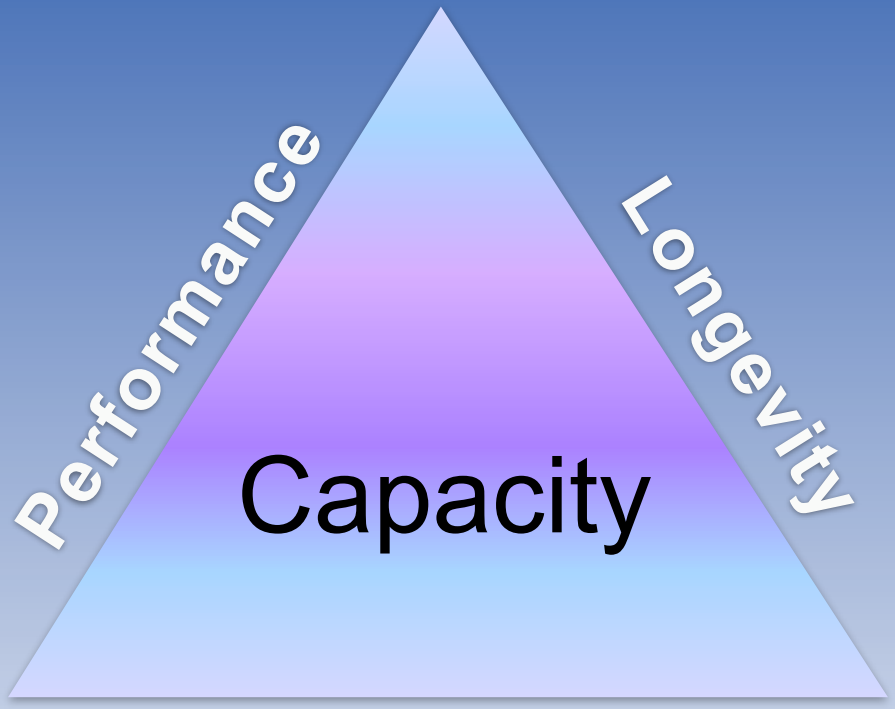
An In-Depth Examination of the Workings of an Enterprise-Class SSD

David Flynn, CTO Fusion-io

Enterprise-Class SSD Design

- Basic dimensions
  - Reliability & Data Integrity
  - Capacity
  - Performance
  - Longevity
- For each discuss...
  - Metrics
  - Raw Media capabilities (today & tomorrow)
  - Integration approaches (pros & cons)
- Scalability

Reliability cannot be compromised



Other requirements vary by workload

Raw Media Reliability

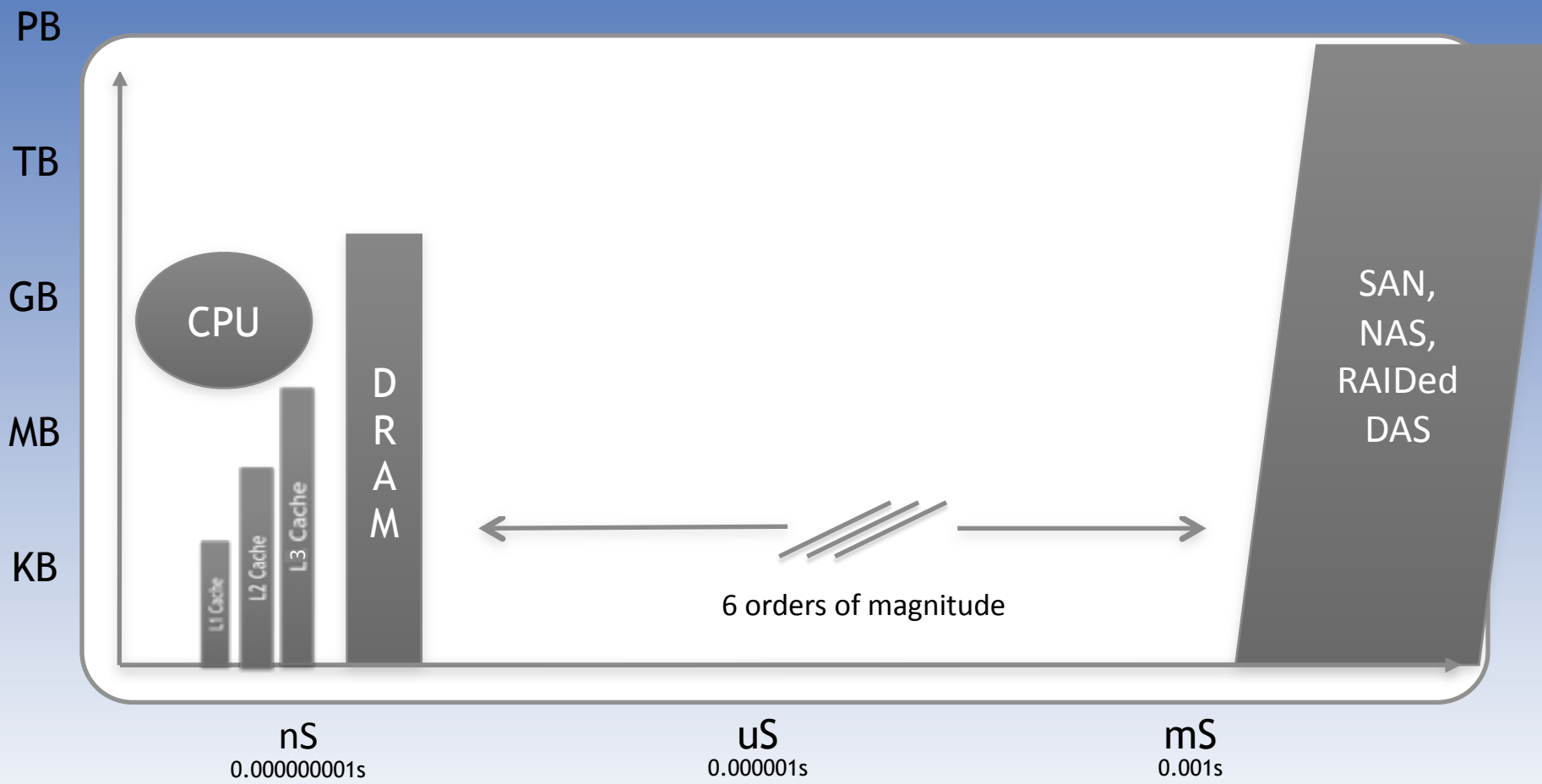
- **The GOOD**
  - No moving parts
  - Post infant mortality catastrophic device failures are rare
  - Predictable wear out
- **The BAD**
  - Relatively high bit error rate, which increases with wear
  - Higher density and MLC increases bit error rate
  - Program and Read Disturbs
- **The UGLY**
  - Partial Page Programming
  - Data retention is poor at high temperature and wear
  - Infant mortality is high (large number of parts…)

Controller Reliability Management

- **In-Flight**
  - Corruption upstream disk controllers
  - Corruption in SSD controller itself
  - Flush at power loss
- **At-Rest**
  - ECC
  - Scanning & scrubbing
  - Redundancy
- **Meta-data**
  - Error correcting memory
  - Data integrity field

Poor Media + Great Controller = Great SSS Solution

Capacity Performance Relationship



Access delay in time

Performance is about ROI

Lower CapEx

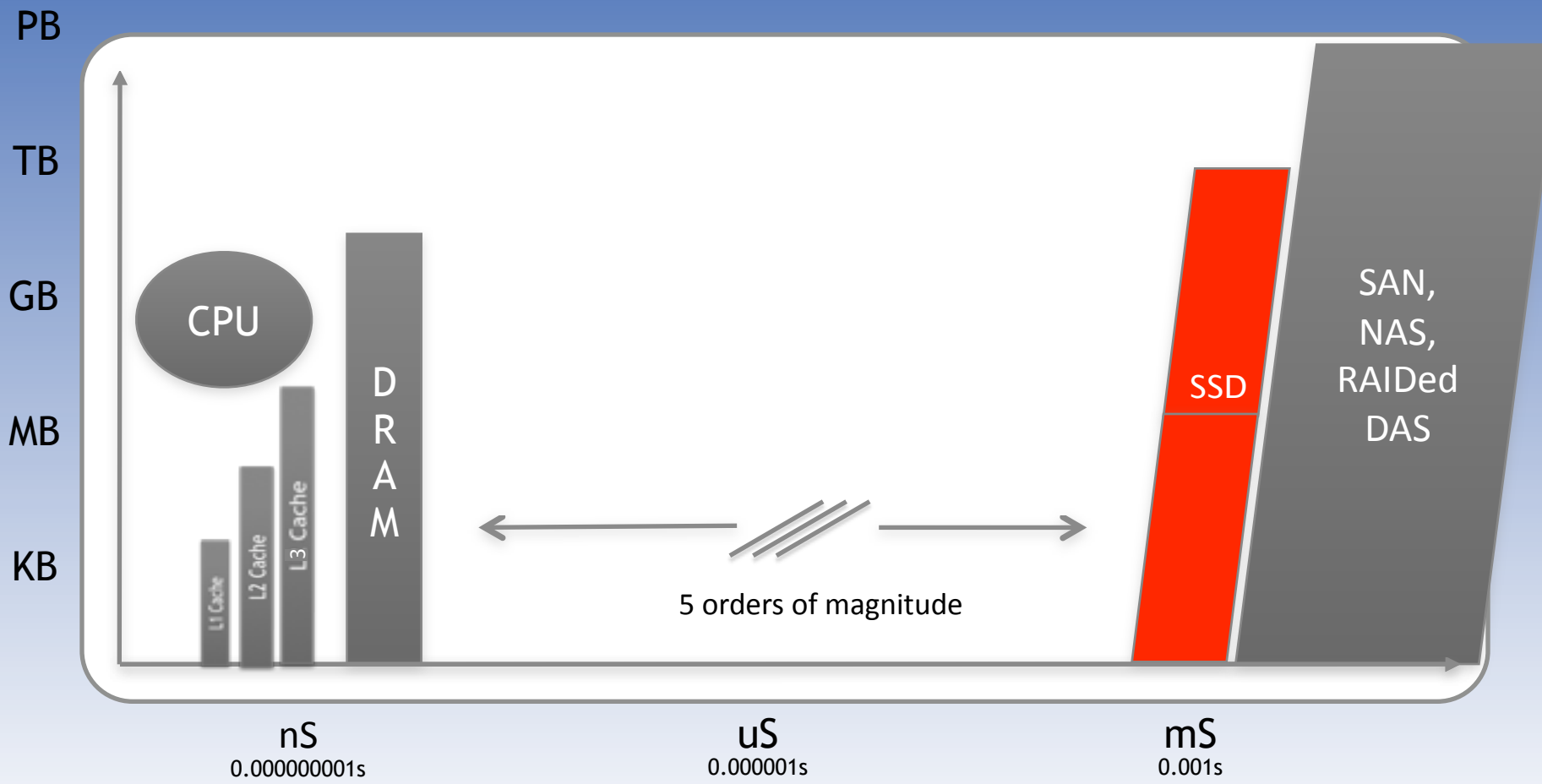
- Fewer CPUs
- Less RAM
- Less Network Gear
- Fewer SW Licenses
- Less Space

