Parallel Optical Interconnection Network for Address Transactions in Large-Scale Cache Coherent Symmetric Multiprocessors

Ahmed Louri, *Member, IEEE,* and Avinash Karanth Kodi

***Abstract*—The authors address the primary limitation of bandwidth demands for address transaction in future cache coherent symmetric multiprocessors (SMPs). As a solution, the authors propose a scalable address subnetwork called symmetric multiprocessor network (SYMNET) in which address requests and snoop responses of shared memory multiprocessors are implemented optically. As the address phase of the transaction is linked to the address bandwidth, which is the major bottleneck in SMPs, they focus only on the address subnetwork in this paper. SYMNET has the capability to pipeline address requests from successive processors, which results in increasing the available address bandwidth and lowering the latency of the network. An optical token is implemented to achieve mutual exclusion to the shared channel. This enables collisionless broadcast of multiple address requests. The simultaneous insertion of multiple address requests into the address subnetwork complicates cache coherence. A modified coherence protocol, called COSYM, was introduced to solve the coherence problem. The authors evaluated SYMNET with a subset of Splash-2 benchmarks running from 4–32 processors. Their simulation studies have shown 10%–67% improvement in execution time for various applications. It is also shown that the average latency for a transaction to complete using COSYM was 85% better than the electrical case. An overview of the proposed optical implementation of SYMNET is presented along with the theoretical power budget and bit-error rate analysis. This analysis shows that SYMNET can scale up to hundreds of processors while still using fast snoopy-based cache coherence protocols and that additional performance gains may be attained with further improvement in optical device technology.**

***Index Terms*—Cache coherence, parallel optical interconnects, scalable optical networks, symmetric multiprocessors (SMPs).**

I. Introduction

SYMMETRIC MULTIPROCESSORS (SMPs) dominate the server market as the most prevalent form of parallel computer commercially available because of the following reasons. First, a global physical address space and a symmetric access to the entire memory space offers increased flexibility and ease of programming. Second, users have legacy applications and are likely to prefer a single system image to manage as SMPs provide the “shared everything” model. Third, SMPs use fast snooping protocols to maintain the caches coherent. In SMPs, each address request is broadcast to all processors/memory modules connected to the network using a shared bus. This address request is *snooped* by all the processors enabling simultaneous update or invalidation of cache blocks, thereby maintaining the caches coherent with low latency. As the number of processors grows in the network, contention to acquire the bus also increases. The evolution of faster processors further aggravates the situation because the shared bus cannot run at speeds comparable to that of faster processors. This is because shared buses running at greater than 100 MHz face some fundamental problems such as wave reflection, impedance mismatch, and parasitic capacitance, which significantly limit the speed improvements [1], [2]. Therefore, the bus speed and the coherence overhead limit the rate at which address requests can be broadcast to all the processors/memory modules connected to the network [3], [4]. This, in turn, limits the number of processors that can share the bus, affecting the scalability of SMP systems [4]. This address rate/bandwidth is the main scaling limit, which cannot follow the increasing demands of faster and larger numbers of processors, limiting the scalability of shared-bus-based SMPs.

In order to increase the address bandwidth, several techniques have been introduced to build high-bandwidth low-latency buses. These techniques include split transaction buses [5], multiple address buses, physically separate address and data subnetworks [4], and moving from physically shared buses to logical buses which are implemented as point-to-point links. For example, each address bus is implemented as a point-to-point link in the Sun Enterprise 10 000 [4], which uses up to four address buses and a data crossbar to scale up to 64 processors. More aggressive solutions using multiple crossbars have been adopted to increase the address bandwidth by using a combination of snooping and directory cache coherence protocols in the FirePlane [6] design from Sun. Directory protocols are more scalable than snooping protocols, since the requests, responses (acknowledgment), and data responses need not be broadcast as in snooping protocols. The drawback of directory protocols is the higher unloaded latency because of the directory indirection and the higher storage overhead required for maintaining the directory information. Hence, new shared-memory architectures [6], [7] have moved away from implementing pure-snooping or pure-directory protocols to hybridization of cache coherence protocols by implementing both snooping and directory protocols within a single architecture model. Therefore, using current electrical technology, we

Manuscript received September 25, 2002; revised January 20, 2003. This work was supported by the National Science Foundation under Grant CCR-0000518.

The authors are with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ-85721 USA (e-mail: louri@ece.arizona.edu).

Digital Object Identifier 10.1109/JSTQE.2003.814189

cannot have a large number of processors and at the same time implement fast pure-snooping cache-coherence protocols to improve the limited address bandwidth in SMPs.

A. Optical Interconnects for Address Bandwidth Limitation

One technology that can provide high communication bandwidth and low latency is optical interconnection technology [8]. The ability to transmit data at 5–10 GHz is possible because optical fibers or waveguides are free from load adaptation, reflection, or capacity problems and are mainly limited by the complexity of the opto-electronic (OE) interfaces. Two unique properties of optics, namely unidirectional propagation and predictable path delays [9], are exploited in this paper to significantly reduce latency and increase the address bandwidth. Optical pulses can coexist on the same optical line without interference if they are sufficiently separated, which is not possible using electrical interconnects. *This enables multiple address requests to propagate within the same waveguide/fiber*. The data transmission rate of a vertical-cavity surface emitting laser (VCSEL) is approximately 3–5 Gb/s. An array of such VCSELs enables address transmission with data rates in excess of 200–300 Gb/s [10]. This could satisfy the bandwidth demands of future SMPs. These advantages provide us the impetus to look at optical technology for developing an interconnection network for scalable SMPs.

B. Our Approach

This paper proposes an integrated solution to solve the address bandwidth requirements of large scalable SMPs and still use fast snooping protocols to maintain cache coherence with low latency using optical technology. An optical address subnetwork called SYMNET using parallel optical interconnects is proposed using one-to-many communication using a single wavelength. Parallel optical interconnects provide higher bandwidth–density product as compared to serial interconnects which provide higher bandwidth–distance product. An optical token is implemented to achieve mutual exclusion to the shared channel. The simultaneous insertion of multiple address requests complicates cache coherence. We have introduced a modified snooping coherence protocol, called COSYM, and verified its correctness. The use of transient states is not a new concept as it has been widely documented, but the transient states in our architecture is used to solve the write atomicity along with the snoop response requirements. COSYM relies completely on snooping protocols and avoids directory protocols altogether. COSYM is compared against electrical-bus-based networks using Splash-2 benchmarks [11]. Our simulation studies have shown a 10%–67% improvement in execution time for various applications. It is also shown that the average latency for a transaction to complete using COSYM was 85% better than the electrical case. SYMNET can scale up to 128 processors which is based on theoretical power budget and BER analysis using current optical technology. Greater scalability can be expected with improvement in optical device technology.

