Workload Characterization of a Leadership Class Storage Cluster

Technology Integration Group

National Center for Computational Sciences

Presented by Youngjae Kim

Youngjae Kim, Raghul Gunasekaran, Galen M. Shipman, David A. Dillow, Zhe Zhang, Bradley W. Settlemyer

A Demanding Computational Environment

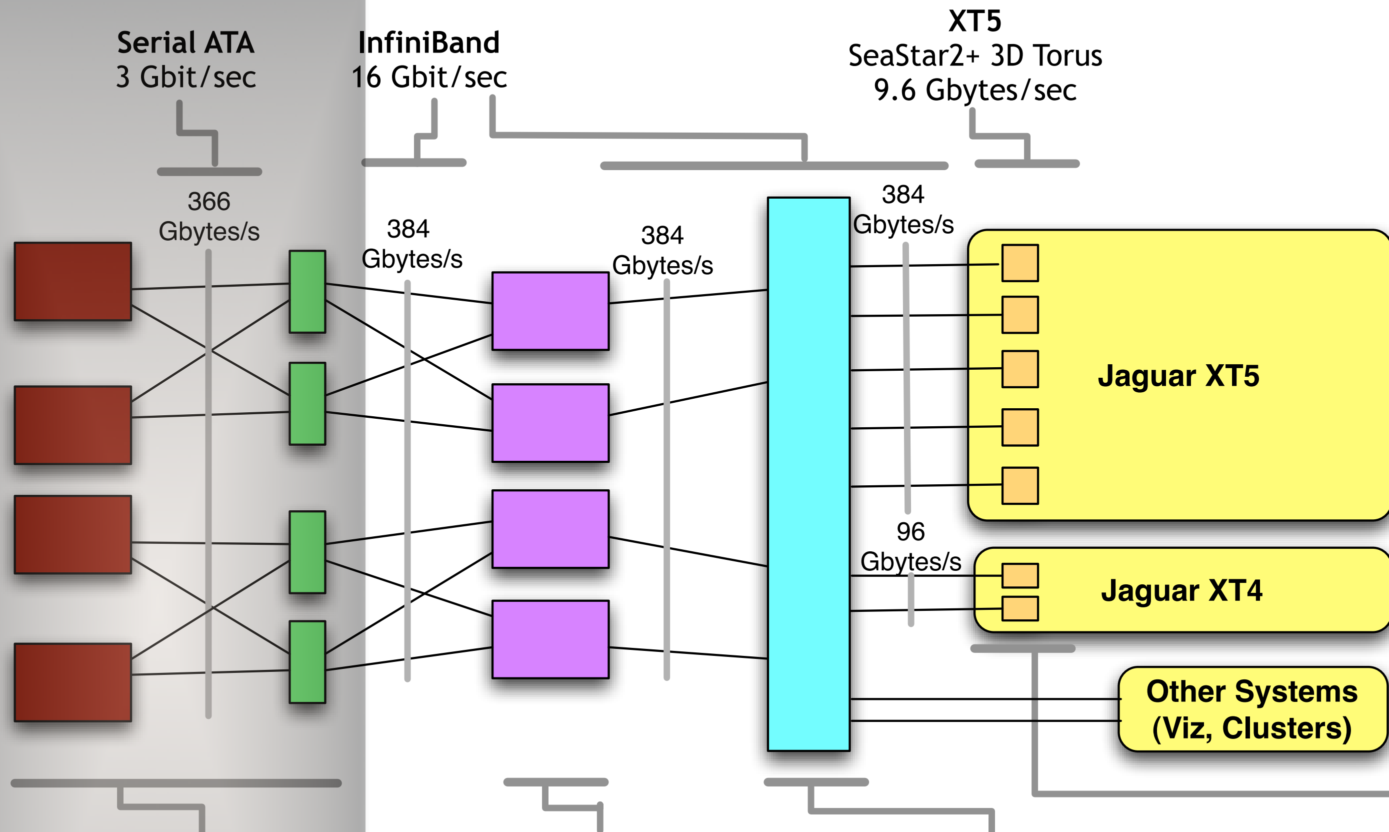
Jaguar XT5	18,688 Nodes	224,256 Cores	300+ TB memory	2.3 PFlops
Jaguar XT4	7,832 Nodes	31,328 Cores	63 TB memory	263 TFlops
Frost (SGI Ice)	128 Node institutional cluster			
Smoky	80 Node software development cluster			
Lens	30 Node visualization and analysis cluster			

Spider: A Large-scale Storage System

- **Over 10.7 PB of RAID 6 formatted capacity**
- **13,400 x 1 TB HDDs**
- **192 Lustre I/O servers**
- **Over 3TB of memory (on Lustre I/O servers)**
- **Available to many compute systems through high-speed IB network**
  - **Over 2,000 IB ports**
  - **Over 3 miles (5 kilometers) cable**
  - **Over 26,000 client mounts for I/O**
  - **Peak I/O performance is 240 GB/s**

OAK RIDGE National Laboratory

Spider Architecture



Enterprise Storage
controllers and large racks of disks are connected via InfiniBand.

96 DataDirect S2A9900 *controller with 1TB drives and 2 active InfiniBand connections per controller*

Storage Nodes
run parallel file system software and manage incoming FS traffic.

192 dual quad core Xeon servers with 16 Gbytes of RAM each

SION Network
provides connectivity between OLCF resources and primarily carries storage traffic.

3000+ port 16 Gbit/sec InfiniBand switch complex

Lustre Router Nodes
run parallel file system client software and forward I/O operations from HPC clients.

