Elastic WDM Switching for Scalable Data Center and HPC Interconnect Networks

Adel A. M. Saleh, Akhilesh S. P. Khope, John E. Bowers, Rod C. Alferness
Electrical and Computer Engineering Department, University of California Santa Barbara, California, USA
AdelSaleh@ece.ucsb.edu

Abstract: *An elastic WDM switch, suitable for silicon photonic integration, was recently proposed by the authors. Here, we summarize its characteristics, and show how it can enhance the scalability and performance of data centers and HPCs.*

Keywords: *(Optical Interconnects, Data Centers, HPC, Silicon Photonics, Dynamic Networks)*

I. Introduction

The use of fast optical circuit switching in the interconnect networks of data centers and high performance computers (HPCs) has been advocated by various researchers [1]-[5]. Recently, we proposed an elastic, microring-based, WDM, crossbar switch suitable for such applications [6]. The switch is compatible with silicon photonic integration, and thus, given the steady progress in that field [7], promises to have small size, cost and power requirements. The purpose of this paper is to show how one can exploit these properties to cost-effectively achieve enhanced performance and scalability of future data centers and HPCs, by incorporating a large number of such switches in their interconnect fabrics. The architecture, operational principles and unique characteristics of the WDM switch are summarized in Sec. II. Examples of its application in folded Clos and HyperX interconnect networks are presented in Secs. III and IV, respectively.

II. Elastic, WDM Crossbar Switch

Figure 1a shows the $N\times N$, M-wavelength, WDM switch proposed in [6], which has N^2 cross-points (L-blocks), each having L microrings, as shown in Fig. 1b. The microrings are individually tuned to drop up to L wavelengths from any input to any output. This elastic connectivity is a key feature of the switch, which leads to a greatly reduced latency [6]. Generally, $1 \leq L \leq M$ and $M \leq N\,L$; and, typically, $L \ll M$. An input-buffered transmitter and a receiver are shown in Fig. 1c. The optical spectrum is depicted in Fig. 1d, showing open slots, to which the microrings would be tuned when they are not required to drop any wavelength. The M operating wavelengths and the open slots are assumed to fall within one free spectral range of the microrings to enable individual control of each wavelength.

The switch operation requires central control, and, as shown in Fig. 1e, assumes equal time slots that are much longer than the switching time (which includes both the time needed by a centralized algorithm to compute the input-output wavelength assignments for each time slot, and the time needed to tune and stabilize the microrings). Our wavelength assignment algorithm is illustrated in Fig. 2, which generalizes known algorithms for determining the port connectivity in input-buffered packet switches [8]. For each time slot, the algorithm receives an admissible $N\times N$ connection matrix $\boldsymbol{A}$, which indicates the input-output wavelength traffic demands. The matrix $\boldsymbol{A}$ is expanded, using a *graph matching algorithm*, into the sum of M *permutation matrices*, each representing the input-output connectivity for one of the M wavelengths. The algorithm runs in polynomial time, which is estimated to take less than 1 μs. In addition, since each microring needs to be tunable across the entire wavelength spectrum, *thermal tuning* is required, which is estimated to need on the order of 1 μs. Thus, the transmission time slot shown in Fig. 1(e) can be on the order of 10 μs (or longer).

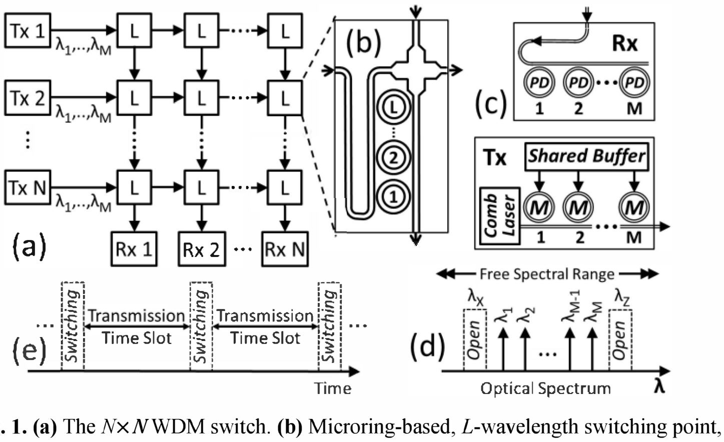
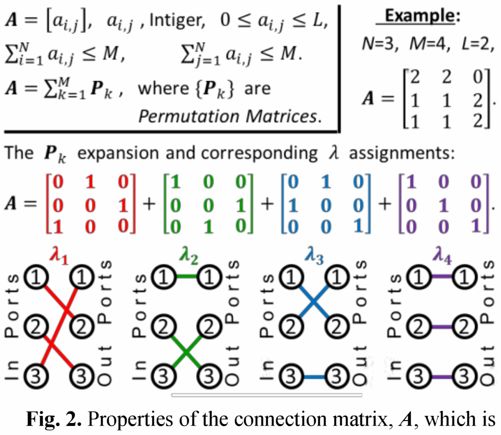


Fig. 1. (a) The $N\times N$ WDM switch. **(b)** Microring-based, L-wavelength switching point, with a low-crosstalk multimode crossing. **(c)** M-wavelength transmitter (Tx) & receiver (Rx) - ***PD***: Photodetector, ***M***: Modulator. **(d)** Optical spectrum. **(e)** Timing diagram.