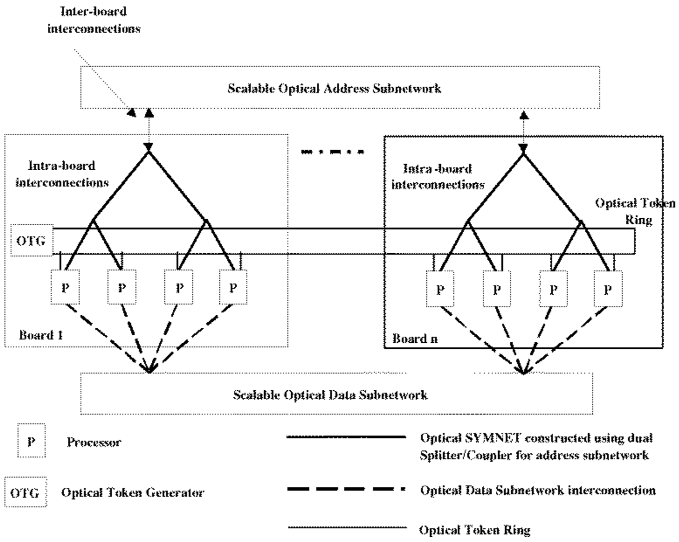
Inter-board interconnections
Scalable Optical Address Subnetwork
Intra-board interconnections
Intra-board interconnections
Optical Token Ring
OTG
P
P
P
P
P
P
P
P
Board 1
Board n
Scalable Optical Data Subnetwork
P Processor
OTG Optical Token Generator
Optical SYMNET constructed using dual Splitter/Coupler for address subnetwork
Optical Data Subnetwork interconnection
Optical Token Ring

Fig. 1. Proposed optical symmetric multiprocessor network (SYMNET) in which the processors are connected to two subnetworks: address subnetwork and data subnetwork.

C. Related Work

Optical-bus-based multiprocessor systems using coincident pulse technique provided optical solutions to the problems of bus design in areas of data transfer, bus arbitration, and device addressing [12]. The photobus smart pixel interconnection system for shared-memory multiprocessors used optical buses for address requests and broadcasts, but arbitration was implemented using electronic buses leading to buffering of address requests at the smart pixel very large-scale integration (VLSI) chip [13]. The constraints of access arbitration are eliminated in the U-bus [3] design for SMPs, which extends the address bandwidth, but a new coherence protocol has to be designed to maintain consistency across the caches. In the SPEED [14] architecture, write requests are broadcast using the snooping protocol and read requests are unicast using the directory protocol. Lightning network [15] uses directory cache coherence protocols in which all transactions are completed in a single hop and is constructed as a tree configuration with wavelength partitioner at each level of the tree. Optical networks discussed so far employ serial links to transmit address requests, address responses and data responses between the source, and the destination using wavelength division multipexing (WDM) technology. Second, directory cache coherence protocols are used to maintain the coherency, which again increase the latency because of the directory indirection. The optical solutions so far have not been able to integrate fast snooping cache coherence protocols and improve the address bandwidth demands to scale the architecture significantly.

II. Architectural Overview

The proposed optical symmetric multiprocessor network SYMNET is shown in Fig. 1. It consists of the processing elements/memory modules and an interconnection network, which in turn consists of two subnetworks (address and data subnetworks). The address and data subnetworks are separated, reducing the design complexity and enabling the design of large scalable SMPs. Scalable data subnetworks have been studied elsewhere [16]; therefore, this paper focuses only on scalable address subnetwork. In what follows, we describe the SYMNET address subnetwork and then explain how the architecture is implemented.

The address subnetwork follows a two-level hierarchical architecture design. The first level consists of grouping a few processors on the boards using intraboard interconnections and the second level consists of interconnecting these boards by using interboard interconnections. The interboard and intraboard interconnections are constructed using dual Y-splitter/coupler combination to form the address subnetwork. Time-division multiple access (TDMA) protocol is used as a control mechanism to achieve mutually exclusive access to the shared channel. Several TDMA protocols such as preallocation-based protocols, reservation-based protocols with preallocated reservation control, and token-based TDMA protocols have been reported [15], [17], [18]. In this paper, we consider an optical token-based TDMA protocol with preallocation to prevent collision of address requests. The optical token is atmost with one processor, which ensures that address request inserted into the interconnect does not collide with any other request already propagating through the interconnect. The address requests from successive processors are pipelined, which allows multiple requests to be propagating through the address subnetwork. It must be noted that multiple address requests propagating through the same address subnetwork is a unique feature of the proposed architecture in contrast to all electrical shared-bus solutions where only a single address request is issued at a time due to the serial nature of the bus. These address requests move up the hierarchy and then are retransmitted back to all processors and memory simultaneously, which ensures serialization of requests.

The basic building block of the SYMNET address subnetwork is shown in Fig. 2. The address subnetwork is constructed using dual Y-couplers and provides two-way address transmission (see the inset in Fig. 2). The up-stream Y-couplers are used for combining the address requests from the processors. After reaching higher levels, this address request is rerouted through the downstream Y-splitters. This enables broadcasting of the address requests to all the processors and memory modules. The optical token is a single pulse generated by a high-frequency (10 GHz) high-powered token generator. It provides a time reference for insertion of address requests into the subnetwork by each individual processor. The optical token is tapped by the processor and this triggers the electronic interface to drive the address request. The token is delayed by using a delay element and provides sufficient time to drive the electronics and also ensures that the address requests from successive processors are transmitted without collision.

The optical clock and the token generator are synchronized; thus, successive processors receive the token every clock cycle. As shown in Fig. 2, in cycle 1 indicated by square shape, the optical token is received by processor 1, which transmits an address request. During cycle 2, when this address from processor 1 is in propagation at the next level of the address subnetwork, the token is received by processor 2, which can transmit an address request, and is shown in the shape of a diamond. In cycle 3, the address request from processor 1 is being rerouted using the downward Y-splitter and at the same time the address request from processor 2 has moved up the address subnetwork. The optical token is now received by processor 3, which can transmit an address. In cycle 4, the address request from processor 1 has reached all the processors, thereby the address request is broadcast to all the processors simultaneously. In cycle 5, the address request from processor 2 has reached all the processors. Broadcasting the address request results in simultaneous reception of the request by all the processors/memory modules enabling snooping of the same request, after which appropriate coherence action is taken as dictated by the snoopy protocols.

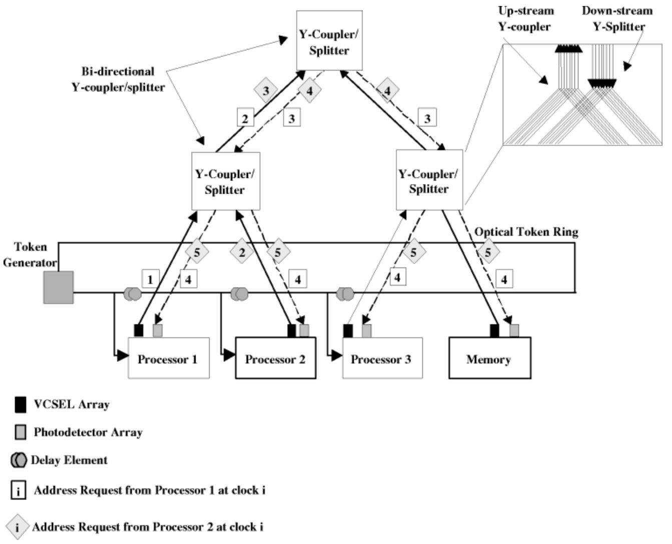

Fig. 2. Overview of a single board of SYMNET, which shows the interconnections for four processors. Processors are connected to dual Y-couplers by optical waveguides/fibers.

