Saridis, G., Yan, Y., Shu, Y., Yan, S. Y., Arslan, M., Bradley, T., Wheeler, N. V., Wong, N. H. L., Poletti, F., Petrovich, M. N., Richardson, D. J., Poole, S. A., Zervas, G., & Simeonidou, D. (2015). EVROS: All-optical programmable disaggregated data centre interconnect utilizing hollow-core bandgap fibre. In *2015 European Conference on Optical Communication (ECOC): Proceedings of a meeting held 27 September - 1 October 2015, Valencia, Spain* [7341960] Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/ECOC.2015.7341960>

Peer reviewed version

Link to published version (if available):

[10.1109/ECOC.2015.7341960](https://doi.org/10.1109/ECOC.2015.7341960)

[Link to publication record in Explore Bristol Research](#)

PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at <http://ieeexplore.ieee.org/document/7341960/?arnumber=7341960>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

EVROS: All-Optical Programmable Disaggregated Data Centre Interconnect Utilizing Hollow-Core Bandgap Fibre

G. M. Saridis⁽¹⁾, Y. Yan⁽¹⁾, Y. Shu⁽¹⁾, S. Yan⁽¹⁾, M. Arslan⁽¹⁾, T. Bradley⁽²⁾, N. V. Wheeler⁽²⁾, N.H.L. Wong⁽²⁾, F. Poletti⁽²⁾, M.N. Petrovich⁽²⁾, D.J. Richardson⁽²⁾, S. Poole⁽³⁾, G. Zervas⁽¹⁾, D. Simeonidou⁽¹⁾

⁽¹⁾ High Performance Networks, University of Bristol, United Kingdom, (George.Saridis@bristol.ac.uk)

⁽²⁾ Optoelectronics Research Centre, University of Southampton, United Kingdom

⁽³⁾ Finisar Australia, Rosebery, Australia

Abstract We present an all-optical flexible disaggregated flat DCN architecture utilizing Hollow-Core Bandgap Fibre, reconfigurable 4x16/8x12 Spectrum Selective Switches and FPGA-based switch and interface intra/inter-blade cards enabling chip-level access while utilizing cut-through low-latency tuneable WDM/WDM-TDM.

Introduction

The growth of cloud services, HPC applications, big data storage/processing along with multi-tenancy, imposes new requirements that challenge current data center networks (DCN)¹. Legacy multi-layer DCNs will have detrimental effects on the distributed and highly-parallelized computing architectures². They lack modularity and flexibility in resource allocation and introduce delays and significant power consumption due to the frequent O/E/O conversions. Optical Top of Rack (ToR) to ToR DCN architectures have been proposed using WDM³⁻⁴ and/or SDM/TDM⁵, showing higher capacity, lower latency and improved scalability. A fully disaggregated DCN with 25/75 local to remote memory balance can be realized with (sub)-μsec latency⁶⁻⁷. Apart from low-latency, a disaggregated architecture should be modular (data center in a box), flexible, programmable and scalable to handle diverse services and requirements.

In this paper, we propose and experimentally demonstrate, for the first time, a programmable, rapidly-tuneable and re-configurable disaggregated flat DCN architecture with all-optical chip-to-chip inter-blade interconnection within and between racks. It delivers ultra-low deterministic latency between blade E/O-I/Os within (44.8–134.4 ns) and between (544 ns) racks, with an additional 316 ns for the FPGA-based electronic blade E/O-I/O from/to memory direct memory access (DMA) driver. Granularity of 100 Mb/s to 142 Gb/s per blade to serve mice and elephant flows is supported together with up to 1-to-77 (# λ and slots) connectivity. Hollow-core photonic bandgap fibre (HC-PBGF)⁸ links offer 30% propagation delay reduction and dimension-programmable single device 20-port NxM spectrum selective switches (SSS) were used for ToR and top of cluster (ToC) switches. An intra/inter-blade programmable FPGA-based Switch & Interface Card (SIC), which performs

Remote DMA and facilitates 2 hybrid 10G OOK fast-tunable ports (to avoid DSP delays) supporting non-disruptive switch-over between fast Ethernet-Lite WDM and WDM/TDM for front-end and 2x10G OOK cut-through back-end ports, allows for pure intra-rack and front-end (hybrid Intra/Inter Rack) networking respectively.

