A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

NICS Lustre Experiences on Cray XT5

Troy Baer

Victor Hazlewood

Junseong Heo

Rick Mohr

John Walsh

NICS Lustre Experience

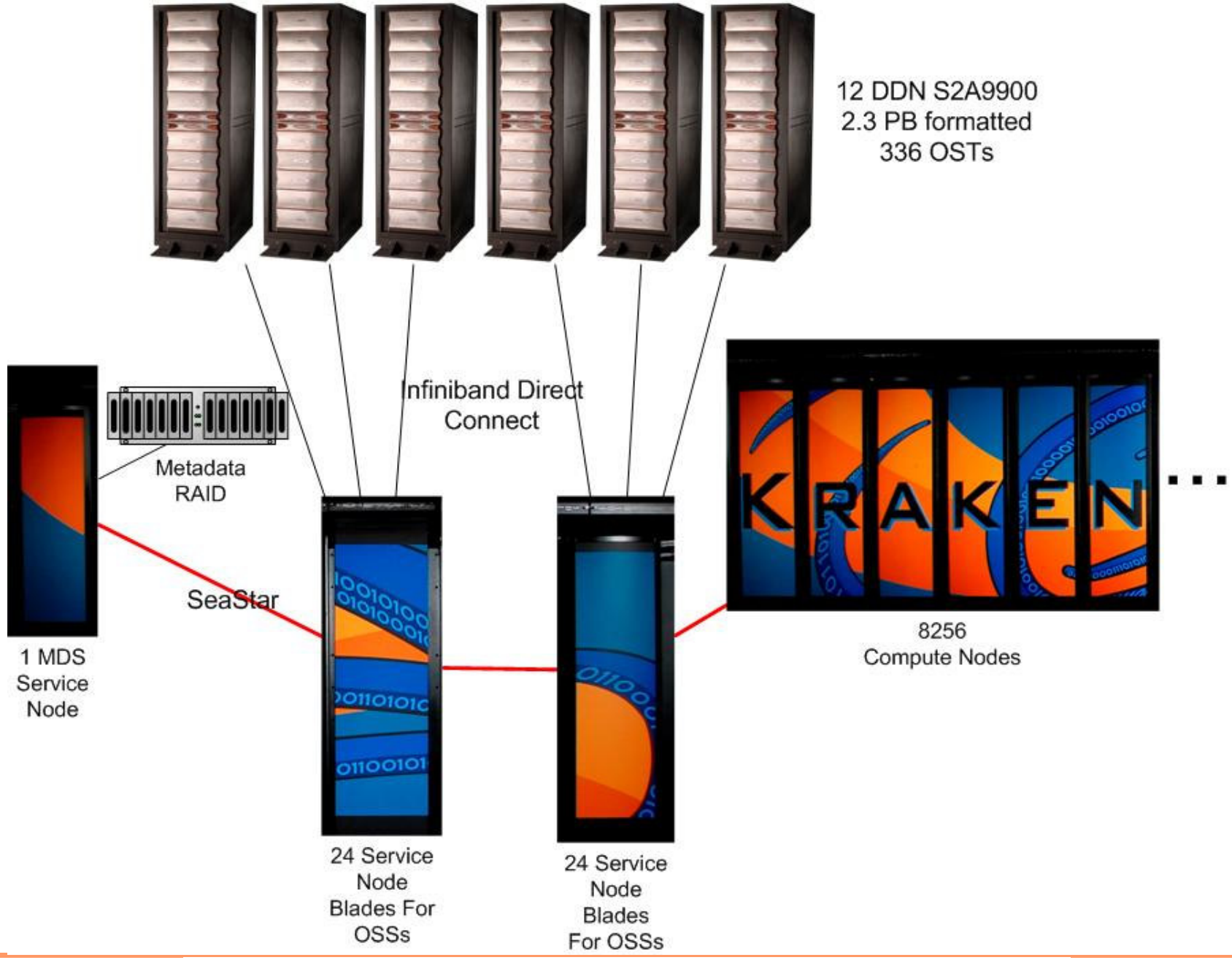
- Overview of NICS and Lustre at NICS
- Building the Lustre file system
- Question of Purging vs. Quotas
- Configuration and Limitations
- Canary in the Coal Mine
- Wrap Up

NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES

NICS

Kraken Cray XT5 Overview

- 88 cabinets
- 8256 compute nodes (66,048 cores)
  - 4416 2GB nodes
  - 3840 1GB nodes
- 72 service nodes (48 OSS/1 MDS/23 other)
- SMW 3.1.10/CLE 2.1.50
- Lustre 1.6.5-2.1.50HD


12 DDN S2A9900
2.3 PB formatted
336 OSTs
Infiniband Direct
Connect
Metadata
RAID
SeaStar
1 MDS
Service
Node
24 Service
Node
Blades For
OSSs
24 Service
Node
Blades
For OSSs
KRAKEN
8256
Compute Nodes

