The Spider Center Wide File System

Presented by:
Galen M. Shipman

Collaborators:
David A. Dillow
Sarp Oral
Feiyi Wang

May 4, 2009

U.S. DEPARTMENT OF
ENERGY

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Jaguar: World’s most powerful computer

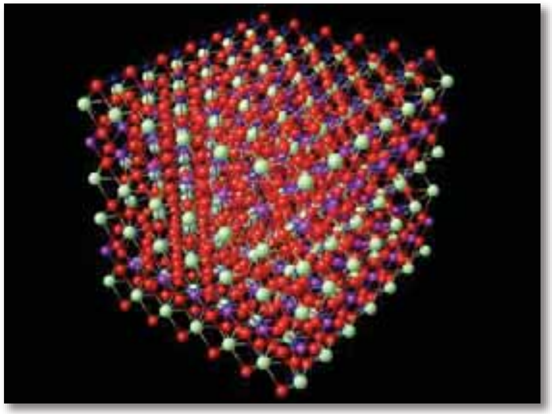
Designed for science from the ground up

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

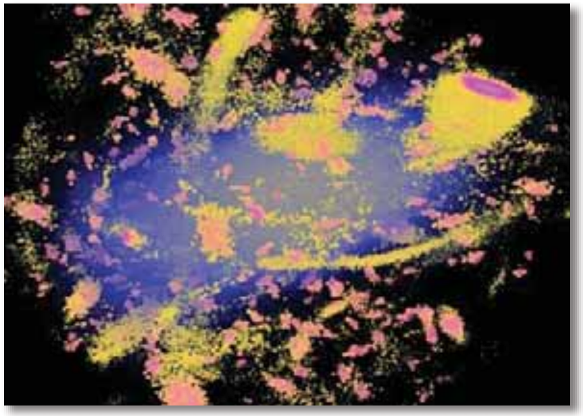
Jaguar
Talk
Tuesday at
10:30

Enabling breakthrough science

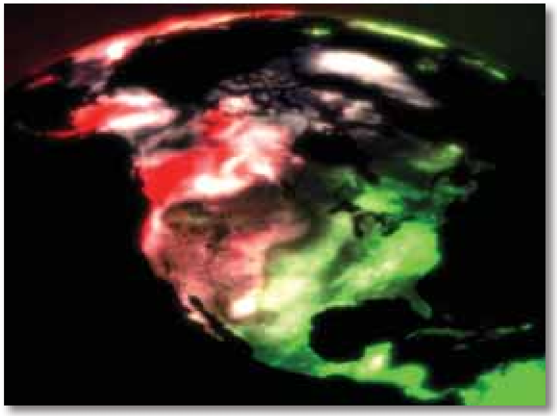
5 of top 10 ASCR science accomplishments in the past 18 months used LCF resources and staff



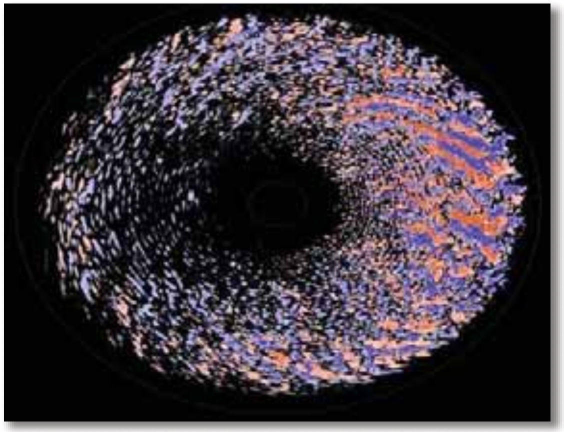
Electron pairing in HTSC cuprates
PRL (2007, 2008)



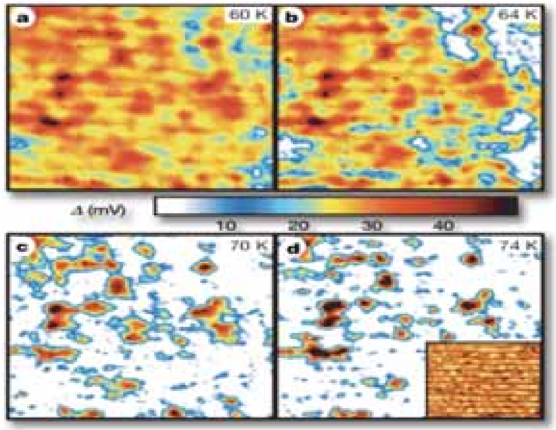
Shining a light on dark matter
Nature 454, 735 (2008)



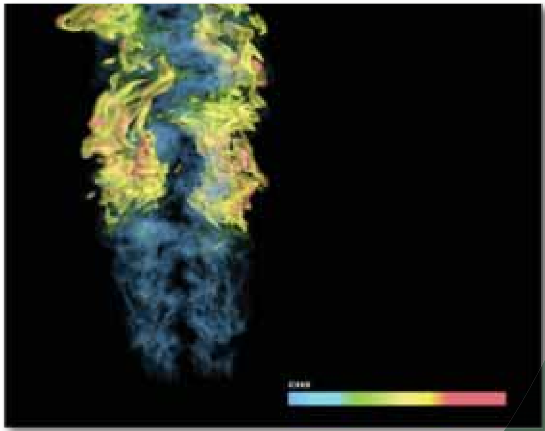
Modeling the full earth system



Fusion: Taming turbulent heat loss
PRL 99, *Phys. Plasmas* 14



Nanoscale nonhomogeneities in high-temperature superconductors
Winner of Gordon Bell prize

