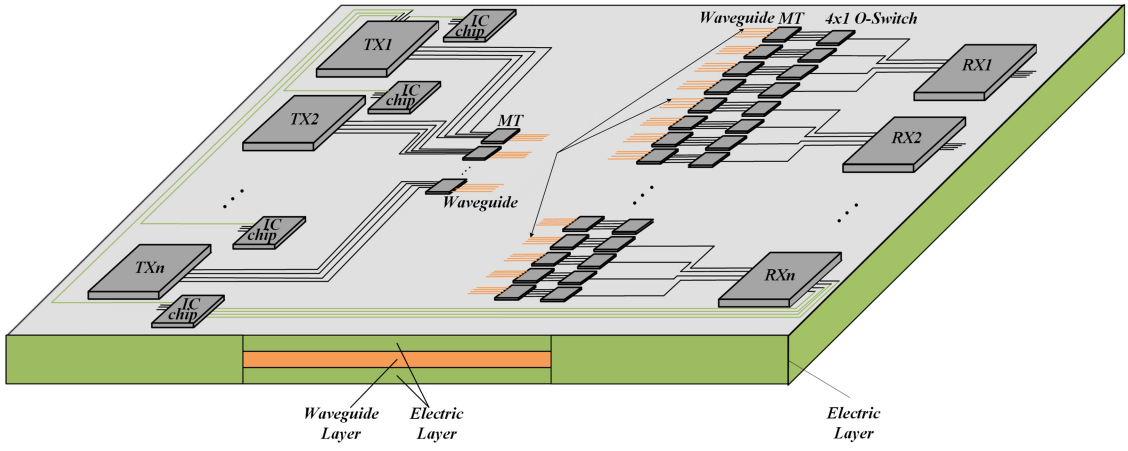
IEEE Photonics Journal
An IEEE Photonics Society Publication

Open Access

Multi-Node All-Optical Interconnect Network Routing for Data-Center Parallel Computers

Volume 11, Number 2, April 2019

Shuailong Yang
Liu Yang
Fengguang Luo
Biyu You
Yao Ni
Danhui Chen



DOI: 10.1109/JPHOT.2019.2897826
1943-0655

IEEE photonics SOCIETY

IEEE

Multi-Node All-Optical Interconnect Network Routing for Data-Center Parallel Computers

Shuailong Yang, Liu Yang, Fengguang Luo, Biyu You, Yao Ni, and Danhui Chen

National Engineering Laboratory for Next Generation Internet Access System, School of Optics and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China

DOI:10.1109/JPHOT.2019.2897826

Manuscript received December 13, 2018; revised January 30, 2019; accepted February 3, 2019. Date of publication February 6, 2019; date of current version February 20, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61471179, and in part by the China Postdoctoral Science Foundation funded Project 2018 M642843. Corresponding author: Liu Yang (e-mail: liuyang89@hust.edu.cn).

Abstract: A multi-node all-optical interconnect network-routing architecture scheme and implementation technology are designed to improve high-speed and large-capacity data transmission in data centers. In the design of the optical interconnect routing electro-optic printed circuit board (EOPCB), the signal generated by the FPGA is directly modulated by the VCSEL arrays, and through the coupler into the parallel optical waveguide. The transmission link and optical switch are operated in optical domain. The designed optical waveguide comprises 12 waveguide arrays with the space of 250 μm. The all-optical interconnection routing system, which is based on the EOPCB high-speed chip multi-node waveguide, is designed, and a 4 $\times$ 4 EOPCB is fabricated. The performance of the waveguide and the entire optical interconnection network is tested. Based on 10 Gb/s VCSEL arrays, the optical interconnect network system can reach a bandwidth of 640 GHz for 4 $\times$ 4 networks. The best bit error rate for the switching system can reach 9.0 $\times$ 10^{-12} with no transmission error.

Index Terms: Optical interconnection, waveguide and optical communications, field programmable gate array, electro-optic printed circuit board.

1. Introduction

The growth in demand and utilization of Internet service bandwidth over the years have prompted data centers to process and manage large amounts of data. To this end, data centers require high-performance computing systems to enable parallel connection to commodity servers (or compute nodes), and the simultaneous high-speed execution of tasks. On the other hand, the increasing parallelization of computing systems has the effect of augmenting the data movements (e.g., data packets) between computing nodes [1]–[3]. Therefore, interconnecting networks that link compute nodes become critical to achieving high computing performance. High performance, scalability and energy efficiency are the key requirements to inter-board and intra-board communications as well as on-chip communications (i.e., on-chip networking) for the next-generation interconnect networks [3], [4].

