Sun Lustre Storage System

Simplifying and Accelerating Lustre Deployments

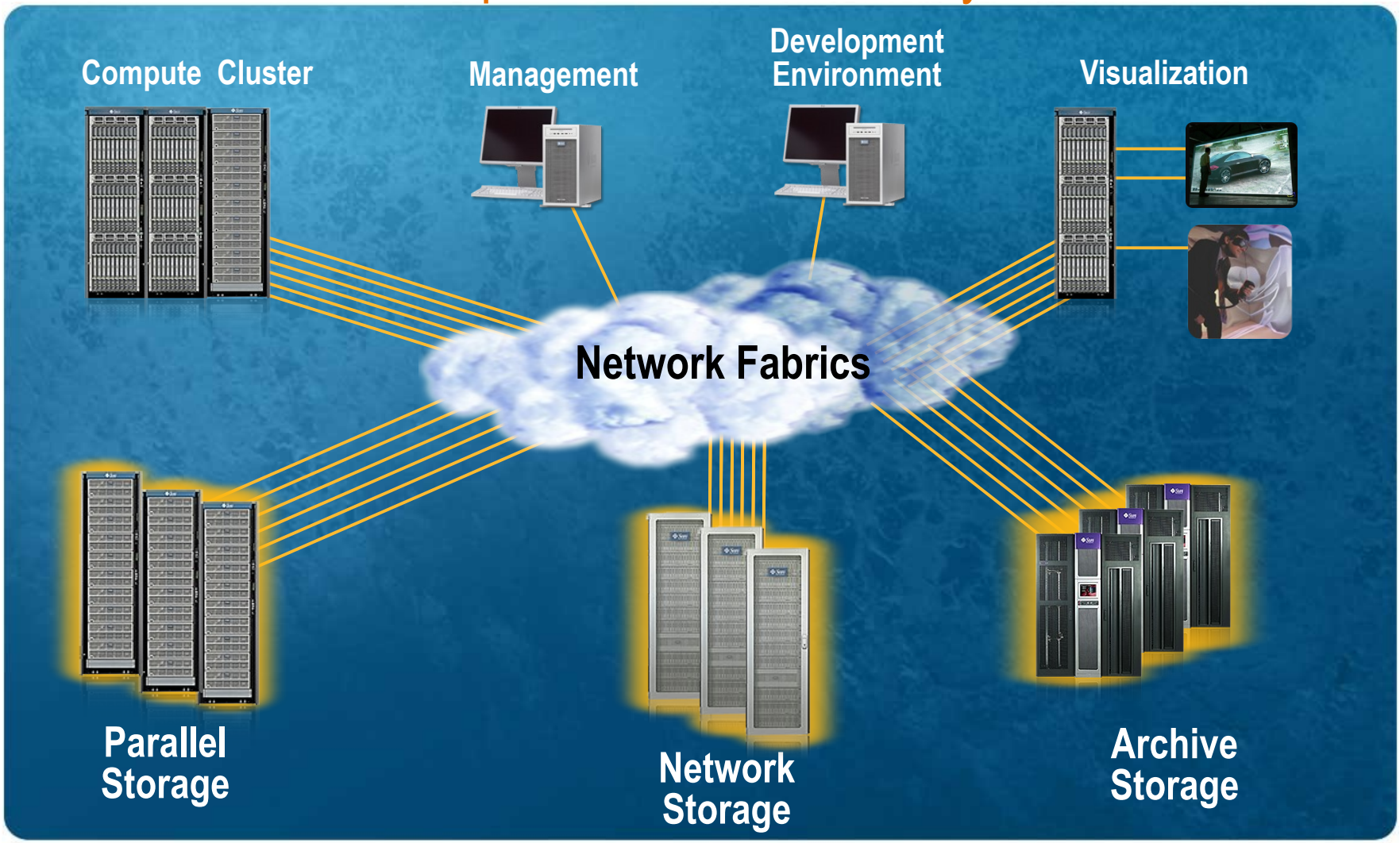
Torben Kling-Petersen, PhD

Principle Field Engineer

HPC &Cloud Computing LoB
Sun Microsystems, Inc.

