SAND2008-2934C

Data Challenges Facing Sandia National Laboratories

Brian Kellogg

brkello@sandia.gov

DICE Alliance '08

May 2008

Sandia National Laboratories

Current System: Red Storm

- Developed with a partnership between Sandia and Cray
- Theoretical Peak: 124.43 TF
- 12,960 compute nodes
- 340 TB of disk storage
- 39.19 TB of memory
- Topology 3-D mesh (27 x 20 x 24)
- MPI Latency 4.78 µs 1 hop, 7.78 µs max
- Bi-Directional link bandwidth: 9.6 GB/s
- 25 GB/s external network bandwidth
- Lustre based file system

SNL CapViz Lustre Configuration

- **Lustre version 1.4.X**
  - **versions of 1.4.X in production for 3 years now. Currently running 1.4.11.1**
- **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**
- **Storage hardware:**
  - **DDN (DataDirect Networks)**
    - **31 - 9550 Controller cuplets (FC4/SATA) for OSS's**
    - **4 - 8500 Controller cuplets (FC2/SATA) for MDS's**
    - **7,440 SATA disks in production!**
      - **Mix of 250 and 500 GB disks.**

Lustre Configuration Cont.

- **File Systems:**
  - **2 main production file systems (Red and Black)**
    - **360 TB: 8 DDN cuplets with 31 OSS's (186 OST's)**
    - **1 PB: 11 DDN cuplets 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: 1,200 clients (growing to 1600 with TLCC)**
  - **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**

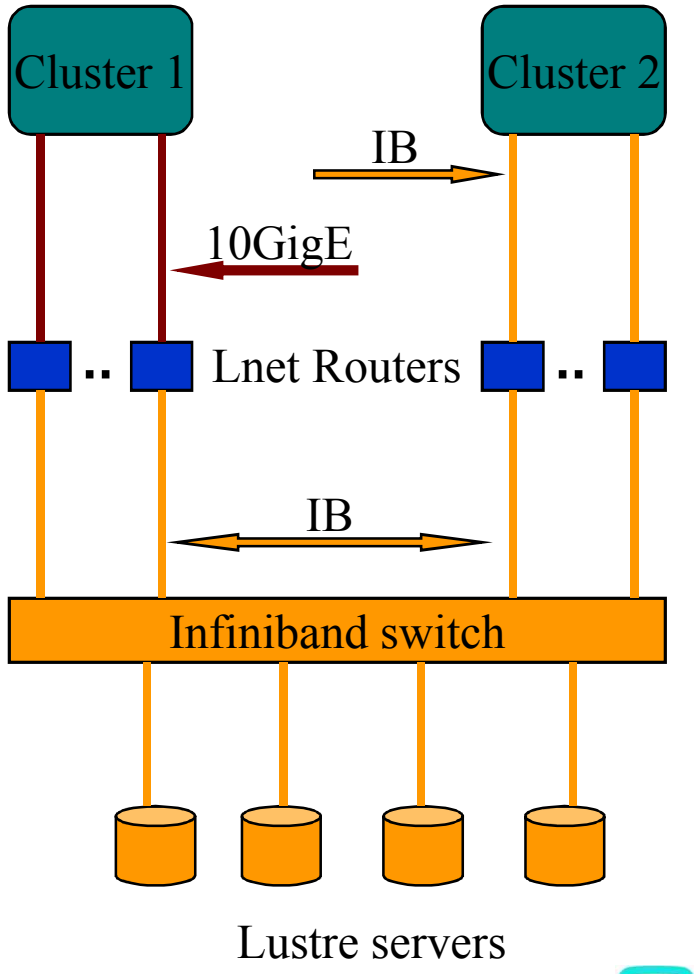
Multi-Cluster Lustre

- **LNET (Lustre NETwork) "routing" is key to sharing a single Lustre file system with several clusters.**

- **Lustre routing provides:**
  - **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!!!**

