scale up
scale out
scale simply

The Ultra-Scalable HPTC lustre Filesystem

R. Kent Koeninger

High Performance Technical Computing Division (HPTC)

ClusterWorld 2003
June 2003

Agenda

1. What is Lustre ?
2. Lustre, Hendrix, ASCI & HP
3. Organizations & Programs
4. Lustre – a little more detail
5. Hendrix Milestones
6. Summary

Lustre filesystem introduction

- Open Source filesystem funded by ASCI Path Forward (the U.S. DOE National Laboratories: Livermore, Los Alamos, Sandia)
- HP is the prime contractor working with Clustered Filesystems Inc. (CFS) and Intel to bring Lustre to market
- single sharable image
  - single name space
  - parallel-coherent access
- high bandwidth transports: GbETN, Quadrics, Myrinet …
- highly scalable
  - scalable bandwidth
    - multiple data servers
    - parallel filesystems
- scalable storage
    - petabytes of disks
- scalable metadata access

Lustre Capabilities Overview

- Lustre is a parallel-scalable-distributed filesystem designed to serve the most demanding high-performance-technical-computing (HPTC) environments
  - Sometimes called the “inter-galactic” filesystem for its extremely high scalability, performance, and availability goals
- Designed for very high scalability:
  - Thousands of compute client nodes
  - Petabytes of storage
  - Hundreds of gigabytes per second bandwidth
  - With full coherence and high reliability

Lustre Capabilities Overview

- Designed for full resiliency
  - No single points of failure
  - Journaling, fail-over, redundancy, etc.
- Designed to manage the storage independently of the client operating systems
  - As with NAS (CIFS or NFS), the compute clients need not know the details of managing the Lustre storage

Lustre Scalability targets: Phased Implementation

- Lustre Lite
  - 100 Clients, 10 OST, 1 MDS (no fail over)
- Lustre Performance
  - 700 Clients, 32 OST, 1+1 MDS
- Lustre Clustered MDS
  - 3000 Clients, 200 OST, 4 MDS
- Lustre T10
  - 3000 Clients, 1000 OST, 16 MDS
- Lustre GNS
  - 10000 Clients, 1000 OST, 100 MDS

Lustre is an Open Source Project funded by the DOE through the Tri-Labs

- HP is investing in Lustre technology
- Lustre is Open Source technology implemented on equipment from many vendors, including HP
  - The code is Open Source – GPL
- Luster projects are well funded by the Department of Energy (US DOE) through the Tri-National Labs (Livermore, Los Alamos and Sandia)
  - HP is the prime contractor for and is co-funding this Lustre ASCI-PathFoward project code-named “Hendrix”

Hendrix

- Open Source TriLabs program – HP, CFS & Intel
  - Funding is via ASCI PathForward grant with HP co-funding its development
  - Series of phased deliverables
- Members
  - CFS: Cluster Filesystem, Inc. technical design lead
    - All Lustre Lite development
  - HP Prime contractor with co-funding
    - Program management
    - Lustre Lite – test development and processes
    - LLP and beyond – part of core Lustre design
    - Active work in NSS & HPTCD HP Divisions
  - Intel focus is on networking & performance instrumentation

