THE SUN STORAGE AND ARCHIVE SOLUTION FOR HPC

The Right Data, in the Right Place, at the Right Time

José Martins
Storage Practice
Sun Microsystems

Agenda

- Sun's strategy and commitment to the HPC or technical computing market
- Storage challenges we hear from HPC customers
- Addressing these challenges with the Sun Storage and Archive Solution for HPC
- Next Steps

Advancements in HPC

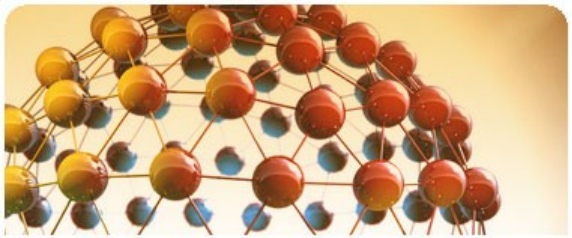
HPC is Now Everywhere

FINANCE

Portfolio risk analysis and optimization, derivatives pricing, fraud detection....

ENGINEERING

Simulate mechanical and electronic systems before building costly prototypes



RESEARCH/DEVELOPMENT

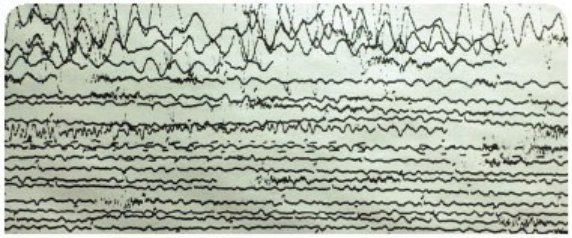
Simulate the effects and prognosis of chemotherapy, radiation, and surgery

ENTERTAINMENT

Render photo-realistic scenes, customer materials, product illustrations and videos

WEATHER AND CLIMATE

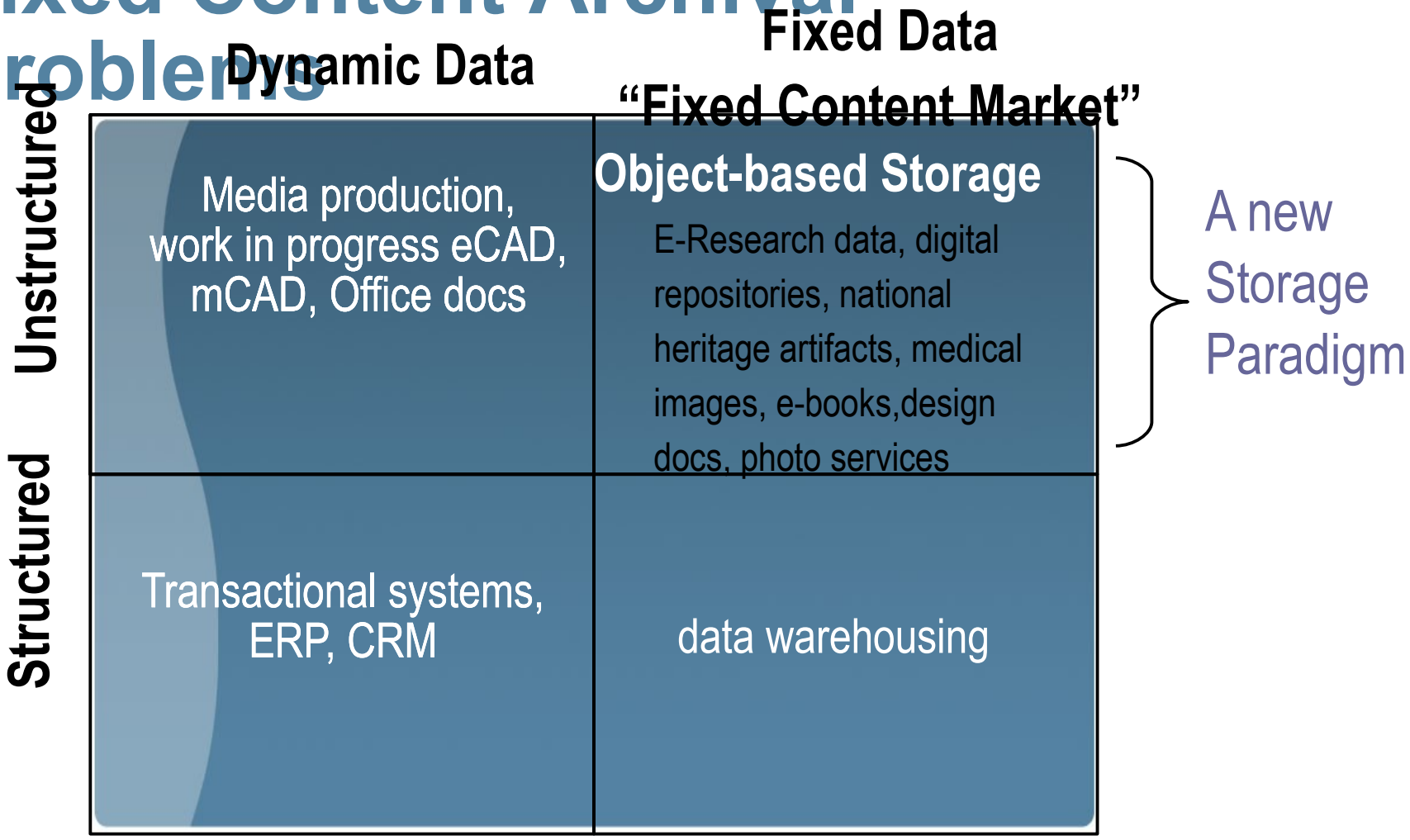
Weather and climate forecasting



OIL AND GAS

Analyze seismic data to find oil and gas and determine how to extract it in the most efficient way

