Numerical and experimental study of a high port-density WDM optical packet switch architecture for data centers

S. Di Lucente,[*] J. Luo, R. Pueyo Centelles, A. Rohit, S. Zou,[1] K. A. Williams, H. J. S. Dorren and N. Calabretta

[1]*Eindhoven University of Technology, Department of Electrical Engineering, 5600MB Eindhoven, The Netherlands*
**s.di.lucente@tue.nl*

Abstract: Data centers have to sustain the rapid growth of data traffic due to the increasing demand of bandwidth-hungry internet services. The current intra-data center fat tree topology causes communication bottlenecks in the server interaction process, power-hungry O-E-O conversions that limit the minimum latency and the power efficiency of these systems. In this paper we numerically and experimentally investigate an optical packet switch architecture with modular structure and highly distributed control that allow configuration times in the order of nanoseconds. Numerical results indicate that the candidate architecture scaled over 4000 ports, provides an overall throughput over 50 Tb/s and a packet loss rate below 10^{-6} while assuring sub-microsecond latency. We present experimental results that demonstrate the feasibility of a 16x16 optical packet switch based on parallel 1x4 integrated optical cross-connect modules. Error-free operations can be achieved with 4 dB penalty while the overall energy consumption is of 66 pJ/b. Based on those results, we discuss feasibility to scale the architecture to a much larger port count.

OCIS codes: (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. Surveys & Tutorials **3**(13), 311–336 (2011).
2. L. A. Barroso and U. Hölze, "The datacenter as a computer: an introduction to the design of warehouse-scale machines," Synthesis Lectures on Computer Architectures **4**(1), 1–108 (2009).
3. R. Luijten, W. E. Denzel, R. R. Grzybowski, and R. Hemenway, "Optical interconnection network: The OSMOSIS project," Lasers and Electro-Optics Society **2**, 563–564 (2004).
4. H. Wang, A. Wonfor, K. A. Williams, R. V. Penty, and I. H. White, "Demonstration of a lossless monolithic 16x16 QW SOA switch," Proceedings ECOC 2009, PD 1.7, Vienna, Austria, 2009.
5. S. C. Nicholes, M. L. Mašanović, B. Jevremović, E. Lively, L. A. Coldren, and D. J. Blumenthal, "The world's first InP 8x8 monolithic tunable optical router (MOTOR) operating at 40 Gbps line rate per port," Proceedings OFC 2009 post deadline paper B1, San Diego, USA, 2009.
6. J. Gripp, M. Duelk, J. E. Simsarian, A. Bhardwaj, P. Bernasconi, O. Laznicka, and M. Zirngibl, "Optical switch fabrics for ultra-highcapacity IP routers," JLT **21**, 2839–2850 (2003).
7. C. Kachris and I. Tomkos, "A survey on optical interconnects for data Centers," IEEE Communication Surveys & Tutorials **PP**, no. 99, 1–16, 2012.
8. 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," Proceedings OFC 2012, OW3J.2, Los Angeles, USA, 2012.
9. OMNeT + + Network Simulation Framework, *http://www.omnetpp.org/*.
10. J. Luo, H. J. S. Dorren and N. Calabretta, "Optical RF tone in-band labeling for large-scale and low-latency optical packet switches," JLT **30**, 16, 2012.
11. S. Di Lucente, N. Calabretta, J. A. C. Resing, and H. J. S. Dorren, "Scaling low-latency optical packet switches to a thousand ports," JOCN **4**, no. 9, 2012.
12. T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," ACM SIGCOMM, 2010.
13. T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in Proc. Internet Measurement Conference (IMC), Melbourne, Australia, 2010.

