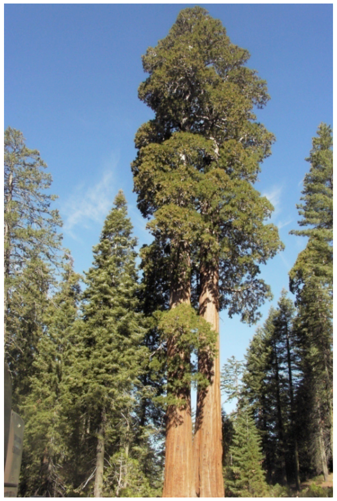
Lawrence Livermore National Laboratory

Sequoiadendron Giganteum: Infrastructure for the Petascale Giants

9/23/08



Mark Gary

File Systems Project Manager

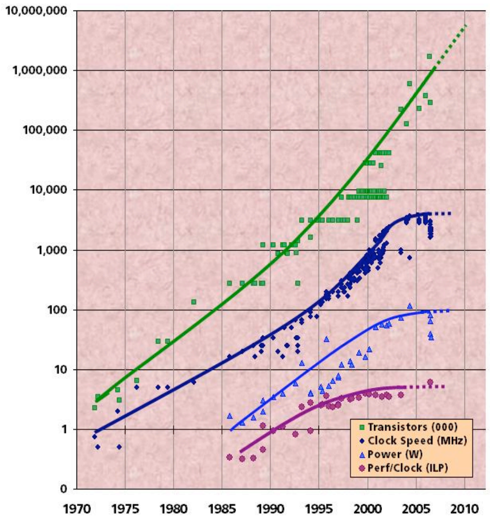
Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, CA 94551

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

LLNL-PRES-406576

How do we prepare a petascale infrastructure?

- ***By learning from the terascale***
- Identifying the hurdles ahead
- Aggressively investing in solutions/strategies

