AN EFFICACIOUS RUNTIME ADAPTIVE HYBRID DRAM/PRAM MEMORY IN FPGA PLATFORM

M. Isaivani[1], Dr. V. Malathi[2] and Dr.E.Sakthivel[3]

[1]Ph.D Scholar - Electrical Engineering, Anna University Regional Campus Madurai, Madurai, Tamil Nadu, India
[2]Professor, Department of Electrical and Electronics Engineering, Anna University Regional Campus Madurai, Madurai, Tamil Nadu, India
[3]Associate Professor, PSR Engineering College, Sivakasi, Tamil Nadu, India

ABSTRACT

Hybrid main memory comprising of DRAM and PRAM becomes quite popular because of the less standby power benefit of PRAM and high performance of DRAM. In this work, the runtime-adaptive control and DRAM bypassing methods are introduced in order to minimize DRAM refresh energy that occupies a considerable portion of total system power. The work is carried out by using Xilinx 12.1 simulation tool and the experimental result proves that in the proposed work power consumption is greatly reduced i.e., only requires 3 %, with less area overhead while maintaining the speed parameter by comparing with the conventional method .

KEYWORDS

Cache memory, Dynamic Random Access Memory, Phase-Change Random Access Memory, Write-back and Fill method

1. INTRODUCTION

Currently, Very Large Scale Integration (VLSI) circuits are miniature in size to promote device functionalities and performance parameters. DRAM becomes a performance impediment in many systems due to the pin count, speed and pin bandwidth are risen extremely slowly. DRAM is one of the highest power consumers in modern computing systems. So that, power budget of DRAM is quite similar to or sometimes, even surpasses the power budget of CPU. It is proven that DRAM in a commercial system uses 25–45% of total system power. With this scenario, energy efficient DRAM has become a crucial design constraint for the system design.

Even DRAM performance emerges as a critical issue in the modern server systems because of the universal utilization of data-centric applications. Since DRAM bandwidth is not scalable and considered as a limited resource, huge contention exists in DRAM that debases total system performance significantly. To enhance DRAM performance and to lower dynamic-energy consumption, the row-level access locality should be enhanced and the count of row activations should be decreased. The row activations are considerably reduced by performing more read and write operations for each activated row. More row activations result in higher wastage of energy.

N. Ramasubramanian et al.(Eds) : SRAIC-2019
pp. 47-57, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.91504

This paper analyzes DRAM activities and implements cache line write-backs and fill methods together with the hybrid DRAM/PRAM memory .In the proposed hybrid main memory, PRAM operates as a background main memory due to its low stand-by power benefit and DRAM serves as a cache owing to its low latency and lower power consumption during read and write operations. This hybrid main memory system takes advantage of both memories while reducing the negative aspect of limitations in both of them.

2. LITERATURE SURVEY

A brief survey of various techniques involved in data compression is presented for main memory and cache memory [1]. A set- and-way management approach is proposed in [2] for a runtime-reconfiguration of a cache's size and associativety. Deterministic nap technique and early miss detection approach are proposed in [3] in order to diminish static as well as dynamic power. "Latency-Programmable System Emulation Memory" is implemented in [4] that permits read-and/or-write latency scale-up capability. Direct data access can be achieved in eDRAM cache by proposing tag-comparison in memory that enhanced energy efficiency [5]. The "eXplicitMulti-Threading (XMT)" paradigm is designed for a Parallel Random Access Model on-chip processor [6].PRAM-On-Chip Phase Change Memory (PCM) provides higher capacity than DRAM and evolves for a greater capacity memory [7,8, 9, 10]. Gen2 Hybrid Memory Cube is characterized from data-centric demands. HMC offers considerable accessing bandwidth while requiring less power [11, 12]. Novel write algorithms for PCM are discussed detail in [13]. Retention-aware placement method is employed in DRAM to lessen refresh power [14]. In [15], statistical populations' method is proposed for DRAM cache and also inherent error-control technique is utilized to minimize refresh rate. Different techniques are discussed for utilizing memory controller in order to enhance energy efficiency of DRAM and also to manage power in DRAM [16]. Another multi-partition based memory controller BIBIM is designed that combines DRAM and PRAM [17]. Memory access scheduling technique is introduced in [18] that reorder memory references to improve the energy efficiency. An effective Collective write-back and fill method is employed for DRAM cache in order to enhance the data throughput [19].

3. EXISTING SYSTEM

Collective Write-back-cum-Fill method [19] consists of the following two approaches: write-back strategy and fill strategy. It is executed by the feedback exists between Dynamic-Write-back-Unit (DWU) and Adaptive DRAM Placement (ADP) entities. The write-back strategy is carried out by DWU. When L3cache replacement occurs, checking is done on the victim line to find if it is dirty. If there is no any modified bit, the victim line is evacuated from L3cache without writing back to the DRAM cache. If the victim line is dirty, DWU will decide if write-back is to be done on DRAM cache. The line is added to the DRAM cache if affirmative decision is taken. Or else, DRAM cache hit/miss will be verified. The line should be introduced into DRAM cache with the condition that is a hit in order to maintain regularity. If not, the line is infused into main memory. Fill strategy is carried out by ADP entity which decides whether to add fill requests to DRAM cache.

