Sun Microsystems

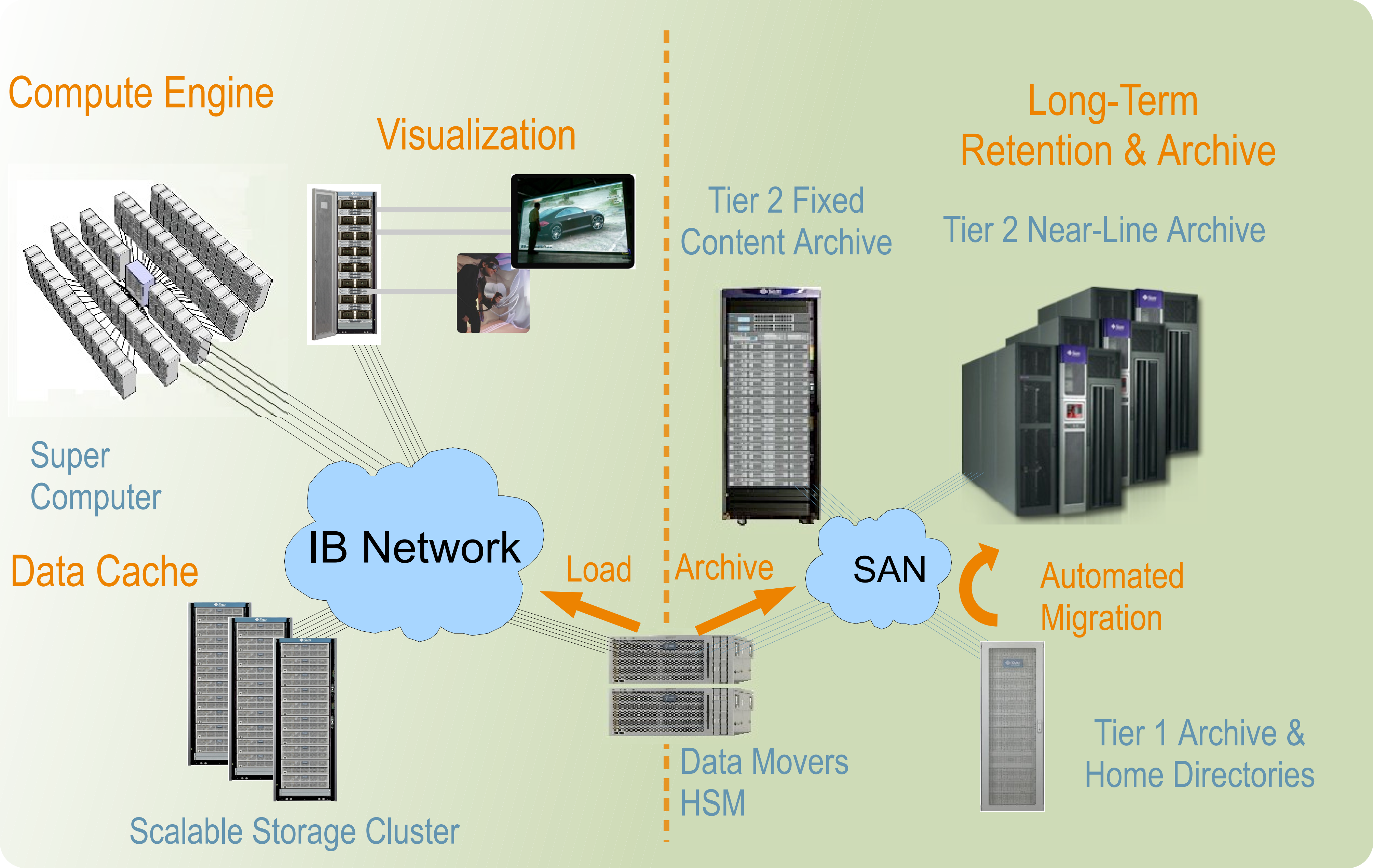
Lustre and Other Open Source Based Storage Projects

Dr Torben Kling-Petersen, PhD

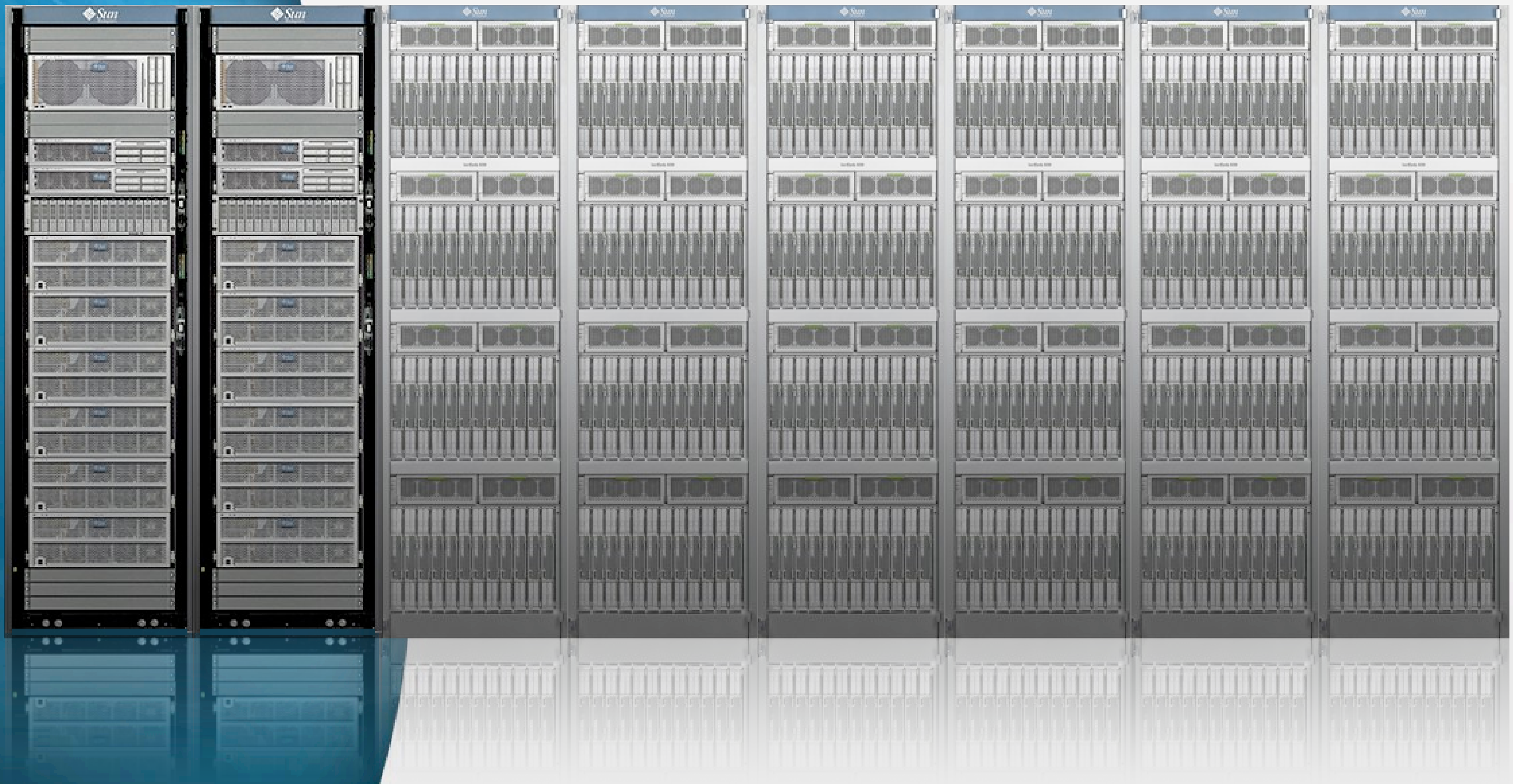
Senior HPC Technical Specialist

Sun Microsystems

Sun Microsystems HPC end-to-end architecture



Lustre and beyond



World’s Fastest and Most Scalable Storage

- Lustre is **Open Source**
- Lustre is the leading HPC file system
  - > 7 of Top 10
  - > 40% of Top 100
- Demonstrated Scalability and Performance
  - > 100GB/sec I/O; 25,000 clients
  - > Many systems with 1000 nodes
- Partners
  - > Bull, Cray, DDN, Dell, HP, Hitachi, SGI, Terascala

