Data Center's Energy Savings for Data Transport via TCP on Hybrid Optoelectronic Switches

Artur Minakhmetov, Cédric Ware, and Luigi Iannone
LTCI, Télécom ParisTech, Université Paris-Saclay
Paris 75013, France
artur.minakhmetov@telecom-paristech.fr

***Abstract*—We report on possible 75% lower energy consumption for packet transport in data center networks replacing Electronic with Hybrid Optical Packet Switching (optical switches with a shared electronic buffer) combined with enhanced Transmission Control Protocol.**

***Index Terms*—Optical Packet Switching, Packet Switching, TCP Congestion Control, Optical Switches, Hybrid Switches, Data Centers, DC Networks**

I. INTRODUCTION

The Optical Packet Switching (OPS) technology seems to be a natural step of evolution from Electronic Packet Switching (EPS) in data networks, offering high reconfigurability, made possible through statistical multiplexing, along with efficient capacity use and limiting the energy consumption of the switches [1]. However, with traffic being asynchronous and in the absence of a technology that would make optical buffers in switches a reality, the contention issue rises, leading to poor performance in terms of Packet Loss Ratio (PLR), thus making OPS impractical. To date, several solutions have been proposed to bring the OPS technology to a functional level [1], among which we propose to combine two approaches: hybrid switches and special TCP Congestion Control Algorithms (CCA).

A hybrid switch couples an all-optical bufferless switch with a shared electronic buffer [2]. In the absence of contention, it is a cut-through all-optical switch, completely avoiding Optical-Electrical (OE) and Electrical-Optical (EO) conversions, achievable if using advanced fast (few ns) switching matrices based on Mach–Zehnder interferometers (MZI) as in [3]. Matrices based on Semiconductor Optical Amplifiers (SOA) [4] also could be used to limit the effect of OE/EO conversions. However, if contention of two (or more) packets occurs, i.e. when a packet has to use an output port already busy transmitting another packet, it is switched to a shared electronic buffer through OE conversion. When the output port is released, the buffered packet is emitted via EO conversion.

Argibay et al. [5] propose to use all-optical switches in OPS networks along with special TCP CCAs, aiming at bringing the OPS network throughput up to the levels of EPS networks, managing to reduce the effect of poor PLR. CCAs provide an intelligent control of packets sending and retransmission, according to the congestion state of the network, ensuring packet delivery. The congestion level is deduced from time-sensitive reception of mandatory packet acknowledgement (ACK) or lack thereof. The proposal leverages two families of TCP CCAs: Stop-And-Wait (SAW) with only one packet at a time in flight, and Additive Increase Multiple Decrease (AIMD) with several packets in flight.

In our previous works [6], [7] we concluded that the throughput of data center (DC) networks can benefit from this combination of TCP CCAs with hybrid switches even with few input/output (I/O) buffer ports, but our studies didn't analyze possible energy savings, compared to EPS, by having fewer OE/EO conversions thanks to OPS. In this letter we aim to address this matter.

It has been shown [8] that transport and switching can represent up to 60% of the total energy consumption in a private cloud storage service, and introducing optics in an EPS network can save about two thirds of the power [9]. An evolution from EPS towards OPS could lead to further improvements in energy consumption through limitation of OE/EO conversions on switches' I/O ports. Taking into account that a transceiver (potentially a switch's I/O port) of 10 Gb/s can spend over 80% of its power on light emission related processes [10], one sees possible energy savings with hybrid switches, that would use these transceivers only for their buffers and not on their main I/O ports. The same conclusion from [10] shows that among OE and EO conversions it is the latter that contributes more to energetic budget, and this is why we choose to base our study measurements on them.

Measurements of "transmission energy cost" in units of "bit transport energy factor" (cf. Sec. II) based on limitation of EO conversions let us conclude that 75% lower energy consumption for packet transport in data center networks is achievable if to use Hybrid Optical Packet Switching instead of EPS.

The letter is composed as follows: Sec. II discusses the choice of energy consumption metrics, Sec. III outlines study conditions, Sec. IV discusses the results obtained and, finally, Sec. V offers our main conclusions.

II. ENERGY CONSUMPTION FOR DATA TRANSPORT

In our work, we consider the following energy-consuming data transport operations to be EO conversions: initial emissions from the servers, reemissions by hybrid switches' buffers and EO conversions by I/O ports of electronic switches.

Manuscript submitted XXX X, 2018.

The authors are with the LTCI, Télécom ParisTech, Université Paris Saclay, 75013 Paris, France (e-mail: artur.minakhmetov@telecom-paristech.fr).

