A High-performance Hybrid Memory Architecture for Embedded CMPs Using A Convex optimization model

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

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

Abstract—In this article, we present a convex optimization model to design a stacked hybrid memory system for 3D embedded chip-multiprocessors (eCMP). Our convex model optimizes numbers and placement of SRAM and STT-RAM memories on the memory layer, and maps applications/threads on cores in the core layer effectively. The detailed proposed model satisfies the power constraint which is the main challenge of dark-silicon era. Experimental results show that the proposed architecture considerably improves the energy-delay product (EDP) and performance of the 3D eCMP compared to the Baseline memory design.

Keywords— Non-Volatile Memory (NVM), Hybrid memory Architecture, Embedded Chip-multiprocessor (eCMP), Convex-optimization.

I. INTRODUCTION

The increase in the number of cores in embedded CMPs comes with an increase in power consumption. Power consumption is a primary constraint in embedded system designs since many of them are generally limited by battery lifetime. Main memory and cache hierarchy can consume a significant portion of the overall energy in memory-intensive embedded applications [1]. On the other hand, leakage power also constitutes a major fraction of power consumption of memory modules. Consequently, architecting new classes of memory systems with the minimum leakage power is essential for embedded systems.

STT-RAM is considered as an attractive replacement for traditional SRAM memories due to its ultra-low leakage power and higher capacity. However, STT-RAM suffers from 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 aforementioned disadvantages of both memory technologies and benefit from their positive features, we need to exploit SRAM and STT-RAM as two different types of memory banks in the memory architecture. This heterogeneous point of view leads us to the best design that benefits from advantages of both memory technologies. In this work, we exploit non-volatile memories (NVMs) and 3D CMP in order to design a dark-silicon-aware CMP. In this wok, we use non uniform memory architecture (NUMA) stacked directly on top of the core layer in a 3D CMP.

In this work, we propose a convex optimization based approach for designing a heterogeneous memory system in order to maximize the performance of the 3D CMP with respect to the peak power budget which is a main constraint in dark silicon era. The proposed convex model chooses efficient numbers and placement of SRAM and STT-RAM memories on the memory layer, and effectively maps applications/threads on cores in the core layer. In the proposed heterogeneous memory system, STT-RAM is incorporated with SRAM banks in the second layer (Figure 1). The rest of this paper is organized as follows. Section II describes the convex optimization model. Evaluation results are presented in Section III, and the paper is concluded in Section IV.

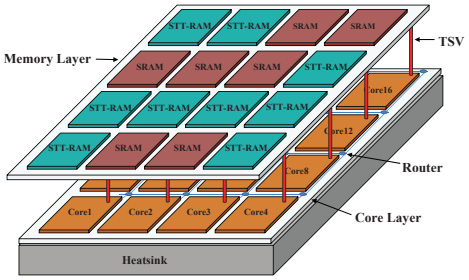

Fig. 1. 3D eCMP where hybrid memory system is stacked onto the core layer

II. OPTIMIZATION PROBLEM

In this section, we propose a convex optimization model with the following outputs: 1) optimal number of SRAM and STT-RAM memory banks based on the memory access behavior of mapped applications with respect to the peak power budget; 2) optimal placement of SRAM incorporated with STT-RAM banks in the memory layer; 3) optimal placement of cores by placing cores with more intense communication closer to each other in the core layer. To solve the models, we use CVX [2], an efficient convex optimization solver.

Our objective function finally achieve as follow:

$$\text{Minimum} \quad J = (X_{Cost-SR} + Y_{Cost-SR}) + \varphi.(X_{Cost-ST} + Y_{Cost-ST}) \qquad (1)$$

Where φ is used as a knob for choosing SRAM versus STT-RAM bank in each x and y coordinate in the memory layer. In this model φ is chosen by the designer to evaluate performance improvement versus energy reduction. In Equation (1), $X_{Cost-SR}$ is the communication cost for accessing to SRAM banks by cores in dimension x:

$$X_{Cost-SR} = \sum_{i=1}^{P}\sum_{l=1}^{P}\sum_{d=1}^{C_x-1}\sum_{j=1}^{M_{ST}}\sum_{k=1}^{C_x-1}((I_{i,l} \times PXdist_{i,l,d} \times d) \times (FREQ_{i,j,r} \times Xdist_{i,j,k} \times k + FREQ_{i,j,w} \times Xdist_{i,j,k} \times k)) \qquad (2)$$

In this equation, M_{sr} is the number of SRAM banks which we want to find its optimal value. C_x is the dimension of the chip in x cordinate. $I_{i,j}$ is communication intensity between cores i and j. $PXdist_{i,l,d}$ is a binary variable and is set to 1 if the distance between cores i and j in x-dimension is equal to d. In this equation, $FREQ_{i,j,r}$ is the number of read accesses of core i to SRAM bank j. Also, $FREQ_{i,j,w}$ is the number of write accesses of core i to SRAM bank j. Note that, these frequencies are known for us because our model is for embedded applications. $Xdist_{i,j,k}$ is a binary variable and is 1 when the distance between core i and memory bank j is equal to k. The three left summations in Equation (2) are for finding the overall cost of communications between the cores and the two final summations consider the distance and communication between the cores and memory banks. Note that, both costs are calculated simultaneously and multiplied with each other in order to find the final cost. Similarly, $Y_{Cost-SR}$ is defined like $X_{Cost-SR}$ for dimension x. Also, $X_{Cost-St}$ *and* $Y_{Cost-St}$ are defined like $X_{Cost-SR}$ for STT-RAM banks.