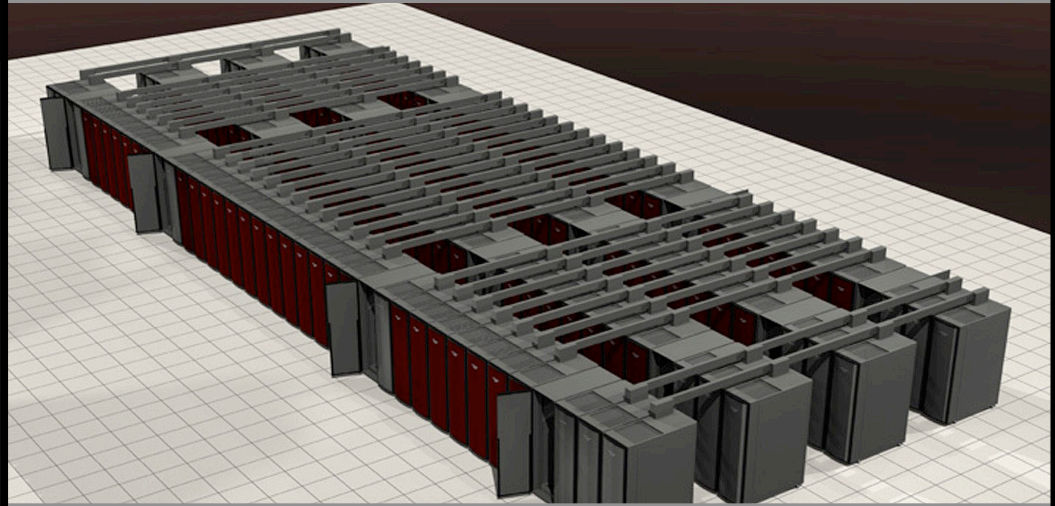
Livermore BlueGene/L

1.9 PB storage; 35.6 GB/s IO throughput; 212,992 client processes

TACC Ranger

1.73 PB storage; 40 GB/s IO throughput; 3,936 quad-core clients

Sandia Red Storm



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

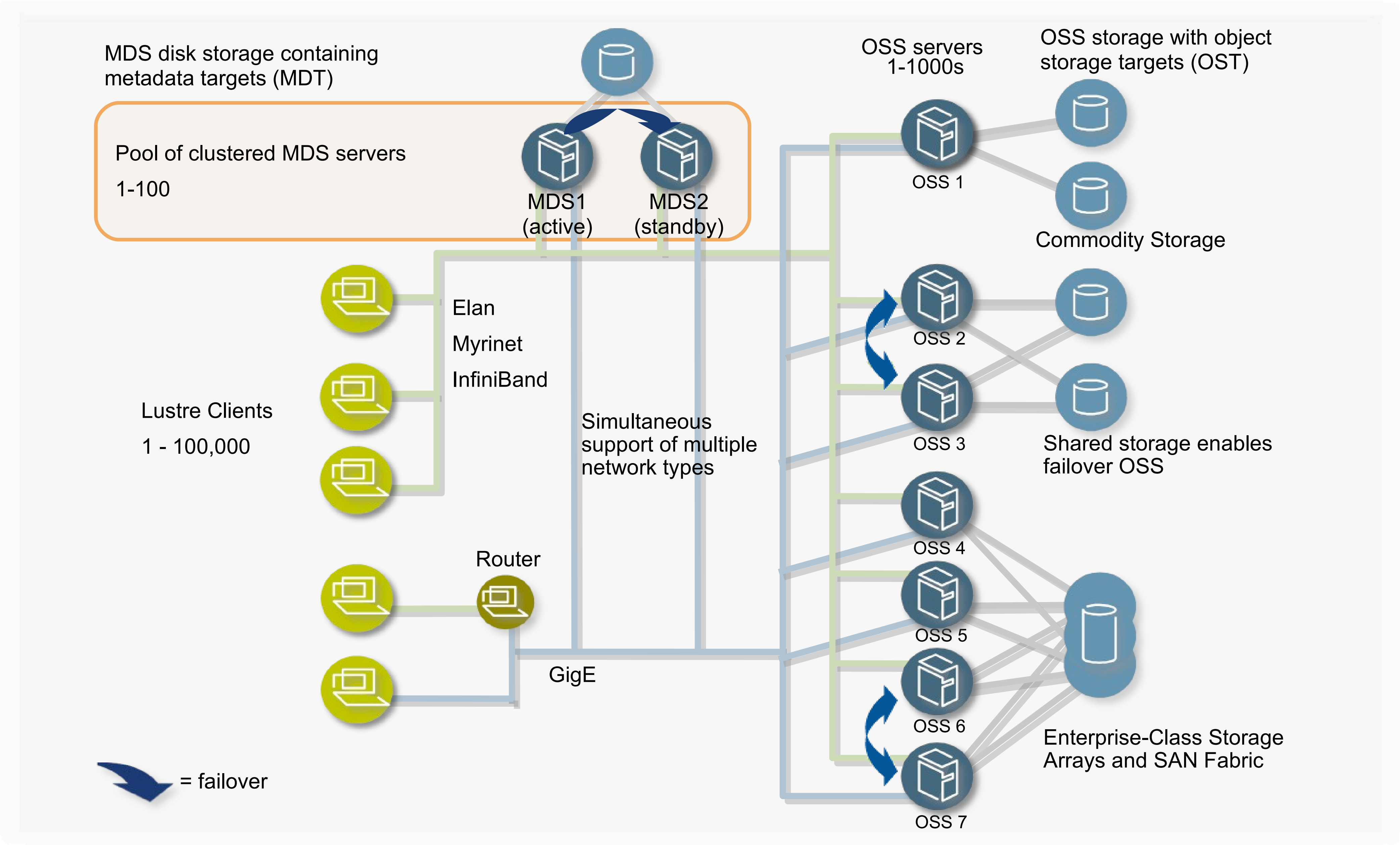
CEA Tera-10

1 PB storage; 100 GB/s I/O throughput; 4,352 dual-core clients

Lustre Today

WORLD RECORD #Clients	Clients: 25,000 – Red Storm Processes: 212,992 – BlueGene/L Can have Lustre root file systems
#Servers	Metadata Servers: 1 + failover OSS servers: up to 450, OST's up to 4000
Capacity	Number of files: 2 Billion File System Size: 32 PB, Max File size: 320 TB
WORLD RECORD Performance	Single Client or Server: 2 GB/s + BlueGene/L – first week: 74M files, 175TB written Aggregate IO (One FS): ~130GB/s (PNNL) Pure MD Operations: ~15,000 ops/second
Stability	Software reliability on par with hardware reliability Increased failover resiliency
Networks	Native support for many different networks, with routing
Features	Quota, Failover, POSIX, POSIX ACL, secure ports
Varia	Training, Level 1,2 & Internals. Certification for Level 1

