Lustre User Group 2009

Lustre as the Core of a Data Centric Best Practice HPC Workflow

Bob Murphy, Open Storage
Sun Microsystems

A recent conversation

"It's not computation I'm worried about, that's been solved, it's storing data, accessing it, and moving it around."

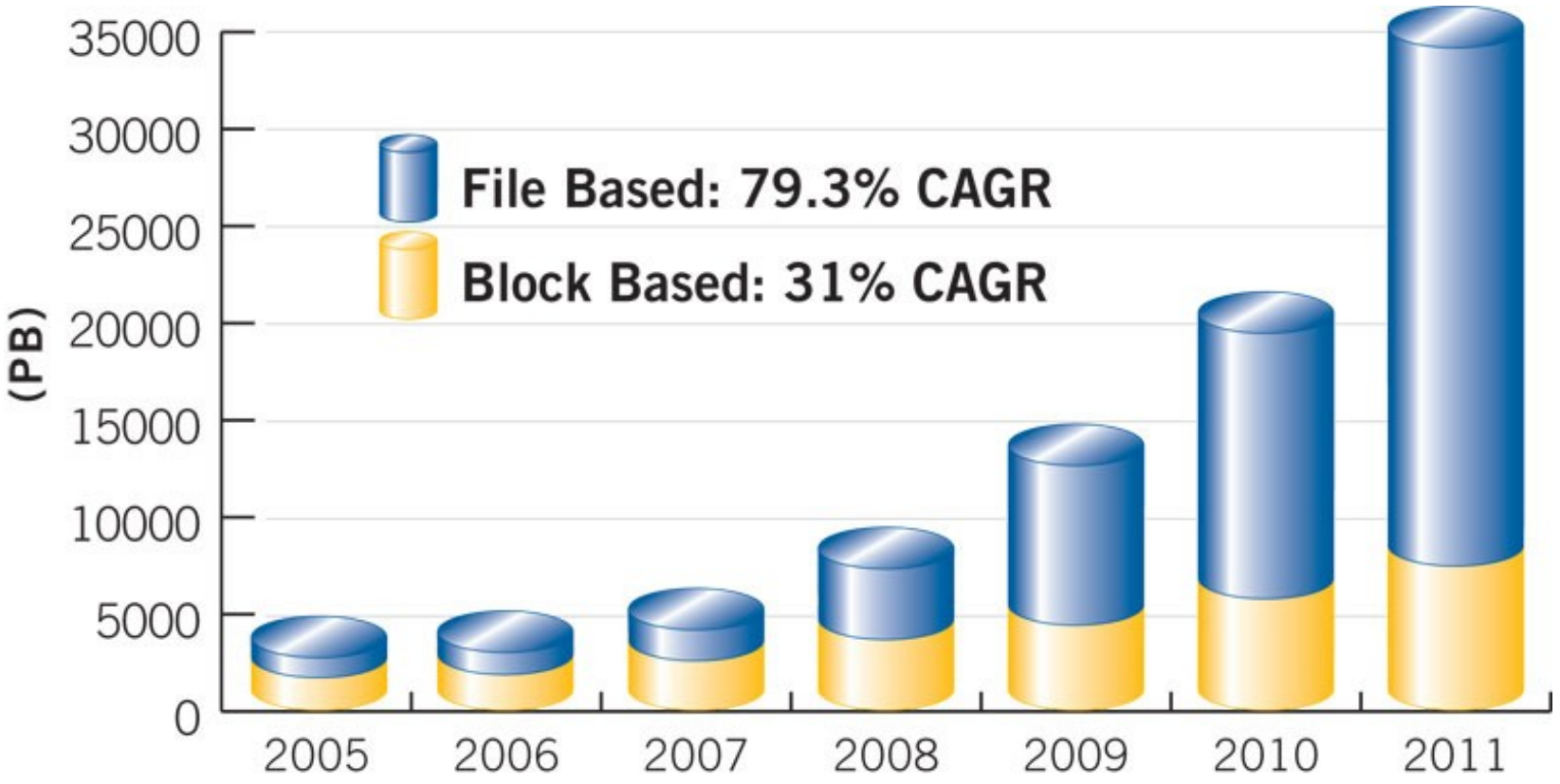
- Rico Magsipoc, Chief Technology Officer for the Laboratory of Neuro Imaging, UCLA

Processing power doubles every 2 years, but not for HPC...

Peak FLOPS Per Rack

In 2005:	84 cores	500 GFLOPS
In 2009:	768 cores	9 TFLOPS
In 2011:	1,536 cores	172 TFLOPS

