LL-PCM: Low-Latency Phase Change Memory Architecture

Nam Sung Kim[#], Choungki Song[*], Woo Young Cho[#], Jian Huang[†], Myoungsoo Jung[‡]
Samsung Electronics[#], University of Wisconsin[*], University of Illinois[†], KAIST[‡]

ABSTRACT

PCM is a promising non-volatile memory technology, as it can offer a unique trade-off between density and latency compared with DRAM and flash memory. Albeit PCM is much faster than flash memory, it is still notably slower than DRAM, which can significantly degrade system performance. In this paper, we analyze a PCM implementation in depth, and identify the primary cause of PCM's long latency, *i.e.*, a long interconnect (high resistance/capacitance) path between a cell and a sense-amp/write-driver. This in turn requires (1) a very large charge pump consuming: ~20% of PCM chip space, ~50% of latency of write operations, and ~2× more power than a write operation itself; and (2) a large current sense-amp with long time to pre-charge the interconnect path. Then, we propose Low-Latency PCM (LL-PCM) architecture. Our analysis shows that LL-PCM can give 119% higher performance and consume 43% lower memory energy than PCM for memory-intensive applications. LL-PCM is only ~1% larger than PCM, as the cost of reducing the resistance/capacitance of the interconnect path is negated by its 4.1× smaller charge pump.

CCS CONCEPTS

• B.3.1 Semiconductor Memories

KEYWORDS

PCM, DRAM, Heterogeneous memory system

1 INTRODUCTION

PCM has been expected to be a few times slower than DRAM. To reduce the notable impact of PCM's long latency on overall system performance, DRAM and PCM can constitute a main-memory system where PCM devices are expected to be drop-in replacement of DRAM devices in Dual-Inline Memory Modules (DIMMs). In such a hybrid main-memory system, fast DRAM can be used as hardware- or software-managed cache (*e.g.*, [1, 2]). DRAM used as hardware-managed cache is software-transparent, but it requires memory space for storing tags and is slower than DRAM used as software-managed cache because of the latency penalty of accessing/comparing tags (*e.g.*, [3]). In contrast, DRAM used as software-managed cache neither requires to store tags nor incurs the latency penalty of accessing/comparing tags. However, the OS or applications need to explicitly handle page placement and migration, which are very challenging [2]. When these challenges are considered, it is important to significantly reduce the latency of PCM through various circuit- and architecture-level techniques without sacrificing density.

In this paper, we first analyze a PCM implementation [4] and observe that I/O circuits of PCM are shared by far more memory cells than those of DRAM to maximize the density. Such an implementation requires a long interconnect path (*i.e.*, high resistance/capacitance) between a cell and a write-driver/sense-amp. The high resistance demands PCM write-drivers to drive a few times higher voltage than the chip nominal supply voltage to provide a necessary amount of current for changing cell states between RESET and SET. This high-voltage requirement in turn demands a large charge pump which consumes ~20% of the PCM chip size, far longer time to pump the voltage to a desired level, and much more current from the chip voltage source than a small charge pump for low-voltage requirement. Note that the long time to pump the voltage contributes to ~50% of latency of write operations, and the high current dissipation limits concurrent write operations under a chip power constraint [4, 5]. Besides, the high capacitance not only requires PCM to adopt current sense-amps which are much larger than voltage sense-amps used by DRAM, but also demands longer time to pre-charge the interconnect path for next operations.

Second, we propose LL-PCM architecture which can provide 2.8× and 3.1× lower resistance and capacitance than PCM for the interconnect path between a cell and a sense-amp/write-driver. This allows LL-PCM to use a 4.1× smaller charge pump which negates the cost to reduce the resistance/capacitance of the interconnect path. As the LL-PCM charge pump needs to supply 49% lower voltages than the PCM charge pump, it can pump the voltage to the target level faster and consume 3.1× less current. Consequently, LL-PCM can support concurrent write operations, whereas PCM cannot under the same maximum current constraint. These aspects allow LL-PCM to accomplish much shorter memory timing parameters such as tRCD, tRP, tCCD, and tWR than PCM.

