Exploring High-Performance and Energy Proportional Interface for Phase Change Memory Systems

Zhongqi Li†, Ruijin Zhou† and Tao Li‡

Intelligent Design of Efficient Architectures Laboratory (IDEAL)
Department of Electrical and Computer Engineering, University of Florida
†{zhongqili, zhourj}@ufl.edu, ‡taoli@ece.ufl.edu

Abstract

Phase change memory is emerging as a promising candidate for building future energy efficient memory systems. To achieve high-performance and energy proportional design, phase change memory devices need to be reorganized so that (1) the relatively long latency of phase change memory devices should be hidden; (2) unnecessary power waste of phase change memory need to be preserved. Previous studies show that conventional memory ranks could be broken down into multiple smaller ranks for increased concurrency and lower power consumption. Nevertheless, the conventional electrical bus is incapable of supporting a large number of memory chips due to its insufficient load capacity and signal traversing speed. In this paper, we propose a phase change memory system design that leverages the state-of-art photonic links to overcome this issue. Moreover, thanks to the flexibility of photonic links, it is possible to amortize the small-rank penalty (e.g. the rank-to-rank switch overhead) by partitioning the channels either statically or dynamically. Our experimental results show that photonically interconnected phase change memory can increase the system performance (IPC) by up to 19% while saving 35% memory system power.

1. Introduction

Current computer systems pose challenges on memory power conservation, especially on the power reduction of main memory. For memory intensive applications, main memory is one of the major power consumers [1]. Recently, several non-volatile memory technologies (e.g. Phase Change Memory or PCM) have emerged as promising alternatives to the conventional DRAMs by using low leakage cells [2, 3, 4].

However, at current technology nodes, PCM faces relatively long latency and high energy writes issues [2]. Our work aims to address these issues without modifying the internal organization of PCM chips. To this end, we borrow one of the prior designs [5], which breaks the conventional ranks into multiple smaller ranks, namely “mini-rank”. The multiple small ranks increase the concurrency and thus hide the long latency of PCMs. Moreover, it preserves power consumption by involving fewer memory chips and narrower row buffers (i.e. sense amplifiers) [2] in each memory access.

Nevertheless, the mini-rank scheme is not naturally supported by the electrical interconnect due to the potential InterSymbol Interference (ISI) problem caused by the load of multiple ranks per channel [6]. As a result, the original mini-rank design adopts a dedicated interface chip between memory chips and the electrical communication bus at the cost of power waste, narrow bus connections, and additional data transmission cycles. The dedicated interface chip exacerbates power consumption and limits bus bandwidth in a system with a larger number of chips. Other solutions to the ISI problem, such as fully-buffered DIMM [7] and “fly-by” topology [8], also introduce intolerant transmission latency in mini-rank style design. The ISI issue will become even more crucial in PCM system design since more chips are necessitated to hide the long PCM access latency [9, 10].

To leverage the merits of mini-rank scheme while minimizing its limitations in PCM systems, we propose OptiPCM, which takes advantage of recent advances in CMOS-compatible nano-scale silicon-photonic integrated circuitry [11, 12]. The photonic channels connect the PCMs and are built on monolithically integrated silicon-photonic waveguides. Compared to conventional electrical bus, the silicon photonic bus is able to drive a large number of memory devices and amortize the ISI effect even under very high frequency given enough injection power. Unlike [5] which supports only 8 mini-ranks, the photonic channels within OptiPCM naturally support 16 or even more ranks. Increasing the number of ranks further lower power consumption owing to finer-grain power management. Meanwhile, more ranks provide higher concurrency to hide the long latency. The high bandwidth and low loss photonic channels also help improve the performance and power on bus.

OptiPCM only requires the modification of the memory interfaces, as opposed to the internal organization of PCMs. To our knowledge, our paper is the first work that uses an emerging (photonic) technology to address the traditional PCM limitations. Note that though our technique can also be applied to DRAMs, we mainly discuss its application for PCMs due to PCMs’ heavy demand for mitigating the long latency and intensive write energy.

This paper makes the following contributions:

1) We propose to apply photonic links to connect large number of independent PCM chips. Our design leverages the high load capacity of optical channels to improve memory performance and save power.

2) We apply the channel division technology to amortize the rank-to-rank turnaround penalty, which is proven to be critical for OptiPCM design with wide bus and numerous ranks.
(3) We introduce a dynamic channel division scheme to address low channel utilization in the static channel division design. Dividing the channels based on traffic characteristics utilizes communication bandwidth more efficiently while amortizing the rank-to-rank turnaround penalty on heavy traffic scenarios.

The rest of this paper is organized as follows: Section 2 provides background on PCM and photonic communication. Section 3 presents our OptiPCM architecture and design. Section 4 proposes OptiPCM performance optimization by using static and dynamic channel division technologies. Section 5 describes the experimental setup and Section 6 presents our evaluation results. Section 7 discusses related work and Section 8 concludes this paper.

