A Heterogeneous Memory Organization with Minimum Energy Consumption in 3D Chip-Multiprocessors

†Arghavan Asad, ‡Salman Onsori
†‡Computer Engineering Department
†Iran University of Science and Technology, Tehran, Iran
‡Bilkent University, Ankara, Turkey
ar_asad@comp.iust.ac.ir, salman.onsori@cs.bilkent.edu.tr

*Kaamran Raahemifar, †Mahmood Fathy,
†Mohammad Reza Jahed-Motlagh
*Electrical and Computer Engineering Department
*Ryerson University, Ontario, Canada
kraahemi@ee.ryerson.ca, {mahfathy, jahedmr}@iust.ac.ir

Abstract— **Main memories play an important role in overall energy consumption of embedded systems. Using conventional memory technologies in future designs in nanoscale era cause a drastic increase in leakage power consumption and temperature-related problems. Emerging non-volatile memory (NVM) technologies offer many desirable characteristics such as near-zero leakage power, high density and non-volatility. They can significantly mitigate the issue of memory leakage power in future embedded chip-multiprocessor (eCMP) systems. However, they suffer from challenges such as limited write endurance and high write energy consumption which restrict them for adoption in modern memory systems. In this article, we propose a stacked hybrid memory system for 3D chip-multiprocessors to take advantages of both traditional and non-volatile memory technologies. For reaching this target, we present a convex optimization-based model that minimizes the system energy consumption while satisfy endurance constraint in order to design a reliable memory system. Experimental results show that the proposed method improves energy-delay product (EDP) and performance by about 44.8% and 13.8% on average respectively compared with the traditional memory design where single technology is used.**

Keywords*— *Heterogeneous memory system, Non-Volatile Memory (NVM), Convex-optimization problem, embedded Chip-Multiprocessor (eCMP), Dark silicon.

I. INTRODUCTION

Chip multiprocessor (CMP) architectures have been widely used to meet growing demands on performance in embedded systems. The increase in the number of cores in eCMPs comes with an increase in energy consumption. Since embedded systems are generally limited by battery lifetime, energy consumption is an essential and important constraint in these systems. It is widely acknowledged that energy consumption of memory systems is a significant contributor in overall system energy due to integration of increasingly larger memory closer to the processor. Therefore, there is a critical need to considerably reduce energy consumption in memory architectures. Memory energy consists of two components: 1) leakage, and 2) energy of the read/write access. In order to reduce memory energy, it is needed to address both the leakage and dynamic energy. Moreover, 42% of overall energy dissipation in the 90*nm* generation is consumed by leakage energy [1] and this value can exceed above 50% in 65*nm* technology [4]. Hence, leakage energy has become comparable to dynamic energy in current generation memory modules and soon exceed dynamic energy in magnitude if voltage and technology are scaled down any further [3]. Consequently, architecting energy efficient memory systems with the lowest leakage energy is especially critical for embedded systems.

Due to physical limitation of two dimensional integration, 3D CMPs receive a lot of attention in these days. 3D integration technology compared with 2D designs reduces global interconnection wire-length which results in low power consumption and short communication latency. To reduce power consumption of CMPs and improve their performance and memory bandwidth, CMP architectures with 3D stacked memory system have been proposed [7]. Stacked traditional memory systems on the core layer may cause a drastic increase in performance degradation, power density and temperature-related problems such as negative bias temperature instability (NBTI) [14].

Non-volatile memories (NVM) as a new emerging memory technology are potentially attractive to design new classes of memory systems due to their benefits such as higher storage density, near zero leakage power consumption and high resilient against soft errors. STT-RAM as a promising candidate of NVM technology combines the speed of SRAM, the density of DRAM and the non-volatility of Flash memory. In addition, excellent scalability and very high integration with conventional CMOS circuits are the other superior characteristics of STT-RAM [2]. Although NVMs have many benefits as described above, drawbacks such as high write energy consumption, long write latency and limited write endurance prevent them from being directly used as a replacement for traditional memories in embedded systems.

In order to overcome the mentioned disadvantages in this paper, we use SRAM and STT-RAM as two different types of memory banks in the stacked memory layer in a 3D eCMP. This heterogeneous point of view leads us to the best design possibility with using benefits of both memory technologies. In this work, we use Non Uniform Memory Architecture (NUMA) stacked directly on top of the core layer in the proposed eCMP.

