Journal of Computer Science

Original Research Paper

Segment-Aware Dynamic Partitioning PCM-DRAM: A Solution to IoT Devices Development Constraints

[1]Qijin Zhu, [1]Shuyi Liu, [2]Zahid Akhtar and [3]Kamran Siddique

[1]School of Computing and Data Science, Xiamen University Malaysia, Sepang, Malaysia
[2]Department of Network and Computer Security, State University of New York Polytechnic Institute, Utica, USA
[3]Department of Computer Science and Engineering, University of Alaska Anchorage, Anchorage, AK, USA

Article history
Received: 07-10-2022
Revised: 12-07-2023
Accepted: 17-07-2023

Corresponding Author:
Kamran Siddique
Department of Computer Science and Engineering, University of Alaska Anchorage, Anchorage, AK, USA
Email: kamransiddique.pk@gmail.com

Abstract: The Internet of Things (IoT) furnishes a visual blueprint for the future internet. It serves up sensors, actuators, and distal devices on the edge of the network, creating a giant interconnected network. The IoT era refers to the future where all the conceivable data streams are integrated into the IoT, granting human-barrier free access to physical entities on the internet. Along with the rapid progress of IoT, pressing issues have emerged. Energy dissipation, limited processing efficiency, and confined memory have become severe constraints for the IoT era. Phase Change Memory with Dynamic Random-Access Memory (PCM-DRAM) is a hybrid memory system that has been proven to reduce energy dissipation. It is known to have a great capacity, higher endurance, and low latency. In this study, we first analyze the significant constraints faced in the IoT development. We then analyze how these constraints can be solved by PCM-DRAM memory. To this end, we propose a PCM-DRAM hybrid memory system called "Segment-Aware and Dynamic Partitioning PCM-DRAM" (SADP PCM-DRAM). Our proposal is grounded in a meticulous evaluation of the specific requirements posed by IoT applications. Furthermore, we also proposed two essential equations for quantifying energy consumption and the overall performance in terms of average memory hit time.

Keywords: IoT, PCM-DRAM, Hybrid Architecture, Main Memory, SADP PCM-DRAM

Introduction

The Internet of Things (IoT) is an extended and expanded application of Networks based on the Internet, aiming to build connectivity among things anyplace, anytime with anyone. The new revolution of the Internet is usually recognized as the future Internet. However, along with the rapid progress of the IoT, some bottlenecks occurred at the current stage of IoT development. The IoT is an interconnected network of transferring and processing data between devices. As the vast quantity of data throughout and written into the memory of devices, the performance of the main memory on IoT devices is an essential factor in determining IoT development (Blaauw *et al*., 2014). Some IoT developers have recognized that one of the critical constraints of IoT elaboration is low energy-conserving process efficiency (Elmangoush *et al*., 2013a). To solve this problem, Phase Change Memory (PCM), low static energy consumption, and non-volatility memory can be of help.

For a general computer, existing main memory DRAM consumes up to amount percent of the energy caused by consistent refresh clock cycles (Nair *et al*., 2014). Therefore, DRAM involving main memory would lead to the bottleneck of energy consumption efficiency. PCM memory not only has comparable performance to DRAM but also has the advantages of low static energy and better scalability. Nevertheless, limited endurance and prolonged latency of writing operations reduce the probability of completely replacing DRAM as the main memory. Hence, to draw on PCM's advantage to offset DRAM's weaknesses and verse visa, hybrid organizations are created to integrate their strength. There are mainly two architectures of the hybrid memory proposed by researchers shown in Fig. 1, both of which intend to reduce the PCM's write operation. The first one confines PCM and DRAM in parallel, commonly applying hot/cold page identification and migration, while the second structure utilizes a small storage component DRAM playing a role as a cache for main memory, that is PCM. Although there are some shortcomings when applying them to IoTs, they both make use of the strength of DRAM and PCM with a beyond-comparison performance. The analysis will be conducted in the following parts.

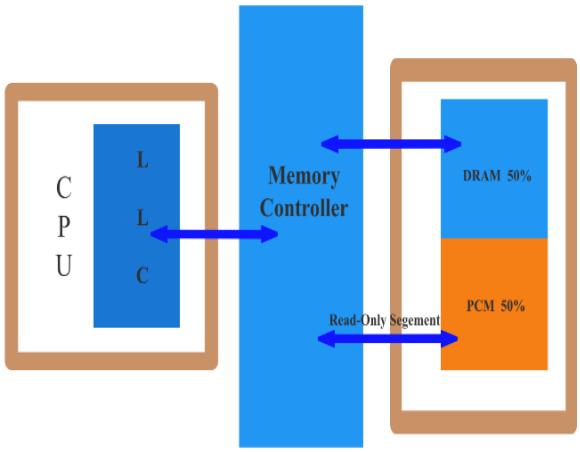

Fig. 1: Parallel organization of PCM-DRAM memory with a Memory Controller (MC)

Materials

IoT Bottlenecks

In terms of IoT development constraints, many scientists were aware of and pointed out in the related work. The constraints could be divided according to the devices involved. Some frontend scientists have discovered the fact that low power is a major constraint of the Internet of Things. Owing to the ultra-large scope of the whole Network and the mobility of devices within might not be able to get a timely power supply (Elmangoush *et al*., 2013a). Scientists trying to find a solution for reducing power consumption, (Blaauw *et al*., 2014) and his team proposed a theoretical hardware solution. They try to reduce energy usage in each action of IoT devices by applying a low-power circuit.