2. Background

2.1 Phase-Change Random Access Memory

Currently, PCM is being considered as one of the most promising technologies for next generation non-volatile memory. The emerging PCM technology has many advantages such as random access, non-volatility, superior scalability, fast read cycle, and manufacturing compatibility with existing CMOS processes. Each PCM cell employs reversible phase change materials to store information. These materials are usually made using a chalcogenide alloy of germanium, antimony and tellurium (GeSbTe) called GST. Figure 1 shows the basic structure of a PCM cell, which consists of a standard NMOS transistor and a phase change device. PCM leverages the differences between the two states in the electrical resistivity of GST to store information. The GST phase can be changed by heating a region of phase change material to a high temperature threshold using electrical-pulse generated Joule heat, which incurs relatively long latency and high write energy [13].

2.2 The Memory Devices Organization and LPDDR2-NVM Protocol

Figure 2 illustrates the contemporary PCM chip organization. The organization of PCM is similar to that of DRAM. In this example, eight PCM chips are ganged together and work in lockstep fashion to respond to the commands issued by the memory controller. The ganged chips form a rank, and the links that the ranks and memory controller use to communicate are called channels.

In our study, we employ Low-Power Double Data Rate Non-Volatile Memory (LPDDR2-NVM) compatible channel [6, 14] to support PCM. The original LPDDR2 is proposed to become the technology of choice for embedded and mobile applications thanks to its low-power characteristics. Recently, [15] leverages LPDDR2 based DRAM in data centers. It is able to save a significant proportion of power (over 50%) under similar device density and performance conditions, which mainly comes from the reduced working voltage (1.2V) and the shrunk pin count. LPDDR2-NVM officially provides support for non-volatile memory such as flash memory and PCM. In LPDDR2-NVM, the signals between memory device and memory controller fall into four categories: command, address, data, and miscellaneous signals. The command and address signals are unidirectional and extend from memory controller to the memory. They provide commands and bank/row/column addresses to the memory chips. The data bus is a bidirectional bus whose bandwidth is the aggregation of each memory device. For example, Figure 2 shows an example of eight 8-bit chips forming a 64-bit data bus. The data bus also contains data strobe signals (DQS) used for data alignment. Apart from the four types of signals, the LPDDR2-NVM protocol also contains miscellaneous signals such as clock signal and temperature sensor signals.

Each PCM chip is organized into multiple banks. A bank is an independently controllable unit and is composed of several PCM arrays. A PCM array consists of cells, which are organized into 2D arrays and accessible through a row address and a column address. In contemporary designs, each read/write access to one rank will activate one row specified by the row address across all PCM arrays in one bank and loads data from the cells in that row to row buffers. The column address specifies one bit from the row buffer of each PCM array, and multiple PCM arrays provide multiple bits. Contemporary memories used in CMP systems usually provide the data from several columns in a burst to meet the size of a single cache line.

2.3. The Principle of Photonic Communication

Silicon nano-photonics have made complete photonic on-stack communication systems a promising alternative to
electrical communication systems. Integrated silicon-photonics technology could be used as an ideal candidate for the connection of the large number of memory chips due to its low-latency and high scalability. Nano-photonics significantly improve the interconnect bandwidth density by approximately two orders of magnitude and yield to over $10^x$ power reduction [11]. Figure 3 shows the basic optical communication components including laser source(s), optical waveguides, modulators, and photo-detectors. The vertical-cavity surface-emitting laser (VCSEL) source(s) multiplex a number of different wavelengths of laser lights into a single waveguide, using dense wavelength-division multiplexing (DWDM). Modulators then modulate laser lights to carry the optical bits using ring-coupled Mach-Zehnder interferometer [16]. The SiGe photo-detectors couple and absorb laser lights at their resonant wavelengths and then convert them into electrical current to be amplified for the final electrical bits.

![Figure 3. A typical optical communication system](image)

Apart from the basic photonic components, the turn resonators which are properly tuned couple the traversing optical signals and drop them to the intersecting waveguides. A turn resonator works in a set of resonant frequencies derived from its material and structural properties. When the resonant frequency of the turn resonator is different from the traversing wavelength(s) of the optical light(s), the light(s) pass through the waveguide intersection uninterrupted (the red light in Figure 3); otherwise they are coupled into the resonator and dropped to the intersecting waveguide (the green light). The material and structural properties of the passive resonators are predetermined when manufacturing and kept constant during run-time, as (e) in Figure 3. Frequency of the active resonator could be tuned during run-time to support different waveguide connections. The frequency tuning is achieved by adjusting the effective refractive index of the resonator and is generally achieved in one of the following ways. Heat tuning applies or cancels the heat on resonator to change the effective index, which usually requires several microseconds. The electrical tuning applies a voltage on the p-n contact and injects electrical current into the resonator to tune the effective refractive index of the waveguide as (c), which requires ~100 ps. Another way is to apply optical pump pulses to inject free carriers through two-photon absorption inside the ring resonator and hence tune the effective index of the ring resonator as (d) [17]. The optical pulse tuning has the lowest latency among the three (~40 ps) [18], and is suitable to control distant resonators which otherwise suffer extra delay from electrical control wire. In our design we apply electrical tuning at the memory controller and optical tuning at the memory device.