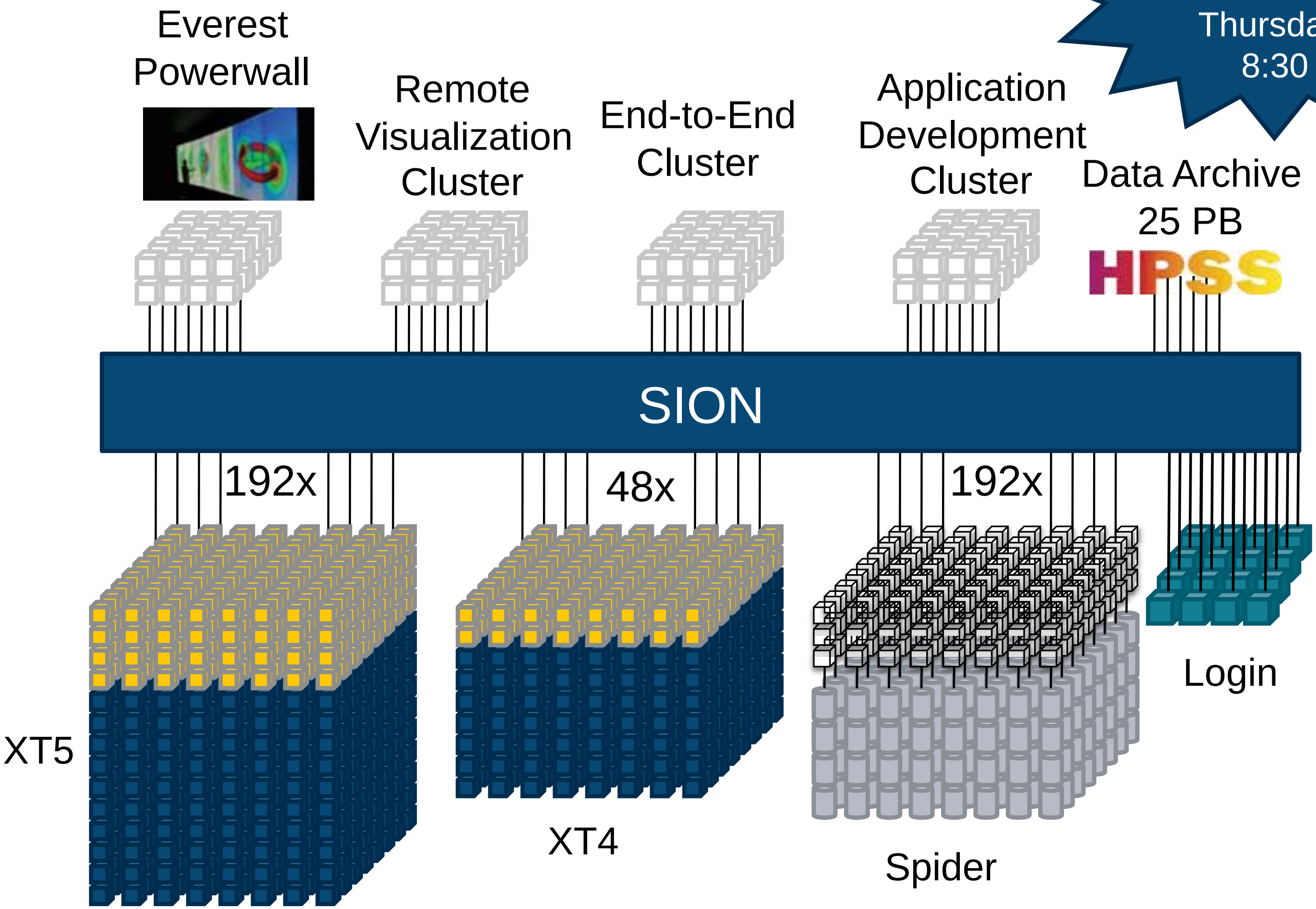


Stabilizing a lifted flame
Combust. Flame (2008)

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

LCF Infrastructure



Talk on integrating XT4 and XT5 Thursday 8:30

Current LCF File Systems

System	Path	Size	Throughput	OSTs
Jaguar XT5				
	/lustre/scratch	4198 TB	> 100 GB/s	672
	/lustre/widow1	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

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	> 50 GB/s	672
	/lustre/widow1	4198 TB	> 50 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	> 6 GB/s	672
	/lustre/widow1	4198 TB	> 32 GB/s	672

Benefits of Spider

- **Accessible from all major LCF resources**
  - **Eliminates file system "islands"**
- **Accessible during maintenance windows**
  - **Spider will remain accessible during XT4 and XT5 maintenance**

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**

How Did We Get Here?

