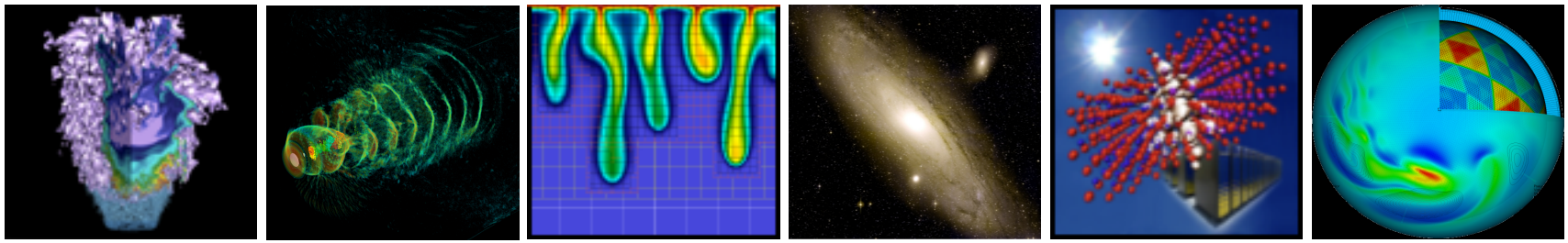


Performance Analysis of Flash Storage Devices and their Application in High Performance Computing

Nicholas J. Wright

With contributions from

R. Shane Canon, Neal M. Master, Matthew Andrews, and Jason Hick

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

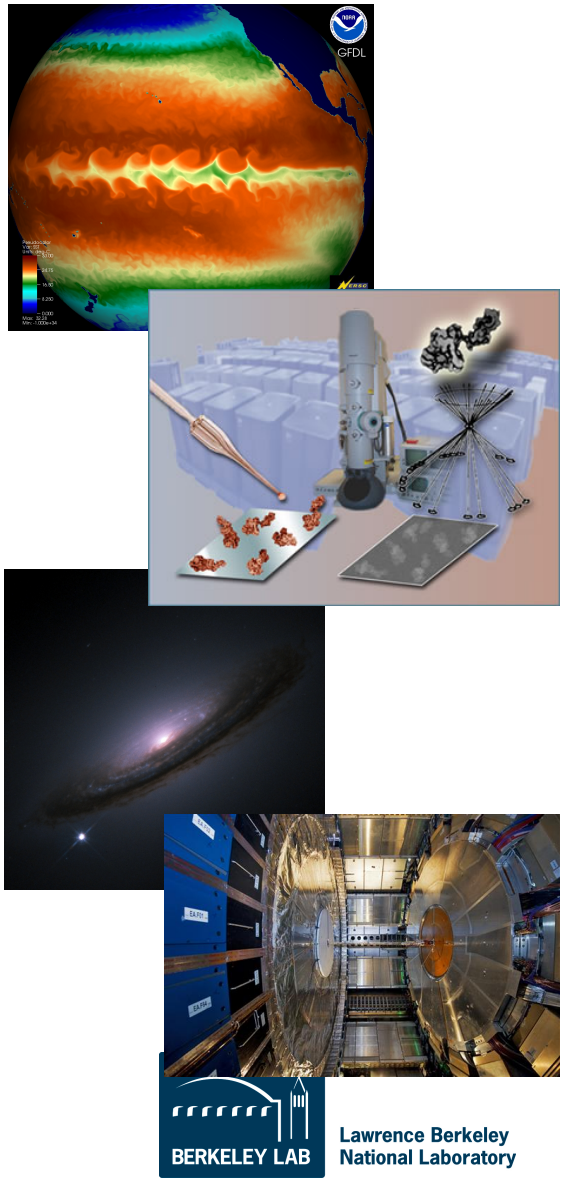
BERKELEY LAB Lawrence Berkeley National Laboratory

Outline

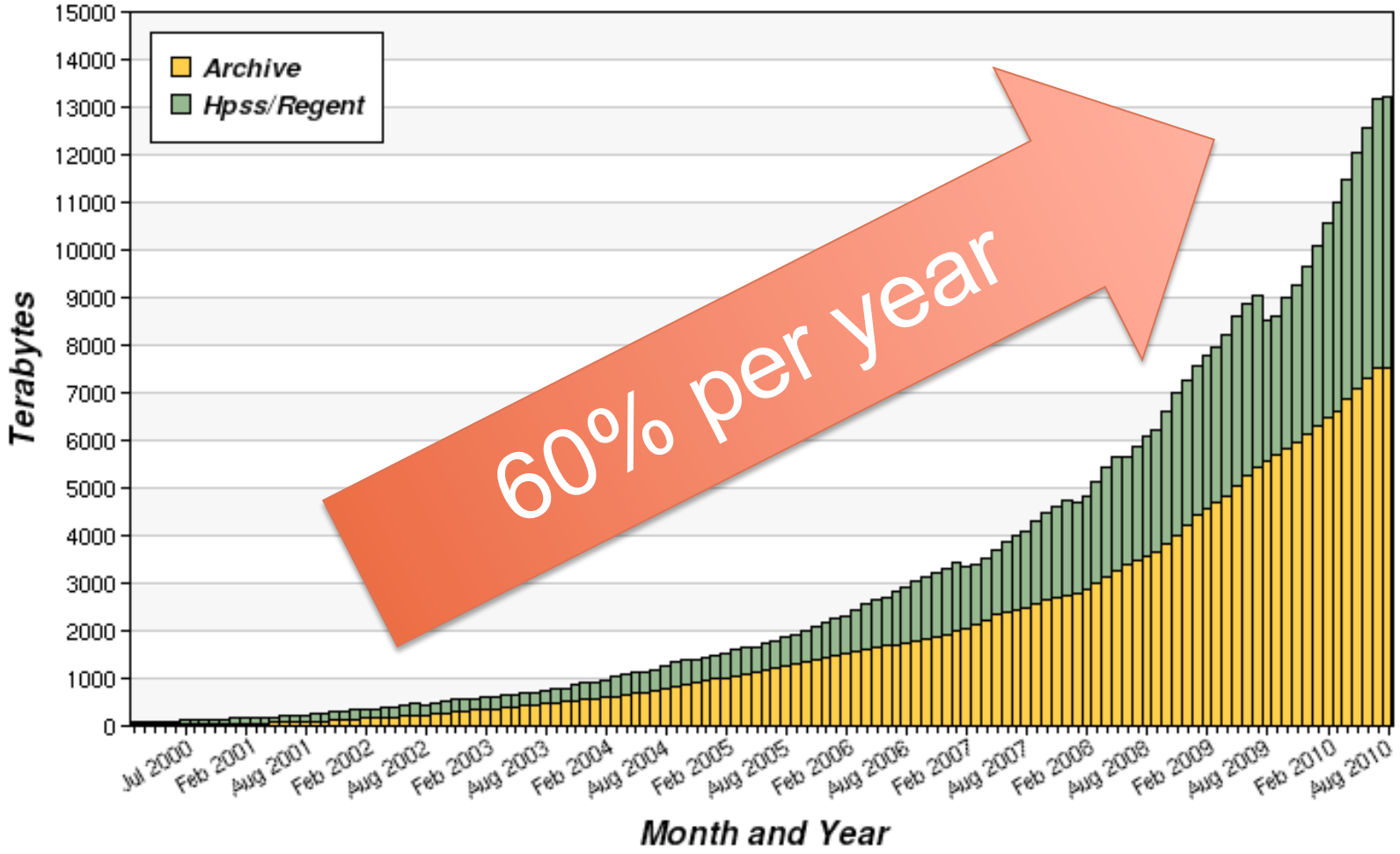
- Why look at flash memory ?
- Performance evaluation of individual flash devices
- What applications is flash going to be good for ?
- Flash in a parallel filesystem
- Summary

Data Driven Science

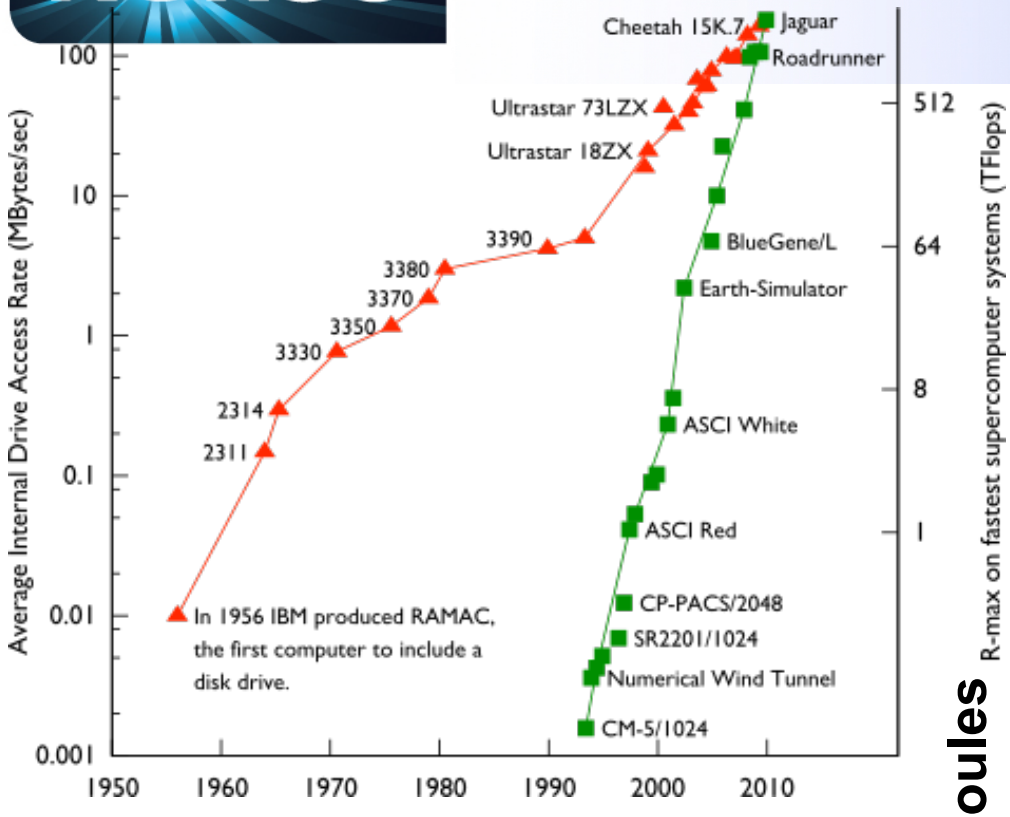
- Ability to generate data is challenging our ability to store, analyze, & archive it.
  - Some observational devices grow in capability with Moore's Law.
  - Data sets are growing exponentially.
- Petabyte (PB) data sets soon will be common:
  - *Climate:* next IPCC estimates 10s of PBs
  - *Genome:* JGI alone will have .5 PB this year and double each year
  - *Particle physics*: LHC projects 16 PB / yr
  - *Astrophysics*: LSST, others, estimate 5 PB / yr
- Redefine the way science is done?
  - One group generates data, different group analyzes
    - 1st Climate 100 paper from a different group than the one collected the data



Data Trends at NERSC



I/O Performance Challenges

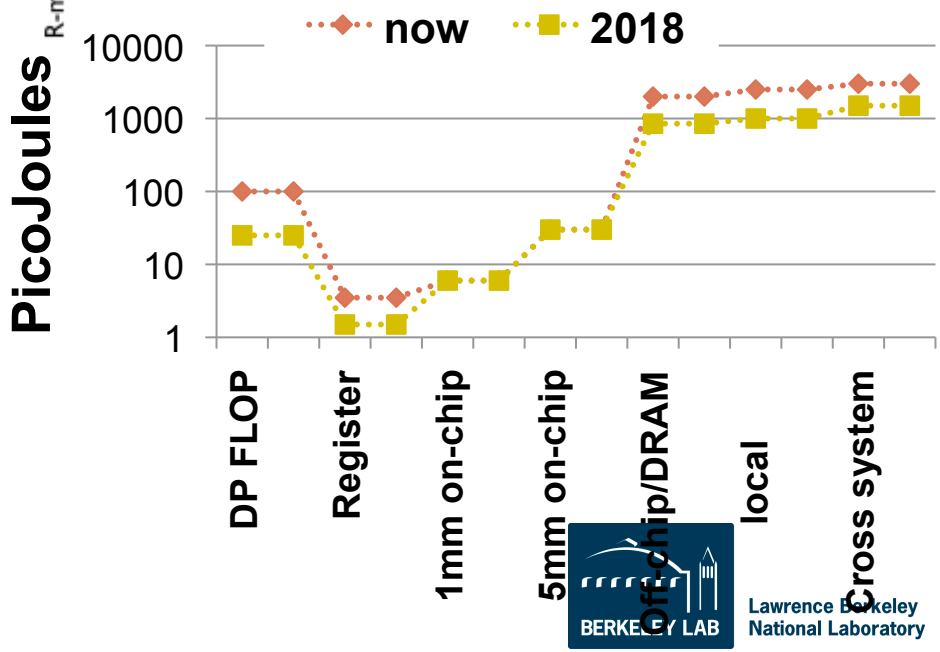


Performance Crisis #1

- **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 !