The CWFP method requires two vital units: Integrated Set Monitors (ISM) and Lose-to-Gain Dispatcher (LGD). Based on the investigation of DRAM cache accessing management, the module ISM determines which activity to be performed on DRAM cache i.e., write-back or fill operation .An another appropriate entity called, Miss-Status Handling Register (MSHR) includes W and F bits to decide for write-back and fill cache operations respectively. This system offers LGD over ISM in the aim to avert needless write-back access to DRAM cache and it also aids to rectify the feasible wrong perception of ISM. Remarkably, LGD supports the

ISM's decision making capability and in the way DRAM interference got reduced. Moreover it evaluates the accuracy of write-back access based on per-core criteria and it curtails the cache resources allotment to useless write-backs.

- Though existing CWFP scheme reduced DRAM interference, inter-core contention and achieved greater speedup, it still suffered with higher power consumption of DRAM. The major amount of total power available in the system is consumed by this unit.
- The following section describes the proposed hybrid DRAM/PRAM scheme to address the above challenge.

4. PROPSED SYSTEM

4.1. PRAM Overview

DRAM has been adopted for the main memory in computer systems for several years. Many non-volatile memories like Phase change RAM (PRAM), Magnetic RAM (MRAM) and Ferroelectric RAM (FRAM) are developed for future generation technologies. In these non-volatile memories (NVM), PRAM becomes popular entity for main memory due to its advantageous properties of low power requirement and high density. A PRAM cell is made up of phase change material for bit representation. PRAM density is very much larger than DRAM (expected to be about four times). Moreover, unvarying property of phase material after power-off, it exhibits petty leakage energy irrespective of memory size.

4.2. Implementation

In this effective hybrid DRAM/PRAM method, DRAM plays as a last-level cache and operates as the working memory while PRAM performs as a massive background main memory. Figure 1 shows the architecture of this hybrid memory, where the memory controller manages the tag structure in DRAM and the counter governs each row in DRAM. By decaying the DRAM contents, power management can be achieved. If the data are clean, they are expelled from DRAM. For the modified data, they should be written back to PRAM before eviction. When the new data are written to DRAM which has been read from PRAM, global time-out value is set to the counter. The counter value is decreased by one periodically. It is again set to the time-out value after row access takes place. When the counter reaches zero, the row will be expelled. That removed row does not need refresh which leads to attain considerable energy saving. The following two methods are utilized to manage the power in the proposed scheme.

