

# The Ultra-Scalable **Lustre** HPTC Filesystem

**R. Kent Koeninger**  
High Performance  
Technical Computing Division  
(HPTC) at HP



# Agenda

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

# What is Lustre?

## ■ A distributed-scalable filesystem for HPTC applications



- High bandwidth: 10s of GB/s of bandwidth (or more)
  - Scalable-parallel file and filesystem access
  - Designed to serve thousands and tens-of-thousands of clients
  - (How many parallel connections/severs can afford?)
- Huge storage: terabytes, petabytes and exabytes
  - (How much storage can afford?)
- Highly responsive
  - (How many metadata servers can you afford?)
- Highly reliable
  - No single points of failure
- Fully sharable data
  - Accessible by Linux nodes with Lustre-specific clients or any node with an NFS client

# 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

The logo for the Lustre filesystem. The word "lustre" is written in a lowercase, sans-serif font. The letters are colored: 'l' is green, 'u' is light green, 's' is light blue, 't' is medium blue, 'r' is light blue, and 'e' is light blue. A horizontal line with black dots passes through the center of each letter, resembling a string of beads or a data bus.

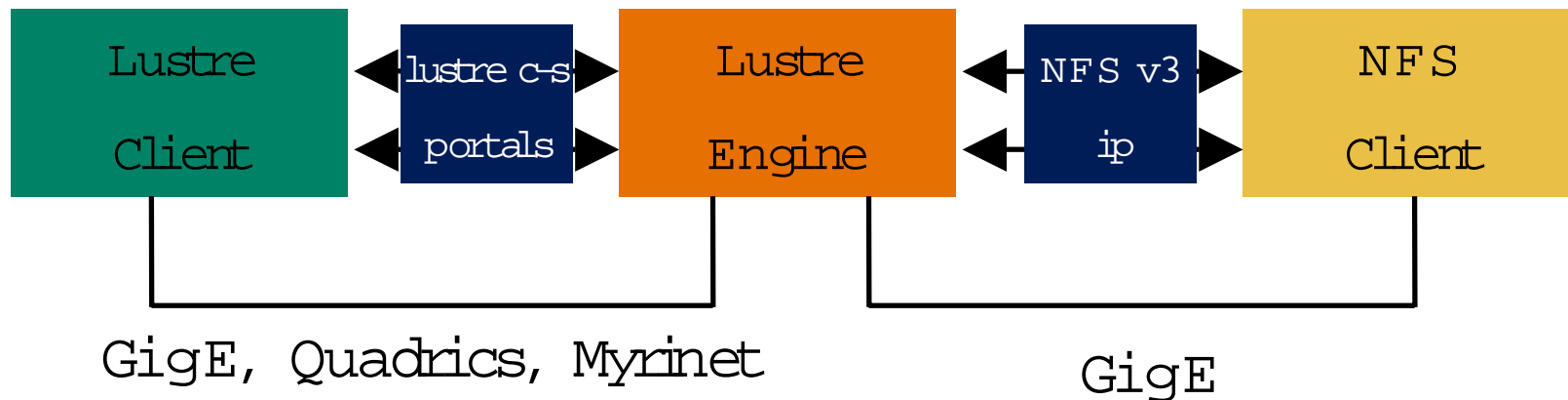
# Lustre Capabilities Overview

- 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

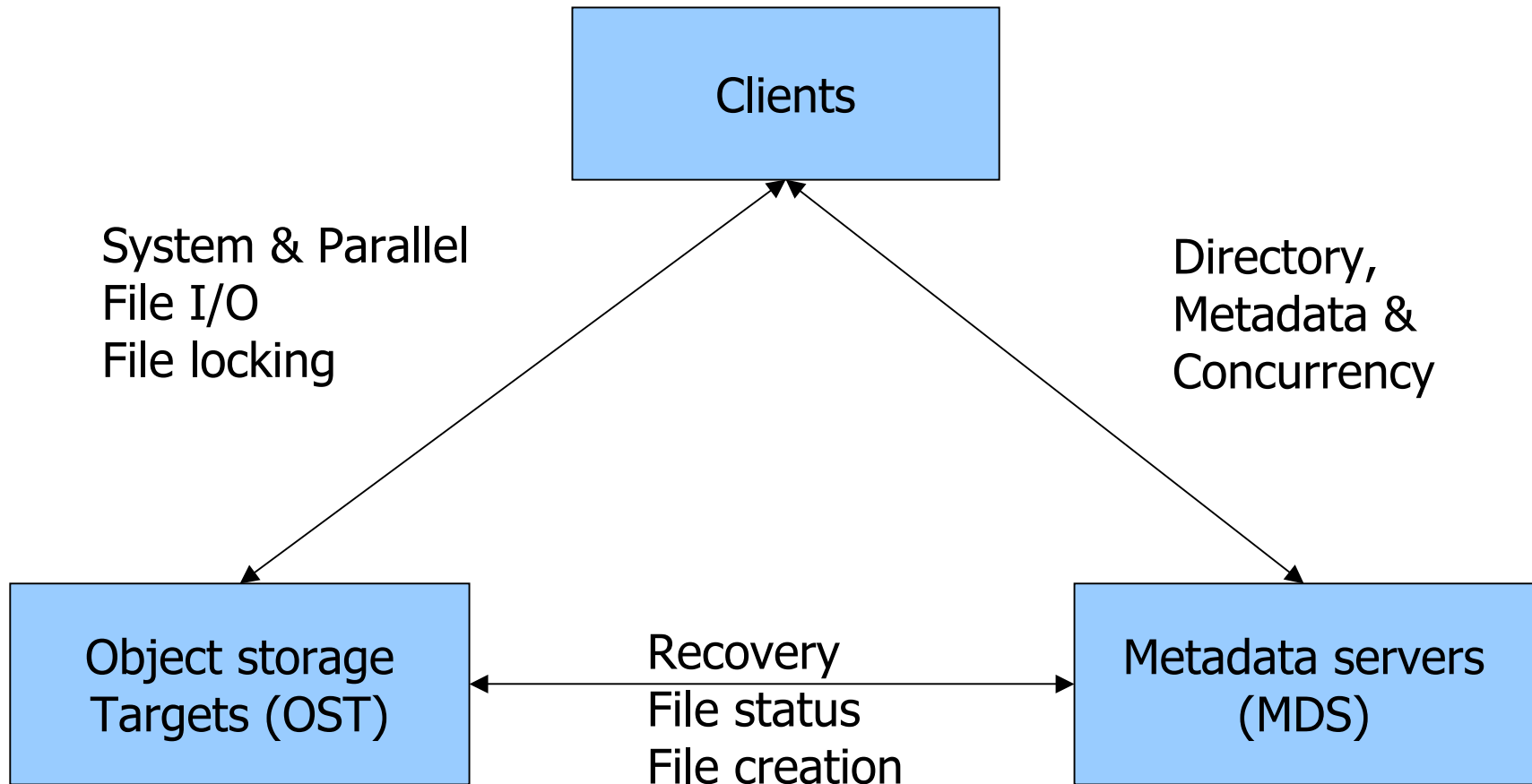
- Designed for full resiliency
  - No single points of failure
  - Journaling, fail-over, redundancy, etc.
- Designed to manage the storage independently of the client operating systems
  - As with NAS (CIFS or NFS), the compute clients need not know the details of managing the Lustre storage

