Optical Switching Performance Metrics for Scalable Data Centers

Keren Bergman and Sébastien Rumley
Department of Electrical Engineering, Columbia University, 530 West 120th Street, New York, NY 10027
rumley@ee.columbia.edu

Abstract: *Optical switching can address some key challenges associated with scaling the communications infrastructure in data centers. We review the critical metrics that would enable significant performance gains and thereby wide adoption in data centers.*

Keywords: *Optical interconnects, Data Centers, Optical Switching*

I. Introduction

The performance of Data Center interconnects (DCIs) is gaining intense interest as cloud based applications span over a growing number of servers, trigger rising traffic volumes, and are increasingly sensitive to network congestion. Until recently, DCIs were primarily designed to ensure the connectivity of single servers with the Internet, and to support "North-South", generally latency tolerant traffic. Nowadays, DCIs are becoming the main backbone of higher-order computing architectures, supporting massive amounts of "East-West" traffic at high throughputs. Data centers are also increasingly operated with virtualization software for sharing servers among multiple users, further amplifying bandwidth demands across DCIs.

These transformations do not only require servers' network interfaces to support increased bandwidths (10Gbps today, 40 and 100Gbps in the near term), but also require the interconnect itself to cope with intensely growing traffic volumes. This means drastically limited over-subscription levels, as illustrated in Figures 1a and 1b. To interconnect 20,000 servers with 32 ports switches (following a generally adopted 3-levels fat-tree typed topology), a minimum of 668 switches are required (646 at the first level, 21 at the second level, and 1 at the last level), as well as 667 internal (i.e. switch-to-switch) links. This infrastructure is sufficient as long as servers communicate very sporadically, i.e. average network utilization of a single server remains below 0.1%. However, as requirements for bandwidth grow, potentially to the point where full-bisectional bandwidth is provisioned to support full utilization of all servers simultaneously, the number of switches and internal links required balloon – in this particular example, by factors up to ~7 for the switches and ~92 for the internal links. For a large data center involving 50,000 servers, the increase in terms of internal links is even higher (~125x). When network utilizations is very low, large data center scaling permits a better amortization of the interconnect cost. By contrast, when the interconnect utilization levels are high, as is the overwhelming applications traffic trend, cost grows super-linearly with data-center scaling [1].

The additionally required internal links, moreover, are predominantly higher-cost long reach since they span between racks (as opposed to short reach, intra-rack links - Fig 1c). Under minimal traffic requirements, these long reach links represent a relatively small portion of all links (less than 4% for the 20,000 servers, 32 port example), but under the highest traffic requirements, the ratio is reversed (75% in the example). This further amplifies the growing cost.

Altogether, the shift to high network throughput cloud applications can raise by orders of magnitude the cost of DCIs. This motivates the development of new technologies that can counter-balance this cost explosion.

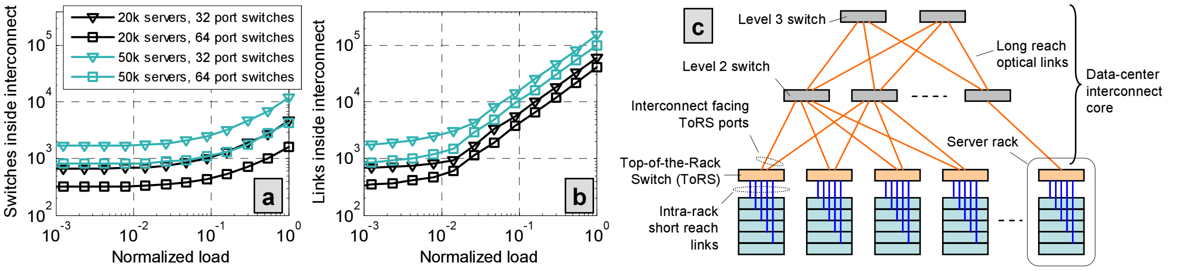

Fig. 1. Resources required in data-center interconnect as a function of expected traffic (normalized to server communication capabilities).

II. Motivation for Optical Switching

Optical cables account for a substantial part of the long-distance links present in current data centers, and in the future this proportion is destined to further rise. Copper based links can span over relatively long distances (10GBase-T: 100m, 40GBase-T: 30m, 40GBase-CR4: 7m), but are subject to higher link latency as they rely on digital signal processing and advanced coding formats [2]. Moreover, as line-rates rise to 40Gbps or beyond, it becomes simply more economical to rely on fiber based cables for distances larger than 5-10m. This break-even distance is expected to decrease in the next years.

