PCRAMsim: System-Level Performance, Energy, and Area Modeling for Phase-Change RAM

Xiangyu Dong
Computer Science & Engineering Department
Pennsylvania State University
xydong@cse.psu.edu

Norman P. Jouppi
Exascale Computing Lab
Hewlett-Packard Labs
norm.jouppi@hp.com

Yuan Xie
Computer Science & Engineering Department
Pennsylvania State University
yuanxie@cse.psu.edu

ABSTRACT

Phase-change random access memory (PCRAM) is an emerging memory technology with attractive features, such as fast read access, high density, and non-volatility. Because of these attractive properties, PCRAM is regarded as a promising candidate for future universal memories, and system-level designers could open up new design opportunities by leveraging this new memory technology. However, the majority of the PCRAM research has been at the device level, and system-level design space exploration using PCRAM is still in its infancy due to the lack of high-level modeling tools for PCRAM-based caches and memories. In this paper, we present a PCRAM model, called *PCRAMsim*, to bridge the gap between the device-level and system-level research on PCRAM technology. The model is validated against industrial PCRAM prototypes. This new *PCRAMsim* tool is expected to help boost PCRAM-related studies such as next-generation memory subsystems.[1]

1. INTRODUCTION

Phase-change random access memory (PCRAM) is an emerging memory technology with many attractive features, which include fast read access, high density, non-volatility, positive response to increasing temperature, superior scalability, and zero standby leakage [3]. These properties make PCRAM a promising candidate for future universal memories. Compared to other emerging non-volatile memories such as Magnetic RAM (MRAM) and Ferroelectric RAM (FeRAM), PCRAM memory has excellent scalability, which is critical to the success of any emerging memory technologies. Consequently, much R&D activity in industry, including at IBM, Intel, Hitachi, ST Microelectronics, and Samsung, has been on PCRAM technology [1, 3, 5, 7, 9, 15, 16, 18, 24].

The unique features of PCRAM technologies open up new opportunities for designers. For example, many traditional memory subsystems or components can be reconsidered and optimized by leveraging this emerging technology. In addition, PCRAM can be an improved replacement for a NAND flash device which is much slower and can only be written about 10^5 times. PCRAM can enable storage-class memory that combines the high performance and robustness of solid-state memories and the low cost of the conventional hard-disk [3].

[1]X. Dong and Y. Xie were supported in part by NSF grants 0702617, 0720659, 0903432 and SRC grants.

ICCAD'09, November 2–5, 2009, San Jose, California, USA.

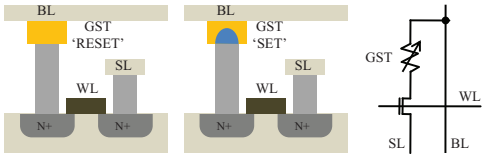

Figure 1: The schematic view of a PCRAM cell with MOSFET selector transistor (BL=Bitline, WL=Wordline, SL=Sourceline)

Many modeling tools have been developed during the last decade to enable system-level design exploration for SRAM- or DRAM-based cache and memory design. For example, CACTI [20, 22] is a tool that has been widely used in the computer architecture community to estimate the speed, power, and area of SRAM and DRAM caches. Evans and Franzon [4] developed an energy model for SRAMs and used it to predict an optimum organization for caches. eCACTI [10] incorporated a leakage power model into CACTI. Muralimanohar *et al.* [14] modeled large capacity caches through the use of an interconnect-centric organization composed of mats and request/reply H-tree networks. In addition, *DRAMsim* [21] is a tool simulating the behaviour of commodity DRAM devices. Similarly, it is imperative to have a high-level model for PCRAM chips or modules, with the extraction of important parameters such as access latency, dynamic access power, leakage power, and die area, to facilitate system-level analysis for PCRAM-based design. Mangalagiri *et al.* [11] extended the CACTI to evaluate the performance, power, and area for PCRAM caches. However, their model used too simplified and optimistic assumptions and didn't get validated.

In this paper we develop a PCRAM model called *PCRAMsim* to bridge the gap between the abundant research activity at the lower level and the lack of a high-level PCRAM model. We also show how to use this model to facilitate system-level performance and power analysis for applications that adopt this emerging technology.

2. PHASE-CHANGE MEMORY

This section gives the background information on *phase-change memory* technology, which has several desirable properties such as fast read access, high density, non-volatility, positive response to increasing temperature, superior scalability, and zero standby leakage.

Phase-Change Mechanism.

Unlike conventional SRAM and DRAM technologies that use electrical charges to store information, PCRAM uses phase-change materials as its name implies. Chalcogenide-based material is one

Table 1: Comparison among SRAM, DRAM, Flash, and PCRAM

	SRAM	DRAM	NAND Flash	PCRAM
Cell size	$> 100F^2$	$6 - 8F^2$	$4 - 6F^2$	$4 - 20F^2$
Read time	$\sim 10ns$	$\sim 10ns$	$5\mu s - 50\mu s$	$10ns - 100ns$
Write time	$\sim 10ns$	$\sim 10ns$	$2 - 3ms$	$100 - 1000ns$
Standby power	leakage	leakage & refresh	zero	zero
Write endurance	10^{18}	10^{15}	10^5	$10^8 - 10^{12}$
Non-volatility	No	No	Yes	Yes

Table 2: Scaling rules for GST parameters (Source: [16])

Parameter	Factor
GST contact area	$1/k^2$
Thermal resistance	k
SET resistance	k
RESET resistance	k
Programming voltage	1
Programming current	$1/k$

of the phase-change materials which can be switched between a crystalline phase (SET or "1" state) and an amorphous phase (RESET or "0" state) with the application of heat. The crystalline phase shows high optical reflectivity and low electrical resistivity, while the amorphous phase is characterized by low reflectivity and high resistivity. The chalcogenide-based materials in recent PCRAM research are usually alloys of germanium, antimony, and tellurium (GeSbTe, or GST) because of their fast crystallization behavior [24]. Also, because of this phase-change mechanism, PCRAM is a non-volatile memory technology.