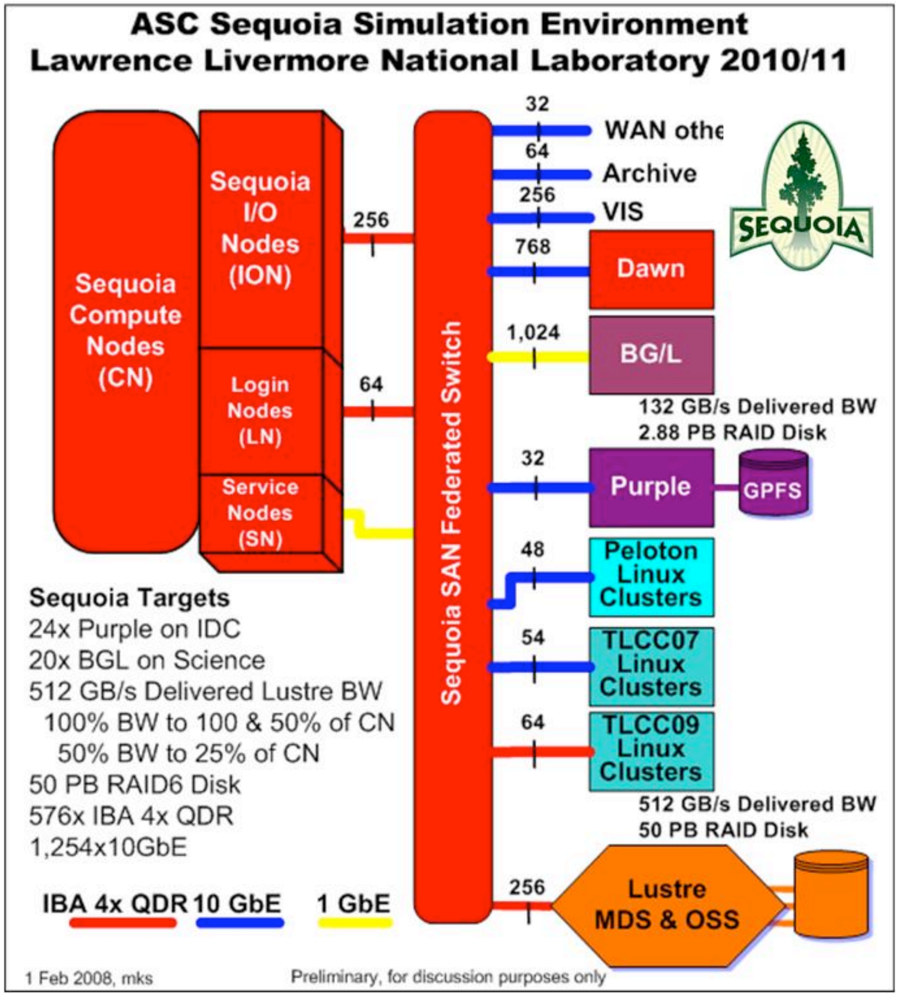


I'd like to use some mantras learned from the terascale as guideposts on a tour of how LLNL intends to provide robust petascale infrastructure

Overview

- **First, our petascale driver – Sequoia**
- **Some of our mantras**
  - *Infrastructure must be balanced*
  - *File systems should be global resources*
  - *I/O is random, small and bursty*
  - *Scalable Units – the only way one can survive*
  - *At-scale testing is not optional*
  - *Archives are forever (and not transparent!)*
  - *Collaborate, collaborate, collaborate*
- **New mantras forming**

Our petascale driver - Sequoia



ASC Sequoia Simulation Environment
Lawrence Livermore National Laboratory 2010/11

Sequoia Targets
24x Purple on IDC
20x BGL on Science
512 GB/s Delivered Lustre BW
100% BW to 100 & 50% of CN
50% BW to 25% of CN
50 PB RAID6 Disk
576x IBA 4x QDR
1,254x10GbE

1 Feb 2008, mks

Preliminary, for discussion purposes only

- **We have a multi-PetaFlop machine arriving going into production in 2012**
- **Furthers our ability to simulate complex phenomena “just like God does it – one atom at a time”**
  - **Uncertainty quantification**
  - **3D confirmations of 2D discoveries for more predictive models**
- **The success of Sequoia will depend on an enormous off-machine petascale storage infrastructure**

Mantra #1: Balanced Infrastructure

Or… Without a balanced infrastructure, a supercomputer is a very expensive doorstop

- Requires very large investment (networks, file systems, archive, software development, customer support, data analysis, *facilities*…)
- Bleeding edge platforms require leading edge infrastructure
  - Off the shelf solutions often cannot deliver the performance we require

Requires that management and funding agents understand this and have the discipline to back it up with resources

To provide balance for Sequoia -

Fortunately, our management and funding agents understand the importance of balance

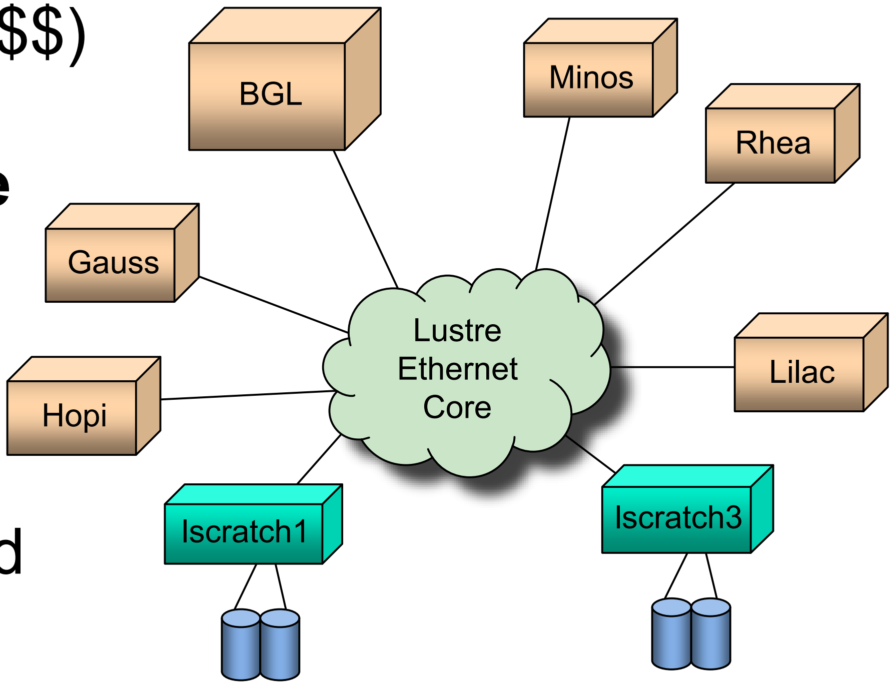
- **Platform procurements are *preceded* by well-funded file system and network procurements**
- **We invest in targeted software development**
  - File system development (Lustre)
  - Operating systems (TOSS)
  - Archive development (HPSS)

