Low-power, transparent optical network interface for high bandwidth off-chip interconnects

Odile Liboiron-Ladouceur,[1,2] Howard Wang,[1] Ajay S. Garg[1], and Keren Bergman[1]

[1]*Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, , NY, 10027, USA*
[2]*Currently with the Department of Electrical and Computer Engineering, McGill University, 3480 University St., Montreal, QC, H3A 2A7, Canada*
[*]*Corresponding author:* howard@ee.columbia.edu

Abstract: The recent emergence of multicore architectures and chip multiprocessors (CMPs) has accelerated the bandwidth requirements in high-performance processors for both on-chip and off-chip interconnects. For next generation computing clusters, the delivery of scalable power efficient off-chip communications to each compute node has emerged as a key bottleneck to realizing the full computational performance of these systems. The power dissipation is dominated by the off-chip interface and the necessity to drive high-speed signals over long distances. We present a scalable photonic network interface approach that fully exploits the bandwidth capacity offered by optical interconnects while offering significant power savings over traditional E/O and O/E approaches. The power-efficient interface optically aggregates electronic serial data streams into a multiple WDM channel packet structure at time-of-flight latencies. We demonstrate a scalable optical network interface with 70% improvement in power efficiency for a complete end-to-end PCI Express data transfer.

©2009 Optical Society of America

References and links

1. S. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, P. Iyer, A. Singh, T. Jacob, S. Jain, S. Venkataraman, Y. Hoskot, and N. Borkar, "An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS," in *Proceedings of IEEE Int. Solid-State Circuit Conference*, Dig. Tech. (ISSCC) (Institute of Electrical and Electronics Engineers, 2007), Paper 5.2.
2. M. Tremblay and S. Chaudhry, "A Third-Generation 65nm 16-Core 32-Thread Plus 32-Scout-Thread CMT SPARC® Processor," in *Proceedings of IEEE Int. Solid-State Circuits Conference*, Dig. Tech. (ISSCC) (Institute of Electrical and Electronics Engineers, 2008), pp. 82-83.
3. M. Kistler, M. Perrone, and F. Petrini, "Cell Multiprocessor Communication Network: Built for Speed," IEEE Micro. **26**, 10-23 (2006).
4. Y. Li, E. Towe, and M. W. Haney, Eds., "Special Issue on Optical Interconnections for Digital Systems," Proc. IEEE **88**, 723-863, (2000).
5. A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, and M. B. Ritter, "Exploitation of Optical interconnects in Future Server Architecture," IBM J. of Res. Dev., **49**, 755-775 (2005).
6. C. L. Schow, F.E. Doany, O. Liboiron-Ladouceur, C. Baks, D. M. Kuchta, L. Schares, R. John, and J. A. Kash, "160-Gb/s, 16-Channel Full-Duplex, Single-Chip CMOS Optical Transceiver," in *Proceedings of Optical Fiber Communication Conference & Exposition and the National Fiber Optic Engineers Conference*, Technical Digest (OFC/NFOEC), paper OThG4, 2007.
7. M. Asghari, "Silicon Photonics: A Low Cost Integration Platform for Datacom and Telecom Applications," in *Proceedings of Optical Fiber Communication Conference & Exposition and the National Fiber Optic Engineers Conference,* Technical Digest (OFC/NFOEC), paper NThA4, 2008.
8. D. M. Kuchta, Y. Taira, C. Baks, G. McVicker, L. Schares, and H. Numata, "Optical Interconnects for Servers," Jpn. J. Appl. Phys. **47**, 6642-6645 (2008).

9. O. Liboiron-Ladouceur, H. Wang, and K. Bergman, "An All-Optical PCI-Express Network Interface for Optical Packet Switched Networks," in *Proceedings of Optical Fiber Communication Conference & Exposition and the National Fiber Optic Engineers Conference,* Technical Digest (OFC/NFOEC), paper JWA59, Mar. 25-29, 2007.

10. O. Liboiron-Ladouceur, H. Wang, and K. Bergman, "Low Power Optical WDM Interface for Off-Chip Interconnects," in *Proceedings of the 20th Annual meeting of the IEEE Lasers and Electro-Optics Society,* Technical Digest (*LEOS*) (Institute of Electrical and Electronics Engineers, 2007), pp. 680-681.

11. J. A. Kash, "Leveraging Optical Interconnects in Future Supercomputers and Servers," in *Proceedings of the 16th IEEE Symposium on High Performance Interconnects*, (Institute of Electrical and Electronics Engineers, 2008), pp. 190-194, Aug. 2008.

12. International Technology Roadmap for Semiconductors (ITRS), 2005.

13. Standard Development Group, http//:www.pcisig.com.

14. 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. **26**, 1777-1789 (2008).

15. E. Dulkeith, F. Xia, L. Schares, W. M. J. Green, and Y. A. Vlasov, "Group Index and Group Velocity Dispersion in Silicon-on-Insulator Photonic Wires," Opt. Express **14**, 3853–3863 (2006).

