dReDBox: Demonstrating Disaggregated Memory in an Optical Data Centre

A. Saljoghei[1], V. Mishra[1], M. Bielski[2], I. Syrigos[3], K. Katrinis[4], D. Syrivelis[4], A. Reale[4], D. N. Pnevmatikatos[5], D. Theodoropoulos[5], M. Enrico[6], N. Parsons[6], G. Zervas[1]

[1] University College London (UCL), United Kingdom. [2]Virtual Open Systems, France. [3]University of Thessaly, Greece. [4]IBM Research Ireland, Ireland. [5]Foundation of Research and Technology Hellas, Greece .[6]Huber+Suhner Polatis, United Kingdom.

a.saljoghei@ucl.ac.uk

Abstract: This paper showcases the first experimental demonstration of disaggregated memory using the dReDBox optical Data Centre architecture. Experimental results demonstrate the 4-tier network scalability and performance of the system at the physical and application layer.

OCIS codes: 060.4253 Networks, circuit-switched; (060.4258) Network topology; (200.4650) Optical interconnects

1. Introduction

Today's conventional data centre (DC) architectures follow a server-centric model, where each server contains fixed amounts of memory and processing power. However, this model is bound to achieve low resource utilisation due to the 4 orders of magnitude disproportionality in demand for CPU over memory resources [1]. This leads to a fragmentation in resource allocation and in turn an increase in the total cost of ownership (TCO) and a reduction in energy efficiency. The TCO will further increase in the server-centric model since the replacement/refresh cycle of each component (e.g. memory, CPU), is locked to that of a whole server.

Disaggregated DCs have been proposed to tackle the shortcomings associated with the server-centric model. The vision of disaggregation creates the new block-as-a-unit paradigm, or the resource-centric model, where a pool of interconnected resources containing a different number of CPU, storage, memory or accelerator units can be dynamically assigned to the incoming service requests. However, the interconnect (on-chip and chip-to-chip) between the various resources residing either within a rack or neighbouring racks needs to achieve the lowest possible latency while being able to support a substantially higher bandwidth at lower cost and power consumption levels. [2] reported IT disaggregation using optical networks through provided communication between direct-attached memories (via CPU) on conventional servers and not disaggregated components while using switch and interface card via PCIe that led to microsecond-scale latencies. To address these challenges we have proposed the disaggregated recursive datacentre-in-a-box (dReDBbox) architecture [3]. In previous studies, it has been shown that the dReDBbox architecture can improve resource utilisation by up to 33% and reduce power consumption by up to 47 % compared to the server-centric models [4].

In this paper, we report on the experimental testbed used in this work, and we provide results from the physical layer, layer 2 and industry standard STREAM memory benchmark. We demonstrate remote memory attachment [5] using the customised operating system and for first time remote memory access over a DC network using on-chip (processor, and memory) switching as well as chip-to-chip optical circuit switching that scales up to 4-tier network.

2. dReDBbox Architecture

The dReDBbox architecture aims to a) bound the round-trip latency between resources to hunderds of nanoseconds, b) deliver up to 400 Gb/s network interconnect capacity for individual disaggregated resources, c) maximise resource utilisation, d) achieve scalability. Also it explores the deployment of accelerated and reconfigurable protocol/function ports on each disaggregated resources. The disaggregated rack in the dReDBbox architecture (dRACK) is shown in Fig. 1 (a). Each dRACK houses 16 rack mounted 2U units (dBOXes), which also contain an electronic cross-point circuit switch, a set of optical circuit switches (dBOSM) and 16 pluggable dBRICKs (individual disaggregated resources). Each dBRICK can support general-purpose processing (dCOMPUBRICK), random-access memory (dMEMBRICK) or application-specific acceleration (dACCELBRICK). All dBRICKs on the same dBOX are interconnected by an electronic cross-connect switch (dBESM) and three optical circuit switches (dBOSM). However, the dBRICKs on different dBOXs are connected to one another using a second tier optical circuit switches (dROSM) positioned at the middle of the rack and the dBOSM. Fig. 1 (b) presents the vertical dReDBbox software-defined architecture that will enable the control and orchestration of the dReDBbox architecture [5]. The system software and the software-defined control will allow for allocation of disaggregated resources to individual dBRICKs by controlling networking elements on individual bricks. The orchestration software features a database of all available resources and based on the VM requests from the application layer it uses the software defined memory (SDM) controller to allocate appropriate resources to