Lower OpEx

- Less HW Maintenance
- Less SW Maintenance
- Greater Uptime
- Less Power/Cooling
- Fewer Diverse Skills

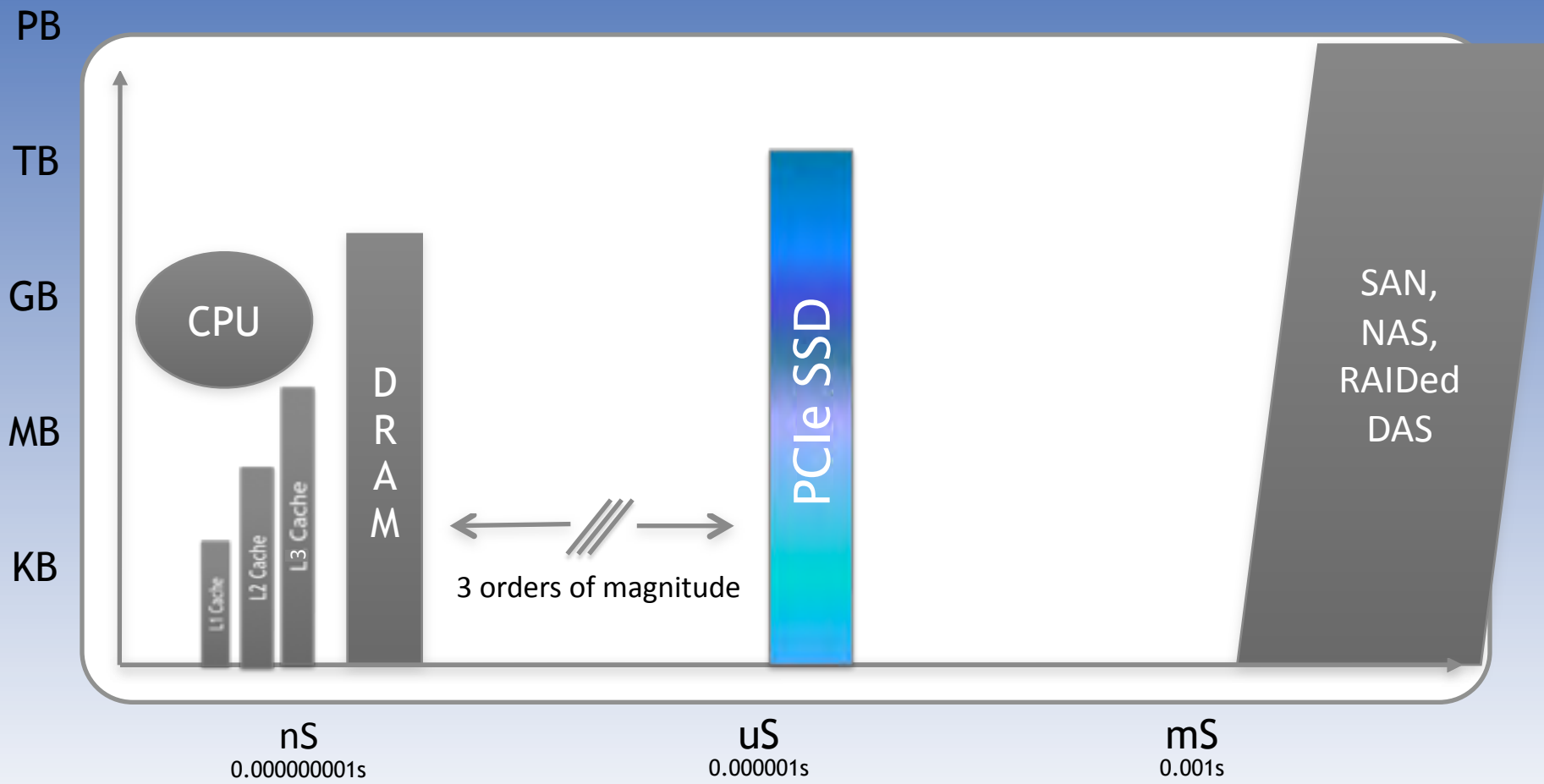
HIGHER Productivity

Traditional SSD's



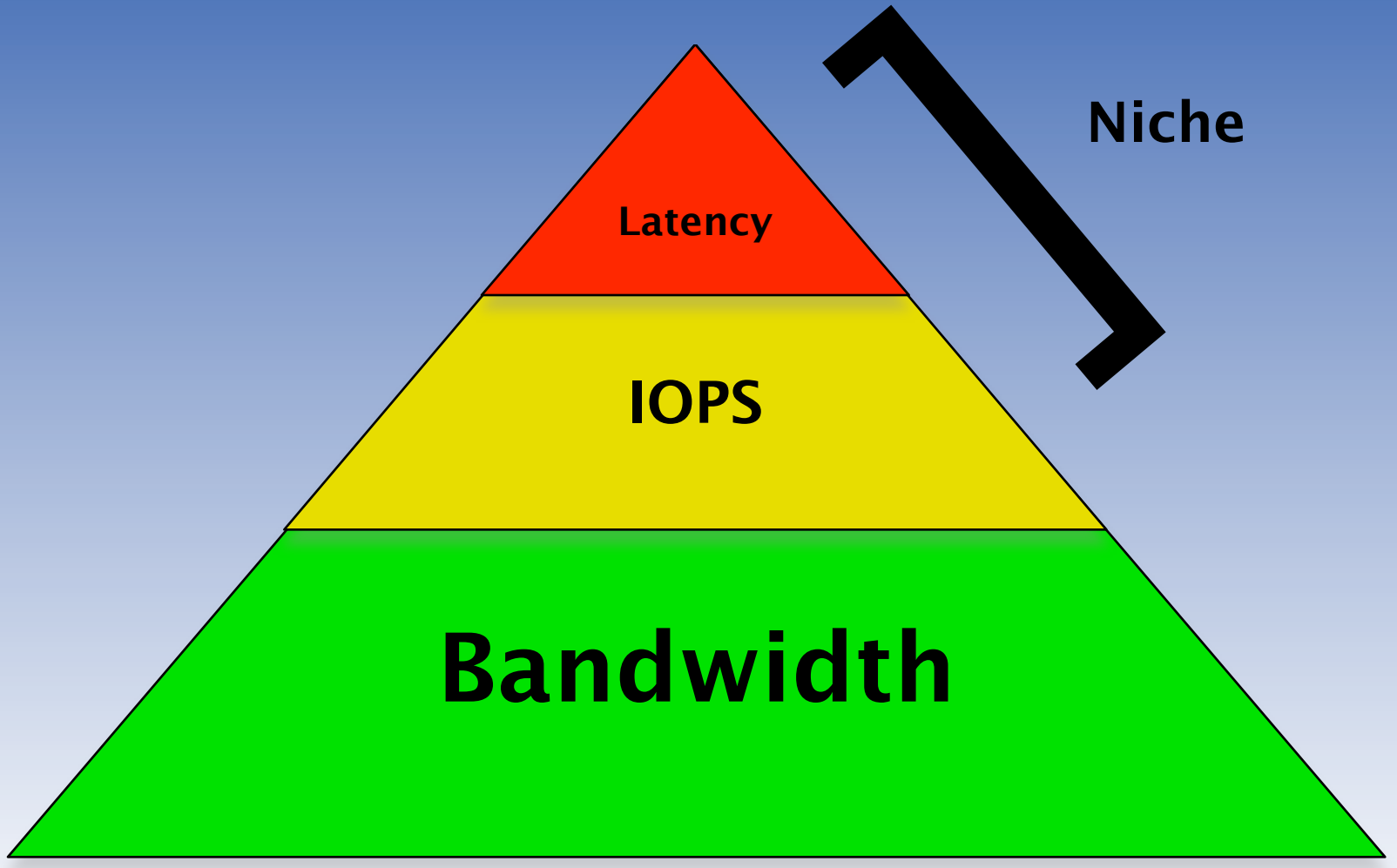
Access delay in time

PCIe Attached SSD's

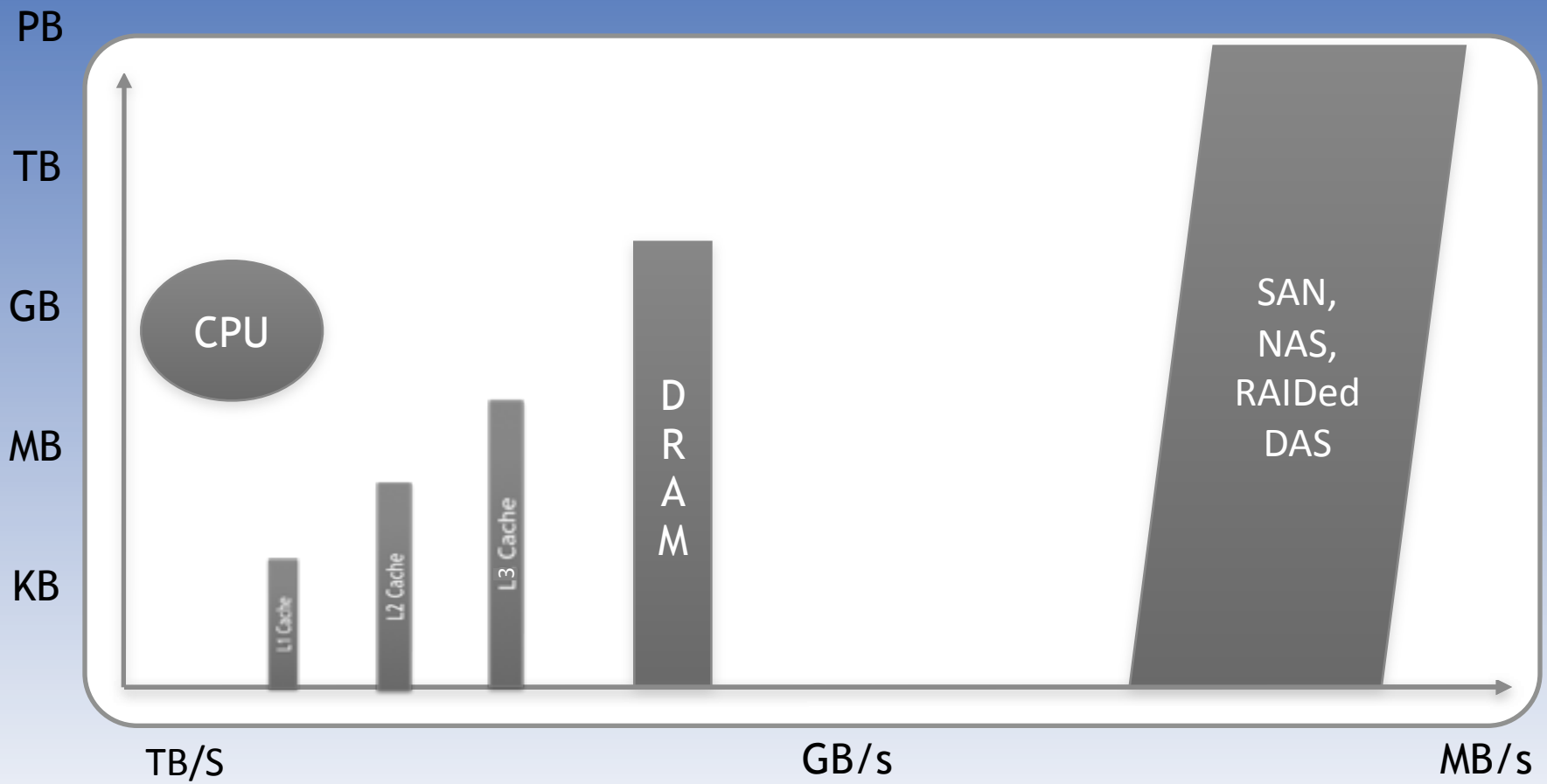


Access delay in time

Performance Dimensions

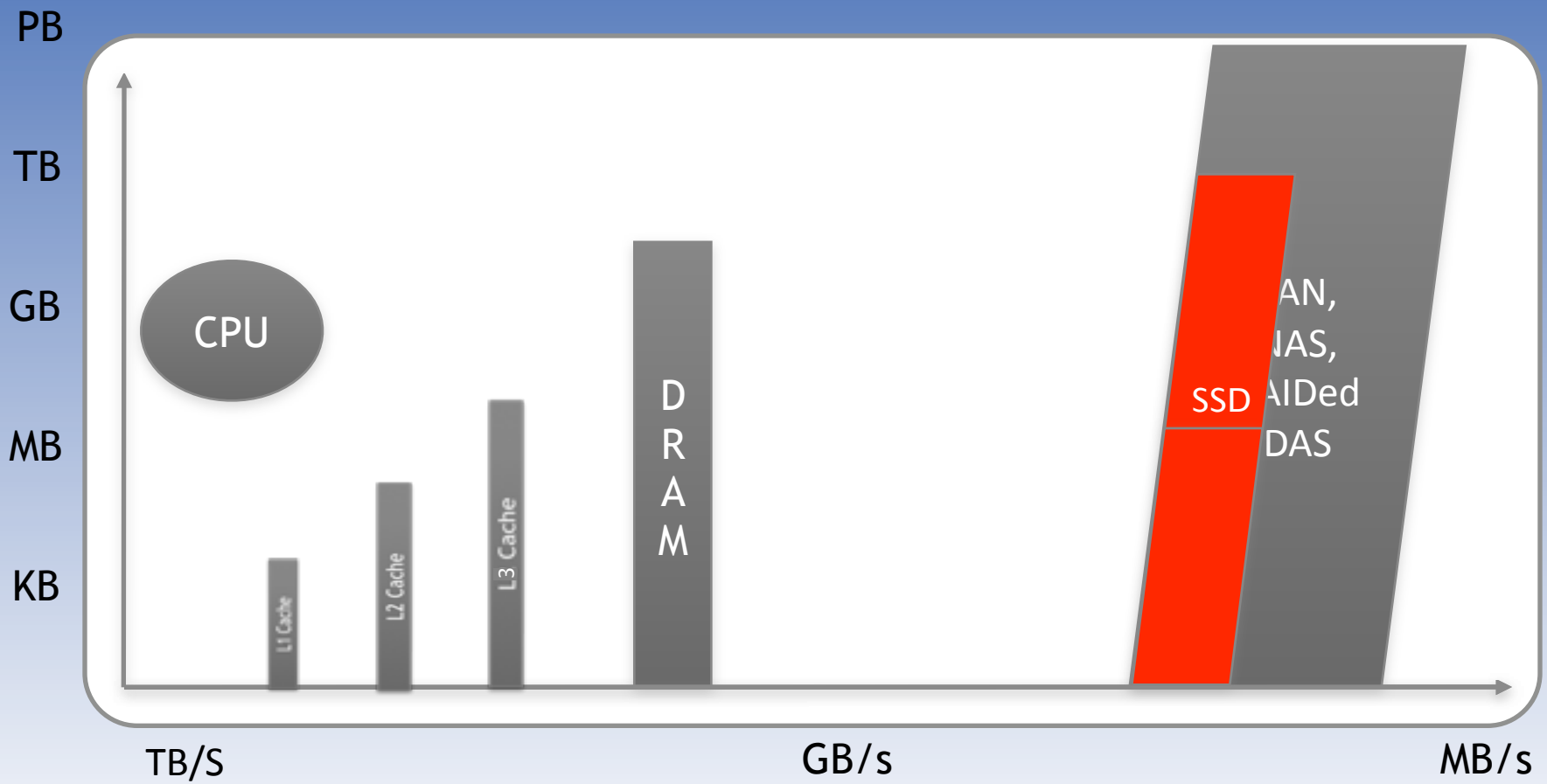


Access Bandwidth



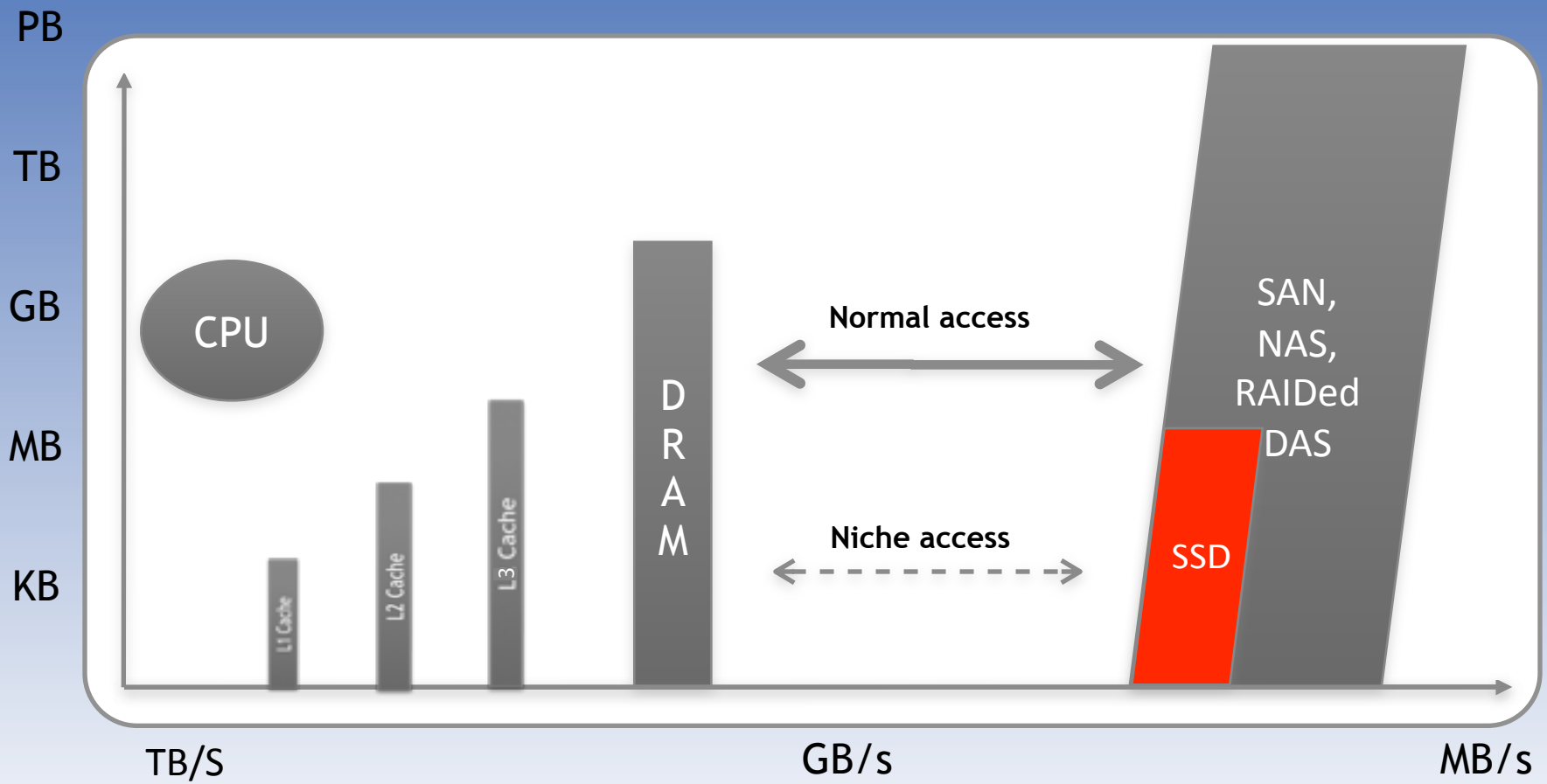
