Journal of Parallel and Distributed Computing 163 (2022) 300–312

Contents lists available at ScienceDirect

Journal of Parallel and Distributed Computing

www.elsevier.com/locate/jpdc

Optically connected memory for disaggregated data centers

Jorge Gonzalez[a,b,*], Mauricio G. Palma[b,**], Maarten Hattink[e,***], Ruth Rubio-Noriega[c], Lois Orosa[d], Onur Mutlu[d], Keren Bergman[e], Rodolfo Azevedo[b]

[a] University of Engineering and Technology, Lima, Peru
[b] University of Campinas, São Paulo, Brazil
[c] INICTEL-UNI, Lima, Peru
[d] ETH Zurich, Switzerland
[e] Columbia University, New York, USA

ARTICLE INFO

Article history:
Received 1 March 2021
Received in revised form 19 October 2021
Accepted 19 January 2022
Available online 29 January 2022

Keywords:
Disaggregated memory
Photonics
Data-centers
DRAM
Memory systems

ABSTRACT

Recent advances in integrated photonics enable the implementation of reconfigurable, high-bandwidth, and low energy-per-bit interconnects in next-generation data centers. We propose and evaluate an Optically Connected Memory (**OCM**) architecture that disaggregates the main memory from the computation nodes in data centers. OCM is based on micro-ring resonators (MRRs), and it does not require any modification to the DRAM memory modules. We calculate energy consumption from real photonic devices and integrate them into a system simulator to evaluate performance. Our results show that (1) OCM is capable of interconnecting four DDR4 memory channels to a computing node using two fibers with 1.02 pJ energy-per-bit consumption and (2) OCM performs up to 5.5× faster than a disaggregated memory with 40G PCIe NIC connectors to computing nodes.

1. Introduction

Scaling and maintaining conventional memory systems in modern data centers is challenging for three fundamental reasons. First, the dynamic memory capacity demand is difficult to predict in the short, medium, and long term. As a result, memory capacity is usually over-provisioned [49,59,26,65,30], which wastes resources and energy. Second, workloads are limited to using the memory available in the local server (even though other servers might have unused memory), which could cause memory-intensive workloads to slow down. Third, memory maintenance might cause availability issues [55]; in case a memory module fails, all running applications on the node may have to be interrupted to replace the faulty module. A promising solution to overcome these issues is to disaggregate the main memory from the computing cores [45]. As depicted in Fig. 1, the key idea is to organize and cluster the memory resources such that they are individually addressable and accessible from any processor in the data center [16]. Memory disaggregation provides flexibility in memory allocation, improved utilization of the memory resources, lower maintenance costs, and lower energy consumption in the data center [60].

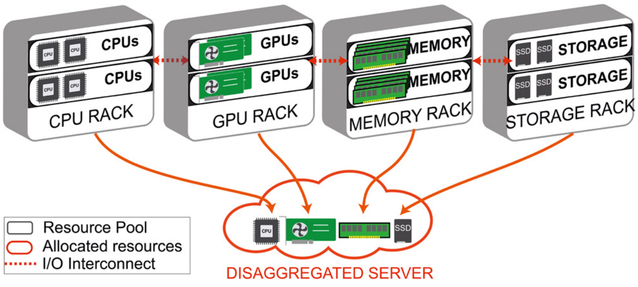


Fig. 1. Disaggregation concept for data centers.

Disaggregating memory and processors remains a challenge, although the disaggregation of some resources (e.g., storage) is common in production data centers [43]. Electrical interconnections in rack-distances do not fulfill the low latency and high bandwidth requirements of modern DRAM modules. The primary limitation of an electrical interconnect is that it constrains the memory bus to onboard distance [70] because the electrical wire's signal integrity loss increases at higher frequencies. This loss dramatically reduces the Signal-to-Noise Ratio (SNR) when distances are large. An optical interconnect is more appealing than an electrical interconnect for memory disaggregation due to three properties: its

[*] Corresponding author at: University of Engineering and Technology, Lima, Peru.
[**] Corresponding author at: University of Campinas, São Paulo, Brazil.
[***] Corresponding author at: Columbia University, New York, USA.
E-mail addresses: jgonzalez@utec.edu.pe (J. Gonzalez), mauricio.palma@ic.unicamp.br (M. G. Palma), mh3654@columbia.edu (M. Hattink).

https://doi.org/10.1016/j.jpdc.2022.01.013
0743-7315/© 2022 Elsevier Inc. All rights reserved.

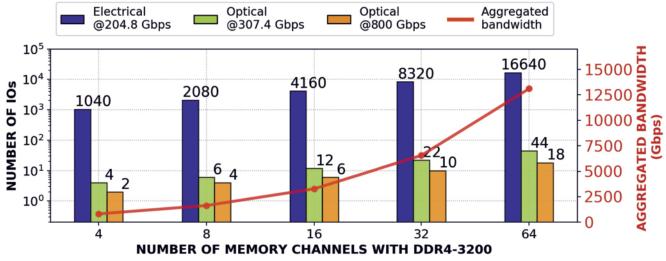


Fig. 2. Required electrical and optical IO counts (lower is better) for sustaining different amounts of aggregated bandwidth.

(1) high bandwidth density significantly reduces the number of IO lanes, (2) power consumption and crosstalk do *not* increase with distance, and (3) propagation loss is low. Silicon Photonic (SiP) devices are likely suitable for disaggregation, delivering ≥ Gbps range bandwidth, as well as efficient and versatile switching.

The **goal** of this work is to pave the way for designing high-performance *optical memory channels* (i.e., the optical equivalent of an electrical memory channel) that enable main memory disaggregation in data centers. Our work provides an optical link design for DDR DRAM memory disaggregation, and it defines its physical characteristics, i.e., i) number of Micro-Ring Resonator (MRR) devices, ii) bandwidth per wavelength, iii) energy-per-bit, and iv) area. We evaluate the performance (see Section 4.2) and energy consumption (see Section 4.3) of a system with disaggregated commodity DDR DRAM modules.

We make three key contributions: (1) we propose the Optically Connected Memory (OCM) architecture for memory disaggregation in data centers based on state-of-the-art photonic devices, (2) we perform the first evaluation of the energy-per-bit consumption of a SiP link using the bandwidth requirements of current DDR DRAM standards, and (3) we model and evaluate OCM in a system-level simulator and show that it performs up to 5.5× faster than a 40G NIC-based disaggregated memory.

2. Motivation

Photonics is very appealing for memory disaggregation because: (1) the integration (monolithic and hybrid) between electronics and optics has already been demonstrated [3], which allows the design and fabrication of highly-integrated and complex optical subsystems on a chip, and (2) optical links offer better scaling in terms of bandwidth, energy, and IO compared to electrical links; e.g., optical switches (o-SW) show better port count scaling [68]).

New electrical interfaces, such as GenZ, CCIX, and OpenCAPI, can disaggregate a wide range of resources (e.g., memory, accelerators) [15]. Optical devices can enable scalable rack-distance, and energy-efficient interconnects for these new interfaces, as demonstrated by a previous work that disaggregates the PCIe interface with silicon photonics [79]. Our OCM proposal extends the memory interface with optical devices and does not require substantial modifications to it, e.g., the memory controllers remain on the compute nodes. It is a direct optical point-to-point approach without additional protocols, such as PCIe. OCM can be used in next-gen disaggregated datacenters [33], as optical transceivers integration on the server motherboard reaches maturity [2].

Fig. 2 shows the IO requirements in the memory controller for electrical [50], and optical interconnects to achieve a specific aggregated bandwidth. We define IO as the number of required electrical wires or optical fibers in the interconnects. We use, for both electrical and optical interconnects, 260-pin DDR4-3200 DRAM modules with 204.8 Gbps maximum bandwidth per memory channel. We make two observations. First, the required number of optical IOs (left y-axis) is up to three orders of magnitude smaller than the electrical IOs because an optical fiber can contain many *virtual channels* using Wavelength Division Multiplexing (WDM) [19,9]. Second, a single optical IO achieves up to 800 Gbps based on our evaluation, requiring 2 IOs for bidirectional communication (see Section 4.3). An optical architecture could reach the required throughput for a 4 memory channel system using only 2 IOs (two fibers) and for a 32-channel system with only 10 IOs.

3. OCM: optically connected memory

To overcome the electrical limitations that can potentially impede memory disaggregation, we introduce an OCM that does not require modifications in the commonly-used DDR DRAM protocol. OCM places commodity DRAM Dual Inline Memory Modules (DIMMs) at rack-distance from the processor, and it sustains multiple memory channels by using different wavelengths for data transmission. OCM uses conventional DIMMs and memory controllers, electro-optical devices, and optical fibers to connect the computing cores to the memory modules. Our work explores the idea of direct point-to-point optical interconnects for memory disaggregation and extends prior works [21,5], to reduce the latency overhead caused by additional protocols such as remote direct memory access (RDMA) and PCIe [78]. OCM is versatile and scales with the increasing number of wavelengths per memory channel expected from future photonic systems [33].

3.1. Architecture overview

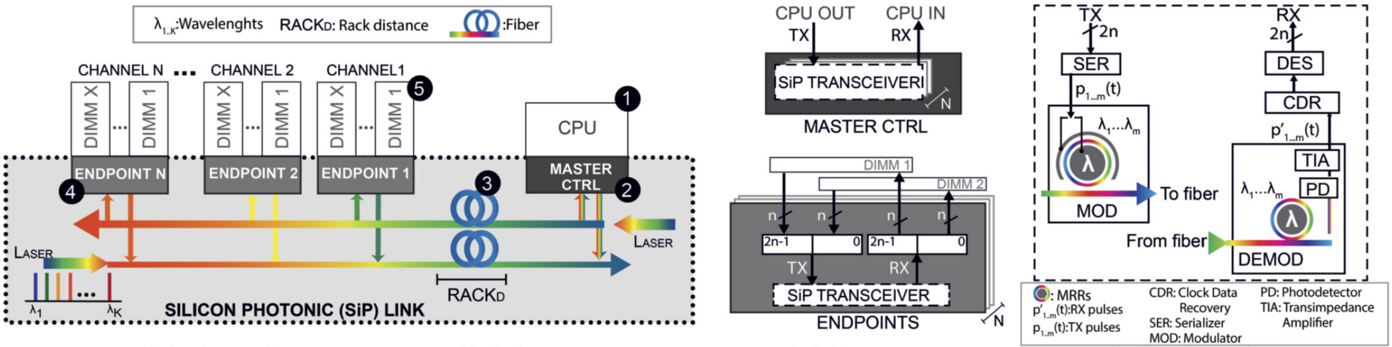
Fig. 3a shows the main components of the OCM architecture configured with state-of-the-art photonic devices and DDR memories. OCM uses N optical memory channels, each one consisting of X memory modules (DIMM 1 to X) operating in lockstep.

OCM uses two key mechanisms to take advantage of the high aggregated bandwidth of the optical domain while minimizing the electrical-optical-electrical conversion latency overhead. First, it implements an optical memory channel with multiple wavelengths that can support multiple DIMMs in a memory channel. Second, it achieves high throughput by increasing the cache line size and splitting it across all the DIMMs in a memory channel. For example, if OCM splits a single cache line between two DIMMs, it halves the bus latency (i.e., data burst duration tBL), compared to a conventional DDR memory. Then, more data will be moved per memory transaction.