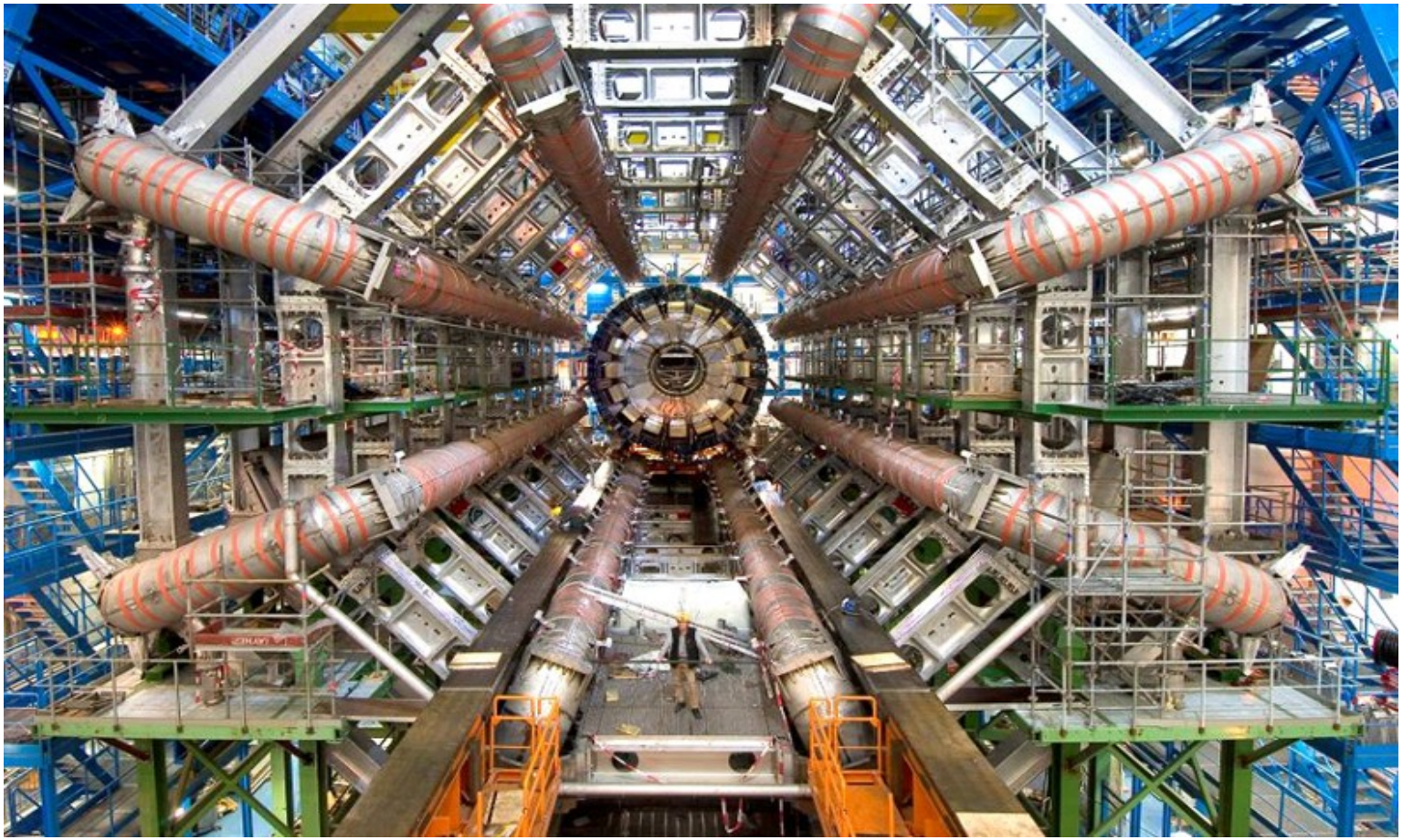
Data Explosion



Source: IDC

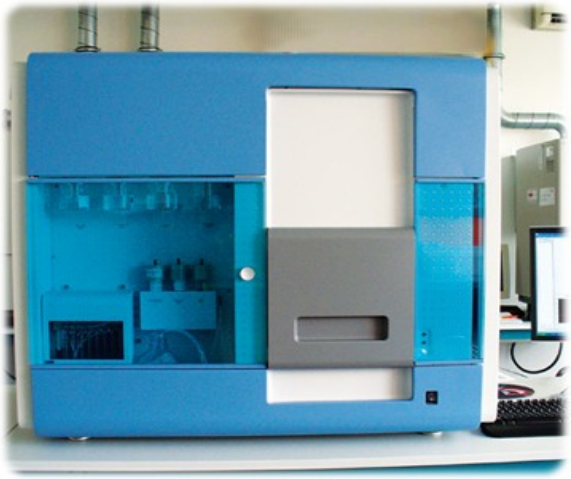
Factors driving the HPC data tsunami

- Increased sensor resolution
  -Cameras, Confocal Microscopes, CT Scanners, Sequencers, etc



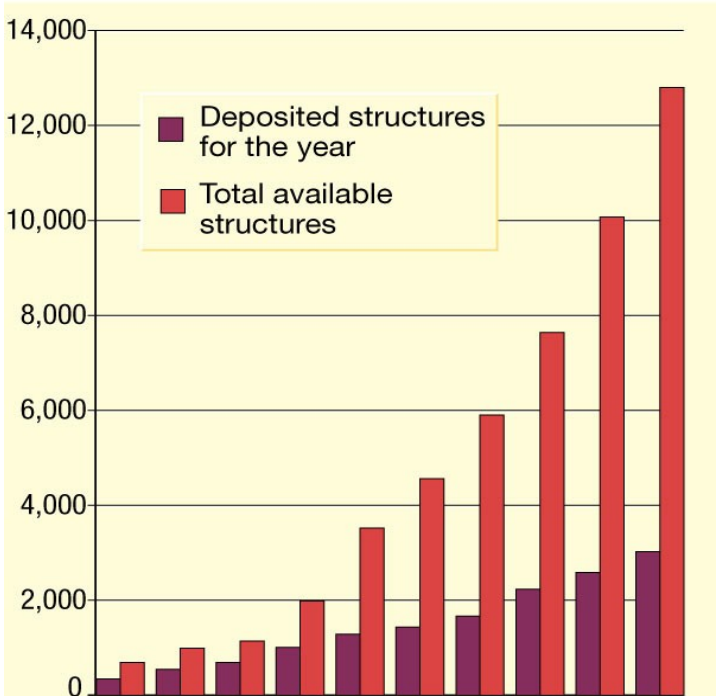
Like computers themselves, the devices creating these datasets have proliferated

- From 1 per institution, to 1 per department, to 1 per lab, to 1 per investigator



- Terabyte-sized datasets from a single experiment are now routine
- 1TB/day/investigator
- Plus more - and higher resolution - simulations per unit time

Oh, all the previously accrued data needs to be stored too...

