Assessment of scalable and fast 1310-nm optical switch for high-capacity data center networks

Citation for published version (APA):

Miao, W., De Waardt, H., van der Linden, R., & Calabretta, N. (2017). Assessment of scalable and fast 1310-nm optical switch for high-capacity data center networks. *IEEE Photonics Technology Letters*, 29(1), 98-101. [7744467]. <https://doi.org/10.1109/LPT.2016.2629087>

DOI:

[10.1109/LPT.2016.2629087](https://doi.org/10.1109/LPT.2016.2629087)

Document status and date:

Published: 01/01/2017

Document Version:

Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Assessment of Scalable and Fast 1310 nm Optical Switch for High-capacity Data Center Networks

Wang Miao, Huug de Waardt, Robbert van der Linden, and Nicola Calabretta

Abstract—The development of high-capacity 1310 nm optical interconnects and the limitation of the electrical switches have necessitated the investigation of 1310 nm optical switches for flat data center networks. We have performed an experimental assessment of the scalable and fast 1310 nm optical switching system with three types of traffic, namely the waveband 4×25 Gb/s OOK, 28 Gb/s 4-level pulse-amplitude modulation (PAM4), and 40 Gb/s discrete multi-tone (DMT) by using the prototyped 1310 nm optical switch. 10 dB input power dynamic range has been achieved at 16 ports for waveband 4×25 Gb/s, 28 Gb/s PAM4 and 40Gb/s DMT with power penalty of 1 dB, 3 dB and 3.3 dB, respectively. Waveband 4×25 Gb/s OOK traffic shows the most promising results with port-count scaling up to 64 ports with limited penalty and allows for the potential scaling to larger port-count and more wavelength channels.

Index Terms— Data center network, optical switching, wavelength division multiplexing.

I. INTRODUCTION

The steady growth in the demand of smart devices with faster connectivity and diverse data-intensive applications have dramatically boosted the cloud-based IP traffic. As the major storage and computing resources, data centers (DCs) are experiencing the rapid upgrade of the servers at higher data rate ($> 100\text{Gb/s}$) [1]. Interconnecting thousands of top-of-the-rack (ToR) switches each grouping 40+ servers would require a DC network (DCN) with Petabit/s capacity, which necessitates the revolution of both optical interconnects and switching fabrics.

Multi-mode (MM) short reach optical interconnects are commonly used due to the power and cost efficiency. The advances in wide band MM fiber (MMF) and shortwave wavelength division multiplexing (SWDM) provide an economic upgrade path for another two to three generations up to 400Gbps and 1Tbps using VCSEL-MMF solution [2]. The high bandwidth distance product and spectral efficiency of the single mode (SM) fiber give rise to the employment of SM optical interconnects supporting link reach longer than 500 m [3]. In particular, 1310 nm optical interconnects are preferred especially at high data rate because of the close-to-zero

dispersion in the standard SM fiber. The continuous emergence of evolved technologies (e.g. silicon photonics) has accelerated the development of low-power and low-cost 1310 nm WDM optical interconnects [4], pushing the migration towards 400Gbps/1.6Tbps bandwidth for DC interconnects applications.

With the pavement to Tb/s interconnecting solutions, more stringent challenges are exposed to the switching fabrics. As high bandwidth is required to efficiently group several to tens of Tb/s links from the ToRs, the implementation with electronic switch is limited by the switch ASIC I/O bandwidth due to the scaling issues of the ball grid array (BGA) packages [5]. In addition, fueled by the demand for higher data-rate, it becomes beneficial for transmitters to employ WDM and multi-level modulation schemes instead of the simple on-off keying (OOK). Besides the power consuming O/E/O, dedicated parallel optics and format-dependent interfaces need to be further included as front-end, which contribute unnecessarily to the latency and power consumption.

The transparent optical switching technologies can avoid the O/E/O as well as format-dependent interface and overcome the I/O bandwidth limits of the electrical switches. Several optical switching schemes mainly based on optical circuit switching (OCS) and optical packet switching (OPS) have been investigated [6]. The millisecond reconfiguration time of OCS results in poor bandwidth utilization. The OPS can offer sub-wavelength granularity for low-latency on-demand connectivity and the lack of optical buffer has been mitigated by the implementation of optical flow control, which makes OPS more practical for DCNs [7]. A 1310 nm flow-controlled OPS with nanoseconds switching time has been proposed and preliminarily investigated with OOK payload [8]. Foreseeing the deployment of high-capacity 1310 nm optical interconnects in large scale DCN, the capability of incorporating WDM multi- λ channel and variable modulation formats with the port-count scaling are essential and should be addressed.

