SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

June 2019, Vol. 62 062401:1–062401:7
https://doi.org/10.1007/s11432-018-9392-x

Error correction for short-range optical interconnect using COTS transceivers

Ziyuan ZHENG, Chuanchuan YANG*, Dan ZHAO & Ziyu WANG

State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871*, China*

Received 15 January 2018/Revised 27 February 2018/Accepted 14 March 2018/Published online 23 October 2018

Abstract Nowadays, quantities of commercial off-the-shelf (COTS) optical transceivers are widely equipped in both exiting and abuilding data centers. As data centers are becoming larger and containing more servers, the method to ensure longer and steady transmission with COTS optical transceivers becomes one of the key issues. Meanwhile, variations in the manufacturing of COTS optical transceivers bring in fluctuation of packet error rate (PER) in communication. In this paper, we present an application layer adaptive forward error correction (AL-AFEC) coding scheme for COTS optical transceivers in optical interconnect to solve the problems mentioned above and test its performance in a 10 Gbps Ethernet (10GbE) link using 10GBASE-SR SFP+ modules. Experimental results show that the proposed scheme enables longer transmission distance and the instability introduced by manufactural variations is well settled. Since our scheme requires no extra hardware on optical transceivers, it is a considerable low-cost alternative to improve system reliability and achieve longer transmission distance for COTS optical transceivers.

Keywords short-range optical interconnect, vertical cavity surface emitting laser (VCSEL), ethernet, application layer adaptive forward error correction, commercial off-the-shelf optical transceiver

Citation Zheng Z Y, Yang C C, Zhao D, et al. Error correction for short-range optical interconnect using COTS transceivers. Sci China Inf Sci, 2019, 62(6): 062401, https://doi.org/10.1007/s11432-018-9392-x

1 Introduction

With the exponential increase in the employment of cloud computing and several emerging web applications, short-range optical interconnect is facing great challenges of being longer and stable for larger data centers [1]. Nowadays, the short-range optical links in data centers mainly rely on low-cost vertical cavity surface emitting lasers (VCSELs), multimode fiber (MMF), and direct detection by photodiodes (PDs). A typical example of short-range optical interconnect, a 70 Gbps NRZ link based on 850 nm VCSEL and FFE+MLSE equalization approach, is shown in [2]. The connections of the servers in data centers commonly use commercial off-the-shelf (COTS) small form-factor pluggable (SFP) modules (10/28 Gbps SFP+ and 40/100 Gbps QSFP+) [3].

The SFP+/QSFP+ module is a kind of COTS optical transceiver targeted for data-com applications specified by the SFF Committee (SFF8402, SFF8431, SFF8432, and SFF8472). Tremendous progress in the use of SFP+/QSFP+ modules has been achieved in the past few years. The creation of as many as 48 ports of 10GBASE-SR links on a single 19-inch rack is achieved in [4]. Nasu et al. [5] proposed a solderable 28 Gb/s × 4-channel VCSEL-based optical transceiver module designed for QSFP28 MPO-type transceiver over 5 m OM2 MMF. Intel Corporation presented a low power, 2 × 25.625 Gb/s optical

* Corresponding author (email: yangchuanchuan@pku.edu.cn)

transmitter and receiver integrated circuit (IC) which can be used in 100 G QSFP28 optical module [6]. However, transmission distance of the above-mentioned SFP+ module is greatly limited as packet error rate (PER) increases significantly with the increase of MMF length transmitted, owing to VCSELs' large spectrum width and multiple transmission modes [7], e.g., normal transmission distance of short-range MMF links with 850 nm COTS SFP+ modules is 300 m OM3 MMF or 400 m OM4 MMF. Meanwhile, variations in manufacturing will introduce individual difference at center wavelength, spectral width and optical modulation amplitude (OMA) of SFP+ modules [4]. With the increase of the scale of data centers, transmission distance and channel loss of short-range optical link becomes larger [8]. As a result, how to realize the target of high reliability and long distance at the same time regardless of variations in manufacturing becomes one of the key issues in the data center.

Forward error correction (FEC) is a good candidate to be introduced to promote transmission performance in longer distance. Although traditional physical-layer FEC (PL-FEC) such as shortened cyclic codes and Reed Solomon codes has been introduced according to IEEE 802.3 [9], a large quantity of COTS SFP+ modules without PL-FEC chips have been widely used in exiting and abuilding data centers. Furthermore, traditional PL-FEC does not suit well dealing with the fluctuation of PER in communication introduced by individual difference among commercial optical transceivers due to its characteristics of fixed-rate.