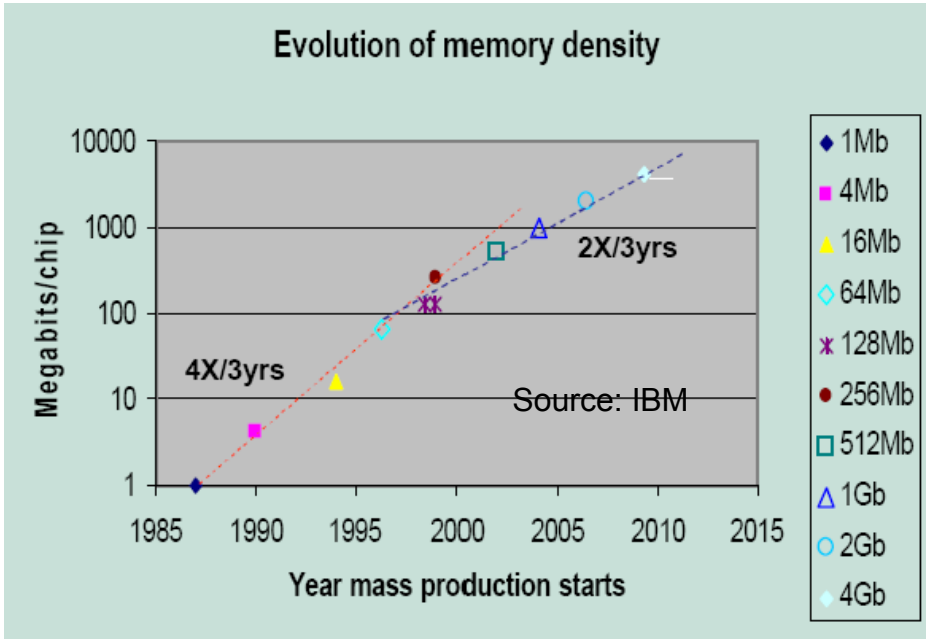


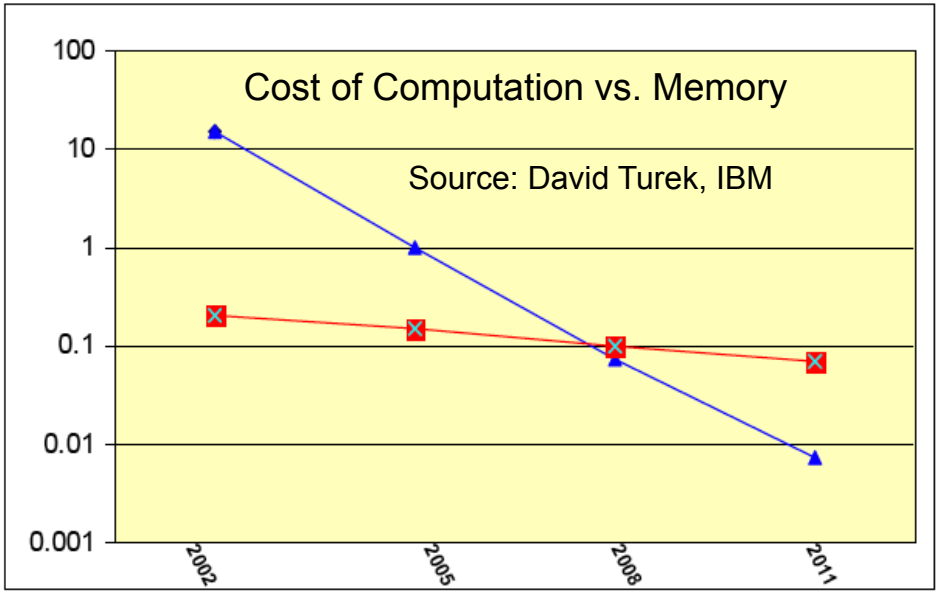
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

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

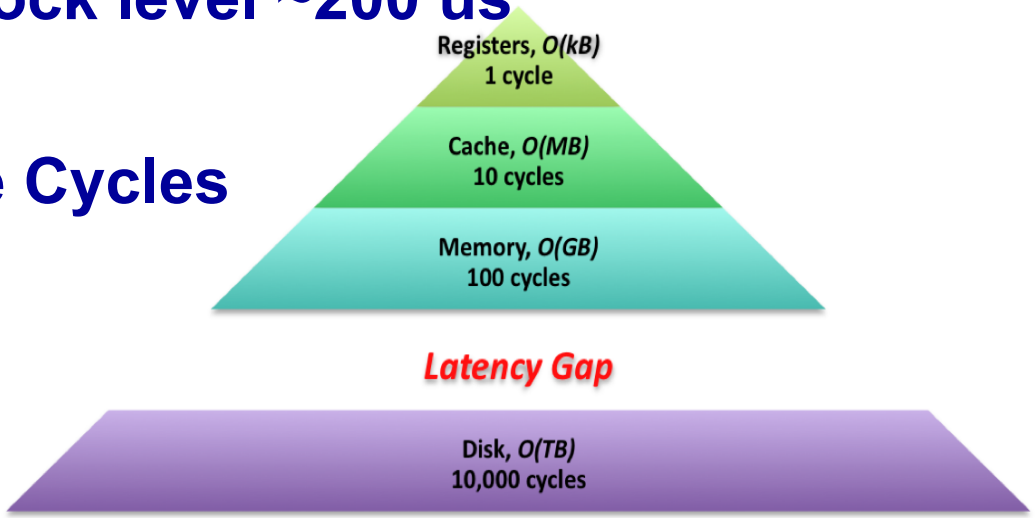
Hardware Trends are exacerbating the issue

- Data volumes exploding!
- Memory Capacity per Flop decreasing
- I/O Bandwidths not keeping pace
- Data movement is expensive
- Will NVRAM save the day ?
  - Let's evaluate Flash Storage !

Flash Memory - Ubiquitous

Flash – What is it good for?

- Read - Word level ~20 us
- Write - Erase/Write - block level ~200 us
- $/GB
- Finite Number of Erase Cycles
- 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

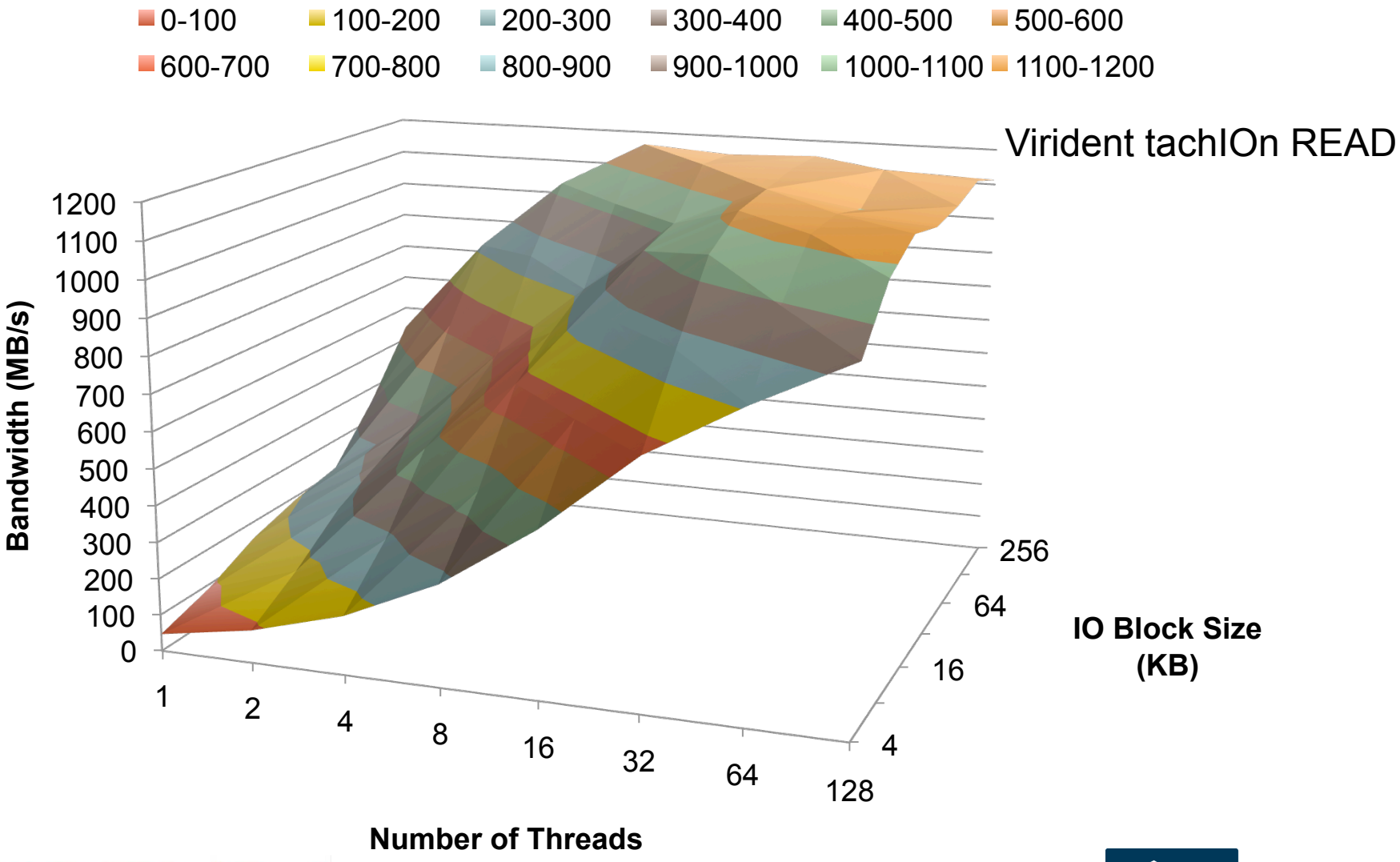
- 3 PCI-e SLC
  - Virident tachIOn 400GB 8x
  - FusionIO ioDrive Duo 2x 160GB 4x
  - Texas Memory Systems RamSan-20 450GB 4x
- 2 SATA MLC
  - Intel X-25M 160GB
  - OCZ Colossus 250GB

Performance Analysis of Commodity and Enterprise Class Flash Devices.
Neal M. Master, Matthew Andrews, Jason Hick, Shane Canon & Nicholas J. Wright
PDSI Workshop, Supercomputing 2010, New Orleans.

