




The Madness of Project George

SSDs, the cloud, and lots of speed



Ben Evans
Principal Lustre Architect
4/2/2013

TERASCALE



The Beginning

- Wandering around SC-11
 - » Discovered 1/2u, blade enclosures
- Later, calculate 42u of QDR
 - » ~250 GB/s/rack
- But what can you do with it?
 - » Drives not fast enough
 - » PCI-E SSDs might have the bandwidth
 - » Storage would be really shallow, but fast

Hints of a solution

- Customers complaining about single users
 - » Creating millions of tiny files
 - » Lots of small I/O
 - » Heavy I/O
- Generally, disruptive behavior
 - » Give them their own filesystem?
 - » Move the power users?
 - » Tuning?

Bricks

- Many 'bricks' make up a file system
- One brick contains:
 - » IB connection
 - » Enough SSD bandwidth to saturate the IB connection
 - » Memory, CPU, etc.

Temporary Filesystems

- Sized based on storage or performance needs
- Allocated by the resource manager from a pool of 'bricks'
- Formatted and tuned for the application
- Named based on Job#
- Initial implementation will be Lustre only, but no reason other file systems can't be used

Data

- Data copied onto the FS before the job starts
- Tar, gzip options for storing data on Primary FS
- Clients need to mount the TempFS
- Job is run as usual
- Results copied off afterwards

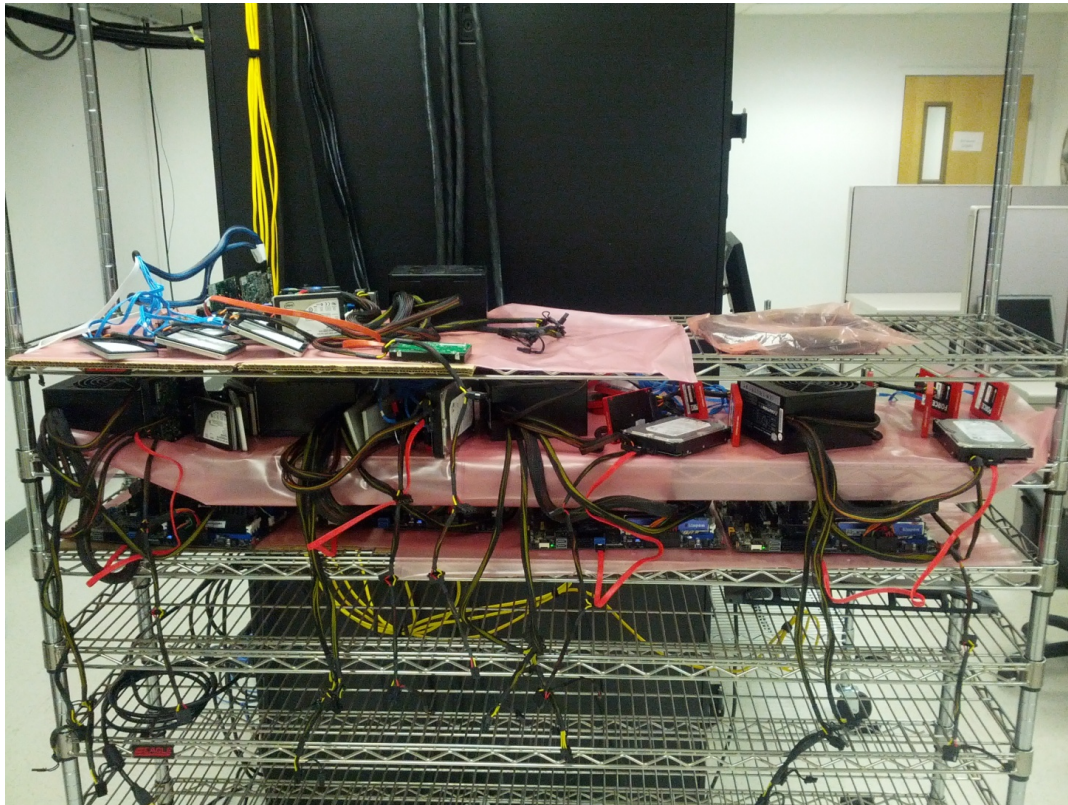
Notes

- Net gain if:
 - » Total job time is reduced (Data copy steps included)
 - » Reduces load on primary FS
 - » Lower PFS performance requirements
- Good fits
 - » Small input files, small results, big intermediate data
 - » High metadata requirements
 - » Unique requirements
 - » Random workloads
 - » High IOPS workloads

Other considerations

- Non-redundant
 - » If a brick fails, only one job is affected
 - » If a brick resets, job should continue, no data is lost
- Monitored
 - » can be removed from the available pool until repaired
- Inherent Job-level QoS, I/O monitoring, resource tracking ...

Current State



- 6 bricks, 18 GB/s
- Mobile design
- Managed airflow
- Energy efficient
- Inexpensive
- Patent filed

QUESTIONS?