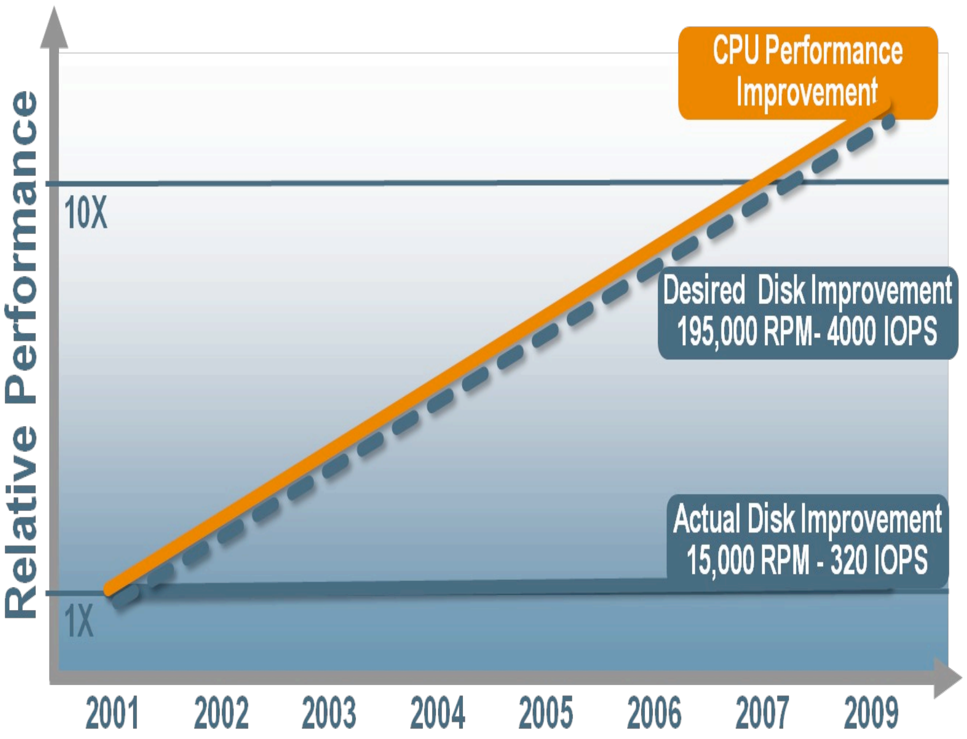
Flash Performance in Storage Systems

Bill Moore
Chief Engineer, Storage Systems
Sun Microsystems

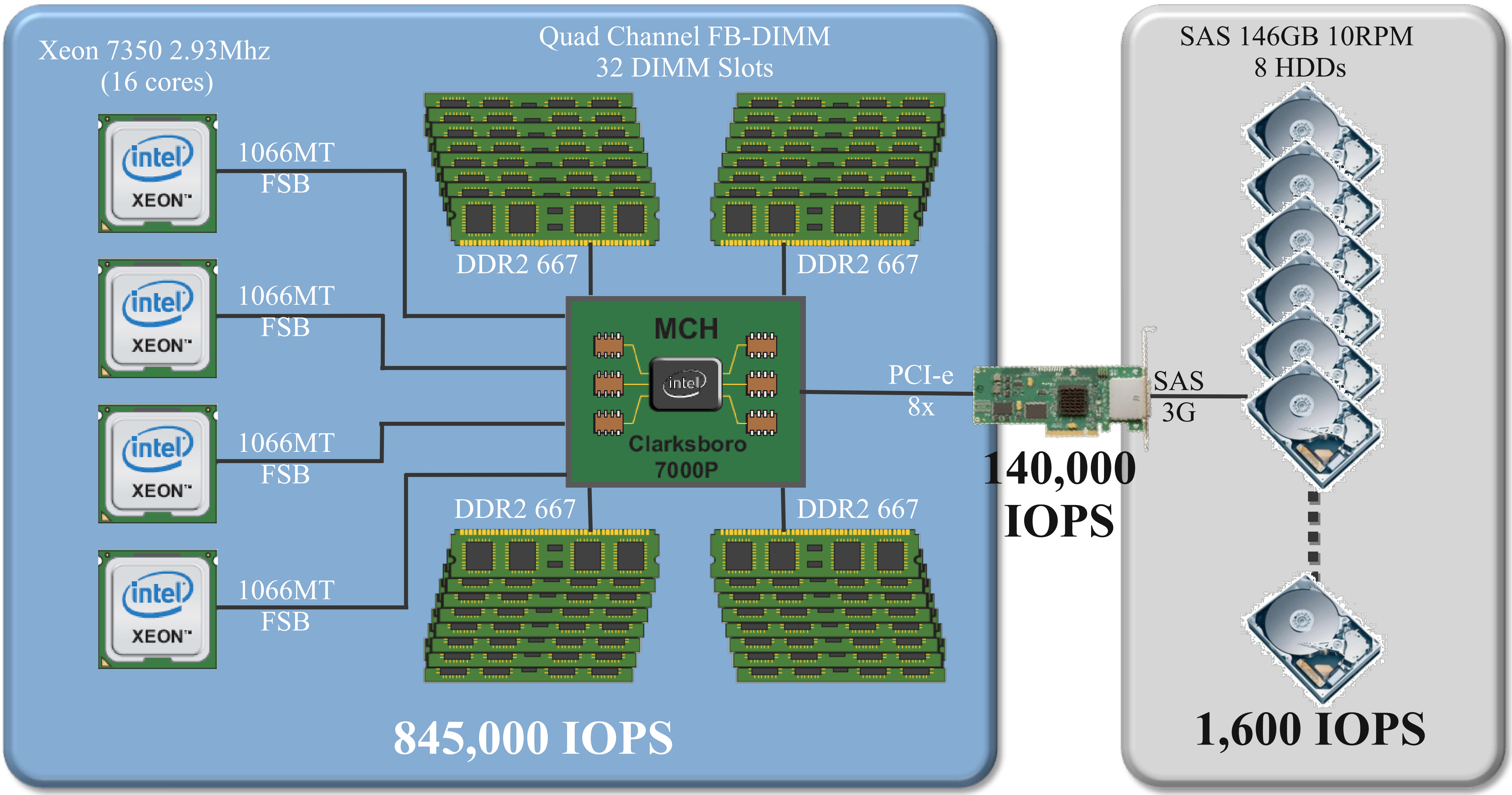
Disk to CPU Discontinuity

- Moore's Law is out-stripping disk drive performance (rotational speed)
- As a result, servers and storage systems are hopelessly unbalanced between CPU/controller capability and storage pool performance
- The objective of modern systems design is to rebalance the CPU-storage ecosystem while optimizing both low $/GB and $/IOPS