The Internet of Things carries its initial intention to connect things in the real world and share information among physical devices. Along with the rapid development, the opportunities come with the challenge. Many issues have been brought to the surface. From a macro view, IoT development bottlenecks can be roughly summarized into two broad levels: The social level and the technological level.

Social level refers to the difficulty and degree of IoT dissemination. The IoT's development is constrained due to the biased development of the different areas. In developing countries, the administrative and financial systems are run mostly without any integrated and automated system. Moreover, the level of technology usage is low and the investment in research and development is very little. Miazi *et al.* (2016) In developing countries, the primary issue is the internet access rate. According to the world bank data, compared to the 80% internet access rate in developed countries, the internet access rate in developing counting is merely 35% (World Bank, 2020). Since these areas still have difficulties facility with the fundamental needed elements such as power, Internet, and end-devices, further development is limited.

Another general aspect is the technical aspect. We can roughly break down the technical aspect into five different aspects. Haroon *et al*. (2016) comprehensively concludes the existing challenges in five aspects, they are security, storage, energy, communication, and standardization.

The IoT accepts various kinds of end-device and shares the data such as private information and index among devices. Zhang *et al*. (2014) conclude the security issue faced in the IoT era has the following aspects: Object identification and locating, authentication and authorization, privacy, lightweight cryptosystems, security protocols, software vulnerability, and malware.

In this study, continuous attention is focused on the power dissipating and the process speed issue faced by the IoT. The majority number of end devices in the IoT network are designed to be small and portable for convenient use. In real-life use, there is no guarantee that the device will be charged in time. The IoT device is constrained by the entity, which is in a physical monitoring state and the position of the entity is frequently changing without access to power (Elmangoush *et al*., 2013b).

It includes sub-question like addressing and sensing issues, network issue, congestion control, data buffering, and so on (Haroon *et al*., 2016).

Communication and standardization are the other two obstacles that lie in the way of IoT development. With a large amount of data transportation and information communication, the control of data congestion limits the efficiency of the IoT network. The constraints caused by standardizations mainly appears when different kind of devices tries to manipulate external data, interoperability sometimes is not available.

Phase Change Memory

The increasing need for main memory capacity has already provoked the research for a scalable, faster, and necessarily less power consumption with each generation. Current DRAM is limited by not only its significant energy consumption but also its low scalability. In contrast, Phase-Change Memory (PCM) sheds new light on the demand for non-volatile memory which has attracted much interest because of its scalability. Nevertheless, it has been noted that PCM is not good at handling write operations, particularly in terms of high latency and energy consumption.

Table 1: Attributes of PCM and DRAM (Khouzani *et al*., 2016)

Attributes	PCM	DRAM
Scalability	45 nm	65 nm
Latency of read	~10 ns	~10 ns
Latency of write	~100 ns	~10 ns
Read energy (pJ/bit)	2.5	4.4
Write energy (pJ/bit)	14 (set)~20 (reset)	5.5
Endurance of write cycles	10^8	10^{15}

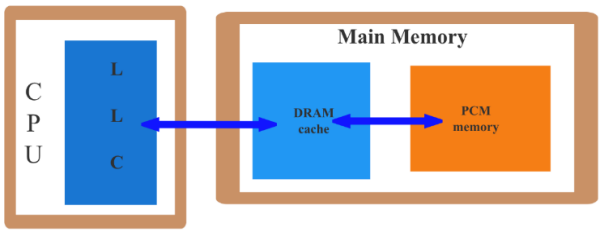

Fig. 2: DRAM-cache organization of PCM-DRAM memory

Table 1 compares the attribute of DRAM memory and PCM memory structures. Since DRAM is a volatile memory, it is a significant problem in consuming energy with consistent refresh cycles. Regarding the research (ITRS, 2013), the low scalability is a result of leakage current and capacitor placement. Conversely, it has been shown that PCM has better scalability and PCM does not need a power-hungry refresh cycle as it is a non-volatile memory. Regarding (Raoux *et al*., 2008), PCM can be developed on a 3*20 nm footprint prototype. Notably, the difference between their performance on the latency and energy consumption of write/read operations is the primary concern in this study. PCM consumes a smaller amount of energy when dealing with reading operations. On the other hand, DRAM is better at writing operations.

Hybrid Architecture of DRAM-PCM

In order to integrate the benefits of PCM & DRAM and to avert their disadvantages as much as possible, several works have been done to build them into a combination of memory with hybrid architecture. These designs are mainly divided into two main varieties, as visualized in Figs.1-2.