16. C. Gunn, "CMOS Photonics for High-Speed Interconnects," IEEE Micro, **26**, 58–66 (2006).

17. R. Soref, "The Past, Present, and Future of Silicon Photonics," IEEE J. Sel. Top. Quantum Electron. **12**, 1678–1687 (2006).

18. M. Lipson, "Guiding, Modulating, and Emitting Light on Silicon-Challenges and Opportunities," J. Lightwave Technol. **23**, 4222–4238 (2005).

19. S. J. Koester, J. Schaub, G. Dehlinger, and J. O. Chu, "Germanium-on-SOI Infrared Detectors for Integrated Photonic Applications," IEEE J. Sel. Top. Quantum Electron. **12**, 1489–1502 (2006).

20. S. Janz, P. Cheben, D. Dalacu, A. Delge, A. Densmore, B. Lamontagne, M.-J. Picard, E. Post, J. Schmid, H. Waldron, D.-X. X. Yap, and W. N. Ye, "Microphotonic Elements for integration on the Silicon-on-Insulator Waveguide Platform," IEEE J. Sel..Top. Quantum Electron. **12**, 1402–1415 (2006).

21. K.-Y. Kim, J. H. Song, J. Lee, S. Y. Kim, J. Cho, Y. S. Lee, D. Hand, S. Jung, and Y. Oh, "Reduction of Insertion Loss of Thin Film Filters Embedded in PLC Platforms," IEEE Photon. Technol. Lett. **17**, 1041–1135 (2006).

22. R. B. Sargent, "Recent advances in thin film filters," in *Proceedings of Optical Fiber Communication Conference*, Technical Digest (OFC) (Optical Society of America, 2004), paper TuD6.

23. R. S. Tucker, K. Pei-Cheng, and C. J. Chang-Hasnain, "Slow-Light Optical Buffers: Capabilities and Fundamental Limitations," J. Lightwave Technol. **23**, 3046–4066 (2005).

24. N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power," Computer **36**, 68-75(2003).

1. Introduction

Parallel computing environments require scalable interconnect solutions that provide the necessary bandwidth and latency to meet the computational needs of the system while maintaining manageable power dissipation figures. The recent emergence of multi-core architectures and chip multiprocessors (CMPs) for driving performance via increases in the number of parallel computational cores has accelerated the need for high bandwidth interconnect solutions in high-performance processors [1-3]. Given the vastly growing numbers of cores, on-chip and, most critically, off-chip communications has become the key bottleneck limiting the performance of parallel computing systems.

Optical interconnects offer a potentially disruptive technology solution by directly addressing the bandwidth, latency, and power limitations of electronic interconnects [4]. Parallel optical links for board-level inter-chip optical communication and inter-board communication have been recently demonstrated to offer impressively high data throughput [5-8]. However, as off-chip bandwidth demands of CMPs continue to accelerate (a current generation IBM Cell processor can require up to 50.6 GB/s of throughput [3]), the power dissipation associated with multiple parallel electro/optical signal conversions grows rapidly.

In this work, we propose a scalable and transparent network interface designed to address the issue of power dissipation by uniquely exploiting the parallelism and capacity of wavelength-division multiplexing (WDM) [9,10]. The proposed photonic network interface extends the capabilities of serial electronic protocols in high-performance computing systems by offering full bandwidth deployment in intra- and inter-chip optical interconnects by directly mapping serial streams onto multiple WDM channels in a highly power efficient

manner. The optical interface exhibits significant power savings by decoupling power dissipation from bandwidth utilization via the use of only one optical modulator, one broadband gate, and one optical receiver regardless of the number of WDM channels employed. The photonic network interface is envisioned to be embedded within the processor boards and other shared high throughput source nodes (Fig. 1). The demonstrated interface translates serial 40 Byte PCI Express (PCIe) encoded electronic packets onto an 8-channel WDM optical link with a measured power penalty of 1.5 dB and a 70 % improvement in power efficiency. We experimentally demonstrate the end-to-end generation of a PCIe link originating from a remote endpoint across the photonic network interface to a host computer in a transparent manner. The remote endpoint is built on a field programmable gate array (FPGA) based device.
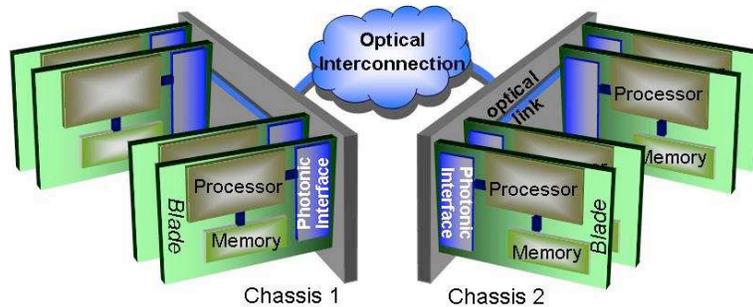

Fig. 1. Transparent optical photonic network interface for high throughput communications infrastructure.

The paper is outlined as follows. In the next section (Section 2), we discuss optical interconnects, their current limitations due to interface power dissipation, and the method in which the proposed photonic interface decouples power dissipation from bandwidth utilization. In Section 3, the experimental implementation of the transparent interface is described and performance results are presented. We follow with a discussion in Section 4 on the importance of timing, the effect of group velocity dispersion on the interface scalability, and the possibility of integration onto a silicon-based platform for intra-chip interconnects. We conclude with a summary of our findings.

