



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

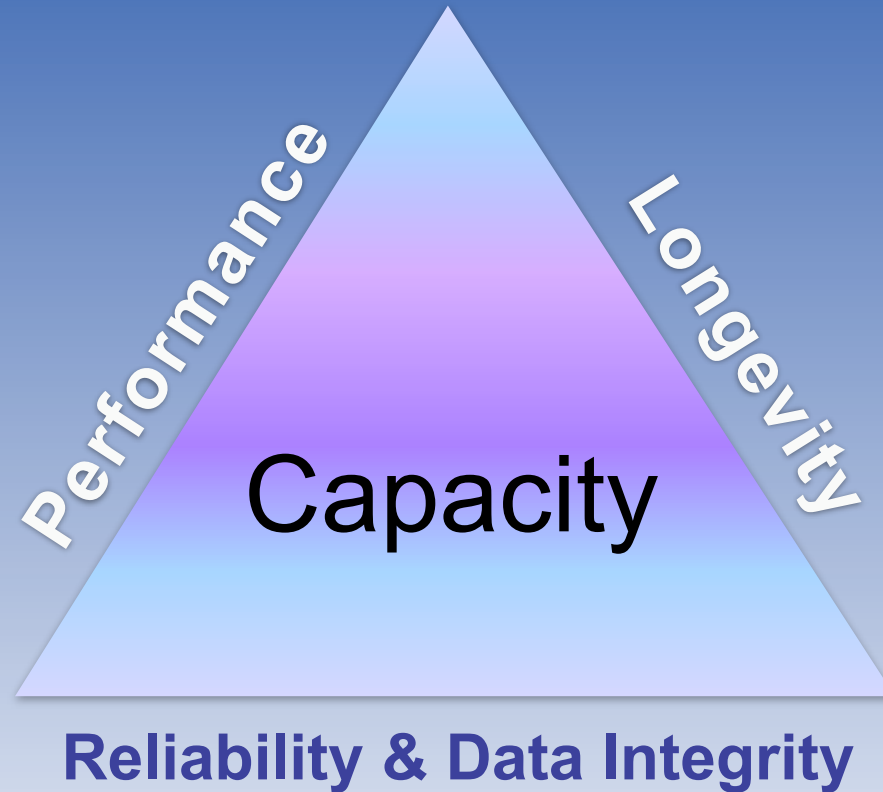
David Flynn, CTO Fusion-io



# 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...)

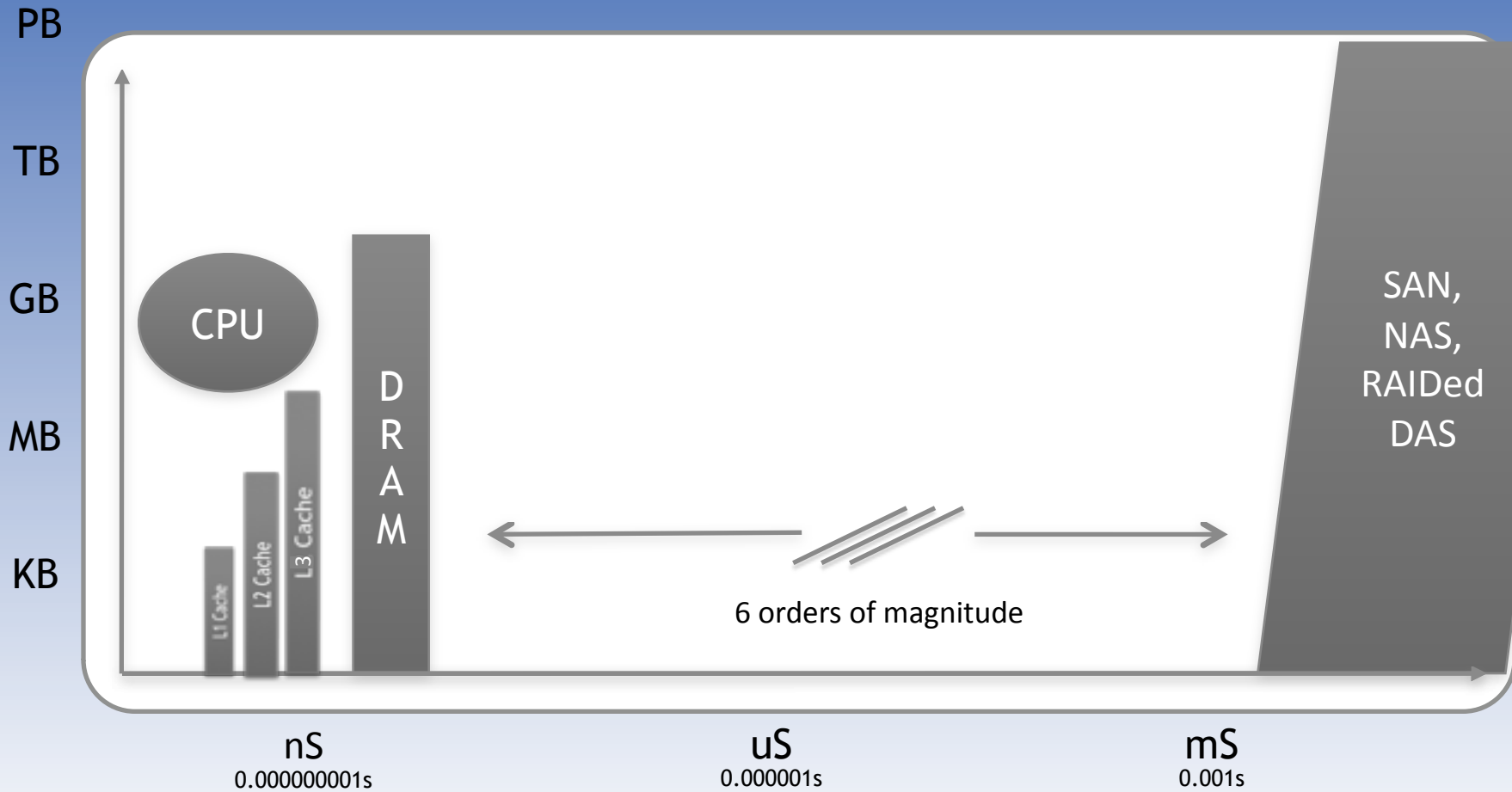


# Controller Reliability Management

- **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



# Performance is about ROI

## Lower CapEx

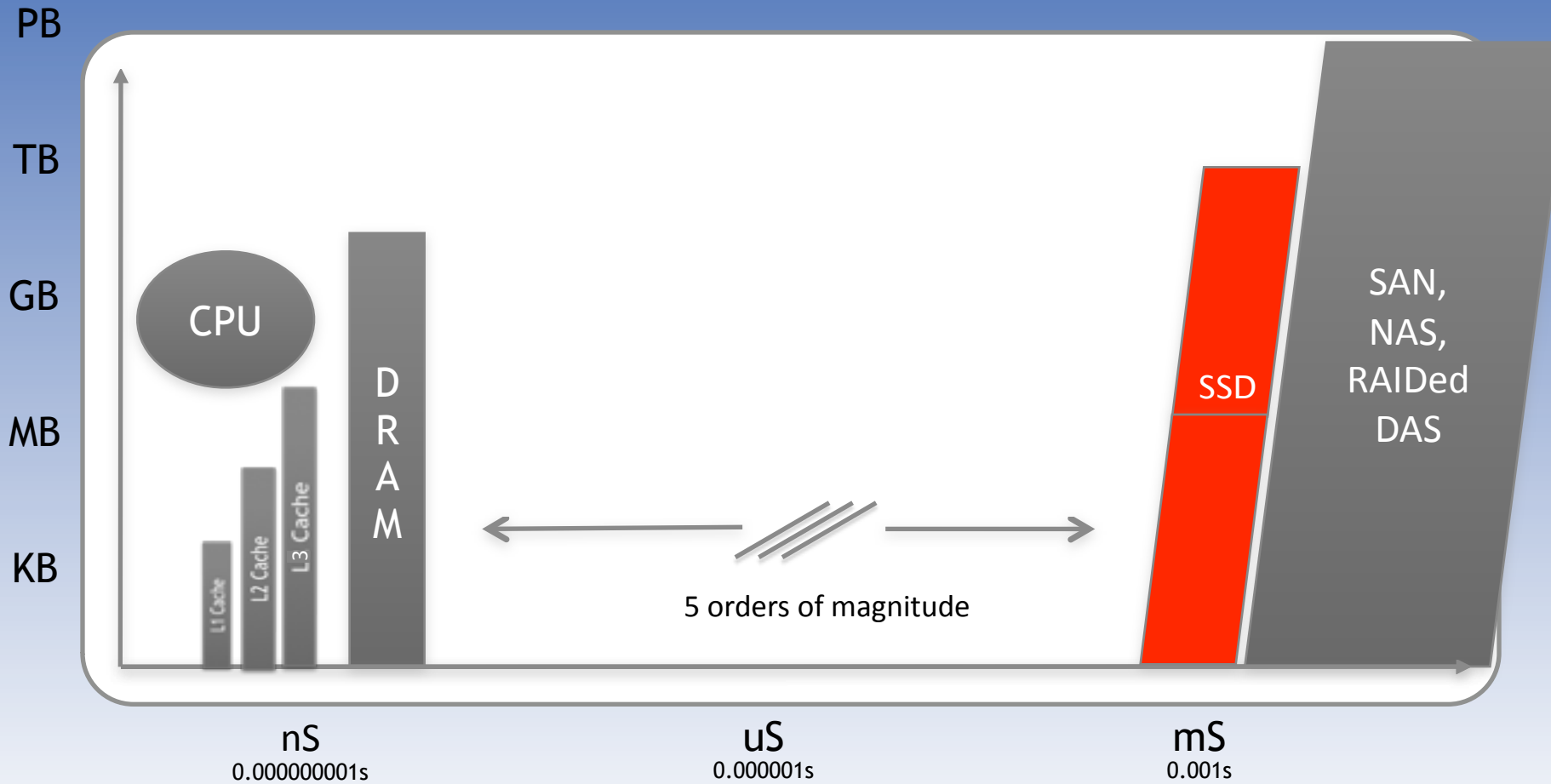
- Fewer CPUs
- Less RAM
- Less Network Gear
- Fewer SW Licenses
- Less Space

