JOURNAL OF SCIENCE
SAKARYA UNIVERSITY

Sakarya University Journal of Science

ISSN 1301-4048 | e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University |

http://www.saujs.sakarya.edu.tr/

Title: A Survey of Hybrid Main Memory Architectures

Authors: Zerrin Yıldız Çavdar, İsa Avcı, Murat Koca, Ahmet Sertbaş

Recieved: 2017-08-14 17:48:27

Revised: 2018-07-05 18:47:48

Accepted: 2018-07-11 15:06:57

Article Type: Research Article

Volume: 23

Issue: 1

Month: February

Year: 2019

Pages: 1-15

How to cite
Zerrin Yıldız Çavdar, İsa Avcı, Murat Koca, Ahmet Sertbaş; (2019), A Survey of Hybrid Main Memory Architectures. Sakarya University Journal of Science, 23(1), 1-15, DOI: 10.16984/saufenbilder.334645
Access link
http://www.saujs.sakarya.edu.tr/issue/38708/334645

Sakarya University Journal of Science 23(1), 1-15, 2019

A Survey of Hybrid Main Memory Architectures

Zerrin Yildiz Çavdar*[1], İsa Avci[2], Murat Koca[3], Ahmet Sertbaş[4]

ABSTRACT

Rapidly evolving technology, increased internet speed and capacity, and the widespread use of mobile technologies have increased the demands for faster applications and less power consumption of modern electronic systems. In modern electronic systems, RAM is as effective as CPU regarding performance and power consumption. Although DRAM is the most used types of main memory today, it has been insufficient in terms of provide increasing demands. One of the issues to be addressed is to improve DRAM in terms of performance and power consumption. Another study to address this increasing demand is the development of hybrid main memory architectures. Hybrid Main Memory is one of the most recent studies on RAM. In this research, we investigate hybrid main memory systems for a more efficient main memory architecture.

Keywords: Hybrid Main Memory, DRAM, Phase Change Memory/PCM, performance, energy saving

1. INTRODUCTION

With the rapid growth and widespread use of Internet speed and capacity, modern electronic systems that we use in many areas of our daily lives are expected to be faster. At the same time, the production of the mobile version of many devices has made the power consumption / battery life of the devices an important issue.

Performance and power consumption are two important factors in all modern electronic systems such as mobile phones, computers, smart home systems, and so on. It is desirable that the devices operate both very fast and with low power consumption (if mobile having long battery life). The biggest problem with developing electronic systems is that speed performance and power consumption have a negative effect on each other. In other words, while efforts to increase the performance of systems often cause to more power consumption, on the other hand, the work done on power saving negatively affects the performance of the systems.

In a standard electronic system, CPU and RAM are the two fundamental components most affecting power consumption and performance. In order to achieve higher performance with low power consumption design goal in modern electronic systems the CPU and RAM related software and hardware enhancements have become the main interest in commercial and academic work. The studies show that RAMs are

* Corresponding Author: zerrinyildizcavdar@sehir.edu.tr
[1] Istanbul City University, Vocational School, Computer Technologies Department, Computer Programming, Istanbul, Turkey
[2] Turkish Airlines, Information Processing Unit, Istanbul, Turkey
[3] Hakkari University, Vocational School of Health Services, Hakkari/Turkey
[4] Istanbul University, Faculty of Engineering, Department of Computer Engineering, Istanbul, Turkey

highly efficient at high power consumption, so that in modern server systems, main memory contains almost 30-40% of total power consumption. [1].

At the present, DRAM is the most widely used main memory type. However, it is now difficult to meet the demand for rapidly evolving technology and increased performance. DRAM uses almost 20% to 40% of the energy consumed by an existing server system [2] [3]. This necessitates the creation of alternative main memory architectures (hybrid for example) for DRAM development. In the current studies, many hybrid main memory architectures are developed by using different memory types together.

Newest two commercial works on improving main memory performance and power consumption are Hybrid Memory Cube and IBM PRAM projects. First one is produced in 2011 by Micron and Samsung, and based on the principle that DRAM main memories are joined vertically. This makes it possible to use DRAM at more capacities without taking up more space. Technical details related to the study are available in [4] and [5]. The second one is developed by IBM in 2016. IBM scientists have worked on PCM (Phase Change Memory) memory and have been able to increase the data storage capacity from 1 bits per cell to 3 bit per cell [6].

In recent years, PCM has become a preferred solution to be used in hybrid main memory modules together with DRAM. DRAM memories are more advantageous than PCM memory types in write operation. PCM memories are more advantageous than DRAM in reading and stand-by operations because they do not need to refresh. Therefore, the Hybrid Main Memory architecture therefore focuses on the solutions used in combination with DRAM and PCM memories.

In this survey study, we are conducting a research on improving RAM technologies, which are very effective in terms of power consumption and performance. In the study, we investigate the academic works on hybrid main memory systems in order to produce a more efficient main memory architecture. In the second section of our work, we briefly mention non-volatile main memory types. In the third section, we examine hybrid main memory architectures. We present architectures under the titles DRAM + PCM, DRAM + PRAM, DRAM + NVM, DRAM + Other Types. The fourth section of the paper conducts a performance review. The work in this section is presented under the titles of energy saving, performance analysis and endurance / lifetime. In these examinations, we have tried to compare techniques in different ways. Some of the studies have been addressed in more than one section, because they deal with multiple categories, such as energy saving and endurance. In the last section, we give conclusions of the work.

2. ELECTRONIC MEMORY TYPES

This section provides basic information about the types of memory that can be used as the main memory.

Memory types are electrically divided into Volatile (temporary) and Non-Volatile (permanent). We showed the memory types together in the following figure. The most common types of memory used in hybrid memory architectures are volatile DRAM and nonvolatile RAM. Also volatile DRAM is well known memory type, so we gave basic information only about Non-Volatile memories (2.1-2.4) below.

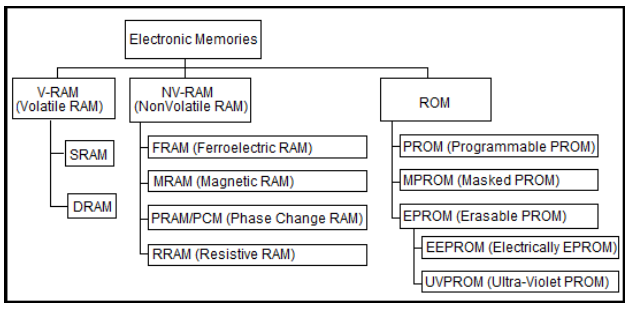

Figure 1: Electronic memory types

2.1. FRAM/FeRAM (Ferroelectric RAM)

FRAM is one of the memory types that are beginning to get attention in the hybrid main memory architectures, which is a low-voltage, non-volatile memory type with fast memory [7]. Ramtron International Corporation presented the first commercial FRAM in 1988 [8]. FRAM is an advanced non-volatile memory that marketed earlier than its counterparts like MRAM, PCRAM, ReRAM [9].

