DataDirect NETWORKS
performance, capacity and innovation

Architecting Storage for Petascale Clusters

Key Challenges in Modeling and Simulation
September 25, 2007

Dave Fellinger, CTO
dfellinger@datadirectnet.com

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Agenda

DataDirect NETWORKS
performance, capacity and innovation

- Storage Challenges
- Parallel Storage Architecture
- S2A 9900
- Future Requirements

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Cluster Storage Requirements

DataDirect NETWORKS
performance, capacity and innovation

"Scratch" storage on simulation clusters has specific requirements

- **Write cycles must be fast and consistant**
- **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**

Access to Data and Metadata requires both large and small transfer requirements

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Agenda

DataDirect NETWORKS
performance, capacity and innovation

- Cluster Storage Requirements
- Drive Roadmap
- Parallel Storage Architecture
- S2A 9900
  - Overview; 9500 vs. 9900 Comparison
  - Performance Highlights

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Storage Trends

DataDirect NETWORKS
performance, capacity and innovation

- Moore's law is still valid for CPU processing power
  - 2X performance every 18 months = CAGR of 45% per annum

- How about storage at the individual disk drive level?
  - Disk areal density = CAGR of 60%/year (capacity is doing fine)
  - Seek time improvement = CAGR of <15%/year (performance is NOT doing fine!)
  - Disk IO/second increase = CAGR of <15%/year (performance is NOT doing fine!)
  - Sustained bandwidth = CAGR of <25%/year (performance is NOT doing fine!)

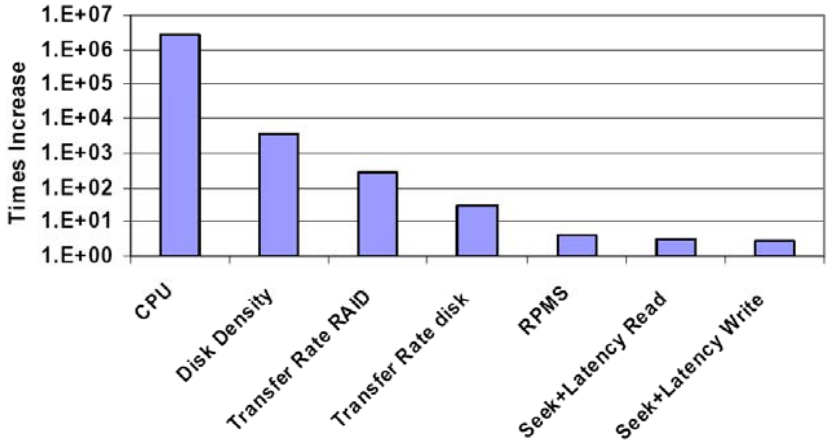
- What does this mean?
  - The more cost effective the system, the lower its performance!!!
  - The ratio of Performance to Capacity has become a significant factor in managing storage subsystem performance

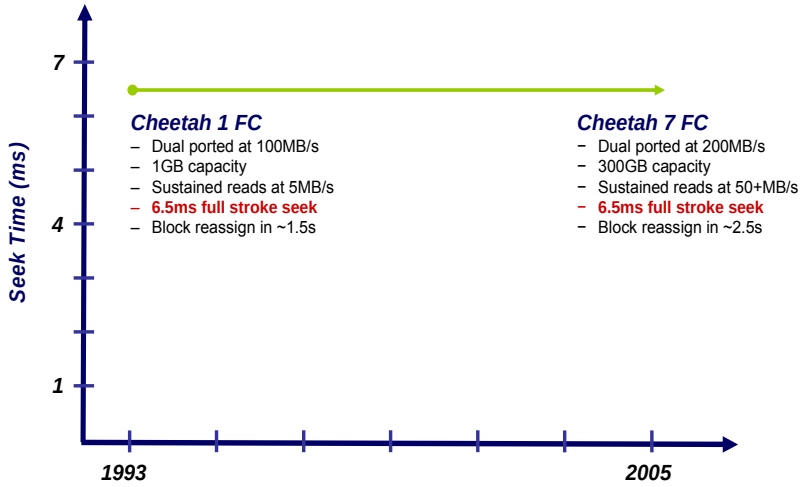
THE PERFORMANCE OF DISK DRIVES IS NOT KEEPING UP WITH OTHER TECHNOLOGICAL TRENDS IN THE CLUSTER!

