Physical-layer network coding for passive optical interconnect in datacenter networks

RUI LIN,1,2 YUXIN CHENG,2 XUN GUAN,3,4,6 MING TANG,1 DEMING LIU,1 CHUN-KIT CHAN,3 AND JIAJIA CHEN2,5

1National Engineering Laboratory for Next Generation Internet Access System, School of Optics and Electronic Information, Huazhong University of Science and Technology, Wuhan, 430074, China
2School of ICT, KTH Royal Institute of Technology, Isafjordsgatan 22, Electrum 229, 164 40 Kista, Sweden
3Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
4Centre d’Optique, Photonique et Laser (COPL), Department of Electrical and Computer Engineering, Université Laval, Québec, QC, G1V 0A6, Canada
5jiajiac@kth.se
6guanxun@ie.cuhk.edu.hk

Abstract: We introduce physical-layer network coding (PLNC) technique in a passive optical interconnect (POI) architecture for datacenter networks. The implementation of the PLNC in the POI at 2.5 Gb/s and 10Gb/s have been experimentally validated while the gains in terms of network layer performances have been investigated by simulation. The results reveal that in order to realize negligible packet drop, the wavelengths usage can be reduced by half while a significant improvement in packet delay especially under high traffic load can be achieved by employing PLNC over POI.

© 2017 Optical Society of America

OCIS codes: (060.2330) Fiber optics communications; (200.4650) Optical interconnects.

References and links

1. Introduction

To meet high capacity and low energy consumption requirements of datacenter networks (DCNs), passive optical interconnect (POI) is a promising candidate for interconnections among different servers [1]. It is because the POI is able to eliminate or reduce the optical-electrical-optical conversions, which are not only power-consuming but also the bottleneck for capacity upgrade. Basically, it mainly employs passive optical components, such as arrayed waveguide grating (AWG) [2,3] and optical couplers [4,5]. Compared to the AWG-based solutions, high insertion loss of the optical couplers may result in inferior scalability of the coupler-based POIs. This shortcoming can be overcome by improving the link power budget, e.g., via digital signal processing techniques [6], or adding amplifiers [7]. On the other hand, without the need of any wavelength-sensitive components for interconnection, the coupler-based POI architectures outperform the AWG-based ones, in terms of flexibility in wavelength resource allocation. In addition, the coupler-based POIs have a broadcast-and-select feature, which inherently supports multicast to handle concurrent traffic flows generated by the servers in datacenters [8]. Meanwhile, to enable non-collision communications, distributed [4] as well as centralized [5,9] control protocols are proposed. The scheduling is carried out in both wavelength and time domains such that short packet delay (in the magnitude of microseconds) can be achieved. However, as discussed in [9], it is difficult to achieve good performance in terms of latency and throughput without a sufficient number of wavelengths in coupler-based POIs. This requires the transceivers of wide operating spectrum range, which increase the cost and the complexity of system design.

Network coding [10] was presented as a method to increase the information flow between different network elements in a multicast environment by exploiting the function of encoding at a node. Physical-layer network coding (PLNC) accomplishing the network coding operation in the physical-layer [11], i.e., taking the advantage of the natural superimposition of electromagnetic waves [12], and can be applied for both wireless [13] and optical fiber communications [14]. In the coupler-based POIs, the optical signals sent by different input ports are naturally power-combined at the optical coupler, which can be utilized as a coding process facilitating the implementation of the PLNC. It allows the same wavelength channel to work for full-duplex communications between a pair of nodes. Thus it has a great potential to save the spectrum resource.

In this paper, we introduce the PLNC in the coupler-based POIs for the DCNs, carry out an experimental validation on transmission performance and quantify its possible gains in network performance in terms of latency and packet drop ratio. The remainder of this paper is organized as follows. In Section 2, we elaborate the implementation of the PLNC in the coupler-based POI architecture for the DCNs. In Section 3, experiments are carried out to verify acceptable transmission performance can be achieved when the PLNC is implemented.
Based on this, we study the tailored bandwidth allocation scheme in Section 4 and demonstrate the benefits of the PLNC in terms of latency and throughputs. Finally, we conclude the paper in Section 5.