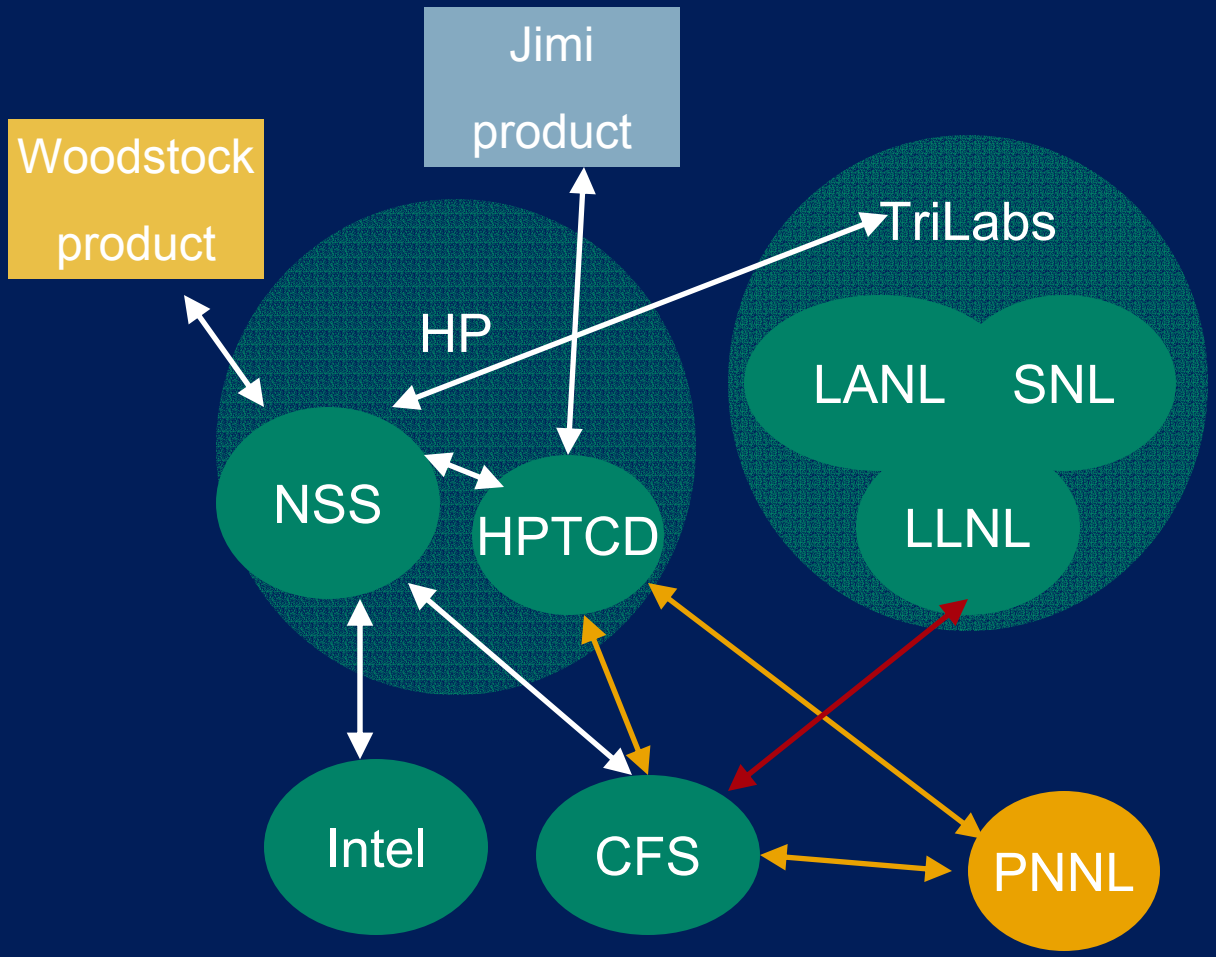
History of Lustre, Hendrix, ASCI & HP

- Original Pathforward RFP generated many proposals:
  - One from HP Labs
  - One from CFS Inc/Intel
- TriLabs requested a joint proposal from HP/Intel/CFS
- DOE money was committed in Fall 01
- HP NSS (storage) saw potential in proposal
  - Provided program management and technical staff
  - Negotiated contract & SOW
- Winston Prather, HPTCD, signs contract in June 02.
  - Joint development between agreed to by HPTCD and NSS.
- Program codenamed Hendrix

Programs & Organizations

- Hendrix – DOE ASCI Path Forward project
- Jimi – HP HPTCD project
  - Productization of Hendrix for HPTCD
- Woodstock – HP Storage project
  - Productization of Hendrix for NSS
- LLNL (Lawrence Livermore National Lab)
  - Large Lustre program on 32-bit cluster (non-HP)
- PNNL (Pacific Northwest National Lab)
  - Early Lustre support on 64-bit cluster

Lustre/Hendrix Eco System



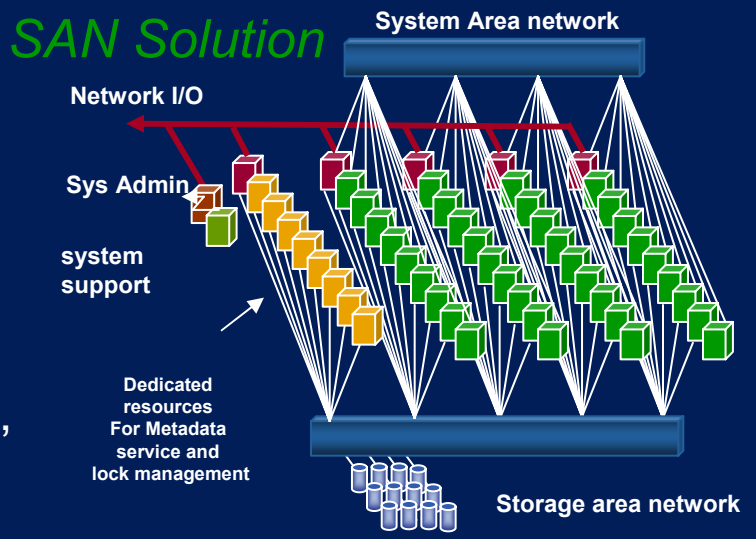
Hendrix

PNNL

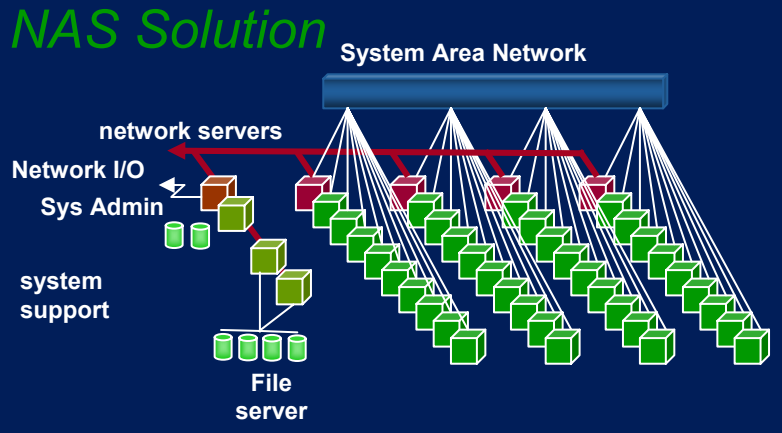
LLNL Lustre

Comparing SAN and NAS

- Storage Area Networks (SANs) provide high bandwidth, low latency connections from multiple hosts to storage
  - Fibre Channel: industry standard SAN
    Pro: High bandwidth with low latency
    Con: Expensive interconnect difficult to scale
    Con: Block level access: hard to share, manage, and secure



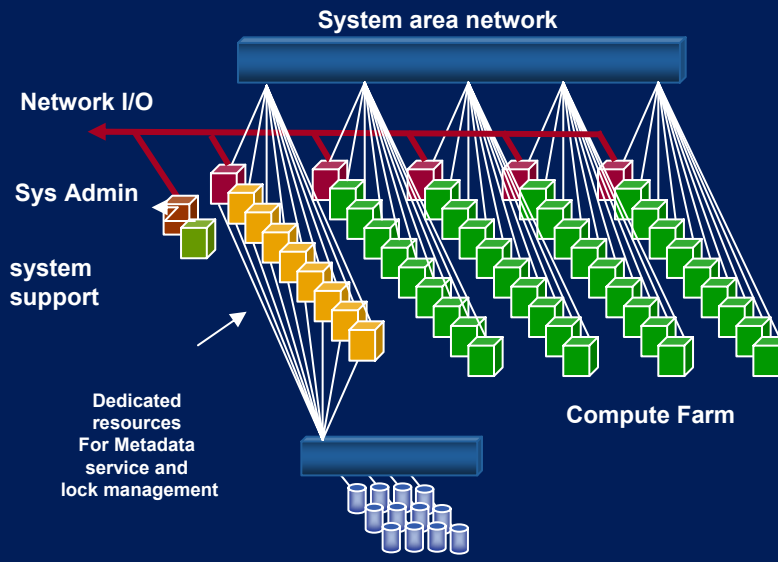
- Network Access Storage (NAS) provides independently managed storage that can be accessed by many clients
  - CIFS and NFS on Ethernet: industry standard NAS
    Pro: File level access: easy to share,manage, and secure
    Con: Low bandwidth, high protocol overhead