978-1-4673-8721-7/16/$31.00 ©2016 IEEE

Recently, dark silicon challenge is emerging as a trend in VLSI technology. The rise of utilization wall due to thermal and power budgets restricts active components and results in a large region of dark silicon. Uncore components such as memory and cache subsystems play an important role in consuming large portion of power consumption. Thereby, power management of these uncore components can be critical to maximize the design performance in the dark silicon era. How to design uncore components to combat dark silicon in future eCMP is largely unexplored in literatures and prior work only focused on energy efficient core design [15][16]. In this regard, heterogeneous architectures can be a promising solution to tackle the challenges of multicore scaling in the dark silicon era because of slight improvement in CMOS technology. NVMs can integrate with CMOS circuits efficiently in energy-efficient designs.

To the best of our knowledge, this paper is the first work targeted at energy efficient heterogeneous memory architecture design based on a convex optimization approach for future eCMPs. We exploit 3D die-stacking and emerging NVMs to design a high performance 3D eCMP architecture for minimizing energy consumption as a solution to combat dark silicon challenge.

Figure 1 shows an overview of the proposed design using an example of a 16 homogeneous core in the lower layer and hybrid memory architecture in upper layer. In the proposed heterogeneous memory system, STT-RAM as a well-known candidate of NVMs is incorporated with SRAM banks in the second layer.

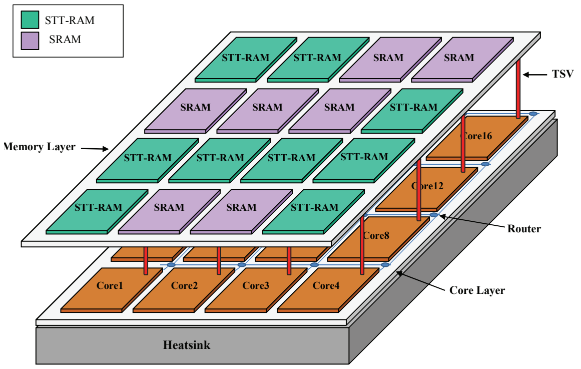

Fig. 1. An overview of the proposed architecure.

This paper makes the following novel contributions:

- We provide a convex optimization based platform to design a heterogeneous memory system consists of NVM and SRAM memory banks.
- Our proposed model can optimally find the number of SRAM and STT-RAM memory banks in the memory layer based on access behavior of mapped applications to minimize energy consumption.
- We propose an analytical endurance model for STT-RAM memory banks to use in our optimization problem for architecting a high-endurance heterogeneous stacked memory system.

The rest of this paper is organized as follows. A brief background is explained in Section II. Section III describes related works. In Section IV, the details of convex optimization-based problem and its formulation are investigated. In Section V, evaluation results are presented. Finally, Section VI concludes the paper.

II. Background

STT-RAM has been one of the most popular NVM structures due to its scalability in sub-nanometer technology and the low writing current in comparison with the conventional Magnetic Random Access Memory (MRAM).

As it is illustrated in Figure 2, for performing a read operation from the STT-RAM cell, the NMOS transistor will be turn ON and a little voltage between bit line and source line will be set. This voltage causes a current in the MTJ. The amount of this current depends on the state of the MTJ. A current sensor senses and compares it with a reference current. As a result, the logic value of that cell will be determined.

For a write operation, with respect to the value of the cell, the amount of the current would be varied. In order to write a the logic value of ‘0’ a positive current and for writing the logic value of ‘1’, negative current is injected between bit line and source line. The amount of the current for a reliable write operation is known as threshold current which is depended on the type of material used to construct the MTJ and its shape[24].

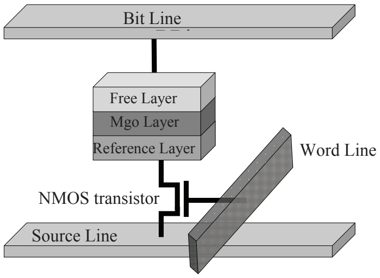

Fig. 2. Structure of a STT-RAM.