# Simplified View: Lustre and NFS Clients



# Lustre File System

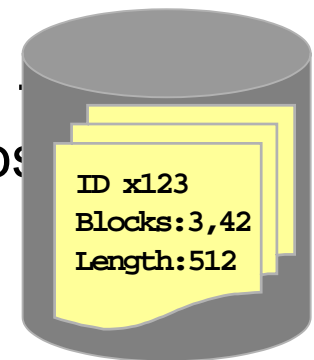
## logical structure





# Lustre - Object Storage Model

- An OST is more intelligent than a block device
- An object is a logical unit of storage
  - Used to store a “piece” of a user file
  - Lives in flat name space with an ID
  - Contains data *and* metadata (similar to an inode)
  - File-like methods: open, close, read, write
- An OST stores objects and could be any of
  - Disk drive, storage appliance, storage controller, ...
- OSTs enable high performance, secure and cross platform access
  - ✓ Uses a higher level abstraction
  - ✓ Offload read & write from the storage server
  - ✓ Can enforce secure client access



# 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
  
- Lustre 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-PathForward 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

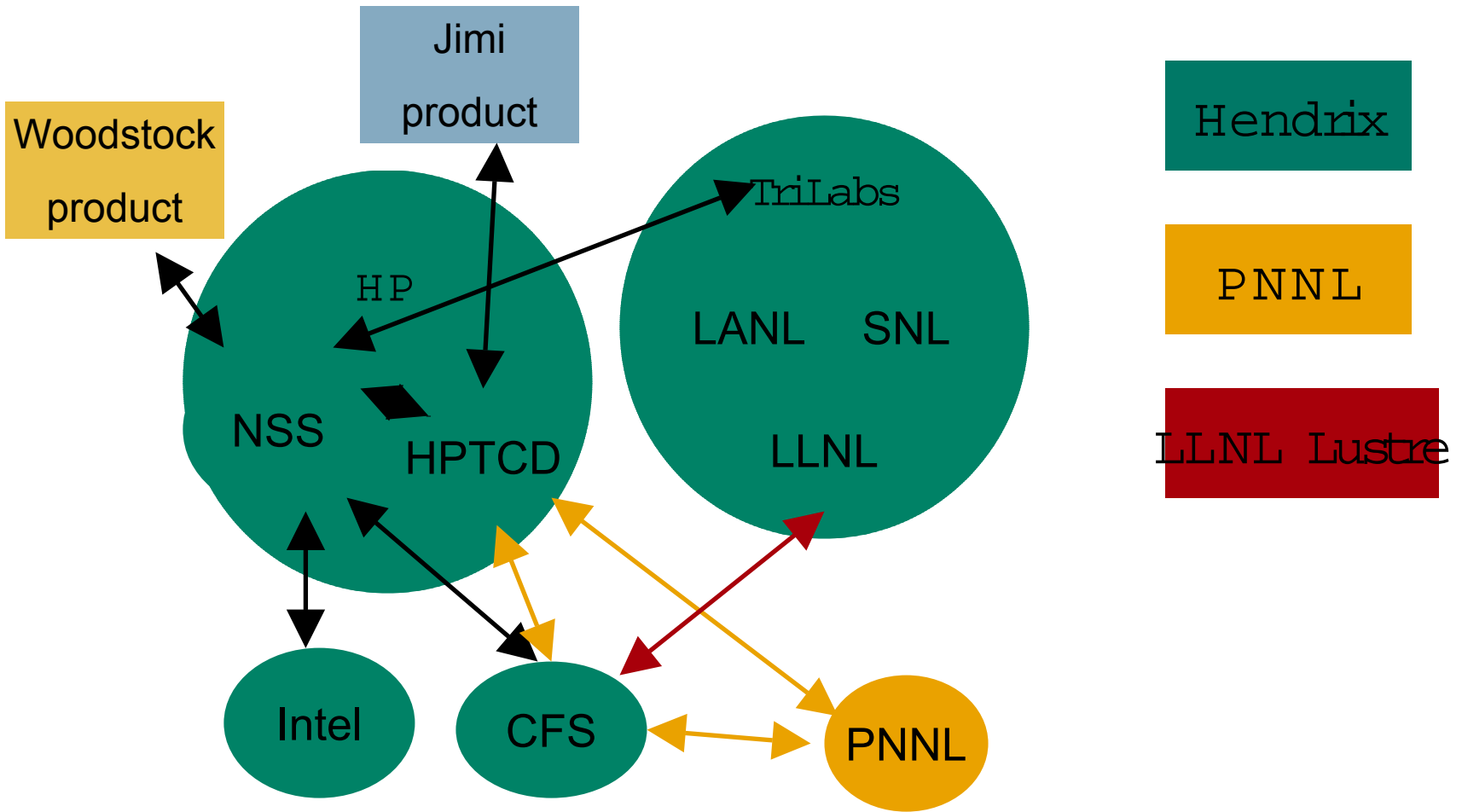


- Original Pathforward RFP generated many proposals:
  - One from HP Labs
  - One from CFS Inc/Intel
- TriLabs requested a joint proposal from HP/Intel/CFS
- DOE money was committed in Fall 01
- HP NSS (storage) saw potential in proposal
  - Provided program management and technical staff
  - Negotiated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

# Programs & Organizations

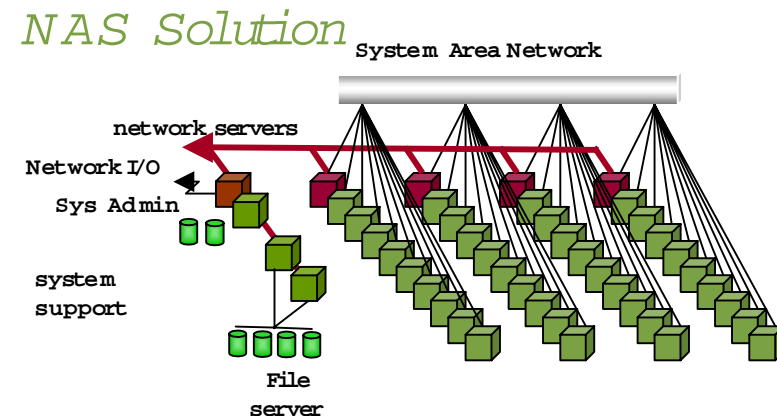
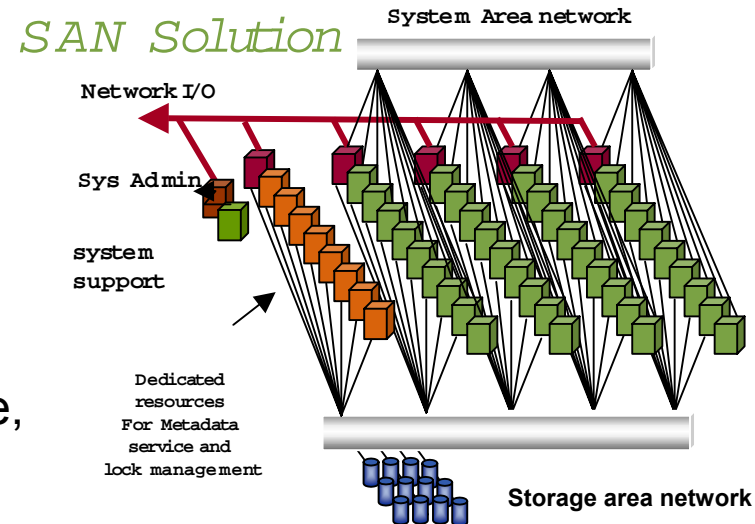
- Hendrix – DOE ASCI Path Forward project
- Jimi – HP HPTCD project
  - Productization of Hendrix for HPTCD
- Woodstock – HP Storage project
  - Productization of Hendrix for NSS
- LLNL (Lawrence Livermore National Lab)
  - Large Lustre program on 32-bit cluster (non-HP)
- PNNL (Pacific Northwest National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