Figure 1 shows DRAM unit and PCM unit are organized in parallel, whereas Fig. 2 shows a relatively tiny DRAM unit as an off-chip cache. Multiple existing PCM-DRAM architectures are proposed to improve a certain aspect of DRAM or PCM's main memory drawback. Both parallel and hierarchical organization PCM-DRAM have a principle to settle write-frequent memory pages in DRAM while putting write-infrequent memory pages in the PCM component, which aims at reducing the energy efficiency and increasing the lifetime of hardware.

Parallel organization of PCM-DRAM memory is an architecture in which the whole address space is divided into two separate areas, as in Fig. 1. Normally, the DRAM's scale is the same as the PCM's. Data is stored exclusively on one of the two devices and migrating between them is expensive. The works directedly against this structure mainly work on identifying and exchanging cold data in DRAN and hot data in PCM. By way of illustration, the PDRAM architecture launched by Dhiman *et al*. (2009) utilizes an additional counter for recoding write frequency. Accordingly, the data migration is based on such frequency. Whereas, it has been proven that PDRAM can lead to some useless writing operations and result in unnecessary costs on energy. Recently, a CLOCK-RWRF algorithm with a parallel focus on using a relative frequency to measure which data should be migrated (Wang *et al*., 2020). Regardless of that, it improves the efficiency of data identification and migration. The cost of data migration still exists and cannot be ignored. In contrast to those architectures, our proposed architecture eliminates the cost of migration significantly. We take advantage of the segment information controlled by the operating system, to identify read-only segments in advance and avert the need for migration. Details will be discussed in part IV.

When it comes to the architecture in Fig. 2, where DRAM is an internal cache unit for main memory just like the Last Level Cache (LLC), its size is generally kept small. Fig. 2 displays that there are two cases of memory accesses. One is that accesses are directed to the DRAM and the other happens when DRAM is being accessed only when DRAM misses occur. However, competition for the DRAM cache resource can increase power consumption and reduce DRAM cache efficiency for low locality workloads. As a result, a high miss rate could generate negative impacts on overall energy consumption and performance, but there is some research recently conducted to resolve such problems and this architecture still has prospects. For instance, Lee *et al* launched a method to reduce DRAM miss penalties in their research (Lee *et al*., 2011). The architecture launched by Bheda *et al*., (2011) eliminates the necessity for memory controllers in PCM-DRAM architecture. His team also elaborated a methodology to not only reduce the mean access time of the PCM system but also perform at a comparable capacity with the DRAM-only system.

A major weakness that exists in the hybrid system is, first, the limited time for reading and writing operations (Dhiman *et al*., 2009). With the unbalance process time between two memories, The PCM's accessing time is 6-10 times that of the DRAM, causing the PCM idle when waiting for the DRAM to finish (Elmangoush *et al*., 2013b). Lee *et al*., (2013) launches a new technology to hide the slow writing performance and proposes a CLOCK-DWF algorithm to enlarge the process efficiency. Thakkar and Pasricha (2017) does some relevant work and launched a Dyphase PCM architecture to reduce the write latency and enhance the process performance; what's more, they also do an experiment that shows that after applying the new architecture, "Resulting in decelerated degradation and increased memory lifetime".

PCM-DRAM and IoT

Hybrid PCM-DRAM and IoT

The IoT is promoting rapid technological changes so that technological changes will bring new challenges and impacts to mainstream data storage technologies. From this point of view, the speed of development and change of the Internet of Things and the innovation of storage technology are mutually restricted. In recent years, many information storage structures have been studied to cater to the development of information technology and promote technological innovation. According to the research of Pimo *et al*. (2021), by comparing the read delay of PCM and DRAM, the average PCM is obtained the result is a growth rate of 48%. Therefore, a single PCM has a barrier at the technical level. Moreover, in order to improve efficiency and reduce runtime consumption, a hybrid memory structure such as PCM-DRAM was born. And in the context of taking advantage of the high density of PCM and the high energy efficiency of DRAM, there is a structure that uses a smaller-sized DRAM as an inclusive cache for PCM.

According to the existing research (Pimo *et al*., 2021), the read delay of PCM is longer than DRAM's. It is precise because PCM has higher memory time, which causes the overall system's delay rate to grow. This element also increases the possibility of having high miss rates. The CPU will initially access the data in the DRAM, which serves as the PCM's cache in this configuration. The CPU will access the data in the PCM if there is no comparable data in the DRAM. We may effectively halt writes by manipulating the data in this way, which decreases writes to the PCM.

The DRAM miss rate also has an impact on the performance of the memory structure. The efficiency of the memory structure is, however, significantly impacted by the miss rate's existence. The block size in the DRAM cache area is dynamically adjusted in an effort to remedy the situation; this technique successfully lowers the DRAM miss rate but disregards the PCM durability (Khouzani *et al*., 2016).

