A Study on the Challenges and Prospects of PCM Based Main Memory Architectures

S. Rajarajan, M. Prabhu, S. Arunkumar and M. P. Karthikeyan

Department of Computer Science and Engineering
Department of Information and Communication Technology,
SASTRA University, Thanjavur, India

Abstract: For several decades, computer's memory has been volatile. High speed memories such as DRAM and SRAM have been used as main memory and cache memory respectively. Magnetic disks are used as the persistent secondary storage devices. The present DRAM based main memory has reached its energy limit. Due to the advent of nonvolatile memories like PCM, MRAM, RRAM, FeRAM and Flash, there is growing interest in incorporating them into the computers as the main memory alternates. Introducing the non-volatile storage devices in memory hierarchy is considered to reduce the energy consumption since non-volatile memories have got no power leakage in memory cells. A major hurdle in this direction is the poor the endurance limit of most non-volatile memory technologies in comparison to the RAM technologies. Among the non-volatile memory technologies, PCM is the most attractive and the most investigated option due to it advantages over other memory types. But there are few issues related to PCM that need to be addressed in order to integrate it as a potential memory for computers. There have been intense researches being carried out to devise new memory architectures based on PCM. There are also methods identified to overcome the issues of PCM. This paper analyses the various challenges in integrating PCM into the system and the various PCM based main memory architectures that have been proposed so far.

Keywords: Non-Volatile - Phase Change - PCM - PCRAM - PRAM - Hybrid Memory - Main Memory

INTRODUCTION

According to Moore’s law, the transistor density doubles every two years. This law has been sustained for several years. But it is now realized that it would be difficult to maintain Moore’s law simply by developments in transistor technology alone. There is a need for new memory structures and new materials to complement the transistor’s improvements. One of the major obstacles in increasing the memory capacity is the energy consumption by the current volatile memory technologies. In the presence of multi-core processors capable of simultaneous execution of multiple applications and with the growing demands for more memory by new applications, there is more pressure on designers to somehow expand the available memory capacity. One of the problems preventing the memory expansion is the leakage of energy that is dissipated by the main memory [1]. The amount of leakage increases with memory capacity. A number of non-volatile memories are being considered as potential solutions for this problem since non-volatile memories do not have the problem of energy leakage. Some of the non-volatile memories which are actively pursued to replace the DRAM and SRAM types are NAND flash, STT-RAM, Phase Change RAM and FeRAM. These memories are more reliable than common random access memories.

Between the PCM and the NAND flash, later have fewer write/erase cycles. Besides that, a block has to be first erased before writing into it. This results in additional delay and energy. Moreover, not being a byte addressable memory is a major limitation of flash and it is not suitable to be considered as a main memory. So it could be rather used as a disk cache [2]. According to [3], flash have been proposed to be a replacement for disk, since writes are relatively lesser and generally occur in blocks. So PCM is regarded as the number one choice for replacing DRAM in future since they are non-volatile,
more energy efficient, byte addressable and having better endurance than flash. PCM also scores when it comes to scalability. In this paper we attempt to analyze the various challenges that are faced for the inclusion of PCM as the main memory and some of solutions suggested for addressing those challenges, the various memory architectures developed on the basis of PCM technology. The remainder of this paper is organized as follows. Next section gives an overview of PCM, then the reasons for considering PCM are discussed, lists of the major challenges are given, description of the various PCM memory architectures are provided, discussion on the results of the review are presented and finally the conclusion of the paper.

**Phase Change Memory (PCM):** Phase Change Memory (PCM) has been a strong contender as a replacement for the conventional DRAM for the main memory. It is also being considered to be used at other levels of the memory hierarchy. For example, there memory models made using PCM as a cache [4]. A PCM cell stores bits not with electrical charge, but switching a fragment of chalcogenide glass between amorphous and crystalline states, which have different resistivity. To program a single bit, PCM memory heats a small amount of chalcogenide and then cools it either slowly or quickly, depending on the value being programmed [5]. The difference in cooling speeds creates different structural patterns in the material, each with varying electrical resistance. Heating the phase change material to crystallize it is called SET operation and melt-quenching it to make it amorphous is called RESET operation. Fig. 1 shows a PCM cell.

