Oak Ridge National Laboratory

Meeting the I/O Demands of the World's Most Powerful Scientific Computing Complex

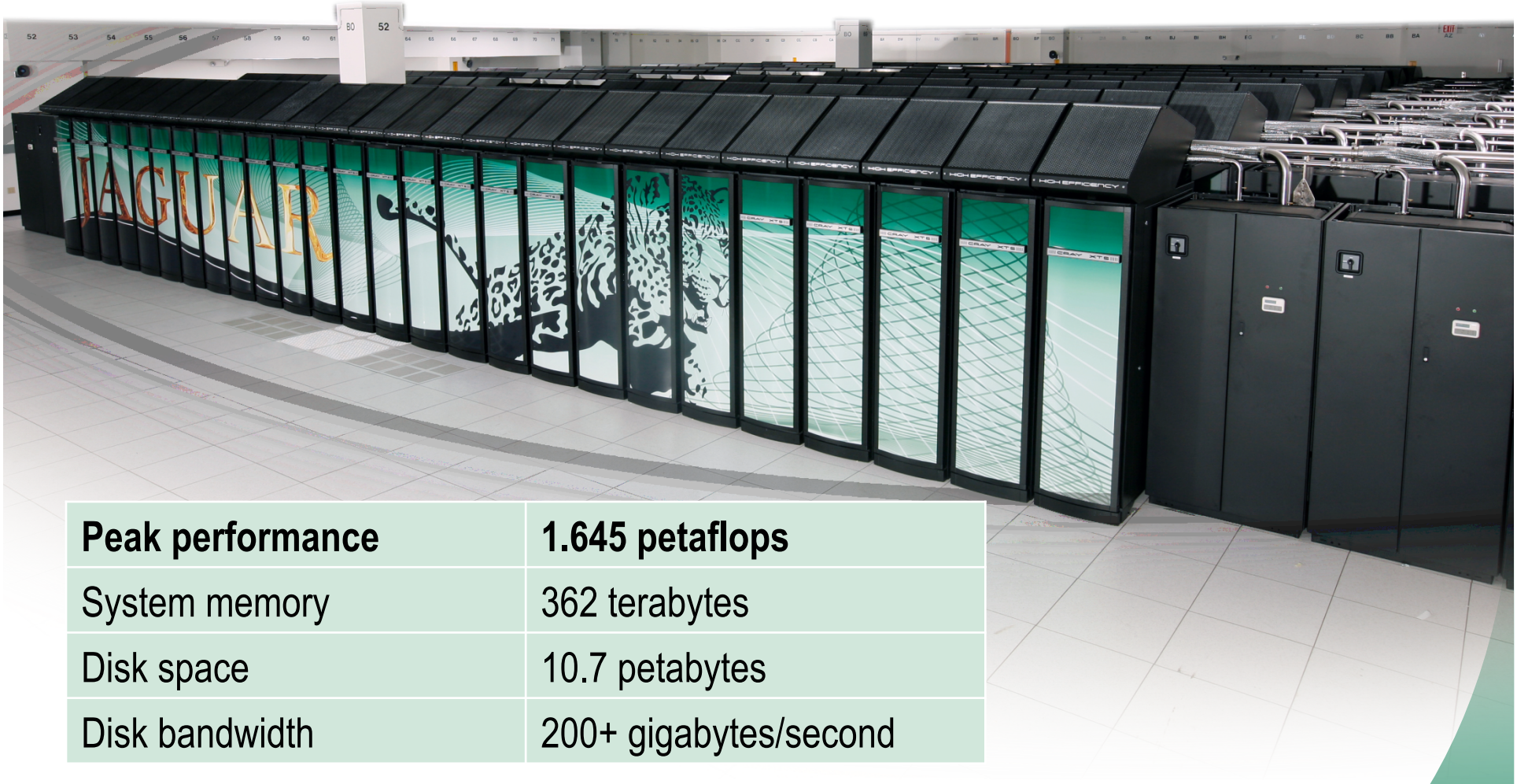
Presented by:
Galen M. Shipman

April 15, 2009

Managed by UT-Battelle for the
U. S. Department of Energy

OAK RIDGE National Laboratory

Jaguar: World's most powerful computer
Designed for science from the ground up

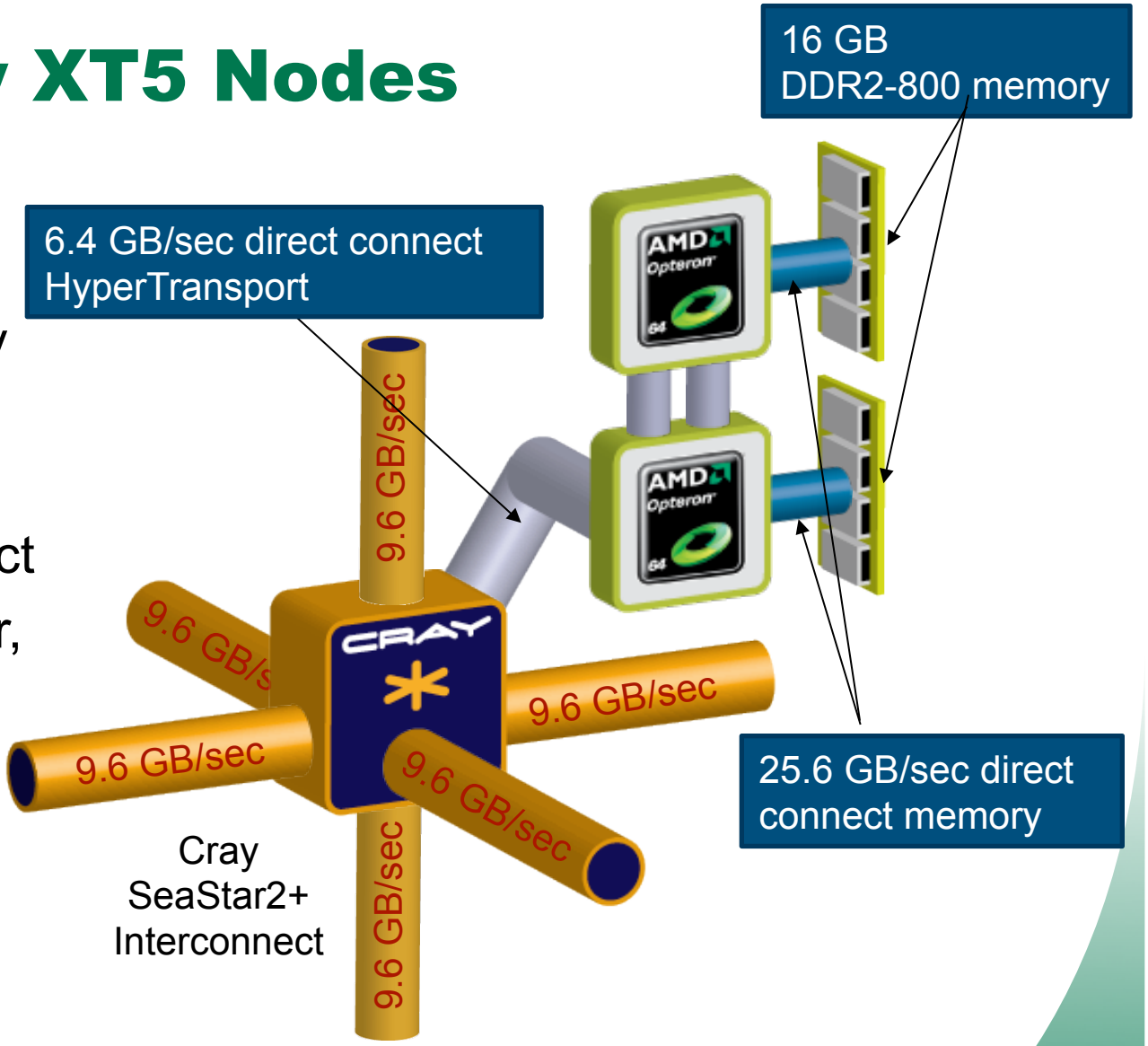


Peak performance	**1.645 petaflops**
System memory	362 terabytes
Disk space	10.7 petabytes
Disk bandwidth	200+ gigabytes/second