Access Bandwidth

Traditional SSD's are no better



Access Bandwidth

Workload Segregation



Access Bandwidth

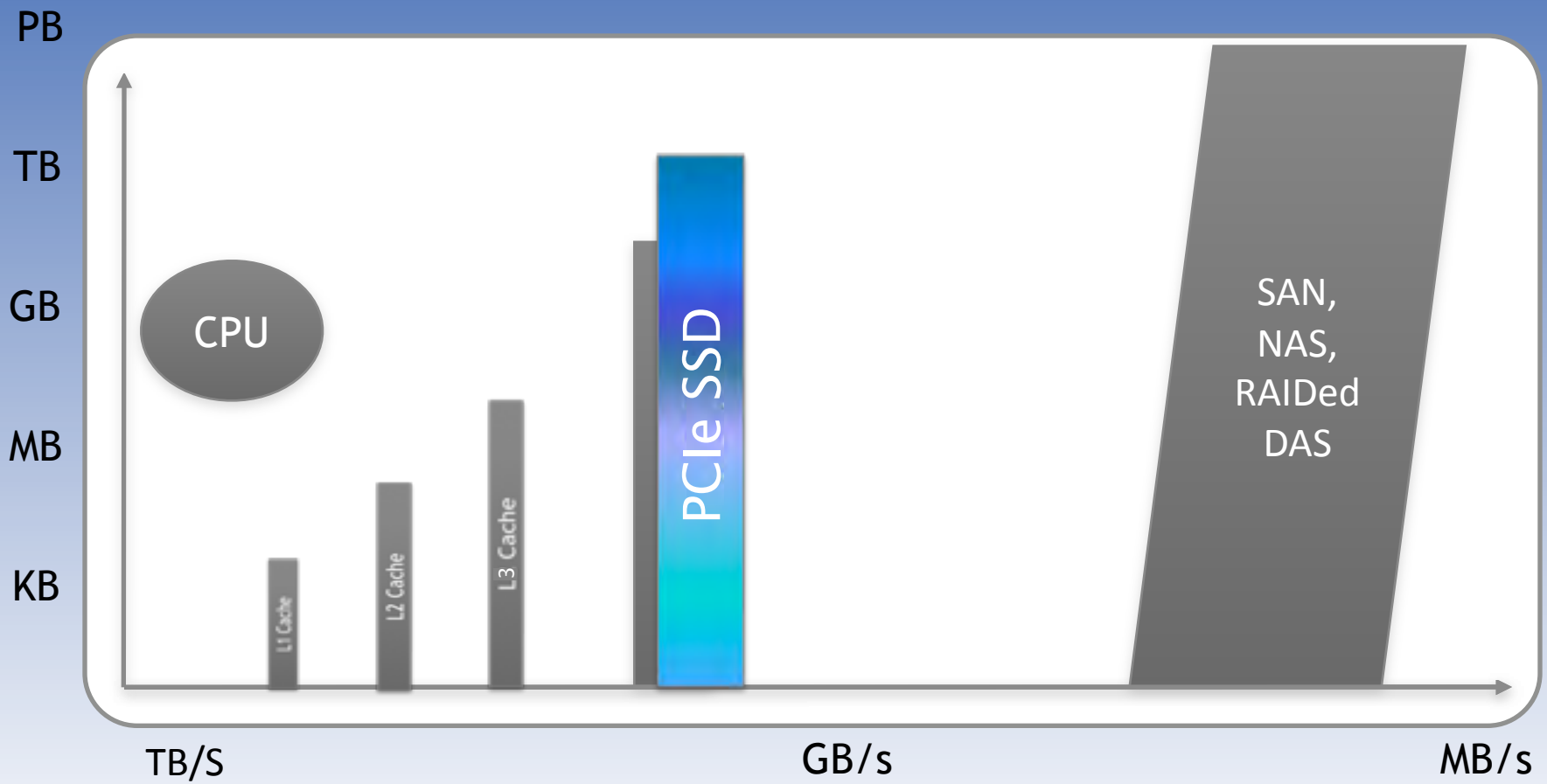
A cache needs…

- Bandwidth
- Mixed reads and writes
- Writes while full (saturated)

That's exactly what SSD's suck at!

(well traditional ones anyway)

PCIe SSD's are more like DRAM



Access Bandwidth

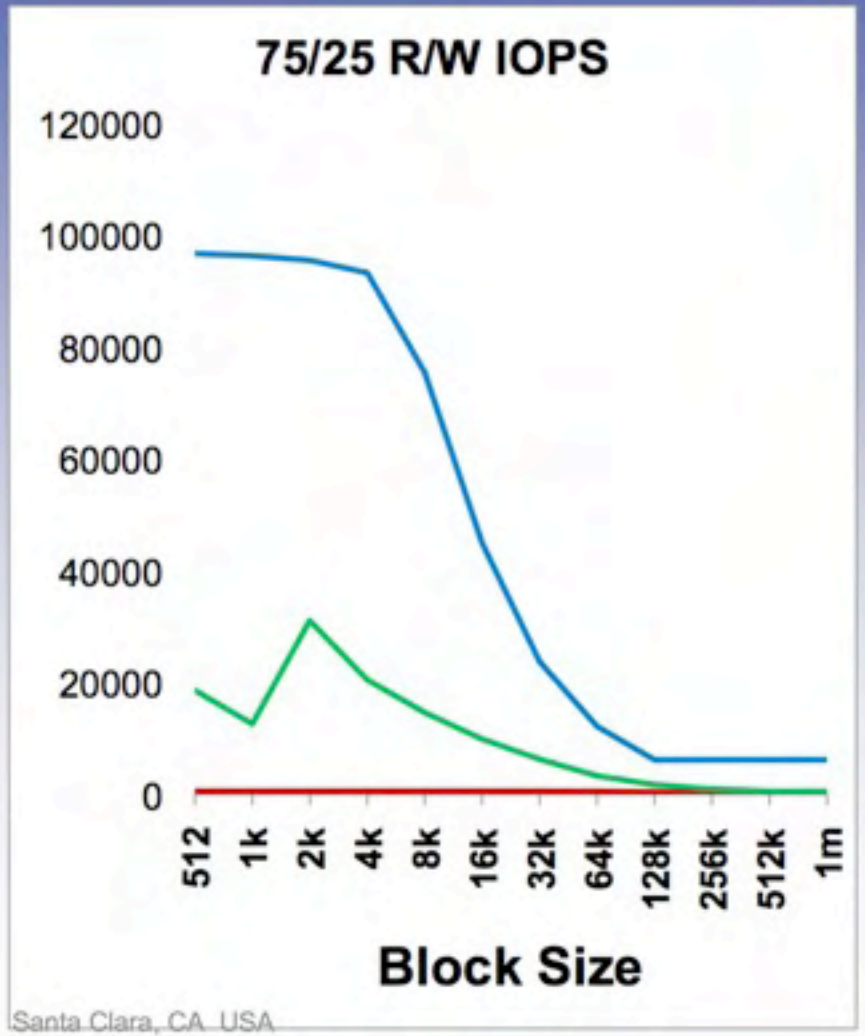
Raw Media Performance

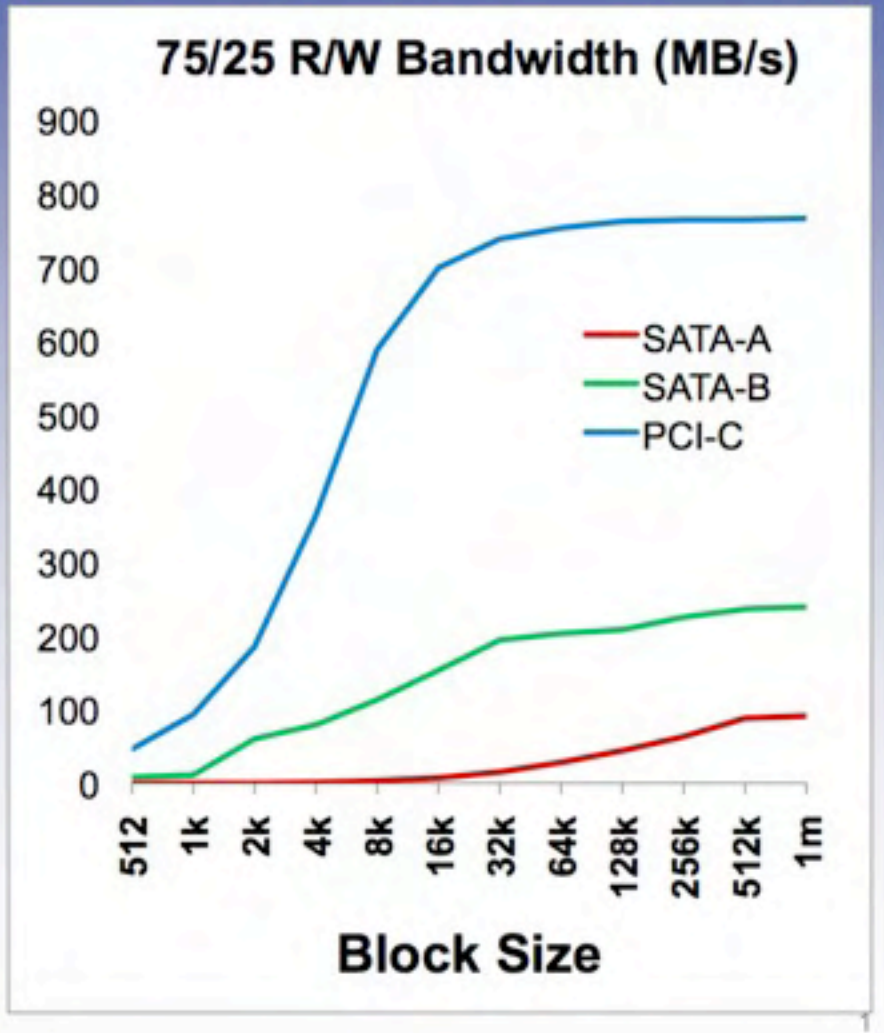
- The GOOD
  - Performance is excellent (wrt HDDs)
  - High performance per power (IOPS/Watt)
  - Low pin count: shared command / data bus → good balance
- The BAD
  - Not really a random access device
    - Block oriented
    - Slow effective write (erase/transfer/program) latency
    - R/W access speed imbalance
  - Performance changes with wear
- The UGLY
  - Some controllers do read/erase/modify/write
  - Others use inefficient garbage collection

