Scalable Many-Core Memory Systems
Lecture 3, Topic 2: Emerging Technologies and Hybrid Memories

Prof. Onur Mutlu
http://www.ece.cmu.edu/~omutlu
onur@cmu.edu
HiPEAC ACACES Summer School 2013
July 17, 2013
What Will You Learn in This Course?

- Scalable Many-Core Memory Systems
  - July 15-19, 2013

- Topic 1: Main memory basics, DRAM scaling
- Topic 2: Emerging memory technologies and hybrid memories
- Topic 3: Main memory interference and QoS
- Topic 4 (unlikely): Cache management
- Topic 5 (unlikely): Interconnects

Major Overview Reading:
Readings and Videos
Course Information

- **Website for Course Slides and Papers**
  - [http://users.ece.cmu.edu/~omutlu/acaces2013-memory.html](http://users.ece.cmu.edu/~omutlu/acaces2013-memory.html)
  - [http://users.ece.cmu.edu/~omutlu](http://users.ece.cmu.edu/~omutlu)
  - Lecture notes and readings are uploaded

- **My Contact Information**
  - Onur Mutlu
  - onur@cmu.edu
  - [http://users.ece.cmu.edu/~omutlu](http://users.ece.cmu.edu/~omutlu)
  - +1-512-658-0891 (my cell phone)
  - Find me during breaks and/or email *any* time.
Memory Lecture Videos

- Memory Hierarchy (and Introduction to Caches)
  - [http://www.youtube.com/watch?v=JBdfZ5i21cs&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=22](http://www.youtube.com/watch?v=JBdfZ5i21cs&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=22)

- Main Memory
  - [http://www.youtube.com/watch?v=ZLCy3pG7Rc0&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=25](http://www.youtube.com/watch?v=ZLCy3pG7Rc0&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=25)

- Memory Controllers, Memory Scheduling, Memory QoS
  - [http://www.youtube.com/watch?v=ZSotvL3WXmA&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=26](http://www.youtube.com/watch?v=ZSotvL3WXmA&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=26)
  - [http://www.youtube.com/watch?v=1xe2w3_NzmI&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=27](http://www.youtube.com/watch?v=1xe2w3_NzmI&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=27)

- Emerging Memory Technologies
  - [http://www.youtube.com/watch?v=LzfOghMKyA0&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=35](http://www.youtube.com/watch?v=LzfOghMKyA0&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=35)

- Multiprocessor Correctness and Cache Coherence
  - [http://www.youtube.com/watch?v=U-VZKMgItDM&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=32](http://www.youtube.com/watch?v=U-VZKMgItDM&list=PL5PHm2jkkXmidJ0d59REog9jDnPDTG6I&index=32)
Readings for Topic 1 (DRAM Scaling)

Readings for Topic 2 (Emerging Technologies)

Readings for Topic 3 (Memory QoS)

Readings for Topic 3 (Memory QoS)

- Ebrahimi et al., “Parallel Application Memory Scheduling,” MICRO 2011.
Readings in Flash Memory


Online Lectures and More Information

- **Online Computer Architecture Lectures**
  - [http://www.youtube.com/playlist?list=PL5PHm2jkkXmidJOD59REog9jDnPDTG6IJ](http://www.youtube.com/playlist?list=PL5PHm2jkkXmidJOD59REog9jDnPDTG6IJ)

- **Online Computer Architecture Courses**
  - Advanced: [http://www.ece.cmu.edu/~ece742/doku.php](http://www.ece.cmu.edu/~ece742/doku.php)

- **Recent Research Papers**
  - [http://users.ece.cmu.edu/~omutlu/projects.htm](http://users.ece.cmu.edu/~omutlu/projects.htm)
  - [http://scholar.google.com/citations?user=7XyGUGkAAAAJ&hl=en](http://scholar.google.com/citations?user=7XyGUGkAAAAJ&hl=en)
Emerging Memory Technologies
Agenda

- Major Trends Affecting Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies
- Conclusions
- Discussion
Major Trends Affecting Main Memory (I)

- Need for main memory capacity and bandwidth increasing

- Main memory energy/power is a key system design concern

- DRAM technology scaling is ending
Trends: Problems with DRAM as Main Memory

- Need for main memory capacity and bandwidth increasing
  - DRAM capacity hard to scale

