Buffer Architecture Study for Effective Service of Time-Sensitive Traffic in Intra-rack Optical Data Center Networks

David Georgantas
Department of Informatics and Telecommunications
University of Thessaly
Lamia, Greece
dageorgantas@uth.gr

Peristera Baziana
Department of Informatics and Telecommunications
University of Thessaly
Lamia, Greece
baziana@uth.gr

***Abstract*—This paper presents a software defined (SD) passive optical intra-rack Data Center Network (DCN) architecture and studies the effect of buffer size and cycle duration variation on the DCN performance. The purpose of this study is to optimize the performance of the intra-rack DCN, especially in terms of average packet delay. After defining the buffer sizes and cycle duration based on the variation study, we present the results of several performance metrics. Specifically, simulation results show that our proposal reaches 90% bandwidth utilization, while the average packet delay remains in a few μs, with the largest increase occurring under the highest possible load.**

***Index Terms*—Optical Networks, DCN, SDN**

I. Introduction

In recent years the increase in traffic in data center networks (DCNs) [1], [2] has put pressure on the organizations that manage data centers (DCs) and the need for the DCN infrastructure to be modernized. Optical technologies in recent years have been the focus of research for DCNs, due to the high transmission rate they offer, and the lower power consumption compared to electrical connection equipment.

Regarding optical switching, many studies of hybrid (electrical/optical) DCNs as well as purely optical ones have been published [3], [4]. In hybrid studies the optical switching is done through the optical circuit switching (OCS). OCSes create a dedicated path between the source and destination node, which requires high reconfiguration time. For this reason, in hybrid DCNs, OCS has been mainly used to transmit large-sized flows and not short and bursty kind of traffic [5].

The burstiness that occurs in DCN traffic characteristics [6], [7] makes OCS unlikely to be a universal solution for switching in DCNs. For this reason, optical packet switching, which can support the transmission of packets of various sizes, has been studied in recent years as a solution to the above problem, with a variety of studies being published [8]–[11].

The preceding research mostly concentrate on inter-rack switching, i.e., inter-rack communication. However, intra-rack communication is of particular relevance since racks service a significant portion of DC traffic, particularly in cloud DCs. Furthermore, the vast majority of intra-rack networks in typical DCNs continue to operate in the electrical domain, resulting in unnecessary energy consumption [12]. Furthermore, the rack is the first point of contact between computing resources in a DC and understanding and optimizing its internal communication is critical.

In this direction, studies such as [13], [14], use passive optical interconnect components to increase the capacity of the intra-rack DCN and at the same time reduce the energy consumption compared to an electrical intra-rack DCN. The specific proposals are based on software defined (SD) principles for determining the transmissions in the intra-rack network with the central controller located in the Top-of-Rack (ToR). Although the above studies have contributed to the research on the application of optical technologies in intra-rack DCNs, they do not contain a study on the size of the buffers at each rack node (buffers that store traffic until its transmission) and the cycle duration of the intra-rack DCN simulation.

The size of the buffers on each node of the rack is a finite number, and considering it to have an infinite capacity can lead to final simulation results that do not correspond to reality. Also, buffer capacity can significantly affect network performance metrics such as dropping probability (probability of a traffic flow to get dropped from the network due to lack of available memory), which by extension may affect other performance metrics. Regarding the cycle duration of the simulation, it should be mentioned that this is also an important parameter for the optimization of the intra-rack DCN for two reasons. On the one hand because the cycle duration of the simulation reflects the clock synchronization in real intra-rack DCNs. On the other hand, because cycle duration affects very important network performance metrics, such as average packet delay.

In this paper, we propose a passive optical intra-rack DCN alongside with a buffer size and cycle duration variation study for performance optimization. We use the SD technique for a high-capacity wavelength division multiplexing (WDM) passive coupler-based intra-rack network architecture with four WDM data channels and an extra synchronized WDM control channel for synchronous transmissions coordination. We also propose a resource allocation strategy implemented centralized

by a field programmable gate array (FPGA)-based controller located at the top-of-rack (ToR) infrastructure.

