A Novel Flat DCN Architecture Adopting Low Radix Optical Packet Switches

Fulong Yan[1,2], Wang Miao[1], Harm Dorren[1], Nicola Calabretta[1]

[1] COBRA Research Institute, Eindhoven University of Technology, Eindhoven, the Netherlands,

[2] Beihang University, Beijing, China. E-mail: f.yan@tue.nl

We investigate the performance of a novel flat data center network (DCN) architecture based on distributed buffer-less optical packet switches and fast flow control. Scalability of the novel DCN is assessed by numerical simulation with OMNeT++ discrete event simulator for medium and large size data centers. The simulation results reports an end-to-end latency of less than 8 µs (including retransmission), a packet loss $<10^{-8}$ and >9Tb/s average network throughput for load of 0.4 and DCN up to 100,000 servers, which therefore provided a promising solution apart from the prevalent electrical data center architectures.

INTRODUCTION

Driven by the cloud computing paradigm and data-rich applications, data centers experience a steady annual increase of over 30% in the amount of traffic [1]. Emerging data center applications and workloads are creating new communication patterns with more than 75% of the total data center traffic processed within data centers (server to server and rack to rack) [2]. This huge increase of intra data center traffic requires architectural and technological changes to the underlying interconnect in order to enable the scalable growth both in communicating endpoints and traffic volume, while decreasing the costs and the energy consumption.

Building data center with scaling out number of servers (>10000) and scaling up data rate (>10Gb/s) demand electrical switches with high radix to prevent hierarchical multi-layers architectures with bandwidth bottleneck, high end-to-end latency as well as poor cost-efficiency. However, the limited I/O bandwidth of the switch ASIC caused by the limited BGA density will prevent the implementation of high radix switch [3]. Multi-stage switching architecture will allow higher radix at the expense of large latency and cost [4]. On the contrary, optical switch technologies can transparently switch high data rate/format signal, reducing the required number of switch ports and avoiding costly O/E/O required in hierarchical multiple stages architecture. However, optical switches

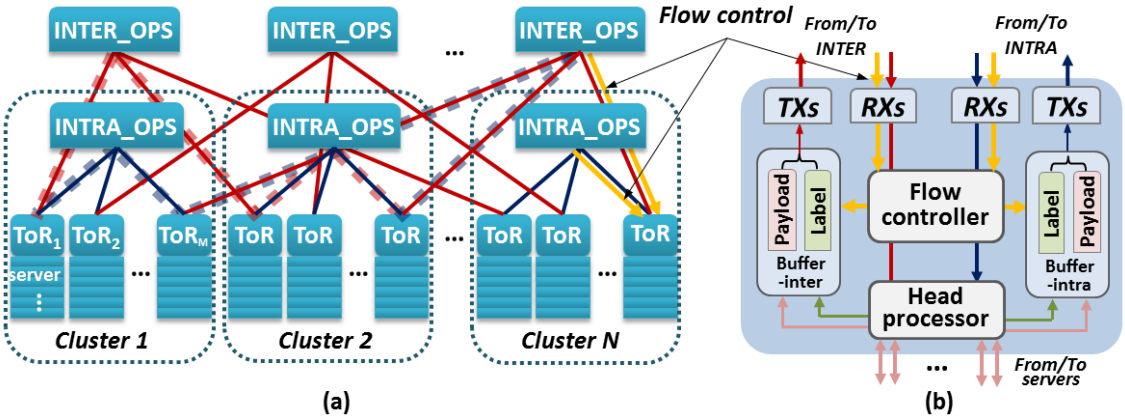

Fig. 1: Novel DCN architecture and ToR schematic diagram with fast flow control

with sub-microseconds reconfiguration time have limited port count [5]. Large port-count optical switches based on multi-stage architecture are prevented by the lack of optical buffers, making fast optical switching technology not practical for DCNs.
In this work, we propose a novel flat DCN architecture based on distributed buffer-less optical switches with fast flow control. Exploiting the fast switching and optical flow control, no optical buffer is required. Numerical simulation based on OMNeT++ is employed to assess the scalability and performance of the novel flat DCN in terms of latency, packet loss, and average throughput under different inter/intra-cluster traffic distribution. Results show that the DCN could interconnect up to 100000 servers with a latency <8 µs, $<10^{-8}$ packets loss and >9Tb/s average network throughput under a uniform load of 0.4 and ToR with 40KB buffer per port.

