ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

2017 Pacific Southwest Section
Meeting: Tempe, Arizona Apr 20 Paper ID #20691

A Study of Emerging Memory Technology in Hybrid Architectural Approaches of GPGPU

Dr. Reza Raeisi, California State University, Fresno

DR REZA RAEISI is a Professor of Electrical and Computer Engineering Department at California State University, Fresno. He is also Chair of the ECE department. His research interests include integrated circuits, embedded systems, and VLSI-CAD technology. He serves as Pacific Southwest regional director of American Society of Engineering Education. He is an entrepreneur with over 20 years of domestic an international experience and professional skills in both industry and academia. Dr. Raeisi may be reached at rraeisi@csufresno.edu

Mr. Vidya sagar reddy Gopala P.E., California State University, Fresno

Vidya sagar reddy Gopala received the B.E. in Electronics and Communication from Visvesvaraya Technological University of India (2015). He is currently perusing M.S. in Computer Engineering at California State University,Fresno. He works as teaching and Graduate Assistant in the Department of Electrical and Computer Engineering at California State University, Fresno. His research interests include NOC, VLSI design, system testing, testable design and verification.

A survey of Hybrid architectural approaches for optimizing power and performance of GPGPU

Dr. Reza Raeisi, Vidya Sagar Reddy Gopala, Electrical and Computer Engineering Department,
California State University, Fresno, CA 93740, USA.

Abstract:

Recent trends of drastic improvement speed of the processors have led to application of General Purpose Computing on Graphics Processing Units (GPGPU). We intend to present an educational study of two different hybrid architectural approaches for designing global memories of GPGPUs using different hybridization techniques. We address a hybrid memory organization and design of GPGPU considering emerging memory technology such as Spin Transfer Torque Random Access Memory (STT-RAM), Resistive Random Access Memory (RRAM) and Phase Change Memory (PCM) with conventional Dynamic Random Access (DRAM) memory. PCM memory technology provides inexpensive, fast access time, and high density nonvolatile storage. STT-RAM and RRAM are the new energy-efficient memory alternate for Dynamic and Static Random Access Memory technologies. The General purpose GPU's global memory consumes a considerable amount of total power of GPGPU. Using hybridization memory organization technique, the leakage power can be scaled to a minimal amount, such that the power and performance can be optimized. We do present the survey on two-hybrid memory architecture approaches that use PCM and combination of RRAM, STT-RAM along with the conventional DRAM memory. We discuss two main issues that designers face while designing hybrid memory architecture approach. A simple flowchart technique will be used to illustrate memory usage behavior and data migration issues that are seen while hybridizing the Global memory of GPGPU. Various memory access patterns will also be discussed in detail. Our primary purpose is to present and share our learning experience obtained through survey with others by highlighting similarities and differences between memory technology, and the current emerging memory technology in GPGPU architecture based on parameters and characteristics.

I. Introduction

In recent times, with the drastic improvement of processors, GPUs have been progressively utilized as general purpose GPU (GPGPU). They are used to improve the performance of many applications such as multimedia [16], EDA [17], numeric algorithm [18]. This paper is written to motivate and inspire engineering students in taking up projects in this particular domain. This domain of study is at the research level in many universities and thus there is very limited scope for teaching in class. However, as the topic is booming in the market there is always a scope for doing wide range of study or projects in this domain. We would like to share our survey on this domain as it may be a guide or motivation for many engineering students. The increasing computational power of GPGPU's makes it a solution for high performance processes, where thousands of threads are executed in parallel by their high computational power and programmability [1]. The power consumption of the processor increases drastically in order to accommodate a high throughput demands [2]. Thus, reducing the consumption of power is a crucial challenge for next-generation GPGPU systems [3]. Research has shown that the GPGPU with DRAM as global memory expends around 20%-40% of the aggregate power consumption of GPGPU [4]. And leakage power makes up to 70% of total memory power utilization [3].

Currently the DRAM based GPGPU's adopt many techniques to reduce the consumption of power and to increase the computational power, but it reduces the bandwidth [11]. Thus, there is a trade of between bandwidth and consumption of power. Therefore, a promising solution for this particular solution is using Non Volatile memory (NVM) techniques such as Phase change memories (PCM) [5], Spin Torque Transfer memory (STT RAM) and resistive memory (RRAM) instead of DRAM [12]. Due to Non-volatility NVMs' do not require a constant refresh operation, thus near zero standby power but few of the main drawbacks of NVMs are their long write latency, high write energy, and endurance problem. The long write latency will affect the memory bandwidth, which is one of the critical resources of a GPU. However, DRAM has its own advantage of low latency and can sustain unlimited writes. Thus, hybridization technique is used where NVMs are used along with the DRAM to get the advantage of both low latency and low power consumption. Hybrid memory systems have a high potential to overcome the power issues related to DRAM based GPGPUs.