# Lustre/Hendrix Eco System



# 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

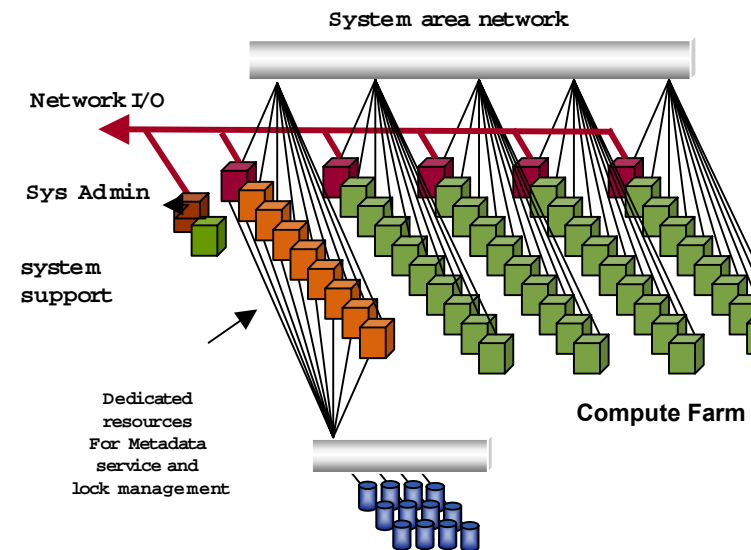




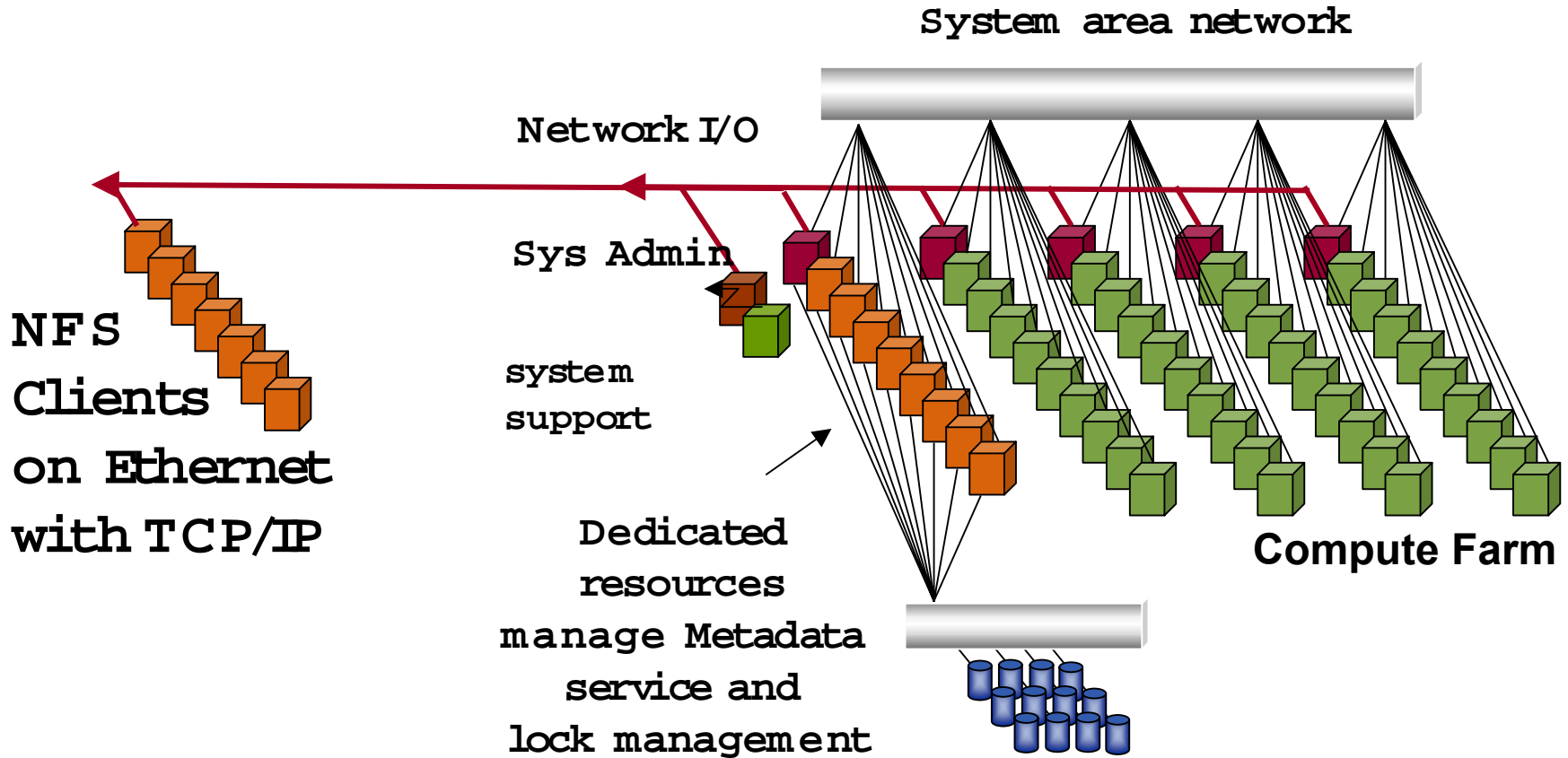
# Lustre Combines the Best of SAN and NAS

