Scaling Optical Interconnects in Datacenter Networks

Opportunities and Challenges for WDM

Hong Liu, Cedric F. Lam, and Chris Johnson
Google Inc.
Mountain View, CA
hongliu@google.com, clam@google.com, cjjohnson@google.com

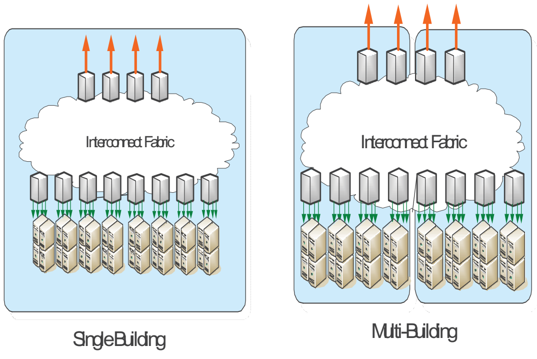
ABSTRACT

We review the growing need for optical interconnect bandwidth in datacenter networks, and the opportunities and challenges for wavelength division multiplexing (WDM) to sustain the “last 2km” bandwidth growth inside datacenter networks.

INTRODUCTION

The rapid growth of Internet and cloud computing applications has resulted in datacenter network bandwidth requirements that outpace Moore’s Law. When most datacenter applications are provided free of charge, datacenter operators are faced with the challenge of meeting exponentially increasing demands for network bandwidth without exorbitant increases in infrastructure cost and power.

Nowadays, a datacenter typically contains tens of thousands of servers that form a massively parallel super-computing infrastructure [1], [2]. Figure 1 shows a typical datacenter cluster, with servers arranged in racks of 20-40 machines each. Servers within a rack are connected to a top-of-rack (TOR) switch, which is in turn connected to layers of cluster switches. These cluster switches provide connectivity between the racks and form cluster-fabrics for warehouse-scale computing. Notice that a datacenter can consist of either one building or multiple buildings, as shown in Figure 1.



(a) (b)

Figure 1. Hierarchies of cluster/switch interconnection fabrics for (a) single-building datacenter, and (b) multi-building datacenter.

Fiber optic technologies play critical roles in datacenter operations. Optical interconnections, with reach between 10m to 2km, are of paramount importance for intra-datacenter connectivity in a warehouse-scale computer. For 10Gb/s data rates and beyond, ultra-low-cost and power-efficient active optical cables, such as Light Peak Modules [3], will soon replace copper cables in the final 10m that connect servers to the TOR switch. Within the same building, vertical cavity surface emitting laser (VCSEL)-based low-power and low-cost short reach (SR) multi-mode optics are already playing an important role. Long reach (LR) optics, at 1310nm wavelength and based on higher-power single-mode distributed feedback (DFB) lasers, are now being used for inter-building interconnections.

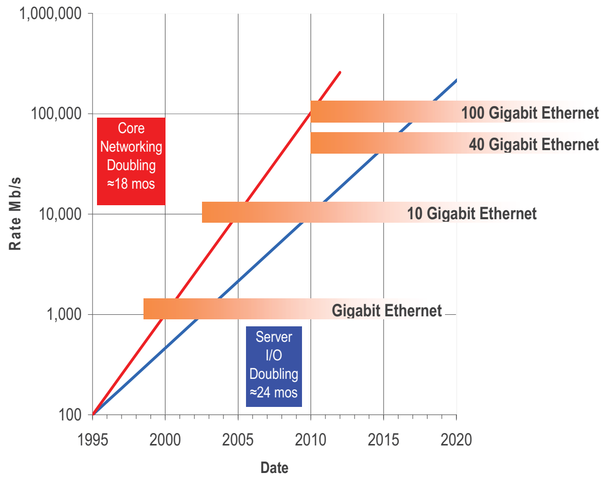


Figure 2. Bandwidth Trend for networking and server I/O [4].