In this work, by exploiting the 1310 nm optical switch prototype, we have experimentally assessed the switching performance of three types of directly-modulated traffic, namely the waveband 4×25 Gb/s NRZ OOK, the 28 Gb/s 4-level pulse-amplitude modulation (PAM4), and the 40 Gb/s discrete multi-tone (DMT) featuring with intensity-modulated direct detection (IM/DD). The BER tests show less than 0.8 dB power penalty and 10 dB input power dynamic range at 16 ports for all three patterns. Port-count can be scaled up to 64 ports for waveband NRZ OOK traffic with limited extra power penalty.

Manuscript received on June 22, 2016. This work was supported by FP7 LIGHTNESS project (n°318606).

Wang Miao, Huug de Waardt, Robbert van der Linden, and Nicola Calabretta are with the Electro-Optical Communications group, COBRA research institute, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands (e-mail: w.miao@tue.nl).

Fig. 1. (a) 1310 nm optical switching system (AWG: arrayed waveguide grating, AP: ACK processor); (b) 1310 nm optical switch prototype.

II. SYSTEM OPERATION

The architecture of the investigated 1310 nm $N \times N$ fast optical switching system is shown in Fig. 1 (a). The traffic generated from the servers is grouped by N ToRs and transferred through the OPS switching node. At each ToR, the aggregation controller assigns the inter-ToR traffic (payload packet) with a certain wavelength (among $\lambda_1, \lambda_2, \dots, \lambda_M$) and attaches it with an in-band optical RF tone label. Due to the lack of the optical buffer, a copy of the transmitted packet is electrically stored at the ToR and a fast optical flow control between the OPS and the ToRs is implemented for packet re-transmission in case of contention [7].

The OPS node consists of N identical modules each operating independently. It allows for a highly distributed control which makes the reconfiguration time of the switch port-count independent. This is especially important to ensure low latency as the switch scales to large number of ports. WDM packets are de-multiplexed and processed in parallel. Payload is fed into the O-band SOA-based broadcast & select $1 \times N$ switch. Most of the SOAs are in idle mode and will only be switched on in case of forwarding packets. The module-based switch fabrics can be further photonic integrated improving the power and cost efficiency [9]. The label is extracted by a narrow-band filter such as a fiber Bragg grating (FBG), and processed by the label processor (LP) to recover the digital label bits. The switch controller detects the label bits, resolves the possible packet contention, and configures the $1 \times N$ switch to forward the packets with highest priority to one or multiple OPS output ports. In case of contention, the packet with lower priority is blocked and an optical flow control signal (negative ACK) is

sent back to the ToR side requesting for packet retransmission. A low-speed directly-modulated laser (DML) at 1310 nm driven by the switch controller is used to generate the flow control signal. This efficient optical flow control significantly improves the throughput by avoiding the loss of the contented packets, with limited deterioration of latency due to the retransmission and waiting time in the buffer [8]. The nanoseconds switching time of the SOA in combination with the parallel processing of the label bits allows a reconfiguration time of few tens of nanoseconds. Fig. 1(b) shows the developed 1310 nm 4×4 OPS prototype integrating the FPGA-based switch controller, label processor, SOA drivers and passive optical components (circulators, couplers, etc.). The gate SOAs and DMLs for generating the flow control signals are mounted on the driver board controlled by the FPGA switch controller.

III. EXPERIMENTAL ASSESSMENT

Benefited from the modular structure, the switching performance of the 1310 nm optical switch would mainly depend on the $1 \times N$ broadcast and select stage and will be limited by the splitting loss experienced by the payload. By using the experiment setup depicted in Fig. 2(a), we experimentally assess the switching performance and the port-count scalability for waveband OOK, PAM4 and DMT traffic. The broadcasting splitter is emulated by a variable optical attenuator (VOA) and another SOA has been added before the $1 \times N$ 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 [9]. The pre-amplifier SOA and gate SOA have similar gain characteristics and the G-I curve for pre-amplifier SOA is

Fig. 2. (a) Experimental set-up employing the emulated $1 \times N$ optical switch; (b) G-I curve for pre-amplifier SOA.

Fig. 3. WDM 4×25 Gb/s traffic: (a) BER curves; (b) input dynamic range; bias current and output OSNR of (c) pre-amplifier SOA and (d) gate SOA.

illustrated in Fig. 2(b).