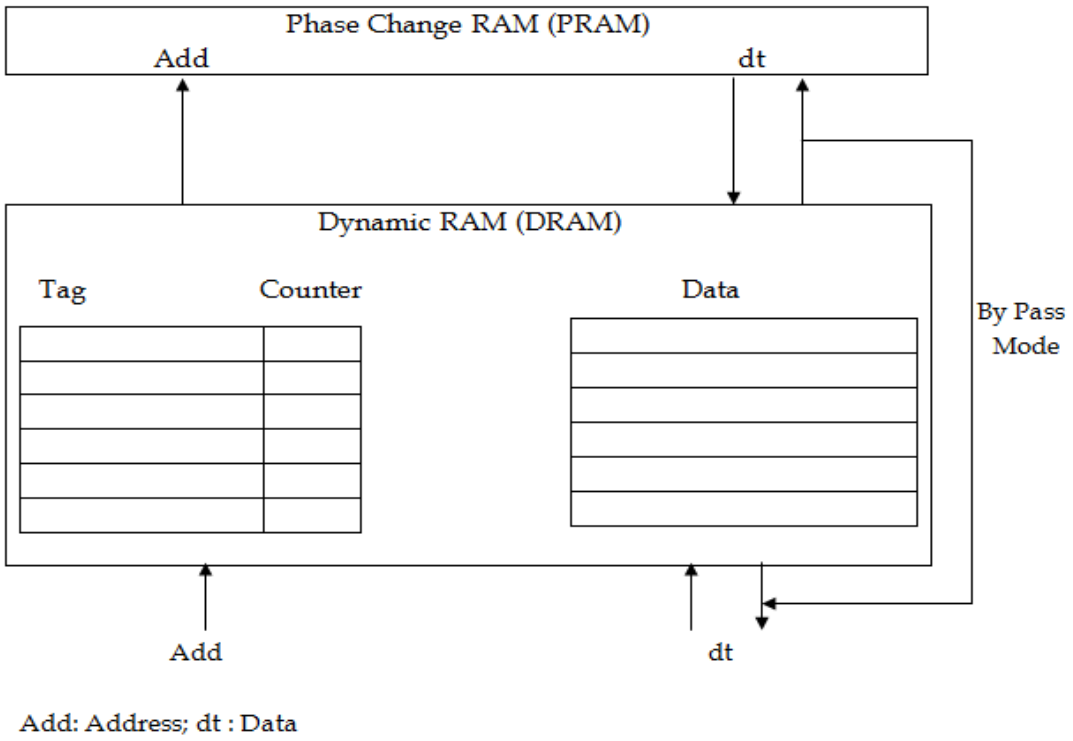

Figure 1. Hybrid DRAM/PRAM Memory Scheme

4.2.1. Runtime-Adaptive Control of Time-Out Value

This method is performed that reduces the total energy while satisfying system performance constraints. When time-out value increases, many rows in DRAM would be live and needs refreshes and so DRAM energy rises. But PRAM energy decreases and PRAM accesses are less because much data are accessed by DRAM with a huge hit rate. There is a best time-out value (BTO) that provides less total energy of the hybrid scheme by averting DRAM refreshes. The BTO value changes because of the dynamically varying manner in memory access.

4.2.2. Bypassing DRAM

DRAM bypassing is performed to the first read operation that fetches a miss in DRAM cache. When in the second access to the corresponding row in DRAM within the time-out value, the data in row are replicated to DRAM from PRAM and that row begins to be refreshed to perform further accesses. This method screens out memory accesses that have poor spatial locality and to enhance the DRAM refresh energy efficiency.

5. RESULTS AND DISCUSSION

The proposed system is implemented in Xilinx 12.1 tool and written in VHDL (Very large scale integration Hardware Description Language) coding to perform evaluation for both existing cache write-back and fill method with proposed runtime-adaptive hybrid DRAM/PRAM schemes.

5.1. Performance Parameters of Existing System

5.1.1. Area Consumption

The area requirements of slice registers, LUT, LUT –flip-flop pairs, bonded IOB buffers are 41, 44, 40, 18,1 respectively as shown in figure 2.

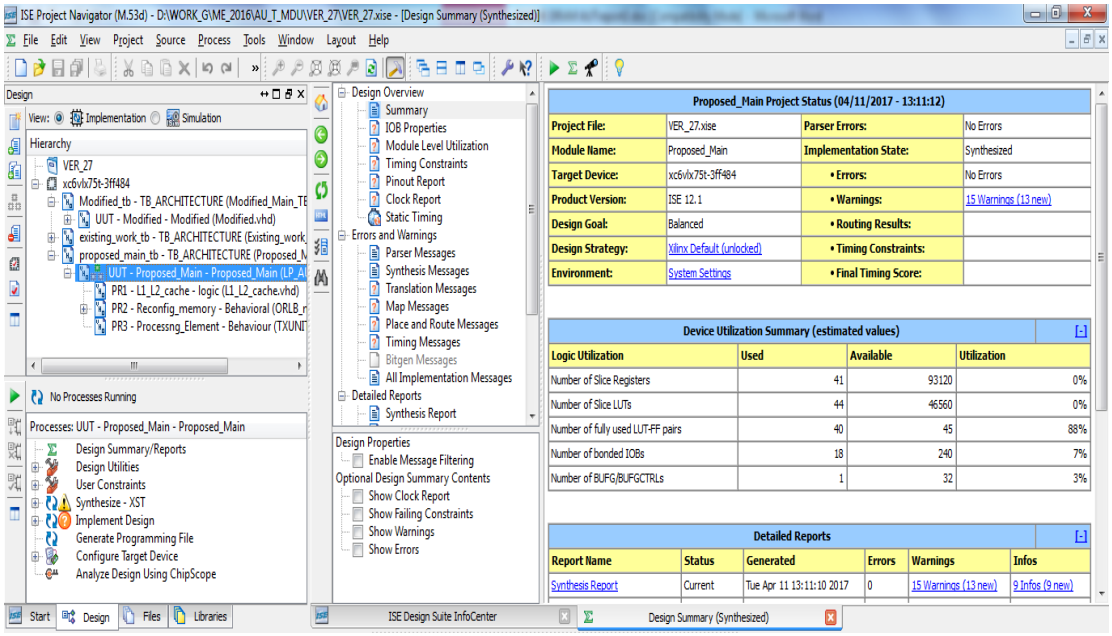

Figure 2. Area consumption of Existing System

5.1.2. Speed Evaluation

The output shows the minimum required time and maximum frequency are 1.425ns and 701.991MHz respectively. The minimum input arrival time before clock, maximum output required time after clock, maximum combinational path delay are 0.929ns, 0.567ns, 0.289ns respectively. It is shown in figure 3.

