

Phase Change Memory (PCM): A New Memory Technology to Enable New Memory Usage Models

Sean Eilert

Director of Architecture, Principal Engineer

Mark Leinwander

Systems Manager, Principal Engineer

Giuseppe Crisenza

Vice President of Strategic Alliances

Micron Technology, Inc.

June 23, 2011

Key Design Requirements for Electronic Systems

Phase change memory (PCM) technology directly addresses the needs of today's electronic systems.

Density

With the convergence of consumer, computer, and communication electronic systems, an exponential growth of code and data is occurring in all electronic systems. To accommodate this growth, memory densities must not only meet current needs but must demonstrate the ability to scale to larger densities as required over time.

Bandwidth and energy

In high-level "convergent" electronic systems, performance is measured in terms of higher bandwidth to speed up internet connection and lower power consumption to enhance mobile use.

The memory system design must support the increasing requirements of bandwidth and reduced power consumption. Nonvolatile solid state memory is the proven way to reduce power consumption, as demonstrated with traditional NOR Flash memory.

Memory subsystem architecture

Memory subsystem architecture is a key challenge for embedded systems designers. Memory parameters such as density, performance, packaging, and interfaces all play a significant role in system-level performance. With the variety of memory types available to system designers, it is viable to partition the memory subsystem according to the specific needs of the higher-level system and application components. In some cases, caching is a reasonable approach to achieve an appropriate balance of performance, power, and cost. In other cases, partitioning according to the unique characteristics of the underlying memories becomes a more reasonable approach. For example, putting bit-alterable content in a bit-alterable memory instead of attempting to manage bit-alterability in a block-alterable memory.



Bandwidth Partitioning

At a high level, there are three main bandwidth categories: Code, data streaming, and data storage.

Code: The main variable for code performance is read speed. Code depends on speed of execution utilizing one of the following modes: Execute-in-place (XIP) using NOR Flash memory to meet high bandwidth and fast random access reads; or store and download (SnD) using NAND Flash memory and DRAM for code densities larger than about 1Gb.

Data Streaming: The main factor for data streaming performance is programming speed. Data streaming is typically based on DRAM technology, but can be implemented using NAND Flash memory and DRAM for densities larger than 4Gb, mainly for density capability and reduced power consumption.

Data Storage: The main considerations for data storage performance are density and data retention. However, because density is growing exponentially, latencies between the different parts of the system have a strong impact on subsystem performance. Data storage usage models typically utilize NAND Flash for densities between 4Gb and 100Gb. Given the strong correlation between cycling and data retention, systems that utilize NAND to achieve high write performance are often faced with the difficulty of ensuring adequate data retention to endure long periods of inactivity.

Comparison of High-Density Memory Technologies

Attributes	DRAM	PCM	NAND	MIC NAND	HDD
Nonvolatile	No	Yes	Yes	Yes	Yes
Erase Required	Bit	Bit	Block	Block	Sector
Software	Simple	Simple	Complex	Very Complex	Simple
Power	~W/GB	100–500 mW/die	~100 mW/die	~100 mW/die	~10W
Write Bandwidth	~GB/s	1–100+ MB/s/die	10–100 MB/s/die	~10 MB/s/die	200–400 MB/s
Write Latency	~20–50ns	~1 μ s	~100 μ s	~800 μ s	~10ms
Write Energy	~0.1nJ/b	<1 nJ/b	0.1-1 nJ/b	<1 nJ/b	>10 nJ/b
Read Latency	50ns	50–100 ns	10-25 μ s	25-50 μ s	~10ms
Read Energy	~0.1nJ/b	<<1 nJ/b	<<1 nJ/b	<<1 nJ/b	>10 nJ/b
Idle Power	~W/GB	<<0.1W	<<0.1W	<<0.1W	<10W
Endurance	–	10 ⁸	10 ⁵ –10 ⁴	10 ⁴ –?	–
Data Retention	ms	Not <i>f</i> (cycles)	<i>f</i> (cycles)	<i>f</i> (cycles)	Not <i>f</i> (cycles)



PCM Scalability

System designers continue to face significant challenges to design reliable embedded and storage systems based on Flash memory. With each new generation, the capabilities of existing memory technologies degrade, requiring significant system-level changes to maintain system-level reliability and performance. Both NOR and NAND Flash rely on memory structures that become increasingly difficult to shrink at smaller lithographies. This scaling effect is often referred to as Moore's Law, where memory densities double with each smaller generation. PCM, however, is based on a physical state change of a chalcogenide material, commonly referred to as GST. Chalcogenide films have already been proven to have stable characteristics to a 5nm node¹. As the PCM memory cell shrinks, the volume of GST material involved in the state change shrinks, resulting in reduced power consumption or higher write performance. This unique feature of PCM technology supports the promise of scalability beyond that of other memory technologies.

PCM in Embedded Systems

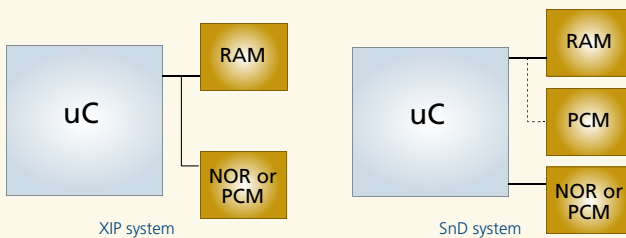


Figure 1: SnD and XIP System Architectures

A common use of memory in embedded systems is code storage. Systems requiring a relatively small amount of memory, less than approximately 2Gb, are architected such that code is executed directly from the NOR Flash (XIP). This memory is often also used as storage memory for an embedded file system. DRAM is often used in these systems as a scratchpad memory.

