DataDirect NETWORKS

performance, capacity and innovation

Storage Architecture and Roadmap
Lustre User's Group
April, 2007

Dave Fellinger, CTO
dfellinger@datadirectnet.com

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

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

Lustre + DDN

DataDirect NETWORKS
performance, capacity and innovation

- 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

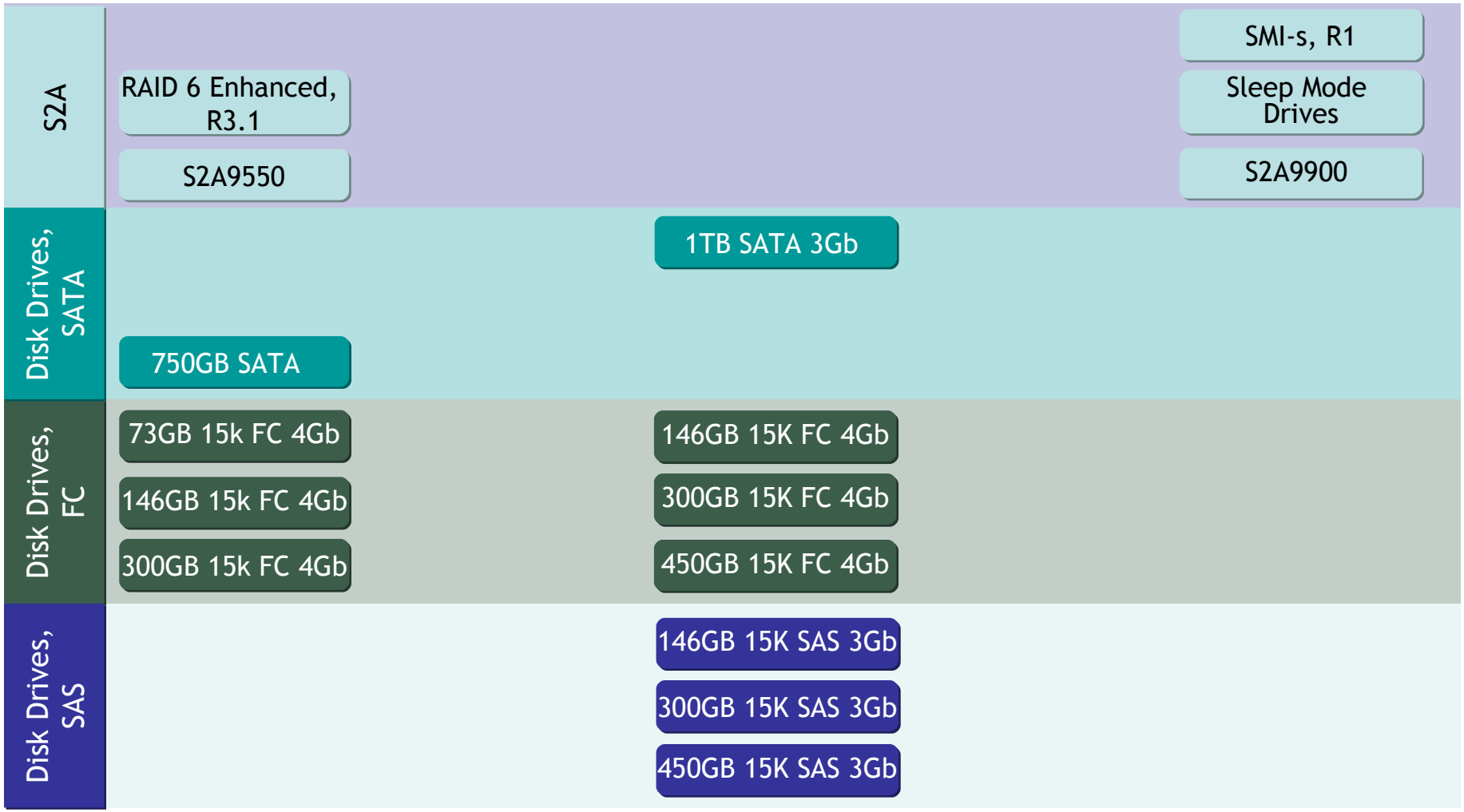
Roadmap

DataDirect NETWORKS
performance, capacity and innovation

- Drive Roadmap
- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - Reliability, Serviceability & Availability
- Dragon Disk Enclosure