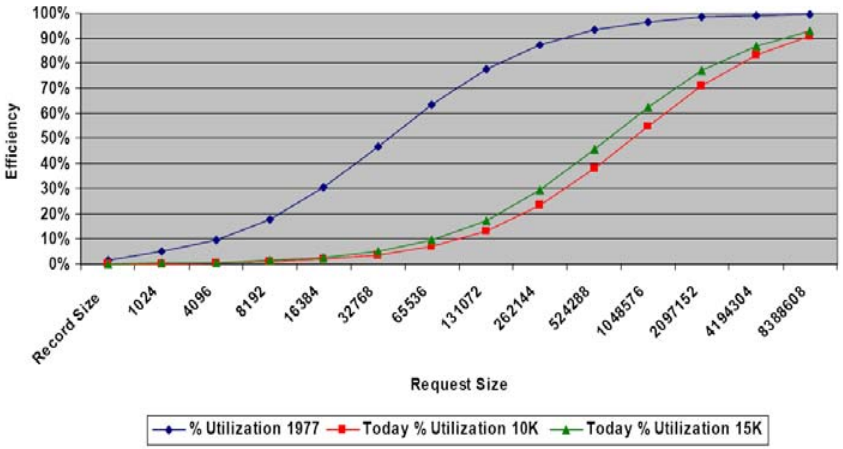
Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

