Evaluating Phase Change Memory for Enterprise Storage Systems
Micron provided a prototype SSD built with 45 nm 1 Gbit Phase Change Memory
Part I. PCM performance
Flash Memory vs. Phase Change Memory

Flash memory traps charges to remember bits

PCM melts and cools material to remember bits

Temperature

Reset (**Amorphous phase**)

Set (**Crystalline phase**)

Time
PCM write latency

*ISCA’09, SOSP’09, FAST’11: 150 ns was used as PCM write latency*

**FAST’11 2015 Projections:**

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read/Write Latency ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flash SSD (SLC)</td>
<td>25,000 200,000</td>
</tr>
<tr>
<td>DRAM (DIMM)</td>
<td>55 55</td>
</tr>
<tr>
<td>PCM</td>
<td>48 150</td>
</tr>
</tbody>
</table>
PCM write throughput and power consumption

PCM chips have limited write throughput because of limited power budget
26.1

An 8Mb Demonstrator for High-Density
1.8V Phase-Change Memories

F. Bedeschi, C. Resta, O. Khouri, E. Buda, L. Costa, M. Ferraro, F. Pellizzer,

STMicroelectronics, MPG and Central R&D
Via C. Olivetti, 2 - 20041 - Agrate Brianza (Milano) - Italia
Tel.: +39-039-6036123, E-mail: ferdinando.bedeschi@st.com

Abstract

An 8Mb Non-Volatile Memory Demonstrator incorporating a novel 0.32 μm² Phase-Change Memory (PCM) cell using a Bipolar Junction Transistor (BJT) as selector and integrated into a 3V 0.18 μm CMOS technology is presented. Realistically large 4Mb tiles with a voltage regulation scheme that allows fast bitline precharge and sense are proposed. An innovative approach that minimizes the array leakage has been used to verify the feasibility of (GST) [5]. The cell formation modules are placed between the CMOS front-end (FEOL) and back-end (BEOL) and all the additional process steps required for the cell formation are compliant with a standard 0.18 μm CMOS technology. The basic CMOS parameters are reported in Tab.1. In order to integrate the PCM cell with the BJT selector, the chalcogenide alloy GST is patterned together with an AlCu Metal0 to form the subbitlines. The heater element is connected to the BJT p-Emitter, while the n-Base is contacted to the Metal1 wordline [5].
along the sub-bitlines). A RESET pulse of 40 ns and a SET of 150 ns have been demonstrated. In order to minimize the parasitic voltage drop along the wordline, only a pair of bits of a byte is programmed at the same time in the RESET state (1 bit per half-tile), while 4 bits are SET at the same time (2 bits per half-tile), considering the lower value of the SET current. On the 8Mb Demonstrator, where the 8 bits are arranged inside the same tile, the Write throughput demonstrated is 2.5MB/s, limited by the 200 ns SET time (SET pulse + 50 ns of circuitry delay).
Write latency and throughput comparison

- Write latency
  - PCM SET operation time: 200 ns
  - Flash page program time: 200 μs

- Write throughput per chip
  - PCM: 4 bits / 200 ns = 2.5 MB/s
  - Flash: 4 KB / 200 μs = 20.5 MB/s
## Table I. Quantitative Parameters of PCM

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DRAM</th>
<th>NAND Flash</th>
<th>PCM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>[Boboila and Desnoyers 2010]</td>
<td>[Atwood 2010] [Pirovano et al. 2004b]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Javanifard et al. 2008]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[MicronFlash 2008]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Nobunaga et al. 2008]</td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>3X nm</td>
<td>2X nm</td>
<td>&lt;1X nm</td>
</tr>
<tr>
<td>[Samsung 2011]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read Latency</td>
<td>60ns</td>
<td>25–200us</td>
<td>50–100ns</td>
</tr>
<tr>
<td>Write Speed</td>
<td>~1Gb/s</td>
<td>2.5 MB/s</td>
<td>~100MB/s</td>
</tr>
<tr>
<td>Endurance</td>
<td>N/A</td>
<td>10^3 to 10^5</td>
<td>10^6 to 10^8 [Atwood 2010], 10^{11}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Pirovano et al. 2004b]</td>
</tr>
</tbody>
</table>
Clarifications about PCM performance

