Architecting Phase Change Memory as a Scalable DRAM Alternative

CSCI 311 (Computer Architecture)

Nile University

Presented by:

Sandra Ezzat Rezk, 19100502

Abdelrahman Nasser, 18192267

Ahmed Ibrahim, 18101650

Hussein Elshazly Eida, 19106038

Presented to:

Dr. Eman Kamel Gawish

Contents

[II. Literature Review 3](#_Toc61647220)

[1 Introduction 3](#_Toc61647221)

[1.1 Flash Memory & DRAM 3](#_Toc61647222)

[1.2 Phase change memory (PCM) 4](#_Toc61647223)

[1.3 PCM’s disadvantages 4](#_Toc61647224)

[2 PCM Technology 5](#_Toc61647225)

[2.1 write 5](#_Toc61647226)

[2.2 Memory Cells 6](#_Toc61647227)

[2.3 Write endurance 8](#_Toc61647228)

[2.4 Reads 9](#_Toc61647229)

[2.5 Process Scaling 9](#_Toc61647230)

[2.6 Array Architecture 10](#_Toc61647231)

[3. BASELINE PCM/DRAM COMPARISON 12](#_Toc61647232)

[3.1 Experimental Methodology 14](#_Toc61647233)

[3.2 Evaluation Baseline 17](#_Toc61647234)

[Methodology 16](#_Toc61647235)

[Implementation and testing 20](#_Toc61647235)

[Code 32](#_Toc61647235)

[Conclusion 36](#_Toc61647235)

[References 38](#_Toc61647236)

# II. Literature Review

## 1 Introduction

[Phase change memory](https://www.zurich.ibm.com/sto/memory/) (PCM) is an emerging non-volatile memory technology that could play a key role in future computing systems, Memory technology scaling drives higher density, higher power, and lower price-performance ratios. Memory scaling, the first-order technology target, is at risk for traditional technologies. Storage mechanisms in prevalent memory technologies include fundamentally non-scalable load positioning and control.

### 1.1 Flash Memory & DRAM

In a non-volatile space, Flash memory must precisely monitor the discrete load imposed on the floating gate. In the volatile primary memory, the DRAM must not only place the charge in the storage capacitor but must also place the charge in the storage capacitor. It also mitigates the leakage of sub-threshold charges through the access system. The space must be high enough to store the fee, Also, Capacitors must be large enough to store the charge for accurate sensing and the transistors must be large enough to exercise efficient control over the tube. In view of these obstacles, processing solutions for DRAM scaling above 40 nm technology remain uncertain.

### 1.2 Phase change memory (PCM)

The Phase Change Memory (PCM) has a non-volatile storage device that can be used for scaling. During writing, the access transistor injects current into the storage material and thermally causes phase shift, which is observed during readings. PCM, based on analogue current and thermal effects, does not require power of isolated electrons. As technology scale and heating contact areas shrink, existing measurements are configured linearly. This PCM scaling process has been demonstrated in a 20nm system prototype and is expected to be scaled to 9nm. As a flexible alternative to DRAM, PCM may have a simple road map to increase key memory density and capability.

### 1.3 PCM’s disadvantages

Present prototypes are not planned to minimize PCM latencies, energy costs and finite durability. This paper rethinks the design of the PCM subsystem to put the technology into the competitive context of DRAM. Since the region translates directly into memory processing costs, we ensure that the suggested solutions are neutral to the field.

Graphical user interface, application

Description automatically generated with medium confidenceTable 1: Technology Survey. \*\* denotes information not available in cited publication. The last column identifies parameters derived for this work.

## 2 PCM Technology

In view of the already speculative state of PCM technology, researchers have taken a variety of various production and design decisions. We looked up for these various attempts to try to give the PCM technology into the competitive context of DRAM.

### 2.1 write

Phase change memory generally operates in two states. The (SET) and (RESET), these are classified as the crystalline (low-resistance) and amorphous (high-resistance) phases of chalcogenide. (Figure 1) shows that the storage element is (RESET) with a high, fast current pulse. The brief pulse suddenly ceases the current surge, rapidly extinguishes the heat generation, and freezes the chalcogenide to an amorphous state. In contrast, the storage element is (SET) by a moderate, long current pulse, which ramps down over the duration of the write. The ramp down slowly cools the chalcogenide and induces crystal formation.

We extract conservative writing latencies and currents, while other prototypes display more extreme parameters. Shorter (SET) latencies of 80 and 100 n are demonstrated for new cell technologies and not for array prototypes. Longer (SET) latencies of 180 to 400 ns emerge from the option of dense yet slow-moving access devices. Chen et al, demonstrates a 90 μA RESET that uses a modern theoretical phase change material.

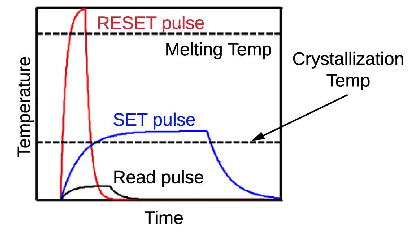


Figure (1)

### 2.2 Memory Cells

The storage factor consists of two electrodes separated by a resistor and a phase change substance, which is usually chalcogenide (Figure 2). Ge2Sb2Te5 (GST) is most widely used, but other chalcogenides give higher resistance and boost electrical characteristics of the system. Nitrogen doping raises resistance and reduces current programming while GS provides lower latency step shifts. We extract parameters for Nitrogen-doped GST due to its widespread use. Phase shifts are caused by the injection of current into the resistor-chalcogenide junction and by the heating of chalcogenide. To 650 a.k.a. The current and voltage properties of the chalcogenide are the same regardless of its source, which decreases the difficulty and latency of the programming. The amplitude and width of the current pulse injected defines the programmed state.

Phase change memory cells are 1T/1R devices, consisting of a resistive storage portion and an entry resistor (Figure 2). Access is usually regulated by one of three devices: a field-effect transistor (FET), a bipolar junction transistor (BJT) or a diode. In the future, FET scaling and significant voltage drops around the cell would adversely affect the durability of gate oxide for unelected wordlines. BJTs are quicker and are supposed to scale more robustly without this vulnerability. Diodes occupy smaller areas and theoretically allow higher cell densities but require higher operating voltages. In view of their compromise between speed and scalability, we draw parameters from BJT access devices.