In our evaluation (Section 4), we use two DDR channels operating in lockstep to get a cache line of 128 bytes with similar latency as a cache line of 64 bytes in a single DDR channel (Section 3.2). OCM benefits from the use of a wide Xn-bit interface, where X is the number of DIMMs, and n is the width in bits of a DIMM bus. OCM transfers depend on the serialization capabilities of the SiP transceiver.

The serialization/deserialization latency increases with the number of DIMMs in lockstep. Notice that a commercial SERDES link supports serialization up to 256 B (i.e., four 64 B cache lines). A larger than 64 B cache line can help overcome serializer underutilization. For example, the latency of 28 serialized transactions of 128 B is only 50% higher than the required latency for 28 serialized transactions of 16 B [37].

As shown in Fig. 3a, on the CPU side, there is a Master controller, and on the memory side, there are N Endpoint controllers that respond to CPU requests. Both controllers have a structure called SiP Transceiver, and Fig. 3b shows a difference in the organization of the SiP transceivers per controller. Fig. 3c shows the SiP transceivers present in the Transmitter (TX) and Receiver (RX) lanes in both Master and Endpoint controllers. A TX lane consists of a serializer (SER) and Modulator (MOD) for transmitting data. An



a) Optically Connected Memory (OCM) b) OCM controllers c) SiP Transceiver

Fig. 3. Optically Connected Memory organization: optical memory channels for disaggregation of the main memory system.

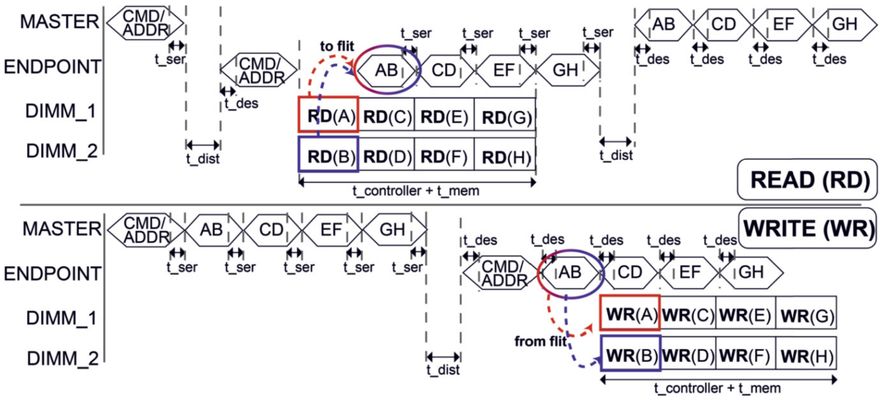


Fig. 4. OCM timing diagram for Read (top) and Write (bottom) requests.

RX lane contains a Demodulator (DEMOD), a Clock and Data Recovery (CDR) block, and a Deserializer (DES) for receiving data. Both TX and RX lanes connect with a Xn-bit (e.g., $X = 2$ and $n = 64$ in our evaluation) bus to the Endpoint controller, which forms the bridge between the lanes and the DRAM module.

3.2. Timing model

OCM transfers a cache line as a serialized packet composed of smaller units called *flits*, whose number depends on the serialization capabilities of the SiP transceiver. Fig. 4 presents the timing diagram of the OCM Read (RD) and Write (WR) operations. For reference, a conventional DDR DRAM memory channel uses 64 B cache lines; a data bus transfers each line as 8 B data blocks in 8 consecutive cycles, and the 1 B Command (CMD) and 3 B Address (ADDR) use separate dedicated buses. In OCM, as depicted in Fig. 4, the cache line is transferred in AB-GH flits. We show OCM timing with a $flit$ size that doubles the width of the memory channel data bus, and is the reason for dividing the cache line between DIMMs 1 and 2 to perform parallel access and decrease latency. OCM splits a single cache line between two DIMMs, which halves the bus latency (i.e., tBL [40]), compared to conventional DDR DRAM memory.

For the RD operation, data A and B are read from different DIMMs to compose a flit (AB). Flit AB serialization and transmission occur after the Master controller receives the CMD/ADDR flit. For the WR operation, the Master controller sends the flit containing data blocks AB immediately after the CMD/ADDR flit. After Endpoint deserialization, DIMM 1 stores A, and DIMM 2 stores B. For example, OCM with a commercial Hybrid Memory Cube (HMC) serializer [37] and 128 B cache line size, transfers 2 × (4 × 16 B of data) with 1 × 4 B CMD/ADDR initiator message (or *extra flit*).

Compared to conventional electrical DDR memory, OCM adds serialization and optical packet transport latency to the overall memory access time (see Section 4). The DIMM interface can support the latency overhead that is imposed by our optical layer integration. In our evaluation, we consider both optimistic and worst-case scenarios. Past experimental works [5] show that the overhead is low in the order of a few nanoseconds, requiring no modification to the memory controller. However, if there is high latency imposed by the optical layer, the signaling interface from the memory controller needs to be adapted. Equation (1) shows the OCM latency model T_{lat}, which is defined as the sum of the DIMM controller latency T_{contr}, DIMM WR/RD latency $T_{mem(A|B)}$ (latency is equal for both DIMMs), serialization/deserialization latency T_{serdes}, modulation/demodulation latencies T_{mod} and T_{demod}, distance propagation latency penalty T_{dist}, and system initialization time (e.g., Clock Data Recovery (CDR) latency, modulator resonance locking [58]) T_{setup}.

$$\begin{aligned} T_{lat}(t) = & T_{setup} + T_{contr} + T_{mem(A|B)}(t) + T_{serdes} + T_{mod} \\ & + T_{demod} + T_{dist} \end{aligned} \tag{1}$$

T_{setup} equals zero because it has no impact on the system once it is configured [5]. In the optical and millimeter wavelength bands, T_{mod} and T_{demod} are in the order of ps [9], due to the small footprint of ring modulators (tens of micrometers) and the high dielectric constant of silicon.

3.3. Operation

Fig. 3a illustrates the five stages of a memory transaction.

Stage ❶: the processor generates a Read/Write (RD/WR) memory request. In the photonic domain, a laser source generates light in

Table 1
Baseline processor, memory, OCM, and NIC.

Baseline	*Processor* *Cache*	3 GHz, 8 cores, 128 B cache lines 32 KB L1(D+I), 256 KB L2, 8 MB L3
MemConf1	*Mem*	4 channels, 2 DIMMs/channel, DDR4-2400 [40]
MemConf2	*Mem* *DRAM cache*	1 channel, 2 DIMMs/channel, DDR4-2400 4 GB stacked, 4-way, 4K pages, FBR [77], DDR4-2400
OCM	*SERDES* *Fiber*	latency: 10/150/340 cycles latency: 30/60/90 cycles (2/4/6 meters roundtrip)
NIC	*40G PCIe [57]*	latency: 1050 cycles

$\lambda_{1,2,\ldots,K}$ wavelengths simultaneously [10].

Stage ❷: the data from the processor is serialized (SER) onto the Master Controller's TX lane, and the generated electrical pulses $p_{1,2,\ldots,m}(t)$ drive the cascaded array of Micro-Ring Resonators (MRRs) for modulation (MOD), represented as rainbow rings. We use non-return-to-zero on-off keying (NRZ-OOK) that represents logical ones and zeros imprinted on the envelope of light [9].

Stage ❸: the optical signal is transmitted through an optical fiber. At the end of the fiber, the combined optical WDM channels are coupled into an optical receiver.

Stage ❹: first, in the RX lane of an Endpoint, the WDM Demodulator (DEMOD) demultiplexes the optical wavelengths using m MRRs. Each MRR works as an optical band-pass filter to select a single optical channel from $\lambda_{1,2,\ldots m}$. Second, these separated channels are then fed to DEMOD's integrated photo-detectors (PD) followed by transimpedance amplifiers (TIA). Together the PD and TIA convert and amplify the optical signal to electrical pulses $p'_{1,2,\ldots,m}(t)$ suitable for sampling. Third, the data is sampled, deserialized (DES), and sent to the memory controller.

Stage ❺: the processor accesses memory with the DDR protocol using a RD or WR command and a memory address. For a RD command, the Endpoint TX transmits to the processor a *cacheline* with the wavelengths $\lambda_{1,\ldots,m}$ (similar to Stages 1 to 4). For a WR command, the data received from the processor is stored in memory.

3.4. Enabling reconfigurability

OCM supports reconfigurability by placing an o-SW between the Endpoints and the Master controller, similar to previous work [5]. OCM uses optical switching to connect or disconnect a master controller from an endpoint. Switching can happen (1) in the setup phase, which is the first time that the system is connected before starting execution, or (2) before executing a workload, to adapt the amount of assigned memory to the requirements of the workload.

As depicted in Fig. 5, an optical switch has multiple ports, through which a set of N processors can be connected to a configurable set of M OCMs, where N and M depend on the aggregated bandwidth of the SiP links. In Section 4, we evaluate OCM with a single CPU, and assume that the setup phase is already completed.

3.5. High aggregated bandwidth

OCM uses WDM [9,19] to optimize bandwidth utilization. WDM splits data transmission into multiple colors of light (i.e., wavelengths, λs).

To modulate data into lightwaves, we use Micro-Ring Resonator (MRR) electro-optical modulators, which behave as narrowband resonators that select and modulate a single wavelength. We use MRRs because they have a small hardware footprint and low power consumption [10], and they are tailored to work in the communications C-band (1530-1565 nm). For more detail on photonic devices, please see [33,69,6].

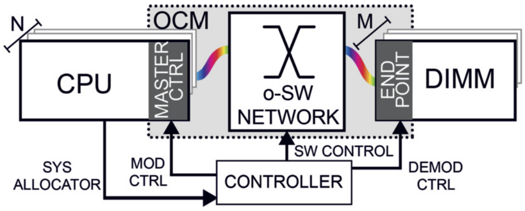


Fig. 5. Reconfigurable OCM with optical switches (o-SW).

OCM achieves high aggregated bandwidth by using multiple optical wavelengths $\lambda_{1,2,\ldots,K}$ (see laser in Fig. 3a) via WDM in a single link. The K wavelengths are evenly distributed among the controllers, where the TX/RX lanes of a single DDR memory channel have the same number (m) of optical wavelengths ($\lambda_{1,2,\ldots,m}$, see Fig. 3c). All wavelengths have the same bit rate b_r, and the aggregated bandwidth for N memory channels is $BW_{aggr} = b_r \times m \times N$. Assuming that BW_{aggr} is higher than the required bandwidth for a single memory channel BW_{mc}, then $BW_{aggr} = BW_{mc} \times N$. The total number of MRRs is $2 \times 2 \times 2 \times N \times m$ because each TX or RX lane requires m MRRs. OCM has two unidirectional links; each link needs both TX and RX lanes, and these lanes are located in both Endpoint controllers and Master controllers.

4. Evaluation

Before showing our evaluations of OCM system-level performance (in Section 4.2), and SiP link energy estimation (in Section 4.3), we describe our methodology for evaluation.