TDMA protocols are useful in optical interconnection where high-communication bandwidth can be exploited. Optical pulse transmitted on a fiber or waveguide have two distinct properties, namely unidirectional propagation and predictable propagation delays [9], that are not shared by electronic signal transmission. As a result, fibers or waveguides support pipelining, i.e., at any instance, several addresses encoded as light pulses can be traveling down a fiber, one address request behind the other. In an optical multiprocessor, connected to a fiber, relationships between the spatial and temporal positions of the transmitted pulses can be established [12]. If two processors transmit a pulse having the same spatial extension at two different times, the difference between the arrival times of these two pulses at any checkpoint is equal to the propagation delay between the two processors. In other words, the spatial separation of the two processors determines the temporal separation between the pulses they transmit. The transmission time of the processor is controlled by the optical token which dictates when the processor can insert an address request. Therefore, by controlling the transmission time of the processor, multiple address requests can propagate through the interconnect simultaneously.

III. Optical Implementation

In this section, we analyze a possible implementation of optical SYMNET for address request propagation using parallel optical interconnects such as VCSEL/photodetector arrays, integrated arrays of 2 × 1 y-splitters and 1 × 2 y-couplers, polymer waveguides/ribbons, and integrated semiconductor optical amplifiers.

A. Components Required for SYMNET

- **Parallel Optical Interconnects**: The key component of SYMNET is the VCSEL/photodiode (PD) arrays (transceiver arrays) capable of transmitting at data rates in excess of 3 Gb/s per channel, which results in providing aggregate data rates of several gigabytes per second. The low-cost linear arrays of VCSEL offer a number of advantages over conventional edge emitting laser diodes [19]. High-performance GaAs- and InGaAs-based selectively oxidized or proton implanted top-emitting bottom-emitting VCSEL arrays emitting at 780 to 980 nm have been widely reported in the literature [19]–[22]. Even commercially, several optical component manufacturers such as Xanoptix, TerConnect, Agilent, Corona, Cielo, Emcore, Infineon, and Picolight [10] have developed one- and two-dimensional transceiver arrays with geometries of (1 × 12, 4 × 12, 6 × 12) operating at the lower spectrum of a 850/980-nm wavelength. Therefore, parallel optical interconnects for short distances are viable options for developing an optical interconnection network.
- **Polymer Waveguides/ribbons**: Optical polymers are increasingly considered as highly versatile elements that can be readily transformed into single-mode, multimode, and micro-optical waveguide/fiber structures as they exhibit excellent thermal stability, low optical loss, humidity resistance, low birefringence, flexibility, mechanical robustness and have demonstrated capability in a variety of demanding applications. Acrylate-based polymers, developed by Allied Signals, have shown optical loss less than 0.1 dB/cm at 0.8 μm [23]. The low loss in these polymers makes them an attractive material for constructing the 2 × 1 couplers, 1 × 2 splitters, and for routing optical pulses from VCSELs to these couplers/splitters in the SYMNET interconnection network. Multimode waveguides with a dimension of $65 \times 65 \mu$m have been fabricated using laser direct patterning with an excellent sidewall control. The ability of this polymer to be fabricated on a variety of substrates makes it suitable to be directly interfaced to micro-optical elements such as micro-optical mirrors, 45° micro-reflectors, micro-optical lenses, and also to fiber-to-waveguide interconnect structures. Optical waveguides constructed using other techniques such as photolithographic techniques, reactive ion etching, excimer laser ablation, molding, and embossing techniques have been reported [23]. Polymer waveguides can be machined using excimer lasers to form mechanical structures with a high degree of accuracy to place on the module and can be connectorized to MT-type ferrule packaged with push/pull housing connected to a standard ribbon fiber [24].
- **Arrays of Couplers/Splitters**: The optical pulses from the VCSELs are routed through polymer waveguides to arrays of integrated 2 × 1 couplers. These couplers can easily be constructed using optical polymer waveguides and these couplers are further connected to the next 2 × 1 coupler to construct the address subnetwork [19], [23], [25].
- **Semiconductor Optical Amplifier**: An oxide confined vertical cavity semiconductor optical amplifier, with a gain of 20 dB, fast response time of 60 ps, and narrow amplification bandwidth at 980 nm, has been reported [26]. Arrays of semiconductor optical amplifiers (SOA) are used only at the root of the network to offset the losses caused by the Y-splitter/coupler combination.

B. Technology for Integrating Electronic and Photonic Components

Optical interconnects based on complimentary-metal-oxide-semiconductor (CMOS)/VCSEL technology have been widely proposed for high-performance computing applications [20], [27], [28]. The approach followed in our design is the most widely used hybrid integration using flip-chip bonding of OE-VLSI components [20], [21]. The VCSEL/PD arrays can be fabricated on a GaAs substrate such that the devices are designed to be back-side emitting because of the desire to flip-chip bond them to CMOS driver circuits. The n-contact and p-contact should then be on the top surface of the wafer to facilitate electrical connectivity with CMOS circuits. The GaAs substrate can then be selectively etched leaving the VCSEL/PD contact pad on the back side of the wafer and all optical sources/detectors on the other side of the wafer. The VCSELs and PDs can now be integrated onto the CMOS driver using flip-chip bonding and substrate removal techniques [27], [28]. The passive alignment of VCSELs to waveguides by placing the alignment pedestals on the assembly surface at the locations in reference to the OE component fiducial marks have been demonstrated, which showed efficient coupling between a waveguide and VCSEL array [29]. VCSEL–waveguide coupling using 45° mirrors has also been demonstrated [30], where the mirror loss was estimated to be 0.2 to 0.8 dB at 0.83 μm.

A possible CMOS–VCSEL–waveguide integration is possible by feeding the address request to the CMOS transmitter IC through the address port controller, which, in turn, drives the VCSELs. The direction of light launched into the waveguide is changed by 90° with a 45° mirror. The address pulses propagate in the waveguides through different levels of couplers and splitters. At the receiving end of the waveguides, the direction of light pulse is once again changed by 90° using the 45° mirrors. The light pulses are detected by the photodetector after amplification and returned back to the address port controller from the CMOS receiver IC for processing the received address request. Fig. 3 shows the overview of connections between two adjacent processors n and $n + 1$. The processor's request for an address in case of a miss is forwarded to the cache controller. The cache controller forwards the request to the address port controller, which will buffer it until the optical token arrives. When the optical token is received by the address port, it forwards the address request to the transmitter IC which