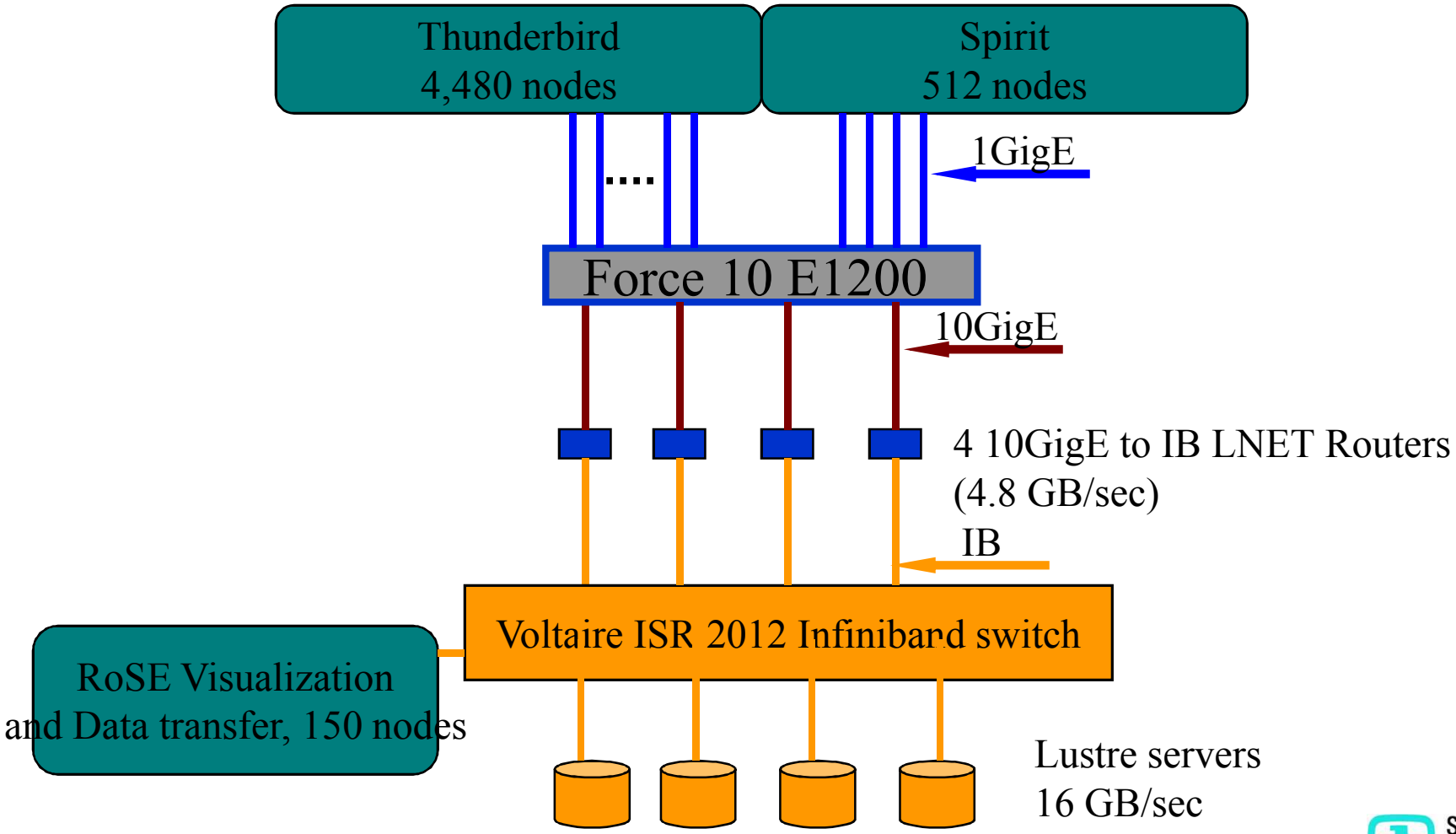
Router configuration

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



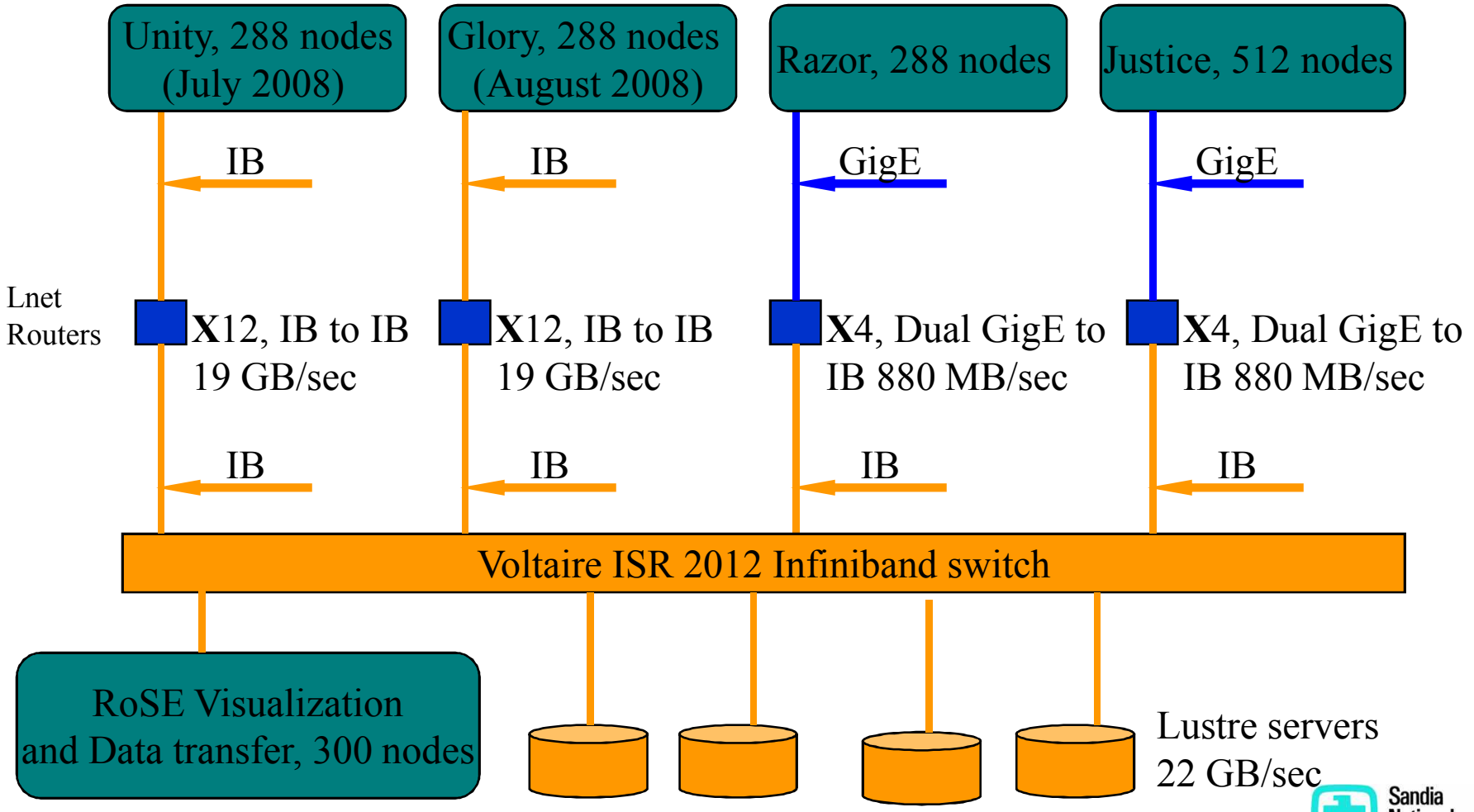
Sandia National Laboratories

SRN Router configuration



Sandia National Laboratories

SCN Router configuration



Sandia National Laboratories

Benefits of a Multi-Cluster file system

- Avoid Islands of storage located within a given cluster.
- 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.