The Solution for Large Scale Fixed Content Archival Problems



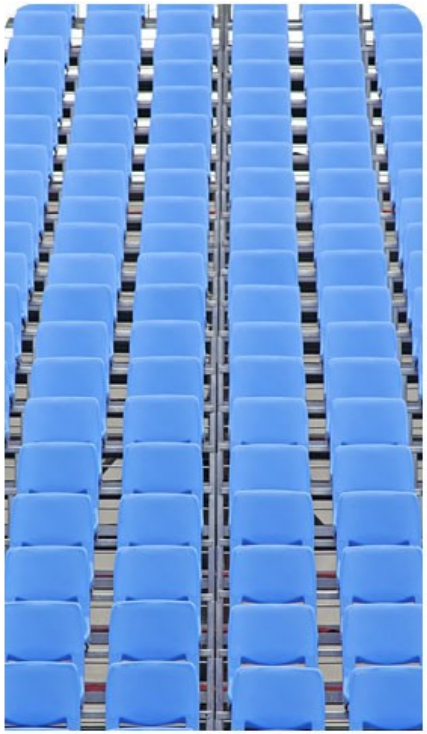
Solution Overview

What We Have Heard:

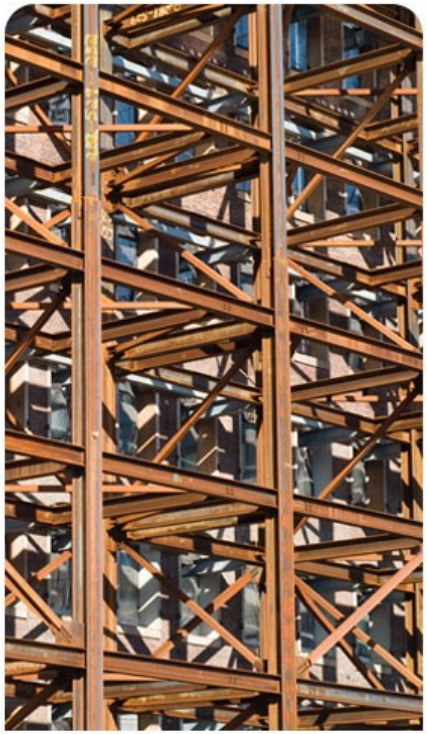
Storage Challenges for HPC

High Performance on a Fixed Budget

Explosive Growth with Limited Staff



Lack of Tuning and Customization



Limited Flexibility

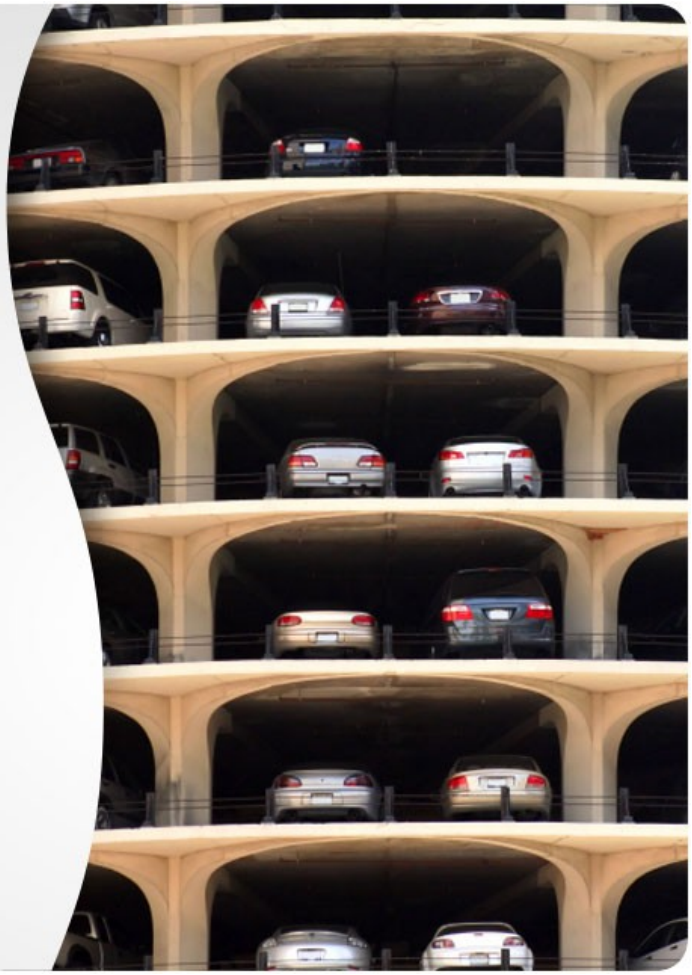
Addressing these Challenges

Sun Storage and Archive Solution for HPC

A Comprehensive Environment That:

- Efficiently addresses all your HPC storage needs by
  - > Integrating multiple storage types
  - > Automating data life cycle and workflow
- Enables breakthrough economics:
  - > Reduce three year TCO up to 44%*
  - > Reduce energy consumption up to 24%*
  - > Or increase cluster compute power up to 23%*, by reinvesting the savings in more cluster nodes

*Based on comparative analysis with typical NAS products serving a compute grid



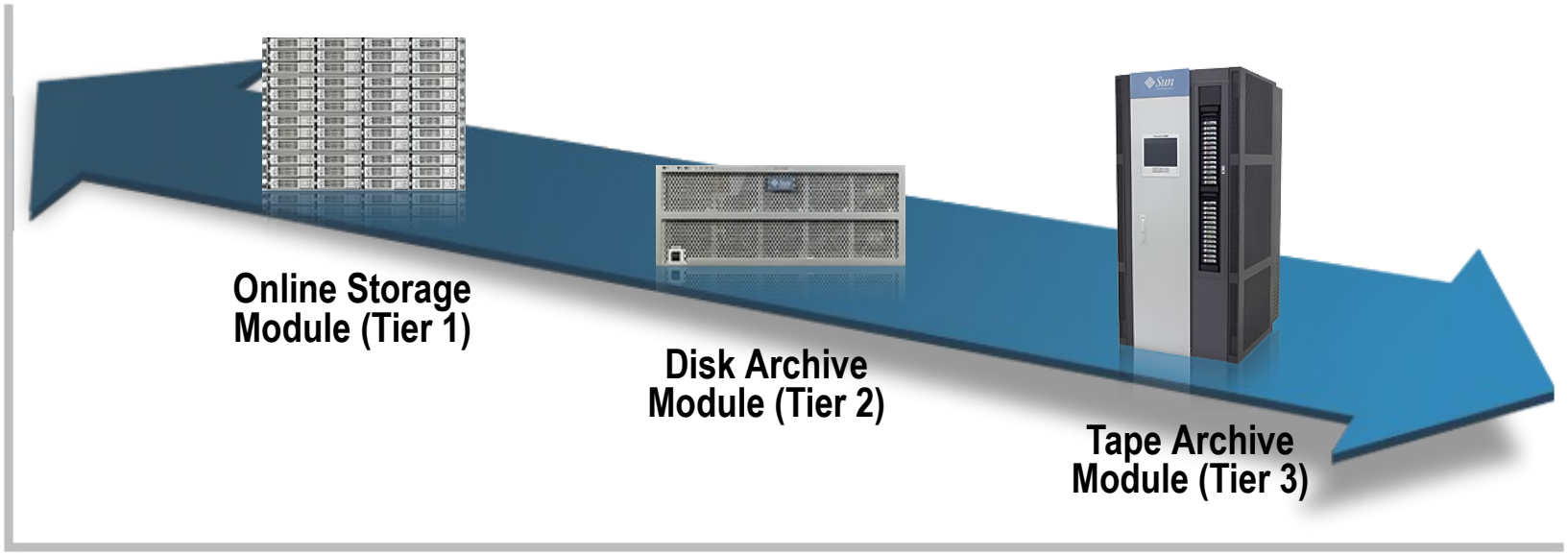
Sun Storage and Archive Solution for HPC