2. Optical interconnects

2.1 Background

Due to fundamental physical limitations, gains in computational performance are suffering from diminishing returns via traditional processor scaling techniques. Multi-core architectures and CMPs have emerged as an industry-wide means to continue the trend toward improved computational throughput while maintaining optimal performance-per-watt characteristics. Consequently, this trend toward the multiplication of processor cores is further accelerating the need for scalable high-performance power-efficient interconnect solutions in parallel computing environments, such as cluster computers and data centers, which are already, by nature, communication-limited.

The performance of current electronic interconnect solutions are heavily limited by physical impairments characteristic of transmission lines. Specifically, the parasitic inductance and capacitance of electrical traces and losses due to the skin effect limit both the bandwidth and propagation distance of electrical signals. These limitations can be somewhat alleviated by increasing the signal power, which further requires the use of power-hungry equalization techniques. However, power dissipation has become one of the most critical figures of merit for large-scale parallel computing systems. As such, it is clear that current electronic techniques will not scale to support future multi-core high-performance computing clusters. Furthermore, as the critical performance bottleneck shifts from the processors to the

interconnect, it becomes particularly challenging to meet the bandwidth demands and power restriction of these high-performance systems.

Optical interconnects have been proposed as a potentially attractive solution to alleviate the bandwidth and power limitations challenging copper interconnect technologies. Parallel optical interconnects are being deployed as a cost effective solution in existing high-performance computers [8,11]. Serial data communication is being replaced by parallel optical interfaces where the data is simultaneously transmitted and captured over multiple fibers in a single link. Unfortunately, as bandwidth demands increase, the scalability of this method becomes power-limited. Each optical channel requires its own set of optoelectronic components for the electrical-to-optical (EO) and optical-to-electrical (OE) signal conversions. A recent demonstration of an optical interconnect consisting of an EO signal conversion, 200 meters of optical fiber, and an OE signal conversion exhibited 15.6 mW/Gb/s per link of power dissipation [6]. Extrapolating this figure to a 10 Tb/s system, the theoretical power dissipation reaches 156 W. This figure does not take into account any electrical processing functions such as data aggregation, clock recovery, or temperature control. In CMOS technology [12], the total power dissipated ($P_{tot,elec}$) by the components for signal conversion corresponds to the number of optical channels (N), multiplied by the sum of the static power and the dynamic power as expressed in Eq. (1).

$$P_{tot,elec} = N\left(P_{static} + CV_{DD}^2 Af_S\right) \qquad (1)$$

Leakage current in digital logic (e.g. drivers) and bias current in analog circuitry (e.g. amplifiers) associated with the optical modulator and receiver contribute to the overall static power dissipation of the interface. The dynamic power is proportional to the gate capacitance (C) and the voltage supply (V_{DD}), as well as the average switching frequency (Af_s). The switching frequency (f_s) is multiplied by the activity factor ($0 \leq A \leq 1$), a constant reflecting the average switching activity. The power consumption of the EO and OE conversions increases with both the number of optical channels and data rate. As such, even an optimized design will exhibit high power consumption. As explained in the following section, the proposed photonic network interface decouples the total power dissipated from the number of channels.

2.2 High-bandwidth, low-power photonic network interface

Current optical interconnect solutions are implemented as one-to-one replacements of copper wire connections. As a result, scalability is limited by the aforementioned bandwidth-coupled power dissipation. Conventionally, a serial data stream is electronically mapped onto multiple optical wavelengths using one modulator per wavelength. In the proposed novel photonic network interface, the serial stream is directly mapped onto multiple optical wavelengths by simultaneously modulating all wavelengths using a single optical modulator. Consequently, each wavelength carries a copy of the same data stream as generated by the processor. A bank of filters subsequently isolates each modulated wavelength. Using optical delay lines, each wavelength is then individually delayed in time with respect to its adjacent channels by an amount corresponding to the packet size as defined by the associated optical interconnect fabric (Fig. 2).

