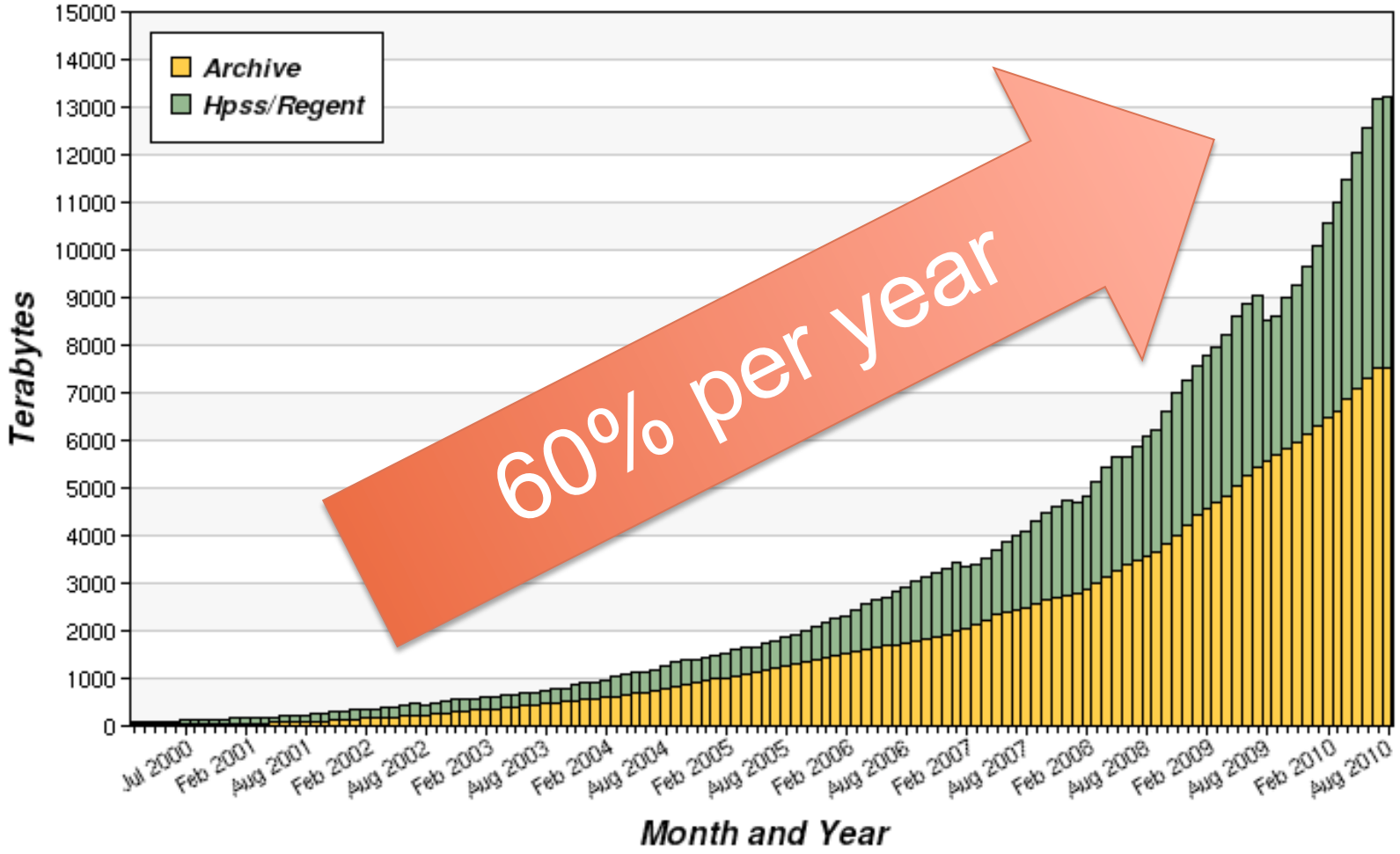
Performance Analysis of Commodity and Enterprise Class Flash Devices

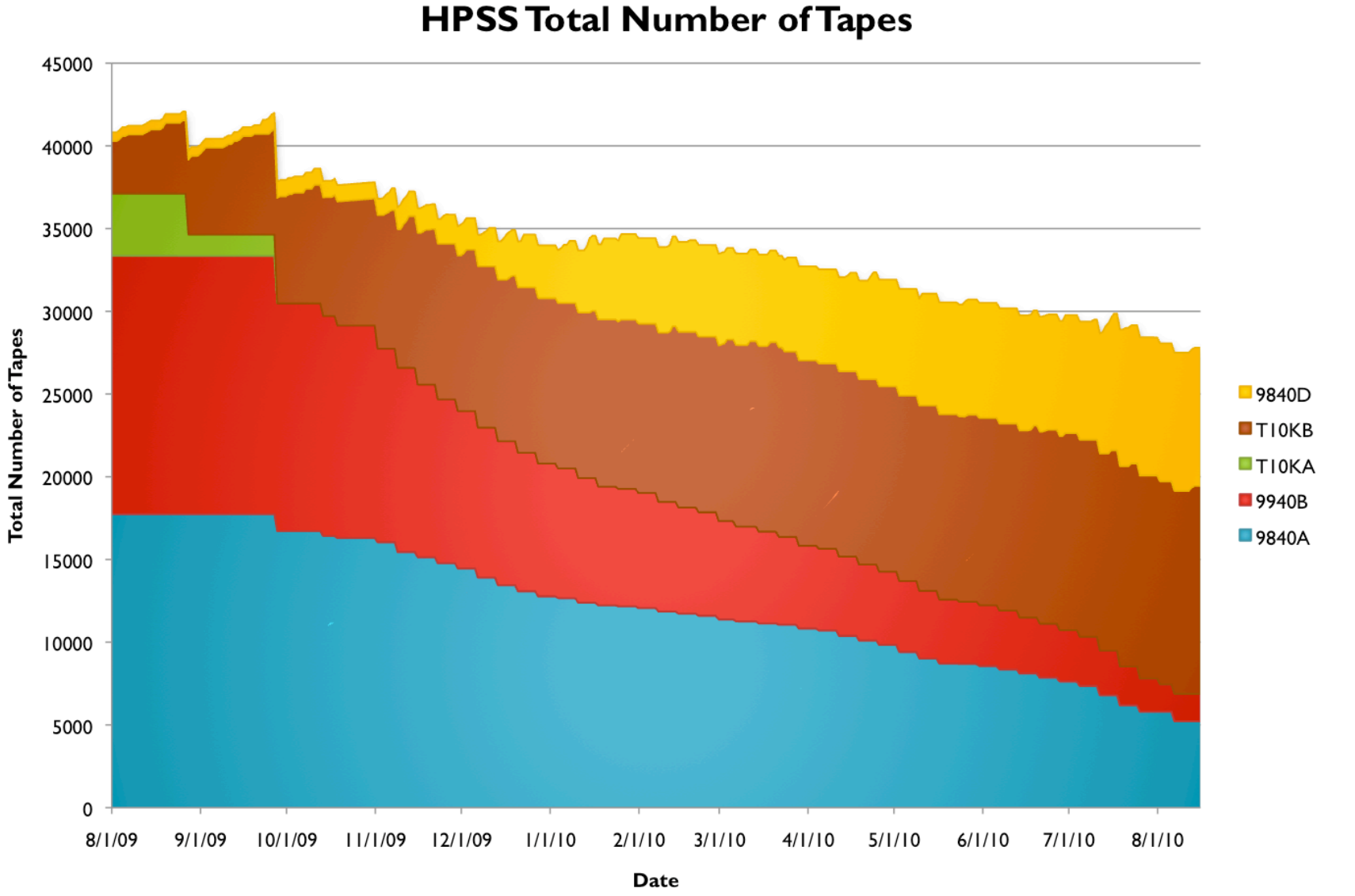
Neal M. Master,
Matthew Andrews, Jason Hick,
Shane Canon &
Nicholas J. Wright

