Error-Free Operation for Fully Connected Wavelength-Routing Interconnect among 8 FPGAs with 2.8-Tbit/s Total Bandwidth

Takanori Shimizu[(1)], Shigeru Nakamura[(1)], Hiroshi Yamaguchi[(1)], Koichi Takemura[(1)], Kenji Mizutani[(1)], Tatsuya Usuki[(1)], and Yutaka Urino[(1)]

[(1)] Photonics Electronics Technology Research Association (PETRA), Japan

t-shimizu@petra-jp.org

Abstract *We developed high bandwidth-density embedded optical modules (EOMs) for C-band dense wavelength-division-multiplexing (DWDM) interconnect. We constructed fully connected wavelength routing network among 8 FPGA nodes with 4 EOMs per node and confirmed error-free operation for total bandwidth of 2.8 Tbit/s.*

Introduction

Since the growth of computing power in a CPU has slowed down, heterogeneous computing with accelerators and parallel computing across many nodes have become main targets in data centers. Performance of these computing systems tightly depends on their inter-node interconnect characteristics. Fully connected networks offer the advantages of low latency, low power consumption and wide bandwidth[1],[2]. To realize the fully connected network topology with high-density fiber optic wires, parallel link network among field-programmable gate arrays (FPGAs) with fiber sheet and Optical I/O Core was demonstrated[2],[3].

Wavelength routing (WR) has been proposed for the fully connected networks with the fewer fiber cables[4]. Comparing with the parallel link network, WR using dense wavelength division multiplexing (DWDM) has large scalable merits in the point of bandwidth density on the front panel of the router board and energy efficiency of the optical amplifier by which all WDM channels in a fiber can be amplified simultaneously[1].

In this paper, we present high bandwidth-density embedded optical modules (EOMs), which are available for DWDM with single-mode fibers (SMFs) in C-band, and demonstration in WR and fully-connected interconnect among eight FPGA cards using the EOMs.

Design of system and its components for WR

Figure 1 shows configuration of the fully connected WR interconnect among eight FPGA cards. WDM light sources are placed out of the EOMs to lock each wavelength at the grid precisely and to be managed centralizedly. This scheme also contributes high energy efficiency and high reliability. An optical amplifier is essential to realize error-free WR interconnect, especially large-scale one. Since an erbium-doped fiber amplifier (EDFA) in C-band is extremely superior to the others in terms of energy efficiency and low noise, we positively use the EDFAs. The wavelength router consists of multiple arrayed waveguide gratings (AWGs) to suppress coherent crosstalk among same wavelength signals input from different ports.

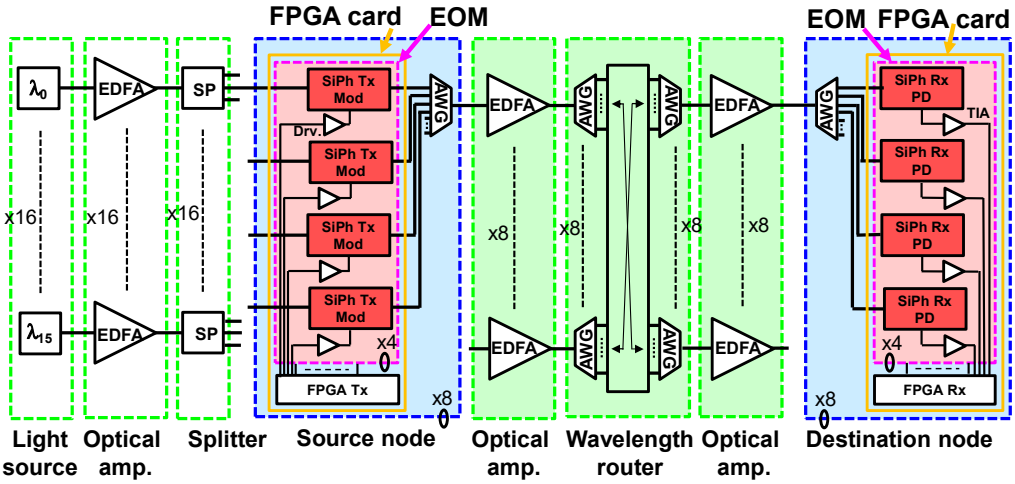

Fig. 1: Configuration of fully connected wavelength-routing interconnect among 8 FPGA cards

