Trends in Solid State Storage

Jeff Kimmel
NetApp

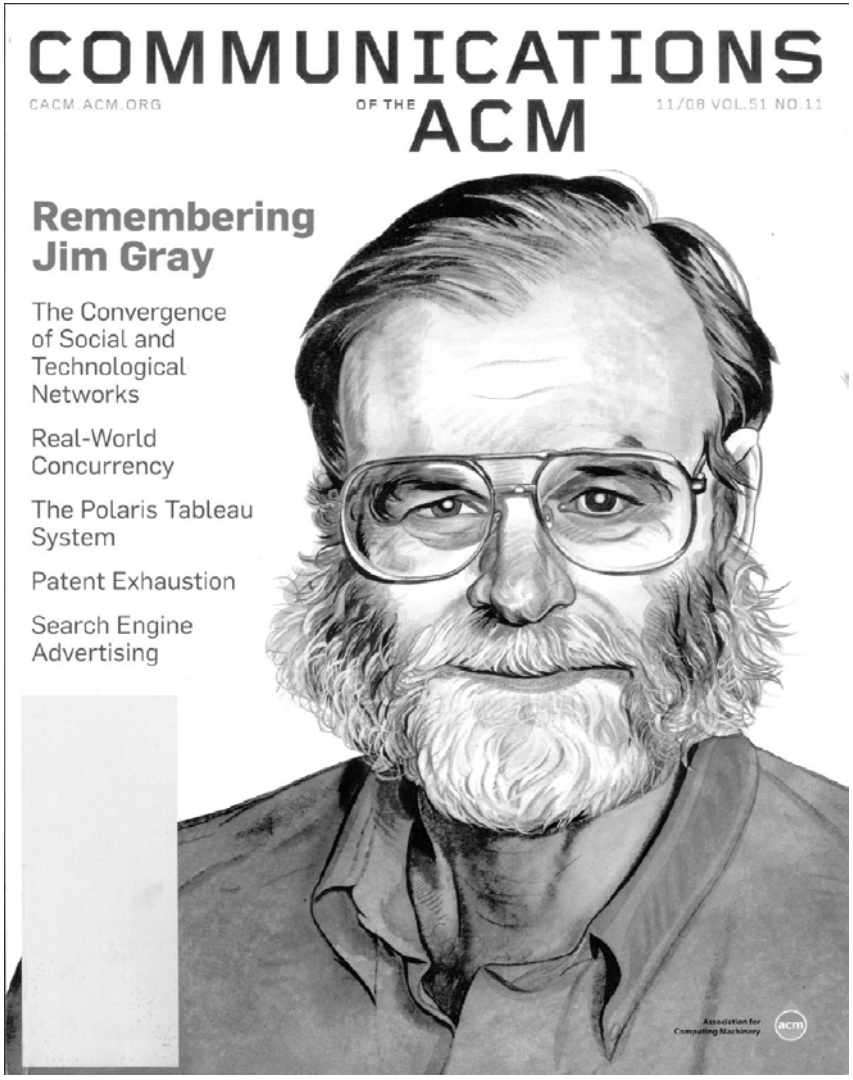
Abstract

- Trends in Solid State Storage
  - This talk explores trends in solid state storage and their likely impacts on system architectures. Included are SLC/MLC considerations, read and write efficiencies vs. HDD, flash caching layers and their placement in systems and tradeoffs between caching and use of SSDs.
  - Qualifiers:
    - "Solid State" focus on NAND flash, not DRAM, PCM, memristors, …
    - "Storage" == Enterprise data center storage

Agenda

- Why flash in the datacenter? Why now?
- Memory, cache and storage
- Application opportunities
- Flash in enterprise storage today
  - SSD storage tier
  - Network cache
  - Storage controller-based cache
- What's next
- Conclusion

Remembering Jim Gray



Database and systems design pioneer, and co-creator of the Five Minute Rule (1987)

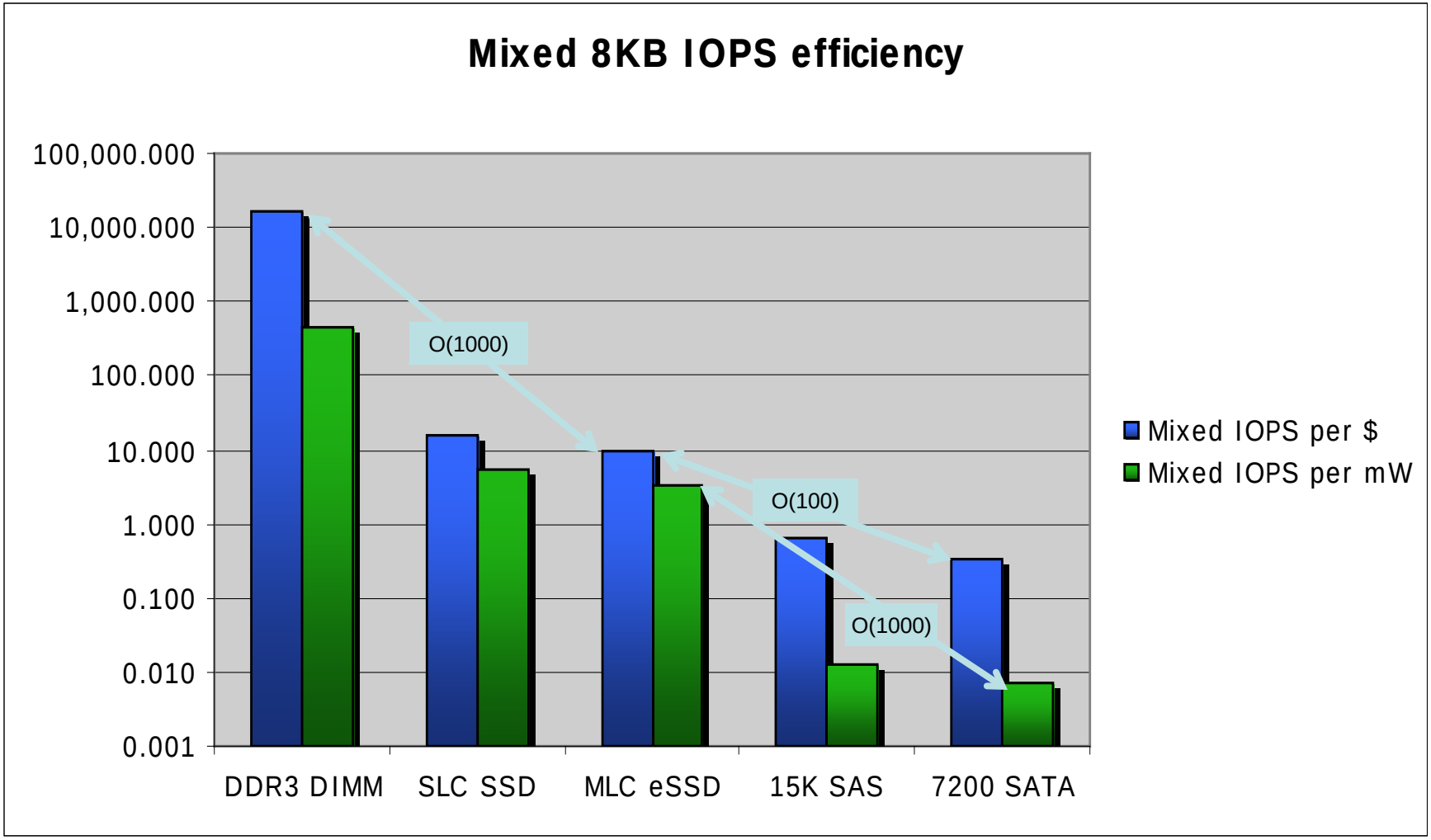
"Flash is a better disk …, and disk is a better tape" ~2006

