A Low-Power Hybrid Non-Volatile Cache with Asymmetric Coding

Omid Hajihassani, Armin Ahmadzadeh, Mohsen Gavahi, Mohammadreza Raei, Dara Rahmati, Saeid Gorgin*

School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
HPC@ipm.ir

***Abstract*— Cache memories such as magnetic ram or phase change memory came a long way in term of their architecture from their earlier models and have marked differences in power, performance, access latency, and dynamic/static energy consumption. In our work, we propose a hybrid cache design that exploits the characteristics of the employed cache technologies to achieve better power and area efficiency alongside the asymmetric coding that increases the ratio of 0s to 1s in the cache data by adding an order of information redundancy to the cache's original data. We benefit from a hybrid cache memory architecture that utilizes the positive aspects of STT-RAM and SRAM technologies to propose a solution that is more energy efficient compared to conventional cache architectures. By the evaluation of programs' cache data from Splash-2 and Parsec suits, it is indicated that alone by the hybrid architecture the total static and dynamic power consumption has dropped by 55% compared to the SRAM and DRAM caches and the area has reduced by 45%. With the aid of the proposed coding scheme, the number of set operations issued to cache has decreased by 47%. This reduces the write power of programs by 24%, leading to an overall 14% reduction in the programs' total static and dynamic power consumption.**

Keywords—Cache; Power and energy efficiency; Asymmetric coding scheme; Hybrid cache architecture; STT-RAM; PRAM

I. INTRODUCTION

With the increase in the number of utilized cores in CPUs and better support for multi-threaded applications, the need for caches that are shared between the existing cores and processes becomes more highlighted [1]. The emergence of new technologies causes multiple changes in the cache architecture. For example, three-dimensional integrated circuit design has found its way in the design of new cache architectures, having multiple layers of different types of memories build on top of one another, taking up less area and scoring better in various performance metrics [2].

As mentioned, the emerging cache architectures need to satisfy the demands of sophisticated and resource demanding multi-threaded algorithms and applications [3] that ask for bigger and faster caches [1]. Hence, designers need to bear in mind many different challenges and factors when coming up with new architectures [4]. Moreover, many devices, nowadays, are embedded systems running on batteries. It is of utmost importance to reduce the active and passive energy consumption of the systems and inherently the cache devices in upcoming generations [5].

We propose and evaluate a hybrid scheme of cache architecture using STT-RAM and PRAM to suggest a power efficient cache device with better performance and reduced cache area compared to conventional caches. The proposed architecture utilizes SRAM as the L1 cache and STT-RAM or PRAM as L2 /L3 caches. We have evaluated these architectures by NVSim [6], which is a modeling tool for non-volatile memory power, area, and performance. We also used CACTI [7], which is a cache performance and power model. Our results prove the dynamic and static power efficiency of the proposed hybrid cache. However, due to the costly set/write operations of STT-RAM and PRAM, we propose an asymmetric coding scheme, that further reduces the number of set/write operations and the leakage power of cells in the cache to reach better power efficiency over previous cache solutions. The employed asymmetric coding scheme generates codewords that have characteristics proven to be beneficial to be written into the non-volatile cache cells. By encoding the user's original data, obtained by Sniper full system multi-core simulator [8], and adding a reasonable order of information redundancy, we have achieved codewords with a higher ratio of 0s to 1s in their data. Hence, we can further reduce the set energy for the data in cache cells.

The rest of this paper is organized as follows. In Section II, a thoroughly detailed specification of the terms used in this work is provided. The related works of the low-power cache are outlined in Section III. In Section IV, we propose the hybrid architecture and build upon it our coding solution that meets the discussed non-volatile cache shortcomings. In Section V, the evaluation results of our proposed techniques are discussed in depth, which indicate the superiority of the hybrid cache design and the asymmetric coding. Finally, Section VI concludes our proposal and discusses the future works.