Lawrence Berkeley
National Laboratory

Data Trends at NERSC

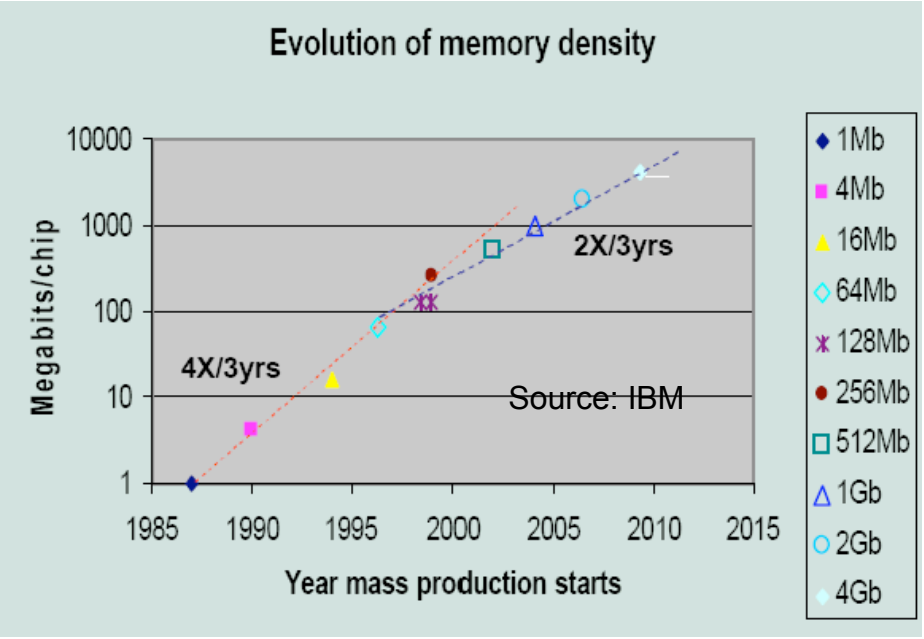


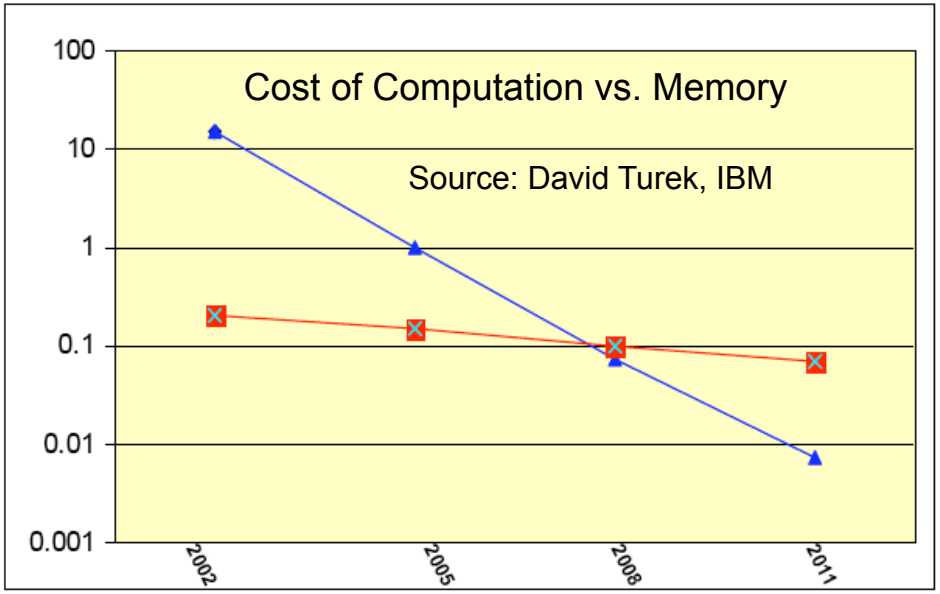
Data Trends at NERSC cont.



Memory Capacity Trends

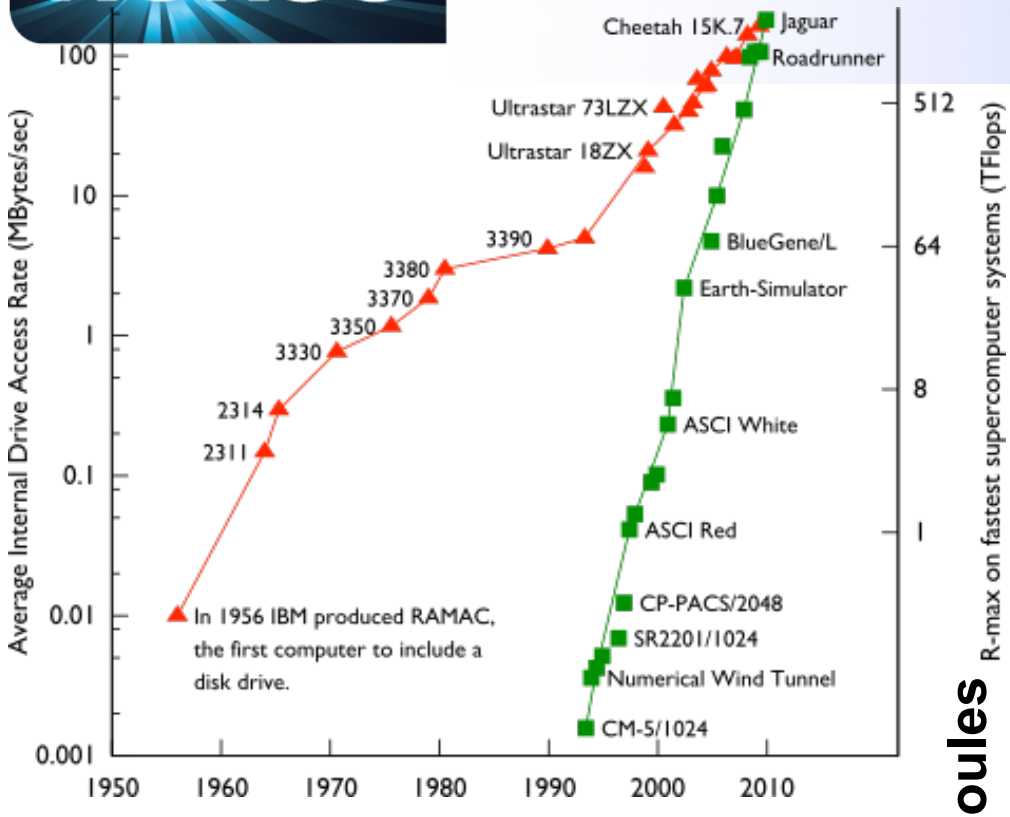
- **Technology trends:**
  - **Memory density 2X every 3 yrs; processor logic every 2**
  - **Storage costs ($/MB) drops more gradually than logic costs**





