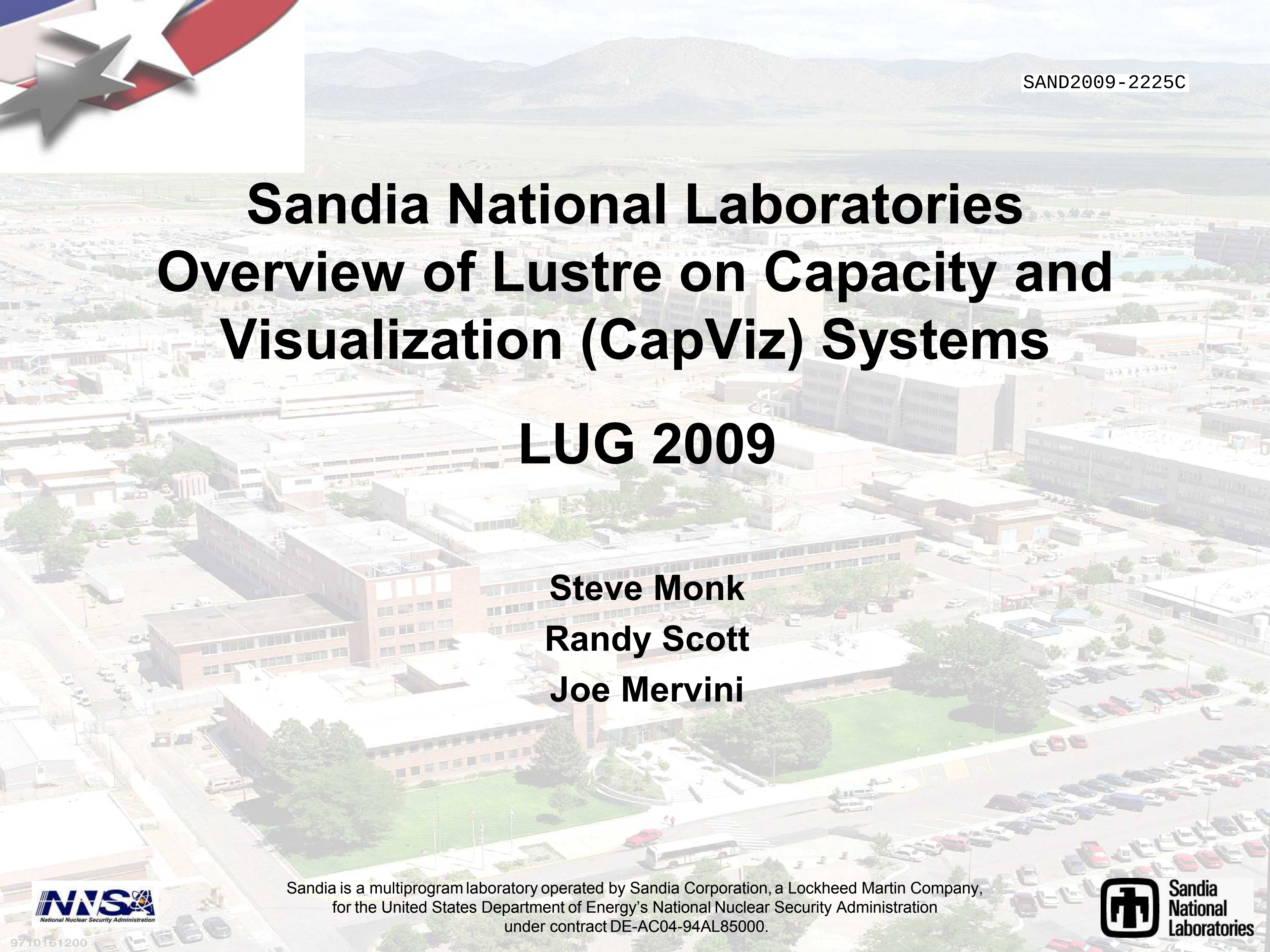


SAND2009-2225C

Sandia National Laboratories Overview of Lustre on Capacity and Visualization (CapViz) Systems

