MCF-SMF hybrid low-latency circuit switched optical network for disaggregated data centers

Arsalan Saljoghei, Hui Yuan, Vaibhawa Mishra, Michael Enrico, Nick Parsons, Craig Kochis, P. De Dobbelaere, Dimitris Theodoropoulos, Dionisios Pnevmatikatos, Dimitris Syrivelis, Andrea Reale, Tetsuya Hayashi, Tetsuya Nakanishi, Georgios Zervas

***Abstract*—This paper proposes and experimentally evaluates a fully developed novel architecture with purpose built low latency communication protocols for next generation disaggregated data centers (DDCs). In order to accommodate for capacity and latency needs of disaggregated IT elements (i.e. CPU, memory), this architecture makes use of a low latency and high capacity circuit switched optical network for interconnecting various end-points, that are equipped with multi-channel Silicon photonic based integrated transceivers. In a move to further decrease the perceived latency between various disaggregated IT elements, this paper proposes a) a novel network topology, which cuts down the latency over the optical network by 34% while enhancing system scalability and b) channel bonding over multi-core fiber (MCF) switched links to reduce head to tail latency and in turn increase sustained memory bandwidth for disaggregated remote memory. Furthermore, to reduce power consumption and enhance space efficiency, the integration of novel multi core fiber (MCF) based transceivers, fibers and optical switches are proposed and experimentally validated at the physical layer for this topology. It is shown that the integration of MCF based subsystems in this topology can bring about an improvement in energy efficiency of the optical switching layer which is above 60%. Finally, the performance of this proposed architecture and topology is evaluated experimentally at the application layer where the perceived memory throughput for accessing remote and local resources is measured and compared using electrical circuit and packet switching. The results also highlight a multi fold increase in application perceived memory throughput over the proposed DDC topology by utilization and bonding of multiple optical channels to interconnect disaggregated IT elements that can be carried over MCF links.**

***Index Terms*—Data center networks, Multi Core Fiber, Network topology, disaggregated data center, optical interconnects, optical circuit switching**

I. Introduction

Today's data center networks (DCNs) follow a server-centric approach whereby the available resources per servers are fixed and limited to the boundaries of the mainboard tray. It has previously been shown that the ratio of demand for resources such as storage and memory to CPU could span over three orders of magnitude given the wide array of tasks which could be arriving at a typical Google data center (DC) [1]. This disproportionality between the demands for different resources in the current DCNs can lead to significant underutilization of available resources, as their allocation is upper bounded by the resources available within the boundary of the mainboard. This issue can cause spare resource fragmentation and inefficiencies which accounts for 85% of total DCN costs [2]. The other factor which also impacts the total cost of ownership in modern operational DCN's since technological upgrades need to be made to every server, even if only a specific component needs to be replaced.

Thankfully, such shortcomings can be mitigated by migrating towards disaggregated data center (DDC) architectures, which would follow resource-centric properties. The deployment of such architectures entails 1) the defragmentation and disaggregation of the IT resources (compute, memory, storage and accelerators) 2) finely interconnecting these defragmented resources by a well-interconnected scalable network. This approach can lead to elevated resources utilization of up to 34% and power savings around 40% [3]. These advantages have spearheaded R&D into DCN disaggregation in both industrial and academic circles. This can be attested to by work carried in Intel Rack Scale architecture [4], Open Compute project [5] and HPE Moonshot / Machine project [6, 7]. In recent years' substantial work has so far been carried in defragmenting long term storage elements within DCNs using either storage area networks or network attached storage [8], requiring peak data rates up to 6-32 Gb/s and response latencies between 10 to 50 µs [9, 10]. The employment of disaggregated accelerator elements in DCN can lead to a significant boost in computation power for tasks such as networks analytics, deep learning or encryption. However, substantial progress towards memory disaggregation has not yet been fully materialized [11].

Despite the advantages inherent to disaggregated systems,

This work was supported by EC H2020 dReDBox project with grant agreement 687632.

A. Saljoghei, H. Yuan, V. Mishra, G. Zervas are with University College London, London, UK.

D. Syrivelis, A. Reale are with IBM research Ireland, Dublin, Ireland.

M. Enrico, N. Parsons are with Huber+Suhner Polatis, Cambridge, UK.

C. Kochis, P. D. Dobbelaere are with Luxtera Inc., California, USA.

D. Theodoropoulos, D. Pnevmatikatos are with FORTH, Greece.

T. Hayashi and T. Nakanishi are with Optical Communications Laboratory, Sumitomo Electric Industries, Ltd., Yokohama 244-8588, Japan (e-mail: t-hayashi@sei.co.jp, nakanishi-tetsuya@sei.co.jp).

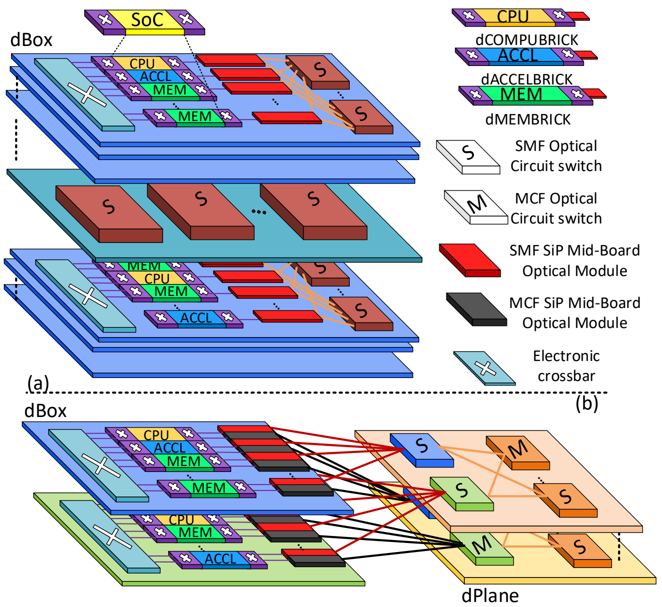

dBox
SoC
CPU
ACCL
MEM
S
M
dCOMPUBRICK
dACCELBRICK
dMEMBRICK
SMF Optical Circuit switch
MCF Optical Circuit switch
SMF SiP Mid-Board Optical Module
MCF SiP Mid-Board Optical Module
Electronic crossbar
(a)
(b)
dPlane

Fig. 1. (a) Rack scale architecture used in dReDBox which interconnects disaggregated Compute, Memory and Accelerator blocks via optically switched links. (b) dReDBox rack structure using hybrid SMF/MCF parallel topology

such architectures would require adhering to several system constraints. These are namely 1) lower latency levels, 2) higher link capacities at lower costs and 3) lower power and space consumption level. To address these challenges, this paper proposes and showcases a novel and fully developed resource centric architecture for disaggregated data centers (DDCs) called the disaggregated recursive data-center-in-a-box (dReDBox) [12]. This architecture allows all IT elements in the topology to act as standalone entities with dynamic on chip packet/circuit switching capabilities, which can independently communicate with one another through a high capacity and low latency circuit switched optical network. Furthermore, to enhance power efficiency and reduce system latency, this paper examines and proposes novel low latency/parallelized and highly modular topology making use of spatial division multiplexing (SDM) through multi core fiber (MCF) based subsystems for DDCs using the dReDBox architecture, in order to accommodate for simultaneous flexibility and higher switching densities. The proposed topology in conjunction to the dReDBox architecture and on-chip logic to offer channel bonding is evaluated experimentally at application as well as the physical layer, moreover, the architecture and the topology are also examined at the network layer using computer simulations. The simulation signifies the benefits of the proposed topology for various virtual machine (VM) demands showcasing power and space benefits. Moreover, the proposed topology and architecture are examined at the physical layer to ensure its feasibility. The developed channel bonding logic which had been implemented on the dReDBox architecture is also examined at the application layer using the proposed hybrid where application perceived memory throughput is assessed for disaggregated memory nodes. The proposed topology allows for 34% latency reduction compared to a three-tier tree topology and in excess of 68-81% power consumption at the switching layer compared to a system only employing SMF based switches.

In DDCs, the accessing of disaggregated remote memory resources as opposed to storage and accelerator elements [9, 10] has the highest demand in terms of latency (10s of nanoseconds) and required link bandwidths (100s of Gb/s) [8] since the perceived memory throughput at the application layer is highly affected by these factors. Unfortunately, today's DCNs are unable to meet these demands [8, 11]. To assess the ability of the dReDBox architecture with the proposed topology for meeting these demands, the application perceived performance for accessing remote DDR4 memory resources is experimentally measured; the results suggest that the proposed architecture can sustain 70% of memory throughput when accessing remote memory. Crucially application perceived memory throughput is analyzed for four different operations (copy, scale, add, triad) following the STREAMS benchmark across local, remote and hybrid (local and remote) access when using single or bonded optical communication channel.

The rest of this paper is organized as follow: section II provides a brief overview of the key enabling elements in the dReDBox architecture, section III describes the characteristics required by the optical subsystems in DDCNs, section IV proposes and describes a novel low latency hybrid topology for DDCNs which can decrease the overall power consumption, section V gives a thorough overview of the experimental system used, and finally section VI presents the experimental results obtained at both the physical and application layers on the proposed architecture and topology.

