An All-Optical PCI-Express Network Interface for Optical Packet Switched Networks

Odile Liboiron-Ladouceur, Howard Wang and Keren Bergman
Department of Electrical Engineering, Columbia University, New York, New York 10027
ol2007@columbia.edu

Abstract: We report on the implementation of a power-efficient, low-latency edge node PCI-Express interface to a WDM optical packet switched network. Scalability is examined under the limits set by PCI-Express jitter specification on the packet propagation.

OCIS codes: (200.4650) Optical interconnects, (060.4510) Optical communications;

1. Introduction

As the need for real time high-bandwidth and on-demand applications continues to grow at immense rates, next generation core optical networks are driven to deliver scalable commensurate capacities [1]. Many applications already require terabytes of data exchange bandwidths including video on-demand, gaming, e-science via grid computing and 100G Ethernet deployments [2]. Optical packet switched (OPS) networks offer a potential high-capacity platform as an integrated part of the backbone network for interconnection of high-performance computing nodes within the distributed network. Some of the key features of OPS networks, namely switching granularity, low latency and scalable bandwidth, are incongruent to the emerging paradigm of ubiquitous computing communications networks and the merging of the data and transport layers. In successfully inserting OPS networks within the optical core backbone, a critical challenge is the design of an efficient and transparent interface between the electronic and the optical data structures. In recent years, the advancement of various standard protocols, such as PCI Express, HyperTransport and RapidIO has enabled interconnectivity among diverse communicating modules. PCI Express (PCIe), now in its third-generation, has emerged as the I/O interconnect for high-speed serial buses supporting chip-to-chip, board-to-board, graphics, and other applications in advance switching [3]. It is therefore likely that PCIe will form the dominant interface protocol for computing nodes at the network edge. The main challenge in I/O interconnection is the latency at the edge nodes associated with traffic grooming, which can create bottlenecks in distributed networks.

In this paper, we present a novel all-optical PCIe edge node interface to OPS distributed networks (Fig. 1a). An all-optical interface at the edge of the OPS network can deliver highly scalable aggregated bandwidth using WDM technology with improved latency and power efficiency over conventional electronic solutions. Packet grooming at the network interface consists of mapping the serial PCIe packet onto parallel WDM channels within the packet slot time structure (Fig. 1b). At the egress node, the PCIe packet is optically reconstructed using passive optical components. The effect of chromatic dispersion on the reconstructed PCIe packet is investigated and network scalability is derived from the PCIe jitter budget specification for the media.