Jaguar's Cray XT5 Nodes

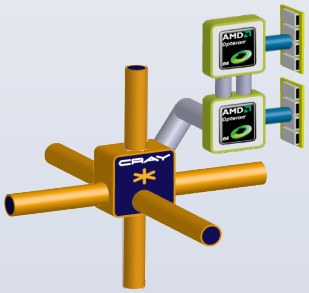
- Powerful node improves scalability
- Large shared memory
- OpenMP Support
- Low latency, High bandwidth interconnect
- Upgradable processor, memory, and interconnect

GFLOPS	76.3
Memory (GB)	16
Cores	8
SeaStar2+	1



Building the Cray XT5 System

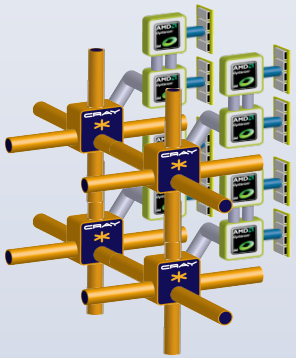
Node
73.6 GF
16 GB



1x1x1

4x

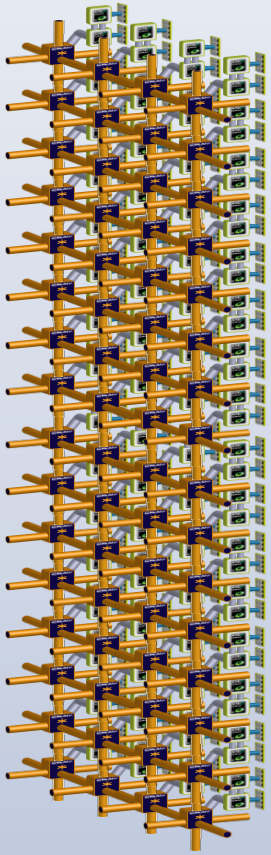
Blade
294 GF
64 GB



1x2x2

24x

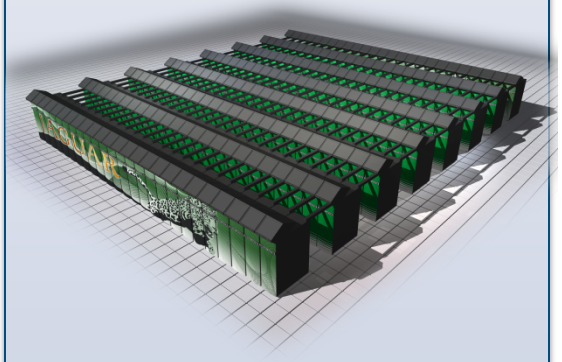
Rack
7.06 TF
1.54 TB



1x4x16

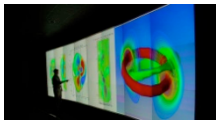
200x

System
1382 TF
300 TB



25x32x16

Current LCF Infrastructure



Everest Powerwall

Lens

Smoky

SION

192x

48x

XT5

XT4

Wolf

Data Archive

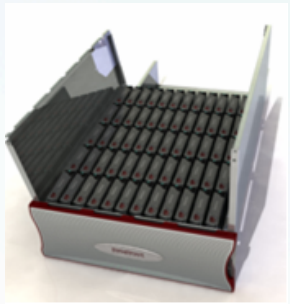
HPSS

Accessible via Ethernet LAN

Current LCF File Systems

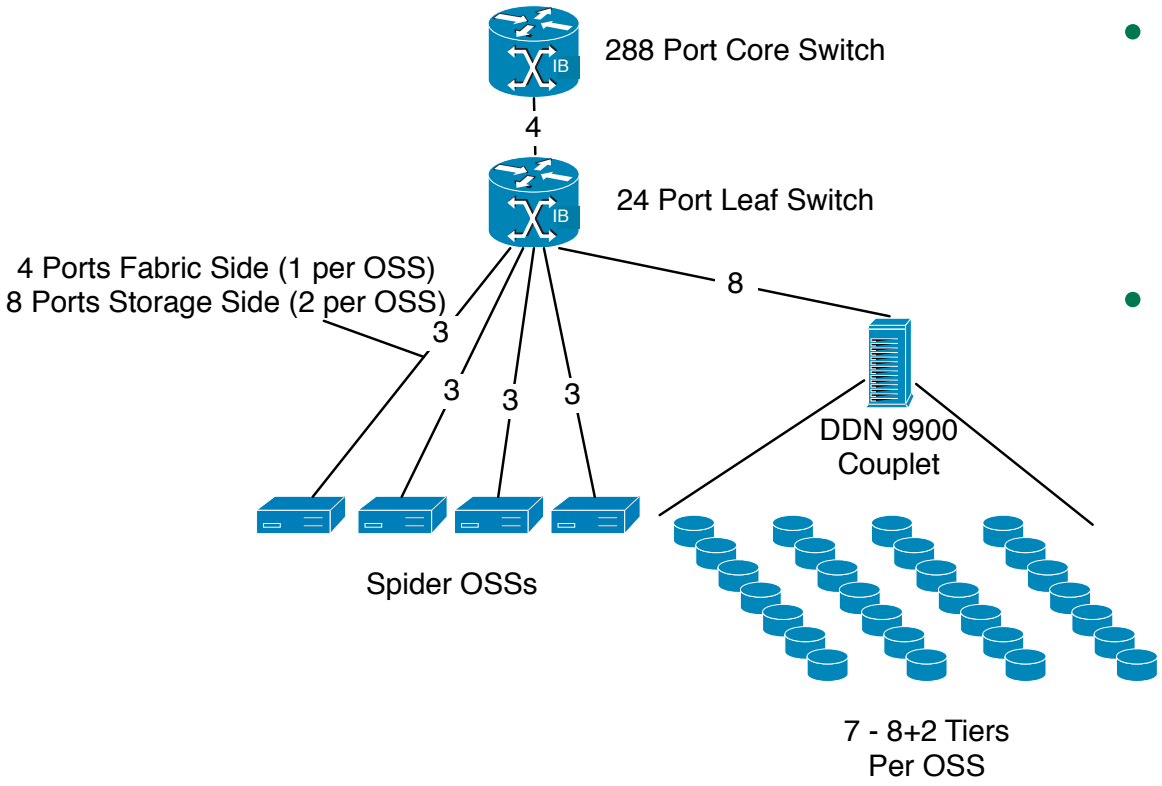
System	Path	Size	Throughput	OSTs
Jaguar XT5				
	/lustre/scratch	4198 TB	> 100 GB/s	672
Jaguar XT4				
	/lustre/scr144	284 TB	> 40 GB/s	144
	/lustre/scr72a	142 TB	> 20 GB/s	72
	/lustre/scr72b	142 TB	> 20 GB/s	72
	/lustre/wolf-ddn (login nodes only)	672 TB	> 4 GB/s	96
Lens, Smoky				
	/lustre/wolf-ddn	672 TB	> 4 GB/s	96