### 3. OptiPCM System Organization

As shown in Figure 4, OptiPCM employs a set of non-volatile memory chips and replaces the conventional electrical channels with optical channels. In contemporary design, multiple PCM chips in a rank work in a tandem and share one DIMM (Dual In-line Memory Modules), timing synchronization circuit, and power supply circuit. OptiPCM breaks the PCM chips from one rank into multiple mini-ranks, each mini-rank is a single PCM chip with its own power supply and timing synchronization circuits. Unlike the original design, which breaks each rank into eight mini-ranks, OptiPCM could break each rank into 16 or even more memory ranks thanks to the low signal loss in the photonic channels. The x16 PCM chips [19] are adopted in OptiPCM to assure access speed. The large number of ranks can hide long PCM access and internal latency. Besides, the power consumption will be further lowered from electrical mini-rank due to finer-grained power management and the removal of interface chip.

The communication channels in OptiPCM consist of two bundles of unidirectional silicon photonic waveguides (shown as purple and blue). These waveguides carry DWDM laser signals and provide highly aggregated pin-bandwidth density, which improves bandwidth density by two orders of magnitude than that of electrical buses. Among them, the outbound channel consists of 64-byte wide data, address, and command buses from the memory controller to the memory while the return data travels through inbound channel. The memory controller modulates the laser lights in the outbound channel by properly tuning the modulators array (“M” in Figure 4).

OptiPCM adopts connected outbound/inbound channel where the laser lights are only injected into the outbound channel. Since one memory chip can only either be read or written at any instant, the connected channel works in half-duplex where the data can only be transmitted in one direction, which is distinct from prior optical memory design with separate laser supplies for outbound and inbound channels [9]. The connected outbound/inbound channel
shares one laser source, hence nearly halves the static laser power from conventional separate outbound/inbound channels.

To support photonic channel, the conventional electrical DIMM is redesigned as a CMOS-compatible integrated photonic interface (PI) in place of the conventional electrical pins. PI converts optical signals from outbound channel to electrical signals for write commands. On read commands, PI recaptures laser lights from outbound channel to inbound channel and then modulates the laser lights to carry the return data. The memory controller generates and distributes optical clock through outbound channel to each memory rank in order to avoid the need for centralized signal retiming unit and global timing synchronization among memory ranks. The clock wavelength parallels the data wavelengths, with the clock signal traveling with the data signals.

The increased memory ranks require the laser source to inject enough laser power for the compensation of waveguide propagation loss. The injection power is constrained at the first modulator, which must be below the threshold (around 10-20 mW [20]) that induces nonlinear effect.

Note that state-of-the-art DDRx protocol does not support non-volatile memory interface. With lower performance, protocols such as LPDDR2-NVM [6, 14] provide PCM interface. The photonic channels in OptiPCM follow LPDDR2-NVM based protocol and recuperate the performance loss.

4. Sub-channel Division Technology

OptiPCM surpasses the original electrical LPDDR2-NVM based PCM systems in terms of performance due to high memory concurrency and wide channel. Nevertheless, the baseline design could be further optimized to reduce the rank-to-rank switch penalty, which is non-trivial when the number of ranks is large. In this section, we propose the technology by dividing the channel bandwidth into several sub-channels to overcome the rank-to-rank turnaround penalty.

4.1 Channel Division

Rank-to-rank turnaround penalty exists in the high-frequency global-synchronous memory system. Unlike the commands to same open memory bank that can be issued and pipelined back to back, consecutive commands to different memory ranks relies on the system-level synchronization operations [8]. Thus, the shared data bus must be idle for some period of time between data bursts from different ranks. In the electrical design, the synchronization operation requires synchronization circuits to align the strobe signal DQSs to DQs on data bus. In LPDDR2 the synchronization circuits incur several cycles of latency even though the DLL circuits in the DDRx protocol is removed in LPDDR2 protocol. Synchronization latency dominates the rank-to-rank synchronization penalty, which depends on system-level synchronization mechanisms and is usually 1 to 3 cycles [14]. The photonic channel suffers from unpredictable photo-detector/modulator delay for optical clock recovery as well, which further exacerbates this issue.