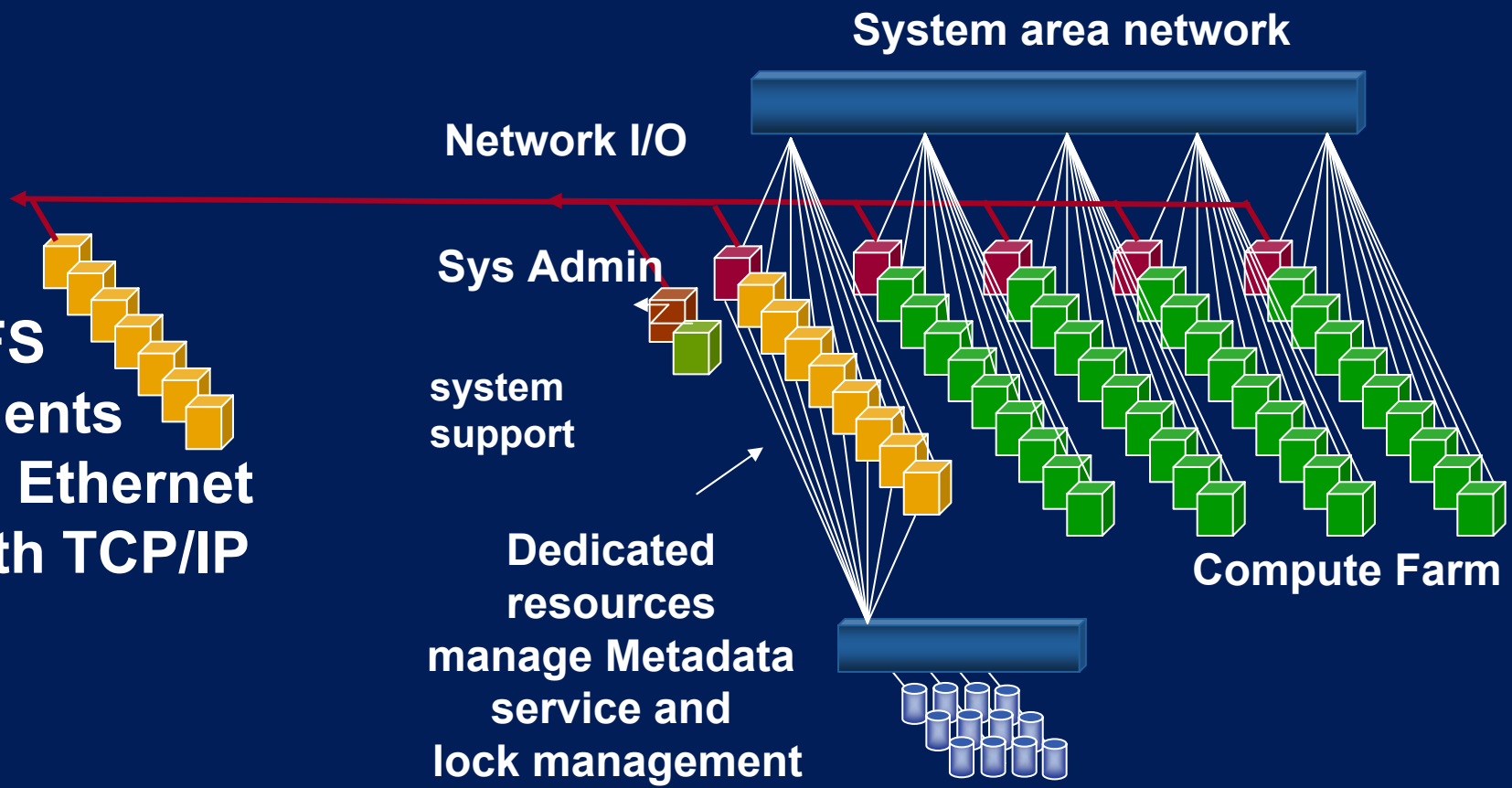
Lustre Combines the Best of SAN and NAS

- Shared data (as with NAS)
- High bandwidth, low overhead access (as with SAN)
- High scalability (even higher than NAS)
- Storage managed independently of client hosts (as with NAS)
- Highly resilient
- Designed to work with multiple interconnects
  - Can use existing message-passing interconnects
  - Gb Ethernet, 10 Gb Ethernet, Quadrics, Myrinet, …
  - Lower cost than connecting Fibre Channel to each of hundreds or thousands of compute clients