The Right Data, in the Right Place, at the Right Time

Key Functions

- High performance I/O for the compute cluster or grid
- Reliable and robust home directory space
- Automated migration of data to the most cost effective storage for the job
- Seamless user access to data that is moved to tape
- Continuous automated backup protection of data

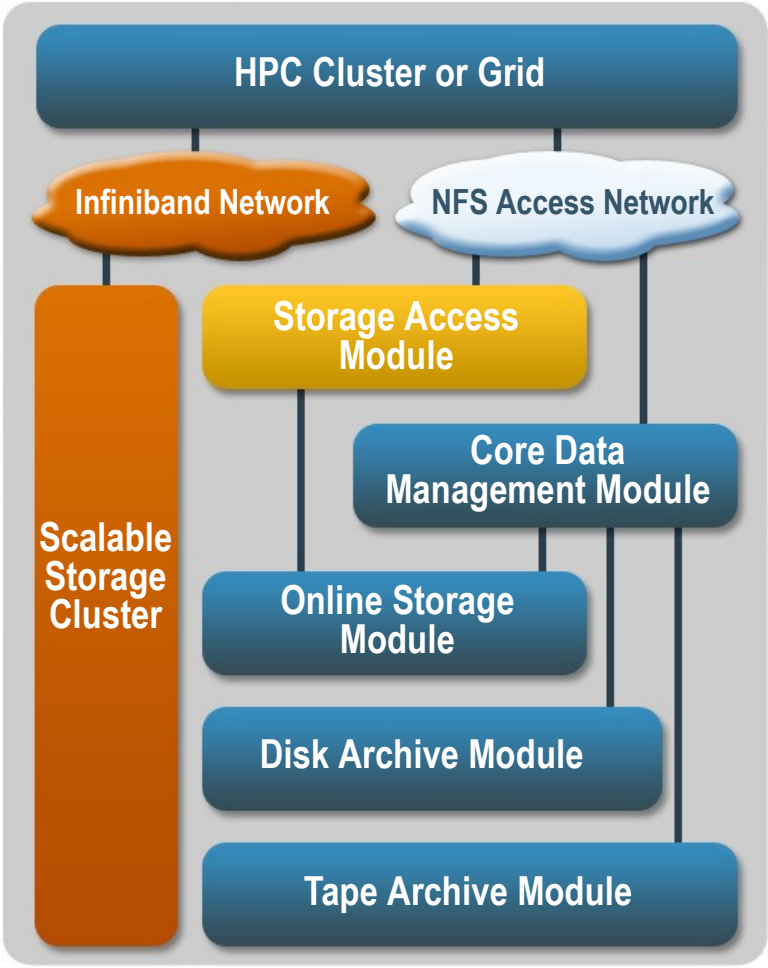
Leveraging the Most Cost Effective Asset



Closer Look

Solution Architecture and Modules

- **Core Data Management Module:** manages the environment, data locations and provides the policy interface
- **Storage Access Module (optional):** provides cluster working space or scratch space access
- **Online Storage Module:** stores cluster working data and user home directories
- **Disk Archive Module:** stores data that has been archived either by policy or by user action
- **Tape Archive Module:** stores data that has been archived on tape at a lower cost and power burden than the disk archive module
- **Sun Customer Ready Scalable Storage Cluster (optional):** very high performance scratch space module, utilizes Lustre and InfiniBand Sun storage; used in place of the Storage Access Module



