All-Optical Programmable Disaggregated Data Centre Network realized by FPGA-based Switch and Interface Card

Yan Yan, George M. Saridis *Student Member, IEEE,* Yi Shu, Bijan R. Rofoee, Shuangyi Yan, Murat Arslan, Tom Bradley, Natalie V. Wheeler, Nicholas H.L. Wong *Student Member, IEEE*, Francesco Poletti, Marco N. Petrovich, David J. Richardson, Simon Poole, George Zervas, *Member, IEEE,* Dimitra Simeonidou, *Member, IEEE*

***Abstract*—This paper reports a FPGA-based Switch and Interface Card (SIC) and its application scenario in an all-optical, programmable disaggregated Data Centre Network (DCN). Our novel SIC is designed and implemented to replace traditional optical Network Interface Cards (NICs), plugged into the server directly, supporting Optical Packet Switching (OPS)/ Optical Circuit Switching (OCS) or Time Division Multiplexing (TDM)/ Wavelength Division Multiplexing (WDM) traffic on demand. Placing the SIC in each server/blade, we eliminate electronics from the top of rack (ToR) switch by pushing all the functionality on each blade while enabling direct intra-rack blade-to-blade communication to deliver ultra-low chip-to-chip latency. We demonstrate the disaggregated DCN architecture scenarios along with all-optical dimension-programmable N×M Spectrum Selective Switches (SSS) and an Architecture-on-Demand (AoD) optical backplane. OPS and OCS complement each other as do TDM and WDM, which can support variable traffic flows. A flat disaggregated DCN architecture is realized by connecting the optical ToR switches directly to either an optical Top of Cluster (ToC) switch or the intra-cluster AoD optical backplane, while clusters are further interconnected to an inter-cluster AoD for scaling out.**

***Index Terms*—Disaggregated Data Center Networking, FPGA-based, Optical Network Interface Card, Optical Packet Switching, Optical Circuit Switching, Time Division Multiplexing, Wavelength Division Multiplexing**

Manuscript received October 15th 2015. Current version published xx.xx. 2015. This work was presented in part as one of top-scored papers at ECOC 2015[1].

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

Yan Yan, George M. Saridis, Yi Shu, Bijan R. Rofoee, Shuangyi Yan, Murat Arslan, Georgios Zervas and Dimitra Simeonidou are with the High Performance Networks Group, University of Bristol, United Kingdom. (e-mail: yan.yan@bristol.ac.uk)

Tom Bradley, Natalie V. Wheeler, Nicholas H.L. Wong, Francesco Poletti, Marco N. Petrovich, David J. Richardson are with ORC, University of Southampton, United Kingdom

Simon Poole is with Finisar Australia, Rosebery, Australia.

I. Introduction

THE ever-increasing demand for cloud services and Big Data imposes a constant increase of data center size and complexity. According to the IDC predictions [2], 90% of IT industry growth from 2013 to 2020 will be driven by Cloud services, Big Data analytics, mobile and social network technologies. Most of current data center network (DCN) [3] architectures follow a multi-tier and fat-tree hierarchy. The infrastructure is based on commodity devices and equipments. One of the challenging issues when scaling out a data center is its network infrastructure. Recent research around cloud data centers (DCs) shows over 80% of traffic originated by servers stays within the rack [4]. By 2018, more than three quarters (78%) of workloads will be processed by cloud data centers [5]. Thus, for next-generation DCNs, special focus is needed on improving the intra-data center communication performance.

There is considerable interest and effort in improving data center network architectures [6,7], however, efforts lack modularity and flexibility to deliver required performance for disaggregated data centers [8-10]. Traditional data centers have a relatively static computing infrastructure [11], normally a number of servers, and each with a set number of CPUs and fixed amount of memory. However, workloads and traffic patterns, in commercial as well as in High Performance Computing (HPC) data centers, show huge variation and a great degree of unpredictability. This erratic traffic inside DCs stems

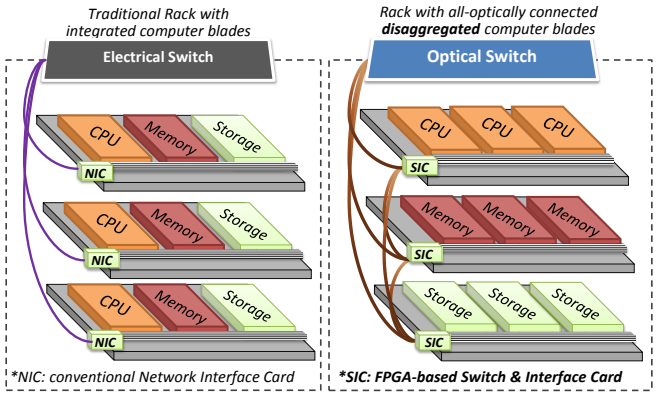

Fig. 1. Conventional integrated servers rack(left) & all-optically connected disaggregated computer blades rack approach with FPGA-based SIC (right)

