Hyojun Kim

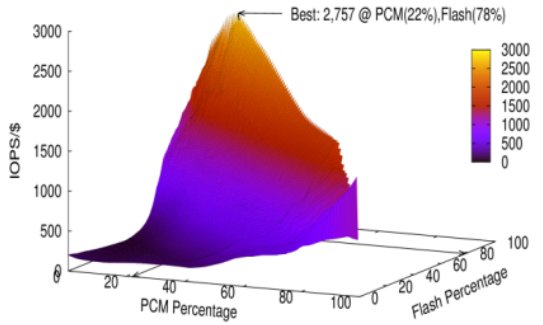
Evaluating Phase Change Memory for Enterprise Storage Systems

IBM Almaden Research

Micron provided a prototype SSD built with 45 nm 1 Gbit Phase Change Memory

Measurement study

Performance Characteristics



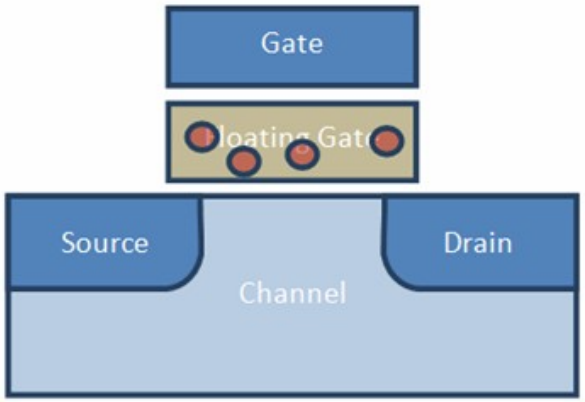
Storage Simulation

Part I

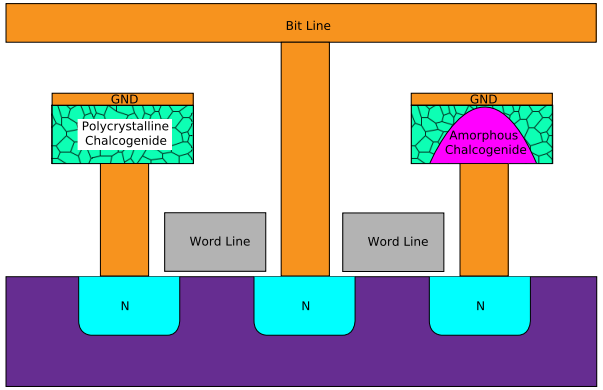
Part II

Part I. PCM performance

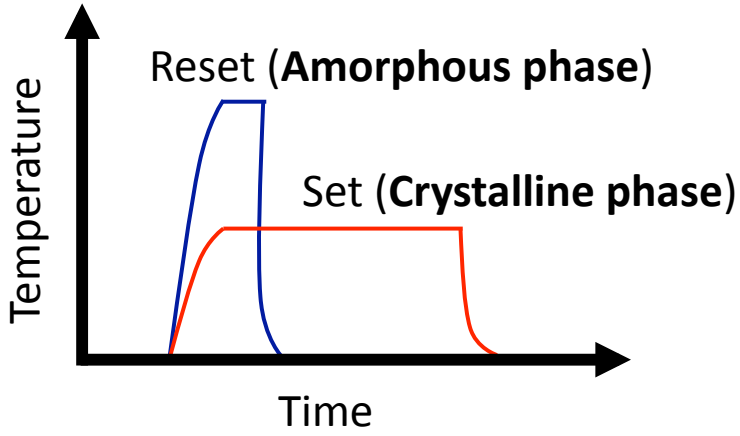
Flash Memory vs. Phase Change Memory



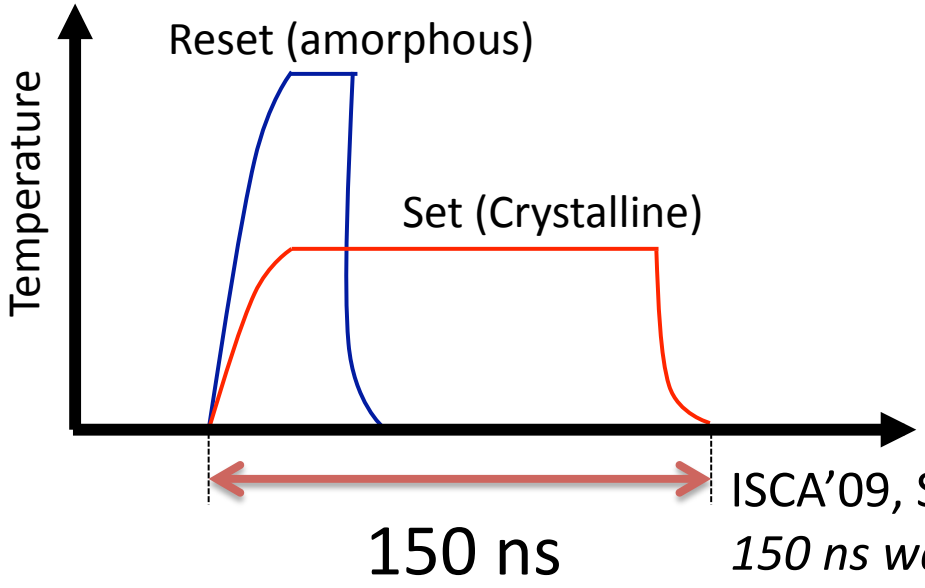
Flash memory traps charges to remember bits



PCM melts and cools material to remember bits



PCM write latency

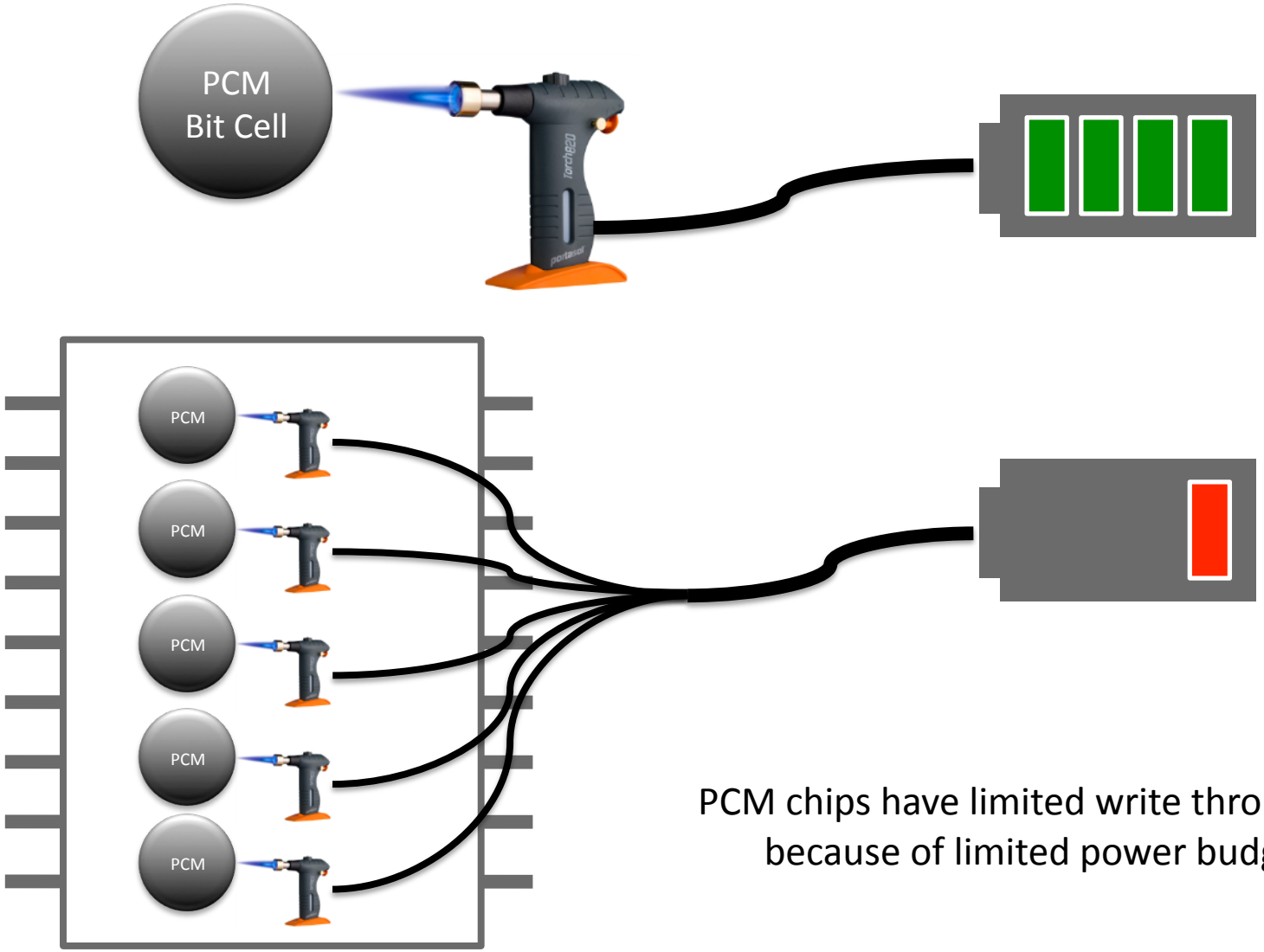


ISCA'09, SOSP'09, FAST'11:
150 ns was used as PCM write latency

FAST'11 2015 Projections:

Technology	Read/Write Latency ns	
Flash SSD (SLC)	25,000	200,000
DRAM (DIMM)	55	55
PCM	48	150

PCM write throughput and power consumption



PCM chips have limited write throughput because of limited power budget

Example: VLSI Circuits, 2004, F. Bedeschi

26.1

An 8Mb Demonstrator for High-Density 1.8V Phase-Change Memories

F. Bedeschi, C. Resta, O. Khouri, E. Buda, L. Costa, M. Ferraro, F. Pellizzer, F. Ottogalli, A. Pirovano, M. Tosi, R. Bez, R. Gastaldi, and G. Casagrande

STMicroelectronics, MPG and Central R&D
Via C. Olivetti, 2 - 20041 - Agrate Brianza (Milano) - Italia
Tel.: +39-039-6036123, E-mail: ferdinando.bedeschi@st.com

Abstract

An 8Mb Non-Volatile Memory Demonstrator incorporating a novel 0.32 μm^2 Phase-Change Memory (PCM) cell using a Bipolar Junction Transistor (BJT) as selector and integrated into a 3V 0.18 μm CMOS technology is presented. Realistically large 4Mb tiles with a voltage regulation scheme that allows fast bitline precharge and sense are proposed. An innovative approach that minimizes the array leakage has been used to verify the feasibility of

(GST) [5]. The cell formation modules are placed between the CMOS front-end (FEOL) and back-end (BEOL) and all the additional process steps required for the cell formation are compliant with a standard 0.18 μm CMOS technology. The basic CMOS parameters are reported in Tab.1. In order to integrate the PCM cell with the BJT selector, the chalcogenide alloy GST is patterned together with an AlCu Metal0 to form the subbitlines. The heater element is connected to the BJT *p*-Emitter, while the *n*-Base is contacted to the Metal1 wordline [5].

along the sub-bitlines). A RESET pulse of 40 ns and a SET of 150 ns have been demonstrated. In order to minimize the

of 150 ns have been demonstrated. In order to minimize the parasitic voltage drop along the wordline, only a pair of bits of a byte is programmed at the same time in the RESET state (1 bit per half-tile), while 4 bits are SET at the same time (2 bits per half-tile), considering the lower value of the SET current. On the 8Mb Demonstrator, where the 8 bits are

arranged inside the same tile, the Write throughput demonstrated is 2.5MB/s, limited by the 200 ns SET time (SET pulse + 50 ns of circuitry delay).