II. BACKGROUND

Many different technologies have been proposed by manufacturers and scholars for non-volatile memories including Flash memory, Phase Change Memory (PCM) [9] and Spin-Torque Transfer RAM (STT-RAM) [10]. These cache technologies differ in the smallest factors such as feature size, density, write/read speed, dynamic power, leakage current, volatility, and scalability.

STT-RAM is a non-volatile memory that is a good substitute for conventional caches. This technology employs

* S. Gorgin is also affiliated with Iranian Research Organization for Science and Technology (IROST), Tehran, Iran.

978-1-5386-0804-3/17/$31.00 ©2017 IEEE

the Magnetic Tunnel Junction, which is composed of two magnets and an insulating layer, as the cell storage [11]. The cells' storage status can be differentiated by the magnetization direction between the ferromagnetic layers. If the two ferromagnetic layers have the same magnetization direction, the cell represents a "1" and vice versa. The cell is programmed by altering the relative magnetic direction of the layers, which is handled by programming one of the layers referred to as the free layer [12].

On the other hand, in volatile memories, after a predetermined time the information in the cache is not valid anymore. This fact demands the periodical intervals of state refreshment of the cache cells in DRAM. The static power of SRAM is much more than that of its non-volatile counterparts, MRAM and PRAM.

Despite the fact that the SRAM has the least cell density compared to MRAM and PRAM, it has the fastest write/read speed. The slowest amongst all is PRAM. Moreover, MRAM has a moderately fast read speed with a slow write [13]. Also, it has a low read dynamic power and high write dynamic power. However, PRAM has a moderate read dynamic power and high write power. Although, both these technologies have higher dynamic write power compared to their counterpart SRAM. Therefore, stored data retrieval for MRAM and PRAM consumes more power compared to SRAM [14].

III. Related Works

In order to decrease the power consumption of an on-chip cache memory a PCM based architecture has been proposed in [15]. The evaluations indicate this architecture reduces the leakage power of L1/L2 caches by 80%. In this paper, the researchers proposed a combination of SRAM and PRAM with the aim of reducing the power consumption. In [16], the authors constructed PDRAM with fusing together the DRAM and PRAM. This work benefits from the low access latency of the PRAM compared to DRAM and the lower program energy consumption of the DRAM compared to PRAM. The PDRAM masks the shortcomings of DRAM and PRAM memories by morphing them into one. This proposed organization of DRAM and PCM resolved the problem of the DRAM's low capacity and decreased the static power consumption of DRAM. Banakar, in [17], investigated the simple data management algorithms for scratch pad memories proposed as an alternative to the cache memories. In [18], it is concluded that NVM is far more efficient in terms of the power consumption when used in the scratch pad memories and proposes a hybrid non-volatile based scratch pad memory.

In [19], by designing new architectures for low-power and low-latency STT-RAM cells, which are optimized for read/write operations, a hybrid cache architecture composed of SRAM and these optimized STT-RAM caches were suggested, evaluated, and compared to an all SRAM conventional cache, which we have used as the basis of our evaluations. However, it should be mentioned that the optimized versions of STT-RAM have lower data retention rate and worsen the STT-RAM's leakage power by 50%. Also, in [20], a read/write aware hybrid architecture is proposed, where the L2 cache is an organization of read/write regions of STT-RAM/PRAM and SRAM caches. It is worth noting that the cache data migration policy imposes more operation time on the cells and increase the cache miss rate.

In our work, without putting any of the mentioned burdens on the cache, we suggest a hybrid cache design of SRAM, PRAM, and STT-RAM alongside a coding scheme that further reduces the number of set/programming operations in the cache, the overall write energy, leakage power, and area.

IV. Proposed Solutions

In this section, a hybrid cache architecture and a coding scheme are proposed to reduce the total power consumption of the cache unit. First, the hybrid cache design is outlined, then the asymmetric coding scheme is introduced. Also, a set of initial parameters for the coding is evaluated by the characteristics of the generated codewords. Finally, the circuit diagram for the coding scheme is described.