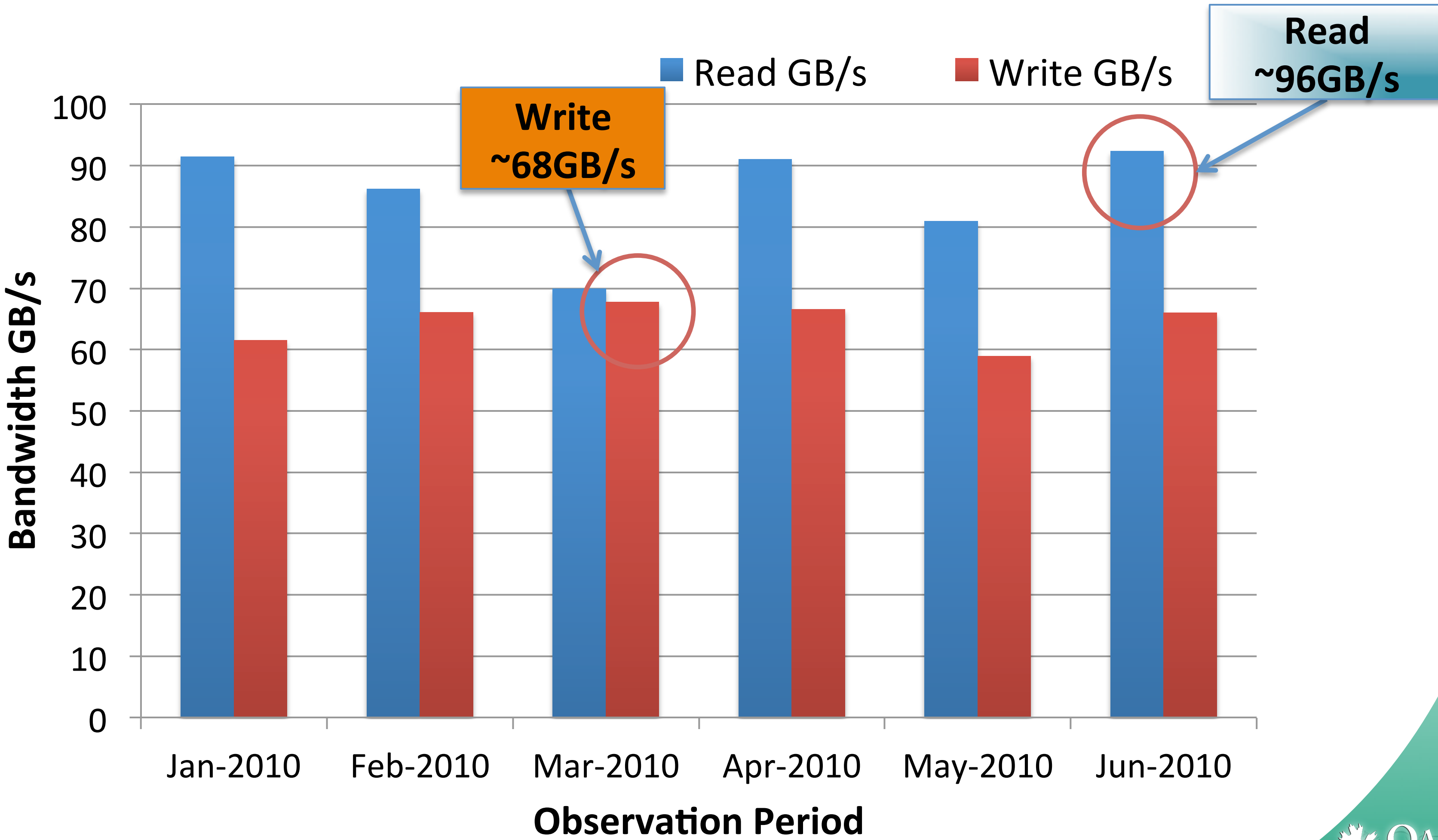
192 (XT5) and 48 (XT4) one dual core Opteron nodes with 8 GB of RAM each

Outline

- Background
- Motivation
- Workload Characterization
  - Data collection tool
  - Understanding workloads
    - Bandwidth requirements
    - Request size distribution
    - Correlating request size and bandwidth, etc.
  - Modeling I/O workloads
- Summary and Future works
  - Incorporating flash based storage technology
  - Further investigating application to file system's behavior

