Carrier Sensing Transmission Technique Study for Performance Improvement in Optical Data Center Networks

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

***Abstract*—Multiple applications based on Artificial Intelligence (AI) and Machine Learning (ML) emerge with an accelerating rate, requiring that various cloud Data Centers (DCs) around the world can process a constantly increasing number of datasets. As a result, the DC networks (DCNs) are requested to serve enormous number of tasks in an effective way. In this framework, the optical switching technology seems to be the most promising one for the DC servers interconnection grace to its flexibility and effectiveness, as compared with the electrical one. In this paper, we investigate the access control requirements in optical DCNs targeting their performance optimization and aiming to effectively serve traffic high bandwidth demands like those of AI applications traffic. Especially, we propose an efficient Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol, called Congestion Sense Medium Access Control (CS-MAC) protocol, for optical DCN environments exploring the bandwidth utilization level, as compared to the conventional CSMA scheme. Our proposal is based on the sensing of the congestion conditions in the network in order for the CS-MAC protocol to force the servers to properly adjust a lower number of transmissions, aiming to guarantee sufficient bandwidth to overcome the network congestion. Limited analytical study combined with simulation results show that the suggested CS-MAC protocol provides almost 100% bandwidth utilization, much higher than the conventional CSMA scheme, being a promising solution for the service of multiple AI and ML tasks traffic in optical DCNs.**

***Keywords*— bandwidth utilization, data center network, optical network.**

I. Introduction

Nowadays, multiple upcoming cloud applications have been arisen covering a wide variety of human activities, such as communication platforms employing real-time and time-sensitive traffic, social networking, internet of things (IoT), artificial intelligence (AI)-based and machine-learning (ML)-based applications, Industrial IoT (IoT) [1], Industry 4.0 [2] etc. The main characteristic of such applications concerns the need for an efficient cloud data center (DC) environment in terms of energy consumption, scalability and limited delay network connectivity among the servers. Especially, the enormous amount of traffic that such applications create by multiple tasks service requirement overloads the DC networks (DCNs) and demands ensured service level of low (in the order of microseconds) end-to-end delay and high scalability.

In support of these facts, one of the main concerns of DCN operators is the network efficiency in order to ensure the high performance needs. In this framework, the optical switching technology gradually gains ground since it provides higher effectiveness in terms of power efficiency and flexibility as compared to the electrical switching one. Therefore, the optical packet switching (OPS) [3-4], the optical circuit switching (OCS) [5-6], and the optical burst switching (OBS) [7-8] technologies have been proposed in literature for optical DCNs' environment. It is obvious that the transition from the electrical to the optical switching era in modern DCNs requests for expenditures and new network architecture solutions that have not been completely introduced due to technical constraints such as optical buffering luck (OPS) and high switching delays (OCS). For this reason hybrid switching solutions are currently used in DCNS providing high power consumption, targeting to the gradual transition to the extended optical switching technology use.

In this context, the implementation of hybrid or optical switching DCNs has to efficiently face important performance challenges like the transmission schemes requirements. Especially, synchronized access schemes under diverse transmission capacities of 10 Gbps, 40 Gbps or even higher, has been proposed [9-12], while experimental clock and data recovery (CDR) results in the range of nanoseconds have been introduced [13]. Under such high data rates that will reach up to 800 Gbps per wavelength in the near future and the complicated CDR control, the design and implementation of medium access control (MAC) algorithms is getting more critical. It is evident that the DCN effectiveness, either in the intra-rack or the inter-rack domain, is closely related to the optical devices technology limitations. Therefore, the requirement for synchronization among the DC servers for transmission is restricted by technology limitations as a function of the actual transmission rate and requires high expenditure, while for these reasons is hardly deployable. On the other hand, asynchronous transmission schemes are seldom proposed in literature and only in small scale intra-rack DCNs since the distance among the communicating servers has to be limited [14], avoiding large scale DCN implementations. For these reasons, MAC algorithms based on the carrier sense multiple access (CSMA) technique is used in DCNs environments [12], having as a precondition that the packets transmission time is higher that the propagation delay time in the fiber.

