Data Mapping for Higher Performance and Energy Efficiency in Multi-Level Phase Change Change Memory

HanBin Yoon*, Naveen Muralimanohar‡, Justin Meza*, Onur Mutlu*, Norm Jouppi‡

*Carnegie Mellon University, ‡HP Labs
Overview

• MLC PCM: Strengths and weaknesses
• Data mapping scheme for MLC PCM
  – Exploits PCM characteristics for lower latency
  – Improves data integrity
• Row buffer management for MLC PCM
  – Increases row buffer hit rate
• Performance and energy efficiency improvements
Why MLC PCM?

• Emerging high density memory technology
  – Projected 3-12× denser than DRAM

• Scalable DRAM alternative on the horizon
  – Access latency comparable to DRAM

• **Multi-Level Cell**: 1 of key strengths over DRAM
  – Further increases memory density (by 2×–4×)

• But MLC also has drawbacks

[1Lee+ ISCA’09]
Higher MLC Latencies and Energy

• MLC program/read operation is more complex
  – Finer control/detection of cell resistances
• Generally leads to higher latencies and energy
  – \(~2\times\) for reads, \(~4\times\) for writes (depending on tech. & impl.)
MLC Multi-bit Faults

- In MLC, single cell failure can lead to multi-bit faults
Motivation

• MLC PCM strength:
  – Scalable, dense memory

• MLC PCM weaknesses:
  – Higher latencies
  – Higher energy
  – Multi-bit faults
  – Endurance

Mitigate through *bit mapping schemes* and *row buffer management* based on the following observations.
Observation #1: Read Asymmetry

- Read latency depends on cell state
  - Higher cell resistance $\rightarrow$ higher read latency

[Qureshi+ ISCA’10]
Observation #1: Read Asymmetry

- MSB can be determined before read completes
- Quicker MSB read \(\iff\) group LSB & MSB separately
Observation #2: Program Asymmetry

- Program latency depends on cell state

[3] Joshi+ HPCA’11
Observation #2: Program Asymmetry

- Single-bit change reduces LSB program latency
- Quicker LSB prog. \(\leftarrow\) group LSB & MSB separately
Observation #3: Distributed Bit Faults

• Bit mapping affects distribution of bit faults
  – 1 cell failure: 2 faults in 1 block vs. 1 fault each in 2 blocks (ECC-wise better)
• Distributed faults ← group LSB & MSB separately
Idea #1: Bit-Decoupled Mapping

- Decoupled bit mapping scheme
  - Reduced read latency for MSB pages (read asym.)
  - Reduced program latency for LSB pages (prog asym.)
  - Distributed bit faults between LSB and MSB blocks
  - Worse endurance
Coalescing Writes

- Assuming spatial locality in writebacks
- Interleaving blocks facilitates write coalescing
- Improved endurance ↔ **interleave blocks between LSB & MSB**
Idea #2: LSB-MSB Block Interleaving

- LM-Interleaved (LMI) bit mapping scheme
  - Mitigates cell wear
Opportunity: Two latches per cell in row buffer
- Use single row buffer as two "page buffers"
Idea #3: Split Page Buffering (SPB)

- Increased row buffer hit rate
Evaluation Methodology

• Cycle-level x86 CPU-memory simulator
  – CPU: 8 out-of-order cores, 32 KB private L1 per core
  – L2: 512 KB shared per core, DRAM-Aware LLC Writeback\(^4,5\)
    – Dual channel DDR3 1066 MT/s, 2 ranks, aggregate PCM capacity 16 GB (2 bits per cell)
• Multi-programmed SPEC CPU2006 workloads
  – Misses per kilo-instructions > 10

\(^4\text{Lee+ UTA-TechReport’10; }^5\text{Stuecheli+ ISCA’10}\)
Comparison Points and Metrics

• **Baseline**: Coupled bit mapping
• **Decoupled**: Decoupled bit mapping
• **LMI-4**: LSB-MSB interleaving every 4 blocks
• **LMI-16**: LSB-MSB interleaving every 16 blocks

• **Weighted speedup (performance)** = sum of thread speedups versus when run alone
• **Max slowdown (fairness)** = highest slowdown experienced by any thread
Performance

Baseline  Decoupled  LMI-4  LMI-16

Decoupled schemes benefit from reduced read latency (MSB) & program latency (LSB)
Fairness

- Baseline
- Decoupled
- LMI-4
- LMI-16

Individual thread speedups and increased row buffer hit rate
Energy Efficiency

Baseline  Decoupled  LMI-4  LMI-16

Lower read energy (dominant case) due to exploiting read asymmetry
Memory Lifetime

5-year lifespan feasible for system design? Point of on-going research...
Conclusion

- MLC PCM is a scalable, dense memory tech.
  - Exhibits higher latency and energy compared to SLC
1. LSB-MSB decoupled bit mapping
   - Exploits read asymmetry & program asymmetry
   - Distributes multi-bit faults
2. LSB-MSB block interleaving
   - Mitigates cell wear
3. Split page buffering
   - Increases row buffer hit rate
- Enhances perf. and energy eff. of MLC PCM
Thank you! Questions?