To overcome the scalability limitation of passive optical interconnects in datacentres

To Overcome the Scalability Limitation of Passive Optical Interconnects in Datacentres

Rui Lin[1,4], Krzysztof Szczerba[2], Erik Agrell[3], Lena Wosinska[1], Ming Tang[4], and Jiajia Chen[1]
[1]*School of Information and Communication Technology, KTH Royal Institute of Technology, Electrum 229, Kista, Sweden,*
[2]*Department of Microtechnology and Nanoscience, Chalmers University of Technology, Göteborg, Sweden,*
[3]*Department of Signals and System, Chalmers University of Technology, Göteborg, Sweden,*
[4] *Next Generation Internet Access National Engineering Lab, School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan, China.*
jiajiac@kth.se

Abstract: We propose to add optical amplifier(s) to passive optical interconnect (POI) at top-of-rack (ToR) in datacentres and validate this approach by introducing impairment constraints into POIs design. It is shown that one amplifier can improve scalability by factor of 16.

OCIS codes: (060.4510) Optical communications; (060.4250) Networks

1. Introduction

The volume of datacentre traffic is ever-growing, causing a serious bottleneck in terms of bandwidth and energy consumption. To address this problem, optical communication and switching techniques are introduced in datacentres. However, commodity switches are still widely used at the top-of-rack (ToR). They consume a lot of energy and also limit capacity upgrading. Therefore, a coupler-based passive optical interconnect (POI) has been proposed [1–3] to replace the commodity switch at ToR, leading to an overall reduction of energy consumption of switching equipment in datacentres by a factor of 10. The gain in energy consumption increases when a higher data rate on a per-server basis is considered [2]. Apart from the advantages of capacity and energy consumption, the coupler-based ToR interconnect can also offer high reliability and cost efficiency [4] thanks to its passive manner. However, the design of existing coupler-based POI architectures [1–4] does not consider the physical layer impairments which may significantly affect the quality of the signals. If the received signal power fails to reach the receiver sensitivity threshold, the transmitted data cannot be detected. As a result, the reserved resources are wasted. Therefore, it is of key importance to take the physical layer impairments into account and analyze their impact on the feasibility of the POI architectures. In this paper, we focus on the intra-rack communications. A thorough analysis of physical layer impairments is carried out considering the coupler-based POI at ToR in order to quantify its scalability. It has been shown that with introduction of an optical amplifier to increase the optical power budget, the interconnect size can be increased by a factor of more than 10, making it possible to meet the scalability requirement at ToR [5].

2. Passive optical interconnect architecture

The authors of [2, 4] proposed a POI architecture based on an Nx2 coupler, as shown in Fig. 1(a). It takes advantage of the high capacity of the dense wavelength division multiplexing (DWDM) technique and the broadcast-and-select nature of optical couplers to handle the bursty and multicast traffic in datacentres. In this approach, the optical network interfaces (ONIs) are dedicated to each server and connected to the side of the coupler with N ports. The transmitted signals pass through the coupler and are selected by a wavelength selective switch (WSS) according to whether the destination is inside or outside the rack. For the intra-rack traffic, see the red arrow in Fig. 1, the signals are looped back via an isolator (ISO) to the coupler and broadcast to all the servers. The optical tunable filter (OTF) at the receiver selects the wavelength of interests and the signal is detected by a positive-intrinsic-negative (PIN) diode. This architecture operates in the 1550 nm band with single-mode fibre to enable DWDM. The anticipated cost reduction of silicon photonic techniques make the C and L bands possible for the practical use in datacom in the near future.

3. Physical layer impairments

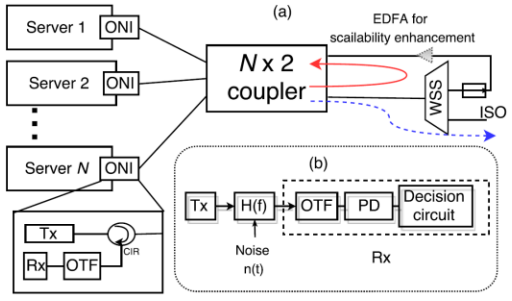

Fig. 1. (a) The basic coupler-based optical interconnect at ToR; (b) The equivalent block diagram for the link within the coupler-based optical interconnects.