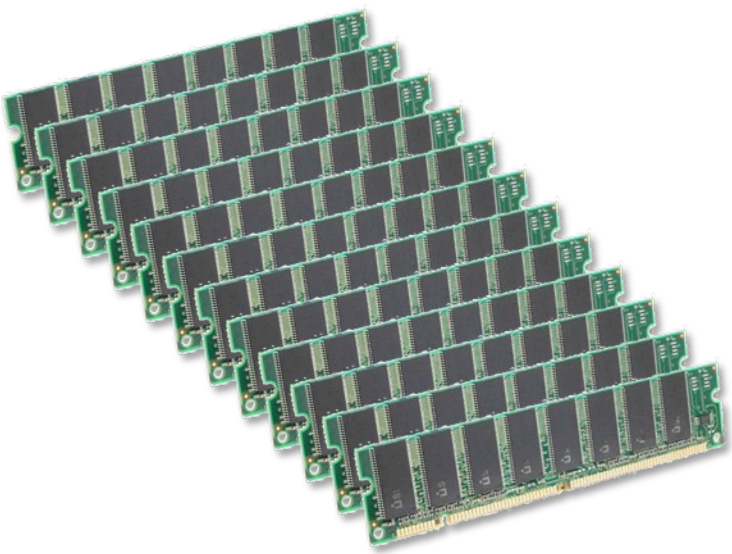


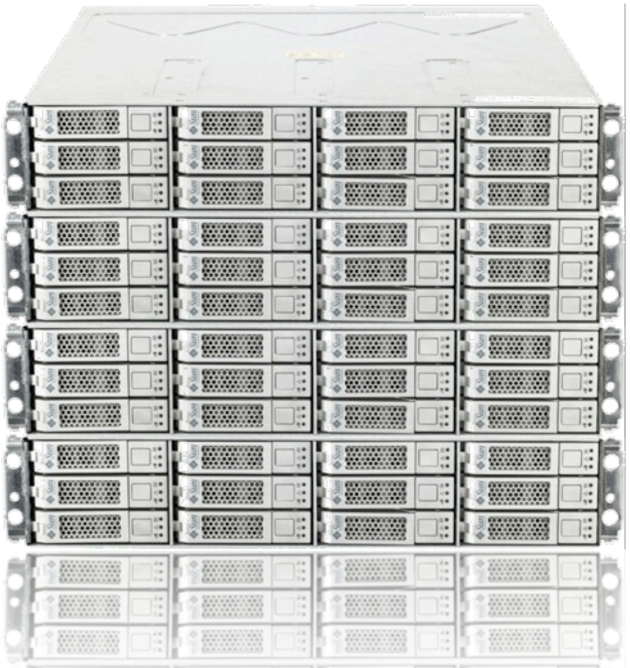
CPU-Storage Bottleneck

SunFire x4450 Memory Hierarchy



Traditional Architecture for Performance





- Store working set in DRAM to avoid latency
- Use 1-2G (Netapp) or 1-2T (EMC DMX) of battery-backed DRAM to buffer synchronous writes

- Huge pool of high RPM, high power spindles
- “Short-stroked” so data lives on outer tracks
- Reduce head seeks, increase IOPS

Example: TPC-C Benchmark

Lots of Disk, Lots of DRAM

BIG BLUES: IBM System x3950 M2*

Total Hardware Cost: **$2,911,464**

The Hidden Performance Cost:

512GB DRAM (**64** x 8GB DIMMs) : **$960,000**

47.7TB HDD (**1344** x 36.4GB 15K RPM HDDs): **$1,198,848**

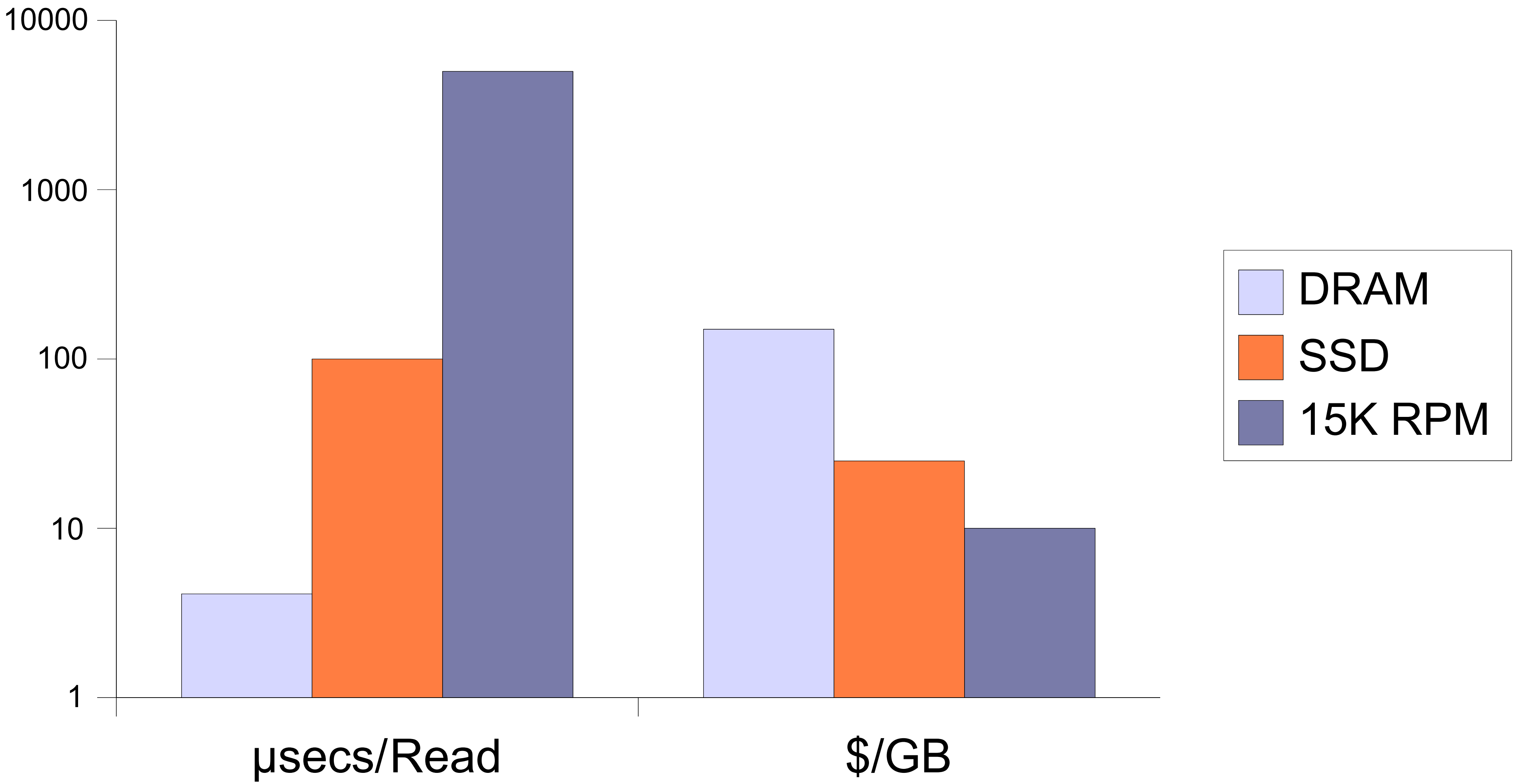
74% of the cost is for Storage Performance

Flash Background

- 2004 flash cost as much as DRAM
- 2008 flash now a fraction of the cost
  - > But still much more than rotating media
- Very low-latency reads (~50μs)
- Relatively slow writes (> 200μs)
- Limited write/erase cycles (100k-1m)
- Compelling for storage market in the right package:
  - > Fronted by DRAM to buffer the write latency
  - > DRAM backed by super-capacitor to drain on power fail
  - > Multiple channels of Flash to increase parallelism
- Complicated to integrate intelligently
  - > Not just another tier in a traditional HSM model