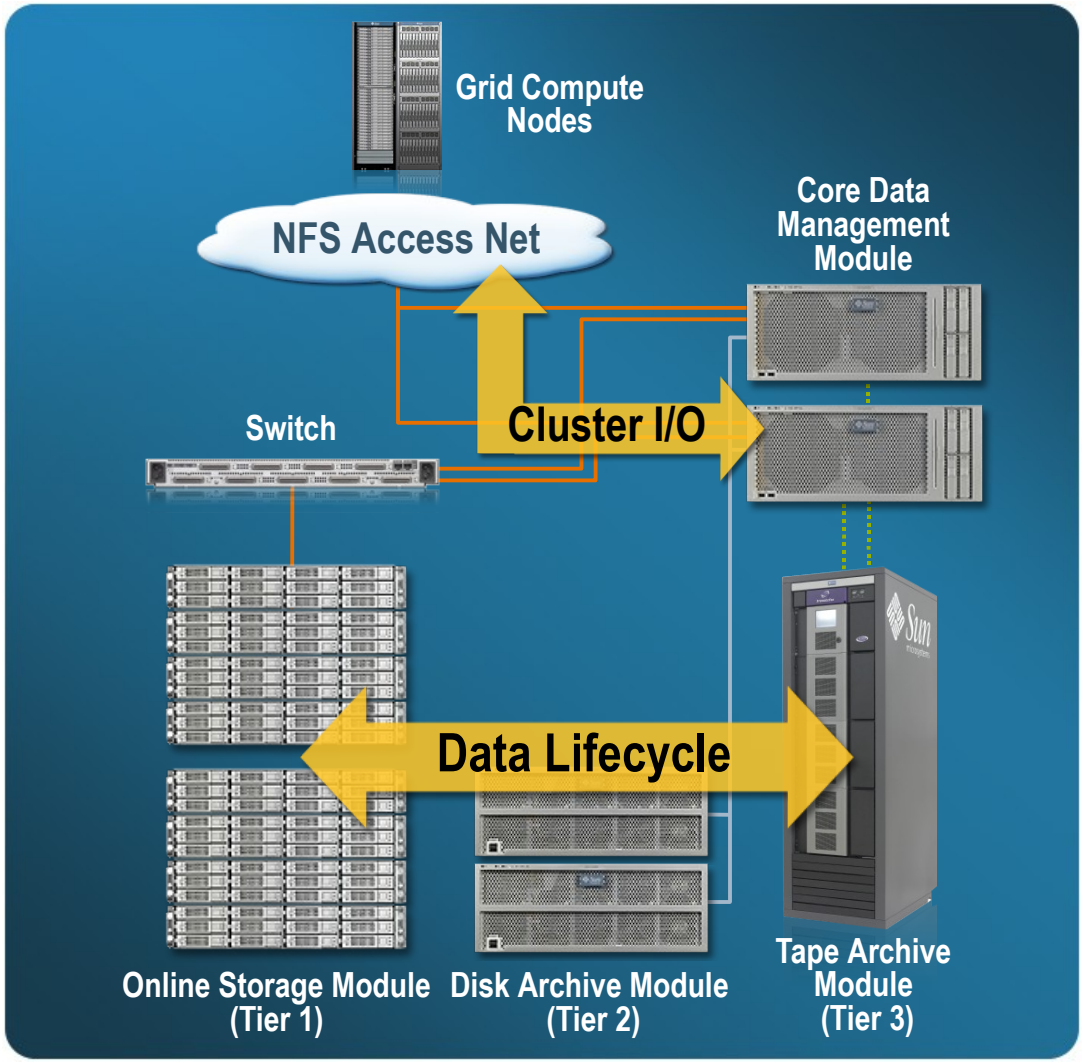
Reference Configurations

	Node Count Guidelines	Aggregate Bandwidth	Data Access	Name Space
Small	32-128 for light I/O per node	100 - 400MB/sec	NFS - Cluster I/O	Combined Scratch and Archive
	32-64 for heavy I/O per node		NFS- Home Directories	
Medium	64-512 for light I/O per node	400MB – 1GB/sec	NFS - Cluster I/O	Combined Scratch and Archive
	32-256 for heavy I/O per node		NFS- Home Directories	
Large	512-1000s for light I/O per node	1 - 50+ GB/sec	IB - Cluster I/O	Separate Scratch and Archive connected by data mover
	256-1000s for heavy I/O per node		NFS- Home Directories	

NOTE – sizing information is provided as a starting guideline; work with your Sun systems engineer to configure for your specific needs

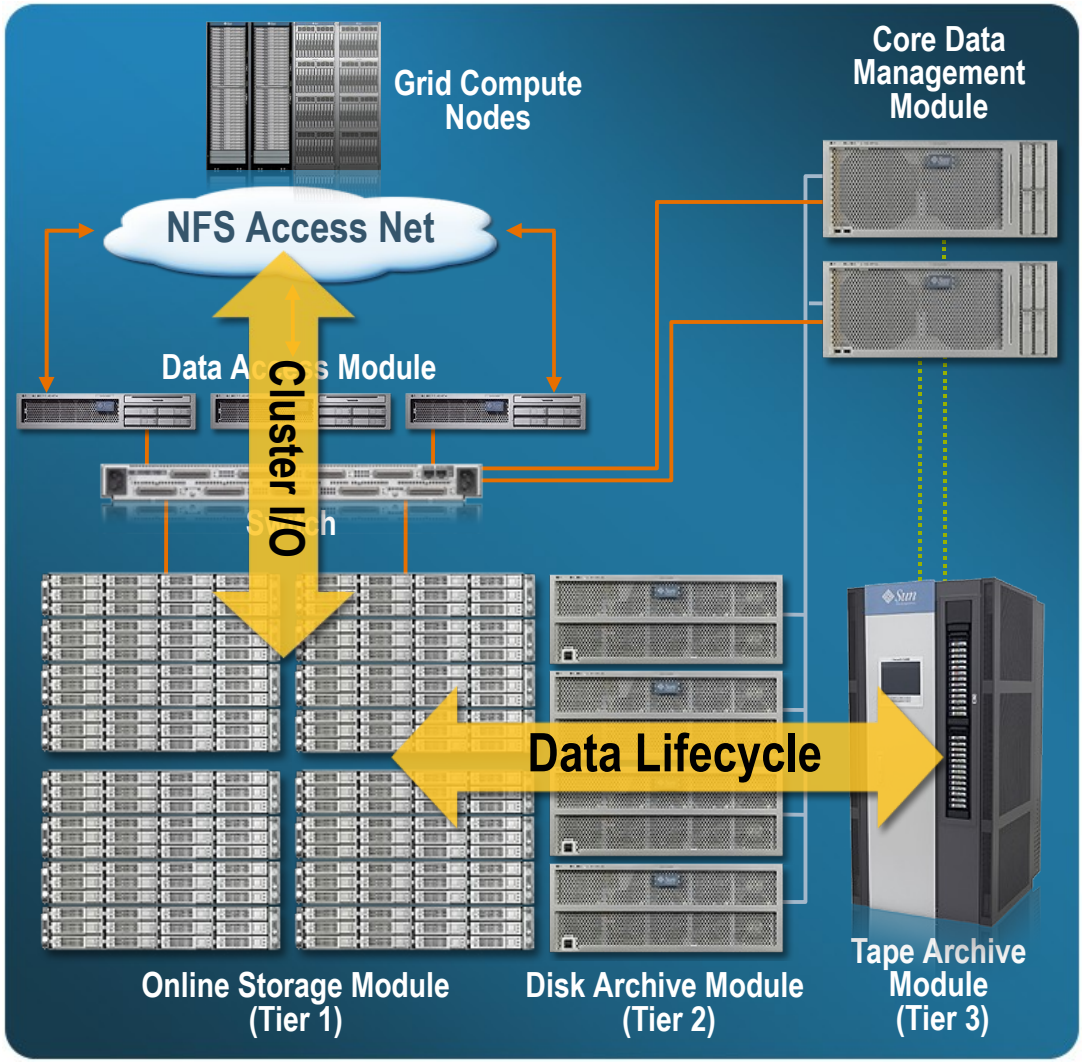
Small Configuration

- Departmental HPC needs
- Typically 32 -128 nodes
- Provides up to ~ 400 MB/sec
- HPC cluster and users are served via NFS (10GE or GE)
- Core modules are clustered via QFS, add more to scale performance
- Core module controls data movement between the Online and Archive modules



