Experimental Demonstration of End-to-End Message Passing for HPC systems through a Hybrid Optical Switch

Conference Paper · October 2011

DOI: 10.1364/ECOC.2011.Tu.6.K.5 · Source: IEEE Xplore

5 authors, including:

Yawei Yin
Alibaba Group
68 PUBLICATIONS **589** CITATIONS
SEE PROFILE

Xiaohui Ye
University of California, Davis
25 PUBLICATIONS **205** CITATIONS
SEE PROFILE

Roberto Proietti
University of California, Davis
106 PUBLICATIONS **941** CITATIONS
SEE PROFILE

Experimental Demonstration of End-to-End Message Passing for HPC systems through a Hybrid Optical Switch

Yawei Yin, Xiaohui Ye, Roberto Proietti, Venkatesh Akella, and S.J.B. Yoo
Department of Electrical and Computer Engineering, University of California, Davis, California 95616, USA
{yyin, rproietti, xye, akella, sbyoo}@ucdavis.edu

Abstract: We experimentally demonstrate end-to-end message passing through a hybrid optical switch and FPGAs emulating a High Performance Computing system. The experiment-driven simulation results verify the highly-scalable and low-latency nature of the switch.
OCIS codes: (200.4650) Optical Interconnects; (200.0200) Optics in Computing.

1. Introduction

Reliable, scalable, high-throughput and low-latency message passing is critical to High Performance Computing (HPC) systems. The Message Passing Interface (MPI) has been defined to facilitate computation in data-intensive applications, such as scientific computing. The interconnection architecture among massive parallel processors plays a significant role in narrowing the gap between the hardware's capabilities and the actual performance can be delivered [1]. In the traditional electrical interconnection architecture, both wire-connected electrical buses and hierarchical electrical switches suffer from high blocking probability, huge power dissipation, and long end-to-end latency [2]. Recently, optical interconnects are emerging as energy-efficient means to provide high capacity communications in data centers and HPC systems [3, 4], and the Arrayed Waveguide Grating Routers (AWGR) based optical switching [5, 6] can effectively exploit the inherent parallelism of wavelength division multiplexing (WDM) to achieve high throughput, high power efficiency, and low latency interconnection.

However, the lack of practical optical buffer technology makes it difficult to seek all-optical solutions that are mature and agile enough to fully prevent packet losses in the interconnection switch. In this paper, we theoretically and experimentally demonstrate very low latency and zero packet loss rate interconnection between emulated parallel processors using a hybrid optical packet switch including the all-optical AWGR and the FPGA RAM based electrical buffer. This paper investigates and analyzes in detail the end-to-end latency of packets with and without contending in a message passing paradigm. The experimental results were fed back into the simulator, which in turn generated the experiment-driven results verifying the scalable and low latency nature of the switch.

2. Hybrid Optical Switching Architecture

There is a theoretical observation that it is possible to emulate purely Output-Queued (OQ) switches with Combined Input Output Queued (CIOQ) Switches running at approximately twice the line-rate, or namely with a "speedup" of two [7]. However, it is not easy to implement an electrical switch with speedup of two due to the limited IO port counts or insufficient memory bandwidth at high line rates (to speed up the line-rate by two, you need either double the clock frequency or double the number of IO wirings, neither of which is practical). The inherent parallelism of WDM technology makes it straight-forward to implement the "speedup" with optical switches. The AWGR based optical hybrid switch effectively takes advantage of the wavelength parallelism to increase the line rate at each output port. As Fig. 1 shows, the AWGR switching fabric can easily realize the speedup of k by providing 1:k optical DEMUX and k receivers at each AWGR output [5]. In this paper, we equipped each output with a 1:2 optical DEMUX and two receivers. Besides the AWGR, the core of the switch contains Tunable Wavelength Converters (TWCs), the FPGA based electrical control plane, the FPGA based electrical distributed loopback buffer, the Label Extractors (LEs), and the Fiber Delay Lines (FDLs). In the experiment, the switching size N is set to be 4, and since $k=2$, we have $N/k=2$ contention groups. Here contention groups were utilized to reduce the complexity of the arbitration process by a factor of k as well as to make the architecture of the control plan distributable [5]. Therefore, as illustrated in Fig. 1, packet 2 and packet 3 can go through the switch to the same destination output port without contention since they belong to different contention group, while on the other hand, packet 1 and packet 4 were contended at output port n since their input port falls into the same contention group. Packet 4 was directed to the distributed loopback buffer through output port 2' (Out(2') in Fig. 1). The loopback buffer then reapplied to the arbiter for the credit after a fixed delay to send packet 4 again. If granted, the packet would be send out through input 2' (In(2') in Fig. 1) and be directed to the destination port n. Otherwise if denied, the loopback buffer will wait and retry again and again until the packet will finally be send out. For fairness, to the same arbiter the application from the loopback buffer will always have higher priority. However, there's still a possibility that the loopback

