Hybrid Cache Architecture Replacing SRAM Cache with Future Memory Technology

Suji Lee, Jongpil Jung, and Chong-Min Kyung
Department of Electrical Engineering
KAIST
Daejeon, Republic of Korea
ssooji555@kaist.ac.kr

Abstract—**Recently, hybrid cache architecture has become illuminated. As heterogeneous memory dies are stacked, it improves the performance of microprocessor enhanced in terms of power consumption and processing speed. This paper analyzed the hybrid cache architecture using different programs and memory types. SRAM is fixed for L1 cache memory, whereas DRAM, MRAM, and PRAM are the candidates for L2 cache memory. Each memory structure has the area satisfying the least Average Memory Access Time (AMAT) under a given area condition. Architecture composed of SRAM and MRAM shows 16.9% reduction in average memory access time and 15.2% of power reduction compared with that composed of homogeneous SRAM. Structure of SRAM and DRAM represents 33.0% reduction in power consumption, and that of SRAM and PRAM shows a potential to reduce area and power consumption due to their high density.**

I. Introduction

Hybrid die stacking is an emerging technology that multiple layers of dies are stacked with through-silicon-via (TSV). It takes improvements of speed, power consumption, and performance. Remarkable strength of 3D integration is additional reduction of area size, wire length, and performance progress. Heterogeneous memory dies in different memory types, such as dynamic random access memory (DRAM), magnetic random access memory (MRAM) and phase change random access memory (PRAM), can be stacked on a microprocessor [1, 2].

Static random access memory (SRAM) is used as cache memory in most microprocessors since SRAM has very high speed. However, SRAM has high leakage power consumption and low density compared with other types of memory. DRAM, MRAM, and PRAM are good candidates to replace SRAM cache. Speed of DRAM and MRAM is comparable with large cache capacity. Power consumption and density of MRAM and PRAM are superior as well [3].

Memory access pattern depends on characteristic of benchmark program. Some programs access memory excessively while other programs do infrequently. Variation of miss rate according to cache capacity also depends on the benchmark program. Some benchmarks show good performance with small cache capacity, while other benchmarks require large cache capacity. Exploiting these features, we design optimum cache architecture in aspects of speed, power and area. With different memory and program types, we obtain the best memory structure satisfying the least average memory access time and power consumption.

II. Related Works

In recent years, new Random Access Memory (RAM) technology has been proposed and developed by numerous companies for limitation of Si-based semiconductor. The most promising technologies are magnetic random access Memory (MRAM) and Phase change Random Access Memory (PRAM).

In the early 1990 MRAM was proposed, which has been improved in speed and power performance. It is operated by a storage element - Magnetic Tunnel Junction (MTJ) that shifts electric resistance by transferring direction of magnetic fields. The feature of MRAM is fast read speed, low power consumption and high density [4, 5, 6].

PRAM is another promising memory technology for non-volatile computer memory. It exploits the unique behavior of chalcogenide glass. With the application of heat applied by an electric current, the material can be altered between two states, crystalline and amorphous. Properties of PRAM are low power, slow speed, and very high density [7, 8]. The comparison of memory technologies are shown in Table I.

TABLE I. Comparison of Memory Technologies

Features	***SRAM***	***DRAM***	***MRAM***	***PRAM***
Density	Low	High	High	Very high
Speed	Very Fast	Fast	Fast read Slow write	Slow read Very low write
Dyn.Power	Low	Medium	Low read: High write	Medium read High write
Leak.Power	High	Medium	Low	Low

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

Hybrid cache architecture has attracted considerable attention recently. For its numerous opportunities, it attracts substantial number of researches from industry and academia of 3D stacking. Bryan *et al*. have evaluated the 3D stacking in terms of power and performance [1], and Loi *et al*. have analyzed the processor-memory hierarchy using 3D technologies for performance and thermal perspectives [2]. Also, Black *et al*. have researched on die stacking 3D micro-architecture made up only SRAM or DRAM [9].