from the large differences that are observed between the requirements of the various applications or tasks that run in such environments. Thus, the data center may safely supply adequate service during peak-demand, yet cannot fully utilize those same resources during non-peak conditions. Furthermore innovative data-intensive applications require sharing of compute and memory/storage resources among large amounts of servers, such as in modern highly parallelized computer architectures. Moving away from the traditional server-rack-cluster architecture, a need for novel arrangement and interconnection approaches has appeared. A promising approach to satisfy the above conditions, while adding modularity and upgradability, is the physical disaggregation of the DC resources (CPU, DRAM, storage) which will then be joined by an all-optical low-power consumption high-capacity interconnect with minimal chip-to-chip delays (as shown in Fig.1). It is quite obvious that ultra-low-latency connectivity and protocols need to be established, especially between processing and remote memory chips/blades, so that maximum performance is maintained [12]. Modular and flat architectures consisting of optical and electrical technologies are able to support this notion, by incorporating dimension reconfigurable capacity-agnostic optical switches, and with fully-programmable, integrated SICs within and among the disaggregated blades. Instead of traditional star topology connecting intra-rack blades with the ToR switch, a full mesh interconnects among the intra-rack servers will further minimize the latency. This disaggregation of server resources can enable fine-grained resource provisioning consequently boosting the overall system performance.

In this paper, we demonstrate a novel FPGA-based optical programmable SIC, and by using the SIC, we show an all-optical, programmable disaggregated DCN architecture with OPS/OCS or TDM/WDM techniques. The SIC could eventually replace the traditional NIC, by being plugged into the server directly and enabling intense intra/inter-rack blade-to-blade chip-to-chip communication. We report on the design and implementation of the SIC. Using the SIC, we enable flat and scalable all-optical intra-DC interconnection scenarios, for intra/inter-rack as well as intra/inter-cluster communication. The features and functions of the SIC enable intra-rack blade-to-blade direct interconnection eliminating the need for electronic devices in Top of Rack (ToR) switches (Fig.1), thus minimizing the intra-rack latency. Simultaneously, it can be used as an optical interface, transferring data traffic among blades and optical ToR switches. Additionally, the SIC is able to aggregate traffic and also perform OCS-to-OPS or WDM-to-TDM conversion and vice versa. Moreover, the SIC has the OPS/OCS or TDM/WDM switch functionality which can be used for OPS/OCS and TDM/WDM multi-hopping, while it also supports Layer2 switching functionality. We experimentally demonstrate two main scenarios with the SICs; one is the back-to-back testbed for the SIC-only measurement; the second one employs the SIC in an all-optical disaggregated DCN architecture, which by also utilizing Hollow-Core Photonic Bandgap Fiber (HC-PBGF) links, shows flexible ultra-low latency intra-DCN interconnection. The measurement results show 360.8 ns and 860 ns chip-level access latency for cut-through intra-rack and inter-rack respectively.

II. Disaggregated DCN inter-cluster and intra-cluster architecture with OPS/OCS techniques

The proposed SIC enables the all-optical programmable DCN inter- and intra-cluster architecture. This section shows an example of all-optical programmable disaggregated DCN architecture based on hybrid OCS and OPS technology [13]. The programmable DCN design supports a synthetic structure that uses each technology on a needs basis. OCS and OPS complement each other since OCS can accommodate long-lived high-capacity smooth data flows with ultra-low latency, and OPS can offer flexible bandwidth capacity for each optical link when facing dynamic and unpredictable traffic demands with either short or long lived data flows. We also employed an AoD [14,15] optical backplane to supply the programmable and flexible optical interconnect.

A. *DCN intra-cluster architecture*

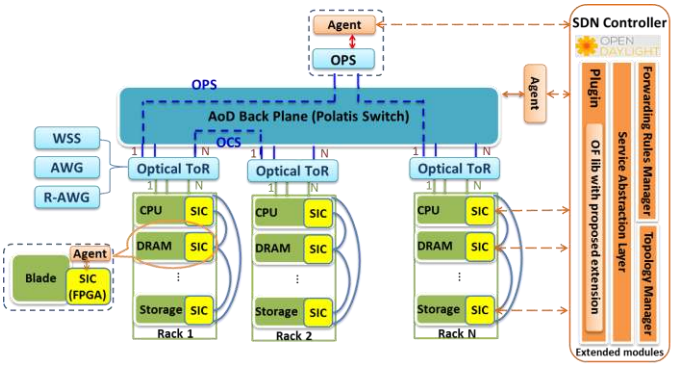

Fig. 2. DCN intra-cluster architecture with SDN controller

We designed and implemented the DCN intra-cluster architecture with the SIC to offer the best intra-rack and inter-rack latency performance.

The intra-cluster architecture is shown in Fig. 2. With the novel SIC plugged into each blade directly, the blades are all-to-all directly connected with each other in the rack, and each blade is capable of communicating with the optical ToR switch, that could be a wavelength selective switch (WSS), an array waveguide grating (AWG) or a routing AWG (R-AWG), through the direct optical link of the SIC.

