5th International Workshop on Grid Computing for Complex Problems
(Vendor session)

Institute of Informatics Slovak Academy of Sciences in Bratislava

Monday 26th Oct. 2009

HIGH PERFORMANCE COMPUTING FROM SUN

Dalibor Kubacek
GSE IT Architect
Sun Microsystems Slovakia, spol. r.o.

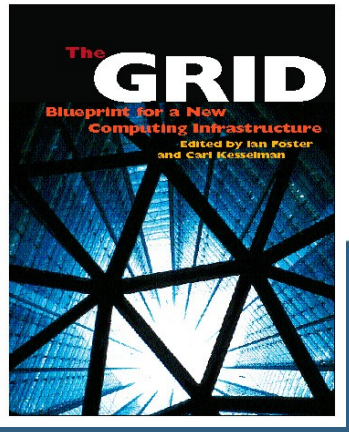
HPC & Grid

High Performance Computing (HPC)

- is about using concentrated compute power to address highly complex problems, business critical analysis, or computationally intensive workloads – faster and more efficiently
- the use of parallel processing for running advanced application programs efficiently, reliably and quickly
- the term applies especially to systems that function above a teraflop or 10^{12} floating-point operations per second

  - Some users need to complete work faster.
  - Some users need to complete more work.
  - Some users need to do both...and more.

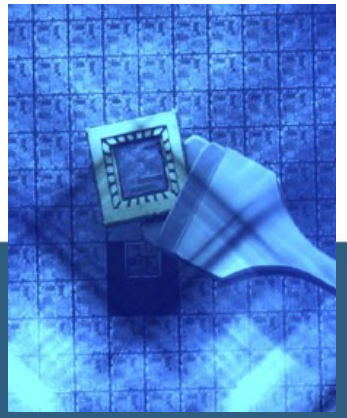
HPC & Grid



Grid

- Originally born out of High Performance Computing but relevant to wider markets
- Oriented to enhancing the network of things rather than making applications run faster
- "A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities"
- A universal computing infrastructure that builds on the power of the **Net** and enables more efficient computation, collaboration, and communication

HPC & Grid

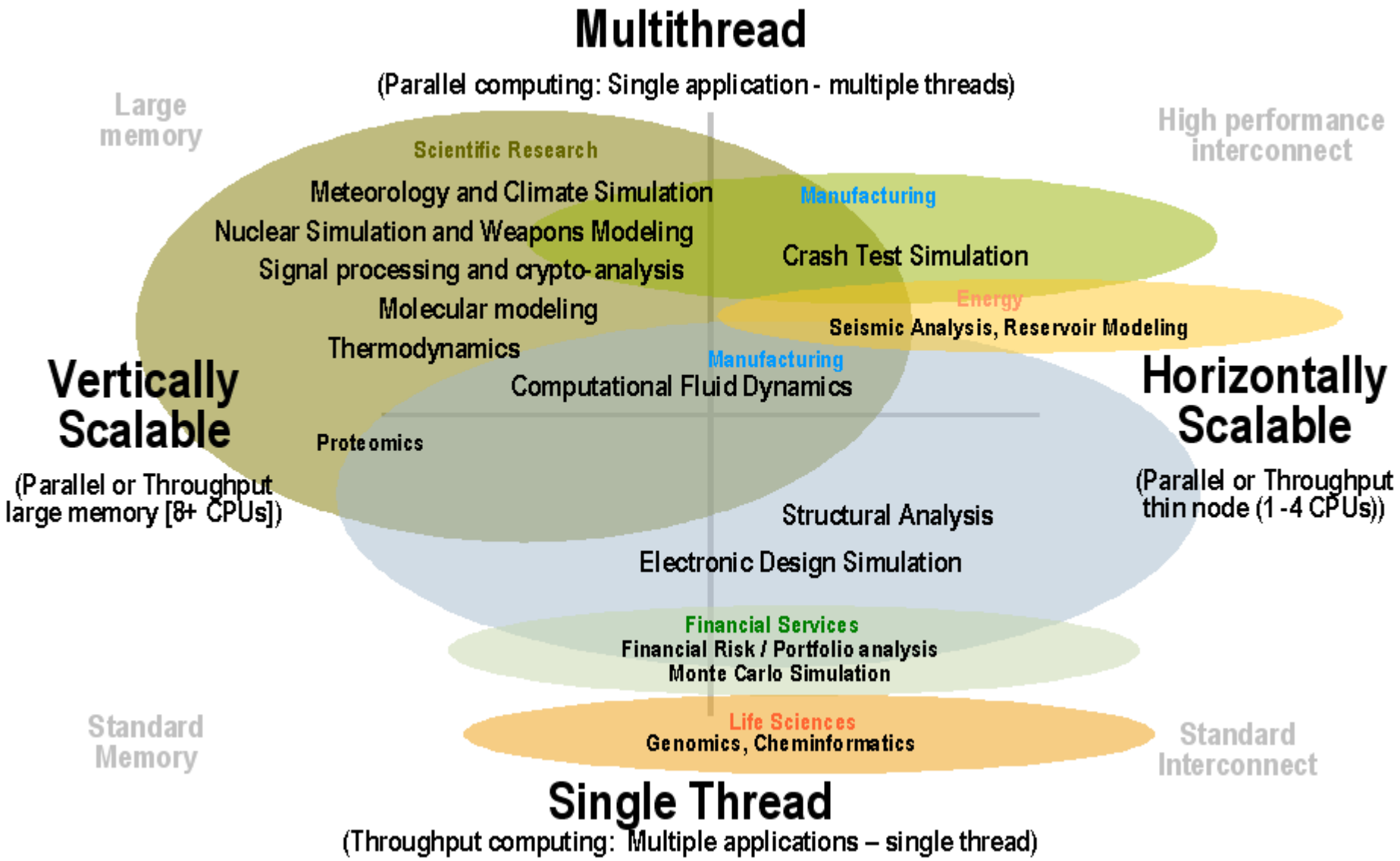


Grid

- Some academics define grid computing as the dynamic sharing of sets of computing resources by virtual organizations who come together for a particular purpose. These resources and their interconnects are not fixed. A way of using many heterogeneous resources to perform many kinds of tasks, accessible from many places by many people
- Grid can mean a variety of things to various organizations and marketers (*).

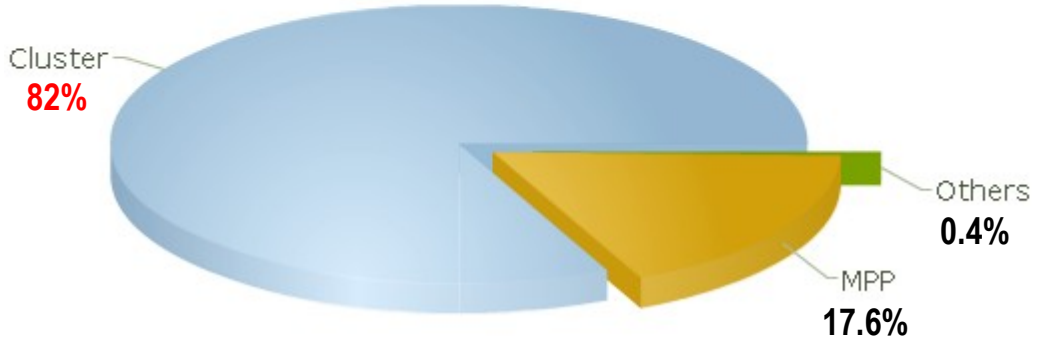
(*) Finance institutions may refer to grid, or distributed environment a Service Oriented Architecture for example

