Multiple granularities switching in intra-data center interconnect

Qian Kong1*, Chongfu Zhang1,2, Zichuan Yi1, Miao Yu1, Limengnan Zhou1, Youxian Chen1 and Yu Miao1

1 School of Electronic and Information Engineering, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan, Guangdong 528400
2 School of Information and Communication Engineering, Zhongshan Institute, University of Electronic Science and Technology of China, Chengdu 611731, China

*Corresponding author’s e-mail: kongqian01@yeah.net

Abstract. We propose a DCN interconnect based on hybrid OCS and OBS for dynamic DCN connectivity provisioning. Numerical results show that the proposed interconnect has preferable performance in dealing with various traffic granularity.

1. Introduction

The rise of the cloud computing and emerging big-data applications has significantly increased the traffic within the data center (DC). The bandwidth bottleneck and growing power requirements have become central challenges for high performance DCN interconnect. Various optical interconnect [1-9] have been proposed to take advantage of the high bandwidth capacity and low power consumption offered by optical switching. K. Chen et al. [10] proposed the optical data center interconnect ‘OSA’, which can dynamically change its topology through configuring the optical circuit switching (OCS) module, each top of rack switches (ToR) can setup light path connection with any other ToR simultaneously. Y. Huang et al.[11] proposed a DCN combining OCS and optical packet switching (OPS) in a unified architecture. In order to support 100 Gbit/s optical packets, hybrid optoelectronic packet routers are used. However, OPS is constrained by its system scalability and complexity which is hard to satisfy the demand of the growing large number of ports in DC. In addition, OPS switches do not have any buffering capability to reduce losses, and they therefore suffer much higher losses than their electronic packet switching (EPS) counterparts. Imran et al.[12] proposed a software defined optical burst switching data center interconnect, which employ a single-hop topology with a two-way reservation protocol that results in zero burst loss.

In summary, how to cope with various traffic granularity [13] (both mice and elephant flow) efficiently become more and more crucial issue in the DC network. In this paper, based on our previous work [14, 15], we propose a hybrid OCS/OBS interconnect for data center network. The services are differentiated and aggregated by the switch and interface card (SIC) embedded in the server, thus completing the separation of OCS services and OBS services. The performance of the interconnect in terms of delay and packet loss rate is also investigated.
2. Principle of design

The hybrid OCS/OBS network interconnect for the proposed DCN is illustrated in Fig. 1. The server is directly connected to the network through the OCS switches. The servers connected to the same OCS switches are called intra-cluster servers, and servers connected to different OCS switches are called inter-cluster servers. Each of the intra-cluster OCS switches are connected by an inter-cluster OCS switches. OBS switches (which can be implemented by a waveguide array grating (AWG) or a semiconductor optical amplifier (SOA)) are connected in both intra- and inter- cluster OCS switches. Each optical component in the network is connected through a proxy and a central controller. In particular, the OBS switches internally implement distributed control and interact resource information with the central controller.

![Diagram of Hybrid OCS and OBS intra-data center interconnect](image)

Fig.1. Hybrid OCS and OBS intra-data center interconnect

The servers in the racks are directly connected to the intra-cluster optical circuit switches by replacing the traditional network interface card by switch interface card (SIC, which can be implemented by a programmable logic device (PLD)), and each server is connected to the optical circuit switches through such an interface. Each of the SIC is connected to the central controller through an agent. The agent is responsible for reporting the service status information of the server, interacting with the traffic request information, and configuring the control command. OBS switches are connected to each of the OCS cluster, and the OBS can be shared by all intra-cluster servers. Each of the intra-cluster optical switches has a corresponding interface to an inter-cluster optical switch, and the inter-cluster optical switch is responsible for communication between the inter-cluster servers. In addition, the inter-cluster optical switch is also connected with an inter-cluster OBS switches, which is responsible for the switching of optical burst packets between the inter-cluster servers. All optical switching components are connected
by a proxy and a central controller, which in charge of providing resource information and configuration of the optical modules.

![Flow chart of traffic processing in SIC](image)

**Fig. 2. Flow chart of traffic processing in SIC**

ToRs are not employed in the proposed optical interconnect, which means the traffic do not have to aggregate on the electrical layer. Therefore, the SIC fully utilizes random-access memory (RAM) in the server to buffer the traffic, and implements MAC address-based routing between the servers. In addition to the function of traditional network interface card, the SIC can send and receive mixed OCS/OBS traffic by reading and writing data on the server or sending and receiving data according to the communication protocol. Different switching scenario (such as OCS scenario, OBS scenario, hybrid OBS and OCS scenario) can be realized by programmable logic control devices. The SIC designed two 10Gbps links for hybrid OCS and OBS. Based on the LUT, Data interaction between servers can select either OCS or OBS mode. The received traffic can be sent directly to the corresponding port without processing. In addition, there are two 10Gbps links serve as the internal communication interface of the OCS racks. This interface enables communication between servers in the same rack. When used as an optical circuit switches, the received traffic can be sent to other OCS module without returning to the server.