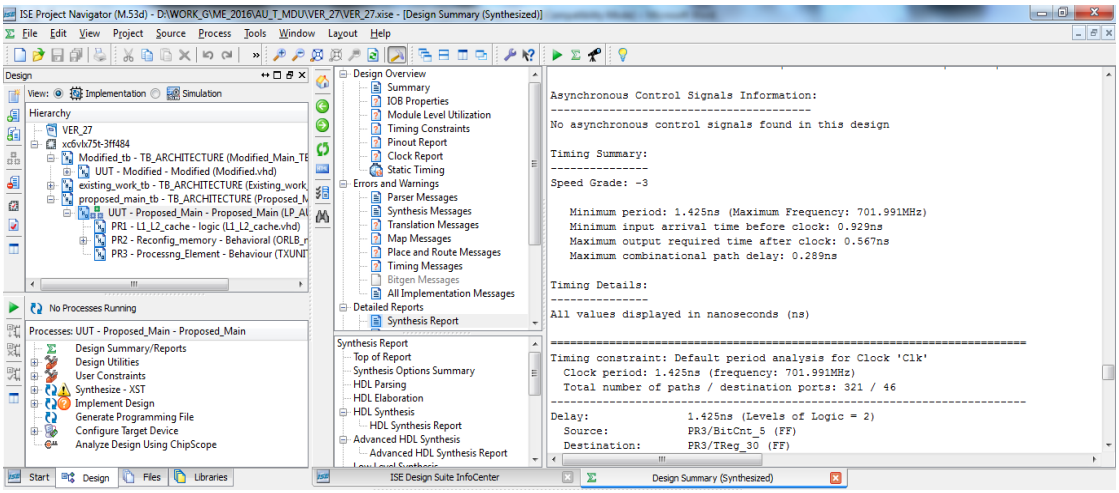

Figure 3. Speed Evaluation of Existing System

5.1.3. Power Consumption

The existing system which involves the collective write-back and fill method, requires 7.6% of total power available. It is computed from figure 4.

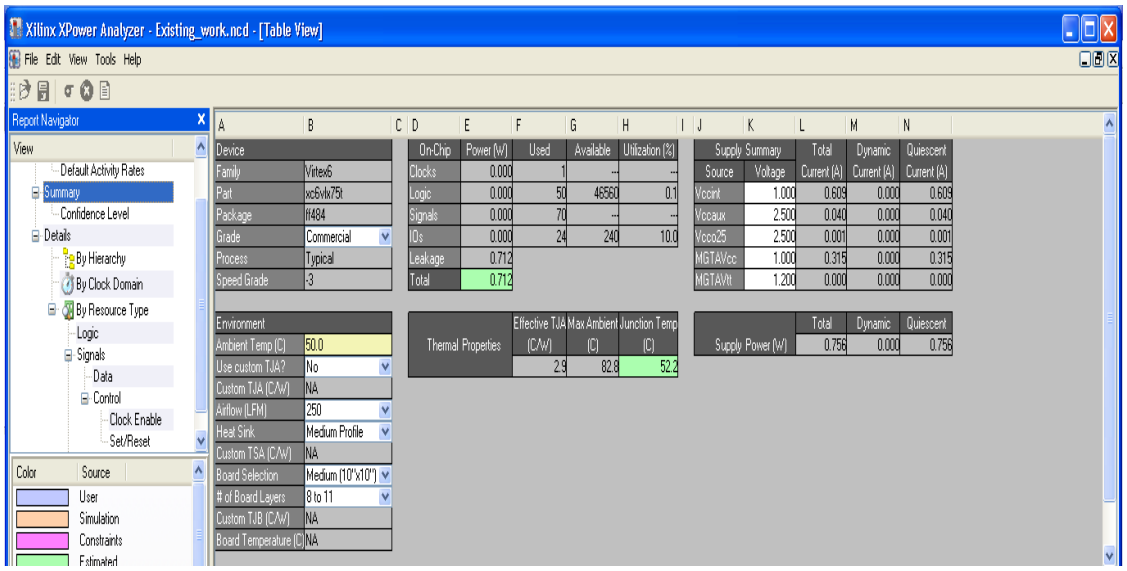

Figure 4.Power Consumption of Existing System

5.2. Performance Measurement of Proposed System

5.2.1. Area Consumption

The area requirements of slice registers, LUT, LUT –flip-flop pairs, bonded IOB buffers are 39,44,39,7,1 respectively as shown in figure 5. Thus the area consumption is reduced from the existing system.

