Modular AWG-based Interconnection for Large-Scale Data Center Networks

Tong Ye, *Member, IEEE,* Tony T. Lee, *Fellow, IEEE,* Mao Ge, and Weisheng Hu, *Member, IEEE*

Abstract—Along with the recent surge in scale expansion of data centers, the interconnection scheme is facing a grave challenge. A huge amount of cables between the switches make the system maintenance and heat dissipation extremely difficult. A promising solution to this problem is using the arrayed waveguide grating (AWG), which can provide a set of wavelength links between its inputs and outputs. However, the scalability of the AWG-based interconnection scheme is restricted by the coherent crosstalk and the wavelength granularity of AWGs. In this paper, we propose a generic modular AWG-based interconnection scheme with scalable wavelength granularity for mega data centers. We first devise a matrix-based method to decompose the AWG into a three-stage network of smaller AWGs, while preserving the nonblocking wavelength routing property of the AWGs. We then introduce the concept of wavelength independency based on the partitioning of the optical connections, such that modular AWGs in the network can reuse the same wavelength set with smaller granularity. We show that the proposed modular AWG-based interconnection network can simplify the cabling complexity of data center networks, while preserving the same function and bandwidth as the original data center network.

Index Terms—Data center networks, arrayed waveguide grating (AWG), interconnection network, modularity.

1 Introduction

MANY emerging internet services call for new demands on storage and processing of massive data [1]. In this revolutionary trend of information technology, data centers recently surge up in both scale and quantity and gradually become a critical part of the internet. In the meantime, each mega data center [2] contains more than one million servers spreading across tens of thousands of racks, which require a large-scale switching network to provide broadband and reconfigurable interconnections. Typical topologies of data center networks are multi-root tree [3], fat-tree [4], or butterfly [5], as illustrated in Fig. 1. Within mega data centers, these regular topologies provide huge bandwidths via their bipartite interconnections with a high complexity of cabling. The mass of long cables between the switches causes great difficulties in system maintenance [6]. First, when the network connection changes or line failure occurs, the system upgrading becomes extremely complicated. Furthermore, it is shown in a report from IBM [7] that the dense cabling will affect the heat dissipation due to inefficient airflow. Thus, the choice of cabling architecture has a major influence on the throughput, scalability and energy efficiency management of data centers, as pointed out in the Cisco white paper [8]. The concern that these issues may become the development obstruction of data centers has ignited much research [4], [5], [9]–[20] to embark on seeking new solutions to cut down the cabling complexity.

This work was supported by the National Science Foundation of China (61271215, 61172065, and 61433009), and Shanghai Committee of Science and Technology No. 13511502302. The work of T. T. Lee was also partially supported by the Hong Kong RGC Earmarked Grant CUHK 414012.
The authors are with the State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University, Shanghai 200030, China. (e-mail: {yetong, ttlee, mao.ge, wshu}@sjtu.edu.cn)

1.1 Previous Works

Wireless interconnect was initially suggested in Ref. [9] to deal with the complexity issue of cabling. The idea of all-wireless data centers using 60 GHz wireless channels was proposed in [14] to provide inter-rack interconnections. Such schemes require restructuring data center layout and have poor bisection bandwidth due to radio interference. A data center with hybrid wireless and wired interconnections was proposed in [10]–[13], of which the key idea is to use the wireless links as supplementary capacities to offload extra traffic from congested wired links. Similarly, the hybrid network suffers from the radio interference and the link blockage induced by the intrinsic line-of-sight limitation of the 60 GHz wireless links. To relax these limitations, a 3D beamforming paradigm was proposed in [15], [16], where wireless connections were established by 60 GHz signals bounce off ceilings. Though this method alleviates radio interference to some extent, the bandwidth of wireless links is too small for mega data centers. Another kind of all-wireless solution is the free space optics (FSO) based data centers reported in [18], where visible or infrared lasers are used to establish point-to-point data links. The implementation of this scheme is difficult in practice since it needs precise control of the lasers and the ceiling mirrors.

Optimal allocation of network devices is an alternative solution to cut down the cabling complexity. A packaging scheme was proposed in [4] to reduce the number of cables in a Fat-tree data center, the key idea is to put the edge switches and the aggregate switches that belong to the same pod together in a centralized rack such that the cables between them can be eliminated. This scheme still requires a lot of copper cables between the pods and the core switches, which makes the network impossible to be constructed and maintained [6]. To ease this cabling issue, it was suggested in Ref. [6] to place the core switches together

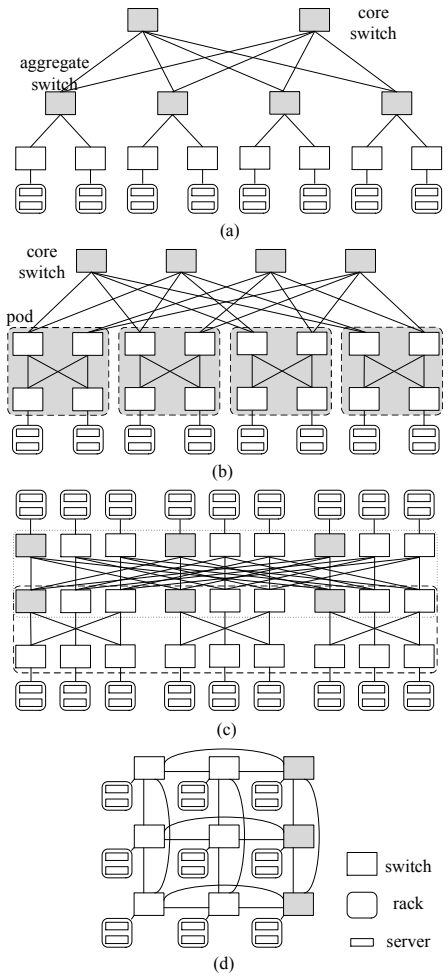


Fig. 1. Several topologies of data center networks: (a) multi-root tree [3], (b) fat-tree [4], (c) butterfly and (d) flattened butterfly [5].

in a central location and use multimode optical fibers. The cable reduction problem in data centers was investigated in Ref. [5] with a Butterfly topology. For an i-level Butterfly topology, it was shown in Ref. [5] that a flattened Butterfly topology can be obtained by replacing i switches at the ith level with a high-radix switch, which substantially reduces the number of interconnection links. However, the cabling complexity of the flatten Butterfly network is still high when the network size becomes very large. Moreover, this specific scheme is not a universal method that can be applied to an arbitrary data center network.

Recent results reported in [19], [20] demonstrated that the optical passive component, called arrayed waveguide grating (AWG), is a promising candidate to significantly reduce the number of links in an interconnection network. The AWG is not only energy efficient because it is a passive component but also it has a small insertion loss. An $N \times N$ AWG is a static wavelength router, which can provide N^2 interconnection links between its inputs and outputs via its internal wavelength set. The $N \times N$ AWG can be used to construct an $N \times N$ fully connected network if each of N switches connects to one input and one output of the AWG. As a matter of fact, $2N$ links together with an $N \times N$ AWG are sufficient to construct an $N \times N$ fully connected network. According to this principle, an AWG-based hierarchical data center network with low cabling complexity was developed in [19], and an AWG-based interconnection scheme was proposed in Ref. [20] to further reduce the number of cables of full meshes in the flattened Butterfly network.

Though the results presented in [19], [20] are very interesting, it remains a challenge to employ large AWGs in a mega data center, as indicated in [21] that a large AWG suffers from serious coherent crosstalk. Furthermore, a large AWG will make the network associated with a lot of wavelengths [19], which are precious resources in the optical communication window [22]. We refer to the number of wavelengths used by an interconnection network as the wavelength granularity of this network. Therefore, a scalable AWG-based interconnection in mega data centers implies that the reductions of the size of the AWG and the wavelength granularity of the network are both achievable.

Recently, our proposed modular scheme [23] demonstrated that a large AWG can be implemented by a network of modular AWGs with smaller sizes. Furthermore, the experimental results reported in Ref. [24] demonstrated that the reduction of the wavelength granularity of the modular AWG-based network can be realized when it is used to interconnect the electronic switching nodes. These preliminary results hinted at the feasibility to construct completely scalable AWG-based interconnections in mega data centers.

1.2 Our Work

In this paper, taking the limitation of wavelength granularity into consideration, we propose a generic principle of modular AWG-based interconnections to simplify the ever-growing complexity of the cabling of data center networks, such as Fat Tree, Butterfly, and BCube, rather than a new design of data center networks [25]–[28]. The design of modular AWG-based interconnection networks in this paper follows the similar idea of classical three-stage switching network, called Clos network [29], which was first formalized by Charles Clos in 1952 to reduce the number of cross-points in a telephone switching network. The aim of this paper is to show that this idea of multistage network remains important in the advent of complex data centers, with huge interconnect structures, each based on optical fiber links.

Our goal is to modularize the AWGs and partition the associated wavelength set while preserving the nonblocking wavelength routing property of the AWGs. We achieve this goal in two steps. We first develop a partitioned matrix scheme to decompose each AWG into a three-stage network of smaller AWGs. We then introduce the concept of wavelength independency based on the partitioned matrix of optical connections. We show that all AWGs in the interconnection network can be associated with the same wavelength set with smaller granularity. We further prove that the nonblocking wavelength routing property is preserved in the wavelength-reused interconnection network. Finally, we show that it is feasible to construct a modular AWG-based interconnection network which possesses the following features:

1) The modular AWG-based interconnection networks can replace the large amount of links within a data center, such as Fat Tree, Butterfly, and BCube, while

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCC.2016.2522417, IEEE Transactions on Cloud Computing

preserving the same function and bandwidth as the original data center network;
2) The cabling complexity can be substantially reduced;
3) The network only employs small-size AWG modules to avoid serious coherent crosstalk and passband deviation;
4) The modules of the AWG-based interconnection networks reuse the same set of wavelengths, which keeps wavelength granularity of the network small.

In summary, the novelties of our contributions are listed as follows:

a) This paper elaborates the relationship between the AWG decomposition and the cyclic Latin square matrix. In particular, we use the partitioned matrix method, or matrix-based AWG decomposition, to show the uniqueness of the decomposition of the cyclic Latin square matrix, which characterizes the intrinsic wavelength routing properties of AWG modules;
b) This paper demonstrates that a smaller set of wavelengths can be repeatedly used by AWG modules in a decomposed interconnection network when it is applied to data center networks. Such wavelength reuse property is important in practice because the cost of data center networks can be substantially reduced by smaller wavelength granularity.

The rest of this paper is organized as follows. In Section 2, we show that a stereotypical data center network comprises of several Banyan-type subnetworks, and illustrate that a Banyan-type interconnection subnetwork can be implemented by an AWG. In Section 3, we propose a scheme to modularize the AWG and reduce the wavelength granularity of AWG-based interconnection networks. We first develop a matrix-based method to decompose the AWG into a three-stage network of smaller AWGs, and then reduce the wavelength granularity of the network by partitioning the optical connections to achieve wavelength independency. We show that the modular AWG-based interconnection network is functionally equivalent to the original network with much less number of interconnection links. Section 4 further extends this scheme to construct the asymmetric AWG-based interconnection networks. Section 5 discusses the application of the proposed AWG-based interconnection network to simplify the cabling complexity of data center networks, and the relevant issues including transmission performance and the trade-off between the network complexity and cost. Finally, Section 6 concludes this paper.

