OptMem: Dark-Silicon Aware Low Latency Hybrid Memory Design

†Salman Onsori, ‡Arghavan Asad
†‡Computer Engineering Department
†Bilkent University
Ankara, Turkey
salmanonsori@bilkent.edu.tr, ar_asad@comp.iust.ac.ir

*KaamranRaahemifar,‡Mahmood Fathy
‡Iran University of Science and Technology, Tehran, Iran
*Electrical and Computer Engineering Department
*Ryerson University, Ontario, Canada
kraahemi@ee.ryerson.ca, mahfathy@iust.ac.ir

***Abstract*—In this article, we present a convex optimization model to design a three dimension (3D)stacked hybrid memory system to improve performance in the dark silicon era. Our convex model optimizes numbers and placement of static random access memory (SRAM) and spin-transfer torque magnetic random-access memory(STT-RAM) memories on the memory layer to exploit advantages of both technologies. Power consumption that is the main challenge in the dark silicon era is represented as a main constraint in this work and it is satisfied by the detailed optimization model in order to design a dark silicon aware 3D Chip-Multiprocessor (CMP). Experimental results show that the proposed architecture improves the energy consumption and performanceof the 3D CMPabout 25.8% and 12.9% on averagecompared to the Baseline memory design.**

Keywords—*Dark silicon, Non-Volatile Memory (NVM), Hybrid memory architecture, Embedded Chip-Multiprocessor (eCMP), Convex optimization, uncore components, 3D integration.*

I. INTRODUCTION

In nowadays multicore architectures, energy efficiency becomes the primary concern during system design. Especially, energy consumption is a primary constraint in embedded system design since many of them are generally limited by battery lifetime. Main memory and cache subsystemsconsume a significant portion of overall energy in memory-intensive embedded applications. Due to the exponential contribution of leakage power in total power consumption in nanoscale era, leakage power can be a major driver of dark silicon in future multicore systems. However, leakage power also constitutes a major fraction of power consumption of memory modules [1]. Consequently, architecting new classes of memory systems with the minimum leakage power is essential for embedded systems.

One of the newest challenges in multicore design is the management ofdark silicon [2-4]. The rise of utilization wall due to thermal and power budgets restricts active components and results in a large region of dark silicon. Among the on-chip components, the cores anduncore components consume most of the power.Uncore components such as memory and on-chip network play a significant role in consuming large portion of power. Power management of these uncore components can be critical to maximize design performance in the dark silicon era. Thus, in addition to the embedded system requirements, dark silicon constraint forces designers to reduce energy consumption.

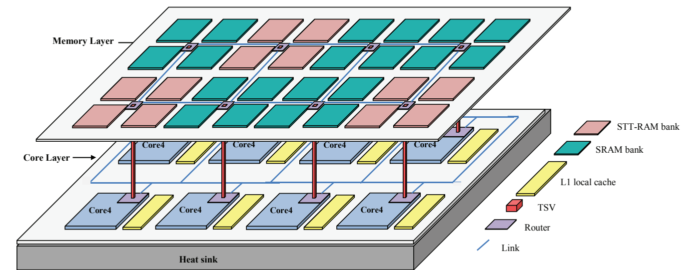

Fig. 1. 3D architecture of the proposed design

Spin Transfer Torque RAM (STT-RAM) as a promising candidate of non-volatile memories (NVMs) are considered to be attractive replacement for traditional SRAM memories due to their favorable characteristics such as highdensity, non-volatility and near-zero leakage power [5,6]. Nevertheless, they sufferfrom a longer write latency, limited write endurance and higher write energy consumption when compared to the traditional SRAM memory technology. In order to overcome the mentioned disadvantages of both memory technologies and benefit from their positive features, we use SRAM and STT-RAM as two different types of memory banks in the proposed memory architecture. This heterogeneous memory design is the best design possibility because it benefits from both memory technologies. Several research have also exploredmicro-architectural heterogeneity to combat the dark silicon problem [4,7].Fig. 1 shows an overview of the proposed architecture using an example of a 16 core homogeneous CMP in the lower layer and hybrid memory architecture in upper layer. In the proposed heterogeneous memory system, STT-RAM memory banks are incorporated with SRAM memory banks.

