Optically-Connected Memory: Architectures and Experimental Characterizations

Daniel Brunina

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

Abstract

Optically-Connected Memory: Architectures and Experimental Characterizations

Daniel Brunina

Growing demands on future data centers and high-performance computing systems are driving the development of processor-memory interconnects with greater performance and flexibility than can be provided by existing electronic interconnects. A redesign of the systems' memory devices and architectures will be essential to enabling high-bandwidth, low-latency, resilient, energy-efficient memory systems that can meet the challenges of exascale systems and beyond.

By leveraging an optics-based approach, this thesis presents the design and implementation of an optically-connected memory system that exploits both the bandwidth density and distance-independent energy dissipation of photonic transceivers, in combination with the flexibility and scalability offered by optical networks. By replacing the electronic memory bus with an optical interconnection network, novel memory architectures can be

created that are otherwise infeasible. With remote optically-connected memory nodes accessible to processors as if they are local, programming models can be designed to utilize and efficiently share greater amounts of data. Processors that would otherwise be idle, being starved for data while waiting for scarce memory resources, can instead operate at high utilizations, leading to drastic improvements in the overall system performance.

This work presents a prototype optically-connected memory module and a custom processor-based optical-network-aware memory controller that communicate transparently and all-optically across an optical interconnection network. The memory modules and controller are optimized to facilitate memory accesses across the optical network using a packet-switched, circuit-switched, or hybrid packet-and-circuit-switched approach. The novel memory controller is experimentally demonstrated to be compatible with existing processor-memory access protocols, with the memory controller acting as the optics-computing interface to render the optical network transparent. Additionally, the flexibility of the optical network enables additional performance benefits including increased memory bandwidth through optical multicasting. This optically-connected architecture can further enable more resilient memory system realizations by expanding on current error dectection and correction memory protocols.

The integration of optics with memory technology constitutes a critical step for both optics and computing. The scalability challenges facing

main memory systems today, especially concerning bandwidth and power consumption, complement well with the strengths of optical communications-based systems. Additionally, ongoing efforts focused on developing low-cost optical components and subsystems that are suitable for computing environments may benefit from the high-volume memory market. This work therefore takes the first step in merging the areas of optics and memory, developing the necessary architectures and protocols to interface the two technologies, and demonstrating potential benefits while identifying areas for future work. Future computing systems will undoubtedly benefit from this work through the deployment of high-performance, flexible, energy-efficient optically-connected memory architectures.

Contents

List of Figures

List of Tables

Acknowledgements

I am grateful for the support of my advisor, Professor Keren Bergman, for her leadership and vision that have guided me through my doctoral studies. I have learned a great deal from her, and none of this work would have been possible without her.

Thank you to the members of my dissertation committee: Professor Richard Osgood, Professor Luca Carloni, Professor Ken Shepard, and Dr. Jeffrey Kash. Professor Richard Osgood, having taught several of my optics classes, was an integral part of my Columbia education. Professor Luca Carloni has always been an eager source of feedback, and our discussions contributed greatly to my research. Thank you to Professor Ken Shepard for his humor and contributions to this work.

I am especially grateful to Dr. Jeffrey Kash for having had the opportunity to intern with his amazing research group. While at IBM Research, I had the opportunity to work with Clint Schow, Petar Pepeljugoski, Alexander Rylyakov, Dan Kuchta, Marc Taubenblatt, Laurent Schares, Fuad Doany, and Ben Lee. All of these kind, brilliant individuals made the internship an invaluable learning experience and I am grateful to them all. During that period, Ben Lee, whom I only knew briefly during his final year in the Lightwave Research Laboratory, took the time to impart to me his wisdom and experience and I thank him immensely.

Thank you to all the members of the Lightwave Research Laboratory, past and present. I had the pleasure of being mentored by Howard Wang and Ajay Garg, and the majority of my lab skills were learned from them. Throughout my time at Columbia,

they were always willing to engage in research discussions and provide valuable feedback on any ideas. Thank you to Caroline Lai, who worked with me on multiple experiments and provided mentorship both in the lab and in the subsequent research papers. I would like to thank Noam Ophir, Kishore Padmaraju, Michael Wang, Johnnie Chan, and Gilbert Hendry for their friendship and research discussions throughout the years. I have enjoyed mentoring, and hopefully have been useful to, the younger students: Atiyah Ahsan, Cathy Chen, Robert Hendry, Qi Li, Lee Zhu, Christine Chen, and Gouri Dongaonkar. Also, thank to you to Cedric Ware, Balagangadhar Bathula, Lin Xu, Wenjia Zhang, and Dawei Liu for your contributions. To all of you: I look forward to continuing our relationships in the future, both as colleagues and as friends. I wish everyone the best of luck.

I am especially thankful to my friends and family. I am most of all thankful to my mother - without your endless support and hard work I would of course not be here (literally and figuratively). Thank you to my brothers, grandmother, and great-grandmother. Thank you, Brian, for almost 20 years of friendship. It is not possible to complete a dissertation, while remaining sane, without family and a good friend.

This thesis is dedicated to my mother and grandmother.

Chapter 1

Introduction

Next-generation large-scale high-performance computing (HPC) systems and data centers will require microprocessors to support unprecedented off-chip bandwidths to memory, with low access latencies and interconnect power dissipation. However, today's off-chip electronic interconnects face performance challenges with low bandwidth densities, as well as distance- and data-rate-dependent energy dissipation. As a result, large-scale systems have experienced an exponentially growing performance gap between the computational performance of microprocessors and the performance of off-chip main memory systems [1]. This communications bottleneck will undoubtedly limit the overall system performance and scalability of future large-scale systems. Due to the trade-offs among the requirements in communication bandwidth, latency, and energy efficiency, the systems' high-performance microprocessors will be starved for memory data [2, 3]. Furthermore, growing data sets and server virtualization are placing demands on the system, limiting the types of applications that may be supported.

Although it is feasible for electronic interconnection networks to reach per-channel data rates up to 25 Gb/s [4], the power dissipation at such high bandwidths becomes overwhelming and contributes greatly to increased overall system cost and complexity. Currently, microprocessors are estimated to dissipate half of their energy in the interconnect alone [5]. Scaling interconnect performance using traditional approaches would continue to exacerbate this imbalance. For example, a typical main memory system consists of multiple chips of synchronous dynamic random access memory (SDRAM) packaged together onto a circuit board called a dual in-line memory module (DIMM), which is capable of providing over 120 Gb/s of peak bandwidth [6]. Multiple DIMMs must be accessed in parallel, requiring an extremely complex electronic bus, to provide the many terabits-per-second of memory bandwidths required by data-intensive applications. However, the scaling challenges facing electronic interconnects limit the number of DIMMs that can be accessed, and, consequently, the total memory bandwidth. Increasing the per-channel SDRAM data rate has been attempted [7]; however, the resulting system remains limited in use due to its significantly higher energy consumptions.

1.1 Trends in Large-Scale Computing

HPCs and data centers 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 and reduce the total system-wide performance. Next-generation large-scale computing

systems must therefore leverage novel physical-layer technologies in order to close the processor-memory performance gap and enable future microprocessors to achieve their full potential.

1.1.1 High-Performance Computers

HPCs, or supercomputers, are designed to optimize processing capacity and are typically used for calculation-intensive such as: climate modeling [8], oil and gas exploration [9], quantum physics [10], and nuclear research [11]. Today's supercomputers employ tens of thousands of processors and petabytes of memory to achieve the targeted performance, as rated in floating-point operations per second (FLOPS), on the order of petaFLOPS. Given the trend in HPC performance (Figure 1.1), an exaFLOP machine should be created within this decade. Such extreme-scale computers rely heavily on efficient interconnects to provide the high-performance processing cores with a steady stream of data. The interconnects for these systems are typically electrically-interconnected three-dimensional (3D) torus topologies [12]. However, with the ever-increasing demands being placed on the interconnect, future HPCs require a redesign at the physical layer, potentially capitalizing on the benefits presented by optical interconnects [13].

A recent example of this trend is IBM's Power 775 [15], which was created in response to the Defense Advanced Research Projects Agency (DARPA) initiative for High Productivity Computing Systems. The goal of this initiative was to enable the creation of HPCs that are economically viable. With this goal in mind, a significant amount of inter-node communication was moved from the traditional electrical domain

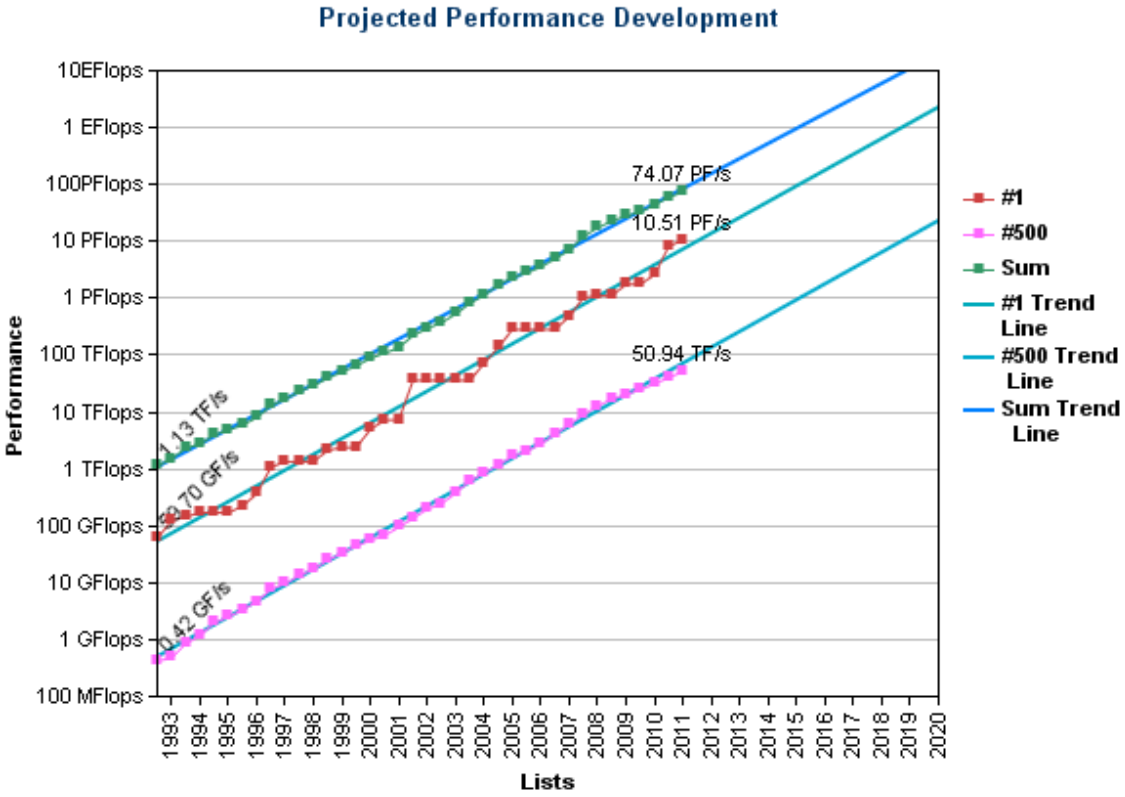

Figure 1.1: Performance Trends in Top500 Supercomputers - The performance of the top supercomputers has increased by roughly 10× every 4 years [14].

to the more efficient optical domain. Each Power 775 drawer contains up to eight nodes of four POWER7 [16] processors, resulting in up to 256 POWER7 cores and 128 DIMMs (yielding 2 TB of total memory capacity). The DIMMS (Figure 1.2) are very densely packed, requiring water cooling, due to the immense required memory bandwidth and the high power and writing complexity of electrically connecting those DIMMs to the cores. Twelve of these drawers can be combined into one rack, which requires over 1,500 DIMMs, for a total performance of 96 TFLOPS. The most notable advancement on the Power 775 drawer is the hub controller, which contains a combination of optical and electrical transceivers to achieve an aggregate 1.1 TB/s communication bandwidth.

The primary reason for a shift to optical interconnects within HPCs is the need to both overcome electrical bandwidth limitations and achieve a high FLOPS/Watt ratio. As supercomputers continue to grow, incorporating more processors and larger interconnection networks, the operating costs due to power have become a significant design consideration. This constraint results in the following trend: maximizing computational performance is no longer the only focus when designing HPCs. With this in mind, four of the top 10 supercomputers are based on the optically-interconnected Blue Gene/Q [17] architecture; at 2097.19 MFLOPS/Watt [18], this design is the most energy-efficient HPC architecture. In contrast, the electrically-connected Tianhe-1A supercomputer [19] in China consumes over 4 megawatts. If this system was located in the US, it would cost an average of 10 cents per kilowatt hour [20], and would therefore result in an electricity bill of over $3.5 million per year; this is typical of the Top 10 systems listed on the Top500 [14].

The current top supercomputer, IBM Sequoia [14], is a Blue Gene/Q-based16.32

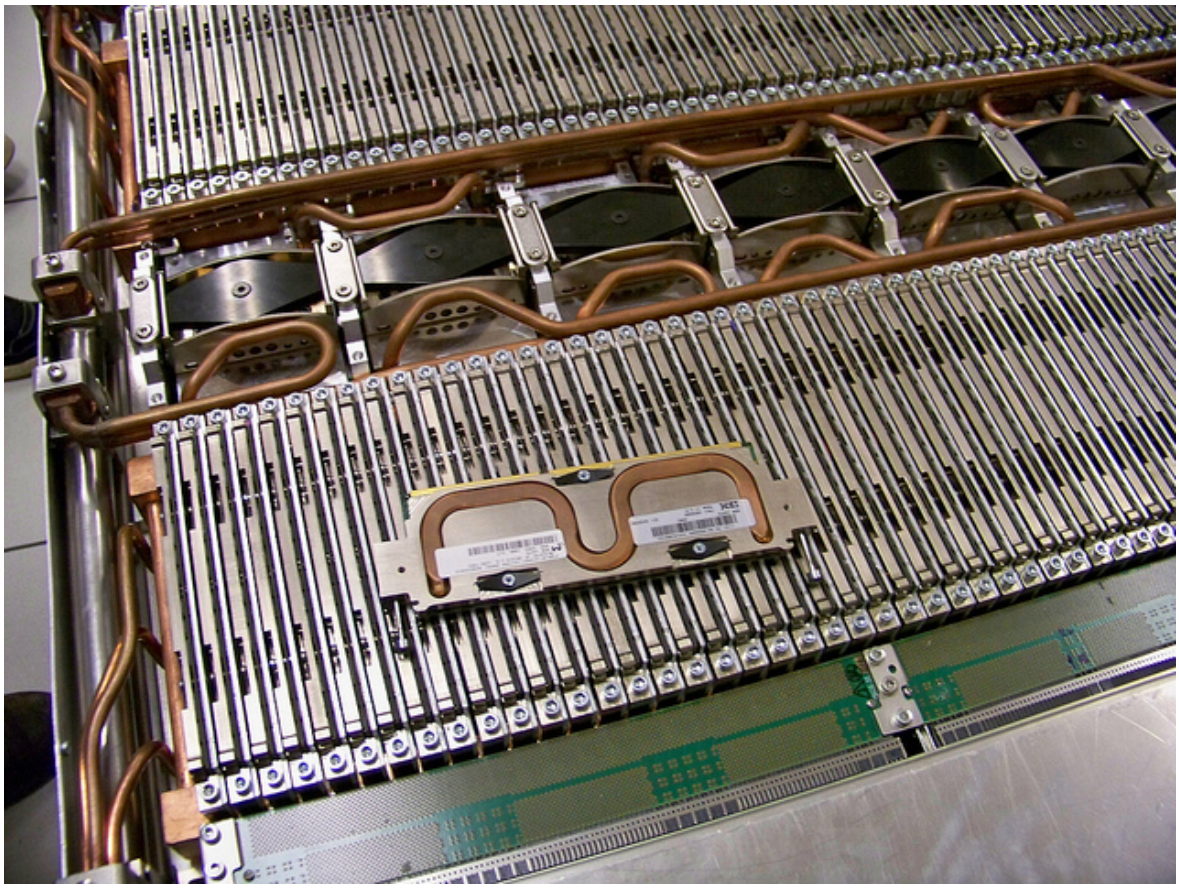

Figure 1.2: Photograph of Power 775 Drawer - IBM's Power 775 drawer contains up to 256 cores and 128 DIMMs. The components are densely packed to maximize performance, which necessitates water cooling. Hub modules enable high-performance optical links between drawers [15].

petaFLOPS machine that uses over 1.5 million cores. With a total power consumption of 7.9 MW, Sequoia is not only 1.5 times faster than the second-ranked supercomputer, the K computer [21], but also 150% more energy efficient. The K computer, which utilizes over 80,000 SPARC64 VIIIfx processors [22], results in the highest total power consumption of any Top500 system (9.89 MW). IBM Sequoia achieves its superior performance and energy efficiency through the use of custom compute chips and optical links between compute nodes. Each compute chip (Figure 1.3) contains 18 cores: 16 user cores, 1 service, and 1 spare [17]. The chips contain two memory controllers, which enable a peak memory bandwidth of 42.7 GB/s, and logic to communicate over a 5D torus that utilizes point-to-point optical links.

These trends demonstrate that current HPCs are already pushing the limits of traditional, electrical interconnection networks, not only in terms of performance but also energy. Future exascale HPCs will require unprecedented levels of processor-memory and inter-node communication, and thus require substantially higher bandwidth-density than is available in today's HPCs. The distance-dependent energy dissipation and low bandwidth-density of electrical interconnects will thus make such next-generation HPCs too expensive to operate or simply impractical to build, and a shift to optical interconnects is therefore required.

1.1.2 Data Centers

Data centers are large-scale computing systems with high-port-count networks interconnecting many servers, typically realized by commodity hardware, which are designed to support diverse computation and communication loads while minimizing

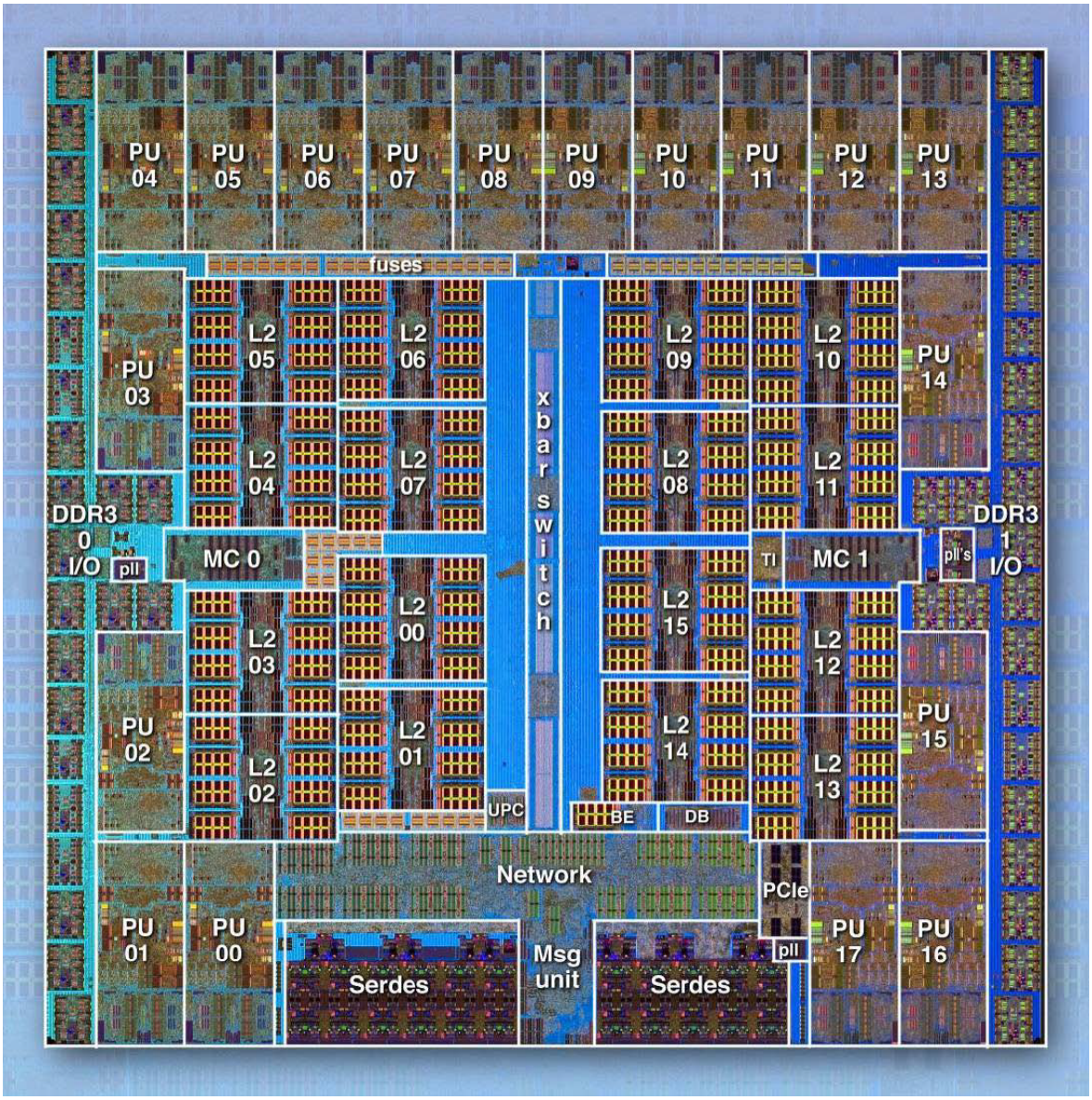

Figure 1.3: Blue Gene Q Compute Chip - IBM's Blue Gene Q compute chip contains 18 cores and dual DDR3 memory controllers for 42.7 GB/s peak memory bandwidth. Compute nodes are connected by a 5D torus with optical links [17].

hardware and maintenance costs. Contemporary data centers consist of tens of thousands of servers, or nodes, and new mega data centers are emerging with over 100,000 nodes [23]. The tremendous rise of cloud computing has caused a dramatic increase in the need for a greater number of larger data centers in order to handle the diverse, unpredictable communication between the computational nodes/servers. Simultaneously, these data centers must remain inexpensive to build and operate in order to maximize the profitability of the underlying applications. The design constraints on data centers therefore focus on the conflicting goals of minimizing cost while meeting the high-performance requirements of future cloud services.

The performance scaling of data centers is primarily hindered by the unpredictable, communication-intensive workloads that can consist of both bursty and long-lived traffic [24, 25]. Typically, data centers are required to support data-intensive applications, such as internet searches, which require vast amounts of data to be readily available at all times. These targeted applications are most efficiently realized with the search data stored in main memory. This is due to: the relative slow speed of hard disks, which have high data capacity but unacceptably high access latencies; and the relative small storage size of on-chip memories, which could deliver the data in a fast way but cannot possibly contain all the search data [26]. Main memory therefore possesses the ideal capacity-speed ratio, at an acceptable price, for data center applications. The result is that data center networks must deliver varying amounts of memory data between servers in in a low-cost, high-performance, unpredictable manner.

Existing attempts to optimize networks for such heterogeneous traffic [27, 28] have only increased bandwidth at the expense of added complexity and power dissipation,

and therefore cost, due to the use of electronic-based switches. The resulting trade-off makes fat trees [29] a desirable topology for data centers; however, the resulting configuration is one with significant oversubscription of the data center interconnection network, reducing overall performance to maintain an adequate performance/cost ratio.

Recent developments in data center architectures have focused on improving the performance/cost ratio by addressing the underlying physical-layer technology within the network. High-radix microelectromechanical systems- (MEMS) based optical circuit switches have been proposed as an attractive addition to data center interconnection networks [30, 31] due to the superior energy efficiency and bandwidth density of optics as compared to electrical switches. However, the main drawback of a MEMS-based approach is the relatively high switching latency associated with the technology, leading to an inflexible network. Furthermore, such implementations are inadequate for handling diverse and unpredictable data center traffic, and therefore additional physical-layer advancements are necessary for next-generation data centers.

1.2 The Memory Wall

The growing disparity between processor performance and memory bandwidth has been termed the "Memory Wall" [32]. Between the years 1980 and 2000, the typical processor improved in performance (measured in MFLOPS) by approximately 50% per year, while memory bandwidth only improved by 35% per year [33]. Figure 1.4 illustrates the continuation of this trend with modern processors. The trend line for modern processors is compared to a reference 1 B/FLOP metric, which would indicate an equal amount of memory bandwidth and processor performance, and highlights the

continued growth of the processor-memory gap.

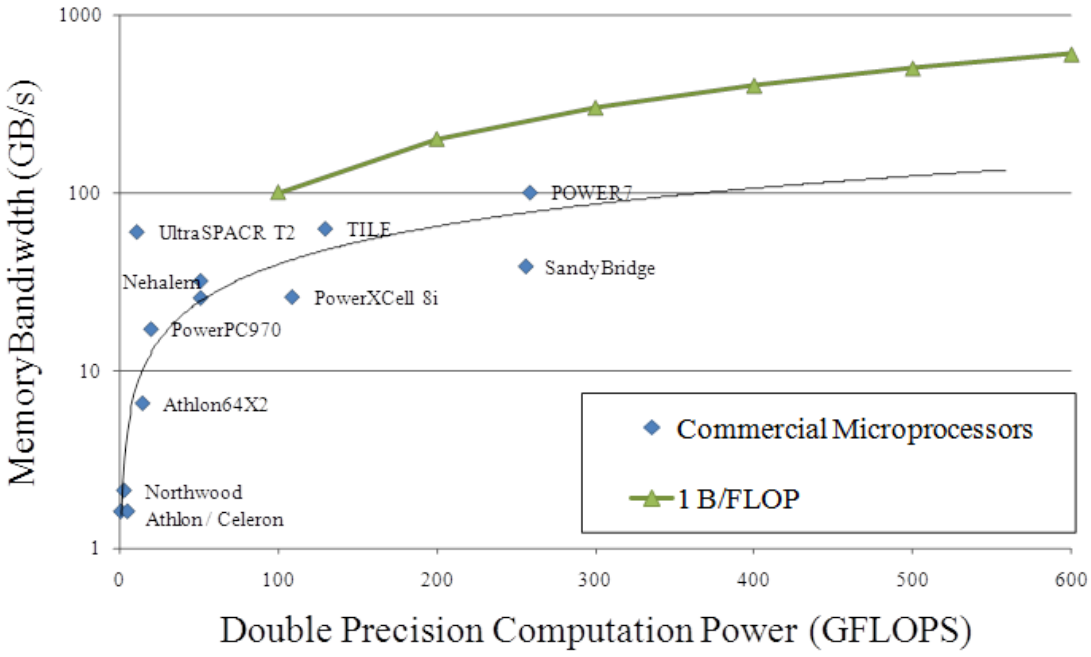

Figure 1.4: The Memory Gap: Graph of Processing Power vs Memory Bandwidth - A comparison of processor performance (GLOPS) to memory bandwidth for recent high-performance processors. The 1 B/FLOP reference line illustrates the increasing gap between performance and memory bandwidth.

The primary cause of the Memory Wall is a significant drop in bandwidth when moving from on-chip to off-chip communication, combined with the relatively high latency of SDRAM access compared to processor clock speeds. For example, the Sony-IBM-Toshiba Cell processor inter-core network can achieve bandwidths of 100 GB/s [34], while a high-performance off-chip memory link can only reach 10 GB/s [35]. A primary cause of this gap is that SDRAM technology operates at a fraction of the clock frequency of modern processors [6]. Therefore, increasing the

bandwidth of memory systems requires increasing 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 electrical traces to be path-length matched [13]. The wire length and routing complexity combine to limit the minimum memory access latency and maximum bus signaling rate. The result is that while SDRAM peak bandwidth has increased roughly 10-fold over the past decade, the clock speeds and latencies of the SDRAM devices have remained almost unchanged (Figure 1.5) [36, 37].

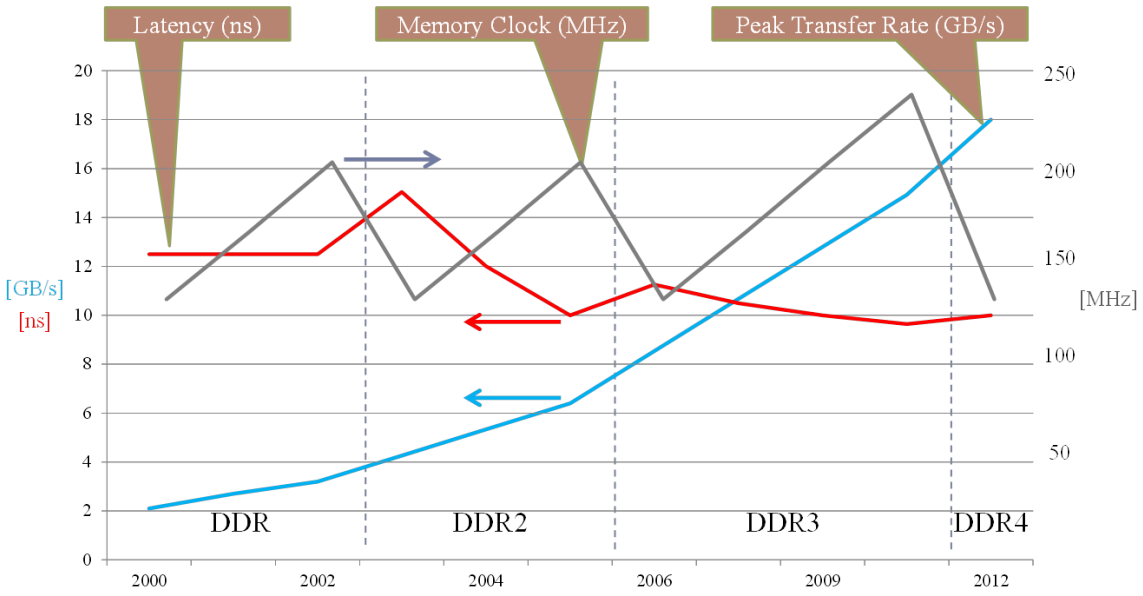

Figure 1.5: Graph of SDRAM Technology Trends - Over the last four generations of SDRAM technology, the bandwidth has increased approximately10-fold while latency has decreased very little. [36, 37].

Even assuming the latency of SDRAM itself can be improved to match the speed of

processors, the Memory Wall will persist due to limited off-chip bandwidth. Figure 1.6 illustrates how future systems will be constrained by pin-channel data rates and pin counts. The low bandwidth-density of electronics has limited memory per-channel data rates to at most 1 GHz [6] and therefore a typical DIMM utilizes a 64-bit wide electronic bus. Increasing memory bandwidth by accessing multiple DIMMs in parallel requires the processor to dedicate more pins to the memory bus, with 64 bits per parallel access (i.e. each memory channel). International Technology Roadmap for Semiconductors (ITRS) projections indicate that even next-generation high-performance processors cannot dedicate more than a few thousand pins to memory [3] even in the most aggressive estimates. Increasing the per-channel data rates is prohibitively expensive in terms of energy efficiency due to the data rate- and distance-dependent nature of electronic wiring power dissipation. The two combined factors place an upper limit on potential memory bandwidth of future processors. This trend necessitates either a shift in HPC and data center roadmaps to account for higher costs and lower performance systems, or the deployment of novel optical interconnect and switching technologies.

