SiP-enabled FPGA Network Interface for Programmable Access to Disaggregated Data Centre Resources

Qianqiao Chen [1,4], Vaibhawa Mishra [1], Peter De Dobbelaere[2], Michael Enrico[3], Nick Parsons[3], Jose Nunez-Yanez[4], Georgios Zervas [1]

[1]University College London, London WC1E 6BT, UK, [2]Luxtera Inc., Carlsbad CA 92011, USA, [3]HUBER+SUHNER Polatis, Cambridge CB4 0WN, UK, [4]University of Bristol, Bristol BS8 1TH, UK

qianqiao.chen@ucl.ac.uk

Abstract: This paper demonstrates a FPGA-based network interface with reconfigurable switched ports that access to disaggregated resources. Silicon-Photonics on-board transceivers and miniaturized optical switches deliver bandwidth density and FEC-free scalability of up to 5-tier network.

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

1. Introduction

Cloud computing enables organizations and enterprises to rent IT services from data centres housing large amounts of computing resource. Recently, data centre memory disaggregation [1] was proposed to maximize resource utilization [2]. By interconnecting them with a dynamic and flexible network, these resource-disaggregated servers can form on-demand runtime systems [3].

Resource access of disaggregated servers (i.e. CPU-to-memory, CPU-to-Internet, and VM-to-VM) each require a distinct set of services such as specific protocols, low deterministic latency, statistical multiplexing among others. In addition, the bandwidth requirement of these resource access services may vary with the applications. For example, database applications require high bandwidth on the storage access services, while HTTP software servers need more on the services of Internet access. As the existing network interface card (NIC) or switch interface card (SIC) is only accompanied with fixed resource access services, the bandwidth will be wasted when the deployed applications do not need the associated resource. Furthermore, the dedicated NIC chipset substantially increase the latency, price and power consumption of the servers.

To increase bandwidth of the optical data centre network (DCN), recent Silicon-Photonic (SiP) mid board optics (MBO) have been fabricated. Table 1 compares the single-mode fibre based transceivers with on-board optical transceivers. SiP-based transceivers are able to double the bandwidth while offering higher density and lower energy consumption.

Tab. 1: SMF-based 1.3 μm 100Gbps + transceivers

Module	QSFP28	CFP4	OptoPHY®
Line rate (Gbps)	4x25	4x25	8x25
BW volume (Mbps/mm^3)	12.6	4.9	15.5
Energy Efficiency (pj/bit)	35	50	23

This paper demonstrates an FPGA-based network interface for resource disaggregated servers with the following novelties: (a) It enables reconfigurable access services to any type of disaggregated data centre resource on each port. An on chip interconnect has been implemented to enable the dynamic resource access reconfiguration. (b) The SiP MBO transceiver (OptoPHY) has been integrated to further increase the bandwidth density. As opposed to multimode fibre based MBOs, we demonstrate that the single mode fibre based SiP MBO can deliver up to a 5-tier all-optical network. By using miniaturized and ultra-low insertion loss circuit switches with 10E-12 BER, the proposed network interface does not require FEC and thus reduces latency by ~100nsec. Network layer results demonstrate resource access reconfiguration on each physical port. The performance for memory access (0.6 μsec), storage access (support 9.6G burst traffic) and layer 2 Ethernet switch (maximum 1 μsec) is reported.

2. Run-time Reconfigurable Network interface for disaggregated resource access

The network interface is made of a high bandwidth SiP MBO optical transceiver connected to a FPGA high-speed I/Os as shown in Fig. 1. An on chip hybrid packet/circuit interconnect module has been implemented to link on-chip high-speed I/Os with local resource (i.e. CPU, memory). Furthermore, a set of functional blocks for resource access services can be added on the interconnect module (i.e. protocol engine, packet parser, buffer). Required access services can be added between local resource and high speed I/Os. The services can be runtime reconfigured by reprograming the port map (control registers) of the interconnect module.

All these network interfaces and their attached data centre resource are connected by an optical switch to build up a complete resource disaggregated data centre network (DCN). A central network controller is responsible to

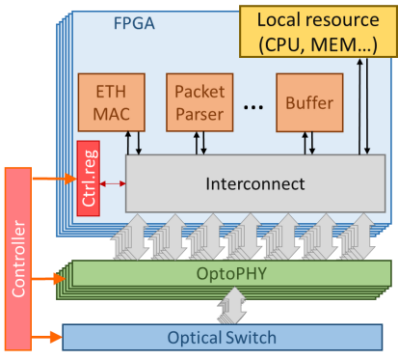

Fig. 1: The network interface for the access of disaggregated data centre resource

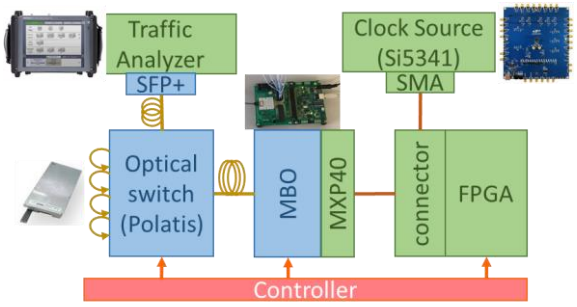

Fig. 2: Experimental setup. The Xilinx VCU109 with BullsEye connector is used for BER results. The NetFPGA with FMC SMA-daughter card is set up for network interface.

manage the resource access reconfiguration. The port map of the on-chip interconnect and optical switch can be controlled by a central controller to build on demand network topologies.