2 Preliminaries

A stereotypical data center network consists of several Banyan-type subnetworks. As shown in Fig. 2, a Banyan-type network, denoted by $\mathcal{N}_A$, possesses two disjoint switching node sets U and V, in which each node in the set U connects to each node in the set V via exactly one link. In the $N_1 \times N_2$ Banyan-type network $\mathcal{N}_A$ shown in Fig. 2, there are $N_1 N_2$ links connecting the nodes $u_0, u_1, \cdots, u_{N_1-1}$ in the set U and the nodes $v_0, v_1, \cdots, v_{N_2-1}$ in the set V.

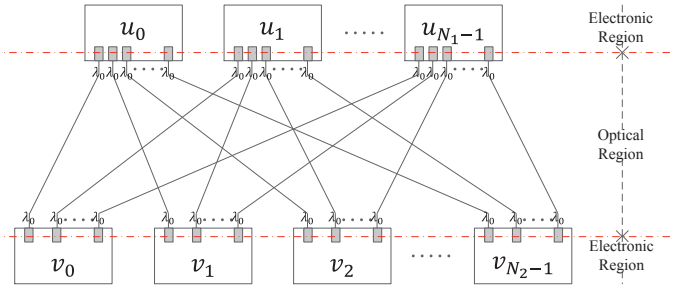


Fig. 2. A Banyan-type network $\mathcal{N}_A$.

Several data center networks with Banyan-type topologies are illustrated in Fig. 1. In the multi-root tree network shown in Fig. 1(a), the Banyan-type network between the core switches and the aggregate switches is a 2×4 $\mathcal{N}_A$. The core switches and the pods enclosed by dashed cycles in the fat-tree network shown in Fig. 1(b) are connected by a 4×4 $\mathcal{N}_A$. In the butterfly network shown in Fig. 1(c), either the network in the dotted cycle or that in the dashed cycle can be divided into 3 3×3 $\mathcal{N}_A$s, of which one of them connecting the gray nodes is shown in Fig. 1(c) as an example. In the flattened butterfly network shown in Fig. 1(d), the switches in each row (or column) are interconnected by a full-mesh network, which can be considered as a 3×3 network $\mathcal{N}_A$ if the switches and their mirrors form the node sets U and V respectively [21]. Furthermore, it is straightforward to show that the interconnection networks employed in other types of data centers, e.g., BCube in [30] and DCell in [31], can also be partitioned into multiple Banyan-type subnetworks.

It is clear that the cabling complexity of an $N_1 \times N_2$ Banyan-type $\mathcal{N}_A$ will grow rapidly with the increasing of N_1 and N_2. In this paper, we explore the modular AWG-based interconnection scheme to cut down the cabling complexity $N_1 N_2$ of $\mathcal{N}_A$. In the following, we first briefly introduce the optical devices to facilitate our discussions, and then illustrate the principle of AWG-based Banyan-type interconnection scheme.

2.1 Optical Transceiver

Optical fibers have been widely employed in the data center to fulfill bandwidth requirements of high-speed transmissions. Optical transceivers are key components installed at the end points of optical fiber links to transmit and receive optical data. That is, optical transceivers serve as the terminals of optical connections, and define the boundaries between the electronic region and the optical region, as illustrated in Fig. 2. We show in Section 3 that, with the help of these boundaries, it is possible to reuse the wavelengths in AWG-based modular interconnection networks.

The transceivers are represented by gray boxes in Fig. 2, in which the jth transceiver in the node u_i connects to the ith transceiver in the node v_j via an individual optical fiber. Since all connection links are disjoint, it follows that all transceivers in the $N_1 \times N_2$ Banyan-type interconnection $\mathcal{N}_A$ can be associated with the same wavelength, say λ_0, as shown in Fig. 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCC.2016.2522417, IEEE Transactions on Cloud Computing

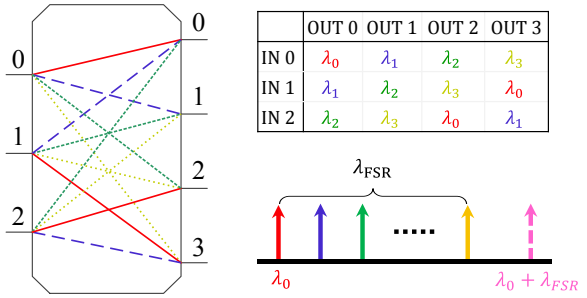


Fig. 3. A 3×4 AWG and its routing table [32].

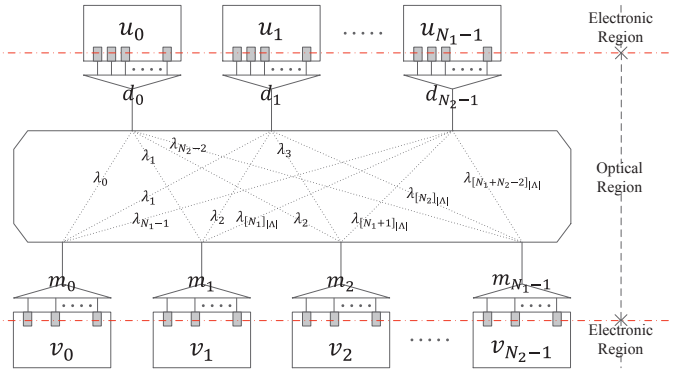


Fig. 4. AWG-based Banyan-type interconnection network $\mathcal{N}_B$.

2.2 Arrayed Waveguide Grating (AWG)

An $N_1 \times N_2$ AWG is associated with a set of equally spaced wavelengths $\Lambda = \{\lambda_0, \lambda_1, \cdots, \lambda_{|\Lambda|-1}\}$ in its principal free spectrum range (FSR), where $|\Lambda| = \max\{N_1, N_2\}$. The cyclic wavelength routing property of the AWG is defined as follows: the input i connects to the output j via the wavelength $\lambda_k \in \Lambda$ if

$$[i+j]_{|\Lambda|} = k, \tag{1}$$

where $[x]_y \triangleq x \bmod y$. It is easy to show that the wavelength assignments of AWGs according to (1) are contention-free [23], meaning that any two connections terminated on the same input or output will not use the same wavelength. As an example, the wavelength routing table of the 3×4 AWG shown in Fig. 3 is contention-free. For $N_1 = N_2 = N$, the routing table of a symmetric AWG becomes a cyclic Latin square matrix $\mathbf{M} = \{m_{ij}\}_{N \times N}$ defined by $m_{ij} = \lambda_k$, where $[i+j]_N = k$. For easy reference, we denote the wavelength subset associated with the ith input port as follows:

$$\Lambda_i = \{\lambda_{[i+j]_{|\Lambda|}} | j = 0, 1, \cdots, m-1\}.$$

for $i = 0, 1, \cdots, r-1$. Similarly, the wavelength subset associated with the jth output port is denoted by

$$\Lambda'_j = \{\lambda_{[j+i]_{|\Lambda|}} | i = 0, 1, \cdots, r-1\}.$$

for $j = 0, 1, \cdots, m-1$. It is clear that $\cup_{i=0}^{r-1} \Lambda_i = \cup_{j=0}^{m-1} \Lambda'_j = \Lambda$. For example, the output 2 of the 3×4 AWG in Fig. 3 is associated with the wavelength subset $\Lambda'_2 = \{\lambda_0, \lambda_2, \lambda_3\}$.

According to the wavelength and route assignment defined in (1), the $N_1 \times N_2$ AWG provides a total of $N_1 N_2$ fixed interconnection channels between N_1 inputs and N_2 outputs via a group of $|\Lambda|$ wavelengths. For example, the 12 interconnection channels of a 3×4 AWG are shown in Fig. 3. Larger AWGs can provide much richer interconnection channels. However, the results in [21] show that the AWG with a large port count suffers serious coherent crosstalk, which eventually imposes a limitation on the scalability of AWG-based interconnection network. Besides, the AWGs perform poorly outside its principal FSR due to channel loss imbalance and center frequency deviation [33], and thus we only consider the wavelengths in the principal FSR in this paper.

2.3 Wavelength Multiplexer (Mux) and Demultiplexer (DeMux)

The Mux is an essential component in modular AWG-based interconnection networks. An Mux can be considered as an $n \times 1$ (or $1 \times n$) AWG associated with N wavelengths, where $N = kn$ and k is a positive integer. An $n \times 1$ Mux can multiplex together n wavelength sets, and each set contains k wavelengths. A DeMux is an Mux in the reverse direction, which partitions a set of nk wavelengths into n disjoint wavelength subsets. It should be noted that, unlike AWG, neither Mux nor DeMux are affected by the coherent crosstalk [21]. In this paper, we use a triangle to represent a Mux or DeMux, as illustrated in Fig. 4.

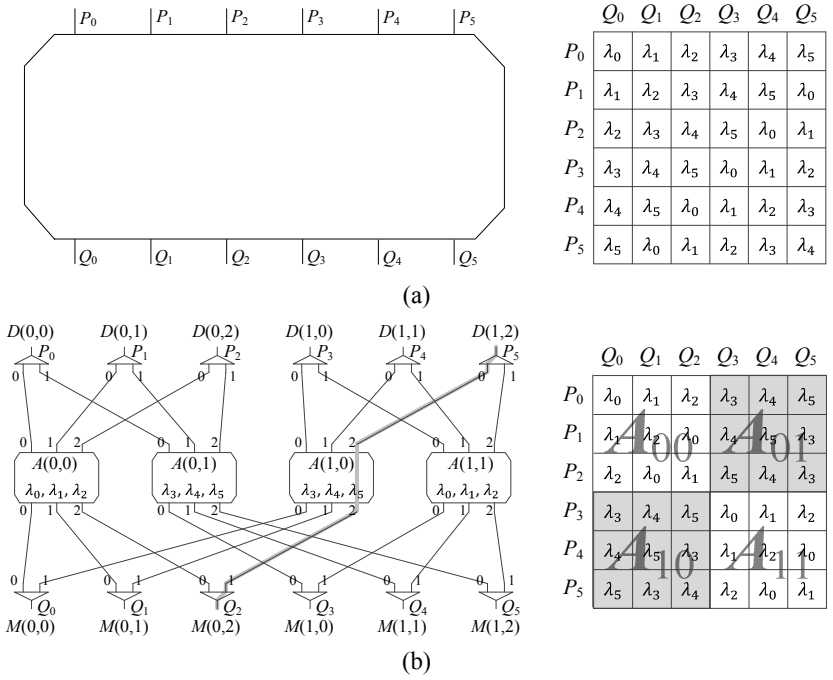
2.4 AWG-based Interconnection Networks

An $N_1 \times N_2$ AWG equipped with N_1 $N_2 \times 1$ Muxs at inputs and N_2 $1 \times N_1$ DeMuxs at outputs, as illustrated in Fig. 4, can be used to replace an $N_1 \times N_2$ Banyan-type $\mathcal{N}_A$. If a unique distinct wavelength is assigned to each transceiver in a node, then this AWG-based interconnection network, denoted by $\mathcal{N}_B$, can provide $N_1 N_2$ equivalent links based on the wavelength routing property (1).

