# Switching Performance Assessment for High-capacity and Large-connectivity OPSquare Data Center Network

W. Miao, F. Yan, and N. Calabretta

COBRA Research Institute, Eindhoven University of Technology, Eindhoven, the Netherlands

*The switching performance and scalability of the OPSquare data center network are investigated for different modulation formats, namely 4×25 Gb/s NRZ-OOK waveband, 28 Gb/s PAM4 and 40 Gb/s DMT traffic. By exploiting the fast WDM optical switching technology, the optical switch can achieve large port-count with lower broadcasting ratio and thus lower OSNR deterioration for the traffic. Experimental results show >8 dB input power dynamic range with <3 dB penalty for optical switches at scale of 64×64, which potentially enables a high-capacity and large-connectivity OPSquare data center network.*

## Introduction

The rise of cloud and 5G mobile computing, and big-data applications is steadily increasing the traffic within the data centers (DCs) [1]. Much more powerful servers with 100 Gb/s I/O bandwidth are expected to be deployed to provide higher processing capability [2]. As a result, the top-of-rack (ToR) switches that group tens of servers would have a multi-Tb/s aggregation bandwidth. This will put a tremendous pressure on DC network (DCN) infrastructure in which each switching node has to handle tens of Tb/s aggregation traffic. Electronic switches would hardly afford this [3], and would require stacking several switches in multi-tier structure at the expense of bottlenecked bandwidth, larger latency, and higher cost.

Switching the packets in optical domain has the advantage of data rate/format transparency, overcoming the I/O limitation of the electrical switches, the format-dependent front-end, and the power consuming O/E/O [4]. In view of this, a flat optical DCN architecture OPSquare based on buffer-less optical switches with fast nanoseconds reconfiguration time has been recently proposed [5]. By employing moderate-radix and distributed optical switches, the parallel intra/inter-cluster switching networks effectively provide large-connectivity, high-capacity and low-latency switching capability. The performance has been investigated for single wavelength OOK traffic using 4×4 optical switches. However, fueled by the demand for higher capacity and efficiency, multiple WDM channels and multi-level modulation schemes are currently highly considered for scaling the capacity of the DCN. Therefore, the impact of the OPSquare optical switches on the signal integrity and potential OSNR degradation should be properly assessed, especially when scaling the port-count of the optical switches to realize a large scale DCN.

In this work, we experimentally assess the scalability and switching performance of the OPSquare DCN architecture under different data formats, namely the waveband 4×25 Gb/s NRZ-OOK, 28 Gb/s 4-level pulse-amplitude modulation (PAM4), and 40 Gb/s discrete multi-tone (DMT) traffic. Benefitted from the fast WDM optical switching technology, large port-count can be achieved with lower broadcasting ratio and thus lower OSNR deterioration for the traffic. Experimental results show a > 8dB input power dynamic range when scaling up the optical switch port-count to 64×64 with

limited performance degradation. By using the aforementioned interconnect solutions, an OPSquare architecture interconnecting 4096 ToRs and providing Petabit/s capacity can be potentially built up for next-generation DCN.

## System operation

The OPSquare flat DCN architecture is shown in Fig. 1(a). It consists of $N$ clusters and each cluster groups $M$ racks. Each rack contains $K$ servers interconnected via an electronic ToR switch. The $N$ $M\times M$ intra-cluster optical switches (IS) and $M$ $N\times N$ inter-cluster optical switches (ES) are dedicated for intra-cluster and inter-cluster communication, respectively. The i-th ES interconnects the i-th ToR of each cluster, with i = 1,2,…,$N$. At most two-hop is sufficient to interconnect ToRs residing in different clusters and multiple paths are available for each pair of ToRs, increasing the network fault-tolerance. The network scales as $N\times M$ (square of radix), thus with $32\times32$ IS/ES, up to 1024 ToRs (40,960 servers in case $K$ = 40) can be interconnected.

The header information associated to the data packets coming from the servers is processed and the ToR will directly forward the intra-rack traffic to the destination server. For intra-cluster and inter-cluster traffic, $p$ and $q$ WDM transceivers (TRXs) with dedicated electronic buffers are utilized to interconnect with the IS and ES, respectively. The multiple ($p$ and $q$) TRXs allow for scaling the bandwidth and the number of TRXs can be tailored according to desired oversubscription and intra-/inter-cluster traffic ratio. Each TRX is only dedicated for the communication with a certain group of ToRs. For intra-cluster network, the $M$ ToRs are thus divided into $p$ groups and each group contains $F = M/p$ ToRs. One of the $p$ TRXs addresses $F$ (instead of $M$) possible destination ToRs, in cooperation with the $1\times F$ switch at IS. Therefore, each IS has a scale of $p$ $F\times p$ $F$ benefited from the grouping mechanism of the WDM TRXs.