As data center systems demand higher bandwidth density, lower latency, and lower power consumption, optical interconnect switching is an alternative technology to meet the requirements of the high-speed and large-bandwidth parallel processing for computer systems. Therefore, optical interconnect networks can be utilized between chips, on-board and on-chip with the better performance. Different optical domains can be used for switching and multiplexing (e.g., space, time, wavelength, mode domain) and the combination to increase network scalability [5]–[7]. A control mechanism is added to ensure high network throughput, low latency and good scalability for scheduling data packets switched correctly between input and output ports in a synchronous manner [8]–[10]. Recently, many solutions have been proposed to develop and optimize optical interconnection networks from hardware implementation or software optimization. In terms of software optimization, a centralized optical resource management be proposed for evaluating multipoint-to-multipoint optical communication systems [11], [12]. The longest queue-first algorithm (ipLQF) and the iSLIP algorithm with low latency and proper computational efficiency are proposed to optimize the performance of the interconnected network [1], [13]. As for the hardware implementation, injection-enhanced Si-emitting devices using avalanche mode have been proposed to provide very large-scale integrated compatible light emitters for inter-chip or intra-chip signal transmission [14]. It is indicated that both spatial modulation of light emitting pattern and light intensity modulation can be realized by using these devices (Gate-controlled diode MOS-like and carrier injection BJT-like multi-terminal) [15]. The high cost and high-power consumption characteristics of traditional data centers introduce severe scalability limitations. A passive optical top-of-rack (ToR) interconnect architecture is proposed to replace the traditional electronic packet switch (EPS) in the data center networks (DCNs) access layer [16]–[18]. Meanwhile, an optical interconnection structure based on a fiber link is proposed to realize long-distance optical interconnection transmission [19], [20]. However, the introduction of optical fiber has caused serious dispersion damage, or the dispersion damage can be mitigated by the additional in-phase quadrature (IQ)-dual binary modulation scheme [21]–[24]. Due to the superiority of optical interconnects, a dual optical architecture has been proposed with optics on the front plane (the server connected to the rack) and the backplane (connected to the TOR switch) to further increase the transmission rate and bandwidth of the system [25]–[27]. Therefore, the cost of optical interconnects has become an important part of the cost of data center networks.

Inter-chip optical interconnects on printed circuit boards (PCB) are important techniques for increasing the data rate and bandwidth of signal transmission between chips. To address the bandwidth requirements between data center processors, we design and manufacture a chip-to-chip all-optical interconnect platform based on FPGA control, in which chips are connected to the chip through waveguides. The all-optical interconnect platform consists of four nodes, each with 8 rounds. Based on 10 Gbit/s VCSEL arrays, the bandwidth of system is 640 GHz for 4×4 network.

2. Operation Principle

2.1 Structural Design of EOPCB

A multi-node all-optical interconnect network routing structure scheme is proposed for high-performance computer applications with optical waveguide interconnect layers. The structure based on parallel optical signals and optical waveguide interconnections between EOPCB chips is designed and verified. High-speed interconnection between chips is realized through optical interconnection, and the routing function of the $n \times n$ node all-optical interconnection switching network is realized to overcome the electromagnetic interference between high-speed electrical signals. Fig. 1 is a schematic diagram of an overall structure of a $n \times n$ node all-optical interconnect switching network routing system with multicast function.

The EOPCB consists of Integrated Circuit (IC) chips, n 1×4 TX, which are Vertical Cavity Surface Emitting Laser (VCSEL) parallel optical transceiver integrated array module with 10 Gbit/s, parallel optical waveguide arrays (orange line), n 1×4 RX, which are PIN arrays, MT-compatible coupling interfaces (MT) and N ($N = 4 \times n$) 4×1 optical switches (O-Switch). The EOPCB contains four processing IC chips for generating parallel data and processing data. The VCSEL/PIN array modules

