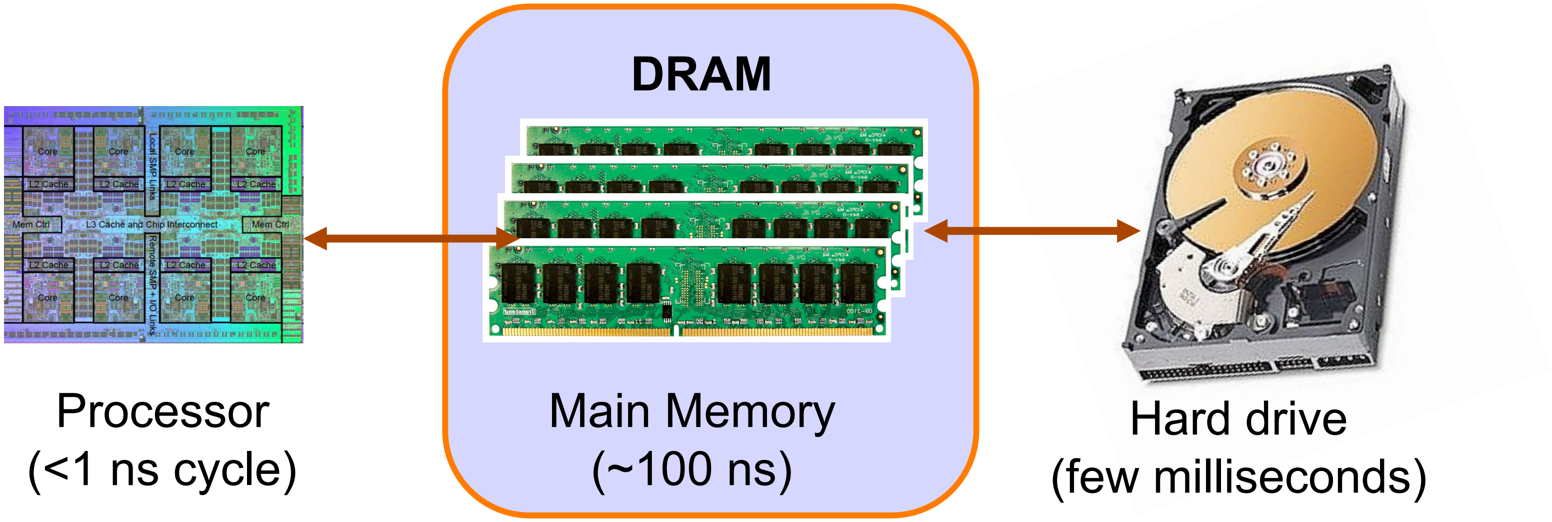
PCM (NVM) as Main Memory: Opportunities and Challenges

Onur Mutlu
Carnegie Mellon University

CMU PDL Retreat
October 25, 2010

The Main Memory System



DRAM

Processor
(<1 ns cycle)

Main Memory
(~100 ns)

Hard drive
(few milliseconds)

- Main memory is a critical component of all computing systems: server, mobile, embedded, desktop, sensor

- Main memory system must scale (in size, technology, efficiency, cost) to maintain performance growth and technology scaling benefits

State of the Main Memory System

- Recent technology, architecture, and application trends
  - lead to new requirements from the memory system
  - exacerbate old requirements from the memory system

- DRAM alone is (will be) unable to satisfy requirements

- Some emerging non-volatile memory technologies (e.g., PCM) appear promising to satisfy these requirements
  - and enable new opportunities

- We need to rethink the main memory system to enable emerging technologies

Talk Agenda

- Major Trends Affecting DRAM-Based Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies (PCM)
- Research Challenges: PCM as Main Memory
- Preliminary Ideas and Results
- Open Questions
- Summary

Major Trends Affecting Main Memory (I)

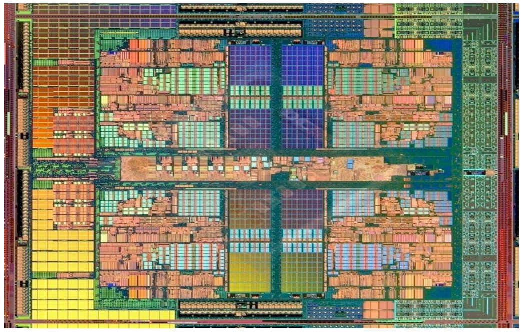
- Need for main memory capacity and bandwidth increasing

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

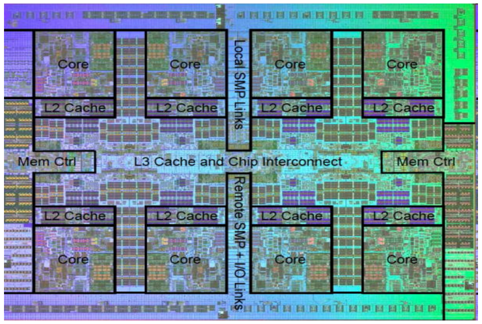
- DRAM technology scaling is ending

Demand for Memory Capacity

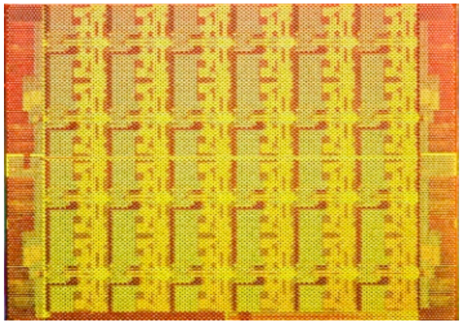
- More cores → More concurrency → Larger working set



AMD Barcelona: 4 cores



IBM Power7: 8 cores



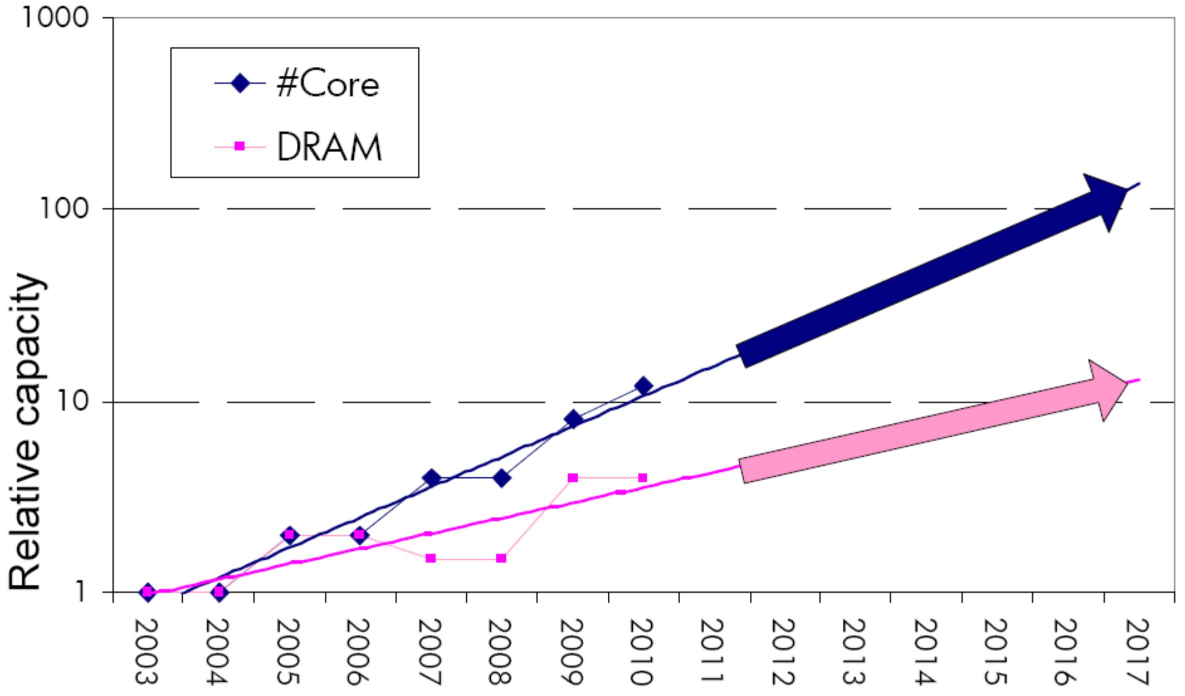
Intel SCC: 48 cores

- Emerging applications are data-intensive

- Many applications/virtual machines (will) share main memory
  - Cloud computing/servers: Consolidation to improve efficiency
  - GP-GPUs: Many threads from multiple parallel applications
  - Mobile: Interactive + non-interactive consolidation

The Memory Capacity Gap

Core count doubling ~ every 2 years
DRAM DIMM capacity doubling ~ every 3 years



Source: Lim et al., ISCA 2009.

