Thermal-aware, heterogeneous materials for improved energy and reliability in 3D PCM architectures

Heba Saadeldeen
eng.hmohsen@gmail.com
University of California, Santa Barbara

Zhaoxia Deng
zhaoxia@cs.ucsb.edu
University of California, Santa Barbara

Timothy Sherwood
sherwood@cs.ucsb.edu
University of California, Santa Barbara

Frederic T. Chong
chong@cs.uchicago.edu
University of Chicago

ABSTRACT

The properties of Phase-Change Memory (PCM) are defined in large part by the different chalcogenide materials employed. As the GeTe and Sb_2Te_3 ratios in the materials are changed, the operating temperatures needed for the phase change are also variable. Motivated by this phenomenon, we study the potential of exploiting different material compositions to achieve different trade-offs among the optimal operating temperatures, energy efficiency, write endurance and write latency. Specifically, we study the trade-offs for energy efficiency and lifetime in the scenario of using PCM materials for all layers of a 3D stack memory. Rather than a "one-memory-fits-all" approach, we propose Heterogeneous 3D PCM architectures by tailoring the Ge-Sb-Te ratios of PCM in concert with both the location and the intended function of these memories within the 3D stack. By varying the material compositions and their operating temperatures in correspondent with the non-uniform heat distribution across the stack, the heterogeneous PCM architectures improve the programming energy by up to 3.5X compared to the best homogeneous configuration. Moreover, the diversity in material compositions can also be exploited to protect error-correcting codes (ECC) by storing them in PCM materials with lower operating temperatures, which drastically reduces ECC early failures and brings a 30% improvement in the lifetime of the entire memory system. This architectural study attempts to make the case for exploring the whole material spectrum and the manufacturing cost associated with that.

CCS CONCEPTS

• **Hardware → Emerging architectures**; **Memory and dense storage**;

KEYWORDS

PCM; 3D stack memory architecture; Thermal-aware, phase change materials

1 INTRODUCTION

As concerns mount about the end of DRAM scaling, researchers are exploring alternative emerging memory technologies such as RRAM, PCM, MEMS/NEMS, etc. Phase-change memory (PCM) is one of the most promising emerging memory technologies suitable for incorporation into main memories. Particularly, due to its high operating temperature, PCM is a good candidate for integration in 3D memory stacking as compared to DRAM [31]. The increase in power density of 3D technology leads to elevated on-chip temperature. This causes DRAM to operate at double (or higher) the current refresh rate as the temperature elevates [6]. PCM operation, on the other hand, is high-temperature friendly due to the way PCM cells are written which requires heating the phase-change material to high temperature thresholds [31].

However, different phase-change temperatures are observed with different chalcogenide materials in PCM. Figure 1 shows the chalcogenide materials that lie along the pseudo-binary line $(GeTe)_x(Sb_2Te_3)_y$ in the ternary system Ge-Sb-Te. Moving down the pseudobinary line from GeTe to Sb_2Te_3 gives PCM materials that exhibit decreasing operating temperatures. Yamada et al. [25] demonstrated the feasibility of varying the chemical composition of the materials with small percentages. The variation of operating temperatures in different materials motivates us to explore the potential trade-offs of using heterogeneous materials in the 3D stack architectures with respect to system properties such as energy efficiency, write endurance and write latency.

In a traditional 3D PCM architecture, when the power density increases, the heat is distributed non-uniformly across all layers in the 3D stack, with layers near the heat sink exhibiting lower ambient temperatures than layers far from it [31]. Due to the "One memory fits all" approach, the PCM material used in a 3D stack is optimized for the worst-case (highest) temperature. This in turns makes the material energy inefficient for memory configurations that operate under lower temperature than the maximum possible (caused by temperature-friendly operation of PCM). So, memory layers close to the heat sink (which operates at lower temperatures) are less energy efficient than those far from it. Instead of using the same PCM material for all layers of the 3D stack, we propose to vary the material compositions in correspondent with the ambient temperatures in different layers. Intuitively, using the same PCM material composition for all layers forces the designer to optimize for the highest peak temperature in a stack (the layer furthest from the heat sink). In a PCM stack composed of *heterogeneous materials*, however, the designer can optimize for lower peak temperatures in layers nearer the heat sink.

Furthermore, for PCM to be considered as a candidate replacement of DRAM in 3D stacking, we have to address its major limitation of limited endurance. The average lifetime of a memory cell is 10^7-10^8 writes [10]. This problem is further exacerbated by the high variability in lifetime across different cells due to process variations [30]. Although many architectural techniques have been proposed to improve the lifetime of emerging memory technologies [4, 27, 29],

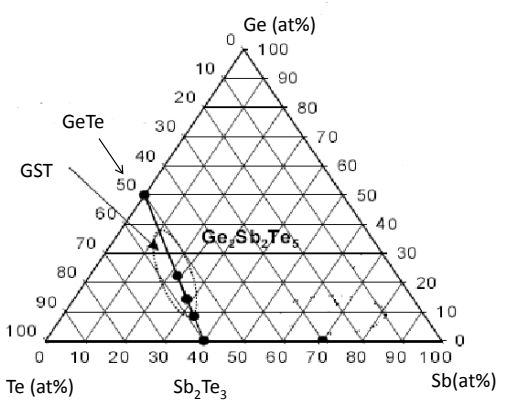


Figure 1: Phase diagram of GeSbTe ternary alloy system

the wearout problem remains a major roadblock to their wide application as the main memory system, yet inspired applications to cache directories [28], branch predictors [19], hardware security architectures [2] and hardware accelerated data structures [3]. In addition, PCM cells are prone to soft errors caused by resistance drift. This kind of error is shown to be problematic for PCM, especially for multi-level cells (MLC) [1] [32]. Ensuring reasonable system lifetime in presence of errors requires that the design must provision large amounts of error correction for PCM lines. Recent studies have proposed write-efficient error-correction schemes such as error-correction pointers (ECP) [20], stuck-at-fault error recovery (SAFER) [21], and pay as you go (PayG) [17] to tolerate a large number of hard faults in memory lines. Fine-grained remapping with ECC and embedded pointers (FREE-p) [26] and wear-out aware page allocation [31] have been proposed as more general techniques that can tolerate both hard and soft errors using strong ECC codes with reduced complexity.

However, due to the high entropy of Hamming-based ECC schemes, they can fail earlier than the data bits they are designed to repair [20]. This problem indicates various lifetime demands for cells storing ECC bits and cells storing data bits. In order to improve the reliability of the entire memory system, we address the ECC early failure problem by storing the ECC bits in a different PCM material than that used to store the data bits. According to the endurance test on PCM memories performed by Kim and Ahn [12], endurance failures are correlated with the reset energy needed for the material melting. Using PCM materials that operate at lower temperature for ECC bits reduces the reset energy, which has the potential to provide higher endurance for these cells.

In this paper, we propose *Heterogeneous 3D PCM architectures* to achieve potential trade-offs between operating temperatures and system properties by varying material compositions in concert with both the location of those memories (where in the memory stack?) and the intended function of these memory blocks (storing ECC bits or storing data bits?). In particular, we focus on the trade-offs for improving *energy efficiency* and extending the *lifetime* of the 3D memory system.

Rather than using a single PCM material that is optimized for the worst-case temperature, we propose using different PCM material for each layer. The PCM material is chosen such that its operating temperature is equal to or slightly higher than the layer's peak temperature. We show that this improves the energy efficiency of the memory system by 1.21X to 3.5X over the case where all layers use the same PCM material that is optimized for the maximum possible operating temperature. Moreover, our experiments show that the programming current is preserved similar across layers in the heterogeneous 3D architecture. This simplifies the write circuit as compared to the approach used by Zhang and Li [31], where the programming current has to be adjusted according to the ambient temperature.

We exploit heterogeneous PCM materials to provide longer lifetime for cells used to store ECC bits than that used to store data bits. We propose to store ECC bits in the logic layer used for peripheral circuits and memory controller [31]. The heterogeneous material composition provides a system lifetime that is 30% more than that provided by Hamming-based ECC, where both data and ECC bits are stored in the same material. In addition, it provides comparable results to ECP [20] but protects against both hard and soft errors. Finally, it can be combined with other error-correcting techniques to improve storage overhead and/or lifetime. For example, using the reliable implementation of ECC1 instead of ECP1 for local error correction in PayG [17] reduces the storage overhead of PayG by 23%.

The cost overhead associated with the fabrication and testing of these materials is not well-known at this point. However, we attempt to make the case for further exploration of the material space. We believe that exploring the whole material spectrum will open opportunities for using the material that is right for the use-case. In this study, we provide an example and demonstrate the advantages associated with it. However, there are a lot of other use-cases that can further emphasis the benefits of material exploration. For example, one can use a material, such as Sb_2Te_3, that trades retention for better energy. In such case, you will need to refresh the memory system, however, the overhead will be relatively low (refresh every couple of days).

The rest of this paper is organized as follows. Section 2 provides a background on PCM materials trade-offs, describes its use in 3D stacking, and presents its major limitations and the open opportunities for energy-efficient design. Section 3 presents a detailed study of the trade-offs among different PCM materials. Section 4 describes the methodology used for 3D thermal modeling. Section 5 describes our proposed heterogeneous 3D PCM memory stack used to improve the energy efficiency of the memory system. Section 6 describes ECC implementation that uses a heterogeneous PCM stack to improve the reliability of the memory system and/or storage overhead of the ECC scheme. Section 7 describes the architectural implications for using a multi-material PCM memory stack. Section 8 discusses the manufacturability of tailoring the chemical composition of PCM and their integration in 3D stack. Finally, we conclude in section 9.