- **We didn’t just pick up the phone and order a center-wide file system**
  - **No single Vendor could deliver this system**
  - **Trail Blazing was required**
- **Collaborative effort was key to success**
  - **ORNL**
  - **Cray**
  - **DDN**
  - **SUN**

A Phased Approach

- **Conceptual design - 2006**
- **Early Prototypes - 2007**
- **Small Scale Production System (wolf) - 2008**
- **Storage System Evaluation - 2008**
- **Direct Attached Deployment - 2008**
- **Spider File System Deployment - 2009**

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**

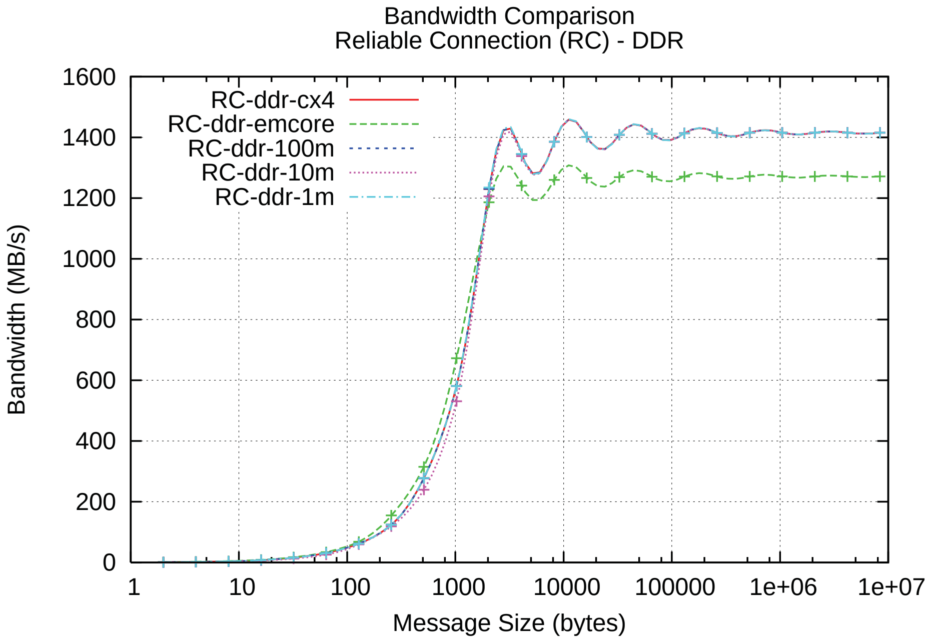
- **Early access on Jaguar XT5 today!**
  - **General Availability this Summer**

Snapshot of Technical Challenges

- **Fault tolerance**
  - **Network**
  - **I/O Servers**
  - **Storage Arrays**
  - **Lustre File system**
- **Performance**
  - **SATA**
  - **Network congestion**
  - **Single Lustre Metadata server**
- **Scalability**
  - **26,000 file system clients and counting**

InfiniBand Support on Cray XT SIO

- **LCF effort; required system software work to support OFED on the XT SIO**
- **Evaluation of a number of optical cable options**
- **Worked with Cray to integrate OFED into stock CLE distribution**



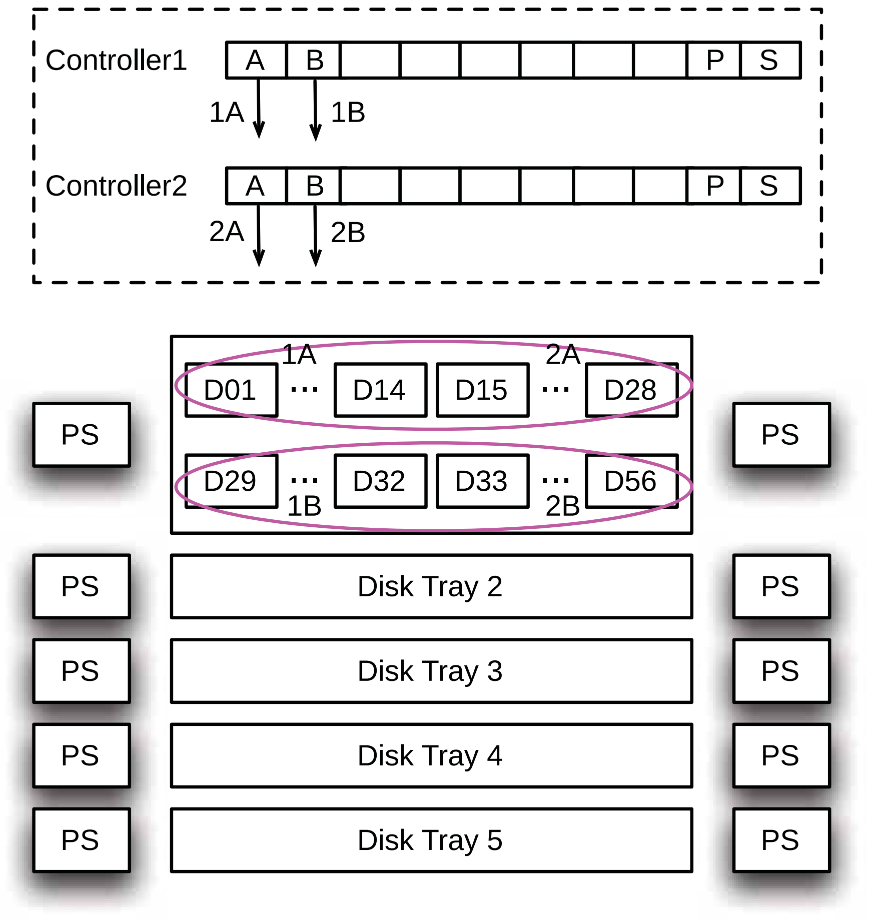
*InfiniBand Based Cable Comparison, Makia Minich, 2007

