Experimental Demonstration of Hitless OCS-based DCN Reconfiguration to Steer Multi-Class Traffic

(*Invited Paper*)

Qian Lv, Zhihuang Ma, and Zuqing Zhu[†]
School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[†]Email:{zqzhu}@ieee.org

***Abstract*—To realize hitless optical datacenter network (ODCN) reconfiguration for improving the specific QoS of multi-class traffic flows, we propose a novel topology engineering (TPE) and traffic engineering (TE) scheme, and demonstrate its effectiveness experimentally in a real ODCN testbed.**

I. Introduction

The recent surge in network services has put great pressure on data-center networks (DCNs) [1–4]. Compared with electrical inter-rack architectures, optical interconnects can potentially offer larger throughput, higher energy efficiency, and shorter latency [5]. Therefore, researchers have proposed various all-optical interconnects and explored dynamic topology engineering (TPE) and traffic engineering (TE) with them to deal with the heavy, skewed and highly-dynamic multi-class traffic in today's DCNs [6]. For instance, the optical DCN (ODCN) in Fig. 1(a) is built by interconnecting pods with optical circuit switches (OCS') [7], and dynamic TPE and TE can be realized in it to steer traffic. However, an OCS reconfiguration can induce millisecond-level service interruptions [6], degrading the quality-of-service (QoS) of traffic in ODCNs severely.

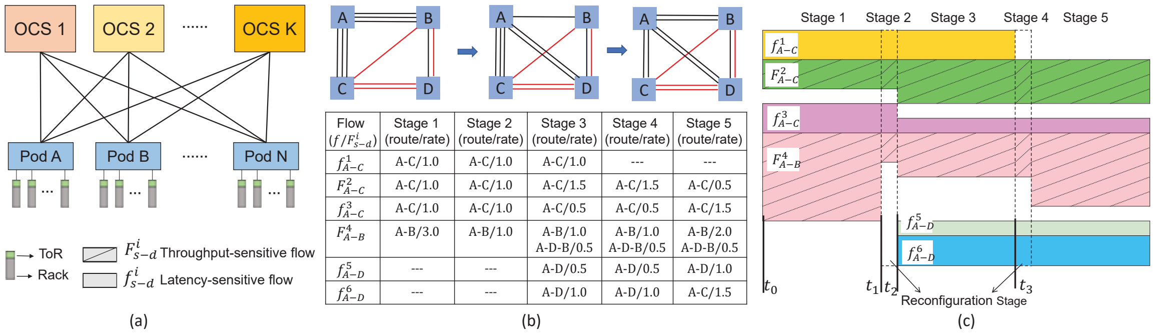
Several studies [6, 7] have tried to achieve TPE and TE with reduced numbers of service interruptions. However, these existing approaches just treated all the traffic in ODCN equally to reduce service interruptions during OCS reconfigurations, but did not differentiated the QoS demands of multi-class traffic [8]. For example, the traffic flows in a DCN can be roughly classified into two categories [9]: *throughput-sensitive* and *latency-sensitive* flows, whose QoS demands are significantly different. A throughput-sensitive flow usually tries to occupy as much bandwidth as possible to minimize the flow completion time (FCT), but cares less about the end-to-end (E2E) delay between source and destination as its contribution to FCT is negligible [10]. The QoS of a latency-sensitive flow depends mainly on the E2E delay, since untimely messaging is unacceptable for it. Therefore, we should not simply treat these two types of flows equally when planning the OCS reconfigurations to improve their QoS, and only minimizing their service interruptions will not be good enough.