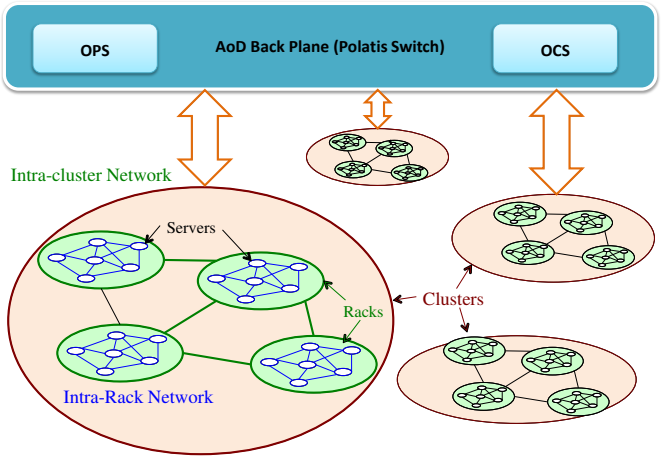

Fig. 3. DCN inter-cluster architecture based on OPS/OCS technology

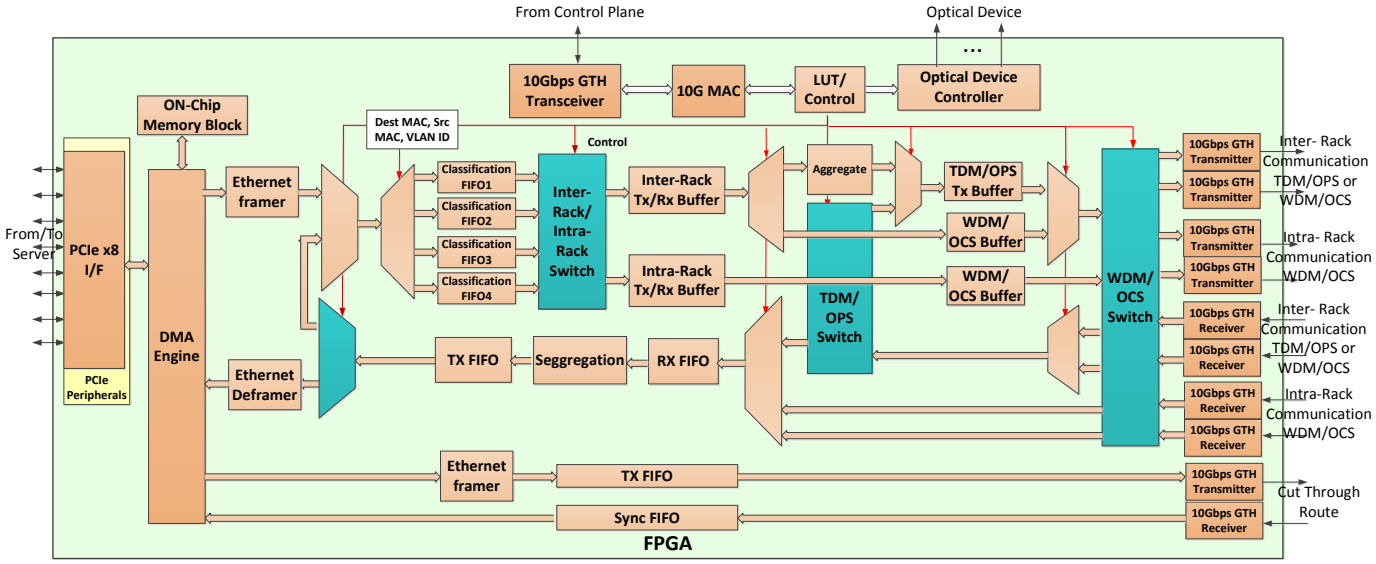

Fig. 4. FPGA-based Optical Programmable SIC design and implementation functional blocks architecture

For the intra-cluster DCN architecture, an AoD node interconnects all the input and output ports of ToR switches through the OCS and OPS modules, and traffic from/to other clusters as well, benefiting the flexibility and programmability on demand of AoD.

Our solution is SDN-enabled. To fulfil the controlling mechanism enabled SDN framework, we extended the Openflow library with proposed extensions and also developed our own OpenDayLight agent to communicate with the optical devices and our SIC [16]. The OpenDayLight agent is capable of translating the information between the SDN controller and the optical devices.

B. DCN inter-cluster architecture

Further to the disaggregated DCN intra-cluster architecture, we also designed the disaggregated DCN inter-cluster architecture, shown in Fig.3. The blades are full mesh connected in the racks, while all the racks in the cluster are connected to an intra-cluster AoD node. A group of clusters are further interconnected by an inter-cluster AoD.

III. FPGA-BASED OPTICAL PROGRAMMABLE SIC DESIGN AND IMPLEMENTATION

Our SIC is designed and implemented in FPGA for flexibility and programmability. The Hitech global HTG-V6HXT-X16PCIE board was used for the prototyping, which features with Xilinx Virtex-6 HX380T FPGA, SFP+ interfaces and Gen2 Peripheral Component Interconnect Express (PCIe) x8 interface.