In this paper, two different techniques of hybridization using different NVMs has been studied. Our study reviews the key issues which arise while hybridizing memory. The techniques, working and results of the both techniques used are compared. The hybrid design has two key concerns to be looked into: first, how to decide the memory capacity between DRAM and NVM; second, how to migrate the data between the available memories. These two issues will be discussed, then the two techniques of hybridization will be discussed in detail. In the first technique, PCM is the non-volatile memory used along with the DRAM. In the second technique, STT-RAM and RRAM, both are used in combination as non-volatile memories along with the DRAM. However, these cannot replace the DRAM as they consume higher dynamic power and have longer latency. Thus, the hybrid technology takes the advantages of both the memories.

II. Hybridization using PCM as NVM

The research on hybridization of memory shows that maximum part of the power consumption in a GPGPU is due to leakage current and power dissipation in standby mode. As stated earlier, the key issues seen while hybridizing: deciding the memory capacity between memories and migration of data between memories. To solve the first issue, GPGPUs benchmarks memory usage is characterized. Based on the execution time, the execution of benchmark is equally divided into small intervals. Survey says that memory size used by each interval is very small compared to large global memory of GPGPU [9]. Thus, a small DRAM is used for executing short intervals of benchmark. Therefore, the capacity of DRAM is reduced heavily in the hybrid technology and major portion of memory is replaced with PCM. For the second issue, an effective algorithm is used to analyze data criticality. The data which are more frequently accessed and accessed earlier are more critical, thus these data are stored in DRAM. A memory controller which is employed will update the position and data criticality of the data placed.

2.1 Characteristics of PCM

Phase change memory uses a Phase changing material called GST. GST is an alloy of germanium, tellurium and antimony. There are two phases of GST amorphous and crystalline. When the alloy is heated up to high temperature and quickly cooled down, then it takes up an amorphous form. If its heated up to temperature between crystallization and melting point and cooled down slowly, it crystallizes. Amorphous phase has higher electrical resistivity than crystalline form. This resistive difference is used to represent the binary digits such as logic ‘0’ and ‘1’ [5].

The two main challenges of PCM: limited write endurance and high write latency. The write endurances of PCM is only near 10^8 writes [19- 20]. PCM if used as a cache can lead to failure in less than an hour. Thus, it is suitable for memories such as main memories and lower level cache. The access time of PCM is four times longer than DRAM [8]. Due to the high density, the capacity of PCM based memory is larger than DRAM based memory. The dynamic power consumption of DRAM is much lesser than PCM. Thus, replacing DRAM with PCM would increase the system power consumption. PCM also has endurance issue, thus limited number of writes whereas DRAM has an unlimited number. It is likely to sustain (10^8) ~ (10^9) writes whereas DRAM can sustain unlimited writes [5]. Thus, Hybridization of memory is the solution to gain the advantage of both the memory technologies and avoid their disadvantages. The characteristics of PCM memory is tabulated in table 1 [21-28]. In the table shown, F denotes the smallest lithographic unit in a particular given dimension.

	PCM
Cell size (F^2)	120-200
Write Endurance	10^{16}
Speed (R/W)	Very fast
Leakage Power	High
Dynamic Energy (R/W)	Low
Retention Period	N/A

Table. 1. Characteristics of PCM.

2.2 Memory usage Behavior

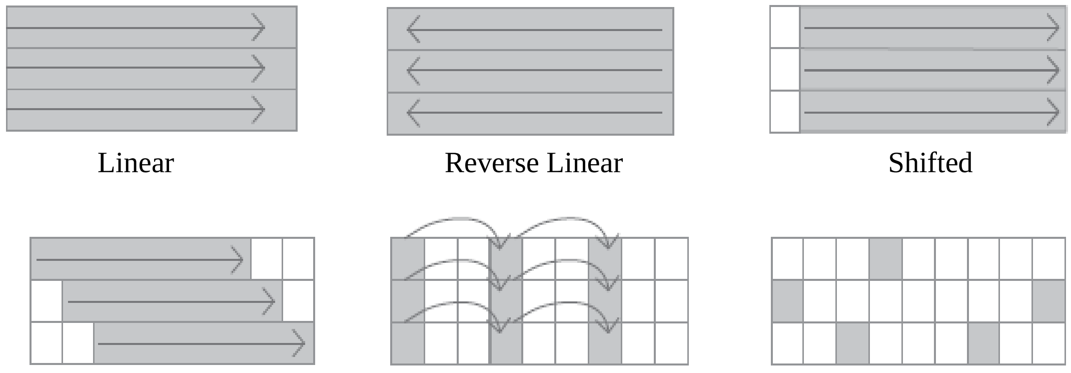
To characterize memory usage behavior, a wide set of typical benchmarks of GPGPUs are analyzed [9]. The execution of the benchmark in terms of execution time is divided into sequence of continuous intervals to check how much of memory is required for the execution. Research shows that small memory is required when compared to large-capacity global memory. Thus, small DRAM is enough to store data within the interval. The small interval is used to analyze the maximum of memory access space of all intervals which is abbreviated as MMASI. Table 2 shows the tabulated result of the analysis done by Kai Chen in his work [9]. One can observe that maximum memory space used by any benchmark is maximum of 166KB, which is very small compared to the 4GB or more capacity of the global memory. The mean or the average MMASI of all the benchmarks is 80KB. Table 2 indicates that only a small fraction of the whole memory is in use at a particular execution time. Thus most part of the memory is idle most of the time but still consuming high power. Thus, this idle memory can be of PCM which has almost zero standby leakage power. Thus, the major part of the memory comprises of PCM which decreases the leakage current significantly [6]. Secondly, for the migration of data between DRAM and PCM, a concept of data criticality is discussed as a solution for this issue.