The delayed wavelengths are then multiplexed, and gated appropriately to meet the WDM packet structure of an optical interconnection network [14]. The resulting WDM packets consist of periodic intervals of payload data, which are time-compressed by a factor proportional to the number of WDM channels. The resultant dead-time between WDM packets can be interleaved with packets generated from other lanes in a time-division multiplexed manner to maximize link utilization. At the destination, the packet repartitioning is performed all-optically by filtering and delaying each wavelength with respect to its adjacent wavelength in a manner complementary to that performed at the source node. The channels are then multiplexed prior to a broadband receiver enabling the reconstruction of the original electronic serial stream.
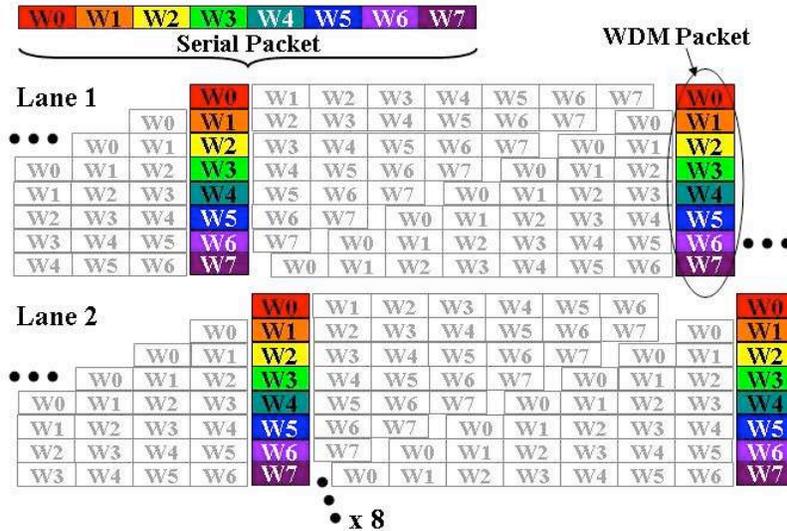
Fig. 2. Serial electronic packets mapped onto 8 WDM channels (W0 to W7) and TDM interleaved by up to 8 lanes to maximize link utilization.

The use of passive optical components for partitioning the serial data stream allows for very low power consumption while taking advantage of the format and data rate transparency of the optical fiber as a low loss medium. Additionally, only one modulator-receiver pair is needed regardless of the number of wavelengths used. Therefore, the power dissipation is decoupled from the bandwidth capacity of the interface, allowing for tremendous scalability. The static power dissipation is significantly reduced by a factor corresponding to the number of optical channel used (N). However, since the activity factor is scaled by the number of channels N, the dynamic power remains the same as in the standard parallel optical interface. As transistor feature sizes continue to scale in the nanometer regime, the overall dynamic power dissipation remains relatively constant while static power dissipation arising from leakage currents increases exponentially. Therefore, static power dissipation is expected to be the dominant contributor to power consumption figures in future electronic designs [24]. While modulators themselves draw a nominal amount of power for DC biasing, it is expected that the majority of the power dissipated to be attributed to the modulator driver circuitry, which will be heavily affected by the trend towards increased static power consumption. The ratio of the total power dissipation for the photonic network interface ($P_{tot,WDM}$) as compared with the standard parallel interface ($P_{tot,elec}$) is given by:

$$\frac{P_{tot,WDM}}{P_{tot,elec}} = \frac{P_{static} + N \cdot CV_{DD}^2 f_s}{N \cdot P_{static} + N \cdot CV_{DD}^2 f_s} \tag{2}$$

Thus, by decoupling the power dissipation from the bandwidth utilization, the novel design enables a 10-100x reduction in power dissipation compared to traditional parallel methods.

3. Experimental demonstration

3.1 PCI Express link

For successful integration within current state-of-the art computing systems, the advantages offered by photonic interconnection solutions must be leveraged in a way that is complementary to existing electronic standards. Several data communications standards, such as PCI Express (PCIe), HyperTransport and RapidIO, have emerged based on high-bandwidth serial architectures enabling communication among shared source ICs. The differences between each protocol are the tradeoffs between flexibility and extensibility versus latency

and overhead. PCIe, now in its third generation, has emerged as the I/O protocol of choice for high-speed serial buses supporting chip-to-chip and board-to-board applications in modern computing systems. As a result, the experimental interface is demonstrated to support for the PCIe protocol.

The typical PCIe network consists of four main components connected in a tree topology [13]. The root complex allows the CPU to control the PCIe network and is connected via a switch to endpoints, which are any peripheral devices connected to the PCIe network. The switch converts a single link into multiple links and the link itself is a point-to-point serial connection between the root complex, switches, and endpoints. Any link is comprised of 1, 2, 4, 8, or 16 lanes, where each lane denotes two differential transmitter-receiver (Tx-Rx) pairs: one to send data upstream towards the root complex, and one to send data downstream away from the root complex. PCIe uses a packetized and layered protocol structure and uses 8b/10b encoding with sufficient transitions in the data to maintain proper clock recovery. PCIe packets may contain as little as 4 bytes to as many as 4096 bytes, which include the link layer and transaction layer information. The lane data rate for PCIe version 1.0 is 2.5 Gb/s and can scale up to a width of 32 lanes. PCIe 2.0 runs at a lane rate of 5.0 Gb/s up to 16 lanes. The third generation will increase the transmission rate to 8.0 Gb/s and remove 8b/10b encoding, doubling the throughput per lane from PCIe 2.0. As data rates increase, the photonic network interface has sufficient capacity to support future PCIe generations due to its scalable bandwidth capabilities and its transparency to data rates.

