Xyratex ClusterStor6000 & OneStor

Proseminar „Ein-/Ausgabe – Stand der Wissenschaft“

von Tim Reimer
Structure

• OneStor
  • OneStorSP
  • OneStorAP
  • "Green" Advancements
• ClusterStor6000
  • About
  • Scale-Out Storage Architecture
  • Software Architecture
  • Hardware Architecture
• Features & General Information
• Conclusion
OneStor - OneStor SP

- Xyratex OneStor SP-2584 delivers ultra dense storage capacity
- Petabytes of storage
- 3TB drives
- 6Gb/s I/O modules offers support for longer cable lengths
  → reduces cable complexity
- For enterprise-class applications
  - Big data, High Performance Computing (HPC)
OneStor - OneStor SP

- Includes Xyratex’s Unified System Management (USM)
  - USM embedded software is tightly coupled to OneStor hardware

- Ensures maximum availability
  - Comprehensive fault diagnosis, monitoring etc.
  - N+1 Power Cooling Modules
  - Dual I/O modules and dual data path to all drives
OneStor - OneStor SP

• Supports OEMs
  • Simplifies development and testing
  • Accelerates market introduction
  • Tailor brand requirements
  • Data protection features
OneStor - OneStor AP

- Xyratex OneStor AP-2584 delivers storage server building block
- For cloud computing
- Scale-out storage server architecture
- Application performance scales along with capacity increases
OneStor - OneStor AP

- Single or dual Embedded Server Modules (ESMs)
  - Server-level processing capabilities directly on-board
  - Colocated with OEM developed scale-out storage applications

- Unified Systems Management API (USM)
- For enterprise reliability, availability and serviceability
- OEMs can design management systems for their product line
- Help for market introduction
OneStor – "Green" Advancements

- Individual drive power control
- Advanced adaptive cooling technology
- "green" design meeting worldwide recycling requirements
- SP – 80+ % efficient power transformation
- AP – 92% efficient power transformation at 50% load
ClusterStor6000 - About

- "ClusterStor™ 6000 provides the ultimate integrated HPC data storage solution delivering optimized time to productivity"
  Xyratex about ClusterStor6000 * 2)

- Integrated Lustre storage solution
- Efficient petascale solutions for HPC applications
  - Scientific research, simulations etc.
- Linear performance scalability in less space
- Up to 1TB/s file system throughput
- Storage capacity up to tens of petabytes
- ClusterStor distributed by Cray as Sonexion
ClusterStor6000 – Scale-Out Storage Architecture

- Scale-out Storage Architecture combined with the Lustre file system delivers
  - Simplified system installation and operation
  - Optimized HPC performance
  - Not disturbing cluster expansion
Scale-Out Storage Architecture

Traditional storage systems:

- Made of unequal building blocks
  - Servers to run file system and software
  - High-speed storage interconnect
  - A RAID controller
  - High-density storage systems housing the disk
- Each subsystem adds complexity and potential bottlenecks
## Scale-Out Storage Architecture

### Storage subsystem

<table>
<thead>
<tr>
<th>Potential Bottlenecks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under-provisioned networks</td>
</tr>
<tr>
<td>Unbalanced fabrics</td>
</tr>
<tr>
<td>SD or DDR InfiniBand</td>
</tr>
<tr>
<td>Gigabit Ethernet</td>
</tr>
</tbody>
</table>

### Interconnect fabric

- IB/10GbE/FC ...

### File system server(s)

- Old and slow systems
- Lack of memory
- Too few servers for the underlying storage system

### RAID controller(s)

- Too many disks behind each controller
- Slow disk connectivity
  - (3Gb SAS, 4Gb FC, SATA)

### Disk system

- Too many for each expander
- Too little bandwidth available to each drive
- SAS dongle
- SATA drives

Bottlenecks named by Xyratex * 3)
Scale-Out Storage Architecture

- Consolidated hardware and software environment
- For simple Lustre scalability
- Integrated Scalable Storage Unit (SSU)
  - Each supports two industry-standard x86 ESMs
  - Common midplane to all drives in the SSU
    → High-speed interconnect
- ESMs can run industry-standard Linux distributions
Scale-Out Storage Architecture

Traditional HPC storage scaling

ClusterStor HPC storage scaling

Xyratex' Comparison: traditional scaling – ClusterStor scaling * 3)
ClusterStor6000 – Software Architecture

ClusterStor Manager

Lustre File System

Data Protection Layer (RAID)

System Management Software (GEM)