## Lower OpEx

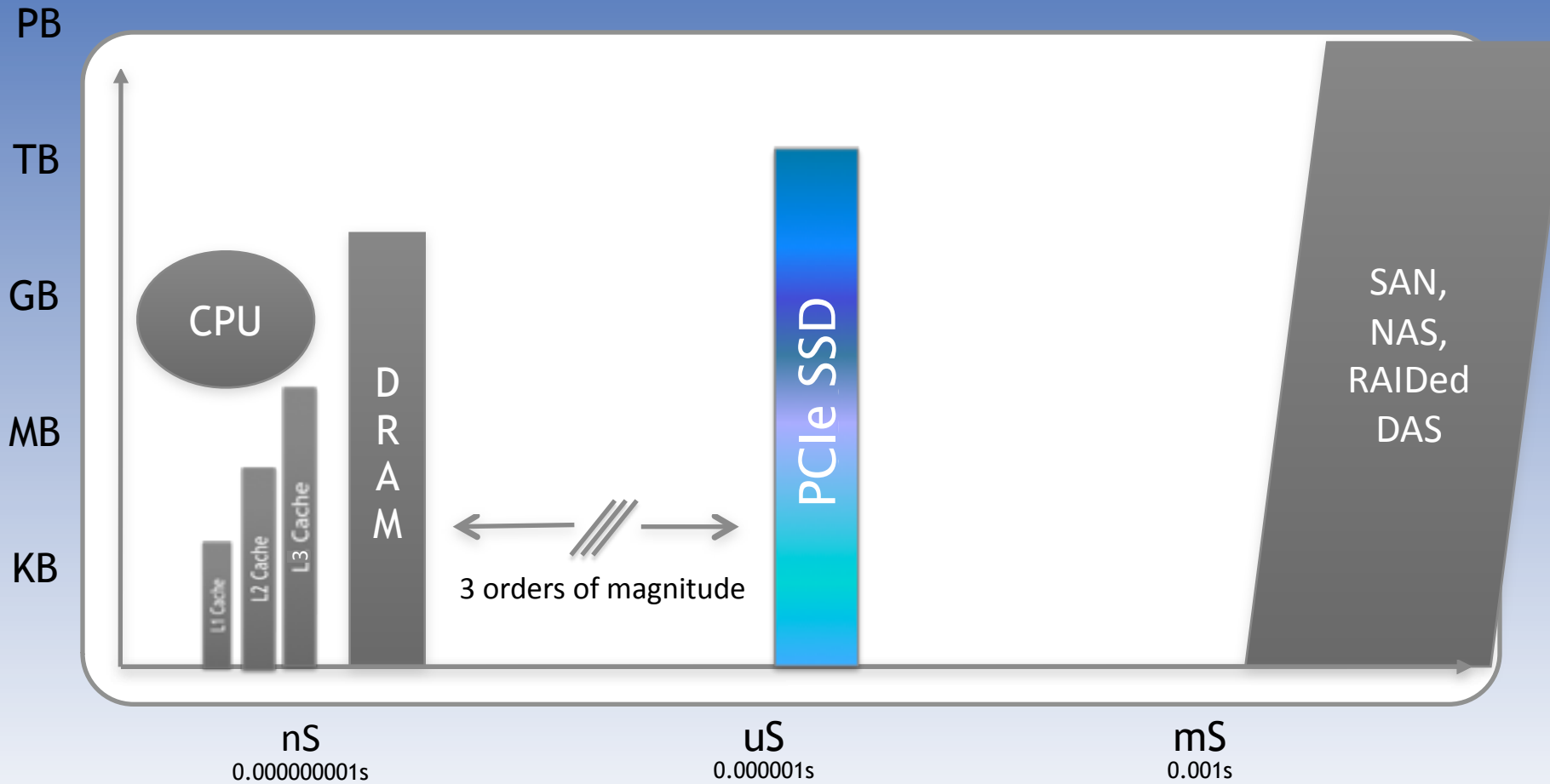
- Less HW Maintenance
- Less SW Maintenance
- Greater Uptime
- Less Power/Cooling
- Fewer Diverse Skills

