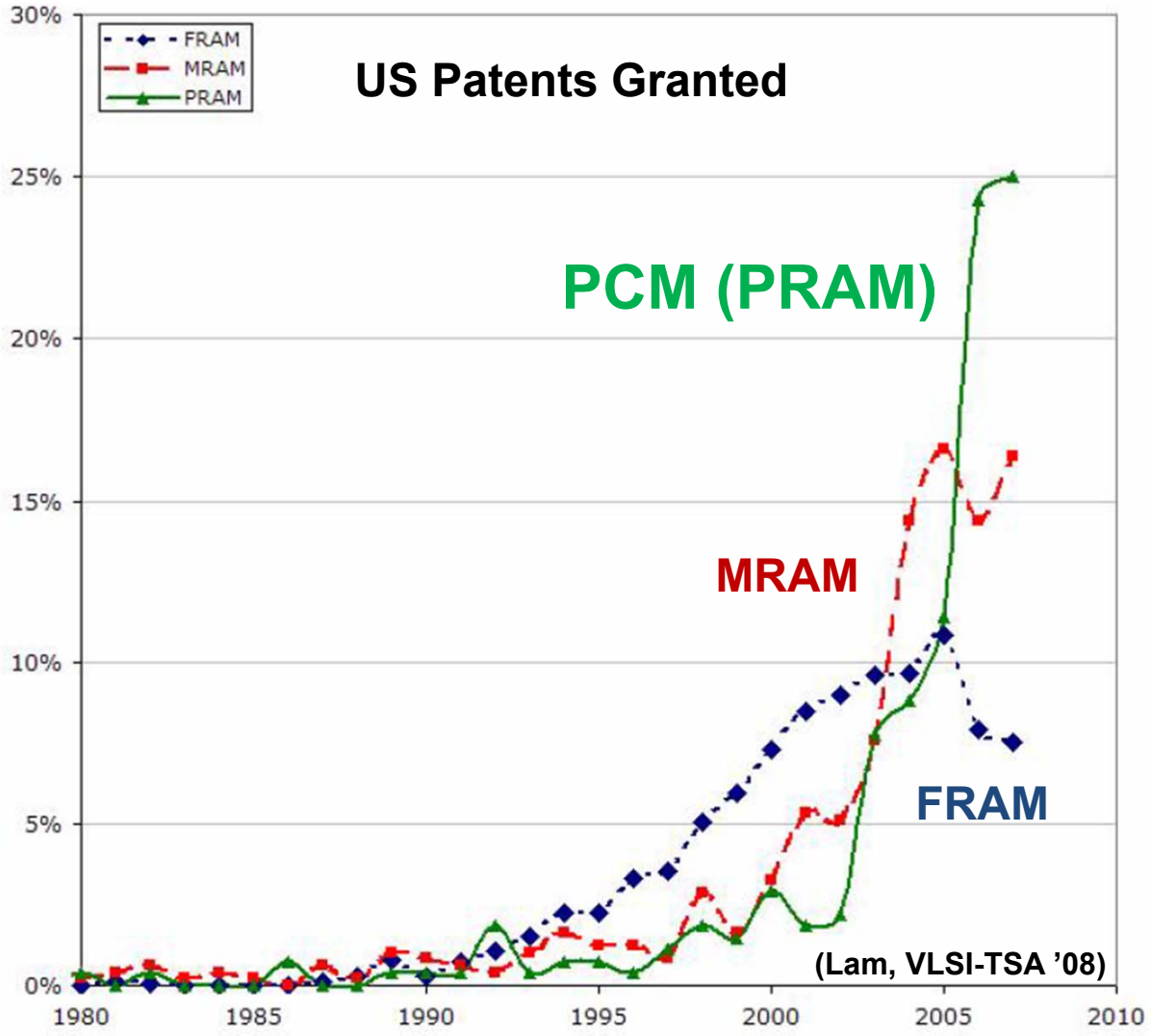
Use of PCM in Computer Systems: We need an End-to-End Exploration

Sangyeun Cho

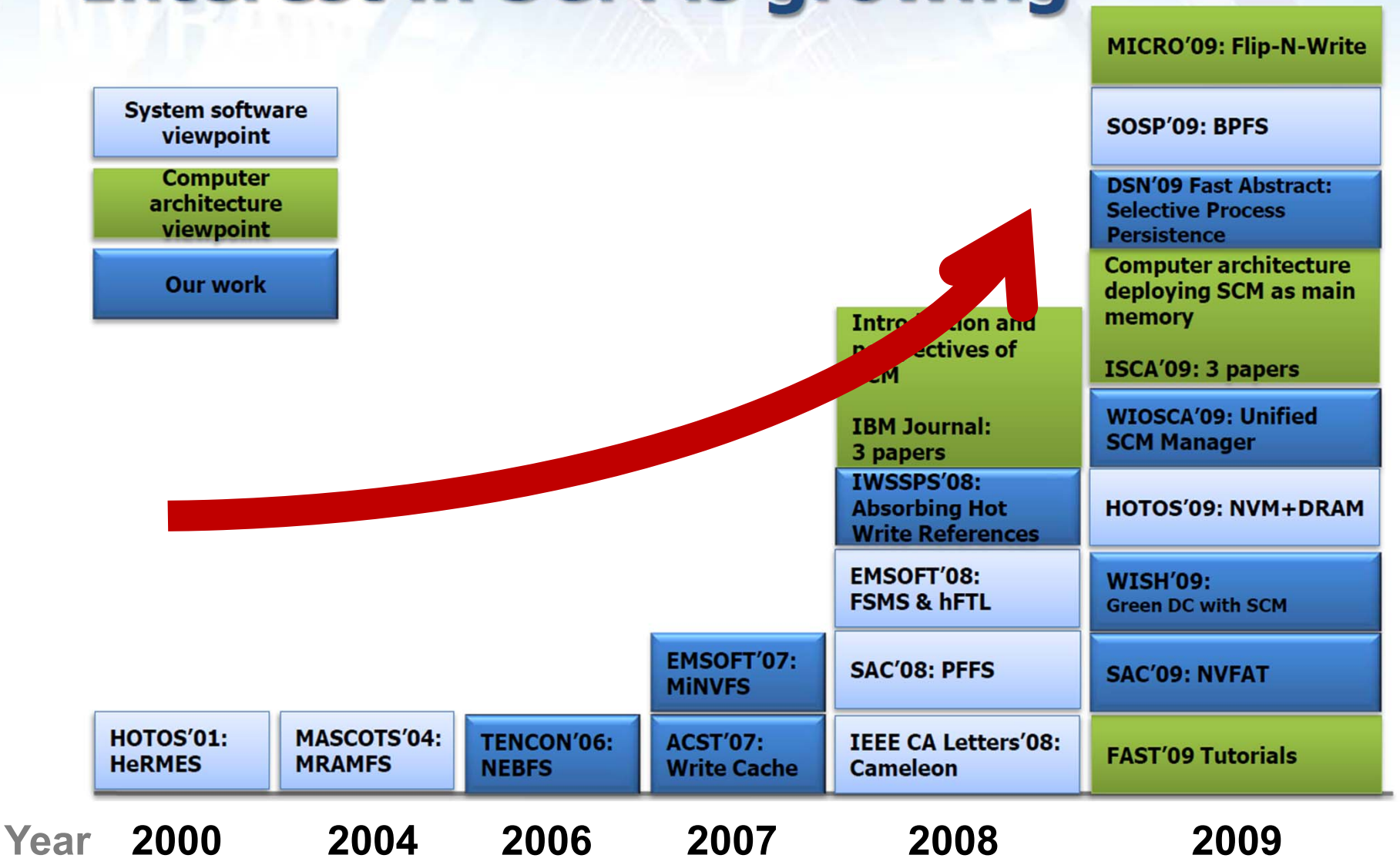
Computer Science Department

University of Pittsburgh

Era of new memories near



Interest in SCM is growing



(S. H. Noh, NVRAMOS '10)

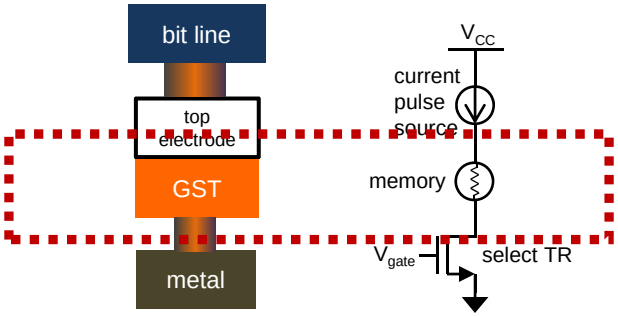
Since then (arch. & design community)

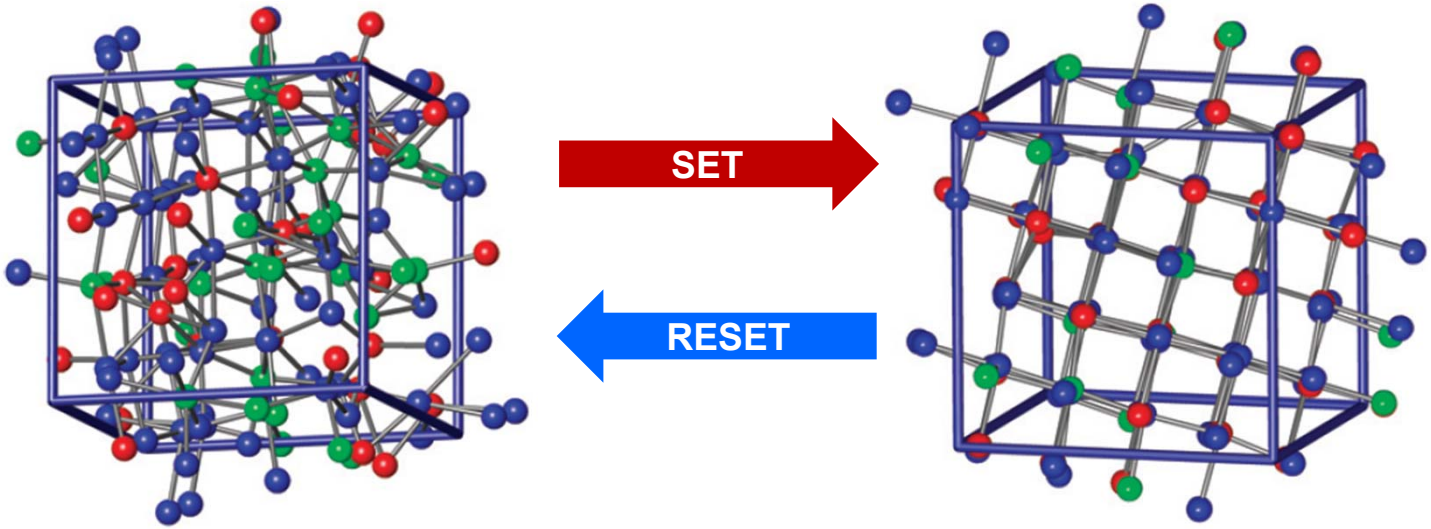
- Wear leveling & memory attack handling
  - Start-Gap [MICRO ‘09]
  - Security Refresh [ISCA ’10]
  - On-line attack detection [HPCA ‘11]
- Fault masking
  - ECP [ISCA ‘10]
  - SAFER [MICRO ‘10]
  - FREE-p [HPCA ‘11]
- Process variation awareness
  - Characterization & mitigation [MICRO ‘09]
  - Mercury [HPCA ‘11]
  - Variation vs. endurance [DATE ‘11]
- DAC-2011 has three papers
  - “Power Management” (Prof. Yoo), “Wear Rate Leveling” (ICT, China), “Variable Partitioning” (Hong Kong City Univ.)

