Proceedings of CECNet 2021
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210451

ReDCN: A Dynamic Bandwidth Enabled Optical Reconfigurable Data Center Network

Xinwei ZHANG[1], Zuoqing ZHAO, Yisong ZHAO, Yuanzhi GUO, Xuwei XUE, Bingli GUO and Shanguo HUANG
The State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract. A reconfigurable optical data center network is proposed, in which the optical bandwidth can be automatically reconfigured by reallocating time slots based on the real time traffic. Numerical investigations validate that the network performance of packet loss after reconfiguration decreases by 58.5%, and the end-to-end latency decreases by 63.8% with comparison to the network with rigid link interconnections, and thereby increasing the 9.4% of throughput at load of 0.8.

Keywords. Reconfigurable bandwidth, optical data center network, time slots

1. Introduction

With the rapid development of traffic boosting applications, the traffic presents a rapid growth trend in data centers (DCs) [1-2]. This poses unprecedented challenges to the existing data center network (DCN) based on electric switching technology from the aspects of both switching technology and network architecture [3-4]. The development and application of optical switching technology in DCN has been extensively investigated to overcome the bandwidth bottleneck of electrical switches [5-6]. However, the high bandwidth between the top-of-racks (TORs) in optical DCNs is uniform after the network building, which cannot reallocate bandwidth according to the variety of traffic in real-time [7-8].

In the latest researches, a feasible solution is to configure specific transmission links or devices for specific applications using load distribution algorithms [7]. However, the deployment of load distribution mechanism will increase the complexity of network control and management, and then increase the cost, especially for large networks. Another method is to build a flexible reconfigurable network with the capability to dynamically adjust optical bandwidth [9-10]. However, the proposed schemes need complicated network interconnections and control mechanism. This is hard to afford the large-scale network with requirements of low management cost and power-consumption.

In this paper, we propose a reconfigurable optical DCN, ReDCN, which can allocate time slots according to the traffic proportion to then reconfigure the bandwidth. In the

[1] Corresponding Author, The State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China; E-mail: zhangxinwei@bupt.edu.cn.

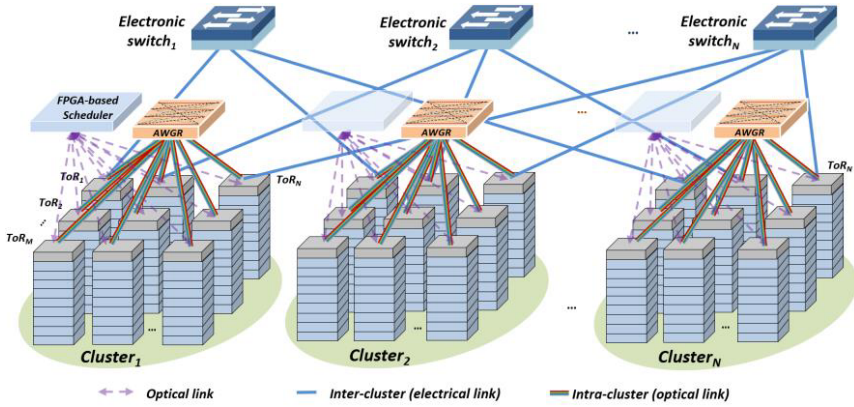

Figure 1. The structure of the proposed reconfigurable optical DCN.

proposed ReDCN network, we implement a field programmable gate array (FPGA)-based scheduler, which is used to deploy the reconfigurable instructions. The scheduler reconfigures the time slots and bandwidth according to the collected ToR traffic and topology information to provide adaptive optical bandwidth for the links in the cluster. The network can provide adaptable optical bandwidth to the hosted application with variety of traffic, and thereby reduce the packet loss rate, improves the latency and throughput performance.

2. Reconfigurable Architecture

Figure 1. shows the proposed network which is divided into *n* clusters. There are *k* servers interconnected through ToR in each rack, and each cluster contains *n* racks. The traffic (Ethernet frame) generated by the server is divided into three types (intra ToR, intra cluster and inter cluster) according to its destination. Ethernet frames are first processed at the Ethernet switch of each ToR. The intra-cluster links are interconnected by the AWGR, while the *i*-th electronic switch interconnects the *i*-th ToR of each cluster, where $i = 1, ..., N$. Therefore, inter cluster communication only needs two hops. This interconnection improves the flexibility of the network.

