Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system

Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system

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

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

Abstract: We propose and demonstrate an optical flat datacenter network based on scalable optical switch system with optical flow control. Modular structure with distributed control results in port-count independent optical switch reconfiguration time. RF tone in-band labeling technique allowing parallel processing of the label bits ensures the low latency operation regardless of the switch port-count. Hardware flow control is conducted at optical level by re-using the label wavelength without occupying extra bandwidth, space, and network resources which further improves the performance of latency within a simple structure. Dynamic switching including multicasting operation is validated for a 4x4 system. Error free operation of 40 Gb/s data packets has been achieved with only 1 dB penalty. The system could handle an input load up to 0.5 providing a packet loss lower that 10^{-5} and an average latency less that 500ns when a buffer size of 16 packets is employed. Investigation on scalability also indicates that the proposed system could potentially scale up to large port count with limited power penalty.

OCIS codes: (060.4259) Networks, packet-switched; (060.6719) Switching, packet; (200.4650) Optical interconnects.

References and links

1. S. Sakr, A. Liu, D. Batista, and M. Alomari, "A survey on large scale data management approaches in cloud environments," IEEE Commun. Surv. Tutorials **13**(3), 311–336 (2011).
2. M. Meeker and L. Wu, "2013 internet trends," Kleiner Perkins Caufield & Byers, Technical Report (2013).
3. G. Astfalk, "Why optical data communications and why now?" Appl. Phys. A Mater. Sci. Process. **95**(4), 933–940 (2009).
4. S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: measurements and analysis," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement* (ACM, New York, 2009), pp. 202–208.
5. J. Oltsik, B. Laliberte, and B. Lundell, "Research report: Data center networking trend," Enterprise Strategy Group (2006).
6. L. A. Barroso and U. Hölze, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* (Morgan and Claypool, 2009).
7. P. Gill, N. Jain, and N. Nagappan, "Understanding network failures in data centers: measurement, analysis, and implications," in *Proceedings of the ACM SIGCOMM 2011 Conference* (ACM, New York, 2011), pp. 350–361.
8. C. Kachris, K. Bergman, and I. Tomkos, *Optical Interconnects for Future Data Center Networks* (Springer, 2013), Chap. 1.
9. A. Benner, "Optical interconnect opportunities in supercomputers and high end computing," in *Optical Fiber Communication Conference*, Technical Digest (CD) (Optical Society of America, 2012), paper OTu2B.4.
10. C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," IEEE Commun. Surv. Tutorials **14**(4), 1021–1036 (2012).
11. O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, and K. Bergman, "The data vortex optical packet switched interconnection network," J. Lightwave Technol. **26**(13), 1777–1789 (2008).
12. 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. 1–12.

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014

13. 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.
14. 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).
15. W. Miao, S. Di Lucente, J. Luo, H. Dorren, and N. Calabretta, "Low latency and efficient optical flow control for intra data center networks," in *European Conference and Exhibition on Optical Communication* (Optical Society of America, 2013), paper Th.1.A.2.
16. 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.
17. T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," ACM SIGCOMM Comput. Commun. Rev. **40**(1), 92–99 (2010).

1. Introduction

Emerging services such as cloud computing and social networks are steadily boosting the Internet traffic [1,2]. The huge volumes of packetized data travelling to and from the data centers (DCs) are generated to satisfy users requests which present only a small fraction of the total traffic handled by these systems [3], putting a tremendous pressure on the DC networks (DCNs). In large DCs with 10.000's of servers, merchant silicon top-of-the rack (TOR) switches are used to interconnect servers in a group of 40 per rack with 1 Gb/s link (10 Gb/s expected soon). To interconnect the 100's of TORs, with 10/40 Gb/s aggregated traffic (100Gb/s is expected soon) per TOR, current DCN is built up on multiple switches, each with limited port count and speed, organized in fat-tree architecture [4,5]. This multi-layer topology has intrinsic scalability issues in terms of bandwidth, latency, costs and power consumption (large number of high speed links) [6,7] and they are becoming critically limiting figures of merit in the design of future DCNs.