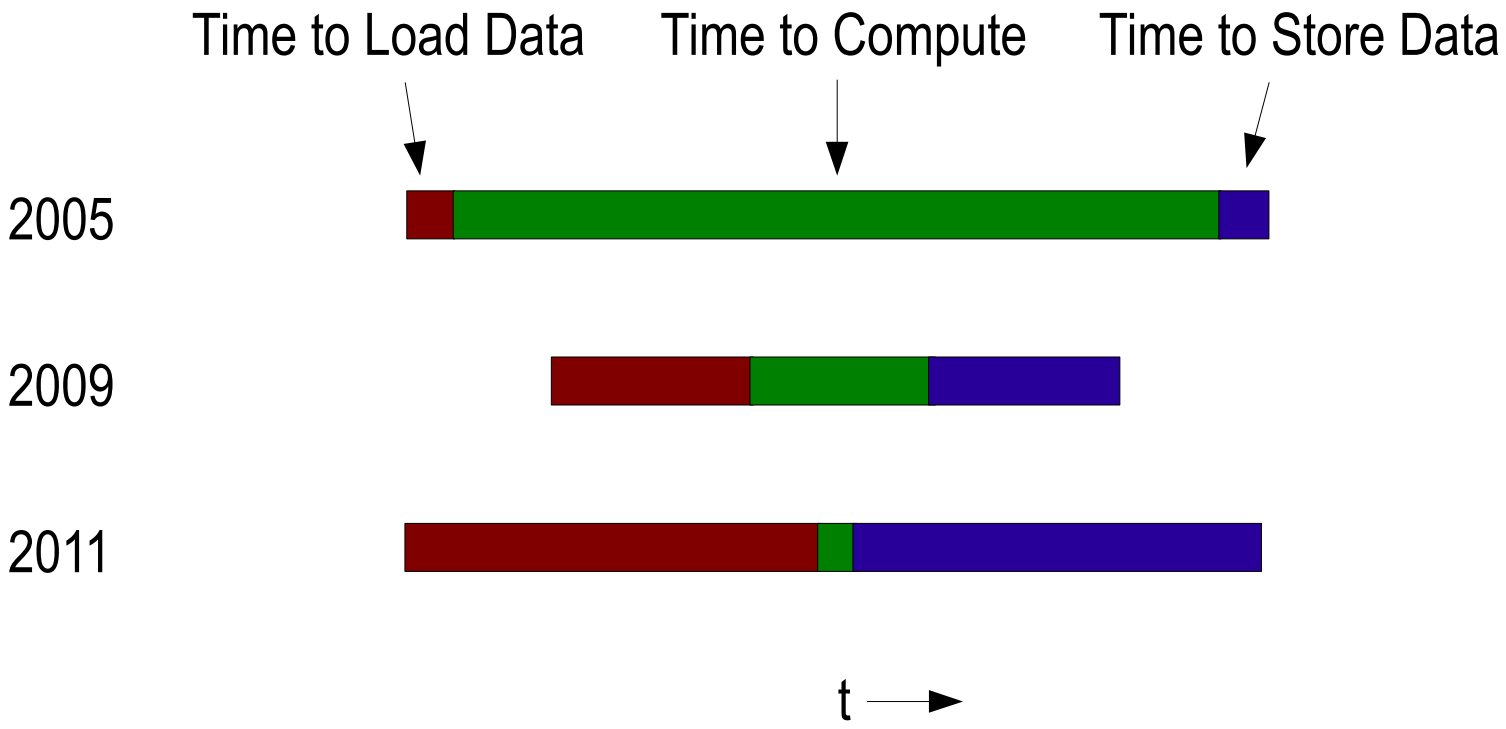


Conclusion

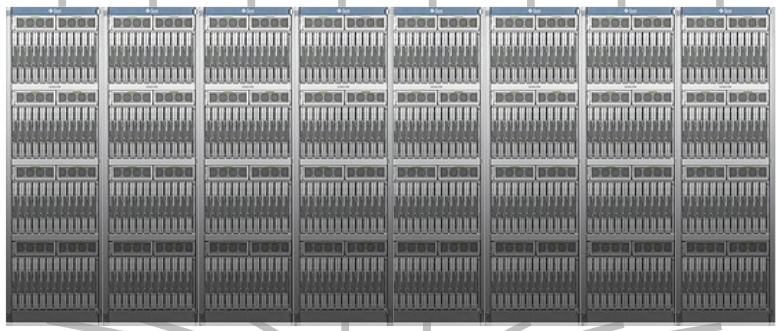
- You can only compute the data as fast as you can move it