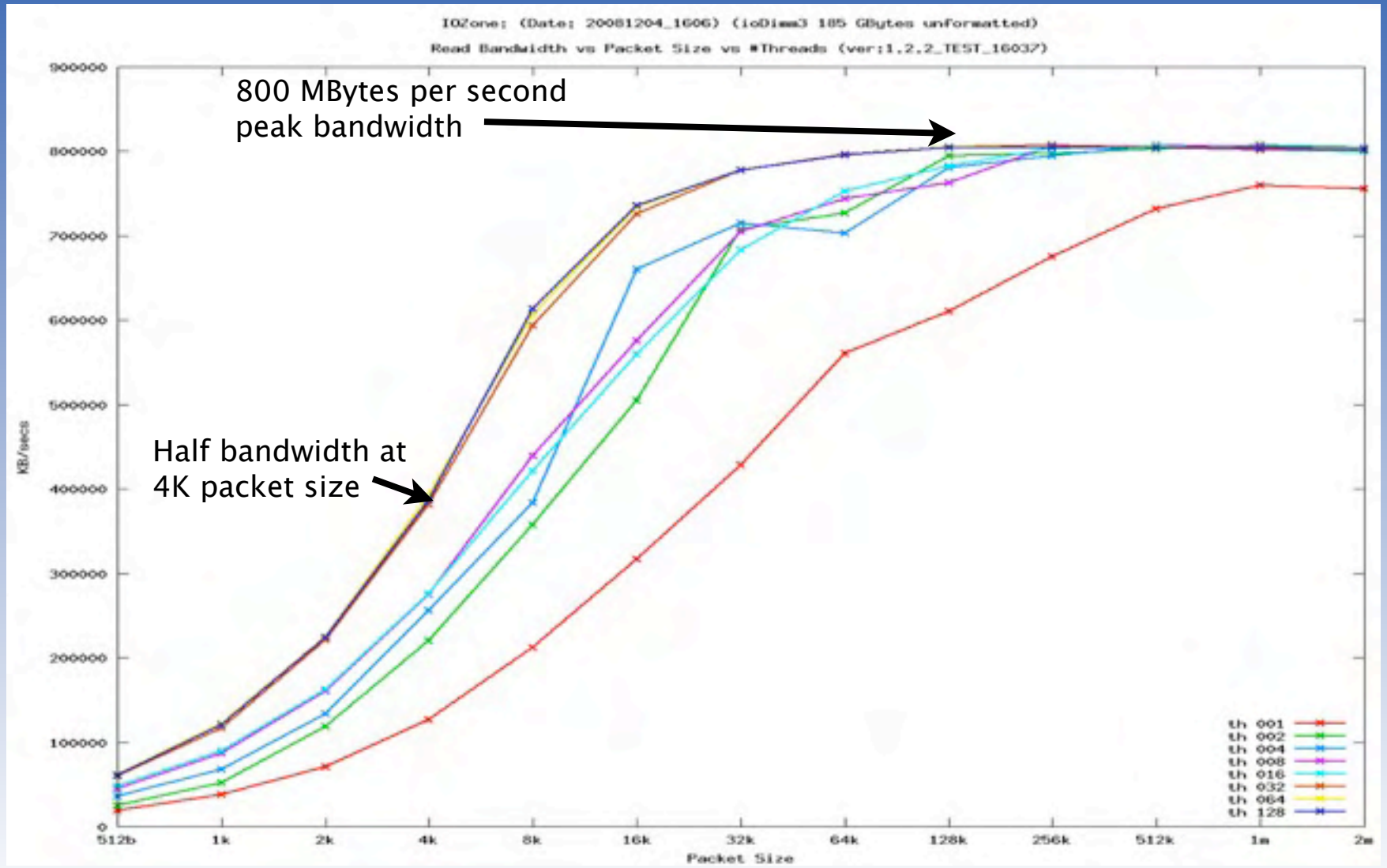
Controller Performance Drivers

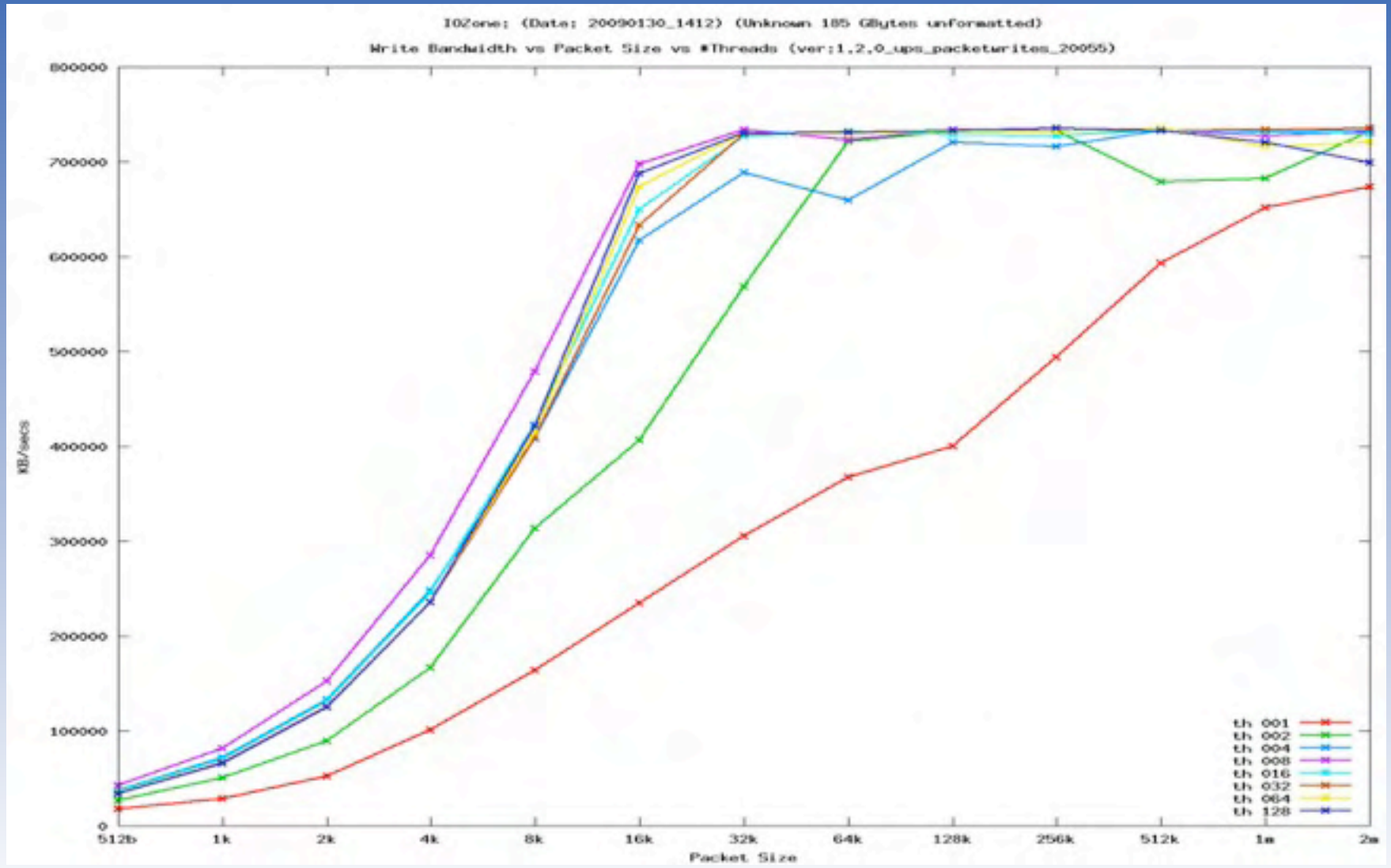
- Interconnect
- Number of NAND Flash Chips (Die)
- Number of Buses (Real / Pipelined)
- Data Protection (internal/external RAID; DIF; ECC…)
- SLC / MLC
- Effective Block (LBA; Sector) Size
- Write Amplification
- Garbage Collection (GC) Efficiency
- Buffer Capacity & Mgmt
- Meta-data processing

Performance vs Block Size (75/25)





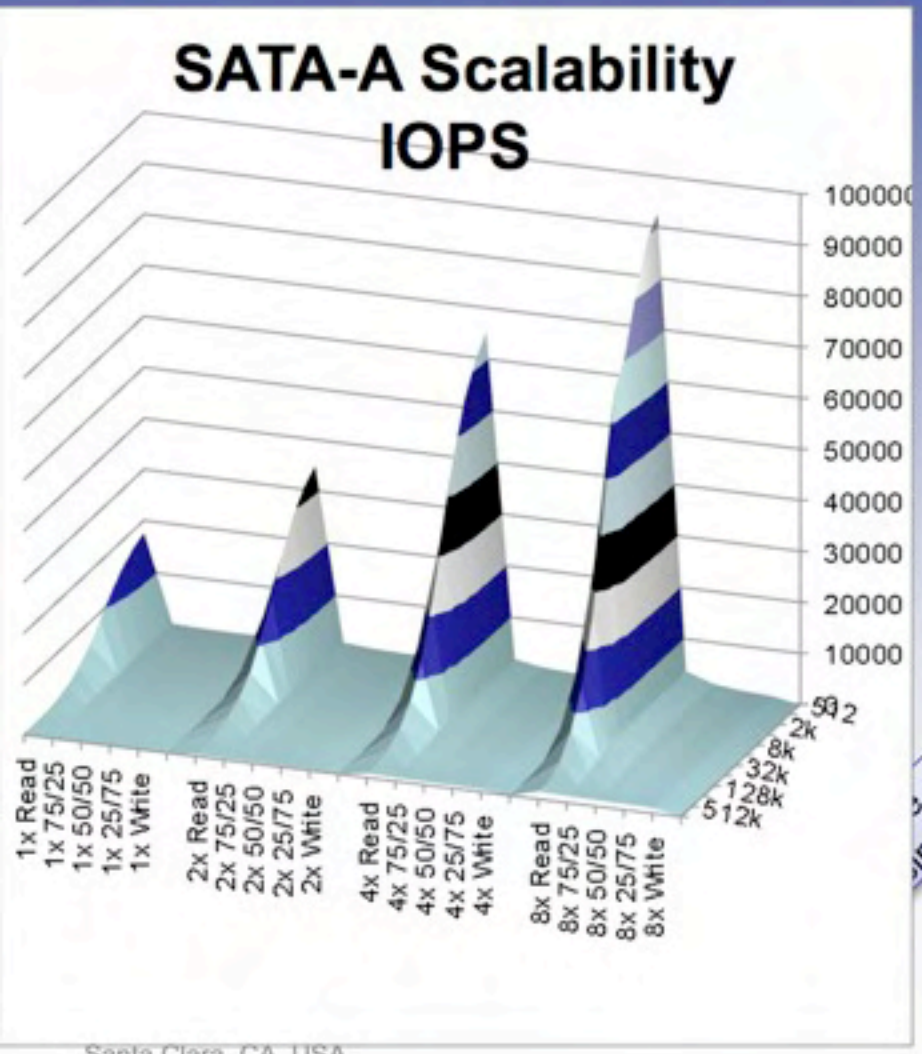

IOZone: (Date: 20081204_1606) (ioDimm3 185 GBytes unformatted)
Read Bandwidth vs Packet Size vs #Threads (ver:1.2.2_TEST_16037)
800 MBytes per second
peak bandwidth
Half bandwidth at
4K packet size
KB/secs
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
512b
1k
2k
4k
8k
16k
32k
64k
128k
256k
512k
1m
2m
Packet Size
th 001
th 002
th 004
th 008
th 016
th 032
th 064
th 128