2. Principle of operation

In this section, a coupler-based POI architecture, implemented at top-of-rack (ToR) shown in Fig. 1(a), is considered to interconnect different servers directly. C-band transceivers and single-mode fiber (SMF) are deployed in the POI to support DWDM and provide high capacity and interoperability. On the data plane, each server is assigned with a two-port optical network interface (ONI), which is composed of a pair of tunable optical transceivers. N ONIs are connected to an \((N + 1) \times (N + 1)\) coupler via two separate pieces of fibers, keeping the signal propagation only in one direction. The coupler reserves one port for connecting with a wavelength selective switch (WSS), which assures that only the wavelengths carrying inter-rack traffic can pass the interface to the outside of the rack. The traffic that destined inside the rack is broadcast to all the servers and selected by the tunable receivers in the ONIs at the servers. Regarding the control plane, either an out-of-band (e.g., [9,15]) or an in-band (e.g., [4]) management can be applied. In Fig. 1(a), the out-of-band scheme, where all the servers are connected to the ToR controller by the transceivers dedicated for the control signaling, is considered. Small factor pluggable (SFP), as a simple solution, is shown as an example of the transceiver for control signaling in Fig. 1(a).

![Fig. 1. (a) A coupler-based POI architecture at ToR, (b) operation principle of the conventional POI, and (c) operation principle with the PLNC over the coupler-based POI.](image)

Figure 1(b) illustrates the operation principle of the POI without implementing the PLNC, taking a six-port POI, which has simultaneous mutual communication requests between Port 1 and 2, Port 3 and 4, Port 5 and 6, as an example. Two wavelength channels are needed for each simultaneous mutual communication request to avoid the conflict at the reception in the POI. With the implementation of the PLNC, as shown in Fig. 1(c), the optical signals sent by different ports are inherently combined at the optical coupler that can be utilized as a coding process. A simultaneous mutual communication request can be served by using a single wavelength channel. For instance, data sent from Port 1 to Port 2, \(D_{1,2}\), and the one from Port 2 to Port 1, \(D_{2,1}\), can be carried by the same wavelength channel. They are combined as \(D_{\text{PLNC}} = D_{1,2} + D_{2,1}\) in the optical coupler. At the reception, Port 1 performs the decoding process by subtracting its transmitted data \(D_{1,2}\) from the received network coded signal, i.e., \(D_{2,1} = D_{\text{PLNC}} + (-D_{1,2})\). Compared to the case without the PLNC, wavelength usage is reduced by half. It should be also noted that a buffer and a decoder are needed for the decoding process. The decoder can be realized by an inverter coupled with an electrical combiner, which are commonly used functions in digital circuits. Besides, the decoding can be realized in analogue domain so that the extra delay and power consumption can be minor.
3. Transmission performance

![Fig. 2. The experimental setup to implement the PLNC over the coupler-based POI architecture. Inset (a) the eye diagram of the original signal $D_{12}$. Inset (b) the combined optical signal. EDL: electrical delay line, IM intensity modulator, VOA: variable optical attenuator, EATT: electrical attenuator, LPF: low pass filter, BERT: bit-error-rate tester.]

There have been a few PLNC schemes for optical fiber communications reported in literature. By the means of multiplexing different polarization states, the PLNC implementation could be realized through carrying the two-way signals on exactly the same wavelength in both direct detection systems [16] as well as coherent systems [14,17]. In [18], the authors demonstrated a PLNC scheme in a time-division multiplexing passive optical network (TDM PON) to realize inter-optical network unit (ONU) communications. The signals from the two ONUs were carried by two optical carriers the channels having central carriers that are not exactly the same. Compared to the other PLNC schemes, e.g [12,14], this scheme does not need to deal with different polarization states and has a simple system configuration, thus is could be more suitable for datacenter applications. However, as presented in [16], the coded two-way signals were required to have a channel spacing of 3.27 nm, which could not fit in one DWDM grid, though DWDM can provide a sufficient number of wavelength channels for high scalability and flexibility in the coupler-based PONs [9]. With this in mind, we carry out the experiments to assess the feasibility of implementing the PLNC scheme as proposed in [16] in a DWDM system. To be more specific, we are exploring the minimum frequency spacing required for the two-way communications while keeping acceptable transmission performance of the decoded signals.