To improve the performance and lower the operation costs, flattened DCN is currently being widely investigated. To this aim, large port count switches with high speed operation, low latency and power consumption are the basic blocks to realize a flat DCN [8]. Photonic interconnect-based technologies have the potential of efficiently exploiting the space, time, and wavelength domains, which leads to significant improvements over traditional electronic network architecture for scaling up the port count while switching high speed data at nanoseconds time scale with low energy per switched bit and small footprint photonic integrated devices [8,9]. Despite the several optical switch architectures presented so far [10–13], no one has been proved a large number of ports while providing a port-count independent reconfiguration time for low latency operation. Moreover, the lack of a practical optical buffer demands complicated and unfeasible system control for store-and-forward operation and contention resolution.

In this work, we propose and experimentally investigate a novel flat DCN architecture for TOR interconnect based on a scalable optical switch system with hardware flow control for low latency operation. Experimental evaluation of a 4x4 optical switch system with highly distributed control has been carried out. The hardware flow control at the optical level allows fast retransmission control of the electrically buffered packets at the edge nodes preventing the need of optical buffers. Moreover, this makes a dedicated flow control network redundant, which effectively reduces system complexity and power consumption. Experiment results demonstrate dynamic switching operation including multicasting and only 1dB power penalty has been observed for 40Gb/s payload. A buffer size of 16 packets sufficiently guarantees $<10^{-5}$ packet loss for 0.5 input load and less than 500ns average end-to-end latency could be achieved within 25m distance. Scalability investigation also indicates that the optical switch can potentially scale up to more than 64 × 64 ports with less than 1.5dB penalty while the same latency is retained.

2. System operation

The proposed flat DCN based on N × M highly distributed controlled optical packet switching (OPS) architecture is shown in Fig. 1. Each cluster groups M TORs and an aggregation

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014

controller is used for balancing the traffic load and aggregating the input data coming from different TORs. Packetized data will be assigned with different wavelength λ_1, λ_2 …, λ_M and transmitted to OPS node. Switching is performed based on the in-band label information carried by each packet [14]. After the packet being sent out, aggregation controller will store the copy in a FIFO until receiving a positive acknowledgment that the packet has been transported to proper destination.

OPS node consists of N identical modules and each of them handles the packets from the corresponding cluster. Label extractor separates the optical label from the optical payload by using a fiber Bragg grating (FBG). The optical payload is then fed into the SOA based broadcast and select 1xN switch while the extracted label is split into two parts. One of them is detected and processed by the label processor (LP) after optical-to-electrical conversion (O/E). The switch controller retrieves the label bits, checks possible contentions and configures the 1xN switch to block the contended packets with low priority and to forward packets with high priority. Moreover, the switch controller generates the acknowledgment (ACK) used to inform the aggregation controller on the reception or re-transmission of the packets. The other part of label power is re-modulated in an RSOA driven with the base band ACK signal generated by the switch controller and sent back to cluster side within the same optical link [15]. This fulfills the efficient optical flow control in hardware which minimizes the latency and buffer size. Baseband ACK signal is easily extracted at the edge node by using a 50 MHz low pass filter, to remove the label information at RF frequencies. The adopted modular structure allows highly distributed control which makes the reconfiguration time of the overall switch port-count independent. In addition, the M channels of each cluster could be processed in parallel, greatly minimizing processing time and thus the latency [16].

