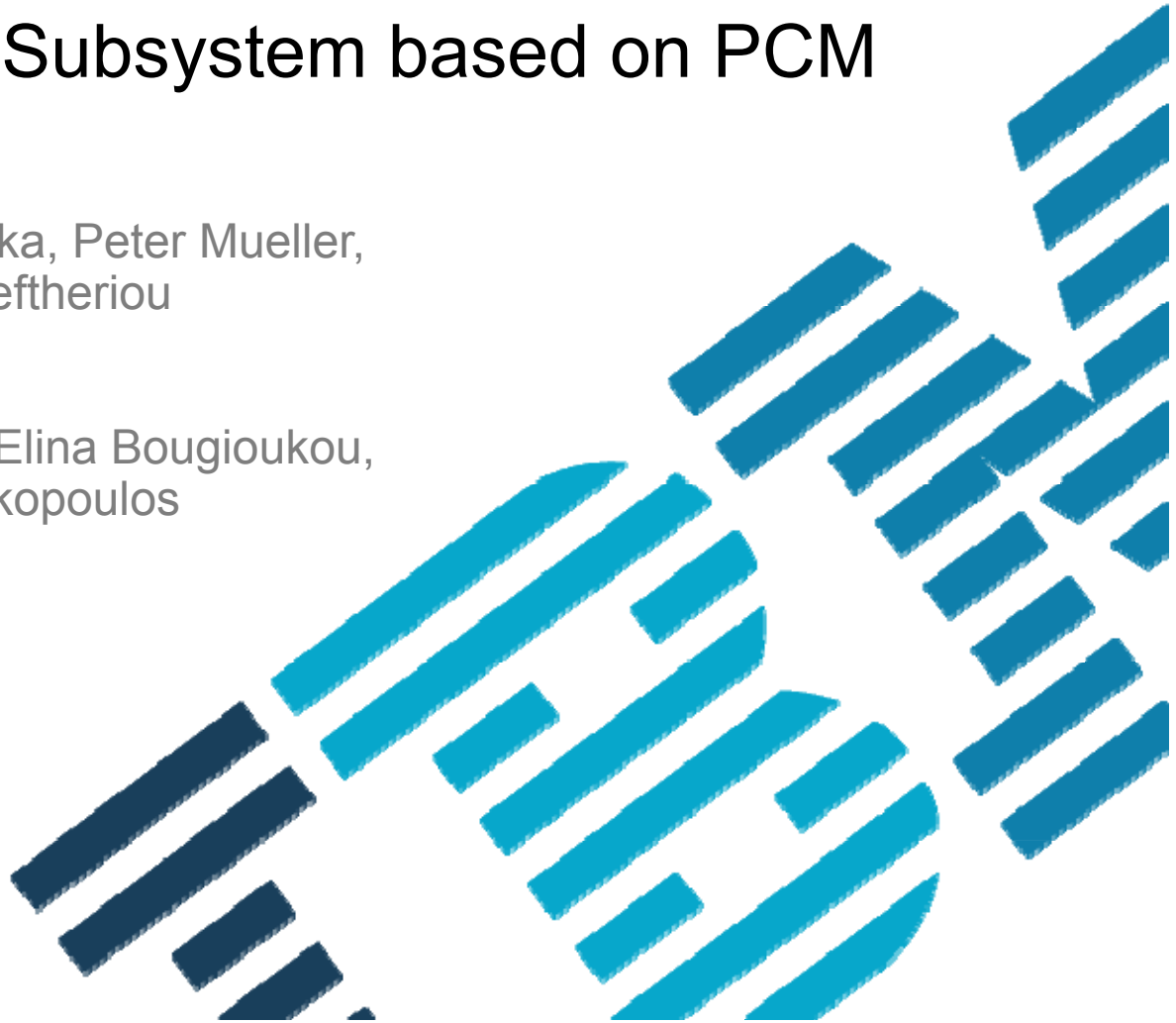
2014 Non-Volatile Memories Workshop
March 9 - 11, 2014

PSS
A Prototype Storage Subsystem based on PCM

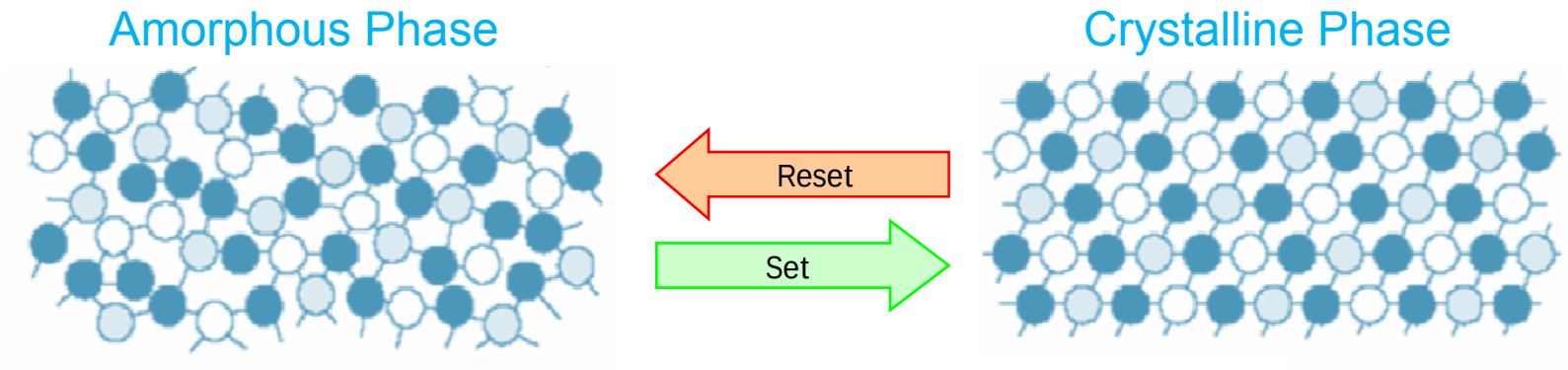
IBM Research – Zurich
Ioannis Koltsidas, Roman Pletka, Peter Mueller,
Thomas Weigold, Evangelos Eleftheriou

University of Patras
Maria Varsamou, Athina Ntalla, Elina Bougioukou,
Aspasia Palli, Theodore Antonakopoulos



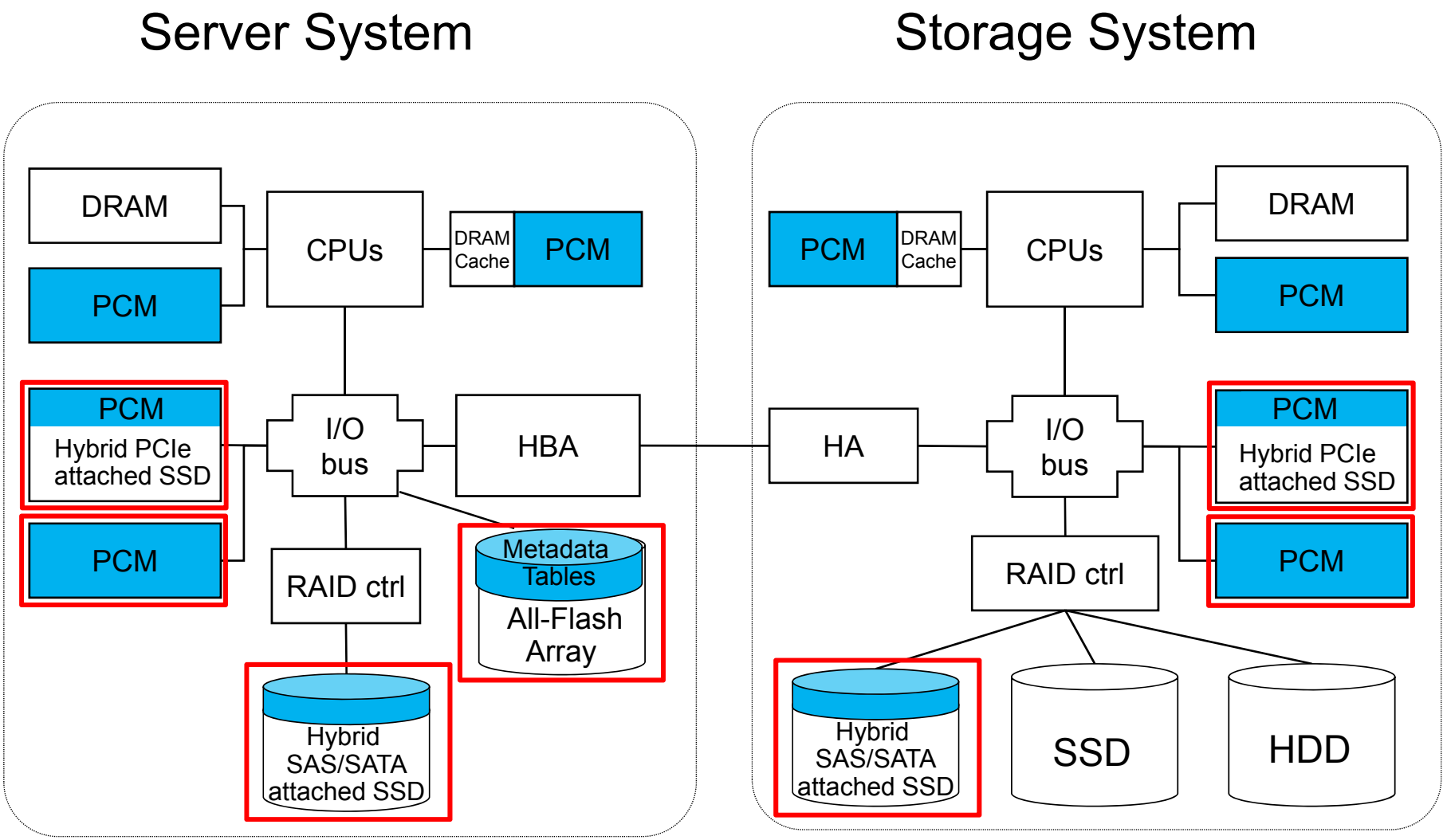
Phase Change Memory (PCM)