Write latency and throughput comparison

- Write latency
  - PCM SET operation time: 200 ns
  - Flash page program time: 200 µs
- Write throughput per chip
  - PCM: 4 bits / 200 ns = 2.5 MB/s
  - Flash: 4 KB / 200 us = 20.5 MB/s

PCM technology survey paper (2013)

Table I. Quantitative Parameters of PCM

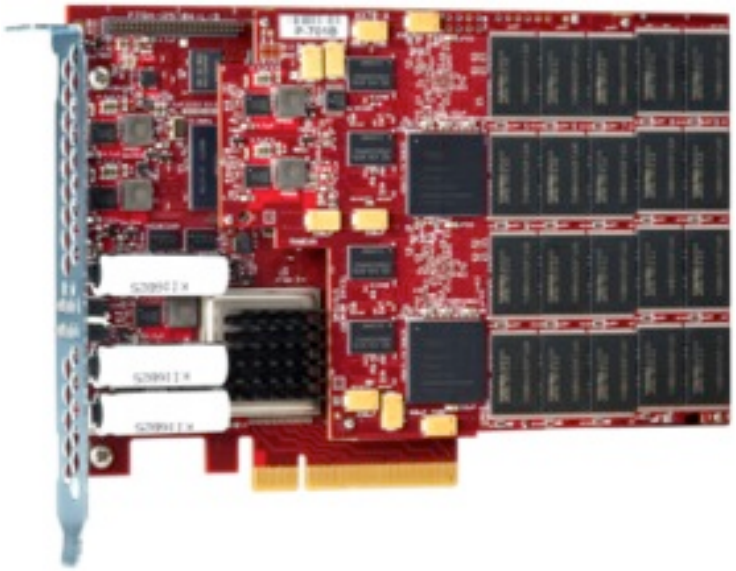
Parameter	DRAM	NAND Flash [Boboila and Desnoyers 2010] [Javanifard et al. 2008] [MicronFlash 2008] [Nobunaga et al. 2008]	PCM [Atwood 2010] [Pirovano et al. 2004b]
Scalability	3X nm [Samsung 2011]	2X nm	<1X nm
Read Latency	60ns	25–200us	50–100ns
Write Speed	~1Gb/s	2.5 MB/s	~100MB/s
Endurance	N/A	10^3 to 10^5	10^6 to 10^8 [Atwood 2010], 10^{11} [Pirovano et al. 2004b]

Clarifications about PCM performance

- Write latency ≠ Write throughput
- 150 ns SET time = Material performance
- Material performance ≠ Chip performance
- Chip performance ≠ SSD performance