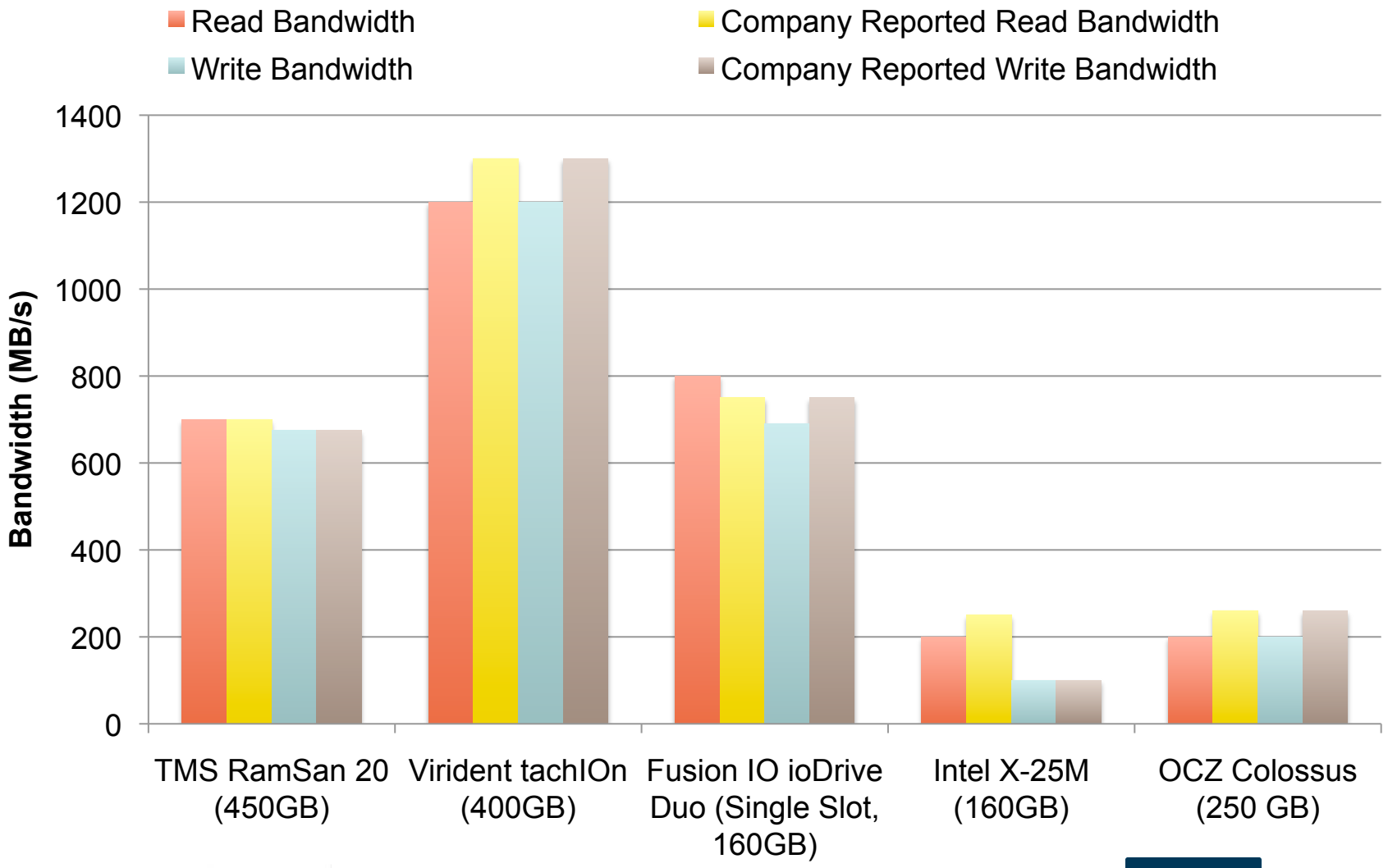
IOZone Experiments

- **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)