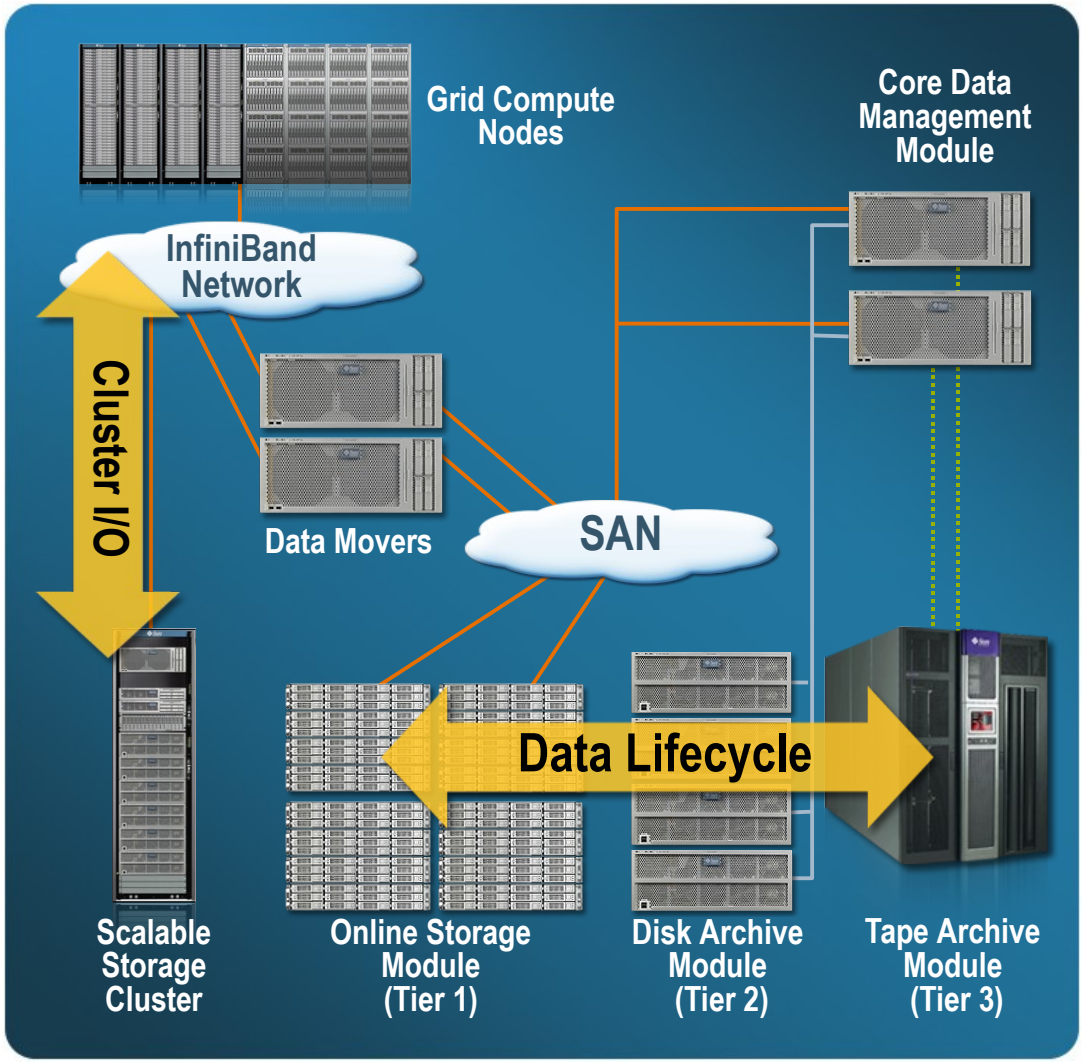
Medium Configuration

- Divisional HPC needs
- Typically 64-512 nodes
- Scales ~ 1 GB/sec
- HPC cluster or grid is served via NFS by the Data access modules
- Data Access modules are clustered via QFS, add more to scale performance
- Core module controls data movement between the Online and Archive modules

