Petascale Storage Solutions

2013 MSST

xyratex

Mike Feuerstein
Field Applications Engineer

Xyratex and HPC Storage

- 1994 Formed as MBO of IBM HDD capital test equipment business (1966)
- 1990s Expanded into HDD enclosure business for leading OEMs
- 2010 Acquires extensive Lustre® expertise: ClusterStor
- 2011 Largest OEM disk provider, >4,000 PB shipped
  - 50% of all disk drives w/w produced with Xyratex technology
- 2011 Introduces Lustre® HPC storage solution: ClusterStor 3000
  - Integrated, pre-configured, pre-cabled, linear scaling, high RAS,
- **2012 Proves CS at extreme scale: NCSA Blue Waters - Cray partnership**
  - CS-6000 introduced
- 2012 $1.1B in revenue; 26% of employees involved in R & D
- 2012 Patents
  - US: 149 71 pending Non-US: 98 52 pending
- 2013 Expands leadership role in the Lustre ® and HPC communities
  - Acquires Lustre® from Oracle: copyright, TM, engineers, support contracts
- 2013 ClusterStor receives Cloud Storage Excellence Award
- *2013 More Lustre ® solutions, plus Big Data Analytics, & Cloud*

Cray Sonexion system at NCSA

Total system throughput of 1.1 TB/s

Blue Waters Scale

Compute

- 237 Cray XE6 cabinets
- 32 Cray XK7 cabinets
- **25,766 clients**
- 1.5 PB memory
- Sustained Petaflop Computing
- 11.6 PF peak

Storage

- 25 PB total Lustre® storage on Cray Sonexion hardware
- 1.1 TB/sec total, **1.0 TB/sec** /scratch (**22 PB**)
- /scratch: 360 OSSs, 1440 OSTs => **14,400 HDDs**

Xyratex Petascale Solution Approach

- Software
  - Capable of performing at extreme scale
- Hardware
  - Capable of scaling but with efficiency and reliability required
- Management
  - Comprehensive view of every component of a petascale system
- RAS
  - Hardware, software, monitoring, with HA design & processes

Why Lustre®?

- At 10+ years old, still the fastest & most scalable file system for HPC
- Model for other petascale storage solutions
- POSIX compliant
- Runs on a large variety of hardware
- Un-matched scalability
  - 50,000+ clients >1 TB/sec bandwidth Billions of files
  - 31 PB max file size multi-PB file systems
- Active Community of Development, accelerating progress on a wider feature set
  - 12 contributors in 2.4 ~200K LOC (35K in 2.1)
  - Intel, Xyratex, EMC, CEA, IU, ORNL, LLNL…

Large Network I/O
Expanded use of flash storage
Wide striping and data placement
Large volume support for Lustre®
End-to-end integrity with T10-DIF
Data Migration, HSM

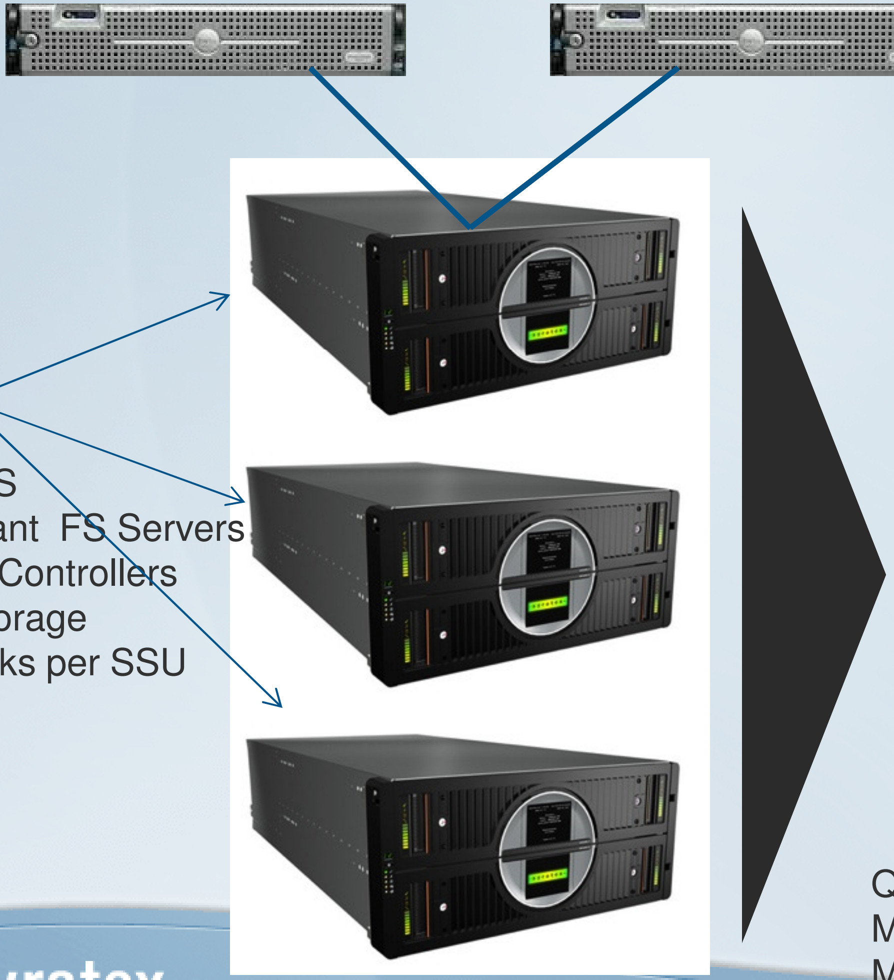
Distributed Metadata, MDS threading
Network Request Scheduler (NRS)
LNET Channel Bonding , IPV6
Increase Maximum file counts
Data Replication
Optimized CIFS, NFS exports