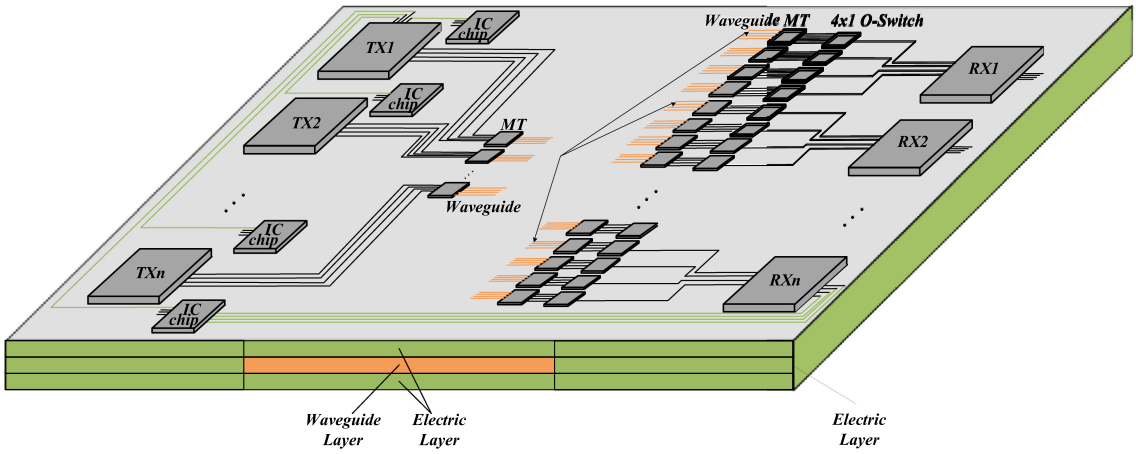


Fig. 1. $n \times n$ node EOPCB all-optical interconnect network routing structure.

are used to modulate parallel baseband signals and receive signals. The polymer waveguides are optical interconnect layer as the optical transmission links on the EOPCB board. The MT couplers are used to couple the optical path between the VCSEL/PIN and the waveguide. As shown in Fig. 1, the four input signals of TX1 are generated by four different IC chips. We reorder the parallel electrical signals of each IC chip at the input end to avoid the crosstalk of the optical fiber array and the optical waveguide array on the EOPCB board, as shown in Fig. 1. The original parallel signal is re-divided into n sets of parallel electrical signals. The first electrical signals of each chip are recombined into a set of parallel electrical signals, and the second signals are recombined into a set of parallel electrical signals, and the signals of each chip are grouped according to this rule. The re-grouped electrical signals respectively drive 1×4 VCSEL arrays. The signal is directly modulated to form N parallel optical signals. The communication capacity of the entire optical interconnection network depends on the transmission rate of the VCSEL and the degree of parallelization of the data center. The parallel computer data is connected to the IC chips to enable high-speed optical interconnect transmission in parallel computing data centers. In the design, we choose FPGA instead of IC chip. The FPGA controls the optical switch to implement chip-to-chip bidirectional parallel high-speed interconnect network communication on the EOPCB board. According to the requirements of optical interconnection, we can choose the four optical signals needed. The PIN arrays receive the optical signal and demodulate the signal. The electrical signals are received and stored by the FPGA, followed by DSP processing and error analysis. Increasing the speed of the VCSEL and the degree of parallelism in the data center can increase the bandwidth of the entire interconnect network. The interconnect structure can improve high-speed and large-capacity data transmission between data centers.

2.2 Optical Interconnect Layer PCB Board Production

The optical waveguide is fabricated by the doctor blade method, which has the advantage of obtaining the large-area polymer optical waveguide layer. The method can directly apply the polymer optical waveguide material on the FR4 board as an interlayer in the multilayer PCB board, and the process conditions are compatible with the conventional PCB process, which is convenient for low-cost production.

A multimode optical waveguide core layer and a cladding film are fabricated using a low loss rate poly-siloxane polymer material. The preparation process of polymer multimode optical waveguide layer materials be studied. First, a SU-8 glue optical waveguide structure mold frame is fabricated. The optical waveguide polymer material is injected. The core layer and the cladding optical waveguide film layer of the required thickness are respectively obtained by using a doctor blade at a