The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it

I/O Performance Challenges

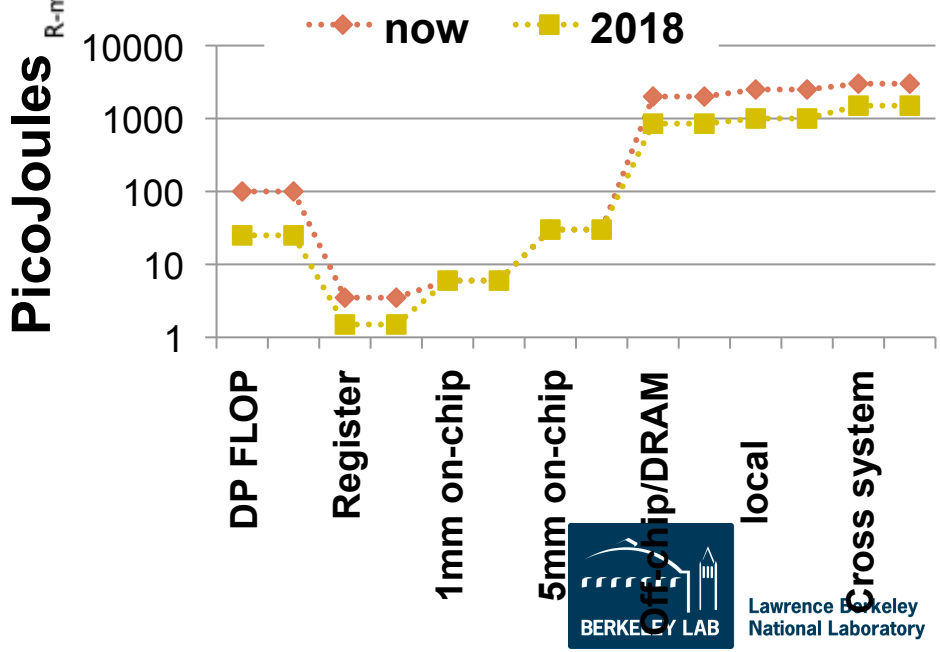


Performance Crisis #1

- **Disks are outpaced by compute speed.**
- **To achieve reasonable aggregate bandwidth many spindles needed – 10^3 spindles = 1PB but only ~0.1 TB/s !**

Performance Crisis #2

Data Motion on an Exascale Machine will be expensive – both in terms of energy & performance !



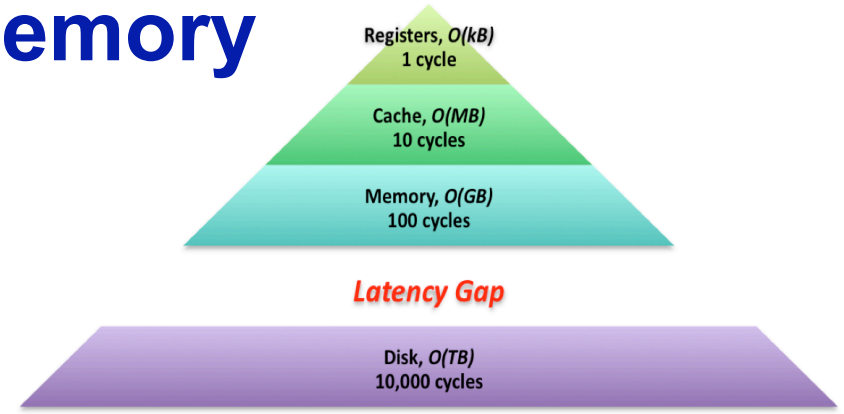
U.S. DEPARTMENT OF ENERGY | Office of Science

Lawrence Berkeley National Laboratory

Flash Memory - Ubiquitous

Flash – What is it good for?

- Fits nicely into latency gap between spinning disk and memory
- Lots of open Q's:
  - PCI vs SATA vs ?
  - SLC vs MLC
  - Write requires block erase - performance dependent upon previous IO pattern
  - Correct algorithm in software at all levels
  - ….



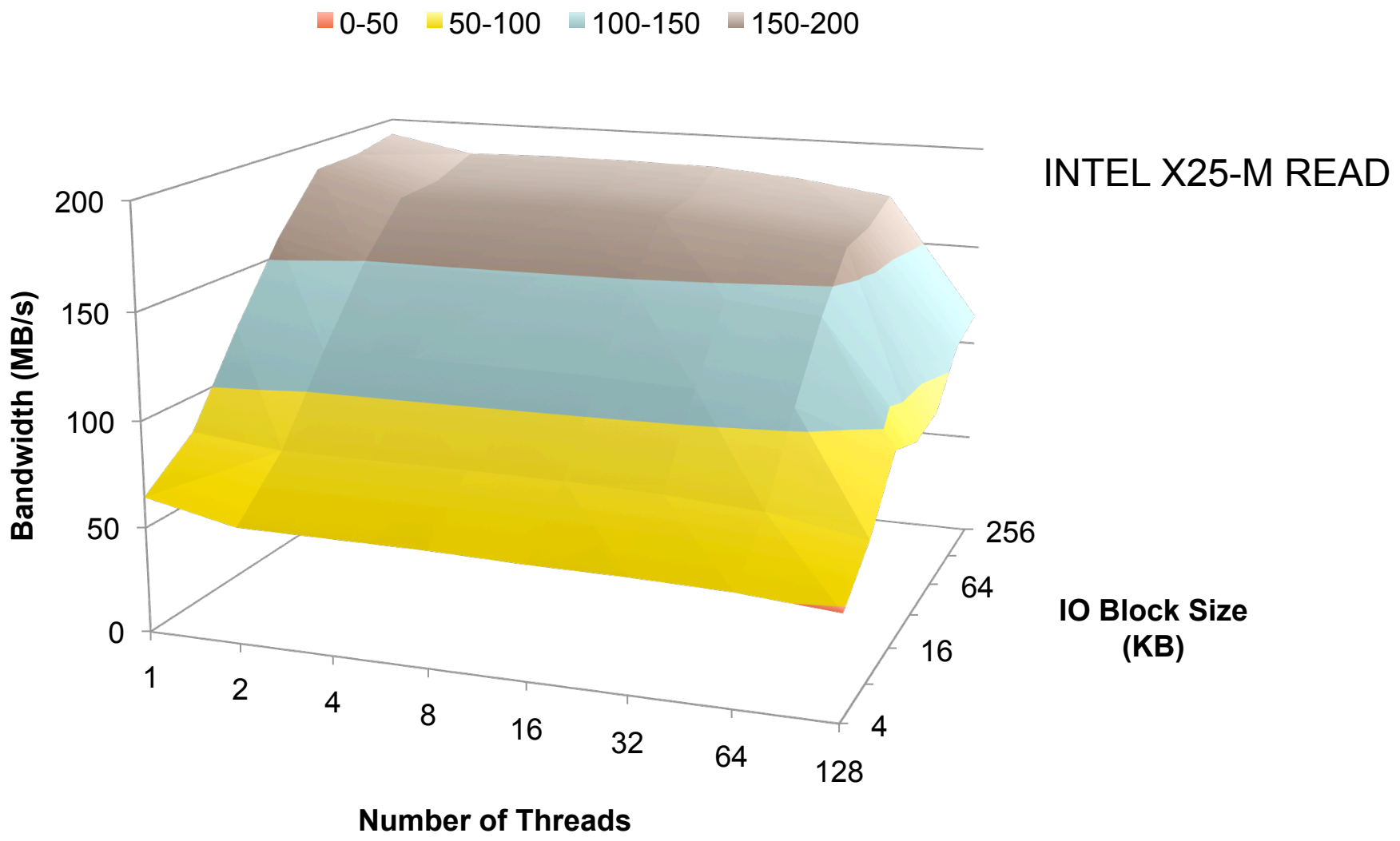
Devices Evaluated

- 3 PCI-e SLC
  - Virident tachIOn 400GB 8x
  - FusionIO ioDrive Duo 2x 160GB 4x
  - Texas Memory Systems RamSan-20 450GB 4x
- 2 SATA MLC
  - Intel X-25M 160GB
  - OCZ Colossus 250GB

