Abstract—This paper proposes and demonstrates a flexible-bandwidth optical interconnect architecture for data centers exploiting wavelength routing in arrayed waveguide grating routers 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 tested with hierarchical all-to-all interconnection shows $1.77 \times$ throughput increase and $1.19 \times$ network energy efficiency improvement in the case of intercluster hot-spot traffic, while guaranteeing more than $97\%$ throughput for the portion of the traffic with uniform random distribution.

Index Terms—Arrayed waveguide grating routers, data center networking, elastic optical networks, flexible bandwidth, optical interconnects.

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 Fig. 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 $75\%$ 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 allocate multiple non-shortest paths to the hot-spot traffic. However, such a solution has two drawbacks: (1) hot-spot traffic spreads over multiple multi-hop paths and potentially increases the number of network congestion points; (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.

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 arrayed waveguide grating router’s (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 AWGRs [10], and fast tunable lasers (TL) [11].

This paper proposes, 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 $1.77 \times$ throughput increase for hot-spot traffic, and $1.19 \times$ 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 FB networks and AWGR-based networks. Section III introduces the proposed dynamic

Fig. 1. 3-tier tree-based datacenter network [9].
CB technique that enables fast FB 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], FB 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. HILION [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 fast reconfiguration 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 CB TECHNIQUE

Dynamic 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 TLs (<10 ns switching time) are key enabling technologies to achieve dynamic bandwidth adjustment for the proposed core network. As shown in Fig. 2, a N × 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.

When hot spots form, the network control plane will tune certain transceivers’ (TRXs) wavelengths to increase the number of connections between the hot Clusters. Fig. 3 illustrates the concept of CB in details. Originally, with proper wavelength assignment, the four Clusters (C₀ → C₃) are connected with each other in an all-to-all fashion. When the bandwidth requirement between C₀ and C₃ exceeds the peak bandwidth of a single link, TRX for C₀ → C₂ with λ₁ (blue link) is tuned to λ₂ (red link) for C₀ → C₃. Then we perform similar tuning procedure for C₃ → C₀. Eventually, the bandwidth between C₀ and C₃ is doubled, and the two red TRXs are bonded to perform the transmission. As a consequence, the direct connection C₀ → C₂ is no longer available anymore, and the traffic between C₀ and C₃ needs to be relayed by C₁ or C₃.

IV. FLEXIBLE FLAT ALL-TO-ALL OPTICAL CORE NETWORK

A. System Architecture

We group Clusters into μ Regions and each Region contains p Clusters and one AWGR. Each intra-cluster top level switches (TLS) uses p – 1 TRXs for intra-region communication and μ – 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 TLSs which can achieve fast wavelength tuning in <10 nanoseconds. A control plane interfaces with all the TLSs in the Clusters and 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. Fig. 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 Fig. 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 \( \mu \) AWGRs with \( \mu - 1 \) fibers in an all-to-all pattern (blue lines in Fig. 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\(_2\), ..., AWGR\(_p\)\} 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:

\[
\begin{align*}
  a_{ij} &= a_{ji} & 1 \leq i \leq \mu, 1 \leq j \leq \mu, i \neq j \\
  a_{ij} &= NULL & i = j \\
  a_{ij} \cap a_{2j} \cap a_{\mu j} &= \emptyset & 1 \leq j \leq \mu \\
  a_{i1} \cap a_{i2} \cap a_{i\mu} &= \emptyset & 1 \leq i \leq \mu
\end{align*}
\]

\[(1)\]

**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 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 \( \lambda_{[(i+j) \mod M]} \). So, in Fig. 4(a), \( C(4,1) \) uses two TLSs (grey ones) with \( \lambda_{[(11+1) \mod 18]} \) (\( \lambda_{12} \)) and \( \lambda_{[(12+6) \mod 18]} \) (\( \lambda_{0} \)) to communicate with \( C(4,2) \) and \( C(4,3) \) respectively. Regarding the inter-region communication, \( C(4,1) \) uses the other three TLSs (green ones) with \( \lambda_{11}, \lambda_{13}, \) and \( \lambda_{15} \) 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 \( \lambda_{13} \). Then, \( C(2,1) \) forwards this packet to \( C(2,3) \) with \( \lambda_{0} \). 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

---

Fig. 4(b) is the symmetric matrix for the network in Fig. 4(a). For example, AWGR\(_2\) and AWGR\(_3\) are using port 16 to connect with each other.
be no direct connection between certain pairs of Clusters. In this case, additional forwarding is needed. For example, in Fig. 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 CB 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.