Compare with the Banyan-type interconnection network $\mathcal{N}_A$ shown in Fig. 2, the equivalent AWG-based network $\mathcal{N}_B$ only requires $N_1 + N_2$ optical fibers to connect the nodes to the AWG, which implies that the interconnection complexity is reduced from $O(N_1 N_2)$ to $O(\max\{N_1, N_2\})$. This reduction is attributed to the fact that the interconnections in the $\mathcal{N}_B$ were absorbed by the wavelengths associated with the $N_1 \times N_2$ AWG according to the routing property defined in (1). Since the network $\mathcal{N}_A$ only requires one wavelength but the equivalent network $\mathcal{N}_B$ needs $\max\{N_1, N_2\}$ wavelengths, the complexity reduction is achieved at the expense of increasing the number of associated wavelengths. In a mega data center, the equivalent network $\mathcal{N}_B$ may suffer from a serious coherent crosstalk if the required wavelength granularity $\max\{N_1, N_2\}$ is too large. In the next section, we will propose a modular AWG-based interconnection scheme to cope with the scalability issue of network $\mathcal{N}_B$.

3 Modular AWG-based Interconnection Networks

In this section, we propose a method to construct an $N_1 \times N_2$ interconnection network by using a set of $r_1 \times r_2$ AWGs, where $r_1 \leq N_1$ and $r_2 \leq N_2$, such that the network is nonblocking and all AWGs use the same set of wavelengths with a smaller granularity $\max\{r_1, r_2\}$. Unlike space-division networks, in an AWG-based interconnection network, each link carries multiple wavelength channels. Thus,



	Q_0	Q_1	Q_2	Q_3	Q_4	Q_5
P_0	λ_0	λ_1	λ_2	λ_3	λ_4	λ_5
P_1	λ_1	λ_2	λ_3	λ_4	λ_5	λ_0
P_2	λ_2	λ_3	λ_4	λ_5	λ_0	λ_1
P_3	λ_3	λ_4	λ_5	λ_0	λ_1	λ_2
P_4	λ_4	λ_5	λ_0	λ_1	λ_2	λ_3
P_5	λ_5	λ_0	λ_1	λ_2	λ_3	λ_4

Fig. 5. A 6×6 AWG and its three-stage decomposition.

a nonblocking AWG-based interconnection network should fulfill the following condition [32]:

Nonblocking and Contention-free Principle of WDM Interconnection Networks: *Within the optical region, two connections share the same link cannot share the same wavelength, or two connections can use the same wavelength if they are link-disjoint.*

In this section, we demonstrate the construction of a modular AWG-based interconnection network in two steps. In the first step, we develop a systematic process to construct a three-stage $N_1 \times N_2$ AWG network, which is functionally equivalent to an $N_1 \times N_2$ AWG. The goal is to modularize the AWGs and partition the wavelength set while preserving the same routing property of the original AWG. In the second step, we show that the same wavelength set can be reused in each modular AWG by physically isolating the optical connections within the optical region. In the following, we first discuss the symmetric case when $N_1 = N_2 = N$, and then extend the method to the asymmetric case $N_1 \neq N_2$ in the next section.

3.1 AWG Decomposition

A modular decomposition of an AWG is valid only if it preserves the same routing property of the original AWG, as demonstrated by the example shown in Fig. 5. The routing table of a symmetric 6×6 AWG displayed in Fig. 5(a) is a cyclic Latin square, as described in Section 2.2. In the three-stage modular AWG network shown in Fig. 5(b), the upper ports are labeled by $P_0, P_1, \cdots, P_5$ and the lower ports are labeled by $Q_0, Q_1, \cdots, Q_5$. Each port connects to two 3×3 AWGs associated with two wavelength subsets $\{\lambda_0, \lambda_1, \lambda_2\}$ and $\{\lambda_3, \lambda_4, \lambda_5\}$, respectively. The routing table of the three-stage modular AWG network is also a Latin square, which is functionally equivalent to that of the original symmetric 6×6 AWG. The generalized decomposition of a cyclic Latin square is defined as follows.

Definition 1. Let $\mathbf{A}$ be an $N \times N$ cyclic Latin square matrix, where $N = rn$, then it can be partitioned into n^2 $r \times r$ cyclic Latin squares as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{00} & \cdots & \mathbf{A}_{0(n-1)} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{(n-1)0} & \cdots & \mathbf{A}_{(n-1)(n-1)} \end{pmatrix} = \begin{pmatrix} \mathbf{M}_0 & \cdots & \mathbf{M}_{(n-1)} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{(n-1)} & \cdots & \mathbf{M}_{(n-2)} \end{pmatrix} \quad (2)$$

where $\mathbf{A}_{ab} = \mathbf{M}_k$, for $a = 0, 1, \cdots, n-1$, $b = 0, 1, \cdots, n-1$ and $k = [a+b]_n$. It should be noted that these $r \times r$ cyclic Latin squares $\mathbf{M}_0, \mathbf{M}_1, \cdots, \mathbf{M}_{n-1}$ are respectively defined on n disjoint subsets of symbols.

The above decomposition can be further illustrated by the example shown in Fig. 5, where each block of the routing table of the AWG network corresponds to a 3×3 AWG in the middle stage. For instance, the routing table of the first 3×3 AWG that connects with the ports $P_0 \sim P_2$ and $Q_0 \sim Q_2$ is the block $\mathbf{A}_{00}$, which is the intersections of the rows $P_0 \sim P_2$ and the columns $Q_0 \sim Q_2$ in the original routing table $\mathbf{A}$. This example also clearly demonstrates that the topological decomposition of an $N \times N$ AWG into an equivalent $N \times N$ three-stage AWG network can actually be formulated by the proper partitioning of a cyclic Latin square.

Motivated by this observation, we propose a method, referred to as *matrix-based AWG decomposition*, to construct a nonblocking AWG network. Suppose that the wavelength set $\Lambda = \{\lambda_0, \lambda_1, \lambda_{N-1}\}$ associated with the AWG network is partitioned into n subsets $\Lambda_l = \{\lambda_{lr}, \lambda_{lr+1}, \cdots, \lambda_{(l+1)r-1}\}$, and $\mathbf{M}_l$ is the Latin square defined over the subset Λ_l,

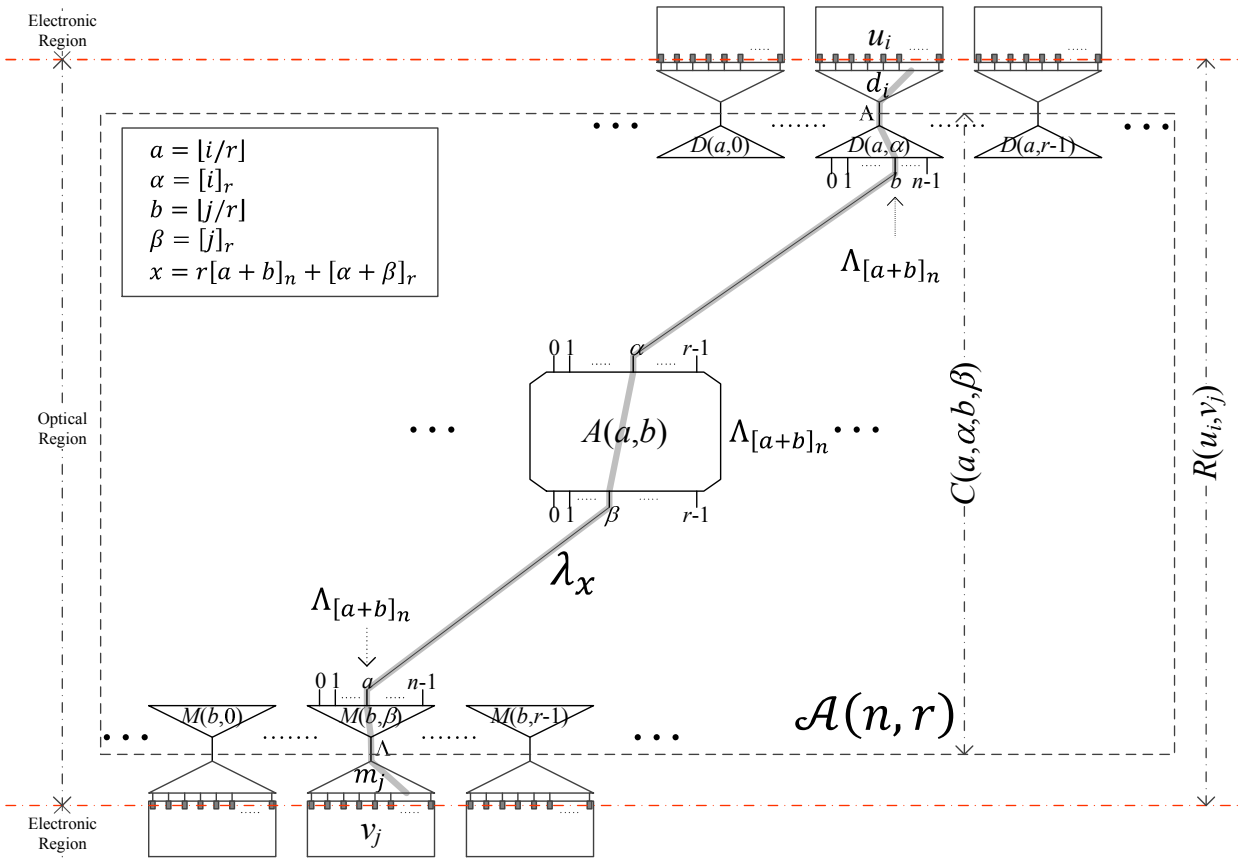


Fig. 6. The network $\mathcal{N}_C(n,r)$ with a three-stage AWG network $\mathcal{A}(n,r)$ in the center.

for $l = 0, 1, \cdots, n-1$. According to definition (2), the n Latin squares $\mathbf{M}_0, \mathbf{M}_1, \cdots, \mathbf{M}_{n-1}$ could immediately give rise to the $N \times N$ cyclic Latin square matrix $\mathbf{A}$, and the corresponding three-stage AWG network, which can be constructed by following three steps:

- Step 1 Layout n^2 $r \times r$ AWGs from left to right, and associate the kth AWG with the routing table $\mathbf{A}_{ab}$ and label it by $A(a,b)$, for $k = an + b = 0, 1, \cdots, n^2-1$. Note that $\mathbf{A}_{ab} = \mathbf{M}_{[a+b]_n}$ according to the definition given by (2), and thus the $A(a,b)$ is associated with the wavelength subset $\Lambda_{[a+b]_n}$.
- Step 2 Label the ith Mux at the upper layer by $D(a,\alpha)$ and connect its bth output port to the αth upper port of the $A(a,b)$, if the ith row of the matrix $\mathbf{A}$ is the αth row of the block $\mathbf{A}_{ab}$, where $a = \lfloor i/r \rfloor$, $\alpha = [i]_r$, and $b = 0, 1, \cdots, n-1$.
- Step 3 Label the jth DeMux at the lower layer by $M(b,\beta)$ and connect its ath input port to the βth lower port of the $A(a,b)$, if the jth column of the matrix $\mathbf{A}$ is the βth column of the block $\mathbf{A}_{ab}$, where $b = \lfloor j/r \rfloor$, $\beta = [j]_r$, and $a = 0, 1, \cdots, n-1$.