- Based on the thermal threshold switching effect of chalcogenidic meterials
- Two Phases:



- Phases have *very* different electrical resistances (ratio of 1:100 to 1:1000)
- Transition between phases by controlled heating and cooling
- Read time: 100-300 nsec
- Program time: 10-150 µsec
- PCM cells can be reprogrammed at least 10^6 times
- Performance and price characteristics between DRAM and Flash

Placing PCM in Servers and Storage Systems

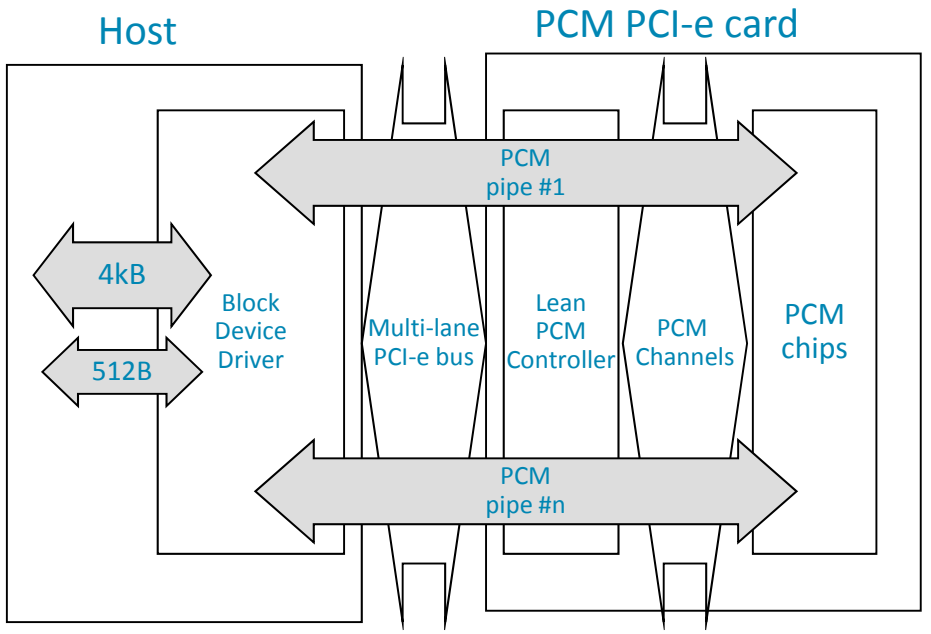


PSS: A PCM-based PCI-e Prototype Card

Goal:

Architect a PCM-based device and implement a fully-functional, high-performance PCI-e card

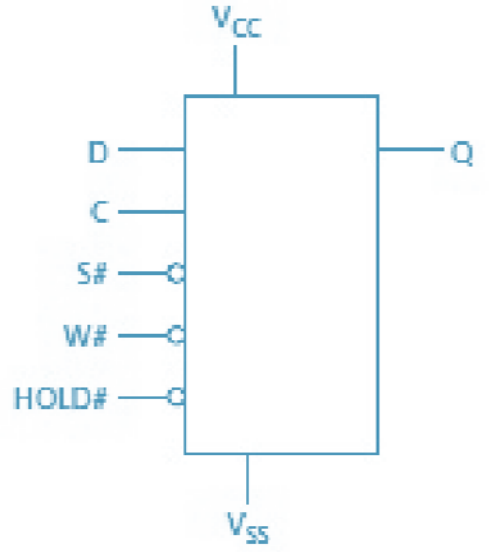
- Take advantage of the characteristics of PCM and mitigate its limitations
- Target is workloads dominated by 4kB requests
- Simple, lightweight hardware design
- System integration of multiple cards through software
- Emphasis on consistently low, predictable latency



- Limited in capacity due to the density of commercially available PCM parts (as of early 2013)
- Use cases:
  - Caching device
  - Metadata store
  - Backend for low-latency Key-Value store
  - Tiered storage device in a hybrid configuration with Flash

PCM Parts: Micron P5Q

- 90nm technology node
- 128 Mbit devices (NP5Q128AE3ESFC0E)
- SPI bus compatible serial interface
- Maximum clock frequency: 66 MHz
- 64-byte write buffer
  - 120 µsec average program time
  - about 0.5 MB/s write bandwidth



Block transfer time: 8.24 usecs (64+4 Bytes)

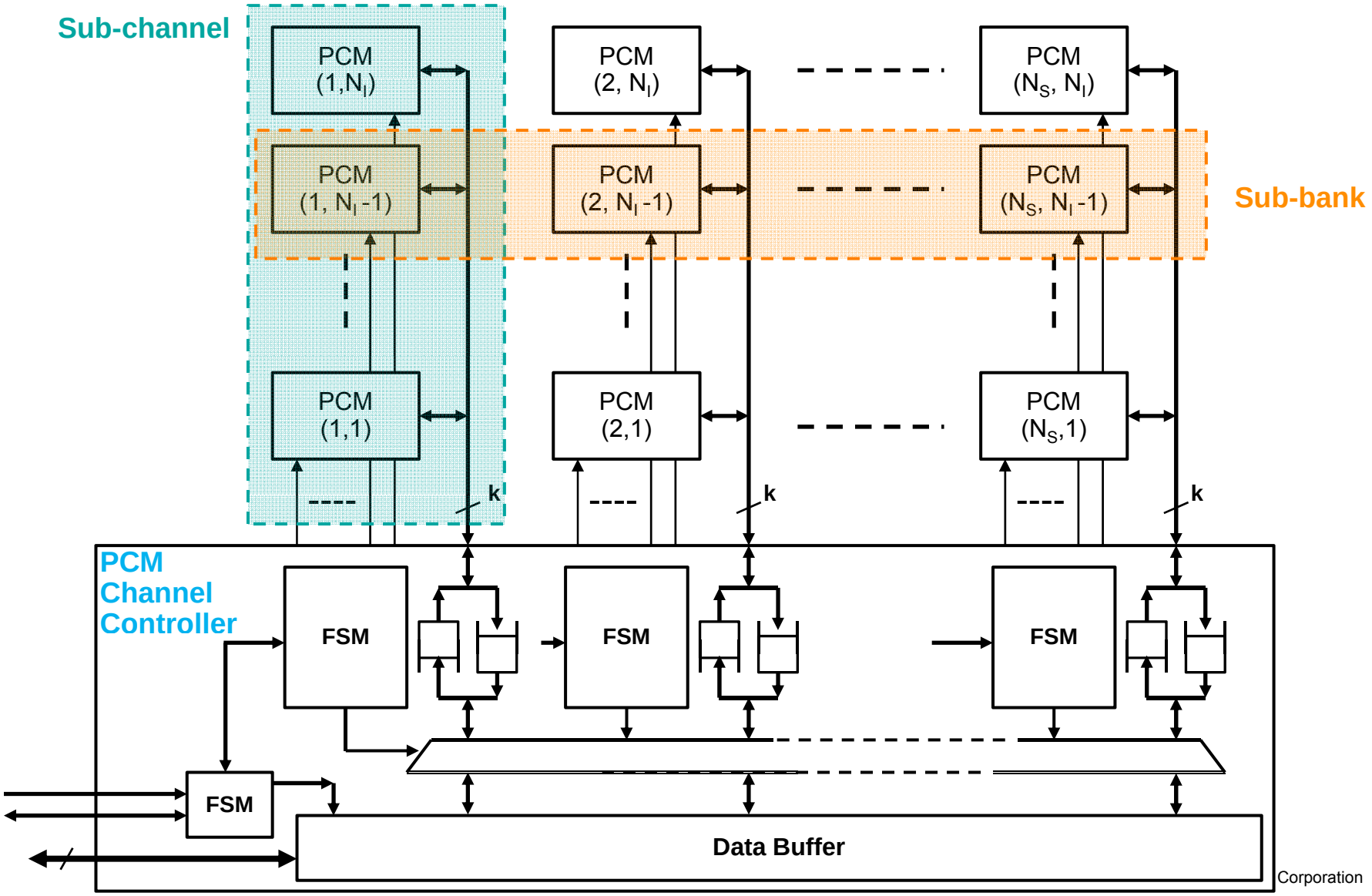
Sector I/O (512B + 64B):