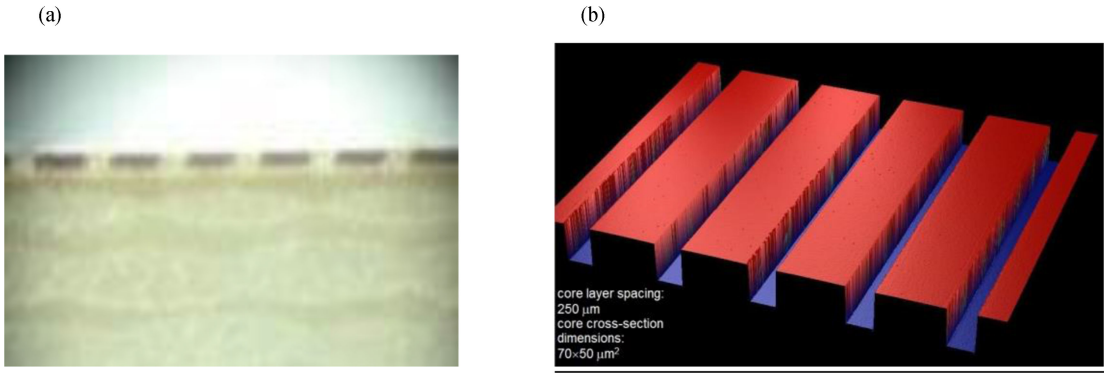


Fig. 2. (a) SU-8 glue optical waveguide mold cross-section photomicrograph and (b) three-dimensional image of SU-8 glue optical waveguide core layer mold.

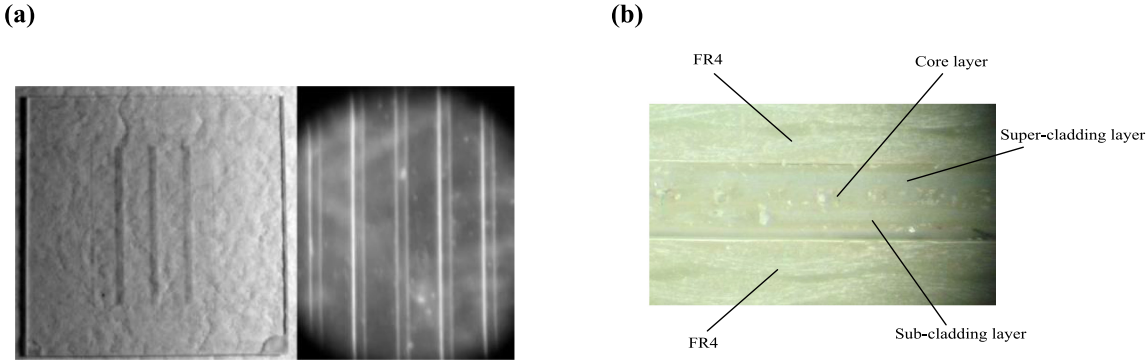


Fig. 3. Photograph of optical waveguide (a) Parallel optical waveguide magnified photograph, (b) Cross-section molding photo of optical waveguide layer in EOPCB.

certain pressure and rate. The optical waveguide core layer and cladding layer are fabricated by a doctor blade method. The optical waveguide core layer has a refractive index of 1.43. The cladding refractive index is 1.41. Optical transmission loss is 0.08 dB/cm. The number of polymer optical waveguide arrays is 1 $\times$ 12. The core layer spacing is 250 μm. The core cross-section dimensions are approximately 70 $\times$ 50 μm^2. The upper and lower cladding layers have a thickness of about 150 μm. As shown in Fig. 2, SU-8 glue optical waveguide mold cross-section photomicrograph and three-dimensional image of SU-8 glue optical waveguide core layer mold are given.

In the intermediate layer of the multilayer printed circuit board material FR4, the formation technique of the optical waveguide optical interconnect layer is studied. The process steps are: making FR4 cladding frame - injecting poly-siloxane under cladding material - laminating with core layer - heating curing - demoulding - injecting polysilicon over-cladding material - and laminating under cladding - warm curing - embedded in FR4 multilayer PCB for warm rolling - end treatment - EOPCB optical waveguide layer - compatible with conventional PCB processes. Fig. 3(a) is a photograph of three sets of 1 $\times$ 12 polymer optical waveguide array PCB boards, and Fig. 3(b) is a photograph of a cross-section of an optical waveguide layer in an EOPCB. In the figure, the EOPCB is FR4 layer, polymer waveguide upper cladding layer, optical waveguide core layer, polymer waveguide lower cladding layer and FR4 layer from top to bottom.