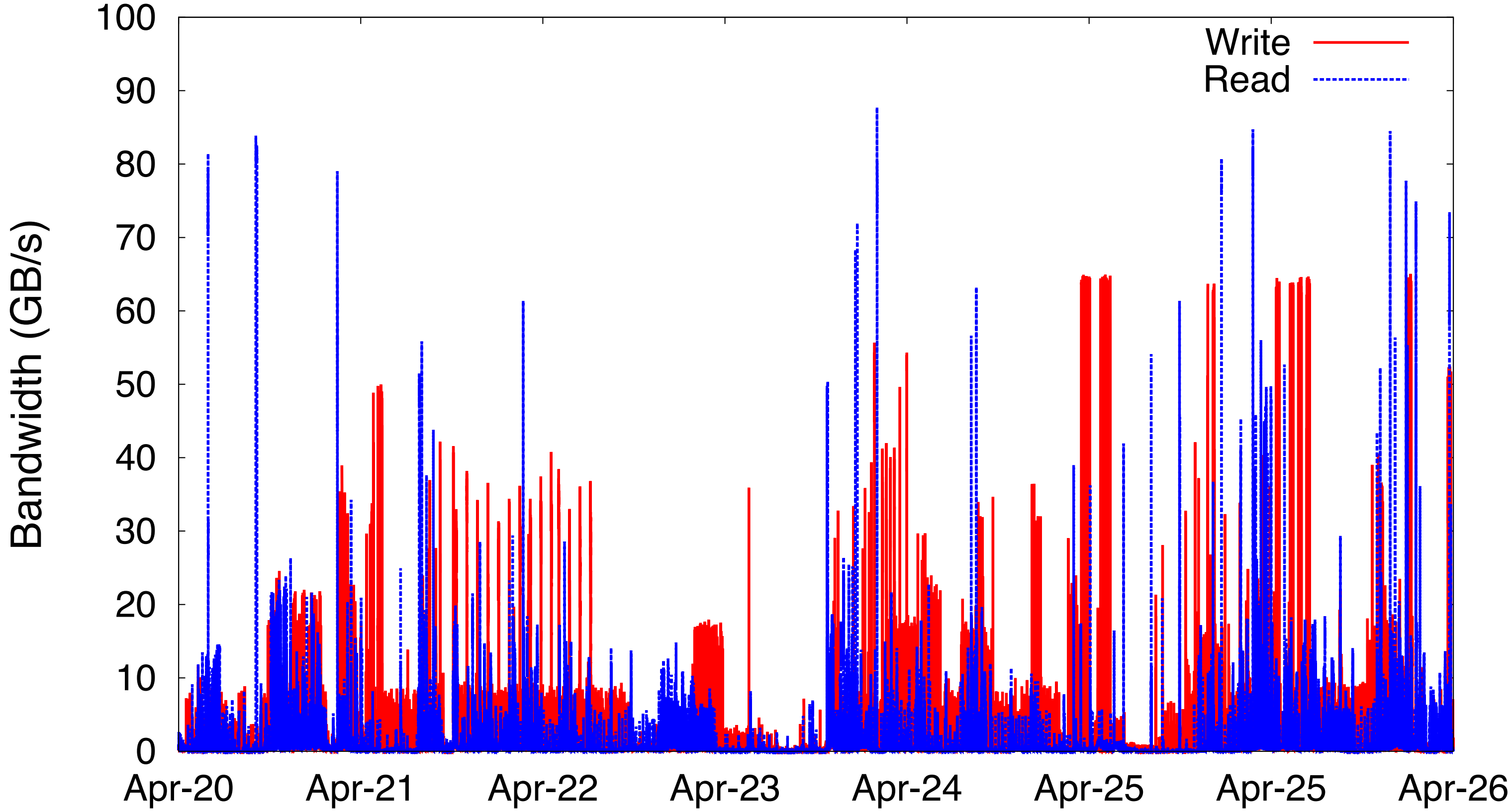
Monthly Peak Bandwidth

- Measured monthly peak read and write bandwidth on 48 controllers (half our capacity)



Snapshot of I/O Bandwidth Usage

- Observed read and write bandwidth for a week in April

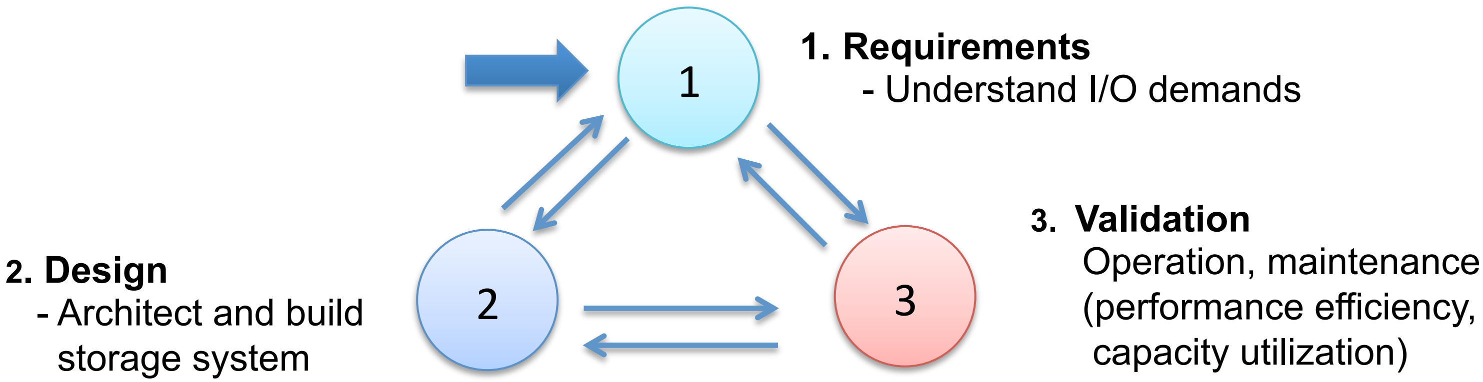


Data sampled every 2 seconds from 48 controllers (half our capacity)

Motivation

Why Characterize I/O Workloads on Storage Clusters?

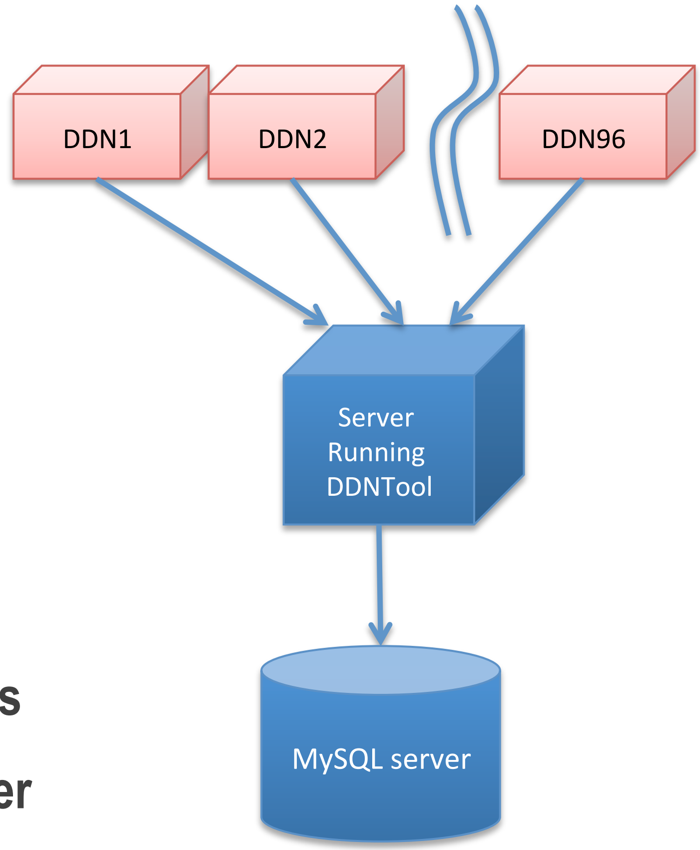
- Research Challenges and Limitation
  - Understanding I/O behavior of such large-scale storage system is of importance.
  - Lack of understanding on I/O workloads will lead under- or over-provisioned systems, increasing installation and operational cost ($).
- Storage System Design Cycle



- Goals
  - Understanding I/O demands of large-scale production system
  - Synthesizing the I/O workload to provide useful tool to storage controller, network, and disk-subsystem designers

Data Collection Tool

- Monitoring Tool
  - Monitors variety of parameters from the back-end storage hardware
  - Metrics: Bandwidth (MB/s), IOPs



- Design Implementation
  - DDN S2A9900 API for reading controller metrics
  - A custom utility tool* on the management server
    - Periodically collects stats from all the controllers
    - Supports multiple sampling rates (2, 60, 600) seconds
  - Data is archived in a MySQL database.