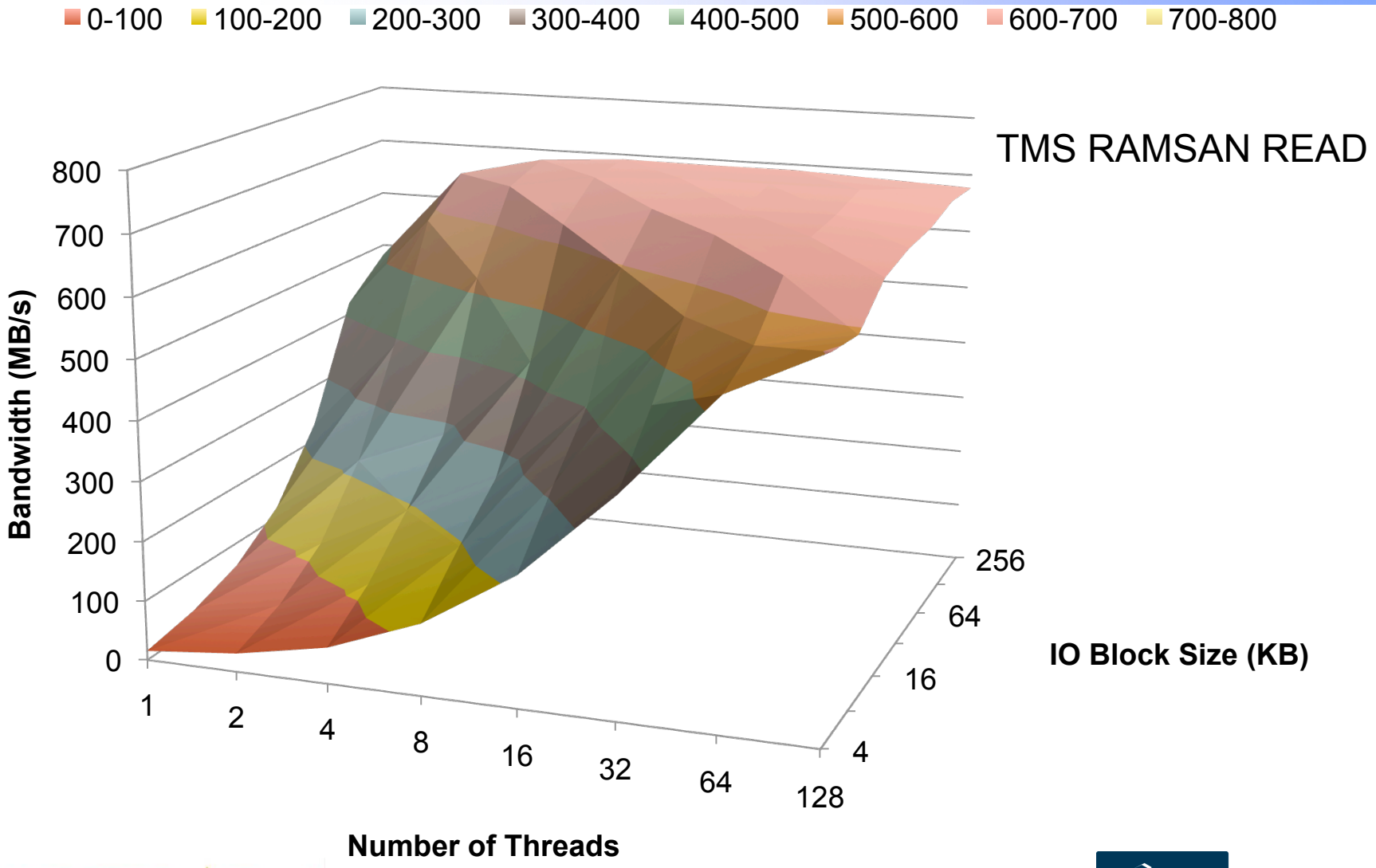
IOZone Experiments

- **Bandwidth**
  - Vary block size: 2^n KB, n =2-8
  - Vary concurrency: 2^n threads, n=0-7 (1-128)
  - Vary IO Patterns: Sequential Write/Re-write, Sequential Read/Re-read, Random Write, Mixed Random Write/Read, Random Read
- **IOPS**
  - 4KB block size
  - Vary concurrency: 2^n threads, n=0-7 (1-128)