In these types of systems, PCM can be used as a code execution memory. With its bit-alterable feature, PCM is able to displace some or all of the DRAM required in the system. (See Figure 1.)

In SnD memory systems, PCM can reduce the density requirements for DRAM while fulfilling the density requirement of the NAND Flash. At the same time, the presence of PCM simplifies and improves the performance of file systems stored in the PCM due to the bit-alterability and low latency features.

PCM in Wireless Systems

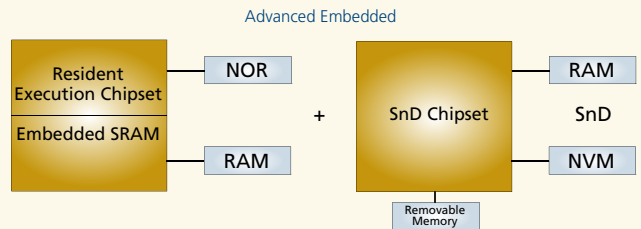


Figure 2: PCM in High-End Wireless Systems

It is common to find nearly independent subsystems for baseband and application processing in wireless systems. At the highest level, these can be considered independent embedded systems. Generally speaking, both subsystems have the need for a resident execution memory and for storage of small data structures. In many cases, the application subsystem is also expected to store and perform operations on larger multimedia content. (See Figure 2.)

The low read latencies and fast memory overwrite capabilities of PCM make it an ideal nonvolatile XIP solution that scales from low- to high-density wireless solutions. With read latencies that are slower but on the same order of magnitude as the latencies of DRAM, albeit on smaller page sizes, PCM can serve as an outstanding code execution memory and outstanding read-mostly memory for all but the most frequently manipulated data structures. The bit-alterability of PCM eliminates the need for block erase, which reduces the DRAM requirements even further, resulting in a lower cost memory subsystem.

PCM promises a scalable memory subsystem solution that provides the best overall cost while meeting the increasing performance demands of high-end, multimedia wireless devices.



PCM in Solid State Storage Subsystems

Managing NAND Flash in solid state storage (SSD) subsystems is a challenge due to the block-alterable nature of the NAND technology. It is also challenging to handle increasing levels of error management required when the memory is heavily program/erase cycled or frequently read.

PCM can be used in SSD systems to store frequently accessed pages and to store those elements which are more easily managed when manipulated in place. Examples of these types of elements include parity bits for data stored in NAND; bad block tables; and block and page mapping tables. In this scenario, a small amount of PCM could be used to enhance the manageability of NAND. (See Figure 3.) By minimizing the stress on the NAND memory, higher-density MLC NAND is enabled, thus leveraging the capability of PCM to lower the cost of the NAND Flash in the subsystem. This caching with PCM will improve the performance and reliability of the subsystem.

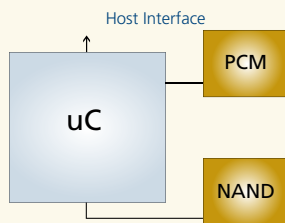


Figure 3: Hybrid Solid State Storage

Additionally, when erased pages are scattered across many blocks (near full state), PCM can provide further reliability improvement. Managing a block-alterable memory in a near-full state implies that multiple erase cycles are likely required to free space to store the new data being written to the device. This increases the number of cycles on the device and further accelerates the time until the maximum endurance limits are reached.

The bit-alterable nature of PCM solves the issue of increased WRITE cycles when the device is full. Higher endurance of PCM addresses the needs of these systems when heavy use is expected.

micron.com

Products are warranted only to meet Micron's production data sheet specifications. Products and specifications are subject to change without notice.

Micron and the Micron logo are trademarks of Micron Technology, Inc. All other trademarks are the property of their respective owners. ©2011 Micron Technology, Inc. All rights reserved. 06/23/11 EN.L

PCM in Computing Platforms

As a volatile memory, DRAM consumes power to simply maintain the contents of the memory. As a nonvolatile memory, PCM banks can be turned off when they are not in use to provide reduced power in idle states. More importantly, turning off the banks decouples the relationship between density and power consumption. This results in a PCM subsystem density envelope that is not limited by the power envelope constraints of that system. In addition to nonvolatility, PCM offers endurance and write latencies that are compelling for this type of application. This is a key advantage over read-mostly solutions that have been attempted until now.

Conclusion

PCM can be exploited by the memory system and by the convergence of consumer, computer, and communication electronic systems. The caching of existing memory technologies and reduction in overall system cost and complexity will be compelling motivation for PCM adoption. Bandwidth will drive the sustaining side of PCM in code and data transfer applications, while reduced power dissipation will provide further value.

PCM is today's memory breakthrough. Like Flash, PCM is a nonvolatile memory that can store bits even without a power supply. But unlike Flash, data can be written to cells much faster, at rates comparable to the dynamic and static random access memory (DRAM and SRAM) used in all computers and cell phones today. Quite simply, PCM blends together the best attributes of NOR Flash, NAND Flash, EEPROM, and RAM—delivering a new category of memory for new usage models.

For more information on PCM, please visit micron.com or contact your local Micron sales representative.

References:

¹ PCM scalability as referenced by C. D. Wright, M. M. Aziz, M. Armand, S. Senkader, and W. Yu, "Can We Reach Tbit/sq.in Storage Densities with Phase-Change Media," EPCOS 2004, and C. Lam, SRC NVM Forum 2004.