A. Multi- λ (WDM 4×25 Gb/s) waveband switching

The promise of scaling the front-panel density, capacity and reach needed for the next-generation DCN has accelerated the deployment of WDM in DC applications (e.g. QSFP28). The waveband switching of 4×25 Gb/s data payload enabled by the broadband operation of the SOA-based switch is first analyzed. Two DMLs (NEL NLK5B5E2KA) and two electro-absorption modulated lasers (EMLs) (Multiplex MTX312EW) designed for operating at 10Gb/s and allocated at 1302.20 nm, 1303.86 nm, 1307.74 nm and 1309.43 nm are used. They 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. The extinction ratio (ER) of the output from DMLs and EMLs is 8 dB and 8.8 dB, respectively. Then it is fed into the emulated $1 \times N$ 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) with 3 dB bandwidth of 1 nm and sent to the 40 Gb/s photo-receiver.

The BER curves and eye diagrams of the four wavelength channels for a 4-port optical switch configuration (6 dB splitting loss) are presented in Fig. 3(a). The total input optical power is 0 dBm (without pre-amplifier SOA). In DCs the target BER for 100 Gigabit Ethernet (GbE) standards is 1×10^{-12} , for which less than 0.8 dB power penalty can be achieved for all four wavelength channels.

The pre-amplifier SOA is then added for the investigation of the optical power dynamic range. The splitting loss set by the VOA is 9 dB, 12 dB, 15 dB, and 18 dB to emulate the scale of 8, 16, 32, and 64 ports, respectively. The power penalty measured at BER of 1×10^{-12} when varying the input optical power is plotted in Fig. 3(b). The optimal input optical power is around

-5 dBm and both lower and higher input power result in higher penalty mainly due to the noise and saturation, respectively. Scaling the port-count from 8 to 64 results in < 0.8 dB extra penalty and smaller dynamic range due to the inadequate power level at the switch output caused by the increased splitting loss. For the 64-port case, 12 dB dynamic range is achieved with < 1.5 dB power penalty. Figs. 3(c) and (d) depict the optimal bias current and the output OSNR for both SOAs. For pre-amplifier SOA, lower input power requires higher current to boost the signal and thus avoiding further OSNR degradation due to the low input power at the gate SOA. The gate SOA allowing loss-less operation provides amplification especially for lower input power and larger scale cases. The OSNR of the gate SOA output also confirms the deterioration of the signal (from 32 dB to 27.5 dB at 0 dBm input) caused by the broadcasting stage. Scaling to larger port-count and more wavelength channels are expected to further increase the capacity at the expense of limited additional power penalty and smaller dynamic range.

B. Switching of 28 Gb/s PAM4 traffic

PAM4 with forward error correction (FEC) can effectively increase the transmission bitrate in an IM/DD manner with low-complexity and real-time DSP, making PAM4 a promising candidate for short-reach interconnects solutions. Due to the bandwidth limitation of the DML, the switching performance of 14 Gbaud/s (28 Gb/s) PAM 4 traffic has been evaluated, instead of 50 Gb/s being considered for 200 GbE and higher speed. As shown in Fig. 2(a), 28 Gb/s PAM4 traffic with ER of 9.2 dB is generated by driving the DML from a 60 GSa/s arbitrary waveform generator (AWG). At the switch output, the traffic is received by PIN 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 further offline digital signal processing (DSP). The DML and PD+TIA are commercially mature 10G

Fig. 4. 28 Gb/s PAM4 traffic: (a) BER curves; (b) input dynamic range; bias current and output OSNR of (c) pre-amplifier SOA and (d) gate SOA.

Fig. 5. 40 Gb/s DMT traffic: (a) bit allocation per subcarrier; (b) BER curves; (c) input dynamic range; (d) throughput vs. received optical power.

class components. Digital pre-compensation is applied at the AWG and digital filtering is applied at the receiver side.

The BER curves of the switched 28 Gb/s PAM4 traffic as well as the eye diagrams for a 4-port switch configuration are shown in Fig. 4(a). The input optical power is 0 dBm. A power penalty of 0.8 dB at typically targeted BER (pre-FEC) of 2.4×10^{-4} has been observed, which is also confirmed by the eye diagrams. The optical power dynamic range has been investigated for the 8, 16, and 32 ports cases. The power penalty measured at BER of 2.4×10^{-4} when varying input power is depicted in Fig. 4(b). The optimal bias current and the output OSNR for the pre-amplifier and the gate SOAs are presented in Figs. 4(c) and (d). Due to the low OSNR tolerance, the OSNR degradation caused by the high splitting loss has greatly affected the power penalty that ~ 3 dB extra penalty has been observed from 8 to 32 ports. 8 dB dynamic range is obtained within 3.2 dB power penalty for the 32-port case and larger port-count is achievable with large penalty (>3 dB).

