Supporting Bandwidth Guarantee for a Fast Optical Switching Network with Micro Buffer Switching Fabric

Fulong Yan[1*], Chongjin Xie[2], and Nicola Calabretta[3]

1 Alibaba Cloud, Alibaba Group, Beijing, China
2 Alibaba Cloud, Alibaba Group, Sunnyvale, California 94085, USA
3 Eindhoven University of Technology, Eindhoven, the Netherlands
yanfulong.yfl@alibaba-inc.com

Abstract: We propose a micro buffer fast optical switch (MFOS) fabric for a data center network. MFOS highly improves the network performance, and achieves 6.7 μs latency and 99.9% throughput at a load of 0.8. © 2022 The Author(s)

1. Introduction

Data centers (DCs) witness the booming of various bandwidth thirsty applications including big data, artificial intelligence and cloud computing in the past decade. In the era of supporting predictable network service, emerging applications such as virtual reality, car networking and smart city require not only huge bandwidth, but also extreme low latency or even bounded latency [1]. To support the guarantee of quality of service (QoS), data center networks (DCNs) need to evolve from the current best effort paradigm to expected service paradigm.

Various optical switching based DCNs have been proposed to solve the high bandwidth and low latency challenges [2–5]. To achieve packet switching as an electrical switch performs, fast optical switch (FOS) based DCNs capable of switching on the order of nanoseconds must be considered. Unfortunately, the fabric of current FOSes without buffer normally adopts failing and re-transmission mechanism to address contentions, which unavoidably results in a low throughput as the network saturates [4], and does not provide QoS support [5].

Nakano et al demonstrated the implementation of large capacity compact optical buffer with coiled fiber delay lines (FDL). The fibers with total lengths of 1.2 km were coiled onto a single bobbin (40 mm in diameter and 20mm in height) [6]. Considering the limitation on the bufferless switching fabric and the ineffectiveness of the scheduling mechanism, novel FOS fabric with corresponding scheduling mechanism must be developed to support the differentiated QoS.

In this paper, we propose a micro buffer based FOS fabric (MFOS). MFOS adopts a flow fair queue (FFQ) scheduling algorithm to address the transmission of optical packets. The simulation results show that MFOS improves the performance of conventional FOS based DCN and achieves bandwidth guarantee. Moreover, theoretic analyses and simulations validate that the MFOS based DCNs support bandwidth guarantee and bounded latency.

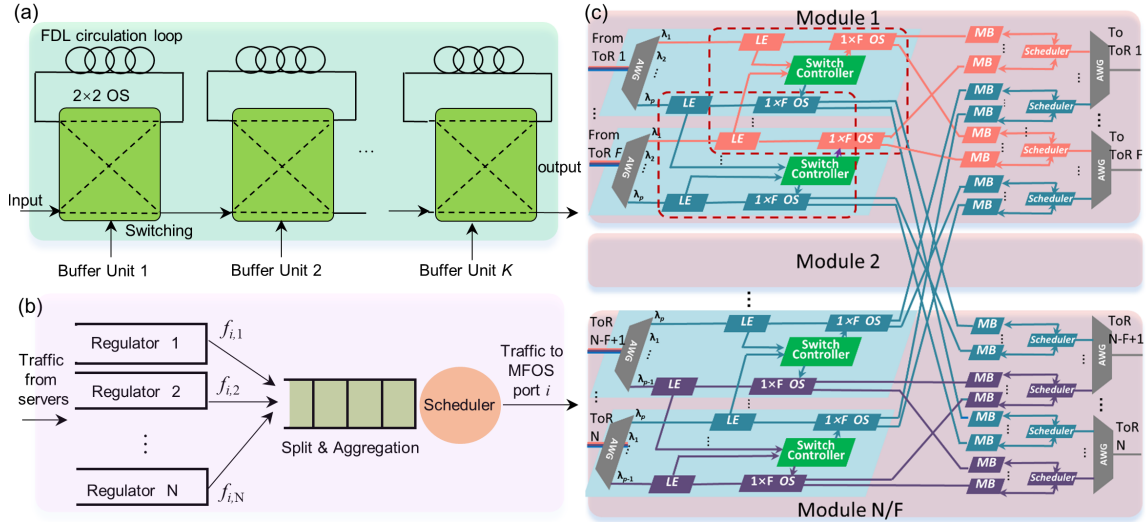

Fig. 1. The schematic blocks of (a) MB, (b) ToR traffic processing modules, (c) MFOS

