xyratex

Advancing Digital Storage Innovation

ClusterStor
xyratex

Increased Reliability of Large HPC Storage Deployments

Torben Kling Petersen, PhD

Principal Solution Architect, HPC

Xyratex Leading Provider of Data Storage Technology to OEMs

- **> 4,000 Petabytes of storage shipped in 2011**
- **Largest OEM Disk Storage System provider**

Enterprise Data Storage Solutions

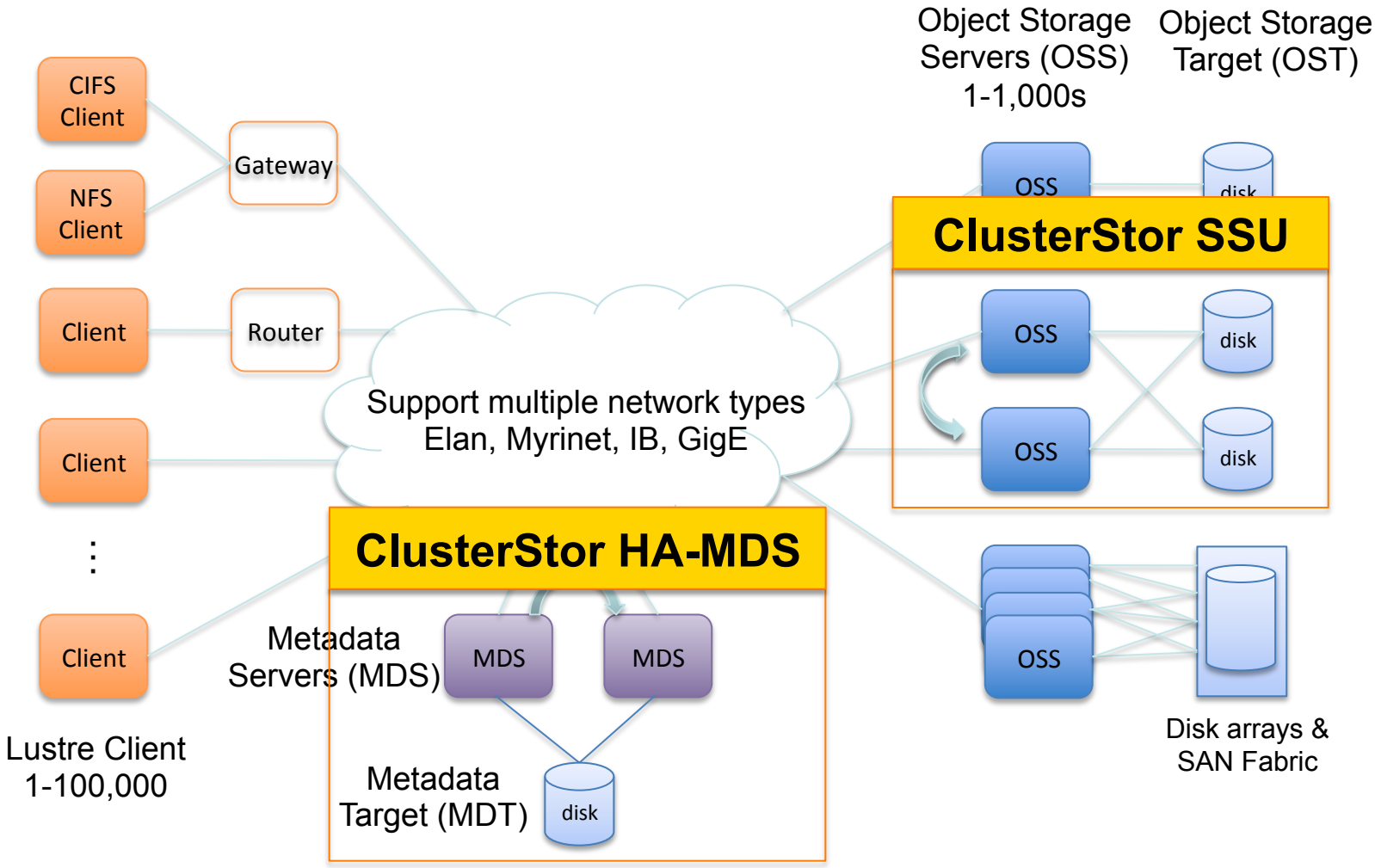
- **~ 50%** of w/w disk drives are produced utilizing Xyratex Technology*
- **Largest independent** supplier of Disk Drive Capital Equipment

