Intra-datacenter links exploiting PCI Express Generation 4 interconnections

A. Gatto[(1)], P. Parolari[(1)], M. Brunero[(1)], F. Corapi[(1)], V. Costa[(2)], C. Meani[(2)] and P. Boffi[(1)]
[(1)] POLITECNICO DI MILANO, Dip. Elettronica, Informazione e Bioingegneria, PoliCom Lab, via Ponzio 34/5 – 20113 Milan (ITALY)
[(2)] ITALTEL S.p.A., via Reiss Romoli - 20019 Settimo Milanese (ITALY)
pierpaolo.boffi@polimi.it

Abstract: We demonstrate few-km reaches for PCIe-based optical fiber interconnections according to latency limitations, characterizing 16-Gb/s per lane Generation4 up to 10 km and confirming the Generation3 compliance of 2-km links employing suitable PCIe cards.
OCIS codes: (060.2330) Fiber optics communications; (200.4650) Optical interconnects

1. Introduction

Peripheral Component Interconnect Express (PCIe) is the de facto protocol, which supports high-performance communications between the host central processing unit (CPU) and the peripheral interfaces. PCIe implements the best features of the previous high-speed serial Input/Output (I/O) buses [1], establishing a serial I/O point-to-point interconnection through dual simplex differential pairs to communicate without I/O sharing, transport level and higher level functions. PCIe bandwidth has been scaled by means of multiple lanes (x4, x8, x16, and x32) and interconnections rates have increased from 2.5 Gb/s of the first generation (Gen1), to 5 Gb/s (Gen2) and 8 Gb/s (Gen3). However, communications are limited to few meters in case of copper cable links, thus PCI special interest group (PCI SIG) released the external cabling specifications [2] including fiber optical communications to extend the PCIe-based links, achieving remote data interconnection. PCIe Gen3 8x8 Gb/s transmission over 8 parallel OM4 multi-mode fibers (MMFs) (each one corresponding to a single lane) has been already experimented, demonstrating 150-m reach operation [3]: commercially available active optical MMF cables with 850-nm multi-mode VCSEL sources were employed. As PCIe next generation (Gen4) is expected to operate at 16 Gb/s [4], targeting 512-Gb/s capacity over 32 lanes (more than 1 Tb/s bidirectional) the exploitation of optical fiber communication is mandatory. Moreover, it has been recently demonstrated that adopting proper network interface cards two or more servers can be clustered together through PCIe interconnections realizing a very efficient PCIe fabric [5]. In this case thus the exploitation of fast optical PCIe links would provide an effective coverage of the entire datacenter area with an efficient low-power technology, achieving a low-latency connectivity.

These high-speed intra-datacenter communications will highly benefit from the employment of single-mode fibers (SMF) and sources in the second and third fiber windows, allowing to achieve hundreds/ thousands of meters of interconnection reach. On the other hand, longer distances add delays in propagation and impact on the connection latency: in case of PCIe-based optical links, the maximum reach appears not limited by the fiber propagation impairments (such as attenuation and dispersion), but by the delay affecting the ACK/NACK operation [6].

In this paper we provide a detailed analysis on PCIe maximum optical link, limited by the ACK/NACK operation latency for PCIe Gen3 and Gen4. Although Gen4 protocol is still under standardization, exploiting Gen4 preview hypotesis [4], we demonstrate Gen4 interconnections at 16-Gb/s per lane over 2 and 10 km of SMF with commercially available 1310-nm SFP+ transceivers, without limitations due to the fiber propagation impairments. Finally, we demonstrate the real-time operation of a PCIe Gen3 end-to-end fiber link between two PCIe cards developed within the ADVENT project [7] for interconnection reaches of few kms, typical of the intra-datacenter communications, showing the compliance with the PCIe standard.