LUG 2009

Steve Monk

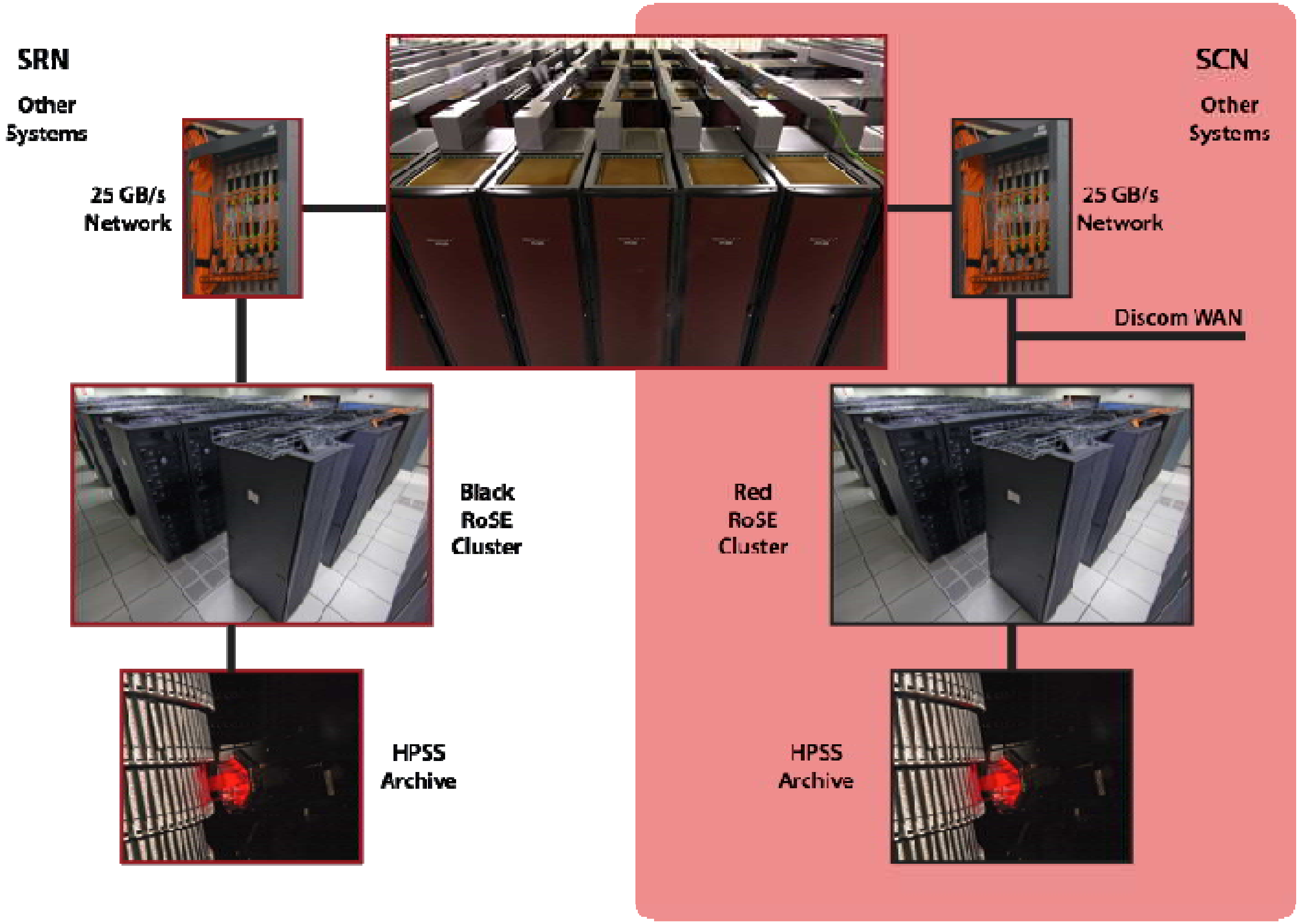
Randy Scott

Joe Mervini

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

SNL’s Capacity Lustre has its Roots in SNL’s Capability machine:

Architected Red Storm Environment



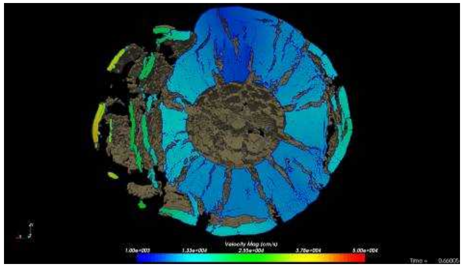
Circa 2004
Red Storm needed a post processing environment
RoSE= Red Storm Environment

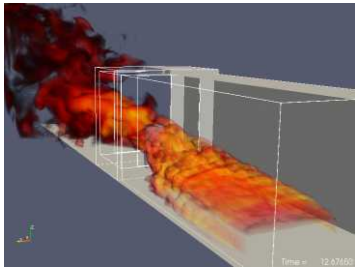
Two requirements to support Red Storm:

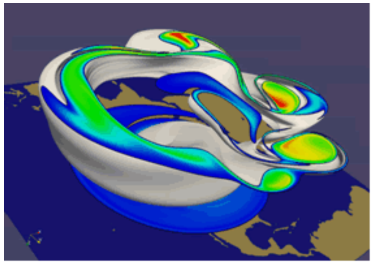
- **Vis Power**
  - for highly interactive visualization and analysis of large (300 mega-cell and larger) datasets

- **I/O Power**
  - for accessing terascale data within RoSE cluster at interactive rates
    - 25 GB/s parallel file system
    - 1 second to access one time-step of a 300-mega-cell calculation
  - for moving terascale data from Red Storm
    - 25 GB/s (90 TB/hour) parallel file transfer, to minimize impact on Red Storm file system

- **To get to this performance we ended up with bunch of storage capacity**
  - The idea to create a multi-cluster Lustre was born
    - enable other compute clusters to write output directly to the RoSE file systems and eliminate disk to disk transfers
    - Buy more compute power for capacity clusters and use our disk systems more efficiently