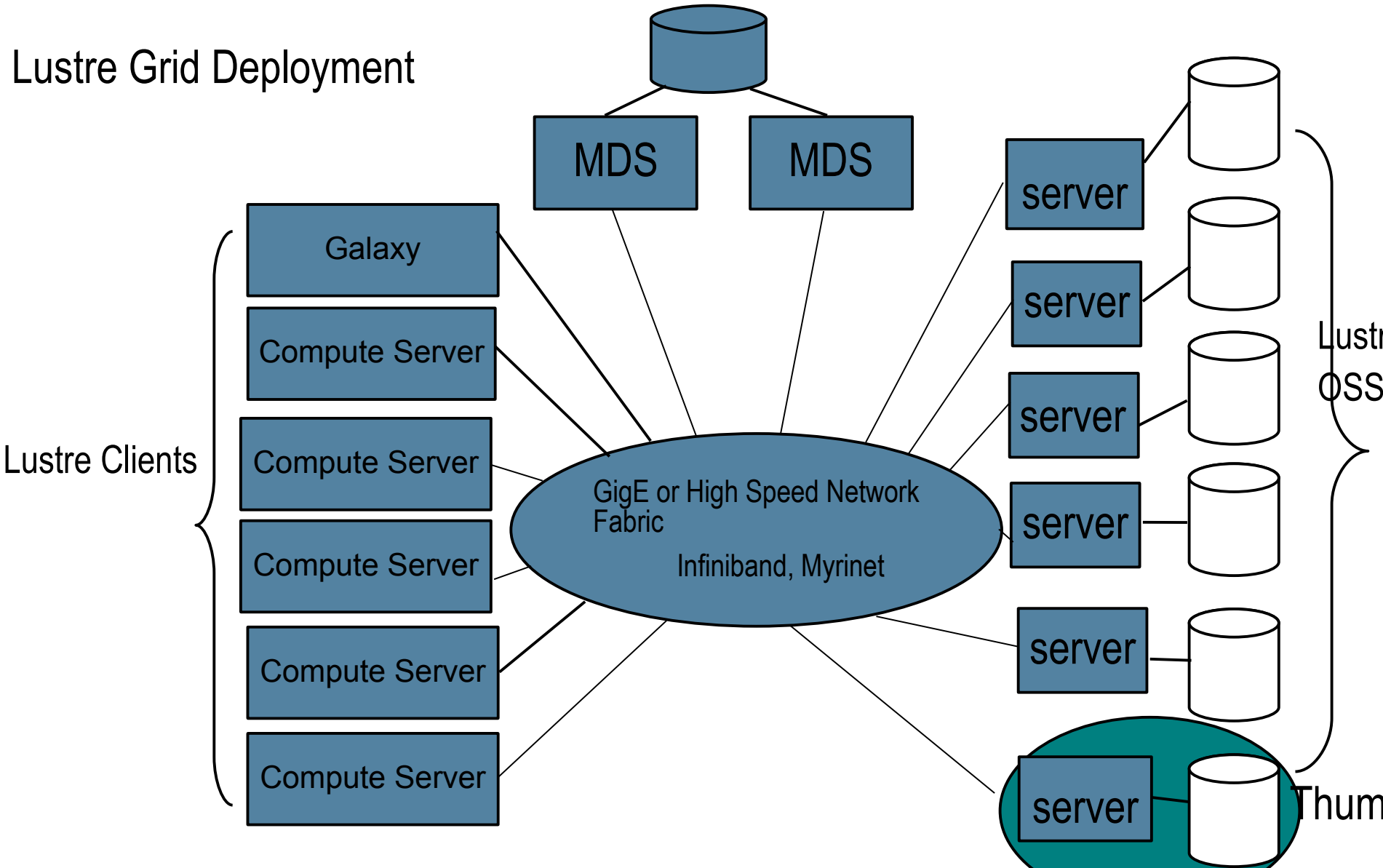


Large Configuration

- Large scale HPC needs
- Up to 1000s of nodes
- 1-50GB/sec and beyond
- HPC cluster or grid is served via InfiniBand and Lustre
- Users home directories are served via NFS
- Data movers bridge between the high performance cluster scratch space and archive space
- Core modules controls data movement between the Online and Archive modules



Thumper – Perfect Grid Computing Storage No

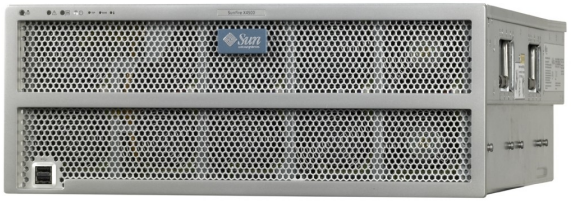


Sun Customer Ready Scalable Storage Cluster

- Addresses all those challenges:
  - Fast, incredibly fast
  - Scalable to the nth degree
  - Industry-leading storage density and power-efficiency
  - Exceptional price/performance
  - Field proven design
  - Customer specifics accomodated
  - Configured, built, tested by Sun, ready to run off the pallet

Components and Connectivity

Object Storage Servers



MetaData Servers

Data Movers

Standard on large, else optional

Compute Cluster

Infiniband Network

Archive Network

Archive / Permanent Storage

Object Storage Server

Sun Fire X4500 Server

Manages storage targets, each object contains whole file or partial stripe, controls locking for it's own I/O

solaris redhat

Compute

- 2 x Dual Core Opteron processors
- 16GB Memory

Storage

- 48 SATA II drives
- 24TB raw capacity

I/O

- 6 x SATA channels
- 2 dual port 4x IB HCAs

Availability

- Hot-swap/plug power, fans, disks

Management

- Same management as other Galaxy servers

O/S

- RHEL 4u4, Lustre OSS

Metadata Server

Sun Fire X4200 Server

- Metadata server manages file layout, metadata locking and access/security control

Compute

- 2 AMD Opteron (200 series) Processors Dual-core
- 4 GB DDR1 RAM

I/O

- 5x PCI-X slots
- 4x Gigabit Ethernet ports
- 1 x 73GB SAS disk drive
- 2 dual port 4x IB HCAs

High Availability

- Redundant hot-swap Power supplies and Fans, hot-swap disk drives
- Hardware RAID 0 & 1
- Dual MDS configured for high availability

Management

- IPMI 2.0; remote KVM, floppy/CDROM with dedicated 10/100 Ethernet port
- RHEL 4u4 with Lustre MDS software installed

Data Mover

Sun Fire X4600 M2 Server

Data mover managers the transfer of data between Lustre and SAM-QFS, proven policy-based archive

Compute

- 8 Next-Generation AMD Opteron Processor 8000 Series (dual core) processors
- Support Multi-generation single, dual-core AMD Opteron
- 32 DIMM Slots (4 per socket) DDR2-667
- 16 GB RAM

I/O

- 20GB/s (160Gb/s) bi-directional I/O
- 6 x PCI-Express slots (4 @ 8X, 2 @ 4X), 2 x PCI-X
- 4 x GigE standard, USB 2.0, Video, Serial ports
- 2 x 73 GB disk drives

Availability

- Hot-swap disk; RAID 0 or 1 built-in
- 4 power supplies, 2+2 redundant, hot-swap
- Redundant, hot-swap fans
- Sun Cluster, MS Cluster support

Management

- ILOM remote power on/off/status, browers+CLI, IPMI 2.0, SNMP
- Remote KVM, Floppy/CDROM
- N1 System Manager, N1 Service Provisioning System

O/S

- RHEL 4u4, Lustre client software and shared QFS client for Linux

Infiniband Switches

- Voltaire ISR 9024
- 10-20 Gbps performance for clusters and grids
- Ultra-low latency: under 140 nanoseconds
- Available bandwidth of up to 960 Gbps
- Powerful CPU to allow management of fabrics, as well as device management capabilities

