Experimental Demonstration of Flexible Bandwidth Optical Data Center Core Network With All-to-All Interconnectivity

Experimental Demonstration of Flexible Bandwidth Optical Data Center Core Network with All-to-All Interconnectivity

Z. Cao, R. Proietti, M. Clements and S. J. B. Yoo, *Fellow, IEEE, Fellow, OSA*

***Abstract*— This paper proposes and demonstrates a flexible-bandwidth optical interconnect architecture for data centers exploiting wavelength routing in arrayed waveguide grating routers (AWGR) and fast tunable lasers. The proposed architecture provides hierarchical all-to-all connectivity with low contention and dynamic interconnection reconfiguration for higher bandwidth provisioning between hot spots.**

An eight-cluster core network experiment testbed with hierarchical all-to-all interconnection shows 1.77× throughput increase and 1.19× network energy efficiency improvement in the case of inter-cluster hot-spot traffic, while guaranteeing more than 97% throughput for the portion of the traffic with uniform random distribution.

***Index Terms*—Data center networking, optical interconnects, arrayed waveguide grating routers, flexible bandwidth, elastic optical networks.**

I. Introduction

Scalability of networks interconnecting beyond tens of thousands of servers inevitably leads to introducing hierarchical network architectures. The 3-tier tree-based network architecture shown in Figure 1 is one of the most commonly used in data centers due to its scalability and cost-effectiveness.

The highest tier network (core network) design is the most critical for the full system network performance among all layers. Numerous research results [1, 2] have shown that the core network is the most utilized layer, containing hot-spot links. The hot-spot traffic usually occupies around 25% of the links and changes over time [1]. The problem is even more severe if data center networks are based on topologies utilizing switches with relatively small radix numbers, incapable of supporting many-to-many or all-to-all interconnection. Since the hot-spot traffic can cause network congestion and seriously degrade the global communication performance, it is important to optimize the network resources to cope with the hot-spot traffic in the core network. Dynamically allocating more bandwidth between hot–spot links can reduce the network congestion and improve the overall network performance in terms of latency, throughput, and energy-consumption.

Legacy electrical core networks make use of multi-path routing [3, 4] to allocates multiple non-shortest paths to the hot-spot traffic. However, such solution has two drawbacks: (1) hot-spot traffic spreads over multiple multi-hop paths and potentially increases the number of network congestion point; (2) the issue of the out-of-order transmission/arrival becomes more serious.

In the context of telecom networks, flexible bandwidth (FB) transceiver technologies [5, 6] have been widely studied and experimentally demonstrated. For data center applications, ref. [7, 8] proposed flexible-bandwidth optical-interconnect architectures which can achieve variable bandwidth by using optical orthogonal frequency division multiplexing (OFDM) technique. However, the link bandwidth cannot still exceed the maximum transceiver bandwidth, making it not possible to account for the high peak-to-average-traffic-ratio.

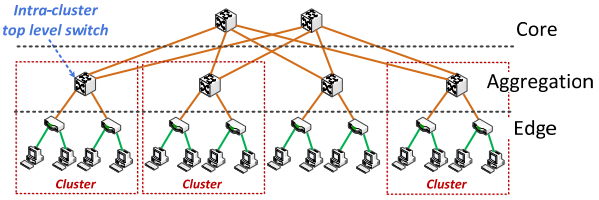

Figure 1. 3-tier tree-based datacenter network [9].

Here, we propose a new flexible-bandwidth optical core network that can dynamically increase the number of direct links between hot spots, thus increasing the communication bandwidth. The proposed architecture provides hierarchical all-to-all low-contention communication for average bandwidth traffic by using AWGR's intrinsic all-to-all connectivity. Moreover, the proposed network can dynamically perform connectivity reconfiguration at nanoseconds scale by using a channel bonding (CB) technique. The CB technique (see Section III) exploits wavelength routing in Arrayed Waveguide Grating Routers (AWGRs [10]), and fast tunable lasers [11].

This paper provides, for the first time to our knowledge, an experimental demonstration of flexible-bandwidth optical networking in data centers with an all-to-all interconnection topology. Networking experiments show that the CB technique leads to a 1.77× throughput increase for hot-spot traffic, and 1.19× improvement in energy efficiency without reducing the background (cold) traffic performance.

The remainder of this paper is organized as follows: Section II introduces the related work of both flexible bandwidth networks and AWGR-based networks. Section III introduces

Manuscript received October 27, 2014. This work was supported in part under DoD Agreement Number: W911NF-13-1-0090.

Z. Cao, R. Proietti, M. Clements and S. J. B. Yoo are with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: zhcao@ucdavis.edu; sbyoo@ucdavis.edu).

the proposed dynamic channel bonding technique that enables fast flexible bandwidth adjustment. Section IV introduces the proposed network architecture. Section V discusses network experiment studies using a hardware prototype of eight Clusters emulated with field programmable gate array (FPGA) boards. Section VI concludes the paper.

II. Related Work

Since data center networks account for 23% of the total IT power consumption [12], flexible bandwidth networks [7, 8, 13] have been proposed to reduce power consumption. Ref. [7, 8] used OFDM technology to dynamically adjust the links' line rate according to the real-time bandwidth requirements. Ref. [13] proposed a mechanism to detect and turn off the idle links. These works focus on saving power only for the lightly-loaded or idle links. The solution proposed in this paper aims to improve the overall network energy efficiency by providing more bandwidth for the hot-spot traffics. Many works [3, 4, 14-16] proposed adaptive/multi-path routing mechanisms to find additional existing paths for hot-spot traffics. With a total different approach, the architecture proposed in this paper dynamically creates new paths for hot-spot traffics. In other words, lightly-loaded or idle links can be reconfigured to boost the bandwidth of the hot-spot links. Furthermore, the proposed architecture is based on simpler wavelength assignment scheme rather than to utilize relatively complex OFDM optoelectronics.