In this study, we propose an effective CSMA with collision detection (CSMA/CD)-based MAC protocol, called congestion sense medium access control (CS-MAC) protocol, to access the multiple wavelength division multiplexed (WDM) channels of an optical fiber that is used by the servers to be interconnected via a passive optical coupler in an intra-rack DCN. Since the conventional CSMA technique, even the CSMA/CD one, results to high retransmitted traffic load due to the increasing number of packet collisions under increasing offered load conditions, our proposed CS-MAM protocol adopts a performance improvement criterion to adjust the servers' transmission rate under congestion conditions in order to allow for the traffic service balance to be gradually return after any inefficient increase of packets collisions. Taking into account the high diversity of DCNs' traffic (in terms of packets interarrival time distribution, packet size etc.) due to the multiple cloud applications that they serve [15-17], the proposed CS-MAC protocol seems to be a promising solution for effective access in optical intra-rack DCNs. This is because the proposed CS-MAC protocol does not require any synchronization control that would result to high cost and technology restrictions, providing a cost, energy and performance effective intra-rack optical DCN solution. In order to compare the suggested CS-MAC protocol performance to that of a conventional CSMA scheme, we adopt a simple analytical model to derive analytical results for the average successful transmission rate, i.e. the average throughput. The comparison shows that the proposed CS-MAC protocol achieves effective congestion control improving the throughput around 100% under high offered traffic conditions, providing high bandwidth utilization.

The study is organized as follows: Section II presents the proposed DCN model and the CS-MAC protocol. Section III gives the analytical model for the CSMA-based scheme. The performance evaluation and the comparison is given in Section IV. Finally, section V concludes our study.

II. DCN Model and CS-MAC Protocol

The proposed DCN architecture is illustrated in Fig. 1. We assume that a number M of servers are interconnected within a rack using a passive optical coupler. Each server is connected with the passive coupler using an optical fiber where (W+1) wavelengths (i.e. λ_0, λ_1… λ_W) are multiplexed based on the wavelength division multiplexing (WDM) technique. In this way, a multi-channel intra-rack optical system is formed. Each wavelength runs @ data rate of R Gbps. We assume that each server lays at a distance of L meters away from the passive coupler and we denote as T_p the propagation delay in the intra-rack network. On the other hand, a server's communication with servers of different racks is performed using a separate set of wavelengths, over the inter-rack DCN beyond this study focus.

Two network interfaces are used by each server, as Fig. 1 presents. The first one is dedicated for the intra-rack communication including (W+1) fixed optical transceivers each of which is always tuned to a specific channel from the set {λ_0,λ_1, λ_2… λ_W}. The second one is dedicated for the inter-rack communication including an optical tunable transceiver that operates over different waveband for the inter-rack DCN communication. Also, each server is equipped with a multiplexer/demultiplexer (MUX/DEMUX) to adapt all intra-rack channels on the optical fiber to the passive coupler, and a separate MUX/DEMUX for the connection to the upper DCN switching equipment through the top-of-rack (ToR) switch.

The assumption of two separate optical network domains for the intra-rack and inter-rack communication, which operate independently one from the other using different wavelengths and server's network interfaces, provides many advantages: it avoids any interference between the two networks and supports the adoption of difference MAC protocols to each of them.

In the following, we focus on the intra-rack DCN domain. Based on the broadcast nature of the proposed intra-rack network architecture, a packet transmitted by any server on any channel in the same rack, is distributed to all servers in the rack by the passive coupler. We assume that the aggregated traffic in a rack obeys Poisson statistics. Also, the generated intra-rack traffic by each one server follows the Poisson statistics, while packets of constant size L_p are generated. We denote by T a packet's transmission time, i.e T= L_p/R. The generated packets by each server are stored in an electrical output buffer which is served in a first-in-first-out (FIFO) way, while the incoming from the intra-rack network traffic is stored in a relative electrical input buffer, as Fig. 1 illustrates.