A Lustre Cluster

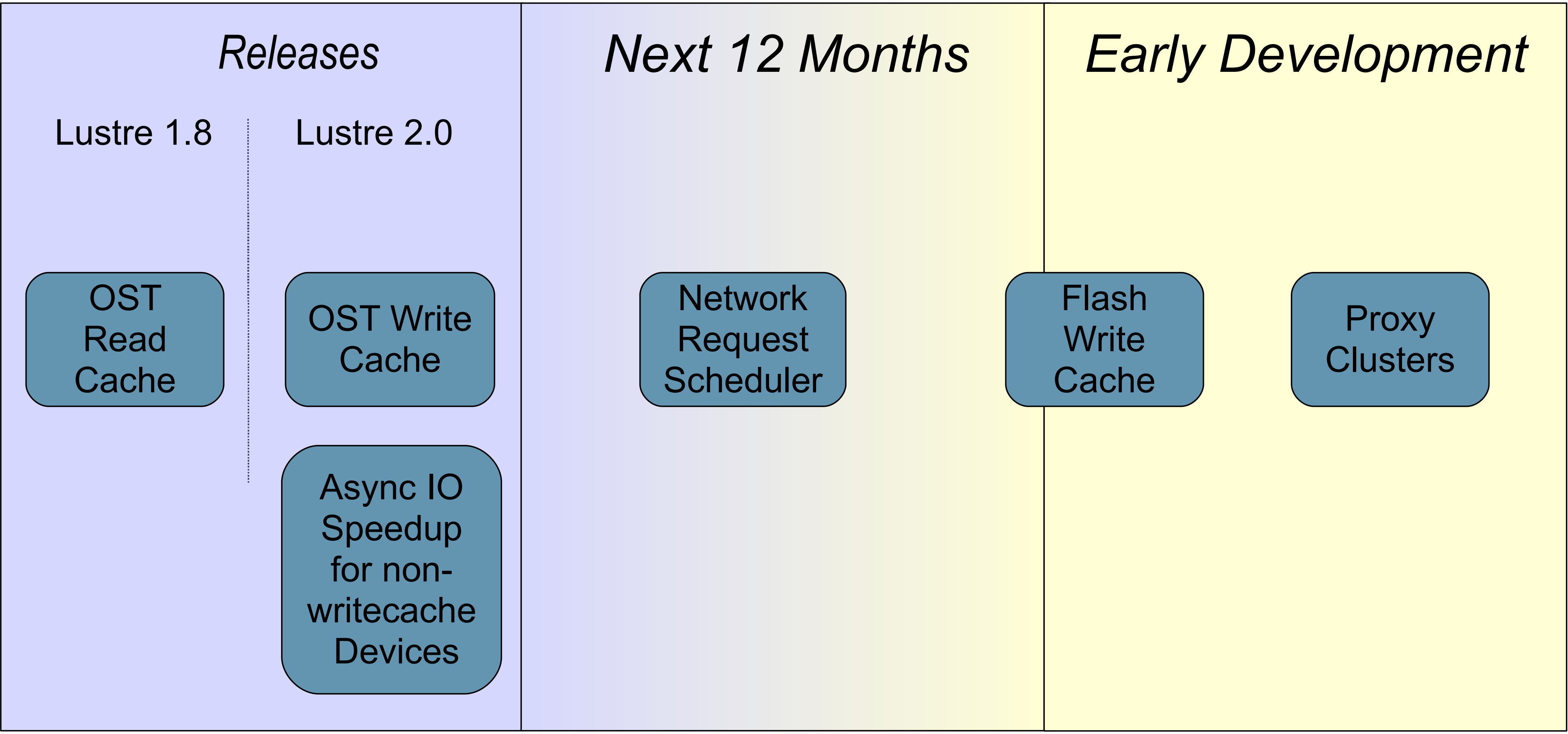

MDS disk storage containing metadata targets (MDT)
Pool of clustered MDS servers
1-100
MDS1 (active)
MDS2 (standby)
OSS servers 1-1000s
OSS storage with object storage targets (OST)
OSS 1
Commodity Storage
Elan
Myrinet
InfiniBand
Lustre Clients
1 - 100,000
OSS 2
OSS 3
Shared storage enables failover OSS
Simultaneous support of multiple network types
OSS 4
Router
OSS 5
GigE
OSS 6
Enterprise-Class Storage Arrays and SAN Fabric
OSS 7
= failover

Lustre requirements & futures

- Pools
- ZFS OSS & MDS
- Solaris
- Kerberos
- Migration
- Windows pCIFS
- Clustered MDS
- 1 PFlop+ Systems
- 1 Trillion files
- 1M file creates / sec

- Solid State Disk
- 30 GB/s mixed files
- 1 TB - 10 TB/sec throughput
- WB caches
- Small files
- Proxy Servers
- Disconnected Operation
- HSM
- WAN based Lustre

IO Performance



Releases

Lustre 1.8

OST Read Cache

Lustre 2.0

OST Write Cache

Async IO Speedup for non-writecache Devices

Next 12 Months

Network Request Scheduler

Flash Write Cache

Early Development

Proxy Clusters

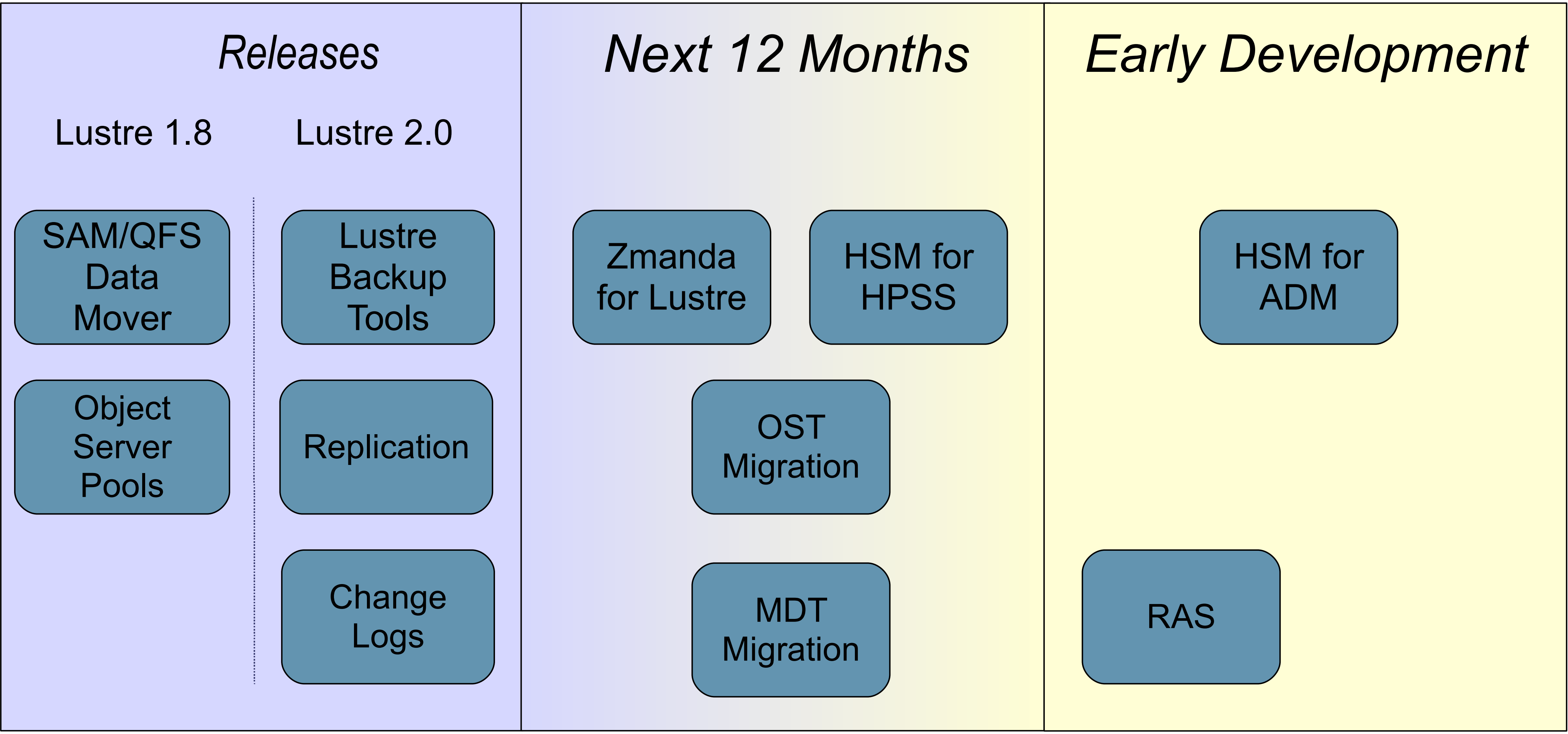
NB. Dates and feature sets are subject to change

End to End Data Integrity

Releases		*Next 12 Months*	*Early Development*	
Lustre 1.8	Lustre 2.0			
Lustre network checksums	Kerberos and GSS API	Checksum for ldiskfs	Checksum off-loading	Data integrity with ZFS

NB. Dates and feature sets
are subject to change

Information Lifecycle Management



Releases

Lustre 1.8

- SAM/QFS Data Mover
- Object Server Pools

Lustre 2.0

- Lustre Backup Tools
- Replication
- Change Logs

Next 12 Months

- Zmanda for Lustre
- HSM for HPSS
- OST Migration
- MDT Migration

Early Development

- HSM for ADM
- RAS

NB. Dates and feature sets
are subject to change

Lustre OEM Partners

BULL

CRAY
THE SUPERCOMPUTER COMPANY

DELL

DataDirect™
NETWORKS
Performance. Capacity. Innovation.