Therefore, a network for runtime reconfigurable resource access with the following features has been built: a) Instead of memory mapped processing, FPGA based parallel stream processing can support high bandwidth and low latency. b) The interconnect module is runtime reprogrammable with the standard AMBA AXI4-lite control interface which can be accessed by local or remote ARM processors. c) High bandwidth SiP MBO transceivers with low-insertion loss circuit switches to deliver scalable, low-power and low-latency optical DCN.

3. Experiment set up

The setup of the experiment is shown in Fig. 2. The reference clock of the on chip transceiver (GTY) is generated by the Si5341 clock generators from Silicon Lab (200MHz and 312.5MHz for 16Gbps and 25Gbps respectively). The Xilinx® iBERT® has been configured on the FPGA for the bit error rate (BER) measurement. And the OptoPHY is controlled by the Luxtera Control Suite (LCS). The received power can be monitored through the LCS and the clock data recovery (CDR) on the OptoPHY can be enabled here.

4. Physical layer results

The measured BER for 16Gbps and 25Gbps back-to-back and through a number of optical switch cross-connections have been shown in Fig. 3. Enabling the CDR at 25Gbps offers 1 dB improvement. As shown on Fig. 3b, a network of 4 hops and 7 hops was supported at 25 Gbps (with CDR) and 16 Gbps while achieving 1E-12 BER. This crucially eliminates the need for FEC that contributes ~100nsec latency critical for compute to memory transactions. Considering that the 12-fibre MPO-to-LC fan-out harnesses were used between the SiP MBOs and optical switch with ~1dB loss each it is expected to achieve 2 more hops (1dB/hop loss) with same performance when using same connector type. This corresponds to 3-tier and 5-tier networks respectively that scale to 100,000 of end-points.

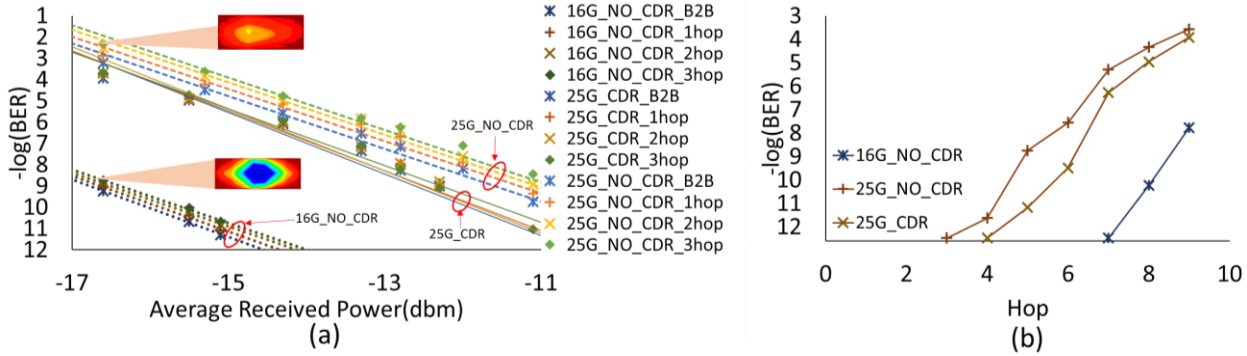

Fig. 3: Physical layer MBO (OptoPHY) result: (a) Average Received Power vs BER (b) Hop vs BER

5. Network layer results

The NetFPGA has been set up as the FPGA platform. An SMA FMC daughter card has been added on the board to interface 8 channels at 10Gbps each to SiP MBO.

The dynamic port-level resource access reconfiguration is demonstrated in Fig. 4. All the physical ports have been initially configured as memory access ports with on-chip circuit switching to support maximum remote access and minimum deterministic latency between one CPU and 5 remote memories. After the reconfiguration process, the

FPGA on-chip interconnect LUT is configured to offer 2 of the ports as a Layer2 Ethernet switch and 2 of the ports performing access to a storage area network (SAN). The update request over 10Gbps interface is generated from the controller and the respond received by the controller has been demonstrated below using Wireshark.