In this paper, we propose a convex optimization based approach for designing the heterogeneous memory system in order to maximize performance of the three dimensional(3D) CMP with respect to the peak power budget which is the main constraint in the dark silicon era. The proposed model maps applications/threads with more dependency and communication intensity closer to each other while at the same time it finds optimal distance of these applications/threads to each memory banks in order to reduce latency of the3D CMP design. More specifically, the proposed convex model optimally chooses efficient number and placement of SRAM and STT-RAM memory banks on the memory layer, and maps applications/threads on cores in the core layer.

978-1-5090-0033-3/16/$31.00 ©2016 IEEE

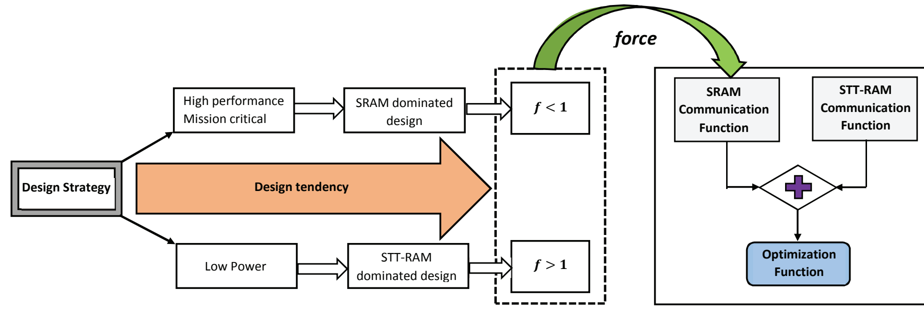

Fig. 2. Impact of force coefficient on the design strategy.

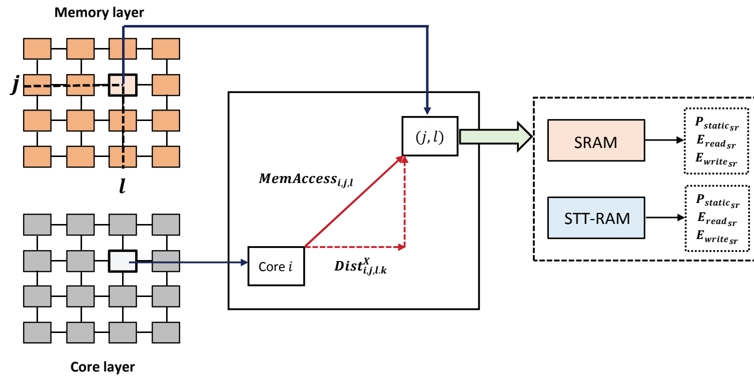

Fig. 3. Define parameters of a sample core and a memory bank and the communication between them.

The rest of this paper is organized as follows. Section II describes the optimization problem and its formulation. In Section III, excremental evaluation is presented. Finally, Section IV concludes the paper.

II. Optimization Problem and Formulation

In this section, we propose a convex optimization model for achieving the following goals: 1) satisfying the dark silicon constraint with exploiting emerging technologies like NVM memories and 3D integration technology, 2) designing a hybrid memory architecture in order to use maximum advantages of SRAM and STT-RAM memories 3) efficient arrangement of memory banks with different technologies in the memory layer to decrease communication overhead on the CMP.

In order to achieve these goals, the proposed objective function is presented as follow:

$$Minimum \quad Overal_{com} = f.CommLatency_{SRAM} + CommLatency_{STTRAM} \quad (1)$$

In Equation (1), the overall communication is comprised of SRAM and STT-RAM communication functions. f is called a *force*coefficientthat is used to show priority of memory layer to be designed by SRAM or STT-RAM banks. In other words, f is a factor which can change impact of STT-RAM or SRAM communication cost in the overall cost function,$Overal_{comm}$. For example, if the goal is designing a mission critical embedded system and it should be highly reliable, SRAM banks are chosen with the cost of additional energy consumption but it is acceptable due to functionality of the device. On the other hand, if the reliability is not that much important or our goal is simply reduce the power consumption, force is changed to design a memory with higher possibility for choosing STT-RAM banks. However, there should be a tradeoff for these decisions since we need to satisfy power constraint of the desirable design.These design strategies and their impact on the optimization function is shown in Fig. 2.Also, it should be mentioned that each of SRAM and STT-RAM communication costs are combination of costs in x and y dimensions.These functions are introduced in the following paragraphs.

