Minimizing the I/O Cycle Time in Simulation Clusters Through the Use of High Performance Storage

Cray Users Group

Dave Fellinger, CTO
dfellinger@datadirectnet.com

Joint DDN/Cray Installations

Customer	Rank	Computer
NNSA / Sandia National Laboratories	2	Sandia/Cray Red Storm, Opteron 2.4 GHz dual core - 88 DDN Couplets
Oak Ridge National Laboratory	10	Cray XT3, 2.6GHz dual core
Atomic Weapons Establishment	15	Cray XT3, 2.6GHz dual core
ERDC MSRC	26	Cray XT3, 2.6GHz
Pittsburgh Supercomputing Center	85	Cray XT3, 2.4GHz
Swiss Scientific Computing Center (CSCS)	94	Cray XT3, 2.6GHz
UK Engineering and Physical Sciences Research Council (EPSRC)	TBD	Cray XT4 Opteron MPP / BlackWidow

Several Top 100 Computing sites use Cray and DataDirect S2A

Agenda

- **Cluster Storage Requirements**
- Parallel Storage Architecture
- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - Reliability, Serviceability & Availability

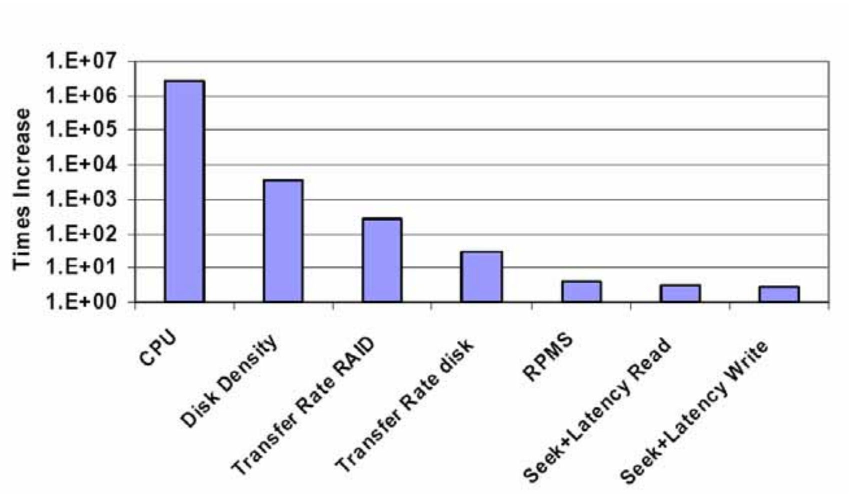
Cluster Storage Requirements

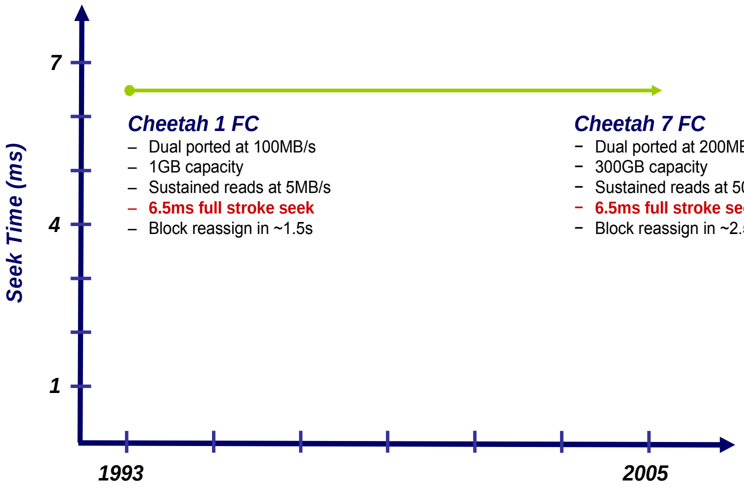
- “Scratch” storage on simulation clusters has specific requirements
- Write cycles must be fast and consistent
- Disk I/O errors and retries cannot affect the performance of writes to the system
- I/O rates must scale well across threads and transfer size
- Storage for visualization clusters has specific requirements
- Read cycles must be fast and consistant
- Disk I/O must be checked for errors if SATA is employed
- Disk I/O errors and retries cannot affect the performance of reads from the system