In this paper, we propose an application layer adaptive FEC (AL-AFEC) coding scheme into short-range optical interconnect to enable longer transmission distance as well as cope with the individual difference. The scheme employs a class of application-layer FEC named raptor codes, which are originated from the family of fountain codes. Raptor codes have the ability that the decoder can recover the source blocks from any set of encoding packets acquired from the limitless stream generated by the encoder [10], which are considered as rateless codes. The encoder generates different number of redundant packets at different PER, making it easy to accommodate different transmission distance, channel loss and SFP+/QSFP+ modules with individual difference. We tested the performance of the AL-AFEC scheme in a 10 Gbps Ethernet (10GbE) link over OM4 MMF using COTS 850 nm 10GBASE-SR SFP+ modules. Longer transmission distance was achieved with the aid of the scheme. At the same time, PER fluctuation in communication introduced by variations in manufacturing is well settled due to raptor codes' rateless characteristic. It is worth noting that the AL-AFEC scheme is easy to be introduced into the exiting and abuilding data centers, which are equipped with COTS SFP modules, as no extra hardware is required on optical transceivers.

2 Mechanism of AL-AFEC coding scheme

In this section, the proposed coding scheme, which is called AL-AFEC, is presented. Raptor codes can be categorized as systematic and non-systematic. Our proposal is systematic raptor codes, i.e., the source packets are among the encoded packets for the reason that when PER of the transmission system is pretty low, receivers are able to obtain all data needed by only using the packets containing original source data without exploit the redundant packets since there is no error in the original source packets in this case.

Figure 1 presents the configuration of the encoder of systematic raptor codes. The encoding progress of systematic raptor code includes a high-rate low-density parity-check (LDPC) pre-coding and a later Luby transform (LT) coding.

At the transmitter, transmitted data is divided into several blocks with each block consisting of k source packets whose length is denoted as L bytes. The raptor encoder encodes the k source packets into n packets for each block. The number of redundant packets in a block is r, which is equal to $n-k$. Each square in Figure 1 represents a packet of L bytes. Firstly, the k source packets are pre-coded into s intermediate packets using a hybrid LDPC-Half systematic linear correction code introduced in [11], which is not traditional error correction code. Then a weakened LT code encodes the s intermediate packets into r redundant packets. Soliton distribution is employed in the weakened LT code scheme. At last, the original k source packets and the generated r redundant packets constitute the total n encoded

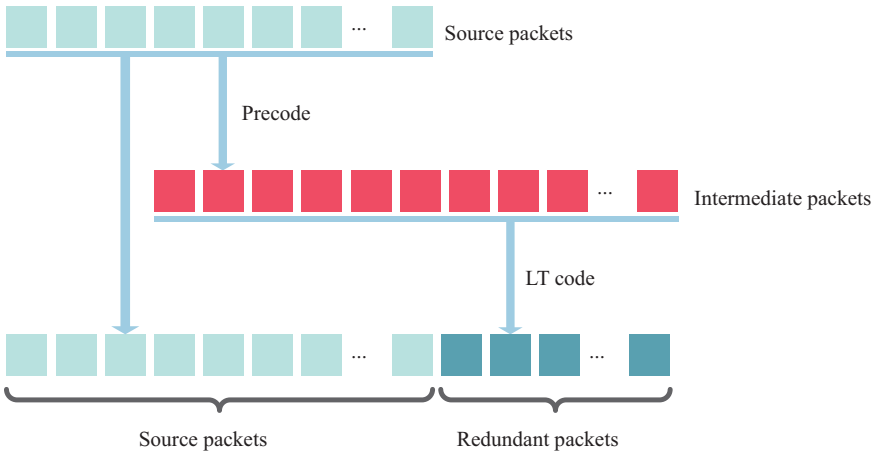

Figure 1 (Color online) Configuration of the encoder of systematic raptor codes.

