Sun Grid & Cluster File Systems

Robert Read

Senior Staff Engineer

Sun Microsystems

May 2008

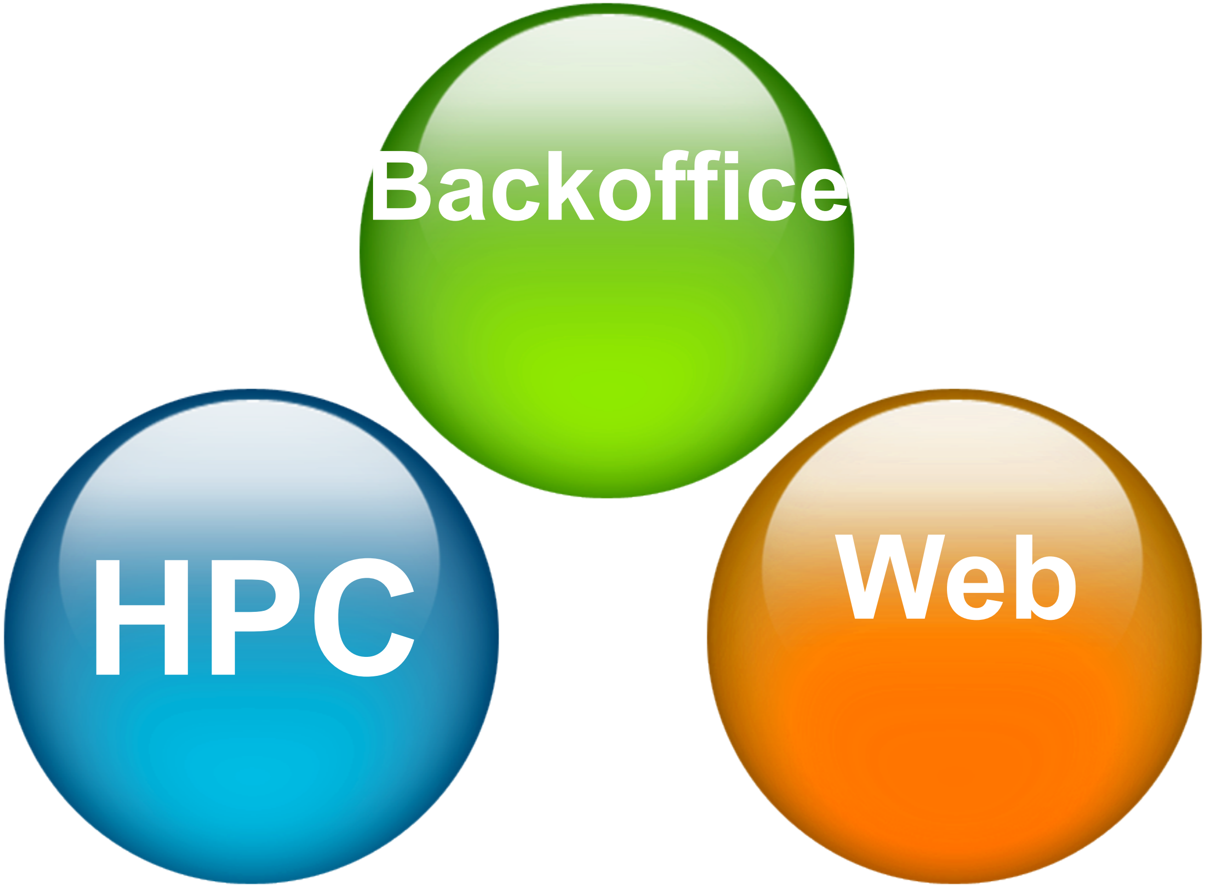
Agenda

- Future of Storage – Sun’s vision
- QFS
- Lustre
- pNFS

Sun’s view on storage

introduction

The IT Infrastructure



Backoffice

HPC

Web

Big Changes

- Everything is a cluster
- Open Source everywhere (Computer, Network, Storage)
- Fully virtualized processing, IO, and storage
- Integration, datacenter as a design center

NOW

COMPUTE:
Many cores, many threads, open platforms

COMING

STORAGE OPEN PLATFORMS:
$/performance
$/gigabyte

NETWORKING:
Huge bandwidth
Open platforms

What's Ahead

Open Servers

- Leveraging innovative product design and packaging
- Common components
- Open source software
- Wide interoperability to deliver breakthrough economics