PCRAM Read Operation.

To read data stored in PCRAM cells, a small voltage is applied across the GST. Since the SET status and the RESET status have a large difference in their equivalent resistances, stored data bits are sensed by measuring the resulting current. The read voltage is set to be sufficiently high to provide a sensible current but remains low enough to avoid write disturbance. As shown in Fig. 1, every PCRAM cell contains one GST and one selector transistor. This structure has a name of "1T1R" where T means the MOSFET transistor, and R stands for the GST. The GST in each PCRAM cell is connected to the drain region of the MOSFET in series so that the data stored in PCRAM cells can be accessed by controlling a wordline. Usually, the source line is connected to the ground, and the voltage of bitlines during read operations is clamped between $0.2V$ to $0.4V$ [5, 15].

PCRAM Write Operation.

In order to change the phase of PCRAM cells from one state to the other, there are two kinds of PCRAM write operations: the SET operation that switches the GST into the crystalline phase and the RESET operation that switches the GST into the amorphous phase. The SET operation crystallizes GST by heating it above its crystallization temperature, and the RESET operation melt-quenches GST to make the material amorphous (see Fig. 2). These two operations are controlled by electrical current: high-power pulses for the RESET operation heat the memory cell above the GST melting temperature; moderate power but longer duration pulses for the SET operation heat the cell above the GST crystallization temperature but below the melting temperature [18]. The temperature is controlled by passing a specific electrical current profile and generating the required Joule heat. The RESET/SET writing current pulses have shapes similar to those shown in Fig. 2. Recent PCRAM prototype chips demonstrate that the RESET latency can be as fast as $100ns$, and the peak SET current can be as low as $100\mu A$ [5, 15].

PCRAM Cell Size & Scalability.

The cell size of PCRAM is mainly constrained by the current driving ability of the MOSFET selector transistor. Studies show the achievable cell size can be as small as $10 - 20F^2$ [1, 7], where F is the feature size. When the MOSFET selector transistor is substituted by a diode, the PCRAM cell size can be reduced to $4F^2$ [9].

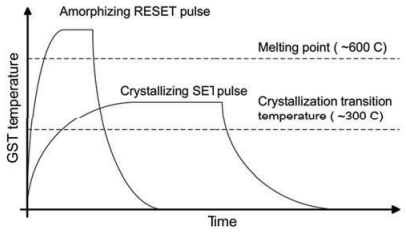

Figure 2: The temperature-time relationship during SET and RESET operations

Pirovano *et al.* [16] show that PCRAM has excellent scalability behaviors. Section 3 discusses more details about the PCRAM cell design.

Thermal Effect.

Since the PCRAM status is reversed by raising the temperature to certain levels, the higher the temperature the easier it is to switch the phase of GST. In contrast, at high temperatures, SRAM suffers from increasing leakage power, DRAM requires higher refresh power, and other non-volatile memories like NAND flash become more unreliable.

Comparison to Other Memory Technologies.

Table 1 compares the properties of PCRAM with other memory technologies. Similar to DRAM and SRAM, PCRAM has superior read latency. The write latency of PCRAM is significantly better compared to other non-volatile devices but slower compared to its volatile counterparts. The drawback of current PCRAM devices are their relative low endurance (the number of write operations in lifetime). However, it has been predicted by ITRS that the endurance could be improved to 10^{12} in the near future [6], making it suitable for memory and cache applications.

3. PCRAMSIM

To evaluate the emerging PCRAM technology, we develop a tool, called *PCRAMsim*, that models PCRAM devices in terms of area, delay, dynamic energy, and leakage power. The work is extended from CACTI [20], a widely-used tool in the computer architecture community for modeling of SRAM/DRAM-based caches and main memories. Since our work on *PCRAMsim* is an enhancement to CACTI, a detailed description of all the components in the CACTI circuit-level model are not covered by this paper. Instead, in this paper, we focus on the major changes we have made for *PCRAMsim*.

