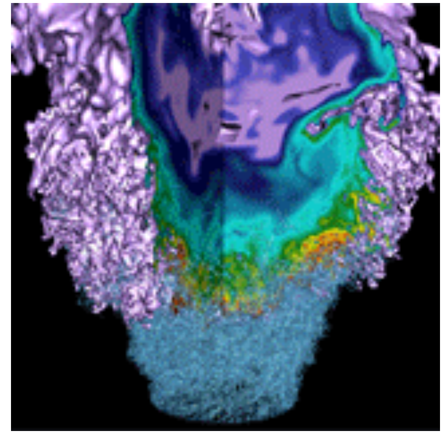
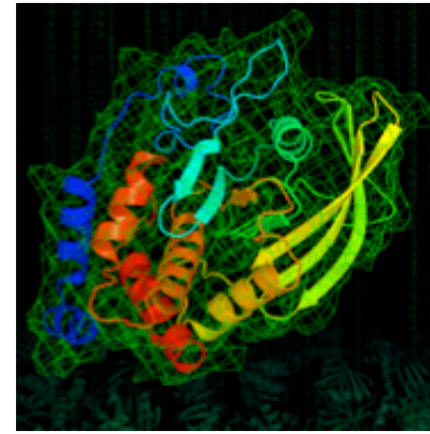
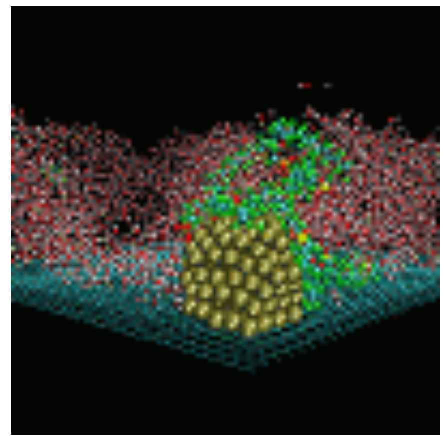
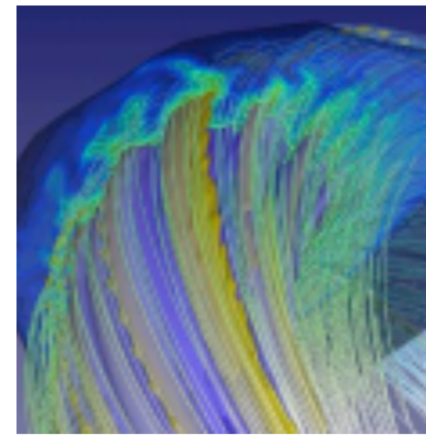
Evaluation of A Flash Storage Filesystem on the Cray XE-6

Jay Srinivasan and Shane Canon

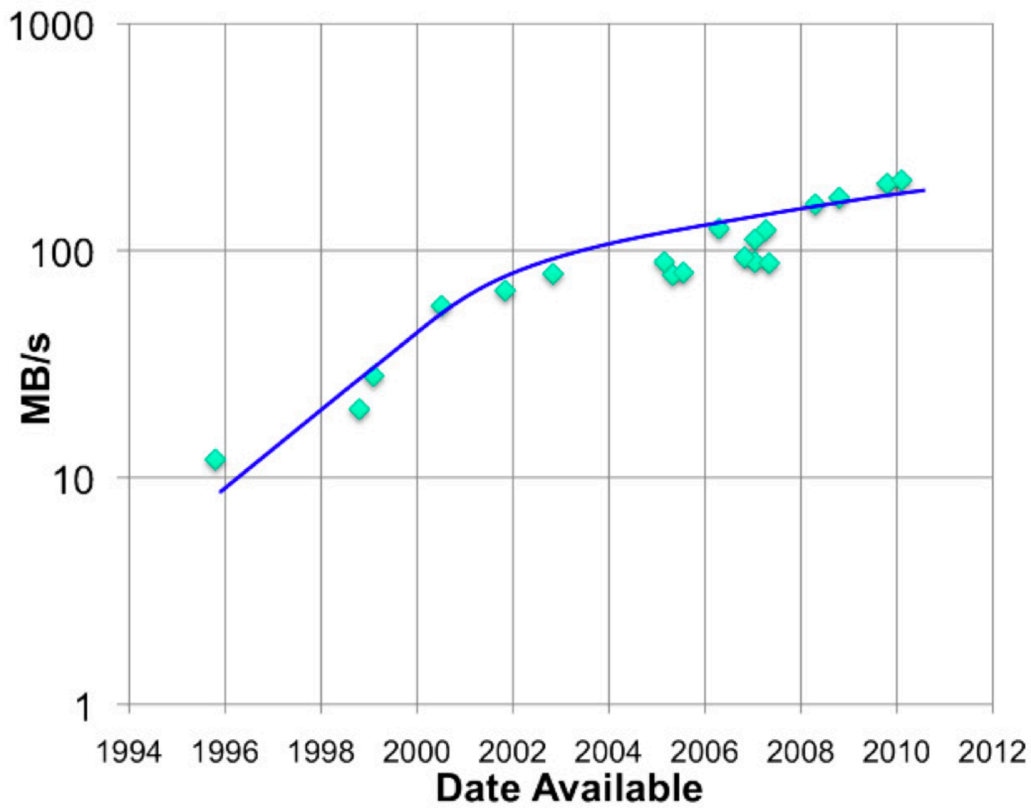
CUG 2013
Napa, California
May 2013

U.S. DEPARTMENT OF ENERGY | Office of Science
BERKELEY LAB Lawrence Berkeley National Laboratory