#177554 - $15.00 USD Received 5 Oct 2012; revised 22 Nov 2012; accepted 23 Nov 2012; published 4 Jan 2013

14. R. Pueyo Centelles, S. Di Lucente, H. J. S. Dorren and N. Calabretta, "On the performance of a large-scale optical packet switch under realistic data center traffic," (submitted).
15. A. Rohit, A. Albores-Mejia, J. Bolk, X. Leijtens, and K. Williams, "Multi-path routing in a monolithically integrated 4x4 broadcast and select WDM cross-connect," Proceedings ECOC 2011, Geneva, Switzerland, 2011.
16. N. Kikuci, Y. Shibata, Y. Tomori, "Monolithically integrated 64-channel WDM channel selector," NTT R D **51**, 11, 2003.
17. N. Calabretta, R. Stabile, A. Albores-Mejia, K. A. Williams, and H. J. S. Dorren, "InP monolithically integrated wavelength selector based on periodic optical filter and optical switch chain," Opt. Express **19**(26 issue 26), B531–B536 (2011).
18. M. Matsuura, N. Kishi, and T. Miki, "Ultra-wideband wavelength conversion over 300 nm by cascaded SOA-based wavelength converters," PDP OFC 2006, Anaheim, USA, 2006.

1. Introduction

Data Centers (DCs) need to face the large traffic generated by emerging bandwidth-hungry internet services like cloud computing, social networking and video sharing [1]. These applications require intense interactions among the servers of a DC. Servers of a current DC are typically interconnected according to a fat tree topology network as depicted in Fig. 1(a). Each rack contains up to forty servers interconnected by Top-of-the-Rack (ToR) switches through 1 Gb/s links (10 Gb/s expected in future). A small fraction (the typical subscription rate is 1/5) of the ToR switches bandwidth is dedicated to communicate with servers positioned in different racks. Racks are then grouped in clusters and ToR switches are interconnected each other by means of cluster switches [2]. In this environment optical technology is used only for point-to-point links between the switches of the DC. ToR switches are interconnected with the higher level by means of 10/40 Gb/s links (higher bit rate is expected in future). The switching is performed by electronic packet switches at each level of the DC tree topology. This results in the need of power-hungry optical-electrical-optical (O-E-O) conversions at each hop of the transmission link. Moreover, packets have to face the high latency of few hundreds of μs for inter-cluster communication delays introduced by the multiple conversions and thus by the multiple store-and-forward processes [2]. In this scenario, flattening the inter-cluster DC network by employing photonic technologies for optically transparent switching and interconnecting operation would avoid the costly and power hungry high speed (40 Gb/s today and higher in future) O-E-O conversions and electronic buffers unavoidable with current electronic switches. This may result in higher DCs network performance in terms of end-to-end latency and power consumption. In our vision, illustrated in Fig. 1(b), the optical packet switch (OPS) interconnects N input /output nodes transparently in the optical domain in a flat configuration. Each of the nodes represents a cluster of servers. Each cluster contains an electronic input/output buffer and a WDM transmitter/receiver. These buffers are needed to store the packets and to allow retransmission in case of output contentions. The switch configuration time, the data rate operation and the port count scalability of the OPS architecture play crucial roles in realizing the flat topology with high connectivity, low packet loss, and high throughput while minimizing the electronic buffer size needed in the cluster switch. In the last few years several research projects have focused on this topic [3–6]. In [7] a qualitative categorization and comparison of optical interconnects for DCs is reported. However, an OPS that scales to a large number of ports, that provides low end-to-end latencies and that is easily controllable does not exist yet.

In [8] we presented a WDM OPS architecture with modular structure and highly distributed control and demonstrated 40 Gb/s operation with discrete components. In this paper we present a numerical and experimental study of the WDM OPS architecture with modular structure and highly distributed control that allows interconnecting high-speed cluster switches in a flat configuration employing an integrated optical cross-connect as basic building block of the modular architecture. Firstly, we study the scalability and the performance of such architecture in terms of packet loss and latency by means of OMNeT + + Network Simulation Framework [9]. Secondly, we prove the feasibility of the proposed OPS architecture in an experimental set-up. The paper is organized as follows. In section 2 we present the proposed WDM OPS architecture. In section 3, we numerically analyze the WDM OPS performance in terms of port-count scalability, packet loss, and latency under a

#177554 - $15.00 USD Received 5 Oct 2012; revised 22 Nov 2012; accepted 23 Nov 2012; published 4 Jan 2013

realistic DC traffic. In section 4, we experimentally demonstrate the feasibility of a 16x16 WDM OPS based on 1x4 WDM switches. Conclusions on the feasibility of larger port-count OPS are given in section 5.