4.1. Evaluation methodology

OCM performance. To evaluate system-level performance, we implement OCM architecture in the ZSIM simulator [67]. Table 1 shows the configuration of our baseline system (a server processor), the two DDR4 memory configurations used in our evaluation (MemConf1 and MemConf2), the latencies of an OCM disaggregated system, and the latencies of a disaggregated system using 40G PCIe NICs. MemConf1 has 4 DDR4 memory channels as in conventional server processors, and MemConf2 has a single DDR4 memory channel, and an in-package DRAM cache on the processor side.

The goal of the DRAM cache is to reduce the optical disaggregation overhead [78], which can have a significant performance impact in memory-bound applications. Our DRAM cache resembles the Banshee DRAM cache [77] that tracks the contents of the DRAM cache using TLBs and page table entries, and replaces pages with a frequency-based predictor mechanism. We configure our

Table 2
Evaluated SPEC06 & SPEC17 benchmark mixes.

SPEC06	mix1	soplex_1, h264, gobmk_3, milc, zeusm, bwaves, gcc_1, omnetpp
	mix2	soplex_1, milc,povray, gobmk_2, gobmk_3, bwaves, calculix, bzip2_2
	mix3	namd, gromacs, gamess_1, mcf, lbm, h264_2, hmmer, xalancbmk
SPEC17	mix1	exchange2, cactus, gcc_2, imagick, fotonik3d, xalancbmk, xz_2, lbm
	mix2	gcc_1, nab, lbm, leela, mcf, xz_1, sroms, omnetpp
	mix3	xalancbmk, nab, cactus, mcf, imagick, xz_1, fotonik3d, deepjeng

Table 3
Optical and electrical models for OCM SiP link devices.

Parameter	Design Criteria	Details	Ref.
Optical power	20 dBm	Max. aggregated	
Center wavelength	1.55 μm		
Laser	10% and 30%	Laser wall-plug efficiency	[24]
Waveguide loss	5 dB/cm 0.02 dB/bend	fabrication roughness waveguide bend loss	[35]
Coupler loss	1 dB	off-chip coupler	[27]
Modulator	Q = 6500	Ring resonator Q factor	[62]
	ER = 10 dB	MRR extinction rate	
	65 fF	Junction capacitance	
	−5 V	Maximum drive voltage	
	1 mW	Thermal-tuning power/ring	[6]
Mod. mux and receiver demux	MRR power penalties	Crosstalk model	[9]
Photodetector	1.09 A/W	Current per opt. power	[32]
Modulator driver	28 nm	Semicond. tech. for OOK-WDM	[62]
SERDES power model	28 nm	Semicond. tech.	[62]
Digital receiver	28 nm	Semicond. tech. for OOK-NRZ	[62]
Element positioning	100 μm	Modulator padding	

DRAM cache to have the same operation latency as commodity DDR4 memory.

We calculate the SERDES link latency values for the upcoming years. We estimate the minimum at 10 cycles, which assumes 3.2 ns serialization/deserialization latency [42]. We use 340 cycles (113 ns) maximum latency reported in a previously demonstrated optical interconnection system [63]. We simulate rack distances of 2 m, 4 m, and 6 m with a 5 ns/m latency [1], which translates into 30, 60, and 90 cycles latency in our system.

For a Network Interface Card (NIC) based system configuration, we evaluate a scenario using a 40G PCIe NIC. The 40G bitrate per link is similar to commercially available devices, e.g., Infiniband HDR [52]. We consider a NIC latency of 1050 cycles (350 ns) [1], a realistic NIC-through-PCIe latency is in the order of μs [57] (e.g., ≈0.5 μs for Infiniband HDR). We dimension both 40G PCIe NIC links and OCM based on latency penalties in our system-level simulator. Both have enough bandwidth to support the DRAM memory used in our evaluation. Notice that the main differences from a commercially available optical 40G to OCM are: (1) the number of fiber links required to sustain the memory bandwidth of the memory pool, (2) the energy-per-bit consumption, and (3) the characteristics of the photonic devices. We custom-tailored the SiP links required to disaggregate memory (see Section 4.3).

We evaluate the system-level performance of OCM with applications from six benchmark suites representing three workload scenarios: (1) multi-program, (2) multithread, and (3) multinode.

(1) The first scenario for **multi-programmed workloads** depicts a mix of benchmark applications executing concurrently. We used SPEC06 [38] with Pinpoints (warmup of 100 million instructions, and detailed region of 30 million instructions), and SPEC17 [23] *speed* with reference inputs. Table 2 lists the content of the used SPEC benchmark mixes.
(2) The second scenario represents **multithreaded workloads,** i.e., a single application with multiple threads executing on a multicore processor. We used PARSEC [18] with *native* inputs, SPLASH2 [17] with *simlarge* inputs, and GAP graph benchmarks [13] executing 100 billion instructions with the *Web* graph input, and 30 billion instructions with the *Urand* graph input. The *Urand* input has very poor locality between graph vertices compared to the *Web* input. We also used five MPI applications from the NAS Parallel Benchmark (NPB) [11] with class C inputs, executing 100 billion instructions for FT, MG and CG, and whole running IS and EP. For these MPI applications, we split them over 8 processes, 1 process per core.
(3) For **multinode workloads**, a single application executes on multiple nodes of a computer cluster. We considered a computer cluster composed of eight nodes with the same characteristics as shown in Table 1. We used the same MPI applications from NPB that we used on the multithreaded workload scenario, but instead of running all eight processes on the same node, we distribute them among eight nodes, one process per node. Our goal is to evaluate workload executions with OCM considering the network overhead. As ZSIM lacks the simulation of the network layer, we used a two-step simulation approach to account the node-to-node communication overhead. On the **first step**, we execute the MPI application on ZSIM, considering only one of the eight created processes. We considered the first process (rank 0) and ignored the others (ranks 1-7). Which process to choose not to ignore is irrelevant because of the evenly distributed workload among the processes of the used benchmarks. This execution allows measuring performance, i.e., speedup, without node-to-node communication overhead of a single process while running on a single node with OCM. On the **second step**, we modeled a computer cluster with the SimGrid simulator [25] and tuned each node's processing power according to the speedup obtained in the first step. We executed the MPI benchmarks on our tuned cluster model using the SimGrid MPI interface [29], obtaining a performance measurement that considers the network overhead. The cluster model we used resembles the topology from a local computer cluster named Kahuna, where the eight nodes are connected via a Mellanox SX6025 switch. We measured the bandwidth and latency on node to node communication in Kahuna using two kernels, *osu_latency* and *osu_bw*, both from the OSU benchmark suite [46]. In the first step of our two-step simulation approach, we executed 9 billion instructions for IS, 24 billion instructions for FT, 8 billion instructions for CG, 7 billion instructions for MG and 21 billion instructions for EP. On the second step, all the applications executed without any limitation on the number of instructions.

Table 4 summarizes the measured memory footprint values for all the benchmarks used in our evaluation, measured using the Massif tool from Valgrind [56]. The measured memory footprint of MPI applications from NPB is for the application code only, and

Table 4
Measured memory footprints.

SPEC06 [38]	*MIX1*: 2.2 GB, *MIX2*: 3.1 GB, *MIX3*: 2.4 GB
SPEC17 [23]	*MIX1*: 19.9 GB, *MIX2*: 36.4 GB, *MIX3*: 34.7 GB.
PARSEC [18]	*canneal*: 716.7 MB, *streamcluster*: 112.5 MB, *ferret*: 91.9 MB, *raytrace*: 1.3 GB, *fluidanimate*: 672 MB
SPLASH [17]	*radix*: 1.1 GB, *fft*: 768.8 MB, *cholesky*: 44.2 MB, *ocean_ncp*: 26.9 GB, *ocean_cp*: 891.8 MB.
GAP [13]	*Urand* graph: 18 GB, *Web* graph: 15.5 GB
NPB [11] Class C	*Integer Sort (IS)*: 2.3 GB, *Fast Fourier Transform (FFT)*: 7.2 GB, *Conjugate Gradient (CG)*: 1.2 GB, *Multi Grid (MG)*: 3.5 GB, *Embarrassingly Parallel (EP)*: 34.4 MB

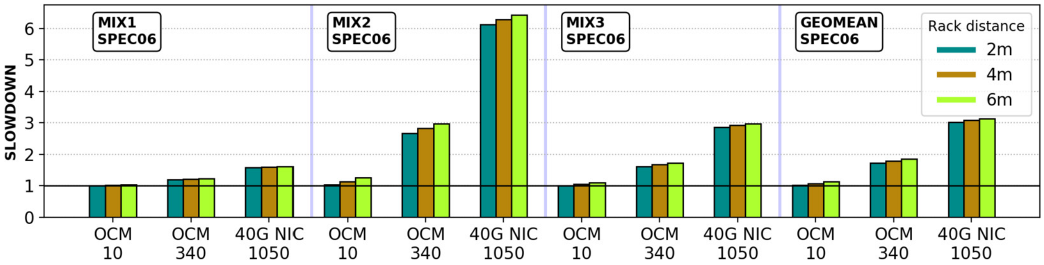


Fig. 6. Slowdowns of OCM and 40G NIC-based disaggregated systems, compared to a non-disaggregated baseline with MemConf1, for three randomly-selected mixes of SPEC06 benchmarks (lower is better).

it does not include the memory footprint from the MPI process manager.

We also used a synthetic benchmark, that resembles the *copy* kernel from the STREAM benchmark [51], to obtain the OCM memory roofs, based on the memory roof concept of the Roofline model [73].

SiP link energy-per-bit. To evaluate the interconnection between processor and memory as a point-to-point SiP link, we use PhoenixSim [66] with parameters extracted from state-of-the-art optical devices [9,62,8]. PhoenixSim considers the physical features of the optical devices and their digital semiconductor drivers to evaluate many SiP link energy-per-bit cases in terms of: (1) the required number of optical wavelengths (λ), and (2) the bit rate per λ. Table 3 lists OCM optical devices and their main characteristics used in our simulation model.

4.2. System-level evaluation

Multiprogrammed evaluation. Fig. 6 shows the slowdown of OCM and 40G NIC-based disaggregated memory systems with MemConf1, compared to a non-disaggregated MemConf1 baseline, for three mixes of SPEC06 benchmarks (Table 2). Notice that a system with disaggregated main memory is expected to perform worse than the non-disaggregated baseline, because of the extra latency introduced by the interconnects (see Eq. (1)).

We make two observations. First, the 40G NIC-based system is significantly slower than our OCM system, even though the Ethernet configuration we evaluate is very optimistic (350 ns average latency, equivalent to 1050 cycles in Table 1). OCM is up to 5.5× faster than 40G NIC for the minimum SERDES latency, and 2.16× faster for the maximum SERDES latency. Second, the results show the feasibility of low-latency disaggregation with OCM as future SERDES optimizations become available. OCM has an average slowdown (across all rack-distances) of only 1.07× compared to the baseline with a SERDES latency of 10 cycles, and 1.78× average slowdown with a SERDES latency of 340 cycles.

Fig. 7 shows the slowdown of OCM and a 40G NIC-based disaggregated system; both compared to a non-disaggregated baseline. We used MemConf2 and evaluated the *MIX2* of SPEC06, which obtained the highest slowdown in our previous experiment with MemConf1 (as shown in Fig. 6). The NIC-based disaggregation shows an improvement of 2.54× when using the DRAM cache. Although it presents the most significant improvement compared to OCM, it still provides a higher slowdown than OCM (with and without DRAM cache). We observe the advantage of OCM over NIC-based disaggregation, even when using DRAM cache.

