Hybrid and Optical Packet Switching Supporting Different Service Classes in Data Center Network

Artur Minakhmetov, Cédric Ware, and Luigi Iannone

LTCI, Télécom ParisTech, Université Paris-Saclay, Paris 75013, France
artur.minakhmetov@telecom-paristech.fr

Abstract. Optical Packet Switching is a prominent technology proposing not only a reduction of the energy consumption by the elimination of numerous optical-electrical-optical conversions in electronic switches, but also a decrease of network latencies due to the cut-through nature of packet transmission. However, it is adversely affected by packet contention, preventing its deployment. Solutions have been proposed to tackle the problem: addition of shared electronic buffers to optical switches (then called hybrid opto-electronic switches), customization of TCP protocols, and use of different service classes of packets with distinct switching criteria.
In the context of data center networks we investigate a combination of said solutions and show that the hybrid switch, compared to the optical switch, boosts the performance of the data center network. Furthermore, we show that introducing a "Reliable" service class improves performance for this class not only in the case of the hybrid switch, but also brings the optical switch to performance levels comparable to that of the hybrid switch, all the while keeping other classes' performance on the same level.

Keywords: Optical Packet Switching · Packet Switching · TCP Congestion Control · Optical Switches · Hybrid Switches · Classes of Service · Packet Preemption.

1 Introduction

The Optical Packet Switching (OPS) technology regained public interest in the mid-2000s [4] in the face of demand for high reconfigurability in networks, made possible through statistical multiplexing along with efficient capacity use and limiting the energy consumption of the switches [15]. However, with traffic being asynchronous and in the absence of technology that would make practical optical buffers in switches, the contention issue arises, leading to poor performance in terms of Packet Loss Ratio (PLR) [10], thus making the OPS concept impractical. To the present moment, several solutions have been proposed to bring the OPS technology to functional level, among which: adding a shared electronic buffer, thus making hybrid opto-electronic switches [21, 19, 17]; intelligent routing of packets of different priorities in the hypothesis that not all of

them would need the same requirements for PLR [16]; and a network-level solution without changing the OPS hardware, introducing special TCP Congestion Control Algorithms (CCA) for packet transmission in order to increase overall network throughput, thus negating the still high PLR [6]. These three solutions are detailed below.

First, the hybrid switch consists in coupling an all-optical bufferless packet switch with an electronic buffer. Several implementations of the idea were already proposed in the last decade [21, 19, 17]. The concept of the hybrid switch considered in this study is: when contention occurs on two (or more) packets, i.e. when a packet requires using an output that is busy transmitting another packet, it is diverted to a shared electronic buffer through Optical-Electrical (OE) conversion. When the destination output is released, the buffered packet is emitted from the buffer, passing Electrical-Optical (EO) conversion. However, in the absence of contention, the hybrid switch works as an all-optical switch, without any wasteful OE and EO conversions. Adding a shared buffer with only a few input-output ports lets us considerably decrease PLR compared to an all-optical switch, and bring its performance up to the level of an electronic switch, but now with an important reduction in energy consumption, since one would save the OE/EO (OEO) conversions for most packets [16].

Second, highlighting an important question of the existence of classes of service in a network, Samoud et. al. [16] propose handling packets depending on their class: high priority packets can preempt low priority ones from being buffered or transmitted. It was shown that the demand for low PLR may be met for high priority packets and relaxed for others, achieving sustainable operation with a number of buffer input/output ports less than half that of optical links in a switch.

Third, Argibay-Losada et. al. [6] propose to use all-optical switches in OPS networks along with special TCP CCAs, in order to bring the OPS network throughput up to the same levels as in Electrical Packet Switching (EPS) networks with conventional electronic switches. Particularly noteworthy in protocol design is the Retransmission Timeout (RTO). This parameter controls how long to wait for the acknowledgment after sending a packet until the packet is considered lost and re-sent. When a transmission is successful and without losses, RTO is set to a value close to the Round-Trip-Time (RTT), i.e. the time elapsed between the start of sending a packet and reception of its acknowledgment. By simple tweaking of initialization value of RTO and reducing it from conventional 1 s to 1 ms, it was shown that both custom and conventional TCP CCAs will boost the performance of the optical packet switched network.

