

Power and Performance of Read-Write Aware Hybrid Caches with Non-volatile Memories

Xiaoxia Wu^{†‡} Jian Li[†] Lixin Zhang[†] Evan Speight[†] Yuan Xie[‡]

[†] IBM Austin Research Laboratory

[‡] Computer Science and Engineering Department, The Pennsylvania State University, University Park, PA 16802

[†] Email: {jianli,zhangl,speight}@us.ibm.com [‡] Email: {xwu,yuanxie}@cse.psu.edu

Abstract—Caches made of non-volatile memory technologies, such as Magnetic RAM (MRAM) and Phase-change RAM (PRAM), offer dramatically different power-performance characteristics when compared with SRAM-based caches, particularly in the areas of static/dynamic power consumption, read and write access latency and cell density. In this paper, we propose to take advantage of the best characteristics that each technology has to offer through the use of read-write aware Hybrid Cache Architecture (RWHCA) designs, where a single level of cache can be partitioned into read and write regions, each of a different memory technology with disparate read and write characteristics. We explore the potential of hardware support for intra-cache data movement within RWHCA caches. Utilizing a full-system simulator that has been validated against real hardware, we demonstrate that a RWHCA design with a conservative setup can provide a geometric mean 55% power reduction and yet 5% IPC improvement over a baseline SRAM cache design across a collection of 30 workloads. Furthermore, a 2-layer 3D cache stack (3DRWHCA) of high density memory technology with the same chip footprint still gives 10% power reduction and boost performance by 16% IPC improvement over the baseline.

I. INTRODUCTION

Different memory technologies exhibits significantly different properties: dynamic/static power consumption, density, read/write latency, reliability features, scalability, etc. Table I lists important qualitative features of three memory technologies: SRAM, Magnetic RAM (MRAM) [12], and Phase-change RAM (PRAM) [10]. Several observations may be made from Table I:

- SRAM has high static power, while MRAM and PRAM have very low static power due to their non-volatile property.
- MRAM and PRAM have very different read and write features in terms of latency and power consumption, with particularly high write latency and write power consumption.
- PRAM has the highest potential density, but it also has the slowest speed. MRAM also have higher density than SRAM, but is slower than SRAM. Depending on the design, MRAM read speed may be comparable to that of SRAM.

TABLE I: Comparison of different memory technologies.

Features	SRAM	MRAM	PRAM
Non-volatility	No	Yes	Yes
Leakage Power	High	Low	Low
Dynamic Power	Low	Low for read very high for write	Medium for read high for write
Density	Low	High	Very high
Speed	Very Fast	Fast for read slow for write	Slow for read very slow for write
Scalability	Yes	Yes	Yes

On the other hand, we observe that different applications have differing read and write behaviors, and read/write access patterns also vary along the time line for one application. Fig. 1 shows that the read and write ratio changes significantly in a

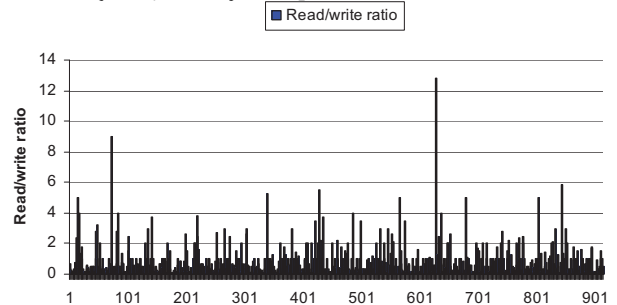


Fig. 1: Read/write ratio for one cache line in *omnetpp* benchmark.

single cache line for different data along the time line in one benchmark (*omnetpp*). Therefore, reads and writes may be distinguished in order to achieve better power-performance. Consequently, a properly designed cache that is made of differing memory technologies with different read/write features may have the potential to outperform its counterpart of single technology. In addition, even though mixed technologies can also be integrated on the same two-dimensional (2D) chip, the emerging three-dimensional (3D) chip integration technologies may provide further design and manufacture cost benefits for on-chip mixed-technology integration.