As the technology has been developed, it is emerging that hybrid cache architecture combined with new memory types. Desikan *et al*. are the first ones who consider on-chip MRAM as replacement for DRAM memories [10, 11]. Dong *et al*. announced research on advanced performance of hybrid MRAM [6]. Hybrid PRAM is another attractive industry. Mounthaan *et al*. proposed hybrid cache architecture composed of PRAM and SRAM for power saving [12].

In this paper, analytical model for access time, power consumption, and area of cache memory is proposed. With this model, we compare various cases adopting different types of memory and benchmark programs. And then we find optimum architecture which results in the lowest average memory access time and low power consumption in limited area.

III. Problem Definition

A. *Target Architecture*

The microprocessor which consists of heterogeneous cache memory is targeted. In aspects of logical structure, it contains a core, 2 levels of on-chip cache, and external memory. Fig. 1 illustrates the target architecture. The forepart is composed of L1 and L2 cache, and the back part is off-chip main memory. L1 cache is fixed at SRAM for its greatly fast property. On the other hand, L2 cache can be selected in SRAM, DRAM, MRAM, or PRAM.

For simplicity of the problem, we assumed that external memory has fixed at access latency - 50*ns*. Also, total area of cache memory is limited in 100mm^2. Cache model of this paper is based on 45*nm* technology.

B. *Problem Definition*

As mentioned in previous section, we find the best memory architecture type which substitute SRAM L2 cache, and area size of L1 and L2 cache memory. SRAM is fixed for L1 cache memory, whereas there are four candidates for L2 cache memory; SRAM, DRAM, MRAM, and PRAM. As controlling area size of L1 and L2 cache, we could search the best cache capacity satisfying minimum average memory access time (AMAT) and small power consumption. To design practically, constraints are given that total power consumption and total area must be less than certain limits.

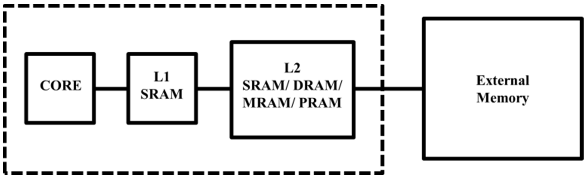


Figure 1. 3D hybrid cache architecture

For this problem, memory access time, power consumption, and area of cache memory are modeled on mathematical formulation. For each type of cache memory, we need to know how memory access time altered depending on cache capacity. We also need to know how the area and power consumption varies for cache capacity. These cache models are explained in section IV.

IV. Cache Model

Average memory access time (AMAT) and power consumption is modeled to evaluate performance of hybrid cache architecture. The model is formulated by combining memory access time (MAT) with a ratio of cache misses. The model of power consumption is composed of dynamic energy, static power, and miss rate.

A. *Modeling of miss rate*

In most cases, miss rate decreases as capacity of cache increase. However, dependency of miss rate to capacity varies according to the characteristic of benchmark program. Equation (1) is a general form to express miss rate in the function of capacity [13].

$$m(c) = \mu_0 \cdot c^{-\mu 1} \qquad (1)$$

where *c* is cache capacity, μ_0 and μ_1 are constants determined by the benchmark programs.

If value of μ_0 increases, it makes an increase of overall miss rate. It means that more data approaches to L2 cache due to less hits on L1 cache. On the other hand, μ_1 determines the dependency of miss rate and cache capacity. If μ_1 increases, it makes the miss-rate curve steeper and impacts of capacity increases. The value of μ_1 is normally between 0.3 and 0.7. Each program results in different AMAT and power consumption due to different miss rate with different values of parameters μ_0, and μ_1.

B. *Modeling of memory access time (MAT)*

The model of memory access time (MAT) is formulated based on the data extracted by tool CACTI 6.0. By curve fitting, the MAT according to cache capacity is expressed by (2). The model tracks original values with error of 1.71% on average, range from 0.08% to 8.36%. α, β, and γ are constants determined by memory technologies.