2.2. MRAM (Magnetic / Magnetoresistive RAM)

MRAM is a non-volatile RAM that stores information with electron spin. It works by turning on and off the magnetic moment, not by electricity to determine the write state. MRAM, which is highly ambitious against other RAM technologies due to its non-volatility and power efficiency, is a candidate for universal memory.

2.3. PRAM/PCM (Phase Change RAM)

PRAM is a type of phase change memory designed to protect the data even when power is lost. While reading data from the PRAM, the power consumption is much lower because no heating is required. There is also no energy requirement for PRAM refresh. While PCM is used in many sources to describe PRAM, phase change RAM, it is possible for PCM to serve outside of the main memory. Non-Volatile PCM uses the phase shifting property of chlorochemical glasses to store data (bit information) [10] [11].

2.4. RRAM (Resistive RAM)

The memory type, which is a type of a persistent / durable (nonvolatile) storage that functions via altering the exclusively designed solid non-conducting substance's impedance, is the resistive random access memory (RRAM). A RRAM embodies a memory resistor unit (memristor) of which shows a diverse level of resistance in accordance with the various voltages applied to it.

3. HYBRID MAIN MEMORY ARCHITECTURES

In this part, we examined the studies between the years 2000-2017 on the development of the hybrid main memory architecture. In Table 1, Group 1 and Group 2 are works using the same architecture (PRAM/PCM), listed them in different groups because of using the different terms in their architectural representations. In fact, both groups used Phase Change Memory. Group 3 and Group 4 are studies based on DRAM + NVM architecture, and DRAM + other memory types architectures respectively.

In the remainder of this section, we tried to examine some selected solutions, which include different hybrid main memory architecture models given in works listed on Table 1.

Table 1: Hybrid main memory architecture types

Group No	Suggested Hybrid Architectural Components	Studies Using This Architecture
1	DRAM + PCM	[12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]
2	DRAM + PRAM	[27], [28], [29], [30], [31], [32], [33], [34], [35], [36]
3	DRAM + NVM	[37], [38], [39], [40], [41]
4	DRAM+Other Memory types	[42], [43], [44], [45], [46], [47], [48]

3.1. DRAM + PRAM/PCM Architectures

In this subsection, we studied on main memory architectures with Phase Change Memory and DRAM. Firstly, an energy-efficient main memory architecture, named PDRAM, using DRAM and PRAM together is suggested in Fig.2. In order to address the challenges of managing such an energy efficient system, the researchers anticipated that a hybrid (hardware + software) solution could be used. Because of the write

endurance problem of PRAM, in this work, a hardware solution is used to store the page-level write frequency information. On the software side, they proposed a page managing program using the writing frequency from hardware at the operating system level [27].

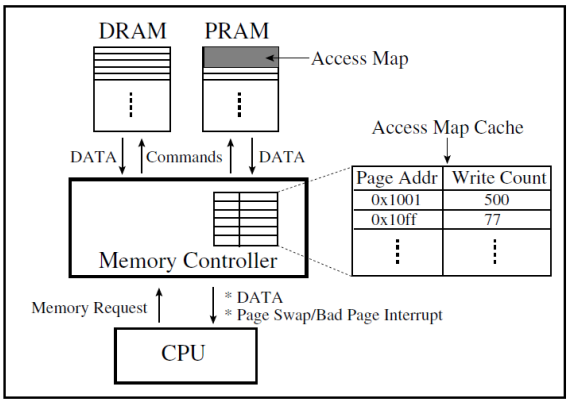

Figure 2: PDRAM diagram [27]

Unlike Dhiman and his friends' PDRAM named hardware and software based suggestion [27], Park and his colleagues implemented the memory management system only with a solution at the operating system level. In the study, the main memory management mechanism at OS (Operating System) level applied to DRAM+PRAM hybrid structure shown in Figure 3, the memory pages are divided into cold and hot and dynamically placed in the suitable memory areas.

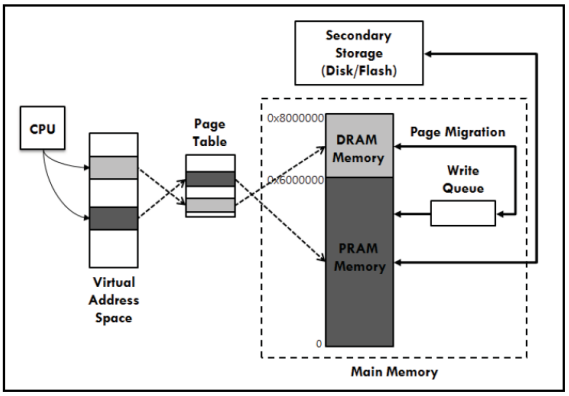

Figure 3: Hybrid main memory [28]

For this process, [28] used three methods called memory access monitoring, page switching and page assignment, cold pages are transferred to PRAM with page transfer, hot pages are allocated to DRAM. They aimed to reduce energy consumption by this separation and by a page deletion structure that allows the DRAM to close when DRAM is not used [28].

Lee and his colleagues proposed a DRAM + PCM memory design, which increases performance without increasing energy consumption [12]. In addition to examining DRAM + PCM designs considering both performance and energy efficiency (Fig.4), a DRAM + PCM architecture is proposed that combines the energy efficiency achieved using DRAM as a chip cache and the performance levels accomplished when DRAM is used as an buffer of PCM. The architects have studied these approaches in terms of energy consumption and performance in their study of DRAM as a write cache and DRAM as cache memory. In the study, unlike previous studies, which emphasized the management of PCM memory, the idea of using DRAM memory in different forms was examined.

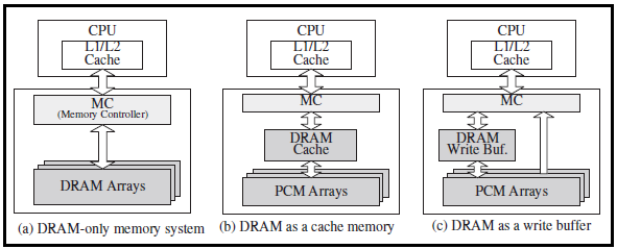

Figure 4: DRAM/PCM configuration [12]

In order to minimize the impact of DRAM's energy consumption, PRAM's write endurance and various limitations, [29] has worked on power control in DRAM + PRAM main memory, in Fig. 5.

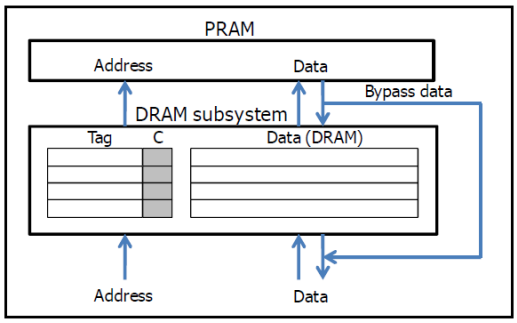

Figure 5: DRAM/PRAM memory [29]

[13] aimed enlarging the PCM's lifespan by reducing the writing number (Fig. 6), therefore they presented a caching scheme called CAR (Cache Address Remapping). In addition, RanCAR (Randomized CAR) also referred to a practical application. In this structure, RanCAR may reduce PCM write-back traffic by evenly distributing single cache's writes to different cache sets. Experiments with the M5 simulator have shown that CAR can decrease the ~ 4600x DRAM cache miss ratio under particular attack and the PCM's lifespan can be extended to 13.8 years, with negligible performance loss.