Outline

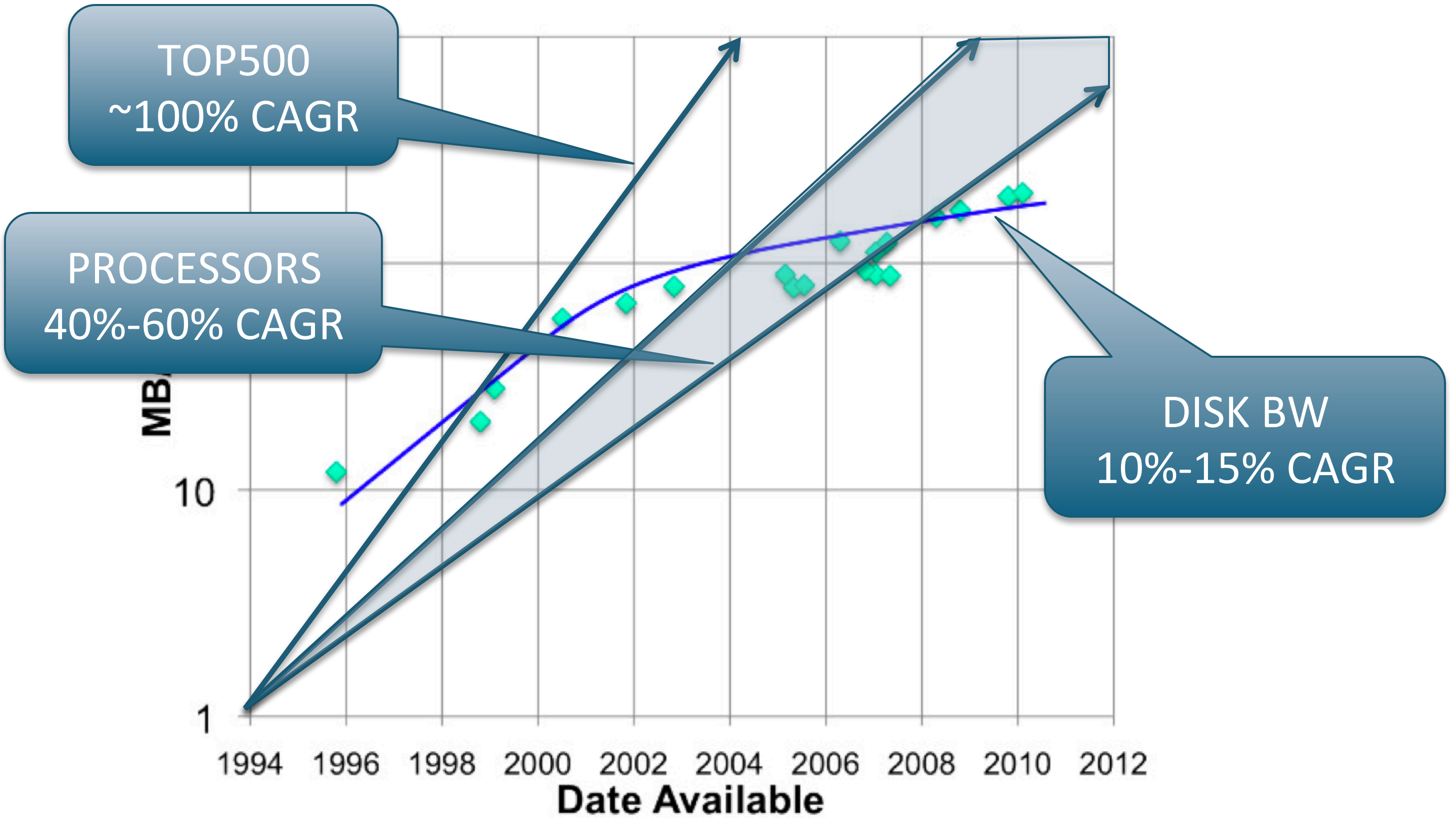
- Motivation
- Flash Characteristics and Approaches
- Testbed architecture
- Results
- Future work & Conclusions