II. dReDBox Architecture

The dReDBox architecture [12] aims at provisioning low latency and high bandwidth links between various disaggregated resources. To achieve lower latency and power consumption, in the dReDBox architecture, FPGAs are embedded with individual IT elements by using a custom designed pluggable card. This enables, each end-point to be equipped with programmable on chip electrical packet/circuit switching capabilities, which eliminates the need for network interface cards. In this architecture, each of this pluggable card is called a disaggregated brick (dBrick). Next, in order to meet the required memory bandwidths at a cost and energy efficient manner, the I/O functionalities for each individual disaggregated element is carried by high bandwidth multi transceiver Silicon photonic mid board optics (MBO). Thus, optical circuit switches will be used to interconnect these disaggregated elements, which further aids the system latency. Furthermore, given the high latencies associated with forward error correction (FEC) schemes that can be of 100s of nanoseconds, the dReDBox architecture aims at restricting their use. In order to guarantee error free performance over extended periods of time (days), 48 specific control parameters inherent to optoelectronic and electronic transceivers are required to be fine-tuned (further details are provided in section V-B). the architecture makes use of adequate electrical and optical transceiver optimization.

Fig. 1 (a) represents the architecture of a typical rack or dRack (disaggregated racks) designed for the dReDBox topology. Individual server's blades referred as dBoxes in the dReDBox architecture can house a various combination of pluggable IT resources or dBricks. Furthermore, dBricks have been classified based on the type of resources such as computing referred to as dCOMPUBRICK, memory as dMEMBRICK and acceleration (dACCELBRICK). Each dBox uses a high port count electronic cross connect switch for supporting a reconfigurable switching between various dBricks residing in the same dBox. To achieve all other intra-dBox (transaction within a server blade) and inter-dBox (transaction between a dBrick in one server blade with other dBricks in other server blades), various distributed dBricks are interconnected using a three tier topology on the original architecture [13]. Each individual dBrick can be utilized as various IT resources by using technologically advanced FPGAs which combine both multi-core processors and configurable logic on same die. Each dBox can house a different ratio of various resources and if required they can be made up of only a single IT resource given the system requirements.

To interconnect the dBricks within a common dBox and towards remote dBoxes, a set of low port count switches called disaggregated box optical switch modules (dBOSMs) are also employed. This revised dReDBox architecture avoids using top of rack switches since the longer optical paths between the dBoxes in the bottom of the rack to the top of the rack can lead to heightened levels of unwanted latencies. Thus, as it can be seen in Fig. 1 (a), the second-tier switches are moved to the middle of the rack. These second-tier optical switches are called disaggregated rack optical switch modules (dROSM).

III. Optical Interconnect and switching technologies

It has been envisioned that by 2030 semiconductor chips will have I/Os which will need to support capacities beyond 1 Pb/s [14]. Moreover, memory architectures such as hybrid memory cube and high bandwidth memory which are key technologies for disaggregation are currently capable of achieving multi-Tb/s bandwidths [14]. Considering these aspects, it becomes apparent that optical transport and switching technologies are the best means of meeting the current and on ongoing bandwidth and latency demands in disaggregated DCNs.

Even though 100GbE technologies are readily available and 400GbE technologies are set to enter the market to sustain the growth in bandwidth for current DCN topologies, these rates still lag behind the Tb/s or Pb/s rates required. Moreover, data centers are envisioned to still heavily rely on 10GbE technologies [15] despite the availability of 40GbE/100GbE/400GbE transceivers due to the high costs associated with these higher capacity technologies. Thus, it is of utmost importance that optical interconnects in DDCs be designed to achieve high capacities while promoting low latency and cost efficiency.

A. *Multi Core vs single mode*

SDM based links for today's DCN are not a new concept as they are being widely employed in the form of SMF ribbons. Today, fiber ribbons are made up of few to hundreds of individual SMF fibers stacked into a common link [16]. An example of commercial transceivers for DCNs using such ribbons are the quad small form-factor pluggable transceivers operating at 40/100G employing a separate SMF links per each transmitter or receiver block of each of the four optical channels in the transceiver.

Employing SDM in such transceivers compared to WDM or dense-WDM can clearly result in a significant reduction in complexity, since individual optical transmitters would not require tighter wavelength control, additional parts such as MUX and DEMUX with their associated link impairments such as insertion loss. Moreover, optical switching in the spatial domain compared to the frequency domain allows for a flexible and low complexity routing scheme, since each granular link can be independently switched.

Fiber ribbons can also lead to a decrease in spatial density, which is due to the increase in space taken up by the collection of fibers interconnecting servers and racks. Moreover, an increase in the number of individual fiber connections to a transceiver translates into the need for a larger space required for housing fiber connectors which in fact is limited to the front panel area of a typical 2U rack mount chassis. Furthermore, an increase in the number of individual fibers in DCNs can also lead to a significant increase in power consumption given that multiple links have to traverse the same path need to be switched individually. To remedy the shortcoming associated with fiber ribbons, the use of MCFs is promising in the context of DCNs. Nevertheless, despite the advantages which can be gained from MCFs, they suffer from inter-core cross talk which can affect the optical to signal noise ratio of the transmission system and place an upper limit on the number of spatial channels which can be employed per individual fiber strand [17]. Moreover, given the small pitch core of these fibers, interfacing these elements to various active and passive optical components may prove difficult [16].

B. *Mid Board Optics*

In a move towards replacing copper-based interconnects and exploiting optical printed circuit boards [18] for interconnecting various on board IT elements, board-detachable optical transceivers in the form of MBOs are seen as an attractive solution. These MBOs look to be set to replace front face pluggable transceivers [18]. This type of transceiver enables better utilization of front panel space, it leads to a better dissipation of heat and it allows for deployment of the optical transceiver ever closer to the IT element hence reducing the electrical interfaces between them which will minimize latency and electrical signal parasitic. Given that MBOs are formed by the Integration of multiple transceivers for achieving high aggregate rates, they are able to achieve a significant level of bandwidth density and energy efficiency (see Table I in [8]). Recent demonstrations [19] have shown a MBO with 168 integrated optical transceivers each operating at 8 Gb/s, accounting for a net transmission rate of 1.34 Tb/s (FEC-Free), bandwidth density of 64 Gb/s/mm^2 and energy efficiency of 10 pJ/bit. A major advantage associated with MBOs is that~~,~~ the net transponder data rate can be parallelized over multiple integrated transceivers running at lower rates. Given the bandwidth limitation associated with opto-electronics, this lower rate transmission can ensure better signal integrity and FEC free operation. Considering these achievements, it is becoming obvious that MBOs can play a significant role in DDCNs. This is provided by their ability to achieve the link capacities required by next generation systems while maximizing bandwidth density and energy efficiency while simultaneously allowing for FEC less operation [19, 20]. So far the majority of MBO demonstrations have relied on either WDM or spatial multiplexing of fiber ribbons (see Table I in [8]). Given the benefits which can be obtained with MCF based DCNs in terms of network utilization and costs, it's understandable that the employment of MCFs in conjunction with these MBOs can be seen as an added advantage.

C. *Optical switching*

The use of optics for interconnecting various elements in a DDC, pushes towards the employment of optical means of switching within the network. Such switches need to accommodate for short reconfiguration time to ensure lower levels of loss of service. Furthermore, they need to offer low latency levels to ensure adequate performance. Compared to conventional electrical switches employed in DCs, optical switches can ensure lower levels of power consumption [21] as well as modulation format/bandwidth transparent characteristics.

As it was found in previous sections, parallelizing the net payload between various IT elements in disaggregated DCs over N optical channels is complemented by the employment of multi transceiver MBOs which can allow for FEC free transmission and promote spatial as well as energy efficiency. However, by considering that optical switches will be used to interconnect the two IT elements in such a network, it can be clearly seen given that the total link payload is parallelized. In this system, each switching element needs to accommodate for Nx4 ports (considering the input/output as-well-as the transmitter and receiver ports). Given the parallelization, some IT elements would require multiple optical channels to be set in between them. In this scenario, it is clear that the use of MCF fibers along with a low loss switch fabric capable of switching all cores in a MCF simultaneously can reduce the power consumption by a factor of N which would be a significant benefit to DDCs.

IV. Network Topology for dReDBox

Given the latency sensitive nature of disaggregated DCN systems, the dReDBox architecture reduces the overall network latency by introducing custom designed communication protocols. Nevertheless, network topologies can also affect the latencies experienced over links interconnecting disaggregated elements. Thus, forming a network topology tailored for disaggregation allowing for a reduction in latency while favouring system scalability and cost are advantageous. Moreover, following the discussion in the previous section, these types of topologies would also need to rely on MCF technologies, thus they require to have the means to support MCF switches. Despite the advantages of having a fully MCF based OCS-DCN architecture [22], the granularity enforced on the I/Os of each dBrick in the dReDBox topology due to the employment of MBOs can lead to significant underutilization of fiber capacity introduced by a SDM rich infrastructure. This is because, a particular dBrick communicating with another may require varying levels of link capacity which can either fully (all cores) or partially (some of cores) utilize the whole capacity of a MCF link given the limited core by core configurability. In case of partial fulfilment, this can be also translated into an inefficient utilization of resources. To account for this, and increase re-configurability in the system, SMF based transceivers and OCS should also be used in similar infrastructures.

