Demonstration of End-to-End Bit-Parallel Memory Transactions Across the Ultra-Low Latency Data Vortex Optical Packet Switch

Howard Wang and Keren Bergman
Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, New York 10027
howard@ee.columbia.edu

Carl Gray and David C. Keezer
School of Electrical and Computer Engineering, Georgia Institute of Technology, 777 Atlantic Drive NW, Atlanta, Georgia 30332

Abstract: We experimentally demonstrate end-to-end memory transactions between a processor and memory node across the data vortex optical packet switch. Successful read and write transactions via the network at 4×2.5 Gb/s are verified.

OCIS codes: (060.4259) Networks, packet-switched; (060.4250) Networks

1. Introduction

The evolution of modern high-performance computing (HPC) systems towards the now ubiquitous massively parallel, distributed paradigm can be attributed to the auspicious scaling of the performance and cost of current generation microprocessors and memories. Furthermore, due to the prevalence of multicore processor architectures, the computational performance and off-chip bandwidth of each node have scaled superlatively. In order to meet the demanding memory requirements of typical supercomputing workloads, contemporary HPC systems are organized in distributed shared memory architectures, which are characterized by substantial physically distributed memories logically shared across upwards of tens of thousands of compute nodes. As these systems scale in size, the increased latency of remote memory transactions and interprocessor exchanges represents a critical communications bottleneck to system scalability [1]. As such, low-latency high-bandwidth message exchange represents one of the key challenges to the realization of these advanced computing systems [2].

Conventional electronic interconnects are fundamentally incapable of supporting the necessary capacities and latencies that will be demanded by future high-performance computing systems. Exemplified by inherently low time-of-flight latencies, high capacities via wavelength division multiplexing (WDM), and low power consumption, optical interconnection networks have been widely proposed to relieve the apparent performance disparity between the nodes and the interconnect [3]. The data vortex optical packet switch architecture [4] represents a novel design for a photonic interconnection network uniquely aimed at leveraging the overwhelming capabilities of photonic media while elegantly addressing the challenges facing all-optically switched fabrics: namely, the lack of sophisticated optical processing techniques and mature optical memories. The architecture supports the ultra-low latency transmission of short messages and leverages WDM to achieve terabit capacities via a wavelength striped message format. The performance of the data vortex has been extensively studied and characterized empirically and through architectural simulation.

Currently, high-speed serial electronic protocol standards prevail in nearly all modern computing platforms. While these standards have enabled the interconnection of current generation computing systems at a small scale, their incongruity with the short, bursty nature of the messages supported by the data vortex architecture represents a critical challenge at the data transport interface between the processors and memories and the optical network. Test electronics and interfaces for transporting data originating from a PCI Express data source to the data vortex optical packet switch have been previously proposed and developed [5]. In an effort to evaluate the applicability of the data vortex in a real-world system, the aforementioned high-speed electronics have been further extended in its interfacing capabilities. Furthermore, the data vortex network is being migrated to a new programmable node architecture, facilitating more sophisticated future empirical architectural evaluations. In this work, we experimentally validate the error-free transmission of data generated from an x86-based computer terminal across five node hops. Furthermore, successful end-to-end memory transactions are established between a processor and memory node via the data vortex optical packet switch and high-speed packet formatter. Error-free read and write transactions across the network at 2.5 Gb/s per channel across four wavelengths are confirmed.

a) b)

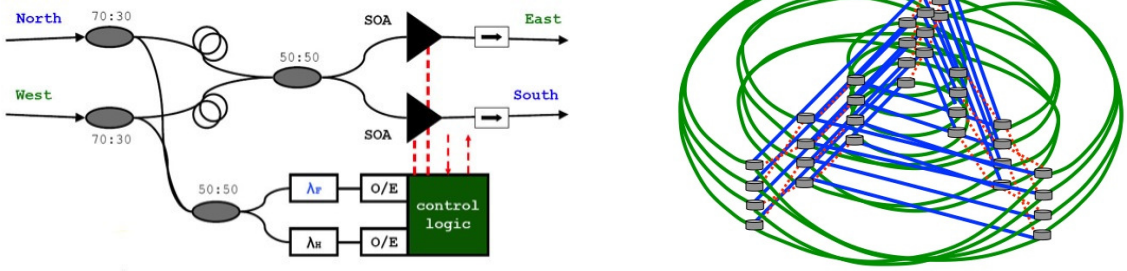

Fig. 1. a) 2×2 data vortex switching node design. b) Topology of a 12×12 data vortex all-optical packet switch consisting of 36 2×2 switching nodes. Green lines represent deflection fibers while blue lines represent ingression fibers.