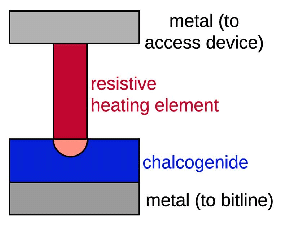


Figure (2)

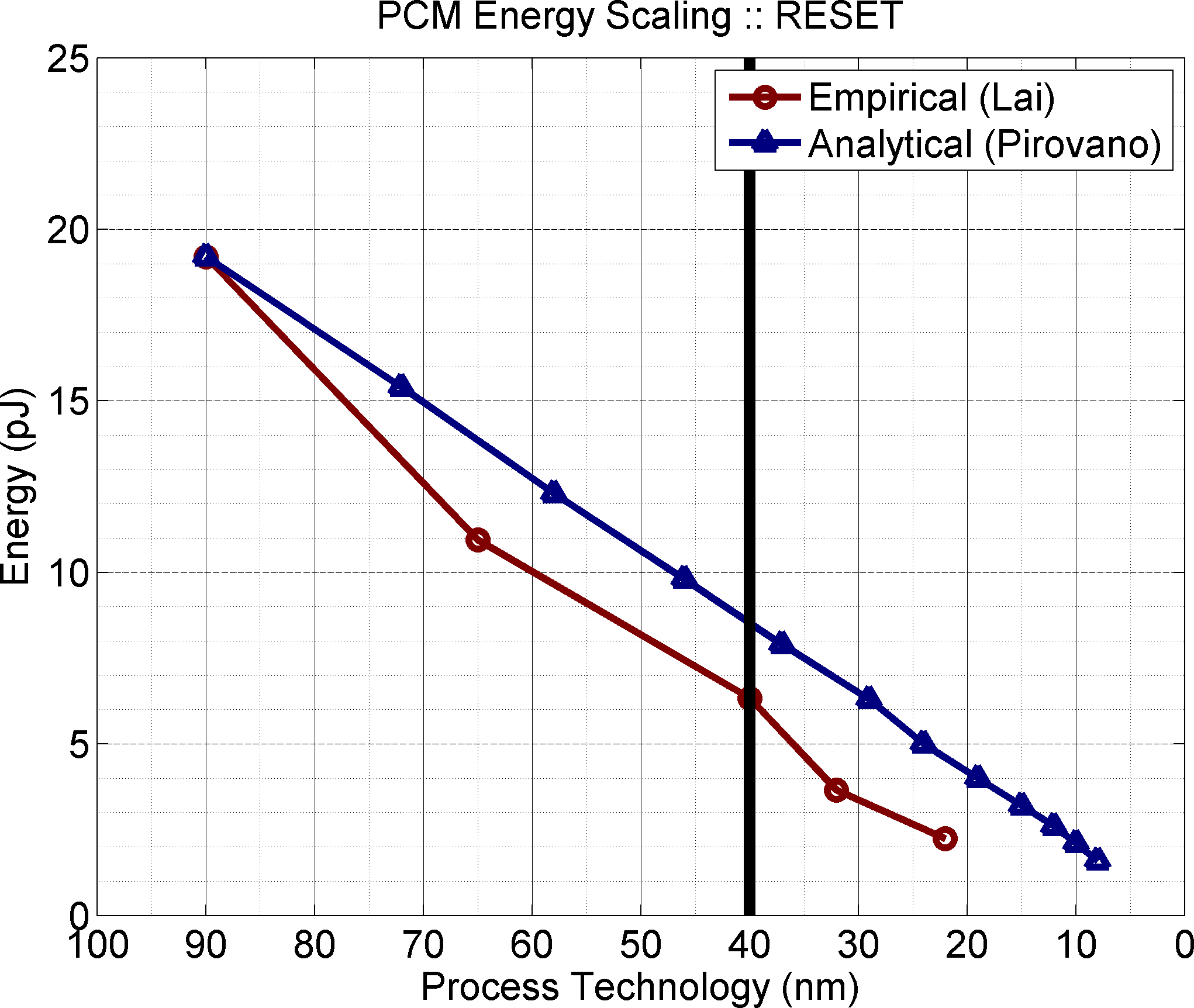
### 2.3 Write endurance

Write stamina, the number of writings done until the cell cannot be accurately configured, varies from 1E+04 to 1E+09. Write durability relies on production processes and varies from one maker to another. 1E+04 is likely to be a particular outlier for speculative, modern cell technologies. 1E+05, the low end of the stamina surveyed, is equivalent to Flash endurance. However, PCM is more likely to show higher write durability by many orders of magnitude (e.g., 1E+07 to 1E+08). The ITRS Roadmap ventures increased the endurance of 1E+12 at 32nm, but we are conservatively modelling 1E+08. We suggest differential writing. This can be paired with the previously suggested Flash memory techniques so that write constraints are not revealed to the device over the lifespan of the memory.

### 2.4 Reads

* The bitline is precharged to the read voltage earlier to perusing the cell. The wordline is energetic less when employing a BJT access transistor. In the event that a chosen cell is in a crystalline state, the bitline is released with current streaming through the storage component and get to transistor. something else, the cell is in an undefined state, anticipating or restricting bitline current. We determine a cell perused inactivity of 48ns.This idleness incorporates bitline precharge and accept BJT get to and current detecting. This same model requires 40 µA of examined current at 1.0V. In this usage, a cell studied disseminates 40 µW for 48ns, consuming approximately 2pJ of vitality. Other models demonstrate higher examined latencies, which run from 55 to 70ns. Yet, these other models actualize FET or diode access devices, which create slower reaction times.

### 2.5 Process Scaling



**Figure 2: Energy Scaling PCM RESET. It is estimated that PCM**

**will scale to 9 nm, while DRAM will scale to 40 nm.**

* PCM scaling diminishes required programming current injected via the electrode-storage contact. As the contact region diminishes with highlight estimate, warm resistivity increments and the volume of stage alter fabric that must be softened to completely square current stream diminishes. These impacts empower littler get to gadgets for current infusion. As shown for vitality in Figure, which are affirmed experimentally in a overview]. Particularly, as include measure scales down by k, contact zone diminishes quadratically 1/k2. Diminished contact area causes resistivity to extend directly (k), that causes programming current to diminish straightly (1/k). These patterns accept set/reset voltage does not scale.Set current is ordinarily 40 to 80 percent of reset current and these streams scale together. Prepare scaling does not affect examined and compose latencies. Compose latencies, in specific, are decided fundamentally by the phase change fabric. Operational issues emerge with forceful PCM technology scaling. As contact zone diminishes, sidelong warm coupling may cause programming streams for one cell to influence the states of adjoining cells. Lai’s study of the state of PCM finds these impacts irrelevant in estimation and simulation. Temperatures drop exponentially with separate from programmed cell, recommending no calculable affect from thermal coupling. Expanding resistivity from littler contact zones may decrease flag quality. However, these flag qualities are well inside the sense circuit capabilities of advanced memory structures.