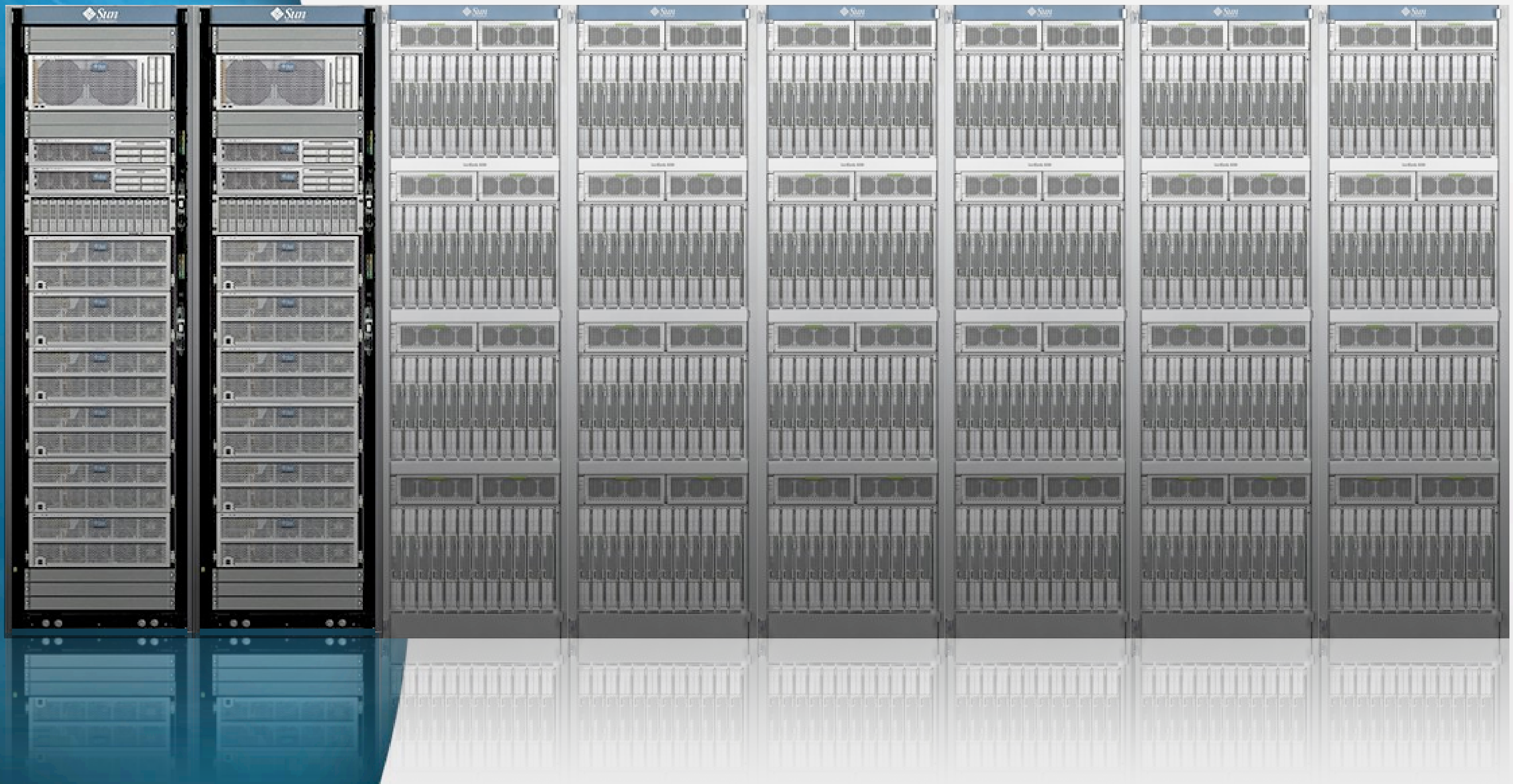
HITACHI
Inspire the Next

hp

sgi®

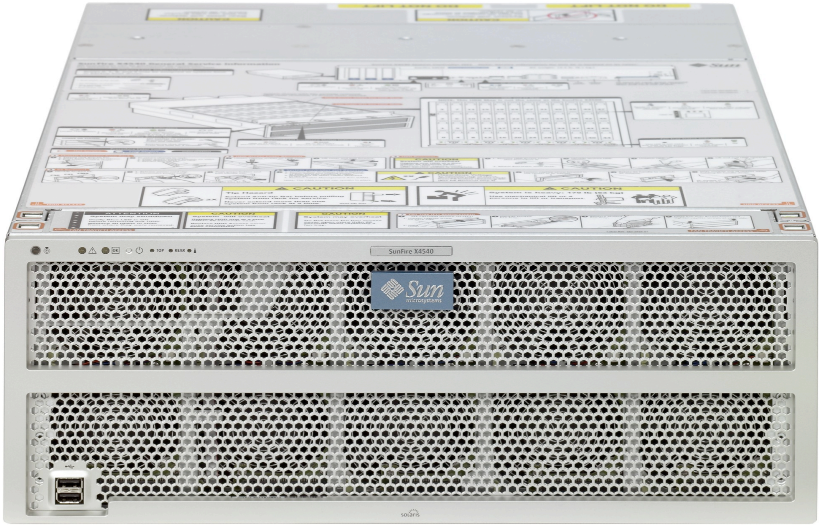
TERASCALA

OpenSolaris Storage Solutions



Sun Fire X4540 Storage Server

2-Socket Quad-Core Enterprise Server with 48 SATA hard drives direct attached



- **Compute**
  - **2 Quad-Core AMD Opteron processors Series 2300**
  - **16 DDR2 slots - 64GB Memory**
- **I/O**
  - **3x PCI-e 8-lane slots**
  - **4x Gigabit Ethernet ports**
  - **48x SATA 3.5" disk drives**
- **Availability**
  - **N+1 redundant hot-swap power supplies and fans**
  - **Software RAID**
- **Management and OS**
  - **Sun ILOM, IPMI 2.0; remote KVM, floppy/CDROM**
  - **Solaris OS (pre-installed)**
  - **Linux and Windows**

Zettabyte File System

InfoWorld 2008 Technology of the Year Awards

Best File System

End-to End Data Integrity

64-bit checksums
Copy-on-write transactions

Immense Data Capacity

World's first
128-bit file system

Easier Administration

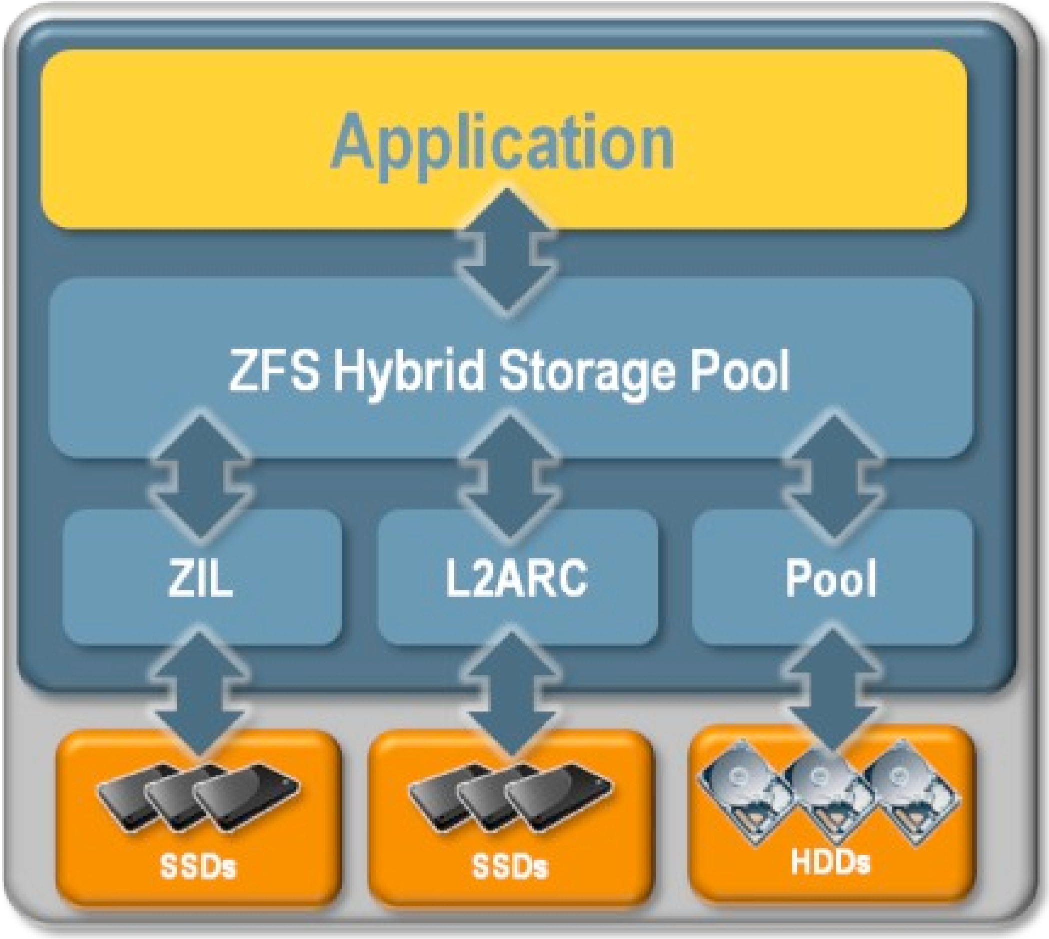
Pooled storage model– no volume manager
Move volumes between systems