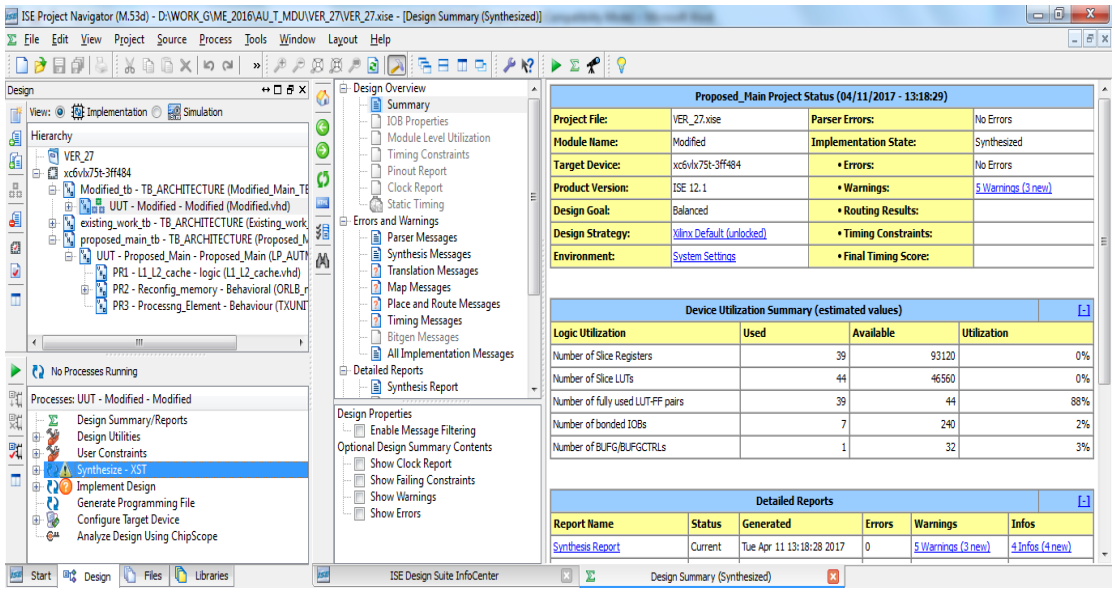

Figure 5. Area consumption of Proposed System

5.2.2. Speed Evaluation

The output shows the minimum required time and maximum frequency are 1.425ns and 701.991MHz respectively. The minimum input arrival time before clock, maximum output required time after clock are 0.920ns, 0.562ns respectively. It is shown in figure 6. The output of speed measurement is as same as the conventional method and this has to be enhanced in the future work. The delay is slightly better than the existing method.

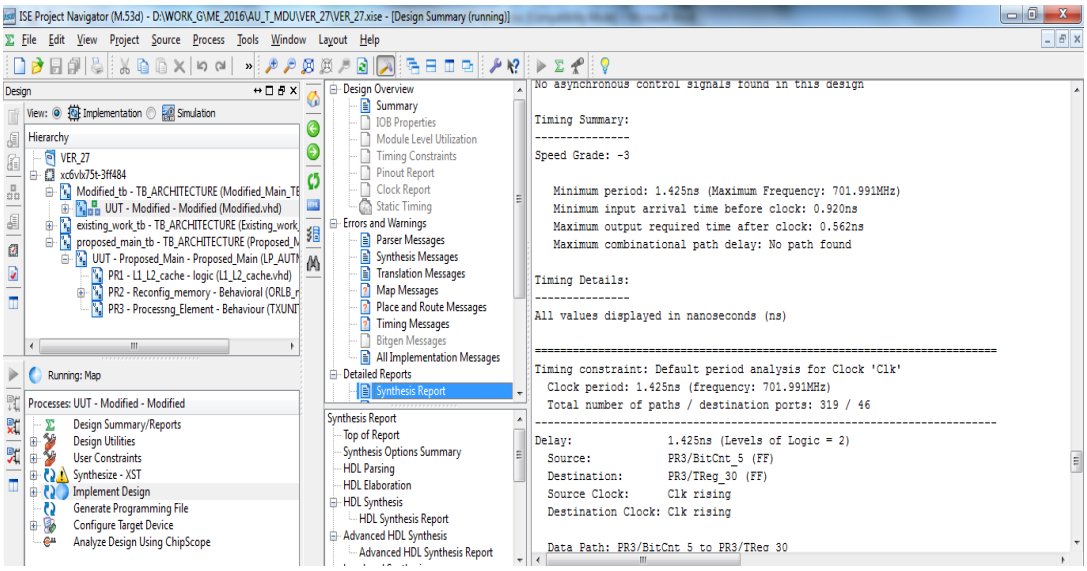

Figure 6. Speed Evaluation of Proposed System

5.2.3. Power Consumption

The proposed run-time adaptive hybrid DRAM/PRAM memory with cache write-back and fill method requires only 3% of total power which is greatly diminished while comparing with the conventional method. It is calculated from figure 7.

