Demonstration of Hitless OCS Provision for Multi-modal Traffic in a Centralized Scheduling Hybrid Optical/Electrical Datacenter Network

Shi Feng[1], Jiawei Zhang[1], Jun Dai[1], Yashe Liu[2], Xiaorun Wang[1], Yuefeng Ji[1]
[1]*State Key Lab of Information Photonics and Optical Communications, Beijing Univ. of Posts and Telecommunications (BUPT), Beijing, China*
[2]*Huawei Technologies Co., Ltd*
Corresponding author Emails: {zjw, jyf}@bupt.edu.cn

Abstract: We demonstrate a hitless OCS provision in a centralized scheduling hybrid optical/electrical datacenter network through a real-time FPGA-based testbed. Experimental results show that it achieves a low packet delay and flow completion time accelerations. © 2024 The Author(s)

1. Introduction

With the rapid growth of cloud-based intra-datacenter traffic, more challenges have been brought to the switching capability of datacenter networks (DCNs). The traffic inside DCNs appears multi-modal characteristics that present diverse flow patterns: the long-lasting bandwidth-intensive transmission belonging to distributed computing frameworks such as Hadoop and the highly dynamic bursty transmission belonging to real-time interactive applications. The hybrid optical/electrical DCN (HOE-DCN) architecture [1-3] has been considered as an attractive alternative by exploiting the advantages of the two switching paradigms, where optical circuit switching (OCS) providing dedicated optical paths suit the former traffic pattern, and electrical packet switching (EPS) with more flexibility suits the latter. However, a big challenge of HOE-DCN is how to guarantee the continuity and quality of packet transmission when setting up and tearing down optical paths (a.k.a "hitless OCS provision"). There are mainly two approaches to realize the hitless feature, as illustrated in Fig. 1(a). One of them is "Make-before-arrival" [1], which relies on traffic prediction to preestablish optical paths before long-lasting traffic arrives. It requires a good "match" between the predicted arrival time of traffic and the optical path setup time, which is to shorten the holding time of the optical path before traffic arrives. Another one is "Make-after-arrival" [2, 3], where the setup of optical paths starts after the long-lasting traffic arrives, which provides the accurate configuration of optical paths. Before the setup is done, the long-lasting traffic is directed towards the EPS network, where the routers and switches inside the EPS network buffer the traffic randomly leading to high tail latency distribution. In light of the above, it is worth developing an HOE network with hitless OCS provision with stable latency performance.

In this paper, we propose a hitless OCS provision for multi-modal traffic in a centralized scheduling HOE-DCN and demonstrate it in an FPGA-based testbed. We exploit a central controller that arranges the sending time and end-to-end routing paths for long-lasting traffic flows inside the EPS network before the setup of optical paths. After the optical path is set up, the traffic flows are able to be transmitted through optical paths with hitless optical provision. We compare the performance of our work with a centralized scheduling electrical DCN. The experimental results

