INTREPID: Developing Power Efficient Analog Coherent Interconnects to Transform Data Center Networks

Clint L. Schow[1] and Katharine Schmidtke[2]
[1]*Dept. of Electrical and Computer Engineering, the University of California Santa Barbara, Santa Barbara, CA 93106, USA*
[2]*Facebook, Inc., 1 Hacker Way, Menlo Park, CA 94025, USA*
schow@ece.ucsb.edu, kschmidtke@fb.com

Abstract: The INTREPID program is developing power efficient coherent optics for package-level integration with future switch ICs as a path to realizing higher-radix switches for flatter networks while enabling new architectures incorporating optical routing and switching.
OCIS codes: (200.0200) Optics in Computing; (060.2330) Fiber optics communications; (060.4265) Networks, wavelength routing

1. Introduction

Datacenters are now an important part of our nation's technical infrastructure. Server-to-server transfers account for >70% of the global datacenter traffic [1] so maximizing the bandwidth and efficiency of intra-datacenter communications is key to improving overall productivity and efficiency. Today, interconnects are implemented as an inefficient concatenation of chip-to-chip links through multiple electrical 50-Ω interfaces. Electrical chip I/O cells are designed for worst-case physical channels and consume on the order of 50% of the total power for switches.

The INTREPID project, part of the ARPA-E ENLITENED program, is a collaboration between UCSB and Facebook focused on developing a technology platform that integrates efficient high-speed photonic interfaces directly into chip packages. The efficiency targets of the co-packaged optical interfaces are aggressive, scaling to sub-pJ/bit for multi-mode VCSEL-based short-reach server links, and to <10 pJ/bit for single-mode analog coherent datacenter-scale interconnects. Achieving these targets will enable highly integrated solutions for the 51 Tb/s switch generation and beyond that can potentially offer substantial expansions in switch radix with simultaneous improvements in efficiency compared to aggressive projections of conventional module-based transceiver technology. Such large, highly efficient switches can enable flatter networks with higher bandwidth to improve the efficiency of datacenters of all scales.

Our focus is on the integration of photonic I/O with electrical switch cores since the network switches are the points of highest bandwidth concentration and where efficient photonic I/O can have the greatest impact. Fat-tree networks, schematically depicted in Figure 1(a), are the workhorse topology for datacenter networks due to their superior performance and scaling properties that fundamentally depend on switch radix, as shown in Figure 1(b). The Top of Rack (ToR) switches require two types of interconnects: short distance (<3 m) for server connections, and longer distance (<2 km) for connections to switches in the next level of hierarchy. For the longer fabric links above the ToR, the use of single mode (SM) fiber is essentially a requirement due to its substantial advantages in operational management, cost, and support for future bandwidth scaling through wavelength division multiplexing (WDM).

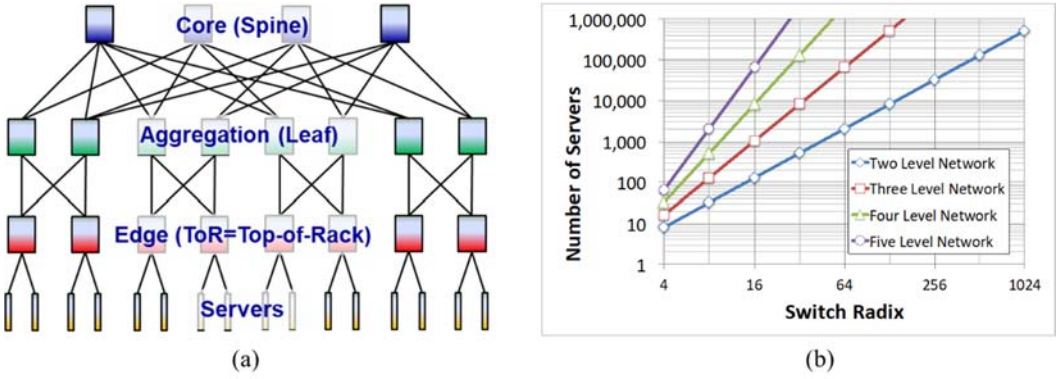

Figure 1. (a) Illustration of a fat tree network and (b) scaling properties: number of connected servers as a function of switch radix.