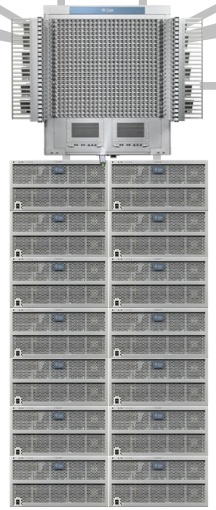
- HPC workflows will become performance bound by the speed of the storage system

So data access time will dominate HPC workflow



Sun end-to-end infrastructure optimized to accelerate data centric HPC workflows





Network Storage

Sun Storage 7000 Unified Storage System

High Availability, Manageability, Shared Access

- Home Directories, Application Code
- Input Data, Results Files

Parallel Storage

Sun Lustre Storage System

High Performance Parallel File System

- Ongoing Computation

Archive Storage

Sun StorageTek Tape Archives

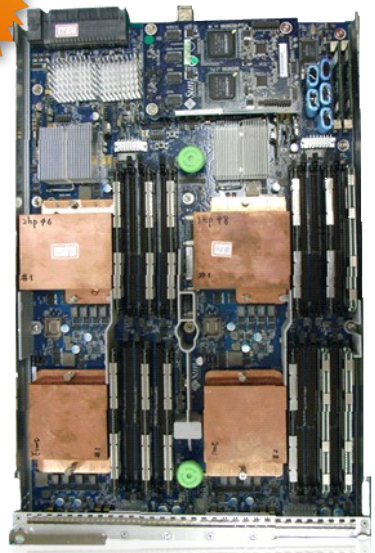
Economic, Green, Long Term Data Retention

- Protection of IP Assets
- SAM Storage Archive Manager HSM

High Bandwidth, High Memory Petaflop-Scale HPC Blade System

- **Sun Blade X6275**
  - New, high density, dual node blade server
  - 2X2 Nehalem-EP 4-core CPUs
  - 2X12 DDR3 DIMM Slots, up to 192GB (8GB DIMMs)
  - 2x1 Sun Flash Module (24GB SATA)
  - 2x1 Dual Port QDR IB HCA
- **Sun Blade 6048 System**
  - 768 Nehalem-EP cores, 9TF, up to 9.2TB memory
  - 96 X Gigabit Ethernet via NEM
  - 96 X PCIe 2.0 Express Module x 8 interfaces
- **Highest Compute Density in the Industry**

New



Sun Blade X6275

Sun Blade 6048 System

Breakaway application performance

Compute-Intensive MCAE Applications Run Faster, with Less Power, Less Heat, Less Space with Integrated QDR Infiniband

