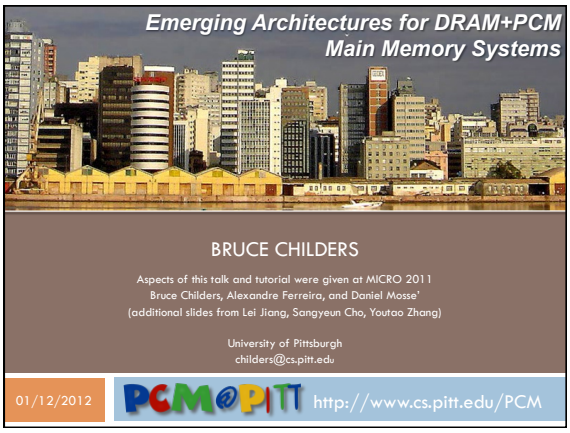


BRUCE CHILDERS

Aspects of this talk and tutorial were given at MICRO 2011
Bruce Childers, Alexandre Ferreira, and Daniel Mosse'
(additional slides from Lei Jiang, Sangyeun Cho, Youtao Zhang)

University of Pittsburgh
childers@cs.pitt.edu

01/12/2012 PCM@PITT http://www.cs.pitt.edu/PCM

Data Center Energy Trends

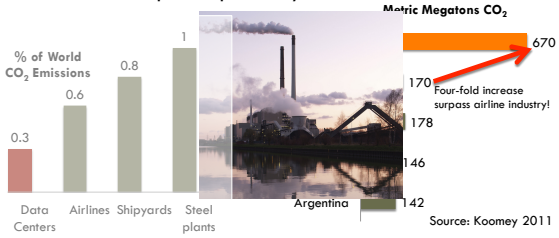
- Data center electricity usage
  - Increased by 56% from 2005 to 2010
  - 1.1% to 1.5% total world electricity usage
  - 1.7% to 2.2% total US electricity

(Note: Includes impact of 2008 recession.)
(Note: 2x increase 2000 to 2005, below prediction.)

Source: Koomey 2011

The Consequence

- At current growth rate (2000-2005) in energy usage for data centers, will need 30 new coal-fired or nuclear power plants by 2015

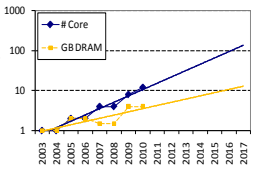


Source: Koomey 2011

Increasing Memory Demand

- Parallelism (core count)
- Larger & complex data sets
- More sophisticated applications
- Virtualization & consolidation

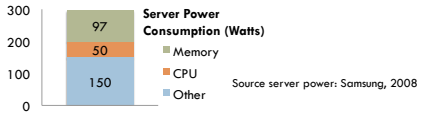
- Today: 10's (to 100's) GB
- Tomorrow: Terabyte and beyond???



Source: Kevin Te-Ming Lim, *Disaggregated Memory Architectures for Blade Servers*, Ph.D. Thesis, University of Michigan, 2010

More Memory

- Energy/power consumption shift



Source server power: Samsung, 2008

- Terabyte in Buffered DRAM or DDR3 SDRAM
  - 8GB: 125 DIMMs, 400W@DDR3, 1.25KW@FBDRAM
- **Up to 4-10x more than already power hungry machines!**

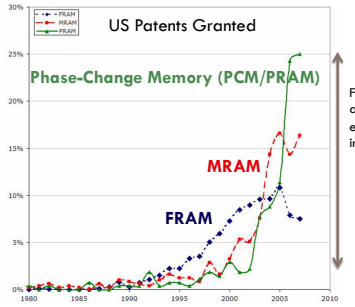
DRAM

- A long-time winner: Decades old!
  - Cost, power, performance trade-offs have favored it
  - ***Massive future capacity leads to a different outcome!***
- Limitations to DRAM
  - Destructive reads: Must replace data after a read
  - Limited data retention: Periodic refresh
  - Susceptibility to errors: Charge can be disturbed
  - Scalability: Projections (ITRS) question below 22*nm*

The Wave Rolling In

- DRAM has long been the best choice until now…
- DRAM ***does*** offer advantages
  - Effectively unlimited write endurance (doesn't wear out)
  - Fast read/write (symmetric) latency
  - (And, of course, it's a commodity, here today, etc.)
- Can we use it judiciously? Just a little bit, please?
  - Combine with ***alternative technology***
  - Small DRAM has reasonable energy, capacity
  - We've seen this before… SRAM cache vs DRAM?

The Wave Rolling In



For an "old technology", a dramatic change of events with tremendous interest!