Technical Workloads and Solutions



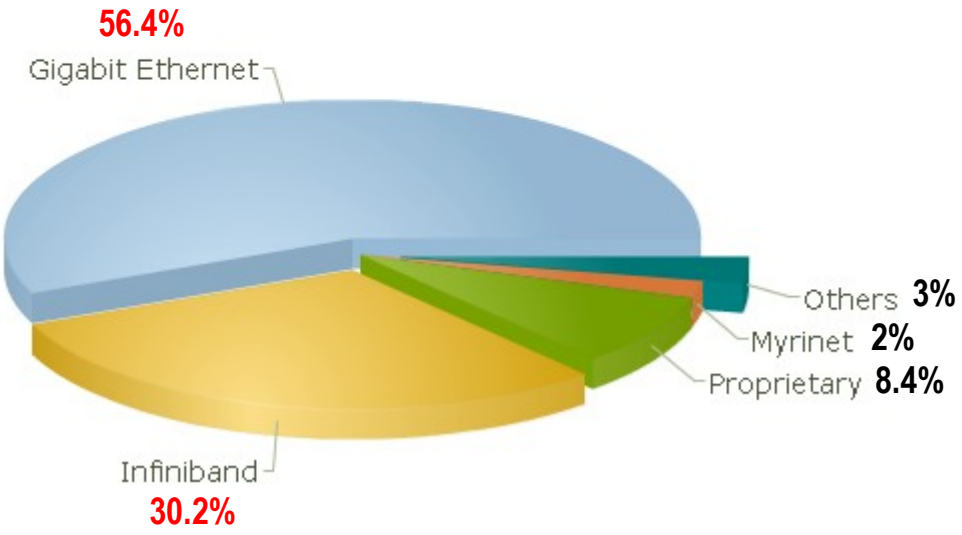
TOP500® SUPERCOMPUTER SITES

Architecture / Systems
June 2009



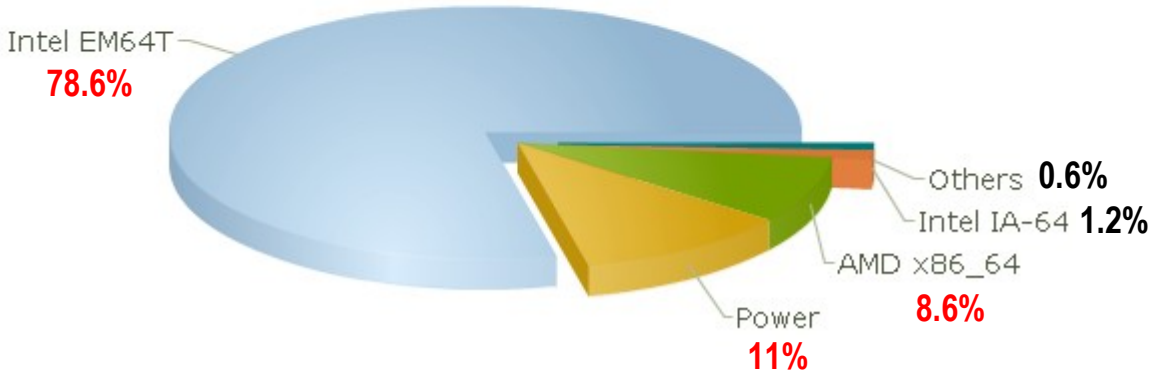
TOP500® SUPERCOMPUTER SITES

Interconnect Family / Systems
June 2009



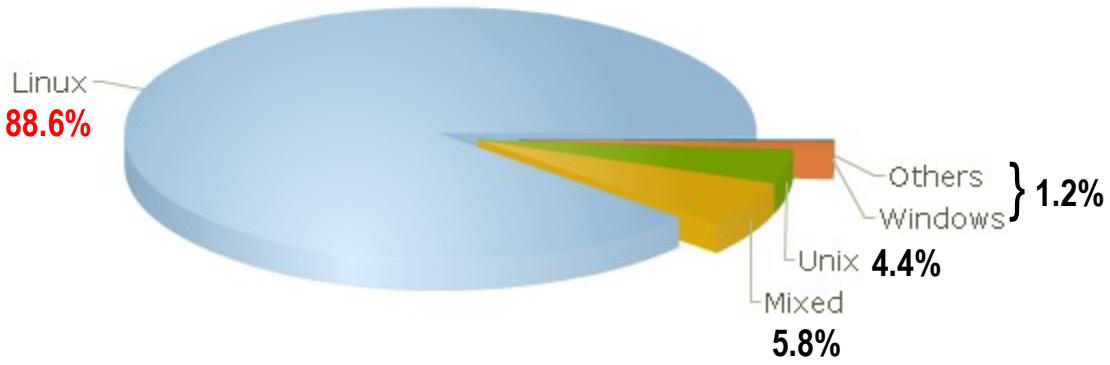
TOP500® SUPERCOMPUTER SITES

Processor Family / Systems
June 2009



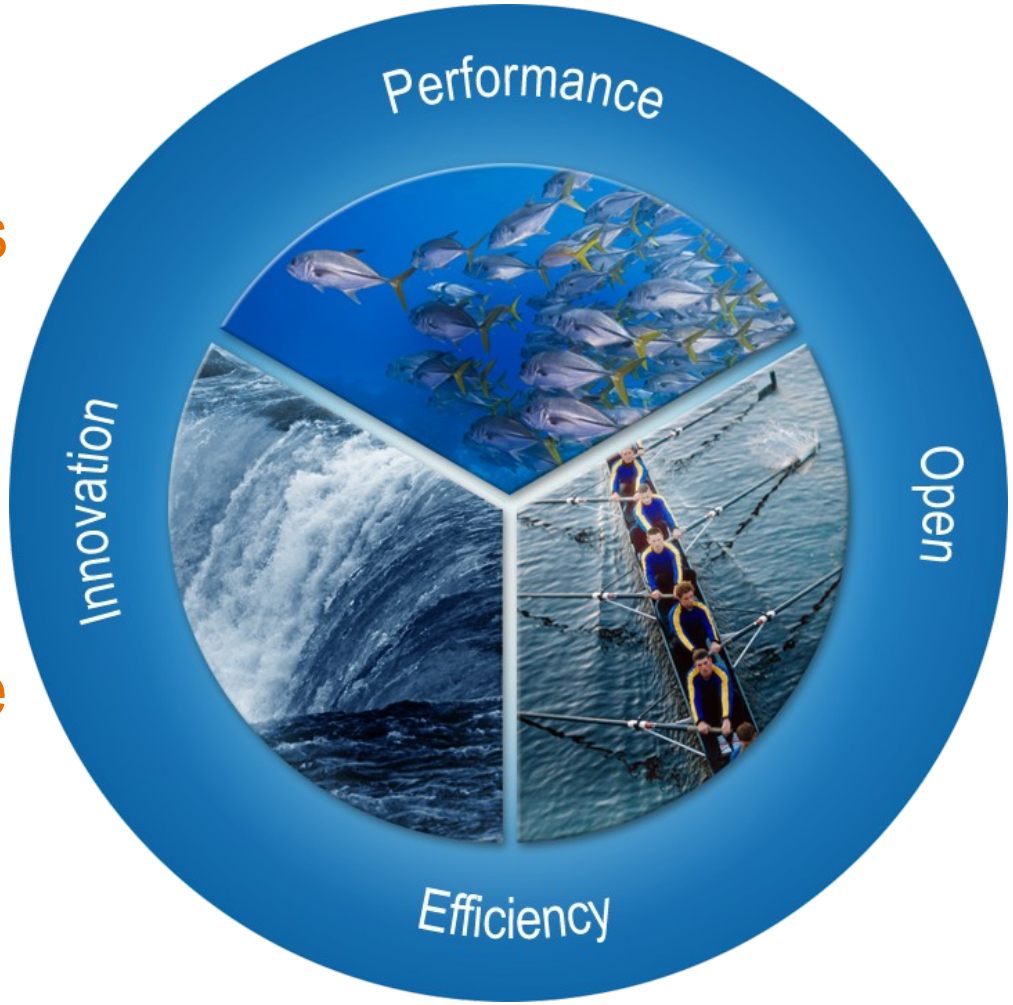
TOP500® SUPERCOMPUTER SITES

Operating system Family / Systems
June 2009



Our Vision

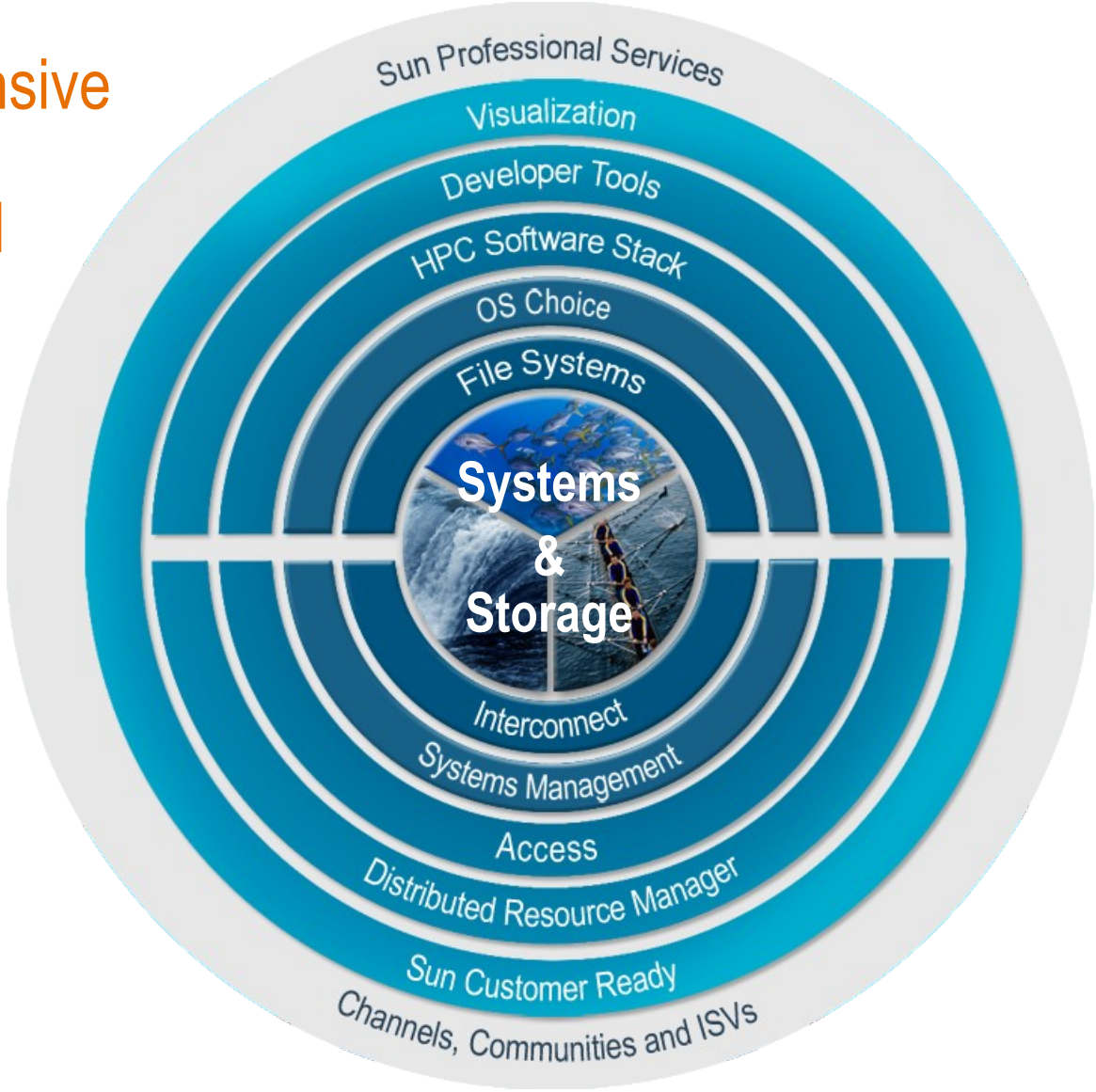
To provide our customers with a competitive edge through innovative infrastructure designed to solve their most data-intensive and computationally-intensive problems



Our Strategy

Building a Comprehensive HPC Portfolio that Delivers Differentiated Customer Value

- Providing complete and Integrated Solutions for targeted markets with best-in-class unique Sun products
- Innovation built on Open Standards at the component, product & solution level
- Making HPC simpler and easier Ensuring design and delivery excellence using strategic engagement teams
- Extending value through partnerships



Game Changers

Sun Fire X4540

Worlds first Storage Server
Up to 48TB in 4RU

Sun Datacenter InfiniBand Switch 648

High-density QDR InfiniBand Switch

Open Storage

Easy-to-use storage appliances based on Open Standards

Multi-Core and Multi-Threading

CoolThread servers and Solaris optimizations for the most demanding thread-based applications

Open Source

Community Driven Development of HPC Software Solutions

GRID ENGINE

Lustre

Sun Open Cloud Platform

Public & Open Clouds with leading Open Technologies

The Complete Picture

Sun Services

Compute Engine
Data Cache

Long-Term
Retention and Archive

Computer Cluster

Management

Development
Environment

Tier 2 Fixed
Content Archive

Tier 2 Near-Line
Archive

IB Network

Load

Archive

LAN / SAN

Automated
Migration

Scalable
Storage Cluster

Data Movers

Tier 1 Archive and
Home Directories

Sun Customer Ready

Choice and Flexibility

Complete Range of Compute Nodes (Blades or Rack Mounts) Based on:

AMD | intel | ULTRASPARC

Choice of OS

solaris opensolaris |  | Microsoft Windows

Racks and Blades

Rack Mount Strengths

- Ideal for smaller environments
