Yan, Y., Shu, Y., Saridis, G. M., Rofoee, B. R., Zervas, G., & Simeonidou, D. (2015). FPGA-based optical programmable switch and interface card for disaggregated OPS/OCS data centre networks. In European Conference on Optical Communication, ECOC. (Vol. 2015-November). [7341957] Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/ECOC.2015.7341957

FPGA-based Optical Programmable Switch and Interface Card for Disaggregated OPS/OCS Data Centre Networks

Y. Yan[(1)], Y. Shu[(1)], G.M. Saridis[(1)], B. R. Rofoee[(1)], G. Zervas[(1)], D. Simeonidou[(1)]

[(1)] High Performance Networks, University of Bristol, United Kingdom, (yan.yan@bristol.ac.uk)

Abstract *We present an FPGA-based optical switch and interface card (SIC) for an optical disaggregated OPS/OCS DCN. It eliminates electronics from ToR switch and enables direct intra-rack blade-to-blade communication to deliver ultra-low intra-rack latency.*

Introduction

The ever-increasing demand for cloud services, and Big Data imposes a constant increase of data centre size and complexity. Most of current data centre network (DCN) [1] architecture follow a multi-tier and fat-tree hierarchy. The infrastructure is based on commodity devices and equipment. One of the challenging issues when scaling out a data centre is its network infrastructure. There is considerable interest and effort in improving data centre network architecture[2,3], however, efforts lack modularity and flexibility to deliver required performance for disaggregated data centres[4,5]. Innovative data-intensive applications require sharing of compute and memory/storage resources among large amount of servers. The disaggregation of server resources enables fine-grained resource provisioning that can be exploited by High Performance Computing (HPC) and other applications. Furthermore, recent research shows in the cloud data centres, over 80% of traffic originated by servers stays within the rack.[6] Thus, for DCNs, such as cloud data centres, special focus is needed on improving the intra-rack communication performance.

In this paper, for the first time, we proposed a FPGA-based optical programmable switch and interface card (SIC) which can replace the traditional network interface card (NIC), plugged into the server directly and enable the intra-rack blade-to-blade communication. We report on the design and implementation of the FPGA-based optical programmable SIC. And by using the SIC, we can enable a flat and scalable all-optical data centre inter- and intra- cluster architecture. The feature and functions of the SIC enables intra-rack blade-to-blade direct interconnection and eliminates the electronic devices in the Top of Rack (ToR) switch, thus minimizing the intra-rack latency, while it can be used as an optical NIC, moving data among blades and ToR. The SIC can also aggregate optical circuit switching (OCS) or optical packet switching (OPS) traffic and enable OCS-to-OPS and OPS-to-OCS conversion. Moreover, the SIC has the OPS/OCS switch function that can be used as a OPS/OCS hop. We setup a back-to-back testbed for the FPGA-based optical programmable SIC, the measurement results show the intra-rack blade-to-blade latency can reach as low as 416ns.

DCN inter, intra-cluster architecture with FPGA-based optical programmable SIC

The proposed FPGA-based optical programmable SIC enables the DCN inter- and intra-cluster architecture, displayed in Fig.1. The architecture is based on hybrid OCS and OPS technologies. The programmable DCN design supports a synthetic structure that use each technologies where and when needed. OCS and OPS complement each other since OCS can accommodate long-lived high-capacity smooth data flows with ultra-low latency, and OPS can