Open Storage

A storage architecture that leverages:

- Open software
- An open architecture
- Common components
- Open interoperability to create innovative storage products
- Delivers breakthrough economics

Open Networks

- Unified datacenter network that utilizes common components
- Open source software
- Seamless integration with exitsting evironments
- Delivers breakthrough ecomonics

QFS

Solaris cluster file system

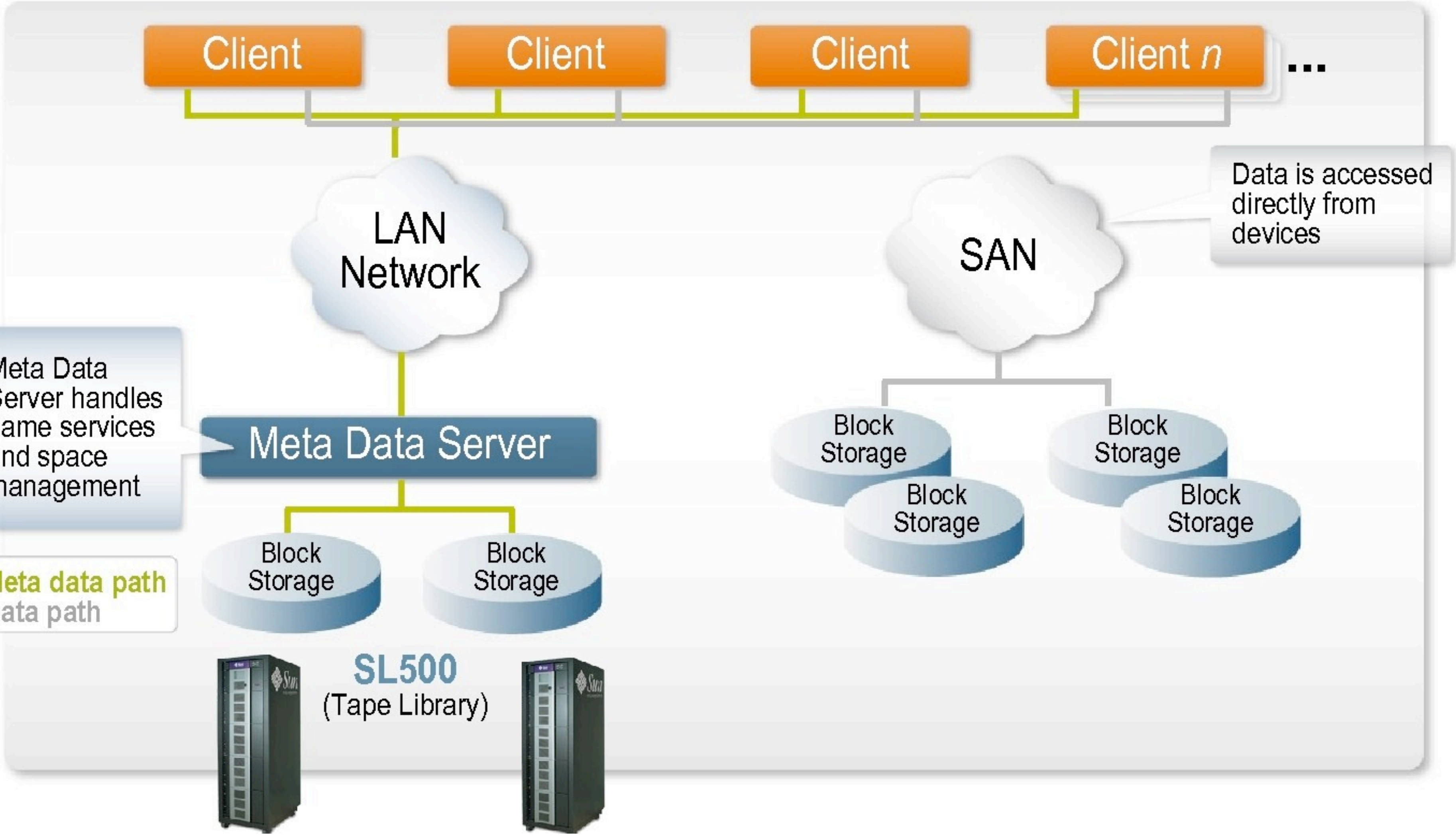
Sun's Advanced HPC Data Management Product Today