** Developed by Ross Miller, et. al., in TechInt group, NCCS, ORNL*

Characterizing Workloads

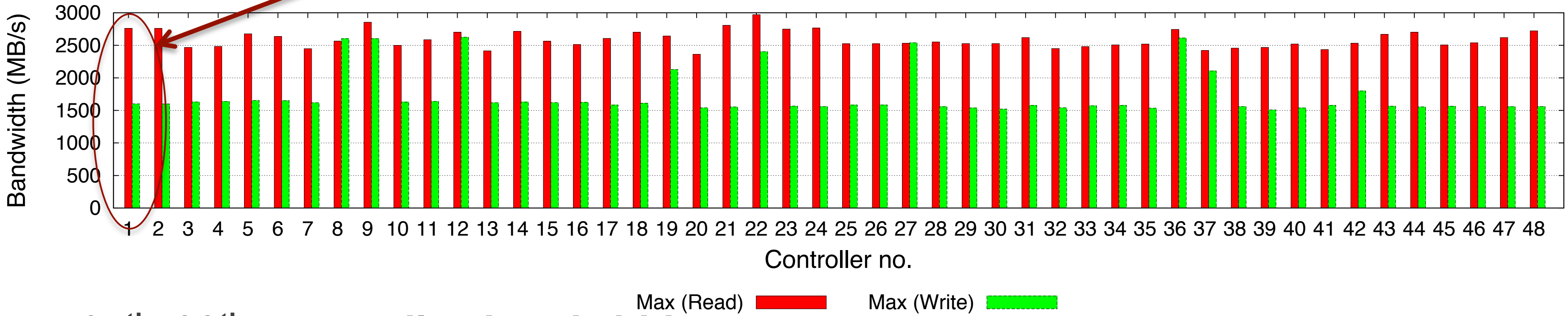
- Data collected from RAID controllers
  - Bandwidth/IOPS (every 2 sec)
  - Request size stats (every 1 min)
  - Used data collected from Jan. to June (around 6 months)

- Workload Characterization and Modeling
  - Metrics
    - I/O bandwidth distribution
    - Read to write ratio
    - Request size distribution
    - Inter-arrival time
    - Idle time distribution
  - Used curve-fitting technique to develop synthesized workloads

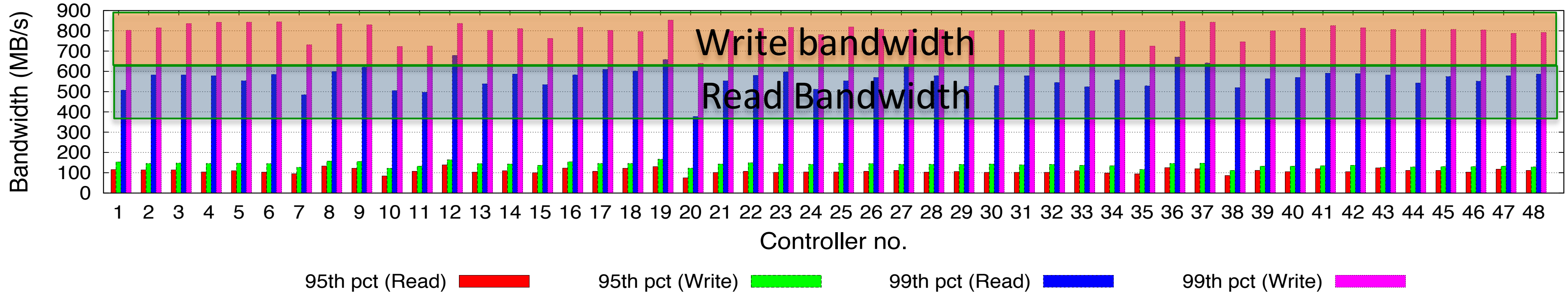
Bandwidth Distribution

- Peak bandwidth

Peak Read BW up to 2.7GB/s >> Peak Write BW up to 1.6GB/s



- 95th, 99th percentiles bandwidth



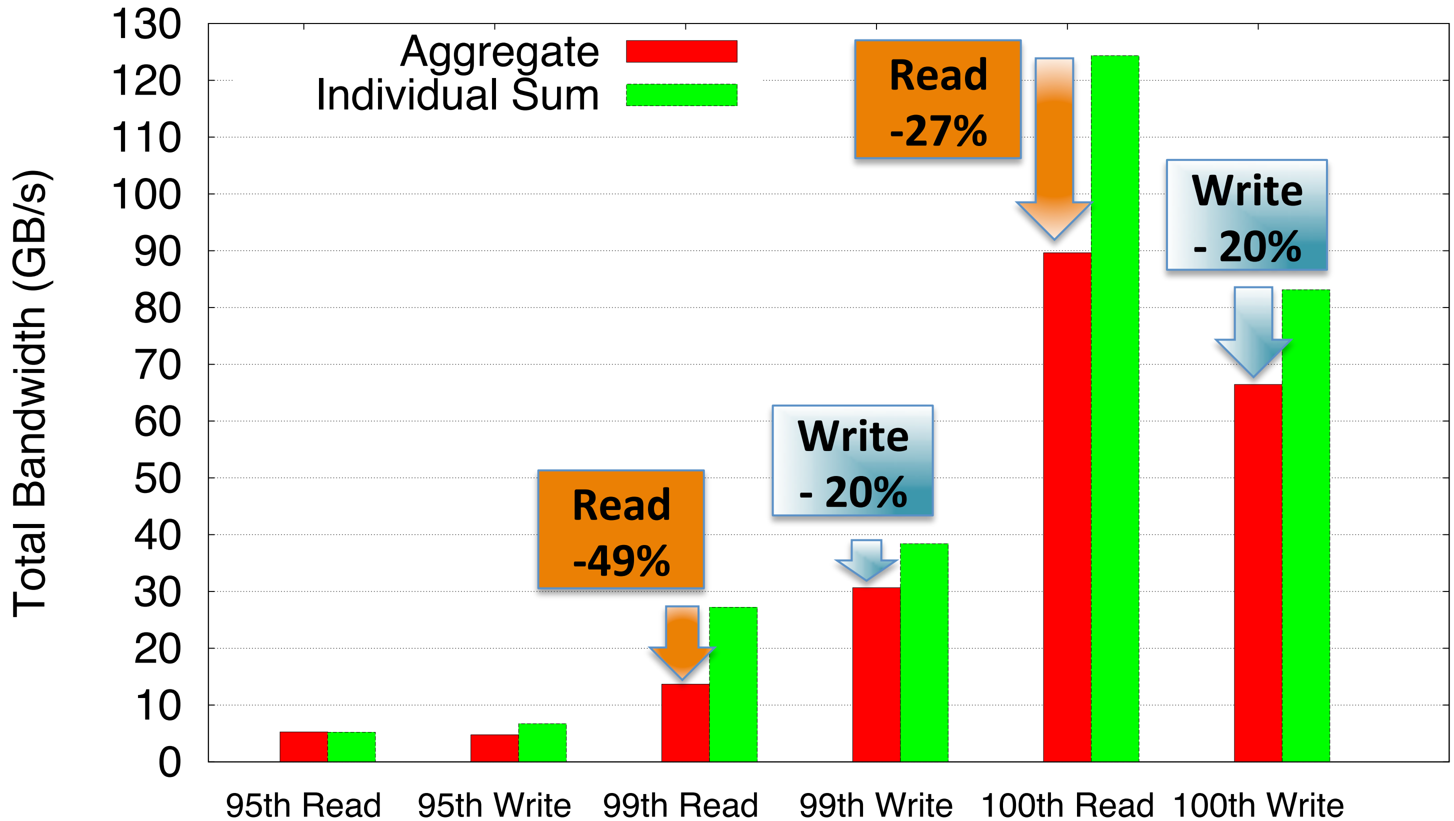
Write bandwidth >> Read bandwidth for both 95th and 99th percentiles bandwidth