- Memory capacity per core expected to drop by 30% every two years

Major Trends Affecting Main Memory (II)

- Need for main memory capacity and bandwidth increasing
  - Multi-core: increasing number of cores
  - Data-intensive applications: increasing demand/hunger for data
  - Consolidation: Cloud computing, GPUs, mobile

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

- DRAM technology scaling is ending

Major Trends Affecting Main Memory (III)

- Need for main memory capacity and bandwidth increasing

- Main memory energy/power is a key system design concern
  - IBM servers: ~50% energy spent in off-chip memory hierarchy [Lefurgy, IEEE Computer 2003]
  - DRAM consumes power when idle and needs periodic refresh

- DRAM technology scaling is ending

Major Trends Affecting Main Memory (IV)

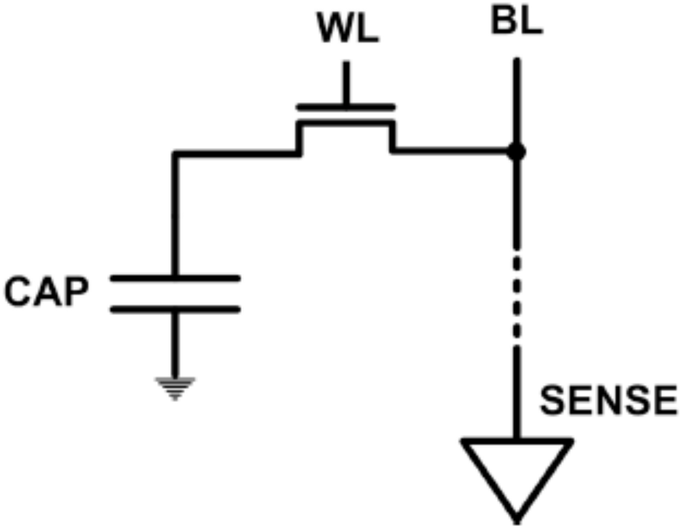
- Need for main memory capacity and bandwidth increasing

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

- DRAM technology scaling is ending
  - ITRS projects DRAM will not scale easily below 40nm
  - Scaling has provided many benefits:
    - higher capacity, higher density, lower cost, lower energy

The DRAM Scaling Problem

- DRAM stores charge in a capacitor (charge-based memory)
  - Capacitor must be large enough for reliable sensing
  - Access transistor should be large enough for low leakage and high retention time
  - Scaling beyond 40-35nm (2013) is challenging [ITRS, 2009]



- DRAM capacity, cost, and energy/power hard to scale

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

Talk Agenda

- Major Trends Affecting DRAM-Based Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies (PCM)
- Research Challenges: PCM as Main Memory
- Preliminary Ideas and Results
- Open Questions
- Summary

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

Talk Agenda

- Major Trends Affecting DRAM-Based Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies (PCM)
- Research Challenges: PCM as Main Memory
- Preliminary Ideas and Results
- Open Questions
- Summary

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?

Charge vs. Resistive Memories

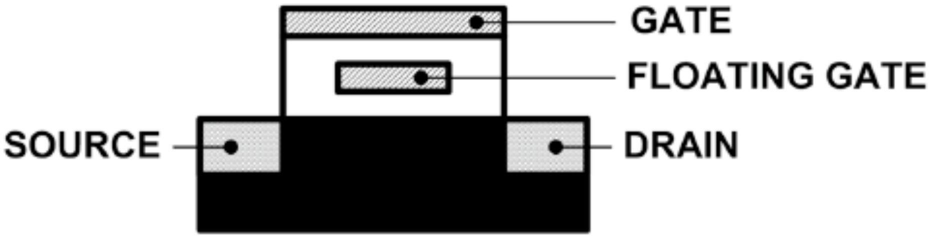
- Charge Memory (e.g., DRAM, Flash)
  - Write data by capturing charge Q
  - Read data by detecting voltage V

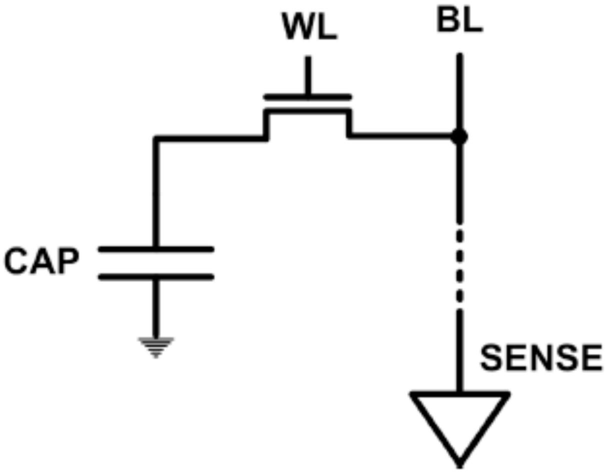
- Resistive Memory (e.g., PCM, STT-MRAM, memristors)
  - Write data by pulsing current dQ/dt
  - Read data by detecting resistance R

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

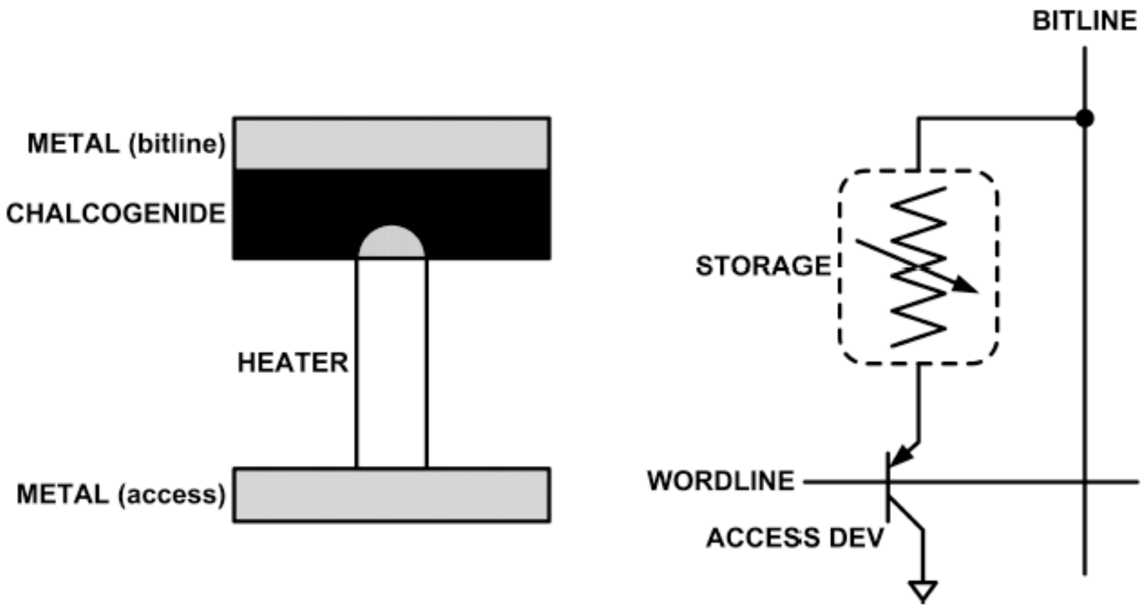
- PCM
  - Inject current to change material phase
  - Resistance determined by phase

- 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?

