

# Building Data Centers With Optically Connected Memory

Daniel Brunina, Caroline P. Lai, Ajay S. Garg, and Keren Bergman

**Abstract**—Future data centers will require novel, scalable memory architectures capable of sustaining high bandwidths while still achieving low memory access latencies. Electronic interconnects cannot meet the challenges presented by the need for multi-terabit off-chip memory data paths. In this work, the electronic bus between main memory and its host processor is replaced with a circuit-switched optical interconnection network. We investigate the impact of our optically connected memory system on large-scale architectures and experimentally validate the protocol using field-programmable gate array based processor nodes and a custom-designed memory controller. The processor communicates all-optically with multiple synchronous dynamic random access memory nodes using  $4 \times 2.5$ -Gb/s wavelength-striped payloads, operating error free with bit-error rates less than  $10^{-12}$ .

**Index Terms**—Memory architecture; Optical communication; Photonic switching systems; SDRAM.

## I. INTRODUCTION

The continued performance scaling of large-scale computing systems and data centers is reliant on efficient high-capacity memory architectures [1]. Processors can only operate on data as quickly as the data can be provided by the memory nodes. A system must therefore maximize the ratio of memory bandwidth to processor performance, as measured in bytes per second per floating point operations per second (FLOPS)—a ratio that has become steadily worse with each successive generation of large-scale computing systems [2]. Furthermore, growing data sets and server virtualization are placing demands on the system, limiting the types of applications that are supported.

Large-scale computing systems magnify the limitations of memory scalability. Their extreme scale requires high sustained memory bandwidth and capacity, while simultaneously maintaining low access latency with energy-efficient interconnects. These factors, however, require trade-offs in an electronically interconnected memory system.

The addition of more memory devices to a system can increase both the memory capacity and the memory bandwidth. Main memory technology operates at a fraction of the clock frequency of modern processors. Therefore, an increase in the bandwidth of memory systems requires an increase in the number of memory devices, which are

subsequently accessed in parallel on a multi-drop bus to improve bandwidth. Each new memory device increases the physical wiring distance and capacitive load on the multi-drop memory bus. Additionally, skew limits require the electronic traces to be path-length matched [3]. The wire length and routing complexity combine to limit the minimum memory access latency and maximum bus signaling rate.

The low-bandwidth density of electronic interconnects [4] further limits the maximum memory bus signaling rate. Current synchronous dynamic random access memory (SDRAM) technology performs a quasi-serialization technique [5] to transmit multiple memory words per clock cycle, but is still limited to about 1 GHz signaling [6]. Hundreds of pins may therefore be required to provide the necessary bandwidth between a processor and memory, and future processors may be pin limited when attempting to meet bandwidth requirements [7].

The need to replace the complex, parallel memory bus is evident in industry; the fully buffered dual in-line memory module (FB-DIMM) was presented as an electronic solution to memory speed and density [8]. The goal of FB-DIMM is to alleviate the limitations of electronic interconnect scaling by incorporating serialization/deserialization (SerDes) to operate the data bus at 12 times the memory clock rate, resulting in electronic serial links clocked at several GHz. The traditional parallel multi-drop bus is replaced by 24 point-to-point links, configured as a daisy chain, with the memory controller (MC) only directly communicating to one FB-DIMM. If a memory transaction is not addressed to the first FB-DIMM in the daisy chain, the transaction is passed along to the next FB-DIMM with a latency penalty of approximately 4 ns [3]. Although the use of FB-DIMM increases scalability by improving bandwidth density and thereby reducing pin count, the electronic interconnect will still limit memory systems in large-scale computer systems. Chaining together too many FB-DIMMs in a single channel results in unacceptable memory access times. The addition of too many memory channels will not only overburden the MC but also leads to the same pin count and wiring problems with traditional parallel electronic wiring.

With the growing number of memory devices and the increasing signaling rate of the memory bus, the electronic bus linking main memory to its processor becomes a key bottleneck to system performance. Each memory module added to the system increases the chip pin count and physical wiring distance, thus limiting the transmission data rate while increasing latency, overall wiring complexity, and power dissipation. Current microprocessors dissipate up to half of their energy in the interconnect alone [9], and memory access

Manuscript received February 1, 2011; revised June 12, 2011; accepted June 17, 2011; published July 14, 2011 (Doc. ID 142131).

The authors are with the Department of Electrical Engineering, Columbia University, New York, New York 10027, USA (e-mail: daniel@ee.columbia.edu).

Digital Object Identifier 10.1364/JOCN.3.000A40

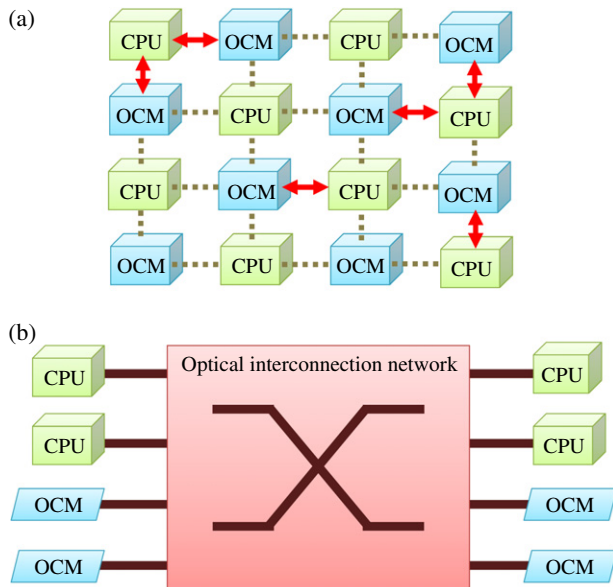


Fig. 1. (Color online) Block diagrams showing (a) processing nodes connected to optically connected memory nodes by optical links (dotted lines), with simultaneous communication between different nodes (arrows), and (b) a subset of the nodes connected by an optical interconnection network.

latencies are currently in the tens of nanoseconds. Due to these trade-offs, electronically connected memory systems cannot scale to meet the requirements of future data centers.