The proposed solution fully exploits the wavelength routing property of AWGR. There have been several studies on AWGR-based datacenter networks. By using the cyclic wavelength routing of AWGR, DOS [17] introduced the use of AWGR to achieve output queuing in the optical domain. TONAK LION [18] further improved DOS' performance and scalability by introducing an all-optical control plane and an all-optical acknowledgement technique. Ref. [19] proposed a passive AWGR to implement flexible inter-rack interconnection by using OFDM technology. In addition to the above works focusing on single-stage networks, there have been also some AWGR-based multi-stage network studies. Petabit [20] proposed a three-stage Clos network based on AWGR and tunable wavelength converters (TWC). Hi-LION [21] proposed an optical interconnection architecture that includes a passive AWGR-based local hierarchical all-to-all network and a global AWGR-based mesh-like network. Ref. [22] proposed a different hierarchical all-to-all architecture to achieve flexible bandwidth adjustment at the network core layer. Its main goal is to boost the interconnection bandwidth between hot spot links while providing ultra-low contention communication for average bandwidth traffic. This paper extends the work in [22] proposing a routing algorithm and TRXs selection algorithm.

III. Dynamic Channel Bonding Technique

Dynamic channel bonding (CB) technique allows to dynamically, rapidly, and flexibly assign additional bandwidth to certain pairs of Clusters upon demand of hot spot traffic. AWGRs and fast tunable lasers (<10ns switching time) are key enabling technologies to achieve dynamic bandwidth adjustment for the proposed core network. As shown in Figure 2, a $N\times N$ AWGR can provide all-to-all communication among N ports in a flat topology without contention when using N wavelengths. Different connectivity between input and output ports can be achieved by injecting different wavelengths into the AWGR input ports.

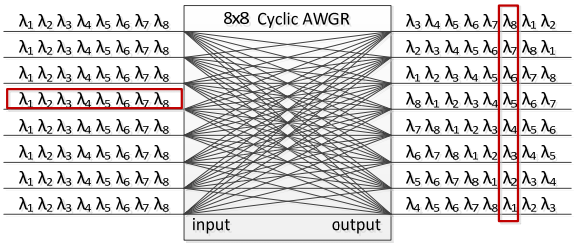

Figure 2. Example of 8×8 AWGR.

When hot spots form, the network control plane will tune certain transceivers' (TRXs) wavelengths to increase the number of connections between the hot Clusters. Figure 3 illustrates the concept of channel bonding in details. Originally, with proper wavelength assignment, the four Clusters (C_0~C_3) are connected with each other in an all-to-all fashion. When the bandwidth requirement between C_0 and C_3 exceeds the peak bandwidth of a single link, TRX for $C_0 \rightarrow C_2$ with λ_1 (blue link) is tuned to λ_2 (red link) for $C_0 \rightarrow C_3$. Then we perform similar tuning procedure for $C_3 \rightarrow C_0$. Eventually, the bandwidth between C_0 and C_3 is doubled, and the two red TRXs are bonded to perform the transmission. As a consequence, the direct connection $C_0 \rightarrow C_2$ is not available anymore, and the traffic between C_0 and C_2 needs to be relayed by C_1 or C_3.

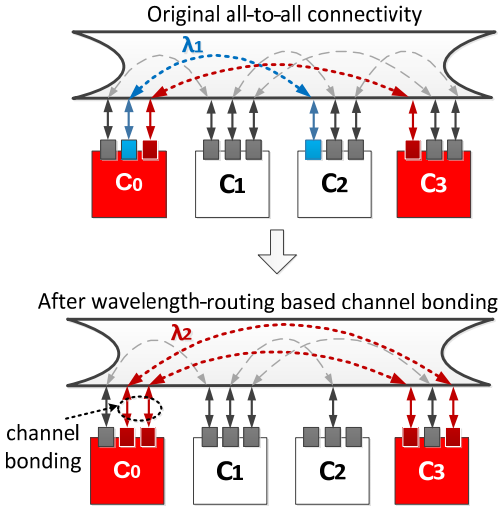

Figure 3. The concept of wavelength-routing-based channel bonding.

IV. Flexible Flat All-to-All Optical Core Network

A. System Architecture

We group Clusters into $\boldsymbol{\mu}$ Regions and each Region contains $\boldsymbol{p}$ Clusters and one AWGR. Each intra-cluster top level switches (TLS) uses $\boldsymbol{p}$**-1** TRXs for intra-region communication and $\boldsymbol{\mu}$**-1** TRXs for inter-region communication. To effectively achieve the flexibility and reconfiguration of the topology between the hot spots, all the TRXs make use of fast tunable lasers (TL) which can achieve fast wavelength tuning in < 10 nanoseconds. A control plane interfaces with all the TLSs in the Clusters and