### 2.6 Array Architecture



**Figure 3: Array Architecture. A hierarchical memory organization includes banks, blocks, and subblocks with local, global decoding for row, column addresses. Sense amplifiers (S/A’s) and word drivers (W/D’s) are multiplexed across blocks.**

* As appeared in Figure 3, stage alter memory cluster structures are comparative to those for existing memory innovations. PCM cells could be progressively organized into banks, pieces, and sub-blocks. Push and column addresses are of- ten decoded at nearby sub-blocks. Environmental circuitry, such as sense intensifiers and. In spite of likenesses to routine memory cluster architectures, PCM-specific plan issues must be addressed. Choice of bitline sense intensifiers influence the studied get to time of the cluster. Voltage sense intensifiers are cross-coupled inverters which require differential releasing of bitline capacitances. In differentiate, current sense enhancers depend on current contrasts to make a differential voltage at the amplifier’s output nodes. In spite of the fact that current detecting is speedier since it does not release bitline parasitic capacitances, these amplifier circuits are bigger [25]. We utilize current detecting to derive 48ns PCM peruses for this work. These bigger sense amplifiers influence PCM range. Within this memory engineering, a push is actuated by reading it from the cluster and locking it in a buffer. Memory gets to examined information from and type in information to the buffer. Accesses that require an unbuffered push must remove the current push and examined the required push. Damaging Measure reads require cluster composes amid each removal to reestablish buffered data. In differentiate, PCM peruses are non-destructive and array writes are required to overhaul the cluster as it were when evicting dirty buffer contents. In Measure, sense enhancers both sense and buffer information utilizing cross-coupled inverters. In differentiate, we investigate PCM architectures with partitioned detecting and buffering; sense intensifiers drive banks of unequivocal locks. These locks provide greater adaptability in push buffer organization by empowering numerous buffered columns. In any case, these hooks cause range overheads, which influence PCM zone.Separate detecting and buffering empowers multiplexed sense amplifiers. Nearby wordline decoders enact lines across multiple sub-blocks. A subset of these sub-blocks’ information pass through neighborhood and worldwide bitline decoders for detecting and buffering. This dispersed bitline interpret empowers buffer widths smaller than the overall number of bitlines. Buffer width may be a basic plan parameter, deciding the desired number of costly current sense speakers.

## 3. BASELINE PCM/DRAM COMPARISON

We express PCM gadget and circuit qualities inside traditional DDR timing and energy boundaries, accordingly, measuring PCM with regards to more natural DRAM boundaries while encouraging an immediate examination.

|  |  |  |
| --- | --- | --- |
|  | **PCM** | **DRAM** |
| Delay & Timing (cy) | | |
| tRCD | 22 | 5 |
| tCL | 5 | 5 |
| tWL | 4 | 4 |
| tCCD | 4 | 4 |
| tWTR | 3 | 3 |
| tWR | 6 | 6 |
| tRTP | 3 | 3 |
| tRP | 60 | 5 |
| tRRDact | 2 | 3 |
| tRRDpre | 11 | 3 |
| Energy (pJ/bit) | | |
| Array read | 2.47 | 1.17 |
| Array write | 16.82 | 0.39 |
| Buffer read | 0.93 | 0.93 |
| Buffer write | 1.02 | 1.02 |
| Background power | 0.08 | 0.08 |

Table [2]: Memory subsystem parameters

### 3.1 Experimental Methodology

A four-core chip multiprocessor is evaluated using the SESC simulator[24]. The 4-way super-scalar, out-of-order cores work at 4.0GHz. This data path is provided by 32KB, direct-map instruction and 32KB, 4-way L1 cache data, which can be accessed in 2 to 3 cycles. A 4MB, 8-way L2cache with 64B lines is shared between the four cores and can be accessed in 32 cycles.

Below the caches is a 400 MHz SDRAM memory subsystem modeled after Micron’s DDR2-800 technical specifications [16]. We consider one channel, one rank, four x16 chips per rank to achieve the standard 8B interface. Structurally, each chip is categorized into four banks to improve throughput as data is interlinked through banks and accessed in parallel. We're designing a burst duration of eight blocks. The memory controller has a queue of 64-entry transactions.

We consider parallel workflows of the SPLASH-2 suite (fft, radix, ocean), SPEC OpenMP suite (art, equake, swim) and NAS parallel benchmarks (cg, is, mg)[3, 4, 27]. Each application is tested to completion. With regard to input sets, we use 1M points for FFT, 514x514 grid for ocean, and 2M integers for radix. SPEC OpenMP workloads run MinneSpec-Large data set and parallel NAS benchmarks run with Class A problem sizes. All programs are compiled using gcc and Fortran compilers at the O3 stage of optimization. Individual applications in each benchmark suite are selected for their memory intensity. We did not consider a benchmark if the system efficiency or energy was not affected by the replacement of DRAM with PCM.

**Delay and timing.** DDR determines its command interface with a set oftime constraints that determine when a command can be given.In Table 2, the DRAM timing parameters are given by the Micronspecifications[16] and the analogue PCM timing parameters are derived from Table 1.

-tRCD defines the delay between the read array and the read/write buffer. This parameter is set to 60ns read latency array, which includes 48ns read (Table 1) and 7.5ns decode row [15]. At 400MHz, the tRCD for PCM is 22 cycles, 4.4x higher than the DRAM value of 5 cycles.

- TCL, tWL, tCCD, and tWTR constrain successive buffer commands and are independent of memory cell technology. tWR,tRTP determines the delay between the read/write buffer commands and the write array of the buffered data. tWR,tRTP maintains data stability in cross-coupled inverters that feed the write array drivers and are independent of the memory cell technology.

- TRP determines the delay between the write array and the read array. As array read begins only after previous buffered data is correctly written back to the array, tRP quantifies array write latency. The longer SET delay of 150ns specifies PCM write latency (Table 1) and tRPis 60 cycles at 400MHz.

- tRRDact,tRRDpre Specify limits on the frequency of accesses to the PCM array to meet power budgets. differentiate between array read (tRRDact) and write (tRRDpre) as reading is non-destructive and writing is needed only when reading removes dirty buffer material. Moreover, given the asymmetrical reading and writing of energy costs, no single timing constraint can fulfil both reading and writing power budgets. PCM reading energy and delay is 2.1x and 4.4x higher than that of DRAM. Since power is energy divided by delay, the reading of the PCM dissipates 0.47x the reading power of the DRAM, which generates 2 cycles of RRDact (0.47x of 3 cycle tRRDin DRAM). Likewise, tRRDpre is 11 cycles written by PCM.

As a result, we estimate PCM read, write intervals are about 4.4x, 12.0x higher than those for DRAM. PCM array readings can occur 2.1x more regularly and array writings may occur 3.6x less regularly than those for DRAM.

**Energy.** The energy costs of DRAM are measured according to the technical notes and specifications of Micron[17]. However, these notes do not specifically distinguish reading and writing energy, because writing must follow any destructive DRAM reading. The current diagram in the technical note shows that array read current is much higher than array write current. From this current diagram, we extract array write and read energy costs, which are 25 and 75 percent of the total 1.56pJper DRAM bit.

Table 1 shows that the PCM array absorbs 2.0 pJ of energy per bit.In addition, we use CACTI to measure energy absorbed by peripheral circuitry to obtain ~2.5 pJ of total array reading energy per bit[18]. The writing array absorbs 13.5pJ or 19.2pJ while writing a zero or one. On average, zeros and zeros are equally likely and writes require 16.35pJ of energy in addition to 0.53pJ of peripheral circuit energy. Thus, we distinguish the read and write energies of the PCM series, which are 2.1x and 43.1x larger than those of the DRAM array.