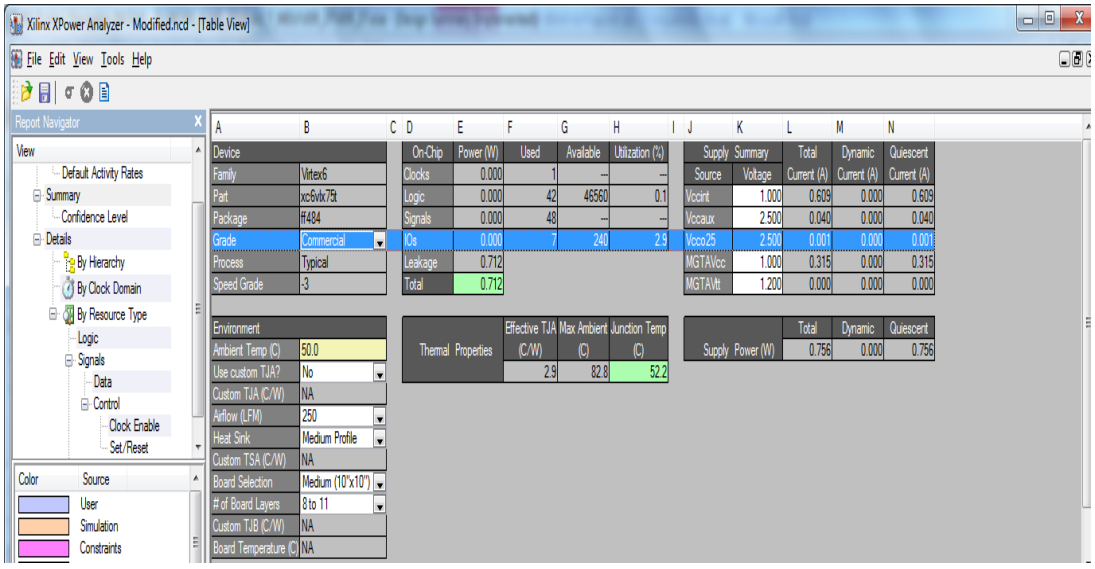

Figure 7. Power consumption of Proposed System

5.3. Simulation Waveforms

5.3.1. Existing System

The simulated result of DRAM cache with collective write-back and fill method is shown in below figure 8.

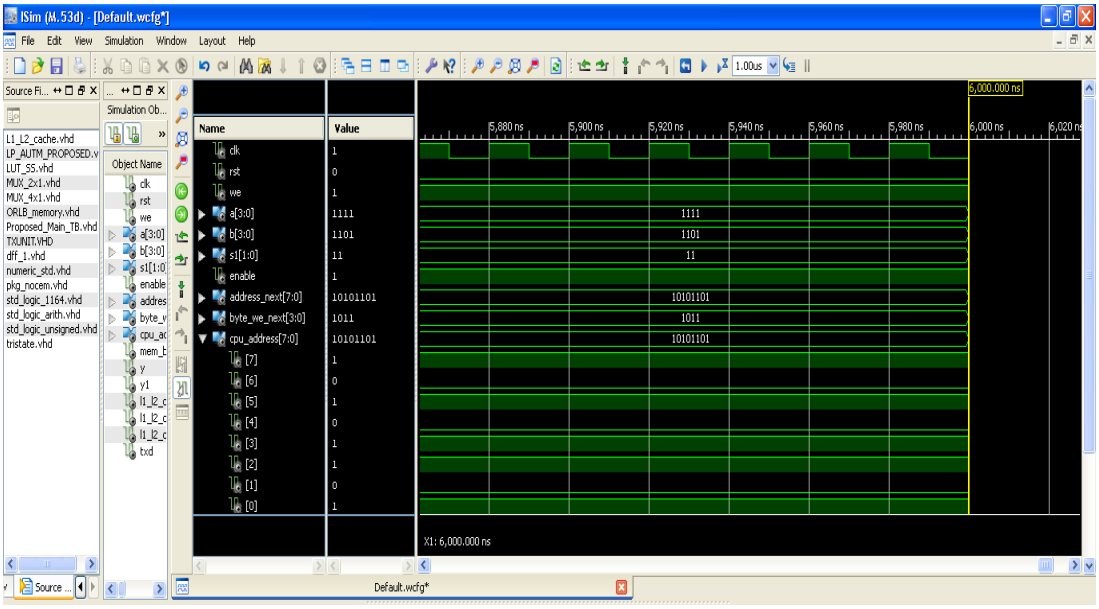

Figure 8. The Simulated Result of Existing System

5.3.2. Proposed System

The simulated result of hybrid DRAM/PRAM memory with cache collective write-back and fill method is shown in below figure 9.