2. Latency limitations of PCIe optical fiber remotization

In the PCIe protocol the error detection is based on the link cyclic redundancy check (LCRC), which is used to ensure the delivery and the integrity of the transaction layer packet (TLP) during the transmission across the link. Several counters and timers are employed to set up the TLP LCRC to identify the sequence number and to support the data link layer retry. In particular, the reply timer (RT) is connected to the ACK/NACK mechanism and limits the PCIe-based link design in terms of reach owing to the propagation latency. RT is conditioned by the maximum payload size (MPS), by the TLP overhead (depending by the PCIe generation), by the ACK factor (i.e. the number of maximum size TLPs which can be received before an ACK is sent back), by the internal system delay (due to processing delay for the transmitted and received TLPs, depending again by the PCIe generation taken into account), and finally by the operating width of the link in terms of number of lanes.

978-1-943580-23-1/17/$31.00 ©2017 Optical Society of America

Starting from the above considerations regarding the limitation of the ACK/NACK mechanism in terms of delay time and by the RT values standardized by the PCI SIG [8], we calculated the maximum length of PCIe-based links for the different generations, MPSs and link width, taking into account the exploitation of the optical fiber as propagation medium. In Tab. 1 left the maximum distance reachable in case of PCIe Gen3 fiber link is reported. Moreover, since PCIe Gen4 RT values of the ACK/NACK mechanism are not yet available, we performed some hypothesis based on [4, 8] and obtained plausible RT values. In particular, analyzing the coding scheme of the older generations and following the Gen4 recommendations reported in [4], it is possible to make a prediction on the maximum reachable distance for the future Gen4. The obtained results are presented in Tab.1 right. As can be seen, for Gen4, about 2-km interconnection can be reached in cases of 8 and 16 lanes for a MPS of 4096 bytes (the reach remains the same as when the number of lanes is halved the ACK factor is doubled). When just a single lane (e.g. a single fiber link) is taken into account, about 10-km reach is achievable.

PCIe Gen3		Link Operating Width						
		x1	x2	x4	x8	x12	x16	x32
Max Payload Size (Byte)	128	1.62	1.09	0.82	0.80	0.75	0.70	0.63
	256	2.50	1.53	1.04	0.99	0.97	0.82	0.69
	512	3.19	1.88	1.22	0.89	1.00	0.89	0.72
	1024	5.69	3.13	1.84	1.20	1.41	1.20	0.88
	2048	10.68	5.62	3.09	1.82	2.25	1.82	1.19
	4096	20.66	10.61	5.59	3.07	3.91	3.07	1.81

PCIe Gen4		Link Operating Width						
		x1	x2	x4	x8	x12	x16	x32
Max Payload Size (Byte)	128	1.25	0.87	0.72	0.67	0.67	0.63	0.57
	256	1.62	1.09	0.82	0.80	0.75	0.70	0.63
	512	1.82	1.22	0.89	0.74	0.80	0.74	0.69
	1024	3.19	1.88	1.22	0.89	1.00	0.89	0.72
	2048	5.69	1.84	1.20	1.41	1.41	1.20	0.88
	4096	10.68	5.62	3.09	1.82	2.25	1.82	1.19

Tab 1. Calculated maximum distance (in km) reachable for PCIe fiber links for Gen3 (left) and Gen4 (right).