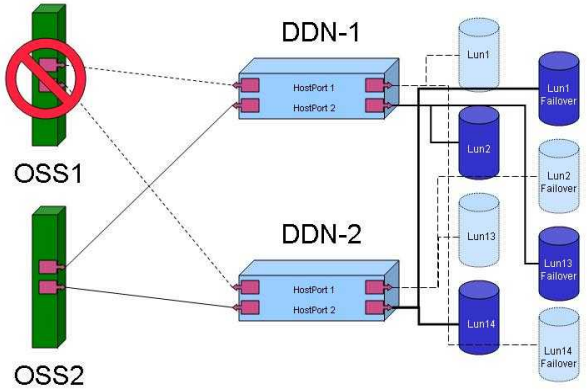
Things We've Learned

- Our biggest failure point has been related to back end storage problems
  - We are currently testing Lustre's failover capability
  - Goal is to have automated failover cover ~80-90% of our failures.
    - Automation is hard and initial deployment may involve manual (sys-admin) intervention.

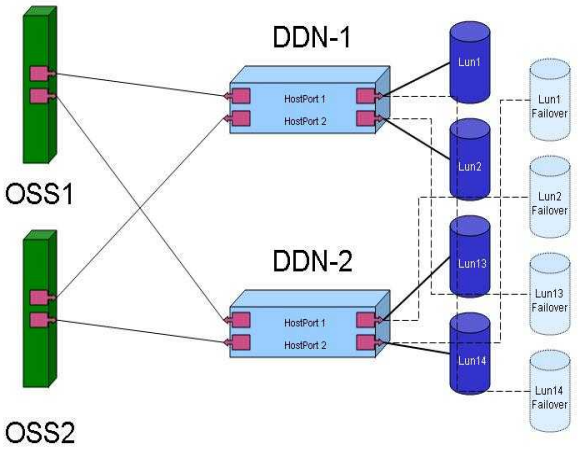
Things We've Learned (cont)