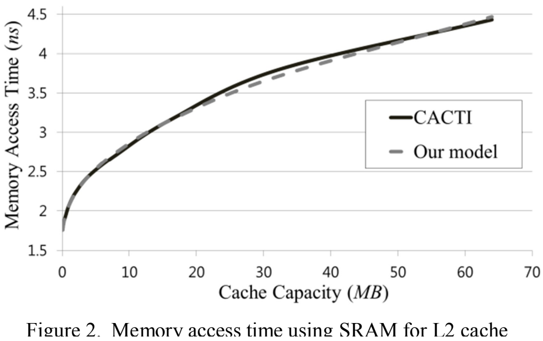


Figure 2. Memory access time using SRAM for L2 cache

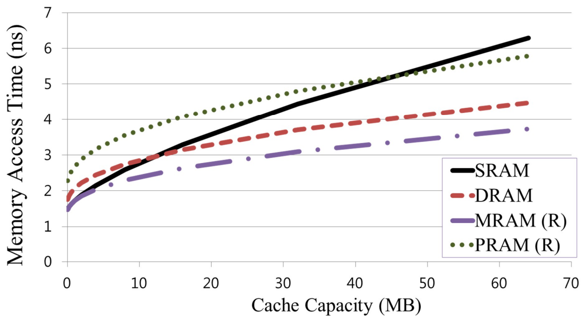


Figure 3. Memory access time of each memory type

$$T(c) = \alpha \cdot c^{\beta} + \gamma \qquad (2)$$

Fig. 3 depicts an elasticity of MAT with respect to cache capacity of each RAM types. SRAM is the fastest memory type in the range of small capacity. However, it becomes slower when capacity increases over few Mbytes. Therefore, speed of MRAM or DRAM could be comparable to that of SRAM due to their large cache size.

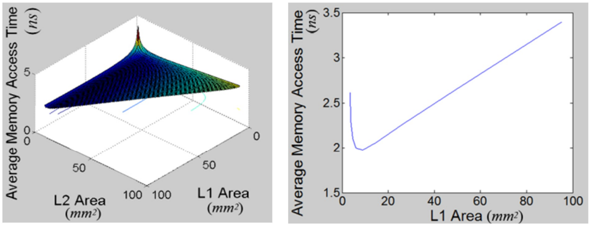
C. Modeling of average memory access time (AMAT)

Average memory access time (AMAT) is formed with MAT and the miss rate as shown in (3). In (3), c_1 and c_2 are capacity of L1 and L2 cache, and h and m are hit rate and miss rate, respectively. Also, T_1, T_2, and T_{ext} are access time of L1, L2 cache, and external memory. In (4), T^r, T^w are reading and writing latency, and ρ is a parameter which depends on program features that shows the ratio of the read access from all memory references. The access time is independent to size of ρ for SRAM and DRAM since the access time of reading and writing is same. On the other hand, access time of MRAM and PRAM is altered since the differences of reading and writing latency are significant.

$$\begin{aligned} AMAT = {} & h(c_1)\cdot T_1(c_1) \\ & + m(c_1)\cdot(h(c_2)\cdot T_2(c_2) + m(c_2)\cdot T_{ext}) \end{aligned} \qquad (3)$$

$$T_i(c_i) = \rho\cdot T^{r}{}_{i}(c_i) + (1-\rho)\cdot T^{w}{}_{i}(c_i) \qquad (4)$$

Fig. 4(a) shows a three-dimensional graph which illustrates elasticity of AMAT in (3) and (4) where DRAM is used for L2 cache. Fig. 4(b) illustrates a section of Fig. 4(a) where total area is fixed for maximum size and lowest AMAT value. The curve shows trend of increasing after sharp reduction. In low L1 range, AMAT decreases significantly as area increases. It means that the smaller size of L1 cache is given, the higher miss rate is formed. Consequently, average access time becomes longer as slow L2 cache accesses more. On the other hand, in high L1 range, AMAT increases as L1 area increases even if L1 hit rate becomes greater. This result shows L1 access latency is significant due to its large cache capacity.