2 BACKGROUND AND MOTIVATION

In this section, we give a brief background on trade-offs between different phase-change materials. In addition, we present the potential use of phase-change memories in 3D-die stacking, highlighting its advantages and discussing some of the potential areas for improvement. Finally, we discuss error-correcting techniques that have been proposed for phase-change materials motivating for Hamming-based ECC schemes.

2.1 Materials Trade-offs

In the ternary system Ge-Sb-Te (shown in Figure 1), different chalcogenide materials lie along the pseudobinary line $(GeTe)_x(Sb_2Te_3)_y$. These materials exhibit different characteristics and present different trade-offs in terms of programming energy, operating temperature, and endurance. Moving down the pseudobinary line from GeTe to Sb_2Te_3 reduces the write energy as well as improves the endurance. However, the operating temperature of the material decreases. Using PCM for memory applications in a 3D stack setting requires that it retains its data at an operating temperature of ≈ 95-100°C [31]. Thus, though materials toward the end of the pseudobinary line provide attractive properties of improved energy efficiency and lifetime, we need to be careful about using the suitable material for the memory application given the operating temperature target. Note that exceeding the material's operating temperature exponentially reduces its reliability [7].

Figure 2 shows a top-level diagram of parameter trade-offs across different chalcogenide materials on the pseudobinary line. The stoichiometry coefficient is the ratio y/(x+y) in $(GeTe)_x(Sb_2Te_3)_y$ and it ranges from 0 (GeTe) to 1 (Sb_2Te_3). An arrow between two parameters indicates the presence of a relationship between them which is either direct (+) or inverse (-). In section 3 we will present a detailed study of different trade-offs between different PCM materials and how such trade-offs could potentially be exploited to improve the energy efficiency and lifetime of the memory system. The numbers in circles are labels for the arrows and will be used in section 3.

2.2 PCM and 3D Die-Stacking

Emerging 3D integration allows the memory to be stacked on top of the microprocessor. This has the advantage of significantly reducing the wire delay between the two and thus alleviating memory bandwidth and latency constraints [14]. However, the increased power density of the 3D technology leads to elevated on-chip temperature.

Using phase-change memory would potentially reduce power consumption and alleviate temperature constraints. This is due to the low standby power and high temperature-friendly operation that PCM exhibits. Writing PCM cells requires heating the phase-change material to a high temperature threshold. Thus, holding material composition constant, the programming energy of PCM cells can be reduced as the chip temperature is elevated [31], i.e. the PCM cell's programming current is dependent on ambient temperature [31]; as the ambient temperature increases, programming current decreases. Varying the material compositions, however, improves the programming energy for materials that are optimized for low ambient temperatures (shown in Figure 4).

In this paper, we propose to improve the energy efficiency of the memory system by exploring the benefits of building a heterogeneous PCM memory stack that leverages PCM temperature friendliness as well as non-uniform temperature across memory layers. This would potentially improve the energy efficiency by up to 3.5X as compared to a homogeneous memory stack (shown in section 5).

2.3 Error-Correcting Techniques for PCM

In this section, we give a brief background on the error categories PCM cells are prone to and the different error-correcting techniques that have been proposed for mitigating such errors. Finally, we provide a motivation for using Hamming-based ECC codes, especially after overcoming early failure problems.

2.3.1 PCM error categories. Phase-change memories are susceptible to both hard and soft errors. Hard errors are due to cell wear-out, which is caused by repeated writing. Writing PCM cells requires heating the phase-change material to high temperature thresholds; melting temperature (600°C) in case of RESET (writing 0) and crystallization temperature (300°C) in case of SET (writing 1). Repeated application of the high temperatures required for writing the cell impacts its lifetime by causing the heating element to separate from the phase-change material, leaving the cell unmodifiable. Kim and Ahn[12] showed that a RESET writing condition is responsible for limited endurance due to material melting and following quenching operation.

Unlike DRAM, PCM cells are not susceptible to particle-induced soft errors. However, they are still prone to soft errors caused by different reasons such as spontaneous crystallization (long-term drift), and resistance drift (short-term drift). Long-term drift is due to slow crystallization of the phase-change material at room temperature, which degrades the cell resistance over time. This type of error can be mitigated by periodically refreshing cells every several days [26]. Short-term drift, on the other hand, occurs when the resistance of the cell continues to grow for a certain period of time before it starts reducing again after the sudden cooling of a PCM cell that triggers the state change. This kind of error is shown to be problematic for PCM, especially for multi-level cells [1].

2.3.2 PCM error-correcting techniques. Recent research has used error-correcting codes to ensure the reliability of the memory system and thus expand its lifetime. Strong ECC codes that protect large data blocks (64 bytes) against multi-bit hard failures (at least 6 bits error) are shown to be necessary for expanding the lifetime of the PCM system [20] [26] with reasonable storage overhead.

The complexity of Hamming-based error-correction codes, however, increases linearly with the correction capability [23]. In addition, ECC bits wear out much faster than the data bits they are designed to repair [20].

In order to address these problems, Schechter et al. [20] proposes to use ECP to keep the position of failing bits within the data block. Each fail cell is replaced by a spare cell. PayG [17] reduces the storage overhead of ECP by allocating error-correction entries in response to the number of errors in the given memory line. SAFER [21] proposes to partition a data block dynamically while ensuring

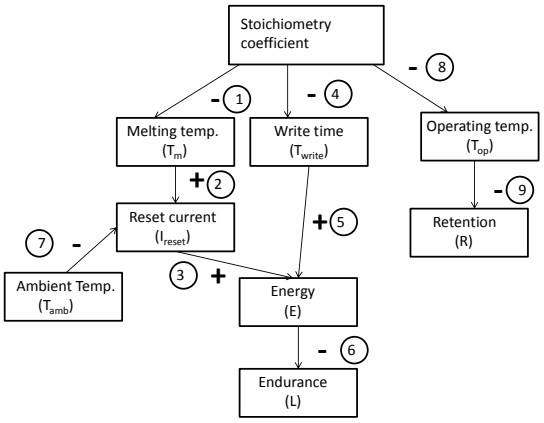


Figure 2: Top-level diagram of parameter trade-offs across different chalcogenide materials. Increasing the stoichiometry coefficient would reduce write energy and increase endurance but at the cost of low operating temperature.

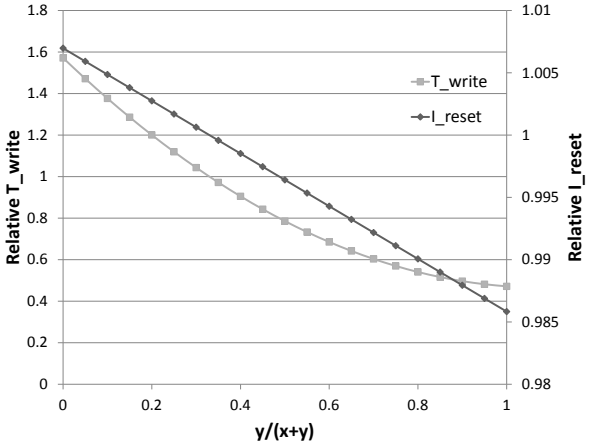


Figure 3: I_{reset} and T_{write} as a function of stoichiometry coefficient $y/(x+y)$ in $(GeTe)_x(Sb_2Te_3)_y$. The values are normalized to $Ge_2Sb_2Te_5$. Increasing the stoichiometry coefficient reduces programming current and duration.

that there is at most one fail bit per partition and uses single error-correction techniques per partition for fail recovery. All previous techniques require a custom-designed PCM, which increases cost per memory bit. In addition, they only protect against hard errors.

In order to account for both hard and soft errors, FREE-p [26] relies on strong ECC codes accompanied by a fine-grained remapping mechanism to tolerate wear-out failures. In addition, it implements multiple ECC logic paths to reduce the complexity of error correction. Zhang and Li [31] propose using strong BCH error-correcting codes along with an OS-level page-allocation scheme that takes wear-out level into consideration, i.e., pages with a lower level of wear-out are allocated first.

While each of the previous techniques that rely on strong ECC [26] [31] proposes a different solution to overcome the increased complexity of having strong codes, these techniques still did not address the fact that ECC bits could fail earlier than the data bits they are designed to protect. In the next section, we will further describe this problem and present the advantages of eliminating it.

2.3.3 Early failure of ECC bits. While Hamming-based ECC schemes are general and can address both hard and soft errors, they have high entropy, i.e., they are likely to flip on every data write (whenever any data bit in its protected region is written). Data bits within a large block, on the other hand, do not always change. Awasthi et al. [1] shows that only around 50% of the data bits within the 64-byte block size are expected to be reprogrammed on every write. This results in early failure of ECC bits, which can be further exacerbated by cell-lifetime variation [20] [30]. Eliminating such problems would improve the lifetime provided by the Hamming-based ECC scheme. We propose to eliminate this problem by storing ECC bits in more reliable storage than that used to store the data bits. However, as we will show in section 3, using different PCM material with higher reliability to store ECC bits does not come for free, since it will have to operate under lower temperature.

The proposed approach has three main advantages. First, it preserves the Hamming-based ECC capability of providing end-to-end reliability. Second, it can improve the reliability of error-correcting schemes that rely on Hamming-based ECC for error correction. Third, it reduces the storage overhead over ECP from 11.9% to 9.1% without sacrificing reliability. Thus, it can replace the use of ECP in various frameworks such as PayG [17] reducing its storage overhead from 3.8% to 2.9%. Detailed evaluation of the proposed scheme is shown in section 6.

Next, we will present a detailed study of trade-offs among different PCM materials. We will focus on those trade-offs that have potential to improve PCM energy efficiency, and lifetime.