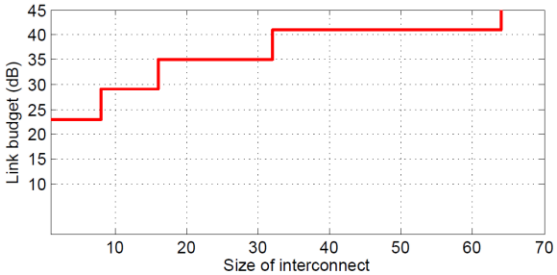

Fig. 2. The optical link budget for intra-rack communication as a function of the size of interconnect N.

In this section, two indicators are defined to assess the scalability of the interconnect architecture for the intra-rack communication: the number of servers that can be connected (i.e., the size of the interconnect N) and the highest data rate that can be supported under a certain transmission quality. As a case study, 10 GHz bandwidth transceiver with multi-level pulse amplitude modulation (M-PAM) are considered.

(a) Power loss

Fig. 1(b) shows the equivalent block diagram of the intra-rack communication links. Servers are connected to the ToR with less than 10 m cables/fiber. With such a short distance, the effects of the chromatic dispersion and nonlinearities on the signal would be negligible. Therefore, the main signal impairment to be considered for transmission is the power loss. For N servers accommodated in a rack, the splitting loss of the coupler in dB is $IL_{\text{coupler}} = 6 \lceil \log_2 N \rceil$. The total power loss in the link can be written as $IL_{\text{total}} = IL_{\text{coupler}} + IL_{\text{WSS}} + IL_{\text{ISO}} + IL_{\text{CIR}} + IL_{\text{OTF}}$. The value of IL_{WSS}, IL_{ISO}, IL_{CIR} and IL_{OTF} are set to 2 dB, 0.4 dB, 0.6 dB and 0 dB respectively. With 2 dB margin, the relationship between the size of the interconnect and the optical link budget is illustrated in Fig. 2.

(b) Receiver noise

Apart from the attenuation in the link, the receiver noise also affects the signal. At the receiver side, the OTF filters out the wavelength of interests and the optical power is then converted to a photocurrent using a PIN diode. The converted photocurrent is $I_{\text{opt}} = R_{\text{d}} \cdot P_{\text{opt}}$, where R_{d} is the responsivity of the photodiode and P_{opt} is the received optical power. In this case the noise has three contributions: thermal noise, shot noise and relative intensity noise (RIN) [6]. The variance σ^2 of these three noises is constant, linear, and quadratic in I_{opt}, resp., as shown in Fig. 3.

Considering a 10 GHz transceiver working at 1550 nm with a laser RIN of −145 dB/Hz and receiver responsivity 1 A/W, when the received optical power is below −4 dBm the total noise is determined by the thermal noise, while the RIN becomes dominant at higher power.

Using regular M-PAM with Gray labeling, the BER can be approximated as a function of the symbol error rate (SER)

$$BER \approx \frac{SER}{\log_2 M} = \frac{1}{M \log_2 M} \sum_{i=0}^{M-1} \sum_{j=0, j\neq i}^{M-1} P_{ij}, \quad (1)$$

where P_{ij} is the probability that transmitted PAM symbol i is received as symbol j. It can be expressed as

$$P_{ij} = \frac{1}{2}\text{erfc}\left(\frac{I_{th,j} - I_i}{\sigma_i \sqrt{2}}\right) - \frac{1}{2}\text{erfc}\left(\frac{I_{th,j+1} - I_i}{\sigma_i \sqrt{2}}\right), \quad (2)$$

where I_i is the photocurrent of symbol i, $I_{\text{th},j}$ denotes the photocurrent threshold between symbols j and j−1 and σ_i^2 is the noise current variance at symbol i5. The thresholds $I_{\text{th},j}$ are chosen as $I_{\text{th},j} = (I_j + I_{j-1})/2$, for j = 1,⋯, M−1, $I_{\text{th},0}$ should be interpreted as $-\infty$ and $I_{\text{th,M}}$ as $+\infty$.