- Main memory energy/power is a key system design concern
  - DRAM consumes high power due to leakage and refresh

- DRAM technology scaling is ending
  - DRAM capacity, cost, and energy/power hard to scale
Agenda

- Major Trends Affecting Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies
- Conclusions
- Discussion
Requirements from an Ideal Memory System

- Traditional
  - Enough capacity
  - Low cost
  - High system performance (high bandwidth, low latency)

- New
  - Technology scalability: lower cost, higher capacity, lower energy
  - Energy (and power) efficiency
  - QoS support and configurability (for consolidation)
Requirements from an Ideal Memory System

- **Traditional**
  - Higher capacity
  - Continuous low cost
  - High system performance (higher bandwidth, low latency)

- **New**
  - Technology scalability: lower cost, higher capacity, lower energy
  - Energy (and power) efficiency
  - QoS support and configurability (for consolidation)

Emerging, resistive memory technologies (NVM) can help
Review: Solutions to the DRAM Scaling Problem

- Two potential solutions
  - Tolerate DRAM (by taking a fresh look at it)
  - Enable emerging memory technologies to eliminate/minimize DRAM

- Do both
  - Hybrid memory systems
Solution 1: Tolerate DRAM

- Overcome DRAM shortcomings with
  - System-DRAM co-design
  - Novel DRAM architectures, interface, functions
  - Better waste management (efficient utilization)

- Key issues to tackle
  - Reduce refresh energy
  - Improve bandwidth and latency
  - Reduce waste
  - Enable reliability at low cost

Solution 2: Emerging Memory Technologies

- Some emerging resistive memory technologies seem more scalable than DRAM (and they are non-volatile)
- Example: Phase Change Memory
  - Expected to scale to 9nm (2022 [ITRS])
  - Expected to be denser than DRAM: can store multiple bits/cell

- But, emerging technologies have shortcomings as well
  - Can they be enabled to replace/augment/surpass DRAM?

Hybrid Memory Systems

Hardware/software manage data allocation and movement to achieve the best of multiple technologies

Yoon, Meza et al., “Row Buffer Locality Aware Caching Policies for Hybrid Memories,” ICCD 2012 Best Paper Award.
Agenda

- Major Trends Affecting Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies
- Conclusions
- Discussion
The Promise of Emerging Technologies

- Likely need to replace/augment DRAM with a technology that is
  - Technology scalable
  - And at least similarly efficient, high performance, and fault-tolerant
    - or can be architected to be so

- Some emerging resistive memory technologies appear promising
  - Phase Change Memory (PCM)?
  - Spin Torque Transfer Magnetic Memory (STT-MRAM)?
  - Memristors?
  - And, maybe there are other ones
  - Can they be enabled to replace/augment/surpass DRAM?
Agenda

- Major Trends Affecting Main Memory
- Requirements from an Ideal Main Memory System
- **Opportunity: Emerging Memory Technologies**
  - Background
  - PCM (or Technology X) as DRAM Replacement
  - Hybrid Memory Systems
- Conclusions
- Discussion
Charge vs. Resistive Memories

- Charge Memory (e.g., DRAM, Flash)
  - Write data by capturing charge Q
  - Read data by detecting voltage V

- Resistive Memory (e.g., PCM, STT-MRAM, memristors)
  - Write data by pulsing current $dQ/dt$
  - Read data by detecting resistance R
Limits of Charge Memory

- Difficult charge placement and control
  - Flash: floating gate charge
  - DRAM: capacitor charge, transistor leakage

- Reliable sensing becomes difficult as charge storage unit size reduces
Emerging Resistive Memory Technologies

- PCM
  - Inject current to change material phase
  - Resistance determined by phase

- STT-MRAM
  - Inject current to change magnet polarity
  - Resistance determined by polarity

- Memristors
  - Inject current to change atomic structure
  - Resistance determined by atom distance
What is Phase Change Memory?

- Phase change material (chalcogenide glass) exists in two states:
  - Amorphous: Low optical reflexivity and high electrical resistivity
  - Crystalline: High optical reflexivity and low electrical resistivity

PCM is resistive memory: High resistance (0), Low resistance (1)
PCM cell can be switched between states reliably and quickly
How Does PCM Work?