A. Cache Design

Our proposed hybrid cache is comprised of three distinct levels. L1's cache architecture is based on the SRAM cache technology. L2/L3 are both either STT-RAM or PRAM caches. We evaluate the power efficiency and the performance gain from both of these architectures and weigh them against a cache design where all three levels are comprised of SRAM caches [19]. Thus, we have two cache architectures wherein both of them L1 is based on SRAM cache and they just differ by the cache technology in their L2/L3 caches. We evaluate the hybrid cache design in our work for area, leakage power, and write power. Based on the aforesaid information concerning the STT-RAM and PRAM lower leakage power compared to SRAM, the hybrid cache schemes, introduced here, are far more efficient in static and dynamic power consumption. The evaluation results for these cache architectures will be thoroughly discussed in next sections.

We use these hybrid cache designs as the basis for our coding scheme. In our suggested work, based on the fact that in STT-RAM and PRAM, submitting a "1" to the cell is more power consuming than submitting a "0", we try to further reduce the write and the leakage power of the cells. We achieve this goal, by changing and favoring the statistics in the data words that make them suitable and more efficient to be written in non-volatile caches.

B. Coding Scheme

The employed coding scheme is intended to increase the ratio of 0s appearing in the generated codewords. Due to the fact that submitting a "0" into the non-volatile cache consumes less programming energy compared to the submitting of a "1", by reducing the number of set operations issued to the cache, the power consumption is reduced. Therefore, we propose a coding scheme that generates codewords having more number of bits valued as "0" compared to "1". This is done by adding redundancy to the original data in the form of flag bits. This coding scheme is an asymmetric coding that encodes users'

original codewords that have less than 50% 0s in their data bits. This coding is applied to words of different segment sizes. We will indicate the characteristics of the codewords generated by the coding segments of sizes 2, 4, and 8 bits. Here, for the sake of brevity and ease of presentation, we decided to discuss the encoding of 4 bit codewords. However, for real-life application, we propose 8 bit asymmetric coding, which not only has a reasonable order of information redundancy but also has a fairly good performance compared to the other parameters.

In the case of encoding 4 bit codewords, at first the bus coming into the cache, from any other unit, is partitioned into 4 bit codewords. Then, each of these segments of 4 bit words is encoded to produce codewords to be submitted into the cache storage medium. In the case of 4 bit segments, the generated codewords from the coding have 1 bit information redundancy added to them. Hence, the flag bit indicating the status of the word's coding state adds 25% of information redundancy to the generated codewords. Table I gives the indication of the aforementioned coding scheme having 4 bit data as input and a 5 bit output that gets to be written in the cache cells.

TABLE I. INDICATION OF THE CODING SCHEME ON 4 BIT CODEWORDS

Decimal	Properties of Coding			
	Original	*Encoded*	*Prob. of Zero*	*Over Head*
0	0000	00000	100%	25%
1	0001	00001	75%	
2	0010	00010	75%	
3	0011	00011	50%	
4	0100	00100	75%	
5	0101	00101	50%	
6	0110	00110	50%	
7	0111	11000	75%	
8	1000	01000	75%	
9	1001	01001	50%	
10	1010	01010	50%	
11	1011	10100	75%	
12	1100	01100	50%	
13	1101	10010	75%	
14	1110	10001	75%	
15	1111	10000	100%	

In Table I, the generated codewords have a total 69% of 0s in their bit patterns. The added flag, as the indication of the status of the asymmetric coding in the codewords, is 25% of the data. Hence, the size of the data submitted and written in the cache is 125% of the size of the user's initial data. The Algorithm 1 indicates the procedure of asymmetric coding on 4 bit codewords, where the generated codewords are 5 bit.

ALGORITHM 1. 4 BIT ASYMMETRIC ENCODING PROCEDURE

```
1:  Procedure Encoding (Input [4])
2:  Codeword [5]
3:  //Asymmetric Data
4:  Codeword←Append a '0' as the MSB to the Input
5:  If (Weight > 2)do
6:      Flip the Codeword's bits
8:  Output(Codeword [5])
```

In Algorithm 1, at first a data of size 4 is passed in as an input argument. The codeword of size 5 is generated by appending a single "0" to the most significant bit of the 4 bit input. Then, if the weight of the generated codeword is bigger than two, we flip the codeword's bits. Also, in the decoding phase of the codewords, we take a look at the 5th bit of the freshly retrieved data from the cache. In case where the value of this bit is "0", we simply truncate this bit and the remaining 4 bit word will be the user's original data. However, if the MSB of the retrieved 5 bit codeword equals "1", all of the bits in the codeword are inverted to wind up with the original data.