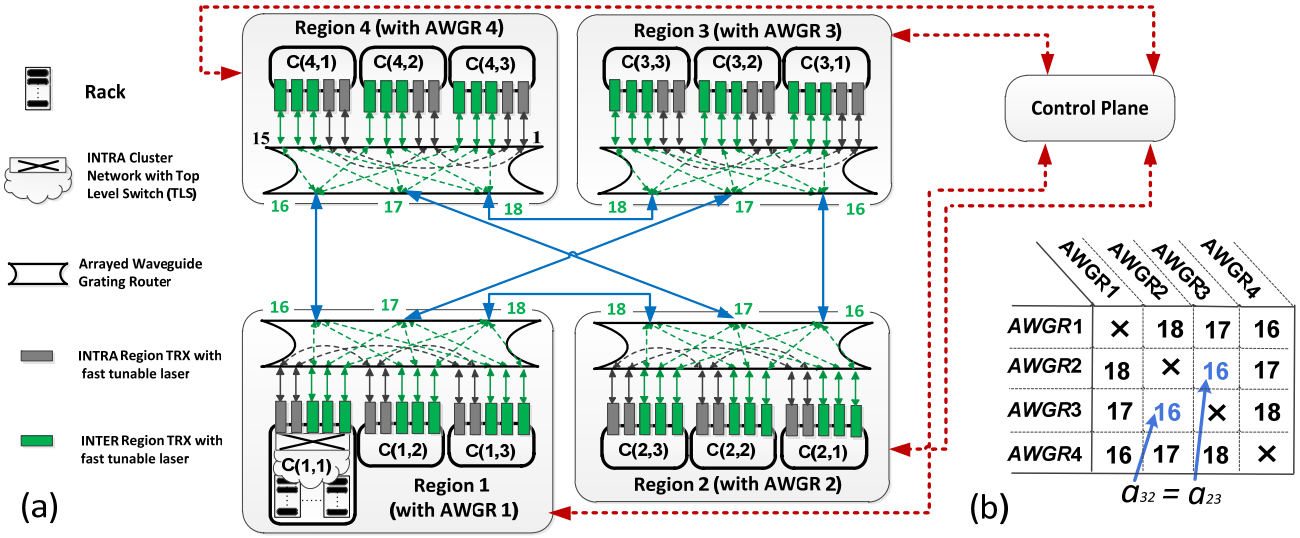

Figure 4. (a) Proposed data center flat all-to-all optical interconnect architecture based on wavelength routing in AWGR (b) an example of symmetric matrix.

controls the TLSs flow tables. The architecture scales to $p \times \mu$ Clusters and the radix of AWGR is $p \times (p+\mu-2)+\mu-1$. The full system can reach, for example, 103,680 servers using six 65-port AWGRs when $p = 6$, $\mu = 6$. The number of Servers per Rack is 40, and there are 72 Racks per Cluster. Figure 4 shows an example of proposed core network with three Regions and three Clusters per Region.

This paper proposes a default all-to-all interconnection for both intra-region and inter-region communication to support both high scalability and connectivity.

Default intra-region all-to-all connectivity: if there is no over-peak hot-spot traffic, p-1 different wavelengths will be assigned to each Cluster's p-1 TRXs (grey TRXs in Figure 4(a)) and all-to-all connectivity between p Clusters is achieved by using wavelength routing in AWGR.

Inter-region all-to-all connectivity is achieved by connecting μ AWGRs with μ -1 fibers in an all-to-all pattern (blue lines in Figure 4(a)). In order to transmit data between two regions without changing wavelengths, the all-to-all topology must be carefully designed by taking advantage of AWGR's routing table. We propose a *symmetric-matrix all-to-all topology* where any AWGR pair interconnects by the same port number. So, the topology's connection matrix is a symmetric matrix. To facilitate the description of the symmetric matrix, we label the AWGRs in Regions as $\{AWGR_1, AWGR_2, \ldots, AWGR_\mu\}$ and define the element a_{ij} in the matrix as the sequence number of the port to connect $AWGR_i$ with $AWGR_j$. Then, the symmetric matrix can be generated as follows:

$$\begin{cases} a_{ij} = a_{ji} & 1 \le i \le \mu \;\; 1 \le j \le \mu \;\; i \ne j \\ a_{ij} = NULL & i = j \\ a_{1j} \cap a_{2j} \cap \ldots \cap a_{\mu j} = \varnothing & 1 \le j \le \mu \\ a_{i1} \cap a_{i2} \cap \ldots \cap a_{i\mu} = \varnothing & 1 \le i \le \mu \end{cases} \quad (1)$$

Figure 4(b) is the symmetric matrix for the network in Figure 4(a). For example, $AWGR_2$ and $AWGR_3$ are using port 16 to connect with each other.

B. Routing under Default Hierarchical All-to-all Connectivity

The routing is performed by a combination of optical wavelength routing in AWGR and electrical packet switching in the intra-cluster top level switch (TLS). Under the default hierarchical all-to-all connectivity, the communication between Clusters in the same Region is performed by the contention-free wavelength routing in AWGR. In terms of the inter-region communication, Clusters attached to the same numbered port of AWGRs can communicate with each other without changing wavelengths. Other communications need at most one-time forwarding performed by the TLSs.

For an M-port AWGR, according to its cyclic routing table, the wavelength to interconnect port i and port j is $\lambda_{[(i+j) \bmod M]}$. So, in Figure 4(a), $C(4,1)$ uses two tunable lasers (grey ones) with $\lambda_{[(11+1) \bmod 18]}$ (λ_{12}) and $\lambda_{[(12+6) \bmod 18]}$ (λ_0) to communicate with $C(4,2)$ and $C(4,3)$ respectively. Regarding the inter-region communication, $C(4,1)$ uses the other three tunable lasers (green ones) with λ_{11}, λ_{13}, and λ_{15} to directly communicate with $C(1,1)$, $C(2,1)$, and $C(3,1)$, respectively. If $C(4,1)$ sends a packet to $C(2,3)$, the packet arrives first at $C(2,1)$ with λ_{13}. Then, $C(2,1)$ forwards this packet to $C(2,3)$ with λ_0. The maximum hop count for inter-region hierarchical all-to-all network is one. In fact, any switching operation happens only in the TLSs while the AWGR is passive and contention-free with a latency only determined by the speed of light. Experimental results in Section V prove that the hierarchical all-to-all network can achieve almost 100% throughput for average-bandwidth traffic with uniform random distribution.

