Hardware Programmable Network Function Service Chain on Optical Rack-Scale Data Centers

Qianqiao Chen[1], Vaibhawa Mishra[1], Nick Parsons[2], Georgios Zervas[1]
[1]University of Bristol, Bristol, UK, [2] Huber-Suhner Polatis, Cambridge, UK
qianqiao.chen@bristol.ac.uk

Abstract: A datacenter network that supports programmable optical and multi-layer service chaining by adopting miniaturized reconfigurable optical backplanes and FPGAs is demonstrated. The end-to-end testbed delivers hitless on-chip service chain switch-over, 9.8G throughput and sub-microsecond latency.

OCIS codes: (060.4253) Networks, circuit-switched; (060.4259) Networks, packet-switched; (200.4650) Optical interconnects.

1. Introduction

Datacenters have been established to support cloud, fog and mist computing services. Research shows that over 70% of data exchange traffic remains within the datacenter because of intra traffic between servers, storage and databases [1]. Current Data Centre Networks (DCNs) are based on 10G/40G Ethernet over power-hungry electronic switches [2].To overcome the requirement of traffic growth, hybrid packet/optical DCNs have been proposed [3].

However, new datacenter architectures require more flexible and programmable network to address the challenges of processing-hungry and diverse set of applications. For example, rack scale architectures need flexible interconnects between disaggregated computing, memory and storage resources so that they can be allocated on demand according to the requirements of applications [4]. Software and processor-based flexible DCN virtual switching solutions often require multiple cores to achieve high bandwidth, however this is at the expense of power consumption as well as high and unstable network latency (100 μsec – 10s msec) [5]. FPGA-based hardware solutions can introduce new levels of flexibility and efficiency to the DCN by introducing high performance runtime programmability for network devices [6].

An agile programmable DCN design is proposed in this paper, which enables high performance optically interconnected Multi-Layer network function service chaining. Servers equipped with open-hardware FPGA platforms are connected to an ultra-low loss (~1 dB), low power consumption (<100mW/port) optically reconfigurable backplane. The architecture and experimental demonstration allows for deep programmable virtual network function service chains. The open-hardware is designed and implemented using a network on chip (NoC) to support hitless (no loss of data) switch-over between Ethernet switching and plug-in functions (e.g. IP parsing, TCP parsing) on each of the 2x2 10Gbps programmable interface and switch nodes. The use of such systems bring advanced physical and virtual switching functions to the server blade where traditionally only static network interface cards (NICs) have been used to interface CPUs with the network. The optical interconnect allows for multi-hop service chaining where Multi-Layer functions (optical circuit switching, Ethernet over optical, IP and/or TCP parsing over optical) can be deployed between servers. The result demonstrates a high performance multi-server system each with a programmable 4-port reconfigurable Multi-Layer switch and interface card delivering 9.8Gbps throughput per 10 Gbps port and ultra-low end-to-end latency (500 nsec to 1000 nsec).

2. FPGA-based Programmable Intra Datacenter Network

As shown in Fig.1 (a), the proposed DCN architecture deploys FPGAs on each server. The FPGA modules can be independent chipsets or they can be part of a system-on-chip integrating CPUs and programmable logic. The FPGA can be configured on demand to deliver diverse network functions such as Ethernet switch, IP router, traffic generator/monitor, filter etc. An optical chip-to-chip network is set up to interconnect electronic on-chip network functions (on-chip service chain) and servers.

High performance FPGAs are used to deliver diverse high throughput and low latency network functions in the proposed DCN architecture. However, re-configuration of an FPGA often needs hundreds of microseconds [7] which introduces interruption during the switch-over. To reduce the duration of interruption, a NoC architecture has been implemented as shown in Fig.1 (b). The NoC is able to forward data at Layer 2 Ethernet between on chip network functions. PHY (PCS/PMA) is also connected by the NoC to interface with the off chip optical network, so that a service chain can be set up with network functions on multiple FPGAs. The routing strategy of the NoC that creates the service chain can be re-configured at runtime by updating tables which are implemented as memory mapped control registers. In this case, during the network function switch-over process, only the registers need to be updated while FPGA configuration bit file is not required, which the reduces the duration of interrupt.