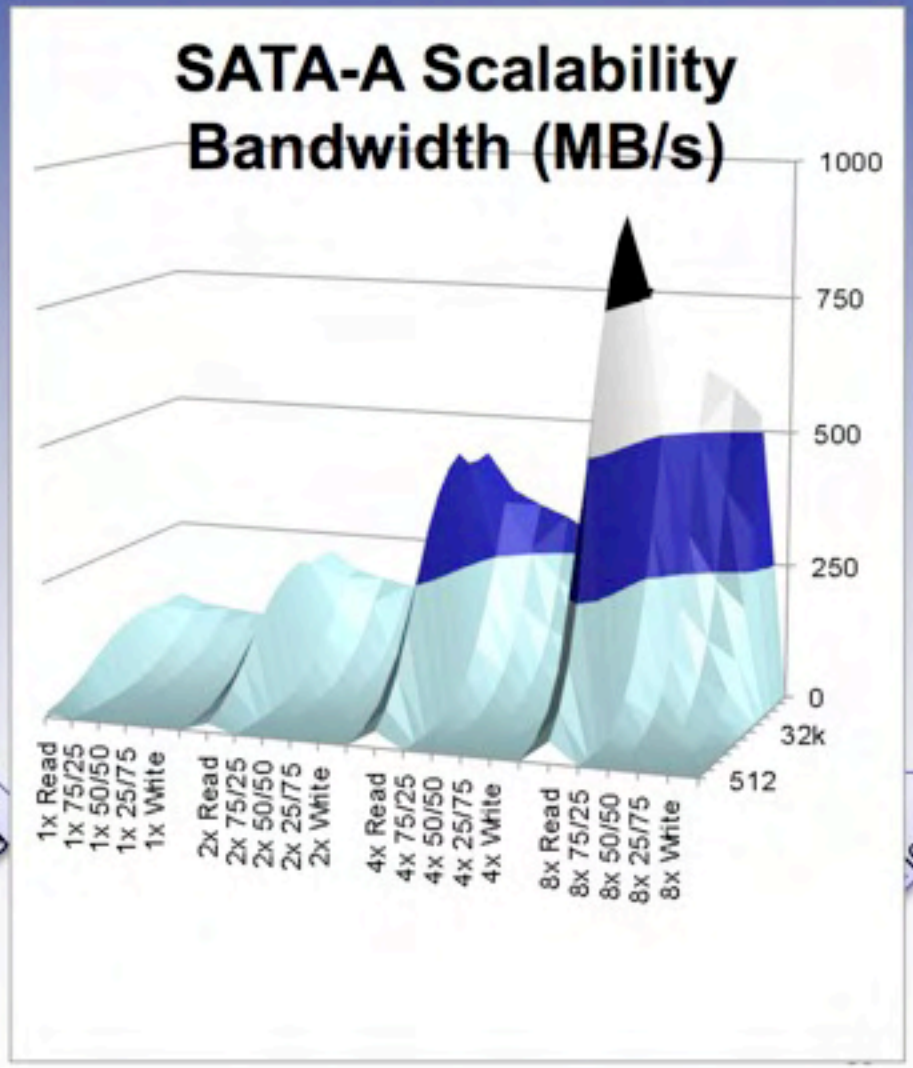

IOZone: (Date: 20090130_1412) (Unknown 185 GBytes unformatted)
Write Bandwidth vs Packet Size vs #Threads (ver:1,2,0_ups_packetwrites_20055)
KB/secs
800000
700000
600000
500000
400000
300000
200000
100000
0
512b
1k
2k
4k
8k
16k
32k
64k
128k
256k
512k
1m
2m
Packet Size
th 001
th 002
th 004
th 008
th 016
th 032
th 064
th 128

Scalability

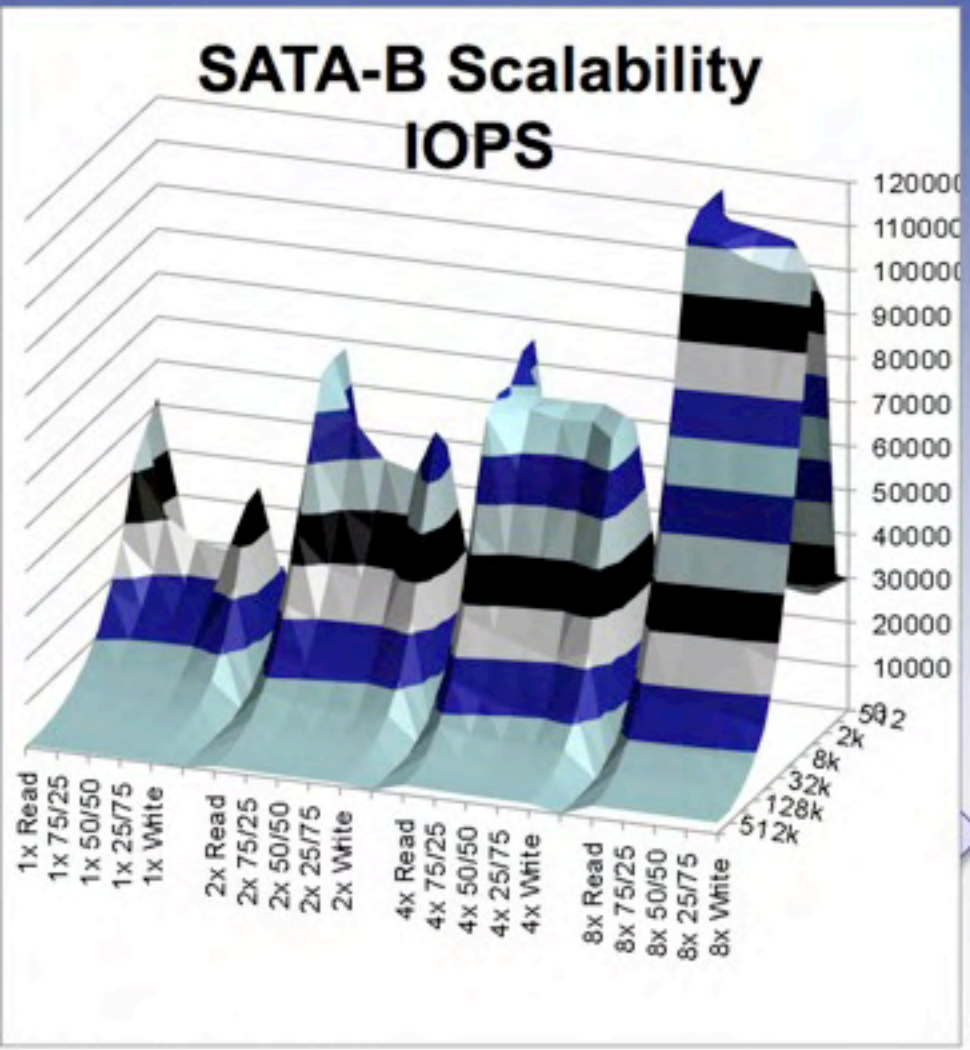
- Following Slides Show
  - Scalability of {1, 2, 4, 8} units
  - Only 1 SATA controller is used – limiting scalability
  - Only 1 thread running
- Measurements taken at Read/Write Ratios of
  - {100/0, 75/25, 50/50, 25/70, 0/100}
  - RMS value is the “root mean square” of these values
- IOPS measurement taken at 512 Byte Transfers
- Bandwidth taken at 128K Byte Transfers
  - Unless shown differently
  - Linux has a 128K limit