(a) (b)

Figure 4. Average memory access time with (a) overall 3D graph, (b) section of maximum area

D. Modeling of power consumption

By curve fitting grounded on extracted data in CACTI 6.0, power consumption of cache memory is modeled as linear functions shown in (5) and (6). δ, θ, ρ, and σ are parameters obtained from memory technologies.

$$E_{dyn}(c) = \delta \cdot c + \theta \qquad (5)$$

$$P_{static}(c) = \rho \cdot c + \sigma \qquad (6)$$

Power consumption is formulated as (7), (8), and (9). P_1, P_2, E_{dyn1} and E_{dyn2} indicate power or dynamic energy consumed by L_1 and L_2 cache per access. N_{access} is the number of access per second. $P_{static1}$ and $P_{static2}$ are static power consumed by L_1 and L_2 cache respectively.

$$\mathrm{P}(\mathrm{c}_1, \mathrm{c}_2) = P_1(c_1) + P_2(c_2) \qquad (7)$$

$$P_1(c_1) = N_{access} \cdot h(c_1) \cdot E_{dyn1}(c_1) + P_{static1}(c_1) \qquad (8)$$

$$\begin{aligned} P_2(c_2) = {} & N_{access} \cdot m(c_1) \cdot h(c_2) \cdot E_{dyn2}(c_2) \\ & + P_{static2}(c_2) \end{aligned} \qquad (9)$$

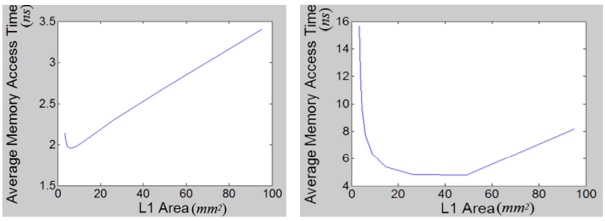
The graphs of power consumption are almost linear to the area of L1 cache for all types of memory (S, D, M, PRAM) due to dominant power consumption of L1 cache (SRAM).

V. Experimental Results

In this section, we examine AMAT and power consumption with respect to program and memory types.

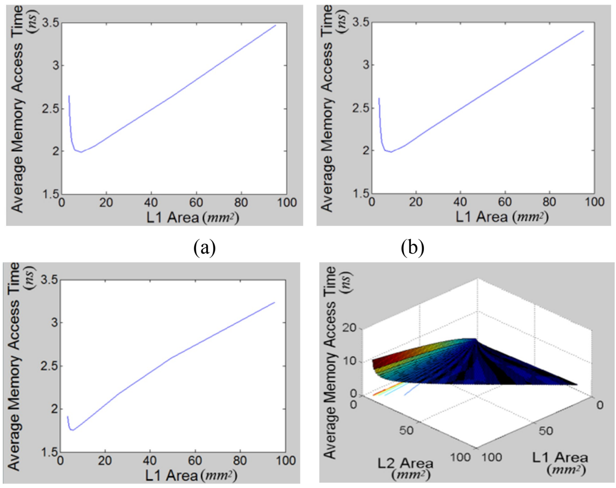
A. Results with program types

Each program determines approaching rate of L1 and L2 cache, which affects value of AMAT consequently [14]. Results of AMAT in different program are shown in Fig. 5. Miss rate of *equake* in Fig. 5(a) is much smaller than that of *face_rec* in Fig. 5(b). If miss rate is low, most of access hits L1 cache regardless of L1 cache size. Thus, AMAT increases as L1 capacity increases. On the other hand, more data hits slow L2 cache if miss rate is high. For the reason, Fig. 7 illustrates optimum area size is located on middle of the graph. These features make possible to design reconfigurable hardware containing characteristic of various programs. It will be helpful for 3D multi-core processor design. If each cache



(a) (b)

Figure 5. Average memory access time for each program: (a) equake, (b) face_rec



(a) (b)

(c) (d)

Figure 6. Average memory access time for L2 cache using (a) SRAM, (b) DRAM, (c) MRAM, and (d) PRAM

level of multi-core system is divided into several partitions, each partition could be adjusted dynamically. Therefore grand average memory access time could be minimized.