Source: Lam, VLSI-TSA 2008

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Fast, non-destructive reads: Nearing parity w/DRAM
Non-volatile, non-destructive, no refresh → low energy

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Density on par with DRAM, 2.5nm prototype
Liang et al, A 1.4uA Reset Current Phase Change Memory Cell with Integrated Carbon Nanotube Electrodes for Cross-Point Memory Applications, *IEEE Symp. on VLSI (VLSIT)*, 2011

Fast, non-destructive reads: Nearing parity w/DRAM

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Write performance limited
Relatively slow bit cell writes but no block erasure required like Flash
Multiple write rounds of bit groups, leading to 1us (Numonyx prototype)

Density on par with DRAM, 2.5nm prototype

Fast, non-destructive reads: Nearing parity w/DRAM

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Repeated writes lead to wear on bit cell
Writes cause stress to bit cells, leading to failure
Limited write cycles but better than Flash

Write performance limited by individual bit and group of bits

Density on par with DRAM, 2.5nm prototype

Fast, non-destructive reads: Nearing parity w/DRAM

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Similar array structure/operation as DRAM: bit (byte) addressability

Repeated writes lead to wear on bit cell

Write performance limited by individual bit and group of bits

Density on par with DRAM, 2.5nm prototype

Fast, non-destructive reads: Nearing parity w/DRAM

Alternative Memory Technology

	Read Speed	Write Speed	Cell Area	Endurance	Addressability
DRAM	20~50ns	20~50ns	$6F^2$	10^{15}	Yes
SRAM	~2ns	~2ns	$146F^2$	10^{15}~10^{16}	Yes
NAND Flash	25us	500us	$5F^2$	10^4~10^5	No
STT-RAM	2ns	10ns	37~$40F^2$	10^{12}	Yes
PCM	30~50ns	~1us	5~$8F^2$	10^7~10^8	Yes

Similar

Write

Nearly ideal complement (*maybe* replacement?) for DRAM (scales, low standby power, bit addressable, fast reads)

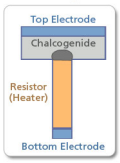
BUT…. must find techniques to overcome limitations

Fast, non-destructive reads: Nearing parity w/DRAM

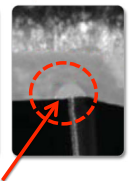
PCM: The Fundamental Idea

- *Similar* process as CD-R
- Chalcogenide (GST)
- Application of ***heat*** changes state of material
- ***Resistance*** associated with each state stores a bit
  - Crystalline (low, SET, 1)
  - Amorphous (high, RESET, 0)
- Operation
  - Write: Heat/cool
  - Read: Measure resistance

Theory



Implementation



Programmed volume of GST (heated and then cooled to change phase)

Diagram/photo: Micron Technology
http://www.micron.com/innovations/pcm.html

PCM Read/Write Operations

- Read
  - Measure resistance
    - Low: logic 1 (SET)
    - High: logic 0 (RESET)
  - Relatively fast
  - Power efficient
  - Non-destructive
- Writes
  - Slow bit writes: heating/ cooling: 50ns ~ 150ns
  - Limited parallel bit writes: large programming current
  - Long latency: 1000ns
  - High write energy
  - Heat stress leads to failure, with limited endurance (10^7)

Consequences of PCM

Asymmetric read/write latency and bandwidth

- Reads projected to reach parity with DRAM
- Writes will remain slow due to heating/cooling

Wear-out and endurance management

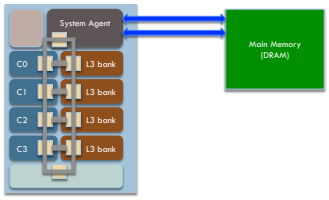
- Integrated relatively near CPU leads to heavy usage
- E.g., one write/second: PCM fails in 110 days
- Memory will quickly fail without precautions

Nonvolatility
Reliability

Important, desirable properties. Most focus has been on making it work first, then find ways to exploit these properties

Rethinking Main Memory for PCM

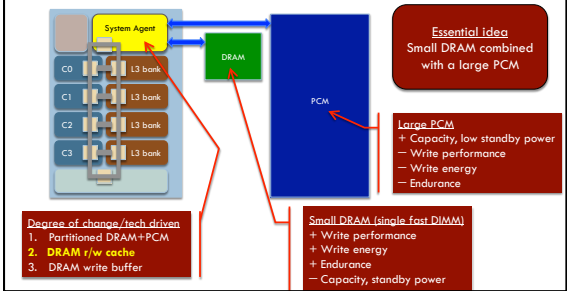
- Starting Point: DRAM Main Memory



Sandy Bridge

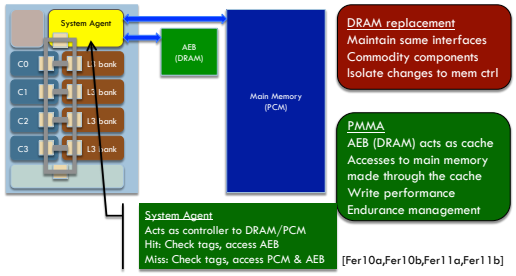
Hybrid Memory Archetype

- Conventional memory adapted to PCM



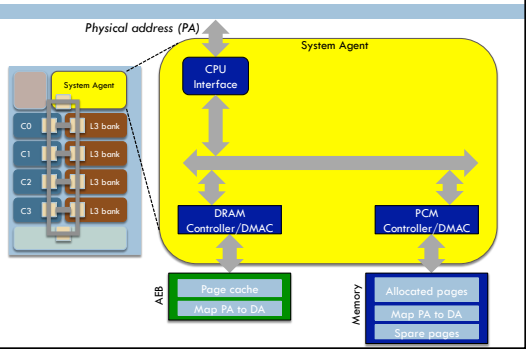
DRAM read/write cache

- Phase-change Main Memory Architecture (PMMA)

