# Hedging Our Filesystem Bet

April 16, 2010

IT&S
information technology and services

# Our Business

- Provide computational resources for Seismic Imaging and Algorithm Research
- Quick turnaround of ad-hoc requests related to seismic acquisition and processing as well as answering questions that arise during drilling
- Direct customer base about 50 geophysicists
- Indirect customer base 2000 users around world
- We are considered an R&D shop so we don't have stringent uptime requirements but a job on 900 nodes wastes a lot of cycles if single node crashes due to NIC problem

# Our Challenges

- Support multiple projects at all times
  - Projects in different phases
  - Various resource needs
    - Memory Size
    - MPI or not
    - Network bandwidth
    - I/O patterns vary by algorithm

- All systems need to see all data

# Principles

- KISS – Keep It Simple Stupid
- Balance bottlenecks of compute, network, storage by adjusting allocation of budget to each
- Maintain competitive environment for suppliers – budgets are always being squeezed
- Maintain revolving door so we always have some of newest toys and always have oldest toys going off lease
- Choose strategic direction where value is added but keep eye on opportunities and don't get boxed in corner
- Maintain vendor neutrality where possible and avoid being boxed in a corner by any vendor

# Compute Resources

- Large Memory
  - 28 IBM Power6/575 – 32 core 256GB ea
  - 1 SGI Altix 4700 – 256 core 1.5TB
  - 5 SGI Altix 4700 – 128 core 512GB ea
  - 6 SGI Altix 4700 – 64 core 384GB ea *
  - 4 SGI Altix 3700 – 64 core 512GB ea
- Small Memory
  - 1240 HP/Dell Nehalem – 8 core 48GB ea
  - 960 ??? Westmere – 12 core 48GB ea **
  - 800 Dell DCS Harpertown - 8 core 32GB ea
  - 720 Dell PE1950 Clovertown – 8 core 16GB ea *

# Network

- Ethernet
  - Foundry (2)MLX-32, RX-16 switches in triangular core
  - Myricom switches 2 @ 512port, 2 @ 256port all full of Nehalem compute nodes, 2 @ 128port (half full with balance of large memory compute nodes), ethernet uplinks to core  (32up, 480node)
  - Testing MLAG and Layer3 features of Arista 7148SX switches with eye to larger switch
  - Direct 10G connections from Panasas/Lustre/GPFS storage to core
  - Assortment of low end Foundry TOR switches for Harpertown and Clovertown nodes

# Storage

- 1.4PB SGI CXFS – Long relationship, great for SAN attached systems but sucks for NFS
- 2.2PB Panasas – Long relationship, great for cluster access, single stream access getting better
- 300TB Lustre – Installed end of 2008, great for cluster access, good single stream access, getting more comfortable with reliability, learning to ignore most error messages – Fundamentally equivalent to Amber Road
- 1.75PB GPFS – Installed late 2009, production Jan 2010 – Already had several patches to code, running into few things as we scale, couple corrupt files we can't clean up without offline fsck – waiting until scheduled outage
- 5.65PB current -> 7.5PB by end 2010

# Lustre filesystem

- Substantially equivalent to Snowbird (X4440 & J4400)
- Pair of servers for MGS/MDS
- 5 sets of OSS pairs
  - 4 primary OSTs per OSS
  - Failover for 4 OSTs per OSS
- Myrinet 10G (IP) connection for each machine
- Initially Lustre 1.6.7.1 then quickly upgraded to 1.6.7.2 to fix couple bugs.  Still had to install couple patches for Hung CPUs, MPTSAS, and RAID5
- Some implementation issues – Snowbird not completely defined when we implemented, couple issues with type of SAS controllers and number required for failover – Sun took care of those

# Storage Concerns

- CXFS - Retiring
  - doesn’t scale outside SAN
  - will SGI continue to support/enhance it
- Panasas - Continuing
  - Expensive – needs competition, current hardware long in tooth, easy management, great support
- Lustre - Limbo
  - Was our stick in dealing with Panasas until Oracle joined the fray
- GPFS – CXFS replacement – Further Potential
  - New stick with no track record (with us).  If it flops, convert to Lustre?  More difficult with just announced Lustre 2 certification model

# People

- Six FTEs deal with everything
  - Architecture
  - Interfacing with building personnel
  - HW/SW Installs/break-fix
  - Network
  - Filesystems
  - OS
  - Support
  - Removal

# Expected changes in 2010

- Clovertowns leave late April
- 1.4PB CXFS retires
  - 800TB converts to Lustre/GPFS
- Replace core network
- Westmere 960 - 1600 nodes
- NehalemEX/Power7
- Panasas
- GPFS

# Good things about Lustre

- Pretty Fast – 7GB/s on small FS with small number OSSes.
- Handles large number of clients – up to 800 clients tested routinely – per client throughput slows but servers keep going
- No data loss
- Seems pretty resilient to network, server, client issues thanks to recovery mechanism
- Filesystem in production for about 16 months – current uptime 130 days since unplanned building cooling “issue”
- More reliable than I expected

# Outage definition – "It Depends"

- Annoyance – filesystem goes offline but comes back and applications resume where they were
- Problem – Filesystem becomes unavailable causing jobs to die, have to restart jobs that may have been running for days (restarts take too long)
- BIG problem – losing data that is already complete and sitting on disk

# Bad things about Lustre - Outages

- MGS failed over to MDS server – we didn’t know it for a while - Annoyance
- OSSes stopped working with stuck CPUs, failover to alternate OSS didn’t quite work but when we got OSS back up, applications resumed – Annoyance
- Few server fan failures – Annoyance

# Ugly things about Lustre

- Log analysis is horrible. Most admins don’t have time to follow all of the logs and some get worried by all the long error codes with no understanding of what they mean without further info.
- User causes severe problem with metadata due to poor job setup (bunch of jobs all writing debug output to single logfile). Can’t blame Lustre for this.
- Poorly written applications (small block read/write) stress out MDS. Becomes very apparent. Can’t blame Lustre for this.

# Suggestions for Lustre/Oracle

- Some good/bad thoughts of new Lustre 2 support model. Options restricted for getting support if for example GPFS flops and we think of converting existing hardware to Lustre.
- If you make certification of Lustre 2 configurations too onerous on vendors, we won't have options – Lustre drops off list of options.
- Understand and agree with trying to reduce combinations of hardware/OS levels being supported. Lustre certification suite for vendors is good start but restricts some options.
- Create tools to easily identify important Lustre errors vs. trivial ones that are informational only – better indication of severity level
- Get your database in order – Related to problems getting correct parts for Sun HW (not Lustre servers thank god) but gives a black eye to Oracle.

# Technical directions

- At LUG2008, question about size and number of files. At that time I was seeing 300GB files. Now I see many 1-3TB files, 1x 6.8TB files. Only see files sizes growing.
- Very few users have >10,000 files in directory.
- Filesystems growing to few PB size to have enough spindles to sustain I/O requirements as clients get faster.
- Large LUN support can't wait. 8/16TB LUNs won't be practical beyond 2TB drives. Ldiskfs needs to keep up. If ZFS on Solaris is only path forward then we won't be able to choose other integrators besides Oracle – therefore, Lustre will not be on our list of filesystem choices for very long.