$CommLatency^{X}_{SRAM}$is the latency of accessing to SRAM banks by cores in dimension x:

$$CommLatency^{X}_{SRAM} = \sum_{i=1}^{C} \sum_{j=0}^{D_X-1} \sum_{l=0}^{D_Y-1} \sum_{k=1}^{D_X-1} MB^{SR}_{j,l} \times (MemAccess_{i,j,l,r} \times Dist^{X}_{i,j,l.k} \times k + MemAccess_{i,j,l,w} \times Dist^{X}_{i,j,l.k} \times k) \quad (2)$$

In Equation (2), j and l are x and y dimension of a memory bank in the second layer. $C$is number of cores and dimension of the chips for x and y coordinates are D_X and D_Y, respectively. $Dist^X_{i,j,l.k}$is a binary variable and is set to 1 if the distance between cores i and a memory bank in position (j,l) is equal to k for x-dimension. $MemAccess_{i,j,l,r}$is number of read accesses of core i to a SRAM bank in position (j,l). Also, $MemAccess_{i,j,l,w}$ is number of write accesses of core i to SRAM bankin position (j,l). In addition, Fig. 3 demonstrates two dimensional communications of a sample core and a memory bank. Similarly, $CommLatency^Y_{SRAM}$ is defined like $CommLatency^X_{SRAM}$just for dimension y.

$$CommLatency^Y_{SRAM} = \sum_{i=1}^{C}\sum_{j=0}^{D_X-1}\sum_{l=0}^{D_Y-1}\sum_{k=1}^{D_X-1} MB^{SR}_{j,l} \times (MemAccess_{i,j,l,r} \times Dist^Y_{i,j,l.k} \times k + MemAccess_{i,j,l,w} \times Dist^Y_{i,j,l.k} \times k) \quad (3)$$

Equation (2) is comprised of four summations. It finds different commination costs between cores and SRAM banks based on possible distances between cores and SRAM banks in x coordinate. With minimizing objective function $Overal_{com}$, we also will achieve minimum distances (k) in $CommLatency^X_{SRAM}$. This procedure again is done for $CommLatency^Y_{SRAM}$ in the y coordinate. Therefore, we can have the best placements for SRAM banks to have minimum communication latencies between cores and these banks.

$CommLatency^X_{STTRAM}$ is another cost function that is the communication cost for accessing to STT-RAM banks by cores in dimension x. More specifically,

$$CommLatency^X_{STTRAM} = \sum_{i=1}^{C}\sum_{j=0}^{D_X-1}\sum_{l=0}^{D_Y-1}\sum_{k=1}^{D_X-1} MB^{ST}_{j,l} \times (MemAccess_{i,j,l,r} \times R_{Cost} \times Dist^X_{i,j,l.k} \times k + MemAccess_{i,j,l,w} \times W_{Cost} \times Dist^X_{i,j,l.k} \times k) \quad (4)$$

Equation (4) models communication costs of cores with STT-RAM banks. Conceptually, this equation is similar to Equation (2) but it has additional parameters to model STT-RAM memory instead of SRAM. Two parameters namely R_{Cost} and W_{Cost} are defined in this part which are the cost of reading and writing to a STT-RAM normalized with SRAM memory. $CommLatency^Y_{STTRAM}$ are defined assame as $CommLatency^X_{STTRAM}$for dimension y.

$$CommLatency^Y_{STTRAM} = \sum_{i=1}^{C}\sum_{j=0}^{D_X-1}\sum_{l=0}^{D_Y-1}\sum_{k=1}^{D_X-1} MB^{ST}_{j,l} \times (MemAccess_{i,j,l,r} \times R_{Cost} \times Dist^Y_{i,j,l.k} \times k + MemAccess_{i,j,l,w} \times W_{Cost} \times Dist^Y_{i,j,l.k} \times k) \quad (5)$$

Note that, as there are only two layers in this work, communication cost in z dimension is assumed to be same for all cores and banks; thereforewe do not consider it in the model.

In order to design the memory layer with SRAM and STT-RAM banks, there should be some constraints to architect the system in order to achieve the optimal placement and positions of SRAM and STT-RAM banks. These constraints are defined at Equation (6) and (7).

$$MB^{ST}_{i,j} + MB^{SR}_{i,j} = 1 \quad \forall\, i,j \quad (6)$$

$$\sum_{i=0}^{D_X-1}\sum_{j=0}^{D_Y-1} (MB^{ST}_{i,j} + MB^{SR}_{i,j}) = C \quad (7)$$

$MB^{ST}_{i,j}$ is a binary variable and is set to 1 if the existing memory bank in (i,j) is a STT-RAM bank. Similarly, $MB^{SR}_{i,j}$ is a binary variable and is set to 1 if the existing memory bank in (i,j) dimension is a SRAM banks. Equation (6) allows only assignment of one SRAM or STT-RAM bank to a single coordinate. In addition, sum of used STT-RAM and SRAM banks in second layer is equal to C that is defined in Equation (7).

