# RETROSPECTIVE: Scalable High-Performance Main-Memory System using Phase-Change Memory Technology

Moinuddin Qureshi
Georgia Tech

Vijayalakshmi Srinivasan
IBM Research

Jude A. Rivers
Independent Consultant

This paper appeared at ISCA-2009, as part of a dedicated first session, which surprisingly had three papers on *Phase Change Memories (PCM)*, a topic that had not previously appeared at a top architecture conference. The hybrid memory architecture proposed in this paper would go on to become a blueprint for memory systems based on emerging memory technologies. There was a bit of serendipity involved in the formation of this paper – mainly being at the right place at the right time and a lucky rejection of an earlier short paper. With this retrospective, we hope to share the context and backdrop in which this research was done, the key contributions, and the impact and legacy of the paper.

## I. The Backdrop

All three authors were Research Staff Members at IBM T. J. Watson Research Center at the time of this work (Aug-Nov of 2008). Viji and Jude had been with IBM for almost a decade, and Moin had joined IBM only about a year earlier. His first year at IBM was focused on capacity sharing for private caches of Power-7, and a paper on this topic was submitted to HPCA-2009 in Aug 2008. Coincidentally, the work on the PCM paper originated right on the day of submission of that HPCA paper.

Back then, the HPCA industrial session used to have a submission deadline that was about 4 weeks away from the deadline for the regular conference, with papers about half the number of pages. The genesis for our paper was in Viji's office, when over coffee, we organically started discussing potential topics for possible submission to the HPCA-2009 industry session. In hindsight, we just happened to be at the right place at the right time, in that, in those days, there were lots of technical seminars at IBM Watson on *Storage Class Memories (SCM)*. So, we were aware that there was some movement in the device community on emerging devices that had better scaling properties than DRAM. However, much of the discussion on the topic was restricted to the devices/circuits teams with no studies on the impact this would have on system performance, or how one would go about architecting memory systems for these emerging technologies. We decided to pursue this line of research and submit a (short) paper to the HPCA-2009 industrial session on "Architecture Impact of Emerging Memory Technologies". This paper tried to describe the landscape of emerging technologies (PCM, Flash etc.) using an abstract model and reasoning about the performance implications of using memory systems made of these emerging memory technologies.

After the submission to the HPCA industry paper, we realized that PCM was the main contender, and it would make sense to study memory systems for PCM, focusing on overcoming the specific drawbacks of PCM. As our understanding of PCM improved, we were concerned that if the HPCA-2009 industry paper got accepted, we probably would not be able to submit our PCM paper to ISCA-2009 (due to overlap).

HPCA-2009 industry session was canceled due to a lack of submissions and instead we were invited to a panel discussion, allowing us to submit our PCM paper to ISCA-2009. At the time of submission, we had our doubts if this topic of a relatively unknown, new and emerging technology would be of interest to ISCA. Little did we know that there would be three papers on PCM at ISCA, with a dedicated first session, and that our hybrid memory (DRAM+PCM) architecture would become blueprint for future NVM systems, let alone that this paper would go on to become one of the highest cited papers on memory systems.

## II. The Problem and Challenges

The key problem solved by this paper was to increase the capacity of the main memory. DRAM is the primary technology for main memory systems, however, DRAM scaling was slowing down, significantly compared to logic. For example, a contemporary paper at that time from HP showed that the number of cores doubled every 2 years, but DRAM DIMM capacity doubled every 3 years, causing the memory capacity-per-core to go down 30% every 2 years. Furthermore, there were significant concerns about decreasing DRAM reliability which caused poor yield (this eventually resulted in the incorporation of on-die ECC in DRAM chips). It also became important to reduce the power consumed in memory refresh.

We proposed to architect the memory system with a new technology with better scaling and density than DRAM. *Phase Change Memory* was the leading technology candidate at that time, with major companies such as Samsung (and some newer companies like Numonyx) demonstrating gigabit scale PCM chips, serving as proof-points that the technology was becoming relatively mature for system adoption. PCM's better density (2x-4x) relative to DRAM enabled larger capacity memory systems, but it came with drawbacks, namely, higher read latency (about 2x-4x of DRAM), much higher write latency (about 4x-8x of DRAM), and limited write endurance (10-100 million). Incorporating PCM in memory systems required addressing these significant drawbacks.

The goal of our paper was to enable scalable high-performance memory systems using PCM, while addressing the drawbacks of PCM. There were two practical challenges in performing this research: (1) having a reasonable set of assumptions for the PCM devices (2) lack of simulation infrastructure for evaluating large-capacity memory systems, specifically the impact of changing memory capacity from X gigabytes to Y gigabytes, as simulators at that time implicitly assumed that the workload fits in memory thus skipping page faults. As part of our study, we overcame both challenges, by modeling PCM with parameterized numbers for density and latency (default values set based on discussion with experts), and by developing a new memory system simulator that models memory capacity and page faults and generating traces for large workloads with working sets of several gigabytes.