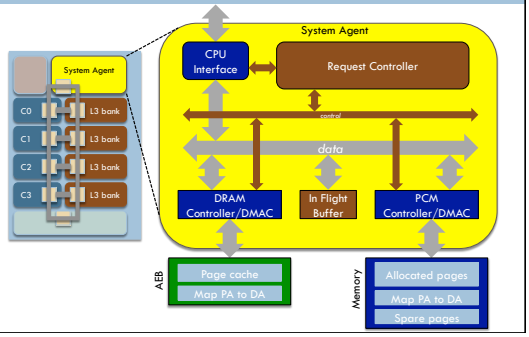


[Fer10a,Fer10b,Fer11a,Fer11b]

PMMA



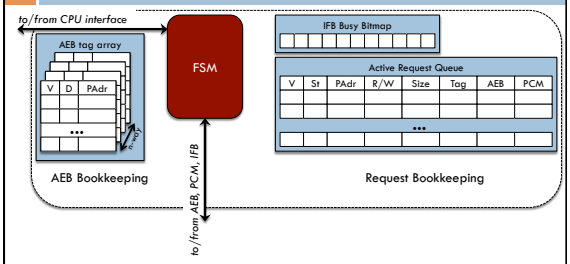
PMMA



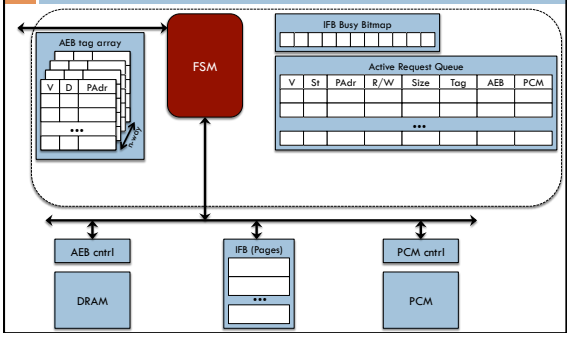
Request Controller

- Operates on ***pages (larger than cache block from CPU)***
- Processes requests & allocates resources
  - Multiple outstanding requests
  - Page allocation & eviction (AEB)
  - Map physical to device address
- Book keeping
  - Track resources used, including what is cached & where
  - Map physical address (PA) to PCM device address (DA)
  - IFB: High speed memory buffers inflight pages (AEB/PCM)