III. Related Work

Recent studies [8], [9] have proposed hybrid architectures, wherein the SRAM is integrated with NVMs to use advantages of both technologies. Energy consumption is still a primary concern in embedded systems since they are limited by battery constraint. Several techniques have been proposed to reduce energy consumption of hybrid memory architectures in embedded systems. Fu et al. [12] presented a technique to improve energy efficiency through a sleep-aware variable partitioning algorithm for reducing the high leakage power of hybrid memories. Hajimiri et al. [11] proposed a system-level design approach that minimizes dynamic energy of a NVM-based memory through content aware encoding for embedded systems. Our work is different from all these prior works as we focus on placement of SRAM and STT-RAM banks in a stacked memory architecture in future eCMPs to minimize energy consumption with using a convex-optimization based approach.

As mentioned before, there are some obstacles for employing STT-RAM without integration with traditional technologies in modern memory systems. One of them is limited number of write operations. After limited number of write operations, it is not possible to write a value into a STTRAM cell, and just the stored value can be read [25]. Number of researches presented different techniques to combat endurance problem of NVMs. Qureshi et al. [10] proposed a wear-leveling technique for a PRAM-based memory system to enhance the lifetime. Wang et al. [5] proposed an algorithm to distribute write events evenly in the address space of scratchpad memory to extend the endurance of NVM. Lue et al. [6] presented a writing technique called Min-Shift to reduce the total number of writes onto NVM and enhance lifetime of NVMs. Hu et al. [13] proposed a software wear leveling technique to extend life time of NVM in hybrid memory structure of embedded systems. For the first time in this work, we present an endurance model for NVM technologies. This endurance model is used as a constraint in the proposed optimization problem to design a high endurance heterogeneous memory system with minimum energy consumption.

IV. Optimization Model

In this section, we formulate our energy optimization problem to design a minimum energy heterogeneous memory structure in a 3D eCMP. Figure 3 shows the block diagram of our model for designing the proposed hybrid memory architecture with minimum energy consumption.

Outputs of our optimization problem are 1) finding optimal number of SRAM and STT-RAM memory banks based on memory access behavior of mapped applications with respect to the endurance constraint, 2) appropriate placement of SRAM incorporated with STT-RAM banks in the memory layer to minimize energy consumption.

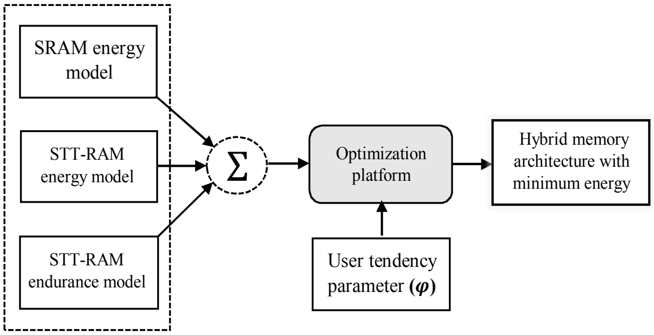

Fig. 3. Overview of our model.

Table I gives the constant terms used in our convex formulation. To solve the models, we use CVX [17], an efficient convex optimization solver.

Assuming that P denotes the total number of processor cores, M_{sr} the total number of SRAM memory banks, M_{st} the total number of STT-RAM memory banks, $(C_X\ , C_Y)$ the dimensions of the chip, $(P_X\ , P_Y)$ the dimensions of the processor core. Our approach uses 0-1 variables to specify the coordinates of each memory bank and processor core.

Note that in this work we do not consider application mapping in our proposed model and applications are randomly mapped onto cores in the core layer.

