A Scalable, Low-Latency, High-Throughput, Optical Interconnect Architecture Based on Arrayed Waveguide Grating Routers

Roberto Proietti, Zheng Cao, Christopher J. Nitta, Yuliang Li, and S. J. Ben Yoo, *Fellow, IEEE, Fellow, OSA*

***Abstract*—This paper proposes, simulates, and experimentally demonstrates an optical interconnect architecture for large-scale computing systems. The proposed architecture, Hierarchical Lightwave Optical Interconnect Network (H-LION), leverages wavelength routing in arrayed waveguide grating routers (AWGRs), and computing nodes (or servers) with embedded routers and wavelength-specific optical I/Os. Within the racks and clusters, the interconnect topology is hierarchical all-to-all exploiting passive AWGRs. For the intercluster communication, the proposed architecture exploits a flat and distributed Thin-CLOS topology based on AWGR-based optical switches. H-LION can scale beyond 100 000 nodes while guaranteeing up to 1.83× saving in number of inter-rack cables, and up to 1.5× saving in number of inter-rack switches, when compared with a legacy three-tier Fat Tree network. Network simulation results show a system-wide network throughput reaching as high as 90% of the total possible capacity in case of synthetic traffic with uniform random distribution. Experiments show 97% intracluster throughput for uniform random traffic, and error-free intercluster communication at 10 Gb/s.**

***Index Terms*—Arrayed waveguide grating routers (AWGRs), datacenter networking, optical interconnects, optical switches.**

I. Introduction

THE landscape of today's cyberinfrastructure is dominated by innovations in data centers and high performance computing (HPC) systems. As of January 2014, Google processes approximately six billion searches per day [1] and Facebook data centers serve 1.3 billion users who spend 640 billion minutes browsing monthly [2]. In scientific computing, analyses, simulations, and visualization of extreme data are enabled by HPCs. Both large scale data centers and HPC systems include many thousands of servers intimately networked with each other. As Amdahl's law [3] suggests, a parallel computing system with balanced processing, memory, and communications can perform optimally across most applications. This indicates that an optimized petascale computing system requires ∼1 PB/s bisection bandwidth in addition to ∼1 PB memory and ∼1 petaFLOPS processors. Typical petascale data centers and computing systems are already consuming on the order of many megawatts [4]. Scalability, efficiency, throughput, and latency of these large systems are largely determined by those of the interconnection networks [5]. Today's interconnection networks typically rely on electronic-crossbar switches that store and forward packets [6], which may lead to limited scalability, capacity, throughput, and power-efficiency. Large-scale data centers and HPC systems inevitably cascade such switches to interconnect a large number of servers, adding complexity, inefficiency, and latency to the systems. In order to support several hundreds of thousands nodes, future data centers and HPCs must undergo complete architectural and technological transformations so that they will meet the necessary performance requirements while operating within the power limits.

Optical interconnects can potentially bring transformative changes to computing system architectures. Compared to electrical interconnects, optical interconnects provide (1) higher transmission bandwidth and lower energy consumption independently of distance, (2) inherent parallelism, (3) low interference and crosstalk, and (4) low parasitic. In addition, optics offers wavelength (frequency) routing capability not available in electronics with comparable size. As a result, optically interconnected computing systems can potentially achieve (1) higher scalability and energy efficiency, (2) high-density parallel links and buses overcoming input/output pin density limits, and (3) low-latency avoiding the need for including repeaters or switches with store-and-forward architectures. Further, optical devices with wavelength (frequency) routing capability can achieve all-to-all interconnection between computing nodes without contention.

An arrayed waveguide grating router (AWGR) [7], [8] is an example of devices with such wavelength routing capability. As Fig. 1(a) and (b) illustrate, the well-known wavelength routing property of an AWGR allows any input port to communicate with any output port simultaneously using different wavelengths without contention. Thus an $N \times N$ AWGR intrinsically provides all-to-all communication among N compute nodes in a flat topology using N wavelengths. Fig. 1(c) illustrates an $N \times N$ AWGR switch with N compute nodes (Node 1-Node N Tx/Rx) with each node having k_t transmitters and k_r receivers, respectively ($1 \le k_t, k_r \le N$). Note that, only the TXs are tunable. This illustration assumes ring-resonator-based transmitters and modulators with frequency selectivity. When $k_t = N$ and $k_r = N$, all-to-all interconnection without contention can be achieved in

Manuscript received September 8, 2014; revised December 14, 2014 and January 16, 2015; accepted January 16, 2015. Date of publication January 21, 2015; date of current version February 20, 2015. This work was supported in part under DoD Agreement Number: W911NF-13–1–0090.

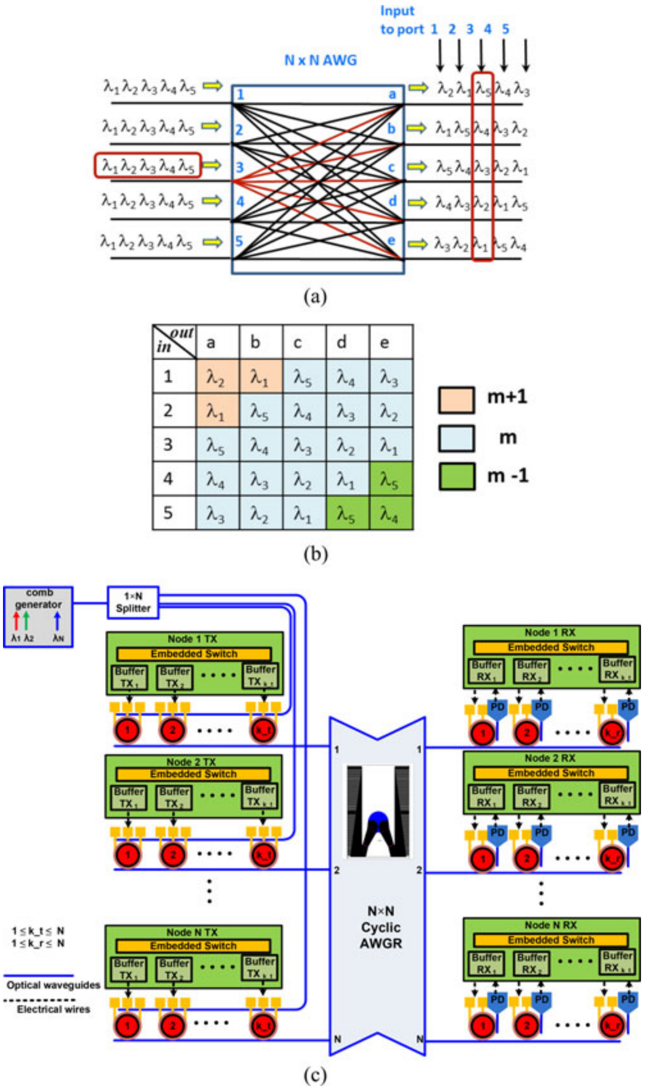
R. Proietti, C. J. Nitta, and S. J. B. Yoo are with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: rproietti@ucdavis.edu; cjnitta@ucdavis.edu; sbyoo@ucdavis.edu).

Z. Cao is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhcao@ucdavis.edu).

Y. Li was with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA. He is now with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089-0781 (e-mail: yuliangl@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2015.2395352



(a)

out / in	a	b	c	d	e
1	λ_2	λ_1	λ_5	λ_4	λ_3
2	λ_1	λ_5	λ_4	λ_3	λ_2
3	λ_5	λ_4	λ_3	λ_2	λ_1
4	λ_4	λ_3	λ_2	λ_1	λ_5
5	λ_3	λ_2	λ_1	λ_5	λ_4

🟧 m+1
□ m
🟩 m -1

(b)

(c)

Fig. 1. (a) $N \times N$ AWGR's wavelength routing property (shown is $N = 5$ example), (b) a wavelength assignment table. Here, m, m + 1 and m-1 represent three different grating orders (Free Spectral Ranges), and (c) LIONS with a $N \times N$ AWGR, N compute nodes, with each node having k_t transmitters and k_r receivers, respectively ($1 + \leq k_t$, $k_r \leq N$). This illustration assumes ring resonator based transmitters and modulators with frequency selectivity.

a flat topology. This case will be called passive AWGR switch or passive Low-latency Interconnect Optical Network Switch (LIONS) since no optical reconfiguration is necessary. If $k_r < N$, contending conditions exist, and Ref. [9] demonstrated buffered architecture and [10] presented bufferless all-optical distributed control plane with performance far greater than typical electronic switches. If $k_t < N$, then the transmitter must be able to tune its wavelength corresponding to the desired destination node. This case will be called an active AWGR switch or active LIONS. Ref. [9] investigated cases where $k_t = 1$ (single tunable transmitter) and $k_r < N$ for a large scale rack-to-rack interconnect network.