Observations:

1. Long-tail distribution of read write bandwidth across all controllers
2. Read peak bandwidth much higher than write peak bandwidth, but majority of bandwidth higher in writes over reads (e.g., 95-99 percentiles of bandwidth)
3. Variation in peak bandwidth across controllers

Aggregate Bandwidth

- **Peak aggregate bandwidth vs. Sum of peak bandwidth at every controller**



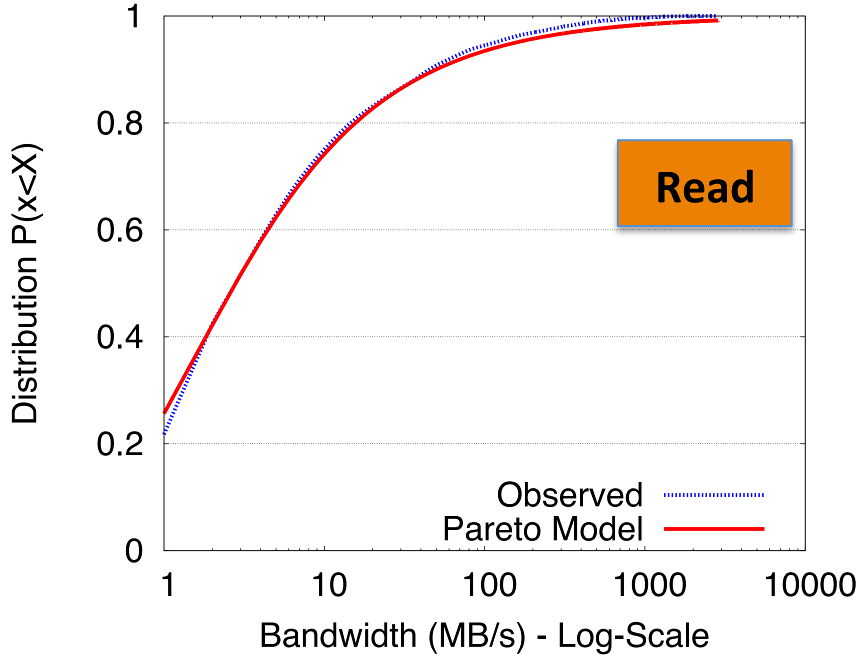
<u>**Observations:**</u>

1. Peak bandwidths of every controller unlikely to happen at the same time
2. Read bandwidth more unlikely to happen at the same time than write bandwidth for 99th and 100th percentiles of bandwidth

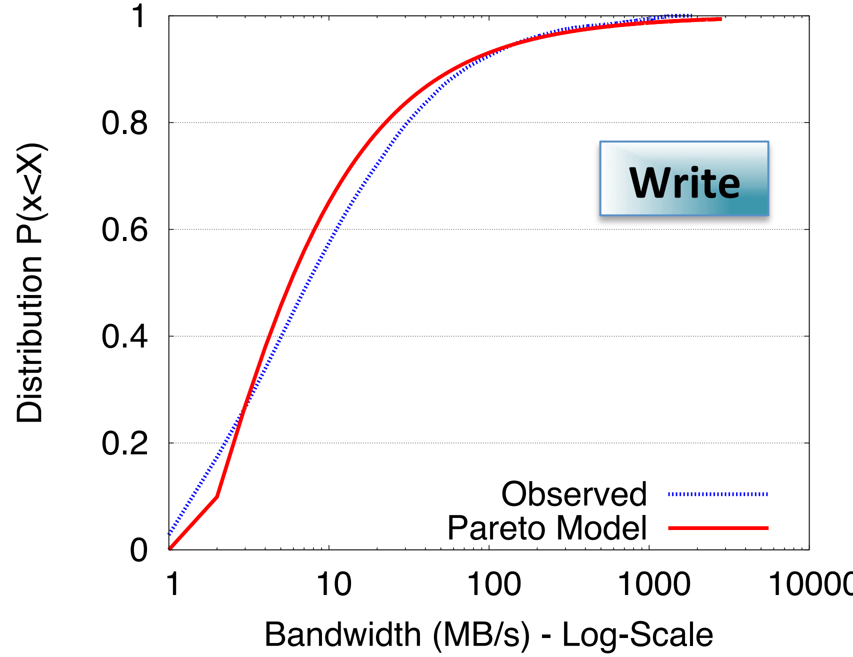
Modeling I/O Bandwidth Distribution

- We observed that read write bandwidth follows a long-tail dist.
- Pareto model is one of the simplest long tailed dist. models.
- Pareto model validation
  - Single controller

$$F_X(x) = \begin{cases} 1 - \frac{x_m^\alpha}{x}, \text{ for } x \geq x_m \\ 0, \text{ for } x < x_m \end{cases}$$