2. Optical DCN based on MFOS

The schematic of the micro buffer (MB) is shown in Fig. 1 (a). The MB is built by cascading K feedback buffer units. The MB buffer unit operates by passing the fiber delay line (FDL) circulation loop repeatedly through a 2×2 optical switch (OS) with reconfiguration time of several nanoseconds [6]. The storage time of an optical packet depends on the counts of recirculations. To ensure the feasibility and practical use of the MB in the MFOS, K should be a small number.

Fig. 1 (b) shows the traffic processing modules of the Top-of-Rack switch (ToR) connected to the MFOS. The arriving traffic in i-th ToR is divided into N flows denoted by $f_{i,j}$ ($j = 1,...,N$) with average rate of $r_{i,j}$ where j is the index of the destination ToR. The electrical packets from serves are split into cells and multiple cells comprise an optical packet to be scheduled out.

The building blocks of the MFOS are presented in Fig. 1 (c). Different from the bufferless FOS [4], the MFOS contains N^2 MBs to store the optical packets contended for the same output port. The switch controller performs the switching of the 1×N OS based on the label from the label extractor (LE). The optical packet contentions are avoided by the MBs, which is handled by the scheduler. The scheduling of the MBs can adopt various output queued scheduling algorithms [7].

For the operation of the MFOS based DCN, the optical packets are first scheduled into the MBs from the input ports by the switch controller. At each output port, the scheduler chooses one optical packet from N MBs and send it out of the MFOS. The scheduling of the input ports and output ports are decoupled and therefore the scheduling process is simplified. By considering the status of arriving traffic matrix $\mathbf{I}$, MB buffer matrix $\mathbf{M}$ and output scheduling matrix $\mathbf{O}$ where the i-th row (j-th column) corresponds to the i-th input (j-th output) port, we present the operation procedure of MFOS. At the start of the MFOS operation, the status of the MB is a zero matrix. Therefore, in the first slot, $\mathbf{M}(1)$ is equal to $\mathbf{I}(1)$. $\mathbf{O}(t)$ is part of $\mathbf{M}(t)$ since the packets are scheduled from the MBs. The optical packets which are buffered at the MBs in the second slot are the sum of the residual optical packets in the MBs after scheduling of the first slot and the arriving optical packets in the second slot. More generally, we have the following Eq. (1) for the scheduling of the two continuous slots t and t+1.

$$\mathbf{M}(t+1) = \mathbf{I}(t+1) + \mathbf{M}(t) - \mathbf{O}(t) \tag{1}$$

3. Flow fair queue scheduling

To minimize the size of the MB in MFOS, the traffic scheduled into the MB should be scheduled out of the MB in the shortest time. Generalized processing sharing (GPS) scheduling is the optimum algorithm due to the fair sharing of bandwidth. In the operation of GPS, considering the i-th ToR, let $W_j(t_1,t_2)$ be the amount of bandwidth received by flow $f_{i,j}$ in the interval $[t_1,t_2]$, and the set of backlogged flows at time τ remains unchanged during any time interval $[t_1,t_2]$. Then

$$W_j(t_1,t_2) = \frac{r_{i,j}}{\sum_{j\in B(t_1)} r_{i,j}} \tag{2}$$

It is easy to find that $W_j(t_1,t_2) \geq r_{i,j}$ since $B(t_1)$ is a subset of the N flows at the i-th ToR. Therefore, $f_{i,j}$ is guaranteed a minimum service rate of $r_{i,j}$. Notice that the GPS assumes that all flows can be scheduled simultaneously in one slot under the condition that the flows are infinitely divisible. However, in a realistic DCN, for each switching port, only one flow can be scheduled in a time slot. Namely, the entire optical packet must be transmitted before another optical packet can be transmitted under FFQ. And FFQ is adopted to approach the GPS [7].