Current Production Lustre Configuration

- **Lustre version:**
  - Running 1.6.6 on all production Lustre servers
  - Migrated from 1.4.12 about 4 months ago
    - We were cautious in moving to 1.6.X as we were pleased with 1.4 stability and didn't want to lead the technology curve on this one
    - 1.6.6 has been very stable for us so far
  - Several Clusters running 1.4.X clients to keep them afloat until they get decommissioned
  - We've had Lustre in some sort of production for 4+ years now

- **Server hardware:**
  - OSS's/MDS's: Dell 1950's
    - 8 GB RAM, Fiber Channel 4, 4X DDR Infiniband
  - LNET routers: Dell 1950's
    - 8 GB RAM,10GigE, 4X DDR Infiniband and 10GigE NICS (Chelsio T310's)

- **Storage hardware:**
  - DDN (DataDirect Networks)
    - 31 - 9550 Controller couplets (FC4/SATA disks) for OSS's, 8+2 RAID configuration
    - 4 - 8500 Controller couplets (FC2/FC disks) for MDS's
    - 7,440 SATA disks in production!
      - Mix of 250 and 500 GB disks
  - LSI IS4600's (part of Dark Storm IO cluster)
    - 6 controller pairs in RAID 5 configuration
    - 744 SATA disks

Current Production Configuration Cont.

- File Systems:
  - 2 main production file systems (Red and Black)
    - 360 TB: 8 DDN couplets with 31 OSS's (186 OST's)
    - 1 PB: 11 DDN couplets with 44 OSS's (264 OST's)
    - ~600 TB in test bed will be deployed soon

- Clients:
  - Black: 5,142 client's (Tbird largest cluster @4300 nodes)
  - Red: 1,600 clients
  - Most clients connect to file system via LNET routers
    - Visualization and Red Storm data transfer nodes are on local file system fabric (Infiniband) to allow for better throughput

Lustre Support and Operations

- Two Lustre administrators and a team lead supporting three environments plus a test-bed.
  - Team also provides support for other file systems (NFS, Panasas etc.) and daily Cluster operations.
- Two people from the CapViz hardware group are responsible for the DDN/LSI maintenance
- "RAS"
  - Notification scripts send out email warnings to Lustre admins and SNL's 24/7 monitoring center
  - real time Syslog monitoring (cat | grep |awk)
  - LMT (newest addition to our tool set)
- Lustre support contract with a single point of contact (Cliff White) and weekly conference calls

Multi-Cluster Lustre

- LNET (Lustre NETwork) “routing” is key to sharing a single Lustre file system with several clusters

- Lustre routing provides SNL with:
  - Network segmentation and location
    - No need for multiple clusters to share the same high-speed interconnect
    - Cluster and storage don’t need to be in same facility
  - Storage resources on a dedicated network fabric
    - Single IB switch fabric has proven to be very stable
  - Tunable performance
    - just add more routers to get more bandwidth
    - Note: we are seeing routers running at near wire speed!

Generic Benefits of a Multi-Cluster file system

- Avoid Islands of storage

- Users see same file system everywhere
  - No need to move data between clusters

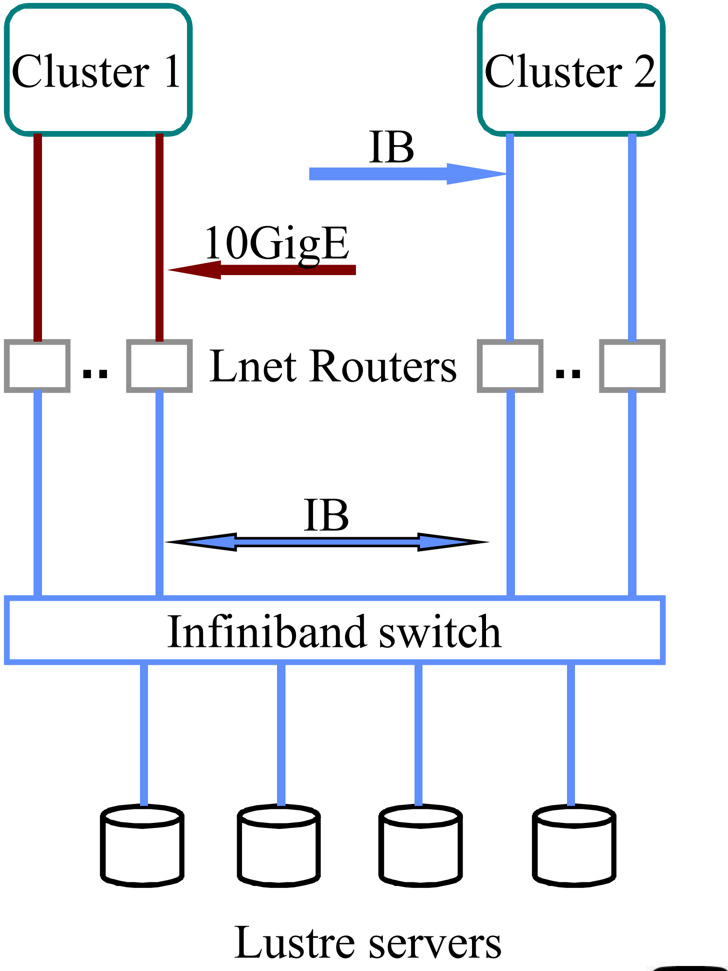
- Central management of storage by storage experts
  - Storage can get the attention it deserves