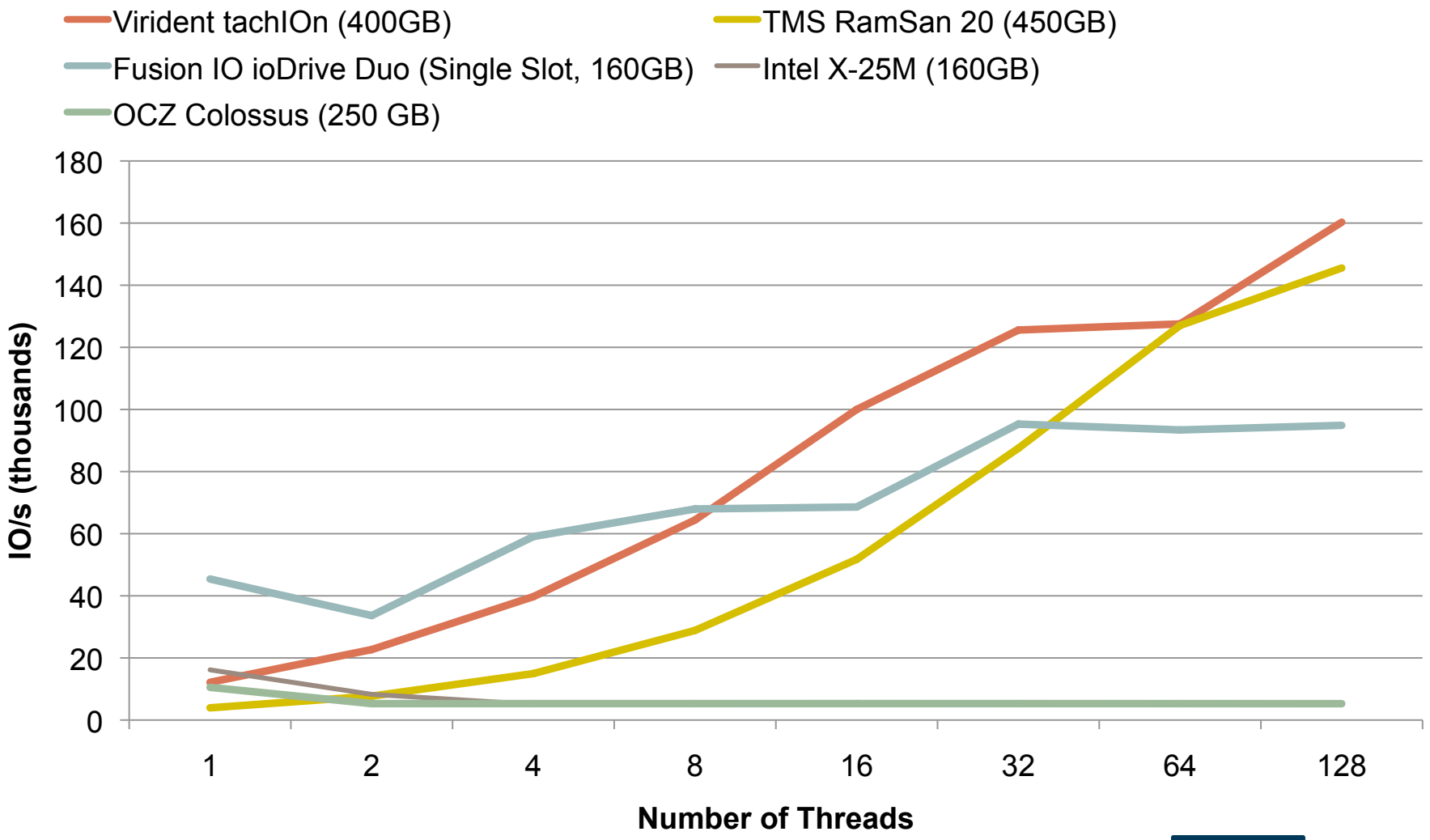
PCI-Bandwidths



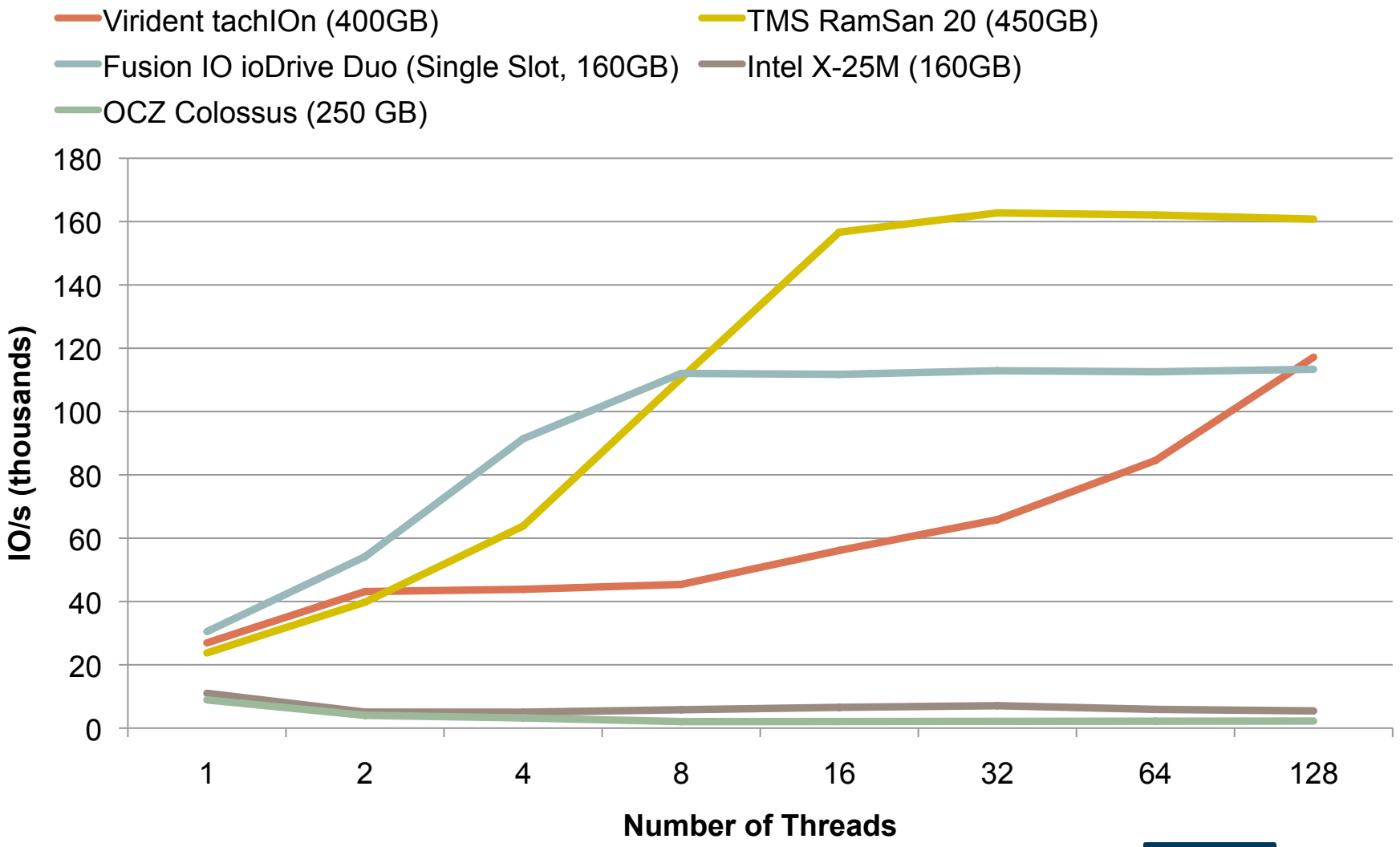
Bandwidth Summary



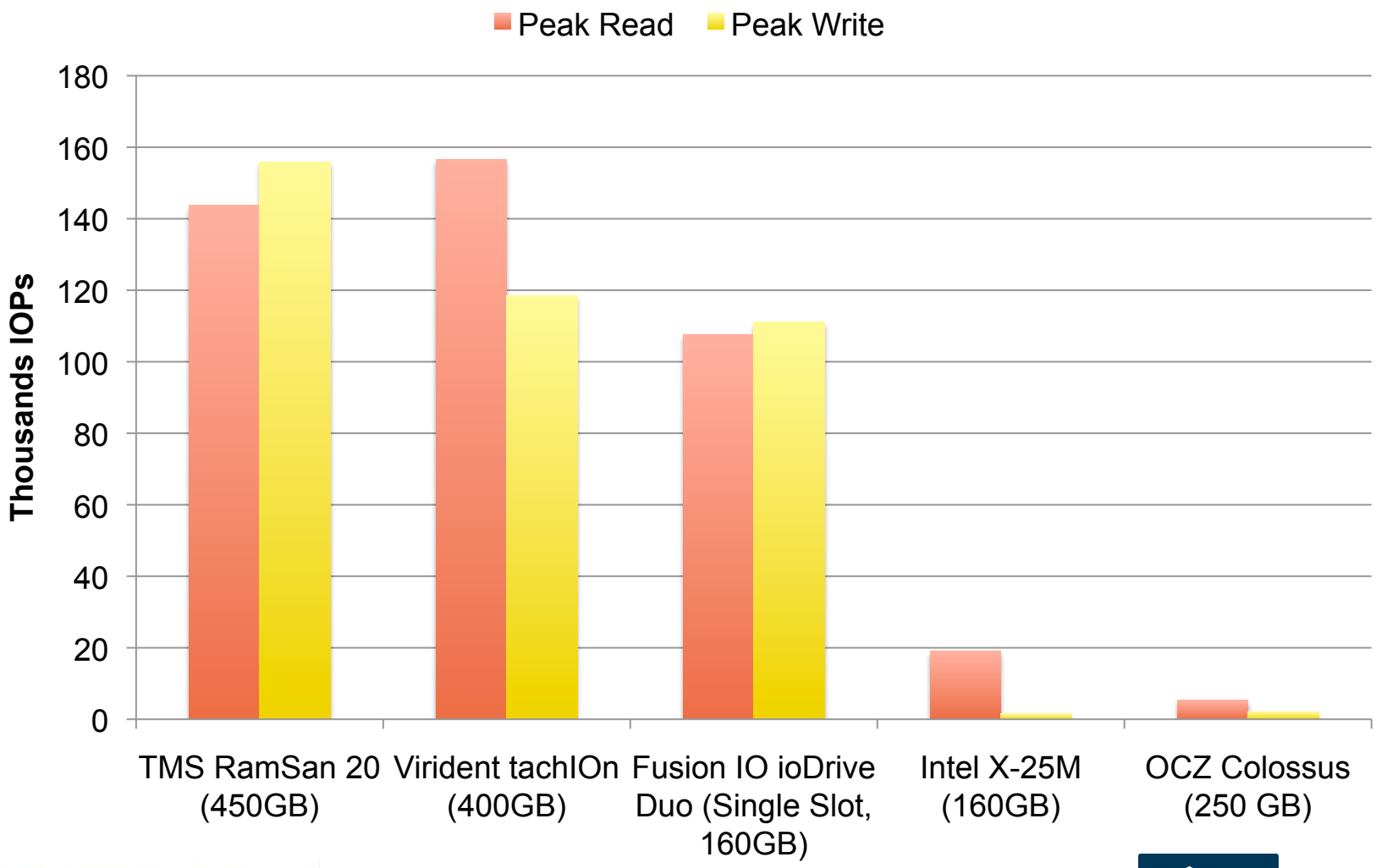
IOPS - READ


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

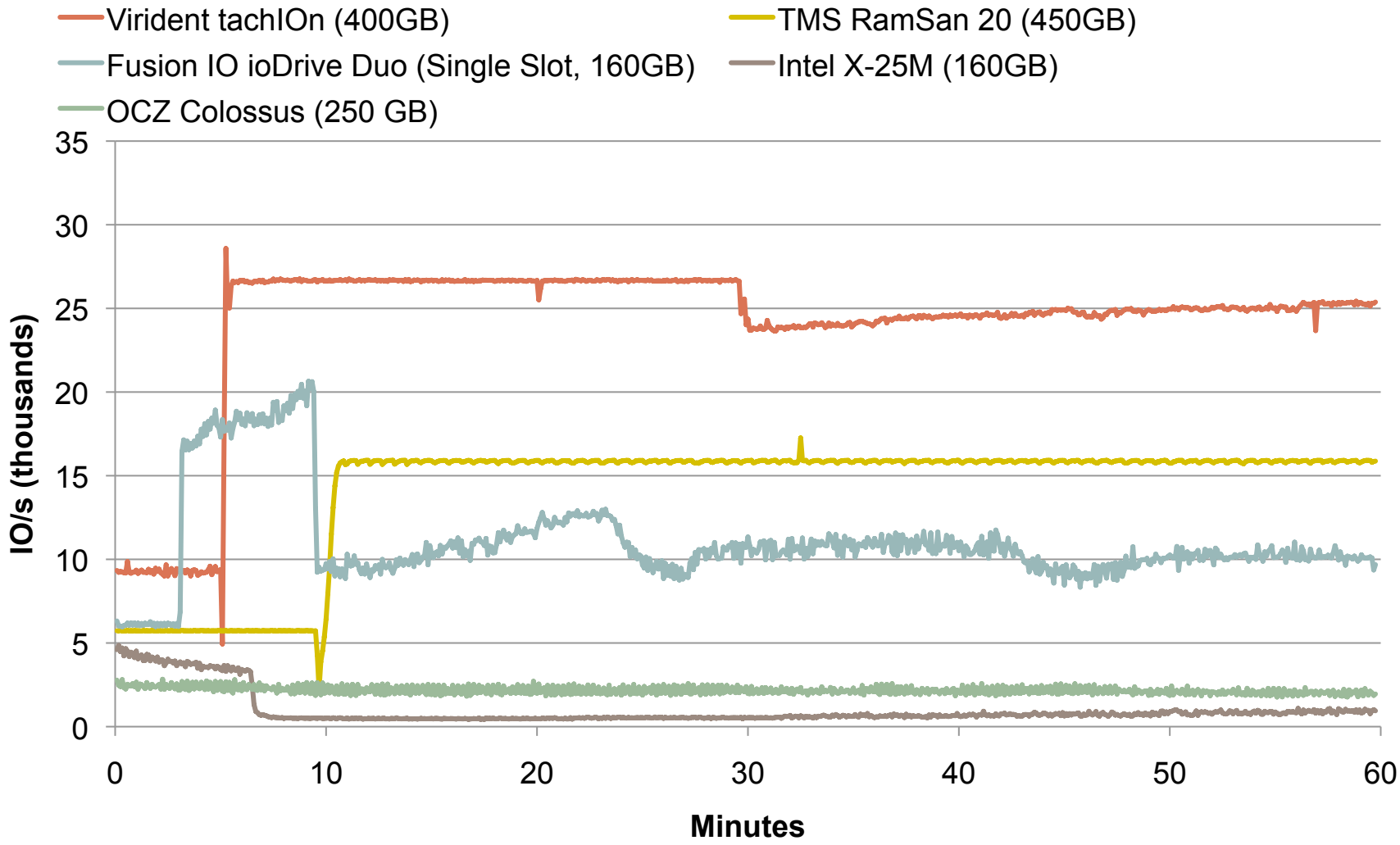
Flash Device Evaluation - IOPS



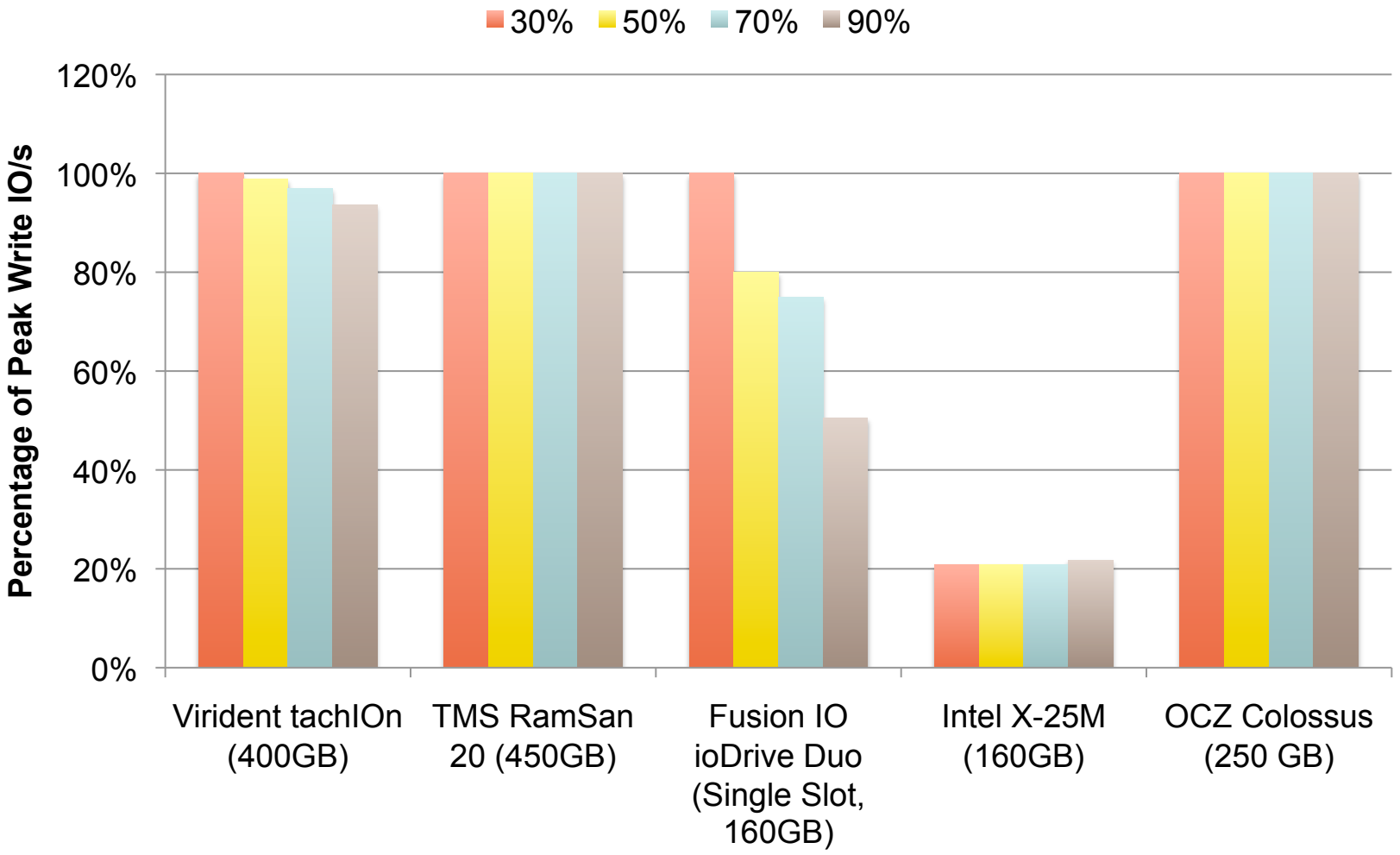
Degradation Experiment