2. Approach

Our approach is conceptually illustrated in Figure 2. Integrating the photonic interfaces into the chip package means connections to the electronics are at C4 pitch (~130 µm) instead of BGA/LGA pitch (~1 mm), avoiding the primary packaging bottleneck and improving bandwidth density up to ~60X. Electrical paths between the photonics and

electronics are also minimized, potentially enabling a >10X improvement in the efficiency of the electrical links connecting the ASIC to the photonic I/O. In fact, general trends of electrical chip I/O show a direct dependence on channel loss, with 30 dB of channel loss (a typical target for general purpose electrical I/O) degrading efficiency by 10X [2]. A recent report of a 2-cm long (low channel loss) chip-to-chip link demonstrated an efficiency of 1.4 pJ/bit, >14X more efficient compared to typical general purpose I/O cells that consume ~20 pJ/bit [3].

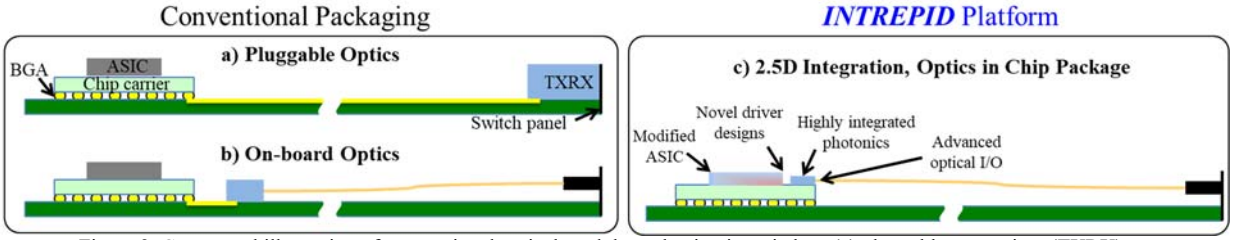

Figure 2. Conceptual illustration of conventional optical module packaging in switches: (a) pluggable transceiver (TXRX) modules, (b) on-board optical modules, and (c) the integrated platform under development.

Figure 3 presents a conceptual view of the modular photonic integration platform we envision applied to a ToR switch, encompassing the co-design of interface circuitry to the digital switch core, the I/O bridge, electronic/photonic interposers, and single-mode (SM) and multimode (MM) photonics with array fiber coupling. Switches for the higher levels of the network will integrate only SM photonics for the reasons discussed above.

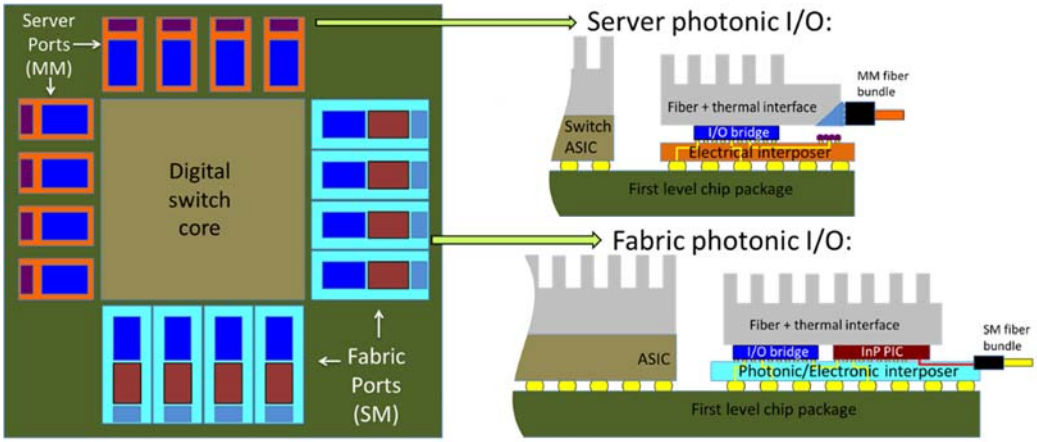

Figure 3. Integration concept for MM and SM optical I/O into a first-level switch chip package.