To meet the exponential growth of bandwidth, mega-datacenter operators are bundling multiple 10GE links in parallel via link aggregations (LAG) [5]. However this approach not only is complex to configure but can also cause load imbalance across multiple links in a LAG group, making it difficult to scale. Pooling multiple server data streams into a LAG group can improve utilization, but can also destroy the natural parallelism of the data streams and introduce performance latency. Interconnection bandwidth must scale

978-0-7695-4208-9/10 $26.00 © 2010 IEEE
DOI 10.1109/HOTI.2010.15

from 10Gb/s through 40Gb/s to 100Gb/s, while the server interconnect at the bottom of datacenter networks is moving up to 10Gb/s (Figure 2).

Scaling Interconnect Bandwidth

In addition to bandwidth growth, the power, cost, and space density must scale at the same time to satisfy the needs of mega-datacenter computing.

Given the huge optical interconnect bandwidth needed, it is unlikely that a single stream of data will meet the growth requirement. Instead, time division multiplexing (TDM), spacing division multiplexing (SDM) or wavelength division multiplexing (WDM) will be needed to tap the vast bandwidth of optical fiber.

TDM: To maximize off-chip bandwidth and power efficiency, the I/O data rate will increase with new CMOS technology nodes. Nevertheless, the aggregated data rate can't be scaled arbitrarily high, but is limited by physical impairments in both electrical and optical domains. As a result, TDM faces various scalability challenges.

From cost and power perspectives, the most efficient way to scale per-lane optical data rate is to align the state-of-the-art electrical signal rate with the optical signal data rate. Thus, no expensive and power hungry SerDes or gear box is needed for data-rate conversion between electrical and optical lanes. Moreover, the natural parallelism of computer data lanes is preserved. Parallelism is critical to scaling the interconnect fabric for future low-latency and energy proportional computing [6].

SDM: To scale the bandwidth for intra-building interconnection, SDM uses multiple fibers in parallel to transfer multiple streams of data.

Parallel optics, using VCSEL arrays, have played an important role in scaling the density and bandwidth of intra-datacenter communication. Parallel optics take advantage of integrated electronic driving circuits and the good yield of VCSEL arrays. This presents a lower-cost and lower-power solution than discrete SR transceivers. The disadvantage is the expense of parallel fiber termination such as MTP/MPO, and the ribbon fibers required for external connections. As the data rate increases beyond 10Gb/s per lane, it becomes increasingly difficult to scale the data rate in large VCSEL array to align with next-generation common electrical interfaces (CEI), such as CEI-25 at 25Gb/s or CEI-28 at 28Gb/s, and to cover the reach needed for intra-building interconnection. The large differential-modal-skew among different multi-mode fibers presents design challenges at 40Gb/s and 100Gb/s, when data streams span multiple physical lanes.

WDM: To overcome the limitations of TDM and SDM, wavelength multiplexing, where multiple wavelengths of light run in a single fiber, appears to be a promising technique. The true potential of the vast bandwidth available with single-mode optical fibers is fully exploited. A low-power and low-cost photonic integration circuit (PIC), using wavelength division multiplexing (WDM), holds the promise to further scale the density, reach, and data rates needed for next-generation datacenter networks, and to enable new network architectures and applications.

Other multiplexing techniques, such as optical orthogonal frequency division multiplexing (O-OFDM) [7] and high-order modulation [8], can also scale bandwidth and capacity within a single fiber. But these methods require gear boxes to perform signal encoding, ASICs for digital signal coding and processing, and/or ADCs and DACs for signal conversions between analog and digital domains, all of which consume large amounts of power and are cost-prohibitive for datacenter applications.

WDM for Datacom

Light is intrinsically parallel, and single-mode fiber has tens of terabits of bandwidth. Multiple data streams using WDM could naturally scale the optical interconnect density in datacenters.

WDM has been broadly deployed in wide-area-networks (WANs). Since fiber is scarce between remote datacenters, spectral efficiency is important. This makes WDM a natural fit for WAN interconnection.

Traditionally, the intra-datacenter environment was fiber-rich and spectral efficiency was never an important concern. However, as bandwidth and datacenter size grow, the cost of fiber, together with the expense of termination and management, has become a significant portion of overall link cost. In addition, increasingly large fiber bundles can present problems for mechanical clearances and air flow, making equipment more susceptible to thermal and cooling issues.

