Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement

Xiangyu Dong, Xiaoxia Wu,
Guangyu Sun, Yuan Xie
Pennsylvania State University
University Park, PA 16802
{xydong,xwu,gsun,yuanxie}@cse.psu.edu

Helen Li,
Yiran Chen
Seagate Technology
Bloomington, MN 55435
{helen.li,yiran.chen}@seagate.com

ABSTRACT

Magnetic Random Access Memory (MRAM) has been considered as a promising memory technology due to many attractive properties. Integrating MRAM with CMOS logic may incur extra manufacture cost, due to its hybrid magnetic-CMOS fabrication process. Stacking MRAM on top of CMOS logics using 3D integration is a way to minimize this cost overhead. In this paper, we discuss the circuit design issues for MRAM, and present the MRAM cache model. Based on the model, we compare MRAM against SRAM and DRAM in terms of area, performance, and energy. Finally we conduct architectural evaluation for 3D microprocessor stacking with MRAM. The experimental results show that MRAM stacking offers competitive IPC performance with a large reduction in power consumption compared to SRAM and DRAM counterparts.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types and Design Styles—*Advanced technologies, Memory technologies*

General Terms

Design, Performance

Keywords

MRAM, 3D Stacking

1. INTRODUCTION

Magnetic Random Access Memory (MRAM) has been considered as one of the most promising universal memory technologies due to its non-volatility, fast speed, zero standby power, and high density. The key element of MRAM cells is called Magnetic Tunnel Junction (MTJ), which is used for binary storage. Unlike traditional memory technologies, which use the electric charge as the information carrier, MTJ is based on magnetic storage. MTJ contains two ferromagnetic layers and one tunnel barrier layer (MgO). The direction of one ferromagnetic layer is fixed (reference layer) while the direction of the other one can be changed by passing a driving current (free layer). The relative magnetization direction of two ferromagnetic layers determines the resistance of MTJ. Usually, if two ferromagnetic layers have the same directions, the resistance of MTJ is low, indicating a "0" state; if two layers have different directions, the resistance of MTJ is high, indicating a "1" state. The conceptual view of the MTJ structure is shown in Fig. 1. More details about MRAM are presented in Section 3.

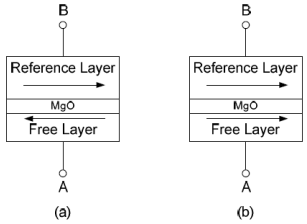

Figure 1: A conceptual view of MTJ structure. (a) Anti-parallel (high resistance), which indicates "1" state; (b) Parallel (low resistance), which indicates "0" state.

MRAM fabrication involves hybrid magnetic-CMOS process, (for example, MRAM process requires growing magnetic stack between metal layers), it may incur extra cost and additional fabrication complexity to integrate MRAM with conventional CMOS logic into a single 2D chip. The three-dimensional (3D) stacking technology is a promising means to solve this problem [1]. In 3D chips, multiple active device layers are stacked together with short and fast vertical interconnects. Among several benefits offered by 3D integrations, the mixed-technology stacking is especially attractive for stacking MRAM memory on top of CMOS logics, providing a mean to mount MRAM layer on top of logic layers, so that designers can take full advantage of the attractive benefits that MRAM provides. In particular, the integration of MRAM with microprocessors via 3D integration would offer a new perspective on the memory hierarchy. However, Most of the current research on MRAM are mainly in the fabrication, device modeling, or memory design areas. The architectural level benefits of using MRAM as a universal memory replacement for SRAM or DRAM are not well quantitatively evaluated yet.

The main objective of this paper is to evaluate MRAM performance/energy/density at the circuit level, as well as perform architectural level evaluation for stacking MRAM memory atop processors. First, we compare MRAM against SRAM and DRAM in terms of area, performance, and energy. Then we explore architectural evaluation for 3D microprocessor stacking with MRAM against SRAM/DRAM as L2 cache, MRAM as L3 cache, and MRAM against DRAM as main memory.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 provides the fundamental concept of MRAM and the circuit design for MRAM. Section 4 presents timing and

DAC 2008, June 8–13, 2008, Anaheim, California, USA.

energy models for MRAM-based cache. Section 5 describes the comparison results of MRAM against SRAM and DRAM in terms of area, latency, and energy. Section 6 shows architectural evaluation results when we stack MRAM on top of microprocessors. Finally, Section 7 presents the conclusion.