Lost at sea January 2007

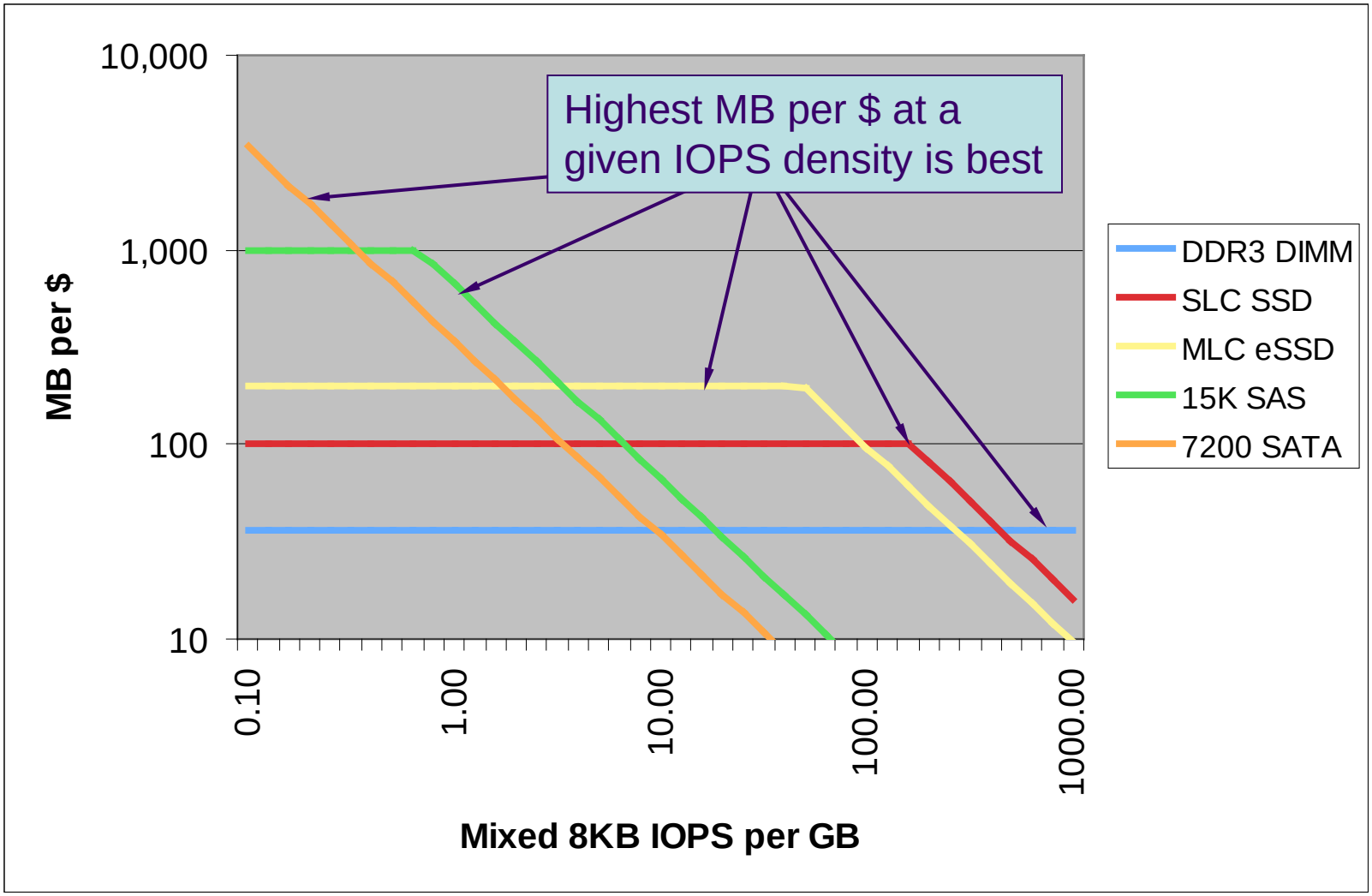
Why flash in the datacenter now?

- Why flash?
  - Capacity efficiency versus DRAM
    - ~5x better $ per GB
    - ~40x better power per GB
  - IOPS efficiency versus HDDs
    - ~40x better $ per IOPS
    - ~600x better power per IOPS
- Why now?
  - Period of rapid density advancements led to HDD-like bit density at lower $/GB than DRAM
  - Innovations in SSD and tiering technology