Benchmark	**Memory space (KB)**
BFS	40
BLK	140
FWT	95
LIB	130
LPS	89
MM	32
MUM	166
NE	58
NN	10
NW	25
RAY	91
MEAN(avg)	80

Table. 2. Maximum of memory access space of all intervals of every benchmark [9].

2.3 Data Migration mechanism

The research shows that after deciding the capacity of the PCM to be used the next important factor is the migration of data between the memories and determining data position in the hybrid memory. It is evaluated by two factors namely First accessed time (FAT) and accessed amount (AA). On considering two-dimensional arrays, based on the type of accessing memory, memory access patterns are broadly classified into six types by Jang et al. [10]. They are linear, reverse linear, shifted, overlapped, non-unit stride and random. Other complex patterns are formed by clubbing any of these patterns. There are certain formulas for calculating the amount of memory accessed. The memory access patterns are shown in the below figure 1. And their related formulas for calculating amount of memory accessed are tabulated in the table 3.



Linear

Reverse Linear

Shifted

Overlapping

Non-unit Stride

Random

Figure. 1. Different memory access patterns [10].

From the above figure, one can see that memory access can be broadly classified into six types. All the formulas given in table are for two-dimensional array access. For multi-dimensional similar patterns can be used.

Pattern	**Formula**
Linear	AA = M x N
Reverse Linear	AA = M x N
Shifted	AA = M x (N-C)
Overlapping	AA = M x (N-C)
Non-unit stride	AA = M x (N/C)
Random	-

Table. 3. Formulas for different memory access patterns [10].

The memory controller initializes data positions based on data criticality. Random is the least accessed type, thus its formula can be neglected. There are two algorithms that help in data criticality. Firstly, initialization algorithm; which initializes the position of an array and then data exchange algorithm is used in migration of data. Initialization algorithm is represented as a flowchart in the figure 2. As the memory controller keep updating the array access amount, at the same time, it will periodically check for the possibility of memory exchange between different memories. The second algorithm, that is the data exchange algorithm; which helps in data migration between memories is depicted using a flowchart as shown in figure 3. Our survey showcases the algorithm suggested by Kai, Zhibin, Chengzhoung, Jin Xiaoke in their work [9].

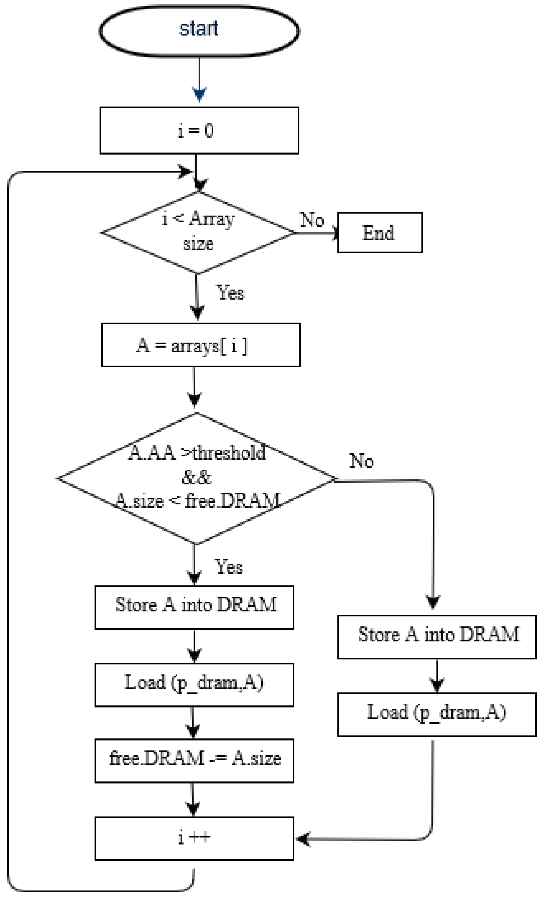


Figure. 2. Flow chart for initialization algorithm [9].

