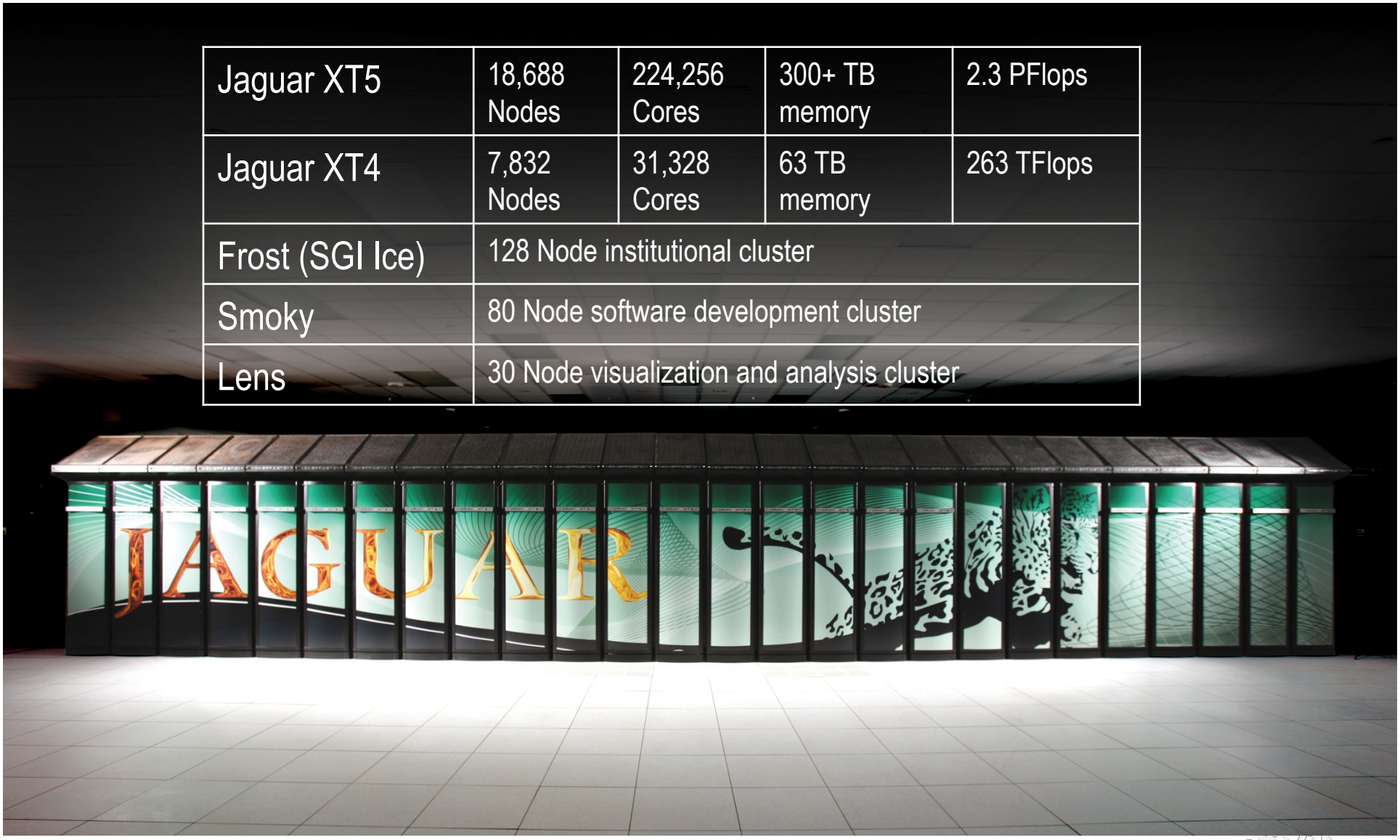
Lustre at the OLCF: Experiences and Path Forward

Galen M. Shipman

Group Leader – Technology Integration

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

Fastest Lustre file system in the world

Demonstrated bandwidth of 240 GB/s on the center-wide file system

Largest scale Lustre file system in the world

Demonstrated stability and concurrent mounts on all major OLCF systems

- **Jaguar XT5**
- **Jaguar XT4**
- **Opteron Dev Cluster (Spider)**
- **Visualization Cluster (Lens)**

