Optically Connected Memory for Disaggregated Data Centers

Jorge Gonzalez* Alexander Gazman‡ Maarten Hattink‡ Mauricio G. Palma* Meisam Bahadori§
Ruth Rubio-Noriega¶ Lois Orosa** Madeleine Glick‡ Onur Mutlu** Keren Bergman‡ Rodolfo Azevedo*

*University of Campinas ‡Columbia University §Nokia ¶INICTEL-UNI **ETH Zürich

***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 1.07 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.**

***Index Terms*—disaggregated computing, disaggregated memory, photonics, data-centers, DRAM, memory systems**

I. 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 [23], [25], [36], [43], [49], 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 [39]; 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 [35]. As depicted in Figure 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 [14]. Memory disaggregation provides flexibility in memory allocation, improved utilization of the memory resources, lower maintenance costs, and lower energy consumption in the data center [44].

Disaggregating memory and processors remains a challenge, although the disaggregation of some resources (e.g., storage) is common in production data centers [33]. Electrical interconnections in rack-distances do not fulfill the low latency and high bandwidth requirements of modern DRAM modules.

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 Sao Paulo Research Foundation (FAPESP) (fellowships 2013/08293-7 and 2014/01642-9), CAPES (fellowships 2013/08293-7 and 88882.329108/2019-01), and CNPq (fellowships 438445/2018-0, 309794/2017-0 and 142016/2020-9).

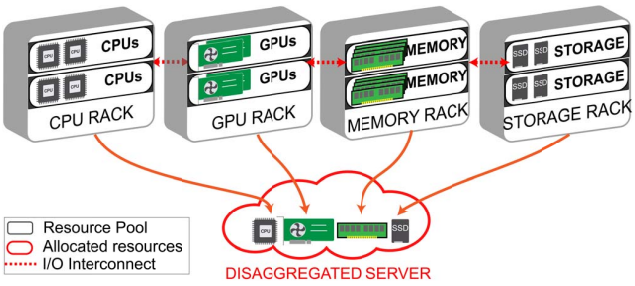


Fig. 1. Disaggregation concept for data centers.

The primary limitation of an electrical interconnect is that it constrains the memory bus to onboard distance [54] 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 (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 $\geq$ 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 IV-A) and energy consumption (see Section IV-B) 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 5.5x faster than a 40G NIC-based disaggregated memory.

2643-3001/20/$31.00 ©2020 IEEE
DOI 10.1109/SBAC-PAD49847.2020.00017

II. 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 [52]).

New electrical interfaces, such as GenZ, CCIX, and OpenCAPI, can disaggregate a wide range of resources (e.g., memory, accelerators) [13]. 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 [62]. 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.

Figure 2 shows the IO requirements in the memory controller for electrical [37], 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) [8], [17]. Second, a single optical IO achieves up to 800 Gbps based on our evaluation, requiring 2 IOs for bidirectional communication (see Section IV-B). 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.

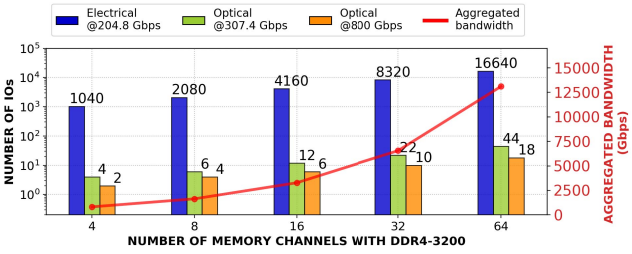


Fig. 2. Required electrical and optical IO counts (lower is better) for sustaining different amounts of aggregated bandwidth.

III. 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 [5], [18], to reduce the latency overhead caused by additional protocols such as remote direct memory access (RDMA) and PCIe [61]. OCM is versatile and scales with the increasing number of wavelengths per memory channel expected from future photonic systems [26].