Figure 5(a) shows an example of such rank-to-rank turnaround penalty incurred by consecutive writes to different memory ranks on the electrical bus. In this example, the data burst (tBL) and penalized synchronization operation

![Figure 5 (a). In LPDDR2-NVM, memory access timing without channel division for consecutive write to different ranks (PA_n = Pre-activate rank n, ACT_n = Activate rank n, WR_n = Write rank n, RAH_n = Raw Address High bits, RAL_n = Raw Address Low bits. [tRP = 5, tRCD = 4, tWL = 1, tBL = 4, tRTRS = 3, unit: cycle])](image)

![Figure 5 (b). Memory access timing with channel division technique for consecutive write to different ranks (part of the PA and ACT commands for rank 5-8 are not shown) [tRP = 5, tRCD = 4, tWL = 1, tBL = 16, tRTRS = 3, unit: cycle]](image)
(tRTRS) interleaves on the data bus. In contemporary CMP systems, typical tBL length is 4 cycles (64-bytes cache line), and tRTRS rises with increasing frequency of bus clock (e.g. approximately 3 cycles in DDR3 [8]) and therefore wastes considerable bus bandwidth. In OptiPCM, the tBL is reduced to 1 cycle for 64-bytes cache line due to the 64-bytes wide photonic data bus and the number of ranks is significantly increased, which could incur more significant bandwidth waste.

To address this issue, we propose channel division technique by dividing the data channel into several sub-channels. In the example shown in Figure 5(b), the narrow sub-channel extends the original tBL from 4 to 16, but rank-to-rank switch penalizes only one sub-channel and thus quarters the overall rank-to-rank switch overhead.

We first propose a static sub-channel division. In this method, the data channel is equally divided into several sub-channels; each sub-channel is dedicated to one rank. The memory controller allocates one sub-channel rather than the whole data channel to one memory access. Although the width of each sub-channel decreases, the sub-channel division eliminates the rank-to-rank synchronization penalty since each sub-channel is dedicated for one rank.

4.2. Dynamic Channel Division

Although static sub-channel division eliminates rank-to-rank turnaround penalty, it may fail to utilize the provisioned channel bandwidth sufficiently especially under non-memory intensive applications. To address this issue, we further propose dynamic channel division, which improves static channel division by adjusting the sub-channel width (i.e. number of wavelengths) based on the incoming traffic. In dynamical channel division, if only one command is waiting in the queue, the memory controller allocates the whole channel to it. If two or more commands in the queues compete for the channel, the memory controller seeks to equally divide the available channel width among these commands. However, if the equal division is impossible, one or more commands(s) will be delayed. In non-memory intensive applications the dynamic channel division maintains a low latency data channel; while the rank-to-rank switch penalty is amortized in the memory-intensive applications. Dynamic channel division requires extra design module for memory device and the memory controller as we will discuss in Section 4.4.

4.3. The Structure of PIs

In OptiPCM, the PI is deployed in place of the conventional electrical DIMM as shown in Figure 6(a). The PI directs photonic signals to appropriate ranks and converts between electrical/optical signals. There are two types of optical bits: data + RW bits and connectivity bits. In the first group, one RW bit in each sub-channel indicates the current direction of one sub-channel and travels along with the data bits. All RW bits from outbound waveguides are directed by R1 to the vertical waveguide first, and then separated by R2 resonators as the pulse control signals of R3 by injecting or canceling the free carriers [21]. The R3 resonators then direct the signals to either the ranks (or optical crossbar in dynamic channel division) or the inbound waveguides based on their status. The optical signals are directed to the ranks and the built-in photo-detectors convert the optical signals to electrical signals.

The pattern of channel division is determined by the connectivity bits. The PI simply directs the signals in one sub-channel to the corresponding rank using passive turn resonators (these resonators are not shown in Figure 6). An optical crossbar under the control of the connectivity bits (the red dashed-dotted box in Figure 6(a)) switches the data between the sub-channels and the ranks. A n × w optical crossbar is required to switch the data between n ranks and w wavelengths wide data bus. The organization of the crossbar is illustrated in Figure 6(b). The optical crossbar (shown in the red dashed-dotted box in Figure 6(b)) contains n × w basic units. A basic unit contains a passive resonator used to extract one connectivity bit and an optical tuning resonator that directs the data from one sub-channel to a rank. A 64-bytes wide data bus with 16 sub-channels, for example, requires $512 \times 16 = 8192$ such components. The PI requires no crossbar-like units for the read operation. As Figure 6(a) shows, the modulator arrays from different ranks modulate data to the same waveguide at different wavelengths to avoid interference.

The PI incurs extra photonic components and therefore results in additional power consumption. The reason is that the extra resonators and increased length of waveguides attenuate the traversing laser lights. However, the maximal power that the modulator could support must be below the threshold at which nonlinear effect is induced, which is typically 10-20 mW [20]. In OptiPCM, the optical power experiences up to 1 filter drop loss, 4095 filter through loss, 1 modulator insertion loss, and 1 photo-detector loss when the system employs dynamic wavelength assignment to 64-byte wide data bus with 16 sub-channels. And the photo-detectors require 5 μW sensing power to detect optical
signals [22]. As we will elaborate in Section 5.2, the optical loss on the PI is around 10.1 dB. The propagation loss of the waveguides under current technology node is 0.3 dB/cm [23]. Assuming 10 mW modulation power, the remaining power could support up to ~76 cm waveguides, which is sufficient to route waveguides from the processor to memories.