$\boldsymbol{A} = [a_{i,j}]$, $a_{i,j}$, Intiger, $0 \leq a_{i,j} \leq L$,
$\sum_{i=1}^{N} a_{i,j} \leq M$, $\sum_{j=1}^{N} a_{i,j} \leq M$.
$\boldsymbol{A} = \sum_{k=1}^{M} \boldsymbol{P}_k$, where $\{\boldsymbol{P}_k\}$ are *Permutation Matrices.*

Example:
N=3, M=4, L=2,

$$\boldsymbol{A} = \begin{bmatrix} 2 & 2 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 2 \end{bmatrix}.$$

The $\boldsymbol{P}_k$ expansion and corresponding λ assignments:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Fig. 2. Properties of the connection matrix, A, which is expandable into the sum of *permutation matrices*; and an example showing the expansion and the λ assignments.

This research was supported by the American Institute for Manufacturing of Integrated Photonics (AIM Photonics).

The switching mode described above is actually one of three possible switching modes, which are listed below:

1) Quasi-static circuit switching: In this mode, the switching time is longer than the packet time, which applies to HPC interconnect applications, where the processors interchange very short ($\ll 1\ \mu$s) packets. In this case, the connectivity pattern of the switch is set to best match the workload, and is left fixed for the duration of each computational task. This mode also applies to data centers to respond to relatively slow workload variations.
2) Fast time-slot switching: This is the main switching mode discussed above, which is applicable when the time slot duration (e.g., 10 μs) is much larger than the switching time. The input-output connectivity for each time slot is configured in real time in response to varying input-output packet traffic demands (i.e., matrix A above).
3) Ultrafast time-division-multiplexed (TDM) switching: This mode augments the above modes by using an end wavelength, e.g., λ_1 in Fib. 2d, to add periodic, ultrafast-switched, TDM, all-to-all connectivity within each time slot, to enable the interconnected devices to exchange very short ($\ll 1\ \mu$s) messages. To realize this, one microring is reserved within each L-block to be rapidly tuned, via *current-injection*, between λ_1 and the adjacent open slot, i.e., λ_X. This way, the ultra-fast-switched channel would not affect the other connections.

III. Application in Folded Clos Interconnect Networks

A folded Clos, leaf/spine data center is shown in Fig. 3a, which has $2P$ leaf switches and P spine switches, all having the same capacity of $2P$R (b/s), where R is the bit rate of a fiber link. For R ≥ 100G, WDM fiber links with integrated transceivers [9],[10] are typically employed. For example, for M-wavelength links with r (b/s) per wavelength, R = 100G; 400G and 1,000G can be implemented, respectively, with $M = 4$, r = 25G; $M = 16$, r = 25G and $M = 20$, r = 50G.

Note from Fig. 3a that the total number of servers is limited to $2MP^2$ because one cannot add any more leaf or spine switches to the topology. However, if the different wavelengths from a switch port are fanned-out to different switches, then more leaf and spine switches, can be added, and hence, also servers. This is done in Fig. 3b by adding a layer of $N\times N$, M-wavelength WDM switches, leading to a factor of N increase in the number of servers. (A similar technique was proposed in [3], using arrayed waveguide grating routers (AWGRs) instead WDM switches.) Of course, adding a layer of packet switches instead of the WDM switches would also lead to an increased number of servers, which is the idea behind the multi-level folded Clos architecture. But, there are two problems with that approach: First, latency will be increased due to the added layer of packet switches, while the WDM switches will add no latency when operating in the circuit switching mode discussed above. Second, because the WDM switches are compatible with silicon photonic integration, as mentioned in Sec. I, they will have lower cost and power consumption than the packet switches.

In addition to the above scalability, the WDM switching layer also enables reconfigurability of the data center, e.g., to respond to slow or fast workload variations, using, respectively, the circuit switching mode or the time-slot switching mode (possibly augmented by the TDM switching mode). The goal is to enable elastic bandwidth connectivity between the various port pairs of the packet switches to better match workload variations. This leads to a better utilization of the available resources (e.g., servers and transponders), and thus, improved performance (e.g., throughput and latency).