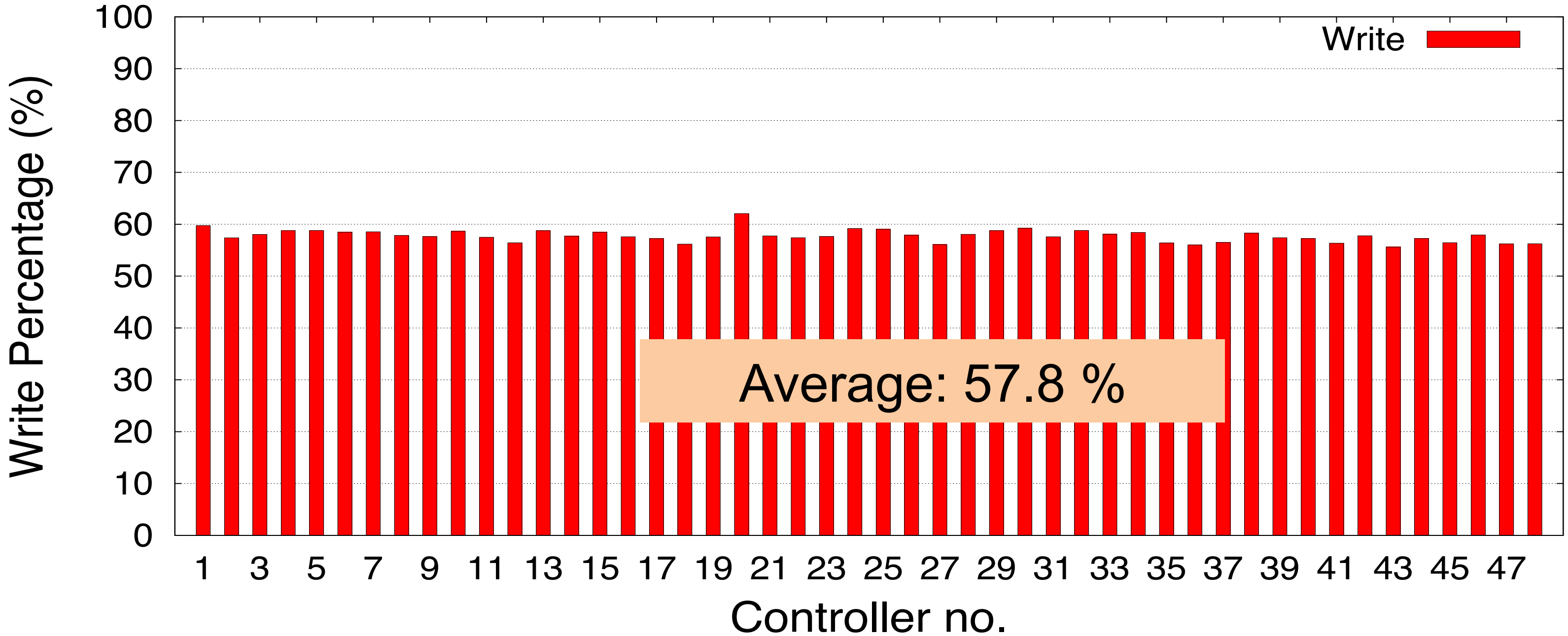
Goodness-of-fit (R^2): 0.98
α = 1.24



Goodness-of-fit (R^2): 0.99
α = 2.6

Read to Write Ratio

- **Percentage of write requests**



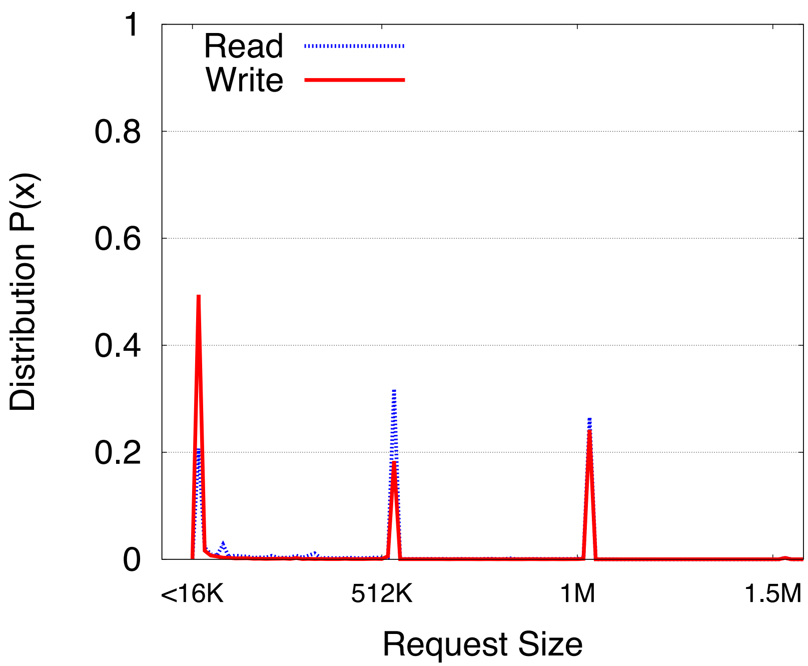
42.2% Read requests ➔ still significantly high!!!

42.2% read requests:

1. Spider is the center-wide shared file system.
2. Spider supports an array of computational resources such as Jaguar XT5/ XT4, visualization systems, and application development.