Disk Bandwidths failing to Keep Pace



Source: IBM/Violin Memory White Paper

Disk Bandwidths failing to Keep Pace



Source: IBM/Violin Memory White Paper

Characteristics of Flash

Good

- Random Read Performance/IOPS
- Bandwidth
- Power

Bad

- Erase Cycle for Writes
- Wear/Endurance
- Grooming Cycle

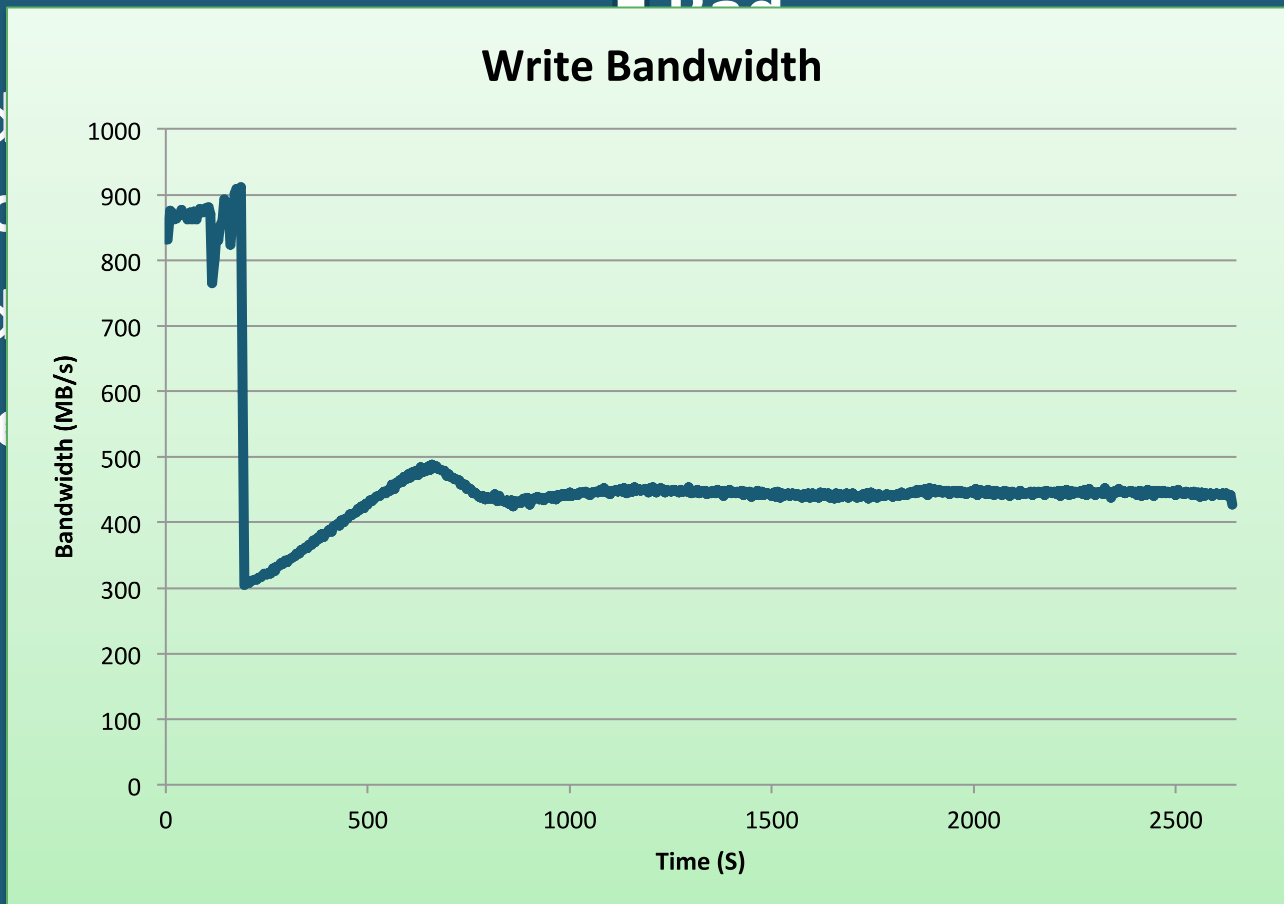
Characteristics of Flash

Good

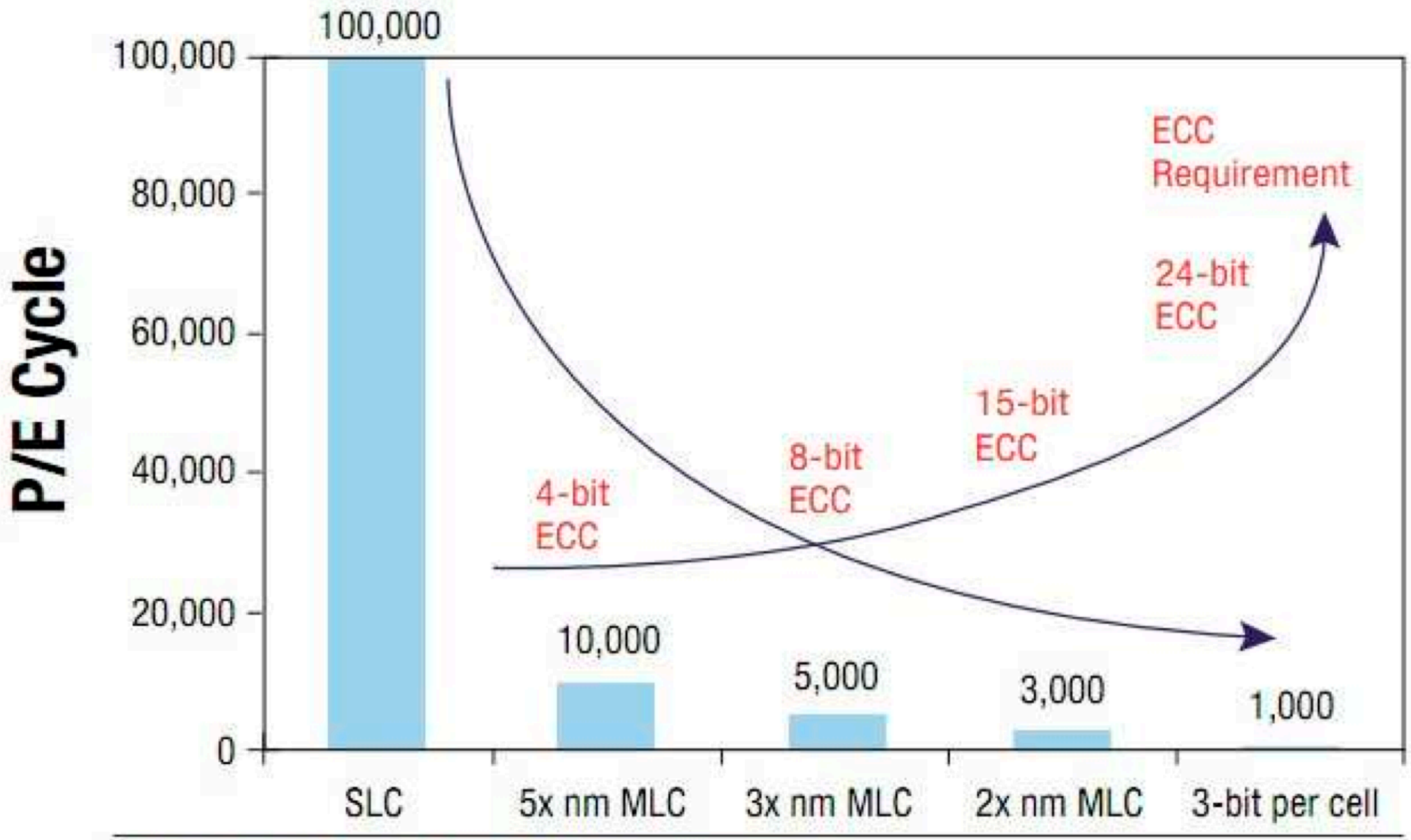
- Rand
  Perfo
- Band
- Powe

Bad

rites



SLC, MLC, and TLC



Source: JMicron, Western Digital, Morgan Stanley Research

Methods of Integration

On-Node

- Scalable BW
- Use as Memory