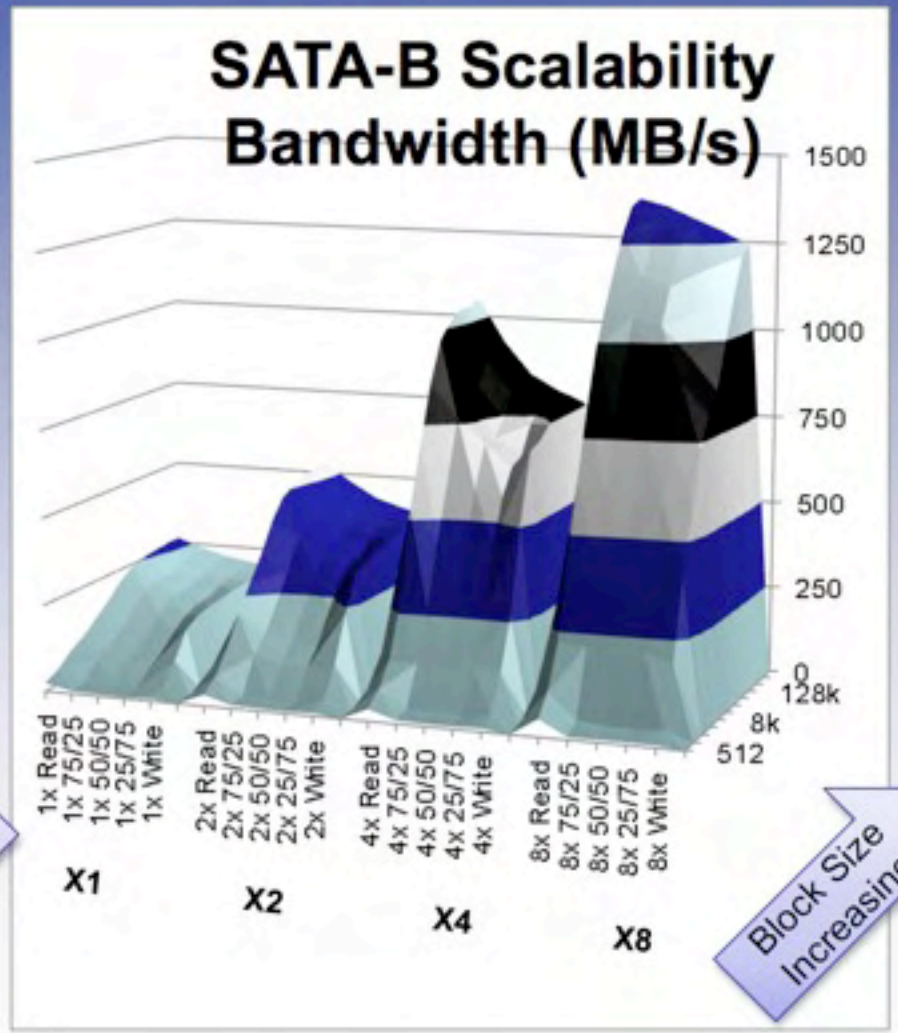
Scalability vs RW Ratio vs Block Size



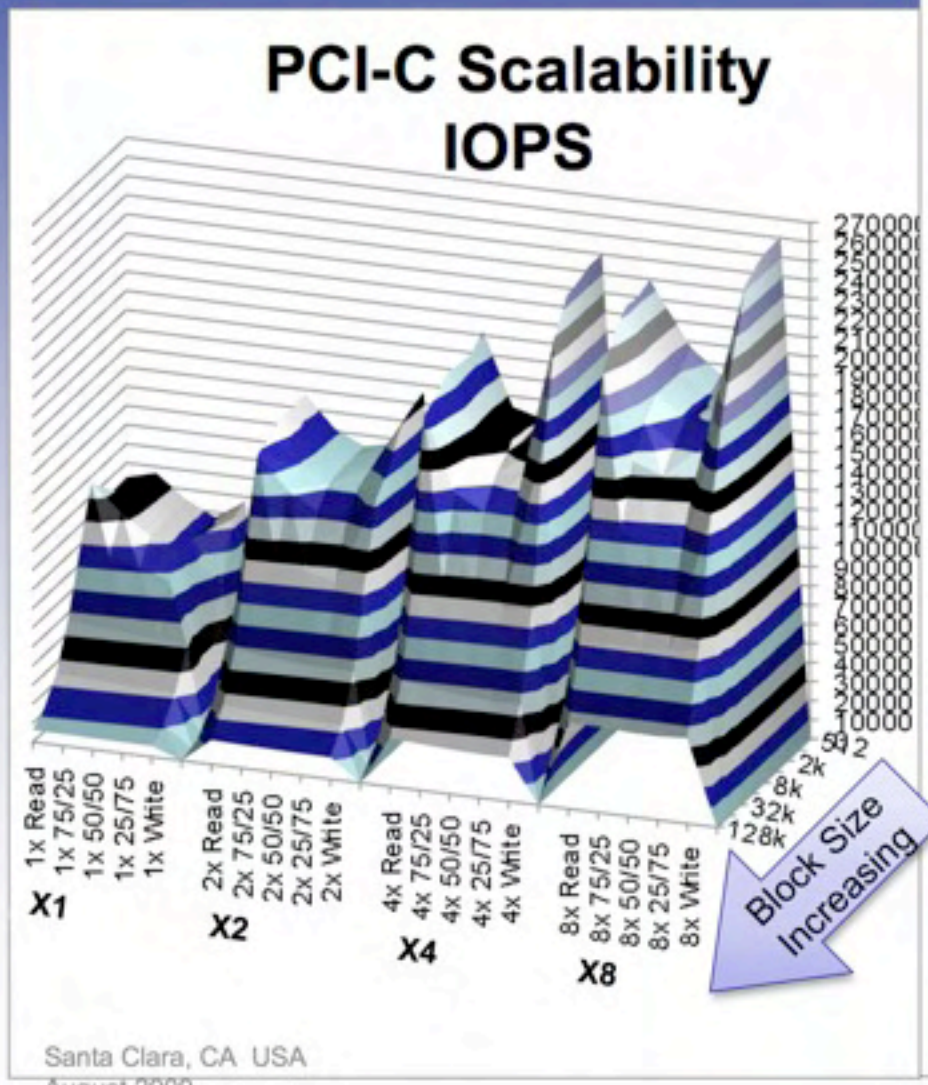


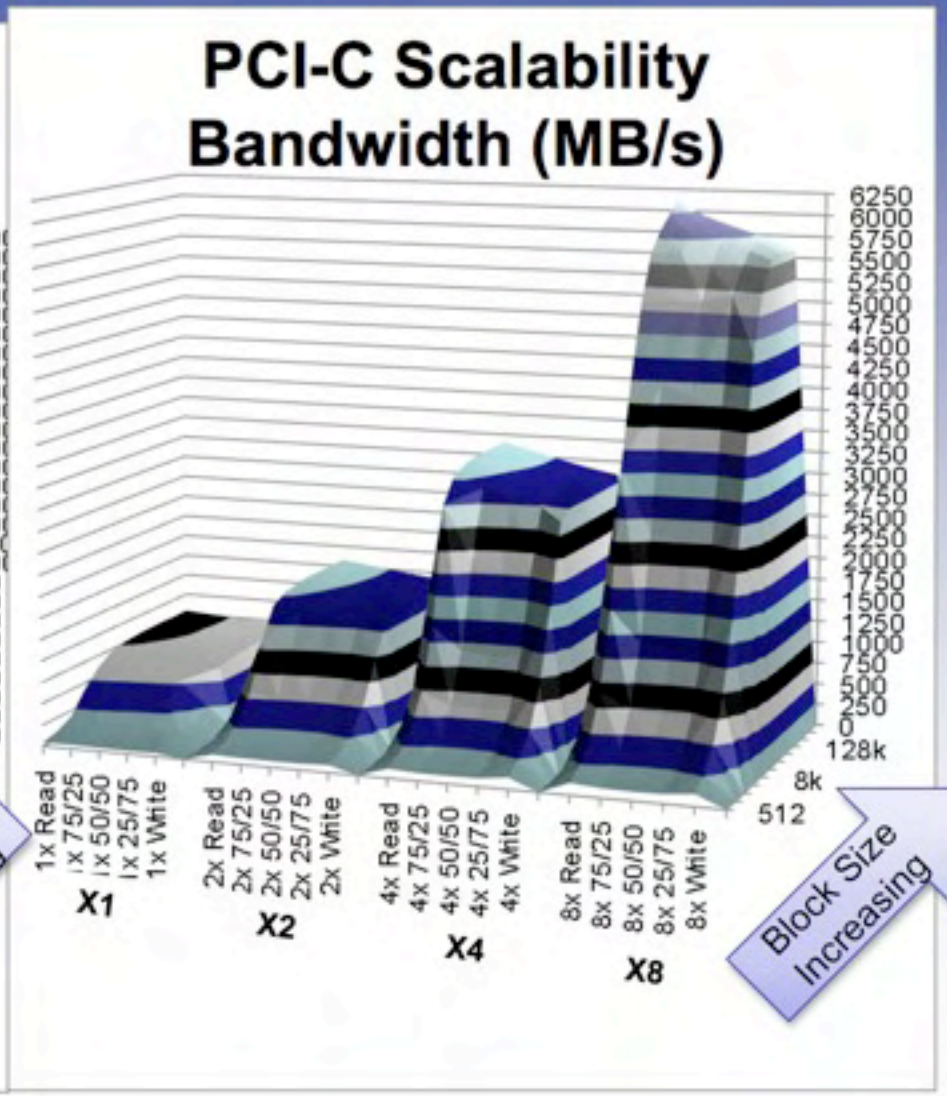
Scalability vs RW Ratio vs Block Size

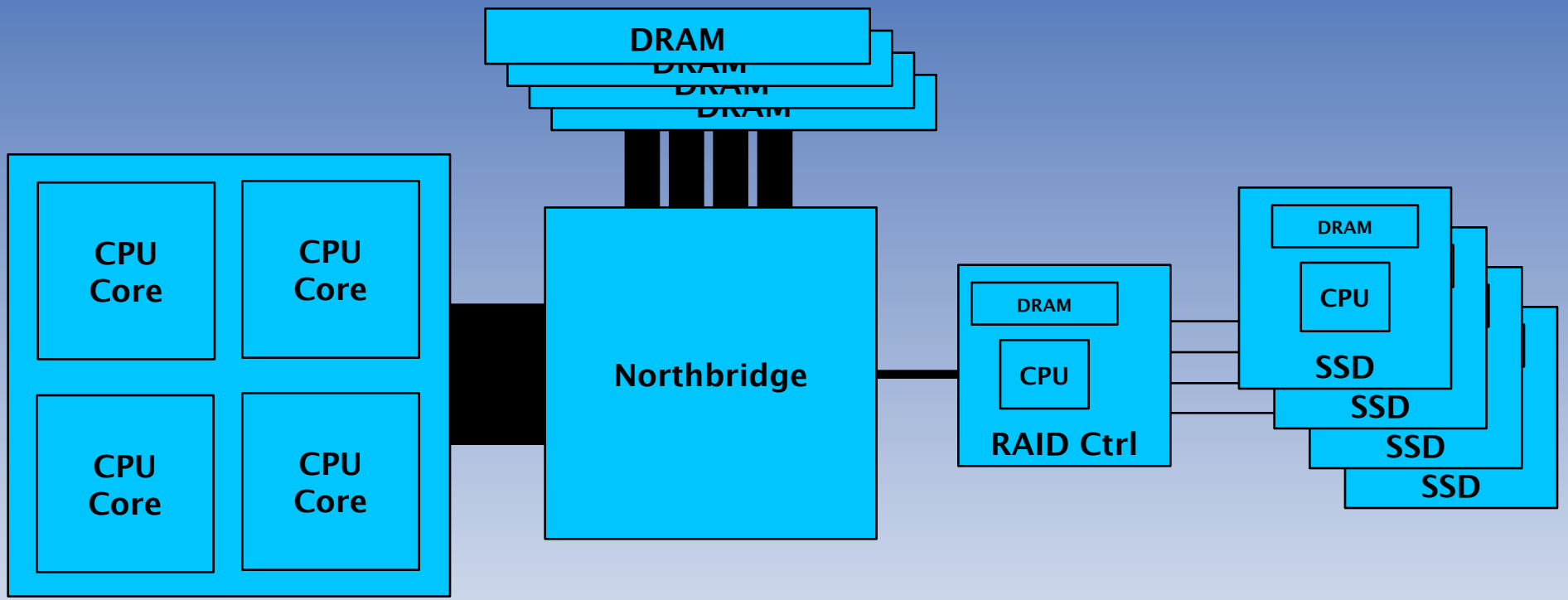


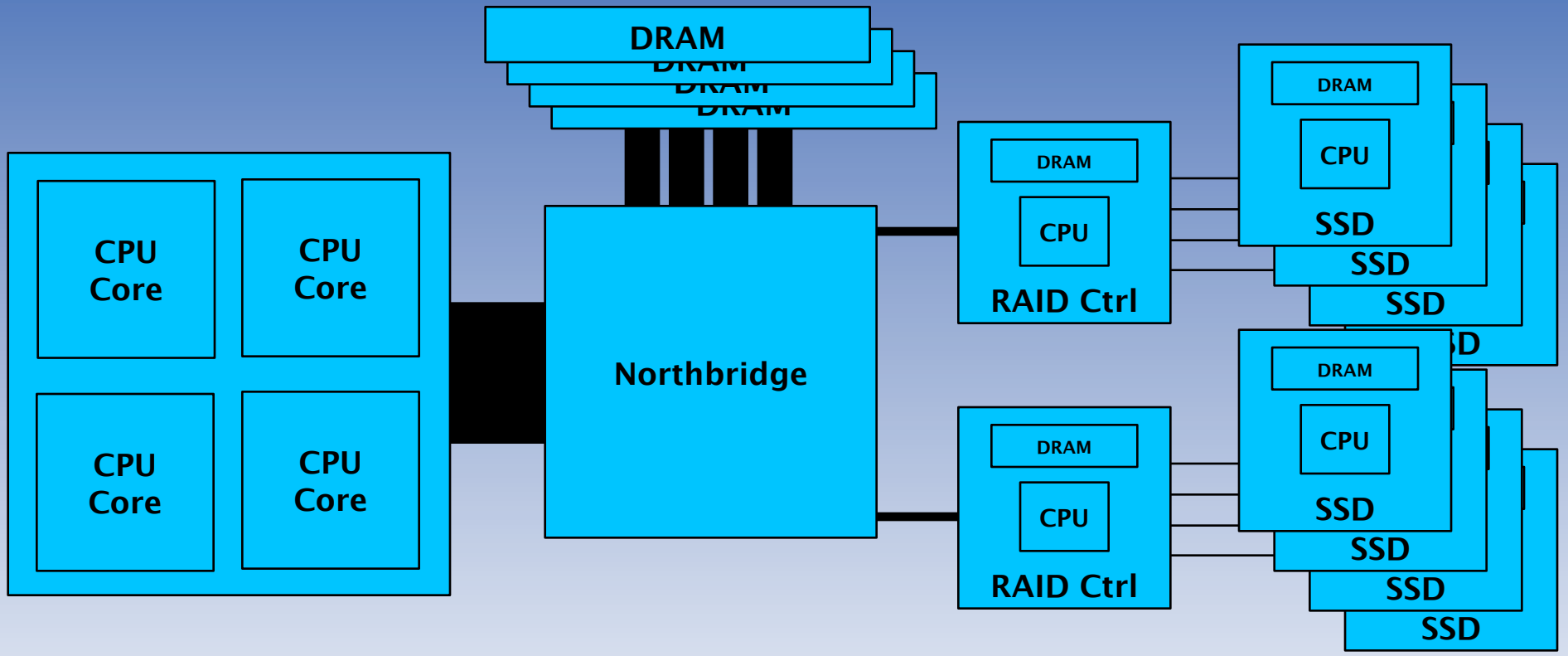


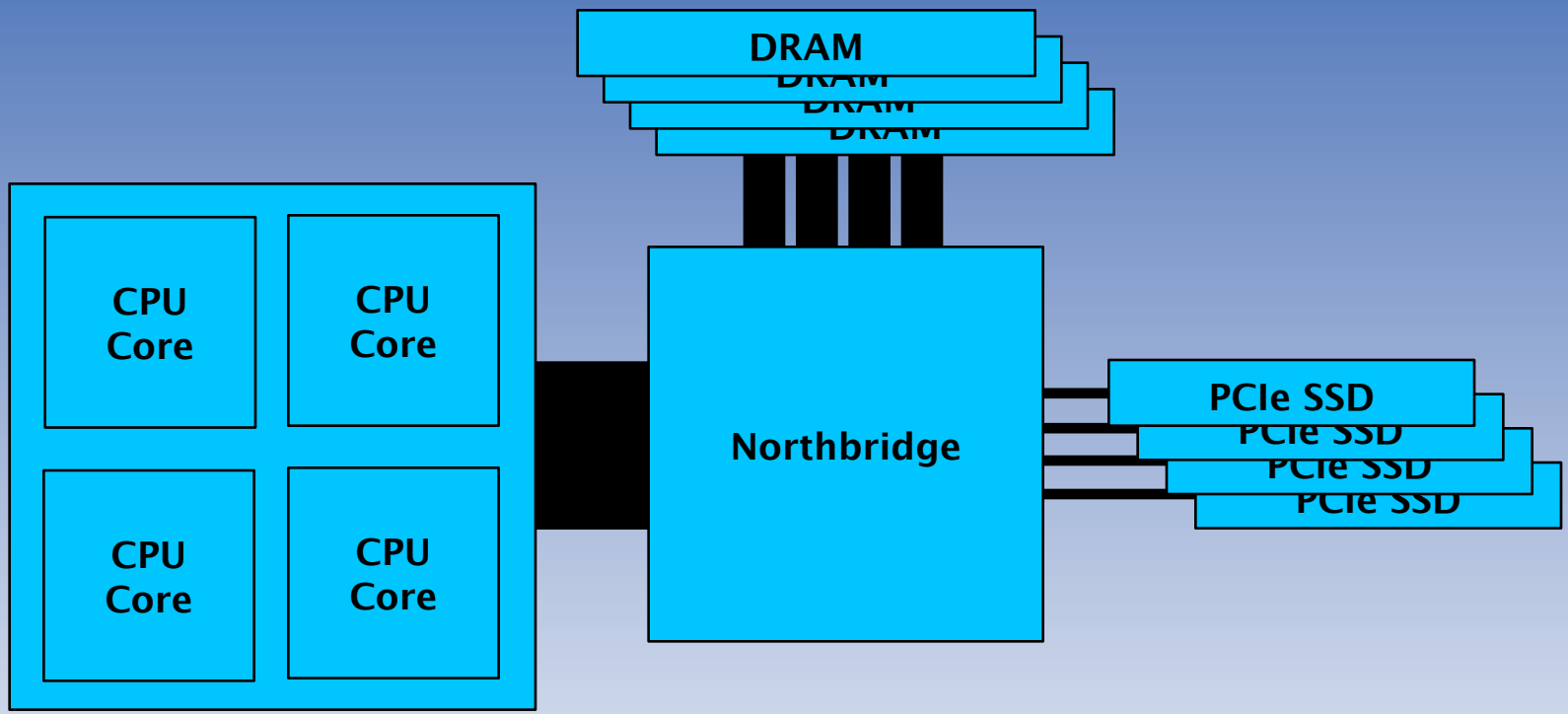
Scalability vs RW Ratio vs Block Size

