Experimental Demonstration of a 64-Port Wavelength Routing Thin-CLOS System for Data Center Switching Architectures


Abstract—This paper reports on the results of the design, fabrication, and experimental demonstration of a wavelength routing Thin-CLOS system based on arrayed waveguide grating routers (AWGRs) for intra-data-center applications. By using \( M^2 \) \( W \)-port AWGRs, this architecture allows increasing by a factor of \( M \) the scalability provided by a wavelength router using a single \( W \)-port AWGR. The fabricated Thin-CLOS system has 64 ports and makes use of \( M^2 = 4 \) AWGRs with \( W = 32 \) ports and 100 GHz channel spacing. The system fits in a 1U rack enclosure and consumes less than 10 W. We demonstrated error-free performance at 10 Gb/s per wavelength with a power penalty of <3 dB under a worst-case crosstalk scenario. The adopted 64-port Thin-CLOS architecture makes use of four 32-port silicon AWGRs to reduce the impact of in-band crosstalk while guaranteeing nonblocking connectivity among the 64 ports and using only \( M = 32 \) distinct wavelengths.

Index Terms—Arrayed waveguide grating routers; Data centers; Optical interconnects; Wavelength routing.

I. INTRODUCTION

Next-generation data center architectures will require transformative changes to sustain the ever-increasing demands for cloud computing services, ubiquitous connectivity, and bandwidth while reducing the amount of power required to operate these warehouse-scale computing systems. With IPv6, data centers can now address every appliance and sensor on Earth; further, this vast set of data will be networked, processed, and accessed on virtual platforms (commonly referred to as cloud architectures), which consist of data systems, networks (optical, electrical/wireless, and wireline), and client interfaces (e.g., terminals or handheld devices). On the physical level, these virtual platforms run in data centers, which are essentially a collection of internetworked servers designed to store, access, and process data for clients. Hundreds of thousands of servers are interconnected by multiple layers of electronic switches organized in tree-style architectures (i.e., Fat Tree [1,2]) with different levels of oversubscription to reduce costs and power consumption. In particular, today’s cloud computing architectures are designed with a fixed topology with fixed patterns of data movements at fixed data rates. However, actual computations have large peak-to-average ratios in processing, bursty data traffic, dynamically changing data movement patterns, and heterogeneous processing threads that may benefit from low-diameter all-to-all topology at certain times and high-diameter topologies but with high-bandwidth neighbor communication at other times.

While current data center architectures already heavily rely on point-to-point optical interconnects for energy-efficient and high-bandwidth inter-rack communication, switching operations, as mentioned above, still rely completely on electronic packet switches. It is necessary to introduce a new class of data center interconnects and switches that can leverage the unique properties of certain classes of optical devices to design new and transformative data center switching architectures. There has already been a significant amount of research on switch and data center architectures exploiting optical devices with wavelength and spatial switching functionalities in conjunction with legacy electronic solutions. The reader can refer to the following papers for solutions based on spatial switches (MEMS [3–5], Semiconductor Optical Amplifiers [6–9], wavelength switches (AWGR [10–15], and micro-ring resonator (MRR)-based switches [16,17]). The reader can also refer to papers [13,18] for comparison between the above-mentioned solutions. All these optical switches share a common aspect: they are bufferless (no buffering operation at the switch input and output ports) and therefore cannot be cascaded. Thus, they could be either used as core switches in folded-CLOS type of architectures (i.e., Fat Tree [1,2]) or they could be used in directly connected architectures, e.g., Torus [19], Flattened Butterfly [20,21], or Dragonfly [22] (where they can interconnect directly computing nodes or ToR switches).

As for the electronic switches, the optical switch solutions mentioned above can have scalability limitations due to different aspects mostly related to loss and coherent in-band crosstalk [23].

