A Memory-Disk Integrated Non-volatile Memory System with its Dual Buffering Adapter

Mei-Ying Bian
Dept. of Computer Science
Yonsei University
Seoul, Korea
jojo04592@gmail.com

Su-Kyung Yoon
Dept. of Computer Science
Yonsei University
Seoul, Korea
mypioioi@naver.com

Shin-Dug Kim
Dept. of Computer Science
Yonsei University
Seoul, Korea
sdkim@yonsei.ac.kr

Abstract

In this paper, conventional main memory and disk storage layers are merged into a single memory layer using a combination of PRAM and NAND Flash memories, which is called as an integrated memory-disk (IM-D) non-volatile memory structure. The IM-D memory architecture consists of a dual buffering IM-D adapter, an array of SLC/MLC PRAM/Flash hybrid IM-D memory, and an associated memory management module as IM-D access translation layer in the operating system. In this paper, we focus on two hardware modules, i.e., a dual buffering IM-D adapter and an array of PRAM/Flash hybrid IM-D memories. Also, even though non-volatile memories such as PRAM and NAND Flash show advantages in low power consumption, higher density, and non-volatility compared with DRAMs, access latencies of such non-volatile memories are too slow to simply replace conventional DRAMs and disks. Thus, the proposed structure is designed to enhance the lifetime and hide asymmetric read/write access latencies. The experimental results show that overall storage capacity remains the same level and overall access latency decreases by around 49.6 times. Moreover, our structure can significantly save the energy consumption by 83.51%. Based on this result, our proposed memory system offers the possibility of replacing conventional combination of DRAM and HDD/NAND Flash structure based mobile device system with IM-D structure.

Keywords-mobile device storage; memory hierarchy; cache and buffering; integrated memory disk structure; dual buffering adapter

1. Introduction

In daily life, we can easily see most of the people are playing with some types of mobile devices. Specifically, mobile device has become a dominant computing device for a wide variety of applications. Furthermore, mobile devices tend to quietly replace the PCs (personal computers) as the most common web accessing device. People can view the weather information on the internet using a mobile smart phone, or to play games with tablet PC. Thus, gradually it requires high demands on mobile device performance, such as shorter response time, lower power consumption, and mass storage capacity.

Recently, many mobile devices, such as smart phones, tablet PCs, and personal digital assistants (PDAs), use NAND Flash memory for their storage system due to its non-volatile, low power consumption, and high random access speed characteristics, and use DRAMs as its main memory. However, as shown in Table 1 [3, 5, 6], poor density of DRAM makes it uncompetitive, compared to the PRAM memory. Moreover, the activation/pre-charge power is required to open and close a row in the memory array. This one is higher for the case of DRAM, since it has to refresh the row data before closing a row, which can be avoided in PRAM structure due to its non-volatile characteristics. The standby power, consumed when it is idle, is also lower for PRAM than DRAM, due to its negligible leakage power consumption. For NAND Flash memory, a write operation can only be performed on an empty or erased unit, so in most cases write operation must be preceded by an erase operation. Also a block is the minimum unit to implement erase operations, and this is not required for PRAM, which supports fast byte or word access capability without erase-before-program requirement like DRAMs. Thus, considering the above aspects, PRAM has emerged as a promising candidate for main memory. However, in terms of price, PRAM has not been reached to achieve general commercialization. Therefore, a hybrid model of PRAM and Flash memories should be considered as a currently feasible model for cost effective mass storage. And also conventional role of main memory and disk layers needs to be reconsidered as new non-volatile memories are emerged.

In this research, conventional main memory and disk storage layers are merged into a single memory layer using a combination of two non-volatile memories, such as PRAM and Flash memory as an example candidate of our first trial, which is called an integrated memory-disk (IM-D) non-volatile memory structure. In this approach, files are stored in the non-volatile IM-D structure and are accessed in place as if they were in conventional DRAM main memory, without any duplication between them. Conventional file access and memory access mechanisms must be changed at the operating system layer to be transparent to the new IM-D memory structure. The overall IM-D architecture is based on a dual buffering IM-D adapter, an array of PRAM and Flash hybrid memories, and an associated memory management module in the operating system with its IM-D access translation layer. In this paper, we focus on the two hardware components, i.e., a dual buffering IM-D adapter and an array of PRAM/Flash hybrid memory. As basic components for the IM-D structure, write throughput for SLC PRAM is much faster than that for MLC PRAM, and MLC NAND Flash density is double of SLC NAND FLASH. Especially, we exploit fast SLC PRAM access feature and large MLC NAND Flash capacity to provide lower power consumption and mass memory/storage space, with a small amount of DRAM buffering space for the IM-D adapter to provide short response time and long storage endurance.