The total power consumption of the proposed stacked heterogeneous memory system during the running phase of the mapped workload must be less than the maximum power budget. In other words, Equation (3) is the dark silicon constraint for the proposed memory architecture.

$$P_{Total} = (P_{static} + P_{dynamic}) \leq P_{budget} \quad (3)$$

The static power dissipation depends on the 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.

$$P_{static} = \sum_{i=0}^{C_X-1}\sum_{j=0}^{C_Y-1}\left(\sum_{k=1}^{M_{sr}} MC_{k,i,j,l} \times P_{static_{sr}} + \sum_{k=1}^{M_{st}} MC_{k,i,j,l} \times P_{static_{st}}\right), \quad l = 2 \quad (4)$$

In Equation (4), $MC_{k,i,j,l}$ indicates whether a SRAM or STT-RAM bank is in (i,j) in layer l which here is equal to 2. This equation finds the static power of hybrid memory by summing static power consumption of each SRAM and STT-RAM bank.

In Equation (5), $P_{read_{sr}}$,$P_{write_{sr}}$, $P_{read_{st}}$and $P_{write_{st}}$indicate the average dynamic power consumed by the SRAM and STT-RAM banks per read and write access, respectively. $P_{dynamic}$ is the dynamic power consumption of the proposed hybrid memory system and is calculated as:

$$P_{dynamic} = \sum_{i=0}^{C_X-1}\sum_{j=0}^{C_Y-1}\sum_{p=1}^{P}\left(\sum_{k=1}^{M_{sr}} MC_{k,i,j,l} \times \left(FREQ_{p,k,r} \times P_{read_{sr}} + FREQ_{p,k,w} \times P_{write_{sr}}\right) + \sum_{k=1}^{M_{st}} MC_{k,i,j,2} \times \left(FREQ_{p,k,r} \times P_{read_{st}} + FREQ_{p,k,w} \times P_{write_{st}}\right)\right), \quad l = 2 \quad (5)$$

Also, sum of STT-RAM and SRAM banks used in the second layer equals to P as follows:

$$\sum_{x=0}^{C_X-1}\sum_{y=0}^{C_Y-1}\left(\sum_{i=1}^{M_{sr}} MMAP_{i,x,y,l} + \sum_{i=1}^{M_{st}} MMAP_{i,x,y,l}\right) = P, \quad l = 2 \quad (6)$$

$MMAP_{i,x,y,l}$ is a binary variable which indicates when the coordinate (x,y) is assigned to SRAM bank in layer $l = 2$. Note that, our model finds the optimal number of SRAM and STT-RAM banks (M_{sr}, M_{st}).

III. Experimental Evaluation

In this section, we evaluate our proposed 3D eCMP with stacked memory in two different cases: the CMP with SRAM-only stacked memory on the core layer (Baseline), and the CMP with proposed hybrid stacked memory on the core layer. In the proposed method, we consider 16 SRAM banks (1MB each) and 16 STT-RAM banks (4MB each) as the maximum available memory which can be used for designing the hybrid memory architecture. In our setup, threads in a given application are randomly mapped to cores to avoid a specific Operating System (OS) policy. For experimental evaluation, P_{budget} and T_{max} are considered $100W$ and 80℃, respectively.

Figure 2 shows the results of normalized energy efficiency, where energy efficiency is energy-delay product (EDP). As shown in this figure, in comparison with the Baseline design, the proposed design reduces EDP by about 45.3% on average.

Figure 3 compares the normalized performance results. As shown in this figure, the proposed design improves performance up to 16% (9.2% on average) compared with the Baseline design.

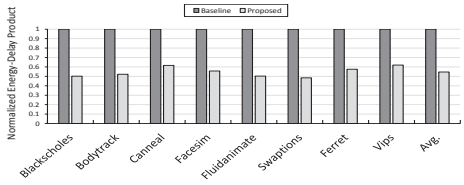

Fig. 2. Normalized energy delay product (EDP) comparison of each application with respect to the Baseline.

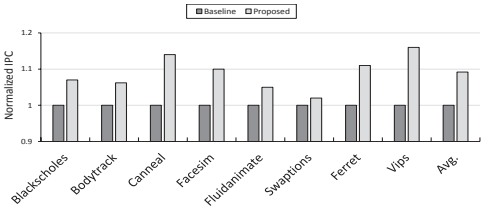

Fig. 3. Normalized performance comparison of each application with respect to the Baseline.

IV. Conclusion

In this work, we proposed a model to design an optimal heterogeneous memory system using SRAM and STT-RAM memory banks. Our proposed convex optimization-based model finds the optimal number and placement of different memory banks to satisfy peak power budget. We maximized the performance of CMP design considering communication intensity of cores in our model. Experimental results show that compared with the traditional memory designs which use a single technology, the proposed method improves energy-delay product (EDP) by 45.3% on average.

References

[1] H. Cheng, et al. "Core vs. Uncore: The Heart of Darkness," Design Automation Conference(DAC '15), USA, 2015.

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

[3] 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.