We define that the intra-rack multi-channel system is idle at a specific time instance if there is no server transmitting over any of the channels at this time instance. In other words, the sensing technique in order to determine if the system is idle does not concern each channel separately but the whole multi-channel system. On the other hand, if the multi-channel system is not sensed idle, it is busy. The busy state of the multi-channel system can be characterized by successful and/or unsuccessful packet transmissions, depending on the number of servers transmitting over each channel.

The proposed CS-MAC protocol does not require any synchronization among the servers in the intra-rack DCN. For this reason is an cost efficient solution avoiding any clock and data recovery (CDR) expenses and implementation complexity. Based on the CSMA/CD access procedure, if a server intents to a data packet transmission, it first asynchronously senses the multi-channel intra-rack network with its fixed receivers to check if an optical signal is transmitted over the (W+1) intra-rack channels. If all of the channels are currently transmitting a packet, then the server continues to sense the multi-channel network and postpones its transmission for another time instant in the future. Otherwise, if the server senses that one or more intra-rack channels are free, it selects randomly one of the free channels with equal and constant probability and starts transmitting its packet over the selected free channel. According to the collision detection procedure, the server monitors the transmission channel and receives the optical signal over it during the transmission time period plus the propagation delay time period in order to determine if a packet collision occurs. In this case, the server transmits a jamming signal over the transmission channel in order to inform all the other servers about the collision, while it increases a counter indicating the server collisions level. After the collision detection, the server waits for a time period according to the binary exponential backoff algorithm before attempting to the next transmission. It is obvious that the carrier sensing technique results to a rapidly increasing number of collisions especially under high

