Building Petabit/s Data Center Network with Sub-microseconds Latency by Using Fast Optical Switches

Wang Miao, Fulong Yan, Harm Dorren, and Nicola Calabretta

COBRA Research Institute, Eindhoven University of Technology, Eindhoven, the Netherlands

We propose a novel flat optical data center network based on fast optical switching systems. Petabit/s capacity and connectivity of 1000 ToRs are enabled by using low-radix optical switches. Demonstration with 4×4 switch prototypes shows dynamic multi-path switching operation at 40Gb/s with <10^{-5} packet loss and 640ns latency with 0.4 load.

Introduction

Data centers networks (DCNs) are experiencing the traffic explosion to support various data intensive applications and virtualization technologies [1]. The demand for high processing capability continuously drives the upgrade of the server blades. Assuming blades with 16 multi-processor cards and each card with 8 quad-core processors at 2.5GHz, a 40-blade rack will provide a processing power of 50THz, in which case a 50Tb/s I/O is required for top-of-rack (ToR) switches to achieve a balanced system performance [2]. This will open new scaling challenges for the implementation of low latency Petabit/s DCN where 100's of ToRs need to be interconnected.

Limited by the package size of ball grid array (BGA) and the power density, the maximum I/O bandwidth of the electronic switch IC is predicted to be around 5Tb/s [3]. To interconnect the ToRs each with tens of Terabit/s interfaces, even multi-layer structure deteriorating the performance of latency and throughput is hardly to meet to the desired scale. In addition, the high speed O/E/O interfaces also introduce undesirable extra costs and energy consumption.

Switching packets in the optical domain has the advantage of transparency to the data rate [4], which therefore overcomes the I/O limitation of the electrical switches. However, optical switches with sub-microseconds reconfiguration time allowing low latency operation have been demonstrated only with limited port count [4]. To ensure the Petabit/s capacity as well as the large connectivity, the optical switches based DCN needs to be intelligently arranged.

In this work, we propose a flat Petabit/s optical DCN architecture with sub-microseconds latency based on distributed fast optical switches with optical flow control

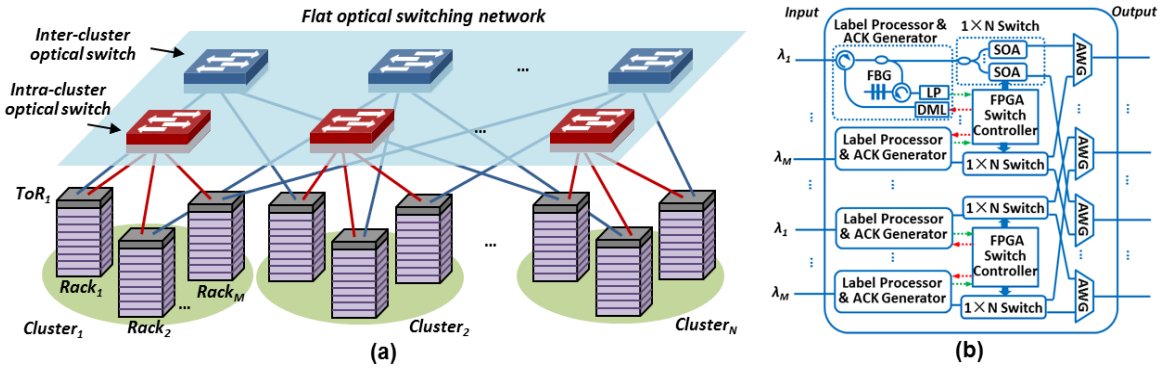

Fig. 1: (a) flat DCN based on two parallel intra- and inter-cluster networks (b) optical switch node

to overcome the lack of optical buffers. Two parallel intra-/inter-cluster networks are introduced which could support multiple parallel and reconfigurable paths for load balancing and fault tolerant protection. Experimental evaluation employing a 4×4 optical switch prototype shows successful dynamic path reconfiguration and error free operation with maximum 1.5dB penalty for the longest paths. Packet loss less than 10^{-5} and an average end-to-end latency less than 640ns have been observed for a load=0.4 with a buffer size of 16 packets.

System operation

The proposed flat DCN architecture based on two parallel intra- and inter-cluster networks is shown in Fig. 1(a). An intra-cluster M×M optical switch interconnects the M ToRs. The inter-cluster network consists of M independent N×N optical switches that interconnect the ToRs of different clusters, namely the m-th optical switch interconnects the ToR_m of each cluster, with m=1,2,..,M. The parallel instead of hierarchical operation makes the DCN bounded in a flat way, and the connectivity scales as the square of the optical switch radix. More than 1000 ToRs could be interconnected by using 32×32 optical switches. The transparency to data rate and data format of the optical switch could allow for scaling up the data rate per link without changing the infrastructure, or adding more switching ports. Sub-microseconds reconfiguration time could properly handle the long flows as well as the short ones by exploiting statistical multiplexing of the optical switches. Moreover, no optical buffers are required and due to the short distance between the ToRs and optical switches, small electrical buffers at the ToRs allow an efficient implementation of the optical flow control to guarantee high performance of the network. And as can be seen from Fig. 1(a), multiple paths are available which improves the fault tolerance and enables the load balancing operation.