2 ANALYSIS OF INDUSTRY PCM DESIGN

2.1 PCM Architecture and Operations

We analyze an implementation of an industry PCM chip [4] complying with the LPDDR2-NVM specification. PCM consists of 8 partitions (Figure 1(a)). Each partition consists of 128 tiles, *i.e.*, left and right sub-partitions, each of which is composed of 64 tiles (Figure 1(b)). Each tile has 4096 word-lines, 2048 bit-lines and

DAC '19, June 2–6, 2019, Las Vegas, NV, USA

ACM ISBN 978-1-4503-6725-7/19/06...$15.00
https://doi.org/10.1145/3316781.3317853

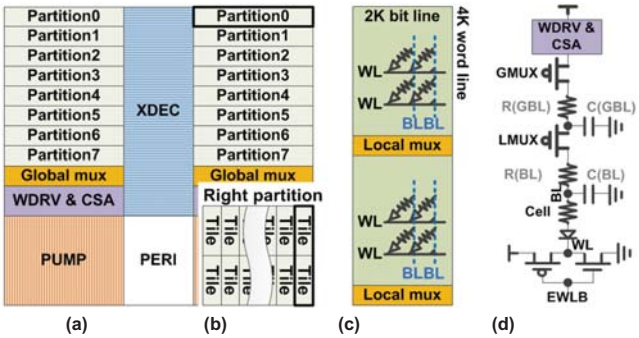

Figure 1: PCM architecture based on LPDDR2-N: (a) overall architecture, (b) right partition, (c) up/down tiles (d) interconnect path between a cell and a sense-amp/write-driver.

local multiplexers (Figure 1(c)). In this implementation, 256 (global) sense-amps ("CSA" in Figure 1(a)) are shared by 8 partitions to maximize the chip density. That is, only 2 cells out of 2048 cells per tile (or 256 out of 128×2048 activated cells in a partition) are connected to sense-amps per row activation. Unlike DRAM, it can share a sense-amp with many bit-lines, because PCM does not destroy the states of cells activated by a word-line. As 256 bits from a partition are sensed at a time, the size of a row buffer per partition is 32 bytes. As each PCM chip has 16-bit I/O, four ×16 PCM chips constitute a rank (*i.e.*, the size of a row buffer per rank is 128 bytes and there are 8 of such row buffers).

When a row is activated, the corresponding word-line ("WL") is driven to V_{SS}. Then, local multiplexers ("LMUX") and global multiplexers ("GMUX") are enabled to establish an interconnect path between an activated cell and a sense-amp through a bit-line ("BL") and a global bit-line ("GBL") (Figure 1(d)). The resistance and capacitance of the interconnect path are approximately 34.6KΩ (= R_{BL} + R_{LMUX} + R_{GBL} + R_{GMUX}) and 1.04pF (C_{BL} + C_{LMUX} + C_{GBL} + C_{GMUX}), respectively. The high resistance and capacitance of the interconnect path greatly affect the latency of PCM.

For read operations, the high resistance and capacitance reduce the sensing margin, requiring PCM to adopt expensive current sense-amps. As a current sense-amp is a few times larger than a voltage sense-amp used by DRAM, the number of sense-amps per PCM chip is limited to 256. The high resistance also requires PCM to provide very high voltage for its write drivers ("WDRV" in Figure 1(a)) for write operations, each of which consists of RESET and subsequent SET operations depending on a given value. For example, a RESET operation needs 150μA [4] per cell. This in turn requires a write driver to drive ~6V based on $V_{RESET} = I_{RESET} \times (R_{BL} + R_{LMUX} + R_{GBL} + R_{GMUX}) + V_{TH\text{-}DIODE}$.

2.2 Charge Pump

PCM requires a charge pump that converts its low input voltage (V_{IN}) to high output voltage (V_{OUT}). As a charge pump is required to supply higher output voltage, it gets larger, dissipates more power for voltage conversion, draw more current from the input voltage source, and consumes more time to reach at a target V_{OUT} from its discharged state. The PCM's charge pump converting 1.8V to ~6V requires four conversion stages comprising large capacitor with driver, clock generator, and control circuits. For example, a PCM charge pump consumes 10mm^2 (*i.e.*, ~ 17% of total PCM chip space) [4]. If the output voltage can be reduced by ~50%, the charge pump requires only two conversion stages and consumes only ~2.5mm^2 based on a model [6].

A 4-stage charge pump with conversion efficiency of 37% dissipates ~1.7W to supply 1W, whereas a 2-stage one with conversion efficiency of 57% consumes only ~0.75W to supply 1W based on a model [6]. Further, the power that the chip input voltage source needs to supply for the charge pump is proportional to the conversion ratio (V_{OUT}/V_{IN}) and inversely proportional to the conversion efficiency. This increases power consumption and thus restricts the number of bit-lines that can be simultaneously driven by write-drivers to 128 per chip under a chip power constraint. This in turn, prevents PCM from supporting concurrent write operations to different partitions [5]. Prior work indicates that a charge pump accounts for 42~75% (250~300ns) of total latency of write operations [6, 7]. The long latency of such a charge pump is because the charge pump needs to be transitioned to a discharged state (~3V) after write operations are completed [5]. This decreases the damaging effect of continuously applying high voltage to the gate oxide of transistors in PCM, but also increases the time to reach the target voltage level from a discharged state and power dissipation. If the voltage can be reduced to ~3V or a lower level, the charge pump does not need to be transitioned to the standby state as charge pumps for word-line drivers in DRAM.