Agenda

- **PCM 101**
- Industry trends
- PCM usage models
- Summary

Phase-change memory (PCM)



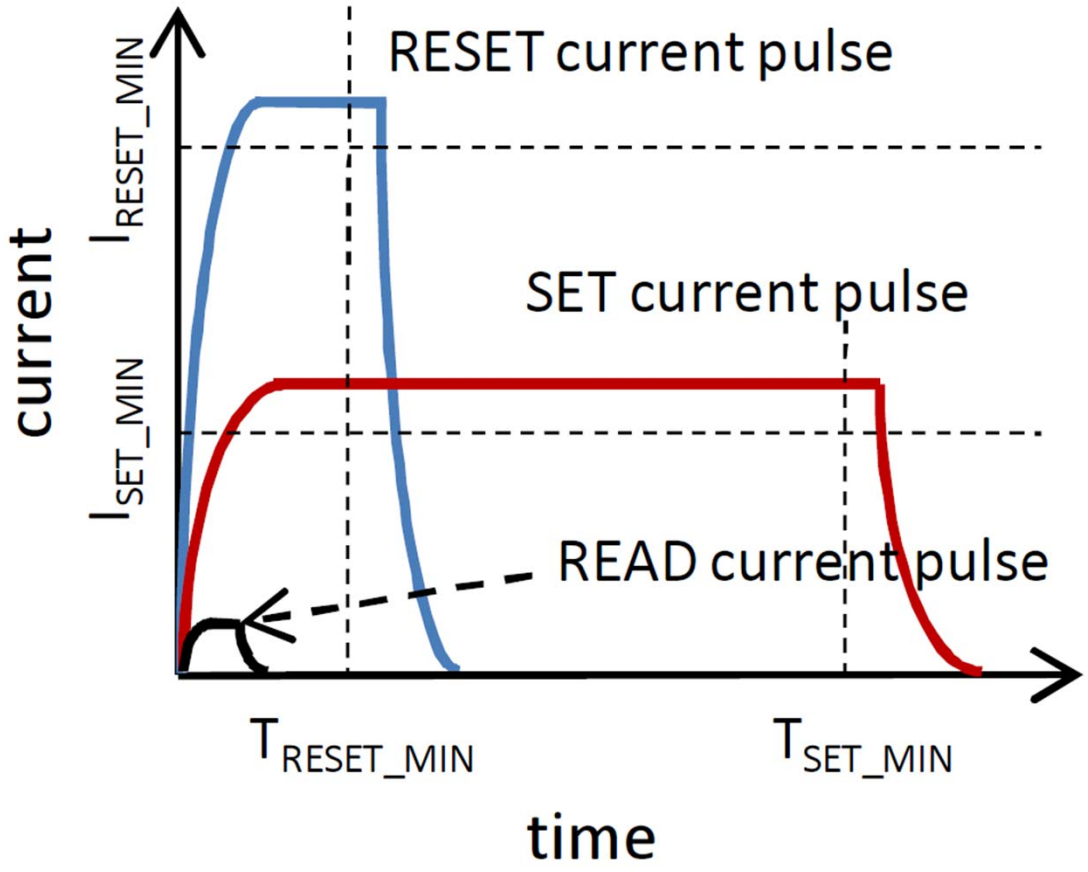


Amorphous = high resistivity

Crystalline = low resistivity

(Pictures from Hegedüs and Elliott, *Nature Materials*, March 2008)

PCM asymmetries

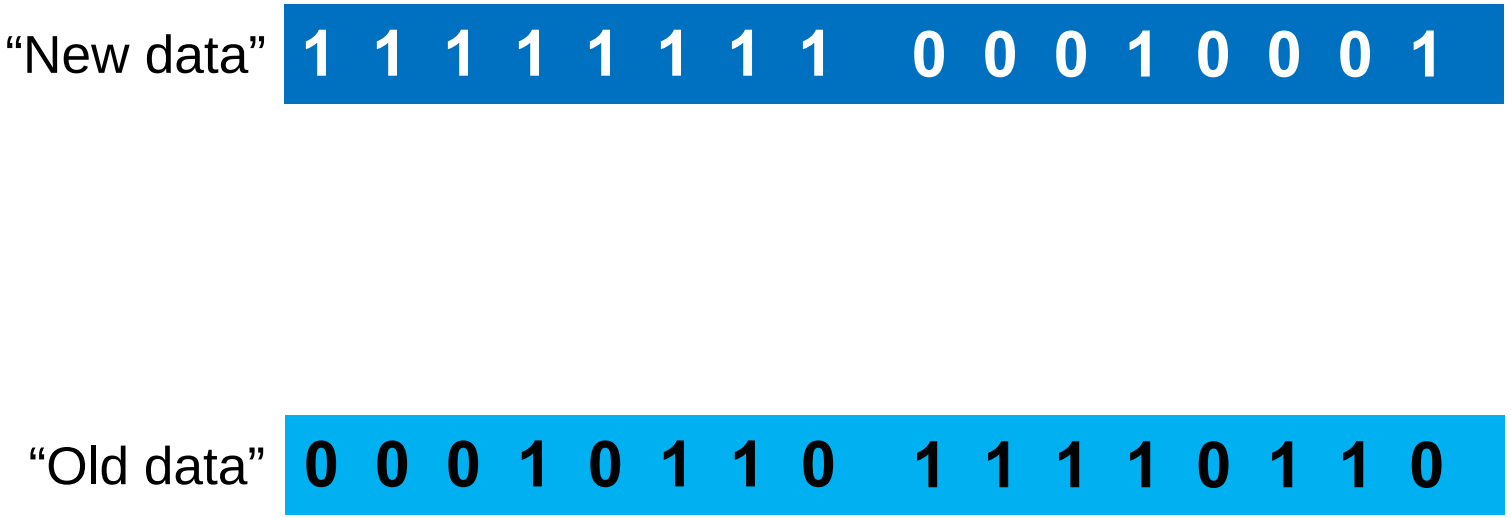


Note: Write is the thing

- Cycling of cell states leads to cell aging
  - Reported write endurance 10^5 to 10^6 (who said 10^{12}?)
- Burdensome to scale write bandwidth
  - High write currents (more bits means higher currents)
  - Reliability problem, added system design costs, …
- ***Theoretically*, scaling helps with both problems**
- ***Architectural techniques* to reduce bit updates**

E.g.: Flip-N-Write [MICRO '09]

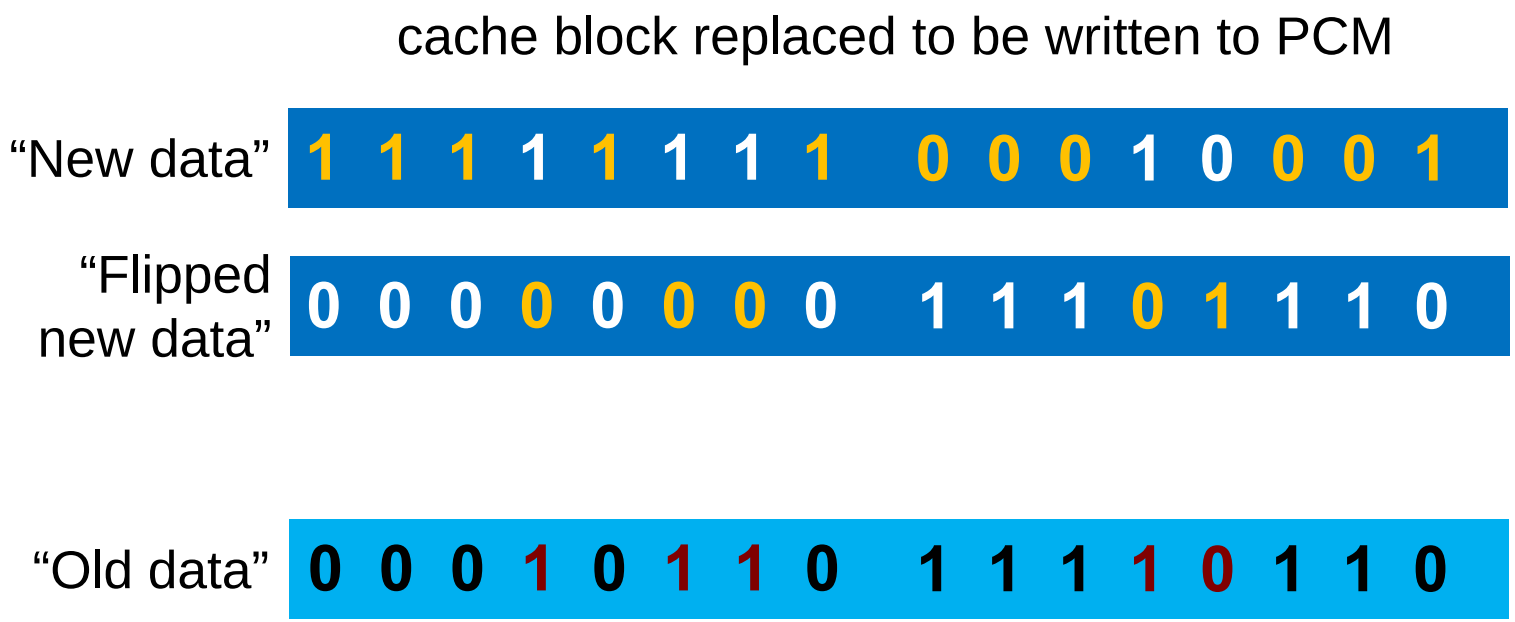
cache block replaced to be written to PCM



E.g.: Flip-N-Write [MICRO '09]

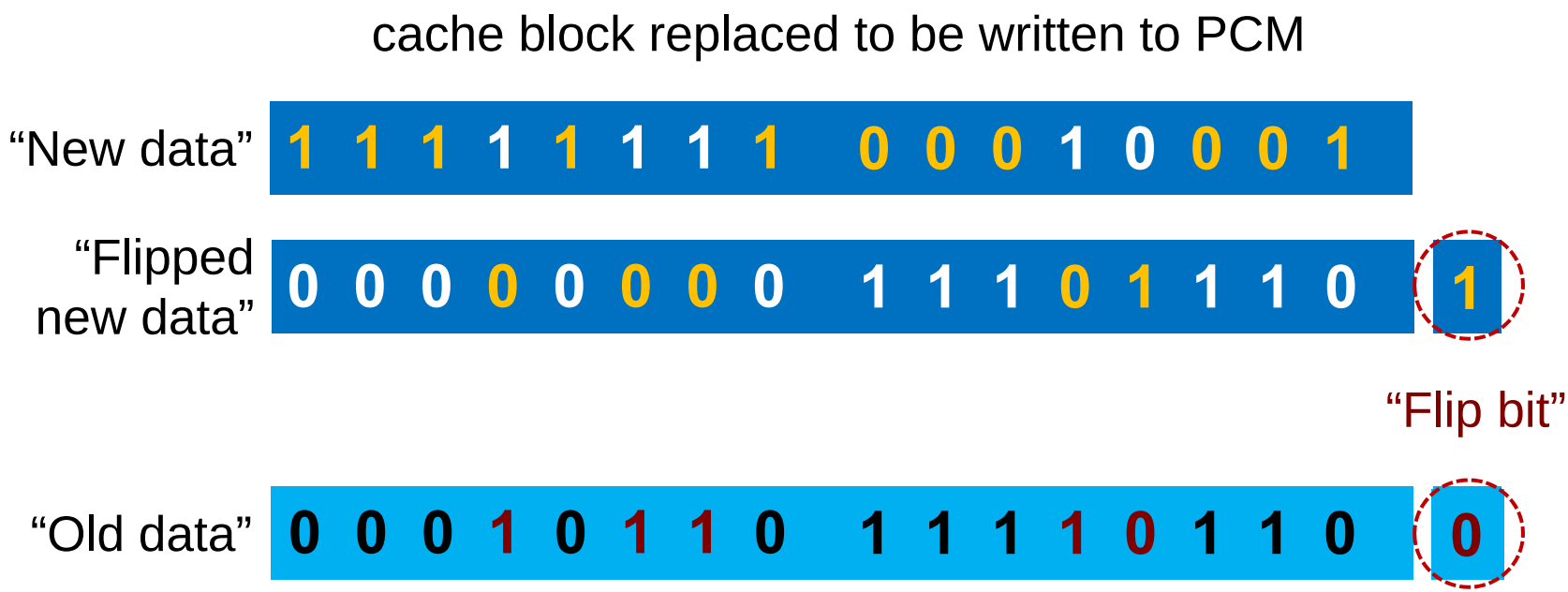
cache block replaced to be written to PCM

E.g.: Flip-N-Write [MICRO '09]



Only five bits are different!