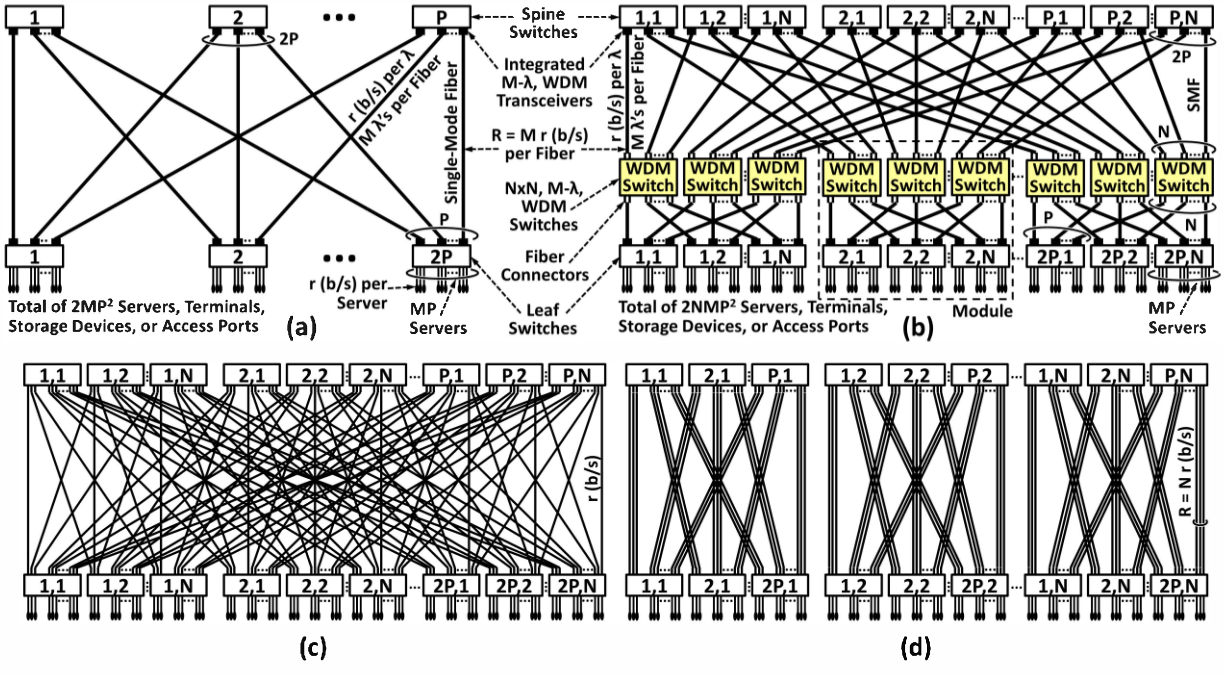


Fig. 3. (a) A folded Clos leaf/spine data center architecture with WDM fiber links. **(b)** An N-fold increase in the number of servers is obtained by adding a layer of $N\times N$, M-wavelength WDM switches (each representing either two $N\times N$ switches, one for each direction, or one bidirectional $2N\times 2N$ switch). For simplicity, we assume that $M = N$, in Figs. c and d. **(c)** With all the WDM switches in Fig. b set to an AWGR routing pattern, a dense, folded Clos topology results. **(d)** With all the WDM switches in Fig. b set to a straight-through pattern, N, uncoupled, folded Clos, topologies result.

Two extreme configurations of the interconnection network of Fig. 3b are shown in Figs. 3c and 3d. The dense folded Clos topology of Fig. 3c results when all the WDM switches are set to an AWGR pattern (as was done in [3]). The same topology would result if the packet switches had N-times as many ports, interconnected with N-times as many single-wavelength fibers, which would not be a scalable solution. The topology of Fig. 3d results when all the WDM switches are configured to pass the input wavelengths straight through the switch. In this case, the data center is divided into N uncoupled, folded Clos partitions with N-wavelength links. One can get a rich set of interconnection topologies, to respond to varying workloads, by configuring the various WDM switches differently from the above two extremes.

As a design example: a leaf/spine system with $P = 16$, r = 25 Gb/s per λ, $M = 16$ λ's per fiber (i.e., packet switches with 32 × 400G ports), requires $4P^2 = 1{,}024$ WDM switches (with $N = 16$-ports) to support $2NMP^2 = 131{,}072$ servers.

IV. Application in HyperX Interconnect Networks

HyperX topology [11] is an important class of multi-dimensional, interconnect networks that encompasses other well know types of data center and HPC interconnect networks, such as the Flattened Butterfly and the Hypercube. In general, it requires a very dense fiber interconnection fabric, as depicted in Fig. 4a, which shows a small, *regular* HyperX network, having $D = 3$ dimensions and $S = 4$ packet switches or routers per dimension, yielding a total of $S^D = 64$ switching nodes. Large data centers and HPCs would require a much larger value for S, which could result in very complex fiber interconnect fabric. Some packaging techniques are proposed in [11] to simplify the system wiring. The wiring can be further simplified and the number of fibers can be greatly reduced by incorporating WDM switching as shown in Fig. 4c. The switches also enable the reconfigurability of the interconnect topology, thus making it possible to adjust the bandwidth connectivity among the various packet switches to better match workload variations.