Figure 1 illustrates the required circuit design for the asymmetric encoding scheme.

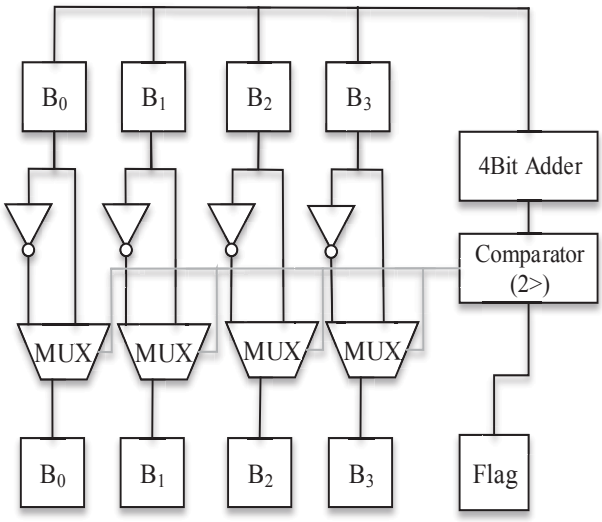

Fig. 1. Schematic Circuit Diagram of the 4bit asymmetric encoding

Other settings of the coding are the encoding of 8 bit and 2 bit codewords. The evaluation of the percentage of 0s generated in the codewords and the order of information redundancy added to them are indicated in Fig. 2.

In our evaluation, we choose the 8 bit asymmetric coding setting where 8 bit codewords are encoded into 9 bit codewords. Since, the overhead of this setting is of trivial order compared to the other codewords and also the overall ratio of 0s to 1s is noticeable. These codewords have an overall 64% of 0s and 12.5% redundancy added to them.

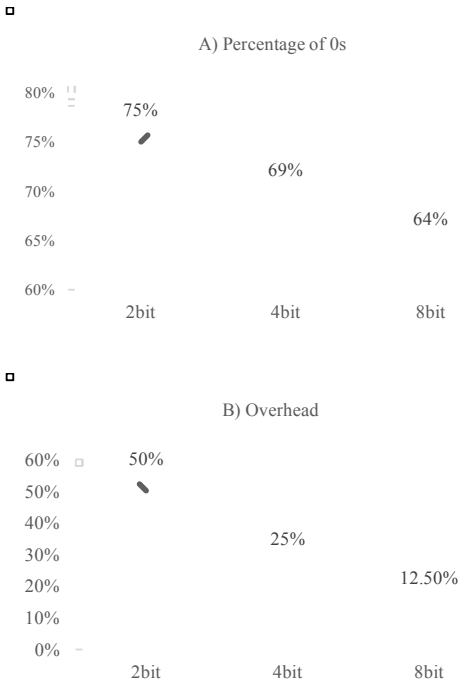

Fig. 2. A) The Percentage of 0s. B) the Overhead of the Generated Codewords in Different Settings

In the following section, after a thorough evaluation of the three different proposed hybrid caches, we obtain and capture the characteristics and crucial statistics of cache data from Splash-2 [21] and Parsec [22] benchmark suits. We obtained the cache data by modifying the Sniper simulator running aforementioned benchmarks. Then, we evaluate the effect of the proposed coding scheme on elevating the calculated statistics from original data to better characteristics that favor our desirable criterion. Finally, the write, the leakage power, and the total cache area are evaluated using NVSim and CACTI modeling tools.