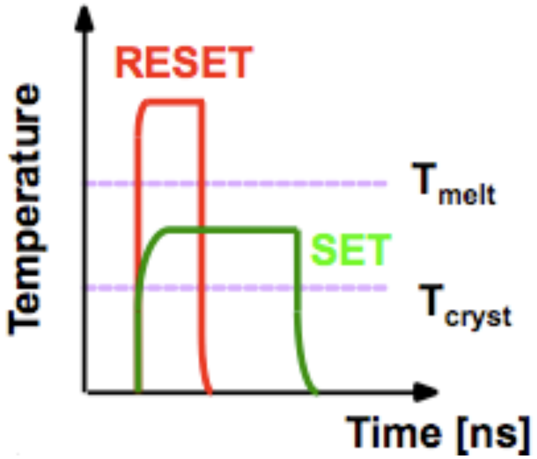
- Phase change material (chalcogenide glass) exists in two states:
  - Amorphous: Low optical reflexivity and high electrical resistivity
  - Crystalline: High optical reflexivity and low electrical resistivity

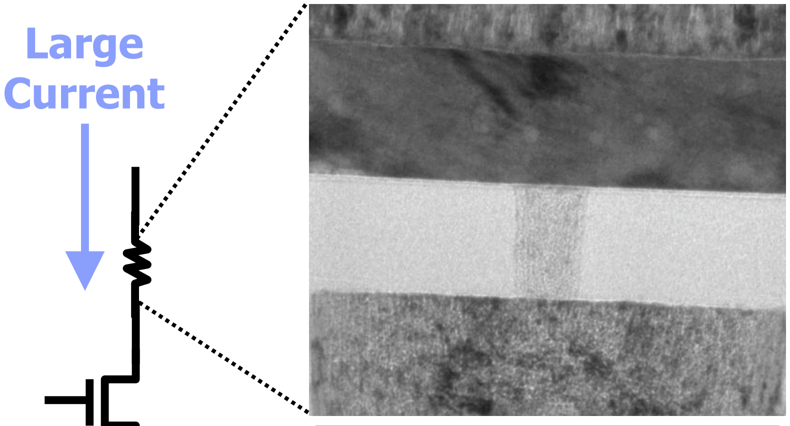


PCM is resistive memory: High resistance (0), Low resistance (1)
PCM cell can be switched between states reliably and quickly

How Does PCM Work?

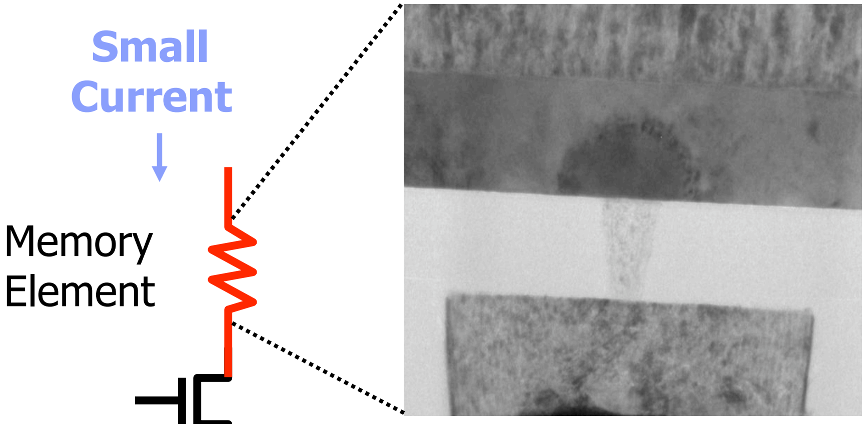
- Write: change phase via current injection
  - SET: sustained current to heat cell above T*cryst*
  - RESET: cell heated above T*melt* and quenched
- Read: detect phase via material resistance
  - amorphous/crystalline





SET (cryst)
Low resistance

10^3-10^4 Ω



RESET (amorph)
High resistance

10^6-10^7 Ω

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

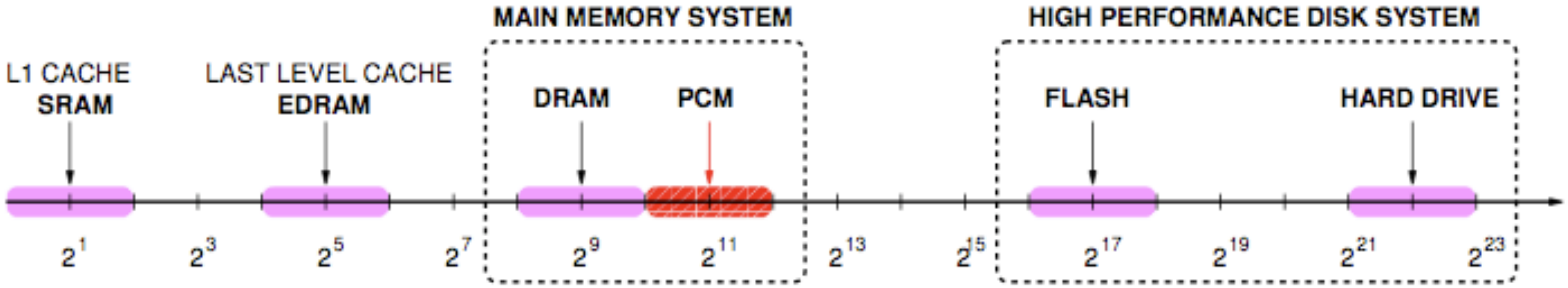
Phase Change Memory Properties

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

- Lee, Ipek, Mutlu, Burger, “Architecting Phase Change Memory as a Scalable DRAM Alternative,” ISCA 2009.

Phase Change Memory Properties: Latency

- Latency comparable to, but slower than DRAM



Typical Access Latency (in terms of processor cycles for a 4 GHz processor)

- 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 650C
  - Contacts degrade from thermal expansion/contraction
  - 10^8 writes per cell
  - 10^{-8}x DRAM, 10^3x NAND Flash

- Cell Size
  - 9-12F^2 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

Talk Agenda

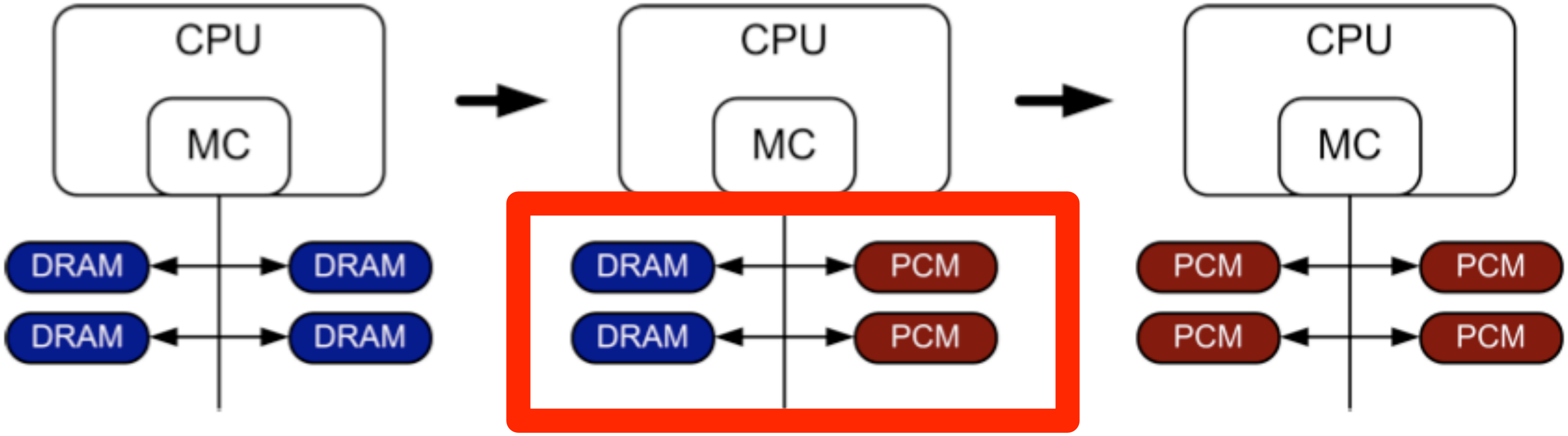
- Major Trends Affecting DRAM-Based Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies (PCM)
- Research Challenges: PCM as Main Memory
- Preliminary Ideas and Results
- Open Questions
- Summary

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?



