Towards Storage Solutions for Petaflops Clusters

Toine Beckers
tbeckers@ddn.com

ZKI-ASC@UH
5.3.2010
DDN = HPC

- DDN provides more bandwidth to the top500 list than all other vendors combined!
- 8 out of Top10 systems choose DDN
- 45 out of Top100
- 5 systems over 120GB/s
- Mix of applications:
  - Government/University
  - Defense/Intelligence
  - Oil Exploration
  - Product Design
  - Archival, Backup

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NNSA/LANL United States</td>
<td>Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8, 3.2 GHz Opteron DC 1.8 GHz, Voltaire Infiniband IBM</td>
</tr>
<tr>
<td>2</td>
<td>Oak Ridge National Laboratory United States</td>
<td>Jaguar - Cray XT5 QC 2.3 GHz</td>
</tr>
<tr>
<td>3</td>
<td>NASA/Ames Research Center/NAS United States</td>
<td>Plaides - SGI Altix ICE 6200EX, Xeon QC 3.0/2.66 GHz SGI</td>
</tr>
<tr>
<td>4</td>
<td>DOE/NNSA/LLNL United States</td>
<td>BlueGene/L - eServer Blue Gene Solution IBM</td>
</tr>
<tr>
<td>5</td>
<td>Argonne National Laboratory United States</td>
<td>Blue Gene/P Solution IBM</td>
</tr>
<tr>
<td>6</td>
<td>Texas Advanced Computing Center/Univ. of Texas United States</td>
<td>Ranger - SunBlade x6420, Opteron QC 2.3 Ghz, InfiniBand Sun Microsystems</td>
</tr>
<tr>
<td>7</td>
<td>NERSC/LBNL United States</td>
<td>Franklin - Cray XT4 QuadCore 2.3 GHz Cray Inc.</td>
</tr>
<tr>
<td>8</td>
<td>Oak Ridge National Laboratory United States</td>
<td>Jaguar - Cray XT4 QuadCore 2.1 GHz Cray Inc.</td>
</tr>
<tr>
<td>9</td>
<td>NNSA/Sandia National Laboratories United States</td>
<td>Red Storm - Sandia/ Cray Red Storm, XT3/4, 2.4/2.2 GHz Cray Inc.</td>
</tr>
<tr>
<td>10</td>
<td>Shanghai Supercomputer Center China</td>
<td>Dawning 5000A - Dawning 5000A, QC Opteron 1.9 Ghz, Infiniband, HPC 2008 Dawning</td>
</tr>
</tbody>
</table>
S2A9900

Purpose-Built

High Performance Streaming & Archiving Storage Platform
S2A Design Architecture

Low Latency - High Performance, Silicon Based Storage Appliance

• Parallel Access For Hosts
• Parallel Access To A Large Number Of Disk Drives
• Quality Of Service
• Scalability
• Drive Error Recovery In Real Time
• True State Machine Control
  – 10 Virtex 4 FPGAs, 16 Intel embedded processors, 8 Data FPGAs
An Implementation of Parallelism w/ Double Parity RAID Protection

- Double Disk Failure Protection
- LUNs can span tiers
- All ports access all storage
- Reed-Solomon Code Implemented in a Hardware State Machine
  - No penalty for RAID 6!
- Parity Computed On Writes AND Reads
- No loss of performance on any failure
- Multi-Tier Storage Support, SSD, SAS, SATA Disks
- Up to 1200 disks total
  - 960 formattable disks

2 x 10 SAS Loops to Disks

Tier 1
Tier 2
Tier 3

8 FC-8 and/or 4 IB 4X Parallel Host Ports

RAID 6, 8+2 Byte Stripe

RAID 0
Single Enclosure Storage Tiering

Put data on the Most Effective Medium!