Write: 1.15 **m**secs (0.86 kIOPs)

Read: 75.24 **u**secs (13.29 kIOPs)

} Very asymmetric read / write performance

2D PCM Channel Architecture



Read vs. Write Performance Trade-Offs

- A high degree of pipelining:
  - Increases the write performance
    - By having the long programming times overlap
  - May reduce the read performance
    - Read times are anyway very short
    - Less parallelism due to fewer I/O pins
- For a given budget of I/O pins:
  - More sub-channels ➔ Better read performance
  - More sub-banks ➔ Better write performance
- Application needs should drive the configuration
  - Channels with different geometry in the same device possible
- We chose a configuration that minimizes write latency without severely penalizing reads

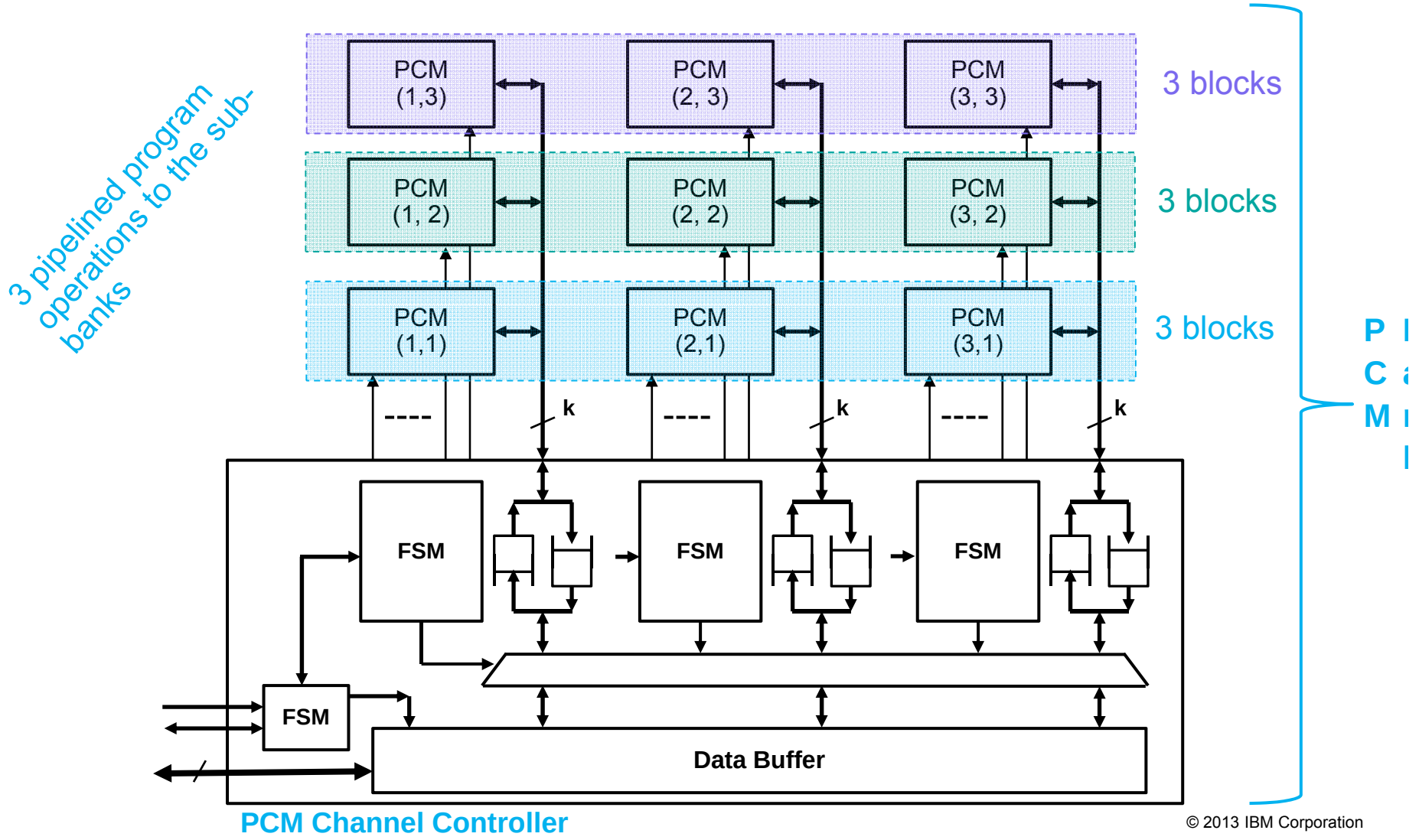
$$G_R = \frac{\left\lfloor \frac{P_B}{N_s N_D + N_C + N_l - 1} \right\rfloor}{\left\lfloor \frac{P_B}{N_D + N_C + N_P - 1} \right\rfloor} N_s$$

$$G_W = \frac{\left\lfloor \frac{P_B}{N_s N_D + N_C + N_l - 1} \right\rfloor}{\left\lfloor \frac{P_B}{N_D + N_C + N_P - 1} \right\rfloor} N_s$$

3x3 Channel Architecture

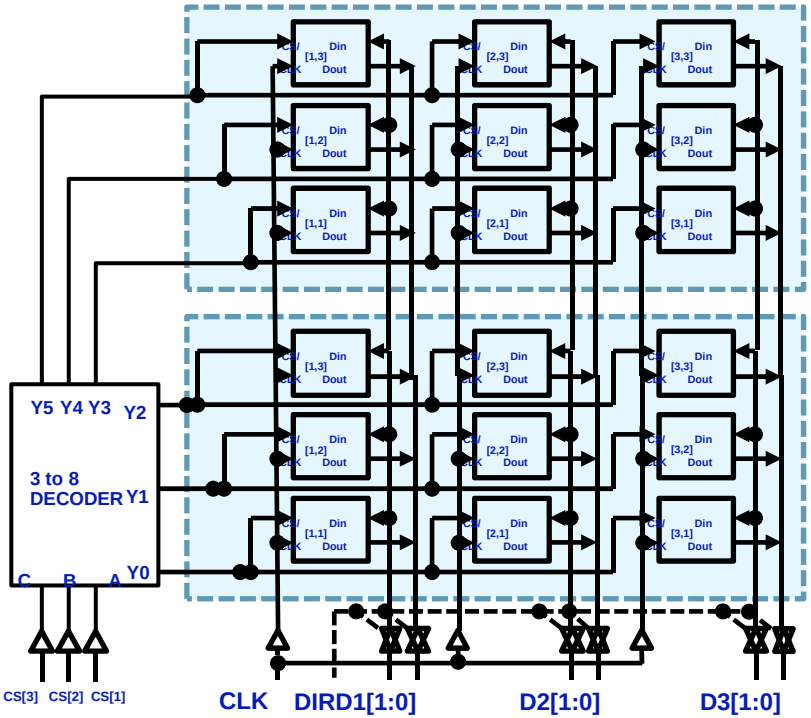
1 Block = 64 bytes

For each user sector (512b= 8x64), we store 64bytes of metadata, *i.e.*, 9 blocks in total



