xyratex

Advancing Digital Storage Innovation

Rock-Hard Lustre
Trends in Scalability and Quality

nathan_rutman@xyratex.com

SCALE

Big Iron

- Lustre dominates the top end today

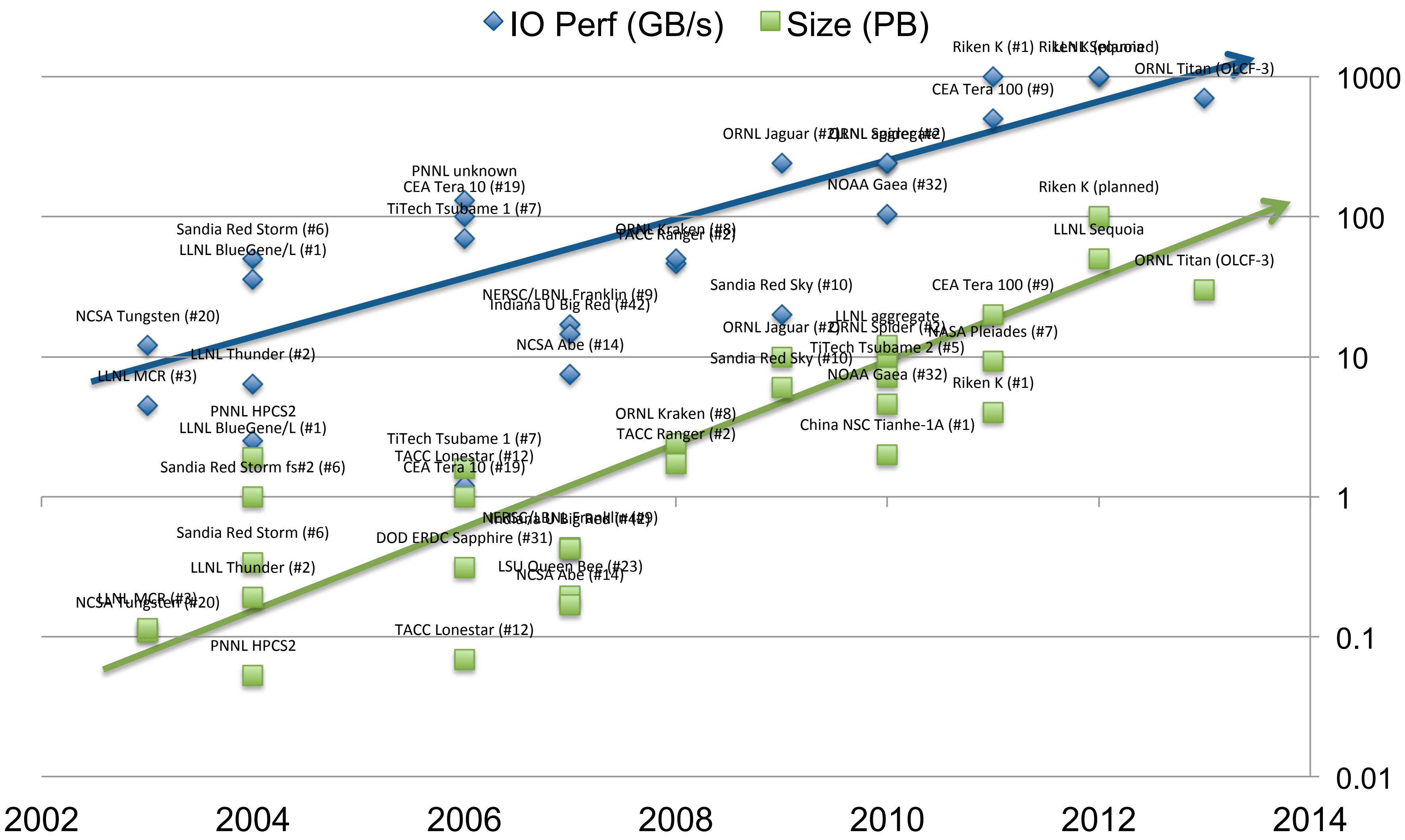
2010 numbers:

- 100% of the top 3
- 70% of the top 10
- 66% of the top 100
  - 59/100 Lustre + 8 suspected
  - 22/100 GPFS
  - 3/100 PanFS
  - 1/100 CXFS
  - 6/100 unknown, suspect non-Lustre