- Write latency ≠ Write throughput
- 150 ns SET time = Material performance
- Material performance ≠ Chip performance
- Chip performance ≠ SSD performance
All PCM SSD vs. eMLC SSD

- 45 nm PCM
- 64 GB capacity

- eMLC NAND flash
- 1.8 TB capacity
Fine-grained I/O latency measurement

<table>
<thead>
<tr>
<th>Device</th>
<th>Driver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linux</td>
<td>Storage Software stack</td>
</tr>
<tr>
<td>Workload Generator</td>
<td></td>
</tr>
<tr>
<td>Measure Fine-grained I/O latency</td>
<td></td>
</tr>
<tr>
<td>Device Driver</td>
<td></td>
</tr>
<tr>
<td>Statistics Collector</td>
<td></td>
</tr>
</tbody>
</table>
4KB Random Read Latency (5M samples)

- **Onyx**: 38 µs
  - Mean: 6.7 µs
  - Maximum: 54.7 ms
  - Standard deviation: 76.2 µs

- **PCM**:
  - Mean: 108.0 µs
  - Maximum: 194.9 µs
  - Standard deviation: 1.5 µs

- **eMLC**
  - Mean: 108.0 µs

*16x shorter average latency*
4KB Random Write Latency (1M samples)

PCM

Onyx: 179 µs

Mean 128.3 µs

Maximum 378.2 µs

Standard deviation 2.2 µs

37.1 µs 3.4x longer 128.3 µs

Mean 37.1 µs

Maximum 17.2 ms

Standard deviation 153.2 µs

eMLC
Summary of Performance Numbers

<table>
<thead>
<tr>
<th></th>
<th>PCM SSD</th>
<th>eMLC SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4KB Read Latency</td>
<td>6.7 μs</td>
<td>108.0 μs</td>
</tr>
<tr>
<td>4KB Write Latency</td>
<td>128.3 μs</td>
<td>37.1 μs</td>
</tr>
</tbody>
</table>

- For read, PCM SSD is about 16x faster
- For write, PCM SSD is about 3.4x slower than eMLC SSD using DRAM as write buffer
Part II. Storage simulation

“Will PCM be useful for enterprise storage systems?”
### PCM / Flash / HDD multi-tiered storage

<table>
<thead>
<tr>
<th></th>
<th>PCM SSD</th>
<th>eMLC SSD</th>
<th>15K RPM HDD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4KB Read Latency</strong></td>
<td>6.7 μs</td>
<td>108.0 μs</td>
<td>5 ms</td>
</tr>
<tr>
<td><strong>4KB Write Latency</strong></td>
<td>128.3 μs</td>
<td>37.1 μs</td>
<td>5 ms</td>
</tr>
<tr>
<td><strong>Normalized Cost</strong></td>
<td>24</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
Multi-tiered storage simulation

1. Assume multi-tiered storage made of \(X \%\) of PCM, \(Y \%\) of Flash, \(Z \%\) of HDD

2. Estimate Performance (IOPS)

3. Estimate Storage Cost (Normalized: HDD = 1)

4. Evaluation metric:

   \[
   \frac{\text{Estimated Performance (IOPS)}}{\text{Estimated Cost (\$)}}
   \]
Tiering simulation methods

- **Static-optimal data placement**
  - Complete knowledge about I/O workload
  - No data movement

- **Dynamic tiering**
  - Reactive data movement based on I/O traffic
Retail Store, 2012 June, one week duration

- Cumulative amount (%)
  - Portion (%) of total accessd capacity (16.1 TiB)
  - Amount of Read: 252.7 TiB
  - Amount of Write: 45.0 TiB

- Read
- Write
Retail Store: IOPS/$ with Dynamic tiering

Best: 2,757 @ PCM(22%), Flash(78%)
Retail Store

- HDD 100%: 200 IOPS/$
- Flash 100%: 1,713 IOPS/$
- PCM 100%: 1,661 IOPS/$
- PCM 30%: 3,220 IOPS/$ (Static optimal data placement)
- Flash 67% HDD 3%: 2,757 IOPS/$ (Dynamic tiering, 61% increase)