Flexible disaggregated DCN architecture

Fig. 1 shows the proposed architecture with the bidirectional intra-cluster interconnection scheme of various blades/racks containing disaggregated resources for increased system modularity and performance. Two complementary physical networks co-exist, the back-end (intra-rack) and the front-end (intra/inter rack) (Fig. 1 inset A) to support diverse interconnection scenarios. The back-end full-mesh intra-rack FPGA-based network provides all-to-all (zero-to-multi-hop) communication. The front-end re-configurable

Fig. 1: Overall proposed DCN architecture (bottom) Inset A: Back- & Front-end description. Inset B: FPGA-based SIC. Inset C: Blade-ToR interconnection

intra/inter-rack network provides one, two or three hop intra-cluster communication. The optical SSS-based ToR switches, which are bandwidth-to-port selection-flexible and able to dynamically reconfigure their I/O dimensions (e.g. 4×16 , 8×12 , 10×10), groom, switch and balance traffic within and between racks. This unique feature enables the architecture to adapt to different networking scenarios with diverse requirements on capacity, latency and connectivity. The identical SSS-based ToC handles and re-balances intra/inter-cluster communication and aggregates traffic towards a high-port count fibre switch, adding scalability to the network. The blade-to-ToR interface is implemented with 4 channels (Fig.1 inset C) and circulators to exploit the bi-directionality of the SSS, thus saving resources. By utilizing 3x fast tuneable 10 Gb/s OOK channels for either low-latency WDM or WDM/TDM and a 28 Gbaud DP-QPSK channel on each SIC, wide (C-band), flexible-bandwidth connectivity with various granularities is achieved.

The highly programmable FPGA-based SIC (Fig.1 inset B) is developed to read/write data blocks via a PCIe Gen2 bus between the CPU memory and remote DRAM chip. Then accordingly, traffic is switched either via the cut-through optically-connected electrically-switched back-end or via the programmable fast-tuneable dual-mode (cut-through WDM/WDM-TDM) to reach remote destinations. Apart from interfacing, SIC can also hitlessly switch traffic from the back-end towards the front-end and vice versa, allowing multi-hop communication to assist congested links. Furthermore, SIC is able to flexibly aggregate TDM traffic based on

network requirements (QoS, latency, connectivity) by dynamically programming the size/number of slots per frame.

Experimental setup, results and evaluation

The experimental test-bed (Fig. 3), contains two FPGA-based SIC boards that support real-traffic scenarios with fully-functional transceivers that act as the interface between the local memory chip and the back/front-end network, three re-configurable $4 \times 16/8 \times 12$ SSSs, a traffic analyser and nine spools of HC-PBGF (6×10 m for intra-rack and 3×100 m for inter-rack). We demonstrate 85 channels in total (77 fast-tuneable + 8 ECL) with variable channel-spacing and baud-rates (Fig. 2 top-left). The maximum capacity-per-blade is 142G (i.e. $3 \times 10G + 1 \times 112G$) and per-rack 568G (i.e. $4 \times \text{blades} \times 142G$). 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). Transmission-wise, 8 channels (from the ECL bank) are placed between the 50 GHz-spaced GCSR fast-tuneable channels, creating a 25 GHz channel spacing spectral slice which feeds the 10G MZM modulator, driven by the FPGA-based SIC. In the middle stage, signals pass through circulators, 10/100 m HC-PBGF links, 1/2/3 SSS and terminate to another FPGA as seen in Fig. 2. The BER is tested with real traffic (with PRBS payload) to evaluate board-to-board scenarios (optics & electronics) in Fig. 3a. In the 25(50) GHz-spaced 10G signals, 17(32) dB OSNR is observed and ≤ 2 dB penalty was obtained for intra-rack interconnection with a further ≤ 3 dB penalty for the two and three-hop inter-rack cases, respectively.

Fig. 2: Experimental setup, HC-PBGF characteristics, spectra at key points of the network, time plots