2.1. Schematic of the ToR

The function block of ToR implemented by the FPGA is shown in **Figure 2.**. Each rack contains *k* servers interconnected through Ethernet switches in ToR. The Ethernet switch processes Ethernet frames with different destinations generated by the server and accordingly forwards them to the AWGR and the electrical switch. In each ToR, *n* transceivers (TRXs) with corresponding electric buffers are deployed to store the traffic to the AWGR (*p*) and electric switch (*n-p*) respectively for intra cluster and inter cluster communication. By processing the MAC address of each frame destined for intra cluster and inter cluster links, ToR can calculate the traffic of each link and sends this traffic statistics (traffic distribution within and between clusters) to the FPGA based controller on the optical link.

2.2. Reconfigurable Scheme

In this work, the distributed scheduling system based on FPGA connects ToR through optical links to monitor and collect traffic and topology information. The controller redistributes the time slot according to the collected traffic information, and reconfigures the optical bandwidth of the connection in the cluster to adapt to the current traffic mode.

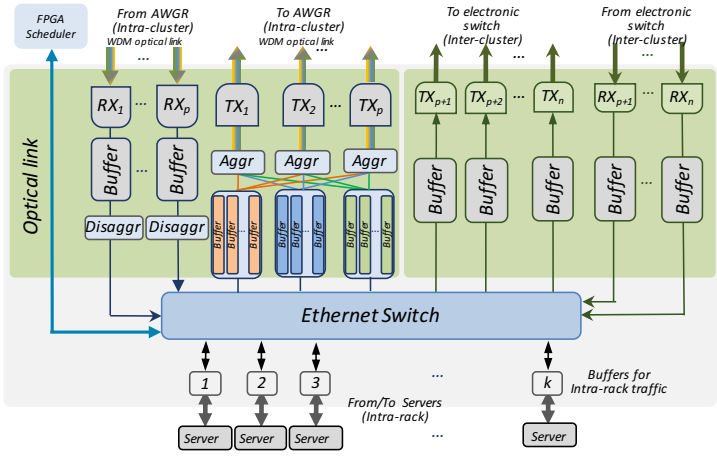

Figure 2. Schematic of the ToR

To highlight the characteristic of the reconfiguration scheme, an example assumes that 4 clusters containing 4 racks in a DCN. Each ToR has 4 TRXs, and they are connected through optical links with AWGR. The *i*-th buffer caches the traffic sent to the *i*-th ToR, *i*=1, 2, 3, 4. The traffic ratio distributing to different destinations is different. We design the reconfiguration scheme by reallocating the different scheduling time for each buffer so that they can forward traffic with corresponding bandwidth. As the

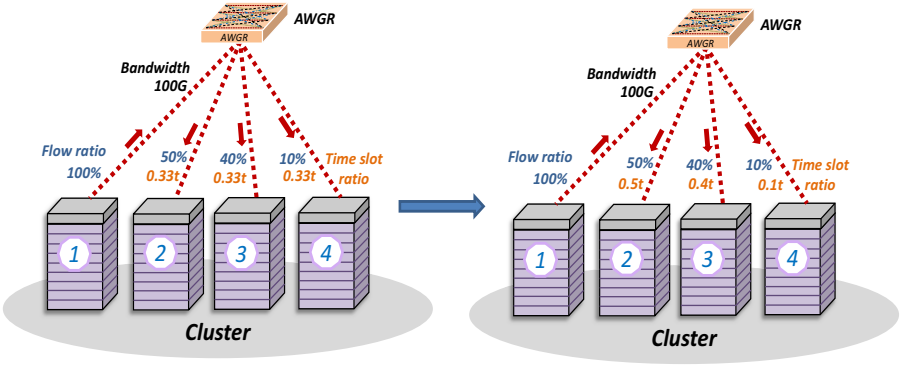

Figure 3. Reconfiguration scheme

reconfiguration scheme is shown in **Figure 3.**, ToR_1 sends 50%, 40%, 10% intra traffic to ToR_2, ToR_3, ToR_4, respectively. Before reconfiguration, the bandwidth is uniformly allocated for the links to the different destination. In our reconfiguration scheme, the controller will adjust the bandwidth allocation by reallocating time slots. So the cycle is accordingly divided into $0.5t$, $0.4t$ and $0.1t$ to schedule corresponding buffers, with t