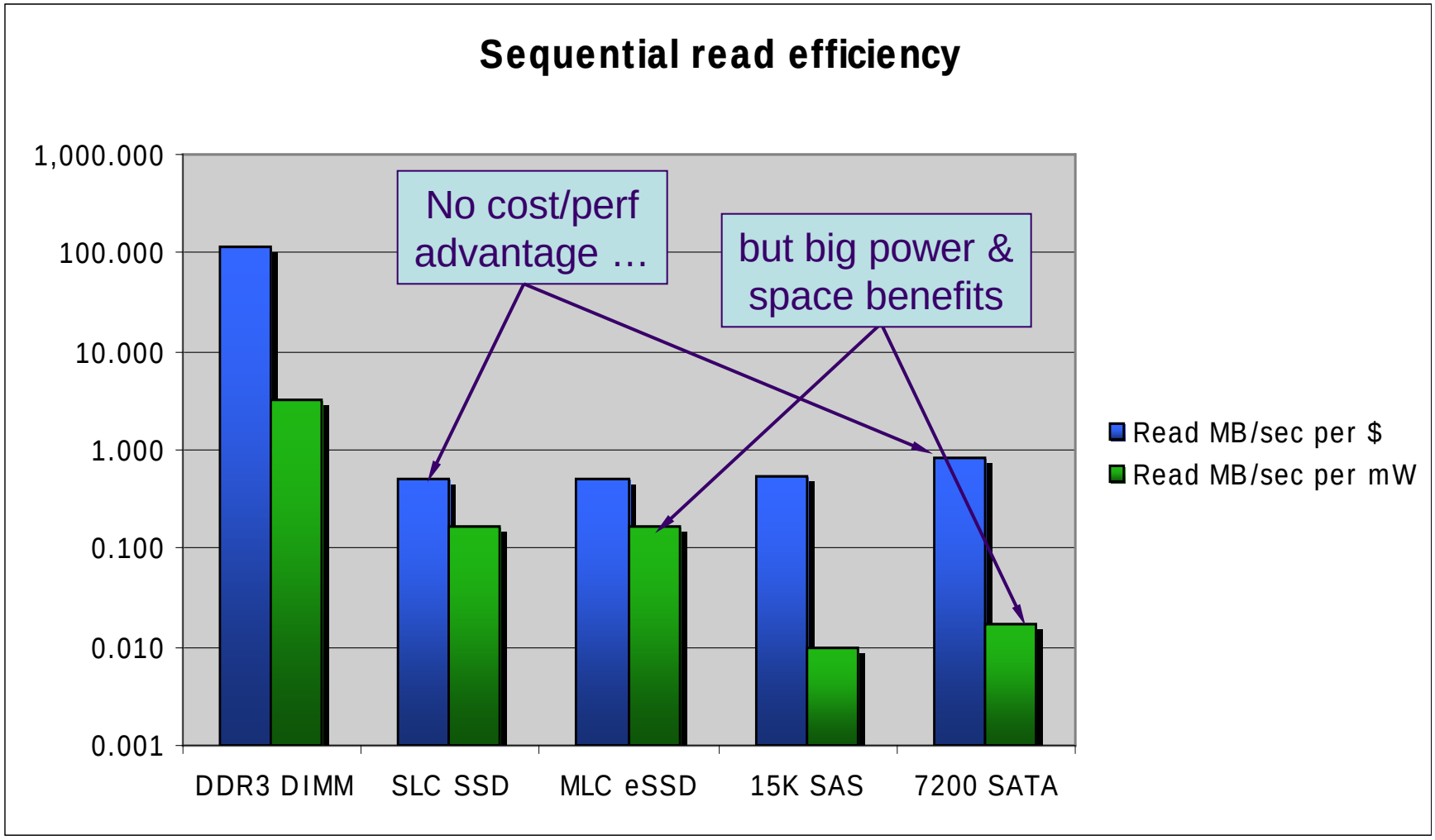
Why flash? IOPS efficiency



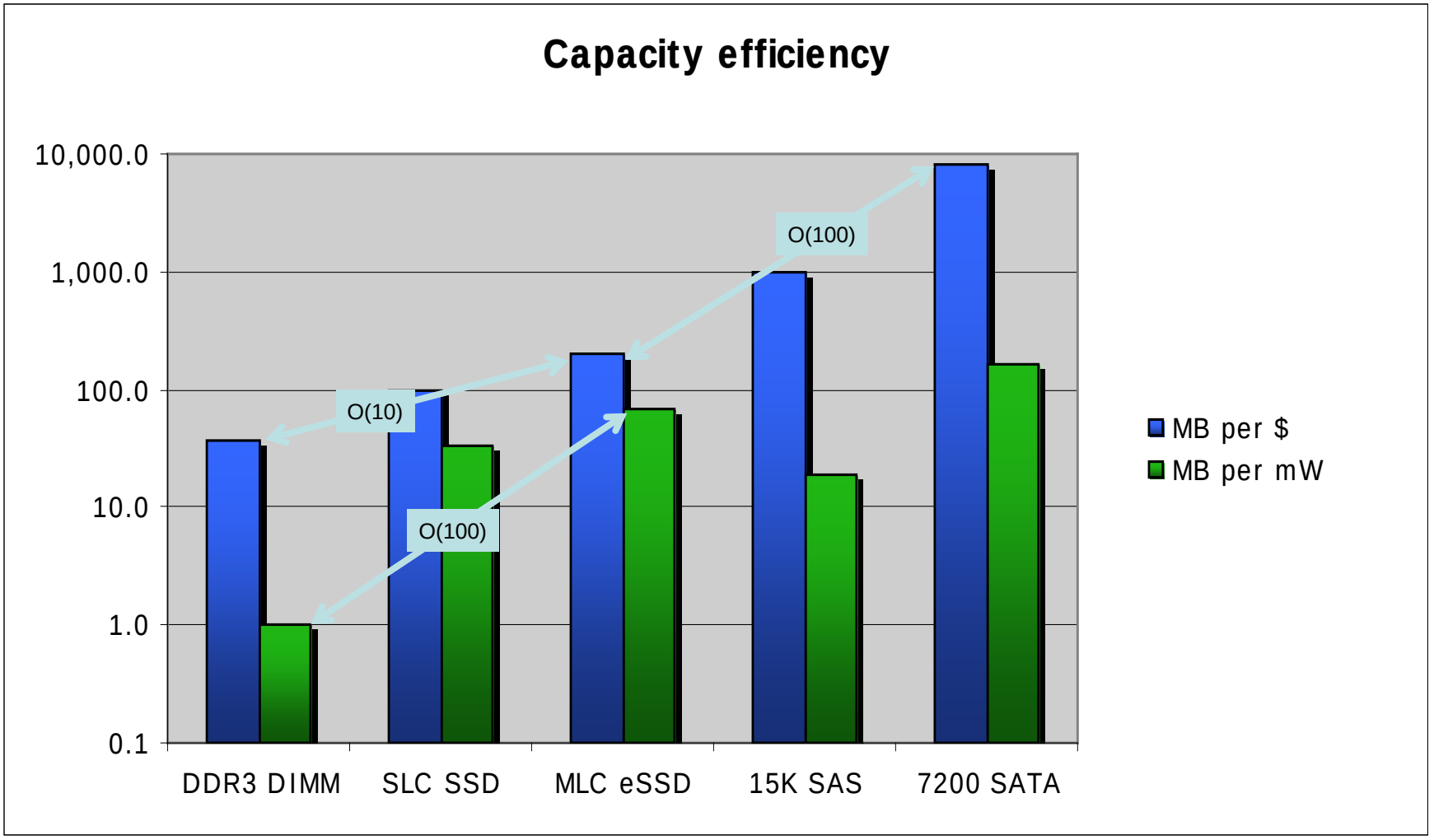
Why flash? An IOPS density view



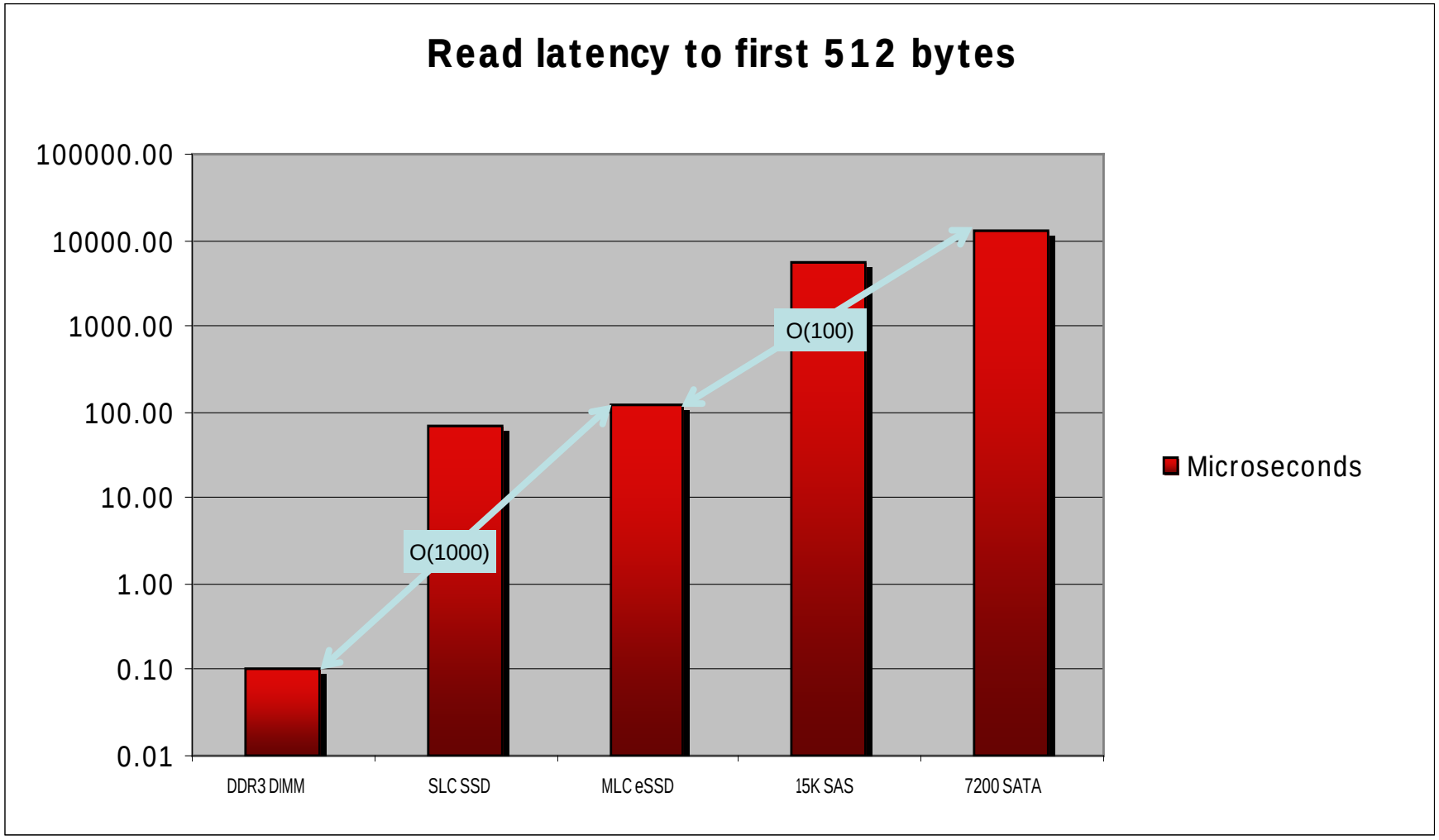
Why flash? Read throughput per watt



Why flash? Capacity efficiency



Why flash? Read latency



Five Minute Rule, 1987

- Assuming that the cost of a cache is dominated by its capacity, and the cost of a backing store is dominated by its access cost (cost per IOPS), then the breakeven interval for accessing a page of data in cache is given by:

  Break-Even-Interval =
  Backing-Store-Cost-Per-IOPS /
  Cache-Cost-Per-Page

- 1987: Disk $2,000 / IOPS; RAM $5 / KB ➔
  1 KB breakeven = 400 seconds ~= 5 minutes

Five Minute Rule, 2010: DRAM & HDD

- Disk $1 / IOPS (2,000x reduction)

- DRAM $25 / GB (200,000x reduction)

→ 100 KB breakeven ~= 5 minutes
→ 8 KB breakeven ~= 1 *hour*
→ 1 KB breakeven ~= 10 hours *as Gray predicted*

- 200,000x / 2,000x = 100-fold decrease in breakeven access rate for a DRAM cache page backed by disk
  → much bigger DRAM caches

Five Minute Rule, 2010: Flash & HDD

- Disk $1 / IOPS
- MLC eSSD ~$5 / GB

→ SSD 100 KB breakeven ~= 30 minutes
→ SSD 8 KB breakeven ~= 7 hours (5x DRAM)

Flash economically caches working sets with 5x longer access intervals than DRAM.