Sun Blade X6275 48 Node Cluster

Radioss Single Car Crash Code

Performance	Space	Perf/Watt
3.2x	50%	3.2x

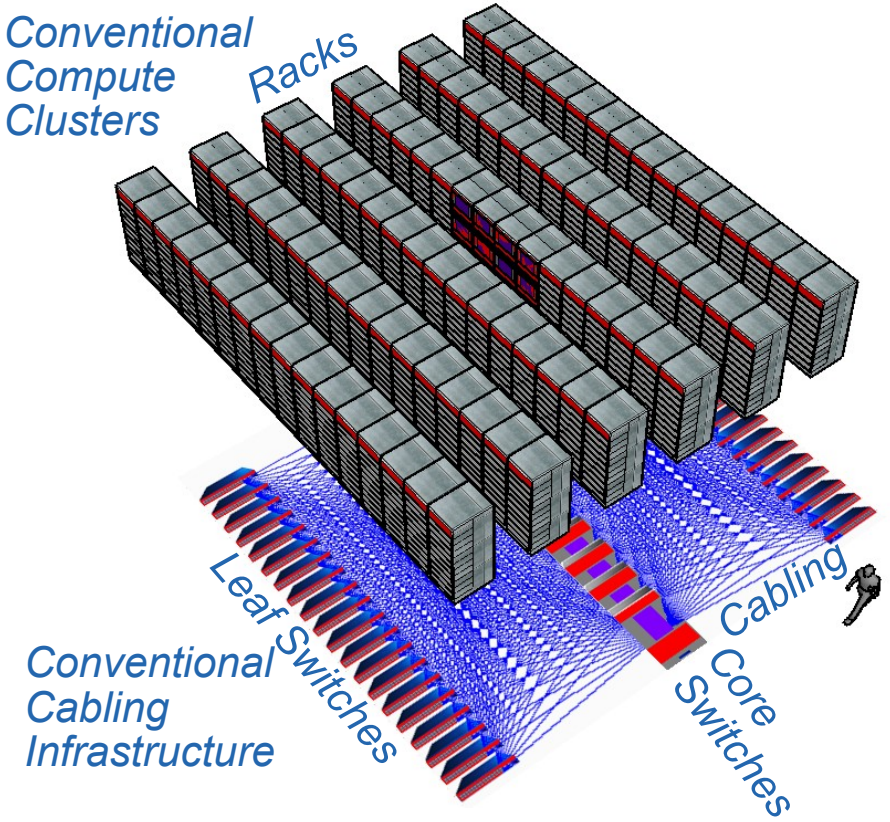
HP BL460c

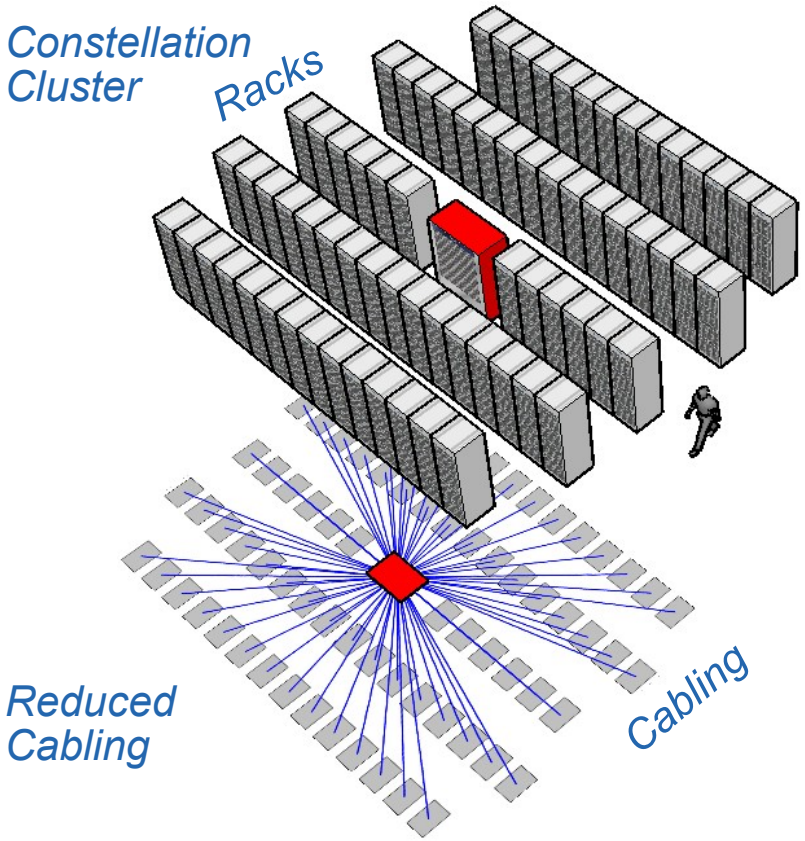
2 x 3GHz Xeon E5456 Processors, SUSE Linux

Sun Data Center 3456 IB Switch

- World's Largest Infiniband Core Switch
- Replaces 300 discrete InfiniBand switches and thousands of cables
- Unparalleled scalability and dramatically simplifying cable management
- Most economical InfiniBand cost/port
- Prelude to “Project M9” doubling InfiniBand performance

Efficient Petascale Architecture





Conventional IB Fabric

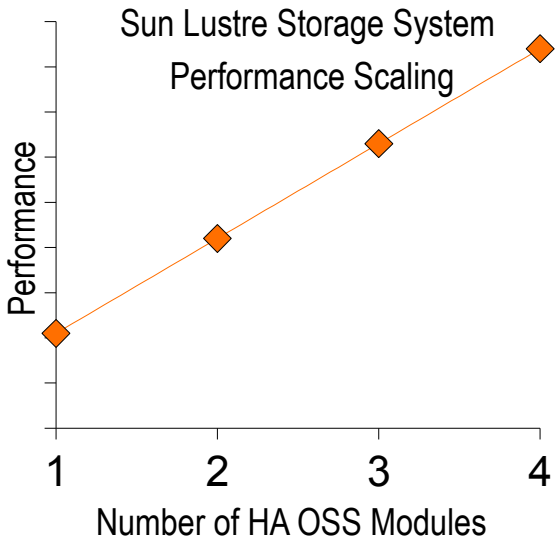
- 300 Switches: 288 Leaf + 12 core
- 6912 Cables: 3456 HCA + 3456 trunking
- 92 Racks

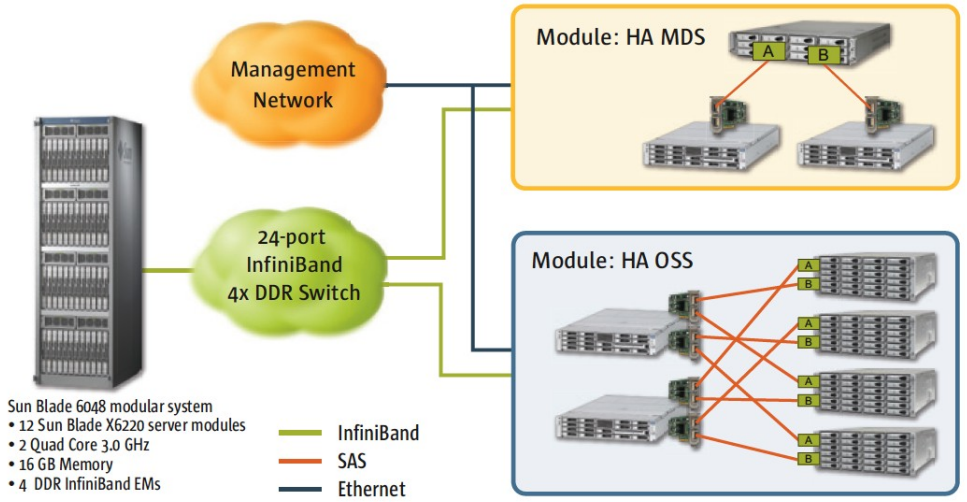
Sun Datacenter Switch 3456

- 1 Core Switch *300:1 Reduction*
- 1152 Cables *6:1 Reduction*
- 74 Racks *20% Smaller footprint*

