<table>
<thead>
<tr>
<th>#</th>
<th>Organization</th>
<th>System Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NNSA/LLNL</td>
<td>eServer Blue Gene Solution</td>
</tr>
<tr>
<td>2</td>
<td>NNSA/Sandia National Laboratories</td>
<td>Sandia/ Cray Red Storm, Opteron 2.4 GHz dual core</td>
</tr>
<tr>
<td>3</td>
<td>IBM Thomas J. Watson Research Center</td>
<td>eServer Blue Gene Solution</td>
</tr>
<tr>
<td>4</td>
<td>DOE/NNSA/LLNL</td>
<td>eServer pSeries p5 575 1.9 GHz</td>
</tr>
<tr>
<td>5</td>
<td>Barcelona Supercomputing Center</td>
<td>BladeCenter JS21 Cluster, PPC 970, 2.3 GHz, Myrinet</td>
</tr>
<tr>
<td>6</td>
<td>NNSA/Sandia National Laboratories</td>
<td>PowerEdge 1850, 3.6 GHz, Infiniband</td>
</tr>
<tr>
<td>7</td>
<td>Commissariat a l'Energie Atomique (CEA)</td>
<td>NovaScale 5160, Itanium2 1.6 GHz, Quadrics</td>
</tr>
<tr>
<td>8</td>
<td>NASA/Ames Research Center/NAS</td>
<td>SGI Altix 1.5 GHz, Voltaire Infiniband</td>
</tr>
<tr>
<td>9</td>
<td>GSIC Center, Tokyo Institute of Technology</td>
<td>Sun Fire x4600 Cluster, Opteron 2.4/2.6 GHz and ClearSpeed Accelerator, Infiniband</td>
</tr>
<tr>
<td>10</td>
<td>Oak Ridge National Laboratory</td>
<td>Cray XT3, 2.6 GHz dual Core</td>
</tr>
<tr>
<td></td>
<td>Organization</td>
<td>Architecture</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>DOE/NNSA/LLNL</td>
<td>eServer Blue Gene Solution</td>
</tr>
<tr>
<td>2</td>
<td>NNSA/Sandia National Laboratories</td>
<td>Sandia/ Cray Red Storm, Opteron 2.4 GHz dual core</td>
</tr>
<tr>
<td>3</td>
<td>IBM Thomas J. Watson Research Center</td>
<td>eServer Blue Gene Solution</td>
</tr>
<tr>
<td>4</td>
<td>DOE/NNSA/LLNL</td>
<td>eServer pSeries p5 575 1.9 GHz</td>
</tr>
<tr>
<td>5</td>
<td>Barcelona Supercomputing Center</td>
<td>BladeCenter JS21 Cluster, PPC 970, 2.3 GHz, Myrinet</td>
</tr>
<tr>
<td>6</td>
<td>NNSA/Sandia National Laboratories</td>
<td>PowerEdge 1850, 3.6 GHz, Infiniband</td>
</tr>
<tr>
<td>7</td>
<td>Commissariat a l'Energie Atomique (CEA)</td>
<td>NovaScale 5160, Itanium2 1.6 GHz, Quadrics</td>
</tr>
<tr>
<td>8</td>
<td>NASA/Ames Research Center/NAS</td>
<td>SGI Altix 1.5 GHz, Voltaire Infiniband</td>
</tr>
<tr>
<td>9</td>
<td>GSIC Center, Tokyo Institute of Technology</td>
<td>Sun Fire x4600 Cluster, Opteron 2.4/2.6 GHz and ClearSpeed Accelerator, Infiniband</td>
</tr>
<tr>
<td>10</td>
<td>Oak Ridge National Laboratory</td>
<td>Cray XT3, 2.6 GHz dual Core</td>
</tr>
</tbody>
</table>
Lustre + DDN

- Blue Gene L @ LLNL: 360TF
  - 130GB/s sustained data transfer rate
- Red Storm @ Sandia National Labs: 101.4TF
  - 110GB/s sustained data transfer rate
- Tera 10 @ CEA: 60TF
  - 100GB/s sustained data transfer rate
- Jaguar @ ORNL: 119TF
  - 45GB/s sustained data transfer rate
- Big Ben @ PSC: 10TF
  - 5GB/s sustained data transfer rate
Drive Roadmap

- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - Reliability, Serviceability & Availability

- Dragon Disk Enclosure
<table>
<thead>
<tr>
<th>Drive Roadmap</th>
<th>S2A</th>
<th>Disk Drives, SATA</th>
<th>Disk Drives, FC</th>
<th>Disk Drives, SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAID 6 Enhanced, R3.1</td>
<td>750GB SATA</td>
<td>73GB 15k FC 4Gb</td>
<td>146GB 15K SAS 3Gb</td>
</tr>
<tr>
<td></td>
<td>S2A9550</td>
<td>1TB SATA 3Gb</td>
<td>146GB 15K FC 4Gb</td>
<td>300GB 15K SAS 3Gb</td>
</tr>
<tr>
<td></td>
<td>S2A9900</td>
<td></td>
<td>300GB 15K FC 4Gb</td>
<td>450GB 15K SAS 3Gb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Today</th>
<th>Q2 '07</th>
<th>Q3 '07</th>
<th>Q4 '07</th>
<th>Q1 '08</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMI-s, R1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep Mode Drives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2A9900</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Drive Roadmap

