Sequoia and the ZFS OSD

April 16, 2013

Christopher Morrone
Sequoia – IBM BG/Q
Sequoia
1,572,864 Compute Cores

- 96 racks
- 1024 nodes per rack
- 16 compute cores per node
32 Compute Nodes on Node Card
Sequoia Naked
Sequoia Clothed
Lustre Client on I/O Nodes (IONs)
Sequoia Filesystem Stats

- 55 Petabytes
- 850 GB/s measured sustained write throughput
- 768 OSS & OST
- Each OST is 72 Terabytes
Filesystem building block
NetApp E5460 + 2 OSS

40Gbit IB

OSS

40Gbit IB

OSS

NetApp Controller

10 X 3TB Drives RAID6

10 X 3TB Drives RAID6

10 X 3TB Drives RAID6

10 X 3TB Drives RAID6

10 X 3TB Drives RAID6

10 X 3TB Drives RAID6
Sequoia's Filesystem Hardware
Sequoia's Filesystem Network

- 768 IONs
- ION leaf switch
- Login/DM leaf switch
- 2 - 648 port QDR IB switches
- Expansion
- Lustre OSS Nodes
- MDS Leaf switch
- Lustre OSS Nodes
- Lustre MDS Node
- 768 OSS Nodes
Lustre OSD and OSD-ZFS in 2.4!

- Lustre OSD & ZFS planning began in 2007
- LLNL contracted w/ Whamcloud to complete
OSD is not just for ZFS!

- Address long-existing technical debt
- Provides foundation for DNE, and other features
ZFS Benefits

- Copy-on-write sequentializes random write
- Single volume size limit of 16 EiB
- Zero off-line fsck time
  - Online data integrity checking and error handling
- Scalable directory contents
  - LLNL tested 1.5+ million files in one directory
ZFS Additional Benefits

- Compression – Average 1.7:1 on Sequoia
- Snapshots
  - Great for investigating Lustre problem!
- Multiple filesystems in same pool
  - E.G. test Lustre FS with less danger to production FS
- Other ideas
  - zfs send/recv for backups or copies
  - Snapshot entire Lustre filesystem
Sequoia Filesystem Status

- **Lustre OST over ZFS**
  - Stability: good
  - Performance: good

- **Lustre MDT over ZFS**
  - Stability: good
  - Performance: could be better

- **Lustre Client**
  - Stability: acceptable
  - Performance: needs work
ZFS Packaging

- Arch Linux
- Debian
- Fedora
- Funtoo
- Generic DEBs
- Generis RPMs
- RHEL / CentOS / SL
- Sabayon
- SprezzOS
- Ubuntu
Lustre+ZFS Installation is easy!

- http://zfsonlinux.org/lustre.html

- RHEL example:
    - Adds `/etc/yum.repo.d/zfs.repo` and signing keys
  - `sudo yum install lustre`
http://zfsonlinux.org