Over 26,000 clients mounting the file system in production

Over 282,000,000 files
Multiple petabytes of data stored

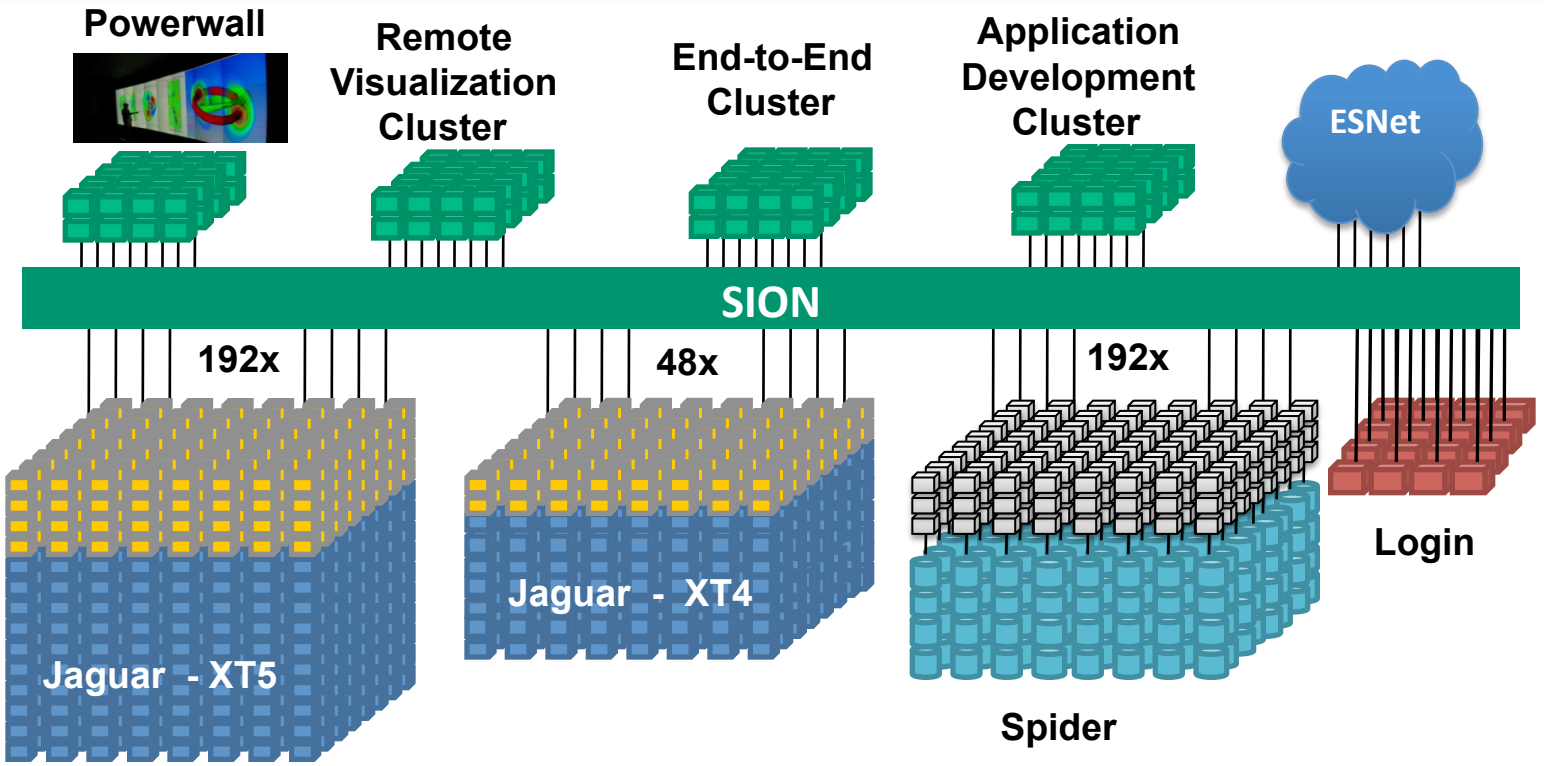
Cutting edge resiliency at scale

Demonstrated resiliency features on Jaguar XT5

- **DM Multipath**
- **Lustre Router failover**

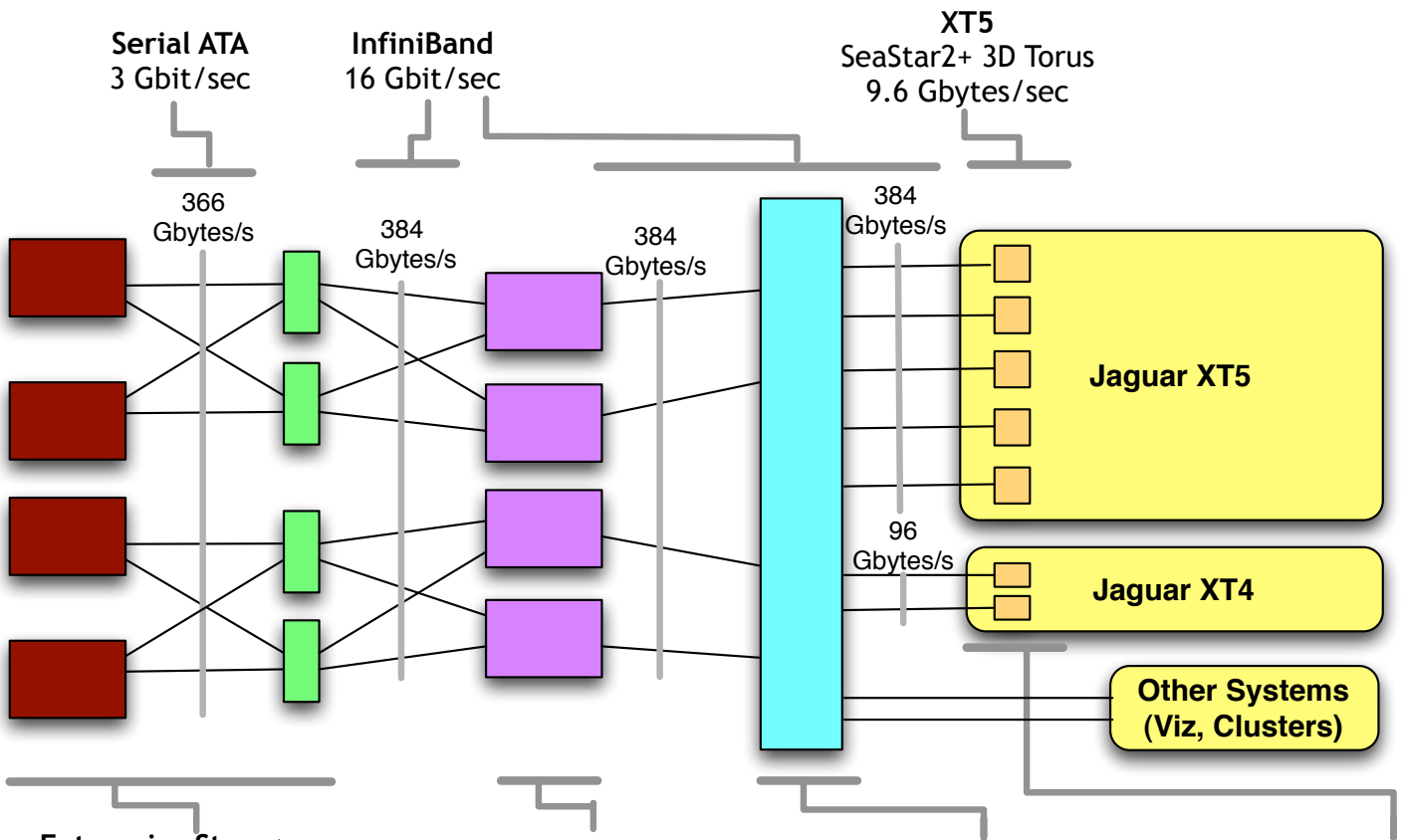
Spider – Helping Users and Staff

- Accessible from all major OLCF resources
  - enables data sharing between systems without resource consuming data copies
- Accessible during maintenance windows
- Decouples simulation platform procurement from storage system procurement
  - Allows file system to take an independent trajectory
  - Procurements can be planned to better coincide with vendor roadmaps