3 LL-PCM ARCHITECTURE

PCM depicted in Figure 1(a) requires a long interconnect (high resistance/capacitance) path between a cell and a sense-amp/write-driver. This in turn requires (1) bulky charge pumps to supply high RESET and SET voltages and (2) large current sense-amps to distinguish states from weak bit-line signals. Such charge pumps and sense-amps incur a significant cost in terms of chip space, power, and/or latency. Building on our deep circuit-level analysis (Section 2), we propose LL-PCM substantially reducing latency without notably increasing the overall chip space in this section.

3.1 Hierarchical Sensing Adapted for PCM

To reduce the latency of PCM, we propose a hierarchical sensing

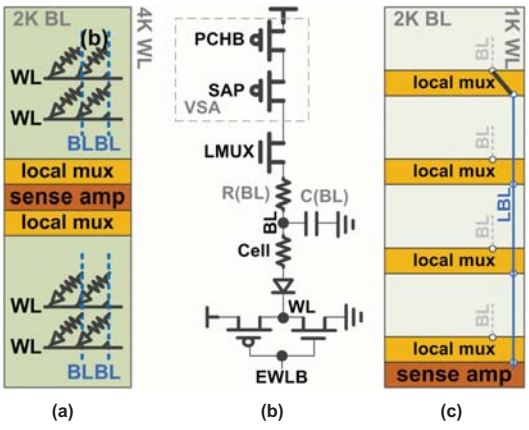

Figure 2: (a) Local sense-amps shared by top and bottom tiles, (b) LL-PCM local read/write path. (c) sub-tile architecture.

architecture. Specifically, we propose to place small local voltage sense-amps between two neighboring top and bottom tiles in a partition (Figure 2(a)). In a tile 1024 bit-lines share one local sense-amp (two local sense-amps per tile). Compared to PCM (Figure 1(a)), LL-PCM does not change the overall architecture (*e.g.*, the number of tiles in each partition and the number of partitions), but LL-PCM significantly reduces the amount of resistance /capacitance of an interconnect path between a cell and a sense-amp/write-driver (*i.e.*, $R_{LBL} + R_{LMUX} + R_{GBL} + R_{GMUX}$ (Figure 1(b)) to $R_{LBL} + R_{LMUX}$ (Figure 2(b)).

Nonetheless, we observe that the capacitance of a tile bit-line is too large to use a voltage sense-amp. If the capacitance is too large, it takes too long for a sense-amp to sense the correct state from a bit-line or the sense-amp may not correctly sense the state. The capacitance of a tile bit-line is primarily contributed by two components: (1) capacitance of diode anodes connected to a bit-line through cells and (2) interconnect capacitance of the bit-line. We observe that (1) is far more significant than (2) (*e.g.*, 552fF for 4096 cells and 27fF for the interconnect across a tile). To tackle this challenge, we propose to divide a tile into sub-tiles and place local multiplexers at each sub-tile (Figure 2(c)) where a tile is divided into four sub-tiles (*i.e.*, 1024 cells per sub-tile bit-line) as an example. This sub-tile architecture requires more local multiplexers (*i.e.*, pass transistors), but it does not increase the number of local sense-amps compared to Figure 2(a), as sub-tiles in a tile share the local sense-amps.

Our circuit analysis shows that LL-PCM reduces the resistance and capacitance seen by a local sense-amp to 36% and 32% of PCM, respectively (Section 5). This in turn allows LL-PCM to use smaller sense-amps and charge pumps than PCM. A voltage sense-amp is ~5× smaller than a current sense-amp for the same bit-line capacitance although it is 20% slower [8, 9]. As the reduced resistance proportionally decreases RESET and SET voltages, a 2-stage charge pump is sufficient for LL-PCM, whereas a 4-stage charge pump is necessary for PCM (Section 2.2). Later, we