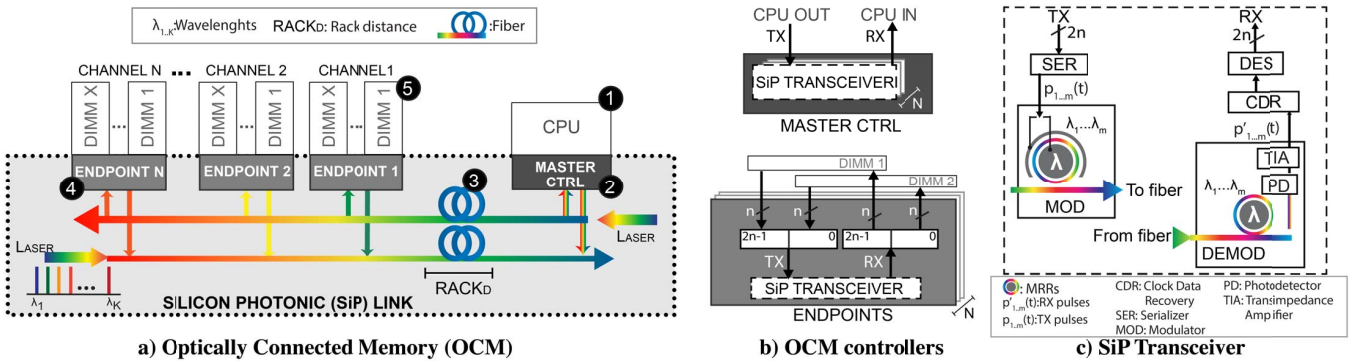
A. Architecture Overview

Figure 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.

In our evaluation (Section IV), 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 III-B). 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 (e.g., [29]) supports serialization up to 256B (i.e., four 64B cache lines). As shown in Figure 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 Figure 3b shows a difference in the organization of the SiP transceivers per controller. Figure 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 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.

B. 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. Figure 4 presents the timing diagram of the OCM Read (RD) and Write (WR) operations. For reference, a conventional DDR DRAM memory channel uses 64B cache lines; a data bus transfers each line as 8B data blocks in 8 consecutive cycles, and the 1B Command (CMD) and 3B Address (ADDR) use separate dedicated buses. In OCM, as depicted in Figure 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



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.

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 [1]), 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 [29] and 128B cache line size, transfers 2 $\times$ (4 $\times$ 16B of data) with 1 $\times$ 4B 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 IV). 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 [42]) T_{setup}.

$$T_{lat}(t) = T_{setup} + T_{contr} + T_{mem(A|B)}(t) + T_{serdes} + T_{mod} + T_{demod} + T_{dist} \quad (1)$$

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 [8], due to the small footprint of ring modulators (tens of micrometers) and the high dielectric constant of silicon.

C. Operation

Figure 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 $\lambda_{1,2,...,K}$ wavelengths simultaneously [9].

Stage ❷: the data from the processor is serialized (SER) onto the Master Controller's TX lane, and the generated electrical pulses $p_{1,2,...,m}(t)$ 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

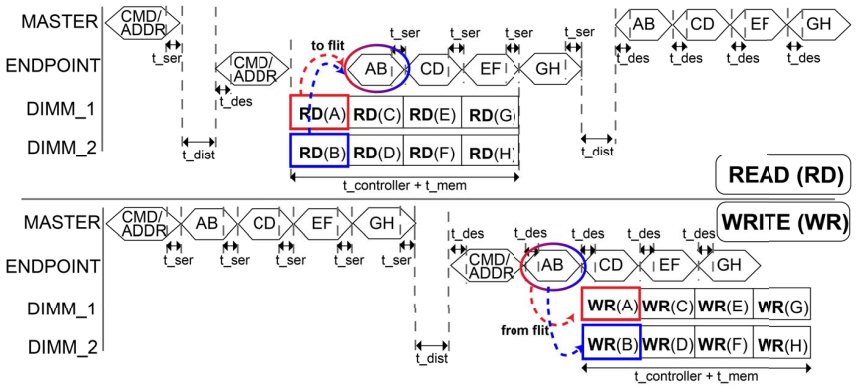


Fig. 4. OCM timing diagram for Read (top) and Write (bottom) requests.