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 $w_{ij} = \text{MAX}$ to label it as a break path (e.g. $\text{MAX} = 1000$). 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 Fig. 5.

Fig. 6 shows the usage of the WCM during the procedure of increasing bandwidth between $C_0 \rightarrow C_3$ (see Fig. 3). In Fig. 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 Fig. 6 (right top), if we tune TRX from $C_0 \rightarrow C_1$ to $C_0 \rightarrow C_3$, then the traffic from $C_0 \rightarrow C_1$ will be added to $C_0 \rightarrow C_2$ and $C_2 \rightarrow C_1$. 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 Fig. 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

Fig. 7 shows an eight-cluster experimental setup for the proposed core network architecture, with $p = 4$ and $\mu = 2$. Eight Xilinx VC709 boards with high-speed Rocket I/O TRXs at 10 Gb/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 ~ 1561.04 nm on a 0.4 nm (50 GHz) grid.

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 TLs with tuning time as short as few nanoseconds (see Fig. 8). These TLs allow for fast FB 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 Fig. 9).

Each FPGA acts as an intra-cluster TLS and traffic generator. As shown in Fig. 10, each TLS contains four network ports (each one connected with one of the four TRXs), one injection port with four independent 10 Gb/s traffic generators and one $20 \times 8$ 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

---

**Fig. 5.** Pseudo code of TRX selection algorithm.

```
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=1}^{N-1} w_{i}/(N-1)
Foreach (w_s ≤ avgRate): Push(LowRateSet, i)

# From LowRateSet, select the proper TRX to be tuned
finalTRX = 0;
finalRate = 1000;
finalReachable = FALSE;
Foreach TRX in LowRateSet:
    t = Pop(LowRateSet);
    finalRate = 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 (i=0; i<N; j=i+1)
        Begin
            if ((w_i > 1000) & (w_j > 1000))
                Begin
                    isReachable = TRUE;
                    if (minRate > max(w_i, w_j)) minRate = max(w_i, w_j);
                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 (minRate+w_t) > 1000) 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 && finalReachable = TRUE) && (finalRate < minRate))
            Begin
                finalTRX = t;
                finalRate = minRate + w_t;
            End
        End
    if (finalReachable) return finalTRX;
End
```

---

**Fig. 6.** Shows the usage of the WCM during the procedure of increasing bandwidth between $C_0 \rightarrow C_3$.
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 TL and use the look-ahead table as the new working table.

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. Fig. 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%.

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 Fig. 7, we experimentally demonstrated the following four scenarios: 40 Gbps hot-spot traffic between FPGA 1 and 4 (intra-Region) with and without FB adjustment; 40 Gbps hot-spot traffic between FPGA 1 and 5 (Inter-Region) with and without FB 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 Table 1. BER measurements in Fig. 9 show error-free operation after tuning the tunable TXs for bandwidth increase between the
Fig. 10. Emulated intra-cluster TLS.

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

Fig. 12. Measured accepted hot-spot traffic bandwidth.

Fig. 13. Measured accepted background traffic bandwidth.

Fig. 14. Measured stable points of accepted hot-spot traffic bandwidth.

Fig. 15. Measured overall network accepted bandwidth.

hot-spot points. Figs. 12–15 show the experimentally measured statistics for the four scenarios described above.

Fig. 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 FB adjustment, the performance of hot-spot traffic will cease to decrease, and it stabilizes while achieving up to $\sim 1.77 \times$ improvement in accepted hot-spot traffic.

Fig. 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 Fig. 14 shows, the network with flexibility can achieve similar latency performance, since the queuing time caused by the hot traffic is reduced.

Fig. 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.

The significant improvement in throughput for the hot-spot traffic is also beneficial from an energy efficiency standpoint. Table II 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 TL [23], SiP receiver (3.95 mW) [24] and SiP modulator [25]. According to ref. [23], we can utilize a TL 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
Fig. 14. Measured latency of background traffic.