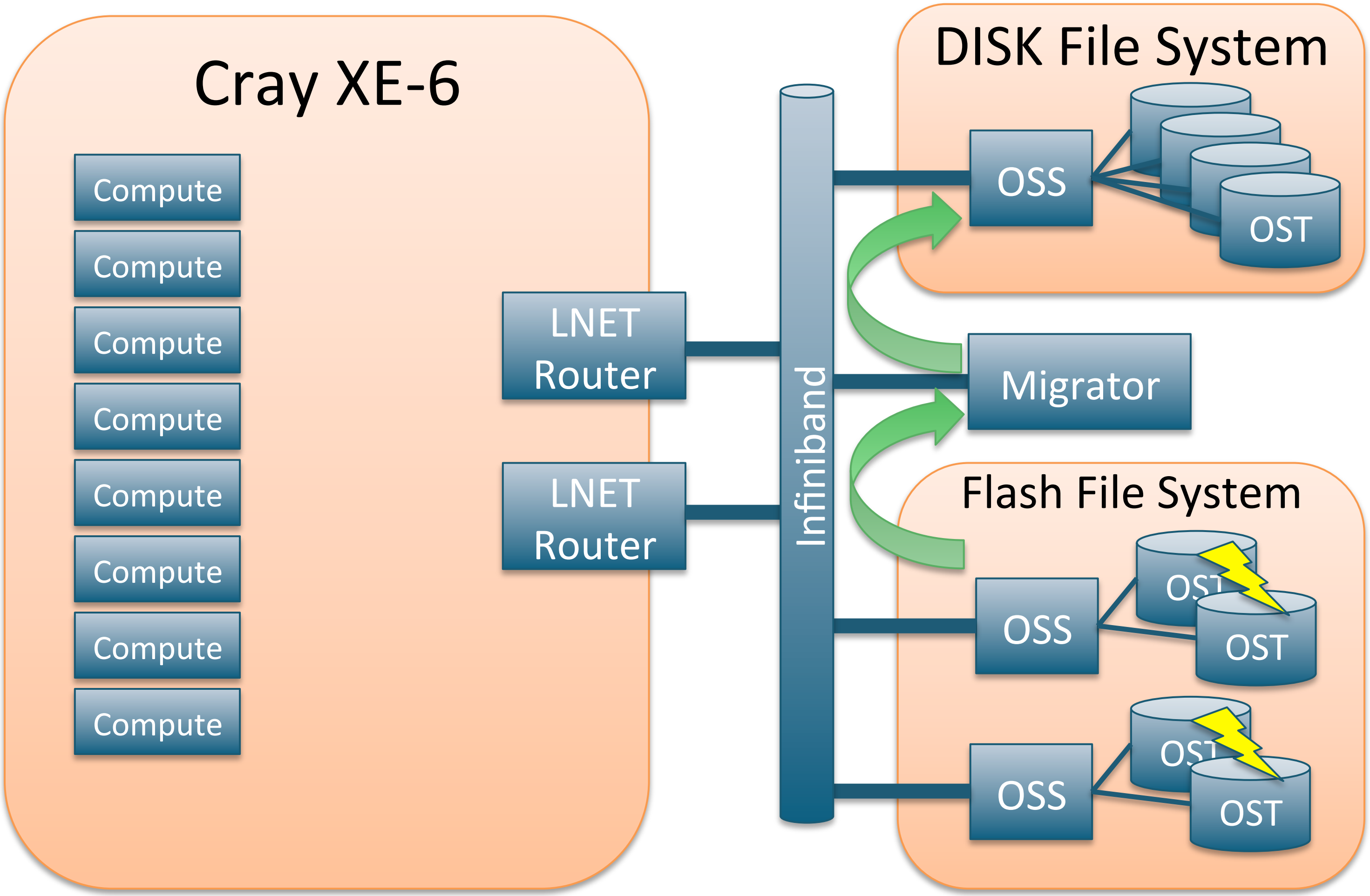
Integrated Hierarchy

- Transparent to upper layers

On-Edge/Shared

- Works in systems without local-node support

Testbed Architecture



Testbed Architecture

- We used a pool of high-performing storage along with a pool of lower-performing storage and migrated files between the two.
- Both Flash and Disk filesystems are mounted on the compute/service/login nodes as "external" Lustre filesystems.
- The "Migrator" program runs in the background looking for specific "checkpoint" files to move – this is done on a login node that has the Flash and Disk filesystems mounted.