ClusterStor: H/W Scaling Complements Lustre®

Performance Density Enables Dynamic Scaling

Balanced Performance & Scalability

Clients



SSU
- OS
- Lustre FS
- Redundant FS Servers
- Storage Controllers
- RAID Storage
  - -- 84 disks per SSU

Network I/O Ports: 🟥

Compute & RAM: 🟨

Total HDDs: 🟩

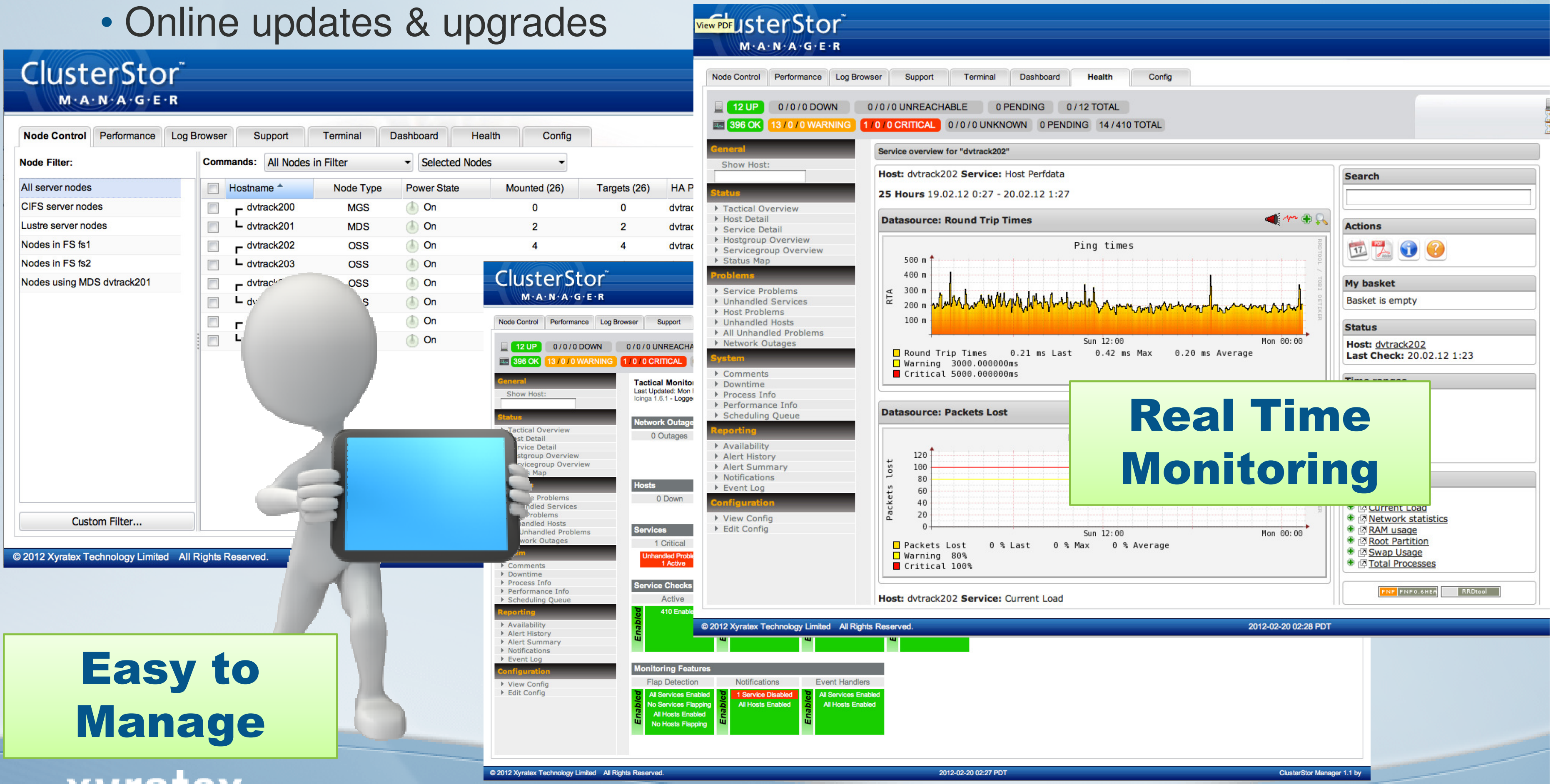
=

CS-6000
per rack
~**36-42 GB/sec**
File System
Throughput
up to **1.5 PB**
usable

QDR/FDR IB or 10/40GbE
MDS, MGS servers
Management Servers & Networks

ClusterStor Manager

- Fully Integrated End-to-End File System Visibility & Management
  - Low level diagnostics, embedded monitoring, logging, proactive alerts
  - Xyratex development and proven open source infrastructure components
  - Online updates & upgrades



Node Filter: All server nodes; CIFS server nodes; Lustre server nodes; Nodes in FS fs1; Nodes in FS fs2; Nodes using MDS dvtrack201

Commands: All Nodes in Filter | Selected Nodes

□	Hostname	Node Type	Power State	Mounted (26)	Targets (26)	HA P
□	dvtrack200	MGS	On	0	0	dvtrac
□	dvtrack201	MDS	On	2	2	dvtrac
□	dvtrack202	OSS	On	4	4	dvtrac
