Low latency and efficient optical flow control for intra data center networks

Low latency and efficient optical flow control for intra data center networks

Wang Miao,* Stefano Di Lucente, Jun Luo, Harm Dorren, and Nicola Calabretta

COBRA Research Institute, Eindhoven University of Technology, PO Box 512, 5600MB Eindhoven, Netherlands
w.miao@tue.nl

Abstract: We experimentally demonstrate a highly spectral efficient optical flow control technique for intra data center networks. A bi-directional system is implemented for generating flow control signal by reusing label wavelength and the transmission link within the same WDM channel. Dynamic operation shows high-quality flow control signal with 265ns latency including 220ns propagation delay and 500mV amplitude with low input power and low bias current. Error free operation with 0.5dB penalty for 40Gb/s payload indicates that no distortion has been caused due the transmission of label and flow control signal.

OCIS codes: (060.6719) Switching, packet; (200.4650) Optical interconnects.

References and links

1. T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proceedings of 10th annual conference on Internet measurement*. (ACM, New York, 2010), pp. 267–280.
2. S. Sakr, A. Liu, D. Batista, and M. Alomari, "A survey on large scale data management approaches in cloud environments," IEEE Commun. Surveys Tutorials **13**(3), 311–336 (2011).
3. *Cisco Data Center Interconnect Design and Deployment Guide*
4. L. A. Barroso and U. Hölze, *The datacenter as a computer: an introduction to the design of warehouse-scale machines* (Morgan and Claypool Publishers, Los Angeles, 2009).
5. K. Chen, C. Hu, X. Zhang, K. Zheng, Y. Chen, and A. V. Vasilakos, "Survey on routing in data centers: insights and future directions," IEEE Netw. **25**(4), 6–10 (2011).
6. X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: A scalable optical switch for datacenters," in *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*. (ACM, New York, 2010), pp. 1–24.
7. A. K. Kodi and A. Louri, "Energy-Efficient and Bandwidth-Reconfigurable Photonic Networks for High-Performance Computing (HPC) Systems," IEEE J. Sel. Top. Quantum Electron. **17**(2), 384–395 (2011).
8. J. Gripp, J. E. Simsarian, J. D. LeGrange, P. Bernasconi, and D. T. Neilson, "Photonic Terabit Routers: The IRIS Project," in *Optical Fiber Communication Conference*, Technical Digest (CD) (Optical Society of America, 2010), paper OThP3.
9. S. Di Lucente, J. Luo, R. P. Centelles, A. Rohit, S. Zou, K. A. Williams, H. J. S. Dorren, and N. Calabretta, "Numerical and experimental study of a high port-density WDM optical packet switch architecture for data centers," Opt. Express **21**(1), 263–269 (2013).
10. J. Luo, S. Di Lucente, J. Ramirez, H. J. S. Dorren, and N. Calabretta, "Low latency and large port count optical packet switch with highly distributed control," in *Optical Fiber Communication Conference*, Technical Digest (CD) (Optical Society of America, 2012), paper OW3J.2.
11. J. Luo, H. J. S. Dorren, and N. Calabretta, "Optical RF tone in-band labeling for large-scale and low-latency optical packet switches," J. Lightwave Technol. **30**(16), 2637–2645 (2012).
12. P. De Heyn, S. Verstuyft, S. Keyvaninia, A. Trita, and D. Van Thourhout, "Tunable 4-Channel Ultra-Dense WDM Demultiplexer with III-V Photodiodes Integrated on Silicon-on-Insulator," in *Asia Communications and Photonics Conference*, Technical Digest (CD) (Optical Society of America, 2012), paper ATh2B.1.
13. S. Kandula, S. Sengupta, A. Greenberg, A. Patel, and R. Chaiken, "The Nature of Datacenter Traffic: Measurements & analysis", in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. (ACM, New York, 2009), pp. 202–208.
14. T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding Data Center Traffic Characteristics," ACM SIGCOMM Computer Communication Review **40**(1), 92–99 (2010).
15. Analyzing data using eye diagrams," Agilent, 2009. *http://na.tm.agilent.com/plts/help/WebHelp/Analyzing/Analyzing_Data_using_Eye_Diagrams.html*.