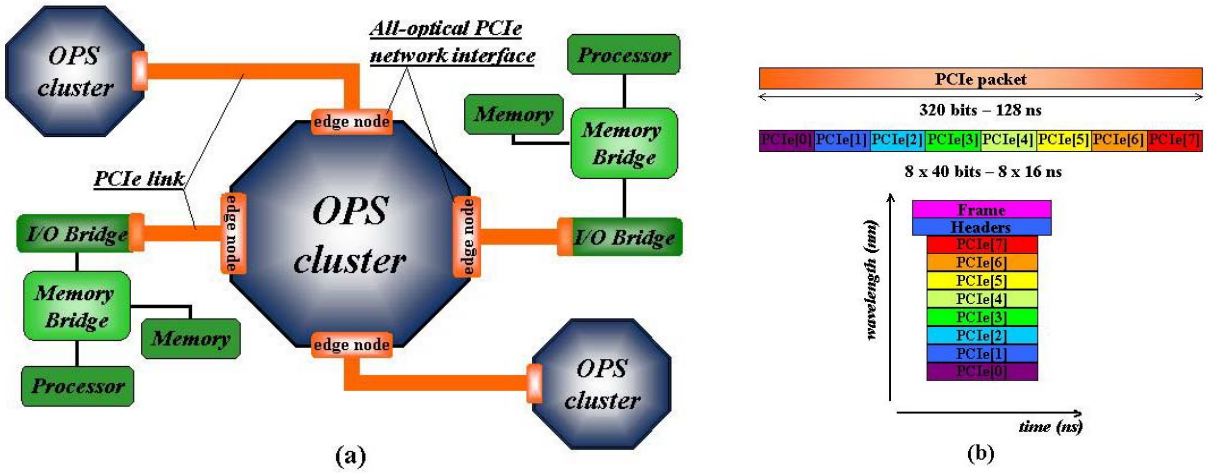

Fig. 1: (a) Distributed network of computer nodes or clusters with PCIe network interface at the edge nodes to other computer nodes, networks and shared memory and processors. (b) Illustration of a PCIe packet to a WDM packet structure mapping.

2. PCI-Express Network Interface

The implemented network interface maps PCIe packets for the Data Vortex, a time-slotted self-routing OPS network architecture. A 12×12 I/O port prototype was recently implemented demonstrating the ability of routing high-bandwidth multiple-wavelengths packets [4]. The packet structure is composed of multi-wavelength payload channels with lower bit-rate wavelength-striped header and frame signals for proper routing (Fig. 1b). At the ingress network interface node, the PCIe packet is first stripped and encoded onto multiple WDM channels. Longer PCIe packets can be mapped on demand using more WDM channels. In this work, the PCIe packet is a 320-bit long sequence mapped onto eight WDM channels. To minimize the effect of chromatic dispersion, the channels are closely spaced by 0.8 nm from 1543.72 nm to 1549.33 nm. The PCIe packet bit sequence used is a training sequence with bytes BC, F7, F7, 14, 02, 00 followed by ten bytes of 4A. This represents 160 bits with 8b/10b encoding. The packet is then repeated twice for a 320-bit long sequence. This particular training sequence is therefore DC balanced and provides sufficient transitions for clock recovery. The eight payload channels contain 40 bits each (16 ns-long packet). The beginning of the PCIe packet is encoded on the shortest wavelength (PCIe[0], 1543.72 nm), the following 40 bits are encoded on the next wavelength (PCIe[1], 1544.48 nm) and so on until the last 40 bits of the 320-bit long PCIe packet are encoded on the longest wavelength (PCIe[7], 1549.33 nm). The patterns are programmed onto an Agilent ParBert E81250, which externally modulates each of the eight cooled DFB lasers individually at 2.5 Gb/s. The modulated wavelengths are then coupled onto one fiber to create an optical packet containing eight payload channels of 40 bits (PCIe[0..7]) with frame and header signals for proper routing. The packet is then launched into the Data Vortex network.

At the egress network interface node, the main challenges are to correctly sample very short packets (tens of ns) and maintain synchronization within the distributed network. One approach consists of using a source-synchronous clock reference encoded on a separate channel that samples every payload channels individually [5]. A novel, more power-efficient approach is employed here where a low-latency all-optical interface containing passive components reconstructs the 8b/10b encoded packet such that conventional clock recovery can be used. The all-optical network interface is presented in Fig. 2. The multi-wavelength packet emerging from the Data Vortex network is first amplified by an EDFA, and then split and filtered into eight optical channels using DWDM bandpass filters. Each channel is then delayed by 16 ns with respect to the shorter adjacent channel using optical fiber delay lines. The filtered wavelengths are multiplexed onto one fiber (Fig. 2 top-right) to a 10.7-Gb/s DC-coupled optical receiver module with an integrated limiting amplifier. The optical power difference between channels is apparent (Fig. 2 bottom-right), but the limiting amplifier alleviates the optical power difference by quantifying the optical signal to a 450 mV_{p-p} signal. The broadband signal of eight optical channels PCIe[0..7] optically delayed with respect to each other represents the reconstruction of the 320-bit long PCIe packet initially launched into the Data Vortex network, which can then be used by the standard PCIe interface for proper clock and data recovery.