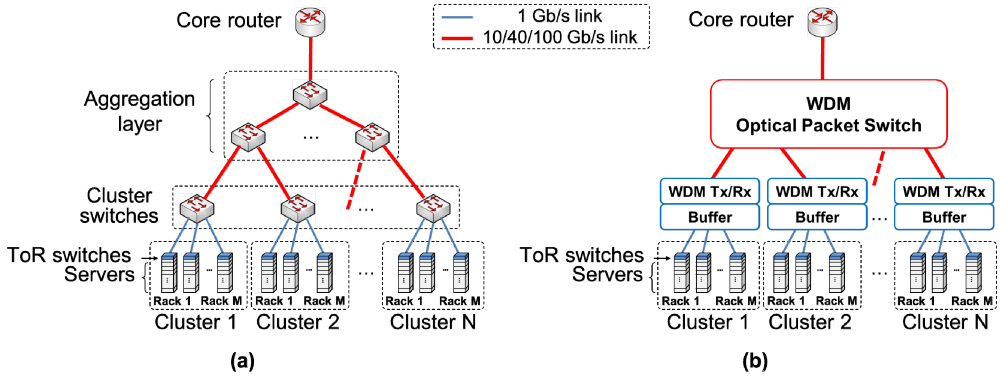

Fig. 1. Typical data center network topology (a) and a possible data center topology employing a large port-count optical packet switch at cluster level (b).

2. WDM modular OPS architecture with highly distributed control

The proposed strictly non-blocking WDM OPS architecture is shown in Fig. 2(a). The WDM OPS has N input fibers, N independent optical modules, and N output fibers. Each of the N input/output fibers carries M WDM wavelength channels, indicated by ch_1 to ch_M. Thus, the total number of logical input/output ports is NxM. Each of the N optical modules has one input and N different outputs, destined to M fixed wavelength converters (FWCs) at the outputs of the OPS architecture. The FWCs 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 output the same output fiber. As shown in Fig. 2(b), each optical module consists of a label extractor, a 1xN Broadcast Stage (BS) and the switch controller that drives N independent Wavelength Selective Stages (WSSs), one for each optical module output. The optical module operates as follows. The label extractor separates the packet labels from the packet payloads. The labels are O-E converted and processed by the switch controller.

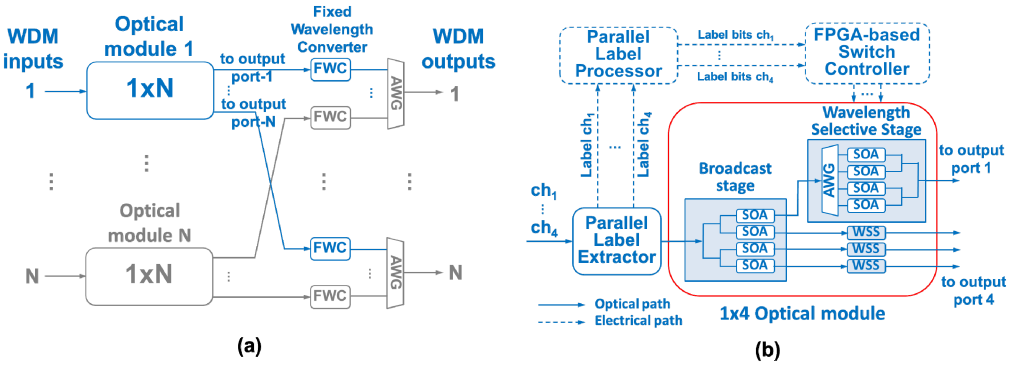

Fig. 2. Modular OPS architecture with highly distributed control (a) and 1x4 optical module layout and functionalities (b).