B. Results with memory types

Each memory has intrinsic characteristic of AMAT. Fig. 6 illustrates AMAT when S, D, M, and PRAM are adopted for L2 cache. Fig. 6 (a) to (c) represents AMAT with maximum area, whereas (d) displayed configurable area. These result shows the optimum partitioning is changed for memory types. By altering types of memory in 3D architecture, we can design better architecture with improved performance.

As a result of simulation, the least AMAT is given when MRAM is used as L2 cache. Fig. 7 illustrates the average memory access time of several benchmark programs. Values are normalized with respect to the values of SRAM. On average, it shows 16.9% reduction in AMAT and 15.2% reduction in power consumption. Using DRAM is rational alternatives for lowest power consumption. It shows 33.0% reduction in power consumption on average, while AMAT is not reduced remarkably (2%).

In the case of PRAM, minimum point of AMAT is located with minimum area of PRAM. The reason is that PRAM is too slow for a cache memory as shown in Fig. 3. PRAM is not suitable for on-chip L2 cache. PRAM has very high density, thus it is valuable for mass storage or very small area.

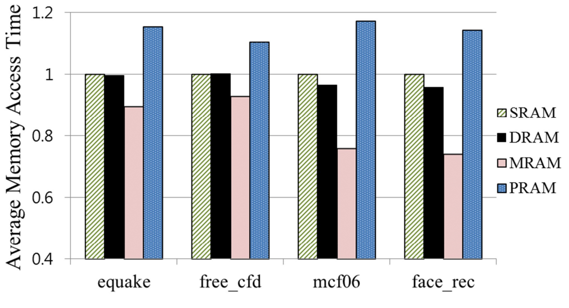


Figure 7. Average memory access time for each memory type with various program. Values are normalized with respect to SRAM.

VI. Conclusions

In this paper, we have compared different types of hybrid cache architecture. SRAM is used for level 1 cache, whereas S, D, M, and PRAM are selected for level 2 cache memories, respectively. On each case, Average memory access time (AMAT) and power consumption are examined with several benchmark programs. The cache memory architecture (L1: SRAM, L2: MRAM) offers 16.9% AMAT reduction and 15.2% power saving than homogenous SRAM architecture. The architecture (L1: SRAM, L2: DRAM) offers 33.0% power saving than homogenous SRAM architecture, which is the most effective structure for reducing power consumption. PRAM is not suitable for on-chip L2 cache in this model.

Acknowledgment

This work was supported by the Smart IT Convergence System Research Center funded by the Ministry of Education, Science and Technology as Global Frontier Project (SIRC 2011-0031863) and IT R&D program of MKE/KEIT [KI002134, Wafer Level 3D IC Design and Integration].

References

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

[2] 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.

[3] W. Xiaoxia, L. Jian, Z. Lixin, “Hybrid Cache Architecture with Disparate MemoryTechnologies,” in International Sysmposium on Computer Architecture, 2009, pp 34-45.

[4] 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.

[5] 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.

[6] D. Xiangyu, W. Xiaoxia, L. Helen, “Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement,” in Design Automation Conference, 2009, pp 554-559.

[7] S.R. Ovshinsky, “Reversible Electrical Switching Phenomena in Disordered Structures,” Physical Review Letter, Vol.21, No.20, 1968, pp.1450.

[8] S. Lai and T. Lowrey, “OUM - A 180nm Nonvolatile Memory Cell Element Technology, For Stand Alone and Embedded Applications,” IEDM Tech. Dig., 2001, pp.803.

[9] B. Black, M. Annavaram, E. Brekelbaum, J. DeVale, L. Jiang, G. Loh, D. McCauley, P. Morrow, D. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb. Die Stacking (3D) Microarchitecture. In Proceedings of MICRO-39, December 2006.

[10] 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.

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

[12] M. Mouthaan. “Mass production phase-change RAM in June,” http://www.hardware.info, May 2009.

[13] A. Hartstein *et al.*, "Cache miss behavior: is it √2," in Proc. Computing frontiers, pp. 313-320, 2006.

[14] R. W. Quong, and W. Lafayette, “Expected I-cache miss rates via the gap model,” in International Symposium on Computer Architecture, 1994, pp 374-383