The active AWGR will have to deal with contention resolution, even though the contention probability is minimized by the use of k_r RXs per output port and by the unique wavelength routing property in AWGR. Ref. [9], [10] show the performance of different LIONS switch architectures interconnecting a number of nodes ≥ 64 with $k_r = 4$. Even with a limited amount of RXs, the switch can sustain a throughput above 80% in case of uniform random traffic. Round-robin arbitration and an all-optical first-come first-served arbitration schemes are used in [9], [10], respectively.

This paper investigates and demonstrates a scalable, low-latency, high-throughput, flat optical interconnect architecture employing passive and active AWGR switches in a hierarchy. The proposed architecture, Hierarchical Lightwave Optical Interconnect Network (H-LION), leverages wavelength routing in AWGRs, and computing nodes (or servers) with embedded routers and wavelength-specific optical I/Os.

Within the racks and clusters, the interconnect topology is hierarchical all-to-all exploiting passive AWGRs and embedded switches in the servers. For the inter cluster communication, the proposed architecture exploits a flat and distributed thin-CLOS topology based on AWGR-based optical switches. As explained above, the active AWGR will have to handle some amount of contention, which will affect its performance in terms of throughput and latency. This could be overcome by using all-to-all passive interconnection also for inter-cluster communication, at the expenses of a smaller scalability due to the limited number of TRXs available for inter-cluster interconnection.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III introduces the proposed scalable interconnect architecture. Section IV presents network performance studies for the proposed architecture, and Section V demonstrates system emulation and experimental testbed studies by prototyping a cluster network involving AWGRs and FPGA boards. Section VI concludes the paper.

II. Related Work

Regarding network topologies based on electrical switches, Fat Tree (FT) [5], [11] is one of the most common architecture to build large-scale computing systems and it is usually built with some level of oversubscription to reduce the number of inter-rack switches and cables. For a three-tier FT network with 64-port electrical switches, the scalability of FT is well below 100 000 nodes.

Flattened Butterfly (FB) [12], [13] is another possible interconnect architecture. While FB can guarantee a smaller diameter compared to FT network (three hops instead of five), its scalability is also limited by the radix of the electrical switches, and the maximum number of interconnected nodes is also well below 100 000 nodes, assuming to use 64-port electrical switches.

3-D Torus [14] is a well-known network topology for HPC systems. Its scalability is not limited by the switch radix but the network diameter grows quickly with the number of nodes.

The dragonfly topology [15] is a hierarchal network topology that creates groups of nodes with an extremely high virtual radix. Ref. [15] does not provide details regarding the intra or the inter-group network, but it specifies that there is at least one global link between any pair of groups. The authors also developed routing algorithms that assure that the minimal path between any pair of nodes uses at most one global link. While any arbitrary network can be used for the inter-group and intra-group networks, to minimize the worst-case hop count these networks should be all-to-all, i.e. all the routers in a group should be connected to each other directly (otherwise, some local forwarding is required via the intra-group network, which increases the hop count).

A recent trend in large-scale interconnection architectures is to use servers/computing nodes with embedded switches with a large number of ports to create directly connected networks such as PERCs from IBM [16]. Embedding the switch/router directly in the node reduces the latency of the initial/final hop. Keeping the diameter of the network small helps to reduce the network's average latency which is strongly correlated with the average number of hops between each source and destination pairs. Researchers at HP have even proposed [17], [18] integrating photonics into high radix switches in order to meet the needs of future high performance systems.

Concerning the use of optical technologies to improve performance or scalability of electrical interconnect architectures, Helios [20] is a popular hybrid electrical/optical switch architecture for modular datacenters that uses slow optical switches for optical circuit switching and a legacy FT topology for electrical packet switching. In addition, Helios' design does not address the challenges of scalability. The main focus is to offload the high-bandwidth traffic to the optical network.

Calabretta *et al.* proposes and experimentally demonstrates a flat and distributed optical packet switching interconnect architecture for inter-cluster communication [20]. The intra cluster network still makes use of legacy electrical networks.

The AWGR based optical switches and optical routers with packet switching capability have been investigated for a number of years [21]–[23], [11], [24]–[28]. Previous works [11], [21], [24]–[28] mainly focused on applying AWGR techniques in access and telecommunication networks. However, the wavelength parallelism on AWGR outputs to achieve output queuing was not explored in those designs due to the need to ensure wavelength consistency on the input and output fibers. Ref. [9] introduced LIONS (initially named as DOS) and the use of AWGR to achieve output queuing in the optical domain. Ref. [10] presented a significant improvement over the architecture originally proposed in [9] by introducing the all-optical distributed control plane based on the optical-TOKEN technique and the all-optical NACK. The simulation comparison in [29] shows that LIONS provides low-latency and high-throughput switching without saturation even at very high (∼90%) input load. Ref. [30] introduces an optical switch architecture based on a combination of wavelength routing in AWGR and delivery-coupling-switch fabricated on a single planar lightwave circuit. The architecture can potentially scale beyond 512-ports, but no network performance studies have been reported.

Research efforts in silicon photonics and photonic integrated circuits can bring advantages in terms of power consumption and bandwidth. For instance, researchers at HP [31] proposed photonic I/Os to increase the per-port switch bandwidth up to 320 Gb/s by using 32 wavelengths per fiber and having each wavelength operating at 10 Gb/s. Ref. [32] shows a commercial development of a router chip with 168 optical I/Os and an aggregate bandwidth of 1.34 Tb/s. Ref. [33]–[35] report advancements and demonstrations of silicon photonics (SiP) WDM solutions. Silicon photonic LIONS in the configuration shown in Fig. 1(c) with 32 transmitters and 32 receivers utilizing 8×8 AWGR with $k_t = k_r = 4$ have been recently demonstrated [35] on a compact $1.2\ \mathrm{mm} \times 2.4\ \mathrm{mm}$ silicon-on-insulator platform.

Traffic statistics and patterns in data centers are rarely discussed in most publications, but some recent papers show that a strong communication locality within racks exists [36] and that some applications (e.g. map reduce) running in data centers exhibit all-to-all communication patterns [37]. In this paper we discuss a scalable and flat interconnection topology supporting both locality and all-to-all communications.