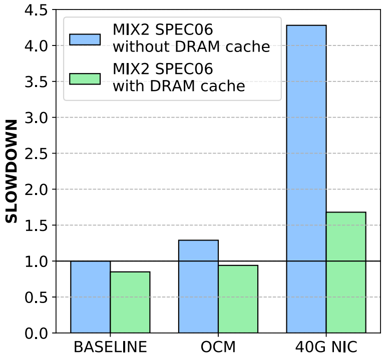


Fig. 7. Slowdown of OCM (150 cycles SERDES time) and 40G NIC-based disaggregated systems (4 meters distance), compared to a non-disaggregated baseline with MemConf2, for the randomly-selected *MIX2* of SPEC06 (lower is better).

Fig. 8 shows the speedup of a disaggregated OCM system (green bars) compared to a non-disaggregated baseline, both configured with MemConf1. Fig. 8 also shows the speedup of OCM with MemConf2 (red bars), and the speedup of a non-disaggregated system with MemConf2 (blue bars), both compared to a MemConf2 baseline without a DRAM cache and without disaggregation. OCM has a conservative SERDES latency of 150 cycles, and a distance of 4 m.

Fig. 8 (left) shows the results for SPEC17 mixes (see Table 2). We make two observations. First, the average slowdown of OCM without DRAM cache (green bars) is 17%, which is in the same order as the SPEC06 results (Fig. 6). Second, with a DRAM cache, the performance of the OCM disaggregated system (red bars), and the non-disaggregated system (blue bars) is very close, as the memory intensity of these benchmarks is not very high. As expected, the performance of the disaggregated system is always lower than the non-disaggregated system.

Multithreaded evaluation. Fig. 8 (right) shows the results for multithreaded graph applications. We make two observations. First, the maximum slowdown of OCM without a DRAM cache (green bars) is up to 45% (*pagerank* (*PR*)), which is in the same order as SPEC17 results, despite the *Web* input having very high locality. The extra latency of the OCM disaggregated system has a clear negative effect on performance. Second, graph workloads dramat-

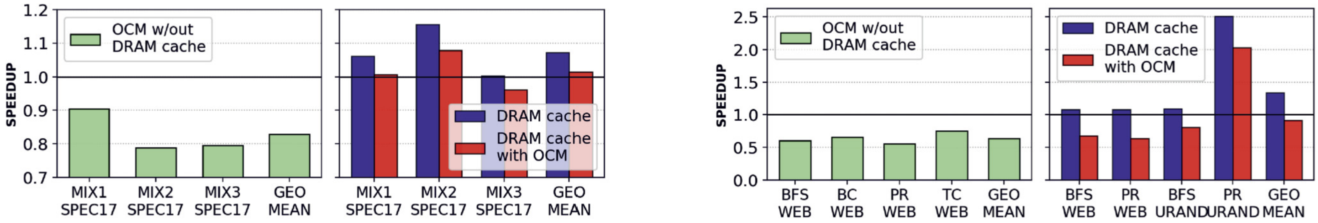


Fig. 8. OCM speedup results with 4 m distance and a SERDES latency of 150 cycles (higher is better), compared to a disaggregated baseline, with or without a DRAM cache. Left: Speedup for SPEC17. Right: Speedup for GAP [13] graph benchmarks.

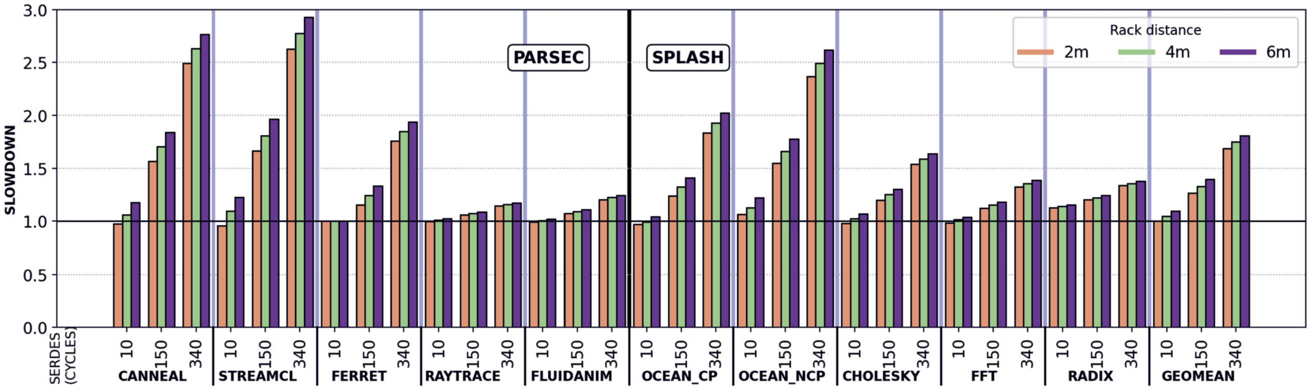


Fig. 9. OCM slowdown compared to the baseline for PARSEC and SPLASH2 benchmarks (lower is better).

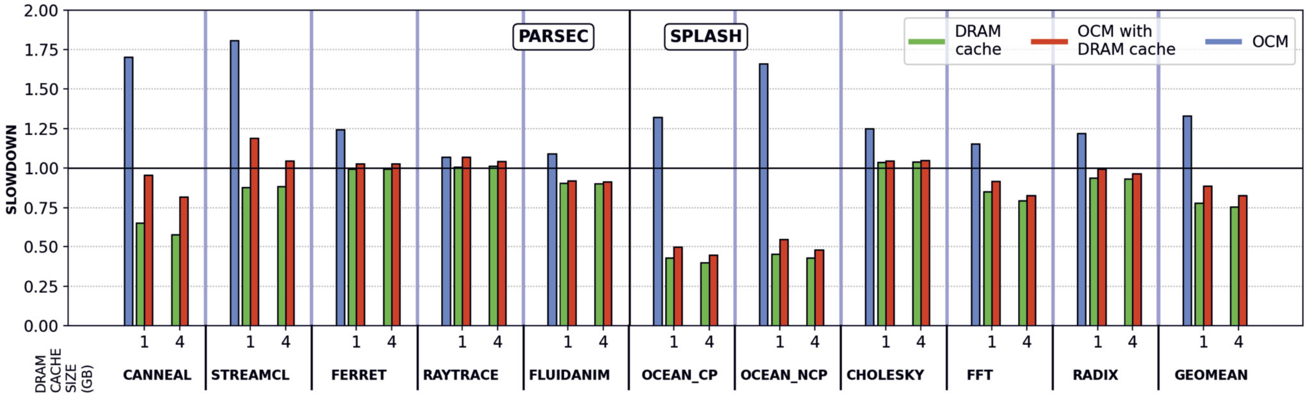


Fig. 10. OCM slowdown results with a DRAM cache for PARSEC and SPLASH2 benchmarks on a system with 150 SERDES latency and 2 m rack distance (lower is better).

ically benefit from using a DRAM cache (red and blue bars), e.g., *PR* with *Urand* input shows a speedup of 2.5× compared to the baseline, which is 50% lower speedup than the non-disaggregated scenario. We believe that the performance degradation of OCM with DRAM cache is still reasonable. However, adding a DRAM cache also brings new challenges that need further investigation in a disaggregated setting, such as page replacement mechanisms and caching granularity [77,44,53,76,54,75,64,39].

Fig. 9 shows the slowdown of OCM compared to the baseline, using MemConf1 with PARSEC and SPLASH2 benchmarks. We show results for the memory-bound benchmarks only. We also test other compute-bound benchmarks (not shown in the figure) that show less than 5% slowdown. We make three observations. First, with the lower bound SERDES latency (10 cycles) and lowest rack distance (2 m), applications such as *streamcluster*, *canneal* and *cholesky*, experience an average 3% speedup. This small improvement occurs as a result of T_{mem} reduction (tBL related) due to splitting of a cache line into two DIMMs. Second, the slowdowns increase slightly as distance increases. Third, with large rack-distance and maximum SERDES latency, the slowdown is significant. The highest slowdown measured is 2.97× for *streamcluster* at 6 m and 340 SERDES cycles; the average slowdown is 1.3× for SPLASH2 and 1.4× for PARSEC.

Fig. 9 shows the slowdown of OCM compared to the baseline, using MemConf1 with PARSEC and SPLASH2 benchmarks. We show results for the memory-bound benchmarks only. We also test other compute-bound benchmarks (not shown in the figure) that show less than 5% slowdown. We make three observations. First, with the lower bound SERDES latency (10 cycles) and lowest rack distance (2 m), applications such as *streamcluster*, *canneal* and *cholesky*, experience an average 3% speedup. This small improvement occurs as a result of T_{mem} reduction (tBL related) due to splitting of a cache line into two DIMMs. Second, the slowdowns increase slightly as distance increases. Third, with large rack-distance and maximum SERDES latency, the slowdown is significant. The highest slowdown measured is 2.97× for *streamcluster* at 6 m and 340 SERDES cycles; the average slowdown is 1.3× for SPLASH2 and 1.4× for PARSEC.

Fig. 10 shows the slowdown of OCM with DRAM cache in a conservative scenario, i.e., medium rack distance (4 m) and SERDES latency (150 cycles), using MemConf2 with memory-bound benchmarks of PARSEC and SPLASH2. We additionally explore MemConf2

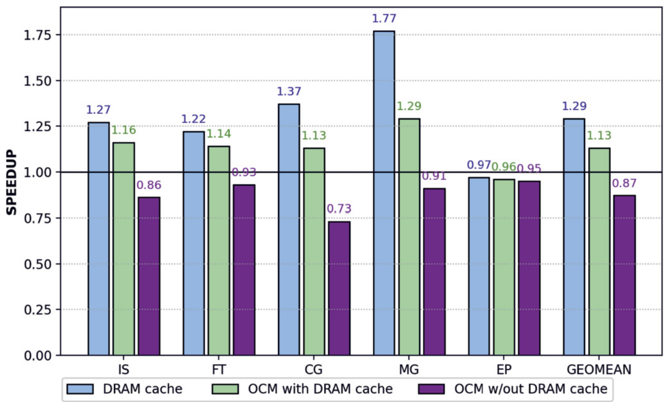


Fig. 11. Speedup of the usage of DRAM cache (with and without OCM) and OCM compared to the baseline for NPB benchmarks (higher is better), using eight processes all on a single node.

with a 1 GB DRAM cache. We make two observations. First, using a DRAM cache reduces the latency overhead caused by OCM disaggregation. The average slowdown is 0.89× for OCM with a 1 GB DRAM cache and 0.83× for OCM with a 4 GB DRAM cache. However, OCM with a 1 GB and 4 GB DRAM cache performs faster compared to the 1.33× slowdown of an OCM system without DRAM cache. Second, OCM can also benefit from a lowersized DRAM cache depending on the workload memory footprint and access behavior. The *ocean_ncp* and *streamcluster* benchmarks have the highest slowdown with OCM. Both benchmarks exhibit a similar performance improvement using a 1 GB DRAM cache compared to a 4 GB DRAM cache. The *ocean_ncp* benchmark performs only 3% slower in an OCM system with a 1 GB DRAM cache than an OCM system with a 4 GB DRAM cache. While executing *ocean_ncp* benefits from a 1 GB DRAM cache because of its large memory footprint of ≈ 27 GB, benchmarks with low memory footprint such as *cholesky* (≈ 44 MB) does not benefit from a DRAM cache due to the TLB overhead. Using a smaller DRAM cache can help reduce area and electrical energy consumption on an OCM system's processing side.