A. FPGA-based optical programmable SIC functionalities

The idea of SIC is to design and implement an interface card that can replace the traditional NIC, plugged into the server through the Peripheral Component Interconnect Express (PCIe) socket. Therefore, the functionality of communicating with the server is the first priority. The SIC is capable of initiating the copy of data between the memories of the blades and SIC, besides, it is also able to accept the copy commands initiated from the server. All the data copied to the block RAM of the SIC are processed and sent out in particular ports of TDM/OPS or WDM/OCS, based on the instruction of the control plane.

Another important functionality of the SIC is switching. Following the commands of the control plane, the SIC can act as an OCS/WDM switch, an OPS/TDM switch or a Layer 2 switch. With the switch functions, the SIC can work as a hop to supply maximum flexibility and programmability in the DCN architecture.

The SIC is also capable of aggregating one or more channels, and sending them out in required format. As a result, the SIC supports 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 FPGA-based design and implementations are considered based on the functionalities that need to be realized. For the PCIe interface, exchanging Transaction Layer Packet (TLP) directly is not bandwidth or latency efficient due to the overhead and PCIe communication mechanism, therefore, we employed Direct Memory Access (DMA) to access the blades' memory. Furthermore, we designed our own Ethernet framer/deframer, instead of using Ethernet MAC core, to get the required functions and minimize the latency. Lastly, we implemented hitless switches to enable the hitless switchover on-the-fly.

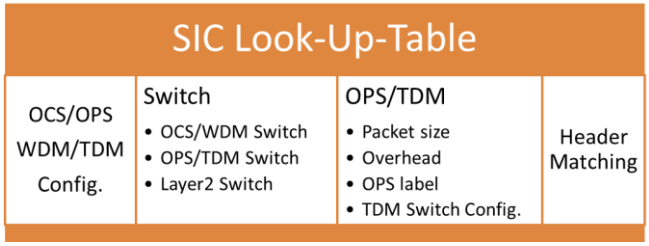

Fig. 5. FPGA-based SIC LUT

B. FPGA-based optical programmable SIC design and implementation architecture

The SIC design and implementation architecture are shown in Fig. 4. The data flow follows the arrow direction in the

design. When receiving from server side, the SIC can send the formatted traffic through inter-rack or intra-rack interfaces on demand, and vice versa. There are four interfaces communicating out of FPGA: the interface connecting with the SDN control plane, the PCIe interface talking with the disaggregated blades, the inter-rack and intra-rack interfaces, and the optical device controller interface.

For the interface with the control plane, the SDN agent sends commands encapsulated in Ethernet frames via a 10Gbps interface. With the same interface and method, the SIC sends feedback with its status back 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. The content of the LUT is displayed in Fig. 5, which mainly has four major information parts. Firstly, there is OCS/OPS and WDM/TDM configuration information; the output ports transmit out OPS/OCS or WDM/TDM traffic based on the commands stored here. Secondly, there is switching information, as explained above, the SIC can act as an OCS/WDM switch, an OPS/TDM switch or a Layer 2 switch, when set as '1', the SIC performs the corresponding switch function as required. Thirdly, there is OPS/TDM information that includes the optical packet size, the Quality of Transmission (QoT) overhead size, and the OPS label, which can get updated hitless on-the-fly. Lastly is traffic flow header matching information. The matching mechanism is shown in Fig. 6. The SIC can match based on the destination MAC address, source MAC address, and VLAN ID of the Ethernet frame. The traffic flow header matching LUT stores both the header and the mask. As shown in Fig. 6, the mask can be '1' (which means select) or '0' (which means do not care), then in the example LUT of Buffer1:

$$Matching = Header\ AND\ Mask$$

With the 16 bits header, we can match 65536 flows in total.

Traffic Flow Header Matching Look-Up-Table

Byte	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Header	Destination MAC Address						Source MAC Address						VLAN			
MASK	XX	XX	XX	XX	XX	XX	XX	XX	XX	XX	XX	XX	XX	XX	XX	XX

Example LUT of Buffer 1

Byte	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Header	11.	22.	33.	44.	55.	66	77.	88.	99.	AA.	BB.	CC	11	22	33	44
MASK	00	00	00	00	00	0F	00	00	00	00	00	00	00	00	00	01
Matching	XX	XX	XX	XX	XX	X6	XX	XX	XX	XX	XX	XX	XX	XX	XX	XXXXXXX0

Fig. 6. FPGA-based SIC traffic flow header matching LUT

For the PCIe interface with the blade, the SIC communicates with the blade through ×8 lanes of Gen2 PCIe interface. The design employs a DMA engine to generate memory addresses and initiate memory read/write cycles and offload memory operations from the CPU to the dedicated DMA engine. This enables efficiently copying the data between the blade (i.e. memory DIMM) and SIC on-chip RAM.