- **Failover Needs to cover:**
  - **Host Failures (OSS)**
  - **RAID Controller failures**
- **To avoid Data corruption:**
  - **we 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**

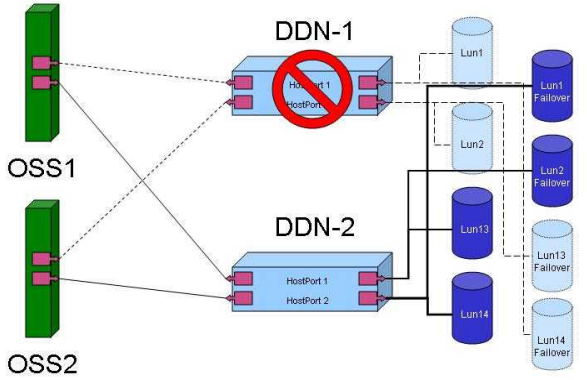
Host (OSS) Failure



Normal operation with Zoning



RAID Controller Failure



Sandia National Laboratories

Things We've Learned (cont)

- **Routing is not perfect**
  - It took several months for CFS/SNL to figure out a client crash issue while trying to deploy to Thunderbird cluster (2005)
  - Routing is much more stable now, but it still has some corner failure cases that we are working with Sun(CFS) to fix
- **Storage and recoverability issues**
  - We 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 capacity users 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**
  - This is one of our highest priorities.
- **Lustre 1.6.x**
  - Recently went through a pre-deployment test on 1.6.3 and we were pleased, but are holding back on official cutover until Summer of 2008
  - OST leveling, network checksums and ability to easily add OST's are some of the appealing features
- **Lustre as a NAS (NFS) replacement**
  - 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.

Future (cont)

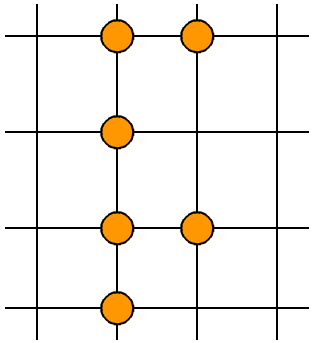
- Simplification of storage infrastructure
  - "all in one" storage appliances with 3 cables: power, Ethernet and high-speed interconnect
    - Current solution involves separate server nodes with IB interconnect and Fiber Channel connecting to RAID controllers that then have Fiber Channel connections to disk trays which then connect to SATA disk drives..
      - Complicated topology with many failure points!
  - 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)

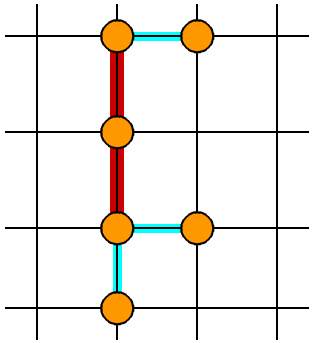
Autonomics

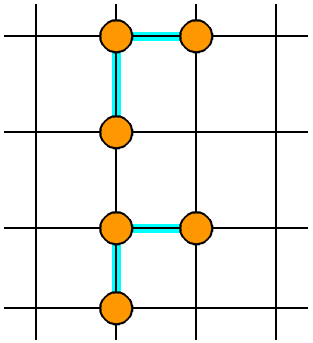
- Current machine is 13k nodes
- System has decent support for operating on the machine as an aggregate
- Much less support for problem mitigation/recovery
- Automatic diagnosis of problems and repair
- Storage can have its own subsystems
  - How can you make sure everything is configured the same
- We need help to automate configuration tasks and make it more deterministic
- The future is scary
  - Machines with 100,000s of nodes

Intelligent Job Schedulers

- We need topologically aware schedulers
  - Products like these exist, but no one is currently working with us







High Performance Storage

- We don’t need fail stop
- Not looking for 5 9s but 3 9s would be nice
- Don’t need bells and whistles
- Need the basics done well
  - Fast and reliable
  - Cheap
  - Redundant Paths
  - Command line interface over the net
- DDN very fast, but we lose one a week

Distributed Parallel File Systems

- Red Storm capable of 50 GB/s to applications
  - Can only make use of 20 GB/s at best

- New Los Alamos machine (Zia) spec is 1 TB/s
  - Should be built in 2009
  - Follow on machine to come in 2014/2015
  - NRE funding available (talk to Lee Ward)

Questions?

- Not me!
- For future-looking research at Sandia and Zia contact Lee Ward
  - lee@sandia.gov
- For info regarding Multi-cluster Lustre contact Steve Monk
  - smonk@sandia.gov