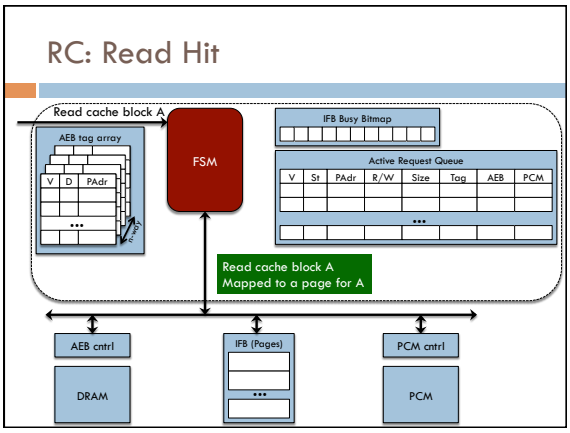
Request Controller



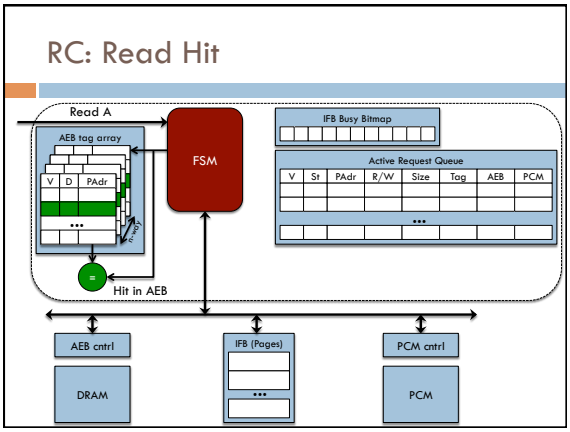
Request Controller



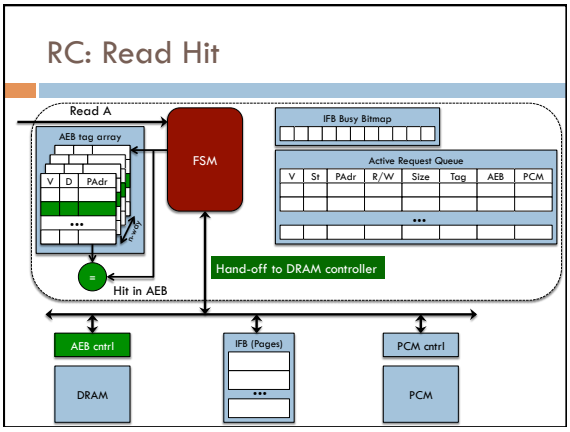
RC: Read Hit



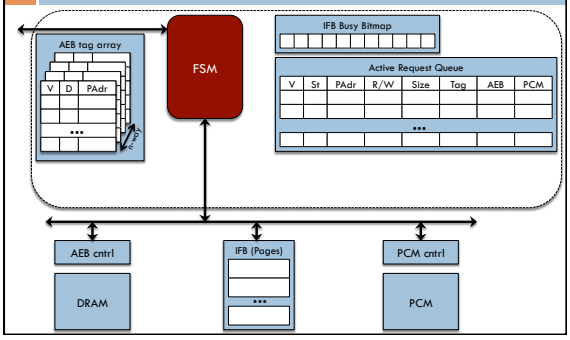
RC: Read Hit



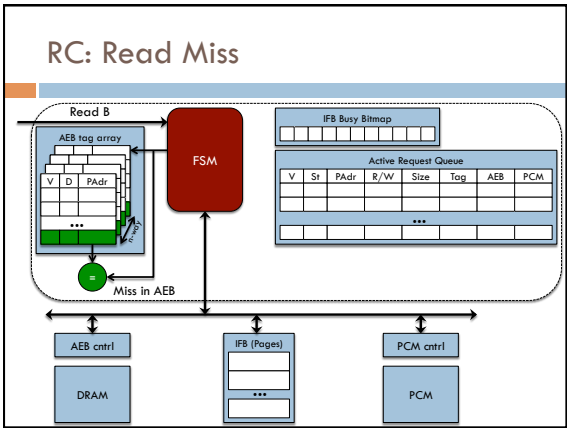
RC: Read Hit



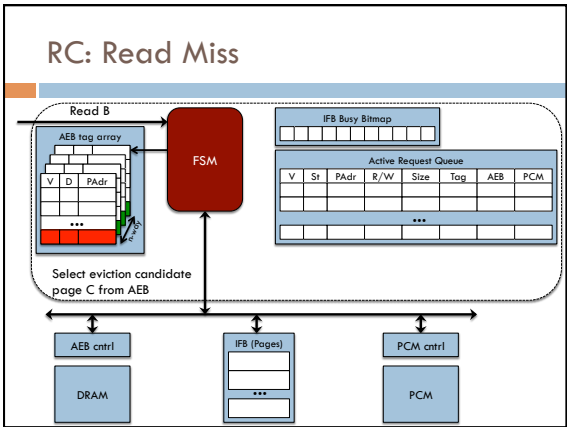
Request Controller



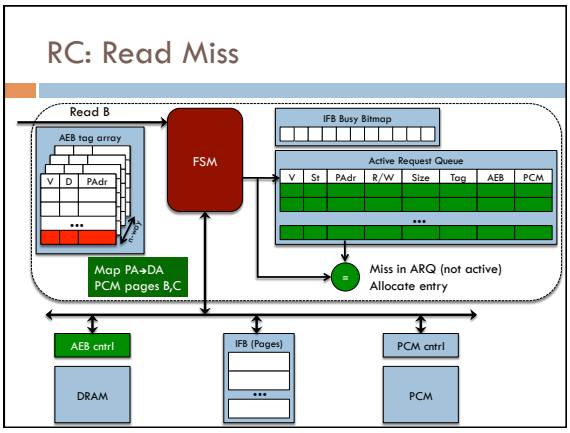
RC: Read Miss



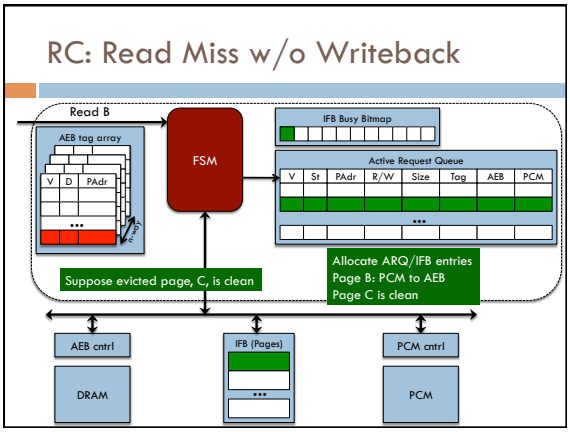
RC: Read Miss



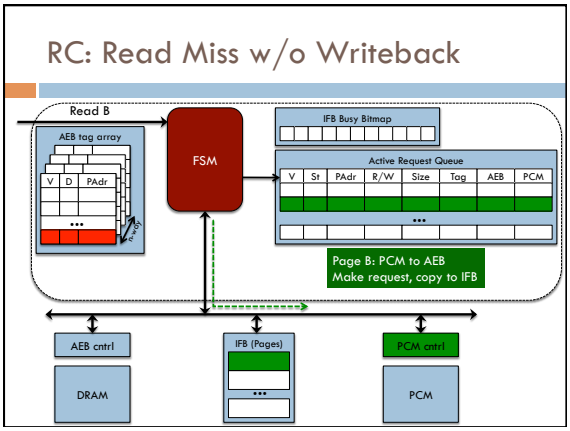
RC: Read Miss