The total power consumption of the proposed memory architecture during run time period of the mapped workload must be less than the available power budget. More specifically,

$$P_{Total} = (Power_s + Power_d) \leq P_{budget} \quad (8)$$

Equation (8) is the dark silicon constraint for the proposed memory architecture.Power consumption is the main constraint of the dark silicon era and uncore components such as on-chip memories are responsible for significant amount of power consumption [1]. On the other hand, satisfying power budget,P_{budget}, which in the dark silicon erais well-known to Thermal Design Power budget (TDP),is a main factor of the proposed model. Therefore, the achieved memory architecture based on the proposed model mitigatesthe dark silicon challenge by reducing power of the memory system as one of the most important uncore components. Focusing on uncore components architectures as a solution to combat dark silicon is unexplored in these days [1].

Since this optimization approach is solved at design time and static power dissipation depends on temperature, we consider pessimistic worst-casescenario and calculate P^{SR}_{Static} and P^{ST}_{Static} at maximum temperature limit.

$$Power_s = \sum_{i=0}^{D_X-1}\sum_{j=0}^{D_Y-1}(MB_{i,j}^{SR} \times P_{Static}^{SR} + MB_{i,j}^{ST} \times P_{Static}^{ST}) \quad (9)$$

Equation (9)finds static power of the hybrid memory by summing static power consumption of each SRAM and STT-RAM bank.

In Equation (10), P_{read}^{SR}, P_{write}^{SR}, P_{read}^{ST} and P_{write}^{ST} indicate average dynamic power consumed by the SRAM and STT-RAM banks per read and write access, respectively. $P_{dynamic}$ as the dynamic power consumption of the proposed hybrid memory system is calculated as bellow:

$$Power_d = \sum_{i=0}^{D_X-1}\sum_{j=0}^{D_Y-1}\sum_{c=1}^{C}(MB_{i,j}^{SR} \times (MemAccess_{i,j,c,r} \times P_{read}^{SR} + MemAccess_{i,j,c,w} \times P_{write}^{SR}) + MB_{i,j}^{ST} \times (MemAccess_{i,j,c,r} \times P_{read}^{ST} + MemAccess_{i,j,c,w} \times P_{write}^{ST})) \quad (10)$$

To summarize, objective function $Overal_{com}$ is minimized under constraints (2) through (10). We only mentioned main constraints and their related variables in this section for brevity.

III. Exprimental Evaluation

A. Experimental Setup

We use GEM5[8] as a full system simulator to implement memories and cores. To simulate accurate behaviour of the 3D CMP design and its NoC architecture, we integrated GEM5 with a NoC simulator[9]. In addition, to calculate power consumption of the design, mention platform is integrated with McPAT [10]. The cache capacities and energy consumptions of SRAM and STT-RAM are estimated from CACTI [11] and NVSIM [12], respectively. The simulation platform of the work is shown in Fig. 4. Also, detailsof the memory parameters and the baseline system configurationwhichwe used in our experiments for SRAM and STT-RAM banks are shown in Table I and Table II, respectively.

We use multithreaded workloads for performing our experiments. The multithreaded applications with small working sets are selected from the PARSEC benchmark suit [13]. In our setup, programs in a given workload are randomly mapped to cores to avoid a specific OS policy. For the experimental evaluation, P_{budget} and T_{max} are considered $100W$ and 80 ℃, respectively. Furthermore, we use CVX [14]to model the proposed convex optimization problem and solve it.

TABLE I. Different memory technologies comparison at32nm

Technology	Area	ReadLatency	WriteLatency	LeakagePower at80 ℃	ReadEnergy	WriteEnergy
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

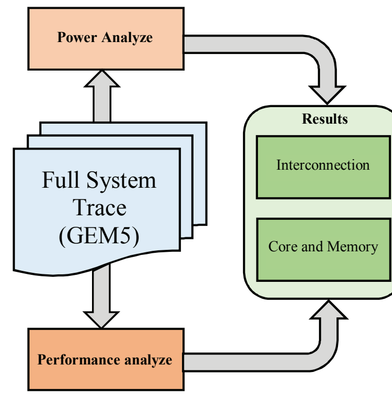

Fig. 4. Simulation platform of the work.