- **Write:** change phase via current injection
  - **SET:** sustained current to heat cell above $T_{cryst}$
  - **RESET:** cell heated above $T_{melt}$ and quenched
- **Read:** detect phase via material resistance
  - amorphous/crystalline

---

![Diagram](image)

**SET (cryst)**
- Low resistance
- $10^3-10^4 \, \Omega$

**RESET (amorph)**
- High resistance
- $10^6-10^7 \, \Omega$

---

Photo Courtesy: Bipin Rajendran, IBM  
Slide Courtesy: Moinuddin Qureshi, IBM
Opportunity: PCM Advantages

- **Scales better than DRAM, Flash**
  - Requires current pulses, which scale linearly with feature size
  - Expected to scale to 9nm (2022 [ITRS])
  - Prototyped at 20nm (Raoux+, IBM JRD 2008)

- **Can be denser than DRAM**
  - Can store multiple bits per cell due to large resistance range
  - Prototypes with 2 bits/cell in ISSCC’08, 4 bits/cell by 2012

- **Non-volatile**
  - Retain data for >10 years at 85C

- **No refresh needed, low idle power**
Phase Change Memory Properties

- Surveyed prototypes from 2003-2008 (ITRS, IEDM, VLSI, ISSCC)
- Derived PCM parameters for F=90nm

<table>
<thead>
<tr>
<th>Parameter*</th>
<th>Horri⁶</th>
<th>Ahn¹²</th>
<th>Bedeschi¹³</th>
<th>Oh¹⁴</th>
<th>Pellizer¹⁵</th>
<th>Chen⁵</th>
<th>Kang¹⁶</th>
<th>Bedeschi⁹</th>
<th>Lee¹⁰</th>
<th>Lee²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process, F(nm)</td>
<td>**</td>
<td>120</td>
<td>180</td>
<td>120</td>
<td>90</td>
<td>**</td>
<td>100</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Array size (Mbytes)</td>
<td>**</td>
<td>64</td>
<td>8</td>
<td>64</td>
<td>**</td>
<td>**</td>
<td>256</td>
<td>256</td>
<td>512</td>
<td>**</td>
</tr>
<tr>
<td>Material</td>
<td>GST, N-d</td>
<td>GST, N-d</td>
<td>GST</td>
<td>GST</td>
<td>GST</td>
<td>GS, N-d</td>
<td>GST</td>
<td>GST</td>
<td>GST</td>
<td>GST, N-d</td>
</tr>
<tr>
<td>Cell size (μm²)</td>
<td>**</td>
<td>0.290</td>
<td>0.290</td>
<td>**</td>
<td>0.097</td>
<td>60 nm²</td>
<td>0.166</td>
<td>0.097</td>
<td>0.047</td>
<td>0.065 to 0.097</td>
</tr>
<tr>
<td>Cell size, F²</td>
<td>**</td>
<td>20.1</td>
<td>9.0</td>
<td>**</td>
<td>12.0</td>
<td>**</td>
<td>16.6</td>
<td>12.0</td>
<td>5.8</td>
<td>9.0 to 12.0</td>
</tr>
<tr>
<td>Access device</td>
<td>**</td>
<td>**</td>
<td>BJT</td>
<td>FET</td>
<td>BJT</td>
<td>**</td>
<td>FET</td>
<td>BJT</td>
<td>Diode</td>
<td>BJT</td>
</tr>
<tr>
<td>Read time (ns)</td>
<td>**</td>
<td>70</td>
<td>48</td>
<td>68</td>
<td>**</td>
<td>**</td>
<td>62</td>
<td>**</td>
<td>55</td>
<td>48</td>
</tr>
<tr>
<td>Read current (μA)</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>40</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Read voltage (V)</td>
<td>**</td>
<td>3.0</td>
<td>1.0</td>
<td>1.8</td>
<td>1.6</td>
<td>**</td>
<td>1.8</td>
<td>**</td>
<td>1.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Read power (μW)</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Read energy (pJ)</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>2.0</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Set time (ns)</td>
<td>100</td>
<td>150</td>
<td>150</td>
<td>180</td>
<td>**</td>
<td>80</td>
<td>300</td>
<td>**</td>
<td>400</td>
<td>150</td>
</tr>
<tr>
<td>Set current (μA)</td>
<td>200</td>
<td>**</td>
<td>300</td>
<td>200</td>
<td>**</td>
<td>55</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Set voltage (V)</td>
<td>**</td>
<td>**</td>
<td>2.0</td>
<td>**</td>
<td>**</td>
<td>1.25</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>1.2</td>
</tr>
<tr>
<td>Set power (μW)</td>
<td>**</td>
<td>**</td>
<td>300</td>
<td>**</td>
<td>**</td>
<td>34.4</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Set energy (pJ)</td>
<td>**</td>
<td>**</td>
<td>45</td>
<td>**</td>
<td>**</td>
<td>2.8</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>13.5</td>
</tr>
<tr>
<td>Reset time (ns)</td>
<td>50</td>
<td>10</td>
<td>40</td>
<td>10</td>
<td>**</td>
<td>60</td>
<td>50</td>
<td>**</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Reset current (μA)</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>90</td>
<td>600</td>
<td>300</td>
<td>600</td>
<td>300</td>
<td>600</td>
</tr>
<tr>
<td>Reset voltage (V)</td>
<td>**</td>
<td>**</td>
<td>2.7</td>
<td>1.8</td>
<td>1.6</td>
<td>**</td>
<td>1.6</td>
<td>**</td>
<td>1.6</td>
<td>**</td>
</tr>
<tr>
<td>Reset power (μW)</td>
<td>**</td>
<td>**</td>
<td>1620</td>
<td>**</td>
<td>**</td>
<td>80.4</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>480</td>
</tr>
<tr>
<td>Reset energy (pJ)</td>
<td>**</td>
<td>**</td>
<td>64.8</td>
<td>**</td>
<td>**</td>
<td>4.8</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>19.2</td>
</tr>
<tr>
<td>Write endurance</td>
<td>10⁷</td>
<td>10⁹</td>
<td>10⁶</td>
<td>**</td>
<td>10⁸</td>
<td>10⁴</td>
<td>**</td>
<td>10⁵</td>
<td>10⁵</td>
<td>10⁸</td>
</tr>
</tbody>
</table>