V. Evaluation

In this section, the evaluation of our proposed architecture is proposed in details and the power aspects are thoroughly investigated. Then, the asymmetric coding scheme is utilized to encode the data written in the cache by applications from the Splash-2 and Parsec benchmark suits. Finally, the capabilities of the proposed coding scheme to reduce the cells' write power and energy on the cache data obtained from the aforesaid benchmarks are studied. In Table II, the specification of the machine utilized to obtain the cache data from the benchmark suits using Sniper simulator is described.

TABLE II. Specification of the Machine Used for Evaluation

Device	Specification
Processor	Nehalem Core model (Sniper Config), 2.66 GHz Frequency, 64bit processor
L1 Cache	8-Way, 256KB dcache, 64 Cache Block Size, LRU
L3/L2 Cache	8-Way, 4MB L2 + 16MB L3 (STT-RAM), 64 Cache Block Size, LRU, 16MB L2 + 32MB L3 (PRAM)

A. Evaluation of the cache design

The NVSim [6] is employed to simulate the caches and also the cache cells. The results of evaluating different build technologies used in our proposal are given in Table III.

TABLE III. Details of Cache Technologies Employed

Factor	Technology		
	SRAM (256 KB)	*PRAM (16 MB)*	*STT-RAM (4 MB)*
Read (nj)	0.276	0.136	0.528
Write (nj)	0.254	0.65	1.04
Leakage (mw)	342	27	24
Latency R (ns)	0.706	11.2	1.98
Latency W (ns)	0.572	152	11.15

The capacity of the memories and their technologies have an important role in the amount of power consumed by them. Therefore, in Table III, the size of the SRAM is limited to 256KB. The feature size in these memories is set to 45nm and 90nm for PRAM. The 32MB PRAM and 16MB STT-RAM specifications are not indicated in Table III. We evaluated numerous benchmarks and programs with the aid of the Sniper simulator. The set of the selected applications from the Splash-2 and Parsec benchmarks are presented in Table IV.

TABLE IV. Utilized Programs from Parsec and Splash-2

Benchmark	Program	Domain
Parsec	Facesim	Animation
	Blackscholes	Financial Analysis
	Bodytrack	Computer Vision
	x264	Media Processing
	Vips	Media Processing
	Swaptions	Financial Analysis
	Freqmine	Data Mining
	Fluidanimate	Animation
	Ferret	Similarity Search
Splash-2	FFT	Signal Processing
	FMM	High-Performance Computing
	Cholesky	High-Performance Computing
	Barnes	High-Performance Computing
	Radiosity	Graphics
	Radix	General
	Volrend	Graphics

In our evaluations, three distinct organizations of cache memory architecture are evaluated. The first solution employs three levels of SRAM. The second implementation is using the SRAM for the first level and PRAM for others. This combination is named SRAM-PRAM cache. The third implementation employs SRAM in L1 and STT-RAM in the other two levels, which is called SRAM-STT-RAM. The differences of the total dynamic and static power consumption for the proposed implementations are given in Figure 3.

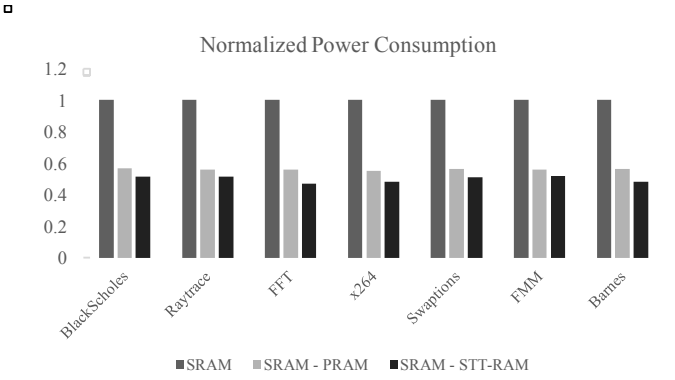

Fig. 3. Normalized power characteristics of different programs with proposed cache designs

