PROPH: A Contention-free and Elastic Bandwidth Scheduling Scheme for AWGR-based Optical DCN

Xinwei Zhang, Xuwei Xue, Bingli Guo, Xiaoyue Su, Yisong Zhao, Yuanzhi Guo and Shanguo Huang
State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, China
Author e-mail address: x.xue@bupt.edu.cn

Abstract: A contention-free scheduling scheme PROPH is proposed for AWGR-based optical DCN to elastically allocate adaptable bandwidth. Assessments validate that the network deploying PROPH decreases 73.6% packet loss and 11.9% server-to-server latency, improves 11.6% network throughput. ©2022 The Author(s)

1. Introduction

With the rapid increase of applications traffic in data center networks (DCNs), electronic switch based DCNs encounter bottleneck in terms of both bandwidth and power consumption. The emergence of optical switching technology overcomes these shortcomings with benefits of data rate and format transparency [1]. In particular, Arrayed waveguide grating router (AWGR), a passive optical component, which can work as the fast optical switch with wavelengths from a given input port are routed cyclically across the output ports. Various AWGR based DCNs have been extensively investigated. However, to avoid the packet contention shown in Fig. 1(a), where one destination port has two or more sources sending traffic to it at the same timeslot, static bandwidth scheduling schemes are utilized in these proposed structures [2, 3]. These static schemes usually deploy a Round Robin style to arrange the forwarding timeslots and wavelength, which cannot flexibly satisfy the real-time forwarding requirements of heterogeneous traffic. For instance, Sirus in [2] repeats every "epoch" comprising G timeslots. H-lion in [3] is a fully connected structure, where the transmitters send the first coming flow. In case of contention, the solution is to retransmit the later data packets. This will heavily increase the network latency when large flows need to be retransmitted.

In this paper, we propose PROPH, a contention resolution and elastic bandwidth scheduling scheme, to both prevent the packet contention and elastically provide the adaptable optical bandwidth to the various traffic volume. Based on the real-time monitored traffic volume at each buffer, the traffic at the most-occupied buffer will be transmitted under the premise of no packet contention. Moreover, to avoid the "Request-Response" scheduling style caused extra waiting delay, PROPH use a "Dislocation Pipeline" method to allocate the bandwidths of the N+1th time slot at the Nth time slot. Numerical investigations show that the PROPH strategy, compared with static scheme, decrease 73.6% packet loss, 11.9% server-to-server latency and improve 11.6% network throughput at the traffic load of 1.0.

2. Principle of PROPH

We proposed an AWGR based optical DCN in [4]. This network is divided into individual intra- and inter-cluster interconnections. Due to the intra-cluster and the inter-cluster interconnect network are two same independent sub-networks, here we explicitly introduce the intra-cluster network as shown in Fig. 1(b). The top of racks (ToRs) in each cluster are connected by a group of AWGRs. The corresponding scheduler for each AWGR is connected each ToR through optical links. Fig.1(c) illustrates the structure of the ToR. Each ToR has various buffer blocks storing

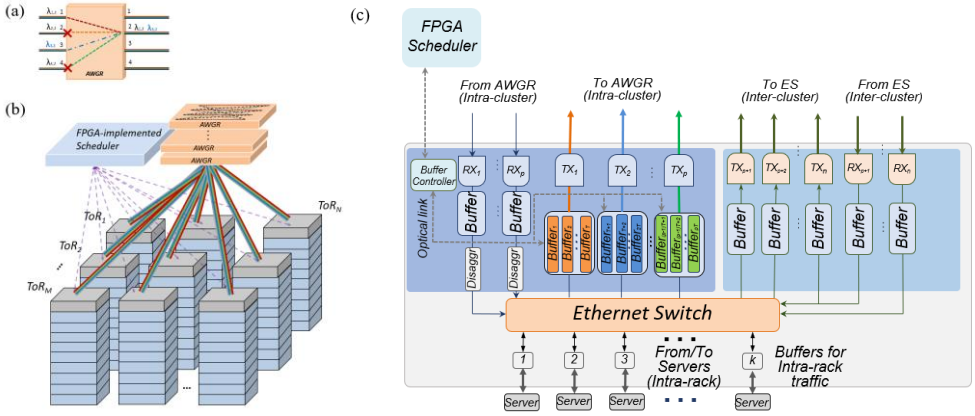

Fig. 1: (a) Case of packet contention (b) Interconnections of a cluster in AWGR-based optical DCN (c) Functional blocks of ToR

data packets destined to different destinations. The buffers are divided into groups and each group sends flows through the same Tx. The advantage is that we can use less wavelengths to connect all the ends in large scale networks. However, the flows in the source end need to be scheduled since each transceiver in the ToR can send and receive only one flow at the same time. We spilt time into timeslots to schedule flows to guarantee that all the ends can communicate with others without contention. In each time slot the buffer controller at ToR will send the traffic information including the source and destination of the flows and the buffers ratio to the FPGA based scheduler through the optical link. The scheduler then will choose flows according to the presented strategy and send scheduling instructions to the buffer controller used to decide which buffers can release packets through Txs.