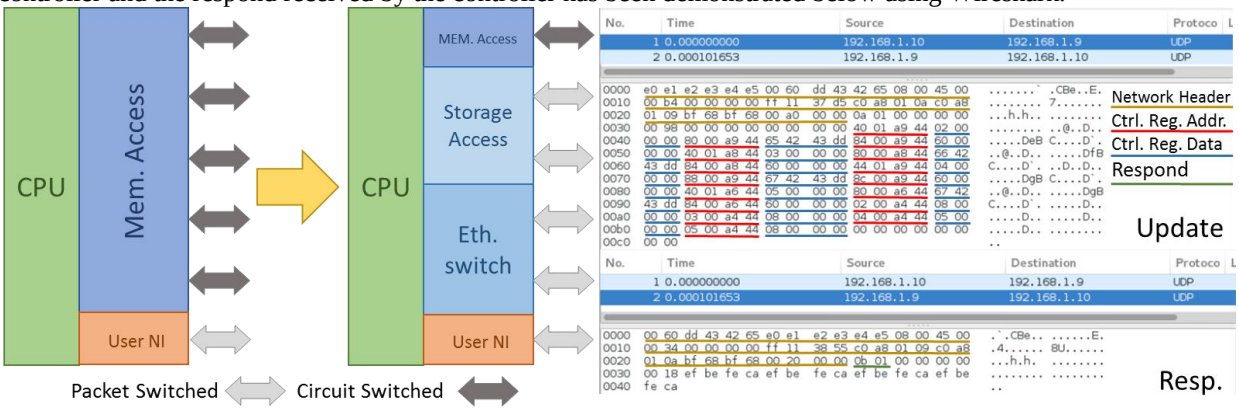

Fig. 4: Demonstration for the reconfiguration from memory access to storage access and Ethernet switch service

The performance of the Ethernet switch, storage access and memory access ports are reported in Fig. 5. The traffic pattern for the Ethernet switch is a continuous bi-modal traffic coming from two physical ports and forwarding to 1 port with packet size distribution shown in Fig. 5a. The traffic pattern for storage access is a burst traffic emulating bulk data transfer coming from 2 physical ports with throughput distribution shown in Fig. 5b. The period of the burst is set as 10 µsec. And the data rate of the traffic is controlled by changing the duty cycle. The latency is shown in Fig. 5c. The system can achieve ultralow deterministic latency (0.6 µsec) using circuit-switched on-chip and chip-to-chip memory access. And the proposed network interface is able to deliver Ethernet switch service with low latency (less than 1 µsec). The network interface can also work at high throughput (9.6G burst mode traffic) to deliver access to storage.

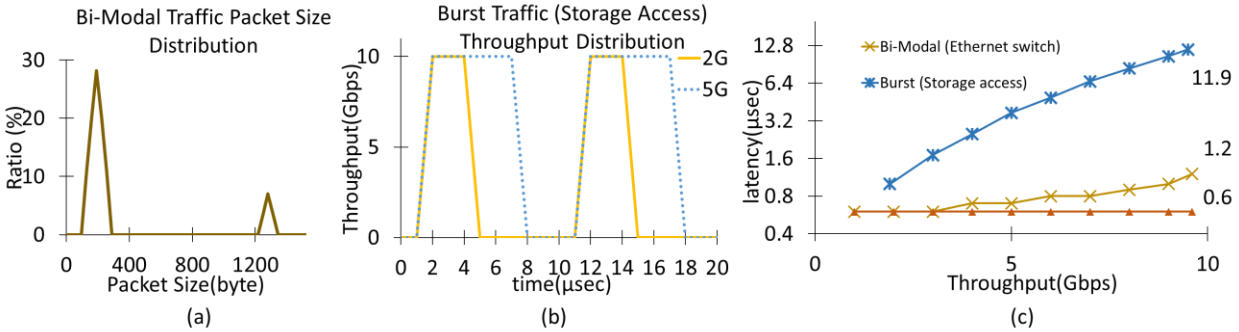

Fig. 5: The latency of the proposed network interface. (a) The Bi-Modal traffic pattern for Ethernet switch. (b) The burst traffic pattern for storage access. (c) Latency vs throughput

4. Conclusions

A FPGA network interface for disaggregated data center resource access is proposed and demonstrated. It matches network access requirements of disaggregated resources at port level using runtime reconfiguration of on-chip packet/circuit and other services Using SiP MBO transceivers. It is possible to achieve up to a 3-tier and 5-tier all-optical network with FEC-free 10E-12 BER operation. The interface delivers low latency and high bandwidth services.

Acknowledgements

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

References

[1] Lim, K. et al., 2009. Disaggregated Memory for Expansion and Sharing in Blade Servers. In Proceedings of the 36th Annual Inter-national Symposium on Computer Architecture. ISCA '09. New York, NY, USA: ACM, pp. 267–278J.

[2] G Zervas, et. al., "Disaggregated Compute, Memory and Network Systems: A New Era for Optical Data Centre Architectures" OFC 2017, W3D. 4

[3] Qianqiao Chen, Vaibhawa Mishra, Nick Parsons, Georgios Zervas, "Hardware Programmable Network Function Service Chain on Optical Rack-Scale Data Centers", OFC 2017