Reading and writing to buffered data can consume comparable energy costs for PCM and DRAM as buffer access mechanisms are independent of memory technologies. Although there are different power modes for DRAM, in reality, only one power mode is observed when the application is running; there are no possibilities to reach low power modes during simulation. This phase uses 0.08pJ per buffered bit per memory cycle when clocks are allowed and memory is ready for commands[17]. This background energy is used by peripheral circuitry typical to both PCM and DRAM.

### 3.2 Evaluation Baseline

We consider a PCM baseline architecture, which implements DRAM-style buffering with a single 2048B-wide buffer. Figure 4L illustrates end-to-end application performance when PCM replaces DRAM as main memory. Application delay increases with penalties relative to DRAM ranging from1.2x (radix) to 2.2x (ocean, swim). On average, we find a penalty of 1.6x delay. Energy penalties are greater, ranging from 1.4x (cg) to 3.4x (ocean) due to the very costly variety of writings needed when buffer contents are eliminated. On average, we find an energy penalty of 2.2x.

**Figure 4: Application delay and energy when using PCM as a DRAM replacement (L). With non-destructive PCM reads, only a fraction of reads first require a write for dirty data evicted from buffer (R)**

End-to-end delays and energy losses are more moderate than the underlying technical criteria would imply. Even memory-intensive workloads combine memory access with computing. In comparison, the long delay, high-energy array writes are somewhat less manifesting in PCM than in DRAM; non-destructive PCM readings do not require subsequent writings, whereas destructive DRAM readings do. Figure 4R reveals that just 28 percent of the PCM array reads require a dirty buffer array first.

To allow PCM to be used below the lowest-level processor cache in general-purpose systems, the delay and energy gap between PCM and DRAM must be closed. Figure 4 displays non-destructive PCM readings that help minimize underlying delays and energy drawbacks by default. We aim to remove the remaining differences between PCM and DRAM with architectural solutions. In particular, a single 2048B-wide buffer per bank is considered in the baseline analysis. Such broad buffering is cheap in DRAM, but incurs excessive energy costs in PCM due to the costly current injection needed when writing buffer material back to the array.

# Methodology

In order to make PCM much more useful, and comparable to that of DRAM in terms of performance, it becomes essential to mitigate these drawbacks. Several methods have been explored in this regard.

A. Improving Lifetime

The following techniques help to improve the write endurance of a PCM cell and thereby increase the lifetime. The lifetime of a PCM can be viewed as the time from the start (of usage) until the first cell of the PCM starts to wear out.

|  |  |  |  |
| --- | --- | --- | --- |
| Method | | Results | Comments |
| 1. Eliminating Redundant Writes | As a first step in improving the lifetime of PCM, we suggest reducing the write frequency to a single PCM cell. In a typical DRAM write operation, the write update writes the entire row. It has been observed that most of these writes are redundant. A write does not change all the bit values. This means that the redundant writes could possibly be eliminated. | The below figure shows the results of experiments on number of redundant writes, after testing with various memory benchmark programs.  Text  Description automatically generated  The above figure shows that all the benchmark programs exhibit high level of memory write redundancy. Here MLC-2 and MLC-4 stands for Multi Level Cell, where a single PCM cell can hold two values and four values respectively. Removing the redundant writes can be done by implementing a read before a write. The read operations are much faster than write operations in PCM. Therefore, implementing a read before a write takes lesser time than what a complete write operation takes. Therefore, it is very beneficial to do a read before a write. This is illustrated in the next figure. This can be implemented by a simple XNOR gate on the ‘write’ path of the cell.  Text  Description automatically generated | It appears at the first sight that the same technique could be used for DRAM as well. But DRAM doesn’t benefit from this technique because, the read and write operations in DRAM take about the same time and energy. However, in PCM write operations take about 5.x to 10.x times that of a read operation. Therefore, this technique is very useful. |
| 2. Row Shifting  The above method reduces the redundant writes up to 5 times. This results in a life time improvement to ~1.4 to 2.2 years. Yet, this is very short for a main memory. This is because; most of the writes happen locally and therefore, certain hot cells get worn out soon. To avoid this, a technique called Row Shifting is used. | Row shifting mechanism aids in spreading out writes that tend to be localized to a few specific cells of a given row. After certain number of writes to a specific number of cells, new set of cells are chosen to write the data. This process helps to write evenly in all the cells in a row. Experimental results have proven that, shifting a byte at a time improves the performance very well. However, the frequency of shifting also influences the lifetime. For instance, shifting quite often is not preferred because row once shifted, is difficult to be brought back into place and it involves a lot of overhead. This is due to the property of temporal locality. Therefore, the frequency at which shifting is done, is to be carefully chosen. Furthermore, not all pages of the memory are written quite frequently. The pages can be sorted according to the number of times they are accessed. Therefore, the best row shifting algorithm varies from page to page. Page classification is done based on the total write counts on the page and the standard deviation of writes among all lines in a page. | On varying row shift interval from 0 to 256 writes and averaging the resulting lifetime from all the sample pages, it was found that results for various benchmarks varied greatly. On plotting the results, it was observed that the shift interval of 256 writes generates the highest lifetime for all the means for mcf benchmark as shown in the next figure.    The write intervals were not extended for 2 reasons:   1. Geometric and harmonic means have leveled off 2. Increasing the write count will increase the hardware complexity. |  |
| 3. Segment Swapping  The row shifting mechanism only improves the lifetime of each row. However, this technique has to be implemented on a granularity level big enough to be applicable for memory segments such as hot pages that are written quite often. This can be done with a technique called segment swapping. However, the important parameters to be considered here are the size and swap interval of the segments. | The main problem lies in choosing the page size is the metadata that has to be sorted every time in order to determine the cold pages and hot pages. For instance, having a 4GB memory with 4KB page size may require 1MB page counter size. Although this is not a big memory overhead, it requires long latency running times for running through the entire page counter. Therefore it is better to have bigger page sizes. | Different benchmark programs were run with different page sizes with varying swap intervals. The averages of all the results are plotted in a graph with harmonic mean, as in the next figure. It is inferred that a segment size of 1 MB with swap interval of 2X is the most efficient. This is because; the bigger page sizes incur more overhead for the extra writes. For example, the overhead for 1MB, 4MB, and 16MB segments on their base swap intervals are 2.8%, 5.6% and 5.2% respectively. This has been shown in the following figure The swap interval is in terms of base interval ‘X’ because; the swap size should be based on the page size. Larger pages should use larger page swap intervals.  A picture containing application  Description automatically generated |  |
| 4. Partial writes | In case of a main memory system involving PCM, partial write technique reduces the number of writes by tracking the dirty data in the L1 cache. An extra state is added to each cache line which keeps track of stores using fine grained dirty bits. The data is written back to PCM only when the data in the cache is modified or evicted from the cache. | Consequently, number of writes to the PCM, are mitigated. This incurs a small overhead of latches which are used for this implementation. | The partial level writes can be done in two levels of granularity, i.e. cache line size and word size. |

B. Improving latency and Power

PCM suffers from very high access latencies (5-10 times that of a DRAM) in its operation. This limits the performance of the system. During write operation, latency in PCM is mainly attributed to the time taken by the phase change material to undergo a transition in its state i.e. from crystalline to amorphous or vice versa. The ‘GS’ phase change material, offers the best in terms of achieving a lesser latency. These latencies can be hidden or tolerated to a certain extent by bringing about changes at the architectural level.