Sun Lustre Storage System

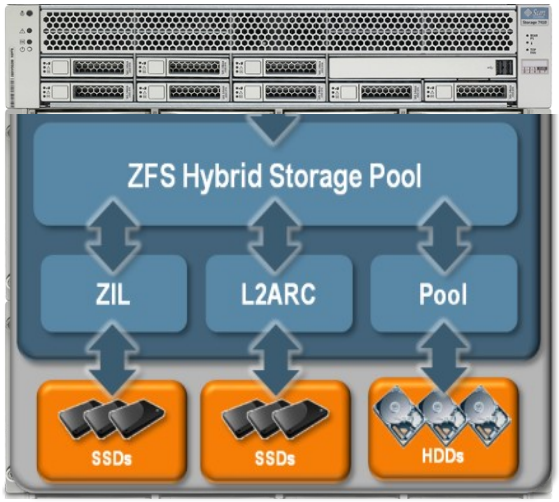
- Scales to over 100 GBs/sec and capacity to petabytes
- Accelerated deployment though pre-defined modules
- Easy to size and architect configurations
- Automated configuration and standard install process
- Compelling price/performance with Sun components
- Deployed by Sun Professional Services to ensure success





Sun Storage 7410 Unified Storage System

Breakthrough NAS performance at low cost and power with Hybrid Storage Pool

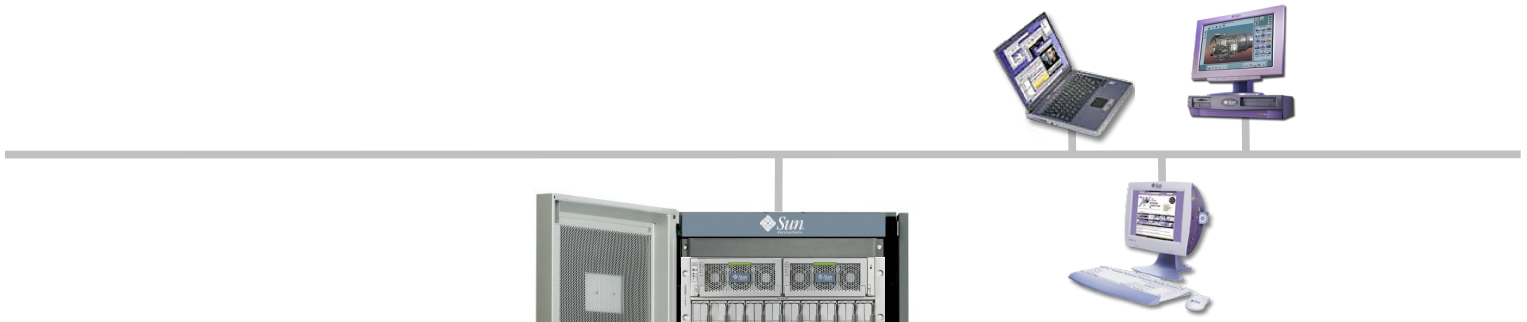


Sun Storage 7410 Unified Storage System

- Delivers over 1GB/s throughput
- 2GB/s from DRAM
- 280K IOPS
- At ¼ the price of competitive systems
- And 40% of the power consumption

- ZFS determines data access patterns and stores frequently accessed data in DRAM and FLASH
- Bundles IO into sequential lazy writes for more efficient use of low cost SATA mechanical disks

Scratch storage for a small HPC cluster



Pre-configured Sun MCAE Compute Cluster

Sun Storage 7410 Unified Storage System

High Availability, Manageability, Shared Access

- Home Directories, Application Code
- Input Data, Results Files

Ongoing Computation Scratch storage

7410 scratch performance good for single rack MCAE clusters

Reduces install/operational complexity

Expands market for HPC clusters

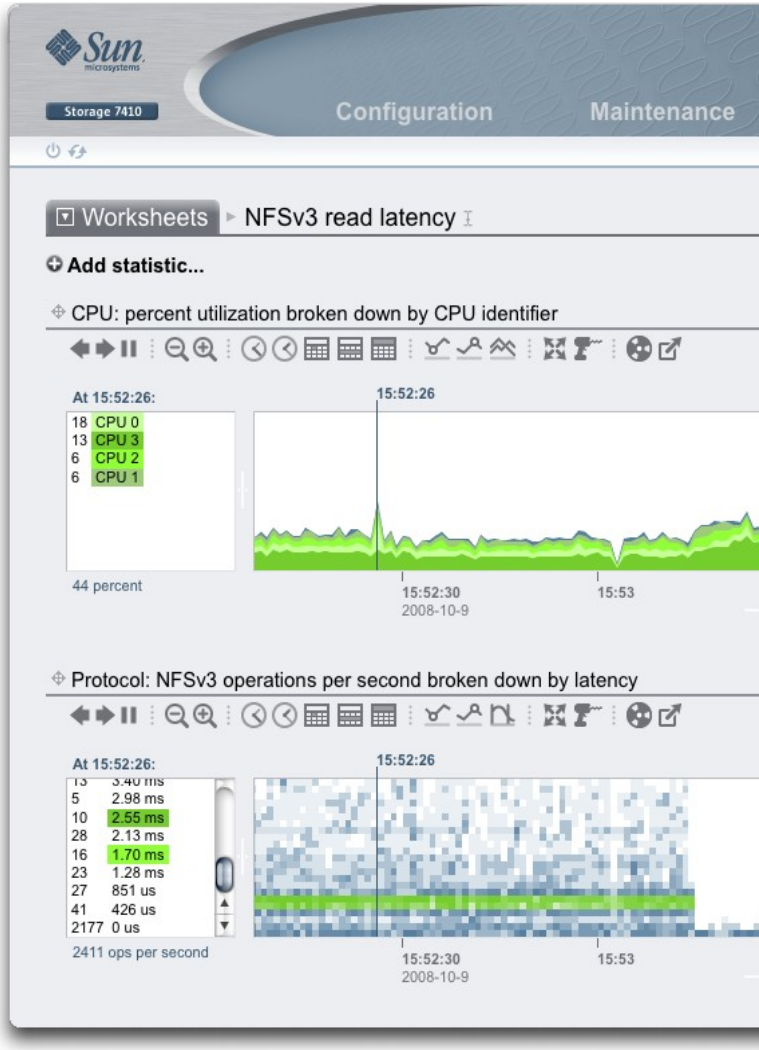
-Many organizations “stuck on the desktop”