In our previous works we analyzed the gain from use of the hybrid switch in a Data Center (DC) network by introducing Hybrid Optical Packet Switching (HOPS): we showed that HOPS with a custom designed TCP can outperform OPS and EPS in throughput [13, 12]. Furthermore, in [11] we have managed to show the possibility of 4 times reduction in DC energy consumption for data transport coming from OEO conversions while using HOPS compared to EPS. In this study we aim to investigate not only a combination of HOPS with custom

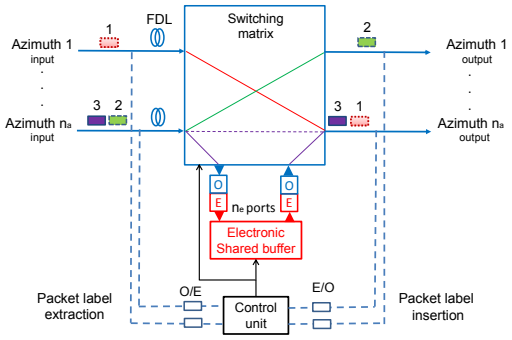

Fig. 1: General architecture of hybrid optical packet switch

design of TCP, but also the influence of the introduction of Classes of Service, i.e. switching and preemption rules for packets of different priorities.

Considering the general interest in the scientific and industrial communities to implement different packets priorities in Data Centers (DCs), as well as the problem of traffic isolation for tenants in DC [14], we implement the idea presented by Samoud et. al. [16] and investigate the benefits of application of such technology in a DC network. We successfully show that one can considerably improve the performance of network consisting of hybrid switches with a small number of buffer inputs for high priority connections while keeping it on a good level for default connections. Additionally, we show that high priority connections in OPS network also can profit from the introduction of classes of service, matching or even surpassing the performance of the network consisted of hybrid switches with a small number of buffer inputs without classes of service.

The paper is composed as follows: Sec. 2 presents hybrid switch's architecture and packets preemption policy, Sec. 3 outlines simulation conditions, Sec. 4 discusses the results obtained and, finally, Sec. 5 offers our main conclusions.

2 Hybrid Switch Architecture and Packets Preemption Policy

2.1 Hybrid Switch Architecture

The first concept of a hybrid switch was proposed in 2004 by R. Takahashi et al. [20], and the scientific community has kept its attention on the implementation of the idea since then [17]. In 2010 X. Ye et al. [21] presented a Datacenter Optical Switch (DOS), an optical packet switch, that could be seen as a prototype of a hybrid switch: switching was performed through a combination of Arrayed Waveguide Gratings switching matrix with Tunable Wavelength Converters (TWC), contentions were managed through the shared electronic buffer, storing contending packets. In 2012 R. Takahashi et al. [19] presented a similar concept, called Hybrid Optoelectronic Packet Router (HOPR). DOS and HOPR,

despite the name, are not quite what we call hybrid switches, as all the packets undergo OEO conversions by TWCs.

In 2016 T. Segawa et al. [17] proposed a switch that performs switching of optical packets through a broadcast-and-select switching matrix and then re-amplification by Semiconductor Optical Amplifiers (SOAs). This switch splits the incoming optical packet into several ways corresponding to output ports, blocks those that don't match the packet's destination, and then re-amplifies the passed packet with a SOA. A shared electronic buffer is there to solve packet contention. The OEO conversion is made only for contending packets, unlike DOS or HOPR where all the packets undergo OEO conversions.

All of the presented solutions above have common main blocks, that we are emulating in our study in order to approach hybrid switch functions. The general structure of a hybrid switch is presented in Fig. 1 with the following main blocks: an optical switching matrix; an electronic shared buffer; and a control unit that configures the latter two according to the destination of the packets, carried by labels. The hybrid switch has n_a inputs and n_a outputs, representing non-wavelength-specific input and output channels, or Azimuths, thus making n_a channels for a switch. Another important parameter is n_e: n_e inputs and n_e outputs of a buffer. These are the channels through which a packet is routed/emitted to/from a buffer.

In our study we make the following assumptions. The optical matrix has a negligible reconfiguration time, on the ns scale [8]. The labels can be extracted from the packet and processed without converting the packet itself to electronic domain, e.g. by transmitting them out of band on dedicated wavelengths as in the OPS solution presented by Shacham et al. [18]. This solution allows label extraction via a tap coupler, requiring an OE conversion only for the label, and short Fiber Delay Lines at the inputs of the optical switch. We are not considering any particular technology for the Control Unit, and implement our simulations focusing on the supposed ideal optical matrix, and on a store-and-forward buffer.