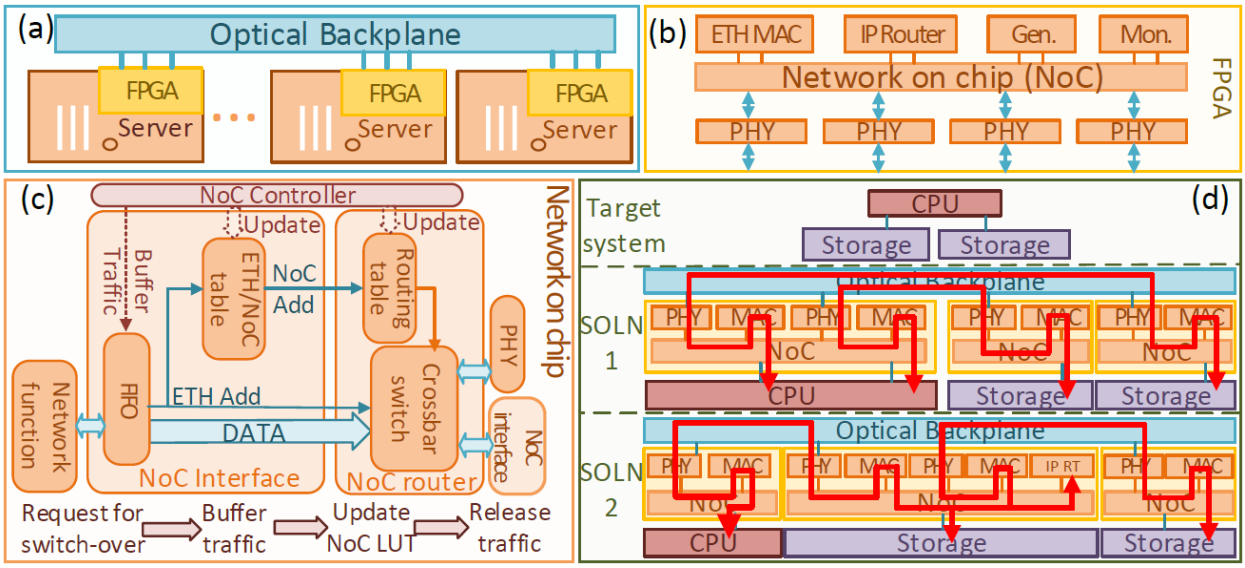

Fig. 1. The design of the proposed DCN. (a) The overall architecture. (b) The implemented flexible system based on open-hardware FPGA platform. (c) The design of the NoC and the process of the hitless-switch over (d) Rack scale architecture use case for the proposed DCN

The implementation of NoC and the process of hitless function switch-over are shown in Fig.1 (c). Every connected function block is attached with a NoC interface. The NoC interfaces are interconnected by NoC routers. All the NoC interfaces and routers are exposed to users from the NoC controller. The NoC interface is able to buffer and release traffic on demand. To perform a switch-over, the routing strategy needs to be updated so that it can forward traffic to requested function block. The NoC controller has been implemented to make the switch-over hitless by performing the following steps: buffer the traffic first, then update all the tables in NoC, and lastly release the traffic. The NoC interface also translates Ethernet address into local NoC address according to its matching table. The NoC routers forward data by looking up the related output port according to the translated NoC address.

A use case for rack scale architecture is given in Fig.1 (d). There are two solutions to achieve the target system. Direct optical circuit switched connections are established in SOLN1, which introduce minimal latency. A packet switch is set up in SOLN2, where flexible layer 3 and 4 network functions can be plugged into the service chain between CPU and storage. As both the FPGA and the optical switch are reconfigurable, the system can switch-over between SOLN1 and SOLN2 on the demand of the deployed software applications.

