Flexible-Bandwidth Power-Aware Optical Interconnects with Source Synchronous Technique

Roberto Proietti, Christopher J. Nitta, Zheng Cao, Matthew Clements, Georgios Tzimpragos, S.J.B. Yoo
Department of Electrical Engineering, University of California, Davis, One shields avenue, Davis, CA, 95616, USA
rproietti@ucdavis.edu, sbyoo@ucdavis.edu

Abstract: This paper investigates energy savings in data centers with flexible-bandwidth power-aware source-synchronous optical interconnects. Network simulations show ≥ 5× energy savings. Link experiment shows error-free operation from 625 Mb/s to 10 Gb/s.

OCIS codes: (200.4650) Optical Interconnects; (200.6715) Switching.

1. Introduction

Today's warehouse-scale computing systems (data centers) are facing challenges in processing exponentially increasing data due to limitations in power consumptions. While optical interconnects can support scalable bandwidth interconnection independently of communication distance with low energy requirements at ~ 1 pJ/b, the power consumption from communications alone can exceed 100 kW in large data centers today. A significant part of power inefficiency in communication comes from poorly adapting to the communication and traffic patterns, which are known to be bursty [1] with high peak-to-average ratios (typically > 10). Hence, most of the time, conventional communication systems are consuming energy even when no meaningful bits need to be transported. On the other hand, we can design an energy efficient communication system by employing flexible-bandwidth optical communication. The dynamic power of CMOS transistors in the transceivers scales as $\propto V_{dd}^2 f$, where V_{dd} is the drive voltage and f is the clock speed. Since V_{dd} can be lowered for circuits with low f, there can be a significant improvement in energy efficiency when the clock speed can be lowered in combination with the dynamic voltage scaling (DVS) technique [2] (nearly 2× improvements in power efficiency for 20% underclocking). There has been research on power-aware interconnects that utilize frequency scaling and DVS techniques [2-6] relying on standard RX architectures with clock and data recovery (CDR) with limitations due to the range of clock frequencies and resynchronization latency. We propose and study flexible-bandwidth power-aware (FB-PA) optical interconnects exploiting a ***source synchronous*** technique [7] where clock is transmitted together with the data for receivers without CDR. We design an accurate power and energy model for the FB-PA source synchronous link and investigate the energy saving for a fat-tree network. Finally we report a proof-of-concept link experiment demonstration with error-free performance over a range 625 Mb/s - 10 Gb/s.

2. Source Synchronous Link Power with Frequency and Voltage Scaling

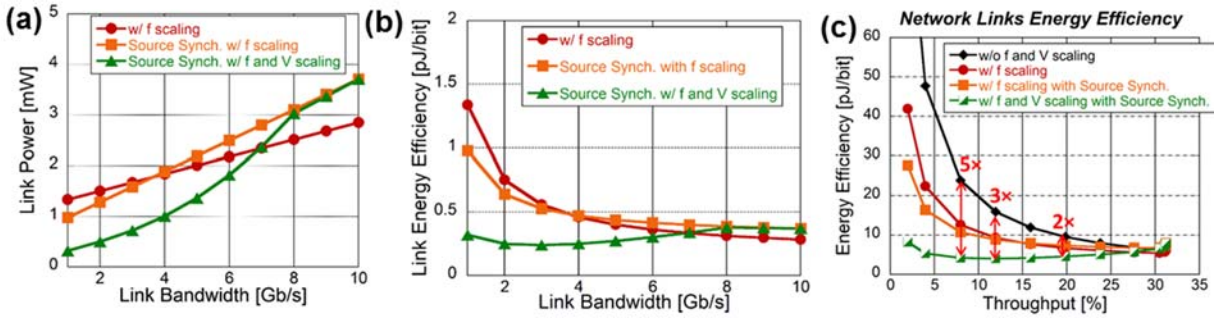

Figure 1. Link power consumption (a) and energy efficiency (b) as function of link bandwidth. Red curve is with frequency scaling only and standard CDR scheme up to 10 Gb/s. Orange curve is with frequency scaling only and source synchronous technique up to 10 Gb/s. Dark green curve is with adaptive frequency and voltage scaling with source synchrnous technique. Included in the electrical model were SERDES, TIAs, modulators, and required support circuit such as clocking network. Transistor leakage power and photonic power were also included in the power link simulations. Transistor and wire technology data were obtained from ITRS 2011 [8] with a target year of 2016 (corresponding roughly to a 16 nm technology node). Note that below 2 Gb/s static power starts to dominate as the Vdd can only marginally be reduced as it is approaching near threshold values. (c) Network links energy efficiency by using the frequency-voltage scaling techniques in a fat-tree network with 3:1 oversubscription.

Figure 1(a) and (b) show the simulated link power consumption and energy efficiency for a source synchronous link with voltage and frequency scaling. We observe significant savings for link bandwidth below 6 Gb/s, with the source synchronous paying only a small penalty beyond ~ 7 Gb/s due to the fact that it requires one additional TX and RX for the optical clock distribution. Figure 1**(c)** shows the overall network links energy efficiency in terms of picojoules (pJ) per bit of throughput for a two-tier Fat Tree network interconnecting 1296 servers running up to 10 Gb/s. The

oversubscription ratio is 3:1 and the network makes use of 24-port switches at the first level and 12-port switches at the higher level. Traffic is uniform random and packet size is 256 Byte. By using the proposed source synchronous technique with voltage and frequency scaling (dark green curve), we can achieve an energy efficiency improvement of ~ 2×, 3×, and 5.4× for ~ 20%, 12% and 8% throughput values, respectively.