This work studies how to optimize TPE with OCS reconfiguration and TE with rerouting to realize hitless ODCN reconfiguration for ensuring the QoS of throughput-sensitive and latency-sensitive flows simultaneously. We first design the approach to optimize the TPE and TE in an ODCN according to its network status, such that the E2E delays of latency-sensitive flows can be ensured when fixed bandwidth is provisioned to each of them and the FCTs of throughput-sensitive flows can be minimized. Then, we build a small-scale but realistic ODCN testbed and experimentally demonstrate hitless OCS-based DCN reconfiguration to steer multi-class traffic with QoS guarantee. Experimental results show that our proposal can achieve QoS-aware transmission of flows with zero packet loss, and minimize the FCTs of throughput-sensitive flows and the E2E delays of latency-sensitive flows. Specifically, compared with the scheme that treats all the flows equally, our proposal can reduce the maximum E2E delay of latency-sensitive flows by 51% without increasing the average FCT of throughput-sensitive flows significantly.

II. Hitless OCS-based DCN Reconfiguration to Steer Multi-Class Traffic

Fig. 1(a) shows the architecture of the OCS-based DCN that is considered in this work. The ODCN equips K independent OCS' to interconnect N pods, each of which contains several top-of-rack (ToR) switches that are directly connected to the OCS'. Then, to explain our proposal of optimizing the TPE and TE in such an ODCN clearly, we introduce an example in Figs. 1(b) and 1(c), where the ODCN consists of 4 pods and 2 OCS' and each pod has 3 direct connections to each OCS through its ToR switches. Specifically, Fig. 1(b) shows the TPE (top) and TE (bottom) schemes, while Fig. 1(c) illustrate the transmission rates of flows. Here, each black link in Fig. 1(b) denotes a connection between pods that can be reconfigured, and each red connection is one that is carrying traffic and thus cannot be reconfigured. The table at the bottom of Fig. 1(b) shows the routing paths and rates of in-service flows at each stage of TPE, and the blocks in Fig. 1(c) explain the transmission scheme of each flow in a more illustrative way.

In Fig. 1(b), we denote the i-th throughput-/latency-sensitive flow from s to d as $F^i_{s,d}/f^i_{s,d}$, respectively. There are 6 flows in the ODCN, including 4 latency-sensitive flows (hop-count ≤ 1) $f^1_{A\text{-}C}$, $f^3_{A\text{-}C}$, $f^5_{A\text{-}D}$ and $f^6_{A\text{-}D}$, and 2 throughput-sensitive flows $F^2_{A\text{-}C}$ and $F^4_{A\text{-}B}$. The capacity of each connection to/from an OCS is 1 unit of bandwidth. *Flows* $f^1_{A\text{-}C}$ and $f^3_{A\text{-}C}$ start at time t_0 and both of them use one units of bandwidth until t_2. After t_2, *Flows* $f^1_{A\text{-}C}$, $f^3_{A\text{-}C}$, $f^5_{A\text{-}D}$ and $f^6_{A\text{-}D}$ respectively require 1, 0.5, 0.5 and 1 units of bandwidth until their



Flow (f/F^i_{s-d})	Stage 1 (route/rate)	Stage 2 (route/rate)	Stage 3 (route/rate)	Stage 4 (route/rate)	Stage 5 (route/rate)
f^1_{A-C}	A-C/1.0	A-C/1.0	A-C/1.0	---	---
F^2_{A-C}	A-C/1.0	A-C/1.0	A-C/1.5	A-C/1.5	A-C/0.5
f^3_{A-C}	A-C/1.0	A-C/1.0	A-C/0.5	A-C/0.5	A-C/1.5
F^4_{A-B}	A-B/3.0	A-B/1.0	A-B/1.0 A-D-B/0.5	A-B/1.0 A-D-B/0.5	A-B/2.0 A-D-B/0.5
f^5_{A-D}	---	---	A-D/0.5	A-D/0.5	A-D/1.0
f^6_{A-D}	---	---	A-D/1.0	A-D/1.0	A-C/1.5

Fig. 1. (a) Architecture of ODCN based on OCS', (b) TPE (top) and TE (bottom) to enhance QoS of multi-class traffic with hitless OCS reconfigurations, and (c) Traffic scheduling schemes of throughput-/latency-sensitive flows.

transmissions are completed. As for the throughput-sensitive flows, $F^2_{A\text{-}C}$ and $F^4_{A\text{-}B}$ try to take all the bandwidth that is available to them so as to finish their data-transfers as soon as possible. Hence, apart from the initial inter-pod topology in Fig. 1(b), our TPE/TE scheme designs 2 other topologies and 5 stages of flow routing to enhance the QoS of the flows.