Lustre Software Architecture

- Separate file system services allow complete parallelization of I/O functions - enabling massive I/O scalability
- Deployed in clusters > 256 nodes using IB or other low-latency interconnect
- Very high metadata and I/O performance
  - 8,000 file creations/sec in 1 dir, 1,200 nodes
  - Single clients/servers up to 2.5GB/sec.
  - Aggregate up to 11GB/sec
- Scalable to 1,000's of nodes, petabytes per file system
- 200-300 commercially supported worldwide deployments including 7 Top10 Supercomputers (according to the November 07 Top500 list):

Scalable Configurations

- Small configuration
  - > 48TB[1], 1 IB switch
- Large Configuration
  - > 144 TB[1]
  - > 2 IB switches
  - > Data Mover
- Expansion Rack
  - > Up to 192[1]TB
  - > 2 IB switches
  - > Data Mover

[1] Raw Data Size

Customer Examples

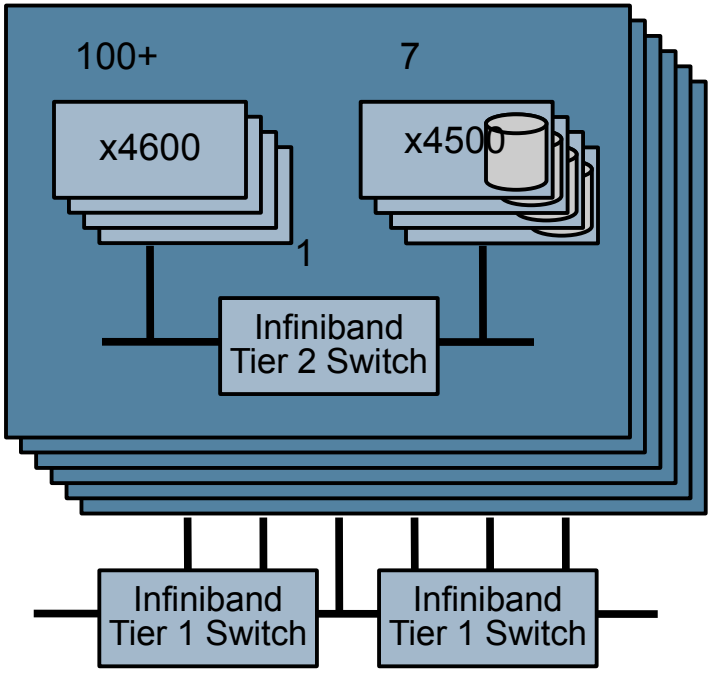
Tsubame Supercomputer

Tokyo Institute of Technology

#7 Fastest Supercomputer in the world

Tsubame By The Numbers

38.18 TFLOPS in 31 Days



6 linked sub-clusters
3 Voltaire Infiniband switches
42 Sun Fire x4500 Data Server
80+ racks
360 Clearspeed FP accelerators
655 Sun Fire x4600 Servers
10480 Opteron cores
21 TeraBytes RAM
1 PetaByte storage
N1 Grid Engine
N1 System Manager
Lustre parallel file system

Asia's fastest supercomputer

Our Customers Say It Best:

"The Arctic Region Supercomputing Center needed a flexible, cost-effective, high-performance supercomputing environment in order to facilitate a wide range of scientific research, such as ocean and climate modeling, tsunami analysis, regional weather forecasting and applications requiring basic computational fluid dynamics. Sun's new AMD Opteron-powered products, with huge memory, large disk bandwidth and a fast cluster interconnect, comprise an overall cost savings and energy-efficient architecture, making them the ideal systems to drive our compute-intensive work."

Frank Williams, Director
The Arctic Region Supercomputing Center

Recent Sun Achievements in HPC

- Deployed Texas Advanced Computing Center:
  - World's largest supercomputer using general purpose hardware
- Created the world's first open petascale architecture:
  - The Sun Constellation System
- Produced a complete solution portfolio for HPC:
  - Clustering, visualization, storage and software solutions
- Designed the most open blade platform:
  - Simultaneous support for AMD, Intel and UltraSparc blades; Solaris, Linux and Windows OS
- Acquired Cluster File Systems, developer of Lustre:
  - The world's fastest and most scalable parallel file system
- Developed the most scalable & efficient Tape Library portfolio:
  - Over 37% of the world's data is storage on Sun libraries

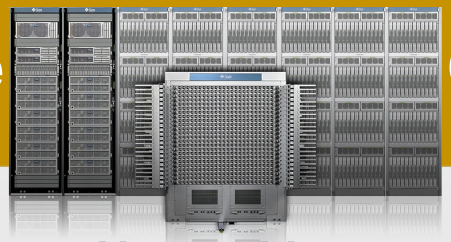
The Sun Cluster Portfolio

Open, Seamless and Comprehensive

Access	Developer	Manage-ment	OS	Inter-connect	Storage/ Archive	Systems
Visualization, Workstation, Thin Clients, Remote	Compilers, Debuggers, Optimization Tools, Libraries	Workload, Systems and Cluster Management	Linux or Solaris	InfiniBand or Ethernet	Cluster Storage, Backup, Archive, File Systems, HSM	Racks or Blades Variety of CPU Architectures

Sun Services

Storage



Networking

Compute

Sun Customer Ready

- **AMD** (EDA) – Global supplier of processors
  - Accelerate electronic chip designs
- **Metaldyne** (MCAE) – Global designer of engine, driveline and chassis products
  - Improved mechanical design and time to market
- **University College London** – Medical Research establishment
  - Improved visualization to advance biotechnology research
- **Paramount Pictures** – Media and entertainment company
  - Accelerate video rendering and animation – brought **Barnyard** release date forward three months
- **Cedars Sinai** – Medical Center
  - Developing treatments for Cancer & Heart disease
  - 400 SunFire servers generating 4TB data daily
  - Saved $60K and two months deployment time by using CRS