Drive Roadmap

DataDirect NETWORKS
performance, capacity and innovation



- Drive Roadmap
- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - Reliability, Serviceability & Availability
- Dragon Disk Enclosure
- Janus Storage System

S2A Storage Technology Difference

DataDirect NETWORKS
performance, capacity and innovation

- **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

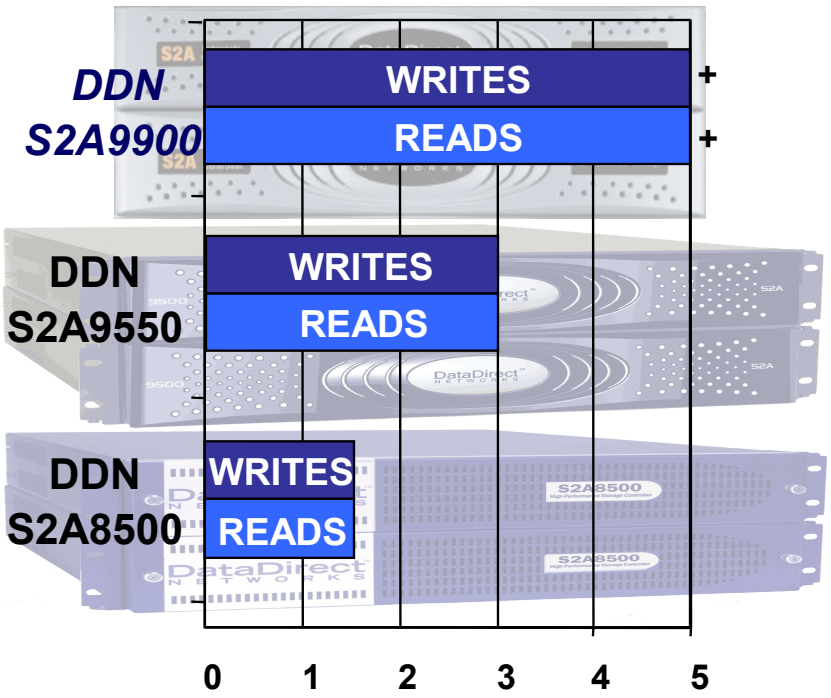
DataDirect NETWORKS
performance, capacity and innovation