OAK RIDGE National Laboratory

Spider - Speeds and Feeds



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

48 DataDirect S2A9900 controller pairs with 1 Tbyte drives and 4 InifiniBand connections per pair

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

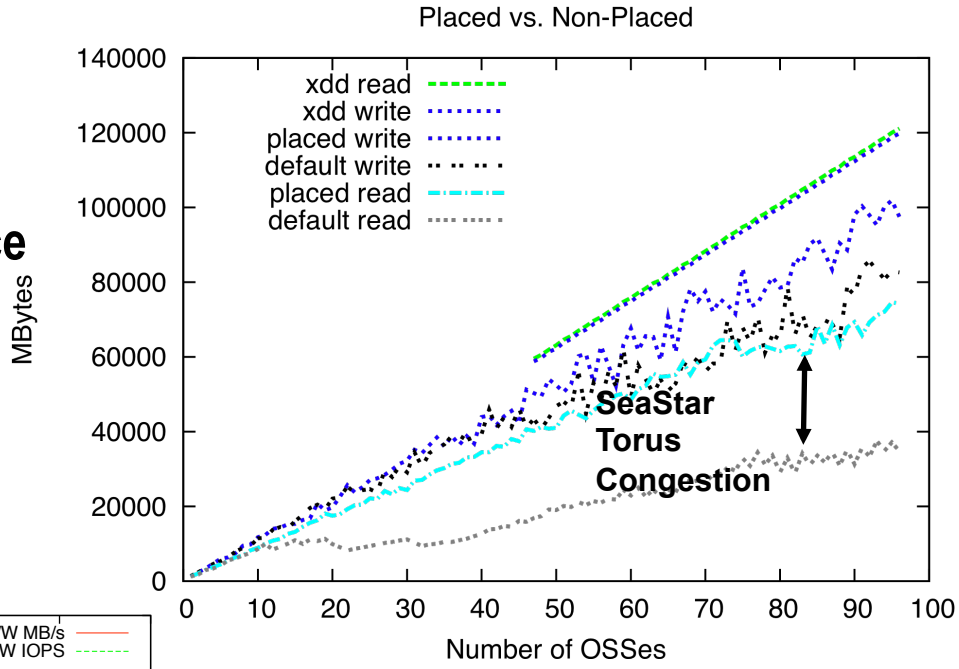
Lustre Router Nodes
run LNET router software forward requests from Lustre file system clients on compute nodes.

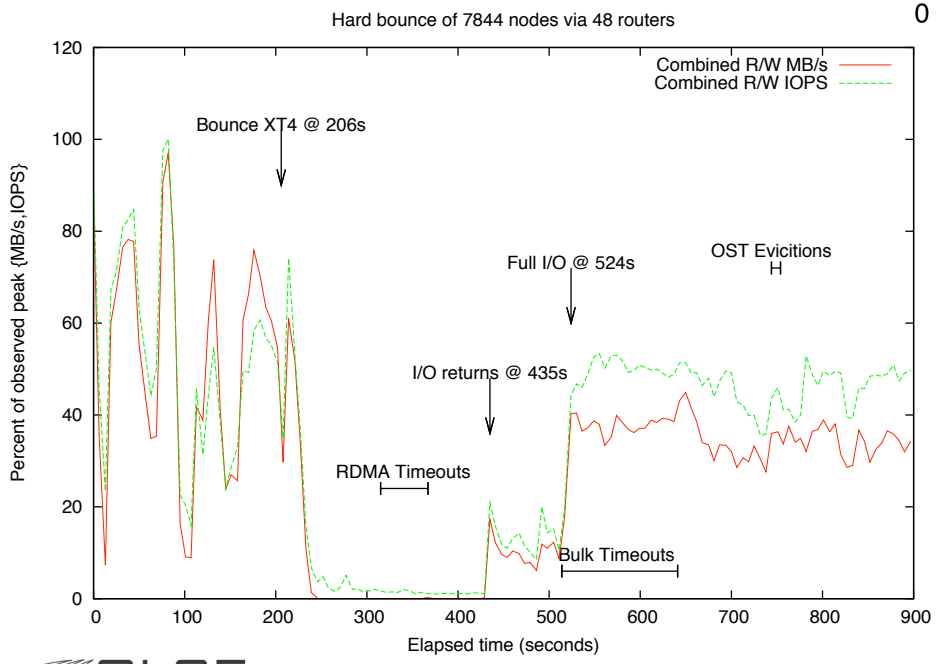
192 (XT5) and 48 (XT4) one dual core Opteron nodes with 8 GB of RAM each

OAK RIDGE National Laboratory

Snapshot of Technical Challenges Solved

- **Performance Innovations**
  - Improved SATA performance (2x Improvement)
  - Network congestion avoidance (2x improvement)
- **Scalability**
  - 26,000 file system clients and counting





- **Fault tolerance Design Features**
  - System Area Network
  - I/O Servers
  - Storage Arrays
- **Infiniband Support on XT SIO**