**HIGHER Productivity**

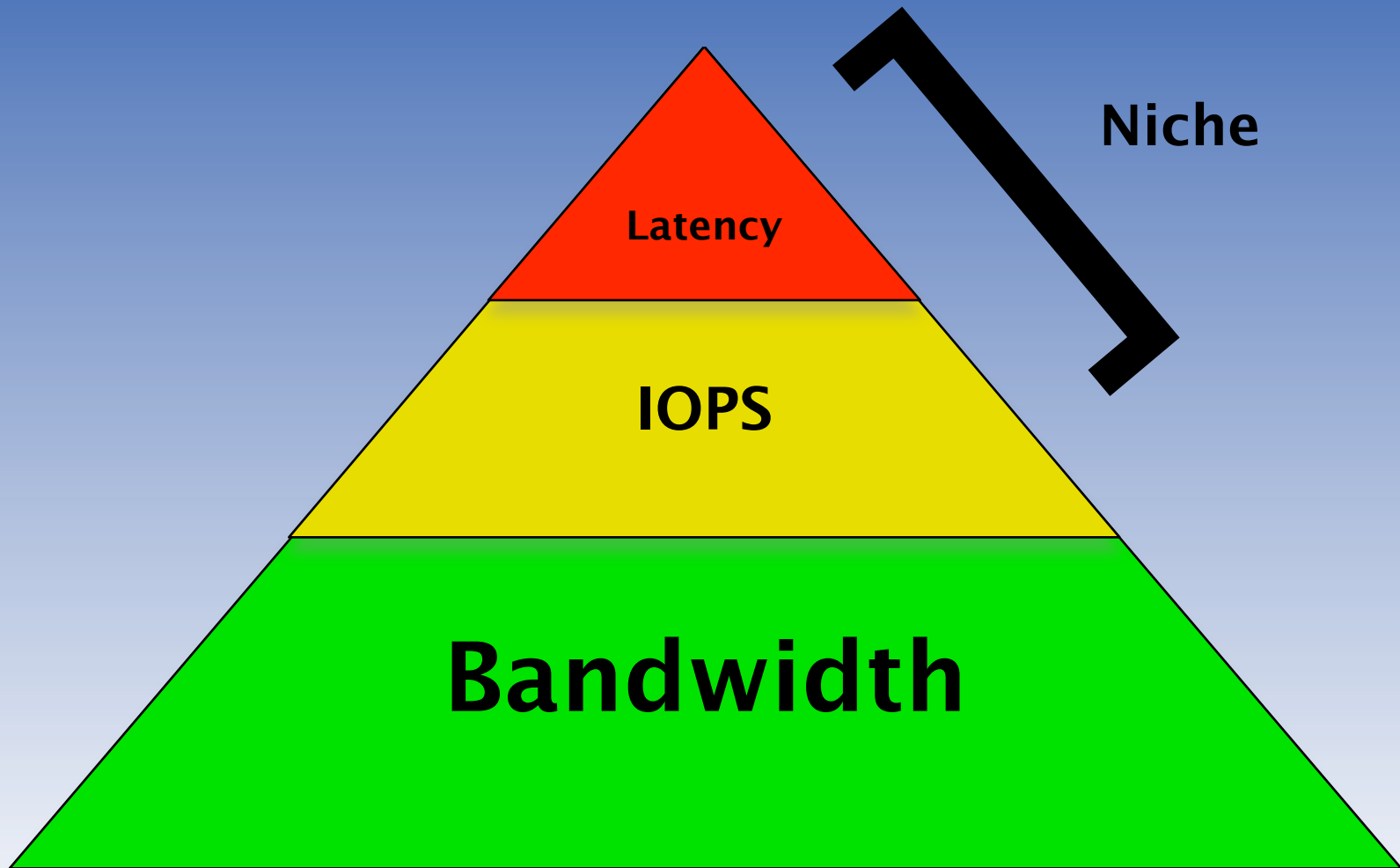
# Traditional SSD's



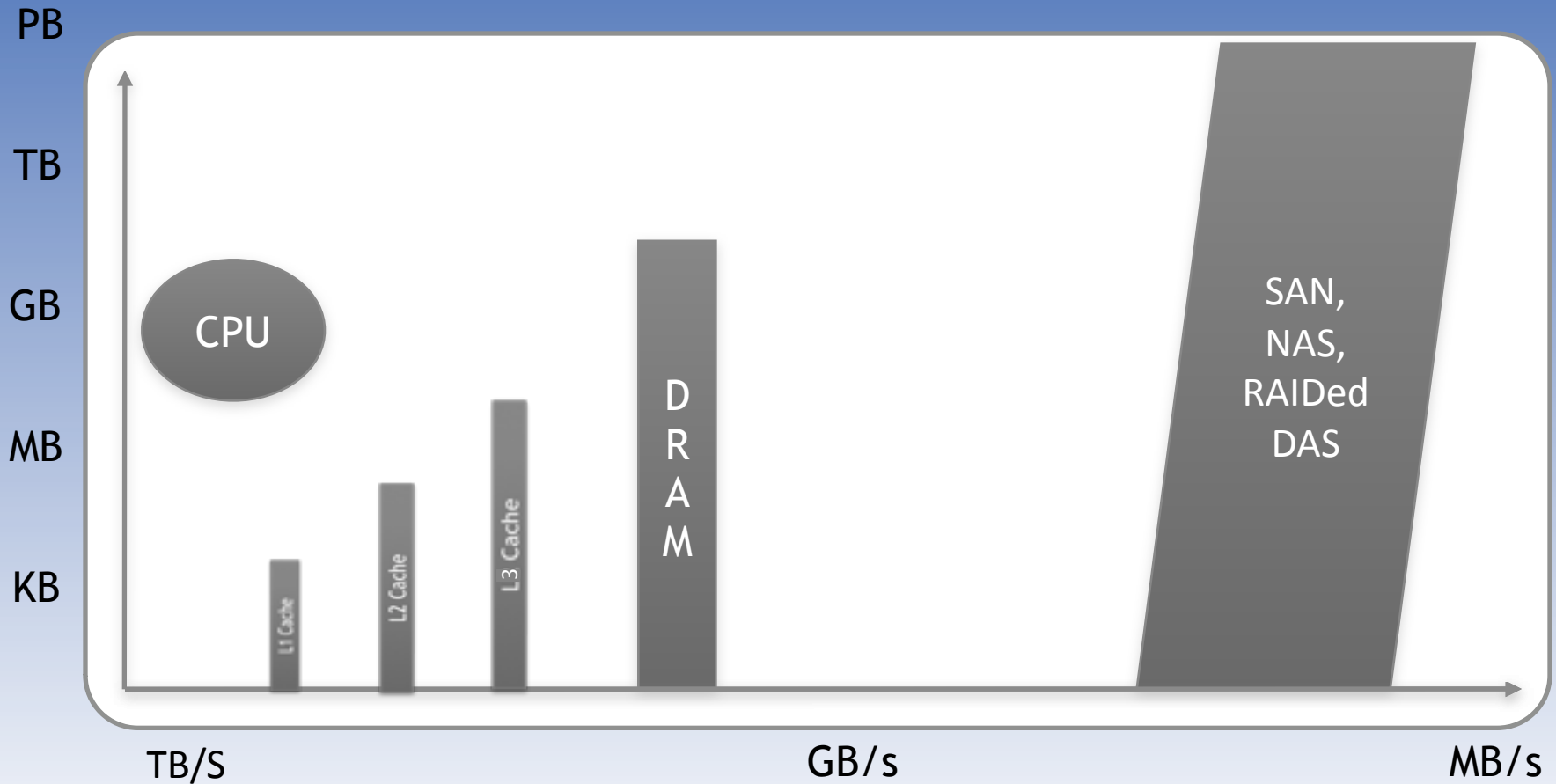
# PCIe Attached SSD's



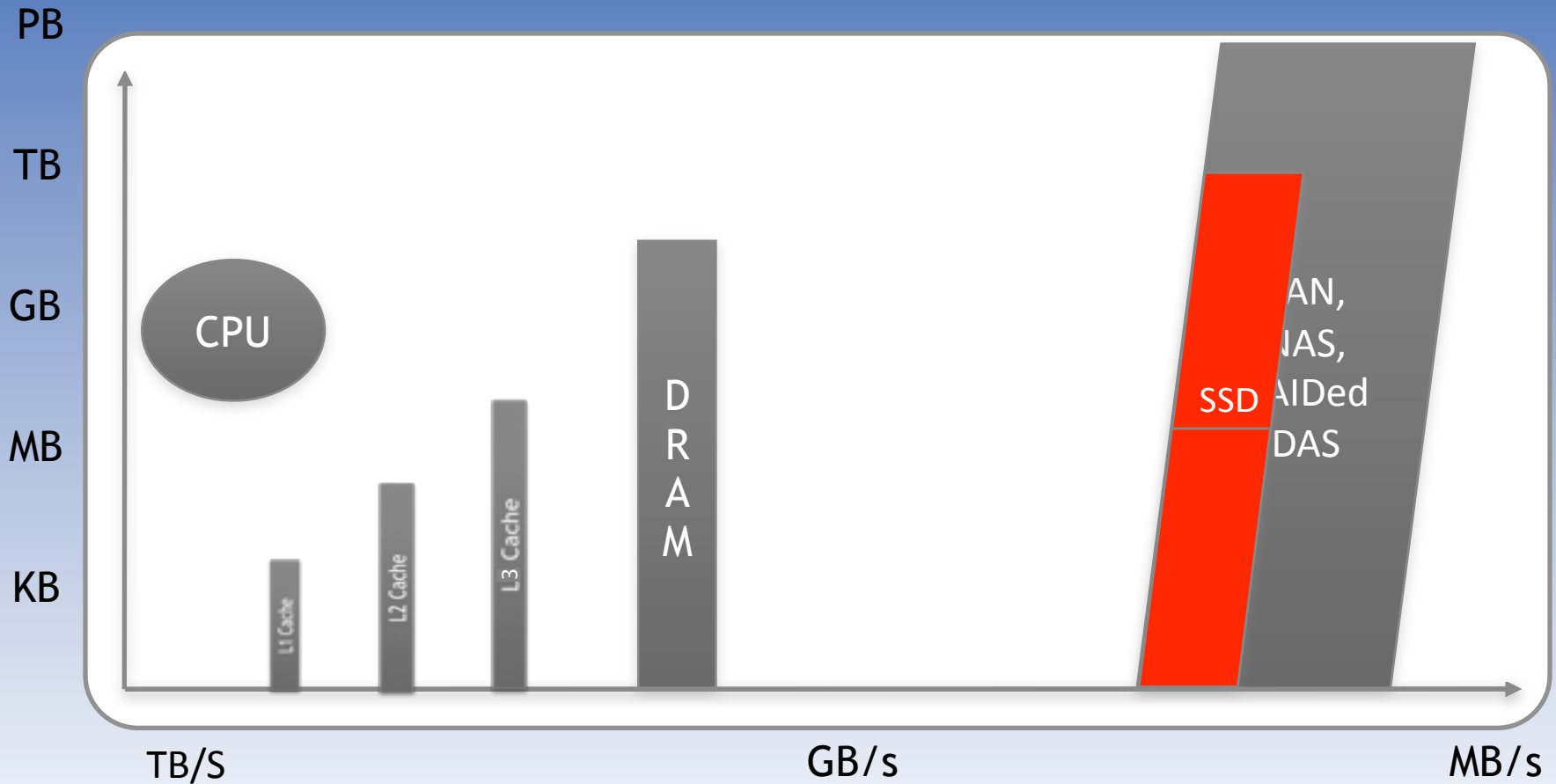
# Performance Dimensions



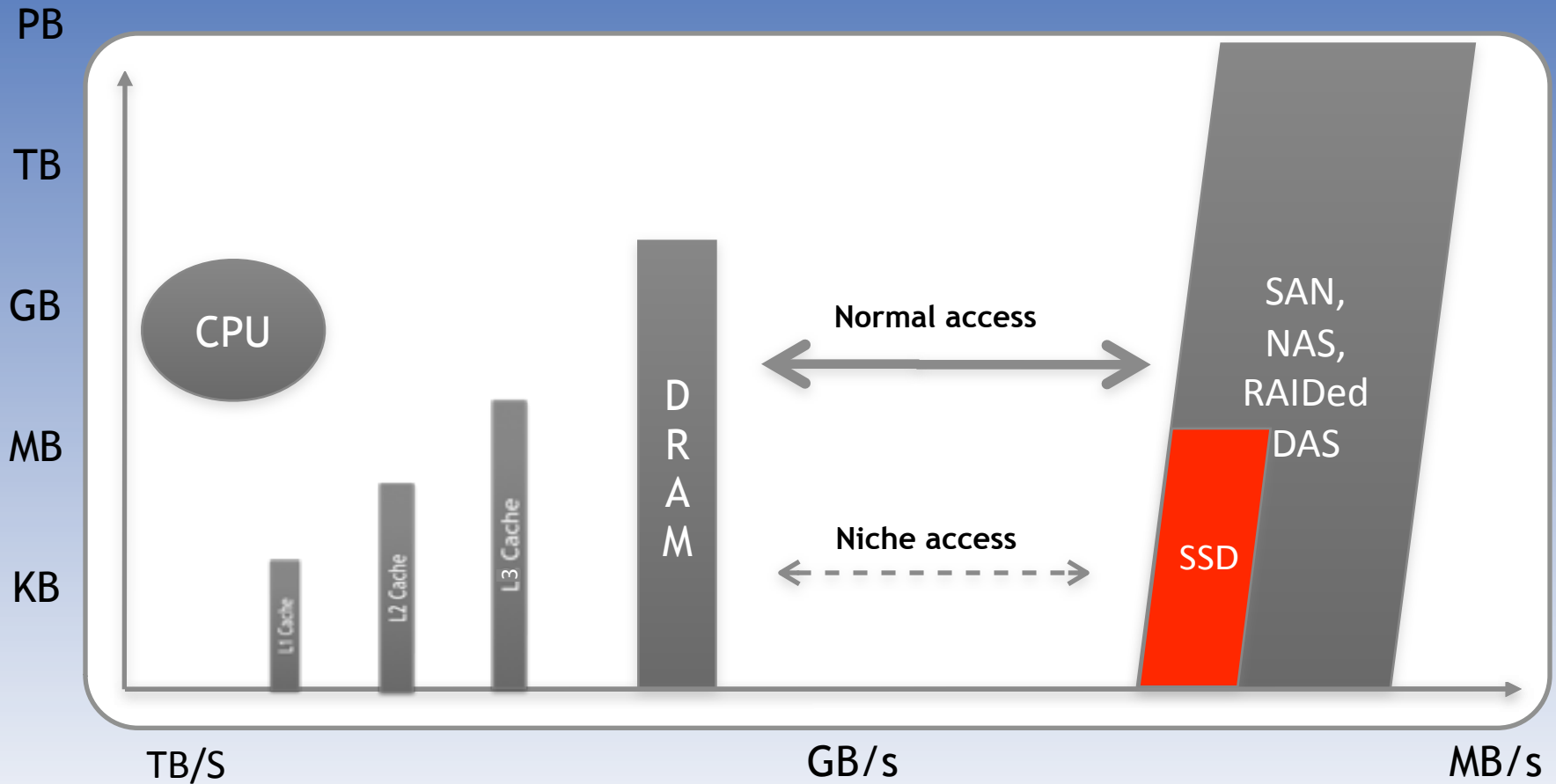
# Access Bandwidth



# Traditional SSD's are no better



# Workload Segregation





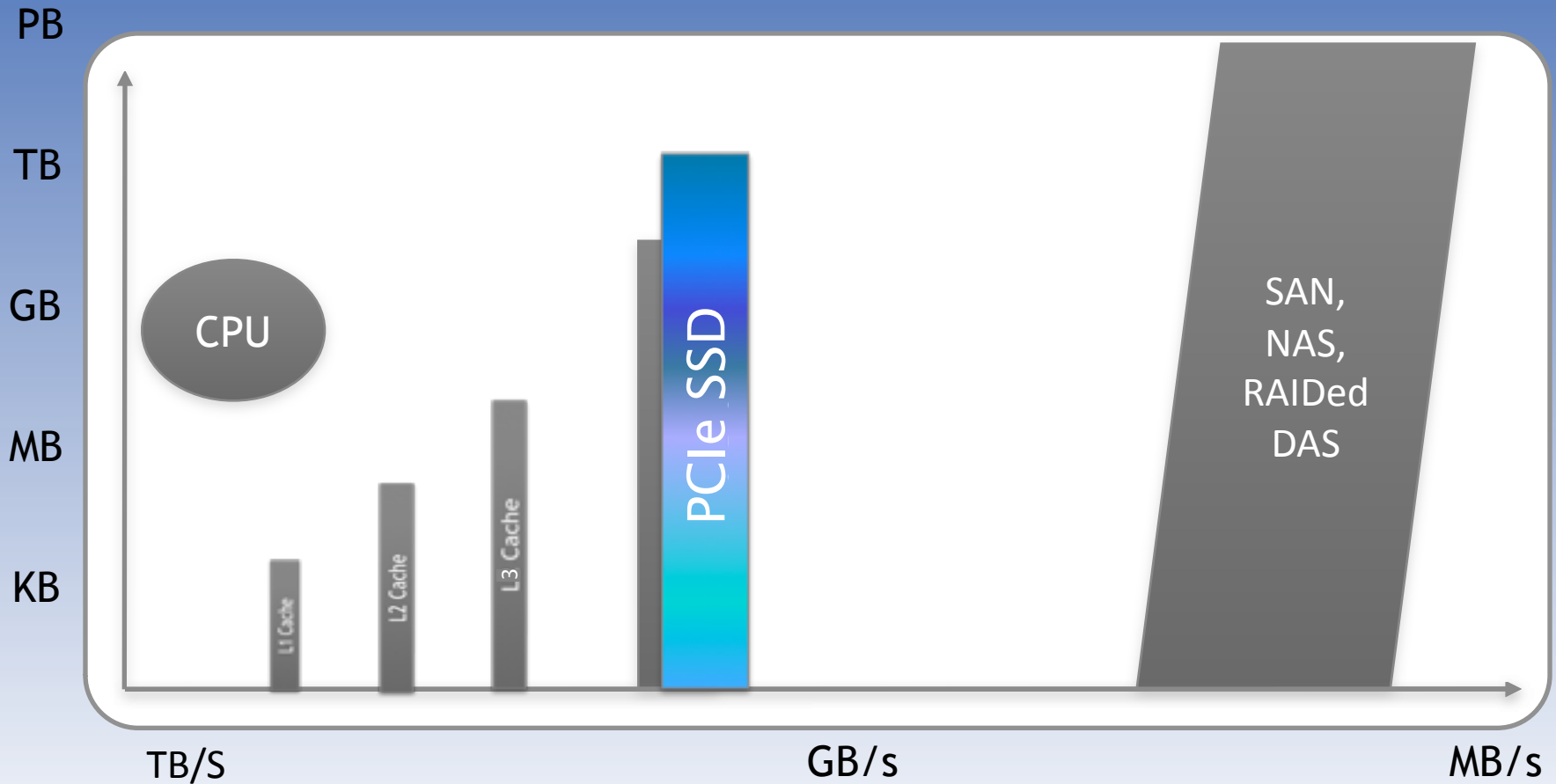
# A cache needs...

- 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