(NRZ-OOK) that represents logical ones and zeros imprinted on the envelope of light [8].

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 $\lambda_{1,2,...m}$. 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 $p'_{1,2,...,m}(t)$ 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 $\lambda_{1,...,m}$ (similar to Stages 1 to 4). For a WR command, the data received from the processor is stored in memory.

D. 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 Figure 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 IV, we evaluate OCM with a single CPU, and assume that the setup phase is already completed.

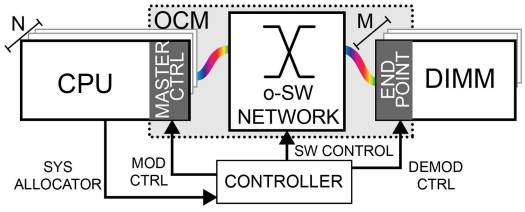


Fig. 5. Reconfigurable OCM with optical switches (o-SW).

E. High Aggregated Bandwidth

OCM uses WDM [8], [17] 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 [9], and they are tailored to work in the communications C-band (1530-1565 nm). For more detail on photonic devices, please see [7], [26], [53].

OCM achieves high aggregated bandwidth by using multiple optical wavelengths $\lambda_{1,2,...,K}$ (see laser in Figure 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 ($\lambda_{1,2,...,m}$, see Figure 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.

IV. Evaluation

To evaluate system-level performance, we implement OCM architecture in the ZSIM simulator [51]. To evaluate the interconnection between processor and memory as a point-to-point SiP link, we use PhoenixSim [50] with parameters extracted from state-of-the-art optical devices [6], [8], [46]. The SiP link energy-per-bit modeling allows us to find: (1) the required number of optical wavelengths (λ), and (2) the bit rate per λ. Table I lists OCM optical devices and their main characteristics used in our simulation model.

TABLE I
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	30%	Laser wall-plug efficiency	[22]
Waveguide loss	5 dB/cm	fabrication roughness	[27]
	0.02 dB/bend	waveguide bend loss	
Coupler loss	1 dB	off-chip coupler	[24]
Modulator	Q = 6500	Ring resonator Q factor	[46]
	ER = 10 dB	MRR extinction rate	
	65 fF	Junction capacitance	
	-5 V	Maximum drive voltage	
	1 mW	Thermal-tuning power/ring	[7]
Mod. mux and receiver demux	MRR power penalties	Crosstalk model	[8]
Photodetector	1 A/W	Current per o-power	[55]
Modulator driver	28 nm	Semicond. tech. for OOK-WDM	[46]
SERDES power model	28 nm	Semicond. tech.	[46]
Digital receiver	28 nm	Semicond. tech. for OOK-NRZ	[46]
Element positioning	100 μm	Modulator padding	

Table II 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 [61], which can have a significant performance impact in memory-bound applications. Our DRAM cache resembles

the Banshee DRAM cache [60] 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 DRAM cache to have the same operation latency as commodity DDR4 memory.

TABLE II
BASELINE PROCESSOR, MEMORY, OCM, AND NIC.

Baseline	*Processor*	3 GHz, 8 cores, 128B cache lines
	Cache	32KB L1(D+I), 256KB L2, 8MB L3
MemConf1	*Mem*	4 channels, 2 DIMMs/channel, DDR4-2400 [1]
MemConf2	*Mem*	1 channel, 2 DIMMs/channel, DDR4-2400
	DRAM cache	4GB stacked, 4-way, 4K pages, FBR [60], DDR4-2400
OCM	*SERDES*	latency: 10/150/340 cycles
	Fiber	latency: 30/60/90 cycles (2/4/6 meters roundtrip)
NIC	*40G PCIe [41]*	latency: 1050 cycles

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 [32]. We use 340 cycles (113ns) maximum latency reported in a previously demonstrated optical interconnection system [47]. We simulate rack distances of 2m, 4m, and 6m with a 5 ns/m latency [2], which translates into 30, 60, and 90 cycles latency in our system.

