Summary High-performance computing (HPC) has penetrated into various research fields, yet the increase in computing power is limited by conventional electrical interconnections. The proposed architecture, NEST, exploits wavelength routing in arrayed waveguide grating routers (AWGRs) to achieve a scalable, low-latency, and high-throughput network. For the intra pod and inter pod communication, the symmetrical topology of NEST reduces the network diameter, which leads to an increase in latency performance. Moreover, the proposed architecture enables exponential growth of network size. Simulation results demonstrate that NEST shows 36% latency improvement and 30% throughput improvement over the dragonfly on an average.

Key words: high-performance computing, optical interconnects, optical switches, network topology

1. Introduction

Exascale computing systems have an important influence on many scientific fields such as biotechnology, fluid dynamics, and materials science. These scientific disciplines are getting more and more "computational" today. Demand for computing power in these fields impose great pressure on the network size of high-performance computing (HPC) systems. With the extraordinary growth in the number of computing nodes, performance is increasingly determined by how data is transmitted among the numerous compute systems [3]. The interconnection network largely determines the efficiency, throughput, and latency. Innovative data movement technology is necessary for future exascale computing systems to break through the limitations of conventional electrical networks. By contrast, optically interconnected computing systems have higher scalability and energy efficiency, high-bandwidth achieved through wavelength division multiplexing, and low-latency without store-and-forward process [6]. Further, optical devices with wavelength (frequency) routing capability enable non-blocking switching between compute nodes. The arrayed waveguide grating router (AWGR) [7], [8] is such a device that forwards the signal at an input port to an output port depending on the wavelength of the signal. However, the commercial AWGR has a limited number of switch ports and we also need to concern about how to use a restricted number of ports to build a large-scale network.

2. Proposed Architecture

The proposed architecture, NEST, is organized with servers, racks, and pods. Figure 1 illustrates the NEST which exploits passive AWGR for intra pod and inter pod communication, while intra rack communication is provided by top-of-rack switch (TOR).

The NEST network is defined by a 3-tuple \((S, R, P)\) where \(P\) is the total number of pods, \(R\) is the total number of racks per pod, and \(S\) is the sum of servers per rack. To achieve good scalability, the above parameters should satisfy \(S = R = P = k\) and \(k\) is an even number because of the symmetrical structure of NEST. Therefore, the total number of servers in NEST is the multiplicative factor \(N = S \times R \times P = k^3\), and network scale achieves exponential growth. The NEST with \(k = 48\) can meet the demand of exascale computing system with 100 000 servers and the

![Fig. 1](image-url) The proposed architecture NEST is composed of servers (S), racks (R), and pods (P) \((S = R = P = k = 4)\). There are \(k^3\) \((S \times R \times P = k^3)\) servers in the network. The AWGRs with red edge and yellow edge are responsible for inter pod and intra pod communication, respectively. Blue links represent the optical links.
48-port AWGRs are commercially available.

In Fig. 2, TOR switch has one embedded electrical switch with 2× commodity optical transceivers. The NEST architecture uses commodity optical transceivers with fixed wavelength, because tunable wavelength transceivers are quite expensive [9] and the time of wavelength conversion increases the communication latency. It is more suitable for NEST to use fixed wavelength transceivers in the network without optical relay. The embedded switch performs the routing function within the TOR switch to transmit the packets to the proper output buffer, which connects to the optical transceivers. The optical transceivers with different wavelengths correspond to different destination addresses, and packets with the same destination will be sent to the same output buffer. Moreover, the communication between servers in the same rack is completed by this embedded switch, while half of the optical transceivers are in charge of transmitting packets among different racks within the same pod and the other half of the transceivers are responsible for the communication among the different pods.

In the pod, k TOR switches are directly connected to the AWGR for intra pod communication, which is marked in yellow edge in Fig. 2. The red edge AWGR numbered AWGR_i is only connected to the TOR switches with the same number. Routing table for a m-port AWGR is described by the following formula:

$$\lambda_n \rightarrow n = \begin{cases} n = i + o - 1, & i + o - 1 \leq m \\ n = (i + o - 1) \mod m, & i + o - 1 > m \end{cases}$$

where $\lambda_n$ is the wavelength obtained from input port $i$ and output port $o$. Intra pod (local) and inter pod (remote) communication have been separated from one another to achieve the full utilization of the wavelengths. In NEST, the wavelength that used for local communication are completely reused for remote communication, which facilitates the scalability of network and saves wavelength resources. Therefore, the NEST with 100,000 nodes only needs 48 wavelengths. It is worth noting that up to 160 (and theoretically more) separate wavelengths or channels of data can be multiplexed into a lightstream transmitted on a single optical fiber using dense wavelength division multiplexing (DWDM). Moreover, a 512 × 512 AWGR with a channel spacing of 25 GHz has been demonstrated in 2014 [10].