- 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

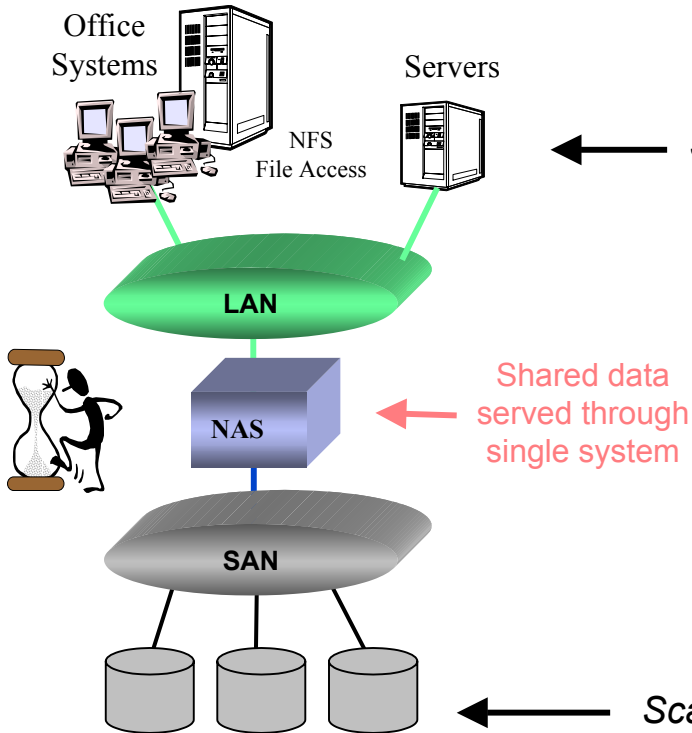


# Lustre is also a Scalable NFS Server



# Scalable NAS File Sharing

## The NAS Scaling Problem

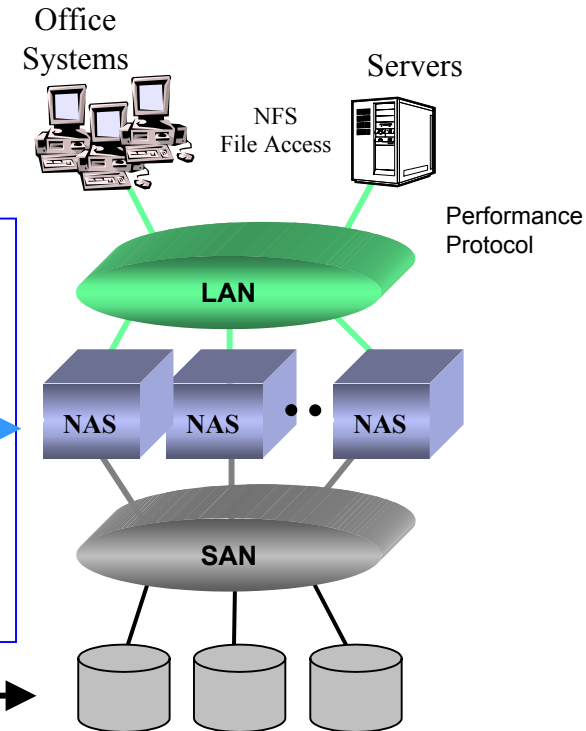


## Scalable NAS Solution

← Scalable number of client Systems sharing files →

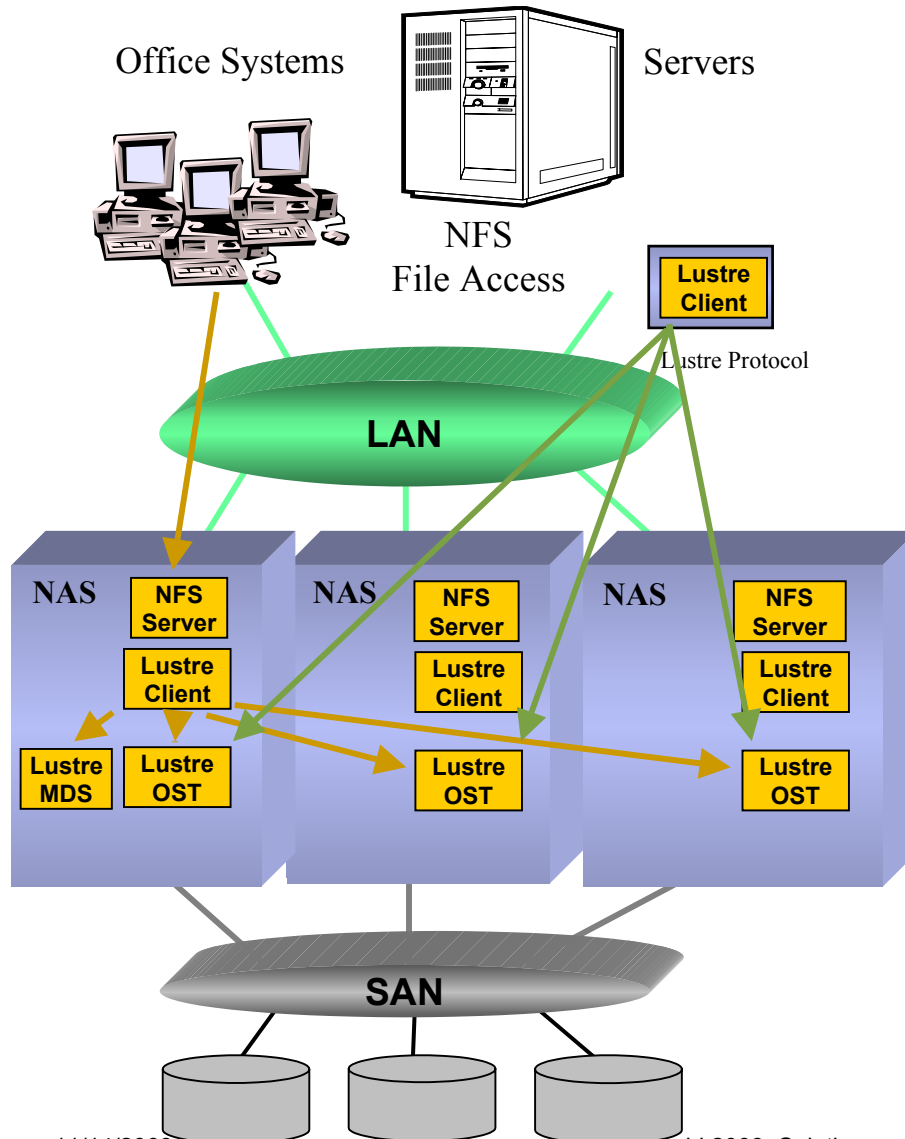
### Scalable NAS cluster

- Single System Image
- Scalable Performance - aggregate & single client
- High availability
- Managed as a single entity
- Enables commodity hardware products