The above construction procedure and the numbering scheme of the decomposed three-stage AWG network, refer to as $\mathcal{A}(n,r)$, are illustrated by the connection pattern encircled by the dashed box shown in Fig. 6. It is easy to show from the routing property given in (1) that the AWG $A(a,b)$ can provide a unique connection, denoted by $C(a,\alpha,b,\beta)$, between the Mux $D(a,\alpha)$ and the DeMux $M(b,\beta)$ through the wavelength λ_x, along the following path:

$$\begin{aligned} \text{output } b \text{ of } D(a,\alpha) &\rightarrow \text{upper port } \alpha \text{ of } A(a,b), \\ &\rightarrow \text{lower port } \beta \text{ of } A(a,b) \\ &\rightarrow \text{input } a \text{ of } M(b,\beta) \end{aligned}$$

where

$$a = \lfloor i/r \rfloor, \tag{3}$$

$$\alpha = [i]_r, \tag{4}$$

$$b = \lfloor j/r \rfloor, \tag{5}$$

$$\beta = [j]_r, \tag{6}$$

and

$$x = r \times [a+b]_n + [\alpha+\beta]_r. \tag{7}$$

An example to illustrate the above decomposition procedure and numbering scheme is given in Fig. 5(b). The second 3×3 AWG is labeled by $A(1,0)$, because $2 = 1 \times 2 + 0$. The fifth row of the matrix $\mathbf{A}$ is the second row of the block $\mathbf{A}_{10}$ and $\mathbf{A}_{11}$, thus we denote the fifth upper Mux as $D(1,2)$, and connect its input 0 to the upper port 2 of the AWG $A(1,0)$ and its input 1 to the upper port 2 of the AWG $A(1,1)$. Similarly, the second column of the matrix $\mathbf{A}$ is the second column of the block $\mathbf{A}_{00}$ and $\mathbf{A}_{10}$, thus we denote the second lower Mux as $M(0,2)$, and connect its input 0 to the lower port 2 of the AWG $A(0,0)$ and its input 1 to the lower port 2 of the AWG $A(1,0)$. The $D(1,2)$

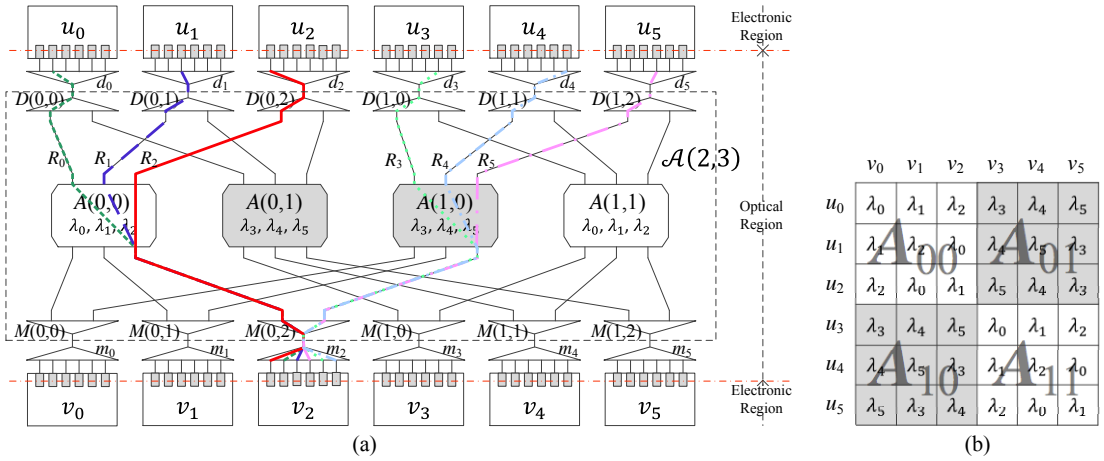


Fig. 7. An AWG interconnection network: (a) $\mathcal{N}_C(2,3)$ and (b) its routing table.

and the $M(0,2)$ are connected by the wavelength λ_4 since $[1+0]_2 \times 3 + [2+2]_3 = 4$.

The above example show that an ingress port in the three-stage AWG network $A(n,r)$ can be connected to an egress port via a wavelength in the set Λ, which is equivalent to the function of an $N \times N$ AWG. This equivalence relationship is formally stated in the following theorem.

Theorem 1. The three-stage network $\mathcal{A}(n,r)$ is nonblocking and functionally equivalent to an $N \times N$ AWG.

Proof: Suppose that two connections $C_1(a_1, \alpha_1, b_1, \beta_1)$ and $C_2(a_2, \alpha_2, b_2, \beta_2)$ originate from the same upper layer Mux $i = D(a, \alpha)$:

$$\begin{cases} a_1 = a_2 = a \\ \alpha_1 = \alpha_2 = \alpha \end{cases},$$

and use the same wavelength. According to (7), we have

$$r \times [a + b_1]_n + [\alpha + \beta_1]_r = r \times [a + b_2]_n + [\alpha + \beta_2]_r,$$

which implies

$$[a + b_1]_n = [a + b_2]_n,$$

and

$$[\alpha + \beta_1]_r = [\alpha + \beta_2]_r.$$

We thus have $b_1 = b_2$ and $\beta_1 = \beta_2$. It follows that C_1 and C_2 must be the same connection, which implies that there is no wavelength contention at each upper layer Mux. Following the same argument, we can prove that each lower layer DeMux is also contention-free. The routing property of AWGs given in (1) guarantees that all $r \times r$ AWGs in middle are contention-free. Thus, the three-stage network $\mathcal{A}(n,r)$ is nonblocking. □

Replacing the $N \times N$ AWG in the network $\mathcal{N}_B$ with the $\mathcal{A}(n,r)$, we obtain an equivalent network, denoted by $\mathcal{N}_C(n,r)$, as shown in Fig. 6. A 6×6 network $\mathcal{N}_C(2,3)$ is plotted in Fig. 7(a) as an example, of which the Latin square routing table is shown in Fig. 7(b). In the interconnection network $\mathcal{N}_C(n,r)$, the node u_i connects to the $D(a,\alpha)$ through the $1 \times N$ Mux d_i if $i = ar + \alpha$, and the node v_j connects to the $M(b,\beta)$ through the $1 \times N$ Mux m_j if $j = br + \beta$. To set up a connection from the node u_i to the node v_j, denoted by $R(u_i, v_j)$, a path $C(a, \alpha, b, \beta)$ should be established in network $\mathcal{A}(n,r)$.

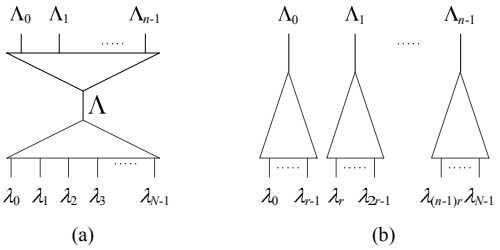


Fig. 8. The equivalence between (a) a pair of back-to-back $n \times 1$ DeMux and $1 \times N$ Mux and (b) n $1 \times r$ Muxs.

Unlike the single AWG within the network $\mathcal{N}_B$, the scheme $\mathcal{N}_C(n,r)$ employs modular AWGs and partitions the wavelength set Λ into n subsets $\Lambda_0, \Lambda_1, \cdots, \Lambda_{n-1}$ to enhance the scalability of the network when N is large. Though the number of the long fiber links in the $\mathcal{N}_C(n,r)$ is n times of that in the $\mathcal{N}_B$, but it is only about $1/r$ of that in the $\mathcal{N}_A$.

Despite that, the wavelength granularity of the $\mathcal{N}_C(n,r)$ remains the same, namely N. This is due to the fact that N connections terminated on the same node require N different wavelengths in order to cross the same pair of back-to-back $N \times 1$ Mux and $1 \times n$ DeMux within the optical region. For example, the connections $R_0 \sim R_5$ in Fig. 7 use 6 different wavelengths because they cross the same Mux $M(0,2)$ and DeMux m_2. The issue on the reduction of wavelength granularity will be addressed in the next subsection.

3.2 Wavelength Reuse

According to the nonblocking and contention-free condition of WDM interconnection networks, only disjoint optical connections in the network can use the same wavelength set. The reduction of wavelength granularity of the network $\mathcal{N}_C(n,r)$ can only be achieved by wavelength reuse, which requires spatial separation of optical connections terminated on the same switching node.

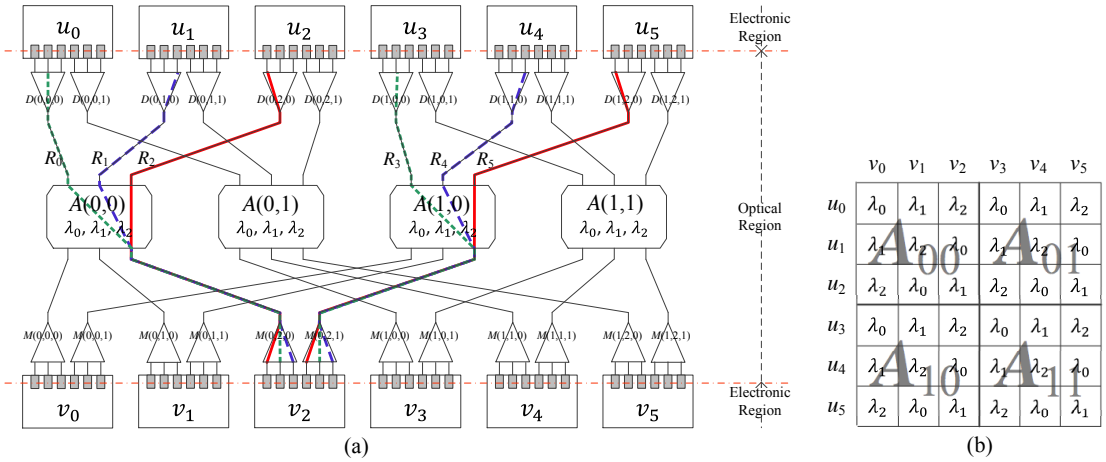


Fig. 9. An wavelength-reused AWG interconnection network: (a) $\mathcal{N}_D(2,3)$ and (b) its routing table.

As illustrated in Fig. 7, each node in the network $\mathcal{N}_C(n,r)$ is attached by a pair of back-to-back $N \times 1$ Mux and $1 \times n$ DeMux whose function is to fully demultiplex the wavelength subsets $\Lambda_0, \Lambda_1, \cdots, \Lambda_{n-1}$. As shown in Fig. 7, the Mux $M(0,2)$ multiplexes the wavelength subsets $\Lambda_0 = \{\lambda_0, \lambda_1, \lambda_2\}$ and $\Lambda_1 = \{\lambda_3, \lambda_4, \lambda_5\}$ to form the complete set $\Lambda = \{\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$, which is then demultiplexed by the Mux m_2. It is obvious that this function can also be separately achieved by n $1 \times r$ Muxs, as shown in Fig. 8, each of which independently demultiplexes one wavelength subset.