Figure 2 shows the experimental setup for the bi-directional communications between a pair of servers that employ the PLNC technique. At Server 1 (S1) and Server 2 (S2), the continuous wave (CW) lights are respectively intensity-modulated with non-return-zero (NRZ) pseudo random binary sequences (PRBSs) to form the signals $D_{12}$ and $D_{21}$. They are combined at the optical coupler realizing the coding process inherently. Here, we do not put any extra fiber in between the two servers since the test case is to simulate for the intra-rack communications, which typically have a rather short transmission distance (less than 10 meters). The variable optical attenuator (VOA) is employed for measuring the receiver sensitivity. The decoding process is realized at the receiver. We form the two servers in peer and validate the performance for one of the receivers, i.e., S1 in the experiment. The combined signal $D_{PLNC}$ is broadcast and then detected by a $p$-$i$-$n$ diode before being fed into an electrical inverter.

This output electrical signal is then combined with a copy of the original transmitted signal, i.e., $D_{12}$, via an electrical combiner (EC). An electrical attenuator (EATT) is used to adjust the voltage of $D_{12}$ and an electrical delay line (EDL) is used to mitigate the misalignment between $D_{12}$ and $D_{PLNC}$. The decoded electrical signal is then fed into a low pass filter (LPF) before its signal performance is evaluated at a bit-error-rate tester (BERT). The wavelength separation between the two optical carriers is adjusted by changing the center
carrier of the laser at S1 while keeping that at S2 as 1550.12 nm unchanged. Both lasers are external cavity lasers (ECLs) working at C-band with the linewidth of 100 kHz. We first characterize the two-way communication at a bit rate of 2.5 Gb/s. The eye diagrams of the original signal \( D_{12} \) and the combined optical signal \( D_{\text{PLNC}} \) are shown in insets (a) and (b) in Fig. 2(a) and 2(b). The measurement is achieved with PRBS length of 7.

The lowest required received optical power to achieve a BER of \( 10^{-9} \) is \(-16 \) dBm. The relationship between the BER performance of the decoded signal and the frequency difference of the two optical carriers at the two transmitters is shown in Fig. 3(a). We can notice that \( D_{21} \) cannot be recovered when \( D_{12} \) and \( D_{21} \) are carried by exactly the same wavelength due to strong beating noise. BER = \( 10^{-9} \) can be achieved with a frequency difference of 6 GHz. The larger the central frequency difference is, the better the signal quality can achieve. On the other hand, the frequency difference cannot be too large. Otherwise, the full-duplex mutual communications cannot be carried out in the same DWDM grid. Considering the signal bandwidth of 2.5 GHz, the feasibility to put a two-way connection in one spectral grid of DWDM defined by ITU [19] is demonstrated, which means, this PLNC scheme can realize full-duplex communication, via a single 50-GHz DWDM grid at a line rate of 2.5 Gb/s.

Similar full-duplex communication at 10 Gb/s has also been investigated. The measured BER performance as a function of the frequency difference between the two optical carriers for a pair of servers is shown in Fig. 3(b). We can observe that a BER under \( 3.8 \times 10^{-3} \) can be obtained when there is about 18 GHz frequency difference. The error free full-duplex communication can be realized by applying HD-FEC technique [20]. In addition to the double-sided modulated 10Gb/s OOK signals, the spectrum of the combined signal is 38 GHz. Assuming 6 GHz guard band at the edge of the grid, the simultaneous mutual communications can be performed in the same 50 GHz DWDM grid.