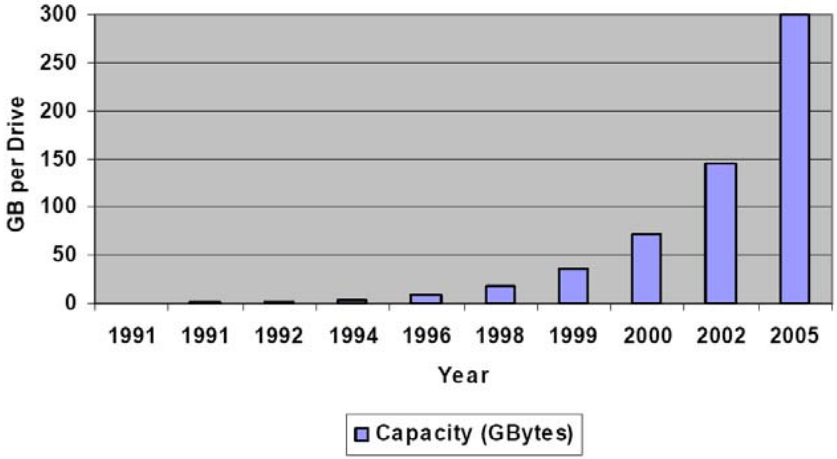
I/O and Storage Challenges

DataDirect NETWORKS
performance, capacity and innovation







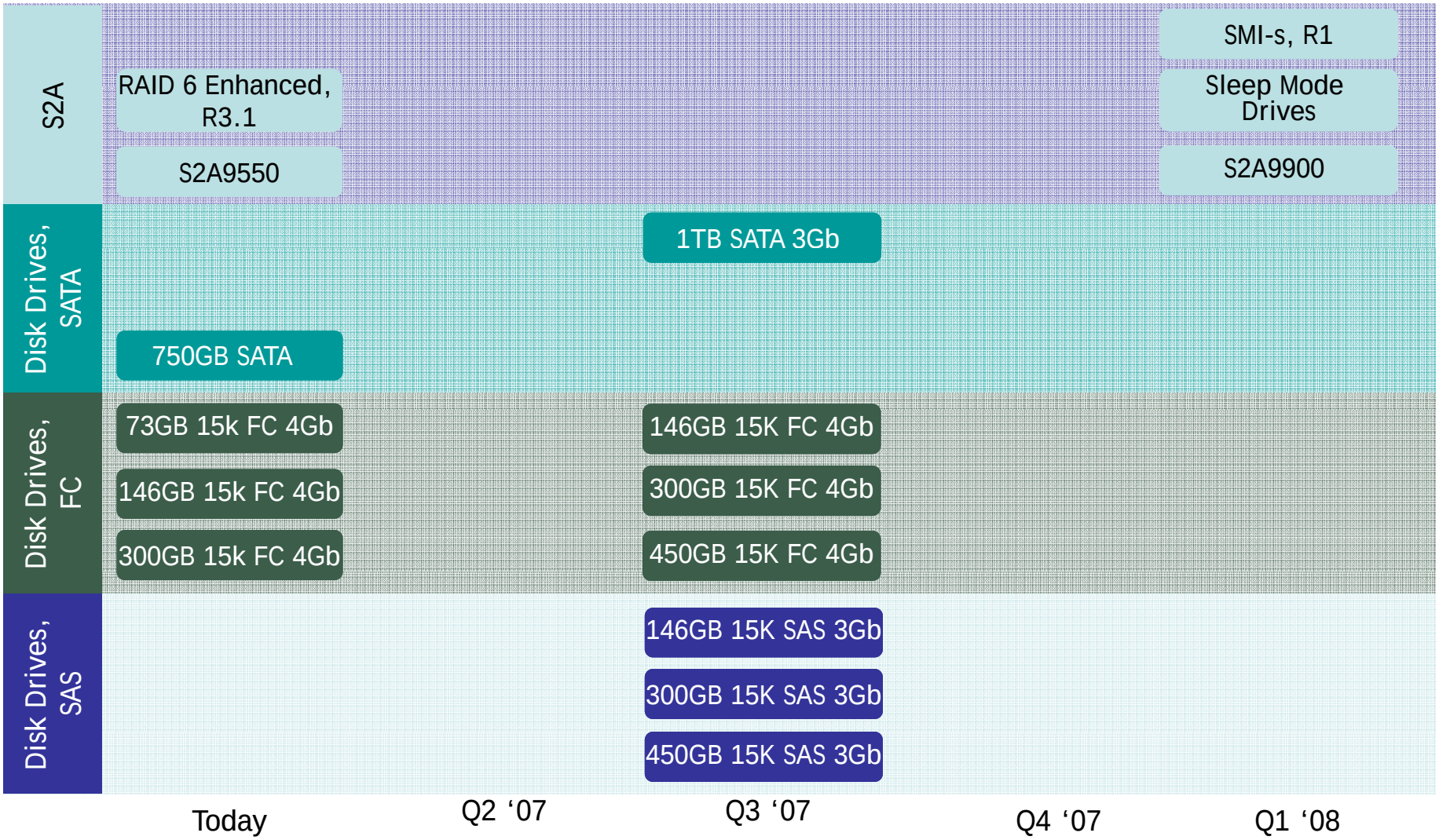


Source: David Koester, Ph.D. and Henry Newman @ HPCS I/O Workshop, July 12, 2005

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Drive Roadmap

DataDirect NETWORKS
performance, capacity and innovation



Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Agenda

DataDirect NETWORKS
performance, capacity and innovation

- Storage Challenges
- Parallel Storage Architecture
- S2A 9900
- Future Requirements

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Large Scale Parallel Storage

DataDirect NETWORKS
performance, capacity and innovation

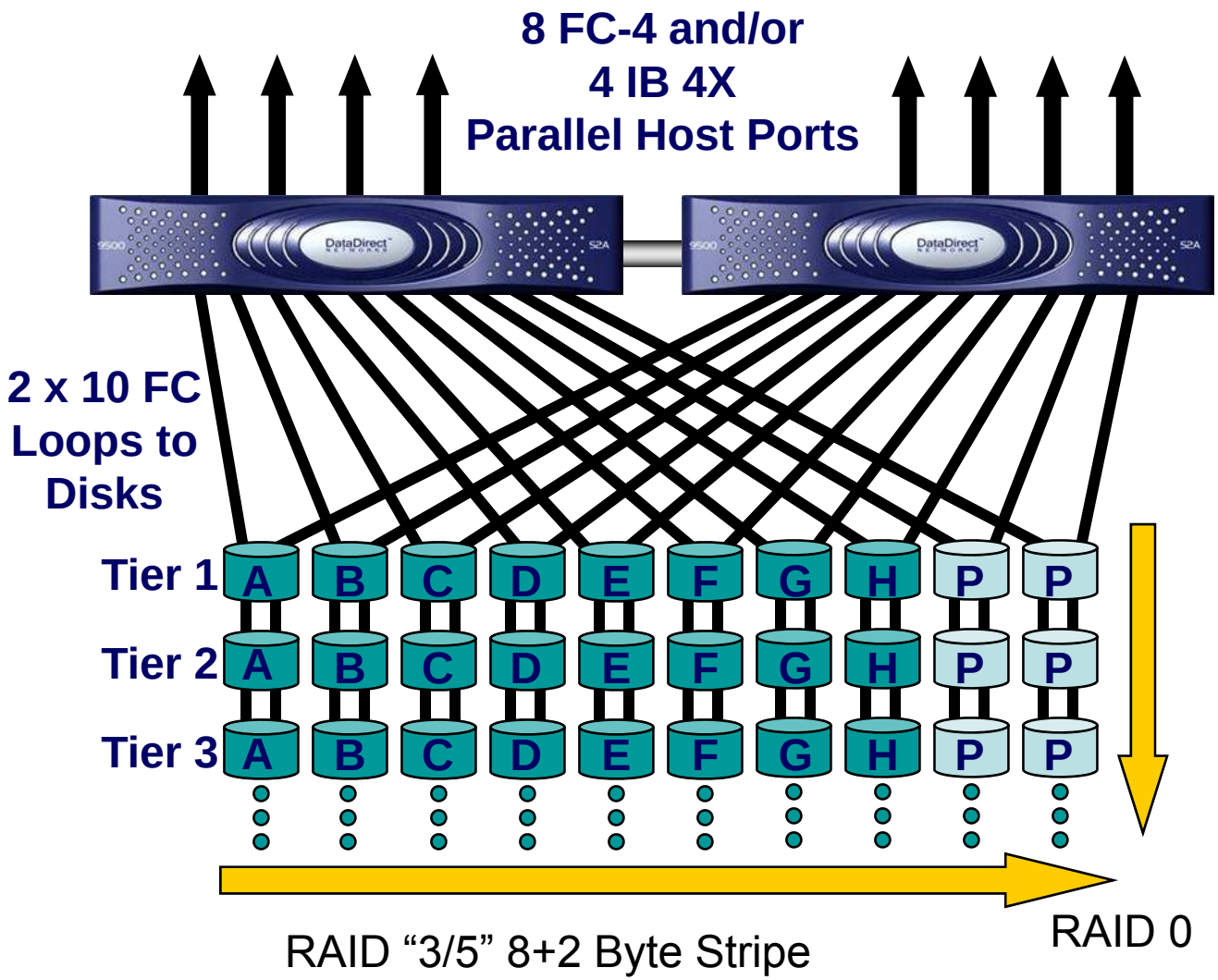
Low Latency High Performance Silicon Based Storage Controller Addressing Root Cause Disk Access Anomalies