Optical interconnects provide the necessary bandwidth density to create low-latency, energy-efficient communication infrastructures for next-generation computing systems [10,11]. We propose to apply the benefits of optical systems to memory interconnects to create a high-performance, energy-efficient memory system that can meet the scaling challenges presented by data centers. The integration of optics into memory systems allows microprocessors to efficiently access vast amounts of memory, for example, using an optical interconnection network (Fig. 1), to maintain the continued scaling of overall system performance.

In this work, our approach focuses on the interface between processors and memory devices, including the development and experimental demonstration of an optically connected memory (OCM) module and an MC that is able to access remote OCM nodes across an optical interconnection network. We develop the memory access protocol necessary for an optical network-based memory architecture, and analyze the resulting system architecture.

We experimentally demonstrate multiple OCM modules communicating with a microprocessor across a circuit-switched,  $4 \times 4$  optical interconnection network test-bed. We emulate a microprocessor using a field-programmable gate array (FPGA), which interfaces to the optical network using four differential-pair 2.5 Gb/s serial transceivers ( $4 \times 2.5$  Gb/s) that modulate four wavelength channels. The microprocessor contains a custom MC that is optimized for accessing SDRAM across an optical interconnection network. Two OCM modules are created using identical FPGA-based circuit boards that each contain four chips of Micron DDR2 SDRAM and

$4 \times 2.5$ -Gb/s transceivers. The four modulated wavelength channels are combined on a single-mode fiber via wavelength-division multiplexing (WDM) and injected into each network input port using a wavelength-striped format.

## II. RELATED WORK

A large body of research investigates the leveraging of optical interconnects to address the limited scalability of main memory. The work in [12] explores the impact of accessing large banks of remote memory across optical links. The authors focus on symmetric multiprocessing (SMP) systems, and demonstrate that, within large SMP systems, the improved bandwidth from optical links outweighs the increased time-of-flight latency incurred from moving optically attached memory physically away from processing elements.

The authors in [13] present a photonic network-on-chip (NoC) that accesses off-chip main memory, in the form of SDRAM, through a circuit-switched optical interconnect. Circuit-switched, optical access to memory was demonstrated to improve performance and reduce power consumption when compared to electrical interconnects.

OCDIMM [14] is an OCM module based on FB-DIMM, in which silicon photonics realized near the processors and memory modules allow the electronic memory bus to be replaced with waveguides as a shared, wavelength-routed bus. The unique advantage here is improved memory access latency when compared to traditional FB-DIMM. In [15], the authors propose a substantial redesign of memory devices through monolithic integration of silicon photonics. This implementation is shown to reduce power consumption by a factor of 10.

The main contributions of this work are the experimental demonstration of our OCM system, the architectural analysis, and the novel memory access protocol that was required to physically implement the experiment. To the authors' knowledge, this work presents results from the only existing implemented OCM system; the steps required to move beyond simulation provide unique insight into the challenges facing future memory systems.

## III. ELECTRONIC MEMORY ACCESS PROTOCOL

Contemporary memory systems are arranged in a hierarchy that is designed to balance data capacity and access efficiency. At the top of the hierarchy, i.e., closest to the processor, the on-die caches provide access to data at near-processor speeds but must remain limited in size to a few megabytes to minimize latency. At the bottom of the hierarchy, hard disks can support terabytes of data storage per device, as required by data centers, but their low bandwidth and millisecond access times can easily limit overall system performance. The optimal balance between speed and capacity currently lies in main memory (typically SDRAM), which is in the middle of the memory hierarchy.

Current commercial memory modules are packaged as dual in-line memory modules (DIMMs), which contain multiple SDRAM chips and can provide access to gigabytes of

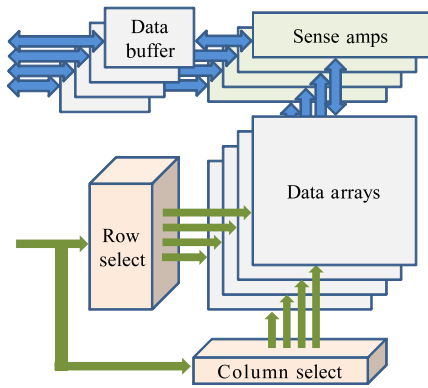


Fig. 2. (Color online) Anatomy of one bank of SDRAM. Four data arrays share common row and column selects for concurrent access.

capacity with over 100 Gb/s memory bandwidth and tens of nanoseconds access time [6]. Each SDRAM chip is comprised of several independent memory banks (Fig. 2), which are accessed independently to allow the MC to pipeline memory accesses to different banks. The main components of a bank are the data buffers, sense amplifiers, and data arrays. The sense amplifiers are the interface between the data arrays and the data buffers, and the data buffers function similarly to a SerDes for the MC-SDRAM data path. At each SDRAM clock cycle, typically 100–300 MHz, the data arrays store multiple kilobits of parallel data in the data buffers. The data buffers then transfer 64-bit data words (or 72-bit if error correction is enabled) to the MC at clock frequencies up to 1 GHz with two transfers per clock cycle. This double pumping of the 64-bit, 1 GHz memory bus allows high-end SDRAM systems to achieve peak transfer rates of over 100 Gb/s.

Access to SDRAM is a multi-step process that requires tens of nanoseconds [5]. All main memory accesses are handled by an MC. The memory devices themselves, typically DIMMs, remain essentially idle in the absence of instructions from the memory controller. The microprocessor issues read-to- or write-from-memory requests to the MC, which translates the requests into a series of SDRAM-specific commands that are reordered to maximize memory bandwidth. Typically, the first command sent is an activate (ACT) instruction along with bank and column address bits. This step causes the sense amps to transfer all bits from the activated row into data buffers (thousands of bits over multiple modules). Next, the MC sends a read (RD) or write (WR) command along with the column address. For a write, the MC will transmit data on the data bus and overwrite any data stored at the corresponding address. A read command will use the column address to select the desired bits from the data buffers, causing the read data to be transmitted back to the MC.