Lustre Installations

Rank	Machine	Storage	Speed	Clients	
1	Riken “K”	4PB	1TB/s	64,512	Fujitsu Exabyte File System
2	Tianhe-1A	2PB		7168	
3	ORNL Jaguar	10PB	240GB/s	18,688	Biggest U.S.
5	Tsubame 2	11PB		1408	
7	NASA Pleiades	5.1PB		11776	
9	CEA Tera 100	20PB	500GB/s	4324	Biggest EU, Lustre 2.0
Planned	Riken HPCI	30PB 90 OSS	720GB/s		
	Riken “K”	100PB 4000 OSS	1TB/s		upgrade
	LLNL Sequoia	50PB	500GB/s	98,304	expect 1TB/s

Lustre Systems Over Time



Continuing Systems Growth

- Biggest systems get bigger
  - Lustre scale barriers continue to fall
  - Nothing new
- Larger systems become more affordable
  - More customers need Lustre

New Scaling Features in Lustre 2.1

- Complete rework of MDS and Client IO stacks
- EXT4
  - 16TB file size (=stripe object limit)
  - 4B files
  - unlimited files/dir
  - faster fsck
    - skip uninitialized bitmaps
    - skip unused inodes
    - checksum group descriptors
  - faster mkfs
  - MMP
- MDRAID improvements
- Large LUN
  - 128TB LUNs