S2A 9900
- Overview; 9500 vs. 9900 Comparison
- Performance Highlights
- Reliability, Serviceability & Availability

Dragon Disk Enclosure

Janus Storage System
**High Performance Scalability**
- 5+ GB per Second per Couplet
- Active/Active Controllers
- Parallel Shared Data Access Architecture
  - 8 IB-4X DDR and/or 8 FC-8 Host Ports to 20 SAS Disk Loops
  - Host Parallelism and PowerLUNs
- No Performance Loss in Degraded Mode
- RDMA Enabled — Low Latency Application Access

**Large Capacity, High Density Scalability**
- 600TB in one Rack: Scale Up to 1.2PB in Two Racks!!!
  - SAS or SATA Storage
  - RAID 6 (8+2) and Read & Write Parity Checking

**Best $ per Performance**

**Best $ per Capacity per Sq.Ft.**
## S2A 9900 Hardware Specifications

<table>
<thead>
<tr>
<th>Specification</th>
<th>S2A9900 Couplet</th>
<th>S2A9550 Couplet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported Disk Technology</td>
<td>SAS &amp; SATA</td>
<td>FibreChannel &amp; SATA</td>
</tr>
<tr>
<td>RAID Parity Protection</td>
<td>RAID6 8+2 Only</td>
<td>RAID3 (8+1+1), RAID6 8+2</td>
</tr>
<tr>
<td>Sustained Throughput</td>
<td>5.6GB/s – 6.0GB/s</td>
<td>2.4 GB/s – 2.8GB/s</td>
</tr>
<tr>
<td>Maximum Cache</td>
<td>5.0 GB ECC Protected</td>
<td>2.5GB RAID Protected</td>
</tr>
<tr>
<td>Minimum Cache</td>
<td>2.5 GB ECC Protected</td>
<td>2.5GB RAID Protected</td>
</tr>
<tr>
<td>Disk Side Ports</td>
<td>20 x SAS 4 Lane</td>
<td>20 x FC-2</td>
</tr>
<tr>
<td>Host Side FC Ports</td>
<td>8 x IB 4x DDR or 8 x FC-8</td>
<td>8 x FC-4 or 8 x IB 4x</td>
</tr>
<tr>
<td>Dimensions</td>
<td>7 x 19 x 28 in. (4U)</td>
<td>7 x 19 x 25 in. (4U)</td>
</tr>
<tr>
<td>Certifications</td>
<td>UL, CE, CUL, C-Tick, FCC</td>
<td>UL, CE, CUL, C-Tick, FCC</td>
</tr>
<tr>
<td>Release Date</td>
<td>1Q/2008</td>
<td>September 2005</td>
</tr>
</tbody>
</table>
Performance & Capacity Scalability

**Performance, GB/sec**

- **DDN S2A8500**
  - READS
  - WRITES

- **DDN S2A9550**
  - READS
  - WRITES

- **DDN S2A9900**
  - READS
  - WRITES

**Raw Capacity, TBs**

- **DDN S2A8500**
  - SATA

- **DDN S2A9550**
  - SAS
  - FC

- **DDN S2A9900**
  - SATA
  - FC

---

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION
- Singlet Failover Maintains Realtime Disk Access During Singlet Loss
- PowerLUNs can span arbitrary number of Tiers
- directRAID
  - Equivalent READ & WRITE performance
  - No performance degradation in crippled mode
  - Tremendous back-end performance for detection, very low-impact rebuild, disk scrubbing, etc.
- RAIDed Cache
- Parity Computed Writes
- Read Parity Checking for Each I/O Corrects Silent Data Corruption
- Double Disk Failure Protection Implemented in Hardware State Machine
- Multi-Tier Storage Support, SAS or SATA Disks
- Up to 1200 disks total
## S2A 9900 Capacity

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Storage Capacity</th>
<th>Drives Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five 60-Slot JBODs</td>
<td>300 Disks</td>
<td>300TB SATA using 1TB Drives, 135TB SAS using 450GB Drives</td>
</tr>
<tr>
<td>Ten 60-Slot JBODs</td>
<td>600 Disks</td>
<td>600TB SATA using 1TB Drives, 270TB SAS using 450GB Drives</td>
</tr>
<tr>
<td>Twenty 60-Slot JBODs</td>
<td>1200 Disks</td>
<td>1.2PB SATA using 1TB Drives, 540TB SAS using 450GB Drives</td>
</tr>
</tbody>
</table>

*Confidential Information*
Improvements

- Faster Intel Main CPU
- Faster Interface
  - SDR IB -> DDR IB
  - FC4 -> FC8
- PCI Express Bus Architecture
- Faster Intel Host Processors
- Doubled Cache Size & Cache Rate
- Faster Backend
  - FC2 -> SAS
- Optimized Drive Health Management
- Increased Component Reliability
  - Cooling
  - Connection
Additional Enhancements