2. RELATED WORK

MRAM has been under development since the 1990s. In the last several years, MRAM has been proposed and developed by different companies. Since conventional toggle mode MRAM suffers slow write speed and high write power consumption, a second-generation MRAM design, called Spin-Torque Transfer MRAM (STT-MRAM, or SP-RAM) becomes the most popular design due to its better scalability property, higher speed, and lower power consumption [2, 3]. Recently, Zhao et al. proposed a macro-model for this hybrid magnetic-CMOS SPRAM design [4]. However, previous work mainly focus on device fabrication and device model. There are few research targeted at the architectural level for MRAM evaluation.

3D technologies have attracted considerable attention recently. With 3D manufacturing becoming mature, 3D architecture exploration attracts substantial number of researchers from industry and academia [1, 5, 6]. Bryan et al. have evaluated the 3D stacking (SRAM Cache or DRAM stack on top of a microprocessor) in terms of power and performance [5]. Loi et al. have analyzed the processor-memory hierarchy using 3D technologies from performance as well as thermal perspectives [6]. However, almost all the previous research on 3D stacking are based on traditional SRAM and DRAM technologies. Desikan et al. are the first one to consider on-chip MRAM as the replacement for DRAM memories [7, 8], but the MRAM they have used is based on previous generation of MRAM technology, which has scalability issue, and the physical characteristic is different from the one we discuss in this paper.

In this paper, we use the up-to-date generation of MRAM technology (Spin-Torque Transfer), and explore the MRAM and 3D stacking benefit by first comparing MRAM against SRAM and DRAM in aspects of area, power, and performance, and then by performing architectural evaluation when we replace the cache and the main memory with MRAM and stack MRAM on the top of microprocessors.

3. MRAM CIRCUIT DESIGN

This section introduces the physical mechanism of MRAM, and discusses the circuit design for MRAM.

3.1 Fundamental of MRAM

MTJ (Magnetic Tunnel Junction) is the storage element of MRAM cells. Normally, there are two ways to form an MRAM cell – the cross-point ("XPT") [9] structure and one transistor, one MTJ ("1T1J") structure [2, 3]. Because current XPT structure has very poor read performance [10], we only focus on the second one – 1T1J styled MRAM cell. For a 1T1J MRAM cell, as illustrated in Fig. 2, each MTJ is connected in series with a NMOS. The gate of the NMOS is connected to the word line (WL), and the NMOS is turned on if its connected MTJ needs read or write operations. As shown in Fig. 2, the source of the NMOS is connected to the source line (SL), and the free ferromagnetic layer is connected to the bit line (BL). The read and write operations are explained as follows:

- **Read Operation**: On a read operation, a negative voltage difference is applied on BL relative to SL [2]. This negative voltage is usually very small, which is -0.1V in our design. This voltage difference will lead to a current passing through the MTJ, which is small enough and will not invoke a disturbed write operation. The value of the current is mainly dependent on the resistance of the MTJ. Finally, a sense amplifier compares this current with a reference current and then decides whether a "0" or a "1" is stored in the selected MRAM cell.
- **Write Operation**: When writing "0" state into MRAM cells, positive voltage difference is established between SL and BL; When writing "1" state, vice versa. The current amplitude required to reverse the direction of the free ferromagnetic layer is determined by the size of MTJ and the writing pulse duration. The smaller the MTJ is or the longer the writing pulse is applied, the less the critical switching current is needed.

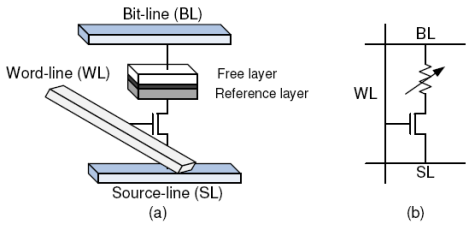

Figure 2: Demonstration of a MRAM cell. (a) Structural view. (b) Schematic view.

3.2 Design and Simulation of MRAM cell

To simulate the performance of a single MRAM cell, it is important to estimate its area first. As mentioned before, each 1T1J MRAM cell is composed of one NMOS and one MTJ. The size of NMOS is constrained by the current needed by the write operation, and the current amplitude is related to the writing pulse width. Because the critical current will increase dramatically when writing pulse is shorter than 10ns [2], we confine the writing pulse width within 10ns. According to the formula in [4] and scaled from the related 0.18μm MRAM work [2], the current amplitude we need for 90nm technology about 216μA.