Center-wide File System



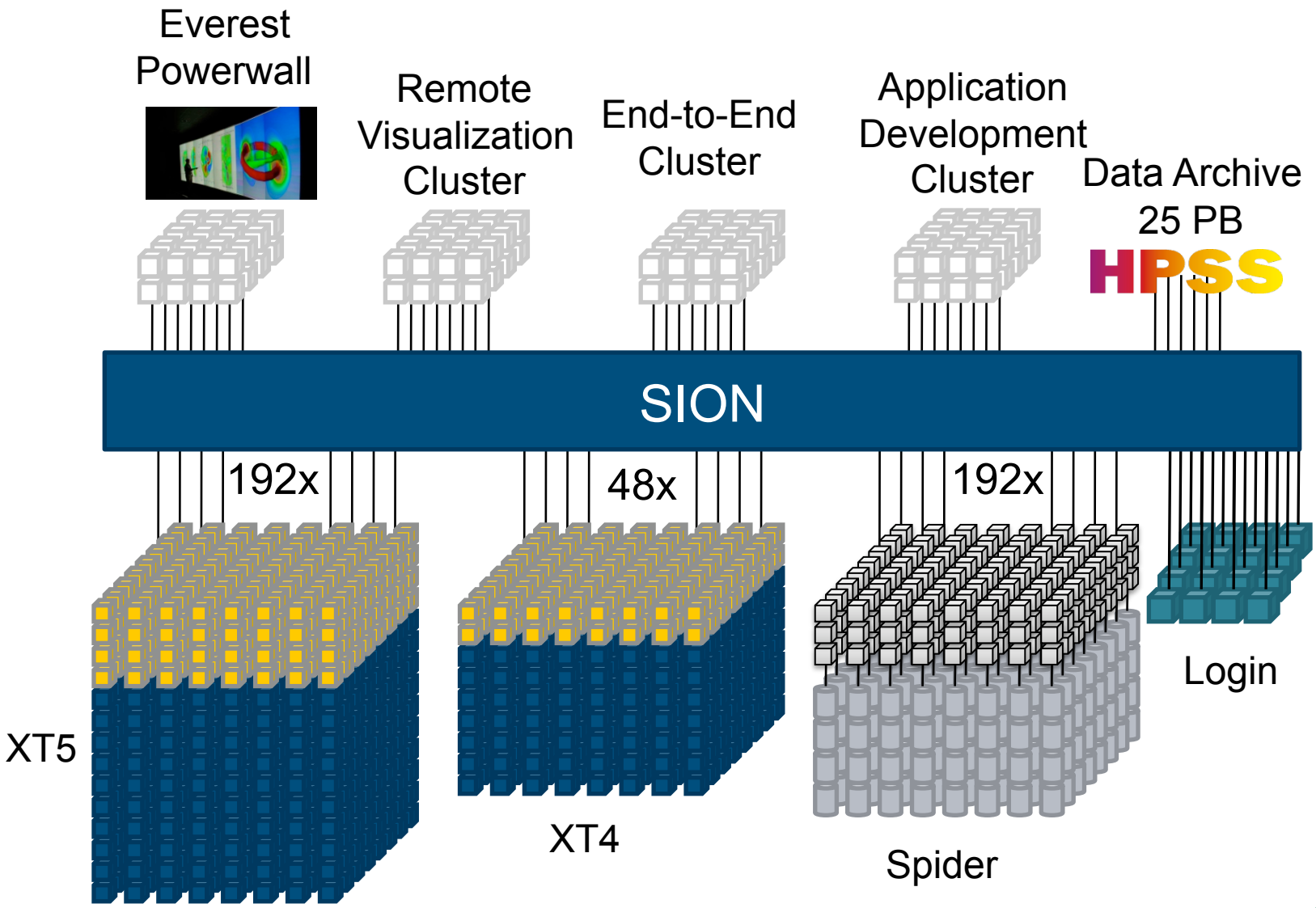
- "Spider" will provide a shared, parallel file system for all systems
  - Based on Lustre file system
- Demonstrated bandwidth of over 200 GB/s
- Over 10 PB of RAID-6 Capacity
  - 13,440 1 TB SATA Drives
- 192 Storage servers
  - 3 TeraBytes of memory
- Available from all systems via our high-performance scalable I/O network
  - Over 3,000 InfiniBand ports
  - Over 3 miles of cables
  - Scales as storage grows
- Undergoing system checkout with deployment expected in summer 2009

Spider Couplet View



- **3 Couplets complete with 12 OSSes and 3 IB Leaf switches in 2 DDN 9900 cabinets**
- **16 Scalable Storage Clusters for a total of 48 couplets and 192 OSSes in 32 racks**

Future LCF Infrastructure



Future LCF File Systems

System	Path	Size	Throughput	OSTs
Jaguar XT5				
	/lustre/widow0	4198 TB	> 100 GB/s	672
	/lustre/widow1	4198 TB	> 100 GB/s	672
Jaguar XT4				
	/lustre/widow0	4198 TB	> 100 GB/s	672
	/lustre/widow1	4198 TB	> 100 GB/s	672
	/lustre/scr144	284 TB	> 40 GB/s	144
	/lustre/scr72a	142 TB	> 20 GB/s	72
	/lustre/scr72b	142 TB	> 20 GB/s	72
Lens, Smoky				
	/lustre/widow0	4198 TB	> 100 GB/s	672
	/lustre/widow1	4198 TB	> 100 GB/s	672

Benefits of Spider

- Accessible from all major LCF resources
  - Eliminates file system “islands”
  - Eliminates the need for data transfers between XT4, XT5, Lens and Smoky
    - Currently limited to Ethernet LAN bandwidth constraints
- Accessible during maintenance windows
  - Spider will remain accessible during XT4 and XT5 maintenance
  - Users will be able to access the file system from other LCF systems such as Lens and Smoky as well as remotely via GridFTP or bbcp

Benefits of Spider

- Unswept Project Spaces
  - Will provide larger area than $HOME
  - Not backed up, use HPSS
  - The Data Storage council is working through formal policies now
- Higher performance HPSS transfers
  - XT Login nodes no longer the bottleneck
  - Other systems can be used for HPSS transfers which allow HTAR and HSI to be scheduled on computes
- Direct GridFTP transfers
  - Improved WAN data transfers