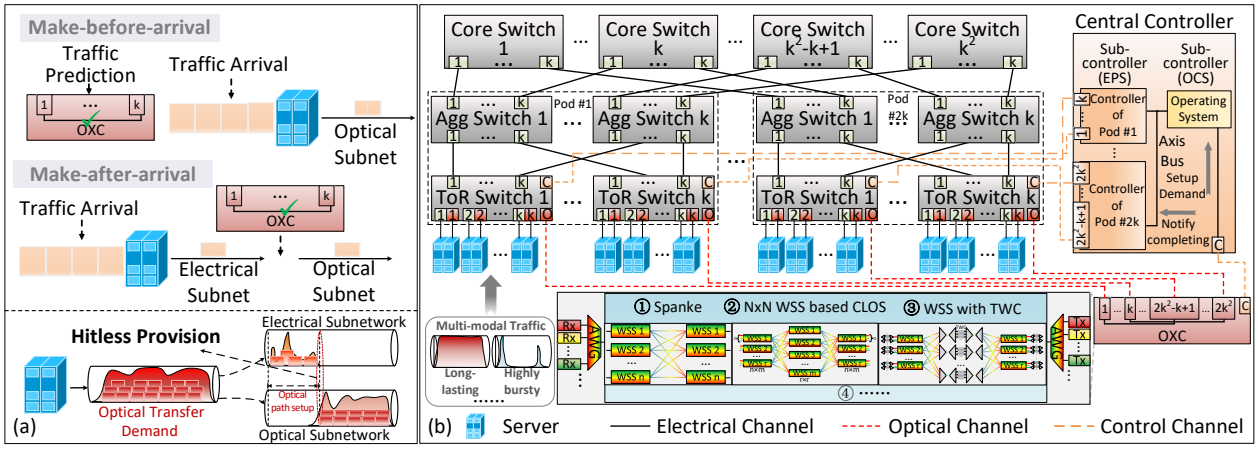

Fig. 1. (a) Mainstream approaches for hybrid electrical/optical datacenter networks; (b) Hybrid optical/electrical fat-tree networks topology

show that our work can achieve better performance on packet latency and flow completion time (FCT).

Disclaimer: Preliminary paper, subject to publisher revision

2. Architecture and Operating Mechanism

The centralized scheduling HOE-DCN architecture consists of three parts, as shown in Fig. 1(b), a central controller that interconnects all ToR switches with dedicated channels, an OCS network that interconnects ToR (Top-of-rack) switches through a multi-ports OXC (optical cross-connector), and a three-tier EPS network with a 2*k*-ary fat-tree topology [4], where 2*k* represents the total port number of each switch. To realize centralized scheduling, the central controller integrating electrical sub-controllers with an optical sub-controller receives traffic requests and sends scheduling results, where each electrical sub-controller schedules the traffic of a pod on the electrical network, and the optical one configures the optical paths. For the electrical network topology, there are 2*k* pods in total. Each 2*k*-port ToR switch is connected to *k* servers by two separate channels (electrical/optical). Besides, there are k^2 core switches for inter-pod communication. For the optical sub-network, there are several possible solutions to realize the function of OXC, such as Wavelength Selective Switches (WSSs) and Micro-Electro-Mechanical Systems (MEMSs). In this work, we consider the WSS-based OXC because of its potential advantage for a large number of switching ports by combining space dimension and wavelength dimension. The non-blocking is an important feature of OXC, and there are some promising structures of WSS-based OXC to realize it: (1) 1×N WSS based Spanke [5] (used in this work), provides strictly non-blocking on space dimension; (2) N×N WSS based CLOS [6], guarantees strictly non-blocking on space and wavelength dimensions under the condition of meeting the requirements for the number and scale of WSSs; (3) wavelength convertor based OXC [7], can realize fully non-blocking on space and wavelength dimensions by using wavelength convertors at each input ports of OXC.

All incoming Ethernet frames with variable lengths are encapsulated into data cells with a fixed length. The time is also divided into timeslots of fixed length relying on the whole network time synchronization. The scheduled data cells are transmitted according to the allocated timeslots. To realize the hitless provision, the flows with optical transfer demands utilize the electrical network before the setup of optical paths and transmit by OCS immediately when the setup is done. Based on the above, Fig. 2 shows the operating mechanism of our work: Consider a traffic demand between server 1 and server k^2 that requires long-lasting high bandwidth transmission. (1) Firstly, before the transmission, the network initialization starts with time-synchronization. The central controller sends Schedule (SCHD) frames, and the servers respond with request-to-send (RTS) frames. After several interactions, the time synchronization becomes stable. (2) Secondly, after the arrival of the flow, server 1 divides the flow into data cells with a fixed length and sends RTS frames including the traffic demands of each cell to the controller. To avoid the collision that multiple RTS frames compete for one output port towards the controller at ToR switches, the electrical controllers let servers belonging to the same ToR switch stagger the RTS frames sending time so that RTS frames from various servers arrive at the same ToR switch successively. (3) Thirdly, after the central controller receives the requests, the electrical controller decides to build the new optical channel from server 1 to server k^2 in OXC because there are the most cells queuing belonging to that source-destination pair currently. Before the setup of a new optical channel between ToR switch 1 and ToR switch *k*, all flows will be scheduled by the electrical controllers, the controllers allocate the transmission timeslots and end-to-end routing paths for each data cell by a first-fit algorithm. (4) Fourthly, the controllers send the SCHD frames to the servers at each timeslot similarly to RTS frames. (5) Finally, when the setup of the new optical path is accomplished, server 1 is able to communicate with server k^2 directly through the optical channel, where no other servers compete for the bandwidth slice belonging to server 1.