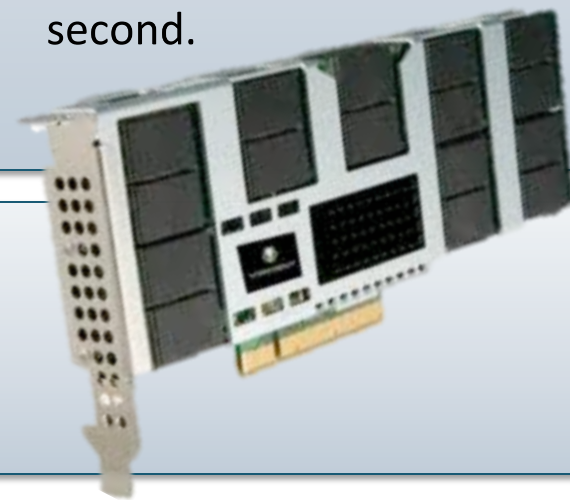
Configuration Details

Disk File System

- **Single Object Storage Server**
- **Four Object Storage Targets**
- **Single Dell R710**
- **Two LSI 8600 storage arrays.**
  - (Very Small, just for testing)

Flash File System

- **Two Object Storage Servers**
- **Four Object Storage Targets**
- **Two IBM x3650 M2**
- **Four Virident tachIOn cards**
  - 400 GB of SLC-class NAND
  - ~1.1 GB/s bandwidth
  - 160k Read IO operations per second.



Common to Both

- **12 Node TDS System (288 cores)**
- **Two Lustre Network (LNET) Routers**
- **QDR InfiniBand Network**