1.3 Main Memory

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' targeted applications, but their low bandwidth and millisecond access

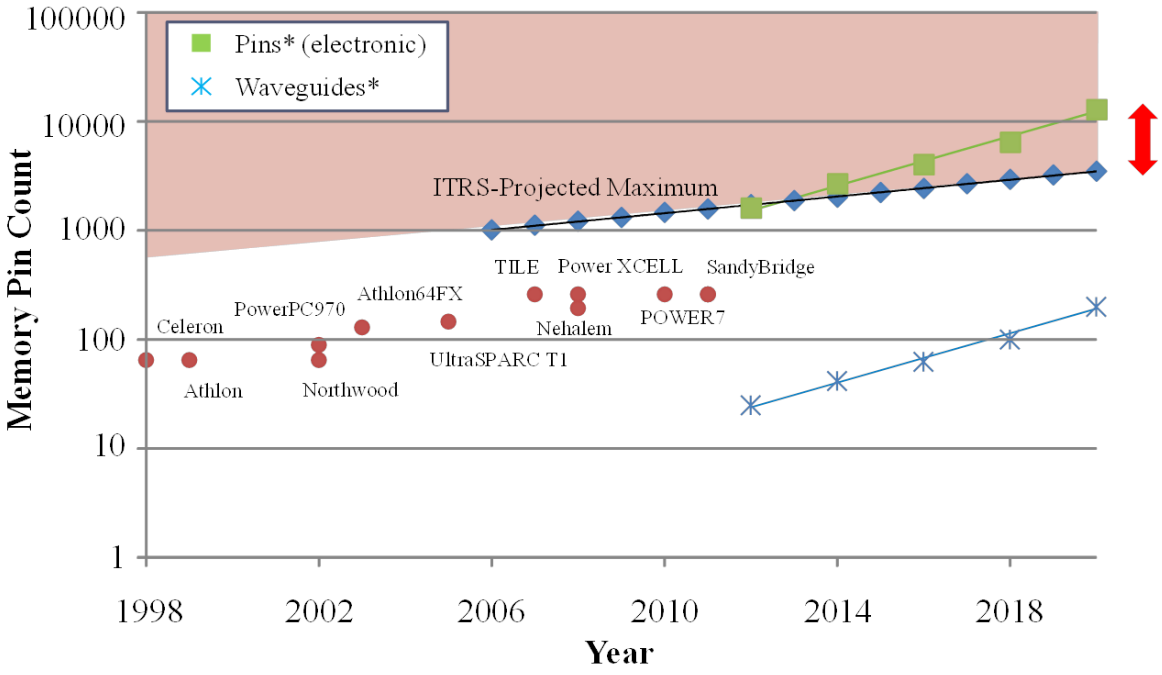

Figure 1.6: Illustration of Memory Scalability Challenges - The number of pins on commercial processors dedicated purely to memory has increased with time to meet growing memory bandwidth demands. Future systems using electronic pins will soon require a greater number of pins than is physically possible to implement [3].A shift to optical interconnects can accomodate the bandwidth requirements with a small number of waveguides. This graph is for a hypothetical system assuming: 1 TFLOPS in 2012 and performance doubling every 2 years, 1 B/FLOP required memory bandwidth, and per-channel data rate scaling from 5 Gb/s (2012) to 12.5 Gb/s (2020).

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 DIMMs, which contain multiple SDRAM chips and can provide access to gigabytes of capacity with over 100 Gb/s memory bandwidth and tens of nanoseconds access time [37]. Each SDRAM chip is comprised of several independent memory banks (Figure 1.7), which are accessed independently to allow the processor's memory controller (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 serializer/deserializer (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 words if error correction is enabled) to the memory controller 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.

Accessing SDRAM is a multi-step process that requires tens of nanoseconds [6]. All main memory accesses are handled by a memory controller. 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 memory controller, which translates the requests into a series of SDRAM-specific commands that are re-ordered to maximize memory bandwidth.

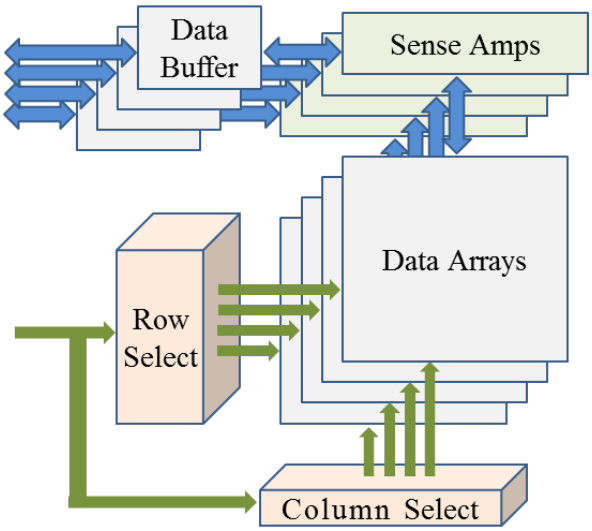

Figure 1.7: Block Diagram of SDRAM Bank - Anatomy of one bank of SDRAM. Four data arrays share common row and column selects for concurrent access.

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 memory controller sends a read (RD) or write (WR) command along with the column address. For a write, the memory controller 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 memory controller.

Bandwidth is optimized by re-ordering 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 re-ordering memory accesses, the memory controller 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. The burst size has doubled with each new generation of SDRAM technology, which has allowed the memory bandwidth to increase without real gains in the actual speed of SDRAM devices, and thus comes at the cost of memory access granularity.

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 [7]. The goal of FB-DIMM is to alleviate the limitations of electronic interconnect scaling by incorporating 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 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 [26]. Although using 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. Adding too many memory channels will not only overburden the memory controller but will also lead to the same pin count and wiring problems with traditional parallel electronic wiring.

Overall, advancements in main memory technology have lagged well behind those

of processors. The resulting Memory Wall threatens to stifle future processor advancements to due the lack of high-bandwidth, energy-efficient access to memory data. Thus far, attempts to increase the number of memory channels or scale up per-channel data rates have required unacceptable trade-offs between bandwidth, latency, and power. It is abundantly clear that a shift from electrical wires will be necessary as the next step for achieving improved processor-memory links.

1.4 The Optics-Computing Interface

The use of photonic technology can enable high-bandwidth links, with novel functionalities to reduce off-chip data access latency and power dissipation [38]. The integration of on-chip silicon photonic transceivers [39] will further enable processor-memory communication with the off-chip bandwidth and energy-efficiency performance equal to that of on-chip communications [39, 40]; this would be impossible using conventional electronic interconnects. In addition to achieving high per-channel data rates, optical interconnects can significantly improve communication bandwidths through wavelength-division multiplexing (WDM), and can therefore support many terabits-per-second of optical bandwidth using a single waveguide or optical fiber [41].

Active optical cables serve as an intermediate step in the shift toward optics in computing; however, such implementations still rely on traditional, inefficient electrical transceivers and therefore provide only modest performance or energy improvements. Meanwhile, the development of optical switching [42, 43, 44, 45] can drastically improve the overall performance and energy efficiency of HPCs and data centers. The challenge lies in combining the disparate technologies of computing and optical

networks to maximize the benefits achieved within a future optically-enabled computer system. The electronic technology within microprocessors has been optimized for short-distance, bursty communication that relies heavily on point-to-point links where data is frequently buffered and retransmitted. In contrast, a lack of optical buffer technology has resulted in the development of optical networks that utilize unique communication protocols that do not directly map to existing electronic systems. It is therefore necessarily to develop an optics-computing interface that can enable future processors to fully leverage the benefits of optical interconnects without significantly modifying the underlying processor functionalities.

1.4.1 Silicon Photonics

Close integration of photonic and processor/memory hardware is essential to realizing the full potential benefits of off-chip optical interconnects. Recent advances in silicon photonics have yielded high-performance, compact, energy-efficient, CMOS-compatible nanophotonic devices [46, 47]. Integrating these silicon photonic components with processors and memory (Figure 1.8) will eliminate the need for power-hungry off-chip electronic wires, alleviate pin-count constraints, and maximize memory bandwidth with WDM. The resulting alignment of off-chip memory bandwidth with on-chip bandwidth enables processors to access remote memory as if it were local, which is unachievable with electronic interconnects due to pinout limitations.

Silicon photonics remains a relatively immature technology, and as of yet, no process exists to fully integrate photonic transceivers directly with processors or memory devices. However, recent advances have shown optical transceiver modules that can be

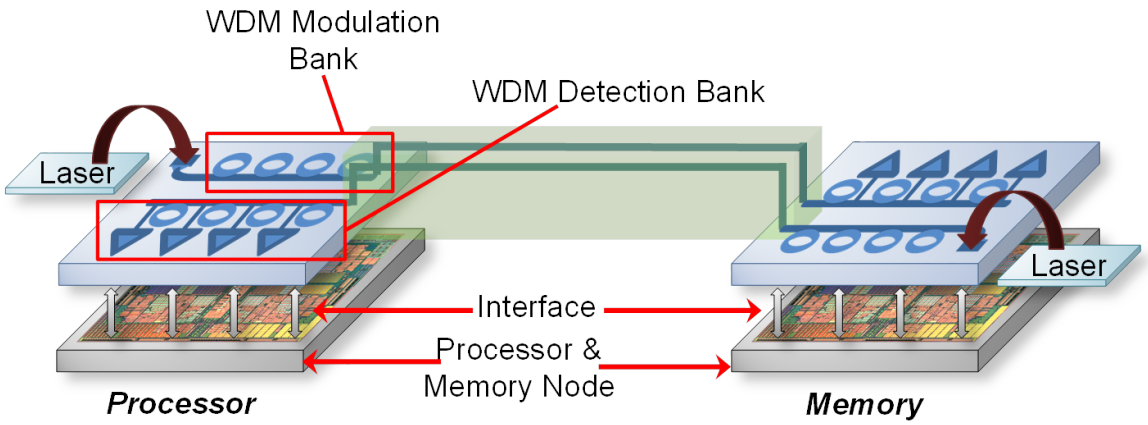

Figure 1.8: Illustration Depicting the Integration of Optics and Computing - Silicon photonic modulators and detectors can be integrated with processors and memory to enable energy-efficient, high-bandwidth communication.

closely packaged with electronic transceivers [48] and processor/memory devices can communicate via non-integrated silicon photonic components [49]. These critical steps demonstrate the potential for optically-enabled processors and memory devices, and motivate the continued research of this area.

1.4.2 Optical Interconnection Networks

Optical interconnection networks (OINs) comprise an attractive solution to the communication bottleneck within future large-scale computing systems [13, 44, 45, 50].

Optical MEMS-based switches, such as those considered for deployment in data centers today, suffer from high switching latency and are thus unsuitable for most networks. Meanwhile, semiconductor optical amplifiers (SOAs) have been demonstrated to provide high-bandwidth, low-latency switching [43, 51] for optical switches. Additionally, silicon photonic devices can create high-bandwidth, low-

latency, energy-efficient switches [52] that can then be used to create large-scale optical networks [53]. However, as mentioned in the previous section, silicon photonics in relatively immature and is not yet suitable for the creation of large-scale optical networks. SOA-based OINs thereby serve as the basis for this dissertation.The developed protocols remain compatible with both SOA-based or silicon photonic switches.

1.4.3 Optically-Connected Memory

Memory interconnect architectures are especially well-suited for the deployment of optical interconnects, and especially optical networks, owing to the performance and energy requirements of main memory systems, as well as the necessary flexibility within a network to support potentially diverse and unpredictable traffic patterns. Leveraging the bandwidth-density and distance-immunity of optics will alleviate pin-count constraints in microprocessors and enable optically-connected memory (OCM) to be more physically-distant from the processor, thus yielding the potential for a greater number of memory devices to be connected and accessed in parallel (Figure 1.9). The latency of these links are dictated purely by the fiber's time-of-flight latency [54], allowing efficient, transparent optical networks to provide low memory access latency. In this way, an OCM system can be constructed with greater performance and capacity, while achieving lower memory access latencies and reduced power consumptions as compared to traditional electrically-connected memory systems.

A large body of research investigates leveraging optical interconnects to address the limited scalability of main memory. The work in [55] explores the impact of

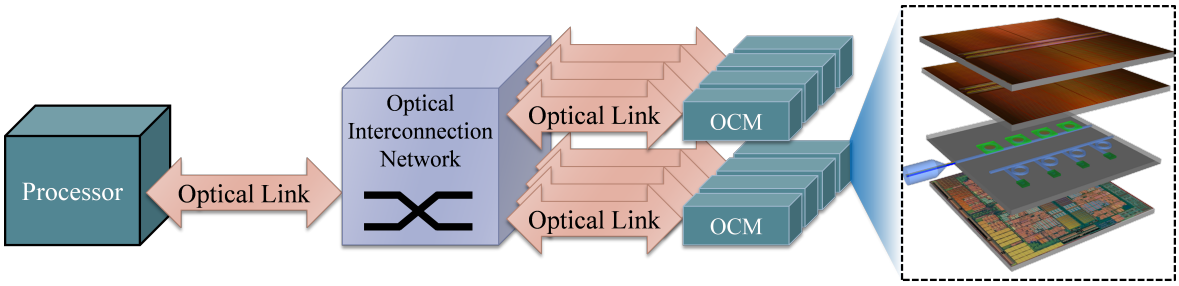

Figure 1.9: Optically-Connected Memory Block Schematic - Block schematic depicting how a next-generation processor can be connected to many optically-connected memory modules across an optical interconnection network. Inset illustrates a memory module that uses 3D-stacking to integrate SDRAM with photonic transceivers and associated driver circuitry.

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 [56] 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 for applications with long, predictable memory access patterns. The authors later explored the use of time-division multiplexed arbitration [57] as a means to improve routing efficiency for more diverse message sizes.

OCDIMM [58] is an optically-connected memory 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 the improved memory access latency when compared to traditional FB-DIMM. The main drawback to this work is the lack of an optical interconnection network. This work effectively creates an optical daisy chain, which suffers from non-uniform latencies that could become unacceptably high as the number of memory devices increases. Each optical message must also pass through several devices in the chain to reach its destination, incurring additional insertion loss, which can also limit scalability.

In [59], the authors propose a substantial redesign of memory devices through monolithic integration of silicon photonics. This implementation assumes that silicon photonic modulators and receivers can not only be successfully monolithically integrated into memory devices, but that they can be operated at per-channel data rates of at least 10 Gb/s. The memory fabrication process is optimized for DRAM cell density, as opposed to processor fabrication that is optimized for speed, and it is therefore unlikely that any devices, electrical or photonic, can operate as proposed in [59].

The most plausible near-term method for integrating memory and photonics is the use of photonic transceiver modules packaged at the processors and memory. This is similar to FB-DIMM and OCDIMM; however, significant performance and scalability gains can be achieved from using an OIN to link many memory devices to the processor as nodes on a network. Each optically-connected memory module (OCMM) could then be accessed with uniform, low-latencies as dictated by the optical network architecture. From the perspective of the processor, each path through the optical network would

be analogous to a traditional memory channel. For example, a packet-switched optical network can provide low memory access latency for short messages, while a circuit-switched optical network delivers greater performance for longer messages (e.g. in a streaming application) [56]. To optimally route heterogeneous memory traffic, a hybrid packet-and-circuit-switched optical network [60] could be deployed between processors and memory.

In the long term, 3D integration can enable optical transceivers to be incorporated directly with processors and memory devices without the need for monolithic integration. For example, multi-layer deposited materials [61] can realize high-performance optical modulators and receivers adjcaent to standard CMOS-fabricated cores. 3D-stacked memory devices, called hybrid memory cubes (HMCs), are currently under development [62] to address the lack of high-speed logic within DRAM fabrication technology. The HMC is an ideal insertion point for photonics into memory because of their 3D-integrated nature and the presence of a high-speed logic layer that can not only interface the memory cells to high-speed logic, but also the high-speed logic to photonic modules and receivers.

1.5 Scope

The primary contribution of this dissertation is the development and implementation of the first ever OCM system. In this system, the electrical bus between a processor and its memory is replaced by an optical interconnection network; to achieve this, processors and memory devices interface with local photonic transceivers. Additionally, this work encompasses the creation of a novel optical-network-aware memory controller

that serves as the optics-computing interface. A series of experiments is presented to characterize the OCM system across the four key metrics that must be addressed by next-generation HPCs and data centers:

- **Bandwidth**: The low bandwidth-density of electrical interconnects [2] and pint-count constraints [3] limit memory bandwidth. OCM must enable a roadmap to exascale memory architectures;
- **Latency**: The lower bound of memory access latency is the access time of each SDRAM cell [37]. Optical networks must be designed to minimize, or eliminate, any additional latency as compared to a traditional electrical memory link;
- **Energy**: Microprocessors dissipate up to half of their energy in the interconnect alone [5], and the electronic bus linking main memory to its processor has become a significant source of power dissipation. Significant power savings can be achieved by eliminating both the electronic memory bus and the associated input/output (I/O) buffers at each end of the link;
- **Resilience**: Increasing the number of memory devices in a system, as with any commodity hardware, increases the probability of a failure within the memory system as a whole [63, 64]. OINs enable novel memory protocols that can improve reliability beyond what is possible using today's electronic interconnects.

This dissertation is organized as per the following. Chapter 2 details the optical switching fabric architectures and their uses within OCM systems. Additionally, experimental results are presented to demonstrate the scalability of the switching fabric

and thus the OCM system. Chapter 3 presents the prototype of the world's first optically-connected memory module (OCMM), which was created based on analysis from Chapter 2 and serves as the remote OCM node within the subsequent detailed experiments. The optical-network-aware memory controller is described in Chapter 4, with experimental characterizations of its impact on memory bandwidth, latency, and energy. Resilient OCM architectures are discussed and experimentally characterized in Chapter 5. This demonstrates that the flexibility provided by optical interconnects provides not only improved bandwidth, latency, and energy performance but additional novel techniques to build resilient large-scale computers. In Chapter 6, the OCM system is demonstrated using silicon microring modulators in place of discrete $LiNbO_3$ modulators. This critical step characterizes for the first time not only the importance of integrated silicon photonics within processor-memory systems but also the impact of unpredictable memory data and traffic patterns on silicon photonic devices. Finally, Chapter 7 summarizes the contributions of this dissertation and describes ongoing and future work toward developing OCM for HPCs and data centers.

Chapter 2

Optical Interconnection Networks for OCM

This chapter details the optical network architecture utilized throughout this dissertation. Optical networks are the fundamental enabling technology for future, optically-connected HPCs and data centers. Here, the network functionalities and protocols are analyzed with regard to their impact on future computing and memory systems. Additionally, an experimental characterization is presented to demonstrate that such optical networks can scale to the sizes and bandwidths required for next-generation large-scale computers.

2.1 Optical Network Architecture

The optical networks utilized throughout this dissertation make use of one or more SOA-based optical switching nodes [65] to implement 2×2 (Figure 2.1) or 4×4 (Figure 2.2) switching fabric testbeds. The switching nodes are modular and may

be linked together to create larger, multi-stage networks. For example, four 2×2 nodes have been configured to implement a 4×4 as a 2-stage Banyan network topology. In this way, an optical network can be implemented with hundreds of thousands of nodes [66].

Figure 2.3 shows the wavelength-striped message format that enables messages to be transparently routed through each switching node. The use of wavelength-striping simplifies routing and minimizes switching latency: routing information is encoded on dedicated header wavelengths that are combined with the memory payload data using WDM. Each header wavelength remains constant, either logical 1 or 0, for the duration of each message. The minimum number of address headers required for an N×N network is $\log_2(N)$. Additional network functionalities, such as optical multicasting [67], can be added to the network by increasing the number of header wavelengths.

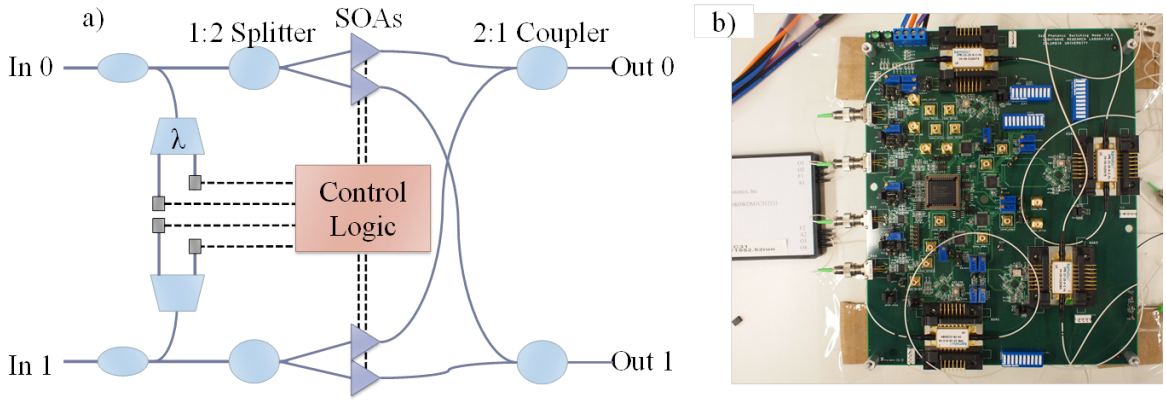

Figure 2.1: 2×2 Photonic Switching Node - (a)Schematic and (b)photograph of the 2×2 SOA-based photonic switching node.

All switching nodes utilize the same broadcast-and-select architecture: wavelength-striped messages enter at an input port where passive optical components (filters, couplers, fiber) direct the appropriate header wavelengths to low-speed (155-Mb/s)

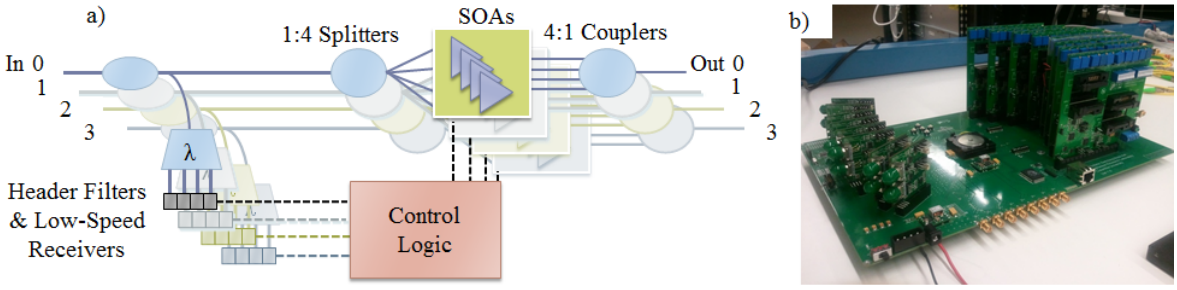

Figure 2.2: 4×4 Photonic Switching Node - (a)Schematic and (b)photograph of the 4×4 SOA-based photonic switching node.

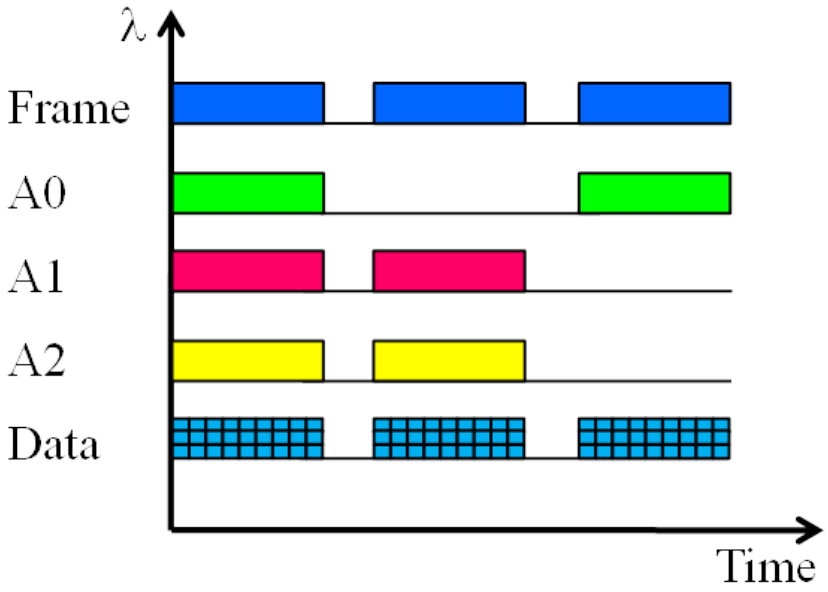

Figure 2.3: Wavelength-Striped Message Format - Low-speed header wavelengths (frame and address data) are combined with high-speed payload wavelengths using WDM.

photodetectors (PD). The electrical header data is then passed to simple, high-speed control logic that in turn gates SOA on or off. The SOAs act as broadband optical gates at each output port, and the gain provided restores power equivalent to any optical losses incurred through the switching node. The payload data, meanwhile, passes through a fiber delay line (FDL) to match the time required to filter, receive, and process the header information (approximately 10 ns). The FDL feeds the payload data into the SOAs, which are enabled just-in-time to allow transparent, low-latency routing. Each 2×2 switching node implements the control logic using a complex programmable logic device (CPLD), while the 4×4 switching nodes instead use a FPGA that allows for more diverse and advanced network functionalities.

2.1.1 Switching Protocols

The photonic switching nodes have been demonstrated to operate as packet switches [51], circuit switches [68], or hybrid packet and circuit switches [60]. Optical packet switching (OPS) is achieved by modulating the header wavelengths such that they are constant for the duration of each packet plus a small amount of guard time at the start and end of each packet. In this way, optical packets ranging from tens of nanoseconds up to milliseconds have been demonstrated with payload data ranging from a single 10-Gb/s wavelength channel up to eight 40-Gb/s WDM channels per packet [65]. The simple, distributed control logic within the switching fabric makes optical packet switching ideal for smaller messages or random traffic patterns. However, large numbers of larger messages may need to be broken down into multiple packets that may then overload the network or incur undesirable latency

Optical circuit switching (OCS) is ideal for larger, less random messages. Optical circuit paths, or light paths, can be created through the network through the use of a low-speed electrical control plane, such as over ethernet [4] links between the end-point nodes and the optical switching nodes. By opting for a single, up-front latency penalty to each transaction, a circuit switch eliminates the many, smaller latency penalties associated with packet-switched networks. Simulations of integrated, silicon photonic optical circuit-switched networks [53, 56] have shown the potential for significant power and performance benefits as compared to electrical packet- or circuit-switched networks. This is especially true for streaming applications, which exhibit predictable communication patterns involving large amounts of data.

OCM systems may exhibit unpredictable communication patterns involving small or large messages. It is therefore desirable to allow each end-node to analyze its own outgoing traffic and dynamically use either packet- or circuit-switching. The control logic for both previously described packet- and circuit-switching implementations can operate simultaneously within the photonic switching nodes, which requires only minimal additional control logic to arbitrate between conflicting packet and circuit traffic. Using this protocol, a memory controller can analyze the pattern of memory access requests from a processor to transmit each request (or aggregated group of requests) as packets or circuits [69].

2.2 Implications for Memory Architectures

The latency characteristics of an optical interconnect approach to main memory 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 the photonic switching nodes reduces the overall latency to the time-of-flight between a processor and OCMMs. Each additional meter of single-mode fiber (SMF) adds approximately 5 ns [54] 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. One goal of this OCM design is 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 [70] expect a global memory address space but already exploit locality to maximize references to a local memory. In this dissertation, local memory is the memory with the shortest optical path.

The OIN described here 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 [65, 66]. Meanwhile, high-speed transceivers operating at high per-channel 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 [4]. Using multiple high-speed transceivers with WDM creates the bandwidth density necessary for OCM nodes with memory bandwidth in the 100's of gigabits per second. Many OCM nodes could be further combined using an optical interconnection network for petabit system bandwidths.

The reconfigurability of the proposed OCM system supports the diverse applications

that data centers and HPCs may run. For example, a streaming application [71] 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. Additionally, a web search application with unpredictable communication can leverage a hybrid packet- and circuit-switched OCM system to optimize communication on a message-by-message basis.

The network nodes can be configured to support wavelength-striped payload multicasting to enable access to multiple OCM nodes simultaneously [72]. 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 multicast-capable OCM ssytem can be configured for resilience to tolerate the failure of an entire OCMM while still maintaining error-free performance [69]. The use of multicasting can also be used to transmit along several redundant paths to the same destination node.

Reducing the power consumption of the memory system can have significant impact on overall data center and HPC designs, and huge power savings can be achieved by incorporating integrated photonic components [47] in future OCM nodes by gaining benefits from the tighter integration of optical components with electronic driver and receiver circuitry [56, 58, 59]. 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, each memory link can

operate at 1 pJ/bit energy efficiency [39, 40]. Continued development of silicon photonic technology is likely to improve this further. Meanwhile, with the growing use of high data rate serial links in HPCs and data centers today, the SerDes power consumption will continue to decrease [73] to reduce the cost of the electrical drivers for the optical transceivers.

2.3 Scalability

Adopting high data rates and advanced modulation formats in optical interconnection networks will enable future computing systems to achieve greater performance and scalability. Traditionally, the simplicity of non-return-to-zero on-off keying (NRZ-OOK) has made it a popular modulation format for optical communication links. However, phase-shift keying (PSK), specifically differential-binary-phase-shift-keying (DBPSK or DPSK), has long been known as a potential method for improving resilience to nonlinearities, with a 3-dB improved receiver sensitivity (with balanced detection) as compared to OOK. Such benefits of advanced modulation formats are becoming more significant as data rates now exceed 10 Gb/s and even 40 Gb/s, and the optical network elements are beginning to exceed the abilities of the driver and receiver electronic circuitry. Current microprocessors already dissipate up to half of their energy in the interconnect alone [5], and scaling up per-channel data rates without optimizing overall communication efficiency would result in an undesirable increase in energy dissipation. Therefore, next-generation computer systems may require optical interconnection networks that utilize not only high per-channel data rates, but also advanced modulation formats with improved resilience and spectral efficiency [74].

This section presents an experiment that demonstrates the scalability of the photonic switching nodes utilized throughout this dissertation through ability to transparently route data modulated at 10 Gb/s to 40 Gb/s, modulated as OOK or DPSK, with negligible impact on power penalties.

2.3.1 OOK Characterization

The first step in demonstrating the OIN scalability is to determine the performance impact, in terms of power penalty, of increasing per-channel data rates from 10 Gb/s to 40 Gb/s while maintaining OOK modulation. To accomplish this, four 2×2 photonic switching nodes are configured as a 2-stage 4×4 optical network test-bed (Figure 2.4). Payload data consists of 8 separate wavelengths combined intro a single WDM optical packet. Each 64-ns optical packet is created using eight distributed feedback (DFB) lasers, which are passively combined and modulated with OOK data by a $LiNbO_3$ modulator. The 10-Gb/s OOK data is generated by a pulse pattern generator (PPG), using a $2^{15}-1$ pseudo-random bit sequence (PRBS), and a 10-Gb/s $LiNbO_3$ modulator. Separately, the 40-Gb/s OOK data is generated by using an identical PPG, using $2^{15}-1$ PRBS, and an electrical multiplexer to create 40-Gb/s PRBS. Each optical packet is combined with three, independently-generated network control signals, creating wavelength-striped optical packets that pass transparently through the optical test-bed.