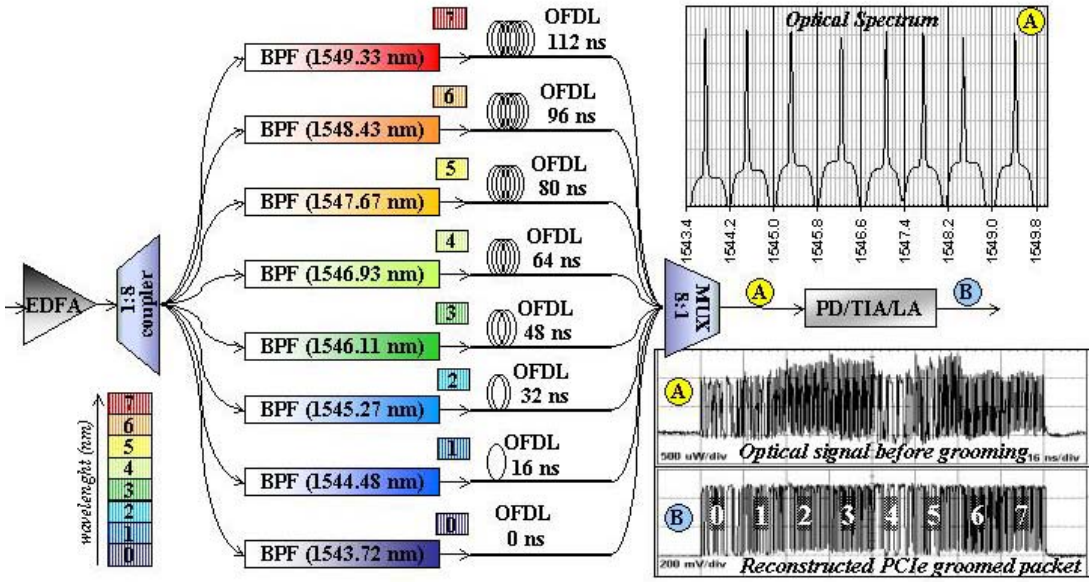

Fig. 2: Schematic of the all-optical PCIe network interface with packet reconstruction using bandpass filters (BPF) and optical fiber delay lines (OFDL). Top-right inset shows the optical spectrum at the output of the multiplexer with all eight payload channels after propagating through 3 nodes of the 12 × 12 I/O Data Vortex. The bottom-right inset shows the corresponding broadband optical packet before the optical receiver module (PD/TIA/LA), and the groomed and reconstructed 320-bit long PCIe packet.

3. Timing Alignment

The all-optical PCIe packet reconstruction approach relies heavily on timing alignment. Timing misalignment will occur due in part to path variations, unequally cut fiber lengths, and most importantly chromatic dispersion. To investigate the tolerance of the all-optical network interface to timing misalignment, the electrically reconstructed PCIe packet is sampled using a 2.7-Gbit/s analyzer (Agilent E4861A). The threshold voltage is set to the common mode voltage of the optical receiver module. In Fig. 3a, the three possible cases are investigated by changing the timing delay of channel PCIe[3] with respect to its adjacent channel at a shorter wavelength (PCIe[2]). The three possible cases examined are (1) large timing skew in PCI[2] data, (2) no significant skew, and (3) large timing skew in PCIe[3] data with respect to PCI[2]. The analyzer output data of the PCIe packet is shown for all three cases. As reported in Fig. 3b, sampling errors occur for cases (1) and (3) where the timing skew is purposely set to be exceedingly large. The measured timing skew tolerance for case (2) is ± 155 ps.