There is at most one packet service difference between the FFQ and the GPS [7]. The length of optical packet is denoted by L, and we take the maximum length of transmitted packets in the link between the ToR and the MFOS as L_d. Under the scenario where both the input and output ports adopt GPS scheduling, it has been shown in [8] that the packets scheduling difference between the input and output are bounded by $2L$ since the distance between the input and output schedulers is neglectable inside an electrical switch. While in the MFOS based DCN, the input scheduler locates inside the ToR resulting in a packet scheduling difference of $2L+L_d$ under GPS.

Considering the traffic scheduling differs L between GPS and FFQ in both input and output scheduling, we immediately know that the packets scheduling difference between the input FFQ and output FFQ are bounded by $4L+L_d$. The MB at most needs to store $4+L_d/L$ optical packets. Therefore, the MB size normalized by the optical packet length, namely the value of K, is bounded by $4+L_d/L$. As a deduction, K could be only 5 under FFQ scheduling as $L_d = L$.

4. Simulation setup and results discussion

We use OMNeT++ platform to carry out the simulations of the optical DCN supporting 256 servers equipped with 100Gb/s NIC. The MFOS(FOS) radix is set as 16 with 4 wavelengths ($N = 16$, $p = F = 4$). Each server generates 10^5 packets independently based on ON/OFF Pareto distribution model [4]. The intra-ToR traffic ratio σ is set as 0.5, while the other 50% inter-ToR traffic destines to servers in the rest 15 ToRs. The inter-ToR traffic distribution $d_{i,j}$ is defined in the Eq. (3). Both uniform and nonuniform inter-ToR traffic patterns with ω of 0 and 0.5, respectively, are considered. We have $r_{i,j} = d_{i,j} \times \rho$ for any flow $f_{i,j}$ where ρ is the network load.

$$d_{i,j} = \begin{cases} (1-\sigma)(\omega + (1-\omega)/(N-1)), & i = j-1 \\ \sigma, & i = j \\ (1-\sigma)(1-\omega)/(N-1), & \text{else} \end{cases} \tag{3}$$

Each ToR is equipped with 4 transceivers operating at 800 Gb/s. We take the optical packet size as 12288 Bytes comprising of 48 cells with a size of 256 Bytes. A guardband of 8.5 ns including switching and optical packet

preamble is considered to lead an optical packet time slot of 500 ns [5]. The fiber length between the ToR and the MFOS is 100 m, resulting in link delay of 500 ns. The total fiber length of a MB will be 500m buffering at most 5 optical packets, which can be coiled onto a single bobbin supporting fiber of 1.2km length as demonstrated in [6].

Firstly, we investigate and compare the performance of MFOS based DCN using the FFQ scheduling with FOS based DCN using the fail and transmission mechanism under both uniform and nonuniform inter-ToR traffic as described in Eq. (3). Figure 2 (a) shows the average latency and normalized throughput of the MFOS and FOS based DCNs. As the load is less than 0.4, the ToR to ToR latency of FOS based DCN is lower than that of the MFOS based DCN due to low contention and direct transmission of the arriving packets. However, as the load increases beyond 0.4, the increasing contentions of the FOS result in a large amount of re-transmissions, and the performance of the MFOS based DCN outperforms the FOS based DCN due to the eliminated retransmissions.

Secondly, we expand the relative load of flow $f_{1,3}$ from 1 to 10 in non-uniform traffic to simulate a flow that sends traffic exceeding its allocated bandwidth, and the network load is 0.5 when there is no expansion on $f_{1,3}$. Figure 2 (b) shows the average latency of $f_{1,3}$ and $f_{1,4}$. It is clearly shown that in MFOS DCN, the increased load of $f_{1,3}$ does not have influence on $f_{1,4}$, namely, the $f_{1,4}$ obtains guarantee bandwidth no matter the load of other flows. Therefore, the average latency of $f_{1,4}$ is almost a constant. While in FOS based DCN, the FOS could not differentiate flows, the excessive sending behaviour of $f_{1,3}$ could not be identified and deteriorates the performance of well behaved flows that send traffic based on the negotiated bandwidth.