C. Tuning TRX Selection Algorithm

As anticipated in Section III, after performing channel bonding, the default all-to-all connectivity is broken, and there will be no direct connection between certain pairs of Clusters. In this case, additional forwarding is needed. For example, in Figure 3(bottom), the background traffic from $C_0 \rightarrow C_2$ must be forwarded by C_1. So, to guarantee the background traffic performance, the TRXs to be tuned are selected based on following rules:

1) All the Clusters are still reachable after the channel bonding operation using the TRX;

2) Choice of the TRXs will introduce little additional forwarding to non-hot Clusters and will not overload the forwarding links;
3) The TRXs have relatively light traffic load.

As introduced in Section III.B, some Clusters can communicate with each other with a direct link, while others need one-time forwarding path (some inter-region communications). If the hot-spot path is for an inter-region communication with one-time forwarding (contains two links), we will perform the TRX selection twice, one selection for each link.

```
TRXSelect (WCM(NxN), s, d)
WCM(NxN): weighted connection matrix for N Clusters
s: source Cluster ID
d: destination Cluster ID
Begin
# Select a group of light utilized (lower than average utilized rate) TRXs
  from all the TRXs attached to Cluster s as the candidates to be tuned
    avgRate = sum_{i=0}^{N} w_si / (N-1)   i != s
    Foreach (w_si ≤ avgRate): Push(LowRateSet, i)

# From LowRateSet, select the proper TRX to be tuned
    finalTRX = 0;
    finalRate = 1000;
    finalReachable = FALSE;
    Foreach TRX in LowRateSet:
    Begin
        t = Pop(LowRateSet);
        miniRate = 1000;
        isReachable = FALSE;
        # Check if Cluster s can still reach Cluster t after tuning the TRX
        # Get the forwarding path with minimum utilization rate
        for (j=0; j< N; j = j+1)
        Begin
            If ((w_sj != 1000) && (w_jt != 1000))
            Begin
                isReachable = TRUE;
                if (miniRate > max(w_sj, w_jt)) miniRate = max(w_sj, w_jt);
            End
        End
        # Calculate utilized rate of forwarding link after adjustment
        # If the forwarding link will be overloaded, then current
          TRX used for communication between Cluster s and
          Cluster t cannot be tuned.
        If ((miniRate+w_st) >100) isReachable = FALSE;
        # Select the one that can still maintain Cluster reachable and introduce
            less load on the forwarding path
        If (isReachable) finalReachable = TRUE;
        If (isReachable == TRUE & ((miniRate+ w_st) < finalRate))
        Begin
            finalTRX = t;
            finalRate = miniRate + w_st;
        End
    End
    If (finalReachable) return finalTRX;
End
```

Figure 5. Pseudo code of TRX selection algorithm

Implementation of the selection algorithm for the tuning TRX utilizes a weighted connection matrix (WCM). If the system contains N Clusters, then the WCM is a $N \times N$ matrix and each element w_{ij} is the utilization rate of the direct connection between Cluster i and Cluster j. Hence, normally, $w_{ij} \in [0,100]$, but if there is no direct link between Cluster i and Cluster j, then we set w_{ij} = MAX to label it as a break path (e.g. MAX=1000). Based on this matrix, we propose the TRX selection procedure for hot spots on direct path (source Cluster ***s*** and destination Cluster ***d***), as shown in Figure 5.

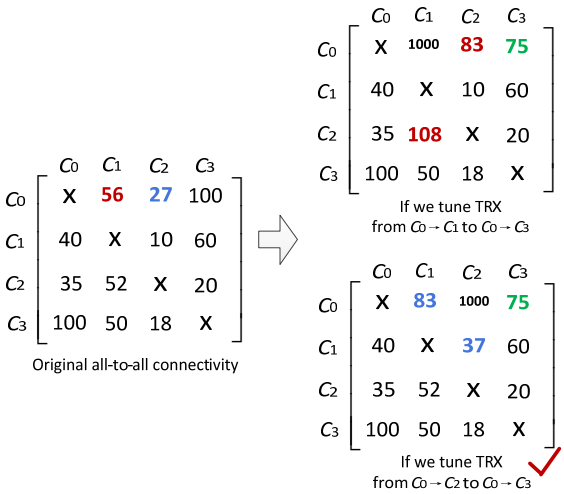

Figure 6. An example of using weighted connection matrix

Figure 6 shows the usage of the weighted connection matrix during the procedure of increasing bandwidth between $C_0 \rightarrow C_3$ (see Figure 3). In Figure 6 (left), the average utilization rate of TRXs in C_0 is (27+56+100)/3 = 61, hence TRXs to C_1 and C_2 are candidates for tuning. However, as shown in Figure 6 (right top), if we tune TRX from $C_0 \rightarrow C_1$ to $C_0 \rightarrow C_3$, then the traffic from $C_0 \rightarrow C_1$ (56) will be added to $C_0 \rightarrow C_2$ (52+56=108) and $C_2 \rightarrow C_1$ (27+56=83). Since the link $C_0 \rightarrow C_2$ is overloaded, the TRX for $C_0 \rightarrow C_1$ should not be tuned. Instead, as shown in Figure 6(right-bottom), the TRX of $C_0 \rightarrow C_2$ is the correct one to be tuned.