I/O and Storage Challenges

DataDirect NETWORKS

performance, capacity and innovation



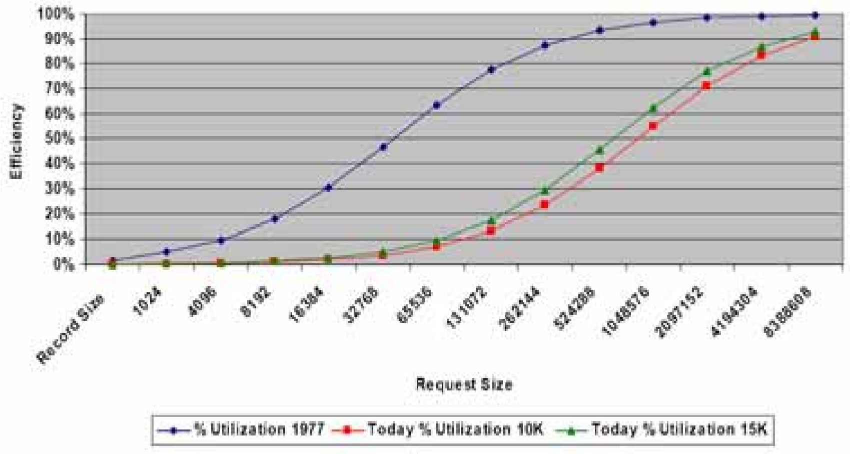


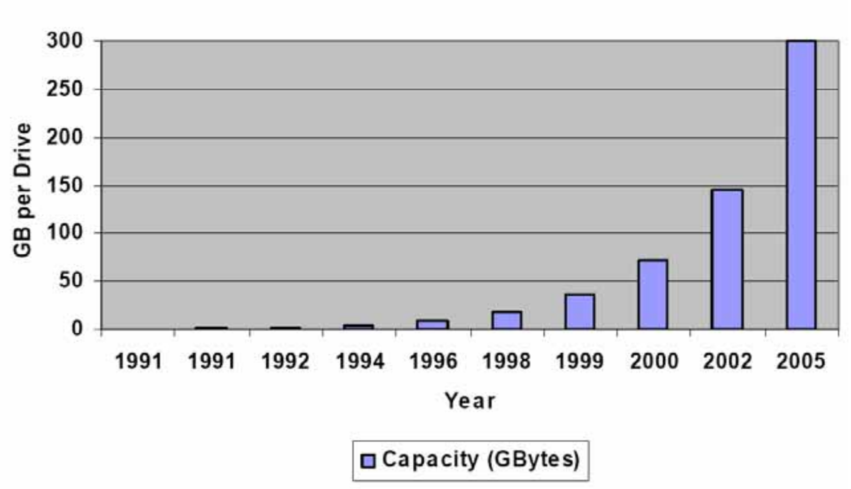
Cheetah 1 FC

- Dual ported at 100MB/s
- 1GB capacity
- Sustained reads at 5MB/s
- 6.5ms full stroke seek
- Block reassign in ~1.5s

Cheetah 7 FC

- Dual ported at 200MB/s
- 300GB capacity
- Sustained reads at 50+MB/s
- 6.5ms full stroke seek
- Block reassign in ~2.5s