Huge Performance Gains

Especially architected for speed

ZFS Turbo Charges Applications

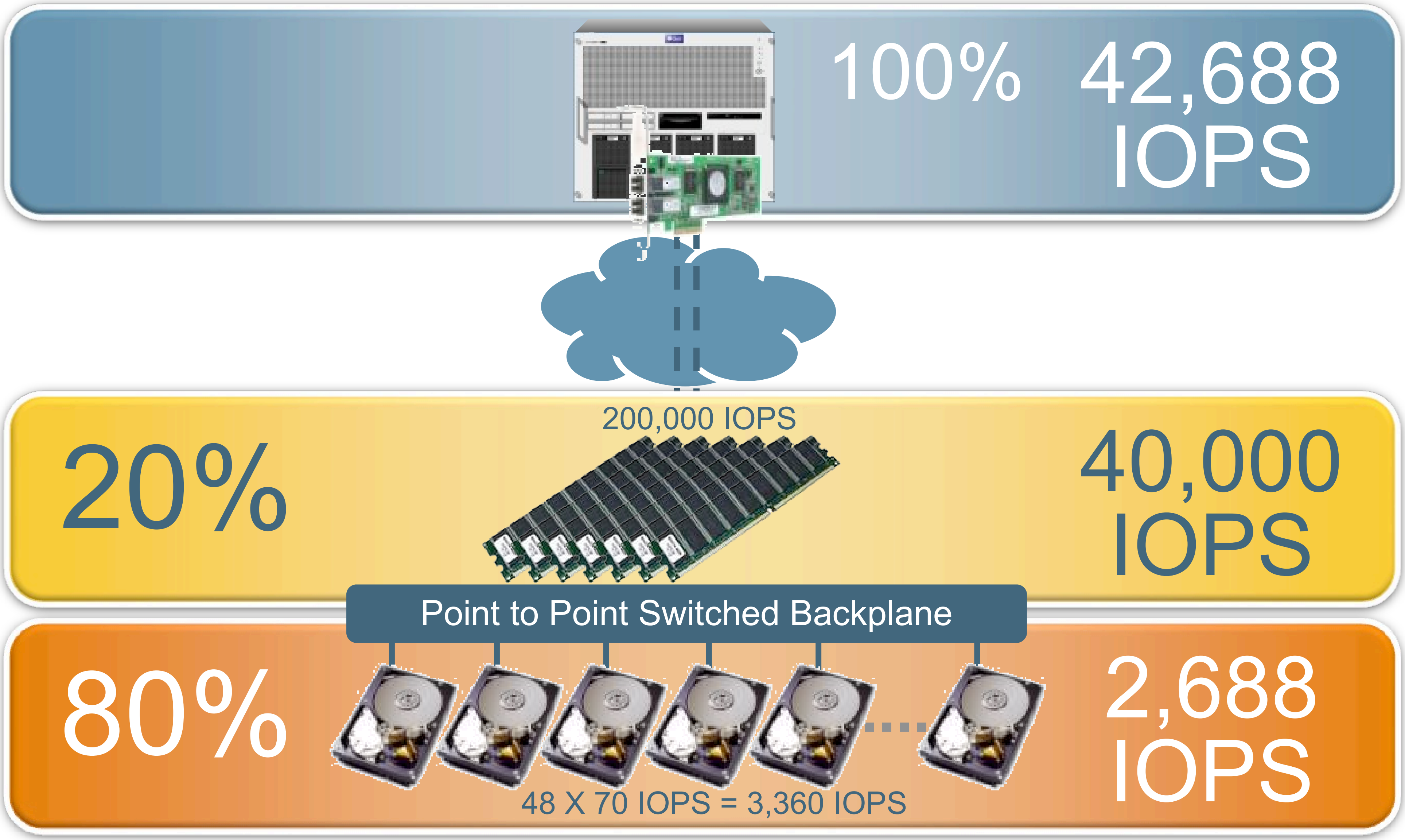
The Hybrid Storage Pool Data Management



- ZFS automatically:
  - >Writes new data to a very fast SSD pool (ZIL)
  - >Determines data access patterns and stores frequently accessed data in the L2ARC
  - >Bundles IO into sequential lazy writes for more efficient use of low cost mechanical disks