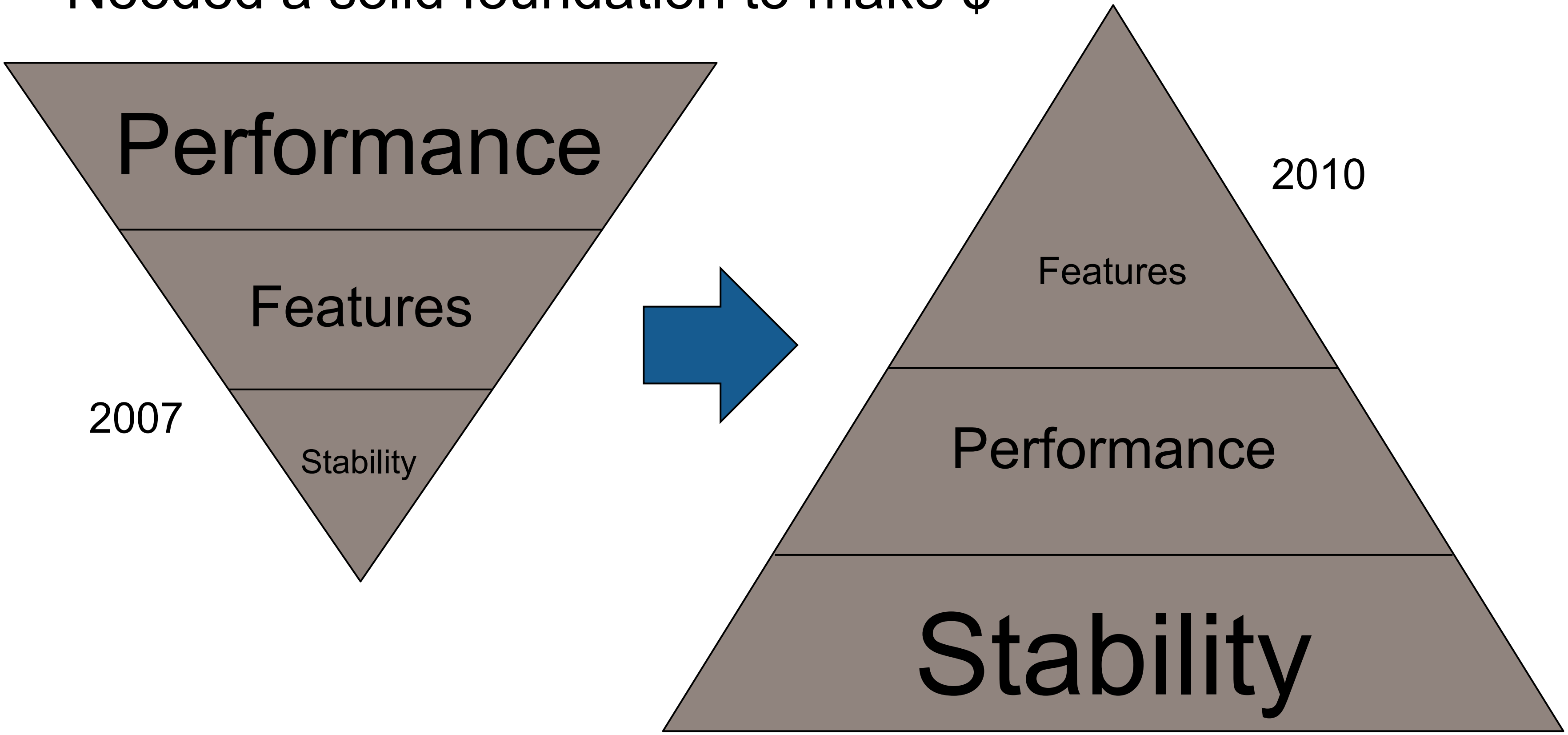
Future Community Scaling Projects

- Wide striping
    - 1350+ OSTs
- SMP scaling
    - Vastly improved MD rates
- Simplified SOM
    - ‘ls –l’
- Flash cache
- DNE

Quality

Lustre Quality

- Early perceptions of Lustre: “it’s a science project”
- Sun Microsystems purchased CFS in 2007
- Needed a solid foundation to make $



Performance

Features

2007

Stability

2010

Features

Performance

Stability

“Lustre Development” LUG 2010

Stability / Quality Improvements

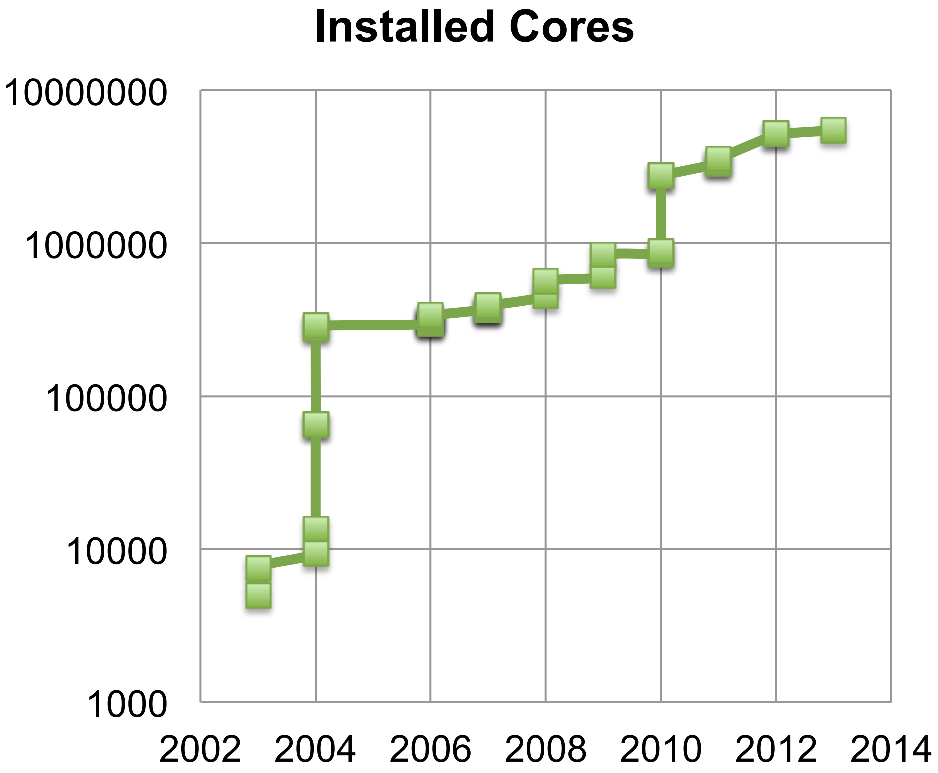
- Testing
  - LBATS, YALA automated build and test
  - Hyperion
  - feature-specific tests
  - scale tests
- Process
  - Short development cycles
  - Strict inspections
  - Branch gatekeepers
  - Train model
- Documentation
  - LID
  - Oak Ridge’s Lustre Internals
  - Subsystem map
  - Doxygen