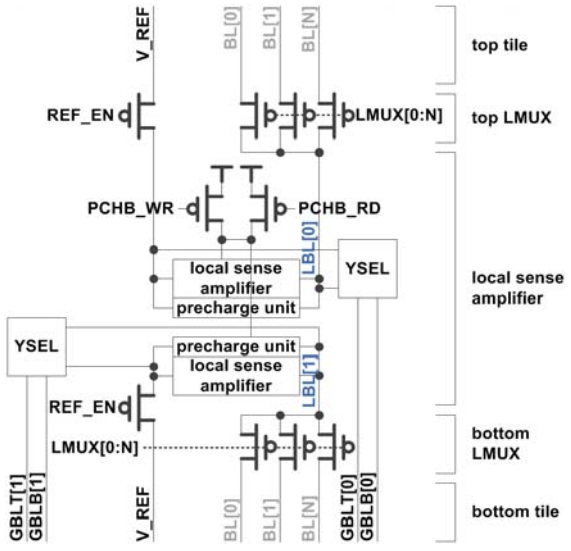

Figure 3: Local voltage sense-amp circuit.

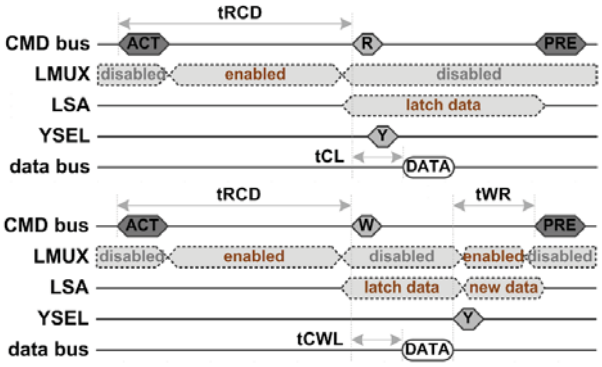

Figure 4: LL-PCM read (top) and write (bottom) operations.

demonstrate that the cost of placing these local multiplexers per sub-tile and a local sense-amp per tile can be mostly negated by a smaller charge pump of LL-PCM (Section 5).

Figure 3 describes a local voltage sense-amp circuit in LL-PCM. A sense-amp is shared by 1024 bit-lines from a tile. A signal from one bit-line chosen by a local multiplexer is amplified by a local sense-amp instead of being sent to a global bit-line/sense-amp as PCM. Unlike DRAM, PCM can amplify signals from only selected bit-lines, as its read operation is not destructive. The sensed state remains latched until a pre-charge command is received. Note that a "pre-charge" command in fact discharges bit-lines and sense-amp nodes to V_{SS} in both PCM and LL-PCM. Disconnecting the bit-line from the sense-amp right before sensing a state has been commonly adopted by SRAM to improve the speed and reduce the energy consumption of fully swinging a bit-line to the full V_{DD} or V_{SS} level. It also reduces tRP in LL-PCM when sensing other bit-lines connected to other cells from the same activated row in subsequent cycles, as only the sense-amp needs to be pre-charged.

Figure 4(top) depicts a read operation. For a given row address, the corresponding word-line is activated. That is, the word-line driver drives the word-line to V_{SS} to make the connected diodes conduct the current. At the same time, a tile bit-line selected by an enabled local multiplexer is connected to a local voltage sense-amp. Then the transistor connected to "PCHB_RD" in Figure 3 provides the voltage for the sense-amp to drive the selected bit-line. When the voltage level of the bit-line is sufficiently developed after tRCD, the local multiplexer disconnects the bit-line from the sense-amp, and then the sense-amp is triggered by a sense-amp enable signal. If the voltage level of a bit-line driven by PCHB_RD does not rise above the reference voltage ("V_REF") due to low resistance of the connected cell (1KΩ), the sense-amp senses it as 0. Otherwise, the sense-amp sense the bit-line voltage level as 1. We provide a circuit-level analysis on the impact of sub-tile size (*i.e.*, cells per sub-tile bit-line) on the sensing delay, size, and power (Section 5).

Finally, we propose to place global "voltage" sense-amps at the same location as current sense-amps in PCM. After completing a sensing operation, the local sense-amps start to drive the global bit-lines connected to the global multiplexers and voltage sense-amps. Such a deployment of global sense-amps is similar to DRAM where global sense-amps (*i.e.*, row buffers) in a bank re-amplify weak signals sent from local sense-amps in mats [10]; the global bit-lines shared by all the mats in a bank are long while each mat has limited

column-direction space (or pitch) to put many strong drivers.