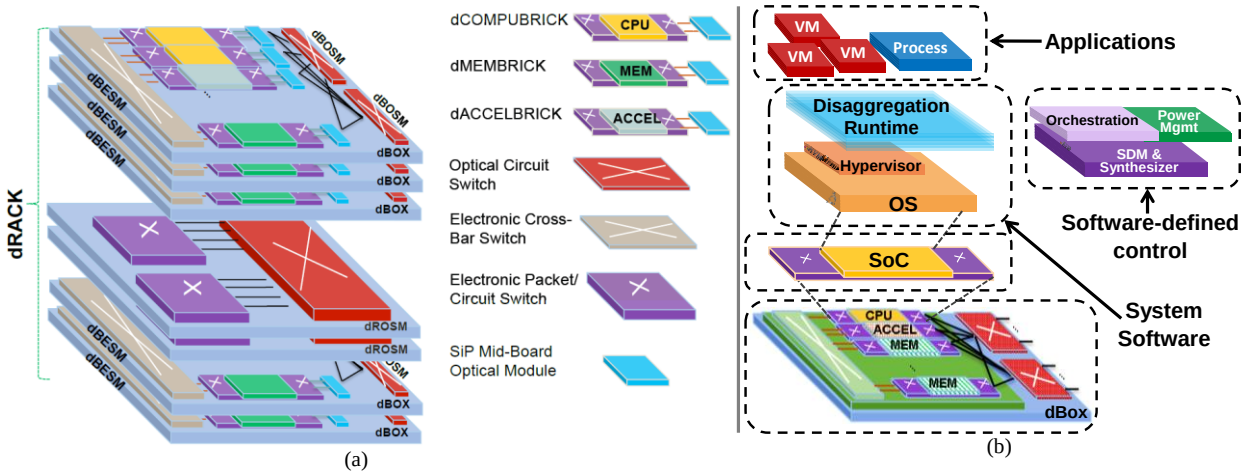


Fig. 1. (a) dRack architecture of dRedDbox, (b) vertical software-defined dRedDbox architecture

hypervisors running on the individual system on chips (SoC). The OS running on dCOMPUBRICK is responsible for the management of the virtual machine (VM) that will be deployed on the disaggregated resources.

The individual dBRICKs are implemented using commercial Multi-Processor System on Chip (MPSoC) based hardware solution. Each brick also features a reconfigurable system on chip to perform networking functions beyond interfacing as traditional network interfaces do. This allows each brick to support forwarding, switching and aggregation at either the packet or circuit level [6]. To minimise footprint and power consumption as well as to maximise bandwidth density and front panel density each dBRICK uses mid-board optics (MBOs) based on SiP technology with a capacity of up to 200 Gb/s. By considering the electrical ports connected to the dBESM switch, each dBRICK can achieve the 400 Gb/s network interconnection capacity.

3. System setup

The experimental setup that showcases disaggregated memory using the dReDBox architecture is shown in Fig. 2. Each resource brick is implemented on a Xilinx ZYNQ Ultrascale+ FPGA. The compute resources in the dCOMPUBRICK are represented by a 4-core ARM processor and the disaggregated memory resources in the dMEMBRICK are represented by a 256 MB DDR4 module. The glue logic (GL) in each brick translates the physical memory addresses (seen by Linux OS) to remote physical memory and maps the disaggregated memory resources in network encapsulated outgoing transcactions. Once memory resources are identified by the glue logic, the resulting read/write memory requests and data transactions are sent to a network on chip (NoC) element, which forwards them (at either packet or circuit switched level) to the appropriate physical port. Each of these physical incoming/outgoing ports for each PHY block on the FPGAs is attached to a different channel on the multi-channel SiP MBO. For optical circuit switching a 48-port optical switch is used, however, this switch is logically split to realise multi-tier network topology (Fig. 2). In the dReDBbox architecture, dCOMPUBRICK runs a customised Linux based operating system (OS). This OS is capable of (a) running the hypervisor to virtualise the available resource, (b) hot-plugging the remote memory resources and (c) managing and instructing the glue logic through a custom driver.