978-1-6654-3868-1/21/$31.00 ©2021 IEEE

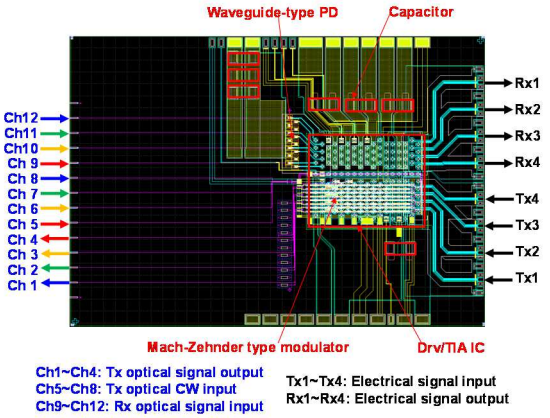

Fig. 2: Layout of 25 Gbit/s × 4 channel transceiver chip

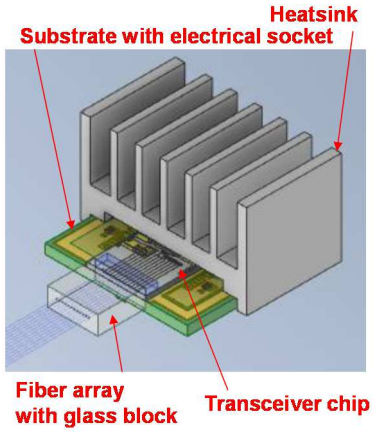

Fig. 3: Overview of embedded optical module

Figure 2 shows layout of the 100-Gbit/s bidirectional (25-Gbit/s × 4-channel) transceiver chip. The chip fabricated by silicon photonics technology[5] includes silicon optical waveguide with core size of 400 nm wide and 200 nm high, four Mach-Zehnder-type modulators for transmitters (Tx), and four waveguide-type photodiodes (PD) for receivers (Rx). An IC chip including modulator drivers and trans-impedance amplifiers is mounted by flip-chip bonding on the transceiver chip. The chip size is 5 mm × 7.1 mm.

Since optical modulators for DWDM should be wavelength independent to reduce inventory, we do not take polarization diversity by grating couplers, but take butt-coupling with polarization maintaining fibers to feed CW light from the off-module light sources because the optical modulator can operate only for TE mode. On the other hand, PD should be also polarization independent to couple with ordinary SMF. Since a jitter due to polarization mode dispersion (PMD) along the silicon optical waveguide is serious for signal deterioration at high bit rate such as 25 Gbit/s or higher, we take waveguide-type PDs and introduced 1-μm-wide waveguides between the SMF and PD to reduce the PMD. Inverse tapered spot-size converters with 240-nm tip width for efficient butt-coupling to SMF are adopted at silicon waveguide endface.

Figure 3 shows an overview of the EOM, which is socketable to an FPGA card. The EOM size is 12 mm × 12 mm × 9 mm without fiber array, resulting high bandwidth density of 69.4 Gbit/s/cm^2.

Demonstration of fully connected WR interconnect among 8 FPGAs

Firstly, optical insertion losses in the Tx and Rx of the EOM were evaluated. The Tx loss was about 13.0 dB, including 2-facet coupling loss and modulator loss. The Rx loss was about 3.5 dB, including a facet coupling loss and waveguide loss estimated from power conversion efficiency of 1 mA/mW at waveguide-type PD. These measured losses in both Tx and Rx were low enough for our optical power budget of WR link design. The bit-error-rate (BER) curves of the Rx on the transceiver chip was also measured with psuedo-randam bit sequence (PRBS) 2^{31}-1 at 25 Gbit/s, as shown in Fig. 4. The penalty by polarization dependence was about 1 dB. Since the BER curve of 45-degree linear polarization is placed between those of TE and TM, we confirmed that the jitter by PMD was well suppressed.