In addition to eliminating PCM's drawbacks, such as asymmetric read-write, the integration of DRAM and PCM also makes up for PCM's limited storage capacity. By implementing this architecture, IoT devices may efficiently address the issue of energy consumption and lower the system decay rate. At the material level, PCM is mainly dependent on a physical change from a highly resistive amorphous state to a more conductive crystalline one. PCM mostly uses chalcogenide glass, but as technology changes, underlying demand will also shift. In addition, some essential materials associated with emerging technologies have issues with skyrocketing raw material prices and supply security. However, the availability of raw materials will immediately impact the manufacture of PCM hardware, which will also have an impact on the adoption of pertinent storage designs on IoT devices. This will directly constrain the rate of development of IoT devices (Ku, 2018). As a result, in order to further promote the use of PCM-DRAM storage structure in the development of the IoT, further research should be done in order to improve storage density and lower the cost of raw materials, if the durability of PCM needs to be improved.

Parallel PCM-DRAM

There are two primary hybrid PCM-DRAM organizations proposed by the previous scientist to solve the power dissipation and process acceleration problem.

The parallel organization is one of the earliest proposed structures that consider the respective pros and cons of the PCM and the DRAM and innovatively combine them to achieve higher energy efficiency.

The key idea of the parallel structure is to keep the hot data within PCM while storing cold data within the DRAM. The memory address is exposed to the Operating system and in the later implementation, this will help to migrate page content among different memory easily. In this parallel hardware structure (Dhiman *et al*., 2009), the DRAM and PCM are split into two equal-sized pages. PRAM is holding the access map. Unlike the traditional memory system, DRAM here acts like a cache to store page information for DRAM. The data placement is based on the write-in frequency. If the write-in frequency of the PCM is exceeded the threshold, the hot data in PCM is transferred into DRAM.

Since DRAM has more read and write latency than PCM, it could handle massive hot data more efficiently to achieve wear-leveling. In the latter experiments, the structure has shown good performance in aspects such as power characteristics and read-write latency. According to Dhiman *et al*. (2009), the proposed structure is said to have a benchmark performance of energy saving.

Though the hybrid parallel structure has low-energy consumption and relatively higher process speed, the disadvantages still exist. First, energy efficiency could still be optimized. As there is a threshold limit to write in data volume and demands an entire page-swap strategy, the energy could be dissipated and wasted.

Second, in this structure, we mainly depend on PCM when dealing with the hot data; however, unlike DRAM, PCM has no more than 10^8 write operations to perform. Once the write operation exceeds the threshold, the inability of PCM will cause data loss for the whole PCM-DRAM hybrid structure with the high-performance ability and the energy-saving outcome, such a hybrid memory system might be effectively applied to IoT end devices. Such a hybrid system has been proven to have better energy efficiency; it could be of help in solving power

constrain, where devices are unable to cause power dissipation and timely replenishment.

Overall Benefits of PCM-DRAM to IoT

IoT devices' limitations, including energy dissipation, working lifetime, and size, have been the subject of numerous solutions in recent years. With IoT devices' increasing popularity, being used to monitor various scenarios, such as an office's energy consumption or a big shopping mall's condition. when they cannot be plugged into a power supplier, their main problem is their very unpredictable battery life. PCM-DRAM has been proven to have better energy efficiency; and high-performance ability. All of these characteristics are granted with high expectations when they are applied in practice. PCM-DRAM could be of help in solving power constrain, where devices are unable to cause by power dissipation and timely replenishment. Once the power problem is improved, the popularity and the ubiquity will be increased, it will greatly help with the IoT's development.

Another barrier that could be moved is the issue with the device. In order to gain wider use, the end devices must not be bulky. If carrying on the traditional memory hierarchy, Hay *et al*. (2011) found, that it is almost impossible to keep minimize DRAM size while maintaining the same functionalities. It is tough to find a tradeoff between better storage performance and place refinement. The Hybrid system introduces PCM, which has a higher storage density and a size as small as 3 nm (Raoux *et al*., 2008).

Methods

Out of concerns about IoT applications' diverse variety and drawbacks of the hybrid PCM-DRAM architectures, our team proposed a parallel PCM-DRAM architecture with segment-aware and dynamic partitioning. This is based on diverse application algorithms (SADP PCM-DRAM). The proposed memory structure mainly aims at reducing write operation in PCM, to avoid high energy consumption when doing page migration. Moreover, since our target structure is the parallel organization, a high energy consumption led by a high miss rate caused by DRAM-cache architecture can be averted.

We outline our main contributions as follows:

1. Analyze bottlenecks of current IoT development in terms of memory and evaluate what PCM-DRAM can contribute to solving the problems
2. Propose a Read-only Segment Aware and Dynamic-Partitioning hybrid-parallel memory architecture to eliminate the high miss rate of hierarchical architecture, reduce cost on page migration of parallel architecture and write operations on PCM
3. Devise an approach to evaluate overall performance and energy consumption

We layout our methodology into the following five stages:

Stage 1: We collected information from related research on how IoT is suffered from constraints. We came to the finding that IoT needs a low energy consumption and comparable performance memory to break the bottleneck

Stage 2: After reviewing the current memory technology, we analyzed different architectures of PCM-DRAM proposed in the last eight years. We qualitatively summarized DRAM-cache and Parallel organization's strengths and drawbacks. A DRAM-cache architecture called Segment-Aware memory access proposed by Khouzani *et al*. (2016) arouses our interest