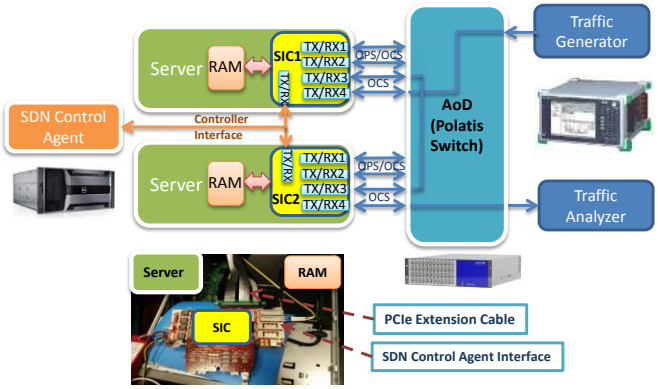

Fig. 7. FPGA-based optical programmable SIC back-to-back testbed setup

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

Meanwhile, there are also two 10Gbps links implementing OCS or WDM intra-rack communication interfaces. This implementation enables the intra-rack blade-to-blade communication. Similar to inter-rack interface, when the SIC is used as a switch, traffic can be directly forwarded to other ports without returning back to the blade.

Furthermore, to get a minimum inter-blades latency result, we designed and implemented a cut-through option for port4 which eliminates all the store-forward delays in multiple first-in-first-outs (FIFOs). Compared to store-forward FIFO, the cut-through FIFO has better latency result benefits from not needing to wait for the complete frame/packet ready checking the Frame Check Sequence (FCS) then sending out. However, it suffers from low correction ability, which forwards the error frames/packets to the network. We use the cut-through FIFO for one intra-rack (OCS or WDM) port to deliver ultra-low latency service for communication of disaggregated memory and processing blade.

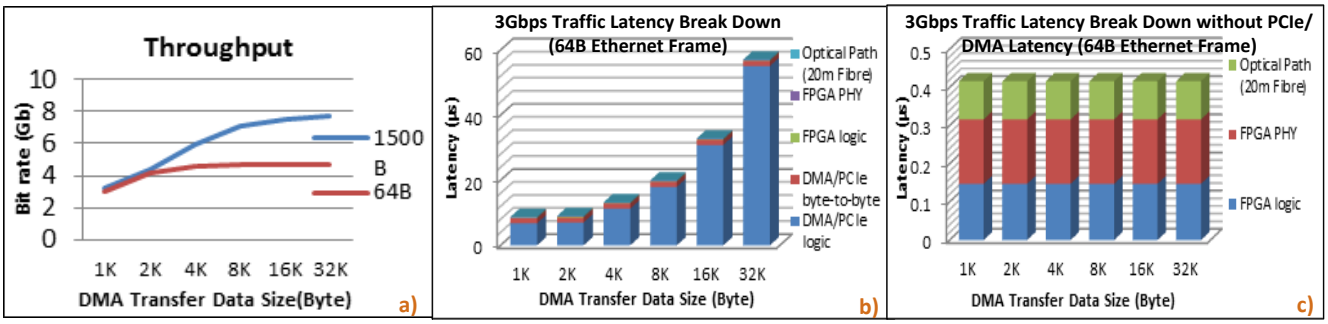

Fig. 8. Measurement result: a) Throughput, b) Latency break down, c) Latency break down without PCIe/DMA latency

IV. Testbed Setup and Experimental Result

To test the performance of the SIC, we setup a testbed with back-to-back interconnect to measure the SIC-only performance and also setup a testbed utilizing HC-PBGF to check the performance in the network based on TDM and WDM technologies.

A. FPGA-based optical programmable SIC back-to-back test

We tested the performance of SIC using the testbed shown in Fig. 7. The SICs are connected to the DELL Poweredge 710 servers PCIe Gen2 sockets through PCIe extension cables. We used a Polatis 192×192 fibre switch as an 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 the port4 of the SIC1. The port4 was set as cut-through mode. A traffic analyzer (Anritsu MD1230B) was used to collect the results from the output port4 of the SIC2.

When the 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 SIC. After receiving the data, the SIC processes it and transmits it out.

In this experiment, we measured the maximum throughput and the latency. The maximum throughput measurement result is shown in Fig. 8a. For throughput measurement, as described above, data was written to RAM, and after the block of RAM was filled, data was read back. The maximum throughput is limited because of this non-duplex transmission. For latency measurement, our previous work [15] has included TDM/WDM latency, in this paper we focused on measured latency in cut-through mode since cut-through FIFO (compared to store-forward FIFO) helps minimize the latency. Fig. 8b shows the 3Gbps 64B traffic latency break down by DMA/PCIe latency, FPGA logic latency, FPGA PHY latency and optical path (20m fiber) 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, displayed in Fig. 8c.