V. Two-Region Eight-Cluster Network Experiment Testbed

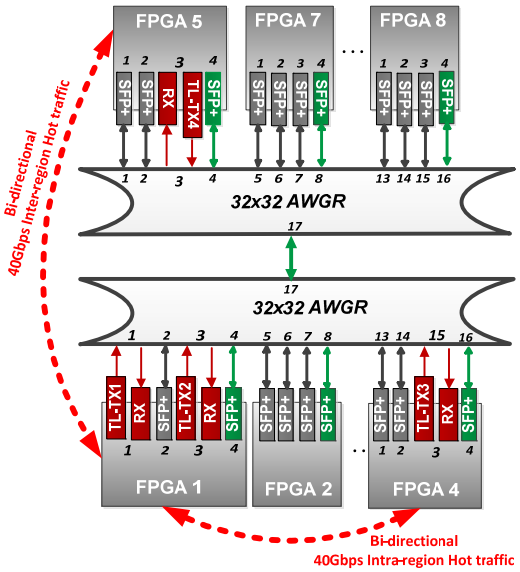

Figure 7. Experiment setup of the full system interconnection network. TL-TX: tunable transmitter (10 Gb/s); RX: receiver (10 Gb/s); SFP+: small form pluggable transceiver (10 Gb/s); AWGR: arrayed waveguide grating router; FPGA: field programmable gate array.

Figure 7 shows an eight-cluster experimental setup for the proposed core network architecture, with p=4 and μ=2. Eight Xilinx VC709 boards with high-speed Rocket I/O TRXs at

10Gb/s emulate the eight Clusters. The TRXs are connected to two 32-port AWGRs. The two AWGRs interconnect by a single fiber (carrying WDM signals). The AWGRs' channel spacing is 50 GHz, and their insertion loss is 8 dB. The wavelengths used in the experiment are in the range 1546.04~1561.04nm on a 0.4nm (50 GHz) grid.

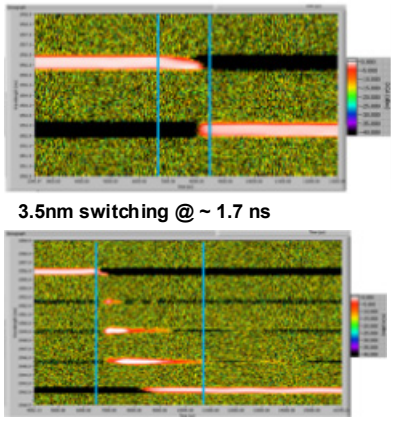

Figure 8. Measured fast tunable laser tuning time.

Each FPGA board has three TRXs for all-to-all intra-Region communication and one TRX for inter-Region communication. TX1 and TX3 in FPGA1 and TX3 in FPGA4 and FPGA5 are implemented with fast tunable lasers with tuning time as short as few nanoseconds (see Figure 8). These tunable lasers allow for fast flexible bandwidth adjustment between hot spots as explained below. All the other TXs and RXs are commercial small form pluggable (SFP) TRXs at 10 Gb/s with -26 dBm RX sensitivity (see Figure 9).

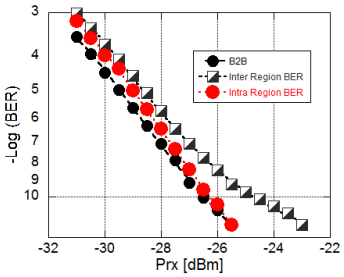

Figure 9. BER of tuned intra-Region and inter-Region links.

Each FPGA acts as an intra-cluster top level switch and traffic generator. As shown in Figure 10, each TLS contains four network ports (each one connected with one of the four TRXs), one injection port with four independent 10Gb/s traffic generators and one 20×8 crossbar. Each network port has a virtual output queuing architecture (four virtual channels) to avoid the Head-of-Line blocking issue. Each injection port can generate up to 40 Gb/s traffic. The crossbar performs switching among 20 input channels and eight output channels.

In order to perform seamless network reconfiguration by channel bonding, the TLS makes use of two routing tables: a working table and a preparing table. The working table is a table used for forwarding the packets in the default hierarchical all-to-all scenario, while the look-ahead table is used for accepting the new table content containing the routing information for the newly reconfigured network. During the network reconfiguration, all the updated routing information is written into the look-ahead table. After the look-ahead table has been updated, we tune the tunable laser and use the look-ahead table as the new working table.

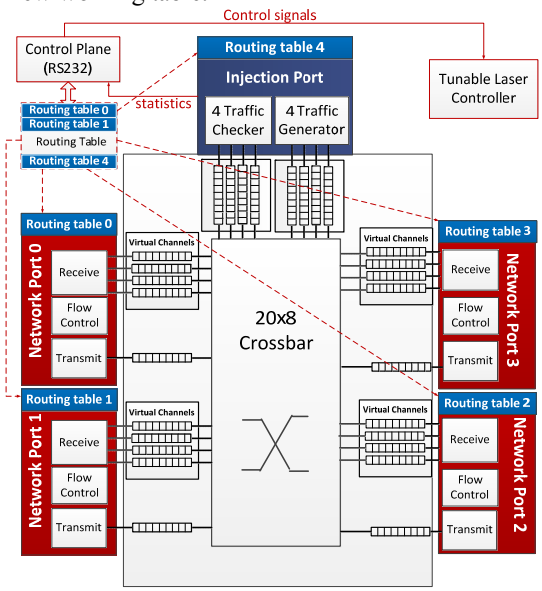

Figure 10. Emulated intra-cluster top level switch

A. *Experimental Network Performance Measurements*

As first, we measured the average bandwidth traffic performance under the default hierarchical all-to-all connectivity by using the experimental platform. Figure 11 shows the simulation results obtained for uniform random traffic. The network saturation point is at around 91% throughput, with a maximum throughput of 97.1%.