3.1 PCRAM Cell Modeling

The most significant difference between PCRAM technology and SRAM/DRAM is their distinct memory cells. PCRAM cell is typically a "1T1R" structure, while SRAM cell is a conventional "6T" structure and DRAM cell is usually a "1T1C" structure. The difference of cell structures directly leads to different cell sizes. The

Table 3: Current driving abilities of MOSFETs and estimated minimum PCRAM cell sizes

Technology node	$130nm$	$90nm$	$65nm$	$45nm$	$32nm$
I_{DS} per micron of channel width	$823\mu A$	$1005\mu A$	$1169\mu A$	$1360\mu A$	$1560\mu A$
Estimated minimum RESET current [6]	$500\mu A$	$294\mu A$	$202\mu A$	$125\mu A$	$70\mu A$
Estimated SET current [6]	$67\mu A$	$39\mu A$	$27\mu A$	$17\mu A$	$9\mu A$
Required channel width	$4.67F$	$3.25F$	$2.66F$	$2.04F$	$1.40F$
Required minimum cell size	$22.68F^2$	$17.00F^2$	$14.64F^2$	$12.16F^2$	$9.60F^2$

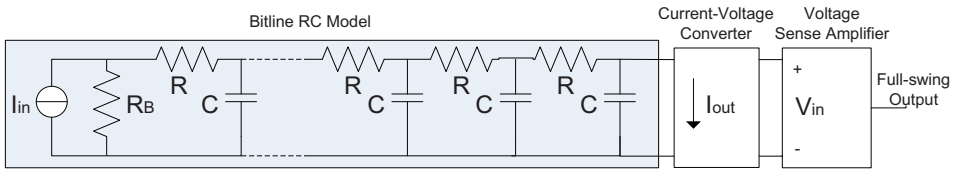

Figure 3: Analysis model for PCRAM current-mode sensing scheme. The model has three separate parts: bitline RC model, current-voltage converter, and voltage sense amplifier.

SRAM and the embedded DRAM cells presented in [20] have areas of $120 - 150F^2$ and $19 - 26F^2$, respectively, and the commodity DRAM cell area is about $6 - 8F^2$ [6]. The PCRAM cell area is constrained by two factors: the size of chalcogenide-based phase-change materials (GSTs) and the size of the selector device that could be a MOSFET [5, 7], a BJT [15], or a diode [9]. Basically, the size of GSTs determines the minimum required programming current, which further decides the size of the selector device[2].

For the scaling rule of GSTs, Pirovano *et al.* [16] reported a detailed scaling analysis using a physics-based electrothermal model of a cell verified by measurements conducted on sample devices. Their study shows that the RESET current can be scaled downward by scaling the contact area of the GST. A generic scaling rule with constant voltage is listed in Table 2. This scaling rule implies that a smaller GST size is usually preferred because it can lead to a lower requirement on the programming current amplitude.

The consideration of the selector device sizing is mainly focused on how to drive the RESET current of GST programming, since the saturation current of the selector device has to be greater than the required RESET current. The traditional MOSFET, BJT, and diode can all be used as the selector device for PCRAM cells. Although BJTs and diodes usually have stronger current driving abilities than a MOSFET of the same size, using a diode or a BJT has the disadvantages of parasitic currents to neighboring cells as well as being incompatible with conventional CMOS fabrication. Thus, in *PCRAMsim* tool, MOSFETs are chosen as the default PCRAM cell selector device at the expense of larger area, though the BJT-selected and the diode-selected designs are kept in the *PCRAMsim* model. In order to estimate the current driving ability of MOSFET devices, a small test circuit using HSPICE with PTM models [25] is simulated, and the values of I_{DS} per micron are listed in Table 3. The minimum RESET current is projected by ITRS [6], and therefore the required channel width of MOSFET devices can be calculated. If the distance between two MOSFETs is assumed to be $1F$, the minimum required PCRAM cell size can also be estimated, as tabulated in Table 3. We find that, as the technology shrinks, the PCRAM cell has an area ranging from $22.68F^2$ to $9.60F^2$. This cell size estimation is well consistent with some PCRAM prototype designs: Kang *et al.* [7] showed a $0.10\mu m$ MOSFET-selected PCRAM design with $16.6F^2$ cells; Ahn *et al.* [1] showed a $0.12\mu m$ design with $20F^2$ cells. Our default PCRAM cell model is built by using the cell sizes shown in Table 3.

[2]PCRAM cell properties have large variations in the literature. In our model, we provide an approximation of PCRAM cell parameters by default. For users who have accurate numbers, we provide an interface for customization.

Besides the PCRAM cell area, another important part of the PCRAM cell model is the inherent properties of PCRAM cell itself. These properties include the resistivity of the GST, the duration and shape of programming pulses during SET and RESET operations, and so on. To avoid GST material-level analysis, the default GST parameters are set based on the published literature [15] and the scaling rule shown in Table 2, although users are allowed to input their accurate numbers to *PCRAMsim*. Our default parameters to model the GST inherent properties are listed in Table 4.

Table 4: Default GST inherent parameters used in *PCRAMsim*

	RESET	SET
Pulse Duration	$40ns$	$100ns$
$130nm$	$3M\Omega$	$10K\Omega$
$90nm$	$4.3M\Omega$	$14.4K\Omega$
$65nm$	$6M\Omega$	$20K\Omega$
$45nm$	$8.7M\Omega$	$28.9K\Omega$
$32nm$	$12M\Omega$	$40K\Omega$

3.2 Sensing Scheme Modeling

The sensing scheme for the PCRAM module is different from the one modeled in CACTI, which is originally designed for SRAM data reading. The most significant difference is that the sense amplifier modeled in CACTI is voltage-mode while the sensing scheme for PCRAM data reading is current-mode. As mentioned in Section 2, the bitline voltage is clamped to 0.2 - $0.4V$ during the PCRAM read operation. The PCRAM status is read out by measuring the resulting current: the current on the bitline is compared to the reference current generated by reference cells, the current difference is amplified by current-mode sense amplifiers, and they are eventually converted to voltage signals. Fig. 3 shows the current-mode sensing scheme modeled in *PCRAMsim*, which contains three major parts: the bitline RC model, the current-voltage converter, and the conventional voltage-mode sense amplifier. The current-mode bitline RC model and the current-voltage converter are the models we have newly developed in *PCRAMsim*.