3.2 Hierarchical Driving Adapted for PCM

Building on the hierarchical sensing architecture, we propose a hierarchical driving architecture for write operations in this section. A PCM write operation requires two different high voltages: RESET and SET voltages. Figure 1(a) shows that the charge pump located at the bottom ("PUMP") can directly supply these two voltages to the write drivers ("WDRV") in PCM. In LL-PCM, however, the write drivers at the bottom drive the nominal voltage (= 1.8V) to the local sense-amps of a target partition through the global bit-lines and global multiplexers. After the local sense-amps receive and latch the values represented by the nominal voltage levels, the drivers in the local sense-amps start to "locally" drive the high RESET or SET voltage to the activated cells through the sub-tile bit-lines selected by the local multiplexers.

With this architecture, a driver needs to drive either RESET or SET voltage through only a local multiplexer and a sub-tile bit-line. This in turn allows LL-PCM to use much lower RESET and SET voltages than PCM to deliver the same amount of required RESET or SET current to a target cell. Our circuit-level analysis shows that LL-PCM needs 2.9V and 2.2V, whereas PCM requires 5.9V and 4.1V for RESET and SET voltages, respectively. That is, a 2-stage charge pump is sufficient for LL-PCM, whereas a 4-stage charge pump is necessary for PCM. These lower RESET and SET voltages can reduce the write latency (tWR) because of two reasons. First, it takes less time for a charge pump to raise the voltage to a desired level. Second, the charge pump needs to be turned off when there is no pending write operation, because of a reliability concern at high voltage [5]. This in turn increases tWR because it takes long time for a charge pump to raise the voltage to the target level from a discharged state. We see that this contributes to 42% of tWR in PCM [5]. In contrast, the charge pump does not need to be transitioned to a discharged state in LL-PCM, since it is demanded to supply slightly lower output voltage (2.9V) than the output voltage of the PCM charge pump in a discharged state (~3V).

A transistor connected to "PCHB_WR" in the local sense-amp depicted in Figure 3 is to provide the RESET or SET voltage for the write-driver to drive the necessary current to a target cell through a sub-tile bit-line selected by a local multiplexer. As the RESET voltage and current of LL-PCM are 2× and 3× lower than those of PCM (Section 5), LL-PCM can deliver these lower voltages to the drivers in local sense-amps with smaller loss through a low-resistance power delivery network with a small cost.

Figure 4(bottom) depicts a write operation of LL-PCM when a write buffer in a PCM chip is full or absent. First, a row is activated and then a write command is placed on the command (CMD) bus after tRCD. After tCWL a value placed on the data bus by the MC arrives at the (global) write driver located at the bottom of PCM. This write driver differs from ones in PCM as it drives the value to the sense-amp (latch) of a target tile at the nominal voltage. As soon as the sense-amp latches the value, the transistor connected to PCHB_WR starts to provide the RESET voltage/current. As the RESET and SET voltages are sequentially applied, we locally step-down RESET voltage to SET voltage at each tile using the same current mirror adopted by prior work. Then an appropriate local multiplexer is enabled to connect the latch output to a target cell through a sub-tile bit-line. After the RESET or SET time has elapsed, the local multiplexer is disabled. The value written to the cell remains in the latch until a pre-charge command is executed.

Table 2 : PCM technology parameters in 22nm.

Cell size [4]	$4F^2$
Current for RESET and SET [4]	150μA/100μA
Diode resistance at 4V [17]	2.5KΩ
Resistance of PCM in reset/set states [18]	1MΩ/1KΩ
Local bit-line resistance [15]	93.25Ω/μm
Global bit-line resistance [16]	4Ω/μm
Pulse width for RESET and SET [4, 5]	50ns/150ns

Table 1: Timing parameters.

	PCM	LL-PCM
tRCD	120ns [5]	40n
tRP	15ns	5ns
tCL	10ns [9]	15ns
tRTP	2nCK [9]	2nCK
tCCD_RD	4nCK [9]	4nCK
tCCD_WR	338ns	44ns/220ns
tWR	638ns	220ns

4 METHODOLOGY

4.1 Circuit Model

Technology parameters. Table 2 tabulates the technology parameters used for our analysis. We also perform SPICE simulation with transistor parameters from a 22nm Low-Power (LP) Predictive Technology Model (PTM).