3 PCM MATERIALS TRADE-OFFS

In this section, we present a detailed study of trade-offs among different PCM materials. In our study, we will focus on the most commonly used materials in chalcogenide-based phase-change systems; GST systems such as $GeSb_4Te_7$, $GeSb_2Te_4$, $Ge_2Sb_2Te_5$; and others. Although other PCM materials are possible, we focus on these GST systems because there exists some extrapolatable data and to provide an example where heterogeneous PCM can be useful. These GST materials are close on the pseudobinary line and thus are expected to have only slightly different characteristics. In addition, we will show how we could potentially exploit such trade-offs to improve PCM energy efficiency, and lifetime.

3.1 Reset Energy

In this section we will discuss the effect of using different chalcogenide materials on reset energy. Reducing the reset energy improves the endurance of the memory system [30]. Assuming constant resistivity irrespective of the change in stoichiometry [13], energy is directly proportional to the square of reset current and write time $E \propto I_{reset}^2 * T_{write}$. As we approach the Sb_2Te_3 composition along the pseudo-binary line, the melting temperature T_m decreases, thus affecting the reset current I_{reset} required for melting a chalcogenide volume for amorphization. Refer to arrow 1

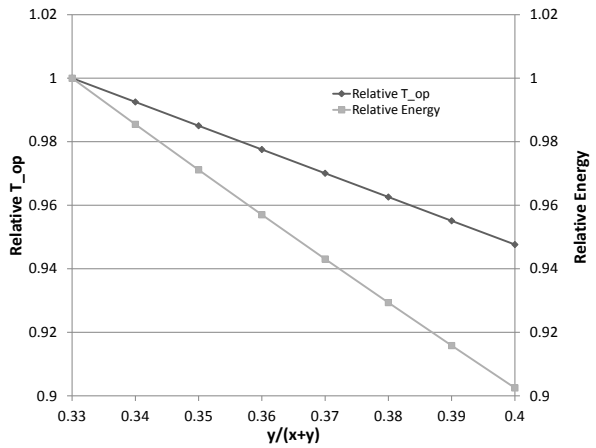


Figure 4: Relative T_{op}, and relative energy with respect to $Ge_2Sb_2Te_5$ for subset of stoichiometry coefficient $y/(x+y)$ whose T_{op} ranges between 100 and 95°C. Increasing the stoichiometry coefficient improves programming energy but reduces operating temperature.

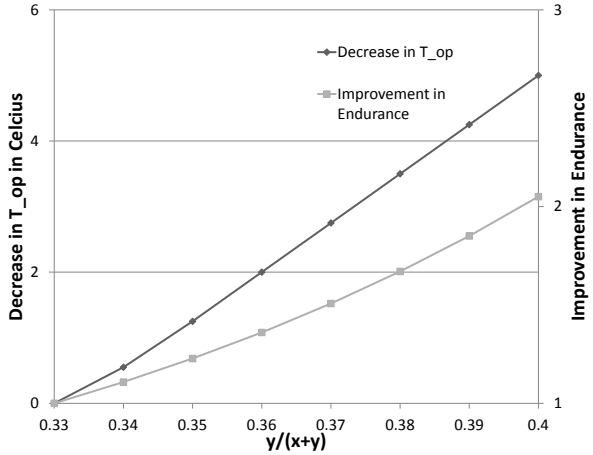


Figure 5: Decrease in T_{op}, and improvement in endurance over $Ge_2Sb_2Te_5$ for different stoichiometry coefficient $y/(x+y)$. At stoichiometry coefficient = 0.387, 1.8X improvement in endurance over $Ge_2Sb_2Te_5$ is achieved at the cost of reduced operating temperature by 4°C.

and arrow 2 in Figure 2. Figure 3 shows the decrease in I_{reset} current as the stoichiometry coefficient $y/(x+y)$ in $(GeTe)_x(Sb_2Te_3)_y$ increases, i.e., moving toward Sb_2Te_3. The decrease in I_{reset} is relatively insignificant. This data was obtained from Lacaita and Ielmini[13]. The I_{reset} values are relative to $Ge_2Sb_2Te_5$ (the most commonly used phase-change material) at stoichiometry coefficient 0.33. These first-order estimates assume a constant electrical resistivity and thermal conductivity irrespective of the change in stoichiometry [13].

In addition, different chalcogenide materials along the pseudobinary line have different write time (T_{write}). The threshold for pulse durations for crystallization monotonically increases from 30ns to 100ns with increasing the GeTe content (arrow 4 in Figure 2). This range of pulse duration is suitable for both crystallization and amorphization [25]. The interpolated values of T_{write} from the three stoichiometric compositions (GeTe, $GeSb_2Te_4$, and Sb_2Te_3) using the data points provided in Yamada et al. [25] are shown in Figure 3. T_{write} is relative to $Ge_2Sb_2Te_5$. In conclusion, increasing the stoichiometry coefficient over $Ge_2Sb_2Te_5$ would decrease the reset energy by up to 2.18X as we move towards Sb_2Te_3.

3.2 Operating Temperature

In this section, we will discuss the effect of using different chalcogenide materials on the operating temperature (T_{op}). In this paper, T_{op} of a PCM cell is the maximum (peak) temperature its materials can tolerate and still maintain a 10-year data retention period.

In addition, we will show the relationship between T_{op} and reset energy for different materials across the pseudobinary line. On one hand, reducing the operating temperature restricts the applications that the phase-change material can be used for. On the other hand, reducing the programming energy improves the energy efficiency of the memory system. As shown in Figure 2, reducing the programming energy is accompanied by reduction in operating temperature.

Successfully using a PCM material for memory applications requires that it retains its data at an operating temperature of ≈ 80-90°C [18]. In the case of 3D memory stacking, the operating temperature would increase further to approach ≈ 100°C [14]. Thus, materials that fall toward the end of the pseudobinary line with low operating temperatures are not suitable for memory applications. In our study, we assume a 3D memory stack with peak temperature of 100°C. Due to temperature variance across different layers in the memory stack (temperature generally increase as we move away from the heat sink), slightly lower operating temperatures than 100°C are possible at memory layers close to the heat sink. We identify that range of materials that are suitable for 3D stacking to be within $0.33 \leq y/(x+y) \leq 0.4$. This provides a range of operating temperatures between 95°C and 100°C. Thus, in our parameter trade-off study, we will focus on that range.

Moving towards Sb_2Te_3 would decrease the operating temperature T_{op} (arrow 8 in Figure 2). Figure 4 shows the extrapolated values for T_{op} obtained from Lacaita and Ielmini [13]. The results are for the subset of materials that are suitable for 3D memory stacking applications, assuming a maximum temperature difference of ≈ 5°C between top and bottom memory layers (T_{op} ranging from 100°C at stoichiometry coefficient 0.33 to 95°C at stoichiometry coefficient 0.4). The T_{op} values are relative to $Ge_2Sb_2Te_5$. In addition, we show the write energy for the same subset of materials. Within this range, write energy can decrease by up to 10% at stoichiometry coefficient of 0.4.

In section 5, we will show that using multiple PCM materials with different operating temperature, reset current, and write time across the memory stack could potentially be used to improve the energy efficiency in 3D PCM memory stacking.

3.3 Endurance

Now that we have shown how T_{op} and reset energy varies across different materials, and identified the plausible range of materials that are suitable for 3D memory stacking, we want to show the relationship between reset energy (and thus T_{op}) and endurance.

The endurance test performed by Kim and Ahn [12] identified reset energy to be responsible for endurance failure due to material melting and the following quenching operation. In addition, they show a power law relationship between the cycle lifetime of a PCM device and its programming energy. Thus, increasing the reset energy will result in lower endurance (arrow 6 in Figure 2). Since GST materials within our range of interest ($0.33 \leq y/(x+y) \leq 0.4$) exhibit very similar characteristics [25] (shown in previous section), we would expect the same trade-off between reset energy and endurance to hold across different materials within that range. Thus, as the stoichiometry coefficient increases towards 0.4, the reset energy decreases, leading to an increase in endurance.

Figure 5 shows the improvement in endurance over $Ge_2Sb_2Te_5$ as we increase the stoichiometry coefficient $y/(x+y)$ to 0.4. In conclusion, using PCM materials with larger y/(x+y) would increase the endurance over $Ge_2Sb_2Te_5$ by up to 2X. However, this comes at the cost of having such materials operate at lower temperatures; up to 5°C less than that of $Ge_2Sb_2Te_5$.

In section 6, we will show that such improvement in reliability is enough to eliminate early failure of ECC bits. Furthermore, a memory stack with a number of layers within the range projected by ITRS [10] would provide the temperature difference required to enable the use of such material with higher endurance to store ECC bits. ITRS projected a memory stack of more than 9 layers by 2015.

4 THERMAL ANALYSIS OF 3D MEMORY MODEL

In order to provide an estimate of the difference in temperatures between top and bottom memory layers in a 3D stacked PCM system, we use the analytical model described in Im and Banerjee [9]. However, we modify its assumptions to account for unique PCM properties like temperature friendliness, i.e., being power efficient at elevated temperatures [31]. The temperature rise above ambient temperature of the j^{th} active layer in an n-layer 3D chip is given by:

$$\Delta T_j = \sum_{i=1}^{j} \left[R_i \left(\Sigma_{k=i}^{n} P_k / A \right) \right] \tag{1}$$

where R_i represents the thermal resistance between the i^{th} and $(i-1)^{th}$ layers, P_k is the power dissipation of the k^{th} layer, A is the area, and n is the total number of active layers. This model does not take into account interconnect joule heating.