Secondly, experiment of fully connected and

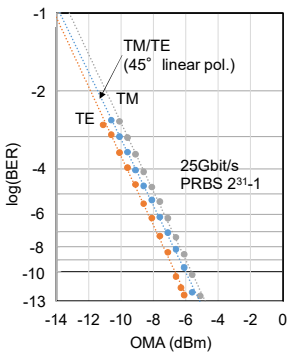

Fig. 4: Bit-error-rate curve of Rx on transceiver chip including 1-μm-wide waveguide

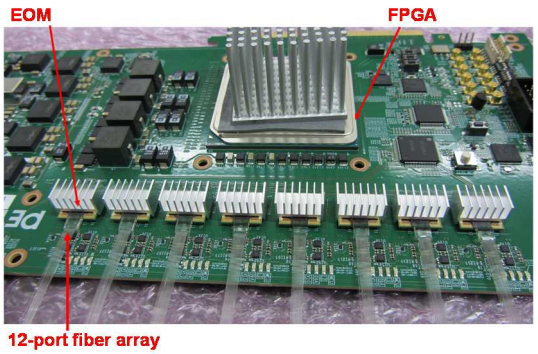

Fig. 5: Photograph of 8-EOMs on FPGA mounted card

Tab. 1: Error free operation range at a FPGA card

WDM channel#	Optical frequency (THz)	Tx		Rx		Input power to destination node (dBm)									
		EOM#	channel#	EOM#	channel#	-5	-4	-3	-2	-1	0	1	2	3	4
1	192.0	1	Tx1	1	Rx2			■	■	■	■	■	■	■	■
2	192.2		Tx2		Rx3						■	■	■	■	■
3	192.4		Tx3	4	Rx4			■	■	■	■	■	■	■	■
4	192.6		Tx4		Rx1			■	■	■	■	■	■	■	■
5	192.8	2	Tx1		Rx2			■	■	■	■	■	■	■	■
6	193.0		Tx2	3	Rx3				■	■	■	■	■	■	■
7	193.2		Tx3		Rx4		■	■	■	■	■	■	■	■	■
8	193.4		Tx4		Rx1				■	■	■	■	■	■	■
9	193.6	3	Tx1		Rx2			■	■	■	■	■	■	■	■
10	193.8		Tx2	2	Rx3				■	■	■	■	■	■	■
11	194.0		Tx3		Rx4							■	■	■	■
12	194.2		Tx4		Rx1							■	■	■	■
13	194.4	4	Tx1		Rx2			■	■	■	■	■	■	■	■
14	194.6		Tx2	1	Rx3					■	■	■	■	■	■
15	194.8		Tx3		Rx4						■	■	■	■	■
16	195.0		Tx4		Rx1							■	■	■	■

WR interconnect among four EOMs on a single FPGA card was performed. Figure 5 shows a photograph of eight EOMs on an FPGA mounted card. The size of the FPGA card is compliant of standard PCIe form factors. The experimental setup was shown as Fig.1 except the number of FPGA cards. The number of wavelength channels was 16 with 200-GHz spacing and each signal to transmit was PRBS 2^{31}-1 at 25 Gbit/s. BER characteristics were measured by a transceiver tool kit installed in the FPGA. Table 1 shows error-free conditions under various input power to the destination node (see Fig.1), where the rows correspond to the WDM channels. The blue-filled cells in Tab.1 show that the BERs were less than 10^{-12}. We confirmed error-free operation of all 16 channels simultaneously at 1-dBm input power. Since the output power from the final EDFA is designed to be 6 dBm per channel, the network has an enough margin.