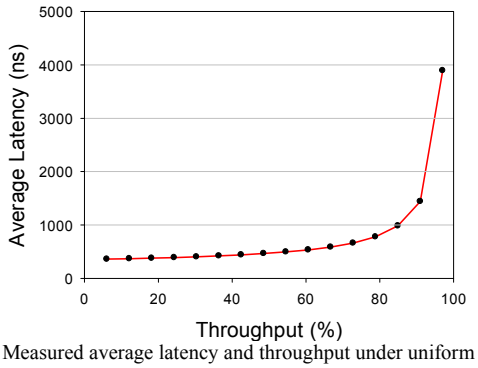

Figure 11. Measured average latency and throughput under uniform random traffic

To demonstrate the hot-spot traffic performance, each FPGA board generates up to 10 Gb/s of background (cold) traffic with uniform random distribution and 40 Gb/s bi-directional hot-spot traffic between two hot-spot points in the network. As shown in Figure 7, we experimentally demonstrated the following four scenarios: 40 Gbps hot-spot traffic between FPGA 1 and 4 (intra-Region) with and without flexible bandwidth adjustment; 40 Gbps hot-spot traffic between FPGA 1 and 5 (Inter-Region) with and without flexible bandwidth adjustment.

The Intra-Region case makes use of TL-TX1 and TL-TX3 in FPGA1 and FPGA4, respectively, while TL-TX2 in FPGA1 and TL-TX4 in FPGA5 are used for the inter Region case. In this experiment, the wavelengths used for TL-TXs are shown in

错误!未找到引用源。. BER measurements in Figure 9 show error-free operation after tuning the tunable TXs for bandwidth increase between the hot-spot points. Figure 12, Figure 13, Figure 14 and Figure 15 show the experimentally measured statistics for the four scenarios described above.

Table 1. Wavelength allocation of the TL-TXs.

	w/o flexibility	w/ intra-Region flexibility	w/ inter-Region flexibility
TL-TX1	1561.41 nm	1552.50 nm	1561.41 nm
TL-TX2	1559.79 nm	1559.79 nm	1548.08 nm
TL-TX3	1546.04 nm	1552.50 nm	1559.79 nm
TL-TX4	1559.79 nm	1559.79 nm	1548.08 nm

Figure 12 shows that under the original hierarchical all-to-all network, limited by the single link between the hot spots, the accepted hot-spot traffic will keep decreasing as the bandwidth of background traffic increases. On the contrary, with flexible bandwidth adjustment, the performance of hot-spot traffic will cease to decrease, and it stabilizes while achieving up to ~1.77× improvement in accepted hot-spot traffic.

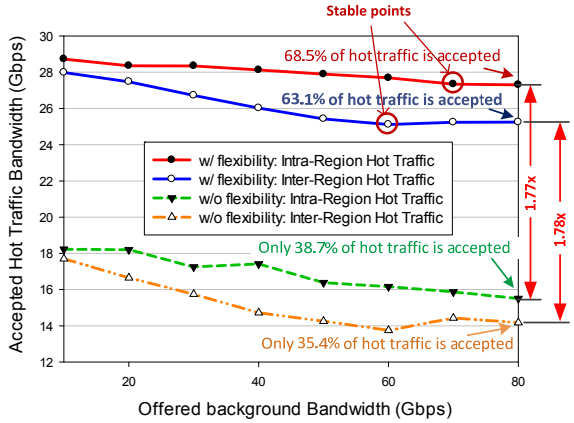

Figure 12. Measured accepted hot-spot traffic bandwidth.

Figure 13 shows that the links reconfiguration dedicated to certain Clusters does not reduce but can increase the accepted background bandwidth by leasing the congestion caused by the hot traffic. As mentioned in Section II, additional forwarding will happen among the other Clusters after reconfiguration. However, as Figure 14 shows, the network with flexibility can achieve similar latency performance, since the queuing time caused by the hot traffic is reduced.

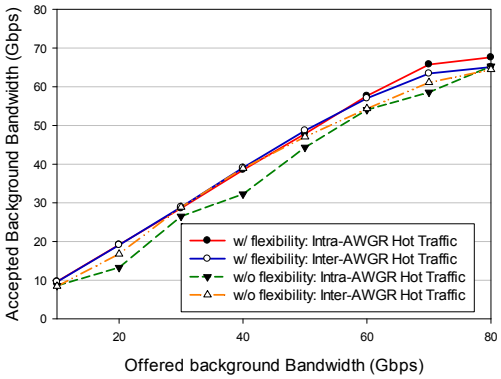

Figure 13. Measured accepted background traffic bandwidth.

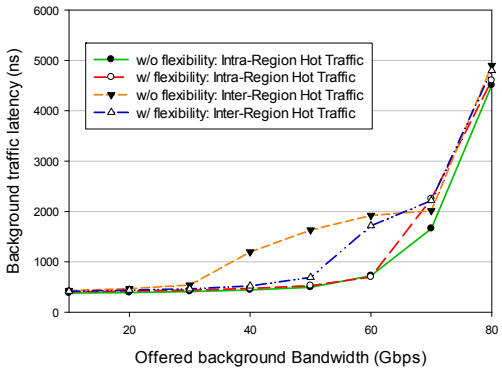

Figure 14. Measured latency of background traffic.

