An AWGR-Based High Performance Optical Interconnect Architecture for Exascale Systems

Shi Xu[1], Lei Zhang[2(✉)], and Zhiling Li[3]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2] School of Computer, National University of Defense Technology, Changsha, China
leizhang@nudt.edu.cn
[3] Cadre Institute for Nationnalities of Urumqi, Urumqi, China

Abstract. The next milestone objective of HPC is exascale computing, which includes millions of nodes in the system. One of the key critical barrier toward realizing exascale computing is the fundamental challenge of communication networks. We propose a high performance optical interconnect architecture based on Arrayed waveguide grating router (AWGR) with WDM wavelength routing, the inherent parallelism in AWGRs and multi-hop switching provide high scalability of the network. Theoretical analysis and simulation show its better performance compared with fat-tree architecture.

Keywords: Exascale computing · Optical interconnect · AWGR · Wavelength routing · Performance evaluation

1 Introduction

The most recent world leading supercomputer, the Tianhe-2, realized 50 PetaFlops/s (PF) peak performance in June 2013. The next milestone objective of supercomputer is exascale computing, while just an scale expanding of current architecture cannot reach exascale computing, because it encounters several complex problems that range from the management of hardware failures at runtime, memory wall constraint, power consumption, identifying the adequate massively parallel programming approaches, and etc. One of the key critical barrier toward realizing exascale computing is the fundamental challenge of communication networks. It's very difficult to meet the high bandwidth, high scalability and low-latency communication requirements using conventional interconnects techniques. Furthermore, as the interconnections between supercomputer racks are usually optical fibers, it is very difficult to deploy the network with the extraordinary growth of fiber interconnections in exascale computing.

Regarding network topologies of current high performance computer systems, Fat Tree is one of the most common architecture to build large-scale computing systems and it is usually built with some level of oversubscription to reduce the number of inter-rack switches and cables. For a 3-tier Fat Tree network with 64 port electrical

W. Xu et al. (Eds.): NCCET 2016, CCIS 666, pp. 157–167, 2016.
DOI: 10.1007/978-981-10-3159-5_15

switches, the scalability of Fat-Tree is well below 100,000 nodes. A recent trend in large-scale interconnection architectures is use high radix switches with a large number of ports to create directly connected networks such as dragonfly from Cray.

Optical switches attract much attention due to its low power and high radix features. The Arrayed waveguide grating router (AWGR) based optical switches and optical routers with packet switching capability have been investigated for a number of years [1, 2]. The AWGR allows the signal from one input to reach one output only on a particular wavelength, it is unique in that all-to-all communication can be realized if every node is equipped with multiple receivers and multiple transmitters working on different wavelengths so that signals on different wavelengths can be transmitted and received concurrently.

The AWGR is a passive device and low-loss at all data rates, but it encounters scalability problem when using AWGR to construct large scale interconnection network for supercomputers. One is the transceiver scalability, a N×N ports AWGR need N2 transceivers, when the number of nodes are very large, we cannot equip so many transceivers for each node. The other is interconnection scalability, because of the interior characteristics of AWGR, they cannot be directly connected as traditional routers to extend the network scale, although the number of ports for a single AWGR has reached 1024 now, the port number cannot increase unlimitedly due to the crosstalk problem, it's far more to meet the exascale computing scalability requirements.

To solve these problems, we propose a nested 2D-Tree topology for exascale computer network, which exploit the unique wavelength routing capability of AWGR to implement all-to-all interconnection in the 2D dimension, and utilize nested structure for hierarchical all-to-all interconnections, the inherent parallelism in AWGRs and multi-hop switching provide high scalability of the network.

2 Related Work

Optical links in most recent supercomputers are primarily in the form of active optical cables [4]. The main drawback is that power hungry electrical-to-optical (E/O) and optical-to-electrical (O/E) transceivers are required since the switching is performed using electronic packet switches. Recent breakthroughs in silicon photonics offer the possibility of integrating optical devices with traditional electronic logic devices in a single chip. In the past decades, passive and active silicon photonics devices capable of all of the operations required for transmission and switching have been demonstrated, such as wave guide, micro-ring resonator and modulator, arrayed waveguide grating router (AWGR), filters, detectors, and etc.