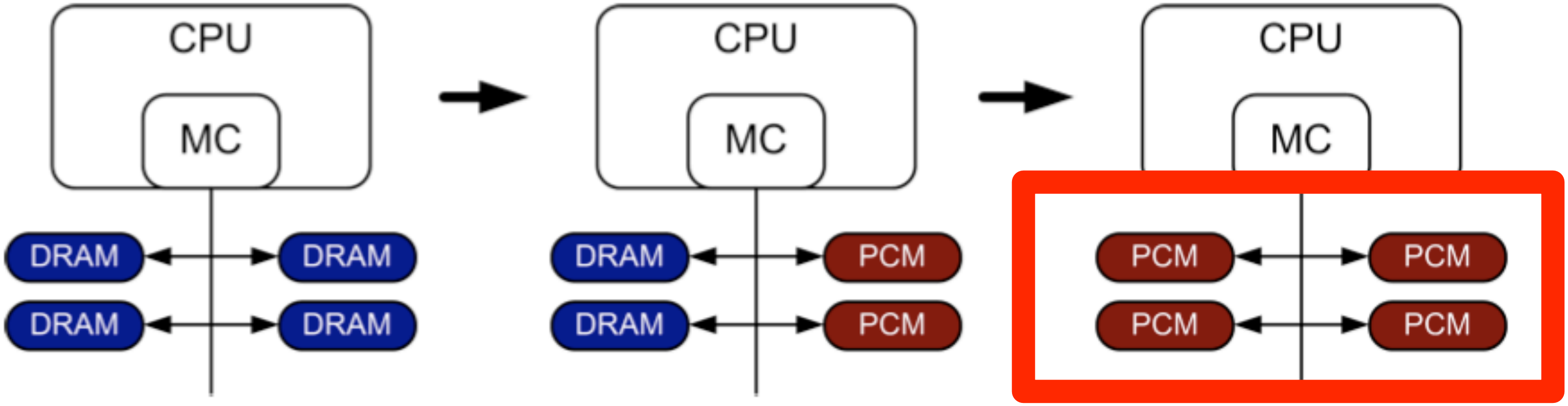
- Hybrid PCM+DRAM [Qureshi+ ISCA'09, Dhiman+ DAC'09]:
  - How to partition/migrate data between PCM and DRAM
  - Is DRAM a cache for PCM or part of main memory?
  - How to design the hardware and software
    - Exploit advantages, minimize disadvantages of each technology

Hybrid Memory Systems: Research 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?
    - Latency-critical, heavily modified → DRAM, otherwise PCM?
    - Preventing denial/degradation of service
- Design of cache hierarchy, memory controllers, OS
  - Mitigate PCM shortcomings
- Design of PCM/DRAM chips
  - Rethink the design of PCM/DRAM with new requirements

PCM-based Main Memory (II)

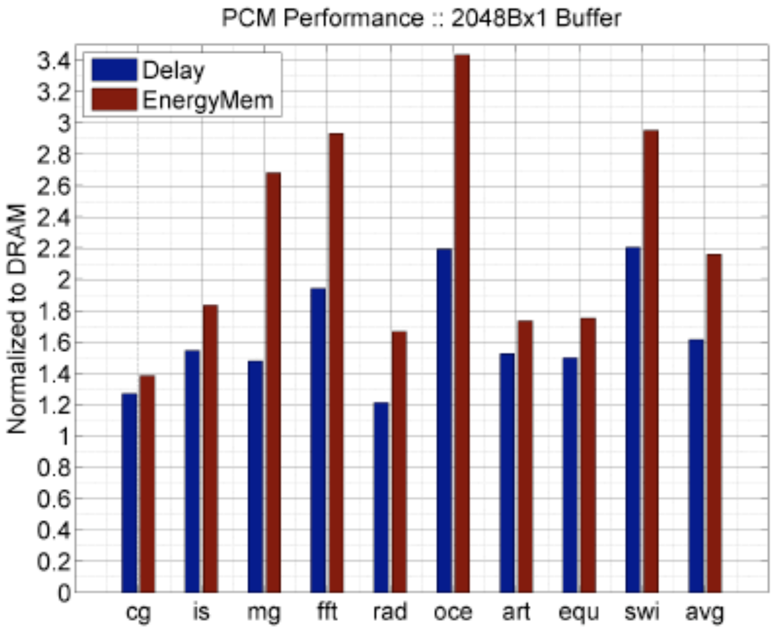
- How should PCM-based (main) memory be organized?

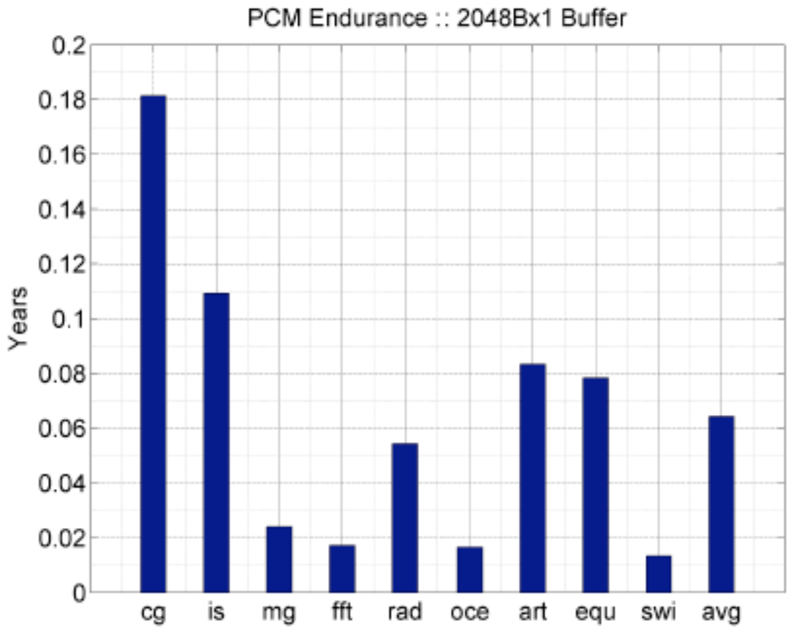


- Pure PCM main memory [Lee et al., ISCA'09, IEEE Micro'10]:
  - How to redesign entire hierarchy (and cores) to overcome PCM shortcomings
    - Latency, energy, endurance