- **We collaborate with peers, vendors, academia…** <more on this later>
- **We investment in testbeds** <more on this later>
- **We plan together – I/O Blueprint process**

Mantra #2: File systems should be global

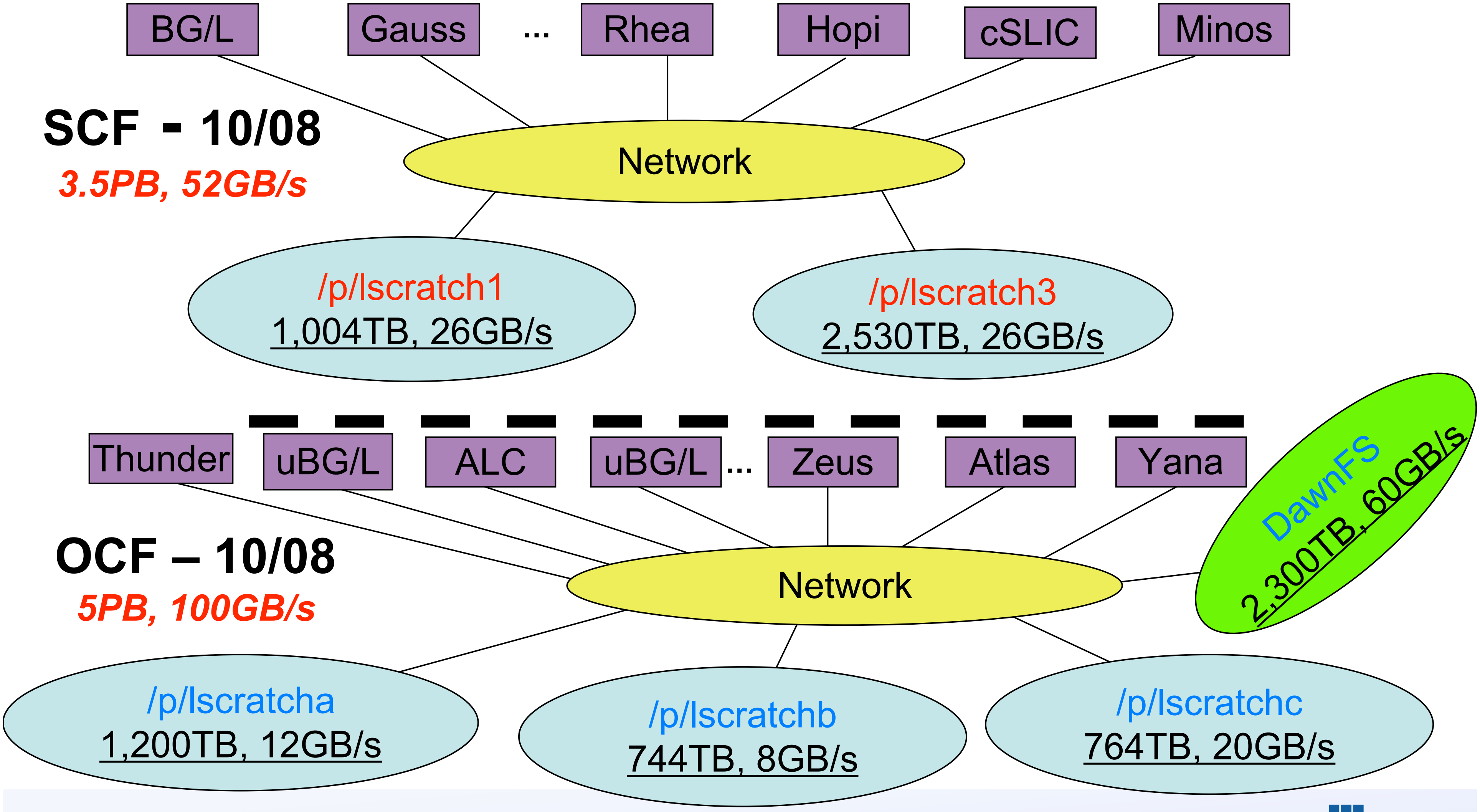
- **Field file systems as shared, multi-cluster resources***
  - No need to move data
  - Intelligent use of resources ($$)