TABLE I. Constant Terms Used in Our Optimization Problem

Constant	Definition
P	Number of processor cores
M_{sr}	Total number of SRAM memory banks
M_{st}	Total number of STT-RAM memory banks
C_X, C_Y	Dimensions of the chip
P_X, P_Y	Dimensions of a processor core
SR_X, SR_Y	Dimensions of a SRAM memory bank
ST_X, ST_Y	Dimensions of a STT-RAM memory bank
l	Index of layers
N	The Number of lines in STT-RAM memory bank
$FREQ_{p,m,r}$	Number of read access to memory bank m by processor p
$FREQ_{p,m,w}$	Number of write access to memory bank m by processor p
φ	Using STT-RAM versus SRAM ratio
$E_{read_{sr}},\ E_{write_{sr}}$	Dynamic energy consumption per read and write access by the SRAM memory bank
$E_{read_{st}}, E_{write_{st}}$	Dynamic energy consumption per read and write access by the STT-RAM memory bank
$P_{static_{sr}}$	Static power consumed by each SRAM memory bank at maximum temperature limit
$P_{static_{st}}$	Static power consumed by each STT-RAM memory bank at maximum temperature limit
$\tau_{sr}^{r}, \tau_{sr}^{w}$	Read latency and write latency of SRAM cache bank
$\tau_{st}^{r}, \tau_{st}^{w}$	Read latency and write latency of STT-RAM cache bank
$Endurance_{STT-line}$	Maximum write number for each line of the STT-RAM memory bank

We use STC and SRC to identify the coordinates of a memory bank. We have two types of memory banks, SRAM and STT-RAM, so we have two variables.

- $SRC_{sr,x,y,l}$: indicates whether a SRAM bank is in (x, y) in layer $l = 2$.
- $STC_{st,x,y,l}$: indicates whether a STT-RAM bank is in (x, y) in layer $l = 2$.

The mapping between coordinates and blocks is ensured by variable $MMap$ for the memory banks in second layer. That is,

- $SRMAP_{sr,x,y,l}$: indicates whether coordinate (x, y) is assigned to a SRAM bank in layer $l = 2$.
- $STMAP_{st,x,y,l}$: indicates whether coordinate (x, y) is assigned to a STT-RAM bank in layer $l = 2$.

A memory bank needs to be assigned to a unique coordinate. In Equation (1), i and j correspond to the x and y coordinates, respectively:

$$\sum_{i=1}^{C_X-1} \sum_{j=1}^{C_Y-1} (SRC_{sr,i,j,l} + STC_{st,i,j,l}) = 1, \quad \forall sr, \forall st, l = 2 \qquad (1)$$

$$STMAP_{st,x,y,l} \geq STC_{st,x1,y1,l}$$
$$\forall p, x1, y2, y1, y2 \text{ such that}$$
$$x1 + ST_X \geq x > x1 \ \ and \ y1 + ST_Y \geq y > y1, \ \ l = 2 \qquad (2)$$

$$SRMAP_{sr,x,y,l} \geq SRC_{sr,x1,y1,l}$$
$$\forall p, x1, y2, y1, y2 \text{ such that}$$
$$x1 + SR_X \geq x > x1 \ \ and \ \ y1 + SR_Y \geq y > y1, \ l = 2 \qquad (3)$$

Also, sum of used STT-RAM and SRAM banks in second layer is equal to P as follow:

$$\sum_{x=0}^{C_X-1} \sum_{y=0}^{C_Y-1} (\sum_{i=1}^{M_{sr}} SRMAP_{i,x,y,l} + \sum_{i=1}^{M_{st}} STMAP_{i,x,y,l}) = P \quad , l = 2 \ (4)$$

In this work, the size of memory banks in the upper layer is same as processor cores in the lower layer.

In order to prevent multiple mappings of a coordinate in our grid, we force a coordinate in first layer to belong a single processor core and a coordinate in second layer to belong a memory bank (SRAM or STT-RAM).

$$\sum_{i=1}^{M_{sr}} SRMAP_{i,x,y,l} + \sum_{i=1}^{M_{st}} STMAP_{i,x,y,l} = 1, \quad \forall x, y, l = 2 \qquad (5)$$

The static power dissipation depends on temperature. Since this optimization approach is solved at design time, we consider pessimistic worst-case temperature assumption and calculate $P_{static_{sr}}$ and $P_{static_{st}}$ at maximum temperature limit.

$$P_{static} = \sum_{i=0}^{C_X-1} \sum_{j=0}^{C_Y-1} (\sum_{k=1}^{M_{sr}} SRC_{k,i,j,l} \times P_{static_{sr}} + \sum_{k=1}^{M_{st}} STC_{k,i,j,l} \times P_{static_{st}}), \qquad l = 2 \quad (6)$$