3. Experiment and Results

Experiments have been conducted to evaluate the optical backplane. As shown in Fig.2 (a), a Polatis 48 port single-sided all-optical switch module was set up as the optical backplane. NetFPGA SUME with low cost SFP+ (LRC 10km, 1310nm) has been set up as the traffic analyzer. Xilinx® iBERT® has been configured on the FPGA to collect the results. The bit error rate (BER) eye diagrams at different locations were collected. As shown in Fig.3 (b), the system can deliver error-free (< 1E-9) over 4 hops through the optical switch (1.1dB average loss per hop), which is sufficient margin to create a 2 tier all-optical cloud network that scales to over 10,000 endpoints. The BER curve is also collected as shown in Fig.2 (c) [8]. The receive power is around 14.3 dB at 10^{-9} BER.

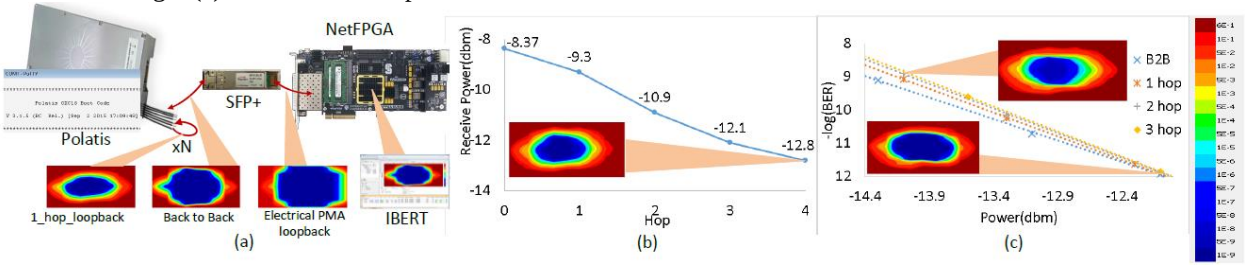

Fig. 2. (a) Experiment set up and BER eye diagram at different location (b) Relationship between hop and receive power (c) BER curve

The experimental set up for the electrical layer evaluation is shown in Fig.3. The two solutions for the target system in Fig.2 (d) was realized. Traffic analyzer was set up as the emulator of the CPU servers in the rack scale architecture. Read and write process to the storage server (i) and (ii) have been performed. NetFPGAs were set up as the emulator

of the storage servers. Data flows from the CPU emulator were stored by the storage to emulate the write process. The storage send its received data flow back to CPU emulator to emulate the read process. The latency is then measured by the traffic analyzer. The SOLN1 in Fig.2 (d) was set up as Fig.3 (a). Direct optical circuit switch was established between CPU and storages. The SOLN2 in Fig.2 (d) was set up as Fig.3 (b). In addition to optical circuit switch, an Ethernet packet switch was set up in storage (i) by interconnecting MACs and PHYs with NoC. Ethernet packets that target to storage (i) are stored and Ethernet packets that target to storage (ii) are forwarded accordingly. Fig.3 (c) shows the latency of the data flow that have access to storage (i) of the two solutions. The packet switch solution (b) introduces additional maximum 0.5 µsec. However, packet level network function can be performed at storage (i). A hitless switch-over between functions connected with NoC has also been demonstrated. The switch-over is shown in Fig.3 (d). A 9G traffic is generated by the traffic analyzer. The initial state of the flow is the green line where all the received packets are sent to the IP parser. After switch-over, the generated flow goes through the IP+UDP parser marked as brown. The number of counted packets in every 10 seconds was recorded and shown in Fig.3 (e). The total number of packets sent out of traffic analyzer is 77,518,546 which equals to the number of IP+UDP parser (42,708,766) plus IP parser (34,809,780).