The AWGR based optical switches and optical routers with packet switching capability have been investigated for a number of years [8–10]. The AWGR allows the signal from one input to reach one output only on a particular wavelength, it can realize all-to-all communication when each node is equipped with multiple receivers and multiple transmitters. Signals from different transceivers can be transmitted and received concurrently. For example, if we have a set of wavelengths λ_i, $0 \leq i < k$, we can use $\lambda_{\text{mod}}(1 - i - j, k)$ to deliver signals from input i to output j. Figure 1 shows a

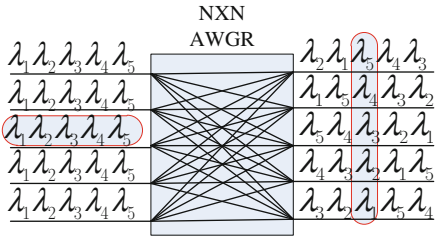

Fig. 1. Arrayed waveguide grating router example

wavelength routing map of a 5 × 5 AWGR with 5 input ports, 5 output ports, and 5 wavelengths per port.

References [2, 11, 12] demonstrate the some silicon photonics WDM solutions. Recently a silicon photonic LIONS Switch with 32 transmitters and 32 receivers utilizing 8 × 8 AWGR with kt = kr = 4 demonstrated [2] on a compact 1.2 mm × 2.4 mm silicon-on-insulator (SOI) platform.

Z. Cao [13] proposes a scalable AWGR topology for High Performance HPC Architecture, it utilizes three level interconnection to improve the network scalability. On the first level all the nodes in a chassis are connected by a single AWGR, on the second level all chassis in a cabinet are directed connected by all-to-all interconnection between AWGRs, on the third level, all cabinets are connected by two orthogonal AWGR arrays. To reduce the number transceivers, each node only has two inter-cabinet TRXs, they are multiplexed/de-multiplexed to connect to the orthogonal AWGR arrays, which need each node has a different wavelength in the same cabinet, this is very complicated for implementation and maintenance.

3 System Architecture

The whole interconnection architecture composes of five levels: CPU, node, frame, cabinet group and system.

A computing node consists of m CPUs and a switch, a switch has m electrical or optical ports that connected to the CPUs, there are also n pair of DWDM optical ports for the switch, denoted as 0, 1, 2, …, n–1, each pair DWDM optical port include a horizontal interconnection port h a vertical interconnection port v, we denote the number of wavelength in each pair DWDM optical ports is h_0, v_0, h_1, v_1, …, h_{n-1}, v_{n-1}, as depicted in Fig. 2. The switch DWDM ports are designed based on micro ring optical photonics technology, so that the power consumption are much lower than traditional serdes based electric switches.

A frame includes h_0 * v_0 computing nodes, as depicted in Fig. 3. In the horizontal direction h_0 nodes in each row is physically connected by a h_0 port AWGR, meanwhile in the vertical direction v_0 nodes in each column is physically connected by a v_0 port AWGR to form a 2D-tree connection, logically all the nodes in each row and each column are connected in an all-to-all connection.

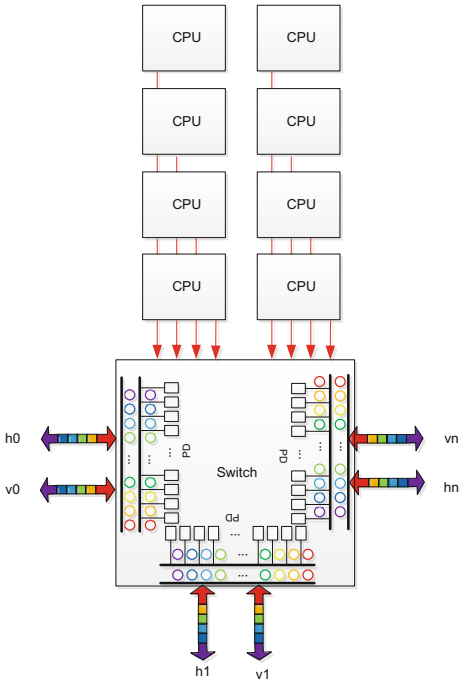

Fig. 2. Node interconnection

The number of h_1 port in the horizontal direction for each frame is $h_0 * v_0$, so does the number of v_1 port, all these ports are multiplexed/de-multiplexed into ports which include x_1 and y_1 wavelengths, so the number of x_1 port and y_1 port are $h_0 * v_0 * h_1/x_1$ and $h_0 * v_0 * v_1/y_1$ respectively, we can arrange the frames as x_1 columns and y_1 rows, so that x_1 frames in each row and y_1 frames in each column can be connected by $h_0 * v_0 * h_1/x_1$ and $h_0 * v_0 * v_1/y_1$ AWGR arrays, which construct the second level 2D-tree structure, as depicted in Fig. 4.

