A Read-Write Aware Replacement Policy for Phase Change Memory

Xi Zhang, Qian Hu, Dongsheng Wang, Chongmin Li, and Haixia Wang

Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science & Technology, Tsinghua University,
Beijing 100084, China
{zhang-xi06,hq07,lcm03}@mails.tsinghua.edu.cn,
{hx-wang,wds}@tsinghua.edu.cn

Abstract. Scaling DRAM will be increasingly difficult due to power and cost constraint. Phase Change Memory (PCM) is an emerging memory technology that can increase main memory capacity in a cost-effective and power-efficient manner. However, PCM incurs relatively long latency, high write energy, and finite endurance. To make PCM an alternative for scalable main memory, write traffic to PCM should be reduced, where memory replacement policy could play a vital role.

In this paper, we propose a Read-Write Aware policy (RWA) to reduce write traffic without performance degradation. RWA explores the asymmetry of read and write costs of PCM, and prevents dirty data lines from frequent evictions. Simulations results on an 8-core CMP show that for memory organization with and without DRAM buffer, RWA can achieve 33.1% and 14.2% reduction in write traffic to PCM respectively. In addition, an Improved RWA (I-RWA) is proposed that takes into consideration the write access pattern and can further improve memory efficiency. For organization with DRAM buffer, I-RWA provides a significant 42.8% reduction in write traffic. Furthermore, both RWA and I-RWA incurs no hardware overhead and can be easily integrated into existing hardware.

1 Introduction

Chip multiprocessor (CMP) is proposed to maintain the expected performance advances within the power and design complexity constraints. Future CMP will integrate more cores on a chip, and the increase in the number of concurrently running applications (or threads) increases the demand on storage. However, using DRAM as main memory is expensive and costs a significant portion of system power. Furthermore, the scaling of traditional memory technology is already hitting the power and cost limits [1]. It is reported that scaling DRAM beyond 40 nanometers will be increasingly difficult [2]. Hence, exploiting emerging technology is crucial to build large-scale memory system within the power and cost constraint.

* This work is supported by Nature Science Foundation of China (No. 60833004, 60970002), and the National 863 High Technology Research Program of China (No.2008AA01A201).
Two promising technologies that fulfill these criteria are Flash and Phase Change Memory (PCM). NAND flash can be only used as a disk cache or an alternative for disks for it is not byte-addressable and is about 200X slower than DRAM. PCM, on the other hand, is only 2x-4x slower than DRAM and can provide up to 4x more density than DRAM, which makes it a promising candidate for main memories. PCM has been backed by key industry manufacturers such as Intel, STMicroelectronics, Samsung, IBM and TDK as both off-chip and on-chip memory. Table 1 lists features of several memory technologies: Static RAM (SRAM), Dynamic RAM (DRAM) and PCM based on the data obtained from literature.

There are several challenges to overcome before PCM can become a part of the main memory system. First, PCM being much slower than DRAM, makes a memory system comprising exclusively of PCM have much increased memory access latency. Thus, several studies proposed a properly designed main memory that is made of different memory technologies. Secondly, for its non-volatile property, PCM is unfriendly to write access. Compared to read access, write access to PCM costs too much power and latency to accomplish. In addition, the limited system lifetime due to endurance constraint is still a concern, for PCM devices can only tolerate $10^8 - 10^9$ writes per cell. Therefore, the write traffic to PCM should be optimized in purpose of decreasing access latency, power and increasing lifetime.

In this paper, we focus on a memory replacement policy to decrease write traffic to PCM. Nowadays most systems use LRU or its approximations for on-chip cache replacement policy and main memory replacement policy. However, systems with PCM as main memory require a new replacement policy which considers not only hit rate but also the replacement cost. Since write access to PCM necessitates more time and energy than read access, a new replacement policy should be designed for memory level above PCM to filter out more write access requests than existing policies. Inspired by the recently proposed RRIP, we propose a Read-Write Aware (RWA) replacement policy, which takes into consideration the imbalance of read and write costs upon eviction. The basic idea of this method is to preserve a certain amount of dirty data lines to stay longer than clean data lines in upper level memory of PCM, thereby decreasing write backs to the lower level PCM. However, this imbalance policy may lead to

---

**Table 1. Comparison of Different Memory Technologies**

<table>
<thead>
<tr>
<th>Features</th>
<th>SRAM</th>
<th>DRAM</th>
<th>PCM</th>
<th>NAND FLASH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>1X</td>
<td>4X</td>
<td>&lt;16X</td>
<td>16X</td>
</tr>
<tr>
<td>Read Latency *</td>
<td>2°</td>
<td>2°</td>
<td>2^11</td>
<td>2^11</td>
</tr>
<tr>
<td>Dyn. Power</td>
<td>Low</td>
<td>Medium</td>
<td>Medium read</td>
<td>Very high read</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High write</td>
<td>Even high write</td>
</tr>
<tr>
<td>Leak. Power</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Non-volatile</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Scalability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Endurance</td>
<td>N/A</td>
<td>N/A</td>
<td>$10^8 - 10^9$</td>
<td>$10^*$</td>
</tr>
<tr>
<td>Retention</td>
<td>Constant power</td>
<td>Refresh</td>
<td>10yrs</td>
<td>10yrs</td>
</tr>
</tbody>
</table>

*Order of magnitude, in terms of processor cycles for a 4GHz processor.