C. Switching of 40 Gb/s DMT traffic

DMT can make optimal utilization of the bandwidth by adapting the modulation format and power of each subcarrier to the characteristics of the transmission channel. As illustrated in Fig. 2(a), an AWG with a sample rate of 24 GSample/s drives the DML to generate the DMT traffic. The switch output is received and captured by the DPO for offline DSP. A probe signal with uniform power and QPSK bit loading is first sent through the optical switching system and the signal-to-noise ratio (SNR) at each subcarrier is estimated. Then the bit loading algorithm is used to obtain the bit and power allocation for the evaluated DMT signal. Here 512 subcarriers are used, targeting for a line rate of 40 Gb/s. An example of the optimal bit allocation after bit loading is shown in Fig. 5(a). It can be seen that the modulation format changes from 128-QAM to BPSK as the SNR decreases towards high-frequency subcarrier.

The BER curve for the 4-port configuration is depicted in Fig. 5(b). The power penalty at BER of 1×10^{-3} is less than 0.6 dB after the switching system. The optical power dynamic range for scales of 8 and 16 ports targeting a line rate of 40 Gb/s and BER of 1×10^{-3} is then investigated. The power penalty as a function of input optical power is presented in Fig. 5(c) which experiences a fast increase when reducing the input optical power lower than -6 dBm. 16 ports could assure an 8 dB dynamic range within 3 dB power penalty. Despite the loss-less operation guaranteed by the gain compensation, the high power

level needed at the receiver side has led to a limited dynamic range and poor power-efficiency. Larger port-count would cause worse performance exposing more pressure on the link power budget. Fig. 5(d) reports the flexible data rate adaption enabled by the DMT system. The target BER has been set to 1.5×10^{-3} and the received optical power is varied from -3 dBm to 0 dBm. By updating the optimal bit allocation, the bit rate can be dynamically adjusted with an increase from 41 Gb/s to 43.5 Gb/s (corresponding to 40.455 Gb/s net bit rate with 7% hard-decision forward-error-correction).

IV. CONCLUSIONS

By exploiting the 1310 nm optical switch prototype, the switching performance and port-count scalability of waveband 4×25 Gb/s, 28 Gb/s PAM4, and 40 Gb/s DMT traffic are investigated. Results show 10 dB input power dynamic range at 16 ports for all three patterns with power penalty of 1 dB, 3 dB and 3.3 dB, respectively. The waveband 4×25 Gb/s NRZ-OOK traffic exhibits most promising capability facing port-count scaling with less than 0.8 dB extra power penalty from 8 ports to 64 ports. Larger port-count and more wavelength channels are potentially achievable. PAM 4 and DMT traffic still hold the promise of further increasing the capacity but at the expense of larger power penalty. The obtained results can be used as reference for designing photonic integrated switches with smaller footprint and lower power consumption.

REFERENCES

- [1] The 2015 Ethernet Roadmap, Ethernet Alliance white paper, 2015.
- [2] Y. Sun, R. Shubochkin, and B. Zhu, "New development in optical fibers for data center applications," in *Proc. SPIE*, vol. 9387, p. 938705, 2015.
- [3] P. Dong, G. Duan, Y. Chen and D. T. Neilson, *Silicon Photonics III* (Springer, 2016), Chap. 14.
- [4] B. R. Koch, E. J. Norberg, B. Kim, J. Hutchinson, J.-H. Shin, G. Fish and A. Fang, "Integrated silicon photonic laser sources for telecom and datacom," *OFC 2013*, Paper PDP5C.8, Anaheim, CA, USA, 2013.
- [5] A. Ghiasi, "Large data centers interconnect bottlenecks," *Opt. Express*, vol. 23, no. 3, pp. 2085-2090, 2015.
- [6] C. Kachris, K. Bergman, and I. Tomkos, *Optical Interconnects for Future Data Center Networks* (Springer, 2013), Chap. 1.
- [7] W. Miao, S. D. Lucente, J. Luo, H. Dorren, and N. Calabretta, "Low latency and efficient optical flow control for intra data center networks," *Opt. Express*, vol. 22, no. 1, pp. 427-434, 2014.
- [8] W. Miao, F. Agraz, H. de Waardt, S. Spadaro, H. J. S. Dorren, and N. Calabretta, "1.3 μ m SDN-enabled Optical Packet Switch Architecture for High Performance and Programmable Data Center Network," *OFC 2015*, Paper Th2A.66, Los Angeles, CA, USA, 2015.
- [9] N. Calabretta, K. Williams, and H. Dorren, "Monolithically Integrated WDM Cross-Connect Switch for Nanoseconds Wavelength, Space, and Time Switching," *ECOC 2015*, Paper Mo.3.2.2, Valencia, Spain, 2015.