With the initial topology, the flow routing is as that in *Stage* 1, where $f^1_{A\text{-}C}$, $F^2_{A\text{-}C}$, $f^3_{A\text{-}C}$ and $F^4_{A\text{-}B}$ are directly transmitted (one hop) using 1, 1, 1 and 3 units of bandwidth, respectively. Then, to better serve the upcoming latency-sensitive flows $f^5_{A\text{-}D}$ and $f^6_{A\text{-}D}$, TPE should be performed. Hence, the TPE/TE scheme comes up with the second topology and the flow routing in *Stage* 2 to realize a hitless OCS reconfiguration, where $f^1_{A\text{-}C}$, $f^3_{A\text{-}C}$ and $F^2_{A\text{-}C}$ use the same rates as those in *Stage* 1, while the rate of $F^4_{A\text{-}B}$ is reduced to 1 unit of bandwidth. After the OCS is accomplished and the ODCN is operating stably with the second topology, the flow routing is updated to that in *Stage* 3, where $F^2_{A\text{-}C}$ takes two paths, one of which is relayed at a ToR switch of *Pod B* to get additional 0.5 unit of bandwidth. After the transmission of $f^1_{A\text{-}C}$ is completed, our TPE/TE scheme design the third topology to speed up the throughput-sensitive flows. Similarly, we introduce *Stage* 4 to ensure the transition of OCS reconfiguration is hitless to in-service flows. After the OCS reconfiguration, TPE has moved one connection from A-C to A-B, and thus $F^4_{A\text{-}B}$ can occupy 2 units of bandwidth over A-B and take 0.5 unit of bandwidth over A-D-B. Each of other in-service flows ($F^2_{A\text{-}C}$, $f^3_{A\text{-}C}$, $f^5_{A\text{-}D}$ and $f^6_{A\text{-}D}$) uses a direct connection, and they respectively use 1.5, 0.5, 0.5 and 1 units of bandwidth until flow completion.

III. Design of TPE/TE Scheme for Hitless ODCN Reconfiguration

To realize hitless ODCN reconfiguration that can improve the QoS of throughput-sensitive and latency-sensitive flows, we need to optimize when and how to implement TPE/TE. Specifically, TPE/TE should be invoked when the operator finds that the current topology of the ODCN and the flow routing in it cannot satisfy the QoS of flows or TPE/TE can improve their QoS effectively. Then, we find the TPE scheme that will not affect in-service flows by only reconfiguring the connections between the free ports on ToR switches, while such connections can be obtained in three ways: 1) locating the connections that do not carry any traffic, 2) rerouting throughput-sensitive flows or squeezing their bandwidth without violating their QoS demands, and 3) rerouting latency-sensitive flows without exceeding their hop-count limits. After a new topology has been got by the TPE, our TE scheme lets throughput-sensitive flows use all the routing paths that are available to them and fully utilize the bandwidth there to minimize their FCTs. Among the throughput-sensitive flows that share one path, we allocate bandwidth to them in proportion to their data volumes. Meanwhile, our TE scheme serves latency-sensitive flows with single-path routing and finds the paths that can satisfy both their bandwidth and hop-count requirements. In short, we design the TE scheme to satisfy the QoS demands of all the multi-class flows in the new topology.

In the following, we will refer to our proposed TPE/TE scheme as hitless ODCN reconfiguration considering multi-class traffic (H-ClassFlow). We also consider three benchmarks: 1) one-stage direct ODCN reconfiguration (1-D-ClassFlow), 2) multi-stage direct ODCN reconfiguration (m-D-ClassFlow), and 3) hitless ODCN reconfiguration treating all the flows equally (H-eqFlow). The 1-D-ClassFlow scheme just directly reconfigures the ODCN to a new topology (for improving the QoS of flows) without addressing the service interruptions during the reconfiguration, m-D-ClassFlow breaks an ODCN reconfiguration into multiple stages to mitigate the resulting service interruptions, and in H-eqFlow, we treat all the flows equally without addressing their specific QoS demands and invoke hitless ODCN reconfiguration.

