Re-architecting DRAM with Monolithically Integrated Silicon Photonics

Scott Beamer
Chen Sun
Yong-jin Kwon
Ajay Joshi
Christopher Batten
Vladimir Stojanovic
Krste Asanovic

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-179
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-179.html

December 17, 2009

Re-architecting DRAM with Monolithically Integrated Silicon Photonics

Scott Beamer[*], Chen Sun[†], Yong-Jin Kwon[*]
Ajay Joshi[‡], Christopher Batten[§], Vladimir Stojanović[†], Krste Asanović[*]

[*] *Department of EECS, University of California, Berkeley, CA*
[†] *Department of EECS, Massachusetts Institute of Technology, Cambridge, MA*
[‡] *Department of ECE, Boston University, Cambridge, MA*
[§] *Department of ECE, Cornell University, Ithaca, NY*

Future manycore computing systems will only deliver increased performance if memory bandwidth improves also. Projected scaling of electrical DRAM architectures appears unlikely to suffice, being constrained by pin count and pin bandwidth, and by total chip power, including off-chip signaling, cross-chip interconnect, and bank access energy.

In this work, we redesign the main memory system using a proposed monolithically integrated silicon photonics technology and show that it provides a promising solution to all of these issues. Photonics can provide high aggregate pin-bandwidth density through dense wavelength-division multiplexing (DWDM). Photonic signaling also provides energy-efficient long-range communication, which we exploit to not only reduce chip-chip link power but also extend on to the DRAMs to reduce cross-chip interconnect power. To balance these large improvements in interconnect bandwidth and power, we must also improve the energy efficiency of the DRAM banks themselves, as we now require many concurrently accessed banks to support the available bandwidth. We explore decreasing the number of bits activated per bank to increase the banks' energy efficiency. Since DRAM chips are a very cost-sensitive commodity, we weigh these efficiency gains against any area overhead. Due to the immature nature of photonics technology, we explore a large design space to capture plausible designs for a range of different technology assumptions. Our most promising design point yields approximately a 10× improvement in power for the same throughput as a projected future electrical-only DRAM, with slightly reduced area.

Finally, we propose a new form of optical power guiding which increases the scalability of our memory architecture and allows for a single chip design to be used in both high capacity or high bandwidth memory configurations.

1. Introduction

The move to parallel microprocessors would appear to continue to allow Moore's Law gains in transistor density to be converted to gains in processing performance. Unfortunately, off-chip memory bandwidths are unlikely to scale in the same way, and could ultimately bottleneck achievable system performance.

Current microprocessors are already pushing the pin bandwidth limits of their sockets, and it seems unlikely that the pin bandwidth will increase dramatically without a disruptive technology. The number of signal pins is limited by the area and power required for high-speed signal drivers and package pins. Improving per-pin signaling rates is possible, but only at a significant cost in energy-efficiency, and so will not necessarily improve aggregate off-chip bandwidth.

Even if we remove pin bandwidth limitations, memory system performance could be constrained by the energy consumption of other components in current DRAM architectures. Apart from the I/O energy required to send a bit to the CPU, considerable energy is required to traverse the DRAM chip from the memory bank to the I/O pin, and to access the bit within the bank. Most bank access energy is due to the wasteful sensing of bits that are never read out.

In this paper, we propose using a monolithically integrated silicon photonics technology to attack all of these issues. Dense wavelength division multiplexing (DWDM) allows for multiple links (wavelengths) to share the same media (fiber or waveguide) for a huge bandwidth density advantage, eliminating pin bandwidth as a constraint. Silicon photonics also demonstrates significantly greater energy efficiency, supporting far larger off-chip bandwidths at a reasonable power budget. Monolithic integration allows energy-efficient photonic links to extend past the chip edge and deep into the banks to greatly reduce cross-chip energy. By redesigning the DRAM banks to provide greater I/O bandwidth from an individual array core, we can supply the bandwidth demands with much smaller pages thereby improving the energy efficiency of activating a bank. Surprisingly, this does not come with an area penalty, as the higher bandwidth from each array core means we can reduce area overheads by employing fewer larger array blocks.

DRAMs are commodity parts, and ideally a single mass-produced part should be useable in a wide variety of system configurations. We propose a technique of *optical power guiding* to enable greater scalability in DRAM configurations. We use photonic point–to–point links to implement a logical bus that allows a designer to change the ratio between capacity and bandwidth while using the same DRAM part, thereby reducing cost by increasing manufacturing volumes.

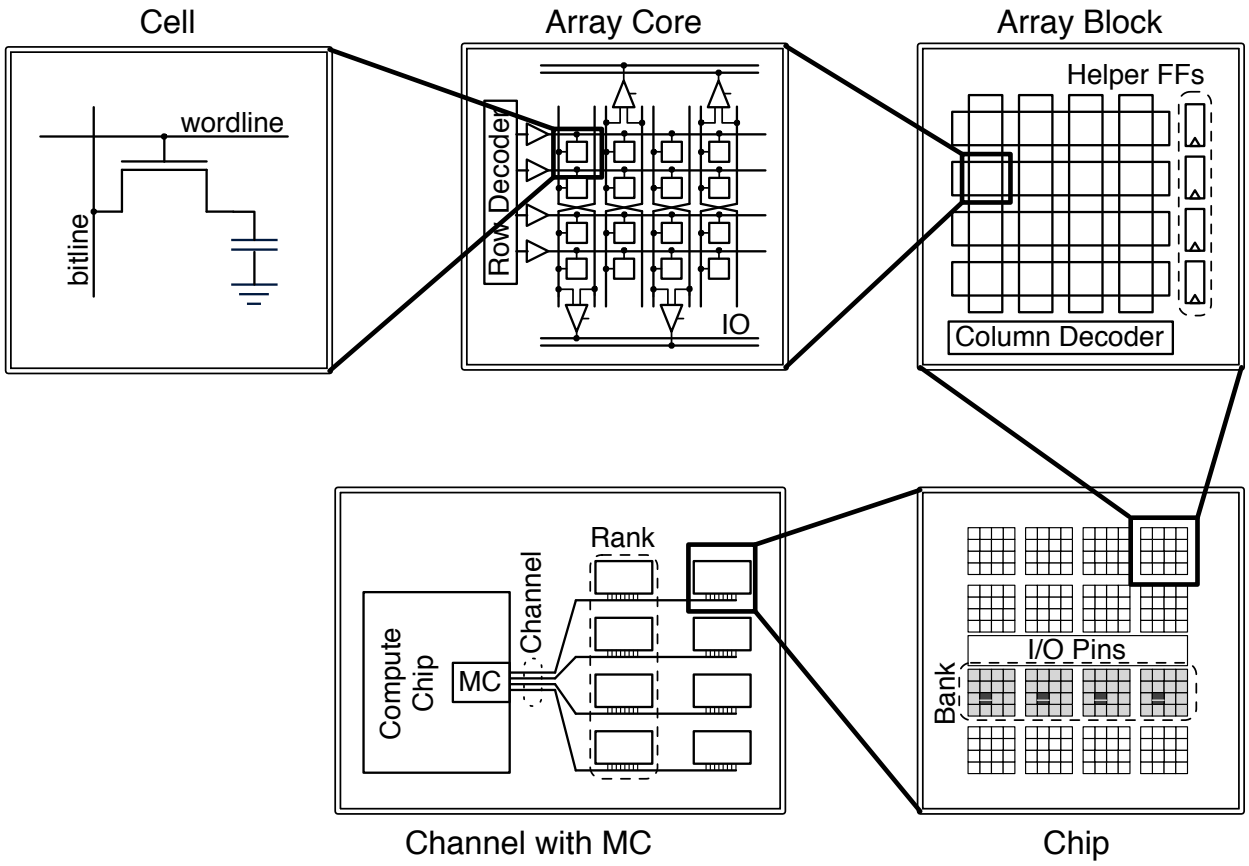


Figure 1: DRAM Hierarchy - Each inset shows detail for a different level of the current electrical DRAM organization.

Our results show a promising photonic design point that provides a 10× improvement in bandwidth at the same power while reducing the DRAM die area, and while supporting a wide variety of DRAM system configurations with a single DRAM part.

2. DRAM Technology

We begin by reviewing the structure of a modern DRAM chip, as shown in Figure 1. Modern DRAMs have multiple levels of hierarchy to provide fast, energy-efficient access to billions of storage cells.