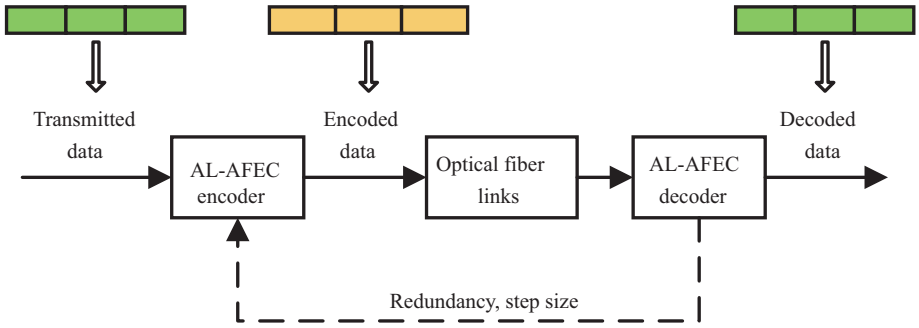

Figure 2 (Color online) Schematic of short-range optical interconnect using AL-AFEC coding.

packets.

In the systematic raptor decoder, following the known relationships amongst the intermediate packets and the source packets, the reverse of the generation matrix can be calculated by Gaussian elimination [11], so the original k source packets can be decoded. The pre-code can provide protection to the source packets by correcting erasures not recovered by the weakened LT code. The weakened LT code can guarantee the complexity of $O(\log k)$, which is better than the normal LT code complexity of $O(k \log k)$. So this concatenated code can provide both high protection and low complexity at the same time. The reception overhead is denoted as ε, which satisfies $r = k\varepsilon$. This suggests that the decoding success probability can be larger when larger k and ε is used.

Based on the description above, parameters of raptor codes must be carefully designed for reliable transmission. Unlike other fixed-rate coding scheme, AL-AFEC coding scheme adaptively determine the redundancy of raptor codes according to the link conditions.

Figure 2 shows the schematic of short-range optical interconnect using AL-AFEC coding. At the transmitter, for a specific block, the AL-AFEC encoder encodes it into a new block with redundant packets. In order to adapt to the changes in link conditions, the redundancy should be adjusted accordingly. Short-range optical link in data centers is a relatively stable link which means link conditions hardly change sharply. As a result, we do not adjust redundancy every block, which avoids high delay in the optical link. Instead, the redundancy is adjusted every step, and it is denoted as r_i $(i = 1, 2, \ldots)$. After encoding, encoded packets are transmitted through optical fiber links. At the receiver, the AL-AFEC decoder recovers the lost packets using the redundant packets. At last, the redundancy and step size is updated and sent to the transmitter every step.

Figure 3 illustrates the schematic of the redundancy adjustment method. The step size, i.e., the number of blocks received at each step is denoted as N_i $(i = 1, 2, \ldots)$.

At the i-th step, we denote number of lost packets in each block as m_j $(j = 1, 2, \ldots, N_i)$. Therefore,

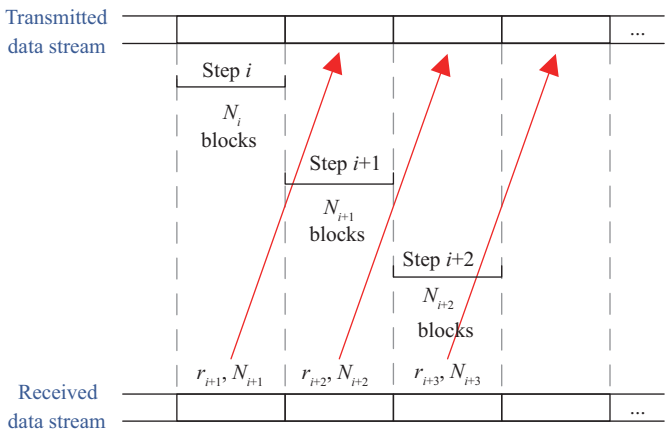

Figure 3 (Color online) Schematic of the redundancy adjustment method.

the PER at this step can be calculated by

$$p_i = \sum_{j=1}^{N_i} \frac{m_j}{(k + r_i) \cdot N_i}. \tag{1}$$

The redundancy r_i at each step depends on the number of source packets in each block k as well as the PER denoted as p_i $(i = 1, 2, \ldots)$. PER is a function of packets length and link conditions which change according to link length, channel loss, fluctuation of transmitters' optical launch power and COTS optical modules' manufactural variations.