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014

1. Introduction

Data center network (DCN) is continuously growing in size and complexity to accommodate the increasing demand of high performance computers (HPCs) and data-intensive applications [1,2]. Most current DCNs are configured in a hierarchical structure due to limited port-count and speed of electrical switches. However, within this architecture, reaching higher interconnectivity level aiming at future DCN will greatly deteriorate the performance of the bandwidth and the latency [3–5]. Flat network architecture that supports any-to-any connectivity is a promising solution for the above issues. Several research projects [6–8] are investigating large port-count and low latency optical packet switches (OPS) in order to flatten the DCN and thus to eliminate the communication bottleneck of current DCNs fat-tree topology. The numerical and experimental demonstration of a novel Wavelength Division Multiplexing (WDM) OPS architecture with highly distributed control (see Fig. 1) was reported in [9]. Port-count independent reconfiguration time and good scalability are benefited from the proposed modular structure. The WDM OPS forwards the packets to the output ports based on the information carried by the labels. Flow control between the WDM OPS and the edge nodes with packet retransmission is considered when packet contentions occur at the OPS. During the packet processing and eventually the packet retransmission, packets are stored in costly high speed electronic buffers at the edge nodes and will only be released from the buffers in response to a positive acknowledgment of transmission. Minimizing the overall buffer size will result in a system not only with lower cost and power consumption, but also with a lower latency.

At physical layer, the size of the electronic buffer at the edge nodes depends on the WDM OPS packet forwarding time (reconfiguration of the switch), the delay of the flow control signaling, and the retransmission time. Although the highly distributed control of the WDM OPS architecture allows the forwarding of the packets within 25 ns [10], the flow control latency will largely affects the buffer size. In particular, the latency associated with the flow control hardware can be substantially minimized if the flow control is implemented directly at the optical level.

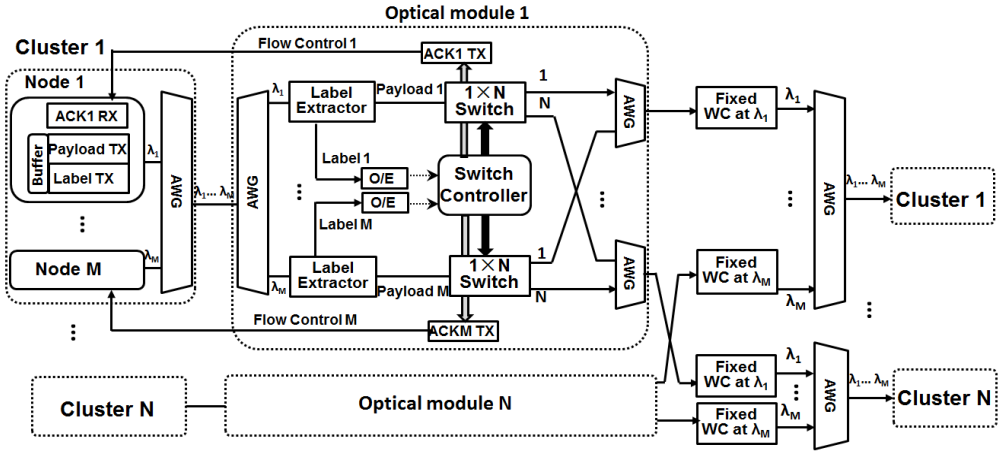

Fig. 1. WDM OPS architecture with highly distributed control.

In this work we experimentally demonstrate a highly spectral efficient and low latency optical flow control based on an optical bi-directional system that exploits direct re-modulation of the optical in-band label wavelength to transmit the flow control signaling back to the edge nodes. Only one single laser is needed to transmit both the optical label information and the flow control signaling. The data plane network will transport also the flow control signaling, making a dedicated flow control network redundant. The proposed bi-directional system can effectively reduce system complexity, cost and power consumption.

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014

The paper is organized as follows. In section 2, we present the proposed optical flow control system. In section 3, we experimentally demonstrate the dynamic operation of flow control and evaluate the performance of payload and acknowledgment signal. Conclusions are given in section 4.