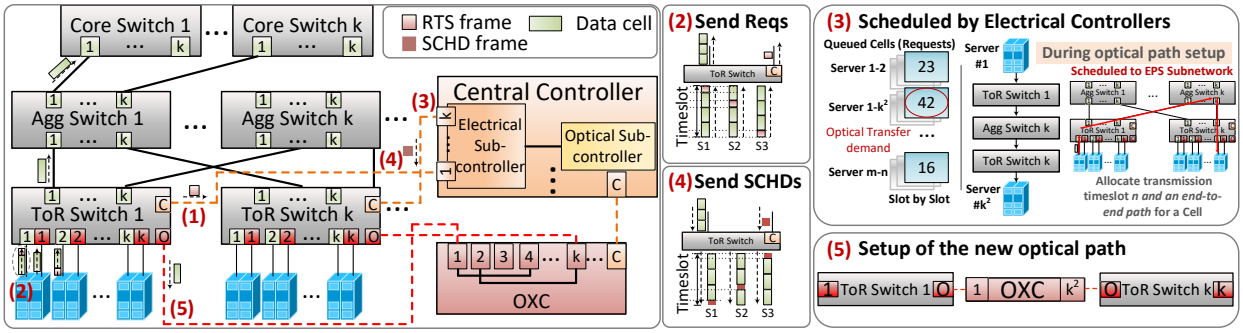

Fig .2. The operating mechanism of the centralized scheduling HOE-DCN fat-tree network

3. Experimental Setup and Results

We set up a one-pod scale HOE-DCN experimental testbed based on FPGA (Xilinx Virtex UltraScale+ VCU118) boards with two 100GbE Ethernet interfaces. As shown in Fig. 3(a). The overall testbed includes four servers, two ToR switches, two Agg switches, an OTU (Optical Transform Unit), an OXC, and a centralized controller. We separate each 40GbE interface of each FPGA board into four independent 10Gbps Ethernet interfaces. Four NICs of

Disclaimer: Preliminary paper, subject to publisher revision

servers are implemented with two FPGA boards, where each two NICs is deployed on a board. Each NIC connects with the Ethernet test center (ETC) through a 10Gbps Ethernet interface. ETC is used to generate the traffic from servers and analyze the flow performance. Each server communicates with a ToR switch through two independent 10GbE interfaces, one for the electrical subnetwork and another for the optical subnetwork. Two ToR switches and two Agg switches are implemented with four FPGA boards. The central controller is also deployed on one FPGA board connecting with all ToR switches. The OXC and OTU are connected together, and OTU interconnects two ToR switches providing wavelength transformation between grey and colored lights. Multi-modal traffic data flows consist of two intra-DCN traffic flows: the Hadoop cluster traffic from Facebook emulating the long-lasting flows with high bandwidth, and the search application traffic from Google emulating the high dynamic bursty flows [8].