- Fat or Thin nodes for mixed environments
- Range of equipment configurations in a rack

Blade Strengths

- Less cabling
- Lower power requirements
- More reliable
- Chassis investment protection

Sun Blade 6048 Modular System

- The first blade platform designed for extreme density and performance
  - > 7 TFLOPS, 768 cores per chassis/42U
    - — 50% more compute power than HP C-Class
    - — 71% more compute power than IBM BladeCenterH
  - > 4 InfiniBand Leaf Switch Network Express Modules
    - — Lowest cost per port with ultra-dense switch solution
- Pay as you grow platform ideal for fast growing businesses
  - > Choose among SPARC, AMD Opteron and Intel Xeon CPU technologies
- Runs general purpose software
  - > Custom compiles and tuning are not required
- Realize economies of scale savings in power and cooling

Sun Constellation System – Massive Horizontal Scale

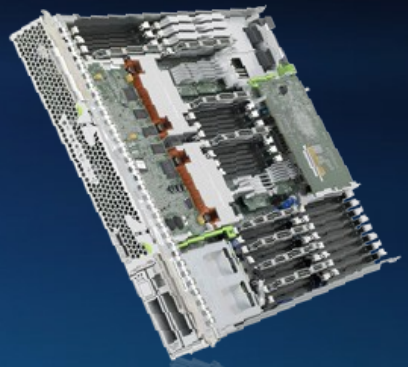
Versatile Sun HPC Blade Portfolio

Sun Blade X6440

- 4 high performance Six-Core or enhanced Quad-Core AMD Opteron Processors
- 32 memory slots supporting DDR2, ECC DIMMs with 2x 2 GB, 2x 4 GB or 2x 8 GB memory kits; Up 256 GB of main memory using 8 GB DIMMs

Sun Blade X6275

- 2 quad-core Intel Xeon Processors 5500 Series per compute node; for a **total of four processors** per server module
- Up to 96GB of main memory (per compute node); Up to 192GB (total) per server module



Sun Blade T6340

- 2-socket 6 or 8-core UltraSPARC T2 Plus processor
- Up to 128 simultaneous execution threads
- One floating-point processor per core and integrated, on-chip crypto accelerator
- 32 DIMM slots with maximum memory capacity of 256 GB

intel

vmware

Microsoft Windows

solaris

Sun Blade Independent IB I/O

Sun Blade 6048 InfiniBand QDR Switched Network Express Module

Industry's Only Chassis QDR Integrated Leaf Switch

- Two onboard 36-Port QDR IB switches

3:1 Reduction in Cabling Simplifies Cable Management

- 30 ports of 4x IB QDR connectivity realized with only 10 physical 12x connectors
- Two Passthru GigE ports per server moduler module

QDR IB PCIe ExpressModule

- Two QDR InfiniBand QSFP ports
- x8 PCI Express Base 2.0 server connection
- Ideal for constructing "dual rail" InfiniBand clusters