- Expanded log capability
- Rebuild write journaling
- Power Down Archiving of writeback data (coupled with UPS)
- Power Consumption Reduction
  - Sleep Mode Drives (SATA)
  - DC Power
Drive Roadmap

S2A 9900
  – Overview; 9500 vs. 9900 Comparison
  – Performance Highlights
  – Reliability, Serviceability & Availability

Dragon Disk Enclosure

Janus Storage System
12GB/s potential backend bandwidth

10 x 4-lane SAS Channels per Singlet

Disk Channel Controller
- Provides Cache to SAS Connectivity
- Provides 2.5GB/5GB Cache Memory Segment via DCC FPGA
- Cache Controller Interface
- Interfaces to Main CPU via Dual Port SRAM
Front-end Throughput

- Maximum **4GB/s** Singlet Front-end Bandwidth
- 4 x 8-lane PCI Express Ports per Singlet

**Host Interface**
- Dual Protocol
  - Fibre Channel (FC8 when available)
  - Infiniband (DDR x4 IB SRP target (iSER tbd))
- DMA Capable
  - Enables Zero-Copy Interfacing
Target: 2-3X 9550 Performance

- Robust Processors:
  - Intel Chevelon Host CPU
  - Intel Sunrise Lake Main CPU
- Faster Cache Controller/Stage Buffer FPGA
- Faster processor DRAM: 512Mb DDR2
  - 3.2GBytes/sec processor to memory bandwidth & reduced latencies
Roadmap

 vídeoDrive Roadmap

 S2A 9900
 – Overview; 9500 vs. 9900 Comparison
 – Performance Highlights
 – Reliability, Serviceability & Availability

 Dragon Disk Enclosure

 Janus Storage System
SATA technology has enabled great cost economies but can significantly jeopardize data integrity without proper controls
- DDN has the experience (a recognized leader in SATA)
- DDN has the understanding (multi-faceted SATA protections)

The Challenge: to maintain QOS regardless of drive retry, reset, and internal recovery issues.

The Solution: All devices will be constantly monitored through HW and SW for excessive errors or defect growth and system software can begin rebuilds to spares before a failure occurs.
The Hardware Solution

- Check parity for every read and correct it in real time.
- Use RAID 6 to identify individual drives that have read corrupt data through Reed-Solomon data recovery algorithms.
- Exercise total control over the array including the ability to power cycle each drive.
Increased Data Availability

The Software Solution

- Take a questionable drive offline immediately.
- Begin a journal of all writes that have been made to the array since the moment that a specific element was taken offline.
- Utilize a series of recovery techniques including command retries, drive resets, and finally power cycling to confirm the status of the specific device.
- If the device cannot be revived it can be replaced.
- If the device can be revived it can be rebuilt from the journal in a short time.
Simplified Design

- PCI-E Serial Bus Structure Enable Significant Connection Reduction
  - 10x-100x Reduction in Component Connections
    - Less Controller Failures/Errors
  - All while increasing performance by 2x!
  - By-Products:
    - Flip-Chip BGAs for all High I/O FPGAs
    - PCI Express has less connector pins and BGA pins
    - DDR2 DRAM eliminates termination requirements
Simplified Design

- Improved Power Management
  - Enhanced Power Supplies
    - Higher Reliability Technology
    - Increased Supportability
    - Better Power Supply Fault Isolation & Monitoring
  - Use Two Supplies instead of Four

- Increased Cooling
  - Moving to 2 power supplies allows full width cooling in 1U
    - Increase potential airflow from: 50CFM to: 75CFM
  - Newer ICs deliver enhanced thermal monitoring
Drive Roadmap

S2A 9900
- Performance Highlights
- Reliability, Serviceability & Availability

Dragon Disk Enclosure

Janus Storage System
Dragon Enclosure

- 4U 60-Bay Enclosure
  - 3.5” Drives
  - Redundant Power & Cooling
    - Drives vertically organized for maximum cooling
  - Dual SAS I/O slots provide dual-channel access
  - Supports SATA & SAS Drives
    - Muxes added to SATA drives for dual-porting
Dragon Enclosure

- 2 Passive Baseboards
- 8 active SAS expander cards (4- “A” & 4 “B”)
  - Groups of 15 drives
- All expander cards are located in the middle of the enclosure drive section.
- Cards are top removable.
- IO modules are SBB compliant and plug into the rear of the enclosure.
- Redundant Power Supplies
  - Hot-swappable
  - Plug into the rear of the enclosure
  - Provides system cooling

Figure 2. Dragon Top View
Power Cycling Capabilities Will Increase System Reliability - Reduce Drive Replacements

- Not all unresponsive drives are dead drives
- 9900+ will implement a series of recovery techniques including command retries & drive resets
- If unsuccessful, Dragon enclosure will have ability to power cycle individual drives to confirm the status of the specific device.
- If the device cannot be revived it can be replaced online.
Storage Architecture and Roadmap
Lustre User’s Group
April, 2007

Dave Fellinger, CTO
dfellinger@datadirectnet.com