We consider endurance problem of STT-RAM in our convex model. Hence, we exploit an endurance constraint for optimal placement of SRAM and STT-RAM memory banks. In our model, if placing a STT-RAM memory bank in the special position leads to destruction of more than half lines of that memory due to writing frequency of cores, STT-RAM memory bank is not chosen for that position. Figure 4 illustrates workflow of the endurance model. This endurance constraint can be expressed as follows:

$$\frac{\sum_{i=1}^{P} FREQ_{i,st,w}}{Endurance_{STT-line}} \times STC_{st,x,y,2} < \frac{N}{2}, \qquad \forall x, y, st \qquad (7)$$

Having specified the necessary constraints in our convex formulation, we next give the objective function. The goal of our objective function is to minimize energy consumption of the stacked heterogeneous memory architecture in the target 3D CMP with respect to the endurance constraint. A weighted objective function is considered to capture the potential effects on power consumption and overall performance. This is achieved by the φ constant which is used as a knob for choosing SRAM versus STT-RAM bank in each x and y coordinates in the memory layer. As mentioned before, STT-RAM in comparison with SRAM technology is slower with higher density and near-zero leakage power. In this regard, STT-RAM banks are applicable for memory-intensive blocks and SRAM banks are applicable for computation-intensive blocks. Therefore, with changing φ value, there is a possibility for having an optimized design based on the tendency of designer.

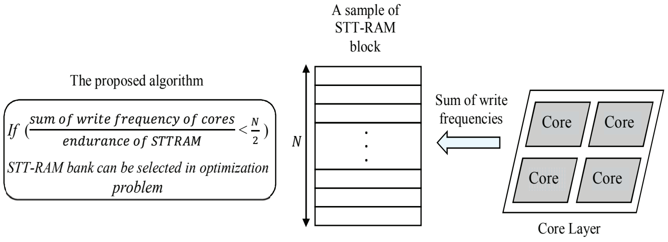
A sample of STT-RAM block
The proposed algorithm
If (sum of write frequency of cores / endurance of STTRAM < N/2)
STT-RAM bank can be selected in optimization problem
N
Sum of write frequencies
Core
Core
Core
Core
Core Layer

Fig. 4. Overview of endurance model.

Static energy of SRAM and STT-RAM banks for each write and read operations are defined as multiplication of their static power consumptions and read and write durations.

$$E_{static_{sr}} = (\tau_{sr}^{r} + \tau_{sr}^{w}) \times P_{static_{sr}} \qquad (8)$$

$$E_{static_{st}} = (\tau_{st}^{r} + \tau_{st}^{w}) \times P_{static_{st}} \qquad (9)$$

In Equation (10), $E_{read_{sr}}$, $E_{write_{sr}}$, $E_{read_{st}}$ and $E_{write_{st}}$ indicate dynamic energy consumed by SRAM and STT-RAM banks per read and write access. $E_{dynamic}$, the dynamic energy consumption of the proposed heterogeneous memory system calculated as bellow:

$$E_{dynamic} = \sum_{i=0}^{C_X-1} \sum_{j=0}^{C_Y-1} \sum_{p=1}^{P} (\sum_{k=1}^{M_{sr}} SRC_{k,i,j,l} \times (FREQ_{p,k,r} \times E_{read_{sr}} + FREQ_{p,k,w} \times E_{write_{sr}}) + \sum_{k=1}^{M_{st}} STC_{k,i,j,2} \times (FREQ_{p,k,r} \times E_{read_{st}} + FREQ_{p,k,w} \times E_{write_{st}})), \quad l = 2 \qquad (10)$$

Consequently, our objective function can be expressed as:

$$minimize \quad E_{Total} = (E_{static_{sr}} + E_{dynamic_{sr}}) + \varphi. (E_{static_{st}} + E_{dynamic_{st}}) \ \ (11)$$

To summarize, objective function E_{Total} is minimized under constraints (1) through (10).

In Equation (11), the overall energy cost is divided to two distinct cost functions related to SRAM and STT-RAM memories as it is shown in Figure 5. φ is used as a knob for choosing SRAM versus STT-RAM bank in the memory layer.