Thus, if each Mux pair is replaced with n $r \times 1$ Muxs or n $1 \times r$ DeMuxs in the $\mathcal{N}_C(n,r)$, then connections terminated on the same node become disjoint and can use the same wavelength sets. Specifically, the pair d_i and $D(a,\alpha)$ are replaced by n $r \times 1$ Muxs $D(a,\alpha,0), D(a,\alpha,1), \cdots, D(a,\alpha,n-1)$, and the pair m_i and $M(b,\beta)$ by n $1 \times r$ DeMuxs $M(b,\beta,0), M(b,\beta,1), \cdots, M(b,\beta,n-1)$. In the resulting network, denoted by $\mathcal{N}_D(n,r)$, the connections originated or terminated on the same node are divided into n mutually link-disjoint groups, who can use the same wavelength subset, says $\Lambda_0 = \{\lambda_0, \lambda_1, \cdots, \lambda_{r-1}\}$.

For example, the network $\mathcal{N}_D(2,3)$ shown in Fig. 9(a) is obtained by replacing each pair of back-to-back 2×1 Mux and 1×6 DeMux in $\mathcal{N}_C(2,3)$ by two 3×1 Muxs. As a result, the connections $R_0 \sim R_2$ and $R_3 \sim R_5$ in Fig. 9(a) are link-disjoint and they can use the same wavelength set $\{\lambda_0, \lambda_1, \lambda_2\}$.

If all $r \times r$ modular AWGs in the network $\mathcal{N}_D(n,r)$ are associated with the same wavelength set Λ_0, the $\mathcal{N}_D(n,r)$ has a much smaller wavelength granularity r, and the routing table is degenerated to a matrix of blocks of cyclic Latin square M_0. For example, the wavelength granularity of the $\mathcal{N}_D(2,3)$ is 3, and its routing table is a block matrix given by Fig. 9(b).

In the network $\mathcal{N}_D(n,r)$, as illustrated in Fig. 10, all $r \times 1$ Muxs, or all $1 \times r$ DeMuxs, attached to the same node use the same wavelength set. Therefore, in the connection $R(u_i, v_j)$ from the node u_i to the node v_j, a Mux attached to the node u_i and a DeMux attached to the node v_j must be specified. The following sequence exhibits the complete path from the node u_i to the node v_j:

$$\begin{aligned} u_i &\to D(a,\alpha,b) \\ &\to \text{upper port } \alpha \text{ of } A(a,b), \\ &\to \text{lower port } \beta \text{ of } A(a,b) \\ &\to M(b,\beta,a) \\ &\to v_j, \end{aligned}$$

via the wavelength $\lambda_{x'}$, where

$$x' = [\alpha + \beta]_r. \tag{8}$$

The following theorem demonstrates that the network $\mathcal{N}_D(n,r)$ is also contention-free, similar to that of the three-stage network $\mathcal{A}(n,r)$ given in Theorem 1.

Theorem 2. The interconnection network $\mathcal{N}_D(n,r)$ is nonblocking.

Proof: Suppose that two routes $R_1(u_{i_1}, v_{j_1})$ and $R_2(u_{i_2}, v_{j_2})$ originate from the same Mux $i = D(a,\alpha,b)$ attached to the node u_i:

$$\begin{cases} a_1 = a_2 = a \\ \alpha_1 = \alpha_2 = \alpha \\ b_1 = b_2 = b \end{cases}.$$

If we use the same wavelength, according to (8), we have

$$[\alpha + \beta_1]_r = [\alpha + \beta_2]_r,$$

which implies

$$\beta_1 = \beta_2.$$

It follows that R_1 and R_2 must be the same connection, since they both reach the same Mux $M(b,\beta,a)$ that is attached to the node v_j. Therefore, wavelength contention will never occur at each upper-layer $r \times 1$ Mux. Similarly, we can show that each lower-layer $r \times 1$ Mux is also contention-free. The wavelength routing property given in (1) guarantees that those $r \times r$ AWGs in the middle are all contention-free. Thus, the network $\mathcal{N}_D(n,r)$ is nonblocking and contention-free. □

Note that the three-stage interconnection network $\mathcal{N}_A$ and the single AWG-base network $\mathcal{N}_B$ are two extreme cases of the network $\mathcal{N}_D(n,r)$:

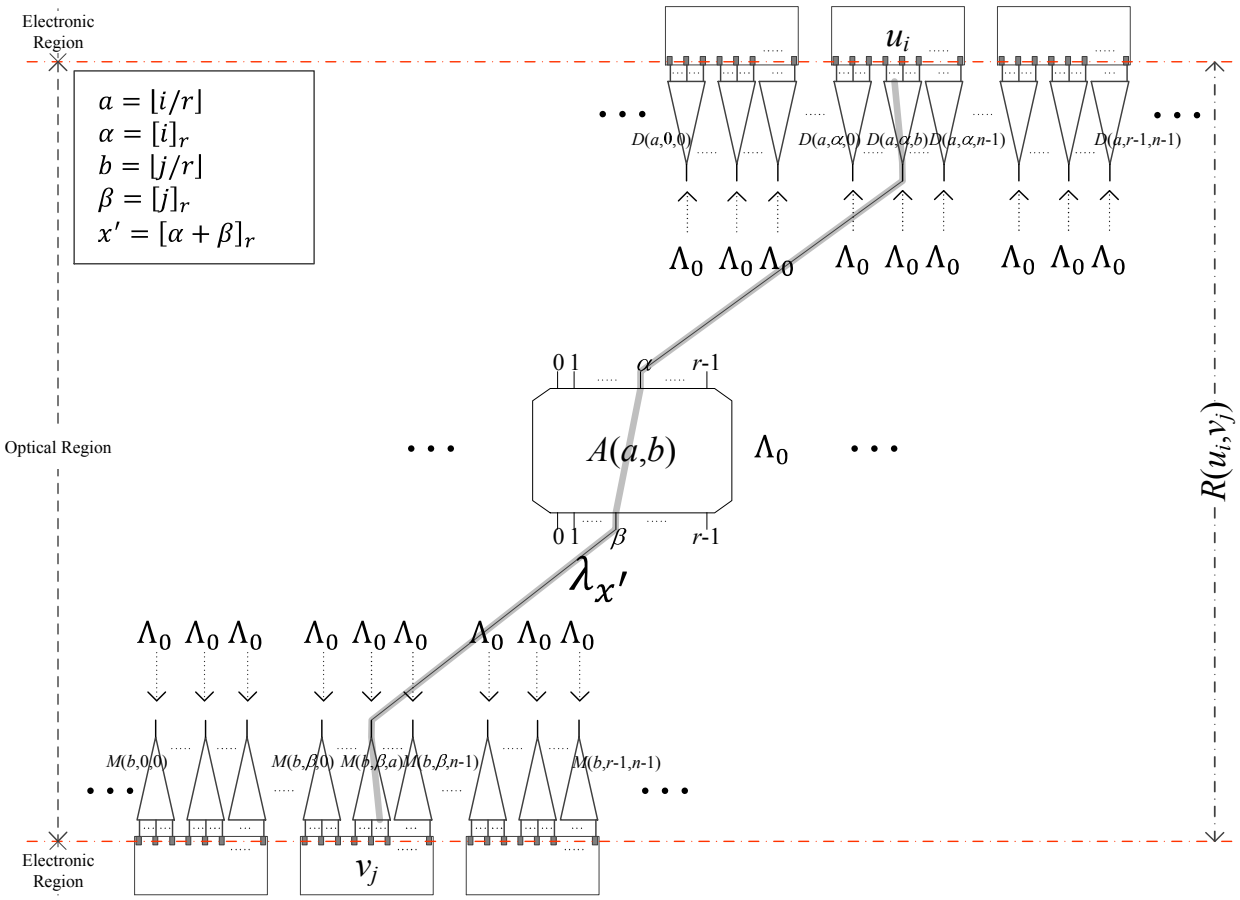


Fig. 10. The network $\mathcal{N}_D(n,r)$: numbering scheme and a connection.

1) The three-stage interconnection network $\mathcal{N}_A$ can be regarded as the network $\mathcal{N}_D(N,1)$. If $r = 1$ and $n = N$, not only each $r \times r$ AWG but also each $1 \times r$ Mux shrinks to a link, in which case the network $\mathcal{N}_D(n,r)$ reduces to an $\mathcal{N}_A$.
2) The AWG-base network $\mathcal{N}_B$ is the same as network $\mathcal{N}_D(1,N)$. If $r = N$ and $n = 1$, then there is only a single $N \times N$ AWG, and each $1 \times r$ Mux becomes a $1 \times N$ Mux, in which case the network $\mathcal{N}_D(N,1)$ becomes the single AWG-base network $\mathcal{N}_B$.

The network $\mathcal{N}_D(n,r)$ is actually a compromise of the interconnection network $\mathcal{N}_A$ and the single AWG-based network $\mathcal{N}_B$. The network $\mathcal{N}_D(N,1)$, or equivalently $\mathcal{N}_A$, establishes the N^2 connections through the optical fibers. Hence, the $\mathcal{N}_D(N,1)$ has the highest interconnection complexity, but the smallest wavelength granularity. On the other hand, the $N \times N$ AWG in the $\mathcal{N}_D(1,N)$, or equivalently $\mathcal{N}_B$, can provide N^2 connections between its inputs and outputs through a group of N wavelengths. Thus, the $\mathcal{N}_D(1,N)$ requires the least number of optical fibers. However, it suffers from most serious coherent crosstalk and its wavelength granularity N is the largest. This comparison is listed in Table 1.

If the parameter r is still too large in the compromised network $\mathcal{N}_D(n,r)$, current state-of-art technology of photonics integrated circuits allows substituting each $r \times r$ AWG by an encapsulated three-stage $r \times r$ AWG network to further enhance the system scalability.

TABLE 1
Comparison of three interconnection networks.

	AWG Size	Wavelength Granularity	Cabling Complexity
$\mathcal{N}_A$ ($\mathcal{N}_D(1,N)$)	1×1	$O(1)$	$O(N^2)$
$\mathcal{N}_B$ ($\mathcal{N}_D(N,1)$)	$N \times N$	$O(N)$	$O(N)$
($\mathcal{N}_D(n,r)$)	$r \times r$	$O(r)$	$O(Nn)$

4 Asymmetric Interconnection Networks

Our results can also be applied to the asymmetric interconnection networks. Without loss of generality, we assume that $N_2 = N > N_1$. In general, the asymmetric network can be obtained by removing $N - N_1$ nodes from the node set U of the symmetric network $\mathcal{N}_D(n,r)$. We consider two particular cases as follows.