Sun™ Datacenter Infiniband Switch 648

Unparalleled HPC Fabric Scale-out for QDR (40Gbps) Attached Servers

Market Segment

- Divisional and Supercomputing HPC

Design Center

- 648 QDR ports
- 216 physical ports
- Solution scales to 5,000+ nodes

Management

- ILOM agents for xVM
- Host based fabric management
- Host based subnet management
- Open Fabrics management stack

Sun™ Datacenter Infiniband Switch 72 / 36

Market Segment

High performance data switch **for small to mid-sized** cluster deployments

Design Center

72 ports of QDR or DDR or SDR IB

Data throughput: 4.6 Tb/s

300 nS port-to-port latency

Scales up to 576 servers and storage systems

Market Segment

High performance data switch for **enterprise** applications

Design Center

36 ports of QDR or DDR or SDR IB

Data throughput: 2.3 Tb/s

100 nS port-to-port latency

Building block for hierarchical fabric topologies

Sun Constellation System Success

TACC World's largest Open Supercomputer

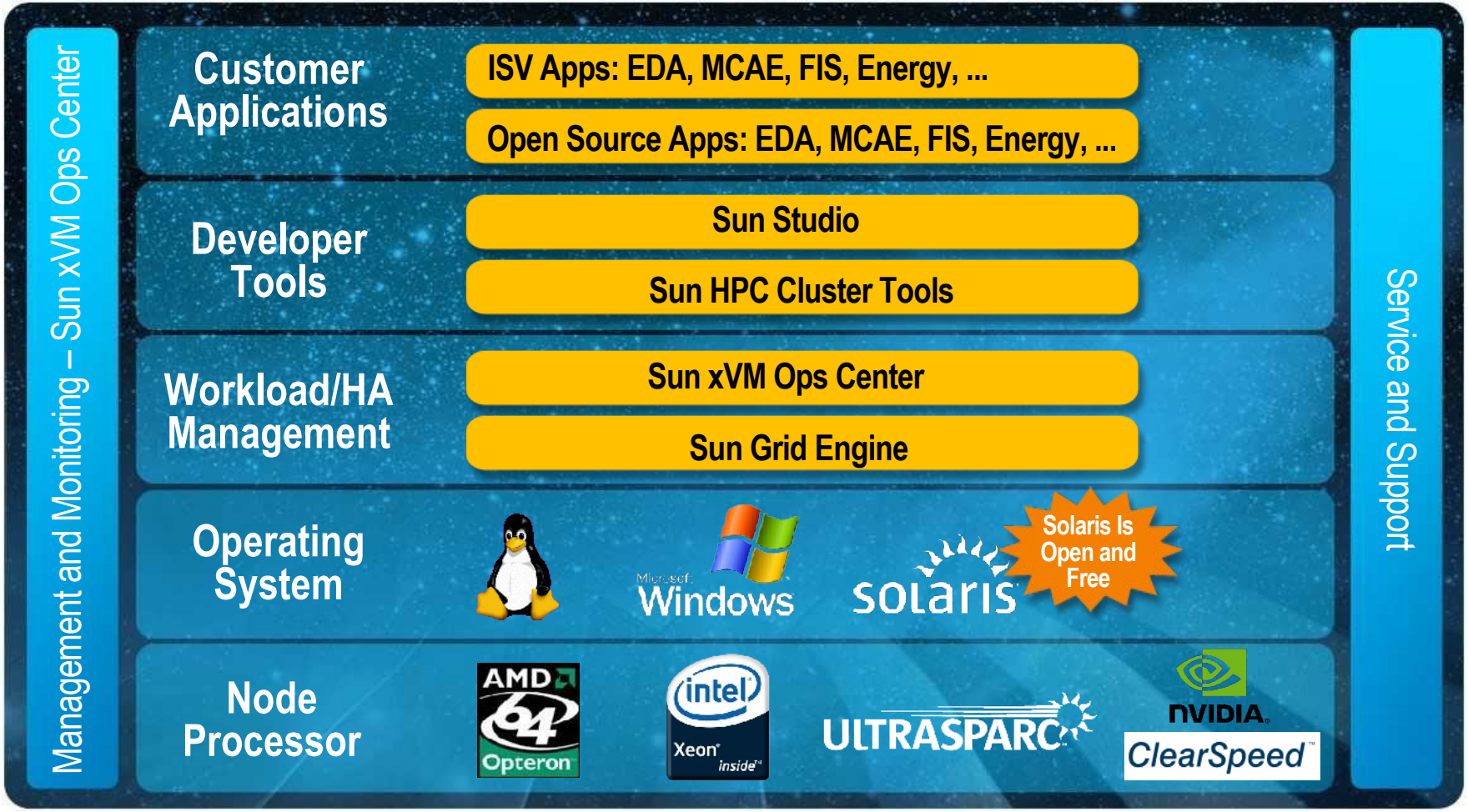
KISTI Largest supercomputing center in Korea

Jülich Germany's largest HPC center

Sandia Worlds first & largest QDR Torus configuration

The Sun HPC Software Stack

A Full Deck (Summary)



solaris

"Sun possesses a deep understanding of the requirements for HPC users and applications and has an unmatched portfolio of technology suitable for HPC, in which Solaris plays a key role"

Source: IDEAS International – Making Solaris Stick with High Performance Computing Systems

Platform Choice

- SPARC, AMD and Intel

Extreme Performance

- Optimized for large memory, NUMA, CMT, large core counts
- Real-time and fixed-priority schedulers
- Dynamic Tracing (DTrace)

Relentless Availability

- Robust and stable – 15 years of continuous improvement
- Predictive self-healing
- Extreme observability

Open Storage

- Massive scale and performance
- Protect/archive critical data
- Leading open source storage stack

Sun Studio Software Overview

Download at:http://developers.sun.com/sunstudio

Integrated Developer Tool Chain

- C/C++/Fortran compilers
- Highly-tuned math and performance libraries
- OpenMP v3.0
- Sun Performance Analyzer (finds bottlenecks)
- Sun Thread Analyzer (detects data races and deadlock)
- NetBeans-based IDE

Why Sun Studio?

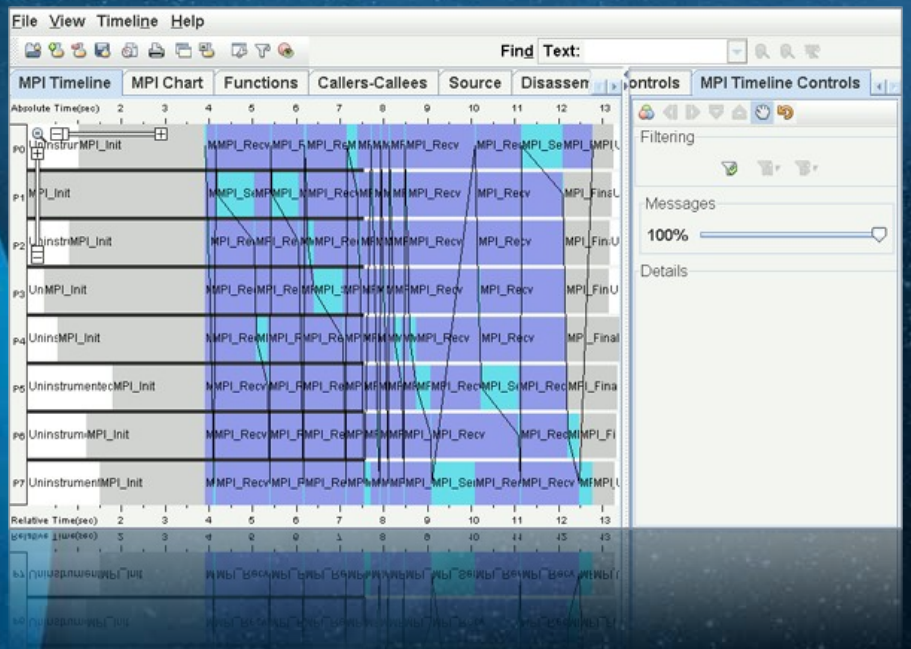
- Record-setting performance: Excellent auto-parallelization and advanced multi-core optimizations
- Dozens of recent industry-based benchmarks across Intel, AMD, Sun, Fijitsu architectures
- Heterogeneous development – Solaris (x86, SPARC), Linux (x86)
- Developer services – Community and/or Sun