In this model, φ is a coefficient which can change impact of STT-RAM or SRAM costs in the overall energy function, E_{Total}. As the target of optimization function is minimizing the overall cost, $\varphi > 1$ results SRAM intensive design, and $\varphi < 1$ results STT-RAM intensive design. Therefore, a designer can tune the model using φ parameter to design a hybrid memory layer with dominant SRAM or STT-RAM banks.

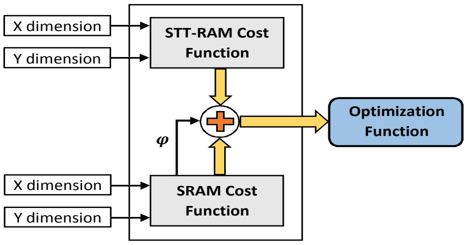

Fig. 5. Structure of the optimization function.

This proposed memory system is very flexible. For example in the proposed architecture, we can use other types of NVM technologies such as PCM instead of STT-RAM and DRAM instead of SRAM banks in the memory layer.

V. Evaluation

We use GEM5 [18], McPAT [19] and a SystemC-based NoC simulator, 3D-Noxim [20], to setup the system platform. The detailed of baseline system configurations used in this evaluation is listed in Table III. The cache capacities and energy consumption of SRAM and STT-RAM are estimated from CACTI [21] and NVSIM [22], respectively. The simulation platform for evaluation of our proposed architecture in this work is illustrated in Figure 6.

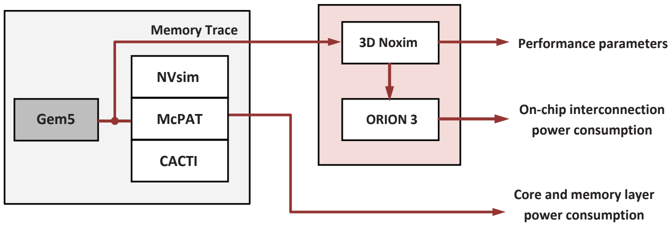
Memory Trace
Gem5
NVsim
McPAT
CACTI
3D Noxim
ORION 3
Performance parameters
On-chip interconnection power consumption
Core and memory layer power consumption

Fig. 6. Simulation infrastructure used in this work.

The parameters we used in our experiments for SRAM and STT-RAM cache banks are shown in Table II. We use multithreaded workloads for performing our experiments. The multithreaded applications with small working sets are selected from the PARSEC benchmark suite [23]. This selected benchmark suit consists of emerging workloads suitable for next generation shared-memory programs for CMPs. For experimental evaluation, P_{budget} and T_{max} are considered $100W$ and 80℃ , respectively.

Figure 7 shows the results of normalized energy efficiency for each PARSEC application, where energy efficiency is energy-delay product (EDP). As shown in this figure, the proposed hybrid stacked memory architecture improves EDP by about 44.8% on average compared with the Baseline memory design.

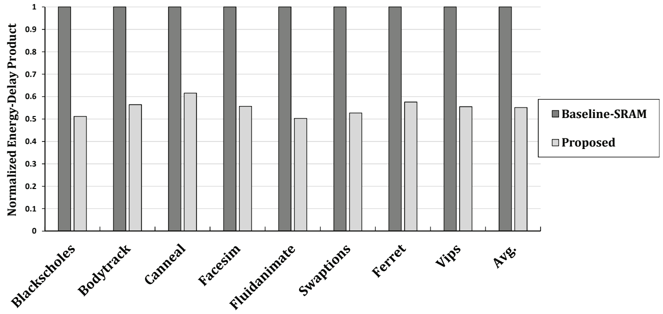

Fig. 7. Energy Delay Product (EDP) of each PARSEC application normalized with respect to the Baseline.