- Sun Storage Tek QFS – SAN File System
  - High performance parallel file system
  - Transparent user interface
  - Production ready
  - http://www.sun.com/storagetek/management_softare/data_management/qfs
- Sun Storage Tek Storage and Archive Manager (SAM)
  - Policy based automatic data migration and protection
  - Full device streaming
  - Tiered storage
  - http://www.sun.com/storagetek/management_software/data_management/sam

Shared QFS (SQFS)

- Large, existing and royal customer base
  - stable base, shipping since Aug 2002
- Target large enterprise, grid and HPC
  - Clients run on Solaris (SPARC & X64) & Linux
  - Metadata server run on Solaris (SPARC & X64)
  - HA option with SunCluster
- Built in HSM with SAM
- SQFS currently supports 256 nodes
- Next release, SQFS will support thousands of nodes
  - Targets HPC clusters

SAM-QFS Shared File System with Tiering



Data is accessed directly from devices

Meta Data Server handles name services and space management

Meta data path
Data path

SL500
(Tape Library)

Shared QFS Customer Benefits

- Data consolidation with SAN file sharing
  - HBO - 5000 hours of programming to manage
    - “Provided the scalability to store and manage large files created by program-length video with the performance necessary to meet HBO’s demanding throughput goals “
- Performance and scalability
  - Tune file system to the application
  - Near raw I/O performance
    - File system I/O performance scales linearly with the hardware
- Parallel processing W/multi-node read/ write access
- SAM provides automatic data protection with tiered storage

Shared QFS Certified w/SunCluster

- SunCluster HA failover support
  - Standalone QFS
  - HA-NFS over QFS
  - Shared QFS Metadata Server failover
    - Supports clients outside the cluster
- Oracle RAC runs on Shared QFS with SunCluster for high availability
  - Oracle certified on 9i and 10g
  - Shared QFS license is free for this configuration
- Shared QFS transactional performance matches raw

Lustre

introduction

World's Fastest and Most Scalable Storage

- Lustre is the leading cluster file system
  - 7 of Top 10 HPC systems
  - Half of Top 30 HPC systems
- Demonstrated Scalability and Performance
  - 100 GB/sec I/O
  - 25,000 Clients
  - Many systems with 1000s of nodes