2.2 Packets Preemption Policy

The switching algorithm for a hybrid switch is adopted from [16] and implements different bufferization and preemption rules for different packets classes. We consider three of them: Reliable (R), Fast (F) and Default packets (D). R packets are those that attempted to be saved by any means, even by preemption of F or D packets on their way to buffer or switch output. F packets could preempt only D packets on their way to the switch output. D packets cannot preempt other packets.

The priority distribution in the DC network is adopted from [16] and taken from the real study on core networks [1]. This may seem improper for DCs, however, we seek to study the performance of the hybrid switch in the known context. Also, it will be shown below that the distribution considered lets us organize a pool of premium users (10%) of R connections in DCs that could

Algorithm 1 Preemption Policies in a Hybrid Switch

```
1: procedure ROUTE (PACKET p)
2:     prio ← p.priority_class
3:     switch_out ← get_destination_azimuth(p)
4:     if switch_out.is_free() then
5:         switch_out.receive(p)
6:     else if buffer_in.is_free() then
7:         buffer_in.receive(p)
8:     else if prio==R and buffer_in.receiving(D) then
9:         buffer_in.preempt_last_packet(D)
10:        buffer_in.receive(p)
11:    else if prio==R and switch_out.receiving(D) then
12:        switch_out.preempt_last_packet(D)
13:        switch_out.receive(p)
14:    else if prio==R and buffer_in.receiving(F̃) then
15:        buffer_input.preempt_last_packet(F̃)
16:        buffer_input.receive(p)
17:    else if prio==R and switch_out.receiving(F̃) then
18:        switch_out.preempt_last_packet(F̃)
19:        switch_out.receive(p)
20:    else if prio==F̃ and switch_out.receiving(D) then
21:        switch_out.preempt_last_packet(D)
22:        switch_out.receive(p)
23:    else
24:        drop(p)
```

profit from the best performance, while other users almost wouldn't be influenced by performance loss. F packets can preempt D packets only on the way to switch output, while R packets first would consider preemption of D packet being buffered. Thus F packets had lower delay than R packets [16]. However, further it will be shown that this device-level gain doesn't translate to network-level gain in a DC network in terms of Flow Completion Time (FCT), and R connections perform better than F. That's why here we refer to Fast (F) as Not-So-Fast ($\tilde{\text{F}}$) packets and connections. Eventually, in this study we consider, that 10% of connections have R priority, 40% of connections have $\tilde{\text{F}}$ priority, 50% of connections have D priority.

When a packet enters the switch it checks if required Azimuth output (i.e. switch output) is available. If yes, the packet occupies it. Otherwise, the packet checks if any of buffer inputs are available. If yes, it occupies one and starts bufferization. If none of the buffer inputs are available, in the case of absence of preemption policy in a switch the packet would be simply dropped. Here, we consider a switch with preemption policy that would follow the steps of algorithm presented in Alg. 1. If a packet of any type is buffered, it is re-emitted FIFO, as soon as required switch output is available.

3 Study Methodology

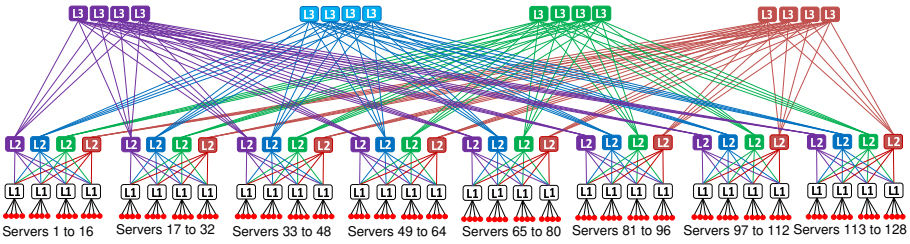

Fig. 2: Fat-tree topology network, interconnecting 128 servers with three layers of switches.

As in our previous work [12, 13], we simulate the communications of DC servers by means of optical packets. We study DC network performance for two groups of scenarios: DC with classes of service using preemption policy outlined in Sec. 2.2, and DC with switches that don't have any preemption rules. For each scenario we consider OPS and HOPS case.

Communications consist of transmitting files between server pairs through TCP connections. The files' size is random, following a lognormal-like distribution [3], which has two modes around 10 MB and 1 GB. We simulate transmission of 1024 random files (on the same order as 1000 in [6]), i.e. 8 connections per server. File transmission is done by data packets using jumbo frames with a size of 9 kB. This value defines the packet's payload and corresponds to Jumbo Ethernet frame's payload.