PSS Channel Card

1 PCM Channel = 2 Banks (2x3x3)



Channel card

One PCM channel per side

Two PCM channels per card

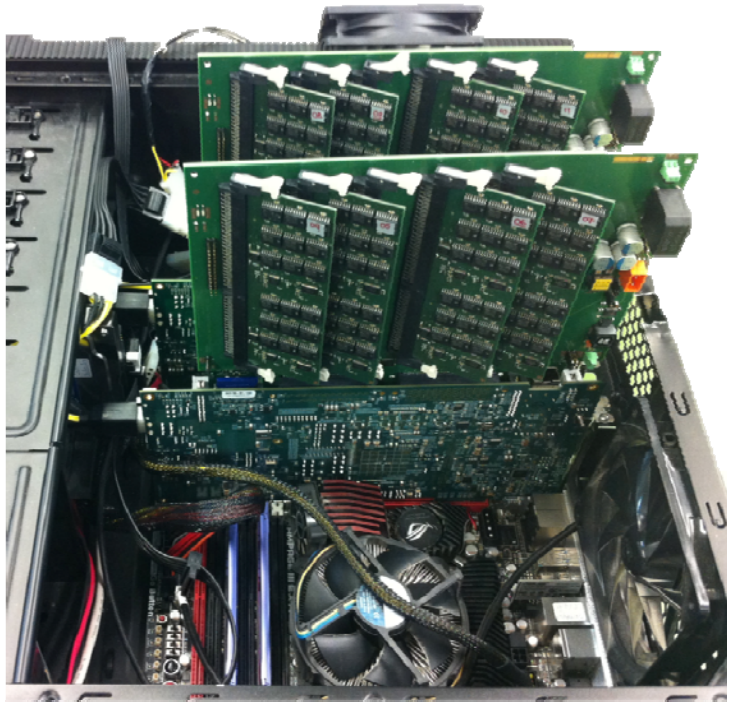
PCM channel specs

Data transfer Rate:	49.5 MBps
Sector read time:	13.8 usecs
Sector read rate:	61.6 ksectors/sec
Sector write time:	133.8 usecs
Sector write rate:	14.8 ksectors/sec

PSS PCI-e Card

Xilinx Zynq-7045 FPGA Board

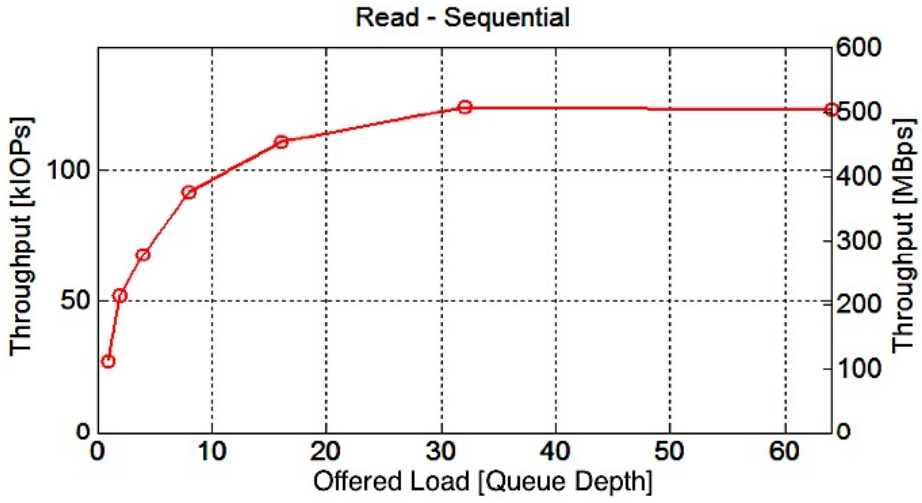
8 PCM channels with pipeline support per card

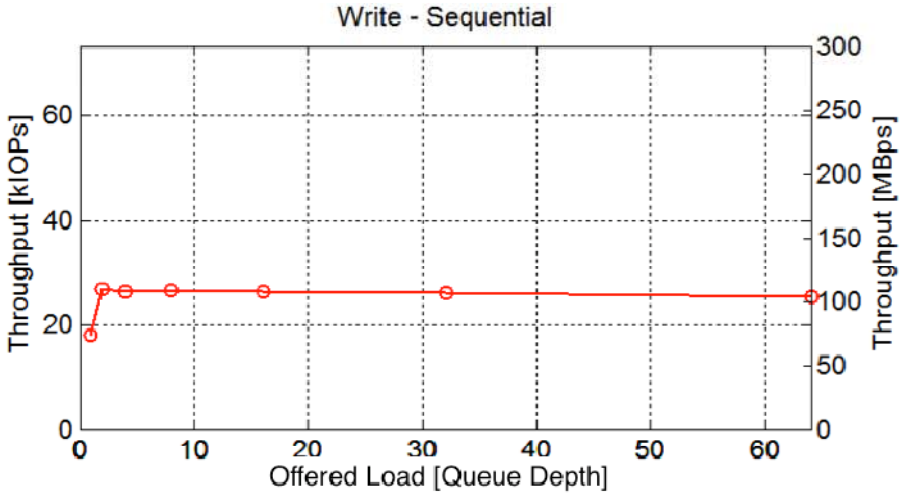


- Error correction based on simple BCH codes
  - 6 BCH codewords per 512 bytes sector with 4 bits error correction capability per codeword.
- Wear leveling using a Start-Gap scheme
- 512MB of DRAM, mostly used as a write cache
- Support for cached writes, direct writes with early completion, direct writes with late completion