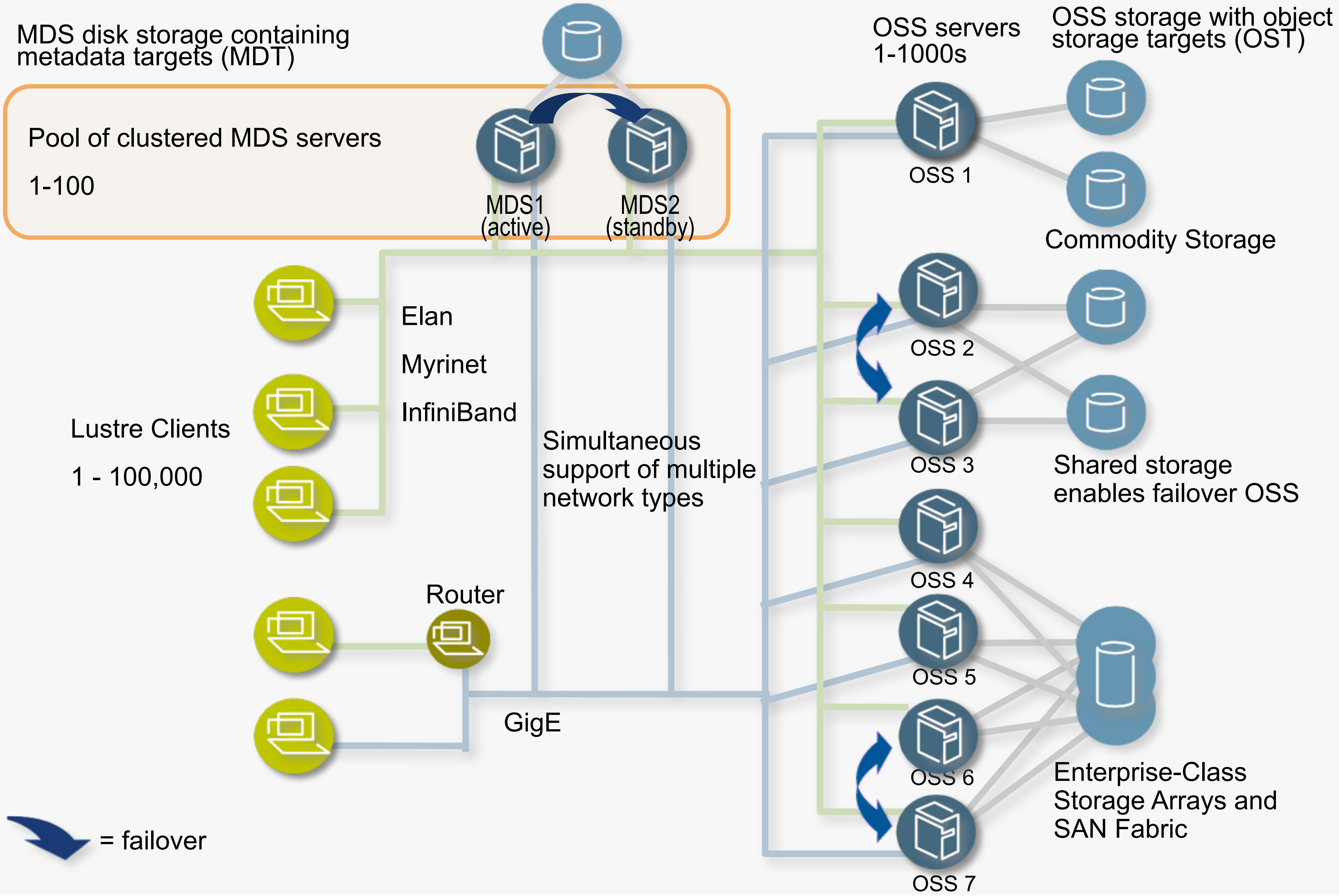
Lustre – scalable file system

- Lustre is a shared file system
  - Software only solution, no hardware ties
  - Developed as company – gvmt lab collaboration
  - Open source, modifiable, many partners
  - Extraordinary network support
  - Smoking performance and scalability
  - POSIX compliance and High Availability