Sun HPC ClusterTools

Download at: http://www.sun.com/clustertools

High-Performance MPI Libraries and Parallel Job Launcher

- Full MPI 2.1 standard
- Based on Open MPI and fully supported by Sun
- Solaris and Linux support
- Sun Studio and GNU support
- DTrace providers on Solaris
- Processor affinity support
  - **IB, GbE, 10GbE and Myrinet MX**
- MPI application profiling support
- Plug-ins for Sun Grid Engine, PBS Pro and Torque



Sun Grid Engine (6.2 Update 4)(*)

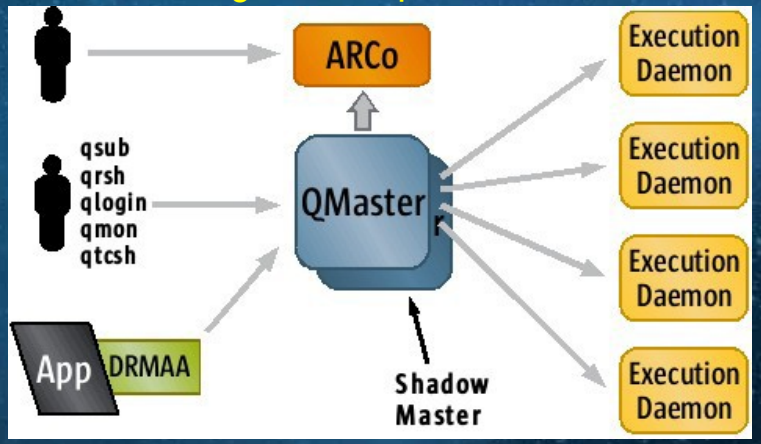
Download at: http://www.sun.com/sge

Dynamic Resource Management

- Job Scheduling
- Advanced reservation
- Resource monitoring
- Policy administration
- User authentication and access control
- Accounting and reporting

Used @ **TACC; TITECH; Mentor Graphics; Rising Sun Pictures; Complete Genomics; DE Shaw** and many, many others

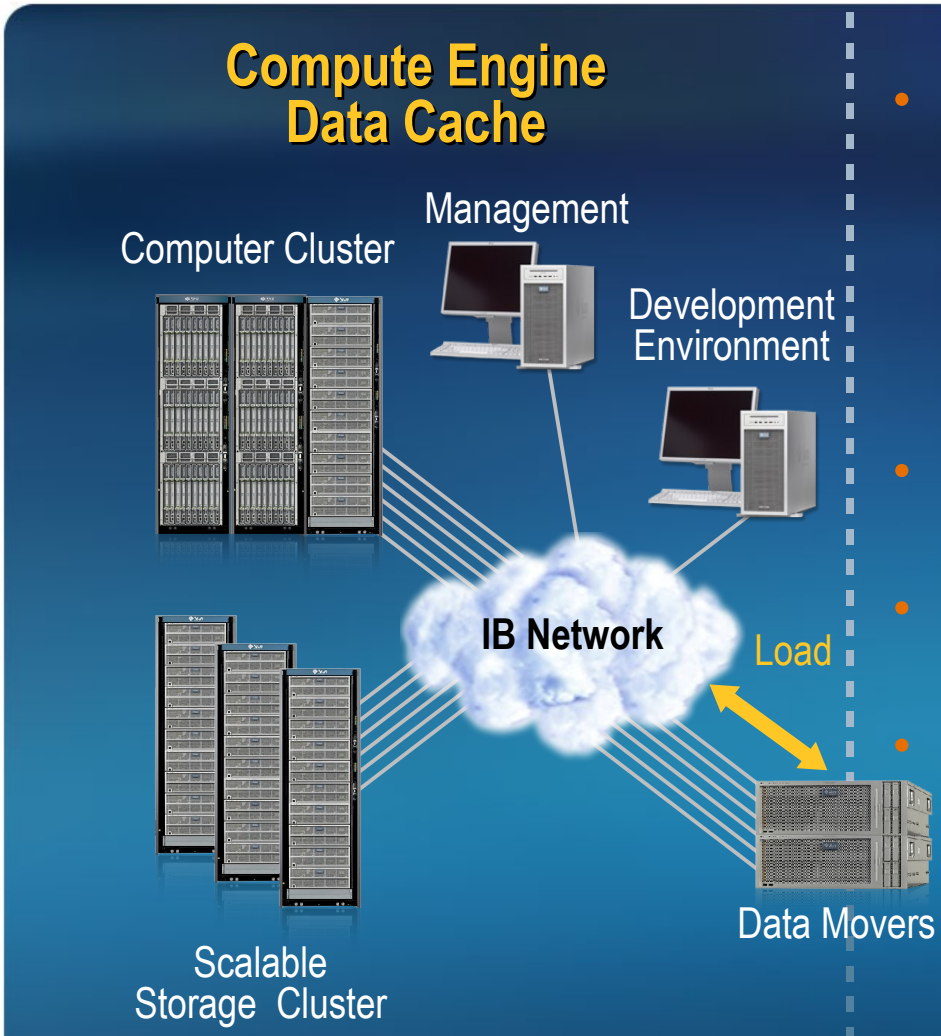
Sun Grid Engine Component Architecture



- **DRMAA** - Distributed Resource Management Application API
- **QMaster** – the central component of a cluster
- **Shadow Master** – a daemon that manages QMaster fail-over
- **Execution Daemon** – the component responsible for executing jobs using compute resources
- **ARCo** - Accounting and Reporting Console

(*) In July 2000 Sun Microsystems, Inc. acquired GRIDWARE, Inc.

Scalable Storage Cluster



- Extreme Multi-Petabyte scalability and aggregate bandwidth delivered with leading density and cost
  - > Density ~240 Terabytes per rack
  - > Bandwidth ~7.8 GB/second sustained per rack
  - > Cost – List price for under $1 per GB
- Architected with Open Standards - Infiniband and Open Source software
- Ideal for cache and feeding the compute cluster
- Leverages revolutionary **SunFire x45x0** hybrid data servers, low-latency IB interconnects, and Lustre – the industry leading parallel file system

Why Parallel Storage?

- Traditional storage architectures “break” under current HPC application demands
  - NFS not designed to scale to level needed
  - SANs not designed for data sharing, costly at scale
  - Performance does not scale beyond a device
- Parallel Storage provides
  - Easy growth through horizontal scaling
  - Very high aggregate I/O bandwidth and capacity for a single name space
  - Support for 100s to 1000s of nodes and users

Unique Capabilities from Sun

Sun Lustre Storage File System(*)

High performance and broad scaling

- From a few to over 100 GB/second
- From dozens to thousands of nodes

Breakthrough economics, up to 50% savings

- Built with standardized hardware and open source software vs. proprietary products

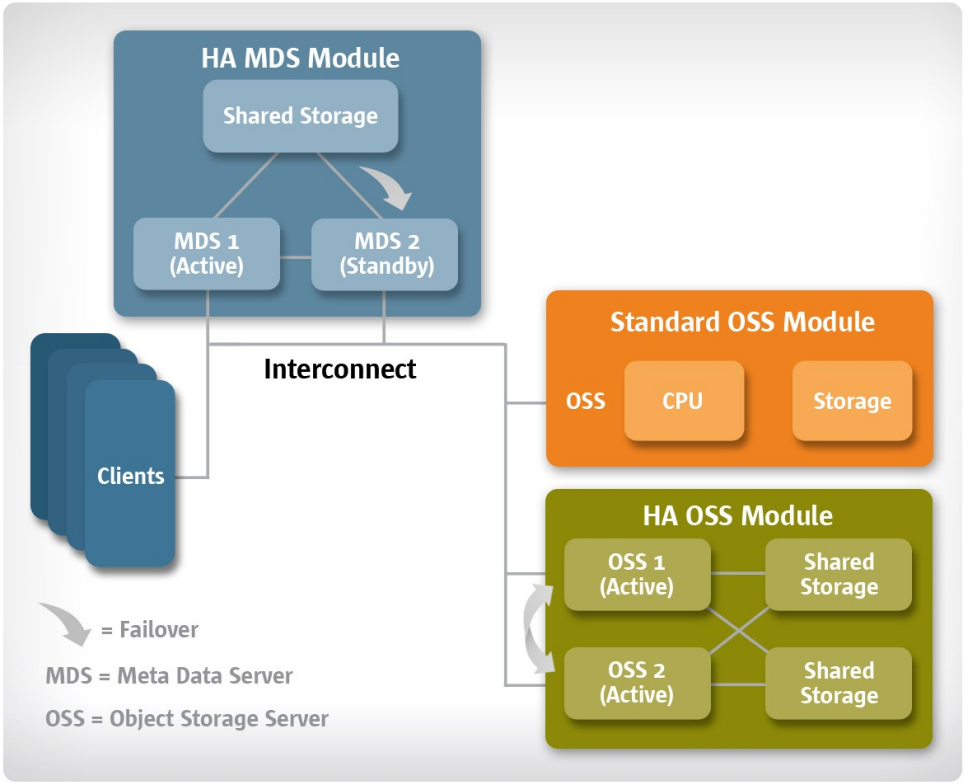
Simplifying Lustre for the mainstream

- Pre-defined modules speed deployment of parallel file system, available factory integrated
- Ease of use improvements – patchless clients, integrated driver stack, LNET self-test, mount- based cluster configuration, many more