HDD Capital Equipment Solutions

**Company estimates*

xyratex

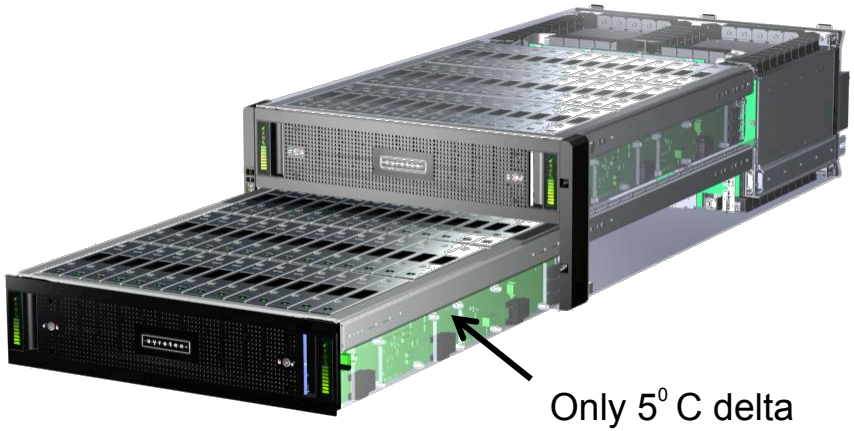
A Lustre Cluster

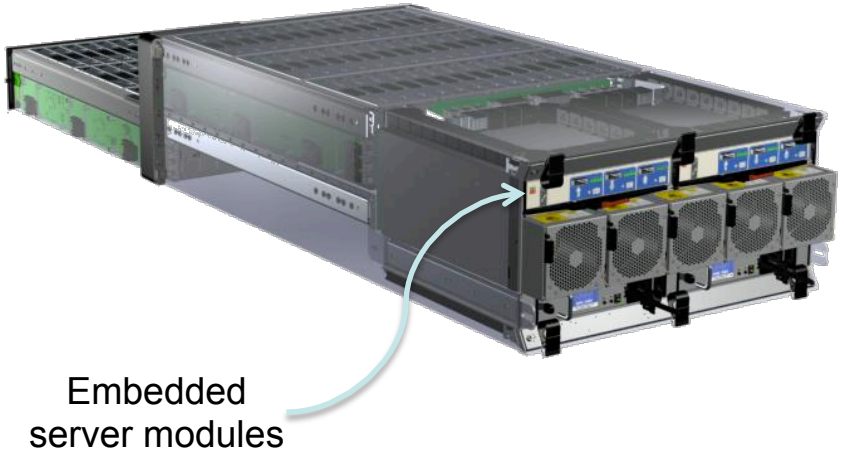


xyratex

CS-2584 - Scalable Storage Unit (SSU) – Lustre OSS

- **Ultra HD - CS-2584 SSU - OSS**
  - 5U84 Enclosure – completely H/A
    - Two (2) trays of 42 HDD's each
    - Dual-ported 3.5" FatSAS & SSD HDD Support
    - 150MB/s SAS available bandwidth per HDD
  - Pair of H/A Embedded Application Servers
    - CS-3000: =3.5GB/sec IOR over IB
  - IB QDR or 10GbE Network Link
  - Data Protection/Integrity (RAID 6, 8+2)
    - 2 OSS's per SSU
    - 4 OST's per OSS
  - 2x SSD OSS journal disks for increased performance
    - 2X Hot Spare HDD's
  - 64 Usable Data Disks per SSU
    - 1TB x 64 – 64TB usable per SSU
    - 2TB x 64 - 128TB usable per SSU
    - 3TB x 64 - 192TB usable per SSU
    - 4TB x 64 - 256TB usable per SSU