3.2.1 Bitline RC Model

The bitline RC delay and power are re-modeled in *PCRAMsim* because the input resistance of ideal voltage-mode sensing devices is infinite but it becomes zero for an ideal current-mode one. Seevinck *et al.* [19] have analyzed the RC delay for both voltage-mode and current-mode signals. Their analysis is performed entirely in the time domain and results in a simplified expression, which is consistent with the Elmore delay model used in CACTI.

Table 5: The delay and power look-up table of current-voltage converter

	$130nm$	$90nm$	$65nm$	$45nm$	$32nm$
Delay	$0.49ns$	$0.53ns$	$0.62ns$	$0.80ns$	$1.07ns$
Dynamic energy per operation	$8.52 \times 10^{-14}J$	$8.72 \times 10^{-14}J$	$9.00 \times 10^{-14}J$	$10.26 \times 10^{-14}J$	$12.56 \times 10^{-14}J$
Leakage power	$1.40 \times 10^{-8}W$	$1.87 \times 10^{-8}W$	$2.57 \times 10^{-8}W$	$4.41 \times 10^{-8}W$	$12.54 \times 10^{-8}W$

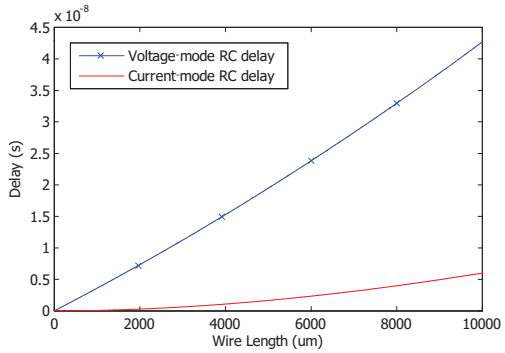

Figure 4: The difference between voltage-mode and current-mode signal delay models.

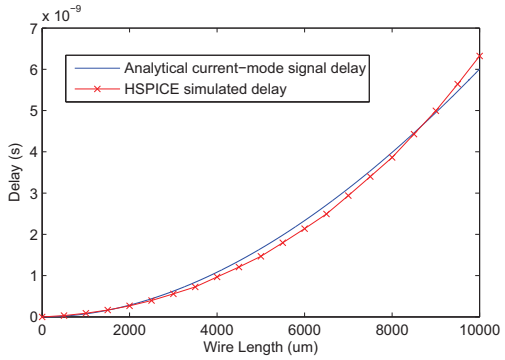

Figure 5: The current-mode signal delay model verification comparing to HSPICE simulations.

Using Seevinck's delay expression, the voltage-mode and the current-mode delays are given by Equation (1) and (2):

$$\delta t_v = \frac{R_T C_T}{2} \cdot \left(1 + \frac{2R_B}{R_T}\right) \quad (1)$$

$$\delta t_i = \frac{R_T C_T}{2} \cdot \left(\frac{R_B + \frac{R_T}{3}}{R_B + R_T}\right) \quad (2)$$

In these expressions, C_T and R_T are the total line capacitance and resistance, respectively, R_B is the pull-down resistance of the PCRAM cell, and δt_v and δt_i are the RC delays of voltage-mode signals and current-mode ones, respectively.

It can be seen from Equation (1) and (2) that, when $R_B > R_T$, which is usually the case for memory arrays, the voltage-mode delay is considerably larger than the intrinsic line delay $R_T \cdot C_T/2$ while the current-mode delay is much smaller. Fig. 4 demonstrates the estimated bitline delay by using these two signal delay models. It shows that the current-mode bitline delay model can have an almost $10X$ delay reduction when the wire length is very long. In PCRAMsim, the current-mode RC delay expression, Equation (2), is used to model the bitline delay.

The bitline delay analytical model is verified by comparing it with the HSPICE simulation result. As shown in Fig. 5, the delay derived by our analytical RC model is consistent with the HSPICE simulation results.

The original CACTI bitline power model is also modified in *PCRAMsim*. Because the voltage of bitline is ideally clamped to a fixed level, the negligible voltage swing on bitlines nearly eliminates the dynamic power dissipation. In our bitline power model, a $10mV$ voltage swing is assumed to take into account the load effect of PCRAM cells. The Joule heat dissipated on bitlines and PCRAM cells is also included in the total dynamic power of the PCRAM read operations.

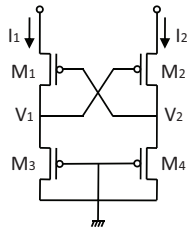

Figure 6: The current-voltage converter modeled in PCRAMsim.

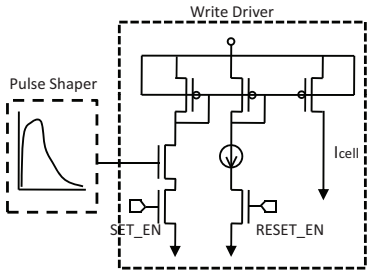

Figure 7: The slow quench pulse shaper used in [9].