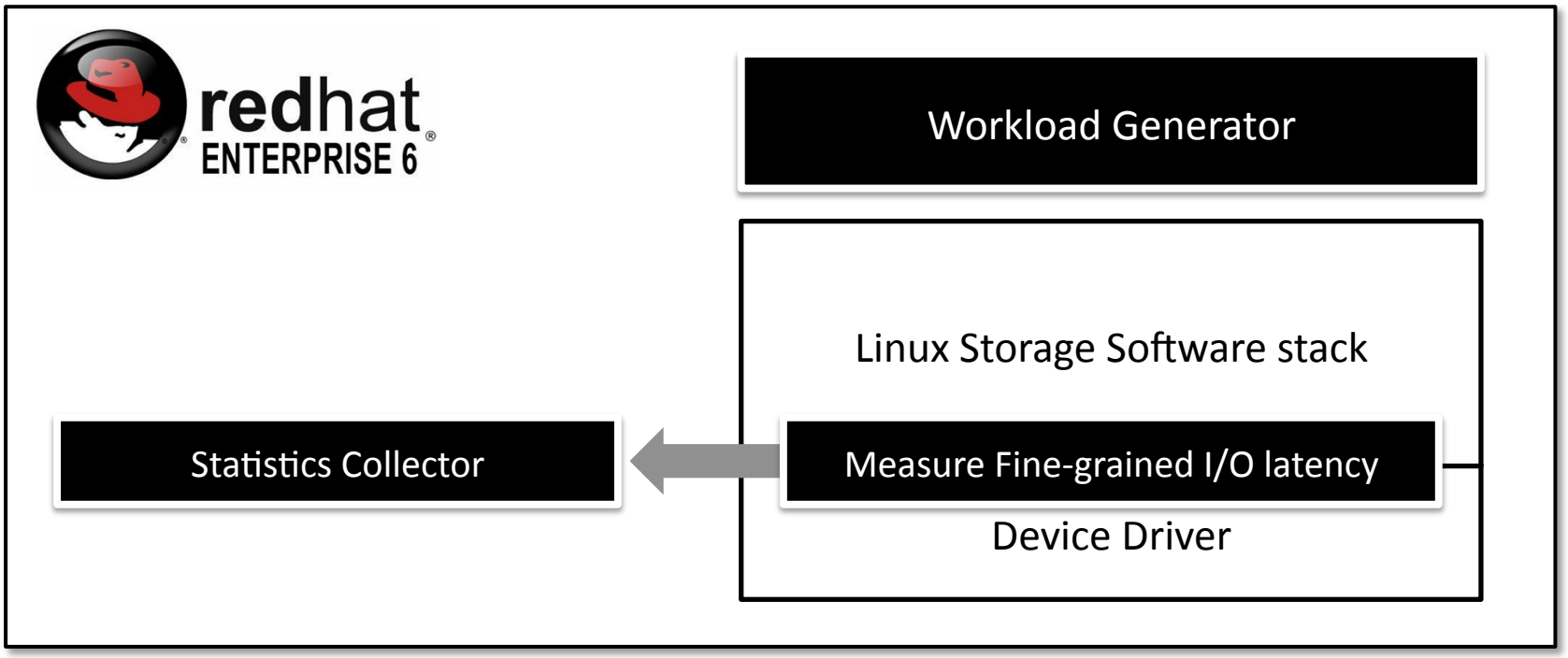
All PCM SSD vs. eMLC SSD

- 45 nm PCM
- 64 GB capacity

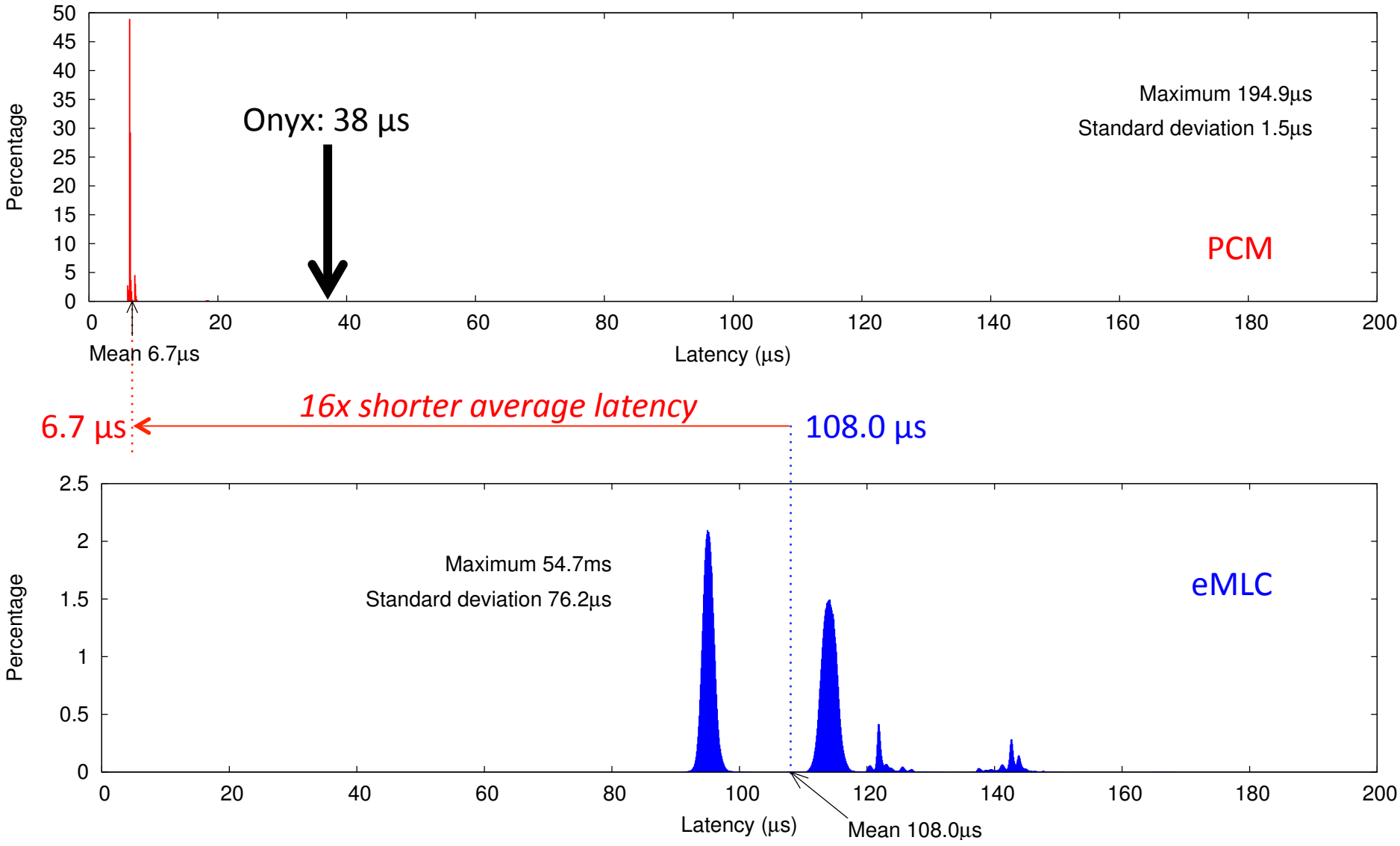


- eMLC NAND flash
- 1.8 TB capacity

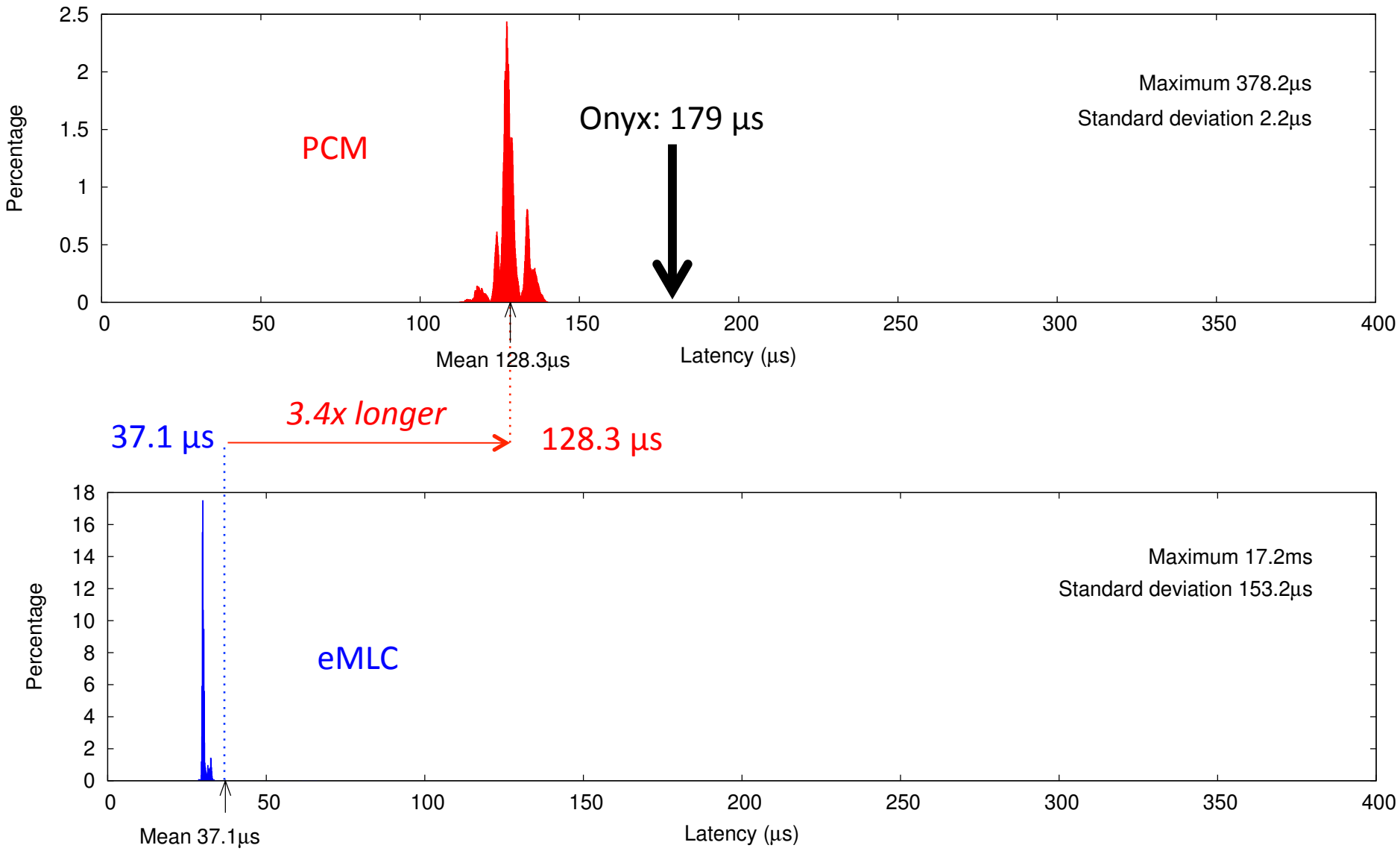
Fine-grained I/O latency measurement



4KB Random Read Latency (5M samples)



4KB Random Write Latency (1M samples)



Summary of Performance Numbers

	PCM SSD	eMLC SSD
4KB Read Latency	6.7 µs	108.0 µs
4KB Write Latency	128.3 µs	37.1 µs

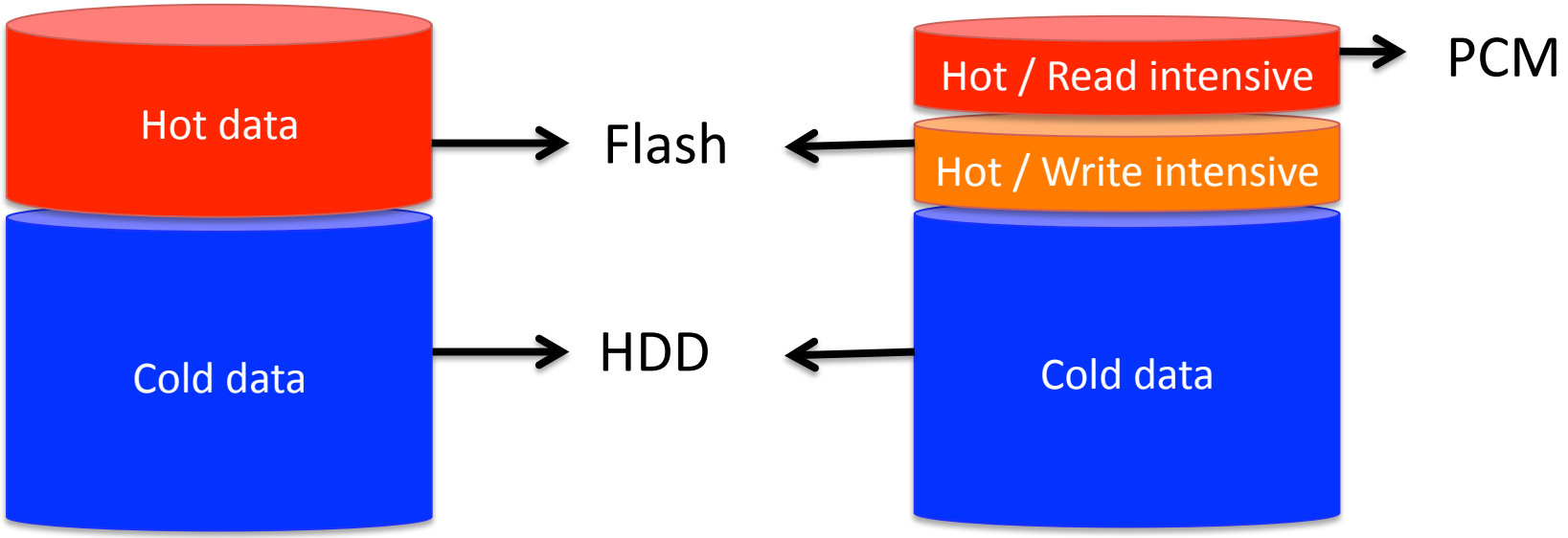
- For read, PCM SSD is about 16x faster
- For write, PCM SSD is about 3.4x slower than eMLC SSD using DRAM as write buffer

Part II. Storage simulation

"Will PCM be useful for enterprise storage systems?"

PCM / Flash / HDD multi-tiered storage

	PCM SSD	eMLC SSD	15K RPM HDD
4KB Read Latency	6.7 µs	108.0 µs	5 ms
4KB Write Latency	128.3 µs	37.1 µs	5 ms
Normalized Cost	24	6	1



Multi-tiered storage simulation

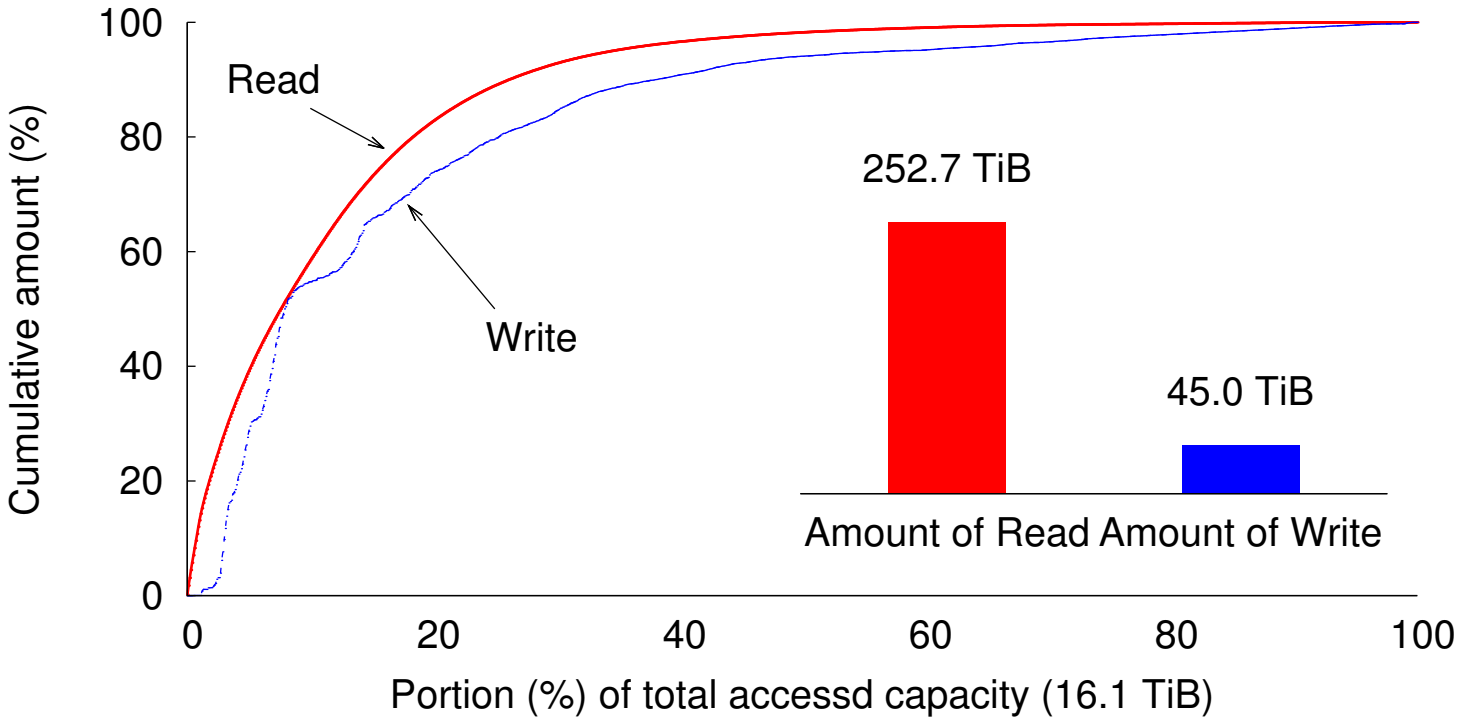
1. Assume multi-tiered storage made of **X** % of PCM, **Y** % of Flash, **Z** % of HDD
2. Estimate Performance (IOPS)
3. Estimate Storage Cost (Normalized: HDD = 1)
4. Evaluation metric:

$$\frac{\text{Estimated Performance (IOPS)}}{\text{Estimated Cost (\$)}}$$

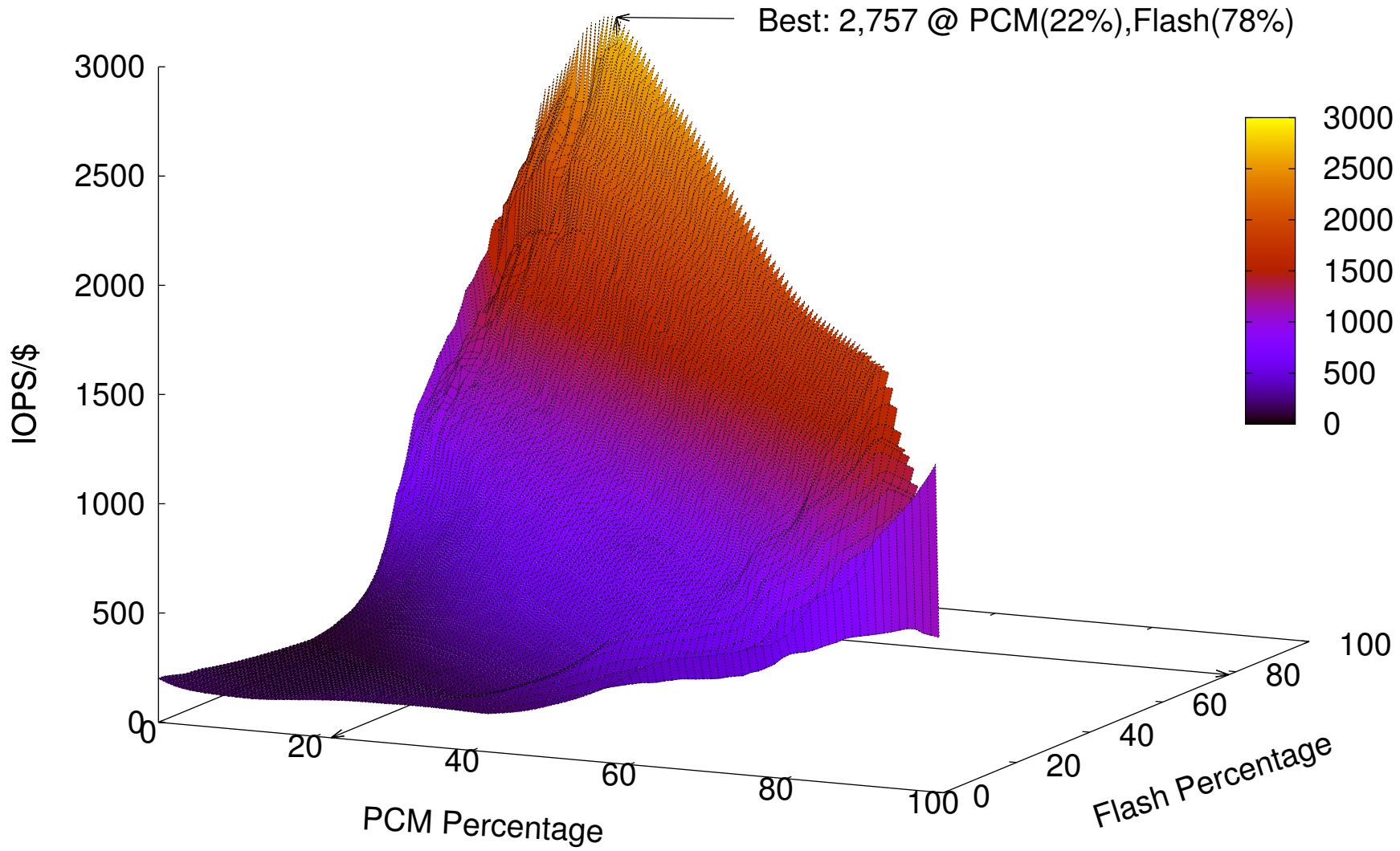
Tiering simulation methods

- Static-optimal data placement
  - Complete knowledge about I/O workload
  - No data movement
- Dynamic tiering
  - Reactive data movement based on I/O traffic

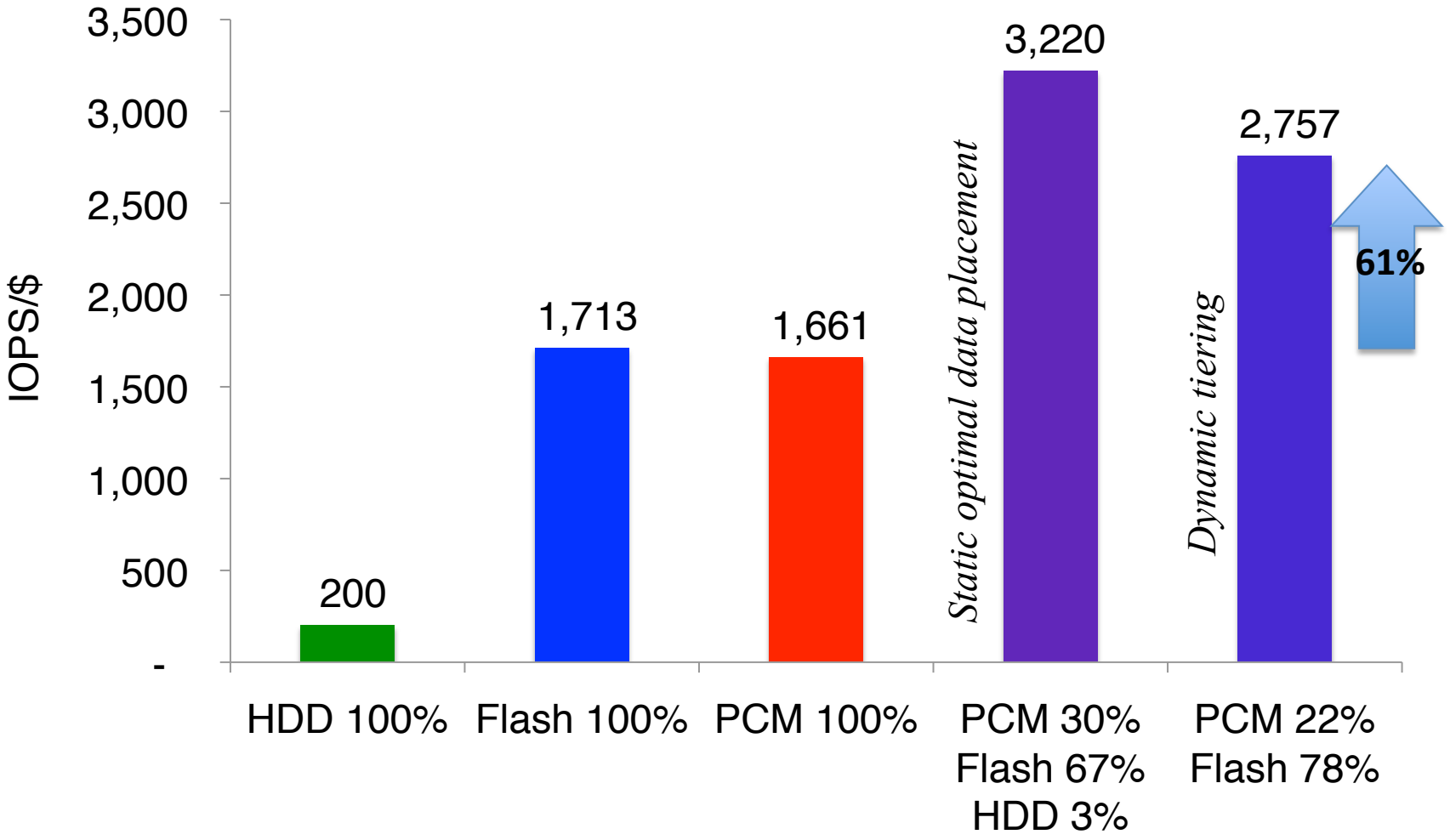
Retail Store, 2012 June, one week duration



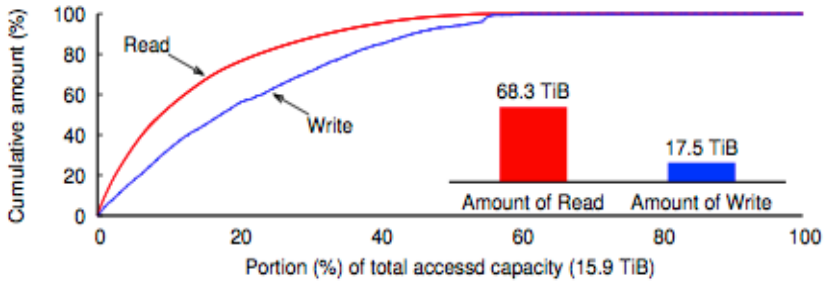
Retail Store: IOPS/$ with Dynamic tiering