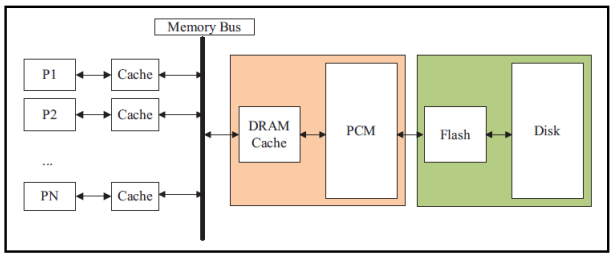

Figure 6: System architecture of DRAM/PCM [13]

An evaluation framework for improving performance in the DRAM + PRAM memory architecture, named the OPAMP is suggested in [30]. In such a structure that emphasizes that hybrid memory design must be done carefully, the frame aggregates the environmental parameters of the combined main memory and evaluates the best performance under relevant circumstances. Here the mentioned above study proposed a method of obtaining the most suitable value with the hybrid master memory profiling.

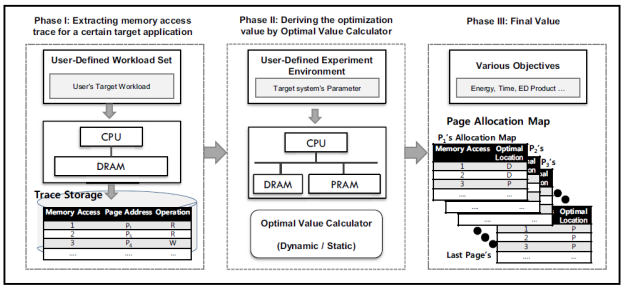

Figure 7: OPAMP process [30]

As an alternative to OPAMP process in [30], [16] proposed to design a new DRAM+PCM hybrid main memory architecture shown in Fig. 8 that takes the recently used metadata cache. Thus they have set a different caching policy for the combined memory, which actuates the transition size causing the low access latency and low pass.

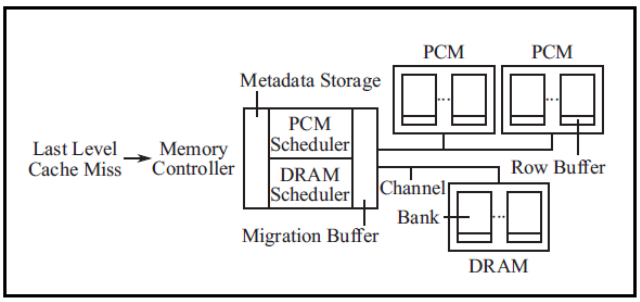

Figure 8: DRAM/PCM overall architecture [16]

Hu and his colleagues proposed SWL (Software Wear Leveling) algorithm and software techniques for correcting abrasions for prolonging the PCM life. They based on the architecture shown in Fig. 9 for their studies in the simulator environment. As a result of those studies, they could achieve a reduction about 80% of the number of writing by greedy algorithm and by ODA (Optical Data Allocation) algorithm a reduction of about 60% of memory access under 6% memory access overhead [17].

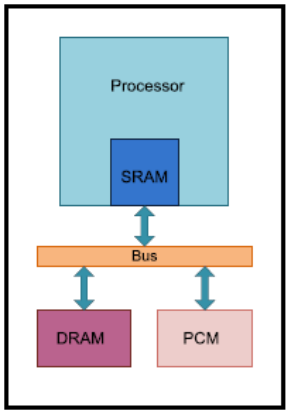

Figure 9: Hybrid memory architecture [17]

As similar to previous work focusing on software solutions, [31] examined the problem of task allocation. For to minimize energy consumption and memory size and extend the life of the hybrid memory, task allocation techniques have been studied in two stages, emphasizing PRAM's energy efficiency and DRAM write durability. Firstly, they designed ILP formulas to solve different objectives best. Then, they proposed 2

different heuristic algorithms, namely 3 polynomial time offline heuristics and 3 online heuristics. At the end of the studies, they conclude that the offline heuristics they offered performed better than the simple ones. Furthermore, when compared to ILP formulations, it has been shown that the offline exploratory methods offer similar solutions, but have lost much less time [31].

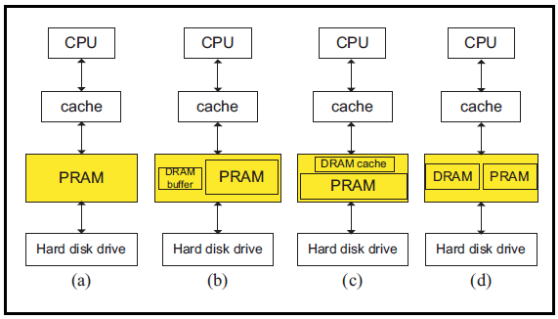

Figure 10: Steps from PRAM to hybrid DRAM + PRAM [31]

To increase the hybrid memory's reliability, Mao and his colleagues have set up three separate low cost ECC-based schemes in Fig. 11. In their work on a hybrid memory architecture [33], they used DRAM's PRAM buffer because it has some advantages like small standby power, speedy access time and big storage density. Fig.11(a) shows a conventional system in which only the PRAM's data is protected by ECC. For this reason, this system only runs if there are relatively low errors in DRAM. Systems 2 and 3 provide ECC protection for the PRAM as well as for the data in the DRAM cache (Figures 11(b) and 11(c)). In the second scheme, the same ECC unit protects the DRAM cache and the data in the PRAM, so the ECC scheme could be more powerful. System 3 uses other layouts for PRAM and DRAM so that 2 memories' error characteristics can be better identified [33].