When the dominant noise is the thermal noise, which is independent of the optical signal level, the minimum BER is achieved with equidistant symbols and decision thresholds halfway between the symbols. The BER performance of M-PAM as a function of the received optical power is shown in Fig. 4. Four cases are calculated, 2-PAM (OOK), 4-PAM, 8-PAM, and 16-PAM. For error-free transmission, the receiver sensitivity of a 10 GBaud OOK signal is around −16 dBm. An error floor can be observed when 16-PAM is used. That is because the RIN dominates the noise when the received power is high.

4. Scalability analysis

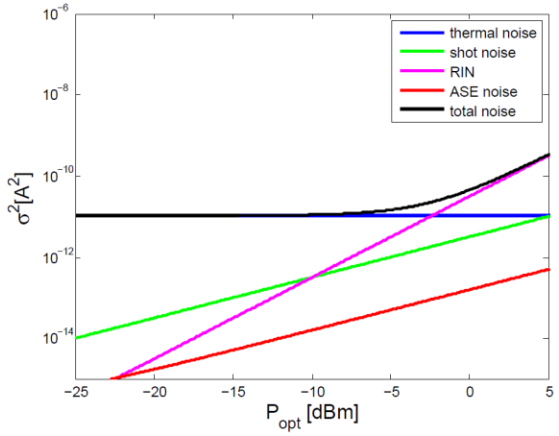

Fig. 3. Thermal noise, shot noise, relative intensity noise, ASE noise and total noise for the given system with a 10 GHz bandwidth PIN diode at the receiver.

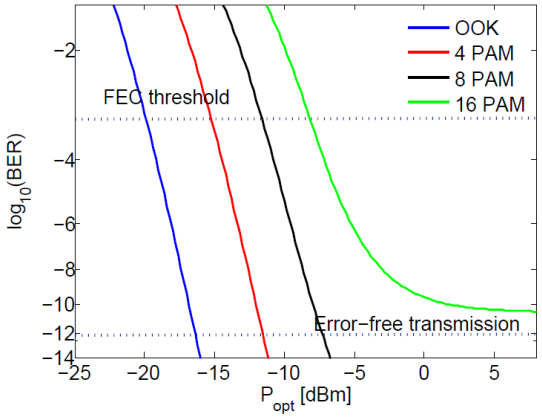

Fig. 4. BER performance of 10 GBaud M-PAM with a PIN diode at the receiver.

Based on the transmission performance evaluation results, we carry out a scalability analysis to identify the maximal data rate and the interconnect size that can be handled by the coupler-based POI. The maximum allowed power budget is equal to the difference between the launch power and the lowest receiver sensitivity that can be achieved by any modulation format. Given an interconnect size N, the scalability of the interconnect is demonstrated when the maximum allowed power budget is larger than the link budget. Assuming a launch power of 10 dBm, we illustrate the relationship between the size of the rack and the maximum supported data rate in Fig. 5. In the error-free transmission case, concurrent 20 Gb/s 4-PAM signals from 8 servers reaches the upper bound of the scalability of the interconnect. The use of forward error correction (FEC) can improve the power budget leading to doubled rack size when the BER reaches the FEC limit of 10^{-3}. However, without additional consideration the original architecture is not feasible, because it is unable to host 40 to 60 servers which is a typical size of a rack in current datacentres.

5. Scalability analysis

To improve the scalability of the architecture, we introduce an erbium-doped fibre amplifier (EDFA) at the ToR right after the WSS in the loop-back link for the intra-rack signal (see Fig. 1). The increase of cost and energy consumption of the architecture is not significant since only one additional EDFA is needed for the whole rack. In a rack hosting 60 servers, the cost increases by only 0.6% and power dissipation increases by only 1% according to the input data from2, 7. On the other hand, the very bursty traffic in the rack may lead to fast on/off changes in the channels, which potentially may result in power/gain excursion, transmission performance degradation and overload of the receivers. Fortunately, the transients can be suppressed by some techniques, e.g., automatic gain control8.