Source: David Koester, Ph.D. and Henry Newman @ HPCS I/O Workshop, July 12, 2005

Drive Roadmap

	Today	Q2 ‘07	Q3 ‘07	Q4 ‘07	Q1 ‘08
S2A	RAID 6 Enhanced, R3.1 S2A9550				SMI-s, R1 Sleep Mode Drives S2A9900
Disk Drives, SATA	750GB SATA		1TB SATA 3Gb		
Disk Drives, FC	73GB 15k FC 4Gb 146GB 15k FC 4Gb 300GB 15k FC 4Gb		146GB 15K FC 4Gb 300GB 15K FC 4Gb 450GB 15K FC 4Gb		
Disk Drives, SAS			146GB 15K SAS 3Gb 300GB 15K SAS 3Gb 450GB 15K SAS 3Gb		

Agenda

- Cluster Storage Requirements
- **Parallel Storage Architecture**
- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - Reliability, Serviceability & Availability

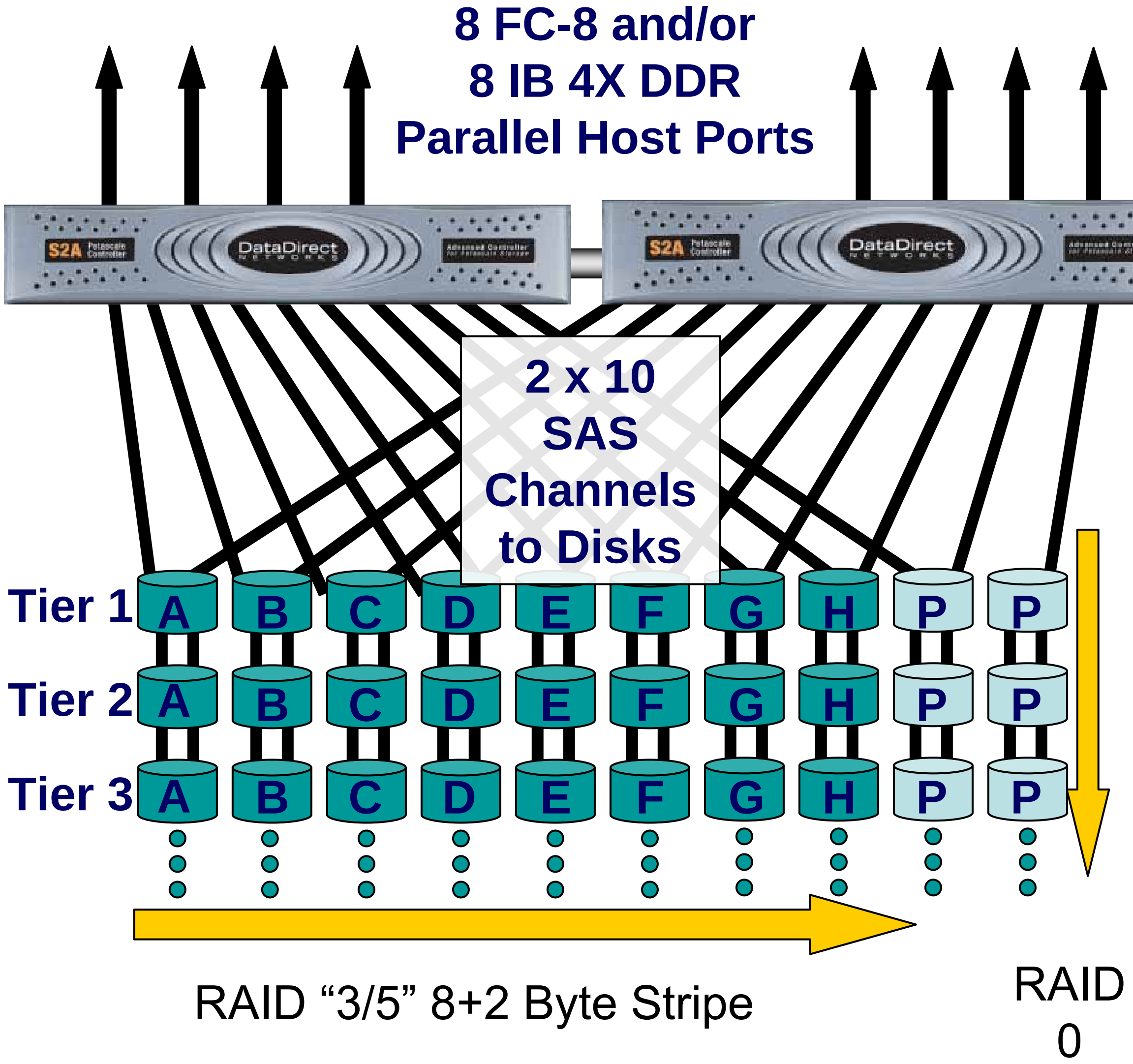
Low Latency High Performance Silicon Based Storage Controller with RDMA