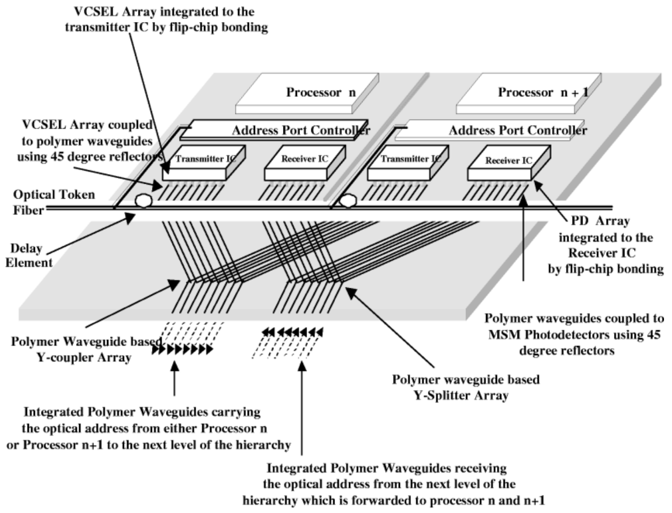
VCSEL Array integrated to the transmitter IC by flip-chip bonding
Processor n
Processor n + 1
Address Port Controller
Address Port Controller
VCSEL Array coupled to polymer waveguides using 45 degree reflectors
Transmitter IC
Receiver IC
Transmitter IC
Receiver IC
Optical Token Fiber
Delay Element
PD Array integrated to the Receiver IC by flip-chip bonding
Polymer waveguides coupled to MSM Photodetectors using 45 degree reflectors
Polymer Waveguide based Y-coupler Array
Polymer waveguide based Y-Splitter Array
Integrated Polymer Waveguides carrying the optical address from either Processor n or Processor n+1 to the next level of the hierarchy
Integrated Polymer Waveguides receiving the optical address from the next level of the hierarchy which is forwarded to processor n and n+1

Fig. 3. Interconnections between two processors having an array of VCSELs and detectors. Polymer waveguides are used to route the address requests from the processors which are fed to a Y-coupler. Y-splitter feeds the return signal back to the individual processors.

drives the VCSEL arrays. The address request encoded as light pulses propagates through 2×1 up-stream couplers, reaches the root, and is transmitted back using 1×2 down-stream splitters. When the address request is received by the receiver IC, it is forwarded to the cache controller through the address port controller. At the intraboard level (level $k = 0$), two to four processor/memory modules are connected using arrays of up-stream y-couplers and down-stream y-splitters. At the next level (level $k = 1$), similar couplers and splitters are connected from different intraboards. At the highest level (level $k = 2$), these up-stream and down-stream couplers are connected to an array of semiconductor optical amplifiers. The optical pulses are routed through waveguides at the intraboard/interboard levels and through polymer fibers between boards. MT Ferrule type push–pull connectors [19] are used for connecting the waveguides and fibers. The addition of newer boards is facilitated by disconnecting the root board (level $k = n$) and inserting another board between level $k = n$ and level $k = n - 1$ and then reconnecting the root board to the newly inserted board. Hence, SYMNET provides an easy addition of boards without significantly altering the existing architecture, which facilitates scalability of the address subnetwork.

IV. Cache Coherence in SYMNET

Coherence and consistency are different aspects of the memory system behavior, both of which are critical for executing correct shared-memory programs [31]. In shared-memory systems, if a processor modifies the shared memory location, then this information should be propagated to all caches existing in the multiprocessor system either by invalidating or updating the shared location; thus, every read should return the latest write to it [32]. Write propagation and serialization are the two requirements of coherence; program order and write atomicity are the two requirements of consistency [33]. We discuss how these two requirements are satisfied in SYMNET.

In SYMNET, each processor inserts an address request upon receiving the optical token. This address request reaches all the processors/memory modules simultaneously as explained in Section III. The processors/memory modules snoop on the address request, make state transitions, and respond by giving the snoop response. Hence, the first criteria of coherence that write requests be propagated to all processors is satisfied since at any request read/write is broadcast to all processors. All processors make requests in program order, hence program order criteria for consistency is also satisfied. The write atomicity requirement of memory consistency extends the write serialization required by coherence [31]. The optical token can be atmost with one processor at any given clock cycle. This ensures mutual exclusion to the shared channel and, hence, address requests from different processors can be injected into the network at different clock cycles. The token propagation maintains a total serial order among all operations from different processors to the same location, therefore satisfies the atomicity requirement.

We begin with a write-invalidate modified, owned, exclusive, shared, invalid (MOESI) [34] cache coherence protocol as the base coherence protocol and modify it accordingly to suit our requirements. When a new request is issued, the memory controller should know whether to supply the data or not, and if it is a read request, then the requesting processor should know whether to load the data in E or S state. In an electrical bus, snoop responses are implemented using two wired-OR lines, *shared* and *owned*. The processors sharing the block assert the shared line if the block is in the shared state or the owned line if the block is in any of the following states: E, M, or O. The shared snoop line could be asserted by more than one processor. In an optically interconnected multiprocessor system, if more than one pulse is inserted into the network as snoop response by multiple processors, collision of snoop responses from several processors results in erroneous response being received by the requestor as they operate at a single wavelength. Therefore, the constraint for the snoop response in our architecture is that it should be *a single response from a single processor*. To achieve this, we maintain an owner for every cache block shared, which is responsible for providing the snoop response. In case of a dirty block, the owner is the most recent processor which wrote to that block. In case of reads, if the block is clean, then there could be several processors sharing the block. In order to determine a single owner the protocol is modified such that if a read-only request is issued to an E block, the block is upgraded to O, instead of S, thereby becoming the owner of the block. Hence, a single snoop response (HIGH or LOW) can determine all the relevant information required such as the following.

- *Snoop High*: Dirty block exists, memory need not respond to the requestor, and if it was for a read request, the block is loaded in S state.
- *Snoop Low*: No dirty block exists, memory responds with the data to the requestor, and if it was for a read request the block is loaded in E state.

In the COSYM protocol, cache controllers react to all kinds of transactions such as read, write, and write-back requests and change states accordingly. COSYM handles write-backs differently when compared to MOESI protocol. The snoop response is always provided by the owner of the block. When the owned block is replaced, there are caches that share the block in the

TABLE I
BENCHMARKS

Benchmark	Description of Application	Input Data Set
FFT	Complex 1-D radix-$\sqrt{n}$ 6-step FFT	64K points
LU	Blocked dense matrix LU factorization	256 × 256, 16 × 16 blocks
RADIX	Integer radix sort	1M integers, radix 1024
OCEAN	Simulates large scale ocean movements	130 × 130 ocean
CHOLESKY	Blocked sparse matrix Cholesky factorization	tk16.0
WATER-NQ	Quadratic-time simulation of water molecules	512 molecules

system. Therefore, the ownership of the block has to be transferred to the next processor sharing the block. In order to do so, each cache block in addition to tag, address bits, and cache state, maintains *the next sharer for the block*, if the block is in either an Owned or Shared state only. If the block is in E or M state, then it is the owner and, therefore, there is no next sharer for the block. This results in an owned block being written back if the block is unshared. The details of the COSYM protocol and the verification of its correctness are a subject of a different publication [35].