Getting Started Today

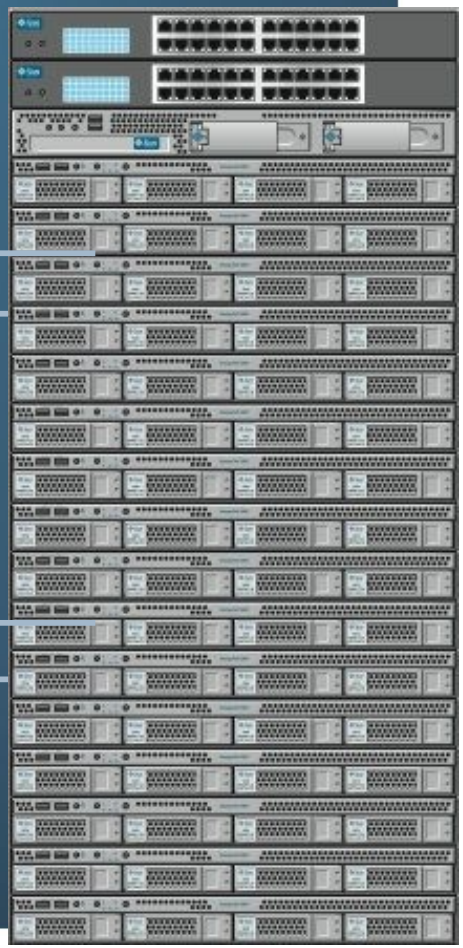
- For more information:
  - > sun.com/hpcdata
- Learn more at Radio HPC:
  - > hpcradio.blogspot.com/
- Learn more about Sun's HPC solutions:
  - > sun.com/hpc
- Customers that have deployed Sun HPC solutions:
  - > sun.com/servers/hpc/customer_references.jsp
- Contact Sun:
  - > sun.com/servers/hpc/start.jsp

The Sun StorageTek 5800 will enable you to

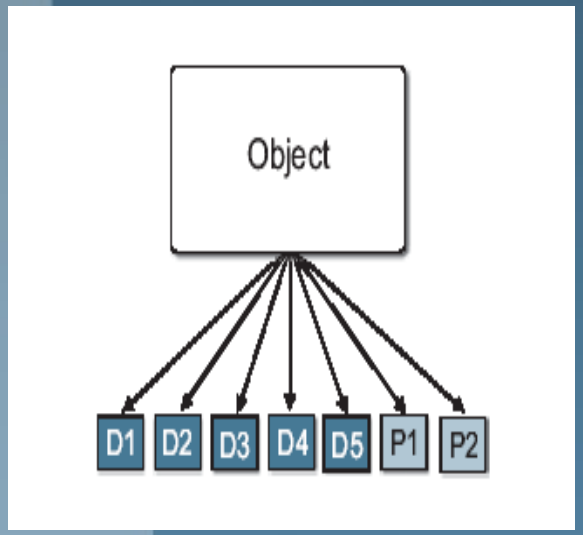
Preserve & Protect Data Assets For Long Term

Reduce Cost & Complexity

Efficiently Manage & Scale More Data Assets

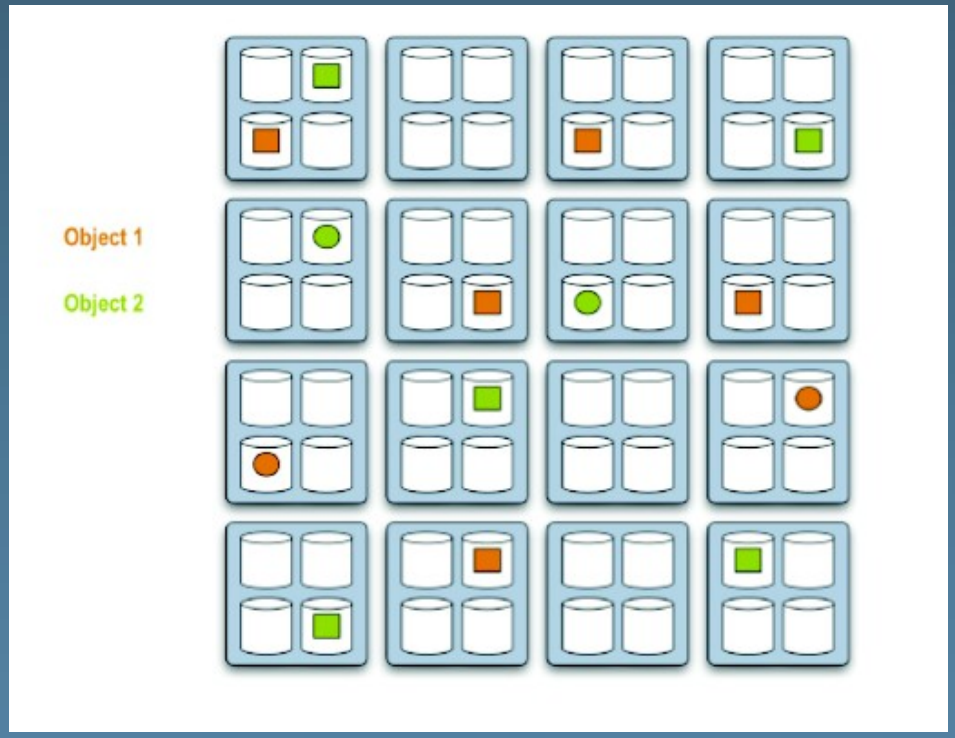


Double Parity Protection and Intelligent Data Placement Algorithm



Data object is broken into 5

data and 2 parity fragments



Data is evenly spread across all disks

Virtual Views Example: Medical Imaging

Doctor = *Smith*

Patient = *BrianParks*

Patient_sex = *M*

Modality = *MRI*

Disease = *Dropping ST5320 on hand*

Vendor = *GE*

Date = *20061001*

Define View “for_doctor” = by Doctor, Patient, Modality, Date, Area+“.jpg”

/for_doctor/Smith/BrianParks/MRI/20061001/Hand.jpg

Define View “for_researcher” = by Patient_sex, Disease, Area, Date+“.jpg”

/for_researcher/male/injury/hand/20061001.jpg

Define View “for_technician” = by Vendor, Modality, Date+“.jpg”

/for_technician/GE/MRI/20061001.jpg

THE SUN STORAGE AND ARCHIVE SOLUTION FOR HPC

The Right Data, in the Right Place, at the Right Time

jose.martins@sun.com