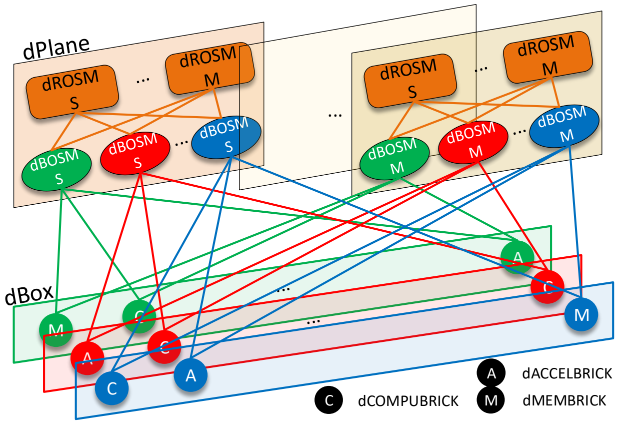


Fig. 2. Proposed low latency hybrid MCF/SMF topology employing circuit switching. S: SMF, M: MCF

A. *Low latency hybrid/parallel topology*

In [23] we showcased a novel highly scalable two-tier network topology for DDCNs which allowed for a 34% reduction in overall system latency compared to three tier tree based topology. In this section, the topology previously proposed is tailored for the dReDBox architecture relying on Hybrid SMF/MCF, Fig. 2 shows the overall architecture of this topology. As it can be seen, it follows a general spine-leaf topology to interconnect all dBricks in every individual dBox in the network, however with some modifications. In order to reduce the size of the port count on each deployed OCS in the 1st and 2nd tiers, instead of having a fully meshed network of spine-leaf switches, these are grouped into a parallel collection of non-conjoined entities (dPlanes).

In this topology, the total number of dPlanes is equal to the total number of individual bi-directional fiber channels per dBrick. The number of dBOSMs per dPlane equals the total number of dBoxes in the network, and the number of dROSM switches per dPlane is half of the number of dBOSMs per plan for a 1:1 subscription ratio. This topology has a high level of scalability favouring east-west communications since the number of dBoxes in the network can be incremented just by increasing the number of dBOSM and dROSM switches in each plane accordingly. This parallelization of switching elements is enabled as a result of employing multi transceiver MBOs, which necessitates full connectivity between all dBricks in this parallel network where each optical I/O of the MBO is routed towards one individual dPlane.

In order to accommodate hybrid SMF/MCF subsystems in this topology, each dBrick will need to be equipped with both SMF and MCF based MBOs. The ratio of total transceivers accounted for by MCF or SMF links, is a pure design choice, however, once this is decided upon, all dBricks need to adopt the same ratio. If M individual SMF and K individual MCF links exist at one dBrick. The topology will need to have M dPlanes only with SMF and K dPlanes with only MCF dBOMSs. However, provided the level of configurability required within the system, dROSMs within a dPlane can have a mix of MCF or SMF OCSs.

To accommodate for this topology the dBox structure initially proposed for the dReDBox architecture (as shown in Fig. 1 (a)) needs to be altered. Fig. 1 (b) represents the new dBox physical structure along with its connectivity with the dPlanes.

B. Power consumption

Results in Fig. 3 showcase the reduction in power consumptions achievable by the proposed hybrid low latency SMF/MCF topology compared to a fully SMF based topology. For the calculations used, a DDCN with 128 dBoxes is envisioned where each dBox is equipped with 16 dBrick. Furthermore, each dBrick is equipped with 16 individual optical channels. By considering the power rating of commercial optical switches, Fig. 3 shows the total power consumption of the optical switches used in this DDC network (DDCN) for different ratios of MCF/SMF channels hired at the first and second tiers of the topology. The ratio of MCF to SMF optical switches at the second tier is variable and it also dictates the level of configurability required within the system, thus, Fig. 3 also analysis the total switching power consumption as function of total number of optical channels carried over MCFs at the second tier.

As the results suggest, introducing MCF switches at the first tier only while keeping the second tier fully SMF based can allow for up to 68-81% reduction in the total power consumption of the optical switching. This is made possible by the fact that a single 16 core MCF can carry all 16 optical channels per each individual dBrick and the full configurability can be provided like a normal SMF based topology by the SMF switching fabric at the second tier. Nevertheless, further reduction in power consumption will require the integration of MCF switches at the second tier, where different ratios of MCF/SMF at this tier can provide anywhere between 15% (high configurability) to above 90% (low configurability) reduction in power consumption.

Fig. 3 also, shows the total power consumption at the switching layer for various types of MCF interfaces with different core counts used on the MBOs to deliver the required optical channels assigned for delivery over MCF links (i.e. exact, two, four, six, eight). As it can be seen, the choice which can achieve the best configurability (i.e. 2 cores) consumes the most power and using an MCF with total number of cores equivalent to that which is required by the dBrick leads to the least power consumption. The lower bounds achieved in Fig. 3 are as result of employing only multi core fibers as well as switches between various dBrick capable of carrying all 16 channels over the lowest number of individual fibers. Since each dBrick had 16 optical channels, the lowest bounds were achieved by employing 16 core MCFs. However, this choice will result in the least amount of flexibility as it assumes all 16 channels on one dBrick will be directed to the other dBrick pair.

Regardless of MCF/SMF switching used in this architecture the power consumption associated with the main electronic as well as to the opto-electronic subsystems can be assumed to be constant. The power consumption of the dBoxes, dCOMPUBRICKs, dMEMBRICKs are 35, 19 and 23 Watts, respectively [24]. Moreover, commercial 8 channel MBOs such as those manufactured by Luxtera [25] are typically rated at 3 W for 10G applications. For the sake of simplicity, it can be assumed that a 16 channel MBO required by each dBrick consumes 6 Watts of power. Considering a 50/50 mix of CPU dBricks and memory dBricks in the 128 dBoxes envisioned, the total power consumption of the electronic and active optoelectronic subsystems of these systems can be calculated to be approximately 60000 Watts which is double the switching power if only SMF based switches were employed. Thus, the switching platform accounts for 30% of the total power consumption of the system.

C. MCF-SMF Port utilization

To evaluate the effectiveness of hybrid MCF/SMF topology

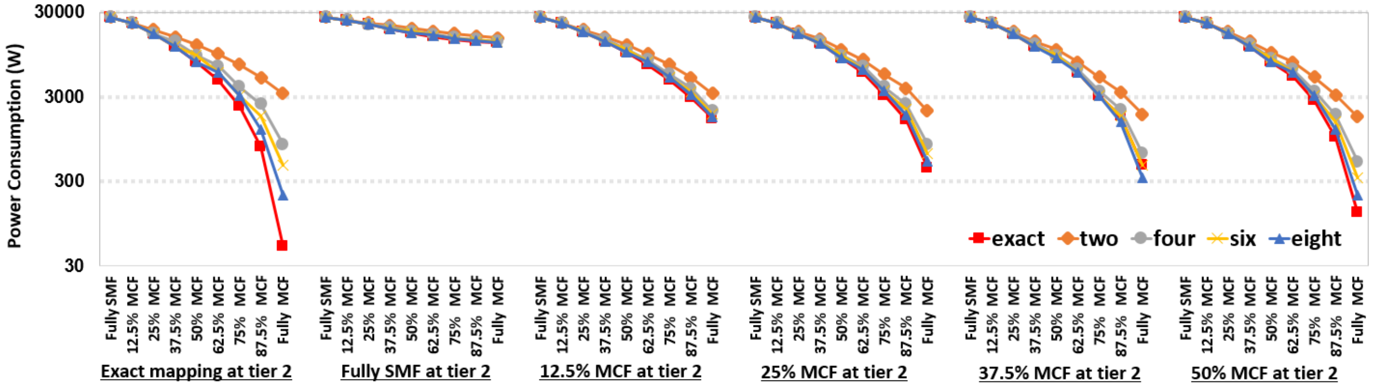
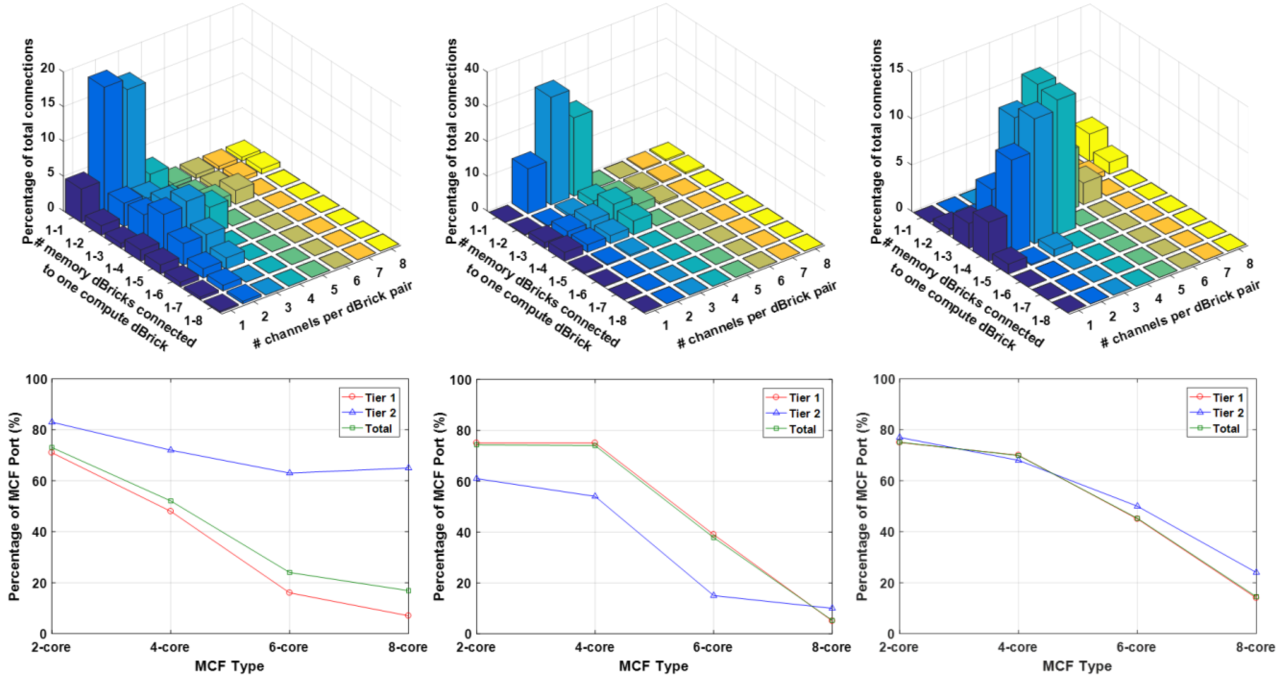