V. Performance Evaluation

A. Simulation Methodology and Architectural Assumptions

We have chosen Limes [36] (Linux Memory Simulator) which is an event-driven execution simulator to evaluate the performance of SYMNET with electrical-bus-based networks considering realistic delays for address and data transactions. Limes models a single-level cache and a blocking bus. We have extended the simulator to implement a two-level cache with a split transaction bus by merging or delaying conflicting requests for the electrical system. The interconnection network consisting from 4 to 32 processors was simulated to evaluate the performance of SYMNET. We used as workloads six applications from the SPLASH2 suite benchmarks, namely FFT, LU, Ocean, Radix, Water-nsquared, and Cholesky which are described in Table I. Throughout this evaluation, we have considered processor clock cycles (pcc) as the base time unit for all measurements. Each node of the simulated network contains a 1-GHz processor and 16-KByte direct mapped first level (L1) cache blocks with 32 bytes block size. The second level cache (L2) is a 64 KByte, four-way set associative with 32-byte block size. All instructions and first-level cache reads/writes hits are assumed to take 1 pcc. First-level cache misses stall the processor until the request is satisfied. L1 to L2 line size is assumed to be 8 bytes, which will result in four cycles for data transfer between the two levels.

In the Gigaplane architecture from Sun [4], the electrical bus runs at 83.3 MHz (12 ns), and it takes two cycles to broadcast a single address request. With a 1-GHz simulated processor, it takes 24 pcc to broadcast a single address request in our simulated address bus. The number of cycles required for data transfer is fixed at two electrical network cycles, irrespective of whether the memory or some cache responds as in Gigaplane [4] design. This results in 24 pcc for data transfer in our simulated electrical system. In SYMNET, the optical token is implemented such that the optical signal, generated by a high-frequency (10-GHz) laser source, is split as shown in Fig. 2. One part of the optical signal is detected by the address port controller and the other part is delayed at the delay element implemented using a fiber loop. Now, we calculate the delay D in transmitting an address request into the address subnetwork by the address port controller. This delay should account for the signal detection, OE conversion, and the rise time of address pulses driven by VCSEL arrays. The delay D is given by

$$D = \frac{S_p}{v_c} + 2.O_e + G_d + \frac{b}{m.V_d} \tag{1}$$

where S_p is the distance of separation between the delay element at processor n and the detector at processor $n+1$, v_c is the velocity of light in fibers, O_e is the latency of OE conversion, G_d is the gate delay faced by the token at the address port controller, m is the number of parallel links, b is the number of address bits (including one bit for snoop response), and V_d is the VCSEL data rate. O-E conversion takes place when the optical signal is detected by the address port controller and E-O conversion takes place when the address bits encoded as optical pulses are driven by the VCSEL array. It is assumed that a single gate delay is seen by the address port controller when it receives the token. Assuming that $S_p = 4$ cm, $v_c = 2 \times 10^8$, $O_e = 75$ ps, $G_d = 0.2$ns, $V_d = 3$ Gb/s and with $m = b, D$ is estimated to be 0.88 ns. The optical token should be seen by the next processor with a delay greater than 0.88 ns to prevent collision of address requests. Therefore, the other part of the optical signal at the delay element should be delayed by more than 0.88 ns. Adding guard time to D, we assume the delay to be 1 ns. Considering 1 ns as the required delay, we can estimate the length of the delay element to be 20 cm$(= (2 \times 10^8) \times 1\ \text{ns})$. The delay element is implemented by using a fiber loop 20 cm in length. Therefore, the time taken by each processor to insert its address request is estimated to be 1 ns or 1 pcc. The delay encountered by a address transaction to be visible is equivalent to the number of stages in the address subnetwork. This is assumed to be twice the logarithm of the number of processors connected in the address subnetwork. The snoop response also takes a similar number of cycles. The delay in data transfer for the optical network depends on the data rate of current multiwavelength VCSEL arrays. At 5-Gb/s VCSEL data rate, to transmit a 32-byte block, it takes around 52 ns$(= (32 \times 8)/(5 \times 10^9))$ and this corresponds to 52 pcc. The data network considered for all simulation is the SOCN [16] network which is an optical crossbar constructed using VCSEL/PD arrays and diffraction grating. The number of wavelengths is assumed to be equal

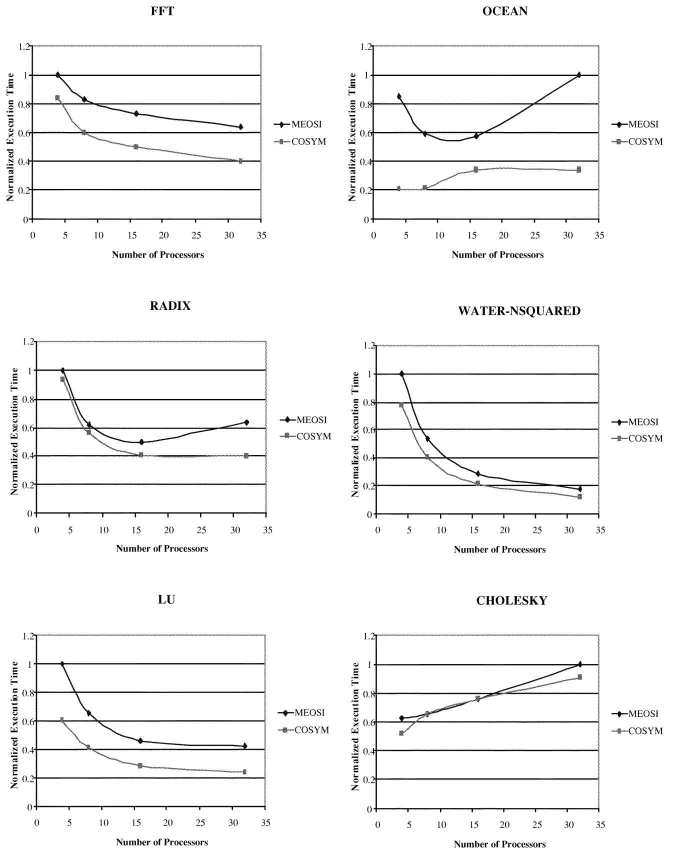

Fig. 4. Normalized execution time for processors varying from 4 to 32 for MOESI and COSYM protocols.

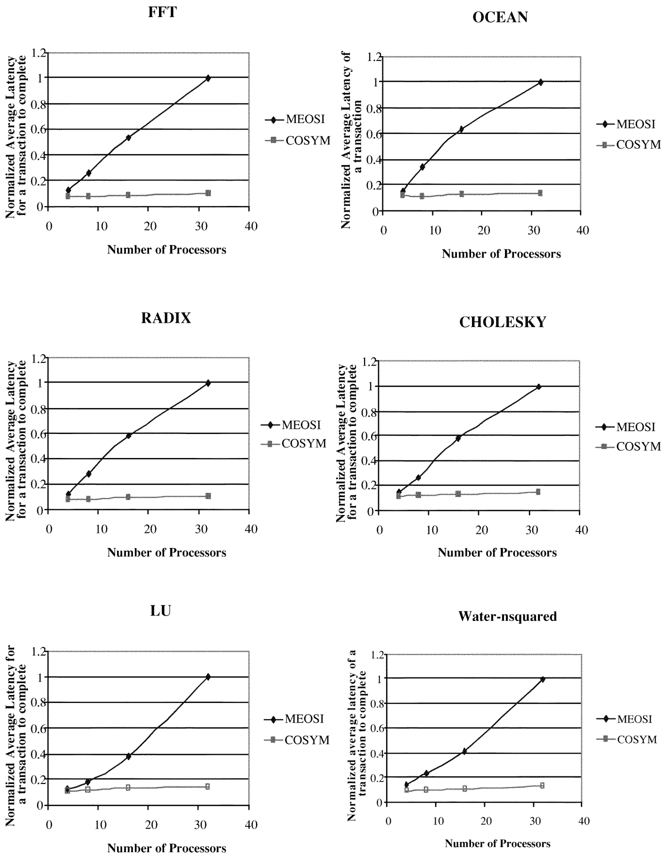

Fig. 5. Normalized average latency of a transaction for processors varying from 4 to 32 for MOESI and COSYM protocols.

to the number of processors, with each processor allocated the wavelength k, such that $k = N \bmod m$, where m is the total number of wavelengths available.