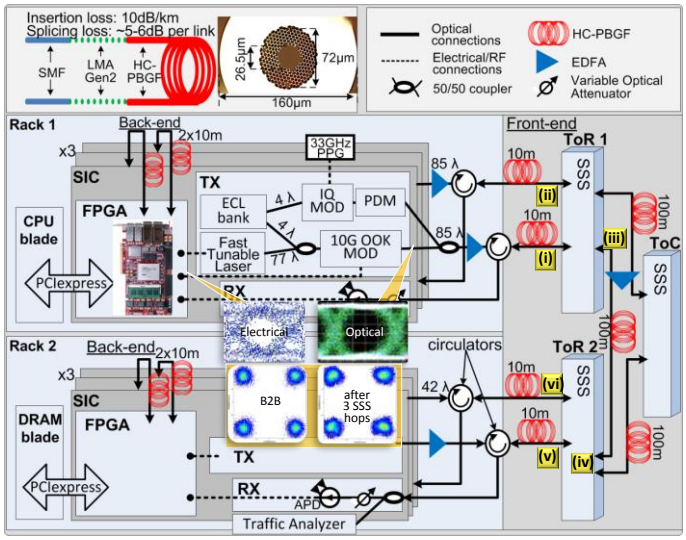

Fig. 9. Experimental setup with FPGA-based optical programmable SIC. Insets: HC-PBGF details, electrical & optical eyes, DP-QPSK constellations

B. All-Optical Programmable Disaggregated Data Center Interconnect experiment using SIC & HC-PBGF

The experimental test-bed (shown in Fig. 9), is comprised of two FPGA-based SIC boards which support real-traffic scenarios with fully-functional transceivers and perform as the interface between the in-board memory chip and the back/front-end network. Three re-configurable and multi-dimensional (4×16/8×12) 20-port spectrum selective switches (SSS) were utilized as all-optical ToR and ToC switches. In addition, a traffic analyzer is used to supply the system with real-traffic by scrabbling Pseudorandom Binary Sequence (PRBS).

Moreover, nine spools of HC-PBGF [16] were used for interconnection purposes in the different scenarios. In order to maintain single-mode propagation within the generally multi-moded HC-PBGF, each spool was first of all spliced with Large Mode Area (LMA) fiber and then with single mode fiber (SMF) pig-tails (as seen in inset of Fig. 9), From the total of nine spools, six of them are 10m long and used for intra-rack (both for the back-end SIC-to-SIC direct connections and the front-end SIC-to-ToR ones) while the remaining three are 100m long and used for inter-rack communication. HC-PBGF is suitable for all the interconnection links due to its physical properties and more specifically due to its reduced propagation delay compared to standard single model fiber (SSMF). Light in HC-PBGF propagates in almost vacuum conditions, about 1.46 times faster than silica-cored SSMF [17]. A hollow-core fiber can demonstrate approximately 30% less delay (3.5 ns/m) than a conventional SMF (5 ns/m), offering significant advantages when used in network links in terms of latency reduction.

Transmission-wise, we demonstrate 85 channels in total, all of them with various channel-spacings and baud-rates. The 77 of those channels are sourced by the grating coupler sampled reflection (GCSR) fast-tunable laser which is fully-controlled by the SIC, while the 8 remaining channels come from our

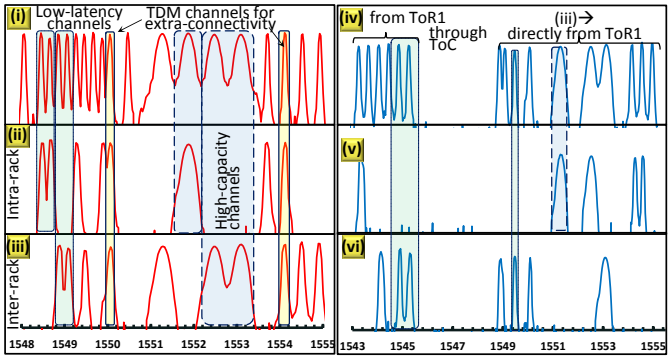

Fig. 10. Spectra plots of low-latency, WDM/TDM & DP-QPSKsignals in different parts of the network directly associated with various interconnection scenarios (as shown in Fig. 9)

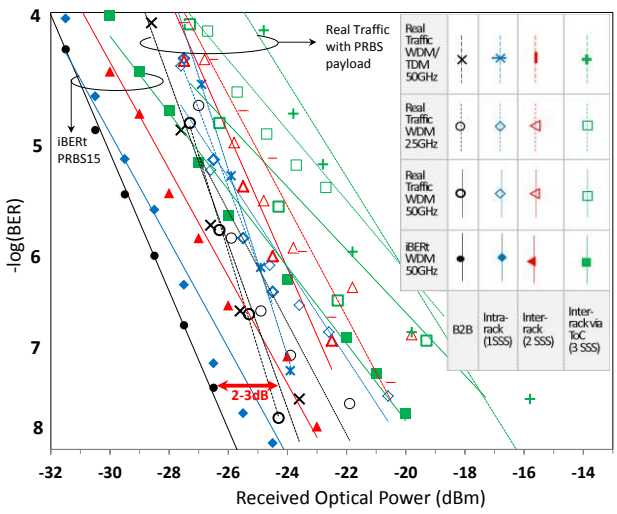

Fig. 11. BER curves for various intra-cluster scenarios with 25/50GHz-spaced WDM & WDM/TDM channels

external-cavity laser (ECL) bank. We place 4 of the ECL channels between the 50GHz-spaced fast-tunable ones, creating a 25GHz channel spacing spectrum slice which then feeds the 10G Mach-Zehnder Modulator (MZM), which is driven again by the SIC. The remaining 4 channels from the ECL bank are modulated in 28Gbaud Quadrature Phase Shift Keying (QPSK) with an IQ modulator and then multiplexed in the polarization domain, leading to 112Gb/s overall capacity per channel. Furthermore, optical circulators are used between the SICs and the SSS-based ToR switches, in order to exploit bi-directionality. The use of circulators aims to take advantage of the N×M dimensionality and bi-directionality of the SSS and save its ports by using one port for both transmission and reception simultaneously. Alternatively, two 1×N SSS per rack would have been needed. It is reasonable that channels need to operate on non-overlapping center frequencies for Tx and Rx. This constraint is adding an element of blocking in very specific cases, which can be resolved with TDM statistical