OAK RIDGE National Laboratory

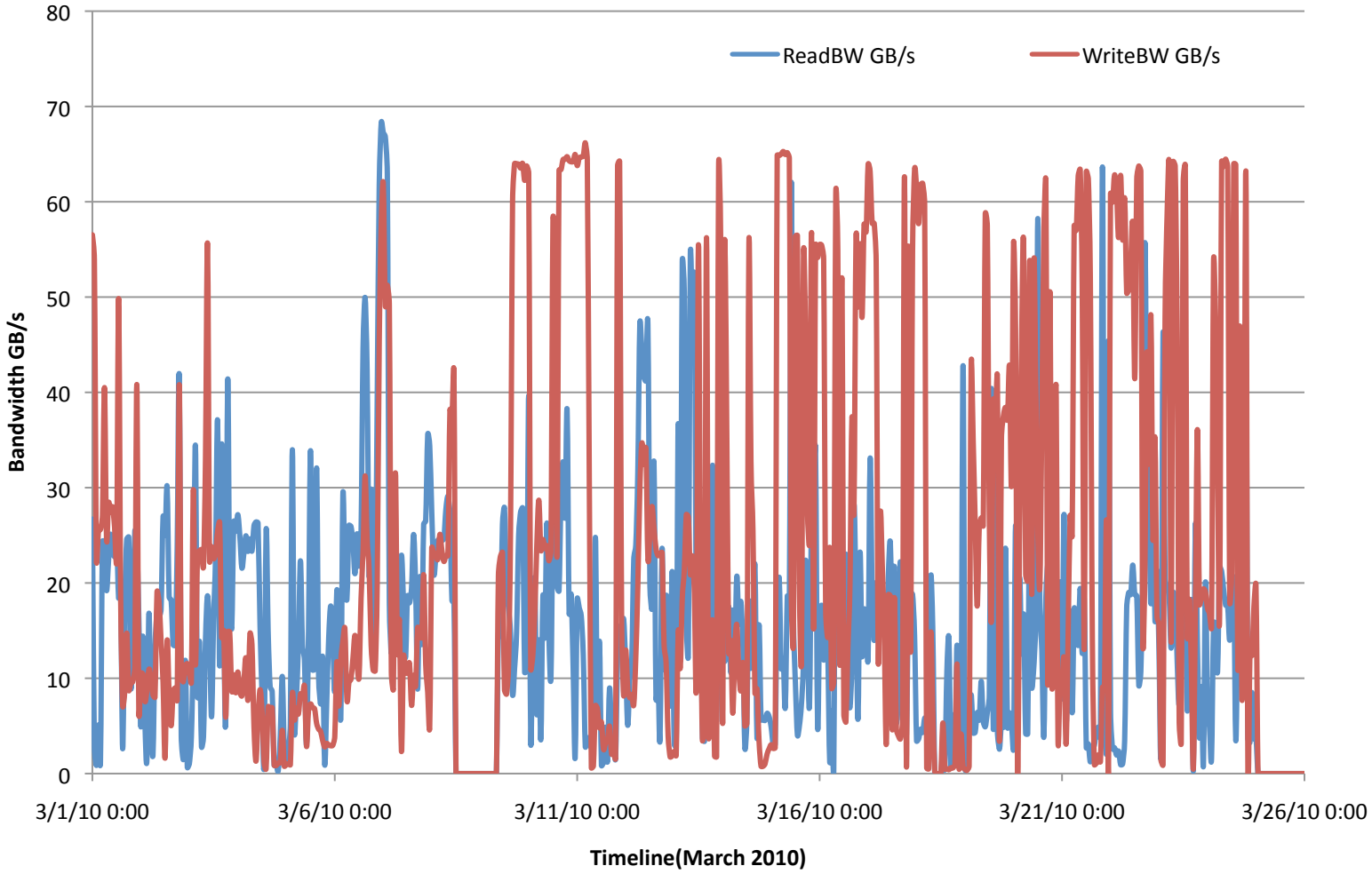
Operational Experiences

- Mixed workloads can achieve ~75% of peak read performance and ~60% of peak write performance
  - Write back caching disabled - limits aggregate performance
  - Network (IB and SeaStar) congestion limits aggregate performance
- Single scratch and project directories across all platforms have been well received
  - Project areas are heavily used
  - Very few users are using Jaguar XT4 direct attached storage
- System resiliency has been a win
  - Engineering for failure pays untold dividends
  - Over 25 saves by DM-multipath alone

Operational Challenges

- Management of a system of this scale can be daunting
  - ¼ billion+ files make this challenging
  - Many custom tools developed, more on their way
- Scalable management tools are lacking
  - Diagnosis and recovery utilities to respond to failures
  - System log analysis
    - Log parsing from over 200 servers
    - Log correlation to compute node and application
  - InfiniBand fabric monitoring
- Single MDS performance
  - Single application collectively creating 220K files
  - Unscalable I/O can impact the entire user population