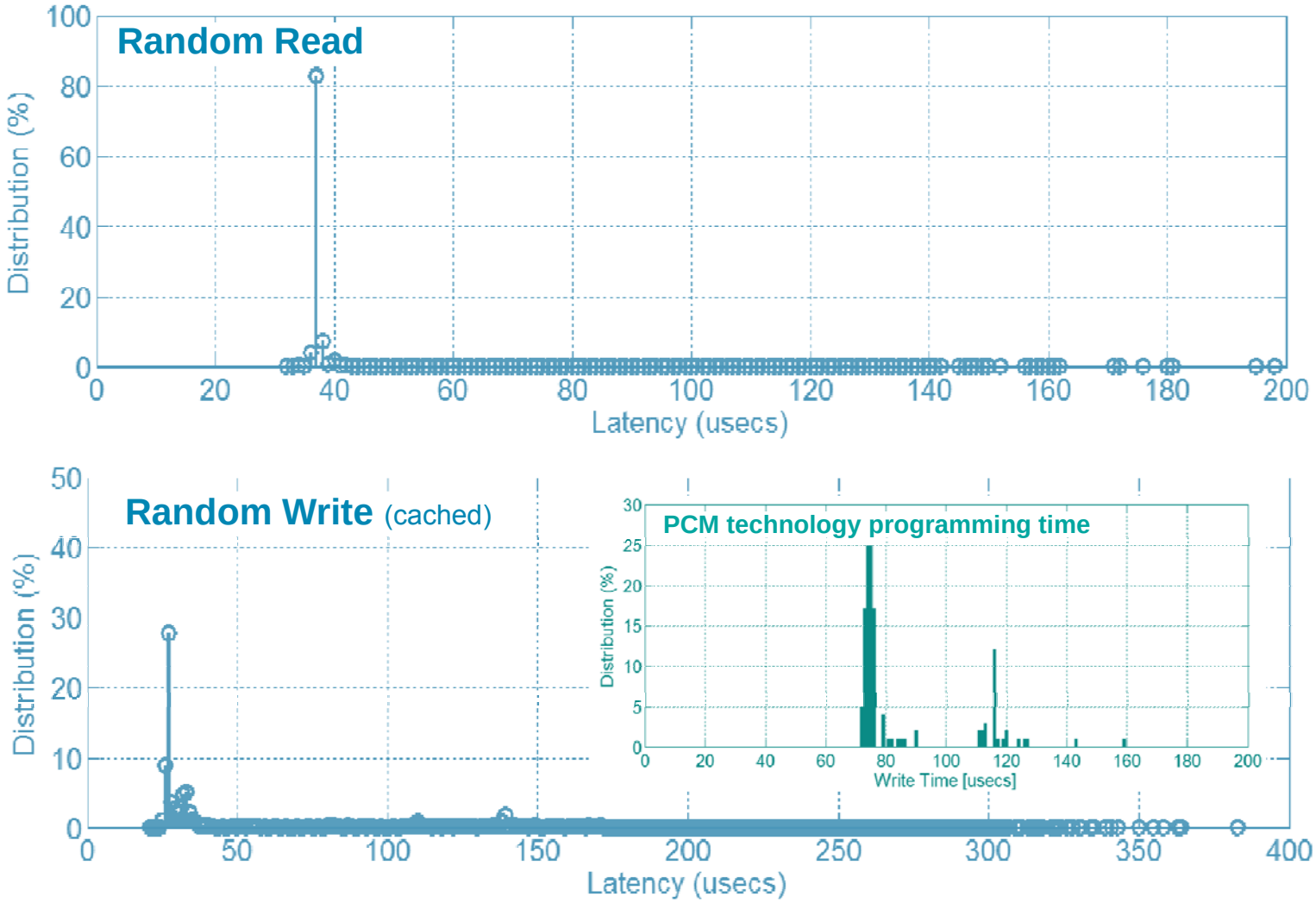
PSS Experimental Results

Throughput versus Offered Load (4kB pages)





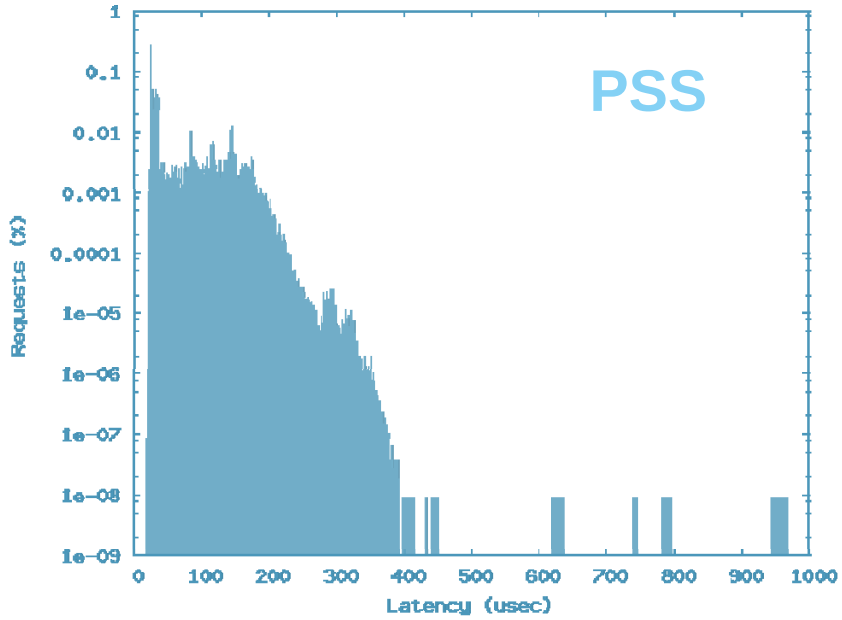
PSS Experimental Results
I/O Completion Latency Distribution

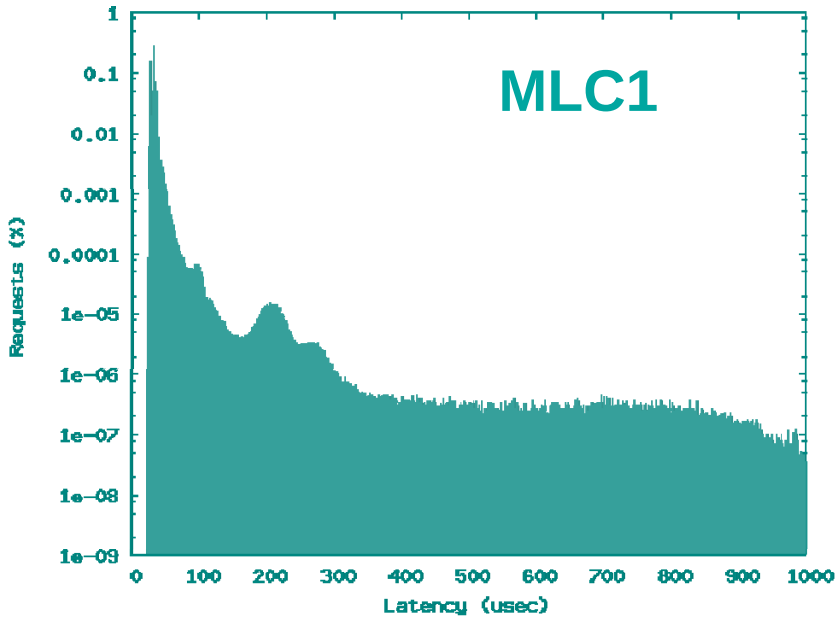


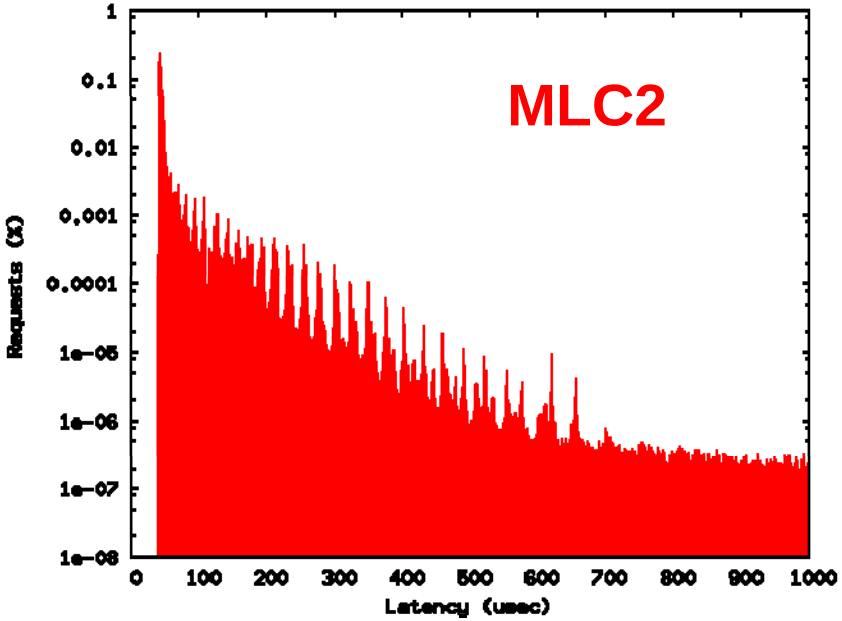
Latency distribution comparison to Flash-based devices

- Devices
  - PSS PCI-e Card
  - MLC Flash PCI-e SSD 1
  - MLC Flash PCI-e SSD 2
  - TLC Flash SATA SSD
- Experiment
  Per-I/O latency measurements for 2 hours of uniformly random 4kB writes at QD=1
  (after 12 hours of preconditioning with the same workload)