## III. The Contributions

The performance limitations of PCM were significant roadblocks. Naively replacing DRAM with PCM would get a larger capacity memory system but with 2x-4x higher latency. Commercially, such a design would be a non-starter, as existing workloads, that fit in the memory system, would run much slower than before. Ideally, we wanted the latency of DRAM and the capacity of PCM. Hence, our first key contribution, a **Hybrid Memory System**, combines the PCM memory with a DRAM buffer. The DRAM buffer would be much smaller in size (we used 1 GB DRAM for 32GB PCM), however, given memory locality, most of the accesses would be satisfied by the DRAM buffer, thus avoiding the high-latency accesses to PCM and the associated slowdown. Such a hybrid memory system offers a blueprint for incorporating any new memory technology that has higher latency than DRAM.

Our second key contribution was **Write-Filtering with DRAM Caches**. Our design assumed a page-granularity hardware-managed DRAM cache (this work was done in 2008 when the understanding of DRAM cache architectures was relatively low). While hits in the DRAM cache reduced both read-traffic and write-traffic to PCM memory, we paid special attention to further reducing the write traffic, due to the limited endurance. One of the write filtering schemes we proposed had line-level dirty bits in the DRAM cache tags to write only the lines that get modified. We also had a scheme that placed the incoming page directly into DRAM cache and kept the PCM-page stale. The DRAM cache entry was marked as dirty to ensure that the page is written to PCM on the eviction. This avoided an extra write of install to PCM for dirty pages.

The third key contribution of the paper was **Fine-Grained Wear Leveling.** We observed that some lines (such as line-0) were written more frequently than other lines in the page. We proposed a simple scheme that used a single counter per page to make write traffic uniform. We developed an analytical model to show the impact of non-uniform traffic on system lifetime and showed that our proposed design can increase system lifetime from 3 years to 9.7 years. Our evaluations showed that our design provided 3x speedup, coming within 10% of an idealized design that had 4x the DRAM capacity.

## IV. The Initial Reception and Impact

There were three PCM papers at ISCA: two of them proposed to replace DRAM with PCM, whereas our paper was the only one that argued not for the replacement of DRAM with PCM but rather configuring both DRAM and PCM in a hybrid fashion. Given that memory latency matters, the higher latency of emerging memories over the last decade makes it clear that any design that replaces DRAM with another technology of higher latency will cause a significant slowdown for existing workloads (that fit within memory), making it not a viable choice. The hybrid memory design proposed in our paper would go on to become the de-facto design for both academic studies and industrial proposals. For example, the 3D X-Point memories introduced by Intel were not meant to replace DRAM but rather to augment DRAM (for latency) with NVM (for capacity), forming a hybrid memory system that offers both low-latency and high-capacity.

While ISCA-2009 was the first top-tier architecture conference to have PCM papers, this topic would go on to become quite popular over the ensuing 3-4 years, with almost each top-tier architecture conference having 1-2 sessions on emerging memory technologies (PCM or NVM). The focused topics included low-cost wear leveling (with our IBM team proposing the Start-Gap algorithm), mitigating high write latency (write-cancellation and write pausing), trade-offs in Multi-Level Cells (morphable memory system), error correction (e.g. ECP, PAYG, and drift tolerance schemes), and reduced read latency (early-read and turbo-reads). Together, the solutions produced by the research in the architecture community significantly changed the landscape of PCM-based memory systems, transforming this relatively unknown and quirky technology into something that became well-understood with practical solutions to mitigate the shortcomings.

## V. The Legacy

This paper has received significant recognition for its contributions. It received the 2019 NVMW Persistent Impact Prize with the citation ”in recognition of its contribution to hybrid memory systems that combine phase change memory and DRAM. It was one of the very first papers to propose such a system and describe how careful design can overcome the limitations of both technologies to build fast, scalable, reliable memory systems. It is especially exciting to see how the paper anticipated and influenced the structure of many subsequent proposals for hybrid memory systems.” With nearly 1800 citations, this is currently one of the highest-cited papers in memory systems, and top-15 highly cited papers of all ISCA.

While PCM and subsequently 3D-XPoint have not enjoyed much commercial success, the device community continues to explore new technologies (FeRAM and variants of MRAM) with the potential to provide lower-cost memory. These technologies tend to have similar challenges as PCM making the solutions proposed in our paper applicable for them as well. As DRAM faces scalability and reliability challenges (e.g. Rowhammer), there is still a significant need for new scalable memory technologies to enable future large-capacity systems.