2.1. Current DRAM Organization

At the very core of a DRAM is the *cell*, which contains a transistor and a capacitor and holds one bit of storage. Cells are packed into 2D arrays and combined with the periphery circuitry to form an *array core*. The array core contains a grid of cells such that each row shares a wordline and each column shares a bitline. Each row has wordline drivers to drive these heavily loaded wires. Each column has sense-amplifiers which are used to read and write an entire row of cells at a time. Differential sense amps are used to amplify and

latch low-swing signals when reading from the bitlines, and to regenerate full rail voltages to refresh the cell or write new values into the cell. Even though every cell in an activated row of an array core is read or written on any activation, only a handful will go in or out of the array core on the I/O lines during any column access. Although some row hits are possible for some workloads, most of the other bits read from a row are never accessed before a different row is read.

The array core is sized to get maximum cell density for a reasonable delay and energy per activation or refresh. In this paper, we model a folded bitline DRAM, which provides better common mode noise rejection for the sense amp. However, our general assumptions are also valid for the open bitline architecture which is making a comeback due to better scalability and area efficiency. We assume an array core size of 512 wordlines by 1024 bitlines. Array cores are limited to a modest size that grows very slowly with respect to technology scaling due to intrinsic capacitances.

An *array block* is a group of array cores that share circuitry such that only one of the array cores is active at a time. Each array core shares its sense amps and I/O lines with the array cores above and below it, and the array block provides its cores with a global predecoder and shared helper flip-flops for latching data signals entering or leaving the array block.

A *bank* is an independently controllable unit that is made up of several array blocks working together in lockstep. The access width of an array block is determined by the number of I/O lines that come from its array cores, and there are enough array blocks per bank to achieve the bank's access width. Array blocks from the same bank do not need to be placed near each other and, as shown in Figure 1, they can be striped across the chip to ease interfacing with the I/O pins. When a bank is accessed, all of its array blocks are activated, each of which activates one array core, each of which activate one row. The set of activated array cores within a bank is the *sub-bank* and the set of all activated rows is the *page*.

A *channel* is a bus that connects a group of banks to a memory controller. Each DRAM chip can only provide so much bandwidth due to pin bandwidth and area constraints, so multiple DRAM chips are typically ganged together in a *rank* to increase the bandwidth of the channel. In this situation, a slice of a bank is present on every chip in the rank. For greater bandwidth, the system can have multiple memory controllers and channels. To increase capacity, multiple ranks can be placed on the same channel, but only one of them can be accessed at a time.

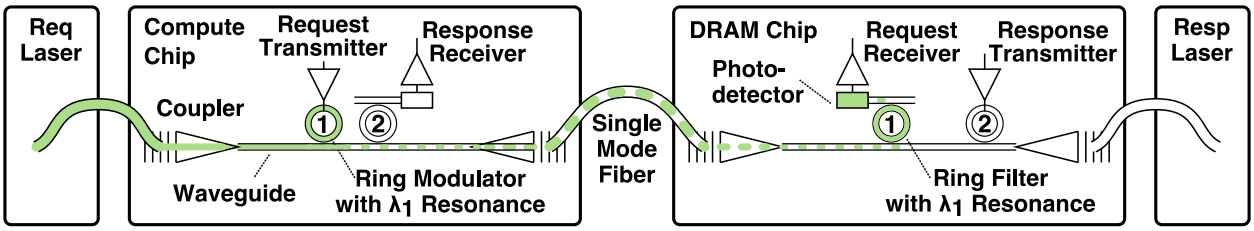


Figure 2: Photonic DRAM Channel – Two wavelength-division multiplexed links which create a bidirectional channel between a memory controller in a compute chip and memory banks in a DRAM chip. λ_1 is used for the request and λ_2 is used for the response in the opposite direction on the same waveguides and fiber.

3. Silicon Photonic Technology

Monolithically integrated silicon photonics is a promising new technology for chip-level interconnects. Photonics offers improved energy-efficiency and bandwidth-density compared to electrical interconnect for intra-chip and especially inter-chip links. In this section, we first describe our assumed photonic technology, and then discuss the various costs involved in implementing a unified on-chip/off-chip photonic memory channel.

The various components in a silicon photonic link are shown in Figure 2. For a memory request, light from an external broadband laser source is coupled into a photonic waveguide on the compute chip. The light passes along a series of ring resonators which each modulate a unique wavelength. The modulated light is then transmitted to the DRAM chip on a single-mode fiber. At the receiver side, the light is filtered by the tuned ring filters and dropped onto the photodetector. The photodetector converts light into electrical current which is sensed by the electrical receiver. For the response, light is sent in the reverse direction on the same waveguides and fiber from the DRAM chip back to the compute chip. In this example, two wavelengths are multiplexed onto the same waveguide, but the real potential of silicon photonics is in its ability to support dense wavelength division multiplexing with dozens of wavelengths per waveguide. There are times when it as advantageous to filter a set of wavelengths at a time, and a comb filter is a single, large diameter ring that can accomplish this.

Both 3D and monolithic integration of photonic devices have been proposed in the past few years to implement processor-to-memory photonic networks. With 3D integration, the processor chips, memory chips, and a separate photonic chip are stacked in a variety of configurations. The photonic devices can be implemented in monocrystalline silicon-on-insulator (SoI) dies with thick layer of buried oxide (BOX) [6], or in a separate layer of silicon nitride (SiN) deposited on top of the metal stack [3]. Since the photonic devices are on a separate layer, engineers can employ customized processing steps to improve photonic device

performance (e.g. like introducing ridge waveguides or epitaxial Ge for photodetectors). With monolithic integration, photonic devices have to be designed using the existing process layers of a standard logic and DRAM process. The photonic devices can be implemented in polysilicon on top of the shallow-trench isolation (STI) in a standard bulk CMOS process [8, 14] or in monocrystalline silicon with advanced thin BOX SoI. Photodetectors can be implemented using silicon-germanium (SiGe) present today in the majority of sub-65 nm processes (and proposed for future DRAM processes [9]). Although monolithic integration may require some post-processing, its manufacturing cost can be lower than 3D integration. Monolithic integration decreases the area and energy required to interface electrical and photonic devices, but it requires active area for waveguides and other photonic devices. In a DRAM process, it also requires an additional step of depositing undoped polysilicon (since, unlike in a logic process, all polysilicon layers in a DRAM process are deposited as heavily doped to minimize fabrication cost and resistivity of polysilicon interconnect).

A photonic link requires several types of power consumption: static laser power, active and static power in the electrical transmitter and receiver circuits, and static thermal tuning power which is required to stabilize the frequency response of the thermally sensitive ring resonators. Table 2 shows our predictions for the active and static power spent in the electrical circuits and in the in-plane heaters for thermal tuning. Static laser power depends on the amount of optical loss that any given wavelength experiences as it travels from the laser, through various optical components, and eventually to the receiver (see Table 1). Some optical losses, such as coupler loss, non-linearity, photodetector loss and filter drop loss, are independent of the main-memory bandwidth and layout. We will primarily focus on the loss components (waveguide loss and through ring loss) that are affected by the required main-memory bandwidth and the layout of the photonic interconnect. These components contribute significantly to the total optical path loss and set the required optical laser power and correspondingly the electrical laser power (at a roughly 30-50% conversion efficiency) requirements.