Reorganization of buffers would be the best approach. Buffers can be reorganized to reduce application execution time from 1.6x to 1.2x, considering that it takes 1.0x in case of DRAM. The buffers are made narrower and arranged in multiple rows. This is done in such a way that the total area remains the same, because area directly translates into cost involved. Multiple rows exploit locality to coalesce writes and hence hides their write latency to a certain extent. We can also suggest using narrow buffers. Narrow buffers also contribute in mitigating the energy. This is because number of sense amplifiers, required decreases linearly, with buffer width. During development of PCM, nitrogen doping helps in increasing the resistivity and lowering programming current. Process scaling also helps in saving energy i.e. as the size of the memory cell scales down, the volume of the material stored inside the cell also decreases. Consequently, lesser amount of material has to undergo a phase transition during the write operation. Decrease in area, also contributes to increase in the resistance value (since resistance is proportional to ratio of length and area). As feature size scales down by a factor say ‘k’, the area decreases by a factor of 1/k², there by leading to an increase in resistance value, by a factor ‘k’. The injection current value decreases by a factor 1/k. This leads to decrease in energy consumption. On the other hand, increasing resistivity by decreasing contact area also has the risk of reducing the signal strength. However, the sense circuitries are capable enough of sensing these signals successfully.

**Implementation and Testing**

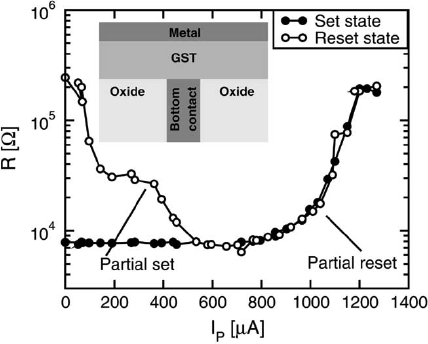
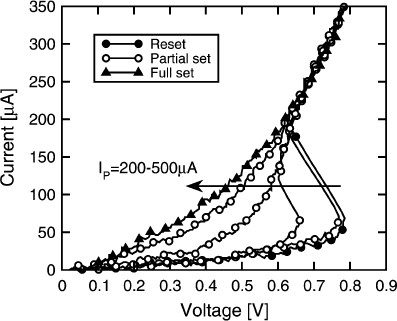
1. **QUICK INTRODUCTION**

In this phase, we have successfully established a compact phase change memory (PCM) model using Verilog-A. As PCM has shown its potential for next generation memory devices, a predictive but easy-to-use circuit model is crucial to growth. Since Verilog-A modelling is versatile and portable for many circuit simulators, the proposed modelling technique can be widely used compared to traditional modelling schemes like DRAM.

**1.2. HOW IT SHOULD WORK**

PCM has two stable phase states: the crystalline phase (Set) and the amorphous phase (Reset). On the figure. 1, presents the difference between the two states: the fixed state is low-resistance (~7 K ohm), and the reset state is high-resistance (~200 K ohm).

The current pulse will alter the phase state. In this example [Figure. 1 and Figure. 2], the current pulse width was held at 100 ns, and if the current pulse amplitude is 700 μA, specified as ISet, the PCM phase will be modified to the crystalline state. In another case, if the current pulse amplitude is more than 1200 μA, known as IReset, the PCM phase will become an amorphous state. When the current programmed pulse falls between zero and ISet or between ISet and IReset, the PCM is in an incomplete (partial) transition. The relationship between the resistance and the current amplitude of the pulse is shown in Figure. 2



**Figure. 1 PCM (I-V curve)** **Figure. 2 PCM (R-I curves)**

1. **MODELING THE PCM MODULE USING VERILOG-A**

Our proposed PCM modelling is based on the HSPICE Verilog-A function. In order to illustrate the modelling principle, we first extract the relevant model parameters from the figure. 1 and Figure 2. The values for the model parameters are shown in the next table. In the next section, we will discuss the model in more detail.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| The Static resistance of Set (RSet) | 7K Ohm |
| The Programming current of Set (ISet) | 600 µA |
| The Programming time of Set (TSet) | 100 ns |
| The Static resistance of Reset (RResett) | 200K Ohm |
| The Programming current of Reset (IReset) | 1200 µA |
| The Programming time of Reset (TReset) | 100 ns |
| The Holding voltage (Vh) | 0.45 V |
| The Threshold voltage (Vth) | 0.78 V |
| The Dynamic-on resistance (Ron) | 1K Ohm |

We present a basic modelling method, accounting for only two stable states, with partial set and partial set conditions ignored. The model consists of the following three modules:

**p**

**PCMR**

**module**

**Iin**

**Decision**

**module**

**n**

**Mout**

**Memory**

**module**

**Qs**

**Qr**

* **PCMR module**
* **Decision module**
* **Memory module**

As in Figure. 3.  **Figure 3**

* 1. **PCMR module**

This module consists of two separate I-V curves (Figure. 4), only one of which can be selected by the "Memory Module" (section 2.2). The PCMR module will then respond to the next decision module, depending on the current pulse width and amplitude.

IReset

Iset

Ron

(Vx,Ix)

RSet

Rsn

(Vth,Ith)

RReset

Vh

**Voltage (V)**

**Figure. 4 Set and reset the I-V curves for the PCMR module.**

**SET Curve:**

This curve blends two states of resistance: R(Set) and R(on). R(Set) represents the resistance of the crystalline state and R(on) represents the resistance (ON) during programming. We set the boundary for I(x). The (V(x), I(x)) is the intersection between the R(Set) and R(on) curves. If the input current pulse amplitude is greater than I(x), the resistance will be set to R(on). The equation for the calculation of R(on) is based on the following relationship:

**V(*I*) = V(h) + *I* × R(on)**

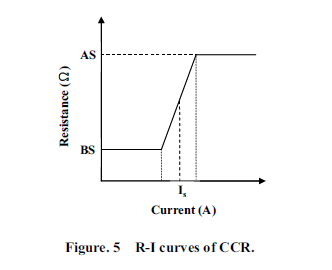
**RESET Curve:**

The reset curve consists of three states of resistance: R(Reset), R(sn) and R. (on). R(Reset) represents the resistance of the amorphous state and R(sn) represents the negative resistance of the snapback. We set two limits at I(th) and I(x) respectively. The R(sn) is determined on the basis of the linear relationship between (V(th), I(th)) and (V(x), I(x)) as:

**V(*I*) = *(I – I(th)) × ()+V(th)***

**3. DECISION MODULE**

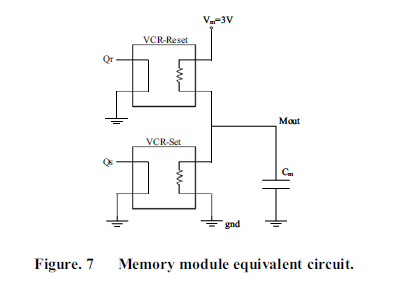
At the point when the choice module gets the memory component current pulse, it checks its width and sufficiency. We use the current control resistance (CCR) to decide the pulse abundancy and utilize a RC circuit to evaluate the pulse width. The RC circuit module gives a straightforward intend to assess the pulse width by means of RC fundamental as expected.