- Compute and Vis clusters can focus on what they do and be "customers" of the file system

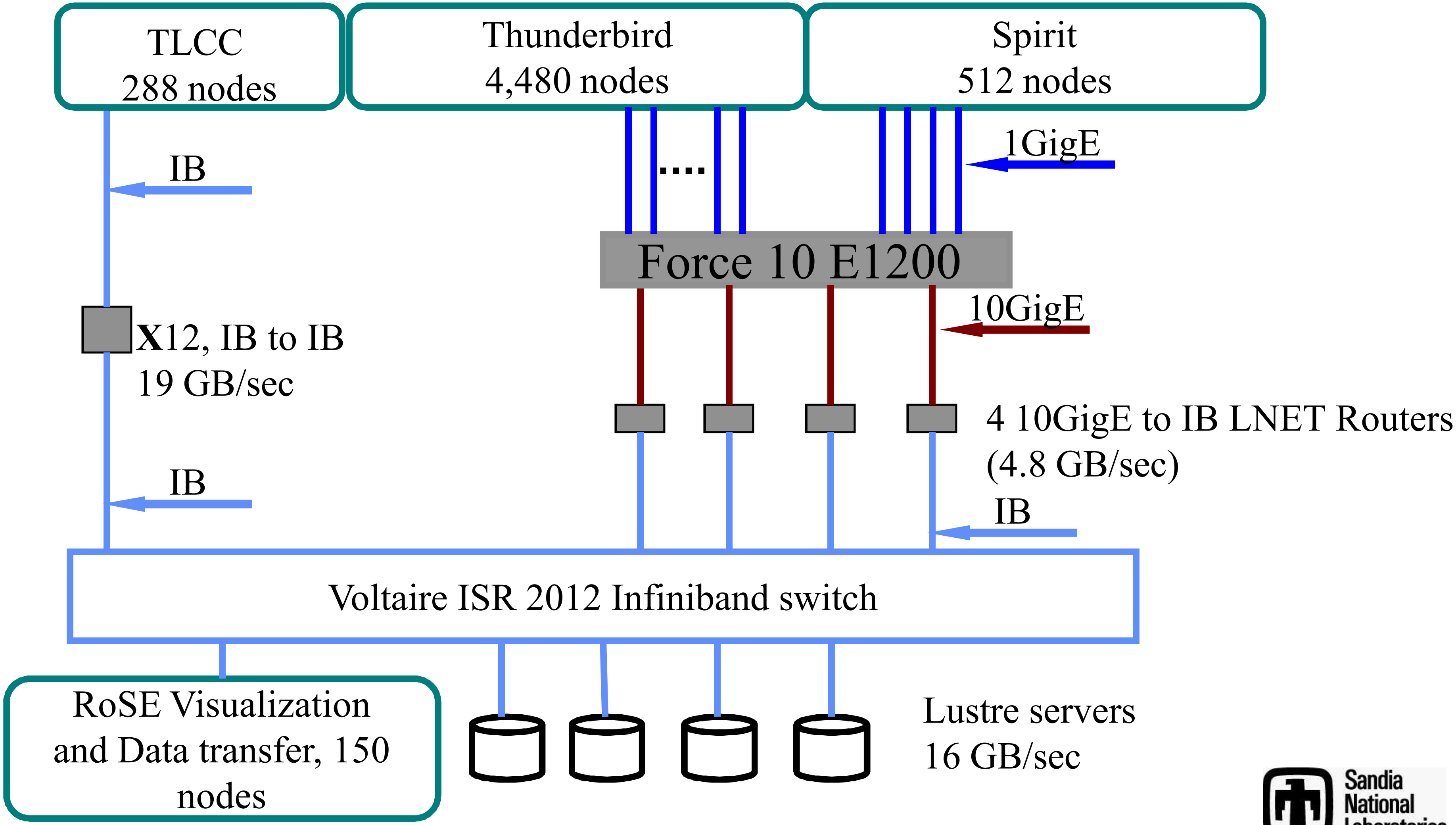
- Hardware utilization: quickly provide better utilization of existing storage resources
  - e.g. offer the old storage combined with older servers as a "slower" file system for long term storage etc.