We evaluate our proposal through a discrete-time simulation model after we define the cycle duration and the buffer size via the variation study. Under the maximum offered load, the simulation results show that the network model we propose achieves a packet delay of just 59 μs and a bandwidth utilization of 90%.

Our paper's remaining content is categorized as follows: Section II describes the proposed SD intra-rack DCN architecture and the resource allocation strategy. Section III provides the buffer size and cycle duration variation study and the evaluation of our proposal. Section IV brings our efforts to a close.

II. Software Defined Intra-rack Architecture

The network architecture shown in Fig. 1 concerns for the communication within a rack that connects N servers. In this setup, the servers are interconnected using a passive optical coupler. The optical fiber supports five WDM wavelengths, namely λ_0, λ_1, λ_2, λ_3, and λ_4, while L is the length of fiber connections among the servers and the passive coupler. These wavelengths serve as the shared communication channels for intra-rack networking. Among these wavelengths, λ_0 is exclusively allocated for control communication, functioning as the control channel. The remaining four wavelengths, λ_1, λ_2, λ_3, and λ_4 are considered as data channels and are utilized for data communication purposes. Moreover, the Top-of-Rack (ToR) switch network interface employs a pair of a fixed transmitter/receiver tuned in λ_0 providing connectivity between the FPGA controller and the passive coupler.

Five sets of fixed wavelength burst mode transmitters (BMTs) and receivers are included in the network interface of each server, as it is presented in Fig. 2. We use semiconductor optical amplifiers (SOAs) to quickly turn on and off the burst mode lasers (with a switching time of 900 ps) [15], [16].

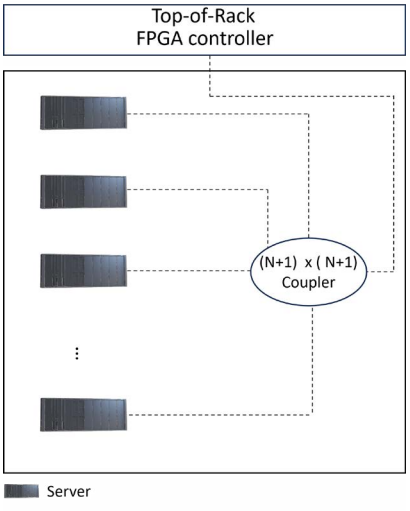

Fig. 1. Intra-rack DCN architecture.

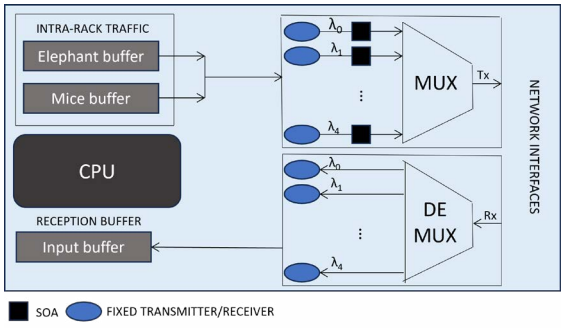

Fig. 2. Server architecture.

Moreover, each server has three buffers: one for receiving data and two for outgoing data. These buffers are used for the data flows to be stored. We employ two distinct electrical buffers for outgoing traffic—one to handle elephant and the other to handle mice flows. This method is intended to account for the significant diversity in traffic characteristics encountered in modern DCs. Additionally, each server uses a central processing (CPU) unit that supports server communication with the ToR, for the SDN control needs presented below.

The suggested architecture for the intra-rack DCN follows SDN principles. This three-layer model shown in Fig. 3 describes the operation of the intra-rack SDN. The cloud services that produce the traffic that the DCN must handle are included in the top layer, which is also referred to as the application layer. The SDN controller (FPGA controller) is part of the middle layer, also known as the control layer, which manages network traffic inside the rack. The FPGA controller specifically manages the coordination of data transmission through the use of Southbound Interfaces and the resource allocation strategy, while Northbound Interfaces facilitate service requests from cloud applications.