The packet of the inter-rack traffic is buffered and a copy combined with an optical label is sent out towards the ToR destination via the optical switch. The optical switch has a modular structure and each module handles the WDM traffic from a single group, as shown in Fig.1 (b). The label extractor (LE) separates the optical label from the payload and processes it on-the-fly, while the payload is fed into the broadcast and select $1\times F$ switch. It is worth to notice that the broadcasting losses are $p$ and $q$ times lower than $1\times M$ and $1\times N$ cases which lead to less OSNR degradation and significant improvement of the scalability and feasibility. With the recovered label bits, the controller enables the SOA gates in the $1\times F$ switch to forward the packets. For each group, the wavelengths used for addressing the same destination ToR are arranged in a round-robin scheduling and will be statistically multiplexed by an F-port AWG. The contention due to the time domain collision is resolved by the switch controller in a distributed manner that the packet with higher priority is forwarded to one/multiple outputs and the others are blocked. Flow control signals (ACK/NACK) generated by

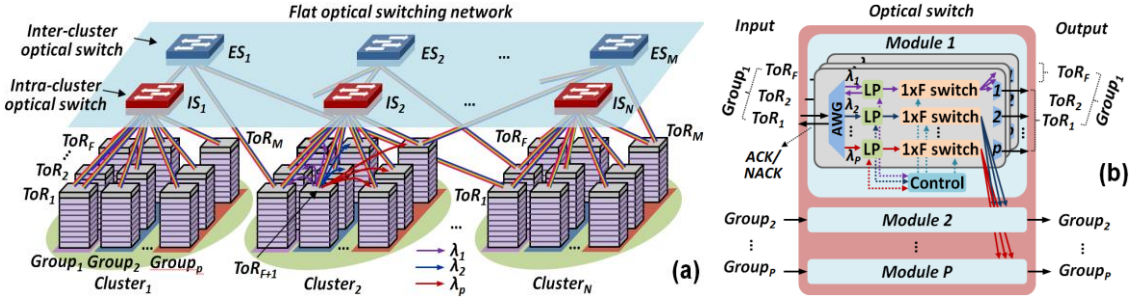


**Fig. 1:** (a) OPSquare data center network architecture. (b) Optical switch node.

controlling a direct-modulated laser (DML) are sent back to the ToRs within the same WDM channel. The packets stored in the buffers will be accordingly released (for ACK) or retransmitted (for NACK).

## Experiment results

Exploiting the flat topology and the identical modular structure of the optical switch, the switching performance in OPSquare architecture would mainly depend on and will be limited by the 1×*F* broadcast and select optical switch. By using the prototyped optical switch shown in Fig. 2(a) and the experiment setup depicted in Fig. 2(b), we assess the switching performance and the port-count scalability under different modulation formats, namely the waveband OOK, PAM4 and DMT. The broadcasting splitter is emulated by a variable optical attenuator (VOA) and another SOA has been added before the 1×*F* optical switch as pre-amplifier of the input signal. It could improve the input dynamic range and can be potentially photonic integrated with the optical switch [6]. The gating SOA in the switch prototype guarantees nanoseconds operation and compensates the broadcasting loss.

First, the waveband switching of 4 × 25 Gb/s data payload enabled by the broadband operation of the SOA-based switch is analyzed. Two DMLs and two electro-absorption modulated lasers (EMLs) with 1.6 nm spacing are driven by the 25 Gb/s PRBS $2^{11}$-1 NRZ-OOK packetized payload from the bit pattern generator (BPG) and then de-correlated to generate the waveband 4×25 Gb/s data. It is then fed into the emulated 1×*F* optical switch after power adjusted by the input VOA. The bias current of the pre-amplifier and the gate SOA have been varied to optimize the performance. At the output port, each channel is filtered out by an optical band-pass filter (OBPF) and sent to the 40 Gb/s photo-receiver.

The power penalty at BER of $1\times10^{-9}$ with different input optical power at scales of 1×8, 1×16, and 1×32 optical switch is reported in Fig. 3(a), which is corresponding to the actual scale of 32×32, 64×64, and 128×128 when employing 4 wavebands ($p$=$q$=4), respectively. 16 dB input dynamic range is achieved with less than 2 dB power penalty. Here each waveband has a 100 Gb/s capacity which can be increased by inserting more wavelength channels. The maximum number of wavebands (equivalent to $p$) and channels inside each band depends on the operation bandwidth of SOA gate and the wavelength channel spacing. With four 100 Gb/s wavebands per ToR and 64×64 optical switches (1×16 broadcasting ratio), an interconnectivity of $64^2$ = 4096 ToRs and a capacity > 3.2 Petabit/s can be achieved benefitted from the transparent optical switching and WDM TRX grouping mechanism featured by OPSquare DCN.