Standard HDDs Starve Servers

Cache vs. SATA Storage Pool



100% 42,688 IOPS

200,000 IOPS

20% 40,000 IOPS

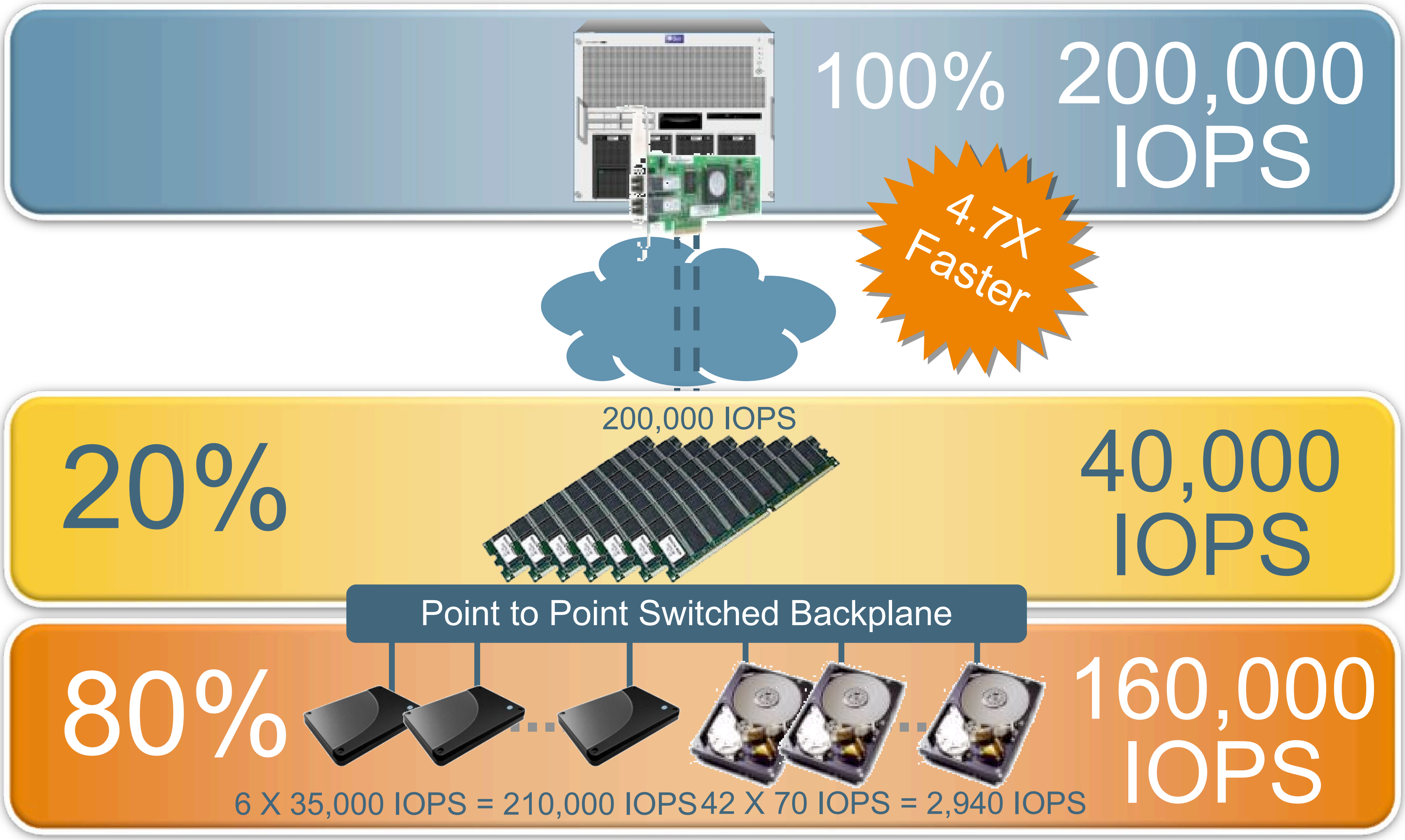
Point to Point Switched Backplane

80% 2,688 IOPS

48 X 70 IOPS = 3,360 IOPS

SSDs Turbo Charges Servers

Cache vs. Hybrid Storage Pool



100% 200,000 IOPS

4.7X Faster

200,000 IOPS

20% 40,000 IOPS

Point to Point Switched Backplane

80% 160,000 IOPS

6 X 35,000 IOPS = 210,000 IOPS

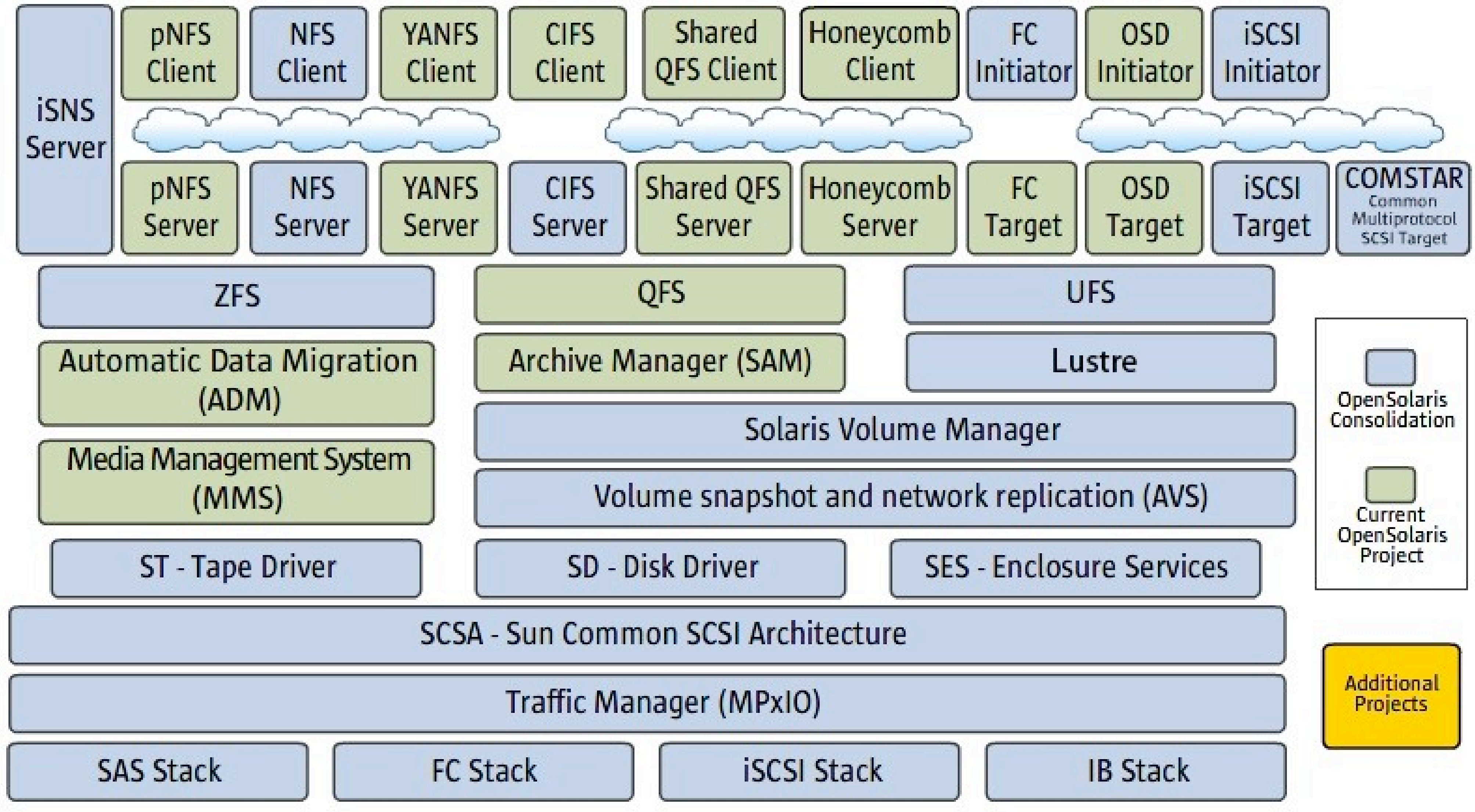
42 X 70 IOPS = 2,940 IOPS

What is Open Storage?

- **Open Architecture**
  General purpose hardware & software implementing storage functions that scale higher at lower TCO than proprietary alternatives
- **Open Software**
  Open sourced code and APIs to maximize the developer opportunity
- **Open Interoperability**
  Simple, predictable integration in heterogeneous environments (open standard)

- **OpenSolaris OS**
  - First OS with ZFS as default file system
  - Enhanced DTrace with D-Light
  - Fast in kernel CIFS server
  - Fully supported enterprise solution

OpenSolaris Storage Software



iSNS Server

pNFS Client
NFS Client
YANFS Client
CIFS Client
Shared QFS Client
Honeycomb Client
FC Initiator
OSD Initiator
iSCSI Initiator

pNFS Server
NFS Server
YANFS Server
CIFS Server
Shared QFS Server
Honeycomb Server
FC Target
OSD Target
iSCSI Target
COMSTAR Common Multiprotocol SCSI Target

ZFS
QFS
UFS

Automatic Data Migration (ADM)
Archive Manager (SAM)
Lustre

Media Management System (MMS)
Solaris Volume Manager
Volume snapshot and network replication (AVS)

ST - Tape Driver
SD - Disk Driver
SES - Enclosure Services

SCSA - Sun Common SCSI Architecture

Traffic Manager (MPxIO)

SAS Stack
FC Stack
iSCSI Stack
IB Stack

🟦 OpenSolaris Consolidation
🟩 Current OpenSolaris Project
🟨 Additional Projects

Innovate in Real Time with DTrace

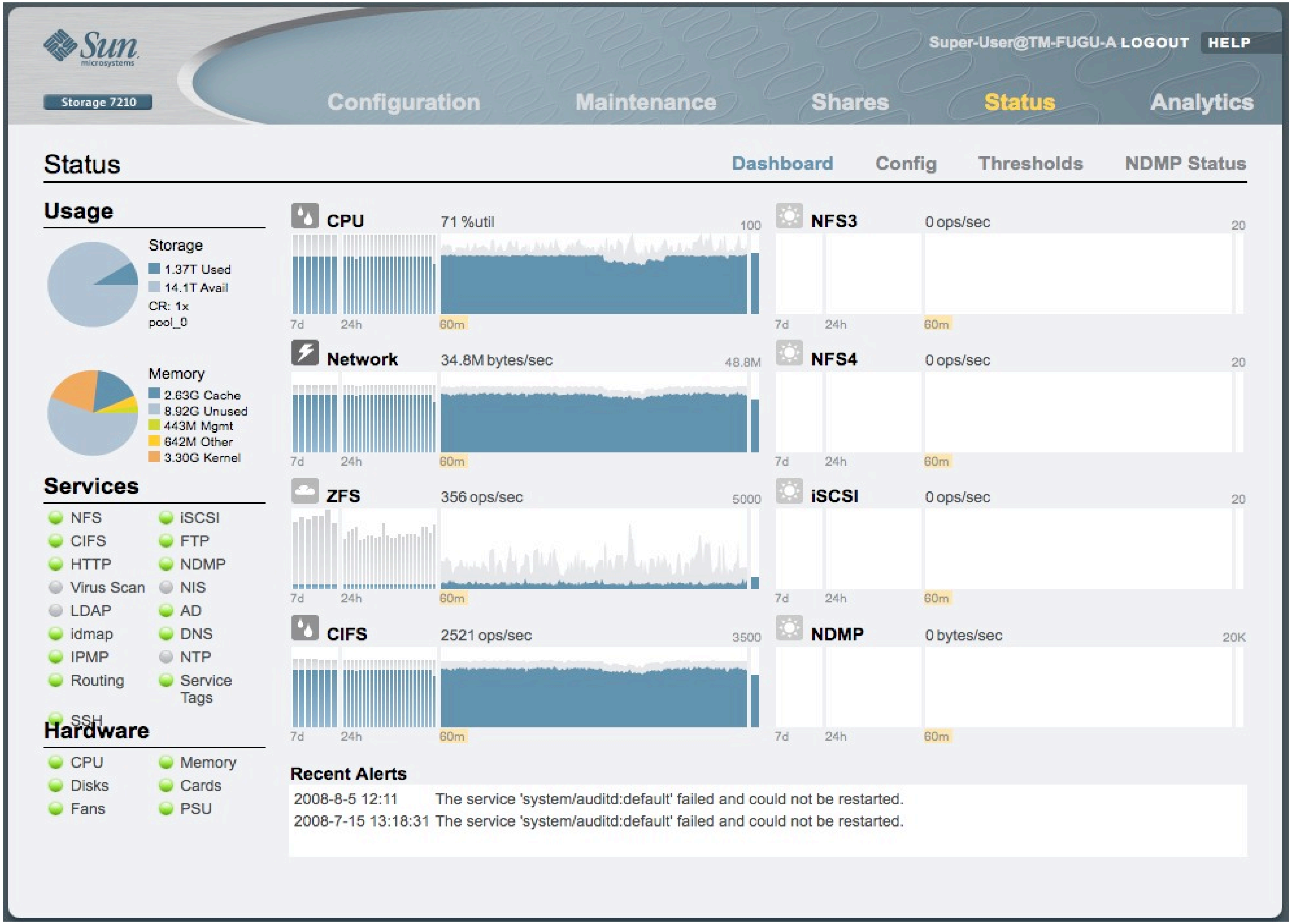
DTRACE -N 'PROC:::EXEC-SUCCESS { TRA
# FILES OPENED BY PROCESS,
DTRACE -N 'SYSCALL::OPEN*:ENTRY { PR
# SYSCALL COUNT BY PROGRAM,
DTRACE -N 'SYSCALL:::ENTRY { @NUM[E
# SYSCALL COUNT
DTRACE -N 'SYSCALL:::ENTRY { @NUM[E
# SYSCALL COUNT BY PROCESS,
DTRACE -N 'SYSCALL:::ENTRY { @NUM[P
# DISK SIZE BY PROCESS,
DTRACE -N 'IO:::START { PRINTF("%D %S
# PAGES PAGED IN BY PROCESS,
DTRACE -N 'VMINFO:::PGPGIN { @PG[E
# NEW PROCESSES WITH ARGUMENTS,
DTRACE -N 'PROC:::EXEC-SUCCESS { TRA
# FILES OPENED BY PROCESS,
DTRACE -N 'SYSCALL::OPEN*:ENTRY { PR