Fig. 3. Total power consumed by the optical switching layer in Watts (log scale) over the proposed hybrid topology employing different ratios of MCF/SMF based OCSs at the first and second tiers for routing optical channel from each dBrick using MCFs with 2,4,6,8 cores or the exact number of cores required.



(a) Random (b) High CPU (c) High Memory

Fig. 4. Network level Simulation showing distribution of optical channel assignments per dBrick basis in a proposed topology (top row), and the required ratio of MCF/SMF links given the use of MCF with specific core number and distribution of MC/SMF links in first and second tiers (bottom row) for a) Random, b) High CPU and c) High memory request scenarios

shown in Fig. 3, the network level simulator developed in [8, 23] is employed and restructured accordingly to represent the topology in Fig. 2. For these simulations, we assume similar architectural parameters which were used for power calculations in Fig. 3. Thus, total of 128 dBoxes were assumed. Each dBox hosts pluggable 8 dCOMPUBRICKs and 8 dMEMBRICKs and every dBrick has 16 individual ports. Moreover, each dCOMPUBRICK is assumed to have 64 CPU cores and each dMEMBRICK contains 64 GB RAM. In the simulation, each VM request requiring a certain number of CPU cores and memory arrives dynamically following a Poisson distribution with a 10-time units average inter-arrival time, which contains the information of the CPU core number, memory size, CPU-memory latency & bandwidth requirements as-well-as resource holding time. The holding time starts from 6300 time units and increases 360 time units for every 100 requests.

To analyze the situation for different workload scenarios, three types of request described in [8] are considered, 1) random request: *1–32 CPU cores and 1-32 GB RAM;* 2) high CPU request: *24–32 CPU cores and 1-8 GB RAM*; 3) high memory request: *1–8 CPU cores and 24-32 GB RAM*. Fig. 4 (a, b, c) shows the results which represent the number of optical channels required between various dBrick pairs making up a CPU dBrick connected to N memory dBricks for the various workload scenarios. As it’s clear in such architectures, a single optical channel between various dBrick accounts for the least number of connections, where at extreme cases they only account for only 10.9%, 4.9%and 8.7% of total connections required in the random, high CPU and high memory scenarios, respectively. However, as it can be seen 2-6 optical connections between dBrick pairs account for 86.7, 93.5, 86.6% of total connections required for the random, high CPU and high memory scenarios. Thus, MCF can play a significant rule in such systems for reducing power consumptions and also increasing connector density. The second row in Fig. 4, demonstrates the percentage of MCF connections required for each work load type considered, if MCFs with a certain core number is employed. It’s evident that the high memory and CPU scenarios can benefit the most from high core count MCFs, since many memory nodes are required to attach to each CPU, thus requiring a larger number of connections between a memory and CPU brick. The random case can also benefit from MCFs, where the integration of a 4 core MCF results into approximately 50% of all connections becoming dependable on multi core fibers and switches. These figures also show the resultant percentage of MCFs in either tier. As for the high CPU/memory scenarios, 95% of the traffic happens intra dBoxes (83% for the random scenario), the total percentage of MCFs in Fig. 4 (b, c) is close to the value at Tier 1. By comparing these ratios to Fig. 3, possible power consumption reductions for each scenario as result of using MCF switches can be derived. For the random scenario the employment of a 2 core MCF, can result into approximately 74% of reduction in power consumption at the switching layer. For the high CPU case the integration of 2 or 4 core MCFs can result into approximately 68 and 87% reduction in total power consumption at the switching layer. For the high memory scenario, the integration of 2, 6 and 8 core MCFs can result into approximately 81, 68 and 28% reduction in the power consumption at the switching layer compared to case using a purely SMF based system. Thus, the employment of 2 core MCF in the proposed topology for various workloads have the potential to reduce the power consumption at the application layer by 68-81%.

V. Experimental System

In order to evaluate the DCN disaggregation based on the dReDBox architecture and the hybrid MCF/SMF topology presented in the previous section. We make use of a fully developed and integrated hardware prototype which was designed and manufactured by the dReDBox consortium. The current prototype can only house up to three individual dBricks or IT elements, however, the final system will house up to 16 individual dBricks. The heart of each dBrick is the MPSoC equipped FPGAs which provide the computing power along with all networking and controlling functionalities of each dBrick.

A. System Setup

In order to evaluate the performance of the hybrid SMF/MCF topology experimentally, in conjunction with dReDBox's disaggregated topology and hardware the setup shown in Fig. 5 is used. For this demonstration, only two dBricks in the dBox prototype are employed, where one will act as the CPU (dCOMPUBRICK) element and the other as the memory element (dMEMBRICK). The FPGAs used on each IT element here is the Xilinx Zynq Ultrascale+ MPSoC equipped with GTH transceivers operating at 10 Gb/s. The compute resources in the dCOMPUBRICK are represented by a quad 4-core ARM processor and the memory resources in the dMEMBRICK are represented by a DDR4 module. The networking and routing functionalities on each dBrick are implemented on FPGAs. The glue logic (GL) in each dBrick translates between the physical memory address seen by CPU and remote memory address to access remote memory elements in the network, and it further maps the memory resources in network-encapsulated outgoing transactions. The ARM processing system (PS) is connected to two separate GLs by two master ports in parallel. The network on chip (NoC) is responsible for providing a link between the GL and one of the on chip electrical transceivers which allows it to forward read/write memory requests and data transactions to the appropriate physical ports of the optical transceiver. The NoC can carry out both circuit and packet switching capabilities. Given the limitations such as system latency, and limited transceiver bandwidths, a single optical channel will not be capable of sustaining the required memory throughput at the application layer. Thus, to meet the link bandwidths for accessing remote memory resources and also to increase the application perceived memory throughput, two or more individual on-board transceiver channels need to be used in conjunction with one another to create access between a common master port to the ARM processor and memory resources. To achieve this, channel bonding logic implemented immediately after NoC (Fig. 5) on the dCOMPUBRICK. Channel bonding block splits GL memory-encapsulated transactions and parallelizes them over 2 links per GL port for high capacity transactions in the egress direction and serializes them in the ingress direction. Thus, each VM can be served dynamically by up to two channels for memory transactions. Nevertheless, this can be scalable to 4, 8 or more channels. However, scalability depends on data-widths of each memory read/write request. On the dMEMBRICK side the same order of logical blocks are used on the programmable section (PL) section to route the read/write commands to the appropriate address of DDR4 memory elements through a DDR4 controller. Given that these bonded channels will have a common destination, the employment of MCF subsystems for these channels can ensure a reduction in system power consumptions as discussed previously.