We can see the $x_1 * y_1$ frames in second level interconnection as a frame group, then there are $h_0 * v_0 * x_1$ h_2 type ports in horizontal direction and $h_0 * v_0 * y_1$ v_2 type ports in vertical direction, we can further multiplexed/de-multiplexed these ports to x_2 and y_2 type ports, which include x_2 and y_2 wavelengths in each port, then the number of x_2 and y_2 type ports are $h_0 * v_0 * x_1 * h_2/x_2$ and $h_0 * v_0 * y_1 * v_2/y_2$. Arrange the frame group in a 2D dimension with x_2 columns and y_2 rows, the frame groups in each row and each column are connected by $h_0 * v_0 * x_1 * h_2/x_2$ and $h_0 * v_0 * y_1 * v_2/y_2$ AWGR arrays, which is the third level 2D-tree connection, as depicted in Fig. 5.

The higher level connections can be deduced by analogy to complete the whole system interconnections. The AWGR physical connection on each level is like a two dimension tree, so we named the topology as nested 2D-tree.

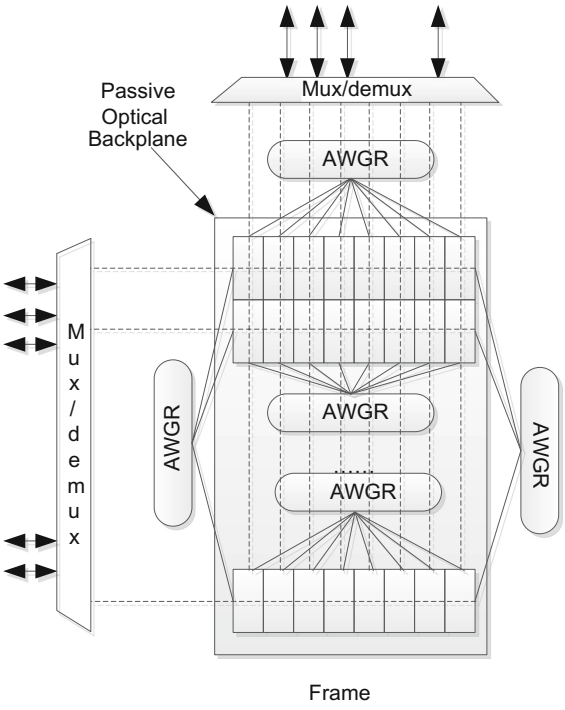

Fig. 3. Frame interconnection

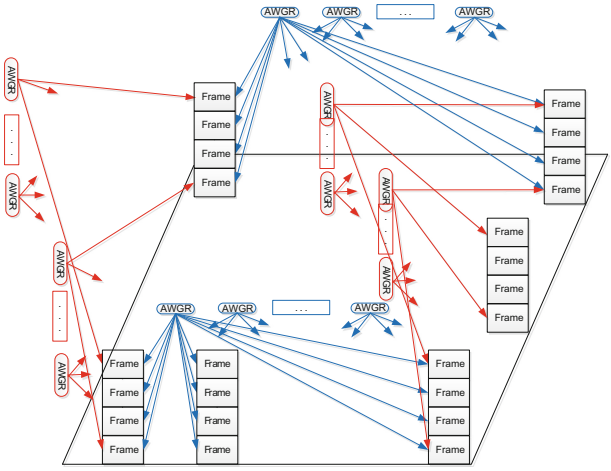

Fig. 4. Cabinet group interconnection

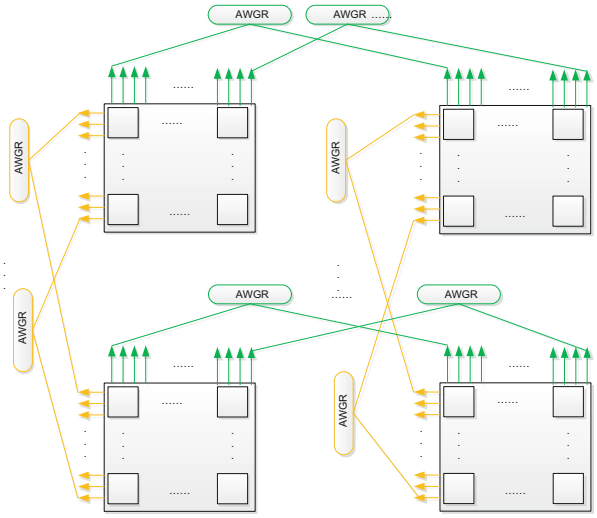

Fig. 5. System interconnection

4 Performance Evaluation

4.1 Scalability

There will be millions of nodes to be connected in future exascale supercomputers, therefore the network scalability is an important feature. Although in laboratory a 1024 port AWGR has been demonstrated, but the number of ports for commercially available AWGRs is only about 32 due to the crosstalk and insertion loss problem; the switch logic should be implemented for each wavelength, even with the emerging micro-ring and optical switch techniques, the number of ports for the switch is limited by the power consumption and die area, not more than 100 in near future. Therefore our scalability analysis is based on these two limitations.

In our proposed architecture, the total number to nodes N in the system is:

$$N = h_0 * v_0 * x_1 * y_1 * \ldots * x_{n-1} * y_{n-1}$$

The number of ports M for the switch is:

$$M = m + h_0 + v_0 + h_2 + v_1 + \ldots + h_{n-1} + v_{n-1}$$

Table 1 gives some of the feasible parameters for exascale supercomputer interconnections, which shows that the propose architecture is easily to scale to millions of nodes.