In our study we also use SYN, FIN, and ACK signaling packets. We choose for them to have the minimal size of the Ethernet frame of 64 B [2]. We assume that this minimal size would contain only the relevant information about Ethernet, TCP/IP layers. As we still need to attach to the jumbo frames all the information of these layers, for simplicity, we just attach to it a header of 64 B discussed previously. Thus we construct a packet of maximum size 9064 B to be used in our simulations, with a duration τ dependent on the bit-rate. Servers have network interface cards of 10 Gb/s bit-rate. Buffer inputs and outputs used by a hybrid switch support the same bit-rate.

The actual transmission of each data packet is regulated by the DCTCP CCA [7], developed for DCs, which decides whether to send the next packet or to retransmit a not-acknowledged one. CCA uses next constants: $DCTCP_{threshold} =$ 27192 B, $DCTCP_{acks/pckt} = 1$, $DCTCP_g = 0.06$, as favorable for HOPS. We apply the crucial reduction of the initialization value of RTO towards 1 ms, as advised in [6]. To be realistic, the initial 3-way handshake and 3-way connection termination are also simulated.

We developed a discrete-event network simulator based on an earlier hybrid switch simulator [16], extended so as to handle whole networks and include TCP

emulation. The simulated network consists of hybrid switches with the following architecture: each has n_a azimuths, representing the number of input/output optical ports, and n_e input/output ports to the electronic buffer, as shown in Fig. 1. The case of the bufferless all-optical switch (OPS) corresponds to $n_e = 0$, for the case of the hybrid switch (HOPS) we consider $n_e = 2$.

We study the DC fat-tree topology, interconnecting 128 servers by means of 80 identical switches with $n_a = 8$ azimuths, presented in Fig. 2, a sub-case of a topology deployed in a Facebook's DCs [5]. All links are bidirectional and of the same length $l_{link} = 10$ m as typical link lengths for DC. The link plays the role of device-to-device connection, i.e. server-to-switch, switch-to-server or switch-to-switch. The link is supposed to represent a non-wavelength-specific channel. Paths between servers are calculated as a minimum number of hops, which offers multiple equal paths for packet transmission allowing load-balancing and thus lowering the PLR.

The network is characterized by the network throughput (in Gb/s) and average FCT (in µs) for each type of connections and general case as a function of the arrival rate of new connections, represented by the Poissonian process. We have chosen FCT as a metric considered to be the most important for network state characterization [9].

4 Evaluation Results

We present here the results of our study and their analysis. To reduce statistical fluctuations, we repeated every simulation a hundred times with different random seeds for $n_e = 0$ (OPS) and $n_e = 2$ (HOPS). The mean throughput and mean FCT are represented in Fig. 3 and in Fig. 4 with 95% t-Student confidence intervals, for three types of connections: R, $\tilde{\text{F}}$ and D connections. We take as a reference results from the network without packet preemption policy: the division of connections to classes is artificial and just represent corresponding to classes' percentage of connections in the network. We define high load as more than 10^5 connections per second.

While comparing just OPS and HOPS, it is seen that in general HOPS outperforms or has the same performance as OPS, but with the cost of only $n_e = 2$ buffer inputs.

R connections benefit the most from the introduction of the Classes of Service and preemption policy as it seen on Fig. 3a) and Fig. 4a) both in the cases of OPS and HOPS. Throughput for R connections in HOPS network rises by around 25% (Fig. 3a), while in OPS case it rises by a factor 2.5 at least on high load, matching the performance of HOPS network. We would like to bring readers attention on the fact that it seems to be low throughput, compared to other classes of service, but this is the mere effect of the fact that in the network only 10% of connections are of type R. However, if one considers the FCT, which is comparable with other types of classes and lowest among them, then the preemption policy's benefits are more evident: on the highest considered load OPS reduces its FCT almost by a factor of 8, while HOPS reduces it by at least