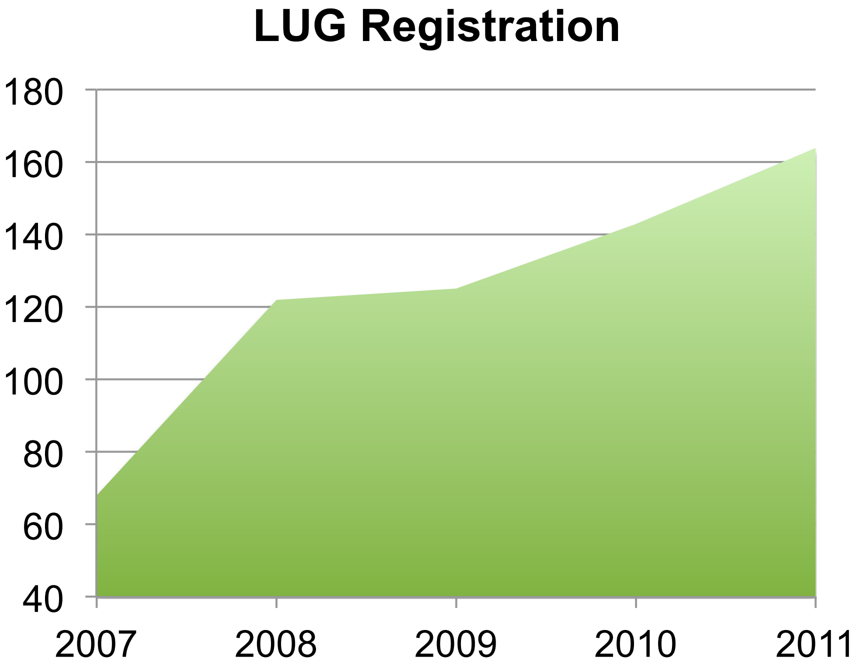
Post-Oracle era

- The quality focus remains post-Oracle
  - Xyratex, Whamcloud both follow the same quality processes
  - OpenSFS has added community inspectors to the SOW acceptance criteria
  - Community-based testing

- Recognition that the quality initiatives have paid off
  - Among developers first
  - Among users
  - Growing user base

Growth of Lustre Installations and Users

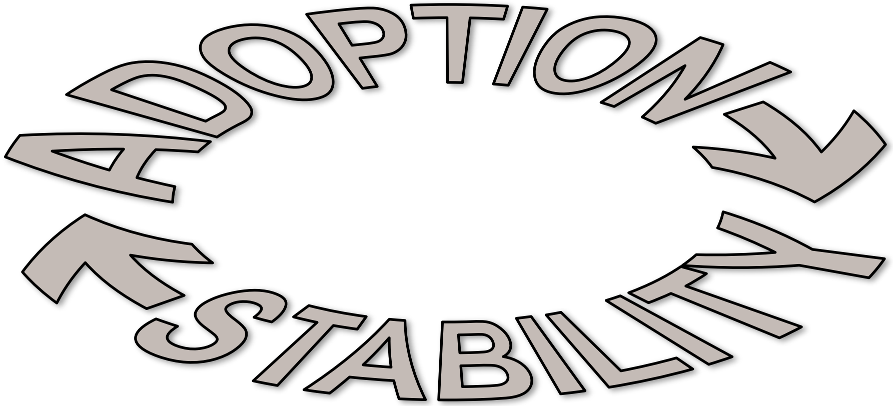




- Many significant contributors
  - Whamcloud, Xyratex, Cray, LLNL, ORNL, CEA, Bull, TACC, DDN

Positive Feedback Loop

- Two facts
  - –Top end hardware moves down
  - –Lustre is open source
- Imply two trends
  - –Community base will continue to broaden
    - Community contributions will increase
  - –Quality will continue to improve



Future Quality Features

- T10-DIF
  - Prevents server-to-drive corruption
- End-to-End Data Integrity
  - Prevents client-to-drive corruption, including network
- On-line LFSCK
  - Continuous verification and repair of metadata
- Imperative Recovery
  - Accelerate recovery for big systems

Improved High Availability in Integrated Solutions

- Integrated solutions like the ClusterStor 3000
  - Reduce the complexity
  - Encapsulate the HA
- Data access always provided for any single point failure
  - Switch, OSS controller, RAID, management server
- Eliminated controller-drive cabling
- Separate management network
- Integrated monitoring
- Integrated HA software

xyratex

Advancing Digital Storage Innovation

Thank You

nathan_rutman@xyratex.com