In order to read back a bit, current passed into the material and depending on how quickly the chalcogenide gets cooled, it encounters either high resistance or low resistance. PCM is bit alterable in the sense that each of the bits can be toggled independently, eliminating the need for the extra erase operation that is needed in the case of a flash memory.

Two of the most encouraging memory technologies are the flash and the PCM. PCM overcomes many of the restrictions that flash poses. Unlike Flash memory, PCM is byte addressable and needs no erase operation. Even though PCM’s durability is 10-100 that of flash, it still requires some wear leveling. But due to the lack of an erase operation and hence the avoidance of the read/write, wear leveling in PCM is much easier.

There are schemes that consider the process variations of the PCM’s physical dimensions for alleviating the write endurance issue [6]. Similar to flash memory, PCM’s write operations are longer than its reads, but the differences are much smaller. In most respects PCM is best thought of as a slightly slower DRAM. PCM is expected to eventually match DRAM’s latency and bandwidth, continuing recent trends of fast density and performance increases and may become a viable technology to replace DRAM as a main memory technology [7]. PCM’s bit alterability makes it ideal for replacing main memory in systems. Nevertheless, since writes to storage often occur in larger blocks, making use of byte addressable memory may not be an advantage, especially if it incurs additional cost.

In a flash, data is stored in memory cells that are made by floating gate transistors. A PCM on the other hand stored information in a PCM cell which is in a cell material state. It can be amorphous or crystalline [8]. A PCM cell consists of a chalcogenide alloy material along with a small heater. This is shown in Fig. 1. Reading a PCM cell incurs less power consumption, since no heating is involved. The requirement of refresh power is not there as it retains its content persistently and [9] shows that it dissipates much lesser power in its standby due to its meager leakage. Fig. 2 shows an array of PCM cells.
An important property of PCM is Threshold switching [10]. High voltage is required to convert from a crystalline state to the amorphous. But above a certain threshold voltage, the material's conductivity which is in the amorphous state elevates drastically. This result in generation of large current and the material is heated by that. If the current pulse is turned off at a threshold voltage, the material relapses to its high-resistance amorphous state.

Why PCM?: Two of the memory technologies that are widely considered to be an alternate for DRAM as main memory are the PCM and the flash. In comparison with a flash, PRAM has the potential offer much higher performance when it is needed that writing should occur fast. The first reason for this is that the memory element can be switched instantly and changing bits does not require erasing of an entire block of cells. With the Flash, cell degradation occurs whenever voltage flows across the cell. Although PRAM devices also happen to degrade with use, the degrading occurs much more slowly. A PRAM device may sustain around 100 million write cycles in its life time. There is also a power issue that advocates PCM as a potential DRAM replacement. DRAM found to be poor energy efficient technology. This is not just due to its periodic refresh which takes place to retain the stored bit in the capacitor, but also due to the simultaneous addressing of multiple banks within the chip. Whenever a bit passes through a DRAM chip, 8 or even 16 devices are being internally accessed to read and to write again But the inherent need to rewrite after each read access is inevitable for a volatile memory like DRAM. So being a non-volatile memory, PCM could potentially offer a lower-power alternative to DRAM. PCM also competes favorably with DRAM in terms of forward scaling into future generations, as DRAM has reached various scaling limits. PCM exhibits lot of highly desirable characteristics, such as rapid state transition, better data retention and potential to be scaled for ultra-small devices in future. PCM also has the capability to store multiple bits per cell.

Challenges of Phase Change Memory

Lower Write Endurance: Phase-change memory has poor write endurance. Repeated writes to PCM cells cause them to wear, reducing its lifetime. The current flow required to write a PCM cell has a degrading effect on the contact between the electrode and storage area. This increases the current variance and subsequently increasing the resistance variability. As a result, the PCM can only endure a limited number of writes. The actual endurance varies based on manufacturing techniques, but normally on the order of \(10^7\) to \(10^9\) writes. The lifetime of the PCM varies with its applications. This is termed as a hard error. This quality of PCM makes it a poor candidate as a replacement for main memory. Going by the locality of references property, memory writes are tend to cluster. This further decrease the lifetime of PCM, since a certain group of cells go through frequent writes, causing damage to those cells. Once a single PCM cell can no longer be considered reliable, it may render an entire page or even the entire memory unusable, depending on whether some error recovery scheme is in use. So some mechanism is needed to distribute the write operations evenly across the whole memory, so that wear will also gets distributed, increasing overall PCM lifetime significantly. The requirement for write endurance improvement through wear leveling can be computed using with the help of the following Eq.(1) [12].

