Wavelength-Striped Multicasting of Optically-Connected Memory for Large-Scale Computing Systems

Daniel Brunina, Caroline P. Lai, Ajay S. Garg, and Keren Bergman
Department of Electrical Engineering, Columbia University, 500 West 120th St, New York, New York 10027
daniel@ee.columbia.edu

Abstract: We demonstrate the broadband multicasting of optically-connected memory between multiple SDRAM nodes and an emulated microprocessor on an optical network test-bed. 4×2.5-Gb/s wavelength-striped memory messages are multicasted error-free (BER<10^{-12}).
OCIS codes: (200.4650) Optical interconnects; (060.4255) Networks, multicast; (200.0200) Optics in computing

1. Introduction

The recent growth of high-performance computing systems is driven by high-bandwidth applications such as Internet searches, high-frequency trading, and high-definition video streaming. A significant challenge in the performance of future computing infrastructures lies in accessing main memory. Synchronous dynamic random access memory (SDRAM) offers the optimal balance of data capacity and speed as compared to other storage mediums. However, the performance gap between processors and memory is increasing steadily due to the scaling limits of electronics. Currently, increasing bandwidth requires using more SDRAM devices that are accessed in parallel, resulting in greater electronic wiring complexity and power consumption. Overall, electronic memory communication architectures cannot continue to meet the needs of next-generation large-scale computing systems.

Optical interconnects offer a high-bandwidth, energy-efficient approach to continue scaling the size and performance of computing systems [1], while mitigating the electronic memory bottleneck. We propose optically-connected memory systems, leveraging optics' power-efficient distance immunity at rack-to-rack computer scales, high-bandwidth density, and time-of-flight latencies [2,3]. Future optically-connected memory designs should also effectively support the redundancy and overprovisioning functionalities that current systems typically use to improve performance and reliability [4]. By allowing one source node (CPU rack) to multicast a single high-bandwidth message to multiple destinations (several SDRAM devices), multiple communication paths can be efficiently established to simultaneously transmit several copies of data. This will enable a more fault tolerance design and reduce memory access times by storing local copies of data closer to distributed computing nodes.

Future optical networks will be required to seamlessly support the multicasting of broadband optical messages comprising of memory information. The previous optically-connected memory system [2] is expanded to enable the multicast of a single memory transaction to multiple remote memory nodes across an optical network. This will allow an architecture wherein a computation-dedicated server rack can simultaneously access memory racks filled with many SDRAM devices (Fig. 1). Hence, computational nodes can efficiently access memory network nodes as if the data were stored locally. The ability to multicast high-bandwidth memory transactions to multiple memory nodes will further improve the reliability and performance of the optically-connected memory system as a whole.

We experimentally demonstrate a processing node communicating simultaneously to multiple memory nodes by multicasting high-bandwidth, wavelength-division multiplexed (WDM) optical messages on a 5-stage 4×4 multicast-capable optical interconnection network test-bed. A high-speed field-programmable gate array (FPGA) is used to realize the processing core and Micron DDR2 SDRAM devices act as the remote memory nodes.

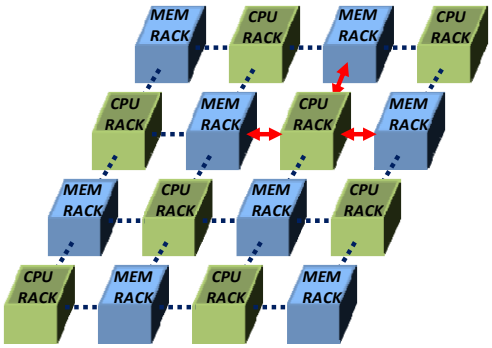

Fig. 1. CPU racks (green) are connected to memory racks (blue) by optical links (dotted lines); red arrows show a CPU multicasting to three memory racks.

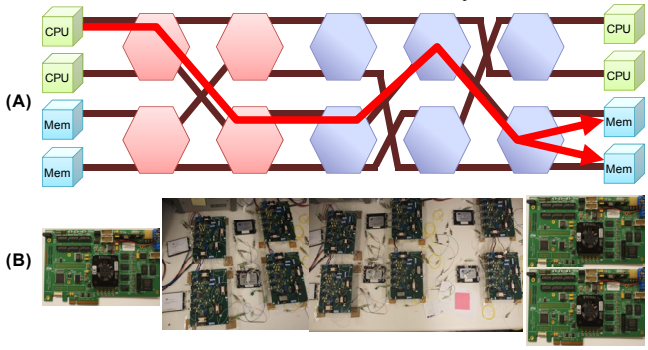

Fig. 2. (A) Block diagram of setup showing a CPU (green) multicasting to two memory nodes (blue) across a 4×4 multicast-capable test-bed. (B) Photograph of implemented FPGA boards and optical network test-bed.