As promising interconnect solutions in DCN, PAM4 and DMT can effectively boost the link capacity. The 28 Gb/s PAM4 and 40 Gb/s DMT traffic are generated by driving the DML from the arbitrary waveform generator. At the switch output, the traffic is

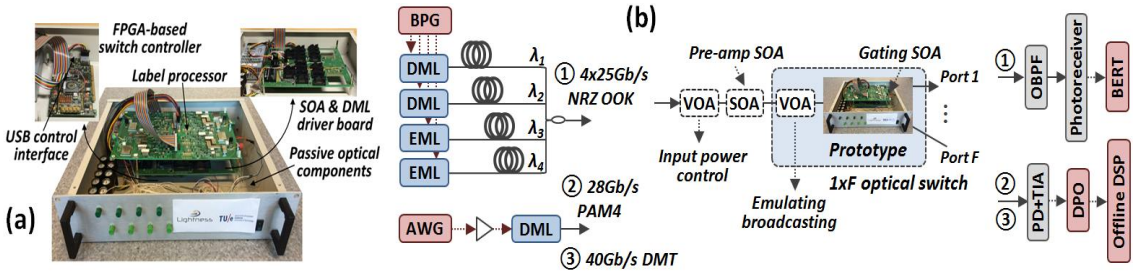


**Fig. 2:** (a) Prototyped optical switch node. (b) The experiment set-up.

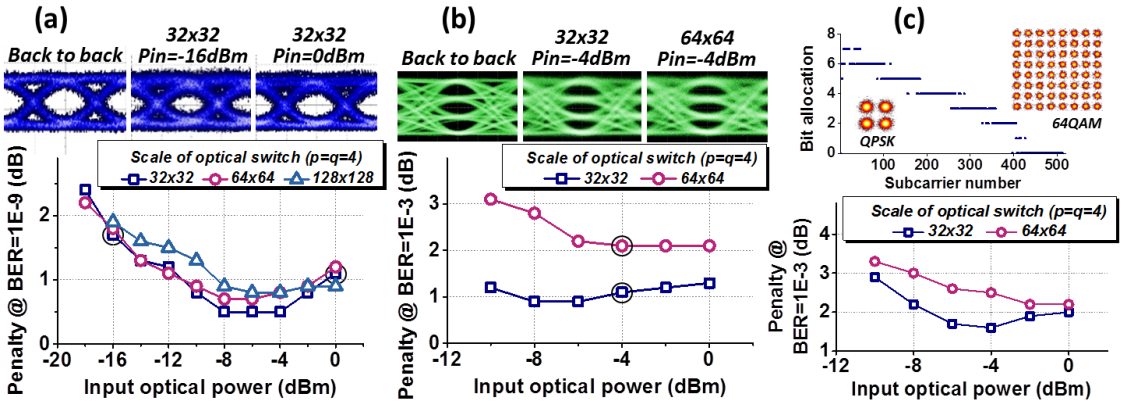


**Fig. 3:** Penalty at different scales for (a) 4×25 Gb/s waveband (b) 28 Gb/s PAM 4 and (c) 40 Gb/s DMT

received by photodiode (PD) integrated with trans-impedance amplifier (TIA), after which the signal is captured by a real-time 50 GSa/s digital phosphor oscilloscope (DPO) for offline DSP. The power penalty at BER of $1\times10^{-3}$ of the two traffic patterns measured at different input optical power within scale of 32×32 and 64×64 optical switch is depicted in Figs. 3(b) and (c), respectively. An example of the optimal bit allocation after bit loading is included in Fig. 3(c). In 32×32 scale, 10 dB dynamic range has been measured with < 3 dB penalty, while for 64×64, 8 dB dynamic range has been obtained for both traffic with 3 dB penalty. With 4 WDM transceivers ($p$=$q$=4) per ToR each operating at 40 Gb/s and 64×64 optical switches (1×16 broadcasting ratio) used, an OPSquare DCN comprising 4096 ToRs each with 320 Gb/s aggregation bandwidth would have a capacity > 1.3 Petabit/s. Larger interconnectivity can be achieved either by increasing the broadcasting ratio of the 1×$F$ switch with limited performance degradation as shown in Fig. 3, or increasing the number of transceivers per ToR which could also improve the bandwidth performance.

## Conclusion

The capability of the OPSquare architecture to switch waveband OOK, PAM4 and DMT traffic when scaling up the port-count is experimentally assessed. Experimental assessment of a 64×64 optical switch ($p$=$q$=4 and 1×16 broadcasting ratio) confirms > 8 dB dynamic range for 4×25 Gb/s waveband, 28 Gb/s PAM4, and 40 Gb/s DMT data traffic. Exploiting 64×64 optical switches, the OPSsquare could support an interconnectivity between $64^2$ = 4096 ToRs and a capacity > 3.2 Petabit/s.

## Acknowledgements

Authors would like to thank FP7 COSIGN project (n °619572) for supporting this work.

## References

[1] Cisco Global Cloud Index: Forecast and Methodology 2013–2018, Cisco Systems, USA. (2013).
[2] C. DeCusatis, “Optical Interconnect Networks for Data Communications,” J. Lightw. Technol., 32(4), 2014.
[3] A. Ghiasi, “Large data centers interconnect bottlenecks,” Opt. Express, vol. 23, no. 3, pp. 2085-2090, 2015.
[4] C. Kachris et al, Optical Interconnects for Furture Data Center Networks, Chap. 1, Springer, 2013.
[5] F. Yan et al, “Novel flat DCN architecture based on optical switches with fast flow control,” Proc. PS 2015.
[6] N. Calabretta et al, “Monolithically Integrated WDM Cross-Connect Switch for Nanoseconds Wavelength, Space, and Time Switching,” Proc. ECOC 2015.