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





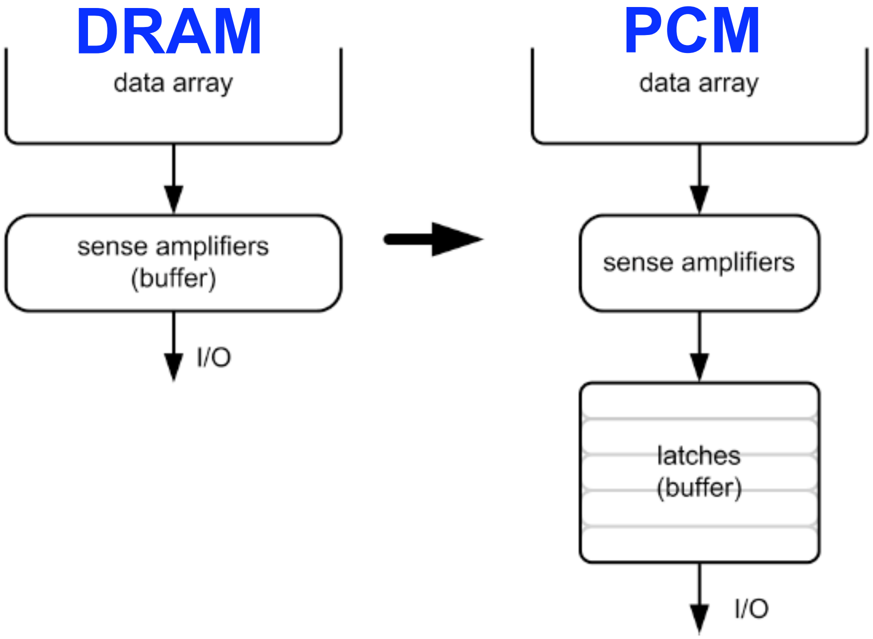
- Lee, Ipek, Mutlu, Burger, "Architecting Phase Change Memory as a Scalable DRAM Alternative," ISCA 2009.

Architecting PCM to Mitigate Shortcomings

- Idea 1: Use narrow row buffers in each PCM chip
  → Reduces write energy, peripheral circuitry

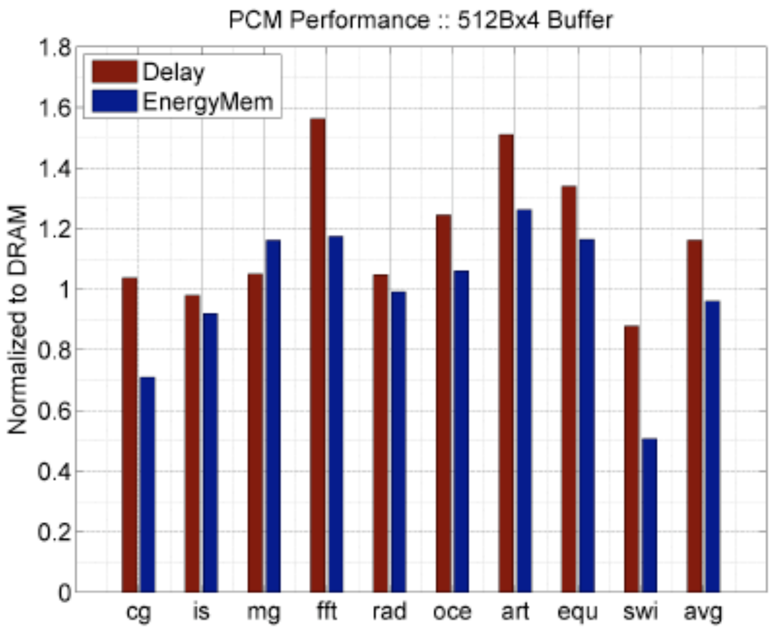
- Idea 2: Use multiple row buffers in each PCM chip
  → Reduces array reads/writes → better endurance, latency, energy

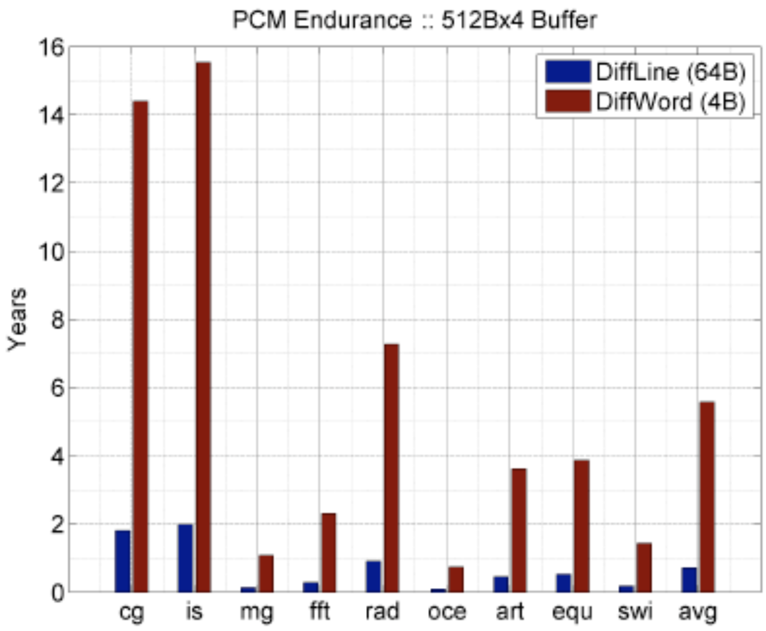
- Idea 3: Write into array at cache block or word granularity
  → Reduces unnecessary wear