Timing parameters. The timing parameters, tRCD, tWR, tRP, and tCL of PCM are from prior work [4, 5] and NVSIM [11] that takes the parameters tabulated in Table 2. Those of LL-PCM are estimated by NVSIM and SPICE simulation that take the parameters tabulated in Table 2. Some models in NVSIM are adapted to match the area and timing parameter such as tRCD and tCL of PCM [4]. The tCL of PCM (= 10ns) [5] is shorter than that of DRAM. The row buffers of PCM are located at the I/O peripheral region, whereas those of DRAM are located at each bank. tWR (= 338ns) is from NVSIM. Specifically, tWR of PCM comprises RESET and SET times since every write operation has a RESET operation followed by a SET operation in PCM. Based on prior work [7], we also incorporate the latency of turning on/off the charge pumps (= 300ns) into tWR (638ns = 338ns + 300ns). tRP (= 15ns and 5ns) of PCM and LL-PCM is from NVSIM and our SPICE simulation, respectively. Table 1 summarizes the timing parameters.

Table 3: Energy parameters.

	PCM	LL-PCM
ACT (nJ)	5.9	4.0
RD (pJ/bit)	5.5	16.8
WR (pJ/bit)	3286.0	459.0
PRE (nJ)	3.5	3.2
PRE_SA (nJ)	-	0.4

Table 4: The resistance of the write path with high voltages.

	PCM	LL-PCM
Global bit-line	11.0KΩ	-
Local bit-line	-	0.5KΩ
Bit-line	15.3KΩ	3.8KΩ
Transistors	4.0KΩ	6.0KΩ
Cells +diode	3.5KΩ	3.5KΩ
Word-line	0.8KΩ	0.8KΩ
Total	34.6KΩ	14.4KΩ

Energy parameters. We estimate energy consumption of various operations of PCM and LL-PCM with NVSIM that takes the technology parameters from Table 2.Table 3 tabulates the estimated energy parameters of PCM and LL-PCM. As LL-PCM uses local sense-amps, it increases the energy consumption of read operations. However, LL-PCM decreases energy consumption for activation and pre-charge operations because of a shorter interconnect path between a cell and a sense-amp.

4.2 Architecture

Simulator. We take *gem5* and adapt it to model MC to support PCM and LL-PCM with the DDR3-1600 protocol with the timing and energy parameters from Table 1 and Table 3, respectively. Lastly, we model a 4-core out-of-order processor operating at 3.2GHz with private 32KB L1 caches, private 256KB L2 cache, shared 4MB L3 cache and 2 DDR3 DRAM channels.

Benchmark. We use the same SPEC 2006 benchmark mixes as prior work [12]. MIX1-3, MIX4-6, and MIX 7-9 incur high, medium, and low levels of last-level cache misses per kilo instructions (MPKI), respectively. To measure performance, we fast-forward 1B instructions, and run until the slowest core executes 1B instructions.

5 EVALUATION

Circuits. The hierarchical architecture of LL-PCM eliminates the resistance of global bit-lines and reduces the resistance of local bit-lines in the interconnect path that needs to conduct high supply voltage for write operations. We conservatively use the same transistor size for the write path although we can decrease the transistor size. Table 4 tabulates the resistance values of various components affecting the RESET and SET voltages in PCM and LL-PCM using the technology parameters tabulated in Table 2 and the dimension of PCM components tabulated in Table 5. As the resistance of the interconnect path is reduced from 34.6 KΩ to 14.4 KΩ, we estimate that the RESET voltage can be reduced from 5.9V (=150μA×34.6KΩ+$V_{TH\text{-}DIODE}$) to 2.9V (=150μA× 14.4KΩ+$V_{TH\text{-}DIODE}$). That is, PCM and LL-PCM need 4- and 2-stage charge pumps, respectively. To calculate the total current that the 1.8V voltage source needs to supply for the charge pumps, we first estimate the efficiency of charge pumps to supply 19.2mA (= 150μA ×128) for 128 cells. We simulate the efficiency of 2-, 3-, and 4-stage charge pumps for different output voltages based on a model [6]. Our simulation shows that the efficiency of 2- and 4-stage charge pumps at V_{OUT} = 5.95V and 2.89V for PCM and LL-PCM is 0.37 and 0.57, respectively. That is, the PCM and LL-PCM charge pumps require the 1.8V voltage source to supply 179mA (= 5.9V/1.7V/0.37×19.2mA), which is aligned with a value in prior work [5] and 57mA (= 2.9V/1.7V/0.57×19.2mA), respectively, where we use 1.7V instead of 1.8V for the worst-case voltage.

Table 5: Height (μm) of components with 9.4mm width.

	PCM	LL-PCM
Charge pump	1064	257
Local multiplexer	256	1024
Voltage sense-amp/write-drivers	-	64
Array and I/O peripheral circuits	4980	4980
Total	6300	6325

Table 6 : Evaluated configurations.

