Demonstrating Optically Interconnected Remote Serial and Parallel Memory in Disaggregated Data Centers

Vaibhawa Mishra, Joshua L Benjamin, Georgios Zervas
University College London, United Kingdom
vaibhawa.mishra@ucl.ac.uk

Abstract: Remote serial and parallel memory using memory-over-network bridge and optical switched interconnect is demonstrated. Remote memory bandwidth of 93% (HMC) and 66% (DDR4) of the local 3.2 and 3.7 GB/s bandwidth is showcased.
OCIS codes: 060.4253 Networks, circuit-switched; (060.4258) Network topology; (200.4650) Optical interconnects

1. Introduction

Disaggregation of computational resources in Data Center networks can tackle the challenges associated with resource wastage that exist within today's conventional server-centric data center models [1]. Low resource utilization is the result of the four orders of magnitude worth disproportionality in demand for CPU over memory resources [2]. However, disaggregation of memory from the CPU is most challenging; as CPUs demand tight coupling with memory using high speed memory or bus interconnect such as PCIe. We previously proposed an architecture that demonstrated remote parallel memory (DDR4) and showcased the remote memory access maximum performance of 0.62 GB/s [3] and ~1 GB/s [4]. Serial memory technologies such as Hybrid Memory Cube (HMC) and High Bandwidth Memory (HBM) modules are already supporting serial I/Os with communication bandwidth of >1Tb/s [5]. Such memory technologies with serialized I/Os can lead to low round-trip system latency. Thus, we introduce a memory over network (MoNet) system/bridge that can support serial and parallel memory elements in an optically switched disaggregated data center network that can (a) minimize latency, (b) maximize memory bandwidth and (c) scale out to many memory modules, locally or remotely. We evaluate the performance of serial memory communication using industry standard STREAM [6] and custom baseline memory benchmark. By comparing with existing memory-disaggregated architecture [7], we achieve 3.3x more DDR4 (2.43 GB/s) memory and 4.16x HMC (3 GB/s) bandwidth due to our pipelined light weight MoNet hardware and use of multiple (up to 8) channels. The best-case round-trip remote latencies achieved over 8-meter links (Rack-Scale) for DDR4 and HMC are 958 ns (14.9 times longer than local) and 403 ns (only 17% longer than local latency) respectively.

2. Architecture

Disaggregated tray architecture, depicted in Fig. 1, shows the installation of multiple interconnected resources such as (a) CPU Micro server card, (b) ASIC/FPGA hosted parallel memory card and (c) optically connected serial memory card. In case of parallel-based memory an ASIC/FPGA processing system is required to perform de/serialization and protocol stack as well as host a memory controller. The serial memory card only requires opto-electronic transceivers as the de/serialization and protocol stack is embedded in the HMC chip. This substantially reduces latency, power consumption and cost. However, HMC memory is supporting up to 2 GB DRAM per chip and doesn't scale as the parallel memories such as DDR. We use optically circuit switched (20msec reconfiguration) interconnect in order to pull together such heterogeneous compute and memory resources that operate a virtual-machine (VM) timeframes (seconds-hours). A dedicated resource management network is used to manage, monitor and orchestrate compute and network resource provisioning. The individual CPU and memory cards are implemented using commercial, off-the-shelf Xilinx based Multi-Processor System on Chip (MPSoC) based hardware solutions. To minimize footprint and power consumption and to maximize bandwidth density and front panel density, each card uses mid-board optics (MBOs) based on SiP technology with a capacity of up to 200 Gb/s (8x25). The HMC memory module supports up-to 64 channels each up-to 15 Gb/s/ (1.9 Tb/s bi-directional line

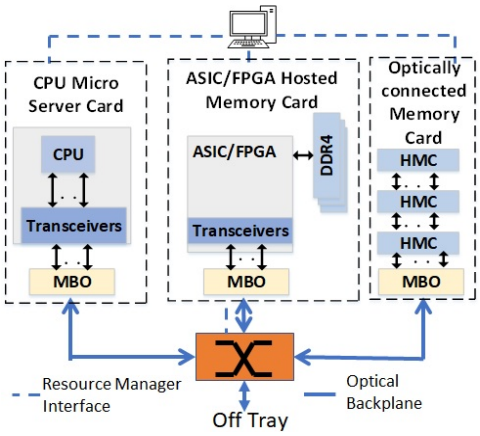


Fig. 1. Disaggregated Data Centre Tray Architecture.