4.1 N and N_1 Are Co-prime

An example of this case is demonstrated in Fig. 11, where a 5×6 AWG-based interconnection network can be constructed by removing u_5 from the $\mathcal{N}_D(2,3)$ in Fig. 11(a). Accordingly, the routing table of the 5×6 network is obtained by deleting the fifth row from the table displayed in Fig. 11(b). Furthermore, some AWGs in the middle stage become asymmetric after the node deletion. In this example, both $A(1,0)$ and $A(1,1)$ become a 2×3 AWG.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCC.2016.2522417, IEEE Transactions on Cloud Computing

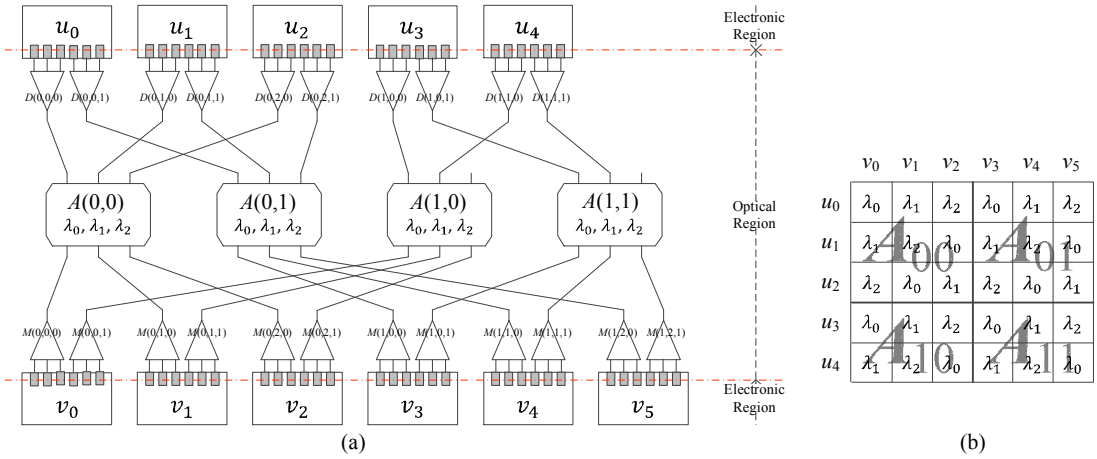


Fig. 11. An asymmetric AWG-based interconnection network: (a) 5×6 network and (b) its routing table.

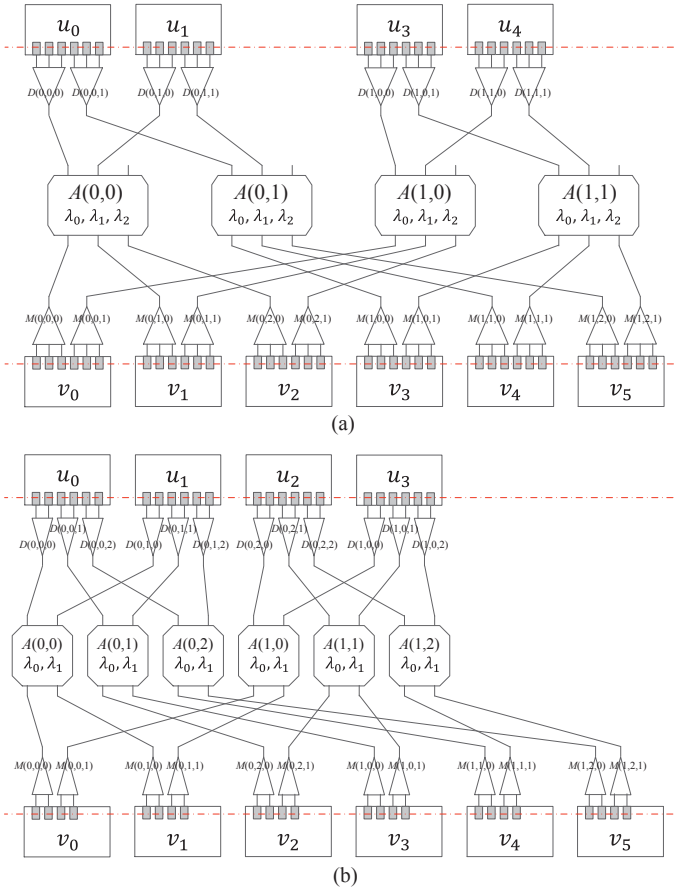


Fig. 12. Construction of a 4×6 AWG-based interconnection network, using (a) the first method and (b) the second method.

4.2 N and N_1 Are Not Co-prime

In this case, we suppose that N and N_1 have a non-trivial greatest common divisor (gcd) n, and assume that $N = rn$ and $N_1 = (r-d)n$, where r and d are positive integers. We show that an $N_1 \times N$ AWG-based asymmetric network can be constructed in two ways, which are similar to that in [23].

4.2.1 *First Method*

We first construct an AWG-based network $\mathcal{N}_D(n, r)$, and repeatedly remove the set of nodes

$$\{u_{ir-j} | i = 1, 2, \cdots, n; j = 1, 2, \cdots, d\},$$

such that each $r \times r$ AWG reduces to an $(r-d) \times r$ AWG. In the 4×6 network plotted in Fig. 12(a), $N_1 = 4$ and $N = 6$ have a gcd $n = 2$. Thus, this asymmetric network can be constructed by deleting the nodes u_2 and u_5 from the network $\mathcal{N}_D(2, 3)$ in Fig. 9, in which each 3×3 AWG reduces to a 2×3 AWG.

4.2.2 *Second Method*

We first construct an AWG-based network $\mathcal{N}_D(r, n)$, and remove dn nodes from right to left, such that the number of $n \times n$ AWGs in the middle can be minimized to $(r-d)r$. For example, the 4×6 network shown in Fig. 12(b) is obtained by deleting the nodes u_4 and u_5 from the network $\mathcal{N}_D(3, 2)$, in which the number of 2×2 AWGs is 6.

5 Applications to Data Center Networks

In this section, we discuss the application of the modular AWG-based interconnection to reduce the cabling complexity of data center networks, while preserving the function and bandwidth of the original network. We also show that the deployment of this scheme in practice is feasible in respect to physical layer transmission performance and system cost.

5.1 Reduction of Cabling Complexity

As an example, we take the flatten Butterfly data center network to quantitatively evaluate the effectiveness of the proposed interconnection scheme in this section. The results can also be applied to other data center networks, such as Fat Tree and BCube. An N^2-node 2-D flatten Butterfly network [34] is a network with N rows and N columns, as shown in Fig. 13, in which each row or each column contains N nodes and they are fully connected via N^2 links. Totally, there are $2N^3$ links in a flatten Butterfly network with N^2 nodes. The wavelength granularity of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCC.2016.2522417, IEEE Transactions on Cloud Computing

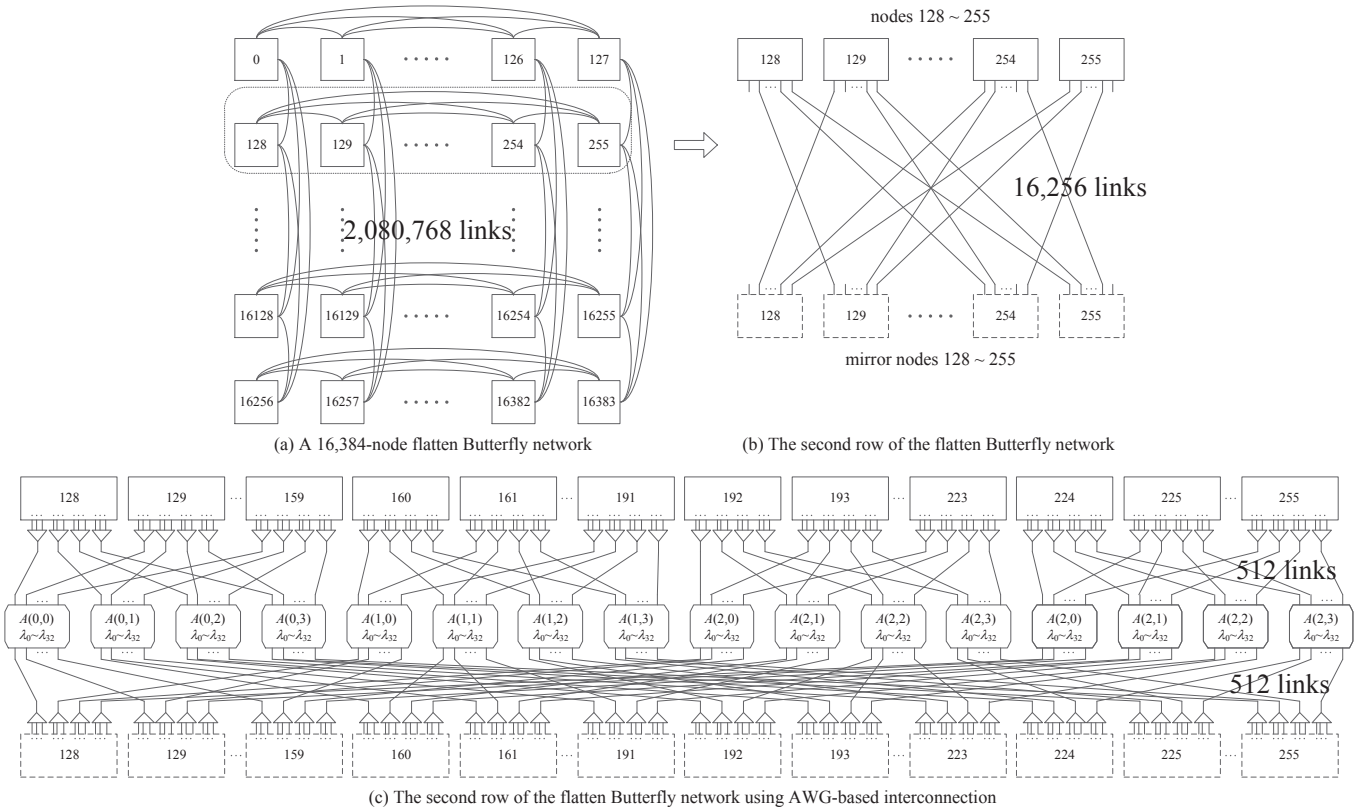


Fig. 13. Application of the AWG-based interconnection scheme in a 16,384-node 2-D flatten Butterfly data center network.

network is 1, because each link carries only one wavelength. For example, the network plotted in Fig. 13(a) is a flatten Butterfly network with $N = 128$. In this network, there are 16,384 nodes and 2,080,768 links, each of which only carries one wavelength channel.

If a modular AWG-based interconnection network $\mathcal{N}_D(n, r)$ is used to replace the N^2 links in each row and each column, where $nr = N$, then the total number of links will be reduced by a factor of $r/2$ at the expenses of increasing the wavelength granularity of data center networks. An example is illustrated in Fig. 13, where the links in each row (or each column) of a flatten butterfly network with 16,384 nodes is replaced by a three-stage AWG network $\mathcal{N}_D(4, 32)$. As plotted in Fig. 13(c), the total number of links is reduced by a factor of 16. At the same time, the wavelength granularity of the network will increase to 32, because each AWG in the $\mathcal{N}_D(4, 32)$ network is a 32×32 AWG which requires transceivers associated with 32 wavelengths.

As we mentioned in Section 3, there is a trade-off between the port count r of AWGs in the central stage and the number of links of the network. The trade-off between the port count r of AWGs in the central stage and the number of links in a flatten Butterfly network is plotted in Fig. 14(a), which shows that the number of links can be substantially reduced when the parameter r is large. In Fig. 14(b), we show that the wavelength granularity of flatten Butterfly networks is a constant determined by the parameter r of the AWG-based interconnection network $\mathcal{N}_D(n, r)$.

It should be noted that the coherent crosstalk and the passband deviation of $r \times r$ AWGs in the $\mathcal{N}_D(n, r)$ network would be pronounced when the parameter r becomes too large, say $r \geq 128$. This problem can be solved by recursively replacing each $r \times r$ AWG with a modular AWG-based network $\mathcal{N}_C(n', r')$ where $n'r' = r$. The links of each $\mathcal{N}_C(n', r')$ network will not increase the cabling complexity of entire data center networks because these internal links are encapsulated on a chip of each module.