We also evaluated the latency on the waveguides and PI. The calculation results from optical latency models [24, 25, 26, 27] indicate that the PI incurs less than 100 ps latency, which is negligible under 533MHz clock (the highest frequency supported by JEDEC LPDDR2).

From the memory side, OptiPCM uses the native internal row buffers within PCM chips to convert data between different widths of sub-channel and memory arrays. Data from the sub-channel fill the row buffer in one or more cycles and is written into the memory cells in one batch. Data read out from memory cells is also temporarily stored in the buffers and then consecutively sent to the sub-channels. Therefore, another overhead of the dynamic channel division mainly comes from the extra internal width conversion circuitry. We synthesize the conversion circuitry using Synopsys Design Compiler [28] and obtain the numbers of the power consumption under different traffic.

### 4.4. The Memory Controller for Dynamic Channel Division

In order to support dynamic channel division, OptiPCM adopts an enhanced First-Ready First-Come-First-Serve (FR-FCFS) memory controller [29]. In traditional FR-FCFS memory controller, the transaction arbiter accepts requests (i.e. transactions) from the processors and arbitrates for them. Once a transaction wins arbitration and enters into the memory controller, it is decomposed into a sequence of memory commands and mapped to a command queue. The command queues are arranged in such a way that there is one queue per bank. Then, commands are scheduled to the memory devices through the optical signaling interface depending on the command scheduling policy. The memory controller in OptiPCM uses the same transaction scheduling and address mapping mechanism like conventional FR-FCFS memory controller, while introduces data modulators and enhances the command scheduler to support dynamic channel division, as shown in Figure 7 (a).

The basic idea of the dynamic channel division is to equally divide the data channel among all issuable read/write memory commands. Doing so requires an enhanced command scheduler, as shown in Figure 7(b). In the enhanced command scheduler, each of the bank queues has a pre-charge manager and row arbiter. They cooperate with address arbiter to decide when the associated bank should be pre-charged or activated.

The workflow of enhanced command scheduler is illustrated in Figure 7(c). The command scheduler checks the commands in the command queues for each bank in a round-robin fashion and allocates them. For the pre-active and active commands, since the command channel is not divided as the data channel, OptiPCM can only assign one command per cycle. For the read/write commands, once the scheduler finds more than one issuable read/write commands from the command queues, it stores them in a pool and seeks to equally divide the available channel width among them provided that two or more commands are in the pool. The pool, which incurs extra overhead for the dynamic channel division, could be efficiently implemented to only store the addresses of the memory commands in the command queues. The channels are then divided into sub-channels in the unit of byte, and each read/write will occupy one sub-channel.

### 5. Experimental Setup

#### 5.1. Simulation Methodology

We evaluate the power consumption and performance improvement of OptiPCM using Simics [30], a multi-
Table 1. Simulation benchmarks

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Workloads</th>
<th>Traffic (Mbps)</th>
<th>Channel Width (Bytes)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC-1</td>
<td>zip * 2, gcc * 2, sjeng * 2, ibm * 2</td>
<td>399.9 1027.7</td>
<td>Electrical 8</td>
<td>The conventional electrical LPDDRx-compatible channel with eight chips in one rank (baseline case)</td>
</tr>
<tr>
<td>SPEC-2</td>
<td>hsp, zip * 2, gcc * 2, sjeng * 2, ibm</td>
<td>772.1 1182.3</td>
<td>Electrical 8</td>
<td>The Mini-rank configuration electrical LPDDRx channel with small ranks as in [5]</td>
</tr>
<tr>
<td>SPEC-3</td>
<td>hsp, GemsFDTD, hsp * 2, gcc * 2, sjeng, ibm</td>
<td>822.0 1181.0</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM design (Section 3)</td>
</tr>
<tr>
<td>SPEC-4</td>
<td>hsp, GemsFDTD, hsp, zip * 2, gcc, sjeng, ibm</td>
<td>1049 1967.3</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM design with static (fixed) channel division (Section 4.1)</td>
</tr>
<tr>
<td>SPEC-5</td>
<td>hsp, GemsFDTD, mcf, cachusADM, hsp, gcc, sjeng, ibm</td>
<td>963.2 1842.0</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with dynamic channel division (Section 4.2)</td>
</tr>
<tr>
<td>SPEC-6</td>
<td>hsp, GemsFDTD, mcf, cachusADM * 2, hsp, gcc</td>
<td>798.8 2039.8</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
<tr>
<td>SPEC-7</td>
<td>hsp, GemsFDTD, mcf, cachusADM * 2, hsp, gcc</td>
<td>608.6 2590.8</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
<tr>
<td>SPEC-8</td>
<td>hsp, GemsFDTD, mcf, cachusADM * 2, hsp</td>
<td>595.9 2556.5</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
<tr>
<td>SPEC-9</td>
<td>hsp * 2, GemsFDTD, mcf, cachusADM * 2</td>
<td>654.2 2831.7</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
<tr>
<td>blackscholes</td>
<td>Financial Analysis, 65,536 options</td>
<td>2.1166 5.1035</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
<tr>
<td>swaptions</td>
<td>Financial Analysis, 16 swaptions, 20,000 simulations</td>
<td>8.7270 2.0904</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
<tr>
<td>freqmine</td>
<td>Data Mining, 990,000 transactions</td>
<td>24.978 16.141</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
<tr>
<td>d264</td>
<td>Media Processing, 128 frames, 640x360 pixels</td>
<td>14.088 13.840</td>
<td>Electrical 8</td>
<td>Prototype OptiPCM with narrow channel (Section 4.1)</td>
</tr>
</tbody>
</table>