Fig. 15. Measured total accepted bandwidth.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>WAVELENGTH ALLOCATION OF THE TL-TXs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w/o flexibility</td>
</tr>
<tr>
<td>TL-TX1</td>
<td>1561.41 nm</td>
</tr>
<tr>
<td>TL-TX2</td>
<td>1559.79 nm</td>
</tr>
<tr>
<td>TL-TX3</td>
<td>1546.04 nm</td>
</tr>
<tr>
<td>TL-TX4</td>
<td>1559.79 nm</td>
</tr>
</tbody>
</table>

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 10 Gb/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 II) will be slight lower than the one in network without flexibility (2nd column in Table II). Overall 1.19× energy efficiency improvements can be achieved by introducing FB assignment. If we can turn off the idle TRXs after changing the connectivity, the system will achieve higher energy efficiency.

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 (Fig. 11) would be the same, but in case of hot-spot traffic, the performance would be the same as shown in Fig. 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 CB 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 TL in the proposed network consumes 2× higher power than the multi-mode TRX in the electrical network as described in [26].

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>POWER COMPARISON ON NETWORK WITH INTRA-REGION TRAFFIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w/o flexibility</td>
</tr>
<tr>
<td>Power</td>
<td>112 W</td>
</tr>
<tr>
<td>Total accepted bandwidth</td>
<td>80 Gb/s</td>
</tr>
<tr>
<td>Energy Efficiency (SFP TRX)</td>
<td>1.400 nJ/bit</td>
</tr>
<tr>
<td>Power(SiP TRX)</td>
<td>251.2 mW</td>
</tr>
<tr>
<td>Energy Efficiency (SiP TRX)</td>
<td>3.14 pJ/bit</td>
</tr>
</tbody>
</table>

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 TLs. 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 FB 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 CB 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.
REFERENCES


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 a Postdoctoral Research Scientist in the Next Generation Networking Systems Laboratory, University of California, Davis, CA, USA. His research interests include low-latency and scalable optical interconnects for data centers and high performance computing.

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 optical communication systems and networking from Scuola Superiore Sant’Anna, Pisa, 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 systems.

Matthew Clements received the B.S. degree from the Department of Electrical Engineering, University of California, Santa Cruz, CA, USA, in 2013. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of California, Davis, CA, USA. His research interests include optical coherent communications, optical-arbitrary waveform generation and measurement, and optical comb generation.

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, Davis, CA, USA. His research interests include 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 in ATD/MONET tested 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. He is a Fellow of the Optical Society of America, and is a recipient of the DARPA Award for Sustained Excellence (1997), the Bellcore CEO Award (1998), the Outstanding Mid-Career Research Award (UC Davis, 2004), and the Outstanding Senior Research Award (UC Davis, 2011).
Abstract—This paper proposes and demonstrates a flexible-bandwidth optical interconnect architecture for data centers exploiting wavelength routing in arrayed waveguide grating routers 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 tested with hierarchical all-to-all interconnection shows 1.77× throughput increase and 1.19× network energy efficiency improvement in the case of intercluster hot-spot traffic, while guaranteeing more than 97% throughput for the portion of the traffic with uniform random distribution.

Index Terms—Arrayed waveguide grating routers, data center networking, elastic optical networks, flexible bandwidth, optical interconnects.

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 Fig. 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 allocate 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.

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 arrayed waveguide grating router’s (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 AWGRs [10], and fast tunable lasers (TL) [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 FB networks and AWGR-based networks. Section III introduces the proposed dynamic
CB technique that enables fast FB 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], FB 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 a 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. 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 FB 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 CB TECHNIQUE

Dynamic 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 TLs (<10 ns switching time) are key enabling technologies to achieve dynamic bandwidth adjustment for the proposed core network. As shown in Fig. 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.

When hot spots form, the network control plane will tune certain transceivers’ (TRXs) wavelengths to increase the number of connections between the hot Clusters. Fig. 3 illustrates the concept of CB in details. Originally, with proper wavelength assignment, the four Clusters (\( C_0 \sim C_3 \)) 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 \( C_0 \to C_2 \) with \( \lambda_1 \) (blue link) is tuned to \( \lambda_2 \) (red link) for \( C_0 \to C_3 \). Then we perform similar tuning procedure for \( C_3 \to C_0 \). 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 \( C_0 \to C_2 \) is not available anymore, and the traffic between \( C_0 \) and \( C_3 \) needs to be relayed by \( C_1 \) or \( C_3 \).