The number of wavelength is also a key factor affecting system scalability. In our proposed architecture, the wavelength can be reused on each level connection. So, the number of wavelength of the system is W:

Table 1. System scalability

Level	m	h0	v0	h1/x1	v1/y1	h2/x2	v2/y2	M	N
2	8	8	8	8/32	8/32	0/0	0/0	88	65536
2	8	16	16	16/32	16/32	0/0	0/0	72	262144
2	8	16	16	32/64	32/64	0/0	0/0	104	1048576
3	8	8	8	4/8	4/8	4/8	4/8	40	262144
3	8	8	8	16/16	16/16	8/8	8/8	72	1048576
3	8	8	8	16/16	16/16	16/16	16/16	88	4194304

$$W = max(max(h_0, v_0), max(x_1, y_1), max(x_2, y_2), \ldots, max(x_{n-1}, y_{n-1}))$$

For example, to build a system with 262144 nodes, if we use two level 2D-tree ($h_0 = v_0 = 16$, $x_1 = y_1 = 32$), 32 wavelengths are needed; if we use three level 2D-tree ($h_0 = v_0 = 8$, $x_1 = y_1 = 8$, $x_2 = y_2 = 8$), only 8 wavelengths are needed. The more levels of the 2D-tree, the fewer wavelengths are needed, which means we can use fewer ports AWGRs to build the same size system.

4.2 Inter Frame Links and Switches

Inter frame links is usually connected by optical fibers, which has important impact on system deployment and cost. In our proposed architecture, the total number of inter cabinet links are:

$$L = h_0 * v_0 * h_1 * y_1 + h_0 * v_0 * v_1 * x_1 + h_0 * v_0 * x_1 * y_1 * h_2 * y_2 + h_0 * v_0 * x_1 * y_1 * v_2 * x_2 + \ldots + h_0 * v_0 * x_1 * y_1 * \ldots * x_{n-2} * y_{n-2} * h_{n-1} * y_{n-1} + h_0 * v_0 * x_1 * y_1 * \ldots * x_{n-2} * y_{n-2} * v_{n-1} * x_{n-1}$$

The number of inter frame switches is:

$$S = h_0 * v_0 * h_1/x_1 * y_1 + h_0 * v_0 * v_1/y_1 * x_1 + h_0 * v_0 * x_1 * y_1 * y_2 * h_2/x_2 + h_0 * v_0 * x_1 * y_1 * x_2 * v_2/y_2 + \ldots + h_0 * v_0 * x_1 * y_1 * \ldots * x_{n-2} * y_{n-2} * y_{n-1} * h_{n-1}/x_{n-1} + h_0 * v_0 * x_1 * y_1 * \ldots * x_{n-2} * y_{n-2} * x_{n-1} * v_{n-1}/y_{n-1}$$

Figures 6 and 7 compare the inter frame links and switches of our proposed architecture and fat-tree, which show that our proposed architecture needs only about 50% cables and 35% switches of fat-tree in a 100000 nodes system. Even compared to the far-tree network with 3:1 oversubscription ratio, our architecture can save up to 25% cables and 50% switches.