We also investigate the maximum latency of the MFOS based DCN under both uniform and non-uniform traffic. To obtain bounded latency, we set the size of the leaky bucket as the length of 10 optical packets in the regulator shown in Fig. 1 (a). The allocated minimal bandwidth is 1/30 and 1/60, respectively, for all flows of uniform and non-uniform traffic from Eq. (3). Therefore, the theoretic latency upper bound is 225.5 and 450.5 μs, respectively for uniform and non-uniform traffic. As shown in Fig.2 (c), the simulated maximum latency is far less than theoretical bound.

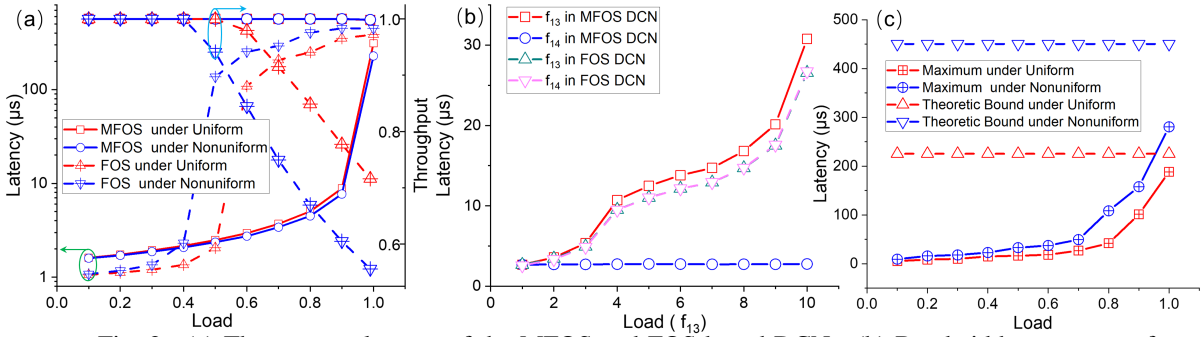

Fig. 2. (a) The average latency of the MFOS and FOS based DCNs, (b) Bandwidth guarantee of MFOS based DCN, (c) theoretic bounded latency and simulated maximal latency

5. Conclusions

We propose a novel FOS fabric adopting MB for an optical DCN and FFQ is adopted for MFOS to minimized the buffer size of MB. The simulation results show that the MFOS based DCN highly improves the network performance, and achieves 6.7 μs ToR to ToR latency, 99.9% throughput at a load of 0.8. Moreover, theoretic analyses and simulations validate that the MFOS based DCN achieves bandwidth guarantee and bounded latency.

Acknowledgements

The authors thank China Postdoctoral Science Foundation (2021M690179) and Beijing Postdoctoral research Foundation for supporting this work.

References

1. C. Mobile, "5G application scenarios white paper," 2019.
2. S. B. Yoo, "Prospects and challenges of photonic switching in data centers and computing systems," *Journal of Lightwave Technology*, vol. 40, no. 8, pp. 2214–2243, 2022.
3. W. M. Mellette *et al.*, "Rotornet: A scalable low-complexity optical data center network," Sigcomm 2017, pp. 267–280.
4. F. Yan, W. Miao, R. Oded, and C. Nicola, "OPSquare: A flat dcn architecture based on flow-controlled optical packet switches," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 4, pp. 291–303, 2017.
5. H. Ballani *et al.*, "Sirius: A flat datacenter network with nanosecond optical switching," Sigcomm 2020, pp. 782–797.
6. T. Tanemura, I. M. Soganci *et al.*, "Large-capacity compact optical buffer based on inp integrated phased-array switch and coiled fiber delay lines," *Journal of Lightwave Technology*, vol. 29, no. 4, pp. 396–402, 2010.
7. J. C. Bennett and H. Zhang, "WF^2Q: worst-case fair weighted fair queueing," in *Proceedings of IEEE INFOCOM'96. Conference on Computer Communications*, vol. 1. IEEE, 1996, pp. 120–128.
8. H. Jin, D. Pan, N. Pissinou, and K. Makki, "Achieving flow level constant performance guarantees for cicq switches without speedup," in *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*. IEEE, 2010, pp. 1–5.