We will use the above equation to estimate the difference in temperature between last PCM memory layer n, and first PCM memory layer y. Assuming identical thermal resistance (R) between PCM memory layers, $\Delta T_n - \Delta T_y$ is given by:

$$\begin{aligned} \Delta T_n - \Delta T_y = R \Big(& P_{y+1}/A + 2 * P_{y+2}/A \\ & + 3 * P_{y+3}/A + \ldots + (n\_mem - 1) * P_n/A \Big) \end{aligned} \tag{2}$$

where n_mem is the total number of PCM memory layers (y - n + 1).

In order to characterize the power density for each memory layer, we made the following assumptions. As the temperature varies across different layers of PCM memory, the minimal required programming power varies as well. Lower programming current is required to RESET/SET a PCM device at elevated temperature, resulting in smaller programming power [31]. Thus, we assume a different I_{reset} current for each PCM memory layer and that the I_{reset} current decreases as we move towards PCM memory layers away from the heat sink. We explore a range of current differences among adjacent layers. NVSim [5] is used to calculate PCM write power for a single memory layer by modeling a 256Mb phase-change memory similar to the prototype developed by Kang et al. [11]. Though PCM has a slightly higher density than DRAM ($4.8F^2$ versus $6F^2$ [10]), here we conservatively assume the same density for both PCM and DRAM.

	Layer Thickness (μm)	Thermal Parameters (mK/W)
Metal Layer	6	0.0833
Active Silicon	1	0.0083
Bulk Silicon	20	0.0083
D2D resistivity (accounts for air cavities and copper)	2	0.0166

Table 1: Thermal and layer thickness parameters

Figure 6 shows the overall structure of the 3D-stacked memory. Table 1 shows the other parameters required by the thermal modeling (obtained from Loh [14]).

In the next two sections, we will present two potential scenarios of using different PCM materials for different purposes and across different layers in a 3D memory stack. Different PCM materials could be used across the 3D memory stack to improve the energy efficiency and lifetime of the memory stack.

5 ENERGY-EFFICIENT 3D MEMORY STACK

In the "One memory fits all" approach, the PCM material is chosen such that it can safely operate under the maximum expected temperature in 3D stacking, regardless of the configuration of the stack, such as the number of layers. This design optimized for the maximum expected temperature is energy inefficient, because programming the cell is most energy efficient at the assumed maximum temperature, which might never be reached. In addition, as described in section 2.2, programming the cells at PCM memory layers close to the heat sink is less energy efficient than programming their counterparts far from the heat sink. This leaves room for improving the energy efficiency of the memory system.

We propose to exploit such potential as follows. First, in order to improve the energy efficiency of the whole memory stack, we use a PCM material that is optimized for the considered 3D stack setting, i.e., taking into consideration the number of layers and the maximum expected temperature for that setting. Second, in

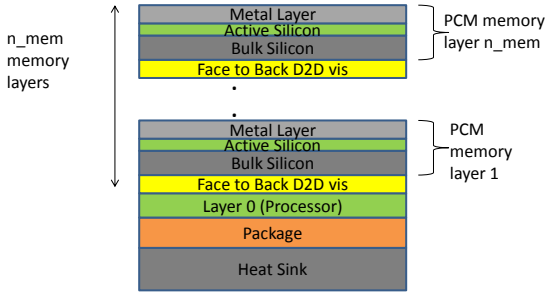


Figure 6: 3D stack structure

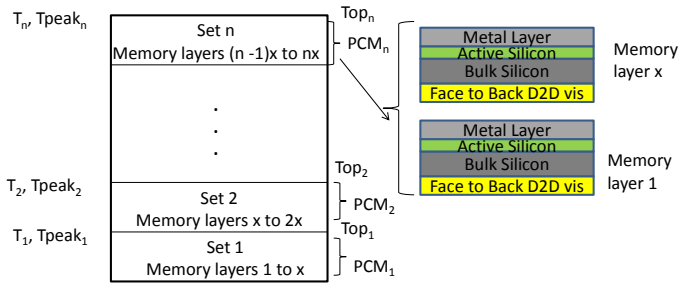
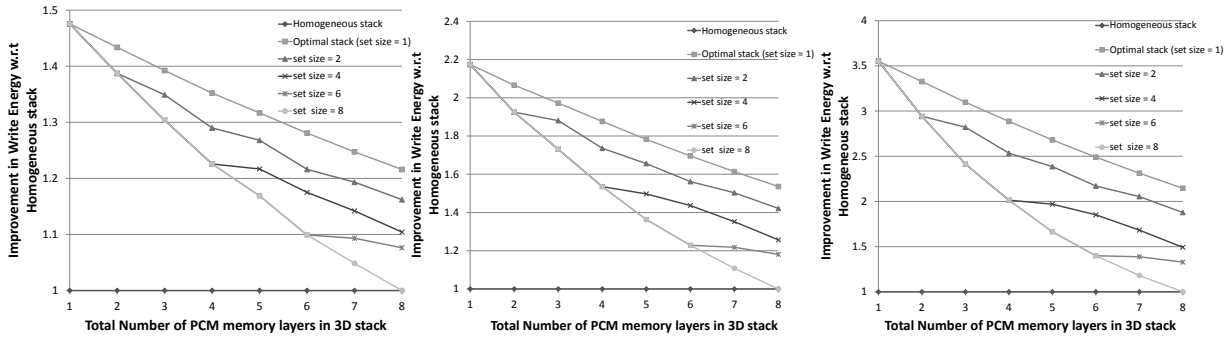


Figure 7: Memory layout using different PCM materials



(a) 0.01mA **(b) 0.02mA** **(c) 0.03mA**

Figure 8: Improvement in write energy by heterogeneous PCM memory stack. The results are relative to a homogeneous PCM memory stack where a different PCM material is used for each layer. The graphs assume various differences in programming current between adjacent memory layers (a) 0.01mA, (b) 0.02mA, and (c) 0.03mA.

order to improve the energy efficiency of memory layers operating under reduced temperature due to proximity to the heat sink, we use different PCM material for different layers in the memory stack, each material being optimized for the maximum temperature of that layer. In our discussion we assume a true 3D organization, where bitcell arrays are stacked in 3D fashion [14]. In addition, we assume that temperature increases as we move away from the heat sink.

In the **homogeneous stack base case** where only a single PCM material is used for the whole stack, the write energy is proportional to $\Sigma_{i=1}^{n\_mem} {I_{reset}}_i^2 * T_{write i}$ where n_mem is the total number of memory layers. Since less programming current is required to reset a PCM device at elevated temperature [31], $I_{reset 1} \geq I_{reset 2} \geq \ldots \geq I_{reset n\_mem}$. In addition, since the PCM material is chosen for the highest temperature, $I_{reset n\_mem} \geq I_{reset opt}$ where $I_{reset opt}$ is the reset current used to program memory cells when operating at a temperature that is equal to the material's maximum allowed operating temperature. T_{write} is the same for all layers as it is largely dependent on the property of the PCM material [30].

On the other hand, in the **heterogeneous setting**, if each layer is made of different PCM material such that the T_{op} of the used material is equal to or slightly higher than the layer's temperature, then each layer will use the $I_{reset opt}$ current for its material. As shown in Figure 3, different PCM materials have a similar reset current. This leads to all layers using a similar programming current that is equal to $I_{reset opt}$. Thus, the write energy is proportional to $\Sigma_{i=1}^{n\_mem} {I_{reset}}_{opt}^2 * T_{write i}$. In addition, write time decreases for material with lower operating temperatures. Thus, $T_{write 1} \leq T_{write 2} \leq \ldots \leq T_{write n\_mem}$. Consequently, the write energy in this case is less than the write energy of the base case where a single material is used for the whole memory stack.

This can be generalized into logically partitioning the memory stack into a number of sets as shown in Figure 7. The 3D memory stack has n sets each having x or fewer memory layers, depending on the total number of layers. Each set uses a different PCM material. Intuitively, as the number of layers within a set increases, the improvement in write energy over the base case decreases. A single layer per set is optimal because every layer is optimized for its own maximum temperature. Next, we will formulate our assumptions about each set S_i and the properties that the PCM material PCM_i must have to be used within that set in order to achieve better write energy.

Each set S_i has a peak temperature ${T_{peak}}_i$ equals to the temperature of the furthest layer from the heat sink within that set. T_{peak} increases for higher sets, i.e., ${T_{peak}}_1 \leq {T_{peak}}_2 \leq \ldots \leq {T_{peak}}_n$. For each set S_i, we choose a PCM material PCM_i. Each PCM_i has a different operating temperature ${T_{op}}_i$. T_{op} increases for sets that are far away from the heat sink, i.e., ${T_{op}}_1 \leq {T_{op}}_2 \leq \ldots \leq {T_{op}}_n$. Within set S_i, PCM_i is used, whose ${T_{op}}_i$ is equal to or slightly higher than ${T_{peak}}_i$.

As the T_{op} decreases for sets closer to the heat sink, we choose PCM materials with larger stoichiometry coefficient for these sets (arc 8 in Figure 2) as compared to sets far from the heat sink. Increasing the stoichiometry coefficient for these sets would lead to decreasing the melting temperature T_m (arc 1 in figure 2). Thus, T_{mi} decreases for sets close to the sink, i.e., $T_{m1} \leq T_{m2} \leq \ldots \leq T_{mn}$. Decreasing T_{mi} would affect $I_{reset\,i}$. However, as shown in Figure 3, there is a negligible difference in I_{reset} across different PCM materials, i.e., $I_{reset1} \approx I_{reset2} \approx \ldots \approx I_{resetn}$. This leads to each set S_i having similar I_{reset} at T_{peak_i} but with a smaller number of layers. In addition, increasing the stoichiometry coefficient leads to decrease in the write time. Thus, $T_{write1} \leq T_{write2} \leq \ldots \leq T_{writen}$.