In this paper, we propose and evaluate a Read-Write aware Hybrid Cache Architecture (RWHCA) to accommodate on-chip cache hierarchies. To fully take advantage of the benefits from varied memory technologies, an RWHCA allows one level of cache to be partitioned into read and write regions of different memory technologies. In addition, we propose techniques such as low-overhead intra-cache data movement to improve cache performance in a RWHCA system. Utilizing a full-system simulator that has been validated against real hardware we demonstrate that a RWHCA design with a conservative setup can provide a geometric mean 55% power reduction and yet 5% IPC improvement over a baseline SRAM cache design across a collection of 30 workloads. Furthermore, a 2-layer 3D cache stack (3DRWHCA) of high density memory technology with the same chip footprint still gives 10% power reduction and boost performance by 16% IPC improvement over the baseline.

This paper makes the following contributions:

- We propose read-write aware region-based hybrid cache architecture (RWHCA) made of differing memory technologies.
- We evaluate RWHCA made of combinations of SRAM, MRAM and PRAM under similar area constraint and show that SRAM-MRAM based RWHCA achieves dramatic power savings while still improves performance. We observe that SRAM-PRAM based L2 RWHCA is not promising but

PRAM is promising for lower level cache.

- We extend the RWHCA technique to 3D stacking and evaluate the power-performance benefits.

II. BACKGROUND

A. Magnetic RAM (MRAM)

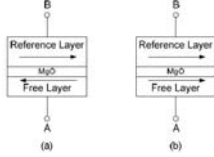


Fig. 2: MTJ structure. (a) Anti-parallel (high resist.), “1” state; (b) Parallel (low resist.), “0” state.

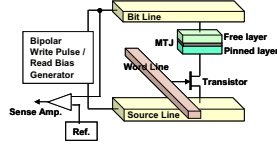


Fig. 3: An illustration of an MRAM cell

The basic difference between MRAM and conventional RAM technologies (such as SRAM/DRAM) is that the information carrier of MRAM is a Magnetic Tunnel Junction (MTJ) instead of electric charges [12]. Each MTJ contains *two ferromagnetic layers* and *one tunnel barrier layer*. Figure 2 shows a conceptual illustration of an MTJ. One of the ferromagnetic layers (the reference layer) has a fixed magnetic direction while the other one (the free layer) can change its magnetic direction via an external electromagnetic field or a spin-transfer torque. If the two ferromagnetic layers have different directions, the MTJ resistance is high, indicating a “1” state (the anti-parallel case in Figure 2(a)); if the two layers have the same direction, the MTJ resistance is low, indicating a “0” state (the parallel case in Figure 2(b)).

The MRAM technology to be discussed in this paper is called *Spin-Transfer Torque RAM* (STT-RAM), which is a new generation of MRAM technologies. STT-RAMs change the magnetic direction of the free layer by directly passing a spin-polarized current through the MTJ structure. Compared to the previous generation of MRAMs that used external magnetic fields to reverse the MTJ status, STT-RAMs have the advantage of scalability, as the *threshold current* to make the status reversal will decrease as the size of the MTJ becomes smaller.

In the STT-RAM memory cell design, the most popular structure is composed of one NMOS transistor as the access controller and one MTJ as the storage element (“1T1J” structure) [12]. As illustrated in Figure 3, the storage element, MTJ, is connected in series with the NMOS transistor. The NMOS transistor is controlled by the the word-line (WL) signal. The detailed read and write operations for each MRAM cell is described as follows:

- **Write Operation:** When a *write operation* is performed, a positive voltage difference is established between the source-line (SL) and bit-line (BL) for writing for a “0” or a negative voltage difference is established for writing a “1”. The current amplitude required to ensure a successful status reversal is called the threshold current. This current is related to the material of the tunnel barrier layer, the writing pulse duration, and the MTJ geometry.
- **Read Operation:** When a *read operation* is desired, the NMOS is turned enabled and a voltage ($V_{BL} - V_{SL}$) is applied between the BL and the SL. This voltage is negative and is usually very small (- 0.1V as demonstrated in [12]).

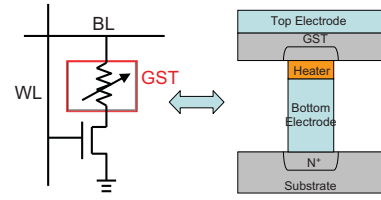


Fig. 4: An illustration of a PRAM cell. When phase change material GST is in an amorphous phase, it indicates “0” state; when GST is in a crystalline phase, it indicates “1” state.

The voltage difference will cause a current to pass through the MTJ, but it is small enough to not invoke a disturbed write operation. The value of the current is determined by the equivalent resistance of MTJs. A sense amplifier compares this current with a reference current and then decides whether a “0” or a “1” is read from the selected MRAM cell.