In the context of AWGR-based architectures, this paper focuses on the first experimental demonstration of a wavelength routing enabling technology (originally proposed in [11,24] but never demonstrated experimentally) called
“Thin-CLOS.” Thin-CLOS aims at increasing the scalability of wavelength routing interconnect systems based on arrayed waveguide grating routers (AWGR) [25,26], which represent the best example of a compact and passive optical device with such wavelength routing capability. Figure 1 illustrates the well-known wavelength routing property of an AWGR allowing any input port to communicate with any output port simultaneously using different wavelengths without contention. Thus, a $W \times W$ AWGR can intrinsically provide strictly nonblocking all-to-all communication among $W$ compute nodes in a flat topology using $W$ wavelengths (see Section II.B for more details about the use of AWGR for optical interconnection and switching solutions).

By using $M^2$ $W$-port AWGRs, the proposed Thin-CLOS architecture allows increasing by a factor of $M$ the scalability provided by a single $W$-port AWGR, thus effectively mimicking the functionality of a larger $N = M \times W$-port AWGR.

In this paper, we designed, fabricated, and experimentally demonstrated for the first time a 64-port Thin-CLOS AWGR system composed of four 32-port AWGRs that fits into a 1U rack and consumes only 10 W ($W = 32, M = 2$).

The remainder of this paper is organized as follows. Section II describes the working principle of Thin-CLOS architecture and the main design trade-offs for an $N$-port system. Section III describes the details of the design, fabrication, and experimental verification of the Thin-CLOS 64-port system. Section IV concludes the paper, and Section V provides a list of references.

II. Thin-CLOS AWGR Architecture for Data Center Switching

A. Principle of Operation

As Fig. 1(c) demonstrates, the wavelength routing functionality in an AWGR enables (strictly) nonblocking communication between any input and output port.

Three limiting factors prevent such an AWGR system from being practically deployed in a large scale ($\geq 32$). First, in-band (coherent) crosstalk increases significantly as the number of wavelength channels grows [27]. This can significantly impact the bit error rate (BER) of the optical links. Second, device size and fabrication constraints can also limit the port count of a single AWGR. Limitations are mainly due to the highly precise control needed for the channel spacing during fabrication as well as accurate wavelength registration required for all channels after fabrication [28]. Third, increasing the port count linearly increases the number of wavelengths, but the limited spectral range results in narrow channel spacing. Therefore, to achieve scalability, it would be desirable to use many smaller AWGRs using a smaller number of wavelengths that can be combined to provide the same interconnectivity offered by a single larger AWGR. This can be achieved by using the proposed Thin-CLOS AWGR.

Figure 2(a) shows a generic $N$-port Thin-CLOS architecture achieving the same functionality of a single $N$-port AWGR with $N$ wavelengths by using smaller $W$-port AWGRs. The architecture is strictly nonblocking and consists of a single layer of $M$ groups of $M$ AWGRs with $W$ ports, being $N = M \times W$. In summary, as shown in Fig. 2(b), there are $M^2$ AWGRs and $2 \times M^2 \times W$ fiber connections in a Thin-CLOS architecture ($M^2 \times W$ input ports and $M^2 \times W$ output ports). Although this translates to a larger number of fibers, connectors, and more complex fiber management inside the enclosure, there are several significant advantages compared with the single AWGR solution:

- it greatly reduces the in-band crosstalk due to the lower port count of the AWGRs ($W - 1$ number of crosstalk sources instead of $N - 1$);
- it offers lower optical losses and reduced optical loss non-uniformity and frequency deviation issues;
- it allows larger channel spacing, relaxes manufacturing tolerance, and relaxes the temperature control requirement (thus reduces power consumption); and
- it improves the yield and reduces the manufacturing cost of the many, identical, and smaller AWGRs.

B. Use of Thin-CLOS AWGR for Interconnection and Switching

As explained above, Thin-CLOS architecture mimics the functionality of an $N$-port AWGR and can, therefore, be used as enabling technology to implement different AWGR-based interconnection and switching solutions already proposed in the literature [12,29–34], which could benefit from a large port-count wavelength routing system.
In this section, we want to discuss the differences and trade-offs of using a Thin-CLOS AWGR to implement either a passive all-to-all wavelength interconnection (as in [30]) with WDM lasers or an optical switch with tunable lasers (as in [12,14,15,32]).