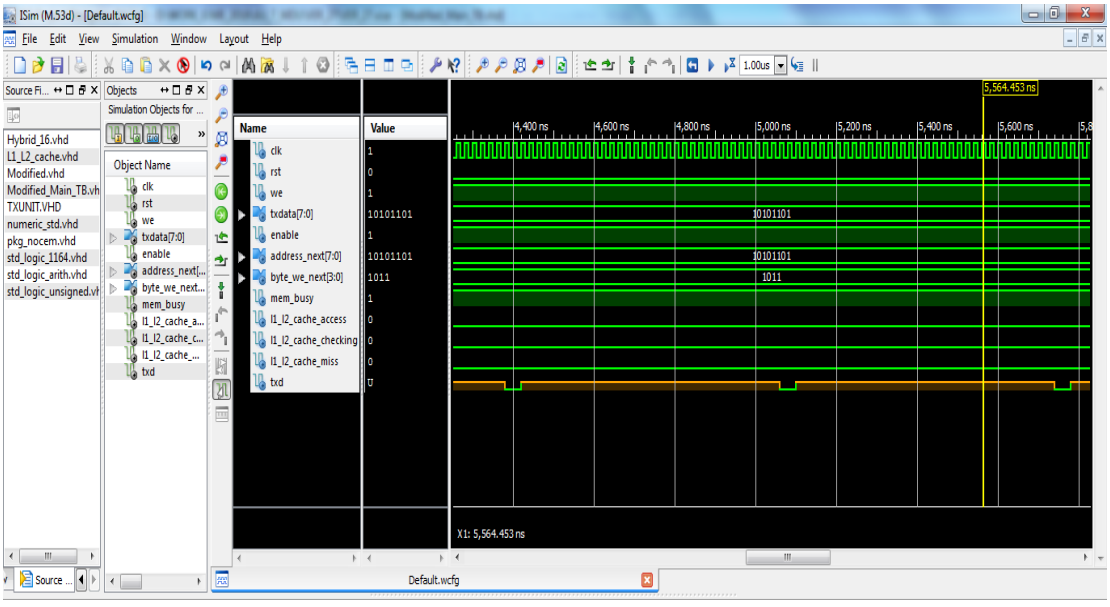

Figure 9. The Simulated Result of Proposed System

5.4. Performance Comparison

Table 1 Comparison Analysis

S. No	Approaches	Area	Power	Latency
1	Latency-Programmable System Emulation Memory [4]	NA	NA	120 ns
2	PRAM-On-Chip Processor [6]	NA	NA	193 ms
3	Hybrid Memory Cube[11]	NA	11w	68 ns
4	3D Stacked Memories [12]	NA	105 w	24 µs
5	BIBIM [17]	NA	NA	0.13 µs
6	CWFP [19]	SR - 41 LUT - 44 LUT-FF- 40 IOB - 18 Buffer - 1	71 mw	0.929 ns
7	Proposed Run-time Adaptive Hybrid DRAM/PRAM memory	SR - 39 LUT - 44 LUT-FF - 39 IOB - 7 Buffer - 1	21 mw	0.920 ns

Table 1 shows the comparison of area, power and latency parameters for various methodologies with the proposed approach. It is evident that the proposed methodology outperforms all other mentioned approaches in terms of the stated performance metrics.

6. CONCLUSION AND SCOPE OF FUTURE WORK

6.1. Conclusion

Thus, the system focuses on the new approach to reconfigure optimizing Hybrid DRAM/PRAM performance and energy consumption. Under this technique, dirty cache lines which are not been accessed recently have been written to DRAM while the respective row is activated by a read operation. This permits both read and write operations to target the particular one. The proposed hybrid memory offers lower standby power as well as higher performance for the server systems. In the proposed work power consumption isgreatly reduced i.e., only requires 3%, and it results with less area overhead while maintaining the speed parameter by comparing with the conventional method.

6.2. Future Scope

Future work should be done to verify the suitability of using the scheme proposed in this approach qualitatively and quantitatively on multi-core systems. Thus, the proposed scheme would need to be qualitatively and quantitatively evaluated, analyzed and improved in more detail to be efficiently working in multi, or even many-core systems.

REFERENCES

[1] Sparsh Mittal , Jeffrey S. Vetter, (2016) "A Survey Of Architectural Approaches for Data Compression in Cache and Main Memory Systems", IEEE Transactions on Parallel and Distributed Systems,Vol. 27, No. 5, pp1524 – 1536.

[2] Karthik T. Sundararajan, Timothy M. Jones, Nigel P. Topham, (2013) "The Smart Cache: An Energy-Efficient Cache Architecture Through Dynamic Adaptation", International Journal of Parallel Programming, Vol. 41, No. 2, pp305–330.

[3] Oluleye D. Olorode, MehrdadNourani, (2015) "Improving Cache Power and Performance Using Deterministic Naps and Early Miss Detection", IEEE Transactions on Multi-Scale Computing Systems, Vol. 1, No. 3, pp150 – 158.

[4] Mu-Tien Chang, I. Stephen Choi, DiminNiu, Hongzhong Zheng, (2018) "Performance Impact of Emerging Memory Technologies on Big Data Applications: A Latency-Programmable System Emulation Approach",GLSVLSI'18, Chicago, IL, USA.

[5] Keng-Hao Yang, Hsiang-Jen Tsai, Chia-Yin Li, Paul Jendra, Meng-Fan Chang, Tien-Fu Chen, (2017) "eTag: Tag-Comparison in Memory to Achieve Direct Data Access based on eDRAM to Improve Energy Efficiency of DRAM Cache", IEEE Transactions on Circuits and Systems, Vol. 64, No. 4, pp858 – 868.