parameter	DRAM	NAND flash	PRAM
Density	1X	4X	4X
Read speed	60ns	25us	200-300ns
Write speed	1Gbps	2.4MB/s	100MB/s
Endurance	N/A	10^4	10^6 to 10^8
Read Energy	0.10uJ/(4KB)	9.4uJ/(4KB)	0.07uJ/(4KB)
Write Energy	0.14uJ/(4KB)	59.6uJ/(4KB)	3.80uJ/(4KB)
Erase Energy	N/A	1056/(256KB)	N/A

Table 1: Characteristics comparison among DRAM, NAND flash and PRAM

Thus, the first goal of this work is to evaluate the effect of non-volatile IM-D structure in terms of performance, power consumption, and overall cost. Experiments show that overall storage capacity remains the same level and overall access latency decreases by around 49.6 times. Moreover, our structure can significantly save the energy consumption by 83.51%. Therefore, the proposed non-volatile IM-D structure can achieve more effective memory capacity and performance.

The rest of this paper is organized as follows. Section 2 describes background and related work. And non-volatile IM-D adapter with a small amount of DRAM space based dual buffering structure is designed and related operational flow is presented in Section 3. Also evaluation is provided in Section 4. Finally, Section 5 concludes this research.

2. Background and Related work

In this section, we briefly describe the characteristics of non-volatile memory usage in mass storage systems, in terms of NAND Flash and PRAM.

2.1 Basic Concepts of NAND Flash

A NAND Flash memory chip is based on a fixed number of independent blocks, and each block holds a fixed number of pages. Pages are the smallest programmable units, each page has a data area as well as spare area. The spare area is often used to store management information or error correction codes. There are three basic operations available in NAND Flash memory, such as read, write, and erase respectively. NAND flash memory can be read in a page unit at a time, a write operation can only be performed on an empty or erased unit, so in most cases write operation must be preceded by an erase operation, and a block is the minimum unit to implement erase operations. To avoid performance degradation caused by frequent erase operations, a common solution is to write an updated page to a pre-erased block under the management of a flash translation layer (FTL) receives read and write requests and maps a logical address to a physical address in NAND Flash [8]. The block that undertakes the updated written page is called the log block, and the original block is known as the data block [1]. When there are no more log blocks to be allocated the system triggers reclamation, which has three types of merge operations, such as full merge, partial merge, and switch merge respectively [4].

When using NAND Flash memory to support mass storage, an example is flash-aware write buffer scheme (FAWB) [2] which proposed by Lu et al. FAWB based on the superblock-based NAND Flash memory storage system architecture, FAWB monitors the current running information of the NAND storage system to manage buffer coordination with the flush operation and obtain optimized writing performance.

2.2 Basic Concepts of PRAM

PRAM's high performance, such as 4x denser than DRAM and non-volatile characteristics makes it particularly interesting in non-volatile memory. However, there are still several challenge required to be overcome before PRAM can be replaces DRAM as a next generation main memory. For example, the read latency of PRAM is 2x-4x slower than DRAM and the write latency is about an order of magnitude slower than read latency. Moreover, PRAM cell can only endure a maximum of 10^8 writes. There has been much subsequent work on chip technology, devices and materials for PRAM, in order to overcome long access latency and wearable lifetime.

Qureshi et al [3] proposed a PRAM-based hybrid main memory system, the PRAM storage is managed by the operating system (OS) using a page table, in a manner similar to current DRAM main memory systems. The DRAM buffer is organized similar to a hardware cache which is not visible to the OS, and is managed by the DRAM controller. Afterwards, a lazy-write organization is proposed when a page fault is serviced, where the page fetched from the hard disk is written only to the DRAM buffer. Specifically only if it is evicted from the DRAM storage, only dirty pages will be written back to PRAM main memory.

Jung et al [4] proposed a PRAM-based memory using a superblock-based adapting buffer, located between on-chip cache and main memory. The adapting buffer is comprised of decoupled dual buffers: one is to utilize spatial locality aggressively (SLSB), and the other exploits temporal locality adaptively (TLAB). SLSB memory adapter aggressively prefetch a set of pages from the main memory, and the blocks that are accessed once before being evicted from the SLSB will promote into the TLAB. For write request to the SLSB, the blocks are moved to the TLAB, where can hide the write latency of the slower PRAM memory medium while increasing the probability of a hit on access by last level of cache.

3. Main Architecture

In the proposed IM-D structure, conventional main memory and disk storage layers are merged into a single memory layer by using PRAM-Flash hybrid approach. Namely, file accesses are converted into memory accesses via virtual address space. Thus, files are stored in the non-volatile IM-D structure. Conventional file access and memory access mechanisms must be changed at the operating system layer to be transparent to the new IM-D structure. This approach also avoids conventional page swapping without the data duplication as in conventional systems, where programs are accessed in place as if they were in conventional DRAM main memory. But conventional file access mechanism is still maintained as the same externally at the operating system level, even though accessing functions are transformed into IM-D storage accesses. The proposed IM-D structure can be formed as three major components as in the following subsections.