B. Phase-Change RAM (PRAM)

PRAM is a another promising memory technology [4], [10]. It has a wide resistance range, which is about three orders of magnitude; therefore, multi-level PRAM allows the storage of multiple bits per cell. Two to four bits per cell have already been demonstrated [10]. The basic structure of a PRAM cell consists of a standard NMOS access transistor and a small volume of phase change material, GST (Ge₂Sb₂Te₅), as shown in Figure 4. The phase change material can be switched from an amorphous phase (reset or “0” state) to a crystalline phase (set or “1” state), or vice versa, with the application of heat. The read and write operations for a PRAM cell is described as follows:

- **Write Operation:** there are two kinds of PRAM write operations: the *SET* operation that switches the GST into crystalline phase and the *RESET* operation that switches the GST into amorphous phase. The *SET* operation crystallizes GST by heating it above its crystallization temperature, and the *RESET* operation melt-quenches GST to make the material amorphous [10]. These two operations are controlled by electrical current: high-power pulses for the *RESET* operation heat the memory cell above the GST melting temperature; moderate power but longer duration pulses for the *SET* operation heat the cell above the GST crystallization temperature but below the melting temperature. The temperature is controlled by passing through a certain amount of electrical current and generating the required Joule heat.
- **Read Operation:** To read the data stored in PRAM cells, a small voltage is applied across the GST. Since the *SET* status and *RESET* status have a large variance on their equivalent resistance, the data is sensed by measuring the pass-through current. The read voltage is set to be sufficiently strong to invoke detectable current but remains low enough to avoid write disturbance. Like other RAM technologies, each PRAM cell needs an access device for control purpose. As shown in Figure 4, every basic PRAM cell contains one GST and one NMOS access transistor. This structure has a name of “1T1R” where “T” stands for the NMOS transistor and “R” stands for GST. The GST in each PRAM cell is connected to the drain-region of the NMOS in series so that the data stored in PRAM cells can be accessed by wordline controlling.

As described, MRAM and PRAM memory technologies are made of different materials than SRAM and have different read/write operations. However, caches constructed from these

technologies have similar structure from a logic designer’s point of view due to the similarity of the peripheral circuits.

III. METHODOLOGY

A. System Configuration

TABLE II: Cache parameters of memory technologies (45nm).

Cache	Den.	Lat.(cycles)	Dyn. eng/op.(n.J)	Stat. pow.(W)
SRAM(1MB)	1	8	0.388	1.36
MRAM(4MB)	4	read:20;write:60	read:0.4;write:2.3	0.15
PRAM(16MB)	16	read:40;write:200	read:0.8;write:1.5	0.3

We based our parameters on searches of appropriate literature [10], [12] for typical density, latency, and energy numbers for the studied memory technologies, and then scale these to 45nm technology. All cache parameters used in this study were obtained either from CACTI [1] or its modified versions [11] and are shown in Table II. Since MRAM and PRAM are emerging memory technologies, the projection of their features tends to be more varied than the ones for established technologies such as SRAM, however, we have chosen cache parameters in-line with other researchers’ assumptions. Note that, multi-level PRAM can store four bits per cell [10] while the other memory technologies store one bit per cell.

TABLE III: System configuration.

Processor	Eight-way issue out-of-order, 4GHz
L1	32KB DL1 + 32KB IL1, 128B, 4-way, 1 R/W port, 2 cycles
L2/L3	See corresponding design cases
Memory	400 cycles, mem. cntrl vs. core speed 1:2, 16MB page

In this work, we study the power savings of RWHCA on a *chipllet* of a multi-core chip, which contains one core and its associated private caches. We assume an eight-way issue out-of-order cores representing future PowerPC-based processors. The experiments are conducted using a full system simulator that has been validated against existing POWER5® hardware. In this paper, we keep the configurations of processor core, L1 caches, on-chip interconnect, and memory system the same, and only study the design of different low-level caches (e.g., L2 or L3) under similar chip area constraint or similar footprint in the case of 3D chip stacking. In fact, we conservatively set the chip area of RWHCA configurations to be slightly larger than their pure-SRAM counterparts, so that pure-SRAM caches tend to have lower leakage power due to smaller capacity. Table III gives our system configuration.