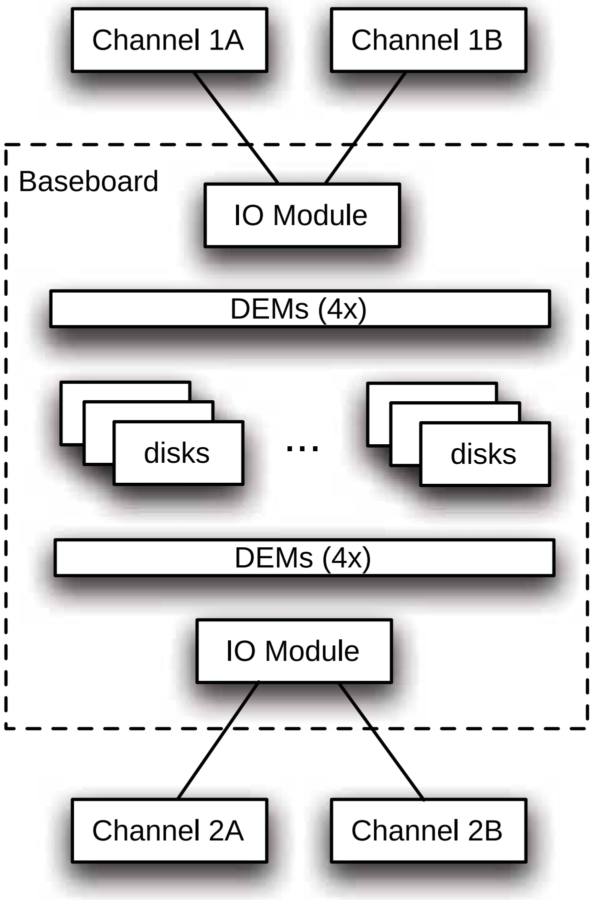
Reliability Analysis of DDN S2A9900

- **Developed a failure model and a quantitative expectation of the system's reliability**
- **Particular attention was given to the DDN S2A9900's peripheral components**
  - **3 major components considered**
    - **I/O module**
    - **Disk Expansion Modules (DEMs)**
    - **Baseboard**
- **Analysis of RAID 6 implementation**

Details to appear in: A Case Study on Reliability of Spider Storage System

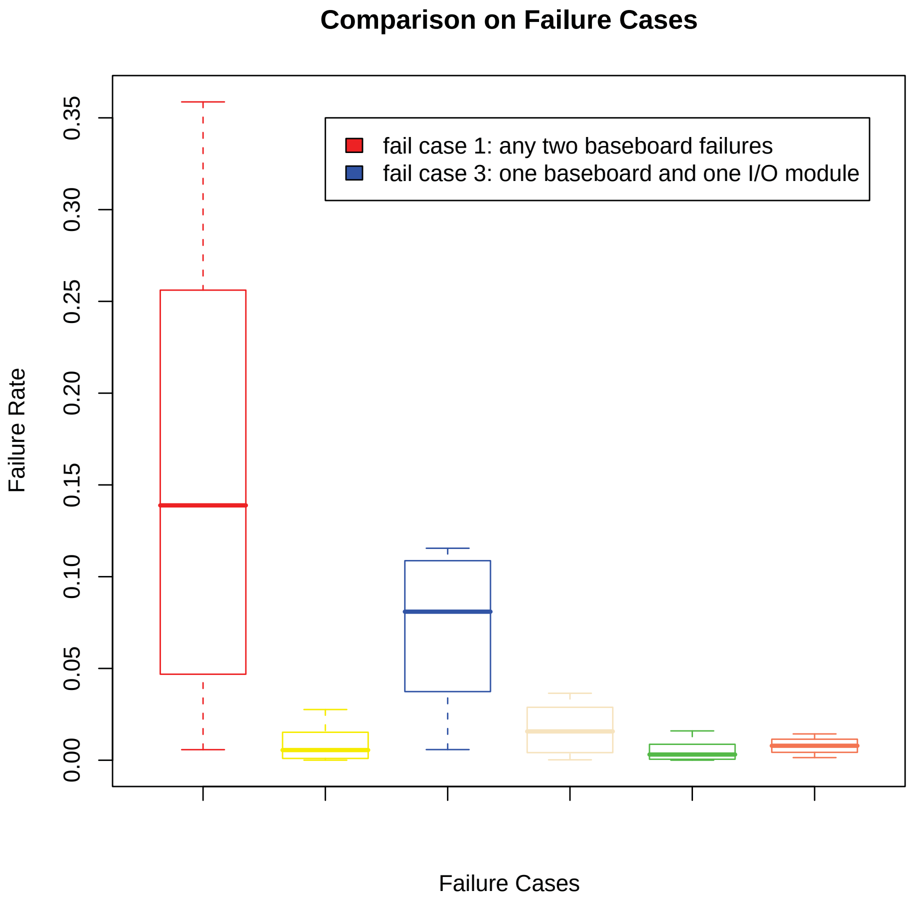
DDN S2A9900 Architecture





DDN S2A9900 Failure Cases

- **Case 1: two out of the five baseboards fail**
- **Case 2: three out of ten I/O modules fail**
- **Case 3: one baseboard fails, and another I/O module fails on a different baseboard**
- **Case 4: any two I/O modules fail and any other baseboard failure**