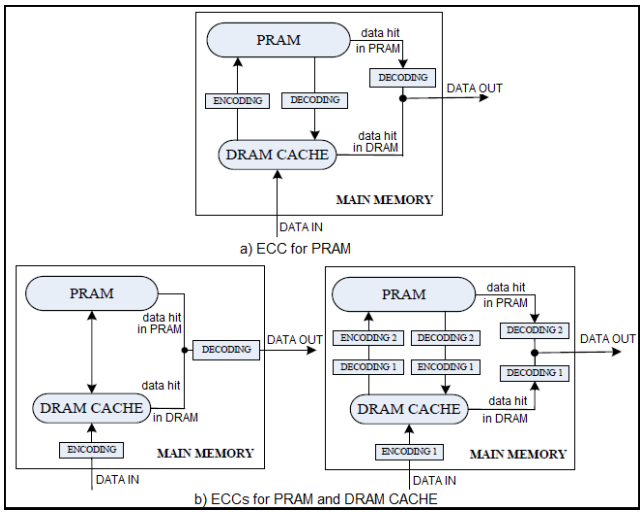

Figure 11: ECC for hybrid DRAM/PRAM [33]

For achieving high performance of DRAM + PRAM hybrid architecture, an architecture that can be accessed in parallel with memories (shown in Fig. 12) has been designed to support selective caching in DRAM to expand writing buffer. For resolving overmuch stall time the problem at the memory controller's data queue, they proposed caching data optionally, that incur the stall thereabout decreasing latency in memory access and ameliorating fairness. Their results presented that the DQSA (Data Queue Stall Aware) approach, achieves 21% better performance and 2.1 times melioration in fairness examined with the best of existing ways in a multi-care system consisting of collective GPU and CPUs [34].

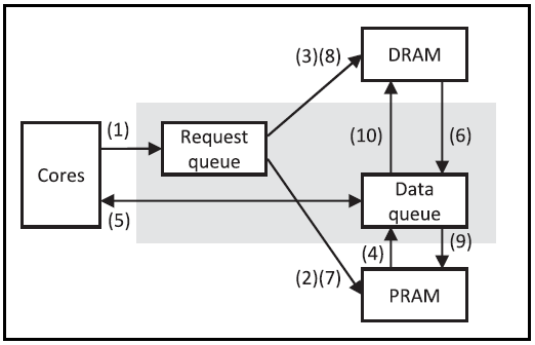

Figure 12: Hybrid main memory block diagram [34]

[35] introduced a main memory system based on PRAM including DRAM converter, CMT (Convert Management Table) and MLC(Multi Level Cell)/SLC(Single Level Cell) PRAM. This work showed that the DRAM converter, which

consisting AFSB(Aggressive Fetching Superblock Buffer) and SFB (Selective Filtering Buffer), could improve access latency and improves endurance. While the AFSB utilizes the regional location effectively / actively by bringing super blocks from the MLC PRAM, the SFB uses the temporary zone by adapting it. The CMT uses data classification management to ensure that pages are assigned and counting information is entered. The SLC PRAM takes on the task of extending the write buffer and the lifetime of the MLC PRAM to mask asymmetric write / read latencies [35].

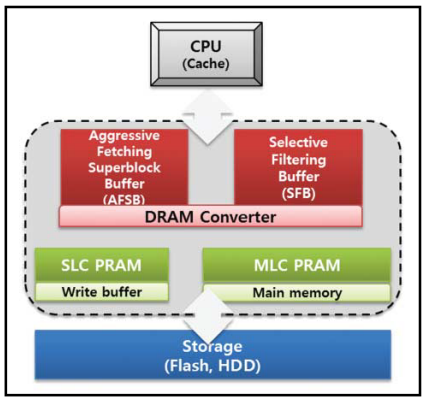

Figure 13: SLC/MLC PRAM + DRAM main memory [35]

For the sake of energy optimization of the PCM and DRAM combined main memory structure shown in Fig. 14, Wang and colleagues aimed to use the PCM completely to ensure the performance of real-time applications and reduce energy consumption [22]. This study proposed an optimal address-mapping algorithm for mapping an appropriate memory address for each address. First they calculated time cost and energy cost for every single address being founded on task types. Then they formulated an ILP model for the timeline issue on distinct memory types and given timing constraints. So they obtained an optimal solution. They concluded that the introduced approach could meaningly decrease energy consumption at the minimum cost when analyzed by the conventional techniques [22].

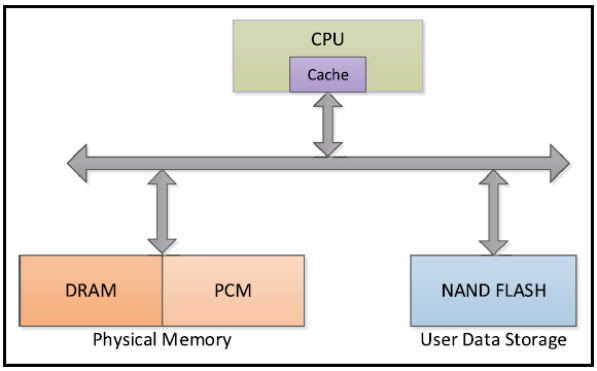

Figure 14: Hybrid DRAM/PCM memory in embedded system [22]

For real-time task-timeline problem for energy efficiency, [23] introduced static table-based and dynamic time lining algorithms for an extraordinary set of tasks. They also introduced 4 real-time programming algorithms based on RM (Rate-Monotonic) and EDF (Earliest Deadline First) timers for real-time embedded systems with a combined PCM-DRAM main memory (in Fig. 15) for a periodic task set. Finally, they have designed a dynamic-RM algorithm that takes advantage of the nearest idle time and a dynamic-EDF algorithm that recovers the entire available free time to enhance the results of static solutions. This architecture addresses DRAM and PCM consecutively in order to provide a direct CPU access to DRAM and PCM. At the end of this study, they presented the timing algorithms that reduce real time constraints and energy consumption [23].

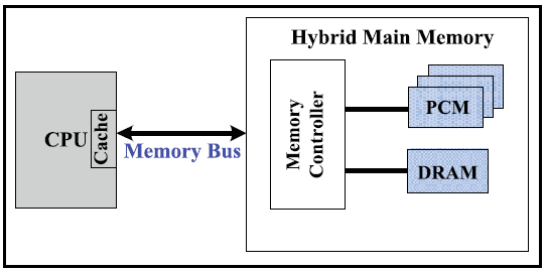

Figure 15: Hybrid DRAM/PCM memory [23]

In [36], Cai and his colleagues examined the task sharing issue for DRAM + PRAM combined main memory structure. To improve memory performance and decrease energy consumption of the memory subsystem shown in Fig.16, they assigned distinct memory devices for every task.

They have designed an ILP offline-ASA (Adaptive Space Allocation) algorithm to access ideal task distribution for an embedded system with a static periodic task set. Furthermore, they proposed an online-ASA algorithm for dynamic task set where accessions of tasks are unknown before [36].

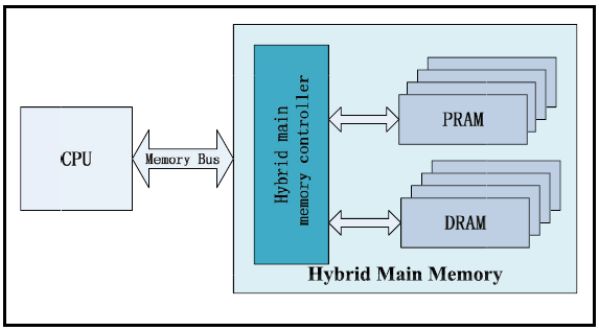

Figure 16: DRAM/PCM main memory [36]

The study in [25] described a multi-page concurrent transition (CMMP - Concurrent Migration of Multiple Pages), a new structure by using hardware and software to manage the construction shown in Fig. 17. In this work, it is indicated that CMMP carries multiple pages simultaneously independently the available memory bandwidth for programs significantly, brings a basic interface for the operating system to monitor memory access models. They also state that CMMP reduces the transfer bandwidth from PCM to DRAM by copying blocks at calls, decreases the bandwidth to the PCM from DRAM by preventing the blocks from being transferred back to the PCM [25].