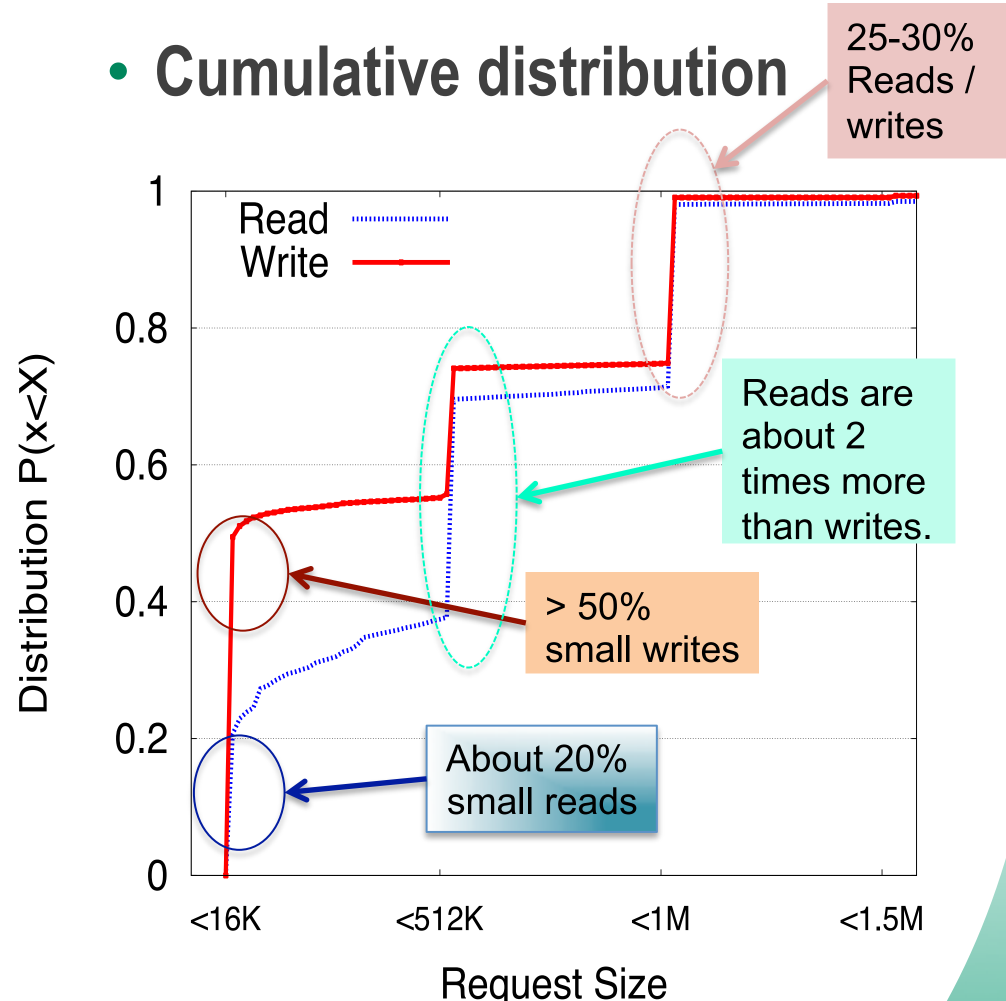
Request Size Distribution

- Probability distribution



Majority of request size (>95%)
- <16KB
- 512KB and 1MB

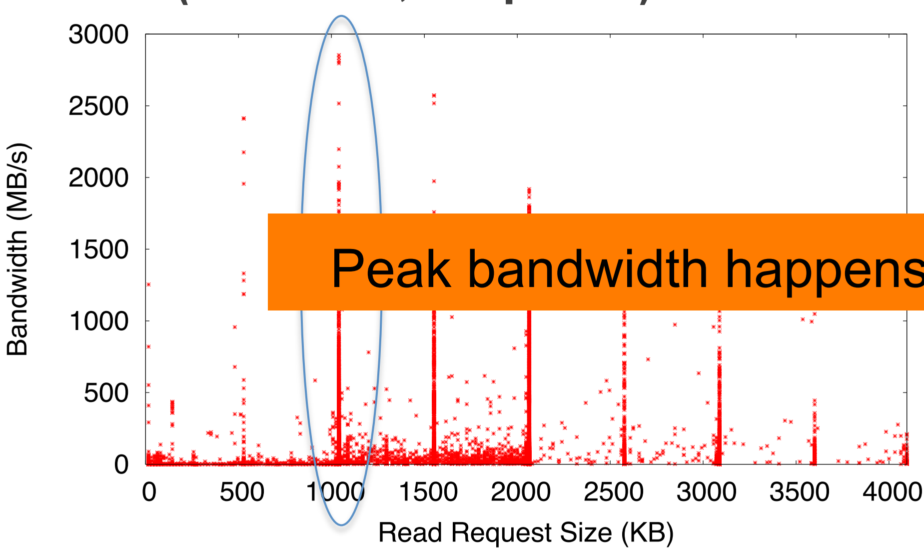
- Cumulative distribution

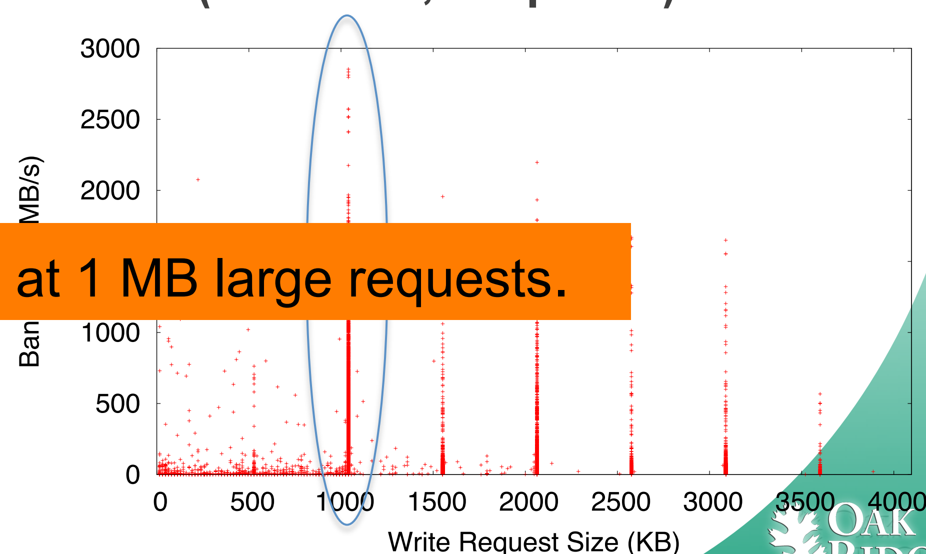


1. Linux block layer clusters near 512KB boundary.
2. Lustre tries to send 1MB request.

Correlating Request Size and Bandwidth

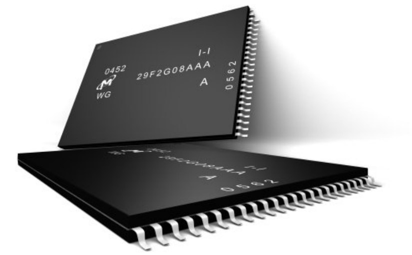
- Challenges: different sampling rates
  - Bandwidth sampling @ 2 second intervals
  - Request size distribution @ 60 seconds intervals
- Assumption
  - Larger requests are more likely to lead to higher bandwidth.
- Observed from 48 controllers
  - (Read BW, Req. Size)
  - (Write BW, Req. Size)





Peak bandwidth happens at 1 MB large requests.