For the 40G NIC-based system configuration, we evaluate a scenario using a PCIe Network Interface Card (NIC) latency of 1050 cycles (350 ns) [2] (a realistic NIC-through-PCIe latency is in the order of microseconds [41]). We evaluate the system-level performance of OCM with applications from four benchmark suites: (1) SPEC06 [30] using Pinpoints (warmup of 100 million instructions, and detailed region of 30 million instructions), (2) PARSEC [16] with *native* inputs, (3) SPLASH2 [15] with *simlarge* inputs, (4) SPEC17 [21] *speed* with reference inputs, and (5) GAP graph benchmarks [11] 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. Table III lists the SPEC benchmark mixes we use in our multiprogrammed workload evaluation. Table IV summarizes the measured memory footprint values for all the benchmarks used in our evaluation.

A. System-level Evaluation

Multiprogrammed Evaluation. Figure 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 III).

TABLE III
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 IV
MEASURED MEMORY FOOTPRINTS.

SPEC06	[30]	*MIX1*: 2.2 GB, *MIX2*: 3.1 GB, *MIX3*: 2.4 GB
SPEC17	[21]	*MIX1*: 19.9 GB, *MIX2*: 36.4 GB, *MIX3*: 34.7 GB.
PARSEC	[16]	*canneal*: 716.7 MB, *streamcluster*: 112.5 MB, *ferret*: 91.9 MB, *raytrace*: 1.3 GB, *fluidanimate*: 672 MB
SPLASH	[15]	*radix*: 1.1 GB, *fft*: 768.8 MB, *cholesky*: 44.2 MB, *ocean_ncp*: 26.9 GB, *ocean_cp*: 891.8 MB.
GAP	[11]	*Urand* graph: 18 GB, *Web* graph: 15.5 GB

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 II). OCM is up to $5.5\times$ faster than 40G NIC for the minimum SERDES latency, and $2.16\times$ 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 $1.07\times$ compared to the baseline with a SERDES latency of 10 cycles, and $1.78\times$ average slowdown with a SERDES latency of 340 cycles.

Figure 7 shows the speedup of a disaggregated OCM system (green bars) compared to a non-disaggregated baseline, both configured with MemConf1. Figure 7 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 4m.

Figure 7 (left) shows the results for SPEC17 mixes (see Table III). We make two observations. First, the average slowdown of OCM without DRAM cache (green bars) is

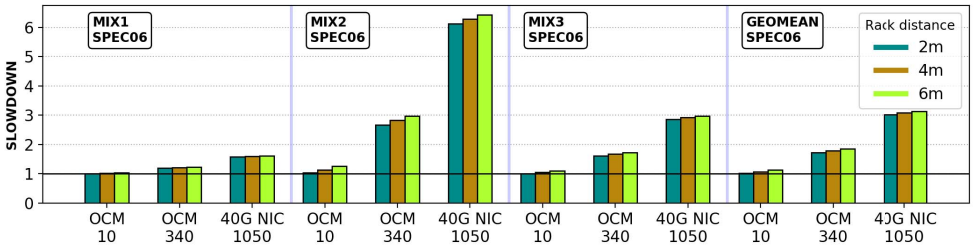


Fig. 6. Slowdowns of OCM and 40G NIC-based disaggregated systems, compared to a non-disaggregated baseline with MemConf1, for three randomly-selected mixes of SPEC06 benchmarks (lower is better).