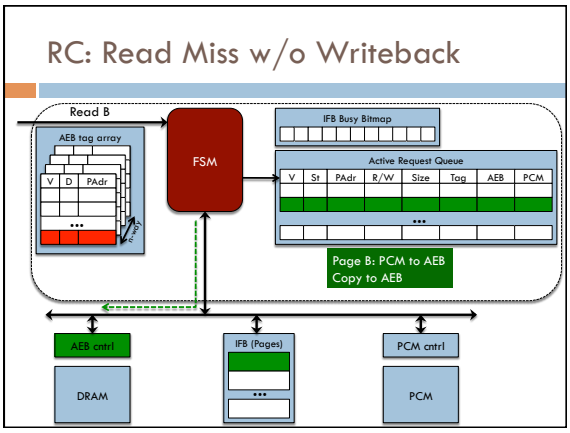
RC: Read Miss w/o Writeback



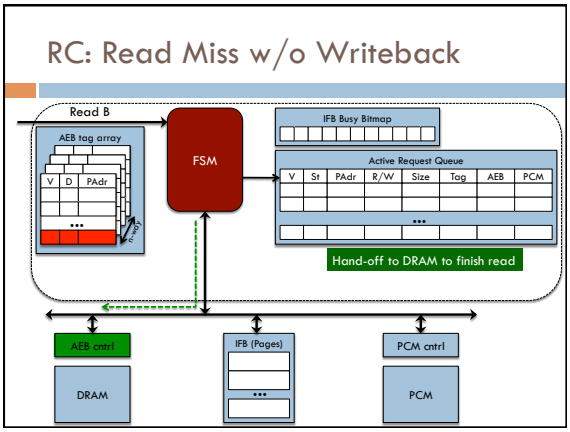
RC: Read Miss w/o Writeback



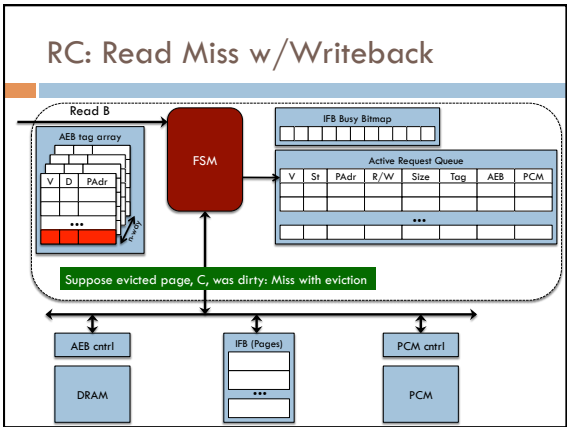
RC: Read Miss w/o Writeback



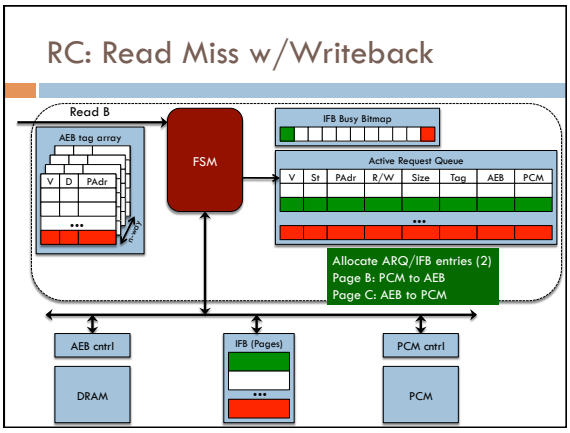
RC: Read Miss w/o Writeback



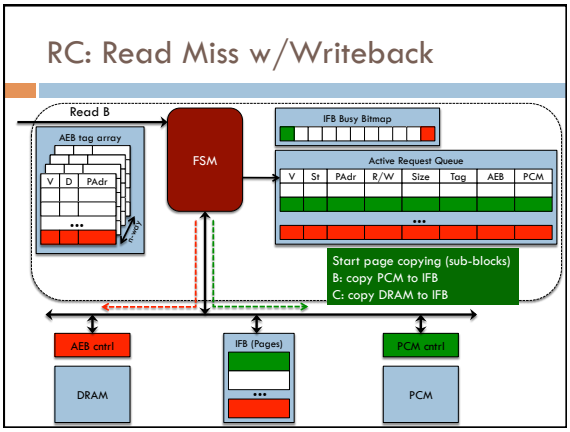
RC: Read Miss w/Writeback



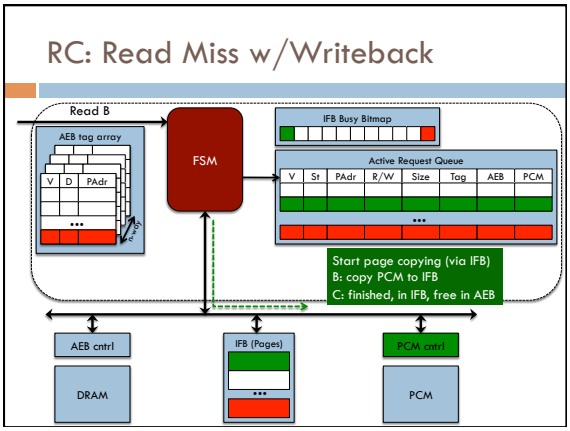
RC: Read Miss w/Writeback



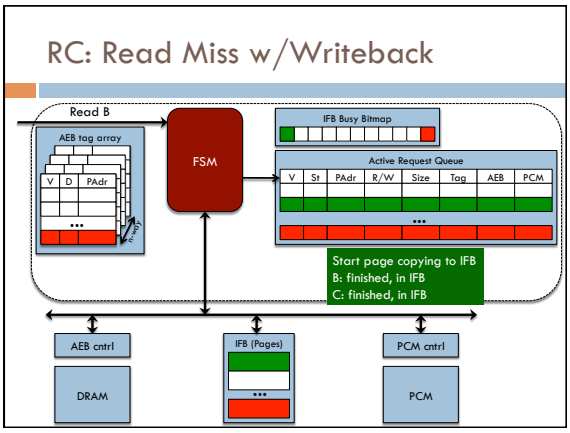
RC: Read Miss w/Writeback



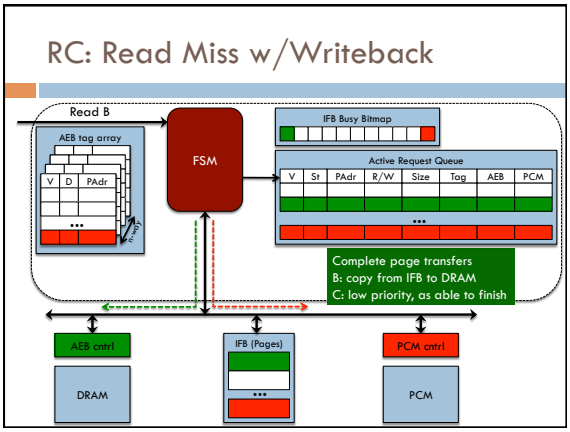