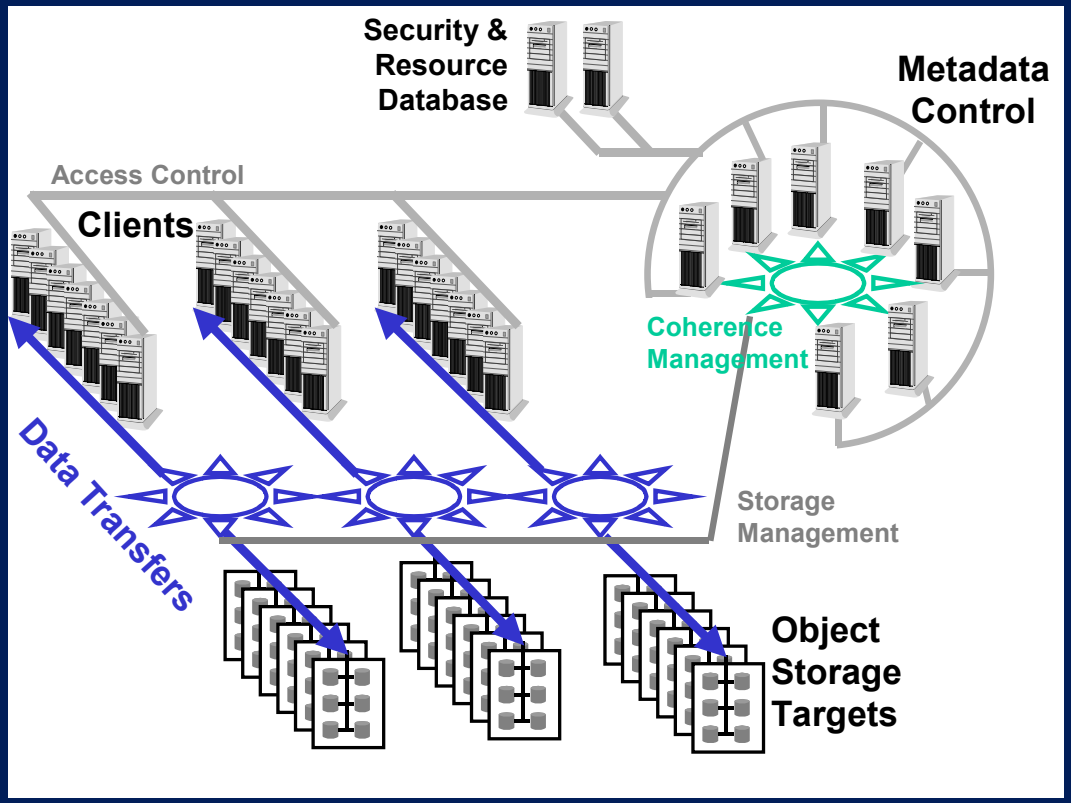
Lustre Solution



Lustre is also a Scalable NFS Server

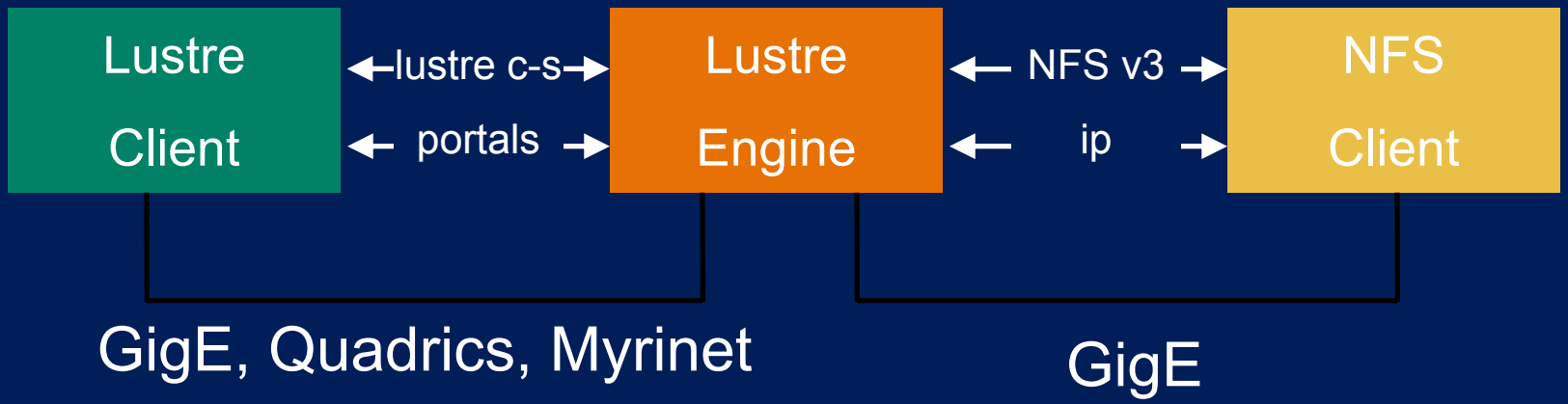


Lustre Topology

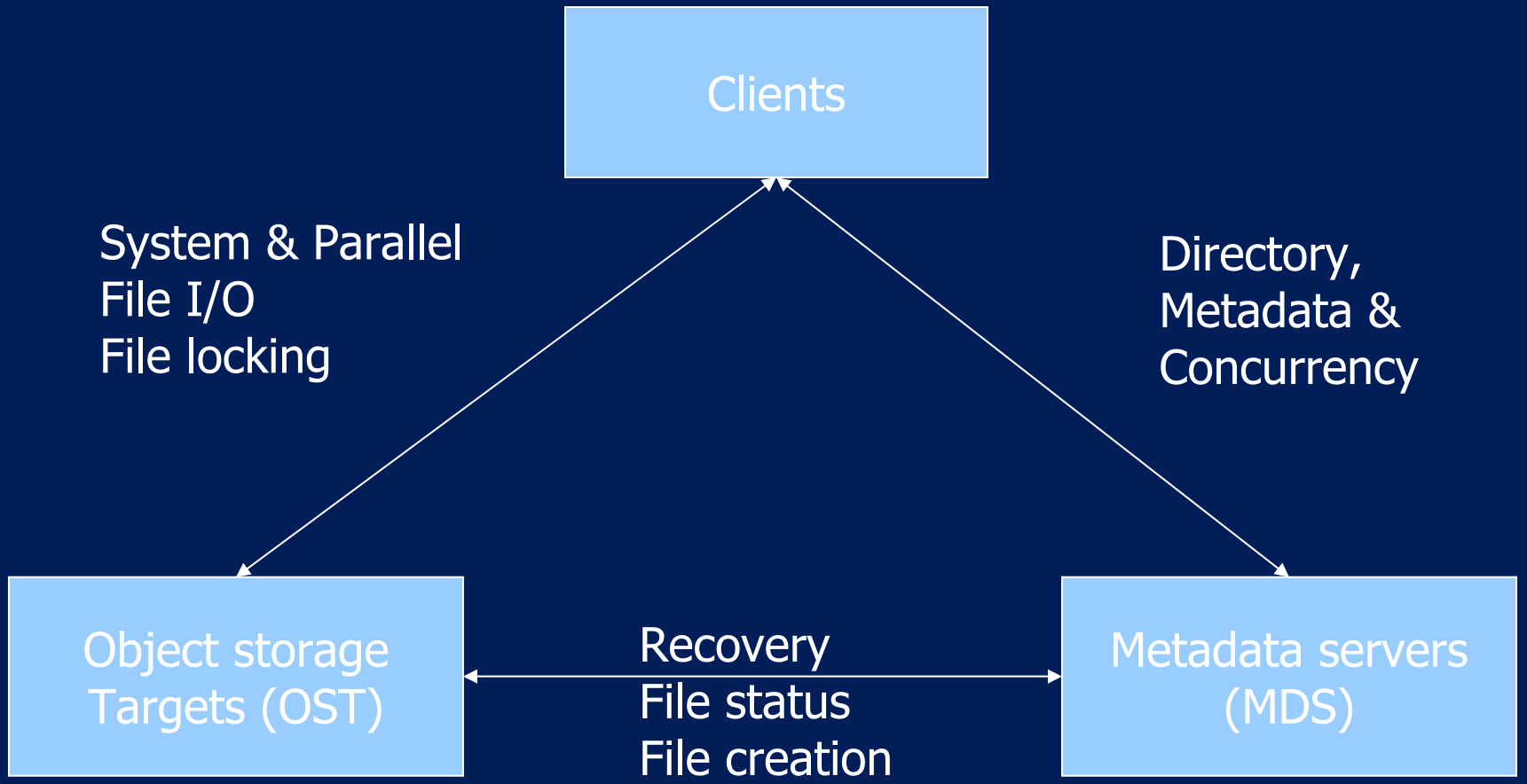


- Metadata controllers manage file system metadata
- Direct client I/O to object storage targets
  - RDMA
  - Storage fabric
- Metadata protocol
  - RDMA
  - Aggressive client caching