SATA
Cost Effective
Capacity

SAS
High Performance
Simple, Reliable Configuration

Direct Connection and RAID Striping Provides Maximum Data Availability

- Direct cabling avoids daisy chaining
- Data is striped across channels/enclosures
- Drive Channels are RAIDed 8+2
- Drive Enclosures are RAIDed 8+2

Only DDN Enclosure RAIDing can withstand the loss of 20% of system enclosures & drives while delivering full data availability!!
Data Corruption Error Handling

Host Data Striping

First step isolates error

Second step corrects error

FPGA

Cache

Protocol/PHY

Disks

Protocol/PHY: SCSI, FC - or - SAS
The data is flushed to the disk and the disks have now correct data on channel F. The cache has been repaired by the FPGA using the parity information.
SATAssure Data Integrity

• The Parallel Data Recovery Engine allows data reconstruction and integrity checking
  – S2A hardware enables SATAssure software to verify all data read from the disks (no silent data corruption)
  – S2A hardware allows SATAssure to send hosts “fixed” data (data integrity is assured)
  – S2A hardware enables SATAssure to correct data on the disk for future accesses (self-healing array)
  – Multiple levels of disk recovery attempted before failing drives (replace fewer drives)
  – S2A controller journaling allows partial rebuilds (less time in degraded mode)
Data Center Efficiency

• Leading Power Efficiency
  - Only 4 x 30A 220V Drops per 600TB
  - 1/4th the components of competing storage: less power supplies, fans
  - Additional D-MAID reduces consumption even further

• Truly Green Storage!

• S2A D-MAID™
  - Intelligent Power Management
  - Optimized for Backup/VTL/Archive
  - Spin Down Tiers of Inactive Drives
    - 12 seconds to spin up
    - No Application Change

<table>
<thead>
<tr>
<th></th>
<th>Active</th>
<th>Dynamic MAID*</th>
</tr>
</thead>
<tbody>
<tr>
<td>300TB (300 x 1TB SATA)</td>
<td>7.1 kW</td>
<td>4.5 kW</td>
</tr>
<tr>
<td>600TB (600 x 1TB SATA)</td>
<td>13.5 kW</td>
<td>8.29 kW</td>
</tr>
<tr>
<td>1.2PB (1200 x 1TB SATA)</td>
<td>26.1 kW</td>
<td>15.8 kW</td>
</tr>
</tbody>
</table>

1.2PB Dynamic-MAID Savings: Up to $36,000/yr

* S2A D-MAID Savings results assume 80% data dormancy for online archive, $0.20 kWhr
Worst Case Recovery

• Disks can become completely unresponsive to all commands
  – The internal OS can enter a loop that does not enable external commands
  – A power cycle always recovers the drive
  – S2A 9900 automatically power cycles a drive in place

• Drives that issue SMART warnings or grow defects at an increasing rate are copied to spare drives
S2A9900 Configuration with 10 Enclosures

- 8 FC-8 and/or 4 IB 4X Parallel Host Ports
- 2 x 10 SAS Loops to Disks

Tier 1: Enclosures 1-10
Tier 2: Enclosures 11-20
Tier 3: Enclosures 21-30
Tier 60: Enclosures 31-40
S2A9900 Configuration with 5 Enclosures

8 FC-8 and/or
4 IB 4X
Parallel Host Ports

2 x 10 SAS Loops to Disks

Tier 1
Tier 2
Tier 3
Tier 30

Enclosure 1
Enclosure 2
Enclosure 3
Enclosure 4
Enclosure 5

8 FC-8 and/or
4 IB 4X
Parallel Host Ports

2 x 10 SAS Loops to Disks

Tier 1
Tier 2
Tier 3
Tier 30
S2A9900 Configuration with 3 Enclosures

2 x 10 SAS Loops to Disks

Tier 1
Tier 2
Tier 3
Tier 15

8 FC-8 and/or 4 IB 4X Parallel Host Ports

Enclosure 1
Enclosure 2
Enclosure 3

Tier 15

DataDirect
EXTREME
STORAGE
Scalability & Density

- **Simple Cabling**: All Enclosures are direct connected (up to 10 enclosures) to the S2A Appliances for easy configuration and maximum reliability.
- **Maximum Availability**: S2A Storage Systems can lose up to 20% of the available drive enclosures without impacting host performance or data availability.