SYSTEM OPERATION

The proposed DCN architecture is shown in Fig. 1(a) and consists of N clusters, each with M electronic ToRs. The functional blocks of ToR are illustrated in Fig. 1(b). Each ToR switch interconnects 40 servers and has two optical links connected to the intra-cluster optical switch (INTRA_OPS) and inter-cluster optical switch (INTER_OPS). The INTRA_OPS interconnects the M ToRs of the cluster, while the INTER_OPS interconnects N ToRs of different cluster. More precisely the *i*-th INTER_OPS interconnects the *i*-th ToRs of each cluster, with $i = 1,\ldots,N$. This architecture allows interconnecting all the ToRs with at most two OPS hops.

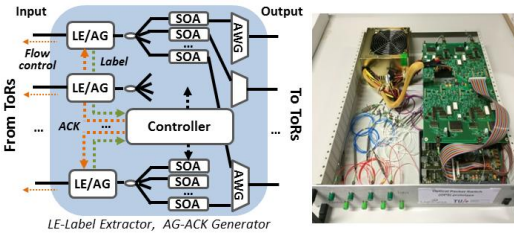

Fig. 2: Prototyped OPS

TABLE I: TRAFFIC PATTERNS AND DISTRIBUTIONS

Traffic distributions \ Traffic patterns	Large inter-cluster traffic (case A)	Light inter-cluster traffic (case B)
Inter-cluster	40%	25%
Intra-cluster	20%	25%
Intra-ToR	40%	50%

When the packets arrive at the ToR, the destination will be checked by the header processor. If the packet destines to a server within the same cluster or in a different cluster, the packet is forwarded to the INTRA_OPS or INTER_OPS output optical link, respectively. An optical label determining the ToR destination is attached to the packet. Moreover, a copy of the packet is buffered in case of retransmission due to contentions in the OPS. The architecture and prototype of the INTRA_OPS and INTER_OPS based on optical switch technology is shown in Fig. 2. More details on the operation and port count scalability of the OPS are reported in [6]. At the OPS, the label is processed to determine the packet destination. The OPS can switch packets within few tens of nanoseconds. Due to the lack of optical buffer, a fast optical flow control is implemented. In case of contention, a negative ACK signal is transmitted to the ToR to ask for packet retransmission; otherwise the positive ACK indicates the successful forwarding and thus release the packet from the electronic buffer.

For intra-cluster connection, the packet undergoes a single INTRA_OPS hop to the ToR destination and then the server within the rack. In case of inter-cluster connection, it might be that the packets has to undergo two OPS hops (INTER_OPS → ToR → INTRA_OPS → ToR destination) before reaching the final destination (see example in

Fig. 1(a)). In this case the packet has to be detected and processed by the intermediate ToR, adding extra latency.

There are multiple advantages of this architecture. First, the number of interconnected ToRs (and servers) scales as N × M with N and M the switch radix of the INTER_OPS and INTRA_OPS, respectively. As an example, by using 64×64 INTRA_OPS and INTER_OPS, up to 4096 ToRs can be interconnected. If each ToR groups 40 servers, up to 160000 servers can be interconnected. Second, 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. Third, the sub-microseconds reconfiguration time of the optical switches allows operating on large as well as on small flows exploiting statistical multiplexing. Fourth, the optical switches are fully distributed allowing the use of electrical buffers at the ToRs in combination with fast optical flow control [6] to efficiently solve packet contention. Moreover, the proposed DCN provides multiple connection paths that can be exploited to improve the resilience of the DCN.

SIMULATION SETUP

OMNeT++ has been employed as the simulation tool to fully investigate the performances of the proposed DCN under packet transmission. To investigate the scalability, DCNs consisting of 144 ToRs interconnected by 12 × 12 INTRA_OPS and INTER_OPS and 2500 ToRs interconnected by 50×50 INTRA_OPS and INTER_OPS have been considered. Each ToR connects 40 servers with 10Gb/s link, resulting in DCNs of 5760 (medium size) and 100000 (large size) servers. The packet size is set to 1000 Bytes at the ToR output that each slot is 800ns. The total data rate of the optical links to the INTRA_OPS and INTER_OPS is 200Gb/s (5 × 40Gb/s WDM channels). The round trip time (RTT) is 560 ns, including the delay to process the labels, control the switches (60ns) and the optical link (2 × 50 = 100m) latency (500 ns).