5.1 Methodology

We use NVSim [5] to calculate PCM write energy for a single memory layer by modeling a 256Mb phase-change memory similar to the prototype developed by Kang et al. [11]. The programming current and write time fed to the NVSim model is determined as follows. The programming current used for each layer varies depending on the assumed current difference among adjacent layers. We explore a range of current differences from 0.01mA to 0.03mA. The time required to write a memory cell depends on the PCM material used. We choose a PCM material whose operating temperature is equal to or slightly higher than the layer's temperature. The operating temperature of each layer with respect to $Ge_2Sb_2Te_5$ is calculated using the analytical model in section 4. We then use the operating temperature with respect to $Ge_2Sb_2Te_5$ to get the write time of the material using data from Lacaita and Ielmini [13], which is illustrated in Figure 5. The homogeneous baseline assumes manufacturing uniform dies that are made from a single material that is suitable for all use-cases. Therefore, we assume that the base-case material, $Ge_2Sb_2Te_5$, is optimized for the most general configuration, an 8-layer memory stack that operates at 100°C. Finally, we assume a true 3D organization where bitcell arrays are stacked in 3D fashion [14]. Thus, total write energy is equal to the sum of write energies across all layers.

5.2 Results

5.2.1 Energy Savings at Peak Temperature. Figure 8 shows the effect of using a heterogeneous PCM memory stack where a different material is used for every set (consisting of up to x memory layers) on the energy efficiency of the memory system. The graphs assume various differences in programming current between adjacent memory layers. Homogeneous stack is the energy consumed by the base case where a single PCM material that is optimized for the highest operating temperature in 3D stack setting is used for all memory layers. The PCM material for the base case in our experiment is optimized for a 8-layer memory stack. Optimal stack (set size = 1), on the other hand, assumes a set size of one memory layer, i.e., a different PCM material is used for each memory layer. For maximum energy gains, each set uses a material that is optimized for the maximum temperature within this set, i.e., temperature of the furthest layer from the heat sink within that set. In addition, we explored the potential of increasing the set size, i.e., the number of layers within a set, on energy efficiency. For example, a set size = 2 in a 8-layer memory stack means that the memory stack is partitioned into 4 sets each having 2 layers and that these sets are of different materials.

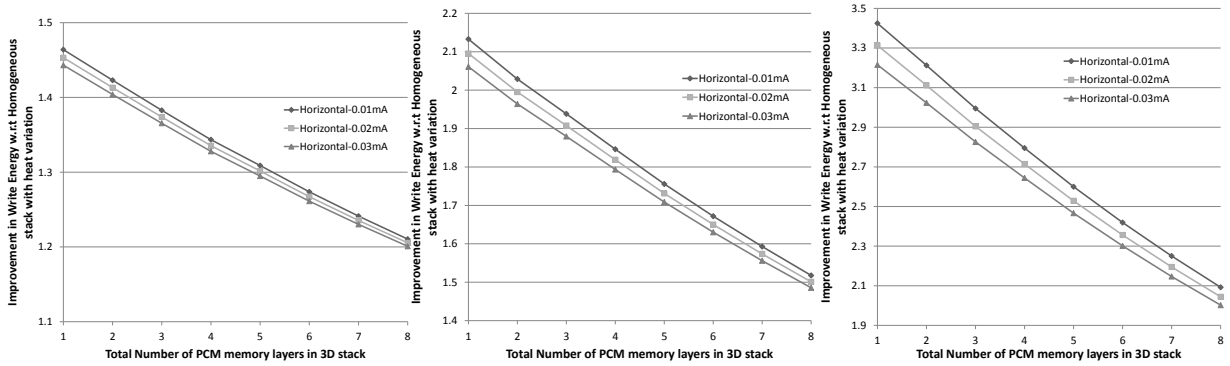
The energy gains achieved by our scheme varies depending on the total number of memory layers in the stack, the number of layers within each set, and the difference in programming current between adjacent memory layers. Using different material for each memory layer (Optimal stack) provides from 1.21X (8-layer 3D stack with 0.01mA programming current difference between adjacent layers) up to 3.5X (1-layer 3D stack with 0.03mA programming current difference between adjacent layers) reduction in the write energy. The reduction in energy consumption decreases as the maximum number of layers within each set increases (i.e., reducing the number of different PCM materials).

Within a given scheme (e.g., set size = 2), the reduction in write energy over the homogeneous stack setting energy consumption decreases as the total number of memory layers in the memory system increases. The reason is that the used PCM material for the homogeneous stack is optimized for a 8-layer stack and thus its energy efficiency improves as the number of layers in the memory stack approaches 8.

5.2.2 Energy Savings in Presence of Heat Variations. In previous section we showed the improvement in programming energy due to the use of heterogeneous stack. Our results assume that all cells within a layer are written under peak temperature (equals to the operating temperature of the material). This corresponds to the maximum energy savings due to temperature friendliness of PCM operation. However, temperature varies within the same memory layer, as well as across layers due to the heat dissipated from the processing layer which exhibit hotspots. Operating under temperature less than the peak temperature will require the use of programming current that is larger than that used at peak temperature ($I_{resetopt}$). This will result in reduction in energy improvement with respect to the heterogeneous memory system with no heat variations.

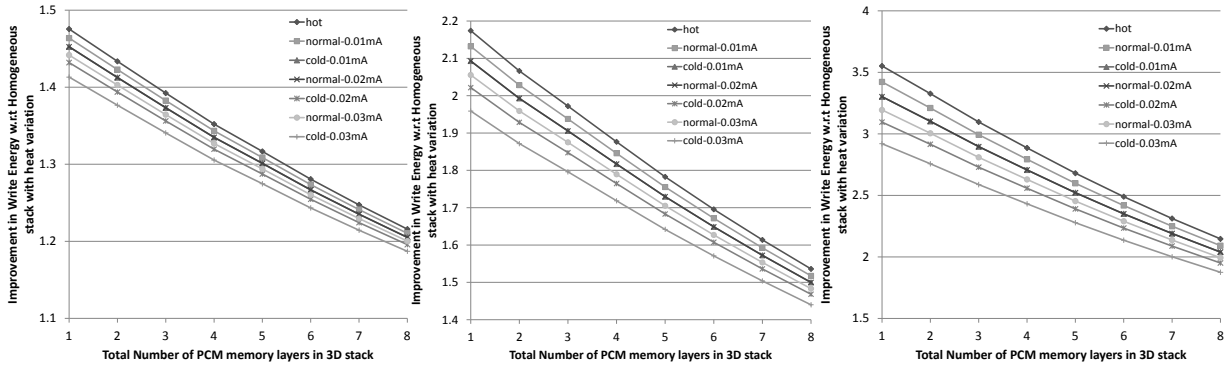
Due to heat variations among different regions within the same memory layer, the improvement in write energy depends on which region in the memory system you are writing to. The higher the temperature of this region, the more energy efficient writing to that region. We divide the memory system into three regions: hot, normal and cold. Each of these regions use different programming current magnitude to program their cells to account for *horizontal* heat variation.

Figure 9 shows the overall improvement in write energy over the homogeneous memory system in the presence of heat variations. We provide sensitivity analysis to the difference in current used to programming cells that reside in different regions within the same layer as compared to the current used to program the cells in the hot region. For example, horizontal-0.01mA means that a current difference of 0.01mA is used to program cells between adjacent regions within the same layer, i.e., the current used to program the cells in the normal region is 0.01mA more than that used to program the cells at the hot region and the current used to program the cells in the cold region is 0.01mA more than that used to program the cells at the normal region. We present the improvement in energy for the same three cases presented in figure 8 which accounts for *vertical* current difference among adjacent memory layers. The



(a) 0.01mA **(b) 0.02mA** **(c) 0.03mA**

Figure 9: Improvement in write energy by heterogeneous PCM memory stack due to heat variation. The results are relative to a homogeneous PCM memory stack with heat variation. The graphs assume three regions: hot, normal and cold with various differences in horizontal programming current between different regions within the same memory layer. The size of the hot region is equal to 1/4 of the memory layer. The graphs assume various differences in vertical programming current between adjacent memory layers (a) 0.01mA, (b) 0.02mA, and (c) 0.03mA.



(a) 0.01mA **(b) 0.02mA** **(c) 0.03mA**

Figure 10: Detailed improvement in write energy by heterogeneous PCM memory stack for writing to each of the regions: hot, normal and cold. The results are relative to a homogeneous PCM memory stack with heat variation. The graphs assume various differences in horizontal programming current between different regions within the same memory layer; 0.01mA, 0.02mA and 0.03mA. The graphs assume various differences in vertical programming current between adjacent memory layers (a) 0.01mA, (b) 0.02mA, and (c) 0.03mA.

results assume that writes are uniform across the memory layer. In addition, they assume that the size of the hotspot, i.e., the region that operates under peak temperature, is equal to one quarter of the memory layer. The size of each memory region, hot, normal and cold is obtained from running hotspot simulations on SPEC2000 benchmarks[8]. We also did sensitivity analysis on the size of the hotspot region by varying the temperature range that is considered hot and dividing the rest of the temperature range equally between normal and cold. We varied the size of the hotspot from 1/4 down to 1/8.

Heat variation has negligible reduction (< 1%) in the minimum energy improvement identified at the previous section as 8-layer 3D memory stack with 0.01mA vertical current difference after accounting for the worst heat variation where there is a horizontal current difference of 0.03mA between adjacent regions. This percentage increases to 1.5% for smaller hotspot region whose size is 1/8 of the memory layer. On the other hand, the best case energy improvement identified at the previous section as 1-layer 3D memory stack with 0.03mA vertical current difference reduces from 3.55X to 3.2X (9.8%). This percentage increases to 11.5% for smaller hotspot region whose size is 1/8 of the memory layer.