> “[expletive deleted] It’s like they saw *inside my head* and gave me the One True Tool.”

- View everything—from high level scripts to low level hardware
- Solve the gnarliest problems on the fly
- Safe enough to use in production, any time
- Serious fun for developers

Amber Road

Unprecedented real-time observability, first and only in market



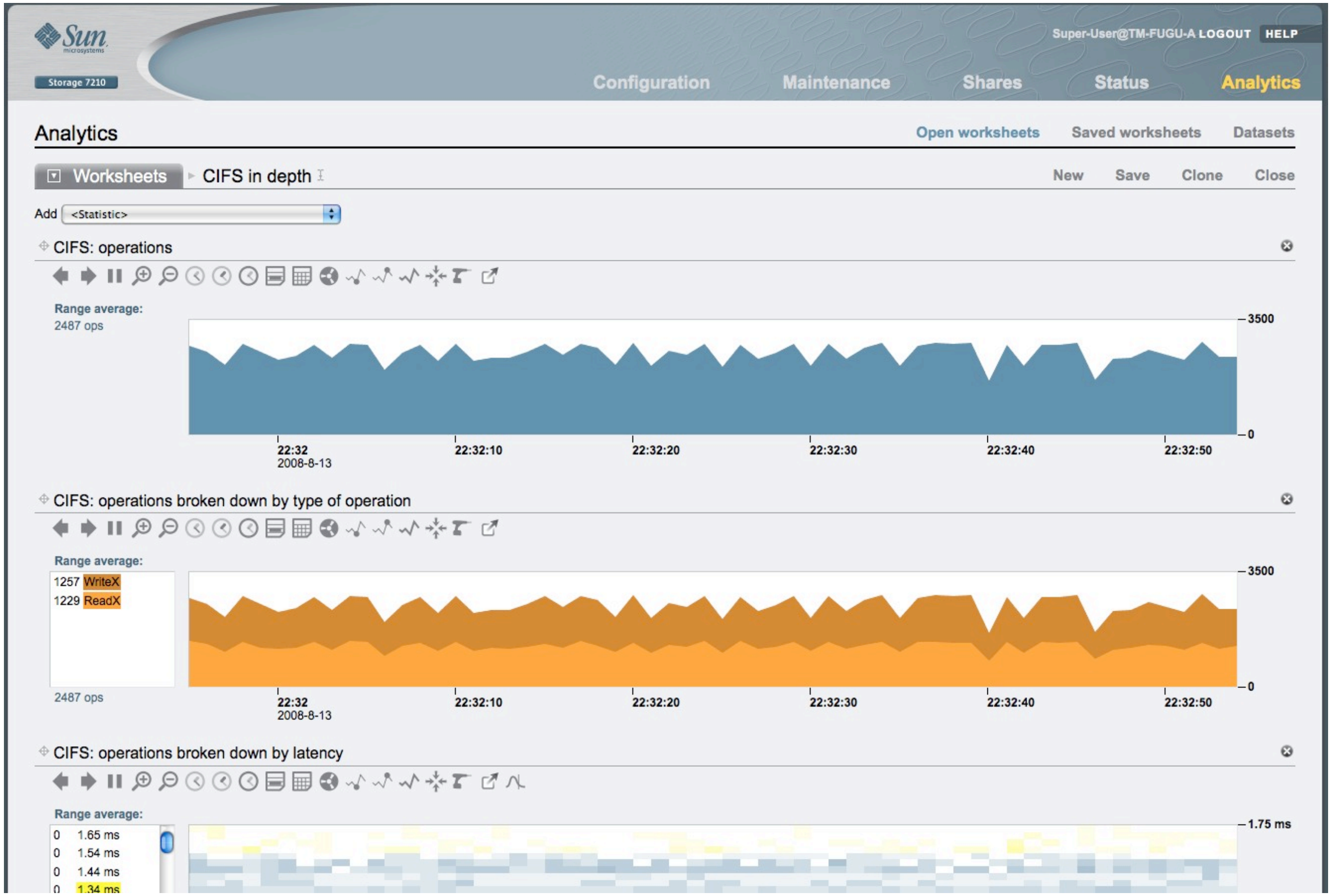
2008-8-5 12:11 The service 'system/auditd:default' failed and could not be restarted.

2008-7-15 13:18:31 The service 'system/auditd:default' failed and could not be restarted.

Amber Road

Key storage subsystems instrumented with DTrace

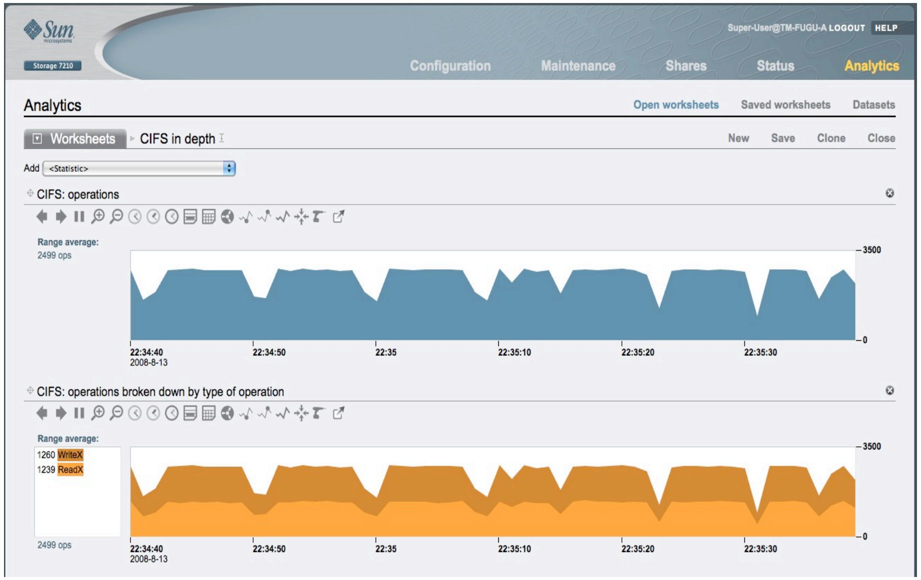
- NFS v3 and V4
- CIFS
- ISCSI
- ZFS
- CPU
- Memory Utilization
- Networking