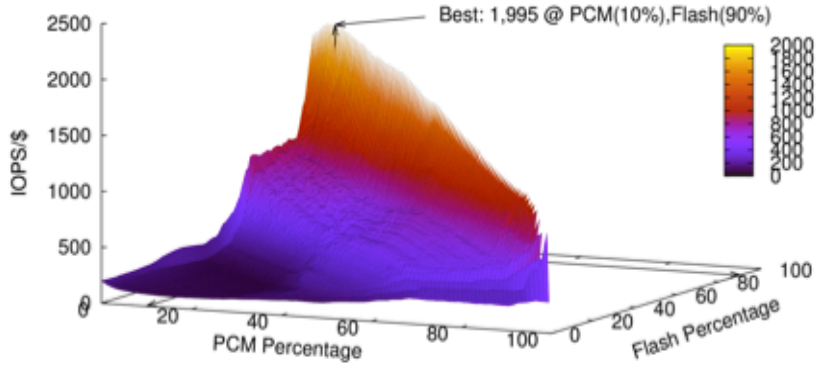
Retail Store



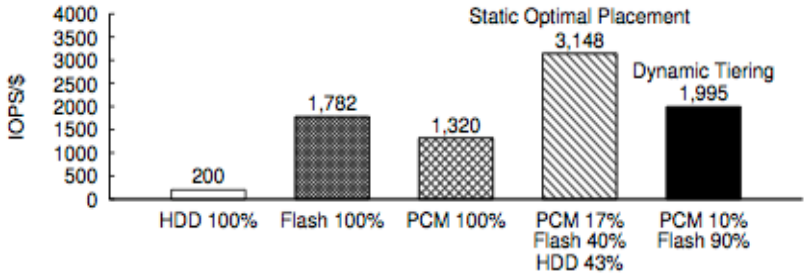
Bank



(a) CDF and I/O amount

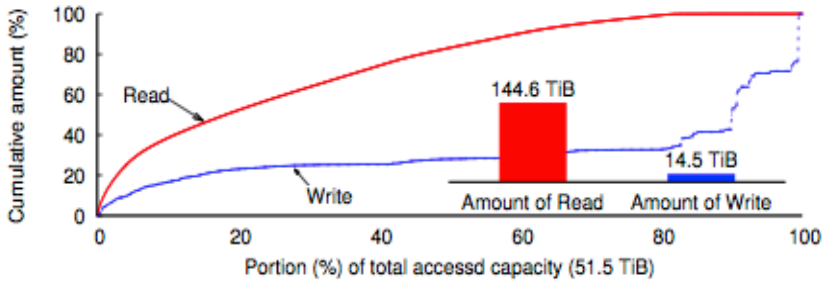


(b) 3D IOPS/$ by dynamic tiering

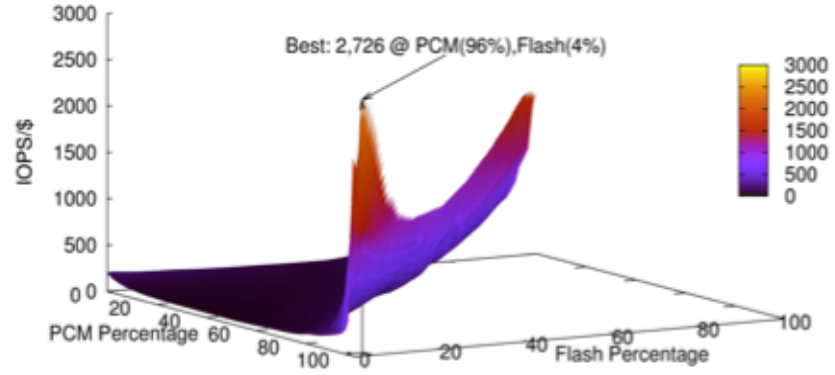


(c) IOPS/$ for key configuration points

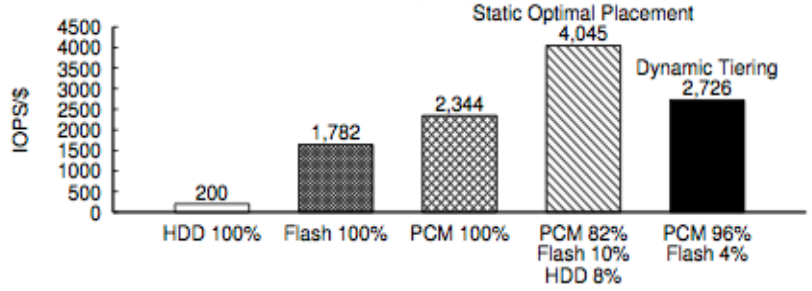
Telecommunication



(a) CDF and I/O amount



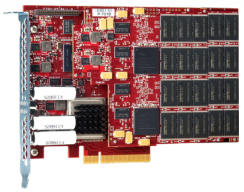
(b) 3D IOPS/$ by dynamic tiering



(c) IOPS/$ for key configuration points

Application Server

IBM Easy Tier Server
EMC XtremCache
NetApp FlashAccel
FusionIO IO Turbine



Flash Cache

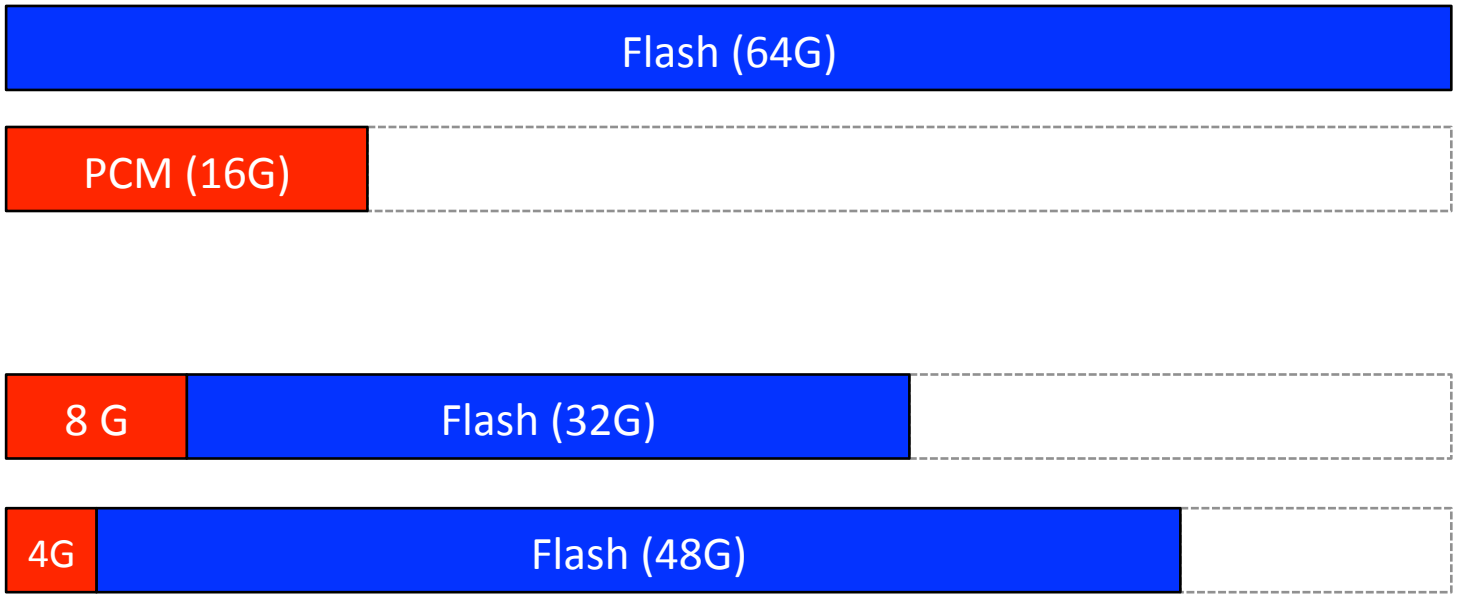
Shared Storage

Application server side cache simulation

- IO by IO storage traces
  - From real customer's production systems (over 24 hour duration)
  - Manufacturing, media, medical companies
- Cache Simulation methods
  - Write-through, LRU replacement
  - Evaluation metric: average read latency
- Performance parameters

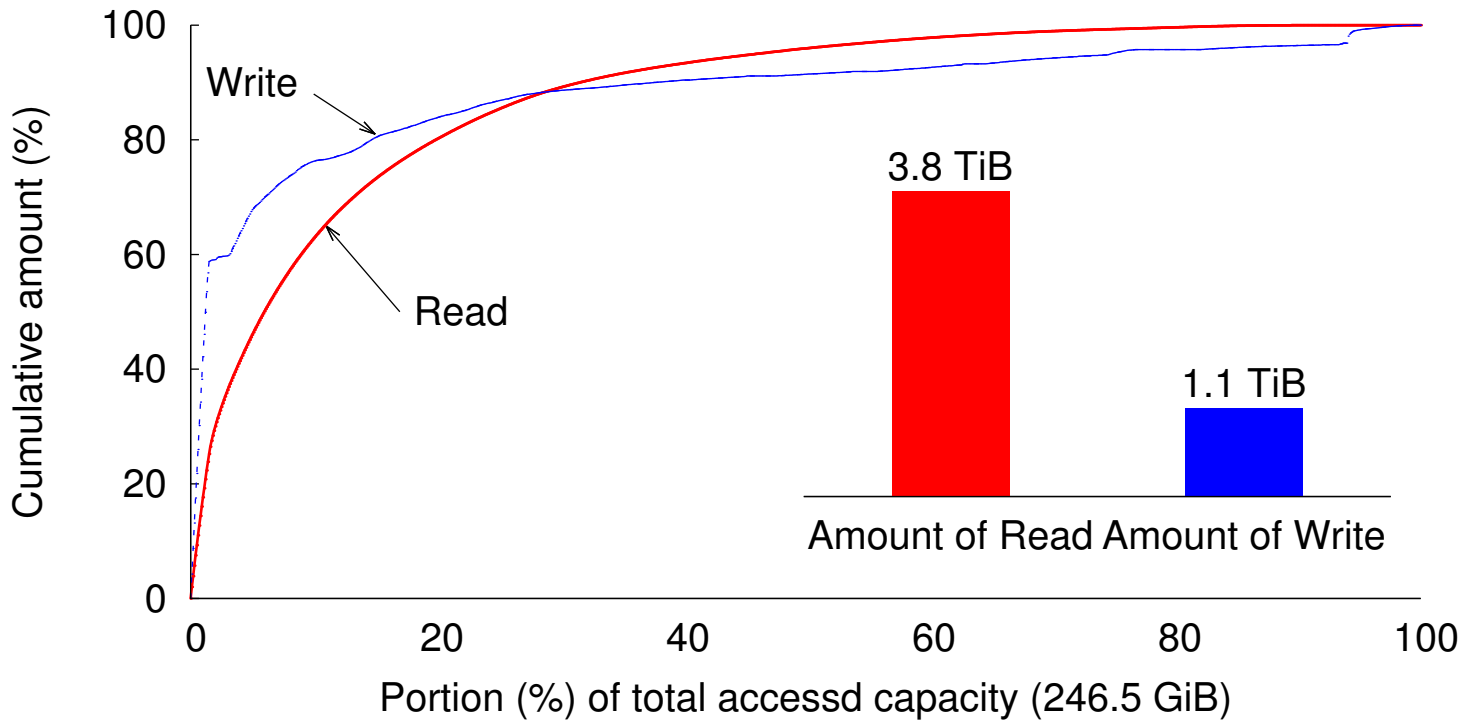
	PCM	eMLC	Net. Storage
4 KiB R. Lat.	6.7 µs	108.0 µs	919.0 µs
4 KiB W. Lat.	128.3 µs	37.1 µs	133.0 µs
Norm. Cost	4	1	–