3.2 Low-power photonic interface implementation

The demonstrated photonic network interface performs the mapping of a 40 byte-long PCIe stream over eight WDM channels. In Fig. 3, the interface on the transmission side is shown along with the receiver side. Eight cooled distributed-feedback lasers are emitting at the ITU channels (labeled W0 to W7). To minimize the effect of chromatic dispersion, the WDM channels are closely spaced by 0.8 nm (100 GHz) from 1543.73 nm to 1549.32 nm, corresponding to ITU channels C35 to C42. The channels are multiplexed onto a single fiber with an 8:1 multiplexer and modulated simultaneously with a 10 Gbps lithium niobate amplitude modulator. The extinction ratio of the optical modulator is approximately 15 dB and is maximized for each individual wavelength by using polarization controllers prior to the 8:1 multiplexer. In this demonstration, the modulator is driven by an amplified bit sequence, representing one PCIe packet (x1), generated by a 2.5 Gbps pattern generator (ParBert). The PCIe packet bit sequence is a training sequence containing bytes BC, F7, F7, 14, 02, 00 followed by ten bytes of 4A. This represents 160 bits with 8b/10b encoding, ensuring a DC-balanced sequence with sufficient transitions for clock recovery. The packet is then repeated twice for a 320-bit long sequence (40 bytes).
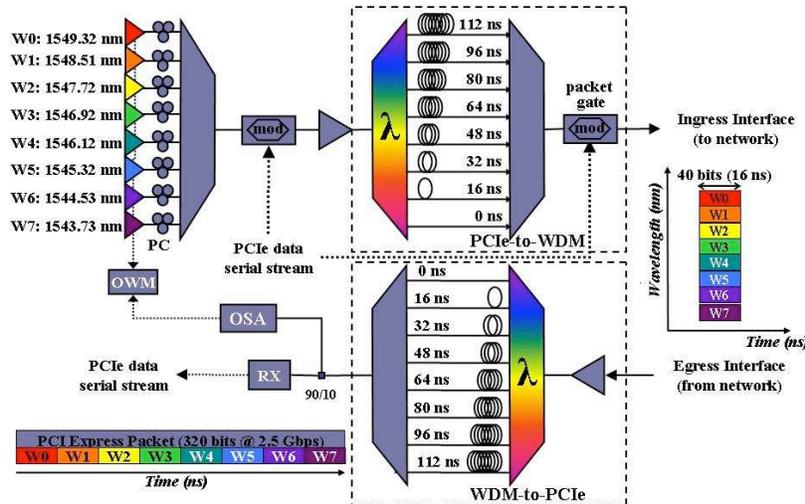
Fig. 3. Detailed schematic of the all-optical PCIe photonic network interface experimental demonstration. The PCIe packet is partitioned onto eight payload channels (W0 to W7) using bandpass filters and optical fiber delay lines in the PCIe-to-WDM block of the interface. The PCIe packet is reconstructed in the WDM-to-PCIe block before being converted to an electrical signal using a broadband receiver (RX). The power and wavelength of each payload channel is monitored and adjusted using an optical wavelength-monitoring (OWM) scheme employing an optical spectrum analyzer (OSA) and controls the cooled DFBs (W0 to W7).

The simultaneously modulated optical wavelengths are amplified by an EDFA and launched into the PCIe-to-WDM block of the interface, grooming the packets according to the physical layer protocol of the associated optical interconnection fabric. The grooming is achieved by means of 100-GHz thin film filtering systems and fiber delay lines. The passive optical devices introduce a total insertion loss of 4 dB. Following the bank of filters and delay lines, each modulated optical signal is subsequently delayed by 16 ns with respect to its adjacent shorter-wavelength channel using precise lengths of fiber. In a 16 ns time window, each wavelength represents a 40-bit segment of the PCIe packet (Fig. 4(a)). An electrical gating signal is used to create the final WDM packet by removing the residual PCIe information outside the segmented payload data. To reconstruct the serial PCIe packet, the WDM interface approach relies heavily on the relative timing between each section of the PCIe packet, which is modulated on separate wavelengths. Hence, a high-speed optical modulator is required to gate the packet with fast transition times to ensure that the first and last bits of each channel in the packet are not truncated. To achieve this, a 13 Gbps amplitude differential driver is used with a dual-drive 10 Gbps lithium niobate modulator characterized by rise and fall times of 30 ps. The gating signal is synchronous with the PCIe packet generator and is configured to take into account the PCIe-to-WDM time-of-flight latency.

At the output of the PCIe-to-WDM block of the interface, the resulting WDM packet contains the first 40 bits of the PCIe packet encoded on the longest wavelength (W0, 1549.32 nm), the following 40 bits encoded on the adjacent wavelength (W1, 1548.51 nm) and so on with the last 40 bits of the 320-bit long PCIe packet encoded on the shortest wavelength (W7, 1543.73 nm). The segmented multi-wavelength data is shown in Fig. 4(a) with each wavelength (W0 to W7) carrying a subset of the PCIe packet.