Spider Status

- Demonstrated stability on a number of LCF systems
  - Jaguar XT5
  - Jaguar XT4
  - Smoky
  - Lens
  - All of the above..
    - Over 26,000 clients mounting the file system and performing I/O
- System Checkout is Ongoing
  - General Availability this Summer

Leadership role in Lustre Scalability

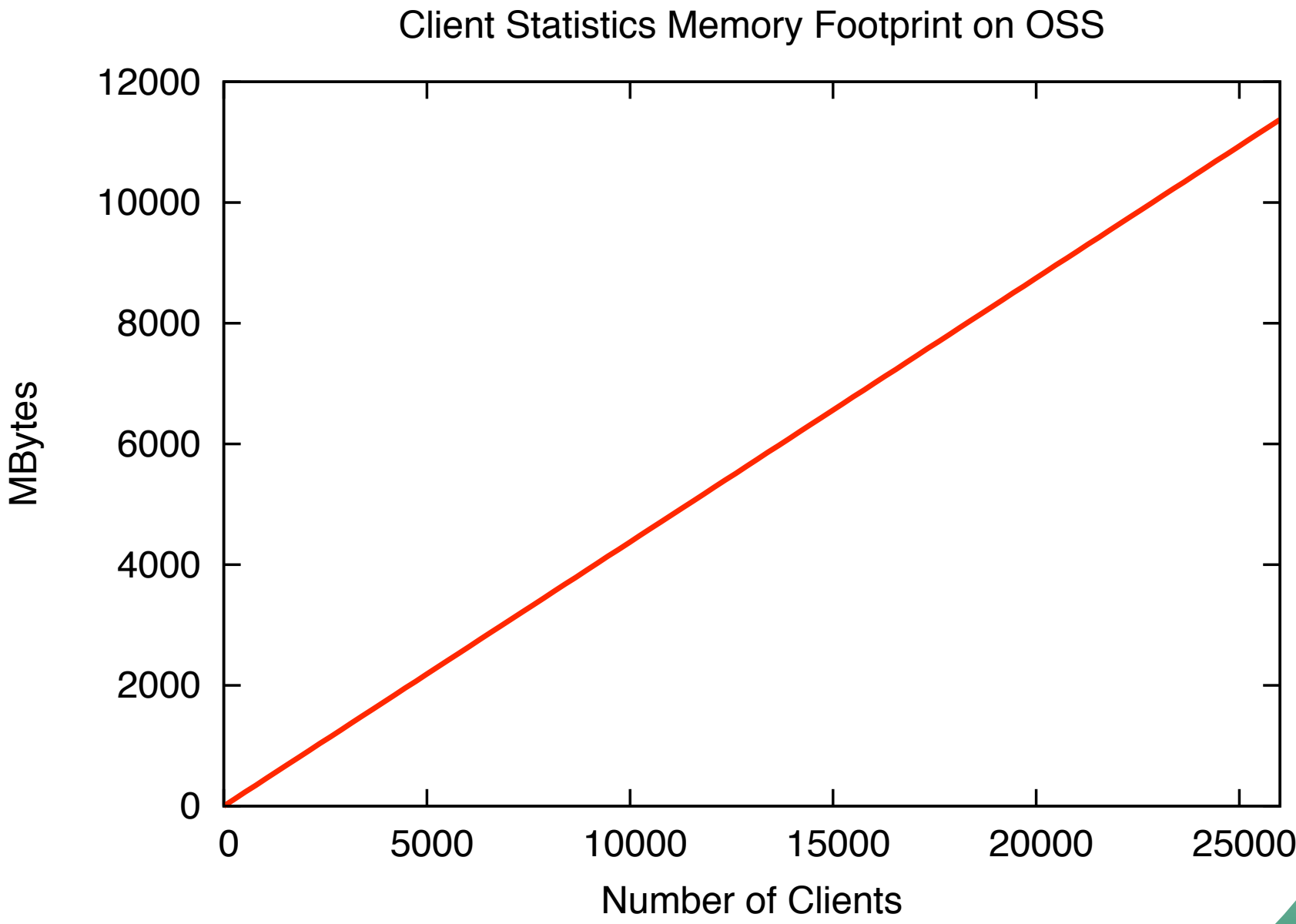
- Through the Lustre Center of Excellence we are driving the Lustre file system to unprecedented scale and performance
- 3 Onsite Lustre engineers work directly with Technology Integration Staff
- Driving collaboration with other stakeholders on improving Lustre performance, scalability and stability
  - 2 workshops this year to tackle these issues

Scaling to More Than 26,000 Clients

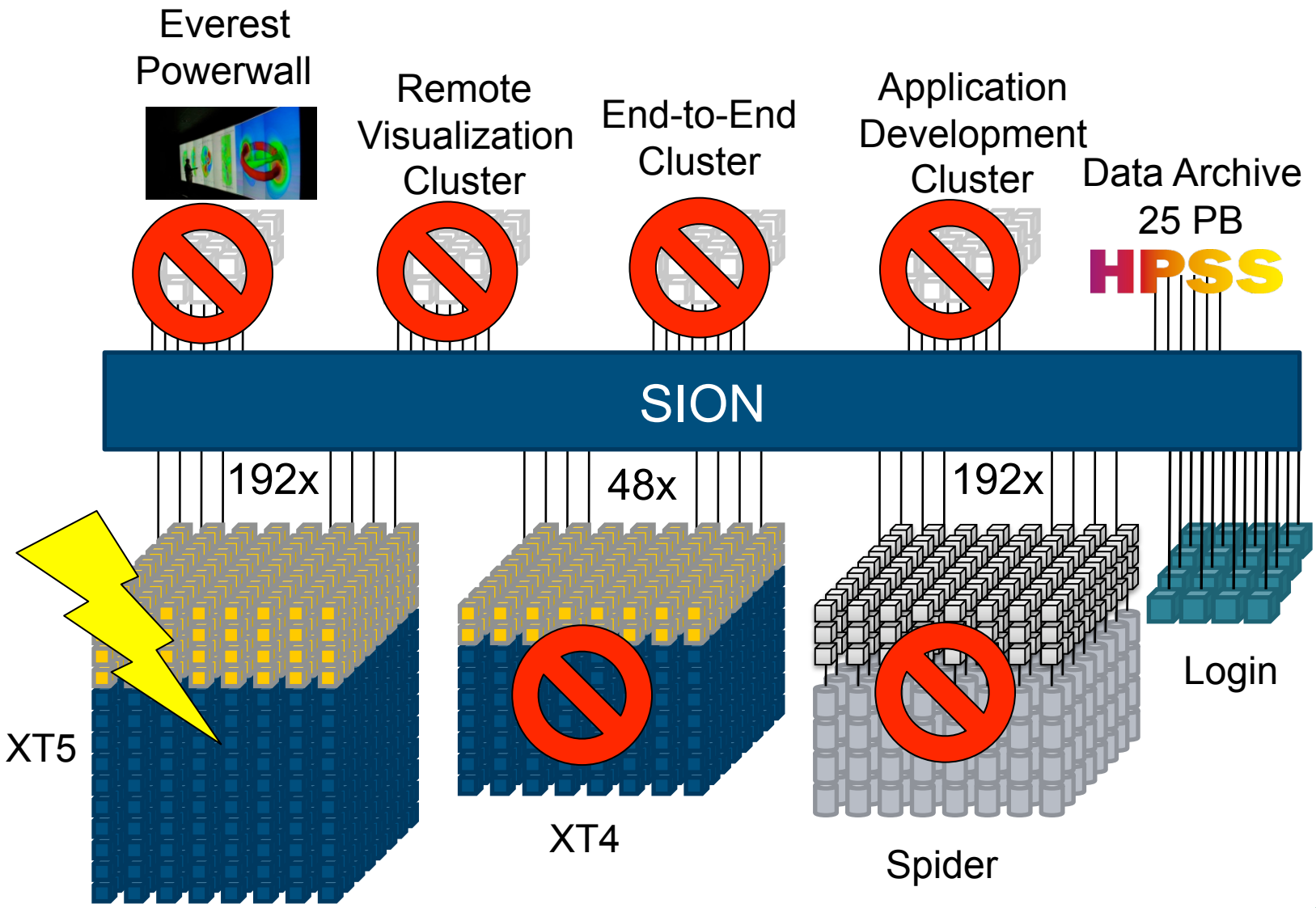
- 18,600 Clients on Jaguar XT5
- 7,840 Clients on Jaguar XT4
- Several hundred additional clients from various systems
- System testing revealed a number of issues at this scale