The main differences between the two architectures are noted as follows.

The first approach requires each node to use $M$ banks of $W$-wavelength WDM TRXs [see Fig. 2(c)]. This architecture is all-to-all; therefore, there is no contention in the optical domain. Switching operation is done at the edges, in the electronic switches (these switches could be either ToR switches or embedded switches in a computing node [10]). This realization has been named a passive AWGR switch or passive low-latency interconnect optical network switch (LIONS) [32] because no optical reconfiguration is necessary. One limitation of this passive approach is the fact that the number of TRXs required to interconnect $N$ nodes is $N^2$. This can be expensive in terms of energy and cost, especially when using off-the-shelf WDM TRXs. A more suitable approach for this solution would be to use emerging silicon photonic (SiP) WDM TRXs with integrated optical frequency combs [35]. In terms of performance, in the case of uniform random traffic, the throughput is 100%, and the latency is constant with the load and is simply equal to the transmission time and propagation delay because the interconnection is flat and single-hop without contention among the $N$ nodes. To scale the number of nodes beyond what is possible with a single Thin-CLOS AWGR, a hierarchical approach (as in [30]) could be used.

In the second approach with tunable lasers [see Fig. 2(d)], each node requires using only $M$ tunable TRXs. Therefore, the solution is more affordable because the total number of TRXs is now $M \times N$. However, this solution requires careful centralized scheduling or contention resolution schemes in the optical domain [we indicated this with a generic block called “control unit” in Fig. 2(d)]. Overall, the complexity of this approach is higher. The switch performance in terms of latency and throughput depends on the specific contention resolution scheme, laser switching time, and packet sizes. Assuming fast tunable lasers with switching time in the order of a few nanoseconds (as in [11]) and burst-mode RXs, the throughput can be still above 70% for packet sizes as small as 256 bytes, even if the number of RXs per node ($M$ in this case) is $\ll N$. This is because the AWGR wavelength routing principle naturally implements output queuing architecture with a speedup of $M$. Therefore, under a uniform random traffic profile, the contention probability is strongly reduced.

III. DESIGN, FABRICATION, AND EXPERIMENTAL DEMONSTRATION OF A 64-PORT THIN-CLOS

A. Architecture and System Design Trade-offs

As mentioned above, Thin-CLOS can reduce the number of wavelengths and crosstalk components at the expense of increasing the number of fibers and AWGRs. As shown in Fig. 2, the number of fibers scales as $2 \times M^2 \times W$ [4]. Therefore, to limit the number of fibers and AWGRs, we chose to investigate a design with $M^2 = 4 \times 32 \times 32$ AWGRs, so that
all the components could fit into a 1U rack unit. The 64 ports are divided into two groups: 1 to 32 and 33 to 64. Each port has two pairs of fibers that would connect to a node.

Each node has two sets of transmitters (Tx) and receivers (Rx) with 32 wavelengths. Intra-group interconnects are implemented by the first and fourth AWGRs while cross-group interconnects are supported by the second and third AWGRs.

While $M = 2$ represents the best solution regarding the number of AWGRs and fibers required, it is essential to verify that the 32-port AWGR from Enablence could guarantee error-free condition under the worst-case in-band crosstalk scenario in all-to-all configuration. To this aim, we performed a detailed experimental analysis of a crosstalk-induced power penalty. The section below reports the experiment setup and results showing that Enablence 32-port AWGR can deliver error-free performance for all-to-all communication. Note that the crosstalk penalty in a 64-port Thin-CLOS system solely depends on the 32-port AWGRs inside the Thin-CLOS enclosure. Our experiment in Section III.A used 32 signals at the same wavelength with aligned polarization and for worst-case loss paths. Therefore, the power penalty measurements can be considered representative of a 64-port system at full-scale.