Results: Architected PCM as Main Memory

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

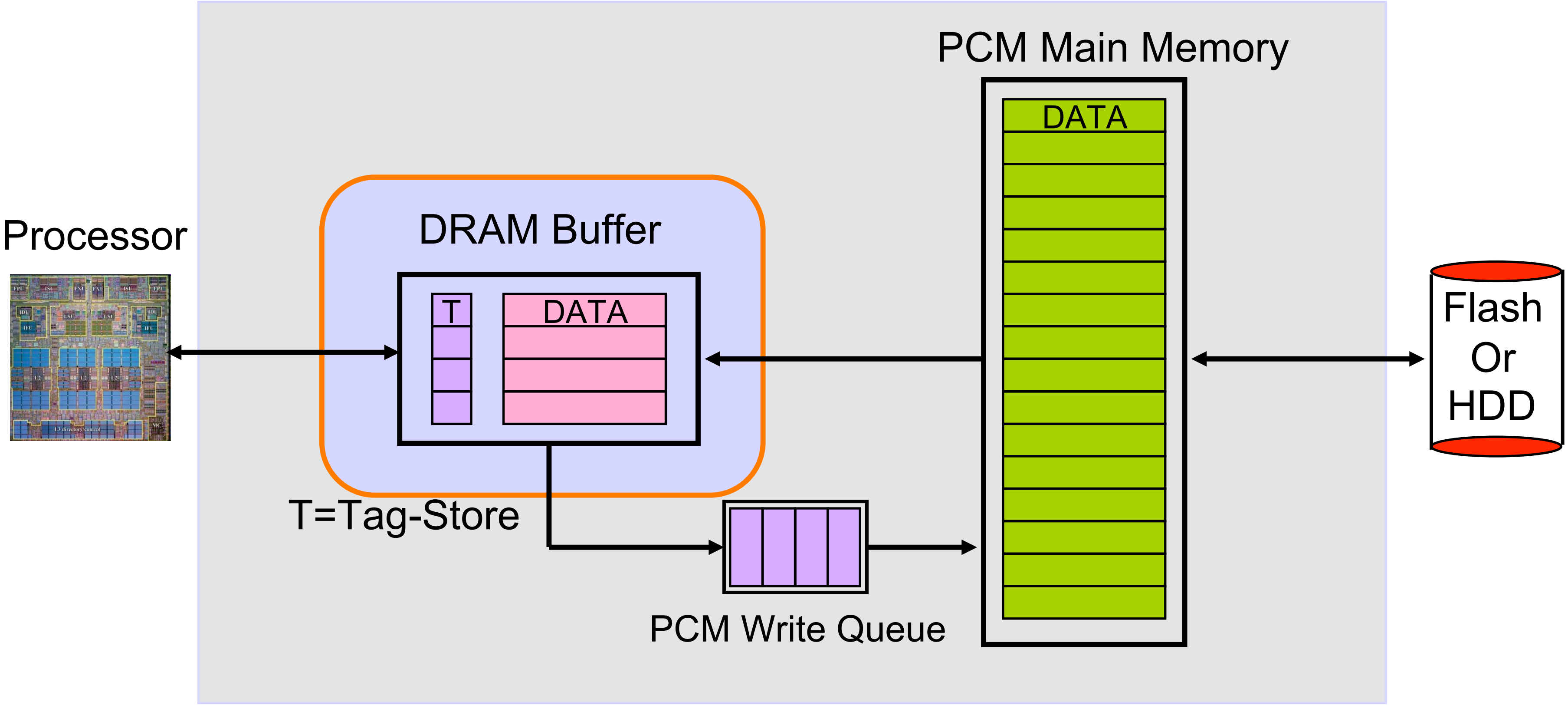




- Caveat 1: Worst-case lifetime is much shorter (no guarantees)
- Caveat 2: Intensive applications see large performance and energy hits

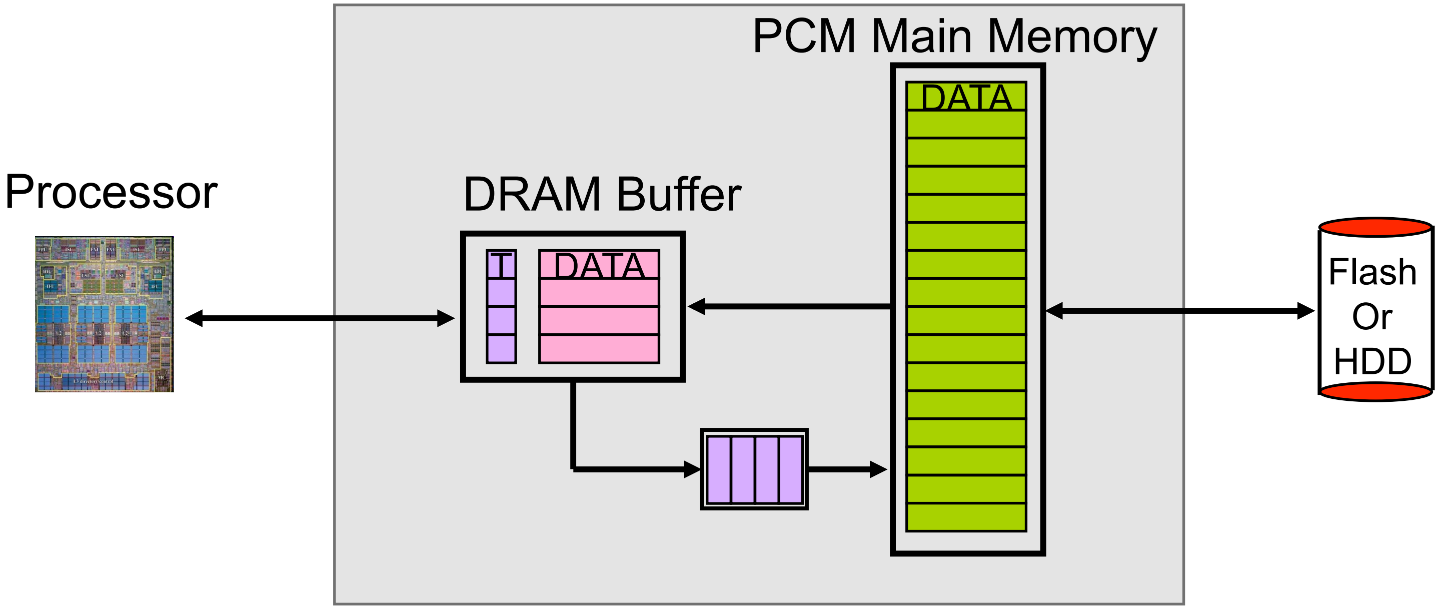
Another Alternative: DRAM as PCM Cache

- Goal: Achieve the best of both DRAM and PCM/NVM
  - Minimize amount of DRAM w/o sacrificing performance, endurance
  - DRAM as cache to tolerate PCM latency and write bandwidth
  - PCM as main memory to provide large capacity at good cost and power



Write Filtering Techniques

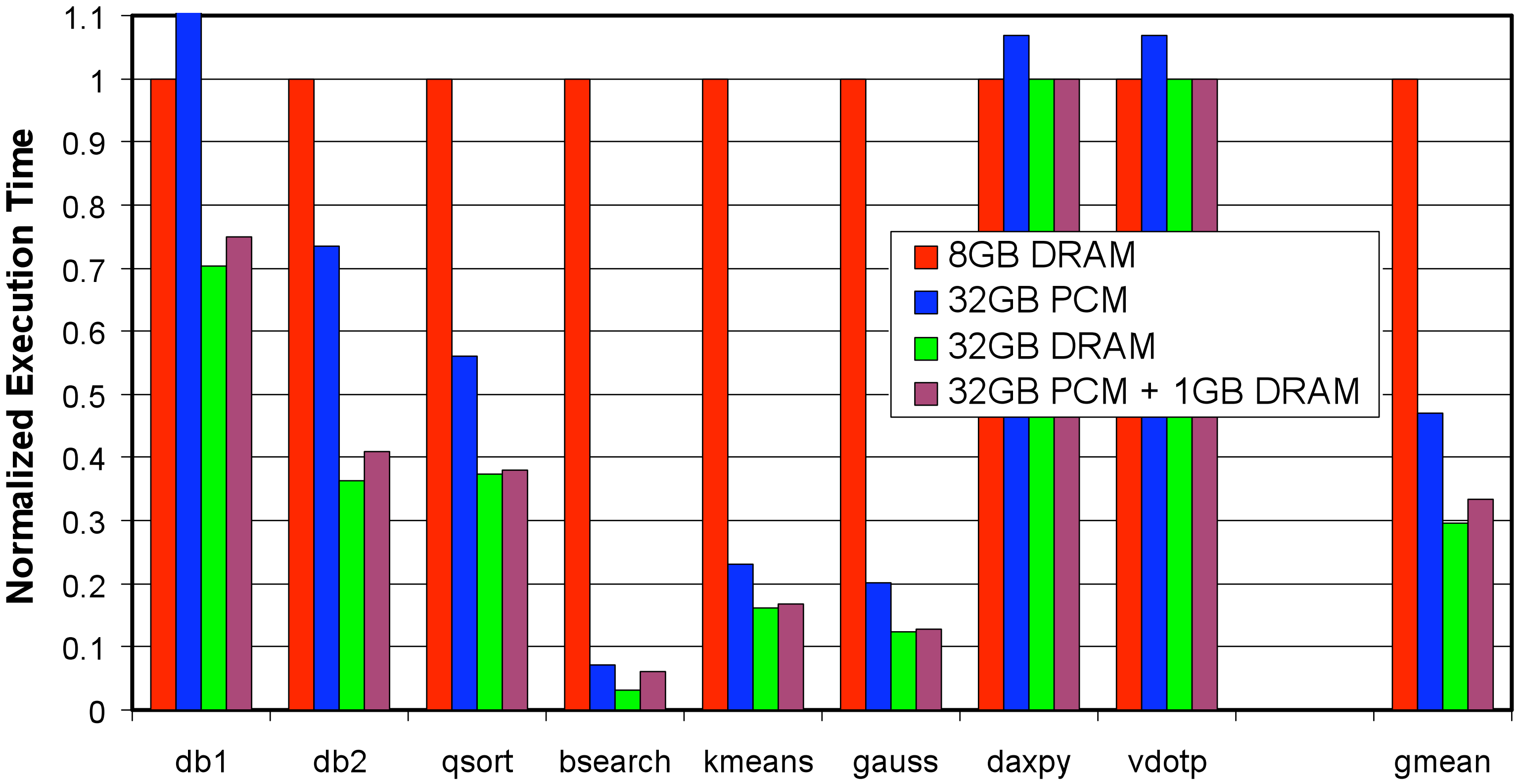
- Lazy Write: Pages from disk installed only in DRAM, not PCM
- Partial Writes: Only dirty lines from DRAM page written back
- Page Bypass: Discard pages with poor reuse on DRAM eviction



- Qureshi et al., "Scalable high performance main memory system using phase-change memory technology," ISCA 2009.