The SiP MBO used in this work has a total of 8 transceivers using external modulation and a shared laser operating at 1310 nm. Each channel on average has an optical output power of -3.7 dBm and uses OOK modulation. These individual channels can operate at up to 25 Gb/s, although, in this work given the capacity of the FPGAs we limit the operation of each channel to 10 Gb/s. The MBO is connected to a 48-port optical switch module. Each hop through the optical switch module accounts for approximately 1 dB of attenuation (Fig. 3(a)), and the optical switch module has an approximate power consumption rating of 100 mW/port.

4. Results & Discussions

A key requirement of such disaggregated architectures is the need for a FEC free interface between various

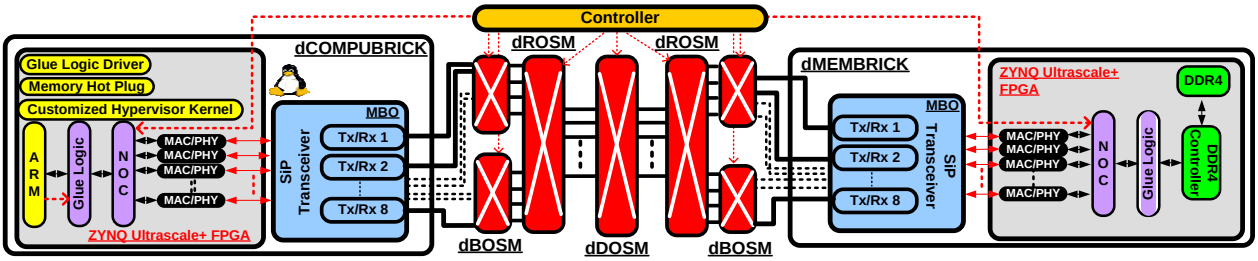
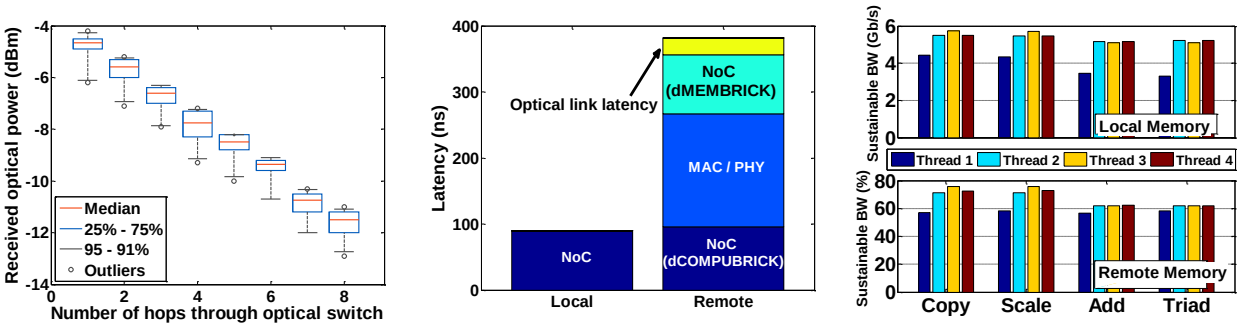


Fig. 2. Experimental system. (dDOSM: top of clustter optical switch, used to connect dBRICKs on different dRACKs)



(a) (b) (c)

Fig. 3. (a)Received optical powers from 10 random optical links after traversing multiple switching hops (BER $1x10^{-12}$ achieved), (b) One-way system latency breakdown for remote or local memory access, (c) sustainable bandwidth for various operations for the local memory in terms of Gb/s, and in terms of % for the local memory (Ratio: Remote sustainable BW / Local sustainable BW).