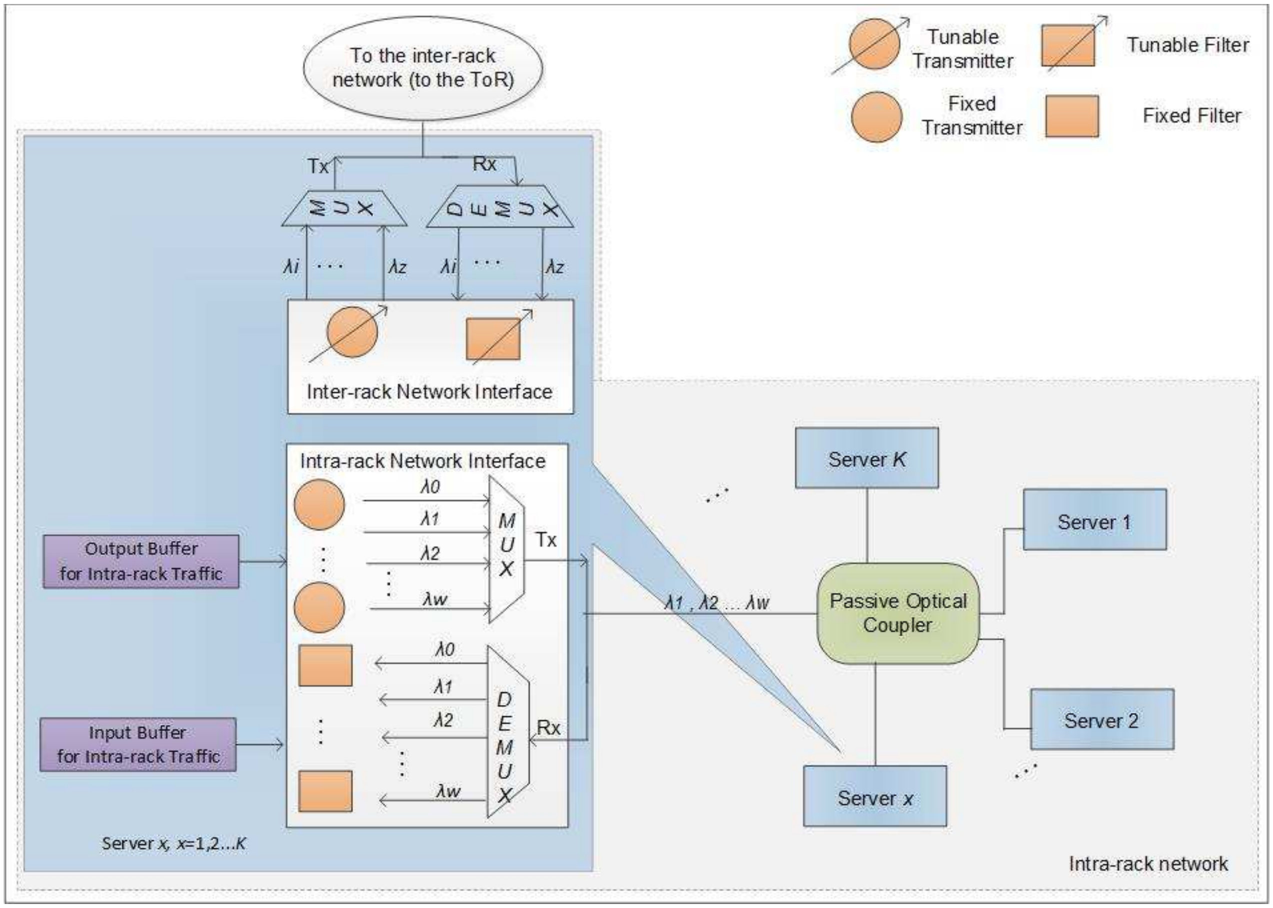

Fig. 1 – Proposed intra-rack DCN and server architecture.

offered load conditions, which consequently gets worse the congestion conditions in the intra-rack network. In order to avoid the escalation of congestion, the proposed CS-MAC protocol keeps for each server a network performance parameter, i.e a counter indicating the collisions level. Each server updates this counter each time its transmitted packet is involved into a collision and informs all the other servers when the value of this counter gets higher than a predefined accepted level forcing all the servers to stop any new packet transmission until the congestion conditions pass. After applying this transmission embargo (and eliminating the collision counter) and when network balance is achieved again, the proposed CS-MAC protocol lets all servers to start a new packet transmission according to the above described procedure.

The asynchronous operation of the proposed CS-MAC access scheme determines time periods of random length during which the intra-rack multi-channel system is idle. These idle periods are followed by transmission periods during which successful and/or unsuccessful transmissions occur over the multi-channel system. During this time periods the multi-channel system is said to be busy. The busy period has fixed length (T+T_p) since it has to host a packet transmission with transmission time T and each channel requires time interval equal to the propagation delay T_p to be cleared. We define as cycle the time interval that includes an idle period followed by a transmission period, as Fig. 2 depicts. It is obvious that a cycle has a random size depending on the random duration of the relative idle period.

III. Analysis of Simplified CS-MAC Protocol

In order to have a quantitative base for comparison for the proposed CS-MAC protocol performance, we first study the performance of the conventional simplified version of the CSMA/CD access scheme based on an analytical model. In this way, we can derive the improvement level provided by the proposed CS-MAC protocol as compared to the conventional simplified CSMA/CD one.

Thus, for the CSMA/CD protocol study, as previously adopted, we also assume that the aggregated offered traffic by all servers in the intra-rack DCN obeys Poisson statistics. The average number of packets offered by all servers over all channels during a propagation delay T_p time period (which is defined as time unit) is denoted by G. In other words, we denote as G the mean rate of total packet offered by all servers over the multi-channel system during a time unit. The average successful transmissions rate S_C from the multi-channel system is given by:

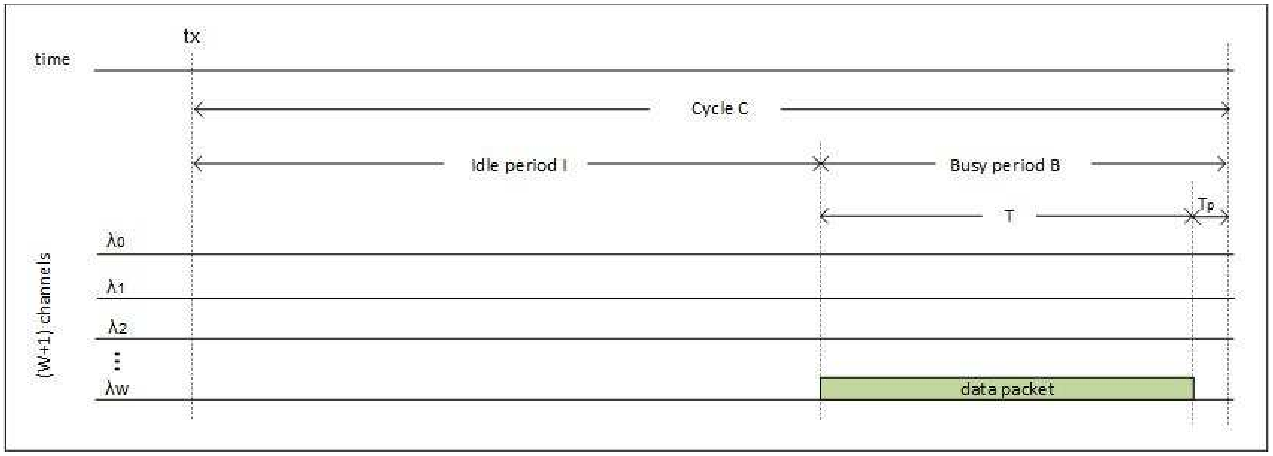

Fig. 2. Cycle definition.

$$S_v = Ge^{-G/(W+1)} \tag{1}$$

We denote as B and I the expected duration of the busy and idle period respectively during a cycle. Therefore, the expected duration C of a cycle is:

$$C = B + I \tag{2}$$

The probability P_1 that an idle period ends at a given time instant is given by:

$$P1 = 1 - e^{-G} \tag{3}$$

It is obvious that:

$$I = \frac{1}{(1-e^{-G})} \tag{4}$$

and:

$$B = T + 1 \tag{5}$$

Moreover, we denote as U the expected time period during a cycle that the multi-channel system is used without collisions. It is:

$$U = \frac{TGe^{-G/(W+1)}}{(1-e^{-G})} \tag{6}$$

Thus, we define the average expected throughput S from the multi-channel system as the cycle percentage during which the multi-channels system is used for successful transmissions, i.e.:

$$S = \frac{U}{C} \tag{7}$$

Substituting (2), (4), (5) and (6) to (7), we get:

$$S = \frac{TGe^{-G/(W+1)}}{1+(T+1)(1-e^{-G})} \tag{8}$$

IV. Performance Evaluation

For the suggested CS-MAC protocol performance assessment, we built a C-based simulation environment in order to derive the performance measures of average throughput and average packet delay. In the following, we assume that L_p = 1500 Bytes, R =100 Gbps, T = 120 ns, L = 5 m, and T_p= 25 ns.

Figure 3 shows the normalized throughput versus the normalized offered load for M=10 servers and W=6 intra-rack channels for the proposed CS-MAC protocol. As it is depicted, the proposed CS-MAC protocol manages to exploit all the available bandwidth in order to serve almost all of the informing traffic. This is shown since there is an one to one correspondence among the offered load and the throughput normalized values, as Fig. 3 illustrates. This fact is based on the efficient transmission adjustment algorithm that the proposed CS-MAC protocol follows in order to face the congestion conditions in the multi-channel network and to manage transmission balance. For example, for the high offered load value of 0.9 the throughput value achieved is almost 0.88, resulting in an almost 100% bandwidth utilization even under highly loaded conditions.