Xyratex ClusterStor (a.k.a. Sonexion) – Breaking new ground

- **Up to 1.8 PB/rack**
  - Up to 588 disks per rack
  - Supports 1, 2, 3 and 4TB SAS disks
- **More than 24 GB/s throughput per rack**
  - Lustre file system performance
- **Supports full QDR IB or 10 GbE fabrics**
- **All active components redundant and hot swappable**
- **Engineered, balanced solution for extreme density and performance**
- **Dedicated management utility**
- **Lustre 2.x based solution**
  - Active development of new features and fixes

ClusterStor
xyratex

xyratex

Advancing Digital Storage Innovation

“What do you mean, the file system is down ??”

“Again !!!”

Let's do the numbers – HAL 9000

- **Problem: Fly a large space craft to Saturn while preparing to kill all of it's astronauts and find the monolith ……***
- **Solution:**
  - Compute system capable of 10 PFLOPs
  - Storage capable of doing 10% of Compute -> 1 000 GB/s
  - Energy efficient
  - Incredible reliability (well, let's settle for decent ….)
  - Supportable for 3-5 years …

Throughput reqs (GB/s)	1000
Embedded Server	CS3000
SSU Performance (GB/s)	3
Volume requirements (TB)	300
Disk size (TB)	2
Rack size (42 or 48RU)	42
Power (SSUs) kW	2,08
SSUs per Rack (8 max)	8

	# SSUs	Total usable volume	Agg. throughput	IB Uplink ports	# Racks	# OSTs	# HHDs	Power reqs (kW)	Weight (T)	Floor space (m2)
Solution (performance)	334	42 752 TB	1002 GB/s	670	42	2 672	27 388	696,7	48,3	50,4
Solution (Full racks)	335	42 880 TB	1005 GB/s	672	42	2 680	27 470	698,8	48,3	50,4

* Thanks goes to A. C. Clarke for inspiration

Petascale Availability Simulation Results (HAL 8000)

Based on the current ClusterStor 3000 solution featuring:

- Lustre file system delivering 640 GB/s
- Usable volume: 26.8 PB
- 27 racks with a total of 17 280 nearline SAS 2 TB drives

Time period of interest	Number of simulations	Mean Availability (across 720 hours)	Instantaneous Availability (at 720 hours)
30 days (720 hours)	100	99.51%	98.00%
	100,000	99.56%	99.55%

Key Take-Aways from 30-day simulation:

- Monte Carlo analysis using Reliasoft BlockSim software
- Only 11 out of 17280 would fail (0.0636%)
- Probability of 1 or more OSTs rebuilding within a 5U/84 = 4.9607%
- Probability of 2 or more OSTs rebuilding within a 5U/84 = 0.1097%

So how do we get there ??

- **Testing of every component and the entire system is key**
  - Disk drives
  - Enclosures
  - Embedded server modules
  - All software
    - GEM (General Enclosure Management)
    - Linux/HA/MD-RAID/Software Components
    - Lustre
    - ClusterStor Manager (Scale-Out Management Solution)
  - Rack integration
  - Cabling
  - System Configuration tests
  - File system deployment tests
  - Client based testing
  - Soak testing of complete system

xyratex

Advancing Digital Storage Innovation

ClusterStor Factory Pre-Integration & Test

Scalable Storage Unit (SSU) Build / Configuration

- Tested Drives, Embedded Application Servers (EAS) and SSU build is received in the area

- The product is configured into a SSU, with the installation of the tested components and custom bezel

