Assessment of Reconfigurable OPSquare for Flexible and Cost-effective Optical Data Center Networks

Xuwei Xue, Fulong Yan, Kristif Prifti, Bitao Pan, Xiaotao Guo, Nicola Calabretta

IPI-ECO Research Institute, Eindhoven University of Technology, Eindhoven, the Netherlands

Elastic bandwidth allocation of the reconfigurable OPSquare DCN supporting multi-tenant is assessed. End-to-end latency <2µs and packet loss <10^{-4} are guaranteed for different traffic patterns by dynamically configuring the WDM ToRs and optical switches.

Introduction

Various applications such as big data, cloud computing, and artificial intelligence offered by several enterprises and service providers are hosted in a multi-tenant approach in large data centers (DC) [1]. The DC network (DCN) resources, the switches capacity, and the optical bandwidth are shared among tenants through dedicated network slices (NSs) [2]. Multi-tenant applications with heterogeneous traffic characteristics sharing the same infrastructure can drastically deteriorate the DCN performance.

Optical DCN architectures have been widely investigated due to their advantages of data rate and format transparency as well as overcoming the high cost and power consumption of O-E-O conversions [3]. Recently we have proposed an optical DCN architecture, OPSquare that supports the generation and reconfiguration of multiple independent NSs and novel load balancing algorithms [4, 5]. However, the bandwidth of the optical links interconnecting the Top-of-the-rack switches (ToR) is static by the pre-allocated WDM transceivers between the ToRs and the optical switches. This prevents the possibility to dynamically allocate more bandwidth on specific links to enable the interconnection of a larger pool of servers located in different racks or to support a larger traffic load for specific application.

In this work, we empower the reconfiguration of the OPSquare architecture by elastically controlling the WDM transceivers (TRXs) and optical switches in order to support the optical interconnect links and shared infrastructure with variable bandwidth allocation. The performance improvements of the reconfigurable OPSquare is assessed under realistic traffics from multi-tenant virtual networks. Numerical results confirm that 2 µs end-to-end latency and 10^{-4} packet loss can be achieved for different traffic shapes based on the dynamic reconfiguration of the WDM TRXs at each ToR. Moreover, the deployments of more WDM TRXs at each ToR to reconfigurable OPSquare network will not cause dramatic increase in cost and power consumption.

Reconfigurable OPSquare DCN

The reconfigurable OPSquare DCN is shown in Fig. 1(a). The electrical ToR switch interconnects H-server in each rack, and M racks are grouped into one cluster. There are

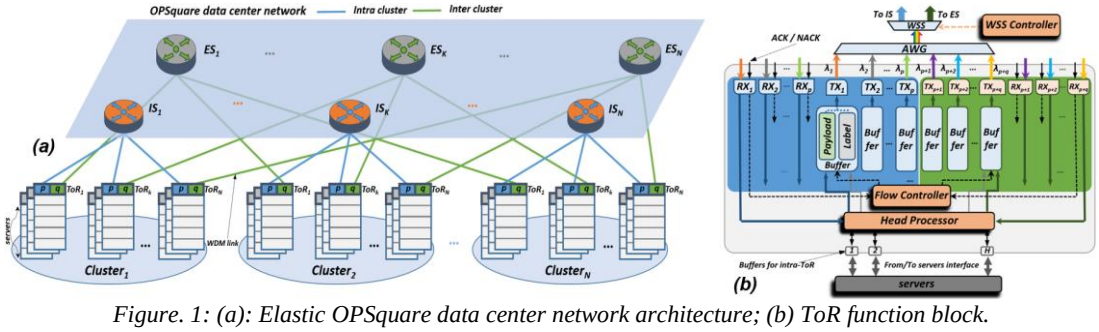

Figure. 1: (a): Elastic OPSquare data center network architecture; (b) ToR function block.

N clusters in the proposed DCN. The traffic generated by the servers can be classified into three categories (intra-ToR, intra-cluster, inter-cluster) and will be handled by the ToRs. The N N×N intra-cluster optical switches (IS) and N N×N inter-cluster optical switches (ES) are dedicated for intra-cluster and inter-cluster communication, respectively. The *i*-th ToR in each cluster are interconnected by the i-th ES (1≤i≤N). For intra cluster and inter-cluster traffic, p and q WDM transceivers with dedicated electronic buffers are utilized to interconnect the ToR with the IS and ES, respectively. The amounts of p and q can be elastically allocated on-demand according to desired capacity, oversubscription, and intra-/inter-cluster traffic ratio.