E.g.: Flip-N-Write [MICRO '09]



(5+1) bits need be updated…

E.g.: Flip-N-Write [MICRO '09]

- Savings in bit updates can improve energy and endurance

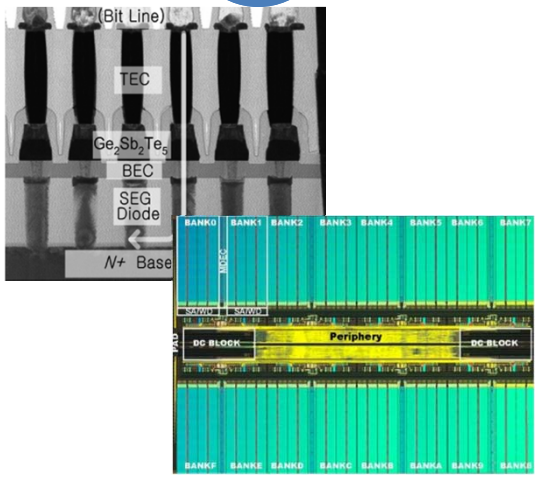
- Flip-N-Write updates *N*/2 bits maximum
- Write-current limited write time (M bits, S bps)
  - Conventional: $(M/S)\times T_{SET}$
  - Differential write: $T_{READ} + (M/S)\times T_{SET}$
  - Flip-N-Write: $T_{READ} + (M/2S)\times T_{SET}$

Agenda

- PCM 101
- **Industry trends**
- PCM usage models
- Summary

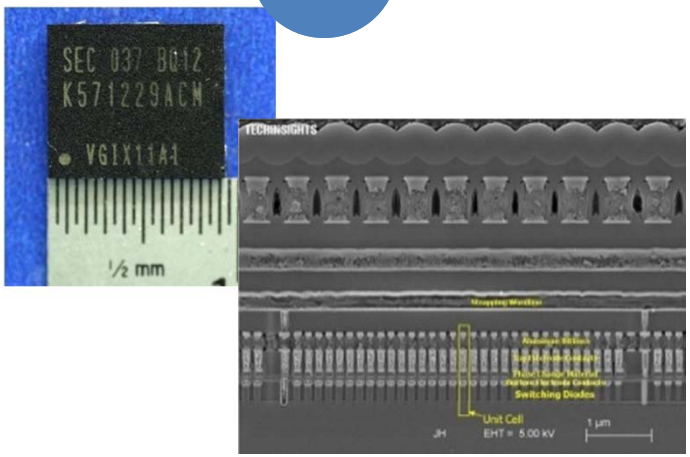
Samsung

Lee et al. ISSCC ’07
Lee et al. JSSC ’08



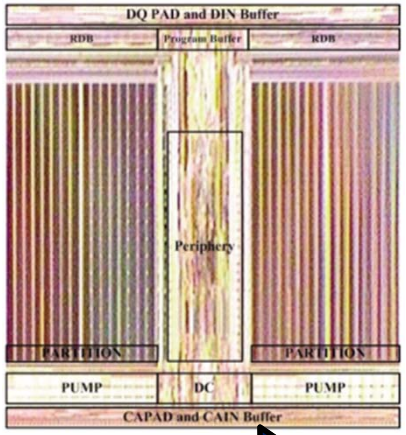
512Mb @90nm
Diode switch design
266MB/s read
4.64MB/s write (x16)

Techinsights decap ’10



512Mb @60nm?
Diode switch design
Believed to be a tech.-migrated design

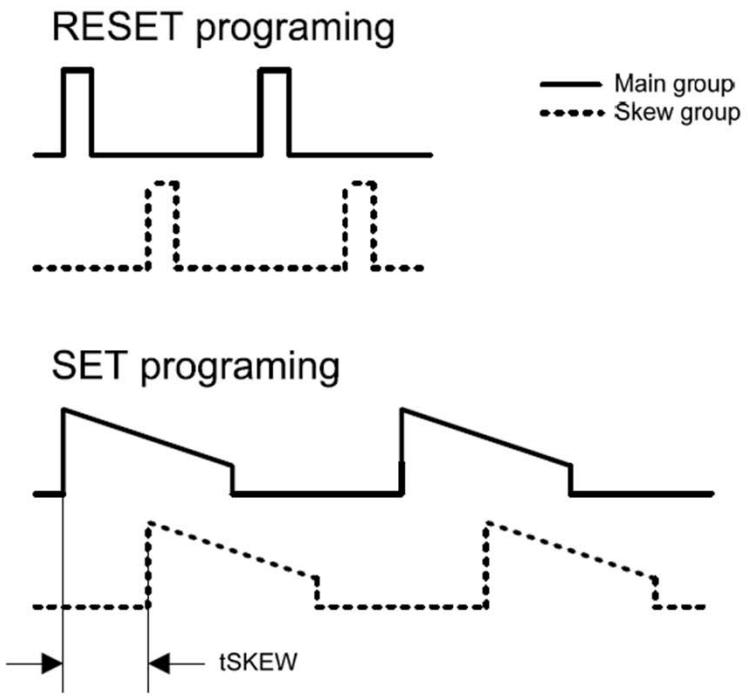
Chung et al. ISSCC ’11



1Gb @58nm
LPDDR2-N
“Write skewing”
6.4MB/s write
“DCWI” (~Flip-N-Write)

Write skewing

- To distribute program current
  - Main and skewed group
- Reduced peak current injected to the write driver
  - ~70% of the conventional simultaneous-write scheme



(H. Chung et al. ISSCC ’11)

Data comparison write w/ inversion

- Concept

Set state: 1

Reset state: 0

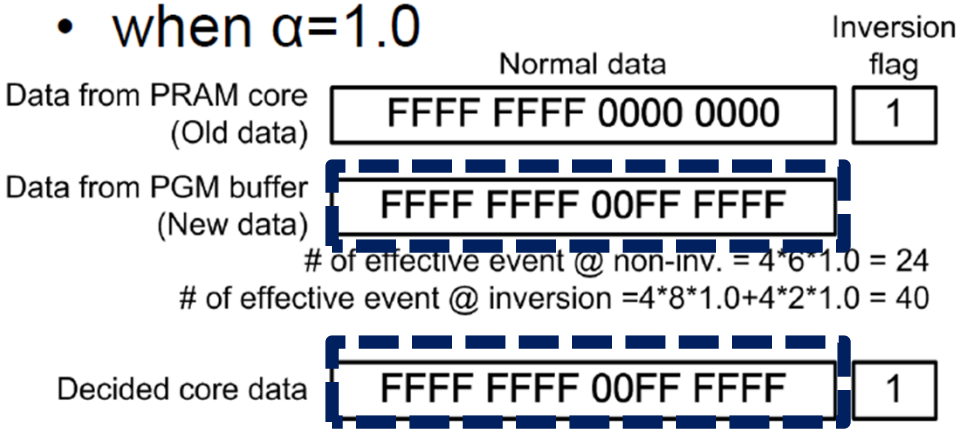
OLD (Core)	NEW (PB)	EVENT 1 → 0	EVENT 0 → 1	Eff. # of event	Eff. # of event
0	0	0	0	0	0
0	1	0	1	0	1*α
1	0	1	0	1	0
1	1	0	0	0	0

α: the ratio of energy to change the state

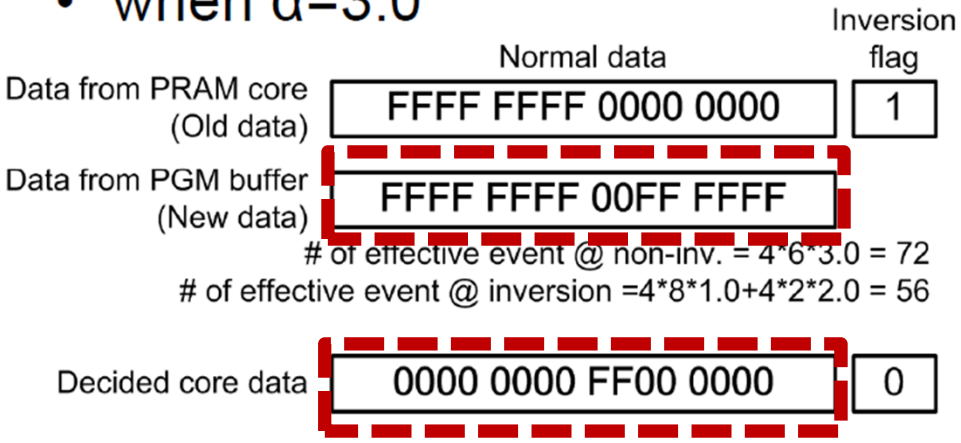
(H. Chung et al. ISSCC '11)

Data comparison write w/ inversion

- when α=1.0

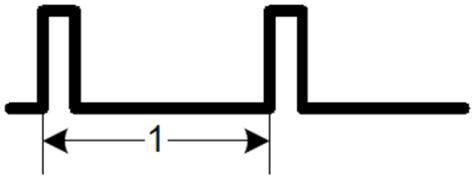


- when α=3.0

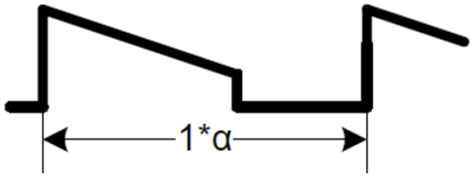


flag 1: non-inverted
flag 0: inverted

RESET programing



SET programing



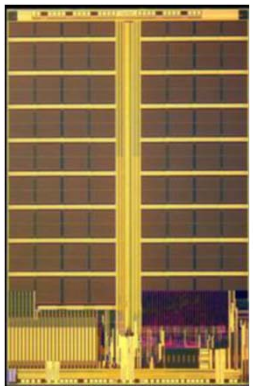
(H. Chung et al. ISSCC '11)

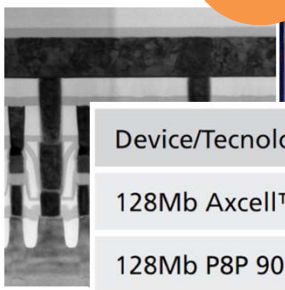
Numonyx (now Micron)

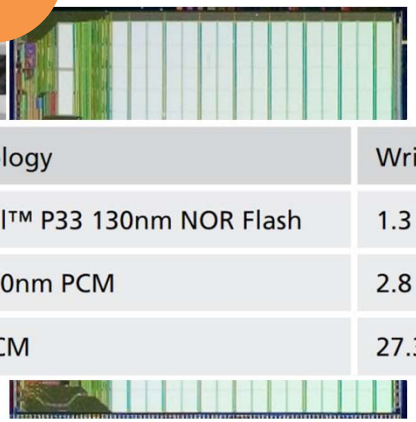
Early access program (2009)

Numerous press releases (slated for MP in 2011)

(2011~2012?)







Device/Tecnology	Write Bandwidth (Mb/s)
128Mb Axcell™ P33 130nm NOR Flash	1.3
128Mb P8P 90nm PCM	2.8
1Gb 45nm PCM	27.3

(www.micron.com)

(Servalli, IEDM '09)

"Alverstone" (OMNEO)
128Mb @90nm
TR switch design
40MB/s read (?)
<1MB/s write (?)