2. Overview of the Data Vortex Architecture

The data vortex architecture, specifically designed to be implemented as an all-optical packet switched topology, is comprised of simple 2×2 all-optical switching nodes (Fig. 1a). Each node utilizes two semiconductor optical amplifier (SOA) devices which perform the switching operation. Given the wide gain-bandwidth of the SOAs, the network utilizes a multi-wavelength striped packet format (Fig. 1b), with high bit-rate payload data segmented across multiple channels and low bit-rate addressing information encoded on dedicated wavelengths, one bit per wavelength per packet. Passive optical splitters and filters within the node extract the relevant routing information (a frame bit to denote the presence of a packet and a header bit to determine the switch's configuration), which are subsequently detected by low speed receivers. The SOA pair is controlled via high-speed electronic decision circuitry, and routes the packet based on the recovered header information.

In a data vortex topology, the 2×2 switching nodes are organized as concentric cylinders and addressed according to their location within the topology, represented by their cylinder, height, and angle (C, H, A)(Fig 1c). Data propagates from the outermost cylinder toward the innermost cylinder and contentions within the network are handled via deflection routing, thus forgoing the need for optical buffering. Nodes occupying the same height in adjacent cylinders are interconnected via a set of equal-length ingression fibers while nodes within a common cylinder of different height are connected by equal-length deflection fibers. The crossing patterns of the deflection fibers are organized in a banyan topology, facilitating load balancing and ensuring that a packet reaches the correct height in a minimal number of hops. As all data is maintained in the optical domain, the network is capable of achieving ultra-low latencies.

3. High-Speed Packet Formatter and Memory Emulator

The system used for this demonstration is built upon a modular test and verification platform presented in [6]. The base platform (Fig. 2) integrates a field-programmable gate array (FPGA), utilized for data storage, processing, and control, as well as general purpose I/O interfaces including USB and a PCI Express card edge connector and a bank of channels directly accessible by the FPGA to the north edge of the board. The platform includes two arrays of modular interfaces to the east and west of the board which are tailored for this demonstration with a total of 18 data channels capable of supporting the target data rate of 2.5 Gb/s each. Transmission modules serialize a slower wide data word into a high-speed bit stream with support for variable amplitude output and a range of data skew up to 10ns in 10ps programmable increments. Additional resolution through the programmable delay device is available if required through the use of an analog vernier and a serial DAC. Receiver modules reverse the process through a similar processing path, including deskew capability, returning the serial channel to a wide word for input to the FPGA.

The FPGA formats the data in time and adjusts the delay on the channels to construct an output packet that is compatible with the Data Vortex in structure and timing alignment. The FPGA is also utilized in this demonstration to implement an emulated memory structure to serve as a target destination for traffic across the network. This memory node is designed to respond to write and read transactions from processing nodes on the network.

4. End-to-End System Demonstrator and Experimental Results

For this demonstration, only one tester board was utilized, as such the FPGA was programmed to support the functions of both a processing node and a memory node. Figure 2 depicts the layout of the experimental system

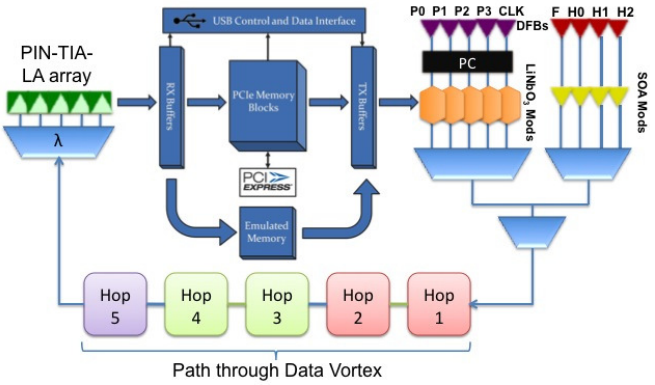

Fig 2. Experimental system diagram: memory transactions originate from the demonstrator board and traverse five switching node hops. Red denotes ingression cylinder nodes. Green denotes middle cylinder. Violet denotes egression cylinder.