- Lustre is for “extreme storage”
  - Horizontal scaling of IO over all servers
    - parallelizes I/O, block allocation and locking
  - Similar for metadata over MDS servers
  - add capacity by adding servers
  - Example: week1 of LLNL BG/L system: 75M files, 175TB

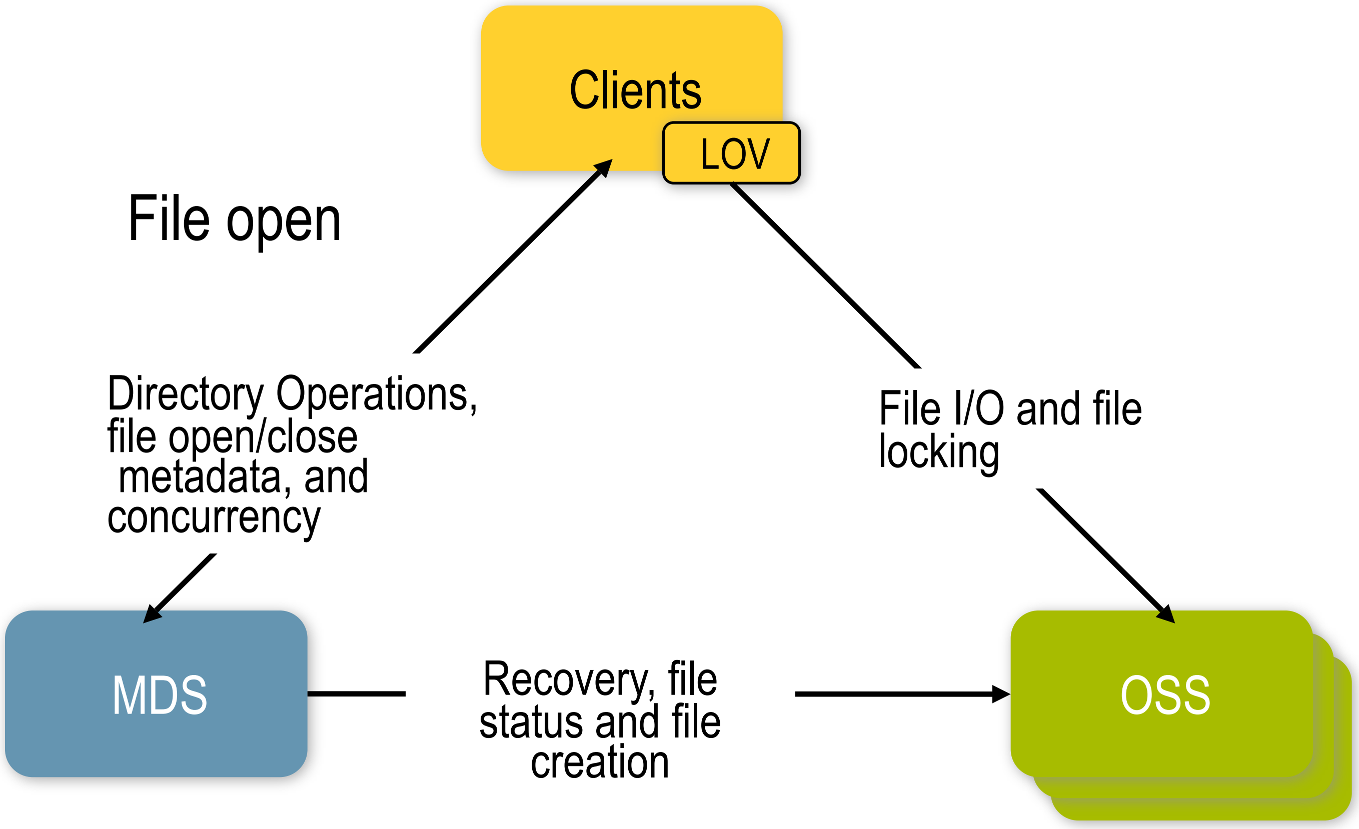
What kind of deployments?