Bandwidth is optimized by reordering SDRAM accesses such that all reads and writes are addressed to the same row within the SDRAM data arrays. Due to the processor-SDRAM performance gap, each access to a new row requires multiple SDRAM clock cycles and hence incurs tens of nanoseconds of latency. Therefore, in addition to reordering memory accesses, the MC also requires the microprocessor to access SDRAM in blocks of eight data words, known as bursts, to guarantee a minimum number of accesses to a row.

#### IV. OPTICALLY CONNECTED MEMORY

We envision a system architecture in which the traditional electronic memory bus interconnect is replaced by an optical interconnection network. As a result of the network configuration, we refer to processor nodes and memory nodes that contain exclusively processors or memory, but not both.

Single-mode optical fibers [16] can provide data centers with communication links that are immune to physical distance. The system must only expend energy to generate and receive the optical signal, which can traverse a rack or potentially the entire data center without added power cost. Each optical link is also agnostic to bit rate, which can allow the transmission of terabits of data on a single fiber. Recent advances in silicon photonics have also enabled efficient coupling [17] for high-bandwidth density, as well as the fabrication of energy-efficient transceivers [18,19].

This work explores the architectural impact of optical interconnects enabling unprecedented growth and flexibility in large-scale computer systems. A photonic transceiver module can aggregate the bandwidth of many individual SDRAM chips, currently up to 20 Gb/s each. Without the limitations imposed by transporting such high bandwidth across long electronic traces, many more SDRAM chips could be used at each OCM node than is possible in electronically connected DIMMs. This also has the benefit of allowing greater memory capacity at each node, which improves system performance by enabling larger or more pages to be stored in main memory. Accessing the OCM nodes across an optical interconnection network further provides greater flexibility in accessing more OCM nodes, either independently or concurrently, thus improving bandwidth and memory capacity compared to point-to-point or multi-drop memory links [20]. The memory protocol design incorporates a circuit-switched optical interconnect [21], which has been shown to provide higher bandwidth density, improved power transmission, and reduction in overall application execution time [13].

Relocation of the memory devices to be physically distant from the processor also frees up board space near the processor itself without significant impact on system performance [12]. This will give system designers more flexibility in designing board layouts: remaining board components, such as on-board routers or caches, can be placed closer to the processor or increased in number. Additionally, more processors can be added in the space formerly occupied by DIMMs and the associated wiring, or, rather than increasing density, the packaging can be made less costly and more easily cooled.

#### V. EXPERIMENTAL DEMONSTRATION AND RESULTS

The experimental system configuration allows us to explore the architectural implications of remote OCM modules on future systems. Replacement of the point-to-point memory link with a network configuration maximizes the scalability of our memory system by greatly expanding the memory address space. The optical network can also add new functionality to the memory system, such as memory multicasting [22], where a single memory transaction can multicast from one processor

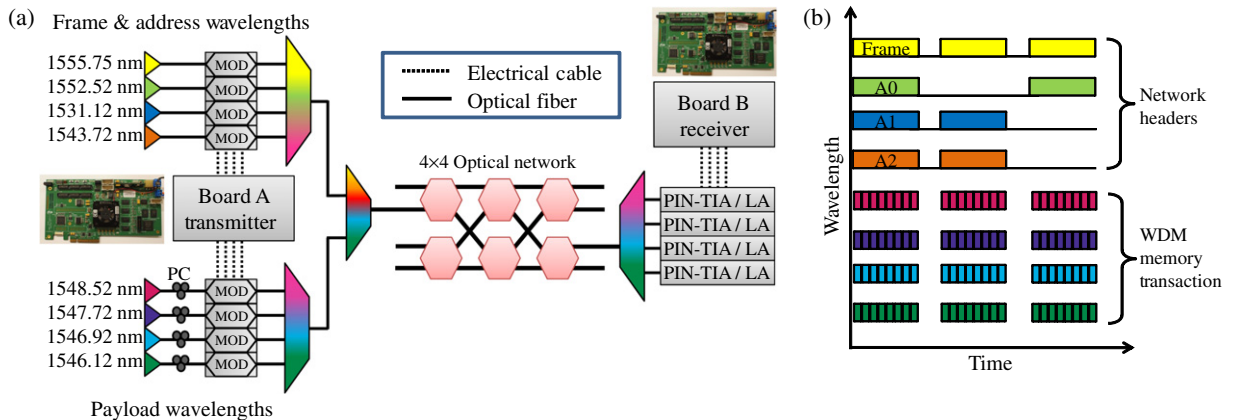


Fig. 3. (Color online) (a) Experimental setup of OCM system illustrating circuit board A modulating eight wavelengths (one frame, three address, and four 2.5 Gb/s payload channels). The wavelengths are combined using WDM and traverse a  $4 \times 4$  optical network before being received by circuit board B. The path from board B to board A is identical. (b) Wavelength-striped optical format with the  $4 \times 2.5$ -Gb/s WDM memory transaction alongside four low-speed network header wavelengths (frame and three address wavelengths).

to multiple memory nodes, or the ability to dynamically reconfigure the memory visible to a given processor.