Sun Storage for HPC

Solutions for the Complete HPC Data Life Cycle



Sun HPC Storage Leadership

- **62%:** Number of top 50 supercomputers using Sun Lustre
- **48%:** Number of top 50 supercomputers connected to Sun libraries for archiving data
- **Awarded #1 in Quality:** Diogenes Labs and Storage Magazine
  - For 2006–2008 Sun StorageTek SL Series
- **Winner:** InfoWorld Technology of the Year Award 2008
  - For 2008 Storage Servers – Sun Fire X4500
- **1st:** To integrate flash technology in network storage
  - Sun Storage 7000 Unified Storage System

Sun Storage Usage in HPC

	Large-Scale	Divisional	Workgroup
Sun 7000 Unified Storage	Application code, home directories, input data	Application code, home directories, input data, cluster working space if <1 GB/sec	Application code, home directories, input data, cluster working space
Sun Lustre Storage Storage	Cluster working space Large Single namespace	Cluster working space for >1GB/sec needs	Not Applicable
Sun Archive Storage	Data backup & low cost deep repositories – move in or out of cluster	Data backup & low cost deep repositories – move in or out of cluster	Data backup

Sun Lustre Storage System

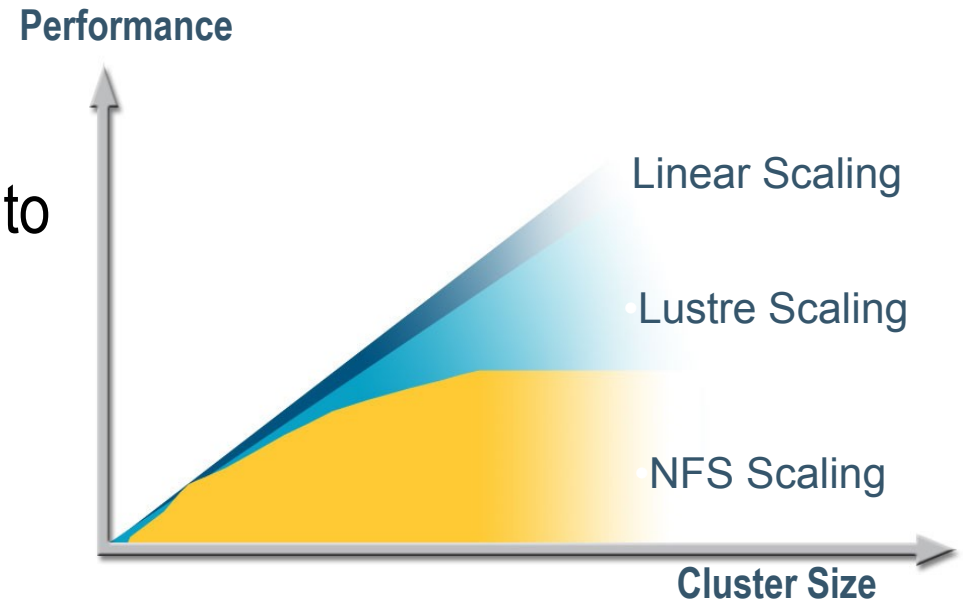
Sun Lustre Storage System

Complete Solution

- Lustre-based, high performance, parallel storage system for clusters
- Complete hardware and software solution with tested building blocks
- Leverages Sun open storage
  - > Based on, high volume, industry hardware instead of high cost, proprietary devices

Simplified Scaling

- Scaling by live addition of new modules
- Grow performance from ~2 to over 100+ GB/sec with the same architecture
- File systems scale up to 2 billion files and 32 PBs
- Cluster scaling from hundreds to thousands of nodes



Lustre is an ideal fit when NFS performance scaling is not possible or too complex

Reduced Complexity

- Aggregates many storage devices in to a single large namespace – no need break apart data sets and manually load balance data
- Predefined modules avoids trial-and-error performance guesswork of custom deployments
- Standard modules simplify planning and budgeting for future growth
- Automated scripts set up the system

Compelling Value

Through Sun Open Storage

- Cost effective solution delivered through Sun Open Storage products
- Easy scaling across storage devices avoids rip and replace upgrades
- Simpler deployments save time and money