B. Transceivers

As it can be seen, each dBrick houses both a SMF and MCF MBO. The SMF MBO has a total of 8 transceivers operating at 1310 nm [25]. Each channel employs OOK modulation and it can operate by up to 25 Gb/s per channel. The fiber interface on the photonic die on these MBOs comprises of 8 Tx and 8 Rx Grating Couplers (GCs) [26]. The MCF MBO is also based on a similar design to the SMF MBO and it has the same capabilities. However, it operates at 1490 nm and makes use of two MCF links each with 8 individual cores (cross sectional diagram in inset (a) in Fig. 5). Furthermore, in order to accommodate for MCF's small pitch core, the GCs are rearranged in a more compact formation on the chip's photonic die [27]. The cladding diameter of the MCF fiber employed at the output of the transceiver is 180 µm, and these

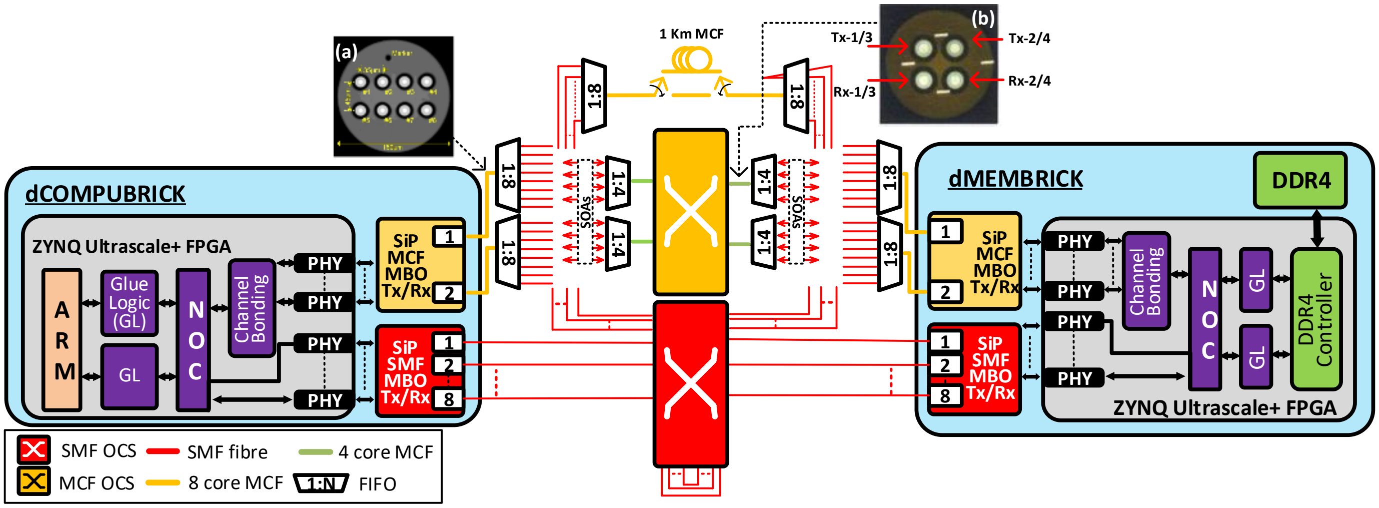


Fig. 5. Experimental setup, (a) cross section view of the 8 core fiber used to interface the MCF-MBOs and the 1km MCF, (b) cross sectional view of the fiber used in the MCF-switch.

are terminated by regular LC connectors. It should be noted that in a particular system individual MCF/SMF MBOs are not required where a single multi-channel MBO can be allowed to have both MCF and SMF interfaces. The employment of MCF links in such transceivers has the added advantage of increasing the connector density, in [28] it was shown that a move from SMF ribbons to 7 core MCFs can reduce connector density and lead to a 5-fold increase in the total number of cores which can be accommodated by 70% of front panel of a 1U panel. This is essential for disaggregated systems as the throughput between various IT elements can reach Tb/s and even Pb/s scales [14] which will require the employment of many individual optical channels in order to promote cost-effectiveness. Thus, the large number of optical connections require an increase in connector density. Moreover, the electronic as well as the optical circuitry on the MBOs used in this work account for a small percentage of the overall size of the chip, whereas the section of the photonic die responsible for coupling of light into optical fibers consumes a large space of the chip [27]. The integration of MCFs with these MBOs has the potential to enhance the throughput of each individual MBO my multiple folds without effecting the footprint of the device [27]. This can in turn lead to even higher switching density as the proposed switching architecture can adapt to higher core counts [29].

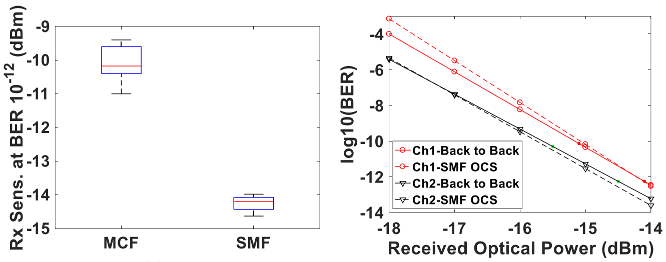
In this system, the electrical and optical transceiver has over 6 individual control parameters for each channel which controls the operation at the physical layer. Careful optimization of these parameters can allow a reduction in the receiver sensitivity and guarantee error free operation over an extended period without the need for FEC. Thus, a total of 48 control parameters exist per each MBO-FPGA pair which needs to be optimized, each have between 5-32 possible values. The electrical transceiver on the Xilinx FPGA has 3 distinct parameters which control the operation at the physical layer, one parameter defines the differential drive level that is directly delivered to the RF links of the MBO and the other two determine the transmitter side de-emphasis (pre- and post-) equalization. On the other hand, the opto-electronic transceiver supports three control parameters per channel, one for the transmitter side continuous time linear equalizer (CTLE), one for the receiver side pre-emphasis equalizer and one for the output driver. There are a total of $5x10^{36}$ possible combination of 48 control parameters, examining each individual parameter will not be feasible, therefore various combination of three control parameters inherent to the electrical transceiver on the FPGA are examined via an automated measuring system and the rest of the control parameters on the optical transceiver are manually tuned for optimized parameters identified for the electrical transceivers.

C. Optical switches

To switch SMF links, in this work we make use of a commercial optical switch module which is based on the patented DirectLight® [30] beam-steering optical switch technology which can provide low loss non-blocking connectivity between 2D arrays of fiber-coupled lenses in free space. using piezoelectric actuators for beam steering [31]. Switching is carried out completely independent of the power level, wavelength of operation and directionality of light and it achieves switching in the millisecond range. This switch has a total of 48 ports, moreover, this switch can be logically split by interconnecting some of its ports, in order to replicate the multi-tier topology which was proposed in Fig. 2. The loss encountered after going through each switching hop is 1dB on average [20]. For achieving switching over the SDM links, the MCF switch employed here also operates based on the beam steering concept. However, in order to accommodate switching MCFs, its architecture slightly defers from the latter [32]. The switch currently only accommodates four MCF ports each with 4 individual cores (cross section shown in inset (b) of Fig. 5), Nevertheless, this module is scalable to up to 96 ports [32]. The switch attains core-core losses below 2.2 dB after switching. The discrepancy in the core count between the MBO and the MCF switch was due to the architecture of the switch being tailored to operate with lower core counts. Alterations to the free space optics used in the MCF switch can allow for the employment of the 8 core fiber architectures used by the MBO (inset (a) of Fig. 5) which can be scalable to 32 – 40 ports within the current design.

In order to allow for an interface between the 8 core MCF connections and the 4 core connections of the MCF switch, fan-in-fan-out (FIFO) elements based on the waveguide coupling principle are used [9, 33]. The FIFOs connected at the interface of the MBO bring about an average loss of 1-1.5 dB and the FIFO connected at the interface of the optical switch have a total of 1.5-3.2 dB. Thus the total loss through the MCF switch can range between 7.2-11.6 dB, due to combined high losses associated with the MCF switch itself, it’s FIFOs and the FIFOs of the MCF MBOs used in this experiment. Thus, in order to enhance the power budget of the MCF interconnect, SOAs are used at the ingress core of each port of the MCF switch prior to the FIFOs. It should be noted that in a practical system such high losses would not be experienced, as all MCF interfaces will follow a similar core pitch and core count to that of the MCF switch.

VI. RESULTS AND DISCUSSIONS



(a) (b)

Fig. 6. (a) Box plot for the receiver sensitivity of multiple optical channels between the MCF or SMF MBOs of the dCOMPUBRICK and dMEMBRICK without bypassing through an optical circuit switch. (b) Performance of a bi-directional channel between the SMF MBOs of the dCOMPUBRICK and dMEMBRICK in back to back and after being switched by the SMF OCS in terms of received optical power (dBm) vs log10(BER)

A. *Physical Layer – Receiver sensitivities*

To determine the performance of the MCF and SMF MBOs, their receiver sensitivity is determined by directly connecting multiple channels of the MBOs on each dBrick to that of the other dBrick without passing through any OCSs. The measured receiver sensitivities are presented in the box plot in Fig. 6 (a). It should be noted for all testes carried out here a PRBS of length 2^{31}-1 is employed. As it can be seen in Fig. 5, on average the SMF and MCF MBOs achieve a receiver sensitivity of -14.1 and -10.2 dBm. As it can be clearly seen, there is a 4dB performance penalty associated with the MCF MBO. On average 1 dB of this penalty can be associated to the higher coupling losses associated with interfacing MCFs to the MBOs. The other 3 dB penalty, on the other hand can be associated to the poor performance of the GCs used in these MBOs at the higher wavelength (1490 nm) used.