B. Workloads

The benchmarks we used in this study are chosen from a wide spectrum of workloads: SpecInt2006 [3], NPB [6], SPLASH2 [14], PARSEC [8], BioPerf [5], and SpecJBB [2]. Four PARSEC workloads covering the range of memory footprints of the whole PARSEC suite are selected. Table IV gives the problem size and other parameters of the benchmarks. For all workloads except SPLASH2, we use either sampled reference or native input sets to represent a real-world execution scenario. For SPLASH2, we increase the input set of some of the workloads such that the total number of dynamic instructions are more than 10 billion.

In order to reasonably evaluate large cache designs, we construct each simulation in three phases with decreasing simulation speed: (1) we fast forward to a meaningful application phase, which may take 10s - 100s billion of instructions; (2) we warm up the caches by 10s billion of instructions; and (3) we simulate the system cycle-by-cycle for a few billion of

TABLE IV: Workloads.

Workload	Applications and Program size
SpecInt06	reference input: astar, bzip2, gcc, gobmk, h264, hmmer-sp, libquantum, mcf, omnetpp, perl, sjeng
SPECJBB	IBM JVM version 1.1.8, 16 warehouses
NAS	Class C: cg, lu, mg, sp, ua
BioPerf	reference input: blast, clustalw, hmmer
PARSEC	native input: dedup, fluidanimate, freqmine, streamcluster
SPLASH2	barnes(64K particles), fmm(128k particles), fit(4M data points) lu(2048*2048 matrix), ocean(2048*2048 grid)

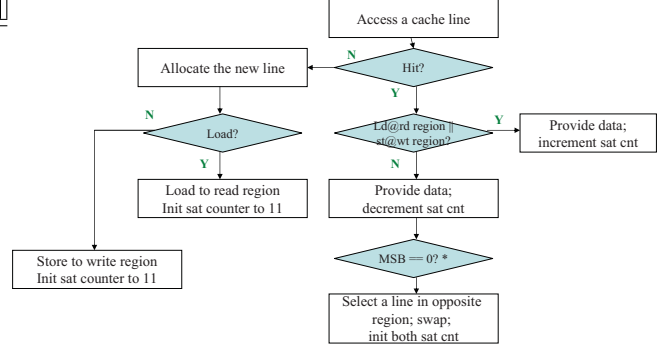


Fig. 5: Cache line allocation and migration policy in the RWHCA.

instructions and collect simulation results. Both performance and power statistics are collected from cycle mode execution. Our cache power model adds the static and dynamic power of the caches used by a workload in the simulation. The static power is obtained from CACTI or its modified versions, as shown in Table II. The dynamic power factors in the number of read and write accesses and their corresponding per-access energy values are given in Table II.

C. Design Methodology

Throughout the RWHCA studies presented here, we assume the chip area, or the chip footprint in the 3D integration scenario, is similar for all the design cases. In our 2D baseline system, each processor core has two levels of private caches (L1 configuration is listed in Table III, L2 is 1MB cache with 4 banks). Both two levels of caching are comprised of SRAM. This configuration serves as the *baseline configuration* in this work.

In a 2D chip design scenario, one can construct a hybrid, coarse-grained Non-Uniform Cache Architecture (NUCA) cache with L2 write- and read-regions made of SRAM and MRAM/PRAM, respectively. The cache regions are mutually exclusive. We discuss this scenario in Section IV. Furthermore, we study stacking L3 cache made of PRAM on top of RWHCA, which forms an L2 cache with read and write regions comprised of SRAM and MRAM, with an additional PRAM-based L3 cache. This design option embodies the 3D RWHCA (3DRWHCA) and is evaluated in Section V.

IV. READ-WRITE AWARE HYBRID CACHE

A. Cache Line Migration Policy

The hybrid L2 cache consists of one small write (SRAM) region and one large read (MRAM or PRAM) region. Fully exploring its potential requires proper cache line replacement and data migration policies between the read and write regions.

Figure 5 depicts the cache line migration policy we use in our RWHCA design. A new cache line is allocated when there is a miss in the cache. If the miss is load miss the data is allocated to the read region based on LRU policy and the saturated counter is initiated to be 11. If it is a store miss the data is allocated to the write region based on LRU policy and saturated counter is initiated to be 11. When there is a hit, we check whether or not it is a load hit in read region or store hit in write region. If so we provide data and increment the saturated counter, otherwise we provide data first and decrement the counter. If the counter is decreased and the MSB bit is 0 it means that there are consecutive hits in the wrong region. Therefore, we select a line in opposite region based on LRU policy, swap with this line, and initiate the counters for both lines. The reason we provide data first and then update the counter is that we want saturation counter update and swap to be in non-critical paths. The swap threshold and the initial value of the counter can be changed according to the read/write ratio of the applications so that it can also provide different priority for reads and writes.