"Bonelli"
1Gb @45nm

1.8V I/O

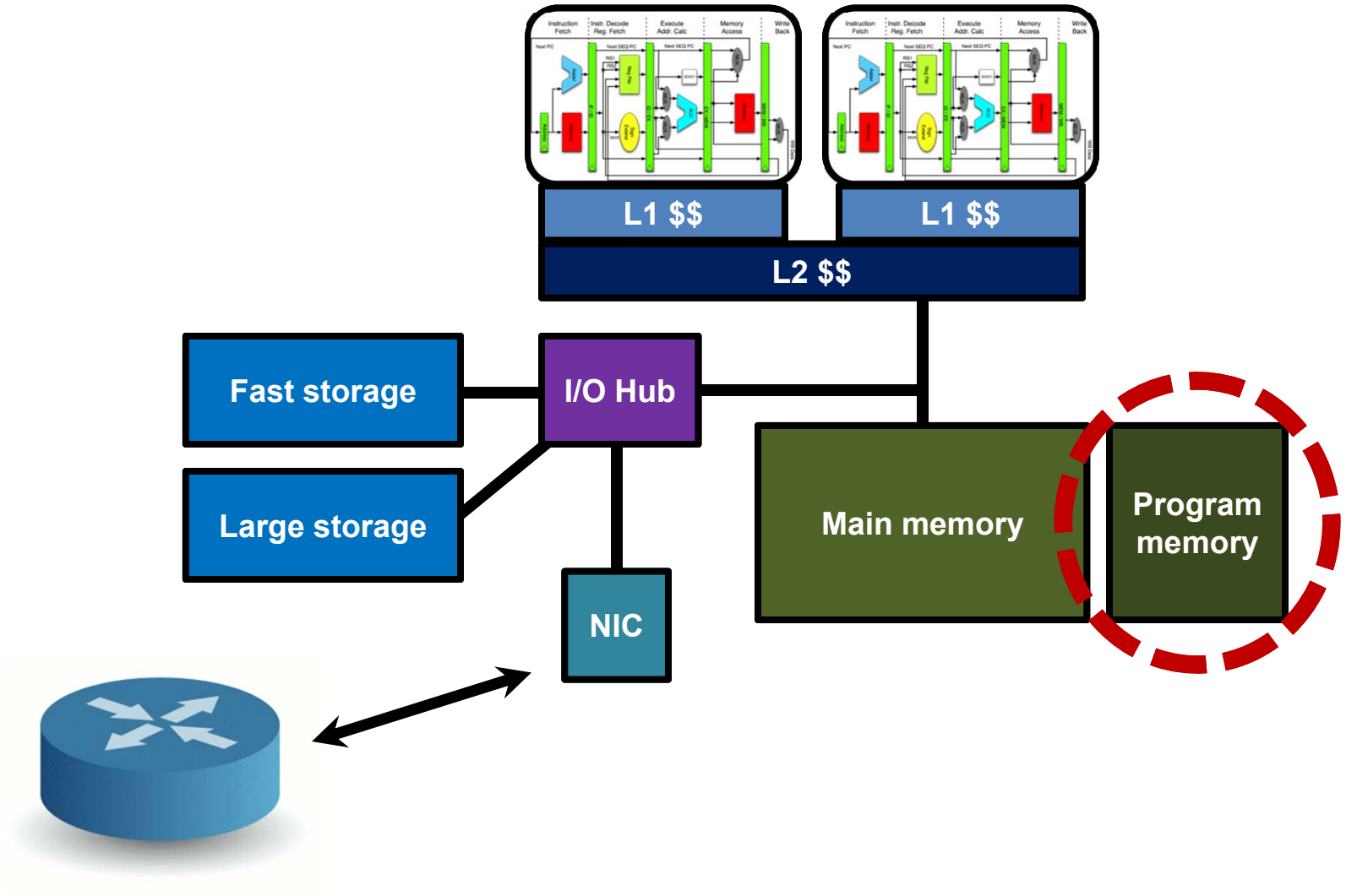
"Imola" and "Mandello"
2Gb & 4Gb @45nm

1.2V & 1.8V I/O
LPDDR2-NVM &
DDR3-NVM

Agenda

- PCM 101
- Industry trends
- **PCM usage models**
- Summary

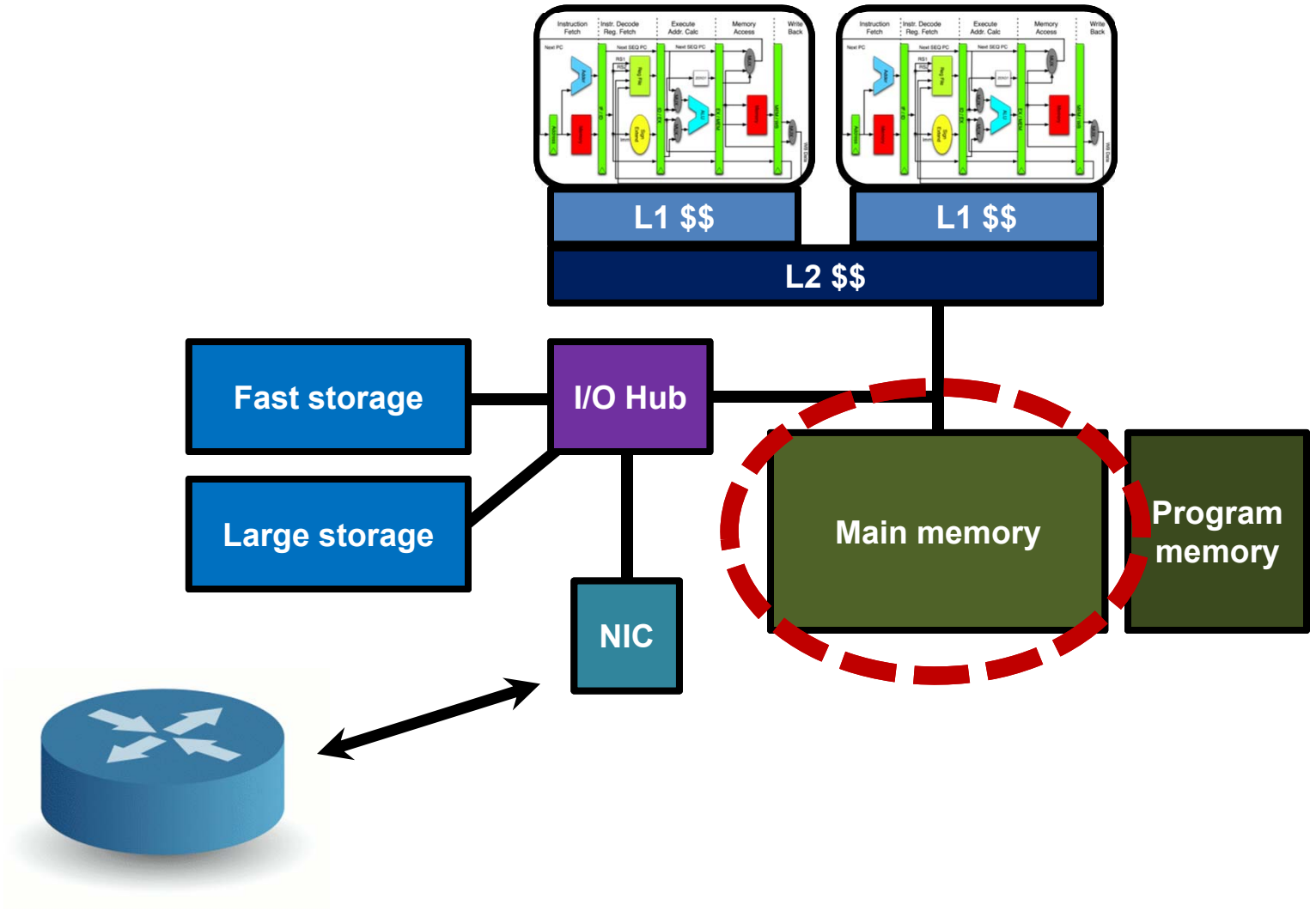
Where does PCM fit?



PCM as program memory

- **“Replace NOR in embedded platforms”**
  - Fast read speed, good retention, reasonable write bandwidth (a few MB/s)
  - First target of both Micron & Samsung
- PCM has an edge due to density, scalability, and write speed (use scrubbing to improve reliability)
- Today, common NOR parts are 64Mb~512Mb
- Initial PCM offerings
  - Micron: 128Mb (x8, x1) moving to 1Gb (x16?)
  - Samsung: 512Mb (x16) moving to 1Gb (x16)

Where does PCM fit?



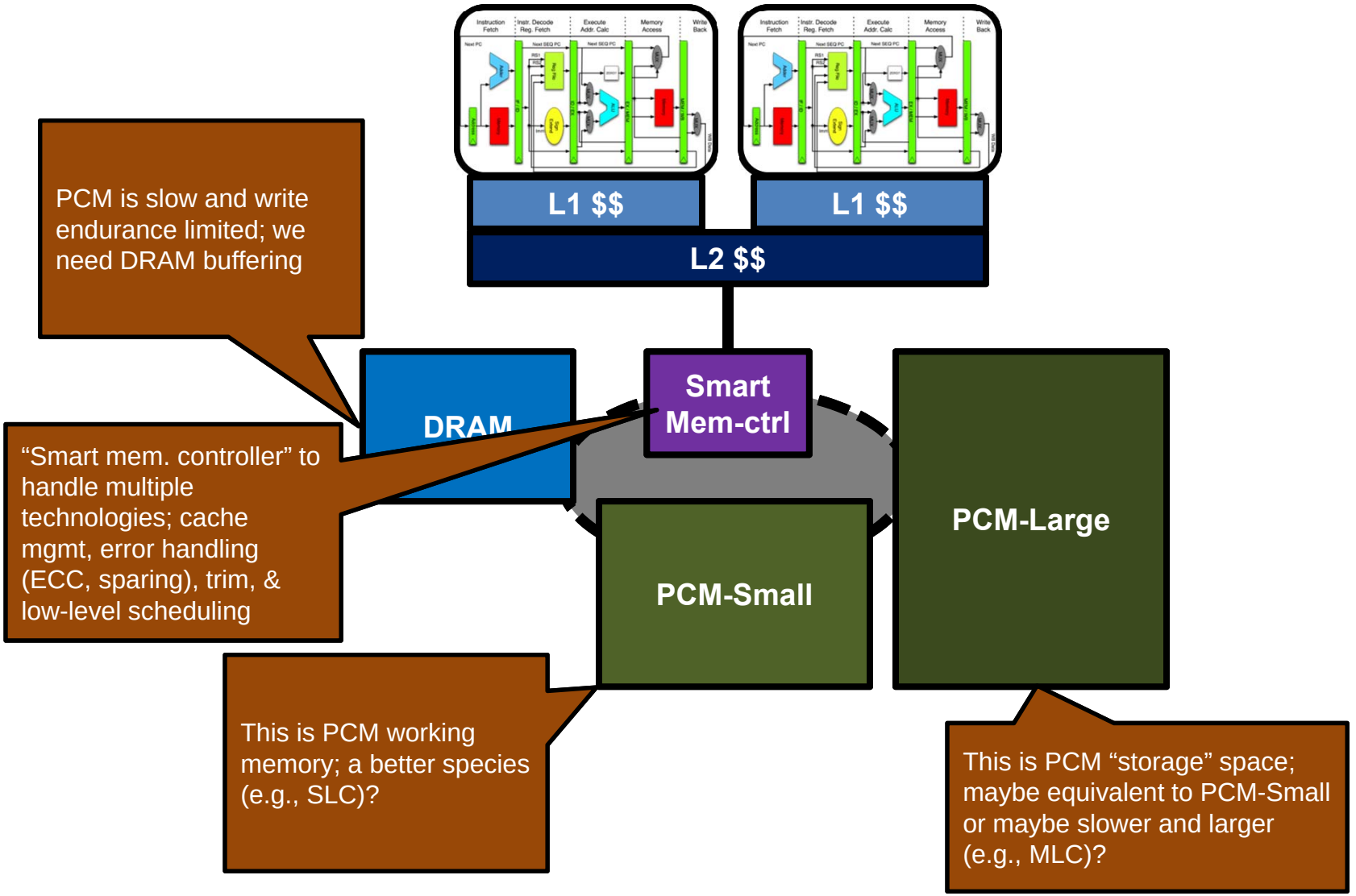
PCM main memory?

- **"Replace (a good chunk of) DRAM"**
- **Why this makes sense**
  - DRAM scaling is hard (no known solutions at < 20nm)
  - DRAM consumes more power than wanted, even at idle time
  - PCM can scale better; PCM is power-efficient on reads & at idle time
- **Why this may NOT happen (easily)**
  - PCM has poor write bandwidth (as of now)
  - DRAM camp has been capable of overcoming hurdles
    - E.g., new DRAM designs and interfacing schemes under consideration to improve on power & reliability
  - PCM is not getting enough attention (~investment)
  - Other competing technology maturing in the mean time?

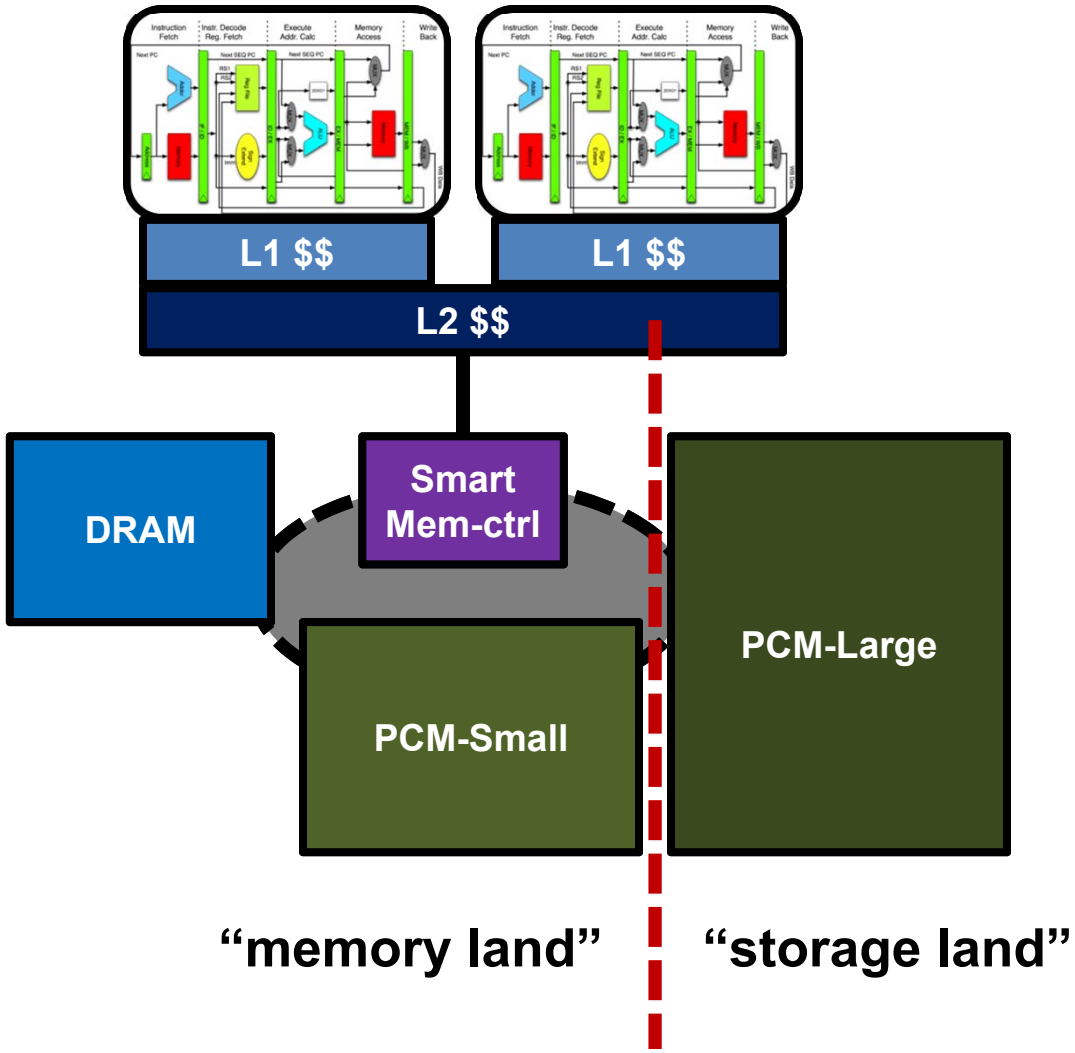
PCM main memory?