- 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

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

An Implementation of Parallelism w/ Double Parity RAID Protection

DataDirect NETWORKS
performance, capacity and innovation

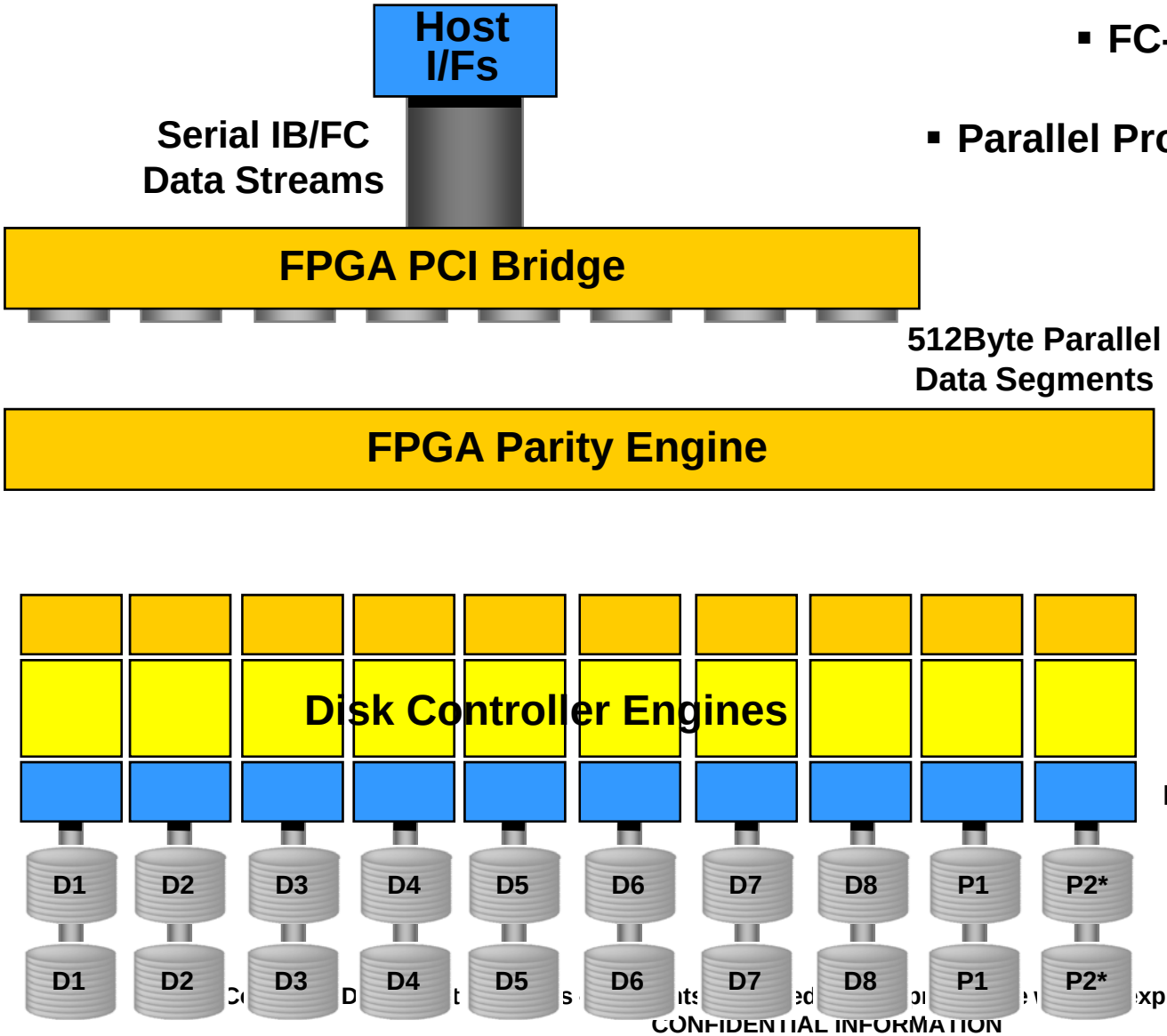


- Double Disk Failure Protection
- LUNs can span tiers
- All ports access all storage
- Implemented in Hardware State Machine
  - Equivalent 8+1 & 8+2 performance
- Parity Computed On Writes AND Reads
- No loss of performance on any failure
- Multi-Tier Storage Support, Fibre Channel or SATA Disks
- Up to 1250 disks total
  - 1000 formattable disks