Fig. 11 shows the OCM speedup using MemConf2 with the NPB benchmarks. These results present a maximum slowdown of 27% with the CG benchmark, while with the other benchmarks, the performance loss stays within 14%. Using a DRAM cache exhibits a performance improvement on all benchmarks except on EP. This occurred due to the extremely low memory footprint from this benchmark, as depicted in Table 4.

Multinode evaluation. Fig. 12 shows the results of a multinode scenario, obtained through a two-step simulation method described in Section 4.1, running NPB benchmarks on eight different nodes, one process per node. This multinode execution case exhibits a reduced variation in performance compared to the multithreaded workloads. The difference between the average performances of the three configurations stays within 7%. This is due to two factors that diminish the impact of the memory system. The first factor is that the memory footprint of each benchmark was also split among the eight nodes. Considering the 34.4 MB memory footprint from EP, the eighth part is around 4.3 MB, which entirely fits the 8 MB L3 Cache. The second factor is that network performance has a significant impact on these applications. We considered a node to node latency of 8 microseconds (around 24000 cycles). As a comparison, our worst SERDES overhead consideration for OCM was 113 nanoseconds (340 cycles). Depending on how the applications can overlap computation and network communication, the performance bottleneck may shift from the memory system to the network performance.

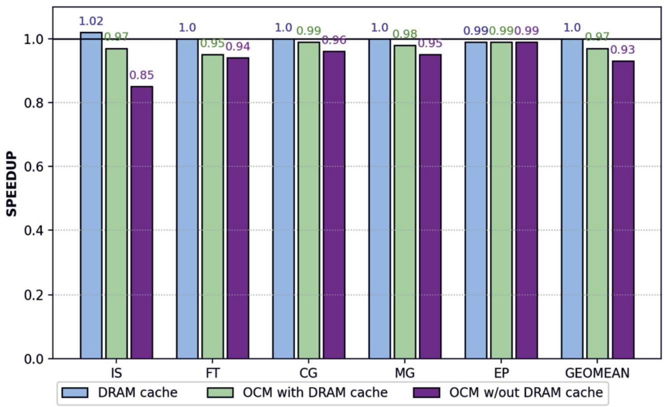


Fig. 12. Speedup of the usage of DRAM cache (with and without OCM) and OCM compared to the baseline for NPB benchmarks (higher is better), using eight processes distributed among eight nodes (1 process per node).

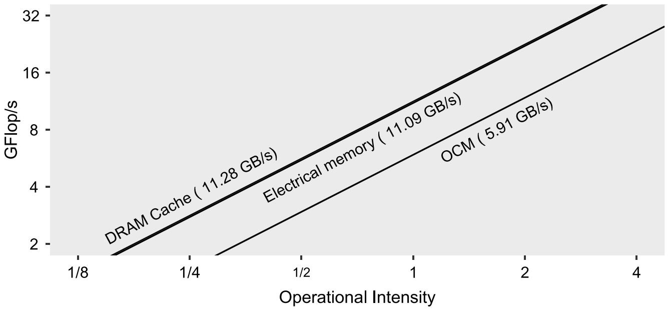


Fig. 13. Memory roof, using MemConf2, of DRAM cache, electrical memory and OCM using a single threaded application.

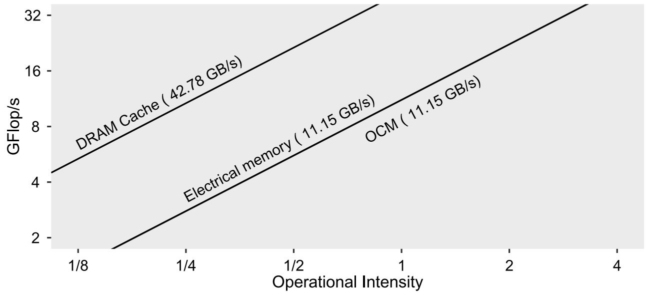


Fig. 14. Memory roof, using MemConf2, of DRAM cache, electrical memory and OCM using a multithreaded application (8 threads).

Memory roofs. Fig. 13 and 14 present the memory roofs obtained with MemConf2 configuration. Fig. 13 represents the memory roofs obtained from a single core (1 thread), while Fig. 14 represents the memory roofs obtained from a multithreaded execution (8 threads, one per core). OCM with DRAM cache shows an increase in bandwidth performance according to the bandwidth demand. They increase their bandwidth on the multithreaded case (3.79× for DRAM cache and 1.88× for OCM), while the electrical memory exhibits no variation on its performance. With the higher bandwidth demand from the multithreaded application, OCM compares to the electrical memory in bandwidth performance, and the DRAM cache exhibits an advantage over the electrical memory. This behavior shows that OCM can achieve similar performance to the electrical memory bandwidth on the best case (cache-friendly memory accesses). Concurrently, a DRAM cache may become only an additional level on the memory hierarchy, without any gain of performance, on lower bandwidth demands.

We conclude that OCM is very promising because of its reasonably low latency overhead (especially with the use of a DRAM

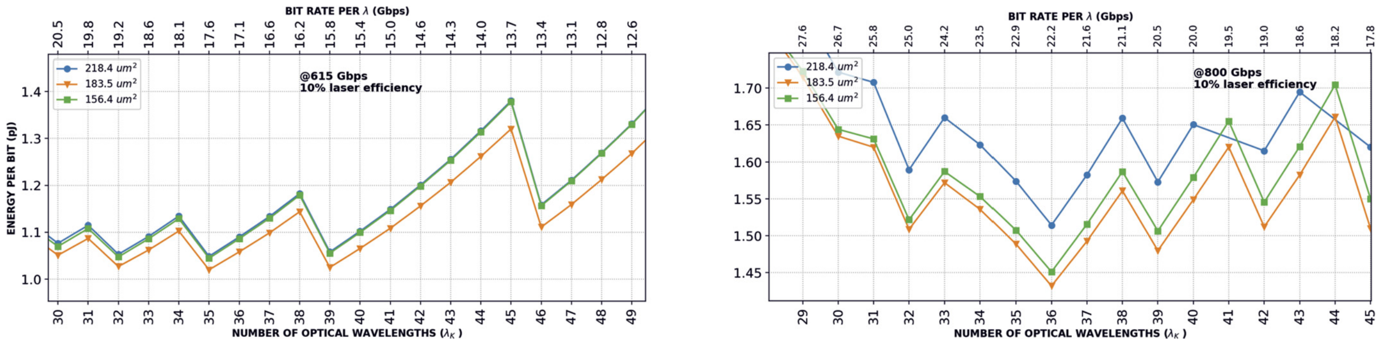


Fig. 15. SiP link energy-per-bit using a laser with 10% efficiency. Left: at 615 Gbps bandwidth, Right: at 800 Gbps bandwidth.

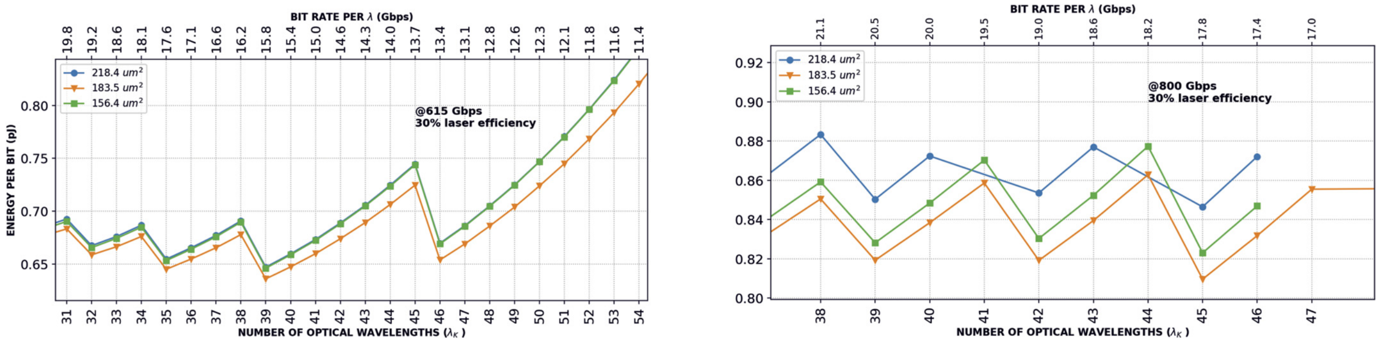


Fig. 16. SiP link energy-per-bit using a laser with 30% efficiency. Left: at 615 Gbps bandwidth, Right: at 800 Gbps bandwidth.

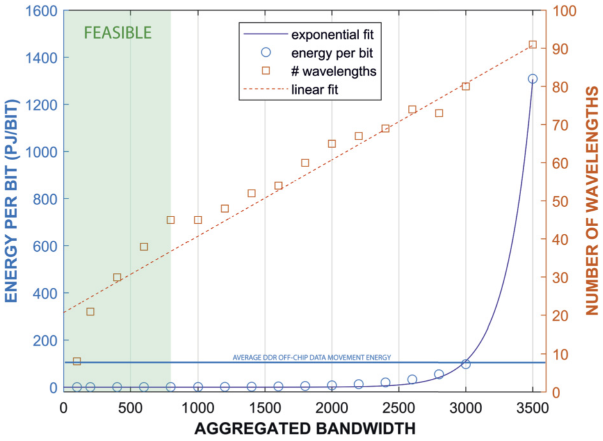


Fig. 17. Minimum SiP link energy consumption and number of wavelengths tendency as a function of the aggregated bandwidth. Results that are currently feasible are highlighted. Aggregated bandwidth is measured in Gbps.

cache), and the flexibility of placing memory modules at large distances with small slowdowns.

4.3. SiP link evaluation

We evaluate the energy and area consumption of the SiP link to allow the system designer to make tradeoffs about the use of SiP devices in the computing system. It is enough to consider a single unidirectional modeled SIP link using PhoenixSim [66] with the input parameters shown in Table 3 to estimate the energy efficiency. We estimate the minimum energy-per-bit consumption and the required number of MRRs for our model, given an aggregated optical bandwidth equivalent to the bandwidth required by DDR4-2400 DRAM memory.

A single DDR-2400 module requires 153.7 Gbps bandwidth [40]. 4 memory channels, with 2 DIMMs per channel in lockstep, require ∼615 Gbps/link. OCM's maximum feasible bandwidth (while remaining CMOS compatible) is 802 Gbps using the parameters in Table 3. More advanced modulation formats, such as PAM4 [69], can be used to achieve higher aggregated bandwidth. Figs. 15 and 16 show the energy-per-bit results (y-axis), and the aggregated bandwidth. The aggregated link bandwidth is the multiplication of the number of λ (bottom x-axis values), and the aggregated bitrate (top x-axis values), i.e., a higher number of λs implies a lower bitrate per λ. We consider three feasible and efficient MRR sizes in our model: 156.4 (green), 183.5 (orange), and 218.4 μm^2 (blue).