The driving current of NMOS, I_{DS} is calculated by

$$I_{DS} = K\frac{W}{L}\left[\left(V_{GS} - V_{TH}\right)V_{DS} - \frac{V_{DS}^2}{2}\right] \quad (1)$$

if NMOS is working at the linear region; or calculated by

$$I_{DS} = \frac{K}{2}\frac{W}{L}\left(V_{GS} - V_{TH}\right)^2\left(1 + \lambda V_{DS}\right) \quad (2)$$

if NMOS is working at the saturation region. No matter in which region NMOS is working, the current driving ability of NMOS is proportional to its W/L ratio, which determines the size of NMOS transistors.

In this paper, we model the MRAM cell based on the equivalent schematic shown in Fig. 2(b). There are several ways to model the MTJ. In order to simplify the model, we use a conservative method treating the MTJ as a static resistor. To get the worst-case W/L ratio of NMOS, we assume that the equivalent resistance of MTJ remains at the high-resistance status during both "1" and "0" write operations. All MTJ-related parameters are scaled from [4]. We use HSPICE model to achieve the minimum W/L ratio of the NMOS which can drive 216μA current under 90nm technology.

According to the simulation result, in order to drive more than 216μA current, the W/L ratio of NMOS in each MRAM cell should be greater than 4.33. Therefore, the minimum size of NMOS should be 390nm$\times$90nm. The size of MTJ, which is the other part of each MRAM cell, can be aggressively scaled to F^2, where F is the feature size. Since the area is small and usually MTJs are built above

Table 1: MRAM Cache Read Latency

Cache size	H-tree input	Decoder+word-line	Bit-line	Sense Amplifier	Comparator	H-tree output	Total Read Latency
4MB MRAM data	$1.091ns$	$0.196ns$	$0.130ns$	$0.240ns$	N/A	$1.138ns$	$2.795ns$
4MB MRAM tag	$0.404ns$	$0.190ns$	$0.140ns$	$0.240ns$	0.033	$0.455ns$	$< 2.795ns$
16MB MRAM data	$2.401ns$	$0.274ns$	$0.130ns$	$0.240ns$	N/A	$2.461ns$	$5.506ns$
16MB MRAM tag	$0.557ns$	$0.215ns$	$0.140ns$	$0.240ns$	0.033	$0.608ns$	$< 5.506ns$

the drain regions of NMOS transistors, MTJ area is not included when we estimate the area of MRAM cells. After consideration of 90nm technology design rule, we set the layout area of single MRAM cell to 8.4F×4.4F.

4. MRAM-BASED CACHE

4.1 Organization and Area Model of MRAM Cache

As shown in Fig. 2(b), MRAM and SRAM have similar electrical interfaces from circuit designers' point of view. Both of them have word lines that are used to select the targeted storage elements, and bit lines that are used to transfer data. For SRAM cells, two differential bit lines (BL and BLB) are connected to each cell. For MRAM cells, although bit lines are single-ended, we can simply regard the combination of bit-lines (BL) and source-lines (SL) as the substitution. Since SRAM-based sense amplifier cannot be used directly in MRAM due to the single-ended bit line, a reference signal is required in MRAM-based sense amplifier. Usually, this reference signal is offered by a dummy MRAM cell. If the cache size is large enough, the layout overhead of dummy cells is negligible.

Thus, the organization of MRAM cache is almost the same as that of SRAM cache. A large MRAM array is divided into several small sub-arrays, and the traditional cache structure is applied to each sub-array. Sub-arrays are connected together with an H-tree. The number of rows and columns of sub-arrays and the size of each sub-array are all the parameters, which will impact the performance of the entire MRAM-based cache. In order to get the optimal partitioning pattern without building a new simulation tool, we modify the widely-used cache simulator, change all the SRAM-related parameters to those of MRAM. The timing and energy models in CACTI are also modified, which will be explained in the following subsections.

4.2 Timing Model of MRAM Cache

Read Timing:

Being aware that the read and write latencies of MRAM cell are asymmetric, we consider the read latency at first. Consistent with the SRAM cache timing model in CACTI, we divide the entire MRAM cache read latency into the following components:

- H-tree input delay,
- Decoder + word-line delay,
- Bit-line delay,
- Sense Amplifier delay,
- Comparator Delay (for tag part only),
- H-tree output delay.