The employed parallel in-band labeling processing technique with few nanoseconds operation regardless the bit-count has been presented in [10]. The packet payloads remain in the optical domain and are fed into the 1xN BS. The switch controller reads the packet labels and drives the N WSSs. A WSS consists of a 1xM array waveguide grating (AWG), M

#177554 - $15.00 USD Received 5 Oct 2012; revised 22 Nov 2012; accepted 23 Nov 2012; published 4 Jan 2013

optical gates based on semiconductor optical amplifiers (SOAs) and an Mx1 combiner. It may happen that two or more packets coming on different wavelength channels from the same input fiber are destined to the same output port. It is the switch controller that solves these output contentions. The controller drives the SOA gates of the WSSs in order to connect only one of the WDM channel per output of the optical module. The packets that lose the contention are dropped at this stage and need to be retransmitted. On the contrary, the selected packet is fed into the appropriate FWC. Main advantage of the described architecture is its modular structure that allows for a simple and highly distributed control. Each of the optical modules is completely independent from the others and thus the configuration time of the entire switch is strictly dependent only on the configuration time of a single optical module. This makes the overall switch configuration time almost independent of the number of ports. As we have already shown in [11], in order to have low configuration time and thus low extra-latency added to the system the employment of a distributed control is necessary. Scaling an OPS based on rearrangeable non- blocking architectures, like Beneš or Banyan, to a large number of ports would require long configuration times due to the need of a centralized controller. Moreover, large port-count OPS can be built starting from low port-count modules. For example a 1024x1024 (N = 32, M = 32) OPS can be implemented employing 32 1x32 optical modules.

3. Numerical investigation

We evaluate the performance of an OPS based on the architecture described in section 2 in a DC environment. OMNeT + + Simulation Framework software is employed to investigate the scalability of the proposed architecture (up to 4096 with NxM = 64x64 = 4096) and to determine its performance under DC realistic traffic conditions. Input traffic sources are programmed to generate the load of a large number of servers. Inter-arrival time distribution and packet length distribution are selected in order to match the few available publications on the DC traffic characteristics [12, 13]. A more detailed description of the traffic generator used in this work can be found in [14]. Due to the modular structure of the proposed OPS architecture and due to the independency of each optical module, simulation of a single optical module provides performance information of the overall OPS. We set the distance between the cluster nodes and the OPS to 50 meters. Each of the M wavelength channels at 40 Gb/s has an independent traffic generator. The packet length ranges between 64 B and 1500 B. The input buffers of each M channels have a capacity of 20 kB. The software emulates the optical module operation as described in section 2. During the simulations the system operates as follows. The traffic generator modules create packets that are stored in the electronic input buffers and transmitted to the OPS. At the OPS, the optical module controller processes the packets label, solves the contention, and drives accordingly the WSS modules in order to forward the packets to the appropriate output port. Packets dropped at this stage need to be retransmitted. Positive acknowledgment message (ACK) in case of successful reception is generated and sent back to the right input. Only when the ACK is received at the cluster buffer the packet is removed from the buffer queue. This means that the minimum latency for a packet in the system (the time in which the packet is stored in the input buffer) results to be equal to the RTT of the system, which is 560 ns including the processing and the configuration of the OPS (2x30 ns). If no ACK is received at the input after the RTT, the packet is re-send. If a new packet arrives at the input buffer while its queue is full, the new packet is lost. At the OPS output the statistics are collected in order to compute the packet loss, the throughput and the average end-to-end latency of the system. Figures 3(a)-3(c) show the packet loss, the throughput and the latency obtained for a system with N = M = 2, 4, 8, 16, 32, 64 that correspond to systems with 4 to 4096 logical input/output ports. The figures clearly indicate that the performances worsen increasing the port-count. This behavior is due to the fact that the contention probability increases increasing the port-count as explained in [11]. Simulation results indicate a packet loss lower than 10^{-6} with an average latency around 900 ns, and 100% throughput is assured for input traffic up to 0.3 regardless the switch port-count. Considering that 99% of the time the traffic load is below 0.3 [12], those results

#177554 - $15.00 USD Received 5 Oct 2012; revised 22 Nov 2012; accepted 23 Nov 2012; published 4 Jan 2013

confirm that the proposed OPS architecture could handle DC realistic traffic. More complicated contention resolution approaches could improve the performance in terms of packet loss, while increasing the controller complexity and thus compromising the average system latency and throughput [14].