Provided that each transceiver has an average output power of -3 dBm, it can be clearly seen that on average the MCF and SMF MBO can provide a total power budget of approximately 7 and 11 dB. Nevertheless, these values can be further enhanced by better optimization of transceivers which can be achieved by using evolutionary algorithms or reconfiguring the MCF MBO to operate at the 1300nm range.

B. *Physical Layer – Switching*

To analyze the impact of the SMF OCS on the performance of the interconnects, bi-directional connections are made between the two SMF MBOs on each of the dBricks used in Fig. 4, either directly or through the SMF OCS. The performance of these bi-directional links is shown in Fig. 6 (b) in terms of BER vs received optical power, where two directional paths are denoted as Ch1 and Ch2. As it can be seen in this diagram, the propagation through the SMF switch leads to a negligible loss in performance. This factor was further highlighted in [20], where it was shown that even passing through up to 8 SMF OCSes results in no loss in performance.

Next, in order to study the possible degradations that can result after the employment of MCF OCS, the experimental setup is rearranged such that 4 bi-directional optical channels are setup between the two MCF MBOs in the setup shown in Fig. 5. For achieving OCS, all four ports of the MCF switch are utilized, with two ports connected to the dCOMPUBRICK and the other two connected to the dMEMBRICK's MCF MBO through FIFOs. Each port of MCF OCS is set to house for 2 bi-directional channels, the layout of the cores are presented in inset b) of Fig. 5. The performance of the two bi-directional channels which would use two ports of the MCF switches for interconnecting the MCF MBOs are shown in Fig. 7. As it is apparent, the back to back connection for both of these bi-directional optical channels exhibits a similar trend. However, once the SOAs, are integrated into the system up to 5 dB of performance degradation is observed for some channels. This can be attributed to the fact that the SOAs employed here were designed to operate in the C-band and the MCF MBOs operated outside this. Nevertheless, comparing the scenarios where SOAs were used for both back to back and transmission through one hop of the MCF switch, it's

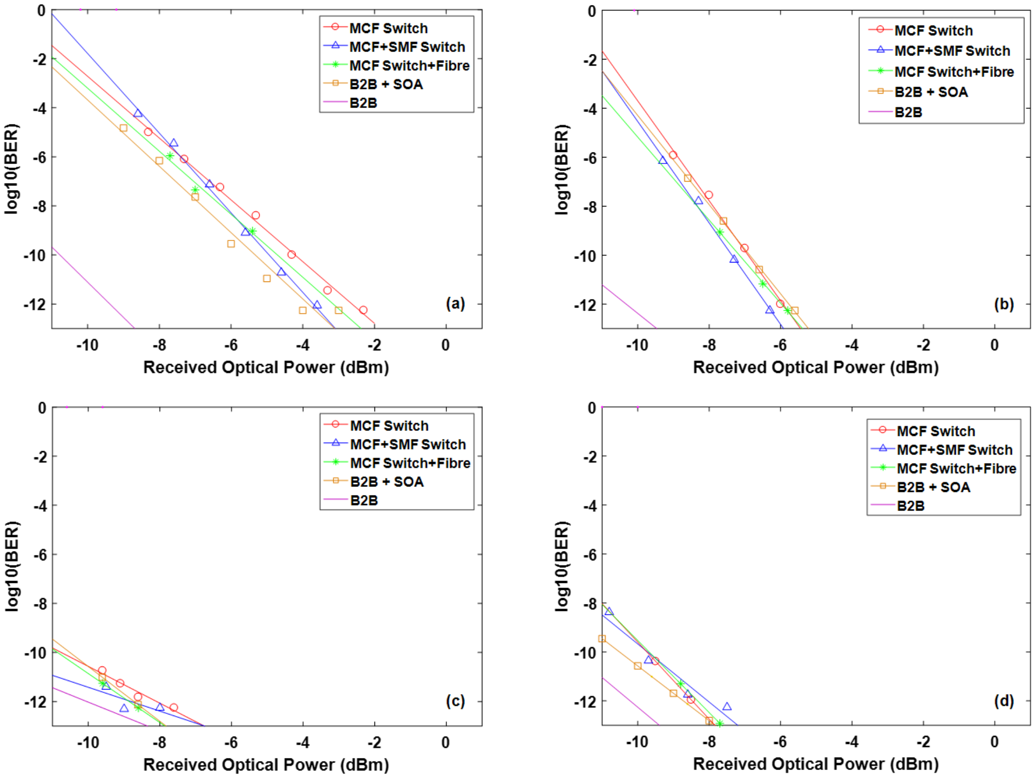


Fig. 7. Performance of two bi-directional channels connecting the MCF MBOs on two dBricks in terms of received optical power vs BER for Back to Back, Back to Back and amplification using SOA, a hop through the MCF switch, a hop through MCF switch and the SMF switch and a hop through the MCF switch and the MCF fiber. dCOMPUBRICK to dMEMBRICK connection on (a) Channel 1 (b) Channel 2, dMEMBRICK to dCOMPUBRICK connection on (c) Channel 1 and (d) channel 2

TABLE I
POWER BUDGET FOR VARIOUS dBrick TO dBrick LINKS IN HYBRID TOPOLOGY

No. Tiers	Switching path	Power budget req.(dB)
1	dBrick/***S-OCS***/dBrick	*1*
1	dBrick/***M-OCS***/dBrick	*2.2*
2	dBrick/***S-OCS***/***S-OCS***/***S-OCS***/dBrick	*3*
2	dBrick/***M-OCS***/***M-OCS***/***M-OCS***/dBrick	*6.6*
2	dBrick/***S-OCS***/***M-OCS***/***S-OCS***/dBrick	*7.2*
2	dBrick/***M-OCS***/***S-OCS***/***M-OCS***/dBrick	*8.4*

S-OCS: SMF-optical circuit switch, M-OCS: MCF-optical circuit switch

noticeable that for all 2 bidirectional channels, little or no discrepancies were observed. The slight discrepancy seen in these results can be related to the sensitivity of the MCF MBO to temperature variations, which can affect its performance. Given the high losses associated with this switch, it was not possible to propagate through more than one MCF switch, nevertheless, multiple hops thorough this MCF switch similar to the SMF OCS can also lead to degradation free transmission.

As it was shown in the proposed topology in Fig. 2, an interconnection between two particular dBricks within the proposed topology can pass through only MCF or SMF switches or a combination of two MCF switches and one SMF switch or two SMF switches and a MCF switch given the granularity required. To examine for possible performance degradation when SMF and MCF OCSes are used in conjunction with one-another, 2 bi-directional channels from the MCF MBOs on each dBrick are connected to one-another after being switched by a MCF switch with core mapping as shown in the inset (b) in Fig. 5 and a SMF switch. The performance of these links are also shown in Fig. 7, where it's clear that this also leads to no degradation in performance. Furthermore, in order to determine the possibility of degradation following transmission through long MCF links the four bi-directional channels surpassing a MCF switch are also routed through a 1 km span of an 8 core MCF. As results in Fig. 7 suggest, once again no penalties in performance are observed for this case.

C. Physical Layer – Power budgets

The OCSes in the topology proposed do not introduce distortion and only introduce insertion loss into the system, thus the combined loss through a series of OCSes determines the power budgets required. Table I shows the typical losses expected for connecting two dBrick through various routes in the hybrid topology proposed. In most cases no FIFOs are required, but, paths employing both MCF and SMF switches would require at least one pair of FIFOs. Given the typical losses across waveguide coupled FIFOs being around 1.5 dB [8], these paths will have at least 3 dB extra penalty. According to this, as stated in Table I, the power budgets required across the proposed topology would range from 1-8.4 dB. However, the SMF and MCF MBOs achieved power budgets of 7.2 and 11.2 dB respectively. Nevertheless, as discussed previously the power budget of the MCF MBOs can be further enhanced to that of the SMF MBOs.