Stage 3: Based on the information on memory architectures from the second stage, we analyzed the feasibility of whether Segment-aware memory access can be applied to parallel organization

Stage 4: According to the knowledge acquired from the previous two stages, we proposed a parallel PCM-DRAM architecture with segment aware memory access technique. The architecture is also designed to be compatible with IoT applications

Stage 5: Finally, in order to evaluate our model's performance qualitatively, we devised two equations and implemented a simulator called DPHSim (Jin *et al*., 2017). However, one of the main obstacles in the research is the time limitation. We only finished the first four stages. In the final stage of the experiment, we simply proved that our architecture is applicable in DPHSim. We still need more comparison experiments to prove the strength of our structure

Implementation

Proposed Architecture

Motivation

1) DRAM cache architecture leads to a high miss rate, resulting in energy waste
2) Traditional parallel hybrid architecture memory needs to identify whether memory paging is cold or hot to migrate them between PCM and DRAM, which is a costly process
3) Text segment-aware memory access can identify read-only text segments before allocating memory page

Parallel Hybrid Architecture to Avoid a High Miss Rate

Currently, exploiting DRAM and PCM aggregates architectures to cover the drawbacks of DRAM and PCM attracts excellent attention. There are two main structures proposed by researchers. The first structure in Fig. 2 proves that DRAM is used as an up-degree of cache visible to the Operating System (OS). Just like the L1 and L2 cache. In this structure, the pages in DRAM are the maximum viable accessed sub-pages in PCM and PCM can be best accessed while the failed memory page access happens in the DRAM cache. However, for workloads that have a lousy locality, the conflict misses when accessing DRAM cache aid might also additionally boom power intake and degrade the performance of the DRAM cache.

Although some researchers have already figured out some approaches to enhance DRAM-cache's efficiency of energy consumption, the dependency of its performance on miss rate is still non-negligible. For instance, Khouzani *et al*., proposed a hierarchical DRAM-cache architecture to reduce the hit rate and write on PCM (Khouzani *et al*., 2016). However, the miss rate caused by DRAM-cache competition cannot be ignored when applied to IoT applications.

On the opposite hand, due to a few pages' repetition placements in the main memory, the usage of memory space is low. Figure 1 depicts the opposite structure that DRAM and PCM are positioned inside the parallel degree of reminiscence hierarchy and the percentage of the identical bodily address of memory. For the CPU, it is best to treat PCM and DRAM as memory together because the memory gets entry to each of them directly. In order to make use of the benefits of DRAM and PCM, OS wants to undertake the correct coverage to soak up the write operations into DRAM. To avert the high energy consumption out of conflict misses, we focus on the structure displayed in Fig. 1, named "parallel hybrid structure" in this study and our structure is present in Fig. 4. Read-only segment aware memory access-to avert page migration.

Traditional parallel hybrid memory architecture has two non-overlapping regions with the same size as DRAM and PCM. Data is stored either in the PCM or DRAM region, along with the program's execution by CPU. Write and read operations are continuously happening within the main memory. To fulfill the principle of "write-intensive memory in DRAM and read-frequent memory in PCM," page's write frequency, including relative (Dhiman *et al*., 2009) or absolute (Khouzani *et al*., 2016) frequency detection and migration between two regions, are destined to happen. However, such architectures have already proved that they could not only generate unnecessary write operations to PCM but also may cause high latency.

It will be better to exclude read-only operations in program contexts from the DRAM unit and save space for writing operations, as our primary goal is to avoid the drawbacks of the two memory units. Commonly, the main three segments of a given program are: One houses the code, namely the code segment, one houses the data, which is the data segment and the third one is the stack segment. It is clear that the text segment is read-only and the stack segment is used for temporary store data; therefore, PCM is where the text segment accesses and DRAM is where the stack segment accesses. As for the data segment, it consists of two parts of the initialized segment and the uninitialized segment, which include constants such as strings and Read/Write memory for variables declared. Our structure will access DRAM when the Memory Controller (MC) encounters Read/Write memory as variables could be modified at run time, which is in contrast to constants in the data segment.

The hybrid architecture we proposed by us settles read-only segments directly in PCM, while DRAM stores other pages. This implementation is straightforward as the Operating System (OS), which is in control of memory address allocation and holds the segmentation information. A flowchart elaborating how memory access is processed in Fig. 3 is provided. Memory access is triggered by LLC, either a miss or writeback. Since writebacks are operations caused by evicting dirty blocks in the LLC and do not target read-only segments, writeback accesses are directly routed to DRAM. When it comes to the miss from LLC, if it is a write miss, it will be routed to DRAM directly. CPU accesses utilizing virtual addresses can lead to a cache miss. Hence the segmentation information will be acquired from the virtual addresses. If the processed segment is a code segment, it will directly access PCM to handle the read-only code segments. Finally, the memory controller needs to identify the part of data segments that keep constants based on segment information, since CPU accesses that use virtual addresses cause cache misses (Lopriore, 1988; Khouzani *et al*., 2016).