Provided that the majority of links involved in the DCI core (Fig. 1c) are (or will be) fiber optics based, the opportunity to use transparent optical switches instead of conventional, electrical packet based ones has been investigated. Such an all-optical architecture can drastically limit the number of hops taken by messages in the interconnect core. In the ideal case, each message can be optically emitted once at the Top-of-the-rack Switch (ToRS), kept in the optical domain at all times inside the DCI core, and received at the ToRS it is destined to. Transparent switching does not provide benefits in terms of number of installed fibers, but can help to drastically limit the number of optical emitters and receivers (transceivers). Assuming negligible optical switching time and ideal arbitration, this number can be reduced to the number of interconnect facing ToRS ports (Fig. 1b), plus several units attached to the data-center internet gateway. Under the highest traffic requirements, these interconnect facing ToRS are at most as numerous as servers. For the 20,000 servers, 32 port example, the transparent switching can represent a factor 3 decrease in the number of required transceivers.

III. Optical Switching Caveats

The alluring picture of transparent switching is incomplete, however, as several hurdles must be taken into account before considering a transparent DCI core to be a viable option.

A. Switching time. Optical switches replacing electrical ones in the interconnect core must provide negligible switching times. To be considered as negligible, a switching time must be a) significantly shorter than the median data exchange duration and b) smaller than the latency tolerated by cloud applications. By data exchange, we mean an uninterruptible sequence of bits sent by one client to another over an optical, circuit switched, link.

If condition a) is not fulfilled, the DCI throughput will be impacted: no data can be transmitted while an optical switch is adapting its state. Furthermore, after each switch state adaptation, the end receiver must also recover its synchronization. If, after each data exchange of duration d_{data} the link is silent for a time $d_{switching+sync} = d_{data}$, the link throughput is limited to 50%. A second link operated in parallel becomes thus necessary. If optical switches configurations are based on standard Ethernet frames, a data exchange is at most ~1500 bytes long. At 10Gbps and 40Gbps, the *maximum* duration is thus of 1.2 us and 0.3 us respectively. This requires switching times <120ns and <30ns respectively to keep throughput above 90%. Furthermore, as revealed by Benson et al. [3], median *flow* sizes in data-center fall generally in the 100 to 1000 byte range. Each flow being composed of one Ethernet frame at least, this further pushes switching time requirements to 80ns (10Gbps) and 20ns (40Gbps) if median size is 1000 bytes, and to 8ns and 2ns if most data exchanges are smaller than 100bytes. Electrical packet switches are not subject to this throughput issue as long as the bandwidth of the internal crossbar connecting input to output queues is high enough, which is generally the case. Also note that even shorter switching times are required if link rates rise to 100Gbps or more.

As for condition b), one can take the latency displayed by current interconnects as a reference for “tolerated” latency, although applications may tolerate higher latencies. An Ethernet 10G switch adds an average latency of 0.5 to 1us at each hop [4]. This translates in a typical server-to-server latency of the order of several microseconds. As a point of comparison, node-to-node latency in supercomputing environments falls below or close to the microsecond [5].

Nanosecond scale switching times (1ns – 100ns) have often been demonstrated with electro-optically tuned switches [6], yet these devices generally inflict significant power penalties to optical signals [7]. This threatens scalability as detailed hereafter. If median data exchange durations can be made larger, microsecond scale switching times are acceptable. Microsecond switching times can also be tolerated if cost of optical switches is kept low. Hence, in this case, the cost of the extra transceivers required to keep throughput unchanged can be counter-balanced. Microsecond switching times can be achieved with thermal effects or with MEMS-actuated planar switches [8]. Switching times above 10us, in contrast, may affect application performance too much to be considered.

B. Arbitration. In Section 2, an ideal arbitration has been assumed: schedules of all data exchange, as well as their routing, are ideally picked up to minimize average latency, maximize throughput, or a combination thereof, *while ensuring that no network resource is allocated more than once.* Realizing such an ideal arbitration is an extremely complex optimization problem. Optical switches are unable to buffer data exchange for arbitration durations: consequently, end-to-end reservations are necessary to ensure that all the resources required for a data exchange to happen are free at the same time. Operating with end-to-end reservations is not only more demanding in terms of optimization efforts, but also lead to a fragmentation of the occupancies over time, synonym of crippled throughput. Furthermore, collecting resource availabilities, and dispatching reservation decision from and across the DCI is challenging at scale.

As a result, arbitration cannot be ideal and necessarily translates into diminished throughput and increased latency. The former can be compensated by provisioning links and switches (in which case the necessity of having low cost switches applies again). The latter can be improved by maintaining circuit on after use [9], and/or by mean of circuit prefetching [10]. If circuit maintenance strategy is applied, provisioning more links also helps to reduce latency.

C. Cost. By transitioning to a transparent DCI *while keeping the same bandwidth*, the number of transceiver required can be reduced by a potentially significant factor. However, to compensate for switching times and arbitration latency, bandwidth must be boosted which comes at the expense of additional transceiver. Furthermore, the insertion of one or more switches modifies the optical budget, as each switch might induce a power penalty ranging from 1 to 10dB or more. More expensive transceivers providing extra optical budget, e.g. longer-reach ones, might thus be required.