Error-free transmission of all optical packets is experimentally measured at the output of the network using a bit-error-rate tester (BERT), with bit-error rates (BERs) less than 10^{12}, for all eight payload wavelengths using 10-Gb/s and 40-Gb/s

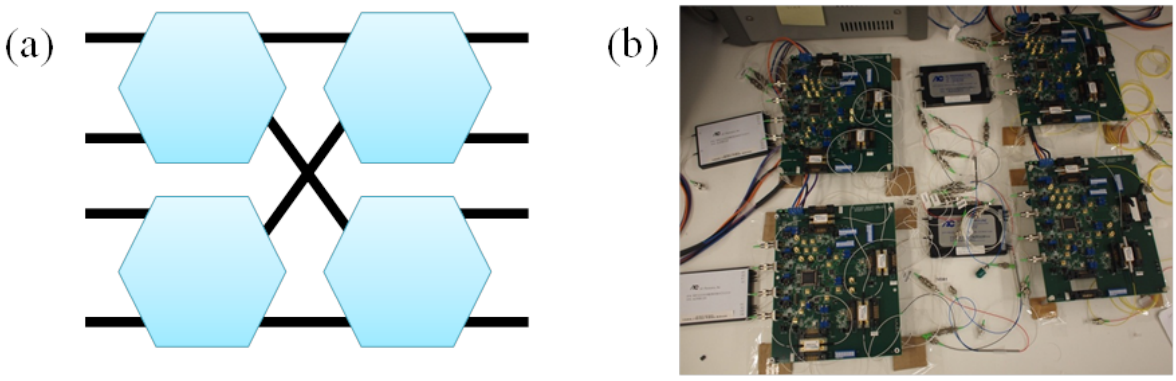

Figure 2.4: 2-Stage 4×4 Network Test Bed - (a)Block diagram of 4×4 optical network (b)photograph of implemented optical interconnection network.

OOK modulation. Figure 2.5a shows the optical packets at the input and output of the network. Figure 2.5b shows the optical eye diagrams for the 40-Gb/s payload wavelengths.

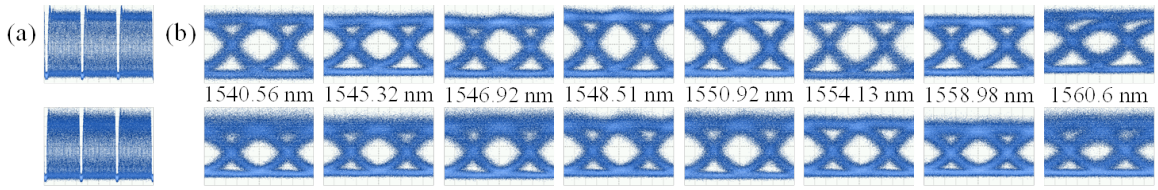

Figure 2.5: Scalability Characterization: 40-Gb/s OOK Optical Eye Diagrams - Optical eyes at the input (top) and output (bottom) of the optical network showing (a) packets; (b) each 40-Gb/s OOK payload wavelength (5 ps/div.

The sensitivity curves shown in Figure 2.6 demonstrate a power penalty of 1 dB for the best-case payload wavelength of 1550.92 nm (0.5 dB per SOA hop) and 1.1 dB for the worst-case payload wavelength of 1560.6 nm (0.55 dB per SOA hop). The small variation in power penalties is likely due to the slight gain curve of the erbium-doped fiber amplifier (EDFA) used in BER testing. In comparison, 10-Gb/s data typically has a power penalty of 0.45 dB per SOA hop [75]. These values are similar, however the

slight increase in power penalty for 40-Gb/s data could become a design consideration for future networks as network sizes and per-channel data rates continue to increase.

2.3.2 DPSK Characterization

The next step in the scalability study is to demonstrate network transparency with regard to modulation formats by comparing amplitude modulation (OOK) to phase modulation (DPSK). With that aim, optical packets with 8×40-Gb/s DPSK data are then generated and transmitted through the test-bed. This second experimental setup (Figure 2.7) consists of the above transparent, 2-stage 4×4 optical test-bed, again correctly routing 64-ns wavelength-striped optical packets. A 40-Gb/s PPG drives a single 40-Gb/s phase modulator to encode $2^{15} - 1$ PRBS DPSK data onto eight payload wavelength channels, each with an average power of -14 dBm at the network input and with wavelengths ranging from 1533.47 nm to 1560.6 nm. Immediately after modulation, and before entering the network, the WDM data passes through a 2-km span of single mode fiber for decorrelation. The three separate header wavelengths (frame, address0, address1) are modulated at the packet rate using three external SOAs, which are controlled by an Agilent ParBERT.

In the experiment, the combined eight payload wavelengths are gated using a single SOA, controlled by the ParBERT, to create 64-ns optical packets. The resulting configuration is such that each optical packet consists of 8×40-Gb/s DPSK wavelength channels and 3 low-speed header wavelength channels. As before, all eleven wavelengths traverse the optical interconnection network concurrently as a single wavelength-striped packet. The amplification provided by the SOA-based switching nodes within the

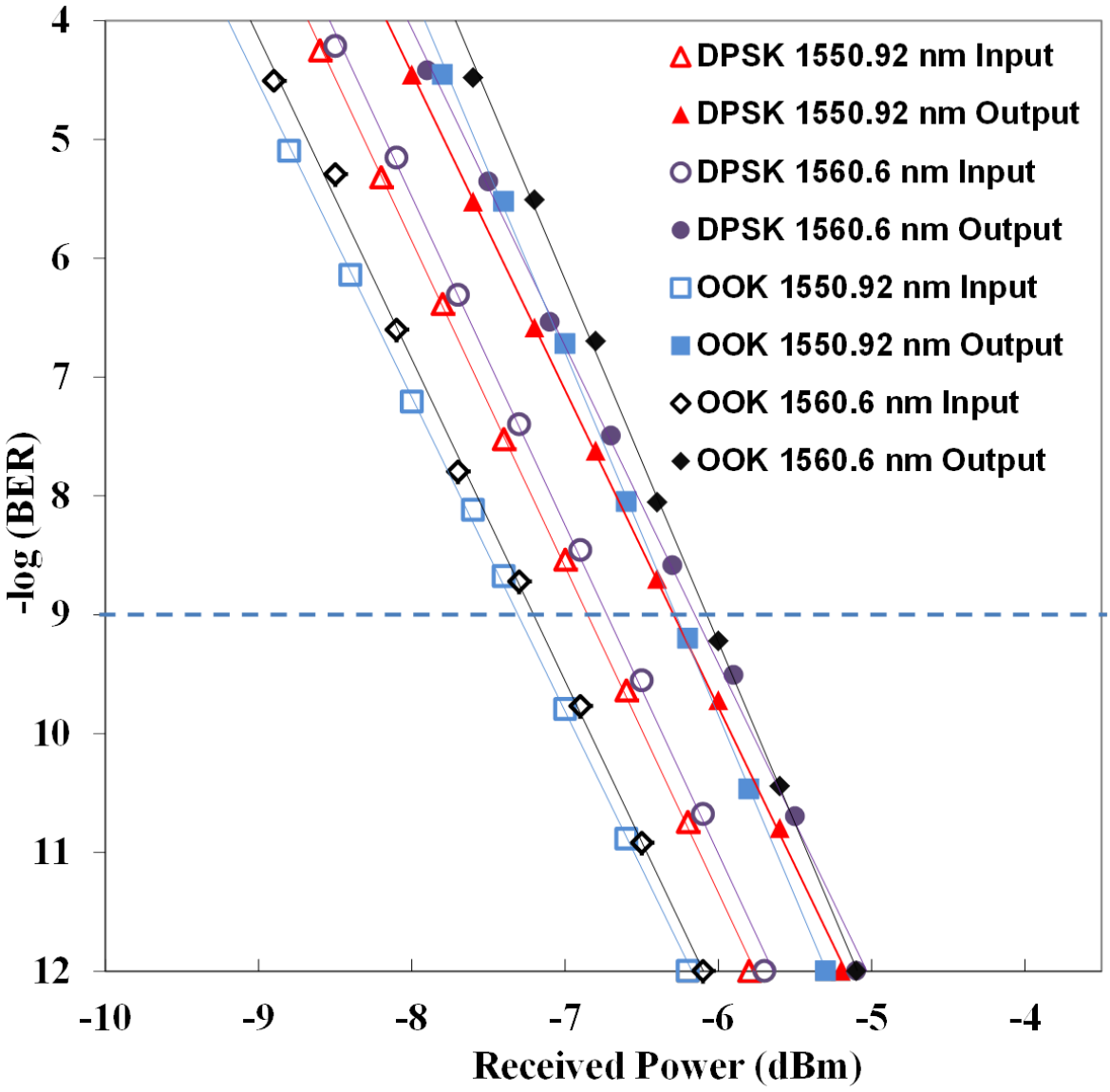

Figure 2.6: Sensitivity curves for 40-Gb/s OOK and DPSK Scalability Characterization - BER curves for best case (1550.92 nm) and worst case (1560.6 nm) 40-Gb/s OOK and DPSK payload wavelengths. Input measurements are shown with open points, while output measurements are depicted with filled points. The power penalty spread from best- to worst-case is 0.1 dB for OOK and >0.1 dB for DPSK at 10^{-9} BER.

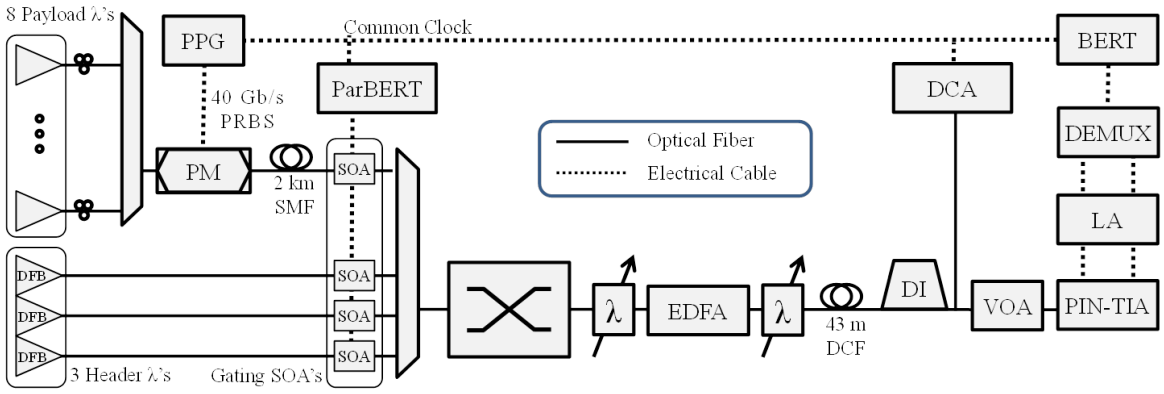

Figure 2.7: Scalability Characterization: Experimental Setup for DPSK Traffic - Experimental setup for 40-Gb/s, 8-channel DPSK network characterization.

optical test-bed maintains this average power during propagation through the network and to the network output.

At the output, the eight payload wavelengths are sent to a tunable filter to select a single 40-Gb/s DPSK signal, which is amplified using an EDFA and filtered again. The single wavelength is then sent to the constructive port of a delay-line interferometer (DLI), a variable optical attenuator (VOA), and a 40-Gb/s photodiode with transimpedance (TIA) and limiting amplifiers (LA). This received electrical data is time-demultiplexed and verified using a BERT, which is gated for packetized data by the ParBERT. The 40-Gb/s optical signals are simultaneously inspected using a digital communications analyzer (DCA).

Error-free operation is confirmed for all eight 40-Gb/s DPSK payload wavelengths with BERs less than 10^{-12} at the output of the optical network test-bed. The optical packets at the input and output are shown in Figure 2.8a, and Figure 2.8b shows the input and output 40-Gb/s optical eye diagrams for all 8 payload wavelengths at the

output of the constructive port of the DI. A best-case power penalty of 0.52 dB was measured at a BER of 10^{-9} (Figure 2.6) for the 2-stage network at 1550.92 nm. The worst-case power penalty of 0.56 dB was measured at 1560.6 nm. This demonstrates an average power penalty of 0.26 dB per SOA switch hop when transmitting 40-Gb/s DPSK data at 1550.92 nm, and a power penalty spread >0.1 dB for all eight wavelengths ranging from 1533.47 nm to 1560.6 nm.

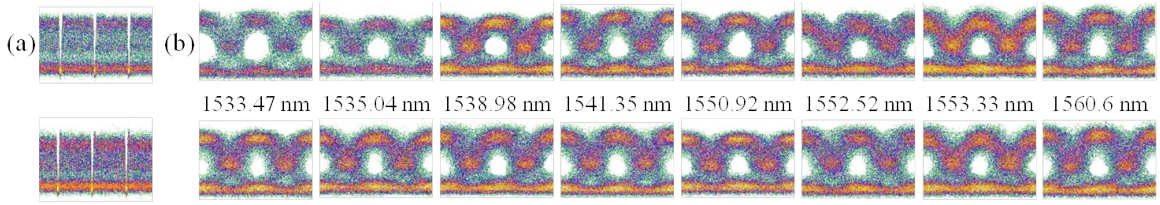

Figure 2.8: Scalability Characterization: 40-Gb/s DPSK Optical Eye Diagrams - Optical eyes at the input (top) and output (bottom) of the optical network showing (a) packets; (b) each 40-Gb/s DPSK payload wavelength (5 ps/div).

2.4 Discussion

Optical interconnection networks are critical components of future optically-connected memory systems that allow high-performance, energy-efficient integrated optical links to leverage low-latency, transparent routing through the large-scale networks required for next-generation HPCs and data centers. This, in turn, enables the deployment of OCM systems with the necessary memory capacity and bandwidth to close the memory-processor performance gap.

The optical network test bed described above is experimentally demonstrated as a data rate- and modulation format-transparent, multi-terabit optical interconnection

network [76]. The comparable power penalty for 40-Gb/s DPSK data (0.26 dB per SOA switch) versus 40-Gb/s OOK (0.5 dB per SOA switch) shows that this network can continue to operate in future HPCs in data centers regardless of the modulation type required for improved spectral efficiency and overall network scalability. In this way, low-latency optical networks can be deployed between processors and main memory to enable high-bandwidth memory access with ultra low-latencies. For this reason, the remainder of this dissertation makes use of the above optical network test bed for all experiments.

Chapter 3

Optically-Connected Memory Modules

In this chapter, the first prototype high-performance OCMM (Figure 3.1) is presented, which enables the implementation and characterization of diverse OCM architectures. The main advantage of using this OCMM is the close integration of DDR3 SDRAM with high-speed serial transceivers by means of a high-performance FPGA. The FPGA provides inexpensive, flexible, and fast prototyping functionalities that are not possible using application-specific integrated circuits (ASICs) or off-the-shelf processors. This, in turn, creates a low-latency interface between the electrical domain within the physical memory devices and the optical domain within the network.

The OCMMs may be located physically distant from the associated processors, leveraging the distance-immunity (at computer-scales) of single-mode optical fibers [54]. Relocating the memory devices to be physically distant from the processor frees up board space near the processor itself without significant impact on system performance [55]. This will give system designers more flexibility in designing board

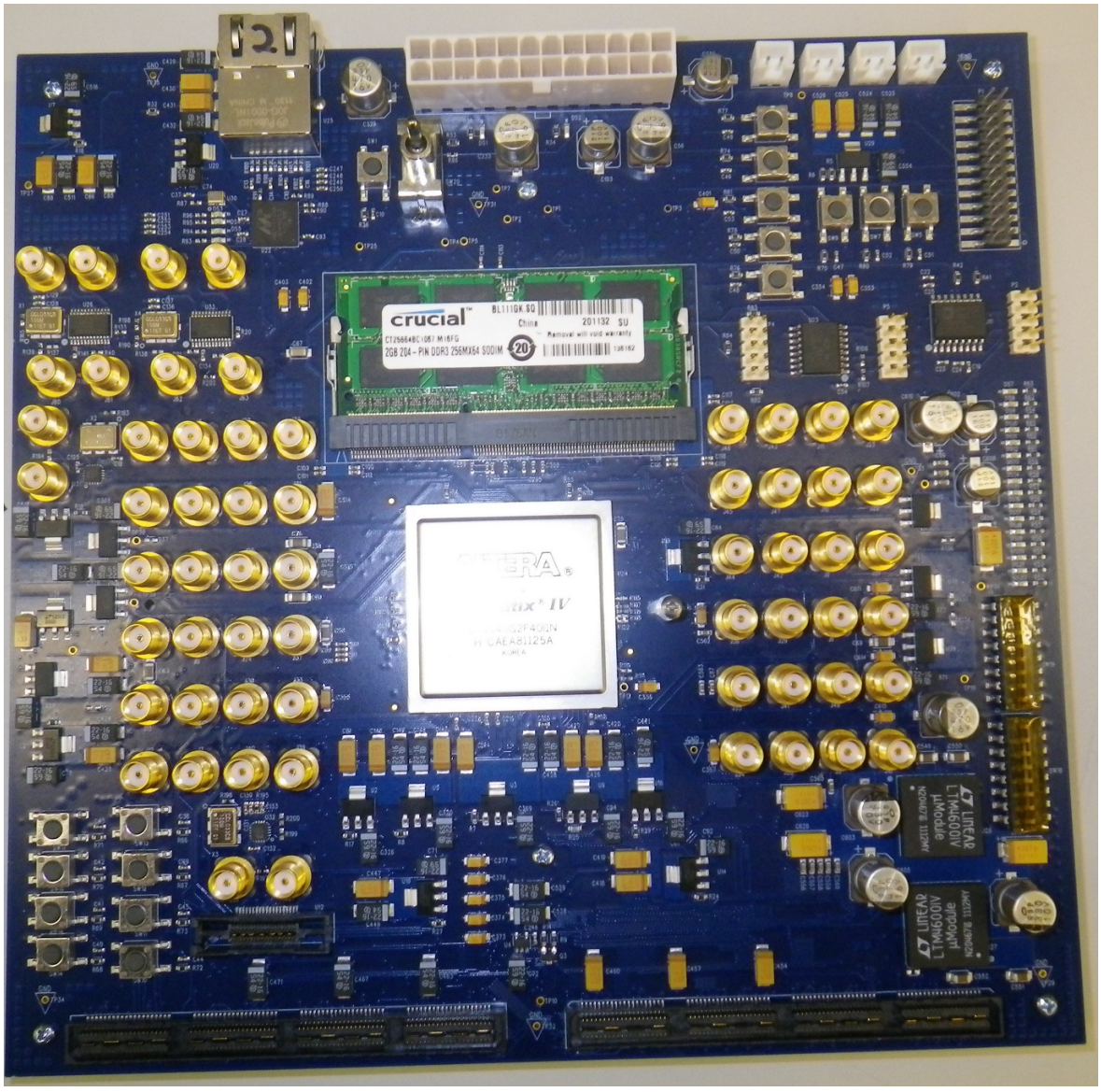

Figure 3.1: Photograph of OCMM - Photograph of the first OCMM prototype.

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.

3.1 Lessons from Preliminary Work

The OCMM design draws upon preliminary work on OCM [68, 72, 77, 78] and seeks to optimize all necessary functionalities required for high-performance operation of optical-network-accessed memory. Throughout these earlier experiments, the overall memory performance was severely limited by the use of an off-the-shelf FPGA board that operated at relatively slow speeds. Additionally, the on-board SDRAM was a fixed component on the board and therefore it was not possible to explore different memory types, sizes, or speeds.

3.1.1 Memory Interface Latency

As explained in Chapter 1, microprocessors and FPGAs were not designed to interface with optical components or networks. Similarly, due to the limitations of electrical wires, the memory on commercial boards, including off-the-shelf FPGA boards, was not intended to directly interface with high-speed transceivers for off-board communication. When combined with the relatively slow speed of the FPGAs used within the early OCM experiments, these factors resulted in unacceptable memory access latencies of hundreds of nanoseconds. For comparison, standard memory communicating to

a nearby, on-board processor will incur latencies of tens of nanoseconds [6, 36]. The high latencies within early OCM demonstrations limited the amount of useful characterizations that could be performed on the implemented memory systems, such as network utilization analysis, and motivated the development of an integrated memory-transceiver interface.

3.1.2 Bandwidth Matching

The ideal configuration for an OCM system is to match the high-speed serial transceiver bandwidth to the maximum bandwidth delivered by the DIMMs, which in turn should be matched to the bandwidth capacity of the processor's memory controller. In contrast to an electrical network, as described in Chapter 2, the OIN that serves as the processor-memory link can operate efficiently for a wide range of message formats and does not need to be modified for bandwidth matching.

Preliminary OCM experiments relied on the Altera Stratix II FPGA [79] and contained only sufficient serial transceivers to create an aggregate 10-Gb/s optical memory link between a processor and its main memory. The DDR2 memory present on the board was capable of double that amount of bandwidth, and therefore the FPGA and electrical serial transceivers were limiting the overall system performance and required additional complex logic and buffering to operate the memory link at one half the offered memory bandwidth. To address this issue, the OCMM contains an Altera Stratix IV [80] FPGA with twelve bi-directional 11.3-Gb/s transceivers to provide 135.6 Gb/s aggregate bandwidth from the on-board memory to the optical network, which is approximately the same as the maximum bandwidth rating of a

single state-of-the-art SDRAM DIMM [81]. By matching memory bandwidth to the network communication bandwidth, overall system performance is improved and more accurate OCM system characterizations can be performed.

3.1.3 Burst-Mode Transceivers

Early OCM implementations also suffered from unacceptable clock and data recovery (CDR) overhead. CDR is the process by which high-speed serial links operate without a shared clock [82], and requires a receiver to align a phase-locked loop (PLL) based on transitions (logical 0's or 1's) in the incoming serial data stream. The time required to lock the PLL and thus recover a clock to read the incoming data is referred to as the CDR time, which adds latency to any communication requiring a lock to be established. Traditional electrical links establish a link once, often during system power-up, and maintain the link indefinitely by transmitting either useful data or "idle data" (such as a stream of alternating ones and zeroes) at all times. As a result, little focus has traditionally been placed on faster CDR techniques and the overhead is typically on the order of several microseconds.

Recently, however, the trend toward energy efficient computing has raised interest in shutting down, or drastically reducing the data rates, of links when data is not being transmitted. While potentially saving energy, the process of repeatedly disabling and re-establishing a serial link will require frequent CDR overhead instead of only a single occurrence during system power-up. Additionally, the move toward optical links and switching [30, 65, 83, 84] may prevent the transmission of idle data and therefore increase CDR overhead. In order to leverage the performance and energy benefits

of optical interconnects without suffering from frequent CDR overhead, each instance requiring several microseconds, recent attention has focused on high-speed burst-mode receivers [85, 86]. Burst-mode receivers are designed to operate with unpredictable traffic patterns and in the presence of "dead time" (i.e. no idle data) between messages, and therefore reduce the CDR overhead to less than 10 ns [87] in laboratory settings.

Commercial processors and FPGAs do not contain such fast burst-mode receivers, and therefore integrating any existing processors or FPGAs with optical switching will suffer from unacceptably high CDR overhead. To address this issue within the OCMM prototype, high-bandwidth expansion slots were created on the OCMM circuit board to enable the implementation of fast burst-mode transceivers such as those in [87]. A separate daughter card can therefore be created that implements the burst-mode receivers and serializer-deserializer (SerDes) functionality, which would process the serial data and then deliver the equivalent parallel data and clock to the FPGA. This process bypasses the serial links of the FPGA, and therefore avoids the relatively high CDR overhead inherent in the FPGA's serial transceivers. The use of expansion slots in place of an integrated, on-board burst-mode receiver allows for various burst-mode receiver implementations, which is necessary due to the nature of ongoing burst-mode receiver research and the lack of commercial options.

3.2 OCMM Implementation Details

The OCMM consists of: an Altera Stratix IV FPGA [80] with twelve bi-directional transceivers each capable of up to 11.3-Gb/s operation, a replaceable DDR3 DIMM [6], a 10/100-Mb/s ethernet port [4], expansion ports capable of supporting burst-mode

transceivers with over 135-Gb/s aggregate bandwidth, and multiple banks of general purpose inpput/output (GPIO) pins. Figure 3.2 shows the physical mapping of each component within the OCMM. The DIMM utilized here is a DDR3-1066 module that can provide approximately 70-Gb/s peak memory bandwidth. This memory module was selected to match the peak bandwidth capacity of the memory controller, which is implemented as a separate node on the OIN and is detailed in Chapter 4.

The configuration of the OCMM (Figure 3.3) is such that off-the-shelf SDRAM chips communicate to a local photonic transceiver chip, which serves are the interface between the OIN and memory. Here, the photonic transceiver chip is a combination of the FPGA, with its high-speed serial transceivers and SerDes logic, and discrete optical components. The purpose of the FPGA at the OCMM is therefore only to relay data, while converting between parallel and serial, between the memory devices and the photonic transceivers with minimal latency. This latency is approximately 30 ns, which is the time required for the memory data to enter the FPGA, pass through the SerDes logic, and exit the FPGA with logic operating at 353.125 MHz. An additional 30 ns latency from FPGA logic effectively doubles the memory access latency as compared to a processor accessing local electrically-connected DIMM, however this is a significant improvement over the hundreds of nanoseconds incurred by the earlier, pre-OCMM implementations.

Future, commercial OCM implementations could further integrate the memory and photonic transceiver chip [48] to drastically reduce time-of-flight latency. Additionally, replacing the FPGA with a high-speed ASIC, operating at gigahertz speeds rather than hundreds of megahertz, could thus reduce the total electronic/optical (E/O) interface

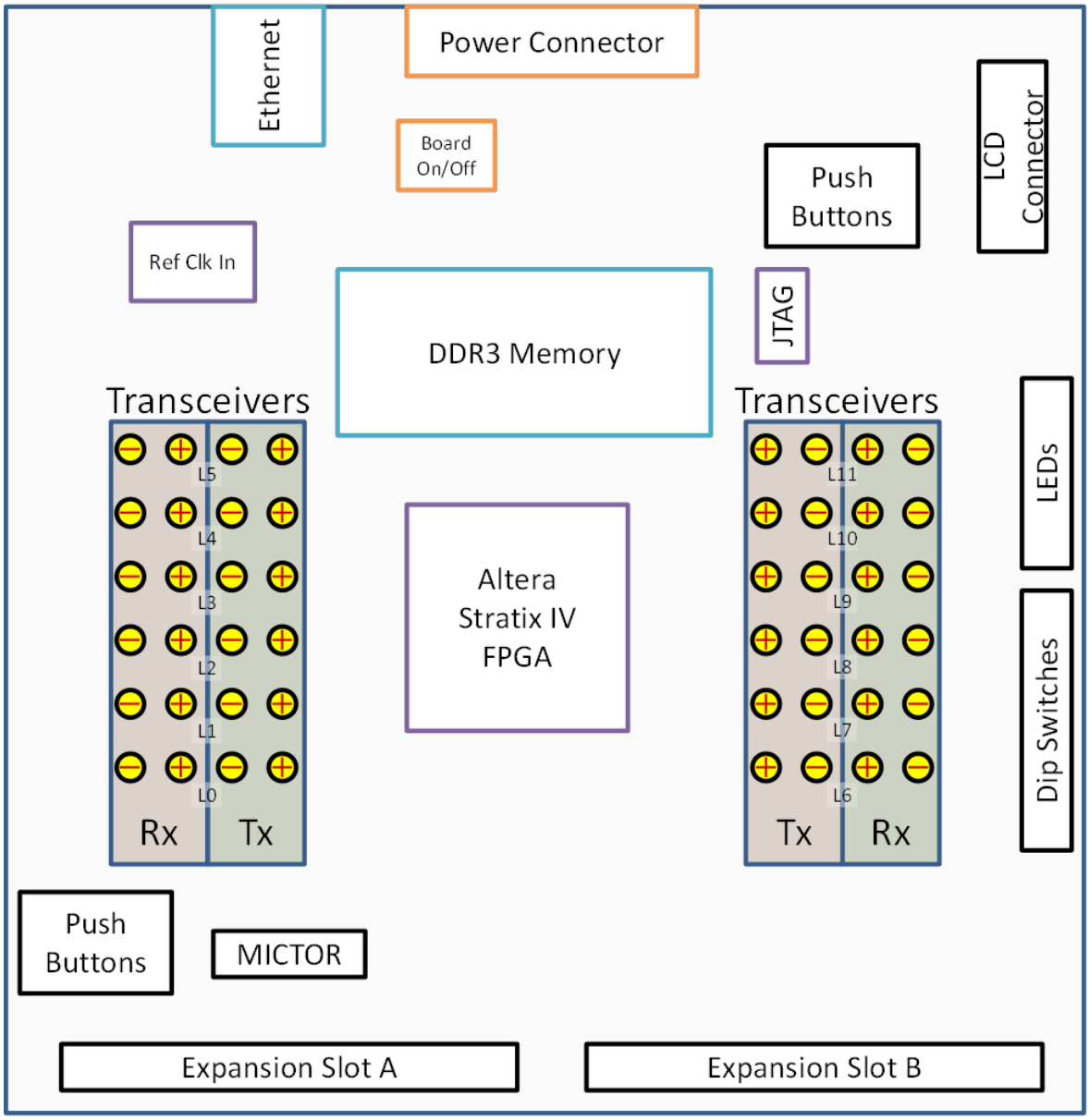

Figure 3.2: Diagram of OCMM Highlighting Significant Components - Diagram of OCMM identifying the locations of the Altera Stratix IV FPGA, DDR3 DIMM, 10/100-Mb/s ethernet port, bi-directional transceivers, expansion ports, and various banks of GPIO.

latency to below 10 ns. An optimized OCMM would therefore provide high-bandwidth, energy-efficient communication across an optical network while only incurring negligible additional latency, which is only a fraction of the tens of nanoseconds latency inherent to SDRAM.

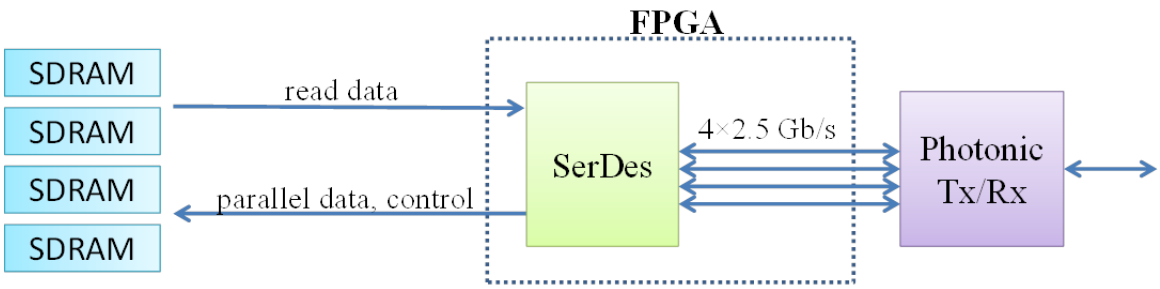

Figure 3.3: Architectural Diagram of OCMM Functionality - An FPGA and discete photonic components serve as a logical photonic transceiver chip for the OCMM's memory devices (SDRAM).

3.2.1 FPGA Hardware Structures

The hardware structures implemented within the FPGA are created using the Verilog Hardware Description Language (HDL) [88]. The top-level hardware modules are the SerDes and SDRAM data path logic (Figure 3.4). As previously described, the SerDes module receives serial data to be de-serialized and relayed to the SDRAM, or serializes SDRAM data to relay to the photonic transceivers, with minimal latency. However, in order to insure that the FPGA-SDRAM communication operates at the proper voltage levels and according to SDRAM specifications [6], an SDRAM data path module is implemented within the FPGA between the SerDes and the SDRAM output pins.