IV. FLEXIBLE FLAT ALL-TO-ALL OPTICAL CORE NETWORK

A. System Architecture

We group Clusters into \( \mu \) Regions and each Region contains \( p \) Clusters and one AWGR. Each intra-cluster top level switches (TLS) uses \( p-1 \) TRXs for intra-region communication and \( \mu-1 \) 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 TLs which can achieve fast wavelength tuning in <10 nanoseconds. A control plane interfaces with all the TLs in the Clusters and controls the TLs 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. Fig. 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 Fig. 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 \( \mu \) AWGRs with \( \mu - 1 \) fibers in an all-to-all pattern (blue lines in Fig. 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 \( \{AWGR_1, AWGR_2, \ldots, AWGR_p\} \) 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:

\[
\begin{align*}
    a_{ij} &= a_{ji}, & 1 \leq i \leq \mu, 1 \leq j \leq \mu, i \neq j \\
    a_{ij} &= NULL & i = j \\
    a_{1j} \cap a_{2j} \cap a_{\mu j} &= \emptyset & 1 \leq j \leq \mu \\
    a_{i1} \cap a_{i2} \cap a_{i\mu} &= \emptyset & 1 \leq i \leq \mu 
\end{align*}
\]

(Fig. 4(b) is the symmetric matrix for the network in Fig. 4(a). For example, \( AWGR_5 \) 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 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 \( \lambda_{[(i+j) \mod M]} \). So, in Fig. 4(a), \( C(4,1) \) uses two TLs (grey ones) with \( \lambda_{[(11+1) \mod 18]} \) and \( \lambda_{[(12+6) \mod 18]} \) to communicate with \( C(4,2) \) and \( C(4,3) \) respectively. Regarding the inter-region communication, \( C(4,1) \) uses the other three TLs (green ones) with \( \lambda_{11}, \lambda_{13}, \) and \( \lambda_{15} \) 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 \( \lambda_{13} \). Then, \( C(2,1) \) forwards this packet to \( C(2,3) \) with \( \lambda_{0} \). 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 Fig. 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 CB 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 (contains two links), we will perform the TRX selection twice, one selection for each link.

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 \(w_{ij} = \text{MAX}\) to label it as a break path (e.g. \(\text{MAX} = 1000\)). 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 Fig. 5.

Fig. 6 shows the usage of the WCM during the procedure of increasing bandwidth between \(C_0 \rightarrow C_3\) (see Fig. 3). In Fig. 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 Fig. 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 Fig. 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

Fig. 7 shows an eight-cluster experimental setup for the proposed core network architecture, with \(p = 4\) and \(\mu = 2\). Eight Xilinx VC709 boards with high-speed Rocket I/O TRXs at 10 Gb/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 ~ 1561.04 nm on a 0.4 nm (50 GHz) grid.

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 TLs with tuning time as short as few nanoseconds (see Fig. 8). These TLs allow for fast FB 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 Fig. 9).

Each FPGA acts as an intra-cluster TLS and traffic generator. As shown in Fig. 10, each TLS contains four network ports (each one connected with one of the four TRXs), one injection port with four independent 10 Gb/s traffic generators and one \(20 \times 8\) 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

---

**Algorithm:**

TRXSelect \(\{\text{WCM}(N \times N), s, d\}\)

\(\text{WCM}(N \times N)\): weighted connection matrix for \(N\) Clusters

\(s\): source Cluster ID

\(d\): destination Cluster ID

**Begin**

1. 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

\[
\text{avgRate} = \frac{\sum_{i \neq s} w_i/(N-1)}{(N-1)}
\]

Foreach \((w_i, \text{avgRate})\): Push(LowRateSet, \(i\))

2. From LowRateSet, select the proper TRX to be tuned

\(\text{finalTRX} = 0;\)

\(\text{finalRate} = 1000;\)

\(\text{finalReachable} = \text{FALSE};\)

Foreach TRX in LowRateSet:

\(t = \text{Pop(LowRateSet)};\)

\(\text{minRate} = 1000;\)

\(\text{isReachable} = \text{FALSE};\)

3. Check if Cluster \(s\) can still reach Cluster \(t\) after tuning the TRX

If \((s = j \&\& j \neq t + 1)\):

\(\text{isReachable} = \text{FALSE};\)