For server links, VCSEL technology provides a viable path to low-cost, short distance links with sub-pJ/bit efficiency. Past results from our team illustrate the potential of VCSEL links, including demonstrations of some of the highest performance VCSEL devices in the world [4], multi-channel transceivers that set records for efficiency, speed, bandwidth density, and total bandwidth, [5] and a record wall-plug efficiency of 1 pJ/bit at 25Gb/s [6]. The focus in our current effort is to demonstrate similar link efficiencies at 56 Gb/s through a combination of fast and efficient VCSELs and circuit designs that incorporate equalization techniques proven to maximize speed and efficiency [7,8]. The VCSEL links we develop will be applicable to ToR to server links in the rack (<3m) and will also support migration from ToR to end of row (EoR) switches (<30m) as data center architectures evolve.

For interconnects above the ToR/EoR tier, we are developing low-cost, low-power coherent WDM photonic interconnects purpose-built for the longer fabric links required in the datacenter. Tailoring the links to datacenter requirements requires optimization for a different set of metrics compared to current long-haul and metro coherent technology, specifically: 1) low power consumption, 2) expanded link budgets, 3) low cost, and 4) low latency. Future scalability to higher data rates is possible through higher-order modulation formats, polarization modulation, and additional wavelengths. For datacenters, the significantly larger link budget is a key advantage for coherent links, enabling reduced link power (lower required source laser power), lower cost (relaxed alignment tolerances/device specifications), and novel network architectures that incorporate all-optical routing/switching. Expansions of link budget on the order of 20 dB are possible [9], and our analysis shows that link budgets of 13 dB can be achieved with wall-plug link efficiencies better than 10 pJ/bit. This level of tolerance to link loss allows for the incorporation of an AWGR (arrayed waveguide grating router) or active photonic switching layer without requiring complex and costly integrated optical gain in such components. Furthermore, the high selectivity offered by coherent reception significantly reduces the optical crosstalk requirements between channels for photonic routing/switching devices.

The *INTREPID* analog coherent links under development are drastically lower in power and complexity compared to current digital coherent technology that relies heavily on digital signal processing (DSP) to compensate for chromatic dispersion (CD), polarization mode dispersion (PMD), and nonlinear effects in DWDM links. In contrast, the *INTREPID* links will operate in the O-band near the zero-dispersion wavelength for standard single-mode fiber (1264-1338 nm), meaning CD and PMD will not have to be compensated as they contribute negligible performance penalties. Furthermore, to eliminate the need for DSP-based carrier recovery, we are developing optical phase locked loops that lock and track the phase and frequency of the receiver local oscillator (LO) to the incoming signal. Previous work by our team has established the viability of highly-integrated OPLLs to enable robust and high-performance "analog coherent" receivers that operate with very low BER ($< 10^{-12}$ up to 35 Gb/s) and do not rely upon costly, high-power ADCs and DSPs [10]. The analog coherent receivers we are developing are optimized for power efficiency through photonic device and circuit co-design, choice of modulation format (QPSK), and close integration of electronics and photonics to minimize loop delays and maximize noise tolerance. We are targeting 200 Gb/s/λ, achieved through a base symbol rate of 50 Gbd/s, QPSK modulation, and polarization multiplexing to maximize the bandwidth per fiber. A four-wavelength link would therefore offer 800 Gb/s/fiber, with further future scaling possible.

3. Conclusion and Outlook

We aim to simultaneously improve the performance and efficiency of datacenters by developing and demonstrating a platform that directly integrates efficient photonics, specifically designed for the datacenter, into first-level chip packages. To ensure our solutions will be viable in the marketplace, we focus on technologies capable of high-volume, low cost production. By replacing today's power-hungry electrical chip I/O with low power photonic interconnects integrated directly into electronic chip packages, switches with a much larger port count (radix) can be realized compared to today's conventional packaging. *INTREPID* spans both establishing highly integrated electronic/photonic platforms with aggressive bandwidth and energy efficiency targets as well as developing novel, power-optimized network architectures to leverage the efficiency and performance improvements of the underlying link technology. This combination of technology and architectural advancement has the potential to eliminate layers of network hierarchy, directly and significantly reducing the power, latency, and cost of datacenter networks, while improving server utilization to boost the performance and efficiency of the datacenter as a whole.