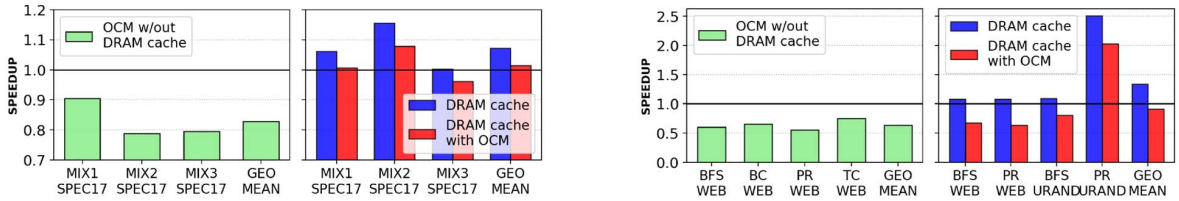


Fig. 7. OCM speedup results with 4m distance and a SERDES latency of 150 cycles (higher is better), compared to a disaggregated baseline, with or without a DRAM cache. Left: Speedup for SPEC17. Right: Speedup for GAP [11] graph benchmarks

17%, which is in the same order as the SPEC06 results (Figure 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. Figure 7 (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 dramatically benefit from using a DRAM cache (red and blue bars), e.g., *PR* with *Urand* input shows a speedup of $2.5\times$ 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 [31], [34], [38], [40], [48], [58]–[60].

Figure 8 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 $2.97\times$ for *streamcluster* at 6m and 340 SERDES cycles; the average slowdown is $1.3\times$ for SPLASH2 and $1.4\times$ for PARSEC.

We conclude that OCM is very promising because of its reasonably low latency overhead (especially with the use of a DRAM cache), and the flexibility of placing memory modules at large distances with small slowdowns.

B. 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. We consider unidirectional SiP links using PhoenixSim [50] using the parameters shown in Table I. 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 [1]. 4 memory channels, with 2 DIMMs per channel in lockstep, require $\sim$615 Gbps/link. OCM's maximum feasible bandwidth (while remaining CMOS compatible) is 802 Gbps using the parameters in Table I. More advanced modulation formats, such as PAM4 [53], can be used to achieve higher aggregated bandwidth. Figure 9 shows 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).

In OCM with 615 Gbps links, the minimum energy consumption overhead compared to the electrical memory system is 1.07 pJ/bit for 35 optical wavelengths (λ) per link, each λ operating at 17.57 Gbps. In OCM with 802 Gbps links, the minimum energy consumption is 1.57 pJ/bit for 39 λs per link, each λ operating at 20.56 Gbps.

We make three observations from Figure 9. 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 [45]; 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 [9]. 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.

We conclude that a bidirectional SiP link, formed by two

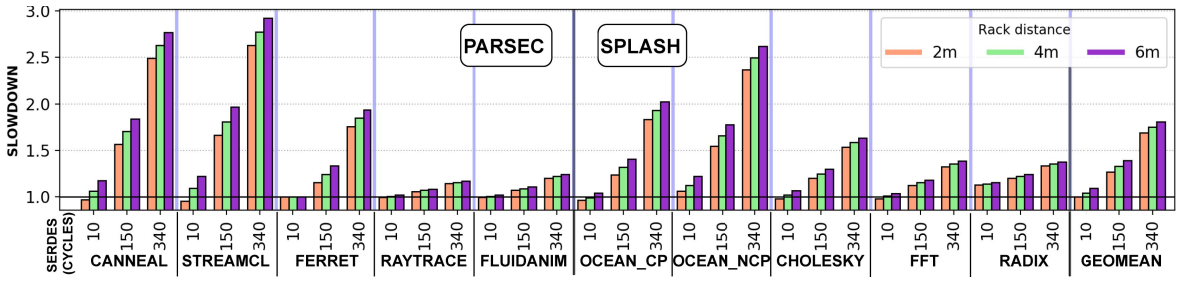


Fig. 8. OCM slowdown compared to the baseline for PARSEC and SPLASH2 benchmarks (lower is better).