Same cost cache configurations

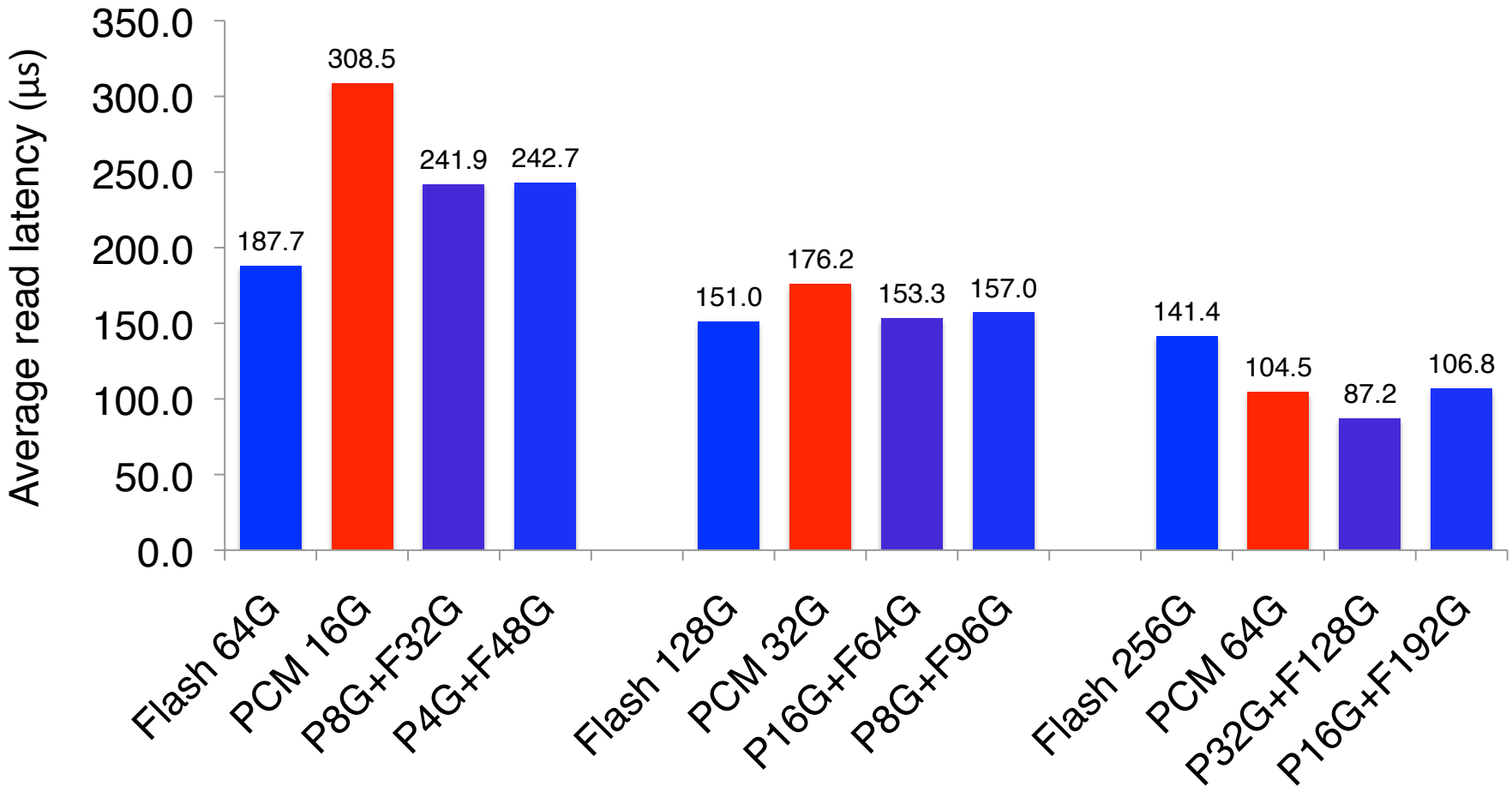


Flash 64G	PCM 16G	P 8G + F 32G	P 4G + F 48G
Flash 128G	PCM 32G	P 16G + F 64G	P 8G + F 96G
Flash 256G	PCM 64G	P 32G + F 128G	P 16G + F 192G

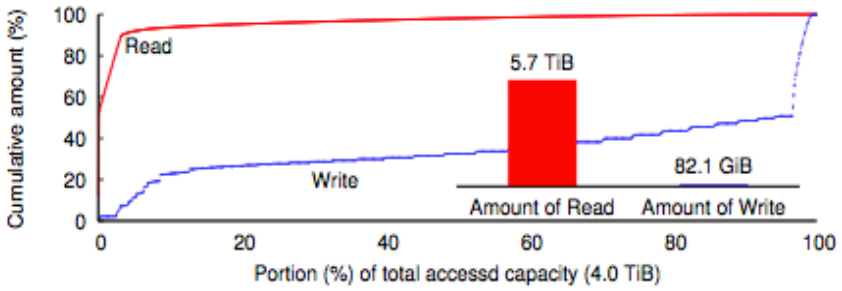
Manufacturing: I/O distribution (24 hours)



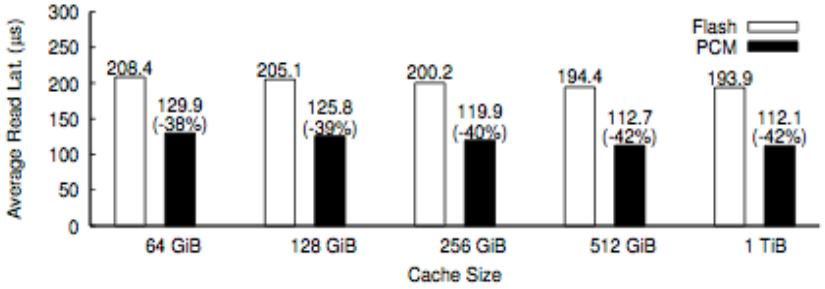
Manufacturing: cache simulation results



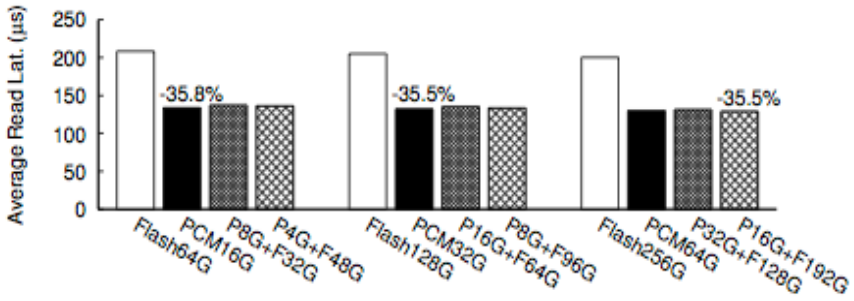
Media



(a) CDF and I/O amount

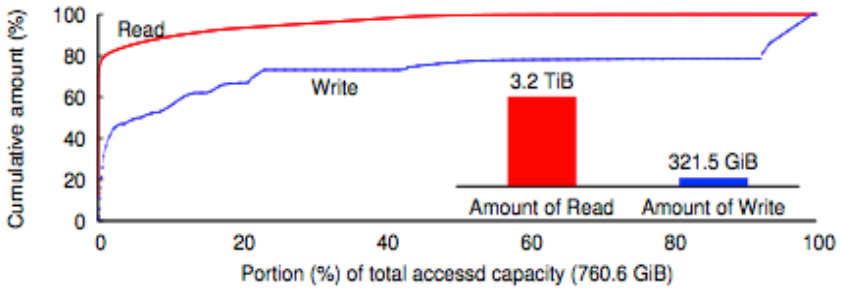


(b) Average read latency

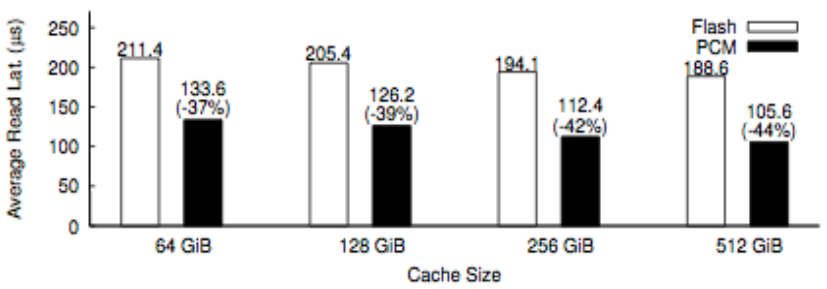


(c) Average read latency for even cost configurations

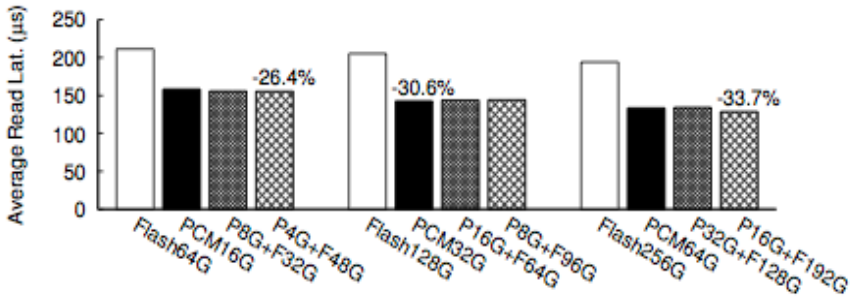
Medical



(a) CDF and I/O amount



(b) Average read latency



(c) Average read latency for even cost configurations

Summary and Conclusion

Summary

- Let's be careful to pick "right performance number" for PCM
- Performance measurement results

	PCM SSD	eMLC SSD
4KB Read Latency	6.7 µs	108.0 µs
4KB Write Latency	128.3 µs	37.1 µs

- Storage simulation results show
  - About 12-66% improved IOPS/$ for tiered storage
  - Up to 35% reduced average read latency for server caching

Concluding question

Phase Change Memory for enterprise storage…

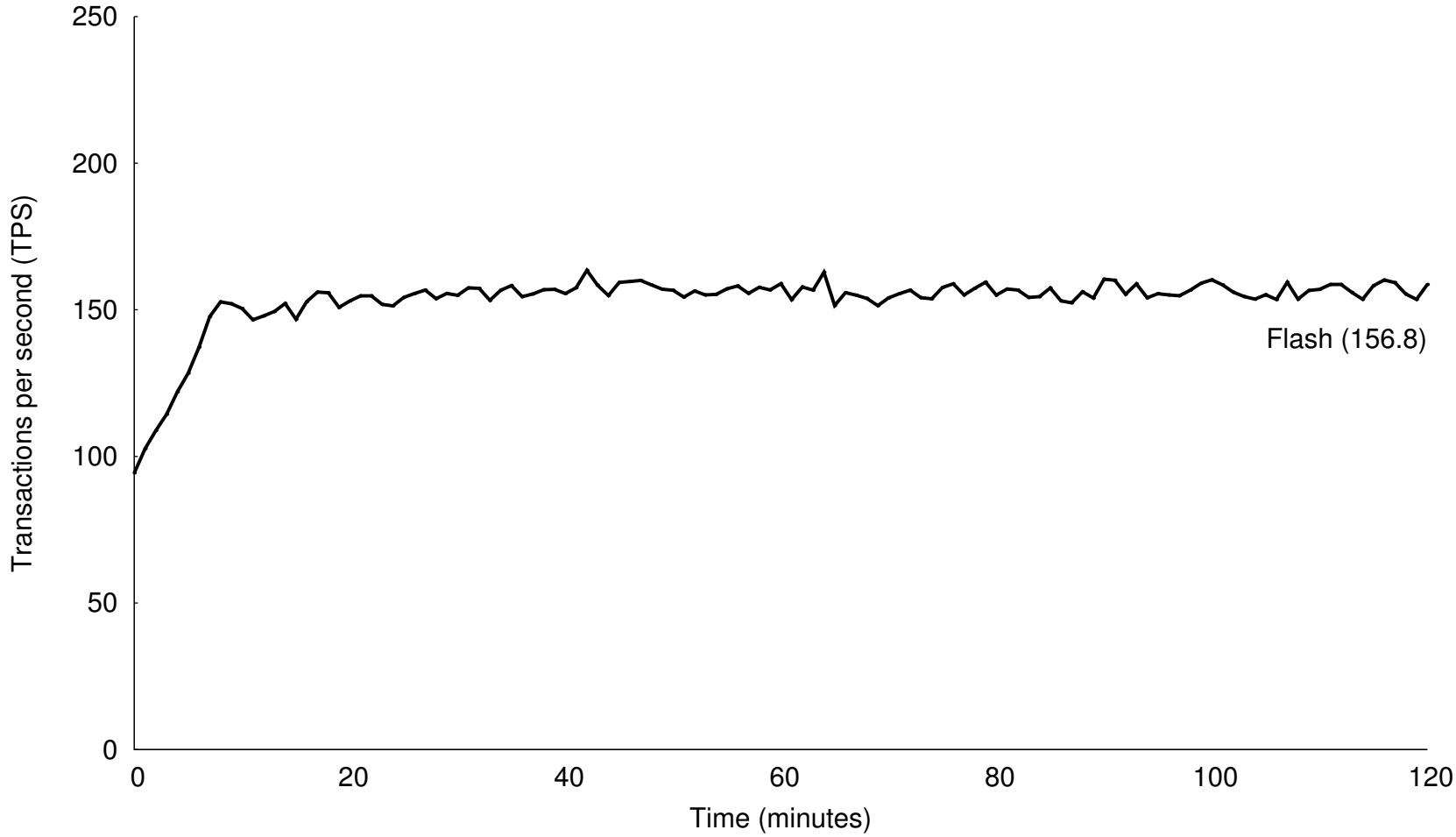
Silver bullet or Snake oil?