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Data Flow, To Disk

DataDirect NETWORKS
performance, capacity and innovation

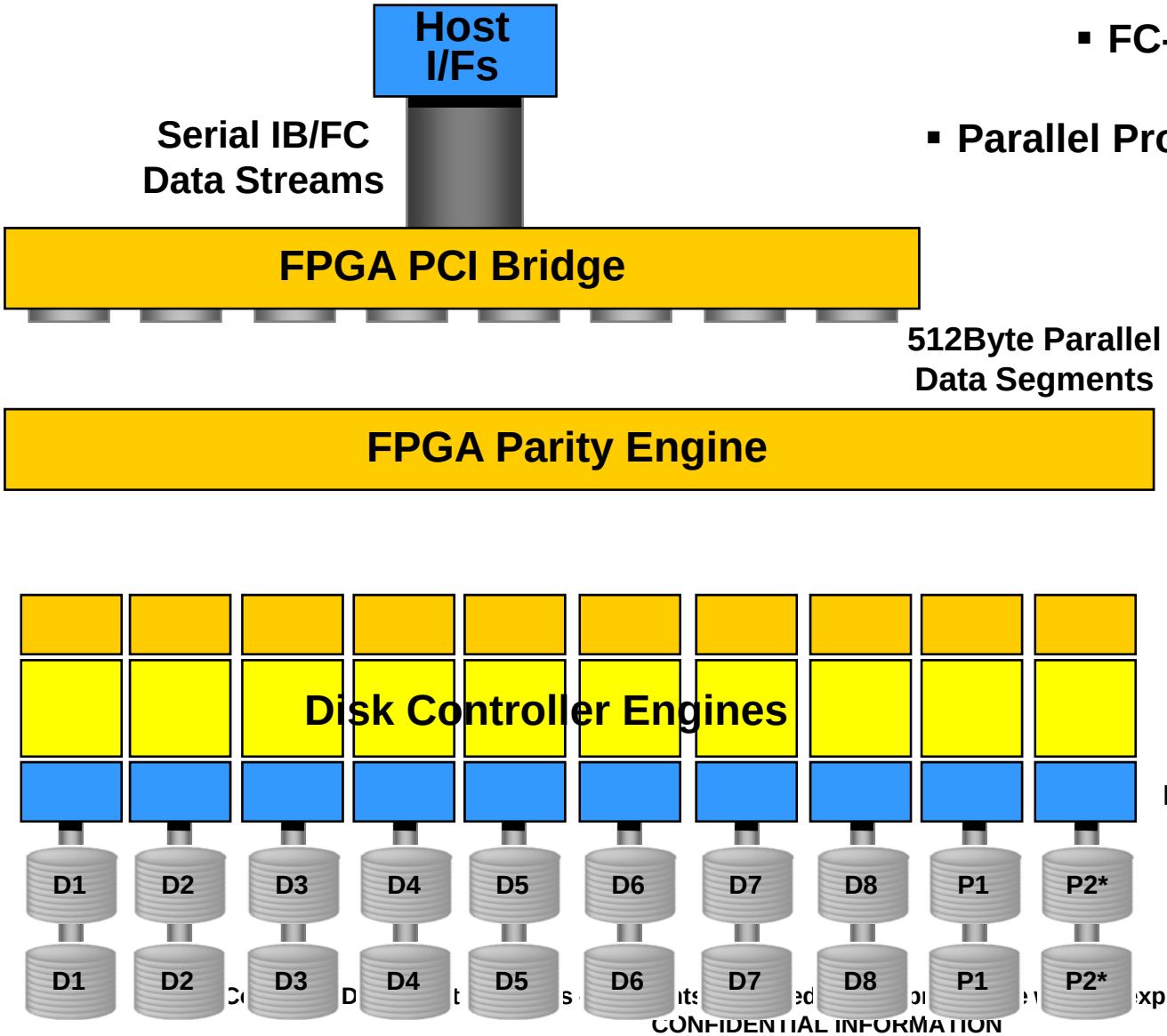


- FC-4 and/or IB Host Interface
- Parallel Processing Parity Engine
- FPGA PCI Bridge
  - Generates 512B Parallel Segments
- FPGA Parity Engine
  - Generates One or Optionally Two Parity Segments Synchronously
- Disk Controller Engines
  - Queue Command Re-Ordering in Queue Cache
  - Vertical Striping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

C D t s nts ed or e v xpress written permission
CONFIDENTIAL INFORMATION

Data Flow, From Disk

DataDirect NETWORKS
performance, capacity and innovation



- 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

express written permission
CONFIDENTIAL INFORMATION

SATAssure Data Integrity

DataDirect NETWORKS
performance, capacity and innovation

- **SATAssure makes SATA drives behave like more expensive enterprise-class drives**
  - S2A hardware enables SATAssure software to verify all data read from the disks
  - 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)