- Parallel Access For Hosts
- Parallel Access To A Large Number Of Disk Drives
- True Performance Aggregation
- Reliability From A Parallel Pool
- 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

RAID 6 Architecture

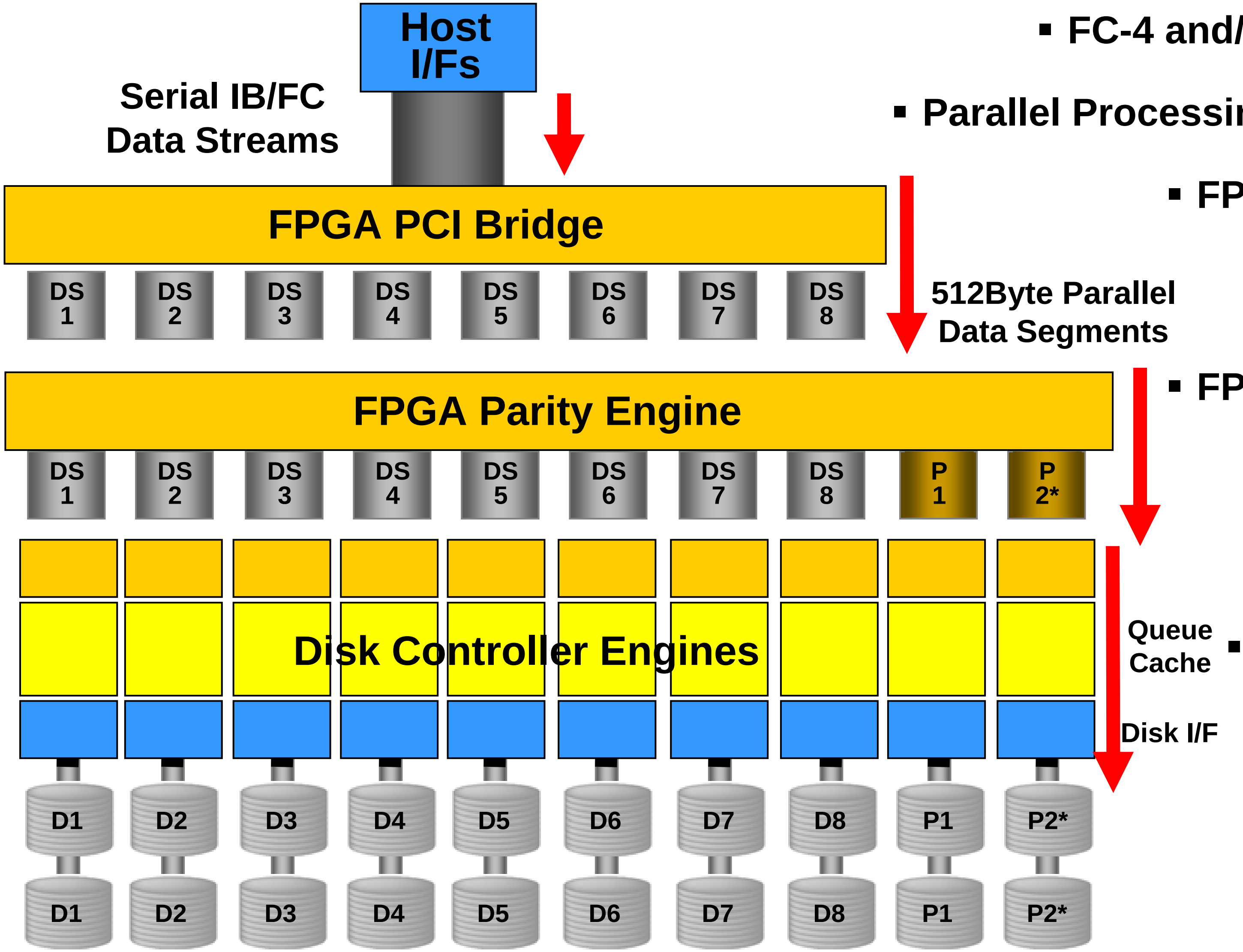
DataDirect NETWORKS

performance, capacity and innovation



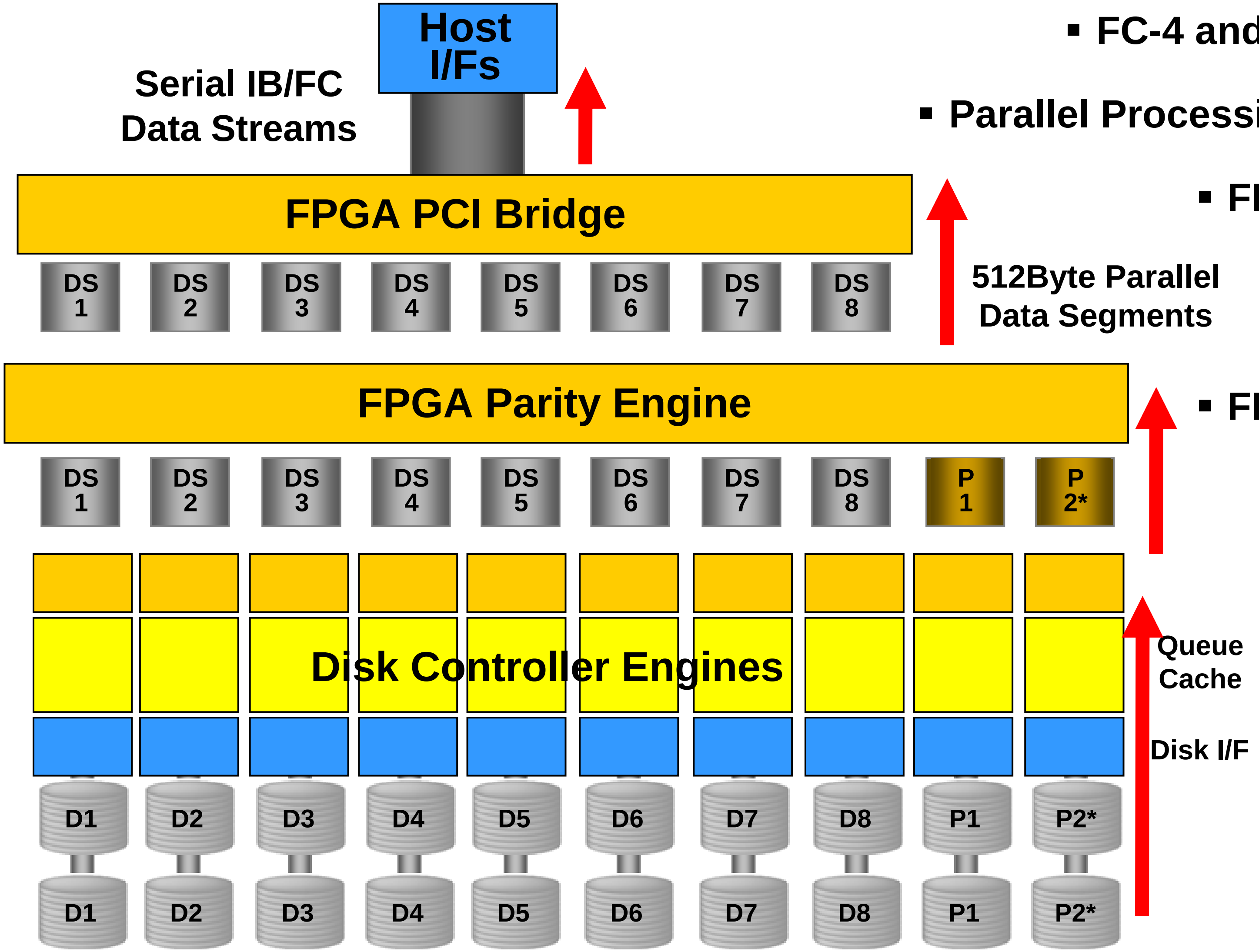
- Singlet Failover Maintains Realtime Disk Access During Singlet Loss
- PowerLUNs can span arbitrary number of Tiers
- directRAID
  - Equivalent READ & WRITE performance
  - No performance degradation in crippled mode
  - Tremendous back-end performance for detection, very low-impact rebuild, disk scrubbing,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