Rack servers and the passive coupler make up the lower layer of the architecture. Within this layer, servers send messages known as demand messages across the control channel. These messages are sent in each cycle in appropriate time slots and are used to seek access permissions to the

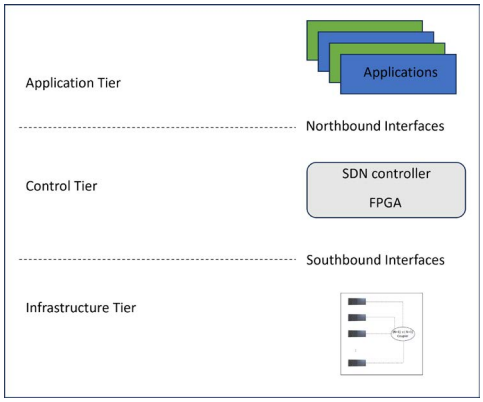

Fig. 3. SDN layers.

intra-rack data channels. The FPGA controller collects demand messages from all servers during each cycle, then during the same cycle it executes the suggested resource allocation strategy to assign certain time slots (each one is a subset of the cycle) for data transmissions. After the execution of the resource allocation strategy, the FPGA controller informs servers about their access rights in the next cycle with grant messages, over λ_0.

The suggested resource allocation strategy is designed to provide priority to the transmission of traffic flows by taking into account how long they have been waiting in the output transmission buffers. The key objectives are to reduce average packet delay, effectively utilize the available bandwidth, and avoid collisions among data transmissions. Our resource allocation strategy necessitates that each traffic flow be transmitted on a certain data channel at a specific time slot (subset of the cycle). Additionally, depending on prior allocations and the amount of bandwidth that is available, our technique chooses which channel to utilize for each transmission.

Consider a scenario with five traffic flows awaiting transmission. In this context, the FPGA controller has already assigned portions (according to the resource allocation strategy) of each data channel's bandwidth to the first four flows—36%, 27%, 42%, and 51%, respectively. Now, if the next flow in line requires 69% of the cycle duration for its transmission, the resource allocation strategy will allocate bandwidth for this flow in the fourth channel. This choice is made to enhance bandwidth utilization, as this channel is currently the most heavily used compared to the others. In a scenario where the available bandwidth on a data channel equals or exceeds 69%, the FPGA controller will allocate the entire available bandwidth of that channel to ensure the full transmission of this flow within a single cycle. This approach avoids splitting the flow across multiple cycles, which could adversely affect average packet delay. Furthermore, in a scenario that there is enough bandwidth for a flow to be fully transmitted in more than one of the available channels, the suggested resource allocation strategy successfully balances bandwidth utilization among data communication channels. In this instance, our approach distributes bandwidth to the less-used channel on purpose. This method enables a more equitable distribution of resources, minimizing channel overload and encouraging a smoother and fairer allocation of bandwidth across all channels.

Based on the aforementioned scenarios for the bandwidth allocation by the resource allocation strategy, it is clear that the operation of our proposed network model is highly related on the cycle duration. In particular, the operation of the resource allocation strategy requires time synchronization between the servers in the rack and the FPGA controller, while the cycle duration must exceed the overall time for the exchange of the control messages (demand and grant messages) and the execution of the resource allocation strategy. Moreover, one more parameter that probably has an effect on the performance of the network model we propose is the buffer sizes of each server for the storage of the outgoing flows. The capacity of buffers, their congestion levels, and the loss of traffic load from the network due to lack of memory can significantly affect several network performance metrics. Therefore, we consider that it is necessary to provide a variation study of cycle duration and buffer sizes in Section III, so as to estimate the effect these two parameters have on the performance of the network model and to define the best performance parameter values for our proposal.