3.1 Overall Structure

The overall IM-D structure consists of a dual buffering IM-D adapter, an array of SLC/MLC PRAM/Flash hybrid IM-D memory, and an associated memory management module as IM-D translation layer. In this paper, we focus on the two hardware modules, i.e., a dual buffering IM-D adapter and an array of PRAM/Flash hybrid IM-D memories as shown in Figure 1. Also this research is to propose how to design basic IM-D structural model with its operational flow for simple mobile devices. To design an effective single IM-D layer by using current non-volatile memories, similar performance and cost to those of conventional DRAM main memory and disk hierarchy, should be achieved, where non-volatile memory advantage can be provided additionally. If advanced non-volatile memory components become reality in the future, then significant performance and cost advantages can be achieved by the proposed approach. However, by using currently available non-volatile memories, we must enhance the read access latency of IM-D structure which is about 2x-4x slower than that of DRAM. We must also hide the write access latency which is about 10x slower than DRAM. In addition to achieving read/write access latency comparable to the DRAM, we must extend the life endurance of IM-D structure to be similar to that of DRAM and disk combination. In this paper, we focus on only two hardware modules, e.g., a dual buffering IM-D adapter and an array of

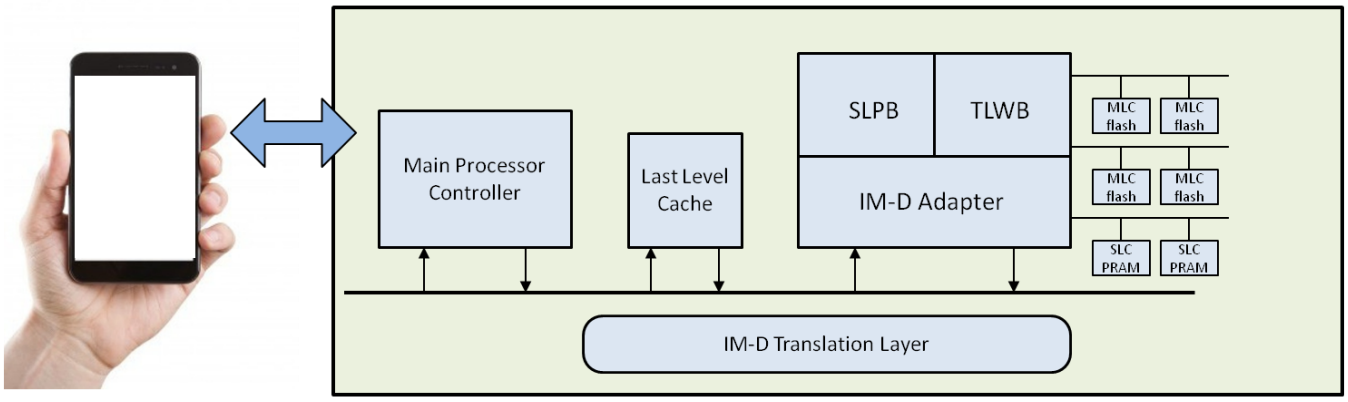

Figure 1. PRAM-Flash based integrated memory disk hierarchy

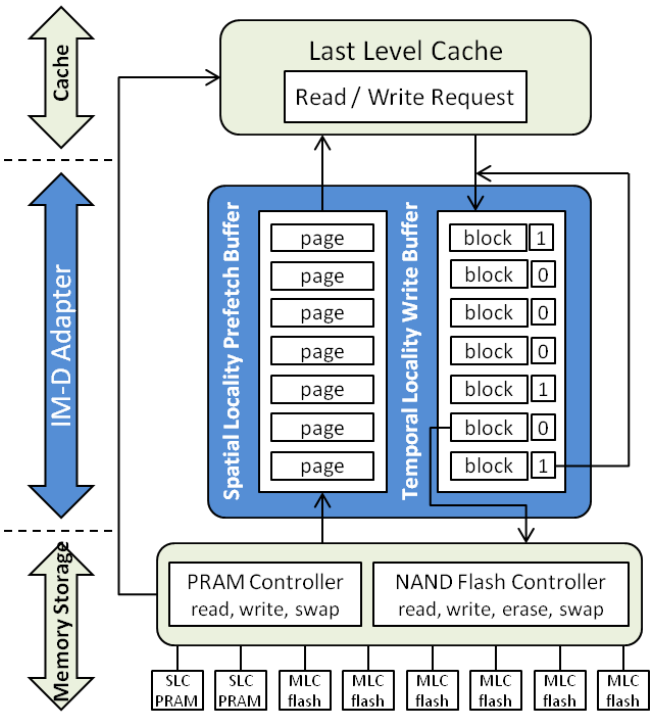

Figure 2. DRAM based memory adapter

PRAM/Flash IM-D memories as shown in Figure 1.