In order to achieve a reliable error-free transmission, the redundancy at step $i + 1$ should satisfy

$$r_{i+1} \geqslant \frac{k \cdot p_i}{1 - p_i}. \tag{2}$$

Thus, the redundancy at the $(i + 1)$-th step is the function of k and p_i at the i-th step, which can be denoted as $r_{i+1}(k, p_i)$.

In the rare case that link conditions change sharply, the redundancy updated from previous step $r_i(k, p_{i-1})$ may fail to recover all the lost packets. The receiver will send a feedback message to the transmitter to increase r_i immediately until the receiver recover all blocks successfully, after which another feedback message will be sent to the transmitter and then the transmitter will stop increasing r_i and send the remaining data with the newly decided r_i.

Step size N_i should also be adjusted to suit the pace of change in link conditions. We employ two parameters β and γ to update the step size, where $\beta > 1$ and $0 < \gamma < 1$, as follows:

$$N_{i+1} = \begin{cases} \lceil N_i \cdot \gamma \rceil, & \text{for decoding failure,} \\ \lceil N_i \cdot \beta \rceil, & \text{for decoding success,} \end{cases} \tag{3}$$

since small step size fits quick change in link conditions better. Values of β and γ depend on the characteristic of link and we suggest a pair of empirical values of $1 < \beta < 2$ and $0.5 < \gamma < 1$ considering short-range optical link is relatively stable.

The remaining work need to be conducted is the select of packet length L and block size k. In the optical interconnect link among servers in data centers, it is an efficient setting that the packet size should be less than the size of an Ethernet frame [12, 13]. As a consequence, the length of source packet L is chosen from 64 to 1500 bytes according to Ethernet protocol. After L is settled, as the decoding complexity of the proposed code is only a function of k which can be denoted as $O(\log k)$ for each block, it can be seen that larger k means higher complexity. Meanwhile, larger k will incur larger latency in communication due to high complexity. However, a larger k will decrease the coding overhead ε needed at the same PER [14]. Hence, a trade-off between complexity, latency and overhead should be made according to communication system's requirements.

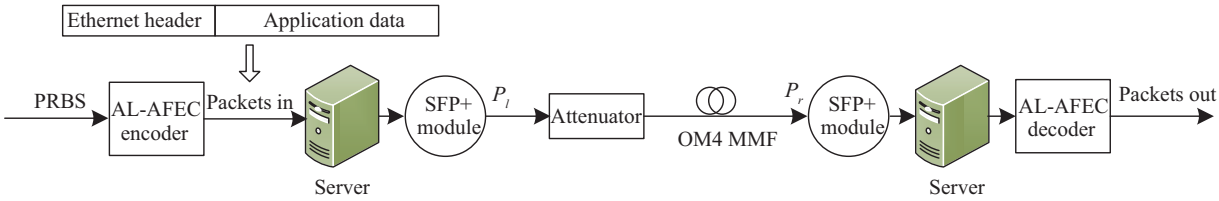

Figure 4 (Color online) Experimental system of the 10GBASE-SR transmission via OM4 MMF.

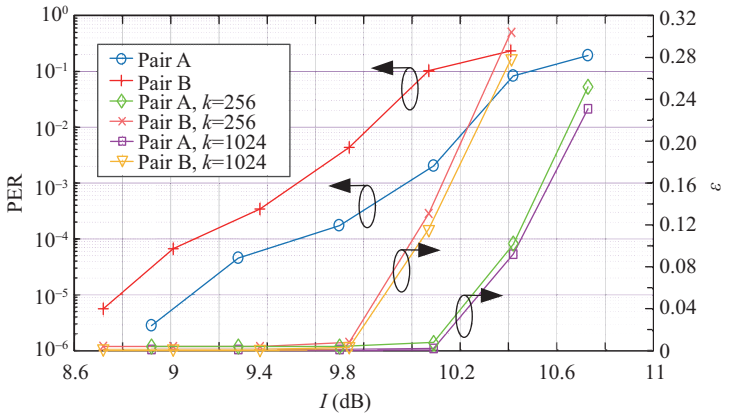

Figure 5 (Color online) PER and reception overhead needed versus channel loss via 600 m OM4 MMF using different SFP+ modules.