Linux OS

Multi-layer software stack * 3)
Software Architecture

ClusterStor Manager

- Single-pane-of-glass view of infrastructure
- Browser based GUI simplifies cluster installation
- GUI can be used to manage the storage environment:
  - Start and stop file systems
  - Manage Mode failover
  - Monitor node status
  - Collect and browse performance data
- Dashboard reports errors and provides system snapshots
Software Architecture
Software Architecture

ClusterStor Manager – Health

ClusterStor Manager 1.1 by
© 2012 Xyratex Technology Limited. All Rights Reserved.
Software Architecture

Data Protection Layer (RAID)

- RAID 6 array to protect against double disk failures
- 8 + 2 RAID sets support hot spares
  - when disk fails data rebuilds on a spare disk
- Write intent bitmaps (WIBS) to aid the recovery of RAID parity data
- WIBS reduces parity recovery time from hours to seconds
Software Architecture

Unified System Management Software (GEM-USM)

• Runs on each ESM in the SSU
• Monitors and controls SSU's hardware infrastructure
• Key features
  • Management system health
  • Power control of hardware subsystems
  • Monitoring of status
  • Efficient adaptive cooling
  • Extensive event capture for post failure analysis
Software Architecture – Lustre

Highlights of the Lustre File System

- Server based architecture for large-scale computing
- Powering world's top HPC clusters
- Petabytes of storage, hundreds of GB/s of I/O throughput
- Lustre cluster is an integrated set of servers that
  - process metadata
  - store data objects
  - manage free space
  - present file systems to clients
Software Architecture – Lustre

Lustre cluster components

• Management Server (MGS)
  • Lustre servers contact MGS to provide Information
  • Lustre clients contact MGS to retrieve Information

• Metadata Server (MDS)
  • Makes metadata from Metadata Target (MDT) available to Lustre clients
  • MDT stores metadata on disk

• Object Storage Server (OSS)
  • provides file I/O service for Object Storage Targets (OSTs)
ClusterStor6000 – Hardware Architecture

The principal hardware components:

- Cluster Management Unit
- Scalable Storage Unit
- Network Fabric Switches
- Management Switch
Hardware Architecture

Cluster Management Unit (CMU)

- ClusterStor Manager – central point of management
- MDS – storing file system metadata
- MGS – manages network request handling
Hardware Architecture

Scalable Storage Unit (SSU)

- Hosts two OSS nodes
- Contains two ESMs
- Can directly access all drives
- If ESM fails the other one manages its OSTs
- Else I/O is balanced
Hardware Architecture

Network Fabric Switches
- Manages I/O traffic
- ESMs connected to several network switches → maximize network reliability
- IB or 10GbE or 40GbE

Management Switch
- Consists of local network used for configuration management
- Enables the ClusterStor Manager to power-cycle the ESMs
- 1GbE
## Features & General Information

<table>
<thead>
<tr>
<th>Product</th>
<th>Maximum System Configuration</th>
<th>Maximum throughput</th>
<th>Usable File System Capacity</th>
<th>File System Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>OneStor</td>
<td>4 enclosures max. 336 drives</td>
<td>14.4 GB/s</td>
<td>Over 2 PB</td>
<td></td>
</tr>
<tr>
<td>ClusterStor 6000</td>
<td>7 SSUs with a max. 588 drives</td>
<td>1 TB/s</td>
<td>Up to 93.4 PB</td>
<td>42 GB/s per rack sustained reads and writes</td>
</tr>
</tbody>
</table>
Features & General Information

OneStor Enclosure

93 cm
22 cm
48 cm

Weight: 128 Kg
Features & General Information

ClusterStor6000 rack * 6)

- Height: 1.99 m
- Depth: 1.2 m
- Width: 60 cm
- Weight: 1,141 Kg
Conclusion

• OneStor
  • Maximum availability
  • Helpful for OEMs
  • "green" advancements
• OneStor SP – enterprise class applications, e.g HPC
• OneStor AP – cloud computing, big data analytics
• ClusterStor6000
  • Scale-Out Storage architecture – performance
  • Lustre – simplifies
  • Software – overview, Bb GUI simplifies interaction, GEM
  • Hardware – CMU, SSU, MS
• Features – PB of storage, high throughput

"Data storage is our business. Innovation is our passion."

Xyratex slogan  * 4)
Sources

1) http://www.xyratex.com/products/onestor-sp-2584
2) http://www.xyratex.com/products/clusterstor-6000
4) http://www.xyratex.com/
5) http://www.xyratex.com/products/onestor-ap-2584
7) http://www.xyratex.com/sites/default/files/files/field_inline_files/OneStor_SP2584_DS_1-0_0.pdf
8) http://www.ecmwf.int/newsevents/meetings/workshops/2012/high_performance_computing_15th/Presentations/pdf/Kling_Petersen.pdf