2. Optical flow control system

The modular WDM OPS architecture with highly distributed control is shown in Fig. 1. The WDM OPS has N independent optical modules and interconnect N different clusters. Each of the N input/output fibers carries M WDM wavelength channels, indicated by λ_1 to λ_M. At the OPS node, the WDM packets are de-multiplexed and processed by a label extractor that separates the optical label from the optical payload. The label information is detected and processed by the switch controller. The optical payload is fed into the 1xN optical switch. The switch controller processes the label information, checks possible contentions, and sets the 1xN optical switches to block the contended packets with low priority and to forward packets with high priority. The fixed wavelength converters (FWCs) in the WDM OPS architecture allow the optical modules to operate independently by converting the input packets of the FWC to a distinct wavelength output, avoiding therefore contention of the packets coming from different modules and destining the same output fiber [10]. Moreover, the switch controller generates the acknowledgment (ACK) signals to acknowledge the edge nodes on the reception or re-transmission of the packets. A typical flow control currently implemented between the WDM OPS and the edge nodes is also shown in the figure. Notice that flow control signal will greatly affect the overall performance of system latency and the minimum buffer size.

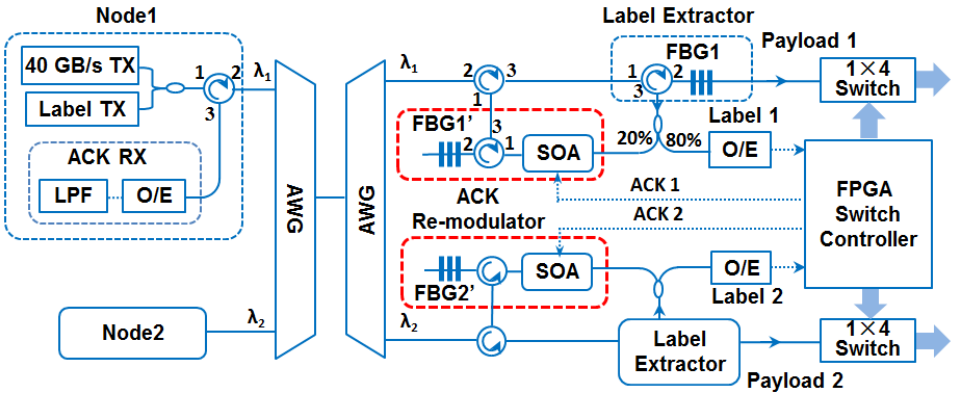

Fig. 2. Experimental set-up of optical flow control system.

We employed a novel optical flow control technique to efficiently transmit the in-band optical label information and flow control signals on a single wavelength, avoiding multiple wavelengths or optical connecting fibers, as schematically reported in Fig. 2. The optical packets consisting of synchronized data payload and in-band RF tone label signal are transmitted to OPS node. The advantage of the in-band RF tone label technique is that it saves bandwidth resources and since label wavelength carries several binary coded RF tones, parallel processing for each bit is allowed which greatly reduce the OPS processing time to few nanoseconds regardless of the bit-count [11]. RF tones are placed at high-frequency region (>100MHz) while the base-band is virtually empty (employed to transmit back the ACK signal). Moreover, the in-band label can be extracted by using narrow band filter such as a fiber Bragg grating (FBG) or an integrated micro-ring resonator (MRR) [12]. After label extraction, the power of the extracted label is split into two parts. The first part is used for label detection after optical-to-electrical conversion (O/E) and then processed by the switch controller. The other part is sent to a SOA-based ACK Re-modulator driven by the base-band ACK signal generated by the switch controller. To transmit the ACK, we exploit the available

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014

base-band bandwidth avoiding the potential crosstalk with the RF tones that are transmitted at frequencies > 100MHz. This allows to simply re-use the same label wavelength without any additional and complicated label eraser or the need of extra lasers and the corresponding wavelength registration circuitries. The re-modulated label wavelength at λ_{L1} and λ_{L2} carrying the ACK signals are sent back to the edge node via a second FBG1' and FBG2' with similar characteristic as the FBG1 and FBG2, respectively. The use of two FBGs is to investigate the effect of the double pass filtering in case a photonic integrated MRR [12] would have been used in the set-up. At edge node, ACK receiver (RX) detects the ACK signal and then conducts buffer manager to fulfill the flow control.