First module is a dual buffering IM-D adapter, where two heterogeneous buffers are located between the last level of cache and the IM-D structure. In order to provide low read latency, when a miss occurs from the last level of cache, a page unit of data, corresponding to the requested data block, will be directly prefetched into the first buffer and the requested data block will be fetched into the last level of cache. Also to hide the long write latency of IM-D structure, when a dirty data block is evicted from last level of cache, it will not be written back directly to the array of PRAM-Flash hybrid IM-D memory, but to the second buffer first. Detailed operation is given in the next subsection.

Second module is an array of PRAM/Flash hybrid IM-D memory. Several types of PRAM/Flash memories can be integrated together to support proper performance and endurance, as basic components for the IM-D structure. Specifically, write throughput for SLC PRAM is much faster than that for MLC PRAM, and MLC NAND Flash density is double of SLC NAND FLASH. Thus, fast SLC PRAM is chosen to provide shorter access time and longer endurance, and the cost-effective MLC Flash memory is chosen to provide mass storage space in the proposed IM-D structure.

3.2 Structure of A Dual Buffering IM-D Adapter

Two heterogeneous buffers are designed to hide longer access latency and support proper endurance. One is the spatial locality based prefetch buffer (SLPB) and another one is the temporal locality based write buffer (TLWB), where both can be implemented by using DRAMs and managed as in first in first out (FIFO).

The SLPB is designed to prefetch a page unit of data from the array of PRAM/Flash hybrid IM-D memory. When a miss occurs in the last level of cache, a single page corresponding to the requested data is prefetched into the SLPB and its associated data block will be fetched into the last level of cache. Moreover, when a hit occurs in the last level of cache, then check if its next one block ahead exists in either cache layer or adapter layer. If there is not its next block, a single page corresponding to next one block ahead is prefetched into the SLPB.

To extend the endurance of the IM-D structure, we exploit the TLWB as a write buffer by decoupling the buffers in the IM-D adapter, where management unit is a cache block size. As shown in Table 1, PRAM writes are much slower than read operations and need more power consumption, when handling write operations. In addition, both PRAM and Flash memories show limited write endurance, where each write back wears the device endurance. To compensate these negative impacts on the write access latency and limited endurance, we need some mechanisms to minimize a certain degree of write count to the array of PRAM/Flash hybrid IM-D memory structure. In order to reduce the number of write operations, we design the TLWB management in a more flexible way. Especially, dirty blocks being evicted from the last level of cache will first be written back to the TLWB, where long write latency for the array of PRAM-Flash hybrid IM-D memory can be hidden, while increasing the probability of buffer hits to be accessed by last level of cache. To keep track of each block existing both in the last level of cache and the TLWB, the tag directory for the TLWB is extended with a “hit” (H) bit. If a last level cache miss occurs and its associated block exists in the TLWB, then its corresponding H bit in the TLWB tag directory is set to 1. Figure 2 shows an example. When the TLWB is filled with blocks completely, then a victim process is performed, where the most recently accessed blocks can be maintained long enough in the TLWB. Therefore, if a candidate victim block is identified as the accessed one (Hit bit is 1), it will be moved to the beginning location of the TLWB, and then the H bit is changed as 0. If this block is accessed again before its eviction, it will be turned on as 1. Thus, any candidate block having the H bit as 0 will be written back to the array of PRAM/Flash hybrid IM-D memory structure.

3.3 PRAM-Flash Hybrid Integrated Memory Storage

As shown in Table 1, PRAM offers a density advantage similar to that of NAND Flash, and shows almost 100X