Hot region	0.25	0.19	0.1552	0.1366	0.125
Normal region	0.5	0.53	0.5194	0.5192	0.521
Cold region	0.25	0.28	0.3254	0.3442	0.354

Table 2: The ratio of regions: hot, normal, and cold. The size of the hot regions varies from 0.25 (1/4) to 0.125 (1/8).

In order to understand the results from figure 9, we show in figure 10 the detailed improvement in write energy over the homogeneous memory system when writing to each of the memory regions separately. For example, normal-0.03mA shows the improvement in write energy when writing to the normal region which requires 0.03mA more current than writing to the hot region. As shown, 17% worst case reduction in energy is identified at a 1-layer 3D memory stack when writing to the cold-0.03mA region. The overall energy improvement shown in figure 9 is obtained from the weighted dot product of the write energy to each region (at figure 10) and the size of that region obtained from hotspot simulations (shown in table 2).

6 RELIABLE ECC IMPLEMENTATION (ECCN-RELIABILITY)

As described in section 2.3.3, Hamming-based ECC codes are likely to fail earlier than the data bits they are designed to repair. Thus, storing the ECC bits in a PCM material that is more reliable than the material used to store the data bits would improve the lifetime of the memory system over traditional Hamming-based ECC implementation where both data and error-correcting bits are stored within a single material. We refer to this implementation as ECCN-Reliability where N is the code strength and Reliability is the improvement in reliability of ECC bits over that of the data bits. For example, ECC6-2X is an ECC6 that stores ECC bits in a 2X more reliable storage than that used to store the data bits.

In order to be able to use a different material with higher reliability to store ECC bits, this material has to operate at a lower temperature than that used to store the data bits (trade-off presented in section 3.3). Thus, we need to store the ECC bits in a memory layer that is close to the heat sink and consequently has a lower operating temperature. In a true 3D stack [14], there exists a logic layer that is used for the memory controller and memory peripheral circuit [14] [31]. This layer resides on top of the processing layer. Thus, we choose this logic layer as a candidate position for storing ECC bits. However, we need to address the following issues. First, since the logic layer is placed on top of the processing layer, we need to be careful about placing the ECC bits in the right position, away from the hotspots. Second, we need to determine the characteristics of the candidate material to store the ECC bits such as its relative reliability and operating temperature to the material used to store the data bits. Finally, we need to determine the 3D stack configuration required to guarantee that the temperature of the location used to store the ECC bits will not exceed the material's operating temperature.

The main hot spots within a processing unit are identified as the register file, load-store queue, and execution units [22]. Thus, assuming a quad core on the processing layer closest to the heat sink for thermal efficiency (floor plan is shown in Figure 11), the hot spots will be concentrated in the middle of the die. Thus, we propose to store ECC codes either toward the left or the right of the logic layer.

Next, we will study the material and 3D stack configuration requirement to enable the use of a more reliable PCM material to store the ECC bits. We will start with our experimental methodology.

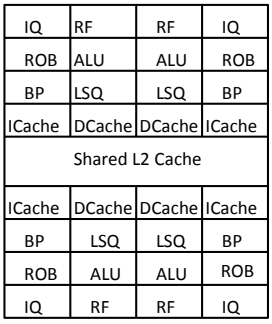


Figure 11: Floorplan of processor layer.

6.1 Methodology

We performed memory-failure simulation with a number of simplifying assumptions in order to quantify the lifetime of the memory system for different error-correction schemes. We define memory-system lifetime by the number of writes before the first data line fails, i.e., it has more errors than can be corrected by the error-correction scheme. The simulation assumptions are as follows:

- Similar to other studies [20] [26], we only focus on modeling hard errors. However, our scheme is general and can be used to correct both hard and soft errors. Due to process variation, we assume the lifetime of each memory cell follows a normal distribution without any correlation between neighboring cells [17] [20]. We assume a mean lifetime of 10^8 writes and a 20% coefficient of variance.
- We assume a 64-byte data block, with an error-correcting code that is able to correct up to 6 errors [20].
- We assume that only 50% of the data bits are expected to be reprogrammed on every write. i.e., 50% of the data bits will be reprogrammed with probability 0.5 [1].
- Each ECC bit will be reprogrammed with probability 0.5 any time any of the data bits are reprogrammed [20].
- We assume perfect wearleveling [17] [20].
- We simulated 2000 pages with page size equal to 4KB [20].

This simulation is used to determine how much more reliable the ECC bits need to be before eliminating the early-failure problem. In order to identify the 3D stack configuration required to enable the use of another material with higher reliability but lower operating temperature to store ECC bits, we use the analytical model described in section 4 to determine the temperature difference among memory layers within the 3D stack.

6.2 Results

6.2.1 Advantages of eliminating early ECC failure: In Figure 12, we compare the memory lifetime and the storage overhead of ECC6-2X with various error-correction techniques. We choose ECC6-2N because using 2X more reliable storage to store ECC bits will eliminate its early failure problem (shown in the next section). For each technique, the lifetime is normalized to an idealized ECP6 that has zero storage overhead, which we refer to as OPT6. ECC6 refers to the Hamming-based coding scheme that corrects up to 6 errors. This implementation stores both data and ECC bits within the same

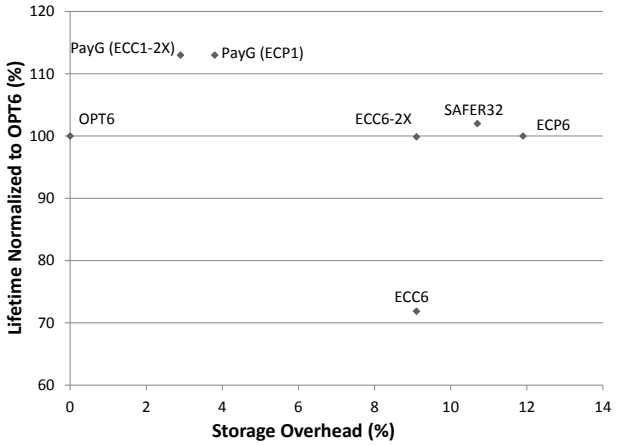


Figure 12: Lifetime and Storage overhead of various Error-correcting schemes.

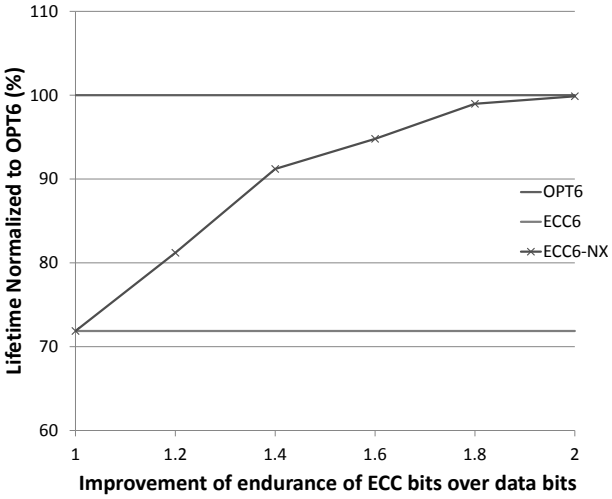


Figure 13: Normalized system lifetime for ECC6-NX where the average endurance of the ECC bits is varied from 1X to 2X the endurance of the data bits it is designed to repair. The results are normalized to OPT6. ECC6-1.8X provides comparable lifetime to OPT6.

material. ECC6-2X improves the lifetime of ECC6 by 30% with the same storage overhead. We also compare to other techniques like SAFER32 [21] and PayG (ECP1)[17]. The lifetime results for various schemes are obtained from our memory-failure simulation, except for SAFER32, whose lifetime with respect to OPT6 is obtained from [21]. In case of PayG, the lifetime results were obtained by simulating ECP8, which provides comparable lifetime to PayG, as shown by Qureshi [17]. The improvement in lifetime over OPT6 provided by our simulations matches closely the results in [17].

A reliable implementation of ECC (ECCN-Reliability) can be combined with PayG (ECP1), which provides a general framework that can be implemented with any error-correcting scheme in order to reduce its storage overhead. PayG (ECC1-2X) replaces ECP1 codes used for local error correction with SEC (single error correction), which in turn reduces the storage overhead to 2.9%. This indicates that combining our approach (ECCN-2X) with other schemes like ECC6 and PayG (ECP1) improves their lifetime and/or the storage overhead. Finally, techniques that rely on strong error-correcting codes like FREE-p [26] are orthogonal to our approach.

6.2.2 Characteristics of the Material to store ECC:. Figure 13 shows the mean time to the first uncorrectable error when the average endurance of the ECC bits is varied from 1X to 2X the endurance of the data bits it is protecting (ECC6-NX). All lifetime numbers are normalized to OPT6 (ECP6 with zero storage overhead). As shown, using ECC bits with average endurance that is 1.8X more than the average endurance of the data bits would provide a memory-system lifetime that is comparable to OPT6 (within 1%). In addition, it improves the lifetime by approximately 30% over ECC6 (implementation where both data and ECC bits are stored within the same material).

As shown in section 3.3, PCM material with stoichiometry coefficient $y/(x + y) = 0.387$ that operates at a temperature that is 4°C less than that of GST with stoichiometry coefficient $y/(x + y) = 0.33$ has 1.8X improvement in endurance. Thus, this material can be used to store ECC bits as long as it operates at a temperature that is 4°C less than the temperature of the material used to store the data bits.