B. Crosstalk Analysis

Figure 3 shows the experimental setup used to assess the crosstalk-induced power penalty when using Enablence 32-port AWGR. One FPGA evaluation board with 10 Gb/s DWDM SFP+ TRX acted as a transmitter and receiver to measure the BER at a given wavelength (we can choose the wavelength by using different TRX modules). To emulate the 31 crosstalk components at the same wavelength of the signal under test, we split the output of the SFP+TX. In this way, we assured that the crosstalk wavelength signals were exactly matching with the one of the test signals for maximum interference condition.

Fiber delay lines of several tens of meters (much longer than the coherence length of the SFP+ laser $\sim 5$ m for 20 MHz linewidth) were used to guarantee that the crosstalk signals acted as independent laser sources. These delay lines also assure that the 31 different copies were decorrelated. Polarization controllers were used together with a polarizer placed at the AWGR output to align all the polarizations for the worst-case crosstalk scenario. The crosstalk-to-signal ratio was measured before and after the PBS to assure the one after the PBS was not improved due to polarization filtering caused by polarization misalignment. When choosing the wavelength of the signal under test, we considered the fact that Enablence AWGR loss is not uniform. Therefore, as the first step, we determined the worst-case loss when using side and center inputs of the AWGR (inputs 17 and 31). When using input 17, the worst-case loss is when the signal must reach output 31. We measured a loss of 4.7 dB, which is in agreement with the data sheet provided by Enablence.

However, when using input 1, we measured a worst-case loss of 6 dB when using output port 32 (Table I). Thus, we selected the wavelengths associated with these input and output combinations to test the BER performance under

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Wavelength (nm)</th>
<th>Insertion loss (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1558.983</td>
<td>5.35</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1558.173</td>
<td>3.95</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>1546.917</td>
<td>3.17</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>1546.119</td>
<td>3.23</td>
</tr>
<tr>
<td>1</td>
<td>31</td>
<td>1560.614</td>
<td>5.39</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>1559.800</td>
<td>6.00</td>
</tr>
<tr>
<td>17</td>
<td>30</td>
<td>1548.515</td>
<td>3.16</td>
</tr>
<tr>
<td>17</td>
<td>31</td>
<td>1547.715</td>
<td>4.68</td>
</tr>
<tr>
<td>17</td>
<td>32</td>
<td>1546.917</td>
<td>3.76</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>Data sheet</td>
<td>1.52-4.53</td>
</tr>
</tbody>
</table>

Fig. 3. Experiment setup using Enablence 32-port OSD based on single 100 GHz spacing AWGR. PM, power monitor; PBS, polarization beam splitter; EDFA, erbium-doped fiber amplifier; VOA, variable optical attenuator; PC, polarization controller.
these worst-case loss and crosstalk scenarios. We also measured the in-band crosstalk contribution given by each of the other 31 ports when using the wavelengths selected (highlighted in red in Table I). Figure 4 shows the results of these measurements. The average per-port crosstalk contribution (normalized to the signal output power or AWGR loss) was $-34.85$ dB and $-37.39$ dB at outputs 31 and 32 for $\lambda_{1-32}$ and $\lambda_{17-31}$, respectively.

According to the analytical results reported in Fig. 5 and based on [23], the crosstalk values above should guarantee error-free operation with limited power penalty. We confirmed this by using the selected wavelengths for the experiment shown in Fig. 3. Figures 6–8 report the BER measurements for the worst cases discussed above. We considered the worst-case path loss from the AWGR datasheet provided for center input 17 as well as the worst-case loss path measured in the lab for side input 1. The power penalty difference is 0.5 dB. We also plotted the BER curve in the case of nonaligned polarizations to emphasize the importance of designing the system for the worst-case scenario.

Note that each measurement has been carried out using a different SFP+ DWDM TRXs for the specific wavelength needed to communicate between two specific ports. We noticed that different TRXs have slightly different RX sensitivities. This explains why the back-to-back curves are slightly different. In any case, the power penalty is always $<3$ dB for BER $= 1E - 12$ with PRBS testing sequence of $2^{31} - 1$.