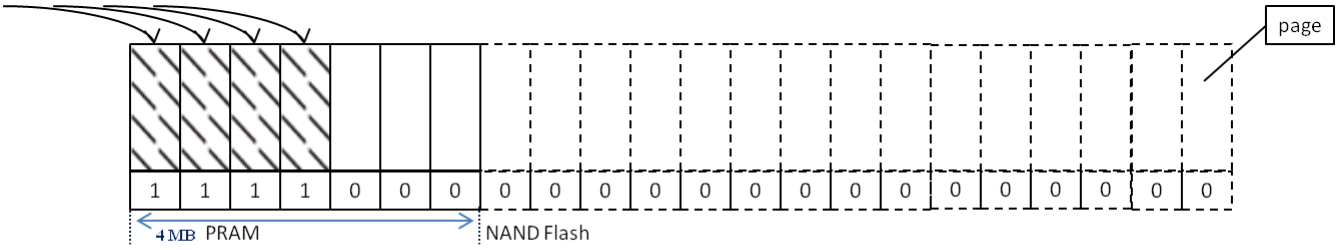

Figure 3. (a) Storage Initial

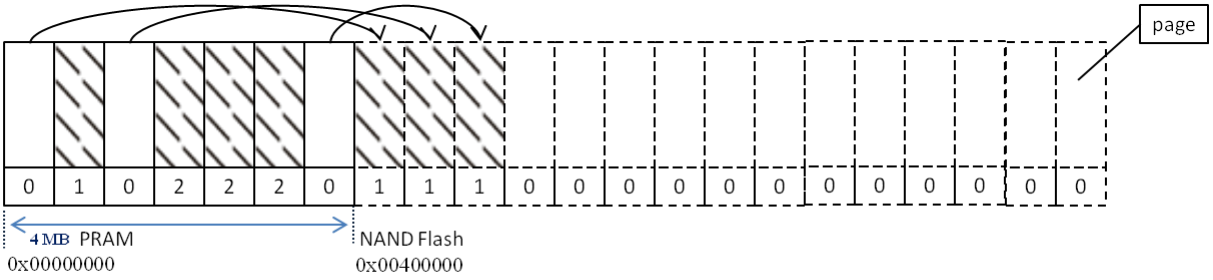

Figure 3. (b) PRAM is Full

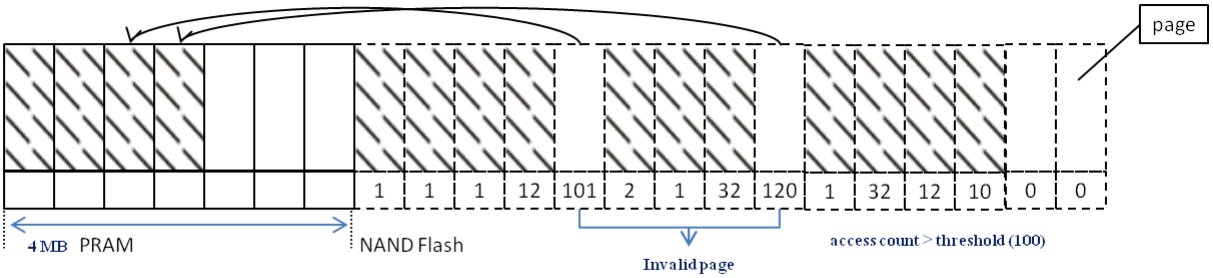

Figure 3. (c) Storage Idle

faster than NAND Flash. However, in terms of price, PRAM has not been reached to achieve general commercialization. Thus, a hybrid model of PRAM and Flash memories should be considered as a currently feasible model for cost effective IM-D structure as in solid state disk mass storage [7]. Here SLC PRAM memory with space *n* and MLC flash memory with space *m* are integrated as a single array of integrated space *n* + *m*. The proportion of *n* over *m* can be determined as an optimal combination for best performance and cost effectiveness. Here, we exploit the fast access latency of SLC PRAM and the large capacity of MLC Flash memory to provide shorter access latency, lower power consumption, and mass IM-D space.

Figure 3 shows management policy for the PRAM/Flash hybrid storage, in step by step. We assume that enough space is available in the NAND Flash storage area. At the initial stage of system usage, namely when the IM-D storage is empty and write requests/file writes are performed, PRAM storage area will be allocated first as shown in Figure 3 (a). And then, when the PRAM storage space is filled with the pages, then the least recently used (LRU) pages will be moved to the pre-erased NAND Flash area, where managing unit is the same as the NAND Flash erase unit as shown in Figure 3 (b). Also, we utilized the flash page spare area to store the number of accesses for its corresponding Flash page. When the IM-D storage is idle, it automatically calculates the threshold value using the number of times accessed on the spare area. By comparing the number of times accessed with the threshold, page swapping can be performed. If the number of times accessed is larger than the threshold, then transmit the page to PRAM storage area, and its status in Flash space is changed as invalid, as shown in Figure 3 (c). Suppose that the IM-D storage is idle and the calculated threshold value is 100. Then, the number of times accessed is compared with the threshold. As in Figure 3 (c), two pages whose access counts are 101 and 120 are chosen as the candidate, and thus, the two pages are moved to PRAM storage area. But if the PRAM storage area is filled completely with the pages, the PRAM LRU pages are

selected and moved to the pre-erased NAND Flash area. Thus, infrequently accessed pages are maintained onto NAND Flash area. Also pages in PRAM are maintained as LRU fashion. The value of the threshold can be determined by formula (1).

$$f(X) = \overline{X} + 2.33\sigma \quad (1)$$

X : the access count which is maintained in the corresponding flash page spare area