What about Flash in Storage?



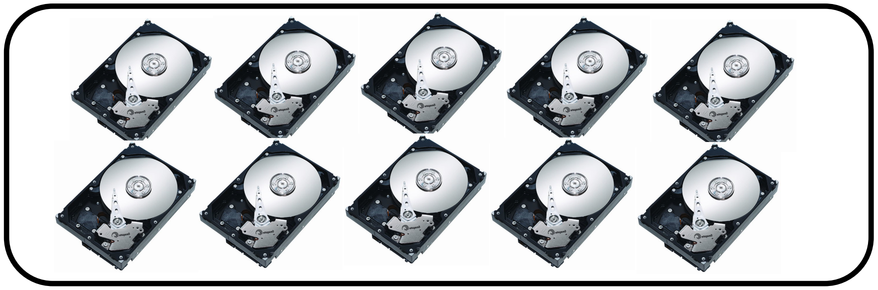
- Major observations from workload characterization
  - Reads and writes are bursty.
  - Peak bandwidth occurs at 1MB large requests.
  - More than 50% small requests and about 20% small read requests

- What about Flash?
  - Pros
    - Lower access latency (~0.5ms)
    - Lower power consumption (~1W)
    - High resilience to vibration temperature
  - Cons
    - Lifetime constraint (10K~1M erase cycle)
    - Expensive
    - Performance variability

OAK RIDGE
National Laboratory

Non-volatile Memory Device

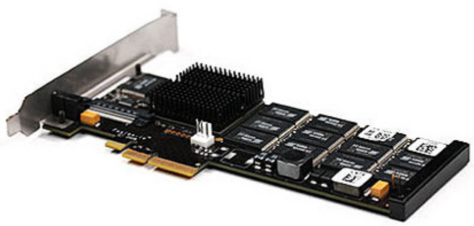
- HDD OST



10 x 1TB Hard drives in RAID-6

(~350 MB/s)

- SSD OST



1 Fusion I/O Duo

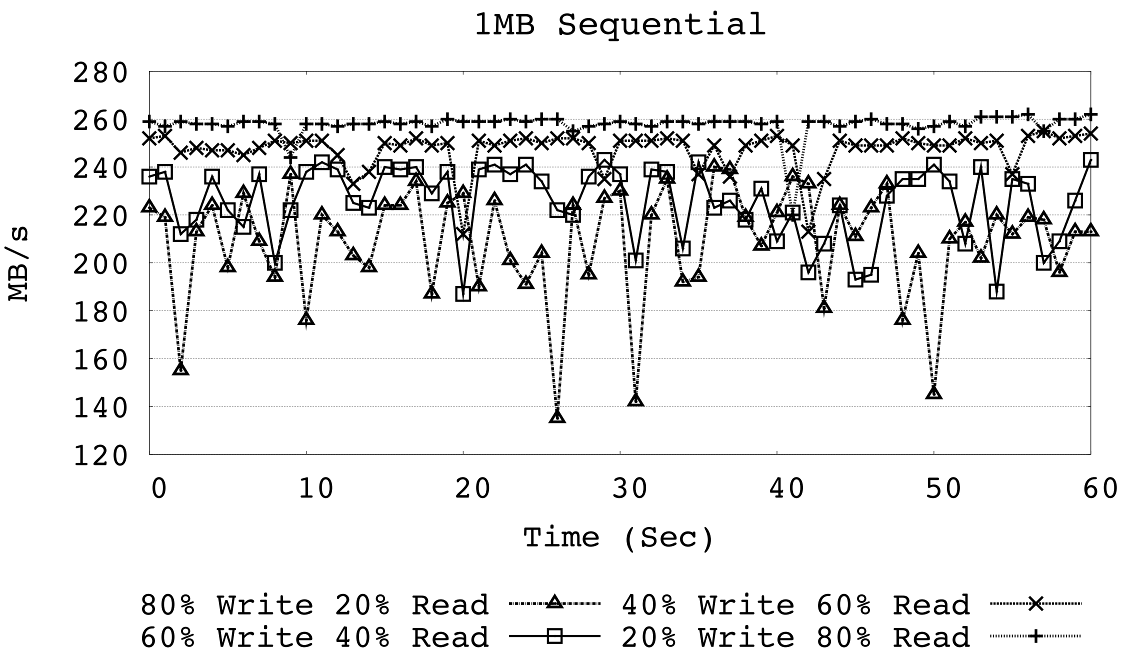
~1.4GB/s (Read)
~ 1GB/s (Write)

6 x Intel SSDs in RAID-0

~1.1GB/s (Read)
~0.8GB/s (Write)

Flash constraints

- Performance variability and lifetime of Flash highly dependent on I/O access patterns of workloads



Intel SSD

- Proper evaluation of Flash requires detailed workload characterization
  - Aggregate IO workload characterization
  - Individual application I/O characterization
  - Duty cycles

Summary and Future Works

- Summary
  - Analyzed 6 months data and still continue collecting data at present
  - From the analysis, we understood:
    - Max bandwidth is much higher than 99th percentile bandwidth.
    - Bandwidth distribution can be modeled in a Pareto model.
    - Read requests (42%) are closely as many as write requests (58%).
    - Peak bandwidth occurs at 1 MB large requests.

- Future works
  - Collecting block-level traces to further understand I/O workloads to the Spider
  - Collecting RPC logs to infer individual applications and profile application I/O access patterns with the block-level traces

Questions?

Contact info

Youngjae Kim (PhD)

kimy1 at ornl dot gov

Technology Integration Group

National Center for Computational Sciences

Oak Ridge National Laboratory