Figure 15 shows the overall network accepted bandwidth. Under the default hierarchical all-to-all interconnection, the maximum bandwidth is very close to the offered background traffic, since the link between hot spots is shared by hot-spot and background traffics. On the contrary, the network with flexibility can break such a bandwidth barrier by dynamically setting up dedicated links for the hot-spot traffics.

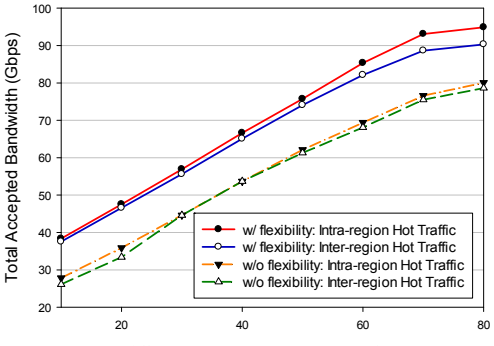

Figure 15. Measured total accepted bandwidth.

The significant improvement in throughput for the hot-spot traffic is also beneficial from an energy efficiency standpoint. Table 2 shows the energy efficiency of our FPGA platform and estimated energy efficiency with silicon photonics (SiP) TRXs. The estimated energy efficiency is calculated with assumption of using SiP tunable laser [23], SiP receiver (3.95 mW) [24] and SiP modulator [25]. According to ref. [23], we can utilize a tunable laser with a tuning span of 38 nm at a tuning power consumption of 26 mW, but utilize only 10 nm maximum tuning, in which case, we expect to consume 3.4 mW (10 nm/38 nm × 26 mW × 50%) on the average for random tuning across 10 nm span. So, each SiP TRX is expected to consume 7.35 mW (3.4 mW + 3.95 mW) static power. In addition, the dynamic power is set to 0.5 mW at 10Gb/s [25]. Thus, each SiP TRX consumes 7.85 mW in total. Since there will be two idle TRXs (no dynamic power) after tuning one pair of TRXs to hot-spot traffic, the power in network with flexibility (3rd column in Table 2) will be slight lower than the one in network without flexibility(2nd column in Table 2). Overall 1.19× energy efficiency improvements can be achieved by introducing flexible bandwidth assignment. If we can turn off the idle TRXs after changing the connectivity, the system will achieve higher energy efficiency.

Table 2. Power comparison on network with intra-Region traffic.

	w/o flexibility	w/ flexibility
Power	112 W	112 W
Total accepted bandwidth	80Gb/s	94.9Gb/s
Energy Efficiency (SFP TRX)	1.400 nJ/bit	1.18 nJ/bit
Power(SiP TRX)	251.2 mW	250.2 mW
Energy Efficiency (SiP TRX)	3.14 pJ/bit	2.64 pJ/bit

Note that, the unique wavelength routing feature of the AWGR is the key factor in support of the above performance improvements. If we replace the AWGR inside each region with fibers to implement all-to-all intra-region connectivity, the performance in case of uniform random traffic (Figure 11) would be the same, but in case of hot-spot traffic, the performance would be the same as shown in Figure 12 for the case named "w/o flexibility". If we replace the two AWGRs in the testbed with a layer of four 8-port electrical switches (basically a one-layer tree network), the network diameter will be the same, but the intra-region all-to-all connectivity will disappear and contending events will take place. As a consequence, the average hop-count and then the average latency will be higher than the proposed network. Again, the electrical network cannot perform dynamic channel bonding in case of hot-spot traffic. In addition, since the electrical switch itself needs additional TRXs at the input/output ports while the AWGR is passive, the power consumption in the electrical network will be higher, even though the tunable laser in the proposed network consumes 2× higher power than the multi-mode TRX in the electrical network as described in [26] .

VI. Conclusion and Future Work

We proposed and demonstrated a core-layer data center optical interconnect architecture with dynamic bandwidth adaptation by wavelength routing in AWGRs and fast tunable lasers. We validated the effectiveness of the proposed solution by experimentally measuring the network statistics on an eight-cluster network testbed with hot-spot traffic. The experimental results show that the proposed architecture can fulfill both average and hot-spot bandwidth requirements, effectively adapting to hot-spot traffic with 1.77× throughput increase for the hotspot links while guaranteeing 1.19× improvements in energy efficiency for the entire network.

The proposed work mainly focused on the optimization for average bandwidth and over-peak hot-spot traffic by simply assigning wavelength channels. This solution achieves coarse-scale optimization. The proposed scheme can also adopt fine-scale optimization and achieve further energy efficiency improvement by exploiting, for example, the previously reported flexible bandwidth techniques [7, 8, 13]. In particular, we can achieve this goal without adding the complexity of OFDM but with a simpler variable-line-rate technique. This work, currently in progress, aims to combine the channel bonding method with variable-line-rate transceivers into the UC Davis testbed to lower the power consumption of lightly-loaded TRXs that cannot be used for channel bonding.

VII. References

[1] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 267-280.

[2] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," *ACM SIGCOMM Computer Communication Review,* vol. 40, pp. 92-99, 2010.

[3] R. Banner and A. Orda, "Multipath routing algorithms for congestion minimization," *IEEE/ACM Transactions on Networking (TON),* vol. 15, pp. 413-424, 2007.