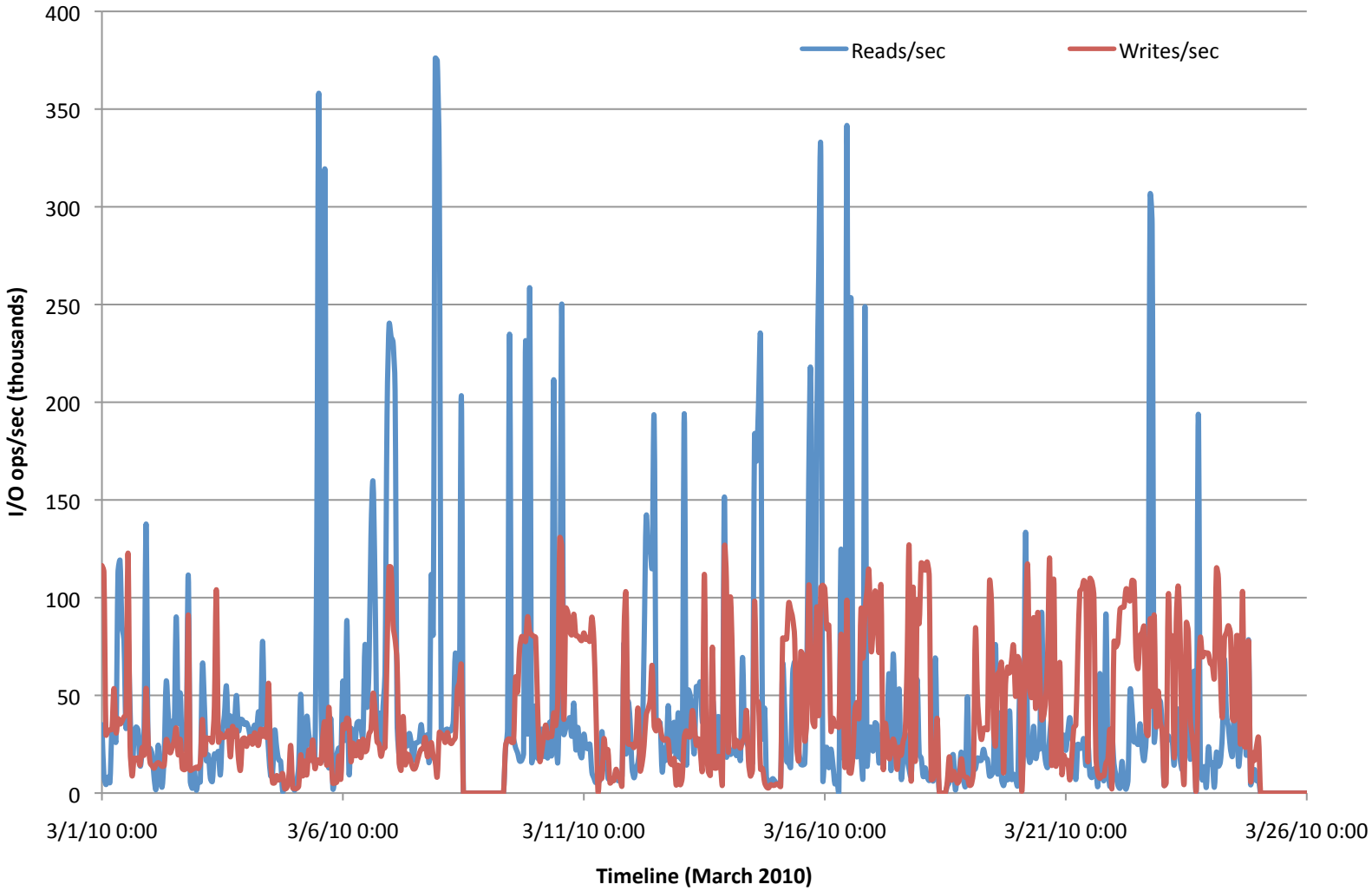
Designed to Support Peak Performance



Max data rates (hourly) on ½ of the storage controllers

OAK RIDGE National Laboratory

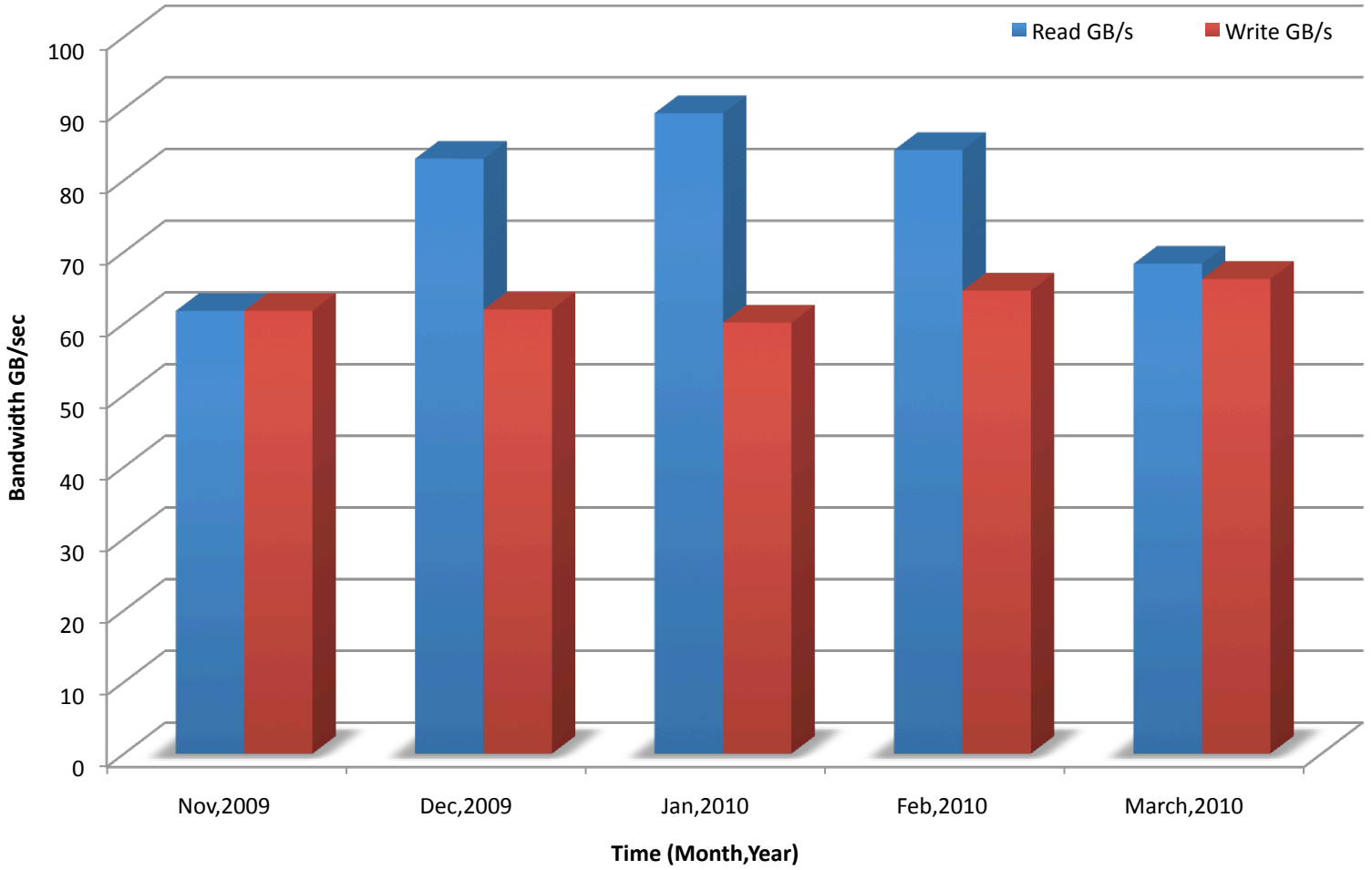
Many Workloads are IOPs Intensive



Max IOPs (hourly) on ½ of the storage controllers

OAK RIDGE National Laboratory

Peak Bandwidths - Past 5 Months

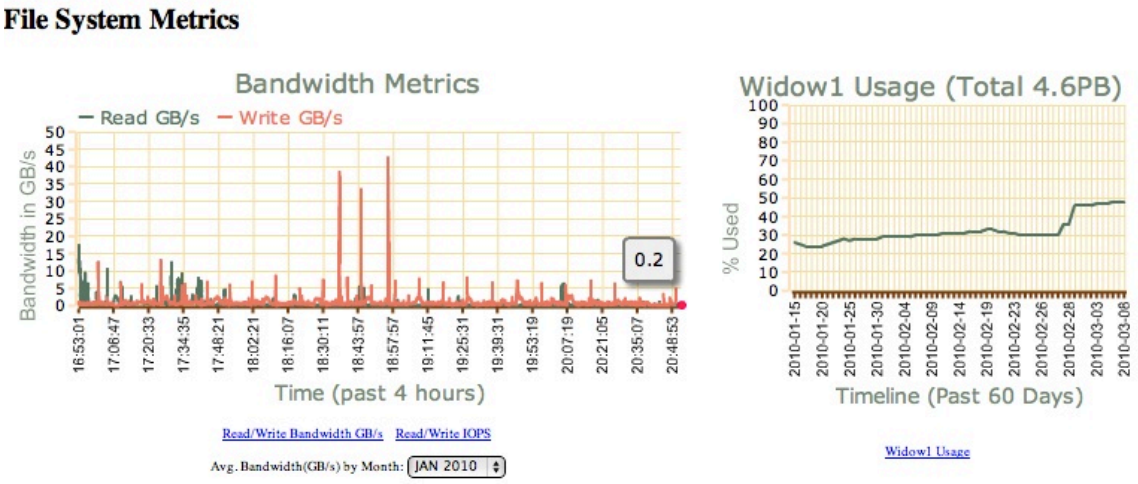


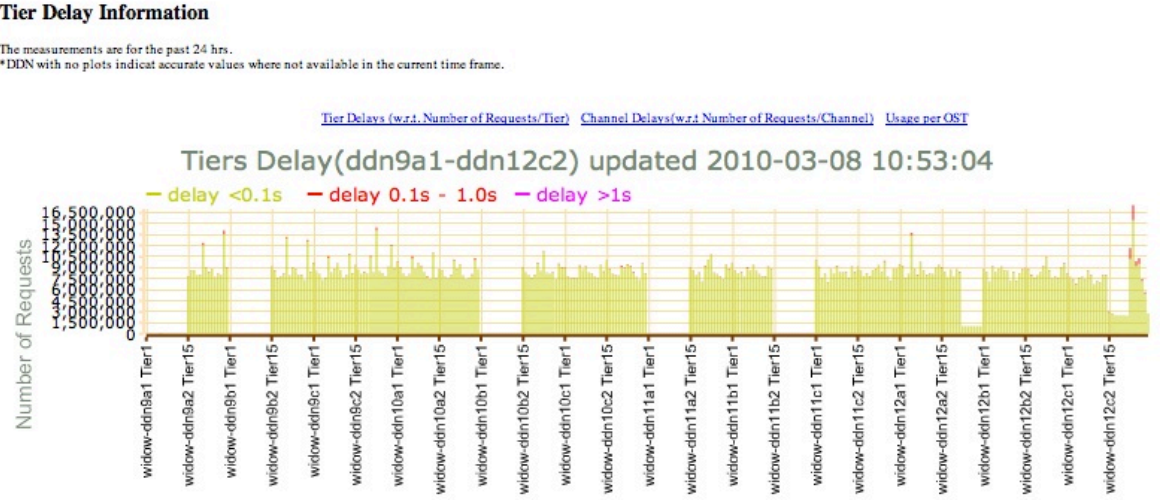
Max data rates (monthly) on ½ of the storage controllers