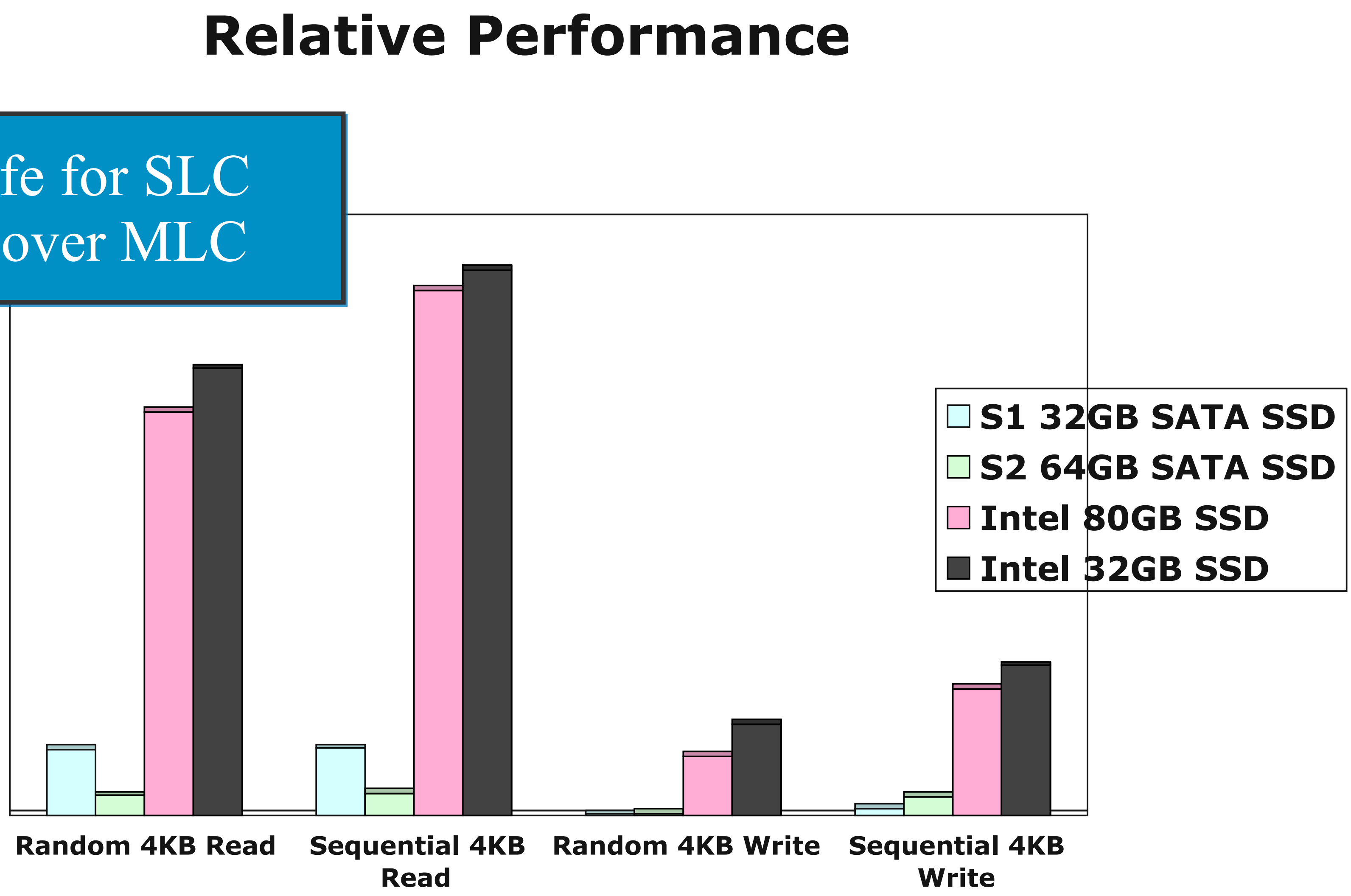
Example: Read-biased Flash in a Memory Hierarchy