# Raw Media Performance

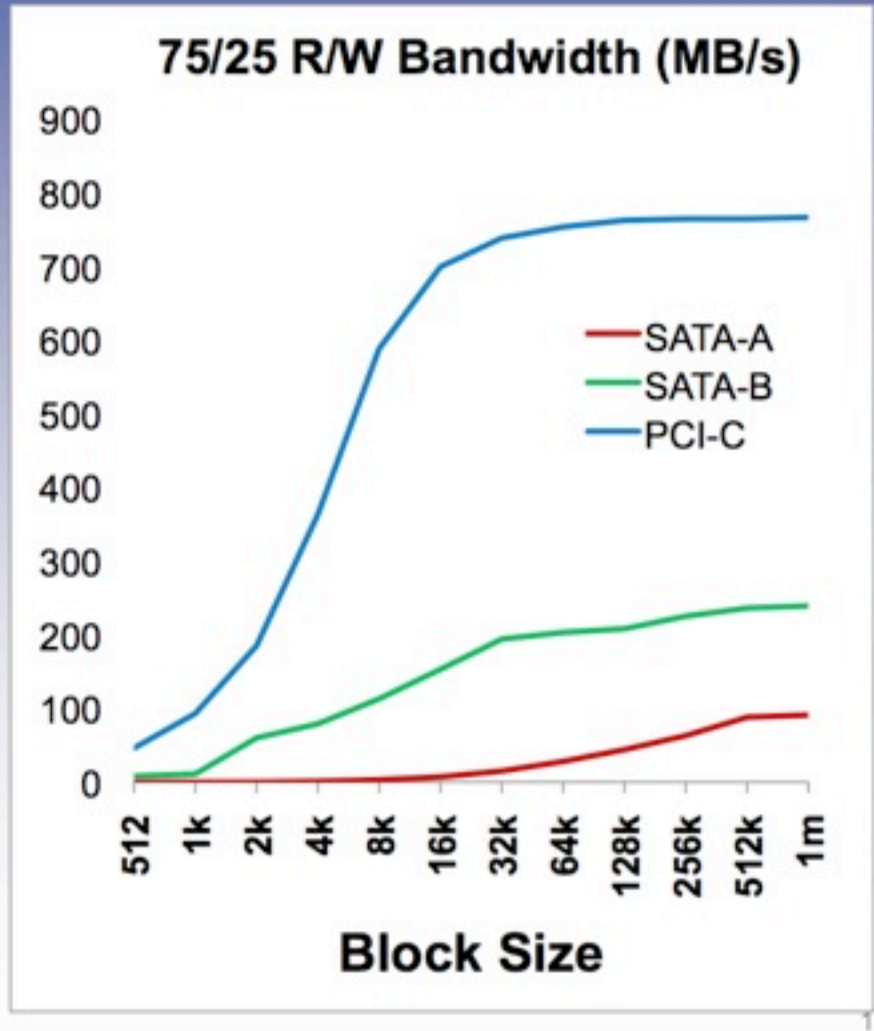
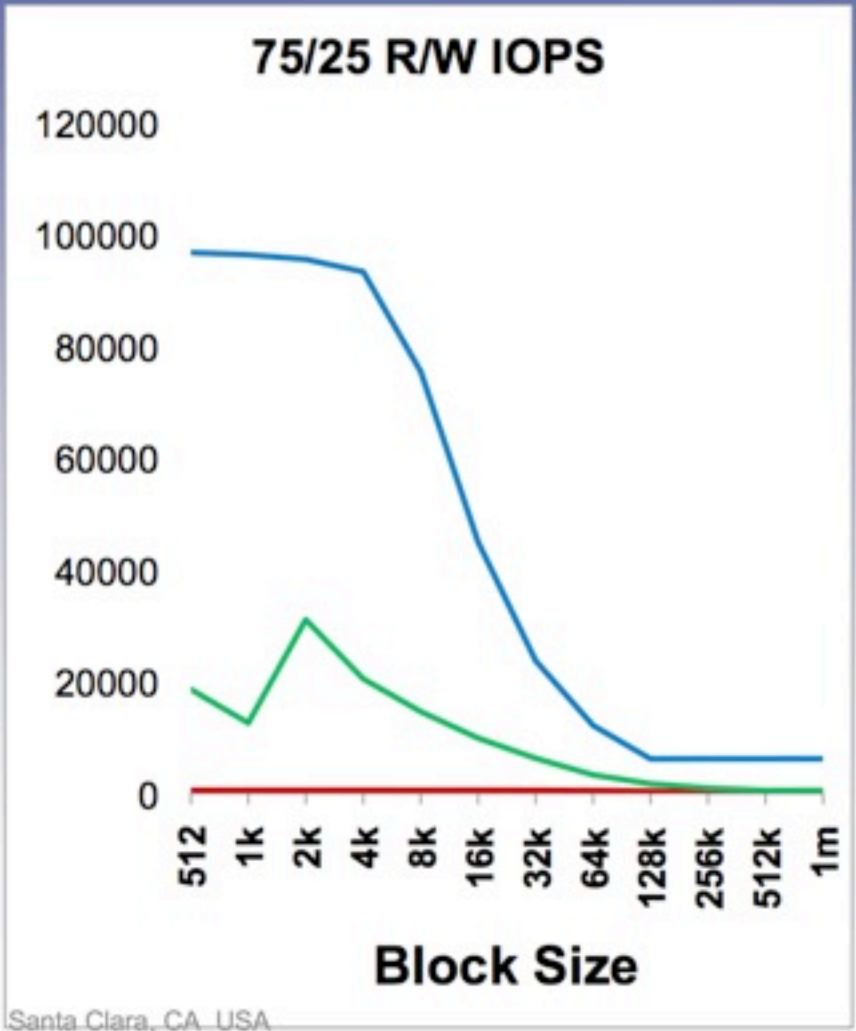
- The GOOD
  - Performance is excellent (wrt HDDs)
  - High performance per power (IOPS/Watt)
  - Low pin count: shared command / data bus → good balance
- The BAD
  - Not really a random access device
    - Block oriented
    - Slow effective write (erase/transfer/program) latency
    - R/W access speed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