![Fig. 3. BER performance of the decoded signal as a function of frequency difference at the transmitters for a pair of the servers, operating at (a) 2.5 Gb/s and (b) 10Gb/s.](image)

The experimental validation has demonstrated the feasibility of employing a simple PLNC scheme over the coupler-based POI architecture. As long as the frequency difference is satisfied by a pair of servers, the full-duplex communications can be realized in one DWDM grid. In other words, compared to the conventional coupler-based POI, applying the PLNC has a great potential to save spectrum resources. It should be noted that the implementation of the PLNC scheme requires a careful control of central carrier of the transceivers to make sure the frequency difference is large enough for achieving acceptable transmission performance. Moreover, strict datacenter environment control is required to maintain the temperature and moisture stable.
4. Network performance

In this section, we first introduce a dynamic bandwidth allocation algorithm designated for the PLNC-based POI and then evaluate the network performance in terms of packet delay and packet drop ratio.

4.1 Dynamic bandwidth allocation algorithm

To implement the PLNC in the POI efficiently, a tailored dynamic bandwidth allocation (DBA) algorithm is required. We propose a DBA algorithm compatible with the centralized cycle-based medium access control (MAC) protocol proposed in [9]. Here we take the coupler-based POI that employs ToR as an example to illustrate the proposed DBA algorithm. As mentioned in Section 2, in the control plane the ToR controller is required for the signaling at the control plane. At the beginning of each cycle all the servers send a request message to the ToR controller reporting the amount of the currently buffered packets targeting different destinations. After the ToR controller receives the information from all the active servers, it generates the traffic demand matrix after receiving the information from all the active servers and then runs a DBA algorithm, which is responsible for assigning wavelengths and time slots for the communication among the servers as well as the interface towards the outside of the rack. The servers are then informed with the assigned wavelengths and transmission window, based on which, the servers tune the transceivers centering to the assigned wavelengths and starts the transmission. Two examples of the generated traffic demand matrix are shown in Fig. 4(a) and Fig. 5(a). The values in each cell represents the packets in bytes that should be sent from the source server (SRC) to the destination server (DST). A wavelength can only be assigned to one element in a row or a column for conflict-free communications.

In this study, we take a greedy algorithm, namely Largest First (LF) which was proposed in [9] without the PLNC implementation as a benchmark. In this algorithm, the matrix elements are sorted in a descending order where the largest one is assigned with an available wavelength first. As shown in Fig. 4(b), LF finds the largest traffic demand, i.e., 1000 bytes from Server 1 to Server 2, and then allocates the first available wavelength \(\lambda_1\). There is another traffic demand of 1000 bytes, i.e., the one from Server 1 to Server 4. However, as the transmitter of Server 1 is being occupied in the coming cycle, this traffic demand is skipped and the next biggest one, i.e., 800 bytes from Server 2 to Server 1 is assigned with the second available wavelength \(\lambda_2\). The iteration stops when all the wavelengths are used or all the traffic demands are satisfied. The transmission window is determined by the largest demand that is being handled in the coming cycle.

![Traffic Demand Matrix](image)

**Fig. 4.** (a) Example 1 of traffic demand matrix, (b) the wavelength assignment based on the largest first algorithm for the case without using the PLNC [2], and (c) the proposed DBA for the case using the PLNC.
Using the PLNC technique, only one wavelength needs to be assigned to a pair of servers that demand simultaneous mutual communications. For example, as shown in Fig. 4(c), when assigning a wavelength to the largest traffic demand from port $p$ to $q$, the same wavelength can be reused to the traffic demand from port $q$ to port $p$. This reveals that fewer wavelengths are required in the case of using PLNC in order to reach the same throughput. In principle, a maximum wavelength reduction of 50% can be achieved by introducing the PLNC.