5.2 Issues on Fault-Tolerance

Since a single failure in the AWG-based interconnection network may disconnect multiple connections, in this subsection, we consider the network survivability issue pertaining to the application of our proposal. Almost all existing fault-tolerance methods in data center networks [30], [31], [35]–[39] are achieved by making a detour around the failure part of the network, and the same principle can still be applied to AWG-based data center networks. Using a Fat-tree network with modular AWG-based interconnection as a particular example, we will illustrate the process to enhance the survivability of AWG-based data center networks by using existing rerouting methods [30], [31], [35]–[39].

In principle, to cope with network failures, the network should provide enough redundant resources and plot alternating paths to reroute affected connections. Up to date, there exist two kinds of methods in published literatures. In the first method, the network will compute a new path for each affected connection to isolate those failed links [31], [35]–[38]. The major drawback of this kind of method is that the affected connections may suffer a delay incurred by on-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCC.2016.2522417, IEEE Transactions on Cloud Computing

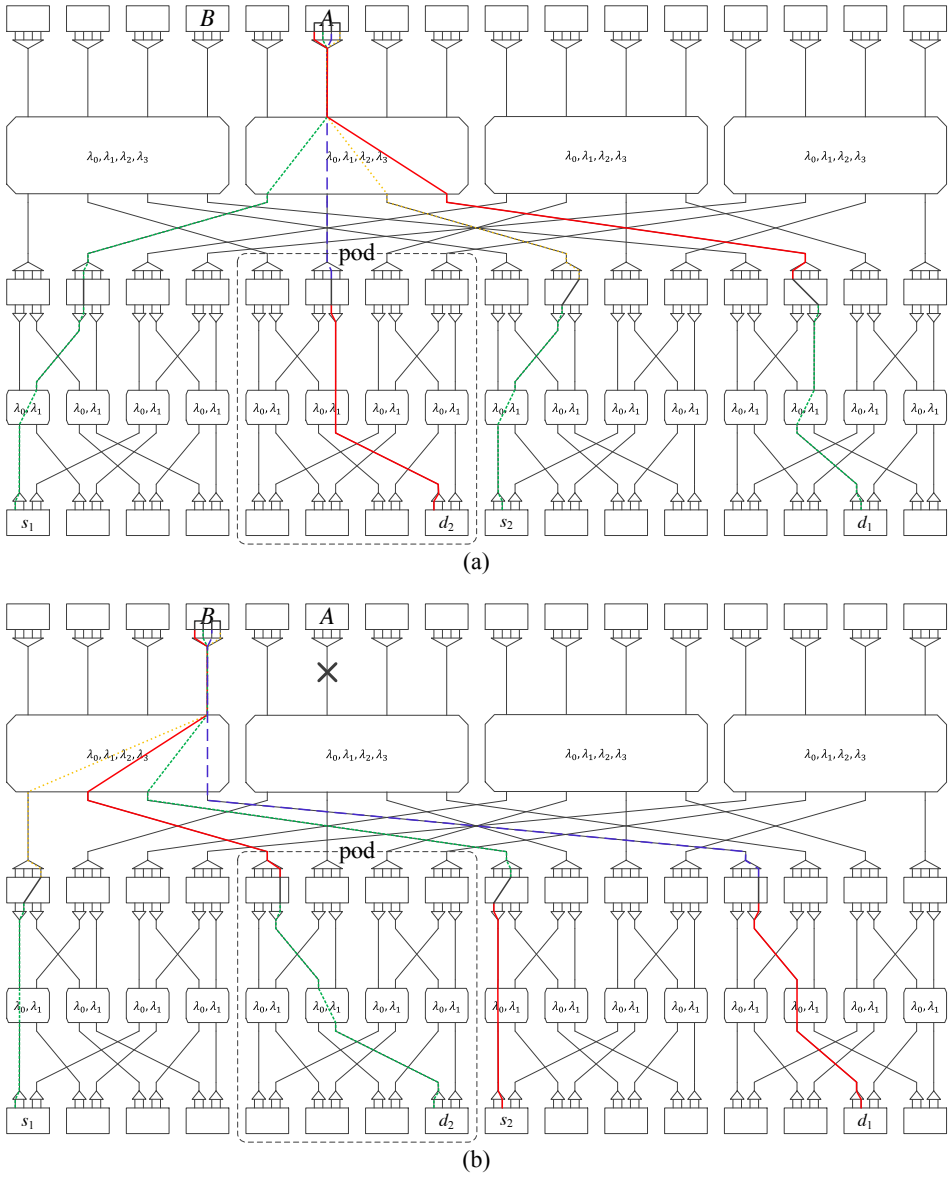


Fig. 15. The detour around a link failure in a Fat-tree network with modular AWG-based interconnection.

line calculation of alternating route. The other kind method is a protection scheme, the backup routes are calculated before the failures occur [30], [39], and thus the communication can be recovered much faster.

Both methods described above can be applied to AWG-based data center networks, as long as they provide the same topological interconnection and communication bandwidth as the original data center networks. We illustrate this point in Fig. 15, which plots a Fat-tree network with modular AWG-based interconnection. In Fig. 15(a), two connections, one from s_1 to d_1 and the other from s_2 to d_2, were established through the node A. If the first method is adopted when the outgoing optical link of node A fails, using the rerouting algorithms proposed in [31], [35]–[38], the network would reroute the connections via another node, say node B, as illustrated in Fig. 15(b). If the second method is employed, the network would calculate backup paths for these two connections by using the methods developed in [30], [39] at the time when they were established via node A.

5.3 Performance and Cost

In this subsection, we demonstrate our modular approach is feasible in terms of physical-layer transmission performance and system cost, when it is applied to a Mega data center. In particular, we evaluate the end-to-end transmission performance of the AWG module $\mathcal{N}_D(4, 32)$, as depicted in Fig. 13(c), using the widely adopted simulation software Optisystem.

As illustrated by the thick dotted link shown in Fig. 13(c), each optical connection may interfere with 31 other connections that share the same 32×32 AWG module. As plotted in Fig. 16(a), we simulate a connection that involves a 32×1 input AWG, a 32×32 AWG module, and a 1×32 output AWG, of which the insertion loss is 4 dB, the channel spacing is 0.08 nm, and each channel has a Gaussian-type passband with 50-GHz 3-dB bandwidth. To mimic a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCC.2016.2522417, IEEE Transactions on Cloud Computing

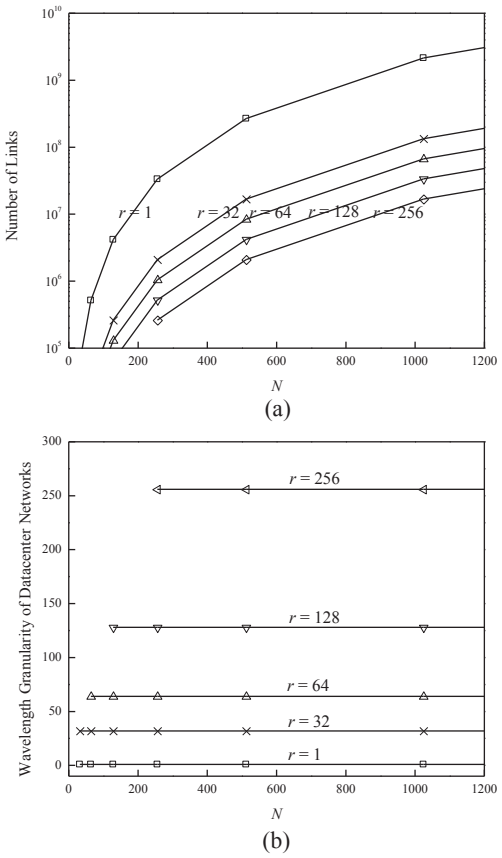


Fig. 14. Number of links and wavelength granularity of flatten Butterfly data center networks.

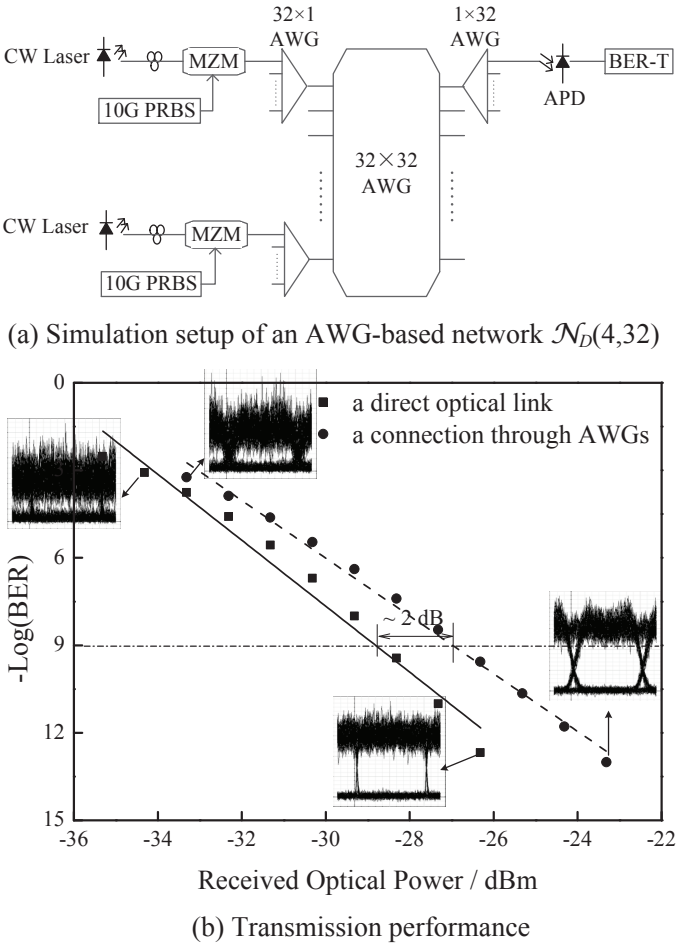


Fig. 16. Transmission performance of a connection in a 32×32 AWG-based network $\mathcal{N}_D(4, 32)$.

connection from a 32×1 input AWG to a 1×32 output AWG, we use a 10-Gbps pseudo random binary sequence (PRBS) non-return-to-zero (NRZ) optical signal at 1552.52 nm, which is finally detected by an avalanche photo diode (APD). To take the coherent crosstalk of the 32×32 AWG into consideration, we also feed 10-Gbps PRBS NRZ signals at 1552.52 nm to other input ports of the 32×32 AWG. The launch power of these optical signals is 6 dBm.

The performance of a connection in the AWG-based network is compared with that of a directly connected optical fiber. The simulation results presented in Fig. 16(b) indicate that the power penalty introduced by the 32×1 input AWG, the 32×32 AWG module, and the 1×32 output AWG is only $\sim$ 2 dB, which confirms that the proposed modular AWG-based interconnection scheme is feasible in practice when it is applied to a Mega data center network, even as large as a 16,384-node flatten Butterfly network.