Data Flow, To Disk



- FC-4 and/or IB Host Interface
- Parallel Processing Parity Engine
- FPGA PCI Bridge
  - Generates 512B Parallel Segments
- FPGA Parity Engine
  - Generates One or Optionally Two Parity Segments Synchronously
- Disk Controller Engines
  - Queue Command Re-Ordering in Queue Cache
  - Vertical Striping
  - Disk Interfaces

Data Flow, From Disk



- FC-4 and/or IB Host Interface
- Parallel Processing Parity Engine
- FPGA PCI Bridge
  - Saturate Multiple Host Ports w/ High Speed Read Data
- FPGA Parity Engine
  - Real-time Parity Checking and Data Correction for each Read I/O Synchronously
- Disk Controller Engines
  - Data Staging with Level One Cache
  - Shared Data Access

Reliability and Performance Solution

Implementation of RAID 6

- Added parity drives effect double redundancy
- Reed Solomon coding in Real Time
- Continuous parity checking in Reads and real time generation in Writes
- Bad block recovery in real time
- Drive error recovery in real time
- Partial rebuilds without affecting host side access

Agenda

DataDirect NETWORKS

performance, capacity and innovation

- Cluster Storage Requirements
- Parallel Storage Architecture
- **S2A 9900**
  - **Overview; 9500 vs. 9900 Comparison**
  - Performance Highlights
  - Reliability, Serviceability & Availability