The experimental OCM system (Fig. 3(a)) is implemented using three FPGA-based circuit boards that communicate all-optically and transparently across a 3-stage optical interconnection network test-bed. Each circuit board contains an Altera Stratix II GX FPGA with  $4 \times 2.5$ -Gb/s electronic transceivers. The transceivers connect to four 10 Gb/s LiNbO<sub>3</sub> modulators to generate wavelength-striped payloads with four wavelengths (1546.12, 1546.92, 1647.72, and 1548.52 nm). General purpose input/output (GPIO) FPGA pins drive four external semiconductor optical amplifiers (SOAs) to modulate the low-speed network control wavelengths (1531.12, 1543.72, 1552.52, and 1555.75 nm). All eight wavelengths (consisting of the 4-channel header and the 4-channel payload) traverse the optical interconnection network test-bed concurrently as a single WDM memory transaction (Fig. 3(b)), and the four payload wavelengths are received at the network output using four 10 Gb/s photodiodes with transimpedance and limiting amplifiers. The four receivers transmit data to the receive-side of the  $4 \times 2.5$ -Gb/s high-speed I/O transceivers, resulting in an aggregate 10 Gb/s memory bandwidth.

### A. Processor Node

Our experimental setup makes use of a single processor node, Fig. 4, which accesses two remote OCM nodes. Here, the FPGA implements the functionality of our emulated processor and custom MC, and all the logic is clocked at 250 MHz. The processor node does not have access to any local SDRAM, and all memory transactions use the  $4 \times 2.5$ -Gb/s memory link. The resulting configuration is a processor node with an on-chip memory controller possessing an aggregate 10 Gb/s memory link. The memory controller handles all overhead necessary for the optical communication, including network arbitration.

**Emulated Processor.** A typical memory system is based on a microprocessor initiating write-to-memory and read-from-memory transactions. Here, the processor functionality is

emulated using a custom memory traffic generator to create programmable, verifiable memory transactions. We assume a parallel programming model that requires sustained high-bandwidth access to memory, such as a streaming application. The resulting memory access pattern is such that most of the application run time involves reads-to- or writes-from-memory.

**Custom Memory Controller.** In order to fully realize the benefits of our OCM system, our MC is designed for efficient control of the optical interconnection network. Therefore, our MC has been implemented with focus on optimizing an optical network aware memory transaction protocol. General MC optimizations, such as intelligent transaction scheduling schemes, are important for both electronically and optically connected memory, but have been investigated elsewhere [23–25] and are not studied here.

A multi-core architecture may require the MC to handle requests from multiple cores, and a high-performance memory system may connect the MC to many independent memory devices. Our implemented MC utilizes the network architecture to enable concurrent accesses from any processor node to any OCM node with minimal access latency.

In order to guarantee a reliable, high-bandwidth memory link, we use a circuit-switched optical interconnection network. The MC is modified to reflect this change, as seen in Fig. 5. The MC manages communication across the network test-bed analogous to the way a standard MC operates an electronic memory bus. All control of OCM devices is from the memory controller, but the key difference here is that the MC establishes MC-to-OCM or OCM-to-MC circuit-switched lightpaths as necessary for write and read operations.

A write-to-memory transaction consists of SDRAM commands and write data streaming from the MC to the appropriate memory node, and thus the MC manages its own lightpath through the network. While the lightpath is being established, the processor must wait before streaming its write data over the network, similar to a processor waiting for a busy memory device in the case of electronically connected memory.

A read-from-memory transaction involves two-way signaling consisting of SDRAM commands sent from the MC to SDRAM

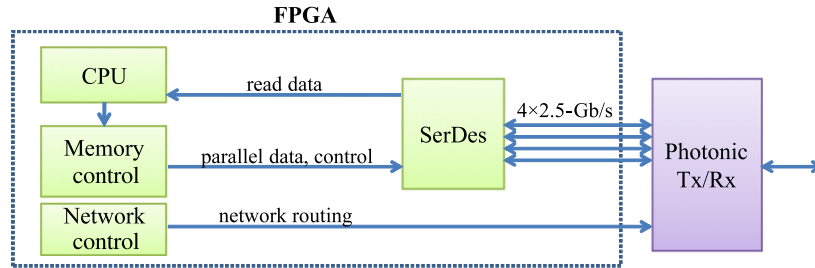


Fig. 4. (Color online) Block diagram of a processor node. An FPGA implements the emulated CPU along with a custom MC (memory control, network control, and SerDes). The high-speed,  $4 \times 2.5$ -Gb/s serial memory transaction combines with the low-speed network routing information at the photonic transceivers using WDM, and traverses the optical interconnection network as a wavelength-striped optical memory transaction.

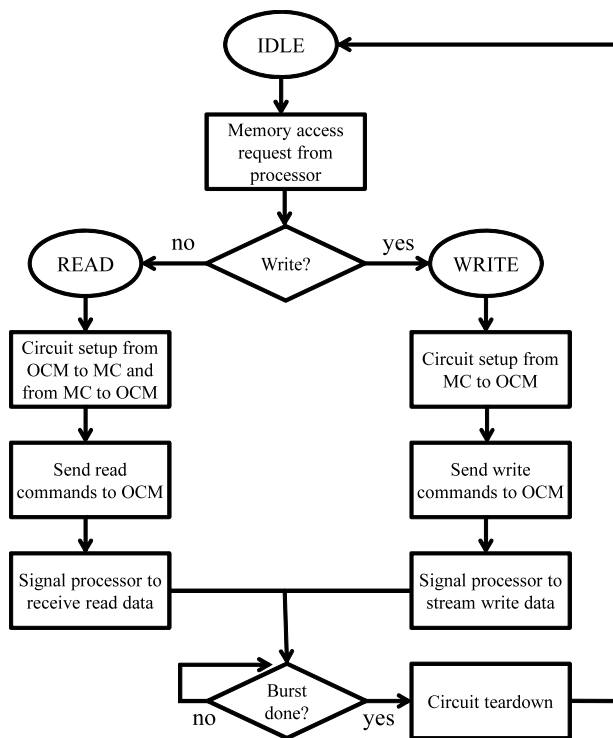


Fig. 5. Flowchart of MC for circuit-switched OCM system.

followed by read data streaming from SDRAM to the MC. In this case, the MC will manage a lightpath from memory back to itself while the read operation is in progress. As in Fig. 5, the first step is for the MC to create lightpaths from the MC to the OCM node and from the remote OCM node back to itself. With a communication link established, the MC sends a small amount of data to memory consisting of read commands and the memory address. The memory controller then tears down the lightpath from itself to the OCM node, while the lightpath from the OCM node to the MC is used to stream read data to the processor. The remaining OCM-to-MC lightpath is torn down upon completion of the burst.