We further assume that the H-tree and the decoder of MRAM caches are consistent with those of SRAM caches, so that the latency of these components can still be obtained from CACTI. The bit line delay is dependent on several intrinsic or parasitical RLC, thus we stick to our own HSPICE model, which is used in the previous section, to calculate the bit-line delay. Moreover, as discussed before, the sense amplifier of MRAM requires a reference signal, which will potentially increase the sensing delay, so we increase the original SRAM sense amplifier delay constant in CACTI by 20%. After these modification, the read latency components of MRAM cache are shown in Table 1.

Write Timing:

For MRAM cache write latency, it consists of the following components, and the delay of each component is shown in Table 2:

- H-tree input delay,
- Decoder + word-line delay,
- Minimum Writing Pulse Duration.

Table 2: MRAM Cache Write Latency

Cache size	H-tree input	Decoder +word-line	Writing Pulse	Total Write Latency
4MB	$1.091ns$	$0.196ns$	$10ns$	$11.287ns$
16MB	$2.401ns$	$0.274ns$	$10ns$	$12.675ns$

The results show that large difference exists between the bit line read time and write time. The reason is that the minimum writing duration is set to $10ns$ [2] while the reading delay is only $0.14ns$. Although this difference can be amortized by other latency components, our simulation shows that for a 4M MRAM cache, write operation latency is still $4X$ of read operation latency, and the gap for a 16M MRAM is about $2X$.

4.3 Energy Model of MRAM Cache

Besides the area/timing model, the energy model of MRAM cache is another critical model for further analysis. The MRAM energy model can be categorized into leakage energy and dynamic energy.

Leakage Energy:

Thanks to their non-volatility nature, MRAM cells do not consume any standby leakage power. Therefore, MRAM cache only has two sources of leakage power. One is the active leakage power consumed by MRAM cells, which is negligible because read and write operations are infrequent and only a small portion of MRAM cells (for a 16MB MRAM cache, the percentage is 0.01% estimated by our simulation) involved in each operation; the other one is the leakage power consumed by peripheral circuits, such as sense amplifiers, multiplexors, and decoders. Our simulation result shows that, for a 4M MRAM cache, the total leakage power is about 0.10W; for a 16M MRAM cache, the value is about 0.21W.

Dynamic Energy:

As for dynamic power, we need to notice that, comparing with SRAM, MRAM cells consume different amount of power during read/write operations. Thus, it is a must for us to replace this part of power estimation in CACTI with the data obtained from my own circuit-level MRAM HSPICe model.

Based on our HSPICE model of MRAM cells, the estimated dynamic power of MRAM cell is:

- For each MRAM read operation: $1.06 \times 10^{-15} J$/cell
- For each MRAM write operation: $2.60 \times 10^{-12} J$/cell

Table 3: Dynamic Energy Budget of a 16MB MRAM Cache

	MRAM cells	Peripheral circuits	Total Energy
Read	$0.026nJ$	$0.129nJ$	$0.155nJ$
Write	$2.833nJ$	$0.109nJ$	$2.942nJ$

Compared to the same parameter of SRAM cells, which is about $3.66 \times 10^{-15} J$, the MRAM read operation consumes slightly less energy; the MRAM write operation consumes three orders of magnitude more energy per operation. Combined the energy consumed by peripheral circuits, we can obtain the power estimation of MRAM cache modules. Table 3 shows an example.

5. COMPARISON OF MRAM, SRAM, AND DRAM

In this section, we compare MRAM with SRAM and DRAM in aspects of density, power consumption, and speed. Note that all caches simulated in this section are 16-way associative, 64-byte block, L2 caches.

5.1 Density

The density of memory media is very important to a memory system's cost. Usually, a single SRAM cell consists of 6 transistors (6T). Extracted from CACTI SRAM parameters, the area of each SRAM cell is about $146F^2$, where F is the feature size. DRAM uses 1T1C structures, whose minimum size of a single DRAM cell is about $21F^2$ under 90nm technology. The minimum layout area of MRAM cell is calculated in Section 3, which is about $37F^2$. If $90nm$ technology is applied, the area of 1 MRAM cell is about 25% of 1 SRAM cell, but it is about 1.7X of 1 DRAM cell. Certainly, if the peripheral circuits needed by DRAM cache is included, the area difference between DRAM cache and MRAM one will be reduced.