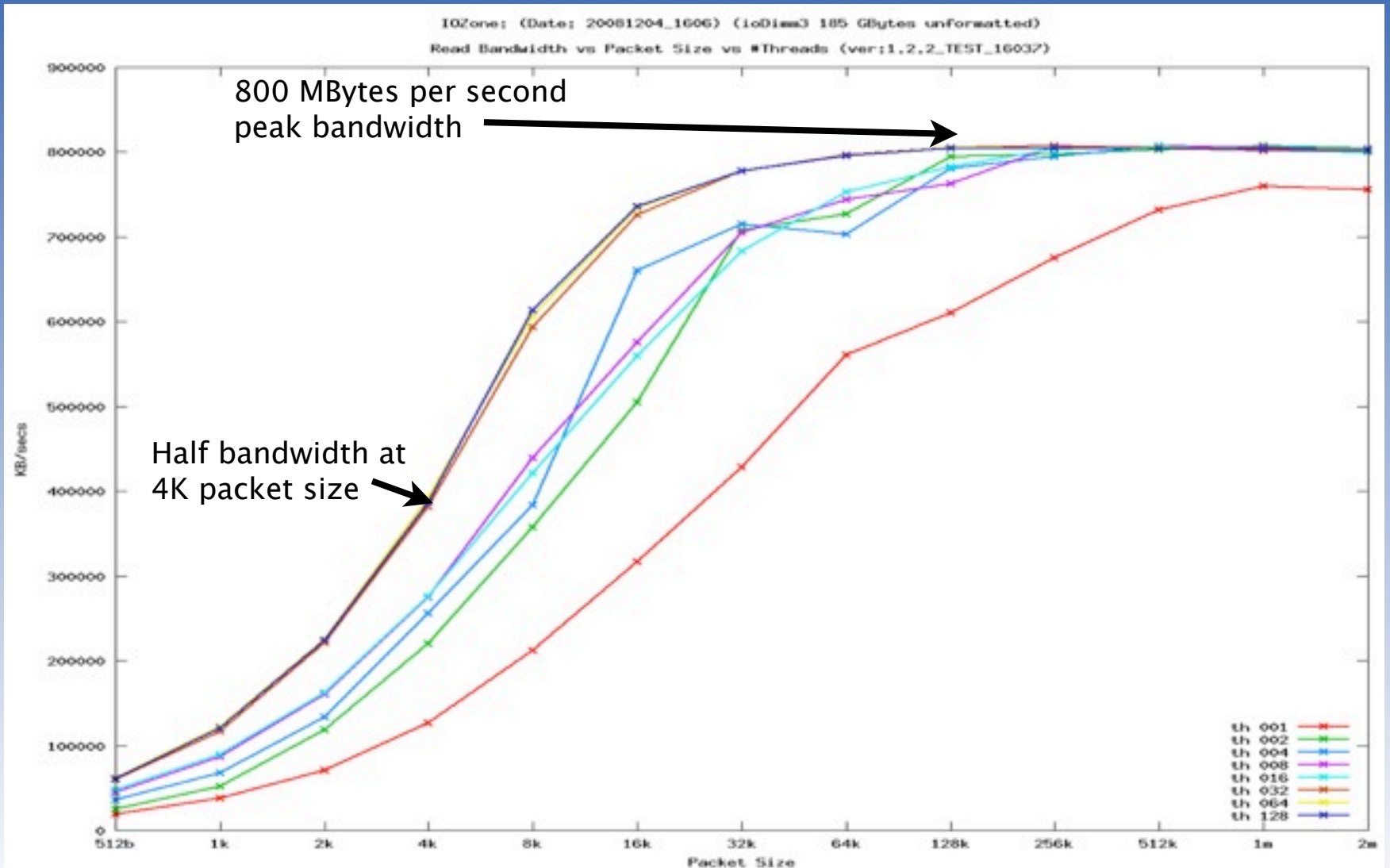


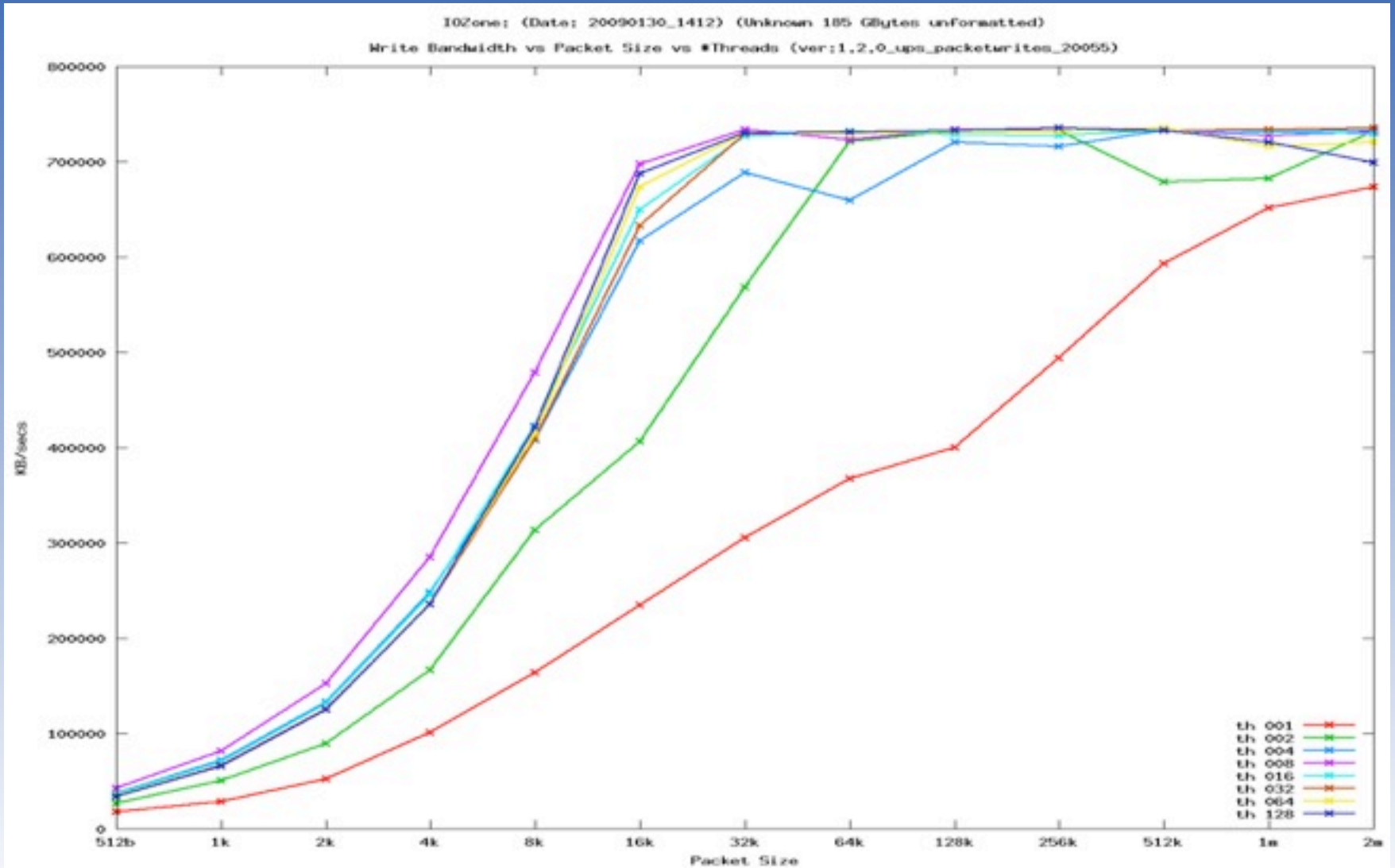
# Controller Performance Drivers

- 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)