Integrated System Testing

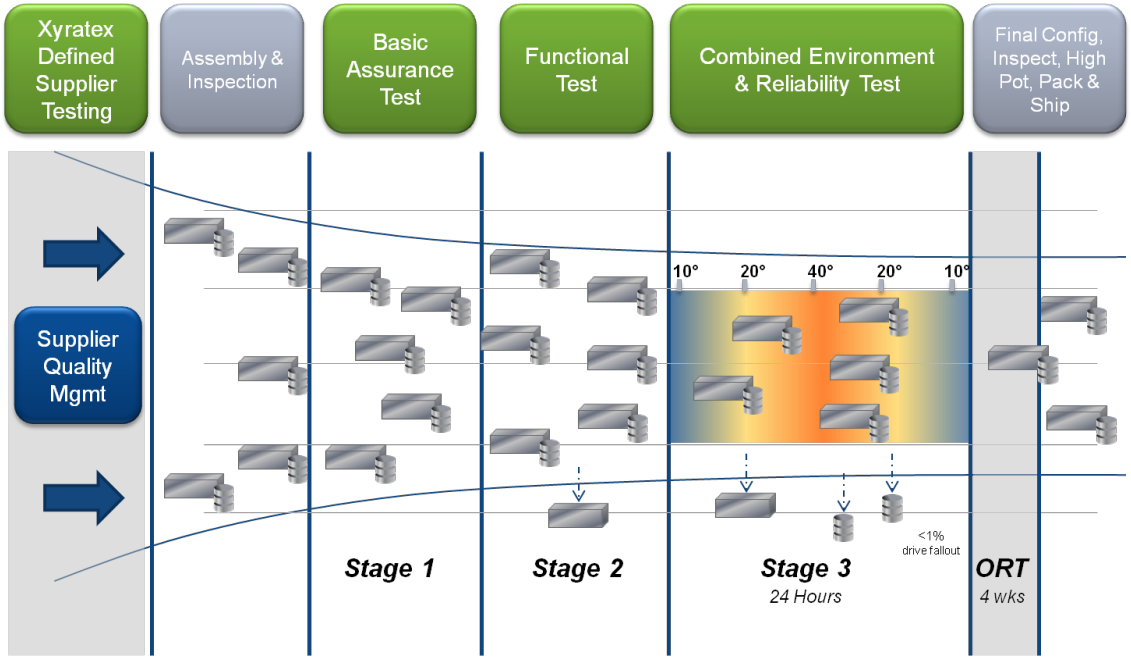
Integrated System Testing (IST) is a patented 3 Stage testing process embedded within manufacturing and designed to remove hidden quality problems