- 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

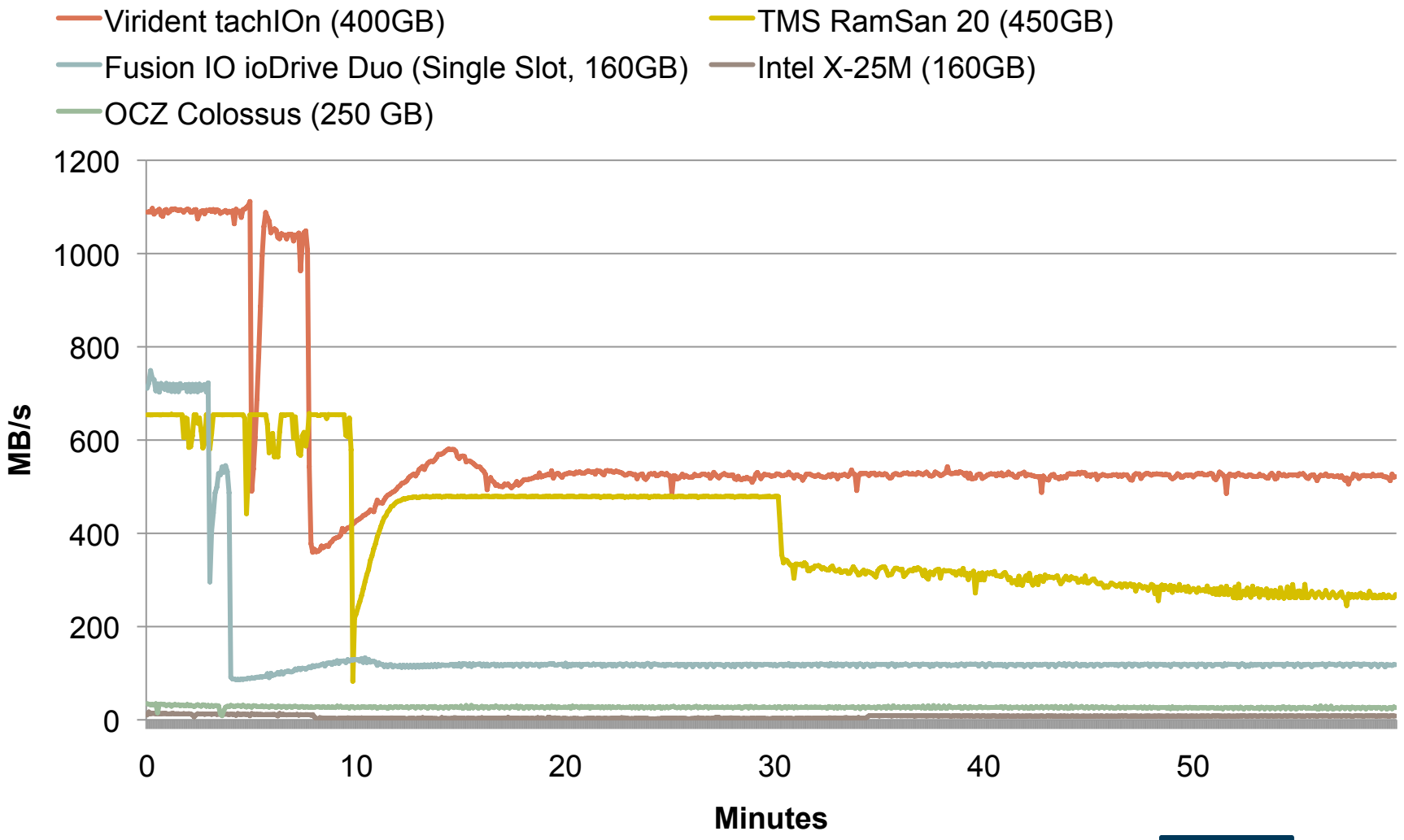
Degradation - IOPS



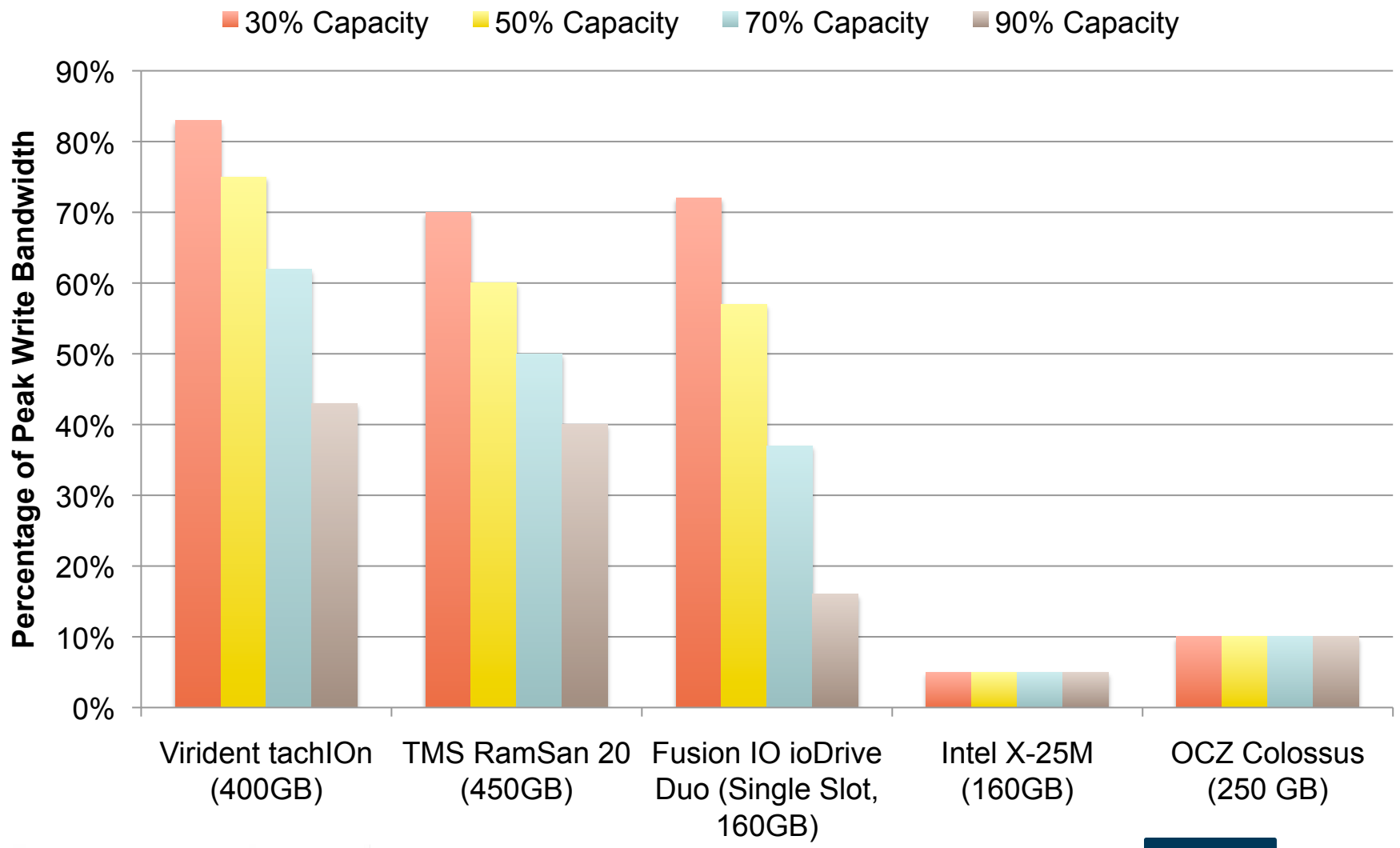
Degradation – IOPS Summary



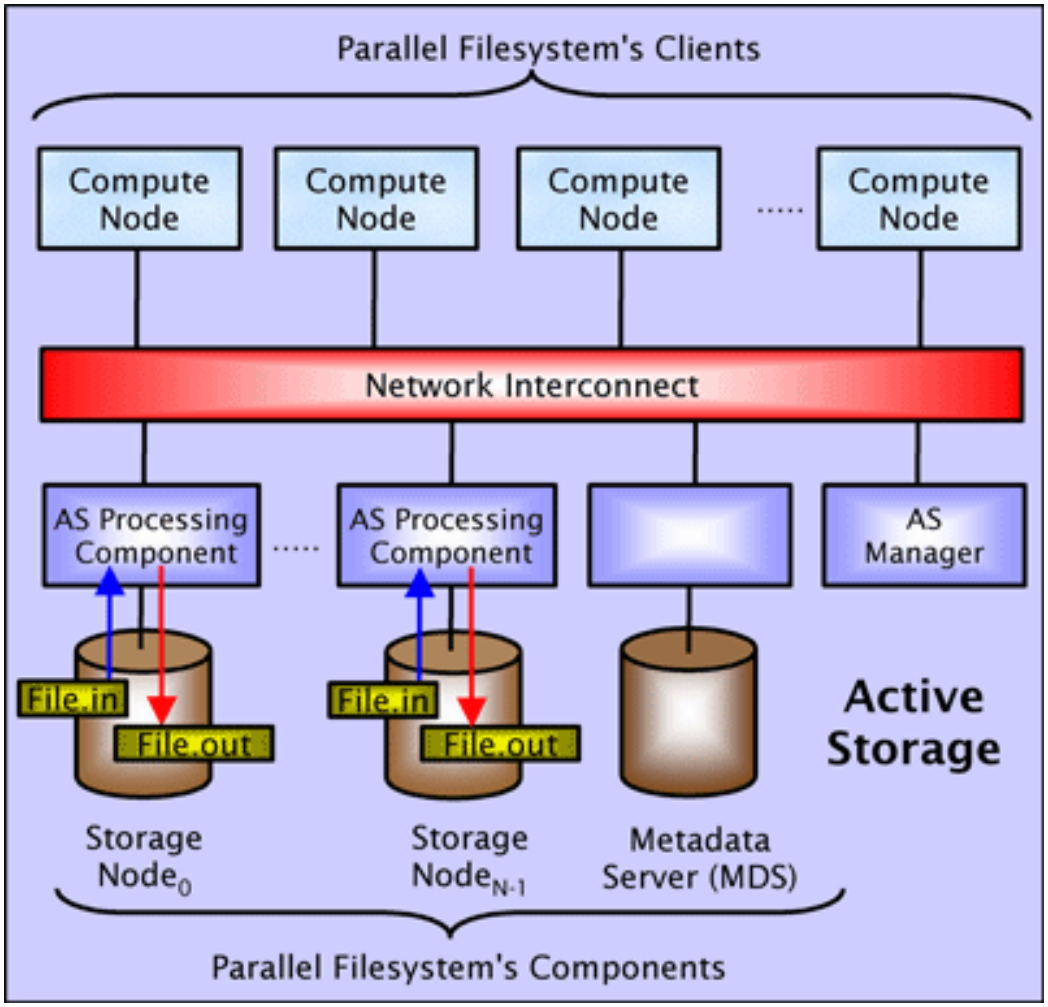
Degradation - Bandwidth



Degradation BW Summary



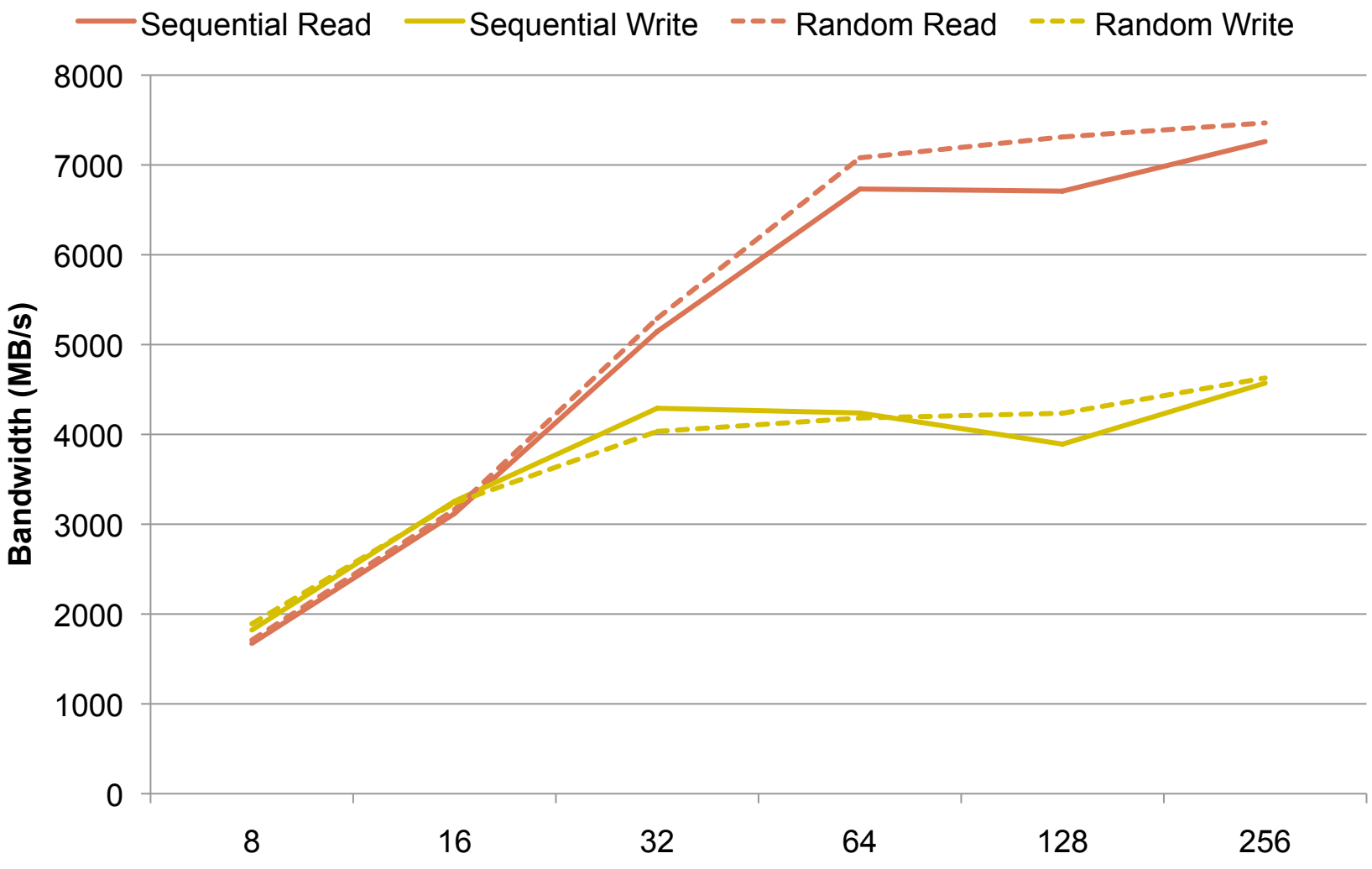
Parallel Filesystems