3. Experiment Setup and Discussions

3.1 Chip-to-Chip Optical Interconnection Network on PCB

We built an experimental demonstration system platform to demonstrate the routing and switching capabilities of the all-optical interconnect network. The system platform includes FPGA routing

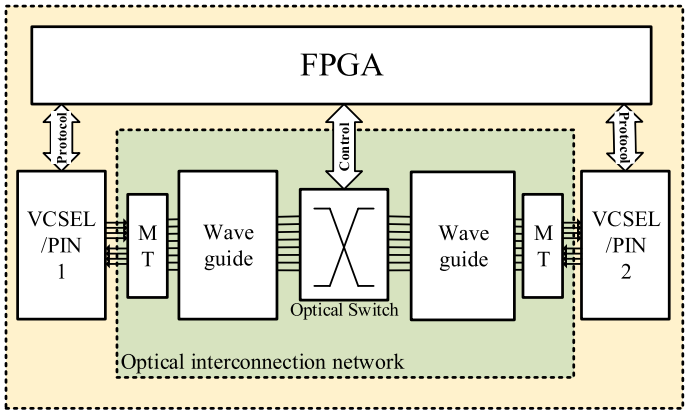


Fig. 4. all-optical interconnect routing exchange experiment demonstration system platform.

switch control board, VCSEL arrays, parallel optical waveguide board, 4 $\times$ 1 optical switch array and other optoelectronic components, as shown in Fig. 4. In the optical interconnect system, a pseudo random bit sequence (PRBS) signal generated by the FPGA is directly modulated by the VCSEL through the serial communication protocol to realize electro-optical conversion and sent to the optical interconnect network. In the optical interconnect network, MT-compatible coupling interface is used for vertical coupling of optical signals between the VCSEL and the waveguide. The communication channel consists of parallel waveguides and optical switch. The optical switch is commercially available with a switching time of 50 μs, and the interface of the optical switch couples the waveguide through the MT-compatible coupling interface. The FPGA generates control commands through the serial communication protocol, and controls the optical switch to select the transmission of the data stream. The VCSEL arrays have 8 ports, four of which are data transmitting ends, and four ports are data receiving ends. The data transmission rate of each port is 10 Gbit/s and the wavelength is 850 nm. The optical interconnect transmission network can be used for routing high-speed and large-capacity data in the data center, and is also made into EOPCB optical interconnect board, which facilitates integration of circuits and waveguides.

We performed a light exchange experiment from node 2 to node 1 to show the performance of each key device in the entire EOPCB optical switching network. The FPGA generates a PRBS of length 10^{12}, which is input into the VCSEL through a serial communication protocol, and directly modulates the parallel transmitted optical signal channel of the input node 2. The FPGA generates a control command to control the optical switch, selects the optical channel to be exchanged, and inputs the parallel optical signal receiving channel to the node 1 to realize the transmission in the optical interconnect network. First, we tested the transmission loss of the waveguide and the optical interconnect network to obtain a more comprehensive performance of the designed optical interconnect network. As shown in Fig. 5, the insertion loss of the waveguide and the optical interconnect network is better than fiber in multi-interface connections of short-range chips. The designed optical interconnect network can be used for high speed and large capacity optical interconnect transmission systems. The average insertion loss of the waveguide and the optical interconnect network is 1.29 dB and 2.24 dB.

The transmission performance of the waveguide, optical switch and the optical interconnect network is measured to obtain the transmission performance of each key component in the EOPCB optical interconnect network. As shown in Fig. 6, the BER performances of both the fabricated waveguide and the optical switch can reach no error transmission under the receiver sensitivity of -9 dBm. There is a performance difference between the four channels, which is about 3 dB for optical waveguide. The difference of the channels is due to the uniformity of the waveguide fabrication and the VCSEL/PIN. The bit error rate of waveguide communication can reach up to 9.0×10^{-12}. In the optical interconnect network, optical switches are important devices for controlling