The evaluations prove the implementation that uses both of the PRAM and STT-RAM technologies in the L2/L3 caches outperform the implementation that utilizes SRAM in all of the three levels of the cache. This is due to the fact that the static power consumption (leakage power) of these two technologies is much lower than that of SRAM. Moreover, because of the high access time latency of the STT-RAM compared to the SRAM, the STT-RAM is not yet a suitable substitute for the SRAM in level 1 cache. It is beneficial to use MRAM and PRAM in L2/L3 caches, where the access rates are lower than L1 cache. Also, it is marked in the above figure that the utilization of STT-RAM in L2/L3 caches is more power efficient. Note that the access rate in the L2/L3 caches is less than that of the L1 cache. Finally, by considering the higher access latency of PRAM compared to STT-RAM, the STT-RAM and SRAM organization for the cache unit is the proper solution.

We have also calculated the total area for the proposed architectures by NVSim. The case in which all of the three cache levels are SRAM based, the area of the cache cells is 53.8 mm^2. The PRAM and STT-RAM solutions take 43 and 23.7 mm^2 chip area, respectively. Hence, the SRAM-STT-RAM solution is the most efficient solution for the area. So, we have a 20% for SRAM-PRAM and 45% for SRAM-STT-RAM area reduction compared to pure SRAM solution. The SRAM-STT-RAM architecture has 50% reduction in the total power consumption compared to the pure SRAM architecture design. Furthermore, SRAM-PRAM solution reduces the read/write dynamic and static power consumption by 43% compared to SRAM. As a result, we choose the SRAM-STT-RAM cache to be our selected solution.

B. Evaluation of the asymmetric coding scheme

In this section, we outline the effect of the asymmetric coding scheme on the data obtained from the aforementioned programs by modifying the Sniper simulator. As mentioned above, the organization of STT-RAM and SRAM proves to be more power and area efficient compared to the other two solutions. However, the write energy of the STT-RAM is considerably higher compared to PRAM and SRAM, which is about 5 times more than SRAM cache. To further reduce the power consumption of the cache unit, we strive to exploit the fact that writing a "0" is less power consuming compared to writing a "1". To propose a solution for increasing the ratio of 0s to 1s in the data and reduce the number of write and set operations issued to the cache. The asymmetric coding discussed earlier, encodes data to increase our favored statistics by adding minimal overhead to the data.

The cache data obtained by our simulations is investigated to calculate the ratio of 1s to 0s prior to the coding and also after the coding. These results are given in Figure 4 and indicate that the asymmetric coding gives an overall of 15% reduction in the ratio of 1s to the data bits. It reduces the number of write and set operations submitted to the cache by 47 percent, subsequently.