Latency Profile up to **1msec**

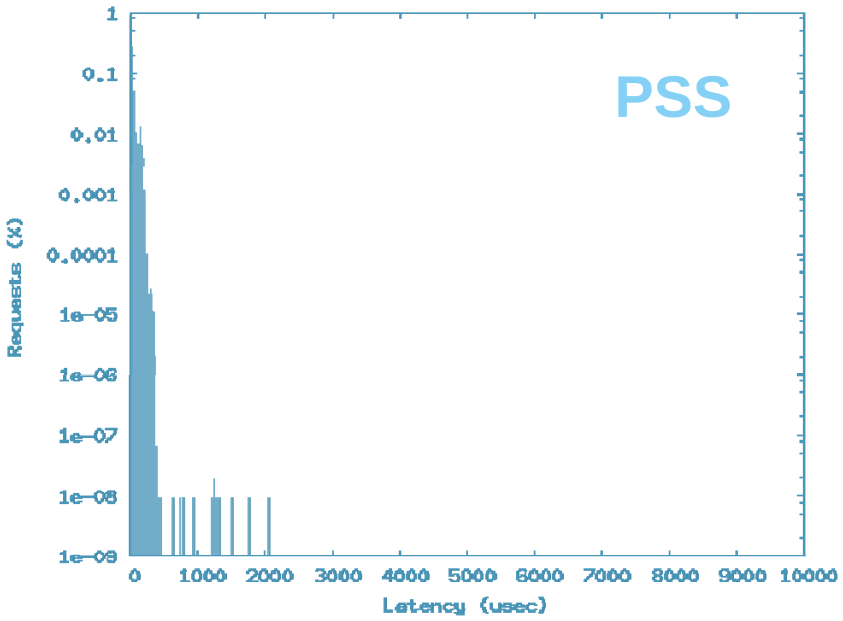


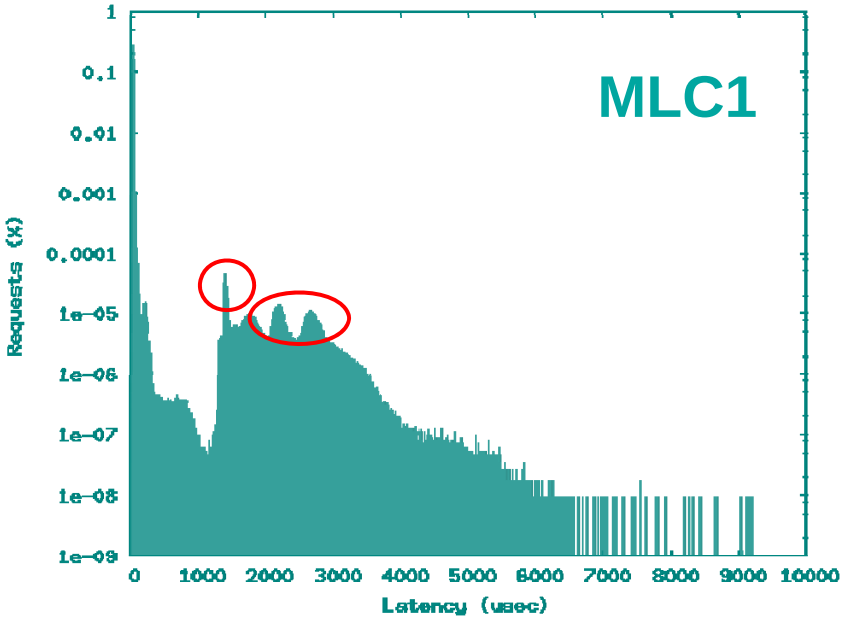


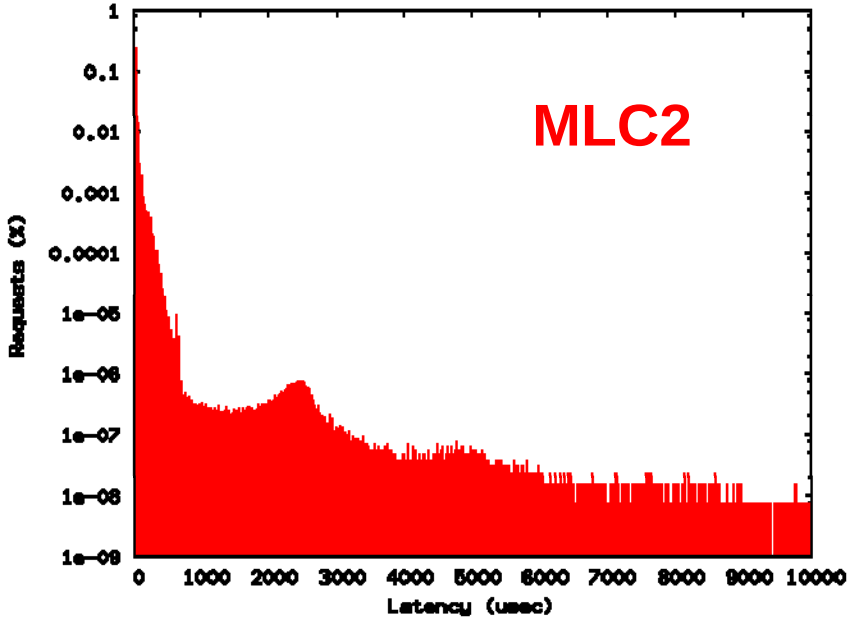


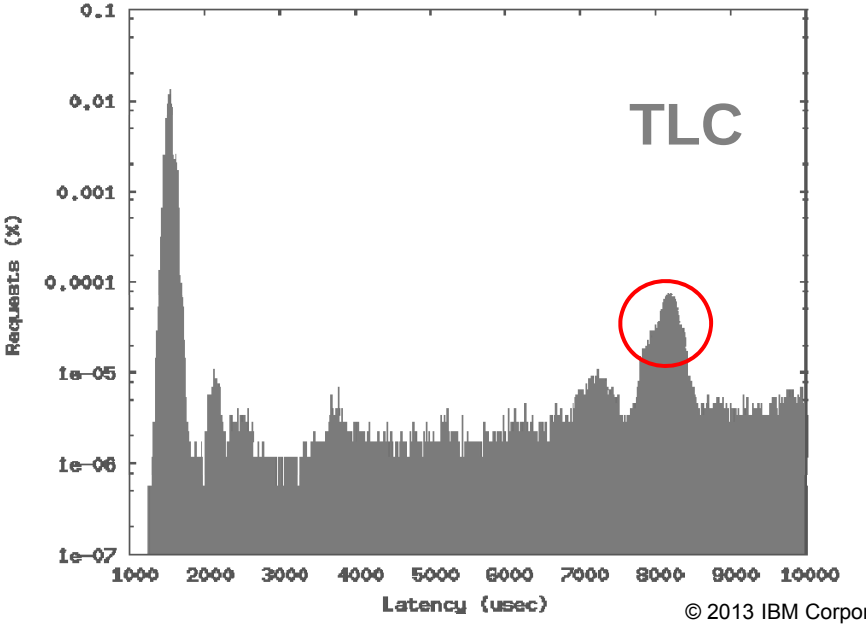
TLC

Latency Profile up to **10msec**

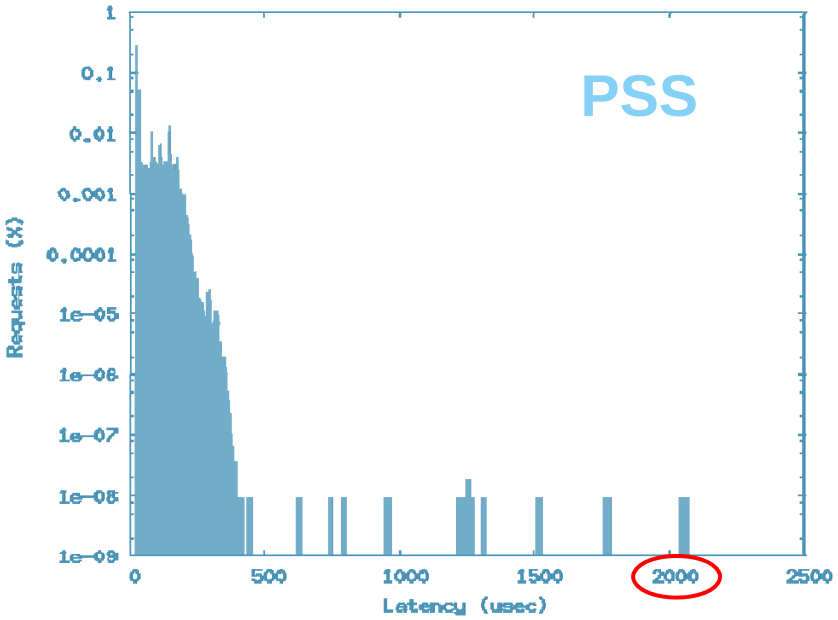