For our photonics links, we assume that with double-ring filters and a 4 THz free-spectral range, up to 128 wavelengths modulated at 10 Gb/s can be placed on each waveguide (64λ in each direction, interleaved to alleviate filter roll-off requirements and crosstalk). A non-linearity limit of 30 mW at 1 dB loss is assumed for the waveguides. The waveguides are single mode and a pitch of 4 µm minimizes the crosstalk between neighboring waveguides. The diameters of a regular modulator/filter ring is $\approx$10 µm and that of a comb filter ring is $\approx$40 µm. In the following photonic layouts, we conservatively assume that these photonic components can fit in a 50 µm STI trench around each waveguide, when monolithically integrated. We also project from

Photonic device	Optical Loss (dB)
Optical Fiber (per cm)	0.5e-5
Coupler	0.5 – 1
Splitter	0.2
Non-linearity (at 30 mW)	1
Modulator Insertion	1
Waveguide (per cm)	2 – 4
Waveguide crossing	0.05
Filter through	1e-4 – 1e-3
Filter drop	1
Photodetector	1
Laser Efficiency	30% – 50%
Receiver sensitivity	-20 dBm

Table 1: Optical Losses

Electrical I/O	Tx+Rx DDE	Tx+Rx FE			
Aggressive	2100 fJ/bt	2900 fJ/bt			
Photonic I/O	Tx DDE	Tx FE	Rx DDE	Rx FE	Tx/Rx TTE
Aggressive	40 fJ/bt	5 fJ/bt	20 fJ/bt	10 fJ/bt	16 fJ/bt/heater
Conservative	100 fJ/bt	20 fJ/bt	50 fJ/bt	30 fJ/bt	32 fJ/bt/heater

Table 2: Aggressive and Conservative Energy and Power Projections for Electrical and Photonic I/O – fJ/bt = average energy per bit-time, DDE = Data-traffic dependent energy, FE = Fixed energy (clock, leakage), TTE = Thermal tuning energy (20K temperature range). Electrical I/O projected from [12] 8 pJ/bt, 16 Gb/s design in 40 nm DRAM process, to 5 pJ/bt, 20 Gb/s design in 32 nm DRAM process. Optical I/O runs at 10 Gb/s/wavelength.

our ongoing circuit designs that the area of the photonic E/O transceiver circuit is around 0.01 mm^2 for modulator driver, data and clock receivers and associated SerDes datapaths. From [12] we also assume that area for an electrical I/O transceiver will be mostly bump-pitch limited at around 0.25 mm^2.

4. Redesigning the Bank for Silicon Photonics

Photonics can supply a very high bandwidth memory interconnect, but this requires a large number of concurrently accessed banks to provide matching memory bandwidth. Photonics also reduces interconnect energy consumption, and so energy costs within the banks begin to dominate overall memory system power. In this section, we describe how we modify the DRAM bank design to reduce energy consumption.

During a bank access, every constituent array block activates an array core, which activates an entire array core row, of which only a handful of bits are used for each access. The energy spent activating the other bits is wasted, and this waste dominates bank energy consumption. Reducing wasted accesses while keeping the bank access size constant requires either decreasing the array core row size or increasing the number of I/Os per array core and using fewer array cores in parallel. Reducing the array core row size will cause greater area penalty, so we propose bringing more I/Os to the array core to improve access efficiency.

Increasing the number of I/Os per array core, while keeping the bank size and access width constant, will have the effect of decreasing the number of array blocks per bank. Currently there is little incentive to make this change because the energy savings within the bank are small compared to the current cross-chip and off-chip energy. Even if there were energy savings, current designs are also pin-bandwidth limited, so there would be no benefit to supporting such a wide bank access width.

We propose the array block as the appropriate level at which to interface the DRAM circuitry to our

photonic data links. Pushing photonics into the array block to directly connect to array cores would be highly inefficient due to the relatively large size of the photonic devices. On the other hand, interfacing photonics at the bank level leaves behind potential improvement, and current designs already connect I/O pins to array blocks rather than banks.

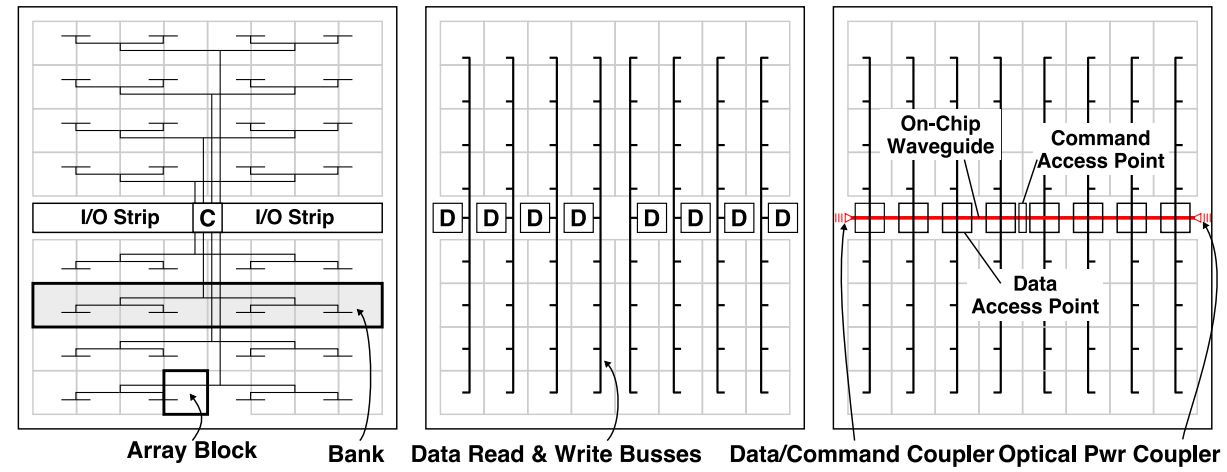
5. Redesigning the Chip with Silicon Photonics

Given improved banks and energy efficient silicon photonic links, the next step is to design the memory chip. The on-chip interconnect connects the banks of a chip to the off-chip links for their channel. Given an off-chip channel implemented with a photonic link, there is a design tradeoff of how far on chip that link should go, with the remaining distance traversed electrically. At one extreme, we only bring photonics to the edge of the chip, and at the other we bring photonics all the way to the array block. The design choice is primarily driven by area and power, as the bandwidth of the on-chip interconnect should meet or exceed the off-chip bandwidth, and its latency should be negligible when considering the end-to-end latency of accessing DRAM.

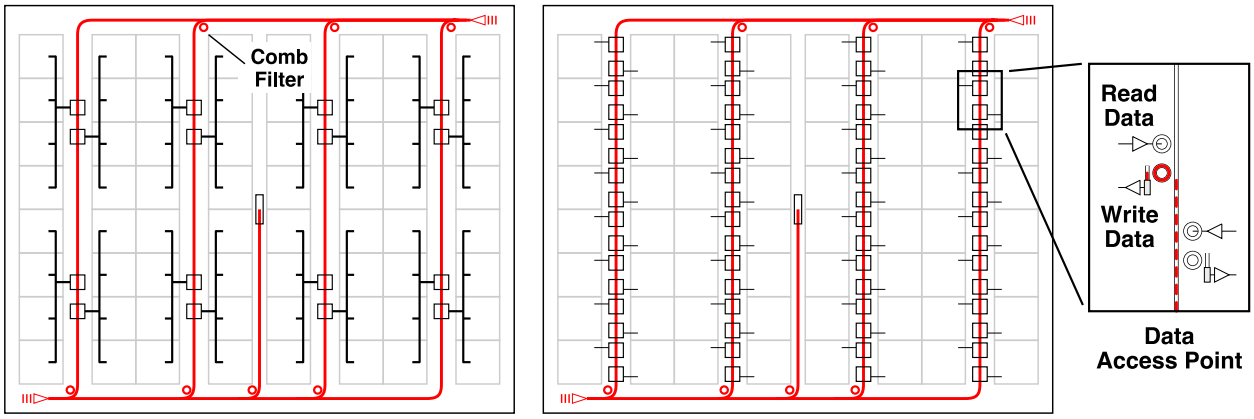
A bank is the collection of array blocks that are accessed in parallel. Each array block will always send and receive its data from the same off-chip link, so the interconnect can be localized to attach only a subset of the array blocks to each off-chip link. Figure 3b shows the all-electrical baseline E1, which contains an I/O strip for off-chip drivers and has the banks spread across the chip, one per row. All of the array blocks in the same column share the same slice of the I/O circuitry, but they each belong to different banks so they will never be accessed at the same time. The control information must be available to every bank in the system. The *address tree* (Figure 3a) takes this control information vertically, but to save power it will only fan out horizontally to the row of array blocks in the bank targeted by a given command.

Converting the off-chip links of E1 to photonic links results in configuration P1. The I/O strip has been replaced with a waveguide with multiple *access points*. Each access point connects a number of wavelengths from the photonic channel to a slice of the on-chip electrical interconnect. Since the photonic channel is full duplex, the electrical on-chip interconnect must be doubled to allow for simultaneous bidirectional communication.