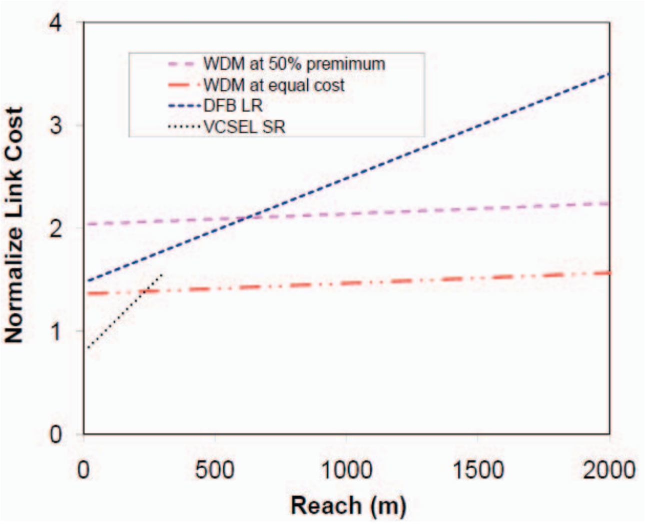


Figure 3. Normalized 100G link cost using SR, DFB LR and WDM

Figure 3 is a simplified presentation of 100Gb/s link cost, as a function of distance, using three different optical-interconnect technologies: (1) VCSEL short reach transceiver, (2) edge-emitting DFB LR transceiver, and (3) integrated WDM transceiver with 10 wavelengths. The starting point at 10m is the combined cost of optical transceivers and fiber terminations. Longer links become more expensive since longer fiber cables cost more. At various reaches, the slope ($/meter) of the WDM links is ten times lower than the slope for single-mode fiber and twenty times lower than the slope for multi-mode fiber. This is a result of (1) the relatively lower cost of single-mode fiber to multi-mode fiber and (2) the bundling of wavelengths within a single-mode fiber, such that only one fiber is needed for transmitting the same amount of bandwidth. When the cost of WDM and an uncolored DFB transceiver is the same for an identical amount of bandwidth, integrated WDM technology can result in drastic reduction of link cost (see Figure 3). From the per-link capital expenditure perspective, even if the WDM transceiver costs 1.5 times that of an uncolored DFB LR for the same amount of bandwidth, there is still a clear advantage to using WDM at reaches beyond 800 meters.

From a cost and manufacturing perspective, when multiple wavelengths are integrated into the same package and transmitted over a single fiber, not only can spectral efficiency be increased , but also the common electronic control circuits can be shared across different channels. The partition length, defined as the length beyond which an integrated WDM package can offer a total-cost savings in comparison to an uncolored discrete LR solution, is shown in Figure 4.