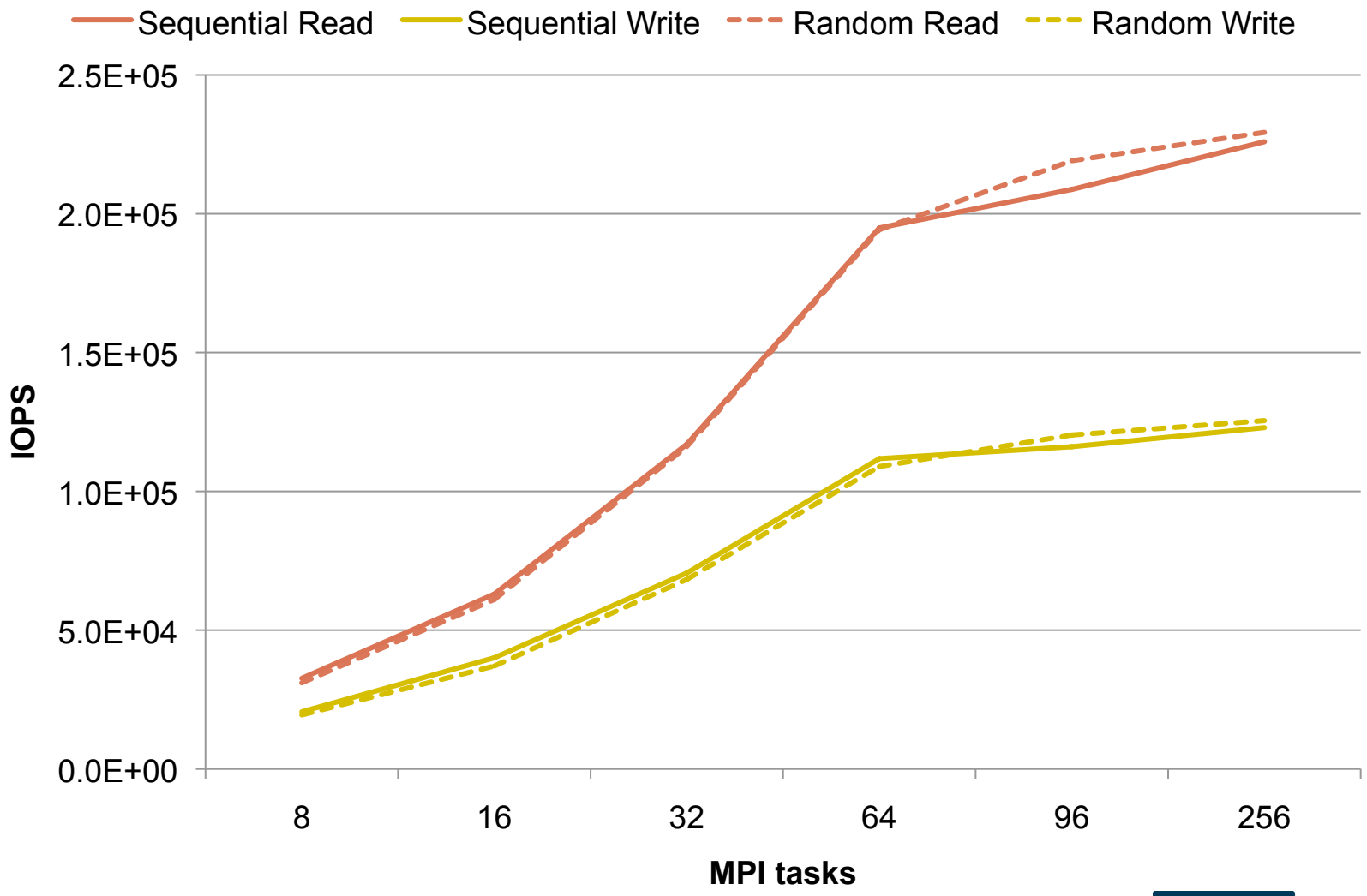
GPFS Flash Filesystem

- 8 Virident Devices
  - 2 per node dual socket Nehalem 2.67 GHz 24 GB QBR-IB
  - v.1.0 Virident Driver Software
- GPFS v 3.2
  - 4 NSD servers – 2 cards per server
  - 256K block size (default)
  - Metadata stored with data
  - Scatter block assignment algorithm
- All measurements made with IOR