TABLE II. Different Memory Technologies Comparison at 32nm

Technology	Area	Read Latency	Write Latency	Leakage Power at 80 ℃	Read Energy	Write Energy
1MB SRAM	$3.03mm^2$	0.702ns	0.702ns	444.6mW	0.168nJ	0.168nJ
4MB STT-RAM	$3.39mm^2$	0.880ns	10.67ns	190.5mW	0.278nJ	0.765nJ

TABLE III. Specification of the Embedded eCMP Configuration

Component	Description
Number of Cores	16
Core Configuration	Single issue in-order Alpha21164, 3GHz, area 3.5mm^2, 32nm
Private Cache per each Core	SRAM, 4 way, 32B line, size 32KB per core
On-chip Memory	Baseline-SRAM: 16MB (1MB SRAM banks on each core) Baseline-STTRAM: 64MB (4MB STT-RAM banks on each core)
Network Router	2-stage wormhole switched, virtual channel flow control, 2 VCs per port, a buffer with depth of 4 flits per each VC, 5 flits buffer depth, 8 flits per Data Packet, 1 flit per address packet, each flit is set to be 16-byte long

Figure 8 compares the normalized performance results. As shown in this figure, the proposed design improves performance by about 13.8% on average compared with the Baseline memory design.

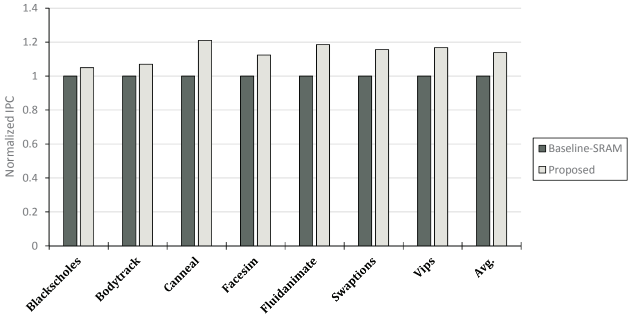

Fig. 8. Normalized performance comparison of each PARSEC application with respect to the Baseline.

Figure 9 shows life time of our novel 3D hybrid memory architecture for each benchmark with respect to baseline which is a memory architecture with only STT-RAM banks. As shown in this figure, life time of our proposed heterogeneous memory architecture is higher than the baseline for all benchmarks. In the other word, our hybrid memory design yields a 9.8× on average up to 24× increase in life time in comparison with baseline memory design. Therefore, our hybrid memory structure results more reliable 3D eCMP and this is because of our novel endurance model for NVM technology in optimization problem.

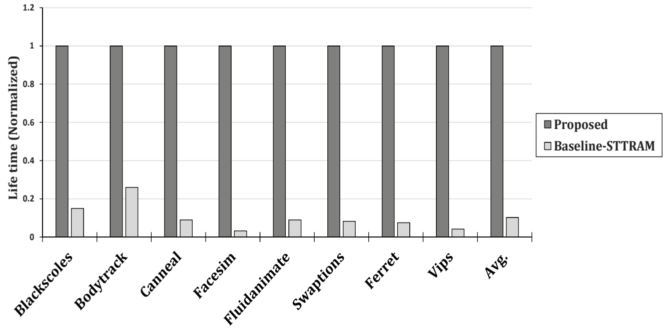

Fig. 9. Life time of our proposed 3D memory architecture for each PARSEC application normalized with respect to the Baseline.

VI. Conclusion

In this work, we proposed a convex optimization based model to design a heterogeneous memory system with using SRAM and STT-RAM memory banks in order to minimize energy consumption of 3D CMP. We propose an endurance model for NVM memory in our optimization problem to design a reliable hybrid memory structure. Our experimental results show that the proposed method improves energy-delay product (EDP) by 44.8% on average compared with the traditional memory design where single technology is used. Furthermore, our 3D eCMP yields on average 13.8% performance improvement in system performance compared with the baseline design.

References

[1] J. Kao, S. Narendra and A. Chandrakasan, "Subthreshold leakage modeling and reduction techniques," In the 2002 IEEE/ACM international conference on Computer-aided design (ICCAD), pp. 141–148, 2002.

[2] A. K. Mishra, T. Austin, X. Dong, G. Sun, Y. Xie, N. Vijaykrishnan and C. R. Das, "Architecting on-chip interconnects for stacked 3D STT-RAM caches in CMPs," In Proc. ISCA, pp. 69–80, 2011.

[3] X. Guo, E. Ipek and T. Soyata, "Resistive computation: avoiding the power wall with low-leakage, STT-MRAM based computing," In Proc. ISCA, pp. 371-382, 2010.