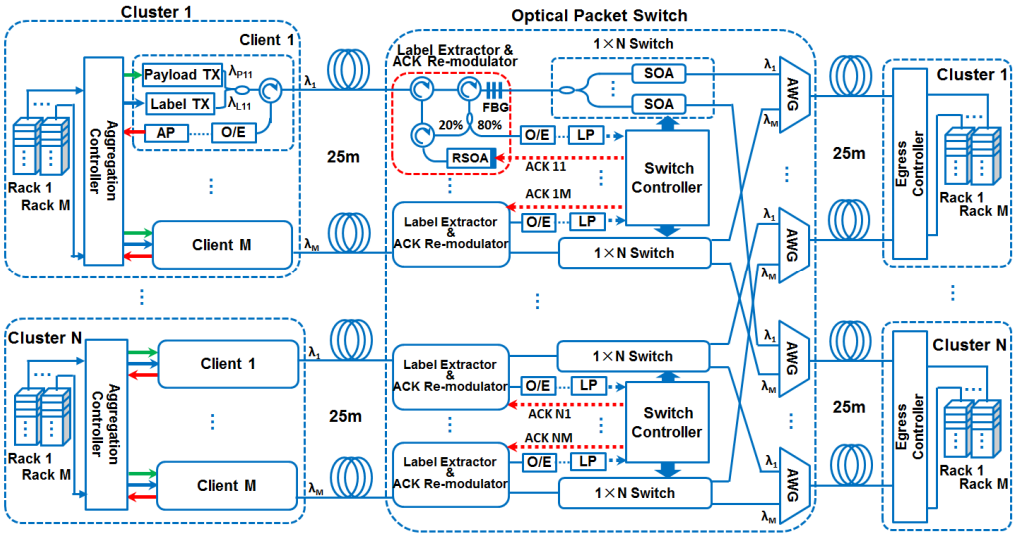

Fig. 1. Proposed flat DCN architecture based on OPS with flow control.

3. Experimental setup and results

For the validation of the DCN, we experimentally investigate the full dynamic operation including flow control of a 4 × 4 system with 25m transmission link. Packetized 40Gb/s NRZ-OOK payloads are generated with 540ns duration and 60ns guard time. The operation of OPS node is actually independent of packet length that shorter or longer duration are both supported. An FPGA acts as aggregation controller that generates for each packet the label according to the port destination, and simultaneously provides a gate signal to enable the transmission of the payload with a certain load. Buffer manager inside FPGA stores the label information in a FIFO queue with a size of 16 packets and removes the label from the queue

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014

in response to a positive ACK. Otherwise the label and payload are retransmitted after re-signaling the input optical gates, implementing the packets retransmission.

RF tone in-band labeling technique and bi-directional optical system are deployed to efficiently transmit the label and ACK in a single fiber. Such labeling technique allows the parallel processing of the label bits which will greatly reduce the OPS processing time [14]. Here we use two RF tones (f1 = 284.2MHz, f2 = 647.1MHz) for coding the 2-bit binary label information. Payload wavelengths are placed at $\lambda_{P11} = \lambda_{P21} = 1544.9$nm and $\lambda_{P12} = \lambda_{P22} = 1548.0$nm. The label wavelengths, each carrying two RF tones, are centered at $\lambda_{L1} = 1545.1$nm and $\lambda_{L2} = 1548.2$nm. The average optical power of the payload and the label at the OPS input is 2.5dBm and −2dBm, respectively. Pass band of FBG is centered at label wavelength and has a −3dB bandwidth of 6 GHz. This narrow bandwidth could avoid spectral distortion of the payload. Optical spectra of the packets before and after label extractor for Cluster1 are shown in Figs. 2(a) and 2(b). A small portion as low as 1% of the label power will be re-used by modulating the ACK signal on the available base-band bandwidth avoiding the potential crosstalk with the RF tones that are transmitted at frequencies > 100 MHz [15]. The generated flow control signal could reach the transmitter side and trigger retransmission without any additional and complicated label eraser or the need of extra lasers and the corresponding wavelength registration circuitries. Considering the overall contributions to the energy consumption given by low speed O/E converter (540mW × 4), label processor (210mW × 4), switch controller (1W × 2), SOA based switch (80mW × 8) and ACK Re-modulator (80mW × 4), the total energy consumption for the 4 × 4 system is 37.25 pJ/bit.