Features

Optimized 36 Hour Manufacturing & Test

Adaptable Test Automation

Standard Across the Globe



Benefits

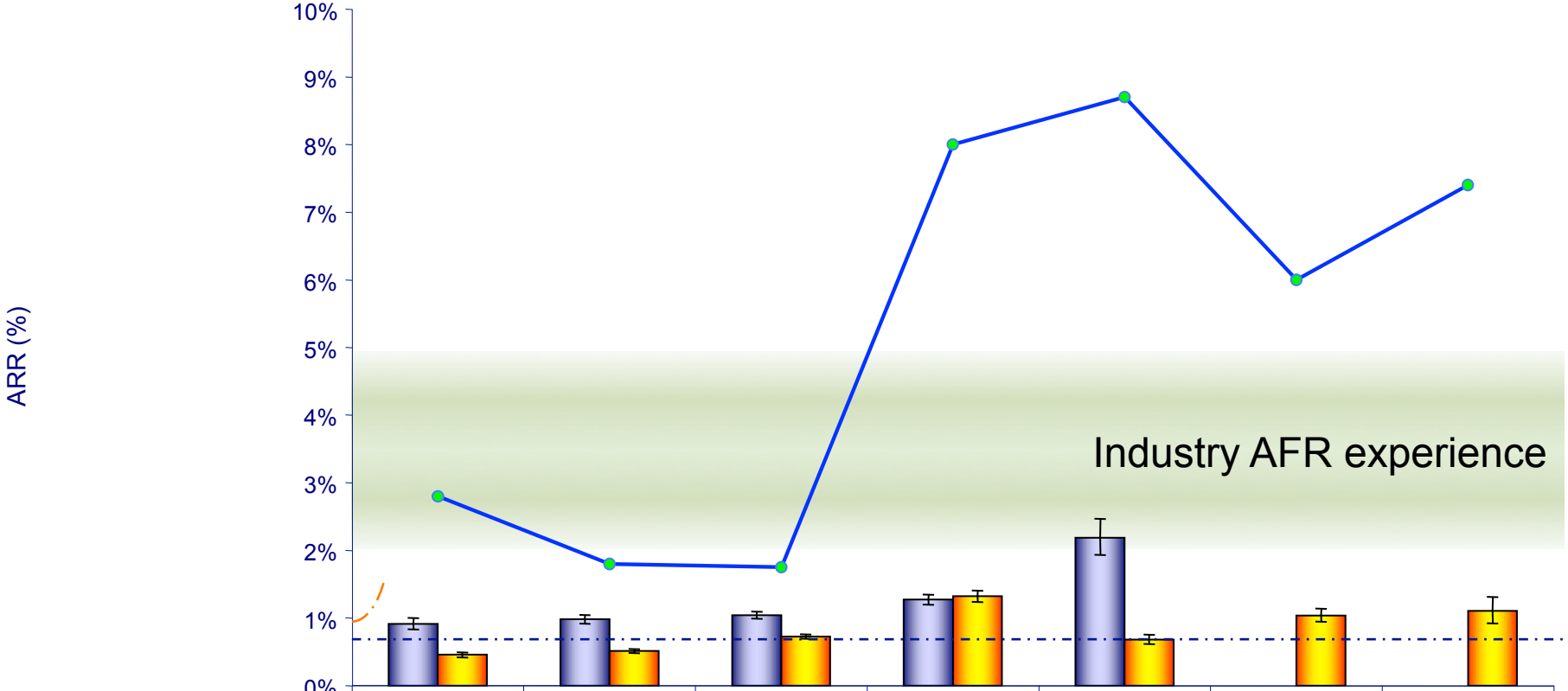
- Reduces solution warranty and service costs
- Reduces Infant Mortality
- Up to 1.5X drive reliability improvement over 3 Yrs.
  - AFR Reduction from ~9% to 2% or less*
  - 67% less disk drive failures in first 3 months
- Accelerates time to market

**Comparison of Google Study and Xyratex CERT*

Xyratex HDD Reliability : Failure Rate Comparison

Annual Failure Rate (AFR) by drive class

NetApp Study, 1.8M HDDs, 155K systems over 44 months, 99.99% reliability



	3 Months	6 Months	1 Year	2 Years	3 Years	4 Years	5 Years
XYR ATA FR	0.92%	0.98%	1.04%	1.27%	2.19%		
XYR Enterprise FR	0.46%	0.51%	0.73%	1.32%	0.68%	1.04%	1.10%
Google paper Base ~AFR	2.80%	1.80%	1.75%	8%	8.70%	6%	7.40%

xyratex

Rack Build / Integration

- All of the rack components are installed and cabled including MDS, SSU's, Network Switches, Management Switches and PDU's
- The assembled rack is installed and fastened into its final shipping crate.
- The shipping crate is positioned into to its test alcove

Test Alcove Infrastructure

- Each test alcove is powered with 4x 32A 3-phase sockets, internal and external IP access.
- Each alcove has a chilled water rear door attached and a transition frame to mate with the product within its crate.

Product Under Test

- Up to 30-day ‘Soak Test’
- Soak test measures:
  - I/O connectivity to (ClusterStor to Lustre clients)
  - I/O performance - read/write/ rewrite (ClusterStor)
- Tests a system with significant load extended over a significant period of time
- Includes "adverse" conditions testing (running HA scenarios for ClusterStor systems)