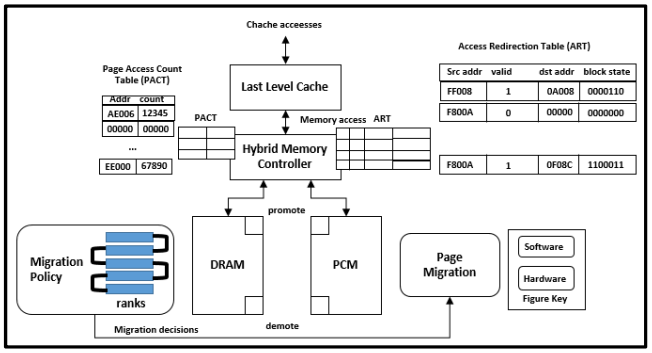

Figure 17: Overview of CMMP [25]

In [26] on power management of DRAM + PCM combined main memory (Fig. 18), it is aimed to decrease energy consumption of combined memory and increase system efficiency. A new page-rated energy consumption strategy and a new data structure road saved the pages as local and global access data, classified according to the entry dates, and then reshaped by deciding which memory (DRAM or PCM) will be located according to the memory structure. Experimental studies on APG (Adaptive Page Grouping) and PDRAM (DRAM + PCM) in the Gem5 simulator environment have shown that they could reduce energy consumption and improve performance with their strategy [26].

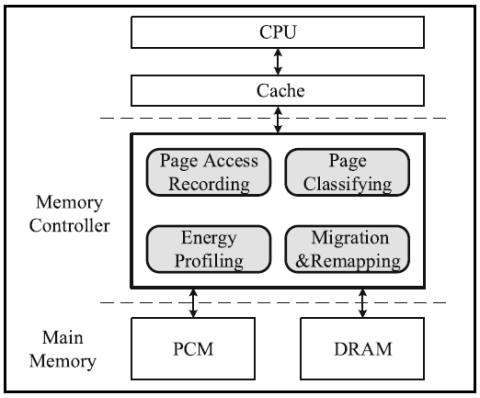

Figure 18: Hybrid DRAM/PCM architecture [26]

3.2. DRAM + NVM Architectures

In hybrid main memory systems, which are currently proposed as an ideal solution for main memory architecture, it is very important to balance the memory types used. Taking this balance into account, Knyaginin and colleagues in [37] proposed Crystal, an analytical approach to resource partitioning at design time for hierarchical hybrid systems based on DRAM + NVM, modeled in Fig. 19. Crystal, which is rapidly identifies the most important quantities and trends of NVM for certain workloads and segmentation targets and obtains targeted design points for detailed evaluation. In this work, it has showed how to achieve system-level performance and energy efficiency using NVM with speed and energy consumption of NAND Flash in place of faster and more energy-efficient NVM like

PCM. Hybrid design calls the data to M2, if a requested page is not in M1(if it is in M2). Each insertion (page entry from disk or a transition from M2) could be result in an evacuation sequence (transition to M2 or page output to disk) because of showing stationary state behavior of curve line. Sending disk only occurs dirty pages of non-memory programs [37].

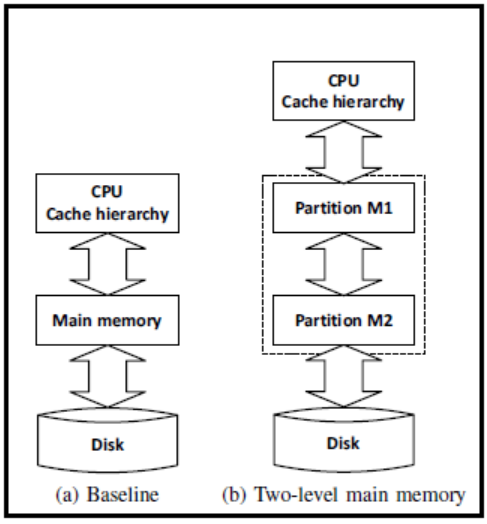

Figure 19: DRAM/NVM system [37]

In addition to the advantages of DRAM + NVM hybrid main memory, which is considered as an ideal next generation architecture, they have significant performance problems due to increased memory traffic, intensive data migration and lack of effective migration. For solving these problems in [40], it is developed a simulator named HMMSim shown in Fig. 20. In addition to introducing the properties of HMMSim, they observed that performance was improved with HMMSim in combined DRAM + NVM architecture in the probatory products. This work shows that hybrid main memory is a promising option with the right software support.

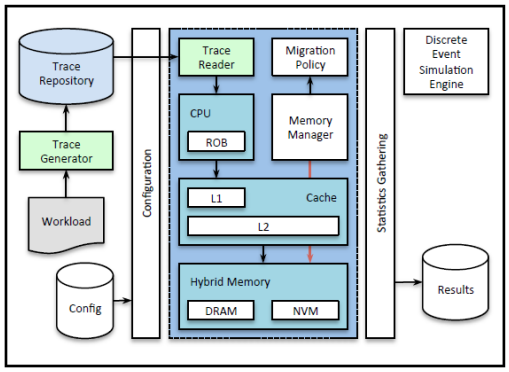

Figure 20: Overview of HMMSim [40]

In [41], Bock and his colleagues suggested the architecture in Fig. 21, which is composed of DRAM and NVM and is supported by the operating system as software. This architecture consists of one or more CPUs with special command and data L1 caching. Assume that requests from the CPU are tailed in L1 tails, all CPUs share an L2 cache, and a single L2 tail handles requests from all CPUs. In the study, working conditions such as application delay and page transition load of the hybrid memory managed by software were analyzed, and factors causing high cost in hybrid construction were determined. As a result of the study, it was seen that the main limiting factor was the delay in the NVM queue and better migration policies could be used to improve performance.

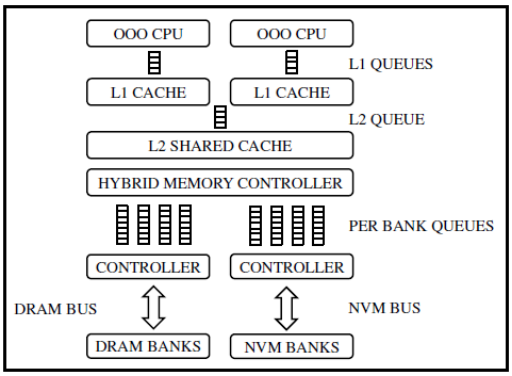

Figure 21: Limiting factors model [41]

3.3. DRAM + Other Memory Types

In regards of the hybrid main memory operations, although PRAM/PCM main memory is preferred well along with DRAM, there are some possible other hybrid solutions. In this section, we

examined other hybrid architectures related to these studies.