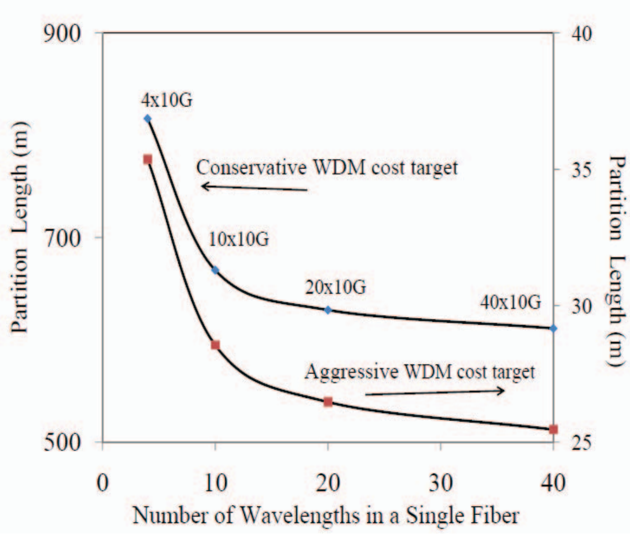


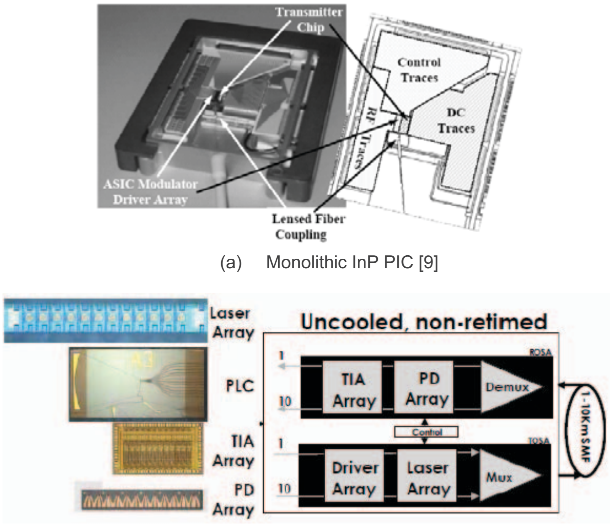
Figure 4. Partition length of WDM links vs conservative LR cost target and aggressive SR cost target.

In the long run, extrapolating from the general cost/bandwidth rule in data communications (datacom) industry development -- that a new generation technology should achieve 10 times the bandwidth with only 4 times the cost increase -- integrated WDM technology would be very cost competitive even in the region (about 20m, as represented by the "aggressive WDM cost target" curve in Figure 4), in which low-cost parallel optics dominate today. Note in Figure 4 that the partition length saturates at around 20 wavelengths. Therefore it is more beneficial to optimize WDM transceiver cost than spectral efficiency, given the reach needed and the datacom cost model.

The greatest challenge with the application of WDM in datacom is scaling the technology in size, power, and cost simultaneously.

It is difficult to meet datacom requirements with discretely packaged single-wavelength WDM transceivers. Photonic Integrated Circuits (PICs) (Figure 5) in monolithic [9] or hybrid [10] forms could meet datacom power, space, and cost requirements of datacom, significantly improve the landscape of next-generation datacenter interconnections.

Monolithic integration aims at lowering costs by taking advantage of streamlined planar circuit processing at wafer scale. Hybrid integration aims to use the most mature and best-performing technologies in optics and electronics, as well as the best-performing passive and active elements. Figure 5 (b) shows a hybrid PIC example [10], [11]. It uses 8nm wavelength-spaced and directly modulated DFB lasers whose outputs are multiplexed into a single fiber using silicon arrayed waveguide grating (AWG). At the receiver, incoming signal streams are de-multiplexed by another AWG into individual wavelengths and accepted by a photo-detector array.



(a) Monolithic InP PIC [9]

(b) Hybrid PIC [10,11]

Figure 5. Photographic and schematic presentation of Photonic Integrated Circuit of 100Gb/s WDM transmitter module (a) monolithic InP PIC [9] (b) hybrid PIC [10,11]

WDM DFB lasers used for long distance telecommunications typically require too much power and have low modulation efficiency for datacom applications. To improve power density, each DFB laser needs to (1) operate at an ultra-low driving current (2) with a very large modulation bandwidth (3) over a wide temperature range. Ideally, a WDM transmitter should operate at the power level of a VCSEL transmitter. Short-cavity DFB lasers have demonstrated high modulation efficiency with ultralow threshold current [12]. Once the threshold current of DFB reaches the same level as VCSELs, the low-cost electrical driving circuits used for VCSEL arrays can be leveraged to drive the low-power DFB lasers.

Temperature control, which is used for output power and wavelength stabilization in traditional long-reach WDM lasers, contributes significantly to power consumption. To meet datacom power consumption limits, the wavelength accuracy and spectral efficiency of lasers can be compromised. In response, the transmitter in [9], uses an uncooled DFB laser with 8nm coarse wavelength spacing. This spacing is wide enough to eliminate the need for a thermal electrical cooler (TEC), yet narrow enough that 10 DFB lasers can be fabricated as an array on a single wafer.

The Quantum dot (QD) laser provides an alternative solution to low-threshold current, as well as providing beneficial temperature stability [13], [14]. As a result of three-dimensional dot confinement, energy levels in individual islands are discrete -- in contrast to the continuous dispersion of bulk and quantum-well active materials. Due to the discrete density of their energy states, QD lasers offer important advantages. The threshold current is lower for QD lasers and they also have wide-temperature stability, thus eliminate the need for expensive and power hungry thermo-electric coolers (TEC) and control circuits. This athermal property allows QD laser placement nearer to the chip/processor without performance degradations. The intrinsic variation in quantum dot sizes, which leads to a broad lasing envelope, can be used to its advantage. For applications that require a broad gain spectrum, such as tunable laser and WDM laser sources, quantum comb lasers with multiple pure low-noise optical wavelength outputs can be obtained from a single Fabry-Perot cavity [15].