The employment of the EDFA boosts the maximum allowed power budget for the POI, but on the other hand adds amplified spontaneous emission (ASE) noise [6]. Considering a commercial EDFA with gain 20 dB and noise figure 6 dB, while the OTF is with 3 dB bandwidth of 10 GHz, the ASE noise can be obtained as a function of the received optical power, which is shown in Fig. 3. As the dominant noise at the receiver is consistent with the case without EDFA, the transmission performance of M-PAM in the updated architecture remains equal to the basic one. Thanks to the gain of the EDFA, the scalability of the coupler-based POI is significantly improved, as shown by the green dashed lines in Fig. 5. In error-free transmission, up to 128 servers transmitting 20 Gb/s 4-PAM signals can be interconnected. When FEC is included, the EDFA boosts the interconnect capability up to 256 servers. The peak switching capacity is improved from 160 Gb/s to over 2 Tb/s.

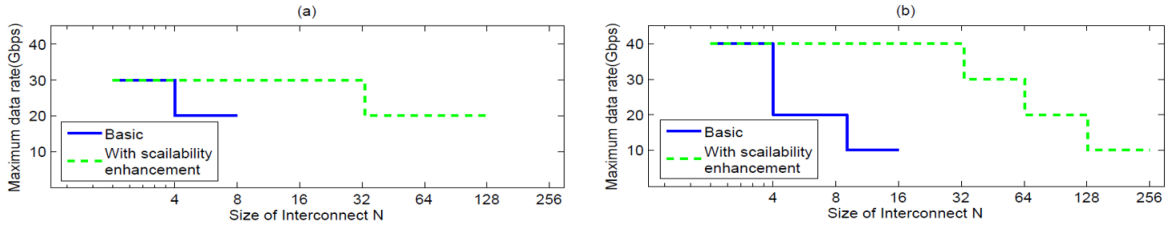

Fig. 5. The relationship between the size of interconnect *N* and the maximum data rate (a) in error-free transmission (BER=10^{-12}) and (b) under a FEC limit threshold (BER=10^{-3})

6. Conclusions

We introduced a physical layer impairment analysis in the design of a coupler-based ToR architecture and quantified its scalability. We found that without any additional strategy to enhance the scalability, the basic POI architecture may become infeasible. Placing a single EDFA improves the POI scalability by a factor of 16 while maintaining the cost and energy consumption in a similar level.

Acknowledgements

This research was supported by the China Scholarship Council, Swedish Research Council, Swedish Foundation for Strategic Research, Göran Gustafsson Foundation and National Natural Science Foundation of China (Grant No. 61550110240 and No. 61331010).

References

[1] W. Ni *et al.,* "POXN: A new passive optical cross-connection network for low-cost power-efficient datacenters," J. *Lightw. Technol.*, vol. 32, pp. 1482–1500, 2014.

[2] M. Fiorani *et al.,* "Energy-efficient elastic optical interconnect architecture for data centers," *IEEE Commun. Letters*, vol. 18, pp. 1531–1534, 2014.

[3] J. Chen *et al.,* "Optical interconnects at top of the rack for energy-efficient datacenters," *IEEE Commun. Mag.*, vol. 53, pp. 140–148, 2015.

[4] Y. Cheng *et al.,* "Reliable and cost efficient passive optical interconnects for data centers", *IEEE Commun. Lett.*, vol. 19, pp. 1913–1916, 2015.

[5] C. Kachris *et al.,* "Optical Interconnects for Future Data Center Networks", Springer, 2013.

[6] G. P. Agrawal, Fiber-optic communication systems. John Wiley & Sons, 2002

[7] MRV Communications Inc., datasheet, available online, http://www.mrv.com/sites/default/files/datasheets/us_pdfs/mrv-fd-edfa_2.pdf

[8] H. S. Carvalho *et al.,* "AGC EDFA transient suppression algorithm assisted by cognitive neural network," Int. Telecomm. Symp., 2014.