3.2.2 *Current-Voltage Converter Model*

As shown in Fig. 3, the current-voltage converter in our current-mode sensing scheme is actually the first-level sense amplifier, and the CACTI-modeled voltage sense amplifier is still kept in the bitline model as the final stage of the sensing scheme. The current-voltage converter senses the current difference $I_1 - I_2$ and then it is converted into a voltage difference $V_1 - V_2$. The required voltage difference produced by current-voltage converter is set to $80mV$, which is the minimum sensible voltage difference of the CACTI-modeled voltage sense amplifier. We use the current-voltage converter design from [19] and the circuit schematic is shown in Fig. 6. This sensing scheme is similar to the hybrid-I/O approach described in [13], which can achieve high-speed, robust sensing, and low power operation.

To avoid unnecessary calculation, the current-voltage converter is modeled by directly using the HSPICE-simulated values and building a look-up table of delay, dynamic energy, and leakage power (Table 5).

3.3 Slow Quench Shaper Model

Due to the different data recording mechanisms used in PCRAM and SRAM/DRAM, PCRAM needs specialized circuits to handle its operations. As mentioned in Section 2, the RESET and SET operation of PCRAM cells needs specific pulse shapes to heat up the GST quickly and to cool it down gradually, especially for SET operations. This pulse shaping requirement is achieved by using a slow quench pulse shaper. As shown in Fig. 7, the slow quench pulse shaper is composed of an arbitrary slow-quench waveform generator and a write driver.

In our PCRAM model, the delay and power impacts of the slow quench shaper are neglected because they are already included in the assumptions made for RESET/SET operations, where the RESET and SET latency are assumed to be $40ns$ and $100ns$ by default, as listed in Table 4. The RESET and SET energy consumption can be further calculated based on our estimated RESET/SET required current, as listed in Table 2. Additionally, the energy efficiency during the RESET/SET operation is assumed to be 35% according to [5]. The area of slow quench shapers is modeled by measuring the die photos of [5] and [9].

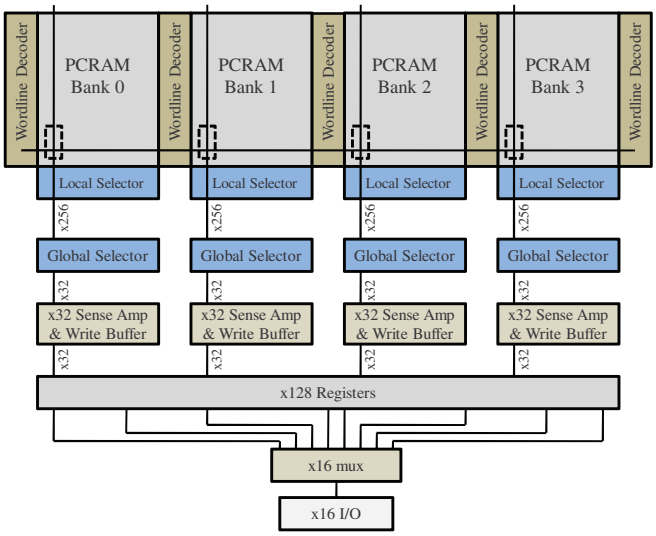

Figure 8: An example of the array organized in PCRAMsim main memory organization.

3.4 PCRAM Organization and Timing Model

In *PCRAMsim*, the PCRAM organization can be configured to support either cache or main memory applications. The memory-style organization is discussed in [20] in detail. The major difference is that a main memory chip usually has a limited pin count, while the on-chip cache module does not. As a result, the output width of a main memory chip is much smaller than the internal block size and prefetching schemes are widely used in the main memory design. As the example shown in Fig. 8, the internal access width is 128 bits. For each read operation, the data is first stored into a set of output registers, and then a 16-bit word is put on pins in the 8X burst mode; for each write operation, the data is first buffered into registers and then the corresponding RESET or SET pulses are generated to record the data into PCRAM cells.

According to this memory organization, the PCRAM timing parameters are defined as follows:

- Read Latency: the delay required to move the data from the memory array to register.
- Write Latency: the delay required to finish a SET operation (because SET takes a longer time than RESET).

In *PCRAMsim*, the original cache-style organization remains the same as the one modeled in CACTI.

4. VALIDATION

The PCRAM model is validated against three PCRAM prototype designs [7], [1], and [9] in terms of area, latency, and energy. We use the information from real chip design specifications to set the input parameters required by *PCRAMsim*, such as capacity, line size, technology node, Ndwl, Ndbl[3], etc.

In order to validate our PCRAM modeling with MOSFET-selected option, the results produced by *PCRAMsim* are compared against a $256Mb$ MOSFET-selected PCRAM chip with $0.1\mu m$ feature size, $50ns$ RESET pulse, and $300ns$ SET pulse [7] and a $64Mb$ chip with $0.12\mu m$ feature size, $10ns$ RESET pulse, and $150ns$ SET pulse [1]. To mimic the array organization in [7] and [1], the number of Ndwl and Ndbl are set to be 4 and 2, respectively. The validation results are listed in Table 6 and Table 7.

In addition, our model is validated against another PCRAM design with diode-selected option, which is a $90nm$ $512Mb$ PCRAM prototype chip with RESET and SET pulse durations of $100ns$ and $400ns$, respectively [9]. Instead of using the PCRAM cell size estimation in Table 3, the diode-selected PCRAM cell size is estimated to be $5.6F^2$. Ndwl and Ndbl are fixed to be 8 and 2, respectively, as these are the values observed from the die photo of their prototype chip [9]. Our tool estimates the write energy with the assumption that the frequencies of RESET and SET operations are the same. The timing and energy model validation results are listed in Table 8.