In a traditional electrically-connected memory system, the SDRAM data path logic

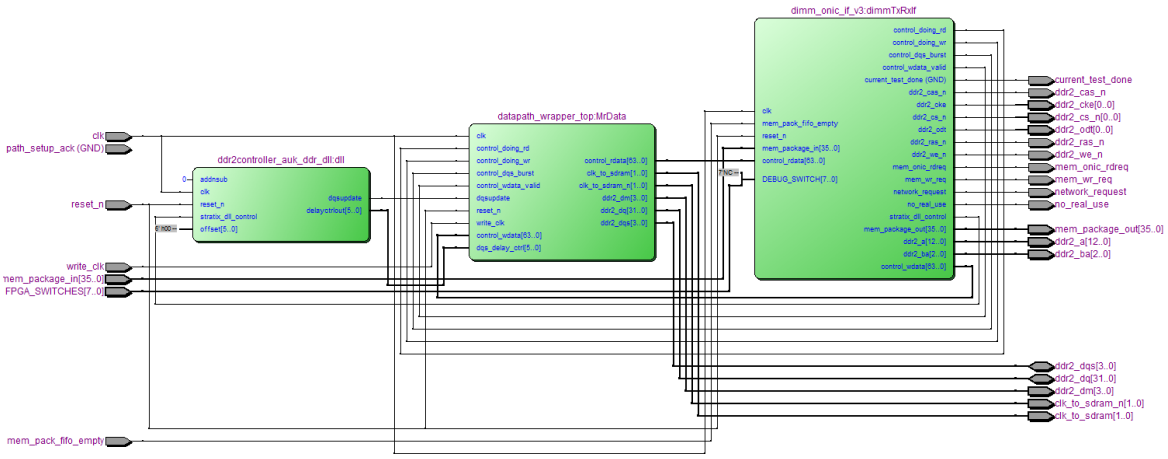

Figure 3.4: Graphical Representation of OCMM Verilog Top-Level FPGA Modules - The primary modules implemented within the FPGA are the SerDes (dimm_onic_if_v3:dimmTxRxIF) and SDRAM data path (ddr2controller_auk_ddr_dll:dll and datapath_wrapper_top:MrData).

would be co-located with the memory controller and possibly the processor itself. Here, however, the SDRAM is physically separate from the memory controller and processor, and the data path logic must therefore be located at the memory side of the optical link. The data path logic is relatively simple (Figure 3.5) and adds negligible complexity to the OCMM. Its main task is to coordinate the "data strobe" signal, which is operated by both the data path logic and the SDRAM to signal that memory data is being transmitted on the electrical memory bus between the FPGA and SDRAM. The FPGA operates the data strobe when it is transmitting memory data to the SDRAM. Otherwise, the SDRAM will operate the data strobe signal when transmitting data to the FPGA. It is therefore necessary to co-locate this small amount of logic with the SerDes at the OCMM.

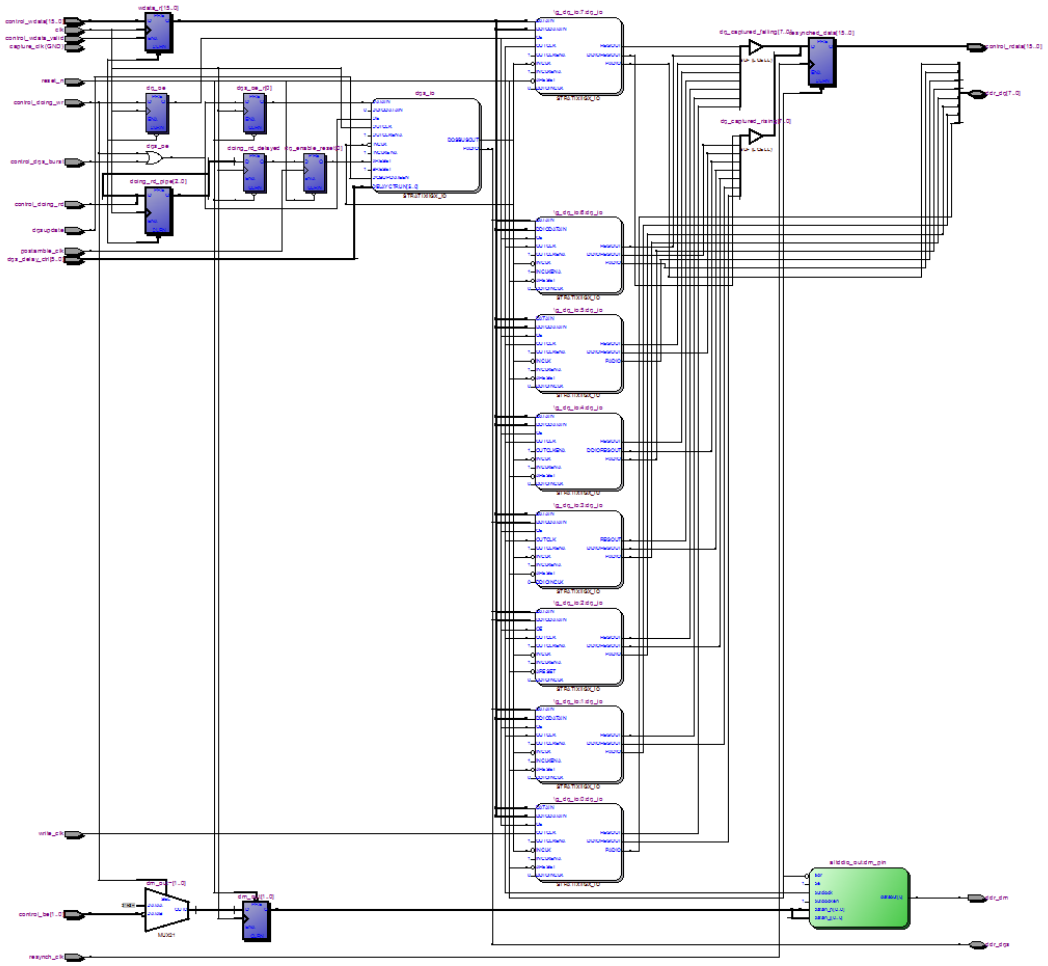

Figure 3.5: Graphical Representation of OCMM Data Path Components - The data path logic controls a "data strobe" signal and synchronizes the memory data to operate the SDRAM according to the DDR3 specification [6].

3.3 Summary

The OCMM is a first-of-its-kind memory module that enables efficient, high-performance all-optical communication between a processor and physically-remote OCM. By drawing on the lessons and experience from experimentally charactering OCM architectures using off-the-shelf FPGA-based circuit boards, the custom OCMM was created to address the challenges facing OCM development that would otherwise remain unknown; including hardware-induced latency and the importance of burst-mode receivers. This, in turn, allows novel memory architectures to be developed, deployed, and experimentally characterized.

Future work will continue to advance and integrate the functionalities of the OCMM with a focus on the use of silicon photonics. This may one day lead to a 3D-integrated OCMM, such as one based on the HMC (Figure 3.6), that can completely eliminate off-chip bandwidth limitations facing memory today. The memory access protocol, which is the optical communication between the processors and OCMMs, may also be modified to optimize performance for diverse optical network architectures and functionalities.

This work represents a first step in the integration of optics into memory devices. The use of optical transceivers and optical networks for memory architectures ensures that future computer systems can efficiently access vast banks of memory storage with bandwidths and energy-efficiencies that are not possible using electronic interconnects. By developing and experimentally characterizing novel optically-connected memory architectures, the challenges that could prevent commercialization and widespread deployment of OCM systems can be identified and overcome.

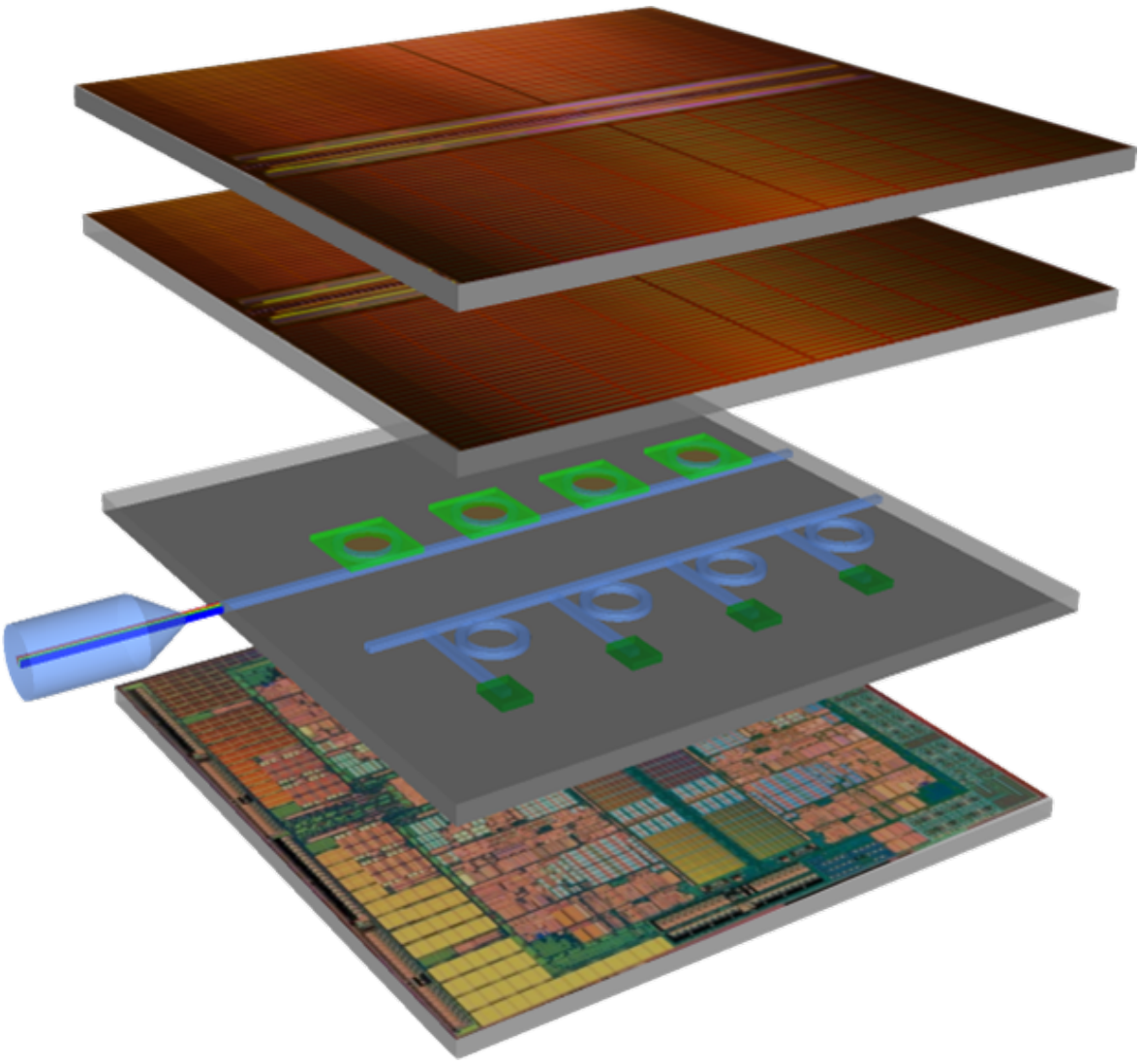

Figure 3.6: Illustration of Future OCMM based on HMC - Future OCMM using 3D integration to combine SDRAM with silicon photonic transceivers.

Chapter 4

Optical-Network-Aware Memory Controller

This chapter describes the network-aware memory controller that was developed to optimize processor-memory communication across a transparent, all-optical memory link by means of an optical interconnection network. This results in a novel memory access protocol that leverages the wavelength-striped network architecture while simultaneously abstracting away the optical network from the processor and memory. As a result of the network configuration, the term "processor node" refers to a network endpoint that contains one or more processing cores without local SDRAM, while the term "memory node" refers to a network endpoint that contains one or more OCMMs without local processing cores.

As in a traditional electronic memory system, the memory controller is co-located on-chip with the processing core(s). This configuration minimizes the latency associated with the round-trip memory request protocol between the processor and memory controller. In this work, the processor and network-aware memory controller

are implemented using a high-speed FPGA (Figure 4.1) identical to the one described in Chapter 3. This enables efficient, low-latency communication between the FPGA-based memory controller and the high-speed transceivers that interface with the optical network.

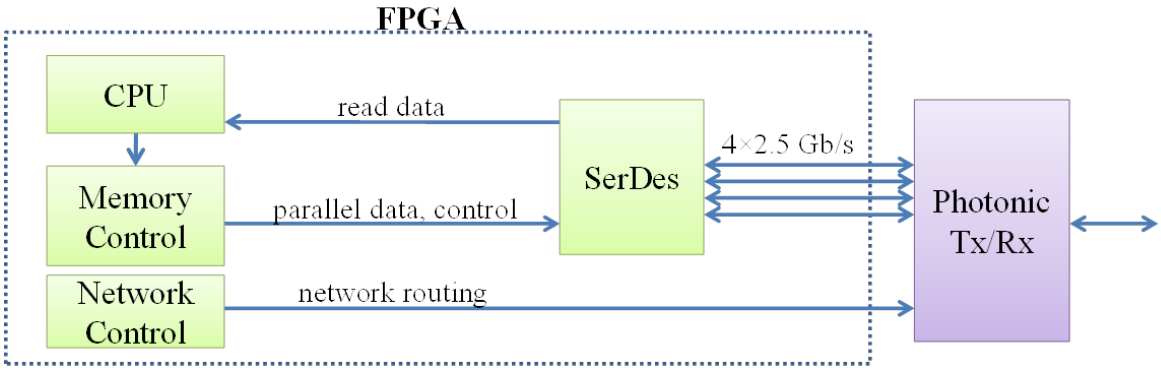

Figure 4.1: Architectural Diagram of a Processor Node - An FPGA implements the processor along with a custom memory controller (Memory Control, Network Control, and SerDes). The high-speed, serial memory transaction combines with the low-speed network routing information at the photonic transceivers using WDM, and traverses the optical interconnection network as a wavelength-striped optical memory transaction.

The replacement of the wide, electronic memory bus with an optical network is made possible by single-mode optical fibers [54], which can provide HPCs and 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 [89] for high bandwidth density, as well as the fabrication of energy-efficient transceivers [39, 40].

This work develops a network-aware memory controller with the goal of exploring the architectural impact of optical interconnects as a means of reaching unprecedented growth and flexibility in large-scale computer systems. Without the limitations imposed by transporting high bandwidth memory data across long electronic traces, many more SDRAM chips can be used at each OCM node than is possible in electrically-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 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 [78]. The memory protocol design first incorporates a circuit-switched optical interconnect [90], which has been shown to provide higher bandwidth density, improved power transmission, and reduction in overall application execution time [56]. Then, a hybrid packet- and circuit-switched implementation is used to allow for greater diversity in communication patterns and therefore applications.

4.1 Novel Memory Access Protocol

A typical memory system is based on a microprocessor initiating write-to-memory and read-from-memory transactions. Throughout the experimental demonstrations presented in this this thesis, the processor functionality is emulated using a custom memory traffic generator to create programmable, verifiable memory transactions. This emulated processor is based on a parallel programming model that is communication-

intensive, and therefore requires sustained high-bandwidth access to memory, while spending a relatively small amount of time processing without active communication. The resulting memory access pattern is such that most of the application run-time involves reads-to- or writes-from-memory.

In order to fully realize the benefits of the optically-connected memory system, the custom memory controller is designed for efficient control of the optical interconnection network. Therefore, the memory controller has been implemented with a focus on optimizing an optical-network-aware memory transaction protocol. General memory controller optimizations, such as intelligent transaction scheduling schemes, are important for electrically- and optically-connected memory, but have been investigated elsewhere [91, 92, 93] and are not the focus on this thesis.

The network-aware memory controller is implemented in the same FPGA as the processor, which creates the industry-standard configuration of a processor with an on-chip memory controller. The top-level hardware modules implemented within the FPGA are the processor module, the network-aware memory controller, and the SerDes interface to the high-speed transceivers (Figure 4.2). The processor module and memory controller maintain a two-way communication protocol to initiate memory access requests, which requires the processor to access all memory data via the memory controller. This allows the memory controller to abstract away the implementation details of the optical network while the processor is oblivious to any physical layer changes, with the exception being that the processor is aware of a higher bandwidth memory link.

A multi-core architecture may require the memory controller to handle requests

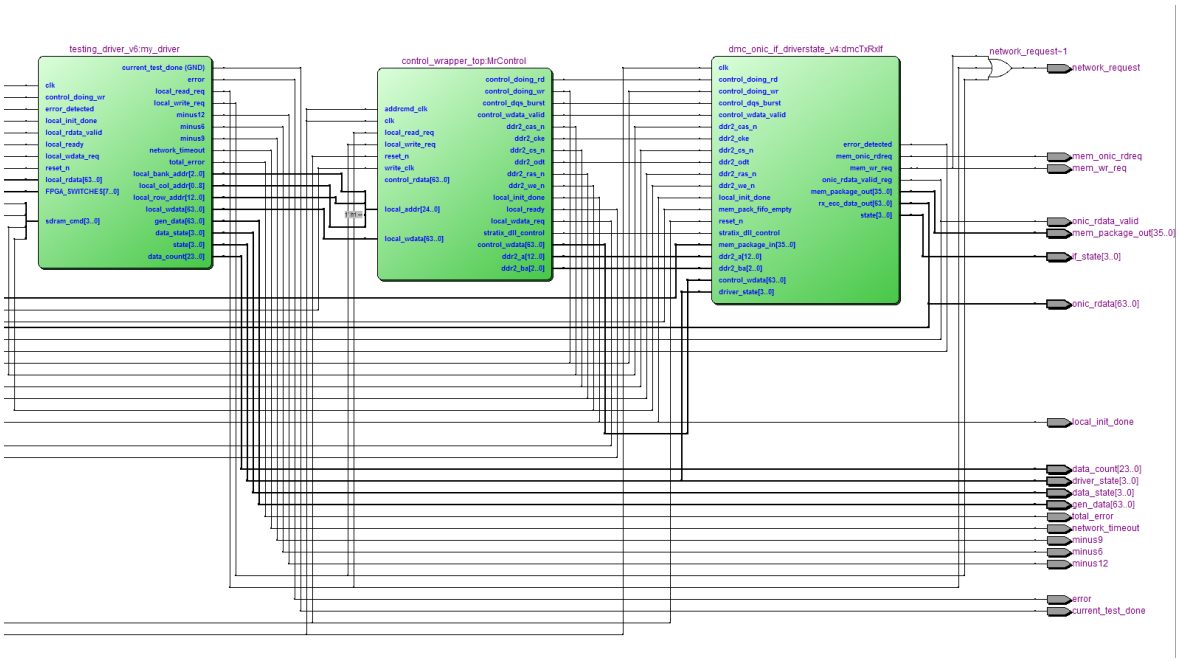

Figure 4.2: Graphical Representation of CPU Verilog Top-Level FPGA Modules - The primary modules implemented within the FPGA are the processor (testing_driver_v6:my_driver), network-aware memory controller (control_wrapper_top:MrControl) and SerDes transceiver interface (dmc_onic_if_driverstate_v4:dmcTxRxIf).

from multiple cores, and a high-performance memory system may connect the memory controller to many independent memory devices. The network-aware memory controller utilizes the network architecture to enable concurrent accesses from any processor node to any OCM node with minimal access latency.

4.1.1 Circuit-Switched Memory Controller

In order to guarantee a reliable, high-bandwidth memory link, the OCM system first uses a circuit-switched optical interconnection network. The memory controller (MC) is modified to reflect this change, as seen in Figure 4.3. The memory controller manages communication across the optical network analogously to the way a standard memory controller operates an electronic memory bus. All control of OCM devices is from the memory controller, but the key difference here is that the memory controller 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 memory controller to the appropriate memory node, and thus the memory controller 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 memory controller to SDRAM followed by read-data streaming from SDRAM to the memory controller. In this case, the memory controller

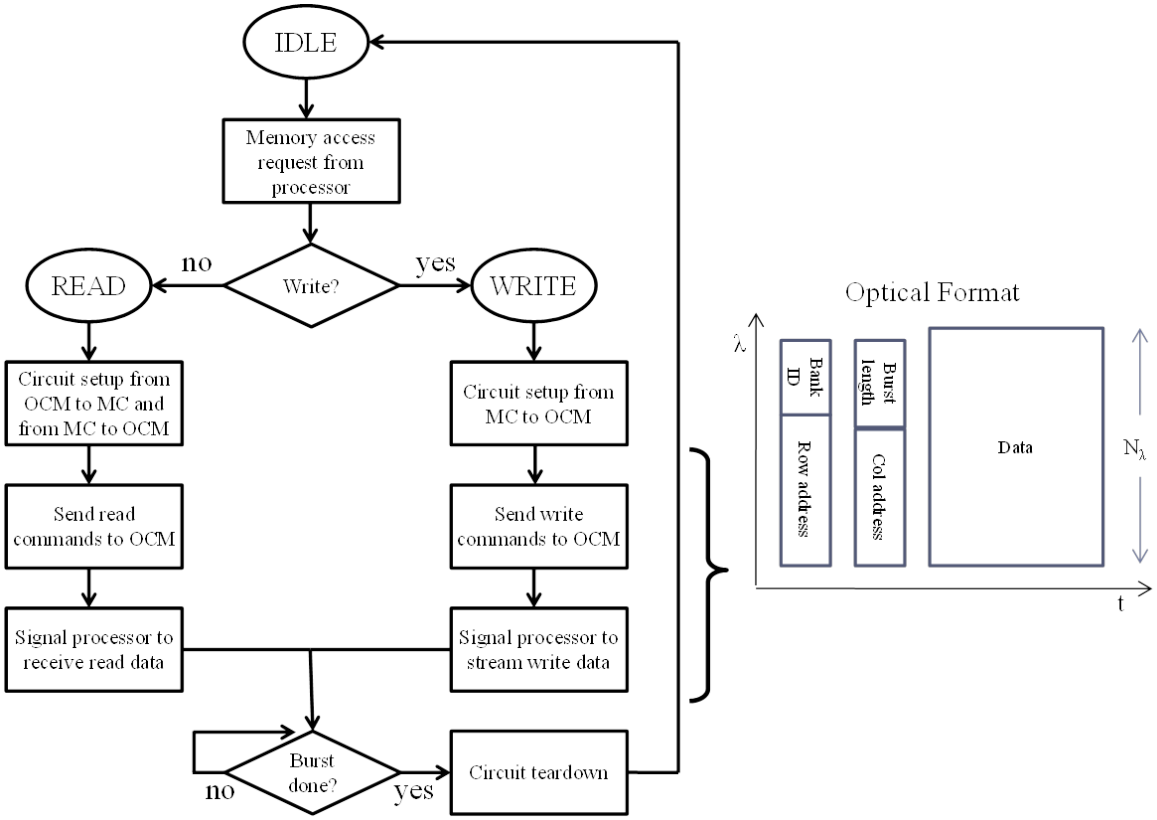

Figure 4.3: Circuit-Switched Memory Controller Flowchart - Flowchart of memory controller for circuit-switched OCM system.

will manage a lightpath from memory back to itself while the read operation is in progress. As in Figure 4.3, the first step is for the memory controller to create lightpaths from the memory controller to the OCM node and from the remote OCM node back to itself. With a communication link established, the memory controller 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 memory controller is used to stream read data to the processor. The remaining OCM-to-MC lightpath is torn down upon completion of the burst.

The optical memory link from the memory controller 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 the serialization of the memory link in the OCM system, dedicating one serial channel to the low-bandwidth control information would be inefficient. The standard SDRAM protocol [6] specifies that memory will only receive one set of commands for each write (or read) burst. This allows the MC 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 the done with current electrically-connected memory.

A typical SDRAM burst is 8 memory words, and therefore streaming a large amount of memory data requires frequent transmission of SDRAM commands from the memory

controller to memory. In data centers, for example, a majority of data traffic may be part of flows over 100 MB [27]. The circuit-switched memory controller therefore increases the memory burst size to a full SDRAM array row, 1024 words, to reflect the nature of the large-scale programming model and improve memory link utilization without overloading the network. Each 1024-length burst therefore writes or reads 64 kilobits of memory data.

The circuit-switched configuration is ideal only for a specific set of applications, such as streaming applications, which require predictable, long streams of memory accesses. A more flexible OCM system can 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.

4.2 Hybrid Packet- and Circuit-Switched OCM

Memory interconnect architectures are especially well-suited for the deployment of optical networks, owing to the performance and energy requirements of main memory systems, as well as the necessary flexibility within a network to support potentially diverse and unpredictable traffic patterns. For example, a packet-switched optical network can provide low memory access latency for short messages, while a circuit-switched optical network delivers greater performance for longer messages (e.g. in a streaming application) [56]. The ideal optical interconnect for microprocessor-to-main-memory communication is therefore a hybrid packet- and circuit-switched network. In this novel hybrid approach, memory accesses exceeding a predetermined size threshold

use circuit-switched lighpaths, while all other accesses are packet-switched. In this way, an OCM system can be constructed with greater performance and capacity, while achieving lower memory access latencies and reduced power consumptions.

Further, in the latest generation of DIMMs, the memory module's energy consumption is reduced by using 'sleep' states, during which the data buffers and transceivers at each node are powered down when not in use [36, 37, 81]. Expanding on this functionality, the OCMMs can enter an equivalent sleep state in which the SDRAM input and output buffers are disabled and the high-speed optical transceivers are disabled. However, the high latency associated with each SDRAM device entering or exiting its sleep state can add significant overhead to each memory access. If not managed efficiently, the additional latency can considerably reduce overall system performance. Therefore, it is necessary to capitalize on innovative interconnect technologies and architectures to redesign processor-to-memory communication.

This section presents an experimental demonstration of the first hybrid packet- and circuit-switched optically-connected memory system. This experiment implements the above-described FPGA-based microprocessor that communicates with three OCMMs across a wavelength-striped 4×4 hybrid packet- and circuit-switched optical network (as in Chapter 2). The processor and OCMMs create wavelength-striped memory messages using eight 10-Gb/s electronic transceivers; eight separate wavelength channels are modulated and then combined using WDM to generate 80-Gb/s messages. The resulting OCM system achieves 240-Gb/s aggregate memory bandwidths through the optical network (80 Gb/s per network port). This work expands on the purely circuit-switch memory controller [68] by adding "sleep state" functionality to the OCMM, as

well as modifying the network-aware memory controller to be capable of issuing both packet and circuit memory accesses. The system optimizes communication for each desired transaction size to support a diverse range of applications. The energy-efficient OCMMs can efficiently enter a low-power 'sleep' state in which the SDRAM and on-board transceivers consume minimal power, and rapidly re-enter normal operation when required.

4.2.1 Experimental Setup

The experimental demonstration characterizes the performance and efficiency of the proposed hybrid packet- and circuit-switched memory access protocol. This eliminates the power-hungry processor-memory electronic bus and leverages the unique functionalities of the optical interconnection network to offer energy-efficient OCMMs. Each OCMM is accessed all-optically and transparently across an implemented optical network using either packet or circuit switching, depending on the memory transaction message sizes. Using circuit switching for smaller memory transactions results in inefficient use of network resources and does not adequately amortize the circuit path setup latency [56]. Smaller messages utilize the previously described wavelength-striped packet switching.

The OCMMs can enter a low-power 'sleep' state in which the SDRAM input and output buffers are disabled and the high-speed optical transceivers are idle (not transmitting any data). In this state, the SDRAM consumes only 20% of its normal operating power [36, 37] and the transceiver logic consumes only minimal static power. Here, the memory controller transmits short optical packets to command an OCMM

to enter or exit its sleep state; the transition requires less than 10 ns [37].

The implemented OCM system uses multiple Altera Stratix IV FPGA-based circuit boards to create a processor and three OCMMs (Figure 4.4). All memory accesses are performed over the 4×4 optical network in the test-bed. Each OCMM consists of SDRAM connected to an FPGA, which contains serialization and deserialization (SerDes) functionality and a bank of 8×10-Gb/s bidirectional electrical transceivers. The processor and custom memory controller are implemented using an identical FPGA with 8×10-Gb/s transceivers. Each transceiver bank interfaces with discrete optical components to generate and receive 8×10-Gb/s WDM memory transactions. Each transceiver bank on the FPGA circuit board drives eight $LiNbO_3$ Mach-Zehnder modulators to modulate eight separate wavelengths (ITU C-band channels C36-C43), which are combined using WDM to create 8×10-Gb/s wavelength-striped memory transactions. The transceiver banks also connect to eight p-i-n receivers with transimpedance amplifiers (TIAs) and limiting amplifiers (LAs), which receive the demultiplexed WDM transactions.

For the packet-switched transactions, the FPGAs use three low-speed general purpose input/output (GPIO) pins to drive three SOAs and modulate the frame and address header wavelengths. The three low-speed header wavelengths and 8×10-Gb/s payload wavelengths are combined before being injected into the 4×4 optical network. For circuit-switched transactions, the three header wavelengths are not used, and only the 8×10-Gb/s WDM payloads are injected into the network. The resulting configuration is such that a processor accesses its main memory across a transparent, hybrid packet- and circuit-switched 80-Gb/s WDM optical memory channel. Memory

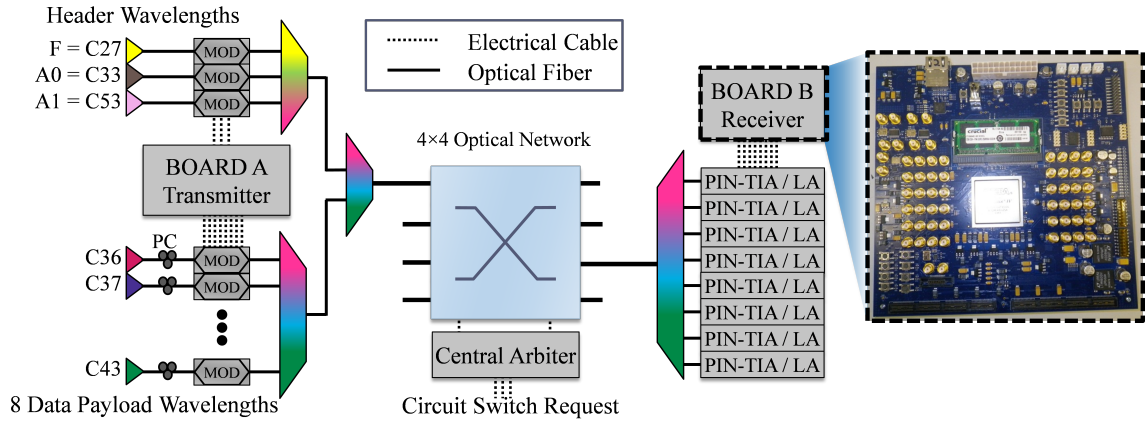

Figure 4.4: Hybrid Packet- and Circuit-Switched OCM: Experimental Setup - Experimental setup showing one direction of processor-memory communication, with a photograph of one FPGA-based circuit board (inset). Packet-switched communication uses the header wavelengths: frame (F), address 0 (A0), and address 1 (A1); circuit-switched communication utilizes the central arbiter.

communication and network utilization are optimized by selecting optical packet or circuit communication based on required memory transaction sizes as specified by the processor.

The custom memory controller analyzes incoming memory access requests from the processor to efficiently schedule all network communication requests and optimize network utilization. The processor can access the memory space at any of the connected OCMMs. Due to the principle of locality, the access is often to the same OCMM as one or many previous memory accesses. The processor may issue memory access requests faster than the memory controller can analyze them, and therefore the memory controller can buffer up to 16 memory access requests. The memory controller will attempt to reorder buffered memory accesses to issue multiple requests to the same OCMM in a single, longer transaction. When a sufficient number of accesses to an

OCMM are combined and exceed a predetermined threshold, the memory access will instead by performed using circuit switching. Any transactions below that threshold are issued as optical packets.