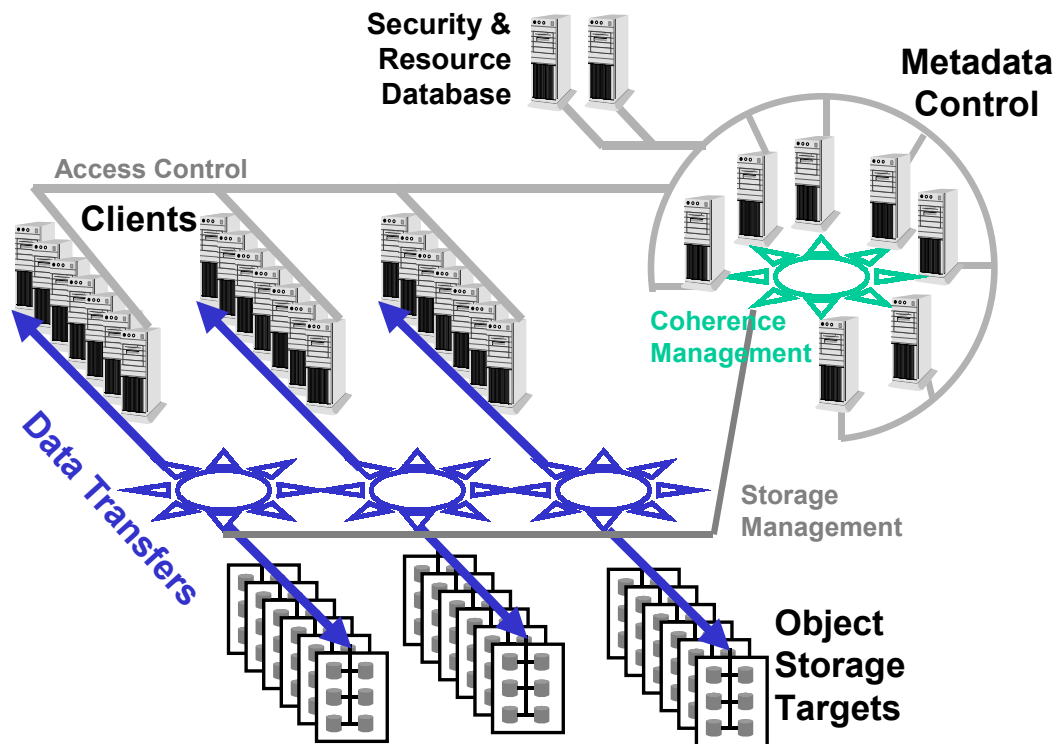
← Scalable NFS storage capacity using HP HPTC Products →

# Scalable NAS File Sharing



- Included in HP HPTC-Lustre Servers
- Multi-protocol NAS cluster supporting:
  - NFS/CIFS
  - Lustre protocol
- SSI file access across servers for NFS/CIFS
- Parallel access across servers for Lustre client access
- High availability - All modules fail-over among servers  
Browser based management GUI
  - Manage whole cluster from single point (including HA config)

# 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

# Lustre technology

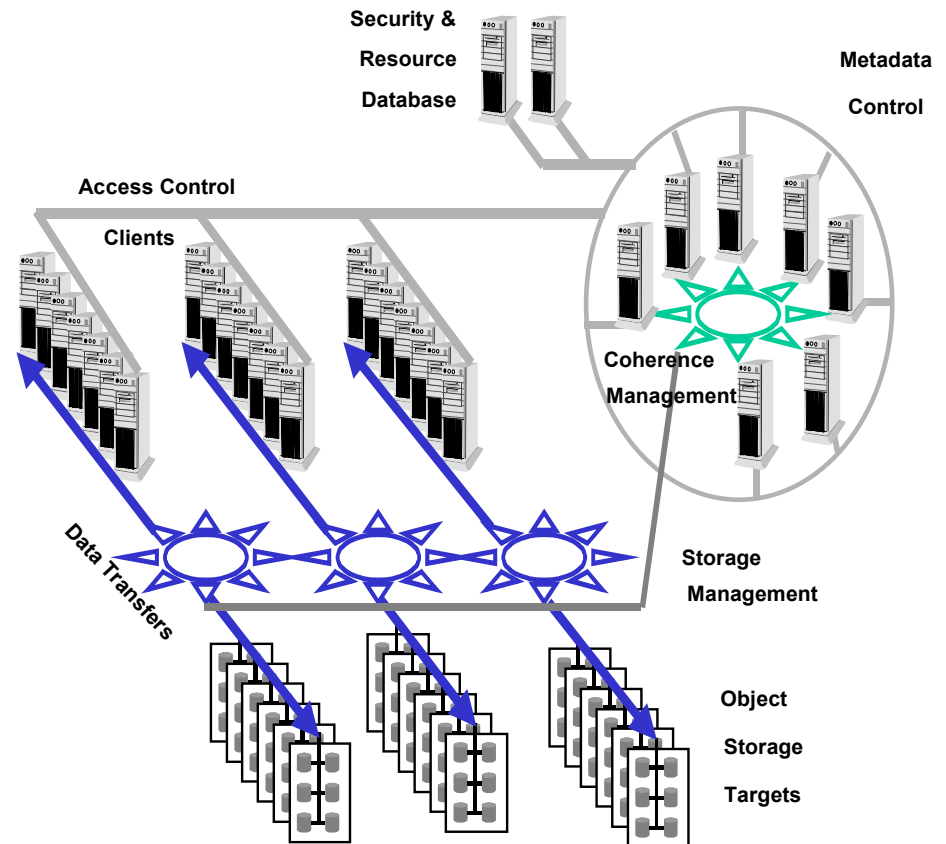
- Key features
  - Scalable/parallel data serving
  - Metadata separation
  - Metadata scaling
- Comparative technologies
  - NFS
  - SANergy
  - TruCluster CFS
  - Petal/Frangiapani
  - PFS
  - GPFS