- Security defined in Resource DB, enforced by storage targets

Simplified View: Lustre and NFS Clients



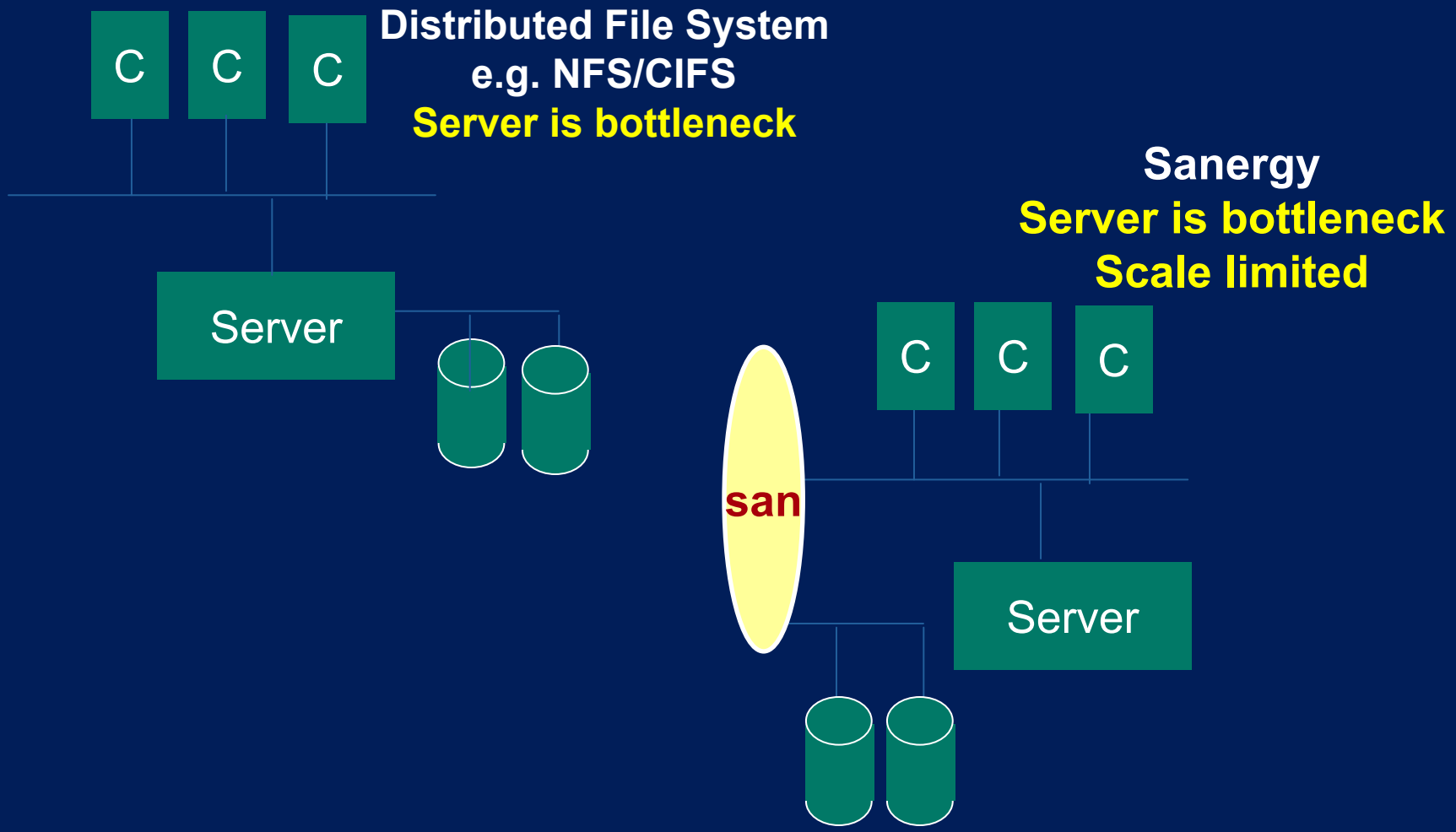
Lustre File System
logical structure



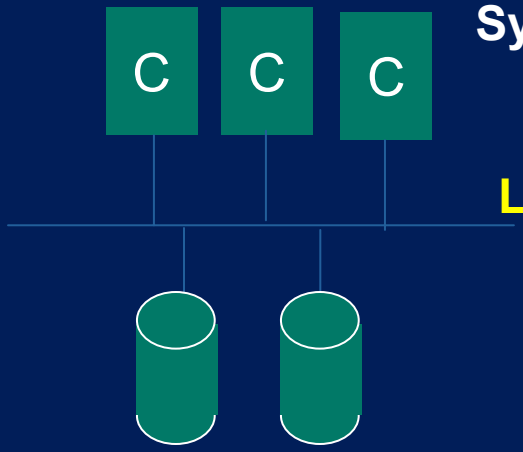
Lustre technology

- Key features
  - Scalable/parallel data serving
  - Metadata separation
  - Metadata scaling
- Comparative technologies
  - NFS
  - SANergy
  - TruCluster CFS
  - Petal/Frangiapani
  - PFS
  - GPFS