- **"Replace (a good chunk of) DRAM"**
- **Why this is attractive**
  - PCM can enable low-power servers [ISCA '09]
  - Instant on/off [Prof. Noh's talk at Pitt, '09]
  - Fast, potentially no-overhead checkpointing and versioning [Venkataraman et al., FAST '11]
  - File system meta-data storage [Park and Park, IEEE TC '11]
- **More usage models**
  - PCM provides working memory space and (very high-speed) storage space
  - Fast application launching via pre-loaded binary image
  - Fast local checkpointing in supercomputing platforms
  - Novel applications that require gigantic memory space

PCM main memory?



Traditional dichotomy

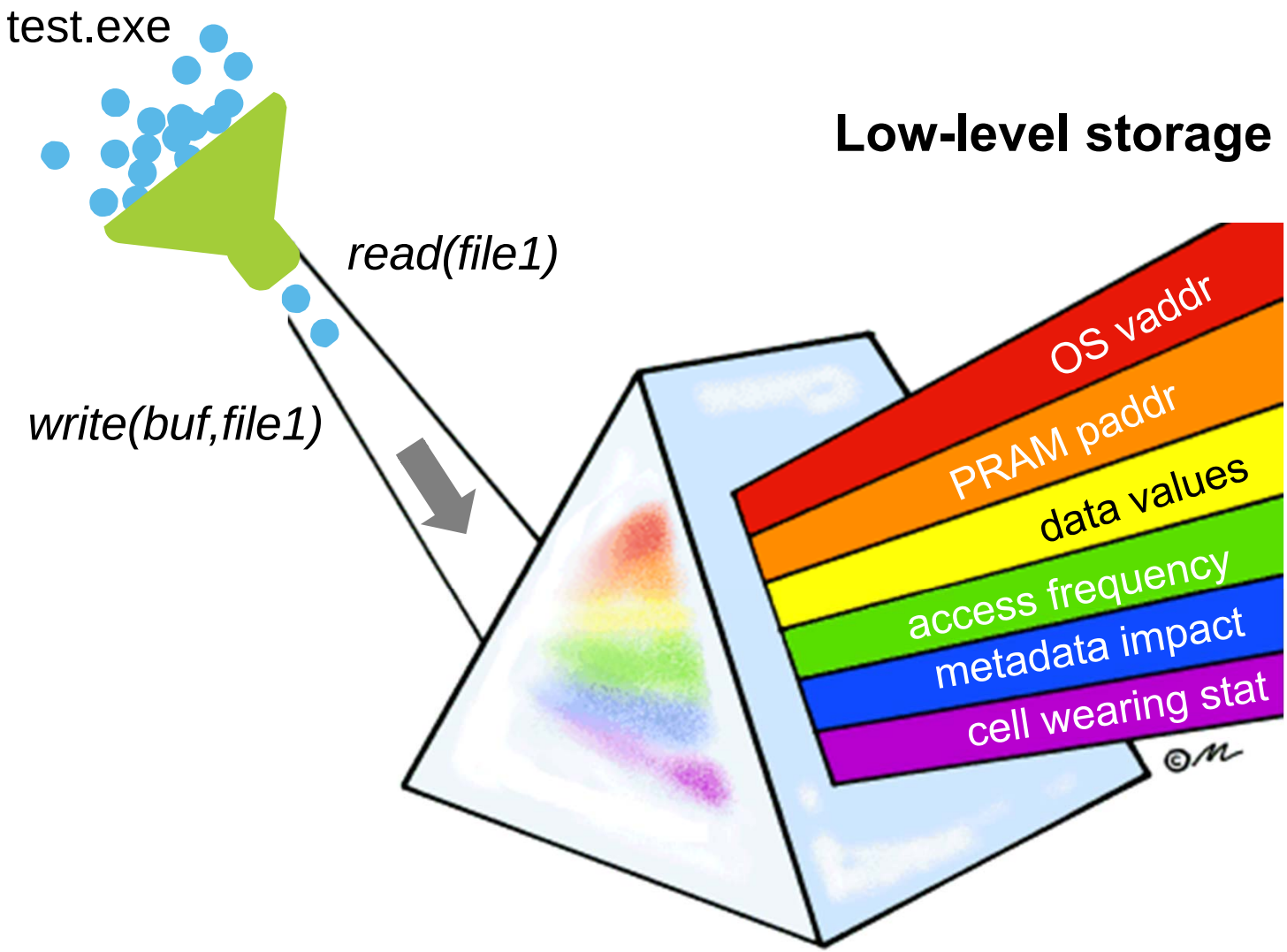


PRISM

- =**P**e**r**s**i**stent RAM **s**torage **m**onitor
  - To study a PRAM storage's low-level behavior
  - To guide PRAM storage designs

- [Jung and Cho, ISPASS '11]

PRISM



Low-level storage behavior

address mapping?

wear leveling?

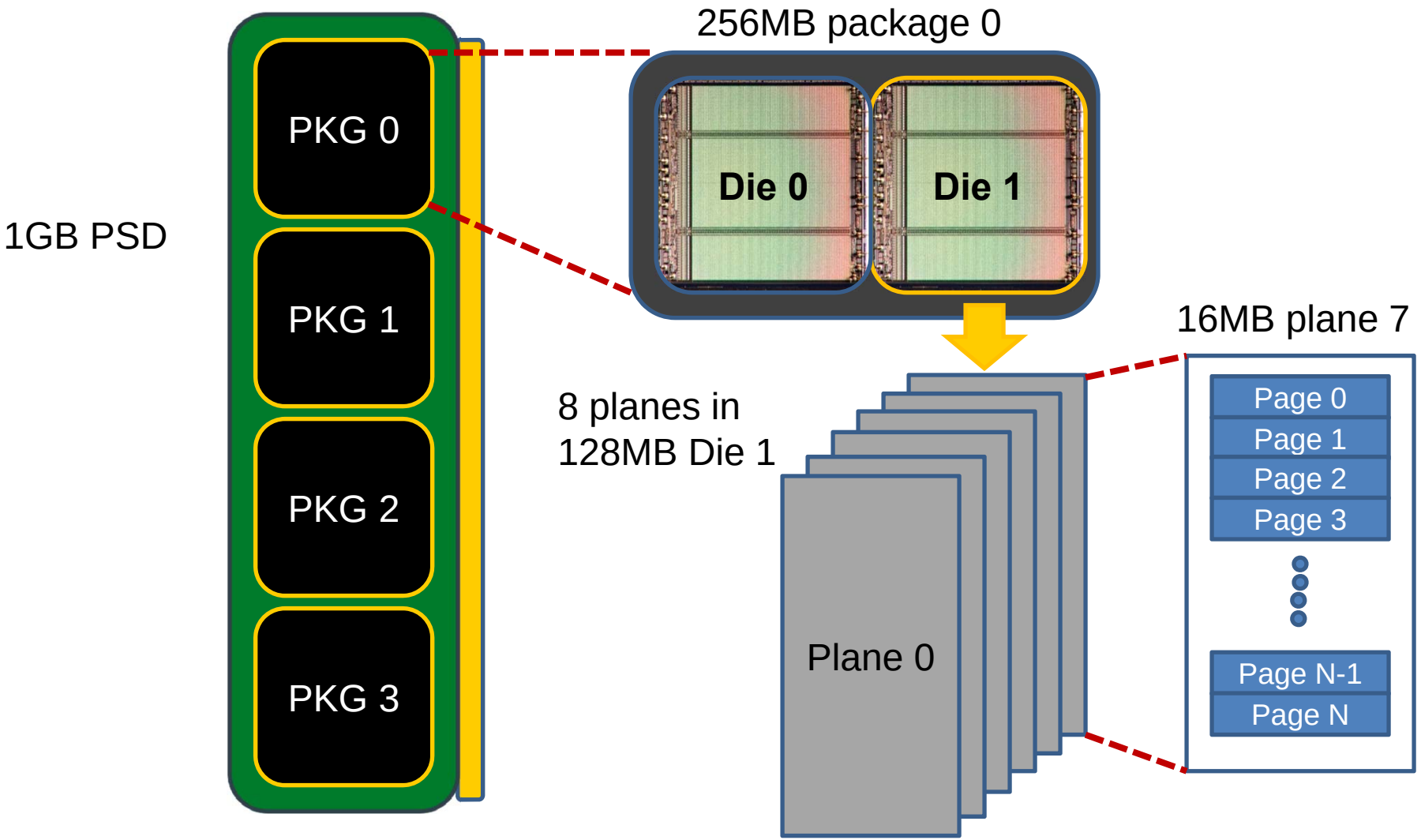
bit masking?

parallelism?

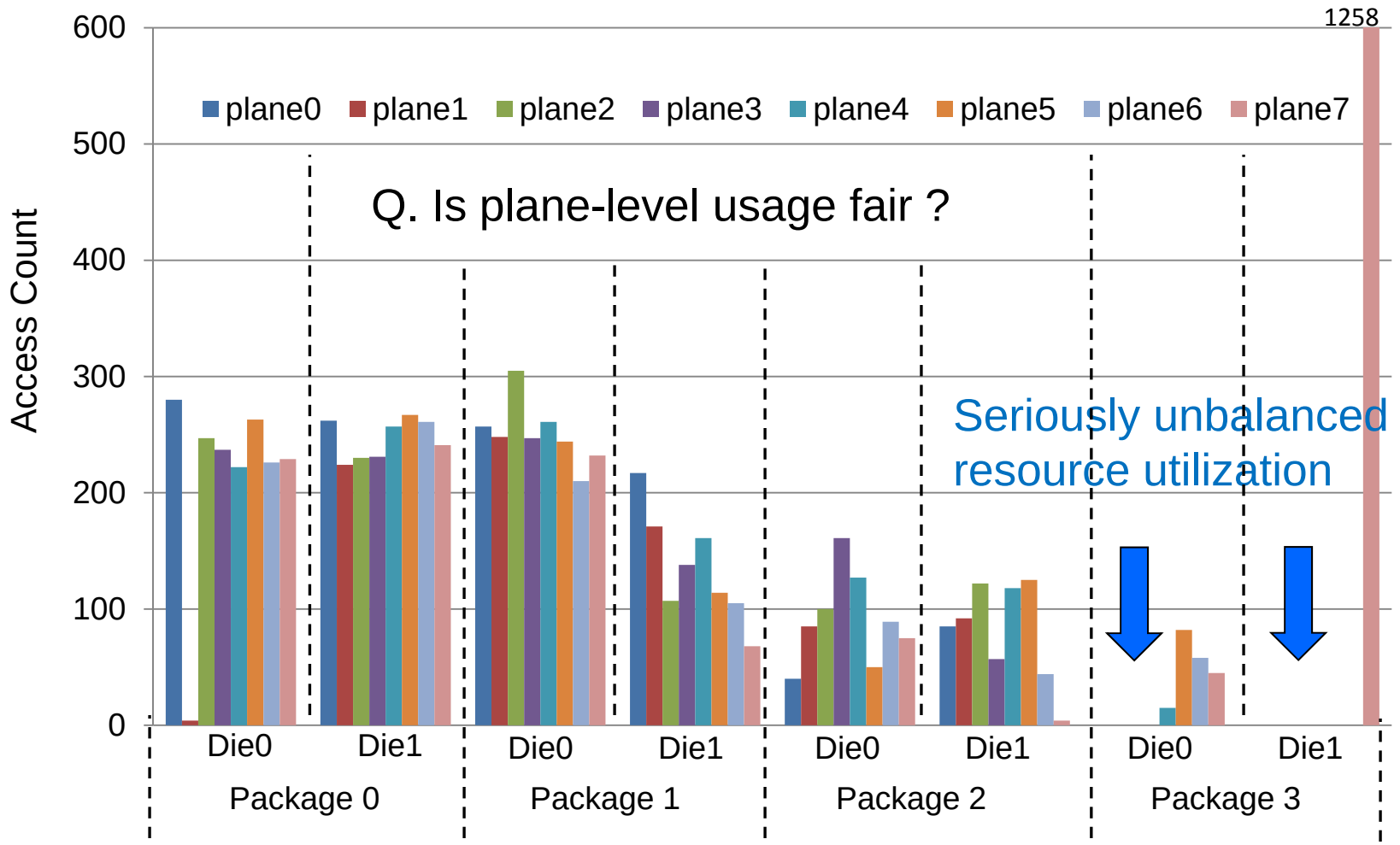
resource conflict?