Taking the photonic links deeper into the chip results in the PBx series of designs, where the waveguide is spread out into a *waterfall* configuration. The horizontal waveguides contain all of the wavelengths, and the comb filters at the base of each vertical waveguide ensure that each vertical waveguide only contains a subset



(a) Command and Address Path (b) Fully Electrical Datapath (c) Photonic Datapath to the Chip



(d) Photonic Datapath to Group of Array Blocks

(e) Photonic Datapath to Each Array Block

Figure 3: Cluster Design Options All of the designs use a similar address tree, so it is shown in Figure 3a and omitted from the rest of the figures. (b) E1; (c) P1; (d) PB2; (e) PB8

of the channel's wavelengths. Each of these vertical waveguides is analogous to the vertical interconnect slices from P1, so a bank can be striped across the chip horizontally to allow easy access to the on-chip interconnect. The x in the name corresponds to how many rows of access points there are. If there are less rows of access points than there are banks, the access points are shared amongst the neighboring array blocks. Photonic access points for array blocks not currently in use must detune their rings to reduce their loss contribution to the channel. To use the same command tree as P1 and E1, another waveguide runs up the middle to the center of the chip. Because command distribution require broadcast, and because broadcast with more than a couple of endpoints would require too much optical power, we retain an electrical broadcast bus for commands.

There should be a vertical waveguide between each pair of array blocks, with at least one wavelength

travelling in each direction, so for some configurations the array blocks will have to communicate farther horizontally to reach an array block.

The motivation for bringing the photonic links deep into the chip to the array blocks is to exploit the distance insensitivity of optical communication. Most of the energy and delay costs of a photonic link arise during the electro-optical and the opto-electrical conversions at the endpoints, so traveling further on-chip is practically free as these conversion costs will have already been paid for the off-chip traversal. Since each access point has to share the same set of wavelengths with other access points on other strips in the same column, all of the photonic circuits have to be replicated at each access point. This can potentially lead to large loss due to the large number of ring-through losses if the devices are not carefully optimized, as we will explore in Section 7. High loss must be overcome by high laser power, so too many photonic access points could result in increased static power. To share this static power, multiple array blocks in the same column can share a photonic access point because they belong to different banks and thus can not be sending or receiving at the same time. The amount of sharing is the number of banks divided by x.

5.1. Clusters

With the tremendous bandwidth provided by silicon photonics, it might even be reasonable to split a chip's bandwidth into two channels. We term a group of banks that share the same channel, a *cluster*, so a chip with multiple channels will have multiple clusters. Splitting up I/O bandwidth across separate channels will increase concurrency, but serialization latency will go up as each channel's bandwidth is decreased. Fortunately photonics provides so much bandwidth that it is possible to have multiple channels with reasonable serialization latency on the same chip. There are also some potential power savings from breaking a chip into multiple clusters, as each cluster now supports fewer banks and lower bandwidth, and will also have lower area overhead.

5.2. Power Delivery Considerations

Current DRAMs have a constraint t_{RRD} which represents the minimum delay between activate commands sent to different banks on the same chip. This constraints reflects the limitations of the power delivery network on the DRAM chip. During an activate, the chip draws high instantaneous power, so these operations must be spread out for the sake of the power delivery grid and the power regulators. This power delivery network is minimally sized such that t_{RRD} does not significantly limit performance on the target workload, but any excess translates to increased cost. Compared to current designs, our memory system has

more banks, and as a consequence they will be doing more activates per unit time. However, because our page size is much smaller, it will actually put less strain on the power delivery network. A smaller page will require less instantaneous power, so our system will have a smaller and less erratic power draw.

6. Redesigning the System to Fully Utilize Silicon Photonics

Building upon the components introduced in the previous sections, we now explain the structure of our proposed memory system. We assume that both the number and silicon area of DRAM chips will be comparable to those of today and that the desired access width is still a cache line, which for this work we set to 64 bytes.

6.1. Structure and Packaging

For economic and packaging reasons, it seems reasonable to assume that each DRAM chip will have only two fibers: one for data and one for optical power. A single fiber can provide 80 GB/s of bandwidth in each direction, which should be sufficient even for applications with the highest bandwidth requirements. Multiple channels can share the same fiber as long as they use different wavelengths, so a DRAM chip could even have multiple clusters. To further increase the capacity of DRAM chips in a system without increasing the fiber wiring overhead, we introduce the concept of a DRAM cube, which is a collection of stacked DRAM chips (e.g., 8 [10]), connected electrically by through-silicon vias and optically by vertical coupling in through-silicon via holes.

In Figure 4, we illustrate this concept on an example board-level single-socket system containing 512 GB of memory. The memory is built from 8 Gb DRAM chips stacked 8 deep in DRAM cubes, with 4 DRAM cubes on each super-DIMM, and 16 super-DIMMs total. On the CPU side, we assume 16 memory channels, each connected to the DRAM cubes in one super-DIMM.

Just as with electrical DRAM channels, each channel has a memory controller, where the controller handles all of the scheduling and arbitration centrally to keep the DRAM chip simple and remove the need for global arbitration. One wavelength heading to the DRAM chip will be reserved to issue commands to the banks in the cluster. Very high bandwidth channels supporting many banks might need to add a second wavelength to ensure sufficient command bandwidth.

Each DRAM chip's power fiber passes through the compute chip to simplify packaging and assembly. Providing one laser per DRAM chip is not economical, so a handful of large but efficient lasers are used

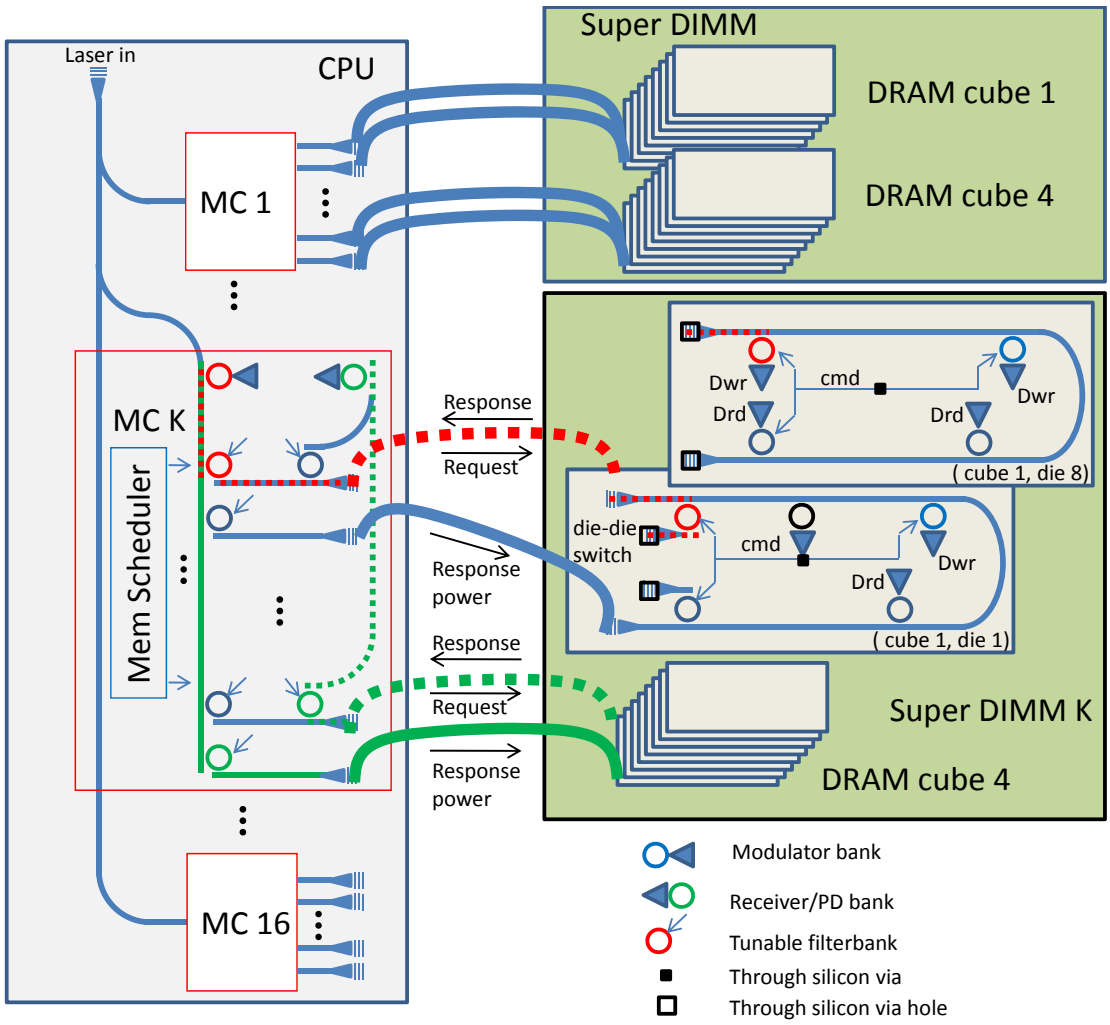


Figure 4: Optical power guiding The highlighted path is the response path from the DRAM chip to the compute chip.

with their optical power brought on to the compute chip and subsequently divided among the DRAM chips. This approach places both fibers destined for the same DRAM chip in physical proximity, allowing them to be grouped into a fiber ribbon to ease assembly.