\[
E = T_{sys} \frac{B}{\alpha C}
\]

Where E is endurance, \(T_{sys}\) is the expected life of the system, B is memory-bandwidth, \(\alpha\) is the efficiency of wear-leveling and C is the system’s memory capacity.

Wear leveling is considered to be the technique for achieving this by enforcing an even usage of all storage cells of the device. The lazy-Organization implemented by [13], decreases the amount of writes that takes place to the PCM. Whenever a page fault error is handled, the page that is fetched from the secondary storage is written only to the DRAM Cache instead of the PCM. This avoids the initial write in case of the dirty pages. Line level writes [13] is another technique in which writing to the PCM memory is reduced to occur in smaller chunks rather than an entire page. This is achieved by tracking a

<table>
<thead>
<tr>
<th></th>
<th>DRAM</th>
<th>PCM</th>
<th>NAND Flash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Volatile</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Erase Mechanism</td>
<td>Bit</td>
<td>Bit</td>
<td>Block</td>
</tr>
<tr>
<td>Software complexity</td>
<td>Simple</td>
<td>Simple</td>
<td>Complex</td>
</tr>
<tr>
<td>Write-Latency</td>
<td>~20–50ns</td>
<td>~1µs</td>
<td>~100µs</td>
</tr>
<tr>
<td>Read-Latency</td>
<td>~20–50ns</td>
<td>50 to100 ns</td>
<td>10-25 µs</td>
</tr>
<tr>
<td>Read-Energy</td>
<td>~0.1nJ/b</td>
<td>&lt;&lt;1 nJ/b</td>
<td>&lt;&lt;1 nJ/b</td>
</tr>
<tr>
<td>Idle-Power</td>
<td>~W/GB</td>
<td>&lt;=0.1W</td>
<td>&lt;=0.1W</td>
</tr>
<tr>
<td>Memory Endurance</td>
<td>(10^4)</td>
<td>(10^6)</td>
<td>(10^5) to (10^9)</td>
</tr>
<tr>
<td>Cell Area</td>
<td>(6F^2)</td>
<td>(5\sim8F^2)</td>
<td>(5F^2)</td>
</tr>
<tr>
<td>Byte-addressable</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
page at the level of a processor’s cache line and then the amount of writes to the PCM can be minimized by writing only the “dirty” lines within a page instead of writing the entire page. In another approach [14], the memory is partitioned into DRAM and PCM. The memory controller maintains a map of the number of writes to the PCM. If the preset endurance limit is reached by the write-count of a page, the controller invokes a ‘bad-page’ interrupt for that page. This is handled by the page handler which moves the page from its free list into a bad-page list. The pages found in the list of bad-page are ignored and not involved for future allocations.

In an implementation of PCM based main memory, [15] achieved wear leveling by row shifting. A shifting mechanism is employed to distribute the write operations within a row to all the cells in the row. For this, any two memory pages from each category of hot, medium-hot, medium-balanced and unbalanced pages are selected for a workload. These pages are chosen based on their write counts and the standard deviations of the writes among all the lines in a page. The row shift interval varied from 0 (no shift) to 256 (shift one byte on every 256 writes) and collected the resulting lifetimes averaged from all selected sample pages. The second level of wear leveling implemented through segment swapping that periodically swaps the memory segments with higher and lower write accesses. Partial writes is a technique proposed by [7]. This aims to decrease the amount of writes to the PCM’s array by tracking dirty-data from the level1 cache to the memory banks. Whenever a buffered row is removed and content written to the PCM array, only the modified data is written. This occurs at two granularities: cache line size and word size. Start-Gap [16] is another technique that makes use of algebraic mapping of logical and real addresses. It does randomization of address space to achieve wear leveling. In spite of the wear leveling techniques, there could still be some cells whose write endurance has dropped lower than a threshold level. In such case, salvaging methods are employed to extend the lifetime of the PCM [17].