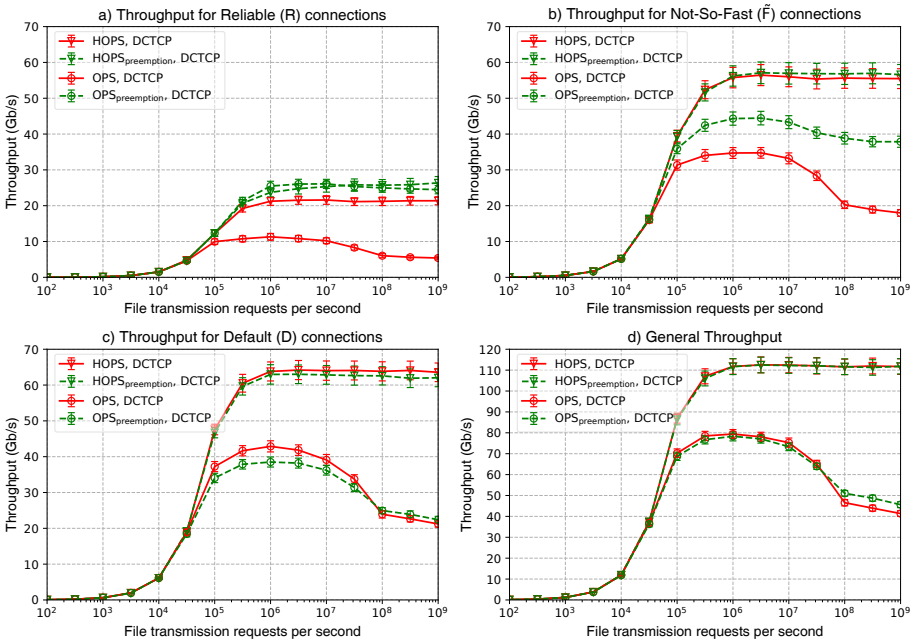

Fig. 3: DC network's throughput for connections: a) Reliable (R) connections, b) Not-So-Fast ($\tilde{F}$) connections, c) Default (D) connections, d) Overall Network Performance

a factor of 2, keeping it on the level of tens of µs. Even if OPS's FCT doesn't match FCT in the case of HOPS while considering Classes of Service, it does match the FCT in the case of HOPS without Classes of Service. While applying preemption policy, connections are indeed Reliable: in Fig. 5 we can see that PLR (ratio of packets lost due to preemption or dropping to packets emitted by servers) decreases by around factor of 10, while for $\tilde{F}$ and D PLR remains around the same level (not shown here).

$\tilde{F}$ traffic benefits less than R traffic from introduction of Classes of Service, but the gain is still there. For OPS we managed to boost the throughput by almost 30-100% on the high load, while for HOPS the gain is less evident. However, when we consider FCT on Fig. 4b) we can see that OPS decreases its FCT by almost a factor of 2 for high load, and HOPS around 25%. HOPS FCT for $\tilde{F}$ packets is bigger than for those of reliable (R), contrary to what may be induced from [16], where they are labeled as Fast (F). This may be explained by the fact that the delay benefits for F packets are on the order of a µs, while here FCT is of an order of tens and hundreds of µs, and is defined mostly by TCP CCAs when contention problem is solved.

D traffic does not benefit from the introduction of Classes of Service, and it is on its account the gains for R and $\tilde{F}$ traffic exists. However, while considering the performance reductions, we notice almost unchanged throughput for HOPS