3 Experimental results

As shown in Figure 4, the experimental system consists of two COTS 10GBASE-SR SFP+ modules which are equipped in two servers, OM4 MMF ranging from 200 to 700 m and an optical attenuator. Pseudo-random binary sequence (PRBS) is divided into several blocks and then transmitted using the proposed AL-AFEC coding scheme. The optical launch power is denoted as P_l and the optical receive power is denoted as P_r. Thus, the channel loss, which is denoted as I, can be calculated by

$$I = -10\log_{10}\frac{P_r}{P_l}. \tag{4}$$

Two pairs of SFP+ modules of the same type, which are named pairs A and B, were used to study the individual difference among SFP+ modules that results from variations in manufacturing as well as research the ratelss characteristic of the proposed coding scheme. P_l of pair A is -0.6 dBm while that of pair B is -1.7 dBm.

In order to investigate the adaptive characteristic of the AL-AFEC scheme, we tested PER of the two pairs of SFP+ modules and obtained the reception overhead ε needed to guarantee an error-free transmission when the channel loss I is gradually adjusted. In this case, L is set as 1500 to evaluate the performance at worst case since larger L means more error probability within a packet. Different values of k are set to analyze k's influence on ε ($k = 256$ or 1024). Transmission distance is set as 600 m. γ, β are set as 0.9, 1.2 and the initial value of step size is set as 50.

As shown in Figure 5, the blue line and the red line refer to the PER of the two pairs of 10GBASE-SR SFP+ modules without AL-AFEC and the remaining four lines show the reception overhead ε for error-free communication at different I. We found that PER of both pairs of SFP+ modules increases significantly as I increases from 8.7 to 10.7 dB. Pair A's PER increases from 2.794×10^{-6} to 0.1927 in the range of 8.9 to 10.7 dB while pair B's PER increases from 5.588×10^{-6} to 0.2339 in the range of 8.7 to 10.4 dB. Same type of SFP+ modules perform differently at the same channel loss as they differ in transfer OMA, spectral width, center wavelength and sensitivity of receiver, which will lead to fluctuation of PER in communication. Then we obtained the reception overhead ε calculated by the proposed decoder.

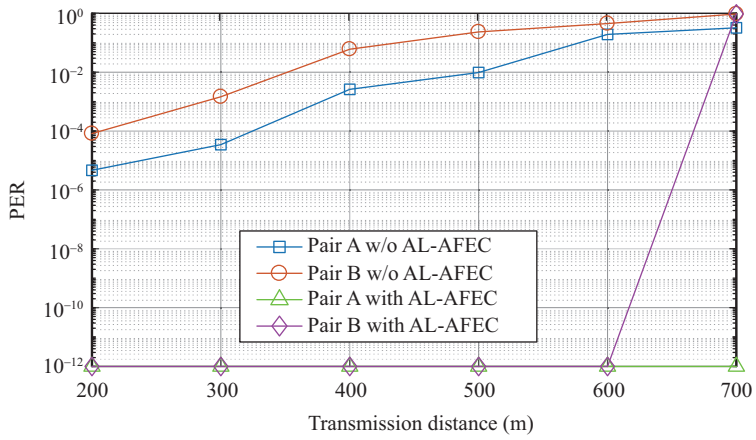

Figure 6 (Color online) PER of the two pairs of SFP+ modules at different transmission distance with/without AL-AFEC.

When $k = 1024$, the reception overhead for pair A to achieve error-free communication increases from 9.771×10^{-4} to 0.231 in the range of 8.9 to 10.7 dB while the reception overhead for pair B increases from 9.771×10^{-4} to 0.278 in the range of 8.7 to 10.4 dB. When $k = 256$, there is a slight increase in the redundancy compared to $k = 1024$ at the same I. However, a great decrease in decoding complexity is acquired at the same time. The AL-AFEC coding scheme can adaptively determine redundancy according to the change in link conditions and variations in manufacturing to guarantee an error-free transmission for the two pairs of SFP+ modules at different channel loss. And it is suggested that smaller k is a better choice since it is more efficient.