III. H-LION: Interconnect Architecture Exploiting AWGRs in Hierarchy

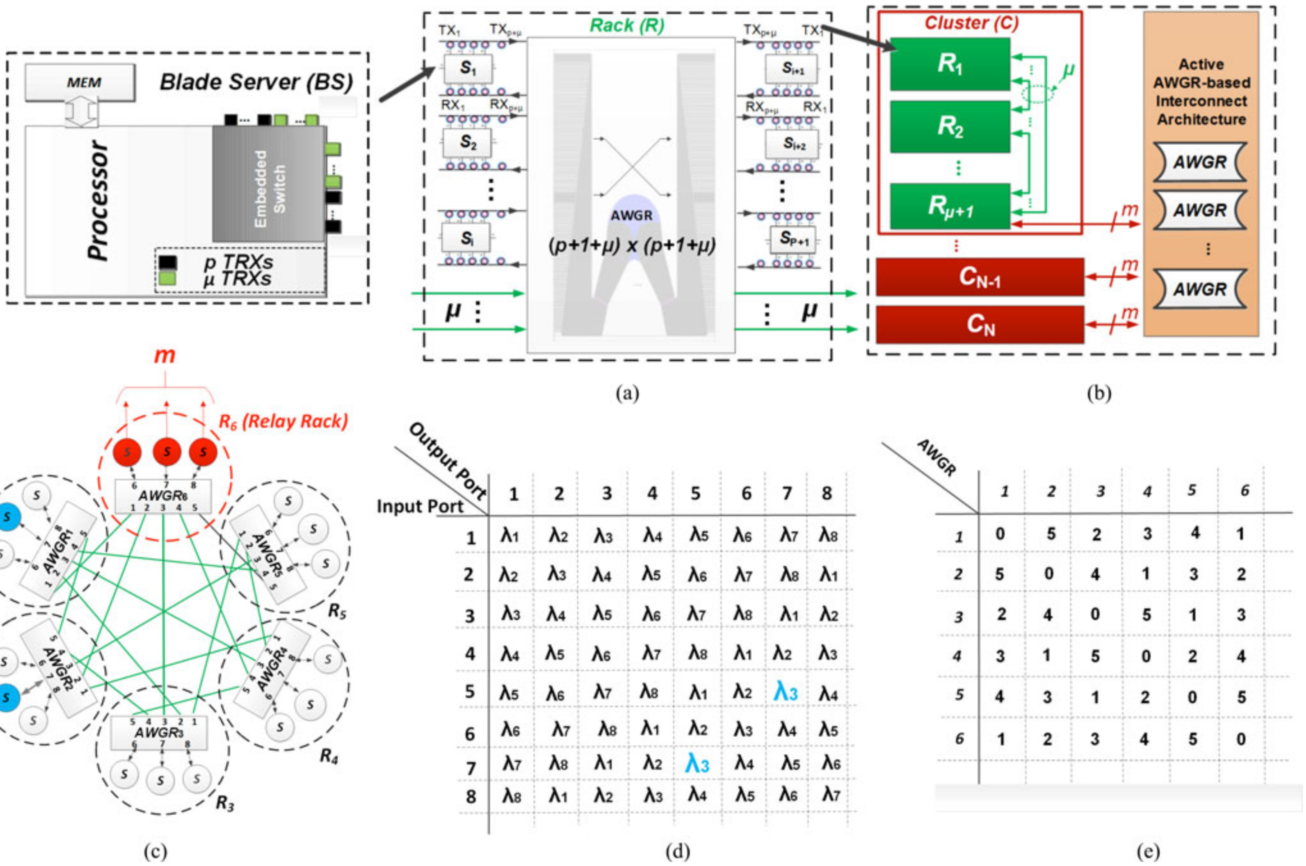
Fig. 2(a), (b), and (c) illustrate the proposed hierarchical optical interconnect architecture (H-LIONS), which is organized in Servers (***S***), Racks (***R***) and Clusters (***C***). The proposed interconnect architecture exploits passive AWGR (passive-LIONS) all-to-all interconnection for intra rack and intra cluster communication, while inter cluster communication is provided by a flat active AWGR switch topology with distributed control (Thin-CLOS, Ref. [10]).

The computing node, i.e. Server (***S***), has one embedded electrical switch with WDM optical I/Os. The embedded switch performs the routing function within the server to forward the packets to the proper TRX port based on the destination address of the packets. The $\boldsymbol{p}$ TRXs are used to connect with other ***S***s in the same ***R*** through all-to-all interconnection of the passive AWGR within the same ***R***. The $\boldsymbol{\mu}$ TRXs are used to connect with ***S***s in different ***R***s within the same Cluster. This interconnection also makes use of wavelength-specific TRXs and each ***S*** can act as a relay (the maximum number of hops in one ***C*** is two).

Different ***R***s in a ***C*** are directly connected with each other supporting all-to-all interconnection using multiple passive AWGRs shown in Fig. 2(c). There are $\boldsymbol{p+1}$ ***S's*** in each ***R***, and $\boldsymbol{\mu+1}$ ***R's*** in each ***C***. Note that, one of the racks in the cluster, i.e. R_6 in Fig. 2(c), acts as a relay rack for the traffic going in and out from the cluster. In total, the relay Rack has $\boldsymbol{\mu} \times (\boldsymbol{p}+1)$ TRXs for distributing the incoming traffic to the intra cluster nodes, and there can be up to $\boldsymbol{p} \times (\boldsymbol{p}+1)$ TRXs for inter cluster communication. We define $\boldsymbol{m}$ to be the number of TRXs for communication with other clusters ($\boldsymbol{m} \leq \boldsymbol{p} \times (\boldsymbol{p}+1)$). Fig. 2(a) illustrates the architecture of ***R*** with all-to-all connectivity among $\boldsymbol{p+1}$ ***S***s. The inset of Fig. 2(a) shows the architecture of the server (*S*) with the embedded switch and $\boldsymbol{p}+\boldsymbol{\mu}$ wavelength-specific transceivers (TRXs).

To perform correct intra-cluster routing, we use an all-to-all topology built following a symmetric matrix shown in Fig. 2(e). If we label AWGRs as $\mathrm{AWGR}_0, \mathrm{AWGR}_1, \ldots, \mathrm{AWGR}_\mathrm{n}$ and define the element a_{ij} in the matrix as the sequence number of the port to connect AWGR_i with AWGR_j, then we define the symmetric matrix as:

$$\begin{cases} a_{ij} = a_{ji} & 1 \leq i \leq n \quad 1 \leq j \leq n \; i \neq j; \\ a_{ij} = 0 & i = j \\ a_{1j} \cap a_{2j} \cap \ldots \cap a_{nj} = \emptyset & 1 \leq j \leq n \\ a_{i1} \cap a_{i2} \cap \ldots \cap a_{in} = \emptyset & 1 \leq i \leq n. \end{cases} \tag{1}$$



Input Port \ Output Port	1	2	3	4	5	6	7	8
1	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
2	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_1
3	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_1	λ_2
4	λ_4	λ_5	λ_6	λ_7	λ_8	λ_1	λ_2	λ_3
5	λ_5	λ_6	λ_7	λ_8	λ_1	λ_2	λ_3	λ_4
6	λ_6	λ_7	λ_8	λ_1	λ_2	λ_3	λ_4	λ_5
7	λ_7	λ_8	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
8	λ_8	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7

AWGR	1	2	3	4	5	6
1	0	5	2	3	4	1
2	5	0	4	1	3	2
3	2	4	0	5	1	3
4	3	1	5	0	2	4
5	4	3	1	2	0	5
6	1	2	3	4	5	0

Fig. 2. Optical interconnect architecture based on passive AWGR interconnection for intra rack and intra cluster communication and active AWGR-based switching topology for inter Cluster communication. S: Server; R: Rack; C: Cluster; AWGR: arrayed waveguide grating router. (a) All-to-all Rack architecture (Inset: Server architecture with embedded switch.-). (b) Cluster and Full System architecture. Green lines represent the intra cluster links. Red lines represent the inter cluster links from/to the relay Rack ($R_{\mu+1}$). (c) Intra Cluster interconnection between racks (all to all): in this example $p = 2$ and $\mu = 5$. (d) Example of routing table for an eight-port AWGR. (e) Symmetric matrix for intra Cluster AWGR interconnections.

For example, Fig. 2(c) shows that the fifth port of AWGR$_1$ connects to the fifth port of AWGR$_2$, and Fig. 2(e) illustrates its corresponding symmetric matrix. In this all-to-all topology, packets from the blue server (attached to the seventh port of AWGR 1) can be correctly forwarded to the blue server (attached to the seventh port of AWGR 2) using the wavelength λ_3 as its routing information.

The orange active AWGR-based interconnect architecture in Fig. 2(b) handles inter cluster interconnection using red links from the relay clusters as shown in Fig. 2(c). Fig. 2(b) further describes this active AWGR-based interconnect architecture for cluster-to-cluster communication.