The intra-cluster ToRs interconnection operates as follow. The ToRs generate optical packets consisting of the payload and the label carrying the destination information. The packets will be transmitted to the intra-(or inter-)cluster optical switch according to the assigned path. As the optical switch has no optical buffer, a copy of the transmitted data is electrically stored. Fast optical flow control between the optical switch and the ToR is implemented for packet re-transmission in case of contention at the optical switch [5]. The modular optical switch architecture shown in Fig. 1(b) is capable of switching multiple WDM channels simultaneously [5]. First the optical label of each channel is detected and processed by the label processor (LP) and the switch controller, while the payload is fed into a 1×N SOA-based broadcast and select stage. The switch controller checks the label bits and according to the look-up table configures the 1×N switch to forward the packets to the destination. In case of contention, the low-priority packet will be blocked and a fast optical flow control signal (negative ACK) using a low-speed directly-modulated laser (DML) is sent back to the ToR to request the retransmission. If no contention occurs, a positive ACK informs the ToR to remove the packet from the buffer. The fast response of the SOA in combination with the parallel processing of the label bits allows few tens of nanoseconds switch reconfiguration time.

As depicted in Fig. 1(a), the interconnection between inter-cluster ToRs can be directly through the inter-cluster optical switch (direct path) or through a combination of intra and inter-cluster switches via an intermediate ToR (indirect and longer path). In case of direct path, the operation is the same as the intra-cluster interconnection. For the indirect path, the intermediate ToR operates as an optical bridge which forwards the optical

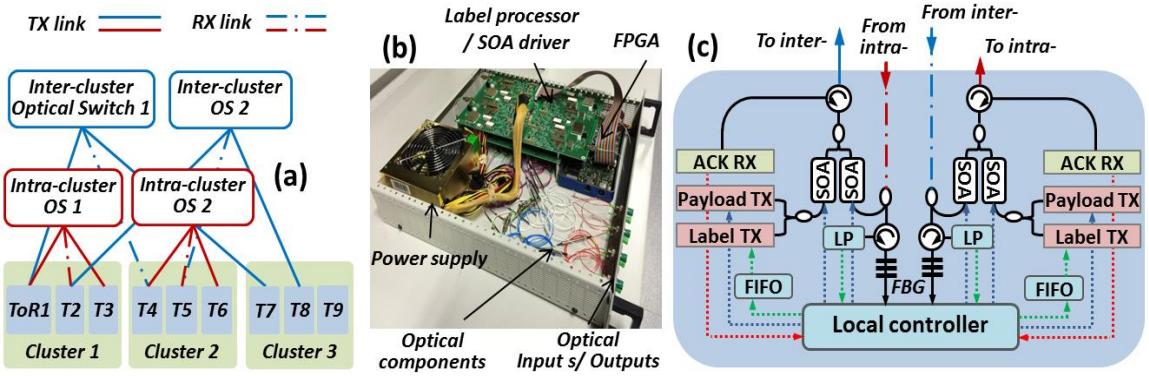

Fig. 2: (a) experiment set-up (b) 4x4 optical switch prototype (c) ToR Schematic

packets towards the next optical switch. As an example, packets transmitted by ToR_1 of $Cluster_1$ and destined to ToR_2 of $Cluster_2$ may go through Intra-cluster $Switch_1$ to the intermediate ToR_2 of $Cluster_1$, then to the Inter-cluster $Switch_2$ and directly headed to the ToR_2 of $Cluster_2$. The intermediate ToR_2 optically forwards the packets to the Inter-cluster $Switch_2$ and stores the packets till the positive ACK sent by the Inter-cluster $Switch_2$ is received.

Experimental results

The experimental set-up to validate the novel DCN consists of 3 clusters, each cluster formed by 3 ToRs, and interconnected by inter- and intra-cluster switches as shown in Fig. 2(a). For inter and intra-cluster switches we employed 4x4 optical switch prototypes shown in Fig. 2(b). The ToRs generate 40Gb/s NRZ-OOK packets with 540ns duration and 60ns guard time at λ_1=1552.55nm, λ_2=1554.17nm and λ_3=1555.78nm for each optical switch input port. As illustrated in Fig. 2(c), an FPGA acting as a local controller stores the look-up table with the packet destination. A 4-bit (2 for inter- and 2 for intra-) in-band RF tone label according to the port destination will be generated for each packet [5]. The label information will be stored in a FIFO queue with a size of 16 packets and removed from the queue in response to a positive ACK. Otherwise the label and payload are retransmitted. When the packet is received at the ToR, the optical label will be first processed to check if it belongs to the current ToR, otherwise it will be optically bridged to the next switch node and a copy will be stored. The locally generated traffic and the packets to be bridged are scheduled by the 1×2 switch and the bridged packets have been given the higher priority when no retransmission is under process. The blocked traffic will wait for the next available time slot.