In current state-of-the-art memory modules, memory accesses in modern processors are of a standard size, called bursts, which are typically eight 64-bit memory words. Each burst incurs a memory access overhead latency on the order of tens of nanoseconds, which is due to the standardized SDRAM access protocol [6]. Before the memory controller can transmit data from the processor to memory, for example, it must first issue low-bandwidth SDRAM-specific commands that are required to operate the SDRAM's internal buffers and addressing hardware. This has resulted in a trend of burst sizes doubling with each new generation of SDRAM [36, 37], reaching the current burst size of 8 words, which amortizes the SDRAM access latency and maximizes memory bandwidth. Hence, to further improve bandwidth within the OCM system, the memory controller assumes a minimum burst size of 32 words for a minimum optical packet size of 2,048 bits.

In order to minimize SDRAM command overhead and optimize communication within the optical network, SDRAM command signals are transmitted as optical packets prior to the transmission of memory data. In the case of packet-switched data, the data packets will be transmitted immediately following the command packets and incur an average of 20 ns SDRAM access overhead per packet. However, for circuit switching, the SDRAM command packets are speculatively transmitted while a circuit path setup request is issued to the central arbiter. This hides a portion of the circuit path setup time (in this implementation, this is approximately 90 ns)

within the SDRAM command overhead to reduce the total latency for circuit-switched transactions. Based on the latency overhead of setting up the packets and circuits, this experiment designates a threshold of five aggregated bursts to a single OCMM as the cutoff between packet and circuit accesses.

The memory controller at the processor node also analyzes the queued memory requests to manage the OCMM sleep states. When no memory requests are pending for a given OCMM, the memory controller will issue an optical sleep command packet to the OCMM. When memory accesses are again desired at a sleeping OCMM, a 'wake up' command packet is issued prior to the SDRAM access command packet. The 'wake up' process requires up to 10 ns, and therefore the worst-case penalty for attempting to access a sleeping OCMM is an additional 10 ns of latency, in addition to the standard SDRAM access overhead. The network-aware memory controller avoids the worst-case penalty by leveraging the 16-deep memory access request queue, which enables the memory controller to predict upcoming memory accesses and 'wake up' appropriate OCMMs just-in-time for each memory access.

Additionally, for circuit-switched accesses, the extra incurred latency (i.e. 10 ns) can also be hidden within the circuit path setup time by the speculative SDRAM commands. As discussed above, the use of sleep states in contemporary systems results in additional latency. Here, a novel aspect of this scheme is that the latency overhead associated with entering/exiting a sleep state can be hidden in the optical circuit setup time. Thus, the system can attain the corresponding power savings without the latency penalty.

To characterize the hybrid optical packet- and circuit-switched OCM system, the

FPGA-based microprocessor is programmed to fill the entire memory address space with predictable bit patterns: a $2^{31} - 1$ pseudo-random bit sequence (PRBS), all 0's, all 1's, and the bit pattern corresponding to the destination memory address. These bit patterns are chosen from both established memory tests and optical system tests. After the memory address space is full, the processor issues 'read from memory' requests to stream all previously stored data back from memory. As the data streams in from memory, a counter within the processor verifies the data test patterns, records the number of correctly verified bits, and calculates the number of bit errors. These counters are used to generate an effective memory-bit-error rate (EMBER) to quantify the functionality and reliability of the hybrid packet- and circuit-switched OCM system. In this way, error-free operation is achieved when the processor correctly verifies over one terabit of memory data from each OCMM, attaining EMBERs less than 10^{-12}.

The order in which the processor accesses each OCMM is random, on an access-by-access basis, such that the memory controller may receive any number of memory access requests for a given OCMM at one time. This enables the memory controller to reorder memory access requests when possible and thus generates both packet- and circuit-switched memory transactions. Randomness is obtained through the use of two linear feedback shift registers (LFSR), one 7 bits and one 8 bits, that are sampled once for each memory access. The serial outputs of the LFSRs are appended together into a 2-bit value that represents which of the three OCMMs will be addressed (00, 01, or 10). The value 11, being an invalid address, causes the processor to address the same OCMM as the previous memory access. This increases the probability that subsequent memory accesses are to the same OCMM, as would be the case for data locality within

an application. Using LFSRs is accepted in computing as an acceptable approximation of randomness.

4.2.2 Experimental Results

Error-free operation of the OCM system is confirmed with EMBERs less than 10^{-12} for all three OCMMs. Figure 4.5 shows the optical eye diagrams for the eight 10-Gb/s memory payload channels, depicting clear open eyes for all data payloads. Figure 4.6 shows an example of how the processor can communicate with the OCMMs using optical packets and circuits.

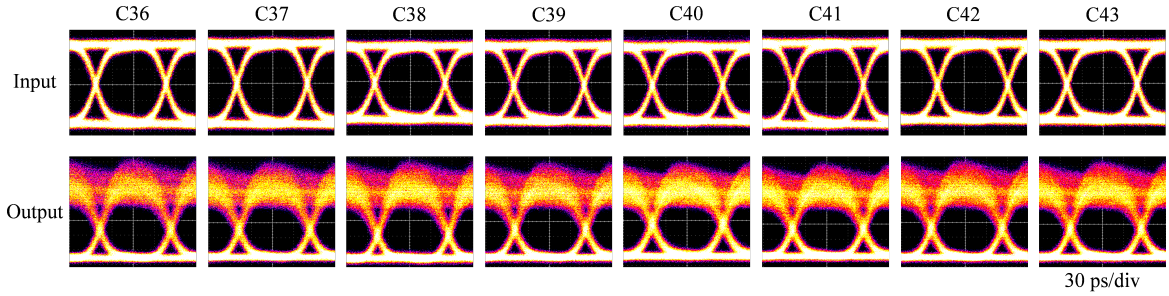

Figure 4.5: Hybrid Packet- and Circuit-Switched OCM: Optical Eye Diagrams - Optical eye diagrams for the 8×10-Gb/s memory payload wavelength channels at one network input port (top) and at one network output (bottom).

The circuit-switching data threshold of 5 bursts (i.e. 10,240 bits) within the 16-deep buffer results in an average of 24% of memory accesses using circuits rather than packets (evaluated for each terabit of memory traffic). Previous OCM studies [56, 57, 68] assume steaming applications with predictable memory access patterns, which are ideal for a purely circuit-switched optical network. Such implementations would penalize the remaining 76% of memory traffic that is comprised of smaller-sized messages. Each

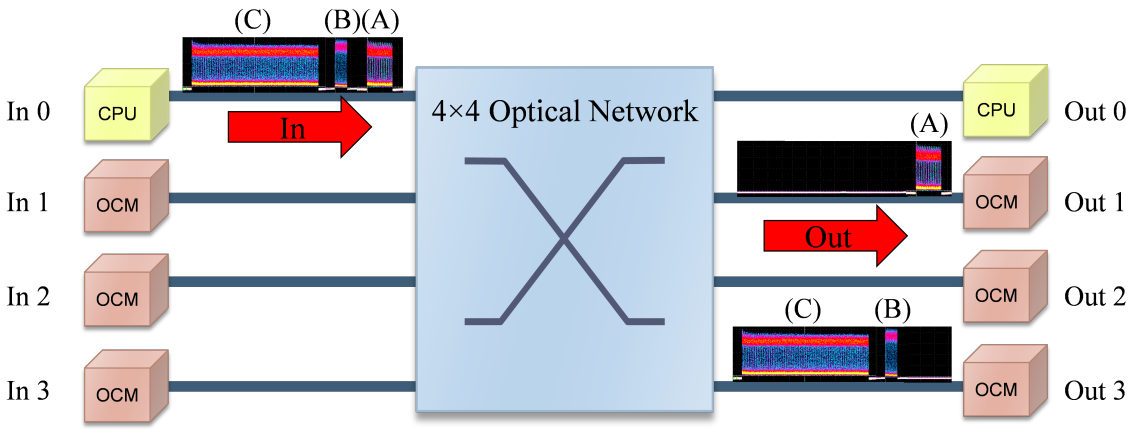

Figure 4.6: Example Packet- and Circuit-Switched Memory Communication - The processor injects an optical packet into the network (A) addressed to the OCM node at network port 1. Subsequently, the processor issues a short control packet (B), followed by a circuit-switched memory access (C) to the OCM node at network port 3.

packet-switched memory access within that 76% reduces latency by up to 70 ns in this implementation. Meanwhile, each memory access within the remaining 24% that use circuit switching reduces latency by >10 ns by amortizing setup latency over several aggregated bursts. Overall, the combined use of packet and circuit switching supports diverse applications and memory access patterns with both short and long data streams.

The use of sleep states allows for a significant reduction in power consumption of the OCMMs. Each OCMM was determined to be set in sleep state for 35% of the experimental characterization, resulting in an overall memory power savings of 28%. This is due to the fact that although each OCMM is addressed 33% of the time, and therefore unused and potentially able to sleep 66% of the time, the random order of memory accesses prevents the memory controller from putting OCMMs to sleep for their theoretical maximum allowable time. It is possible to increase the percentage

of time each OCMM spends in its sleep state, e.g. by allowing the state to persist not only until a request enters the access queue but until just before the access is actually issued. However, this change would trade-off critical memory access latency for improved energy savings within the OCMMs, and would be a design choice specific to each individual system. Deploying the OCMMs within large-scale systems, utilizing terabytes of memory per server and many thousands of servers, could thus achieve the extreme levels of energy efficiency required for next-generation data centers and HPCs.

4.3 Memory Multicasting

Future optical networks will be required to seamlessly support the multicasting of broadband optical messages comprising of memory information. This section describes the expansion of the optical-network-aware memory controller functionality 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 (Figure 4.7). 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.

In the following experiment demonstration a processing node communicates simultaneously to multiple memory nodes by multicasting high-bandwidth, WDM optical messages on a 5-stage 4×4 multicast-capable optical interconnection network test-bed (Figure 4.8). The 5-stage optical network consists of ten 2×2 SOA-based

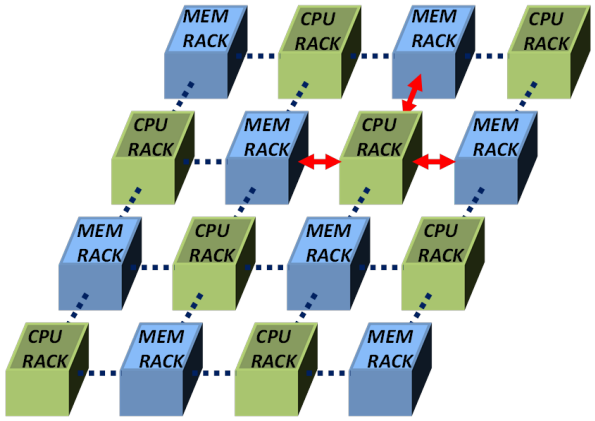

Figure 4.7: Illustration of Memory Multicasting - CPU racks (green) are connected to memory racks (blue) by optical links (dotted lines); red arrows show a CPU multicasting to three memory racks).

switches as described in Chapter 2. A high-speed FPGA is used to realize the processing core and Micron DDR3 SDRAM devices act as the remote memory nodes.

4.3.1 Multicasting Memory Access Protocol

Multicasting of memory accesses leverages the protocol in which 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. 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. 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

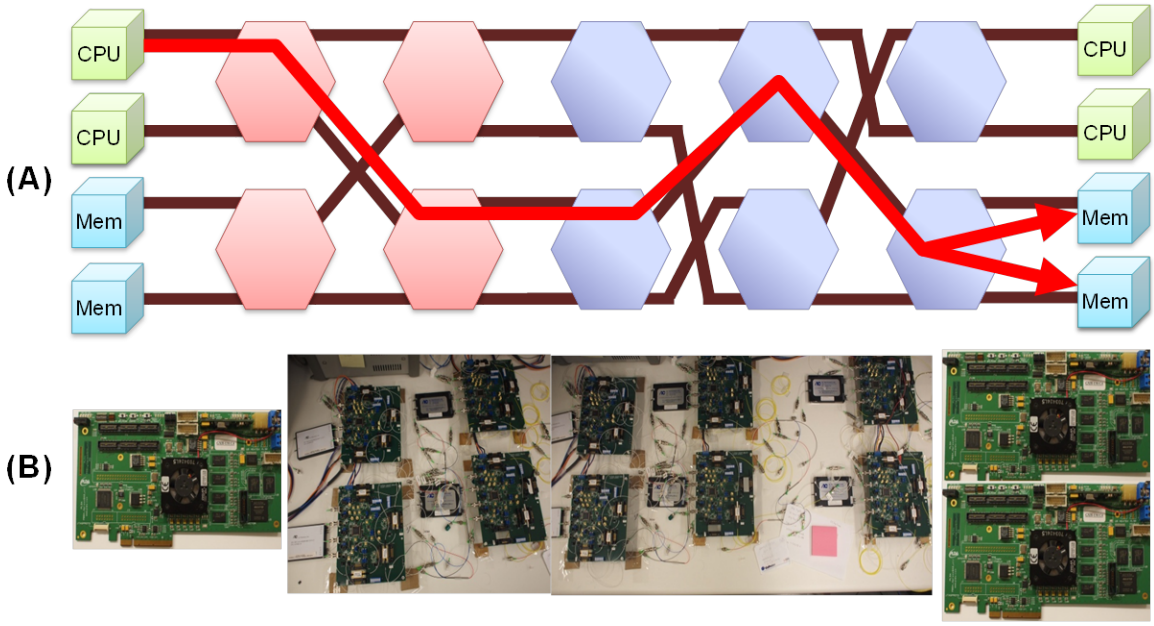

Figure 4.8: Block Diagram and Photograph of Memory Multicasting - (A) Block diagram of setup showing a CPU (green) multicasting to two memory nodes (blue) across a 4×4 multicast-capable test-bed. (B) Photograph of implemented FPGA boards and optical network test-bed.

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.

4.3.2 Experimental Setup and Results

The experimental setup (Figure 4.9) consists of three circuit boards communicating 4×10-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 $LiNbO_3$ 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 FPGA, which is connected to four chips of Micron DDR3 SDRAM and 4×10-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 TIAs and LAs. One circuit board's FPGA is programmed to act as a microprocessor with an on-

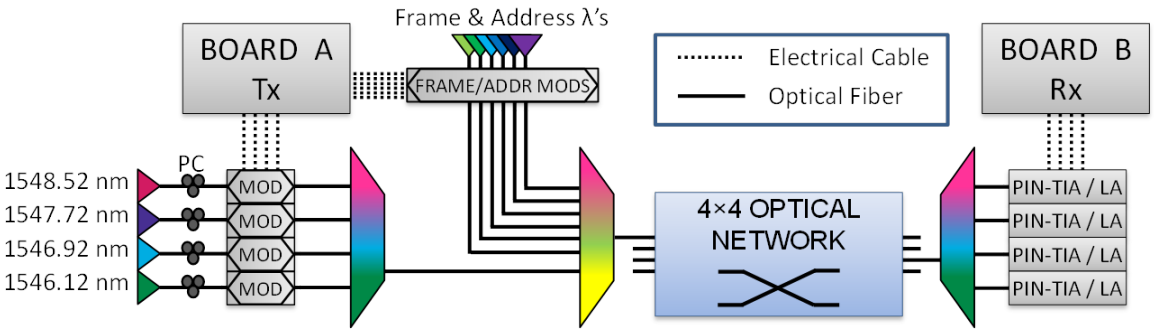

Figure 4.9: Memory Multicasting: Experimental Setup - Experimental setup showing Board A communicating to Board B over a 4x4 optical network test-bed. Board A modulates 4 payload channels (lower left), as well as a frame and 5 address bits (upper center) for network routing. Board B receives the payload from the optical network using four PIN-TIA receivers.

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 DDR3 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 40-Gb/s memory bandwidth; as necessary, the processor node will multicast memory transactions to both SDRAM nodes for an aggregate 80-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, the 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 EMBER is confirmed as less than 10^{-12} once over a terabit of data has been verified by the system. Thus, this demonstrates the correct functionality and stability of the multicast-capable optically-connected memory system. Optical eye diagrams for the four 10-Gb/s payload channels corresponding to processor write-to-memory transactions are shown in Figure 4.10; these eyes were collected using self-triggering.

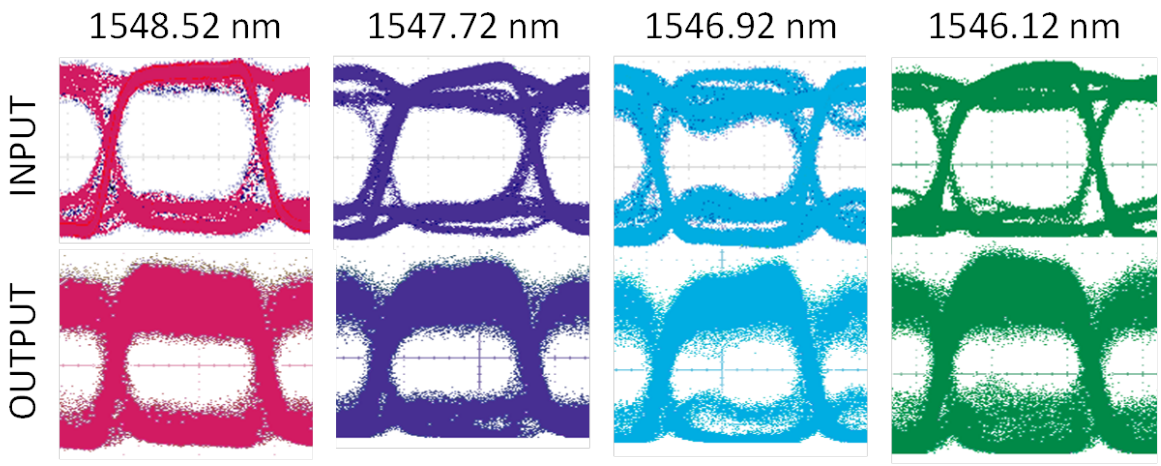

Figure 4.10: Memory Multicasting: Optical Eye Diagrams - Optical eye diagrams for the experimental demonstration, showing the four memory data payload wavelengths at the input of the network test-bed (top), and the same four channels after being multicasted to one network output port (bottom).

4.4 Discussion

This goal of the work presented in this chapter is to move beyond traditional memory controllers implemented in electrically-connected memory systems and enable processors to communicate all-optically, and transparently with main memory. The network-aware memory controller presented here abstracts away the optical network, which allows commercial processors to leverage unprecedented levels of memory bandwidth and novel memory architectures without changing the underlying processor architectures that have been developed over the last several decades.

The first experiment presented in this chapter demonstrates the first hybrid packet- and circuit-switched optically-connected memory system, with error-free (EMBERs $<10^{-12}$) transparent routing of 8×10-Gb/s wavelength-striped memory transactions between processors and OCM nodes. This implementation efficiently optimizes memory communication based on memory data burst lengths and reduces memory access latency by up to 70 ns per memory transaction compared to purely circuit-switched OCM architectures. Additionally, the results show a 28% reduction in the memory node's energy consumption through an optical-packet-controlled OCMM sleep state technique. The second experiment demonstrates 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. This configuration implements a microprocessor with a multicast-capable memory controller that multicasts 4×10-Gb/s wavelength-striped memory transactions using an implemented optical network test-bed to two independent, optically-connected memory modules.

The sum of the work in Chapter 3 and Chapter 4 shows that the integration of optics with main memory architectures can enable memory bandwidths and memory access functionalities that are not possible using electronic memory interconnects. The custom memory controller presented here allows any processor to interface with OCMMs, which have been created and optimized based on rigorous experimental characterizations.

This work demonstrates the need for low-latency, high-performance optical interconnects within future large-scale memory systems. These OCM systems must also be flexible, using hybrid packet- and circuit-switched network architectures and optical multicasting, to enable innovative system architectures for future high-performance computing systems with improved bandwidth, latency, and energy efficiency.

Chapter 5

Resilient OCM architectures

In this chapter, the challenge of resilient memory architectures is addressed by expanding on existing error detection and correction protocols and developing memory access protocols that leverage the optical network architectures described in Chapter 2. Resilience is becoming an increasingly critical performance requirement for future large-scale computing systems. In data center and high-performance computing systems with many thousands of nodes, errors in main memory can be a significant source of failures. As a result, large-scale memory systems must employ advanced error detection and correction techniques to mitigate failures. Here, a resilient OCM architecture is presented and experimentally characterized that can not only continue to operate error-free through the failure of an entire OCMM, but also dynamically redirect memory traffic away from failing memory devices.

5.1 Background

Current scaling trends illustrate that next-generation large-scale computing systems will require many thousands of servers to meet the demands of future applications. In these data center and high-performance computing systems, the performance requirements will undoubtedly strain the limits of main memory with respect to resilience. Main memory is typically composed of several SDRAM chips, which are packaged together onto DIMMs. Current SDRAM technology is designed to optimize density, rather than speed or resilience. In order to achieve the terabytes-per-second memory bandwidths required by future computing applications [1], each system will require a significantly greater number of DIMMs as compared to present systems. As with using any commodity hardware, increasing the number of DIMMs clearly increases the probability of failures within the overall memory system [63, 64].

A failure in memory occurs when any number of bits in memory is retrieved (i.e. read from memory) with a value other than what was stored (i.e. written to memory): this is also designated as a bit error. In a given system, bit errors can cause expensive loss of data or system downtime, which can lead to significant long-term consequences [94, 95]. It was recently reported that SDRAM errors are significantly more common than previously believed: 8% of DIMMs within a modern data center experience errors each year [96]. Further, a more general resilience study, [97], concluded that 8% of servers are in need of servicing each year, resulting in an estimated annual servicing cost of $2,500,000 for a 100,000 server data center. Thus, in order to reduce the operational cost of future data centers, it is clear that future memory systems will require resilience techniques that can identify DIMMs that are more likely

to experience bit errors and can dynamically redirect memory traffic away from failing DIMMs.

Current approaches to achieve resilience in contemporary memory systems leverage the use of error-correction codes (ECC). The most common type of ECC can correct for a single bit error within each memory word, i.e. most servers use an extended Hamming code that is single-error correcting and double-error detecting (SECDED) [98]. For large-scale systems, more robust error protection can be provided by advanced ECC interleaving techniques. However, main memory systems are already unable to keep pace with modern microprocessors [1], a trend that is unlikely to change with current technology [3]. Thus, the obstacles that currently limit SDRAM performance will also constrain efforts to improve resilience in future systems. It is evident that next-generation large-scale computers will not achieve sufficient resilience performance using traditional ECC methods.

Optically-connected memory has been proposed as a solution to scaling memory systems within large-scale computers [68]. The high-bandwidth density of optical interconnects can alleviate pin-count constraints by allowing a single optical fiber to deliver terabits-per-second of memory bandwidth using WDM. Furthermore, the envisioned integration of silicon photonics with processors and memory components will greatly improve the overall bandwidth, latency, and energy efficiency performance [47]. The distance immunity offered by optical networks at computer scales allows these performance benefits to be applied to a greater number of memory devices than is possible using electronic interconnects. This will enable memory systems with greater memory capacity, as well as more flexibility in implementing advanced ECC

functionalities. OCM can replace the point-to-point links between a processor and its memory modules with a multicast-capable optical interconnection network (Figure 5.1), exploiting the distance immunity of optics to allow each processor to address a vast number of OCM nodes [72]. High-bandwidth communication links with multiple OCM devices can be established in parallel to allow processors to attain unprecedented bandwidths without sacrificing sensitive memory access latency. An OCM-enabled approach can protect against the failure of an entire OCM node, as opposed to only a single SDRAM chip in the case of an advanced ECC scheme using electronic interconnects. Optical multicasting further reduces latency and eliminates the need for power-hungry buffering and reordering at the processor, thus creating overall higher performance and a more resilient memory system [99].

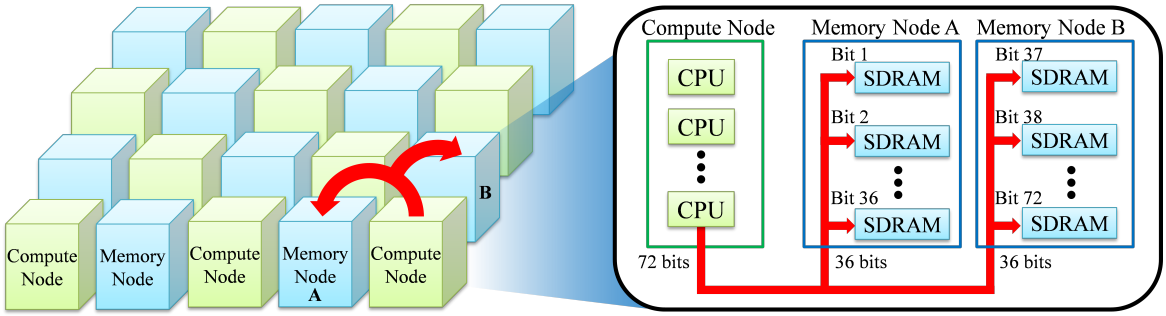

Figure 5.1: Illustration of OCM and Compute Nodes Utilizing Proposed Advanced OCM ECC - Compute nodes and memory nodes are connected using an optical interconnection network. Red arrows show a compute node multicasting to interleave data over two memory nodes. At each memory node, the data is further interleaved so each SDRAM stores a different portion of the data.

5.2 Overview of Error Correction

The resilience of main memory is critical to system stability and hence has been extensively investigated [64, 96, 100, 101, 102, 103, 104]. It is an important performance metric, since failures in main memory can lead to significant bit errors. Such failures can be soft errors, caused by transient effects, or hard errors, caused by permanent hardware defects. Hard errors are more problematic, due to their need for hardware servicing and the potential to cause repeated failures. Both soft and hard errors can lead to either correctable errors (CEs) or uncorrectable errors (UEs); only CEs can be hidden from the processor in real-time by error-correction hardware (if present). Studies show a strong correlation between the two types of errors within a server or DIMM; for example, 80% of all DIMMs with UEs had experienced a CE within the previous month [96].

Improving resilience in main memory is based on redundancy, primarily through the simple duplication of data or through the computation of parity-check bits. The pure duplication of data, such as majority voting systems [105], are cost prohibitive for most main memory applications due to the need for at least three times as much hardware. The preferred solution is to use ECC, such that memory systems can correct a small number of bit errors (typically one error) with the addition of the minimal number of party-check bits.

Traditionally, each basic unit of memory data is comprised of a 64-bit memory word. Each memory word is then protected by an ECC by appending extra parity-check bits, and is subsequently stored in ECC memory. ECC memory differs from normal memory in that it uses extra SDRAM chips packaged together to store the

redundant bits needed for the ECC.

The SECDED Hamming code is the most common error correction protocol for SDRAM [98]. In its basic implementation (Figure 5.2), the SECDED code allows each 64-bit memory word to be protected by eight parity-check bits, creating 72-bit code words known as a (72,64) code. The eight check bits are computed by performing an XOR operation on certain data bits. Only a subset of the 272 possible 72-bit words comprises valid code words, while the remaining words are invalid. The (72,64) code has a minimum Hamming distance of 4, i.e. 4 of the 72 bits must be flipped in order for errors to change one valid code word into another valid code word. It is therefore possible to detect double bit errors and correct single bit errors. The 72-bit ECC protected data is stored in main memory where a bit error may occur.

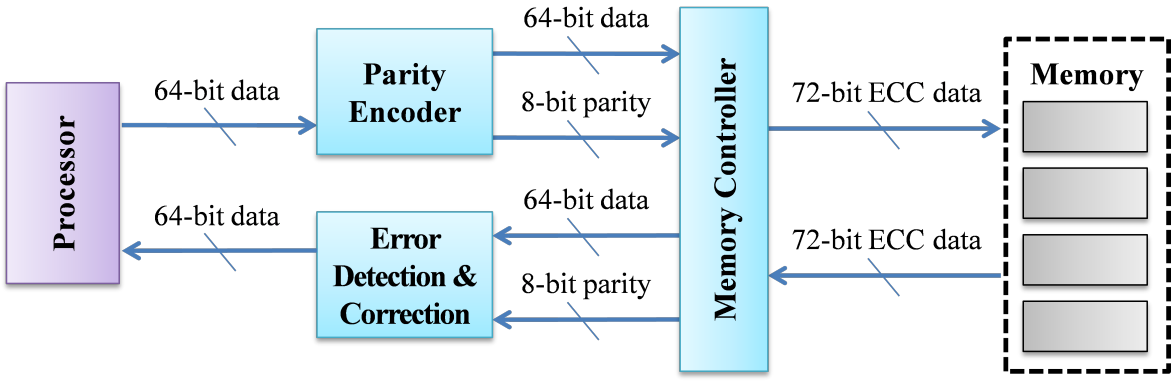

Figure 5.2: Architectural Diagram of Proposed Advanced OCM ECC - Compute nodes and memory nodes are connected using an optical interconnection network. Red arrows show a compute node multicasting to interleave data over two memory nodes. At each memory node, the data is further interleaved so each SDRAM stores a different portion of the data.

A basic SECDED code can adequately mitigate failures to meet the needs of many

small-scale systems. For large-scale computers, greater error protection is provided by advanced ECC interleaving schemes, which distribute the 72-bit memory data over several SDRAM chips on the ECC DIMMs. In this way, the probability of having multiple bit errors within a single chip (and thus corrupting the full memory word) is decreased [106].

Often, SDRAM devices experience only single data bit errors; however, for a small fraction of cases, more than one data bit is erred. These multi-bit errors can affect up to the entire data width of the device, and, for more serious hardware failures, the data for all the SDRAM chips on a DIMM can be corrupted. Since standard ECC memory can only correct a single bit error, multi-bit memory errors can cause an expensive loss of data and/or complete system shutdown. A greater level of protection is therefore required for large-scale systems, as they contain a larger number of devices and therefore more possible points of failure.

Recently, advanced ECC techniques seeking to improve resilience in memory while simultaneously minimizing additional hardware (as compared to traditional ECC) have been shown. The primary developments have been Chipkill [106], SDDC [107], and Chipspare [108]. These three implementations are adequately similar to be discussed as one advanced ECC method. These advanced ECC methods build on standard ECC and add interleaving to protect each memory word from burst errors. Interleaving exploits the distributed structure of memory (i.e. where a single memory module actually contains many SDRAM chips). For instance, in standard ECC using a 72-bit ECC word, the DIMM could contain 18 SDRAM chips, with each SDRAM storing four bits of the 72-bit word. In advanced ECC approaches, the same 72-bit ECC

word would be instead distributed across four separate DIMM modules, each of which contain 18 4-bit-wide SDRAM chips. In this case, each DIMM module receives 18 of the 72 bits, and each of the 18 SDRAM chips receives one bit of the original 72-bit ECC word. The resulting configuration protects the data at each SDRAM chip from multi-bit errors, since each chip only has one of the 72 bits, which can be erred. When the data is subsequently read back from memory, all four DIMM modules are accessed simultaneously, allowing the original 72-bit memory word to be reassembled. Then, standard ECC detection/correction methods are applied.

The resilience improvement provided by advanced ECC is crucial for existing large-scale computers. However, for next-generation systems with many more thousands of compute and memory nodes, these advanced ECC techniques will be as inadequate as standard ECC is within today's large-scale systems. Advanced ECC can protect against the failure of a single SDRAM chip, but not against the failure of an entire DIMM. The correlation between UEs and CEs within a DIMM requires next-generation systems to tolerate and correct for the failure of an entire DIMM [96]. However, the limitations of electronic wiring [ITRS11, Ho01] prevent traditional, electronically-interconnected main memory from scaling to the sizes required for interleaving across 72 separate DIMMs. The advanced ECC protocols also require the processor to issue many individual memory transactions to the many targeted DIMMs, which are meant to be performed serially on a wide electronic bus. For multiple DIMMs, this is clearly a latency-intensive process, with each memory transaction incurring tens of nanoseconds of latency [6].