A. *Inter Cluster Interconnection Network—Thin CLOS*

Inter-cluster communication requires even more scalable architecture as there can be hundreds of clusters. In this case, instead of utilizing the fully connected architecture, whose scalability would be limited by the $\boldsymbol{m}$ TRXs, the inter Cluster interconnection makes use of active AWGR-based switches [10] based on a flat and distributed all-optical interconnect architecture called Thin-CLOS (see Fig. 3). Here, $\boldsymbol{m} \times \boldsymbol{W}$ clusters are divided into $\boldsymbol{m}$ separate groups and are interconnected with each other using $\boldsymbol{m^2}$ active $\boldsymbol{W} \times \boldsymbol{W}$ AWGR switches. As discussed above, each cluster already contains $\boldsymbol{\mu} + \mathbf{1}$ *R*s, with one rack acting as relay rack. There are $\boldsymbol{m}$ links ($\boldsymbol{m} \leq \boldsymbol{p} \times (\boldsymbol{p} + 1)$) coming out from the relay rack. Each cluster has $\boldsymbol{m}$ TXs and $\boldsymbol{m}$ RXs, as shown in Fig. 3. Each cluster's TX/RX pair with index *i* connects with other clusters belonging to the group with index *i,* through one *W*-port active AWGR switch. The topology diameter is one as long as the intra-cluster network routing selects the proper TX that allows to reach the desired cluster without additional relay operations.

The active $\boldsymbol{W} \times \boldsymbol{W}$ AWGR switches used for building the Thin CLOS topology can attain distributed control plane regardless of the scale of the full system by utilizing the TONAK architecture [10] exploiting (a) an all-optical TOKEN technique for contention resolution and (b) an all-optical NACK technique as physical layer flow control to guarantee retransmission and no packet-loss.

B. *Network Oversubscription and Diameter*

As discussed above, there are $\boldsymbol{\mu} \times (\boldsymbol{p} + 1)$ links going from the intra-cluster servers to the relay servers in the relay rack. Therefore, the network oversubscription is equal to $\boldsymbol{\mu} \times (\boldsymbol{p} + 1)/\boldsymbol{m}$. All the simulation results and scalability analysis in Section IV assume a network without oversubscription. Thus,

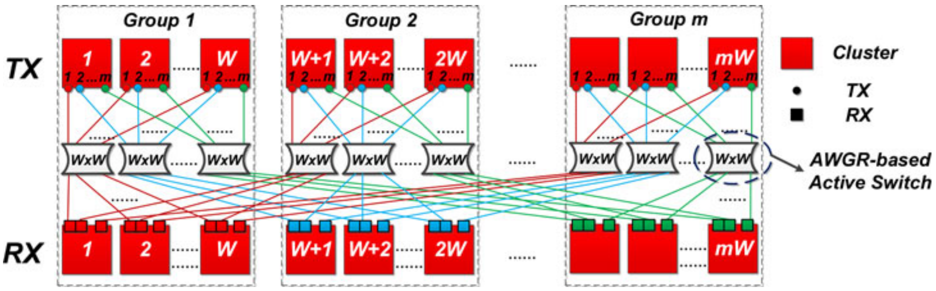


Fig. 3. Thin CLOS topology interconnecting $m \times W$ clusters: m groups, with each group containing W clusters. W is the port-count of active AWGR switches.

TABLE I
DESIGN PARAMETERS TO ACHIEVE A NETWORKS SCALE ≥ 100 000 SERVERS WITHOUT OVERSUBSCRIPTION

W	p	μ	m	N
32	20	3	63	127 008
16	19	4	80	102 400
12	18	5	95	108 300
10	17	6	108	116 640
8	16	7	119	113 288
7	15	8	128	114 688

$\boldsymbol{m} = \boldsymbol{\mu} \times (\boldsymbol{p} + 1)$. For a fair comparison, the scalability analysis and simulation results for the FT network also assume a network without oversubscription.

The network diameter for the proposed architecture is equal to 5 (two hops for intra cluster, one hop for inter cluster, two hops for intra cluster). The FT network has also a network diameter equal to five in the case of a three-tier FT.

C. Scalability Analysis

The radix of the AWGR is an important consideration in determining the scalability of the interconnect networks. While high port count silicon-photonic AWGRs with 512-ports have been experimentally demonstrated [38], in-band crosstalk and the number of wavelengths become challenging. Noting that 32-port AWGRs are commercially available and that 64-port AWGRs were demonstrated in as early as 2003 [7] with <-40 dB crosstalk and $\sim$6 dB insertion loss, we believe that the use of AWGRs with a port count value up to 64 is a viable solution both for passive AWGR interconnection and active AWGR switches (the scalability of active AWGR switches mainly depends on the crosstalk level of the passive AWGR in it). Therefore, we will limit our discussions to AWGRs with port counts not higher than 64.

In the proposed architecture, the number of servers is $\boldsymbol{N} = [(\boldsymbol{p} + 1) \times \boldsymbol{\mu}] \times \boldsymbol{m} \times \boldsymbol{W}$ (we do not count the servers in the relay rack as computing nodes), the number of passive AWGR is $\boldsymbol{N}_{\mathbf{AWGR\_}p} = (\boldsymbol{\mu} + 1) \times \boldsymbol{m} \times \boldsymbol{W}$, and the number of active AWGR switches is $\boldsymbol{N}_{\mathbf{AWGR\_}a} = \boldsymbol{m}^2$. Table I shows some possible combinations of ***p, μ*** and ***m*** parameters to achieve a network scale beyond 100 000 servers. The assumptions here are: (a) oversubscription = 1, which means $\boldsymbol{m} = (\boldsymbol{p} + 1) \times \boldsymbol{\mu}$, and (b) number of wavelengths ≤ 32. The limit on the number of wavelengths translates into the following conditions: 1) $\boldsymbol{W} \leq 32$ for the Thin CLOS architecture; 2) $\boldsymbol{p} + \boldsymbol{\mu} \leq 32$.

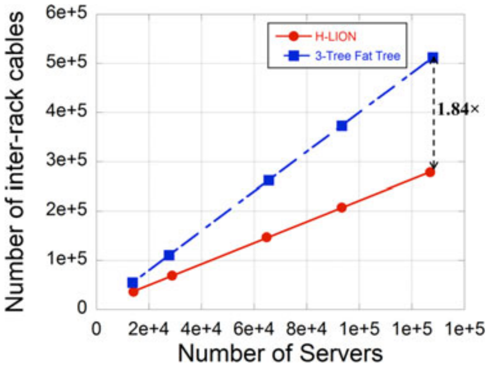


Fig. 4. Number of inter-rack cables as function of the number of servers for H-LION and FT architectures without oversubscription.

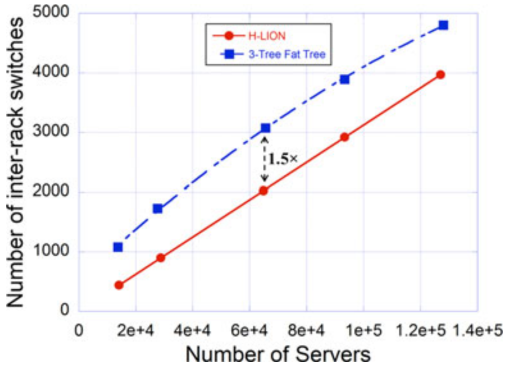


Fig. 5. Number of inter-rack switches as function of the number of servers for H-LION and FT architectures without oversubscription.

Figs. 4 and 5 show the number of inter-rack bidirectional links and switches, respectively. We perform the comparison with a standard m-port n-tree FT networks without oversubscription, where m is the port count of the single electrical switch and the number of layers in the FT network is n = 3. The number of servers, cables and switches in Figs. 4 and 5 are calculated for the following values of m = 38, 48, 64, 72, 80. The total number of servers is $\boldsymbol{N}_{\mathrm{FT}} = 2 \times (\mathrm{m}/2)^{\mathrm{n}}$; the total number of switches is $\boldsymbol{N}_S = (2n - 1) \times (\mathrm{m}/2)^{\mathrm{n}-1}$, while the total number of inter-rack switches is $\boldsymbol{N}_{\mathrm{S}} - \boldsymbol{N}_{\mathrm{FT}}/(\mathrm{m}/2)$. Finally, the number of cables is equal to $\boldsymbol{N}_C = (\boldsymbol{N}_S \times \mathrm{m} - \boldsymbol{N}_{\mathrm{FT}})$.