* BJ T: bipolar junction transistor; FET: field-effect transistor; GST: Ge₂Sb₂Te₅; MLC: multilevel cells; N-d: nitrogen doped.
** This information is not available in the publication cited.
Phase Change Memory Properties: Latency

- Latency comparable to, but slower than DRAM

Read Latency
- 50ns: 4x DRAM, 10^{-3}x NAND Flash

Write Latency
- 150ns: 12x DRAM

Write Bandwidth
- 5-10 MB/s: 0.1x DRAM, 1x NAND Flash
Phase Change Memory Properties

- **Dynamic Energy**
  - 40 uA Rd, 150 uA Wr
  - 2-43x DRAM, 1x NAND Flash

- **Endurance**
  - Writes induce phase change at 650°C
  - Contacts degrade from thermal expansion/contraction
  - $10^8$ writes per cell
  - $10^{-8}$x DRAM, $10^3$x NAND Flash

- **Cell Size**
  - 9-12F² using BJT, single-level cells
  - 1.5x DRAM, 2-3x NAND (will scale with feature size, MLC)
Phase Change Memory: Pros and Cons

- **Pros over DRAM**
  - Better technology scaling
  - Non volatility
  - Low idle power (no refresh)

- **Cons**
  - Higher latencies: ~4-15x DRAM (especially write)
  - Higher active energy: ~2-50x DRAM (especially write)
  - Lower endurance (a cell dies after ~$10^8$ writes)

- **Challenges in enabling PCM as DRAM replacement/helper:**
  - Mitigate PCM shortcomings
  - Find the right way to place PCM in the system
  - Ensure secure and fault-tolerant PCM operation
PCM-based Main Memory: Research Challenges

- Where to place PCM in the memory hierarchy?
  - Hybrid OS controlled PCM-DRAM
  - Hybrid OS controlled PCM and hardware-controlled DRAM
  - Pure PCM main memory

- How to mitigate shortcomings of PCM?

- How to minimize amount of DRAM in the system?

- How to take advantage of (byte-addressable and fast) non-volatile main memory?

- Can we design specific-NVM-technology-agnostic techniques?
PCM-based Main Memory (I)

- How should PCM-based (main) memory be organized?

  ![Diagram showing organization of PCM-based main memory]

- Hybrid PCM+DRAM [Qureshi+ ISCA’09, Dhiman+ DAC’09, Meza+ IEEE CAL’12]:
  - How to partition/migrate data between PCM and DRAM
Hybrid Memory Systems: Challenges

- **Partitioning**
  - Should DRAM be a cache or main memory, or configurable?
  - What fraction? How many controllers?