3. Experimental setup and results

We experimentally evaluate the optical flow control based on the bi-directional system by employing the system set-up shown in Fig. 2. At the edge node, an FPGA acts as buffer manager that stores the transmitted label information and processes the ACK signals that acknowledge whether the transmission is successful. Two sets of 40 Gb/s NRZ WDM packets at λ_{P1} = 1544.9nm and λ_{P2} = 1548.0nm are generated. Packet length in real scenarios is mostly found to be a bimodal distribution around 40B and 1500B which match the Ethernet minimum and maximum lengths [13,14]. We set the packet length at the maximum value of 1500B (300ns) with a guard time of 30ns to test the flow control operation which is actually independent of the packet lengths. The label wavelengths are centered at λ_{L1} = 1545.1nm and λ_{L2} = 1548.2nm to match the pass-band of FBG. For the proof of a 4 × 4 switch system, each label wavelength carries two binary coded RF tones. The number of tones can potentially scale up to at least 30 which will represent a large number of ports with constant processing time [11]. Figure 3(a) illustrates the electrical spectrum of the two RF tones (f1 = 280MHz, f2 = 650MHz) of the label we used in this experiment.

The average optical power of the payload and the label is 2.5dBm and 0dBm, respectively. At the OPS input, the labels are extracted by the FBG1 and FBG2 centered at λ_{L1} and λ_{L2}. The FBGs have a 3-dB bandwidth of 6 GHz to avoid spectral distortion of the optical payload. Figure 3(b) shows the optical spectrum of the signals before and after label extractor for Node 1. Note that integrated MRRs [12] with photo-detectors could replace the FBGs, circulators and discrete photo-detectors.

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014

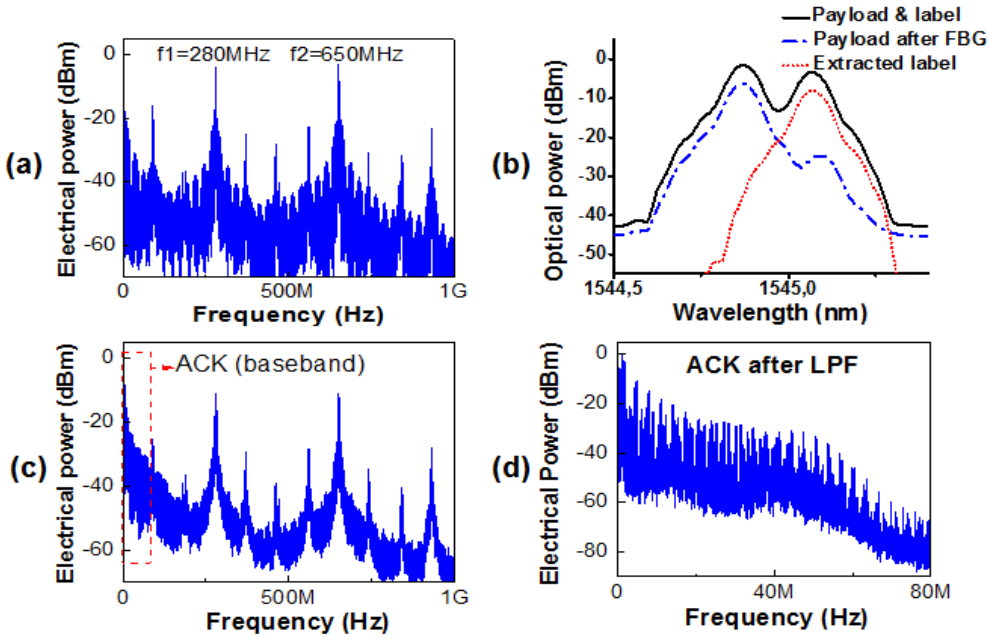

Fig. 3. (a) Electrical spectrum of the RF tone label. (b) Optical spectrum before and after label extractor. (c) Electrical spectrum of detected re-modulated signal. (d) Electrical spectrum of ACK after LPF.