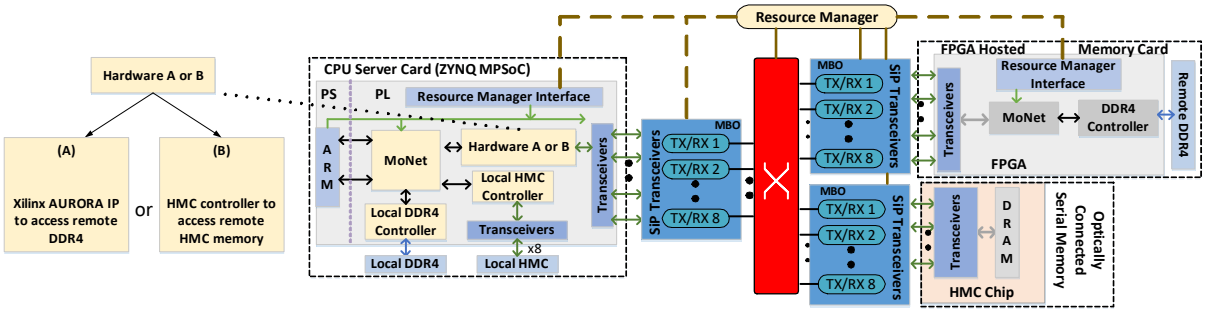


Fig. 2. Experimental system: (Optical switch disaggregates CPU and Memory Cards)

rate or 240 GB/s memory bandwidth) in a 4-lane configuration (one lane can be accessed by a single CPU). In this experiment we use eight channels or (half-width of a lane) capable of 240 Gb/s or 30 GB/s bidirectional memory bandwidth. Considering the electrical I/Os connected to the MPSoC, each CPU card can potentially support up-to 1.8 Tb/s (28 serial link each up-to 32.75 Gb/s) bidirectional line rate [8].

3. System Setup

Fig. 2 shows the experimental test-bed of the optically disaggregated (serial and parallel) memory architecture. CPU and parallel memory cards (DDR4) are implemented on Xilinx ZYNQ MPSoC FPGA, while HMC chips are directly attached to the network with serial transceivers. The CPU card embeds a 4-core processor and the dis-aggregated memory resources in the both memory cards are represented by a 256 MB DDR4 and 2 GB HMC serial memory. The MoNet logic in the CPU card translates and maps the physical memory addresses (seen by Linux OS) to local or remote DDR4/HMC physical memory. MoNet operates in memory mapped mode and is fully pipelined, which requires 2 clock cycles at 312.5 MHz to forward read/write memory requests to the appropriate memory address (pointing to the local or remote controller). Each transceiver port on the FPGA is assigned a different channel on the multi-channel SiP MBO to access remote DDR4 or HMC memory available in the network. To interconnect all the MBOs to an optical circuit switching backplane, a 48-port Polatis switch is used; however, this switch is logically split to realize multi-tier network topology. In our proposed architecture, CPU server runs a customized Linux based OS. This OS is capable of (a) hot-plugging the remote memory resources and (b) managing and instructing the MoNet through a custom driver. The Luxtera 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 -4.1 dBm and uses OOK modulation. Although individual channels can operate at up to 25 Gb/s, the operation of each channel is limited due to the transceiver bit-rates (10, 12.5 and 15 Gb/s) of HMC chip. Each hop through the beam-steering optical switch module has a 1 dB power penalty and it consumes a low power of 100 mW/port [9].

4. Result & Discussion

Fig. 3(a) represents the bit error rate performance of 10, 12.5 and 15 Gb/s bi-directional optical links between CPU and optically interconnected memory, after traversing multiple hops through the optical switch. Provided that each SiP transceiver has an average output power of -3 dBm (-4.81 dBm back-to-back includes MTP to LC fan-in fan-out), the SMF along with optical switch can be allocated a total power budget of approximately 11 dB, thus, enabling our setup to emulate a multi-tier network topology. The multi-tier level disaggregation is dependent on and restricted by the serial link rate. While at 10 Gb/s link connectivity we perform error free operation up to -13.9 dBm,

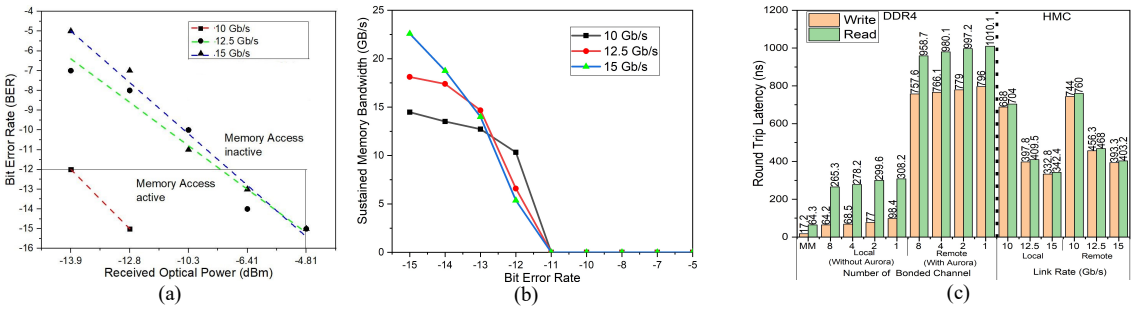
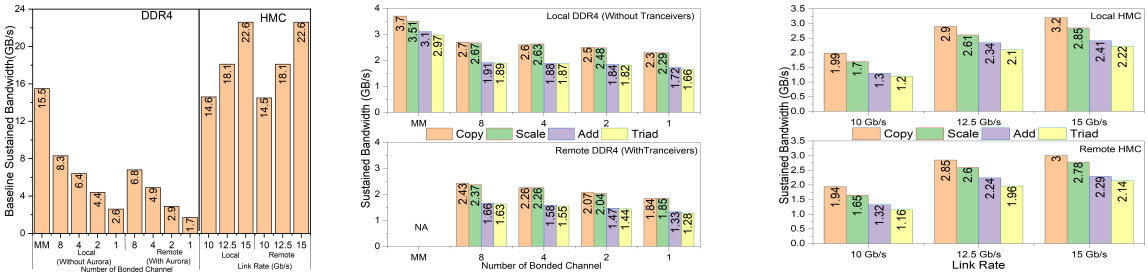