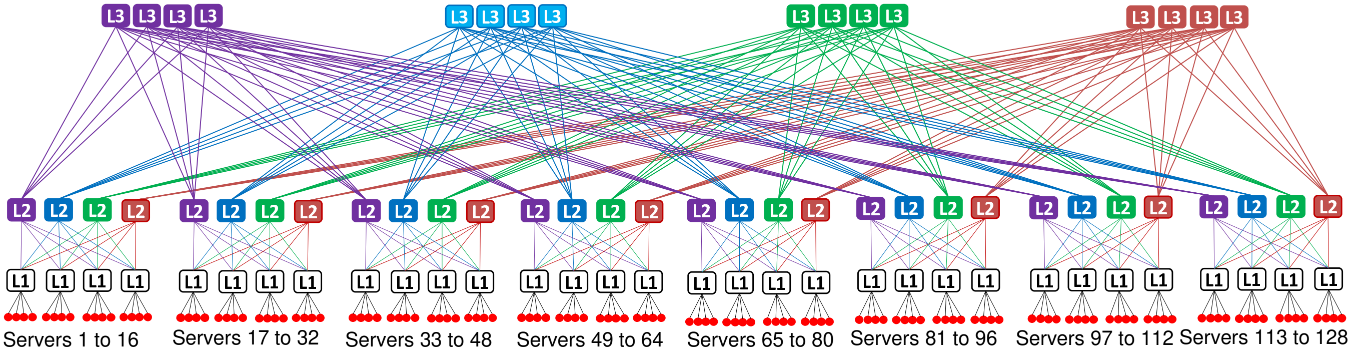

Fig. 1: Fat-tree topology data center network, interconnecting 128 servers with three layers of switches.

The choice of the metric to quantify energy gain due to reducing the number of EO conversions must be made carefully. Measurement of the EO conversions themselves might be not sufficient, as conversion of packets of different sizes will consume a different amount of energy. Measurement of bits that undergo EO conversion may be also not relevant, as it will depend on the amount of data sent on the network. Measurement of power spent on EO conversions of information bits would require choosing a specific emitter, but at the same time we want to stay as general as possible. Therefore, we choose to measure "transmission energy cost" in units of "bit transport energy factor", i.e. how many bits should be physically emitted to ensure the delivery of one bit to the destination. One can think of this value as the number of bits passed through EO conversions within the whole network normalized by the number of total bits that are to be sent into the network over a period of time. We opt for this measure which represents the energy consumption for data transport in the whole network, which can be converted to J/b simply by multiplying it with the emitter-specific consumption value, assuming that all emitters are the same. This may not necessarily be the case in a real data center, as operators can choose different emitters (especially at different bit-rates) for different switches or servers; but for this initial study we chose the simpler assumption of identical emitters everywhere.

Measurement of "bit transport energy factor" based on EO conversions suggests that optical links are active only when there is a packet to transmit. This condition is imposed by OPS and not always true for conventional EPS networks, as usually EPS maintains many point-to-point links that are always active for synchronization purposes. In order to make a fair energetic performance comparison of OPS and EPS networks, we consider the same conditions for both cases: links are active only when there is a packet to transmit, i.e. networks are asynchronous and use burst mode receivers and transmitters. In the scientific literature similar conditions were already considered for EPS point-to-point optical links with sleep mode [11] as the next step after IEEE 802.3az Energy Efficient Ethernet standard implementation. This assumption favors EPS networks in our results, as it limits the use of emitters and transmitters only to when data is being sent.

III. EXPERIMENTAL SIMULATION SETUP

We simulate the communications of DC servers by means of optical packets, for three scenarios when the network is composed of: *i)* all-optical switches, *ii)* hybrid switches and *iii)* conventional store-and-forward electronic switches. The first two cases use OPS while the last one uses EPS technology.

We developed a discrete-event network simulator [6], capable of simulating the switches described in Sec. I and also including TCP emulation. The hybrid switch has the following architecture: it has n_a azimuths, representing the number of I/O optical ports, and n_e I/O ports to the electronic buffer. The re-emission queuing strategy of buffered packets for a given azimuth is First-In-First-Out (FIFO). All-optical switches correspond to the case $n_e = 0$.

The electronic switch has a similar architecture: it has n_a azimuths, but it buffers all of the incoming packets to reemit them FIFO. All the packets undergo OE/EO conversions.