4. Acknowledgments

The authors greatly appreciate the technical contributions and rewarding collaborations with their colleagues at Facebook (Hans-Juergen Schmidtke, Ariel Hendel, Brian Taylor, Todd Hollmann, Jimmy Williams, Gilad Goldfarb, James Stewart) and UCSB (James Buckwalter, Larry Coldren, Jonathan Klamkin, Adel Saleh, Shamsul Arafin, Shireesh Bhat, Sarvagya Dwivedi, Fabrizio Gambini, Sergio Pinna, Hector Andrade, Takako Hirokawa, Aaron Maharry, Thomas Meissner, Stephen Misak, Luis Valenzuela, Yujie Xia).

Funding under the ARPA-E ENLITENED program is gratefully acknowledged and the authors would like to especially thank the ARPA-E team for management and guidance: Michael Haney, James Zahler, and Alan Liu. The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000848. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

5. References

[1] White Paper: "Cisco global cloud index: forecast and methodology, 2016–2021," available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf, accessed 12/12/2018.

[2] D. C. Daly, *et al*., "Through the looking glass—the 2018 edition: trends in solid-state circuits from the 65th ISSCC" *IEEE Solid-State Circuits Mag.*, vol. 10, pp. 30-36, Winter 2018.

[3] T. O. Dickson *et al*., "A 1.4 pJ/bit, power-scalable 16×12 Gb/s source-synchronous I/O with DFE receiver in 32 nm SOI CMOS technology," in *IEEE J. Solid-State Circuits*, vol. 50, no. 8, pp. 1917-1931, Aug. 2015.

[4] Y.C. Chang and L.A. Coldren, "Efficient, high-data-rate, tapered oxide-aperture vertical-cavity surface-emitting lasers," *IEEE J. Sel. Topics in Quantum Electron.,* 15, (3), pp. 704-715, May/June 2009.

[5] F. E. Doany, B. G. Lee, D. M. Kuchta, A. V. Rylyakov, C. Baks, C. Jahnes F. Libsch, C. L. Schow," Terabit/Sec VCSEL-based 48-channel optical module based on holey CMOS transceiver IC," *IEEE J. of Lightw. Technol.*, vol. 31, no. 4, pp.672-680, Feb. 2013.

[6] J. E. Proesel, B. G. Lee, C. W. Baks, and C. L. Schow, "35-Gb/s VCSEL-based optical link using 32-nm SOI CMOS circuits," *OFC 2013*, paper OM2H2, (2013).

[7] A.V. Rylyakov, C. L. Schow, B. G. Lee, F. E. Doany, C. Baks, and J. Kash, "Transmitter pre-distortion for simultaneous improvements in bit-rate, sensitivity, jitter, and power efficiency in 20 Gb/s CMOS-driven VCSEL links," *IEEE J. of Lightw. Technol.*, vol.30, pp.399-405, Feb. 2012.

[8] D. Kuchta, A. Rylyakov, F. Doany, C. L. Schow, J. Proesel, C. Baks, P. Westbergh, J. Gustavsson, A. Larsson, "A 71 Gb/s NRZ modulated 850 nm VCSEL-based optical link," *IEEE Photonics Technol. Lett.*, vol. 27, no. 6, pp 577-580, Mar. 2015.

[9] J. K. Perin, A. Shastri and J. M. Kahn, "Design of Low-Power DSP-Free Coherent Receivers for Data Center Links," in *J.Lightw. Technol.,* vol. 35, no. 21, pp. 4650-4662, Nov., 2017.

[10] M. Lu *et al*., "An Integrated 40 Gbit/s Optical Costas Receiver," *J. Lightw. Technol.,* vol. 31, no. 13, pp. 2244-2253, July, 2013.