OAK RIDGE National Laboratory

Managing Our Environment: System Analytics

- System Metrics
  - Centralized Collection of System Performance, Faults and other Metrics
  - Historical Analysis
  - Snapshots
  - Web Interface
  - Future Integration with DS
- Event Log Analysis (under development)
  - Determine Fault Patterns
  - Cluster-based Approach
  - Correlation of Events
  - Log Visualization





LustreDU (Beta)

- Provide functionality similar to the Unix ‘du’ command.
  - Regular du command could take hours (or days) to run because of the size of the filesystem.
- LustreDU makes use of the output of ne2scan
  - ne2scan is already run regularly as part of the file purge cycle
  - LustreDU ‘piggybacks’ on the ne2scan run and thus doesn’t require a separate scan of the entire file system
- One main executable plus daemons running on each OSS
  - Output from ne2scan (which runs on the MDS) contains all needed metadata except for object size
  - Must query an OST for each object’s size
- Two kinds of output: text file or MySQL database

Parallel Data Utilities

- SPDCP (Released!)
  - A parallel Lustre aware copy tool
  - Uses multiple compute nodes for data movement
  - Provides near cp semantics
  - Preserves Lustre striping
  - Demonstrated speedups of 93x over cp
- PLTAR (Beta)
  - A parallel Lustre aware tar tool
  - Uses multiple compute nodes for data movement
  - Preserved Lustre striping
  - POSIX compliant tar archive
  - Demonstrated speedups of 49x over tar

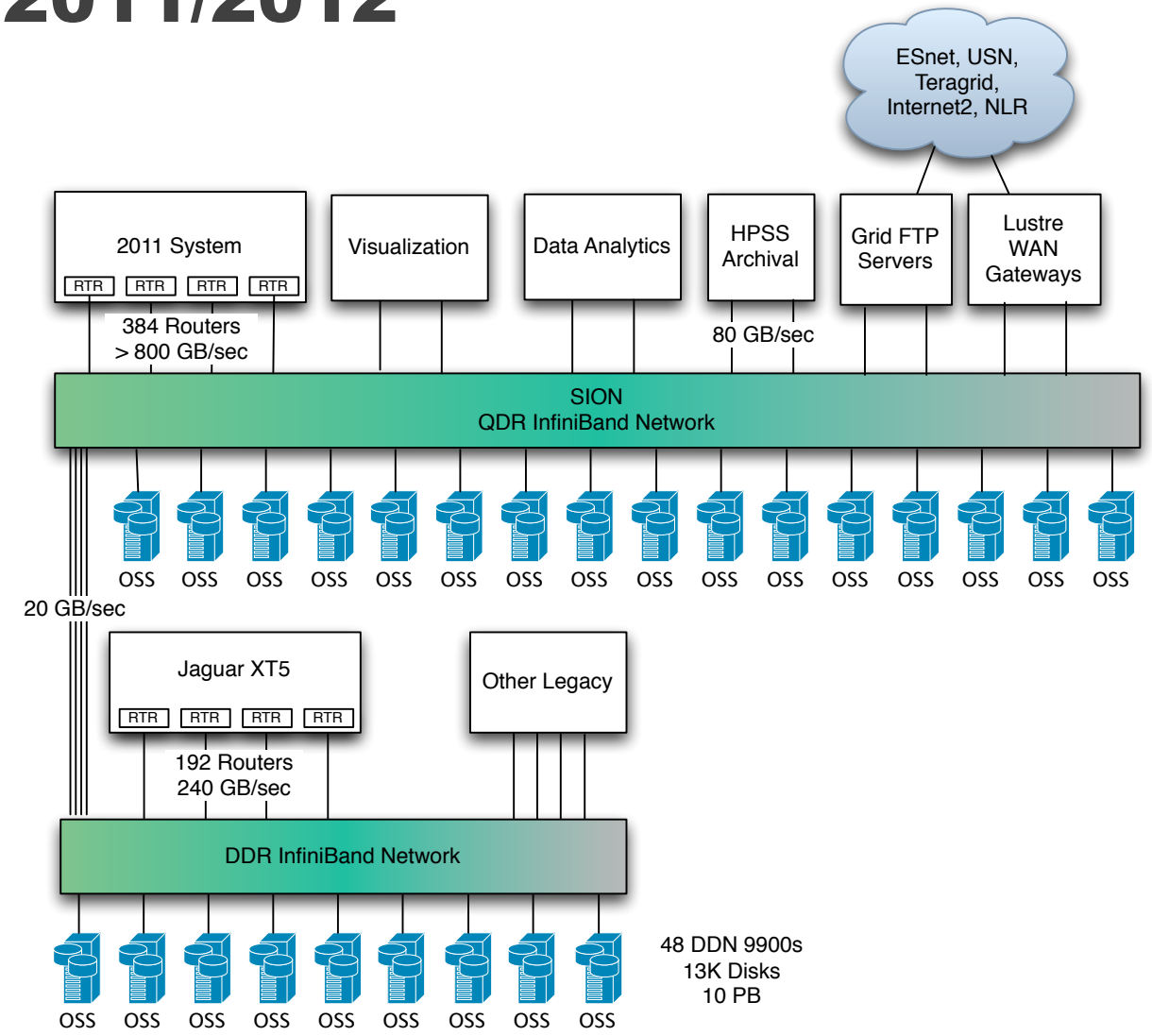
OAK RIDGE National Laboratory

Path Forward for Spider

- Complete at scale testing of fine-grained routing
  - Target deployment during our 1.8.x deployment
- Large EA support still outstanding
  - Shared file limit of 160 OSTs
  - How do we deploy this with 1.8.x?
- Continue to explore MDS performance enhancements
  - Reduce locking overhead
  - Consider multiple center wide file systems as risk mitigation strategy
    - May target 3 shared file systems in final configuration rather than 2.
    - Provides flexibility to isolate by project or workload
- Build and improve existing diagnostic and administrative tools
  - Improved tools to locate candidate files following failure
  - Improved performance monitoring and diagnostic tools
  - Mechanisms to correlate system failures / interrupts with aberrant performance

Challenges for 2011/2012

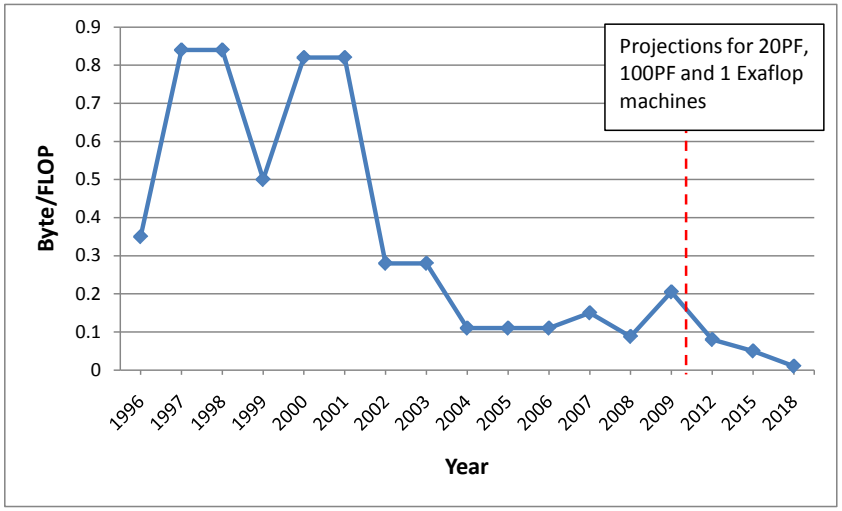
- ~20 Petaflop peak
- > 1 Petabyte system memory
- ~20K Lustre clients
- 3.4x bandwidth Increase
- Up to 7x capacity increase