B. Simulation Results

We evaluated two parameters of interest, namely the execution time and the average delay incurred for a transaction to complete.

- **Normalized Execution Time** Fig. 4 shows the normalized execution time for varying number of processors for different applications. Normalized execution time for a given application is calculated by considering the maximum number of simulated cycles for a particular number of processors and dividing the simulated cycles for all the other cases by that value. COSYM shows almost 36% improvement over MOESI protocol for FFT application. For the LU application, the improvement is 43%. For Radix, as the number of processors increases, the difference also slowly expands. For 32 processors, the improvement in performance is 37%. Ocean application consumes much higher bandwidth compared to other Splash-2 applications. The improvement in performance is almost 67% for the COSYM protocol as compared to MEOSI protocol. Cholesky and Water-nsquared applications show lower improvement in performance. Cholesky shows an improvement of 10% and water shows an improvement of 29% for COSYM as compared to MEOSI.
- **Average Latency** Fig. 5 determines the normalized average delay for a transaction to be completed for varying number of processors. The delay in completing a transaction was calculated from the time the address request was received by the L2 cache to the time the data was received by the L2 cache for each processor. The ratio of the total number of transactions to the total time consumed for all processors was used to determine the average delay. Normalized average delay for a given application is calculated by considering the maximum number of simulated cycles for a particular number of processors and dividing the simulated cycles for all the other cases by that value. The average delay was much higher for the electrical case for all applications which increased linearly with the number of processors. This is directly attributed to the saturation of the electrical bus, as the number of processors increases in the interconnect, the delay to acquire the bus also increases, thereby increasing the latency for a transaction to complete. The SYMNET address subnetwork uses statically predetermined time slots for address insertion by different processors. Now, as the number of processors connected in the system increases, token waiting time will also increase, but the effects of such control mechanisms on delay are minimal for two reasons: 1) higher clocking speeds using optics and 2) pipelining successive address requests on the same fiber. Therefore, even though the delay is increasing, as compared to the electrical case, the delay is minimal

and does not dramatically explode as the number of processors increases in the system. The COSYM protocol with a faster address interconnect and a data crossbar provides much better performance for all the cases. The improvement in latency is as high as 85% for radix, fft,and lu using COSYM.

These simulation results clearly indicate that the proposed optical SYMNET with COSYM as the coherence protocol provide much better support for scalable SMPs than their electrical counterparts. We modeled realistic delays for current electrical systems and for our proposed optical interconnect system making our simulation results significant. The execution time is reduced for all applications ranging from 10% improvement for Cholesky to 67% for ocean running for 32 processors for COSYM protocol as compared to the electrical bus network using MEOSI protocol. The average latency reduction is greater than 85% for all the applications using the COSYM protocol.

C. Power Budget and BER Estimation

Calculation of a power budget and the signal-to-noise ratio at the receiver is important for confirming the realizability and scalability of any optical interconnect implementation. The signal-to-noise ratio at the receiver gives an indication of the expected BER of the digital data stream. For optical communication, the standard BER is taken as 10^{-9}, whereas for a parallel computing interconnect, the BER maybe as low as 10^{-15}. The sensitivity of a digital optical receiver is characterized by the average optical power [37] required for producing a digital output with specified BER

If we assume a Ga–As metal-semiconductor FET receiver with a quantum efficiency h of 80%, an FET channel noise factor G of 0.7, an optical pulse-weighting function I_3 of 0.0868, an FET transconductance gm of 30 mS, a total capacitance C_T of 0.75 pF, and a data rate of 10 Gb/s [38], we can achieve a BER of 10^{-15} with an optical power at the receiver of 4.12 μW or -23.844 dBm or -53.844 dB at 98-nm wavelength.

The total optical loss in the system is the sum total of the losses (in decibels) of all optical components that a beam must pass through from the transmitter (VCSEL array) to the receiver (photodetector array). Optical losses are incurred in the following components of the address subnetwork.

1) **VCSEL-waveguide coupling** (L_{vc}): There are losses incurred while coupling the beam emitted from the VCSELs into the optical waveguides using 45° mirrors. It has been estimated that the mirror loss is to be 0.2 to 0.8 dB at 0.83 μm [39]; therefore, we estimate the VCSEL-waveguide loss to be the same region of -0.2 dB.
2) **Fiber/waveguide** (L_f): Polymer optical fibers/waveguides provide low loss for short reach interconnects, with distances of less than 100 m. Acrylate-based polymer fibers, manufactured by Allied Signal, have a loss of 0.02 dB/cm at 840 nm [23]. Assuming the maximum path of the network from the processors to the root of the hierarchy to be 50 cm, since this path is traversed twice, the fiber loss is estimated to be -0.5 dB.
3) **Y-coupler** (L_y): The loss in an Y-coupler/splitter is a fundamental loss of -3 dB for every Y-coupler. The Y-couplers/splitters are used at the board and interboard levels. For n processors, the loss is -3 dB×$\log_2(n)$. This path is traversed twice and, therefore, the total loss is -6 dB ×$\log_2(n)$.
4) **Fiber-waveguide** (L_{fc}): The address pulses are routed using waveguides on the system boards and are further connected to the next levels of the address subnetwork using MT Ferrule push–pull type connectors. Assuming there are two levels of the address subnetwork, the loss in each pair of connectors is 11 dB, and the fiber-waveguide loss is estimated to be -4 dB.
5) **Waveguide cross-over** (L_c): There are excess losses incurred due to the crossings of several waveguides in the address subnetwork. It has been shown that the excess losses with hundreds of crossings using polymer waveguides at 850 nm are 2.2 dB [30]. We assume the same value for the waveguide crossover losses in our interconnect.
6) **Receiver Coupling** (L_{rc}): There will be some loss incurred when the beam is coupled from the polymer fiber onto the optical receiver. This loss is assumed to be -0.2 dB.