We propose a "Dislocation Pipeline" (DisP) strategy that can schedule the flows in real time and avoid the scheduling "Request-Response" style caused extra waiting delay. As demonstrated in Fig. 2, the DisP scheduling process has two steps. Firstly, the buffer controller sends the traffic volume of each buffer blocks at the N^{th} time slot to the scheduler to request corresponding wavelength. Based on the requests, the scheduler will select adaptable wavelength for each most-occupied buffer blocks of the $N+1^{th}$ time slot under the premise of no wavelength contention. At the start of the $N+1^{th}$ time slot the buffer controller will send flows according the received schedule instructions. The instance illustrated in Fig.2 shows the schedule process for 3 traffic flows at 3 buffer blocks. t1 and t3 are the beginning of time slot N^{th} and time slot $N+1^{th}$, and t2 is the start moment to schedule flows for time slot $N+1^{th}$. At t1 the flow3 will be sent because it has most traffic. The scheduling process for next timeslot starts at t2 and because this process is so fast that we can regard the traffic condition at t2 as that at t3. Then the collected information at t2 can be used to calculate the scheduling rules for time slot $N+1^{th}$ and will choose flow2 as next sent flow. This means the DisP scheduling strategy only needs a little time before the end of one timeslot to schedule flows for next timeslot which guarantees the real time configuration capability.

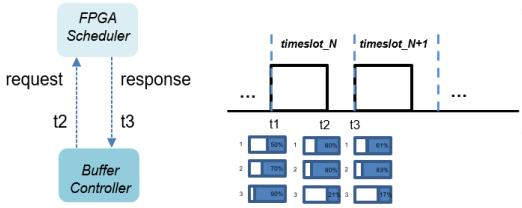

Fig. 2: Principle of "Dislocation Pipeline" strategy

The algorithm in Table 1. illustrates how the scheduling strategy works at the FPGA based scheduler. In the algorithm, n and m mean the number of ToR in the cluster and the number of Tx (Rx) in one ToR respectively. Since a transceiver of the ToR can only send or receive one wavelength at the same time, only when Tx used to send a flow and Rx used to receive the flow are both idle that can the flow to be sent. There are n ToRs in one cluster and each ToR has m Txs so that one cluster can transmit m×n flows at the same time.

Algorithm1

Input:

Traffic information of all the flows in the cluster: flowlist=list[$flow_i$] i=1,2,..n^2;

Output:

flows that can be sent without contention: res=list[$flow_k$] k=1,2,…m⋆n;

1: initialize the lists that record state of the Tx and Rx;
   map_j={1:0, 2:0, …, m:0},map_t={1:0, 2:0,…, m:0}(*0 means idle and 1 means busy*); j,t =1,2,..n;
2: sort flowlist as Q according to each flows' traffic volume;
3: **while**(res.size()<m×n)
4: flow=Q.pop();
5: s=source of flow;
6: **if** Tx and Rx flow uses is idle
7: res.push(flow);
8: Set map_s[Tx and Rx flow uses]=1;
9: **end if**
10: **end while**

Table 1. the scheduling algorithm of PROPH

3. Numerical investigations and discussions

We numerically investigate the PROPH scheme in a 16-cluster DCN in which each cluster has 16 ToRs and each ToR connects 40 servers based on OMNET++ platform. The traffic in DCN is usually not uniformly distributed, the traffic ratio in the rack, intra-cluster and inter-cluster are thus set to 50%, 37.5% and 12.5%, respectively. The link rate of the servers and Txs are 20Gbps and 100Gbps, and buffer size in ToRs is set as 20KB. The simulation compares the static cyclic sending strategy and the presented elastic PROPH schedule scheme on the primes that the simulation runs for equal time. We simulated the results with 1us and 2us timeslot length respectively.