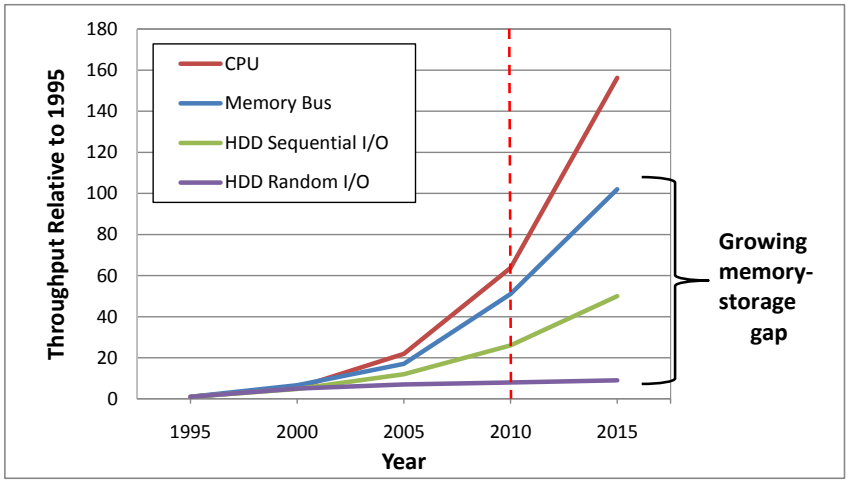


OAK RIDGE National Laboratory

The Memory-Storage Gap Problem

- Memory constraints may place more burdens on the file system (productive I/O)
- Storage system latency unlikely to improve (HDD)
- Sequential bandwidth will continue to improve but at a fraction of the rate of densities





Courtesy Brad Settlemyer & Sudharshan Vazhkudai (ORNL)

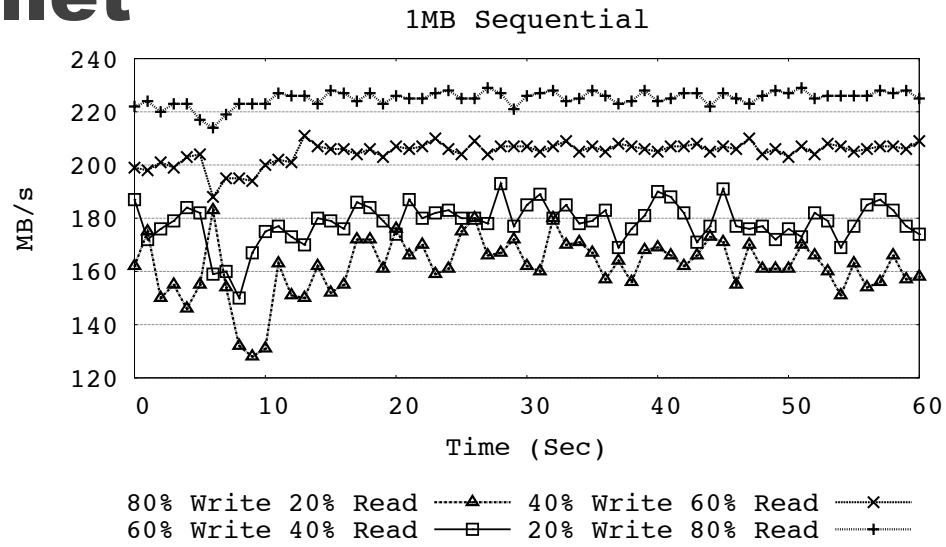
OAK RIDGE National Laboratory

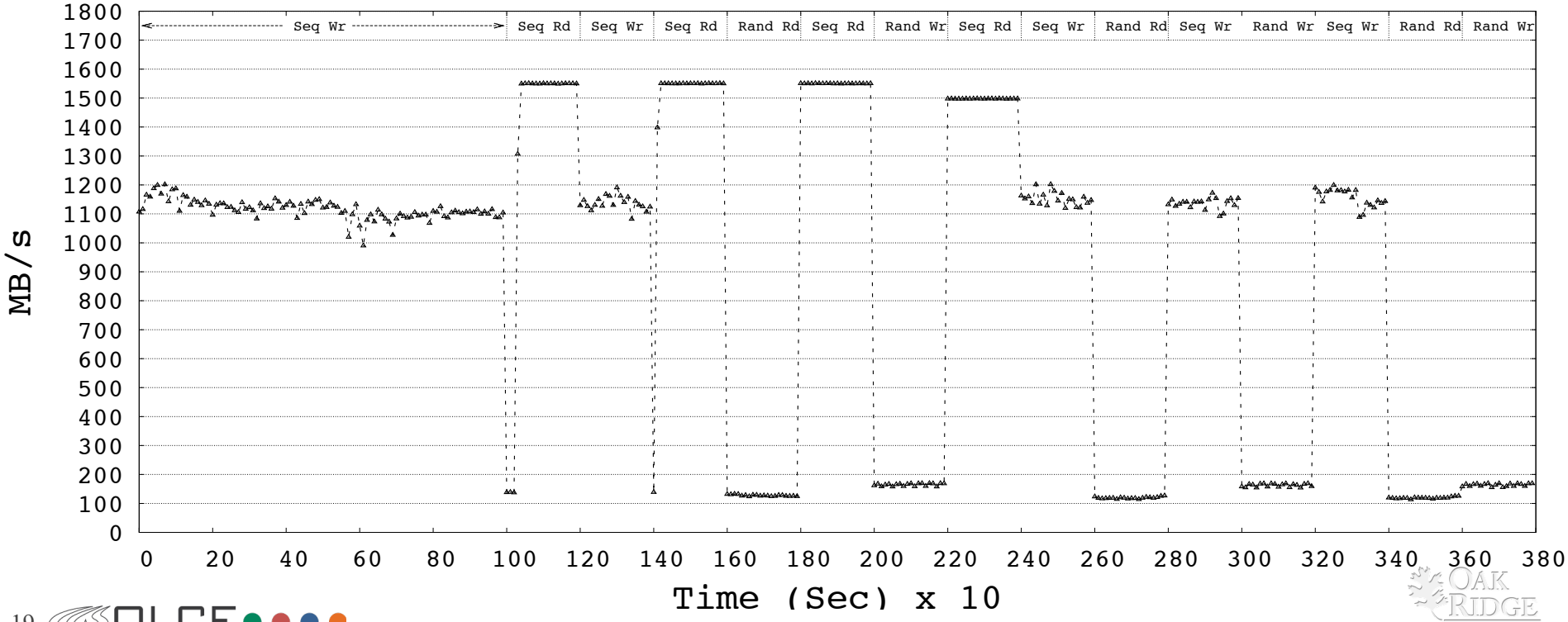
What About Flash?