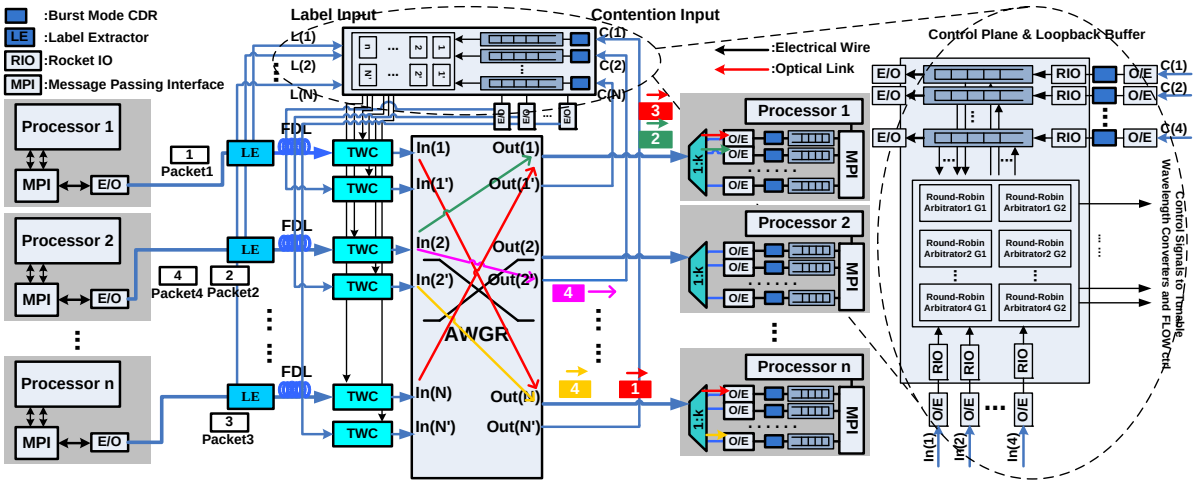

Fig. 1. The Hybid Optical Switch Architecture with FPGA based control plan and distributed loopback buffer

buffer will overflow at high traffic load or hotspot of the traffic. The almost-full status of the buffer will trigger the flow control scheme on the control plane. Thereafter the transmitter will be slowed down or fully stopped until another message was sent to indicate the buffer was ready to receive more packets.

Fig. 1 shows the experiment emulating a HPC system with 4 parallel processors fully connected with each other through the hybrid optical switch. Concurrently executing threads on processor P1, P2 and P4 use MPI for communication. The end to end latency of packet 1, 2, 3 and 4 was measured respectively.