The OCM system proposed here replaces the electronic memory bus with a

multicast-enabled optical interconnect. Using optical multicasting, a processor can issue a single memory transaction to simultaneously read or write data to all (or a subset of all) OCM nodes in parallel. As an example, the 72-bit memory data resulting from the SECDED code can be stored in 72 discrete OCM nodes in a single 'write to memory' command. By configuring the 72 OCM nodes on an optical network to store a different bit from a single stream of multicasted data, the OCM system can then guarantee that no memory device contains more than one bit of each 72-bit ECC word.

Additionally, the optical network can scale to accommodate hundreds of thousands of nodes [66]. It is thus possible to allow each processor to efficiently access all 72 OCM nodes in the above example (or even more nodes if using a different Hamming code). Interleaving the memory data across 72 possible nodes is then sufficient to correct for the failure of an entire OCM node.

In addition, dynamic bit-steering is an accepted technique to further increase the resilience of memory by dynamically reassigning the addressed memory location if a predetermined error threshold is reached [106]. By using an optical interconnection network and a wavelength-striped approach with wavelength-routed streams, this OCM scheme can effectively combined with dynamic bit-steering. Upon the detection of a failing OCM node (i.e. an increase in bit errors that exceeds an error threshold for a specific memory module), the memory transactions' routing information through the optical network can be simply and rapidly changed when a multicast operation is issued to avoid this failing OCM device. By routing around this failure, efficient dynamic bit-steering can be offered, and the problematic OCM node can be quickly replaced in a

hot-swappable fashion without powering down the system.

5.3 Experimental Characterization

The resilient OCM architecture is implementing using the 4×4 photonic switching node detailed in Chapter 2 configured in an Omega topology. Omega networks are common in shared memory systems, as they provide greater memory access efficiency and path diversity than a simple bus [109].

5.3.1 Experimental Setup

In the context of future large-scale memory systems, this section details an experiment to characterize a novel advanced ECC protocol for OCM and analyze its performance. This implementation enables a microprocessor to utilize wavelength-striped multicasting within an optical interconnection network [51] to create a more resilient memory system than is possible using today's electronically-interconnected memory. The optical network also enables fault tolerance through efficient dynamic bit-steering, i.e. when the number of errors at an OCM node exceeds a specified error threshold, the memory data is dynamically redirected to another node, thus allowing hot-swapping of the defective memory module.

The resilient OCM system is implemented using an FPGA-based processor, four OCM nodes, and the 2-stage 4×4 optical network implementation described above (Figure 5.3). The OCM nodes consist of DDR3 SDRAM connected to an Altera Stratix IV FPGA. The FPGA contains 8×10-Gb/s transceivers to provide SerDes and optical interface functionalities for the SDRAM. The FPGA-based processor is implemented

in a second identical Stratix IV FPGA, which creates a processor that issues 'read' and 'write' commands to the OCM nodes. The processor uses the board's 8×10-Gb/s transceivers to establish optical memory communication lightpaths. Each transceiver bank on the FPGA circuit boards is connected to eight $LiNbO_3$ Mach-Zehnder modulators to modulate eight separate wavelengths (ITU C-band channels C36-C43) and create 8×10-Gb/s WDM memory transaction streams. Additionally, the processor uses five low-speed general purpose input/output (GPIO) pins to create the frame, address, and multicast header wavelength signals. Each GPIO pin drives a separate SOA to modulate a discrete network header wavelength. The five header wavelengths and 8×10-Gb/s payload wavelengths are combined using a passive combiner with WDM before being injected into the 4×4 optical network. The network operates as described in Section III, and routes each WDM memory transaction to one or multiple output ports based on the header wavelengths. The wavelengths are demultiplexed at the output of the network, where each payload wavelength is received using a separate p-i-n receiver with transimpedance amplifiers (TIAs) and limiting amplifiers (LAs). The resulting configuration is such that a processor can access its main memory remotely, across a transparent, multicast-capable 80-Gb/s WDM optical memory channel. With only a single transaction, the processor can interleave memory data across multiple OCM nodes for increased resilience. By simply changing the network addresses, the memory traffic can be dynamically redirected (steered) away from a failing OCM node.

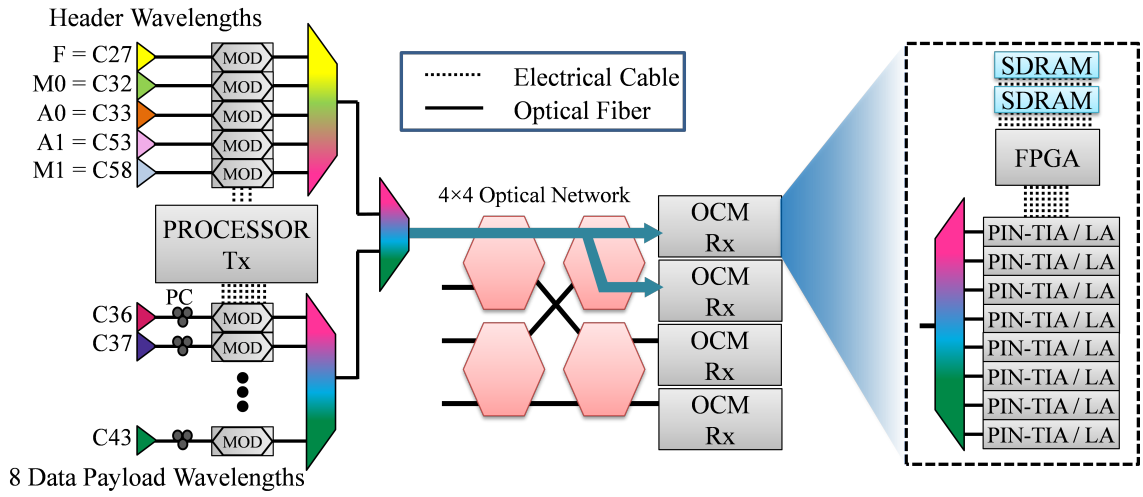

Figure 5.3: Resilient OCM: Experimental Setup - Experimental setup illustrating the communication path from the processor to two OCM nodes via wavelength-striped multicasting. The processor modulates five message-rate header wavelengths, in addition to eight payload wavelengths at 10 Gb/s, which are combined before being injected into the 4×4 optical network. Header wavelengths consist of: frame (F), address bits 0 and 1 (A0, A1), and multicast bits 0 and 1 (M0, M1). Inset shows the contents of an OCM node, including receiver circuitry, FPGA (for SerDes), and SDRAM. The return path from OCM to processor is identical.

5.3.2 Advanced Error Correction for OCM

Here, the advanced ECC protocol for OCM first leverages the (72,64) Hamming code to enable SECDED ECC on any 72-bit block of data. The processor then issues a multicast to a subset of OCM nodes, thus accessing them with the latency of only a single memory access, and interleaves data across the desired number of physically separate OCM nodes. The resulting functionality is similar to that of advanced ECC for traditional memory systems, in which data is interleaved across multiple DIMMs to protect against multi-bit errors; however, the implementation here and the resulting performance are drastically different. This work not only leverages the energy and bandwidth benefits of optics but also the distance immunity and optical multicasting capabilities to improve access latency and system resilience as compared to electronically-connected memory.

With respect to the advanced ECC protocol, this experimental demonstration characterizes the OCM system's resilience performance in three separate cases: for interleaving across two, three, or four OCM nodes, respectively. As a baseline comparison, the results are then compared to an OCM system with standard ECC, which uses SECDED ECC but without interleaving. With no interleaving, each OCM node contains the entirety of each 72-bit memory word, thus rendering it extremely vulnerable to any multi-bit errors within the node.

Each case of interleaving can correct for at least two errors at an OCM node. This worst-case scenario occurs when the OCM node experiences only three errors, but the distribution within the node is such that two of the erred bits are within the subset of 72 bits stored at that node. This results in a multi-bit error in the reassembled 72-bit

memory word read by the processor. The maximum number of correctable errors in each interleaving case is equal to the number of interleaved nodes: two, three, or four errors for the two, three, or four OCM nodes respectively. For example, in the case of interleaving to four OCM nodes, each node contains 18 bits of a 72-bit memory word and can tolerate between two and four errors per node, depending on the error locations. Unlike in an electronically-connected memory system, which is physically limited to four-way interleaving, this OCM system can easily scale to accommodate 72 OCM nodes per processor. Utilizing 72 nodes will not only remove the worst-case limit of two errors but also correct for the failure of an entire OCM node (72 bit errors at a single location).

With the endeavor of achieving dynamic bit-steering, the FPGA-based processor maintains a record of errors for each OCM node to track potentially failing memory devices. By tracking the number and locations of errors, this OCM system can attempt to identify potentially problematic memory devices before they fail completely. For example, using the observed strong correlation of errors within a DIMM [96], any OCM node that experiences an error is statistically more likely to experience another error in the near future. Additionally, such errors are more likely to lead to multi-bit errors that are still correctable by the advanced ECC but remain undesirable. It is therefore necessary to experimentally measure a worst-case error threshold such that the OCM system can allow at any OCM node without substantial risk to system stability. Upon reaching the error threshold, the processor will leverage the optical network's wavelength routing scheme to redirect its memory data away from the failing OCM node and assert a warning signal to indicate which OCM node needs replacing.

Finally, because each OCM node occupies its own independent port on the optical network, a failed OCM node may then be hot-swapped without interrupting normal memory operations.

5.3.3 System Performance

The characterization of the ECC-enabled OCM system involves having a processor repeatedly writing to and reading from memory while a controlled BER is induced within an OCM node. For the baseline control case, the processor accesses each OCM node individually, storing each 72-bit memory word within that node; the system will therefore experience an UE for any multi-bit error.

For advanced ECC interleaving, the processor is programmed to multicast and thus simultaneously stream data to two, three, or four OCM nodes, respectively. For all three experimental cases, the stored data consists of predictable bit patterns: a $2^{31} - 1$ pseudo-random bit sequence (PRBS), all 0's, all 1's, and the bit pattern corresponding to the destination memory address (these bit patterns are chosen from both established memory tests and optical system tests). 72-bit memory words are thus created by appending eight parity-check bits, as calculated by the SECDED Hamming code, to the 64-bit data patterns. The processor is programmed to steam data to memory, continuously performing write operations, until all the memory space is full. The processor then issues repeated requests to read all previously written data back from memory, reassembling the original 72-bit words from the interleaving process, correcting errors using the SECDED ECC, and verifying the post-ECC bit patterns. Pre-ECC and post-ECC errors are documented separately, and a counter

within the processor records the number of correctly verified post-ECC bits read back from memory. This data is used to calculate a pre-ECC BER and a separate post-ECC BER. Error-free operation is defined as achieving BERs less than 10^{-12}, or less than one error for each terabit of memory data.

The FPGA at each OCM node contains circuitry to experimentally induce a controlled BER within the stored memory data, which results in the pre-ECC BER measured by the processor. The errors are generated by randomly flipping a number of data bits with a probability specified by the desired BER, as the data is deserialized at the OCM node and written to SDRAM. Multiple linear feedback shift registers (LFSRs), which generate pseudorandom values, are used to obtain random numbers for the desired BER. A 40-bit LFSR determines if any and how many errors occur in each 72-bit memory word. The bit locations are determined by a series of ten 7-bit LFSRs, of which a subset will be read based on the number of desired errors, and the values of these LFSRs specify the bit locations of the errors. If a selected 7-bit LFSR contains an invalid value (outside the range of 72-bits), then another LFSR is used in its place. The number of 7-bit LFSRs was selected based on the range of induced BERs to be characterized, and therefore sufficient random values will always be available to select bit error locations. Additionally, because the processor FPGA and OCM FPGAs operate independently from each other and are on separate clock domains, each node's LFSRs have different seed values. The LFSRs in each OCM node are only sampled upon receiving a 'write to' memory instruction from the processor. The pseudorandom values generated throughout the OCM system are not correlated.

In this way, BERs are induced in stored memory data at rates up to 10^{-1}. The

wavelength-striped multicasting and interleaving of memory data serve to reduce the probability that even high BERs will cause multi-bit errors within any single 72-bit memory word. At the processor end, the data is disinterleaved as it streams in from the optical network and the SECDED ECC hardware corrects any single bit errors per 72-bit word. The ECC hardware calculates the pre-ECC BER as data streams in from the OCM, while the processor's main program calculates the post-ECC BER based on the output of the ECC hardware.

In the case that the pre-ECC BER exceeds a predetermined threshold, the processor dynamically redirects its memory traffic away from the error-prone OCM node and can then alert system operators of a potentially failing OCM node. In this way, problematic OCM nodes can be identified while their performance is degrading, but while errors are still correctable. Otherwise, the node will fail and errors will be too prevalent for even the proposed advanced ECC protocol to resolve. This bit-steering approach thus guarantees data integrity in the complete memory system.

5.3.4 Results

Using the above setup, a processor communicates with multiple OCM nodes over the optical interconnection network. With the processor evaluating the BER on the read memory data, error-free operation is confirmed with post-ECC BERs less than 10^{-12}. For all three cases of advanced ECC interleaving, the processor verifies that over a terabit of data can be received without bit errors when pre-ECC BERs of $>4\times10^{-3}$ is induced at one OCM node. For interleaving across four OCM nodes, error-free operation is achieved for induced pre-ECC BERs up to 2.4×10^{-2}. The

baseline case, without interleaving, achieves error-free operation for pre-ECC BERs up to 3.5×10^{-5} (Figure 5.4). Figure 5.5 shows the optical eye diagrams for the eight 10-Gb/s memory payload wavelength channels. This demonstrates an approximately 100× improvement obtained from interleaving across two OCM nodes over the baseline case. Interleaving across three OCM nodes yields only an additional 15% improvement over two-node interleaving. The case of interleaving across four OCM nodes resulted in an almost 700× improvement over the baseline case. Envisioned future implementations supporting 72 OCM nodes can correct for the failure of an entire OCM node, which can then achieving error-free post-ECC performance with a pre-ECC BER of 0.5. It is evident that memory systems based on purely electronic interconnects cannot achieve this level of protection.

The significant improvement in resilience achieved by two-way interleaving over standard ECC is due to the introduction of both double error correction capability and the requirement that errors be located in specific bit locations (i.e. locations of errors are important). This means that moving from standard ECC to the proposed advanced ECC OCM provides a substantial increase in resilience over standard ECC. However, it is observed that sequentially increasing the number of OCM nodes (e.g. from two OCM nodes to three) for interleaving provides incremental improvements. Each extra node decreases the probability of a multi-bit error within each SDRAM chip on an OCM node, but does not completely eliminate the threat of multi-bit errors within a single OCM module until reaching 72-way interleaving. The large resilience difference between three-way interleaving and four-way interleaving, as compared to the improvement in moving from two-way to three-way interleaving, is due to the physical

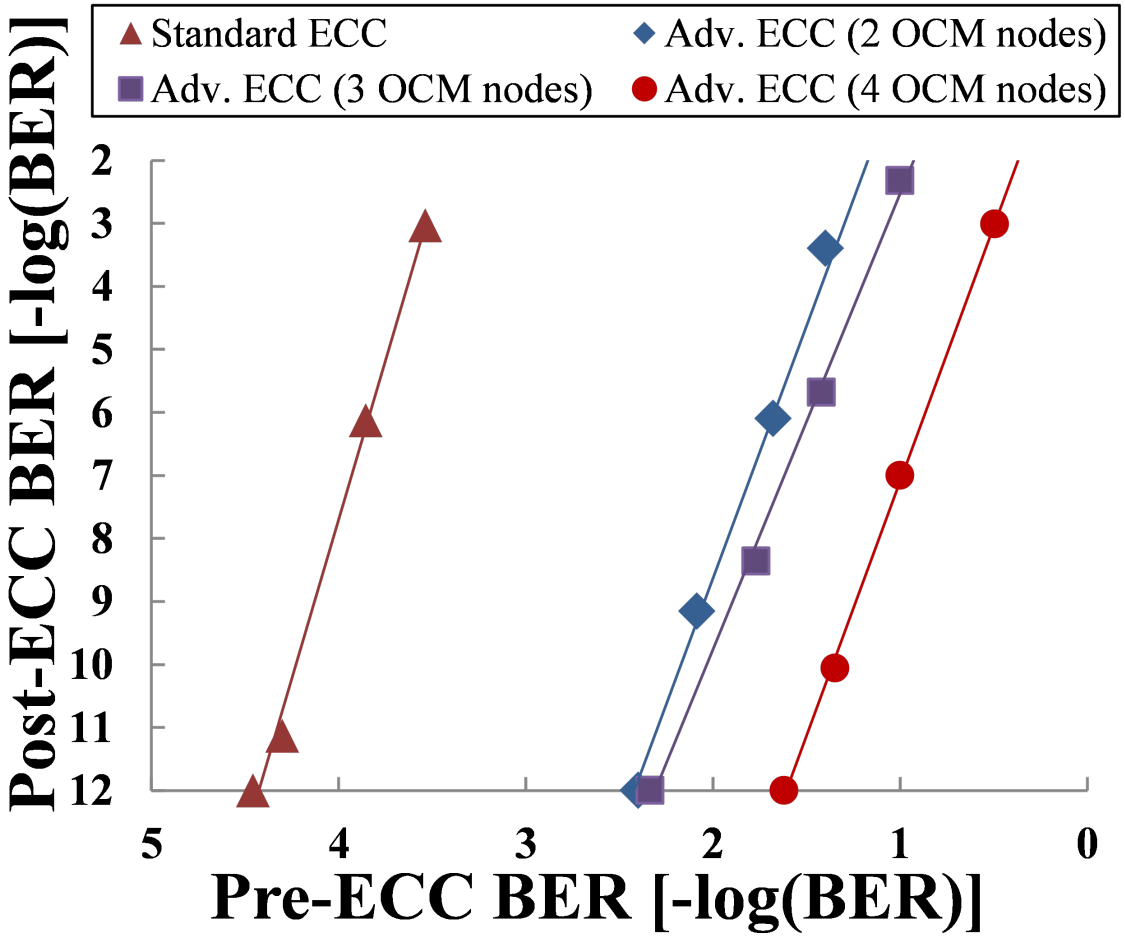

Figure 5.4: Graph of Pre-ECC BER vs Post-ECC BER - Experimentally recovered post-ECC BER as a function of induced pre-ECC BER for the implemented advanced ECC OCM system, for varying number of interleaved nodes. All pre-ECC BERs less than 10^{-5} result in error-free post-ECC BERs (BERs $<10^{-12}$).

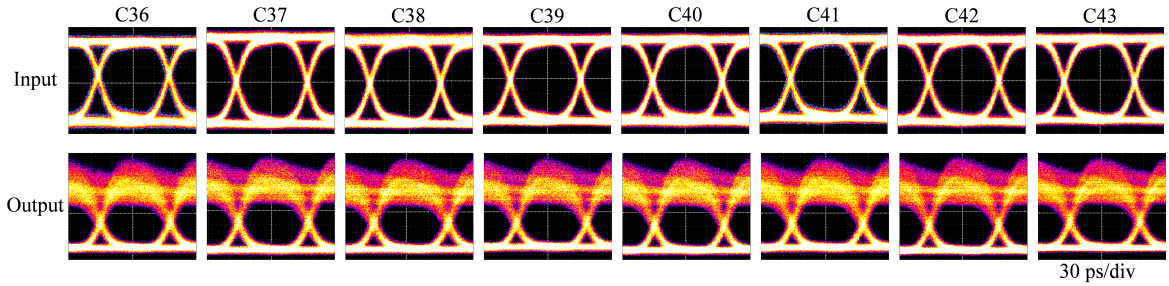

Figure 5.5: Resilient OCM: Optical Eye Diagrams - Optical eye diagrams for the 810-Gb/s memory payload wavelength channels at one network input port (top) and at one network output (bottom).

arrangement of SDRAM chips on the DIMMs used at each OCM node, as previous described. When the memory data is interleaved across four OCM nodes, the SDRAM chips each only contain one bit of the 72-bit ECC memory words, thus providing protection from any multi-bit errors within a single SDRAM chip. Incrementally increasing the number of OCM nodes for interleaving beyond four would therefore result in only small resilience improvements, similar to here when moving from two-way interleaving to three-way. This small resilience increase per node-increase occurs until the system eventually reaches full OCM node failure protection with 72-way interleaving. Using different SDRAM chips, such as those that each store eight bits instead of four, would shift the intermediate "jump" point (i.e. the point between two-way and 72-way interleaving at which an incremental increase in interleaving leads to a large resilience improvement); however, the overall trend would be the same as seen here.

The next step of this experiment uses the 3.5×10^{-5} pre-ECC BER failure point of the baseline case as the error threshold for dynamic bit-steering. This threshold

designates the pre-ECC level beyond which the processor redirects its memory data to a different OCM node. By using the non-interleaving ECC failure point as the error threshold, the system can guarantee zero UEs for each case with interleaving. To demonstrate this functionality, the processor initially communicates with OCM nodes at network ports 0 and 1 with two-way interleaving. The FPGA at OCM node 1 is programmed to steadily degrade the BER of data stored within its memory. When the processor detects a pre-ECC BER at node 1 exceeding 3.5×10^{-5}, the optical headers within the memory transaction are dynamically changed to instantaneously redirect traffic to OCM nodes at ports 0 and 2. This use of dynamic bit-steering enables the processor to continue operating uninterrupted and error-free despite the failure of the OCM node at network port 1.

The use of wavelength-striped optical multicasting reduces overall memory access latency while increasing memory bandwidth. Optical multicasting allows any set of OCM nodes can be accessed in parallel, thus combining the bandwidth of each OCM node to provide the processor with greater aggregate memory bandwidth. Latency is reduced because, as detailed in Section II, each independent memory access incurs tens of nanoseconds of latency, and it is therefore desirable to access all DIMMs simultaneously rather than serially. Constraints from energy dissipation, pin count, and wiring complexity would require any electronically-connected memory system containing the number of memory devices proposed here to perform many memory operations serially rather than simultaneously. For example, 72-way interleaving using electronic interconnects would require up to 18 separate memory access for the best-case electrical memory configuration in which four DIMMs are accessed in parallel

(such as for Chipkill [106]). The memory access power consumption in a Chipkill system can already surpass the processor power [110], and extending such a system to 72 DIMMs is therefore implausible. Furthermore, because retrieving each memory word from main memory would require 18 separate memory accesses, the latency would exceed hundreds of nanoseconds as compared to tens of nanoseconds in existing, less resilient systems.

Here, the total memory access latency is reduced to that of a single memory access, regardless of the number of OCM nodes accessed, by simultaneously accessing any set of OCM nodes with the wavelength-striped optical multicast approach. This experimental demonstrations of interleaving the memory data across two, three, or four nodes all incur the same memory access latency (18 ns). Similarly, the latency incurred for accessing all the OCM nodes in the case of a full 72-node system would also be 18 ns.

5.4 Discussion

This chapter demonstrates how OCM enables novel memory access protocols, and therefore more resilient and flexible memory architectures, that are required for future HPCs and data centers. Large-scale computing systems are continuing to scale and will likely incorporate multiple thousands of servers, leading to many more potential points of failure. As a result, next-generation memory systems will be required to meet and likely surpass today's level of resilience. The limitations of electronic interconnects prohibit modern memory technology from implementing the architectures and ECC protocols necessary for future large-scale computers. To address these challenges, OCM replaces the electronic memory bus with an optical interconnection network, thereby

enabling novel memory architectures and ECC techniques.

This work reports on an advanced ECC protocol that leverages 8×10-Gb/s wavelength-striped multicasting within an interconnection network to allow processors to simultaneously access multiple remote OCM nodes. An FPGA-based processor multicasts and interleaves WDM memory data to multiple OCM nodes where controlled BERs are induced on the stored data. The advanced ECC protocol allows the processor to recover error-free memory data (BERs $<10^{-12}$) for induced bit-error rates up to 2.4×10^{-2}, which would otherwise result in costly loss of data and/or system downtime. This technique is scalable, allowing each processor to simultaneously interleave data across 72 independent OCM nodes, thus protecting memory data from the failure of an entire OCM node. Additionally, the reliable OCM system leverages the concept of dynamic bit-steering to efficiently redirect memory data away from failing memory devices before permanent hardware failures occur.

This work illustrates the novel system architectures and enhanced bandwidth and latency performance attainable by implementing OCM. Such an approach enables robust and dynamic error-correcting techniques that are vital for future large-scale computing systems.

Chapter 6

OCM with Integrated Silicon Photonics

The following chapter presents the closest integration of optics into memory systems to date, and the first characterization of a silicon microring-modulated computer system. In the previous chapters, which outline all progress to-date in designing and implementing OCM systems, discrete, off-the-shelf $LiNbO_3$ modulators served as the E/O interface. While such hardware guarantees high-quality transmission (i.e. large extinction ratios), the resulting photonic transceivers also require large footprints and high power consumptions compared to an integrated, silicon photonic transceiver.

Numerous technological challenges must be overcome before integrated optically-connected memory modules can be commercially implemented. This work seeks to identify challenges that were previously unknown. For example, prior research on silicon microring modulators relies on the use of PRBS data rather than real application traffic, such as memory accesses. The predictable, repetitive nature of PRBS data can provide a close approximation to real-world applications for many laboratory

settings. However, continuous, PRBS-only traffic does not allow for the thorough characterizations required to properly identify and address the challenges in closely integrating silicon photonics with processors and SDRAM.

6.1 Microring Nanophotonic Devices

Silicon photonic components enable CMOS-compatible nanophotonic interconnects [111, 112, 113] with the small footprints necessary for integration with processors and SDRAM. Microring resonators (Figure 6.1) are particularly important for nanophotonic interconnects due to the wide range of functionalities they can provide, including acting as electro-optic modulators, and the small footprint per ring (sub-10 μm^2). The small ring radius yields a large free-spectral range (FSR), which dictates the optical bandwidth separating two adjacent resonator modes [114]. A small FSR allows a single ring to perform optical switching for WDM, wavelength-striped packets as detailed in Chapter 2. A large FSR allows a series of microrings to be cascaded, each aligned to a separate wavelength, to create a WDM modulator bank. Recently, an array of four microring modulators with a footprint of 500 μm^2 was demonstrated to achieve 50-Gb/s modulation [46], which is approximately a 33 Gb/s·μm bandwidth-to-footprint ratio. Overall, microring resonators are ideally suitable as integrated, on-chip photonic components. The remainder of this chapter will focus on the use of microrings as electro-optic modulators for their use within a processor-memory optical interface.

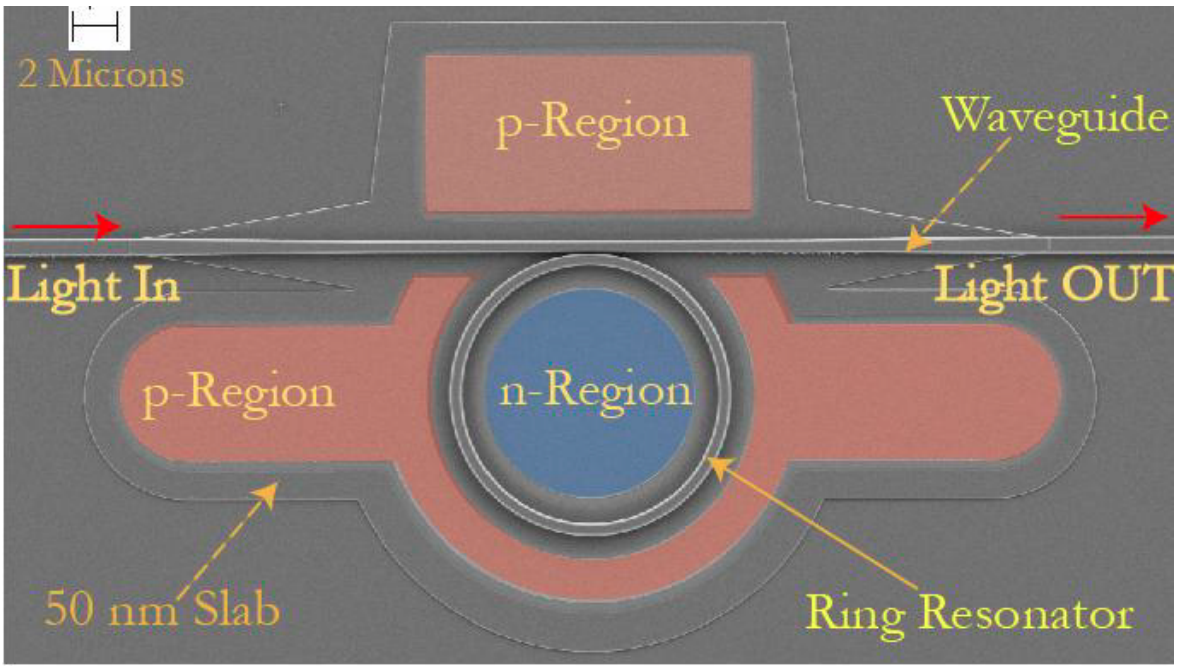

Figure 6.1: SEM Image of Microring Modulator - Scanning-electron-microscope (SEM) image of 6 μm microring modulator.

6.2 Technological Challenges of Integration

The primary performance design constraints for microring modulators are the FSR (which dictates the channel spacing and consequently the total number of channels that can be modulated by the array), and the per-channel data rate. As described above, the FSR is a function of ring size and the maximum spacing (i.e. number of channels to be modulated) is limited by the minimum ring size that can be fabricated. The per-channel data rate is mainly limited by photon lifetime of the optical structures [115], as well as electrical driver circuitry limitations [116] and energy dissipation. These factors are strongly influenced by process fabrication technology and yield, which controls both the optical parameters (such as FSR and carrier lifetime) as well as the electrical drivers (such as signal integrity and energy dissipation).

The fabrication process must also be further advanced to support the accurate creation of a large number of microrings within a single device, or techniques must be developed to efficiently and inexpensively correct for process variations across a photonic chip that would otherwise cause unpredictable behavior (such as incorrect FSRs). Both monolithic integration [117] and 3D integration photonic integration [61] will require currently achievable levels of reliability and repeatability before either process can be incorporated into processor or memory fabrication.

However, even if the fabrication challenges are addressed, the very same resonance characteristics that make microrings attractive solutions for WDM modulator arrays can also cause the rings to become thermally unstable. Microrings are highly sensitive to temperature variations, and it has been demonstrated that even a 1 kelvin (K) temperature fluctuation can sufficiently shift the operating wavelength of a microring

to render it inoperable [118]. Large temperature fluctuations are common within processors, especially within multicore processors that may have "hot spots" from unbalanced computational loads. Alternatively, SDRAM may remain idle while waiting for memory accesses and suddenly heat up during an unpredictable burst of writes or reads.

Research into the topic of microring thermal stability is ongoing. The most common solution in use today is to use local heaters to maintain a constant, elevated temperature at each microring. Integrated temperature monitors have also been proposed [119] as a means of improving heater performance, while wavelength-tracking has been suggested for further improved monitoring [120]. Thermally-insensitive structures have been proposed as a method to isolate the photonic devices from surrounding temperature fluctuations [121], but fail to completely solve the thermal issue. The most promising solution, based on dynamic stabilization [118], utilizes a feedback loop to monitor wavelength shifts and dynamically adjust the ring's bias current to correct for the thermal shift.