- Lifetime constraints
- Performance variability
- Integration
  - Standalone FLASH file system
  - Hierarchical with bleed down strategies
- Packaging
  - SSD RAID
  - PCIE
- Reliability
- Proper evaluation of Flash requires detailed workload characteristics
  - Duty cycles
  - Aggregate I/O characterization
  - Individual application I/O characterization

Flash – Not a Silver Bullet

- Extreme Performance Variability
- Workload dependent





Storage Media

- Expect marginal improvements in disk drive performance (bandwidth)
  - Latency has flat-lined
- Mixed media environment may be required to meet capacity and bandwidth requirements
  - separate file systems but globally accessible
- Integration of Flash needs further evaluation

Media Type	Total Drives
Near-line SAS	20,000
Enterprise SAS	10,000
MLC Flash	4,000

Estimated number of drives required to meet performance of ~800 GB/sec

OAK RIDGE National Laboratory

OLCF-3 File System

- Lustre is our plan of record for the 2011/2012 system
  - Proven scalability and performance on Jaguar XT5
  - Leverages our skills and experience in Lustre
- An evolution of our current parallel file system infrastructure
  - Lower risk, high confidence
  - Component counts (Disks, I/O nodes, etc.) do not increase significantly over our current system (order 2x)
- New features in Lustre will be needed to improve scalability and reliability
  - Improved metadata performance
  - Improved recovery performance
  - Improved manageability

Why Lustre?

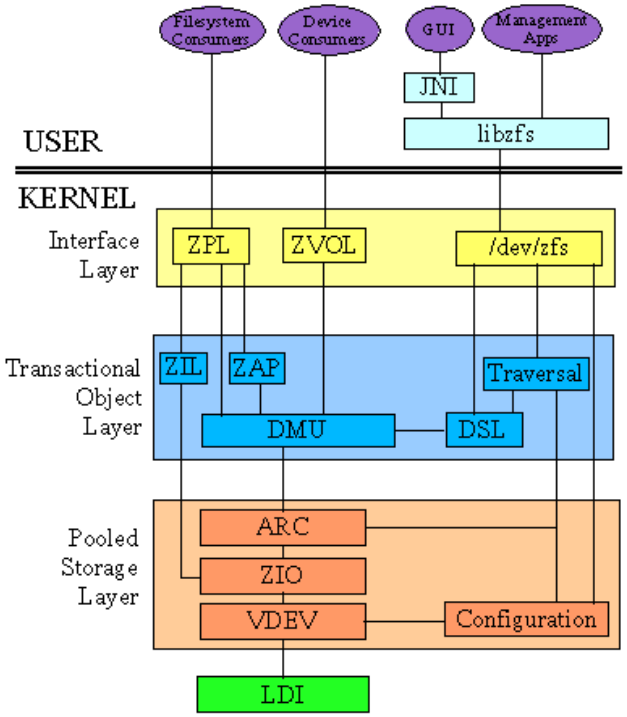
- Meets our requirements today
  - Performance, Scalability, Reliability
  - No other file system meets our current requirements
- Planned features meet our projected requirements
- Deep skills in the Lustre parallel file system
  - Development, Deployment, and Operations
  - Active contributor in the Lustre community
- The only open source parallel file system
  - Provides risk mitigation strategies that closed source does not
  - Broadening community of developers

Lustre Features Needed for OLCF-3

- Improved Metadata Server (MDS) Performance
  - SMP scaling improvements (Single MDS Server)
  - Clustered Metadata Servers (schedule risk)
    - Ordered operations
    - Epochs
- Improved data integrity protection/detection
  - Integrated checksums
  - Network packet protection – LNET
    - End-to-end desirable
- Support for larger LUNS
  - Currently limited to 8 TB
  - ORACLE/ORNL work has increased this to 16 TB (Lustre 1.8.2)
  - May need 32 TB support or beyond
    - ZFS preferred path, ext4 enhancements as fallback
- Predictable performance
  - Network Request Scheduling
  - Quality of Service Mechanisms

Lustre and ZFS

- Lustre currently uses an enhanced ext4 file system for object and metadata storage
- Lustre moving to add support for the ZFS file system for object and metadata storage
- ZFS features
  - Capacity
    - large volume support (512 TB)
    - trillions of files
  - Reliability and Integrity
    - Copy on write
    - Integrated checksums
    - Online integrity checking and reconstruction
    - Snapshots
- Performance concerns
- LLNL porting ZFS – kDMU to Linux
  - Appears to be unsupported at the moment
- Licensing
  - Linux is released under the GPL, ZFS is CDDL
  - Redistribution of Linux + ZFS may have licensing issues



OAK RIDGE National Laboratory

Implications of License Incompatibilities

- Vendors such as DDN and LSI may not be able to sell integrated solutions based on Lustre + ZFS + Linux
- End users can integrate and be in compliance
  - ORNL, LLNL, etc
  - What about support?
- Lustre + enhanced ext4 + Linux will remain an option for vendors
  - Other backend file systems could also be supported (BTRFS)
- Is Lustre + Open Solaris an option for vendors?

Mitigation Strategies for OLCF-3

- Actively engaging Oracle on Lustre
- Continuing our partnerships with key Lustre users
  - ORNL led weekly Lustre community meetings are ongoing
  - Oracle file system engineers attending
- Evaluation of a number of Lustre configurations
  - Lustre 1.8 with enhancements
  - Lustre 2.0 (Linux, Open Solaris)
  - ldiskfs and ZFS testing
- Evaluating alternative file systems (risk mitigation)
  - GPFS
  - Panasas

Preparing for OLCF-3 - Testbeds

- Deployment of next generation file system testbed in 2010
- Representative of OLCF-3 storage system
  - Multiple integrated storage controller technologies
  - Mixed SAS/SATA environment
  - InfiniBand
  - Exploring flash integration
- Capable of testing a variety of file systems and technologies
- Additional build out of testbeds in 2011

Final Thought: Listening to Your Users

- Need for an open forum to discuss
  - User experiences
  - User requirements
  - Next-generation features
- Allows aggregation and concentration of requirements
- Lustre User Group could do this
  - Allow users to run LUG!
    - Run by a LUG board of directors from major Lustre sites
  - Cray User Group and IBM SPXXL are excellent examples

Questions?

- Contact info:

Galen Shipman

Group Leader, Technology Integration

865-576-2672

gshipman@ornl.gov