The 10 Gb/s memory link from the MC to memory is optimized by time multiplexing SDRAM control information with write data. In electronically connected memory systems, the control and data are transmitted on dedicated wires. The electronic data bus has much higher utilization and operates at

a higher data rate than the control bus. Due to our serialization of the memory link into four high-speed channels, the dedication of one channel to the low-bandwidth control information would be inefficient. The standard SDRAM protocol [5] specifies that memory will only receive one set of commands for each write (or read) burst. This allows us to begin each OCM write access with all four channels dedicated to control and address information, and subsequently dedicate all four channels to write data. The OCM node will continue accepting write data until the burst has ended, at which point another memory access may begin, as is done with current electronic memory.

A typical SDRAM burst is 8 memory words, and therefore the streaming of a large amount of memory data requires frequent transmission of SDRAM commands from the MC to memory. In data centers, a majority of data traffic may be part of flows of over 100 MB [26]. We have therefore increased the memory burst size to a full SDRAM array row, 1024 words, to reflect the nature of our large-scale programming model and improve memory link utilization without overloading the network. The physical configuration of SDRAM chips at each OCM node results in 32-bit memory words, where two 16-bit SDRAM chips are accessed concurrently, as compared to 64 bits for commercial DIMMs. Each 1024-length burst therefore writes or reads 32 kilobits, thus each optical memory transaction has an approximate duration of 3.3  $\mu$ s. Future implementations will make use of both packet and circuit switching within the optical interconnection network, which will allow for the efficient packet switching of small transactions and circuit switching of larger transactions, as required by the application.

## B. Optically Connected Memory Node

The OCM node operates as a collection of off-the-shelf SDRAM chips connected to a local photonic transceiver, which interfaces between the SDRAM chips and the high data rate optical interconnection network. The transceiver module handles SerDes functionality as well as wavelength striping of the optical memory data.

Figure 6 shows the FPGA-based implementation of our OCM node. We implement the transceiver module using the Stratix II GX FPGA, which interfaces between the commercially available SDRAM chips and the high-speed, circuit-switched optical interconnection network. The FPGA distributes parallel SDRAM control information and write data to all SDRAM chips. Read data are streamed through the

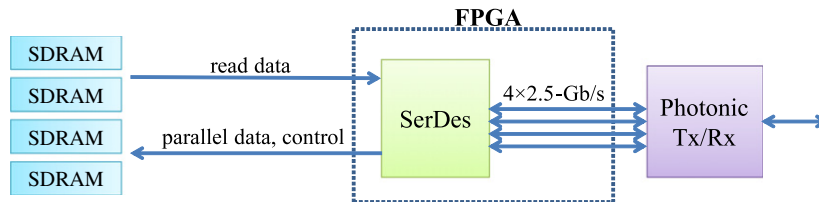


Fig. 6. (Color online) Block diagram of an OCM node. The on-board FPGA acts as a local transceiver, handling SerDes functionality for the SDRAM chips. The  $4 \times 2.5$ -Gb/s serial links are used to send optical read data to the processor node, which is combined using WDM at the photonic Tx/Rx.

FPGA, where they are serialized for striping over four wavelengths. Network control logic is not necessary at the OCM node due to the MC's arbitration of the optical interconnection network. A lightpath from the OCM node to the processor node will already exist when read data are ready, allowing data to stream across the network the moment they become available.

Each OCM node contains 128 MB of SDRAM, and thus the implemented system demonstration contains a total of 256 MB of main memory. The OCM nodes are independently accessible, which enables separate processor nodes to access separate OCM nodes simultaneously across the network. With 10 Gb/s memory bandwidth per OCM node, the aggregate memory bandwidth of our implemented system is 20 Gb/s. In the future, OCM boards can be created with significantly greater capacity and bandwidth.

### C. Optical Interconnection Network

The memory access operations are streamed through a multi-terabit capacity  $4 \times 4$  optical interconnection network test-bed [27]. The test-bed consists of six non-blocking  $2 \times 2$  photonic switching nodes organized as a multistage Banyan network topology. Each  $2 \times 2$  switching node features four SOAs to transparently switch the WDM memory messages. The switching nodes support the wavelength-striped memory message format outlined above, wherein the control header wavelengths are sampled by each routing stage. This requires a minimum number of address wavelengths equal to  $\log_2 N$  for an  $N \times N$  network. The headers are decoded immediately at each switching node using fixed wavelength filters (set to the control wavelengths) and low-speed p-i-n photodetectors. A complex programmable logic device (CPLD) processes the headers, realizes the network routing logic, and makes a lightpath routing decision to gate on the appropriate SOA to route the data stream. Successfully transmitted messages set up end-to-end lightpaths between network ports to support the circuit-switched memory transactions.