However, using PLNC does not always help to improve the network performance, as illustrated in Fig. 5. For the case without using the PLNC, two wavelengths are used to handle the first and the second largest traffic demands (i.e., 1000 bytes from Server 1 to Server 2, and 600 bytes from Server 2 to Server 3). When the PLNC is used, although only one wavelength can be assigned to meet the traffic demands of the bi-directional communications between Server 1 and Server 2 (i.e., 1000 bytes from Server 1 to Server 2, and 100 bytes from Server 2 to Server 1), the second largest traffic demand (i.e., 600 bytes from Server 2 to Server 3) is not able to be handled in the same cycle. With the same cycle length, the overall throughput is 1100 bytes for the case with the PLNC implemented, less than 1600 bytes that can be offered in the case without the PLNC. It can be seen that the benefits of implementing PLNC is related to the traffic load. In general, when the traffic demand matrix is sparse representing low traffic load, it has a higher probability that a bi-directional flow is not balanced, i.e., the traffic in one direction may be much less than that in the opposite direction. In this case, the PLNC cannot be fully utilized and even may make a negative impact on the overall throughput. With this in mind, we introduce an index $R$, which is defined as the ratio of the number of used wavelengths in the original LF algorithm over the total number of available wavelengths in a certain cycle (i.e., $R = \frac{N_{\text{used wavelengths}}}{N_{\text{total wavelengths}}}$, $0 \leq R \leq 1$). To some extent, $R$ can reflect whether the traffic load is high or low in the corresponding cycle. A threshold $R_{\text{max}}$ is determined. In a certain cycle, if $R < R_{\text{max}}$, the PLNC is switched off and the original LF algorithm is employed for wavelength and time slot allocation, while if $R \geq R_{\text{max}}$, the PLNC is switched on and a modified LF algorithm (where one wavelength is assigned to a bi-directional flow) is applied. The pseudocode of the proposed dynamic scheme tailored for the coupler-based POI with the PLNC implemented is shown in Fig. 6(a).
Tailored LF algorithm for implementing PLNC

1: Input: NC, Rmax, w; M: Traffic demand matrix;
2: %Rmax: the threshold to switch on PLNC;
3: w: available wavelengths count;
4: Tx: [None, None, ...]; TxTime = [0, 0];
5: Rx: [None, None, ...]; RxTime = [0, 0];
6: w_used = 0;
7: while Traffic = M.sort():
8: repeat
9:   D = Traffic[0];
10:   if D.Tx is None and D.Rx is None:
11:     Tx[D.src] = w[0];
12:     Rx[D.dst] = w[0];
13:     TxTime[D.src] = D.size/DataRate;
14:     RxTime[D.dst] = D.size/DataRate;
15:     w_used += 1;
16:     delete w[0];
17: until Traffic is empty or w is empty
18: if w_used <= Rmax * w:
19:   if D.Tx is None and D.Rx is None:
20:     Tx[D.src] = w[0];
21:     Rx[D.dst] = w[0];
22:     TxTime[D.src] = D.size/DataRate;
23:     RxTime[D.dst] = D.size/DataRate;
24:     delete w[0];
25: until Traffic is empty or w is empty
32: return Tx, Rx

(a)

Fig. 6. (a) Pseudocode of the modified LF DBA algorithm tailored for implementing PLNC in POI, (b) average packet delay as a function of traffic load, and (c) packet drop ratio as a function of traffic load with different number of wavelengths resources and the thresholds $R_{\text{max}}$ with a line rate of 2.5Gb/s.

4.2 Simulation setup and results

In this section, we evaluate the performance in terms of latency and packet drop for the PLNC over the POI scheme employing the proposed DBA algorithm. For network performance evaluation, a customized discrete event-driven simulator is developed by using Java programming language. In the simulation, the packet inter-arrival time follows a lognormal distribution [21]. The buffer size on each ONI is set to 5 MB. Each cycle consists of the transceiver tuning time and data transmission window. The transceiver tuning time is 200 ns [22] and the data transmission window is determined by the largest traffic request in the current cycle and the maximum transmission time is 5 us. Therefore, the upper bound of the carried load is 0.96. The propagation time is 50 ns (corresponding to 10 m fiber length, which is sufficient for the ToR case). We consider the locality of the data center traffic in the simulation, where 20% traffic of one server is from/to the outside of the rack (i.e., inter-rack traffic) and the remaining stays within the rack (i.e., intra-rack traffic) [23].