**3.1. MEMORY MODULE**

The memory module should keep up the PCM state after the past PCM module decides the phase state. The same circuit model of the memory module is delineated in (Figure 7).

When Qr is at high voltage, the VCR-Reset resistance becomes low (Rmin), VCR-Set resistance becomes high (Rmax), and Cm is charged to Vm (3 V). Then again, when Qs is at high voltage, the VCR-Set resistance turns out to be low, VCR-Reset opposition turns out to be high, and Cm is released.



# **TESTING THE PCM MODEL WITH REAL DEVICE DATA**

We have developed the PCM model using Verilog-A, parameters which the users can set. The proposed model is verified with HSPICE. Basically, there are two approaches to test the model: DC analysis and transient analysis.

## **4.1 DC ANALYSIS**

The PCM model is screened by sweeping current from 0 A to 1 mA for a 0.1 μA phase. Reset and set states are being checked, respectively. The results of the test (Figure 8), where simulation results are compared to actual system data, showed that while the model does not match data perfectly anywhere, it shows strong alignment with data in low and high current regimes for realistic read and write operations, sequentially.

Graphical user interface, chart

Description automatically generated

Figure 8: DC analysis and comparison.

## **4.2 TRANSIENT ANALYSIS**

We input a write/read current pulse to the PCM, and record the readings of the resistance calculated by the model (Figure 9). The very first programming current pulse is 1200 μA/100 ns for the reset condition. The second one is of 600 μA/100 ns for the set condition. Outside the programming current pulses, the read current is of 0.1μA. The initial state of the PCM model is believed to be a crystal state (7KΩ). The resistance of the PCM model ranges from crystalline to amorphous. The resistance returns to the crystal level after the set programming pulse. When the PCM cell is being programmed by the pulse, the resistance of the model is set to .

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

## **4.3 R-I CURVE COMPARISON**

The relationship between the resistance and the current is shown in Figure 10. The model keeps its initial state until the current pulse goes under 600 µA which is . On the other hand, when the model reaches above 1200 µA ( ), it changes to an amorphous state. The reason behind the variance and inconsistency between data and model is because of the ignored partial set and reset conditions in the model.

Graphical user interface, application

Description automatically generated

Figure 10: The R-I curve comparison with real device data.

## **4.4 PRACTICAL MODEL USAGE AND LIMITATIONS**

The model parameters must be gauged first to achieve model consistency and accuracy. Moreover, capricious settings of parameters could lead to a failure in simulation. For example, in Figure 4, the calculated must be smaller than .

# **5. RESULT**

We have presented and implemented a dense PCM model using Verilog-A. The model is easy to use and understand, but more notably, it can be implemented in any circuit simulator which have the Verilog-A option.

# **6. DRAM**

DRAM (pronounced DEE-RAM): is commonly used as the main memory of the computer. Each DRAM memory cell consists of a transistor and a capacitor in an integrated circuit, and a data bit is stored in the capacitor. Since the transistors often leak a small amount, the condensers would slowly discharge, allowing the information contained in the condenser to drain; thus, every few milliseconds, DRAM must be refreshed (given a new electronic charge) to store the data. The key benefits of DRAM are its simple nature and low cost relative to alternative forms of memory. The key drawbacks of DRAM are its high volatility and high-power consumption compared to other choices.

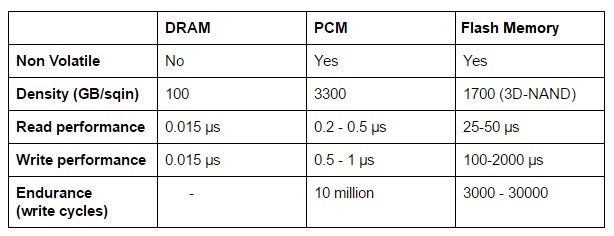
# **6.1. DRAM IMPLEMENTATION**

In order to provide a concrete overview of our proposed solution, we used the specification of the actual commercial high-bandwidth memory (DRAM) portion in **our PCs** and it indicates the following:

|  |  |
| --- | --- |
|  | **DRAM** |
| Non-Volatile | No |
| Density (GB/sqin) | 100 |
| Read performance | 0.015 ps |
| Write performance | 0.015 ps |
| Endurance (write cycles) | - |

**Testing and Verification**

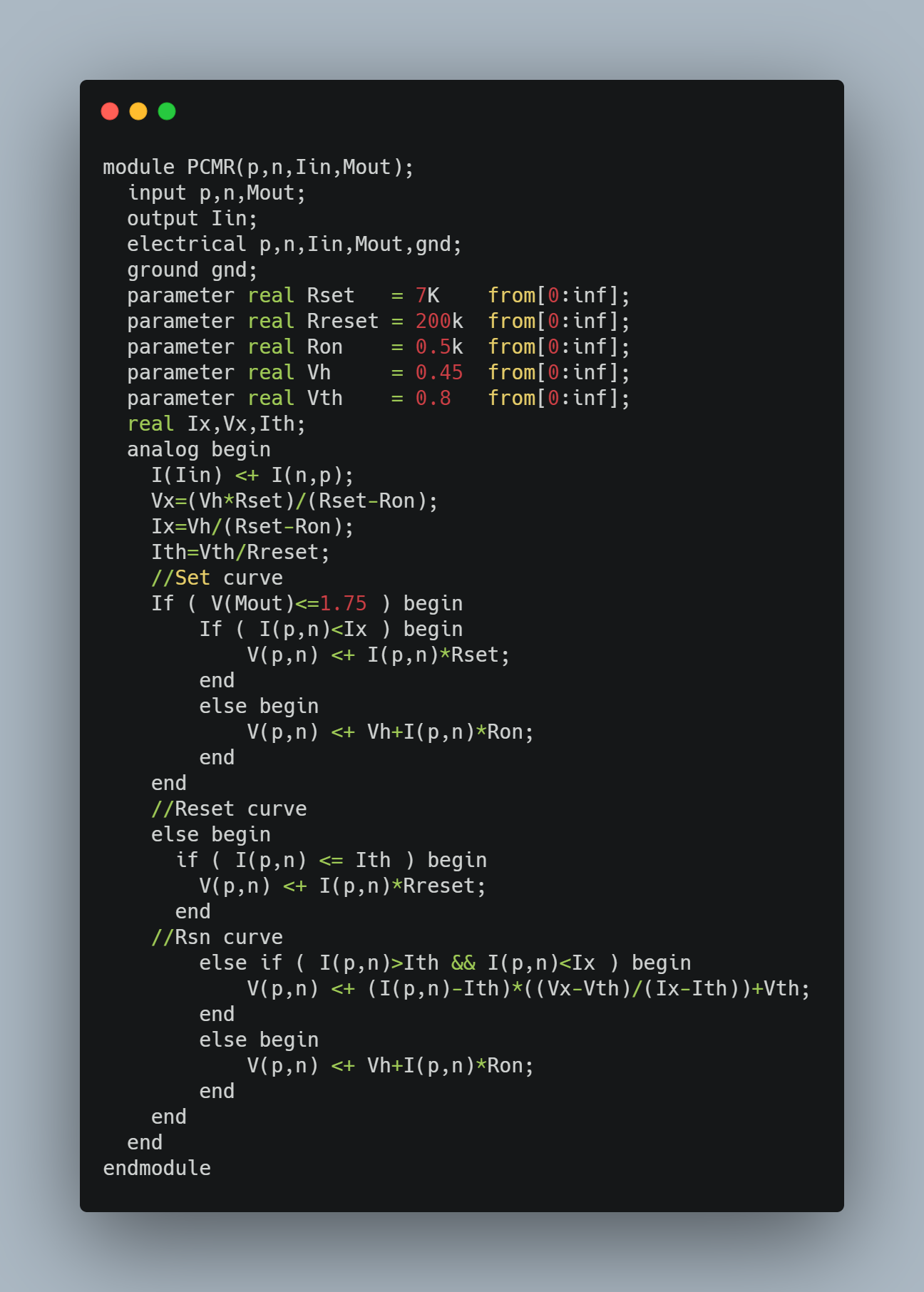
Our testing showed that PCM has a range of advantages over DRAM and Flash Memory. In comparison to DRAM, PCM does not lose data when turned off. It also has a much higher area density than DRAM. Similarly, if we compare PCM with Flash Memory, it beats Flash Memory, both in performance and endurance. In addition, the capacity of PCM to store 3 bits per cell will lower its cost than DRAM and bring it closer to the Flash Memory range.