[3]Ndwl is the number of segments in a wordline; Ndbl is the number of segments in a bitline.

Table 6: Results of PCRAMsim model validation with respect to a $0.1\mu m$ $256Mb$ MOSFET-selected PCRAM prototype chip [7]

Metric	Actual Value	PCRAMsim Projection / Error
Area	$79.2mm^2$	$82.40mm^2$ / $+4.04\%$
Read Latency	$62ns$	$50.87ns$ / -17.95%
Write Latency	$< 500ns$	$323.32ns$ / -

Table 7: Results of PCRAMsim model validation with respect to a $0.12\mu m$ $64Mb$ MOSFET-selected PCRAM prototype chip [1]

Metric	Actual Value	PCRAMsim Projection / Error
Area	$64mm^2$	$55.98mm^2$ / -12.53%
Read Latency	$70ns$	$63.00ns$ / -10.00%
Write Latency	$> 180ns$	$213.00ns$ / -

Table 8: Results of PCRAMsim model validation with respect to a $90nm$ $512Mb$ diode-selected PCRAM prototype chip [9]

Metric	Actual Value	PCRAMsim Projection / Error
Area	$91.5mm^2$	$93.04mm^2$ / $+1.68\%$
Read Latency	$78ns$	$59.76ns$ / -23.40%
Write Latency	$430ns$	$438.55ns$ / $+1.99\%$
Write Energy	$54nJ$	$47.22nJ$ / -12.56%

By observing the comparison between the actual PCRAM parameters and the *PCRAMsim*-generated estimations, we find a 12% to 23% underestimation of PCRAM read latency. First, note these model errors are within the range of chip variation of this emerging technology. We speculate that these differences with the quoted timing parameters originated from the difference between the actual device-level silicon properties and the modeled gate behaviors referred from the conventional CACTI tools. It is also possible that other sources of errors come from the difference between the peripheral circuitry fabricated in the real designs and the one modeled in *PCRAMsim*, which can cause errors in area, latency, and power estimations of sense amplifiers, decoders, and multiplexers. Given the generic nature of PCRAMsim as a system-level design exploration tool and the variation of specific fabrication processes from the ITRS roadmap [6], we consider such validation errors to be reasonable (note that the range of the validation errors are similar to the previous versions of CACTI tools [20, 22]).

5. CASE STUDY ON USING PCRAMSIM

In this section, we conduct two case studies to demonstrate how to use the *PCRAMsim* tool to estimate and optimize the PCRAM memory design and the PCRAM cache design, respectively.

5.1 Embedded EEPROM

One of the promising applications of PCRAM is to replace NAND flash. NAND is widely used as firmware storage or disk in embedded systems. However, NAND flash has limitations: it cannot be accessed randomly because of its page-accessible structure, and

Table 9: Using PCRAM as direct replacement of NAND

A typical $512Mb$ NAND (Source: K9F1208X0C datasheet)	
Access unit	page
Read latency	$15\mu s$
Write latency	$200\mu s$
Erase latency	$2ms$
A $512Mb$ PCRAM (Source: [9], Table 8 for more details)	
Access unit	byte
Read latency	$78ns$ ($59.76ns$, PCRAMsim estimation)
Write latency	$430ns$ ($438.55ns$, PCRAMsim estimation)
A typical $512Mb$ DRAM (Source: K4T51043Q datasheet)	
Access unit	byte
tRCD	$15ns$
tRP	$15ns$

Table 10: New PCRAM parameters after using *PCRAMsim* for speed optimization

Parameter	Before optimization	After optimization
Area	$93.04mm^2$	$102.34mm^2$
Read Latency	$59.76ns$	$16.23ns$
Write Latency	$438.55ns$	$416.23ns$
Ndwl	8	8
Ndbl	2	8

thus it has poor random read access compared to DRAM. As a result, program codes stored in NAND must be copied to random-accessible memory like DRAM before execution.

One can replace NAND flash with PCRAM, with the following two advantages:

- The byte-accessibility of PCRAM[4] makes DRAM shadow mapping unnecessary. Using PCRAM as EEPROM solely instead of NAND+DRAM can eliminate the need of a DRAM module in the system.
- The removal of the shadow mapping in RAM can reduce the system leakage power. As a non-volatile memory, PCRAM does not consume any standby leakage power. DRAM needs refresh power to maintain the data even if the memory is not accessed.

However, replacing the conventional NAND+DRAM architecture with PCRAM without optimization may result in a performance degradation. As shown in Table 9, the PCRAM prototype has a much slower read/write latency than the DRAM memory.

To overcome this obstacle, the PCRAM chip needs to be redesigned for speed optimization at the expense of area efficiency by aggressively cutting wordlines and bitlines or inserting repeaters. Our *PCRAMsim* shows its usefulness in enabling these types of design space trade-offs.