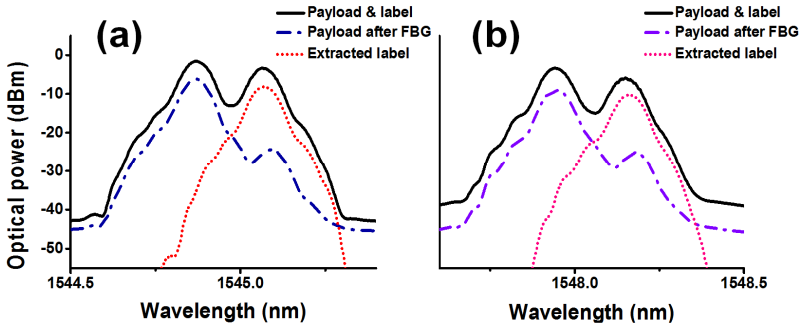

Fig. 2. Optical spectrum before and after label extractor for (a) Client1 and (b) Client2.

3.1 Dynamic operation

To investigate the dynamic operation of the flow control and the payload switching of the system, optical packets are generated with a typical DC traffic load of 0.5 at the clusters side [17]. Figure 3 shows the dynamic generation/retransmission of the label and the payload from both clusters (each color represents one client). The time traces of the label detected by the switch controller and the ACK feed-back detected by the aggregation controller at the transmitter side are reported at the top of Fig. 3. 2-bit label brings up 3 possibilities of switching since “00” represents no packet. “01” stands for output1, “10” for output2, and “11” for multicasting the payload to both ports. To clearly show the contention and switching mechanism, fixed priority has been adopted in our contention resolution algorithm. If two packets from different clients have the same destination, packet from Client 1 will be forwarded at the output while the packet from Client 2 will be blocked and a negative ACK will be sent back requesting packet retransmission. If Client1 is multicast, any data in Client2 will be blocked. Multicasting for Client2 will only be approved if Client1 is not transmitting any packet.

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014

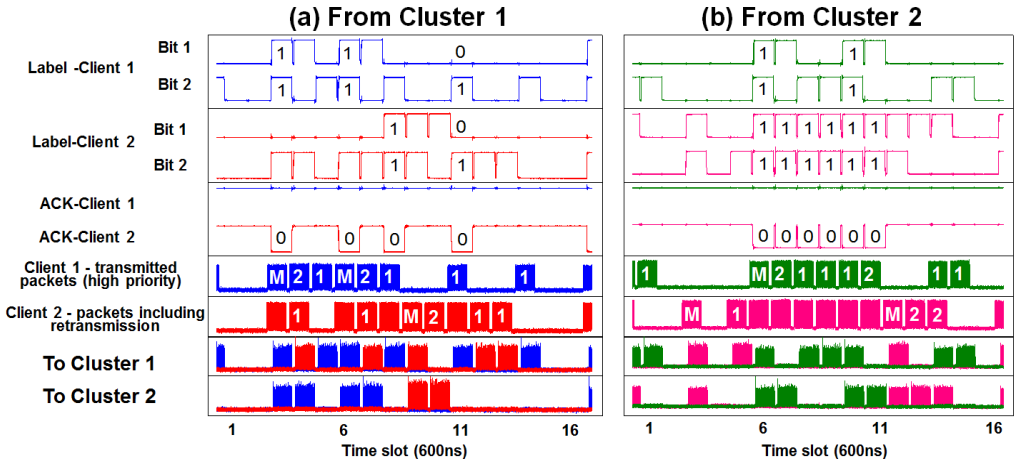

Fig. 3. Dynamic operation of labels and payloads from (a) Cluster1 and (b) Cluster2.

One or both of the SOAs will be switched on to forward the packets to the right destination. The waveforms of the transmitted packets (including retransmitted packets for Client 2) and the switch outputs are shown at bottom of Fig. 3. Flag "M" stands for the packets to be multicast, which should be forwarded to both output ports. If Client 2 contends with Client 1 the packets will be blocked (shown with unmarked packets in Fig. 3). In this case, a negative ACK is generated to inform the buffer manager of the transmitter that the packets have to be retransmitted. Figure 3 clearly shows the successful optical flow control and multicasting operation. The minimum end-to-end latency (no retransmission) is 300ns including 250ns propagation delay provided by 2 × 25m link.