Open storage uses open source software and industry standard server hardware in place of high cost, proprietary storage systems

Plan to Known Performance

Reduce Deployment Risk

- Eliminate guesswork
- Use known performance "increments"

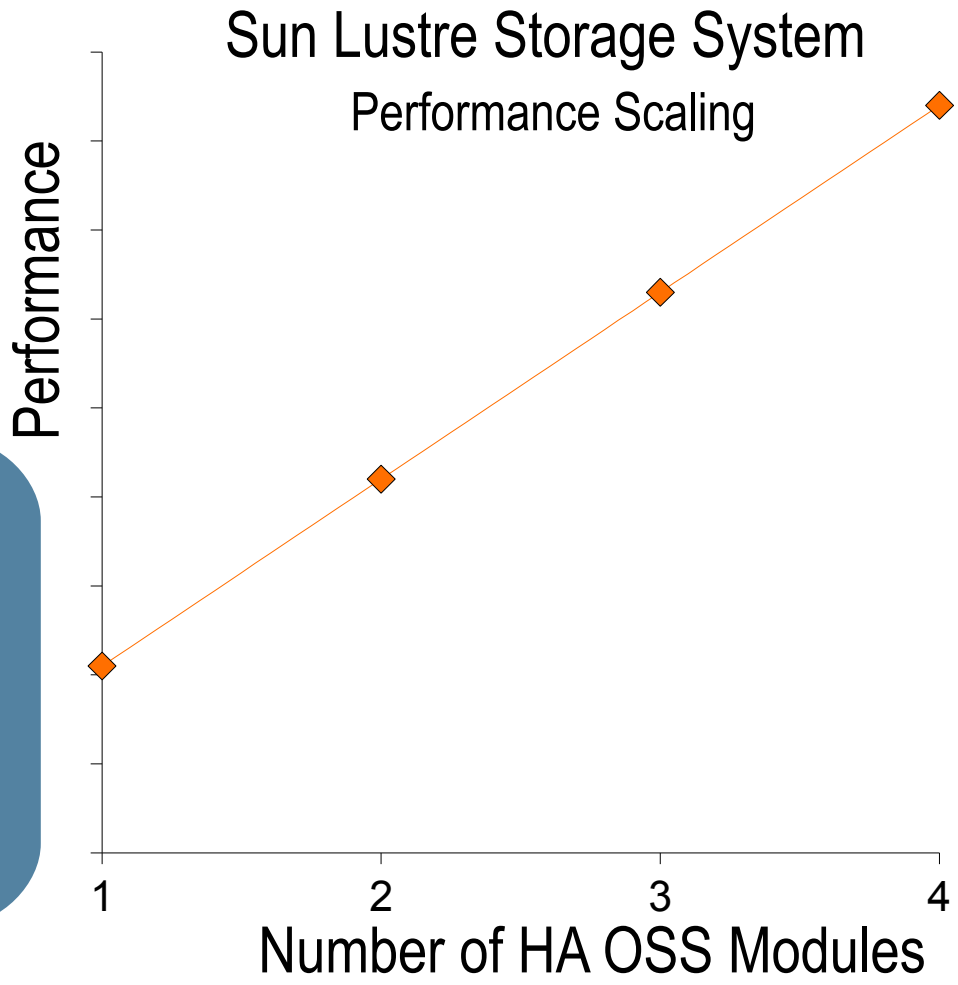
Sizing Example - 7 GB/sec Goal

> Each HA module ~ 2.1 GB/sec

> Lustre scaling ~ 90% linear