Config	tRCD		tCCD_WR	
	Different	Same Row	Same	Different
PCM	120ns	120ns	338ns	338ns
LL-PCM	40ns	40ns	220ns	220ns
LL-PCM/B	40ns	40ns	220ns	44ns

Table 5 tabulates the calculated height values of components in PCM and LL-PCM based on an industry PCM implementation [4]. We first calculate the height of components after determining that the height and width of a die is 6.3mm and 9.4mm, respectively. Based these dimensions, we conservatively calculate the height of LL-PCM components as follows. The charge pump area is proportional to the number of stages and the amount of current for a RESET or SET operation. We estimate the charge pump area of LL-PCM based on a model from [6]. This gives us 2.42mm^2 (or 257μm ×9.4mm) for a 2-stage charge pump of LL-PCM. LL-PCM requires 4× more local multiplexers than PCM when it adopts 4 sub-tiles per tile. Thus, the height of PCM multiplexers are multiplied by four to estimate that of LL-PCM multiplexers. This height increase is mostly negated by the height decrease by a smaller charge pump of LL-PCM. We estimate that the dimension of the voltage sense-amp/write-driver region is 8μm×9.4mm, primarily determined by the number of horizontal metal-2 interconnects needed to deliver the power for the write-drivers. In summary, the height increase of LL-PCM is less than ~1% compared to PCM.

Our simulation shows that the sensing delay decreases from 56ns to 36ns, 23ns, 15ns, and 10ns at the cost of area increase up to 50%, as the number of cells per bit-line decreases from 4096 to 2048, 1024, 512, and 256. The sensing latency is defined from the point that a word-line is enabled to a point that a sense-amp can be enabled when a conservative operating condition is considered (*i.e.*, 90% V_{DD}, 10% device length variation on sense-amps, and 5% of threshold voltage variation). In summary, LL-PCM with four sub-tiles per tile (*i.e.*, 1024 cells per bit-line) offers the best trade-off in terms of the sensing delay, area, and power, reducing the latency and power consumption with only ~1% area increase.

Performance and Energy. To evaluate the performance of LL-PCM at the system level, we first consider three configurations: PCM, LL-PCM and LL-PCM/B. LL-PCM is to evaluate the benefit of tRCD and tCCD_WR reduced by the hierarchical sensing and driving architecture (Section 0). LL-PCM/B is to evaluate the benefit of the write burst atop LL-PCM. The key timing parameters of each configurations are listed in Table 6. Atop these

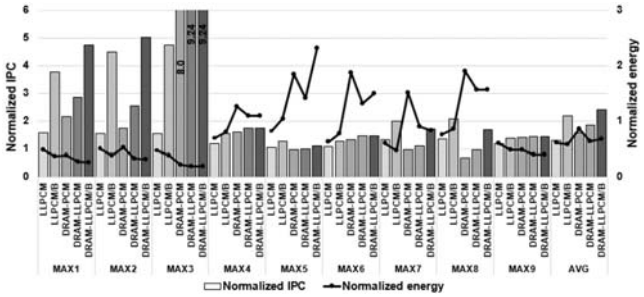

Figure 5: Performance and energy comparison of PCM and LL-PCM.

configurations, we evaluate a hybrid memory system composed of PCM and DRAM (i.e., hardware-managed cache with the line size of 4KB and every tag stored in the memory controllers). This gives three more configurations: DRAM-PCM, DRAM-LL-PCM, and DRAM-LL-PCM/B.

Figure 5 plots the performance and energy of these six configurations, normalized to that of PCM. We demonstrate that LL-PCM and LL-PCM/B provide 30% and 119% higher geometric-mean IPC than PCM. LL-PCM gives higher IPC than PCM because tRCD, tRP, and tWR of LL-PCM is ~3× shorter than those of PCM. LL-PCM/B gives 68% higher geometric-mean IPC than LL-PCM, as it offers almost 5× shorter write latency (tCCD_WR and tCCD_WR) than LL-PCM. DRAM-LL-PCM and DRAM-LL-PCM/B offer 21% and 71% higher IPC than DRAM-PCM. Lastly, we analyze the energy consumption of these configurations. LL-PCM and LL-PCM/B consume 40% and 43% lower energy than PCM, because the power dissipation of the LL-PCM charge pump is 32% lower than that of the PCM charge pump. DRAM-LL-PCM and DRAM-LL-PCM/B consumes 26% and 22% lower energy than DRAM-PCM. These results show that LL-PCM can give a significant performance and energy benefits although it is used with DRAM.