Figures 6(b) and 6(c) show the performances in terms of packet delay and drop ratio of the POI operating at a line rate of 2.5 Gb/s per port. Different numbers of available wavelengths and various values of thresholds $R_{\text{max}}$ are tested. $\gamma$ is denoted as the ratio of the number of the available wavelengths over the total number of the ports in the POI. According to the principle of the PLNC, the maximal required number of wavelengths can be reduced to half, i.e., $\gamma = 50\%$, while for the case without the PLNC implemented $\gamma = 100\%$ provides the sufficient number of available wavelengths and can reach the best network performance. The packet drop is observed at high load (normalized traffic load $> 0.9$), which is slightly less than the upper bound of the carried load. The packet delay consists of queuing time at the source node (servers and interfaces to the outside of the rack), transceiver tuning time, data transmission time, and propagation time. In Fig. 6(b), it can be observed that if the POI without the PLNC implemented with $\gamma = 50\%$ the packet latency degrades seriously compared to that in the case with $\gamma = 100\%$ when the traffic load exceeds 0.4. When the traffic load is
larger than 0.5, an obvious packet drop ratio can be observed, whereas the packet latency increases significantly and then becomes saturated due to the limited buffer size. Meanwhile, an obvious packet drop already appears when the traffic load is 0.5 with $\gamma = 50\%$ if the PLNC is not implemented, while with $\gamma = 100\%$ (i.e., with sufficient available wavelengths provided), the packet drop ratio is significantly improved. When applying the PLNC in the POI, even the case with $\gamma = 50\%$ the packet drop performance is almost the same as the POI without the PLNC, in which the sufficient number of wavelengths is provided (i.e., $\gamma = 100\%$). It should also be noted that the value of $R_{\text{max}}$ set in the proposed algorithm does not affect the packet drop ratio, as shown in Fig. 6(c). From Fig. 6(b) it can be observed that when the traffic load is low (less than 0.45) the PLNC does not bring obvious advantages. However, at a high traffic load, with $\gamma = 50\%$, the latency performance can be significantly improved by applying the PLNC, achieving a reduction by a factor of 100 compared to the case without implementing the PLNC. At a traffic load of 0.8, the average packet delay can still be kept below 1 ms. Moreover, increasing $R_{\text{max}}$ further improves the latency performance when the traffic load is low. As explained in the previous section, the PLNC may have a negative impact when the traffic load is low. By setting $R_{\text{max}} = 1$ in the proposed modified LF algorithm, the latency performance degradation can be minimized. Therefore, $R_{\text{max}} = 1$ is considered in the simulations hereafter.

When the operation data rate is upgraded to 10 Gb/s, it is observed that at a higher line rate, the benefits brought by implementing the PLNC on packet latency performance are more obvious (see Fig. 7(a)). Even with $\gamma = 50\%$ the delay performance for the scheme with the PLNC is very close to that in the scheme without the PLNC but having the number of available wavelengths doubled (i.e., $\gamma = 100\%$). Besides, as shown in Fig. 7(b), a similar performance trend is also observed in terms of the packet drop ratio. We also investigate the impact of the size of the interconnect (i.e., the number of the ports) on the network performance, where all the ports are operating at a line rate of 10 Gb/s. As shown in Fig. 8, we can observe that regardless of the size of the interconnect the network performance benefits introduced by using the PLNC are obvious especially in the cases with a high load.

---

Fig. 7. (a) Average packet latency and (b) packet drop ratio of POI architecture of different scale with and without PLNC at the operation rate of 10Gb/s.
Fig. 8. (a) Average packet latency and (b) packet drop performance of POI architecture of different scales with and without PLNC at the operation rate of 10Gb/s.

5. Summary

In this paper, we have conducted a study of implementing the PLNC in the context of the POI in the DCNs. Proof-of-concept experiments have been carried out to show the feasibility of applying the PLNC with a single DWDM grid to support simultaneous mutual communications between any pair of the ports in the POI. We have also proposed a tailored resource allocation algorithm for implementing the PLNC in the coupler-based POI. The network performance shows that the PLNC has a great potential to reduce the usage of wavelength resource by half while reaching the similar packet drop ratio compared to the case without the PLNC. With a small number of available wavelengths, a significant improvement in packet latency can also be achieved by applying the PLNC, especially at high traffic load.

Funding

Swedish Research Council (VR), the Swedish Foundation for Strategic Research (SSF); Göran Gustafsson Foundation; Natural Science Foundation of Guangdong Province (508206351021); National Natural Science Foundation of China (61550110240, 61671212); Hong Kong Research Grants Council (14200614).