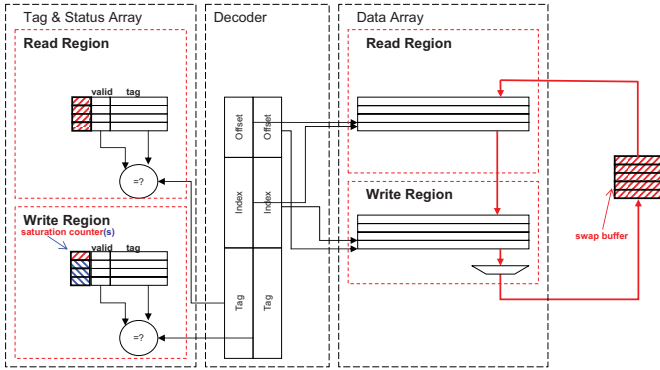


Fig. 6: Block diagram of the proposed RWHCA. Structures with slash patterns are new components.

B. Hardware Support

The hardware support for the swap operations is shown in Figure 6. The read region and the write region each has a tag and status array (left) as well as data array (right) allocated on both sides of the address decoder. The address decoder is replicated to meet timing demands. The trade-offs between the number of replicated decoders and bank partition is beyond the scope of this paper. The main additions are the saturating counters and the swap buffer.

A swap operation involves reading out two cache lines from two regions and writing each to the opposite region. Because of the speed difference between two regions and the contention on the cache arrays, a line read out a region may not be able to go to the opposite region immediately and therefore must be temporarily buffered elsewhere. To simplify logic, we propose to utilize a swap buffer and serialize the swap operation as follows. First the data in the write region is read out and placed into the buffer. Then the data in the read region is read out and written to the write region. Finally, the line in the swap buffer is written to read region. Each of the three steps may take multiple cycles. The swap buffer contains multiple entries and allows multiple outstanding swap operation. Note that the first step is already being done as part of the process of loading the

line into the upper level cache. We simply need save the line in the swap buffer before it can be written to the read region.

An alternative approach is to read both lines in parallel. In this approach, the swap buffer is either double-ported or specially arranged to allow two writes simultaneously.

We have evaluated the sensitivity of swap latency and swap buffer size. The swap buffer is snooped for coherence operations. A snoop hit in the swap buffer will result in a retry response in our simulated system. Our simulation results indicate that such scenario rarely happens and is not a concern for performance degradation. A buffer size of 16 entries is sufficient for all workloads studied.

C. Results

TABLE V: Read-write region hybrid cache parameters in L2.

RWHCA (rd-wrt)	SRAM-MRAM	SRAM-PRAM
L2 size	4MB	16MB
SRAM region (lat)	256KB (6 cycles)	256KB (6 cycles)
M/PRAM region (lat)	read:20;write:60	read:40;write:200
Bank number	16	64
Associativity	16	64
block size	128B	128B
ports	1 rd/wrt	1 rd/wrt

RWHCA can be SRAM-MRAM or SRAM-PRAM based. The RWHCA cache design parameters for the proposed hybrid L2 cache are listed in Table V. We compare the RWHCA designs with the SRAM-only baseline. We also compare our counter-based data migration design with the generational promotion approach first proposed for Dynamic NUCA (DNUCA) by Kim et. al [13]. Generational promotion moves a line to a closer bank on each hit, which does not differentiate read and write operations.

Figure 7.A shows the performance of SRAM-MRAM RWHCA. The RWHCA design has a geometric mean performance improvement of 5% over the SRAM-only design and also is 3% faster than DNUCA. Note that we also include 3-level SRAM (L1 is same with 2-level SRAM baseline, 256KB L2 and 1MB L3) result in the figure, showing that our RWHCA design achieves better performance. We observe that for some workloads RWHCA performs better than the baseline while some are opposite. The possible reason is that for some applications lots of frequently used data are read operations, while the unfrequent write access data migration may not offer the performance benefit but affect the read operations instead. RWHCA outperforms DNUCA because of the difference in the speed of data movement. RWHCA moves frequently-written cache lines directly to the write region from any bank in the read region. DNUCA moves cache lines one bank at a time. Because of the large number of banks in the read region, it often requires many more hits before a line moves into the write region.