□	dvtrack203	OSS	On			
□	dvtrac	OSS	On			
□	dv	S	On			
□			On			
□			On			

© 2012 Xyratex Technology Limited All Rights Reserved.

12 UP | 0 / 0 / 0 DOWN | 0 / 0 / 0 UNREACHABLE | 0 PENDING | 0 / 12 TOTAL

396 OK | 13 / 0 / 0 WARNING | 1 / 0 / 0 CRITICAL | 0 / 0 / 0 UNKNOWN | 0 PENDING | 14 / 410 TOTAL

Service overview for "dvtrack202"

Host: dvtrack202 **Service:** Host Perfdata

25 Hours 19.02.12 0:27 - 20.02.12 1:27

Datasource: Round Trip Times

```
🟨 Round Trip Times     0.21 ms Last     0.42 ms Max     0.20 ms Average
🟨 Warning  3000.000000ms
🟥 Critical 5000.000000ms
```

Datasource: Packets Lost

```
🟨 Packets Lost    0 % Last     0 % Max     0 % Average
🟨 Warning  80%
🟥 Critical 100%
```

Host: dvtrack202 **Service:** Current Load

Search

Actions

My basket

Basket is empty

Status

Host: dvtrack202
Last Check: 20.02.12 1:23

© 2012 Xyratex Technology Limited All Rights Reserved. 2012-02-20 02:28 PDT

Tactical Monitor
Last Updated: Mon
Icinga 1.6.1 - Logge

© 2012 Xyratex Technology Limited All Rights Reserved. 2012-02-20 02:27 PDT ClusterStor Manager 1.1 by