III. PERFORMANCE EVALUATION

We develop a simulation model utilizing a Python-based programming environment to test the performance of the suggested optical intra-rack DCN. Using discrete event simulation methods, we model both the traffic generated by each server and the servers' transmission and reception in the intra-rack DCN. Via simulation, we estimate the average throughput, average dropping rate, dropping probability and average packet delay under various levels of load.

For the generated traffic of data flows, the inter-arrival periods follow a negative exponential distribution. The generated traffic consists of two equal parts (each of 50% of the total traffic): the stable and the bursty traffic part, with both parts consist of 95% small (mice) flows and 5% large (elephant) flows. Mice flows range in size from 1 KB to 10 KB, with a mean of 5.5 KB. The elephant flows vary in size from 100 KB to 10 MB, with a mean value equal to 5.05 MB. Furthermore, The size of the bursty traffic flows follows uniform distribution making the bursty flows less predictable than stable flows, which follow Poisson distribution.

For the performance evaluation of the proposed network model, we consider that N=20 servers within the rack and L=10 meters distance of each fiber connection among the servers and the passive optical coupler, which corresponds to 50 ns propagation delay. The transmission rate of each BMT transmitter equals to 100 Gbps, resulting in 400 Gbps nominal load, since data communication is served over four channels. Moreover, the capacity of mice and elephant flows buffers and the cycle duration will be determined by the buffer size and cycle duration variation study below, in order a full set of performance parameter values to be specified.

A. Buffer sizes and cycle duration variation study

Given the above performance parameters, we provide a study regarding the influence of cycle duration and buffer sizes variation on network performance and the reason why we choose the specific values for each one of these two parameters, in order to specify a full performance parameters set for the evaluation of our proposal. Figures 4(a), 4(b), 4(c) and 5 present the variation study. The labels for each individual bar in each barplot highlight the sizes of the buffers. Specifically, within the parentheses, the first position shows the size of the buffer for mice flows, while the second position shows the buffer size for elephant flows. Additionally, in the labels of each bar in Figures 4(a) and 4(b), the dropping probability of each case is presented next to the sizes of

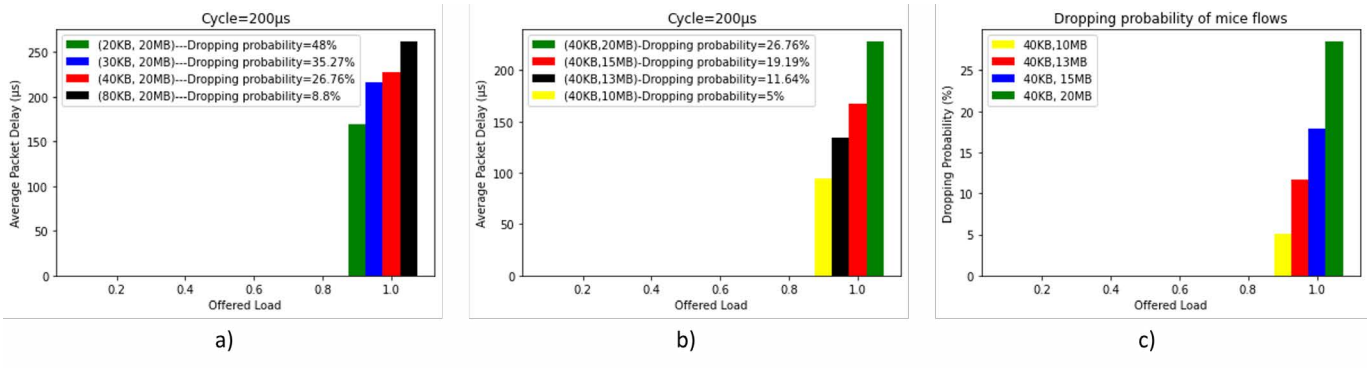

Fig. 4. a) Mice flows buffer size study with the size of elephant flows buffer fixed. b) Elephant flows buffer study with the size of mice flows fixed. c) Dropping probability of mice flows as the elephant buffer size decreases. Packet size=1500 B

buffers, due to the lack of available space for the presentation of more graphs. Moreover, in Fig. 5, next to the sizes of the buffers, the duration of the cycle in μs is also shown. In the first phase, we study the buffer size of mice flows, keeping the buffer size of elephant fixed. For these measurements, we consider a cycle equal to 200μs to achieve a reduced simulation time, leaving the study of the cycle duration for later.