(*) In October July 2007 Sun Microsystems, Inc. acquired Cluster File Systems, Inc.

Sun Lustre Storage System Modules

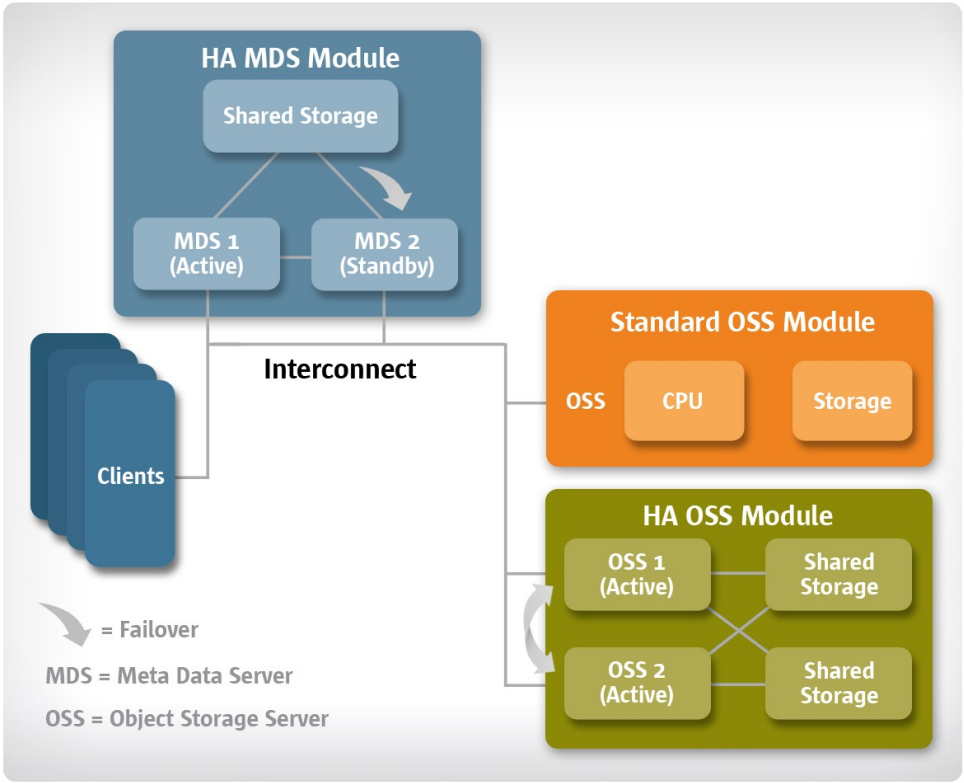
- ***Clients***
  Cluster computation, visualization, or desktop Linux nodes
- ***HA MDS Module***
  Manages and stores metadata, such as filenames, directories, permissions and file layout
- ***Standard & HA OSS Modules***
  Store application data, communicate directly to clients to transfer data



lustre®

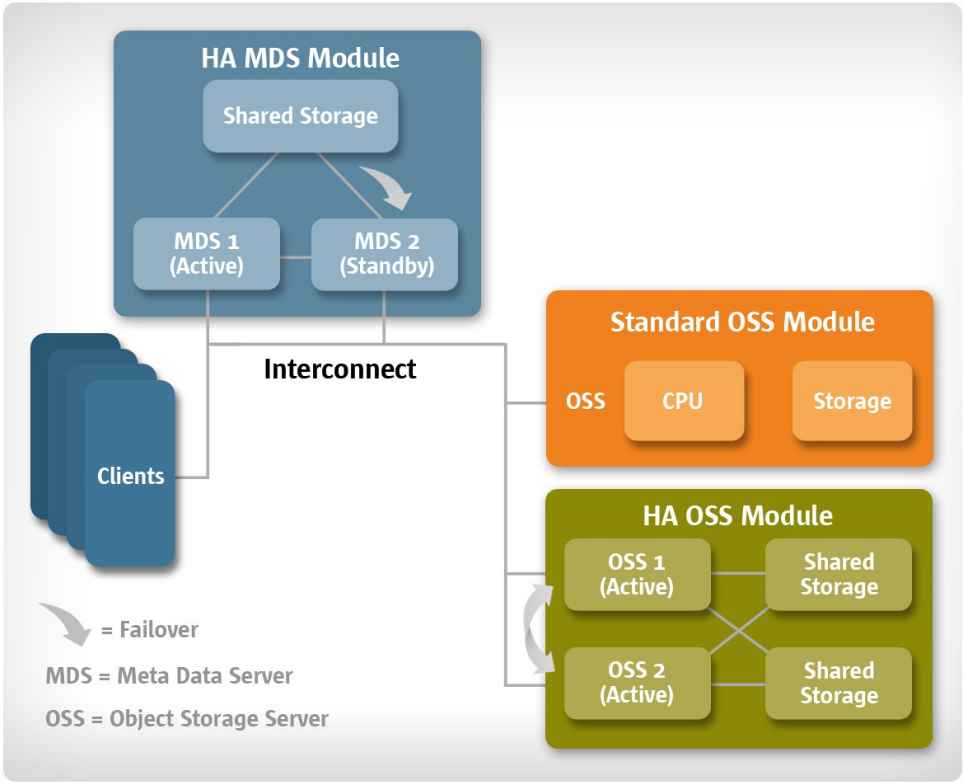
File System Clients

- Clients access storage through a reliable, high speed protocol called LNET
- Typical clients
  - > Are cluster computation, visualization, or desktop Linux nodes
  - > Number from hundreds to thousands, can scale to tens of thousands
  - > Each require tens to hundreds of MB/sec of I/O bandwidth
  - > Connect via InfiniBand, GE or 10GE
- Client nodes **not** included as part of the system



HA MDS Module Functions

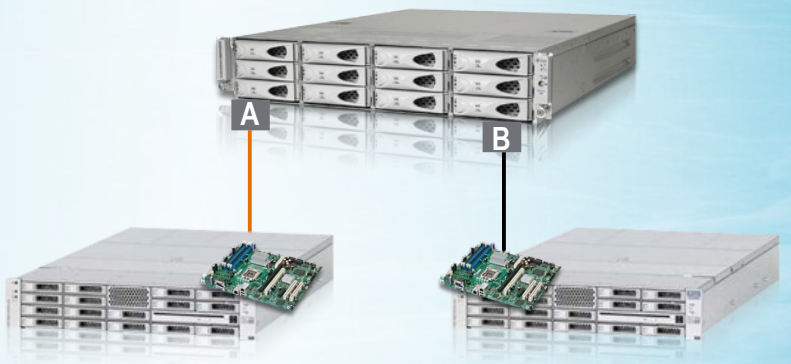
- Manages and stores metadata, such as filenames, directories, permissions and file layout
- Separation of metadata from application data accelerates I/O operations
- Active/Passive server pair with shared metadata storage for availability
- Single HA MDS module supports up to 250 million files* and up to 32 petabytes of file system capacity



lustre®

*Based on typical 4KB metadata entry per file and MDS storage available

HA MDS Module



MDS 1 (Active)