- **Target of 2-3 file systems/side**
  - Failure mitigation
  - Downtime mitigation
  - Simplified administration
  - More capable (bandwidth and capacity) single file systems
  - Large-scale run dedication



* *but there are tradeoffs*

LC file systems - 10/08 – preparing for Sequoia



BG/L

Gauss

...

Rhea

Hopi

cSLIC

Minos

SCF - 10/08

3.5PB, 52GB/s

Network

/p/lscratch1

1,004TB, 26GB/s

/p/lscratch3

2,530TB, 26GB/s

Thunder

uBG/L

ALC

uBG/L

...

Zeus

Atlas

Yana

DawnFS

2,300TB, 60GB/s

OCF – 10/08

5PB, 100GB/s

Network

/p/lscratcha

1,200TB, 12GB/s

/p/lscratchb

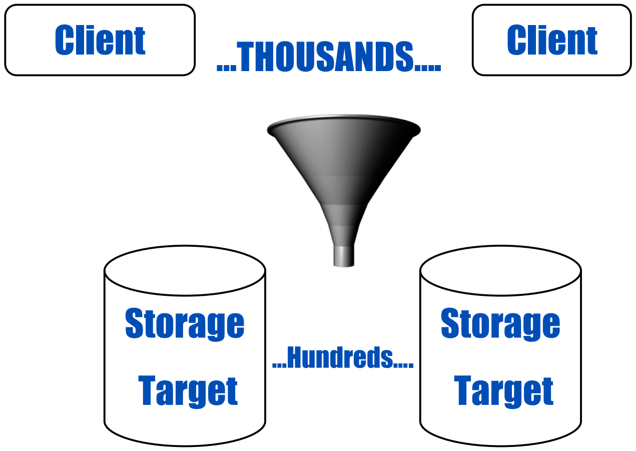
744TB, 8GB/s

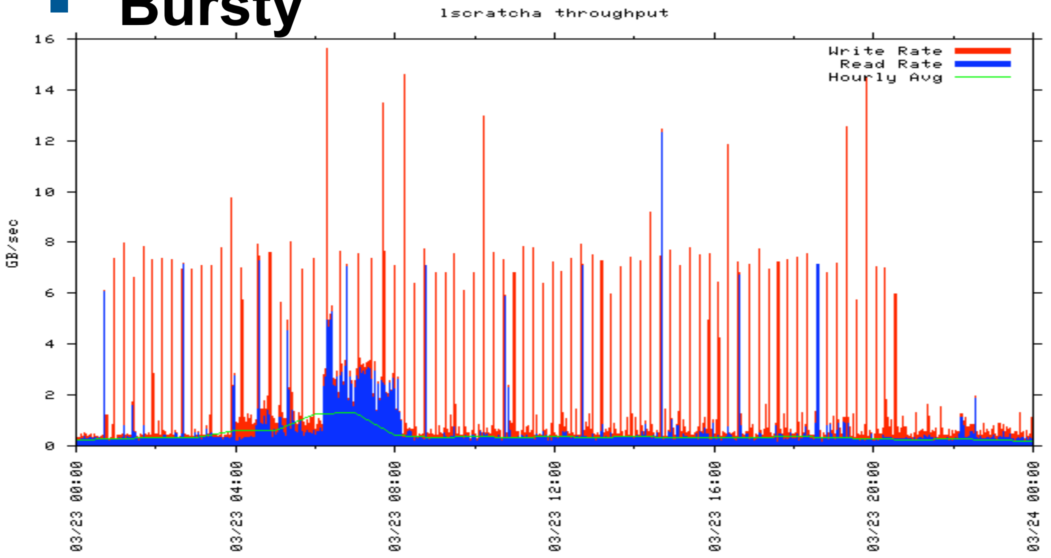
/p/lscratchc

764TB, 20GB/s

Mantra #4: I/O is small, random and …

- **Small** (but linked):
  - File per process and file per core
  - FS/network sweet spot limited
- **Random** (from disk perspective):
  - Product of the scale of compute resources, global file systems
- **Bursty**