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

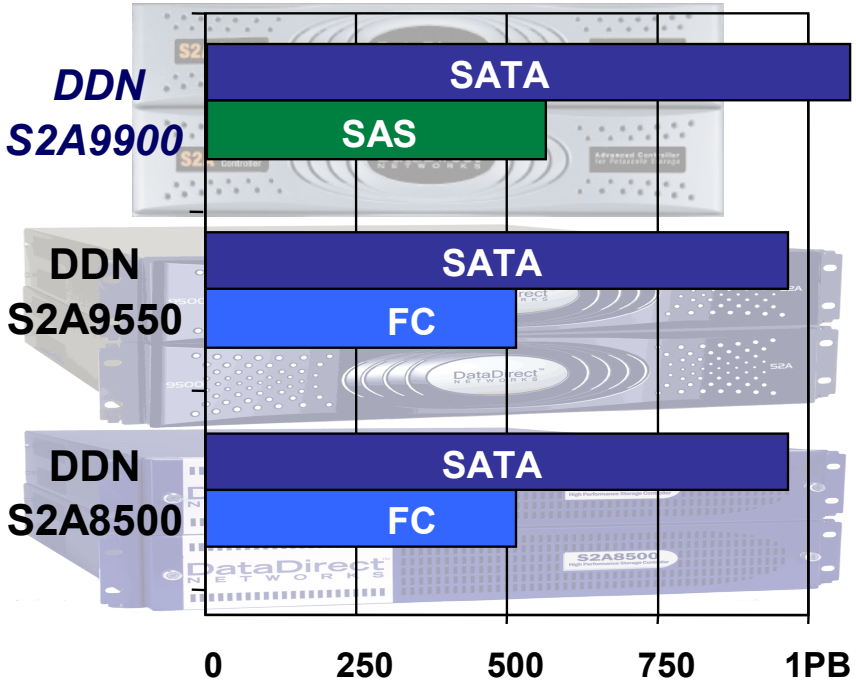
Performance & Capacity Scalability

DataDirect NETWORKS
performance, capacity and innovation

Performance, GB/sec

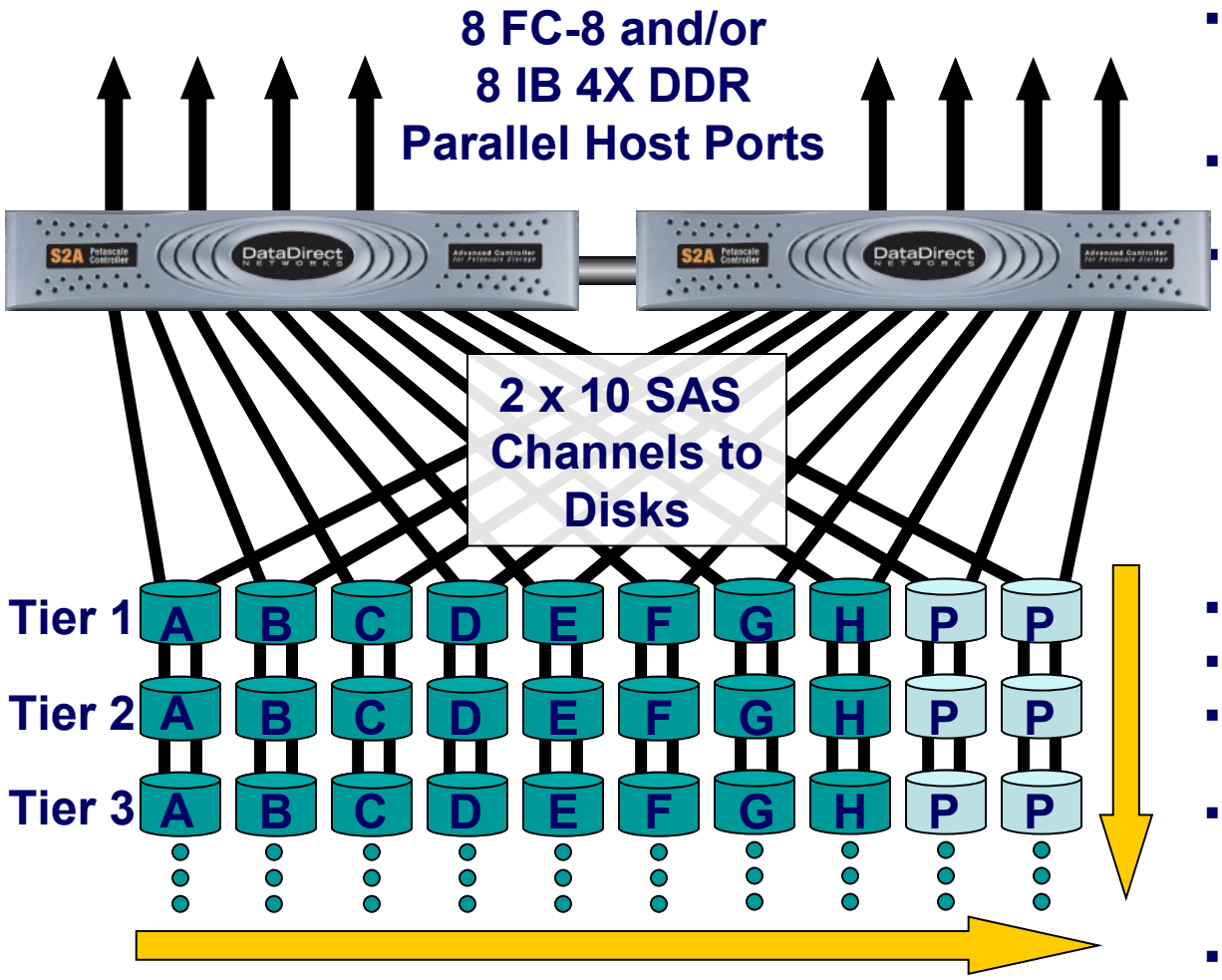


Raw Capacity, TBs



S2A Architecture, 8+2

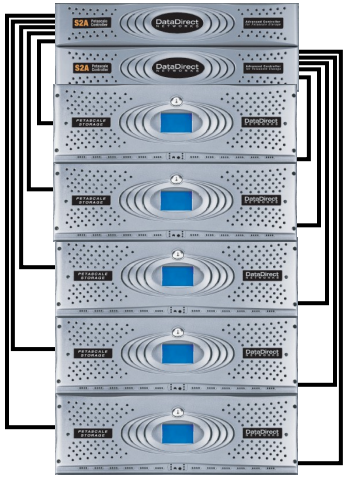
DataDirect NETWORKS
performance, capacity and innovation

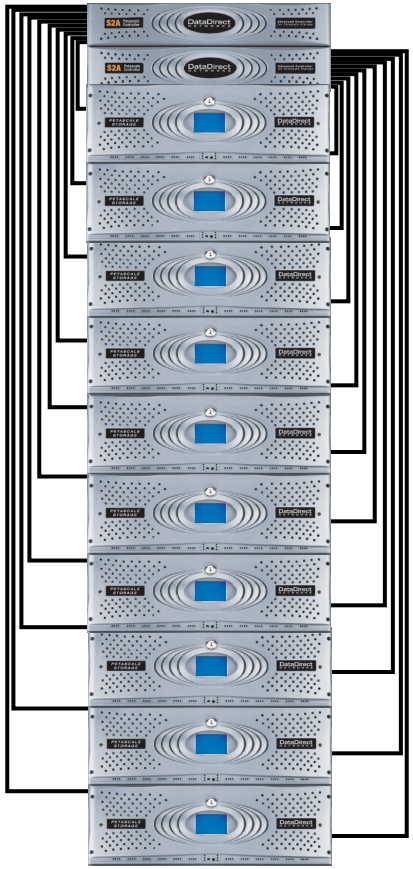