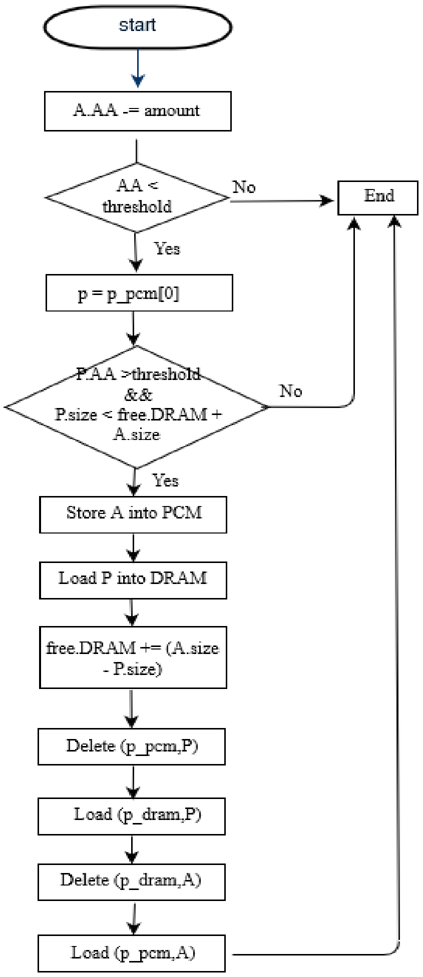


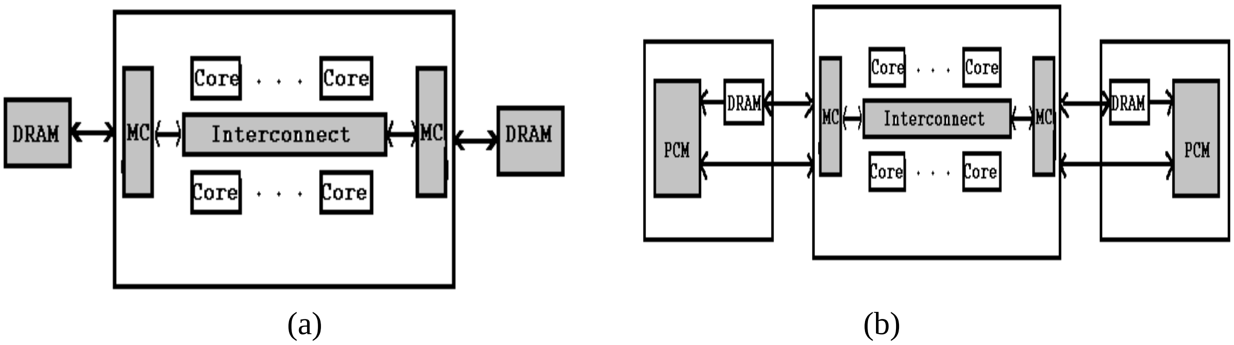
Figure. 3. Flow chart for data migration algorithm [9].

2.4 Hybridization technique

Hybridization of GPGPU memory is with two main memories namely DRAM and PCM. Based on the analysis of data criticality, the data migrate between two memories. The data with high criticality are stored in DRAM and the data with low criticality are stored in PCM. Thus increasing the power and performance of the GPGPU system. Even the high critical data which is accessed later are loaded into PCM and later swapped into DRAM when needed. Therefore, in hybrid memory, the capacity of DRAM is reduced to one tenth of size of global memory and the rest of the memory is made by PCM. The reason is for fixing the capacity at least to one-tenth of capacity to DRAM is to reduce the number of data exchange and this can prolong the lifetime of PCM. As shown in figure 4, GPU access both DRAM and PCM directly. Based on the two algorithms discussed in data criticality, the data are initialized and migrated between the memories.

2.5 Analysis and results

Here the experimental result which were conducted by Kai Chen [9] is discussed, and the results are analyzed. The impact of hybrid memory on performance and power consumption of GPGPU with baseline DRAM memory are discussed. Using 19 typical benchmarks of GPGPU, he has compared the power consumption and performance of both the models. As the DRAM size is decreased, the access time increases. Thus, improving the hybrid memory performance. Based on the research, PCM based hybrid memory saves more than 43% and 15% of power consumption of memory and whole system respectively [9]. By observing the results, main reason for reduction in power consumption is less leakage power of PCM. This paper mainly discusses about two major algorithms, which are needed for hybridization. Firstly, how to split memory capacity between DRAM and PCM. Secondly, data criticality and data migration. And thus, discussed the power consumption of memory using experimental data.



(a) (b)

Figure. 4. (a) Conventional GPGPU with DRAM as global memory. (b) GPGPU with PCM as hybrid memory.

III. Hybridization using STT-RAM and RRAM as NVM

The research on hybridization of memory shows that hybrid memory system with global memories as STT-RAM and RRAM along with DRAM to provide greater bandwidth, lower latency and optimal power consumption [16]. According to our study, only a fraction of memory is frequently accessed during run time. Thus, infrequently accessed can be stored in NVMs which are managed in stand-by mode with near zero power consumption. As shown in the below figure, leakage power of NVM's are lesser than DRAM. It has been reported that the GPU with DRAM as global memory's power consumption is growing linearly with bandwidth [13]. As in the previous technique, frequently accessed data are stored in DRAM and rest in NVMs. Thus the bandwidth will not reduce. According to the analysis done by Zhao, J., & Xie, Y in November 2012, the read and write latency of DRAM, STT-RAM and RRAM are as shown in figure 5 [14].