Router configuration

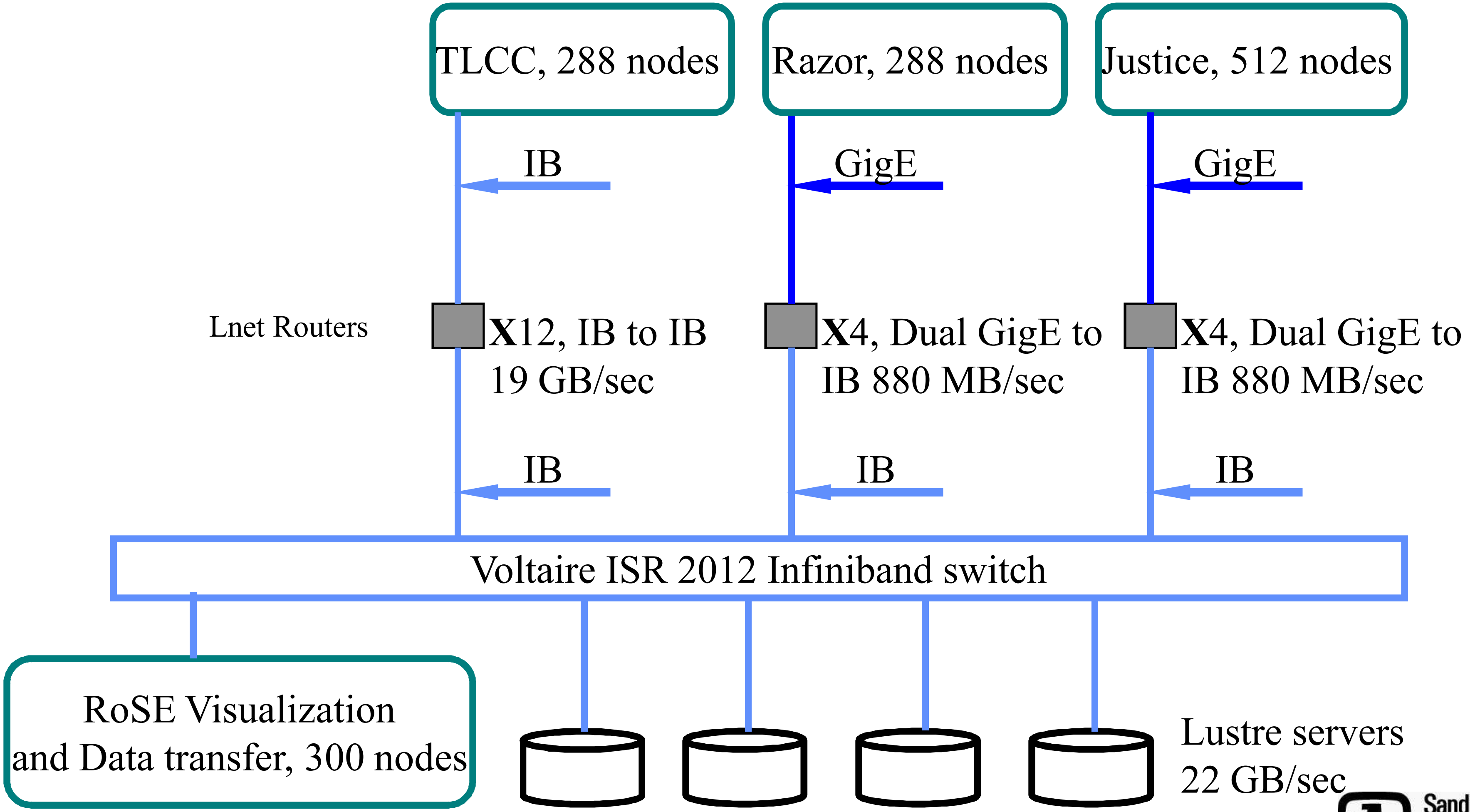
- Lustre Servers and Storage are on Infiniband fabric
- Routers route from networkX to Infiniband
- Currently use:
  - 10GigE to IB (DDR)
    - 1.2 GB/sec
  - Bonded GigE to IB
    - 220 MB/sec
  - IB to IB
    - 1.6 GB/sec (DDR)



SRN LNET configuration



SCN LNET configuration



TLCC, 288 nodes

Razor, 288 nodes

Justice, 512 nodes

IB

GigE

GigE

Lnet Routers

X12, IB to IB 19 GB/sec

X4, Dual GigE to IB 880 MB/sec

X4, Dual GigE to IB 880 MB/sec

IB

IB

IB

Voltaire ISR 2012 Infiniband switch

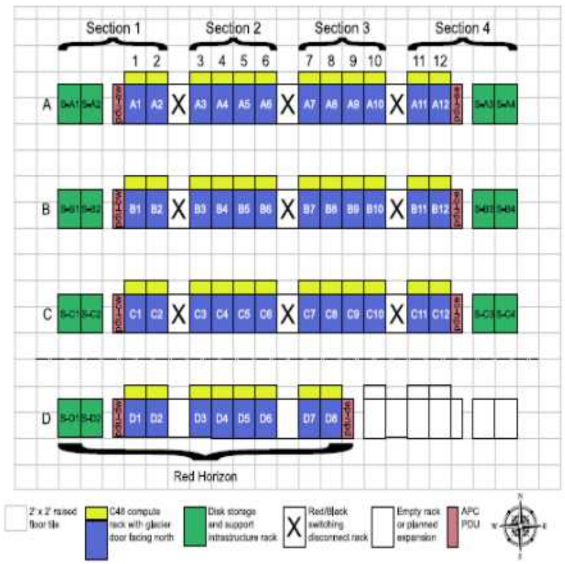
RoSE Visualization and Data transfer, 300 nodes