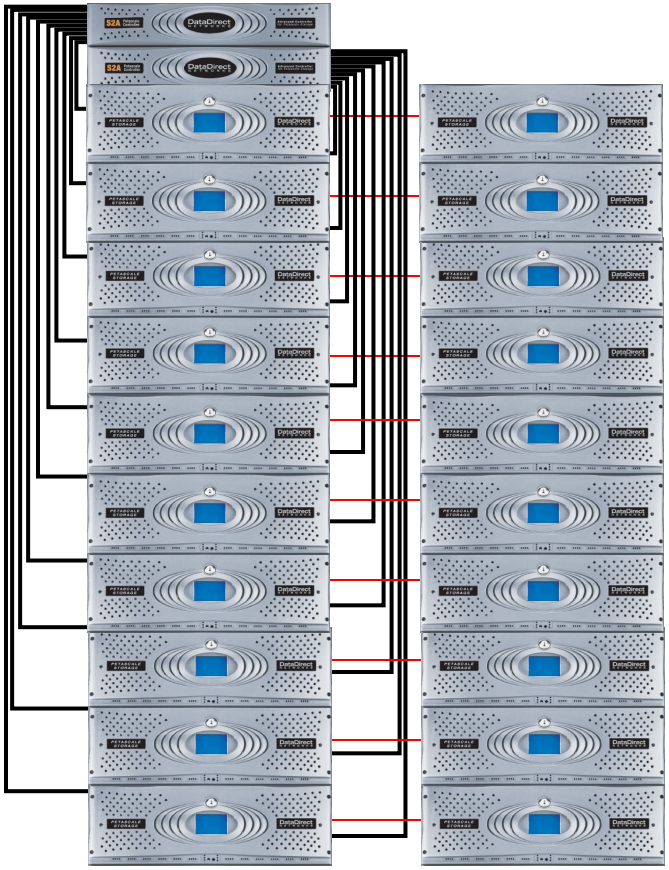
- Singlet Failover Maintains Realtime Disk Access During Singlet Loss
- PowerLUNs can span arbitrary number of Tiers
- directRAID
  - Equivalent READ & WRITE performance
  - No performance degradation in crippled mode
  - Tremendous back-end performance for detection, very low-impact rebuild, disk scrubbing, 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
  - 960 Formattable Disks