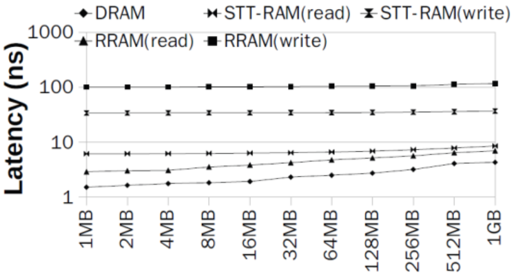


Figure. 5. Latencies of DRAM, STT-RAM and RRAM at different memory capacities [14]

Concept of memory access patterns is discussed in order to explain data migration. Although NVM has high latency than DRAM, proper study of memory access patterns can help in reducing such latencies. The adaptive data migration technique which helps in performance improvement of GPGPU is also discussed below.

Graphics Double Data Rate (GDDR) memories are similar to DDR memories. They are designed for high-performance systems, graphic cards and gaming consoles. Similar to DRAM, GDDR has set of memory banks. Each bank consists of rows and columns of memory cells. Memory cells are connected to sense amplifiers, which senses the bit ‘0’ or ‘1’. They latch the value in row buffer, so that they are available for subsequent access. To provide very high rate of performance GDDR employ high frequencies in order to gain high bandwidth. Thus, the overall power also increases. Therefore, the bandwidth is directly proportional to the power consumption of the system. If tried to reduce power consumption, even bandwidth is reduced [11]. Based on the survey, the characteristics of STT-RAM and RRAM are tabulated as below. Table 4 below displays the characteristics of STT-RAM and RRAM [21-33]. F denotes the smallest lithographic unit in a particular given dimension.

	STT-RAM	RRAM
Cell size (F^2)	6-50	4-10
Endurance	10^{15}	10^5 - 10^{10}
Speed (R/W)	Fast/slow	Fast/slow
Leakage Power	Low	Low
Dynamic Energy (R/W)	Low/ High	Low/ High
Retention Period	N/A	N/A

Table. 4. Characteristics of STT-RAM and RRAM.

3.1 STT-RAM and RRAM technology

Spin Torque Transfer memory (STT RAM) uses a Magnetic tunnel junction (MTJ) as its memory storage. Using this property of MJT, binary value is stored in an STT-RAM cell. Even though STT-RAM has lower density than RRAM, it has been widely used as its endurance rate is high. The endurance of STT-RAM is over 10^{15} which are feasible for hybridization. From figure 5, one can notice that STT-RAM has lower write latency than RRAM across all capacities. Thus it is a better solution to use STT-RAM to store the data which is write only in order to gain maximum benefits.

A resistive memory (RRAM) uses an insulating dielectric. RRAM has high density compared to SRAM, and a smaller leakage current. The drawback of RRAM is its low write endurance of 10^5 – 10^{10}. As seen from the figure 5, the read latency of RRAM is much smaller than STT-RAM and it is comparable with DRAM across all capacities. Thus it is a better solution to use RRAM to store the data which is read only in order to gain maximum benefits.

3.2 Memory access patterns

In the previous technique using PCM as NVM, six types of memory access patterns are discussed. Now, here three types of memory access patterns are discussed based on idle time between the different data accesses of memory. The three access patterns are namely “interleaved access”, “access then idle” and “burst”.

Interleaved access: From the sources [14], figure 6 is taken to demonstrate the memory access pattern called “Interleaved access”. As shown in the figure 6, the memory accesses are sorted based on ascending order of DRAM row addresses. The x-axis represents the index of the memory access. In the first row of the figure y-axis is a row that is accessed by each memory access and in the second row, y-axis is a cycle of each memory access. From the figure for analyzing an example of 550000th memory location access can be taken, which is from the row 2051 at cycle 40000. From keen observation of the figure 6, row 2051 is accessed during the execution of whole application. However, one can observe that idle periods are twice as long as the time taken for accessing the row. From the source [14], the 5/6 part of memory is read only. As RRAM has very small read latency, thus read only memory can be stored in RRAM. Therefore, powering it off during the idle period can reduce the memory power.

Access then idle: Figure 7 shows the “access then idle” pattern. The memory locations between 2300 and 2320 are accessed during the initial time till cycle 1,800,000. After those many cycles, that particular part of memory becomes idle [14]. Thus, the power of that particular part of memory can be turned off which is impossible with DRAM based memory.

Burst: The frequently accessed pattern is represented by “burst” pattern. From the figure 7 it is observed that row 2300 is in burst mode until cycle 2400000 [14]. This part of the data needs to be stored in DRAM as its frequently accessed.