This paragraph describes how the combination of the centralized and distributed control works and the traffic process flow in detail. As shown in Fig. 2, when a new connection request arrives, the network controller is cognitive to the requested bandwidth and latency. After evaluation, the corresponding modules establish demanding optical link according to the requested bandwidth and latency. The SIC can choose to connect the request service to the OCS port or the OBS port to establish the request connection from ethernet to the optical path. Meanwhile, SIC also supports the aggregation function of the ethernet service, which include the traffic to the same destination are packed together, different latency of the incoming traffic can be divide into mice and elephant flow respectively and implement OCS or OBS switching functionality according to the traffic granularity. For delay-sensitive services, SIC supports ethernet frame transmission, ethernet frame is configured to the OBS module through OCS
module, while OCS switches provide a channel for OBS switches through centralized control, and client data perform the switching through distributed control in the OBS module, and switch back through the OCS switches to destination server. In this process, centralized control and distributed control should be corporate with each other to ensure the data from source server to the destination server.

3. Numerical results and discussions
OPNET simulation software is employed to investigate the performance of the proposed optical interconnect. The generated traffic arrival rate obeys to the Poisson distribution and has a random destination address. To capture the burstiness of data at the source nodes, the traffic from a user to the destination server is generated by Poisson packet arrivals rate from 50 to 100 independent traffic sources that follows an ON/OFF model (with or without traffic). We investigated the scalability of the proposed interconnect through enlarge the input/output ports of optical burst switches.

Assuming the ports of optical circuit switches are large enough to support the connection of the corresponding servers and the OBS units, we change the scale optical burst switches. The input/output ports adopt 4×4, 8×8, 16×16 and 32×32 respectively, and the number of connected servers increases linearly, and the number of servers in the cluster is 10, 20, 40, 60 respectively. To OCS unit in the cluster, each of the cluster and the inter-cluster OCS switches are connected to the OBS switch of the port size as described above.

In this simulation, there are four intra-cluster optical circuit switches and one inter-cluster optical circuit switch. The total number of servers is 40, 80, 160, 320. The principle of control and routing is described in the section 2, we assume that the traffic that needs to perform OCS is configured by the central controller before the traffic enters the network. That means client data is connection oriented and switching is performed only on the OBS unit. Therefore, the service latency discussed in this section do not include the configuration time of the optical circuit switches. Since the OCS performance has been analyzed in our previous work [14,15], this section will focus on the impact of the proposed interconnect on OBS.
Distributed control is implemented in the OBS module, the processing unit inside the switching module completes switching according to MAC address. For burst switching process, we use the improved Just-Enough-Time (JET) protocol as the transmission protocol. In our simulation, the average burst size is set to 100 Kbytes, the average packet size is fixed to 256 bytes in the Poisson case. The performance of the delay and packet loss rate of the proposed interconnect is investigated. As illustrated in Fig. 3, the delay of the system increases with the expansion of the burst switching input/output ports. As the traffic load increases, the delay increases more sharply. The reason is higher traffic load leads to more burst traffic in the system, thus increasing the probability of the traffic competing for the same port. The greater probability of port competition leads to an increase in the traffic blocking rate, which results in more packets retransmitted or lost and an increase in service delay. On the other hand, when the load of the service exceeds 0.8, the overflow of server RAM module is unavoidable.

In this case, many packets will count to the network packet loss, other packets will be retransmitted depend on the demands. This process greatly increases the service delay. As can be seen from Fig. 4, after the traffic load reaches 0.6, the packet loss rate increases significantly. This is due to the significant usage of the RAM module as the traffic load increases. When the traffic load reaches 0.7, the server RAM module overflow, and the packet loss rate increases significantly.

4. Conclusion
This paper proposes a hybrid OCS and an OBS interconnect to deal with multiple flows in intra-data center network. A SIC is introduced in the server, which enables directly communicate between servers without the top of the rack switches and fully utilize the internal RAM of the server to buffer the traffic. After the process of SIC, traffic is packaged and aggregated to OCS or OBS modules according to the granularity. A combination of centralized and distributed control scheme is employed in the control plane. The central controller in charge of the configuration of the whole OCS network. Distributed control is implemented in the OBS network. The resource information in the OBS module are uploaded to the central controller, so that the whole network resource is transparent to the central controller. This paper also investigated the performance of OBS module in the proposed interconnect in terms of delay and packet loss rate of traffic load, and evaluate the influence of different port sizes. The results show
that when the payload is not more than 0.6 (normalized to 1), the proposed interconnect indicates preferable performance.