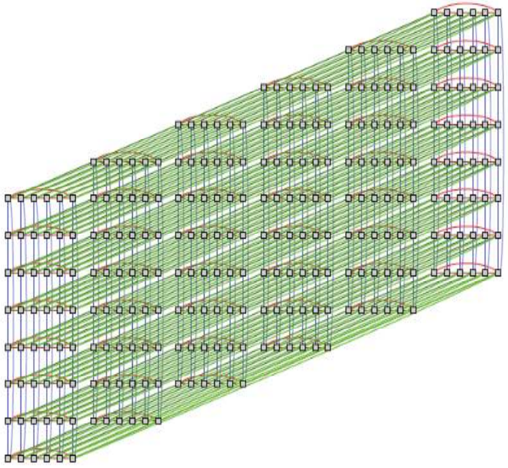
Lustre servers 22 GB/sec

Current Projects : Red Sky

"Mid-Range" Compute Cluster

- Main points:
  - Built on Sun C48 blades
  - 2 dual socket Nehalem nodes/blade
  - QDR IB with 6X6X8 3D Torus
  - Refrigerant cooling doors
  - Security Domain switchable
  - 172 peak TFLOP/s to start,
  - No Ethernet for compute nodes
  - Local Lustre file system with LNET router access to site file systems (multi-hop)
  - System is being built at SNL now!


Section 1
Section 2
Section 3
Section 4
1 2 3 4 5 6 7 8 9 10 11 12
A
S-A1 S-A2 A1 A2 X A3 A4 A5 A6 X A7 A8 A9 A10 X A11 A12 S-A3 S-A4
B
S-B1 S-B2 B1 B2 X B3 B4 B5 B6 X B7 B8 B9 B10 X B11 B12 S-B3 S-B4
C
S-C1 S-C2 C1 C2 X C3 C4 C5 C6 X C7 C8 C9 C10 X C11 C12 S-C3 S-C4
D
S-D1 S-D2 D1 D2 D3 D4 D5 D6 D7 D8
Red Horizon
2' x 2' raised floor tile
C48 compute rack with glacier door facing north
Disk storage and support infrastructure rack
X Red/Black switching disconnect rack
Empty rack or planned expansion
APC PDU
N
E
S
W

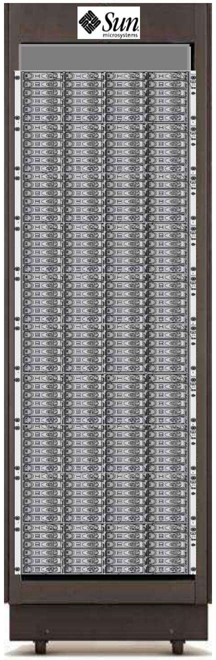


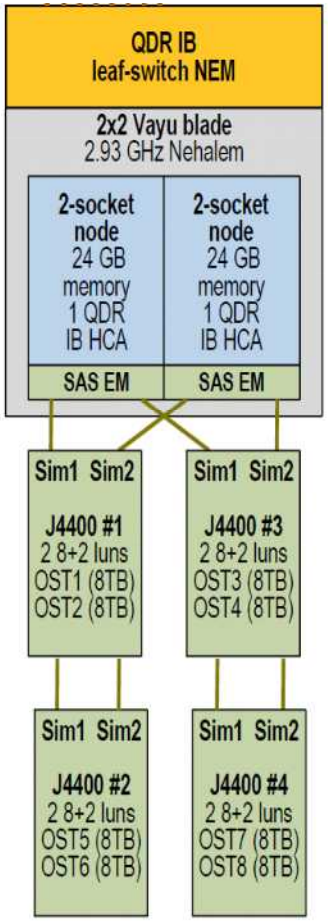
Sandia National Laboratories

Current Projects : Red Sky cont.

Lustre file system

- **Main points:**
  - Software RAID on Sun J4400 Open storage JBODS
  - 1 TB SATA disks for OST's (RAID6 with external mirrored journal)
  - 450 GB SAS disks for MDT's (RAID0+1)
  - 2 scratch file systems at ~ 1 PB each running at ~22 GB/sec
  - /home and /projects on Lustre





QDR IB
leaf-switch NEM

2x2 Vayu blade
2.93 GHz Nehalem

2-socket node
24 GB memory
1 QDR IB HCA
SAS EM

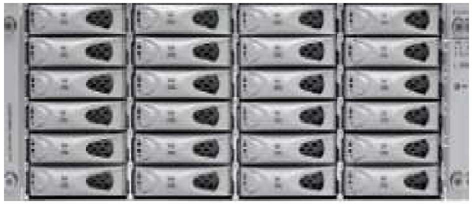
2-socket node
24 GB memory
1 QDR IB HCA
SAS EM