From Table 3, we have considered two cases of lasers, 10%-efficient epitaxially-grown integrated laser, which is widely used in the SiP industry [41], and a state-of-the-art laboratory laser with a nominal efficiency of 30% [24] to demonstrate that improvement of optical features of a single device affects our SiP link energy estimation significantly. Our previous work [34] used off-chip lasers, while in this work, we report results using heterogeneous integration of lasers on silicon [41] and reducing the number of couplers per link.

As shown in Fig. 15, in OCM with 615 Gbps links using lasers with 10% efficiency, the minimum energy consumption overhead compared to the electrical memory system is 1.02 pJ/bit for 35 optical wavelengths (λ) per link, each λ operating at 17.57 Gbps. The SiP link with the 30% laser efficiency achieved and energy consumption of 0.64 pJ/bit with 39 λ's, each operating at 15.8 Gbps, as depicted in Fig. 16.

The energy evaluation of the maximum feasible bit rate of a SiP link is also presented, with an aggregate bandwidth of 800 Gbps. The minimum energy consumption is 1.43 pJ/bit for 36 λ's per link, each λ operating at 22.22 Gbps using a laser with 10% efficiency. The SiP link that has lasers with an efficiency of 30% showed energy consumption of 0.81 pJ/bit for 45 λ's, each operating at 17.77 Gbps.

We make three observations from Figs. 15 and 16. First, as in electrical systems, it is expected that a higher bandwidth per link increases the link energy-per-bit consumption. However, the optical energy-per-bit is lower compared to electrical systems. For reference, the energy-per-bit of a DDR4-2667 DRAM module is 39 pJ [61]; thus, the energy-per-bit caused by an additional SiP link in the memory subsystem is less than 5%. Second, there is a non-smooth behavior on the energy-per-bit curves due to the energy consumption model of the optical receiver, which depends on the data rate. In our model, we set the photodetector current to a minimum value. As the data rate increases, the received signal becomes less distinguishable from noise. Our model forces the photocurrent to step into a new minimum value to avoid this, causing the repeated decrease and increase of the energy-per-bit values [10]. For both SiP links, the 183.5 μm^2 rings consume the lowest energy. The estimated area overhead is 51.4E-3 mm^2 with 2 × 615 Gbps links, and 57.3E-3 mm^2 with 2 × 802 Gbps links. In our case study of 4 DDR4 memory channels, OCM uses fewer physical interconnects (optical fibers) than 40G PCIe NIC links (copper cables). In other words, to achieve the required aggregated link bandwidth, we require 2 optical fibers with OCM or 30 copper cables with 40G PCIe NICs.

From Fig. 17, we make three observations: (i) with the current setup shown in Table 3, the energy per bit grows exponentially for aggregated bandwidths above 2500 Gbps and above the average DDR off-chip data movement energy at 3000 Gbps; (ii) the most energy-efficient number of wavelengths grows approximately linearly with the aggregated bandwidth; (iii) demonstrated fabrication feasibility is highlighted up to 800 Gbps, and the region to the right is estimated with PhoenixSim, yet currently not feasible. Accordingly, we can say that OCM must include new and more efficient optical device models to grow beyond the 3000 Gbps mark. Furthermore, this growth must include optimizing MRRs -or similar devices- to either multiply the number of possible wavelengths or raise the bitrate per wavelength. The physics of this growth will be addressed in the Scaling of optical devices section.

We conclude that a bidirectional SiP link, formed by two unidirectional links using current SiP devices, can fit the bandwidth requirements of commodity DDR4 DRAM modules. OCM incurs a low energy overhead of only 10.2% compared to a non-disaggregated DDR4 DRAM memory (the energy consumption of current DDR4 DRAM technology is ~ 10 pJ/bit [69]).

Scaling of optical devices. Silicon has an indirect energy bandgap in near infrared frequencies. Thus, active devices such as lasers or photodetectors cannot be fabricated using only a single material. Optical active devices on silicon are moving towards monolithic integration of other energy efficient materials. Namely, heterogeneous integration of III-V-group materials epitaxially grown on silicon [71], and the integration of two-dimensional [28] and one-dimensional [47,31] materials to improve modulation, amplification, switching and photodetection. These techniques enable more efficient lasers, improve the sensitivity of photodetectors and reduce modulators driving power. These improvements can affect directly the SiP link estimated energy-per-bit, as shown in the 10% versus 30% integrated laser example discussed previously on this section.

Scaling of lasers on a silicon platform has advanced from epitaxially grown III-V quantum wells, in the scale of tens of nanometers, towards epitaxially grown quantum dots. Although the physics of transversal confinement of light does not change over the years, the cavity length of the device has been shrunk down and its efficiency has improved [31]. In this work we considered as feasible an epitaxially grown on silicon 10% and 30% quantum-well lasers.

Photodetectors require active materials with an direct energy bandgap in the infrared. Germanium has been widely used for this purpose. However, defending the tendency of new materials for smaller footprint, quantum dot photodetors with III-V materials, and the use of two-dimensional materials [48] are also relevant in the literature. In this work, we considered a Germanium optimized photodetector with a high sensitivity [32] which is feasible and CMOS compatible on a high scale.

Lastly silicon modulators are important features of the SiP link, and the ones that require the biggest footprint. Traditionally an electrooptic effect is induced on silicon by doping the MRR optical waveguide slightly [7,6]. This is the modulation method we use in this work. However, as seen in Fig. 17, although there is a predictable linear evolution of the required number of λ's, the energy-per-bit grows exponentially beyond the terahertz aggregated bandwidth, making it nonviable to use all configurations as they are presented in the future. However, it is enough to include new devices in the PhoenixSim platform to estimate a new path for the growth of SiP links in OCM. The reader should also note that works with hybrid Silicon photonics and 2D semiconductor monolayers were demonstrated [28].

5. Related work

To our knowledge, this is the first work to propose an optical point-to-point disaggregated main memory system for modern DDR memories that (1) evaluates a SiP link with state-of-the-art optical devices, (2) demonstrates that OCM incurs only 10.7% energy overhead compared to a non-disaggregated DDR4 DRAM memory, and (3) quantifies the performance implications of the proposed optical links at the system level on commonly-used application workloads.

Brunina et al. [20,22] introduce the concept of optically connected memory in a mesh topology connected with optical switches. Both works propose point-to-point direct access to the DRAM modules using Mach Zender modulators. These works motivate our study in optically connected memory. Brunina et al. [21] also experimentally demonstrate that microring modulators can be used for optically connecting DDR2 memory. Our work builds on [21] to design the microring modulators used in our SiP links. There are several recent works [9,69,10] that propose analytical models of the microring used in our SiP links. Anderson et al. [5] extend the work of Brunina et al. [20,22,21] to experimentally demonstrate the optical switches using FPGAs for accessing memory.

These prior works [5,21,22,20] are all experimental demonstrations to show photonic capabilities. In contrast, our work addresses three important questions prior work does not: (1) How many optical devices (i.e., MRRs) do we need for current DDR technology? (Section 4.3), (2) What is the energy and area impact on the system? (Section 4.3), and (3) How does the processor interact with a disaggregated memory subsystem (system-level)? (Section 4.2).

Some other works, such as [72,79], point out, without evaluation, that existing disaggregation protocols (i.e., PCIe and Ethernet) could lead to high-performance loss. Our work uses system-level simulation to measure the performance overhead of such protocols. We propose to alleviate the optical serialization overhead by using the DDR protocol (Section 3.1). As photonic integration improves, we believe that the optical point-to-point links will become the main candidate for interconnecting disaggregated memory. With our PhoenixSim [66] model, we explore the design of SiP links based on DDR requirements. Our proposal can be used to improve existing PCIe+photonics works, such as [74].

Yan et al. [74] propose a PCIe Switch and Interface Card (SIC) to replace Network Interface Cards (NIC) for disaggregation. SIC is composed of commercial optical devices and is capable of interconnecting server blades in disaggregated data centers. The evaluated SIC shows a total roundtrip latency up to 426 ns. In contrast, the scope of our work is point-to-point DDR DRAM disaggregation without PCIe or other additional protocols.

Other related prior works (1) explore silicon photonics integration with a many-core chip in an optical network-on-chip design [12], (2) propose the design of a DRAM chip with photonic inter-bank communication [14], (3) present an optoelectronic chip for communication in disaggregated systems with 4-λ and an energy consumption of 3.4 pJ/bit [4], (4) evaluate a memory disaggregation architecture with optical switches focusing on re-allocation mechanisms [78], (5) analyze the cost viability of optical memory disaggregation [1], and (6) evaluate memory disaggregation using software mechanisms with high latency penalties in the order of μs [36]. Unlike [78,4,14,1,36], our work evaluates (1) system performance with real applications, (2) the design of the SiP link for DDR DRAM requirements, and (3) SiP link energy for a disaggregated memory system.

6. Conclusions

We propose and evaluate Optically Connected Memory (OCM), a new optical architecture for disaggregated main memory systems, compatible with current DDR DRAM technology. OCM uses a Silicon Photonics (SiP) platform that enables memory disaggregation with low energy-per-bit overhead. Our evaluation shows that, for the bandwidth required by current DDR standards, OCM has significantly better energy efficiency than conventional electrical NIC-based communication systems, and it incurs a low energy overhead of only 10.7% compared to DDR DRAM memory. Using system-level simulation to evaluate our OCM model on real applications, we find that OCM performs 5.5× faster than a 40G NIC-based disaggregated memory. We conclude that OCM is a promising step towards future data centers with disaggregated main memory.

7. Acknowledgments

This work is supported by the LPS Advanced Computing Systems (ACS) Research Program contract HD TAT DO 7 (HT 15-1158)), the Department of Energy (DOE) Small Business Innovation Research (SBIR) ASCR Program (contract DE-SC0017182), the São Paulo Research Foundation (FAPESP) (fellowships 2013/08293-7 and 2014/01642-9), CAPES (fellowships 2013/08293-7 and 88882.329108/2019-01), CNPq (fellowships 438445/2018-0, 309794/2017-0 and 142016/2020-9), and CONCYTEC-PROCIENCIA under the call E041-01 [contract number 015-2018-FONDECYT/BM].

CRediT authorship contribution statement

Jorge Gonzalez: Conceptualization, Investigation, Software, Writing – review & editing. **Mauricio G. Palma:** Investigation, Methodology, Software, Writing – review & editing. **Maarten Hattink:** Investigation, Methodology, Software. **Ruth Rubio-Noriega:** Investigation, Software, Writing – review & editing. **Lois Orosa:** Investigation, Supervision, Writing – review & editing. **Onur Mutlu:** Supervision, Writing – review & editing. **Keren Bergman:** Funding acquisition, Project administration, Supervision. **Rodolfo Azevedo:** Funding acquisition, Project administration, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Abali, R.J. Eickemeyer, H. Franke, C.S. Li, M.A. Taubenblatt, Disaggregated and optically interconnected memory: when will it be cost effective?, arXiv:1503.01416 [cs.DC], 2015.