We study the DC fat-tree topology, interconnecting 128 servers by means of 80 identical switches with $n_a = 8$ azimuths, presented in Fig. 1, a sub-case of the topology deployed in Facebook's DCs [12]. Each server has 10 Gb/s network interface cards. Hybrid switches are studied with a variable number of $n_e \in \{0, 2, 8\}$ with the same bit rate. EPS switches also have the same bit rate per azimuth. All links are bidirectional and of the same length $l_{link} \in \{10, 100\}$ m, typical of DCs. Paths between servers are calculated to have the minimum number of hops, offering multiple equal cost paths, allowing load balancing: a packet has an equal probability to use any available path.

All the simulated communications consist of transmitting files between server pairs through TCP connections. The files' size is random, following a lognormal-like distribution [13]. File transmission is done by data packets of size 9 kB, i.e. Jumbo Ethernet frames, plus a 64 B control overhead. We also use 64 B for SYN, FIN, and ACK signal packets. The transmission of each data packet is regulated by the TCP CCA, which decides either to send a next packet or to retransmit an unacknowledged one. The 3-way handshake and connection termination are also emulated.

We follow a poissonian process of arrivals of new con-

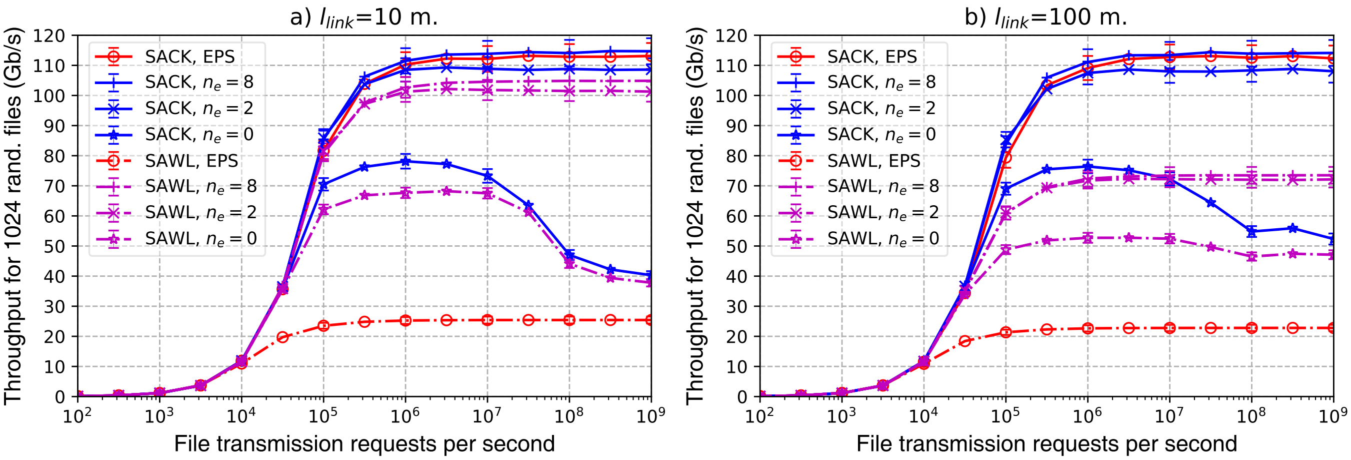

Fig. 2: Network throughput dependence on TCP CCA and switch type: a) $l_{link} = 10$ m, b) $l_{link} = 100$ m.