What do you think?

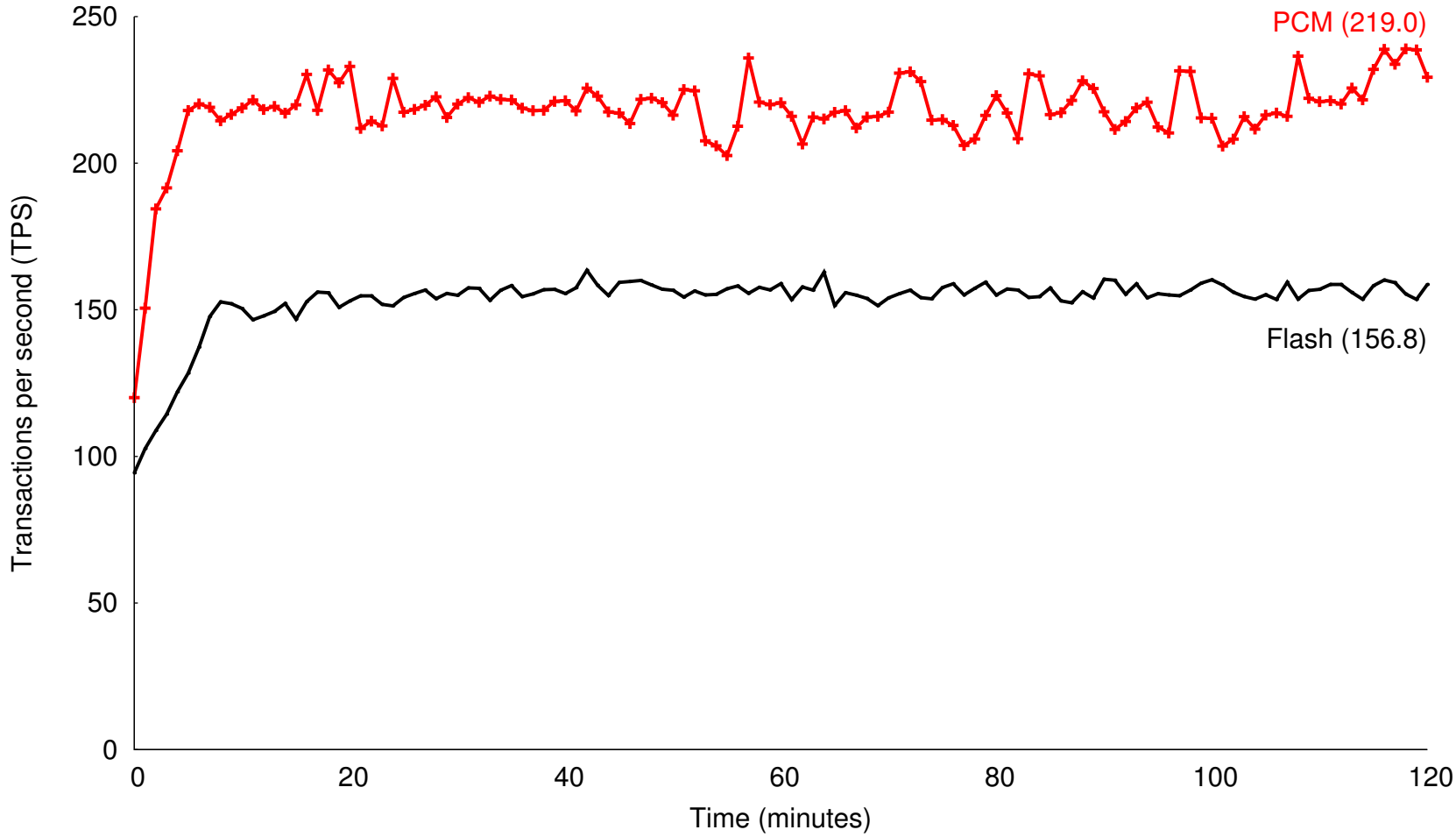
Additional experiments

- The PCM SSD was shipped to Almaden, finally
- Real experiments with Sysbench OLTP benchmark
  - Read-only workload
  - Single thread test / 8 thread test

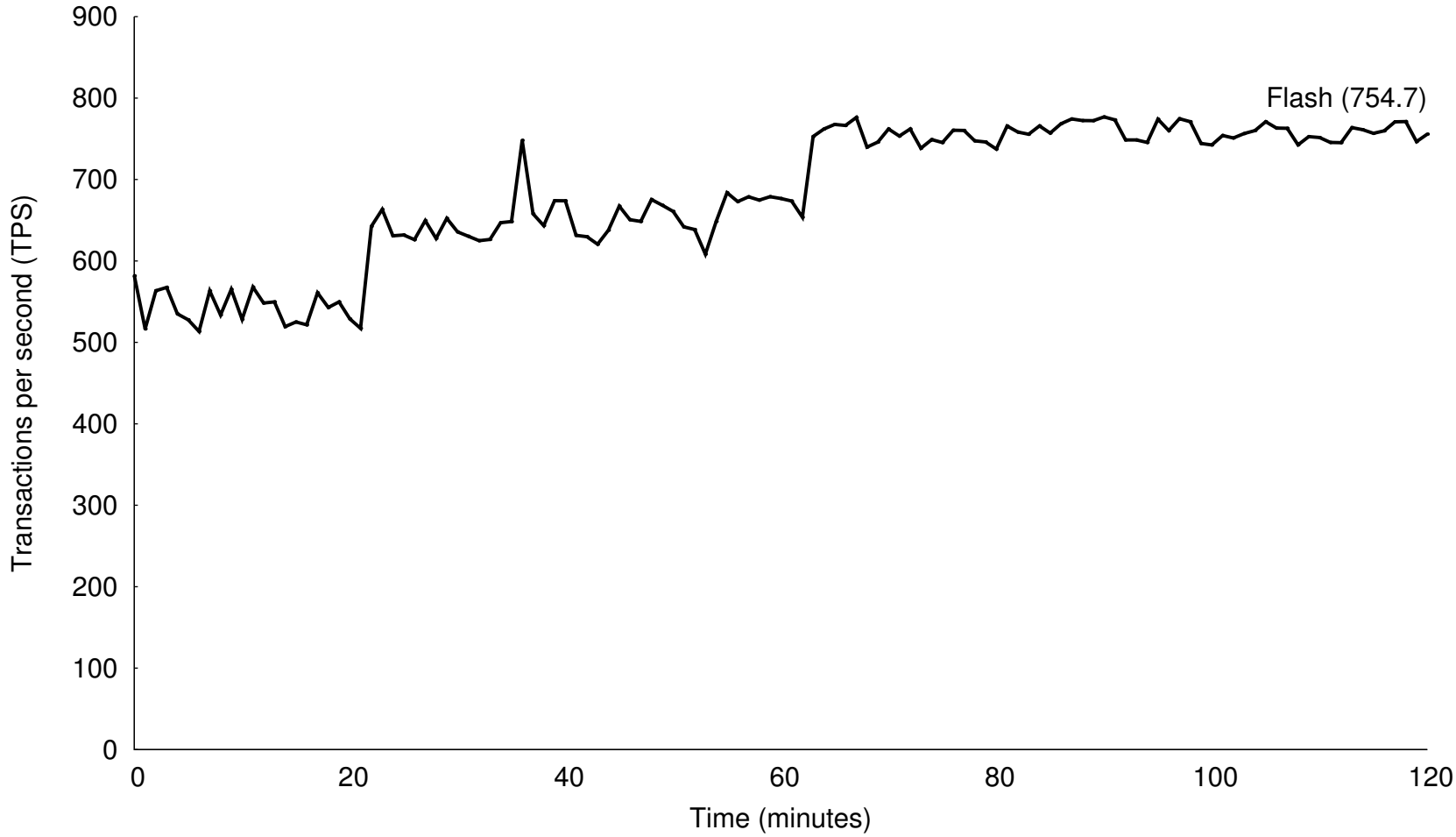
Sysbench OLTP benchmark, read only 1 thread: eMLC SSD