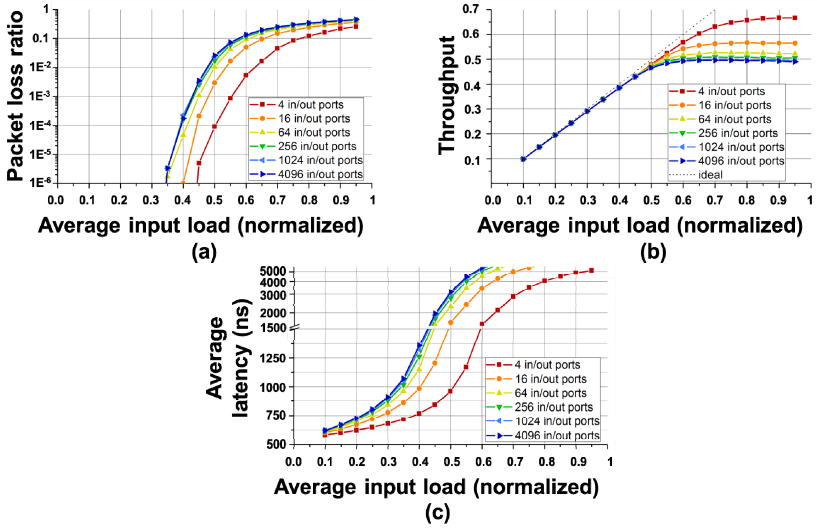

Fig. 3. Packet loss ratio (a), throughput (b) and latency (c) as function of the normalized average input load for systems with 4, 16, 64, 256, 1024 and 4096 input/output ports.

4. Experimental investigation

We demonstrate the feasibility of a 16x16 OPS based on a monolithic integrated 1x4 optical cross-connect module using the experimental set-up shown in Fig. 4(a). The modularity of the WDM OPS architecture allows evaluating its performance studying the operations of a single optical module. Figure 4(b) shows the 1x4 module [15]. The 4.2 mm x 3.6 mm chip integrates a BS, and four parallel WSSs.

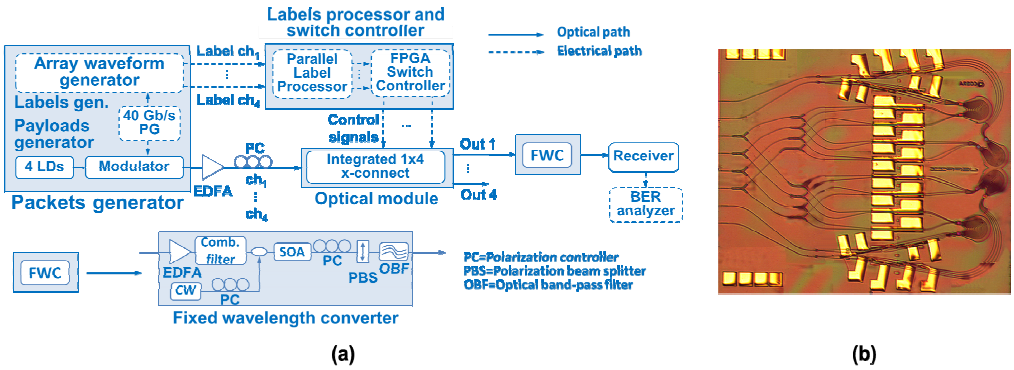

Fig. 4. Experimental set-up (a), photograph of the integrated optical cross-connect module (b).

The BS is realized with two cascaded 1x2 multimode interference splitters (MMI). 750 µm SOA at each output of the BS is used to compensate for losses. Each of the four WSSs consists of a cyclic AWG (channel spacing 400 GHz), four 140 µm SOAs as optical gates, and two cascaded 2x1 MMI couplers. The 750 µm SOAs are DC biased while the SOAs optical gates are directly driven by digital pins of a FPGA-based switch controller. Four 40 Gb/s WDM input packets (ch_1 = 1548.1 nm, ch_2 = 1551.4 nm, ch_3 = 1554.5 nm, and ch_4 = 1557.7 nm) have 290 ns payloads and labels separated by 40 ns guard time. The in-band label