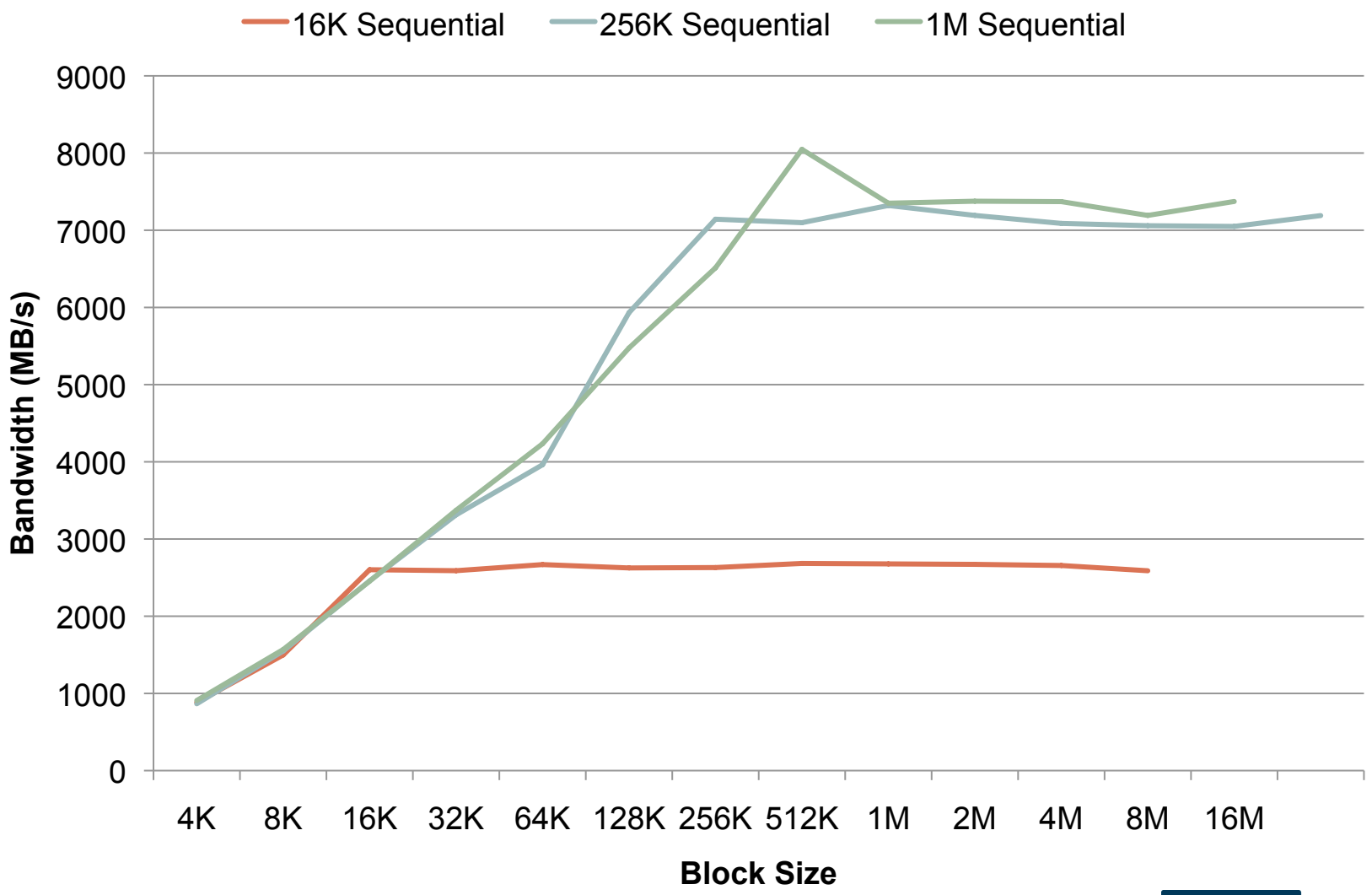
GPFS: Bandwidth Measurements



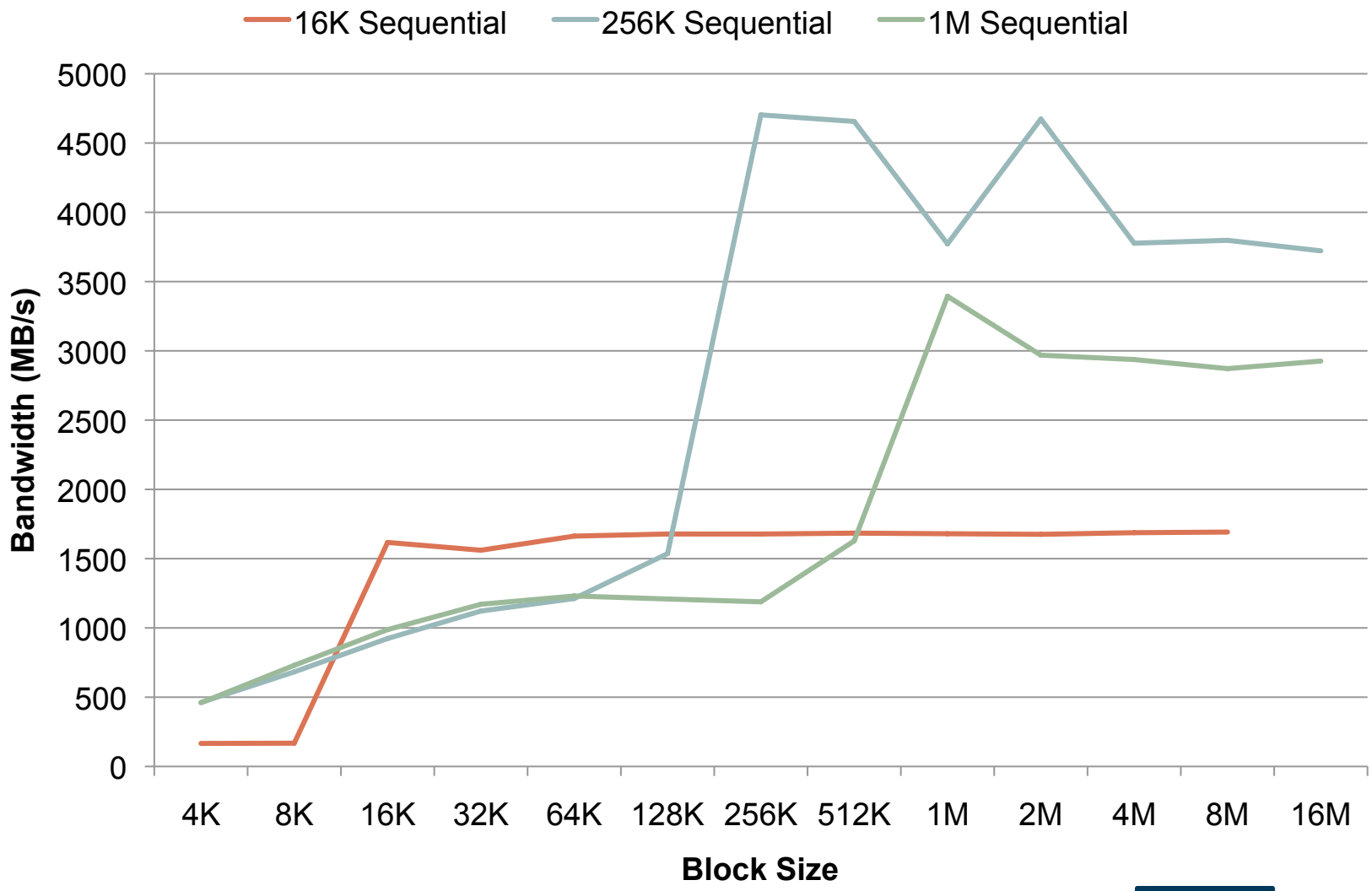
GPFS: IOPS Measurements



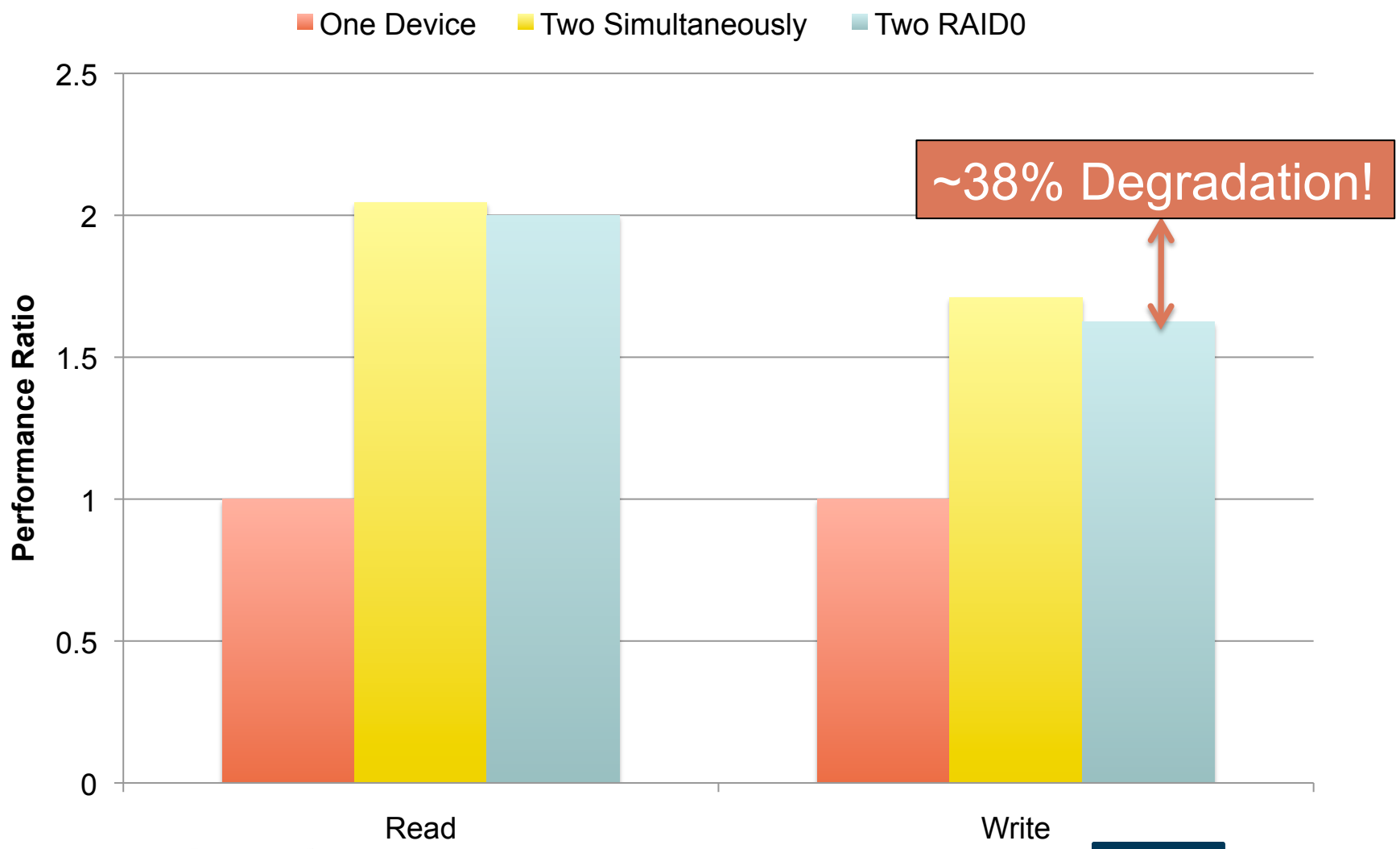
GPFS block size variation: Read



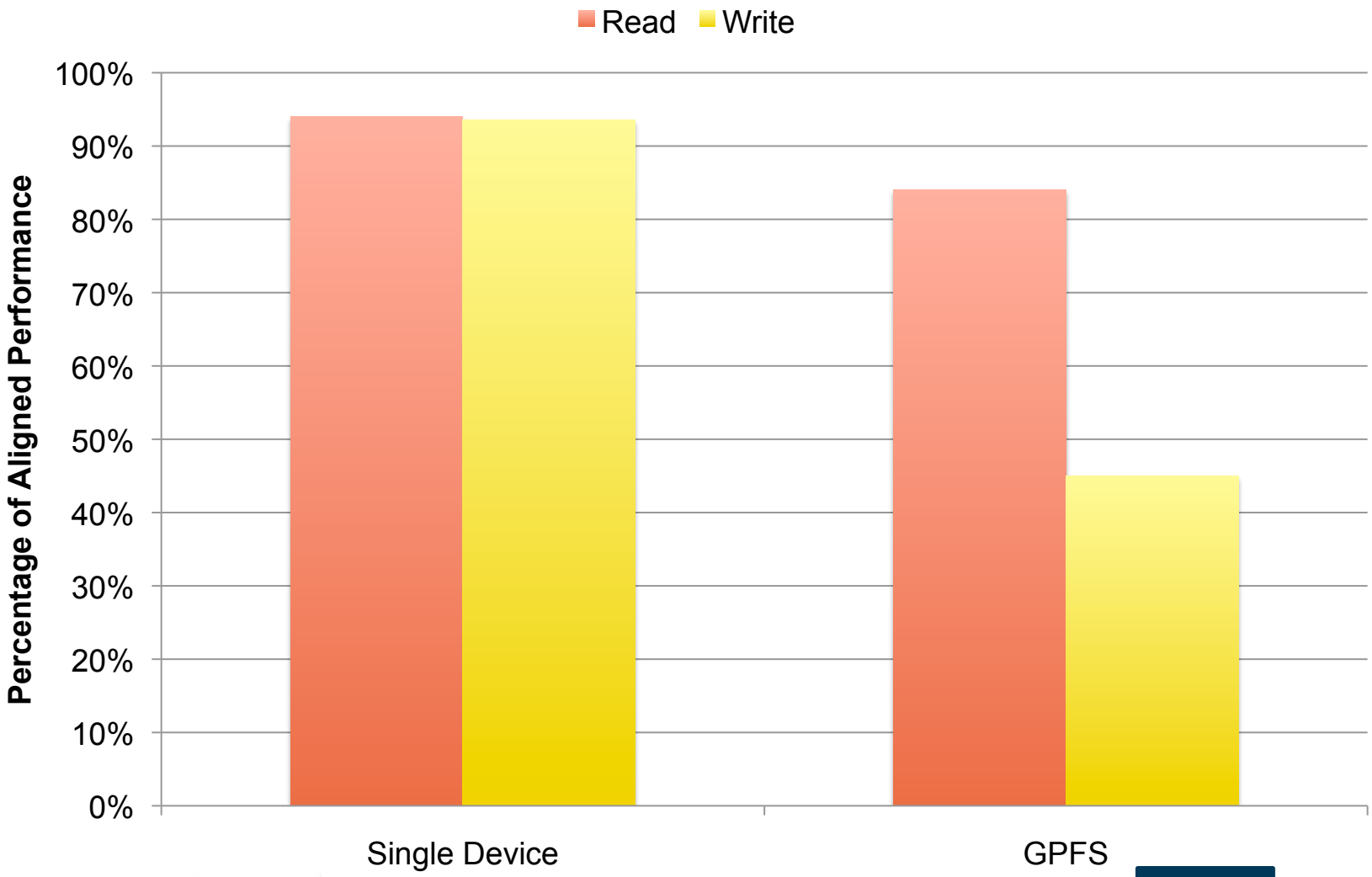
GPFS block size variation: write



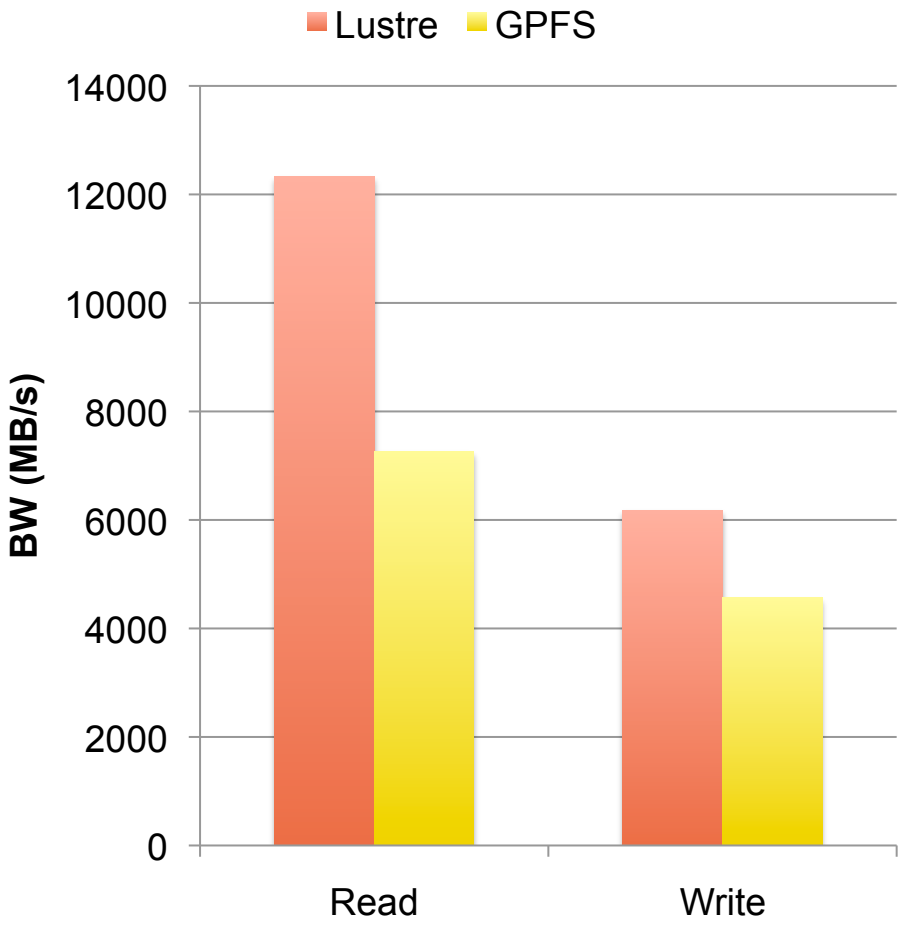
Performance for Two Devices Simultaneously

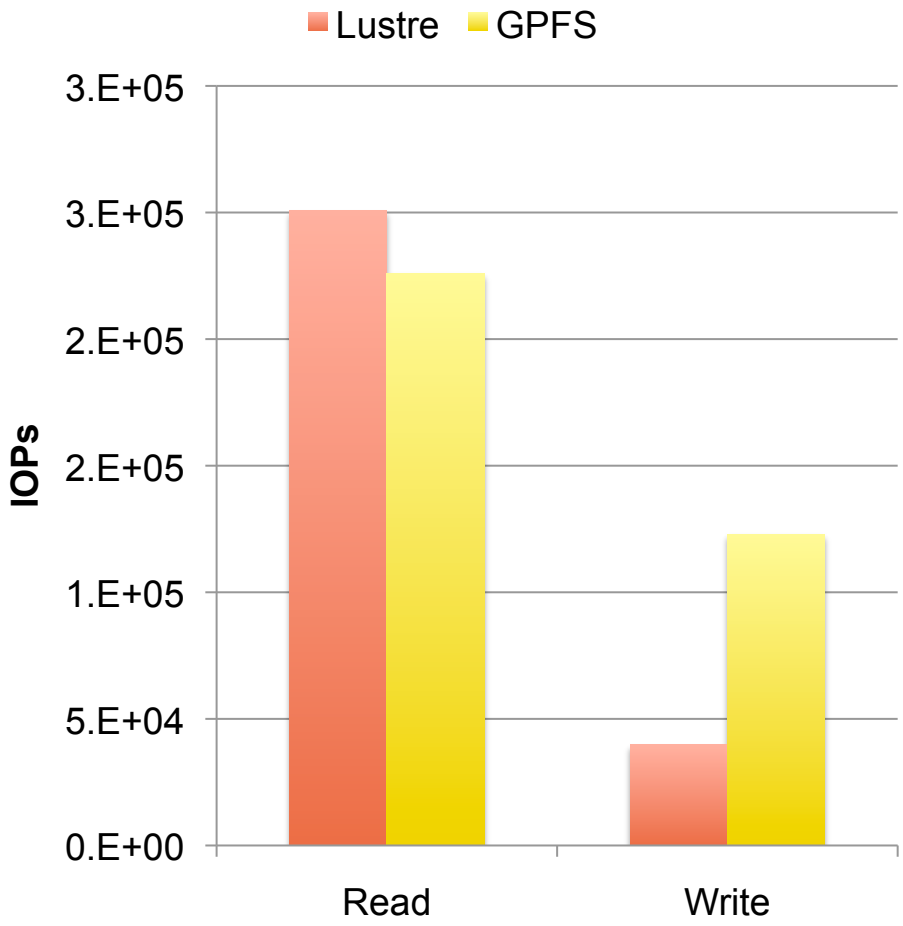


GPFS Unaligned I/O Performance



Lustre - GPFS Comparison





Applications for Flash?