As a result our file system requirements are driven by IOPS performance - not capacity

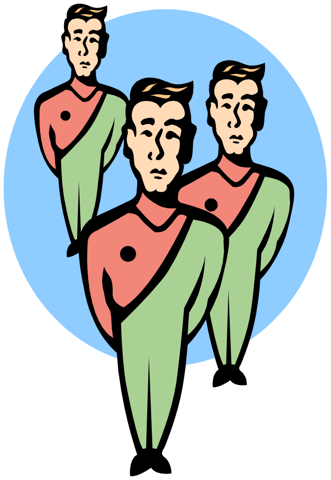
What we are going to do about this?

- **Increasing reliability - reducing defensive I/O**
- ***Working directly with users on their I/O strategies***
- **Focus on device IOPS**
  - FC vs. SATA…
- **Scheduling**
  - I/O request scheduling work
  - Batch scheduler/file system integration, QoS?
- **Copy-on-write (ZFS in Lustre)**
- **Clustered MetaData Servers**
- **Other algorithmic approaches**
  - Size on MDS, lock caching, stat ahead, tape aggregation…

Mantra #5: Scalable Units - mandatory

- **Our size requires a Scalable Unit (SU) deployment philosophy for *all* resources – compute, file system, network, archive…**
  - We deploy in known SUs or widgets – building block style

- **“Same-etry” required for:**
  - Administration
  - Hardware repair/maintenance
  - Spares
  - Ease of expansion, upgrade…
  - Purchasing power



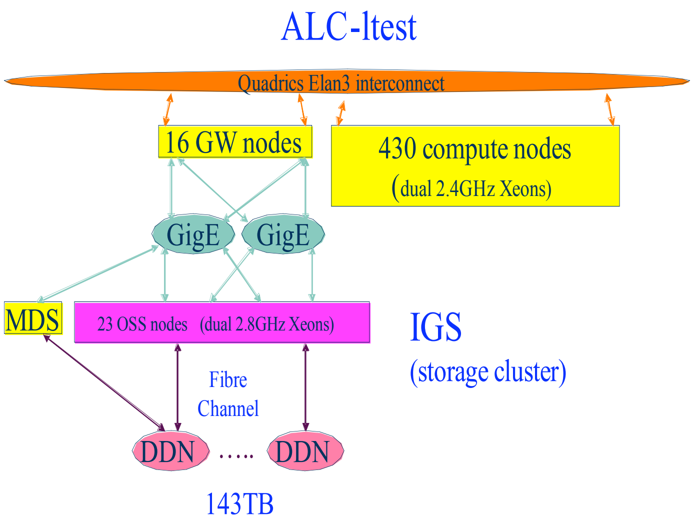
Example - *Storage* Scalable Unit requirements