All of the above thermal stabilization techniques represent important steps in the integration of nanophotonic devices into processors and memory; however, none to-date adequately address the challenge. Thermal heaters are power-hungry and risk undermining the improved energy-efficiency offered by integrated photonic transceivers. Dynamic temperature monitoring is promising; however, this approach may not be able to adapt sufficiently fast to the sources of thermal perturbation, which may include the bit-rate-dependent modulation current that drives the microrings themselves. As a microring modulates data, the electronic driver circuitry must drive the ring with

varying power levels (logic 0 and logic 1) in order to modulate data in the optical domain. The bit pattern and duration of each bit is controlled by the application and can be unpredictable. This source of thermal stabilization renders thermal isolation techniques unusable because the source of thermal fluctuation is the ring itself. Dynamic stabilization techniques will be required and should be improved to provide faster stabilization.

6.2.1 Line Codes

The challenge of addressing thermal stabilization within microring modulators can be assisted by the electronic driver circuitry in the form of line codes. A line code is a system that takes as an input the data to be transmitted over a communication channel, such as an optical link, and outputs a different pattern of voltage or current values, representing logic zeroes and logic ones, that is optimized for the physical layer devices within that channel.

Line codes are commonly used in serial channels to assist in CDR, which requires a receiver to analyze the incoming data stream's logic level transitions to recover a clock signal. For example, by analyzing a series of bits to be transmitted over the channel, a line code can modify that data into a DC-balanced block of data (i.e. a block of data with an equal number of ones and zeroes) with alternating ones and zeroes to provide adequate transitions for CDR circuitry. Without this functionality, the random nature of traffic within most communication channels would make it impossible to guarantee sufficient transitions for CDR and thus distributed serial links would not function.

One of the most common line codes is 8b/10b encoding, which maps blocks of

8 bits to DC-balanced 10-bit blocks with bounded disparity and sufficient transitions to enable CDR [122]. The standard hardware implementation for 8b/10b encoding (Figure 6.2) divides the 8-bit input into a 5-bit group and a 3-bit group, each of which is encoded separately (into 6 bits and 4 bits, respectively) and concatenated together into the 10-bit encoded output. The use of 10-bit output blocks creates a total of 1024 possible codewords, as compared to the 256 8-bit patterns at the encoder input, and therefore many possible codewords are excluded to allow for a run-length limit of 5 consecutive ones or zeroes and a DC-balanced running disparity. Additionally, some 10-bit codewords are reserved as control data, to indicate the start or end of a frame, or the presence of "idle data" that the receiver should use only for CDR.

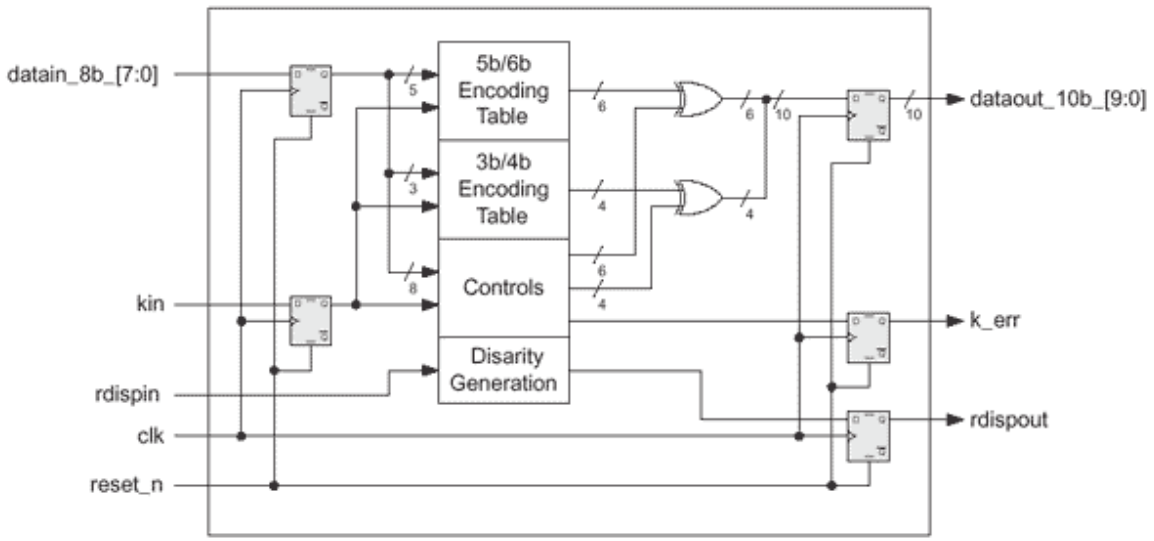

Figure 6.2: 8b/10b Hardware Schematic - Schematic for physical hardware used to implement an 8b/10b line code.

The use of 8b/10b encoding is ideal for systems that are highly sensitive to DC-offsets, which is likely the case for near-term microring modulator-based transceivers.

The guaranteed balance of ones and zeroes within the encoded data driving a microring modulator can provide more predictable thermal conditions. The above thermal stabilization techniques will therefore be less stressed than if attempting to adapt to unpredictable data with potentially long strings of ones or zeroes. The main drawback of 8b/10b encoding is a relatively high overhead, 25%, that lowers the overall channel utilization. As a result, more recent high-speed serial links are migrating away from 8b/10b and toward more efficient line codes.

A more efficient line code is 64b/66b encoding [123] with an overhead of 3.125%. The improved efficiency of 64b/66b encoding has made it a popular choice for 10 Gb Ethernet [4] and InfiniBand [124], of which both leverage electrical and optical transceivers. Analogous to 8b/10b encoding, the 64b/66b line code encodes each 64-bit block of data as 66-bit blocks for the purposes of CDR, run-length limits, and DC balancing. However, unlike 8b/10b, 64b/66b provides only statistical bounds on these criteria as opposed to 8b/10b's guarantee. The 64b/66b encoding is performed by prefixing each 64-bit block of data with two bits: either '01' or '10' depending on the structure of the other 64 bits. If the preamble is '01,' then all 64 remaining bits are data, while a '10' preamble indicates 8 of the remaining 64 bits are reserved control bits while the remaining 56 bits are more control and/or data (as specified by the first 8 control bits).

The two-bit preamble within 64b/66b data guarantees at least one bit transition, either' 0-1' or '1-0', within each 66-bit block. The remaining 64 bits are scrambled using a self-synchronous scrambler function that is typically implemented as a linear feedback shift register (LFSR). The LFSR cannot analyze data in the way that an

8b/10b encoder does; instead, it creates a statistically DC-balance block of data. It is therefore possible, although extremely unlikely, that a block of 64b/66b encoded data will contain 65 ones or 65 zeroes.

However, even a run of ones or zeroes shorter than the maximum 65 bits could be problematic for a microring modulator in a 64b/66b encoded link. A running disparity could potentially shift the temperature of the rings and reduce modulation quality. Additionally, unpredictable run lengths of ones or zeroes, although approximately DC-balanced in the long-term, could also result in frequent, rapid temperature shifts that thermal stabilization techniques will need to correct. It is clear that line codes will be necessary within OCM systems utilizing nanophotonic devices, and both 8b/10b and 64b/66b line codes must be explored.

6.3 Experimental Demonstration

This section presents the first experimental demonstration of an OCM system that utilizes an array of four microring modulators in place of the $LiNbO_3$ components utilized in the previous chapters. The resulting configuration is such that the processor and memory devices operate as if using integrated nanophotonic transceivers. This is the first step in demonstrating that the close integration of photonic and processor/memory hardware is essential to realizing the full potential benefits of OCM. Integrating these silicon photonic components with the proposed OCM system will eliminate the need for power-hungry off-chip electronic wires, alleviate pin-count constraints, and maximize memory bandwidth with WDM. The resulting alignment of off-chip memory bandwidth with on-chip bandwidth enables processors to access

remote memory as if it were local, which is unachievable with electronic interconnects due to pinout limitations.

6.3.1 Silicon Microring Modulators

The microring modulator array (Figure 6.3) utilized in this experiment was fabricated on a silicon-on-insulator (SOI) substrate at the Cornell Nanofabrication Facility [46]. Each ring radius differs such that the circumferences differ by 20, 40, and 60 nm, which allows the four rings to modulate four independent wavelength channels. Each silicon microring is embedded into a PIN diode, and the modulators operate by optical transmission change mediated by free carrier dispersion by injecting or extracting free carriers. The WDM operation of this modulator bank is enabled by the 3 $\pm$ nm inter-channel spacing (Figure 6.4). Previously, the silicon microring modulator array had been characterized at up to 50-Gb/s aggregate bandwidths [46] using electrical amplifiers and pre-emphasis circuitry to achieve high per-channel data rates. The data rates presented here are limited to 2.5 Gb/s, which is achievable through direct modulation of each microring by an FPGA using 1.2 V_{pp} signaling biased at 0.6-0.8 V. This drive voltage conforms to electronic transceiver standards, such as those supported by Altera Stratix FPGAs [79, 80], thus eliminating the need for high-power amplifier circuitry typically required by the large Vπ of $LiNbO_3$ modulators.

The experimental setup presented here creates an OCM system by using the microring modulators to replace the electronic bus between a microprocessor and its main memory. The OCM module consists of four Micron DDR2 SDRAM devices electrically connected to the Altera FPGA, which reformats the memory

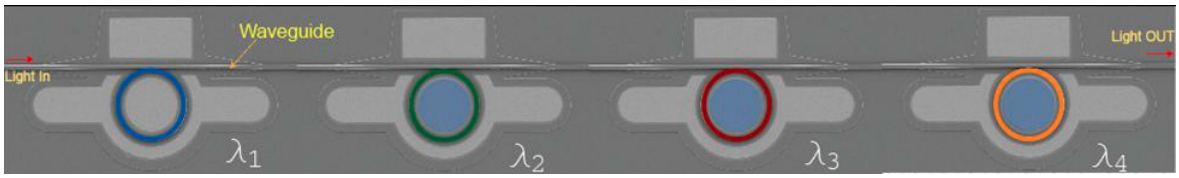

Figure 6.3: SEM Image of WDM Microring Modulator Array - SEM image of the microring modulator array [46]. The color overlay of each ring (blue, green, red, orange) indicates modulation of an independent wavelength channel.

communication into four 2.5-Gb/s data streams to modulate four independent wavelength channels. A second Stratix II GX FPGA implements an emulated microprocessor, along with the memory controller detailed in Chapter 4, which must access the OCM module for all its data. The memory controller optimizes communication across the optical memory link and enables customized communication patterns.

In the first stage of the experiment, the microprocessor's 4×2.5-Gb/s transceivers modulate the four-microring modulator array. The return path from the OCM board is performed electrically. In the second stage of the experiment, the OCM node's transceivers drive the microring modulator array while the microprocessor-to-OCM path is performed electrically. On the receive side, for both stages of the experiment, four PIN-TIA photodetectors with LAs are connected to each FPGA to electrically receive the 4×2.5-Gb/s optical data. This process allows for the characterization of the overall performance of the OCM system in which both processors and memory benefit from integrated nanophotonics. Figure 6.5 shows the experimental setup for one stage with an FPGA-based circuit board, either the processor or memory board, modulating the microring resonator array. Each microring is directly modulated by

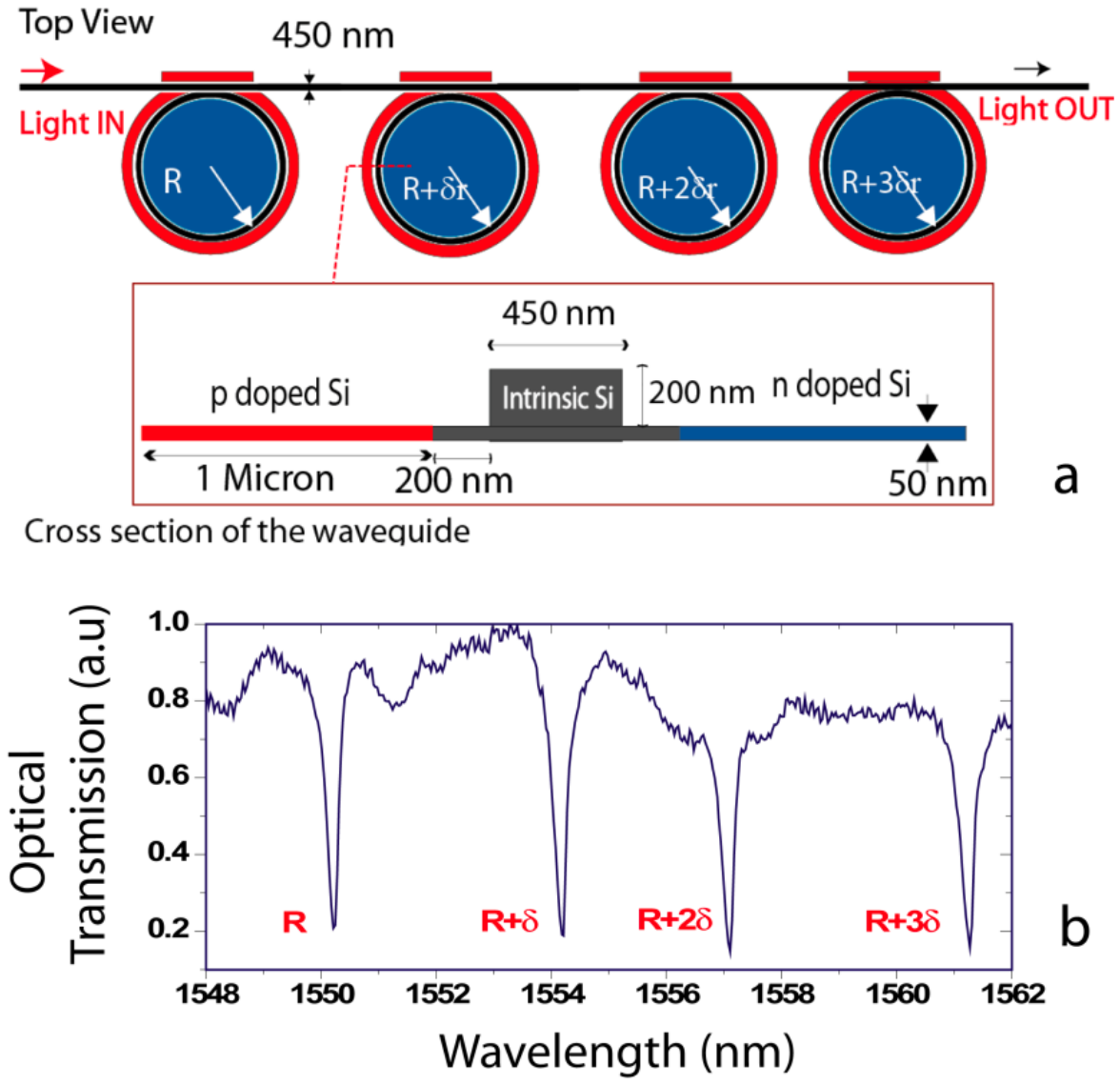

Figure 6.4: Schematic and Spectra of Microring Modulator Array - (a) Schematic of microring modulator array coupled to a single waveguide (inset shows the waveguide cross-section with the doping topology); (b) Transmission spectra of the modulator ring array for quasi-TE polarized light. [46]

the FPGA's high-speed transceivers at 1.2 V_{pp} per channel without pre-emphasis. Four independent wavelength channels (1539.85 nm, 1542.6 nm, 1547.28 nm, and 1551.63 nm) are combined with WDM, amplified by an EDFA, and launched into the chip with an average power of 10 dB per wavelength. Each microring modulates a separate wavelength channel with 2.5-Gb/s OOK data from the FPGA's 4×2.5-Gb/s transceivers. The ring-modulated, WDM memory data exits the chip with an average power of -20 dB before being amplified with an EDFA. The WDM memory data is then demultiplexed to allow each wavelength channel to be electrically received by a separate PIN-TIA-LA for use at the destination processor or memory module.

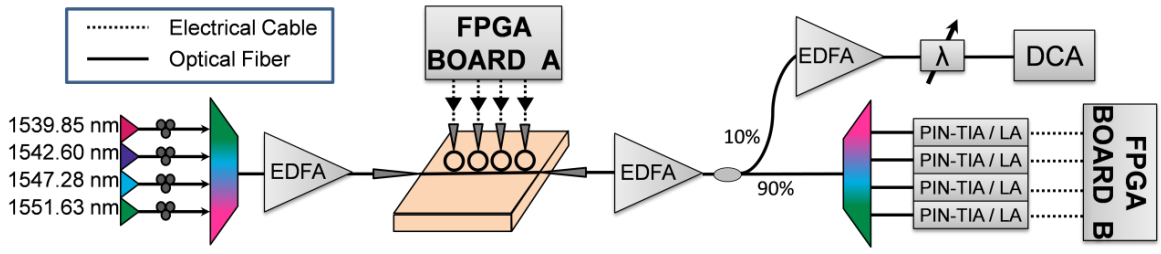

Figure 6.5: OCM With Integrated Silicon Photonics: Experimental Setup - Experimental setup of the processor-memory optical channel communicating via four silicon microring modulators.

To characterize the complete OCM system with the microring modulators, the processor and memory nodes execute a program to repeatedly transmit either 8b/10b- or 64b/66b-encoded memory traffic across the optical memory link. While both line codes guarantee a minimum number of state changes per data block, only 8b/10b is designed to maintain DC balance. Therefore, depending on the memory data, any 64b/66b transmission can contain strings of 65 zeroes or 65 ones for a duration of 26 ns at 2.5 Gb/s. The resulting heating of the microring can potentially impact link

performance and necessitate dynamic stabilization techniques; however, these were not demonstrated here. Since 64b/66b encoding requires only 3.125% communication overhead compared with 25% for 8b/10b encoding, 64b/66b is more desirable and a targeted requirement for any high-performance optical interconnect.

6.3.2 Results

The microprocessor is programmed to generate memory traffic by repeatedly writing to the OCM module with predictable data patterns: all zeroes, all ones, or $2^{31} - 1$ PRBS. These data patterns are chosen from tests used to verify electrically-connected memory systems as well as those for characterizing optical systems. This memory data passes through either 8b/10b or 64b/66b encode/decode hardware at each end of the memory link. Next, after filling the OCM with data, the microprocessor initiates a series of 'read from memory' operations to stream all previously stored data back from the OCM while verifying the data for bit errors. This process repeats for both line codes until one terabit of data has been verified to demonstrate an EMBER less than 10^{-12}.

Figure 6.6 shows the 2.5-Gb/s optical eye diagrams for the 8b/10b and 64b/66b optical memory traffic, as well as for the $2^{31} - 1$ PRBS pattern for comparison. To generate the PRBS eyes, a PPG replaces the FPGA for modulating the four microrings. Table 6.1 contains the measured eye parameters using a DCA to compare the 8b/10b- and 64b/66b-encoded memory data and the PPG-generated $2^{31} - 1$ PRBS, in terms of extinction ratio, rise time, fall time, and root-mean-square (RMS) jitter. The 8b/10b memory data, with its comparatively short consecutive runs of 0's or 1's, results in the

Table 6.1: Comparison of microring-modulated 8b/10b memory data, 64b/66b memory data, and $2^{31}-1$ PRBS.

Measurement	**Data Type**					
	8b/10b Memory		**64b/66b Memory**		$2^{31}-1$ **PRBS**	
	Best (1551.63 nm)	Worst (1539.85 nm)	Best (1551.63 nm)	Worst (1539.85 nm)	Best (1551.63 nm)	Worst (1539.85 nm)
Extinction Ratio	6.52 dB	5.63 dB	5.97 dB	5.78 dB	6.10 dB	6.46 dB
Rise time	106 ps	222 ps	111 ps	311 ps	124 ps	147 ps
Fall time	95 ps	100 ps	102 ps	304 ps	100 ps	104 ps
Jitter (RMS)	20 ps	32 ps	144 ps	169 ps	32 ps	42 ps

overall best performance. The 64b/66b memory data suffers the worst performance, with significantly worse rise/fall times and jitter, due to longer runs of 0's or 1's and resulting microring thermal instability. The performance differences across wavelengths using the 8b/10b or 64b/66b data are attributed to each channel being modulated with different bit patterns (each 2.5-Gb/s data stream is modulating a wavelength as a portion of the total 10-Gb/s memory link). In contrast, each PRBS wavelength contains identical data ($2^{31}-1$ PRBS) and the measured performance is nearly identical for each wavelength.

6.4 Discussion

In order to maximize the performance and energy benefits of optical interconnects, processors and memory devices must leverage integrated photonic components that minimize electrical wiring distances and eliminate the off-chip bandwidth barrier. The challenges facing this level of integration are not only in the realm of fabrication technology, but also in the architectural-level design constraints such as line encoding schemes. Future OCM systems with integrated silicon photonics will therefore require

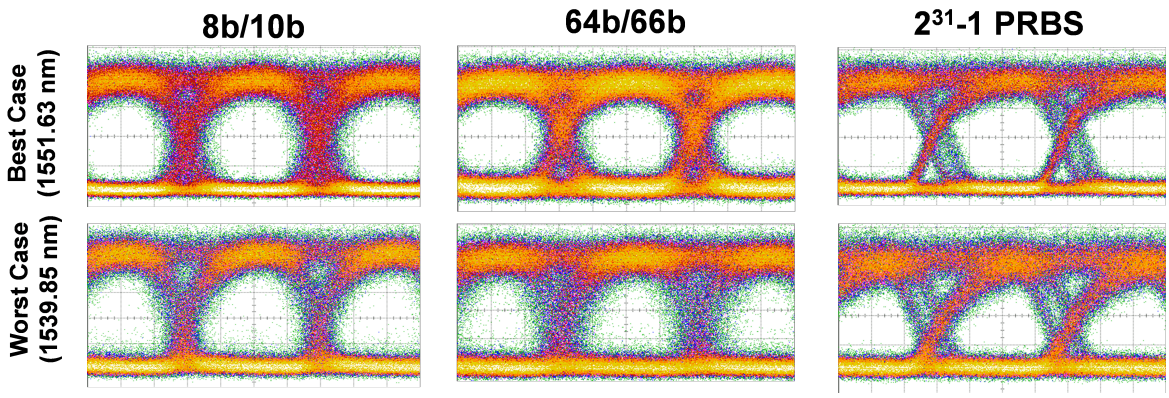

Figure 6.6: OCM With Integrated Silicon Photonics: Microring-Modulated Optical Eye Diagrams - Optical eye diagrams of the 4×2.5-Gb/s wavelength channels showing microring-modulated modulated 8b/10b and 64b/66b memory communication and $2^{31} - 1$ PRBS (100 ps/div).

a combination of physical-layer and architectural optimizations to maintain high-performance optical memory channels.

This chapter presents an experimental demonstration of the first OCM system with integrated silicon photonic microring modulators, which demonstrates the potential for direct interfacing of nanophotonic components with processors and memory and also identifies previously unknown challenges (e.g. the slight signal degradation observed for 64b/66b-encoded data). This work illustrates the energy and performance benefits of close integration of silicon photonics with next-generation memory systems, the flexibility of silicon microrings to operate under the unpredictable communication patterns of large-scale memory systems, and the potential need for thermal stabilization within future silicon photonic transceivers.

Chapter 7

Summary and Conclusions

The work presented in this thesis has focused on the design, implementation, and characterization of optically-connected memory. This final chapter discusses the accomplishments presented here and summarizes ongoing and future work.

7.1 Overview

This work is primarily motivated by the "Memory Wall" facing next-generation computers and the need for a fundamental redesign of processor-memory communication. The use of SDRAM within the memory hierarchy is necessary for its balance of capacity, speed, and cost, which creates a layer of data storage between small, fast on-chip cache and large, slow hard disks. However, the use of SDRAM also creates a bottleneck due to its reliance on off-chip electrical interconnects that limit performance and capacity scalability. The resulting bottleneck has become a limiting factor in the overall performance of large-scale computing systems.

Optically-connected memory has been presented here as a solution to this bottleneck

owing to its ability to eliminate the electronic bus between processors and memory. The need for such a change has become widely apparent to the optical research community and industry, and the work in this thesis takes the first concrete steps in achieving this endeavor. By replacing the wide, electronic memory bus with an optical interconnection network, future processors can access vast amounts of memory and maintain the continued scaling of overall system performance. Throughout this thesis, four key metrics for next-generation memory systems are addressed:

- **Bandwidth**: The low bandwidth-density of electronic interconnects limits the maximum memory bus signaling rate and requires hundreds of processor pins to be dedicated to accessing memory. Furthermore, attempts to increase data rates have imposed limits on the number of SDRAM devices that can be accessed on any one memory channel, thus limiting the overall scalability of memory systems. This work overcomes these challenges through the high bandwidth-density of optics, which allows terabits-per-second of data to traverse a single fiber using WDM. Additionally, optical interconnection networks enable greatly increased bandwidth through simultaneous access of multiple OCM devices with optical multicasting.

- **Latency**: The relatively slow speed of SDRAM compared to modern processors results in that the lower bound of memory access latency is the access time of SDRAM itself, which is on the order of tens of nanoseconds. Future memory systems must therefore implement a memory interface that can provide not only high bandwidths but also ultra-low latencies. The optical network architectures presented here can address this issue through transparent optical routing, in

which high-bandwidth WDM messages traverse an optical network with time-of-flight latency. The use of hybrid packet-and-circuit-switched optical routing enables the custom network-aware memory controller to execute each memory access with optimal latency.

- **Energy**: 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 has become a significant source of power dissipation. Each memory module added to the system increases the chip pin count, necessitating more power-hungry data buffers. Additionally, large numbers of memory devices increases the total physical wiring distance, drastically increasing overall wiring complexity and power dissipation. This work demonstrates how integration of on-chip silicon photonic transceivers will enable processor-memory communication with off-chip bandwidths and energy-efficiencies equal to those of on-chip communication; this is an impossible using electronic interconnects.

- **Resilience**: Increasing the number of memory in a system, as with any commodity hardware, increases the probability of a failure within the memory system as a whole. A single uncorrected error within any system can cause expensive loss of data or system down time, either of which can have significant long-term consequences. The extreme scale of next-generation computers will therefore result in significantly increased probabilities of memory errors, which will require novel ECC protocols to provide the necessary level of reliability. The resilient OCM system presented here can achieve this level of reliability by enabling efficient, simultaneous access to many memory devices. The OCM

system can leverage a combination of ECC and interleaving across many memory devices, which can protect against the failure of an entire OCMM.

7.2 Future Work

With the accomplishments from these first, critical steps in creating optically-connected memory systems, it is important to learn from this existing work and address the remaining challenges.

7.2.1 Photonic Integration

Nanophotonic devices must be integrated as closely as possible with the electronic driver circuitry in order to eliminate power-hungry, bandwidth-limited electrical wires. To achieve this, CMOS-compatible silicon photonic devices are an especially promising technology. The multi-functional microring resonator is especially attractive as a building block for WDM modulators, switches, filters, and photodetectors. The work presented here has focused on the use of ring-based WDM modulators and addresses the challenges of thermal instability with DC-balanced line encodings for serial data. An important next step is the exploration of other microring-based OCM architectures to characterize novel functionalities and identify integration challenges.

Microring-based optical switches are particularly attractive for memory applications. A ring with a large diameter, and therefore a small FSR, can be used to simultaneously switch many wavelengths. As a result, a high-bandwidth, wavelength-striped packet can be switched with the low energy consumption of tuning a single microring. With this configuration, the more data present in each packet,

the more energy efficient the switch becomes. This property is unique to optical switches, and is the opposite trend from electronic switches that consume significantly higher power as bandwidth increases. However, microring-based switches face similar thermal stabilization challenges as the microring modulators investigated in this body of work, and the same solutions used to stabilize microring modulators cannot be directly applied to switches. Although an active microring-based switch is operated at the packet-rate, as opposed to at the bit-rate for microring modulators, thereby experiencing slower temperature changes, the line codes that this work showed to improve microring stability will not apply to ring-based switches. Future dynamic thermal stabilization techniques must therefore correct for thermal perturbations without assistance from higher-level protocols such as line codes.

The fabrication technology challenges faced by integrating optics into processors and memory devices must also be addressed. Until the optical devices can be integrated with monolithically or with 3D integration, the off-chip bandwidth limitations imposed by electronic wiring will continue to bottleneck the memory system. This challenge is further complicated by the high-volume, low-cost commercial market for SDRAM, which will make it difficult for photonics to enter the fabrication process.

7.2.2 Cluster Architectures

The first commercial OCM system will likely be deployed in cluster-scale systems. These systems not only have high performance requirements for their memory systems, but the distributed nature of the physical hardware creates the possibility for novel OCM architectures. For example, one can imagine a configuration in each rack in the

system can contain processors and a small amount of local SDRAM, while separate racks can contain only large banks of remote OCM. The OCM racks would thus have higher time-of-flight latency than the local SDRAM, but the higher capacity and potentially equal memory bandwidth would make this configuration analogous to on-chip caching at the cluster level.

Many such novel system architectures enabled by OCM are not possible using electronic interconnects. A large body of work therefore exists in developing these architectures and analyzing the resulting performance. Similar work is ongoing in the form of chip-scale simulations, whereby a processor with integrated silicon photonics can access SDRAM across an optical link [56, 59], but cluster-scale implementations present a new set of challenges that must be addressed.

7.2.3 Burst-Mode Receivers

The need for burst-mode receivers is made clear throughout the work in this thesis. When processors and memory communicate over an optical network, each message may require costly CDR overhead that can reduce the throughput of the already latency-sensitive memory channel. Additionally, optical network architectures may result in messages arriving at each receiver with different phases and power levels, thus further motivating the need for robust burst-mode receiver circuitry.

A lack of fast burst-mode receivers has been a significant challenge throughout this thesis. The OCMM presented here is capable of incorporating a suitable burst-mode receiver through its high-bandwidth expansion port, and future work must develop such a receiver and characterize its impact on the overall OCM system performance.

7.2.4 Commercial Deployment

By leveraging the work presented in this dissertation, as well as the ongoing work on photonic integration, cluster architectures, and burst-mode receivers, commercially-viable optically-connected memory systems may be realized on a 5-year time scale. High-performance optical transceivers currently exist that may be packaged with processor and memory elements to improve the bandwidth and energy efficiency of large-scale computing systems. The integration of photonic transceivers with memory systems then enables to use of optical switching and optical interconnection networks, which have been demonstrated throughout this dissertation to provide architectural benefits and overall improved system performance. These benefits would provide the motivation to further integrate optical components into processors and memory devices, thus alleviating pin count constraints and off-chip electrical bandwidth limitations, to achieve even greater performance and energy improvements as compared to existing electrically-connected memory systems.

The development of next-generation memory in the form of the HMC provides an ideal insertion point for optical interconnects due to the redesigned memory devices and high-speed interface within the HMC. The HMC is being designed to utilize 3D stacking, with memory cells on separate layers of the stack from high-performance logic, and therefore 3D-integrated photonics may be incorporated into the HMC more easily than with traditional memory devices. In contrast, processor and memory developers today are reluctant to consider 3D integration of optics due to the fact that existing processors and memory do not utilize 3D stacking. Additionally, the HMC interface will require high data rates that are better suited for optical links than power-hungry

electrical busses. The need for SerDes within high-speed optical links is therefore not a drawback within the HMC design space, as SerDes logic would be required in both electrically- and optically-connected implementations. It is therefore possible to leverage the bandwidth density optical components without sacrificing energy efficiency within the electronic driver circuitry, which would risk undermining the overall energy efficiency improved by optical links.