Figure 7.B shows that SRAM-PRAM RWHCA has a 20% performance degradation relative to SRAM-only design, indicating that the SRAM-PRAM RWHCA design is not very promising for an L2 cache due to the long latency of PRAM. However, we will show that it is a promising technology for lower level caches due to its high density in Section V.

Figure 8.A and Figure 8.B illustrate the power comparison for SRAM-MRAM RWHCA and SRAM-PRAM RWHCA, with static power, dynamic power for normal data access and

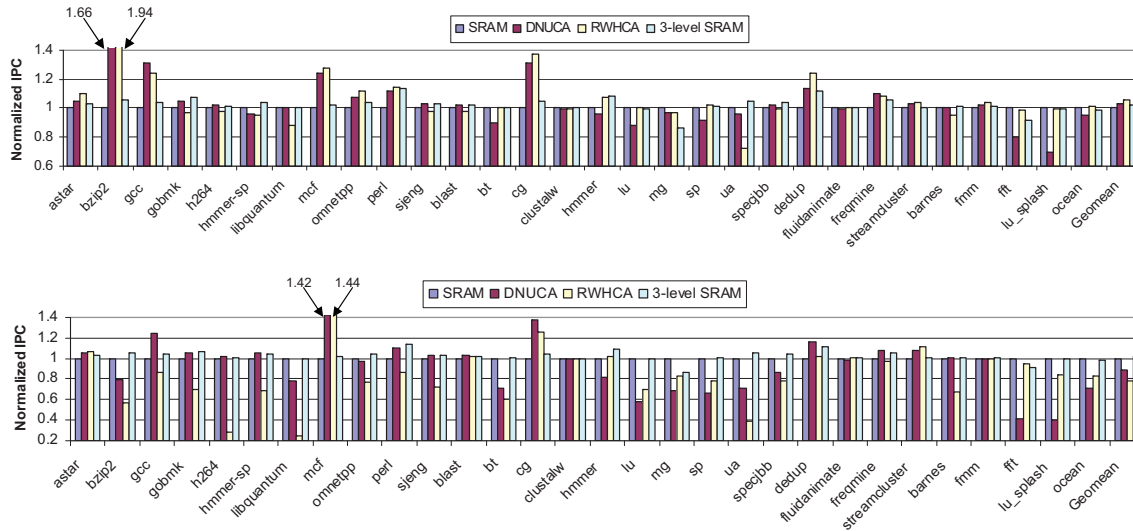


Fig. 7: Performance of SRAM-MRAM (top, A) and SRAM-PRAM (bottom, B) RWHCA.