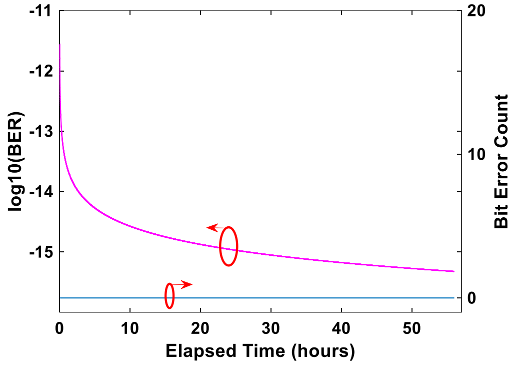


Fig. 8. Recorded BER for 8 bi-directional channels of the MBOs used between dCOMPUBRICK and the dMEMBRICK with various transmission losses over 55 hours. (Measurement interval: 25 seconds)

D. Physical Layer – FEC free performance

As it was stated earlier, the use of FEC encoders are prohibited in the disaggregated topology expressed in this paper. Nevertheless, these interconnects still require to deliver error free performance over the life time of a VM (up to a few days) which could be deployed over a dMEMBRICK-dCOMPUBRICK pair. The transceiver optimization routine, described in the previous section can ensure for this. It's known that MCFs suffer from inter-core crosstalk which is significant for intensity modulated schemes and dependent on environmental changes to the fiber [17, 28]. The crosstalk inherent to MCFs can severely limit the performance of the transmission. Thus, to determine whether the proposed architecture and topology can adhere to the FEC-free transmission over extended periods of time. 8 bi-directional channels were made between the dMEMBRICK and the dCOMPUBRICK, through MBOs used. In order to emulate the diversity of switching insertion losses which can be experienced in a practical implementation of the proposed topology, the bi-directional channels between the two MBOs are passed through lossy optical channels, with similar losses experienced in Table I. Three of these channels have between 8-9 dB of loss, one channel has an insertion loss of 7 dB, and the remaining channels have an insertion loss of either 1 or 3 dB. The performance of these 8 bi-directional optical channels are continuously recorded for up to 55 hours in terms of BER. The results are presented in Fig. 8, and as it can be seen even after 55 hours, no bit errors were recorded on any of the 16 links, proving the possibility of providing an error-free transmission between disaggregated elements.

E. Application Layer

In order to evaluate the performance of the proposed topology at the application layer for accessing memory resources by not using channel bonding or employing channel bonding for exploiting the benefits of MCF links, the perceived memory throughput at the application layer is measured. This is carried by attaching the CPU resource on the dCOMPUBRICK to either the available DDR4 resources on the dCOMPUBRICK itself or the remote DDR4 memory resource available on remote dMEMBRICKs by using the

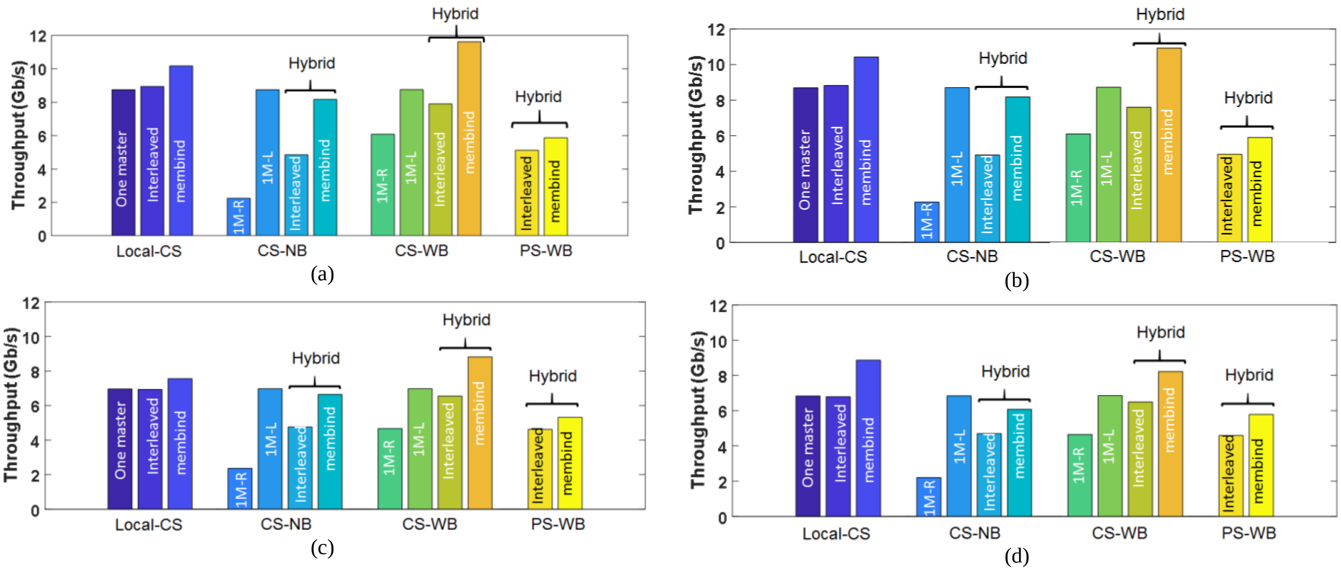


Fig. 9. Application level performance using the STREAMs benchmark for (a) copy (b) scale (c) add and (d) Triad operations for accessing local/remote DDR resource and accessing remote DDR resources in conjuniton with channel bonding with either packet or circuit switching (PS/CS). NB: No bonding, WB: with bonding. M: Master, R: Remote, L: Local. Employing three threads.

logical networking elements on the FPGA (shown in Fig. 5) and the optical circuit switched network. (Only 5 meters of fiber propagation was utilized; in practical implementations, this can increase substantially). To evaluate memory throughputs, a STREAMs benchmark running on customized Linux kernel [33] is used. The STREAM test is an industrial standardized subroutine used to evaluate the sustainable memory bandwidth in high performance computing systems [20]. It achieves this by measuring the perceived throughput from/to the attached memory resource while carrying four logical operations (1. copy, 2. scale, 3. add, 4. triad) on two long vectors.

1) LOCAL-CS: To benchmark the performance, initially, the perceived memory throughput for accessing DDR4 memory available locally on a dCOMPUBRICK is measured. To achieve this the NOC on the dCOMPUBRICK is configured to provide electrical circuit switching without using the on-board electrical transceivers. Moreover, the system is configured such that a connection is made between one or two master ports of the PS section (ARM processor) to access a 256 MB section of the local DDR4 memory. Fig. 9 presents the perceived memory throughputs for various logical operations and setups when three out of four threads of the ARM processor are utilized. For accessing local resources using circuit switching (Local-CS) as it can be seen using only one master port, the performance is saturated to approximately 7-8.5 Gb/s, the lower throughput for add and triad operations can be associated with higher latencies incurred due to back and forth transactions required between the CPU and memory resources for these operations. Furthermore, as it can be seen increasing the number of master ports and using the memory placement policies such as interleaving and membind, increases the total perceived memory throughput to up to 7.5-10.5 Gb/s. As it is shown in Fig. 10, the latency experienced for accessing local DDR4 memory on the dCOMPUBRICK was less than 50 ns which was mainly contributed by the NOC.

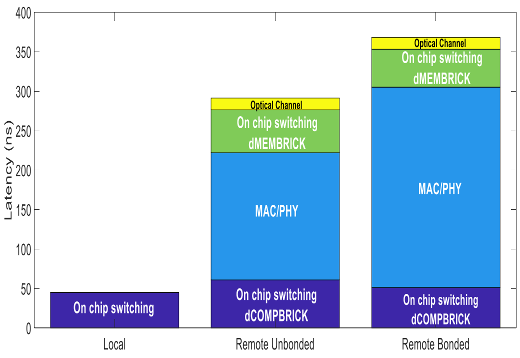


Fig. 10. End to End latency break down for accessing local and remote memory resources

2) CS-NB: Next, to assess the perceived memory throughput for accessing remote resources when channel bonding is not used (CS-NB), the NOC, GL, the MBOs and the OCSs in Fig. 3 are configured to connect one master port on the PS section of the dCOMPUBRICK to two 256 MB sections of the remote DDR 4 memory on the dMEMBRICK using two bi-directional optical circuit switched optical channel (1M-R). As the results in Fig. 9 suggest the 1M-R cases shows a significant reduction in throughput which are saturated at approximately 2 Gb/s (25% throughput sustained compared to the local case). This significant reduction comes at the cost of higher latencies endured for accessing the remote memory resources, where as it can be seen in Fig. 10, it had been measured to be approximately 290 ns which is an order of magnitude higher than what is required for DDR4 memories. The significant portion of this latency comes from the MAC/PHY functionalities of the electrical transceivers on the FPGAs. To increase these perceived memory throughputs, another master ports is also used to attach another 256MB DDR-4 memory section which resides locally on the dCOMPUBRICK (Hybrid) where memory placement policies are once again used. As it can be seen for this hybrid mode, the maximum throughput is recorded for the membind case which ranges between 6-8 Gb/s for various operations, this is a 300-400% increase in perceived throughput compared to the 1M-R case.

3) CS-WB: Next, in order to further increase the system throughput, channel bonding is employed in this system (CS-WB). It’s clear that by using channel bonding while accessing a memory resource remotely using one master port (1M-R) and two serial optical channels bonded together, the average memory throughput increases to 6 Gb/s for the copy operation which is an approximately 300% increase compared to the unbonded remote case (1M-R). This figure sustained for the scaling operation but falls to 4Gb/s for the scale and triad operations which is still higher compared to the non-bonded scenario. Nevertheless, by using both master ports either by interleaving or membinding the memory throughput increases to 8-11.8 Gb/s for the copy operation respectively which 25-33% higher than the non-bonded case (CS-NB). However, this figure slightly falls for the scale, triad and addition operations. As it can be seen in Fig. 10 when bonding is used the total end to end latency raises to 370 ns which is a 22% increase compared to the non-bonded case. This increase in latency comes due to the added logic, which is responsible for splitting and combining data frames at the transmitter and receiver side to make bonding possible. Nevertheless, the memory throughput substantially improves. The increase in this throughput is due to the fact that each individual master port operates at higher rates while channel bonding is employed compared to the non-bonded case which necessitates queuing and a drop in throughput. The increase in application perceived memory throughput as result of channel bonding and using multiple serial channels clearly shows the benefit of employing MCF topologies in DDCs. Furthermore, considering that end points in a contemporary DCN experience latencies in the range of 0.9-5 µs [8], it’s clear that the proposed architecture and topology has enabled a significant reduction in latencies experienced. As it can be seen the use of two master ports in conjunction with membind operation enhances the throughput to up to 8-12 Gb/s.