Figure 6 shows the PER of the two pairs of SFP+ modules at different transmission distance with/without AL-AFEC. In this case, k is set as 256 and L is set as 1500. The maximum channel loss budget for the 10GBASE-SR SPF+ module to acquire error-free link over 400 m OM4 MMF is 9.0 dB, which is illustrated in [15]. So the attenuation of the optical attenuator is set as 9.0 dB which means that $I = 9.0$ dB in B2B case. The blue line refers to the PER of pair A without AL-AFEC and the red line refers to the PER of pair B, respectively. Different performance of the two pairs at the same link conditions is shown and we can see that the PER increases significantly when transmission distance gets longer: pair A's PER increases from 4.642×10^{-6} to 0.3231 in the range of 200 to 700 m while pair B's PER increases from 8.203×10^{-5} to 0.9533 in the same range. Notably, the PER of pair B almost equals to 1 at 700 m which makes it impossible to recover the lost packets by coding technique. The purple and green lines show the PER of the two pairs when AL-AFEC is employed, and we found that error-free transmission up to 600 m over OM4 MMF was achieved with a considerable reception overhead, while pair B's error at 700 m cannot be corrected as nearly all packets going wrong which makes the decoder difficult to get enough useful packets. Compared to the normal case without AL-AFEC, our scheme enables at least 600 m longer error-free transmission distance, which is an obvious promotion on the performance of short-range optical link.

4 Conclusion

We propose an AL-AFEC coding scheme for the optical interconnect system employing COTS optical transceivers. Fluctuation of PER introduced by variations in manufacturing and link conditions is solved using the rateless characteristic of our proposal. Experimental results show that the proposed AL-AFEC coding scheme enables a stable error-free and at least 600 m longer transmission distance, with a considerable redundancy, compared with the case without the aid of our proposal. As no extra hardware is required on optical transceivers, AL-AFEC can be viewed as a flexible and reliable alternative for both exiting and abuilding data centers equipped with COTS optical transceivers, which call for longer transmission distance.

References

1 Xia W F, Zhao P, Wen Y G, et al. A survey on data center networking (DCN): infrastructure and operations. IEEE Commun Surv Tut, 2017, 19: 640–656
2 Tan Z W, Yang C C, Zhu Y X, et al. A 70 Gbps NRZ optical link based on 850 nm band-limited VCSEL for data-center intra-connects. Sci China Inf Sci, 2018, 61: 080406
3 Lam C F, Liu H, Koley B, et al. Fiber optic communication technologies: what's needed for datacenter network operations. IEEE Commun Mag, 2010, 48: 32–39
4 Bhoja S, Ghiasi A, Chang Y F, et al. Next-generation 10 GBaud module based on emerging SFP+ with host-based EDC. IEEE Commun Mag, 2007, 45: 32–38
5 Nasu H, Nagashima K, Ishikawa Y, et al. 28-Gb/s×4-channel solderable optical transceiver module for QSFP28. In: Proceedings of IEEE CPMT Symposium Japan, Kyoto, 2016. 31–34
6 Gao J, Wu H C, Liu G B, el al. 2×25.625G low power optical IC for thunderbolt optical cable technology. In: Proceedings of IEEE Optical Interconnects Conference, San Diego, 2016
7 Lavrencik J, Pavan S K, Thomas V A, et al. Noise in VCSEL-based links: direct measurement of VCSEL transverse mode correlations and implications for MPN and RIN. J Lightwave Technol, 2017, 35: 698–705
8 Kachris C, Tomkos I. A survey on optical interconnects for data centers. IEEE Commun Surv Tut, 2012, 14: 1021–1036
9 IEEE Standards Association. IEEE standard for ethernet. IEEE Std 802.3-2015. http://standards.ieee.org/findstds/standard/802.3-2015.html
10 Shokrollahi A. Raptor codes. IEEE Trans Inf Theory, 2006, 52: 2551–2567
11 Luby M, Shokrollahi A, Watson M, et al. Raptor forward error correction scheme: scheme for object delivery. IETF RFC 6330. https://www.rfc-editor.org/info/rfc6330
12 Kwon O C, Go Y, Park Y, et al. MPMTP: multipath multimedia transport protocol using systematic raptor codes over wireless networks. IEEE Trans Mobile Comput, 2015, 14: 1903–1916
13 Eittenberger P M, Mladenov T, Krieger U R. Raptor codes for P2P streaming. In: Proceedings of Euromicro International Conference on Parallel, Distributed and Network-based Processing, Garching, 2012. 327–332
14 Zhang W Z, Hranilovic S, Shi C. Soft-switching hybrid FSO/RF links using short-length raptor codes: design and implementation. IEEE J Sel Areas Commun, 2009, 27: 1698–1708
15 SFF Committee. SFF-8431 Specifications for Enhanced Small form Factor Pluggable Module SFP+. Revision 4.1, 2009