6.2. Achieving Scalability with Optical Power Guiding

Different workloads and different systems will have different memory bandwidth and capacity requirements, but it would be ideal if all these systems could use the same DRAM part to increase sales volume to drive down cost. A given part has a set amount of bandwidth and capacity, so simply replicating that part will keep the same ratio between the two. If a system wants more bandwidth it will have to also increase capacity, while if a system wants more capacity, it will also have to increase its bandwidth. Being able to decouple these two parameters is essential because excess in either one leads to increased cost and power

consumption.

Current electrical systems have solved this problem by putting multiple ranks on the same channel, which acts as a bus where only one rank can communicate at a time. The system bandwidth is constant, but more ranks can be added to the system to increase capacity.

The challenge is to build an equivalent bus out of a photonic technology proposal. Previous photonic work has proposed daisy-chaining the channel through multiple DRAM chips [17]. This is problematic because the path from the memory controller to the last DRAM chip will have significant optical losses. In addition to the losses associated with rings, each DRAM chip will contribute two coupler losses along with a decent amount of waveguide loss, so this technique will require very low loss components to be feasible. To provide the flexibility of a bus but with less stringent optical loss requirements, we propose *optical power guiding*.

With optical power guiding, we provision the minimum amount of optical power, and use comb filters to guide it to where it is needed. The key components enabling this flexibility are *power and message demultiplexers*. By only allowing one of the comb filters to be on at a time, the input light will only exit on one of the output DRAM waveguides leaving the other DRAM chips dark. This is true not only for the path from the memory controller to the DRAM chip, but also for the response path as shown in Figure 4.

Optical power guiding builds a logical bus from a set of physical point-to-point links. It is a bus since only one of the links can be used at a time, but the loss to any DRAM chip is much lower because it is a direct connection. This technique reaps the static power benefits of a bus, because there is only enough optical power to communicate with one DRAM chip per direction. With power guiding, the channel can be shared amongst multiple clusters, but it requires only a single central memory controller.

Even though the data portion of the channel can be switched between chips, the command wavelengths must always be sent to every cluster. This does impose a static power overhead, but the amount of power for commands is much smaller than for data, e.g., for random traffic a command:data ratio of 1:16 will typically not be command-bandwidth limited.

Figure 4 as drawn provides the ability to switch which DRAM cube receives the channel and further command decoding on the first chip in the cube (labeled (1,1)) configures the intra-cube power/message guiding to the appropriate die. Both directions must be switched in lockstep and are controlled by the memory scheduler on the controller side. The modulation and receive circuits are centralized at the controller side in order to save on background circuits power.

Power guiding requires no changes to the DRAM chip. All of the customization for various system configurations occurs on the compute chip. Systems that are not fully populated can still utilize their full memory bandwidth as long as each channel is attached to at least one DRAM chip.

There is a tradeoff for what the bandwidth to capacity ratio should be for the standard DRAM part. The capacity is set by the chip size and cell density, so the real independent variable is the chip's bandwidth. Setting the bandwidth to be relatively high will enable the greatest flexibility. Most systems will use some optical power guiding to increase capacity and systems that want to increase the bandwidth for the same capacity will just decrease the amount of optical power guiding. At the same time systems that want more capacity can increase the amount of power guiding. On the other hand, setting the bandwidth low will decrease the amount of power guiding most systems will need, which will save power, but it will not support as high bandwidths for the same capacity.

7. Evaluation

7.1. Methodology

To evaluate the impacts of our architectural parameters on the energy and area of a proposed DRAM architecture, we use a heavily modified version of the Cacti-D DRAM modeling tool [16]. Though we were able to use some of Cacti-D's original models for details such as decoder sizing, gate area calculations and technology parameter scaling, the design space we explored required the complete overhaul of Cacti-D's assumed DRAM organization and hierarchy. To this end, we built our own architectural models for the DRAM core, from circuit-level changes at the *array core* level, to *array block* and bank organization at higher levels as shown in Figure 1, while relying on Cacti's original circuit models to handle most of the low-level circuit and process technology details. In addition to covering the baseline electrical DRAM configurations, we account for the overhead of each relevant photonic configuration in our models and developed a comprehensive methodology for calculating the power and area overheads of off-chip I/O for both the electrical and photonics cases of interest. All energy and area calculations were run for a 32 nm DRAM process.

To quantify the performance and energy efficiencies of each DRAM configuration, we use a detailed cycle-accurate microarchitectural C++ simulator. We use synthetic traffic patterns to issue loads and stores at a rate capped by a limited number of in flight messages. We implement a uniform random traffic pattern with varying rates of loads and stores and a few conventional stream patterns such as copy and triad. The

memory controller converts these messages into DRAM commands which are issued based on a round robin arbitration scheme and various timing constraints based on contemporary timing parameters found in the Micron DDR3 SDRAM data sheet [1]. Events and statistics from the simulator are used to compute performance in terms of achieved bandwidth and energy-efficiency. Average power and energy efficiency of each traffic pattern is evaluated for each DRAM configuration using energy and power numbers given by the modified Cacti DRAM models. Since silicon photonics is an emerging technology, we also explore the space of possible results with both aggressive and conservative projections for photonic devices and photonic link circuits.

Using this architectural simulator we simulate a range of different cluster organizations, by varying: the traffic pattern, floorplan, number of I/Os from array block, number of banks per cluster, and the channel bandwidth. For random traffic, the effective bandwidth of a bank with 512 bits access width has approximately the data bandwidth of one 10 Gb/s wavelength, independent of system size, which matches our analytical model. For streaming traffic the effective bank bandwidth was higher, and increased as the channel bandwidth was increased. Serialization latency has a bigger impact on bank occupancy for streaming traffic. For the rest of the design points presented we set the number of banks to at least match the number of wavelengths, in order to have sufficient concurrency to obtain full utilization for random traffic.

For this work, latency was not an important figure of merit because our designs do not change it significantly. We do not change the array core internals, which sets many of the inherent latencies for accessing DRAM. Our bank bandwidths are sufficiently sized such that the serialization latency is not significant, especially for random traffic. As to be expected, as the channel approaches peak utilization, the latency does rise dramatically due to contention.

7.2. Energy

Figure 5 shows the energy efficiency breakdown of each interconnect possibility of three different DRAM configurations. Two of the designs are high bandwidth parts that differ only in the number of IOs per array core, and the third is a low bandwidth part. The results are from a random traffic pattern that achieves high utilizations.

Across all configurations it is clear that replacing the off-chip links with photonics is a clear win, as E1 towers above the rest of the designs. How far photonics is taken on chip, however, is a much richer design space. To achieve the optimal energy efficiency requires balancing both the dynamic and static components

of the overall energy cost. As shown in Figure 5, P1 spends the majority of the energy on getting data across the chip (Write Energy and Read Energy) because the data must traverse long global wires to get to each bank. Taking photonics all the way to each array block with PB64 improves upon the energy efficiency from P1 by minimizing the cross-chip energy, but results in a large number of photonic endpoints (since all the photonic endpoints in P1 are replicated in each access point, i.e. 64 times in case of PB64), contributing to the large static component in the overall energy efficiency (due to background energy cost of photonic transceiver circuits and energy spent on ring thermal tuning). By sharing the photonic endpoints effectively with 8 photonic strips across the chip, an optimal configuration is achieved at PB8, which further improves upon the energy efficiency of reads and writes by not accumulating too much background power. Once these energies have been reduced (as in the PB8 case for the second configuration), however, the activation energy becomes dominant. Expanding the number of data bits we take out from each array core from 4 to 32 has the desired effect, further reducing the activate energy cost, and overall this optimized design is a factor of 10 times more energy efficient than the baseline electrical design.