4) PS-WB: All scenarios presented up to here present the use of electrical circuit switching on the NOC, to evaluate the memory throughput were packet switching capabilities of the NOC are employed the output of two master ports are packet switched in the hybrid mode. The throughputs when packet switching and channel bonding is employed are highlighted in Fig. 9 (PS-WB), as it can be seen the throughput is saturated for all logical operations approximately between 5-6 Gb/s which is a 30-50 % reduction in throughput compared to the circuit switched case (CS-WB). This reduction in throughput can be contributed to the buffering required to employ at the NOC during packet switching.

VII. Conclusions

This paper proposed and experimentally evaluated a novel architecture (called dReDBox) and topology for next generation disaggregated data center networks. In order to provide the high capacities and low latencies required in disaggregated data center networks, the proposed architecture makes use of custom networking protocols implemented on on-board FPGAs in conjunction with high capacity optical interconnects relying on integrated Silicon photonic optics and optical circuit switches, enabling FEC free operation. Moreover, to further reduce the latency and also reduce the overall network power consumption, the proposed topology cut down the experienced latency over the optical network by up to 34% and it reduced the overall power consumption in excess of 60% by utilizing MCF based transceivers, circuit switches and fibers.

The end to end performance of the proposed architecture was evaluated and validated at the physical layer using the proposed low latency network topology hiring MCF subsystems. The results suggest that the system at the physical layer mostly suffers from insertion losses associated with the circuit switches and the possible use of pitch core inversion waveguide, these losses are measured to be between 1-8.4 dB. Nevertheless, these losses can be overcome by the power budgets inherent to the optical transceivers employed.

The performance of the dReDBox architecture is also evaluated at the application layer for accessing remote and disaggregated memory resources over the high capacity optical network using both electrical circuit and packet switching along with mutual bonding of two serial channels which can make use of multi core fibers. It was seen that only 25% of memory throughput was sustainable given the 290 ns latency experienced. However, it was demonstrated that the bonding of two serial ports that can be transmitted over MCF subsystem could enhance the memory throughput over remote resources by 300-400% despite the 22% increase in latency. It was observed that circuit switching results in a higher level of memory throughput compared to packet switching, given the higher latency from queuing, which can be experienced when packet switching is used.

Future works will include the redesign of the MCF switch and transceivers to allow for a larger number of cores per individual fiber link such that reliable and error free transmission would still be possible. Moreover, the logical designs of the FPGAs will be enhanced to maximize the throughputs at the application layer when MCF are employed in the system.

Acknowledgments

This work is supported by the EC H2020 dReDBox project with grant agreement 687632 and in part with EPSRC program EP/R035342/1 TRANSNET.

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] L. A. Barroso and U. Holzle, "The case for energy-proportional computing," (in English), *Computer,* vol. 40, no. 12, pp. 33-+, Dec 2007.

[3] H. M. M. Ali, T. E. El-Gorashi, A. Q. Lawey, and J. M. J. J. o. L. T. Elmirghani, "Future energy efficient data centers with disaggregated servers," vol. 35, no. 24, pp. 5361-5380, 2017.

[4] Intel. *Rack scale architecture*. Available: https://www.intel.com/content/dam/www/public/us/en/documents/guides/platform-hardware-design-guide-v2-1.pdf

[5] O. C. Project. *The Open Compute server architecture specifications*. Available: http://www.opencompute.org/

[6] H. P. Enterprise. *Moonshot System*. Available: https://www.hpe.com/us/en/servers/moonshot.html

[7] H. P. Enterprise. *The machine*. Available: https://www.labs.hpe.com/the-machine

[8] G. Zervas, H. Yuan, A. Saljoghei, Q. Q. Chen, and V. Mishra, "Optically Disaggregated Data Centers With Minimal Remote Memory Latency: Technologies, Architectures, and Resource Allocation [Invited]," (in English), *Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A270-A285, Feb 2018.

[9] P. X. Gao *et al.*, "Network Requirements for Resource Disaggregation," in *OSDI*, 2016, vol. 16, pp. 249-264.

[10] SATA-IO. *SATA-IO releases SATA revision 3.0 specification*. Available: Available: https://sata-io.org/sites/default/files/documents/SATA-Revision-3.0-Press-Release-FINAL-052609.pdf.

[11] M. Bielski, C. Pinto, D. Raho, and R. Pacalet, "Survey on Memory and Devices Disaggregation Solutions for HPC Systems," in *Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on*, 2016, pp. 197-204: IEEE.

[12] dReDBox. Available: http://www.dredbox.eu/

[13] G. Zervas *et al.*, "Disaggregated Compute, Memory and Network Systems: A New Era for Optical Data Centre Architectures," (in English), *2017 Optical Fiber Communications Conference and Exhibition (Ofc)*, 2017.

[14] Available: http://www.itrs2.net/

[15] I. Markit, "Data centre Optics Market Tracker - Worldwide,"

[16] G. M. Saridis, D. Alexandropoulos, G. Zervas, D. J. I. C. S. Simeonidou, and Tutorials, "Survey and evaluation of space division multiplexing: From technologies to optical networks," vol. 17, no. 4, pp. 2136-2156, 2015.

[17] H. Yuan *et al.*, "Experimental Investigation of Static and Dynamic Crosstalk in Trench-Assisted Multi-Core Fiber," in *Optical Fiber Communication Conference (OFC) 2019*, San Diego, California, 2019, p. W4C.2: Optical Society of America.

[18] M. Neitz, M. Wöhrmann, R. Zhang, M. Fikry, S. Marx, and H. Schröder, "Design and Demonstration of a Photonic Integrated Glass Interposer for Mid-Board-Optical Engines," in *Electronic Components and Technology Conference (ECTC), 2017 IEEE 67th*, 2017, pp. 538-544: IEEE.

[19] K. Hasharoni *et al.*, "A high end routing platform for core and edge applications based on chip to chip optical interconnect," in *Optical Fiber Communication Conference*, 2013, p. OTu3H. 2: Optical Society of America.

[20] A. Saljoghei *et al.*, "dRedDbox: Demonstrating Disaggregated Memory in an Optical Data Centre," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, 2018, pp. 1-3: IEEE.

[21] K. Bergman and S. Rumley, "Optical switching performance metrics for scalable data centers," in *OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS), 2016 21st*, 2016, pp. 1-3: IEEE.

[22] Y. Liu, H. Yuan, A. Peters, and G. Zervas, "Comparison of SDM and WDM on direct and indirect optical data center networks," in *ECOC 2016; 42nd European Conference on Optical Communication; Proceedings of*, 2016, pp. 1-3: VDE.

[23] H. Yuan, A. Saljoghei, A. Peters, and G. Zervas, "Disaggregated Optical Data Center in a Box Network using Parallel OCS Topologies," in *Optical Fiber Communication Conference*, 2018, p. W1C. 2: Optical Society of America.

[24] dReDBox, "D5.4–Software and hardware system integration and evaluation," 2018.

[25] Luxtera. Available: http://www.luxtera.com/products/

[26] T. Hayashi *et al.*, "End-to-end multi-core fibre transmission link enabled by silicon photonics transceiver with grating coupler array," in *European Conf. and Exhibition on Optical Communication (ECOC)*, 2017, p. Th. 2. A.

[27] T. Pinguet *et al.*, "Silicon photonics multicore transceivers," in *Photonics Society Summer Topical Meeting Series, 2012 IEEE*, 2012, pp. 238-239: IEEE.

[28] H. Yuan *et al.*, "Space-division multiplexing in data center networks: on multi-core fiber solutions and crosstalk-suppressed resource allocation," vol. 10, no. 4, pp. 272-288, 2018.

[29] C. Deakin, M. Enrico, N. Parsons, and G. J. J. o. L. T. Zervas, "Design and analysis of beam steering multicore fiber optical switches," 2019.

[30] A. Dames, "Beam steering optical switch," 2003.

[31] *Polatis*. Available: http://www.polatis.com

[32] H. C. H. Mulvad *et al.*, "Beam-steering all-optical switch for multi-core fibers," in *Optical Fiber Communication Conference*, 2017, p. Tu2C. 4: Optical Society of America.

[33] J. D. McCalpin. *STREAM: Sustainable Memory Bandwidth in High Performance Computers*. Available: https://www.cs.virginia.edu/stream/