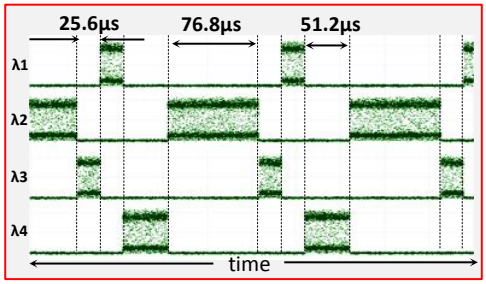

Fig. 12. Time-plots

multiplexing by tuning in another unused wavelength with the fast- tunable lasers.

In the middle transmission stage, signals pass through the circulators, the 10/100m HC-PBGF links, through 1, 2 or 3 SSS, depending on the scenario, and terminate to a receiver SIC as seen in Fig. 9. The maximum capacity-per-blade is 142G (i.e. 3×10G + 1×112G) and per-rack 568G (i.e. 4×blades × 142G). Diverse and variable levels of granularity are supported per-single-carrier and per-blade ranging from 100 Mb/s to 7.8 Gb/s (fast WDM-TDM).

BER is tested with traffic from the traffic analyzer to evaluate board-to-board scenarios. Not only we evaluated the optics separately (with IBERT PRBS15), but also together with electronics (with PRBS payload real-traffic) for the complete demonstration scenario as shown in Fig. 11. In the 25 and 50 GHz-spaced 10G signals, 17dB and 32dB OSNR is observed respectively. A 2-3dB penalty is observed between optics-only and the integrated demonstration scenario (optics & electronics), mostly due to the electronics' inefficiencies. A ≤2dB penalty is obtained for intra-rack interconnection with a further ≤3dB penalty for inter-rack cases, i.e. by passing

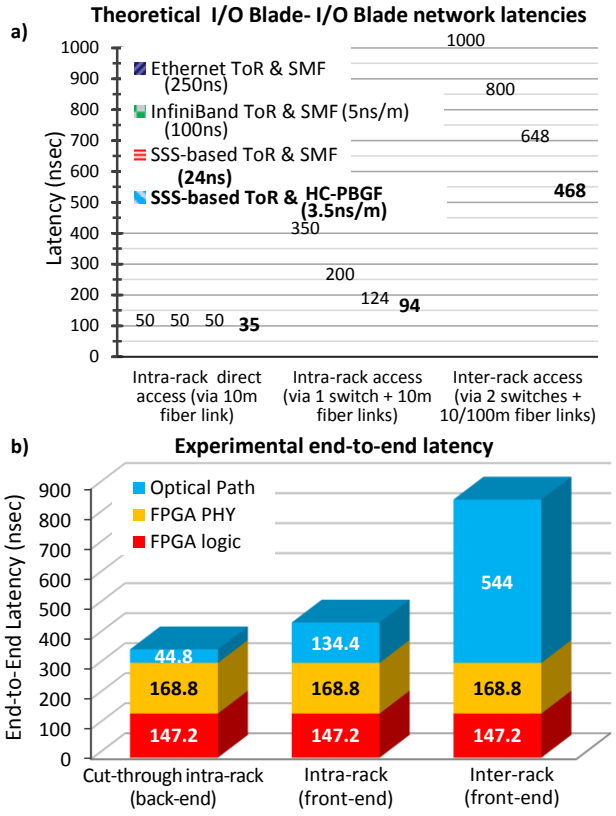

Fig. 13. a) Theoretical network-only latency comparison of inter-connection scenarios utilizing electrical, optical ToR and SMF, HC-PBGF b) Measured blade-to-blade latency break-down – excluding DMA overhead

through two optical SSS-based ToRs or three SSS-based ToR & ToC switches.

By configuring all the dynamic and controlled elements of the SIC in conjunction with the SSS-based optical ToRs and ToC in various frequency-to-port combinations and by also changing between the 4×16/8×12 arrangements, we demonstrate the scenarios of the proposed architecture, as seen in the spectra and time plots of Fig. 10 & Fig. 12, respectively. By looking at both the experimental setup in Fig. 9 and the spectra in Fig. 10, we observe different WDM channels entering ToR1 **(i)** in Rack1, then diverted towards either intra-**(ii)** or inter-rack **(iii)** ports. Then in ToR2 **(iv)**, channels that arrived from ToR1 directly or via ToC, are switched to two separate blades **(v)**, **(vi)** of Rack 2 (or vice versa due to bi-directionality). In the time domain (shown in Fig. 11), λ1-4 start from point **(i)**, are modulated in different time-slots

(25.6/51.2/76.8μs etc.) in order to deliver data in different destinations according to the SSS configuration. λ1 stays intra-rack via ToR1, λ2-3 go directly inter-rack to two different blades of ToR2, whereas λ4 goes to ToR2 via the ToC for congestion avoidance, due to blocking in the direct link between ToR1 and ToR2.