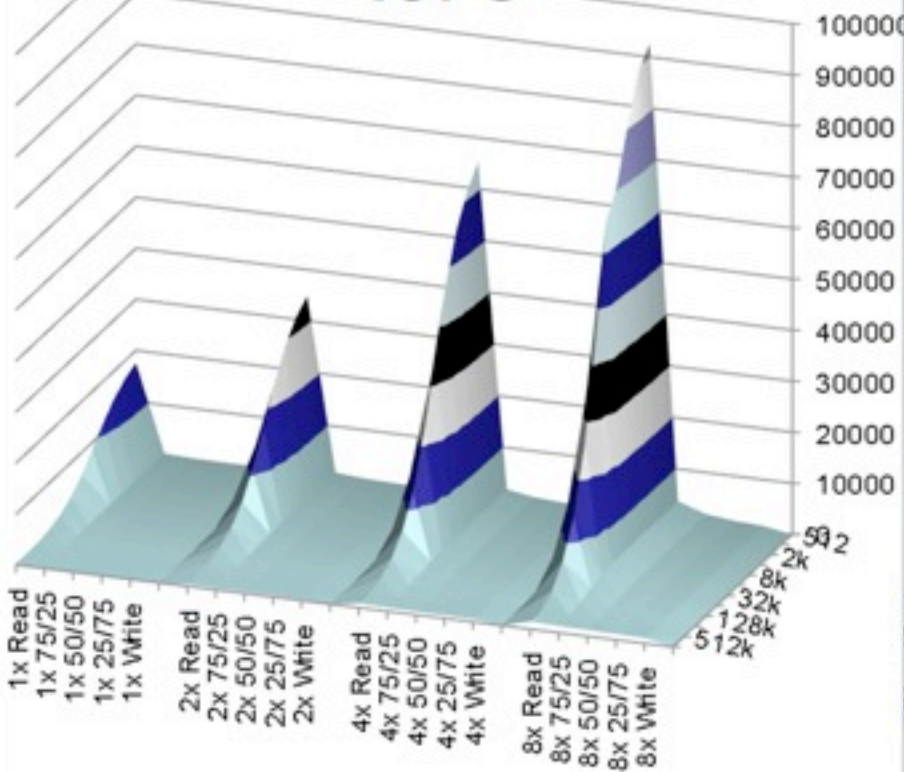


# Scalability

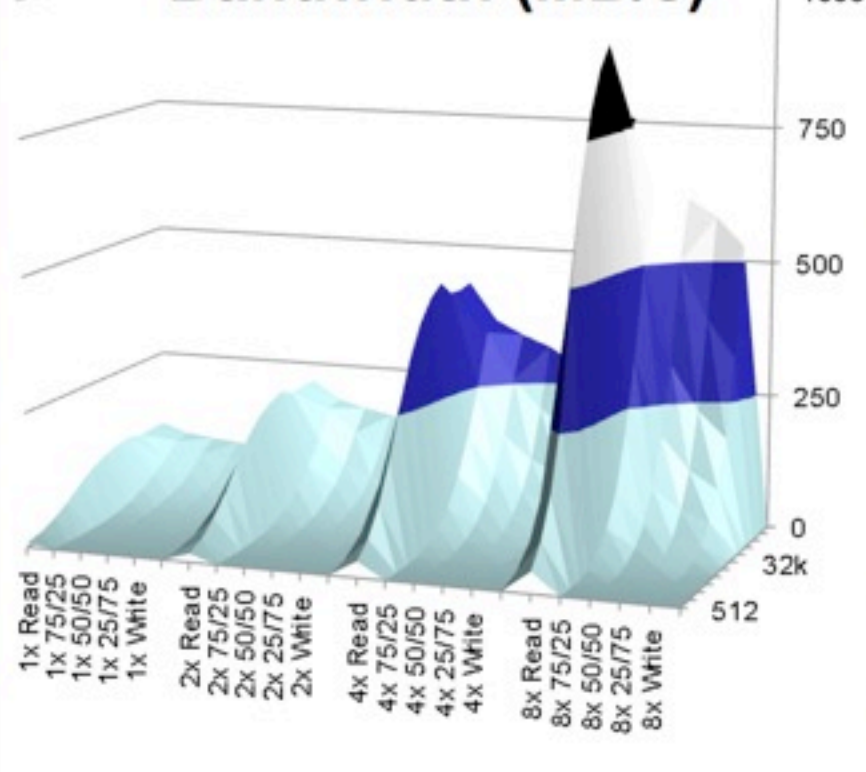
- Following Slides Show
  - Scalability of {1, 2, 4, 8} units
  - Only 1 SATA controller is used – limiting scalability
  - Only 1 thread running
- Measurements taken at Read/Write Ratios of
  - {100/0, 75/25, 50/50, 25/70, 0/100}
  - RMS value is the “root mean square” of these values
- IOPS measurement 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