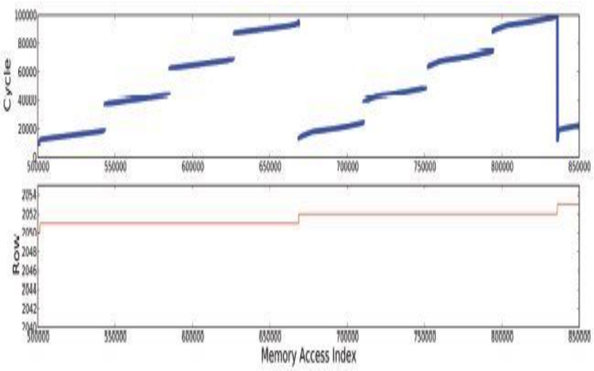


Figure. 6. Pattern of interleaved access [14]

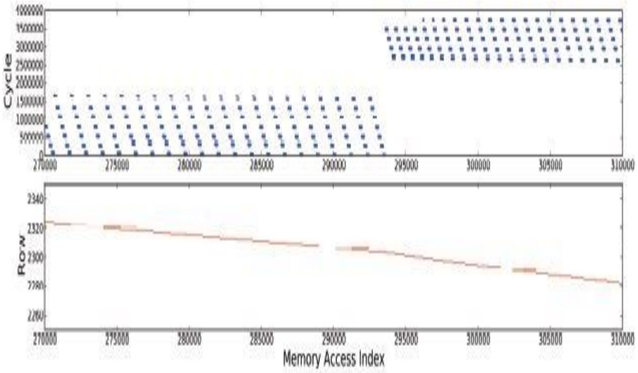


Figure. 7. Pattern of “access then idle” and “burst mode” [14]

3.3 Data Migration mechanism

Data migration between the main memory and the NVMs are based on a particular algorithm which is implemented in the memory controller. The main motto of the hybridization technique is to reduce the consumption of power by memory. Thus in order to reduce the power, the idle part of data is stored in the NVMs in standby mode and actively accessed data in DRAM. Data migration can take place based on the memory access patterns which are discussed in the earlier section. The data migration in “access then idle” is straightforward, but in “interleaved access” start point of data migration should be determined carefully. Data migration mechanism is depicted in the figure 8 and 9. The “Access then idle” is straight forward but in “Interleaved access” starting point needs to be determined.

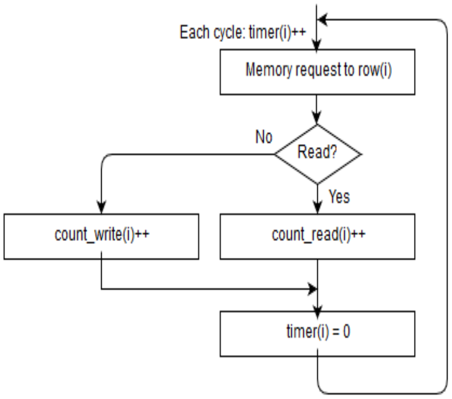


Figure. 8. Loop of DRAM access management [14]

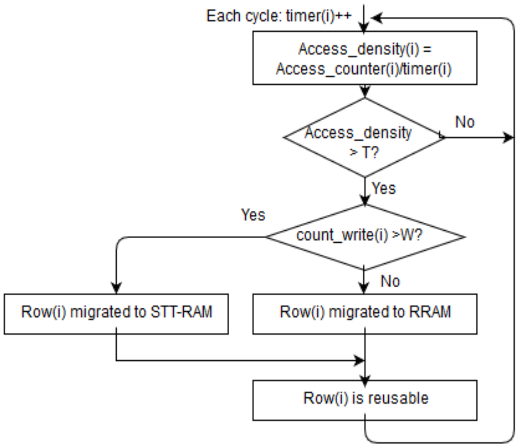


Figure. 9. Data migration algorithm [14]

3.4 Hybridization technique

GPGPU memory is hybridized with two NVMs namely STT-RAM and RRAM along with the main memory DRAM. Based on the analysis of data criticality, the data is migrated between DRAM and NVM memories. The data with high criticality are stored in DRAM which is shown in burst access mode and the data with low criticality are stored in NVMs. Again, the data is migrated between STT-RAM and RRAM based on the type of access, that is read only data is migrated to RRAM and write only data is migrated to STT-RAM. Thus increasing the power and performance of the GPGPU system. Figure 10 projects the overview of hybrid memory technology. Half of the memory is replaced with RRAM and STT-RAM. By reducing the size or the capacity of the DRAM to half of its size, its performance is improved by 25% in memory bandwidth factor. Due to non-volatility characteristics of the RRAM and STT-RAM the memory power consumption can be significantly reduced.