Scaling to More Than 26,000 Clients

- Server side client statistics
  - 64 KB buffer for each client for each OST/MDT/ MGT
  - Over 11GB of memory used for statistics when all clients mount the file system
  - OOMs occurred shortly thereafter
- Solution? Remove server side client statistics
  - Client statistics are available on computes
    - Not as convenient but much more scalable as each client is only responsible for his own stats


Client Statistics Memory Footprint on OSS
MBytes
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
Number of Clients

Surviving a Bounce

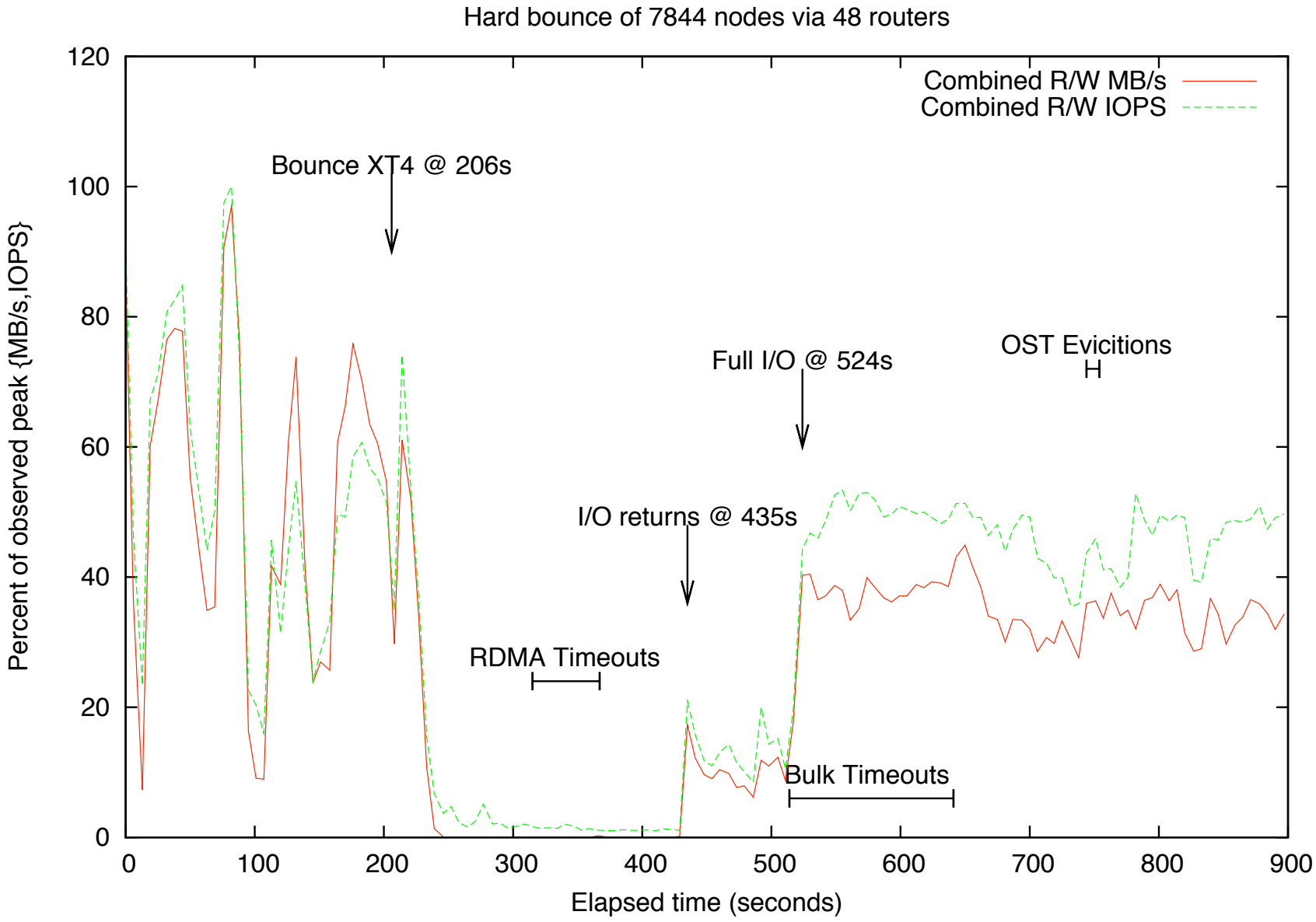


Challenges in Surviving an Unscheduled Jaguar XT4 or XT5 Outage

- Jaguar XT5 has over 18K Lustre clients
  - A hardware event such as a link failure may require rebooting the system
  - 18K clients are evicted!
- On initial testing a reboot of either Jaguar XT4 or XT5 resulted in the file system becoming unresponsive
  - Clients on other systems such as Smoky and Lens became unresponsive requiring a reboot

Solution: Improve Client Eviction performance

- Client eviction processing is serialized
- Each client eviction requires a synchronous write for every OST
- Current fix changes the synchronous write to an asynchronous write
  - Decreases impact of client evictions and improves client eviction performance
- Further improvements to client evictions may be required
  - Batching evictions
  - Parallelizing evictions


Hard bounce of 7844 nodes via 48 routers
Combined R/W MB/s
Combined R/W IOPS
Bounce XT4 @ 206s
Full I/O @ 524s
OST Evicitions
H
I/O returns @ 435s
RDMA Timeouts
Bulk Timeouts
Percent of observed peak {MB/s,IOPS}
Elapsed time (seconds)
0
20
40
60
80
100
120
0
100
200
300
400
500
600
700
800
900