TABLE II. Specification of the baseline embedded CMP configuration

Component	Description
Number of Cores	16, 4× 4 mesh
Core Configuration	Alpha21164, 3GHz, area 3.5mm^2, 32nm
Private Cache per each Core	SRAM, 4 way, 32B line, size 32KB per core
On-chip Memory	Hybrid-fix: 8MB SRAM (8 banks , each 1MB) and 32MB STT-RAM (8 banks , each 4MB)
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

B. Experimental Result

In this section, we evaluate our proposed 3D CMP with stacked memory in two different cases: 1) the CMP with hybrid stacked memory with same number of SRAM and STT-RAM banks in which STT-RAM banks are on the left and SRAM banks are on the right part of the memory layer (Hybrid-fix), 2) CMP with the proposed hybrid stacked memory on the core layer.

Fig.5 shows the results of the normalized energy consumption of the proposed method with respect to Hybrid-fix. As shown in this figure, the proposed design reduces energy consumption by about 25.8% on average compared to Hybrid-fix design.

Fig. 6 compares the normalized performance results. As shown in this figure, the proposed design improves IPC as a best parameter shows performance of the system up to 12.9% (4.87% on average) compared to the Hybrid-fix baseline design.

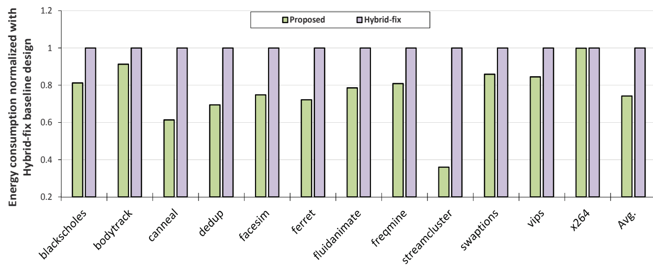

Fig. 5. Normalized energy consumption of the proposed design with respect to Hybrid-fix.

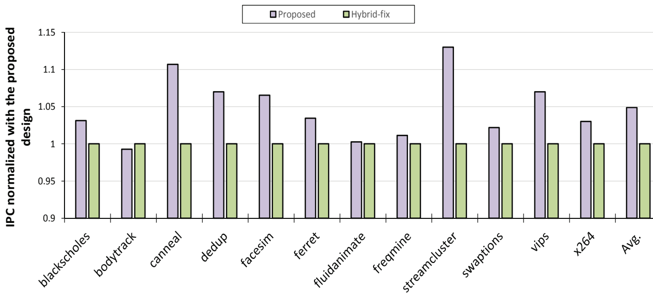

Fig. 6. Normalized performance comparison of each application with respect to the proposed design.

IV. CONCLUSION

In this work, we proposed a model to design an optimal heterogeneous memory system with using SRAM and STT-RAM memory banks. Our proposed optimization model finds optimal number and placement of different memory banks to satisfy peak power budget which is the main challenge in the dark silicon era. Experimental results show that the proposed architecture improves the energy consumption and performance of the 3D CMP on average about 25.8% and 12.9% respectively, compared to the Baseline memory design.

REFERENCES

[1] H. Cheng, et al. "Core vs. Uncore: The Heart of Darkness," Design Automation Conference(DAC), USA, 2015.

[2] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger. "Dark silicon and the end of multicore scaling," In Computer Architecture (ISCA), pp. 365-376, 2011

[3] J. Henkel, H. Khdr, S. Pagani, and M. Shafique. "New trends in dark silicon." In Design Automation Conference (DAC), pp. 1-6, 2015.

[4] J. Allred et al. “Designing for dark silicon: a methodological perspective on energy efficient systems,” In ISLPED, 2012.

[5] 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,” ISCA, pp. 69–80, 2011.

[6] J. Ahn, S. Yoo, and K. Choi, “Dasca: Dead write prediction assisted stt-ram cache architecture,” International Symposium on High Performance Computer Architecture (HPCA), 2014.

[7] Y. Turakhia et al. “Hades: Architectural synthesis for heterogeneous dark silicon chip multi-processors,” In DAC, 2013.

[8] 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.

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

[10] 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.

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

[12] 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, 2012.

[13] 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.

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