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

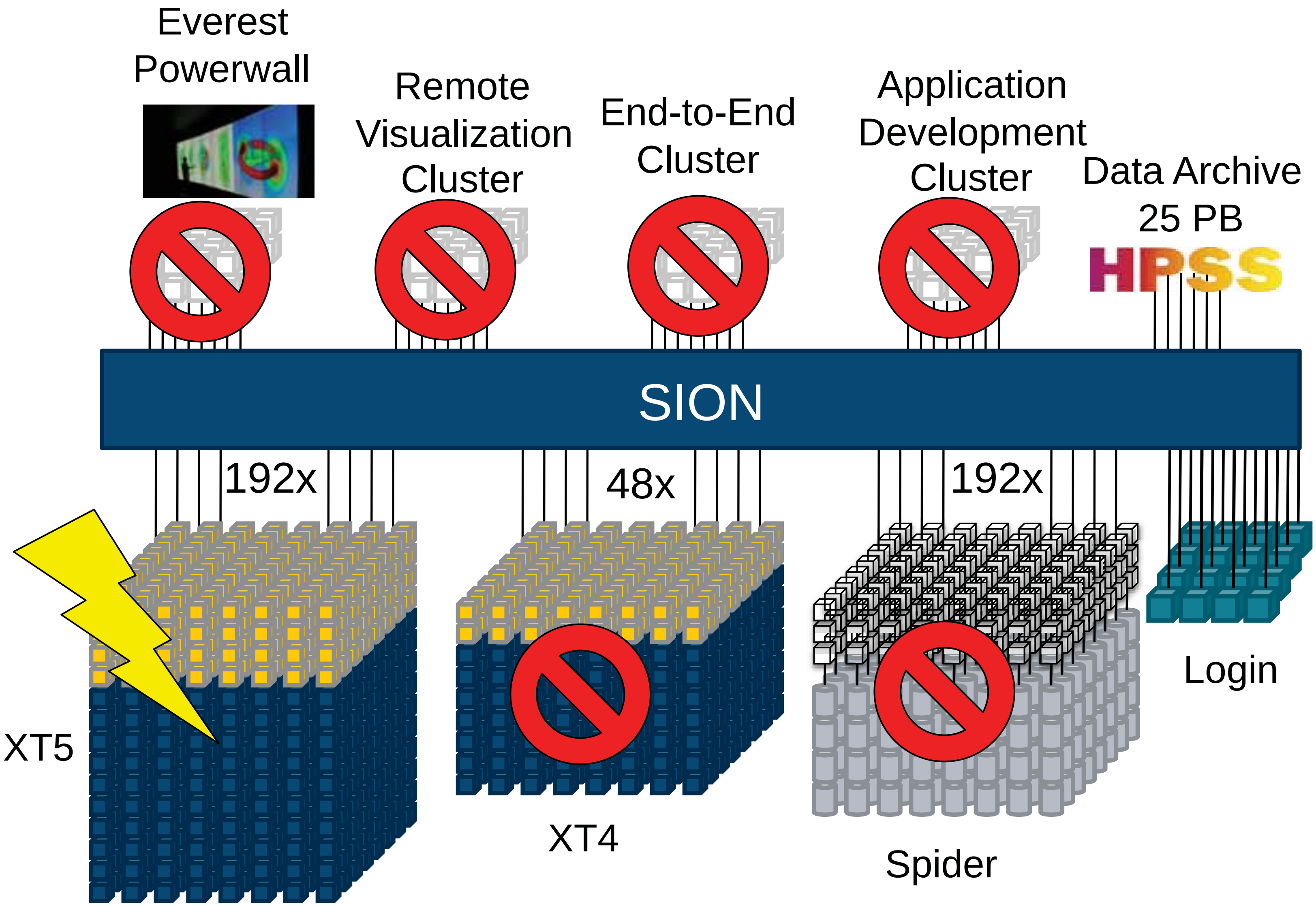
Scaling to More Than 26,000 Clients

- **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**

Surviving a Bounce

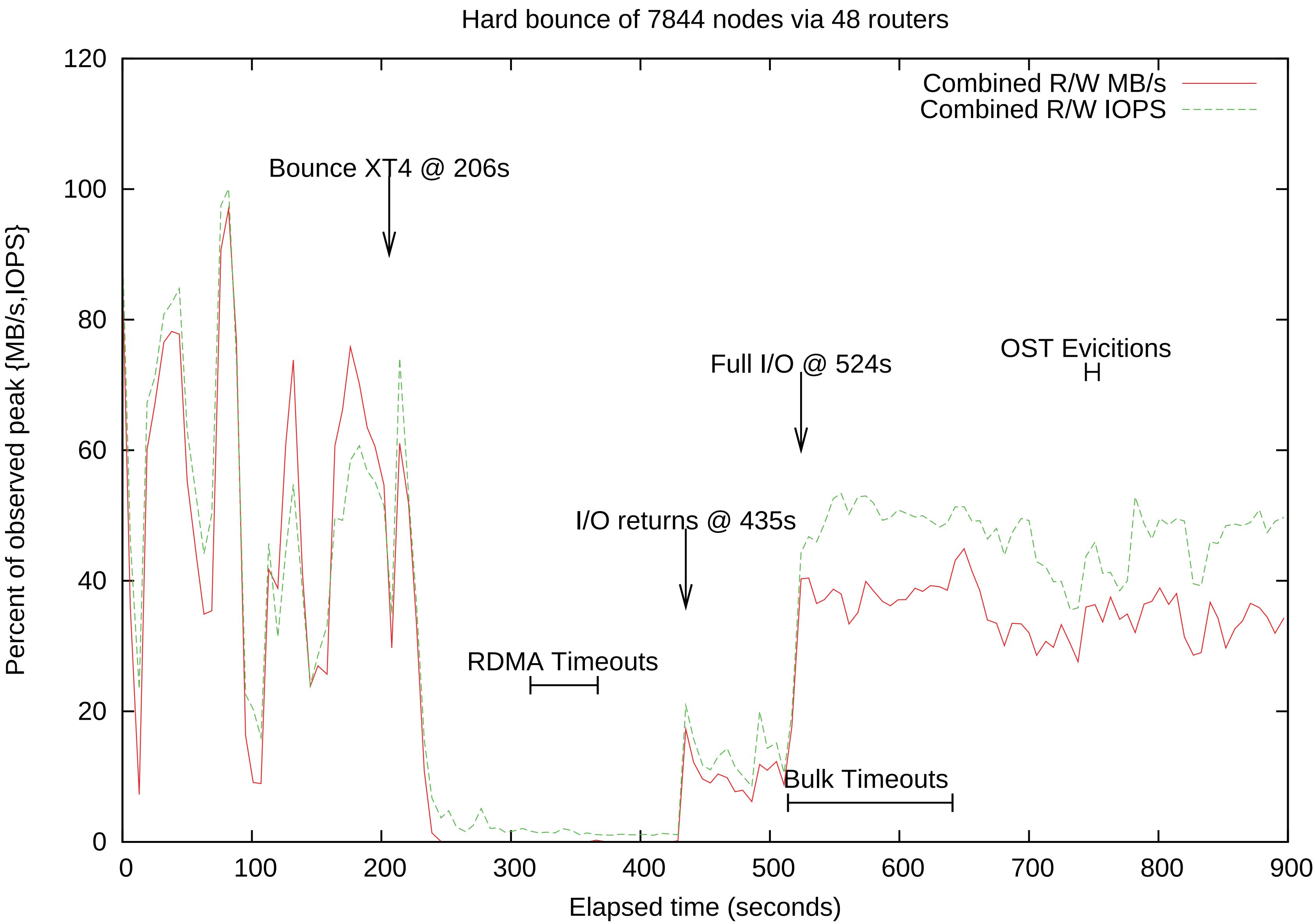


Challenges in Surviving an Unscheduled Jaguar XT4 or XT5 Outage

- **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
Percent of observed peak {MB/s,IOPS}
0
20
40
60
80
100
120
0
100
200
300
400
500
600
700
800
900
Elapsed time (seconds)
Bounce XT4 @ 206s
Full I/O @ 524s
OST Evicitions
I/O returns @ 435s
RDMA Timeouts
Bulk Timeouts