3. Emulated Multiprocessor System

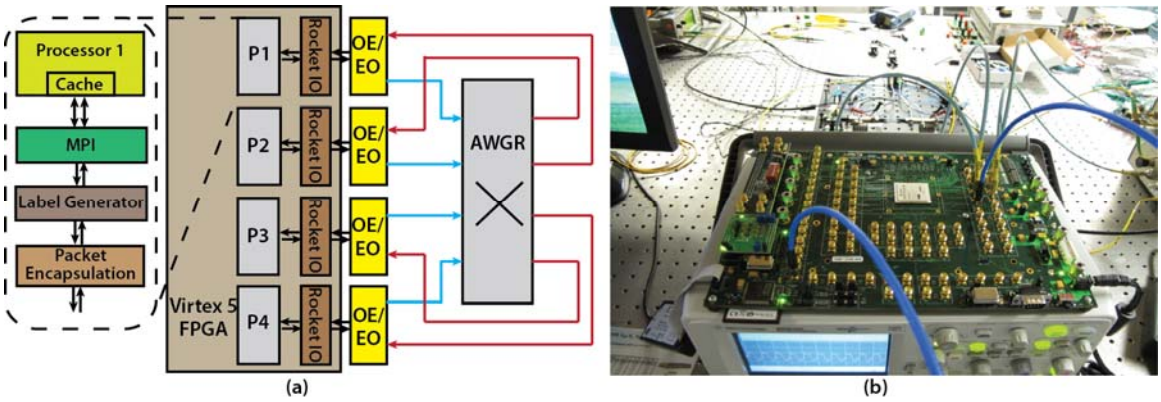

Fig. 2 (a) FPGA emulated parallel multiprocessor system architecture (b) Virtex 5 evaluation board implementing processors & Rocket IO

Fig.2 depicts the FPGA based multiprocessors to emulate the High Performance Computing systems. Four of MicroBlaze Soft Processor Cores [8] were instantiated on Virtex 5 with MPI interfaces. The label generator will extract the destination address from the outgoing message and translate it into the format of a "label" which can be recognized by the optical switch. After the label was generated, the message will be encapsulated into packets with a fixed size of 16 bytes in this experiment scenario. And then the packet will be serialized by Rocket IO [9] and transmitted onto the 1.25 Gbps optical link through E/O convertors. On the reverse direction, the incoming packets after O/E convertion will be de-serialized by Rocket IO and then decapsulated into messages.

4. Experimental and Simulation Results

In the experimental scenario, firstly the emulated processor P2, P4 were passing packet 2 and packet 3 at the same time to processor P1. Since P2 and P4 were connected to input port 2 and 4 respectively (which fall into different contention group), the packets could get through the switch to output 1 on different wavelength without any contention. Their end to end latency was measured respectively. Secondly, processor P1 and P2 sent out packet 1 and packet 4 simultaneously, which were contended for output port 4 in the same arbiter. With the default priority, packet 1 got granted immediately while packet 4 was blocked and directed to the loopback buffer. After a pre-defined delay process in the loopback buffer, the message re-applied and then got grant to be sent out.

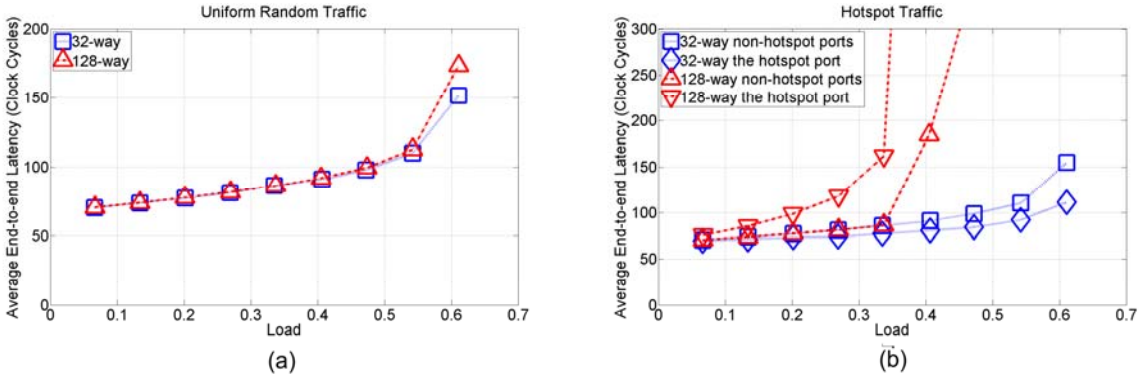

Fig. 3, Experiment-driven simulation results of end-to-end latency under (a) uniformly distributed random traffic, (b) hotspot traffic