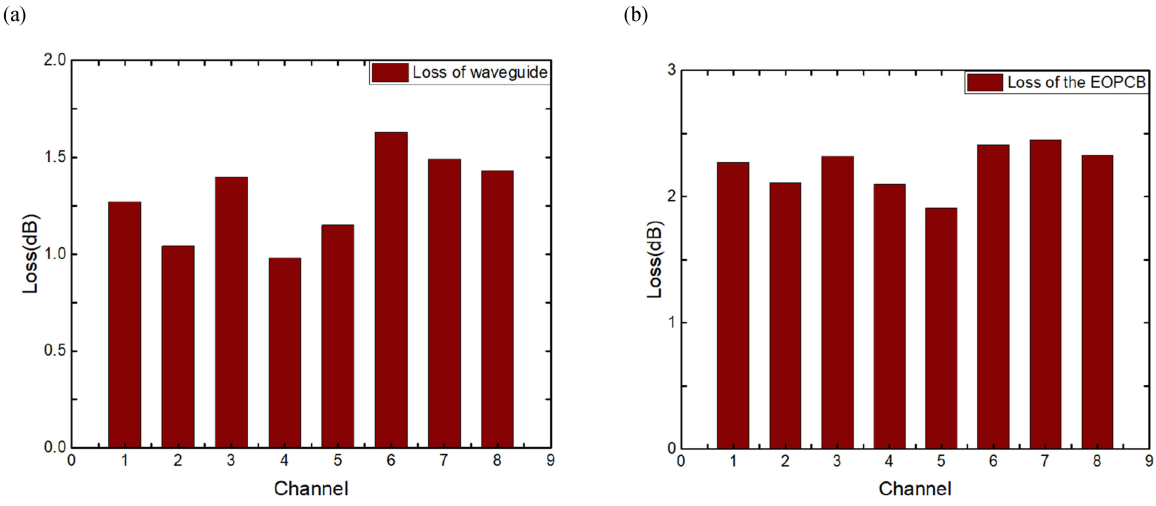


Fig. 5. (a) the transmission loss of the waveguide and (b) the transmission loss of the EOPCB.

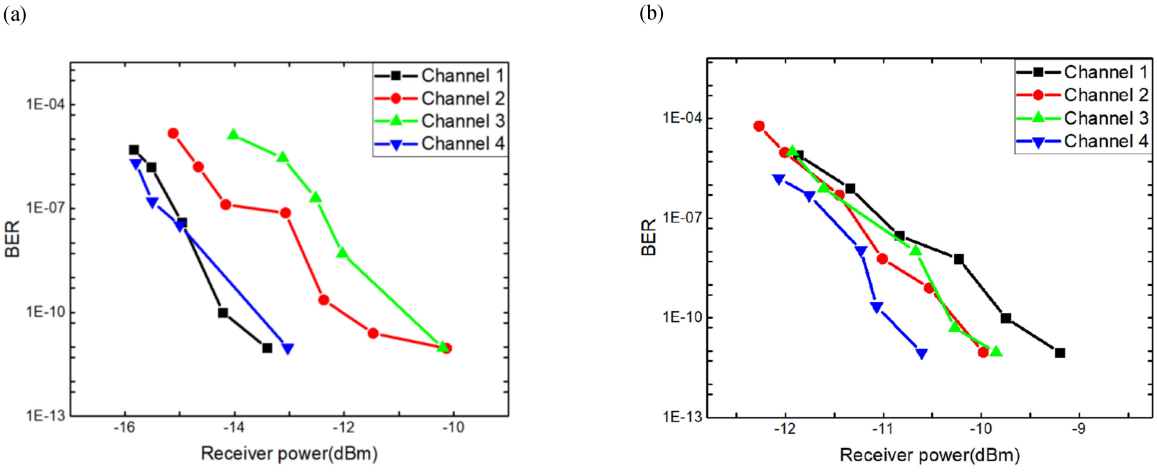


Fig. 6. (a) Waveguide transmission performance. (b) Optical switch transmission performance.

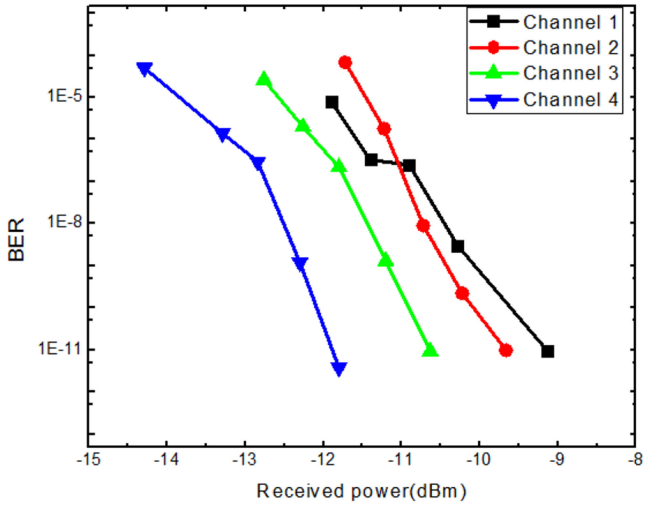


Fig. 7. Transmission performance of the optical interconnect network.