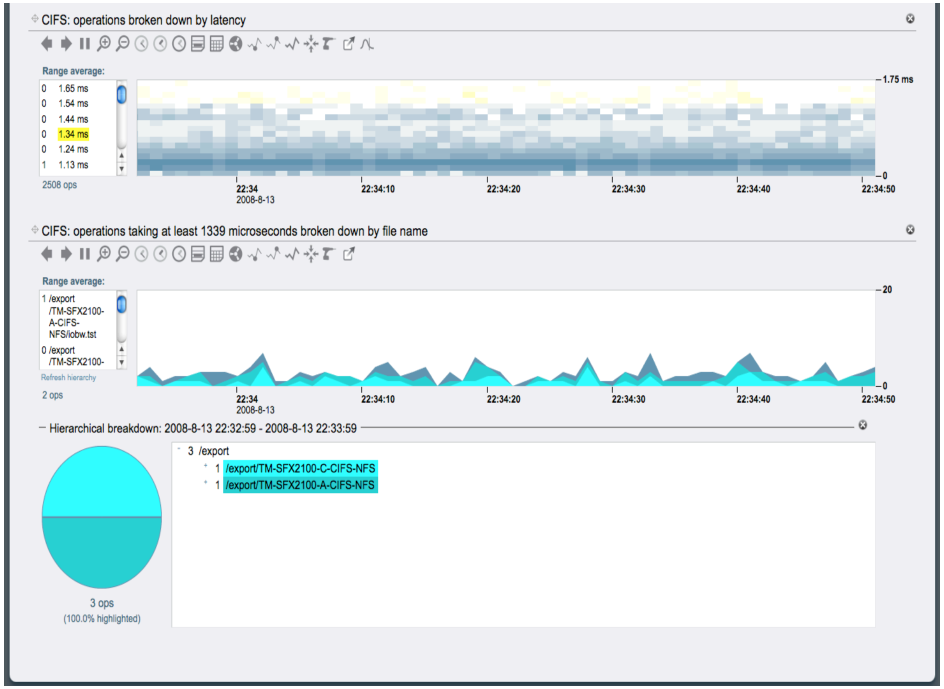


Amber Road

Real-time graphical business analytics

- “What files are hot right now?”
- “What is the distribution of reads and writes?”
- “Which clients are hot right now, displayed by protocol?”





- Group analytics into worksheets, can be made persistent
- Data can be exported for further number crunching, or can be saved on the appliance

pNFS Client/Server (Upcoming)

- Based on the NFS v4.1 spec.
- Separation of a NFS file system's data and metadata paths.
- pNFS features:
  - File virtualization with global namespace.
  - Namespace and data placement scale horizontally
  - Works with heterogeneous clients
    - Sun's pNFS server will serve up CIFS as well
  - Improved performance by striping a file's data across different file servers.

http://opensolaris.org/os/project/nfsv41/

OpenStorage

Storage Configurations

Open File Sharing

- File Sharing
  - NFS
  - CIFS
  - Shared QFS
  - Lustre
- File Services
  - Virus scanning
  - ZFS compression
  - ZFS encryption
- File System
  - ZFS
  - QFS
- Storage Management
  - ZFS pools

Open Data Sharing

- Block Sharing
  - ISCSI target
  - ISNS server
- COMSTAR
  - Fibre Channel target
  - *SAS, iSER, iSCSI targets*
  - *FCoE*
- Data Services
  - ZFS replication
  - ZFS snapshots
- Storage Management
  - *CAMs*

Open Archive

- Fixed Content
  - *Honeycomb*
- Tiered Storage Management
  - *Storage Archive Manager (QFS filesystems)*
  - *Automatic Data Migrator (ZFS file systems)*
- Tape/Storage Management
  - Media Mgmt System

INNOVATION MATTERS !!

”The free lunch is over”

- Compute density
  - Flops/watt
- Interconnect technologies
  - In CPUs, on mother boards etc
  - Between systems (InfiniBand anyone??)
  - Data I/O technologies
- Storage scalability
  - Latency, Performance, Capacity, ILM ...
- Management
  - Hardware (provisioning, upgrade & monitoring)
  - Software (OS, application and patching)
  - Data management
  - Network elements and Storage systems
  - People and procedures
- Servicability and upgradability

Open Source Matters !!

Download, test, deploy and enjoy ...

lustre®

http://wiki.lustre.org

http://opensolaris.org

opensolaris

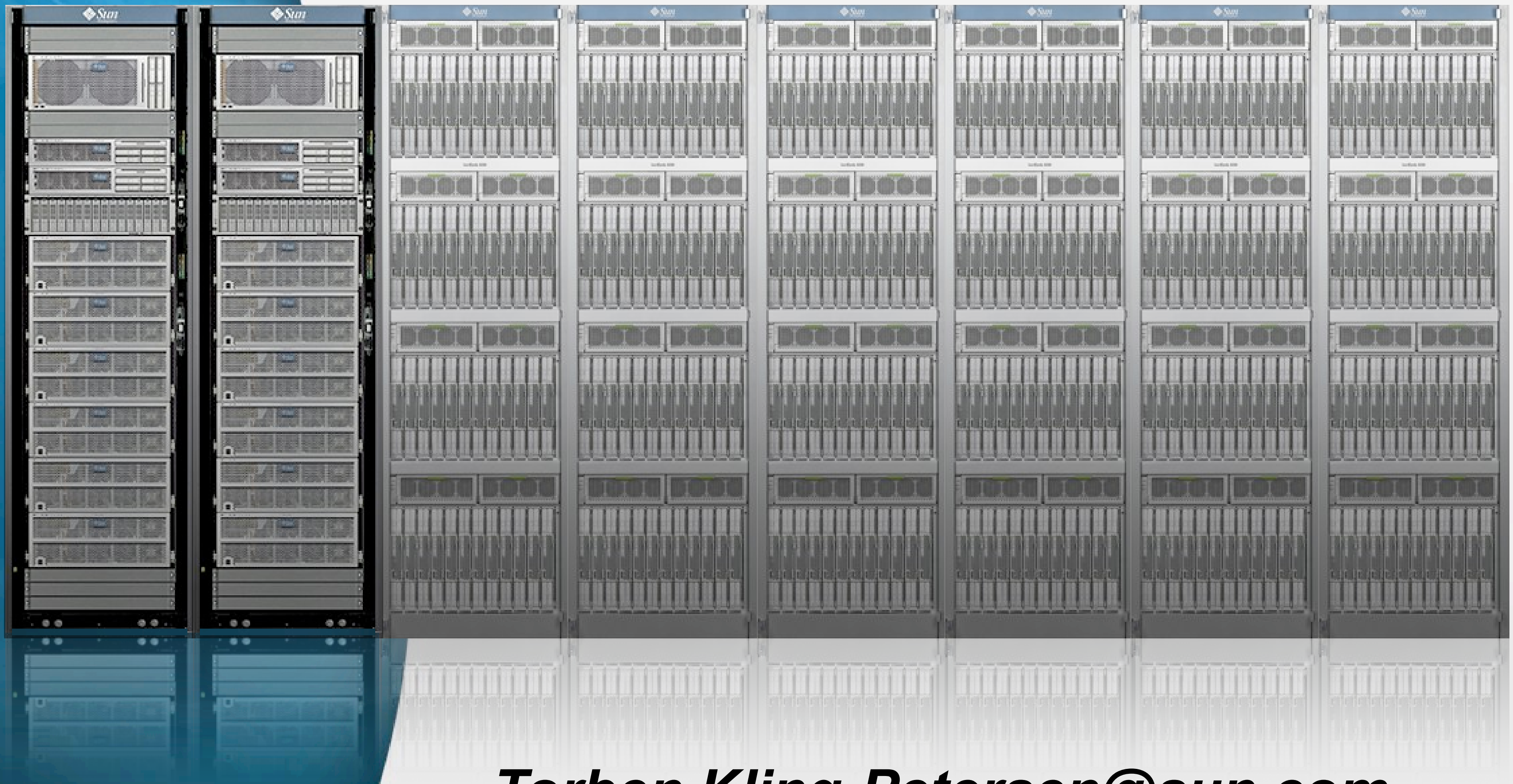
http://opensolaris.org/os/storage

Sun Open Source :

Dtrace, ZFS, GridEngine, MySQL, Glassfish, OpenOffice, NFS etc.

MySQL

Questions ??



Torben.Kling-Petersen@sun.com