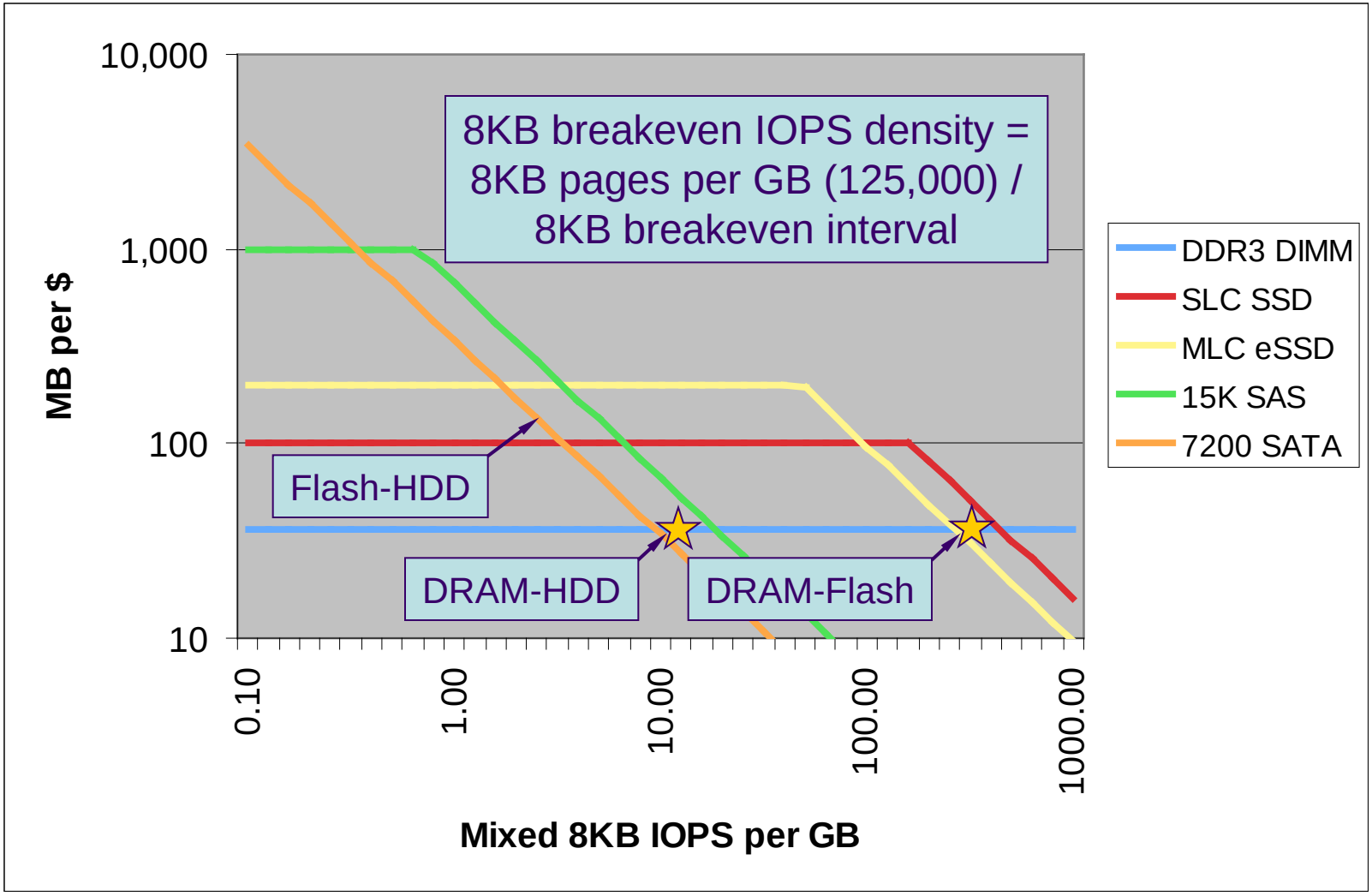
Five Minute Rule, 2010: DRAM & Flash

- MLC eSSD ~$0.10 / mixed 8 KB IOPS
- DRAM $25 / GB

➔8 KB breakeven ~= 8 minutes (1/10th DRAM)

Adding flash between DRAM and HDD reduces the breakeven access interval for DRAM by 10x, indicating that DRAM capacity could be reduced to hold working sets for data accessed 1/10th as often

IOPS density and the Five Minute Rule



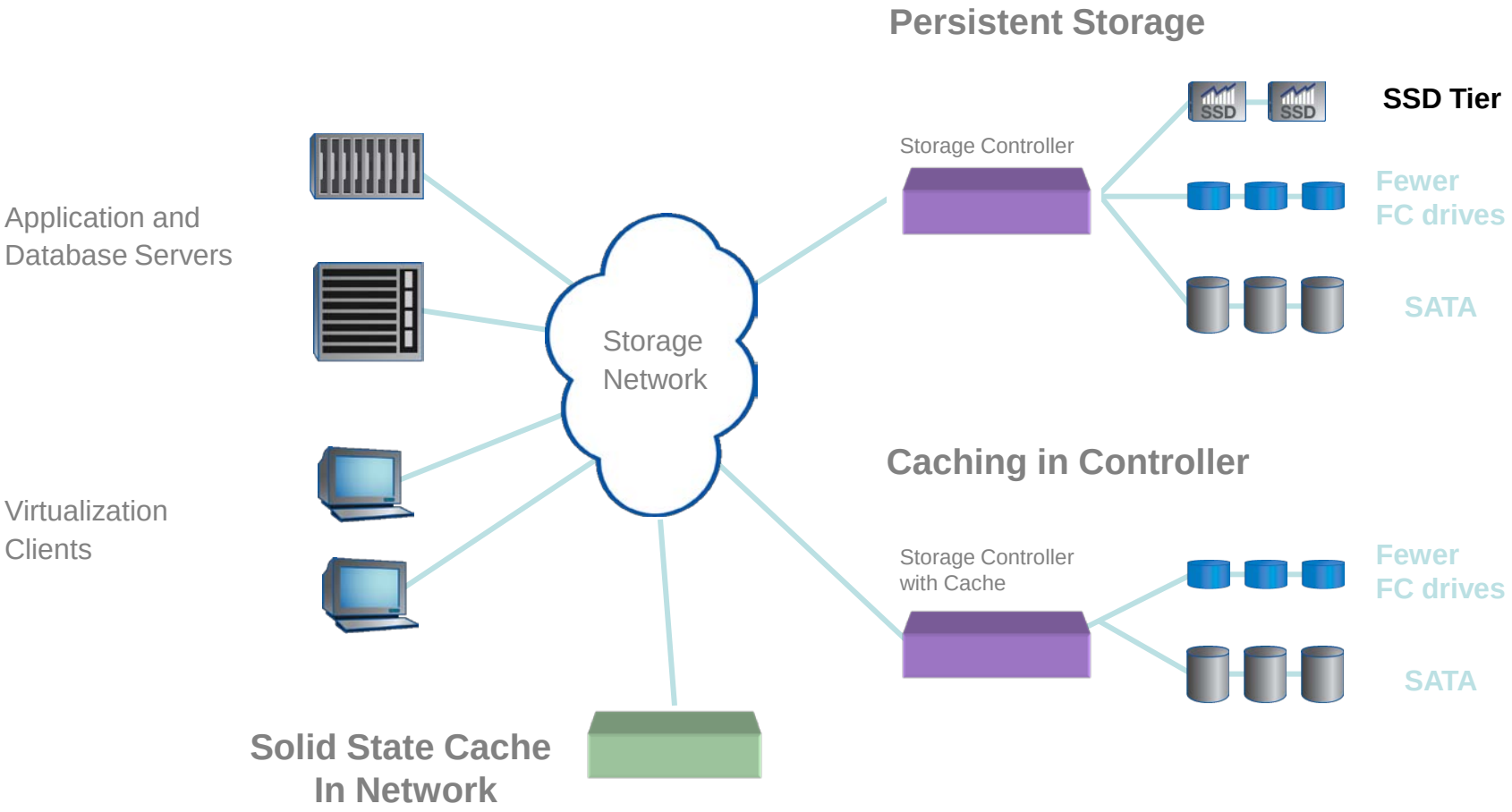
Implications for systems design

- Flash makes it cost-effective to keep more small random data in silicon-based cache versus DRAM:
  ~5+ hour working set versus ~1 hour
- Flash allows small random data working set in DRAM to be reduced, allowing cost, power, space efficiency:
  ~5 minute working set versus ~1 hour
- Assuming appropriate locality of reference, transfer sizes between HDD and flash tiers should increase to preserve expensive HDD IOPS
- Flash tier likely to alter checkpoint processing intervals (shorter), metadata organization (e.g. optimal page size)

Application opportunities

- Intense random reads, e.g. OLTP, metadata
- Sequential read after random write
  - Log-oriented writes convert this to random read after sequential write (e.g. FTL)
- Low read latency (~100x better than HDD)
  - Facilitates DRAM extension by allowing high read throughput with limited read concurrency
  - Paging datacenter apps can be practical again
  - Memory capacity to consolidate more servers with underutilized CPU
- Enabling memory-resident datasets, e.g.
  - OLTP
  - Data warehouses (*viz* TPC-H results)
  - Large metadata