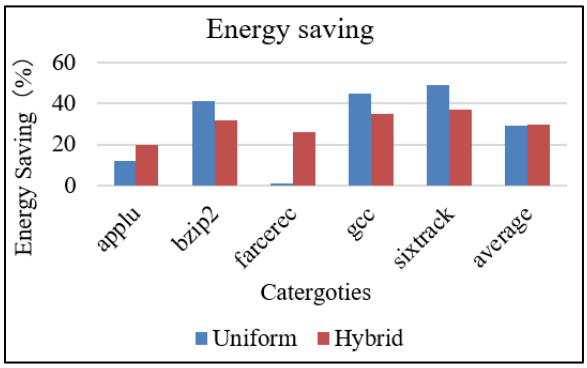

Fig. 3: Energy Savings comparison for hybrid and uniform memory systems

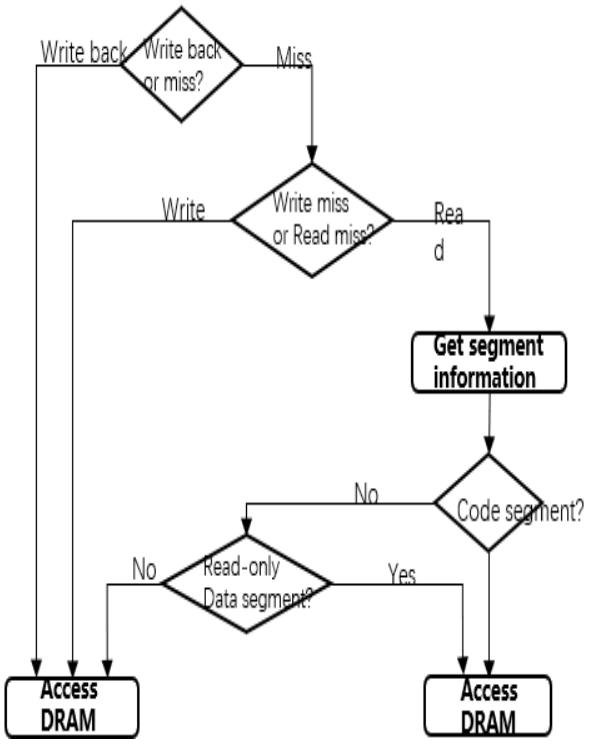

Fig. 4. Flowchart of segment-aware memory access

Dynamic Partitioning PCM Memory-Improve Memory Utilization

Dynamic partitioning refers to the PCM-DRAM partitioning of the structure that can be altered according to different IoT applications by the designers. Since PCM is only in charge of read-only operations and content in PCM is relatively static, reasonably allocating the proportion of PCM memory is worth concern. Specifically, two memory regions are separated by write and read roughly; although there are still read operations in the DRAM region, write operation hardly happens in the PCM region.

The partitioning of PCM and DRAM should be determined by the rate of write and read operation of a program, or more specifically, the rate of text-only segments and non-text-only segments. If we simply adopt the traditional 50% of PCM as the traditional parallel hybrid PCM-DRAM architectures did, it is easy lead to low memory utilization.

As each application of IoT is generally in charge of one particular function, the underlying rate of writing and read operation is approximately fixed. Therefore, it is one energy and performance-efficient way to adopt different PCM and DRAM partitioning rates for different IoT applications. For different IoT applications, engineers are supposed to decide how PCM is more reasonably partitioned according to what is the application of IoT doing.

Table 2: Attributes of DRAM and PCM from the simulator

Attributes	DRAM	PCM
Power of leakage (W)	5.19	1.39
The energy of write (nJ)	99.10	81.50
The energy of read (nJ)	99.10	1.74
Latency of write (ns)	15.20	321.40
Latency of reading (ns)	15.20	63.60
Area (mm^2)	120.00	320.00

Experiment Setup

Since PCM is a new variety of non-volatile memory that be paid attention to, there is evidence showing that current PC chips and devices are not available and it is challenging to evaluate algorithms related to PCM (Jin *et al*., 2017). Out of such motivation, a group from China developed a simulator on PCM and DRAM configurations.

We implement DPHSim with uni-core CPU, multi-programmed and 2-Level cache workloads to evaluate SADP architecture and make a comparison with other existing models, which are LRU-WPAM, CLOCK-DWF (Wang *et al*., 2020) architecture in terms of performance and energy. We set up 3 groups for each model so that we could obtain an average value of hit ratio and energy consumption. In this experiment, the scale of DRAM and PCM is set to 32 MB and 1 GB respectively, 4 KB page size is equipped. DPHSim is utilized for estimating the energy consummation and the latency of different models. We still need more set-up configurations for other components. The size of LLC, last level cache, is 1MB with 128 block size and 4 associativity. Table 2 demonstrates attributes of the PCM and DRAM in the experiment from DPHSim.

The results illustrate that SAPD can improve performance and reduce energy consumption compared to DRAM-only memory. Therefore, SAPD is proven to be applicable.