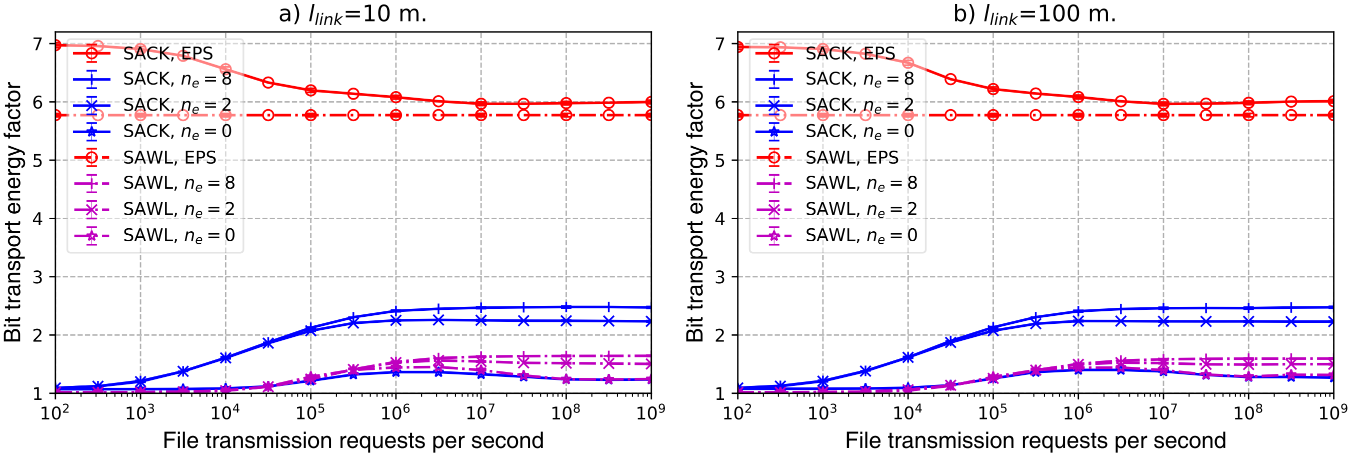

Fig. 3: Transmission energy cost dependence on TCP CCA and switch type: a) $l_{link} = 10$ m, b) $l_{link} = 100$ m.

nection demands with mean rate of given file transmission requests per second between all of the servers, so as to study the performance in terms of throughput and "transmission energy cost" of a network with different switches and protocols under progressively increasing load.

File transmission is regulated by TCP Stop-And-Wait-Longer (SAWL) [6] or TCP Selective ACK (SACK) flavor [7], which we found to be the two best-performing CCAs for hybrid switches from the SAW [6] and AIMD [7] families. For hybrid and all-optical switches, we use 1 ms as the initial value of Retransmission Time Out (RTO, the timer upon expiration of which a packet is considered lost and retransmitted), since it has proven to be favorable [5]. For the EPS case, we use 1 s for RTO initialization, as a relevant recommendation by IETF [14]: it limits unnecessary retransmissions (no packet contention) that would have been induced by 1 ms, still giving the same throughput [7]. EPS usually uses variants of AIMD, so it's sufficient to consider SACK, but for thoroughness of the study we also consider SAWL.

IV. EVALUATION OF RESULTS

To reduce statistical fluctuations, we repeated every simulation with 100 different random seeds for each pair of switch type and protocol case for $l_{link} \in \{10, 100\}$ m. The mean throughput obtained is shown in Fig. 2 and transmission energy cost in Fig. 3, with 95% t-Student confidence intervals at every second point on the graph.

For the case $l_{link} = 10$ m, as expected, EPS incurs the highest energy consumption by far, but gives almost the highest throughput using SACK, only edged out by hybrid switching with $n_e = 8$ and SACK, perhaps thanks to the cut-through nature of OPS. EPS with SAWL, which limits connections to one packet in flight, has a much lower throughput, almost ×4.5, for only a marginal energy gain. Taking EPS with SACK as a reference, at highest load, the transport-energy savings of optical and hybrid switching range between ×2.4 and ×4.8 (58–79%). For hybrid switches, the general result is that SACK gives the highest throughput but highest energy consumption, and vice-versa for SAWL, even with different values of n_e. Nevertheless, it's important to remark that SAWL with $n_e = 2$

loses only 10% of throughput to EPS and saves a factor ×4 (75%) energy-wise. If throughput is a priority, SACK with $n_e = 8$ is slightly better than EPS and still saves up to ×2.4 (58%) in transport energy.

For the case of $l_{link} = 100$ m at highest load, the energy consumption for different switches combined with CCAs is almost the same as with $l_{link} = 10$ m, but the throughput performance of SAWL and hybrid switches drops by 30%, which may make the energy savings less attractive. However, SACK gives the same throughput with hybrid and electronic switches both, allowing the same conclusion: it is still possible to save up to ×2.4 (or 58%) in transport energy consumption without losing network throughput. The drop of throughput in the case with SAWL and its absence in the case with SACK are explained by the fact that SAWL exploits link capacity less efficiently with its only one unacknowledged packet in flight, contrary to SACK with several possible packets in flight. In general, SACK with EPS decreases the energy consumption with load increase: it's explained by features of SACK, with small latencies and in absence of losses capacity of network may be overestimated, leading to congestion and thus retransmissions, adding up to a higher energy consumption.

For the general case of the all-optical switch with $n_e = 0$, we notice that the throughput decreases by $\{60 - 70\}\%$ at highest load compared to the case of $n_e = 2$, without gaining much in energy consumption.