C. 1U Enclosure and Cabling Design

The experiment results reported above demonstrated that it is feasible to use 32-port AWGRs to build a 64-port Thin-CLOS architecture. In this section, we discuss details and solutions adopted regarding connectors and fiber management necessary to guarantee the correct functionality and connectivity required for a Thin-CLOS with $M = 2$ that fits in a 1U rack enclosure.
First, a 64-port all-to-all Thin-CLOS architecture has 128 input fibers and 128 output fibers. To accommodate these 256 connections on the front panel of a 1U rack enclosure, we determined that it was necessary to make use of high-density connectors, i.e., MTP connectors. We used 16 MTP connectors and cables, each one carrying a bundle of 16 fibers (these are custom-made MTP cables because the legacy commercial solutions carry 24 fibers). Each MTP connector carries input and output fibers for four nodes (four fibers per node), as shown in Fig. 9. In fact, each node connected to the designed 64-port Thin-CLOS architecture would have \( M/0.136 \) WDM transmitters and receivers, requiring then \( M/0.136 \) input and \( M/0.136 \) output fibers. Thus, in each MTP connector, the first group of four pins are for the first node, the second group of four pins are for the second node, and so on. Also, pins with odd index numbers (i.e., 1, 3, 5, etc.) connect to the nodes’ TXs, while pins with even index numbers (i.e., 2, 4, 6, etc.) connect to the nodes’ RXs. Figure 9 shows an example of connectivity between MTP1 and nodes 1 to 4.

Once the MTPs’ pins assignment for the connectors facing the end-nodes was determined, it was necessary to carefully determine the connections between each MTP pin inside the enclosure and the AWGRs input and output fibers. Note that, as explained above, in a Thin-CLOS with \( M = 2 \), the ports (nodes) are organized into two groups. The first group of 32 makes use of AWGR 1, 2, and 3. The second group of 32 makes use of AWGR 4, 3, and 2. Therefore, because each MTP serves four nodes, MTP 1 to 8 will belong to the first group, while MTP 1 to 9 will belong to the second group. Table II explicates the rule that needs to be followed to connect MTPs in the first and second groups to the AWGR fibers inside the enclosure. The example in Table II is given for MTP 1 (node 1 and node 2, group 1) and MTP 9 (node 33 and 34, group 2).

### IV. Experimental Demonstration

The intra-band worst-case crosstalk for the fabricated 64-port Thin-CLOS is the same as the worst-case crosstalk scenario for one of the 32 × 32 AWGRs inside the enclosure. The measurements above already show that the 32-port silica AWGR crosstalk rejection value guarantees error-free operation under a worst-case scenario. The goal of the experiment setup of Fig. 10(a) is to verify and demonstrate the correct routing operation in the 64-port
Thin-CLOS system. One FPGA evaluation board with wavelength-specific small form pluggable (SFP+) transceivers at 10 Gb/s acts as the TX and RX to measure the BER curves for different input and output nodes combinations. The FPGA connects to specific ports of the Thin-CLOS enclosure, as shown in Figure 10(a). A power meter is used to monitor the signal power at the receiver. Figure 10(c) plots the BER curves for intra-group (Node 1 to Node 32, Node 64 to Node 63) and cross-group (Node 1 to Node 64, Node 64 to Node 32) interconnects using different AWGRs inside the 1U enclosure. The difference between the four BER curves is simply related to the different sensitivity values of the commercial SFP+ DWDM TRXs used in this experiment. The wavelength values used in the experiment are summarized in Table III and are determined by the wavelength routing table of the four AWGRs (all the AWGRs have the same specifications and wavelengths).