3. PICe Gen4 fiber interconnections experimental validation

The calculated fiber reaches shown in the previous paragraph are related to the protocol limitation due to ACK/NACK operation in terms of latency. To evaluate the impact of the physical fiber propagation impairments and to confirm the above reach values, we experimented optical PCIe Gen4 transmission performance employing SMF to avoid modal dispersion limitations and to save space consumption. Commercial transceiver modules designed for Fiber Channel based applications including 1310-nm DFB lasers (with maximum data rate of 14 Gb/s) and a PIN photodiode in a SFP+ interface were exploited to implement PCIe Gen4 connection. The operation in the second window of optical communications allows negligible chromatic dispersion impairments also at the considered bit rate. The BER performance were measured for interconnection reaches of 2 km and 10 km according to the limits evidenced in Tab. 1 right due to latency limitations. Fig. 1 a) reports the BER curves measured with 2^{23}-1 pseudo random bit sequence (PRBS) at 16 Gb/s; as can be seen no error floors and no significant penalties are introduced by the propagation in the fiber link with respect to the back-to-back (b2b) operation. Moreover, Fig. 1 b) presents the 16-Gb/s detected eye diagram for 2-km SMF: there is a full compliance with the protocol eye-mask deduced from [4] for PCIe Gen4. In Fig. 1 c) the BER bathtub curve is shown as a function of the sample time, reporting the horizontal eye opening of 25 ps at 10^{-12} BER. The exploitation of a transceiver with limited 14-Gb/s bandwidth affects the jitter, measured around 37 ps, but a simple 2-taps decision feedback equalization (DFE), expected in the future PCIe Gen4 standardization [4], would easily compensate this bandwidth limitation.

4. PCIe-based fabric implementation and remote connection characterization

Within the research project ADVENT we have developed small PCIe modules, which are the key elements to design a PCIe-based fabric, where PCIe protocol is used to connect together a cluster of servers exploiting the optical fiber to guarantee the remote connectivity necessary in case of intra-datacenter applications. In this hypothesis, no translation to other protocols is required and the network efficiency is maximized through the physical disaggregation of the cluster resources and the improvement of the scale-up capabilities.

We characterized a real-time PCIe end-to-end fiber link of 2-km SMF by means of two PCIe card modules exploiting PCIe Gen3 data transmission (Fig. 2 a). In our experimentation we employed a host PC housing the two PCIe cards (as show in Fig. 2 b) picture), consuming a few watts and equipped with SFP+ interfaces including commercial 1310-nm transceivers used for 10 Gigabit Ethernet applications. Thanks to Tektronix real-time oscilloscope DPO73304DX, we verified the 2-km SMF transmission compliance with PCIe Gen3 standard both in terms of eye mask (in Fig. 2 d)) and jitter (in Fig. 2 e)). This kind of experimentation was achieved by sending the training sequence between the two PCIe cards in order to achieve the end-to-end connection. BER performance was measured (in Fig. 2 c)) by tapping the received power from a 50/50 fiber coupler placed before one of the two PCIe cards. All the measurements confirm the compliance to the PCIe Gen3 standard, demonstrating the fiber link

capabilities to support 2-km remote connectivity, which is the maximum reach allowed by the latency for 4096-bytes MPS and 32 lanes or for 2048-bytes MPS and 8 or 16 lanes, as demonstrated in Section 2.

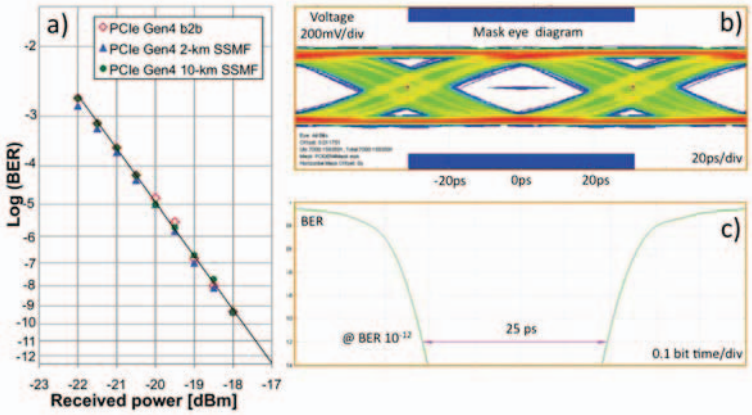

Fig. 1. 16-Gb/s PCIe Gen4 characterization: BER vs. received power up to 10-km (a); eye diagram mask (b) and bathtube curve (c) for 2-km reach, showing the horizontal eye opening for 10^{-12} BER (pink line).