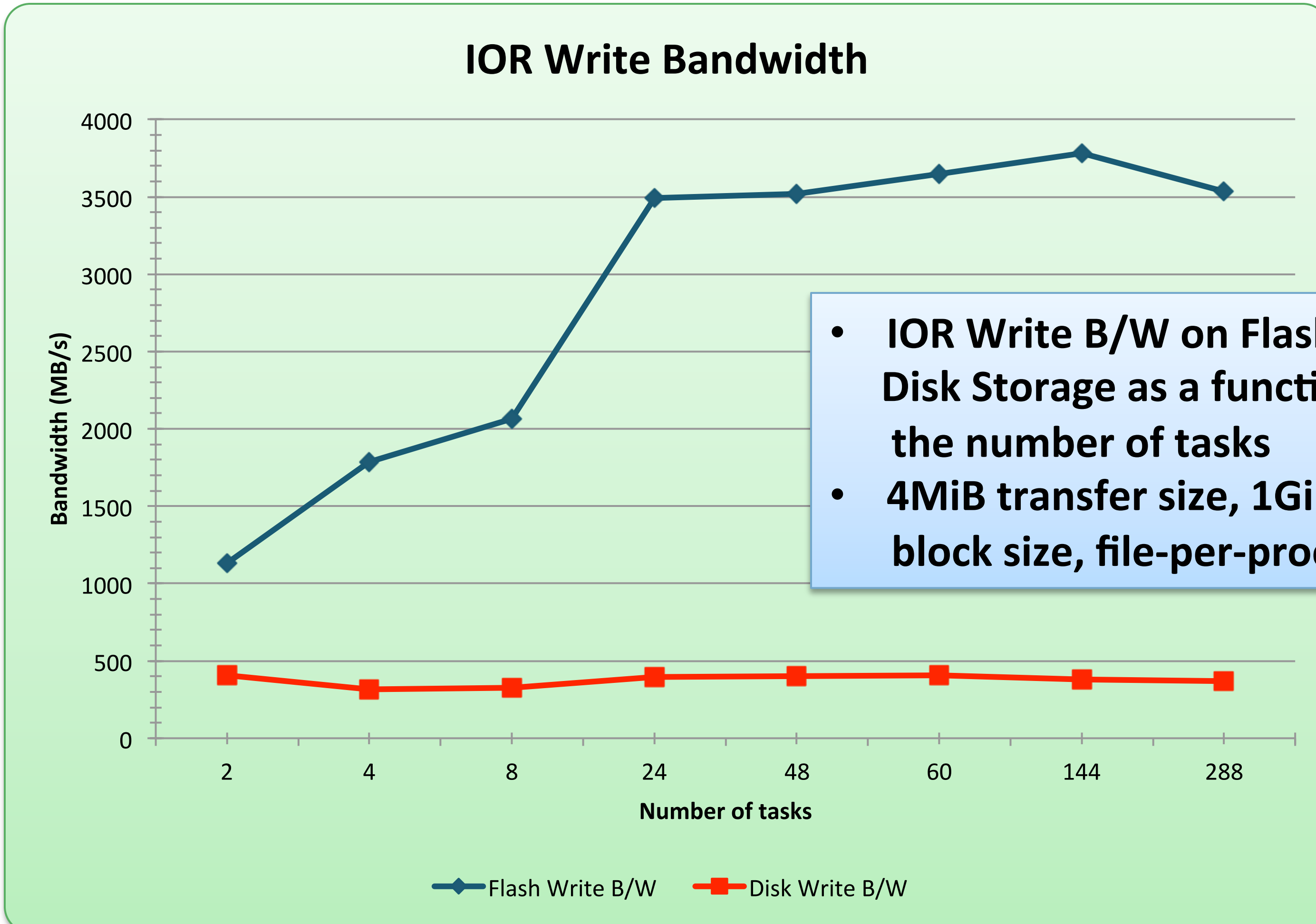
Benchmark code

- **IOR is a standard parallel filesystem benchmark – we use POSIX I/O and a file-per-process**
- **flashio is a benchmark code that mimics checkpoint I/O**
  - Computation (matrix-multiply) followed by “checkpoint” I/O (bursty, short duration)
  - Compute and I/O time can be tuned to ensure I/O time is a small fraction of the overall compute time
  - Code tracks time for compute, I/O and overall run time.

Migrator

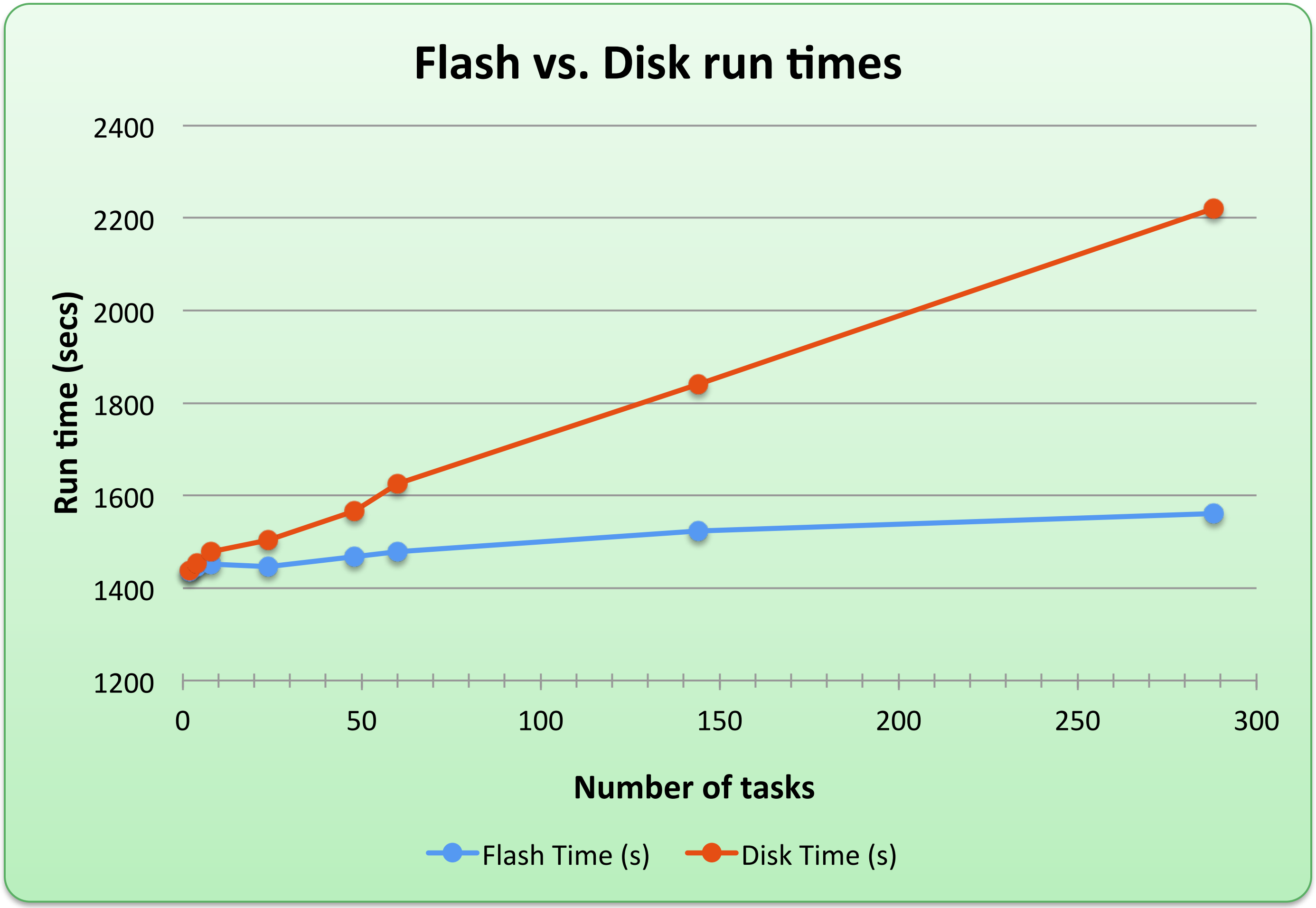
- **Flash storage is a scarce resource and cannot be used for long-term storage, or even for much longer than the duration of the job**
- **I/O acceleration can be explicitly requested by the user or be transparent to the user**
- **I/O path is complicated and some user interaction will be required to ensure it is effective.**
- **We use a migrator task that moves data from Flash to Disk storage automatically**
  - Moves only specific "named" checkpoint files and depends on a semaphore file to determine which one to move.