-Don't have expertise to roll their own cluster

“D-NAS”

Unprecedented Dtrace Storage Analytics

- Automatic real-time visualization of application and storage related workloads
- Supports multiple simultaneous application and workload analysis in real-time
- Analysis can be saved, exported and replayed for further analysis
- Rapidly diagnose and resolve issues
  - > How many Ops/Sec?
  - > What services are active?
  - > Which applications/users are causing performance issues?

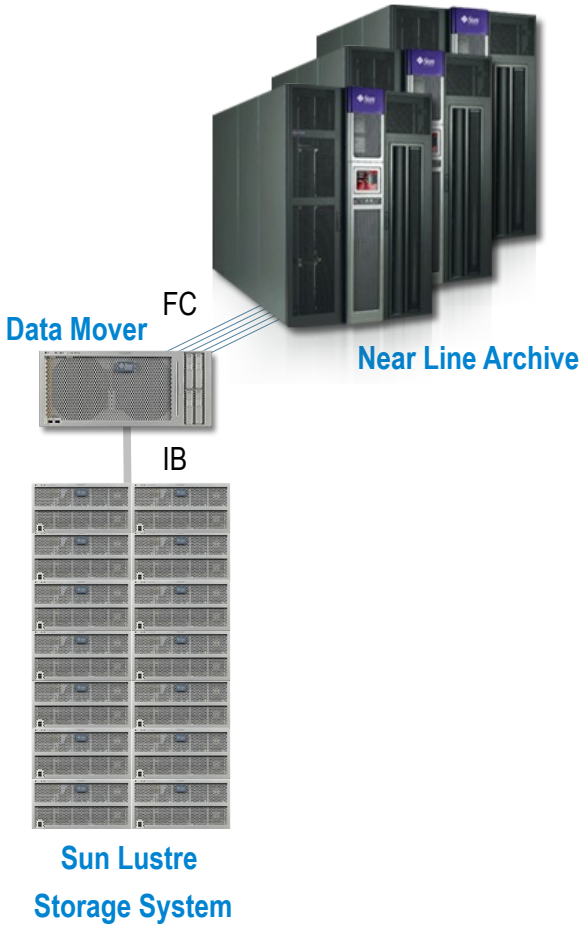


Video

Sun StorageTek Tape Archive

The greenest storage on the planet

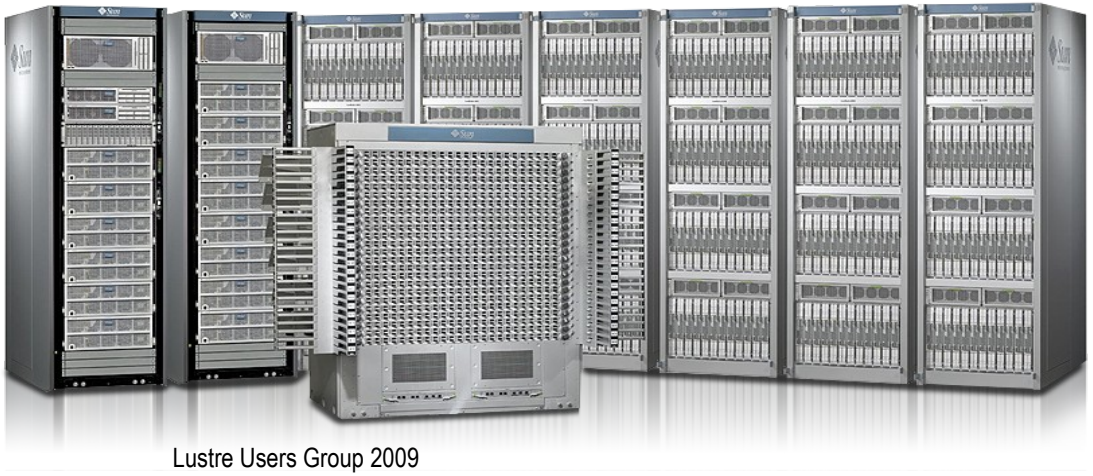
- Provides a massive on-line/near-line repository – up to 70PB with little power or heat
- SAM Storage Archive Manager maintains active data on faster disk and automatically migrates inactive data to secondary storage
- All data appears locally and is equally accessible to multiple users and applications
- Stores data in open formats (TAR) allowing technology refresh and avoiding vendor lock-in
- Tape shelf life of about 30 years
- Manage HPC data at the right cost, with the right performance, on the right media - and deliver it to applications at the right time