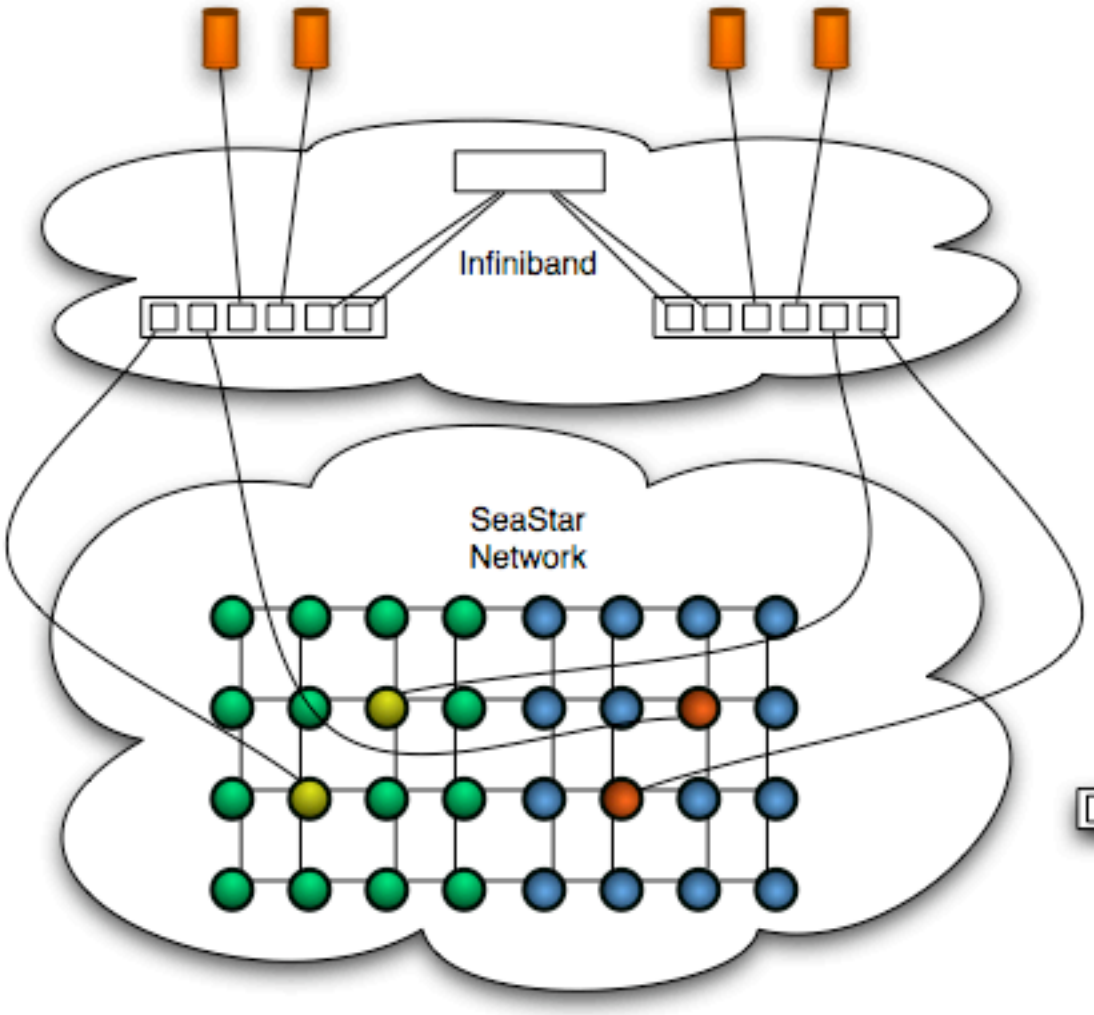
Improving Lustre Performance @ Scale

- **Multiple areas of Network Congestion**
  - **Infiniband SAN**
  - **SeaStar Torus**
  - **LNET routing doesn't expose locality**
    - **May take a very long route unnecessarily**
- **Assumption of flat network space won't scale**
  - **Wrong assumption on even a single compute environment**
  - **Center wide file system will aggravate this**
- **Solution - Expose Locality**
  - **Lustre modifications allow fine grained routing capabilities**

Design To Minimize Contention

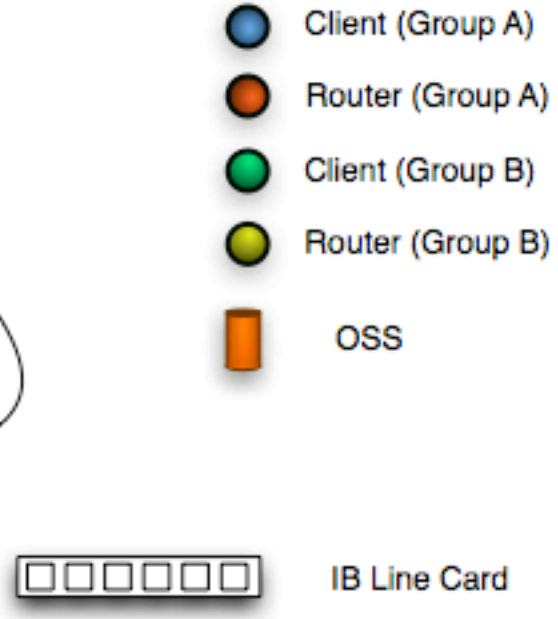
- **Pair routers and object storage servers on the same line card (crossbar)**
  - **So long as routers only talk to OSSes on the same line card contention in the fat-tree is eliminated**
  - **Required small changes to Open SM**
- **Place routers strategically within the Torus**
  - **In some use cases routers (or groups of routers) can be thought of as a replicated resource**
  - **Assign clients to routers as to minimize contention**
- **Allocate objects to "nearest" OST**
  - **Requires changes to Lustre and/or I/O libraries**

Intelligent LNET Routing



Clients prefer specific routers to these OSSes - minimizes IB congestion (same line card)