**High Write Latency:** PCM’s write latency is much higher than the read latency incurring 90ns–150ns for a single cell [5]. The write latency of PCM is based on the SET time which is three times slower than the read. Besides, PCM’s higher write power results in writing operation taking many rounds with a part of a line written in each round. So the write latency of PCM is further extended. This has resulted in several disadvantages in the applications of PCM. There is a need for more parallelism in the implementations of PCM writes to overcome this problem. The method implemented by [4] called read-after-write is aimed at addressing this issue. This eliminates reduces number of bit written by choosing only the bits which are changed and needs to be updated. They have also implemented the technique of data inversion. When ever the new data value is to be written to the cache, its old value is read and a hamming-distance between the two is computed. In case, the size of the hamming distance that is just calculated is larger than half of cache-block, then the new value is inverted and the inversion status bit is set. So the original data value may be recovered based on the inversion status bit.

Fig. 3 depicts the design of read-after write with data inversion [4].
Table 2: RESET/SET Pulse parameters[22]

<table>
<thead>
<tr>
<th>Technology node (nm)</th>
<th>SET current Amplitude(µA)</th>
<th>Duration (ns)</th>
<th>RESET current Amplitude(µA)</th>
<th>Duration (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>500 – 600</td>
<td>500</td>
<td>800 – 1000</td>
<td>50</td>
</tr>
<tr>
<td>120</td>
<td>200 – 500</td>
<td>180</td>
<td>600 – 620</td>
<td>10</td>
</tr>
<tr>
<td>180</td>
<td>N / A</td>
<td>150</td>
<td>600</td>
<td>10</td>
</tr>
</tbody>
</table>

with no energy leakage. Moreover, unabated energy consumption could result in increased temperature. So the power cost controlling of PCM is a very important issue. As presented in Table 2, the SET operation incurs more energy consumption than the RESET operation [21]. This characteristic may be made use of to reduce the high energy cost of write operations.

A common technique to reduce energy consumption during write operation is called read-before-write scheme [4]. It essentially means that the target bit to be written will be first read, then it will be compared with the new value to be written and a write will take place only if the existing value is different from the new value. According to [20], a power token is introduced. A calculation of the maximum watts of power that a PCM chip could supply through its power pins and the power requirement for writing into a single bit is made. Then each bit is associated with a power token which tells the power requirement to write into that bit. The memory controller maintains a collection of power tokens and as writes occur the number of tokens reduces. Once the writes are completed the tokens once again restored. For every write operation, the memory controller should verify whether sufficient power tokens are available to accomplish that operation. If there are not enough tokens available then the write operation will not be carried out. In another approach, [7] made use of several row buffers inside a PCM chip. This decreases the write energy consumption through write coalescing. Multiple writes occurring at the same location are absorbed into the buffers and thereby resulting in much smaller number of write-backs to the PCM array.

Pcm Based Memory Architectures: There are numerous memory architectures proposed involving PCM. But not all of them have proposed PCM as part of or as a sole main memory. For example, Reference [4] has considered PCM as a replacement of the cache memory. But in our analysis, we are focusing on the research works which have proposed PCM either as an alternate main memory or as a hybrid memory along with DRAM. A hybrid DRAM approach adds a PCM with some kind of a DRAM and allows the, to work together. The specific role of each memory differs across research works.

Hybrid Memory System with PRAM and DRAM: A hybrid main memory system consisting of PRAM and DRAM in the name of PDRAM is proposed [14]. In this hybrid architecture, PRAM provides low read and standby power and DRAM provides higher write endurance and lower write power. The two parts of this implementation is the hardware portion which is the memory controller that manages the access to various PRAM pages and the software portion is the operating system’s memory manager which performs the wear leveling to counter write endurance problem. The partitioning of system’s memory between DRAM and PRAM is known to the memory controller. So it will forward the request the appropriate memory based on the address given. Fig 5 shows the model of this architecture.