5.2. Power Model of the Communication Bus

Electrical Links: The electrical link between memory and memory controller is modeled as [34]. The transmission line between the PCM and the processor could be characterized using a simple RC model. LPDDR2-NVM also uses LVCMOS logic level rather than SSTL in DDRx. Figure 8 shows the LPDDR2-NVM communication bus model. The power consumed on the DQ bus could be calculated as: 

\[ P_{(per \ DQ)} = C_L \times V_{DDQ}^2 \times (2 \times f \times K) \]

where DQ data rate frequency is twice the system clock frequency [35]. For the differential transmission line, the voltage supplied to the DQ bus \( V_{EEF} \) is half of \( V_{DDQ} \).

Optical Links: The power consumed on the optical channels is an aggregation of static and dynamic power. The static power consists of the required detection power for each resonator, the tuning power of the resonators when they are

<table>
<thead>
<tr>
<th>Optical loss in various components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical coupler</td>
</tr>
<tr>
<td>Interlayer coupling loss</td>
</tr>
<tr>
<td>Filter drop</td>
</tr>
<tr>
<td>Waveguide loss</td>
</tr>
<tr>
<td>Non-linear loss</td>
</tr>
<tr>
<td>Waveguide crossing</td>
</tr>
</tbody>
</table>

Figure 8. The power modeling of the LPDDR2-NVM (the equivalent driver impedance \( R_{ON} \) is equally devided into two parts: \( R_{OPU} \) and \( R_{ODE} \). The value is typically chosen to be 120Ω when measuring at 0.5 \( \times V_{DDQ} \))
tuned to be ON state (assume OFF state has zero power consumption), the traversing optical power loss, and the power consumed by the heater of the resonator. The dynamic power is consumed by the modulators and photo-detectors when modulating and detecting optical data. The important factor that affects the total static power consumption is the required optical detection power for a single photo-detector. Prior study [5] shows that the power consumption of the photo-detectors is related to the expected Bit Error Rate (BER). We adopted expected BER of $10^{-15}$ [17, 26] to ensure reliable end-to-end communications, which required 5 $\mu$W sensing power per photo-detector [22]. The power consumption for different photonic components [36, 37] is summarized in Table 4. By default, all the ring resonators are set to OFF state. The energy is required when they are tuned to ON state [38] and this in-plane Poly-Si energy per resonator is assumed to be 0.5 mW [39]. Assuming advanced driver circuits with poly-Si carrier lifetimes of 0.1-1 ns, the power consumed by each modulator is approximately 200 $fJ/bits$ [40]. The energy coupling efficiency of the laser source into the chips ranges from 30% [37] to 50% [36, 41]. We use the median value of 40% in our power model.

**PCM Devices**: The main difference between the DRAM and PCM is the structure of the cells. DRAM employs the 1T1C cell while PCM employs the 1T1R cell (shown in Figure 1). In non-volatile memory, the pre-active commands that load the row address buffer replace the pre-charge command in the volatile memory.

PCM has two major power consumers: peripheral circuits and cells. We adopt the power consumption profile from CACTI 6.5 [42] for the peripheral circuits, and the power data of the PCM model extended from [9] in our study. The timing parameters of PCMs are shown in Table 5 [2]. PCM works in different states as defined by LPDDR2-NVM standard [14], which is shown in Figure 9. OptiPCM manages to alleviate the intensive write energy demand for the PCM chips thanks to the smaller chips; besides, existing PCM power management techniques [43, 44] are applicable in OptiPCM as well. Also note that current density does not constrain the performance of OptiPCM [43]. This is because wider power supply pins can now be adopted for PCM chips thanks to the elimination of most electrical I/O pins.