2. Optical Interconnection Network

The memory multicasting demonstration is realized on a 4×4 multicast-capable optical interconnection network test-bed (Fig. 2) [5]. The multistage network topology is composed of two packet routing stages and three packet multicasting stages; each having different electronic control logic. The stages are cascaded to implement a multicasting operation that allows transparent, multiwavelength lightpaths to fan from a single network input port to several output ports. The test-bed itself is built with ten 2×2 photonic switching nodes, each using four semiconductor optical amplifiers (SOAs) that provide a broadband gain spectrum, fast switching times, and data-rate transparency. The switching nodes leverage commercially-available, discretely-packaged components such as passive optical devices, high-speed electronic circuitry, and SOAs. The routing control logic is synthesized in complex programmable logic devices (CPLDs) located at each 2×2 node.

The architecture supports a wavelength-striped optical message format: control bits are encoded on dedicated wavelengths and set constant for the duration of the circuit, while the WDM memory payload data is modulated on separate wavelength channels. The 4×2.5-Gb/s payloads corresponding to the memory information from a processor are multiplexed with control headers and injected in the test-bed. According to the logic within the CPLDs and the extracted headers, the SOAs set up circuit paths through the network to multicast the memory transactions from the processor to multiple SDRAM destination ports. Here, we show a multicast of the wavelength-striped memory data from one processor to two SDRAM modules.

3. Multicasting Memory Access Protocol

Processor-memory communication requires a memory controller (MC), which may be located on-chip or separately close to the processor. The MC translates processor-requested memory addresses into physical memory locations (since the processor itself is typically not aware of physical memory organization), and schedules the memory transactions to optimize memory communication bandwidth. Traditional processor-memory communication uses a wide electronic bus that may limit the scaling of future memory capacity and performance. By replacing the electronic memory bus with an optical interconnect, we allow an architecture in which SDRAM-based memory systems can continue to scale to meet the necessary processor-memory bandwidth requirements. Thus, in this work, all processors and memory nodes exhibit equal bandwidth.

The novel multicast-enabled optically-connected memory architecture enables a structure in which memory devices are addressed by both traditional memory location addresses and optical network addresses [2]. To accomplish this, the custom MC translates a portion of the physical addresses into network addresses corresponding to the appropriate memory node. This process creates wavelength-striped messages using multiple wavelength channels to encode the memory payload and the correct network header information for optical interconnect routing. Furthermore, when a processor must issue a single memory write or read transaction to multiple memory nodes, the MC will allow for the simultaneous transmission of multiple transactions through the multicasting scheme. This will significantly improve performance by increasing the aggregate memory bandwidth, since the processor must only initiate one memory transaction that can then offer the high bandwidth to many memory destinations.

4. Experimental Setup and Results

The experimental setup (Fig. 3) consists of three circuit boards communicating 4×2.5-Gb/s WDM messages across a 4×4 optical network test-bed. In addition to modulating the four memory data payload channels, each board must modulate five network address wavelengths for routing and multicasting through the optical network, and one frame wavelength to indicate that the address information is valid. The payload wavelengths are modulated using four

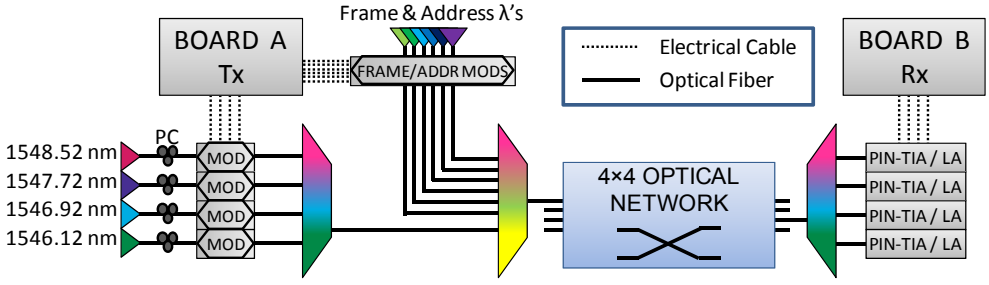

Fig. 3. Experimental setup showing Board A communicating to Board B over a 4x4 optical network test-bed. Board A modulates 4 payload channels (lower left), as well as a frame and 5 address bits (upper center) for network routing. Board B receives the payload from the optical network using four PIN-TIA receivers.

$LiNbO_3$ modulators, while the lower data rate frame and address wavelengths are modulated using SOAs. All ten wavelengths are multiplexed together onto one single-mode fiber before injection into the optical test-bed.