Similar tradeoffs appear for the low bandwidth system, keeping PB8 at the optimum. Interestingly, the number of banks sharing the same photonic access points had no effect on the optimum number of photonic IO stripes. For example, the high bandwidth designs have 64 banks, meaning that each access point is shared among array blocks from 8 different banks for the PB8 floorplan. For the low bandwidth design with 8 banks, the PB8 floorplan has no sharing of access points - each bank has its own dedicated access point - yet it yields the optimal energy efficiency for the design, since the lower number of photonic endpoints prevents the background energies from dominating.

Figure 6 shows the same configurations as Figure 5, but for greatly pessimistic technology assumptions for silicon photonics. Replacing the off-chip links with silicon photonics still helps greatly, but bringing photonics across the chip closer to the array blocks is less helpful. The optimal floorplan still appears to be PB8, but it has a smaller margin over P1. Changing the number of IOs per array core still proved to be beneficial, but this improvement is diluted.

In summary, photonics can get a big win replacing the off-chip links and as the technology improves it will encourage its use deeper in the chip. Increasing the number of IOs per *array core* can help by reducing the activation energy, but the effect is usually minuscule unless the interconnect energies have been minimized as well.

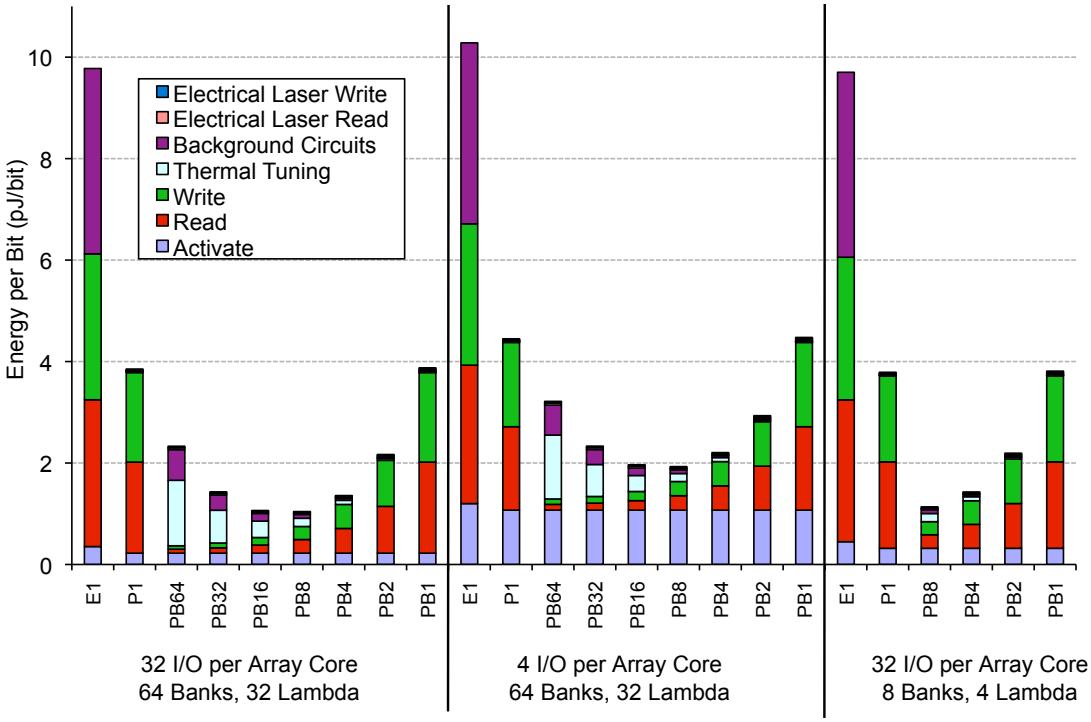


Figure 5: Energy/Bit vs Floorplan (aggressive photonics) – peak achieved bandwidth of ≈500 Gb/s for first two configurations and ≈60 Gb/s for the last configuration. Background Circuits is the IO background energy from Fixed Energy in Table 2. Read Energy consists of IO read, cross-chip read and bank read energy. Write Energy consists of IO write, cross-chip write and bank write energy and Activate Energy consists of IO command, cross-chip row address energy and bank activate energy.

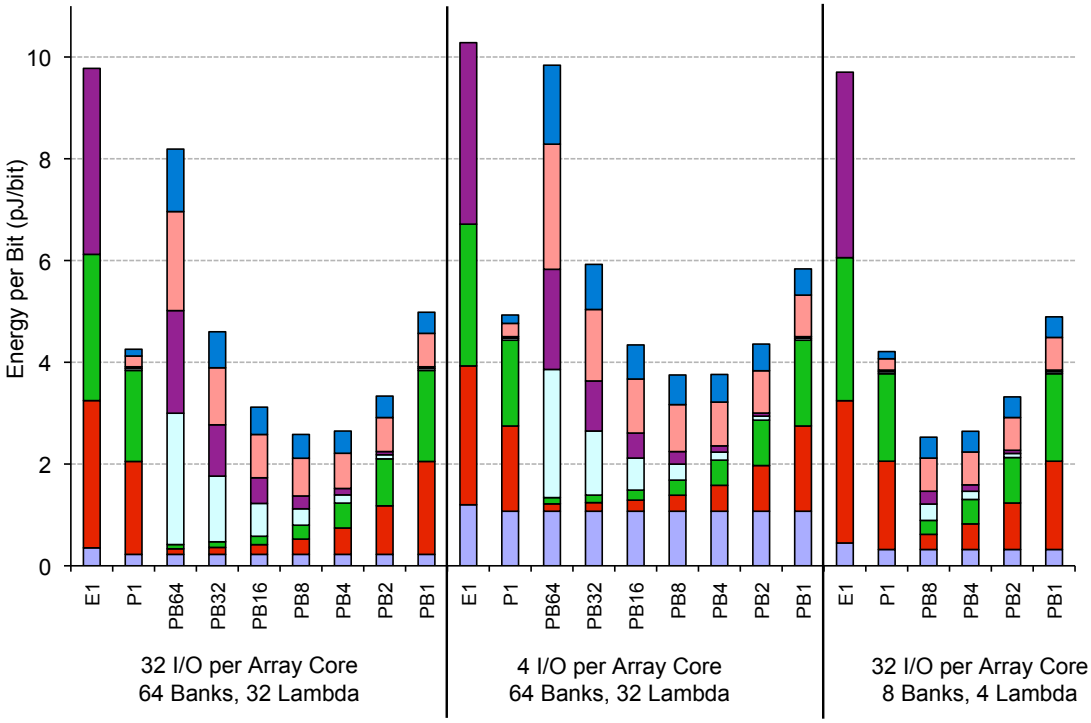


Figure 6: Energy/Bit vs Floorplan (conservative photonics) – same peak achieved bandwidth as Fig. 5.

7.3. Utilization

To obtain true energy efficiency, memory system must still be efficient when not fully utilized, which is often the common case. For a given configuration, we scaled back the utilization by reducing the number of messages that could be in flight at a time, and the results are shown in Figure 7. As to be expected, as utilization goes down, the energy per bit increases because each bit takes longer so it must amortize more static power. Systems with more static power will have a steeper slope, and this tradeoff can clearly be shown when comparing 32 64 2 32 PB8 and 32 64 2 32 PB16 (or any other PB8 and PB16 designs in the plot). The only difference between the two configurations is whether there are 8 or 16 rows of photonic access points per chip. The PB16 configurations do better for the high utilization cases because the global electrical wires connecting the array blocks to the photonic access points are shorter, but do worse than PB8 configurations for low utilization because the static power of their rings and idle photonic circuits add up. Essentially this is a trade-off between the static and dynamic power and the target system utilization will determine the right configuration. The plots also indicate that the most energy-efficient configurations for any target bandwidth are those with two clusters (memory channels) per chip. The two independent channels with half the bandwidth, require less total photonic devices replicated throughout the chip (compared to a single channel with twice the bandwidth), yet still have comparable (and potentially higher) achieved bandwidth under random traffic.