Finally, we demonstrated WR and fully-connected interconnect among eight FPGA cards. The experimental setup is shown as Fig.1 and photograph of the 8-node rack-server system that eight FPGA cards were mounted is shown in Fig 6(a). Thanks to the reduction in the number of optical connectors by DWDM, the wavelength-router fits in the rack with 1U-height. Figure 6(b) shows BERs of all 112 channels for fully connected 8 FPGA cards with 25 Gbit/s bidirectional per channel, 2 channels per link and 7 links per node, namely 350 Gbit/s per node. The open circles indicate that the BERs were less than 10^{-14}. We confirmed error free operation with BER<10^{-12} over all 112 channels. Therefore, we demonstrated fully connected WR interconnect among 8 FPGA cards with 2.8-Tbit/s total bandwidth. In this experiment, 16 channels out of the total 128 channels were not used, because they loopback themselves. These 16 channels are reserved for hierarchical linking in a larger system[1],[4]. In addition, thanks to the design consideration against polarization dependence of the Tx and Rx, we successfully achieved the error-free operation without any polarization controllers.

Since only half of the 8 EOMs mounted on each FPGA card were used in the experiment, we expect to double the network bandwidth by using the all EOMs in the near future.

Conclusion

We developed high bandwidth-density EOMs, which are available for DWDM with SMFs in C-band. We demonstrated error-free operation in WR and fully connected interconnect among eight FPGA cards using the EOMs with 2.8-Tbit/s total bandwidth.

Acknowledgements

The authors thank Junichi Fujikata, Jun Ushida, and Takahiro Nakamura for helpful discussions. This paper is based on results obtained from a project (JPNP13004) commissioned by the New Energy and Industrial Technology Development Organization (NEDO). Part of this work was conducted at the TIA-SCR by AIST.

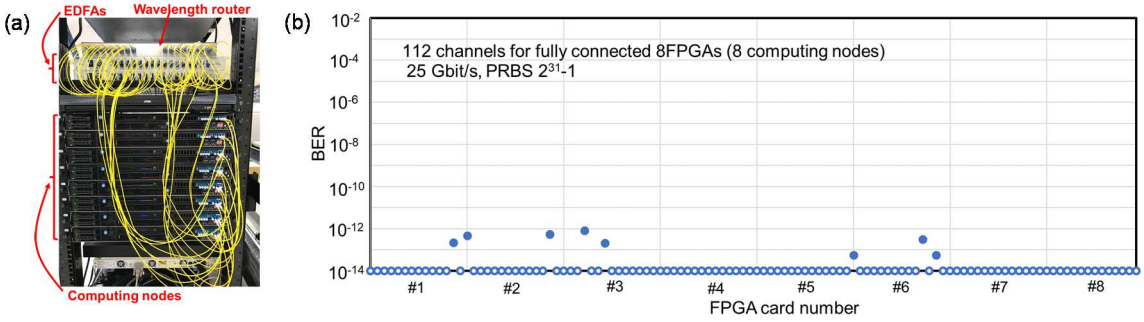

Fig. 6: Demonstration of wavelength-routing and fully connected interconnect among 8 FPGA cards: (a) Photograph of the 8-node rack-server system, (b) Bit error rates of all 112 channels

References

[1] Y. Urino *et al.*, “Wavelength-routing interconnect “Optical Hub” for parallel computing systems”, *in Proc. Int. Conf. on High Performance Computing in Asia-Pacific region (HPC Asia 2020)*, Fukuoka, Japan, Jan. 2020, pp. 81–91.

[2] K. Mizutani *et al.*, “OPTWEB: A Lightweight Fully Connected Inter-FPGA Network for Efficient Collectives”, *IEEE Trans. Computers*, vol. 70, no. 6, pp. 849–862, June 2021.

[3] T. Nakamura *et al.*, “Fingertip-Size Optical Module, “Optical I/O Core”, and Its Application in FPGA”, *IEICE Trans. Electron.*, vol. E102-C, no. 4, pp. 333–339, Apr. 2019.

[4] X. Xiao *et al.*, “Multi-FSR Silicon Photonic Flex-LIONS Module Bandwidth-Reconfigurable All-to-All Optical Interconnects”, *J. Lightwave Technol.*, vol. 38, no. 12, pp. 3200–3208, June 2020.

[5] T. Mogami *et al.*, “1.2 Tbps/cm^2 Enabling Silicon Photonics IC Technology Based on 40-nm Generation Platform”, *J. Lightwave Technol.*, vol. 36, no. 20, pp. 4701–4712, Oct. 2018.