RC: Read Miss w/Writeback



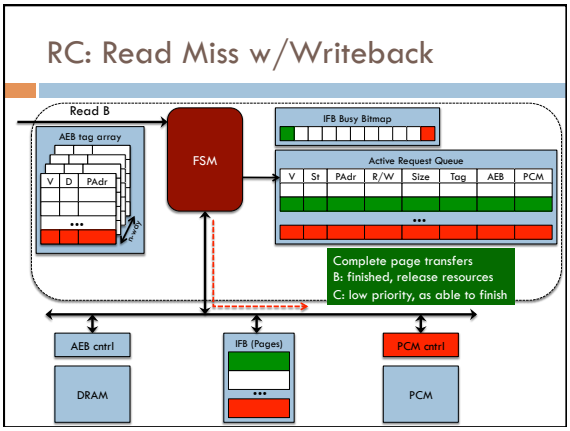
RC: Read Miss w/Writeback



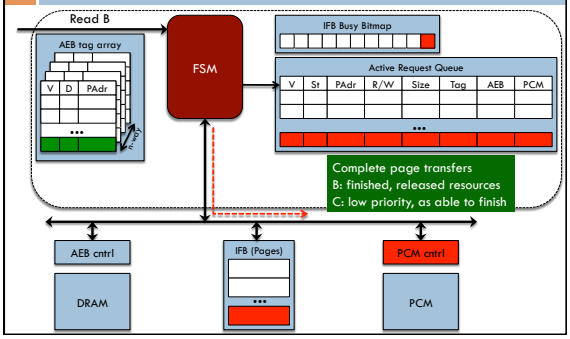
RC: Read Miss w/Writeback



RC: Read Miss w/Writeback



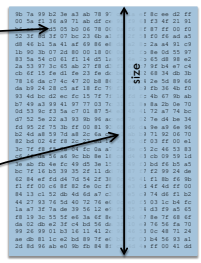
RC: Read Miss w/Writeback



① Optimization: Page Partitioning

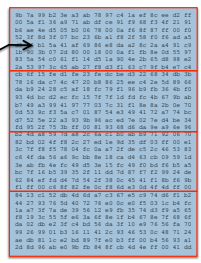
"Page" is data unit
AEB & PCM logical unit
PA-DA map at page level
E.g., 2KB, 1KB, 512B, …

Larger page size
+ smaller tag store
+ smaller mapping table
− unnecessary movement
− writes of clean data