optical switching. We have experimentally evaluated the transmission performance of the optical switch as shown. When only the optical switch is tested in the optical interconnect network, the transmission performance of each channel does not differ much. The effect of the optical switch on the optical interconnect network is 1.5 dB. The error rate of system communication can reach up to 9.0×10^{-12} which is no error transmission at the received power of -9.2 dBm.

We conduct experimental tests on the transmission performance of the entire optical interconnect network, as shown in Fig. 7. There are certain performance differences in different channel switching in the optical interconnect network. The difference in performance is mainly due to the fabrication of the waveguide, but does not affect the transmission of the EOPCB optical interconnect network. Each optical switching transmission can achieve error-free transmission, and the bit error rate can reach 9.0×10^{-12} which is no error transmission at the worst received power of -9 dBm and best of -11.8 dBm, respectively.

4. Conclusion

This research proposes and designs a multi-node all-optical interconnect network routing architecture scheme for high-speed and large-capacity data transmission in data centers. In the optical interconnection network, signals generated by the FPGA are directly modulated by the VCSEL, and optical switches are controlled by FPGA to realize optical interconnection transmission. We design a bidirectional 8-channel (4 transmit/receive) optical waveguide with an optical channel spacing of 250 μm. The transmission loss of the designed waveguide is 1.29 dB and the loss of the EOPCB is 2.24 dB. The transmission performance of the waveguide can reach no error transmission of the 9.0×10^{-12} BER. The optical switch is the key component in the optical interconnect network, and the optical switch can achieve a transmission performance. We develop the EOPCB high-speed chip multi-node waveguide all-optical interconnect routing system and measure the performance of the waveguide and the optical interconnect network. Based on 10 Gbit/s VCSEL arrays, the optical interconnect network system has a bandwidth of 640 GHz for 4×4 network. The best bit error rate for channel transmission can reach 9.0×10^{-12}. In high-capacity and high-rate parallel computer data center, the designed optical interconnect system can serve as a data interconnection transmission hub, which promotes the establishment of a large number of parallel data centers.

References

[1] I. Cerutti *et al.*, "Simulation and FPGA-based implementation of iterative parallel schedulers for optical interconnection networks." *IEEE/OSA J. Opt. Commun. Netw.*, vol. 9, no. 4, pp. C76–C87, Apr. 2017.

[2] T. Barwicz *et al.*, "Optical demonstration of a compliant polymer interface between standard fibers and nanophotonic waveguides," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2015, Art. no. Th3F.5.

[3] R. Dangel *et al.*, "Polymer waveguides for electro-optical integration in data centers and high-performance computers," *Opt. Exp.*, vol. 23, pp. 4736–4750, 2015.

[4] Q. Hu, D. Che, Y. Wang, and W. Shieh, "Advanced modulation formats for high-performance short-reach optical interconnects," *Opt. Exp.*, vol. 23, no. 3, pp. 3245–3259, 2015.

[5] P. Velha, N. Andriolli, and I. Cerutti, "Generating supermodes in uniform and non-uniform arrays of SOI waveguides," in *Proc. Integr. Photon. Res., Silicon Nanophoton.*, 2016, Art. no. ITu3B.5.

[6] J.-J. Liu, *et al.*, "High bit-rate distance product of 128 Gbps/km 4-PAM transmission over 2-km OM4 fiber using an 850-nm VCSEL and a volterra nonlinear equalizer," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2017, pp. 1–3.

[7] A. L. Porta *et al.*, "Optical coupling between polymer waveguides and a silicon photonics chip in the o-band," in *Proc. Opt. Soc. Amer.*, 2016, Paper M2I.2.

[8] F. Gambini, S. Faralli, P. Pintus, N. Andriolli, and I. Cerutti, "BER evaluation of a low-crosstalk silicon integrated multimicroring network-on-chip," *Opt. Exp.*, vol. 23, no. 13, pp. 17169–17178, 2015.