representing one scheduling cycle. FPGA is a parallel processing logic element, the time slot allocation can be faster, so as to reconfigure the optical bandwidth and realize fast switching scheduling.

3. Simulation and Results

In our experiment, the DCN with 2560 servers is simulated on OMNeT++ platform, which is divided into 8 clusters, 8 racks in each cluster and 40 servers in each rack. ToRs have 2 ports, each port connect with 4 destination ToRs in the same cluster through an AWGR. The traffic ratio in the rack, intra-cluster and inter-cluster are shown in **Table 1.**. The Intra-cluster traffic proportion on each link connected to AWGR is the same and the division of different ToR's intra traffic is shown in **Table 2.**. After reconfiguration, the bandwidth is divided according to the traffic proportion. There are 4 time slots in each cycle, and all ToRs schedule a buffer in each time slot. The link rate between server and ToR is 10Gbps, and the link rate from each buffer to AWGR is 100Gbps. The buffer size of ToR is 80KB, and the buffer queue from the ToR to the server is 20KB.

Table 1. Studied traffic pattern.

	Intra-ToR	**Intra-cluster**	**Inter-cluster**
Traffic	50%	37.5%	12.5%

Table 2. The ratio of intra-cluster traffic

	ToR_1	**ToR_2**	**ToR_3**	**ToR_4**	**ToR_5**	**ToR_6**	**ToR_7**	**ToR_8**
ToR_1	0	0.2	0.2	0.1	0.05	0.15	0.2	0.1
ToR_2	0.2	0	0.1	0.2	0.15	0.05	0.1	0.2
ToR_3	0.2	0.1	0	0.2	0.2	0.1	0.05	0.15
ToR_4	0.1	0.2	0.2	0	0.1	0.2	0.15	0.05
ToR_5	0.05	0.15	0.2	0.1	0	0.2	0.2	0.1
ToR_6	0.15	0.05	0.1	0.2	0.2	0	0.1	0.2
ToR_7	0.2	0.1	0.05	0.15	0.2	0.1	0	0.2
ToR_8	0.1	0.2	0.15	0.05	0.1	0.2	0.2	0

Figure 4. shows the simulation results in terms of the packet loss, ToR-to-ToR latency, Server-to-Server latency and throughput before (2560 network) and after reconfiguration (2560 network-R). The packet loss begins to deteriorate when the load is 0.6 when the bandwidth is uniformly allocated, and it begins to deteriorate at the load of 0.8 after reallocating timeslots. Buffers in ToR will be filled faster so that the packet loss in the ToR buffers is more when the load is higher, verifying our reconfiguration scheme is more effective at the high load. Because the bandwidth is reallocated according to the traffic, the buffer can be scheduled in time, so the packet loss rate reduced by at most 58.5%.The ToR-to-ToR latency and Server-to-Server latency begins to deteriorate