[2] N.C. Abrams, Q. Cheng, M. Glick, M. Jezzini, P. Morrissey, P. O'Brien, K. Bergman, Silicon photonic 2.5 d multi-chip module transceiver for high-performance data centers, J. Lightwave Technol. 38 (2020) 3346–3357.

[3] P.P. Absil, P. De Heyn, H. Chen, P. Verheyen, G. Lepage, M. Pantouvaki, J. De Coster, A. Khanna, Y. Drissi, D. Van Thourhout, et al., Imec iSiPP25G silicon photonics: a robust CMOS-based photonics technology platform, in: Silicon Photonics X, 2015.

[4] M.S. Akhter, P. Somogyi, C. Sun, M. Wade, R. Meade, P. Bhargava, S. Lin, N. Mehta, WaveLight: a monolithic low latency silicon-photonics communication platform for the next-generation disaggregated cloud data centers, in: High-Performance Interconnects (HOTI), 2017.

[5] E.F. Anderson, A. Gazman, Z. Zhu, M. Hattink, K. Bergman, Reconfigurable silicon photonic platform for memory scalability and disaggregation, in: Optical Fiber Communication (OFC), 2018.

[6] M. Bahadori, A. Gazman, N. Janosik, S. Rumley, Z. Zhu, R. Polster, Q. Cheng, K. Bergman, Thermal rectification of integrated microheaters for microring resonators in silicon photonics platform, J. Lightwave Technol. (2018).

[7] M. Bahadori, M. Nikdast, S. Rumley, L.Y. Dai, N. Janosik, T. Van Vaerenbergh, A. Gazman, Q. Cheng, R. Polster, K. Bergman, Design space exploration of microring resonators in silicon photonic interconnects: impact of the ring curvature, J. Lightwave Technol. 36 (2018) 2767–2782.

[8] M. Bahadori, R. Polster, S. Rumley, Y. Thonnart, J. Gonzalez-Jimenez, K. Bergman, Energy-bandwidth design exploration of silicon photonic interconnects in 65 nm CMOS, in: Optical Interconnects (OI), 2016.

[9] M. Bahadori, S. Rumley, D. Nikolova, K. Bergman, Comprehensive design space exploration of silicon photonic interconnects, J. Lightwave Technol. (2016).

[10] M. Bahadori, S. Rumley, R. Polster, A. Gazman, M. Traverso, M. Webster, K. Patel, K. Bergman, Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing, in: Design, Automation and Test in Europe (DATE), 2017.

[11] D.H. Bailey, E. Barszcz, J.T. Barton, D.S. Browning, R.L. Carter, L. Dagum, R.A. Fatoohi, P.O. Frederickson, T.A. Lasinski, R.S. Schreiber, et al., The nas parallel benchmarks summary and preliminary results, in: Supercomputing'91: Proceedings of the 1991 ACM/IEEE Conference on Supercomputing, IEEE, 1991, pp. 158–165.

[12] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C.W. Holzwarth, M.A. Popovic, H. Li, H.I. Smith, J.L. Hoyt, et al., Building many-core processor-to-DRAM networks with monolithic CMOS silicon photonics, IEEE MICRO (2009).

[13] S. Beamer, K. Asanovic, D.A. Patterson, The GAP benchmark suite, CoRR, arXiv:1508.03619 [cs.DC], 2015.

[14] S. Beamer, C. Sun, Y.J. Kwon, A. Joshi, C. Batten, V. Stojanović, K. Asanović, Re-architecting DRAM memory systems with monolithically integrated silicon photonics, in: International Symposium on Computer Architecture (ISCA), 2010.

[15] B. Benton, CCIX, Gen-Z, OpenCAPI: overview & Comparison, in: OpenFabrics Workshop, 2017.

[16] K. Bergman, J. Shalf, G. Michelogiannakis, S. Rumley, L. Dennison, M. Ghobadi, PINE: an energy efficient flexibly interconnected photonic data center architecture for extreme scalability, in: Optical Interconnects (OI), 2018.

[17] C. Bienia, S. Kumar, K. Li, PARSEC vs. SPLASH-2: a quantitative comparison of two multithreaded benchmark suites on chip-multiprocessors, in: IEEE International Workshop/Symposium on Workload Characterization (IISWC), 2008.

[18] C. Bienia, S. Kumar, J.P. Singh, K. Li, The PARSEC benchmark suite: characterization and architectural implications, in: International Conference on Parallel Architectures and Compilation Techniques (PACT), 2008.

[19] C.A. Brackett, Dense wavelength division multiplexing networks: principles and applications, IEEE J. Sel. Areas Commun. (1990).

[20] D. Brunina, C. Lai, A. Garg, K. Bergman, Building data centers with optically connected memory, J. Opt. Commun. Netw. (2011).

[21] D. Brunina, D. Liu, K. Bergman, An energy-efficient optically connected memory module for hybrid packet- and circuit-switched optical networks, IEEE J. Sel. Top. Quantum Electron. (2013).

[22] D. Brunina, X. Zhu, K. Padmaraju, L. Chen, M. Lipson, K. Bergman, 10-Gb/s WDM optically-connected memory system using silicon microring modulators, in: European Conference on Optical Communication (ECOC), 2012.

[23] J. Bucek, K.D. Lange, J.v. Kistowski, SPEC CPU2017: next-generation compute benchmark, in: International Conference on Performance Engineering (ICPE), 2018.

[24] B.B. Buckley, S.T.M. Fryslie, K. Guinn, G. Morrison, A. Gazman, Y. Shen, K. Bergman, M.L. Mashanovitch, L.A. Johansson, WDM source based on high-power, efficient 1280-nm DFB lasers for terabit interconnect technologies, IEEE Photonics Technol. Lett. (2018), https://doi.org/10.1109/LPT.2018.2872597.

[25] H. Casanova, A. Giersch, A. Legrand, M. Quinson, F. Suter, Versatile, scalable, and accurate simulation of distributed applications and platforms, J. Parallel Distrib. Comput. 74 (2014) 2899–2917, http://hal.inria.fr/hal-01017319.

[26] W. Chen, K. Ye, Y. Wang, G. Xu, C. Xu, How does the workload look like in production cloud? Analysis and clustering of workloads on Alibaba cluster trace, in: International Conference on Parallel and Distributed Systems (ICPADS), 2018.

[27] X. Chen, C.K. Fung, Y.M. Chen, H.K. Tsang, Subwavelength waveguide grating coupler for fiber-to- chip coupling on SOI with 80 nm 1 dB-bandwidth, in: CLEO, 2011.

[28] I. Datta, S.H. Chae, G.R. Bhatt, M.A. Tadayon, B. Li, Y. Yu, C. Park, J. Park, L. Cao, D. Basov, et al., Low-loss composite photonic platform based on 2d semiconductor monolayers, Nat. Photonics 14 (2020) 256–262.

[29] A. Degomme, A. Legrand, G.S. Markomanolis, M. Quinson, M. Stillwell, F. Suter, Simulating mpi applications: the smpi approach, IEEE Trans. Parallel Distrib. Syst. 28 (2017) 2387–2400.

[30] S. Di, D. Kondo, W. Cirne, Characterization and comparison of cloud versus grid workloads, in: IEEE International Conference on Cluster Computing, 2012.

[31] J. Duan, H. Huang, B. Dong, D. Jung, J.C. Norman, J.E. Bowers, F. Grillot, 1.3-um reflection insensitive inas/gaas quantum dot lasers directly grown on silicon, IEEE Photonics Technol. Lett. 31 (2019) 345–348.

[32] M.M.P. Fard, G. Cowan, O. Liboiron-Ladouceur, Responsivity optimization of a high-speed germanium-on-silicon photodetector, Opt. Express 24 (2016) 27738–27752.

[33] M. Glick, L.C. Kimmerling, R.C. Pfahl, A roadmap for integrated photonics, Opt. Photonics News (2018).

[34] J. Gonzalez, A. Gazman, M. Hattink, M.G. Palma, M. Bahadori, R. Rubio-Noriega, L. Orosa, M. Glick, O. Mutlu, K. Bergman, et al., Optically connected memory for disaggregated data centers, in: 2020 IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), IEEE, 2020, pp. 43–50.

[35] F. Grillot, L. Vivien, S. Laval, D. Pascal, E. Cassan, Size influence on the propagation loss induced by sidewall roughness in ultrasmall SOI waveguides, IEEE Photonics Technol. Lett. (2004).

[36] J. Gu, Y. Lee, Y. Zhang, M. Chowdhury, K.G. Shin, Efficient memory disaggregation with infiniswap, in: Symposium on Networked Systems Design and Implementation (NSDI), 2017.

[37] R. Hadidi, B. Asgari, B.A. Mudassar, S. Mukhopadhyay, S. Yalamanchili, H. Kim, Demystifying the characteristics of 3D-stacked memories: a case study for hybrid memory cube, in: IEEE International Workshop/Symposium on Workload Characterization (IISWC), 2017.

[38] J.L. Henning, SPEC CPU2006 benchmark descriptions, ACM SIGARCH Comput. Archit. News (2006).

[39] X. Jiang, N. Madan, L. Zhao, M. Upton, R. Iyer, S. Makineni, D. Newell, Y. Solihin, R. Balasubramonian, CHOP: adaptive filter-based DRAM caching for CMP server platforms, in: International Symposium on High-Performance Computer Architecture (HPCA), 2010.

[40] Joint Electron Device Engineering Council, JEDEC DDR4 Standard, https://www.jedec.org/, 2012.

[41] S. Keyvaninia, M. Muneeb, S. Stanković, P. Van Veldhoven, D. Van Thourhout, G. Roelkens, Ultra-thin dvs-bcb adhesive bonding of iii-v wafers, dies and multiple dies to a patterned silicon-on-insulator substrate, Opt. Mater. Express 3 (2013) 35–46.

[42] G. Kim, J. Kim, J.H. Ahn, J. Kim, Memory-centric system interconnect design with hybrid memory cubes, in: International Conference on Parallel Architectures and Compilation Techniques (PACT), 2013.

[43] S. Legtchenko, H. Williams, K. Razavi, A. Donnelly, R. Black, A. Douglas, N. Cheriere, D. Fryer, K. Mast, A.D. Brown, et al., Understanding rack-scale disaggregated storage, in: USENIX HotStorage, 2017.

[44] Y. Li, S. Ghose, J. Choi, J. Sun, H. Wang, O. Mutlu, Utility-based hybrid memory management, in: International Conference on Cluster Computing (CLUSTER), 2017.

[45] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S.K. Reinhardt, T.F. Wenisch, Disaggregated memory for expansion and sharing in blade servers, in: International Symposium on Computer Architecture (ISCA), 2009.

[46] J. Liu, B. Chandrasekaran, W. Yu, J. Wu, D. Buntinas, S. Kini, D.K. Panda, P. Wyckoff, Microbenchmark performance comparison of high-speed cluster interconnects, IEEE MICRO 24 (2004) 42–51.

[47] S. Liu, X. Wu, D. Jung, J.C. Norman, M. Kennedy, H.K. Tsang, A.C. Gossard, J.E. Bowers, High-channel-count 20 GHz passively mode-locked quantum dot laser directly grown on si with 4.1 tbit/s transmission capacity, Optica 6 (2019) 128–134.