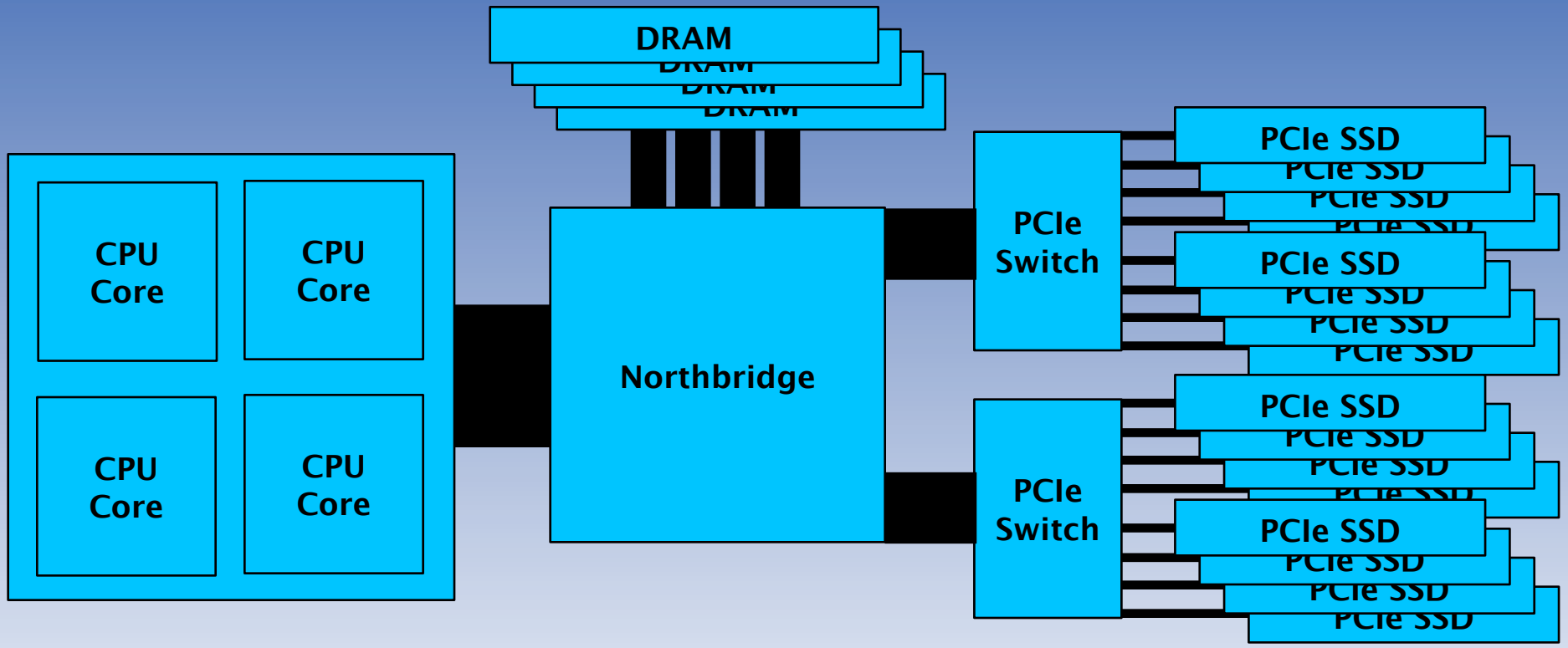





DRAM
CPU Core
CPU Core
CPU Core
CPU Core
Northbridge
DRAM
CPU
RAID Ctrl
DRAM
CPU
SSD
SSD
SSD
SSD


DRAM
CPU Core
CPU Core
CPU Core
CPU Core
Northbridge
DRAM
CPU
RAID Ctrl
DRAM
CPU
SSD
SSD
SSD
DRAM
CPU
RAID Ctrl
DRAM
CPU
SSD
SSD
SSD
SSD


DRAM
CPU Core
CPU Core
CPU Core
CPU Core
Northbridge
PCIe SSD



Thank you!