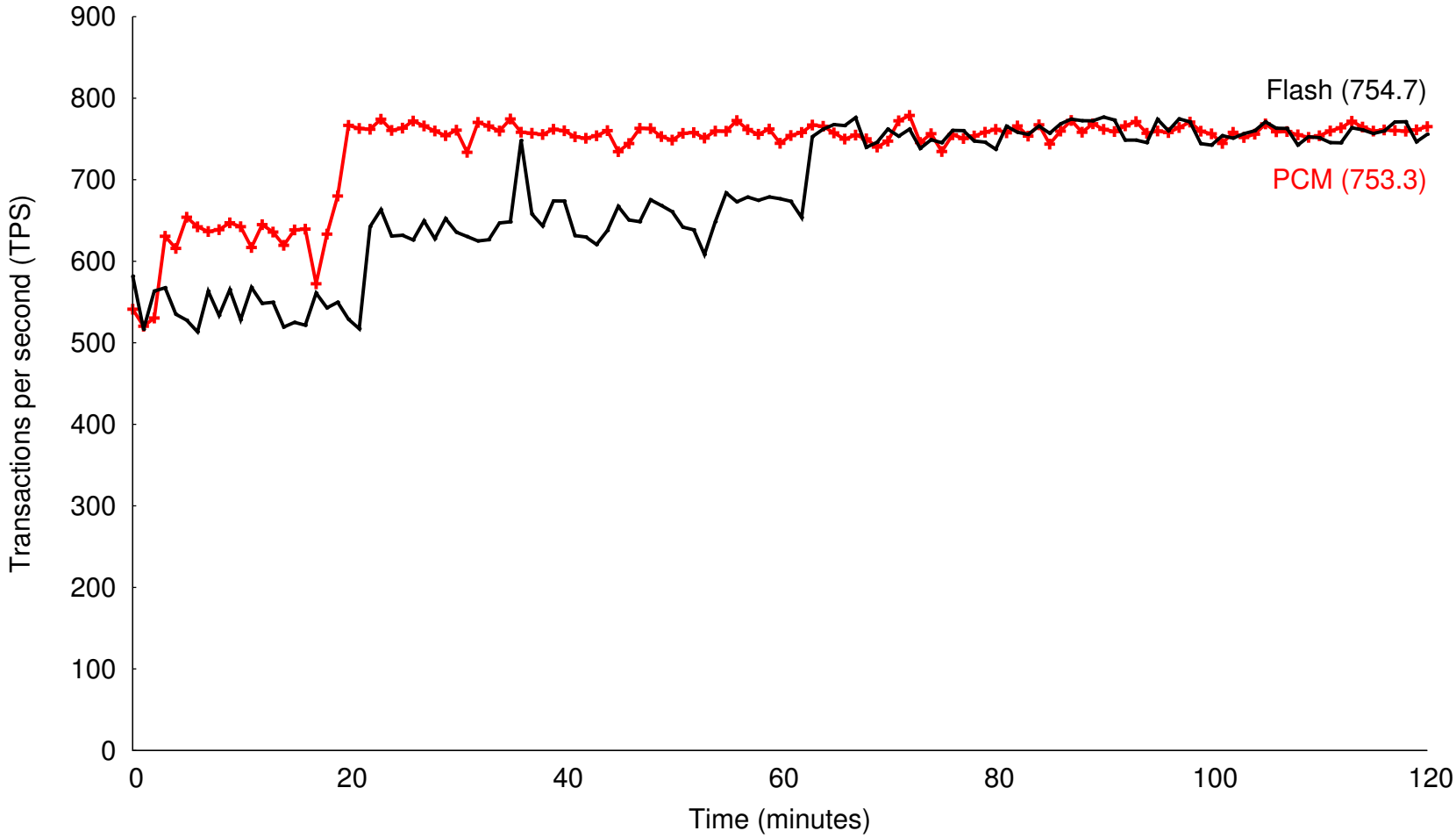
Sysbench OLTP benchmark, read only 1 thread: eMLC SSD vs. PCM SSD



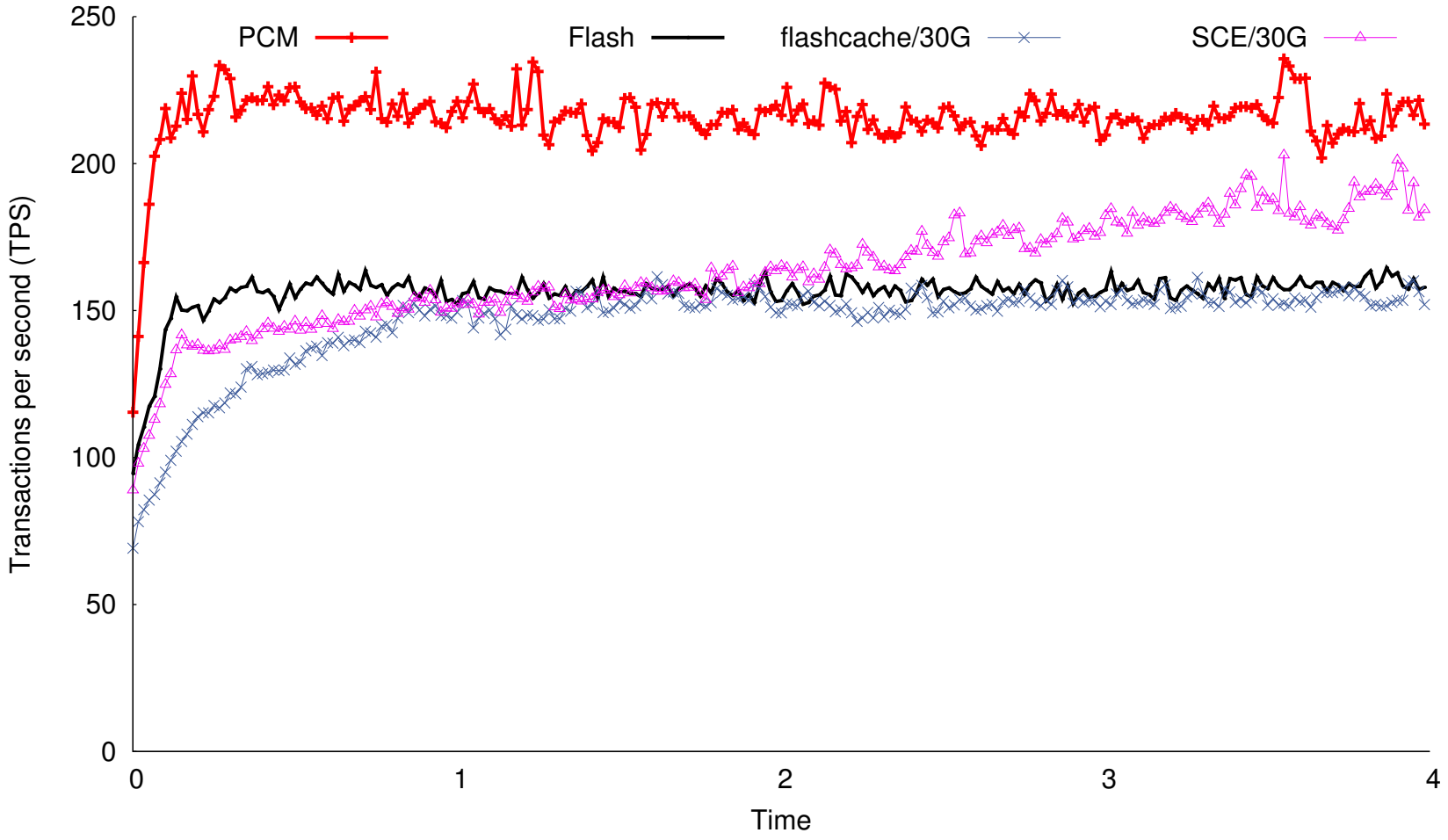
Sysbench OLTP benchmark, read only 8 thread: eMLC SSD

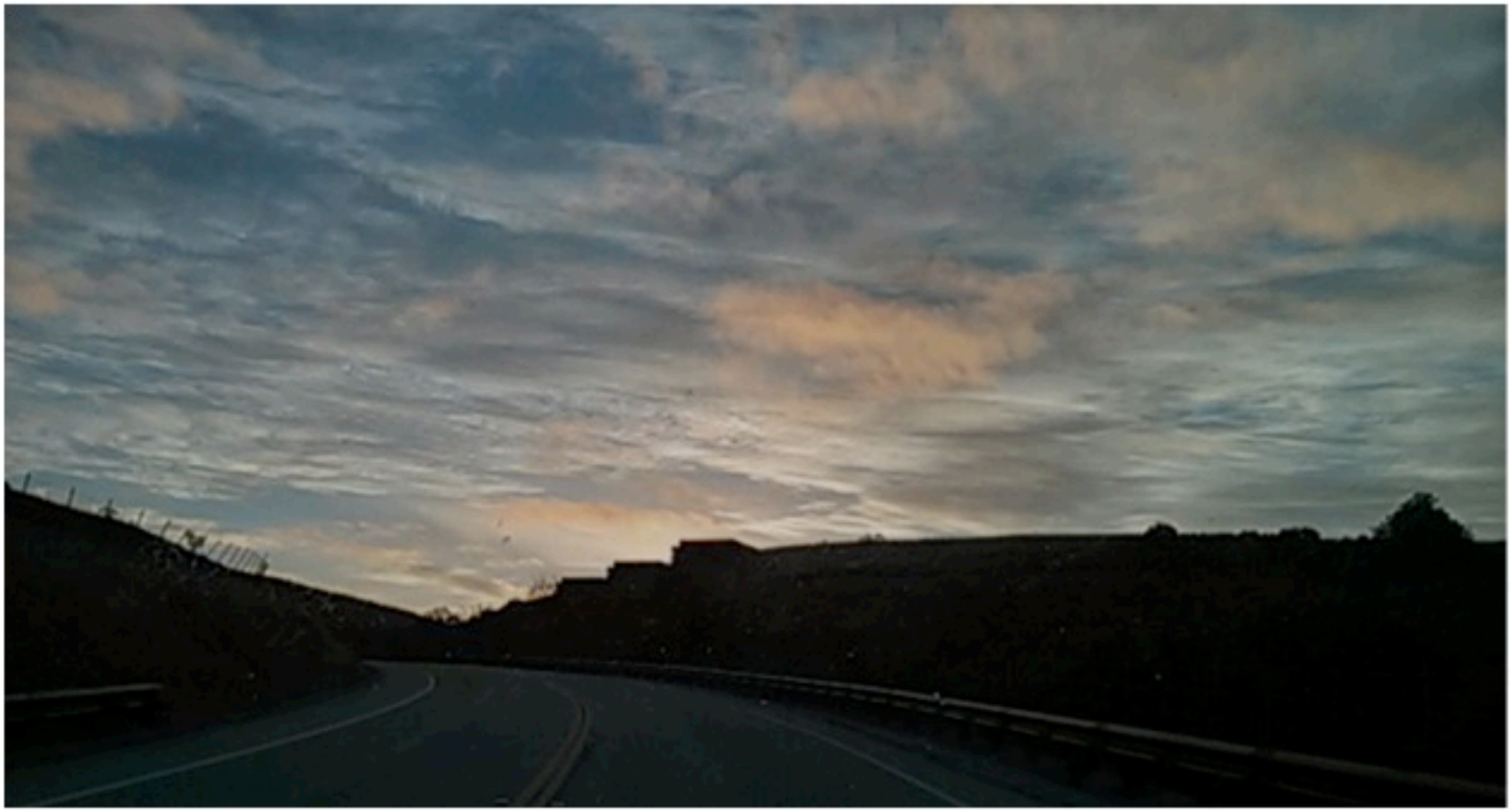


Sysbench OLTP benchmark, read only 8 thread: eMLC SSD vs. PCM SSD



PCM caching over eMLC flash SSD





Thank you