Fig. 3(a) shows the recorded packets and ACK signals to validate the intra and inter cluster dynamic switching operation. As a selected case, intra-cluster (direct path) ToR_4 and ToR_6 traffic, and inter-cluster (indirect path via Optical $Switch_1$ and ToR_4) ToR_1 and ToR_7 traffic with destination ToR_5 are investigated. The traffic generated by the ToRs is reported in Fig. 3(a). Each packet has been labeled with the destination ToR (1 to 9). At the Inter-cluster $Switch_1$, the packets from the ToR_1 and the ToR_7 (higher priority) with destination ToR_5 are transmitted to ToR_4. In case of contention, the traffic from ToR_1 will be blocked, a negative ACK is sent to ToR_1 and the blocked packet is retransmitted in the next time slot. At the ToR_4, packets with destination ToR_5 and ToR_6 will be optically forwarded to the Intra-cluster $Switch_2$ and stored in case retransmission is needed. At the Intra-cluster $Switch_2$, packets coming from ToR_4 and ToR_6 (higher

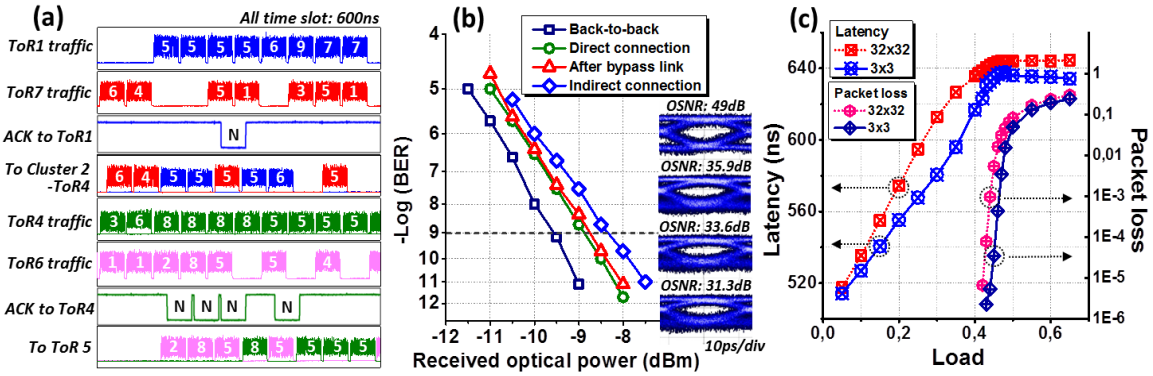

Fig. 3: (a) traces for dynamic switching (b) BER curves and eye diagrams (c) latency and packet loss

priority) and destined ToR_5 might have contention. Optical flow control provides, in case of contention, the ACK signals to ToR_4 as shown in Fig. 3 (a), where the packets received by the ToR_5 are also reported.

The BER performance and the eye diagrams for the 40 Gb/s payload are shown in Fig. 3(b). Error free operation with less than 1.5dB power penalty and 31.3dB OSNR for the indirect connection (longest path) has been obtained. Fig. 3(c) reports the measured experiment results (3×3) in terms of system latency and packet loss under uniform traffic distribution with load varying from 0.1 to 0.7. Considering a Petabit/s DCN, a case deploying 32×32 optical switches with 1024 ToRs being connected is also studied in simulation. Results indicate a packet loss less than 10^{-5} and an average end-to-end latency (including 50m transmission link) less than 640ns for a load=0.4 with a buffer size of 16 packets at each ToR.

Conclusion

We propose a novel flat DCN based on bufferless fast optical switches. Demonstration with 4×4 prototypes shows dynamic switching and error free operation with <1.5dB penalty. Both experiment results and investigation study on the 32×32 system reports less than 10^{-5} packet loss and 640ns end-to-end latency with 0.4 load and 16-packet size buffer. By employing such optical switch based parallel intra- and inter-cluster networks, 100's of ToRs each with tens of Terabit/s interface can be interconnected providing a Petabit/s capacity.

Acknowledgements

The authors would like to thank the FP7 LIGHTNESS project (n° 318606) for supporting this work.

References

[1] G. Bell et al., "Petascale computational systems," IEEE Computer, Vol. 39, no. 1, p. 110 (2006).

[2] K. Lim et al., "System-level implications of disaggregated memory", IEEE HPCA, pp. 1–12 (2012).

[3] A. Ghiasi, "Is there a need for on-Chip photonic integration for large data warehouse switches," Proc. 9th IEEE Int. Conf. Group IV Photon., P. 27 (2012).

[4] C. Kachris et al., Optical Interconnects for Future Data Center Networks, Springer (2013).

[5] W. Miao et al., "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," Opt. Express, vol. 22, pp. 2465 (2014).