resources, as the FEC encoder/decoders can potentially introduce >100s ns of latency, which impacts the performance of the disaggregated system. All bi-directional optical links between the dCOMPUBRICK and dMEMBRICK (Fig. 2) are able to achieve a bit error rate (BER) below 10^{-12} (96% confidence level) while all but one were traversing eight hops through the optical switch, and the remaining channel traversed six hops. Nevertheless, by optimising the system further in terms of equalisation and drive levels, it may also be possible to achieve eight hops through the remaining channel. This demonstrates the ability of this system to scale into a 4-tier architecture. The box plot in Fig. 3 (a) presents the distribution of received optical powers from five random 10 Gb/s bi-directional optical channels between the dCOMPUBRICK and dMEMBRICK after traversing multiple hops through the optical switch (BER below 10^{-12} reached), as it can be seen on average 1 dB of loss is experienced per hop. Fig. 3 (b) provides the breakdown of the latencies encountered in the end-to-end system shown in Fig. 2 for accessing memory resources either locally or remotely. To access local memory resources, the on-chip switch (NoC) is used, and it accounts for 85 ns of one-way latency. By accessing the remote memory, the one-way latency of 356 ns is experimentally measured. This latency refers to contributions of the NoC and the MAC/PHY block on both the dMEMBRICK and the dCOMPUBRICK as well as the optical path propagation delay.

To verify the disaggregated memory access, the GL and the NoC on the dCOMPUBRICK and the optical switch are configured to establish a link with the dMEMBRICK by using a single channel of the MBOs. Next, by using the customised Linux OS distribution loaded onto the dCOMPUBRICK, a total of 256 MB of remote PL-DDR memory on the disaggregated dMEMBRICK is attached to the system. To measure the overhead added by the optical interconnect, we also built a loopback configuration that uses the same GL and NoC to attach 256 MB local PL-DDR to the system. We compare the perceived sustainable application-level memory bandwidth in the two configurations by running the STREAM benchmark [5] and varying the number of cores simultaneously accessing the attached memory. Fig. 3 (c) shows the results: on average a maximum throughput of 5 Gb/s is achieved by the local memory, while remote memory sustains ~68% of that bandwidth due to inferred interconnect latencies.

6. Conclusions

This paper presented the disaggregated dReDBbox architecture. Experimental results showcase the disaggregation of memory over a low latency optical network. It was shown that optical interconnects employed in this architecture can achieve a FEC free operation for a 4-tier topology. Memory disaggregation over an optical network was demonstrated with a throughput of 4 Gb/s and end-to-end network latency of 356 nsec. Compared to the local memory, the disagregated memory access due to the end to end network latency sustains a 68% of throughput.

7. Acknowledgment: This work is supported by the EC H2020 dReDBbox project with grant agreement 687632. We would also like to thank Luxtera for SiP engines and fruitful discussions.

8. References

[1] S. Han, N. Egi, A. Panda, S. Ratnasamy, G. Shi, and S. Shenker, "Network support for resource disaggregation in next-generation datacenters," presented at the Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks, College Park, Maryland, 2013.
[2] Y. Yan and e. al., "All-Optical Programmable Disaggregated Data Centre Network Realized by FPGA-Based Switch and Interface Card," *Journal of Lightwave Technology*, vol. 34, 2016.
[3] K. Katrinis, D. Syrivelis, D. Pnevmatikatos, G. Zervas, D. Theodoropoulos, I. Koutsopoulos, *et al.*, "Rack-scale disaggregated cloud data centers: The dReDBbox project vision," in *Design, Automation & Test in Europe*, 2016.
[4] G. Zervas, F. Jiang, Q. Chen, V. Mishra, H. Yuan, K. Katrinis, *et al.*, "Disaggregated compute, memory and network systems: A new era for optical data centre architectures," in *Optical Fiber Communications Conference and Exhibition (OFC)*, 2017, pp. 471-474.
[5] D. Syrivelis, A. Reale, K. Katrinis, I. Syrigos, M. Bielski, D. Theodoropoulos, *et al.*, "A Software-defined Architecture and Prototype for Disaggregated Memory Rack Scale Systems," presented at the samos-conference, 2017.
[6] Q. Chen, V. Mishra, and G. Zervas, "Reconfigurable computing for network function virtualization: A protocol independent switch," presented at the International Conference on ReConFigurable Computing and FPGAs (ReConFig), 2016.