Lustre Overview

- MDS server w/ metadata RAID
- 48 OSS
- 336 OSTs (7 OSTs per OSS)
- Infiniband direct connect (no IB switch)
- 12 DDN S2A9900 in 6 cabinets
- 2.4 PB formatted / 3.3 PB unformatted

Building the Lustre File System

- All IB rpm's not identified when we installed
- Went through 3 iterations of building Lustre
  - 380TB file system
  - 1.3 PB file system
  - 2.3 PB file system
- Format of the entire 2.3PB was done in stages by a script. A format of the entire space overwhelmed the MDS

Question of Purging vs. Quotas

- "How to manage the space?" was the big question
- Use a purge script?
- Turn on quotas? And how to best make use of quotas, if used
- Different types of performance impacts depending on purging vs. quotas

Question of Purging vs. Quotas

- XT4 experience showed that walking the file system to identify files to purge was prohibitively slow and degraded metadata performance on the file system to an unacceptable degree
- Decided to do performance tests to quantify the impact of quotas just after acceptance of XT5

Question of Purging vs. Quotas

- Three separate tests were run:
  - A threaded file creation and deletion test on a single compute node
  - A file-per-process bandwidth test, using the standard IOR benchmark
  - A shared file bandwidth test, also using IOR
- These tests were run in four situations:
  - Before the file system rebuild
  - After the file system rebuild with quotas disabled
  - After the file system rebuild with quotas enabled but not enforced
  - After the file system rebuild with quotas enforced

Question of Purging vs. Quotas

- Threaded file creation/deletion test showed a substantial improvement in MDS performance of the DDN EF2915 array relative to the LSI RAID boot array used before the rebuild.
- The sustained rate of file creation increased by 59%
- The sustained rate of file deletion increased by a surprising 718%.
- After enabling quotas, these rates did drop slightly, by 12% in the case of file creation and 10% in the case of file deletion.

Question of Purging vs. Quotas

- No measured performance impact on file-per-process I/O by enabling quotas
- Write and read performance to a shared file dropped by 6% and 1% respectively
- Enabling quotas actually improved write performance by 5% but also decreased read performance by 5%, while enforcing quotas effectively reversed the situation. The maximum impact of quotas observed on shared-file I/O performance was 6%.

Question of Purging vs. Quotas

- Results of testing showed
  - a metadata performance penalty of 10-12%
  - a maximum bandwidth impact of 6%
- Therefore we chose to move forward with quotas being enabled but not enforced.
- We suggest that this small performance penalty will be largely offset by not having to traverse the Lustre file system periodically in order to generate a file purge list and the performance impact that goes with it
- However, relies on users to take action…

Configuration and Limitations

- Budget constraints limited our configuration which led to some interesting tradeoffs in performance and capability
  - a minimum number of controllers
  - high number of OSTs per OSS
  - large number of Lustre clients (O8300)

Configuration and Limitations

- 30 GB/s demonstrated sustained performance using IOR benchmark. About 5 GB/s per cabinet
- No redundant paths, therefore, no failover capability ☹
- Ended up with 48 OSS servers, 7 OSTs per OSS

Configuration and Limitations

- Tunable parameters changed:
  - Portals “credits” increased
    - 512 for compute nodes
    - 1024 for OSSs
    - 2048 for MDS
  - Timeout increased to 250 seconds to prevent timeout, eviction and reconnect looping
  - Default stripe count: 4
  - Default stripe size: 1MB (users need training on stripe sizing!)

Canary in the Coal Mine

- Hardware issues affecting the portals network are not always noticed until Lustre generates errors, generally, followed by user complaints of file system “hang”. Need more user education to address this.
- Lustre errors in the logs continues unless the associated hardware issues are resolved, mostly by a system reboot and removal of problematic hardware.

Canary in the Coal Mine

- We typically see half million to seven million lines of Lustre error messages a week
- Once we separate interconnect failure caused error messages, Lustre messages are predictable and consistent
- Failed nodes are identified by the timeout and eviction sequence
- Heavy concurrent I/O patterns beyond the current bandwidth limits manifest themselves as a global delay

Canary in the Coal Mine

- Coordinating Lustre errors with netwatch log messages usually precedes an HSN collapse
- The HSN sometimes recovers by itself eventually ingesting all the portal traffic.
- We did see the self recovery twice during last three months. But it tends to have lingering effects and job performances become unpredictable after such recovery.

Lustre Monitoring

- Lustre error counter: monitors the Lustre warnings, Errors, and ratio of the two.
- Lustre hang sampling: random interval checks on Lustre response time and is logged continuously during production.
- Lustre File system state: number of files generated and total disk space used are recorded hourly.

Wrap Up

- Lustre seems to provide early warning of system failures both detected and undetected. It is our “canary in a coal mine”.
- Quotas enabled but not used seems to provide a decent tradeoff between automated system purging and full quotas. Still have to depend on users to take action.
- NICS is iteratively improving our Lustre monitoring with a combination of log watching, Lustre file system response time and file system state

NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES

NICS

NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES NICS