4.3 Maximum Hop Counts and Average Hop Counts

We can treat switching among the switch ports as one hop, so the maximum hop count to be 5 for 2 level 2D-tree and 7 for 3 level 2D-tree in our proposed architecture. Two hops to destination cabinet group(for 3 level 2D-tree), two hops to the destination

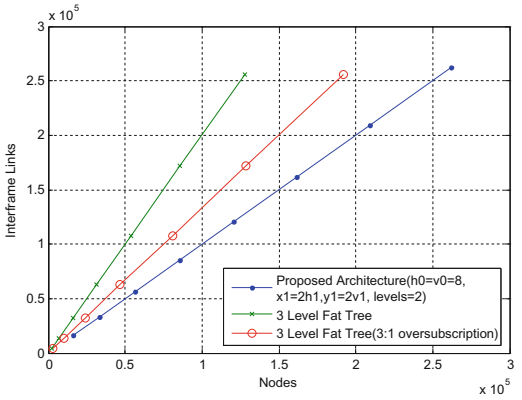

Fig. 6. The number of inter frame links

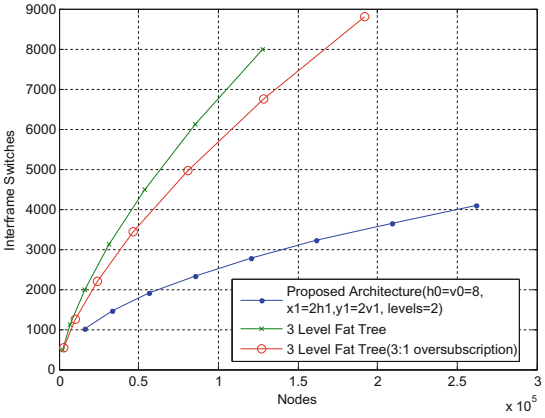

Fig. 7. The number of inter frame switches

frame and three hops to the destination CPU in the frame. For 3 tree fat tree, the maximum hop count is 5, so their maximum hop counts are same.

We simulate the average hop counts under uniform random traffic, the average hop count is of our proposed architecture and 3 tree fat tree is depicted in Fig. 8, which shows that the average hop counts are lower than fat tree architecture.

4.4 Power Consumption

Fat tree is an indirect network, the total power consumption includes network interface in each node and switch power consumption in the network. Our proposed architecture is a direct network, the AWGR power consumption is zero, the network power mainly consumed by the optical switch in each node. To compare the total network power consumption, we can use the number of network ports to estimate the total network

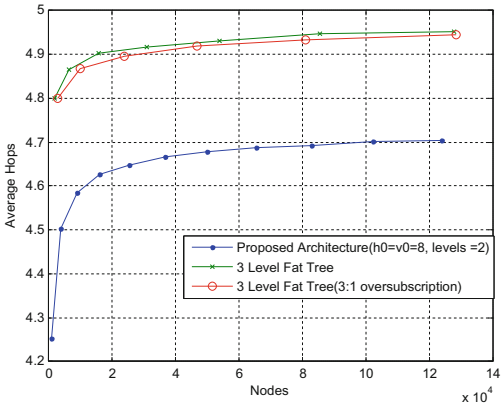

Fig. 8. The average hop counts

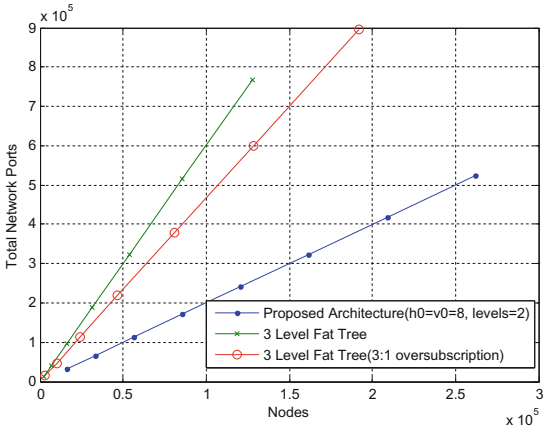

Fig. 9. The average hop counts

power consumption, as depicted in Fig. 9. We can see the power consumption of our proposed architecture is much lower than fat tree. For example, for a 100000 nodes system, the power consumption of our proposed architecture is about 60% and 40% lower than a fat tree and 3:1 oversubscription fat tree architecture.