[4] M.-C. Wang, H. J. Siegel, M. A. Nichols, and S. Abraham, "Using a multipath network for reducing the effects of hot spots," *Parallel and Distributed Systems, IEEE Transactions on,* vol. 6, pp. 252-268, 1995.

[5] F. Vacondio, A. El Falou, A. Voicila, C. Le Bouëtté, J.-M. Tanguy, C. Simonneau, J.-L. Pamart, L. Schoch, and O. Rival, "Real-Time Elastic Coherent Muxponder Enabling Energy Proportional Optical Transport," in *Optical Fiber Communication Conference*, 2013, p. JTh2A. 51.

[6] N. Sambo, A. D'Errico, C. Porzi, V. Vercesi, M. Imran, F. Cugini, A. Bogoni, L. Potì, and P. Castoldi, "Sliceable transponder architecture including multiwavelength source," *Journal of Optical Communications and Networking,* vol. 6, pp. 590-600, 2014.

[7] T. Wang, P. N. Ji, C. Kachris, and I. Tomkos, "Energy Efficient Data Center Network Based on a Flexible Bandwidth MIMO OFDM Optical Interconnect," in *Cloud Computing Technology and Science (CloudCom)*, 2012.

[8] C. Kachris and I. Tomkos, "Optical OFDM-based Data Center Networks," *Journal of Networks,* vol. 8, pp. 1488-1496, 2013.

[9] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *Computer Communication Review,* vol. 38, pp. 63-74, Oct 2008.

[10] B. Glance, I. P. Kaminow, and R. W. Wilson, "Applications of the integrated waveguide grating router," *Lightwave Technology, Journal of,* vol. 12, pp. 957-962, 1994.

[11] A. Bhardwaj, J. Gripp, J. E. Simsarian, and M. Zirngibl, "Demonstration of stable wavelength switching on a fast tunable laser transmitter " *IEEE PHOTONICS TECHNOLOGY LETTERS,* vol. 15, pp. 1014-1016, 2003.

[12] GreenDataProject. (2008, *Where does power go?* Available: http://www.greendataproject.org

[13] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "ElasticTree: Saving Energy in Data Center Networks," presented at the Networked Systems Design and Implementation (NSDI), 2010.

[14] X. Wu and X. Yang, "DARD: Distributed Adaptive Routing for Datacenter Networks," in *International Conference on Distributed Computing Systems*, 2012, pp. 32-41.

[15] L. Huang, Q. Jia, X. Wang, S. Yang, and B. Li, "PCube: Improving Power Efficiency in Data Center Networks," in *IEEE International Conference on Cloud Computing (CLOUD)*, 2011, pp. 65-72.

[16] M. Zhang, C. Yi, B. Liu, and B. Zhang, "GreenTE: Power-aware traffic engineering," in *International Conference on Network Protocols*, 2010, pp. 21-30.

[17] X. Ye, P. Mejia, Y. Yin, R. Proietti, S. J. B. Yoo, and V. Akella, "DOS - A scalable Optical Switch for Datacenters," *ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS),* 2010.

[18] R. Proietti, Y. Yawei, Y. Runxiang, 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," *Lightwave Technology, Journal of,* vol. 31, pp. 4087-4097, 2013.

[19] P. N. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos, "Design and evaluation of a flexible-bandwidth OFDM-based intra-data center interconnect," *Selected Topics in Quantum Electronics, IEEE Journal of,* vol. 19, pp. 3700310-3700310, 2013.

[20] K. Xia, Y.-H. Kaob, M. Yangb, and H. Chao, "Petabit optical switch for data center networks," *Polytechnic Institute of New York University, New York, Tech. Rep,* 2010.

[21] Z. Cao, R. Proietti, and S. J. B. Yoo, "Hi-LION: Hierarchical Large-Scale Interconnection Optical Network With AWGRs," *Journal of*

Optical Communications and Networking, vol. 7, pp. A97-A105, 2015.

[22] Z. Cao, R. Proietti, M. Clements, and S. J. B. Yoo, "Experimental Demonstration of Dynamic Flexible Bandwidth Optical Data Center Network with All-to-All Interconnectivity," in *European Conference on Optical Communications (ECOC 2014)* 2014.

[23] T. Chu, N. Fujioka, and M. Ishizaka, "Compact, lower-power-consumption wavelength tunable laser fabricated with silicon photonicwire waveguide micro-ring resonators," *Optics Express,* vol. 17, pp. 14063-14068, 2009.

[24] X. Zheng, D. Patil, J. Lexau, F. Liu, G. Li, H. Thacker, Y. Luo, I. Shubin, J. Li, J. Yao, P. Dong, D. Feng, M. Asghari, T. Pinguet, A. Mekis, P. Amberg, M. Dayringer, J. Gainsley, H. F. Moghadam, E. Alon, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, "Ultra-efficient 10Gb/s hybrid integrated silicon photonic transmitter and receiver," *Optics Express,* vol. 19, pp. 5172-5186, 2011.

[25] P. Dong, S. Liao, D. Feng, H. Liang, D. Zheng, R. Shafiiha, C.-C. Kung, W. Qian, G. Li, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, "Low Vpp, ultralow-energy, compact, high-speed silicon electro-optic modulator," *Optics Express,* vol. 17, pp. 22484-22490, 2009.

[26] S. J. B. Yoo, "Energy Efficiency in the Future Internet: The Role of Optical Packet Switching and Optical-Label Switching," *Selected Topics in Quantum Electronics, IEEE Journal of,* vol. 17, pp. 406-418, 2011.