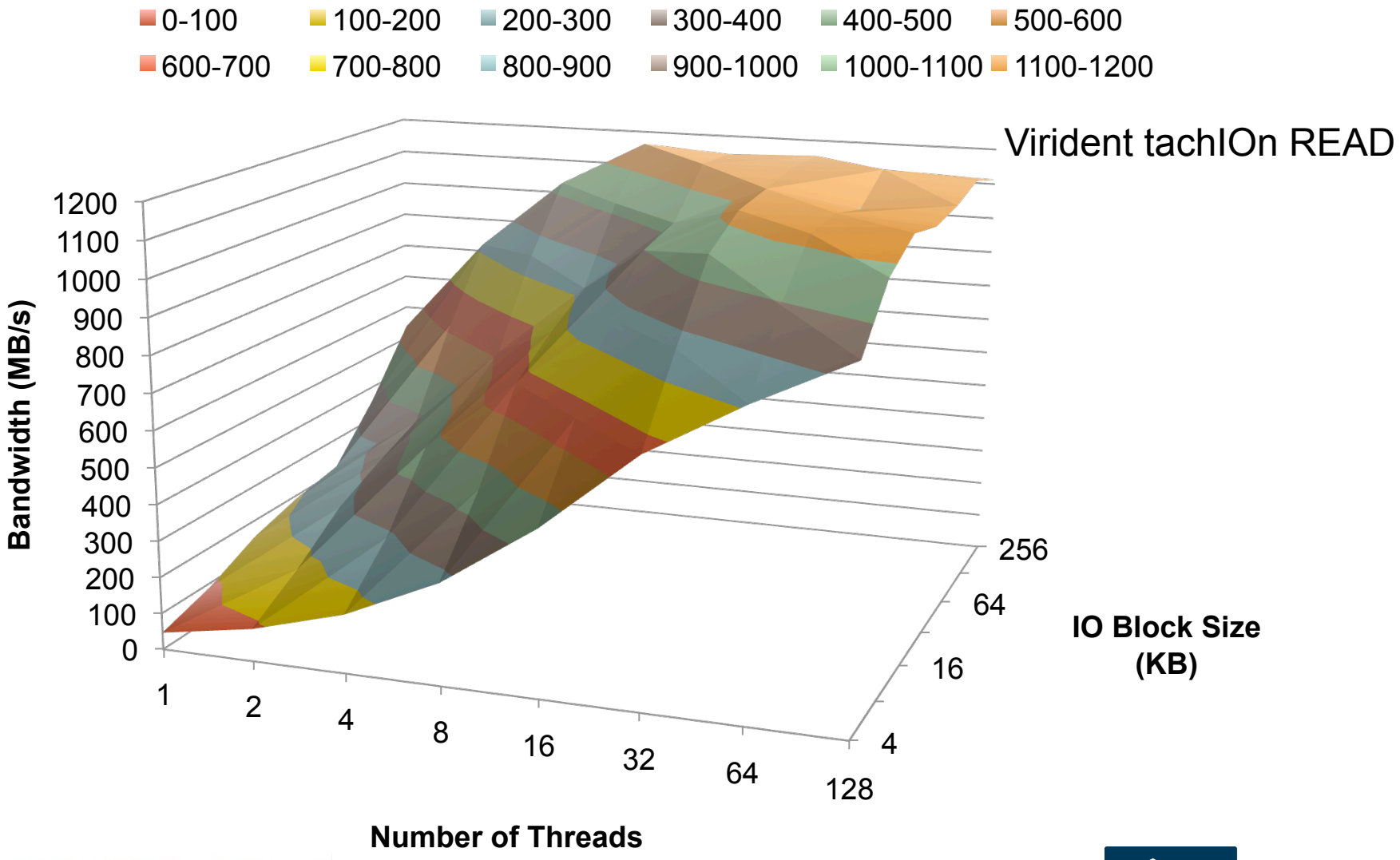
SATA Bandwidths



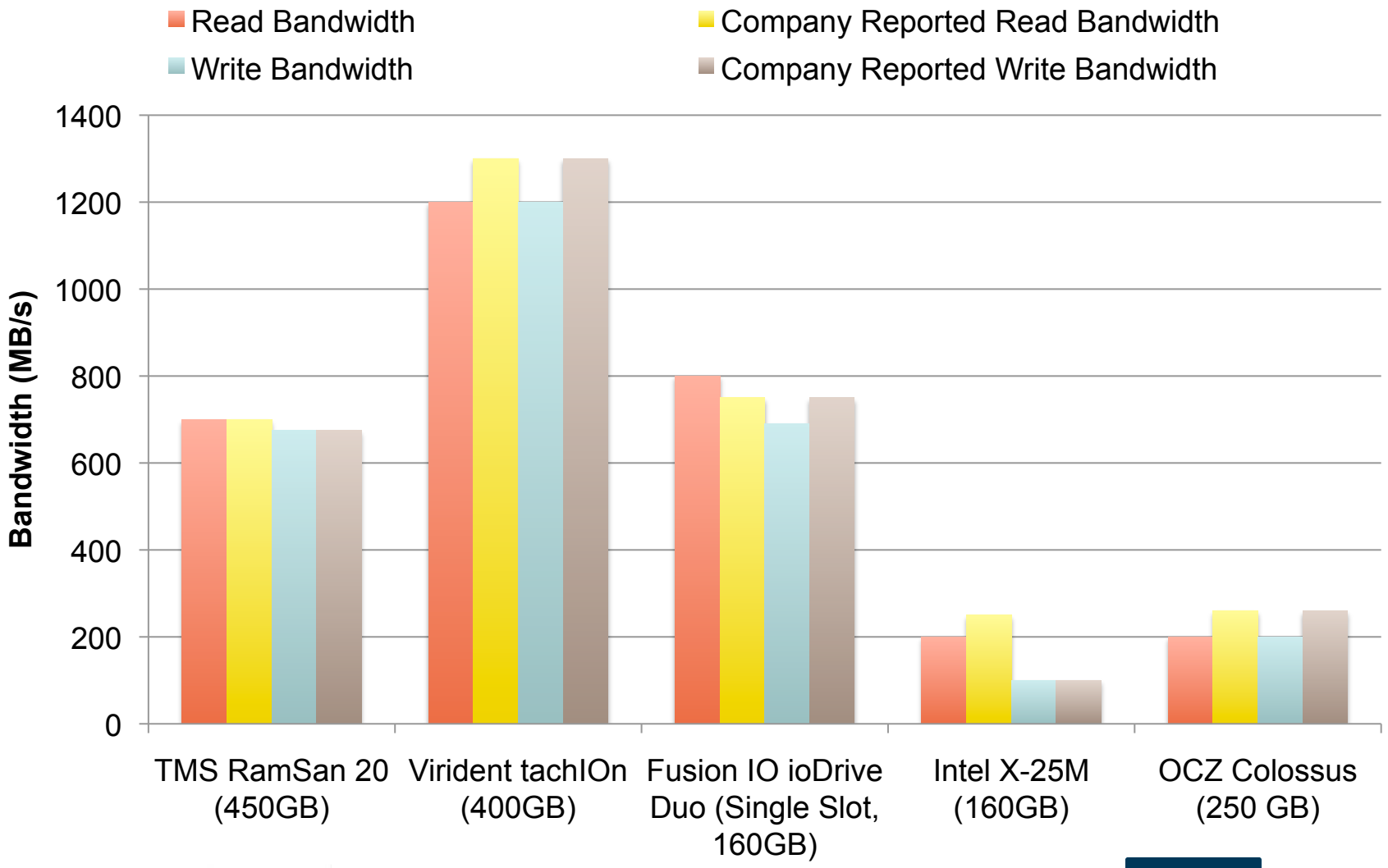
PCI-e Bandwidths



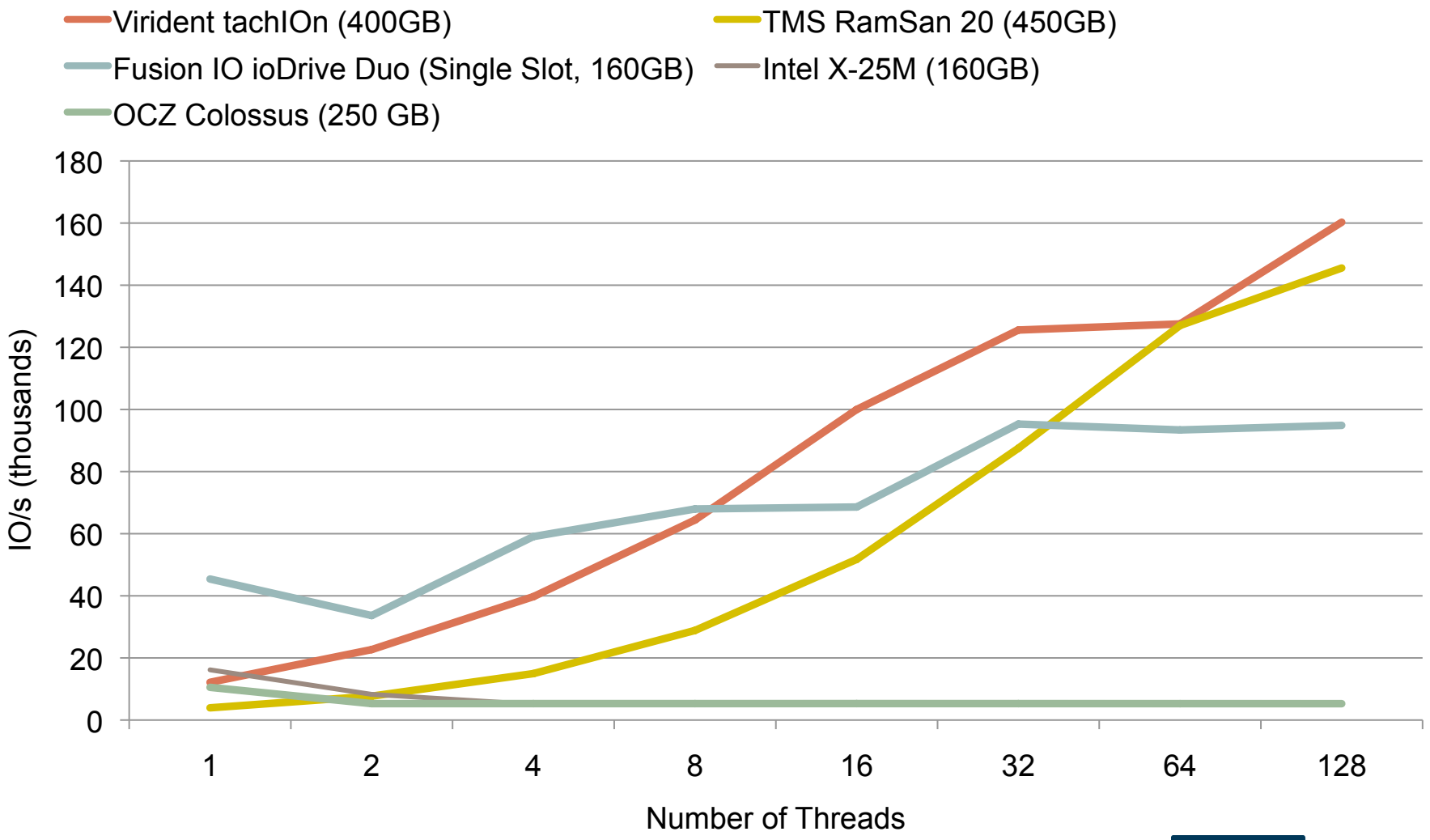
PCI-Bandwidths continued



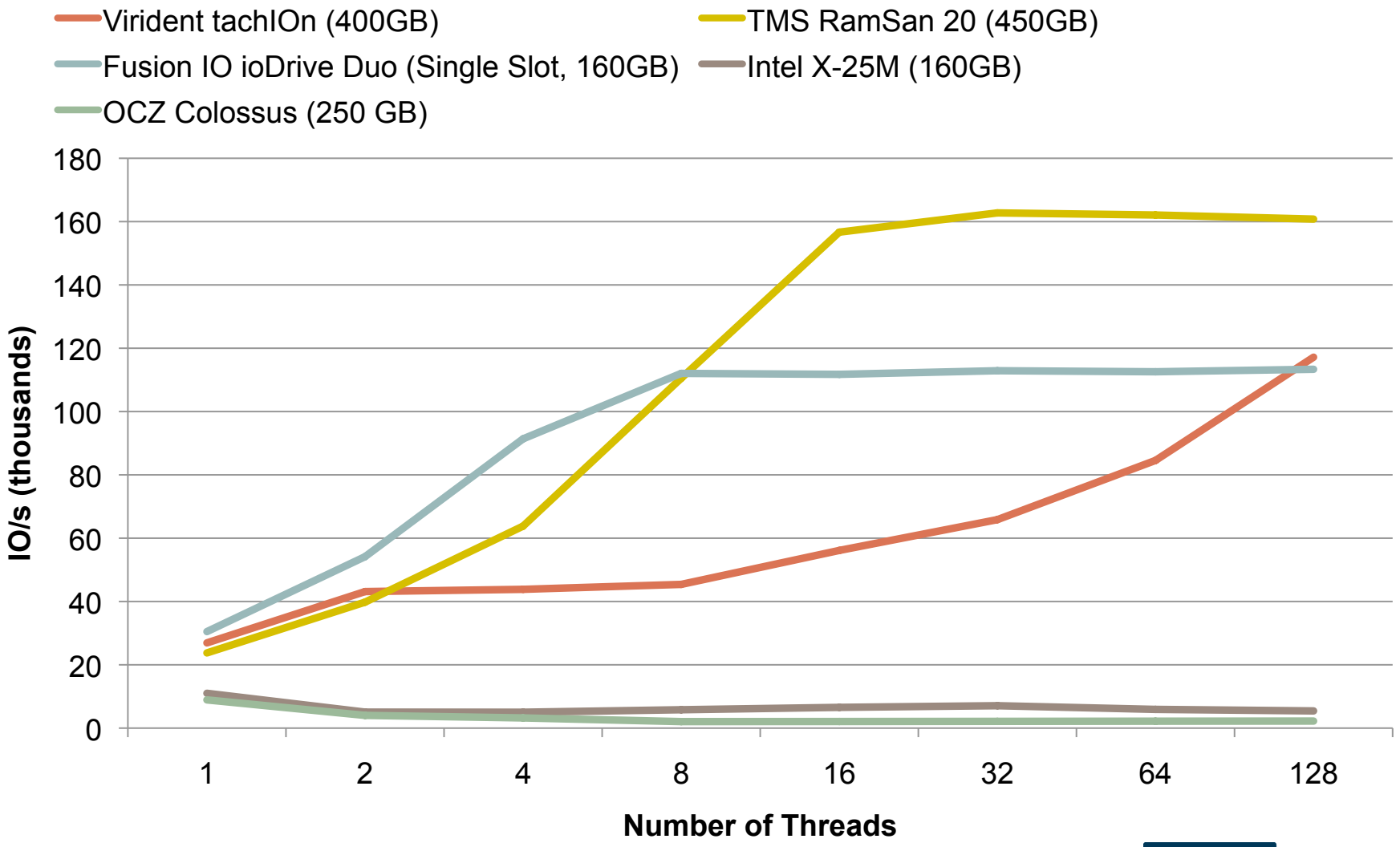
Bandwidth Summary



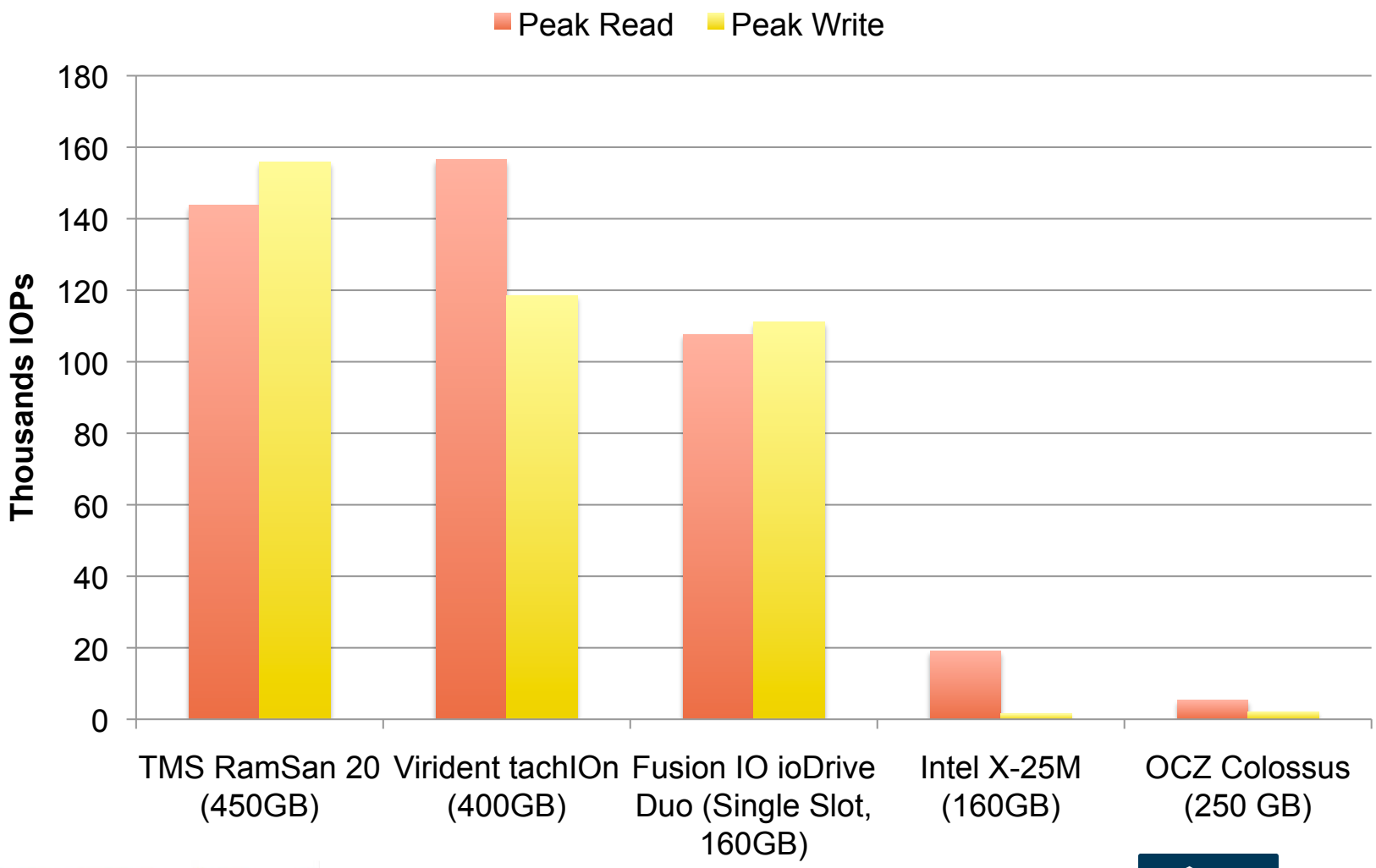
IOPS - READ


Virident tachIOn (400GB)
TMS RamSan 20 (450GB)
Fusion IO ioDrive Duo (Single Slot, 160GB)
Intel X-25M (160GB)
OCZ Colossus (250 GB)
IO/s (thousands)
180
160
140
120
100
80
60
40
20
0
1
2
4
8
16
32
64
128
Number of Threads

IOPS - Write


Virident tachIOn (400GB)
TMS RamSan 20 (450GB)
Fusion IO ioDrive Duo (Single Slot, 160GB)
Intel X-25M (160GB)
OCZ Colossus (250 GB)
IO/s (thousands)
180
160
140
120
100
80
60
40
20
0
1
2
4
8
16
32
64
128
Number of Threads