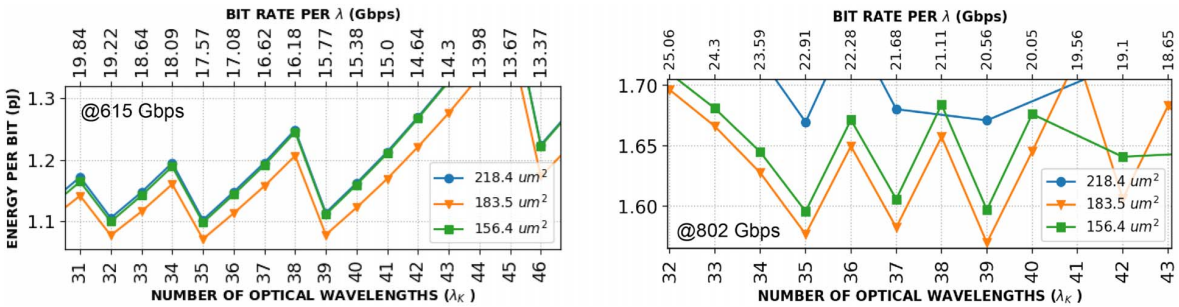


Fig. 9. SiP link energy-per-bit. Left: at 615 Gbps bandwidth, Right: at 802 Gbps bandwidth.

unidirectional links using current SiP devices, can fit the bandwidth requirements of commodity DDR4 DRAM modules. OCM incurs a low energy overhead of only 10.7% compared to a non-disaggregated DDR4 DRAM memory (the energy consumption of current DDR4 DRAM technology is $\sim$ 10pJ/bit [53]).

V. 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. [19], [20] 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. [18] also experimentally demonstrate that microring modulators can be used for optically connecting DDR2 memory. Our work builds on [18] to design the microring modulators used in our SiP links. There are several recent works [8], [9], [53] that propose analytical models of the microring used in our SiP links. Anderson et al. [5] extend the work of Brunina et al. [18]–[20] to experimentally demonstrate the optical switches using FPGAs for accessing memory.

These prior works [5], [18]–[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 IV-B), (2) What is the energy and area impact on the system? (Section IV-B), and (3) How does the processor interact with a disaggregated memory subsystem? (Section IV-A).

Some other works, such as [56], [62], 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 III-A). 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 [50] 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 [57].

Yan et al. [57] 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 [10], (2) propose the design of a DRAM chip with photonic inter-bank communication [12], (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 [61], (5) analyze the cost viability of optical memory disaggregation [2], and (6) evaluate memory disaggregation using software mech-

anisms with high latency penalties in the order of μs [28]. Unlike [2], [4], [12], [28], [61], our work evaluates i) system performance with real applications, ii) the design of the SiP link for DDR DRAM requirements, and iii) SiP link energy for a disaggregated memory system.

VI. 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 times faster than a 40G NIC-based disaggregated memory. We conclude that OCM is a promising step towards future data centers with disaggregated main memory.