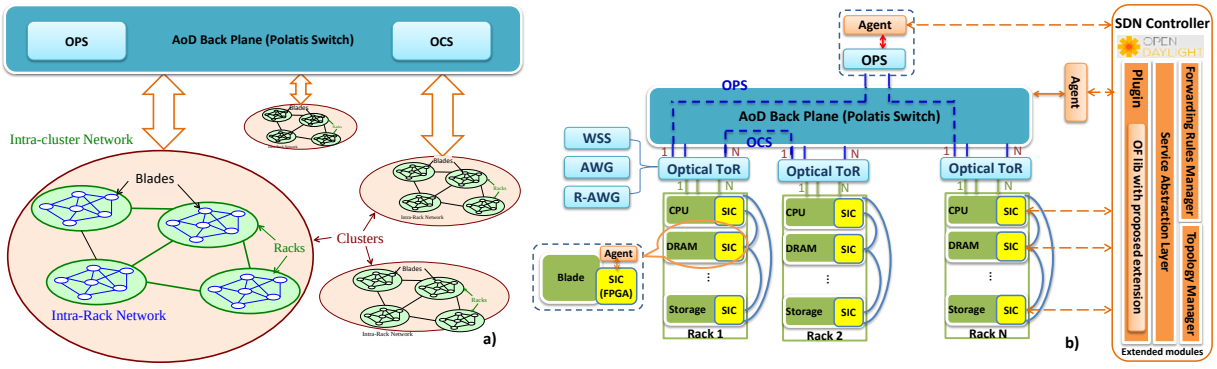

Fig. 1: a) DCN inter-cluster architecture. b) DCN intra-cluster architecture

offer flexible bandwidth capacity for each optical link when facing dynamic and unpredictable traffic demands with either short or long lived data flows. The intra-cluster architecture is shown in Fig.1b. The blades have direct connection to the other blades in the rack, each blade is capable of communicating with the optical ToR switch, which could be wavelength selective switch (WSS), array waveguide grating (AWG) or routing AWG (R-AWG), through direct optical link of FPGA-based optical programmable SIC. For the intra-cluster DCN architecture, an architecture on demand (AoD)[7,8] node interconnects all the input and output ports of different ToRs though the OCS and OPS modules, and traffic from/to other clusters as well, benefiting the flexibility and programmability on demand of AoD. While for the inter-cluster configuration, as shown in Fig.1a, a group of clusters are interconnected by an inter-cluster AoD. To fulfil the controlling mechanism enabled by software defined network (SDN) framework, each switching node and FPGA-based optical programmable SIC have been implemented with an OpenFlow agent that bridges the control plane with each data plane optical device.

FPGA-based optical programmable SIC design and implementation

FPGA-based optical programmable SIC is designed and implemented to replace the traditional NIC, plugged into the server through Peripheral Component Interconnect Express (PCIe) socket, support both intra-rack blade-to-blade communication and blade to optical ToR switch communication with the view to achieve high performance intra-rack evolving to inter-rack communication. The Hitech global HTG-V6HXT-X16PCIE was used for the prototyping, which features with Xilinx HX380T FPGA, SFP+ interfaces and Gen2 PCIe x8 interface.

As demonstrated in Fig.2a, besides the functionality of traditional NIC, reading/writing data from/to the blade, sending/receiving traffic in protocol, the optical programmable SIC is also capable of sending/receiving hybrid OPS/OCS traffic, acting as an OCS switch, an OPS switch, and an OCS/OPS, OPS-to-OCS, OCS-to-OPS aggregation interface. The FPGA-based design and implementation is shown in Fig.2b, the interfaces include interface to the server, interface to the SDN agent, inter-rack interface and intra-rack interface.

For the PCIe interface with the blade, the FGPA-based optical programmable SIC card communicates with the blade through x8 lanes of Gen2 PCIe interface. The design employs a

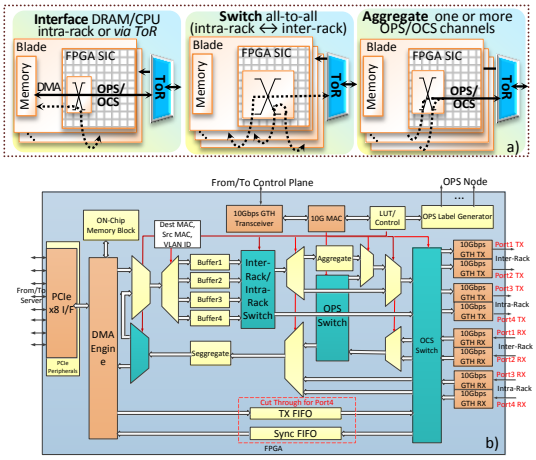

Fig. 2: a) FPGA-based optical programmable SIC functions
b) FPGA-based design and implementation

Direct Memory Access (DMA) engine to efficiently copy the data between the blade (i.e. DIMM RAM) and FPGA-based on-chip RAM.

For the interface with the control plane, the SDN agent sends the commands encapsulated in the Ethernet frame through 10Gbps interface, with the same interface and method, the FPGA-based optical programmable SIC sends feedback with its status to the SDN agent. When receiving the Ethernet frame from the SDN agent, the SIC updates its Look Up Table (LUT) with the commands, and the FPGA-based functional blocks follow the commands in the LUT to achieve certain functions.

There are two 10Gbps links implemented for hybrid OPS/OCS inter-rack communication. Based on the LUT, the traffic can be sent/received as OPS or OCS for inter-rack communication. When used as OPS or OCS switch, the received traffic is directed to the corresponding port without being processed and moved back to the blade.

There are also two 10Gbps links implementing OCS intra-rack communication interface. This implementation enables the intra-rack blade-to-blade communication. Similar to inter-rack interface, when used as OCS switch, the traffic can be directed to the other OCS interfaces without being sent back to the blade.