The optical power of the extracted label is split by a 20:80 coupler. The 80% of the power (−8dBm) is detected and processed by an FPGA-based switch controller. The 20% of the power is fed into a SOA employed as optical modulator driven by the FPGA switch controller to generate the ACK signals. The typical driving current required by the SOA is 30 mA which can be directly provided by the digital pins of the FPGA switch controller. The ACK is then transmitted back on the same optical link to the edge node within the same WDM channel with no extra cost on bandwidth and space. Figure 3(c) shows the electrical spectrum of re-modulated signal after detection at the edge node. It is clearly visible of the two RF tones of the label at 280 MHz and 650 MHz, and the re-modulated base band ACK signal. The electrical bandwidth has been well exploited which also prevents the possible cross talk caused by re-modulation. The ACK signal could be easily retrieved by using a 50 MHz low pass filter (LPF) as shown in Fig. 4(d). According to the positive or negative detected ACK, the buffer manager releases the stored label (positive ACK) or retransmits the packet (negative ACK).

3.1 Dynamic operation

Figure 4 shows the dynamic operation of the flow control. Original label bits that should be transmitted are given at the top. They are sent out and stored in the buffer in case a contention happens. The contention resolution algorithm is based on a fixed priority: packets at λ_{P1} have higher priority. When a contention occurs between the packets from two different edge nodes destined at same output, packets at λ_{P1} will be always forwarded to the destination and a pulse signal with the same packet duration drives the SOA-based modulator to generate the positive ACK signal, while the packet at λ_{P2} will be blocked and a negative ACK will be sent back requesting packet retransmission. If the buffer manager at edge node detects a negative ACK, the corresponding stored label information will be retransmitted. The actual transmitted labels including retransmission are shown in the middle in which the contended packets are marked. At the bottom of Fig. 4, it shows the ACK1 and ACK2 generated by the switch controller and applied as driver signals to the SOA-based modulators. We can also see from the figure that it takes 160ns for labels transmitting and being detected by switch controller which includes

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014

40ns processing time. And 105ns including 5ns processing time is needed from ACK being generated until finally detected by buffer manager. The latency or the round trip time (RTT) is 265ns.

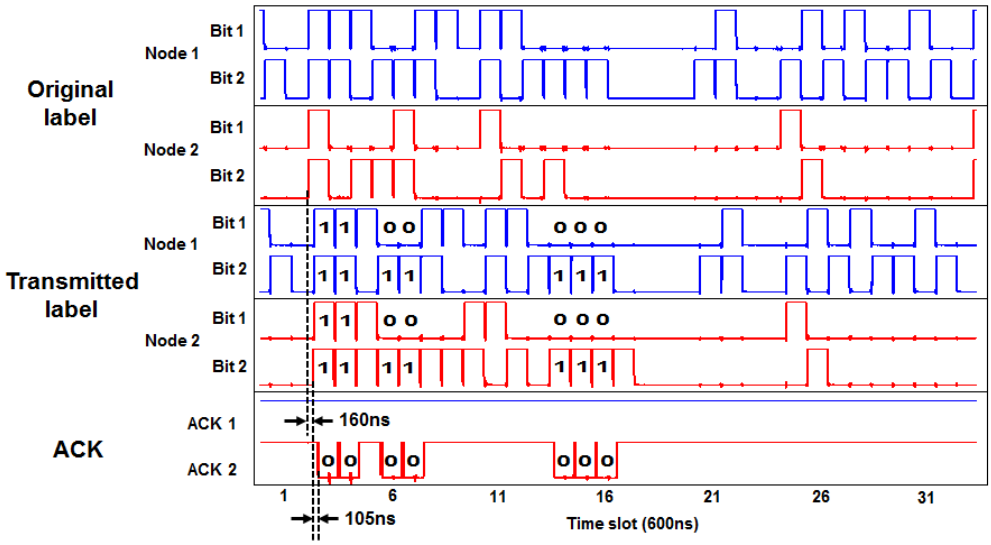

Fig. 4. Dynamic operation including retransmission.