For our network model testing for a variety of mice flows buffer sizes, we give the following values to the buffer of mice flows: 20 KB, 30 KB, 40 KB, 80 KB, while the buffer of elephant has a fixed capacity equal to 20 MB. Moreover, we simulate the intra-rack DCN under 400 Gbps load, in order to test network's performance under the nominal load. However, it makes sense that as the mice buffer size increases, more and more flows remain in the network without being dropped. Fig. 4(a) presents the average packet delay versus the normalized offered load. In Fig. 4(a), it can be seen that as the buffer size increases, the average packet delay also increases, while at the same time the dropping probability decreases. The proposed resource allocation strategy serves flows in a time priority order, which makes the extra flows that remain in the network (their number increases every time the buffer size of mice flows increases) to negatively affect the average packet delay. Based on Fig. 4(a) we choose the 40KB size

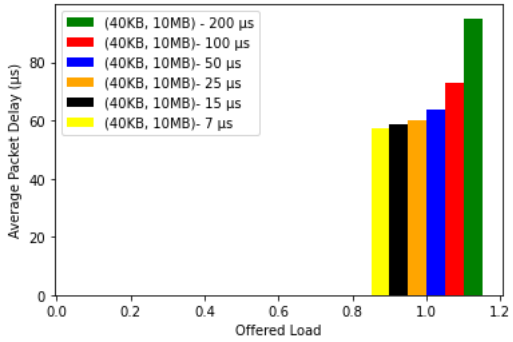

Fig. 5. Cycle duration study based on average packet delay measurements. Packet size=1500 B

buffer, since the overall dropping probability of 48% and 35% was not considered acceptable, despite the less average packet delay achieved. Furthermore, when the mice buffer size equals to 80KB, the average packet delay is almost 200 μs, which is not a performance that satisfies the service of time sensitive traffic.

The next step of the study concerns the buffer size of elephant flows and we will work similarly to the previous case. Keeping the mice flows buffer size fixed at 40 KB, we give the following values to the elephant buffer: 10 MB, 13 MB, 15 MB, 20 MB. In Fig. 4(b), it can be seen that as the buffer size of elephant decreases, the average packet delay decreases. This trend is common with the one shown in Fig. 4(a). However, a noteable difference in network performance between the first and second case is the opposite trend you see in the dropping probability. Specifically, in the first case, as the size of the buffer of mice flows decreases, the overall dropping probability increases simultaneously. On the contrary, in the second case, as the size of the buffer of elephant decreases, a decrease in the overall dropping probability is observed. On the one hand, the reduction in the size of the elephant buffer caused an increase in the dropping probability in elephant flows. On the other hand, since elephant flows constitute only 5% of the overall network flows, an increase in elephant flow drops would not significantly affect the overall dropping probability. Therefore, the reduction of the overall dropping probability, also means a reduction of the dropping probability of mice flows as shown in Fig. 4(c). At the same time the average packet delay is reduced, due to the decongestion resulting from the drops of elephant that would occupy a large part of the bandwidth, thus creating congestion in the buffers of mice flows, since their transmission would be delayed due to the lack of available bandwidth for transmissions. For these reasons, we choose a value of 10 MB for the buffer size of elephant flows and ended up with each server having a buffer size equal to 40KB for mice flows and a buffer size that equals to 10MB for elephant flows.

In the next phase, after finalizing the sizes of the buffers for each type of flow, we study the duration of the simulation cycle. For the needs of this study, we give the cycle duration

the following values: 7μs, 15μs, 25μs, 50μs 100μs, 200μs. We also simulate our network model under 400 Gbps which equals the nominal load, in order to test the network under the max possible congestion. Fig. 5 shows the average packet delay results for the various cycle duration values mentioned above. It is observed that as the cycle duration decreases, the average packet delay decreases.