## SATA-A Scalability IOPS

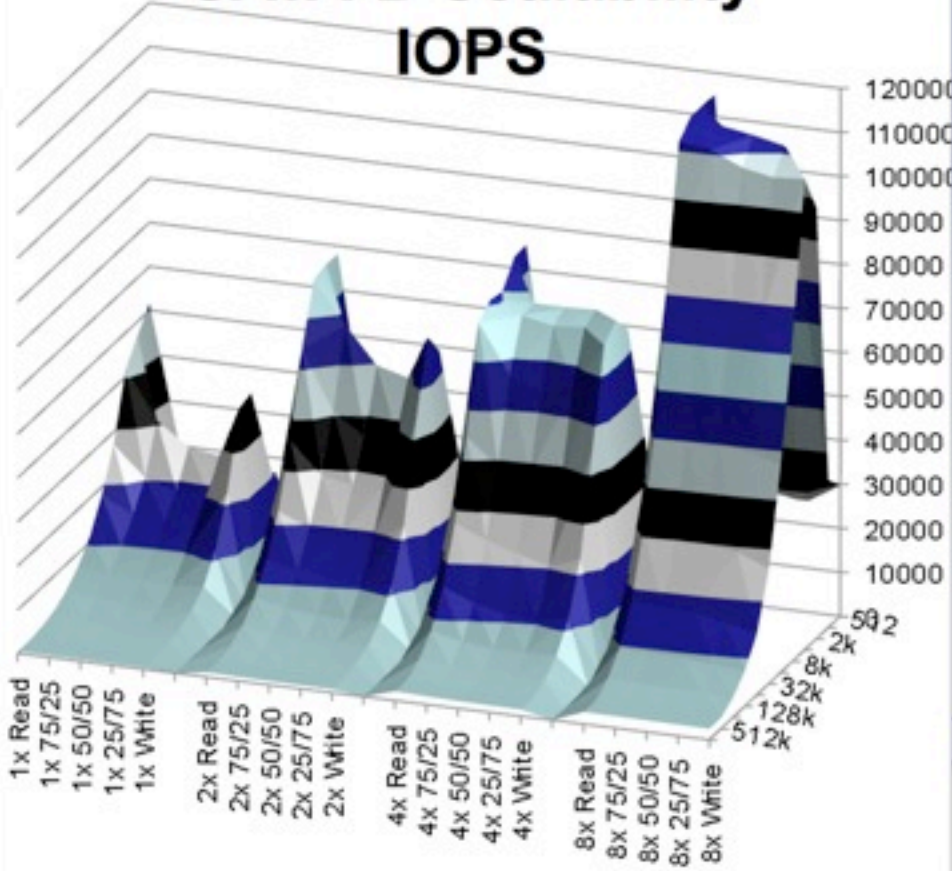


## SATA-A Scalability Bandwidth (MB/s)

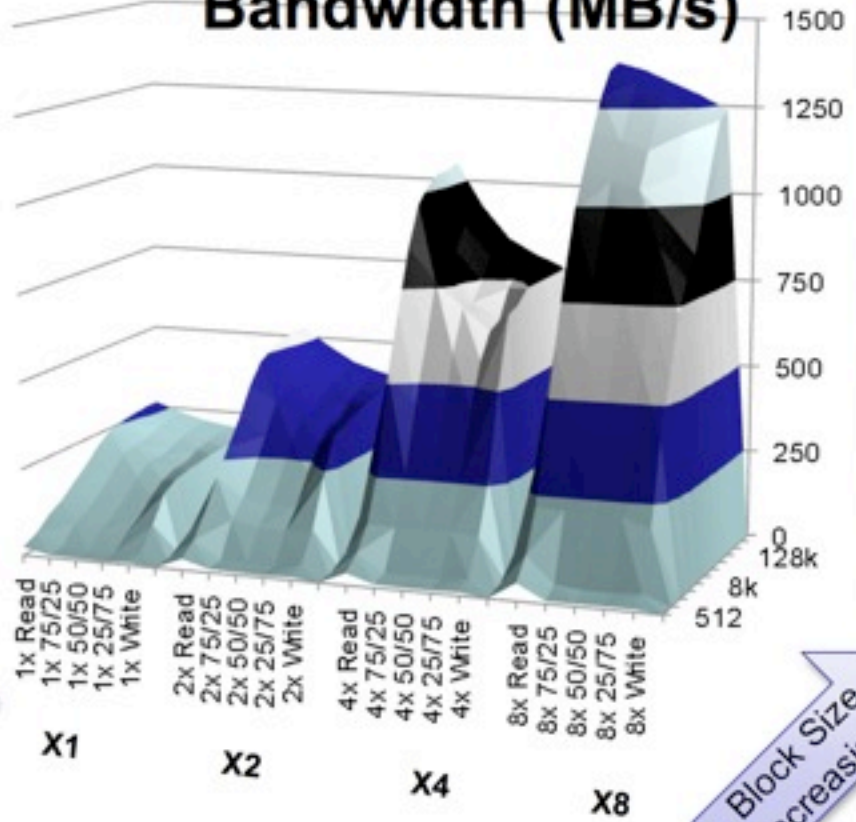


# Scalability vs RW Ratio vs Block Size

## SATA-B Scalability IOPS



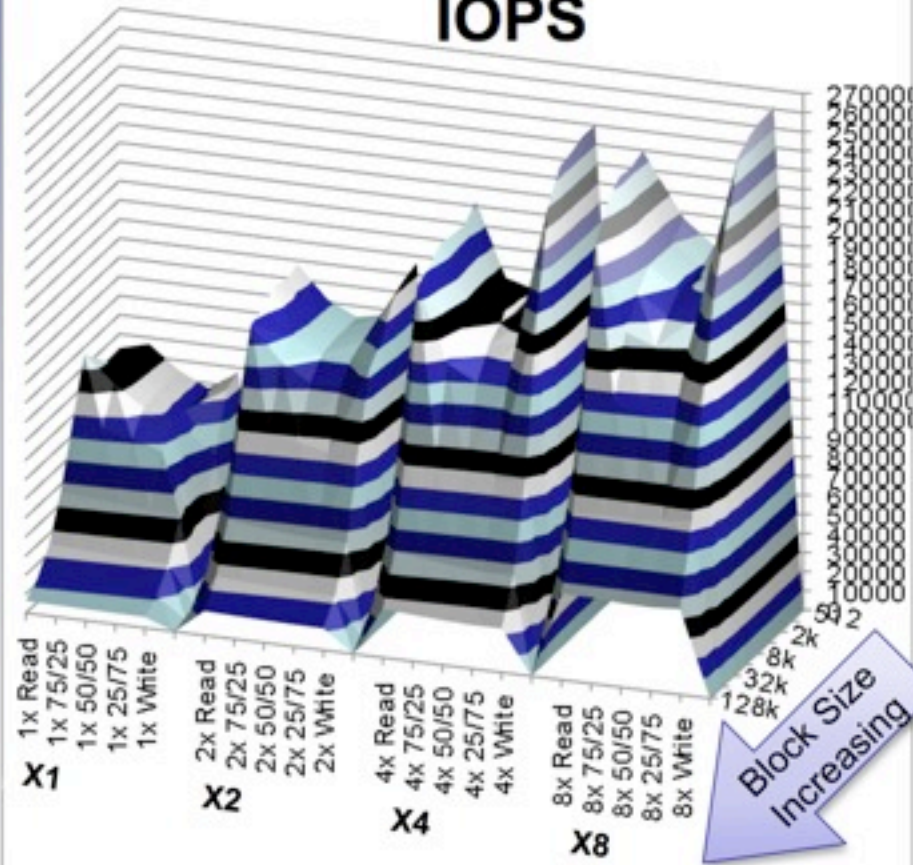
## SATA-B Scalability Bandwidth (MB/s)