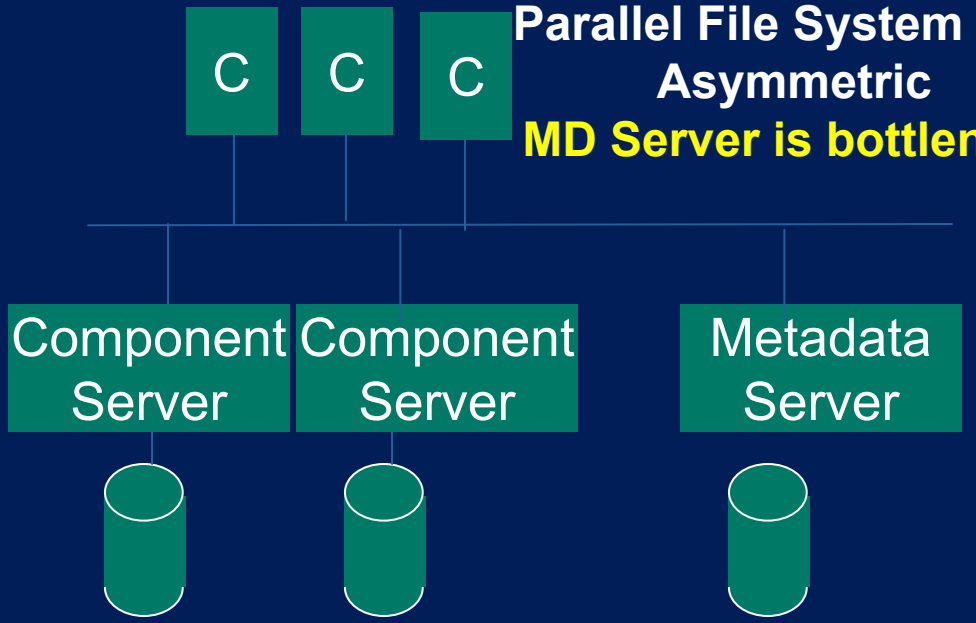
Comparative solutions



Comparative solutions

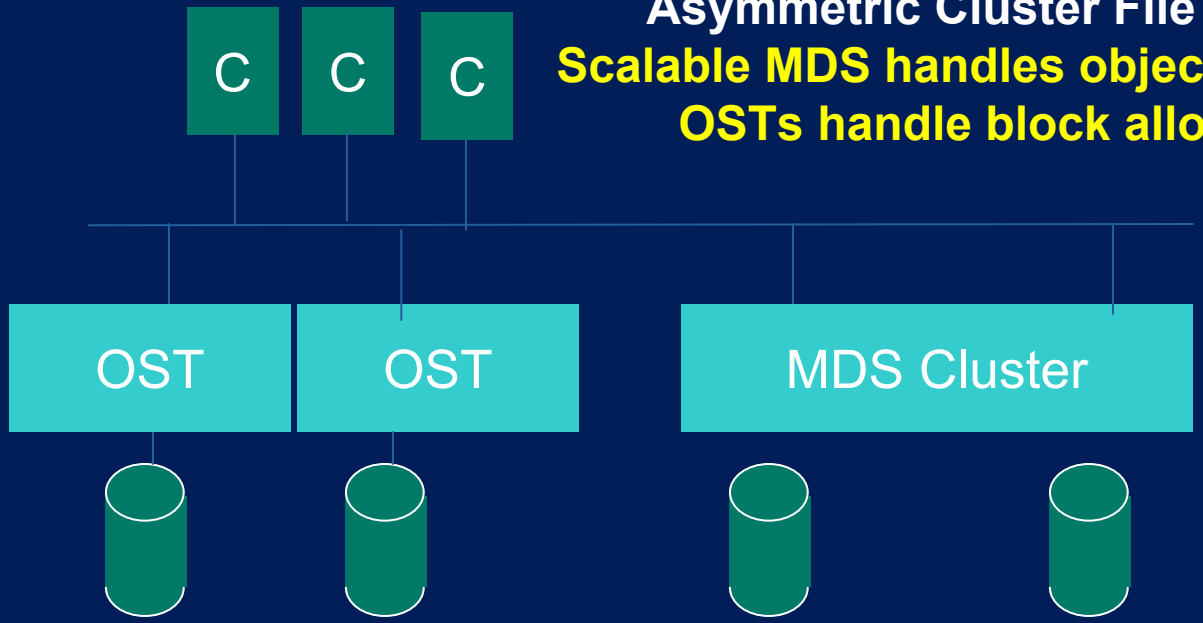


Symmetric Clustered File System
Sistina/GPFS
Lock management
Is bottleneck



Parallel File System (SC)
Asymmetric
MD Server is bottleneck

Lustre Solution



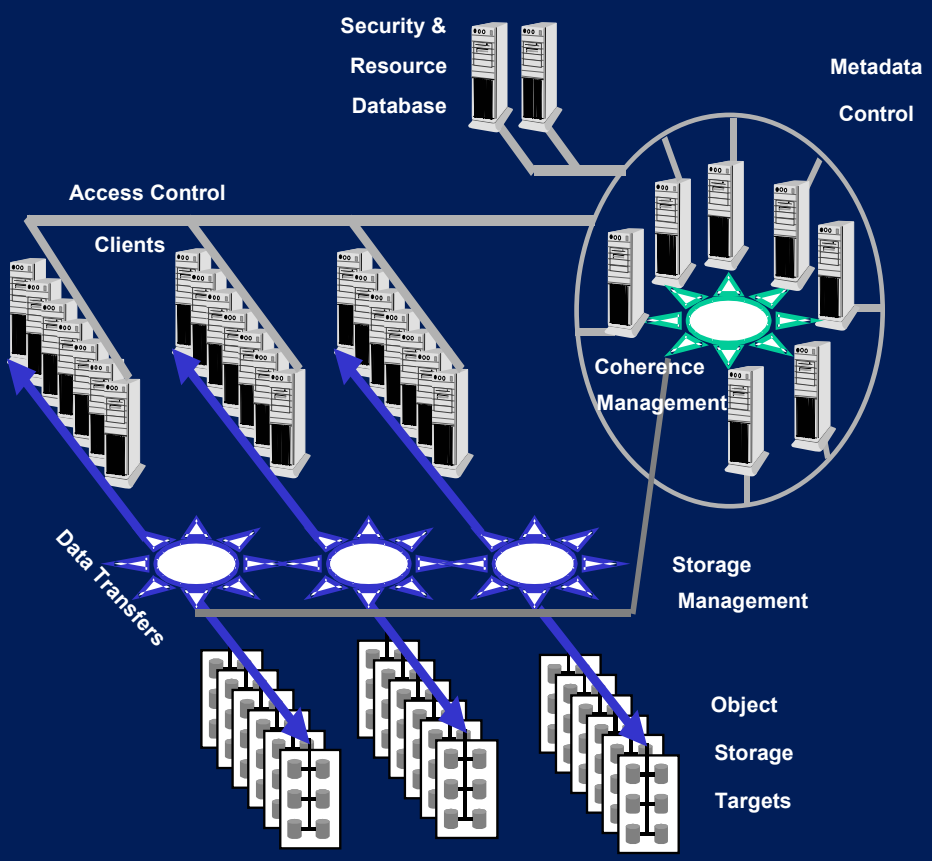
Asymmetric Cluster File System
Scalable MDS handles object allocation, OSTs handle block allocation