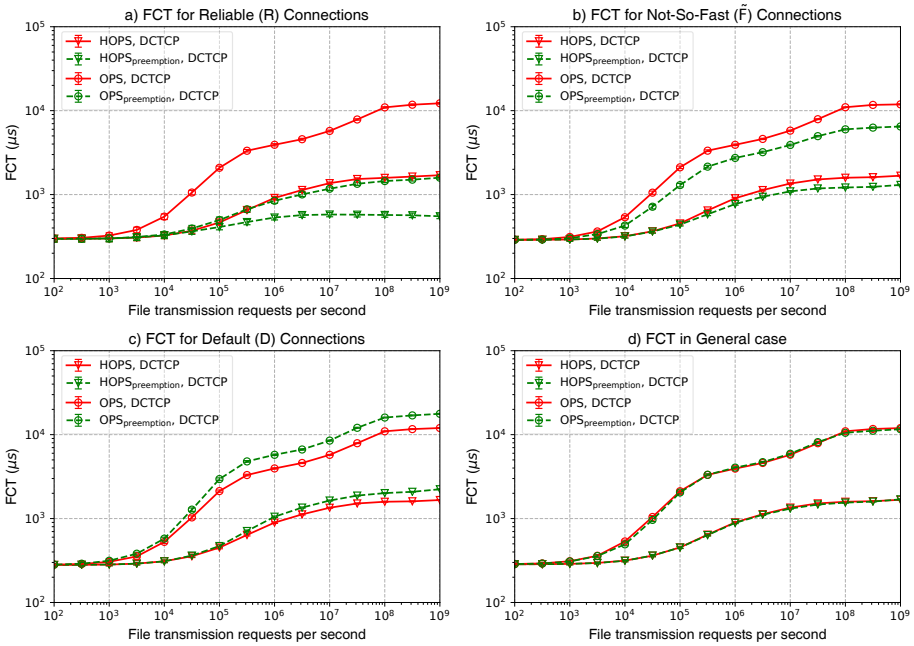

Fig. 4: DC network's Flow Completion Time for connections: a) Reliable (R) connections, b) Not-So-Fast ($\tilde{F}$) connections, c) Default (D) connections, d) Overall Network Performance

case, and for OPS the drop of only 10% at most, which could be seen as a beneficial trade-off in R and $\tilde{F}$ traffic favor with their boost of performance both in throughput and FCT.

The network as a whole, regardless of the presence of Classes of Service, performs the same, which is expected, as connections occupy limited network resources. We can observe that the gain due to introduction of Classes of Service for R and $\tilde{F}$ traffic decreases with the increase of number of buffer inputs/outputs (i.e. from $n_e = 0$ towards $n_e = 2$), and for fully-buffered switch ($n_e = n_a = 8$) the gain would be 0, because no packet would ever require preemption, only bufferization. However, there are technological benefits to use small number of buffer input/outputs as it directly means simplification of switching matrix ($n_a = 8$, $n_e = 2$ means 10x10, $n_a = n_e = 8$ means 16x16 matrix) and reduction of number of burst receivers (inputs) and transmitters (outputs) for buffers. In the case of EPS, the gain would be also 0, but in general EPS entails an increase in energy consumption for OEO conversions compared to HOPS by a factor of 2 to 4 [11] on high load.

While observing the network performance overall, it's seen that introduction of Service of Classes both in OPS and HOPS helps to boost the performance for the R and $\tilde{F}$ connections, while keeping the performance of D connections

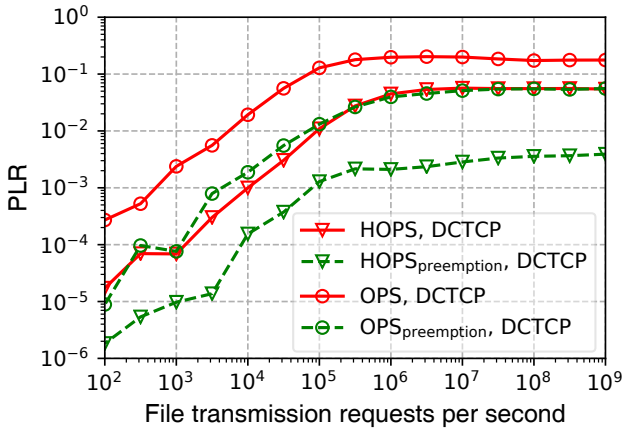

Fig. 5: Mean PLR of Reliable (R) Connections

relatively on the same level. This fact could lead to economic benefits in a Data Center: charge more priority clients for extra performance, almost without loss of it for others. Furthermore, using pure OPS instead of HOPS in DCs may be economically viable, as OPS delivers the best possible performance to R connections, on the level of HOPS performance for $\tilde{\text{F}}$ connections, and relatively low performance for D connections, since high performance may be not needed for D connections.

5 Conclusions

In this study we enhanced the analysis of HOPS and OPS DC networks by applying classes of service in terms of preemption policy for packets in optical and hybrid switches, while solving the contention problem. In the case of HOPS we demonstrated that with custom packet preemption rules, one can improve the performance for Reliable and Not-So-Fast class connections, almost without losing it for Default connections. Furthermore, we showed that Classes of Service can boost the performance of OPS for Reliable and Not-So-Fast class connections, match or bring it on the level of those in HOPS. This proves that OPS could be used in DCs, delivering high performance for certain connections, while Default class connections are still served on an adequate level.

It remains to be seen whether these results remain with a different service class distribution; and whether an actual low-latency service class can be implemented (e.g. using another protocol than TCP).