3.3.1. DRAM + SCM

In a review of the use of SCM (Storage Class Memory) memory in computer systems, in [46] Kwon suggested using a combination of DRAM and SCM memory (Fig. 22) for a combined memory system. Although SCM is one of the memory types having the optimal memory and storage capabilities, it is not widely used because it is only available in small capacity. He pointed out that the operating system have an considerable role in alleviating the imperfections of the SCM and in using the SCM as a working memory. The most common difficulty of a SCM+DRAM combined memory is to estimate the access order of the data locks, for placing the hot data in DRAM during the write operation, and to place the cold write data in the SCM. He noted that in combined memory systems with the SCM + DRAM, it is necessary to investigate how to fit the data between the two memory moments in order to mitigate the imperfections of the SCM and use the persistence of the SCM. He also emphasized that the delay problem for both DRAM and SCM.

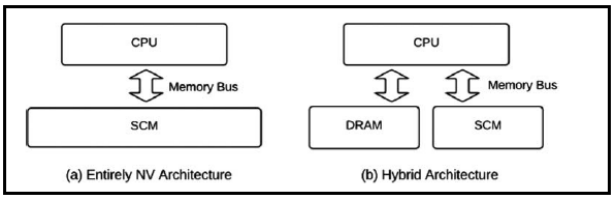

Figure 22: SCM architecture [46]

A study in [48] comparing the architecture of a big DRAM + SCM (small) given with the name of RAHMM (Retention-Aware Hybrid Main Memory) to BSLD (a Big SCM + a Little DRAM hybrid architecture) was performed. In this work, Jing and his colleagues used a small SCM to change the DRAM tail section's data, thereby providing for less refreshing of the DRAM and thus less energy consumption. They have also proposed a HBS (Hidden Buffer Strategy) in order to improve the writing performance and solve endurance problem.

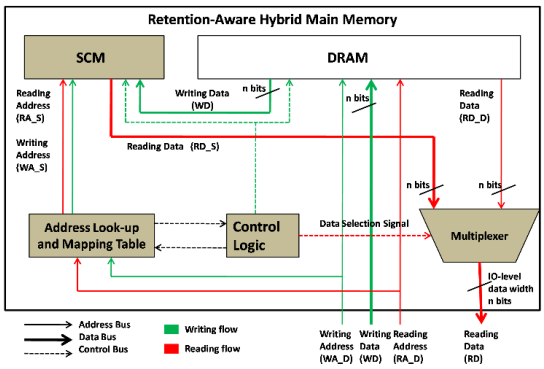

Figure 23: RAHMM architecture [48]

3.3.2. DRAM + PM

In [47], the researchers investigated the influence of hardware stationed page replacement on a combined main memory. The guided model, fixed page swap activity was evaluated by using the suggestions model shown in Fig. 24.

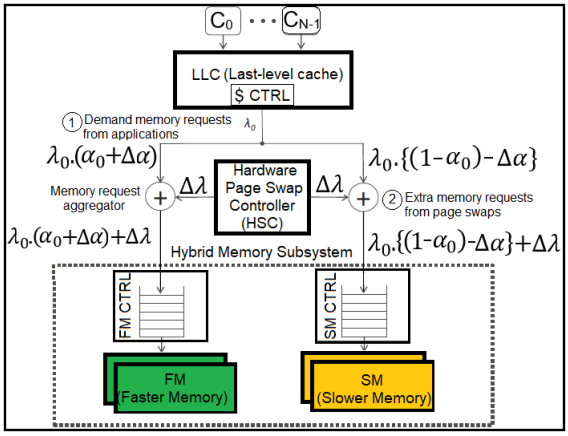

Figure 24: DRAM+PM in hybrid structure [47]

3.3.3. Other Main Memory Architectures

In [42], for performance modeling and analysis of new computer architects, an analysis by using a VMM (Virtual Machine Monitor) was performed for combined main memories includes DRAM + other memory type (DRAM + NAND, ...) shown in Fig. 25. In the study, the hybrid structure mainly consists of reinforced DRAM and a slower second level memory. In [43] using a customized VMM for performance analysis of hybrid main memory, it has been observed that only certain workloads are more suitable for these structures.

The study confirmed that performance evaluation with VMM is a valid and useful technique.

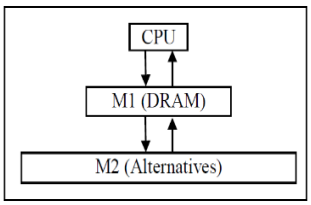

Figure 25: Alternative hybrid options [42]

In [44], a new combined main memory structure with multi-page cache was suggested as shown in Fig.26. In this architecture, Dai and his colleagues proposed GFDP (Global File Data Block Placement) algorithm for placement on file data block issue. They also developed an ILP model for the placement on file data-block issue [44].

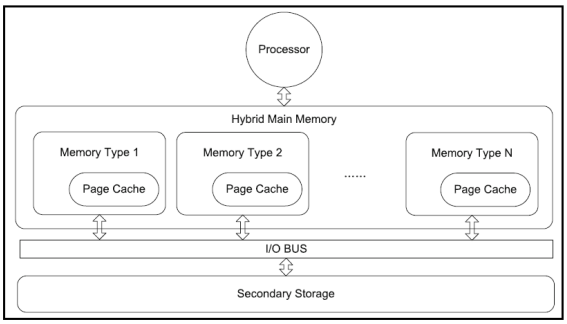

Figure 26: Hybrid memory options [44]

4. PERFORMANCE ANALYSIS

In this part, we examined combined main memory architectures through comparison in terms of energy saving rate, performance rate and endurance/lifetime rate. While the studies shown in Table 2 share mathematical values and measurement results, others interpret them in general, not give numerical results.

Apart from the studies given in Table 2; [31], [37], [18], [39], [22], [23], [26] have indicated that they have improved energy saving; [42], [14], [15], [37], [19], [41], [40], [21], [26] stated that they could provide performance improvement with the solution they proposed ; [28], [38], [35] have pointed out that they achieved positive improvements in durability, in their works.