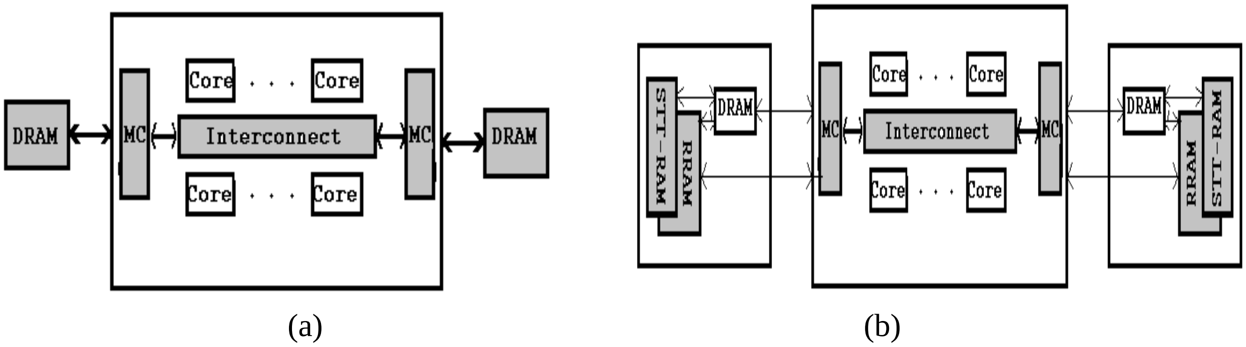


Figure. 10. (a) Conventional GPGPU with DRAM as global memory. (b) GPGPU with STT-RAM and RRAM as hybrid memory.

3.5 Analysis and results

The experimental results which were conducted by Zhao and Xie [14] is discussed and analyzed. The impact of hybrid memory on performance and power consumption of GPGPU with baseline DRAM memory is studied. Using benchmarks of GPGPU, they have compared the power consumption and performance of both the models. As the DRAM size is decreased, the access time increases. Thus, improving the hybrid memory performance. Based on the research and analysis done by Zhao and Xie [14], STT-RAM and RRAM based hybrid memory saves more than 31% and 16% of power consumption of memory and whole system respectively. By observing the results, main reason for reduction in power consumption is less leakage power of NVMs used. In this paper, two major algorithms which are needed for hybridization are reviewed in detail. Firstly, loop of DRAM access management. Secondly, the data migration algorithm which migrates data between STT-RAM and RRAM, in such a way that read only data are migrated to RRAM and write only data are migrated to STT-RAM.

IV. Conclusion

Our research investigation shows that widely used global memory of GPGPUs such as DRAM has various desirable properties such as low latency and high endurance, on the other hand, it also has undesirable properties of leakage power. Thus to overcome the disadvantages and retain the advantages various hybrid architectures are to be used. In this paper, a survey or the learning experience of two hybrid architectural techniques for addressing such issues are presented. An idea over usage of NVMs, algorithms for data migration mechanism, hybridization techniques are discussed. It appears that these emerging hybrid architectural techniques will be effective on the future engineers and this paper will stand as a guide for students in their study or research in the hybrid memory domain. We hope this study will motivate young students towards research, which in turn boosts the technology and benefits the society.

V. Bibliography

[1] Goswami, N., Cao, B., & Li, T. (2013, February). Power-performance co-optimization of throughput core architecture using resistive memory. In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on* (pp. 342-353). IEEE.

[2] Hong, S., & Kim, H. (2010, June). An integrated GPU power and performance model. In *ACM SIGARCH Computer Architecture News* (Vol. 38, No. 3, pp. 280-289). ACM.

[3] Zhao, J., & Xie, Y. (2012, November). Optimizing bandwidth and power of graphics memory with hybrid memory technologies and adaptive data migration. In *Proceedings of the International Conference on Computer-Aided Design* (pp. 81-87). ACM.

[4] GPGPU-Sim. GPUWatch
http://www.gpgpu-sim.org/gpuwattch/

[5] Zhou, P., Zhao, B., Yang, J., & Zhang, Y. (2009, June). A durable and energy efficient main memory using phase change memory technology. In *ACM SIGARCH computer architecture news* (Vol. 37, No. 3, pp. 14-23). ACM.

[6] Xie, Y. (2011). Modeling, architecture, and applications for emerging memory technologies. *IEEE Design & Test of Computers*, (1), 44-51.

[7] Raoux, S., Burr, G. W., Breitwisch, M. J., Rettner, C. T., Chen, Y. C., Shelby, R. M., ... & Lam, C. H. (2008). Phase-change random access memory: A scalable technology. *IBM Journal of Research and Development*, *52*(4.5), 465-479.

[8] Qureshi, M. K., Srinivasan, V., & Rivers, J. A. (2009). Scalable high performance main memory system using phase-change memory technology. *ACM SIGARCH Computer Architecture News*, *37*(3), 24-33.

[9] Chen, K., Yu, Z., Xu, C., Liu, J., & Li, X. (2014, April). Improving power efficiency of GPGPU's global memory by a hybrid memory approach. In *Information Science and Technology (ICIST), 2014 4th IEEE International Conference on* (pp. 660-664). IEEE.

[10] Jang, B., Schaa, D., Mistry, P., & Kaeli, D. (2011). Exploiting memory access patterns to improve memory performance in data-parallel architectures. *Parallel and Distributed Systems, IEEE Transactions on*, *22*(1), 105-118.

[11] Samsung. Green memory solution. *http://www.samsung.com/global/business/semiconductor*

[12] Raoux, S., Burr, G. W., Breitwisch, M. J., Rettner, C. T., Chen, Y. C., Shelby, R. M., ... & Lam, C. H. (2008). Phase-change random access memory: A scalable technology. *IBM Journal of Research and Development*, *52*(4.5), 465-479.