Note that in this experiment since the packet duration is longer than the RTT, the buffer manager at edge node is already detecting the ACK information during the transmission of the same packet. Therefore, sending a new packet or retransmission of the blocked one is decided before the next time slot starts. However, the distance between clusters and OPS node may vary from several meters to kilometers in a DCN. In the case that the RTT is longer than the packet duration, the corresponding ACK will arrive after the entire packet was sent out. A bigger buffer size is needed (typically at least equal to the RTT) and more importantly, the buffer manager should correctly judge to which transmitted packet correspond the received ACK.

To this aim, we implemented a time slot based buffering technique that dynamically configures the size of a shifting buffer area according to the tested RTT. The operation is the following. When the label is being sent out, a copy will be fed into the shifting buffer, and forwarded to next stage at the beginning of each slot. The number of the stages is determined by the RTT and packets duration. Therefore the detected ACK refers to specifically the coming out label from the shifting buffer. A new packet will be transmitted and stored in the buffer in response to a positive ACK otherwise the coming out label will be sent out and go through this procedure again.

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014

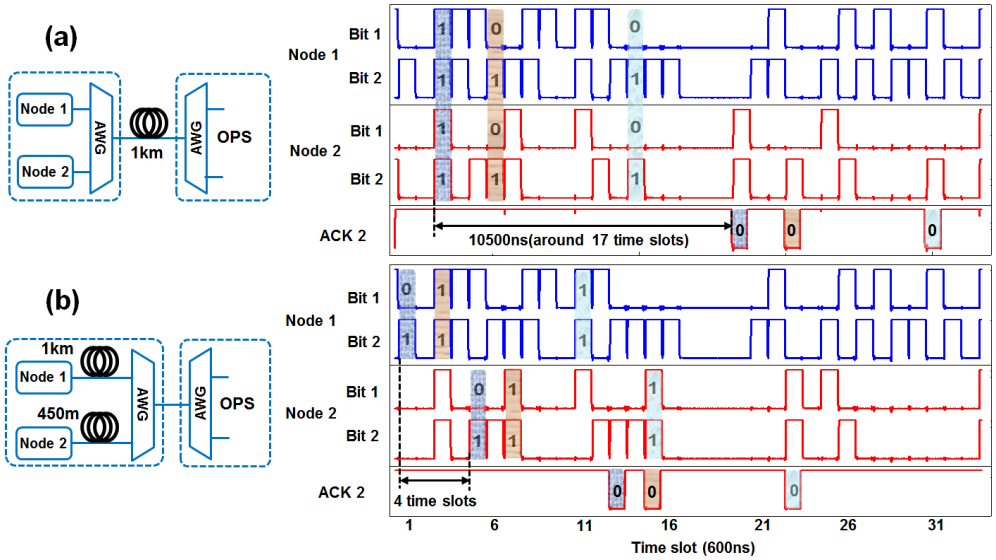

Fig. 5. Dynamic operation with variable distance. (a) Both nodes with1km from OPS. (b) Node1 with1km and Node 2 with 400m.

Using the system set-up in Fig. 2, we increase the distance between the edge nodes and OPS by adding a 1km single mode fiber (SMF) (as shown in Fig. 5(a)) to evaluate the flow control operation with larger RTT. The traces of transmitted label and detected ACK at edge node are given at right side. The reception of ACK is delayed by 10.5μs (around 17 slots) mainly caused by propagation. During this time the transmitted data is stored and shifted in the buffer until the corresponding ACK arrived. At next time slot, a new data or the previously contented one will be transmitted according to the received ACK. Figure 5(b) illustrates the results when two edge nodes have different distance (1km/450m) from OPS node. In this case, the buffer manager at each node sets different number of stages to match the RTT. Since it takes more time which is around 4 slots for packets from Node 1 to reach OPS, the contention may happen between the Node 1 packet and 4-slot later packet from Node 2. From above results we can see that by using shifting buffer, the edge node could handle the situation of variable distance between edge node and OPS. The size of this buffer area is configured according to the RTT to match the timing of detected ACK and corresponding label. The time traces of label including retransmission and ACK in Fig. 4 and Fig. 5 clearly indicate the successful demonstration of the optical flow control within variable distance from OPS node.