$\overline{X}$: average number of accesses

σ : standard deviations

2.33: standard deviation and statistical dispersion percentage outside 3.890592σ

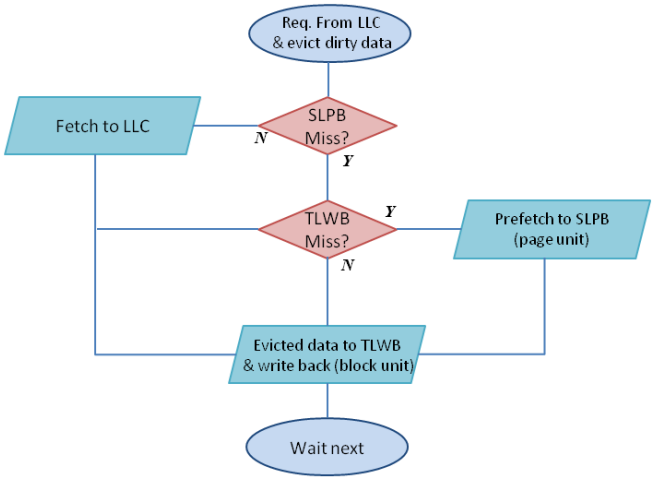

Figure 4. Operational overflow

3.4 Operational Flow

Basic operational flow of the proposed structure is provided as shown in Figure 4. When a data block is requested from the last level of cache and also a victim process is performed, our proposed mechanism is performed as follows:

Step 1: A data request is generated from the last level of cache and simultaneously a data block may be evicted from the last level of cache.

Step 2: First, check its dirty bit, and look over if the dirty bit is set. If the dirty bit is not set, it is simply discarded from the last level of cache, go to Step 6, otherwise continue to the next step.

Step 3: Check the TLWB, and find out if there is any matched data in the TLWB. If the evicted block exists in the TLWB, go to Step 6, otherwise continue to the next step.

Step 4: Check the TLWB, if the TLWB has been filled with blocks, then a block will be written back to the PRAM/Flash hybrid memory.

Step 5: Any highly referenced block in the TLWB will move to the beginning of the TLWB. Other dirty blocks will be written back to the PRAM/Flash hybrid memory, and go to Step 6.

Step 6: Provide the requested block to the last level of cache, and prefetch a page unit of data into the SLPB.

4. Performance Analysis

A dual buffering IM-D adapter based simulator for the PRAM/Flash integrated memory storage was developed to evaluate overall performance. We implemented our simulator to present a comprehensive IM-D performance study of workloads using both SPEC 2006 [9, 10] and SPLASH2 [11], because of their different access characteristics and future expanded applications on mobile environment. We used GEM5 full system mode to generate virtual address traces of CPU requests. We uses 10 scientific applications, e.g., **bzip2** (BM1)**, mcf** (BM2)**, soplex** (BM3)**, sjeng** (BM4)**, wrf** (BM5)from the SPEC 2006 benchmark and **fft** (BM6)**, lu** (BM7)**, barnes** (BM8)**, ocean** (BM9)**, raytrace** (BM10) from the SPLASH benchmark. As a basic system configuration chosen in the experiment, L1 instruction and data caches consist of 32-KByte, 4-way set associative, with a 64 Byte block size. Also, L2 cache memory is configured as a unified 512-KByte, 8-way set associative cache, with a 128 Byte block size. The proposed dual buffering IM-D adapter can be configured: 1-MByte for SLPB and 1-MByte for TLWB and both are managed as in FIFO. The IM-D memory structure is configured as 4-GBytes which includes 4-Mbyte SLC PRAM.

We also implemented a conventional main memory with disk/Flash structure and a simple unified IM-D adapter both having the same space of the dual buffering IM-D adapter. The conventional mobile memory hierarchy consists of a unified L2 cache and a 2-MByte main memory space, called conventional module. The unified IM-D adapter consists of a unified L2 cache and a 2-Mbyte prefetch buffer. When any L2 cache miss occurs, it fetches the requested data to the L2 cache and simultaneously prefech 4 Kbyte unit page to the prefetch buffer, called a unified module. And we call our proposed dual buffering IM-D adapter with unified L2 cache as the proposed module. Several experiments are performed by examining the effectiveness of proposed structure.

DRAM access Speed	1Tbyte/s
PRAM Write Speed	100MByte/s
PRAM Set Cycle Time	50ns
PRAM Reset Cycle Time	100ns
NAND Page Read Latency	50us
NAND page Write Latency	800us
NAND Block Erase Latency	3000us
Page Size	4K
Pages amount in a block	64
Blocks amount in a superblock	32

Table 2: Simulation configuration

4.1 Impact on Miss Rate

This experiment is simply to show the pure impact of the DRAM based dual buffering IM-D adapter management for the same cost. Assuming the same execution environment, miss rate is used to compare four configurations fairly.