5.2 Power Consumption

The power profile of MRAM can be obtained from the energy model we developed in Section 3. In order to have a uniform evaluation metric, we use "energy consumption per operation" for dynamic power and "power per unit area" for leakage power, to make the comparison among MRAM, SRAM, and DRAM.

Compared to SRAM and DRAM, Table 4 lists the comparison. The data of SRAM and DRAM are obtained from original CACTI, and the data of MRAM are calculated from our modified CACTI.

Table 4: Power Comparison (90nm technology)

Cache	Dynamic energy	Leakage Power
SRAM	$0.151nJ$	$25.2mW/mm^2$
DRAM	$0.570nJ$	$8.5mW/mm^2$
MRAM	Read $0.155nJ$, Write $2.942nJ$	$2.7mW/mm^2$

The result shows that the non-volatility nature helps MRAM save lots of leakage power. Although MRAM requires much more power consumption for a write operation, this amount of dynamic power consumption is negligible compared to the large saving from standby leakage power. Thus, MRAM has the advantage on power efficiency over it counterparts.

5.3 Speed

When interconnect delay starts to dominate as the technology develops into ultra-sub-micron region. We compare the read access latency of SRAM/DRAM cache and MRAM cache which have similar area as MRAM-based caches, so that we can make a fair play. The result is shown in Table 5.

Table 5: Read Latency Comparison (90nm)

Cache	Area	Read Latency
4MB SRAM	$87mm^2$	$5.028ns$
16MB DRAM	$83mm^2$	$6.441ns$
16MB MRAM	$79mm^2$	$5.506ns$

The results demonstrate that, in terms of the read latency, MRAM cache is comparable with SRAM cache, and is much better than DRAM cache. However, combined with the relationship of MRAM cache read/write latency concluded in Section 3, the write latency of MRAM is much longer than that of SRAM and DRAM.

5.4 Impact of the Technology Scaling

Unlike the conventional toggle mode MRAM, the new STT MRAM solves the scalability issue [11]. Here, repeating the methodology used above, we evaluate the area, performance, and power of SRAM, DRAM, and MRAM caches under 65nm technology. Table 6 lists the results.

Table 6: Comparison (65nm)

	SRAM	DRAM	MRAM
Cache Size	$4MB$	$16MB$	$16MB$
Area	$44mm^2$	$49mm^2$	$38mm^2$
Latency	$4.659ns$	$5.845ns$	Read $4.693ns$ Write $12.272ns$
Dynamic Energy /operation	$0.103nJ$	$0.381nJ$	Read $0.102nJ$ Write $2.126nJ$
Leakage Power	$5.20W$	$0.52W$	$0.97W$

Compared with Table 4 and Table 5, we observe that like SRAM and DRAM, the density, the performance, and the dynamic power of MRAM can benefit from technology scaling as well. Moreover, despite of the technology scaling, the leakage power of MRAM caches still remains under a reasonable level.

5.5 Summary

The comparison among MRAM, SRAM and DRAM can be concluded in Table 7.

Table 7: Summary (90nm technology)

	SRAM	DRAM	MRAM
Density	Low	High	High
Dynamic Power	Low	Medium	High
Leakage Power	High	Medium	Very Low
Speed	Fast	Slow	Fast Read speed Very Slow Write Speed
Non-Volatility	No	No	Yes
Scalability	Yes	Yes	Yes

6. ARCHITECTURE LEVEL CACHE HIERARCHY DESIGN WITH STACKING MRAM

The 3D stacking technology makes it feasible to stack the memory atop the CPU. And for MRAM-based memory, which is fabricated by a hybrid magnetic-CMOS process, 3D mixed-technology stacking also provides a viable solution to integrate MRAM together with a CMOS-process processor. In this section, we discuss the architectural level benefit of stacking MRAM on a baseline 2D processor instead of stacking SRAM or DRAM memories.