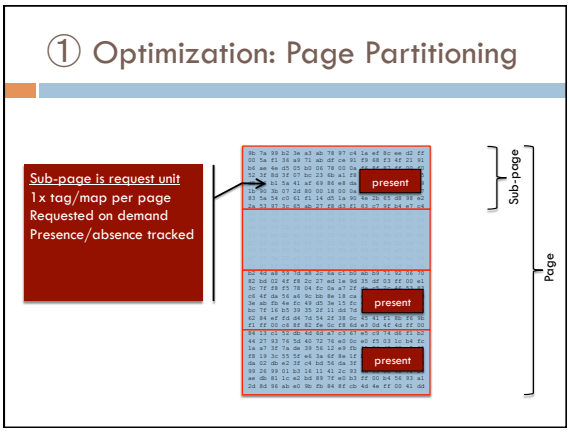


① Optimization: Page Partitioning

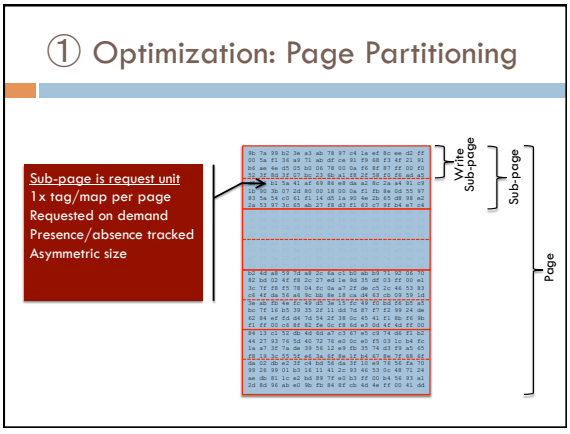
Sub-page is request unit
1x tag/map per page
Requested on demand



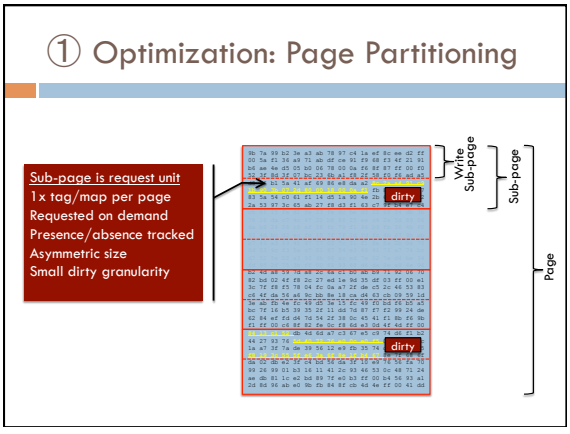
① Optimization: Page Partitioning



① Optimization: Page Partitioning

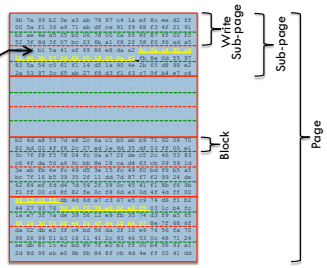


① Optimization: Page Partitioning



① Optimization: Page Partitioning

Block transfer unit
Smallest data transfer
Sized to PCM banks
Higher priority requests pre-empt betw. blocks

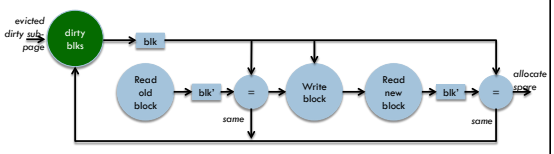


② Optimization: CW + AEB bypass

- Critical block (word) first
  - Deliver block generating miss to CPU
  - Transfer remaining blocks on page
- AEB bypass
  - Inflight pages can service requests, if data available
  - Data delivered directly from AEB

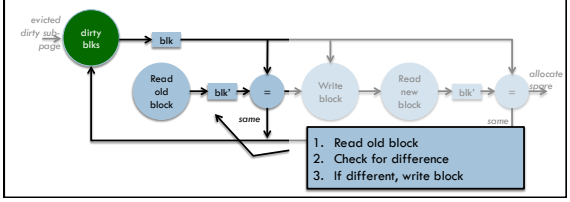
③ Optimization: RWR

- PCM read-write-read (RWR)
  - RWR avoids writing unchanged blocks in sub-page
  - Read verify detects failed page
  - Failed write leads to spare allocation



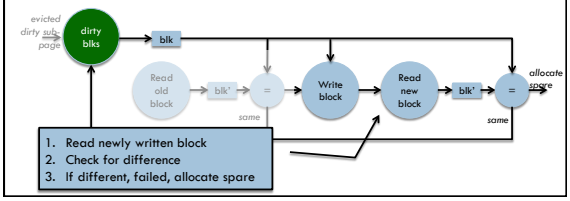
③ Optimization: RWR

- PCM read-write-read (RWR)
  - RWR avoids writing unchanged blocks in sub-page
  - Read verify detects failed page
  - Failed write leads to spare allocation



③ Optimization: RWR

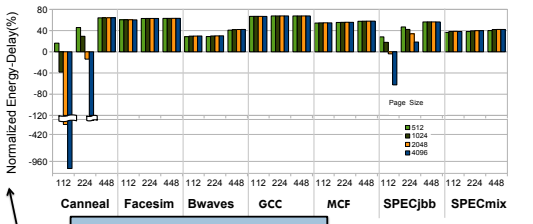
- PCM read-write-read (RWR)
  - RWR avoids writing unchanged blocks in sub-page
  - Read verify detects failed page
  - Failed write leads to spare allocation