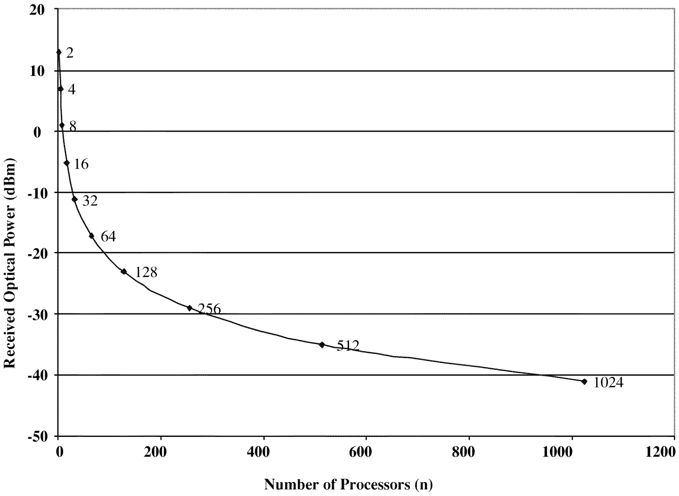

Fig. 6. Received output power in the address subnetwork for varying number of processors. For 128 processors, the received power is -22.57 dBm or -52.57 dB for a BER 10^{-15}

The total transmission loss for the address subnetwork is the total of all the losses

$$\begin{aligned} L_{\text{total}} &= L_{vf} + L_f + L_y + L_{fc} + L_{rc} + L_c \\ &= -6.6\, dB + -6\, dB \times \log_2(n). \end{aligned} \tag{2}$$

The size-dependent parameter of the loss equation is the losses in the Y-coupler. All the other losses are fixed and hence will not vary with the size of the network. High-powered VSCEL arrays delivering output power as high as 4 mE and SOA producing a gain of 20 dB at 980 nm has been reported [10], [26]. For SYMNET, we consider a VCSEL array, which can deliver power of 4 mW at 980 nm. For such values, it is possible to calculate the received power by varying the number of processors connected in the address subnetwork. For 128 processors, the received power is -52.57 dB or -22.57 dBm for a BER 10^{-15}, as shown in Fig. 6. This is significant considering the largest pure-snoopy electrical SMP can support

64 processors [4]. Greater scalability can be expected with further improvement in optical device technology.

VI. Conclusion

In this paper, we addressed the primary limitation of address bandwidth in SMPs. As a solution, we propose a parallel optical interconnect-based SYMNET and a modified cache coherence protocol called COSYM. Using the modified Limes simulator, we simulated the SYMNET using COSYM cache protocol from 4 to 32 processors. The improvement in execution time was seen for all applications ranging from 10% for cholesky to 67% for ocean. The average latency for the transaction also decreased by as much as 85% for various applications of the Splash-2 benchmarks. Using theoretical power analysis, we have shown significant improvement in terms of scalability of SYMNET, as compared to current SMPs. This network architecture provides distinct performance and cost advantages over traditional electronics interconnect and even over other optical networks.

Moore's law expects that the computing speed of electronic systems available at a given cost will increase exponentially. Sun MicroSystems has increased its broadcast-coherency bandwidth from 0.08 to 9.6 GB/s over the last ten years, a scaling rate about the same as Moore's Law (2 × every 18 mo). Optical systems built today are clocked well over 2–5 GHz. It has been seen that short optical clock pulses (100ps-10fs) can be generated and will be available commercially in the future. Improvement in data rates by generating short pulses will increase the available bandwidth for address transactions. It is expected that execution time/latency will further decrease with improved optical device technology and can provide better performance than electronic systems in the future. The available address bandwidth can be further increased by using techniques such as WDM. Each wavelength will be controlled by a separate optical token, which propagates through the token ring. Different tokens are assigned to different memory address spaces and transmission to a particular address space is issued at the wavelength assigned to that address space. These techniques are being pursued to further improve the performance of SYMNET in terms of bandwidth and latency.

Acknowledgment

The authors would like to thank Prof. D. Litaize and Prof. J. H. Collet for their interaction and helpful comments on this paper.

References

[1] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, pp. 728–749, June 2000.

[2] H.S. Stone and J. Cocke, "Computer architecture in the 1990s," *IEEE Comput.*, vol. 24, pp. 30–38, Jan.–Feb. 1991.

[3] J. H. Collet, W. Hlayhel, and D. Litaize, "Parallel optical interconnects may reduce the communication bottleneck in symmetric multiprocessors," *Appl. Opt.*, vol. 40, pp. 3371–3378, 2001.

[4] A. Charlesworth, "Starfire: Extending the smp envelope," *IEEE Micro.*, vol. 18, pp. 39–49, Jan.–Feb. 1998.

[5] M. Galles and E. Williams, "Performance optimizations, implementation and verification of the SGI challenge multiprocessor," in *Proc. 27th Annu. Hawaii Int. Conf. Systems Sciences*, 1996, pp. 134–143.

[6] A. Charlesworth, "The sun fireplane SMP interconnect in the sunfire 3800–6800," *Hot Interconnects 9*, pp. 37–42, Aug. 2001.

[7] D. J. Sorin, M. Plakal, A. E. Condon, M. D. Hill, M. M. K. Martin, and D. A. Wood, "Specifying and verifying a broadcast and a multicast snooping cache coherence protocol," *IEEE Trans. Parallel Distrib. Syst.*, vol. 13, pp. 556–578, June 2002.

[8] J. H. Collet, D. Litaize, J. V. Campenhut, C. Jesshope, M. Desmulliez, H. Thienpont, J. Goodman, and A. Louri, "Architectural approaches to the role of optics in mono and multiprocessor machines," *Appl. Opt., Special Issue Opt. Computing*, vol. 39, pp. 671–682, 2000.

[9] C. Qiao and R. G. Melhem, "Time-division optical communications in multiprocessor array," *IEEE Trans. Comput.*, vol. 42, pp. 577–590, May 1993.

[10] "Parallel links transform networking equipment," *FiberSystems Int.*, pp. 29–32, Feb. 2002.

[11] C. S. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The splash-2 programs: Characterization and methodological considerations," in *Proc. 22nd Annu. Int. Symp. Computer Architecture*, June 1995, pp. 24–37.

[12] D. M. Chiarulli, S. P. Levitan, R. G. Melhem, M. Bidnurkar, R. Ditmore, G. Gravenstreter, Z. Guo, C. Qiao, M. F. Sakr, and J. P. Teza, "Optoelectronic buses for high-performance computing," *Proc. IEEE*, vol. 82, pp. 1701–1710, Nov. 1994.

[13] P. Lukowicz, "The photobus smart pixel interconnection system for symmetric multiprocessing using workstation clusters," in *Proc. 6th Int. Conf. Parallel Interconnects*, 1999, pp. 106–113.

[14] J.-H. Ha and T. M. Pinkston, "The speed cache coherence for an optical multi-access interconnect architecture," in *Proc. 2nd Int. Conf. Massively Parallel Processing Using Optical Interconnections*, 1995, pp. 98–107.

[15] P. Dowd, J. Perreault, J. Chu, D. C. Hoffmeister, R. Minnich, D. Burns, F. Hady, Y. J. Chen, and M. Dagenais, "Lightning network and systems architecture," *J. Lightwave Technol.*, vol. 14, pp. 1371–1387, 1996.

[16] B. Webb and A. Louri, "A class of highly scalable optical crossbar-connected interconnection networks (socns) for parallel computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, pp. 444–458, May 2000.

[17] K. Bogineni and P. W. Dowd, "A collisionless multiple access protocol for wavelength division multiplexed star-coupled configuration: Architecture and performance analysis," *J. Lightwave Technol.*, vol. 10, pp. 1688–1699, 1992.