V. CONCLUSIONS

In this paper we showed how introducing hybrid switches in DC networks related to real world cases (e.g. Facebook [12]) can decrease transport energy consumption at least by ×2 compared to electronic switches, while maintaining the same throughput. It was shown that this factor could be doubled up to ×4 while losing only 10% in throughput, thus letting us claim the beneficial character of DC network migration from EPS towards OPS on hybrid switches. DCs can benefit from hybrid switches that have a lower energy consumption than electronic, and a higher throughput and robustness than all-optical ones, using just a few electric ports and introducing specially-designed TCP protocols.

In future works we plan to investigate the concept of joint use of electronic and hybrid switches in order to study the interest in progressive integration and replacement of electronic switches with hybrid in real standard fat-tree DC networks and how this will influence throughput and energy consumption. Furthermore, we intend to not limit the study to only fat-tree networks, but also include alternative DC networks topologies. Application of hybrid switches on Metropolitan Access Networks (MAN) and influence of Wavelength Division Multiplexing (WDM) as well are in the scope of our future interests.

REFERENCES

[1] C. Ware, W. Samoud, P. Gravey, and M. Lourdiane, "Recent advances in optical and hybrid packet switching," in *Int. Conference on Transparent Optical Networks (ICTON)*, Trento, Italia, Jul. 2016.

[2] S. Ibrahim and R. Takahashi, "Hybrid optoelectronic router for future optical packet-switched networks," in *Optoelectronics - Advanced Device Structures*, S. L. Pyshkin and J. Ballato, Eds., Rijeka: InTech, 2017, ch. 04. DOI: 10.5772/67623.

[3] T. Chu, L. Qiao, W. Tang, D. Guo, and W. Wu, "Fast, high-radix silicon photonic switches," in *Optical Fiber Communication Conference*, Optical Society of America, 2018, Th1J.4. DOI: 10.1364/OFC.2018.Th1J.4.

[4] Q. Cheng, A. Wonfor, J. L. Wei, R. V. Penty, and I. H. White, "Low-energy, high-performance lossless 8×8 SOA switch," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, Mar. 2015, pp. 1–3.

[5] P. J. Argibay-Losada, G. Sahin, K. Nozhnina, and C. Qiao, "Transport-layer control to increase throughput in bufferless optical packet-switching networks," *IEEE J. Opt. Commun. Netw.*, vol. 8, no. 12, pp. 947–961, Dec. 2016.

[6] A. Minakhmetov, C. Ware, and L. Iannone, "Optical networks throughput enhancement via TCP stop-and-wait on hybrid switches," in *Optical Fiber Communication Conference*, Optical Society of America, 2018, W4I.4. DOI: 10.1364/OFC.2018.W4I.4.

[7] ——, "TCP congestion control in datacenter optical packet networks on hybrid switches," *IEEE J. Opt. Commun. Netw.*, vol. 10, no. 7, B71–B81, Jul. 2018.

[8] J. Baliga, R. W. A. Ayre, K. Hinton, and R. S. Tucker, "Green cloud computing: Balancing energy in processing, storage, and transport," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 149–167, Jan. 2011, ISSN: 0018-9219. DOI: 10.1109/JPROC.2010.2060451.

[9] N. Binkert, A. Davis, N. P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, and J. H. Ahn, "The role of optics in future high radix switch design," in *2011 38th Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2011, pp. 437–447.

[10] K. Lee, B. Sedighi, R. S. Tucker, H. Chow, and P. Vetter, "Energy efficiency of optical transceivers in fiber access networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, no. 9, A59–A68, Sep. 2012, ISSN: 1943-0620. DOI: 10.1364/JOCN.4.000A59.

[11] D. Larrabeiti, P. Reviriego, J. Hernández, J. Maestro, and M. Urueña, "Towards an energy efficient 10 Gb/s optical ethernet: Performance analysis and viability," *Optical Switching and Networking*, vol. 8, no. 3, pp. 131–138, 2011, Special Issue on Green Communications and Networking, ISSN: 1573-4277.

[12] A. Andreyev, *Introduction to Facebook's data center fabric*, Online: https://youtu.be/mLEawo6OzFM?t=175, Accessed: 2019-01-28, Nov. 2014.

[13] N. Agrawal, W. Bolosky, J. Douceur, and J. Lorch, "A five-year study of file-system metadata," *ACM Trans. Storage*, vol. 3, no. 3, 2007.

[14] V. Paxson, M. Allman, J. Chu, and M. Sargent, "Computing TCP's retransmission timer," RFC Editor, RFC 6298, Jun. 2011.