TABLE II
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
VC's Buffer size	4 Kbyte	SerDes RX TX Delay (GTH)	113 ns ≈ 14 (Cycles)
Delay in Receiver Module (RX)	4 (Cycles)	Wire Delay per meter	5 ns
Delay in Transmit Module (TX)	4 (Cycles)	Delay in Passive AWGR	1 ns
Switching Delay (Crossbar)	5 (Cycles)	Active AWGR Switching	12 ns

In the proposed architecture, the number of active switches is $\boldsymbol{N}_{\mathbf{AWGR}\_a} = \boldsymbol{m}^2 = (\boldsymbol{p}+\mathbf{1})^2 \times \boldsymbol{\mu}^2$, while the number of cables is $\boldsymbol{N}_C = 2 \times \boldsymbol{W} \times [(\boldsymbol{p}+\mathbf{1}) \times \boldsymbol{\mu}]^2 + [(\boldsymbol{p}+1) \times \boldsymbol{\mu} \times \boldsymbol{W}] \times [\boldsymbol{\mu} \times (\boldsymbol{\mu}+1)]$. The first term represents the number of active switches multiplied for the AWGR port count (the factor "2" account for both input and output ports).

The curves in Figs. 4 and 5 are calculated assuming a constant $\boldsymbol{\mu} = 3$.

For the parameters indicated in Figs. 4 and 5, the proposed architecture at a scale of >60 000 nodes guarantees up to 1.84 × savings in terms of inter-rack cables, and up to 1.5 × savings in terms of inter-rack switches. Note that, based on the formula above, the lower the $\boldsymbol{\mu}$, the higher the savings in terms of number of cables.

IV. PERFORMANCE STUDIES

We developed a cycle-accurate parallel network simulator in C/C + + to study the network performance of the proposed network. The traffic pattern is uniform random following Bernoulli process for the arrival time of the packets. Each transmitter has an input buffer organized as virtual output queues, each one having size of 4 kB. Latency simulation results are shown in terms of average values and 95% confidence intervals. Packet size is 256 Bytes. Line rate is 10 Gb/s. All the latency parameters are listed in Table II. One cycle is equivalent to 8 ns. Each node generates 250 000 packets for each simulation run.

A. Cluster Performance

As explained above, a lower $\boldsymbol{\mu}$ guarantees higher savings in terms of number of cables. Note that, a lower $\boldsymbol{\mu}$ guarantees also better cluster network performance. In fact, when $\boldsymbol{p} > \boldsymbol{\mu}$, there are fewer racks and more servers per rack compared to the case where $\boldsymbol{p} < \boldsymbol{\mu}$. Since intra-rack is fully connected, it should be clear that $\boldsymbol{p} > \boldsymbol{\mu}$ case requires fewer relay operations, then guaranteeing higher performance.

We ran a set of simulations for different combinations of $\boldsymbol{p}$ and μ, with $(\boldsymbol{p}+1) \times \boldsymbol{\mu}$ constant and equal to 128. Fig. 6 shows the Cluster performance in terms of average packet latency with 95% confidence interval as function of the offered load. The packet length used is 256 B. As expected the two configurations with $\boldsymbol{p} > \boldsymbol{\mu}$ perform better than the one with $\boldsymbol{p} < \boldsymbol{\mu}$. For $\boldsymbol{\mu} = 0$, the latency curve would appear flat since this would correspond to the contention-less case where all the 128 nodes are fully connected.

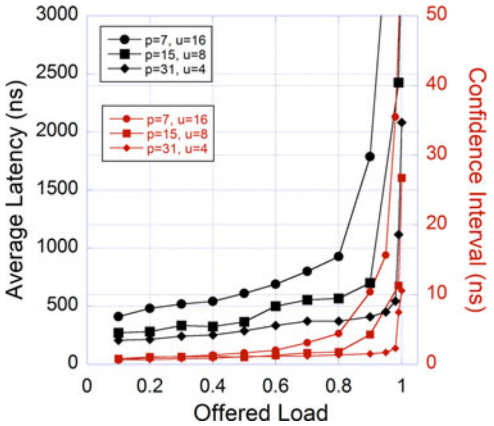


Fig. 6. Intra-cluster average latency with 95% confidence interval as function of the offered load for difference combination of cluster parameters. Number of servers in the cluster is constant and equal to 128. Traffic is uniform random with Bernoulli distribution for the packets arrival time.

B. Full System Performance and Comparison With FT Network

The Thin-CLOS architecture makes use of active AWGR-based switches exploiting an all-optical NACK technique [39] to notify the sender when a packet is not granted so that the sender can prepare for retransmission. The longer the distance, the higher the delay to receive the NACK and retransmit the packet. As a consequence, latency and throughput of the system can be slightly affected by both distance and packet size (see [39] for more details). However, for this simulation we fixed the cable length for inter-cluster connection to 50 m, while the packet size is equal to 256 B. Fig. 7 shows full system simulation results for the proposed architecture (H-LION) and a legacy three-tier FT architecture. Both architectures are without oversubscription. The simulation scale of the proposed architecture is 7200 nodes with the following parameters: $\boldsymbol{p} = 9$, $\boldsymbol{\mu} = 3$ and $\boldsymbol{W} = 8$. The FT network is a 30-port, three-tree network with 6750 nodes. The performance of H-LION and FT architecture are quiet similar, with FT and H-LION saturating at about 80% and 90% of offered load, respectively.

V. EXPERIMENTAL TESTBED STUDIES

A. Intra-Cluster Experiment

Fig. 8(top) shows a hardware testbed for a cluster composed of two racks and a total of 8 $\boldsymbol{S}$s ($\boldsymbol{p} = 3$ and $\boldsymbol{\mu} = 1$). Each $\boldsymbol{S}$ is implemented here with one Xilinx VC709 FPGA board equipped

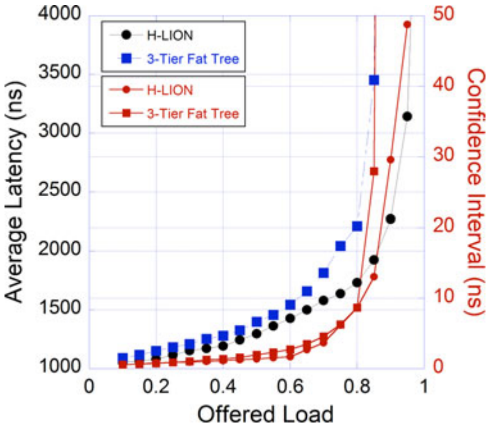


Fig. 7. Full-system H-LION network performance 7200 servers ($p = 9$, $\boldsymbol{\mu} = 3, m = 30, W = 8$). Performance of a 6750-server three-tree FT network without oversubscription is shown for comparison. Traffic has uniform random distribution with Bernoulli process for the packets arrival time. Average packet latency with 95% confidence interval as function of the offered load.

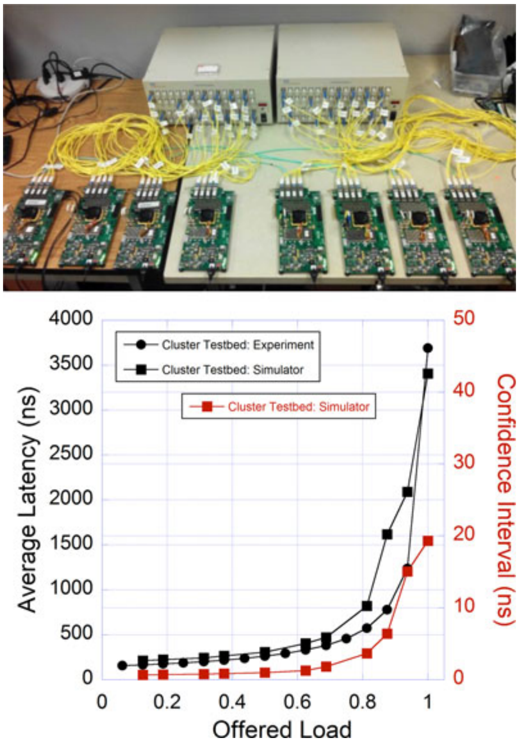


Fig. 8. (Top) Picture of the hardware demo using eight VIRTEX 7 FPGA baords and two 32-port AWGRs. (Bottom) Latency vs. Throughput performance comparison between hardware and simulations.

with a Virtex 7 chip. Each FPGA implements the traffic generation function and the embedded switch (see Fig. 9 for more details) and has four 10 Gb/s WDM SFP$^+$ transceivers (TRXs): three for all-to-all intra rack communication and one for inter rack communication (within the same cluster). The wavelengths used for the experiment are in the range 1546.04–1561.04 nm, with a 0.4 nm (50 GHz) frequency grid. The network traffic generated by the FPGAs is converted in the optical domain, wavelength-routed by two 32-port AWGRs (one per rack) and received by the destination servers (FPGAs). The insertion loss of the AWGRs is 8 dB. The TRXs have an output power of ∼3 dBm and a sensitivity of −26 dBm at BER $= 10^{-12}$, as shown in the black BER curve reported in Fig. 11.