Results



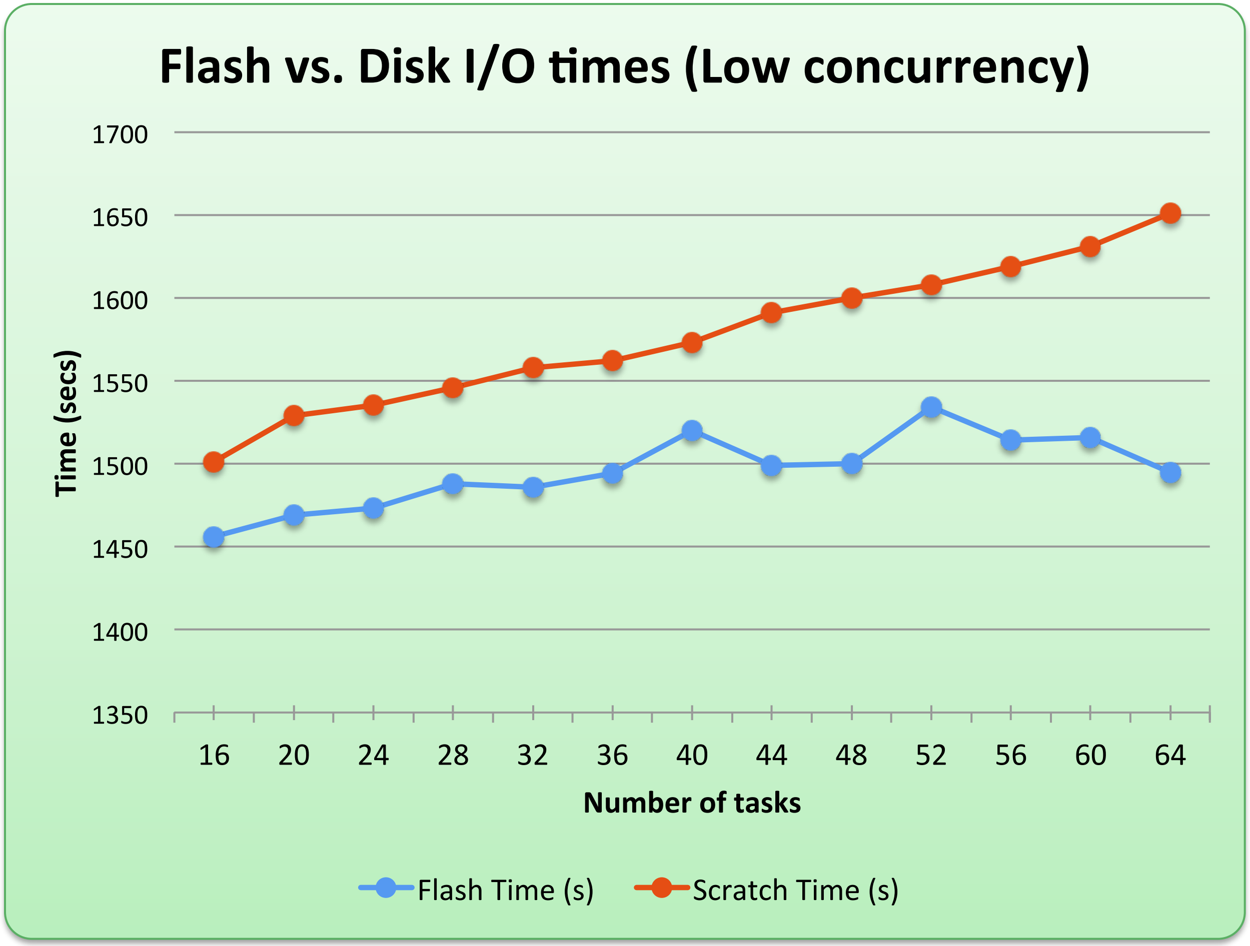
- IOR Write B/W on Flash and Disk Storage as a function of the number of tasks
- 4MiB transfer size, 1GiB block size, file-per-process