Storage networking with flash



Available solutions: Pros and cons

	Pros	Cons
Solid State Drives	❒ Assured performance levels ❒ Low cost per IOPS ❒ Administrator has direct control over data stored in SSD tier	❒ High cost per gigabyte ❒ Requires manual partitioning of hot data ❒ Limited practical applications
Controller Cache	❒ Hot data automatically flows into cache – no administration required ➔ automated efficiency benefit ❒ Deployment can be non-disruptive ❒ Viable for common enterprise applications – cache "just helps"	❒ Cache must be populated before it becomes effective ❒ Less predictable performance than static placement
Network Cache	❒ Hot data automatically flows into the caching tier ❒ Deployment is relatively non-disruptive ❒ Scalable solution for high performance applications	❒ Cache must be populated before it becomes effective ❒ Less predictable performance than static placement ❒ Placement in front of storage may constrain protocols or use cases

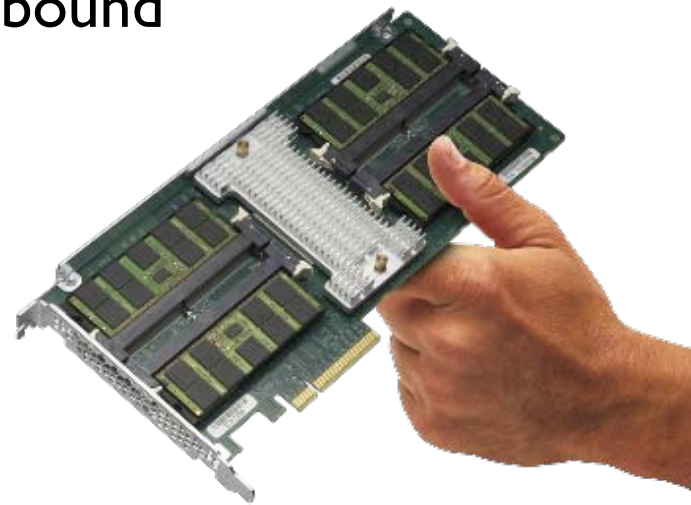
SSD-based Solutions

- **Database acceleration solution**
  - Entire DB on SSD tier
  - Or hot random access files on SSD and rest of DB on standard disk
    - Indexes and temp space
- **Large scale virtual machine environments**
  - Solves "boot storm" problem for large numbers of virtual machines
  - Dedupe of VM data, e.g. virtual desktops, reduces capacity requirements, increasing IOPS density, potentially making SSD economical
- **Network cache solutions**
  - All files on HDD in shared storage array
  - Accelerated by SSD-based network cache
  - Self–tuning write-through cache
  - Applications include
    - Rendering, seismic, financial modeling, ASIC design

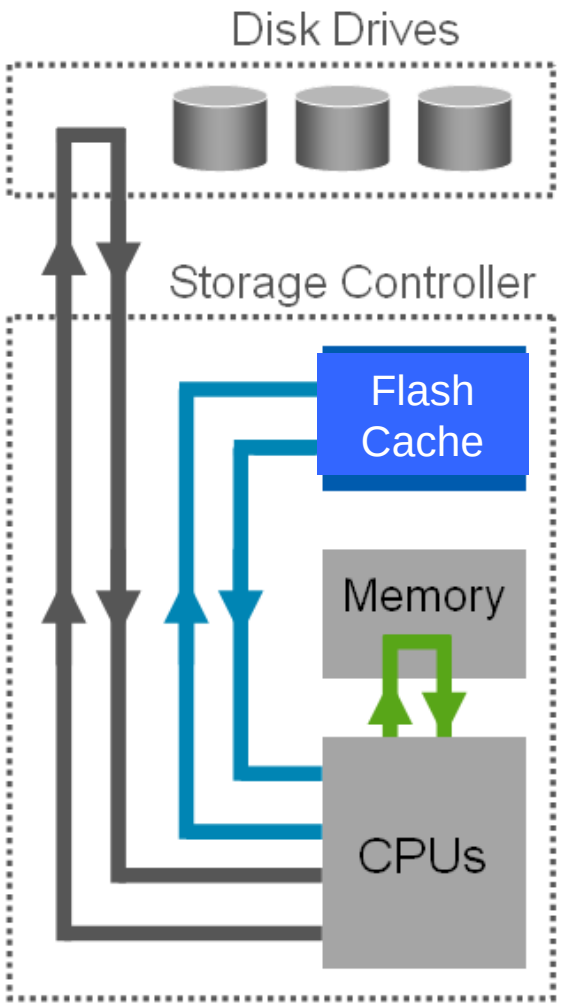
20

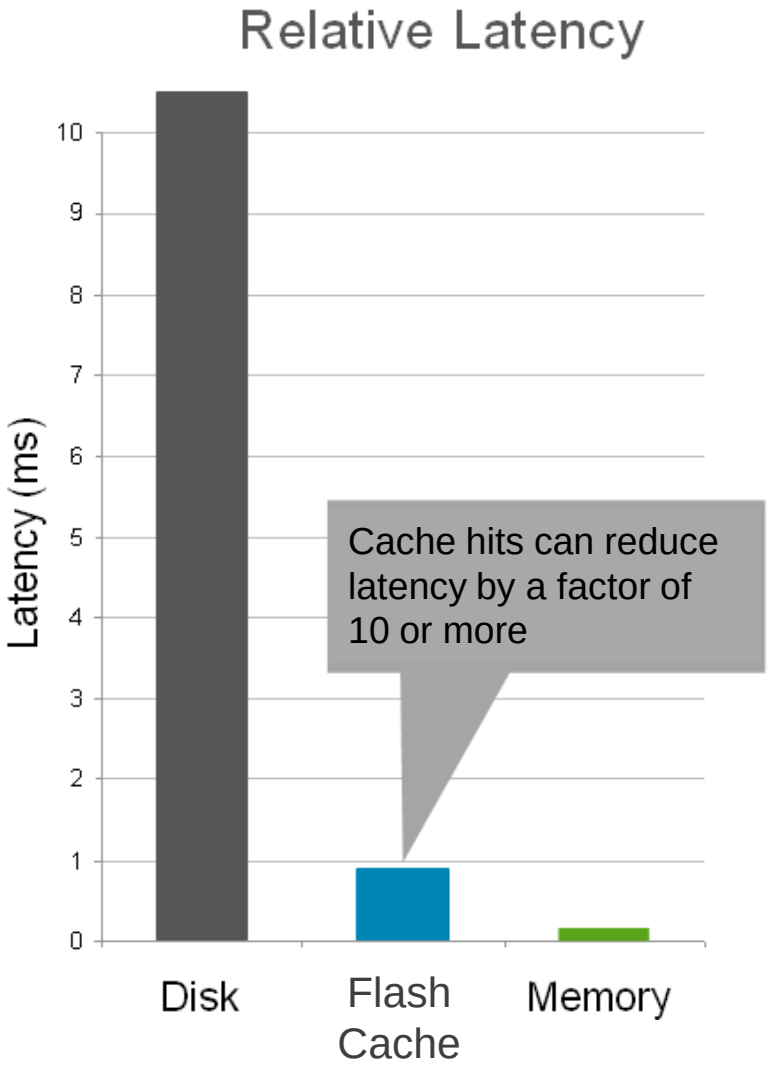
Controller-based Flash Cache

- Functions as an intelligent read cache for data and metadata
- Automatically places active data where access can be fast
- Provides more I/O throughput without adding high-performance disk drives to a disk-bound storage system
- Effective for file services, OLTP databases, messaging, and virtual infrastructure