etc…

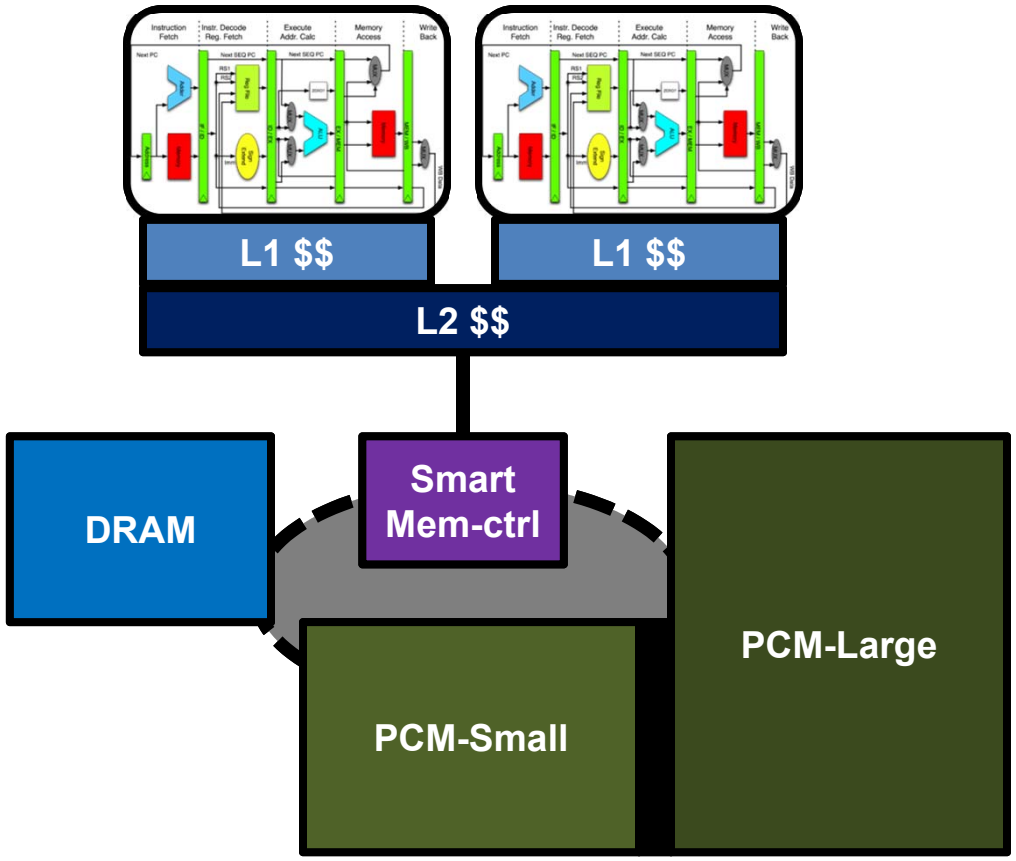
PRISM (example)



PRISM (example)

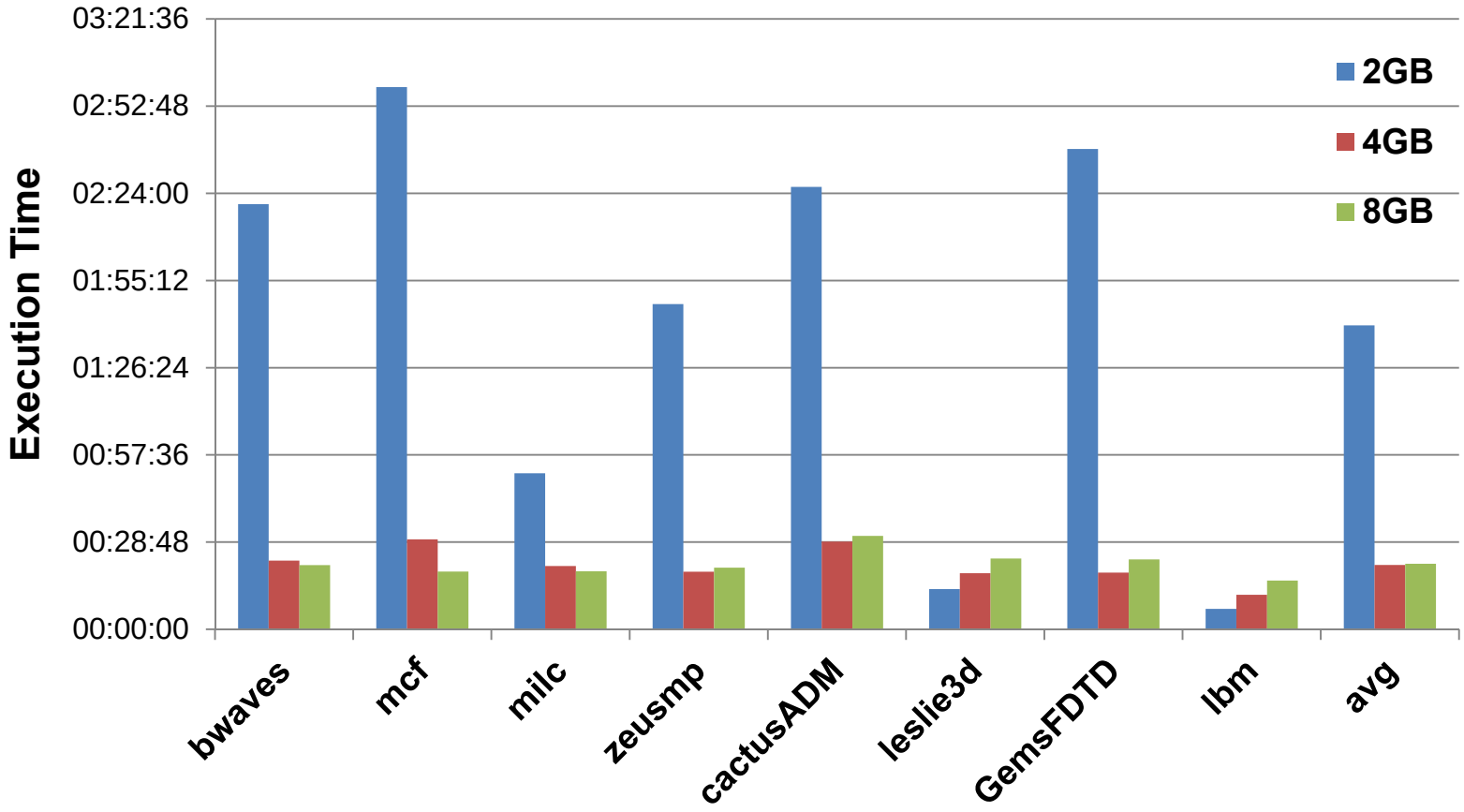


Memory + storage = memorage?

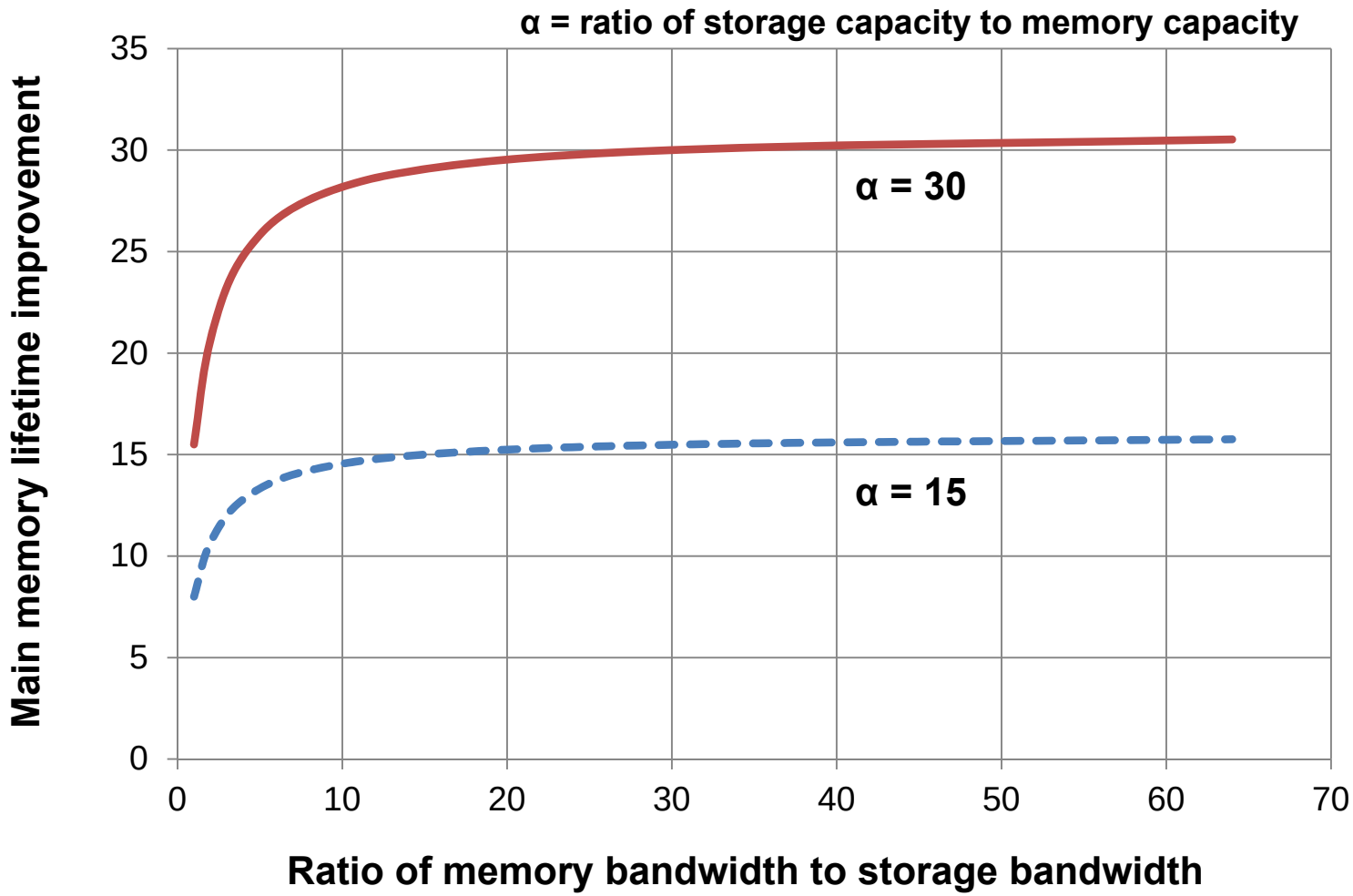


“memorage” [Jung and Cho, CF ’11)

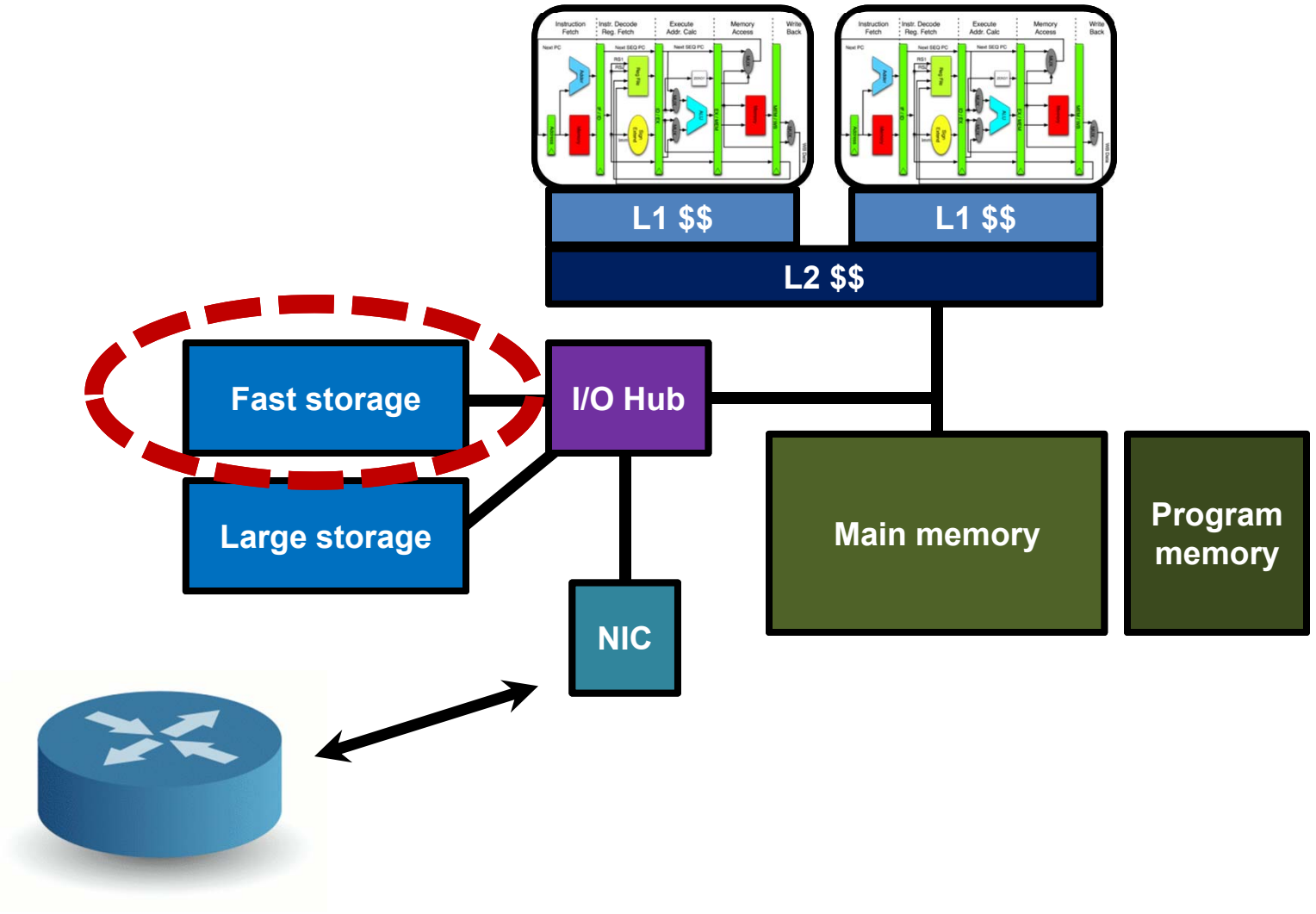
Memorage benefits (elapsed time)