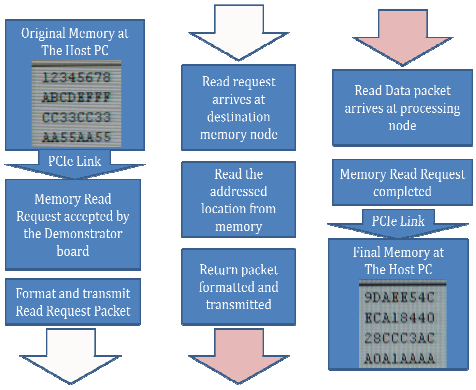

Fig. 3. Flow diagram of end-to-end transaction

demonstrator. Incoming packets are processed and forwarded to the appropriate logic with memory read and write requests being processed by the emulated memory and all other transactions being received by the processing node logic. The data flow for the various transactions begins at a host PC connected to the demonstrator board across a PCI Express connection. This PC writes to or reads from a memory location that is stored at a remote memory node. The FPGA translates this transaction into a network packet that is generated and injected into the system. Upon arrival at the destination, the memory node decodes the request and writes the data from the packet or reads the address and constructs a return packet. The data from this returning packet is ultimately forwarded back to the host processor completing the end-to-end signal path.

While Figure 2 demonstrated the data flow of the system demonstrator, Figure 3 shows the logical progression of a specific end-to-end transaction. This transaction begins at the top left with a memory location within a system processor. The processor issues a read request to update the data at this location, passing the request to the demonstrator system via the PCI Express link. The board processes the request and translates it into a read request packet which is formatted, injected into the network, and routed to the destination memory node through the data vortex (white arrows at the bottom of column 1 and top of column 2, corresponding to the first packet of the transaction). While this packet arrives back at the same originating board, due to the format of the packet it is detected as a memory request and is forwarded to the emulated memory within the FPGA where the addressed memory is read. The data from this address is formatted into a return packet and injected back into the network as the second packet of the transaction (pink arrows, bottom of column 2 and top of column 3). This packet travels through the data vortex, once again arriving at the demonstrator board where it is processed as the response to the read transaction and is routed back through the FPGA to the originating system processor via the PCI Express link.

5. Conclusions

We experimentally demonstrate the error-free propagation of data originating from a computer terminal across the data vortex optical packet switch. End-to-end memory transactions are also successfully transmitted across the network between processor and memory nodes emulated on the custom high-speed electronic evaluation board. We achieve error-free transmission at 2.5 Gb/s per channel across four wavelengths, validating the feasibility of the data vortex as an interconnect for practical HPC systems.

6. References

[1] Dai, D. and Panda, D. K., "How Can We Design Better Networks for DSM Systems?," in *Proc. 2nd Int'l Workshop Parallel Computer Routing and Comm.* pp. 171-184, Jun 1997.

[2] Kodi, A. K. and Louri, A., "Design of a High-Speed Optical Interconnect for Scalable Shared-Memory Multiprocessors," *IEEE Micro* vol. 25, 1, pp. 41-49, Jan 2005.

[3] D.A.B. Miller, ''Rationale and Challenges for Optical Interconnects to Electronic Chips,'' *Proc. IEEE*, vol. 88, no. 6, pp. 728-748, June 2000.

[4] O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, K. Bergman, "The Data Vortex Optical Packet Switched Interconnection Network," *Journal of Lightwave Technology* **26** (13) 1777-1789 (Jul 1, 2008).

[5] Gray, C.E., Liboiron-Ladouceur, O., Keezer, D.C., Bergman, K., "Co-development of test electronics and PCI Express interface for a multi-Gbps optical switching network," *Proc. of the IEEE Intl. Test Conf.* (ITC'07), Paper 22.1,2007

[6] Keezer, D.C., Gray, C., Majid, A., Minier, D., Ducharme, P., "A Development Platform and Electronic Modules for Automated Test up to 20 Gbps," *Proc. of the IEEE Intl. Test Conf* (ITC'09), Paper 14.3, 2009