DRAM, 15K RPM Drives and Read SSD: Price and Performance



SSD Benchmark Relative Performance

Cycle life for SLC
is 10X over MLC

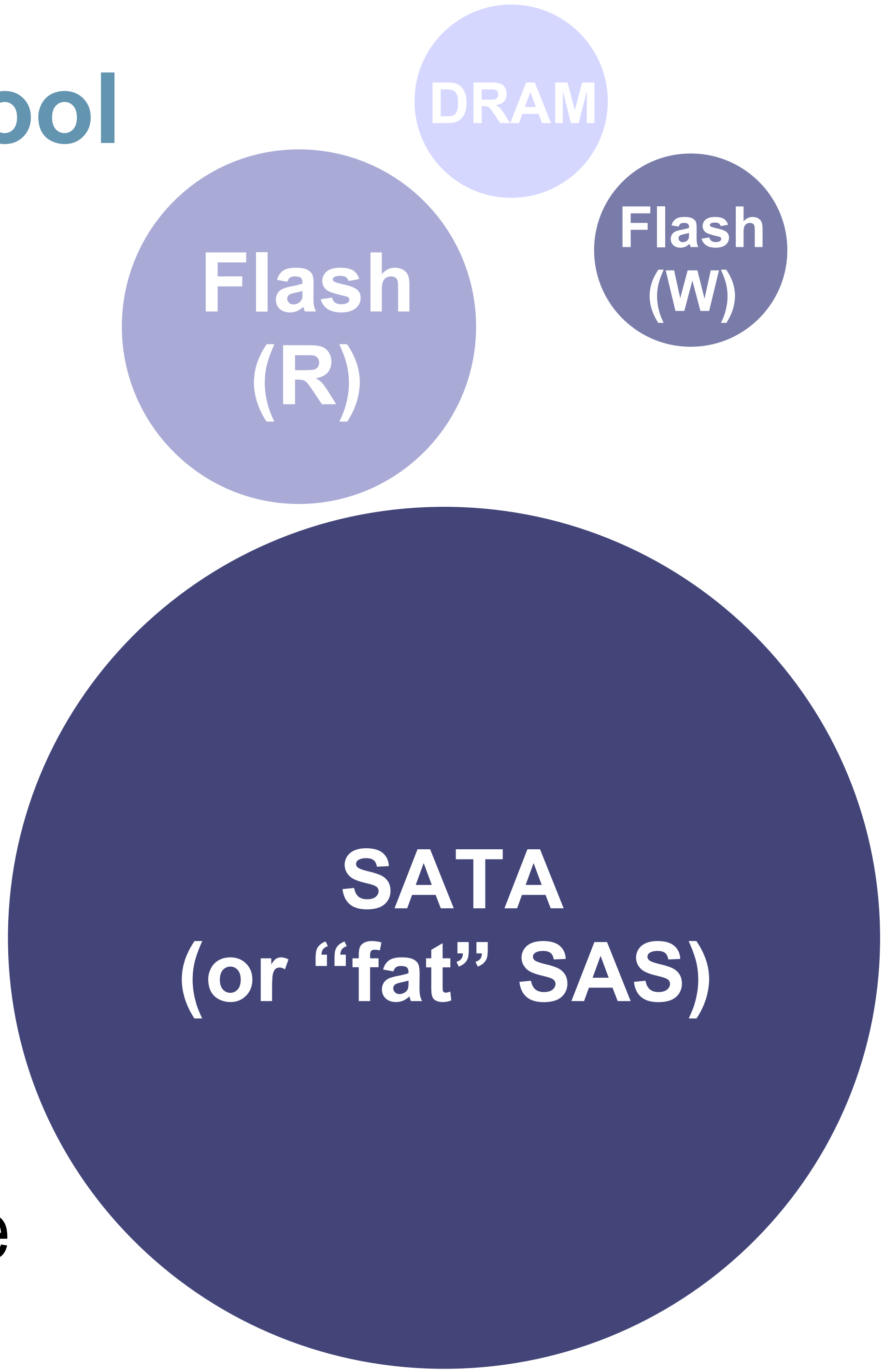


Sun's Flash Architecture

- Flash devices in 2.5” or 3.5” or mini-DIMM form factor
  - Writes buffered by DRAM backed by supercapacitors
  - Devices specifically optimized for read or write
  - Enterprise-grade: 1M w/erase cycles, 3-5 yr lifetime
- Used in place of NVRAM for ZFS Intent Log (ZIL)
  - Write-biased (aka “Logzilla”)
  - Arbitrary scale: put in JBOD instead of slot in head node
  - No battery issues for serviceability
  - First release will support unlimited 24G write-biased SSDs
- Used to extend the ZFS cache (ARCL2)
  - Read-biased (aka “Readzilla”)
  - Extends ZFS DRAM cache for reads and writes
  - First release will support up to 768G of read-biased SSD (!)