[48] M. Long, P. Wang, H. Fang, W. Hu, Progress, challenges, and opportunities for 2d material based photodetectors, Adv. Funct. Mater. 29 (2019) 1803807.

[49] H. Luo, T. Shahroodi, H. Hassan, M. Patel, A.G. Yağlıkçı, L. Orosa, J. Park, O. Mutlu, CLR-DRAM: a low-cost DRAM architecture enabling dynamic capacity-latency trade-off, in: International Symposium on Computer Architecture (ISCA), 2020.

[50] M.D. Marino, Architectural impacts of RFiop: RF to address I/O pad and memory controller scalability, IEEE Trans. Very Large Scale Integr. (VLSI) Syst. (2018).

[51] J.D. McCalpin, Memory bandwidth and machine balance in current high performance computers, Newsl. - IEEE Comput. Soc., Tech. Comm. Comput. Archit. (1995) 19–25.

[52] Mellanox Technologies, Introducing 200G HDR InfiniBand Solutions, Technical Report, 2019.

[53] J. Meza, J. Chang, H. Yoon, O. Mutlu, P. Ranganathan, Enabling efficient and scalable hybrid memories using fine-granularity DRAM cache management, IEEE Comput. Archit. Lett. (2012).

[54] J. Meza, Y. Luo, S. Khan, J. Zhao, Y. Xie, O. Mutlu, A case for efficient hardware/software cooperative management of storage and memory, in: WEED, 2013.

[55] J. Meza, Q. Wu, S. Kumar, O. Mutlu, Revisiting memory errors in large-scale production data centers: analysis and modeling of new trends from the field, in: International Conference on Dependable Systems and Networks (DSN), 2015.

[56] N. Nethercote, J. Seward, Valgrind: a framework for heavyweight dynamic binary instrumentation, ACM SIGPLAN Not. 42 (2007) 89–100.

[57] R. Neugebauer, G. Antichi, J.F. Zazo, Y. Audzevich, S. López-Buedo, A.W. Moore, Understanding PCIe performance for end host networking, in: ACM Special Interest Group on Data Communication (SIGCOMM), 2018.

[58] K. Padmaraju, D.F. Logan, T. Shiraishi, J.J. Ackert, A.P. Knights, K. Bergman, Wavelength locking and thermally stabilizing microring resonators using dithering signals, J. Lightwave Technol. (2013).

[59] G. Panwar, D. Zhang, Y. Pang, M. Dahshan, N. DeBardeleben, B. Ravindran, X. Jian, Quantifying memory underutilization in HPC systems and using it to improve performance via architecture support, in: International Symposium on Microarchitecture (MICRO), 2019.

[60] A.D. Papaioannou, R. Nejabati, D. Simeonidou, The benefits of a disaggregated data centre: a resource allocation approach, in: IEEE Global Communications Conference (GLOBECOM), 2016.

[61] J.T. Pawlowski, Hybrid memory cube (HMC), in: HOTCHIPS, 2011.

[62] R. Polster, Y. Thonnart, G. Waltener, J. Gonzalez, E. Cassan, Efficiency optimization of silicon photonic links in 65-nm CMOS and 28-nm FDSOI technology nodes, IEEE Trans. Very Large Scale Integr. (VLSI) Syst. (2016).

[63] R. Proietti, Y. Yin, Z. Cao, C. Nitta, V. Akella, S.B. Yoo, Low-latency interconnect optical network switch (LIONS), in: Optical Switching in Next Generation Data Centers, 2018.

[64] L.E. Ramos, E. Gorbatov, R. Bianchini, Page placement in hybrid memory systems, in: Proceedings of the International, 2011, conference on Supercomputing (ICS).

[65] C. Reiss, A. Tumanov, G.R. Ganger, R.H. Katz, M.A. Kozuch, Towards Understanding Heterogeneous Clouds at Scale: Google Trace Analysis, Intel Science and Technology Center for Cloud Computing (ISTCCC), 2012, Tech. Rep.

[66] S. Rumley, M. Bahadori, K. Wen, D. Nikolova, K. Bergman, Phoenixsim: cross-layer design and modeling of silicon photonic interconnects, in: International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems (AISTECS), 2016.

[67] D. Sanchez, C. Kozyrakis, ZSim: fast and accurate microarchitectural simulation of thousand-core systems, in: International Symposium on Computer Architecture (ISCA), 2013.

[68] K.i. Sato, Realization and application of large-scale fast optical circuit switch for data center networking, J. Lightwave Technol. (2018).

[69] Y. Shen, X. Meng, Q. Cheng, S. Rumley, N. Abrams, A. Gazman, E. Manzhosov, M.S. Glick, K. Bergman, Silicon photonics for extreme scale systems, J. Lightwave Technol. (2019).

[70] T.N. Theis, H.S.P. Wong, The end of Moore's law: a new beginning for information technology, Comput. Sci. Eng. (2017), https://doi.org/10.1109/MCSE.2017.29.

[71] Z. Wang, A. Abbasi, U. Dave, A. De Groote, S. Kumari, B. Kunert, C. Merckling, M. Pantouvaki, Y. Shi, B. Tian, et al., Novel light source integration approaches for silicon photonics, Laser Photonics Rev. 11 (2017) 1700063.

[72] J. Weiss, R. Dangel, J. Hofrichter, F. Horst, D. Jubin, N. Meier, A. La Porta, B.J. Offrein, Optical interconnects for disaggregated resources in future datacenters, in: European Conference on Optical Communication (ECOC), 2014.

[73] S. Williams, A. Waterman, D. Patterson, Roofline: an insightful visual performance model for multicore architectures, Commun. ACM 52 (2009) 65–76.

[74] Y. Yan, G.M. Saridis, Y. Shu, B.R. Rofoee, S. Yan, M. Arslan, T. Bradley, N.V. Wheeler, N.H.L. Wong, F. Poletti, et al., All-optical programmable disaggregated data centre network realized by FPGA-based switch and interface card, J. Lightwave Technol. (2016).

[75] H. Yoon, J. Meza, R. Ausavarungnirun, R.A. Harding, O. Mutlu, Row buffer locality aware caching policies for hybrid memories, in: International Conference on Computer Design (ICCD), 2012.

[76] H. Yoon, J. Meza, N. Muralimanohar, N.P. Jouppi, O. Mutlu, Efficient data mapping and buffering techniques for multilevel cell phase-change memories, ACM Trans. Archit. Code Optim. (2014).

[77] X. Yu, C.J. Hughes, N. Satish, O. Mutlu, S. Devadas, Banshee: bandwidth-efficient DRAM caching via software/hardware cooperation, in: International Symposium on Microarchitecture (MICRO), 2017.

[78] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, V. Mishra, Optically disaggregated data centers with minimal remote memory latency: technologies, architectures, and resource allocation, J. Opt. Commun. Netw. (2018), https://doi.org/10.1364/JOCN.10.00A270.

[79] Z. Zhu, Y. Shen, Y. Huang, A. Gazman, M. Hattink, K. Bergman, Flexible resource allocation using photonic switched interconnects for disaggregated system architectures, in: Optical Fiber Communication (OFC), 2019.

Jorge Gonzalez is an assistant professor at the Computer Science Department of the University of Engineering and Technology, Peru. He received his Ph.D. degree in 2021 from the University of Campinas in São Paulo, Brazil. His current research interests are in computer architecture, optical interconnects, memory systems, and intra-chip traffic.

Mauricio G. Palma is a Ph.D candidate at the Computing Systems Laboratory of the University of Campinas in São Paulo, Brazil. His current research interests are in computer architecture, optical interconnects, memory systems, and parallel processing.

Maarten Hattink received his B.S. and M.S. from the Eindhoven University of Technology, Netherlands, in 2015 and 2017. While pursuing these degrees he worked at Prodrive Technologies B.V. as a software and FPGA engineer. He is now pursuing a Ph.D. degree and his research interest lies in photonic device integration and optical switching.

Ruth E. Rubio-Noriega is a lecturer at the National University of Engineering Peru and a researcher at the National Institute for Research and Training in Telecommunications in Lima, Peru. She obtained her Ph.D. degree at the University of Campinas in Sao Paulo, Brazil. Her interests include photonic devices, electrooptics, optical interconnects, and applied electromagnetics.

Lois Orosa is a PostDoc in the SAFARI research group at ETH Zürich. His current research interests are in computer architecture, hardware security, memory systems, and ML accelerators. He obtained his PhD from the University of Santiago de Compostela, and he was a PostDoc in the Institute of Computing at University of Campinas. He was a visiting scholar at University of Illinois at Urbana-Champaign and Universidade NOVA de Lisboa, and he acquired industrial experience at IBM, Recore Systems, and Xilinx.

Onur Mutlu is a Professor of Computer Science at ETH Zurich, and is also a faculty member at Carnegie Mellon University. He obtained his PhD and MS in ECE from the University of Texas at Austin (in 2006). His current research interests are in computer architecture, computing systems, hardware security, and bioinformatics. He received the IEEE Computer Society Edward J. McCluskey Technical Achievement Award, the ACM SIGARCH Maurice Wilkes Award, the inaugural IEEE Computer Society Young Computer Architect Award, the inaugural Intel Early Career Faculty Award, US National Science Foundation CAREER Award, Carnegie Mellon University Ladd Research Award, faculty partnership awards from various companies, and a healthy number of best paper or "Top Pick" paper recognitions at various computer systems, architecture, and hardware security venues. For more information, please see his webpage at https://people.inf.ethz.ch/omutlu/.

Keren Bergman is the Charles Batchelor Professor of Electrical Engineering at Columbia University where she also serves as the Faculty Director of the Columbia Nano Initiative. Prof. Bergman received the B.S. from Bucknell University in 1988, and the M.S. in 1991 and Ph.D. in 1994 from M.I.T. all in Electrical Engineering. At Columbia, Bergman leads the Lightwave Research Laboratory (http://lightwave.ee.columbia.edu/) encompassing multiple cross-disciplinary programs at the intersection of computing and photonics. Her work involves the design exploration, architecture, and implementation of photonic systems that incorporate the advantages of manipulating information in the optical domain for accelerating energy efficient high-performance computing and data centers. Bergman serves on the Leadership Council of the American Institute of Manufacturing (AIM) Photonics leading projects that support the institute's silicon photonics manufacturing capabilities and Datacom applications. She is the recipient of the 2016 IEEE Photonics Engineering Award and is a Fellow of the Optical Society of America (OSA) and IEEE.

Rodolfo Azevedo is an associate professor at University of Campinas (UNICAMP). He received his PhD in Computer Science from University of Campinas (UNICAMP) in 2002 and is a member of the Computer Science graduate program where he advises master and PhD students. He got four best papers in conferences (SBAC-PAD 2004, SBAC-PAD 2008, 2018, and WSCAD-SSC 2012). In 2012 he received the Zeferino Vaz Academic Award and the newly created UNICAMP Teaching Award. He has had a CNPq Research Fellowship since 2006. He has been honored 8 times in the Computer Science and Computer Engineering graduations. He was Director of the Institute of Computing from 2017-2019 and currently he is the President of the São Paulo Virtual University (UNIVESP).