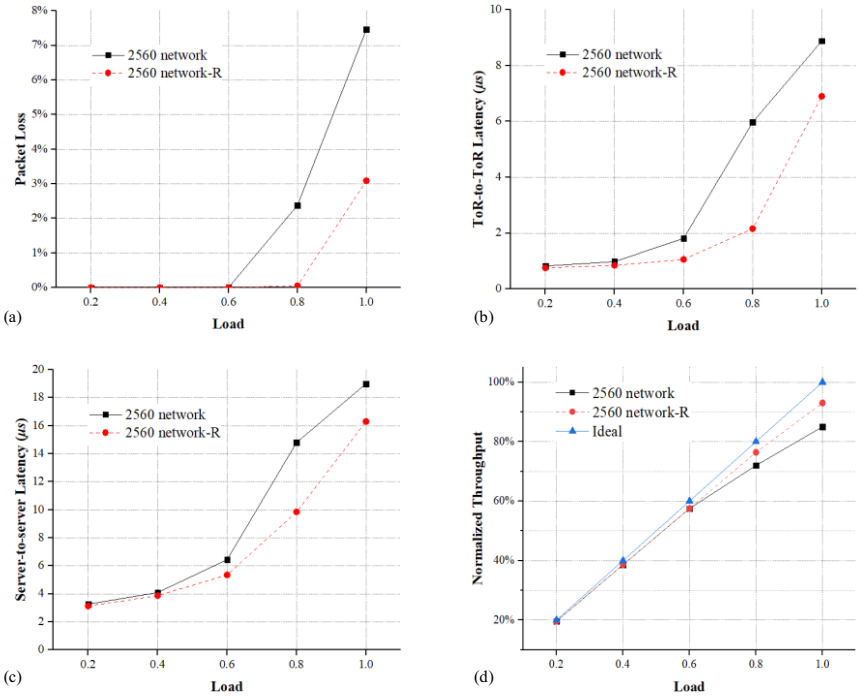

Figure 4. (a)Packet loss (b)ToR-to-ToR latency (c)Server-to-Server latency (d)Throughput

at the load of 0.6, and it gets improvement by 63.8% after reconfiguration since the bandwidth is adapted to real time traffic. When the load is higher, more traffic arrive at server's buffer queues, so the packet loss in the server queues becomes dominant. Because we only calculate the latency of successfully received traffic, so the latency after reconfiguration increases faster at load over 0.8. Due to less packet loss, the throughput is also improved by 9.4% when the link is fully occupied.

4. Conclusions

In this paper, we propose a reconfigurable optical DCN in which the optical bandwidth can be reallocated based on the monitored real time traffic. Numerical results prove the proposed ReDCN improves packet loss by 58.5%, end-to-end latency by 63.8% and throughput by 9.4% with compared to the network with rigid interconnections, validating the good flexibility of the proposed network.

Acknowledgements

The project was supported by Fund of State Key Laboratory of Information Photonics and Optical Communications (Beijing University of Posts and Telecommunications) (No. IPOC2021ZT08), P. R. China.

References

[1] X. Xue et al., "ROTOS: A Reconfigurable and Cost-Effective Architecture for High-Performance Optical Data Center Networks," Journal of Lightwave Technology, 38(13), pp.3485-3494, 2020.
[2] M. Analytics, "The age of analytics: competing in a data-driven world," ed: San Francisco: McKinsey & Company, 2016 (online). See https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.
[3] C. Guo, et al. "Bcube: a high performance, server-centric network architecture for modular data centers." ACM SIGCOMM, 2009.
[4] C. Guo, et al. "DCell: A scalable and fault-tolerant network structure for data centers," ACM SIGCOMM, 38(4):75–86, 2008.
[5] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan, "c-Through: part-time optics in data centers," ACM SIGCOMM Comput. Commun. 40, 327–338 (2010).
[6] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," ACM SIGCOMM Comput. Commun. 41, 339–350 (2010).
[7] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," IEEE INFOCOM, 2010, pp. 1–9.
[8] X. Xue, F. Yan, B. Pan and N. Calabretta, "Flexibility Assessment of the Reconfigurable OPSquare for Virtualized Data Center Networks Under Realistic Traffics," *2018 European Conference on Optical Communication (ECOC)*, 2018, pp. 1-3.
[9] X. Xue, F. Wang, F. Agraz, A. Pagès, B. Pan, F. Yan, X. Guo, S. Spadaro and N. Calabretta. "SDN-controlled and orchestrated OPSquare DCN enabling automatic network slicing with differentiated QoS provisioning," Journal of Lightwave Technology, 38(6), pp.1103-1112, 2020.
[10] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, "OSA: An optical switching architecture for datacenter networks with unprecedented flexibility," IEEE/ACM Transactions on Networking, vol. 22, no. 2, pp. 498-511, 2013.