Figure 8 shows the effects of less capable photonic devices, which result in a relatively large penalty for low utilization of high bandwidth systems. This figure also shows that configurations with less endpoints are now more energy-efficient, both ones with less access points (e.g. PB8 instead of PB16) and ones with two narrower channels instead of one wide channel.

7.4. Area

Since the objective of the memory design in this work is to increase the energy efficiency of accessing DRAM to enable increased memory bandwidth while remaining relatively area neutral, it is important to examine the effects of the various parameters on the total DRAM area, as shown in Figure 9. A point to notice is that since we scaled up the number of bits we take out of each bank per request (512 bits vs. 64 bits in contemporary DRAM), it becomes preferable to scale by increasing the number of bits we get from each array block (i.e. array core) instead of increasing the number of array blocks. The area overhead needed to get more bits from each array block (while maintaining the same bank access size) is much smaller than

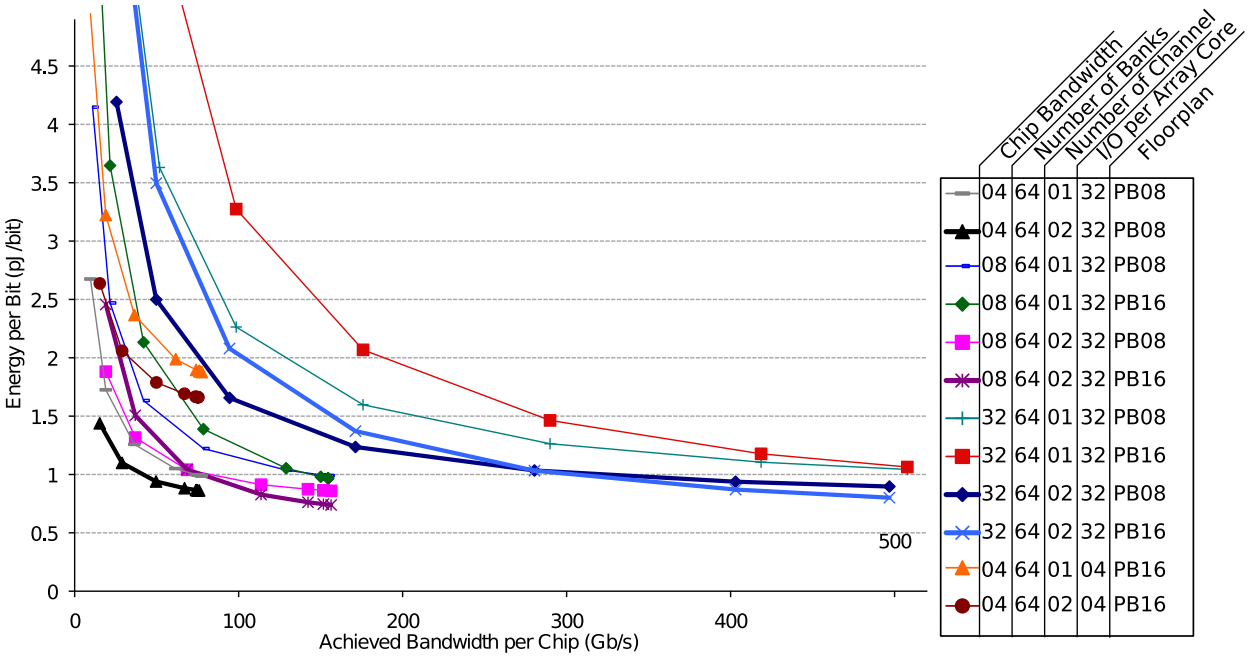


Figure 7: Energy/bit Breakdown vs Bandwidth (aggressive)

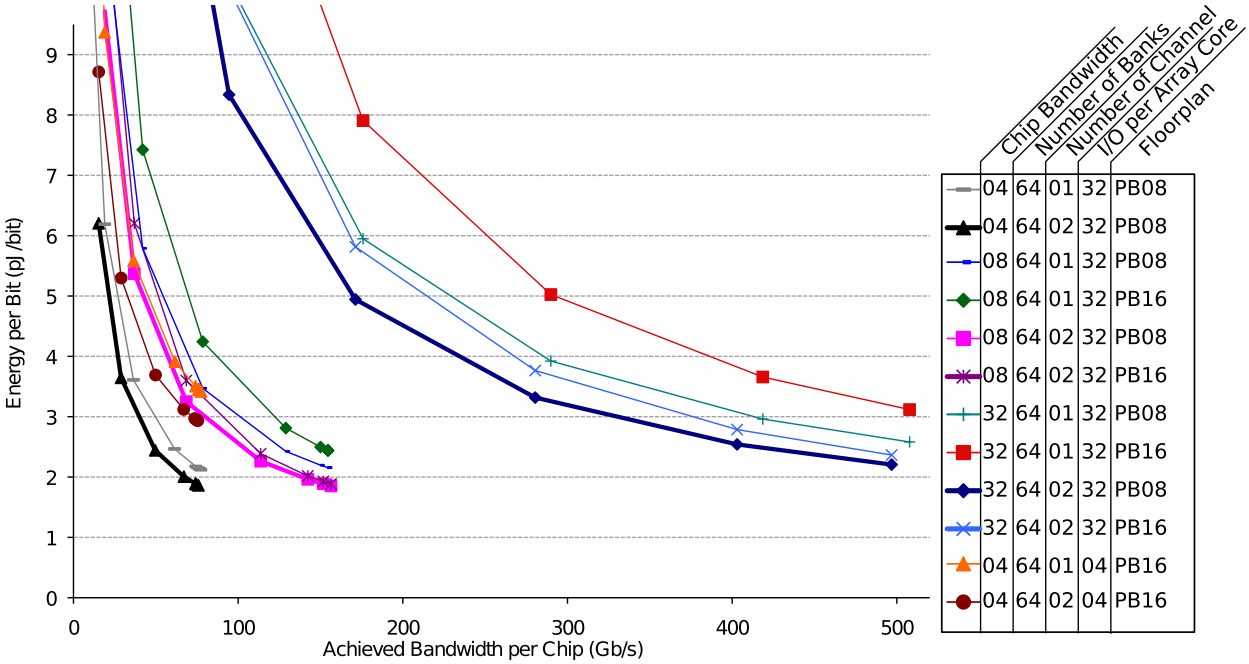


Figure 8: Energy/bit Breakdown vs Bandwidth (conservative)

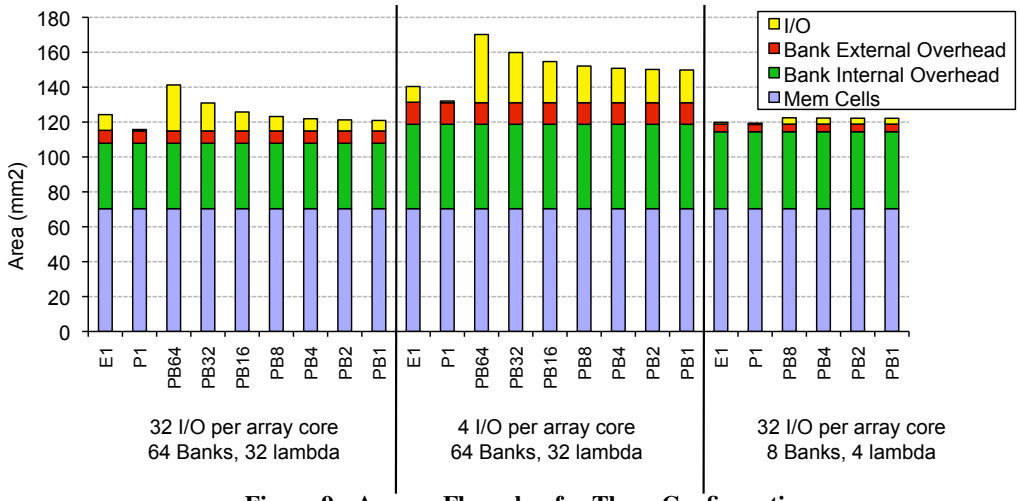


Figure 9: Area vs Floorplan for Three Configurations

having more smaller array blocks. Hence, we see a drastic decrease in total area as we increase the number of bits from each array block from 4 to 32 for our high bandwidth configuration.