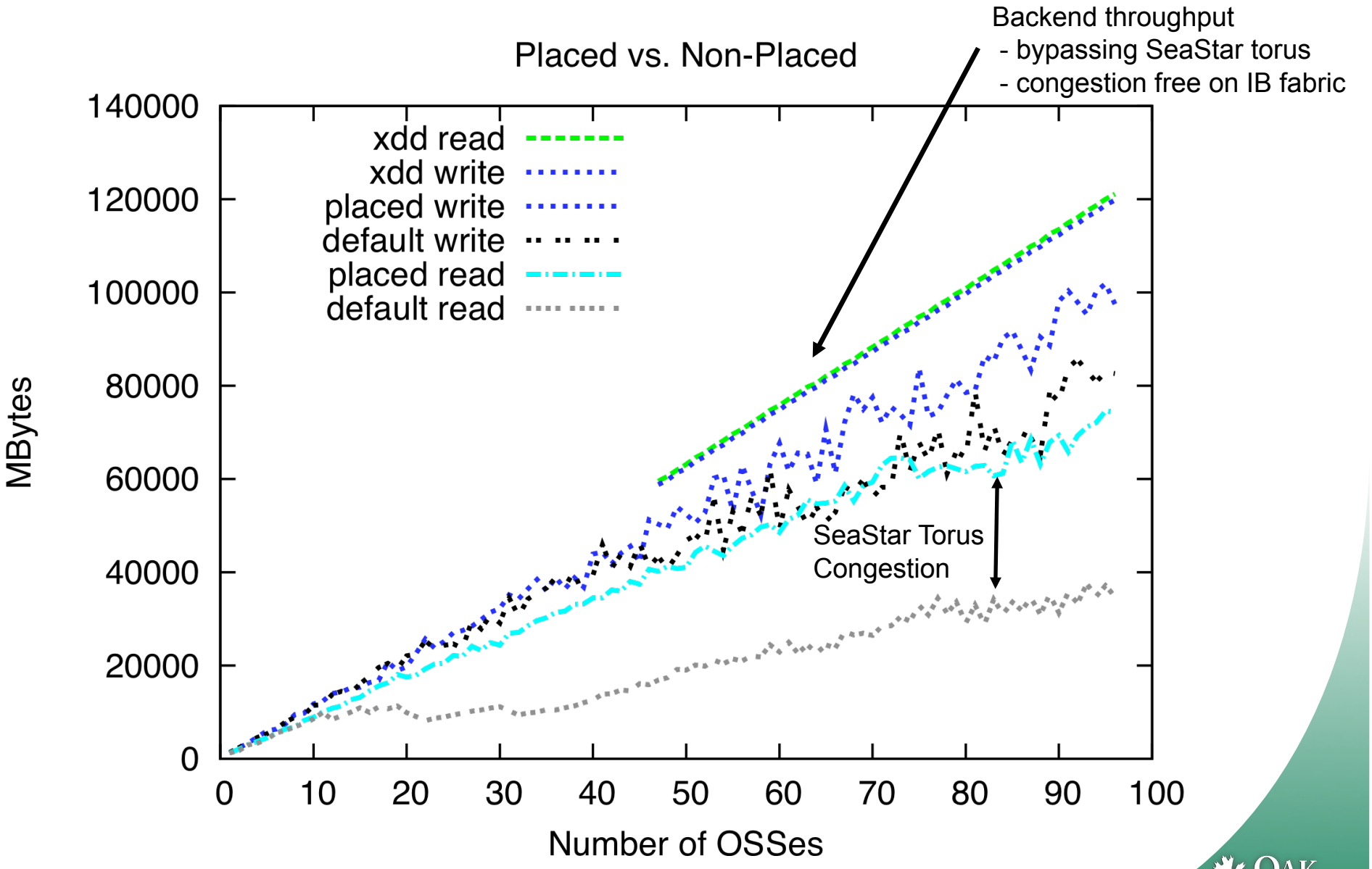
Assign clients to specific Router Groups - minimizes SeaStar Congestion



Performance Results

- Even in a direct attached configuration (no Lustre routers) we have demonstrated the impact of network congestion on I/O performance
  - By strategically placing writers within the torus and pre-allocating file system objects on topologically closest OSTs we can substantially improve performance
  - Performance results obtained on Jaguar XT5 using ½ of the available backend storage

Performance Results (1/2 of Storage)



Lessons Learned: Journaling Overhead

- Even “sequential” writes can exhibit “random” I/O behavior due to journaling
- Special file (contiguous block space) reserved for journaling on ldiskfs
  - Located all together
  - Labeled as “journal device”
  - Towards the beginning on the physical disk layout
- After the file data portion is committed on disk
  - Journal meta data portion needs to be committed as well
- Extra head seek needed for every journal transaction commit

Minimizing extra disk head seeks

- External journal on solid state devices
  - No disk seeks
  - Trade off between extra network transaction latency and disk seek latency
- Tested on a RamSan-400 device
  - 4 IB SDR 4x host ports
  - 7 external journal devices per host port
  - More than doubled the per DDN performance w.r.t. to internal journal devices on DDN devices
    - internal journal 1398.99
    - external journal on RAMSAN 3292.60

Minimizing synchronous journal transaction commit penalty

- Two active transactions per ldiskfs (per OST)
  - One running and one closed
  - Running transaction can't be closed until closed transaction fully committed to disk
- Up to 8 RPCs (write ops) might be in flight per client
  - With synchronous journal committing
    - Some can be concurrently blocked until the closed transaction fully committed
  - Lower the client number, higher the possibility of lower utilization due to blocked RPCs
    - More writes are able to better utilize the pipeline

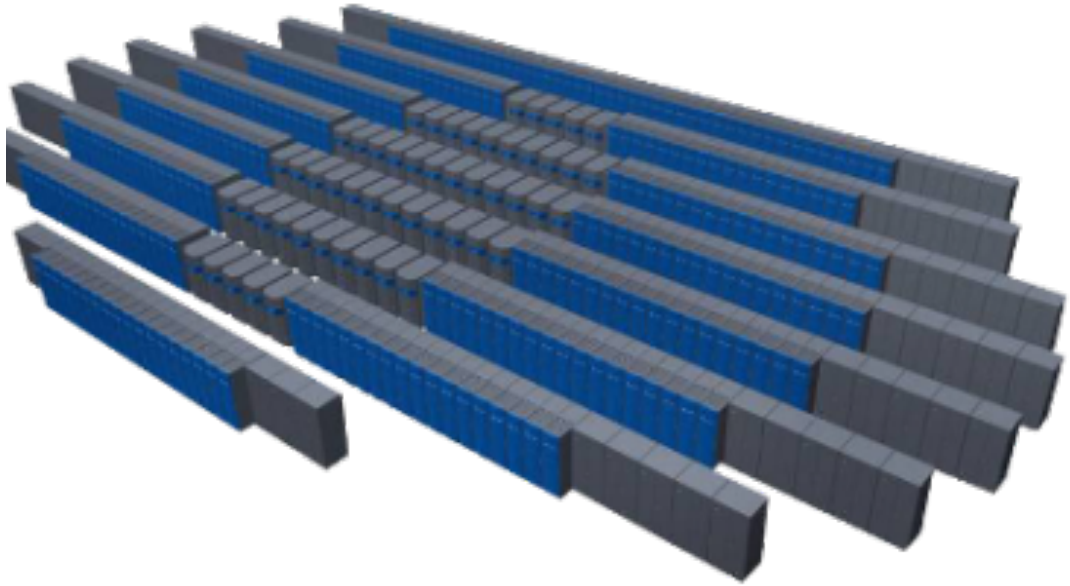
Minimizing synchronous journal transaction commit penalty

- To alleviate the problem
  - Reply to client when data portion of RPC is committed to disk
- Existing mechanism for client completion replies without waiting for data to be safe on disk
  - Only for meta data operations
  - Every RPC reply from a server has a special field in it that indicates “id last transaction on stable storage”
    - Client can keep track of completed, but not committed operations with this info
    - In case of server crash these operations could be resent (replayed) to the server once it is back up
- Extended the same concept for write I/O RPCs
- Implementation more than doubled the per DDN performance w.r.t. to internal journal devices on DDN devices
  - internal, sync journals 1398.99 MB/s
  - external, sync to RAMSAN 3292.60 MB/s
  - internal, async journals 4625.44 MB/s

Overcoming Journaling Overheads

- Identified two Lustre journaling bottlenecks
  - Extra head seek on magnetic disk
  - Blocked write I/O on synchronous journal commits
- Developed and implemented
  - A hardware solution based on solid state devices for extra head seek problem
  - A software solution based on asynchronous journal commits for the synchronous journal commits problem
- Both solutions more than doubled the performance
  - Async journal commits achieved better aggregate performance (with no additional hardware)