[18] J.-H. Ha and T. M. Pinkston, "A new token-based channel access protocol for wavelength division multiplexed multiprocessor interconnects," *J. Parallel Distrib. Comput.*, vol. 60, no. 2, pp. 169–188, Feb. 2000.

[19] Y. S. Liu, R. J. Wojnarowski, W. A. Hennessy, P. A. Piacente, J. Rowlette, J. Stack, M. Kader-Kallen, Y. Liu, A. Peczalski, A. Nahata, and J. Yardley, "Plastic vessel array packaging and high density polymer waveguides for board and backplane optical interconnect," in *Proc. Electron. Components Technol. Conf.*, 1998, pp. 999–1005.

[20] D. V. Plant, M. B. Venditi, E.Emmanuelle Laprise, J. Faucher, K. Razavi, M. Chateauneuf, A. G. Kirk, and J. S. Ahearn, "256-channel bidirectional optical interconnects using vessels and photodiodes on cmos," *J. Lightwave Technol.*, vol. 19, pp. 1093–1103, 2001.

[21] A. V. Krishnamoorthy and K. W. Goossen, "Optoelectronic-vlsi: Photonic integrated with vlsi circuits," *IEEE J. Select. Topics Quantum Electron.*, vol. 4, pp. 899–912, Nov.-Dec. 1998.

[22] R. Pu, C. Duan, and C. W. Wilmsen, "Hybrid integration of vcsels to cmos integrated circuits," *IEEE J. Select. Topics Quantum Electron.*, vol. 5, pp. 201–208, Mar.-Apr. 1999.

[23] L. Eldada and L. W. Shacklette, "Advances in polymer integrated optics," *IEEE J. Select. Topics Quantum Electron.*, vol. 6, pp. 54–68, Jan.-Feb. 2000.

[24] Y. Liu, "Heterogeneous integration of oe arrays with si electronics and micro-optics," in *Proc. Electron. Components Technol. Conf.*, 2001, pp. 864–869.

[25] S. S. Saini, Y. Hu, Z. Dilli, R. Grover, M. Dagenais, F. G. Johnson, D. R. Stone, H. Shen, W. Zhou, and J. Pamulapati, "Integrated 1 (2 loss-lessy-junction splitter on a passive active resonant coupler platform," *Lasers Electro-Opt.*, pp. 423–424, 2000.

[26] D. Wiedenmann, B. Moeller, R. Michalzik, and K. J. Ebeling, "Performance characteristics of vertical-cavity semiconductor laser amplifier," *IEE Electron. Lett.*, vol. 20, 1996.

[27] R. Pu, E. M. Hayes, C. W. Wilmsen, K. D. Ohoquette, H. Q. Hou, and K. M. Geib, "Comparison of techniques for bonding vcsels directly to ics," *JOSA*, vol. 1, pp. 324–329, 1999.

[28] H. J. J. Yeh and J. S. Smith, "Integration of gaas vertical cavity surface emitting laser on Si by substrate removal," *Appl. Phys. Lett.*, vol. 64, pp. 1466–1468, 1994.

[29] Y. S. Liu, R. J. Wojnarowski, W. A. Hennessy, J. P. Bristow, Y. Liu, A. Peczalski, J. Rowlette, A. Plotts, J. Stack, M. Kader-Kallen, J. Yardley, L. Eldada, R. M. Osgood, R. Scarmozzino, S. H. Lee, V. Ozgus, and S. Patra, "Polymer optical interconnect technology (point)-optoelectronic for board and backplane applications," in *Proc. Electron. Components Technol. Conf.*, 1996, pp. 308–315.

[30] T. Sakamoto, H. Tsuda, M. Hikita, T. Kagawa, K. Tateno, and C. Amano, "Optical interconnection using vcsels and polymeric waveguide circuits," *J. Lightwave Technol.*, vol. 11, pp. 1487–1492, 2000.

[31] D. E. Culler, J. P. Singh, and A. Gupta, *Parallel Computer Architecture: A Hardware/Software Approach*. San Francisco, CA: Morgan Kaufmann, 1999.

[32] M. Dubois, C. Scheurich, and F. Briggs, "Memory access buffering in multiprocessors," in *Proc. 13th Annu. Int. Symp. Comput. Architecture*, 1986, pp. 434–442.

[33] S. A. Adve and K. Gharachorloo, "Shared-memory consistency models: A tutorial," *IEEE Comput.*, vol. 29, pp. 66–76, Dec. 1996.

[34] P. Sweazey and A. J. Smith, "A class of compatible cache consistency protocols and their support by the IEEE futurebus," in *Proc. 13th Annu. Int. Symp. Comput. Architecture*, May 1986, pp. 414–423.

[35] A. Louri and A. Kodi, "SYMNET: An optical interconnection network for large-scale, high-performance symmetric multiprocessors," *Appl. Opt.*, vol. 42, no. 17, pp. 3407–3417, 2003.

[36] I. Ikodinovic, A. Milenkovic, V. Milutinovic, and D. Magdic, "Limes: a multiprocessor simulation environment for PC platforms," in *PPAM*, Sept. 1999.

[37] Y. Li, T. Wang, and R. A. Linke, "Vcsel-based angle-multiplexed optoelectronic crossbar interconnects," *Appl. Opt.*, vol. 35, pp. 1282–1295, 1995.

[38] T. V. Moui, "Receiver design for high-speed optical fiber systems," *IEEE J. Lightwave Technol.*, vol. LT-2, pp. 234–267, June 1984.

[39] M. Hikita, R. Yoshimura, A. Kaneko, M. Usui, S. Tomaru, and S. Imamura, "Free-standing polymeric optical waveguide films for optical interconnections," in *Proc. ECOC*, Sept. 1997, pp. 285–288.

Ahmed Louri (M'94) received the M.S. degree in computer engineering in 1984 and the Ph.D. degree in computer engineering in 1988, both from the University of Southern California (USC), Los Angeles.

He is currently a Full Professor of electrical and computer engineering at the University of Arizona, Tucson, and the Chairman of the Computer Engineering Program. He is also the Director of the Optical Networking and Parallel Processing Laboratory. His research interests include computer architecture, parallel processing, optical computing systems, and optical interconnection networks. He has published numerous journal and conference articles on the above topics. Prior to joining the University of Arizona, he worked as a Researcher with the Computer Research Institute, USC, where he conducted extensive research in parallel processing, multiprocessor system design, and optical computing.

Dr. Louri has served as a Member of the Technical Program Committee of several conferences including OSA Topical Meetings on Optics in Computing, OSA/IEEE Conference on Massively Parallel Processors using Optical Interconnects, IEEE High-Performance Computer Architecture, and others. He is a member of OSA.

Avinash Karanth Kodi received the B. Engg. degree in electronics and communication from Manipal Institute of Technology (Mangalore University), Manipal, India, in 1999 and the M.S. degree in computer engineering, in 2003, from the University of Arizona, Tucson, where he is currently pursuing the Ph.D. degree.

His research interests include high-speed optical interconnects, parallel processing, and cache coherence protocols.