MDS 2 (Standby)

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

Hardware

- 2x Sun Fire X4250s each with
  - > Dual, quad-core Intel Xeon CPUs, 8GB RAM, dual SAS 2.5" boot drives, Sun 8-port SAS HBA
  - > Sun DDR 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

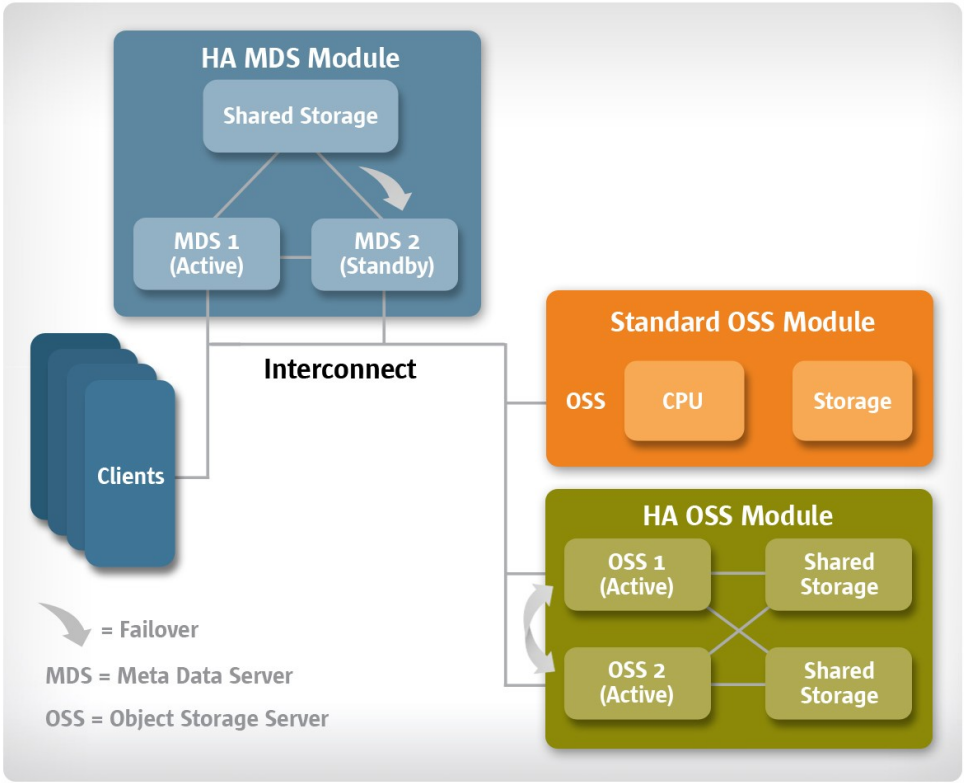
- Linux RAID 1+0 for metadata
- MDS configured in Lustre active/ passive pair
- Hot-swap redundant power and cooling

Management

- Integrated LOM Service Processor

Standard & HA OSS Module Functions

- Store application data, communicate directly to clients to transfer data
- Lustre stripes file systems across OSS to aggregate bandwidth
- Quantity typically ranges from 1s to 10s, but can scale 100s
- Standard OSS Module
  - > Simple, high density, internal storage
- HA OSS module
  - > Strong price/performance on shared storage with fail over for improved availability



lustre®

Standard OSS Module

Hardware

- Sun Fire X4540 Server with dual quad-core AMD Opteron processors
- 32 GB memory
- Sun DDR 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

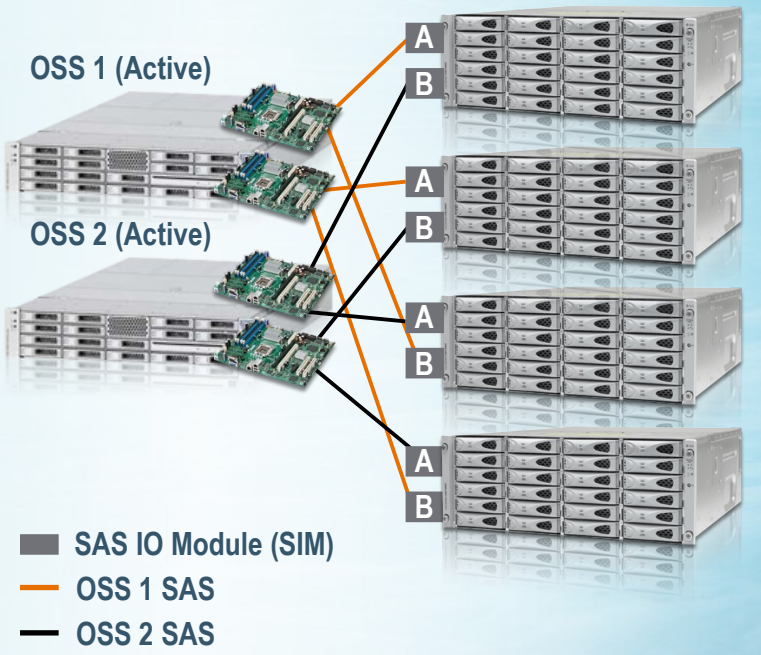
Availability

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

Management

- Integrated LOM Service Processor

HA OSS Module



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

Hardware

- 2x Sun Fire X4250s each with
  - > Dual, quad-core Intel Xeon CPUs, 8GB RAM, dual SAS 2.5" boot drives, Sun 8-port SAS HBA
  - > Sun DDR 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

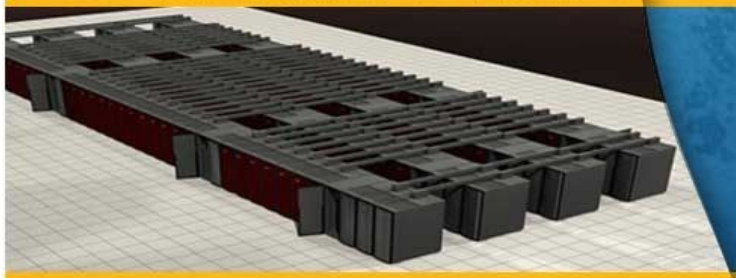
Management

- Integrated LOM Service Processor

Lustre in Action

TACC Ranger

1.73 PB storage, 35GB/s I/O throughput, 3,936 quad-core clients



Sandia Red Storm

340 TB Storage, 50GB/s I/O throughput, 25,000 clients

Framestore

"The Tale of Despereaux"