Global File System evolution

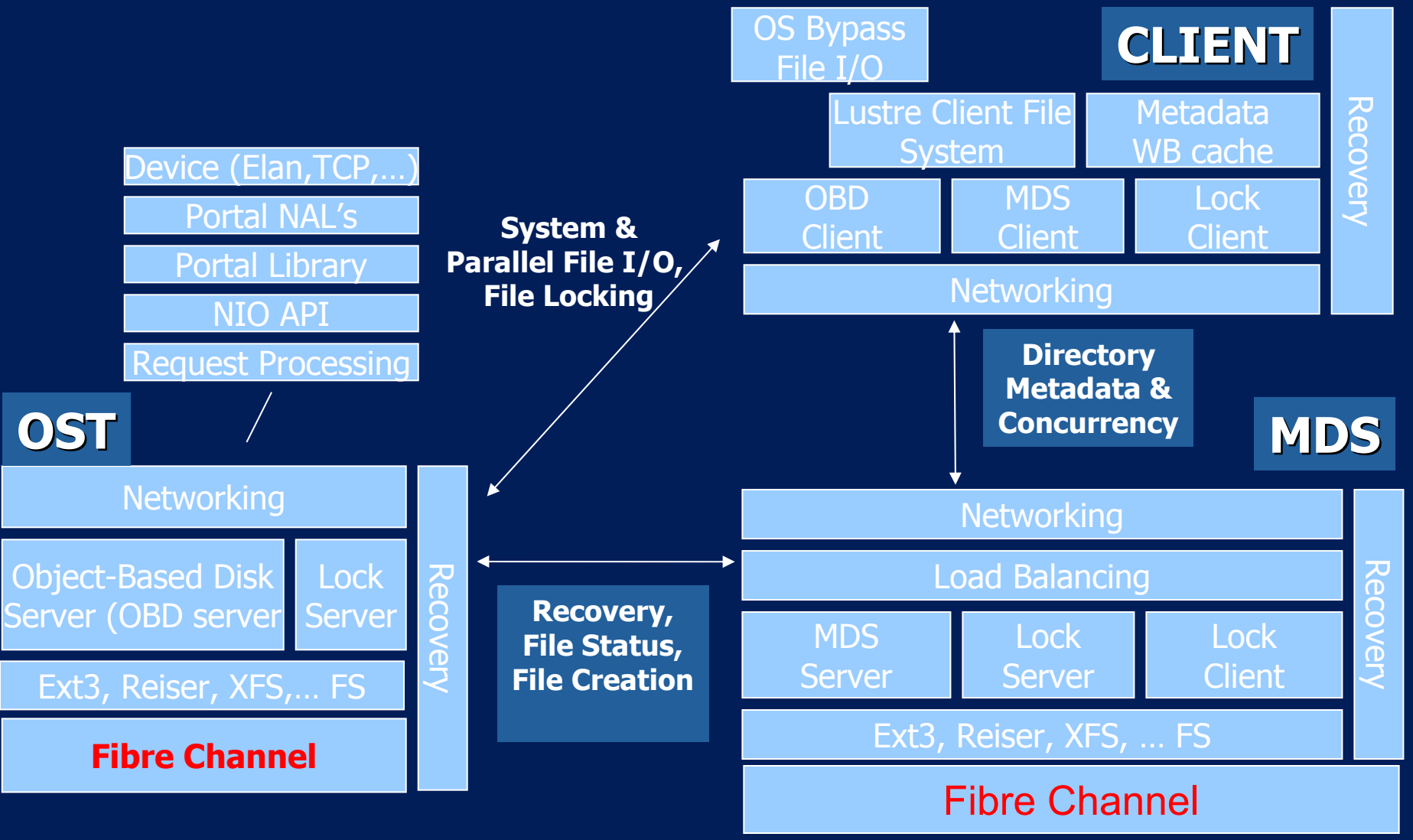
- Per SMP file system – not sharable
- NFS has been traditional solution to sharing
  - Workable with shortcomings
  - Not scalable
- Slow evolution
- Equally a problem for Enterprise & HPTC
  - HPTC more demanding
- Initial solutions in Enterprise space
  - E.g.: SANergy, TruCFS
  - Limited value to HPTC
- GPFS in HPTC
- All are deltas to classic file system design

Lustre is a new approach using best practices in a new ultra-scalable design

- Not encumbered by existing architecture
- Scalability at inception
  - Separation of Metadata & file data
  - Scalable Metadata
  - Scalable file data
    - Block management at OST level
  - Efficient locking
- Object architecture



Lustre File system protocol view



Lustre/Hendrix (phased development)

2003

Phase 1 | Phase 2 | Phase 3 | Phase 4 | Phase 5

Lustre Lite

Key Features

- Functional file system
- Parallel I/O, Distributed benchmarks can be run.
- 100 clients, 10 OSTs, 1 Meta-Data server
- Performance at 20% theoretical (single node) 10% theoretical (parallel nodes)
- Recover from system failure
- Multiple clients, single MDS
- MDS serving more than 1 file system
- Client node failure recovery automatic.
- Metadata journaling.