IV. Experimental Demonstrations

We build a small-scale but realistic ODCN testbed to experimentally demonstrate the advantages of our proposed TPE/TE scheme. The testbed includes 4 pods and 2 OCS', each pod has 3 direct connections to each OCS through its ToR switches. The ToR switch of each pod equips six 1GbE optical ports and each OCS is based on an optical cross-connect (OXC). The TPE/TE scheme is implemented in the ODCN by leveraging the OpenFlow-based software-defined networking (SDN) [11], and we realize traffic generation and analysis with the Data Plane Development Kit (DPDK) in Linux system. We design 5 experimental scenarios, each of which uses a different initial ODCN topology and will run flows

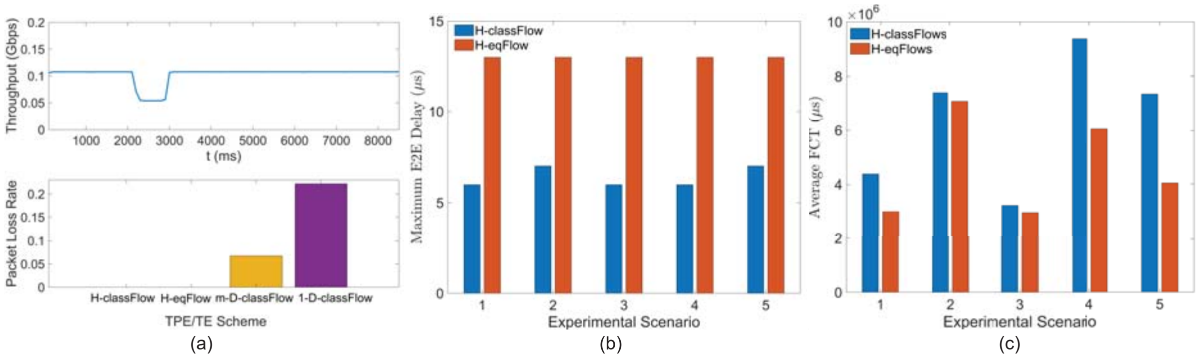


Fig. 2. Experimental results, (a) Service interruptions caused by TPE/TE schemes, (b) Maximum E2E delay of latency-sensitive flows, and (c) Average FCT of throughput-sensitive flows.

with different data volumes (within $[10, 50]$ GB) and skews (the ratio of throughput-sensitive flows to latency-sensitive flows ranges within $[1.5, 10]$). The hop-count limit of latency-sensitive flows is set to 1 and the minimum bandwidth demand of each throughput-sensitive is 1 Gbps. For each experimental scenario, we run the experiments 5 times and average the results to get each data point, for ensuring statistical accuracy.

Fig. 2(a) shows the service interruptions caused by the TPE/TE schemes. Specifically, the top subplot of Fig. 2(a) visually demonstrates the impact of direct ODCN reconfiguration on an in-service flow. We can see that when the ODCN reconfiguration starts at $t = 2$ s, the flow's throughput drops rapidly and stays at a low level for several hundred milliseconds until the reconfiguration is completed. The bottom subplot of Fig. 2(a) compares the average packet loss rates of the 4 TPE/TE schemes. It can be seen that the design of hitless reconfiguration in H-classFlow and H-eqFlow is effective because their packet loss rates are both zero. However, the schemes based on direct ODCN reconfiguration (1-D-ClassFlow and m-D-ClassFlow) cannot avoid packet losses, and as m-D-ClassFlow breaks each ODCN reconfiguration into multiple stages, its packet loss rate (6.75%) is lower than that of 1-D-ClassFlow (22.2%) but it is still too high to be acceptable for most of the data-driven applications in DCNs.