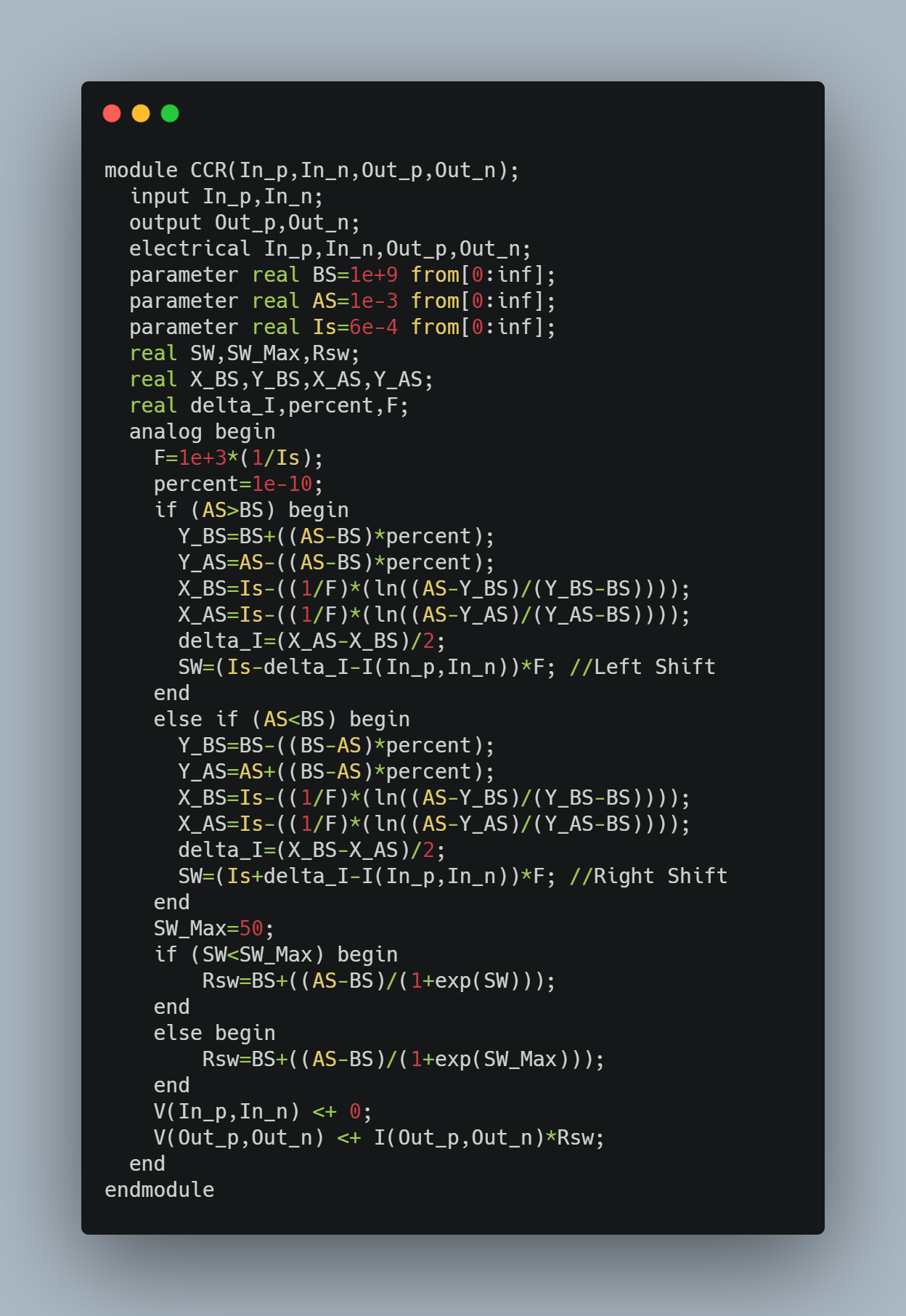
This shows that PCM is potentially the most advanced resistive memory technology in the world today. Materials have been thoroughly researched and mass-produced, for example on DVDs and Blu-Ray discs, and have already appeared on the market as a digital memory product (Intel Optane). Its desirable properties such as multi-level storage, short read/write latency, non-volatility, good cycling endurance and good scalability make it the perfect candidate for applications in modern computing paradigms.