Sim1 Sim2
J4400 #1
2 8+2 luns
OST1 (8TB)
OST2 (8TB)

Sim1 Sim2
J4400 #3
2 8+2 luns
OST3 (8TB)
OST4 (8TB)

Sim1 Sim2
J4400 #2
2 8+2 luns
OST5 (8TB)
OST6 (8TB)

Sim1 Sim2
J4400 #4
2 8+2 luns
OST7 (8TB)
OST8 (8TB)



Current Projects: Dark Storm

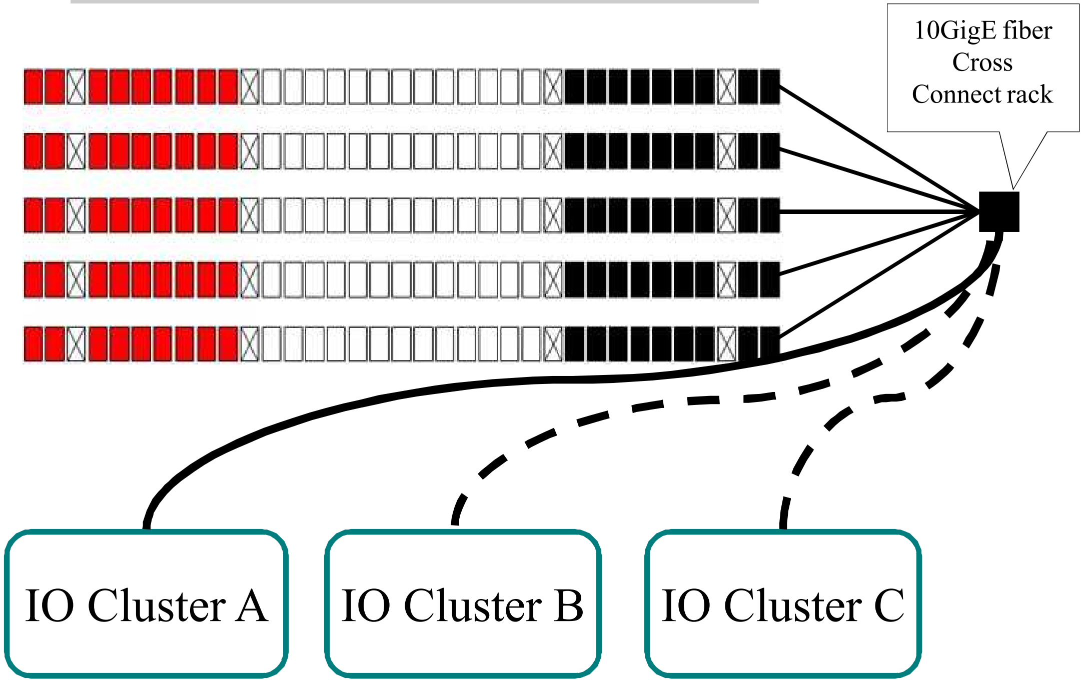
- Red Storm providing capability cycles with separated clusters providing the file systems
- Fiber Cross connect allows switchable links to IO cluster based on customer needs
- 50 SeaStar to 10GigE routers on Red Storm
- Woven EFX1000 switch on each IO cluster
- All storage is located on the IO clusters
- Challenges:
  - Catamount client, Liblustre, routing
  - “Home” file system on Lustre
- Friendly users are running on the system with good success.

Red Storm:

-Up to 12,960 nodes (38,400 Cores)

-Unicos 2.04.1

-SeaStar 2.1 in a 27x20x14 mesh



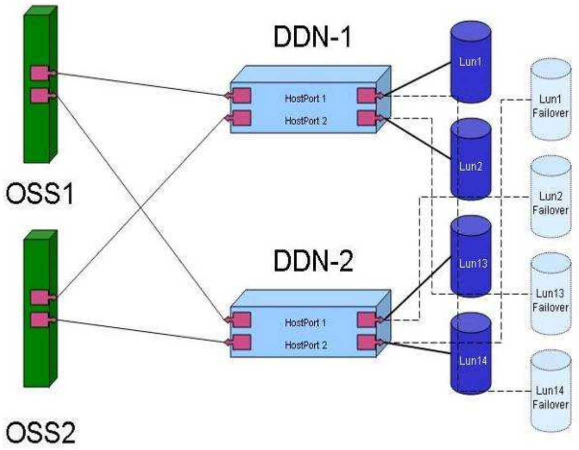
Failover

- **Many early challenges were related to back end storage failure, which prompted us to investigate Lustre Failover**
  - **Currently we have one file system deployed with a Lustre Failover configuration.**
    - **It’s a manual operation and we have not done a failover while in production**
    - **So far its been easier to fix the failing component and avoid the failover**
  - **Goal is to have automated failover cover ~80-90% of our failures**
    - **Automation is hard and initial deployment may involve manual (sys-admin) intervention**

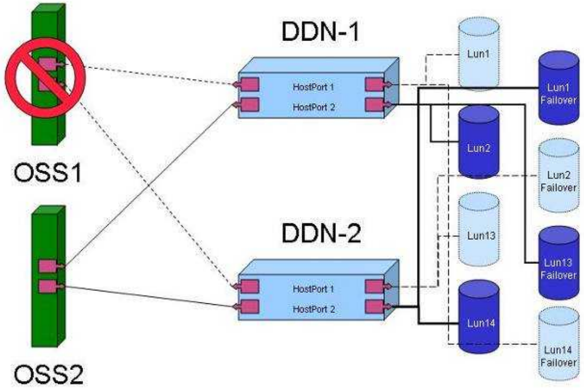
Failover cont.