- **Data allocation/movement (energy, performance, lifetime)**
  - Who manages allocation/movement?
  - What are good control algorithms?
  - How do we prevent degradation of service due to wearout?

- **Design of cache hierarchy, memory controllers, OS**
  - Mitigate PCM shortcomings, exploit PCM advantages

- **Design of PCM/DRAM chips and modules**
  - Rethink the design of PCM/DRAM with new requirements
PCM-based Main Memory (II)

- How should PCM-based (main) memory be organized?

- **Pure PCM main memory** [Lee et al., ISCA’09, Top Picks’10]:
  - How to redesign entire hierarchy (and cores) to overcome PCM shortcomings
Aside: STT-RAM Basics

- **Magnetic Tunnel Junction (MTJ)**
  - Reference layer: Fixed
  - Free layer: Parallel or anti-parallel

- **Cell**
  - Access transistor, bit/sense lines

- **Read and Write**
  - Read: Apply a small voltage across bitline and senseline; read the current.
  - Write: Push large current through MTJ. Direction of current determines new orientation of the free layer.

Aside: STT MRAM: Pros and Cons

- **Pros over DRAM**
  - Better technology scaling
  - Non volatility
  - Low idle power (no refresh)

- **Cons**
  - Higher write latency
  - Higher write energy
  - Reliability?

- **Another level of freedom**
  - Can trade off non-volatility for lower write latency/energy (by reducing the size of the MTJ)
Agenda

- Major Trends Affecting Main Memory
- Requirements from an Ideal Main Memory System
- **Opportunity: Emerging Memory Technologies**
  - Background
  - PCM (or Technology X) as DRAM Replacement
  - Hybrid Memory Systems
- Conclusions
- Discussion
An Initial Study: Replace DRAM with PCM

  - Surveyed prototypes from 2003-2008 (e.g. IEDM, VLSI, ISSCC)
  - Derived “average” PCM parameters for F=90nm

### Density
- $9 - 12F^2$ using BJT
- $1.5 \times$ DRAM

### Latency
- 50ns Rd, 150ns Wr
- $4 \times, 12 \times$ DRAM

### Endurance
- $1E+08$ writes
- $1E-08 \times$ DRAM

### Energy
- $40 \mu A$ Rd, $150 \mu A$ Wr
- $2 \times, 43 \times$ DRAM
Results: Naïve Replacement of DRAM with PCM

- Replace DRAM with PCM in a 4-core, 4MB L2 system
- PCM organized the same as DRAM: row buffers, banks, peripherals
- 1.6x delay, 2.2x energy, 500-hour average lifetime

Architecting PCM to Mitigate Shortcomings

- Idea 1: Use multiple narrow row buffers in each PCM chip → Reduces array reads/writes → better endurance, latency, energy

- Idea 2: Write into array at cache block or word granularity → Reduces unnecessary wear
Results: Architected PCM as Main Memory

- 1.2x delay, 1.0x energy, 5.6-year average lifetime
- Scaling improves energy, endurance, density

Caveat 1: Worst-case lifetime is much shorter (no guarantees)
Caveat 2: Intensive applications see large performance and energy hits
Caveat 3: Optimistic PCM parameters?
Agenda

- Major Trends Affecting Main Memory
- Requirements from an Ideal Main Memory System
- **Opportunity: Emerging Memory Technologies**
  - Background
  - PCM (or Technology X) as DRAM Replacement
  - Hybrid Memory Systems
- Conclusions
- Discussion
Hybrid Memory Systems

Hardware/software manage data allocation and movement to achieve the best of multiple technologies

One Option: DRAM as a Cache for PCM

- PCM is main memory; DRAM caches memory rows/blocks
  - Benefits: Reduced latency on DRAM cache hit; write filtering
- Memory controller hardware manages the DRAM cache
  - Benefit: Eliminates system software overhead

Three issues:
- What data should be placed in DRAM versus kept in PCM?
- What is the granularity of data movement?
- How to design a low-cost hardware-managed DRAM cache?

Two idea directions:
- Locality-aware data placement [Yoon+, ICCD 2012]
- Cheap tag stores and dynamic granularity [Meza+, IEEE CAL 2012]