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**

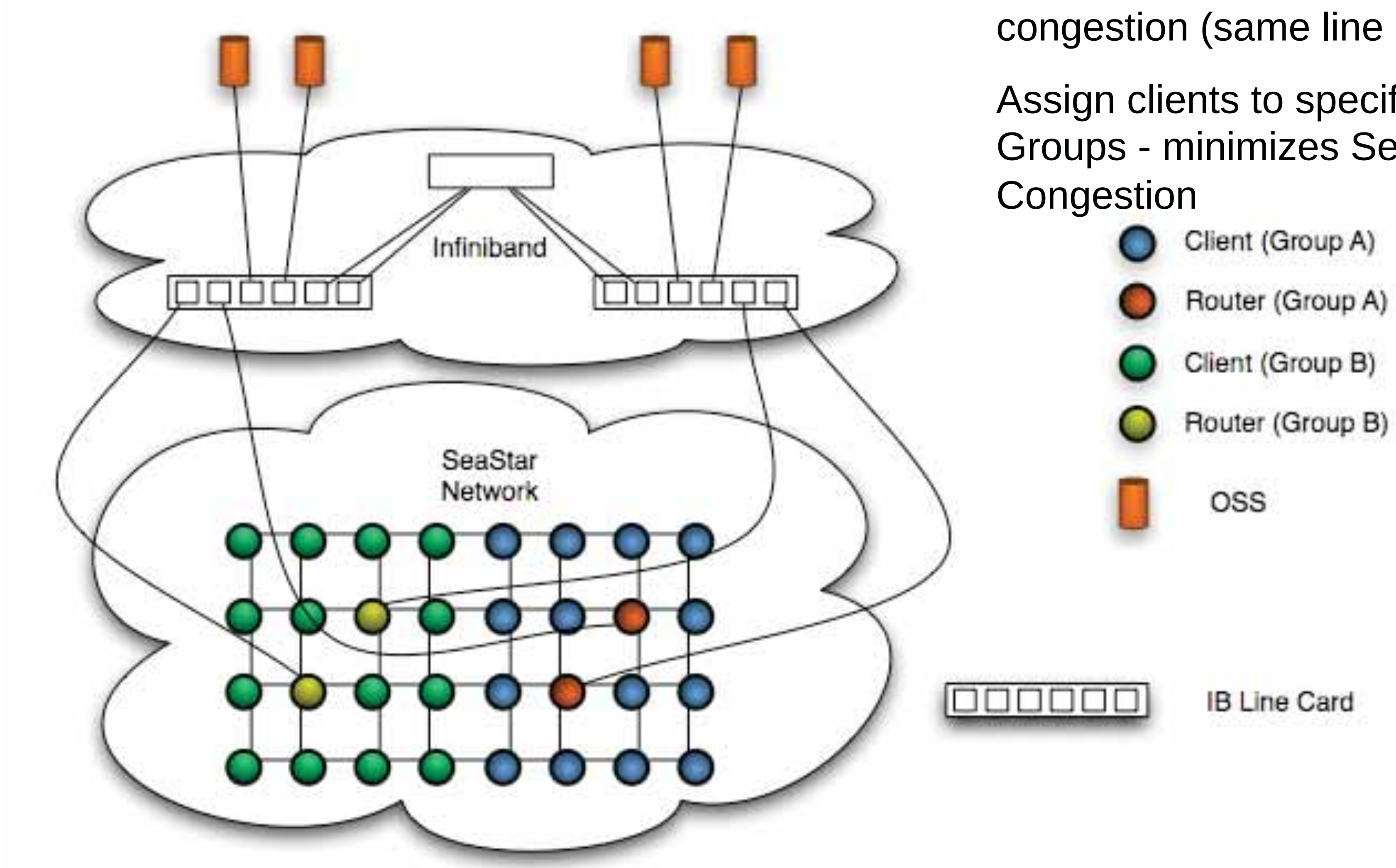
Design To Minimize Contention

- **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**

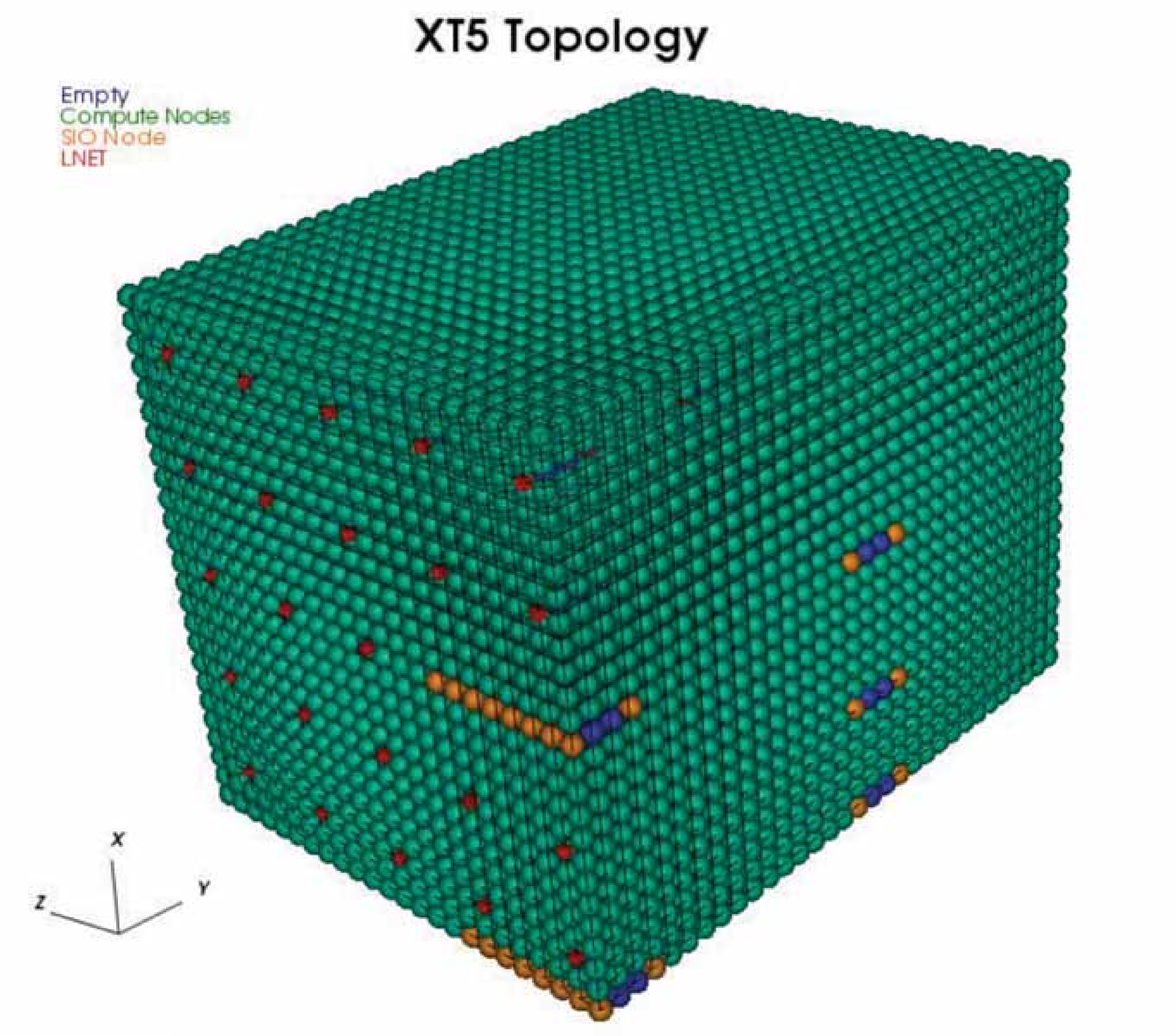
Intelligent LNET Routing

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

Assign clients to specific Router Groups - minimizes SeaStar Congestion



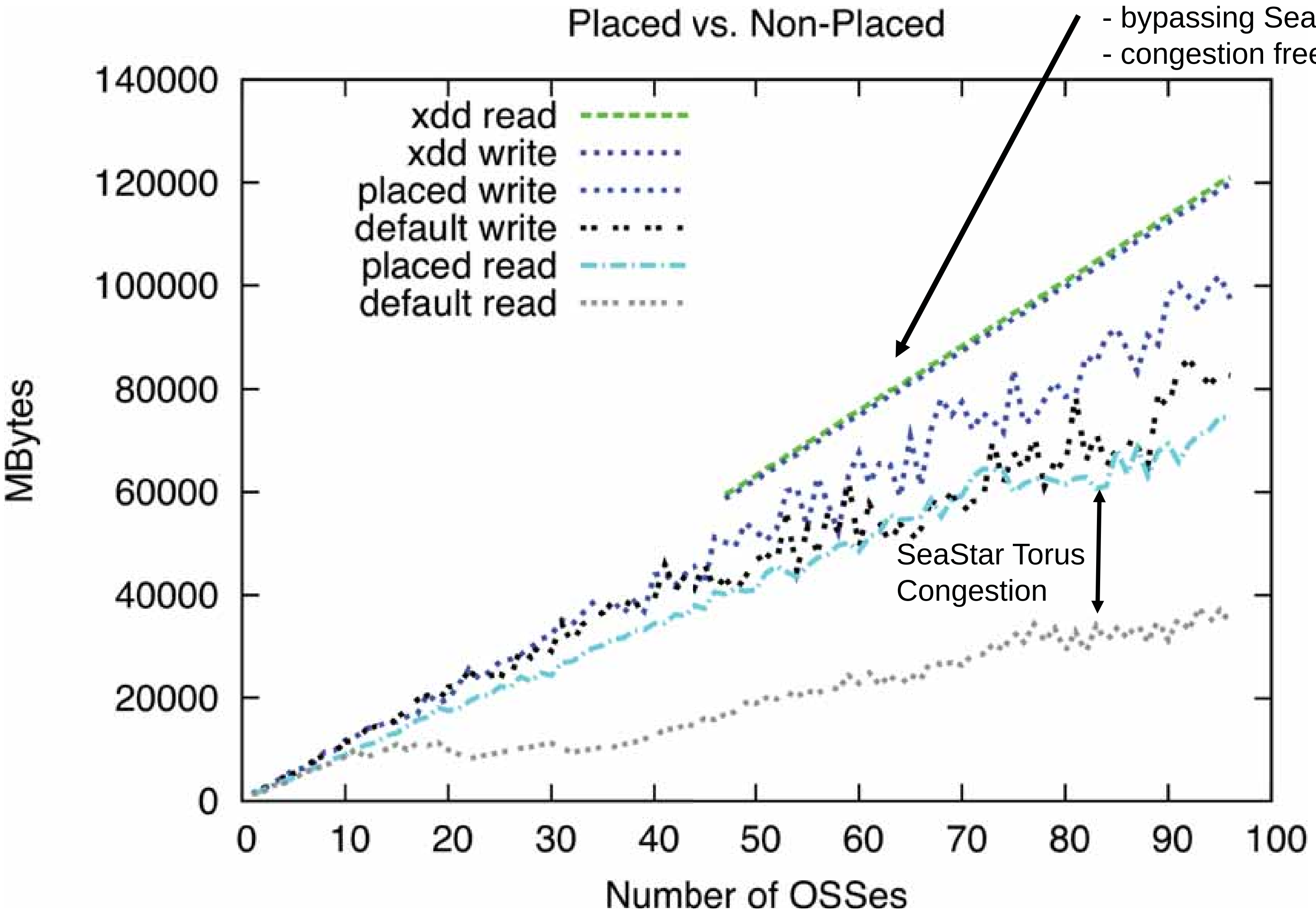
XT5 Router node placement



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)



Backend throughput
- bypassing SeaStar torus
- congestion free on IB fabric

SeaStar Torus Congestion

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
- Asynchronous Journal Commit
  - Lustre – software only change
  - Reply to client when data portion of RPC is committed to disk

Configuration	Bandwidth MB/s