In Fig. 13a, two all-optical (SSS+SMF/HC-PBGF) interconnection schemes are theoretically compared with two electronic (Mellanox Ethernet/Infiniband ToR+SMF) according to their typical values of latency (network only), showing the advantage of our proposed architecture. In Fig. 13b, end-to-end latency is measured from memory chip-to-chip via DMA, including FPGA PHY, logic and optical network delays. The DMA driver latency (approx. 1.5μs) is excluded due to server, OS and driver dependencies, showing total 360.8, 450.4, 860ns latency for cut-through intra-rack, intra- and inter-rack via front-end respectively.

V. Conclusion

This paper reports the FPGA-based optical programmable SIC design, implementation and the application scenario in the disaggregated data centre network. We demonstrated the inter-cluster and intra-cluster disaggregated data centre network architecture with our novel SIC. The SIC is featured with multi-functionality, flexibility, programmability and the ability for supporting multiple techniques like OPS, OCS, TDM and WDM. We setup the testbed with back-to-back interconnects and measured the back-to-back throughput and latency results. We also experimentally demonstrated an all-optical chip-to-chip inter-blade and inter-rack interconnect for highly-connected flexible communication between disaggregated resources (processing/memory/storage) for modern modular Data Centers. By utilizing FPGA-based SICs for real-traffic chip-scale memory access and switch, fast-tunable TDM transceivers, re-configurable flexi-grid optical SSS switches along with HC-PBGF, the proposed network offers variable capacity and granularity (from 100 Mb/s to 158 Gb/s per blade), high-spectral efficiency, high-connectivity (1-to-77 per port) and ultra-low latency interconnection (360.8 ns intra-rack & 860 ns inter-rack), programmable to support diverse and unpredictable Data Center services.

For the future work, we are going to explore the design with 100Gbps ports to achieve higher bandwidth and lower latency. We plan to implant the FPGA-based design to the new generation FPGA-based development board with 16 lanes Gen3 PCIe interface and 100Gbps CFP2/CFP4 interfaces.

References

[1] Y. Yan, Y. Shu, G. M. Saridis, B. R. Rofoee, G. Zervas, D. Simeonidou, "FPGA-based Optical Programmable Switch and Interface Card for Disaggregated OPS/OCS Data Center Networks," in Proc. ECOC2015, Valencia, 2015.

[2] R. Villars, J. Koppy, and K. Quinn "IDC Predictions 2013: The new data centre dynamic," in IDC, Framingham, MA, USA, Dec. 2012.

[3] R. Branch, H. Tjeerdsma, C. Wilson, R. Hurley, S. McConnell, "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.

[4] T. Benason, A. Akella, D. A. Maltz., "Network Traffic Characteristics of Data Center in the Wild" IMC'10, November 1-3, 2010.

[5] "Cisco Global Cloud Index: Forecast and Methodology, 2012–2017," Cisco, 2012.

[6] M. Al-Fares et al., "A scalable, Commodity Data Center Network Architecture", SIGCOMM, pages 63-74, 2008.

[7] S. Han, N. Egi, A. Panda, Sylvia Ratnasamy, G. Shi, S. Shenker, "Network Support for Resource Disaggregation in Next-Generation Datacenters", ACM, 2013.

[8] G. Saridis et al., "DORIOS: Demonstration of an All-Optical Distributed CPU, Memory, Storage Intra DCN Interconnect," Proc. OFC, W1D.2, Los Angeles, 2015.

[9] G. M. Saridis et al., "EVROS: All-Optical Programmable Disaggregated Data Center Interconnect utilizing Hollow-Core Bandgap Fiber", Proc. ECOC, Tu.3.6.5, Valencia, 2015.

[10] 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.

[11] M. P. Kassner, "Disaggregated data centers: great idea, but not just yet," 2015.

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

[13] W. Miao et al., "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system", 2014

[14] B. Rofoee, et al, "Programmable on-chip and off-chip network architecture on demand for flexible optical intra-datacentres", Journal of Optical Express, Vol. 21, 2013.

[15] Y. Yan et al., "FPGA-based Optical Network Function Programmable Node", in Proc. OFC, 2014

[16] B. Guo et al., "SDN-enabled Programmable Optical Packet/Circuit Switched Intra Data Centre Netowrk", in Proc. OFC, 2015

[17] F. Poletti, N. V. Wheeler, M. N. Petrovich, N. Baddela, E. N. Fokoua, J. R. Hayes, D. R. Gray, Z. Li, R. Slavík, and D. J. Richardson, "Towards high-capacity fibre-optic communications at the speed of light in vacuum," Nat Photon, vol. 7, no. 4, pp. 279–284, Apr. 2013