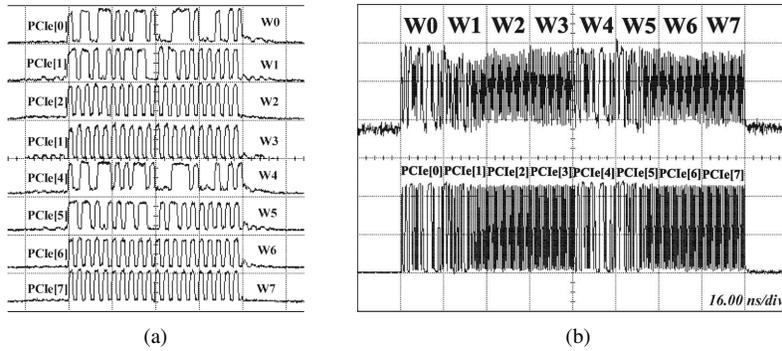| (a)	(b)

Fig. 4. (a) Output packet in the time-domain with each wavelength (W0 to W7) carrying a segment of the PCIe packet. (b) The reconstructed packet at the destination node: (top) optical signal with each packet segment encoded on a different wavelength (W0 to W7); (bottom) electrical PCIe data stream with each PCIe segment annotated.

After propagation through the optical interconnect, the packet is first amplified by an EDFA prior to reconstruction at the WDM-to-PCIe block (Fig. 3). The eight optical channels representing the PCIe packet are individually filtered using eight 100 GHz thin film optical bandpass filters. The delaying process is similar to the approach used for the WDM packet generation in the PCIe-to-WDM section, but in a reverse manner. Each channel is delayed by 16 ns with respect to the longer adjacent channel in a FDL system with the shortest channel at 1543.73 nm (W7) experiencing the most delay. The wavelengths are finally multiplexed into a single fiber representing the original serial PCIe packet of 320 bits. The reconstructed data is converted to an electrical signal using a 10.7 Gbps DC-coupled broadband optical receiver module with an integrated limiting amplifier. Small optical power differences between each channel due to the polarization dependence of the components and the optical modulators can be seen in Fig. 4(b) (top signal). The limiting amplifier of the optical receiver alleviates the small optical power difference by quantifying the optical signal to a 450 mV$_{p-p}$ signal (Fig. 4(b), bottom signal). Hence, small interchannel differences in optical power do not affect the optical to electrical signal conversion of the PCIe reconstruction process.

A reduction of 73.5% (N=8) in the static power dissipation was achieved when taking the broadband optical modulator into account, excluding the EDFAs and the modulator gating the packet. In Fig. 5, the power penalty of the transparent WDM interface was measured to be 1.5 dB (BER < 10^{-12}) with respect to a back-to-back link where the entire electronic packet is serially encoded onto one WDM channel (W0). The source of the power penalty can be attributed to the variations in power and extinction ratio across the WDM channels under a fixed receiver threshold voltage.
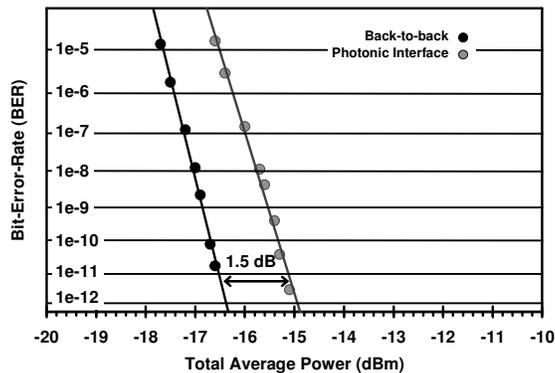

Fig. 5. Measured power penalty of the photonic interface.

The optical eye diagram at the output of the ingress interface is shown in Fig. 8 as well as the reconstructed electrical eye diagram at the output of the egress interface. A training

sequences originating from the host is used as input to the interface. Unfortunately, the relatively high phase jitter of the original signal (attributable to the low quality of consumer devices) necessitated the need to trigger on the recovered signal itself, thus leading to degradation in the optical eye measured at the ingress block.
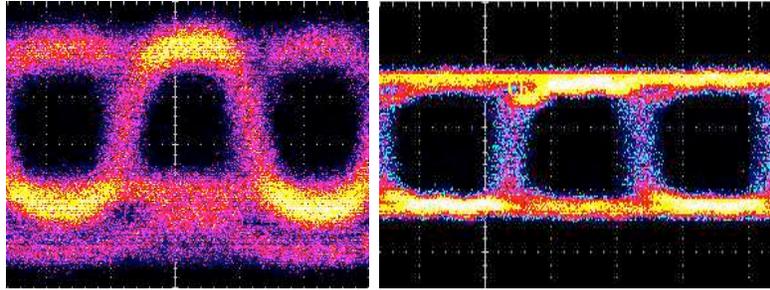

Fig. 8. (a) Optical eye diagram of PCIe training sequence at 1544.53 nm at the output of the egress node and (b) electrical eye diagram at destination.