Fig. 3 (a) Performance of bi-directional channel CPU and HMC in terms of received optical power vs bit error rate (BER), (b) Impact on throughput due to error in optical link. (c) Latency comparison between local/remotely attached DDR4 and HMC on various rate.



(a) (b) (c)

Fig. 4 Sustainable bandwidth achieved for DDR4 and HMC in (a) baseline and (b) and (c) STREAM benchmark (4 CPU threads).

at 12.5 and at 15 Gb/s we can only achieve error free operation up to -6.4 dBm (allowing for 3-hop network i.e spine-leaf). HMC instead of using forward error correction (FEC) offers re-transmission of memory transactions in case of packet loss. However, as shown in Fig. 3(b) error free interconnection is essential as memory bandwidth drops to zero at BER of 10^{-11}. In order to evaluate the optically connected HMC performance such as latency and throughput, MoNet initiates memory transaction (fully pipelined, simultaneous write/read operation, burst size =1, memory granularity = 128 Byte) to assess performance introduced by transceivers, optical data-path and memory technology. Remote serial memory offers 403 ns remote latency (purely impacted by propagation delay) compared to 958 ns of parallel memory (mostly caused by Aurora transceiver protocol stack deployed on both compute and memory cards) as shown in Fig. 3(c). While comparing between local and remote attachment for same memory type, remote access additional latencies vary between 50ns to 60ns (at various link rates) for HMC however on the other hand, this leads to 740 ns extra latency for DDR4 memory.

Remote access latency and bit-rate per channel also impacts the memory throughput as shown in Fig. 4(a), where HMC achieves 22.6 GB/s while DDR4 sustains at 6.8 GB/s. Next, by using customized Linux OS distribution loaded onto the CPU server card, the parallel (DDR4) memory from memory card and serial memory from standalone HMC card are attached to the system. To measure the overhead added by the optical interconnect, we also attached DDR4 and HMC memory locally to CPU server card via MoNet logic as shown in Fig. 2. We compare the perceived sustainable application-level memory bandwidth in these configurations by running the STREAM benchmark [6] using quad core ARM and simultaneously accessing the attached memory. Fig. 4 (b) and Fig. 4(c) shows the results: a maximum throughput of 3.7 and 3.2 GB/s is achieved by the local DDR4 and HMC, while remote memory sustains 66% (DDR4) and 93% (HMC) of that due to inferred interconnect latencies.

5. Conclusion

This paper presented architecture and memory over network logic, which introduce serial and parallel memory to disaggregated Data Centers over a low latency optical network. It was shown that optical interconnects employed in this architecture can achieve a FEC-free operation for a 2-tier and 4-tier topology at 12.5/15 and 10 Gb/s respectively. Parallel and serial memory disaggregation over an optical network was demonstrated with a throughput of 2.43 and 3 GB/s respectively when using 8-channel bonded links at application layer. Compared to the local serial memory, the disaggregated HMC memory access experiences only 40ns extra round-trip latency (pure propagation delay over 8 meter – Rack Scale) and sustains a 93% of throughput.

6. Acknowledgement

The work is supported by EPSRC TRANSNET (EP/R035342/1) programme grant.

References

[1] Georgios Zervas, et al, "Optically Disaggregated Data Centers With Minimal Remote Memory Latency: Technologies, Architectures, and Resource Allocation [Invited]," J. Opt. Commun. Netw. 10, A270-A285 (2018).
[2] S. Han, et. al.," presented at the Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks, College Park, Maryland, 2013.
[3] A. Saljoghei, et. al., “dreddbox: Demonstrating disaggregated memory in an optical data centre,” OFC, March 2018, pp. 1–3.
[4] D. Syrivelis, et al., "A Software-defined Architecture and Prototype for Disaggregated Memory Rack Scale Systems," presented at the samos-conference, 2017
[5] [Online].Available:http://www.eejournal.com/article/20170102-hbm-hmc/.
[6] John D. McCalpin, “STREAM: Sustainable Memory Bandwidth in High Performance Computers,” University of Virginia, Tech. Rep., 1991-2007.
[7] D. Theodoropoulos, et. al., "REMAP: Remote mEmory Manager for disAggregated Platforms," 2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP), Milan, 2018, pp. 1-8.
[8] [Online].Available: https://www.xilinx.com/support/documentation/selection-guides/zynq-ultrascale-plus-product-selection-guide.pdf
[9] [Online]. Available: https://www.polatis.com/