Next, we compare the performance of the two TPE/TE schemes based on hitless reconfiguration (H-ClassFlow and H-eqFlow). Fig. 2(b) shows the results on the maximum E2E delay of latency-sensitive flows in different experimental scenarios. As H-eqFlow treats all the flows equally to ensure hitless reconfiguration and reduce their FCTs, it cannot properly address the QoS demands of latency-sensitive flows. Therefore, the maximum E2E delay from H-eqFlow is much longer than that from H-classFlow. Specifically, by jointly considering the specific QoS demands of latency-sensitive and throughput-sensitive flows, H-classFlow achieves an average reduction of the maximum E2E delay by 51% over H-eqFlow. Note that, E2E delay is an important performance metric for many network services, especially for those of high-frequency trading applications in the financial markets, where a small edge of a few microseconds per transaction can translate to a fairly large amount of loss or profit. As H-eqFlow treats all the flows equally to ensure hitless reconfiguration and to reduce their FCTs, its average FCTs are shorter than those from H-classFlow in Fig. 2(c). However, the gap between the results of the two TPE/TE schemes is not significant, and the shorter average FCT from H-eqFlow is obtained by sacrificing the maximum E2E delay of latency-sensitive flows.

V. Conclusion

We proposed a novel TPE/TE scheme that can realize hitless ODCN reconfiguration for ensuring the specific QoS demands of multi-class traffic flows. Experimental demonstrations in a small-scale but realistic ODCN testbed verified the effectiveness of our proposal, and the experimental results indicated that the maximum E2E delay of latency-sensitive flows can be reduced by 51% without increasing the average FCT of throughput-sensitive flows significantly.

References

[1] P. Lu *et al.*, "Highly-efficient data migration and backup for Big Data applications in elastic optical inter-datacenter networks," *IEEE Netw.*, vol. 29, pp. 36–42, Sept./Oct. 2015.

[2] L. Gong and Z. Zhu, "Virtual optical network embedding (VONE) over elastic optical networks," *J. Lightw. Technol.*, vol. 32, pp. 450–460, 2014.

[3] Z. Zhu *et al.*, "Impairment- and splitting-aware cloud-ready multicast provisioning in elastic optical networks," *IEEE/ACM Trans. Netw.*, vol. 25, pp. 1220–1234, Apr. 2017.

[4] L. Gong *et al.*, "Efficient resource allocation for all-optical multicasting over spectrum-sliced elastic optical networks," *J. Opt. Commun. Netw.*, vol. 5, pp. 836–847, Aug. 2013.

[5] Z. Zhu, W. Lu, L. Zhang, and N. Ansari, "Dynamic service provisioning in elastic optical networks with hybrid single-/multi-path routing," *J. Lightw. Technol.*, vol. 31, pp. 15–22, Jan. 2013.

[6] M. Zhang *et al.*, "Gemini: Practical reconfigurable datacenter networks with topology and traffic engineering," *arXiv preprint arXiv:2110.08374*, Oct. 2021. [Online]. Available: https://arxiv.org/abs/2110.08374.

[7] L. Poutievski *et al.*, "Jupiter evolving: Transforming Google's datacenter network via optical circuit switches and software-defined networking," in *Proc. of ACM SIGCOMM 2022*, pp. 66–85, Aug. 2022.

[8] J. Liu *et al.*, "On dynamic service function chain deployment and readjustment," *IEEE Trans. Netw. Serv. Manag.*, vol. 14, pp. 543–553, Sept. 2017.

[9] Z. Yang *et al.*, "Achieving efficient routing in reconfigurable DCNs," in *Proc. of ACM SIGMETRICS 2020*, pp. 1–30, Jun. 2020.

[10] W. Lu *et al.*, "AI-assisted knowledge-defined network orchestration for energy-efficient data center networks," *IEEE Commun. Mag.*, vol. 58, pp. 86–92, Jan. 2020.

[11] Z. Zhu *et al.*, "Demonstration of cooperative resource allocation in an OpenFlow-controlled multidomain and multinational SD-EON testbed," *J. Lightw. Technol.*, vol. 33, pp. 1508–1514, Apr. 2015.