#177554 - $15.00 USD Received 5 Oct 2012; revised 22 Nov 2012; accepted 23 Nov 2012; published 4 Jan 2013

consists of two bits encoded using the RF tones labeling technique presented in [10]. Time-slotted operation is considered. The WDM packets with an input power of 6 dBm/ch are synchronously fed into the 1x4 optical module by using lensed fiber array. The labels are generated and transmitted in the electrical domain by an arbitrary waveform generator. The two-bit label determines to which of the four output ports the packet is destined. The labels are detected and processed by the switch controller that consists of a label parallel processing circuit and a Virtex 4 FPGA working at 100 MHz clock-speed. Further details on the label processor can be found in [10]. The switch controller performs the label processing and the contention resolution functionality, enabling only one gate of each WSS per time slot, within 25 ns. The digital pins of the FPGA provide 7.5 mA that are sufficient to enable the SOA gates. The switch controller is programmed in order to give higher priority to ch_1, then ch_2 and so on. The selected packet at the outputs of the optical module are amplified to 3 dBm and converted to a fixed wavelength (1549 nm) by the FWC. The 40 Gb/s FWC is based on non-linear polarization rotation in an SOA (CIP-XN-1550) that has a recovery time of 10 ps when biased at 500 mA. At the FWC output, a 40 Gb/s receiver and a BER analyzer are used to analyze the quality of the packets. Figure 5(a) shows the label bits pattern detected by the parallel label processing circuit. Label "1 0" indicates the packets directed to port three of the optical module. Figure 5(b) shows the control signals for the WSS generated by the FPGA-based controller after decoding the label bits and after solving all the possible contentions. In the fifth time slot studied it is visible that the contention resolution scheduler performs as expected, in fact only the SOA gate associated with ch_1 is biased. The switched packets at the output port 3 are shown in Fig. 5(c).

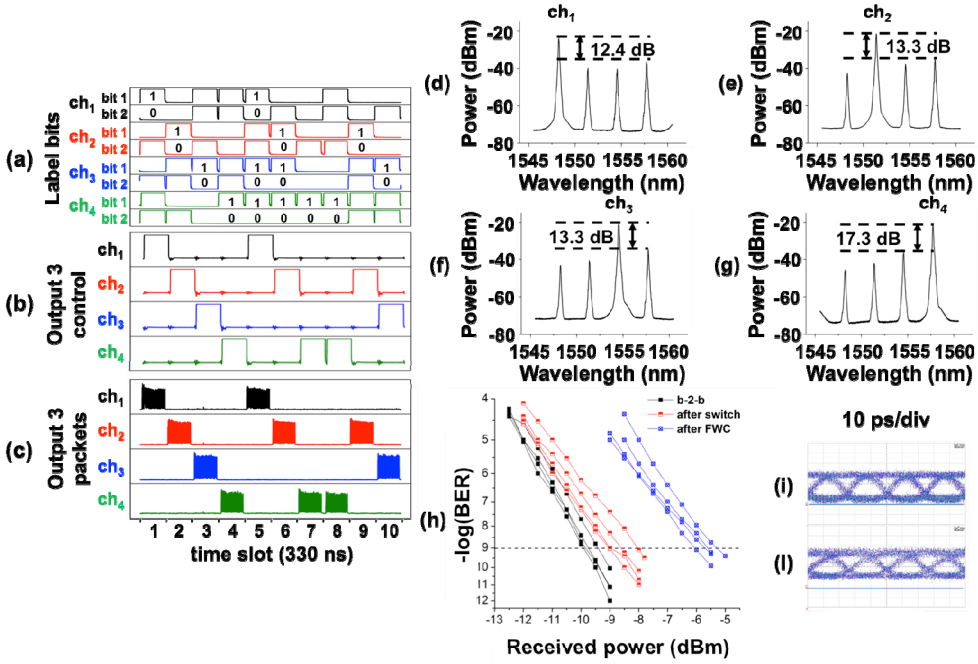

Fig. 5. Label bits of the four WDM channels (a), control signals driving the optical module (b), switched packets (c); optical spectra after static switching (d-g); BER curves (h); eye diagram after switching (i) and after wavelength conversion (l).