As a design example, consider a HyperX system with $D = 3$, $S = 16$, $K = 1$, $T = 16$, which supports $S^D = 4{,}096$ routers of a radix of $KD(S\text{-}1)+T = 61$, and $TS^D = 65{,}536$ servers. This requires $(D\text{-}1)\,S^{(D-1)} = 512$ WDM switches, with $N = 2KS = 32$ ports and (with one λ per channel) $M = D(S\text{-}1) = 45$ λ's per port, which are challenging but achievable numbers.

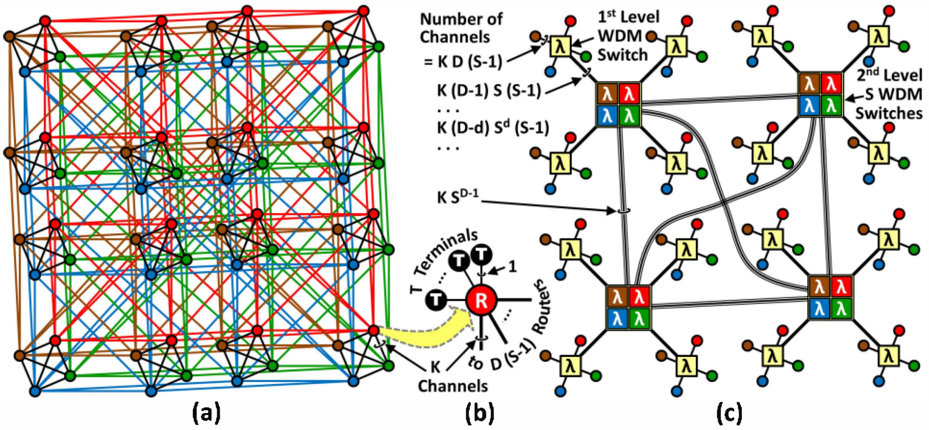


Fig. 4. (a) A small, *regular*, HyperX interconnect network with $D = 3$ dimensions and $S = 4$ packet switches or routers per dimension. **(b)** Details of a switching node, showing a connectivity of K channels between corresponding routers, and T terminals, processors or servers connected to each router. **(c)** Inserting WDM switches (labeled '**λ**') results in greatly simplifying the wiring, reducing the number of fibers, and enabling reconfigurability.

V. Conclusions

A flexible WDM switch of potentially small size, cost and power was presented. It was argued that incorporating a large number of such switches in the interconnect networks of data centers and HPCs can enhance their performance and scalability in a cost-effective way. For example, adding WDM switching was shown to simplify the system wiring, greatly reduce the number of required fibers, scale-out the system, and enable reconfigurability of the system. One can also use spatial switching with fiber ribbons or multi-core fibers instead of, or in addition to, WDM switching [3],[12].

References

[1] G. Porter, et al, "Integrating microsecond circuit switching into the data center," ACM SIGCOMM '13, Aug 2013.
[2] H. Liu, et al, "Circuit switching under the radar with REACToR," 11th USENIX Symposium on NSDI, 2014.
[3] A.A.M. Saleh, "Scaling-out data centers using photonics technologies," Photonics in Switching Conference, July 2014.
[4] R. Yu, et al, "A scalable silicon photonic chip-scale optical switch for HPC systems." Optics Express, Dec 2013.
[5] D. Nikolova, et al, "Scaling silicon photonic switch fabrics for data center interconnection networks," Optics Express, Jan 2015
[6] A.S.P. Khope, et al, "Elastic WDM crossbar switch for data centers," Optical Interconnects Conference, May 2016.
[7] T. Komljenovic, et al, "Heterogeneous silicon photonic integrated circuits," Journal of Lightwave Technology, Jan 2016.
[8] P. Giaccone, et al, "Randomized scheduling algorithms for high-aggregate bandwidth switches," IEEE JSAC, May 2003.
[9] C. Cole, "Future datacenter interfaces based on existing and emerging optics technologies," IEEE Summer Topicals, 2013.
[10] B. R. Koch, et al, "Integrated silicon photonic laser sources for telecom and datacom," OFC/NFOEC, 2013.
[11] J.H. Ahn, "HyperX: topology, routing, and packaging of efficient large-scale networks," ACM SC '09, Nov 2009.
[12] A.A.M. Saleh, "Evolution of the architecture and technology of data centers towards exascale and beyond," OFC 2013.