3.2 Payload performance

The separated payload at λ_{P1} and λ_{P2} to be fed into the 1x4 optical switches (not employed in the experiment) are detected and analyzed by a BER tester to evaluate the filtering and the possible crosstalk caused by the bi-directional system. Figure 6(a) shows the BER curves of the back-to-back and after the label extractor as well as the eye diagrams. Error free operation with only 0.5dB power penalty for 40Gb/s payload has been measured indicating no distortion has been caused due to the label and ACK transmission.

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014

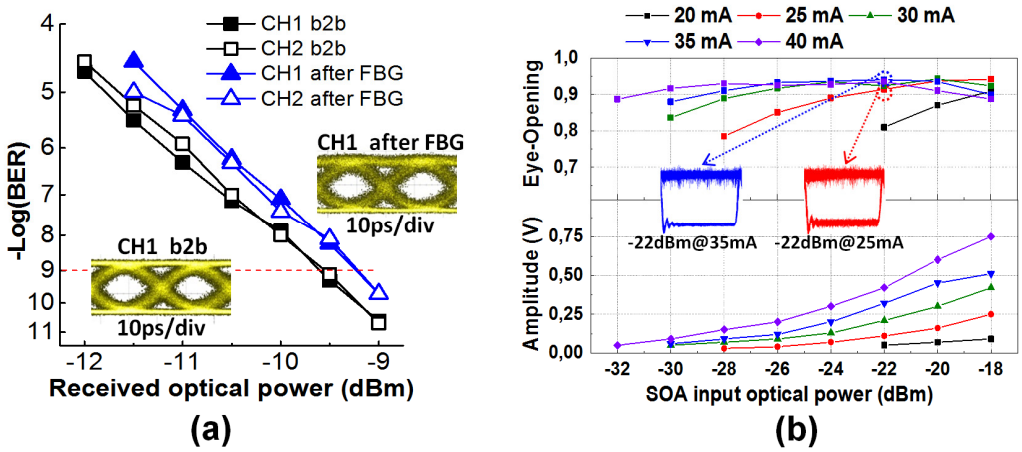

Fig. 6. (a) 40 Gb/s payload BER curves and eye diagrams. (b) Eye-opening and amplitude of detected ACK.

3.3 ACK performance

Although the DCN is a close environment with respect to a telecom network, variation in distance and in power could still occur. Therefore we further estimated the eye-opening [15] and amplitude of detected ACK as a function of the label power fed into the SOA re-modulator and the SOA driving current. Figure 6(b) indicates that at 30mA, the eye-opening factor is higher than 0.8 with input power ranging from −30dBm to −18dBm. An even larger dynamic range is possible with higher driving current. As a result, the optical bi-directional system is quite robust to power fluctuation and thus to distance variation within the DCN. Low input optical power indicates that no impact has been caused on label detection that only 1% of label power is enough for the flow control operation. Moreover, low SOA driving current shows that this technique is not just spectral efficient but also power efficient that the FPGA pins could directly drive the SOA. Considering the contributions to the energy consumption given by the low speed O/E converter for ACK detection (540mW) and SOA for ACK re-modulation (80mW), the flow control operation will introduce only 15.5 pJ/bit more power consumption compared with the WDM OPS systems presented in [9] and [10].

4. Conclusion

We propose and experimentally demonstrate a highly efficient optical flow control technique for intra DCN based on label wavelength re-modulation. A small portion of label power is re-modulated by ACK signal in SOA and transmitted back within same WDM channel. Thus one single laser source is used for both label and ACK generation. Bi-directional system which removes a dedicate flow control link greatly reduce the system complexity.

Experimental results validate the error free bi-directional transmission with low latency and variation in distance form edge node to OPS node has also been considered. Investigation on ACK signal shows that our system is of low power consumption and robust in facing the problem of power fluctuation. The parallel transmission of the label and the ACK signals allows an asynchronous and parallel processing leading to a substantial reduction of the system complexity and latency.

Acknowledgment

This work has been supported by the FP7 European Project LIGHTNESS (FP7-318606).

#197054 - $15.00 USD Received 14 Oct 2013; revised 2 Dec 2013; accepted 3 Dec 2013; published 2 Jan 2014