At switch output, a bit-error-rate (BER) analyzer is used to evaluate the quality of the detected 40Gb/s payload. Figure 4 shows the BER curves and eye diagrams for packets from 4 different clients. Test results for back-to-back (B2B) as well as the signal after the transmission gate are also reported as reference. It is clear that the transmission gate used to set the traffic load does not cause any deterioration of the signal quality. Error free operation has been obtained with only 1dB penalty after switch which is mainly due to the in-band filtering caused by label extractor and noise introduced by SOA switch. It proves that high data-rate operation is supported by our system and no distortion has been introduced by the bi-directional transmission of label and flow control signal.

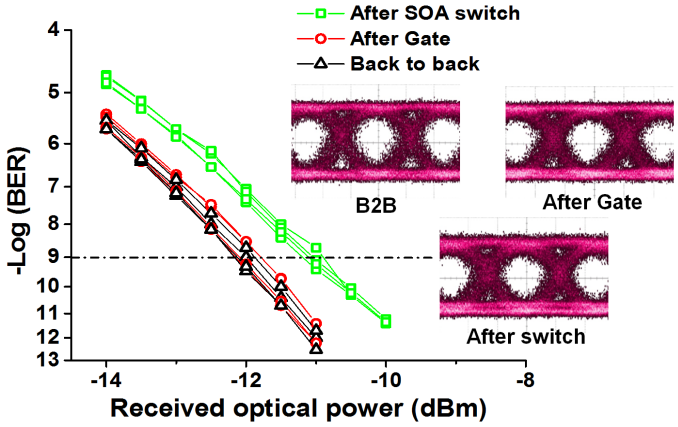

Fig. 4. BER curves and eye diagrams for 40 Gb/s payload.

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014

3.2 Packet loss and latency

To further investigate the performance of the 4 × 4 system with the flow control mechanism, the packet loss and the average latency are tested. As discussed in the previous section, the label that represents the packet's final destination is generated by the aggregation controller and stored in the finite-size FIFO queue. It will be released from the FIFO once the packet has been successfully forwarded. In this case the aggregation controller will receive a positive flow control signal. Otherwise, the packet will be retransmitted. However, if the FIFO is already fully occupied and there is a new packet to be served at the next time slot, this packet will be instantly dropped and considered lost due to buffer overflowing. The packet loss is then calculated as the ratio of the number of lost packets to total number of generated packets.

For the 4 × 4 system, at each time slot, the aggregation controller will generate a packet for each different client with the same average traffic load. The destinations decided by the label pattern are chosen randomly between the two possible outputs according to a uniform distribution. Based on the label information, the switch controller forwards the packets to the right output and if a contention occurs, only the packet with higher priority will be properly delivered. Instead of using a fixed priority for the contention resolution algorithm, a round robin scheme is employed as priority policy to efficiently balance the utilization of the buffer and the latency between the two clients. This means that the priority will be assigned slot by slot. As a result, a packet in the FIFO will be definitely sent to the proper destination within two time slots, and the respective buffer cell will be released.

Figure 5(a) shows the packet loss for different input loads and buffer sizes. The total amount of time considered is 2×10^{10} time slots. As expected the packet loss increases with the input load. Larger buffer size could improve the packet loss performance for input loads smaller than 0.7. Larger buffer capacity does not bring significant improvement when the load ≥0.7 because the buffer is always full and overflowing causing high packet loss. Figure 5(b) presents the buffer occupancies when traffic load equals to 0.5, 0.6, 0.7 and 1, respectively. For the first 200 time slots, it is clear that for load = 1, the 16-packet buffer is rapidly filled up and for load = 0.7 the buffer is fully occupied most of the time which will cause the buffer overflowing.