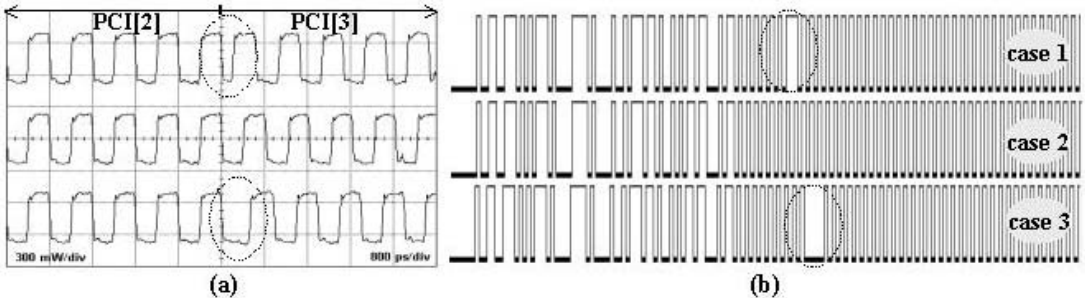

Fig. 3: (a) Receiver output signal showing the timing alignment between PCIe[2] and PCIe[3] data for the three possible cases. (b) Corresponding digitized output signal of the analyzer for all three cases, exhibiting an error in the case of excessive timing skew.

In a practical implementation, the major contributor to timing skew creating the misalignment between the data at different wavelengths is chromatic dispersion within the optical fiber. Hence, timing skew between each PCIe data translates into an overall phase jitter in the converted electrical signal representing the entire PCIe packet. PCIe standard specification has a jitter budget that allocates 90 ps of total jitter for the media [6]. To estimate the design limitations imposed by the all-optical interface we consider a $10k \times 10k$ I/O port Data Vortex which would incorporate approximately 200 m of interconnected single-mode fiber. For this large-scale network, eight channels distributed across 5.6 nm will experience a total of approximately 20 ps of timing skew difference between the shortest and longest channels. Hence, up to 31 wavelengths spaced by 0.8 nm could be used in this design such that the total timing jitter is maintained within the PCIe standard specification.

4. Conclusions

We have shown an innovative approach to alleviating the bottleneck of edge grooming in the backbone of a distributed high-bandwidth network by proposing an all-optical network interface to the PCI-Express standard. PCIe packets are mapped onto the WDM packet structure of an OPS network and optically reconstructed for a low-power and low-latency network interface. The effect of chromatic dispersion on the overall phase jitter of the reconstructed PCIe packet is found to not be a limitation on the network scalability.

The authors would like to thank Aldo Hoyt at Agilent for helping with the PCIe patterns. This work was supported in part by the National Science Foundation under grant ECS-0322813 and by the Department of Defense (subcontract B-12-664).

5. References

[1] A.A.M Saleh, and J.M. Simmons, "Evolution Toward the Next-Generation Core Optical Network," IEEE J. Lightw. Technol., **24**, 9, 3303-3321, (Sept. 2006).

[2] L. Smarr *et al.*, "The OptiIPuter, quartzite, and starlight projects: A campus to global-scale testbed for optical technologies enabling Lambda-Grid computing," OFC/NFOEC, OWG7, Anaheim, CA, (Mar. 2005).

[3] D. Mayhew and V. Krishnan, "PCI express and advanced switching: evolutionary path to building next generation interconnects," 11th Symposium on High Performance Interconnects Proceedings, 21-29, (Aug. 2003)

[4] A. Shacham, B.A. Small, O. Liboiron-Ladouceur, and K. Bergman, "A Fully Implemented 12x12 Data Vortex Optical Interconnection Network," IEEE J. Lightwave Technol., **23**, 10, 3066-3075, (Oct. 2005).

[5] O. Liboiron-Ladouceur, C. Gray, D.C. Keezer, and K. Bergman, "Bit-Parallel Message Exchange and Data Recovery in Optical Packet Switched Interconnection Networks," IEEE Photon. Technol. Let., **18**, 779-881 (Mar. 15, 2006)

[6] Standard Development Group, "PCI express Random Jitter, Deterministic Jitter and Bit Error Rates Specifications," http//:www.pcisig.com.