In reading and writing states, the memory returns the latched data from the row buffers in response to the column access command. We obtain the sense amplifier power and the row buffer power data from CACTI. The pre-active operation is analogous to pre-charging in DRAMs, which resets the row buffer to the idle state once the minimum pre-active latency (tRP) is satisfied. In the pre-active operation, data stored in the row buffer is written back to the memory cells using partial write where only the modified bits will be written [2]. In the active state, a row of data from the PCM cells are sensed, amplified, and then latched into the sense amplifiers. All but power-down states consume leakage power. Idle PCM devices in power-down state could save leakage power, however this incurs longer exit latency going back to the idle or active state to serve the incoming requests. By leveraging mini-rank design, it is possible to fine-tune the state of each rank without impeding data operation in other ranks. In power-down state, the memory device is still supplied with power; however most of the peripheral circuits like input/output buffers are deactivated [45]. The LPDDR2-NVM protocol supports power-down state with stopped clock. The only overhead of resuming the clock is a NOP command before the next access command could be applied [14]. Deep power-down state eliminates power to both the peripheral circuitry and memory array and will be supported in future LPDDRx protocol. In this study we choose to use the power-down state with stopped clock. Applying this state effectively reduces the power consumption of the PCM device, while incurring limited entering and exiting overhead. We use CACTI to estimate leakage power consumed by the peripheral circuits. PCM belongs to the class of non-volatile memory; hence almost no leakage power is consumed by the cells [9]. The data from Micron P8P PCM datasheet [19] shows that less than 100 uA current per memory chip will be consumed in the low power state.
So we assume that the power consumption in power-down state is negligible, which is consistent with [4].

6. Performance Evaluation

6.1. Power Consumption Breakdown

The power consumption of OptiPCM comes from the PCM devices and the photonic channels. The power consumption of PCM devices could be categorized into three groups according to the abovementioned power analysis: background, operation, and read/write power. Background power is consumed all the time when the memory chip powers on except for the power-down state. The device consumes operation power when performing activation or pre-active operations. The read/write power is consumed when the device reads or writes data. Figure 10 shows that most of the overall power consumption is decreased when more ranks are deployed in OptiPCM. For example, the PRT-64, FCD-64 and DCD-64 reduce the overall power by 34.8%, 27.8% and 21.8% respectively compared with the baseline case. Note that we only show half of the benchmarks in Figure 10 due to page limitation, the other half exhibits the similar trend.

The overall power reduction could be broken into different groups. Memory cell R/W power is caused by write/read operations, which is proportional to the number of memory accesses and increases with improved memory utilization efficiency. Memory utilization efficiency increases with the number of ranks by hiding long PCM latency. For example, PRT-64, FCD-64 and DCD-64 increase power by 10.1%, 26.3%, and 40.3%. The memory operation power accounts for a large portion of the overall power consumption thanks to the low-power LPDDRx interface and low PCM leakage power. The operation power is reduced significantly by 84.8%, 83.5% and 82.1% in PRT-64, FCD-64 and DCD-64 modes due to the reduced width of activated sense amplifiers in smaller ranks. The negligible cell and peripheral circuitry leakage contributes to the power saving as well since the number of devices involved per access is reduced compared with BASE. Thus, the PRT-64, FCD-64 and DCD-64 consume 42.7%, 40.2%, and 36.7% less background power on average. Moreover, the photonic channel saves 44.1% power on average compared with electrical channels despite the fact that its power increases with number of ranks.

6.2. Small Rank Benefits

OptiPCM reduces memory latency, as shown in Figure 11. Overall, SPEC 2006 benchmark suite exhibits much higher latency than that of PARSEC owing to the significantly larger memory foot-print. Among the memory configurations, we observe that the conventional mini-rank design incurs an average of 7.8% performance degradation compared with the BASE, especially on the low-traffic PARSEC benchmarks (on average of 14.9%) due to the insufficient channel utilization. The photonic channels reduce memory latency (i.e. PRT-32) by 13.9% compared with the BASE. This is due to the increased number of ranks.
that hides the long PCM access latency, and wider photonic channels. The static channel division, e.g. FCD-32, improves the latency by 30.6% on average in SPEC and 3.6% on average in PARSEC benchmarks. The rank-to-rank turnaround penalty is significantly amortized in the high-traffic benchmarks (SPEC) owing to static channel division; however, the slight performance improvement on low-traffic benchmarks (PARSEC) is mainly due to the increased bus width.

DCD compensates for this performance degradation since it demonstrates similar low-traffic behavior while it amortizes the rank-to-rank turnaround penalty when the traffic is heavy. The dynamic channel division (DCD-32) gains an average performance improvement of 32.0% and 30.9% in SPEC and PARSEC compared with BASE. The performance improvement comes from three aspects: 1) more ranks provide better concurrency; 2) increased width of photonic communication channel; 3) dynamical division of channel width reduces rank switch penalty while preserving the low latency when traffic is low.

Apart from the latency reduction, applying more ranks increases memory concurrency and therefore hides the long latency of PCM cells. We evaluated the IPCs under different configurations. For the sake of fairness, the evaluations are performed under the same channel widths (64 bytes), as shown in Figure 12.

The IPC improvements are highly related to the characteristics of the benchmarks. For example, the IPC improves by 2.8% from PRT-8 to PRT-64 for SPEC benchmarks and 1.3% for PARSEC benchmarks on average compared with BASE. FCD-8, FCD-16, FCD-32, and FCD-64 improve the SPEC IPC by 9.5%, 10.2%, 13.2%, and 13.7%. However, on PARSEC benchmarks, the increased ranks hurt the performance under FCD due to the reduced sub-channel width. FCD-8, FCD-16, FCD-32, and FCD-64 only speed up the BASE by 3.8%, 2.4%, 1.9%, and 1.5% respectively.