Table 2: Performance analysis

Group	**Paper**	**Energy Saving Rate**	**Perf. Rate**	**Endurance / Life Time Rate**	**Method / Explanation**
1	[12]	-	%42	-	Chip cache
1	[16]	%18	-	-	A new caching policy
1	[24]	%11.7	%4.2	-	Refree
1	[25]	%29	%14	-	Concurrent Migration of Multiple Pages)
2	[27]	%30	-	-	Hybrid (hardware + software)
2	[28]	%35	-	-	Hot and cold pages method
2	[29]	%23.5 - %94.7	-	-	- Runtime-adaptive time out control - DRAM bypass - Keep dirty data cleaner
2	[30]	%20	-	-	OPAMP
2	[32]	%49	-	%88	ILP and graphic model for DSP systems
2	[34]	-	%21	-	Data Queue Stall Aware
2	[36]	%27	-	-	ASA (Adaptive Space Allocation)
4	[47]	-	%13 - %28.9	-	Instruction Per Cycle perform. with fixed page swap
4	[48]	%45	%30	-	Retention-Aware Hybrid Main Memory

5. CONCLUSION

As DRAM has begun to widespread use to meet increasing demand, studies on new main memory

solutions continue increasingly. It is desirable to use a second / more memory with DRAM to meet the increasing demand without giving up the advantages of DRAM. Instead of using PCM as an alternative to DRAM, adding PCM memory to DRAM (which is faster than other types of memory in writing) is seen as the most preferred method to provide read and stand-by improvements. Due to factors such as the fact that PCM memory does not need refreshing and that there are some very successful studies on eliminating the negativities such as lifetime, PCM memory is preferred to other memory types in hybrid main memory architectures, We expect that a hybrid main memory with DRAM and PCM memory and managed with a hybrid solution consisting of software + hardware can be produce very successful results in terms of endurance, performance and energy saving.

In this survey paper, the combined/hybrid main memory architectures are examined in order to show that combined/hybrid main memory solutions are promising. As well known, the use of non-volatile PRAM in conjunction with DRAM is one of the most ambitious solutions for combined/hybrid main memory operations. So, in the future, it is expected that the main memory technology can be improved much more with new works regarding the hybrid main memory architectures.

REFERENCES

[1] L. A. Barroso ve U. Hölzle, «The Case for Energy-Proportional Computing,» *IEEE Computer Society,* pp. 33-37, 2007.

[2] C. Lefurgy, K. Rajamani, F. Rawson, W. Felter, M. Kistler and T. W. Keller, "Energy Management for Commercial Servers," *IEEE Computer Society,* pp. 39-48, 2003.

[3] A. N. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis and N. P. Jouppi, "Rethinking DRAM design and organization for energy-constrained multi-cores," in *ISCA '10 Proceedings of the 37th annual international symposium on Computer architecture*, Saint-Malo, 2010.

[4] Micron, "Micron Technology, Inc. - Hybrid Memory Cube | Memory and Storage," 16 04 2017. [Online]. Available: https://www.micron.com/products/hybrid-memory-cube.

[5] H. M. C. Consortium, "Hybrid Memory Cube Consortium - Home," 16 04 2017. [Online]. Available: http://hybridmemorycube.org/.

[6] A. Sammons and C. Sciacca, "IBM New room," 17 05 2016. [Online]. Available: https://www-03.ibm.com/press/us/en/pressrelease/49746.wss.

[7] S. Bagheri, A. A. Asadi, W. Kinsner and N. Sepehri, "Ferroelectric random access memory (FRAM) fatigue test with Arduino and Raspberry Pi," in *2016 IEEE International Conference on Electro Information Technology (EIT)* , Grand Forks, 2016.

[8] Cypress, "Microcontrollers, Connectivity, Memory Solutions," 22 05 2017. [Online]. Available: http://www.cypress.com.

[9] T. Eshita, W. Wang, K. Nakamura, S. Mihara, H. Saito, Y. Hikosaka, K. Inoue, S. Kawashima, H. Yamaguchi and K. Nomura, "Development of ferroelectric RAM (FRAM) for mass production," in *Applications of Ferroelectrics, International Workshop on Acoustic Transduction Materials and Devices & Workshop on Piezoresponse Force Microscopy (ISAF/IWATMD/PFM), 2014 Joint IEEE International Symposium on the*, State College, PA, 2014.

[10] N. Yamada, E. Ohno, K. Nishiuchi, N. Akahira and M. Takao, "RAPID-PHASE TRANSITIONS OF GETE-SB2 TE3 PSEUDOBINARY AMORPHOUS THIN-FILMS FOR AN OPTICAL DISK MEMORY," *AMER INST PHYSICS,* pp. 2849-2856, 1991.

[11] J. Tominaga, T. Kikukawa, M. Takahashi and R. Phillips, "Structure of the optical phase change memory alloy, Ag-V-In-Sb-Te, determined by optical spectroscopy and electron diffraction," *AMER INST PHYSICS,* pp. 3214-3218, 1997.

[12] H. G. Lee, S. Baek, C. Nicopoulos and J. Kim, "An Energy- and Performance-Aware DRAM Cache Architecture for Hybrid DRAM/PCM Main Memory Systems," in *2011 IEEE 29th International Conference on Computer Design (ICCD)*, Amherst, MA, 2011.

[13] G. Wu, H. Zhang, Y. Dong and J. Hu, "CAR: Securing PCM Main Memory System with Cache Address Remapping," in *2012 IEEE 18th International Conference on Parallel and Distributed Systems*, Singapore, 2012.

[14] L. Ramos and R. Bianchini, "Exploiting Phase-Change Memory in Cooperative Caches," in *2012 IEEE 24th International Symposium on Computer Architecture and High Performance Computing*, New York, NY, 2012.

[15] S. Kwon, D. Kim, Y. Kim, S. Yoo and S. Lee, "A Case Study on the Application of Real Phase-Change RAM to Main Memory Subsystem," in *2012 Design, Automation & Test in Europe Conference & Exhibition (DATE) Design*, Dresden, 2012.

[16] J. Meza, J. Chang, H. Yoon, O. Mutlu and P. Ranganathan, "Enabling Efficient and Scalable Hybrid Memories Using Fine-Granularity DRAM Cache Management," *IEEE Computer Architecture Letters,* pp. 61-64, 2012.

[17] J. Hu, Q. Zhuge, C. J. Xue, W.-C. Tseng and E. H.-M. Sha, "Software enabled wear-leveling for hybrid PCM main memory on embedded systems," in *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE) Design*, Grenoble, 2013.

[18] Z. Wang, Z. Gu and Z. Shao, "Optimizated Allocation of Data Variables to PCM/DRAM-based Hybrid Main Memory for Real-Time Embedded Systems," *IEEE Embedded Systems Letters,* pp. 61-64, 2014.

[19] L. Ramos and R. Bianchini, "Robust performance in hybrid-memory cooperative caches," *Parallel Computing,* p. 514–525, 2014.

[20] J. Hu, M. Xie, C. Pan, C. J. Xue, Q. Zhuge and E. H.-M. Sha, "Low Overhead Software Wear Leveling for Hybrid PCM + DRAM Main Memory on Embedded Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* pp. 654-663, 2014.

[21] K. Kavi, S. Pianelli, G. Pisano, G. Regina and M. Ignatowski, "Memory organizations for 3D-DRAMs and PCMs in processor memory hierarchy," *Journal of Systems Architecture,* pp. 539-552, 2015.

[22] G. Wang, Y. Guan, Y. Wang and Z. Shao, "Energy-aware assignment and scheduling for hybrid main memory in embedded systems," *Computing. March 2016,* p. 279–301, 2016.