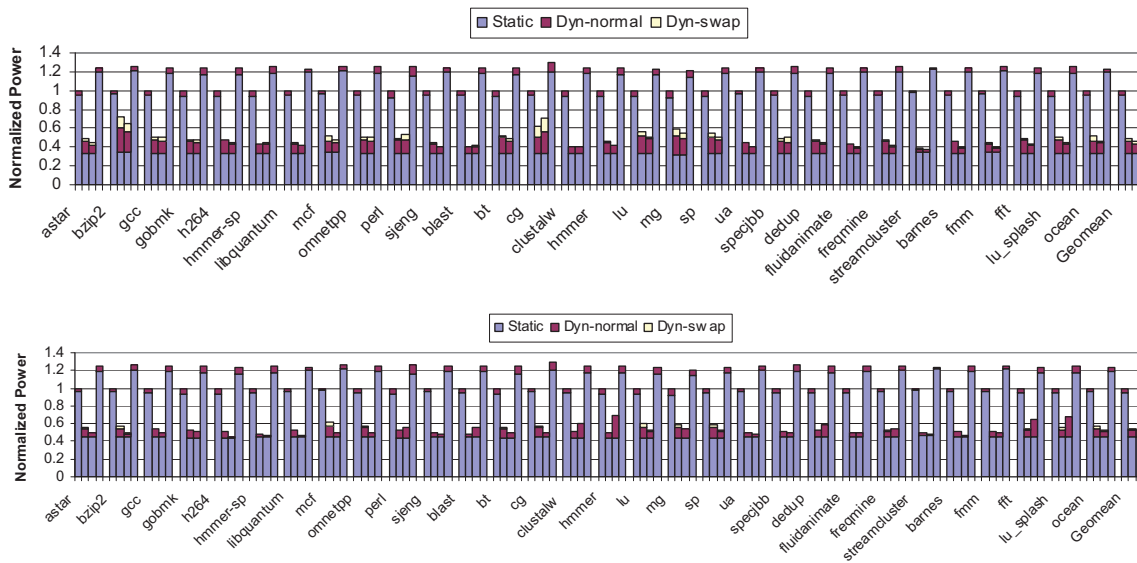


Fig. 8: Power of SRAM-MRAM (top, A) and SRAM-PRAM (bottom, B) RWHCA.

dynamic power for data migration. We see that SRAM-MRAM RWHCA and SRAM-PRAM RWHCA achieve about 55% and 45% power reduction compared to SRAM baseline, respectively, due to their low leakage power. Another observation is that RWHCA consumes moderately less power than DNUCA because of their different data migration policies.

V. 3D HYBRID CACHE STACKING

3D cache stacking enables the addition of more cache levels without sacrificing the number of cores. These extra cache levels should be at least a few times larger than the cache level above it in the cache hierarchy to effectively reduce miss rate. We assume the 3D cache layer has the same footprint as its corresponding 2D chiplet, which consists of a processor core and its original caches. If a memory technology of the same density is used, then multi-layer 3D cache stacking is anticipated. However, multi-layer 3D stacking may incur mounting problems in power delivery, cooling, and TSV

efficiency. Therefore, we expect a denser memory technology to be an alternative approach to multi-layer 3D cache stacking. In this paper, we consider PRAM. Besides its high density, PRAM also has very low static power, which further helps address the cooling issues with 3D. We use the latency and scale power parameters of PRAM as shown in Table II. We assume the processor and memory domain clock frequencies of 3D are the same as its 2D counterpart.

We stack PRAM L3 cache on top of RWHCA architecture, in which the configuration includes a 4MB SRAM-MRAM read-write region RWHCA L2 (Section IV) and 32MB L3 cache. Figure 9A illustrates the performance comparison of 3DRWHCA with SRAM baseline and SRAM-MRAM RWHCA in 2D case. The results show that 3DRWHCA exhibits large improvement (16% and 11%) over SRAM baseline and SRAM-MRAM RWHCA. Figure 9B illustrates the power comparison of 3DRWHCA with SRAM baseline

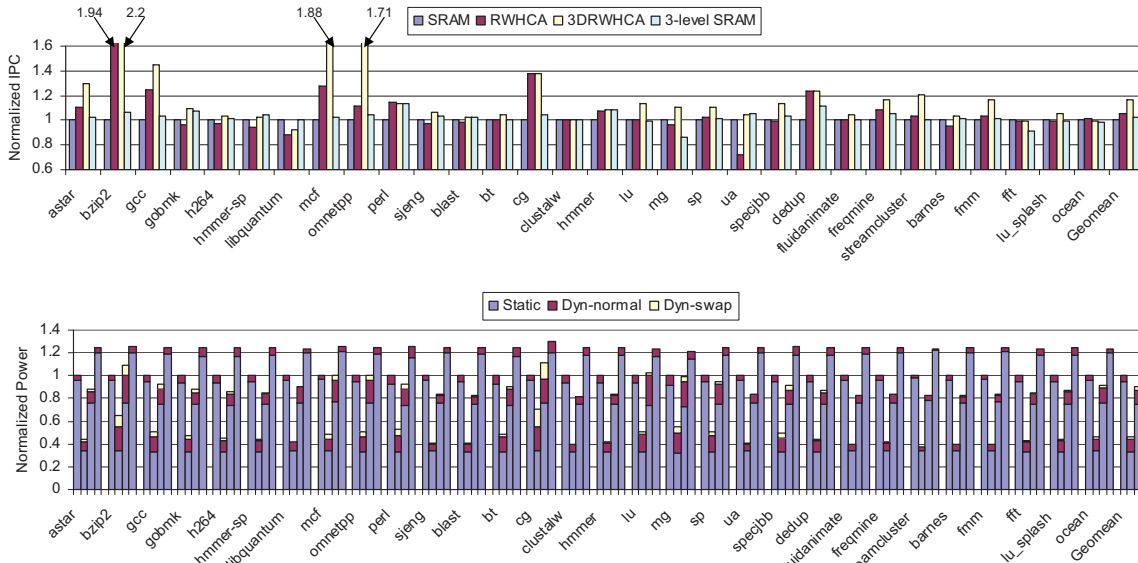


Fig. 9: 3DRWHCA performance (top, A) and power (bottom, B) comparisons.