Scalable High Performance Main Memory Based on PCM: Another memory model was implemented as a hybrid memory system using a PCM memory with a
DRAM buffer [13]. In this memory organization, the operating system administrates the PCM. OS maintains page tables for this purpose. In this organization, DRAM is used like a cache that is invisible to the OS. The performance gain achieved is by reducing the amount of page faults against a base-line DRAM memory of 1/4th capacity. Such a system employs a DRAM module as a cache for the PCM memory and attempts considerably reduce the relatively long response time of PCM memory. But still the access latencies of the hybrid memory cannot be made on par to a DRAM only system. There are independent memory-controllers used for DRAM and PCM subsystems of the main memory. This organization tolerates an order of magnitude slower write latency of PCM with a write queue and tackles the endurance problem using the techniques of Lazy Write, Line Level Write-back and Page Level Bypass to limit the number of writes to the PCM memory. This model is depicted in Fig. 6.

**Energy Efficient Main Memory System Based on PCM with DRAM Acting as the Cache:** In another PCM based main memory system design, a DRAM is used as a cache replacing the row buffers of the PCM memory [21]. There is a reduction in the high latencies of PCM and subsequently the power consumption of the main memory is also reduced. There is also no need for a memory controller. A logical cache with the line size of one kilo bytes is physically dispensed across eight devices in a single rank. This result in individual caches of smaller size and the directory of cache tags is maintained at the memory controller. For each memory access request, a search in the cache is carried out to verify for the availability of data in the PRPB cache. If it is a hit, then the request for accessing the data is treated like a in a normal memory module. An additional control signal line that is present between the memory controller and the PCM is used for signaling the PCM device in order to provide the data from the cache instead of doing a PCM access. An advantage of this approach is that it does not require a separate memory controller. Fig. 7 represents this model.

**Energy Efficient Main Memory Using PCM by Redundant Write Elimination:** Reference [15] have studied the usage of PCM as alternate main memory for a three dimensional stacked-chip. They have also effectively addressed issues of PCM endurance limitation, higher access latencies and higher dynamic power by various techniques. Through 3D integration, the designers are able to place cache or main memory on a die that consist of multiple processors. This is expected to decrease the memory access latency and helps to improve the memory bandwidth in CMPs. But a 3D integrated chip is restrained by tight power and thermal constraints. So using PCM which has almost no energy leakage can greatly reduce the energy consumption in the main memory cores. To reduce the memory writes in order to improve the endurance of PCM, the redundant bit writes are identified and they are eliminated. This is done by comparing the existing bit values with the values to be written and avoiding the write when the values are the same. Since reads are much faster than the writes in a PCM, the additional delay incurred the read before write is not significant. There is also wear leveling row shifting and segment swapping. They have reported that the PCM lifetime is increased from a short duration of 176 days to a longer duration of 22 years for the single-level PCM cells because of their methods. Fig 8. is the graph of the improvement made in PCM energy consumption according this model.
In this paper we reviewed the potential of PCM replacing or partnering the DRAM which is today the sole form of main memory. We have found that there are few challenges to be addressed before contemplating on a PCM based memory architecture. Many of those challenges are effectively mitigated by many researchers using various techniques. There is a larger question of whether PCM based independent main memory or PCM based hybrid memory with DRAM is more viable. We feel that in the present situation, it is more appropriate to consider PCM based hybrid memory architectures. There is still need for more research to further improve the write endurance limits and the energy savings to make PCM competitive with conventional RAM technologies. Another area where there is need for more research is in the development of non volatile memory compatible software and algorithms. Reference [23] have illustrated that many of the common algorithms are not suitable for PCM. A file system designed specifically for non-volatile and byte addressable PCM memory [24]. The algorithms meant for PCM have to be economical with energy consumption [25].

CONCLUSION

In the pursuit of uplifting the system performance, the limitations of DRAM have become an obstacle. Replacing the DRAM by a non-volatile memory appears to be a viable solution and several researches are going on in this direction. In our paper we reviewed a number PCM based main memory architectures that have been proposed so far. We identified the major challenges and discussed the proposed solutions too.

REFERENCES