Table 10 shows the impact of the speed optimization performed by *PCRAMsim*. It shows that the PCRAM read latency after speed optimization is about 3.68X faster than the speed before optimization. The new read latency, $16.23ns$, is already very close to the DRAM read/write latency ($15ns$, shown in Table 9), so the performance degradation is greatly alleviated. Although the PCRAM write latency doesn't reduce too much due to the inherit SET/RESET pulse duration, write latency is typically not in the critical path and can be tolerated using buffers. As a result, the optimized PCRAM chip projected by *PCRAMsim* can properly replace the traditional *NAND plus DRAM* structure in the embedded system design. The speed optimization is at the expense of increasing chip area, which rises from $93.04mm^2$ to $102.34mm^2$. The example shows that *PCRAMsim* explores the entire design space and finds the speed-optimized values of $Ndwl \times Ndbl$ to be 8×8, rather than the original value of 8×2, which is used in the prototype chip with smaller area.

In this case study, we demonstrated how *PCRAMsim* can optimize the PCRAM design so that PCRAM replacements for *NAND plus DRAM* structures become feasible. In this example, the replacement can be translated into a power saving of $72mW$ ($512Mb$ DRAM background power [12]). If a PCRAM replacement for an entire DRAM of $128MB$ (the RAM capacity of iPhone) is assumed, a $0.14W$ power reduction can be expected (Note that the total power consumption of an iPhone ranges from $0.5W$ to 1W [2]).

5.2 Memory and Cache Optimization

Although the current PCRAM technology has a limited write endurance of around 10^8 [8], it has been predicted by ITRS [6] that the PCRAM write endurance will be improved to more than 10^{12} in the near future. Consequently, there has been a recent trend to study PCRAM-based cache [23] and PCRAM-based memory [8, 17], with techniques to mitigate the endurance issues.

Our *PCRAMsim* tool can also be used in such studies to explore different design options, and find the optimized PCRAM-based embedded memory organization and last-level cache designs, depending on the optimization goals (such as speed, area, or leakage optimizations). Table 11 shows how *PCRAMsim* explores the design space for a $32nm$ $16MB$ last-level PCRAM cache design with different optimization targets. The estimation results for SRAM, eDRAM, and DRAM are generated by CACTI-5.3 cache model [20] with its default optimization setting. The *PCRAMsim* tool is used to optimize the PCRAM design by targeting different goals, such as read latency, chip area, and leakage power minimization. With *PCRAMsim* leakage optimizations and using the low leakage device model, the active leakage power of this PCRAM design can be reduced under $45mW$ with a performance penalty of 109%. Combined with the zero standby leakage property of PCRAM, the leakage-optimized PCRAM design has significant power savings compared to its SRAM, eDRAM, and DRAM counterparts. In addition, if the design target is for speed or area optimization, *PCRAMsim* would suggest a different design option (such as different Ndwl and Ndbl or different types of transistors) in designing fast-read or high-density PCRAM modules.

6. CONCLUSION

Phase-change random access memory (PCRAM) is an emerging memory technology for future universal memories, due to its non-volatility, fast speed, zero standby power, and high density. The versatility of the upcoming PCRAM technology makes it possible to use PCRAM at other levels in the memory hierarchy, such as execute in place (XIP) memory, main memory or even on-chip cache. PCRAM design options can vary for different applications by tuning circuit structure parameters such as Ndwl, Ndbl, or by using devices and interconnects with different properties. To enable the system-level design space exploration of PCRAM, we developed a PCRAM performance, energy, and area model called *PCRAMsim*, which is validated against industrial PCRAM prototypes. The new *PCRAMsim* tool is expected to help evaluation and design exploration for system-level research determine how to best leverage this emerging technology for performance and power improvements.

7. REFERENCES

[1] S. J. Ahn, Y. J. Song, C. W. Jeong, J. M. Shin, Y. Fai, et al. Highly Manufacturable High Density Phase Change Memory

[4]NOR flash is also byte-accessible. However, NOR has an erase-before-write problem. The typical time to erase a block in NOR flash is about 1 or 2 seconds. Therefore, we don't consider NOR replacement here.

Table 11: 32-nm 16MB 8-way L3 caches (with different PCRAM design optimizations)

	SRAM	eDRAM	DRAM	PCRAM (speed opt)	PCRAM (area opt)	PCRAM (leakage opt)
Area (mm^2)	37.4	11.9	4.19	3.77	**2.49**	3.74
Read latency (ns)	3.43	2.21	2.25	**2.54**	4.55	5.33
Write latency (ns)	$3.43ns$	2.21	2.25	RESET: 42.54 SET: 102.54	RESET: 44.55 SET: 104.55	RESET: 45.33 SET: 105.33
Read energy per access (nJ)	0.83	0.57	0.61	0.73	0.75	0.70
Write energy per access (nJ)	0.75	0.57	0.61	RESET: 6.66 SET: 2.28	RESET: 13.13 SET: 4.37	RESET: 13.09 SET: 4.33
Leakage power (mW)	10,678	885	1,924	194	96	**44**

Note: The SRAM, eDRAM, and DRAM estimations are generated by CACTI-5.3 cache model [20] with its default optimization settings. The PCRAM estimations and optimizations are all generated by *PCRAMsim* with different optimization goals as listed in the parenthesis.

of 64Mb and Beyond. In *IEDM '04. Proceedings of the 2004 IEEE International Electron Devices Meeting*, pages 907–910, 2004.

[2] Apple. iPhone Technical Specification. `http://www.apple.com/iphone/specs.html`.

[3] G. W. Burr, B. N. Kurdi, J. C. Scott, C. H. Lam, K. Gopalakrishnan, et al. Overview of Candidate Device Technologies for Storage-Class Memory. *IBM Journal of Research and Development*, 52(4/5), 2008.