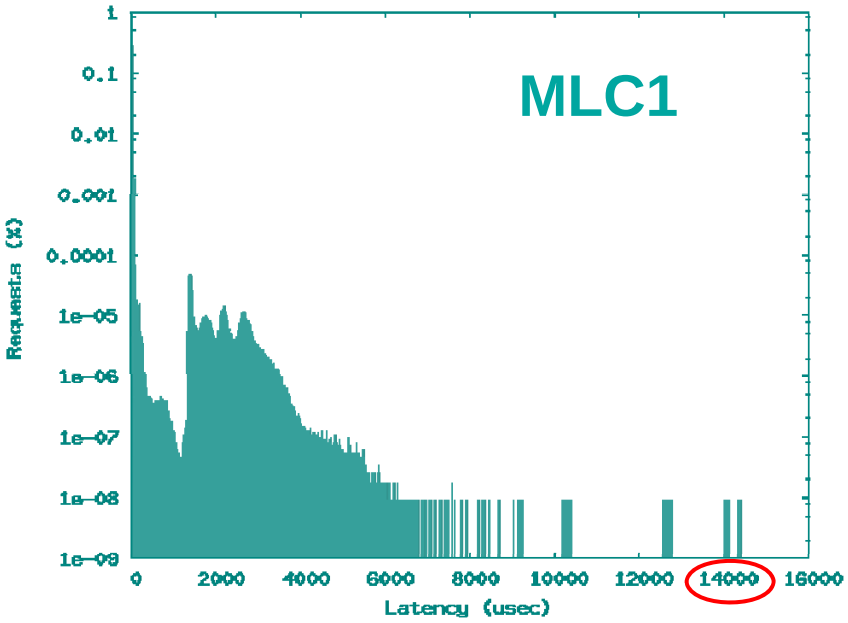


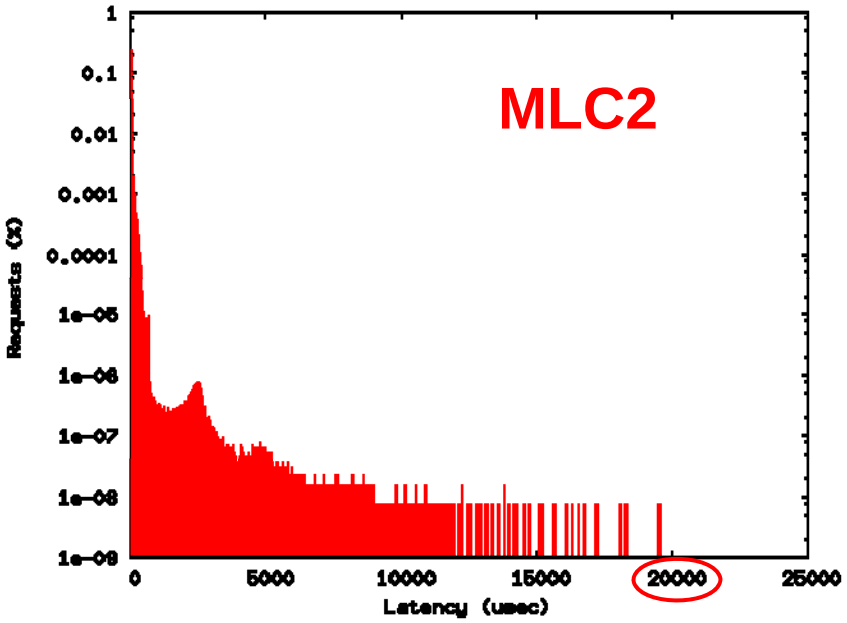


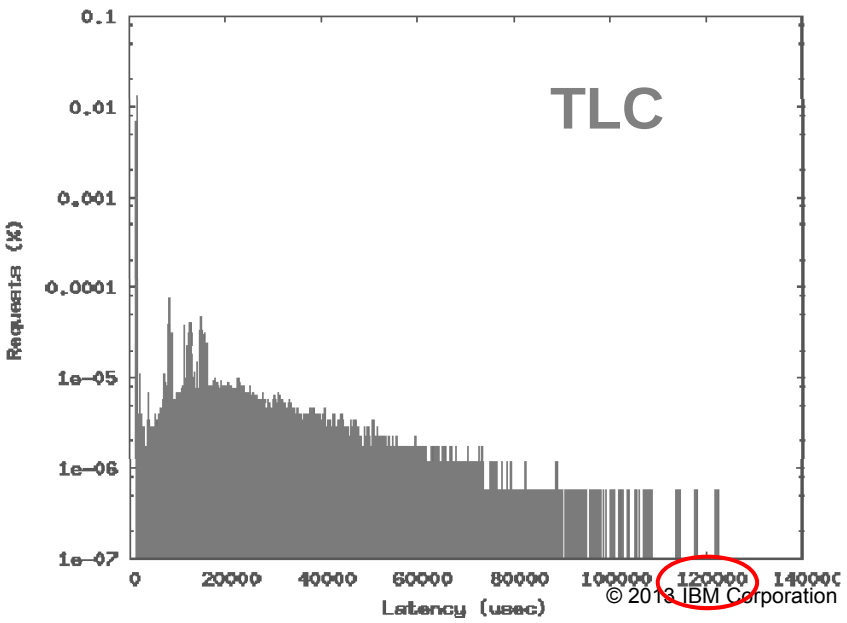


Total Latency Profile









PCM Endurance Measurements

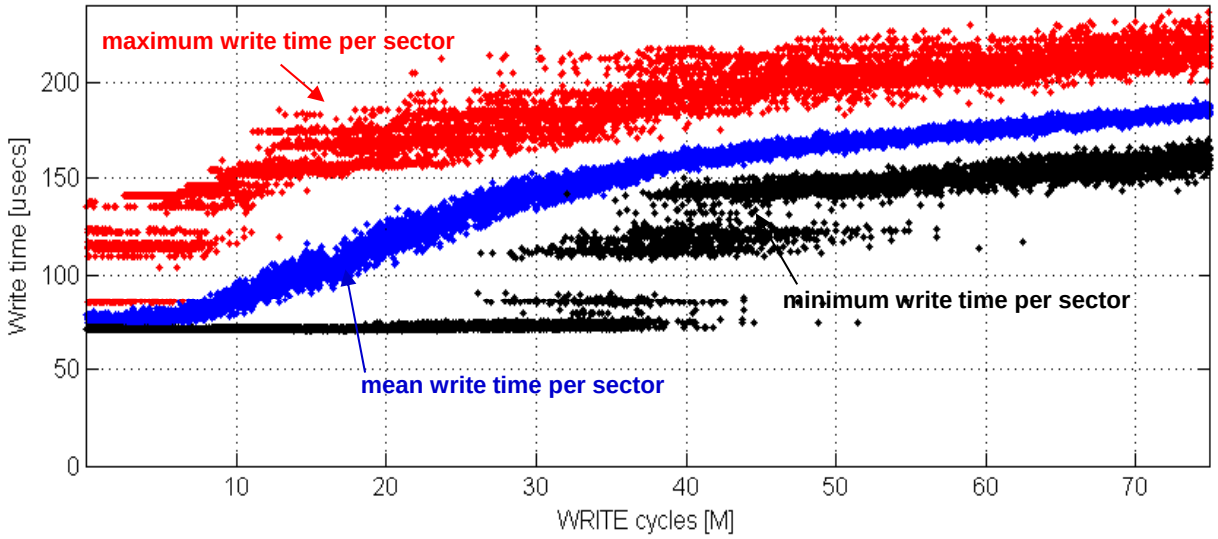
The effect of PCM aging on write time and BER