Memorage benefits (lifetime)



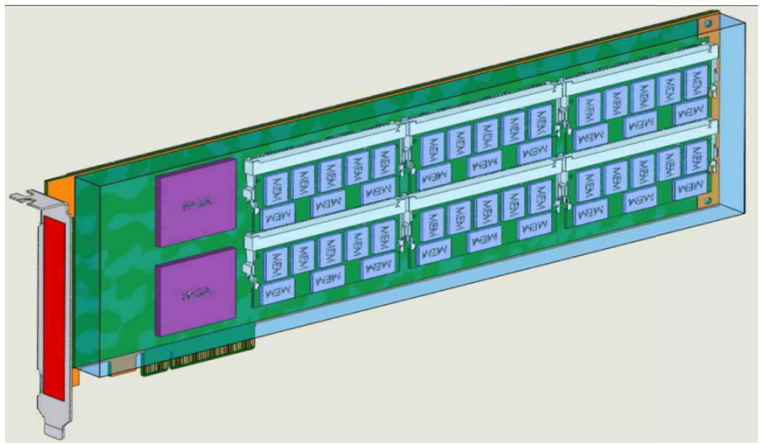
Where does PCM fit?



PCM as main storage medium

- **“Replace NAND in high-speed SSDs”**
- PCM has good potential (theoretically)
  - Lower latency than NAND (~100ns vs. ~100μs)
  - More scalable than NAND (~10nm vs. ~20nm)
  - Much simpler management (e.g., in-place update)
  - Potentially good bandwidth
  - Fast paging storage?
- **Huge challenges ahead**
  - NAND density improving, at least for now (scaling & TLC + better error handling)
  - NAND bandwidth (not latency) improves
  - NAND momentum ensures continued investment

E.g.: PCM SSD



	2010
Min Density	64GB
Max Density	512GB
Interface	PCIe 2.0x8
Read Bandwidth	4.0 GB/s
Write Bandwidth	400 MB/s
Input Voltage	Up to 12V
Power	20W
Read Latency	5µS (hw)
Write Latency	150µS
Physical Dimensions	Full Size PCIe Card
Temperature	0°to +55°C

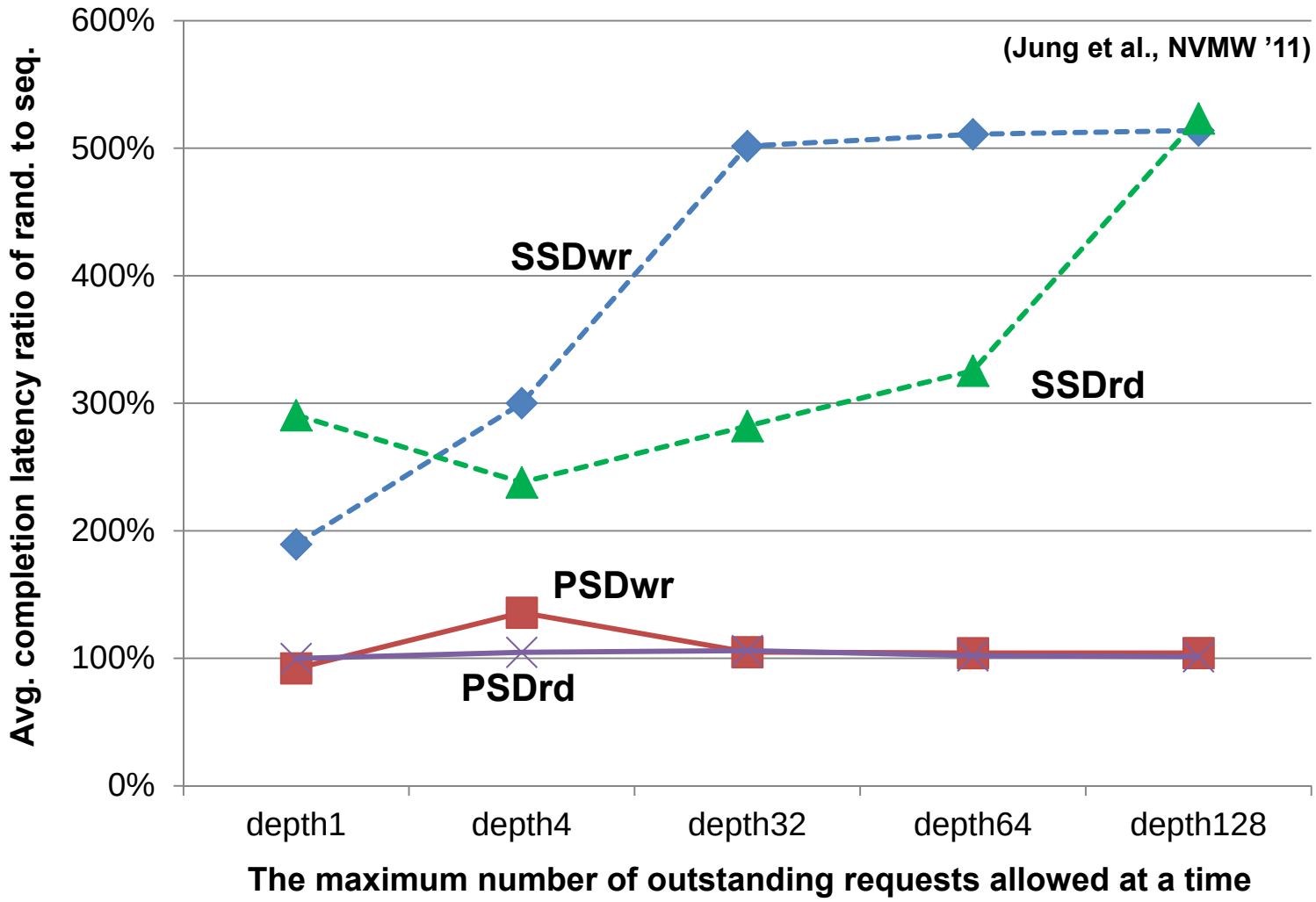
(Numonyx)

(Fusionio)

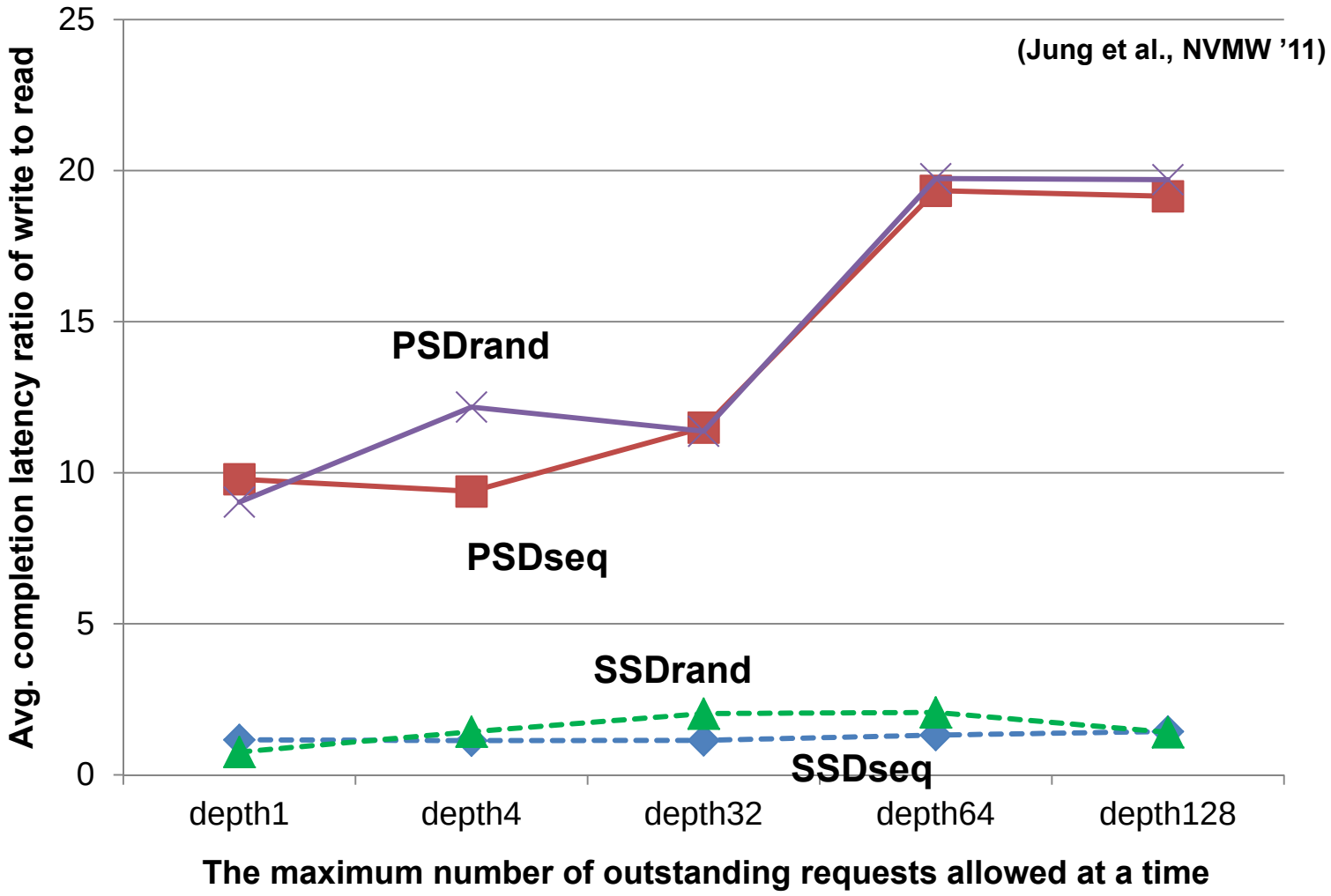
ioDrive Duo Capacity	320GB	640GB	640GB
NAND Type	SLC (Single Level Cell)	SLC (Single Level Cell)	MLC (Multi Level Cell)
Read Bandwidth (64kB)	1.5 GB/s	1.5 GB/s	1.5 GB/s
Write Bandwidth (64kB)	1.5 GB/s	1.5 GB/s	1.0 GB/s
Read IOPS (512 Byte)	261,000	273,000	196,000
Write IOPS (512 Byte)	262,000	252,000	285,000
Mixed IOPS (75/25 r/w)	238,000	236,000	138,000
Access Latency (512 Byte)	26 µs	26 µs	29 µs
Bus Interface	PCI-Express x4/x8 or PCI Express 2.0 x4		

ioDrive Octal Capacity	5.12TB
NAND Type	Multi Level Cell (MLC)
Read IOPS (512 B)	1,190,000
Write IOPS (512 B)	1,180,000
75/25 Mix IOPS (512 B)	729,000
Read Bandwidth (64 kB)	6.0 GB/s
Write Bandwidth (64 kB)	4.4 GB/s
Access Latency (512 Byte)	30µs
Bus Interface	PCI-Express x16 Gen2.0
Operating Systems	64-Bit Microsoft Server 2003/2008, 64-Bit Microsoft Windows XP/Vista/Win7, RHEL 4/5, SLES 10/11, OEL v4/v5