### A. Power Budget Analysis

The system crosstalk power penalty discussed above is of fundamental importance for performing a power budget analysis and determining whether the average optical power at the RX will be enough to meet the requirements in terms of BER (i.e., BER $\leq 1E^{-12}$). The equation below represents the lower bound for the average optical power of each optical transmitter connecting to the all-to-all system:

$$P_{TX} \geq P_{RX\ Sens} + AWGR_{IL} + AWGmux_{IL} + AWGdemux_{IL} + Connector_{IL} + Penalty_{XT}.$$ 

$P_{RX\ Sens}$ is the RX sensitivity, defined as the minimum required optical power to guarantee error-free performance in the absence of any source of impairments other than shot noise and thermal noise at the RX; $AWGR_{IL}$ is the AWGR insertion loss between each input and output port; $AWGmux_{IL}$ is the insertion loss of the optical wavelength multiplexer located at each end-point; $AWGdemux_{IL}$ is the insertion loss of the optical wavelength demultiplexer located at each end-point; $Connector_{IL}$ is the insertion loss of all the connectors located between any pair of TX and RX; $Penalty_{XT}$ is the power penalty (PP) as defined in the above sections.

Based on the values reported in Table IV, it is easy to calculate that the required $P_{TX}$ is only $-9$ dBm for the fabricated $K_{64,64}$ OSD with $[W = 32, M = 2]$. This requirement is about 10 dB lower than the typical output power of the commercial SFP+ DWDM TRXs used in the experiment, meaning that we have a power margin $\geq 10$ dB.

### TABLE III

**WAVELENGTH VALUES USED IN THE EXPERIMENT**

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Node 32</td>
</tr>
<tr>
<td>Node 1</td>
<td>1557.41 nm</td>
</tr>
<tr>
<td>Node 64</td>
<td>1535.98 nm</td>
</tr>
</tbody>
</table>

### TABLE IV

**TYPICAL VALUES MEASURED IN THE EXPERIMENT**

<table>
<thead>
<tr>
<th>Definition</th>
<th>Typical Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{TX}$ Transmitter optical power</td>
<td>0 dBm</td>
</tr>
<tr>
<td>$P_{RX\ Sens}$ Receiver sensitivity</td>
<td>$-25$ dBm</td>
</tr>
<tr>
<td>$P_{RX\ Max}$ Maximum receiver optical input power</td>
<td>$-10$ dBm</td>
</tr>
<tr>
<td>$AWGR_{IL}$ AWGR insertion loss</td>
<td>4 dB</td>
</tr>
<tr>
<td>$AWGmux_{IL}$ Multiplexer insertion loss</td>
<td>2 dB</td>
</tr>
<tr>
<td>$AWGdemux_{IL}$ Demultiplexer insertion loss</td>
<td>2 dB</td>
</tr>
<tr>
<td>$Penalty_{XT}$ Power penalty induced by in-band crosstalk</td>
<td>$&lt;3$ dB</td>
</tr>
<tr>
<td>$Connector_{IL}$ Loss due to connectors</td>
<td>$&lt;1$ dB</td>
</tr>
</tbody>
</table>
V. CONCLUSIONS

In this paper, we experimentally demonstrated a novel scalable AWGR-based wavelength routing architecture for intra-data-center applications. The fully packaged system, called Thin-CLOS AWGR, has been fabricated and assembled from Enablence Technologies Inc. The system has a 1U rack size and consumes 10 W. It allows interconnection in all-to-all fashion 64 end-points, with each end-point using two sets of $W = 32$ wavelengths. It can also be used as enabling technology for other optical switching architectures (as in [29] or [11,30]). The architecture is also promising for scalable on-chip silicon SIJ [33] integrated solutions, where state-of-the-art values for AWGR cross-talk are not low enough to scale beyond 16 ports with a single AWGR. The possibility of using a smaller number of wavelengths is also particularly attractive for such integrated solutions using compact optical frequency combs, which can typically cover a limited wavelength range of operation. Future works include (1) scaling the demonstration to 256 ports (still making use of 32-port AWGRs but with reduced size) and (2) demonstrating the Thin-CLOS AWGR on a SiP integration platform.

ACKNOWLEDGMENT

This work was supported in part by DoD contract # H98230-16-C-0820 and NSF grant # 1611560.

REFERENCES