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Agenda

DataDirect NETWORKS
performance, capacity and innovation

- Storage Challenges
- Parallel Storage Architecture
- S2A 9900
- Future Requirements

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

S2A 9900 Hardware Specifications

DataDirect NETWORKS
performance, capacity and innovation

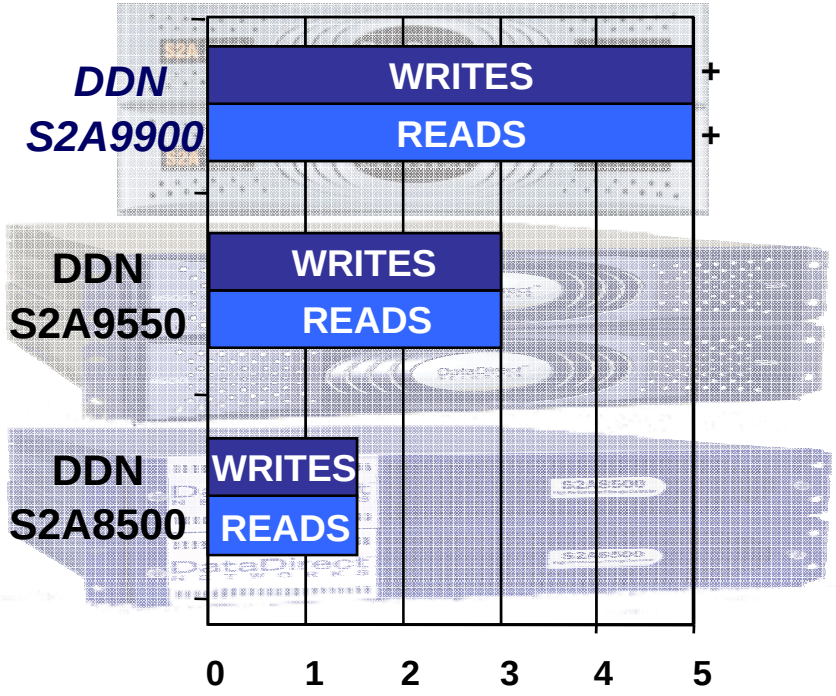
Specification	S2A9900 Couplet	S2A9550 Couplet
Supported Disk Technology	**SAS** & SATA	FibreChannel & SATA
RAID Parity Protection	**RAID6 8+2 Only**	RAID3 (8+1+1), RAID6 8+2
Sustained Throughput	**5.6GB/s – 6.0GB/s**	2.4 GB/s – 2.8GB/s
Maximum Cache	**5.0 GB ECC Protected**	2.5GB RAID Protected
Minimum Cache	**2.5 GB ECC Protected**	2.5GB RAID Protected
Disk Side Ports	**20 x SAS 4 Lane (19.2GB/s)**	20 x FC-2 (3.2GB/s)
Host Side FC Ports	**8 x IB 4x DDR or 8 x FC-8** **With File Service Capability**	8 x FC-4 or 8 x IB 4x
Dimensions	7 x 19 x 2**8** 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