[13] Xu, C., Dong, X., Jouppi, N. P., & Xie, Y. (2011, March). Design implications of memristor-based RRAM cross-point structures. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011* (pp. 1-6). IEEE.

[14] Zhao, J., & Xie, Y. (2012, November). Optimizing bandwidth and power of graphics memory with hybrid memory technologies and adaptive data migration. In *Proceedings of the International Conference on Computer-Aided Design* (pp. 81-87). ACM.

[15] Dong, X., Xie, Y., Muralimanohar, N., & Jouppi, N. P. (2011). Hybrid checkpointing using emerging nonvolatile memories for future exascale systems. *ACM Transactions on Architecture and Code Optimization (TACO)*, *8*(2), 6.

[16] Liang, Y., Cui, Z., Zhao, S., Rupnow, K., Zhang, Y., Jones, D. L., & Chen, D. (2012, March). Real-time implementation and performance optimization of 3D sound localization on GPUs. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012* (pp. 832-835). IEEE.

[17] Chen, X., Ren, L., Wang, Y., & Yang, H. (2015). GPU-accelerated sparse LU factorization for circuit simulation with performance modeling. *Parallel and Distributed Systems, IEEE Transactions on*, *26*(3), 786-795.

[18] Sao, P., Vuduc, R., & Li, X. S. (2014). A distributed CPU-GPU sparse direct solver. In *Euro-Par 2014 Parallel Processing* (pp. 487-498). Springer International Publishing.

[19] Wang, J., Dong, X., Xie, Y., & Jouppi, N. P. (2013, February). i 2 WAP: Improving non-volatile cache lifetime by reducing inter-and intra-set write variations. In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on* (pp. 234-245). IEEE.

[20] Joo, Y., Niu, D., Dong, X., Sun, G., Chang, N., & Xie, Y. (2010, March). Energy-and endurance-aware design of phase change memory caches. In *Proceedings of the Conference on Design, Automation and Test in Europe* (pp. 136-141). European Design and Automation Association.

[21] Li, Y., Chen, Y., & Jones, A. K. (2012, July). A software approach for combating asymmetries of non-volatile memories. In *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design* (pp. 191-196). ACM.

[22] Syu, S. M., Shao, Y. H., & Lin, I. C. (2013, May). High-endurance hybrid cache design in CMP architecture with cache partitioning and access-aware policy. In *Proceedings of the 23rd ACM international conference on Great lakes symposium on VLSI* (pp. 19-24). ACM.

[23] Mittal, S. (2013). Energy saving techniques for phase change memory (PCM). *arXiv preprint arXiv:1309.3785*.

[24] Chen, Y. T., Cong, J., Huang, H., Liu, B., Liu, C., Potkonjak, M., & Reinman, G. (2012, March). Dynamically reconfigurable hybrid cache: An energy-efficient last-level cache design. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012* (pp. 45-50). IEEE.

[25] Zhao, J., Xu, C., & Xie, Y. (2011, November). Bandwidth-aware reconfigurable cache design with hybrid memory technologies. In *Proceedings of the International Conference on Computer-Aided Design* (pp. 48-55). IEEE Press.

[26] Dong, X., Wu, X., Sun, G., Xie, Y., Li, H., & Chen, Y. (2008, June). Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement. In *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE* (pp. 554-559). IEEE.

[27] Venkatesan, R., Kozhikkottu, V., Augustine, C., Raychowdhury, A., Roy, K., & Raghunathan, A. (2012, July). TapeCache: a high density, energy efficient cache based on domain wall memory. In *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design* (pp. 185-190). ACM.

[28] Mittal, S., Zhang, Z., & Cao, Y. (2013, January). CASHIER: A cache energy saving technique for QoS systems. In *VLSI Design and 2013 12th International Conference on Embedded Systems (VLSID), 2013 26th International Conference on* (pp. 43-48). IEEE.

[29] Chang, M. T., Rosenfeld, P., Lu, S. L., & Jacob, B. (2013, February). Technology comparison for large last-level caches (L 3 Cs): Low-leakage SRAM, low write-energy STT-RAM, and refresh-optimized eDRAM. In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on* (pp. 143-154). IEEE.

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

[31] Kryder, M. H., & Kim, C. S. (2009). After hard drives—what comes next?. *Magnetics, IEEE Transactions on*, *45*(10), 3406-3413.

[32] Wu, X., Li, J., Zhang, L., Speight, E., Rajamony, R., & Xie, Y. (2009, June). Hybrid cache architecture with disparate memory technologies. In *ACM SIGARCH computer architecture news* (Vol. 37, No. 3, pp. 34-45). ACM.

[33] Kim, K. H., Hyun Jo, S., Gaba, S., & Lu, W. (2010). Nanoscale resistive memory with intrinsic diode characteristics and long endurance. *Applied Physics Letters*, *96*(5), 053106.

[34] Hynix. GDDR5 SGRAM datasheet.
http://www.hynix.com/products/graphics/