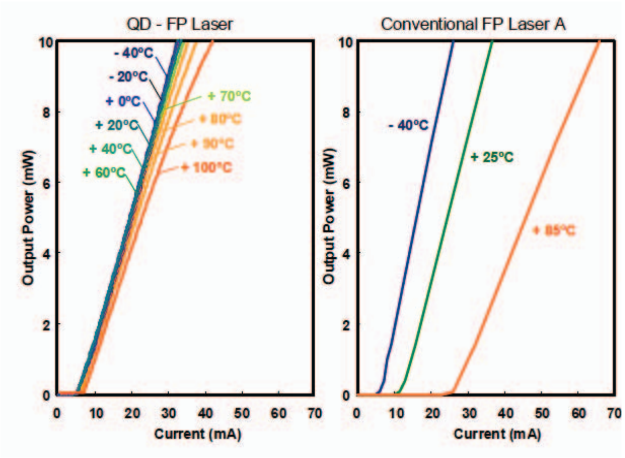


Figure 6. Comparison of the L-I curve of (a) quantum dot laser and (b) quantum well laser [13].

Conclusions

As bandwidth grows in datacenter networks, WDM technology, which taps into the terabit bandwidth of single-mode fiber, as well as the intrinsic parallelism of both light and the computing data streams, holds the promise of delivering scalable optical interconnects with low power consumption, high data throughput, long transmission distance, and the cost-effectiveness needed for future warehouse-scale datacenter networks.

Acknowledgement

We would like to thank Kevin Kissell for his valuable feedback on earlier drafts of this paper.

References

[1] L.A. Barroso and U. Hölzle, The Datacenter as a Computer – an Introduction to the Design of Warehouse-Scale Machines, Morgan & Claypool Publishers, 2009.

[2] C.F.Lam, et al, Fiber Optic Communication Technologies - what's needed for datacenter network operations, IEEE Optical Communications, July 2010.

[3] Light Peak Technology: http://www.intel.com/go/lightpeak/

[4] HSSG IEEE 802, An Overview: The Next Generation of Ethernet, http://www.ieee802.org/3/hssg/public/nov07/HSSG_Tutorial_1107

[5] R. Seifert, The Switch Book: The Complete Guide to LAN Switching Technology, John Wiley & Sons, 2000.

[6] W. Shieh and I. Djordjevic, OFDM for Optical Communications, Academic Press, 2009.

[7] D. Abts, et al, Energy Portional Datacenter Networs, ISCA, 2010.

[8] M. Seimetz, High-Order Modulation for Optical Fiber Transmission, Springer Series in Optical Sciences, 2009.

[9] R. Nagarajan, et al, Large-Scale InP Photonic Inegrated Circuits, IEEE J. Selected Topics Quantum Electronics, Jan/Feb 2007

[10] T. Schrans, et al, 100Gb/s 10km Link Performance of 10x10Gb/s Hybrid Approach with Integrated WDM Array of DFB Lasers, OFC/NFOEC, 2009.

[11] T. Schrans, et al, 10x10G WDM 10km SMF PMD Proposal, http://grouper.ieee.org/groups/802/3/ba/public/may08/, 2008.

[12] T.R. Chen, et al, Ultra High Modulation Efficiency of Ultralow Threshold Current Single Quantum Well InGaAs Lasers, Electronic Letters, Vol.27, No.17, 1993.

[13] http://www.qdlaser.com/

[14] D. Bimberg, "Semiconductor Quantum Dots: Genesis – The Excitonic Zoo – Novel Devices for Future Applications," Chapter 2 in Optical Fiber Telecommunications - V, Volume A, edited by I. Kaminow, et al, Academic Press, 2007.

[15] N. Yamamoto, et al, "Quantum Dot Optical Frequency Comb Laser with Mode-Selection Technique for 1-µm Waveband Photonic Transport System", Japanese Journal of Applied Physics, Volume 49, Issue 4, 2010.