④ Optimization: Endurance

- AEB eviction policy (N-chance) to *minimize* writes
- Non-uniform writes to memory
  - Uneven writes cause pages to fail before others
  - Failed page(s): memory is now broken
- Wear-leveling to uniformly distribute writes
  - Wear pages at same level
  - Pages will fail at approximately same time
- Spare capacity
  - Replace failed pages on-demand

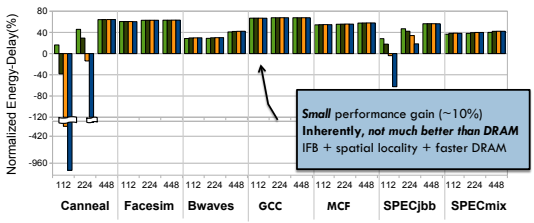
PMMA Energy-Delay



- Compared to equivalent capacity in DRAM-only system (16GB, 4 core)
- PMMA: small DRAM (speed optimized) with large PCM

?

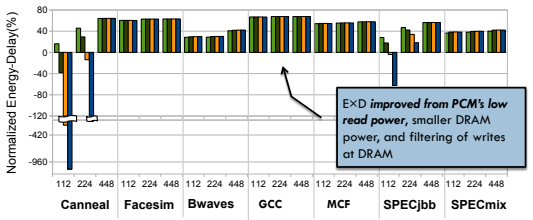
PMMA Energy-Delay



- E×D improved (small losses/gains are wins, e.g., bwaves)

?

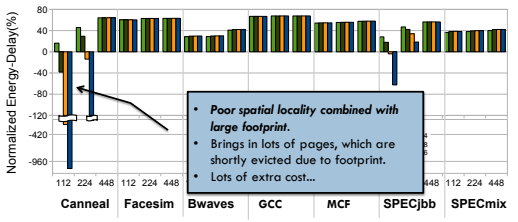
PMMA Energy-Delay



- E×D improved (small losses/gains are wins, e.g., bwaves)

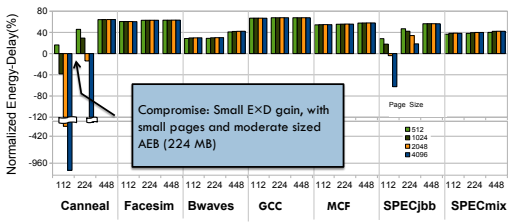
?

PMMA Energy-Delay



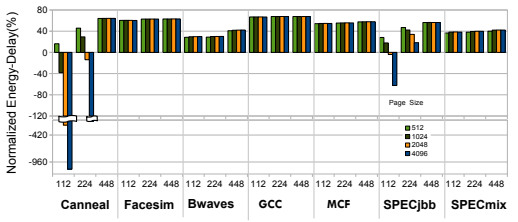
- E×D improved (small losses/gains are wins, e.g., bwaves)

PMMA Energy-Delay



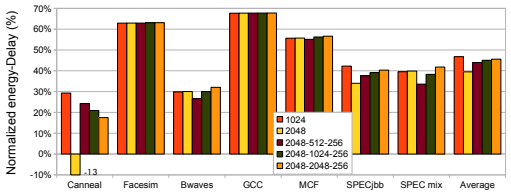
- E×D improved (small losses/gains are wins, e.g., bwaves)

PMMA Energy-Delay



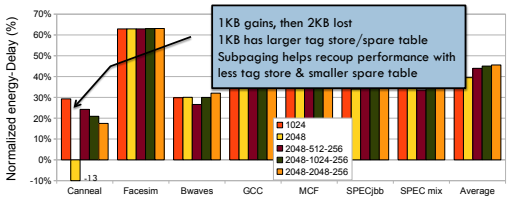
- E×D improved (small losses/gains are wins, e.g., bwaves)
- 256MB DRAM (224MB AEB+32MB meta) is good compromise
- 1024B vs 2048B page trades tag/spare table vs. locality

Read-Write Page Partitioning



- Results for AEB size 224 MB (+32MB meta data)

Read-Write Page Partitioning



- Results for AEB size 224 MB
- 1024B best overall result but larger metadata storage
- R/W page partitioning recoups losses from 2048B

Lifetime: Cumulative Impact

Technique	Lifetime	Cumulative Gain
Baseline (LRU)	0.47 month	
7-Chance	0.86	1.83X
+RWR	3.36 months	3.91X
+GC512-Random	97.29 months	28.91X

- Wear-leveling is essential to achieve 8 years
- 7-chance and RWR also have a large impact

Summary

- PCM architectures
  - ***DRAM complement for main memory?***
  - Flash replacement
  - Memory + storage combination
- Current front-runners share essential idea
  - Small DRAM + Large PCM
- Endurance on the way to being solved?
- Write bandwidth and energy likely to persist