# 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

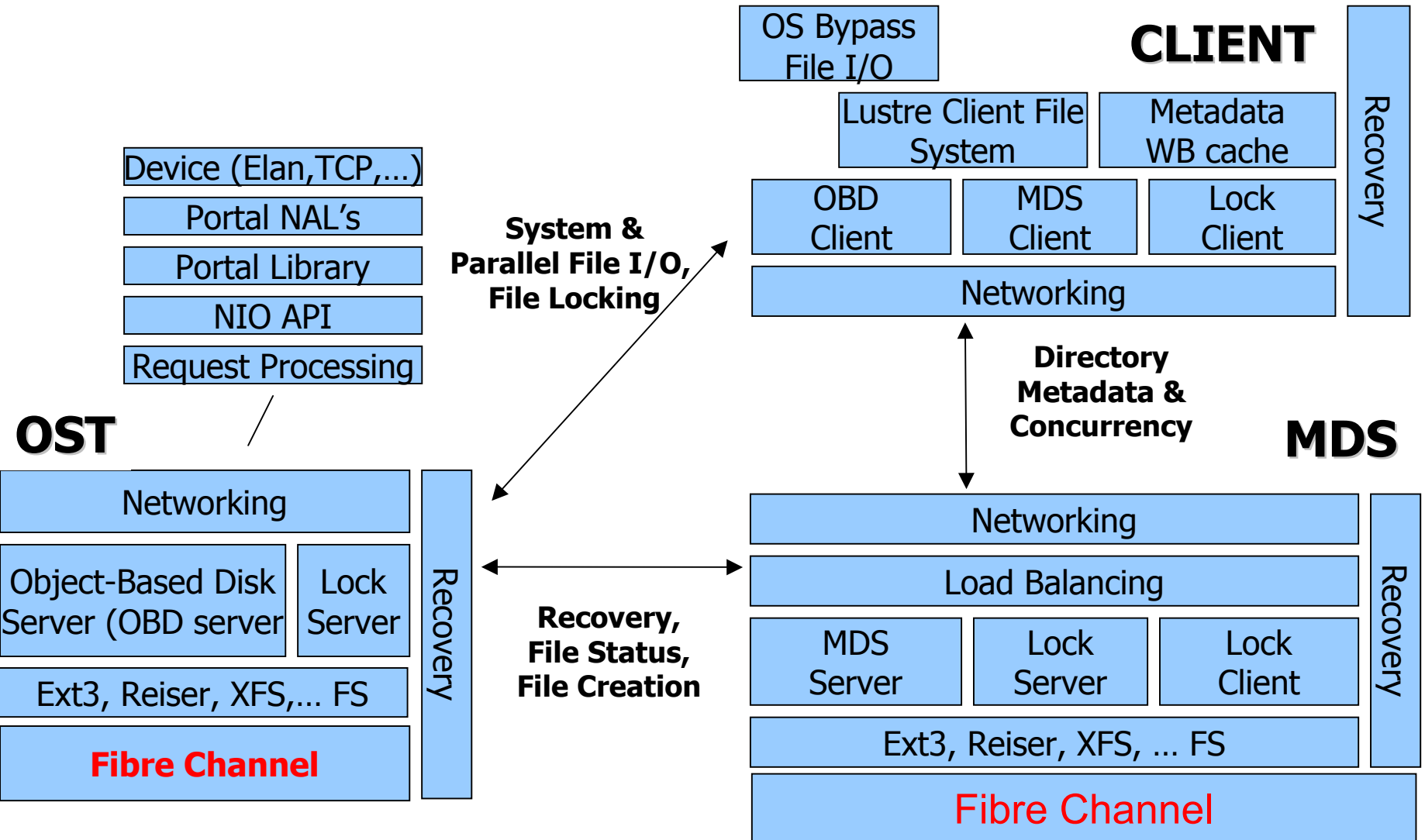
- Not encumbered by existing architecture
- Scalability at inception
  - Separation of Metadata & file data
  - Scalable Metadata
  - Scalable file data
    - Block management at OST level
  - Efficient locking
- Object architecture