On the other hand, the optimum throughput performance is achieved paying the cost of increasing end-to-end delay. In

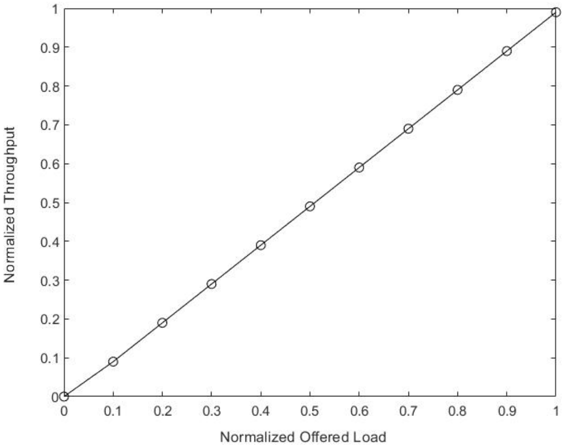

Fig. 3. Average normalized throughput vs average normalized offered load, for M=10, W=6 (i.e. 7 channels).

other words, the transmission embargo that the proposed CS-MAC protocol forces the rack servers to follow in order to face the congestion conditions, results in an increase of the packet delay. This is depicted in Fig. 4 that presents the average packet delay versus the normalized offered load for *M*=10 servers and *W*=6 intra-rack channels for the proposed CS-MAC protocol. In fact, the previous results are validated. As it is depicted, for loads up to 0.4 where the collisions level is low enough, the average packet delay is almost zero. This means that at this low load range the proposed CS-MAC protocol is able to serve the incoming traffic without delay since the congestion conditions in the multi-channel intra-rack network are not met. As the offered load to the multi-channel system increases, the packets collisions level gradually increases too, while the proposed CS-MAC protocol creates more transmission embargo alarms in order to face the network congestion. This is the reason why, as load increases at loads higher than 0.4 the average packet delay gradually increases too. For example, for load values 0.6, 0.8 and 0.9 the respective average packet end-to-end delay values are 9 μs, 45 μs and 90 μs respectively.

In Fig. 5, we study the average throughput variation as a function of the number of channels in the intra-rack network. Especially, Fig. 5 presents the average throughput versus the average normalized offered load for *M*=10 and *W*=1, 2, 3. As it is illustrated, for a fixed load value the throughput achieved is an increasing function of the number of channels (*W*+1). For example, for load 0.9 the throughput value is 143 Gbps for 2 channels, 232 Gbps for 3 channels and 372 Gbps for 4 channels. This is an immediate result of the increasing available bandwidth for transmission as the number of channels increases. An interesting result comes from the observation of the bandwidth utilization level reached at high offered load conditions as the number of channels increases. In fact, the bandwidth utilization is 71.5% for 2 channels, 77.3% for 3 channels and 93% for 4 channels as Fig. 5 depicts. At this point, it is remarkable that the bandwidth utilization increases to almost 100% for 7 channels, as Fig. 3 illustrates. In other words, the bandwidth utilization is an increasing function of the number of channels in the intra-rack multi-channel network. This behavior is explained by the fact that the load value at which congestion is met is an increasing function of the number of channels too, i.e. 0.6 for 2 channels, 0.7 for 3 channels and 0.88 for 4 channels. This is because as the number of channels increases there is more available bandwidth to be exploited for successful transmissions providing lower probability of packet collisions and lower congestion level.

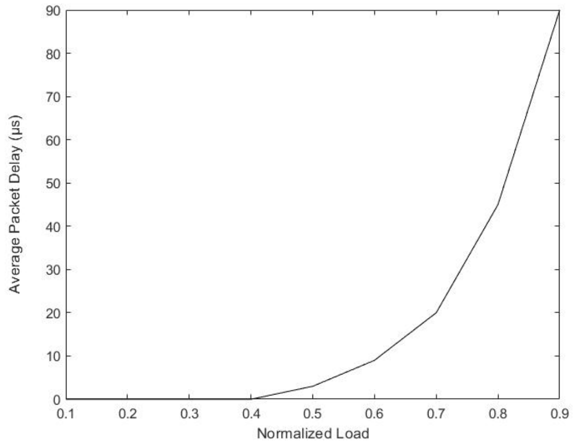

Fig. 4. Average packet delay vs normalized offered load, for *M*=10, *W*=6 (i.e. 7 channels).