When the AWG-based network is used to simplify the cabling complexity of data center networks, the network cost may be increased due to the following two reasons:

1) Currently, the existing WDM devices, such as AWG, Mux, and DeMux, are much more expensive than optical fibers.
2) The wavelength granularity of a data center network is one if it is interconnected by single wavelength optical fibers. However, the wavelength granularity will be much larger if the data center is interconnected by WDM networks, which require transceivers associated with different wavelengths. This extra requirement will also increase network cost.

Since the cost of each individual WDM component does not grow with the network size, and the additional cost incurred by WDM devices is linearly proportional to the port count, the cost of AWG-based interconnections can be justified for large data center networks when the cabling complexity is very high. That is, the modular AWG-based interconnection scheme proposed in our paper provides a compromise solution to strike a balance between the network cost and complexity. The key idea employed in our scheme is to adjust the design parameter r, which represents the port count of AWGs in the central stage. A large parameter r can be used if the complexity of cabling is the critical issue, otherwise, AWG-based networks with smaller r can be adopted if the network cost is the service provider's primary concern.

6 Conclusion

To reduce the cabling complexity of mega data center networks, this paper proposes a modular scheme to construct the AWG-based interconnection network, which possesses the following features. First, compared to the original interconnection network, it provides the same interconnection

function with a much smaller cabling complexity. Second, it has an improved scalability and reliability, since the interconnection scheme comprises small-size AWGs, which are all associated with the same wavelength set with small granularity. Furthermore, our results show that it can provide needed flexibility in the design of ultra-large data centers in the future.

ACKNOWLEDGMENT

We would like to thank Mr. Kuo Zhang for his help on physical-layer simulations.

REFERENCES

[1] C. Kachris and I. Tomkos, "A survey on optical interconnections for data centers," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 1021–1036, Oct. 2012.
[2] X. Zhao, V. Vusirikala, B. Koley, V. Kamalov, and T. Hofmeister, "The prospect of inter-data-center optical networks," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 32–38, Sep. 2013.
[3] M. F. Bari *et al.*, "Data center network virtualization: a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 909–928, May 2013.
[4] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. ACM SIGCOMM*, Aug. 2008.
[5] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: a cost-effective topology for high-radix networks," in *Proc. ACM ISCA*, Jun. 2007.
[6] N. Farrington, E. Rubow, and A. Vahdat, "Data center switch architecture in the age of merchant silicon," in *Proc. IEEE HOTI*, Aug. 2009.
[7] "Optimized airflow assessment for cabling," Accessment, IBM, 2007. [Online]. Available: http://www-935.ibm.com/services/us/its/pdf/oaac-ds-gtd01332-usen-00-053107.pdf
[8] "Data center top-of-rack architecture design," White Paper, Cisco, 2009.
[9] K. Ramachandran, R. Kokku, R. Mahindra, and S. Rangarajan, "60 GHz data-center networking: Wireless => worry less?" Technical Report, NEC, 2008.
[10] S. Kandula, J. Padhye, and P. Bahl, "Flyways to de-congest data center networks," in *Proc. IEEE HOTI*, Oct. 2009.
[11] D. Halperin *et al.*, "Augmenting data center networks with multi-gigabit wireless links," in *Proc. ACM SIGCOMM*, Aug. 2011.
[12] Y. Cui, H. Wang, X. Cheng, and B. Chen, "Wireless data center networking," *IEEE Wireless Commun. Mag.*, vol. 18, no. 6, pp. 46–53, Dec. 2011.
[13] L. Shan *et al.*, "Relieving hotspots in data center networks with wireless neighborways," in *Proc. IEEE GLOBECOM*, Dec. 2014.
[14] J. Shin, E. G. Sirer, H. Weatherspoon, and D. Kirovski, "On the feasibility of completely wireless datacenters," *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 1666–1680, Oct. 2013.
[15] W. Zhang *et al.*, "3D beamforming for wireless data centers," in *Proc. ACM Workshop on Hot Topics in Networks (HotNets)*, Nov. 2011.
[16] X. Zhou *et al.*, "Mirror mirror on the ceiling: flexible wireless links for data centers," in *Proc. ACM Sigcomm*, Oct. 2012.
[17] Y. Cui, H. Wang, and X. Cheng, "Channel allocation in wireless data center networks," in *Proc. IEEE INFOCOM*, Apr. 2012.
[18] N. Hamedazimi *et al.*, "Firefly: A reconfigurable wireless data center fabric using free-space optics," in *Proc. ACM SIGCOMM*, Oct. 2014.
[19] Z. Cao, R. Proietti, M. Clements, and S. J. B. Yoo, "Demonstration of scalable, flat, and high-throughput data center architecture based on arrayed waveguide grating routers," in *Proc. ECOC*, Sep. 2014.
[20] M. Csernai, F. Ciucu, R. P. Braun, and A. Gulyas, "Reducing cabling complexity in large flattened butterfly networks by an order of magnitude," in *Proc. OFC/NFOEC*, Mar. 2014.
[21] R. Gaudino, G. A. G. Castillo, F. Neri, and J. M. Finochietto, "Simple optical fabrics for scalable terabit packet switches," in *Proc. IEEE ICC*, 2008, pp. 5331–5337.
[22] G. Weichenberg, V. W. S. Chan, and M. Medard, "Design and analysis of optical flow-switched networks," *IEEE J. Opt. Commun. Netw.*, vol. 1, no. 3, pp. B81–B97, Aug. 2009.
[23] T. Ye, T. T. Lee, and W. Hu, "A study of modular AWGs for large-scale optical switching systems," *J. Lightw. Technol.*, vol. 30, no. 13, pp. 2125–2133, Jun. 2012.
[24] Y. Yin *et al.*, "AWGR-based all-to-all optical interconnects using limited number of wavelengths," in *Proc. Optical Interconnects Conference*, May 2013, p. TuB4.
[25] G. Wu *et al.*, "A scalable awg-based data center network for cloud computing," *Optical Switching and Networking*, vol. 16, pp. 46–51, 2015.
[26] K. Wang *et al.*, "ADON: a scalable awg-based topology for data center optical network," *Opt. Quant. Electron*, vol. 47, pp. 2541–2554, 2015.
[27] K. Chen *et al.*, "WaveCube: A scalable, fault-tolerant, high-performance optical data center architecture," in *Proc. IEEE INFOCOM*, 2015.
[28] T. Niwa *et al.*, "Large port count wavelength routing optical switch consisting of cascaded small-size cyclic arrayed waveguide gratings," *IEEE Photon. Technol. Lett.*, vol. 24, no. 22, pp. 2027–2030, 2012.
[29] C. Clos, "A study of nonblocking switching networks," *Bell System Technology Journal*, vol. 32, no. 2, pp. 406–424, Mar. 1953.
[30] C. Guo *et al.*, "BCube: a high performance, server-centric network architecture for modular data centers," in *Proc. ACM SIGCOMM*, Aug. 2009, pp. 63–74.
[31] ——, "DCell: a scalable and fault-tolerant network structure for data centers," in *Proc. ACM SIGCOMM*, Aug. 2008, pp. 75–86.
[32] T. Ye, T. T. Lee, and W. Hu, "AWG-based nonblocking Clos networks," *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 491–504, Apr. 2015.
[33] S. Kamei, M. Ishii, A. Kaneko, T. Shibata, and M. Itoh, "NxN cyclic-frequency router with improved performance based on arrayed-waveguide grating," *J. Lightw. Technol.*, vol. 27, no. 18, pp. 4097–4014, 2009.
[34] Z. Zhu, S. Zhong, L. Chen, and K. Chen, "Fully programmable and scalable optical switching fabric for petabyte data center," *Opt. Express*, vol. 23, no. 3, pp. 3563–3580, Feb. 2015.
[35] F. O. Sem-Jacobsen, T. Skeie, O. Lysne, and J. Duato, "Dynamic fault tolerance in fat trees," *IEEE Trans. Comput.*, vol. 60, no. 4, pp. 508–525, Apr. 2011.
[36] V. Liu, D. Halperin, A. Krishnamurthy, and T. Anderson, "F10: A fault-tolerant engineered network," in *Proc. NSDI*, 2013, pp. 399–412.
[37] A. Greenberg *et al.*, "VL2: A scalable and flexible data center network," in *Proc. ACM SIGCOMM*, 2009, pp. 51–62.
[38] R. N. Mysore *et al.*, "PortLand: a scalable fault-tolerant layer 2 data center network fabric," in *Proc. ACM SIGCOMM*, 2009, pp. 39–50.
[39] H. Wu, G. Lu, D. Li, C. Guo, and Y. Zhang, "MDCube: a high performance network structure for modular data center interconnection," in *Proc. CoNext*, 2009, pp. 25–36.

Tong Ye received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 1998 and 2001, respectively, and the Ph.D. degree in electronics engineering from Shanghai Jiao Tong University, Shanghai, China, in 2005. He was with the Chinese University of Hong Kong for one and half years as a postdoctoral research fellow. Currently, he is an Associate Professor at Shanghai Jiao Tong University, where he is with the State Key Laboratory of Advanced Optical Communication Systems and Networks. His research interests include the design of optical network architectures, optical network systems and subsystems, and silicon-ring-based optical signal processing.

Tony T. Lee received his BSEE degree from National Cheng Kung University, Taiwan, and his MS and PhD degrees in electrical engineering from Polytechnic Institute of NYU. Currently, he is a Zhiyuan Chair Professor at the Electronics Engineering Department of Shanghai Jiao Tong University, and an Emeritus Professor of Information Engineering at the Chinese University of Hong Kong. From 1993 to 2013, he was a Chair Professor at the Information Engineering Department of the Chinese University of Hong Kong. From 1991 to 1993, he was a Professor of Electrical Engineering at Polytechnic Institute of NYU, Brooklyn, New York. He was with AT&T Bell Laboratories, Holmdel, NJ, from 1977 to 1983, and Bellcore, currently Telcordia Technologies, Morristown, NJ, from 1983 to 1993. He has been served as an Editor of the IEEE Transactions on Communications, and an area Editor of Journal of Communication Network. Tony is a fellow of IEEE and HKIE. He has received many awards including the 1989 Leonard G. Abraham Prize Paper Award from IEEE Communication Society, the 1999 Outstanding Paper Award from IEICE of Japan, and the 1999 National Natural Science Award from China.

Mao Ge received the B.S. degree in communication engineering from Hohai University, Nanjing, China, in 2013. Currently, he is working towards the M.S. degree at Shanghai Jiao Tong University, Shanghai, China. His research interests mainly focus on optical switching networks.

Weisheng Hu received his B.S, M.S, and Ph.D. from Tsinghua University, Beijing University of Science and Technology, and Nanjing University in 1986, 1989 and 1997 respectively. He joined Shanghai Jiao Tong University as a Postdoctorate fellow from 1997 to 1999, and as a full professor in 1999, where he was promoted to national second-level professor and distinguished professor in 2009. He was the deputy director and then director of the State Key Laboratory of Advanced Optical Communication Systems and Networks from 2002 to 2012. He was a member of the expert team of two national project task forces of CAINONET and 3TNET from 1999 to 2006. He serves on five journal editorial boards including Optics Express, Journal of Lightwave Technology and Chinese Optics Letters, and a number of conference committees, including OFC, ICC, INFOCOM, OPTICSEAST, etc. He has published over 200 peer-reviewed journal and conference papers.

His research interests include optical communication and networking, optical fiber access.