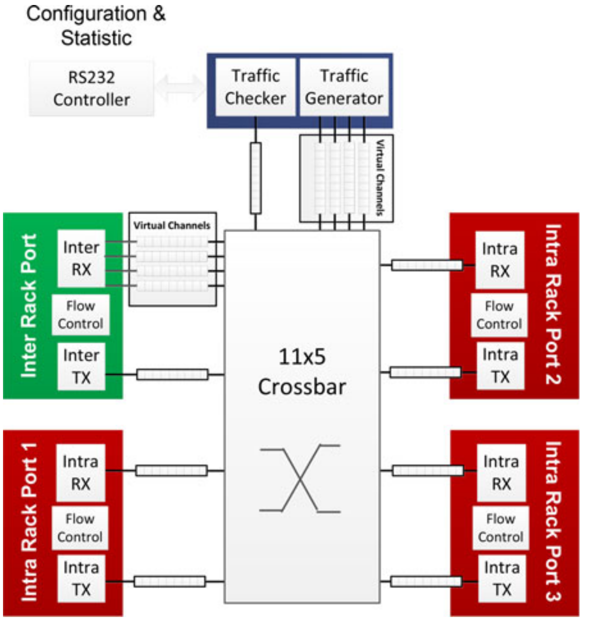


Fig. 9. Architecture of the embedded switch inside the VIRTEX 7 FPGA.

Fig. 8(bottom) shows that, under uniform random traffic, the normalized system-wide network throughput of our cluster testbed is higher than 97% with the latency below 364 ns, only limited by the FPGA speed. In addition, using the parameters measured from the testbed (see Table II), a very similar performance curve can be obtained from the cycle-accurate simulator. This is an important fact that validates the accuracy of the simulator. The packet size used for the simulation and experiment is 256 B. The slight difference between the simulation and experiment results is likely due to the differences in the characteristics of the random traffic generation in the servers. Flow-control mechanism implemented in the testbed and simulator avoids any packet loss.

Fig. 9 shows the architecture of the emulated embedded switch and traffic generator inside the FPGA VIRTEX 7 chip. The core switching fabric is an 11 × 5 crossbar [including the virtual channels (VCs)]. We use a matrix arbiter which can implement one-cycle fair arbitration to perform the arbitration in the 11 × 5 crossbar. The Inter Rack port (the green port in Fig. 9) has four VCs: one for the packets directed to the server, and three VCs for redirecting in the incoming packets to the other three nodes in the same rack. The traffic generator has also four VCs, one for each of the Intra Rack ports and Inter Rack port. In an all-to-all fashion, the Intra Rack ports (red ports) do not need VCs since these ports only communicate with the injection port (traffic checker).

In the demo of Fig. 8, $\boldsymbol{p} = 3$ and $\boldsymbol{\mu} = 1$, while the AWGR radix is 32. In the case where the two racks would be fully populated ($\boldsymbol{p} + 1 + \boldsymbol{\mu} = 32$), it is reasonable to expect some power penalty due to accumulated inband crosstalk at each

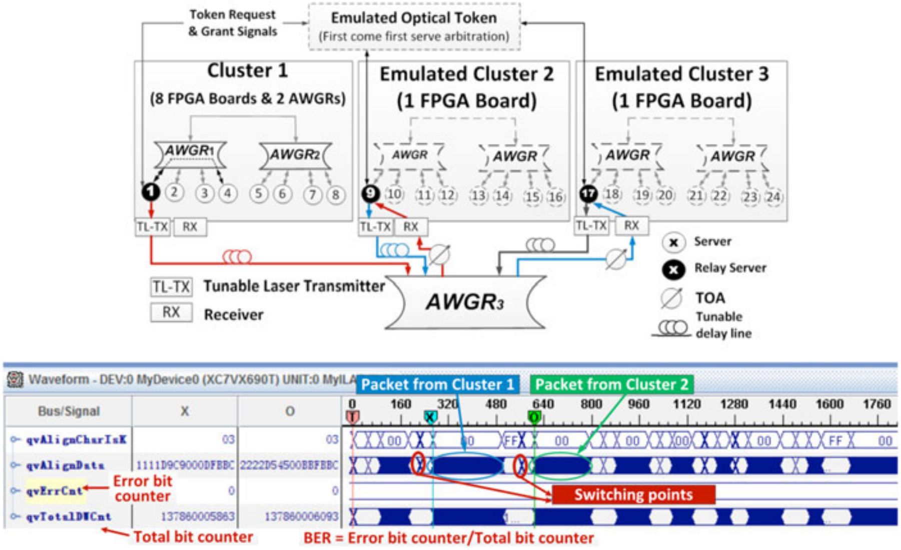


Fig. 10. (Top) Inter-cluster communication experiment setup. TL-TX: tunable transmitter; TOA: tunable optical attenuator; AWGR: arrayed waveguide grating router. (Bottom) Received waveforms for inter-cluster communication from C1 and C2 to C3.

of the AWGR output ports. However, for the 32-port commercial device used in this experiment, the crosstalk level is below −35 dB, which can still guarantee error-free operation [40], [41].

B. Inter-Cluster Experiment

This sub-section shows an experimental demonstration of inter-cluster communications. Fig. 10(top) shows the three-Cluster experiment setup. C1 is the cluster described above with eight FPGA boards. FPGA1 (server #1) acts now also as the relay server for inter-cluster communication. C2 and C3 are emulated with two additional FPGA boards. Each cluster has one 10 Gb/s NRZ-OOK TX with a fast tunable laser (TL) and a 10 Gb/s SFP + RX. Each cluster can communicate with the other ones through wavelength routing in AWGR3 (eight ports, 200 GHz spacing, 8 dB insertion loss). We emulated the optical TONAK system (to solve the network contention) with an FPGA that implements a first-come first-served arbitration scheme. The transmitted packets have random length (64 ∼ 2048 B) and contain different portions of $2^{31} - 1$ PRBS.

Fig. 11 shows the measured bit-error-rate (BER) data. First, we show a contention-less scenario where server #4 in C1 sends packets to server #9 in C2 (red curve) and server #9 in C2 sends packets to server #17 in C3 (blue curve). The BER was measured with a BER tester implemented in the FPGAs. C2-C3 BER curve shows 1-dB penalty due to non-optimum extinction ratio at the tunable TX output. Second, we show a scenario with network contention where C1 and C2 send packets to C3 with random arrival time. To avoid resynchronization in the FPGA GTH receiver, tunable optical delay lines in each transmission path guarantee clock-phase matching of the packets at AWGR output 3. Note that, as for any optical packet switch system, an actual implementation would require to use burst-mode clock recovery technology (as demonstrated in [42]–[44]) to recover the clock on a packet-by-packet basis. Fig. 10(bottom) shows a snapshot of the received packets captured with Xilinx Chip-Scope analyzer tool. Fig. 11 shows measured BER curves for intra cluster communications (black curve) and hierarchical inter cluster communications where the green curve is a plot for inter-cluster C1 and C2 switching to C3, the red curve is for C1 switching to C2, and the blue curve is for C2 switching to C3. Note that, when errors occur at the switching point, the byte misalignment in the GTH RX damages the full packet. As a consequence the BER increases significantly when the RX power is below −26.5 dBm. In all cases error-free (BER $< 10^{-12}$) results have been obtained for RX power >−24 dBm.