Device	Price in $	Capacity in GB	Bandwidth / GB/s	IOPS	$/GB	$/GB/s	$/IOP
SATA MLC FLASH	120	64	0.2	8600	$1.88	**$600**	**$0.01**
SATA SLC FLASH	740	64	0.2	5000	$11.56	**$3,700**	**$0.15**
PCI SLC FLASH	11500	640	1.2	140000	$17.97	**$9,583**	**$0.08**
SATA HDD	80	2000	0.07	90	**$0.04**	$1,143	$0.89
High-Perf Array	250000	240000	5	100000	**$1.04**	$50,000	$2.50

Graph500: Traversing massive graphs with NAND Flash

Roger Pearce[1,2] Maya Gokhale[1] Nancy M. Amato[2]

[1]Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

[2]Department of Computer Science and Engineering

Texas A&M University

June 2011

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE- AC52-07NA27344. LLNL-PRES-487136.

Graph500 Implementation

Semi-External

- $O(|V|)$ data can fit into main memory, $G = (V, E)$
- Read-only from NAND Flash, output and algorithm data kept in-memory
- e.g. store graph in external memory, keep BFS data in memory

Asynchronous Traversal Technique *[SC 2010]*

- Exploit fine-grained path parallelism
- Tolerate data latencies to graph storage (NAND Flash)
- Re-order vertex visitation to improve page-level locality
- Allow over-subscription of thread-level parallelism (256 threads) to maximize NAND Flash IOPS
- Graph stored as CSR on Flash device; read-only

Experimental Setup

Kraken's hardware:

- single node, 32-core Opteron(tm) Processor 6128 @ 2.0Ghz
- 512 GB DRAM
- 4x 640GB Fusion-io MLC; Software RAID0
- Red Hat Enterprise Linux 5.6
- Approximate system cost $71K
  - $25K for base system
  - $46K for 2.56TB of Fusion-io NAND Flash

Result:

Using Fusion-io: 8x larger with 50% performance loss over DRAM only

DRAM + Fusion-io: Scale 34, 55.6 MTEPS
DRAM Only: Scale 31, 104.6 MTEPS

Summary

- Bandwidths per device are impressive
  - But cf. RAID'd set of regular HDD's
- IOP's numbers are very impressive
  - 100x HDD – RAID'ing won't help here
  - Large numbers of threads needed to saturate (Pearce, Gokhale & Amato SC10)
- Previous I/O pattern can effect performance
- Parallel Filesystem Software needs Tweaking to use with Flash
  - Read BW - OK - Write BW – 40% 'peak'
  - Read IOPS – 18% 'peak' - Write IOPS 13% 'peak'

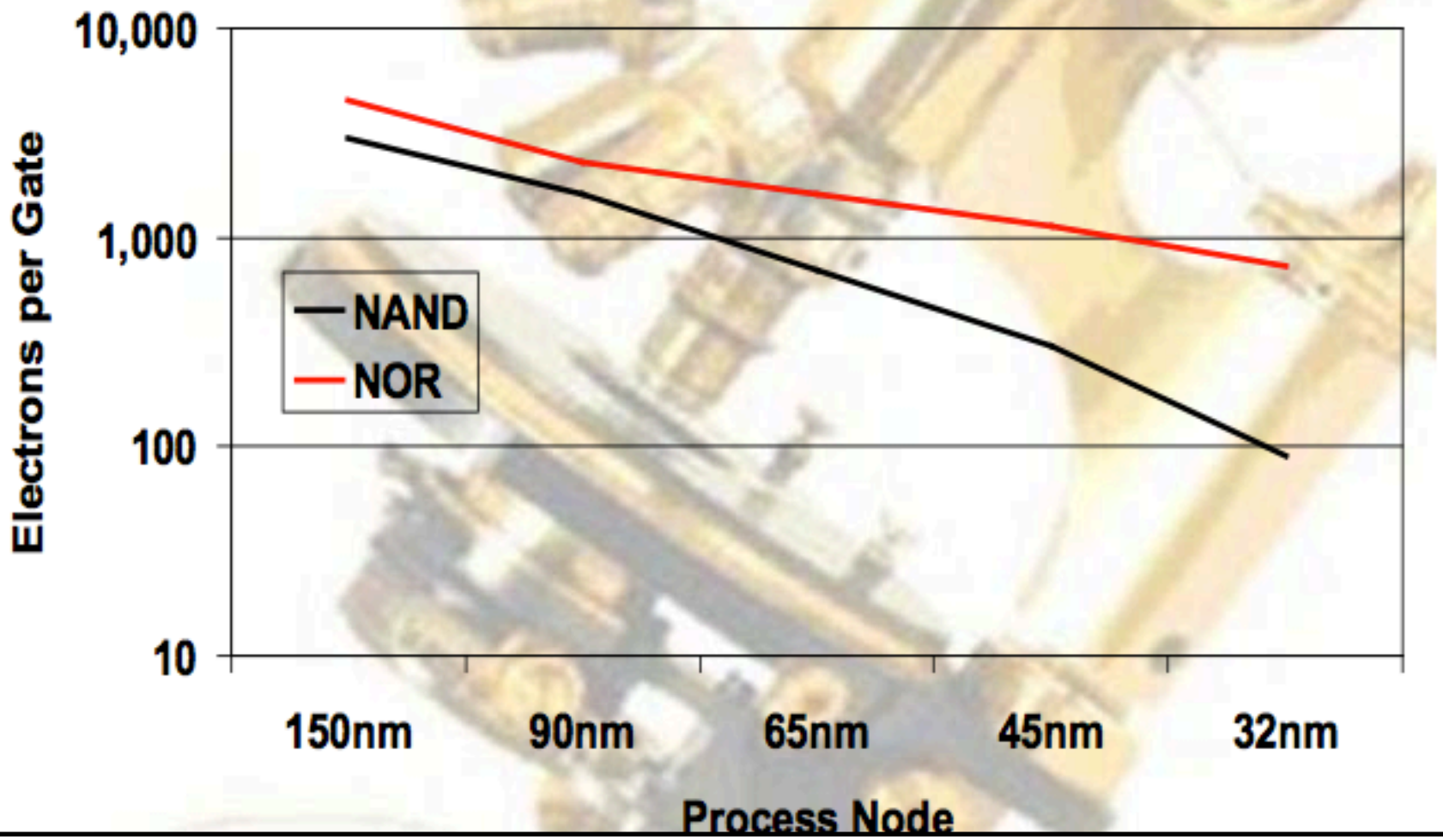
Going Forward…

- ***'When you've got a hammer in your hand, everything looks like a nail.'***
- ***THE FIRST Law of Technology says we invariably overestimate the short-term impact of new technologies while underestimating their longer-term effects* - Dr. Francis Collins**

Flash Going Forward

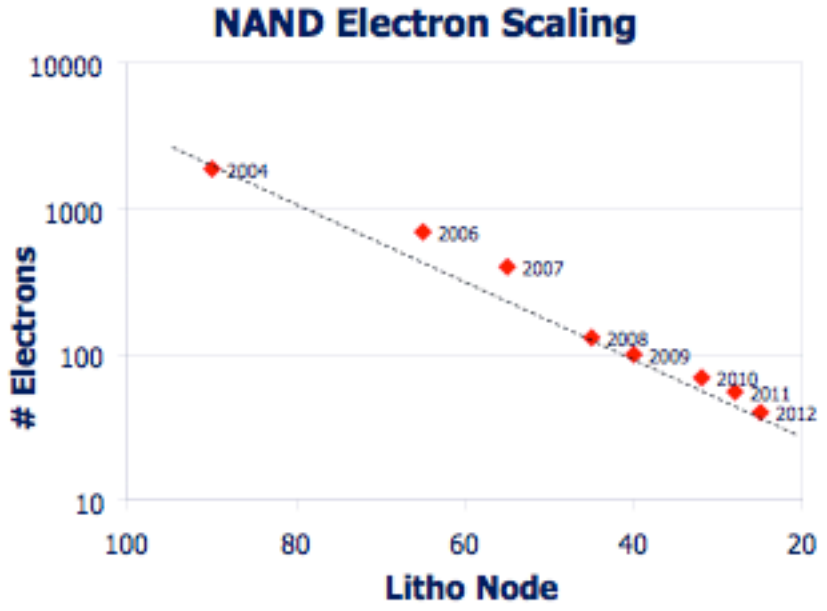
- $/GB unlikely to match regular disk
- $/IOP already significantly better
- Energy costs will be less that a regular HDD
- Usefulness with depend upon the data access pattern

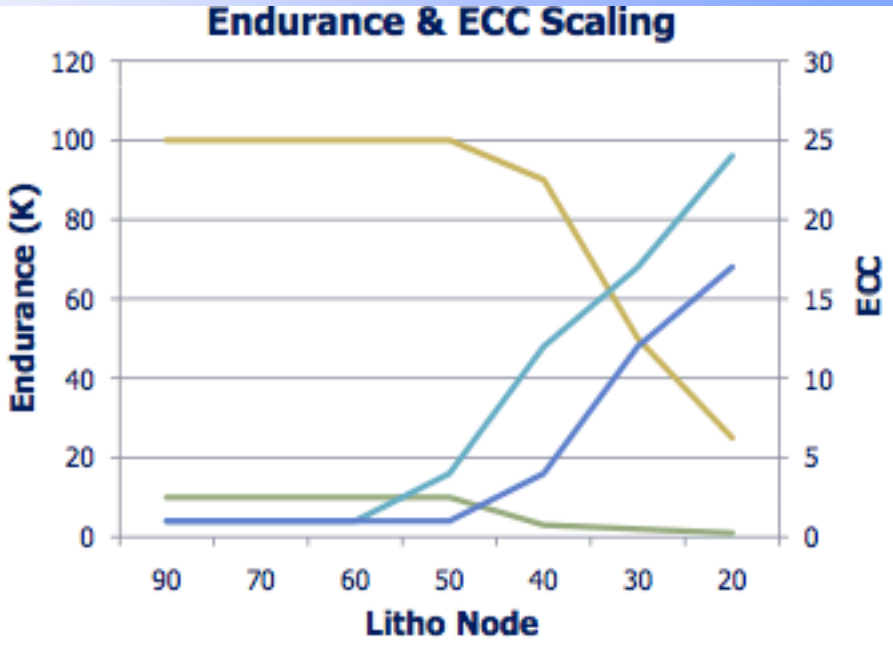
Reliability – Too Few Electrons Per Gate

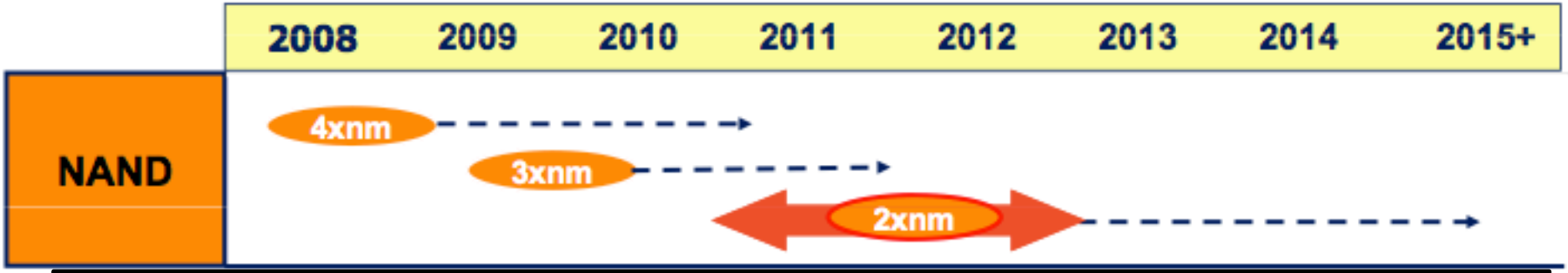


Source: *The Inevitable Rise of NVM in Computing*, Jim Handy, Nonvolatile Memory Seminar, Hot Chips Conference August 22, 2010 Memorial Auditorium Stanford University

Flash Technology Trends







Source: Ed DollerV.P. Chief Memory Systems Architect, Non-Volatile Memory Seminar Hot Chips Conference August 22, 2010 Memorial Auditorium Stanford University