The end-to-end transmission latency for the un-contended packet 1, 2 and 3 were all 68 clock circles (544 ns), while the end-to-end latency for packet 4 was 118 clock circles (944 ns). The delay can be analyzed as follows:

$$\tau_{e2e} = \tau_{cp} + \delta_{lb} \cdot \tau_{lb} + D_{RIO}^{Rx} + D_{RIO}^{Tx} + D_{BCDR} + \sum_{l \in P} T_l \quad (1)$$

$$\tau_{cp} = D_{RIO}^{Rx} + D_{ARB} + D_{TLD} \quad (2)$$

$$\tau_{lb} = D_{BCDR} + D_{ARB} + D_{buf} + D_{RIO}^{Rx} + D_{RIO}^{Tx} \quad (3)$$

$$\delta_{lb} = \begin{cases} 0, & with\ buffering \\ 1, & without\ buffering \end{cases} \quad (4)$$

Where T_l is the traverse delay on the fiber links which belong to path P, τ_{cp} is the delay of the control plane, τ_{lb} is the delay introduced by loopback buffer, D_{RIO}^{Tx} is the delay added by Rocket IO transmitter (*Tx*), D_{RIO}^{Rx} is the delay added by Rocket IO receiver (*Rx*), D_{BCDR} is the delay introduced by the burst mode clock and data recovery chipset, D_{ARB} is the arbitration delay in the control plane, D_{TLD} is the tunable laser switching latency, and D_{buf} is the minimum time duration that a contended packet stays in the electrical loopback buffer.

In the experiment, the measured delay for each parameter is as follows (unit: clks – one clock circles is 8 ns):

$\sum_{l \in P} T_l$: ~12.5 clks, corresponding to around 20 meters of fiber pigtails in the lab, D_{RIO}^{Rx}: 14 clks, D_{RIO}^{Tx}: 6 clks, D_{ARB}: 4 clks, D_{TLD}: 9.5 clks, D_{BCDR}: 8 clks, D_{buf}: 16 ns. Since the experiment was restricted by the lab environment, the exact length of the fiber pigtails/patch cords can't be measured precisely. By feeding these measured parameters into the DOS simulator [5], we can observe the average end-to-end latency in a larger switching scale (say 32-way and 128-way) for uniformly distributed random traffic and hotpot traffic respectively as shown in Fig. 3 (a) and (b). The traffic saturated at load of 0.6 since we were using small size packets with relatively large inter-packet gaps.

5. Conclusion

This paper presented an experimental demonstration of the end-to-end message passing in an emulated HPC systems interconnected by an AWGR based hybrid optical switch. The end-to-end latency of message passing with and without contention were investigated and analyzed in detail. By feeding back the experimental results into the switch simulator, we verified the scalability of the hybrid optical switching architecture as well as the efficient message passing capability with zero loss rate and very low end-to-end latency under different traffic model.

References

1. Darius Buntinas, G.M., William Gropp. *Design and Evaluation of Nemesis, a Scalable, Low-Latency, Message-Passing Communication Subsystem*. in *Proceedings of the International Symposium on Cluster Computing and the Grid, CCGRID 06. Sixth IEEE International Symposium on* 2006.
2. Shalf, J., *The new landscape of parallel computer architecture*. Journal of Physics: Conference Series,, 2007. **78**(1).
3. Liboiron-Ladouceur, O., et al., *The Data Vortex Optical Packet Switched Interconnection Network.* Journal of Lightwave Technology, July 2008. **26**(13).
4. Hemenway, R., et al., *Optical-packet-switched interconnect for supercomputer applications.* Journal of Optical Networks, 2004.
5. Ye, X., et al. *DOS - A scalable Optical Switch for Datacenters*. in *ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*. 2010.
6. Roberto Proietti, X.Y., Yawei Yin, Andrew Potter, Runxiang Yu, Junya Kurumida, Venkatesh Akella, and S. J. B. Yoo, *40 Gb/s 8×8 Low-latency Optical Switch for Data Centers*, in *Optical Fiber Communications Conference (OFC)*2011: Los Angeles, CA.
7. Shang-Tse, C., et al., *Matching output queueing with a combined input/output-queued switch.* Selected Areas in Communications, IEEE Journal on, 1999. **17**(6): p. 1030-1039.
8. http://www.xilinx.com/tools/microblaze.htm
9. http://www.xilinx.com/support/documentation/user_guides/ug196.pdf

This work was supported in part by the Department of Defense through contract #H88230-08-C-0202 and Google Research Awards.