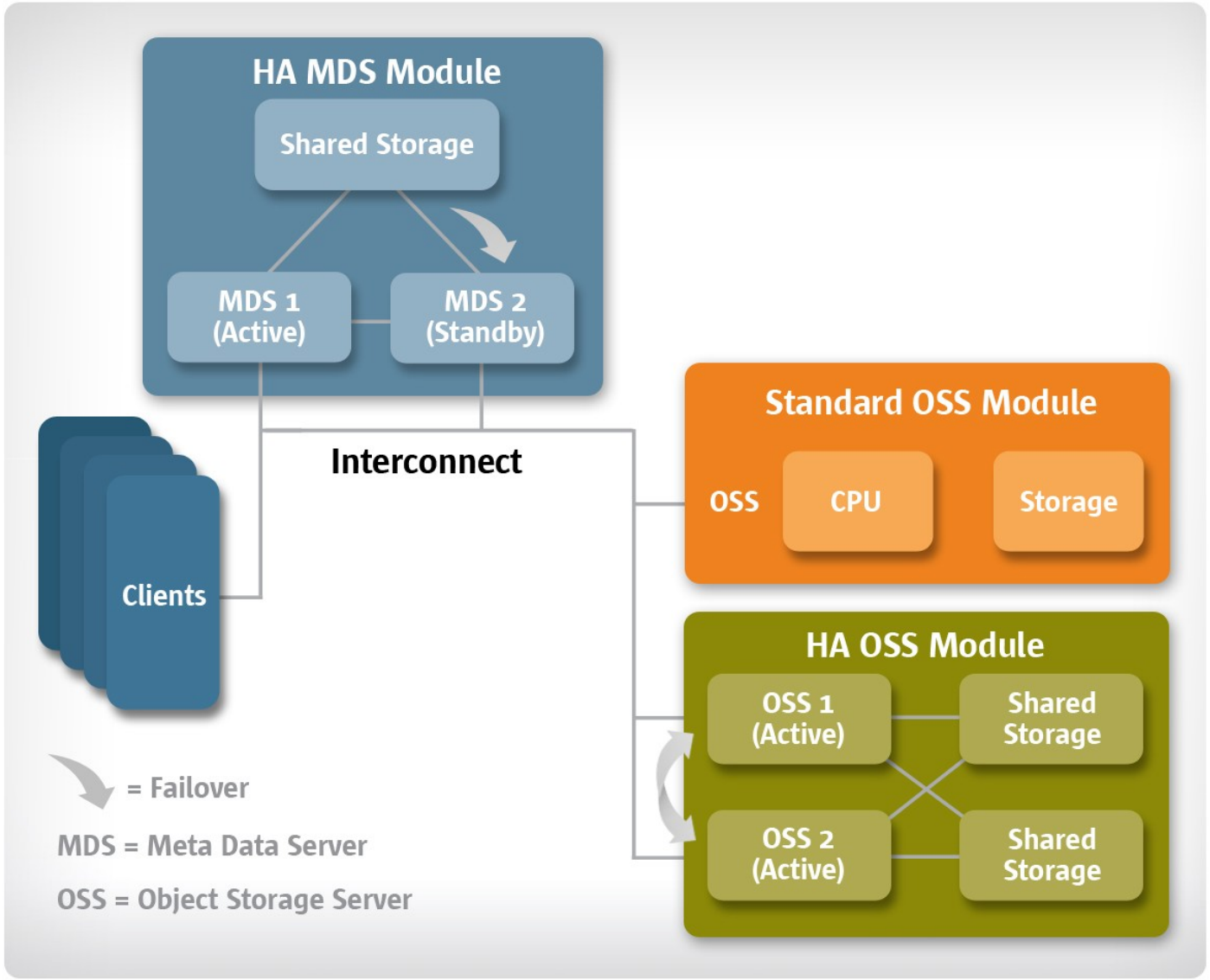
> (7GB/sec) / (2.1GB/sec x 90%) = 3.8

> *4 HA OSS modules required*



Architecture Overview

Sun Lustre Storage System Modules



Sun Lustre Storage System Contents

Defined in the Modules

- OSS & MDS servers, memory and all software
- Storage in the modules including disks
- HBA, HCA & NIC options
- SAS cabling
- RAID configurations