<free bonus mantras>

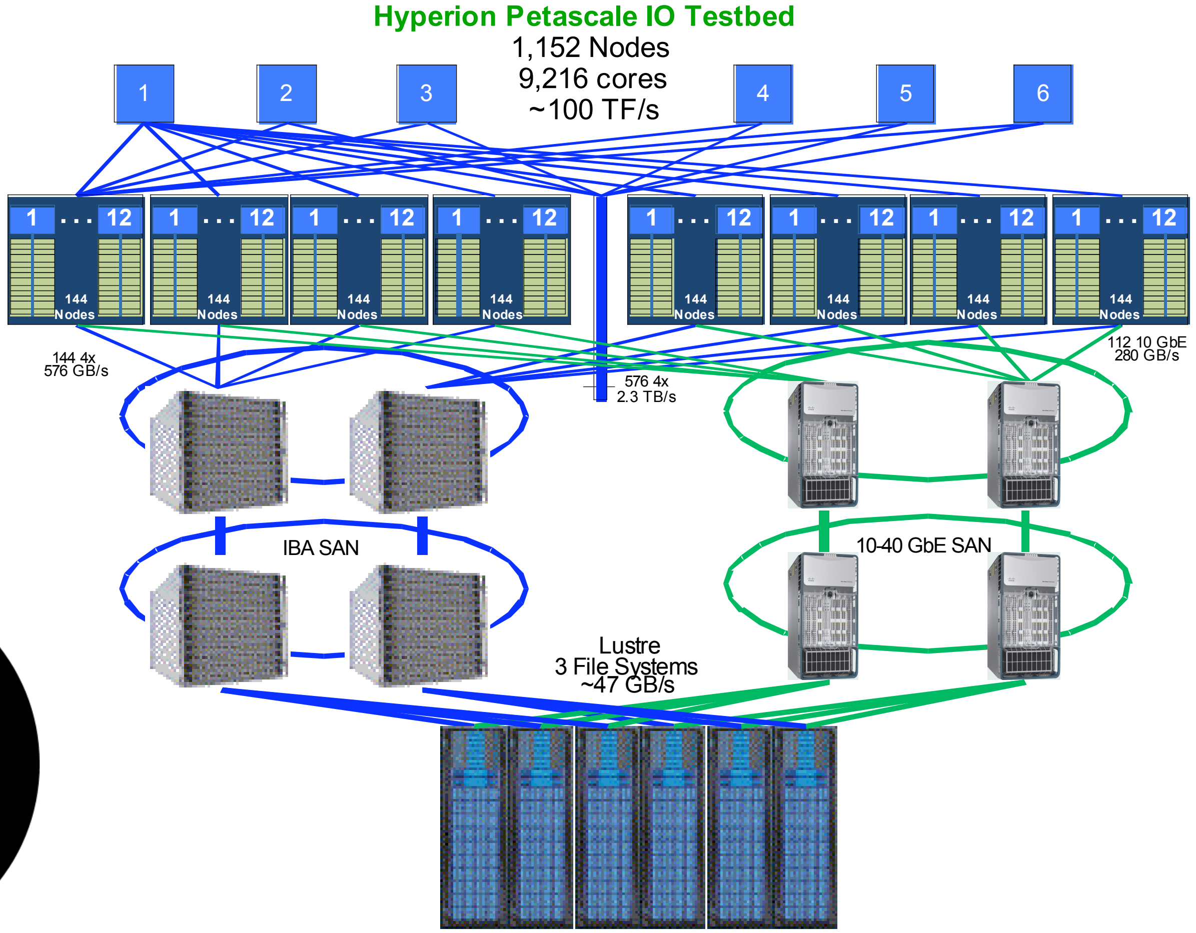
- High IOPs
- Parity on read
- n+2P
- Avoid "enclosure exposure"
- *No single component of SSU should deny production access to data* (redundant power, cooling, bridges…)
- Non-volatile caching or cache protection required
- Failover must be possible at reasonable granularity

Mantra #6: At-scale testing is mandatory

- **At these scales – you will be the first to encounter so much…**
- **As a result, we field a powerful test environment**
  - Particularly useful with Lustre testing
- **We combine this testing with**
  - Dedicated System Times (DSTs)
  - Pre-release discipline
  - Developers "eating their own dog food"
- **And for the petascale…**