Results



Flash vs. Disk run times
Run time (secs)
2400
2200
2000
1800
1600
1400
1200
0
50
100
150
200
250
300
Number of tasks
Flash Time (s)
Disk Time (s)

Results


Flash vs. Disk I/O times (Low concurrency)
Time (secs)
1700
1650
1600
1550
1500
1450
1400
1350
16
20
24
28
32
36
40
44
48
52
56
60
64
Number of tasks
Flash Time (s)
Scratch Time (s)

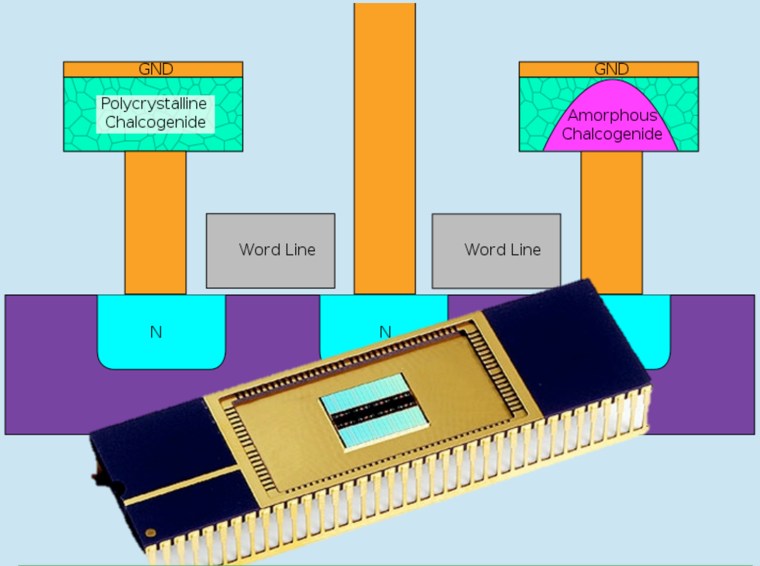
Cost (Enterprise Class)

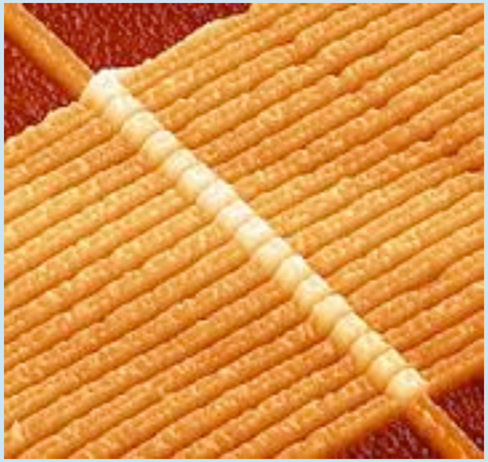
Storage	Bandwidth Cost ($ per GB/s)	Capacity Cost ($ per TB)
Flash Storage (Enterprise PCI-e)	$6,000	$6,000
Disk Storage (Enterprise Array)	$22,000	$400

Cost Comparison – Strawman Config

	Hybrid	Disk Only
Flash Storage BW (TB/s)	2.25	-
Disk Storage BW (TB/s)	0.39	1.00
Flash Capacity (PB)	2.25	-
Disk Capacity (PB)	20.9	53.3
Example Application		
Checkpoint Volume (TB)	1200	1200
Checkpoint Iteration (s)	3600	3600
Time for Checkpoint (s)	533	1200
Time for Compute (s)	3067	2400
Percentage of Time in I/O	15%	33%
Improvement	28%	-
Cost		
Total Cost	$22,100,000	$22,000,000

Beyond NAND





Phase Change Memory

- Changes a material to/from a amorphous/crystalline structure
- O(100ns) switching time
- 100M write cycle endurance
- Limited production at lower capacities

Memristor

- Resistance can be changed which stores the state
- < 100ns switching time
- O(1M) write cycle endurance
- Production pushed back beyond 2015

Future Work

- **Performance at scale**
  - Hundreds or thousands of (filesystem) clients
  - Larger pool of Flash storage
- **Improvements to the "migrator" – allow for "job asynchronous" migration or even staging (for reads)**
- **Evaluate ways to make performance more predictable to users at scale – private storage pools.**
- **Explore ways to expose control of and manage the migrator**

Conclusions

- **Flash and Solid State technologies are promising ways of accelerating I/O on HPC systems today.**
- **I/O acceleration can be done by using a storage hierarchy, and can be achieved all along the I/O path – from the compute element to the storage unit.**
- **I/O acceleration is primarily a Software problem – and there are a number of ways to solve the problem.**
- **We believe an optimal solution should not be hidden from user input and control of its use.**

NERSC

Questions?