References

[1] "JEDEC DDR4 Standard," https://www.jedec.org/, 2012.
[2] B. Abali *et al.*, "Disaggregated and Optically Interconnected Memory: When will it be Cost Effective?" *arXiv*, 2015.
[3] P. P. Absil *et al.*, "Imec iSiPP25G Silicon Photonics: a Robust CMOS-based Photonics Technology Platform," in *Silicon Photonics X*, 2015.
[4] M. S. Akhter *et al.*, "WaveLight: A Monolithic Low Latency Silicon-Photonics Communication Platform for the Next-Generation Disaggregated Cloud Data Centers," in *HOTI*, 2017.
[5] E. F. Anderson *et al.*, "Reconfigurable Silicon Photonic Platform for Memory Scalability and Disaggregation," in *OFC*, 2018.
[6] M. Bahadori *et al.*, "Energy-bandwidth design exploration of silicon photonic interconnects in 65nm CMOS," in *OI*, 2016.
[7] M. Bahadori *et al.*, "Thermal Rectification of Integrated Microheaters for Microring Resonators in Silicon Photonics Platform," *JLT*, 2018.
[8] M. Bahadori *et al.*, "Comprehensive Design Space Exploration of Silicon Photonic Interconnects," *JLT*, 2016.
[9] M. Bahadori *et al.*, "Energy-performance Optimized Design of Silicon Photonic Interconnection Networks for High-performance Computing," in *DATE*, 2017.
[10] C. Batten *et al.*, "Building Many-core Processor-to-DRAM Networks with Monolithic CMOS Silicon Photonics," *MICRO*, 2009.
[11] S. Beamer *et al.*, "The GAP Benchmark Suite," *CoRR*, 2015.
[12] S. Beamer *et al.*, "Re-architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics," in *ISCA*, 2010.
[13] B. Benton, "CCIX, Gen-Z, OpenCAPI: Overview & Comparison," in *OpenFabrics Workshop*, 2017.
[14] K. Bergman *et al.*, "PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability," in *OI*, 2018.
[15] C. Bienia *et al.*, "PARSEC vs. SPLASH-2: A Quantitative Comparison of Two Multithreaded Benchmark Suites on Chip-Multiprocessors," in *IISWC*, 2008.
[16] C. Bienia *et al.*, "The PARSEC Benchmark Suite: Characterization and Architectural Implications," in *PACT*, 2008.
[17] C. A. Brackett, "Dense Wavelength Division Multiplexing Networks: Principles and Applications," *JSAC*, 1990.
[18] D. Brunina *et al.*, "An Energy-Efficient Optically Connected Memory Module for Hybrid Packet- and Circuit-Switched Optical Networks," *JSTQE*, 2013.
[19] D. Brunina *et al.*, "Building Data Centers with Optically Connected Memory," *JOCN*, 2011.
[20] D. Brunina *et al.*, "10-Gb/s WDM Optically-Connected Memory System Using Silicon Microring Modulators," in *ECOC*, 2012.
[21] J. Bucek *et al.*, "SPEC CPU2017: Next-generation Compute Benchmark," in *ICPE*, 2018.
[22] B. B. Buckley *et al.*, "WDM Source Based on High-Power, Efficient 1280-nm DFB Lasers for Terabit Interconnect Technologies," *IEEE PTL*, 2018.
[23] W. Chen *et al.*, "How Does the Workload Look Like in Production Cloud? Analysis and Clustering of Workloads on Alibaba Cluster Trace," in *ICPADS*, 2018.
[24] X. Chen *et al.*, "Subwavelength Waveguide Grating Coupler for Fiber-to-Chip Coupling on SOI with 80nm 1dB-bandwidth," in *CLEO*, 2011.
[25] S. Di *et al.*, "Characterization and Comparison of Cloud versus Grid Workloads," in *CLOUD*, 2012.
[26] M. Glick *et al.*, "A Roadmap for Integrated Photonics," *OPN*, 2018.
[27] F. Grillot *et al.*, "Size Influence on the Propagation Loss Induced by Sidewall Roughness in Ultrasmall SOI Waveguides," *PTL*, 2004.
[28] J. Gu *et al.*, "Efficient Memory Disaggregation with Infiniswap." in *NSDI*, 2017.
[29] R. Hadidi *et al.*, "Demystifying the Characteristics of 3D-Stacked Memories: A Case Study for Hybrid Memory Cube," in *IISWC*, 2017.
[30] J. L. Henning, "SPEC CPU2006 Benchmark Descriptions," *ACM SIGARCH Computer Architecture News*, 2006.
[31] X. Jiang *et al.*, "CHOP: Adaptive Filter-Based DRAM Caching for CMP Server Platforms," in *HPCA*, 2010.