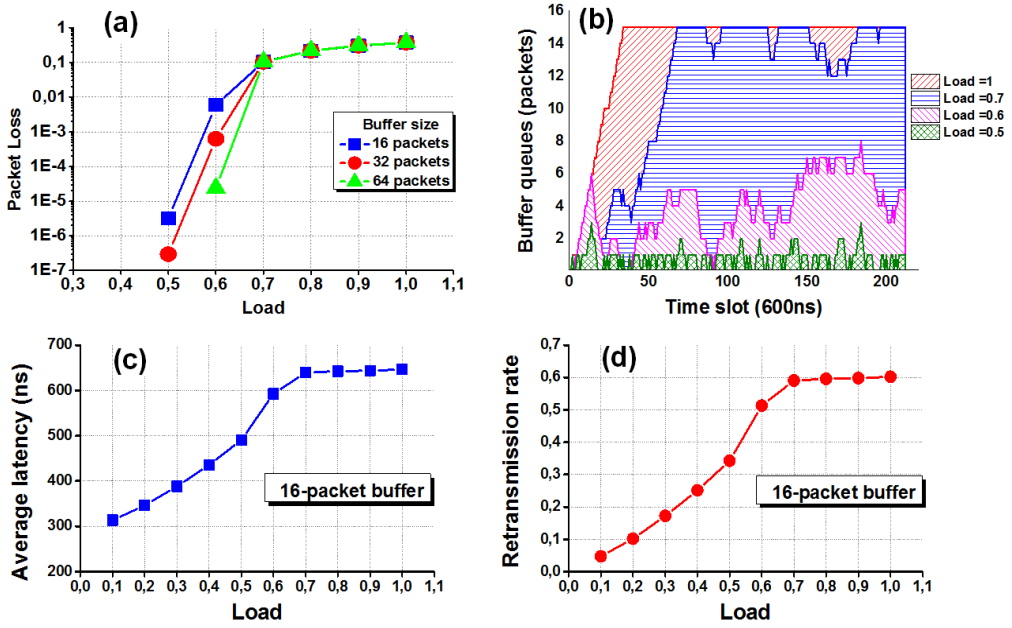

Fig. 5. (a) Packet loss vs. load with different buffer size. (b) Buffer queue occupancy for different input load. (c) Average latency with buffer size of 16 packets. (d) Retransmission rate with buffer size of 16 packets.

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014

Average end-to-end latency for the system with a buffer size of 16 packets is reported in Fig. 5(c). The number of packets that has been successfully forwarded without retransmission and the one that has been retransmitted once are recorded and employed to calculate the average latency. The lost packets are not considered in the latency calculation. Similarly to the packet loss curves, the average latency increases approximately linearly for input loads up to 0.7. As the traffic becomes heavier, the possibilities of contention also increase which results in more retransmissions, and thus larger latencies. However, when the load is higher than 0.7, the buffer is always full but the average latency remains constant since the round robin policy and the lost packets are not considered in the latency calculation. Indeed, due to round robin policy, every packet having entered in the buffer queue will finally win the contention within two time slots. This explains the saturation of the latency curve at 645ns which includes 250ns off-set latency caused by the 25m transmission link. Figure 5(d) shows the average retransmission rate which represents the contention probability as a function of the input load. It is calculated as the ratio of retransmissions to the total number of transmitted packets. The retransmission rate curve keeps the same shape as the latency one and saturates when the input traffic load exceeds 0.7 in which case the actual traffic inside the switch is reaching the maximum due to the retransmissions. From Fig. 5 it can be concluded that the system could handle an input load up to 0.5 providing a packet loss lower than 10^{-5} and an average end-to-end latency lower than 500ns.

3.3 Scalability

In this section we investigate the system scalability in order to support a large port count. The total number of ports supported by the OPS is given by N × M because of the presence of N modules and M clients in each module. The performance of the overall system could be translated into the performance of 1 × N optical switch due to the identical structure of N modules. In this scenario, the main limiting factor for scaling the OPS is the splitting loss experienced by the payload caused by the 1 × N broadcast and select stage. Therefore we employed a variable optical attenuator (VOA) to emulate the splitting losses, as schematically reported in Fig. 6(a). At the output of the SOA switch, the BER and the OSNR are measured to evaluate the payload quality.