3.3 Complete PCI Express data transfer

In order to validate the functionality and transparency of the proposed system in a real-world implementation, we insert the photonic network interface between a PCIe data-generating host and a remote endpoint. A dual-core desktop (x86) serves as the host computer system. As schematically shown in Fig. 7, the host provides a PCIe x16 graphics slot connected to its northbridge and a PCIe x1 slot connected to its southbridge. A Samtec PCIe x1 to SMA adapter is used to gain access to the gigabit Tx and Rx signals at the host. The remote host is implemented on an XpressGXII FPGA evaluation board manufactured by PLD Applications, Inc. (PLDA). The XpressGXII employs an Altera Stratix II GX-class FPGA featuring sixteen gigabit transceivers. Eight transceivers are connected to the PCIe x8 card edge, which is not in use in the setup, while another eight are connected to a daughter card duplicating a PCIe x4 card edge connection. A Samtec PCIe x4 to SMA adapter is used to access the FPGA's Tx and Rx pairs through the daughter card.
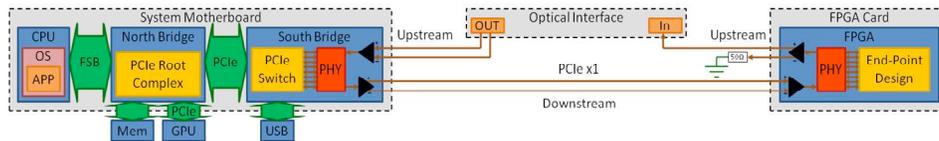

Fig. 7. Schematic representation of host and endpoint system organization as experimentally implemented. The optical interface is inserted inline with the upstream link.

The upstream link is connected single-endedly to the high-speed modulator of the ingress interface. The broadband receiver, which is connected to the output of the egress interface and the input of the gigabit receiver at the host computer, digitizes the recovered PCIe stream as described above and translates the signal to the appropriate voltage levels as specified by the PCIe physical layer protocol. The downstream link is maintained in the electronic domain to achieve a complete dual-simplex PCIe link implementation. The initialization, training, and configuration of the link are performed successfully by the host computer and the endpoint device is correctly recognized across the photonic interface, confirming the transparency of the interface gateway to the PCIe link.

4. Discussion

4.1 Timing jitter

Due to the all-optical method employed in recovering the serial data from wavelength-parallel packets, timing alignment between the data in each WDM channel is critical for error-free reconstruction of the original PCIe stream. In WDM systems, chromatic dispersion is a major contributor to timing misalignments and relative skew. Therefore, the overall scalability will be limited by the timing skew tolerance of the receiver at the destination node. To investigate the tolerance of the WDM interface to timing misalignment, the experimental setup is modified to allow each wavelength to be individually modulated and delayed. In Fig. 8, three possible cases are investigated by artificially 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 PCIe[2] data, (2) no significant skew, and (3) large timing skew in PCIe[3] data with respect to PCIe[2]. The analyzer output data of the PCIe packet is shown for all three cases. As shown in Fig. 8, 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 $\pm\, 0.2$ UI with data rate at 2.5 Gbps.
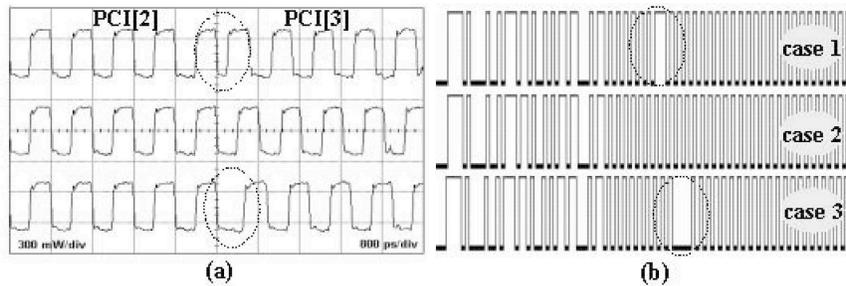
Fig. 8. (a) Receiver output signal showing the timing alignment between PCIe[2] and PCIe[3] data for the three possible cases. (b) Corresponding digitized output signal of the analyzer for all three cases, exhibiting an error in the case of excessive timing skew.

Timing skew between each segment translates into an overall phase jitter in the converted electrical signal representing the entire PCIe stream. The PCIe 1.0 standard specification has a jitter budget that allocates 0.3 UI of total jitter (120 ps) for the media [13]. Given this constraint, we investigate the maximum supported network size given the number of WDM channels. For an optical interconnection with a broadband gain bandwidth of 40 nm, the maximum number of channels supported by the network for a spacing of 0.8 nm and 0.4 nm are 50 and 100 channels, respectively. Dispersive effects are calculated for a network based on SMF-28 single mode fiber (D=17 ps/(nm-km)) and an operating frequency centered at 1550 nm. Fig. 9 shows the relationship between the maximum fiber length supported and the number of channels given the dispersion tolerance of the photonic network interface. To support the longest PCIe packet of 4096 bytes, the necessary 63 wavelengths can be spaced by 0.4 nm. For an integrated photonic network, the GVD in a Silicon-on-Insulator (SOI) waveguide structure has been measured to be 4400 ps/(nm-km) [15]. For a less constrained channel spacing of 1.6 nm, the network size is limited by thermal shift to a waveguide length of 14 cm, which is sufficiently long for intra and inter-chip network applications.
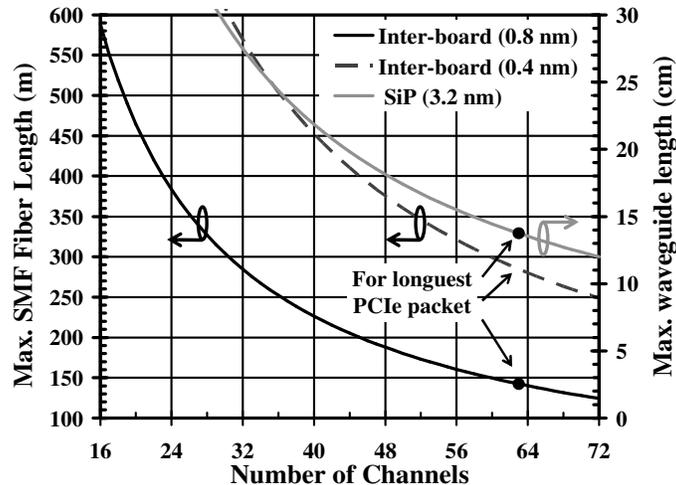
Fig. 9. Maximum interconnection length versus the number of wavelength channels based on chromatic dispersion for 0.8 nm (solid black line), 0.4 nm (dashed line) channel spacing and for a Silicon Photonic (SiP) waveguide (light gray).