In this experimental instantiation, the payload consists of the  $4 \times 2.5$ -Gb/s WDM memory transactions (Fig. 3). The control wavelengths are multiplexed together with the payloads to determine which OCM node is being addressed by each memory access over the optical network test-bed. Here, the numbers of payload and header wavelengths are equivalent due to the limited number of high-speed transceivers on each FPGA board. Future implementations will leverage the broadband nature of the optical network test-bed to increase the number of payload wavelengths, and therefore the memory

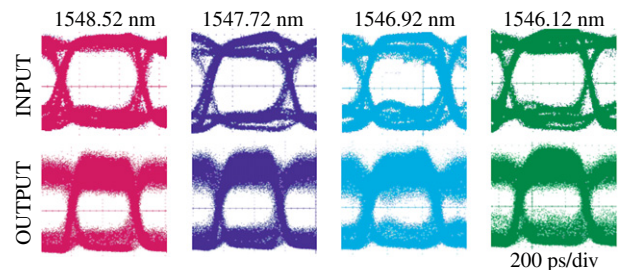


Fig. 7. (Color online) Optical eye diagrams for the  $4 \times 2.5$ -Gb/s memory payload wavelength channels at one network input port (top) and at one network output port (bottom).

bandwidth, while the number of header wavelengths scales efficiently as  $\log_2 N$ . The wavelength-striped format enables the microprocessor to access OCM nodes with time-of-flight latency; this implementation places the processor and memory approximately 24 m apart, and therefore each unidirectional optical transmission requires approximately 120 ns to traverse the network test-bed.

### D. Results

The emulated microprocessor is programmed to iteratively fill all memory addresses with predictable bit patterns—all 1s, all 0s,  $2^{31} - 1$  pseudorandom bit sequence, or a bit pattern corresponding to each destination memory address—and then to read from all memory locations while verifying the received data as they stream in from the optical network. This experiment was performed without error-correction techniques often utilized in large-scale electronic memory systems to allow a counter in the emulated microprocessor to track the number of correctly verified read data using all four test patterns. The counter generates an effective memory-bit-error rate (EMBER) that is used to quantify the functionality and reliability of the implemented OCM system.

Error-free operation of the OCM system is achieved by allowing the microprocessor to correctly verify over one terabit of data from each of the two OCM network nodes, attaining EMBERS less than  $10^{-12}$ . Figure 7 shows optical eye diagrams for the  $4 \times 2.5$ -Gb/s payload channels; these eyes were collected by self-triggering during continuous write-to-memory operations.

The FPGA logic required to access the  $4 \times 2.5$ -Gb/s high-speed FPGA transceivers, including SerDes, adds approximately 100 ns latency to each unidirectional communication.

This additional latency is due to the trade-off between high performance and flexibility presented by the commercial FPGA, which here is operating at a clock frequency of 250 MHz. Future higher performing FPGA implementations may operate at GHz clock speeds; along with tighter hardware integration, the transceiver overhead may then be reduced to a few nanoseconds.

### E. Architectural Analysis and Discussion

The improved latency performance of our optical interconnect approach is a crucial issue to address. Current SDRAM access times are much slower than high-performance processors, thus any additional latency is undesirable. The transparency of our switching nodes reduces the overall latency to the time-of-flight between a processor and OCM. Each additional meter of fiber adds approximately 5 ns [16] of latency to the memory communication path. Though this may be negligible within a single rack, it could become problematic for links spanning a large-scale computing system. With our OCM design, we aim to minimize accesses to memory nodes that are more than a few meters away (similar to the case of today's electronic networks). The main advantage of the OCM system is its sole limitation in distance with regard to time-of-flight latency, whereas electronic systems are limited in distance with regard to latency, power, and bandwidth. Parallel programming models such as PGAS [28] expect a global memory address space but already exploit locality to maximize references to a local memory. In this work, local memory is the memory with the shortest optical path.

Our implemented optical interconnection network enables the bandwidth and latency performance of the OCM system to meet the demands of heavily loaded data centers that exceed hundreds of thousands of nodes [27,29]. The aggregate optical memory bandwidth in this experiment is limited to 10 Gb/s by the FPGA's electronic transceivers. High-end transceivers operating at higher data rates can leverage the available optical bandwidth, as demonstrated by the 10 Gb/s and 25 Gb/s channels in the recent 40- and 100-gigabit Ethernet standards [30]. The use of multiple high-speed transceivers with WDM creates the bandwidth density necessary for OCM nodes with memory bandwidth in the hundreds of gigabits per second. Many OCM nodes could be further combined using an optical interconnection network for petabit system bandwidths.

The reconfigurability of our OCM system supports the diverse applications that data centers may run. For example, a streaming application [31] typically has a predictable traffic pattern with long, sustained memory accesses. The system can be configured to allocate OCM nodes to the appropriate processing nodes before run time, or between memory access stages, which will eliminate the latency associated with circuit-switched lightpath setup.

The network nodes can be configured to support wavelength-stripped payload multicasting to enable access to multiple OCM nodes simultaneously [22]. A processor node can store data at multiple OCM nodes simultaneously with a single memory access, thereby distributing data locally to other processing nodes for distributed computing. Alternatively, the same data

can be distributed for fault tolerance redundancy. The use of multicasting can also be used to transmit along several redundant paths to the same destination node.

This OCM experimental implementation uses commercial components; thus the power consumption analysis does not provide an accurate comparison to electronic memory systems. Further, with the growing use of high data rate serial links, the SerDes power consumption will continue to decrease [32]. We envision huge power savings by incorporating integrated photonic components [33] in future OCM nodes, gaining benefits from the tighter integration of optical components with electronic driver and receiver circuitry [13–15]. Silicon photonics will eliminate the need for any off-chip wiring, such as between an SDRAM chip and a transceiver, improving the overall bandwidth, latency, and energy efficiency. By modulating and receiving the optical memory transactions within the memory and processor chips, we can achieve approximately 1 pJ/bit energy efficiency [18,19]. Continued development of silicon photonic technology is likely to improve this further.