and SRAM-MRAM RWHCA in 2D case. The result indicates that 3DRWHCA achieves 10% power reduction over SRAM baseline even with an extra L3 PRAM cache. Note that we also include the performance and power consumption of 3-level SRAM results in the Figure 9. RWHCA and 3DRWHCA still offer performance improvement over 3-level SRAM design with even more power saving compared to 2-level SRAM design.

VI. RELATED WORK

There are several NUCA studies for single core and chip multi-processors (CMP) in the literature [7], [9], [13]. Kim et al propose the novel NUCA concept for large caches and compare several DNUCA designs [13] in which data movement is based on generational promotion. Subsequently, distance associativity based NUCA, called NuRapid, is proposed in single core and multi-core designs [9]. NuRapid decouples data placement from tag placement by separating it from set associativity. In [7], transmission line based NUCA is presented for multi-core design and a prefetch scheme is evaluated for performance improvement. However, in these NUCA designs, the access latency differences are mainly from interconnect delays. In our RWHCA design, the latency as well as power differences are from disparate memory technologies. Additionally, our RWHCA is a hierarchical design. At a high level, RWHCA is made of cache regions of different sizes with differing memory technologies. At a base level, a cache region itself can be a conventional NUCA.

VII. CONCLUSIONS

In this paper, we have presented a read-write aware hybrid cache architecture to construct on-chip cache hierarchies with differing memory technologies. We have proposed and evaluated low-overhead intra-cache data movement policies and their hardware support to improve cache performance. For a collection of 30 workloads, the geometric mean of simulation results based on a hardware calibrated full-system simulator show that an RWHCA design can provide a geometric mean 55% power reduction and yet 5% IPC improvement over a

baseline SRAM cache design across a collection of 30 workloads. Furthermore, a 2-layer 3D cache stack (3DRWHCA) of high density memory technology within the similar chip footprint still gives 10% power reduction and boost performance by 16% IPC improvement over the baseline.

REFERENCES

- [1] <http://hpl.hp.com/research/cacti/>.
- [2] Specjbb2005 (java server benchmark). In <http://www.spec.org/jbb2005>.
- [3] Standard Performance Evaluation Corporation. 2006.
- [4] G. Atwood and R. Bez. Current status of chalcogenide phase change memory. In *Device Research Conference Digest*, volume 1, pages 29–33, 2005.
- [5] D. A. Bader, Y. Li, T. Li, and V. Sachdeva. BioPerf: a benchmark suite to evaluate high-performance computer architecture on bioinformatics applications. In *Proceedings of the 2005 IEEE international symposium on workload characterization*, pages 163–173, 2005.
- [6] D. Bailey, J. Barton, T. Lasinski, and H. Simon. The NAS parallel benchmarks. In *Technical report RNR-91-002 revision2*, pages 453–464, 1991.
- [7] B. M. Beckmann and D. A. Wood. Managing wire delay in large chip-multiprocessor caches. In *MICRO*, pages 319–330, 2004.
- [8] C. Bienia, S. Kumar, J. P. Singh, and K. Li. The PARSEC benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, October 2008.
- [9] Z. Chishti, M. D. Powell, and T. N. Vijaykumar. Optimizing replication, communication, and capacity allocation in CMPs. *SIGARCH Comput. Archit. News*, 33(2):357–368, 2005.
- [10] L. Chung. Cell design considerations for phase change memory as a universal memory. In *VLSI-TSA*, pages 132–133, 2008.
- [11] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li, and Y. Chen. Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement. In *DAC*, pages 554–559, 2008.
- [12] M. Hosomi, H. Yamagishi, T. Yamamoto, and et al. A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram. In *International Electron Devices Meeting*, pages 459–462, 2005.
- [13] C. Kim, D. Burger, and S. W. Keckler. An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches. In *ASPLOS-X*, pages 211–222, 2002.
- [14] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta. The SPLASH-2 programs: characterization and methodological considerations. *SIGARCH Comput. Archit. News*, 23(2):24–36, 1995.