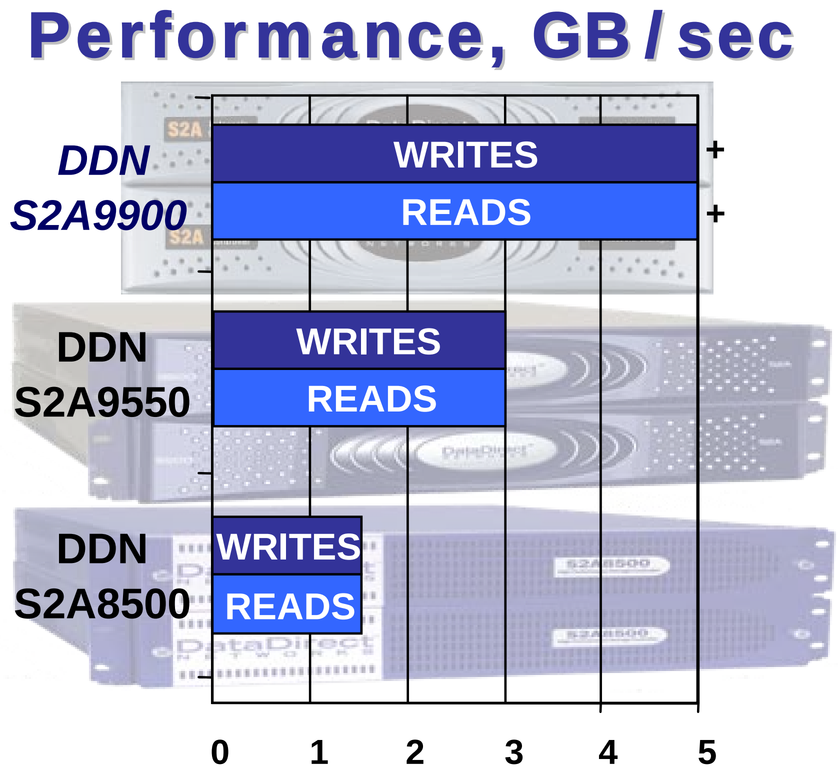
S2A9900 Hardware Specifications

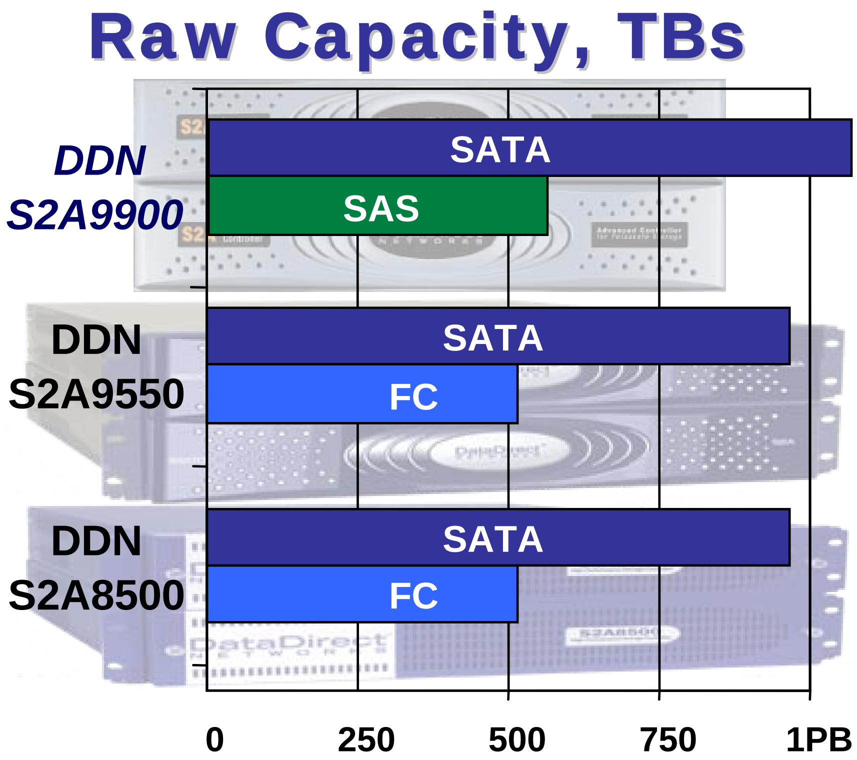
Specification	S2A9900 Couplet	S2A9550 Couplet
Supported Disk Technology	**SAS** & SATA	Fibre Channel & SATA
RAID Parity Protection	**RAID6 8+2 Only**	RAID3 (8+1+1), RAID6 8+2
Sustained Throughput	**5.6GB/s – 6.0GB/s**	2.4 GB/s – 2.8GB/s
Maximum Cache	**5.0 GB ECC Protected**	2.5GB RAID Protected
Minimum 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