From a transceiver cost point of view, if regular transceivers support optical switching and if arbitration and switching time effects require twice more bandwidth to be provisioned, the transition to optical switching, in the 20,000 server case, can provide a 33% cost reduction. However, if twice more expensive "wide budget" transceivers are required, the total transceiver cost rises by 33% in that example.

Transitioning to transparent DCI, however, permits additionally to replace electrical packet switches by optical switches. An inspection of 10Gb Ethernet switch prices [11] shows that every for 10Gbps interconnect every switch port cost around 100$ (350$ for 40Gbps). The cost of an optical port should remain below those prices for 32 ported switches. If much larger port counts are available, the number of levels in the interconnection network that replaces the fat-tree can be reduced, which also limits the number of switches required. For instance, for the 20,000 servers example, having 320 ports switches available instead of 32 reduces the number of ports required by 30%.

D. Power consumption. One can expect the transceivers number reduction and the replacement of electrical routers by transparent switches to diminish the power consumption. Some optical switches, however, do consume a non negligible amount of power to maintain their configuration. The consumption of a future ring resonator based switching fabric with 128 ports, for instance, has been estimated to ~47mW per port. With 10Gbps per port, this translates into 4.7 pJ/bit. For SOA/MZI based switches, hundreds of mW per port must be accounted [12], translating into tens of pJ/bit. As a point of comparison, best-in-class electrical packet routers consume 20-30 pJ/bit. Consumption of optical switches should typically be limited to such values.

On the transceiver side, power consumption is affected by the higher launch optical powers required to overcome extra power penalties introduced by switches. If these power penalties are of 10dB, laser consumption is multiplied by a factor of ten. Lasers are not the dominant energy consumers in transceivers *today*, but this may no longer be the case in the future. If laser is the only consumption of a transceiver, a transceiver facing an additional 10dB power penalty will consume as much as 10 point-to-point transceivers, negating power savings obtained by diminishing the transceiver number.

E. Scalability. Optical power penalties of switches, eventually, limit the scalability of transparent DCI architectures. If end-to-end power penalties exceed the amplest transceiver power budgets, optical amplifiers become necessary in between the switches. This rises both interconnect cost and power consumption.

In general, the power penalty associated to the interconnect core should be limited to 20dB. If three switch stages are sufficient to provide full connectivity, this fixes the fiber-to-fiber power penalty for individual switches to around 6.5 dB. This is typically the case if 320 ports switches are used in a 20,000 server data-center. However, if fives or even seven stages are necessary, constraint on individual switch penalty becomes severe, especially for electro-optical effect based switches, capable of sub-microsecond switching time.

IV. Conclusion

Optical switching in data centers offers potential advantages in terms of cost and power consumption. However to realize these gain, critical design metrics cannot be neglected. Mitigating these photonic technology challenges requires some device level advances, but above all subtle trade-offs to be identified and addressed at the architecture level.

References

[1] S. Rumley, S. D. Hammond, A. Rodrigues, K. Bergman, "Design Methodology for Optimizing Optical Interconnection Networks in High Performance Systems", ISC-HPC conference, 2015.

[2] Z. Zhang, et al. "A 47 Gb/s LDPC Decoder with Improved Low Error Rate Performance", Symp. on VLSI Circuits, 2009.

[3] T. Benson, et al., "Network Traffic Characteristics of Data Centers in the Wild", SIGCOMM, New Dehli, India, 2010.

[4] H. Subramoni, P. Lai, M. Luo and D. K. Panda, "RDMA over Ethernet – A Preliminary Study", IEEE Cluster, 2009.

[5] D. Chen, et al., "The IBM Blue Gene/Q Interconnection Network and Message Unit", Supercomputing, Seattle, WA, 2011.

[6] T. Shiraishi, et al., "Scalability of Silicon Photonic Enabled Opticaly Connected Memory", IEEE Optical Interconnect, 2014.

[7] D. Nikolova, S. Rumley, D. M. Calhoun, Q. Li, R. Hendry, P. Samadi, K. Bergman., "Scaling silicon photonic switch fabrics for data center interconnection networks," Optics Express 23(2), 2015.

[8] T. Joon Seok, N. Quack, S. Han, R. S. Muller, M. C. Wu, "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers", Optica (3) 1, 2016.

[9] K. Wen, et al., "Reuse Distance Based Circuit Replacement in Silicon Photonic Interconnection Networks for HPC", IEEE Symposium on High Performance Interconnects, 2014.

[10] K. Wen, S. Rumley, J. Wilke, K. Bergman, "Latency-avoiding Dynamic Optical Circuit Prefetching Using Application-specific Predictors," ISC-HCP ExaComm Workshop, 2015.

[11] http://www.colfaxdirect.com

[12] S. Liu, et al., "Low Latency Optical Switch for High. Performance Computing With. Minimized Processor Energy. Load", JOCN 3(7), 2015.