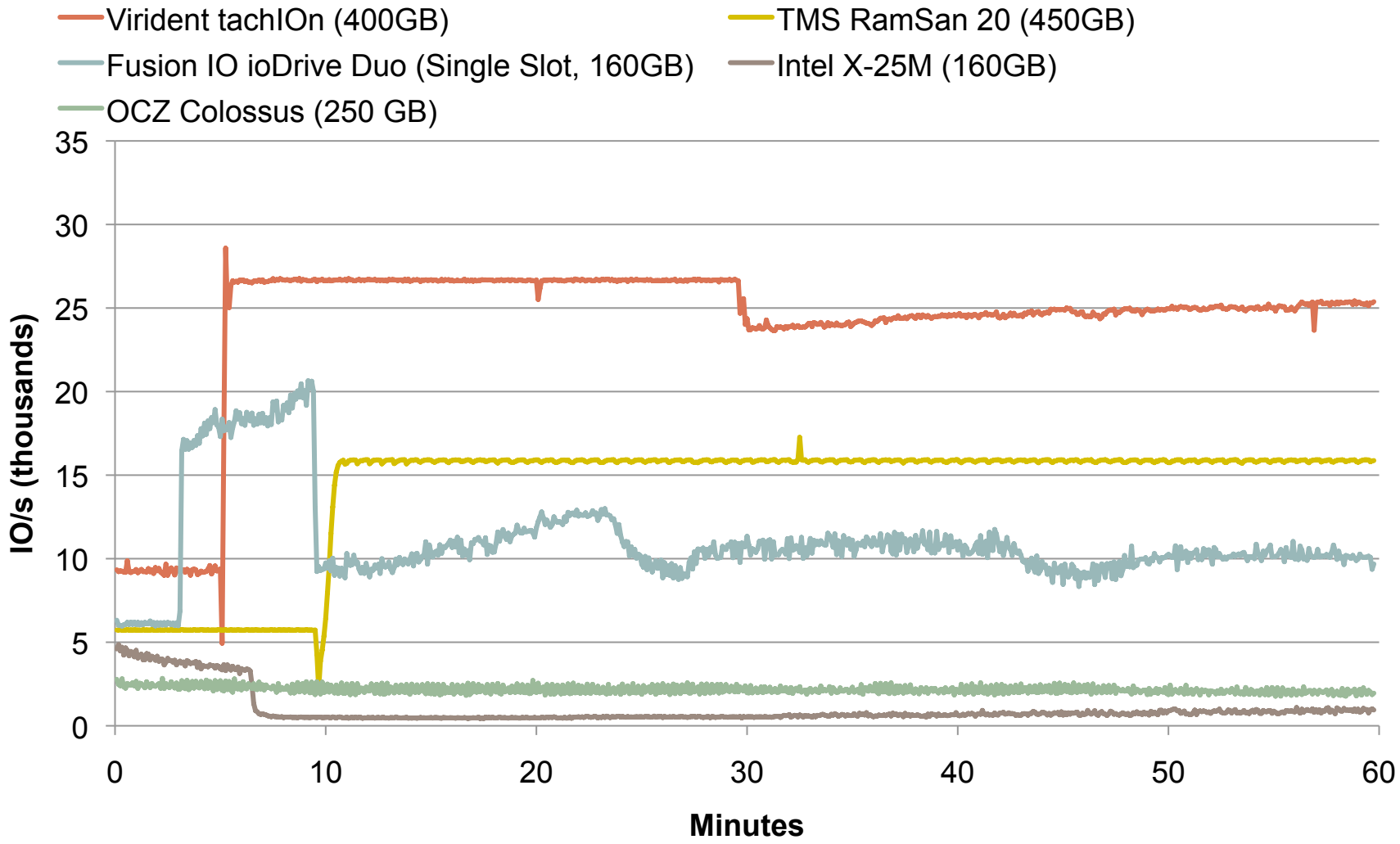
Flash Device Evaluation - IOPS



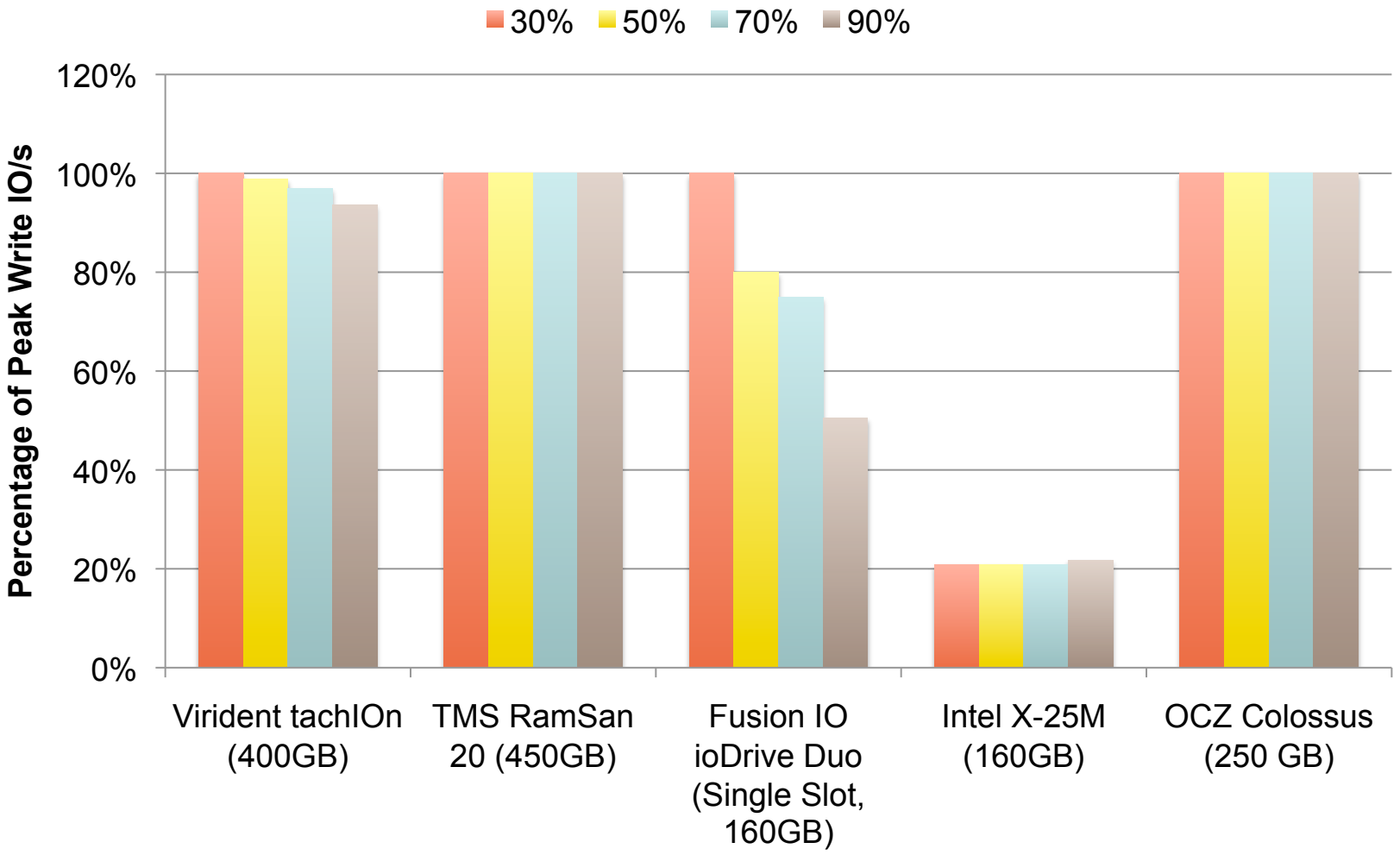
Degradation Experiment

- Create a file using
  - Cat /dev/urandom | dd
  - that fills X% of the drive X=30,50,70,90
- Using FIO randomly write to the file
  - Using 4KB blocks - IOPS
  - Using 128KB blocks - BW