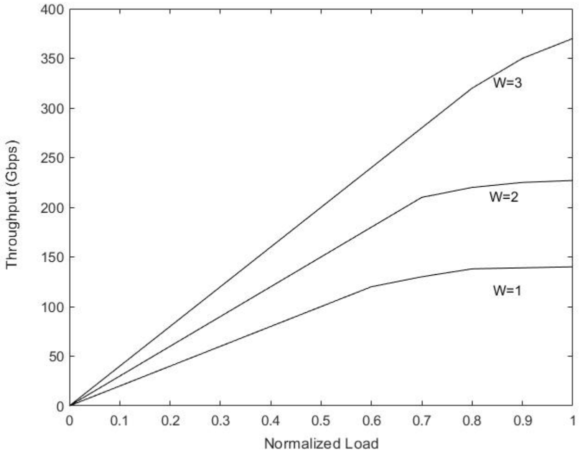

Fig. 5. Average throughput vs average normalized offered load, for *M*=10 and *W*=1, 2, 3 (i.e. 2, 3, 4 channels).

The performance improvement provided by the proposed CS-MAC protocol as compared with the conventional CSMA one, is representatively studied in Fig. 6 that presents the average throughput versus the average offered load, for *M*=15 and *W*=3. As it is depicted, the conventional CSMA reaches congestion conditions at very low loads, lower than 5 Gbps, since the number of collisions is enormous even by this low offered load conditions. On the other hand, the proposed CS-MAC protocol manages to effectively serve the incoming traffic up to loads of 220 Gbps before reaching congestion

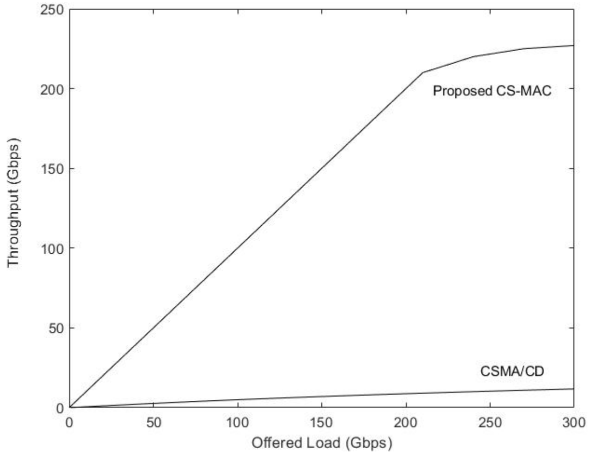

Fig. 5. Probability of destination conflict vs offered load, for *K*=50, *B*=25 *W*=4, 8, 12.

conditions where the slope of the throughput curve decreases sharply. This is because the followed by the suggested CS-MAC protocol transmission embargo algorithm in order to face the congestion conditions efficiently ensures the required bandwidth for reaching again transmission balance. For example for offered load 200 Gbps the throughput improvement that the proposed protocol provides is 2122%.

Finally, our proposal is a power efficient solution since it provides low energy footprint as compared with other intra-rack optical DCN architectures, like that of POTORI [18]. The power consumption is shown in Fig. 6 for the compared DCN architectures. As it is shown, the proposed CS-MAC architecture provides more than 150% power consumption decrease as compared with POTORI. To this result advocates the fact that our solution exploits passive optical devices for the servers interconnection.

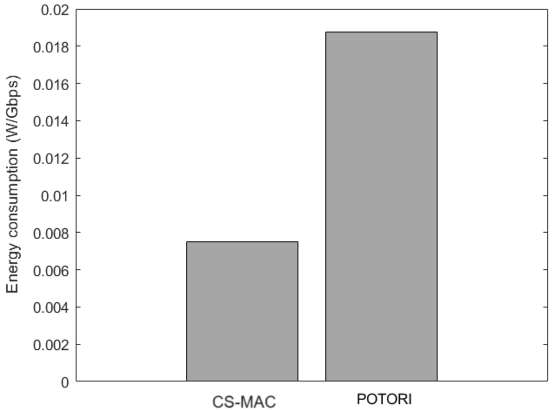

Fig. 6. Power consumption comparison.