We designed and implemented a cut-through

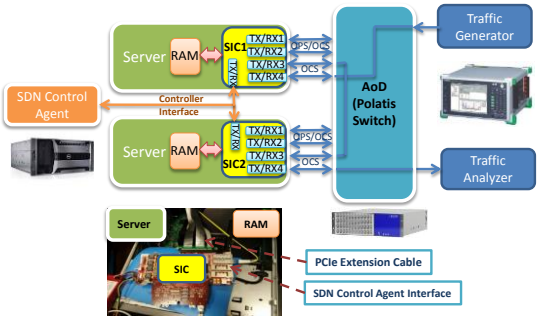

Fig. 3: FPGA-based optical programmable SIC testbed

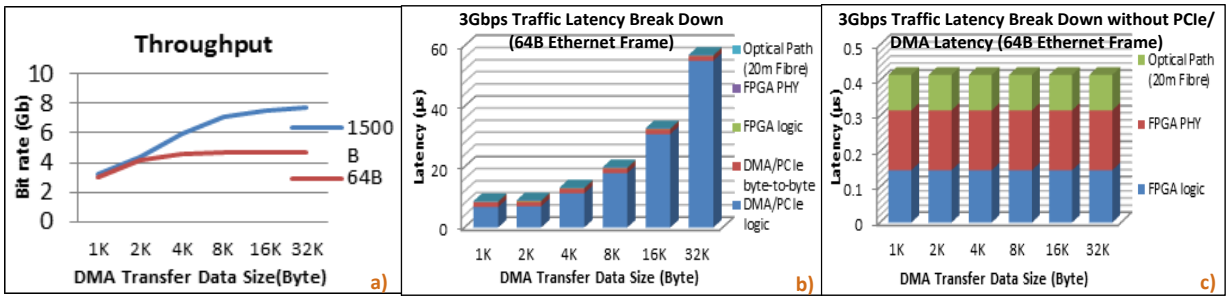

Fig. 4: Measurement result. a) Throughput, b) Latency break down, c) Latency break down without PCIe/DMA latency

option for port4 which eliminates all the store-forward delays in multiple FIFOs to deliver ultra-low latency service for disaggregated memory to/from processing blade communication.

FPGA-based optical programmable SIC testbed setup and experimental result

We tested the performance of FPGA-based optical programmable SIC using the testbed shown in Fig.3. The SIC is connected to the DELL Poweredge 710 server PCIe Gen2 socket through PCIe extension cable. We used Polatis 192x192 fibre switch as AoD optical backplane. The SDN agent hosted in a server is connected with the controller agent interface of the SIC through SFP+ interface. A traffic generator (Anritsu MD1230B) was used to generate the Ethernet traffic and feed to the OCS port4 of SIC1. The port4 was set as cut-through mode. A traffic analyser analysed the measurement result from the output port4 of SIC2.

When FPGA-based optical programmable SIC receives the Ethernet traffic, it processes the data, and enables DMA engine to move the processed data through PCIe to the server RAM. Then when RAM receives a full block of data, the DMA engine initiates the transmission from the server RAM and reads data back to the FPGA. The SIC processes the data and transmits them out.

In this experiment, we measured the maximum throughput and the latency. The maximum throughput measurement result is shown in Fig.4a. When measuring throughput, as described above, data was written to the RAM, and after the block of RAM was filled, data was read back. The maximum throughput is limited because of this non-duplex transmission. We measured the latency on cut-through mode since the cut-through FIFO (compared to store-forward FIFO) helps minimizing the latency. Fig.4b shows the 3Gbps 64B traffic latency break down by DMA/PCIe latency, FPGA logic latency, FPGA PHY latency and optical path (20meters fibre) latency. From the chart, majority of the time were spent on the DMA/PCIe logic. This latency is mostly dependent on the DMA engine core, server CPU respond time and PCIe socket/cable quality. Without considering PCIe/DMA latency, we can get a minimum of 416ns latency for cut-through intra-rack blade-to-blade communication on 3Gbps 64B Ethernet frame traffic.

Conclusions

This paper reports the inter- and intra-cluster data centre network architecture by using FPGA-based optical programmable SIC. We demonstrated FPGA-based optical programmable SIC design, implementation and back-to-back throughput and latency results. The FPGA-based optical programmable SIC is featured with multi-functionality, flexibility and programmability. The measurement results show ultra-low latency (416ns) for intra-rack blade-to-blade communication.

Acknowledgements

This work is supported by LIGHTNESS and COSIGN projects funded by European Commission and SONATAS funded by EPSRC (UK).

References

[1] R. Branch et al., “Cloud Computing and Big Data: A Review of Current Service Models and Hardware Perspectives,” Journal of Software Engineering and Applications, vol.7, no. 8, pp. 686-693,(2014).

[2] T. Benason et al., “Network Traffic Characteristics of Data Center in the Wild” IMC’10, November 1-3, (2010)

[3] M. Al-Fares et al., ”A scalable, Commodity Data Center Network Architecture”, SIGCOMM, pages 63-74, (2008)

[4] S. Han et al., “Network Support for Resource Disaggregation in Next-Generation Datacenters”, ACM,(2013)

[5] G. Saridis et al., “DORIOS: Demonstration of an All-Optical Distributed CPU, Memory, Storage Intra DCN Interconnect,” OFC, W1D.2,(2015)

[6] N. Farrington et al., “Helios: a Hybrid Electrical/Optical Switch Architecture for Modular Data Centers,” ACM SIGCOMM. 2011, vol. 41, no. 4, pp. 339–350, (2011).

[7] B.Rofoee, et al, “Programmable on-chip and off-chip network architecture on demand for flexible optical intra-datacentres”, Journal of Optical Express, v21, (2013).

[8] Y. Yan et al., “FPGA-based Optical Network Function Programmable Node”, OFC, (2014).