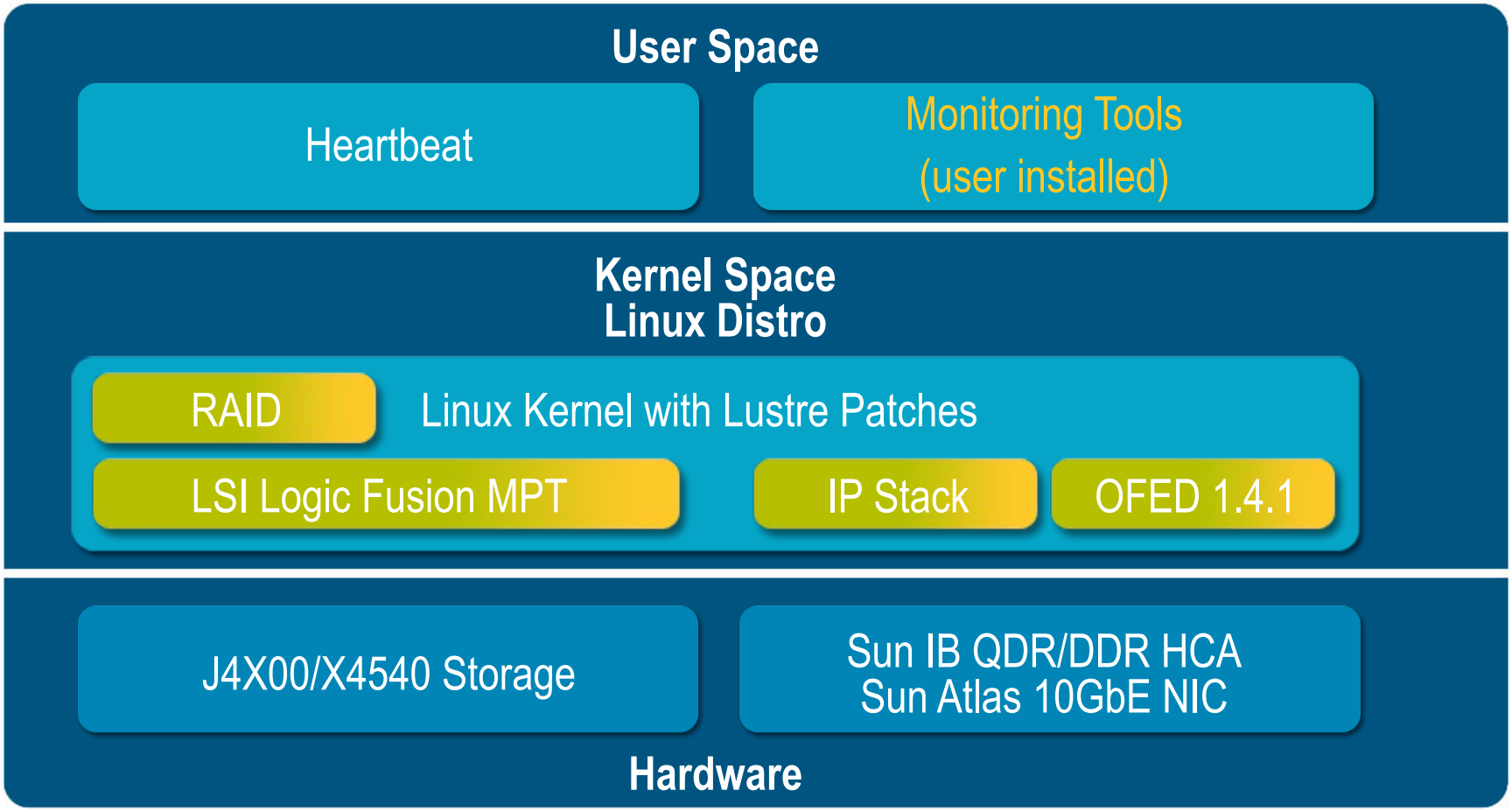
Variable Per Customer Need

- Compute nodes and software
- Networking switches and cabling
- Racks and appropriate mounting hardware
- Data mover to archive
- Implementation and support services