Scaling Lustre for the "Next Big Thing"

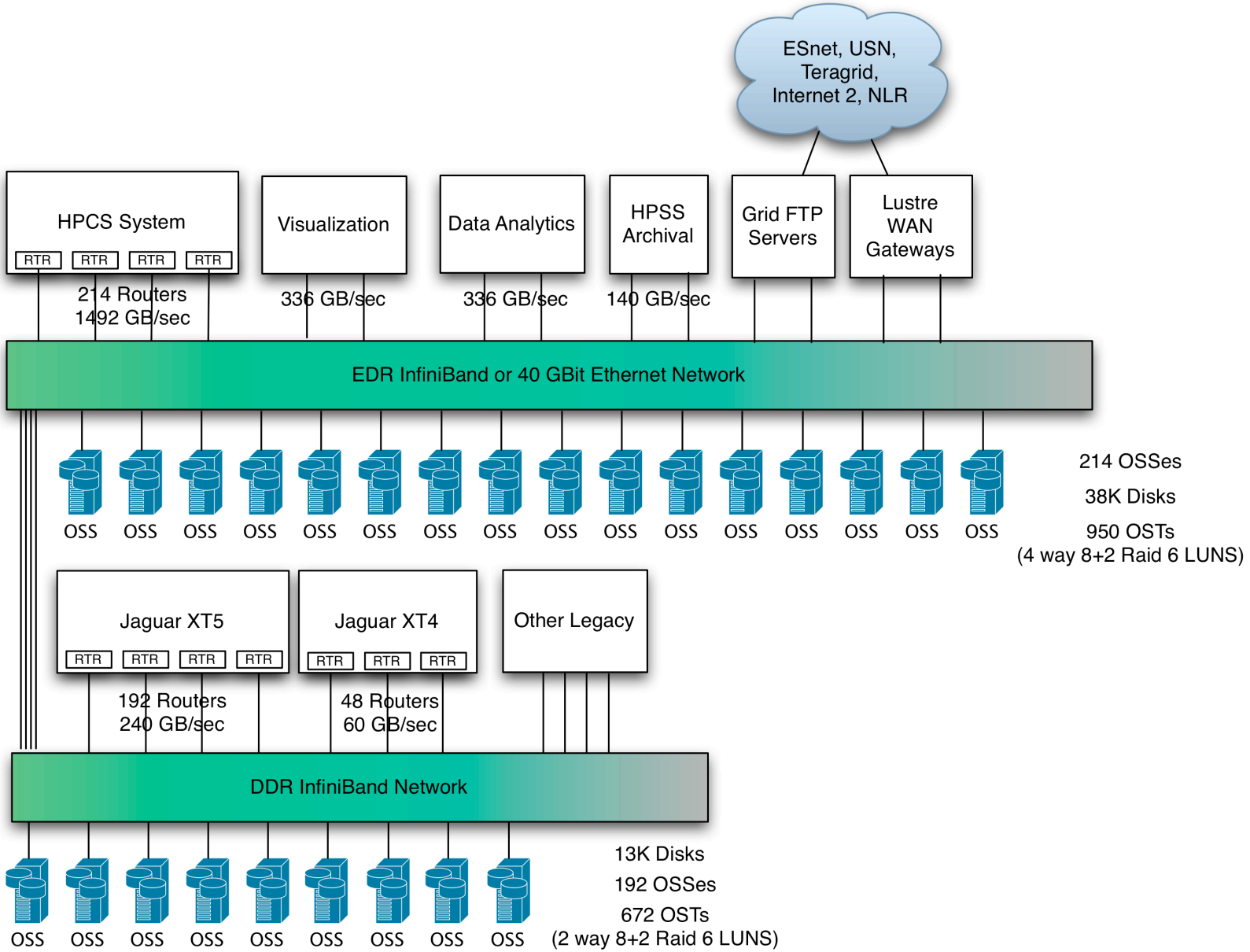


- **20 PF Leadership Class Machine in 2011/2012**

2012 File System Projections

	Maintaining Current Balance (based on full system checkpoint in ~20 minutes)		Desired (based on full system checkpoint in 6 minutes)	
	Jaguar XT5	HPCS -2011	Jaguar XT5	HPCS -2011
Total Compute Node Memory (TB)	298	1,852	298	1,852
Total Disk Bandwidth (GB/s)	240	1,492	828	5,144
Per Disk Bandwidth (MB/sec)	25	50	25	50
Disk Capacity (TB)	1	8	1	8
Time to checkpoint 100% of Memory	1242	1242	360	360
Over Subscription of Disks (Raid 6)	1.25	1.25	1.25	1.25
Total # disks	12,288	38,184	42,383	131,698
Total Capacity (TB)	9,830	244,378	33,906	842,867
OSS Throughput (GB/sec)	1.25	7.00	1.25	8.00
OSS Nodes needed for bandwidth	192	214	663	644
OST disks per OSS for bandwidth	64	179	64	205
Total Clients	18,640	30,000	18,640	30,000
Clients per OSS	97	140	28	47

2012 Architecture



2012 file system requirements

- 1.5 TB/sec aggregate bandwidth
- 244 Petabytes of capacity (SATA - 8 TB)
  - 61 Petabytes of capacity (SAS – 2TB)
  - Final configuration may include pools of SATA, SAS and SSDs
- ~100K clients (from 2 major systems)
  - HPCS System
  - Jaguar
- ~200 OSTs per OSS
- ~400 clients per OSS

2012 file system requirements

- Full integration with HPSS
  - Replication, Migration, Disaster Recovery
  - Useful for large capacity project spaces
- OST Pools
  - Replication and Migration among pools
- Lustre WAN
  - Remote accessibility
- pNFS support
- QOS
  - Multiple platforms competing for bandwidth

2012 File System Requirements

- Improved data integrity
  - T10-DIF
  - ZFS (Dealing with licensing issues)
- Large LUN support
  - 256 TB
- Dramatically improved metadata performance
  - Improved single node SMP performance
  - Will clustered metadata arrive in time?
  - Ability to take advantage of SSD based MDTs

2012 File System Requirements

- Improved small block and random I/O performance
- Improved SMP performance for OSSes
  - Ability to support larger number of OSTs and clients per OSS
- Dramatically improved file system responsiveness
  - 30 seconds for “ls -l” ?
  - Performance will certainly degrade as we continue adding additional computational resources to Spider

Good overlap with HPCS I/O Scenarios

- 1. Single stream with large data blocks operating in half duplex mode
- 2. Single stream with large data blocks operating in full duplex mode
- 3. Multiple streams with large data blocks operating in full duplex mode
- 4. Extreme file creation rates
- 5. Checkpoint/restart with large I/O requests
- 6. Checkpoint/restart with small I/O requests
- 7. Checkpoint/restart large file count per directory - large I/Os
- 8. Checkpoint/restart large file count per directory - small I/Os
- 9. Walking through directory trees
- 10. Parallel walking through directory trees
- 11. Random stat() system call to files in the file system – one (1) process
- 12. Random stat() system call to files in the file system - multiple processes

Questions?

- **Contact info:**

Galen Shipman

Group Leader, Technology Integration

865-576-2672

gshipman@ornl.gov