Table 8: Processor Configuration

Processor Core	Alpha 21264 pipeline, Decode Width - 8 Issue Width - 8, Commit Width - 22
Processor Frequency	3.8GHz
Size of Structures	Fetch Queue - 8, RUU Size - 128 Load/Store Queue - 64
Functional Units	Integer - 8 ALUs, 4 Multipliers Floating Point - 2 ALUs, 2 Multipliers
Branch Predictor	Alpha 21264 Tournament Predictor
L1 D-Cache	16KB, 2-way associative, 64 byte line size, 2 cycles hit latency
L1 I-Cache	16KB, 2-way associative, 64 byte line size, 1 cycle hit latency

The baseline 2D processor simulated in our evaluation is Alpha 21264, whose configuration is shown in Table 8. We determine the size of L2 cache stacking on the processor by estimating the area of Alpha 21264. We examined a die photo of Alpha 21264 in 180nm technology [12]. Scaling to 90nm and removing the original L2 cache, we estimate that the remaining area of processor core and L1 cache is about $80mm^2$.

All simulation results are obtained from SimpleScalar simulator [13], with the benchmark programs from SPEC 2000. After obtaining the statistics of L2 cache access behaviors, we also calculate the power consumption for each configuration using CACTI power model for SRAM/DRAM and our MRAM energy model.

6.1 MRAM as Replacement of SRAM or DRAM L2 Cache

In this section, we evaluation the option of using MRAM as a replacement for SRAM or DRAM in L2 cache.

The first evaluation is to compare MRAM against SRAM and DRAM with similar area constraint. In order to obtain a compatible layout area to the baseline 2D processor, the size of L2 caches are set to be 4MB, 16MB, and 16MB for SRAM, DRAM, and MRAM, respectively. The area and the read latency of each cache configuration is shown in Table 5. In this evaluation, the longer write latency of MRAM-based cache is hidden with the write buffer, as implemented in many state-of-the-art processors.

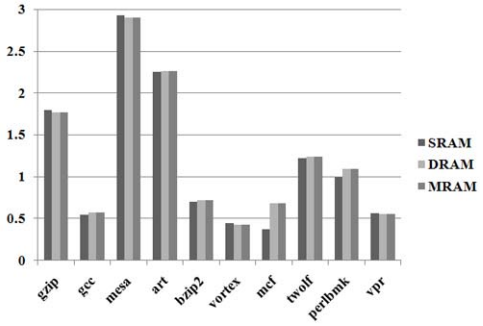

Figure 3: Performance comparison among 4MB-SRAM, 16MB-DRAM and 16MB-MRAM L2 cache (so that the areas are comparable). (Unit: IPC)

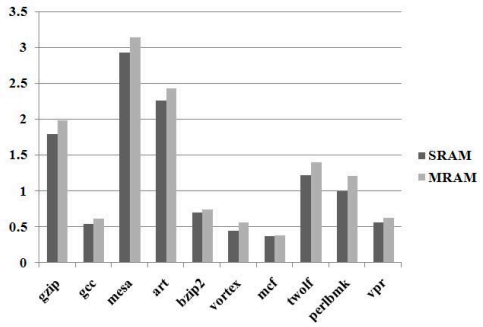

Figure 4: Performance comparison between 4MB-SRAM and 4MB-MRAM L2 cache.(Unit: IPC)

Fig. 3 shows the performance comparison among SRAM, DRAM, and MRAM in terms of IPC (Instruction Per Cycle). We observe that: (1) the performance of DRAM is about 10% lower than that of SRAM or MRAM, because of its relatively long read latency; (2) for most benchmarks, the competition between MRAM and SRAM seems to be a tie, since MRAM has the cache size advantage while SRAM has the access latency advantage.

Another experiment we conduct is to assume the same storage capacity for MRAM, SRAM, and DRAM (so that the specification of the cache does not change). For example, setting the storage size to be 4MB for all of the MRAM, SRAM, and DRAM based cache (the latency is shown in Table 1), we can obtain the IPC comparison result, which is shown in Fig. 4.

This simulation shows that the MRAM cache outperforms about 13% over the SRAM cache of the same size. However, when the MRAM cache size increases and the interconnect delay starts to dominate, the uniform cache access timing model kills the benefits of larger cache size, just like the result we show in Fig. 3.

Although estimated by our current UCA cache model, it seems that there is no obvious benefit to replace the SRAM L2 stacking cache with the MRAM cache of the same size, the real tiebreaker is the power consumption. Fig. 5 and Fig. 6 show the power comparison among SRAM, DRAM, and MRAM. Even though the MRAM has no advantage on dynamic power because of its large power consumption during write operation, MRAM's non-volatility behavior eliminates all standby leakage power consumption of memory cells, and save a large amount of power.