## VI. CONCLUSION

Next-generation data centers will require memory systems with continually increasing capacity and performance, using more SDRAM devices with longer burst sizes and faster bus transfer rates. By replacing the electronic memory bus with a circuit-switched optical interconnection network and modifying the MC accordingly, we enable memory systems to continue scaling without the limiting trade-offs between capacity, performance, and energy efficiency. We first perform a detailed architectural analysis of the impact of OCM on future data centers, in parallel with the functionality and error-free routing of  $4 \times 2.5$ -Gb/s wavelength-stripped optical memory transactions between a microprocessor and OCM with EMBERS  $< 10^{-12}$ . This work illustrates the potential for new system architectures that leverage optical interconnects to create energy-efficient memory systems that are not feasible with today's electronic interconnects. Further integration of photonic and electronic components will be necessary for large-scale, high-performance applications.

## ACKNOWLEDGMENT

This work was supported in part by the Intel Corporation under grant SINTEL CU08-7952 and by the Department of Energy under grant DE-SC0005114.

## REFERENCES

- [1] W. A. Wulf and S. A. McKee, "Hitting the memory wall: implications of the obvious," *ACM SIGARCH Comput. Archit. News*, vol. 23, no. 1, pp. 20–24, 1995.
- [2] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, K. Yelick, and P. Kogge, *Exascale computing study: technology challenges in achieving exascale systems* [Online]. Available: <http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf>.

- [3] B. Jacob, S. W. Ng, and D. T. Wang, *Memory Systems: Cache, DRAM, Disk*. Morgan Kaufmann, 2007.
- [4] R. Ho, W. Mai, and M. A. Horowitz, "The future of wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.
- [5] JEDEC Solid State Technology Association, *DDR3 SDRAM Standard* [Online]. Available: <http://www.jedec.org/standards-documents/docs/jesd-79-3d>.
- [6] Micron Technology Inc., *Product specification. 2 GB DDR3 SDRAM* [Online]. Available: <http://download.micron.com/pdf/datasheets/modules/ddr3/j5f16c256x64h.pdf>.
- [7] The ITRS Technology Working Groups, *International Technology Roadmap for Semiconductors (ITRS) 2009 Edition* [Online]. Available: <http://www.itrs.net>.
- [8] Intel Corp., *Specification Addendum. Fully Buffered DIMM* [Online]. Available: [http://www.intel.com/technology/memory/FBDIMM/spec/Intel\\_FBD\\_Spec\\_Addendum\\_rev\\_p9.pdf](http://www.intel.com/technology/memory/FBDIMM/spec/Intel_FBD_Spec_Addendum_rev_p9.pdf).
- [9] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in *SLIP'04: Proc. of the 2004 Int. Workshop on System Level Interconnect Prediction*, ACM, 2004, pp. 7–13.
- [10] A. F. Benner, D. M. Kuchta, P. K. Pepeljugoski, R. A. Budd, G. Hougham, B. V. Fasano, K. Marston, H. Bagheri, E. J. Seminaro, X. Hui, D. Meadowcroft, M. H. Fields, L. McColloch, M. Robinson, F. W. Miller, R. Kaneshiro, R. Granger, D. Childers, and E. Childers, "Optics for high-performance servers and supercomputers," in *Opt. Fiber Commun. Conf. (OFC 2010)*, San Diego, CA, Mar. 2010, OTuH1.
- [11] B. J. Offrein and P. Pepeljugoski, "Optics in supercomputers," in *Eur. Conf. Opt. Commun. (ECOC 2009)*, Vienna, Austria, Sept. 2009, 3.1.3.
- [12] Y. Katayama and A. Okazaki, "Optical interconnect opportunities for future server memory systems," in *Proc. IEEE HPCA*, 2007, pp. 46–50.
- [13] G. Hendry, E. Robinson, V. Gleyzer, J. Chan, L. P. Carloni, N. Bliss, and K. Bergman, "Circuit-switched memory access in photonic interconnection networks for high-performance embedded computing," in *Supercomputing (SC)*, Nov. 2010.
- [14] A. Hadke, T. Benavides, S. J. B. Yoo, R. Amirharajah, and V. Akella, "OCDIMM: scaling the DRAM memory wall using WDM based optical interconnects," in *Int. Symp. on High-Performance Interconnects*, Aug. 2008.
- [15] S. Beamer, C. Sun, Y. J. Kwon, A. Joshi, C. Batten, and V. Stojanovic, "Re-architecting DRAM memory systems with monolithically integrated silicon photonics," in *37th Int. Symp. on Computer Architecture (ISCA-37)*, June 2010.
- [16] Corning Inc., *Datasheet: Corning SMF-28e optical fiber product information* [Online]. Available: <http://www.princetel.com/datasheets/SMF28e.pdf>.
- [17] B. G. Lee, F. E. Doany, S. Assefa, W. M. J. Green, M. Yang, C. L. Schow, C. V. Jahnes, S. Zhang, J. Singer, V. I. Kopp, J. A. Kash, and Y. A. Vlasov, "20- $\mu$ m-pitch eight-channel monolithic fiber array coupling 160 Gb/s/channel to silicon nanophotonic chip," in *Opt. Fiber Commun. Conf. (OFC 2011)*, San Diego, CA, Mar. 2010, PDP4A.
- [18] W. A. Zortman, M. R. Watts, D. C. Trotter, R. W. Young, and A. L. Lentine, "Low-power high-speed silicon microdisk modulators," in *Conf. on Lasers and Electro-Optics*, May 2010, CThJ4.
- [19] H. D. Thacker, Y. Luo, J. Shi, I. Shubin, J. Lexau, X. Zheng, G. Li, J. Yao, J. Costa, T. Pinguet, A. Mekis, P. Dong, S. Liao, D. Feng, M. Asghari, R. Ho, K. Raj, J. G. Mitchell, A. V. Krishnamoorthy, and J. E. Cunningham, "Flip-chip integrated silicon photonic bridge chips for sub-picojoule per bit optical links," in *Electronic Components and Technology Conf. (ECTC)*, June 2010, pp. 240–246.
- [20] D. Brunina, C. P. Lai, A. S. Garg, and K. Bergman, "First experimental demonstration of optically-connected SDRAM across a transparent optical network test-bed," in *23rd Annu. Meeting of the IEEE Photonics Society*, Denver, CO, Nov. 2010, Th1 1.
- [21] K. J. Barkerm, A. Benner, R. Hoare, A. Hoisie, A. K. Jones, D. J. Kerbyson, D. Li, R. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. Stunkel, and P. Walker, "On the feasibility of optical circuit switching for high performance computing systems," in *Supercomputing (SC)*, Nov. 2005.
- [22] D. Brunina, C. P. Lai, A. S. Garg, and K. Bergman, "Wavelength-stripped multicasting of optically-connected memory for large-scale computing systems," in *Opt. Fiber Commun. Conf. (OFC 2011)*, Los Angeles, CA, Mar. 2011, OWH4.
- [23] S. Rixner, "Memory controller optimizations for web servers," in *37th Int. Symp. on Microarchitecture*, Dec. 2004, pp. 355–366.
- [24] S. Rixner, W. J. Dally, U. J. Kapasi, P. Mattson, and J. D. Owens, "Memory access scheduling," in *Proc. of the 27th Int. Symp. on Computer Architecture*, 2000, pp. 128–138.
- [25] W. Lin, S. K. Reinhardt, and D. Burger, "Reducing DRAM latencies with an integrated memory hierarchy design," in *Proc. of the Int. Symp. on High-Performance Computer Architecture (HPCA)*, 2001, pp. 301–312.
- [26] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: a scalable and flexible data center network," in *Proc. of the ACM SIGCOMM 2009 Conf. on Data Communication (SIGCOMM'09)*, Barcelona, Spain, Aug. 2009, pp. 51–62.
- [27] C. P. Lai, D. Brunina, and K. Bergman, "Demonstration of 8  $\times$  40-Gb/s wavelength-stripped packet switching in a multi-terabit capacity optical network test-bed," in *23rd Annu. Meeting of the IEEE Photonics Society*, Denver, CO, Nov. 2010, ThQ 2.
- [28] W. Carlson, T. El-Ghazawi, B. Numrich, and K. Yelick, "Programming in the partitioned global address space model." *Supercomputing 2003* [Online]. Available: <http://www.gwu.edu/upc/tutorials.html>.
- [29] O. Liboiron-Ladouceur, B. A. Small, and K. Bergman, "Physical layer scalability of WDM optical packet interconnection networks," *J. Lightwave Technol.*, vol. 24, no. 1, pp. 262–270, Jan. 2006.
- [30] "IEEE P802.3ba 40 Gb/s and 100 Gb/s Ethernet Task Force," June 21, 2010.
- [31] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.*, vol. 19, pp. 297–301, 1965.
- [32] N. Nedovic, A. Kristensson, S. Parikh, S. Reddy, W. Walker, S. McLeod, N. Tzartzanis, H. Tamura, K. Kanda, T. Yamamoto, S. Matsubara, M. Kibune, Y. Doi, S. Ide, Y. Tsunoda, T. Yamabana, T. Shibasaki, Y. Tomita, T. Hamada, M. Sugawara, J. Ogawa, T. Ikeuchi, and N. Kuwata, "A 2  $\times$  22.3 Gb/s SFI5.2 SerDes in 65 nm CMOS," in *Compound Semiconductor Integrated Circuit Symp., 2009 (CISC 2009)*, 11–14 Oct. 2009, pp. 1–4.
- [33] L. Chen, K. Preston, S. Manipatruni, and M. Lipson, "Integrated GHz silicon photonic interconnect with micrometer-scale modulators and detectors," *Opt. Express*, vol. 17, no. 17, pp. 15248–15256, Aug. 2009.