4. Experimental Setup and Results

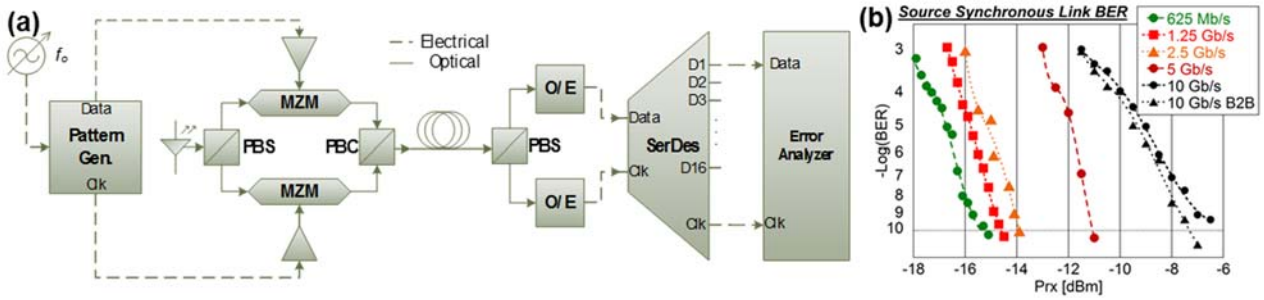

Figure 2. (a) Experiment setup of source synchronous variable vandwidth optical link. PBS: polarization beam splitter; PBC:polarization beam combiner; MZM: Mach Zehnder modulator; SerDes: serializer/deserializer that operates as serial to parallel (S/P) converter. (b) BER measurements for source synchronous link at 625 Mb/s, 1.25Gb/s, 2.5 Gb/s, 5 Gb/s and 10Gb/s after 1:16 S/P.

Figure 3(a) shows the proof-of concept experiment setup for source synchronous link with frequency scaling from 10 Gb/s down to 625 Mb/s. This experiment makes use of polarization multiplexing to carry clock and data over the same fiber. At the RX, a PBS separates the clock and data signals which are converted to the electrical domain by two 10 GHz photoreceivers with TIA. Electrical data and clock feed data and clock ports of a broadband (DC to 17 GHz) 1:16 serial to parallel (S/P) converter, whose clock and data outputs (running at a frequency 16× slower than the input) connect to an error analyzer. Figure 3(b) shows bit error rate (BER) measurements for the source synchronous link running at 10 Gb/s, 5 Gb/s, 2.5 Gb/s, 1.25 Gb/s, and 625 Mb/s. This wide-range of operation it is possible thanks to the use of source synchronous technique and broadband S/P without clock recovery stage.

5. Conclusion

We studied energy link savings in large scale computing systems by source synchronous optical interconnects with voltage and frequency scaling. We derived a power model and applied it to network simulations for a fat-tree network, showing ≥ 5× improvement at low network utilization. Proof-of-concept link experiment shows error-free operation for the proposed source synchronous technique with frequency scaling over the range 625 Mb/s – 10 Gb/s.

[1] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," presented at the Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, Melbourne, Australia, 2010.

[2] A. K. Kodi and A. Louri, "Power-Aware Bandwidth-Reconfigurable Optical Interconnects for High-Performance Computing (HPC) Systems," in *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*, 2007, pp. 1-10.

[3] C. Xuning, W. Gu-Yeon, and P. Li-Shiuan, "Design of low-power short-distance opto-electronic transceiver front-ends with scalable supply voltages and frequencies," in *Low Power Electronics and Design (ISLPED), 2008 ACM/IEEE International Symposium on*, 2008, pp. 277-282.

[4] P. P. Dash, G. Cowan, and O. Liboiron-Ladouceur, "A variable-bandwidth, power-scalable optical receiver front-end in 65 nm," in *Circuits and Systems (MWSCAS), 2013 IEEE 56th International Midwest Symposium on*, 2013, pp. 717-720.

[5] J. E. Proesel, B. G. Lee, A. V. Rylyakov, C. W. Baks, and C. L. Schow, "Ultra-low-power 10 to 28.5 Gb/s CMOS-driven VCSEL-based optical links [Invited]," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 4, pp. B114-B123, 2012.

[6] X. Chen, L.-S. Peh, G.-Y. Wei, Y.-K. Huang, and P. Prucnal, "Exploring the design space of power-aware opto-electronic networked systems," in *High-Performance Computer Architecture, 2005. HPCA-11. 11th International Symposium on*, 2005, pp. 120-131.

[7] C. Gray, D. Keezer, O. Liboiron-Ladouceur, and K. Bergman, "Multi-Gigahertz Source Synchronous Testing of an Optical Packet Switching Network," in *International Mixed-Signals Test Workshop*, 2006.

[8] C. International Roadmap, "International Technology Roadmap for Semiconductors, 2011 Edition," *Semiconductor Industry Association, http://www. itrs. net/Links/2011ITRS/2011Chapters/2011ExecSum. pdf*, 2011.

Acknowledgements

This work was supported in part under DoD Agreement Number: W911NF-13-1-0090.