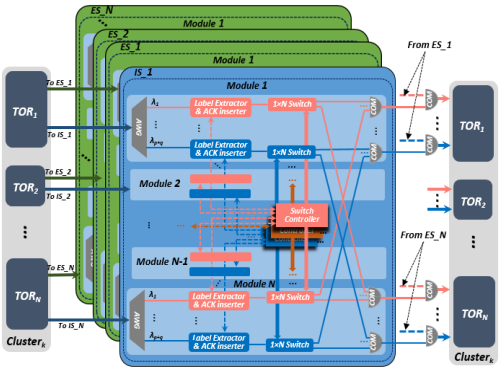

Figure. 2: Schematic of the optical switch

The ToR function blocks are shown in Fig. 1(b). The incoming Ethernet frames with variable lengths and destinations are first processed by the head processor. For the intra-ToR traffic, the frames will be forwarded to and buffered at the ports which connect the destination servers in the same rack. While for the intra/inter-cluster communication, the frames will be forwarded to the ports associated with the intra-cluster TXs (p) or inter-cluster TXs (q) to be aggregated into the optical packets. The copy of the optical packets are sent to the destination ToR via the IE/ES by the WDM TX. An optical label identifying the destination ToR is attached to the packet and will be extracted and processed at the ES/IS nodes. Based on the label information, the amounts of p and q WDM transceivers used to interconnect the ToR to the IS and ES through the optical flow-controlled link can be tuned accordingly by the head processor. A wavelength selective switch (WSS) is used to tune the wavelengths (and thus the bandwidth per link) to ES and to IS. The reason for employing multiple amounts tunable (p + q) WDM transceivers is twofold. First, the communication bandwidth between ToR and IS/ES can be scaled to generate the high capacity channel (waveband super-channel). Second, by elastically tuning the amounts of p and q, the optical bandwidth between the ToR and optical switching network can be flexibly allocated on-demand according to the desired oversubscription and intra-/inter-cluster traffic ratio.

The optical switch featuring a modular structure is shown in Fig. 2 and each module handles the WDM traffic coming from the same ToR. There are p+q wavelengths in each modular. The label extractor separates the optical label from the payload and processes it on-the-fly, while the payload is fed into the broadcast and select 1×N switch. With the recovered label bits, the controller enables the SOA gates in the 1×N switch to forward the packets. The SOA has fast nanoseconds switching speed and can provide optical amplification to compensate the splitting losses caused by the broadcasting architecture. The combiner aggregates the identical wavelength with the same destination. The switch controller resolves the possible contention and controls the 1×N switch accordingly. It is worth to notice that only one wavelength could pass the combiner at every time-slot.

Assessments of elastic bandwidth allocation

OMNeT++ is utilized to simulate the reconfigurable OPSquare DCN. The simulated DCN deploys 23,040 servers interconnected by 576 ToRs. In each rack, 40 servers operating at

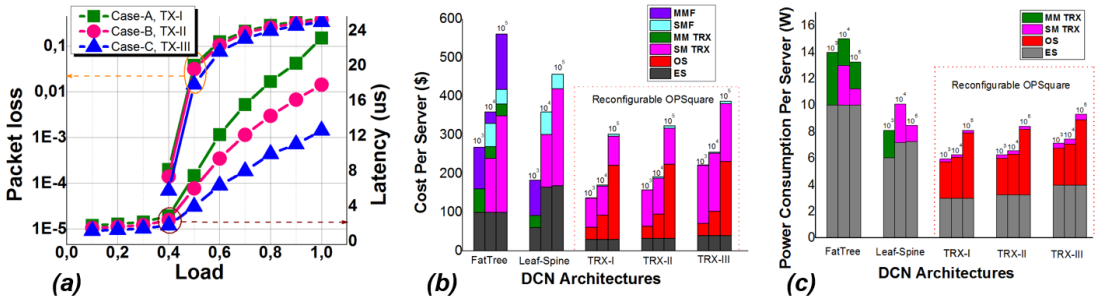

Figure. 3: (a): Packet loss and server-to-server latency versus traffic load for different traffic patterns with specific TX types; (b) Normalized cost (c) power consumption per server of various DCN architectures. Single mode fiber (SMF), Multi-mode fiber (MMF), Single mode transceiver (SM TRX), Multi-mode transceiver (MM TRX), Optical switches (OS), Electrical switch (ES).