Putting it all together, the Sun Constellation System

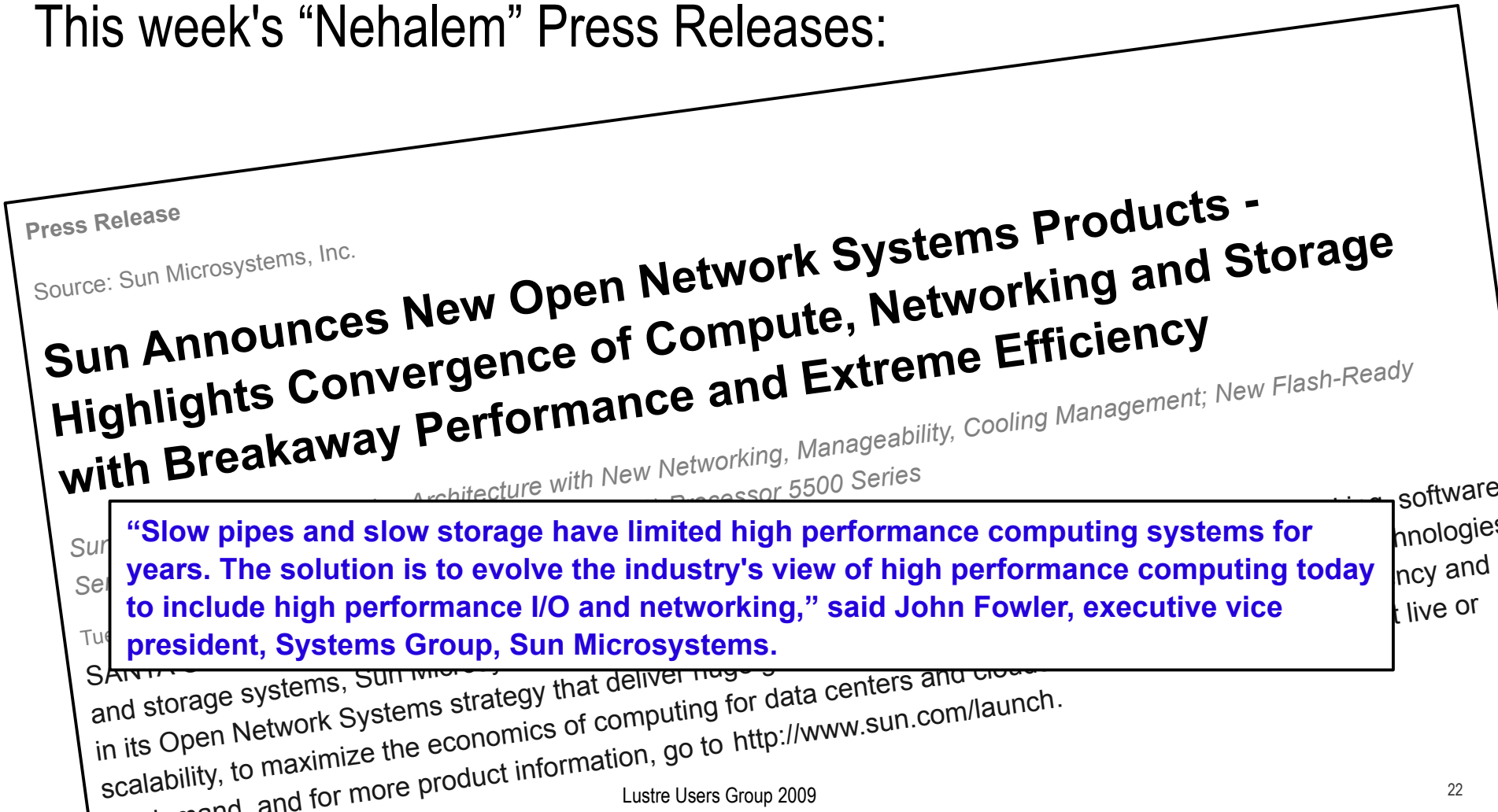
- An open systems super computer designed to scale from Departmental Cluster to Petascale
- Optimized compute, storage, networking and software technologies - and service - delivered as an integrated product
- Integrated connectivity and management to reduce start-up, development and operational complexity
- Technical innovation resulting in fewer components and high efficiency systems in a tightly integrated solution

Easiest Path to Peta-Scale



More than any other computer company, Sun gets it

This week's “Nehalem” Press Releases:



Press Release

Source: Sun Microsystems, Inc.

Sun Announces New Open Network Systems Products - Highlights Convergence of Compute, Networking and Storage with Breakaway Performance and Extreme Efficiency

Architecture with New Networking, Manageability, Cooling Management; New Flash-Ready Processor 5500 Series

“Slow pipes and slow storage have limited high performance computing systems for years. The solution is to evolve the industry's view of high performance computing today to include high performance I/O and networking,” said John Fowler, executive vice president, Systems Group, Sun Microsystems.

and storage systems, Sun
in its Open Network Systems strategy that deliver
scalability, to maximize the economics of computing for data centers and
and for more product information, go to http://www.sun.com/launch.

Summary

- Data is taking over HPC
- Sun has a wide array of complementary IP to Lustre that can cope with that
- Sun is leading the industry in Data-Centric HPC by delivering differentiated customer value in balanced HPC systems, including compute, storage, and networking
- Sun can uniquely deliver a complete end-to-end solution
- You can also integrate parts of that solution into your existing workflow
- Don't yell at your JBODs

THANK YOU