[23] Z. Zhang, Z. Jia, P. Liu and L. Ju, "Energy Efficient Real-Time Task Scheduling for Embedded Systems with Hybrid Main Memory," *Journal of Signal Processing Systems,* p. 69–89, 2016.

[24] B. Pourshirazi and Z. Zhu, "Refree: A Refresh-Free Hybrid DRAM/PCM Main Memory System," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Chicago, IL, 2016.

[25] S. Bock, B. R. Childers, R. Melhem and D. Moss´e, "Concurrent Migration of Multiple Pages in Software-Managed Hybrid Main Memory," in *2016 IEEE 34th International Conference on Computer Design (ICCD)*, Scottsdale, AZ, 2016.

[26] J. Zhang, X. Liao, H. Jin, D. Liu, L. Lin and K. Zhao, "An Optimal Page-Level Power Management Strategy in PCM–DRAM Hybrid Memory," *International Journal of Parallel Programming,* pp. 4-16, 2017.

[27] G. Dhiman, R. Ayoub and T. Rosing, "PDRAM: A Hybrid PRAM and DRAM Main Memory System," in *2009 46th ACM/IEEE Design Automation Conference Design Automation Conference*, San Francisco, CA, 2009.

[28] Y. Park, D.-J. Shin, S. K. Park and K. H. Park, "Power-Aware Memory Management for Hybrid Main Memory," in *The 2nd International Conference on Next Generation Information Technology Next Generation Information Technology (ICNIT)*, Gyeongju, Korea (South), 2011.

[29] H. Park, S. Yoo and S. Lee, "Power Management of Hybrid DRAM/PRAM-Based Main Memory," in *2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Diego, CA, 2011.

[30] J.-H. Choi, S.-M. Kim, C. Kim, K.-W. Park and K. H. Park, "OPAMP: Evaluation Framework for Optimal Page Allocation of Hybrid Main Memory Architecture," in *2012 IEEE 18th International Conference on Parallel and Distributed Systems Parallel and Distributed Systems*, Singapore, 2012.

[31] W. Tian, Y. Zhao, L. Shi, Q. Li, J. Li, C. J. Xue, M. Li and E. Chen, "Task Allocation on Nonvolatile-Memory-Based Hybrid Main Memory," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* pp. 1271-1284, 2013.

[32] T. Liu, Y. Zhao, C. J. Xue and M. Li, "Power-Aware Variable Partitioning for DSPs With Hybrid PRAM and DRAM Main Memory," *IEEE TRANSACTIONS ON SIGNAL PROCESSING,,* pp. 3509-3520, 2013.

[33] M. Mao, C. Yang, Z. Xu, Y. Cao and C. Chakrabarti, "Low cost ECC schemes for improving the reliability of DRAM+ PRAM MAIN memory systems," in *Signal Processing Systems (SiPS), 2014 IEEE Workshop on*, Belfast, 2014.

[34] D. Kim, S. Yoo and S. Lee, "Hybrid Main Memory for High Bandwidth Multi-Core System," *IEEE TRANSACTIONS ON MULTI-SCALE COMPUTING SYSTEMS,* pp. 138-149, 2015.

[35] S.-I. Jang, S.-K. Yoon, K. Park, G.-H. Park and S.-D. Kim, "Data Classification Management with its Interfacing Structure for Hybrid SLC/MLC PRAM Main Memory," *COMPUTER JOURNAL,* pp. 2852-2863, 2015.

[36] X. Cai, L. Ju, X. Li, Z. Zhang and Z. Jia, "Energy efficient task allocation for hybrid main memory architecture," *Journal of Systems Architecture,* pp. 11-22, 2016.

[37] D. Knyaginin, G. N. Gaydadjiev and S. Per, "Crystal: A Design-Time Resource Partitioning Method for Hybrid Main Memory," in *Parallel Processing (ICPP), 2014 43rd International Conference on*, Minneapolis MN, 2014.

[38] G. Nakagawa and S. Oikawat, "Language Runtime Support for NVM/DRAM Hybrid Main Memory," in *2014 IEEE COOL Chips XVII (COOL Chips)*, Yokohama, 2014.

[39] A. Hassan, H. Vandierendonck and D. S. Nikolopoulos, "Energy-Efficient Hybrid DRAM/NVM Main Memory," in *International Conference on Parallel Architecture and Compilation*, San Francisco, CA, 2015.

[40] S. Bock, B. R. Childers, R. Melhem and D. Moss´e, "HMMSim: A Simulator for Hardware-Software Co-Design of Hybrid Main Memory," in *2015 IEEE International Conference on Grey Systems & Intelligent Services (GSIS)*, Leicester, United Kingdom, 2015.

[41] S. Bock, B. R. Childers, R. Melhem and D. Mosse, "Characterizing the Overhead of Software-Managed Hybrid Main Memory," in *IEEE 23rd International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, Atlanta, GA, 2015.

[42] D. Ye, A. Pavuluri, C. A. Waldspurger, B. Tsang, B. Rychlik and S. Woo, "Prototyping a Hybrid Main Memory Using a Virtual Machine Monitor," in *IEEE International Conference on Computer Design Computer Design*, Lake Tahoe, CA, 2008.

[43] J. Stevens, P. Tschirhart, M.-T. Chang, I. Bhati, P. Enns, J. Greensky, Z. Chisti, S.-L. Lu and B. Jacob, "An Integrated Simulation Infrastructure For The Entire Memory Hierarchy: Cache, Dram, Nonvolatile Memory, And Disk," *Intel Technology Journal,* pp. 184-200, 2013.

[44] P. Dai, Q. Zhuge, X. Chen, W. Jiang and E. H.-M. Sha, "Effective file data-block placement for different types of page cache on hybrid main memory architectures," *DESIGN AUTOMATION FOR EMBEDDED SYSTEMS,* pp. 485-506, 2013.

[45] Z. Chen, Y. Lu, N. Xiao and F. Liu, "A hybrid memory built by SSD and DRAM to support in-memory Big Data analytics," *KNOWLEDGE AND INFORMATION SYSTEMS,* pp. 335-354, 2015.

[46] J. B. Kwon, "Exploiting Storage Class Memory for Future Computer Systems: A Review," *IETE Technical Review ,* pp. 218-226, 2015.

[47] J.-Y. Jung and R. Melhem, "Empirical, Analytical Study of Hardware-based Page Swap in Hybrid Main Memory System," in *2016 28th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, Los Angeles, CA, 2016.

[48] W. Jing, K. Yang, Y. Lin, B. Lee, S. Yoon, Y. Ye, Y. Du and B. Chen, "Retention-Aware Hybrid Main Memory (RAHMM): Big DRAM and Little SCM," *IEEE Transactions on Computers,* pp. 912-918, 2017.

[49] S.-I. Jang, C.-G. Kim and S.-D. Kim, "An Efficient DRAM Converter for Non-Volatile Based Main Memory," in *IT Convergence and Security*, Pyeong Chang, Korea, 2012.