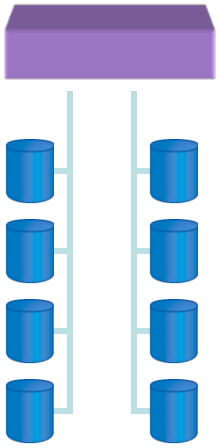


Reduce Latency with Flash Cache



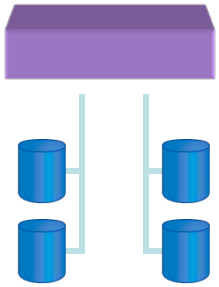


Use case: Scale Performance of Disk-bound Systems



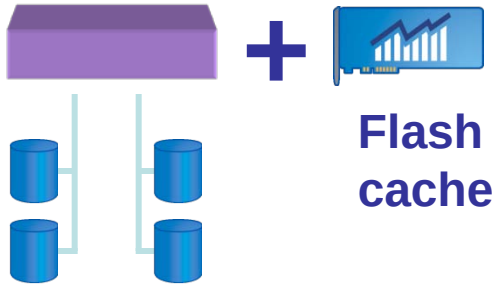
Add Spindles

- Use more disks to provide more IOPs
- May waste storage capacity
- Consumes more power and space



Starting Point:
Need More IOPs

- Performance is disk-bound
- Have enough storage capacity
- Random read intensive workload

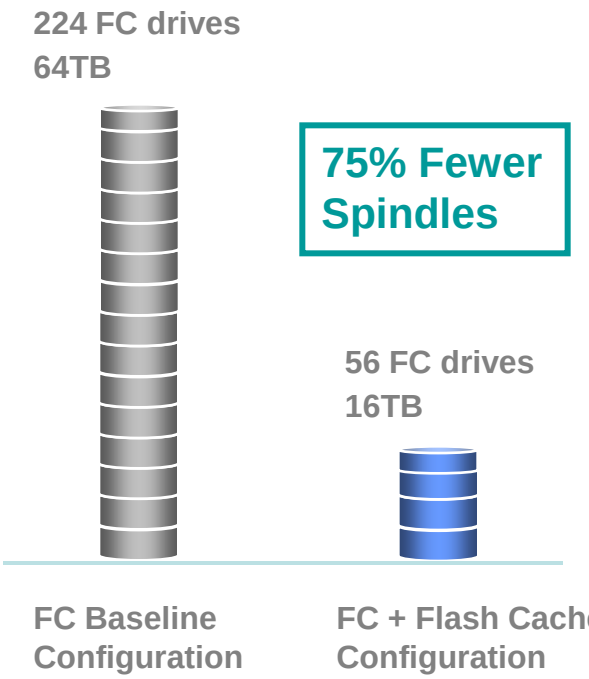


Add Flash Cache

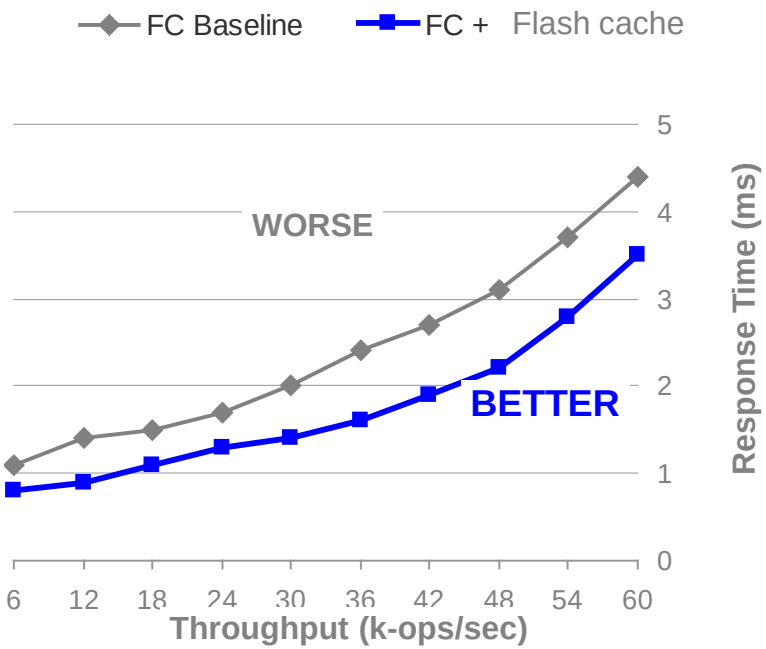
- Use cache to provide more IOPs
- Improves response times
- Uses storage efficiently
- Achieves cost savings for storage, power, and space