- Extremely Large Clusters
  - Deployment: extremely high node count, performance
  - Where: government labs, DoD
  - Strengths: modifiability, special networking, scalability
- Medium and Large Clusters
  - Deployment: 32 – low thousands of nodes
  - Where: everywhere
  - Strengths: POSIX features, HA
- Very large scale data centers
  - Deployments: combine many extremely large clusters
  - Where: LLNL, ISP’s, DoD
  - Strengths: security, networking, modifiability, WAN features

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
OSS 2
OSS 3
Shared storage enables failover OSS
OSS 4
OSS 5
OSS 6
OSS 7
Enterprise-Class Storage Arrays and SAN Fabric
Elan
Myrinet
InfiniBand
Simultaneous support of multiple network types
Lustre Clients
1 - 100,000
Router
GigE
= failover

How does it work?



Clients

LOV

File open

Directory Operations, file open/close metadata, and concurrency

File I/O and file locking

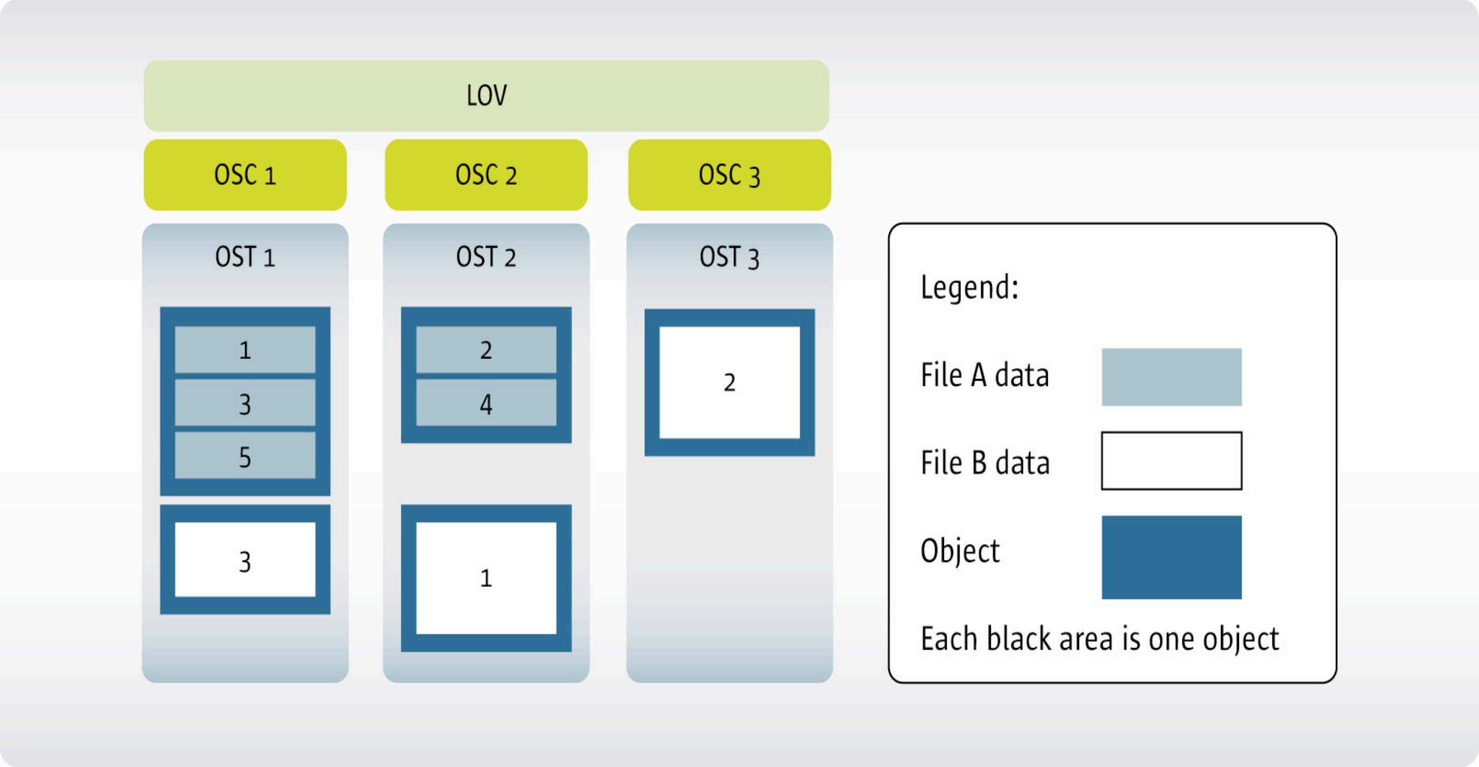
MDS

Recovery, file status and file creation

OSS

Lustre Stripes Files with Objects

- Currently objects are simply files on OSS resident file systems
- Enables parallel I/O to one file
  - Lustre scales that to 100GByte/sec to one file



Lustre – without latency

client write back cache & wide area replicas

Metadata WBC & replication

- Goal & problem:
  - > Disk file systems make updates in memory
  - > Network FS's do not - metadata ops require RPCs
  - > The Lustre WBC should only require synchronous RPCs for cache misses
- Key elements of the design
  - > Clients can determine file identifiers for new files
  - > A change log is maintained on the client
  - > Parallel reintegration of log to clustered MD servers
  - > Sub-tree locks – enlarge lock granularity

Uses of the WBC

- HPC
  - I/O forwarding makes Lustre clients I/O call servers
  - These servers can run on WBC clients
- Exa-scale clusters
  - WBC enables last minute resource allocation
- WAN Lustre
  - Eliminate latency from wide area use for updates
- HPCS
  - Dramatically increase small file performance

General purpose replication

- Driven by major content distribution networks
  - DoD, ISPs
  - Keep multi petabyte file systems in sync
- Implementing scalable synchronization
  - Changelog based
  - Works on live file systems
  - No scanning, immediate resume, parallel
- Many other applications
  - Search, basic server network striping