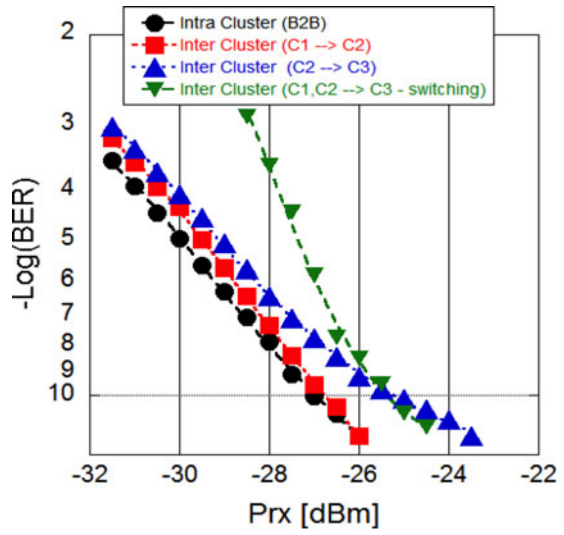


Fig. 11. BER measurements for intra cluster communications and hierarchical inter cluster communications. Black dots: BER for packets going from S4 to S1; Red squares: BER for packets going from S1 to S9. Blue triangles: BER for packets going from S9 to S17. Green triangles: BER for packet switching from C1 and C2 to C3.

VI. Conclusion

We proposed a scalable optical interconnect architecture based on a topology enabled by wavelength routing in AWGRs and by servers with embedded switches and optical I/Os. The proposed architecture exploits AWGR-based all-to-all *passive* interconnection for intra-rack and intra-cluster communication and AWGR-based *active* switching for inter-cluster communication, supporting scalability beyond 100 000 nodes. Network simulation results show throughput as high as 90% in case of traffic with uniform random distribution. A testbed demonstration shows system-wide network throughput at 97% for eight compute nodes and two clusters, with measured Latency-vs.-Throughput data plots closely matching the data plots from the cycle-accurate simulator developed for the AWGR-based interconnect network.

As discussed above, the AWGR radix is limited by the crosstalk, which is also affected by the AWGR channel spacing and the number of wavelengths used in the system. However, Thin-CLOS topology can provide the same all-to-all topology and logical performance as one large $N \times N$ AWGR with N wavelengths but using m groups of $W \times W$ AWGRs with W wavelengths only ($W = N/m$). The fewer number of wavelengths allows to solve the crosstalk issue while requiring more fibers.

Areas of studies to further improve the network performance include control plane and routing algorithms in combination with flexible bandwidth technologies to dynamically adjust the bandwidth of the different links according to the real traffic demand. This can help reducing the congestion between hotspots while saving energy for the links with low utilization.

References

[1] StatisticBrain. (2014). *Google Annual Search Statistics* [Online]. Available: http://www.statisticbrain.com/google-searches/

[2] StatisticBrain. (2014). *Facebook Statistics*. [Online]. Available: http://www.statisticbrain.com/facebook-statistics/

[3] G. Amdahl, "Validity of the single processor approach to achieving large-scale computing capabilities," in *Proc. Spring Joint Comput. Conf.*, 1967, pp. 483–485.

[4] P. M. Kogge and T. J. Dysart, "Using the TOP500 to trace and project technology and architecture trends," presented at the International Conf. High Performance Computing, Networking, Storage and Analysis, Seattle, WA, USA, 2011.

[5] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 38, pp. 63–74, 2008.

[6] F. Abel, C. Minkenberg, I. Iliadis, T. Engbersen, M. Gusat, F. Gramsamer, and R. P. Luijten, "Design issues in next-generation merchant switch fabrics," *IEEE-ACM Trans. Netw.*, vol. 15, no. 6, pp. 1603–1615, Dec. 2007.

[7] S. Kamei, M. Ishii, M. Itoh, T. Shibata, Y. Inoue, and T. Kitagawa, "64 x 64-channel uniform-loss and cyclic-frequency arrayed-waveguide grating router module," *Electron. Lett.*, vol. 39, pp. 83–84, 2003.

[8] B. Glance, I. P. Kaminow, and R. W. Wilson, "Applications of the integrated waveguide grating router," *J. Lightw. Technol.*, vol. 12, no. 6, pp. 957–962, Jun. 1994.

[9] X. Ye, P. Mejia, Y. Yin, R. Proietti, S. J. B. Yoo, and V. Akella, "DOS—A scalable optical switch for datacenters," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst.*, 2010, pp. 1–12.

[10] R. Proietti, Y. Yin, R. Yu, C. J. Nitta, V. Akella, C. Mineo, and S. J. B. Yoo, "Scalable optical interconnect architecture using AWGR-based TONAK LION switch with limited number of wavelengths," *J. Lightw. Technol.*, vol. 31, no. 24, pp. 4087–4097, Dec. 15, 2013.

[11] M. C. Chia, D. K. Hunter, I. Andonovic, P. Ball, I. Wright, S. P. Ferguson, K. M. Guild, and M. J. O'Mahony, "Packet loss and delay performance of feedback and feed-forward arrayed-waveguide gratings-based optical packet switches with WDM inputs-outputs," *J. Lightw. Technol.*, vol. 19, no. 9, pp. 1241–1254, Sep. 2001.

[12] W. J. Dally and B. P. Towles, *Principles and Practices of Interconnection Networks*. Amsterdam, The Netherlands: Elsevier, 2004.

[13] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: A cost-efficient topology for high-radix networks," *ACM SIGARCH Comput. Archit. News*, vol. 35, pp. 126–137, 2007.

[14] S. Horiguchi and T. Ooki, "Hierarchical 3D-torus interconnection network," in *Proc. Int. Symp. Parallel Archit., Algorithms Netw.*, 2000, pp. 50–56.

[15] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," *ACM SIGARCH Comput. Archit. News*, vol. 36, pp. 77–88, 2008.

[16] B. Arimilli, R. Arimilli, V. Chung, S. Clark, W. Denzel, B. Drerup, T. Hoefler, J. Joyner, J. Lewis, L. Jian, N. Nan, and R. Rajamony, "The PERCS high-performance interconnect," in *Proc. IEEE 18th Annu. Symp. High Perform. Interconnects*, 2010, pp. 75–82.

[17] N. Binkert, A. Davis, N. P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, and J. H. Ahn, "The role of optics in future high radix switch design," in *Proc. 38th Annu. Int. Symp. Comput. Archit.*, 2011, pp. 437–447.

[18] R. G. Beausoleil, "Large-scale integrated photonics for high-performance interconnects," *ACM J. Emerging Technol. Comput. Syst.*, vol. 7, p. 6, 2011.

[19] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: A hybrid electrical/optical switch architecture for modular data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 40, pp. 339–350, 2010.

[20] W. Miao, J. Luo, S. Di Lucente, H. Dorren, and N. Calabretta, "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," *Opt. Express*, vol. 22, pp. 2465–2472, Feb. 2014.

[21] S. J. B. Yoo, "Optical packet and burst switching technologies for the future photonic Internet," *J. Lightw. Technol.*, vol. 24, no. 12, pp. 4468–4492, Dec. 2006.

[22] M. J. O'Mahony, D. Simeonidou, D. K. Hunter, and A. Tzanakaki, "The application of optical packet switching in future communication networks," *IEEE Commun. Mag.*, vol. 39, no. 3, pp. 128–135, Mar. 2001.

[23] C. Guillemot, M. Renaud, P. Gambini, C. Janz, I. Andonovic, R. Bauknecht, B. Bostica, M. Burzio, F. Callegati, M. Casoni, D. Chiaroni, F. Clerot, S. L. Danielsen, F. i. Dorgeuille, A. Dupas, A. Franzen, P. B. Hansen, D. K. Hunter, A. Kloch, R. Krähenbühl, B. Lavigne, A. L. Corre, C. Raffaelli, M. Schilling, J.-C. Simon, and L. Zucchelli, "Transparent optical packet switching: The European ACTS KEOPS project approach," *J. Lightw. Technol.*, vol. 16, no. 12, pp. 2117–2134, Dec. 1998.