S2A 9900 Capacity

DataDirect NETWORKS
performance, capacity and innovation







•Five 60-Slot JBODs •Two Dual Loop per JBOD: 300 Disks • 300TB SATA using 1TB Drives • 135TB SAS using 450GB Drives	•Ten 60-Slot JBODs •Two Dual Loop per JBOD: 600 Disks • 600TB SATA using 1TB Drives • 270TB SAS using 450GB Drives	•Twenty 60-Slot JBODs •Two Dual Loop per JBOD: 1200 Disks • 1.2PB SATA using 1TB Drives • 540TB SAS using 450GB Drives

Improvements

DataDirect NETWORKS
performance, capacity and innovation

- 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

DataDirect NETWORKS
performance, capacity and innovation

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

Roadmap

DataDirect NETWORKS
performance, capacity and innovation

- Drive Roadmap
- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - Reliability, Serviceability & Availability
- Dragon Disk Enclosure
- Janus Storage System

Backend Throughput

DataDirect NETWORKS
performance, capacity and innovation

- **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

DataDirect NETWORKS
performance, capacity and innovation

- 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

Increased IOPS

DataDirect NETWORKS
performance, capacity and innovation

- 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

DataDirect NETWORKS
performance, capacity and innovation

- Drive Roadmap
- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - Reliability, Serviceability & Availability
- Dragon Disk Enclosure
- Janus Storage System

Increase Data Availability

DataDirect NETWORKS
performance, capacity and innovation

- 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.

Increased Data Availability

DataDirect NETWORKS
performance, capacity and innovation

- 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

DataDirect NETWORKS
performance, capacity and innovation

- 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

DataDirect NETWORKS
performance, capacity and innovation

- 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

DataDirect NETWORKS
performance, capacity and innovation

- 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

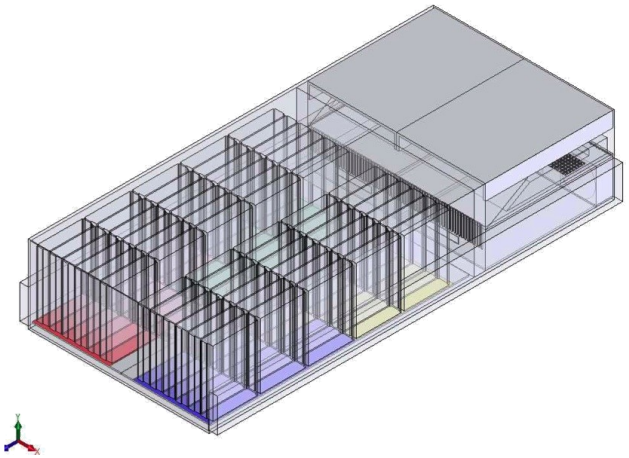
Roadmap

DataDirect NETWORKS
performance, capacity and innovation

- Drive Roadmap
- S2A 9900
  - Performance Highlights
  - Reliability, Serviceability & Availability
- Dragon Disk Enclosure
- Janus Storage System

Dragon Enclosure

DataDirect NETWORKS
performance, capacity and innovation



- 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

DataDirect NETWORKS
performance, capacity and innovation

- 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

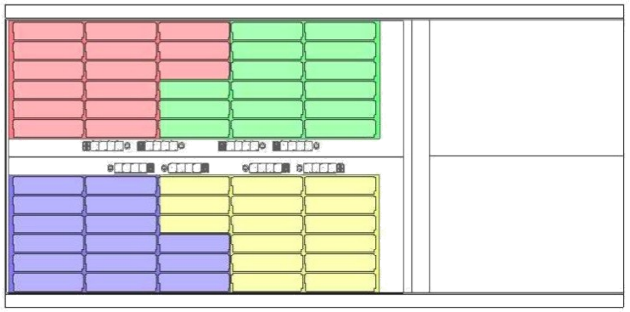


Figure 2. Dragon Top View

Dragon Enclosure

DataDirect NETWORKS
performance, capacity and innovation

- 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.

DataDirect NETWORKS

performance, capacity and innovation

Storage Architecture and Roadmap
Lustre User's Group
April, 2007

Dave Fellinger, CTO
dfellinger@datadirectnet.com