References

1. 100Gb/s Réseau Internet Adaptative (100GRIA) FUI9 project. Tech. rep. (Dec 2012)

2. IEEE standard for ethernet. IEEE Std 802.3-2015 (Revision of IEEE Std 802.3-2012) pp. 1–4017 (March 2016)
3. Agrawal, N., Bolosky, W., Douceur, J., Lorch, J.: A five-year study of file-system metadata. ACM Trans. Storage **3**(3) (2007)
4. de Almeida Amazonas, J.R., Santos-Boada, G., Solé-Pareta, J.: Who shot optical packet switching? In: Int. Conference on Transparent Optical Networks (ICTON). No. Th.B3.3 (Jul 2017)
5. Andreyev, A.: Introducing data center fabric, the next-generation facebook data center network. Online: https://code.fb.com/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/ (Nov 2014), accessed: 2018-07-17
6. Argibay-Losada, P.J., Sahin, G., Nozhnina, K., Qiao, C.: Transport-layer control to increase throughput in bufferless optical packet-switching networks. IEEE J. Opt. Commun. Netw. **8**(12), 947–961 (Dec 2016)
7. Bensley, S., Thaler, D., Balasubramanian, P., Eggert, L., Judd, G.: Data center tcp (dctcp): Tcp congestion control for data centers. RFC 8257, RFC Editor (October 2017)
8. Cheng, Q., Wonfor, A., Wei, J.L., Penty, R.V., White, I.H.: Low-energy, high-performance lossless 8x8 soa switch. In: Optical Fiber Communication Conference. p. Th4E.6. Optical Society of America (2015)
9. Dukkipati, N., McKeown, N.: Why flow-completion time is the right metric for congestion control. SIGCOMM Comput. Commun. Rev. **36**(1), 59–62 (Jan 2006)
10. Kimsas, A., Øverby, H., Bjornstad, S., Tuft, V.L.: A cross layer study of packet loss in all-optical networks. In: Proceedings of AICT/ICIW (2006)
11. Minakhmetov, A., Ware, C., Iannone, L.: Data center's energy savings for data transport via tcp on hybrid optoelectronic switches. In: Optical Fiber Communication Conference 2019. Optical Society of America, submitted, unpublished
12. Minakhmetov, A., Ware, C., Iannone, L.: Optical networks throughput enhancement via tcp stop-and-wait on hybrid switches. In: Optical Fiber Communication Conference. p. W4I.4. Optical Society of America (2018)
13. Minakhmetov, A., Ware, C., Iannone, L.: TCP congestion control in datacenter optical packet networks on hybrid switches. IEEE J. Opt. Commun. Netw. **10**(7), B71–B81 (Jul 2018)
14. Noormohammadpour, M., Raghavendra, C.S.: Datacenter traffic control: Understanding techniques and tradeoffs. IEEE Communications Surveys Tutorials **20**(2), 1492–1525 (Secondquarter 2018)
15. Rouskas, G.N., Xu, L.: Optical Packet Switching, pp. 111–127. Springer US, Boston, MA (2005)
16. Samoud, W., Ware, C., Lourdiane, M.: Performance analysis of a hybrid optical-electronic packet switch supporting different service classes. IEEE J. Opt. Commun. Netw. **7**(9), 952–959 (Sept 2015)
17. Segawa, T., Ibrahim, S., Nakahara, T., Muranaka, Y., Takahashi, R.: Low-power optical packet switching for 100-Gb/s burst optical packets with a label processor and 8 x 8 optical switch. J. Lightw. Technol. **34**(8), 1844–1850 (April 2016)
18. Shacham, A., Small, B.A., Liboiron-Ladouceur, O., Bergman, K.: A fully implemented 12x12 data vortex optical packet switching interconnection network. J. Lightwave Technol. **23**(10), 3066 (Oct 2005)
19. Takahashi, R., Nakahara, T., Suzaki, Y., Segawa, T., Ishikawa, H., Ibrahim, S.: Recent progress on the hybrid optoelectronic router. In: 2012 International Conference on Photonics in Switching (PS). pp. 1–3 (Sept 2012)

20. Takahashi, R., Nakahara, T., Takahata, K., Takenouchi, H., Yasui, T., Kondo, N., Suzuki, H.: Ultrafast optoelectronic packet processing for asynchronous,optical-packet-switched networks, Invited. J. Opt. Netw. **3**(12), 914–930 (Dec 2004)
21. Ye, X., Mejia, P., Yin, Y., Proietti, R., Yoo, S.J.B., Akella, V.: DOS - a scalable optical switch for datacenters. In: 2010 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS). pp. 1–12 (Oct 2010)