V. Conclusion

The CS-MAC protocol is proposed in this study, suitable for optical intra-rack multi-channel DCNs. The CS-MAC protocol belongs to the CSMA protocols family and it provides improved performance as compared with conventional CSMA schemes. This is grace to the suggested transmission adjustment algorithm that senses the congestion level at the DCN and appropriately controls the servers' transmission in order to obtain transmission balance in the multi-channel DCN and sufficiently improve the bandwidth utilization achieved. Numerical analysis and simulation results prove that the throughput improvement that the proposed CS-MAC protocol provides as compared to the conventional CSMA scheme is more than 2000% under high loaded conditions. Moreover, the proposed DCN architecture is power efficient as compared to other intra-rack DCN architectures.

References

[1] Abdellah Chehri, "MAC Protocols for Industrial Delay-Sensitive Applications in Industry 4.0: Exploring Challenges, Protocols, and Requirements", Procedia Computer Science, vol. 192, pp. 4542–4551, 2021

[2] L. Chinchilla-Romero et. al., "5G Infrastructure Network Slicing: E2E Mean Delay Model and Effectiveness Assessment to Reduce Downtimes in Industry 4.0", Sensors, vol. 22, pp. 229, 2022.

[3] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," Commun. Surveys Tuts., vol. 14, pp. 1021–1036, 2012.

[4] C. Kachris et al., "Optical interconnection networks in data centers: Recent trends and future challenges," Communication Magazine, vol. 51, pp. 39–45, 2013.

[5] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, "The emerging optical data center," in IEEE/OSA Opt. Fiber Commun.(OFC), 2011, Paper OTuH2.

[6] N. Farrington et al., "Helios: A hybrid electrical/optical switch architecture for modular data centers," in SIGCOMM, 2010.

[7] M. Imran et al., "Software-defined optical burst switching for HPC and cloud computing data centers," J. Optical Communication and Networking, vol. 8, pp. 610–620, 2016.

[8] M.Y. Sowailem et al., "Contention Resolution Strategy in Optical Burst Switched Datacenters," in OFC/NFOEC, 2013.

[9] Y. P. 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", Trans. Emerging Telecom. Techn., e3764, 2019.

[10] Y. P. Cai, et. al., "Design and Evaluation of a Software Defined Passive Optical Intra-Rack Network in Data Centers", in Proc. IEEE INFOCOM 2019 - ICCN Workshops, 2019, pp.56-61.

[11] W. Ni, et. al., "POXN: A new passive optical cross-connection network for low cost power-efficient datacenters", Journal of Lightwave Technology, vol.32, pp. 1482–1500 , 2014.

[12] Y. Zheng and X. Sun, "Dual MAC Based Hierarchical Optical Access Network for Hyperscale Data Centers", Journal of Lightwave Technology, vol. 38, pp. 1608–1617, 2020.

[13] X. Xue and N. Calabretta, "Nanosecond optical switching and control system for data center networks," Nat. Commun. 13, 2257, 2022.

[14] P.A.Baziana: "i-WABA: An Efficient Wavelengths Arrangement and Bandwidth Allocation WDMA Protocol for Passive Optical Intra-rack Data Center Networks", Trans. on Emerging Telecom. Techn., Vol. 33, No. 9, e4521, pp. 1-16, 2022.

[15] T. Benson et al., "Understanding data center traffic characteristics", in ACM SIGCOMM Comp. Comm. Rev., 2009.

[16] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in ACM SIGCOMM, 2010.

[17] S. Kandula et al., "The nature of datacenter traffic: Measurements & analysis," in ACM SIGCOMM, 2009.

[18] Cheng Y, , et. al., "POTORI: A passive optical top-of-rack interconnect architecture for data centers," J. Opt. Commun. Netw., vol. 9, pp. 401–411, 2017.