If \((\text{minRate} > w_s, w_t)\): \(\text{minRate} = \text{MAX}(w_s, w_t)\);

End

4. Calculate utilized rate of forwarding link after adjustment

If \((\text{finalRate} > w_s) > 100\): \(\text{isReachable} = \text{FALSE};\)

5. Select the one that can still maintain Cluster reachable and introduce less load on the forwarding path

If \((\text{isReachable} \&\& \text{finalReachable}) = \text{TRUE};\)

If \((\text{isReachable} \&\& \text{finalReachable}) = \text{TRUE};\)

\(\text{finalTRX} = t;\)

\(\text{finalRate} = \text{minRate} + w_s;\)

End

6. Get the forwarding path with minimum utilization rate

7. Check if the forwarding link will be overloaded, then current TRX used for communication between Cluster \(s\) and Cluster \(t\) cannot be tuned.

If \((\text{minRate} > w_s); \text{isReachable} = \text{FALSE};\)

8. Select the one that can still maintain Cluster reachable and introduce less load on the forwarding path

If \((\text{isReachable} \&\& \text{finalReachable}) = \text{TRUE};\)

If \((\text{isReachable} \&\& \text{finalReachable}) = \text{TRUE};\)

\(\text{finalTRX} = t;\)

\(\text{finalRate} = \text{minRate} + w_s;\)

End

Fig. 5. Pseudo code of TRX selection algorithm.
Fig. 6. An example of using weighted connection matrix.

Fig. 7. Experiment setup of the full system interconnection network. TL-TX: tunable transmitter (10 Gb/s); RX: receiver (10 Gb/s); SFP+: SFP transceiver (10 Gb/s); AWGR; FPGA.

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. Fig. 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%.

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 Fig. 7, we experimentally demonstrated the following four scenarios: 40 Gbps hot-spot traffic between FPGA 1 and 4 (intra-Region) with and without FB adjustment; 40 Gbps hot-spot traffic between FPGA 1 and 5 (Inter-Region) with and without FB 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 Table 1. BER measurements in Fig. 9 show error-free operation after tuning the tunable TXs for bandwidth increase between the...
hot-spot points. Figs. 12–15 show the experimentally measured statistics for the four scenarios described above.

Fig. 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 FB adjustment, the performance of hot-spot traffic will cease to decrease, and it stabilizes while achieving up to $\sim 1.77 \times$ improvement in accepted hot-spot traffic.

Fig. 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 Fig. 14 shows, the network with flexibility can achieve similar latency performance, since the queuing time caused by the hot traffic is reduced.

Fig. 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.

The significant improvement in throughput for the hot-spot traffic is also beneficial from an energy efficiency standpoint. Table II 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 TL [23], SiP receiver (3.95 mW) [24] and SiP modulator [25]. According to ref. [23], we can utilize a TL 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 ($10 \text{ nm} / 38 \text{ nm} \times 26 \text{ mW} \times 50\%$) 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 10 Gb/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 II) will be slight lower than the one in network without flexibility(2nd column in Table II). Overall 1.19× energy efficiency improvements can be achieved by introducing FB assignment. If we can turn off the idle TRXs after changing the connectivity, the system will achieve higher energy efficiency.

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 (Fig. 11) would be the same, but in case of hot-spot traffic, the performance would be the same as shown in Fig. 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 CB 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 TL 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 TLs. 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 FB 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 CB 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.
REFERENCES


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 a Postdoctoral Research Scientist in the Next Generation Networking Systems Laboratory, University of California, Davis, CA, USA. His research interests include low-latency and scalable optical interconnects for data centers and high performance computing.

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 optical communication systems and networking from Scuola Superiore Sant’Anna, Pisa, 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 systems.

Matthew Clements received the B.S. degree from the Department of Electrical Engineering, University of California, Santa Cruz, CA, USA, in 2013. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of California, Davis, CA, USA. His research interests include optical coherent communications, optical arbitrary waveform generation and measurement, and optical comb generation.

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, Davis, CA, USA. His research interests include 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 in 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. He is a Fellow of the Optical Society of America, and is a recipient of the DARPA Award for Sustained Excellence (1997), the Bellcore CEO Award (1998), the Outstanding Mid-Career Research Award (UC Davis, 2004), and the Outstanding Senior Research Award (UC Davis, 2011).