[6] XingzhiWen , Uzi Vishkin, (2008) "FPGA-Based Prototype of a PRAM-On-Chip Processor", CF'08, Ischia, Italy.

[7] JaeminJang,Wongyu Shin, JungWhan Choi , Yongju Kim, Lee-Sup Kim, (2019) "Sparse-Insertion Write Cache to Mitigate Write Disturbance Errors in Phase Change Memory", IEEE Transactions on Computers, Vol. 68, No. 5, pp752 – 764.

[8] M. Abdulla, and M. Greenberg, (2010)"Will Phase Change Memory (PCM) Replace DRAM or NAND Flash?", Flash Memory Summit.

[9] B. C. Lee, *et al.,* (2009)"Architecting Phase Change Memory as a Scalable DRAM Alternative", ISCA '09 Proceedings of the 36th annual international symposium on Computer architecture, pp2-13, Austin, TX, USA.

[10] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, (2009) "Scalable High Performance Main Memory System Using Phase-Change Memory Technology", ProceedingISCA '09 Proceedings of the 36th annual international symposium on Computer architecture, pp24-33, Austin, TX, USA.

[11] MayaGokhale, Scott Lloyd, Chris Macaraeg, (2015) "Hybrid Memory Cube performance characterization on data-centric workloads", ProceedingIA3 '15 Proceedings of the 5th Workshop on Irregular Applications: Architectures and Algorithms, Article No. 7, Austin, Texas, USA.

[12] RamyadHadidi, BaharAsgari, Burhan Ahmad Mudassar, SaibalMukhopadhyay, SudhakarYalamanchili, and Hyesoon Kim, (2017) "Demystifying the Characteristics of 3D-StackedMemories: A Case Study for Hybrid Memory Cube", In: 2017 IEEE International Symposium on Workload Characterization (IISWC), IEEE, USA.

[13] T.Nirschl, et al., (2007) "Write Strategies for 2 and 4-bit Multi-Level Phase-Change Memory", In: 2007 IEEE International Electron Devices Meeting, IEDM, USA.

[14] R. K. Venkatesan, S. Herr, and E. Rotenberg, (2006) "Retention-Aware Placement in DRAM (RAPID): Software Methods for Quasi-Non-Volatile DRAM", In: The Twelfth International Symposium on High-Performance Computer Architecture, HPCA, USA.

[15] P. G. Emma, *et al.*, (2008)"Rethinking Refresh: Increasing Availability and Reducing Power in DRAM for Cache Applications", IEEE MICRO, Vol. 28, No. 6, pp47-56.

[16] I. Hur and C. Lin, (2008) "A Comprehensive Approach to DRAM Power Management", In: 2008 IEEE 14th International Symposium on High Performance Computer Architecture, HPCA, IEEE, USA.

[17] Gyuyoung Park1, Miryeong Kwon1, Pratyush Mahapatra2, Michael Swift2, and Myoungsoo Jung, (2018) "BIBIM: A Prototype Multi-Partition Aware Heterogeneous New Memory", 10th {USENIX} Workshop.

[18] S. Rixner, *et al*., (2000)"Memory Access Scheduling", Proc. ISCA.

[19] Shouyi Yin, Weizhi Xu, Jiakun Li, Leibo Liu, Shaojun Wei, (2016) "CWFP: Novel Collective Writeback and Fill Policy for Last-Level DRAM Cache", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 24, No. 7, pp2548 – 2561.

AUTHORS

M Isaivani received B.E degree in Electronics and Instrumentation Engineering from Anna University in the year of 2006. She has completed M.E in Embedded System Technologies under Anna University in 2014. She is currently doing Ph.D. in full-time at Anna University Regional Campus Madurai. Her research interests include VLSI, Embedded System applications to Power System Protection.

V Malathi is working as a professor in the department of Electrical and Electronics Engineering and Dean in Anna University Regional Campus Madurai. She completed her Bachelor Degree in College of Engineering Guindy and her Master's Degree in Thiyagaraja College of Engg, Madurai. She Completed her Ph.D. in Anna University Chennai and her areas of interest are Intelligent Techniques and its Applications, Smart Grid, FPGA based Power System and Automation.

E Sakthivelreceived the Bachelor degree in Madurai Kamarajar University, Madurai and the Master's degree in Embedded system Technologies in Anna University, Thirunelveli and he completed his Ph.D. degree in Anna University, Chennai. Currently, he is working as an associative professor at PSR Engineering College. He has more than 10 years of industrial experience in VLSI Technology and Embedded System. His current research interests include Embedded Systems, Low-Power Design, Network On Chip, System-On-Chip, FPGA and ASIC based power electronics control circuits, FPGA based power system, Wireless Networks, Instrumentation and Object Recognition.