7.3 Summary

The growing gap between processor performance and memory bandwidth has been steadily growing each year. In general, the response to this trend has been the development of system architectures that 1) deploy as many memory devices as electronic wiring permits; and 2) optimize processing nodes to cope with a low data/compute ratio (bytes per FLOP). The inability to break out of this paradigm stems from the reliance on electronic wires as the processor-memory communication link, and the resulting high energy dissipation, low off-chip bandwidth, and overall limited scalability restricts design flexibility. The reluctance to adopt new physical-layer technologies, such as optical interconnects, to overcome these challenges is primarily due to the relative maturity of electrical interconnect technology, its pervasiveness from the chip-level to the system-level, and the resulting high cost and uncertainty associated with migrating technologies.

However, currently, the growing performance requirements of HPCs and data centers are exceeding the limits of electrical interconnects. At the same time, computer-oriented photonic technology is reaching the point of maturity that system-level

interconnects (i.e. rack-to-rack) are increasingly optical. This is because the use of system-level optical links is relatively inexpensive and does not require changes to the underlying electrical I/O within the processors and memory. To continue this trend and leverage optics in chip-level interconnects, such as between processors and main memory, an insertion point is required wherein the physical-layer technology is already undergoing such drastic changes that a shift from electrical links to photonic transceivers can provide drastic performance benefits at negligible additional cost. The 3D-stacked hybrid memory cube [62] currently being developed offers such an opportunity. The move toward 3D integration overcomes many challenges that would otherwise prohibit the deposition of nanophotonic devices onto processors or memory, and the high-performance CMOS logic layer within the memory cube is ideal for interfacing between the memory cells and high-speed optical devices. As processors and memory devices move toward 3D structures with high-speed serial links, the use of optical interconnects will become more attractive in terms of both performance and cost.

This thesis takes the first steps in integrating optical interconnects into memory systems. By developing the architectures, protocols, and physical-layer structures, this work demonstrates the many advantages of optically-connected memory, while identifying previously unknown integration challenges. Some of these challenges have been addressed here, such as the development of an optical-network-aware memory controller and the use of established line codes, while others, such as the need for burst-mode receivers, remain as areas for future impactful research.

Glossary

3D Three dimensional

ASIC Application-specific integrated circuit

BER Bit-error rate

BERT Bit-error-rate tester

CDR Clock and data recovery

CE Correctable error

CPLD Complex programmable logic device

CSA Communications signal analyzer

DARPA Defense Advanced Research Projects Agency

DCA Digital communications analyzer

DFB Distributed feedback

DIMM Dual in-line memory module

DLI Delay-line interferometer

DPSK Differential Phase-Shift Keying

DTG Data timing generator

E-O Electronic-optical

ECC Error-correction codes

EDFA Erbium-doped fiber amplifier

EMBER Effective memory-bit-error rate

EU Uncorrectable error

FB-DIMM Fully buffered dual in-line memory module

FDL Fiber delay line

FLOPS Floating point operations per second

FPGA Field-programmable gate array

FSR Free spectral range

GPIO General purpose input/output

HDL Hardware Description Language

HMC Hybrid memory cube

HPC High-performance computing

K Kelvin

LA Limiting amplifier

LFSR Linear feedback shift register

$LiNbO_3$ Lithium niobate

MC Memory controller

MEMS Microelectromechanical systems

NoC Network on chip

NRZ Non-return-to-zero

OCM Optically-connected memory

OCMM Optically-connected memory module

OCS Optical circuit switching

OIN Optical interconnection network

OOK ON-OFF-keyed

OPS Optical packet switching

OSA Optical spectrum analyzer

PD Photodetector

PLL Phase-locked loop

PPG Pulse pattern generator

PRBS Pseudo-random bit sequence

PSK Phase-Shift Keying

RMS Root mean square

SDRAM Synchronous dynamic random access memory

SECDED Single-error correcting and double-error detecting

SerDes Serializer and Deserializer

SMF Single-mode fiber

SOA Semiconductor optical amplifier

SOI Silicon on insulator

TIA Transimpedance amplifier

VOA Variable optical attenuator

WDM Wavelength-division multiplexing

References

[1] 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," 2008. 1, 82, 83

[2] R. Ho, K. Mai, and M. Horowitz, "The future of wires," Proceedings of the IEEE, vol. 89, no. 4, pp. 490 –504, apr 2001. 1, 25

[3] "International Technology Roadmap for Semiconductors (ITRS) 2011 Edition," [Online]: http://www.itrs.net. 1, 13, 14, 25, 83

[4] "IEEE P802.3ba 40Gb/s and 100Gb/s Ethernet Task Force," [Online]: http://grouper.ieee.org/groups/802/3/ba/index.htmll,. 2, 31, 32, 47, 111

[5] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in Proceedings of the 2004 international workshop on System level interconnect prediction, ser. SLIP '04. New York, NY, USA: ACM, 2004, pp. 7–13. 2, 25, 34

[6] "DDR3 SDRAM Standard," [Online]: http://www.jedec.org/standards-documents/docs/jesd-79-3d,. 2, 11, 13, 15, 45, 47, 50, 52, 62, 68, 88

[7] "Intel Corp. Specification Addendum. Fully Buffered DIMM," [Online]: http;//www.intel.com/technology/memory/FBDIMM/spec/Intel_FBD_Spec_Addendum_rev_p9.pdf,. 2, 17

[8] "NOAA Climate Modeling and Research," [Online]: http://www.ornl.gov/adm/contracts/HPC,. 3

[9] B. Weinman, "Linux Supercomputers Power Oil and Gas Exploration," [Online]: http://archive.hpcwire.com/hpc/959103.html,. 3

[10] R. E. Wyatt, "High-speed computing: scientific applications and algorithm design," R. B. Wilhelmson, Ed. Champaign, IL, USA: University of Illinois Press, 1988, ch. Time-dependent quantum mechanics on supercomputers, pp. 166–168. 3

[11] B. T. Rearden and R. A. Lefebvre, "Getting Started with VIBE as a DICE Plug-in Module," [Online]: http://www.ornl.gov/sci/scale/pubs/scalepub.htm,. 3

[12] N. R. Adiga, M. A. Blumrich, D. Chen, P. Coteus, A. Gara, M. E. Giampapa, P. Heidelberger, S. Singh, B. D. Steinmacher-Burow, T. Takken, M. Tsao, and P. Vranas, "Blue gene/l torus interconnection network," IBM Journal of Research and Development, vol. 49, no. 2.3, pp. 265 –276, march 2005. 3

[13] A. Benner, D. Kuchta, P. Pepeljugoski, R. Budd, G. Hougham, B. Fasano, K. Marston, H. Bagheri, E. Seminaro, H. Xu, D. Meadowcroft, M. Fields, L. McColloch, M. Robinson, F. Miller, R. Kaneshiro, R. Granger, D. Childers, and E. Childers, "Optics for high-performance servers and supercomputers," in Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC), march 2010, pp. 1 –3. 3, 12, 20

[14] "The Top500 List - November 2011," [Online]: http://www.top500.org,. 4, 5

[15] "BM Power 775 Supercomputer Specification," [Online]: http://www-03.ibm.com/systems/power/hardware/775/index.html,. 3, 6

[16] "IBM Unveils New POWER7 Systems To Manage Increasingly Data-Intensive Services," [Online]: http://www-03.ibm.com/press/us/en/pressrelease/29315.wss,. 5

[17] R. Haring, M. Ohmacht, T. Fox, M. Gschwind, D. Satterfield, K. Sugavanam, P. Coteus, P. Heidelberger, M. Blumrich, R. Wisniewski, A. Gara, G.-T. Chiu, P. Boyle, N. Chist, and C. Kim, "The ibm blue gene/q compute chip," Micro, IEEE, vol. 32, no. 2, pp. 48 –60, march-april 2012. 5, 7, 8

[18] "The Green500 List - June 2011," [Online]: http://www.green500.org,. 5

[19] R. Stone and H. Xin, "Supercomputer leaves competitionand usersin the dust," Science, vol. 330, no. 6005, pp. 746–747, 2010. [Online]. Available: http://www.sciencemag.org/content/330/6005/746.1.short 5

[20] "U.S. Energy Information Administration - Electric Power Monthly," [Online]: http://www.eia.gov/electricity/data.cfm,. 5

[21] M. Yokokawa, F. Shoji, A. Uno, M. Kurokawa, and T. Watanabe, "The k computer: Japanese next-generation supercomputer development project," in Low Power Electronics and Design (ISLPED) 2011 International Symposium on, aug. 2011, pp. 371 –372. 7

[22] T. Maruyama, "SPARC64 VIIfx: Fujitsu's New Generation Octo Core Processor for PETA Scale Computing," [Online]: http://www.gwu.edu/upc/tutorials.html,. 7

[23] R. Niranjan Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "Portland: a scalable fault-tolerant layer 2 data center network fabric," SIGCOMM Comput. Commun. Rev., vol. 39, no. 4, pp. 39–50, Aug. 2009. 9

[24] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," SIGCOMM Comput. Commun. Rev., vol. 40, no. 1, pp. 92–99, Jan. 2010. 9

[25] S. Kandula, J. Padhye, and P. Bahl, "Flyways to de-congest data center networks," 2009. 9

[26] B. L. Jacob, S. W. Ng, and D. T. Wang, Memory Systems: Cache, DRAM, Disk. Morgan Kaufmann, 2008. 9, 17

[27] 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 Proceedings of the ACM SIGCOMM 2009 conference on Data communication, ser. SIGCOMM '09. New York, NY, USA: ACM, 2009, pp. 51–62. 9, 63

[28] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in Proceedings of the ACM SIGCOMM 2008 conference on Data communication, ser. SIGCOMM '08. New York, NY, USA: ACM, 2008, pp. 63–74. 9

[29] C. E. Leiserson, "Fat-trees: universal networks for hardware-efficient supercomputing," IEEE Trans. Comput., vol. 34, no. 10, pp. 892–901, Oct. 1985. 10

[30] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," in Proceedings of the ACM SIGCOMM 2010 conference, ser. SIGCOMM '10. New York, NY, USA: ACM, 2010, pp. 339–350. 10, 46

[31] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan, "c-through: part-time optics in data centers," SIGCOMM Comput. Commun. Rev., vol. 41, no. 4, pp. –, Aug. 2010. 10

[32] W. A. Wulf and S. A. McKee, "Hitting the memory wall: implications of the obvious," SIGARCH Comput. Archit. News, vol. 23, no. 1, pp. 20–24, Mar. 1995. 10

[33] S. A. McKee, "Reflections on the memory wall," in Proceedings of the 1st conference on Computing frontiers, ser. CF '04. New York, NY, USA: ACM, 2004, pp. 162–. 10

[34] D. Pham, S. Asano, M. Bolliger, M. N. Day, H. P. Hofstee, C. Johns, J. Kahle, A. Kameyama, J. Keaty, Y. Masubuchi, and et al., "The design and implementation of a first-generation cell processor," ISSCC 2005 IEEE International Digest of Technical Papers SolidState Circuits Conference 2005, vol. 10, no. 2, pp. 184–186, 2005. 11

[35] "XDR XIO Data Sheet Summary," [Online]: http://www.rambus.com,. 11

[36] "Micron Inc. DDR2 SDRAM Datasheet, 2008," [Online]: http://www.micron.com/products/dram/ddr2/,. 12, 45, 64, 65, 68

[37] "Micron Inc. Product specification. 2 GB DDR3 SDRAM," [Online]: http://download.micron.com/pdf/datasheets/modules/ddr3/jsf16c256x64h.pdf,. 12, 15, 25, 64, 65, 66, 68

[38] D. Miller, "Device requirements for optical interconnects to cmos silicon chips," in Photonics in Switching. Optical Society of America, 2010, p. PMB3. 18

[39] W. A. Zortman, M. R. Watts, D. C. Trotter, R. W. Young, and A. L. Lentine, "Low-power high-speed silicon microdisk modulators," in Conference on Lasers and Electro-Optics. Optical Society of America, 2010, p. CThJ4. 18, 34, 56

[40] 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 Conference (ECTC), 2010 Proceedings 60th, june 2010, pp. 240 –246. 18, 34, 56

[41] A. Gnauck, R. Tkach, A. Chraplyvy, and T. Li, "High-capacity optical transmission systems," Lightwave Technology, Journal of, vol. 26, no. 9, pp. 1032 –1045, may1, 2008. 18

[42] O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, and K. Bergman, "The data vortex optical packet switched interconnection network," J. Lightwave Technol., vol. 26, no. 13, pp. 1777–1789, Jul 2008. 18

[43] A. Shacham, H. Wang, and K. Bergman, "Experimental demonstration of a complete spinet optical packet switched interconnection network," in Optical Fiber Communication Conference and Exposition and The National Fiber Optic Engineers Conference. Optical Society of America, 2007, p. OThF7. 18, 20

[44] R. Luijten, W. Denzel, R. Grzybowski, and R. Hemenway, "Optical interconnection networks: The osmosis project," in Lasers and Electro-Optics Society, 2004. LEOS 2004. The 17th Annual Meeting of the IEEE, vol. 2, nov. 2004, pp. 563 – 564 Vol.2. 18, 20

[45] R. Grzybowski, B. Hemenway, M. Sauer, C. Minkenberg, F. Abel, P. Muller, and R. Luijten, "The osmosis optical packet switch for supercomputers: Enabling technologies and measured performance," in Photonics in Switching, 2007, aug. 2007, pp. 21 –22. 18, 20

[46] S. Manipatruni, L. Chen, and M. Lipson, "Ultra high bandwidth wdm using silicon microring modulators," Opt. Express, vol. 18, no. 16, pp. 16 858–16 867, Aug 2010. 19, 105, 113, 114, 115

[47] 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. 15 248–15 256, Aug 2009. 19, 33, 83

[48] F. E. Doany, B. Lee, A. Rylyakov, D. M. Kuchta, C. Baks, C. Jahnes, F. Libsch, and C. Schow, "Terabit/sec vcsel-based parallel optical module based on holey cmos transceiver ic," in Optical Fiber Communication Conference. Optical Society of America, 2012, p. PDP5D.9. 20, 48

[49] D. Brunina, X. Zhu, K. Padmaraju, L. Chen, M. Lipson, and K. Bergman, "10-gb/s wdm optically-connected memory system using silicon microring modulators," in Optical Communication, 2012. ECOC '12. 35th European Conference on, sept. 2012, pp. 1 –3. 20

[50] B. Offrein and P. Pepeljugoski, "Optics in supercomputers," in Optical Communication, 2009. ECOC '09. 35th European Conference on, sept. 2009, pp. 1 –2. 20

[51] C. Lai and K. Bergman, "Broadband multicasting for wavelength-striped optical packets," Lightwave Technology, Journal of, vol. 30, no. 11, pp. 1706 –1718, june1, 2012. 20, 30, 90

[52] B. G. Lee, B. A. Small, J. D. Foster, K. Bergman, Q. Xu, and M. Lipson, "Demonstrated 4x4 gbps silicon photonic integrated parallel electronic to wdm interface," in Optical Fiber Communication Conference and Exposition and The National Fiber Optic Engineers Conference. Optical Society of America, 2007, p. OTuM5. 21

[53] G. Hendry, D. Brunina, J. Chan, L. P. Carloni, and K. Bergman, "Photonic on-chip networks for performance-energy optimized off-chip memory access." in High Performance Embedded Computing (HPEC), 2009. 21, 31

[54] "Datasheet: Corning SMF-28e optical fiber product information," [Online]: http://www.princetel.com/datasheets/SMF28e.pdf,. 21, 32, 42, 56

[55] Y. Katayama and A. Okazaki, "Optical interconnect opportunities for future server memory systems," in High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on, feb. 2007, pp. 46 –50. 21, 42

[56] G. Hendry, E. Robinson, V. Gleyzer, J. Chan, L. Carloni, N. Bliss, and K. Bergman, "Circuit-switched memory access in photonic interconnection networks for high-performance embedded computing," in High Performance Computing, Networking, Storage and Analysis (SC), 2010 International Conference for, nov. 2010, pp. 1 –12. 22, 24, 31, 33, 57, 63, 65, 71, 125

[57] G. Hendry, E. Robinson, V. Gleyzer, J. Chan, L. P. Carloni, N. Bliss, and K. Bergman, "Time-division-multiplexed arbitration in silicon nanophotonic networks-on-chip for high-performance chip multiprocessors," J. Parallel Distrib. Comput., vol. 71, no. 5, pp. 641–650, May 2011. 22, 71

[58] A. Hadke, T. Benavides, S. Yoo, R. Amirtharajah, and V. Akella, "Ocdimm: Scaling the dram memory wall using wdm based optical interconnects," in High Performance Interconnects, 2008. HOTI '08. 16th IEEE Symposium on, aug. 2008, pp. 57 –63. 22, 33

[59] S. Beamer, C. Sun, Y.-J. Kwon, A. Joshi, C. Batten, V. Stojanović, and K. Asanović, "Re-architecting dram memory systems with monolithically integrated silicon photonics," in Proceedings of the 37th annual international symposium on Computer architecture, ser. ISCA '10. New York, NY, USA: ACM, 2010, pp. 129–140. 23, 33, 125

[60] H. Wang, A. Garg, K. Bergman, and M. Glick, "Design and demonstration of an all-optical hybrid packet and circuit switched network platform for next generation data centers," in Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC), march 2010, pp. 1 –3. 24, 30

[61] N. Sherwood-Droz and M. Lipson, "Scalable 3d dense integration of photonics on bulk silicon," Opt. Express, vol. 19, no. 18, pp. 17 758–17 765, Aug 2011. 24, 107

[62] "The Hybrid Memory Cube Consortium," [Online]: http://www.hybridmemorycube.org/,. 24, 128

[63] U. Hoelzle and L. A. Barroso, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 1st ed. Morgan and Claypool Publishers, 2009. 25, 82

[64] R. Baumann, "Soft errors in advanced computer systems," Design Test of Computers, IEEE, vol. 22, no. 3, pp. 258 –266, may-june 2005. 25, 82, 85

[65] C. Lai, D. Brunina, and K. Bergman, "Demonstration of 8x40-gb/s wavelength-striped packet switching in a multi-terabit capacity optical network test-bed," in IEEE Photonics Society, 2010 23rd Annual Meeting of the, nov. 2010, pp. 688 –689. 27, 30, 32, 46

[66] 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, p. 262, Jan 2006. 28, 32, 89

[67] C. Lai and K. Bergman, "Network architecture and test-bed demonstration of wavelength-striped packet multicasting," in Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC), march 2010, pp. 1 –3. 28

[68] D. Brunina, C. Lai, A. Garg, and K. Bergman, "Building data centers with optically connected memory," Optical Communications and Networking, IEEE/OSA Journal of, vol. 3, no. 8, pp. A40 –A48, august 2011. 30, 44, 64, 71, 83

[69] D. Brunina, D. Liu, and K. Bergman, "An energy-efficient optically-connected memory module for hybrid packet- and circuit-switched optical networks," Submitted to IEEE Journal of Selected Topics in Quantum Electronics, 2012. 31, 33

[70] B. N. W. Carlson, T. El-Ghazawi and K. Yelick, "Programming in the partitioned global address space model," [Online]: http://www.gwu.edu/upc/tutorials.html,. 32

[71] J. Cooley and J. Tukey, "An algorithm for the machine calculation of complex fourier series," Mathematics of Computation, vol. 19, no. 90, pp. 297–301, 1965. 33

[72] D. Brunina, C. Lai, A. Garg, and K. Bergman, "Wavelength-striped multicasting of optically-connected memory for large-scale computing systems," in Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference, march 2011, pp. 1 –3. 33, 44, 84

[73] 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 2x22.3gb/s sfi5.2 serdes in 65nm cmos," in Compound Semiconductor Integrated Circuit Symposium, 2009. CISC 2009. Annual IEEE, oct. 2009, pp. 1 –4. 34

[74] P. J. Winzer and R.-J. Essiambre, "Advanced modulation formats for high-capacity optical transport networks," Lightwave Technology, Journal of, vol. 24, no. 12, pp. 4711 –4728, dec. 2006. 34

[75] C. Lai, D. Brunina, C. Ware, B. Bathula, and K. Bergman, "Demonstration of failure reconfiguration via cross-layer enabled optical switching fabrics," Photonics Technology Letters, IEEE, vol. 23, no. 22, pp. 1679 –1681, nov.15, 2011. 36

[76] D. Brunina, C. Lai, and K. Bergman, "A data rate- and modulation format-independent packet-switched optical network test-bed," Photonics Technology Letters, IEEE, vol. 24, no. 5, pp. 377 –379, march1, 2012. 41

[77] D. Brunina, A. S. Garg, H. Wang, C. P. Lai, and K. Bergman, "Experimental demonstration of optically-connected sdram," in Photonics in Switching. Optical Society of America, 2010, p. PMC5. 44

[78] D. Brunina, C. Lai, A. Garg, and K. Bergman, "First experimental demonstration of optically-connected sdram across a transparent optical network test-bed," in IEEE Photonics Society, 2010 23rd Annual Meeting of the, nov. 2010, pp. 622 –623. 44, 57

[79] "Altera Stratix II Device Handbook," [Online]: http://www.altera.com/literature/hb/stx2/stratix2_handbook.pdf,. 45, 113

[80] "Altera Corp. Documentation: Stratix IV Devices," [Online]: http://www.altera.com/literature/lit-stratix-iv.jsp,. 45, 47, 113

[81] "Samsung DDR3 Memory," [Online]: http://www.samsung.com/global/business/semiconductor/minisite/Greenmemory/Products/DDR3/DDR3_Overview.html,. 46, 64

[82] J. Costas, "Synchronous communications," Proceedings of the IRE, vol. 44, no. 12, pp. 1713 –1718, dec. 1956. 46

[83] P. Pepeljugoski, J. Kash, F. Doany, D. Kuchta, L. Schares, C. Schow, M. Taubenblatt, B. Offrein, and A. Benner, "Low power and high density optical interconnects for future supercomputers," in Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC), march 2010, pp. 1 –3. 46

[84] J. A. Kash, A. Benner, F. E. Doany, D. Kuchta, B. G. Lee, P. Pepeljugoski, L. Schares, C. Schow, and M. Taubenblatt, "Optical interconnects in future servers," in Optical Fiber Communication Conference. Optical Society of America, 2011, p. OWQ1. 46

[85] P. Ossieur, N. A. Quadir, S. Porto, M. Rensing, C. Antony, W. Han, P. O'Brien, Y. Chang, and P. D. Townsend, "A 10g linear burst-mode receiver supporting electronic dispersion compensation for extended-reach optical links," Opt. Express, vol. 19, no. 26, pp. B604–B610, Dec 2011. 47

[86] J. Nakagawa, M. Nogami, N. Suzuki, M. Noda, S. Yoshima, and H. Tagami, "10.3-gb/s burst-mode 3r receiver incorporating full agc optical receiver and 82.5-gs/s oversampling cdr for 10g-epon systems," Photonics Technology Letters, IEEE, vol. 22, no. 7, pp. 471 –473, april1, 2010. 47

[87] J. M. D. Mendinueta, J. E. Mitchell, P. Bayvel, and B. C. Thomsen, "Digital dual-rate burst-mode receiver for 10g and 1g coexistence in optical access networks," Opt. Express, vol. 19, no. 15, pp. 14 060–14 066, Jul 2011. 47

[88] "Verilog Resources," [Online]: http://www.verilog.com/,. 50

[89] B. G. Lee, F. E. Doany, S. Assefa, W. M. Green, M. Yang, C. L. Schow, C. V. Jahnes, S. Zhang, J. Singer, V. I. Kopp, J. A. Kash, and Y. A. Vlasov, "20-um-pitch eight-channel monolithic fiber array coupling 160 gb/s/channel to silicon nanophotonic chip," in National Fiber Optic Engineers Conference. Optical Society of America, 2010, p. PDPA4. 56

[90] K. Barker, A. Benner, R. Hoare, A. Hoisie, A. Jones, D. 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, 2005. Proceedings of the ACM/IEEE SC 2005 Conference, nov. 2005, p. 16. 57

[91] S. Rixner, W. Dally, U. Kapasi, P. Mattson, and J. Owens, "Memory access scheduling," in Computer Architecture, 2000. Proceedings of the 27th International Symposium on, june 2000, pp. 128 –138. 58

[92] W.-F. Lin, S. Reinhardt, and D. Burger, "Reducing dram latencies with an integrated memory hierarchy design," in High-Performance Computer Architecture, 2001. HPCA. The Seventh International Symposium on, 2001, pp. 301 –312. 58

[93] S. Rixner, "Memory controller optimizations for web servers," in Microarchitecture, 2004. MICRO-37 2004. 37th International Symposium on, dec. 2004, pp. 355 – 366. 58

[94] D. Oppenheimer, A. Ganapathi, and D. A. Patterson, "Why do internet services fail, and what can be done about it?" 2003. 82

[95] A. Modine, "Web Startups Crumble under Amazon S3 Outage," [Online]: http://www.theregister.co.uk/2008/02/15/amazon_s3_outage_feb_2008,. 82

[96] B. Schroeder, E. Pinheiro, and W.-D. Weber, "Dram errors in the wild: a large-scale field study," in Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems, ser. SIGMETRICS '09. New York, NY, USA: ACM, 2009, pp. 193–204. 82, 85, 88, 94

[97] K. V. Vishwanath and N. Nagappan, "Characterizing cloud computing hardware reliability," in Proceedings of the 1st ACM symposium on Cloud computing, ser. SoCC '10. New York, NY, USA: ACM, 2010, pp. 193–204. 82

[98] C. L. Chen and M. Y. Hsiao, "Error-correcting codes for semiconductor memory applications: A state-of-the-art review," IBM Journal of Research and Development, vol. 28, no. 2, pp. 124 –134, march 1984. 83, 86

[99] D. Brunina, C. P. Lai, D. Liu, A. S. Garg, and K. Bergman, "Optically-connected memory with error correction for increased reliability in large-scale computing systems," in Optical Fiber Communication Conference. Optical Society of America, 2012, p. OTu2B.1. 84

[100] T. May and M. Woods, "Alpha-particle-induced soft errors in dynamic memories," Electron Devices, IEEE Transactions on, vol. 26, no. 1, pp. 2 – 9, jan 1979. 85

[101] S. Mukherjee, J. Emer, and S. Reinhardt, "The soft error problem: an architectural perspective," in High-Performance Computer Architecture, 2005. HPCA-11. 11th International Symposium on, feb. 2005, pp. 243 – 247. 85

[102] E. Normand, "Single event upset at ground level," Nuclear Science, IEEE Transactions on, vol. 43, no. 6, pp. 2742 – 2750, dec 1996. 85

[103] T. J. O'Gorman, J. M. Ross, A. H. Taber, J. F. Ziegler, H. P. Muhlfeld, C. J. Montrose, H. W. Curtis, and J. L. Walsh, "Field testing for cosmic ray soft errors in semiconductor memories," IBM J. Res. Dev., vol. 40, no. 1, pp. 41–50, Jan. 1996. 85

[104] J. F. Ziegler and W. A. Lanford, "Effect of cosmic rays on computer memories," Science, vol. 206, no. 4420, pp. 776–788, 1979. [Online]. Available: http://www.sciencemag.org/content/206/4420/776.abstract 85

[105] H. Mine and K. Hatayama, "Reliability analysis and optimal redundancy for majority-voted logic circuits," Reliability, IEEE Transactions on, vol. R-30, no. 2, pp. 189 –191, june 1981. 85

[106] T. J. Dell, "A White Paper on the Benefits of Chipkill-Correct ECC for PC Server Main Memory." 87, 89, 102

[107] "E7500 Chipset MCH Intelx4 Single Device Data Correction (x4 SDDC) Implementation and Validation, Intel Application note AP-726." 87

[108] "Servers and Storage Technology for the Adaptive Infrastructure, Nr. 2 2006," [Online]: http://h40089.www4.hp.com/integrity/pdf/4AA0-7545EEE.pdf,. 87

[109] A. D. Kshemkalyani and M. Singhal, Distributed Computing: Principles, Algorithms, and Systems, 1st ed. New York, NY, USA: Cambridge University Press, 2008. 90

[110] J. H. Ahn, N. P. Jouppi, C. Kozyrakis, J. Leverich, and R. S. Schreiber, "Future scaling of processor-memory interfaces," in Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, ser. SC '09. New York, NY, USA: ACM, 2009, pp. 42:1–42:12. 102

[111] I. Young, E. Mohammed, J. Liao, A. Kern, S. Palermo, B. Block, M. Reshotko, and P. Chang, "Optical i/o technology for tera-scale computing," in Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International, feb. 2009, pp. 468 – 469,469a. 105

[112] R. Beausoleil, J. Ahn, N. Binkert, A. Davis, D. Fattal, M. Fiorentino, N. Jouppi, M. McLaren, C. Santori, R. Schreiber, S. Spillane, D. Vantrease, and Q. Xu, "A nanophotonic interconnect for high-performance many-core computation," in High Performance Interconnects, 2008. HOTI '08. 16th IEEE Symposium on, aug. 2008, pp. 182 –189. 105

[113] A. Shacham and K. Bergman, "Building ultralow-latency interconnection networks using photonic integration," Micro, IEEE, vol. 27, no. 4, pp. 6 –20, july-aug. 2007. 105

[114] B. E. A. Saleh and M. C. Teich, Fundamentals of Photonics, J. W. Goodman, Ed. Wiley, 1991, vol. 5. [Online]. Available: http://doi.wiley.com/10.1002/0471213748 105

[115] S. Manipatruni, Q. Xu, and M. Lipson, "Pinip based high-speed high-extinction ratio micron-size silicon electrooptic modulator," Opt. Express, vol. 15, no. 20, pp. 13 035–13 042, Oct 2007. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-15-20-13035 107

[116] F. Caignet, S. Delmas-Bendhia, and E. Sicard, "The challenge of signal integrity in deep-submicrometer cmos technology," Proceedings of the IEEE, vol. 89, no. 4, pp. 556 –573, apr 2001. 107

[117] M. Georgas, J. Orcutt, R. Ram, and V. Stojanovic, "A monolithically-integrated optical receiver in standard 45-nm soi," in ESSCIRC (ESSCIRC), 2011 Proceedings of the, sept. 2011, pp. 407 –410. 107

[118] K. Padmaraju, J. Chan, L. Chen, M. Lipson, and K. Bergman, "Dynamic stabilization of a microring modulator under thermal perturbation," in Optical Fiber Communication Conference. Optical Society of America, 2012, p. OW4F.2. 108

[119] C. T. DeRose, M. R. Watts, D. C. Trotter, D. L. Luck, G. N. Nielson, and R. W. Young, "Silicon microring modulator with integrated heater and temperature sensor for thermal control," in Lasers and Electro-Optics (CLEO) and Quantum Electronics and Laser Science Conference (QELS), 2010 Conference on, may 2010, pp. 1 –2. 108

[120] C. Qiu, J. Shu, Z. Li, X. Zhang, and Q. Xu, "Wavelength tracking with thermally controlled silicon resonators," Opt. Express, vol. 19, no. 6, pp. 5143–5148, Mar 2011. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-19-6-5143 108

[121] B. Guha, B. B. C. Kyotoku, and M. Lipson, "Cmos-compatible athermal silicon microring resonators," Opt. Express, vol. 18, no. 4, pp. 3487–3493, Feb 2010. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-4-3487 108

[122] A. X. Widmer and P. A. Franaszek, "A dc-balanced, partitioned-block, 8b/10b transmission code," IBM J. Res. Dev., vol. 27, no. 5, pp. 440–451, Sep. 1983. [Online]. Available: http://dx.doi.org/10.1147/rd.275.0440 110

[123] R. Walker and R. Dugan, "64b/66b low-overhead coding proposal for serial links," [Online]: http://grouper.ieee.org/groups/802/3/10G_study/public/jan00/walker_1_0100.pdf,. 111

[124] "The InfiniBand Trade Association," [Online]: http://www.infinibandta.org/. 111