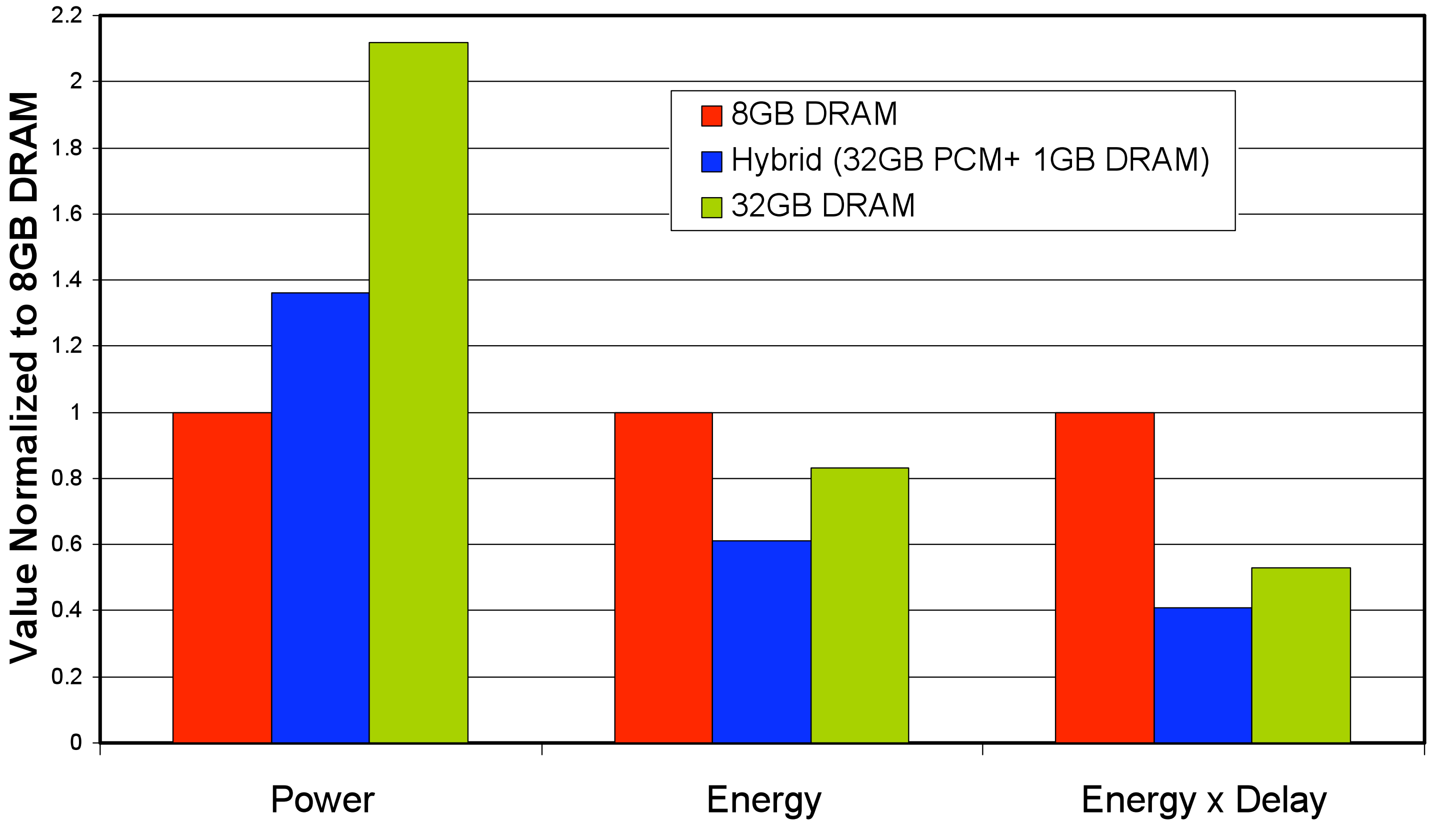
Results: DRAM as PCM Cache (I)

- Simulation of 16-core system, 8GB DRAM main-memory at 320 cycles, HDD (2 ms) with Flash (32 us) with Flash hit-rate of 99%
- Assumption: PCM 4x denser, 4x slower than DRAM
- DRAM block size = PCM page size (4kB)



Results: DRAM as PCM Cache (II)

- PCM-DRAM Hybrid performs similarly to similar-size DRAM
- Significant power and energy savings with PCM-DRAM Hybrid
- Average lifetime: 9.7 years (no guarantees)



Talk Agenda

- Major Trends Affecting DRAM-Based Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies (PCM)
- Research Challenges: PCM as Main Memory
- Preliminary Ideas and Results
- Open Questions
- Summary

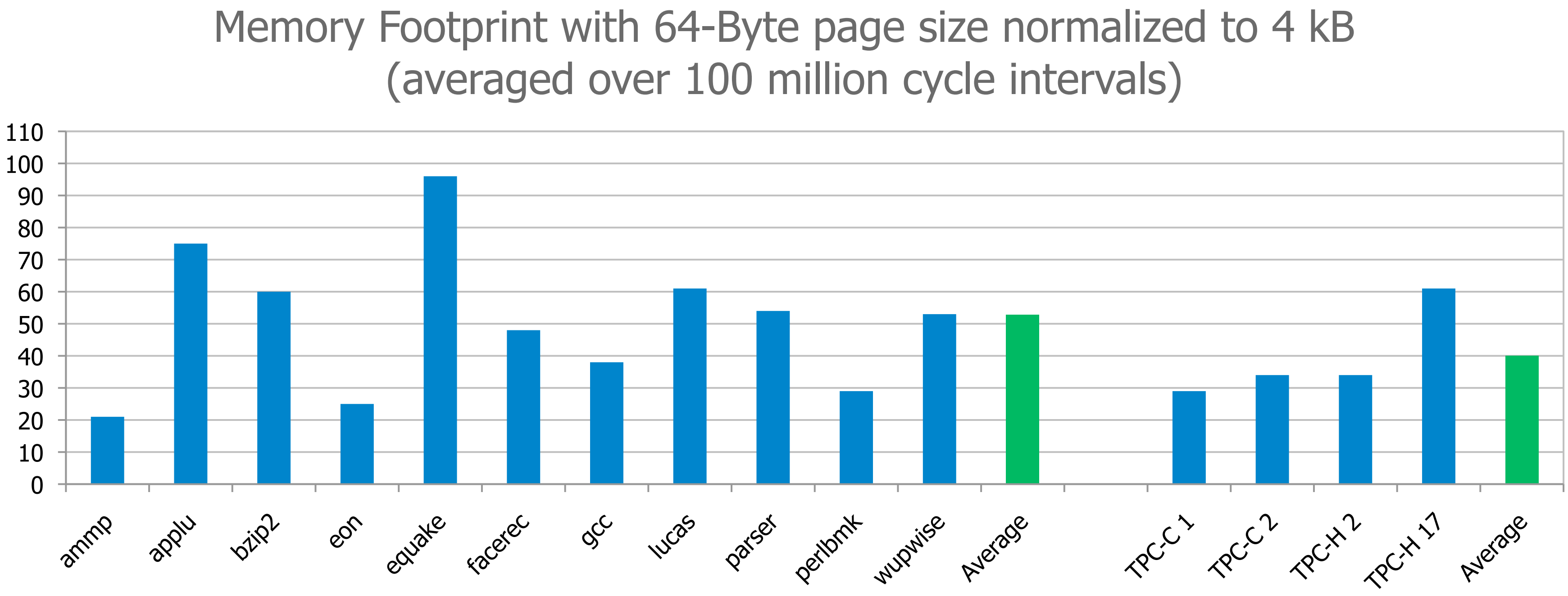
Can We Do Better?