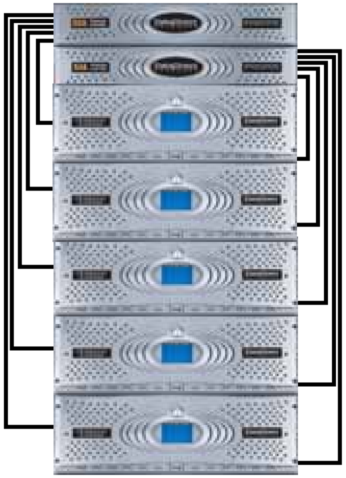
DataDirect NETWORKS

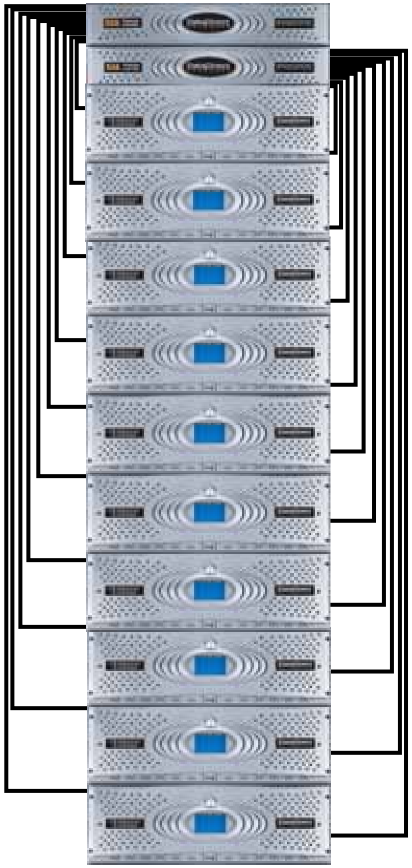
performance, capacity and innovation

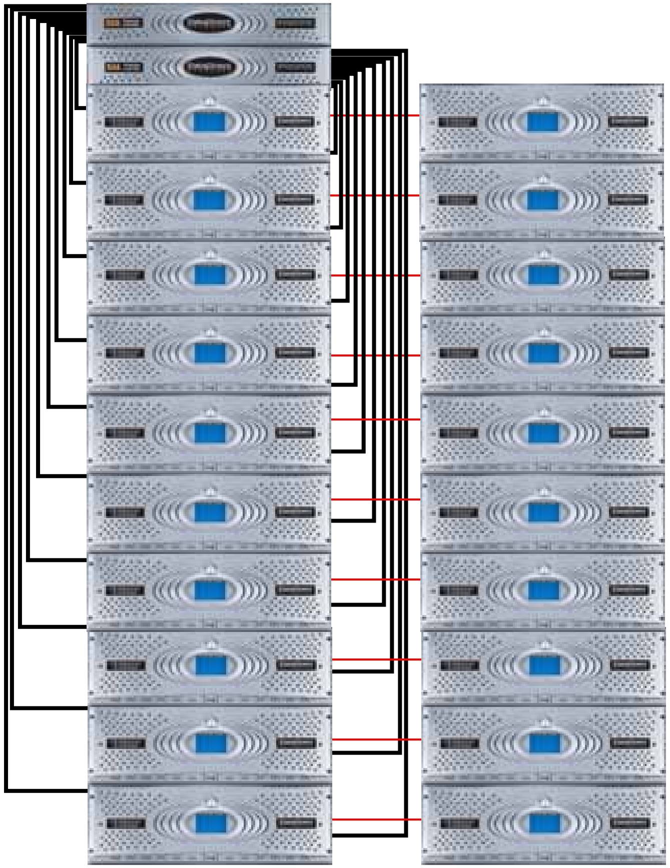




S2A9900 Capacity







•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

- 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

Agenda

- Cluster Storage Requirements
- Parallel Storage Architecture
- **S2A9900**
  - Overview; 9500 vs. 9900 Comparison
  - **Performance Highlights**
  - Reliability, Serviceability & Availability

Backend Throughput

- **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

- 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

Agenda

DataDirect NETWORKS

performance, capacity and innovation

- Cluster Storage Requirements
- Parallel Storage Architecture
- **S2A9900**
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights
  - **Reliability, Serviceability & Availability**

Increase Data Availability

- 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

- 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

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

Cray Users Group

Dave Fellinger, CTO
dfellinger@datadirectnet.com