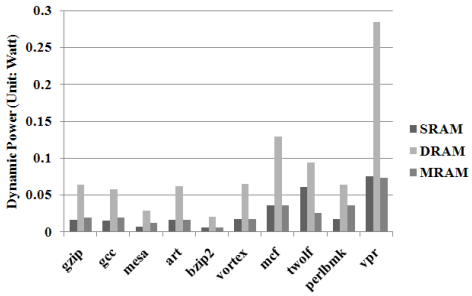

Figure 5: Dynamic Power Comparison among SRAM, DRAM and MRAM L2 cache. DRAM cache usually consumes more dynamic power because it requires refresh.

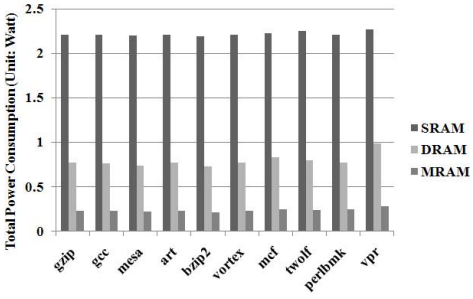

Figure 6: Total Power Comparison. MRAM cache only consumes 0.21W power in average.

On average, our simulation shows that a 16MB L2 MRAM cache only consumes 0.21W total power, and we conclude that using MRAM caches as the replacement of L2 caches can save total power by 89% compared with SRAM caches and 70% compared with DRAM caches.

6.2 Deep Memory Hierarchy with MRAM L3 Cache

Ultra-large L2 cache is not desirable in memory hierarchy because it increases the access latency. An alternative is to deepen the cache hierarchy by adding an L3 cache [14].

MRAM's low-power and high-density characteristics make it very suitable to be the L3 cache storage candidate. Using our MRAM model, we calculate that a 128MB, 4-bank, 16-way, 256-byte block cache only occupies the area of $161mm^2$, which is suitable to be stacked atop today's processors. Furthermore, our timing model shows that its read latency is only $15.82ns$, which is much less than the average memory access time.

To illustrate the performance improvement, we compare two configurations: (1) 2D processor with 4MB L2 SRAM cache; (2) 2D processor with 1MB L2 SRAM cache + 128MB MRAM L3 cache which is on the another die. The rest simulation parameters remain the same.

Simulation results are shown in Fig. 7. The speedup of IPC ranges from 0.03% to 108% for different benchmarks. For those benchmarks with high L2 miss rate, such as mcf and perlbmk, the improvements are incredible.

Being aware that under the real multi-task operation system environment, context switchings will cause more cache misses. In this situation, we believe the MRAM L3 cache will further improve performance. As for the power issues, estimated by our MRAM power

model, this amount of improvement only needs 0.4W more power consumption.

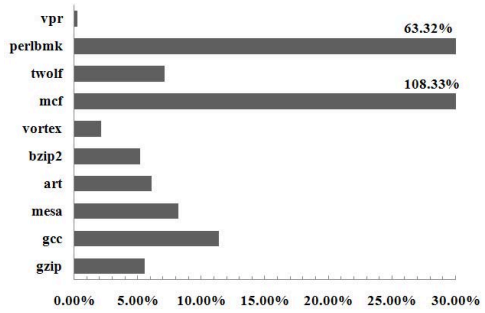

Figure 7: IPC Improvement with MRAM L3 Cache

6.3 Stacking MRAM as Replacement of Main Memory

3D mixed-technology stacking also enables DRAM stacking. Some related work has already explored the design possibility of stacking DRAM on-chip memory to reduce the access latency of off-chip memory module[14, 6].

Compared to DRAM, MRAM does not require periodical refresh time any more. Adapted from the work in [6], the access time of an improved dense DRAM module is about 43ns+cache access time, while estimated from the data part of our MRAM cache model, the read latency of a 128MB MRAM-based main memory is less than 16ns in total.

However, the density of the current MRAM memory still cannot reach that of the densest DRAM. Dense DRAM have smaller cell area, which is $6.6F^2$ demonstrated by [15]. Although XPT MRAM cell is as small as $1.5F^2$ [10], using the current Spin-Torque Transfer (STT) technology, the size of a 1T1J MRAM cell in this paper is $8.4F \times 4.4F$, which is $5.6X$ larger than the densest DRAM.

To consider the performance improvement of stacking main memory, simulations in [6] have already shown that stacking DRAM memories provides 19% and 40% boost in performance for integer and float-point benchmarks. Knowing that the memory footprints of most benchmarks are less than 128MB, we have reasons to believe that stacking a 128MB MRAM memory, whose read latency is shorter than that of DRAM, will lead to additional improvements.