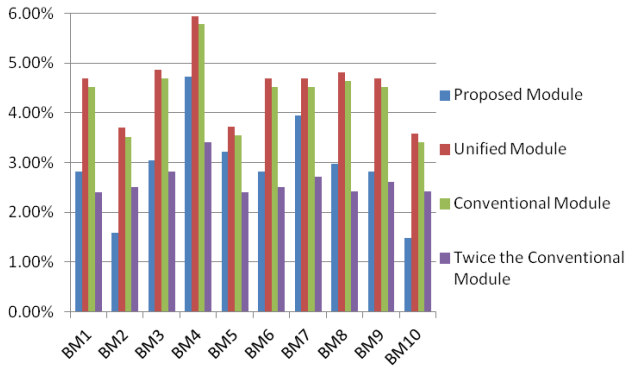

Fig. 5. Miss rates for IM-D adapter and conventional main memory

Figure 5 shows the miss rates on four configurations, where the y-axis shows miss rates to compare the miss rate of the proposed module with those of unified module, conventional module and twice space of the conventional memory module. On the average, the miss rate of proposed dual buffering IM-D adapter is 2.94 %, the miss rate of unified IM-D adapter is 4.54%, the miss rate of conventional memory is 4.37%, and the miss rate of twice space of the conventional main memory is 2.62% respectively. In summary, the IM-D adapter shows similar miss rate as the case of twice space of the conventional main memory, thus, our proposed IM-D structure with a dual buffering IM-D adapter can provide similar performance relative to twice space of the conventional memory structure.

4.2 Reducing Access Latency

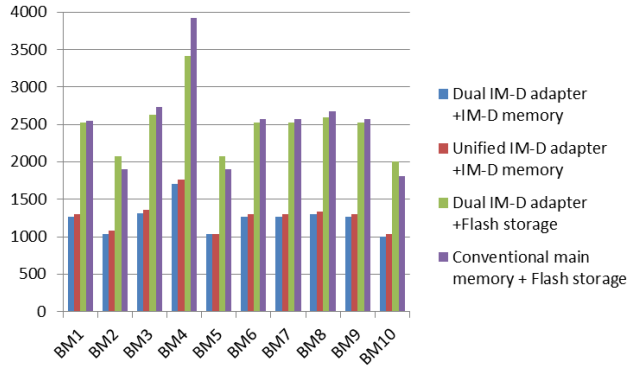

Fig. 6. Comparison of access latency over different configurations.

As shown in Table 1, we assume it has 100X latency difference between NAND Flash and PRAM memory. For comparison, we take conventional Flash storage memory structure as a baseline. Figure 6 shows the access latency on four configurations, where the y-axis shows access time. The dual IM-D adapter with Flash storage shows similar access latency as the case of conventional main memory with Flash storage. However, by using IM-D structure with a dual buffering IM-D adapter and IM-D memory, or with a single unified memory adapter and IM-D memory, the access latency can be enhanced by 49.9% and 48.39% comparing with the conventional main memory with Flash storage based mobile device. And overall storage capacity remains the same level and overall storage access latency (between adapter and storage decreases by around 49.6 and over 40 times, comparing with the conventional Flash storage based mobile device. It means that our proposed PRAM/Flash hybrid IM-D non-volatile memory structure can provide significantly rapid response to the users.

4.3 Impact on PRAM Lifetime

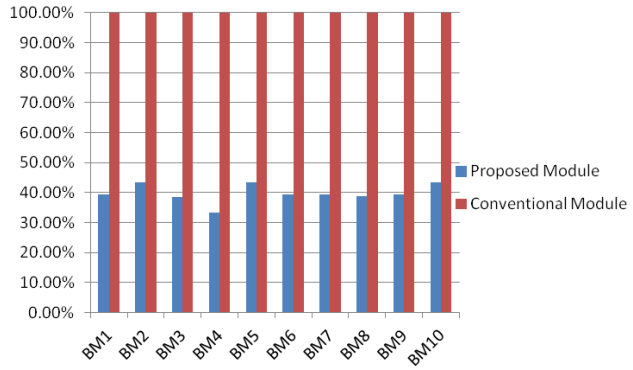

Fig. 7. Comparison of IM-D adapter with conventional memory for write back reduction.

Because of longer access latency and the limited endurance of PRAM, we exploit the TLWB as a write buffer to hide write latency of the slower PRAM memory while increasing the probability of hits accessed by last level of cache. And also the write buffer is to extend the time to stay in the TLWB, showing high probability of re-accessing behavior. We compare the proposed module and a conventional module. We consider conventional module as a baseline.

Figure 7 shows the number of writes reduced over the conventional module, where conventional module is normalized as one, and the y-axis shows the percentage of write back reduction. Our proposed IM-D memory adapter can decrease the average write back count by 60.15%. Thus, by reducing the number of write back operations to the PRAM/Flash memory can prolong the wearable memory endurance.

4.4 Impact on Main Memory Energy Consumption

Energy consumption needs to be clarified to see its impact on our proposed approach. For this, we use basic parameters to evaluate the energy consumption, as given in Table 1. Note that the read power consumption is around two times lower than that of DRAM, while PRAM's write power consumption is significantly greater than that of DRAM. Moreover, the activation/pre-charging power is defined power consumption required to open and close a row in the memory array. This one is higher for DRAM than PRAM, since it has to refresh the row data before closing a row, which can be avoided in PRAM due to its non-volatility. The standby power, consumed when it is idle, is also lower for PRAM than DRAM, due to its negligible leakage power consumption. For NAND Flash, a write operation can only be performed on an empty or erased unit, so in most cases write operation must be preceded by an erase operation, and a block is the minimum unit to implement erase operations. However, this is not required for PRAM.