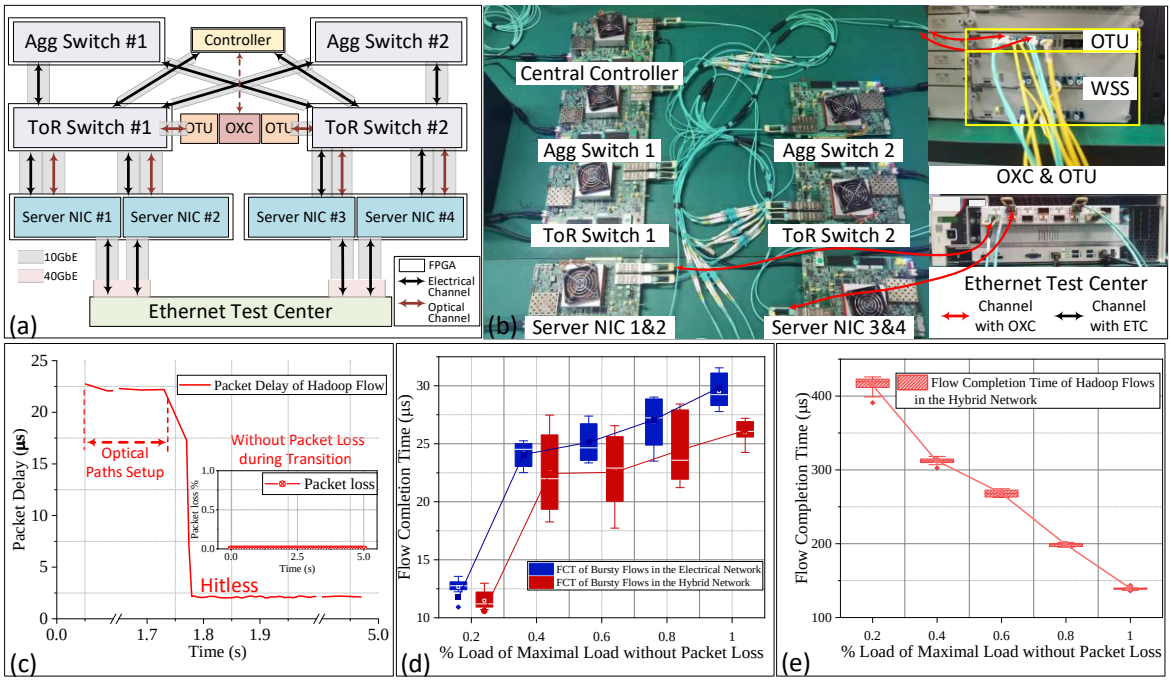

Fig. 3: (a) Overview of testbed; (b) Experimental setup; (c) Variation of packet delay of the Hadoop flow; (d) Flow completion time of the bursty flow; (e) Flow completion time of the Hadoop flow.

Fig. 3(c) presents the variation of the measured end-to-end packet latency of the Hadoop traffic. The packet latency starts with a relatively high value, then instantly decreases and stabilizes at ultra-low latency indicating the hitless provision feature. This is because the packets belonging to Hadoop are firstly scheduled to transfer in the EPS sub-network, and the OCS sub-controller starts to set up the optical path for Hadoop traffic. Once the setup is done, NICs of the servers with Hadoop flows will be notified by the central controller. After executing the EPS scheduling results for Hadoop cells, the servers will transmit the Hadoop packets toward the configured optical paths. The whole process above is hitless without packet loss. We also compare the centralized scheduling EPS network with our scheme in the FCT performance of burst flows. The results shown in Fig. 3(d) show that the hybrid network alleviates the burden for bursty flows achieving lower FCT than the EPS network. Fig. 3(e) indicates the FCT performance of the Hadoop flows in the hybrid network. As the load increases, the FCT decreases because the Hadoop flows utilize the optical path independently without any competition. Therefore, under the multi-modal traffic scenario, the hybrid electrical/optical network performs better than EPS with hitless provision.

4. Conclusions

We proposed and demonstrated hitless OCS provision for multi-modal traffic in a centralized scheduling HOE-DCN. We demonstrated its improving performance on end-to-end latency and flow completion time.

Acknowledgment

This work was supported by the National Key R&D Program of China (No. 2022YFB2903700), Huawei Cooperation Program, and the Xiaomi Young Talents Program。

References

[1] Li, Qinhezi, et al., JOCN 12.2, A113-A122 (2020).
[2] G, et al., ACM.SIGCOMM, 327-338 (2010).
[3] Farrington, et al., ACM.SIGCOMM, 339-350 (2010).
[4] Al-Fares, et al., ACM.SIGCOMM 38.4, 63-74 (2008).
[5] Spanke, IEEE J Quantum Electron 22.6, 964-967 (1986).
[6] Lin, et al., J. Light. Technol 40.17, 5842-5853 (2022).
[7] Ye T, et al., IEEE ACM Trans Netw 23.2, 491-504 (2014).
[8] Montazeri, et al., ACM.SIGOMM, 221-235 (2018).

Disclaimer: Preliminary paper, subject to publisher revision