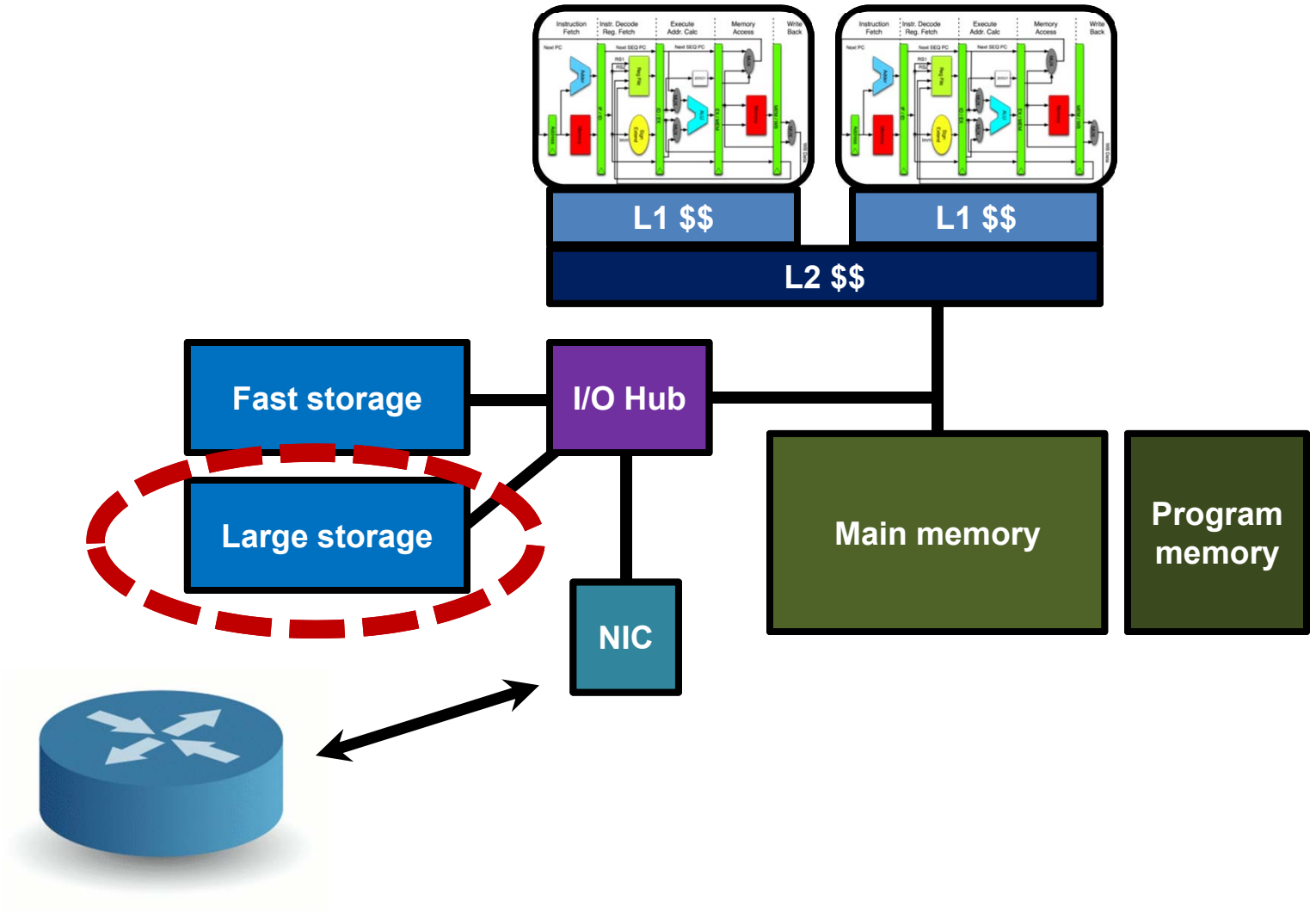
E.g.: PCM SSD


(Jung et al., NVMW '11)
Avg. completion latency ratio of rand. to seq.
600%
500%
400%
300%
200%
100%
0%
SSDwr
SSDrd
PSDwr
PSDrd
depth1
depth4
depth32
depth64
depth128
The maximum number of outstanding requests allowed at a time

E.g.: PCM SSD


(Jung et al., NVMW '11)
Avg. completion latency ratio of write to read
25
20
15
10
5
0
PSDrand
PSDseq
SSDrand
SSDseq
depth1
depth4
depth32
depth64
depth128
The maximum number of outstanding requests allowed at a time

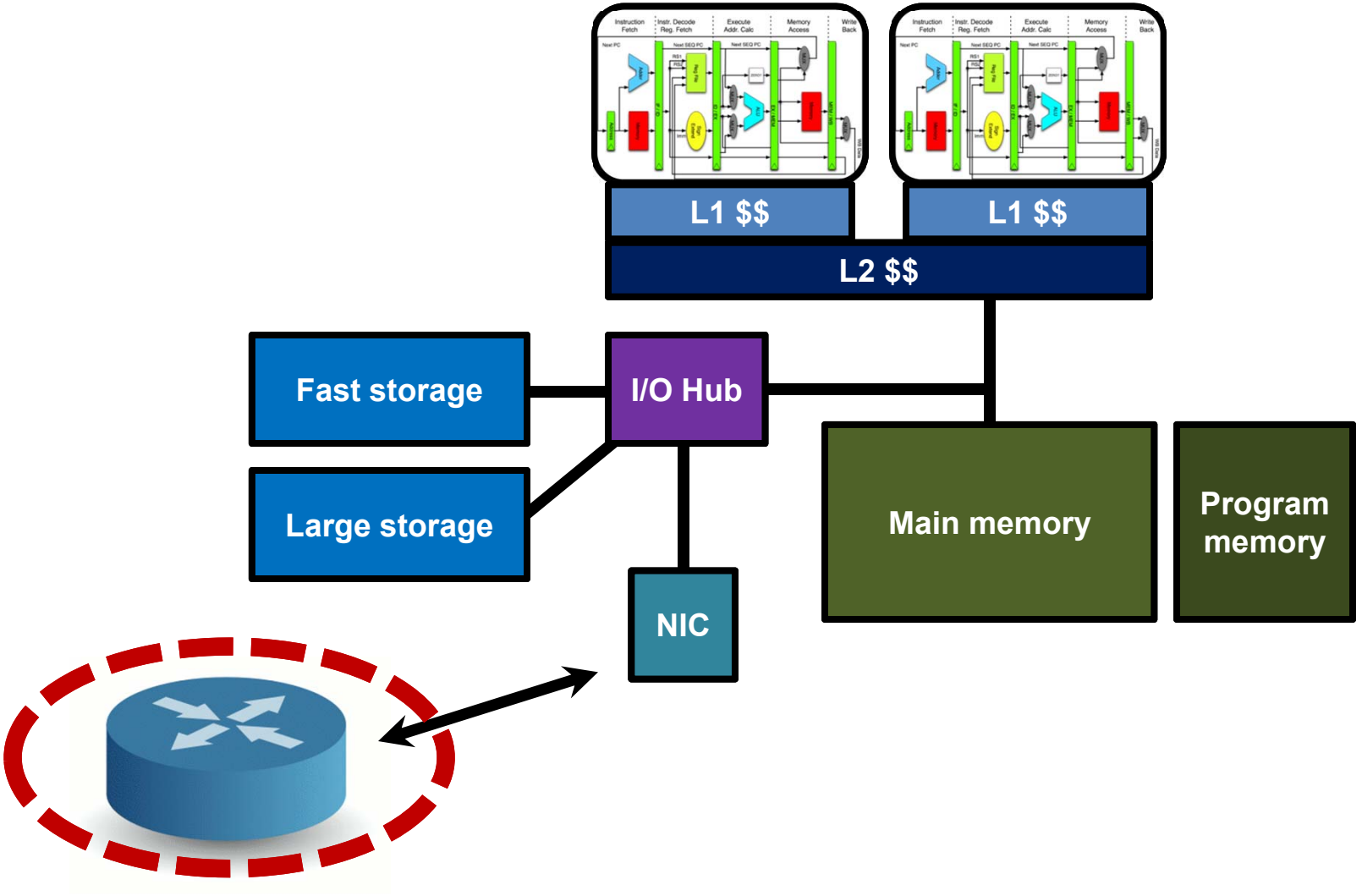
Where does PCM fit?



PCM as hidden specialist in a drive

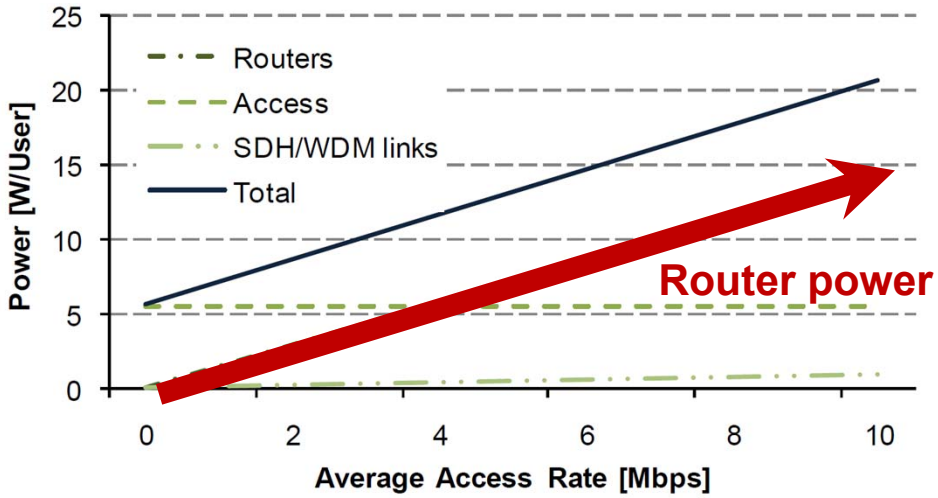
- **“Provide specialized non-volatile capacity”**
- PCM can help boost the performance of HDD
  - Provide fast storage capacity at tier 1
  - Capture small writes, keep working set, and minimize arm movements
  - But can’t NAND do the same? (many hybrid approaches exist)
- PCM can help ease NAND write complexities
  - E.g., [Sun et al., HPCA ’09][Kim et al., EMSOFT ’08]
  - NAND write endurance worse than PCM by orders of mag.
    - Total write data volume = $C(PCM) \times 10^{xxx} + C(NAND) \times 10^{yyy}$
  - NAND latency slower than PCM
    - Similar reasoning about performance is possible
- **For PCM to become the tier 1 within a storage device, how much capacity & bandwidth is needed?**

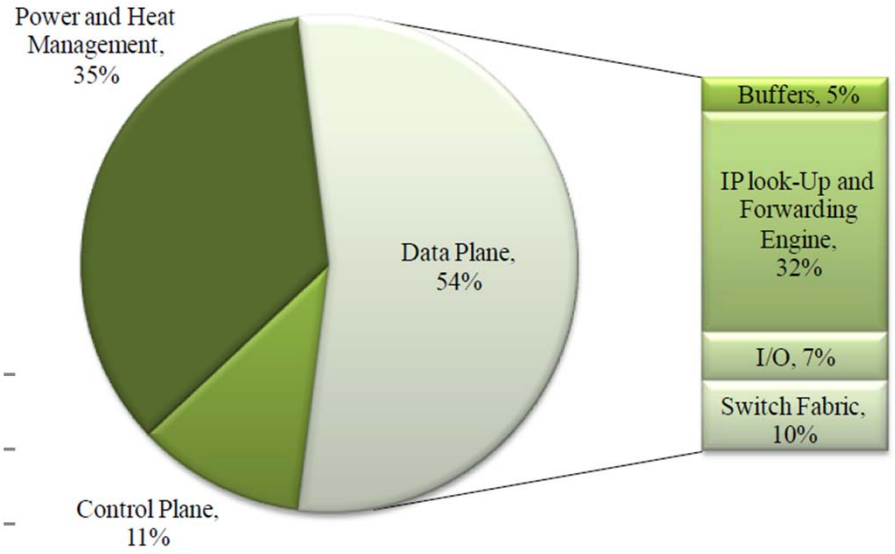
Where does PCM fit?



PCM as low-power search memory

- "Keep inter-networking tables in PCM"





(Bolla et al., IEEE Communications Surveys & Tutorials 2010)

PCM as low-power search memory

- **“Keep inter-networking tables in PCM”**
- Many data structures in inter-networking are read-intensive, e.g., IP lookup table, rules, patterns
  - Updates are relatively low bandwidth and incremental

- PCM could be used to construct TCAMs
- There are algorithms that use more regular RAM structures, e.g., [Hanna, Cho, and Melhem, Networking ’11]

- **This is a niche application domain where some interesting things can be done**

Agenda

- PCM 101
- Industry trends
- PCM usage models
- **Summary**

Summary

- PCM offers new opportunities to improve platform capabilities and end-user experiences
- Drop-in replacement of established memory technology does not work
  - Performance and power asymmetry
  - Errors
  - Falls short of fully utilizing PCM features

- **Optimal solutions will likely resort to horizontal & vertical collaboration of multiple system components**
  - **And the goal should be to improve the whole system**
  - **Are there new ideas?**

Why not …

- **Academic researchers**
  - … we explore system designs end-to-end (both horizontally and vertically) together, identify new opportunities, and specify performance, power, and reliability requirements?
  - Manufacturers will appreciate such a wish list

- **Industry**
  - … you “leak” information on real-world technical challenges you see and a realistic technology roadmap?
  - Researchers will love a laundry list of real problems

PCM@PITT

CAST

(Computer Architecture, Systems, & Technology) Laboratory