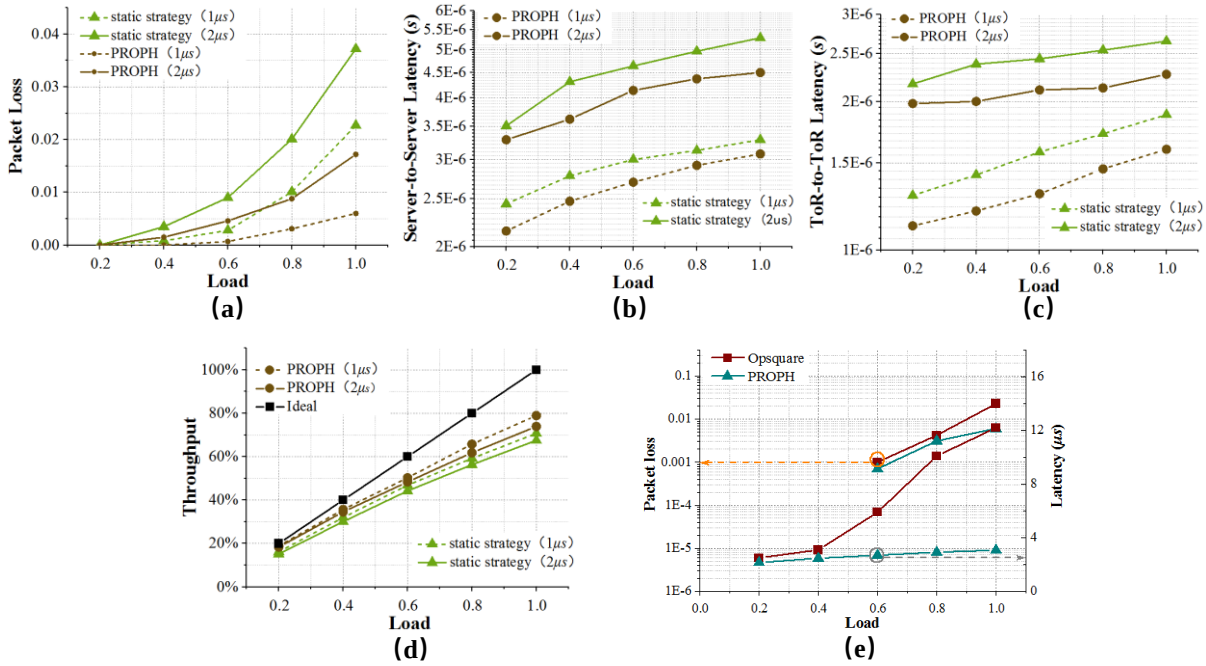

Fig. 3: Numerical results of (a) packet loss, (b) ToR-to-ToR latency, (c) Server-to-Server latency, (d) network throughput. (e) Network performance compared with Opsquare structure.

Fig. 3 shows the numerical results of the static cyclical strategy and the present PROPH strategy. Fig. 3(a) illustrates that the packet loss can be reduced at most 73.6% under the elastic strategy. As the load is higher, all the sources will send more traffic and the buffer will be filled in faster so that the packet loss increases. The traffic sending to different destinations are different from each other at most time, but the static strategy sends all the flows for equal time so that the flows that has more traffic will lose packets. While the dynamic PROPH strategy ensures that most traffic can be scheduled at each timeslot so that the packet loss is much lower. In the 1us length timeslot situation the packet loss is less than the 2us timeslot since the packets can be scheduled faster. Large traffic can be sent in time so their time for queuing is shorter so that the latency is lower. Fig. 3(b) and Fig. 3(c) show the server-to-server latency and ToR to ToR latency are reduced by 11.9% and 14.7% at most respectively and in the shorter timeslot situation the latency is lower because of the faster scheduling. The throughput of the network can be improved due to less packet loss. As is shown in Fig. 3(d), the throughput increases by 11.6 % at most compared with the static strategy. We also compared the performance of the presented strategy with Opsqsuare, which is a flat structure of optical DCN based on Semiconductor Optical Amplifier gates [5]. The comparison of simulation results is illustrated in Fig. 3(e). The end-to-end latency is higher due to the packet retransmission scheme for the conflicted packets in Opsquare structure for contention resolution, while the presented PROPH strategy without retransmission scheme reduces the 74.8% latency and 73% packet loss when the load is 1.0.

4. Conclusion

To avoid wavelength contention in AWGR based optical DCN and to fully utilize the optical bandwidth, we proposed an elastic bandwidth schedule scheme PROPH to allocate adaptable wavelength to traffic flows based on the real-time monitored traffic volume. Numerical results illustrate that compared with existing static bandwidth scheduling schemes, the proposed PROPH reduces 73.6% packet loss, 11.9% server to server latency and 14.7% ToR to ToR latency, improves 11.6% throughput. Moreover, compared with the OPSquare structure with flow control protocol, the proposed PROPH improves the network performance by removing the retransmission packets.

3. References

[1] Xue, X., Calabretta, N. Nanosecond optical switching and control system for data center networks. Nat Commun 13, 2257 (2022).
[2] Ballani, Hitesh, et al. "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching." SIGCOMM '20, pp. 782-797. 2020.
[3] R. Proietti, Z. Cao, C. J. Nitta, Y. Li and S. J. B. Yoo, "A Scalable, Low-Latency, High-Throughput, Optical Interconnect Architecture Based on Arrayed Waveguide Grating Routers," in Journal of Lightwave Technology, vol. 33, no. 4, pp. 911-920, 15 Feb.15, 2015.
[4] Z. Zhao, B. Guo, S. Huang, Y. Zhao, Y. Guo and X. Xue, "FSS: A Fast Switch System Based on AWGR for Optical Datacenter Network," 2021 19th International Conference on Optical Communications and Networks (ICOCN), 2021, pp. 1-3.
[5] F. Yan, W. Miao, O. Raz and N. Calabretta, "Opsquare: A flat DCN architecture based on flow-controlled optical packet switches," in Journal of Optical Communications and Networking, vol. 9, no. 4, pp. 291-303, April 2017.