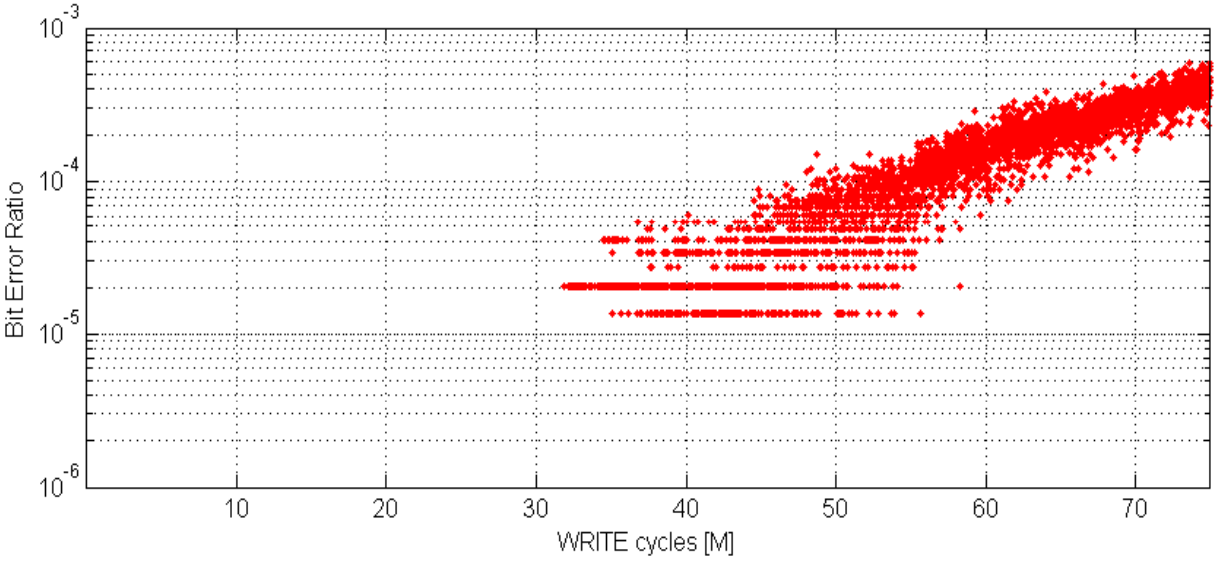
Experimental parameters:

- Random data
- 32 sectors per write cycle
- 4 PCM channels
- 8 PCM banks
- Pipeline is active

Experiment

- Perform 10K write cycles with random data (x32 sectors)
- Write, read and compare a set of 32 sectors (single write cycle)





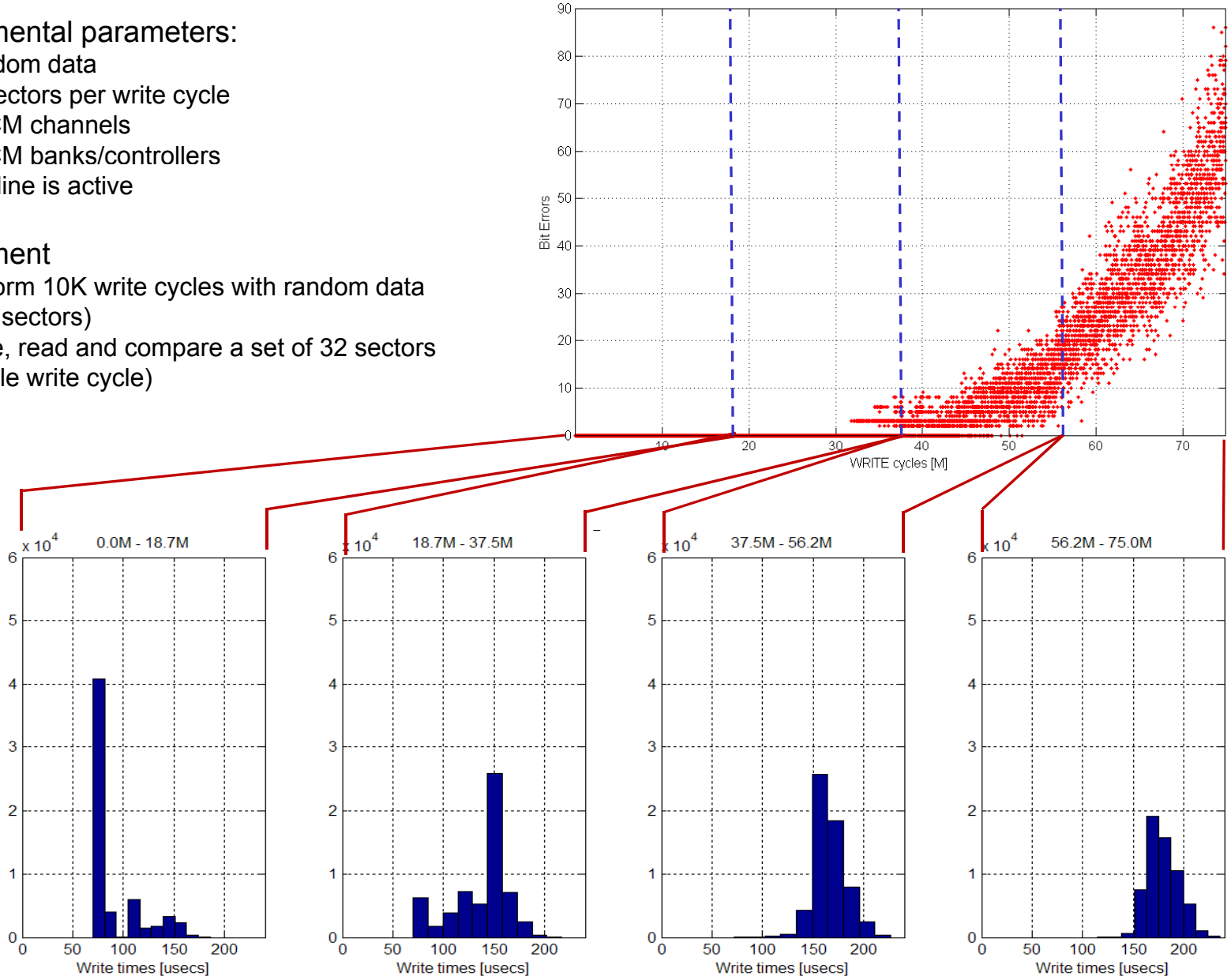
PCM Write Latency Distribution

Experimental parameters:

- Random data
- 32 sectors per write cycle
- 4 PCM channels
- 8 PCM banks/controllers
- Pipeline is active

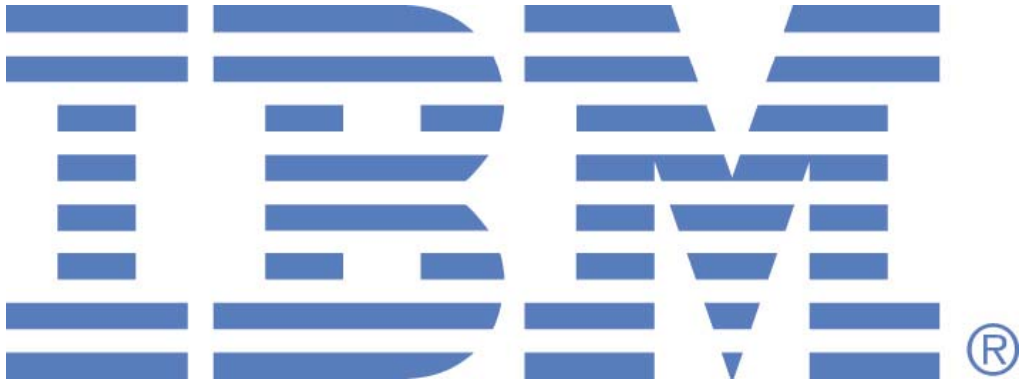
Experiment

- Perform 10K write cycles with random data (x32 sectors)
- Write, read and compare a set of 32 sectors (single write cycle)



Conclusions

- PCM is a promising new memory technology
- PSS is a PCI-e attached subsystem that mitigates the limitations of current PCM technology
- The 2D Channel Architecture allows the designer to trade-off read performance for write performance and vice-versa
- PSS achieved good performace
  - 65k Read IOPS @ 35 µsec
  - 15k Write IOPS @ 61 µsec
- PSS achieved consistently low write latency
  - 99.9% of the requests completed within 240 µsec
    - 12x and 275x lower than MLC and TLC Flash SSDs, respectively
  - Highest observed latency was 2 msec
    - 7x and 61x lower than MLC and TLC Flash SSDs, respectively



UNIVERSITY OF
PATRAS