6.2.3 3D stack configuration: Now that we have identified the material as well as the possible location to store ECC bits, we need to calculate the size of the stack in order to achieve a 4°C temperature difference. Figure 14 shows the minimum number of memory layers required to achieve a 4°C temperature difference between the furthest memory layer from the heat sink and the logic layer. We present the results assuming a range of I_{reset} current differences between adjacent layers. For example, assuming a current difference of 0.03mA between adjacent memory layers, a 3D memory stack of at least 10 memory layers is required to provide at least 4°C difference in T_{op}. This difference in T_{op} would make the use of another PCM material; with lower T_{op} and higher endurance (1.8X) than GST; to store the ECC bits possible.

Figure 15 shows detailed results for the increase in temperature difference between the furthest memory layer from the heat sink and the logic layer as the number of memory layers increase. This assumes 0.03mA current difference between adjacent memory layers.

7 ARCHITECTURE IMPLICATIONS OF HETEROGENEOUS 3D MEMORY STACK

In order to build a multi-material 3D stack, we need to address two issues: First, the implications of heterogeneity on the programming current used. Second, the variation in write time required to successfully program the cells among different materials.

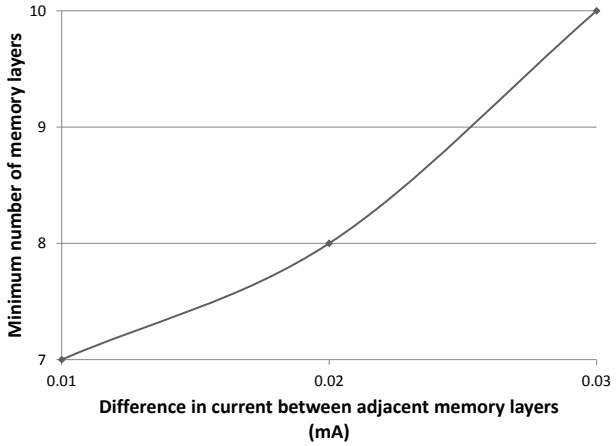


Figure 14: Minimum number of memory layers for achieving 4°C difference in T_{op} between the furthest memory layer from the heat sink and the logic layer. This enables storing ECC bits in PCM material with 1.8X endurance improvement over the PCM material used to store the data bits.

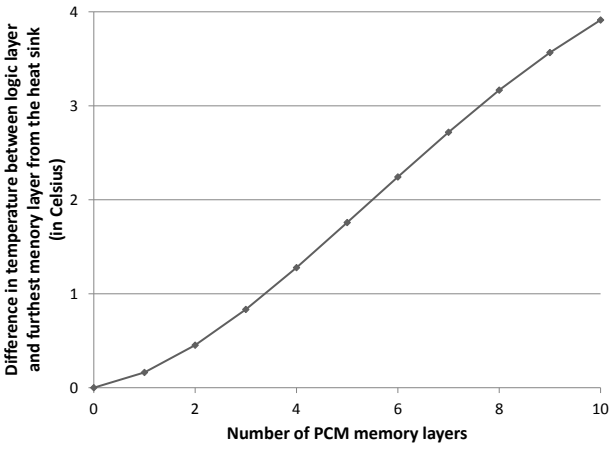


Figure 15: Difference in temperature between the furthest memory layer from the heat sink and the logic layer in 3D stacked memory configuration. 0.03mA current difference between adjacent memory layers is assumed.

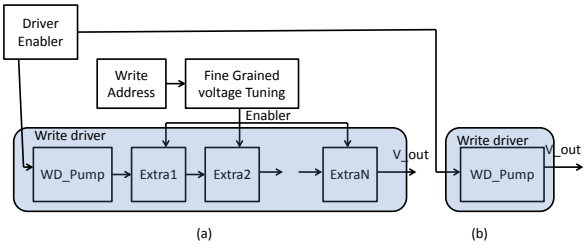


Figure 16: Write driver circuit

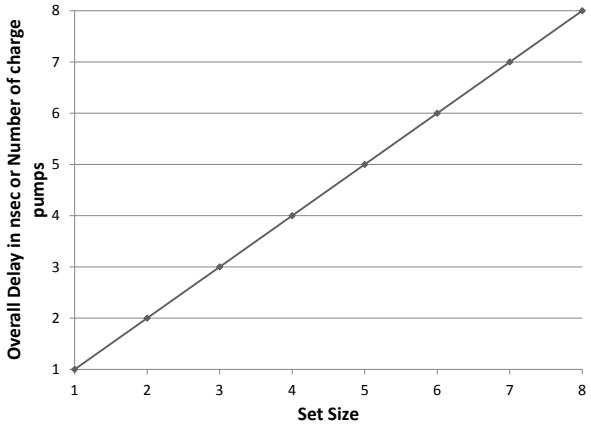


Figure 17: The overall delay in nsec or the number of charge pumps as we vary the set size, i.e., the number of layers per material.

Homogeneous stack requires multiple current magnitudes to program the cells within different layers due to the temperature variation across multiple layers. This complicates the design of the write driver circuit. Zhang and Li [31] propose to adaptively tune the voltage level, which in turn makes the programming current adjustable. This circuit is illustrated in figure 16 (a).They break the charge pump within the PCM write driver circuit down into multiple steps. The initial charge pump is used for pumping and continues until a target voltage sufficient to ensure successful programming of the cells across a layer is reached. Each extra step is then used to provide additional charge to further increase the voltage level. Thus, the number of charge pumps that are enabled is according to the required level of output voltage. The latency introduced by breaking the charge pump into multiple steps is dependent on the number of steps, i.e., the number of memory layers within the stack. Thus, assuming a 1GHz operating frequency for the charge pumps and a maximum of 10 charge pumps/memory layers, the overall introduced delay is 10ns.

In case of building heterogeneous stack, the number of different current magnitudes depends on the size of the set, i.e., the number of layers per material. In the best case where each layer is composed of different material, all layers will use similar programming current. This significantly simplifies the write circuit with only one charge pump and an overall delay of 1nsec as shown in figure 16 (b). The complexity of the driver increases as the size of set increases.

Figure 17 shows the overall delay or number of charge pumps as the size of the set increases.

In order to address variable write latencies for different materials, we explore it in the context of two 3D memory stacking organizations; traditional 3D [15] and true 3D [14] stacking. Traditional 3D stacking organization uses 2D memory dies stacked on top of each other. On the other hand, a true 3D stack has the individual bitcell arrays stacked in a 3D fashion. True 3D stacking reduces the memory access latency due to reducing the lengths of internal buses, wordlines, and bitlines. However, it results in nontrivial through-silicon via (TSV) fabrication challenges as technology scales [24].

Using multiple PCM materials within the 3D stack, whether traditional or true 3D, will result in different write latencies across different materials. In the homogeneous stack base case, a material that operates at the maximum expected ambient temperature T_{max} within the 3D stack is used. The write latency of such material is referred to as W_{max}. On the other hand, using multiple materials within the stack requires that the furthest memory layer from the heat sink safely operates under the same temperature requirement T_{max}. Consequently, the PCM material used for the furthest set from the heat sink will require a write latency W_{max}. Layers close to the heat sink in the memory stack that operates at lower temperatures will have smaller write latencies than W_{max} as shown in Figure 3.

While we tune the write latency for each set to maximize write energy savings, we assume in this work that the total line write latency is equivalent to the maximum write latency across all materials, W_{max}. This gives the same performance as the base case where all the layers use the same PCM material. This assumption is necessary for a true 3D stack since the bitcells are distributed across all memory layers. In a traditional 3D stack, however, a memory write can be fulfilled by a single layer. Thus, one can exploit the difference in write latencies across different materials to optimize memory scheduling such that frequently written pages are scheduled to memory sets with lower write latencies. Since in general PCM write latency is higher than its read latency, scheduling write request for service to a bank can still cause increased latency for later arriving read requests to the same bank [16]. Thus reducing the write latency for popular pages could in turn improve the overall performance but at the cost of stressing those layers with more writes. Similarly, the read latency is also potentially variable. The results could be affected by variable latencies and most architectural designs would need to either assume a worst-case latency for all layers or at least fit the variable latencies into discrete clock cycles. We leave the exploration for performance optimization as future work.

8 MANUFACTURABILITY OF HETEROGENEOUS 3D-STACK

In this section, we will discuss the manufacturability of building heterogeneous 3D-stack. We will focus on the following two aspects. First, the feasibility of tailoring the chemical composition within the material and their integration in 3D stack setting. Second, the manufacturing cost associated with using heterogeneous materials for 3D-stacked PCM.

The feasibility of tailoring the chemical composition of the material has been demonstrated by Yamada et al [25]. They use electron beam coevaporation method to prepare materials across the pseudo-binary line such that the deviation of each material is within 2% of those programmed. Examples of the materials prepared include $Ge_2Sb_4Te_7$, GeSb_2Te_4, $Ge_2Sb_2Te_5$, $Ge_{19}Sb_{25}Te_{56}$, $Ge_{27}Sb_{18}Te_{55}$ and $Ge_{11}Sb_{31}Te_{58}$. Their results are small-scale laboratory results; however, they provide good evidence of the feasibility of varying the chemical composition of those materials with a small percentage of deviation. Furthermore, introducing heterogeneity between layers in 3D-stack should be manufacturable, in principle, as 3D stacking enables combining different technologies [14].

The manufacturability cost of heterogeneous materials is not well-known at this point. Manufacturing dies from different materials would increase the manufacturing and testing cost. In fact, in this study we attempt to make the case for the usefulness and importance of further exploration in the manufacturing area. We think that the exploration of various materials will open opportunities for choosing the PCM material right for the use-case.

9 CONCLUSIONS

Different chalcogenide materials; with different GeTe and Sb_2Te_3 ratios; in the ternary system Ge-Sb-Te exhibit various trade-offs in terms of physical environment of the cells (such as temperature) and the properties of those memories (such as writing speed and endurance). Exploring such trade-offs opens opportunities for addressing system-level problems such as energy efficiency and reliability.