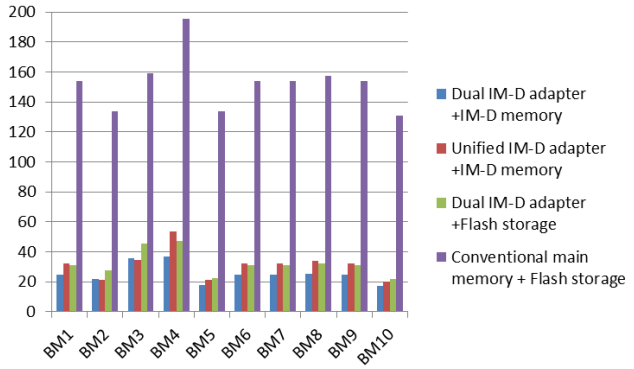

Fig. 8. Comparison of energy consumption over different configurations.

Figure 8 shows the energy consumption on four configurations, and the y-axis shows energy consumption of the proposed structure, comparing with that of conventional module. As shown in Figure 8, on average, energy consumption of the proposed structure is reduced by 83.51%, compared with conventional module. Through experimental results, we consider that a significant reduction in write back operations to the PRAM/Flash IM-D memory can save most of the energy consumption.

5. Conclusions

In this paper, conventional main memory and disk storage layers are merged into a single memory layer using a combination of PRAM and NAND Flash memories, which is called as an integrated memory-disk (IM-D) non-volatile memory structure. The IM-D architecture consists of a dual buffering IM-D adapter, an array of SLC/MLC PRAM/Flash hybrid IM-D memory, and an associated memory management module as IM-D translation layer in the operating system. The dual buffering IM-D adapter comprising of decoupled dual buffers is located between last level of cache and IM-D memory. This component is to utilize spatial locality and temporal locality. One is spatial locality prefetch buffer and the other one is temporal locality write buffer. The IM-D memory structure includes fast SLC PRAM and mass MLC NAND Flash hybrid structure to enhance the lifetime and hide asymmetric read/write access latencies. The experimental results show that overall storage capacity remains the same level and overall storage access latency decreases by around 49.6 times. Moreover, our structure can significantly save the energy consumption by 83.51%. Also, SLC/MLC PRAM/Flash hybrid architecture can show close performance in terms of miss rate, compared to the twice space of the conventional mobile device memory architecture. Based on this result, our proposed memory system offers the possibility of replacing conventional DRAM and HDD/NAND Flash memory based mobile device system with IM-D structure.

References

[1] J. U. Kang, H. Jo, J. S. Kim, and J. Lee, "A Superblock-based Flash Translation Layer for NAND Flash Memory", EMSOFT'06 pp. 22-25, Oct. 2006.

[2] Ning. Lu, In-Sung Choi, Shin-Dug Kim, "A flash-aware write buffer scheme to enhance the performance of superblock-based NAND flash storage", systems, Microprocess. Microsyst, July. 2012.

[3] Qureshi MK, Srinivasan V, Rivers JA, "Scalable high performance main memory system using phase memory technology', ACM SIGARCH Computer Architecture News, Volume 37 Issue 3, June 2009, pp 24-33.

[4] Kwang-Su Jung, Jung-Wook Park, and Shin-Dug Kim, "A Superblock-based Memory Adapter Using Decoupled Dual Buffers for Hiding the Access Latency of Non-volatile Memory" World Congress on Engineering & Computer Secience 2011, pp.802-807, October 19-21, 2011.

[5] Seon-yeong Park, Dawoon Jung, Jeong-uk Kang, Jin-soo Kim, Joonwon Lee, "CFLRU: a replacement algorithm for flash memory", Compilers, Architecture, and Synthesis for Embedded Systems - CASES , pp. 234-241, 2006.

[6] Gaurav Dhiman, Raid Ayoub, Tajana Rosing, “PDRAM: a hybrid PRAM and DRAM main memory system”, Design Automation Conference - DAC , pp. 664-469, 2009.

[7] Hyojun Kim, Nitin Agrawal, Cristian Ungureanu, “Revisiting storage for smartphones”, FAST’12 Proceedings of the 10 USENIX conference on File and Storage Technologies, pp. 17-17, 2012.

[8] Intel, Understanding the Flash Translation Layer (FTL) Specification 1998.

[9] Henning, J. L. 2006 SPEC CPU2006 Benchmark Descriptions.ACM SIGARCH newsletter, Computer Architecture News 34, NO. 4

[10] SPEC CPU2006 : http://www.spec.org/cpu2006/

[11] SPLASH2 : http://www.capsl.udel.edu/splash/Download.html