It is also worth noting that adding photonics to the second configuration with many array blocks results in significant overhead since waveguide trenches are trying to reach each of the many array blocks in a bank. However, by increasing the number of bits from each array block, we reduce the number of array blocks per bank, and hence have less waveguide trenches to array blocks. This leads to much smaller photonic area overheads, as seen in the first configuration where PB8 (which has the best energy-efficiency) has slightly smaller die area than electrical baseline E1. The electrical baseline IO area is mostly bump-pitch limited, and unlikely to scale much in more advanced process nodes, setting the upper bound on memory area efficiency for required IO bandwidth. Photonic access points, on the other hand are allowed to scale due to dense wavelength division multiplexing, small size of vertical couplers and continued scaling of electrical back-end circuits. It is worth noting that our photonic area calculations assumed a very conservative 50 µm trench for each vertical guide.

8. Related Work

Techniques such as microthreading [18] and multicore DIMMs [2] reduce the effective row size in current electrical memory modules by partitioning a wide multichip DRAM interface into multiple narrow DRAM interfaces each with fewer chips. Since both approaches use commodity electrical DRAM chips they result in either shorter bank access sizes or longer serialization latencies. In addition, the energy savings due to a reduced ratio of activated to accessed bits is mitigated by the significant on-chip and off-

chip interconnect energy. Our approach also reduces the ratio of activated to accessed bits, but the energy-efficiency and bandwdith-density advantages of silicon photonics allows us to maintain a similar access size and latency.

Several researchers have proposed leveraging alternative technologies such as proximity communication [5] and 3D stacking to address the manycore memory bandwdith challenge [11, 13, 15]. Both technologies would offer either local DRAM physically packaged with the compute chip, or a tightly integrated multichip memory module connected to the compute via standard electrical interconnect. 3D stacking relies on through-silicon vias (TSVs) to communicate between layers, but monolithically integrated silicon photonics can also use the TSV holes for free-space optical communication. An optical scheme would have significantly higher bandwidth density than recently demonstrated stacked DRAM [11], while improving yield as metal is not required to fill the TSV. Even a more advanced TSV technology with a 10 µm pitch at 20 Gb/s per via, would offer 5-10$\times$lower bandwidth-density compared to an integrated optical vertical coupler. Furthermore, 3D stacking does little to mitigate the horizontal communication required to connect any processing core to any memory bank. More importantly, tightly integrating compute and DRAM chips into a single package eliminates the ability to easily combine commodity processing and memory chips into different configurations suitable for different system workloads. Instead we must leverage 3D stacked memory modules connected to the compute chip through energy-inefficient and comparatively low-bandwidth electrical interconnect. Our proposal provides an energy-efficient memory channel which connects memory controllers deep into the DRAM chip.

Previous studies have illustrated the advantages of using an optical channel between on-chip memory controllers and a *buffer chip* positioned near a rank of DRAM chips. These schemes used either shared buses with arbitration at the memory controller [7] or point-to-point links with arbitration at the buffer chip [4]. Our work examines the bank-level, chip-level, and system-level implications of fully integrating photonics into the actual DRAM chip, and our analysis shows the importance of considering all of these aspects to realize significant energy-efficiency gains. The Corona system briefly mentions a photonic memory-controller to buffer-chip channel, but then proposes using 3D stacked DRAM to mitigate the buffer chip to DRAM energy [17]. Although this mitigates some of the disadvantages of 3D stacking mentioned earlier, the Corona scheme relies on daisy-chained memory modules to increase capacity. We have found that this system-level organization places stringent constraints on the device optical loss parameters, especially waveguide and coupler loss. In this work, we have proposed optical power guiding as a new way to increase

capacity with less aggressive devices. The Corona work assumed a single DRAM organization, but our studies have shown that the advantages of photonics vary widely depending on the bank-level, chip-level, and system-level configurations.

9. Conclusion

Improving the energy-efficiency and bandwidth-density of main memory will be the primary way to achieve higher bandwidth in future manycore systems. In this paper, we have explored how monolithic silicon photonics can impact DRAM design at the bank-level, chip-level, and system-level. Our results show that photonics is a clear win for chip-to-chip communication, but that leveraging this new technology on the DRAM chip itself requires a careful balance between static and dynamic power. In addition, to achieve the full benefit of photonics, one must consider alternative array core trade-offs, such as increasing the number of array core I/O lines. At system-level, we have illustrated a new power guiding technique which allows commodity photonic DRAM parts to be used in both low and high capacity systems without requiring unnecessary bandwidth. Our work helps highlight static power as one of the key directions for future photonic device and circuits work. As always lower-loss components improve laser power requirements, but fine-grain control of laser power would also enable much better static power at low to zero utilization.

References

[1] Micron DDR SDRAM products. *http://www.micron.com/products/dram/ddr3*.

[2] J. Ahn et al. Multicore DIMM: An energy-efficient memory module with independently controlled DRAMs. *IEEE Computer Architecture Letters*, 8(1):5–8, Jan/Jun 2009.

[3] T. Barwicz et al. Silicon photonics for compact, energy-efficient interconnects. *Journal of Optical Networking*, 6(1):63–73, 2007.

[4] C. Batten et al. Building manycore processor-to-DRAM networks with monolithic CMOS silicon photonics. *IEEE Micro*, 29(4):8–21, Jul/Aug 2009.

[5] R. Drost et al. Challenges in building a flat-bandwidth memory hierarchy for a large scale computer with proximity communication. *Int'l Symp. on High-Performance Interconnects*, Aug 2005.

[6] C. Gunn. CMOS photonics for high-speed interconnects. *IEEE Micro*, 26(2):58–66, Mar./Apr. 2006.

[7] A. Hadke et al. OCDIMM: Scaling the DRAM memory wall using WDM based optical interconnects. *Int'l Symp. on High-Performance Interconnects*, Aug 2008.

[8] C. Holzwarth et al. Localized substrate removal technique enabling strong-confinement microphotonics in bulk Si CMOS processes. *Conf. on Lasers and Electro-Optics*, May 2008.

[9] I. Jung et al. Performance boosting of peripheral transistor for high density 4 gb DRAM technologies by SiGe selective epitaxial growth technique. *Int'l SiGe Technology and Device Mtg.*, 2006.

[10] U. Kang, H.-J. Chung, S. Heo, S.-H. Ahn, H. Lee, S.-H. Cha, J. Ahn, D. Kwon, J. H. Kim, J.-W. Lee, H.-S. Joo, W.-S. Kim, H.-K. Kim, E.-M. Lee, S.-R. Kim, K.-H. Ma, D.-H. Jang, N.-S. Kim, M.-S. Choi, S.-J. Oh, J.-B. Lee, T.-K. Jung, J.-H. Yoo, and C. Kim. 8gb 3d ddr3 dram using through-silicon-via technology. In *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pages 130–131,131a, Feb. 2009.

[11] U. Kang et al. 8 Gb 3D DDR3 DRAM using through-silicon-via technology. *Int'l Solid-State Circuits Conf.*, 2009.

[12] H. Lee et al. A 16 Gb/s/link, 64 GB/s bidirectional asymmetric memory interface. *IEEE Journal of Solid-State Circuits*, 44(4):1235–1247, Apr 2009.

[13] G. Loh. 3D-stacked memory architectures for multi-core processors. *Int'l Symp. on Computer Architecture*, Jun 2008.

[14] J. Orcutt et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk CMOS process. *Conf. on Lasers and Electro-Optics*, May 2008.

[15] H. Sun et al. 3D DRAM design and applicaiton to 3D multicore systems. *IEEE Design and Test of Computers*, 26(5):36–47, Sep/Oct 2009.

[16] S. Thoziyoor, J. Ahn, M. Monchiero, J. Brockman, and N. Jouppi. A comprehensive memory modeling tool and its application to the design and analysis of future memory hierarchies. In *Int'l Symp. on Computer Architecture*, June 2008.

[17] D. Vantrease et al. Corona: System implications of emerging nanophotonic technology. *Int'l Conf. on Computer Architecture*, Jun 2008.

[18] F. Ware and C. Hampel. Improving power and data efficiency with threaded memory modules. *Int'l Conf. on Computer Design*, Oct 2007.