[4] R. J. Evans and P. D. Franzon. Energy Consumption Modeling and Optimization for SRAM's. *IEEE Journal of Solid-State Circuits*, 30(5):571–579, 1995.

[5] S. Hanzawa, N. Kitai, K. Osada, A. Kotabe, Y. Matsui, et al. A 512kB Embedded Phase Change Memory with 416kB/s Write Throughput at 100μA Cell Write Current. In *ISSCC 2007. Proceedings of the 2007 IEEE International Solid-State Circuits Conference*, pages 474–616, 2007.

[6] International Technology Roadmap for Semiconductors. Process Integration, Devices, and Structures 2007 Edition. `http://www.itrs.net/`.

[7] S. Kang, W.-Y. Cho, B.-H. Cho, K.-J. Lee, C.-S. Lee, et al. A 0.1μm 1.8V 256Mb 66MHz Synchronous Burst PRAM. In *ISSCC 2006. Proceedings of the 2006 IEEE International Solid-State Circuits Conference*, pages 487–496, 2006.

[8] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger. Architecting Phase Change Memory as a Scalable DRAM Alternative. In *ISCA '09. Proceedings of the 36th International Symposium on Computer Architecture*, pages 2–13.

[9] K.-J. Lee, B.-H. Cho, W.-Y. Cho, S.-B. Kang, B.-G. Choi, et al. A 90 nm 1.8 V 512 Mb Diode-Switch PRAM With 266 MB/s Read Throughput. *IEEE Journal of Solid-State Circuits*, 43(1):150–162, 2008.

[10] M. Mamidipaka and N. Dutt. eCACTI: An Enhanced Power Estimation Model for On-chip Caches. Technical Report TR04-28, Center for Embedded Computer Systems, 2004.

[11] P. Mangalagiri, K. Sarpatwari, A. Yanamandra, V. Narayanan, Y. Xie, et al. A Low-Power Phase Change Memory Based Hybrid Cache Architecture. In *GLSVLSI '08. Proceedings of the 18th ACM Great Lakes Symposium on VLSI*, pages 395–398, 2008.

[12] Micron. System Power Calculator. `http://www.micron.com/support/part_info/powercalc.aspx`.

[13] Y. Moon, Y.-H. Cho, H.-B. Lee, B.-H. Jeong, S.-H. Hyun, et al. 1.2V 1.6Gb/s 56nm 6F2 4Gb DDR3 SDRAM with Hybrid-I/O Sense Amplifier and Segmented Sub-Array Architecture. In *ISSCC 2009. Proceedings of the 2009 IEEE International Solid-State Circuits Conference*, pages 128–129, 2009.

[14] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi. Architecting Efficient Interconnects for Large Caches with CACTI 6.0. *IEEE Micro*, 28(1):69–79, 2008.

[15] F. Pellizzer, A. Pirovano, F. Ottogalli, M. Magistretti, M. Scaravaggi, et al. Novel μTrench Phase-Change Memory Cell for Embedded and Stand-Alone Non-Volatile Memory Applications. In *Proceedings of the 2004 Symposium on VLSI Technology*, pages 18–19, 2004.

[16] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, et al. Scaling Analysis of Phase-Change Memory Technology. In *IEDM '03. Proceedings of the 2003 IEEE International Electron Devices Meeting*, pages 29.6.1–29.6.4, 2003.

[17] M. K. Qureshi, V. Srinivasan, and J. A. Rivers. Scalable High Performance Main Memory System Using Phase-Change Memory Technology. In *ISCA '09. Proceedings of the 36th International Symposium on Computer Architecture*, pages 24–33.

[18] S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y.-C. Chen, et al. Phase-Change Random Access Memory: A Scalable Technology. *IBM Journal of Research and Development*, 52(4/5), 2008.

[19] E. Seevinck, P. J. van Beers, and H. Ontrop. Current-Mode Techniques for High-Speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAM's. *IEEE Journal of Solid-State Circuits*, 26(4):525–536, 1991.

[20] S. Thoziyoor, N. Muralimanohar, J.-H. Ahn, and N. P. Jouppi. CACTI 5.1 Technical Report. Technical Report HPL-2008-20, HP Labs, 2008.

[21] D. Wang, B. Ganesh, N. Tuaycharoen, K. Baynes, A. Jaleel, et al. DRAMsim: A Memory-System Simulator. *SIGARCH Computer Architecture News*, 33(4):100–107, 2005.

[22] S. J. E. Wilton and N. P. Jouppi. CACTI: An Enhanced Cache Access and Cycle Time Model. *IEEE Journal of Solid-State Circuits*, 31:677–688, 1996.

[23] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie. Hybrid Cache Architecture with Disparate Memory Technologies. In *ISCA '09. Proceedings of the 36th International Symposium on Computer Architecture*, pages 34–45.

[24] N. Yamada, E. Ohno, K. Nishiuchi, and N. Akahira. Rapid Phase Transitions of GeTe-Sb_2Te_3 Pseudobinary Amorphous Thin Films for an Optical Disk Memory. *Journal of Applied Physics*, 69(5):2849–2856, 1991.

[25] W. Zhao and Y. Cao. New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration. *IEEE Transactions on Electron Devices*, 53(11):2816–2823, 2006.