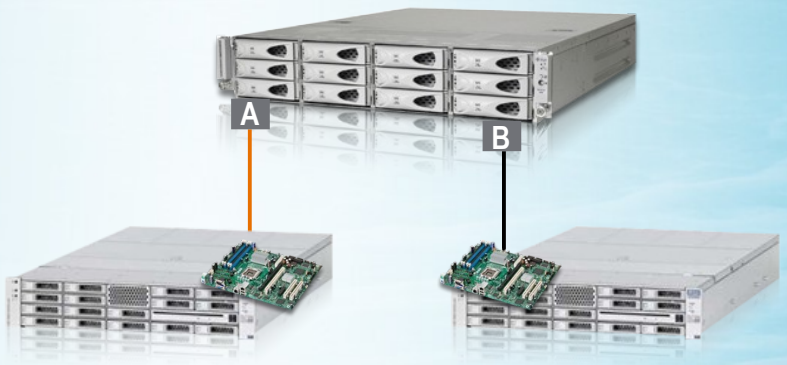
MDS and OSS Software

- Sun Lustre File System
  - Lustre 1.8.1
- Linux distribution
  - Redhat EL 5.3
- Network drivers TCP/IP and OFED
  - OFED 1.4.1
- MDS and OSS configuration tools
  - RAID Configuration Automation
  - Failover tools (Heartbeat)

Logical View



HA MDS Module



MDS 1 (Active)

MDS 2 (Standby)

SAS IO Module (SIM)
Host A SAS
Host B SAS

Hardware

- 2x Sun Fire X4270s each with
  - > Dual Intel Xeon X5570 Quad-Core (2.93 GHz) CPUs, 24GB RAM, dual SAS 2.5" boot drives, Sun 8-port SAS HBA
  - > Sun QDR IB-HCA or Sun 10GE NIC
- Shared Sun Storage J4200
  - > 12x 15k rpm, 300GB SAS drives
  - > 2x SAS I/O Modules (SIM)

Software

- Sun Lustre & tools, network drivers (IB or TCIP/IP), Linux distribution

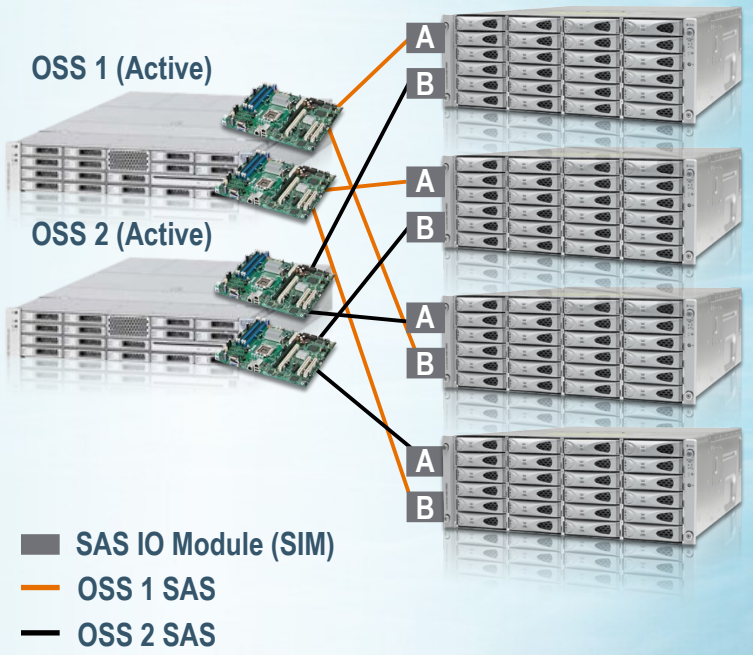
Availability

- Linux RAID 1+0 for metadata
- MDS configured in Lustre active/ passive pair
- Hot-swap redundant power and cooling

Management

- Integrated LOM Service Processor

HA OSS Module



Primary Paths Connect to SIM A
Secondary Paths Connect to SIM B

Hardware

- 2x Sun Fire X4270s each with
  - > Dual Intel Xeon X5570 Quad-Core (2.93 GHz) CPUs, 24GB RAM, dual SAS 2.5" boot drives, Sun 8-port SAS HBA
  - > Sun QDR IB-HCA or Sun 10GE NIC
- 4x Sun Storage J4400 Arrays each with
  - > 24x 7200 rpm, 1 TB SATA drives

Software

- Sun Lustre & tools, network drivers (IB or TCIP/IP), Linux distribution

Availability

- Linux RAID 6 for user data
- Lustre OSS configured in active/active pair
- Hot-swap redundant power and cooling