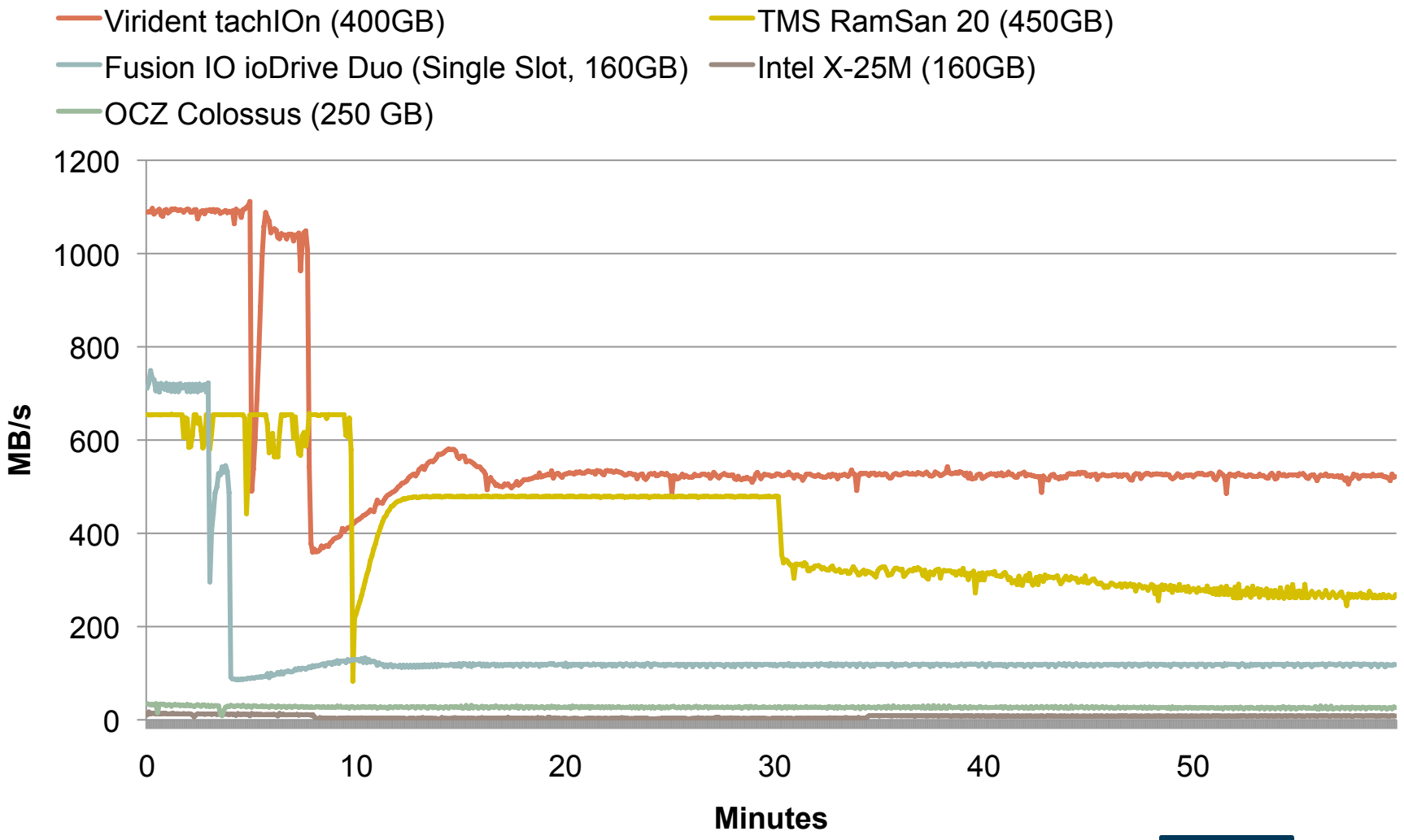
Degradation - IOPS



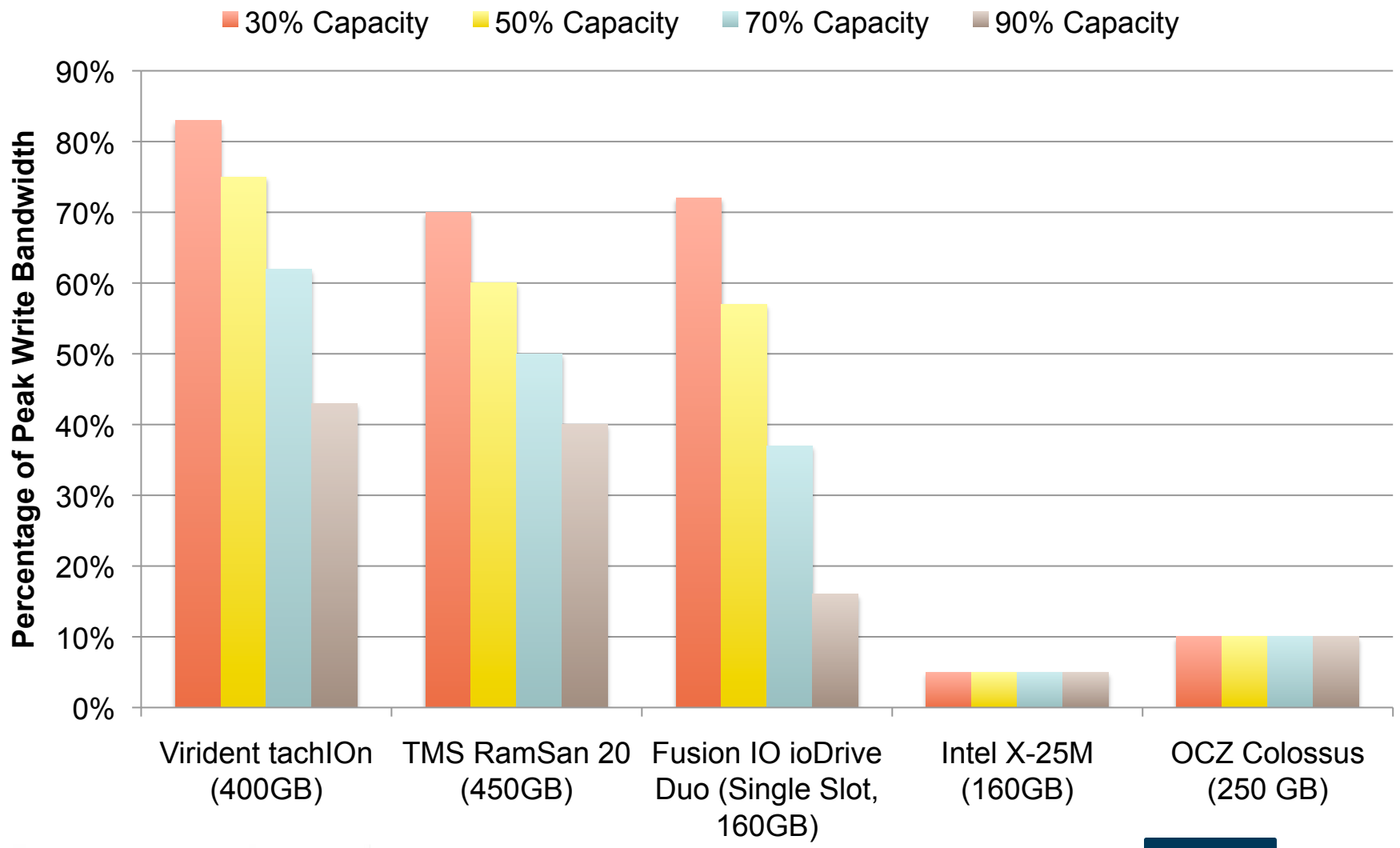
Degradation – IOPS Summary



Degradation - Bandwidth



Degradation BW Summary



Summary

- PCI devices are much more capable than the SATA ones
- For PCI read ~ write for both sequential I/O and IOPS
- It is important to test for *your* workload each device
- The PCI devices especially can be difficult to use……

Future Work

- Testing Flash with Hadoop
- Evaluating various new storage technologies. PCM etc
- Explore other uses for flash
  - Metadata storage

Combining Flash with Hadoop