Lustre Lite Performance

Key Features:

- Improve performance to 80% theoretical (single node) 40% theoretical (parallel nodes)
- 700 clients, 30 OSTs, 1+1 Meta-Data server
- Scalability is much improved
- Locking system improved.
- Exporting NFS
- MPI-IO support.
- OST fail over

Clustered MDS

Key Features:

- First implementation of scalable clustered metadata servers
- 3000 clients, 200 OSTs, 4 node MetaData server
- Improve Parallel I/O performance @ 60% theoretical
- (single node still @ 80%)
- Add global namespace support

Lustre T10 & Security

Key Features:

- Add T10/OSD compatibility and Security pieces.
- Improve Parallel I/O performance @ 70% theoretical
- (single node still @ 80%)
- 3000 clients, 1000 OSTs, 16 node Meta-Data server

Lustre GNS & Management

Key Features:

- Improve Parallel I/O performance @ 85% theoretical
- Single node performance to 95% theoretical
- 10,000 clients, 10,000 OSTs, 100 node Meta-Data server
- Integration with enterprise management tools
- Hooks for HSM application

Lustre Lite

Key Features:

- Functional file system
- Parallel I/O, Distributed benchmarks can be run.
- 100 clients, 10 OSTs, 1 Meta-Data server
- Performance at 20% theoretical (single node) 10% theoretical (parallel nodes)
- Recover from system failure
- Multiple clients, single MDS
- MDS serving more than 1 file system
- Client node failure recovery automatic.
- Metadata journaling.

Lustre Lite Performance

- **Key Features:**
  - Improve performance to 80% theoretical (single node) 40% theoretical (parallel nodes)
  - 700 clients, 30 OSTs, 1+1 Meta-Data server
  - Scalability is much improved
  - Locking system improved.
  - Exporting NFS
  - MPI-IO support.
  - OST fail over

Lustre Lite Performance
(w/Clustered Meta Data Servers)

Key Features:

- First implementation of scalable clustered metadata servers
- 3000 clients, 200 OSTs, 4 node MetaData server
- Improve Parallel I/O performance @ 60% theoretical
- (single node still @ 80%)
- Add global namespace support

Lustre T10 & Security

Key Features:

- Add T10/OSD compatibility and Security pieces.
- Improve Parallel I/O performance @ 70% theoretical
- (single node still @ 80%)
- 3000 clients, 1000 OSTs, 16 node Meta-Data server

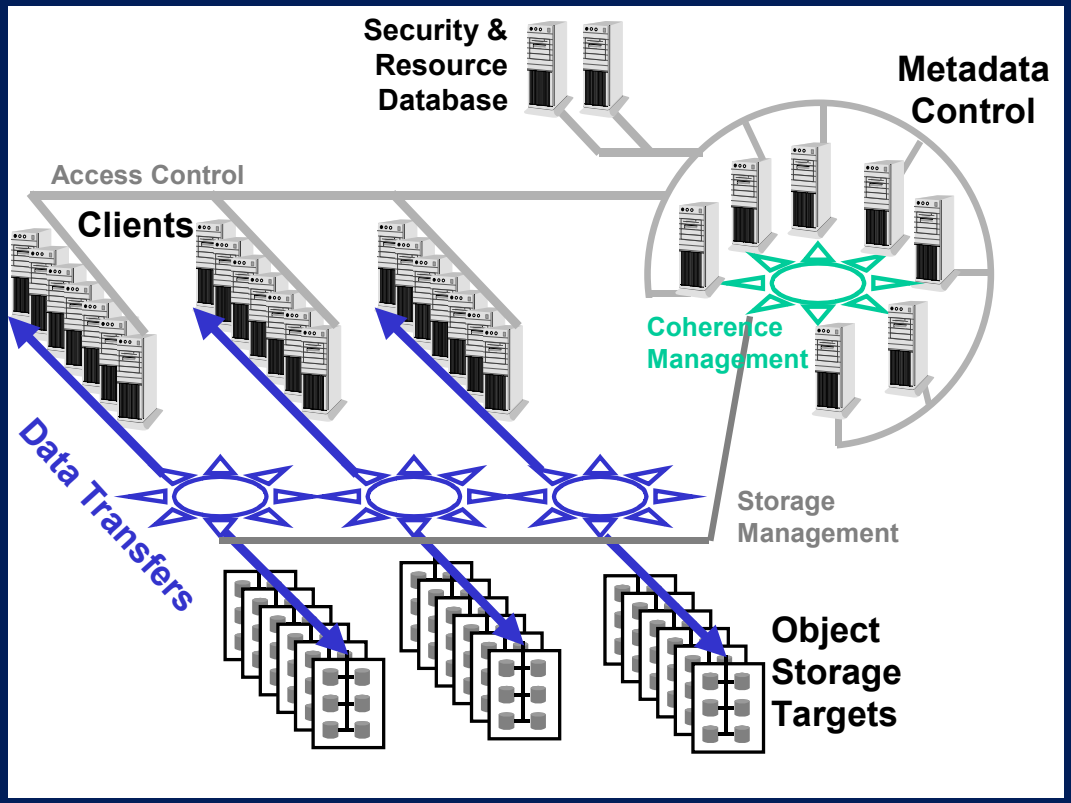
Lustre GNS & Management
(final product!)

- Key Features:
  - Improve Parallel I/O performance @ 85% theoretical
  - Single node performance to 95% theoretical
  - 10,000 clients, 10,000 OSTs, 100 node Meta-Data server
  - Integration with enterprise management tools
  - Hook for HSM application

Recap Lustre and HP

- Next generation global file system
- Open Source
- TriLabs support
- Hendrix program
  - HP, CFS & Intel
  - HP in position to influence design
  - HP knowledge advantage and differentiation opportunity
- HP's goal is to lead with Lustre products
  - HPTC clusters and NSS storage servers
- HP Early field experience at PNNL
- Extensive HP customer interest

Lustre Summary



- The future for scalable, high bandwidth, parallel, high capacity, resilient, filesystems
- Open Source Backed by HP

For More Information

- www.lustre.org
  - The main source for Lustre news, overviews, and technical information
- HP HPTC Web site
  - www.hp.com/techservers
  - All about HP scalable-technical solutions, including Linux clusters and Lustre

Lustre

Hendrix, Jimi, and Woodstock