ZFS Hybrid Storage Pool

- Storage is transparently managed as a single pool with an optimized hierarchy
- ZFS understands how to leverage the attributes of each type of device and function
- Simple administrative policy knobs provide resource control
- Best $/IOP, $/G, Power/G
- Takes full advantage of the SAS/SATA reliability fallacy (see Google disk failure results)



ZFS Hybrid Pool Example
(announced at recent IDF Shanghai)

4 Xeon 7350 Processors (16 cores)
32GB FB DDR2 ECC DRAM
OpenSolaris with ZFS

(1) 32G SSD ZIL Device

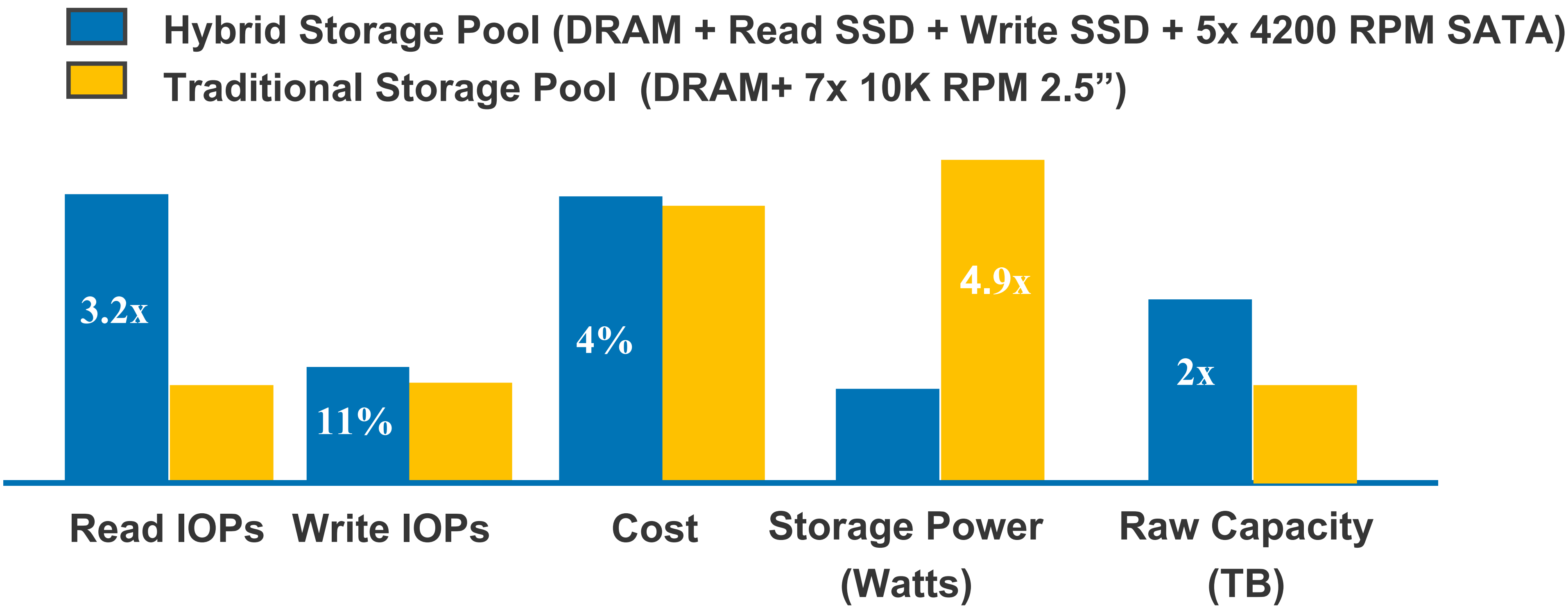
(1) 80G SSD Cache Device

(7) 146GB 10,000 RPM SAS Drives

(5) 400GB 4200 RPM SATA Drives

ZFS Hybrid Pool Example

(announced at recent IDF Shanghai)



- If NVRAM were used, Hybrid wins on cost too
- For large configs (e.g. 48T–750T+) cost is entirely amortized

ZFS: The First I/O Stack Optimized for Flash