Management

- Integrated LOM Service Processor

Standard OSS Module

Hardware

- Sun Fire X4540 Server with dual AMD Opteron Quad-Core 2356 (2.3 GHz) CPUs
- 32 GB memory
- Sun QDR IB-HCA or Sun 10GE NIC
- 4x Gigabit Ethernet ports
- 48x 1TB SATA 3.5” disk drives

Software

- Sun Lustre & tools, network drivers (IB or TCIP/IP), Linux distribution

Availability

- Linux RAID 6 for user data
- Redundant hot-swap power and cooling

Management

- Integrated LOM Service Processor

Optimized RAID sets

- 2 External Linux Journal disks/OST
- Rotated over all disk S-ATA controllers

Controller								
0	c0t0d0	c0t1d0	c0t2d0	c0t3d0	c0t4d0	c0t5d0	c0t6d0	c0t7d0
1	c1t0d0	c1t1d0	c1t2d0	c1t3d0	c1t4d0	c1t5d0	c1t6d0	c1t7d0
2	c2t0d0	c2t1d0	c2t2d0	c2t3d0	c2t4d0	c2t5d0	c2t6d0	c2t7d0
3	c3t0d0	c3t1d0	c3t2d0	c3t3d0	c3t4d0	c3t5d0	c3t6d0	c3t7d0
4	c4t0d0	c4t1d0	c4t2d0	c4t3d0	c4t4d0	c4t6d0	c4t7d0	c4t7d0
5	c5t0d0	c5t1d0	c5t2d0	c5t3d0	c5t4d0	c5t5d0	c5t6d0	c5t7d0
RAID 6 Vol.	OST 1		OST 2		OST 3		OST 4	

Key:

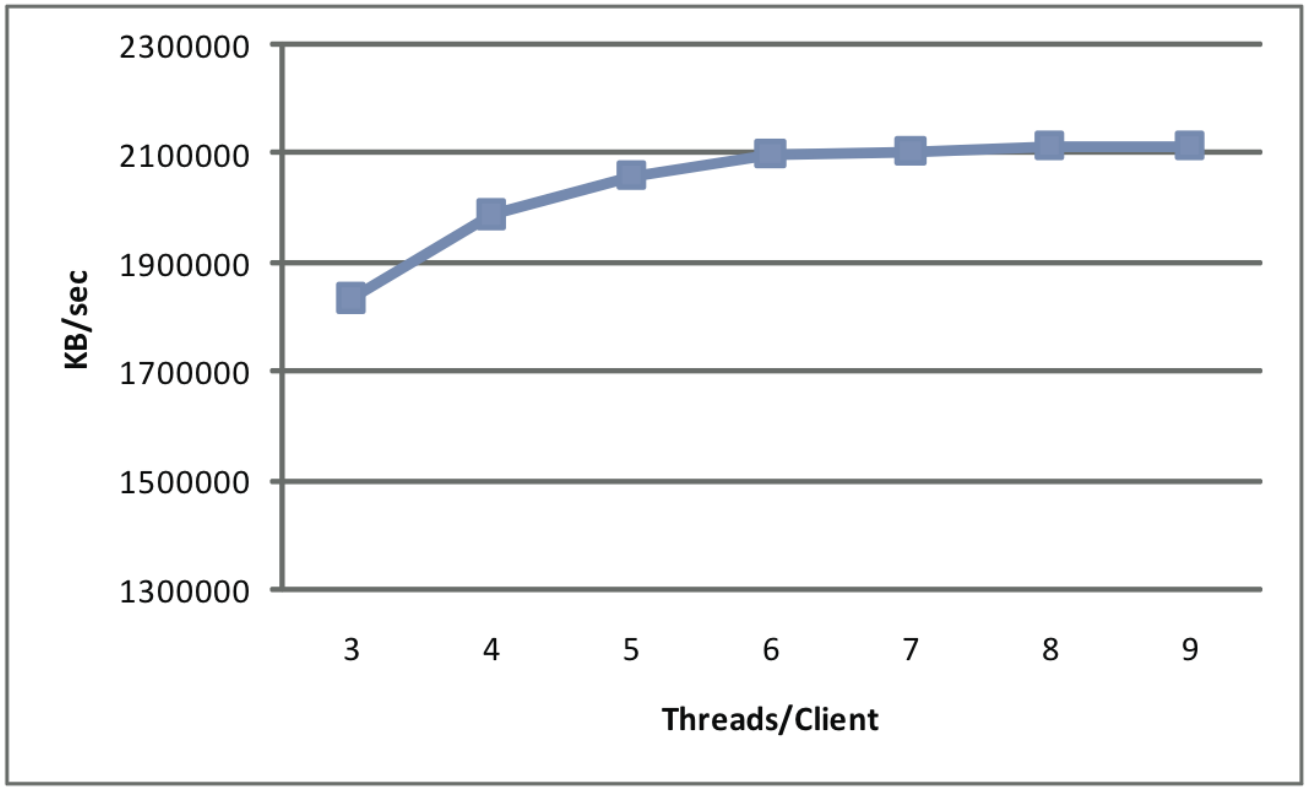
Journal	Data	Journal	Data	Journal	Data	Journal	Data