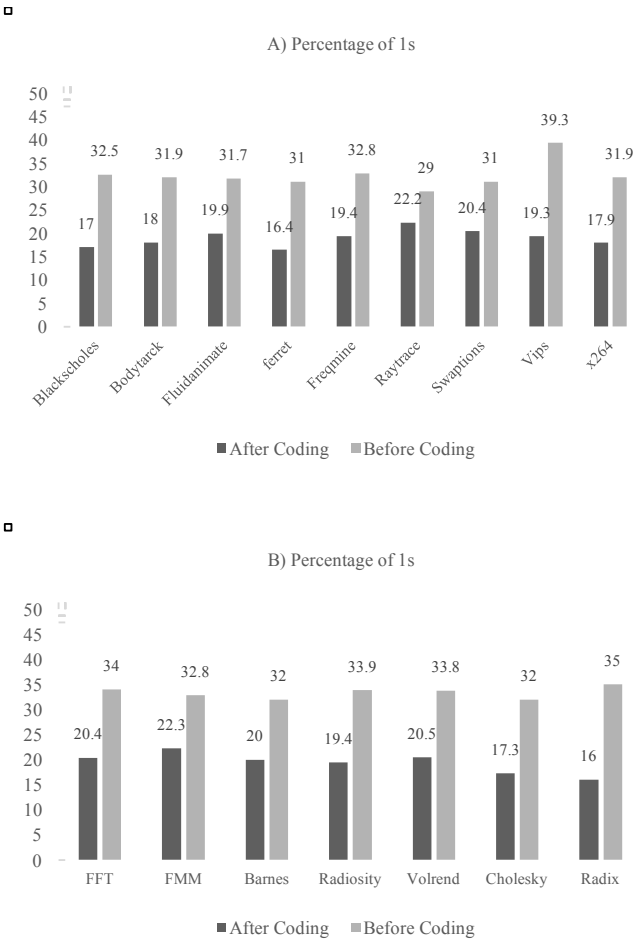

Fig. 4. Evaluation of Coding on Parsec (A) and Splash-2 (B) benchmarks

The percentage of 1s in the data is obtained by evaluating the words that have at least one bit valued "1" since, only these words consume cache set energy. It is proven that the asymmetric coding reduces the number of write operations submitted to the cache by an overall of 47%. As shown in Figure 4, this is accomplished by increasing the ratio of 0s to 1s in the data written in the cache cells. In Figure 5, the evaluation of the asymmetric coding on the read/write power of the programs from the benchmark suits are outlined for STT-RAM and is compared to pure SRAM solution.

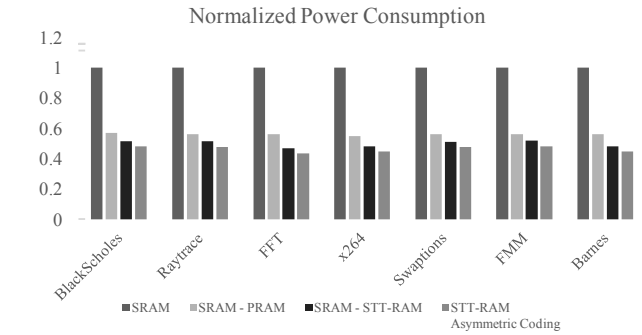

Fig. 5. Asymmetric coding power on Parsec and Splash-2 benchmarks

The asymmetric coding scheme reduces the write energy per access for STT-RAM by 24% and reduces the leakage power of the cache cell by 16%. In the total read/write dynamic and cell's static power the coding causes an average of 14.1% reduction for the benchmarks. The case in which 64 byte words are written in the cache, although 64 flag bits are generated by the coding, this would not pose any problem, since we achieve 45% reduction in the cache area by suggesting a new model for cache architecture.

VI. Conclusion

Based on the evaluation results, the utilization of the NVM in L2/L3 level caches can significantly reduce the power consumption of the caches compared to the other types of technologies by 43-50%. Moreover, the area of the cache on the chip is decreased by 20% and 50% for SRAM-PRAM and SRAM-STT-RAM, respectively. Moreover, by implementing the asymmetric coding scheme on the data and changing the characteristics of the generated codewords, the total power consumption of the cache reduces by 14% and the leakage power of the cells reduces by 16%.

Acknowledgment

We are grateful to Prof. Hamid Sarbazi Azad, Head of the school of computer sciences, for his support and useful guidance. This work has been financially supported by HPC lab, an affiliation of IPM.

References

[1] K. Asanovic, R. Bodik, B. C. Catanzaro, J. J. Gebis, P. Husbands, K. Keutzer, D. A. Patterson, W. L. Plishker, J. Shalf, S. W. Williams and K. A. Yelick, "The landscape of parallel computing research: A view from berkeley," *Technical Report UCB/EECS-2006- 183, EECS Department,* vol. 2, 2006.

[2] S. Das, A. Fan, K.-N. Chen and C. S. Tan, "Technology, Performance, and Computer-Aided Design of Three-Dimensional Integrated Circuits," *Technology, Performance, and Computer-Aided Design of Three-Dimensional Integrated Circuits,* pp. 108-115, Apr. 2004.

[3] S. Rahmani, A. Ahmadzadeh, H. Omid, G. Saeid and P. Mirhosseini, "An Efficient Multi-core and Many-core Implementation of K-Means Clustering," *ACM-IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE),* pp. 128-131, December 2016.