The estimated area of 128MB MRAM memory is about $150mm^2$, which is fit to be stacked atop a 2D baseline processor. The memory capacity can be easily multiplied by stacking several MRAM memories using 3D multiple-layer stacking. Although the low-power nature of MRAM allows this multiple stacking without causing too many temperature issues, the increased latency will be the real overhead, and too many layers stacking together will eventually decrease the yield of 3D stacking chip.

In summary, current MRAM technology is not mature enough to be stacked as the main memory because of its capacity limitation, but a 512MB on-chip MRAM memory with 4-layer stacking is sufficient enough to support embedded applications for handheld processors, where MRAM's low-power nature can be another advantage. Certainly, we expect that the MRAM technology keeps scaling and finally meets the requirement of stacking MRAM as the main memory of high-level processors.

7. CONCLUSION

MRAM is considered as the most proming candidate for future universal memory because of its speed, power, and high density advantages. 3D stacking technology enables the integration of microprocessors and MRAM memories. In this paper, we describe the MRAM circuit design and present a cache model for MRAM-based cache design. We then evaluate the architectural level performance and energy benefits of the MRAM stacking memory. Compared with DRAM and SRAM, our experiments demonstrate that stacking MRAM atop a microprocessor can bring performance improvement, and achieve more than 70% power consumption reduction at the same time. The result of this work shows that MRAM is very promising to be a universal memory replacement in the near future.

8. REFERENCES

[1] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design space exploration for 3D architectures," *J. Emerg. Technol. Comput. Syst.*, vol. 2, no. 2, pp. 65–103, 2006.

[2] M. Hosomi, H. Yamagishi, and T. Yamamoto, "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram," in *International Electron Devices Meeting*, 2005, pp. 459–462.

[3] T. Kawahara, R. Takemura, and K. Miura, "2Mb Spin-Transfer Torque RAM (SPRAM) with Bit-by-Bit Bidirectional Current Write and Parallelizing-Direction Current Read," in *IEEE International Solid-State Circuits Conference*, 2007, pp. 480–617.

[4] W. Zhao, E. Belhaire, and Q. Mistral, "Macro-model of Spin-Transfer Torque based Magnetic Tunnel Junction device for hybrid Magnetic-CMOS design," in *IEEE International Behavioral Modeling and Simulation Workshop*, 2006, pp. 40–43.

[5] B. Bryan, A. Murali, and B. Ned, "Die Stacking (3D) Microarchitecture," in *International Symposium on Microarchitecture*, 2006, pp. 469–479.

[6] G. L. Loi, B. Agrawal, and N. Srivastava, "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," in *Design Automation Conference*, 2006, pp. 991–996.

[7] R. Desikan, C. R. Lefurgy, S. W. Keckler, and D. Burger, "On-chip MRAM as a high-bandwidth low-latency replacement for DRAM physical memories," Tech. Rep., 2002.

[8] R. Desikan, S. Keckler, and D. Burger, "Assessment of MRAM technology characteristics and architectures," Tech. Rep., 2002.

[9] W. Reohr, H. Honigschmid, and R. Robertazzi, "Memories of tomorrow," *IEEE Circuits and Device Mag.*, 2002.

[10] T. Maffitt, J. DeBrosse, and J. Gabric, "Design considerations for MRAM," *IBM Journal of Research and Development*, 2006.

[11] M. Oishi, "Spin Injection MRAM main focus at MMM," Tech. Rep., 2007.

[12] K. Krewel, "Alpha ev7 processor: A high-performance tradition continues," *Microprocessor Report*, 2005.

[13] D. C. Burger and T. M. Austin, "Simplescalar tool set, version 2.0," in *Computer Architecture News*, 1997.

[14] C. Liu, I. Ganusov, M. Burtscher, and S. Tiwari, "Bridging the processor-memory performance gap with 3d ic technology," *IEEE Design and Test of Computers*, vol. 22, no. 6, pp. 556– 564, 2005.

[15] T. Kirihata, G. Mueller, and M. Clinton, "A $113mm^2$ 600Mb/sec/pin 512Mb DDR2 SDRAM with vertically folded bitline architecture," in *IEEE International Solid-State Circuts Conference*, 2001, pp. 382–383.