This trend can be considered to be expected, since the reduction of the cycle duration also results in a reduction of the average queueing delay. For example, a longer cycle duration would result in a longer wait in the buffers for flows that arrived at the beginning of the cycle, until the start of their transmission in a next cycle. If we observe the changes in the average packet delay from the longest cycle duration to the shortest, it seems that the percentage of reduction in the average packet delay becomes smaller. Specifically, the change from cycle duration=200 μs to 100μs, equals 23.38%. The change from 100μs cycle duration to 50μs equals 12.39%, while the change from 50μs to 25μs is 5.8%. Additionally, the changes from 25μs cycle duration to 15μs and 15μs to 7μs are very small, close to 2%. For this reason, we consider that from 25μs cycle duration and less, the gain is minimal. In this way, we avoid the need for a more complex and expensive mechanism needed to implement such a fast clock synchronization and based on the above results, we decided to set the cycle duration equal to 25 μs.

B. Network model evaluation with a fully specified set of performance parameter values

With the values obtained from the buffer sizes and cycle duration variation study, the full set of the performance evaluation parameters is defined as follows : N=20, L=10, BMT transmission rate=100Gbps, mice buffer size=40KB, elephant buffer size=10MB, cycle duration=25μs. We simulate the proposed network model with the above set of performance parameters to infer network performance and present metrics for: average throughput, average packet delay, dropping probability and average dropping rate. In Fig. 6 the average throughput results are presented, as a function of the normalized load. As it is observed, the average throughput slope is approximately constant since load exceed 0.8, which means that the intra-rack DCN is strongly congested for higher loads than 0.8. However, if we notice the graph carefully, a tiny decrease in average throughput slope is also seen when the load exceeds 0.6. At the highest load, average throughput is equal to 357 Gbps, which means that 90% of the offered network load is transmitted successfully. Fig. 7 presents the average packet delay versus the normalized load measurements. In Fig. 7 we can notice that the average packet delay remains at the order of μs, equal to 59 μs.

For example, when the load=0.6 the average packet delay is equal to 12.49 μs, while when the load=0.2 the average packet delay is equal to 3.69 μs. Furthermore, the average packet delay appears to exhibit an exponential behavior, with the largest increase occurring when the load exceeds 0.8 of the nominal load. Another important observation from Fig. 6 is that the average packet delay of the large and mice flows are approximately equal. The measurement concerns the average delay of the packet and not of the flow, so it is considered a reasonable result. The equality between the average packet delay of mice flows and the average packet delay of elephant leads to the conclusion that the resource allocation strategy we propose is fair among the types of flows. This fact is legitimate, since we had designed the resource allocation strategy to give priority based on the arrival time of the flow and not to one of the two types of flows.

Fig. 8 presents dropping probability versus the normalized load of our proposal with the green dotted line. Also, when the load equals 1.0, dropping probability is about 5%, which means that about 5% of the flows are lost from the network during the simulation. The dropping probability of small flows is shown by the red dotted line and is approximately the same as the line of overall dropping probability. This is because the mice flows make up 95% of the overall traffic, which makes the overall dropping probability strongly dependent on the dropping probability of the mice flows. However, it appears that the overall dropping probability is slightly higher than that of mice flows, which is due to the higher dropping probability of elephant flows. However, since elephant flows make up 5% of the overall traffic, the dropping probability of elephant flows

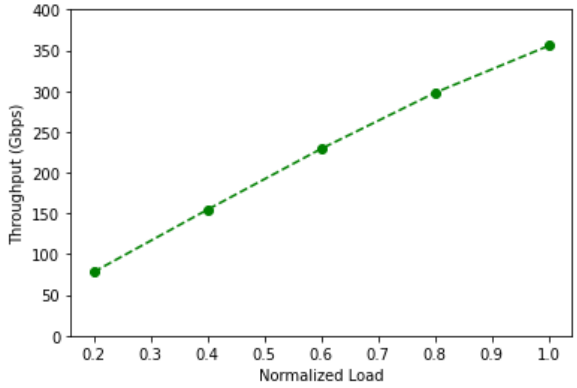

Fig. 6. Average throughput versus normalized load.