- Failover Needs to cover:
  - Host Failures (OSS)
  - RAID Controller failures
- To avoid Data corruption:
  - must be sure that only one host can access a LUN at a time!
  - Host failure:
    - Power off the host (STONITH)
  - RAID Controller Failure:
    - Disable host IO to controller
    - Multi-Path should solve this case!

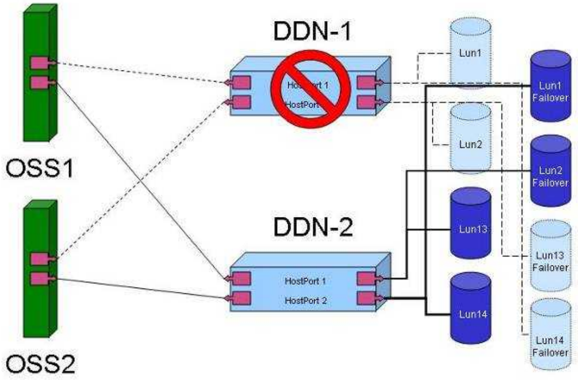
Normal operation with Zoning



Host (OSS) Failure



RAID Controller Failure



Things we've learned

- **LNET Routing has not been easy**
  - Taking Lustre from a local only resource to a "globally" mounted file system was a challenge and took several iterations to get it right
    - It took several months for CFS/SNL to figure out a client crash issue while trying to deploy to Thunderbird cluster (2006) (famous "lost ticks" bug..10375)
    - 16-32 bit LNET routers in early days of Tbird were a source of many of our problems
  - Routing is much more stable now thanks to work by Sun and some nice collaborations with our friends at LLNL and ORNL

- **Storage and recoverability issues**
  - We tend to find corner case issues with early generations of RAID hardware
    - Typically firmware revisions fix these problems
  - turn the DDN controllers write cache off as it is (still) painful to run file system repairs on 2-4 TB LUN's
    - This does have a negative performance impact, but it is important that users get the file system back quickly after a failure
- **SNL's capacity computing users, if given the choice, value file system uptime more than performance**
  - Multi-cluster file systems become the backbone of several clusters…when the file system is down all the clusters are impacted

- **Partnership's with Sun(CFS) and DDN have been very valuable**
  - Weekly conference calls keep the communication levels high and allow for good issue tracking

Future

- Failover operational on all Lustre file systems
  - Looking into using Multi-path to deal with failed RAID controllers
- Lustre 1.8+
  - Will be tested on our permanent test-bed
- Simplified System Administration
  - Transition all “disk-full” Lustre servers to Sandia oneSIS diskless image ensures consistency across the enterprise
  - Currently using the TOSS (Tri-lab Operating System Software) in a test environment
- Lustre as a NAS (NFS) replacement
  - Goal is to have small, highly tuned Lustre file system serve out our /home and /projects areas
  - Appealing for very large traditional linux clusters where NAS/NFS solutions have difficulty with the number of nodes…Lustre scales well out to the 10K clients range
  - Doing this now in support of Dark Storm
  - All user facing storage on Red Sky is Lustre based (scratch and home)

Future cont.

- **Simplification of storage infrastructure**
  - **storage appliances with 3 cables: power, Ethernet and high-speed interconnect**
    - **Current solution involves separate server nodes with IB interconnect and Fiber Channel connecting to RAID controllers that then have Fiber Channel connections to disk trays which then connect to SATA/SAS disk drives..**
      - **Complicated topology with many failure points!**
      - **Proprietary solution with typically good overall performance, but can be very expensive**
    - **Budgetary concerns are driving us to look at commodity based storage systems**
    - **Lustre with ZFS should help with this effort**
  - **IB attached storage**
    - **Replace Fiber channel to RAID controller connections with IB**
    - **Provides relatively low cost SAN solution, simplifies our components (no Fiber channel cards, better server to bandwidth ratio=> fewer servers)**
    - **Have several IB attached DDN 9550's in our test-bed now**
- **Multi-hop routing to help work around some of our facility and existing hardware limitations**
  - **E.g. IB <-> 10GigE to 10GigE <-> IB routers**
  - **We've tested this and it works..this will be happening for Phase 2 of Red Sky (site file system access)**
    - **Would like to see a fix for the "asymmetric failure of a router" (bug 18460) issue**

Questions?

Sandia National Laboratories