3. Analysis and Comparison

In this section, we compare the performance of NEST with several existing network topologies. For the considerate performance evaluation, we study the performance of NEST under various traffic patterns, namely, uniform traffic pattern (UF), hot-spot traffic pattern (HS), and nearest-neighbor traffic pattern (NN). To evaluate the latency and throughput of different network topologies, we use a event-based network simulator OPNET that provides performance management for computer networks and applications.

We consider a NEST with $k = 10$ (1000 nodes), a three-layer fat-tree [11], [12] with 16-port switches (1024 nodes), a 33-group dragonfly [13] with eight 8-port switches in each group (1056 nodes), and a 6x6x6 3D Torus [14] with 5 nodes per switch (1080 nodes). Compared to those networks with the same number of uplinks and downlinks, the configuration of only one node per switch in 3D Torus is unfair. Therefore, each switch in 3D Torus has 5 nodes to achieve similar uplink and downlink bandwidth while maintaining similar size to other networks. The length of each packet is 512 Bytes and line rate is 10Gb/s. Electrical switch delay and cable delay are set to be 40ns and 5ns per meter, respectively. Optical transceivers have a modulation rate of 56Gb/s and opto-electrical (OEO) conversion delay is calculated by dividing the packet length by the modulation rate. However, the OEO conversion only happens at TOR switches, which does not impose latency burden on the network performance. In Fig. 3 (a), the performance metric value is normalized with respect to the dragonfly and that of Fig. 3 (b) is normalized with respect to the fat-tree.

Figure 3 (a) demonstrates the normalized network latency of NEST compared with other existing networks. The symmetrical structure of NEST reduces the network diameter of the system, which improves the network latency on an average of 36% over the dragonfly. In a NEST, the AWGR within a pod can implement non-blocking communication between racks and the AWGRs outside the pod can effectively distribute traffic between pods. Although the network latency of NEST is similar to the fat-tree, NEST has better scalability since it can scale to 100,000 nodes network with 48-port AWGRs while a three-layer fat-tree has to use 74-port switches to achieve the same size.

In spite of the fixed node degree, the scalability of 3D Torus is limited by the network diameter which grows quickly with the network size. From Fig. 3 (a), its perfor-
Fig. 3 (a) Normalized network latency of NEST compared with other three networks. (b) Normalized network throughput of NEST compared with other three networks.

Performance seems to be better than other networks under the uniform traffic pattern. Because the latency performance is mainly determined by the average distance of network before the saturation point. The average distance of 6x6x6 3D Torus with 5 nodes per switch is about 4.5, which is the shortest of the four networks. At the same scale, increasing the number of nodes per switch will reduce the network diameter of 3D Torus. However, this does not mean that the network can be scaled by constantly increasing the number of nodes connected to each switch, because oversubscription will degrade network performance [11]. Moreover, its throughput performance is not good in Fig. 3 (b). This phenomenon shows that the network performance reaches saturation prematurely. The reason why 3D Torus performed better than dragonfly under the nearest-neighbor traffic pattern is that there is only one global link between the adjacent groups in dragonfly. It is more likely for the global links to become the performance bottleneck of the network. However, the number of links between the adjacent groups in NEST is growing with the expansion of network scale, which gives NEST a good path diversity.

In addition to the nearest-neighbor traffic pattern, NEST also performs well under the hot-spot traffic pattern. In the NEST, each rack within the same pod has optical connections to the racks in other pods and the high bandwidth provided by wavelength division multiplexing can better solve the traffic convergence problem. However, scarce global link resources make it difficult for dragonfly to efficiently handle hot-spot traffic and there are many paths overlapping in 3D Torus. Although fat-tree performs best in this situation, it needs more network devices, which leads to more system costs.

The throughput performance of fat-tree and NEST is close and their difference is mainly in scalability and network cost. In the NEST, wavelength multiplexing can not only increase network bandwidth but also reduce the number of links. On an average, the NEST shows 30% and 63% throughput improvement over the dragonfly and 3D Torus, respectively. The details for comparison of these topologies are summarized in Table 1. In contrast, NEST uses fewer devices for better performance, lower complexity and higher practicality.