Easy to Manage

Real Time Monitoring

Extensive Testing -> Reliability -> System Uptime

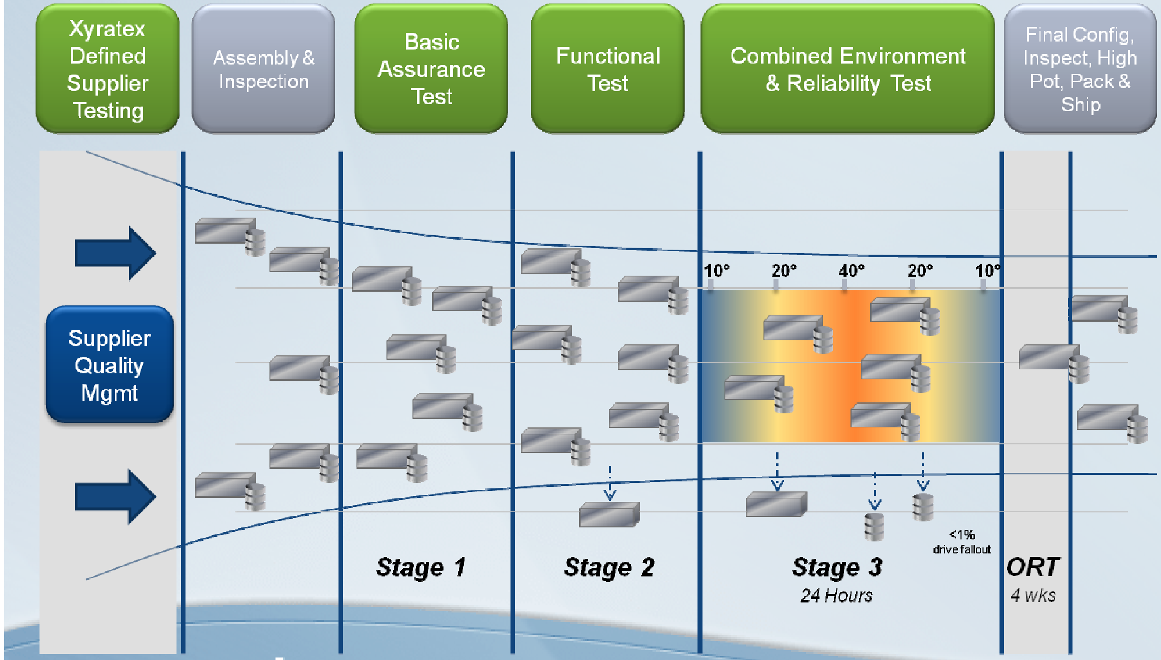
Integrated System Testing (IST) is a patented 3-Stage testing process embedded within manufacturing and designed to remove hidden quality problems

Features

Optimized 36 Hour Manufacturing & Test

Adaptable Test Automation

Standard Across the Globe



Xyratex Defined Supplier Testing

Assembly & Inspection

Basic Assurance Test

Functional Test

Combined Environment & Reliability Test

Final Config, Inspect, High Pot, Pack & Ship

Supplier Quality Mgmt

Benefits

- Reduces solution warranty and service costs
- Reduces Infant Mortality
- Up to 1.5X drive reliability improvement over 3 Yrs.
  - AFR Reduction to < 0.5%, *regardless of disk supplier*
  - 67% less disk drive failures in first 3 months
- Accelerates time to market

ClusterStor High Availability Lustre®

- Goals
  - Detect failures and architect to deal with *any* failure
  - Continuous access to data for applications
  - Multiple redundant components is the basis for Lustre® HA.
- Data Protection Layer
- Individual HA Domains
- HA Event Detection
- Automatic Failover
- Controlled Manual Failback
- Fabric Connectivity & Configuration for HA
- Factory Test & Integration

Scaling Issues & Solutions

- Efforts to scale uncovered problems not seen before
    - HA timings, routing, MDS performance, and more…
- Solution Highlights
    - Fixed Memory Allocation Race
    - Improved utilization of existing buffers and resized
    - Improved thread accounting
    - Improved Callback behavior
    - Fixed LNET for scale
        - Router buffer sizing, Network Priority
        - Unavailable router pass-through and dynamic re-routing
        - Fine grained routing: clients to routers, fs-specific routers

Benefits of BW

- Benefit to customers, Xyratex, entire Lustre community
- Demonstrated linear scaling of ClusterStor
  - Validated large scale integration approach
  - Maximum output per HDD minimizes footprint & power
  - Low HDD failure rate confirmed HDD testing approach
  - Back port strategy minimized risk of new releases
- Validated Lustre® 2.1 at scale
  - Increased understanding of LNET behavior at scale
  - MD operations @100K+ concurrent RPC requests
  - Improved HA timings
  - Identified areas of ongoing need

Thank You

mike_feuerstein@xyratex.com

http://www.xyratex.com

xyratex