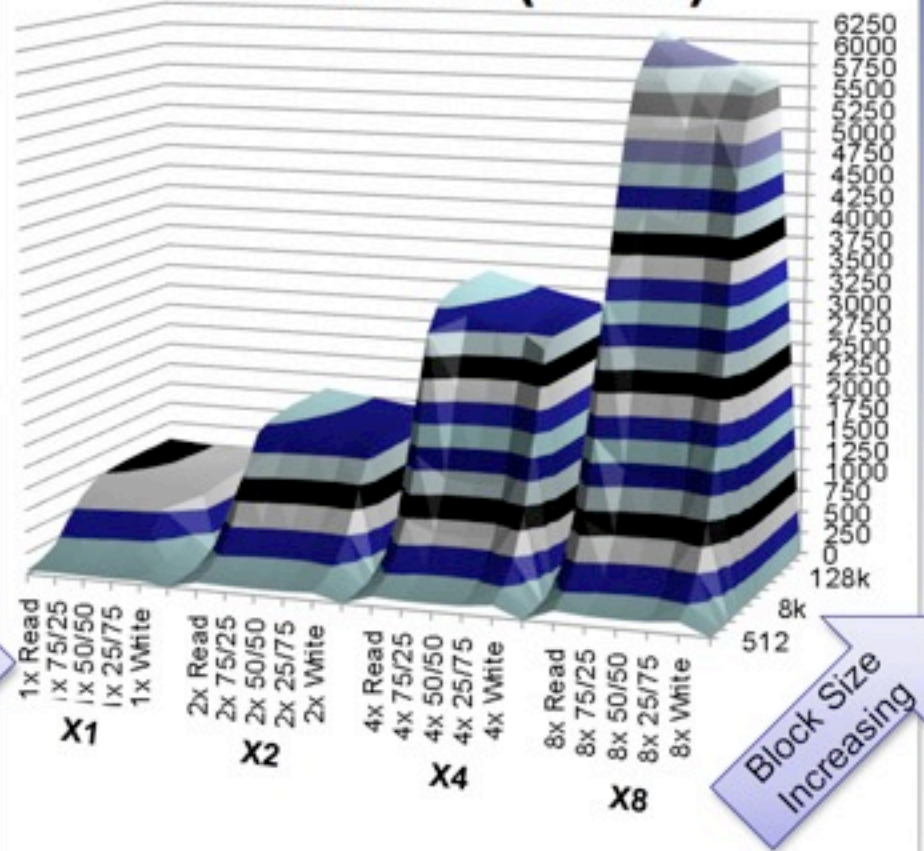
Block Size Increasing

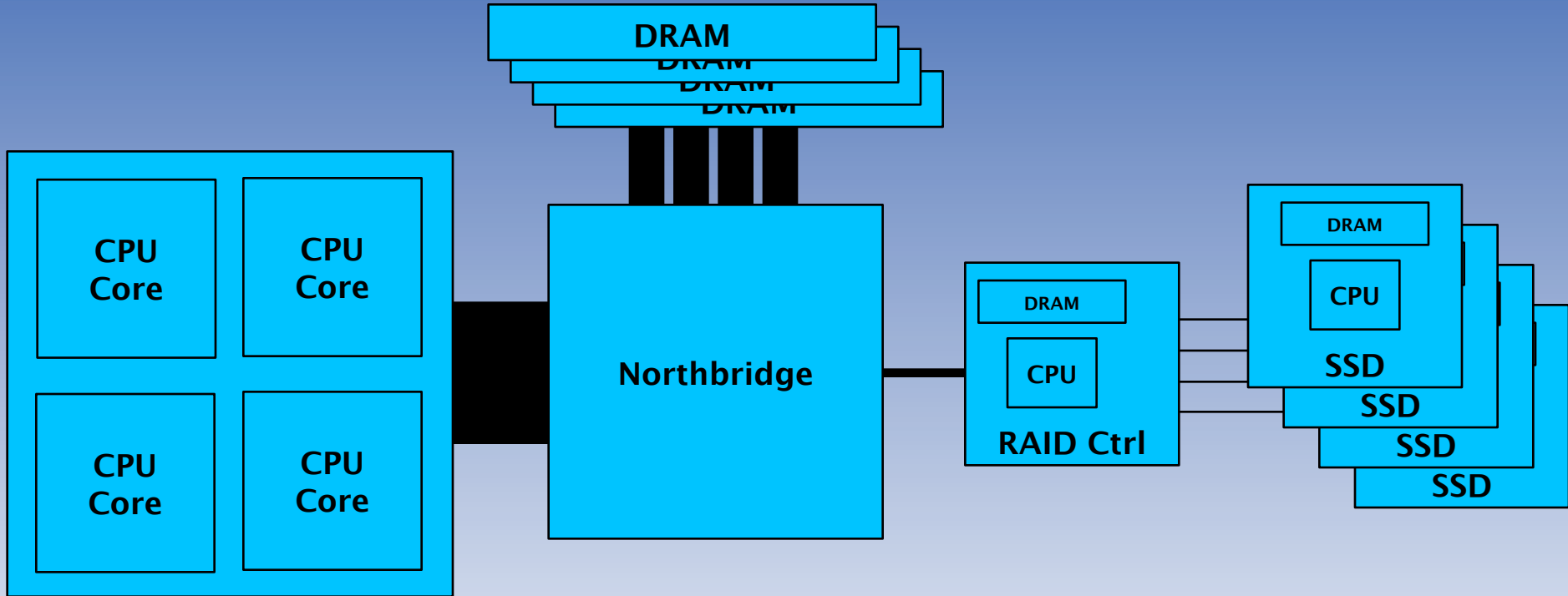
# Scalability vs RW Ratio vs Block Size

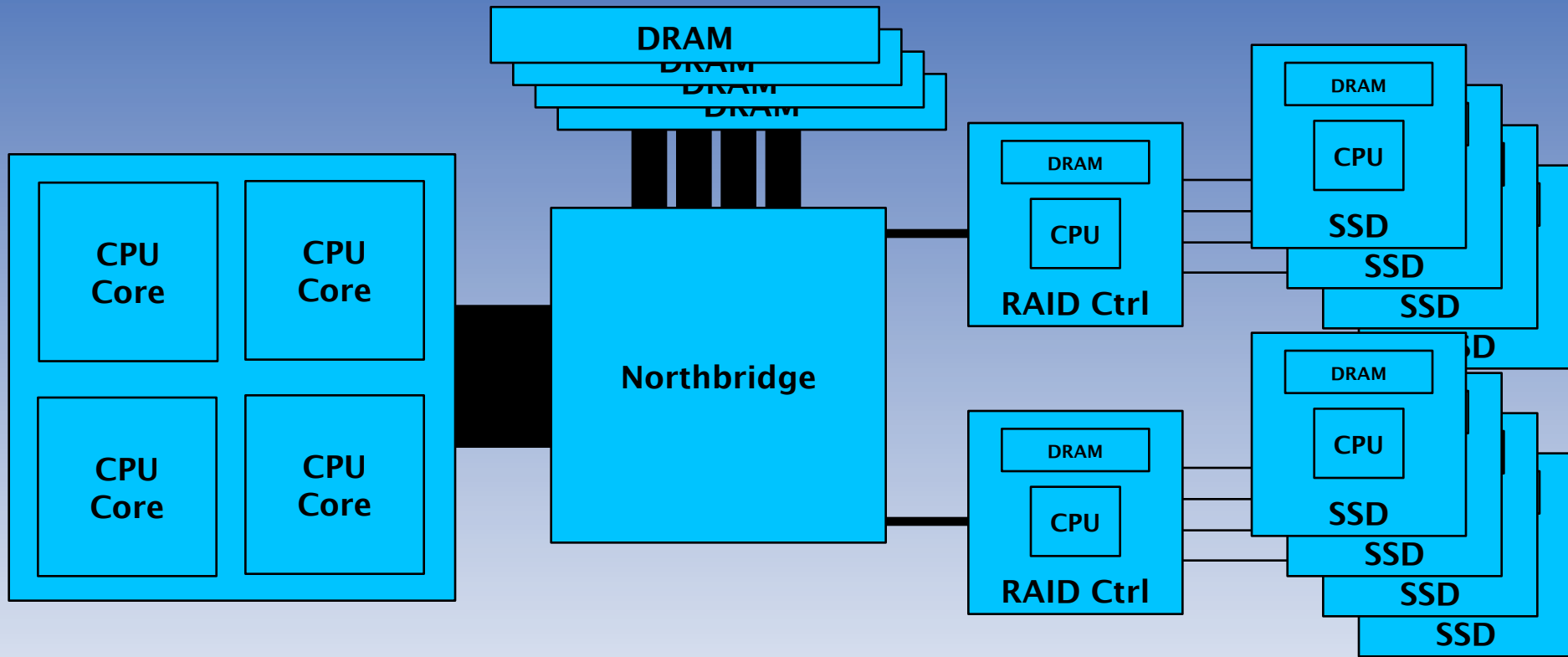
## PCI-C Scalability IOPS

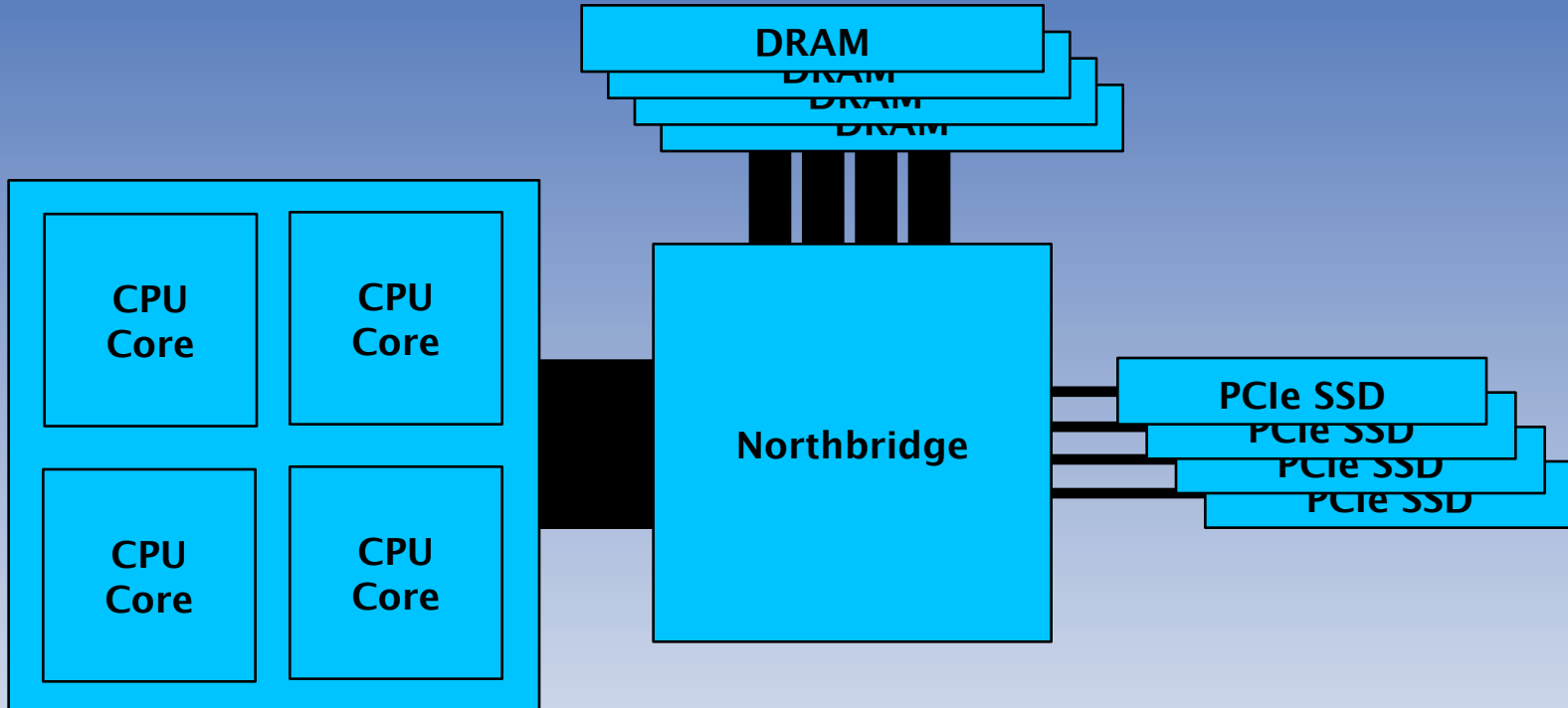


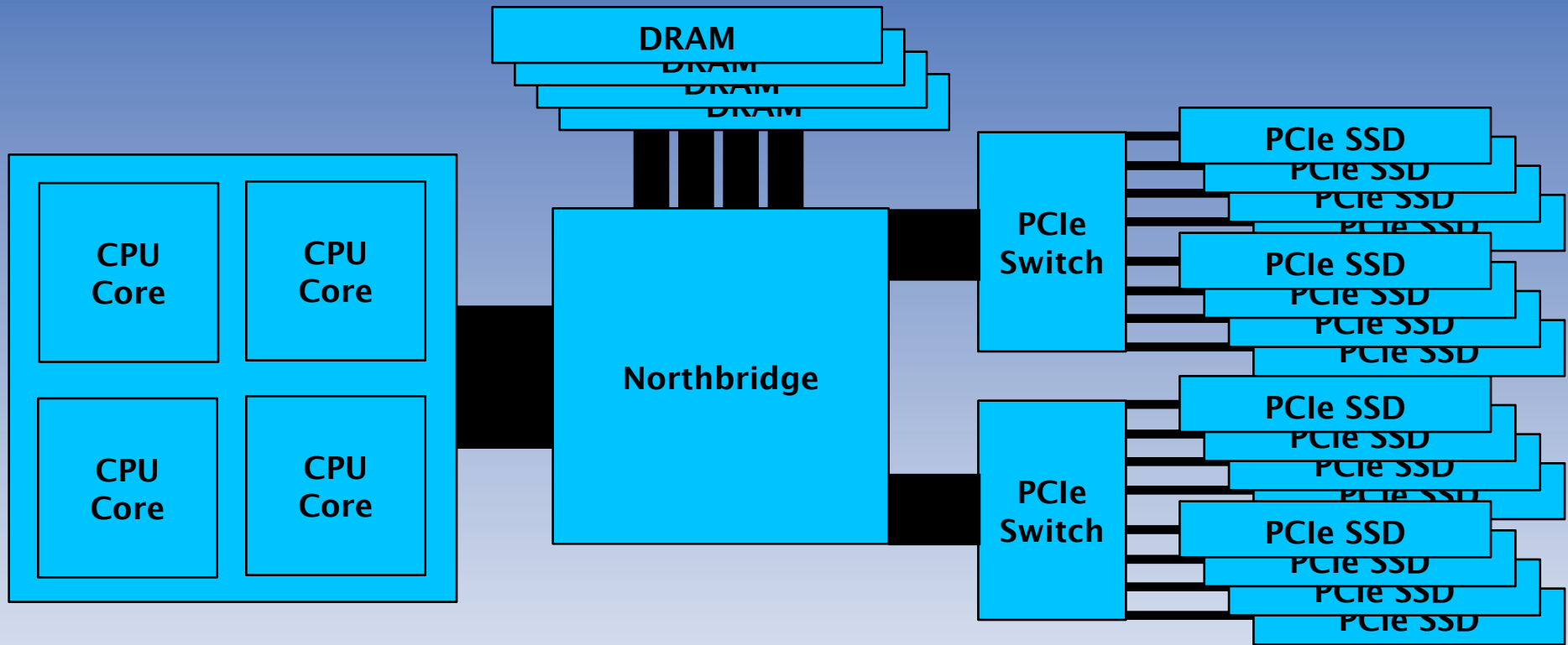
## PCI-C Scalability Bandwidth (MB/s)













***Thank you!***