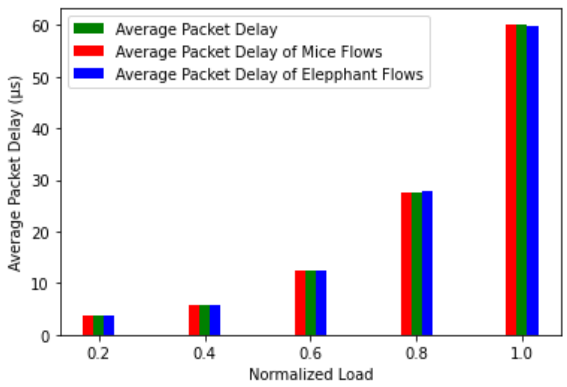

Fig. 7. Average packet delay versus normalized load. Packet size=1500 B

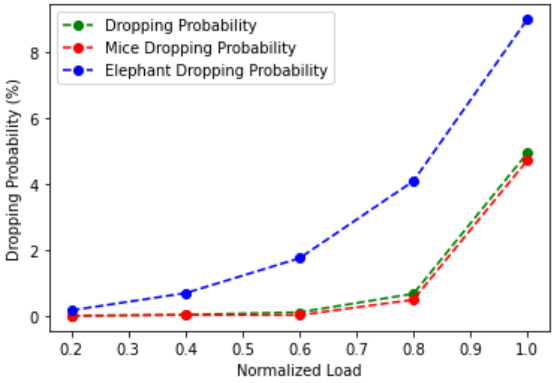

Fig. 8. Dropping probability versus normalized load.

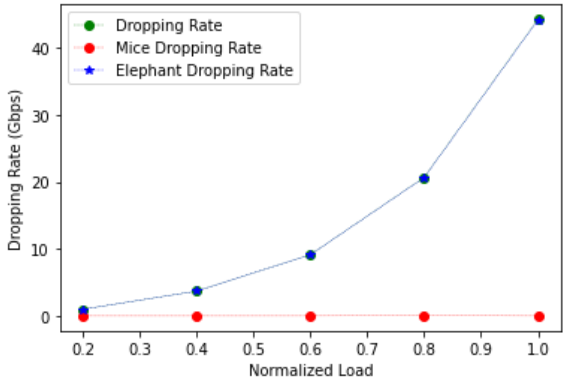

Fig. 9. Dropping rate versus normalized load.

has little effect on the overall dropping probability.

In Fig. 9, the dropping rate versus the normalized load of our proposal is presented. Line colors represent the same types of flows as in Fig 7. It is obvious that the amount of traffic in Gbps lost from the dropping of mice flows is very small. In contrast, the dropping rate in elephant flows has a large, almost universal effect on the overall dropping rate. As it is depicted in Fig. 8, the overall dropping rate is identical to that of the elephant and at the highest load it reaches about 43 Gbps. We consider that the very small dropping rate of mice flows is an important achievement, since they also contain control messages that are crucial for the intra-rack DCN functioning. However, the high dropping rate resulting from elephant flows needs further optimization.

IV. Conclusion

Our research focuses on the study of a passive optical intra-rack DCN. Our proposal works with the principles of SDN, using a resource allocation strategy to coordinate transmissions in the network. To finalize our proposal, we present a study on the size of the buffers and the duration of the cycle, factors which can significantly affect the performance of the network but also raise the cost of the infrastructure. From the simulation results, it appears that our proposal achieves about 90% bandwidth utilization, while the average packet delay remains in the μs scale, with the largest increase occurring at the highest load. However, the study of cycle duration and buffer sizes is strongly dependent on the type of traffic, the average value of the inter-arrival times between flow arrivals, and the average values of flow sizes. This means that a traffic with different characteristics would probably lead us to different results and decisions.