[9] M. Secondini *et al.*, "Optical time-frequency packing: Principles, design, implementation, and experimental demonstration," *J. Lightwave Technol.*, vol. 33, no. 17, pp. 3558–3570, Sep. 2015.

[10] H.-L. Wang, J. Qiu, X. Yu, M. Feng, and N. Holonyak, "85 °C operation of 850 nm VCSELs deliver a 42 Gb/s error-free data transmission for 100 meter MMF link," in *Proc. Opt. Fiber Commun. Conf. Expo.*, 2018, pp. 1–3.

[11] D. P. Van, Matteo Fiorani, L. Wosinska, and J. Chen, "Adaptive open-shop scheduling for optical interconnection networks." *J. Lightwave Technol.*, vol. 35, no. 13, pp. 2503–2513, Jul. 2017.

[12] K. Chen *et al.*, "OSA: An optical switching architecture for data center networks with unprecedented flexibility," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 498–511, Apr. 2014.

[13] Z. Li, I. Shubin, and X. Zhou, "Optical interconnects: Recent advances and future challenges," *Opt. Exp.*, vol. 23, no. 3, pp. 3717–3720, 2015.
[14] K. Xu, "Monolithically integrated Si gate-controlled light-emitting device: Science and properties," *J. Opt.*, vol. 20, no. 2, 2018, Art. no. 024014.
[15] M. d. Plessis, H. Aharoni, and L. W. Snyman, "Two- and multi-terminal silicon light emitting devices in standard CMOS/BICMOS IC technology," *Physica Status Solidi (a)*, vol. 201, pp. 2225–2233, 2004.
[16] Y. Cheng, M. Fiorani, R. Lin, L. Wosinska, and J. Chen, "POTORI: A passive optical top-of-rack interconnect architecture for data centers," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 9, no. 5, pp. 401–411, May 2017.
[17] F. Yan, W. Miao, H. Dorren, and N. Calabretta, "Novel flat data center network architecture based on optical switches with fast flow control," *IEEE Photon. J.*, vol. 8, no. 2, Apr. 2016, Art. no. 0601310.
[18] V. Kamchevska *et al.*, "Experimental demonstration of multidimensional switching nodes for alloptical data center networks," *J. Lightwave Technol.*, vol. 34, no. 8, pp. 1837–1843, Apr. 2016.
[19] L. S. Ronga, S. Jayousi, E. Forestieri, M. Secondini, and F. Cavaliere, "Modulation formats analysis for optical short reach interconnects." in *Proc. 18th Italian Nat. Conf. Photon. Technol.*, Rome, Italy, Jun. 2016, pp. 1–4.
[20] M. Pantouvaki *et al.*, "50 Gb/s silicon photonics platform for short-reach optical interconnects," in *Proc. Opt. Fiber Commun. Conf.*, Los Angeles, CA, USA, Mar. 2016, Paper Th4H.4.
[21] E. Forestieri, M. Secondini, F. Fresi, G. Meloni, L. Poti, and F. Cavaliere, "Extending the reach of short-reach optical interconnects with DSP-free direct detection," *J. Lightwave Technol.*, vol. 35, no. 15, pp. 3174–3181, Aug. 2017.
[22] K. Wang, A. Nirmalathas, C. Lim, K. Alameh, and E. Skafidas, "Space-time coded high-speed reconfigurable free space card-to-card optical interconnects with extended range," in *Proc. Opt. Fiber Commun. Conf.*, Anaheim, CA, USA, 2016, Art. no. Th4D.6.
[23] K. Wang, A. Nirmalathas, C. Lim, K. Alameth, and E. Skafidas, "Space-time coded high-speed reconfigurable free-space card-to-card optical interconnects with extended range," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 9, no. 2, pp. A189–A197, Feb. 2016.
[24] A. Kushwaha, T. Das, and A. Gumaste, "Does it make sense to put optics in both the front and backplane of a large data-center," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2017, pp. 1–3.
[25] J.-J. Liu *et al.*, "High bit-rate distance product of 128 Gbps/km 4-PAM transmission over 2-km OM4 fiber Using an 850-nm VCSEL and a volterra nonlinear equalizer, " in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2017, pp. 1–3.
[26] A. L. Porta *et al.*, "Optical coupling between polymer waveguides and a silicon photonics chip in the o-band," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2016, pp. 1–3.