10 Gb/s generate packets with length ranging from 64-byte to 1518-byte at a given load. Each WDM TX operates at 50 Gb/s and is equipped with 50 KB buffer. The radix of IS and ES is 24. The impact of reconfiguring the optical bandwidth between the intra- and inter-cluster is investigated. The total number of WDM TRXs interfaces of each ToR is fixed to 6, that means p+q=6. The amounts of p and q are allocated on-demand according to the intra/inter-cluster traffic volume. Three kinds of traffic patterns generated by different DCN applications (pool of interconnected servers in different ToRs) are considered: case A (62.5% intra-ToR, 0% intra-cluster and 37.5% inter-cluster traffic), case B (50% intra-ToR, 25% intra-cluster and 25% inter-cluster traffic), and case C (25% intra-ToR, 75% intra-cluster and 0% inter-cluster traffic). The TRX configurations for the intra- and inter-cluster connectivity have been investigated to optimize the provided bandwidth according to the different applications. Fig. 3(a) shows the performance results in term of packet loss and latency. For different traffic patterns, the numerical results show satisfied performance when WDM TX configuration TX-I (p = 3 and q = 3) is employed for serving traffic in case A, WDM TRX configuration TX-II (p = 4 and q = 2) for case B, and TX-III (p = 6 and q = 0) for case C, respectively. For load < 0.4, there are no packet loss for all the three cases, and 10^{-4} packet loss can be guaranteed at load of 0.4. Adapting the link bandwidth for the case A, B and C can provide < 2 µs server to server latency for 0.4 load. Therefore, by building a dynamic and adaptive OPSquare DCN with variable p and q WDM TRXs, the reconfigurable DCN properly adapted to traffic pattern variations can maintain satisfying network performance.

Cost and power consumption calculation

In this section, we compare the cost and power consumption of the electrical (Fat-Tree and Leaf-Spine) DCN and reconfigurable OPSquare architecture DCN deploying different types of WDM TRX configurations. In the calculation, the same number of servers for the three architectures has been considered with the equal contribution to the cost and power consumption. We explore the case of scaling out the DCN with the different number of servers (1000, 10000 and 100,000) operating at 10 Gb/s data rate. Moreover, to have a fair comparison, all DCN architectures have no oversubscription at the ToRs. To efficiently improve the performance of the DCN, the WDM TRXs can be dynamically allocated to vary the link bandwidth and oversubscription as the requirements of various applications. Therefore, to have a comprehensive understanding of the cost and power consumption affected by the elastic amounts of WDM TRXs, three

types of WDM TRX configuration ((TRX-I: p = 3 and q = 1), (TRX-II: p = 4 and q = 1) and (TRX-III: p = 6 and q = 2)) for the reconfigurable OPSquare architecture are considered for these calculations.

Fig. 3(b) shows the normalized cost of different DCN architectures and DC size. As the DC scales out, the costs of the electrical DCN architectures is much higher than the proposed reconfigurable OPSquare optical architecture mainly due to the more cost of the electrical switches (ES). Because of the more usage of multimode components, the cost of Flat-Tree is beyond other architectures. For these three scales (1000, 10000 and 10000 servers) of DC, the reconfigurable OPSquare is the most cost-efficient, especial for the DC scaling 1000 servers. Compared with Fat-Tree, the reconfigurable OPSquare equipped with TRX-I configuration has a cost saving of almost 50%. The cost of fiber is neglectable for OPSquare benefiting from the high switching data rate of fast optical switches (OS) and the WDM technology. The cost of the reconfigurable OPSquare will increase with more usage of transceivers. Fig. 3(c) reports the normalized power consumption per server of the three architectures. The power consumption of Fat-Tree and Leaf-Spine with large size (100,000 servers) is less than medium size (10,000 servers) due to the employment of SM_WDM_4 TRXs. For the reconfigurable OPSquare, the power consumption is the lowest benefiting from the adopting of power-efficient optical switches and the less deployment of electrical switches. More deployments of WDM transceivers to reconfigurable OPSquare architecture will not cause dramatic increase in power consumption.

Conclusions

A reconfigurable OPSquare for virtualized DCN under realistic traffics is proposed and numerically assessed. WDM TRXs deployed at each ToR can be flexibly used to supply the desired oversubscription and support different traffic patterns. Numerical results indicated that elastically allocating 6 WDM TRXs with 50 KB buffer per TX, an OPSquare DCN of 23,040 servers could properly adapt to three different traffic patterns while maintaining a 10^{-4} packet loss and < 2 μs end-to-end latency at load of 0.4. Moreover, with the elimination of a large amount of electrical transceivers, the proposed reconfigurable OPSquare architecture has high cost and power efficiency with respect to the Fat-Tree and Leaf-Spine DCN.

Acknowledgement

The authors would like to thank the H2020 Passion (780326) and H2020 Qameleon (780354) projects for partially supporting this work.

References

[1] Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper, USA, 2018.
[2] Metzger. K, “Data Center Outsourcing Trending Up,” Data Center Solutions Group, CBRE, 2017.
[3] Yan. F and Calabretta. N: HiFOST: A scalable and low-latency hybrid data center network architecture based on flow-controlled fast optical switches. IEEE/OSA JOCN 10, no. 7 (2018): 1-14.
[4] Xue. X and Calabretta. N: Flexibility Assessment of the Reconfigurable OPSquare for Virtualized Data Center Networks Under Realistic Traffics, in ECOC, pp. 1-3, 2018.
[5] Xue. X and Wang. F: Experimental Assessment of SDN-enabled Reconfigurable OPSquare Data Center Networks with QoS Guarantees. OFC, pp. M3F-4, 2019.