### The World Scalability & Density Leaders

<table>
<thead>
<tr>
<th>Enclosures</th>
<th>Height</th>
<th>Storage Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Enclosures</td>
<td>16U</td>
<td>Up to 300TB</td>
</tr>
<tr>
<td>5 Enclosures</td>
<td>24U: 1/2 Rack</td>
<td>Up to 600TB</td>
</tr>
<tr>
<td>10 Enclosures</td>
<td>44U: 1 Rack</td>
<td>Up to 1.2PB</td>
</tr>
<tr>
<td>20 Enclosures</td>
<td>84U: 2 Racks</td>
<td>Up to 2.4PB</td>
</tr>
</tbody>
</table>

- Up to 150 Drives
- Up to 600 Drives
- Up to 600 Drives
- Up to 1,200 Drives
S2A6620

Entry Level Mixed Workload Platform
StorageScaler

Begin – Standalone RAID

Upgrade to…

S2A9900
Replace SBB RAID Modules with SBB SAS Expander Modules

S2A6620
Replace SBB RAID Modules with SBB Storage Server Modules
Up to 350,000 Cache IOPS, 30,000 Disk IOPS
4 x Active/Active Host Ports: FC4, FC8
Scales to Support 120 Hard Drives in 8U
Up to 2.0 GB/s Performance
Mix SSD, SAS + SATA For Storage Tiering
Up to 11 Systems (660 TB) per Rack
RAID 5 and RAID 6 Options
Journaled Fast Drive Rebuild
Active/Active Storage Managers with Failover
Full SATAssure Data Protection
Windows MultiPath Support
Cache Mirroring
SFA

Storage Fusion Architecture
The Next Generation
2010+ Petaflop Systems

- **LLNL**
  - 1TB/sec and 30PB (Lustre)
- **Argonne**
  - 500GB/sec and 60PB (GPFS, PVFS)
- **ORNL**
  - 800GB/sec and 30PB (Lustre)
- **CEA**
  - 500GB/sec (Lustre)
The Case for Change

• CPU technology has moved from increasing speeds to increasing compute cores
  – The latest CPUs, with Hyperthreading, can run 8 threads simultaneously, or 16 threads from a dual-socket server!

• Single computers can now send multiple simultaneous I/O requests to the storage system
  – This looks like transactional or random I/O to the storage
  – Lots of file operations mean lots of file system metadata operations

• New Storage systems must be able to perform well at both high throughput (sequential) and high IOPS (transactional) workloads
SFA10000 Features

Highly Parallelized SFA Storage Processing Engine

- Active/Active Design
- 10GB/s Read & Write Speed
- 1 Million Burst IOPS
- 300K Random Disk IOPS
- 16GB Battery-Backed, Mirrored Cache
- RAID Levels 1, 5 and 6
- Intelligent Block Striping
- SATAssure Data Protection
- GUI, SNMP, CLI
- 16 x FC-8 ports or 8 x QDR-IB ports

Up to 1200 SAS, SATA or SSD Drives with full redundant paths
SFA Embedded: The Vision

- Embed Storage Intensive Applications within the Storage Controller
  - Reduce complexity, infrastructure and administration
  - Reduce cost as well as lower operational cost
  - Increase performance for latency sensitive applications
IO Path Acceleration

Storage Fusion Architecture shortens the IO path from the application to storage, reducing latency and increasing IOPS performance.
Example: Implementing Lustre Today

Infrastructure:
- Multiple OSS servers
- 2 Lustre MDS servers
- 1 Lustre MGS server
- Multiple RAID Arrays for OSTs
- RAID Array for MDTs
- SAN Switching
- Multiple Racks of equipment requiring power, cooling and floor space
EXAScaler HPC Storage on the SFA10000E Appliance

SFA10000E with embedded EXAScaler can result in a 10 to 1 or greater reduction in managed systems.

Storage Fusion Architecture not only reduces complexity, it streamlines IO by reducing latency and protocol conversions.
Multi-Platform Architecture

Block Storage Array
- SFA10000
- Block Storage Target

Clustered Filer
- SFA10000E
- DDN File Storage
- EXAScaler
- GRIDScaler

Open Appliance
- SFA10000E
- Customer Applications
- SFA10000E
- Embedded Storage Server

Flexible Deployment Options: 3 System Modalities
- Block Storage Array
- Clustered Filer
- Open Appliance
Scaling Performance with the SFA10000E Storage Appliance

GridScaler & ExaScaler Clients
IB or 10Gig-E

4.8PB

Integrate multiple appliances to scale to over 100's of GB/s and 10's of Petabytes

Add additional SFA Couplets to Linearly Scale Performance

GridScaler & ExaScaler Servers Embedded

300-600 3.5” Disk Drives

SFA10000E

5GB/s
10GB/s
15GB/s
20GB/s

1.2PB
2.4PB
3.6PB
4.8PB
Thank You

Toine Beckers
tbeckers@ddn.com