6 RELATED WORK

Prior work proposed to improve performance for both writes and read operations of PCM (*e.g.*, [2]. While most of these studies focused on optimizing PCM write operations, two different techniques were proposed to reduce the read latency in an attempt to improve overall system performance [13]. First, the sensing circuitry is enabled to latch the output of sense-amps before the specified sensing time. Second, the read voltage is increased to reduce the sensing time itself. The latency overhead associated with charge pumps was presented [14]. Even though this study modeled the details of current supply, charge/discharge latency, and power dissipation of the charge pumps, it changes neither charge pump or sensing circuits. Instead, it re-orders the RESET operations across different time slots to minimize the peak power consumption of write operations, and generates the target output voltage by utilizing idle times.

7 CONCLUSION

In this paper, we proposed LL-PCM adopting a hierarchical sensing and driving architecture synergistically integrated with memory scheduling policies exploiting some unique aspects of LL-PCM. Our analyses show that LL-PCM (DRAM+LL-PCM) with the memory scheduling policies can offer 119% (71%) higher performance and consume 43% (22%) lower energy than PCM (DRAM+PCM) with ~1% increase in chip size.

ACKNOWLEDGMENTS

This work was supported in part by NRF 2016R1C1B2015312, DOE DEAC02-05CH11231, NRF 2015M3C4A7065645, MemRay grant, and NSF CNS-1850317. This work was completed when Nam Sung Kim was at the University of Illinois, Urbana-Champaign.

REFERENCES

[1] G. Dhiman, et al, "PDRAM: A Hybrid PRAM and DRAM Main Memory System," *ACM/IEEE DAC*, 2009.

[2] H. Lee, et al., "An Energy and Performance-Aware DRAM Cache Architecture for Hybrid DRAM/PCM Main Memory Systems," in *IEEE ICCD*, 2011.

[3] M. Qureshi and G. Loh, "Fundamental Latency Trade-Off in Architecting DRAM Caches: Outperforming Impractical SRAM Tags with a Simple and Practical Design," *IEEE/ACM MICRO*, 2012.

[4] Y. Choi, et al., "A 20nm 1.8 V 8Gb PRAM with 40MB/s Program Bandwidth," *IEEE ISSCC*, 2012.

[5] K. Lee, et al., "A 90 nm 1.8 V 512 Mb Diode-Switch PRAM with 266 MB/s Read Throughput," *JSSC,* 43(1), 2008.

[6] G. Palumbo, et al., "Charge Pump Circuits: An Overview on Design Strategies and Topologies," *IEEE Circuits and Syst Magazine,* 10(1), 2010.

[7] S. Kang, et al., "A 0.1-μm 1.8-V 256-Mb Phase-Change Random Access Memory (PRAM) with 66-MHz Synchronous Burst-Read Operation," *IEEE JSSC,* 42(1), 2007.

[8] K. Sasaki, et al., "A 7ns 140-mW 1-Mb CMOS SRAM with Current Sense Amplifier," *IEEE JSSC,* 27(11), 1992.

[9] P. Nasalski, et al. "SRAM Voltage and Current Sense Amplifiers in Sub-32nm Double-Gate CMOS Insensitive to Process Variations and Transistor Mismatch," in *IEEE ISCAS*, 2009.

[10] Y. Moon, et al., "1.2 V 1.6 Gb/s 56nm 6F2 4Gb DDR3 SDRAM with Hybrid I/O Sense Amplifier and Segmented Sub-Array Architecture," *IEEE ISSCC*, 2009.

[11] X. Dong, et al., "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *ACM/IEEE TCAD,* 31(7), 2012.

[12] T. Ham, et al., "Disintegrated Control for Energy-Efficient and Heterogeneous Memory Systems," *ACM/IEEE HPCA*, 2013.

[13] P. Nair, et al., "Reducing Read Latency of Phase Change Memory via Early Read and Turbo read," *IEEE HPCA*, 2015.

[14] L. Jiang, et al., "A Low Power and Reliable Charge Pump Design for Phase Change Memories," *ACM/IEEE ISCA*, 2014.

[15] "International Technology Roadmap for Semiconductor (ITRS)," 2015.

[16] K. Banerjee, et al., "3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration," *Proceedings of the IEEE,* 89(5), 2001.

[17] S. Lai, "Current Status of the Phase Change Memory and Its Future," in *ACM/IEEE IEDM*, 2003.

[18] G. Burr, et al., "Phase Change Memory Technology," *J. of Vacuum Science & Technology B,* 28(2), 2010.