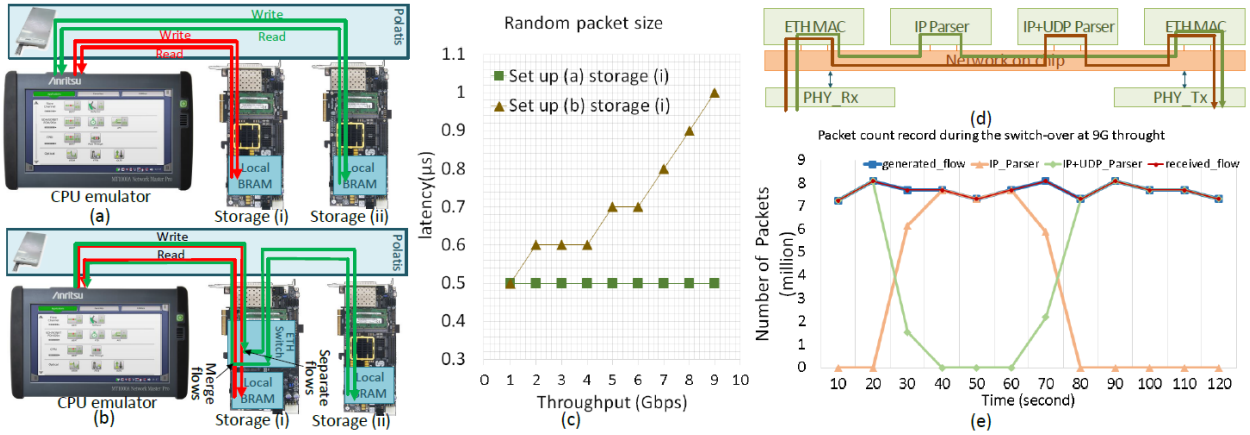

Fig. 3. (a) Experiment set up to emulate the packet solution (b) Experiment set up to emulate the circuit solution (c) Latency for the two solutions (d) Hitless switch-over process (e) Packet count during the switch-over

4. Conclusion

A programmable hybrid intra DCN has been proposed and demonstrated in this paper. It is composed of a flexible optical backplane and reconfigurable FPGA devices embedded on Servers to replace static NICs. Network function service chains have been deployed on the proposed DCN architecture to deliver high performance reconfigurable services between servers that meet the requirement of recent rack scale datacenter architecture and compute-hungry applications. Optical layer experiments have proven the ability to use low-cost transceivers with ultra-low loss reconfigurable optical backplane to realize a 2 Tier network topology. Network level results demonstrate on demand Multi-Layer service (optical circuit, Ethernet over optical, TCP/IP parsing) chaining and hitless switch over in rack scale architecture while delivering maximum throughput and sub microsecond latency.

5. Acknowledgements

The work was supported by European Union's H2020 funded dredbox project with grant agreement NO.687632.

6. References

[1] "Cisco Global Cloud Index: Forecast and Methodology, 2014–2019 White Paper," *Cisco*. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html. [04-Sep-2016].

[2] Greenberg, *et al*., "VL2: A Scalable and Flexible Data Center Network," in *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication*, New York, NY, USA, 2009, pp. 51–62.

[3] N. Farrington, *et al*., "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," in *Proceedings of the ACM SIGCOMM 2010 Conference*, New York, NY, USA, 2010, pp. 339–350.

[4] K. Katrinis, *et. al*., "Rack-scale disaggregated cloud data centers: The dReDBox project vision," in *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2016, pp. 690–695.

[5] B. Li, *et al*., "ClickNP: Highly Flexible and High Performance Network Processing with Reconfigurable Hardware," in *Proceedings of the 2016 Conference on ACM SIGCOMM 2016 Conference*, New York, NY, USA, 2016, pp. 1–14.

[6] R. Rofoee, G. Zervas, Y. Yan, N. Amaya, Y. Qin, and D. Simeonidou, "Programmable on-chip and off-chip network architecture on demand for flexible optical intra-Datacenters," *Opt. Express*, vol. 21, no. 5, p. 5475, Mar. 2013.

[7] M. Liu, W. Kuehn, Z. Lu, and A. Jantsch, "Run-time Partial Reconfiguration speed investigation and architectural design space exploration," in *2009 International Conference on Field Programmable Logic and Applications*, 2009, pp. 498–502.

[8] N. Parsons, R. Jensen, and A. Hughes, "High radix all-optical switches for software-defined data-centre networks," in Proc. ECOC 2016, Invited paper W.2.F.1.