5 Conclusion

Scalability and power consumption are the main challenges for future exascale computing network design. Among all the proposed and existing optical interconnect architectures for HPC and datacenters, the arrayed waveguide grating router (AWGR) based solutions have attracted much attention due to WDM parallelism, dense interconnectivity and unique wavelength routing capability. We propose a high

performance optical interconnect architecture based on AWGR with WDM wavelength routing for exascale systems. By exploiting the unique all-to-all wavelength routing property of AWGRs, the system can scale to millions of node. Compared with fat-tree network with 3:1 oversubscription ratio, 60% power consumption can be reserved with less switches and cables for inter-cabinet communication.

Acknowledgements. The authors would like to thank the anonymous reviewers for the feedback and revision suggestions. Then, we would thank China 863 Program (2015AA015302) and NSFC (61572509) for providing the assistance to make this research possible.

References

1. Kamei, S., Ishii, M., Itoh, M., Shibata, T., Inoue, Y., Kitagawa, T.: 64 × 64-channel uniform-loss and cyclic-frequency arrayed-waveguide grating router module. Electron. Lett. **39**, 83–84 (2003)
2. Yu, R., Cheung, S., Li, Y., Okamoto, K., Proietti, R., Yin, Y., et al.: A scalable silicon photonic chip-scale optical switch for high performance computing systems. Opt. Express **21**, 32655–32667 (2013). 2013/12/30
3. Chia, M.C., Hunter, D.K., Andonovic, I., Ball, P., Wright, I., Ferguson, S.P., et al.: Packet loss and delay performance of feedback and feed-forward arrayed-waveguide gratings-based optical packet switches with WDM inputs-outputs. J. Lightwave Technol. **19**, 1241–1254 (2001). 2013/12/30
4. Al-Fares, M., Loukissas, A., Vahdat, A.: A scalable, commodity data center network architecture. In: Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication, August 2008
5. Rumley, S., Nikolova, D., Hendry, R., Li, Q.: Silicon photonics for exascale systems. J. Lightwave Technol. **33**(3), 547–562 (2015)
6. Binkert, N., Davis, A., Jouppi, N.P., McLaren, M., Muralimanohar, N., Schreiber, R., et al.: The role of optics in future high radix switch design. In: 2011 38th Annual International Symposium on Computer Architecture (ISCA), pp. 437–447 (2011)
7. Beausoleil, R.G.: Large-scale integrated photonics for high-performance interconnects. ACM J. Emerg. Technol. Comput. Syst. (JETC) **7**, 6 (2011)
8. Yoo, S.J.B.: Optical packet and burst switching technologies for the future photonic Internet. J. Lightwave Technol. **24**, 4468–4492 (2006)
9. O'Mahony, M.J., Simeonidou, D., Hunter, D.K., Tzanakaki, A.: The application of optical packet switching in future communication networks. IEEE Commun. Mag. **39**, 1280135 (2001)
10. Guillemot, C., Renaud, M., Gambini, P., Janz, C., Andonovic, I., Bauknecht, R., et al.: Transparent optical packet switching: the European ACTS KEOPS project approach. J. Lightwave Technol. **16**, 2117–2134 (1998)
11. Absil, P.P., De Heyn, P., Dumon, P., Van Thourhout, D., Verheyen, P., Selvaraja, S., et al.: Advances in silicon photonics WDM devices, pp. 90100J-1–90100J-7 (2014)
12. Fang, Q., Liow, T.-Y., Song, J.F., Ang, K.W., Yu, M.B., Lo, G.Q., et al.: WDM multi-channel silicon photonic receiver with 320 Gbps data transmission capability. Opt. Express **18**, 5106–5113 (2010). 2010/03/01

13. Cao, Z., Proietti, R., Yoo, S.J.B.: Photonics Conference (IPC), pp 180–181, October 2014
14. Feng, Q.Y., Sang, X.Z., Dou, W.H.: Demonstration of a 5 Gb/s 24 interchip optical interconnect system. Microw. Opt. Technol. Lett. (2012)
15. Xie, M., Lu, Y.T., Wang, K.F., Liu, L., Cao, H.J., Yang, X.J.: TIANHE-1A interconnect and essage-passing services. IEEE Micro Mag. (2012)