Results

Performance Evaluation

Average Memory Hit Time Evaluates Overall Performance

This part provides an overall performance of the proposed memory architecture determined by the Average Memory Hit Time (AMHT) as per the following equation:

$$AMHT = Rate_{read-only} * T_{PCMread} + \left(1 - Rate_{read-only}\right) * \left(T_{DRAM} + Rate_{DRAMmiss} * T_{PCMread}\right) \quad (1)$$

where, $Rate_{read\text{-}only}$ is the rate of read-only segments, $T_{PCMread}$ and T_{DRAM} are PCM read latency and DRAM access latency, and $Rate_{DRAMmiss}$ is the DRAM miss rate adopted for calculating AMHT of the strategy stated in part VI that applies segment-aware memory access strategy.

Table 3: Estimated hit ratio of three models

	Hit ratio			
Model	1st group	2nd group	3rd group	Avg ratio
DWF	0.34	0.32	0.32	0.327
WPAM	0.41	0.44	0.39	0.413
SADP	0.32	0.36	0.33	0.337

Table 4: Estimated energy consumption of three models

	Energy consumption/KJ			
Model	1st group	2nd group	3rd group	Avg consumption
DWF	224	223	219	222.0
WPAM	150	0151	148	149.7
SADP	70	73	71	71.3

Total Energy Consumption Evaluation Equation

Evaluation of total energy consumption during the memory accesses is determined by the sum of write and read of DRAM and PCM, i.e.,

$$E = E_{DRAMwrite} * N_{DRAMwrite} + E_{DRAMread} * N_{DRAMread} + E_{PCMwrite} * N_{PCMwrite} + E_{PCMread} * N_{PCMread} \quad (2)$$

$$E = E_{DRAMwrite} * N_{DRAMwrite} + E_{DRAMread} * N_{DRAMread} + E_{PCMwrite} * E_{PCMread} * N_{PCMread} \quad (3)$$

$$E = E_{DRAMwrite} * N_{DRAMread} + E_{DRAMread} * N_{DRAMread} + E_{PCMwrite} + E_{PCMread} * N_{PCMread} \quad (4)$$

where, E_x is the energy of one x operation and N_x is the number of x operations. Worthy of note is that $N_{PCMwrite}$ is one, for the first write of read-only memory from disk when executing a program. In order to simplify the calculation, Energy cost by writing operations on PCM can be ignored.

Analysis

We can clearly assess the performance by using the hit ratio shown in Table 3 and compare the energy consumption from Table 4. SADP is not modeled with the highest hit ratio since the CLOCK-WPAM model has a better hit ratio than SADP. Nevertheless, when it comes to the energy consumption of SADP, it has a lower consumption compared with CLOCK-DWF and CLOCK-WPAM, which proves that SADP can efficiently save energy for devices.

Discussion

The proposed SADP architecture proposal demonstrates a noteworthy capacity to enhance energy efficiency for IoT end-devices. This architecture's comprehensive performance is evident across three distinct models. This implies that even though SADP might not outperform in specific performance metrics (hit ratio), its practical value shines through its energy efficiency and its aptitude for catering to the distinctive requirements of IoT end-devices.

Conclusion

This study first addresses the issue of the IoT development bottleneck. Through the analysis of the IoT bottleneck, we attempted to find a solution to the energy dissipation constraints of the IoT. We put our focus on the basic energy consummation unit of the IoT end-devices, that is the memory unit. We then found that the existing main memory unit has many constraints and a new hybrid structure using PCM-DRAM could help improve the performance of the memory thus helping in solving the IoT bottleneck. By analyzing two typical existing hybrid structures, we understand the main logic behind the "hybrid strategy" along with their disadvantages, so we proposed a more advanced new structure. Meanwhile, we proposed two equations to evaluate the performance of the architecture.

Our work is limited in the sense that it lacks practice data support, i.e., the DPHSim simulator may not be able to verify the performance of our architecture with environmental circumstances.

Our work is supported by the data obtained from experiment carried out by the DPHSim simulator, we compare the Hit Ratio and energy consumption of 3 models, including SADP. The result illustrates that although it does not have the eminent performance of hit ratio, SADP has a sterling energy-saving ability.

In future research, we aim to apply our structure to real IoT end devices to assess their performance.

Acknowledgment

The authors would like to thank anonymous reviewers and the editors of this manuscript for their constructive feedback, which has been greatly appreciated.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Qijin Zhu and Shuyi Liu: Conceptualization and methodology, data curation, validation and formal analysis, written original drafted preparation, written reviewed and edited.

Zahid Akhtar and Kamran Siddique: Validation and formal analysis, investigation, written original drafted preparation, written reviewed and edited, supervision, project administration.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all authors have read and approved the manuscript and no ethical issues are involved.