In this paper, we propose building a heterogeneous 3D PCM memory stack. We show that using a different PCM material for each memory layer leads to from 1.21X to 3.5X reduction in write energy over a single-material memory stack, depending on the number of layers in the memory stack and the programming current difference among adjacent layers within the stack. In addition, we propose an error-correcting scheme based on Hamming-based ECC that eliminates its early-failure problem. Our scheme uses different PCM material with higher endurance but lower operating temperature to store ECC bits. This eliminates the ECC early failure-problem and provides 30% improvement in lifetime over Hamming-based ECC schemes. This scheme can be combined with other error-correcting techniques to improve storage overhead and/or lifetime.

REFERENCES

[1] M. Awasthi, M. Shevgoor, K. Sudan, B. Rajendran, R. Balasubramonian, and V. Srinivasan. 2012. Efficient scrub mechanisms for error-prone emerging memories. In *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*. 1 –12. https://doi.org/10.1109/HPCA.2012.6168941

[2] Zhaoxia Deng, Ariel Feldman Stuart A Kurtz Frederic, and T Chong. 2017. Lemonade from Lemons: Harnessing Device Wearout to Create Limited-Use Security Architectures. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*. 361–374.

[3] Zhaoxia Deng, Lunkai Zhang, Diana Franklin, and Frederic T Chong. 2015. Herniated hash tables: Exploiting multi-level phase change memory for in-place data expansion. In *Proceedings of the 2015 International Symposium on Memory Systems*. 247–257.

[4] Zhaoxia Deng, Lunkai Zhang, Nikita Mishra, Henry Hoffmann, and Frederic T Chong. 2017. Memory Cocktail Therapy: A General Learning-Based Framework to Optimize Dynamic Tradeoffs in NVMs. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*.

[5] X. Dong et al. 2012. NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory. *In IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (Jul 2012).

[6] M. Ghosh and H.-H.S. Lee. 2007. Smart Refresh: An Enhanced Memory Controller Design for Reducing Energy in Conventional and 3D Die-Stacked DRAMs. In *Microarchitecture, 2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on*. 134 –145. https://doi.org/10.1109/MICRO.2007.13

[7] B. Gleixner, A. Pirovano, J. Sarkar, F. Ottogalli, E. Tortorelli, M. Tosi, and R. Bez. 2007. Data Retention Characterization of Phase-Change Memory Arrays. In *Reliability physics symposium, 2007. proceedings. 45th annual. ieee international*. 542–546. https://doi.org/10.1109/RELPHY.2007.369948

[8] HotSpot 2015. http://lava.cs.virginia.edu/HotSpot/. (2015).

[9] Sungjun Im and K. Banerjee. 2000. Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs. In *Electron Devices Meeting, 2000. IEDM '00. Technical Digest. International*. 727 –730. https://doi.org/10.1109/IEDM.2000.904421

[10] ITRS 2013. International Technology Roadmap for Semiconductors. (2013). http://www.itrs.net/.

[11] S. Kang et al. 2007. A 0.1- μm 1.8-V 256-Mb Phase-Change Random Access Memory (PRAM) With 66-MHz Synchronous Burst-Read Operation. *Solid-State Circuits, IEEE Journal of* 42, 1 (jan. 2007), 210 –218. https://doi.org/10.1109/JSSC.2006.888349

[12] Kinam Kim and Su Jin Ahn. 2005. Reliability investigations for manufacturable high density PRAM. In *Reliability Physics Symposium, 2005. Proceedings. 43rd Annual. 2005 IEEE International*. https://doi.org/10.1109/RELPHY.2005.1493077

[13] A.L. Lacaita and D. Ielmini. 2007. Reliability issues and scaling projections for phase change non volatile memories. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*. 157 –160. https://doi.org/10.1109/IEDM.2007.4418890

[14] Gabriel H. Loh. 2008. 3D-Stacked Memory Architectures for Multi-core Processors. In *Proceedings of the 35th Annual International Symposium on Computer Architecture (ISCA)*.

[15] Gabriel H. Loh, Yuan Xie, and Bryan Black. 2007. Processor Design in 3D Die-Stacking Technologies. *Micro, IEEE* 27, 3 (may-june 2007), 31 –48. https://doi.org/10.1109/MM.2007.59

[16] M.K. Qureshi, M.M. Franceschini, and L.A. Lastras-Montano. 2010. Improving read performance of Phase Change Memories via Write Cancellation and Write Pausing. In *High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on*. 1 –11. https://doi.org/10.1109/HPCA.2010.5416645

[17] Moinuddin K. Qureshi. 2011. Pay-As-You-Go: low-overhead hard-error correction for phase change memories. In *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-44 '11)*. ACM, New York, NY, USA, 318–328. https://doi.org/10.1145/2155620.2155658

[18] S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y.-C. Chen, R. M. Shelby, M. Salinga, D. Krebs, S.-H. Chen, H.-L. Lung, and C. H. Lam. 2008. Phase-change random access memory: A scalable technology. *IBM Journal of Research and Development* 52, 4.5 (july 2008), 465 –479. https://doi.org/10.1147/rd.524.0465

[19] Hebatallah Saadeldeen, Diana Franklin, Guoping Long, Charlotte Hill, Aisha Browne, Dmitri Strukov, Timothy Sherwood, and Frederic T Chong. 2013. Memristors for neural branch prediction: a case study in strict latency and write endurance challenges. In *Proceedings of the ACM International Conference on Computing Frontiers*. ACM, 26.

[20] Stuart Schechter, Gabriel H. Loh, Karin Straus, and Doug Burger. 2010. Use ECP, not ECC, for hard failures in resistive memories. In *Proceedings of the 37th annual international symposium on Computer architecture (ISCA '10)*. ACM, New York, NY, USA, 141–152. https://doi.org/10.1145/1815961.1815980

[21] Nak Hee Seong, Dong Hyuk Woo, V. Srinivasan, J.A. Rivers, and H.-H.S. Lee. 2010. SAFER: Stuck-At-Fault Error Recovery for Memories. In *Microarchitecture (MICRO), 2010 43rd Annual IEEE/ACM International Symposium on*. 115 –124. https://doi.org/10.1109/MICRO.2010.46

[22] Kevin Skadron, Mircea R. Stan, Karthik Sankaranarayanan, Wei Huang, Sivakumar Velusamy, and David Tarjan. 2004. Temperature-aware microarchitecture: Modeling and implementation. *ACM Trans. Archit. Code Optim.* 1, 1 (March 2004), 94–125. https://doi.org/10.1145/980152.980157

[23] D. Strukov. 2006. The area and latency tradeoffs of binary bit-parallel BCH decoders for prospective nanoelectronic memories. In *Signals, Systems and Computers, 2006. ACSSC '06. Fortieth Asilomar Conference on*. 1183 –1187. https://doi.org/10.1109/ACSSC.2006.354942

[24] Hongbin Sun, Jibang Liu, R.S. Anigundi, Nanning Zheng, Jian-Qiang Lu, K. Rose, and Tong Zhang. 2009. 3D DRAM Design and Application to 3D Multicore Systems. *Design Test of Computers, IEEE* 26, 5 (sept.-oct. 2009), 36 –47. https://doi.org/10.1109/MDT.2009.105

[25] N. Yamada et al. 1990. Rapid-phase transitions of GeTe-Sb_2Te_3 pseudobinary amorphous thin films for an optical disk memory. In *Journal of Applied Physics*.

[26] Doe Hyun Yoon, N. Muralimanohar, Jichuan Chang, P. Ranganathan, N.P. Jouppi, and M. Erez. 2011. FREE-p: Protecting non-volatile memory against both hard and soft errors. In *High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on*. 466 –477. https://doi.org/10.1109/HPCA.2011.5749752

[27] Lunkai Zhang, Brian Neely, Diana Franklin, Dmitri Strukov, Yuan Xie, and Frederic T Chong. 2016. Mellow writes: Extending lifetime in resistive memories through selective slow write backs. In *Computer Architecture (ISCA), 2016 ACM/IEEE 43rd Annual International Symposium on*. IEEE, 519–531.

[28] Lunkai Zhang, Dmitri Strukov, Hebatallah Saadeldeen, Dongrui Fan, Mingzhe Zhang, and Diana Franklin. 2014. SpongeDirectory: Flexible sparse directories utilizing multi-level memristors. In *Proceedings of the 23rd international conference on Parallel architectures and compilation*. ACM, 61–74.

[29] Mingzhe Zhang, Lunkai Zhang, Lei Jiang, Zhiyong Liu, and Frederic T Chong. 2017. Balancing Performance and Lifetime of MLC PCM by Using a Region Retention Monitor. In *High Performance Computer Architecture (HPCA), 2017 IEEE International Symposium on*. IEEE, 385–396.

[30] Wangyuan Zhang and Tao Li. 2009. Characterizing and mitigating the impact of process variations on phase change based memory systems. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*.

[31] Wangyuan Zhang and Tao Li. 2009. Exploring Phase Change Memory and 3D Die-Stacking for Power/Thermal Friendly, Fast and Durable Memory Architectures. In *Parallel Architectures and Compilation Techniques, 2009. PACT '09. 18th International Conference on*. 101 –112. https://doi.org/10.1109/PACT.2009.30

[32] Wangyuan Zhang and Tao Li. 2011. Helmet: A resistance drift resilient architecture for multi-level cell phase change memory system. In *Dependable Systems Networks (DSN), 2011 IEEE/IFIP 41st International Conference on*. 197 –208. https://doi.org/10.1109/DSN.2011.5958219