The operation of the system is synchronous and the simulations run for 10^8 time slots. At each time slot, the server generates a packet according to a binomial distribution with a fixed load, thus there is a certain probability to have a new packet at each server. All servers are programmed to operate independently with the same load. Traffic is classified into 3 categories (inter-cluster, intra- cluster and intra-ToR). In each category, packets destinations are chosen randomly between all possible destinations according to a uniform distribution. Considering the facts that most of the traffic is exchanged within the ToR and cluster, two traffic distributions with different ratios of 3 categories are studied. As shown in Table 1, case A is the large inter-cluster traffic while case B is light inter-cluster traffic. Different buffer size for 200Gb/s link ports are considered (40KB and 60KB).

RESULTS AND DISCUSSION

In this section, the simulations results on the performances of the DCN are presented and discussed.

Fig. 3 shows the end-to-end latency as a function of the load. Results show 8us end-to-end average latency has been obtained at load of 0.4 for both medium and large size DC and different buffer size. For larger load, the effect of the buffer on the system performance is visible. For load above 0.4, latency increases rapidly as it may take several slots for packet being successfully transmitted, before to saturate at high load. When the load is close to 1, the average latency of medium size DC under case B

increases rapidly and exceeds case A for the reason that the packet loss of case A is higher than case B, thus there are more packets going through the inter- cluster in case B, which contributes to the larger latency.

Fig. 4 shows the packet loss of the system. A packet loss less than 10^{-5} for load of 0.5 has been achieved for medium size DC. In agreement with the curve of the latency, the packet loss increases with the load. For load exceeding 0.5 (case A, Large size DC) and 0.6 (case A, Medium size DC), the packets loss is unavoidable as the links and the buffers are fully occupied. Larger buffers do not help and provide extra latency.

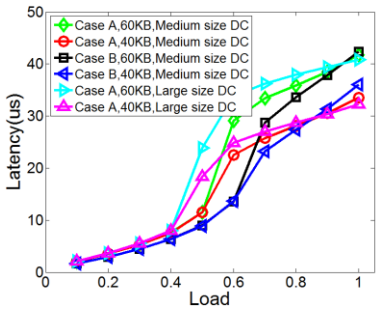

Fig. 3: Average latency

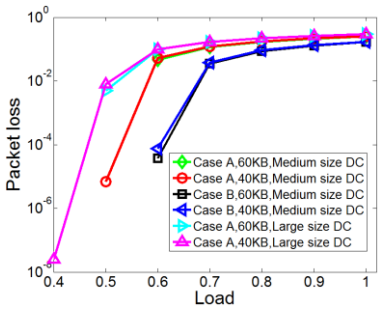

Fig. 4: Packet loss

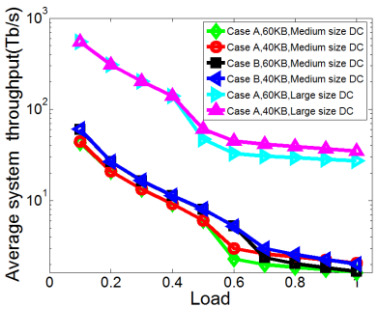

Fig. 5: Average network throughput

Fig. 5 reports the average network throughput calculated by adding all the received packet size (in bits) and divided by the aggregate packets latency. As shown in Fig. 5, the average throughput drops with the load in line with the results shown in Fig. 3 and Fig. 4.

CONCLUSIONS

We propose a novel flat DCN architecture based on distributed buffer-less optical switches and numerically assess the scalability and system performance with OMNeT++. Numerical results show that the prosed DCN architecture allows for 8us end-to-end average latency and less than 10^{-8} packet loss for traffic load of 0.4 and DCN up to 100000 servers. Those results indicate that a scalable DCN can be built by using feasible optical switches with moderate switch radix.

Acknowledgement

The authors would like to thank the FP7 LIGHTNESS project (n°318606), NSFC 61271196 and China Scholarship Council for supporting this work.

References

[1] Cisco global cloud index: Forcast and Methodology, CISCO, (2013).
[2] T. Benson et al., “Network Traffic Characteristics of Data centers in the Wild,” Proc. ACM, New York (2010)
[3] N. Binkertet al., "The role of optics in future high radix switch design." Proc. IEEE ISCA, California (2011).
[4] Cloud Scale Networking Part I: Architecture Considerations, Broadcom Corporation (2014).
[5] C. Kachris et al., Optical Interconnects for Future Data Center Networks, Springer (2013).
[6] 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).