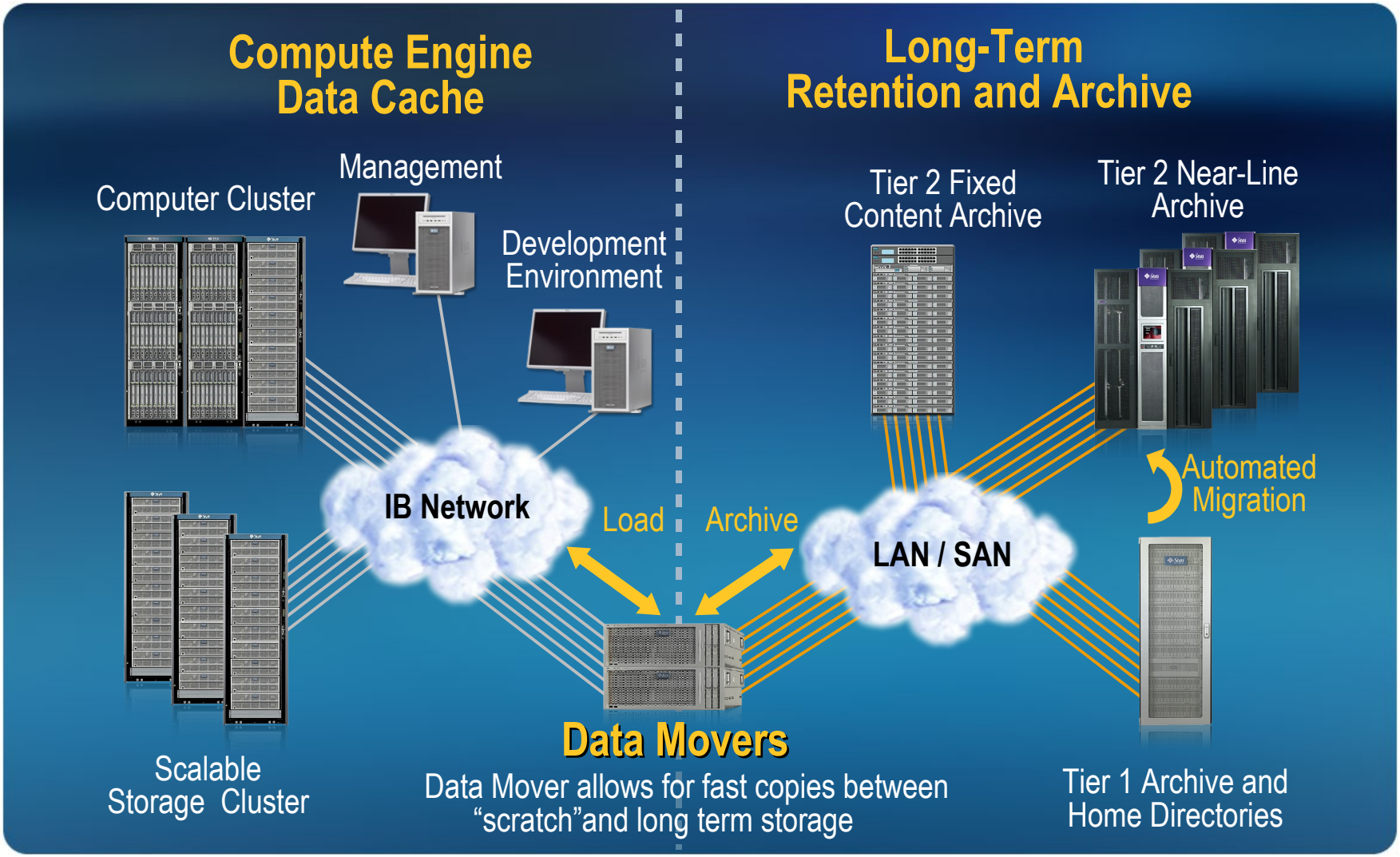
200TB Lustre file system – averaged 1.2GB/s in sustained reads, peaks of 3GB/s, 5TB data generated per night from cluster of 4,000 cores interfacing with Lustre file system

German Climate Research Data Centre (DKRZ)

272 TB, Lustre and Storage Archive Manager, 256 nodes with 1024 cores

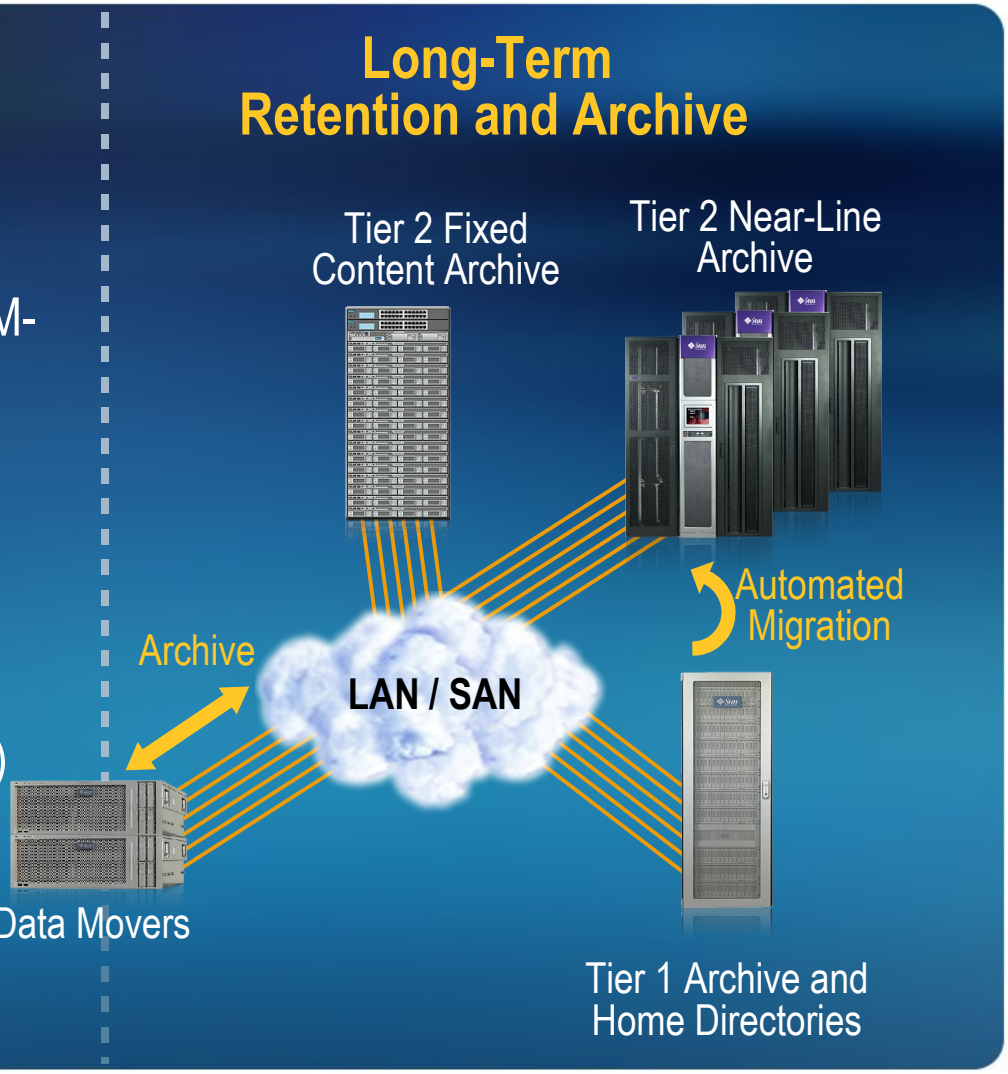
42% of top100 supercomputers using Sun Lustre

Data Movers


Compute Engine
Data Cache
Long-Term
Retention and Archive
Computer Cluster
Management
Development
Environment
Tier 2 Fixed
Content Archive
Tier 2 Near-Line
Archive
IB Network
Load
Archive
LAN / SAN
Automated
Migration
Scalable
Storage Cluster
Data Movers
Data Mover allows for fast copies between
“scratch”and long term storage
Tier 1 Archive and
Home Directories

Long-Term Retention & Archive

- Provides a massive on-line/near-line repository to complement the Scalable Storage Cluster
- Leverages Sun StorageTek Tape Libraries, Modular Arrays and SAM-QFS
- Policy driven engine to automate moving data sets in to and results out of the Object Storage Farm
- Enables Tape Libraries as a large near-line repository
- Stores data in open formats (TAR) allowing technology refresh and avoiding vendor lock-in
- Pre-integrate systems – arrive at your site ready to run



Sun Compute Cluster Pre-Integration

Your Choice of Configuration

Hardware Racked and Cabled

Software Installed and Configured

Sun Customer Ready Program

Delivering "Business Ready HPC" Systems

Agile Development

- Up to 90% faster
- Tools to tailor to specific needs
- You concentrate on your core business

Higher Quality, Lower Risk

- Up to 80% less installation issues
- Integration and testing in the factory
- Uses Sun's IS09000 certified manufacturing processes

Resources

Download

Download and try out:
Lustre
Sun Grid Engine
Sun HPC Software Linux Edition
Sun HPC Developer Preview
Visualization Software

Learn More

Visit: **sun.com/hpc**
hpc.sun.com

Read the latest news, view the latest offers, download the latest white papers and more

Subscribe

Subscribe to Radio HPC via **iTunes** and get regular updates on HPC technology from Sun and our partners

Join

Join the online HPC community at:
hpc.sun.com
and collaborate with Sun engineers and experts

Watercooler

Visit the HPC Watercooler at:
blogs.sun.com/hpc

and get the latest HPC news from around the globe

Try and Buy

Visit: **sun.com/tryandbuy**

to get a free 60-day trial on all of our new systems

THANK YOU.

Dalibor Kubacek
dalibor.kubacek@sun.com