Comparing Static optimal data placement with Dynamic tiering shows an increase of 61%.
Bank

(a) CDF and I/O amount

(b) 3D IOPS/$ by dynamic tiering

(c) IOPS/$ for key configuration points

Telecommunication

(a) CDF and I/O amount

(b) 3D IOPS/$ by dynamic tiering

(c) IOPS/$ for key configuration points
Application Server

IBM Easy Tier Server
EMC XtremCache
NetApp FlashAccel
FusionIO IO Turbine

Flash Cache

Shared Storage
Application server side cache simulation

- IO by IO storage traces
  - From real customer’s production systems (over 24 hour duration)
  - Manufacturing, media, medical companies
- Cache Simulation methods
  - Write-through, LRU replacement
  - Evaluation metric: average read latency
- Performance parameters

<table>
<thead>
<tr>
<th></th>
<th>PCM</th>
<th>eMLC</th>
<th>Net. Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 KiB R. Lat.</td>
<td>6.7 μs</td>
<td>108.0 μs</td>
<td>919.0 μs</td>
</tr>
<tr>
<td>4 KiB W. Lat.</td>
<td>128.3 μs</td>
<td>37.1 μs</td>
<td>133.0 μs</td>
</tr>
<tr>
<td>Norm. Cost</td>
<td>4</td>
<td>1</td>
<td>–</td>
</tr>
</tbody>
</table>
Same cost cache configurations

<table>
<thead>
<tr>
<th>Flash 64G</th>
<th>PCM 16G</th>
<th>P 8G + F 32G</th>
<th>P 4G + F 48G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flash 128G</td>
<td>PCM 32G</td>
<td>P 16G + F 64G</td>
<td>P 8G + F 96G</td>
</tr>
<tr>
<td>Flash 256G</td>
<td>PCM 64G</td>
<td>P 32G + F 128G</td>
<td>P 16G + F 192G</td>
</tr>
</tbody>
</table>
Manufacturing: I/O distribution (24 hours)

- **Cumulative amount (%)**
- **Portion (%) of total accessed capacity (246.5 GiB)**

**Graph Details:**
- **Y-axis:** Cumulative amount (%)
- **X-axis:** Portion (%) of total accessed capacity (246.5 GiB)

**Legend:**
- **Write**
- **Read**

**Notations:**
- 3.8 TiB
- 1.1 TiB

**Data Points:**
- **Amount of Read:** 3.8 TiB
- **Amount of Write:** 1.1 TiB
Manufacturing: cache simulation results

Average read latency (μs)
Media

(a) CDF and I/O amount

(b) Average read latency

(c) Average read latency for even cost configurations

Medical

(a) CDF and I/O amount

(b) Average read latency

(c) Average read latency for even cost configurations
Summary and Conclusion
Summary

- Let’s be careful to pick “right performance number” for PCM
- Performance measurement results

<table>
<thead>
<tr>
<th></th>
<th>PCM SSD</th>
<th>eMLC SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4KB Read Latency</td>
<td>6.7 μs</td>
<td>108.0 μs</td>
</tr>
<tr>
<td>4KB Write Latency</td>
<td>128.3 μs</td>
<td>37.1 μs</td>
</tr>
</tbody>
</table>

- Storage simulation results show
  - About 12-66% improved IOPS/$ for tiered storage
  - Up to 35% reduced average read latency for server caching
Concluding question

Phase Change Memory for enterprise storage...

*Silver bullet or Snake oil?*

What do you think?
Additional experiments

- The PCM SSD was shipped to Almaden, finally
- Real experiments with Sysbench OLTP benchmark
  - Read-only workload
  - Single thread test / 8 thread test
Sysbench OLTP benchmark, read only 1 thread: eMLC SSD
Sysbench OLTP benchmark, read only 1 thread: eMLC SSD vs. PCM SSD
Sysbench OLTP benchmark, read only 8 thread: eMLC SSD
Sysbench OLTP benchmark, read only 8 thread: eMLC SSD vs. PCM SSD

![Graph showing Sysbench OLTP benchmark results for eMLC SSD and PCM SSD](image-url)
PCM caching over eMLC flash SSD

Transactions per second (TPS)

Time

PCM
Flash
flashcache/30G
SCE/30G
Thank you