Acknowledgement
This work was supported in part by National Natural Science Foundation of China (61571092, 61901096), Research projects for young innovative talents of Guangdong Province (2019KQNCX188, 2018KQNCX330, 2018KQNCX332), Zhongshan Science and Technology Bureau (2018B1021), Research Project for Talent of UESTC Zhongshan Institute (418YKQN07) and Project for Innovation Team of Guangdong University (No. 2018KCXTD033).

References
[1] F. Testa, L. Pavesi, Optical switching in next generation data centers, Springer, 2018.
[2] C. Kachris and I. Tomkos, “A Survey on Optical Interconnects for Data Centers,” in IEEE Communications Surveys & Tutorials, vol. 14, no. 4, pp. 1021-1036, Fourth Quarter 2012
[3] Xiang Zhou, Ryohei Urata, Hong Liu Beyond 1 Tb/s Intra-Data Center Interconnect Technology: IM-DD OR Coherent? in Journal of Lightwave Technology, vol. 38, NO. 17, 2020
[4] D. Kong et al., "Intra-Datacenter Interconnects with aSerialized Silicon Optical Frequency Comb Modulator," in Journal of Lightwave Technology, vol. 38, no. 17, pp. 4677-4682, 2020.
[5] L. Zhang, J. Chen, E. Agrell, R. Lin, and L. Wosinska, "Enabling Technologies for Optical Data Center Networks: Spatial Division Multiplexing," J. Lightwave Technol. 38, pp.18-30,2020.
[6] D. Zou, Z. Zhang, F. Li, Q. Sui, J. Li, X. Yi, and Z. Li, "Single λ 500-Gbit/s PAM Signal Transmission for Data Center Interconnect Utilizing Mode Division Multiplexing," in Optical Fiber Communication Conference (OFC) 2020, OSA Technical Digest (Optical Society of America, 2020), paper W1D.6.
[7] X. Xue, F. Yan, K. Prifti, F. Wang, B. Pan, X. Guo, S. Zhang, and N. Calabretta, "ROTEOS: A Reconfigurable and Cost-Effective Architecture for High-Performance Optical Data Center Networks,” J. Lightwave Technol. 38, pp.3485-3494, 2020.
[8] M. Yang, P. Tien, W. Ruan, T. Lin, S. Wen, P. Tseng, C. Lin, C. Chen, C. Chen, Y. Luo, M. Tsai, and S. Zhong, "OPTUNS: Optical intra-data center network architecture and prototype testbed for a 5G edge cloud [Invited],” J. Opt. Commun. Netw. 12, pp.A28-A37, 2020.
[9] Yi Shu, Shuangyi Yan, Chris Jackson, Koteswararao Kondepu, Emilio H. Salas, Yan Yan, Reza Nejabati, Dimitra Simeonidou, Programmable OPS/OCS hybrid data centre network, Optical Fiber Technology, Volume 44, 2018, pp.102-114.
[10] M. Fiorani, S. Aleksic, M. Casoni, L. Wosinska and J. Chen, "Energy-Efficient Elastic Optical Interconnect Architecture for Data Centers," in IEEE Communications Letters, vol. 18, no. 9, pp. 1531-1534, Sept. 2014
[11] Y. Huang, Y. Yoshida, S. Ibrahim, R. Takahashi and K. Kitayama, "Load balancing in switch-fabric type of torus OPS data center networks with hybrid optoelectronic routers," 2016 21st Optoelectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS), Niigata, 2016, pp. 1-3.
[12] Muhammad Imran, Martin Collier, Pascal Landais, Kostas Katrinis, Performance evaluation of TCP over software-defined optical burst-switched data center network, Journal of Computational Science, Volume 24, pp. 44-53, 2018.
[13] S. Kandula et al. The nature of data center traffic: measurements & analysis in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, (2009).
[14] Qian Kong, Shanguo Huang, Bingli Guo, Xin Li, Min Zhang, Yongli Zhao, Jie Zhang, Wanyi Gu, "Numerical and experimental study of a cascaded microelectromechanical system-based all-optical data center interconnect," Opt. Eng. 55(7) 076111
[15] Qian Kong, Yafeng Zhan, Chaowei Duan, “Optical comb enabled flexible-bandwidth single-carrier frequency-division-multiplexed-based intra-data centre interconnect”, IET Communications, 2018, 12, (20), pp. 2623-2629.