[4] W. Wang and P. Mishra, "System-wide leakage-aware energy minimization using dynamic voltage scaling and cache reconfiguration in multitasking systems," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 20, pp. 902 – 910, 2012.

[5] Z. Wang, Z, Gu, M. Yao and Z. Shao, "Endurance-Aware Allocation of Data Variables on NVM-Based Scratchpad Memory in Real-Time Embedded Systems," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2015.

[6] X. Luo, D. Liu, K. Zhong, D. Zhang, Y. Lin, J. Dai and W. Liu, "Enhancing Lifetime of NVM based Main Memory with Bit Shifting and Flipping," Embedded and Real-Time Computing Systems and Applications (RTCSA), 2014.

[7] J. Meng, and A. K.Coskun, "Analysis and runtime management of 3D systems with stacked DRAM for boosting energy efficiency," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012.

[8] Z. Wang, D. A. Jimenez, C. Xu and G. Sun and Y. Xie, "Adaptive Placement and Migration Policy for an STT-RAM-Based Hybrid Cache," In High Performance Computer Architecture (HPCA), pp. 13-24, 2014.

[9] A. Valero, J. Sahuquillo, S. Petit, P. Lopez, and J. Duato. "Design of Hybrid Second-Level Caches," IEEE Transaction on Computers, vol. 64, no. 7, 2015.

[10] M. Qureshi, M. Franceschini, L. A. Lastras-Monta˜no and J. Karidis, "Morphable Memory System: A Robust Architecture for Exploiting Multi-Level Phase Change Memories," In Proc. ISCA, pp. 153–162, 2010.

[11] H. Hajimiri, P. Mishra, S. Bhunia, B. Long, Y. Li and R. Jha,"Content-aware encoding for improving energy efficiency in multi-level cell resistive random access memory," In IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), pp. 76-81, 2013.

[12] C. Fu, M. Zhao, C. J. Xue and Alex Orailoglu. "Sleep-aware variable partitioning for energy-efficient hybrid PRAM and DRAM main memory," In Proceedings of the international symposium on Low power electronics and design, pp. 75-80, 2014.

[13] J. Hu, M. Xie, C. Pan, C. J. Xue, Q. Zhuge and E. H. Sha. "Low overhead software wear leveling for hybrid pcm + dram main memory on embedded systems," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 23, pp. 654 – 663, 2015.

[14] H. Tajik, H. Homayoun, and N. Dutt, "VAWOM: Temperature and process variation aware wearout management in 3D multicore architecture," In Proc. DAC, pp. 1–8, 2013.

[15] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," In: Proc. ISCA, pp. 365–376. 2011.

[16] B. Raghunathan, Y. Turakhia, S. Garg, and D. Marculescu, "Cherry-picking: Exploiting process variations in dark-silicon homogeneous chip multi-processors," In: Proc. DATE, pp. 39–44, 2013.

[17] M. Grant, S. Boyd and Y. Ye, "CVX: Matlab software for disciplined convex programming," Available at www.stanford.edu/ boyd/cvx/.

[18] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness et al. "The gem5 simulator," ACM SIGARCH Computer Architecture News 39, vol. 39, no. 2, May 2011.

[19] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," In Annual IEEE/ACM International Symposium on MICRO-42, pp. 469-480, 2009.

[20] M. Palesi, S. Kumar and D. Patti, "Noxim: Network-on-chip simulator," http://noxim.sourceforge.net, 2010.

[21] N. Muralimanohar, R. Balasubramonian and N. P. Jouppi, "CACTI 6.0: A tool to model large caches," HP Laboratories, Technical Report, 2009.

[22] X. Dong, C. Xu, N. Jouppi, and Y. Xie, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Non-volatile Memory," In Emerging Memory Technologies Springer, pp. 15-50, New York, 2014.

[23] M. Gebhart, Gebhart, Mark, Joel Hestness, Ehsan Fatehi, Paul Gratz, and Stephen W. Keckler. "Running PARSEC 2.1 on M5." University of Texas at Austin, Department of Computer Science, Technical Report, 2009.

[24] L. Wilson, "International Technology Roadmap for Semiconductors (ITRS)."

[25] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie, "Hybrid Cache Architecture with Disparate Memory Technologies," In Proc. ISCA, pp. 34-45, 2009.