5. Conclusions

By exploiting the optical fiber to achieve remote interconnections, new networks based on PCIe standard can be implemented for intra-datacenter applications, guaranteeing high scalability and the functional characteristics of today server clusters. In this paper we have evaluated in details the maximum fiber reach limited by the latency due to the PCIe protocol ACK/NACK operation; we have demonstrated the feasibility of few-km interconnections by exploiting SMF propagation and transmission in the second window of optical communications. In particular, PCIe Gen4 has been experimented up to 10-km SMF by employing commercial transceivers devoted to Fiber Channel applications, demonstrating that potentially 256 Gb/s capacity (512 Gb/s bidirectional) can be targeted over 2 km in case of 16 lanes, according to PCI SIG indications. Moreover, the compliance to the PCIe standard has been shown in case of a 2-km end-to-end fiber link obtained with two PCIe cards operating in real time and exploiting Gen3 data transmission, showing the fiber-link capabilities to support PCIe-based inter-host remote communications with intrinsic advantages in terms of connectivity and applications.

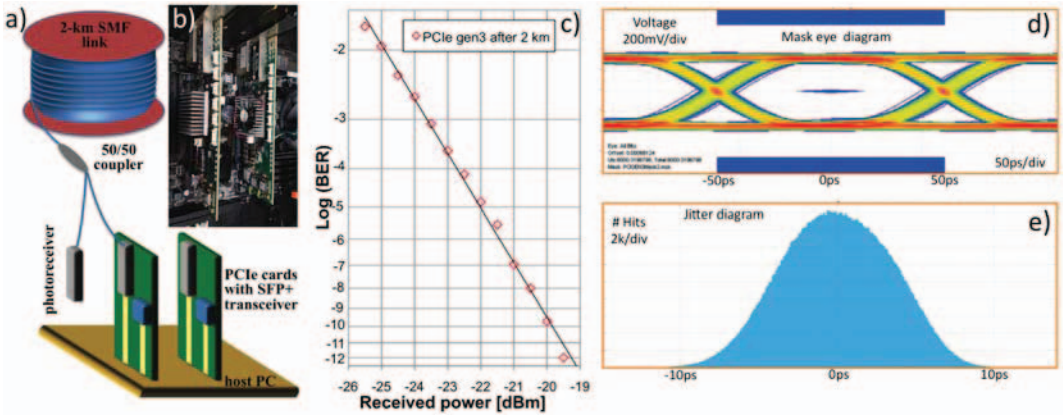

Fig. 2. (a) Experimental layout for the 2-km end-to-end link demonstration. (b) the two employed PCIe cards. PCIe Gen3 link characterization: BER measures (c); eye diagram mask (d) and time jitter histogram (e).

6. Acknowledgements

The authors thank TEKTRONIX for the support to experimentation.

7. References

[1] R. Budruk, D. Anderson, T. Shanley, *PCI express System Architecture* (Addison-Wesley, 2003).
[2] PCI Express External Cabling Specification Revision 1.0, PCI SIG (2007).
[3] www.avagotech.com/docs/AVO2-3245EN, white paper.
[4] D. Gonzales, "PCI Express 4.0 Electrical previews," PCI-SIG Developers Conference, Santa Clara – CA (2015).
[5] E. Billi, "Realizing the next step in storage/converged architectures" Flash Memory Summit 2015, Santa Clara, CA (2015).
[6] D. Percival, "PCIe over Fibre Optics: challenges and pitfalls," PCI-SIG Developers Conference, Tel Aviv – Israel (2015).
[7] P. Boffi et al., "PCIe-based network architectures over optical fiber links: an insight from the ADVENT project," Proc. FOTONICA 2016 AEIT Italian Conference on Photonic Technologies (2016).
[8] PCI Express Base Specification Revision 3.0, PCI SIG (2010)