**Daniel Brunina** (S'08) received his B.S. and M.S. degrees in computer systems engineering from Boston University, Boston, MA, in 2004 and 2005, respectively. He is currently working toward a Ph.D. degree at the Department of Electrical Engineering, Columbia University, New York, NY.

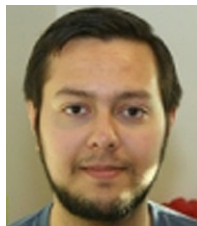
His current research interests include the design of optically connected memory architectures and optical interfaces for large-scale

computing.



**Caroline P. Lai** (S'07) received her B.A.Sc. degree (with honors) in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2006 and her M.S. degree in electrical engineering from Columbia University, New York, NY, in 2008. Currently, she is pursuing a Ph.D. degree in electrical engineering at Columbia University.

Her research interests lie in cross-layer communications for next-generation optical networks, as well as optical interconnects and optically connected memory for high-performance computing systems.



**Ajay S. Garg** (S'07) received his B.S. degree in Electrical engineering and his B.S. degree in computer systems engineering from Rensselaer Polytechnic Institute, Troy, NY, in 2007 and his M.S. degree in electrical engineering from Columbia University, New York, NY, in 2009. He is currently working towards a Ph.D. degree at the Department of Electrical Engineering, Columbia University, New York, NY.

His current research interests include optical interface design for large-scale computing and physical-layer simulation of novel WDM optical networks.



**Keren Bergman** (S'87–M'93–SM'07–F'09) received her B.S. degree from Bucknell University, Lewisburg, PA, in 1988, and her M.S. and Ph.D. degrees from Massachusetts Institute of Technology, Cambridge, MA, in 1991 and 1994, respectively, all in electrical engineering. She is currently a Professor at the Department of Electrical Engineering, Columbia University, New York, NY, where she also directs the Lightwave Research Laboratory.

Her research programs involve optical interconnection networks for advanced computing systems, photonic packet switching, and nanophotonic networks on-chip.

Prof. Bergman is a Fellow of the Optical Society of America and a Fellow of the Institute of Electrical and Electronic Engineers (IEEE). She is the Co-Editor-in-Chief of the OSA/IEEE *Journal of Optical Communications and Networking*.