[24] H. Yang and S. J. B. Yoo, "Combined input and output all-optical variable buffered switch architecture for future optical routers," *IEEE Photon. Technol. Lett.*, vol. 17, no. 6, pp. 1292–1294, Jun. 2005.

[25] M. Maier, M. Scheutzow, and M. Reisslein, "The arrayed-waveguide grating-based single-hop WDM network: An architecture for efficient multicasting," *IEEE J. Select. Areas Commun.*, vol. 21, no. 9, pp. 1414–1432, Nov. 2003.

[26] S. Bregni, A. Pattavina, and G. Vegetti, "Architectures and performance of AWG-based optical switching nodes for IP networks," *IEEE J. Select. Areas Commun.*, vol. 21, no. 7, pp. 1113–1121, Sep. 2003.

[27] W. D. Zhong and R. S. Tucker, "Wavelength routing-based photonic packet buffers and their applications in photonic packet switching systems," *J. Lightw. Technol.*, vol. 16, no. 10, pp. 1737–1745, Oct. 1998.

[28] D. Banerjee, J. Frank, and B. Mukherjee, "Passive optical network architecture based on waveguide grating routers," *IEEE J. Select. Areas Commun.*, vol. 16, no. 7, pp. 1040–1050, Sep. 1998.

[29] X. Ye, P. Mejia, Y. Yin, R. Proietti, S. J. B. Yoo, and V. Akella, "DOS—A scalable optical switch for datacenters," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst.*, 2010, pp. 1–12.

[30] K. Sato, H. Hasegawa, T. Niwa, and T. Watanabe, "A large-scale wavelength routing optical switch for data center networks," *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 46–52, Sep. 2013.

[31] N. Binkert, A. Davis, N. P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, and A. Jung Ho,, "The role of optics in future high radix switch design," in *Proc. 38th Annu. Int. Symp. Comput. Archit.*, 2011, pp. 437–447.

[32] K. Hasharoni, "High BW parallel optical interconnects," presented at the Advanced Photonics for Communications, San Diego, CA, USA, 2014, p. PT4B.1.

[33] P. P. Absil, P. De Heyn, P. Dumon, D. Van Thourhout, P. Verheyen, S. Selvaraja, G. Lepage, M. Pantouvaki, M. Rakowski, and J. Van Campenhout, "Advances in silicon photonics WDM devices," *in Proc. SPIE 9010, Next-Generation Optical Networks for Data Centers and Short-Reach Links* 2014, pp. 90100J-1–90100J-7.

[34] Q. Fang, T.-Y. Liow, J. F. Song, K. W. Ang, M. B. Yu, G. Q. Lo, and D.-L. Kwong, "WDM multi-channel silicon photonic receiver with 320 Gbps data transmission capability," *Opt. Exp.*, vol. 18, pp. 5106–5113, Mar. 1, 2010.

[35] R. Yu, S. Cheung, Y. Li, K. Okamoto, R. Proietti, Y. Yin, and S. J. B. Yoo, "A scalable silicon photonic chip-scale optical switch for high performance computing systems," *Opt. Exp.*, vol. 21, pp. 32655–32667, Dec. 30, 2013.

[36] C. Delimitrou, S. Sankar, A. Kansal, and C. Kozyrakis, "ECHO: Recreating network traffic maps for datacenters with tens of thousands of servers," in *Proc. IEEE Int. Symp. Workload Characterization*, 2012, pp. 14–24.

[37] N. Farrington and A. Andreyev, "Facebook's data center network architecture," in *Proc. IEEE Opt. Interconnects Conf.*, 2013, pp. 49–50.

[38] S. Cheung, S. Tiehui, K. Okamoto, and S. J. B. Yoo, "Ultra-compact silicon photonic 512 x 512 25 GHz arrayed waveguide grating router," *IEEE J. Select Topics Quantum Electron*, vol. 20, no. 4, pp. 1–7, Jul./Aug. 2014.

[39] R. Proietti, Y. Yawei, Y. Runxiang, Y. Xiaohui, C. Nitta, V. Akella, and S. J. B. Yoo, "All-optical physical layer NACK in AWGR-based optical interconnects," *IEEE Photon. Technol. Lett*, vol. 24, no. 5, pp. 410–412, Mar. 1, 2012.

[40] H. K. Kim and S. Chandrasekhar, "Dependence of coherent crosstalk penalty on the OSNR of the signal," in *Proc. Opt. Fiber Commun. Conf.*, 2000, vol. 2, pp. 359–361.

[41] H. Takahashi, K. Oda, and H. Toba, "Impact of crosstalk in an arrayed-waveguide multiplexer on N × N optical interconnection," *J. Lightw, Technol.*, vol. 14, no. 6, pp. 1097–1105, Jun. 1996.

[42] R. Yu, R. Proietti, Y. Shuang, J. Kurumida, and S. J. B. Yoo, "10-Gb/s BM-CDR circuit with synchronous data output for optical networks," *IEEE Photon. Technol. Lett.*, vol. 25, no. 5, pp. 508–511, Mar. 1, 2013.

[43] M. Nogawa, K. Nishimura, S. Kimura, T. Yoshida, T. Kawamura, M. Togashi, K. Kumozaki, and Y. Ohtomo, "A 10 Gb/s burst-mode CDR IC in 0.13 μm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. Digest Tech. Papers.*, 2005, vol. 1, pp. 228–595.

[44] L. Jri and L. Mingchung, "A 20-Gb/s burst-mode clock and data recovery circuit using injection-locking technique," *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 619–630, Mar. 2008.

Roberto Proietti received the M.S. degree in telecommunications engineering from the University of Pisa, Pisa, Italy, in 2004, and the Ph.D. degree in electrical engineering from Scuola Superiore Sant'Anna, Pisa, Italy, in 2009. He is currently a Project Scientist with the Next Generation Networking Systems Laboratory, University of California, Davis, CA, USA. His research interests include optical switching technologies and architectures for supercomputing and data center applications, high-spectrum efficiency coherent transmission systems and elastic optical networking, and radio-over-fiber technologies for access networks.

Zheng Cao received the B.S. degree from the Department of Computer Science and Technology, Shandong University, Jinan, China, in 2003, and the Ph.D. degree in computer architecture from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2009. He is currently an Associate Professor in the Institute of Computing Technology, Chinese Academy of Sciences. His research interest includes large-scale network simulation, low-latency and scalable optical interconnects for data centers, and high-performance computing architectures.

Christopher J. Nitta received the Ph.D. degree in computer science from the University of California, Davis, CA, USA, in 2011. He is an Adjunct Assistant Professor at the University of California, Davis. His research interests include network-on-chip technologies, embedded system and RTOS design, and hybrid electric vehicle control.

Yuliang Li received the B.E. degree in computer science and technology from Tsinghua University, Beijing, China, in 2014. He is currently working toward the Ph.D. degree at the Department of Computer Science, University of Southern California, CA, USA.

S. J. Ben Yoo (S'82–M'84–SM'97–F'07) received the B.S. degree in electrical engineering with distinction, the M.S. degree in electrical engineering, and the Ph.D. degree in electrical engineering with a minor in physics, all from Stanford University, Stanford, CA, USA, in 1984, 1986, and 1991, respectively. He currently serves as a Professor of electrical engineering at the University of California at Davis (UC Davis), CA. His research at UC Davis includes high-performance all-optical devices, systems, and networking technologies for future computing and communications. Prior to joining UC Davis in 1999, he was a Senior Research Scientist at Bellcore, leading technical efforts in optical networking research and systems integration. He participated ATD/MONET testbed integration and a number of standardization activities including GR-2918-CORE, GR-2918-ILR, GR-1377-CORE, and GR-1377-ILR on dense WDM and OC-192 systems. Dr. Ben Yoo received the DARPA Award for Sustained Excellence in 1997, the Bellcore CEO Award in 1998, the Outstanding Mid-Career Research Award (UC Davis, 2004), and the Outstanding Senior Research Award (UC Davis, 2011). He is a Fellow of the Optical Society of America,