References

Bheda, R. A., Poovey, J. A., Beu, J. G., & Conte, T. M. (2011, July). Energy efficient phase change memory based main memory for future high-performance systems. In *2011 International Green Computing Conference and Workshops* (pp. 1-8). IEEE. https://doi.org/10.1109/IGCC.2011.6008569

Blaauw, D., Sylvester, D., Dutta, P., Lee, Y., Lee, I., Bang, S., ... & Choi, M. (2014, June). IoT design space challenges: Circuits and systems. In *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers* (pp. 1-2). IEEE. https://doi.org/10.1109/VLSIT.2014.6894411

Dhiman, G., Ayoub, R., & Rosing, T. (2009, July). PDRAM: A hybrid PRAM and DRAM main memory system. In *Proceedings of the 46th Annual Design Automation Conference* (pp. 664-469). https://doi.org/10.1145/1629911.1630086

Elmangoush, A., Coskun, H., Wahle, S., & Magedanz, T. (2013a, March). Design aspects for a reference M2M communication platform for Smart Cities. In *2013 9th International Conference on Innovations in Information Technology (IIT)* (pp. 204-209). IEEE. https://doi.org/10.1109/Innovations.2013.6544419

Elmangoush, A., Al-Hezmi, A., & Magedanz, T. (2013b, December). Towards standard M2M apis for cloud-based telco service platforms. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia* (pp. 143-149). https://doi.org/10.1145/2536853.2536892

Haroon, A., Shah, M. A., Asim, Y., Naeem, W., Kamran, M., & Javaid, Q. (2016). Constraints in the IoT: the world in 2020 and beyond. *International Journal of Advanced Computer Science and Applications*, *7*(11).

Hay, A., Strauss, K., Sherwood, T., Loh, G. H., & Burger, D. (2011, December). Preventing PCM banks from seizing too much power. In *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 186-195). https://doi.org/10.1145/2155620.2155642

Khouzani, H. A., Hosseini, F. S., & Yang, C. (2016). Segment and conflict aware page allocation and migration in DRAM-PCM hybrid main memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *36*(9), 1458-1470. https://doi.org/10.1109/TCAD.2016.2615845

ITRS. (2013). International technology roadmap for semiconductors, Emerging Research Devices (ERD). http://www.itrs2.net/2013-itrs.html

Jin, P., Wang, X., Zhang, D., & Yue, L. (2017, October). Work-in-progress: effective simulation of DRAM/PCM-based hybrid memory. In *2017 International Conference on Embedded Software (EMSOFT)* (pp. 1-2). IEEE. https://doi.org/10.1145/3125503.3125564

Ku, A. Y. (2018). Anticipating critical materials implications from the Internet of Things (IOT): Potential stress on future supply chains from emerging data storage technologies. *Sustainable Materials and Technologies*, *15*, 27-32. https://doi.org/10.1016/j.susmat.2017.10.001

Lee, H. G., Baek, S., Nicopoulos, C., & Kim, J. (2011, October). 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)* (pp. 381-387). IEEE. https://doi.org/10.1109/ICCD.2011.6081427

Lee, S., Bahn, H., & Noh, S. H. (2013). Clock-DWF: A write-history-aware page replacement algorithm for hybrid PCM and DRAM memory architectures. *IEEE Transactions on Computers*, *63*(9), 2187-2200. https://doi.org/10.1109/TC.2013.98

Lopriore, L. A. N. F. R. A. N. C. O. (1988). Virtual address cache with no reverse address buffering. *Proceedings of the IEEE*, *76*(11), 1538-1540. https://doi.org/10.1109/5.90117

Miazi, M. N. S., Erasmus, Z., Razzaque, M. A., Zennaro, M., & Bagula, A. (2016, May). Enabling the Internet of Things in developing countries: Opportunities and challenges. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 564-569). IEEE. https://doi.org/10.1109/ICIEV.2016.7760066

Nair, P. J., Chou, C. C., & Qureshi, M. K. (2014). Refresh pausing in DRAM memory systems. *ACM Transactions on Architecture and Code Optimization (TACO)*, *11*(1), 1-26. https://doi.org/10.1145/2579669

Pimo, E. S. J., Ashok, V., Logeswaran, T., & Satyanarayana, D. S. S. (2021). Withd Rawn: A comparative performance analysis of phase change memory as main memory and DRAM. https://doi.org/10.1016/j.matpr.2021.01.473

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. https://doi.org/10.1147/rd.524.0465

Thakkar, I. G., & Pasricha, S. (2017). DyPhase: A dynamic phase change memory architecture with symmetric write latency and restorable endurance. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *37*(9), 1760-1773. https://doi.org/10.1109/TCAD.2017.2762921

Wang, H., Shen, Z., Zhao, M., Cai, X., & Jia, Z. (2020, December). CLOCK-RWRF: A read-write-relative-frequency page replacement algorithm for PCM and DRAM of hybrid memory. In *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (pp. 189-196). IEEE. https://doi.org/10.1109/HPCC-SmartCity-DSS50907.2020.00024

World Bank. (2020). Connecting for inclusion: Broadband access for all. https://www.worldbank.org/en/topic/digitaldevelopment/brief/connecting-for-inclusion-broadband-access-for-all

Zhang, Z. K., Cho, M. C. Y., Wang, C. W., Hsu, C. W., Chen, C. K., & Shieh, S. (2014, November). IoT security: Ongoing challenges and research opportunities. In *2014 IEEE 7th International Conference on Service-Oriented Computing and Applications* (pp. 230-234). IEEE. https://doi.org/10.1109/SOCA.2014.58