X4540 OST mapping

Performance ??

- Single HA-OSS
- 12 clients/8 threads
  - 96 threads total
- DDR Infiniband*
- 1 MB blocks
- IOzone

- Max Write: 2.114 GB/s
- Max Read: 1.995 GB/s



* For Snowbird 1.5, interconnect will be QDR, performance tests are on-going

Quick Reference

	Aggregate Bandwidth*	Availability	Capacity** & Rack Units	Disk Type
HA MDS Module	N/A	Active-Passive Lustre MDS servers Shared RAID 1+0 storage, Redundant power & cooling	3.6 TB RAW 6 RU 1.8 TB RAID 1+0 6 RU	3.5” 300GB, 15K rpm SAS
HA OSS Module	Up to 2.1 GB/sec, sustained writes	Active-Active Servers RAID 6 for data RAID 1 for journals Redundant power & cooling	96 TB RAW 20 RU 64 TB RAID 6 20 RU	3.5” 1 TB, 7.2 K rpm SATA II
Standard OSS Module	Up to 970 MB/sec, sustained writes	RAID 6 for data RAID 1 for journals Redundant power & cooling	48 TB RAW 4 RU 32 TB RAID 6 4 RU	3.5” 1 TB, 7.2 K rpm SATA II

*Measured Using DDR InfiniBand **Capacities do not include internal OS drives

Deployment and Services

Sun Lustre On-Site Implementation Services

Quick and efficient Lustre integration into HPC environments

Flexible Offering	Designed to meet your specific needs
Meet Scalability Requirements	OSS implementation available for increased I/O and throughput
Minimize Deployment Time	Proven, tested, & validated procedures
Satisfaction is our priority	4 hr. follow-up TOI to review & educate

- ***For more information: Sun.com/service/implement***

Summary - Why Sun for HPC?

If You Need HPC – You need Sun

- **Delivering a true *HPC System***
  - *A* complete, and tightly integrated HPC ecosystem
- **Industry leading HPC storage solutions**
  - Sun Lustre Storage System, Sun Archive Solution for HPC, Sun Storage 7000 Unified Storage System
- **Industry leading scaling solutions**
  - Scale for the most challenging environments
- **Industry leading innovation**
  - At all levels to provide: performance, scale and efficiency
  - Making HPC Simple & Easy for everyone
- **Award winning global Service organization**
  - Optimizing and supporting your solution

For More Information

- Overview
  - sun.com/scalablestorage
- Solving the HPC Bottleneck: Sun Lustre Storage System
  - https://wikis.sun.com/display/BluePrints/Solving+the+HPC+IO+Bottleneck+-+Sun+Lustre+Storage+System
- Learn more about Sun's HPC solutions
  - sun.com/hpc
- Sun's HPC Customers
  - sun.com/servers/hpc/customer_references.jsp
- Join the HPC Community & Receive HPC news
  - hpc.sun.com

Thank You

Torben Kling-Petersen, PhD
kling@sun.com