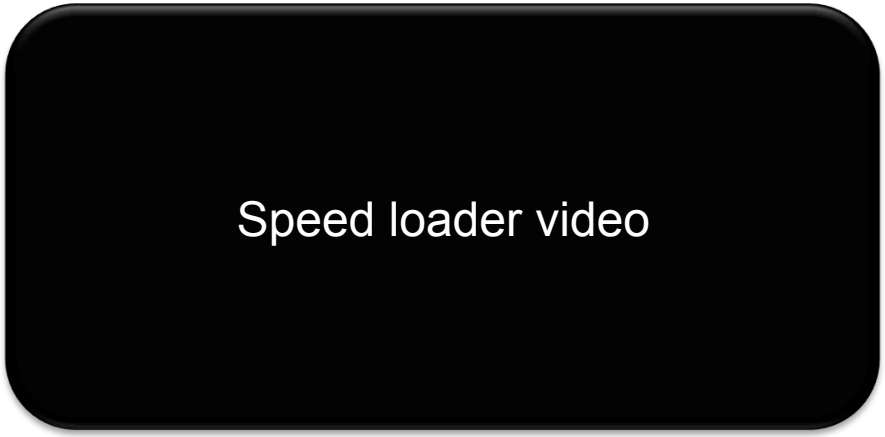
Simplified Installation – Hours vs. Days/Weeks

- **Xyratex delivers a complete ready-to-run ClusterStor solution**
  - Sizing and Configuration optimization
    - Performance centric
    - Capacity centric
  - Factory Integration & Staging
    - Rack integration & Cabling
    - Entire storage software stack factory pre-installed and pre-configured
    - System soak test and benchmark testing area at Xyratex factory
  - Drive speed-loader reduces drive insertion time by 85%

xyratex

Drive Installation / Unloading Process

- The drives are removed from the unit with the use of a speed loader.
- The speed loader allows the user to rapidly remove and install 7 drives at a time.
- The packaging and loader compliment each other, thus significantly reducing the handling time.



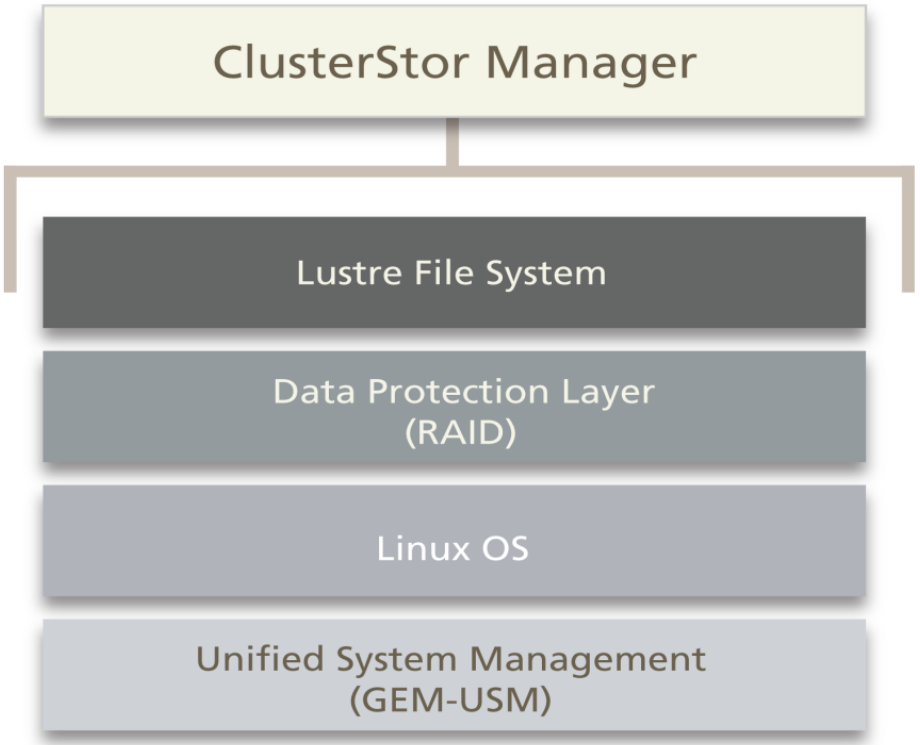
Ensuring Quality of Delivery & OOB Experience

Racks are reinforced with an additional 32 rivets to ensure quality!

ClusterStor Summary

- Architected
- Integrated
- Tested
- Optimized
- Qualified
- Supported

- Factory integration
- Component and system testing
- System shipped to site, not built on site
- Single owner of entire stack
- Global Support capability



xyratex

Advancing Digital Storage Innovation

Thank You - Questions?