References

[1] Inc. Cisco Systems, "Cisco Global Cloud Index: Forecast and Methodology, 2016–2021," White Paper, 2018.

[2] "Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper - Cisco." https://www.cisco.com/c/en/us/solutions/collateral/executiveperspectives/annual-internet-report/white-paper-c11-741490.html (accessed Mar. 23, 2023)."

[3] N. Farrington et al., "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," 2010.

[4] G. Wang et al., "c-Through: Part-time optics in data centers," Computer Communication Review, vol. 40, no. 4, pp. 327–338, 2010, doi: 10.1145/1851275.1851222.

[5] M. Balanici and S. Pachnicke, "Classification and forecasting of real-time server traffic flows employing long short-term memory for hybrid E/O data center networks," Journal of Optical Communications and Networking, vol. 13, no. 5, pp. 85–93, May 2021, doi: 10.1364/JOCN.411017.

[6] T. Benson, A. Akella, and D. A. Maltz, "Network Traffic Characteristics of Data Centers in the Wild," 2010.

[7] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the Social Network's (Datacenter) Network," ACM SIGCOMM Computer Communication Review, vol. 45, no. 4, pp. 123–137, Sep. 2015, doi: 10.1145/2829988.2787472.

[8] F. Yan, W. Miao, O. Raz, and N. Calabretta, "OPSquare: A flat dcn architecture based on flow-controlled optical packet switches," Journal of Optical Communications and Networking, vol. 9, no. 4, pp. 291–303, Apr. 2017, doi: 10.1364/JOCN.9.000291.

[9] K. Chen et al., "OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility

[10] K. Clark et al., "Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network," European Conference on Optical Communication, ECOC, vol. 2018-September, Nov. 2018, doi: 10.1109/ECOC.2018.8535333.

[11] X. Xue, B. Pan, X. Guo, and N. Calabretta, "Flow-Controlled and Clock-Distributed Optical Switch and Control System," IEEE Transactions on Communications, vol. 70, no. 5, pp. 3310–3319, May 2022, doi: 10.1109/TCOMM.2022.3156613

[12] P. Mahadevan, S. Banerjee, P. Sharma, A. Shah, and P. Ranganathan, "On energy efficiency for enterprise and data center networks," IEEE Communications Magazine, vol. 49, no. 8, pp. 94–100, Aug. 2011, doi: 10.1109/MCOM.2011.5978421.

[13] Y. Cai, Z. Yao, T. Li, S. Luo, and L. Zhou, "SD-MAC: Design and evaluation of a software-defined passive optical intrarack network in data centers," Transactions on Emerging Telecommunications Technologies, vol. 33, no. 8, Aug. 2022, doi: 10.1002/ETT.3764.

[14] Y. Cheng, M. Fiorani, R. Lin, L. Wosinska, and J. Chen, "POTORI: A passive optical top-of-rack interconnect architecture for data centers," Journal of Optical Communications and Networking, vol. 9, no. 5, pp. 401–411, May 2017, doi: 10.1364/JOCN.9.000401.

[15] J. L. Benjamin, T. Gerard, D. Lavery, P. Bayvel, and G. Zervas, "PULSE: Optical Circuit Switched Data Center Architecture Operating at Nanosecond Timescales," Journal of Lightwave Technology, vol. 38, no. 18, 2020, doi: 10.1109/JLT.2020.2997664

[16] T. Gerard, C. Parsonson, Z. Shabka, P. Bayvel, D. Lavery, and G. Zervas, "SWIFT: Scalable Ultra-Wideband Sub-Nanosecond Wavelength Switching for Data Centre Networks," Mar. 2020, Accessed: Oct. 20, 2023. [Online]. Available: http://arxiv.org/abs/2003.05489