[4] C.-L. Su and A. M. Despain, "Cache Design Trade-offs for Power and Performance Optimization: A Case Study," *Proceedings of the international symposium on Low power design,* pp. 63-68, Apr. 23 1995.

[5] C. Chakrabarti, "Cache design and exploration for low power embedded systems," *Performance, Computing, and Communications,* pp. 135-139, Apr. 2001.

[6] X. Dong, C. Xu, Y. Xie and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging non-volatile memory," *In Emerging Memory Technologies,* pp. 15-50, 2014.

[7] P. Shivakumar and N. P. Jouppi, "CACTI 3.0: An Integrated Cache Timing, Power, and Area Model," 2001.

[8] T. E. Carlson, W. Heirmant and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," *SC '11: Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis,* Nov. 2011.

[9] C. Lam, "Cell Design Considerations for Phase Change Memory as a Universal Memory," *International Symposium on VLSI Technology, Systems and Applications. VLSI-TSA.,* pp. 132-133, Apr. 2008.

[10] M. Hosomi, H. Yamagishi, T. Yamamoto and K. Bessho, "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram," *IEEE International Electron Devices Meeting, 2005. IEDM Technical Digest,* pp. 459-462, Dec. 2006.

[11] M. Rasquinha, D. Choudhary, S. Chatterjee, S. Mukhopadhyay and S. Yalamanchili, "An energy efficient cache design using spin torque transfer (STT) RAM," *Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design,* pp. 389-394, Aug. 2010.

[12] Z. Diao, Z. Li, S. Wang and Y. Ding, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *Journal of Physics: Condensed Matter,* vol. 19, no. 16, 2007.

[13] A. Pirovano, A. Lacaita, A. Benvenuti and F. Pelizzer, "Scaling analysis of phase-change memory technology," *Electron Devices Meeting. IEDM '03 Technical Digest. IEEE International,* pp. 29-36, Dec. 2003.

[14] R. Bez and A. Pirovano, "Non-volatile memory technologies: emerging concepts and new materials," vol. 7, no. 4-6, pp. 349-355, Nov. 2004.

[15] P. Mangalagiri, K. Sarpatwari, A. Yanamandra, V. Narayanan and Y. Xie, "A low-power phase change memory based hybrid cache architecture," *Proceedings of the 18th ACM Great Lakes symposium on VLSI,* pp. 395-398, May 2008.

[16] G. Dhiman, R. Avoub and T. Rosing, "PDRAM: A hybrid PRAM and DRAM main memory system.," *Design Automation Conference. DAC '09. 46th ACM/IEEE,* pp. 664-669, 2009.

[17] R. Banakar, S. Steinke and B.-S. Lee, "Scratchpad memory: design alternative for cache on-chip memory in embedded systems," *Proceedings of the tenth international symposium on Hardware/software codesign,* pp. 73-78, May 2002.

[18] J. Hu, C. J. Xue, Z. Qingfeng and W.-C. Tseng, "Towards energy efficient hybrid on-chip scratch pad memory with non-volatile memory," *Design, Automation & Test in Europe Conference & Exhibition (DATE),* pp. 1-6, Mar. 2011.

[19] C. W. Smullen, V. Mohan, A. Nigam and S. Gurumurthi, "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," *IEEE 17th International Symposium on High Performance Computer Architecture (HPCA),* pp. 50-61, Feb. 2011.

[20] X. Wu, J. Li, L. Zhang, E. Speight and Y. Xie, "Power and performance of read-write aware hybrid caches with non-volatile memories," *Design, Automation & Test in Europe Conference & Exhibition. DATE '09.,* pp. 737-742, Apr. 2009.

[21] S. C. Woo, m. Ohara, E. Torrie and J. P. Singh, "The SPLASH-2 programs: characterization and methodological considerations," *22nd Annual International Symposium on Computer Architecture,* pp. 24-36, Jun. 1995.

[22] C. Bienia, S. Kumar and J. P. Singh, "The PARSEC benchmark suite: characterization and architectural implications," *Proceedings of the 17th international conference on Parallel architectures and compilation techniques,* pp. 72-81, Oct. 2008.