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

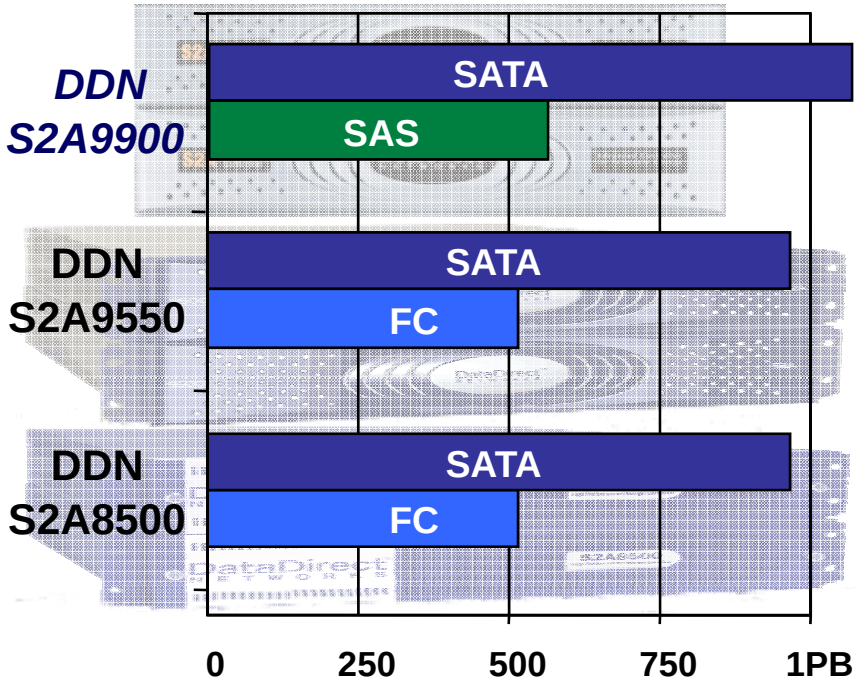
Performance & Capacity Scalability

DataDirect NETWORKS
performance, capacity and innovation

Performance, GB/sec



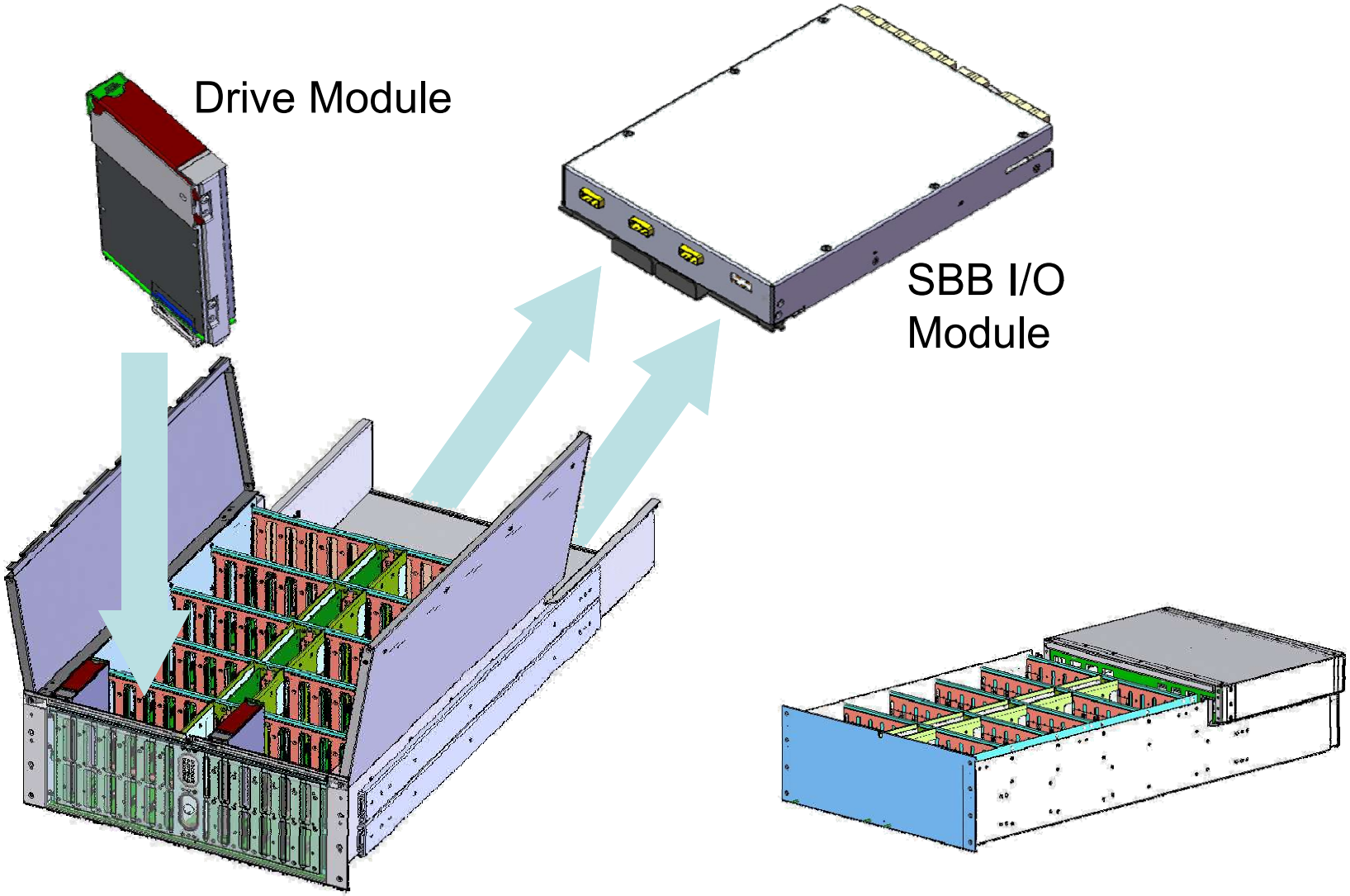
Raw Capacity, TBs



Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

StorageScaler 6000 Enclosure

DataDirect NETWORKS
performance, capacity and innovation

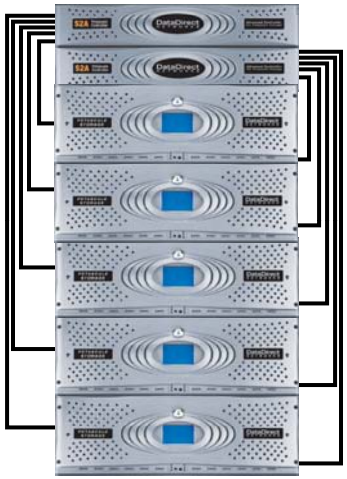


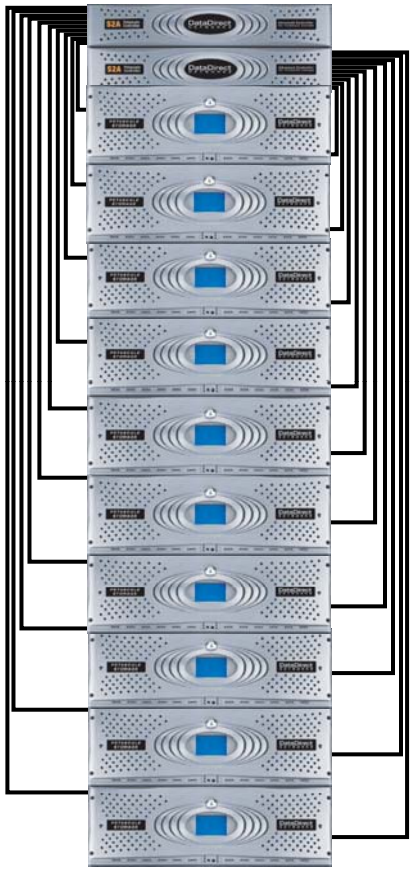
Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

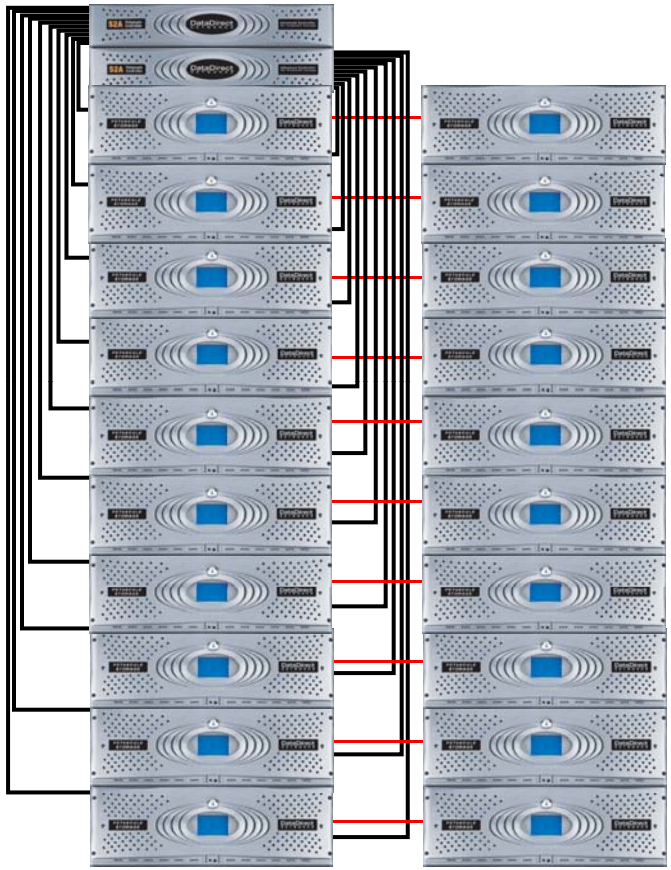
S2A 9900 Capacity

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

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Agenda

DataDirect NETWORKS
performance, capacity and innovation

- Storage Challenges
- Parallel Storage Architecture
- S2A 9900
- Future Requirements

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Future Requirements

DataDirect NETWORKS
performance, capacity and innovation

Storage Challenges

- Data transfer rates will range to TBs/s
- Drive transfer rates will not exceed 120 MB/s
- Average seek times for SAS will remain at 3mS
- Average seek times for SATA will remain at 11ms
- Any random activity greatly diminishes the effective transfer rate

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

Future Solutions

DataDirect NETWORKS
performance, capacity and innovation

Evolving Technology

- Faster physical transfer architectures such as IB 32x
- File systems with better transfer aggregation
  - GPFS at 4MB
  - Lustre at 2MB
- Storage integrated with file services to enable intelligent data transfer reordering
- Commodity storage elements are getting faster, better, cheaper, and lower in power comsumption
  - SSDs are larger and more reliable and can be utilized in the same architecture
  - Smaller form factor disks are larger, cheaper, and more reliable
  - SRAM costs are decreasing with finer pitch implementations

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION

DataDirect NETWORKS
performance, capacity and innovation

Architecting Storage for Petascale Clusters

Key Challenges in Modeling and Simulation
September 25, 2007

Dave Fellinger, CTO
dfellinger@datadirectnet.com

Copyright DataDirect Networks - All Rights Reserved - Not reproducible without express written permission
CONFIDENTIAL INFORMATION