The optical spectra after static switching of the WDM channels shown in Figs. 5(d)-5(g) indicate a contrast ratio of at least 12.4 up to 17.3 dB. The relatively low extinction ratio is mainly due to wavelength misalignments in the AWGs and the limited length of the WSS SOAs. Longer WSS SOAs would improve the extinction ratio providing larger absorption and higher gain while requiring higher driving current. Figure 5(h) shows the BER curves of

#177554 - $15.00 USD Received 5 Oct 2012; revised 22 Nov 2012; accepted 23 Nov 2012; published 4 Jan 2013

the back-to-back, after switching and after wavelength conversion. BER curves are taken switching continuously a single wavelength channel out of the 4 WDM at the time. Error free operation with 1 dB and 4 dB penalty after switching and after wavelength conversion was measured, respectively. The extra penalty is caused by the low extinction ratio (7 dB) as shown in Figs. 5(i), and 5(l) due to the insufficient power at the output of the chip. The total chip insertion losses are around 38 dB including the fiber coupling losses. The input power of each WDM channels was 6 dBm, while at the output of the chip the power was around −21 dBm. The BS SOAs (driven by 80 mA current) provide a gain around 9 dB/ch, while the WSS SOAs gain is around 2 dB. Performance improvement may be anticipated if the excess losses in the optical module are substantially reduced. The total 1x4 WDM module consumes 10.6 W that includes: 0.6 W for the 4 long SOAs of the chip, 1.3 W for the controller, 0.3 W for the chip temperature controller, 2.4 W for the EDFA amplifiers (required to compensate the coupling loss to and from the chip) and 6 W for the 4 FWCs. Considering the total capacity of 4x40 Gb/s, the total energy consumption is 66 pJ/b.

5. Conclusion

We presented a numerical and experimental investigation on the feasibility of a large port-count modular WDM OPS architecture with highly distributed control and its performance in terms of packet loss, throughput and latency. Numerical results confirm the capability of the WDM OPS architecture of handling DC traffic loads providing low packet loss, high throughput and sub-microsecond latency. We experimentally investigate the feasibility of the WDM OPS based on integrated WDM optical cross-connect modules. Error free operation at 40 Gb/s of 1x4 WDM cross-connect module with 4 dB penalty and processing time around 25 ns was measured, indicating that a 16x16 WDM OPS with similar performance can be directly realized by employing four 1x4 WDM integrated modules.

The parallel operation of the label processing and switch control makes the processing time port-count independent. This suggests that it may be feasible to realize a large scale OPS with very low latency and high interconnectivity level. However, scaling the WDM OPS to over thousand ports will require facing other technical challenges. For example, realizing a 1024x1024 (32 input/output fibers carrying 32 wavelength channels) demands 32 blocks of 1x32 WDM optical cross-connect module. This translates in 1x32 BS with 15 dB splitting losses, 32 channel 100 GHz spaced AWGs with flat low loss operation, 1024 SOAs integrated on the chip, and FWC with broadband operation. Considering the 15 dB attenuation after the 1x32 BS and the quasi-constant envelop of the optical power of 32 time-decorrelated channels, the SOAs employed to compensate the splitting losses should not introduce non-linear effects. However, OSNR degradation has to be considered. WSSs with reduced number of SOAs and low cross-talk AWGs, and broadband FWC operation have already been demonstrated [16–18]. Photonic integration of large chip size including all the required devices on a single chip remains technological demanding either for the fabrication and the electrical wiring. Moreover, chip-coupling losses have also to be considered. Improvements in photonic integrated technologies may allow the realization of such large-scale photonic chip.

Acknowledgment

This work is supported by the Netherland Organization of Scientific research (NWO) and the Netherland Technology Foundation (STW) through the VI and NRC Photonic Program. The authors wish to thank the JePPIX for fabricating the 4x4 integrated optical cross-connect.

#177554 - $15.00 USD Received 5 Oct 2012; revised 22 Nov 2012; accepted 23 Nov 2012; published 4 Jan 2013