Each circuit board is identical and contains an Altera Stratix II GX FPGA, which is connected to four chips of Micron DDR2 SDRAM and 4×2.5-Gb/s transceivers. The transceivers drive the $LiNbO_3$ modulators to transmit over the network, while data from the network is received by the transceivers using p-i-n receivers with transimpedance amplifiers (TIA) and limiting amplifiers (LA). One circuit board's FPGA is programmed to act as a microprocessor with an on-chip memory controller that is capable of multicasting to multiple optically-connected memory nodes across the optical network test-bed. The two other circuit boards act as remote memory nodes containing the DDR2 SDRAM. As with traditional memory systems, the MC within the processor FPGA controls the memory node SDRAM. The resulting configuration allows all the processor-memory communication to occur using the optical network with an aggregate 10-Gb/s memory bandwidth; as necessary, the processor node will multicast memory transactions to both SDRAM nodes for an aggregate 20-Gb/s memory bandwidth.

The SDRAM is operated without error-correction techniques, which are frequently used in server applications, to more accurately measure the functionality of the optically-connected memory system. Any uncorrected bit errors during memory communication, either from the SDRAM itself or the interconnect, will cause effects ranging from unpredictable application behavior or data loss, to performance degradation or system failure. Therefore, to verify the correct functionality of the multicast-enabled optically-connected memory system, our processor repeatedly multicasts to both memory nodes and fills all memory addresses with predictable bit patterns: all 0s, all 1s, pseudo-random bit sequence (PRBS), or addresses corresponding to the destination memory locations. The processor then issues read requests for all memory locations, verifying each data bit as it streams in from the network.

An effective memory bit-error rate (EMBER) is confirmed as less than 10^{-12} once over a terabit of data has been verified by the system. Thus, we demonstrate the correct functionality and stability of the multicast-capable optically-connected memory system. Optical eye diagrams for four of the 2.5-Gb/s payload channels corresponding to processor write-to-memory transactions are shown in Fig. 4; these eyes were collected using self-triggering.

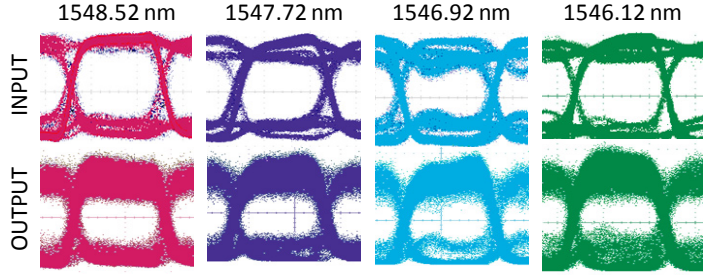

Fig. 4. Optical eye diagrams for the experimental demonstration, showing the four memory data payload wavelengths at the input of the network test-bed (top), and the same four channels after being multicasted to one network output port (bottom). (200 ps/div)

4. Conclusions

We demonstrate a novel memory architecture for large-scale computing systems in which a processing core can communicate to multiple remote memory nodes leveraging a multicast-capable optical interconnection network. Using commercially-available off-the-shelf FPGAs, SDRAM modules, and optical devices, we implement a microprocessor with a custom memory controller that multicasts 4×2.5-Gb/s wavelength-striped memory transactions using an implemented optical network test-bed to two independent, optically-connected memory modules. Error-free multicasting and routing is confirmed (EMBERs < 10^{-12}). This work illustrates the feasibility of multicasting to multiple optically-connected memory devices, enabling innovative system architectures for future high-performance computing systems with improved performance and efficiency.

We gratefully acknowledge partial support for this work from the DARPA MTO under grant ARL-W911NF-08-1-0127.

4. References

[1] R. Ramaswami and K. N. Sivarajan, Optical Networks-A Practical Perspective, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2002, ch.12.

[2] 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 Annual Meeting of the IEE Photonics Society*, Denver, CO, Nov 2010, Paper ThI 1.

[3] Y. Katayama and A. Okazaki, "Optical interconnect opportunities for future server memory systems," in *Proceedings of the 2007 IEEE 13th International Symposium on High Performance Computer Architecture*, pp. 46-50, Feb 2007.

[4] L. A. Barroso and U. Hölzle. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, M. D. Hill, Ed. San Rafael, CA, USA: Morgan & Claypool, 2009.

[5] C. P. Lai and K. Bergman, "Network Architecture and Test-Bed Demonstration of Wavelength-Striped Packet Multicasting," *in Optical Fiber Communication Conference*, San Diego, CA, Mar 2010, Paper OWI4.