pNFS

Standards based HPC file system

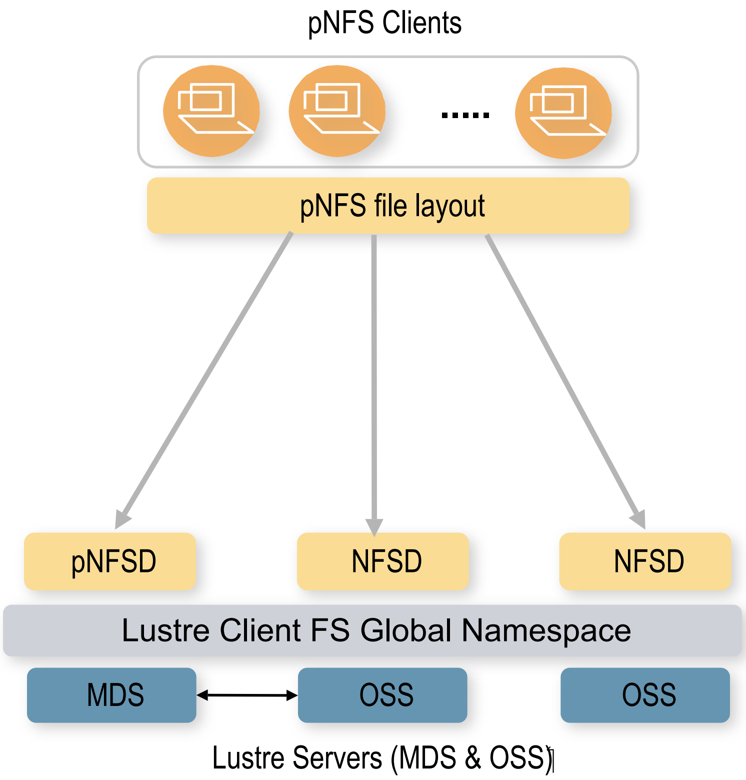
What is pNFS?

- pNFS is a standards based effort to provide I/O with a similar architecture as Lustre.
- Sun expects pNFS to play an important role in commercial HPC and later in the data center.

pNFS & Lustre

- pNFS integration
- Soon – pNFS exports from Lustre on Linux
  - First participation in a Bakeathon by Lustre!
- Longer term possibilities
  - Let Lustre servers offer pNFS & Lustre protocol
    - Requires an interesting Lustre storage layer
  - Make LNET an RDMA transport for NFS?
  - Offer proven Lustre features to NFS standards efforts

Layered & direct pNFS



pNFS Clients

.....

pNFS file layout

pNFSD

NFSD

NFSD

Lustre Client FS Global Namespace

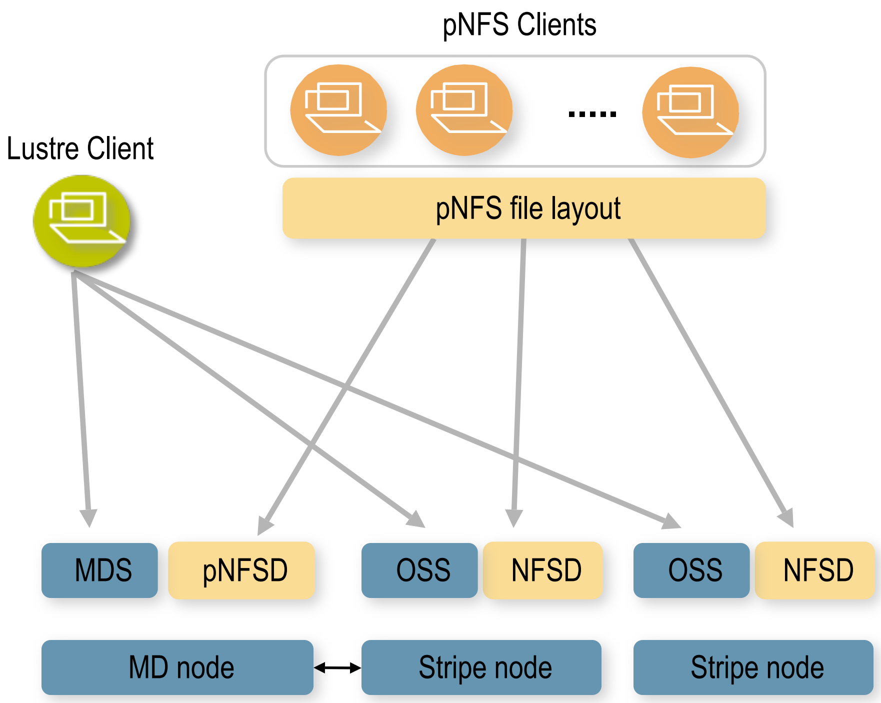
MDS

OSS

OSS

Lustre Servers (MDS & OSS)

pNFS layered on Lustre Clients



pNFS Clients

.....

Lustre Client

pNFS file layout

MDS

pNFSD

OSS

NFSD

OSS

NFSD

MD node

Stripe node

Stripe node

pNFS and Lustre servers on
Lustre / DMU storage system

rread@sun.com