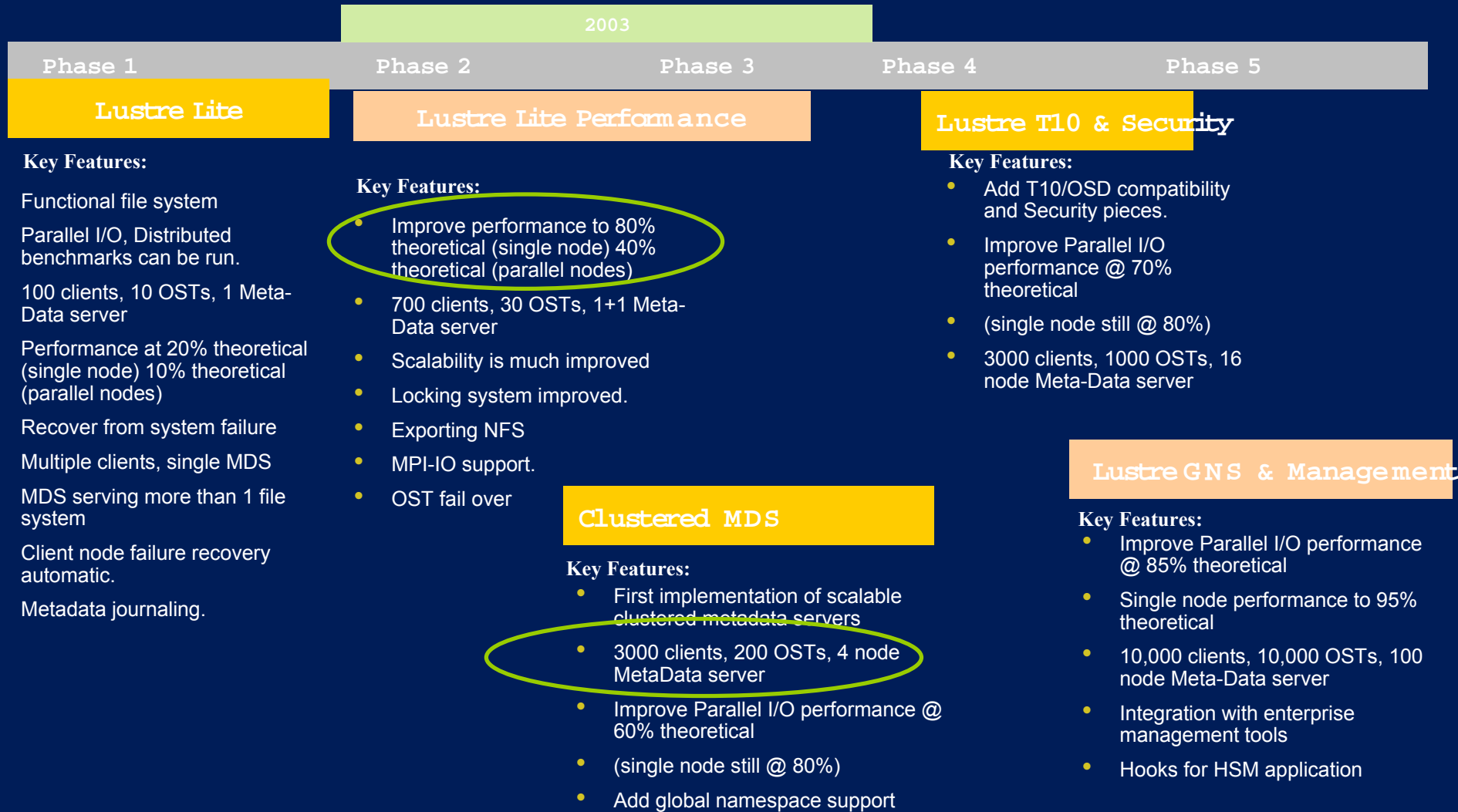


# Lustre File system

## protocol view



# Lustre/Hendrix (phased development)



# 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

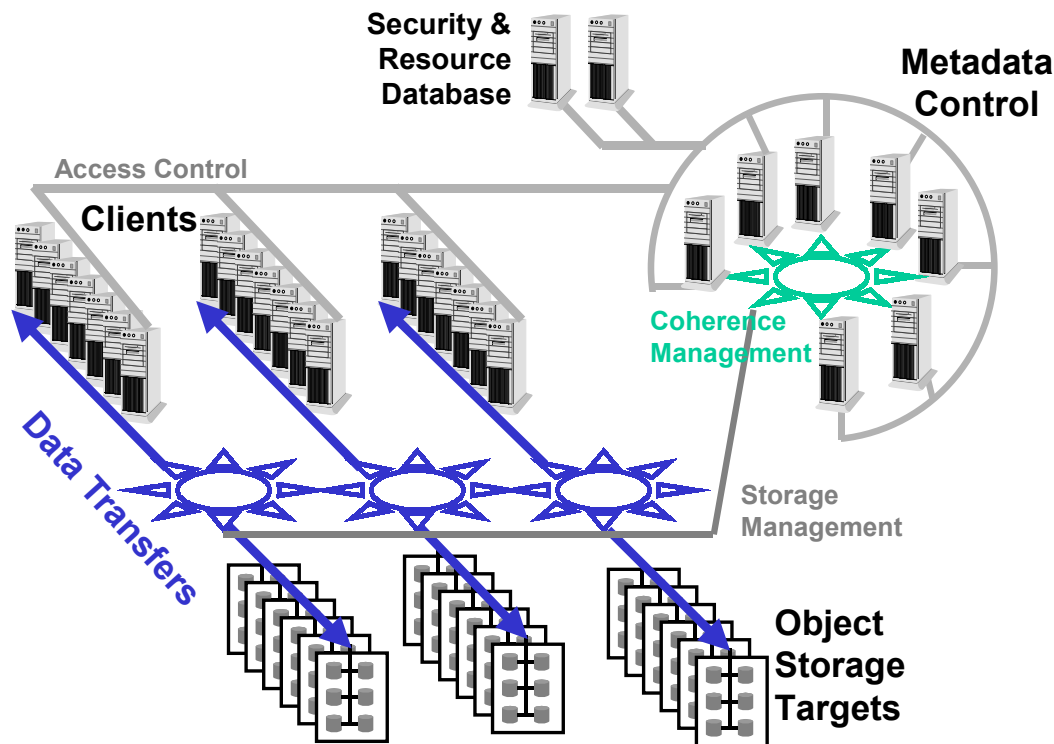
# Recap Lustre and HP

- 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

lustre



- The future for scalable, high bandwidth, parallel, high capacity, resilient, filesystems

- Open Source Backed by HP

# For More Information

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

Lustre



i n v e n t

Hendrix, Jimi, and Woodstock



# HP WORLD 2003

Solutions and Technology Conference & Expo

Interex, Encompass and HP bring you a powerful new HP World.