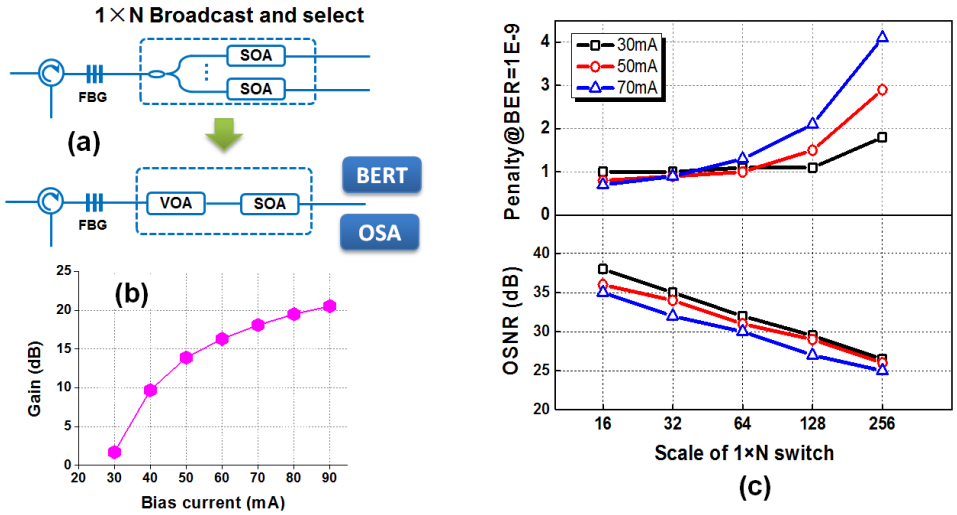

Fig. 6. (a) Set-up for scalability investigation. (b) Gain characteristic with bias current of the SOA switch (c) Penalty and OSNR vs. scale of 1 × N switch.

The input optical power of the 1 × N optical switch is 0dBm and the attenuation caused by the VOA is set to be 3dB × $\log_2$N. The SOA will be switched on to forward the packet, and at the meantime to amplify the signal. Figure 6(b) gives the gain characteristic versus bias

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014

current of the SOA from which we could see that the SOA operates transparently at 30mA and 18dB amplification could be supplied when biased at 70mA. Considering the splitting loss, the SOA could compensate the 18dB loss caused by the 1 × 64 broadcast stage resulting in a lossless 1 × 64 optical switch. Figure 6(c) shows the power penalty (measured at BER = 1E-9), and the OSNR of the switched output as a function of N for different SOA bias currents. A penalty of < 1.5 dB for N up to 64 is measured regardless of the bias current of the SOA. For N > 64 the penalty increases mainly caused by the deterioration of OSNR as a result of splitting loss. The BER performance gets worse when biasing at a higher current due to noise that becomes more prominent. The results clearly shows that when N < 64, less than 1.5 dB penalty is obtained for different driving current which indicates that the OPS under investigation could be potentially scaled up to a large number of ports at the expense of limited extra penalty. In addition, a lossless system without extra amplification could be achieved with the bias current of 70mA.

4. Conclusion

We experimentally demonstrate a fully operational 4x4 OPS system for the implementation of a flat DCN. Exploiting the highly distributed control architecture, the RF tone in-band labeling technique and the efficient optical flow control, we report 300ns minimum end-to-end latency (including 250ns offset introduced by the 25m transmission link) for 40 Gb/s packets. Dynamic switching results including multicasting prove the successful flow control operation. Error free operation with only 1 dB penalty shows that no distortion has been caused by the bi-directional transmission of the in-band label and flow control signal on the same optical link.

Packet loss and average latency are tested under different input load. By employing the round robin algorithm for contention resolution, a packet loss lower than 10^{-5} and an average end-to-end latency less than 500ns could be achieved under relatively high traffic load of 0.5 and limited buffer capacities of 16 packets. Increasing the buffer size could improve the performance in terms of packet loss for load values smaller than 0.7. Investigation on the switch scalability indicates that scaling up to 64 × 64 ports is possible at the expense of 1.5 dB extra power penalty while maintaining the same latency performance. The amplification introduced by SOA switch could compensate the splitting loss of the broadcast stage resulting in a lossless optical switch system.

Acknowledgment

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

#200765 - $15.00 USD Received 7 Nov 2013; revised 23 Dec 2013; accepted 23 Dec 2013; published 29 Jan 2014