[32] G. Kim *et al.*, "Memory-Centric System Interconnect Design with Hybrid Memory Cubes," in *PACT*, 2013.
[33] S. Legtchenko *et al.*, "Understanding Rack-scale Disaggregated Storage," in *USENIX HotStorage*, 2017.
[34] Y. Li *et al.*, "Utility-Based Hybrid Memory Management," in *CLUSTER*, 2017.
[35] K. Lim *et al.*, "Disaggregated Memory for Expansion and Sharing in Blade Servers," in *ISCA*, 2009.
[36] H. Luo *et al.*, "CLR-DRAM: A Low-Cost DRAM Architecture Enabling Dynamic Capacity-Latency Trade-Off," in *ISCA*, 2020.
[37] M. D. Marino, "Architectural Impacts of RFiop: RF to Address I/O Pad and Memory Controller Scalability," *TVLSI*, 2018.
[38] J. Meza *et al.*, "Enabling Efficient and Scalable Hybrid Memories Using Fine-Granularity DRAM Cache Management," *CAL*, 2012.
[39] J. Meza *et al.*, "Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field," in *DSN*, 2015.
[40] J. Meza *et al.*, "A Case for Efficient Hardware/Software Cooperative Management of Storage and Memory," *WEED*, 2013.
[41] R. Neugebauer *et al.*, "Understanding PCIe Performance for End Host Networking," in *SIGCOMM*, 2018.
[42] K. Padmaraju *et al.*, "Wavelength Locking and Thermally Stabilizing Microring Resonators Using Dithering Signals," *JLT*, 2013.
[43] G. Panwar *et al.*, "Quantifying Memory Underutilization in HPC Systems and Using it to Improve Performance via Architecture Support," in *MICRO*, 2019.
[44] A. D. Papaioannou *et al.*, "The Benefits of a Disaggregated Data Centre: A Resource Allocation Approach," in *GLOBECOM*, 2016.
[45] J. T. Pawlowski, "Hybrid Memory Cube (HMC)," in *HOTCHIPS*, 2011.
[46] R. Polster *et al.*, "Efficiency Optimization of Silicon Photonic Links in 65-nm CMOS and 28-nm FDSOI Technology Nodes," *TVLSI*, 2016.
[47] R. Proietti *et al.*, "Low-latency Interconnect Optical Network Switch (LIONS)," in *Optical Switching in Next Generation Data Centers*, 2018.
[48] L. E. Ramos *et al.*, "Page Placement in Hybrid Memory Systems," in *ICS*, 2011.
[49] C. Reiss *et al.*, "Towards Understanding Heterogeneous Clouds at Scale: Google Trace Analysis," *ISTCCC, Tech. Rep*, 2012.
[50] S. Rumley *et al.*, "Phoenixsim: Crosslayer Design and Modeling of Silicon Photonic Interconnects," in *AISTECS*, 2016.
[51] D. Sanchez and C. Kozyrakis, "ZSim: Fast and Accurate Microarchitectural Simulation of Thousand-core Systems," in *ISCA*, 2013.
[52] K.-i. Sato, "Realization and Application of Large-scale Fast Optical Circuit Switch for Data Center Networking," *JLT*, 2018.
[53] Y. Shen *et al.*, "Silicon Photonics for Extreme Scale Systems," *JLT*, 2019.
[54] T. N. Theis and H.-S. P. Wong, "The End of Moore's Law: A New Beginning for Information Technology," *CSE*, 2017.
[55] L. Vivien *et al.*, "42 Ghz Pin Germanium Photodetector Integrated in a Silicon-on-insulator Waveguide," *Optics Express*, 2009.
[56] J. Weiss *et al.*, "Optical Interconnects for Disaggregated Resources in Future Datacenters," in *ECOC*, 2014.
[57] Y. Yan *et al.*, "All-optical Programmable Disaggregated Data Centre Network Realized by FPGA-based Switch and Interface Card," *JLT*, 2016.
[58] H. Yoon *et al.*, "Row Buffer Locality Aware Caching Policies for Hybrid Memories," in *ICCD*, 2012.
[59] H. Yoon *et al.*, "Efficient Data Mapping and Buffering Techniques for Multilevel Cell Phase-Change Memories," *TACO*, 2014.
[60] X. Yu *et al.*, "Banshee: Bandwidth-efficient DRAM Caching Via Software/hardware Cooperation," in *MICRO*, 2017.
[61] G. Zervas *et al.*, "Optically Disaggregated Data Centers With Minimal Remote Memory Latency: Technologies, Architectures, and Resource Allocation," *JOCN*, 2018.
[62] Z. Zhu *et al.*, "Flexible Resource Allocation Using Photonic Switched Interconnects for Disaggregated System Architectures," in *OFC*, 2019.