Hyperion at-scale test environment – being built now



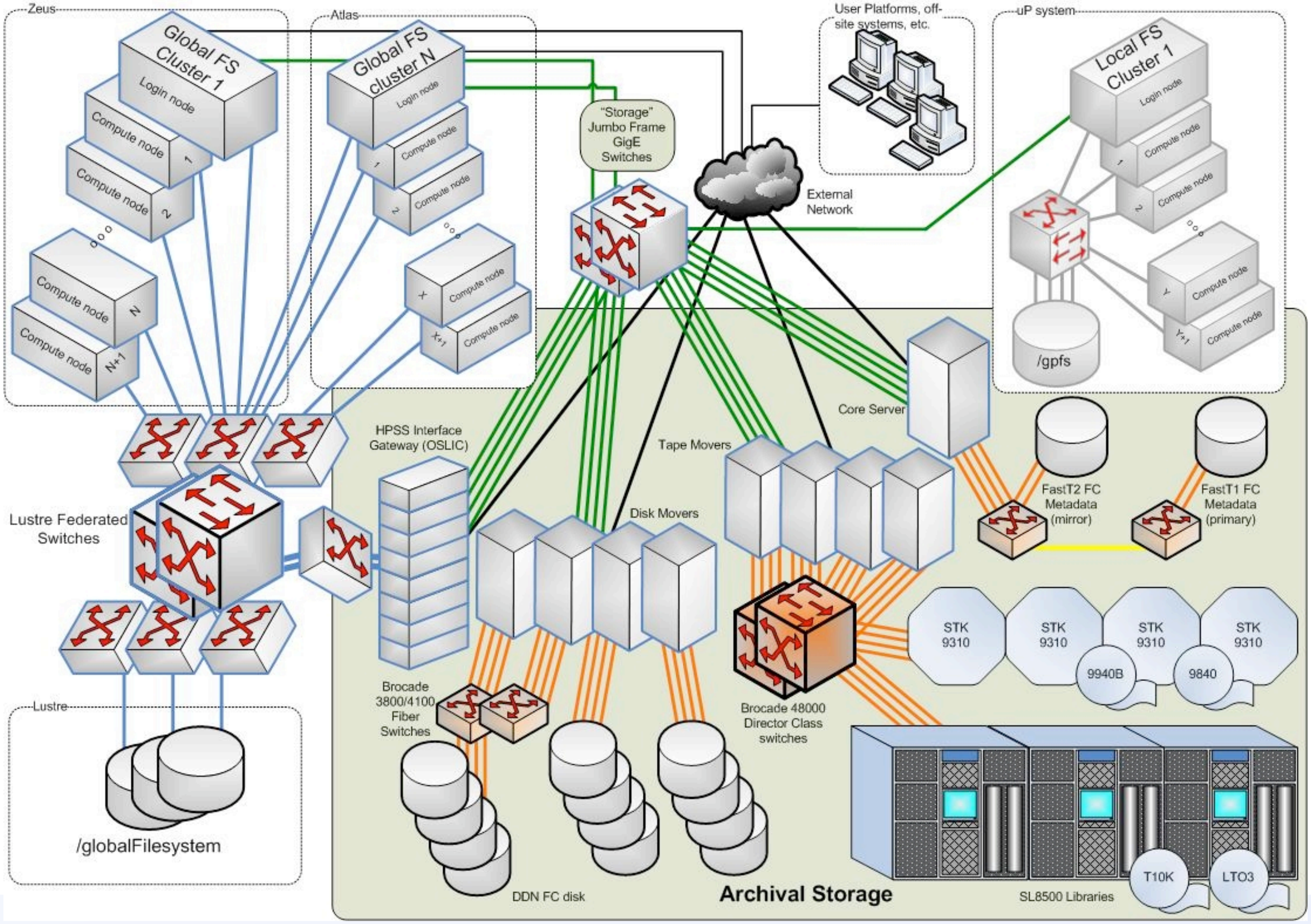
Mantra #7: Archives are forever (and not transparent!)

- **Archives should *not* be part of the file system**
  - Cost – the pain mantra
  - Archives protect billions of dollars of data investment
    - Risk-averse with independent schedule & requirements
    - Outlive vendors, platforms, OSs, file systems, users…
    - 41 years and counting at LLNL
  - Little issues like “find . –exec grep …”
- **But archives need intelligent, fast linkage to file systems**

<another free bonus mantra>

- **½” tape is and will remain king**
  - 6-54x cheaper than disk for purchase (capacity)
  - 300-700x cheaper than disk for power and cooling (energy cost)

Archive infrastructure – scaling for Sequoia



Mantra #8: Collaborate, collaborate, collaborate

- **Collaborate with your peers:**
  - ASC Tri-lab, ORNL, HPC sites
  - Open source!
- **Collaborate with industry**
  - HPSS, Lustre, Hyperion
- **Collaborate with academia**
  - ASC Alliances, GDO…
- **Collaborate/partner with your vendors**
  - Things will go wrong – often
  - When they do you *need* to have partnered with your vendor

Future petascale mantras???

- You really can’t store it all???
- On-platform checkpoint strategies = space savior???
- Post-process before you store???
- End-to-end checksums and encryption required???
- QoS – global resources can’t survive without it???

Summary

- Our terascale journey has provided us with many mantras
- The most powerful is that a well-balanced infrastructure is absolutely critical to the success of an HPC center.
- We are honoring these mantras on our path to the petascale.

What are your mantras? What will they be?