4. Conclusion

We proposed a scalable, low-latency, and high-throughput optical interconnect architecture based on AWGRs. The network size of the proposed architecture has grown exponentially, which suits the rising trend of high-performance computing networks. Low network diameter of NEST reduces the communication latency and optical interconnect accompanied with AWGRs in the network can achieve high throughput. Simulation results show that NEST improves the network latency and throughput on an average of 36% and 30% over the dragonfly under the different traffic patterns. Although the performance of NEST and fat-tree is close, NEST has better scalability than fat-tree.

Acknowledgments

This work was supported by the National Science Foundation of China under Grants 61634004, and Grant 61472300, the Fundamental Research Funds for the Central Universities Grant No. JB180309 and No. JB170107, the key research and development plan of Shaanxi province No. 2017ZDCXL-GY-05-01, and the General Armament Department and Ministry of Education United Fund under Grant No.6141A0224-003.

References

[1] A. Shalaby, I. Fujiwara, and M. Koibuchi, “The case for network coding for collective communication on hpc interconnection networks,” IEICE Trans. Inf. & Syst., vol.E98-D, no.3, pp.661–670, 2015. DOI: 10.1587/transinf.2014EDP7255.
[2] R. Lucas, J. Ang, K. Bergman, S. Borkar, W. Carlson, and L. Carrington, “DOE Advanced Scientific Computing Advisory Sub-committee (ASCAC) Report: Top Ten Exascale Research Challenges,” 2014. DOI: 10.2172/1222713.
[3] C. Kachris, and I. Tomkos, “Power consumption evaluation of all-optical data center networks,” Cluster Computing, vol.16, no.3, pp.611–623, 2013. DOI: 10.1007/s10586-012-0227-6.
[4] A. Grama, Introduction to Parallel Computing, China Machine Press, 2003.

[5] F. Abel, C. Minkenberg, I. Iliadis, T. Engbersen, M. Gusat, F. Gramsamer, and R.P. Luijten, “Design issues in next-generation merchant switch fabrics,” IEEE-ACM Trans. Netw., vol.15, no.6, pp.1603–1615, 2007. DOI: 10.1109/TNET.2007.909727.

[6] M. Koibuchi, I. Fujiwara, K. Ishii, S. Namiki, and F. Chaix, H. Matsutani, H. Amano, and T. Kudoh, “Optical network technologies for hpc: computerarchitects point of view.” IEICE Electron. Express, vol.13, no.6, 2016. DOI: 10.1587/elex.13.20152007.

[7] S. Kamei, M. Ishii, M. Itoh, T. Shibata, Y. Inoue, and T. Kitagawa, “64 x 64-channel uniform-loss and cyclic-frequency arrayed-waveguide grating router module,” Electron. Lett., vol.39, pp.83–84, 2003. DOI: 10.1049/el:20030044.

[8] H. Harai, “Optical packet & circuit integrated network for future networks,” IEICE Trans Commun, vol.95, no.3, pp.714–722, 2012. DOI: 10.1587/transcom.E95.B.714.

[9] C. Kachris, and I. Tomkos, “A survey on optical interconnects for data centers,” IEEE Communications Surveys & Tutorials, vol.14, no.4, pp.1021–1036, 2012.DOI:10.1109/SURV.2011.122111.00069.

[10] S. Cheung, T. Su, K. Okamoto, and S.I.B. Yoo, “Ultra-compact silicon photonic 512 x 512 25 ghz arrayed waveguide grating router,” IEEE J. Sel. Topics Quantum Electron., vol.20, no.4, pp.310–316, 2014. DOI: 10.1109/JSTQE.2013.2295879.

[11] M. Al-Fares, A. Loukissas, and A. Vahdat, “A scalable, commodity data center network architecture,” SIGCOMM Comput. Commun. Rev, vol.38, no.4, pp.63–74, 2008. DOI: 10.1145/1402946.1402967.

[12] M.C. Chia, D.K. Hunter, I. Andonovic, P. Ball, I. Wright, S.P. Ferguson, K.M. Guild, and M.J. O’Mahony, “Packet loss and delay performance of feedback and feed-forward arrayed-waveguide gratings-based optical packet switches with WDM inputs-outputs,” J. Lightw. Technol., vol.19, no.9, pp.1241–1254, Sept. 2001. DOI: 10.1109/50.948271.

[13] J. Kim, W.I. Dally, S. Scott, and D. Abts, “Technology-driven, highlyscalable dragonfly topology,” ACM SIGARCH Comput. Archit. News, vol.36, no.3, pp.77–88, 2008. DOI: 10.1145/1394608.1382129.

[14] S. Horiguchi and T. Ooki, “Hierarchical 3D-torus interconnection network,” Proc. Int. Symp. Parallel Archit., Algorithms Netw., 2000. DOI: 10.1109/ISPAN.2000.900261.