4.2 Photonic integrated interface for CMPs

Recent developments in silicon photonics [16,17] have given the prospect of photonic integration credence as a viable technology platform. Advances in integrated optical devices, such as silicon-based ring resonator modulators, integrated photodetectors, and low loss waveguides, allow us to envision the photonic network interface in a silicon-based platform [15,18,19]. Moreover, compatibility with CMOS SOI fabrication allows for a promising platform for integrated photonic networks for multi-core architectures [20]. Low-loss thin film filters can be embedded in a PLC platform and are capable of low insertion loss when used to build multiplexer and demultiplexer modules configured according to DWDM ITU channel spacings [21,22]. Due to the inter- and intra-chip nature of the applications supporting PCIe as a serial communication protocol (ie. multi-core processors), photonic integration is essential towards achieving optimal performance from the interface. One of the main challenges to overcome in the integration of the interface is in manipulating longer PCIe packets without electronic buffering. Delaying 32 kb (4096 bytes) of data becomes difficult in waveguides, but recent advances in slow light, which is currently capable of buffering up to 235 kb [23], show a promising direction.

4.3 Interface bandwidth scalability

Although the proposed interface allows for significant bandwidth scalability, limitations arise from the restrictions introduced by the physical layer of the optical interconnection network. For a bandwidth of 40 nm, a maximum of 50 channels can be used on the 100 GHz ITU-grid. Therefore, the maximum number of bytes that can be mapped to a WDM packet structure of 20.9 ns is 650 bytes for PCIe 2.0 (5.0 Gb/s). Fig. 10 illustrates the number of wavelengths required with respect to the PCIe packet length. To effectively handle the longest PCIe packets at the maximum lane width, space division multiplexing and longer packets must be considered. To accommodate the maximum PCIe packet length of 4096 bytes on the maximum PCIe width (x16), the packet length can be increased to 209 ns and use 63 wavelength channels over 4 parallel waveguides. Hence, there is a tradeoff between the packet lengths versus the optical bandwidth of the optical interconnect fabric design, but future PCIe generations offer increased data rates, allowing for shorter packets.
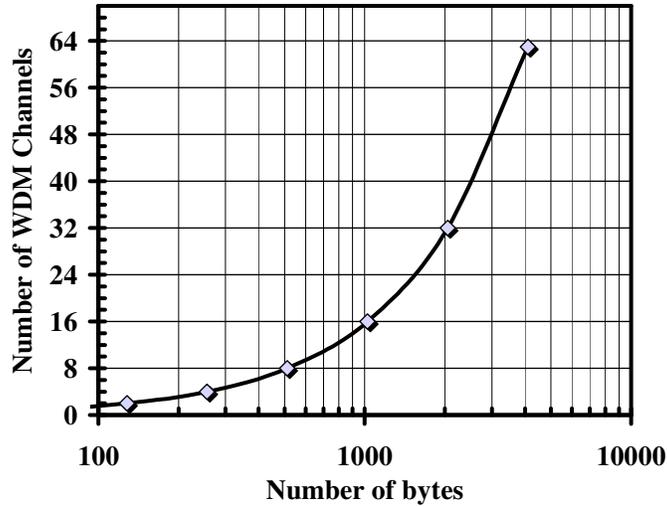
Fig. 10. Number of WDM channels required to map a x16 PCIe packet using space division (4 parallel waveguides).

5. Conclusion

We have shown an innovative approach to alleviating the bottleneck introduced at the off-chip interface by proposing a novel transparent photonic network interface implementation. PCIe packets are mapped onto the WDM packet structure and optically reconstructed in a low-power, low-latency manner. A single modulator/receiver pair is used regardless of the number of wavelengths, thus decoupling the power dissipation from the bandwidth capacity of the interconnection network. A reduction of 73.5% in power dissipation was achieved and a complete PCIe data transfer link was demonstrated. In addition to its characteristically power dissipation, we believe that the scalability and transparency of the photonic network interface will elegantly address the growing demands of off-chip bandwidth in high-performance advanced computing systems and data centers.