FC HDD plus Flash Cache Example

Benchmarked Configurations



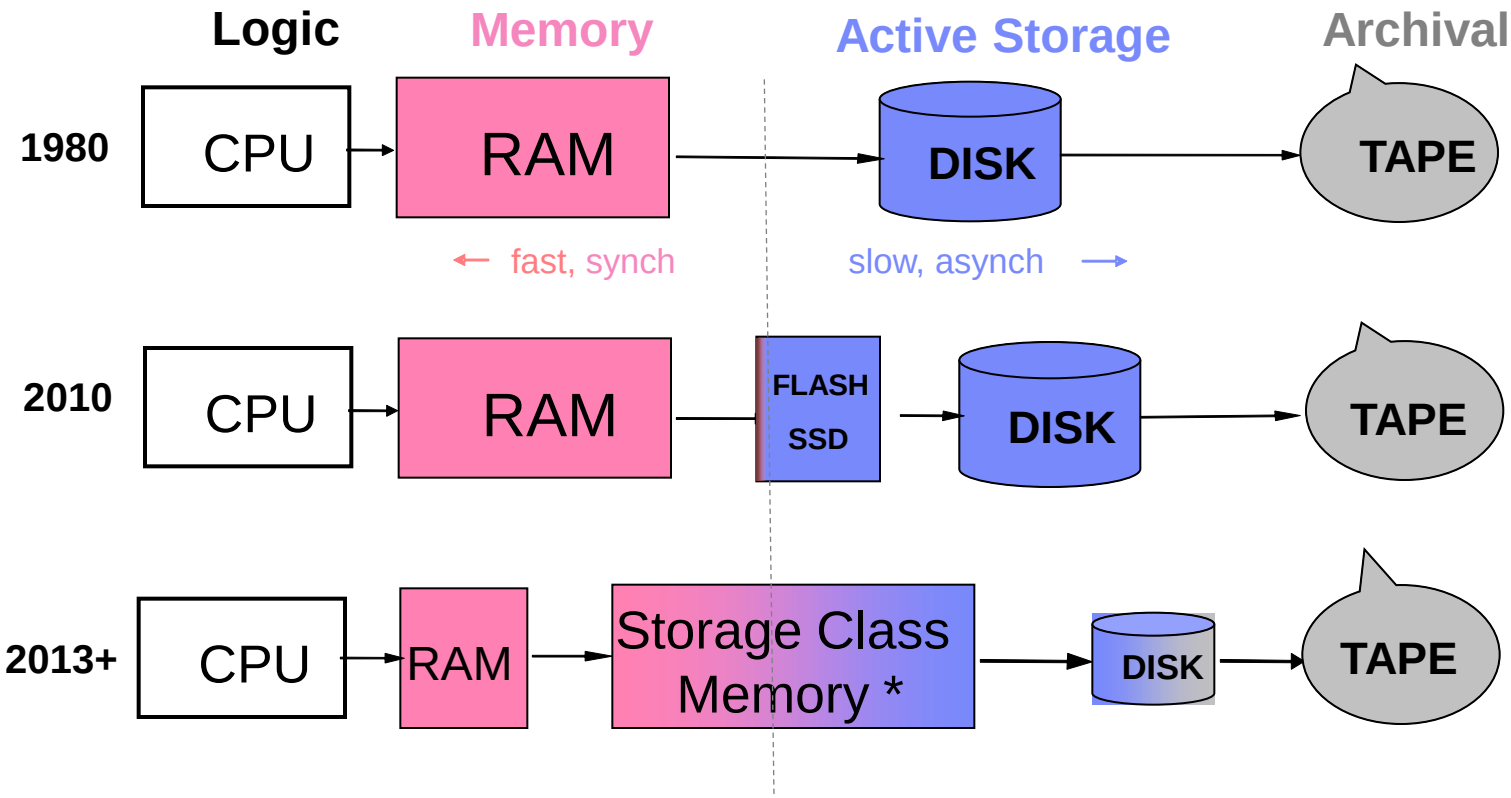
SPECsfs2008 Performance



- Purchase price is **50% lower** for FC + Flash cache compared to Fibre Channel baseline
- FC + Flash cache yields **67% power savings** and **67% space savings**

For more information, visit http://spec.org/sfs2008/results/sfs2008nfs.html.
SPEC® and SPECsfs2008® are trademarks of the Standard Performance Evaluation Corp.

System Evolution



* e.g. Phase change memory
Memristor
Solid Electrolyte
Racetrack memory

Cost Structure of Memory/Storage Technologies

Cost determined by

- cost per wafer
- # of dies/wafer
- memory area per die [sq. μm]
- memory density [bits per $4F^2$]
- patterning density [sq. μm per $4F^2$]

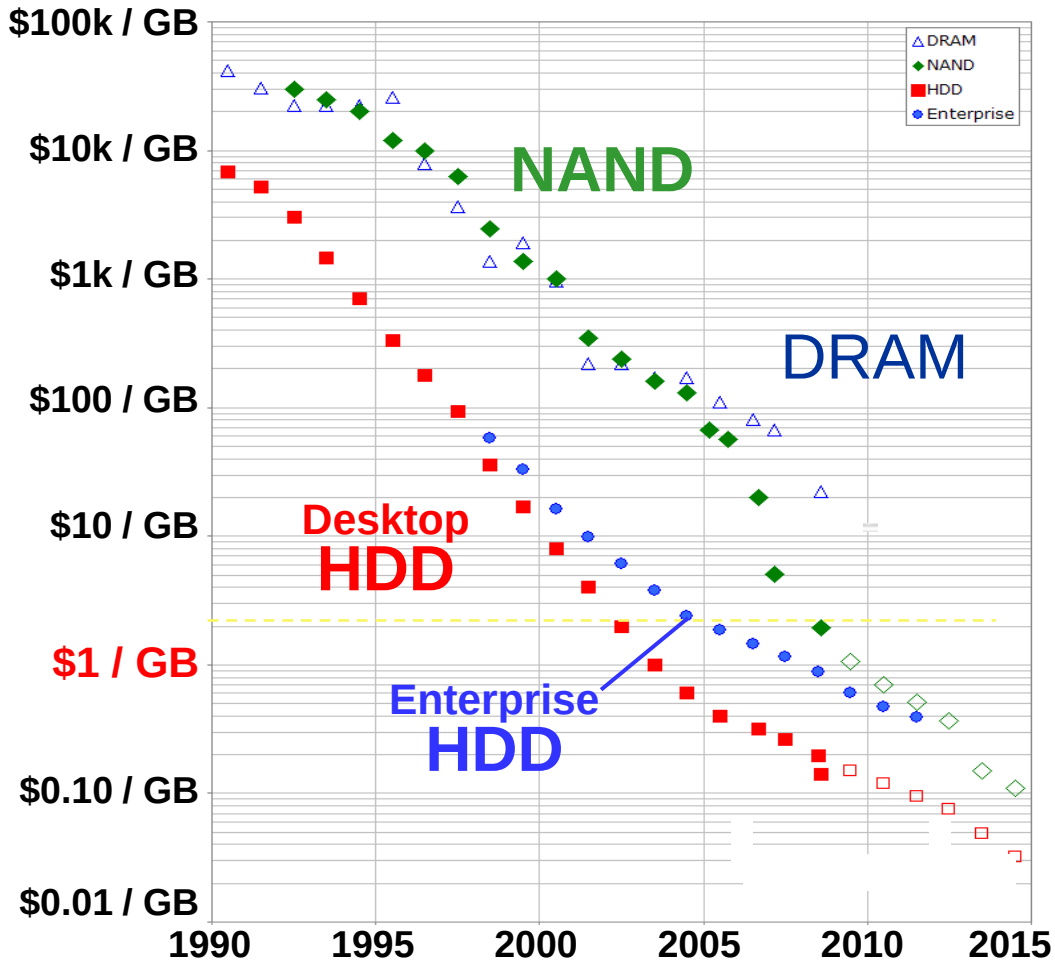


Chart courtesy of Dr. Chung Lam, IBM Research updated version of plot from 2008 *IBM Journal R&D* article

Summary

- Over the next 5 years solid state technologies will have a profound impact on enterprise storage
- It's not just about replacing mechanical media with solid state media
- The architectural balance of memory, cache and persistent storage will change
- Today's solid state implementations in enterprise storage demonstrate these changes
- "Flash is a better disk, and disk is a better tape"