**Code**

**1. PCMR module**



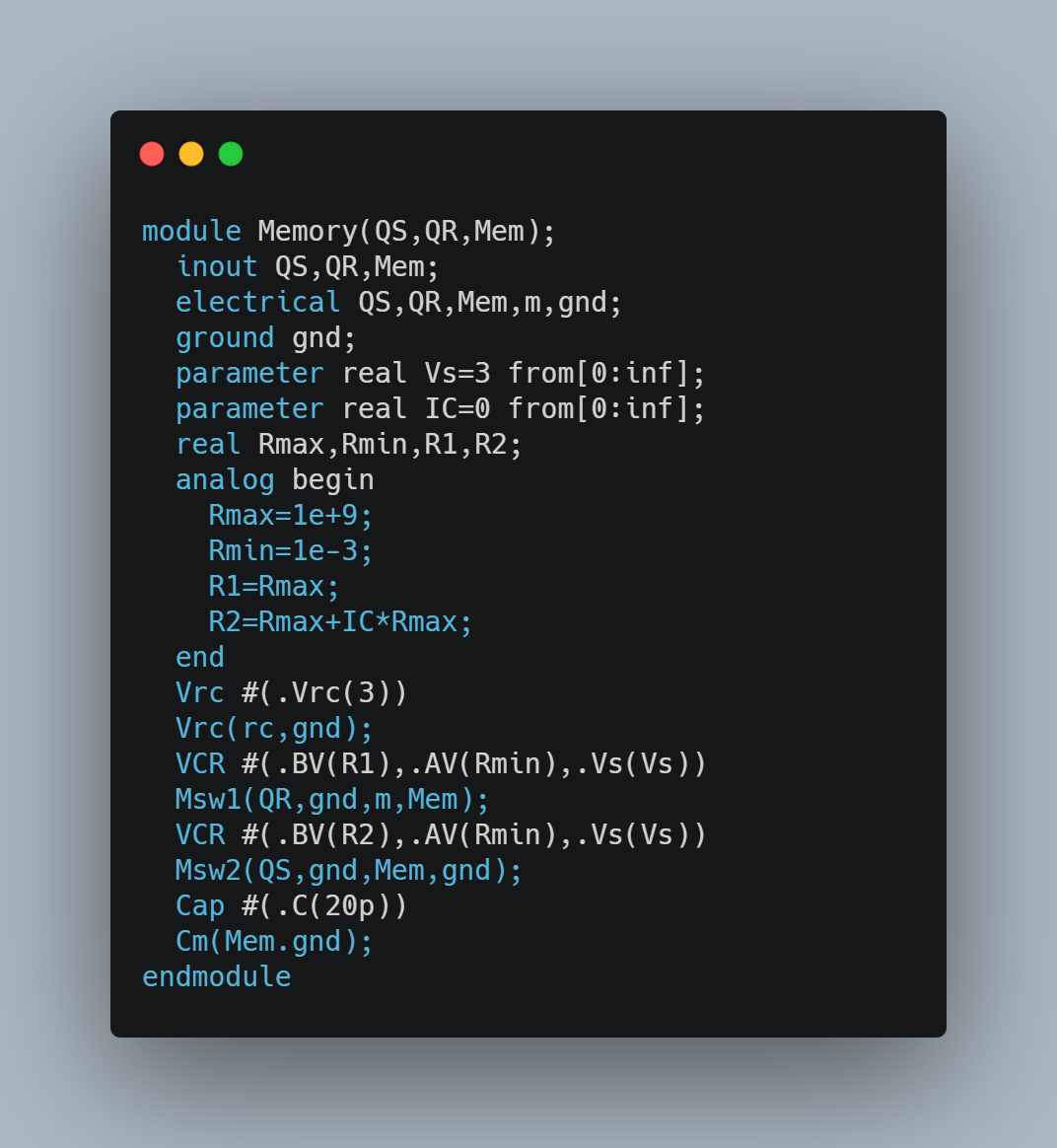
**2.CCR module**

****

**3. RC module**

****

**4. Memory module**



**Final Results**

Phase Change Memory (PCM), a very promising technology, according to IBM, applications range from replacing RAM on modern desktop computers to utilizing hybrid PCM and flash memory to dramatically improving the speed of mobile devices. For example, the cell phone operating system could be stored in PCM, allowing the phone to be launched in a few seconds, the company wrote in a press release. In the business room, entire databases could be stored in PCM for easy query processing for time-critical online applications, such as financial transactions.

There are also cloud-based, artificial intelligence applications that could benefit from PCM. Machine learning algorithms utilizing large datasets can also see a performance increase by reducing overhead latency when reading data between iterations. Compared to flash, which can withstand about 3,000 write cycles, PCM can withstand up to 10 million cycles, making it a potentially industrial-changing technology for years to come.

## Conclusion

We provide a comprehensive survey and variation of the properties of phase change memory to guide architectural studies and upgrades. Architecturally related parameters are presented within the DDR context to enable a DRAM comparison. This comparison finds that long PCM latencies, high energy, and finite resilience can be efficiently mitigated. Efficient buffer organisations and partial writes make PCM competitive with DRAM at existing technology nodes. Moreover, these dynamic, effective solutions are neutral zone, a crucial limitation in memory manufacturing.

The suggested memory architecture lays the basis for leveraging the scalability and non-volatility of PCMs in main memory. PCM scalability ensures lower main memory energy and better write durability. In comparison, non-volatile key memories would radically alter the computing environment.

Software that recognizes this recently founded persistence can have qualitatively new capabilities. For example, the system boot/hibernate will be viewed as instantaneous; the application check point will be inexpensive; the filesystems will have better security guarantees. Thus, the study of this work is a step into a radically new memory hierarchy with significant consequences around the hardware-software interface.

## References

[1]Process integration, devices & structures. International Technology Roadmap for Semiconductors, 2007.

[2] S. Ahn et al. Highly manufacturable high density phase change memory of 64Mb and beyond. InInternational Electron Devices Meeting, 2004.

[3] V. Aslot and R. Eigenmann. Quantitative performance analysis of the SPEC OMPM2001benchmarks.Scientific Programming, 11(2), 2003.

[4] D. Bailey et al. NAS parallel benchmarks. InTechnical Report RNR-94-007, NASA Ames Research Center,March 1994.

[5] M. Bauer et al. A multilevel-cell 32mb flash memory. In International Solid-State Circuits Conference, 1995.

[6] F. Bedeschi et al. An 8Mb demonstrator for high-density 1.8V phase-change memories. In Symposium on VLSI Circuits, 2004

.[7] F. Bedeschi et al. A multi-level-cell bipolar-selected phase-change memory. In International Solid-State Circuits Conference, 2008.

8] Y. Chen et al. Ultra-thin phase-change bridge memory device using GeSb. In International Electron Devices Meeting, 2006

[9] Y. Choi. Under the hood: DRAM architectures: 8F2vs. 6F2.EE Times, February 2008.

[10] R. Hamming. Error detecting and error correcting codes.Bell System Technical Journal, 29(2), April1950.

[11] H. Horii et al. A novel cell technology using N-dopedGeSbTe films for phase change RAM. In Symposiumon VLSI Technology, 2003.

[12] S. Kang et al. A 0.1um 1.8V 256Mb 66MHzsynchronous burst PRAM. In International Solid-State Circuits Conference, 2006.

[13] T. Kgil and T. Mudge. Flash Cache: A NAND flash memory file cache for low power web servers. In International Conference on Compilers on Compilers, Architecture, and Synthesis for Embedded Systems, October 2006.

[14] S. Lai. Current status of the phase change memory and its future. In International Electron Devices Meeting, 2003.

[15] K.-J. Lee et al. A 90nm 1.8V 512Mb diode-switch PRAM with 266 MB/s read throughput .Journal of Solid-State Circuits, 43(1), January 2008.

[16] Micron. 512Mb DDR2 SDRAM component data sheet:MT47H128M4B6-25. Inwww.micron.com, March 2006.

[17] Micron. Technical note TN-47-04: Calculating memory system power for DDR2. In www.micron.com,June 2006