- One idea: Hardware manages DRAM at cache-block (64-byte) granularity instead of page granularity

  + Smaller DRAM footprint → better utilization of DRAM space

  + Smaller read/write granularity into PCM → better latency, endurance

  -- Larger tag store in hardware
- Research challenges: DRAM management algorithms/policies



The Endurance Problem

- Problem: A process can intentionally or unintentionally degrade main memory size
  - Harder problem than in Flash since write bandwidth into main memory significantly higher, latency is lower

- Research Challenge: How to design write-filtering/wear-leveling/attack-detection mechanisms that maximize and guarantee lower bounds on memory lifetime

- Questions/Concerns:
  - Hardware or software?
  - Simplicity: cannot afford large tables
  - Latency: wear-leveling likely cannot have high latency

The Read/Write Latency and QoS Problems

- Write-intensive applications can deny/degrade service of read-intensive applications

- Research challenges:
  - How to tolerate read latency?
  - How to provide QoS in the presence of asymmetric latencies?
  - How to provide QoS in the presence of DRAM and PCM?

- Questions:
  - Can we take advantage of multi-level PCM cells and iterative writes?
  - Can we design intelligent prefetching mechanisms?
  - How do we partition DRAM/PCM capacity and bandwidth to satisfy SLAs?

Taking Advantage of Non-Volatility: Examples

- Application checkpointing [Dong+, SC'09]
  - HPC apps spent significant time (>25%) in checkpointing
  - Major reason of delay: low bandwidth of HDD
  - PCM provides high bandwidth, low latency, high-endurance option to do quick checkpointing (<4% overhead)

- Fast boot and application startup

- Deeply embedded systems
  - Enable continuous computation under unstable power sources

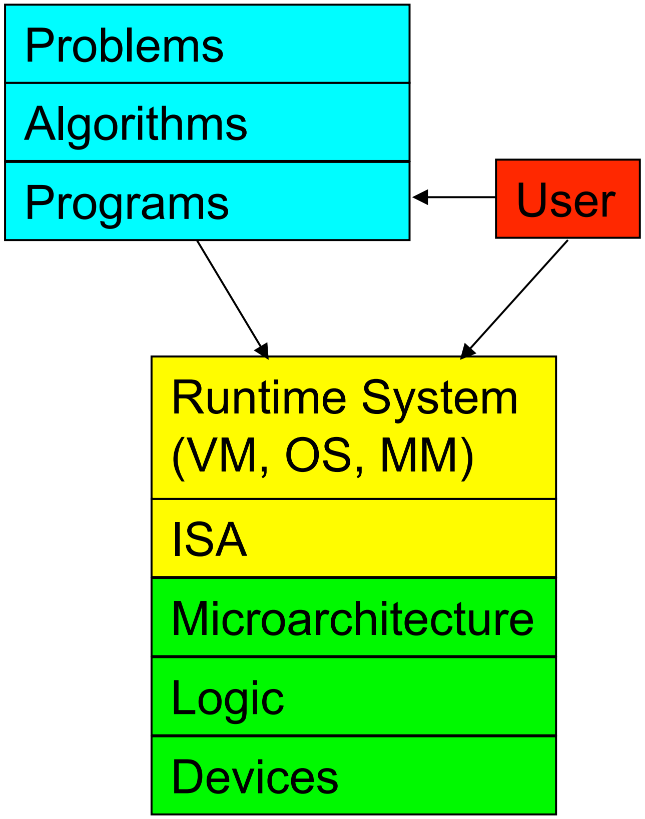
- "Correctness-critical" data storage for applications

NVM as Memory: Research Challenges

- Many research opportunities from technology layer to algorithms layer

- Enabling NVM
  - How to maximize performance?
  - How to maximize lifetime?
  - How to prevent denial of service?

- Exploiting NVM
  - How to exploit non-volatility?
  - How to minimize energy consumption?
  - How to minimize cost?
  - How to exploit NVM on chip?



Problems

Algorithms

Programs

User

Runtime System (VM, OS, MM)

ISA

Microarchitecture

Logic

Devices

Talk Agenda

- Major Trends Affecting DRAM-Based Main Memory
- Requirements from an Ideal Main Memory System
- Opportunity: Emerging Memory Technologies (PCM)
- Research Challenges: PCM as Main Memory
- Preliminary Ideas and Results
- Open Questions
- Summary

Summary

- Key trends affecting main memory
  - End of DRAM scaling (cost, capacity, efficiency)
  - Need for high capacity
  - Need for energy efficiency

- Emerging NVM technologies can help
  - PCM more scalable than DRAM and non-volatile
  - But, it has critical shortcomings: latency, active energy, endurance

- We need to enable promising NVM technologies by overcoming their shortcomings

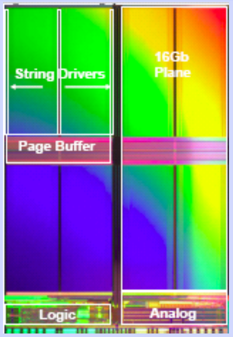
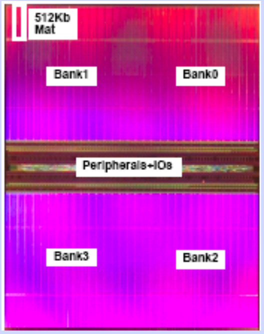
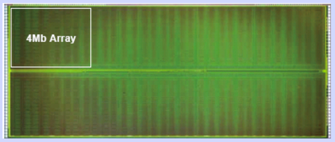
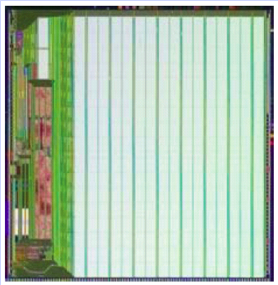
- Many exciting opportunities to reinvent main memory at all layers of computing stack

PCM (NVM) as Main Memory: Opportunities and Challenges

Onur Mutlu
Carnegie Mellon University

CMU PDL Retreat
October 25, 2010

PCM vs. Other NVM Technologies

Metric	Flash	FeRAM	MRAM	PCM
Technology	34 nm	130 nm	150 nm	45 nm
Cell Size	$3.9\ F^2$	$14.9\ F^2$	$44.4\ F^2$	$5.5\ F^2$
Array Size	32 Gb	128Mb	32 Mb	1 Gb
Write Speed	900 μs, 9 MBps	83 ns, 1.6GBps	40ns	10 MBps*
Read Speed	50 μs	43 ns	32ns	<100ns
Vcc	2.7-3.6 V	1.9 V	1.8 V	1.8 V
Company	Intel & Micron	Toshiba	Hitachi & Tohoku Univ	Numonyx
Micrograph (not to scale)	 String Drivers 16Gb Plane Page Buffer Logic Analog R Zeng ISSCC 2009	 512Kb Mat Bank1 Bank0 Peripherals+IOs Bank3 Bank2 H Shiga ISSCC 2009	 R Takemura VLSI 2009	 G. Servalli, IEDM 2009

Slide Courtesy: Moinuddin Qureshi, IBM