Internal Journals	1398.99
external, sync to RAMSAN	3292.60
internal, async journals	4625.44

Future Work

- Increased Metadata performance
  - Improved SMP scalability (10x improvement target from single MDS)
  - Tiger team working this now (ORNL, Cray, SUN)
- Resiliency
  - OSS Failover
  - Router Failover (asymmetric network failure)
- Quality of Service
  - Network Request Scheduler
- Increased Bandwidth
  - 240 GB/sec is not enough
  - Full system checkpoint times need to be reduced
- Changing workloads
  - Data Analytics
  - Visualization
  - No longer a write-once file system for checkpoints

INCITE April 15th call for proposals

Call for large-scale, computationally intensive, high-impact research proposals

In 2010, powerful, leadership-class computing systems at DOE's Argonne National Laboratory and Oak Ridge National Laboratory will provide over one billion processor hours to a limited number of researchers nationwide.

The call is open to scientific researchers and research organizations, including industry; DOE Sponsorship is not required. **Deadline July 1st**.

INCITE awards help advance the state-of-the-art in areas such as

- Accelerator physics
- Astrophysics
- Chemical sciences
- Climate research
- Computer science
- Engineering
- Physics
- Environmental science
- Fusion energy
- Life sciences
- Materials science
- Nuclear physics, and more

For details about the DOE leadership computing facilities, see www.alcf.anl.gov and www.nccs.gov or contact INCITE@DOEleadershipcomputing.org to be added to an announcement distribution list.

Questions?

- **Contact info:**

Galen Shipman

Group Leader, Technology Integration

865-576-2672

gshipman@ornl.gov