For memory-intensive applications under DCD configuration, memory accesses are able to fully utilize the sub-channels and the system achieves better performance. Memory-intensive benchmarks exhibit more significant IPC improvement than non-memory intensive benchmarks. For example, the DCD-64 exhibits up to 19.4% improvement compared with BASE on average for all SPEC benchmarks, while 7.6% improvement on average for PARSEC benchmarks.

6.3. The Link Utilization Improvement

The channel utilization percentage for data transmission is also improved from 33.0% to 44.2% in SPEC and 2.1% to 2.9% in PARSEC as shown in Figure 13(a). However, Figures 13 (a) and (b) show that the percentage of rank-to-
The rank turnaround penalty in channel utilization increases by 33.4% in SPEC and 25.7% in PARSEC from PRT-8 to PRT-64. Our FCD and DCD schemes successfully reduce this penalty. For example, FCD-8, FCD-16, FCD-32, and FCD-64 reduce the percentage of rank-to-rank turnaround penalty by 27.7%, 47.6%, 60.9% and 70.8% in SPEC compared with their PRT counterparts. DCD-8, DCD-16, DCD-32, and DCD-64 reduce the penalty by 27.7%, 28.1%, 38.1%, and 37.5% on SPEC respectively. Although DCD reduces the penalty less than FCD, it does not sacrifice link utilization and thus achieves better overall performance.

6.4 The Impact of Channel Width on OptiPCM

Photonic waveguides could provide much wider data channel than traditional electrical bus. We evaluated memory latency under different data bus widths, from the traditional electrical bus (64-bits) to the single cycle cache line transfer (64-bytes) that the photonic links support. Figure 14 (a) shows that the wide bus contributes to the memory performance across all configurations. In PRT, FCD and DCD, the 64-byte bus outperforms the 8-byte bus by 18.1% and the 64-byte FCD improves 8-bytes FCD performance by 31.1% on average.

Figure 14(b) also shows that on the low traffic benchmarks (PARSEC), the channel width under FCD configurations affects the performance considerably. For example, the 64-bytes FCD shows 27.7% performance degradation compared with the 8-bytes FCD. This is largely because the very narrow channel constrains performance of the memory system. However, we also observe from Figure 14 (b) that increasing the ranks beyond 16 exhibits limited impact on the performance. For example, FCD-32 exhibits a 4.0% improvement over FCD-16 on average of all PARSEC benchmarks. In contrast, the memory system performance under DCD configurations is not as sensitive to the bus widths as the FCD configurations. For example, DCD-32 improves DCD-16 by only 0.9%, compared with the 6.3% and 9.1% improvement in the SPEC.

7. Related work

Prior works focus on leveraging photonic links to support the memory device in order to preserve power and achieve higher performance. For example, Beamer et al. [9] proposed to use photonically interconnected memory by employing a monolithically integrated silicon-photonic technology to replace electrical I/O and electrical links across the memory chips. Hadke et al. [46] leveraged the low latency optical links to replace the multi-hop store-and-forward network in the fully buffered DIMM [7] and successfully support up to 32 DIMMs. Hendry et al. [47] proposed photonic network-on-chip with integrated memory I/O interfaces to build a uniform photonic network-on-chip communication system using low-latency and low-power photonic links. Udipi et al. [48] proposed 3D-stacked memory architecture by adding an interface die to handle the conversion between optics and electronics.

The majority of prior works only leverage photonic channels to reduce the power consumption. The power saving is constrained on communication channels, and is very limited in LPDDR2-NVM environment where the channel consumes less than 20% of the memory system power [1]. The fundamental difference between our work and prior works is: our work takes advantage of the photonic channels to enable more PCM chips connected to the bus and therefore improves power saving and performance. The performance and power advantage come from the management of PCM chips, not merely the channel.

8. Conclusion

In this paper, we propose a high-performance and energy-efficient memory communication infrastructure for PCMs. Our OptiPCM design exploits photonic channels in place of conventional electrical channels to improve performance and power of PCMs. By leveraging the photonic channels, OptiPCM is capable of supporting large number of memory devices and managing the power consumption better. It also improves the memory concurrency and hides the long access latency of PCM. Our simulation results show that PRT, DCD and FCD design save up to 34.8%, 27.8%, and 21.8% of overall power consumption while increasing system performance (IPC) by 11.2%, 13.5%, and 19.4% respectively.
9. Acknowledgements

This work is supported in part by NSF grants 1017000, 0937869, 0845721(CAREER), and by Microsoft Research Safe and Scalable Multi-core Computing Award. Zhongqi Li is also supported by a University of Florida Graduate Fellowship.

10. References