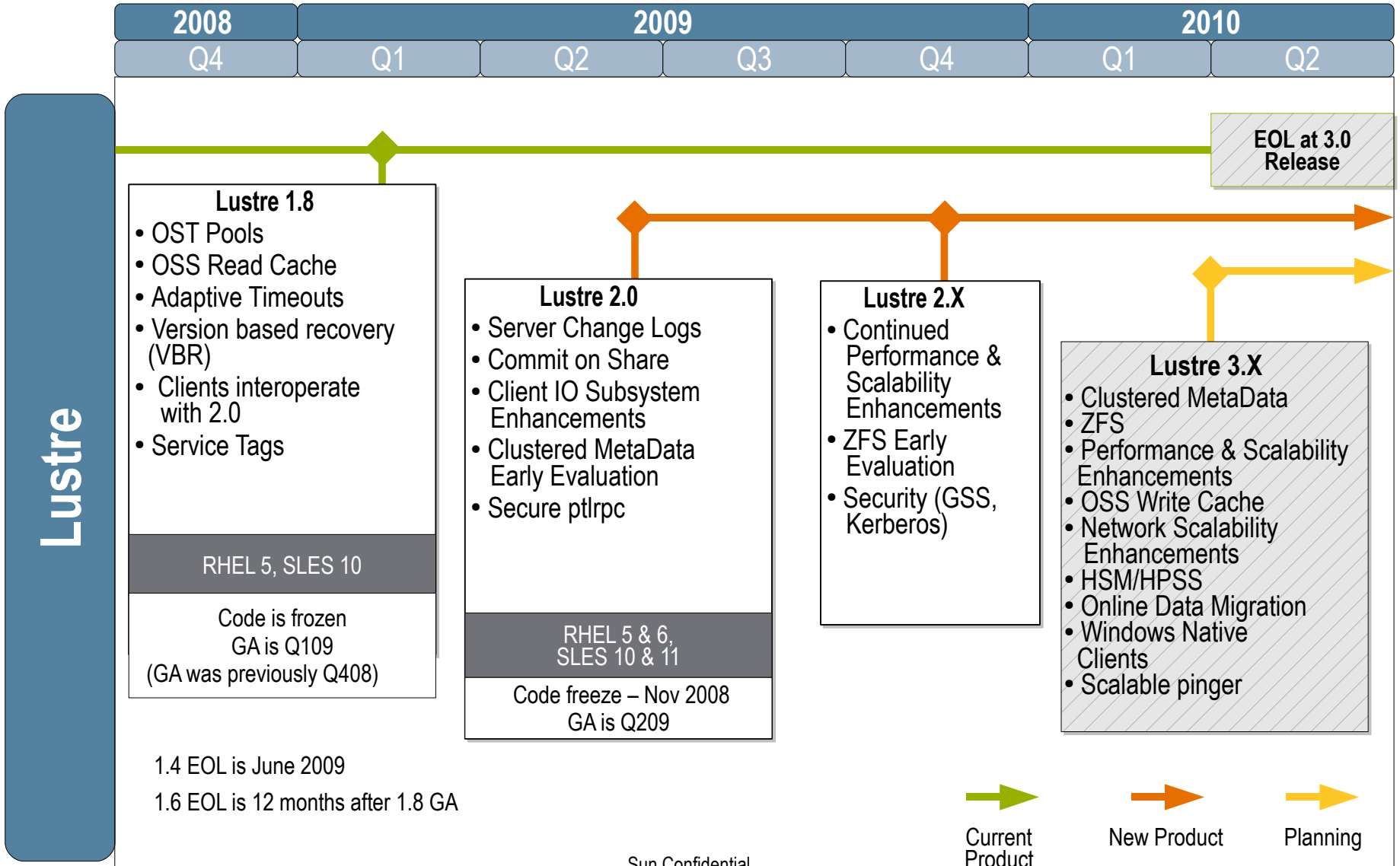
LUSTRE MULTI- PETAFLOP ROADMAP

February, 2009

Eric Barton
Lead Engineer, Lustre Group
Sun Microsystems, Inc.

Lustre Feature Releases



Lustre Projects

Active Development

(Features planned to be released, but not scheduled for a specific release yet)

- HPCS (see next page)
- Network Request Scheduler
- pNFS Export
- Platform Portability Enhancement

Research

(May or may not be scheduled for release)

- Flash Cache
- Dynamic LNET Configuration
- IPv6
- Local Stripe 0-copy IO
- Proxy Servers and IO Forwarding
- Backup Solution

DARPA HPCS Project

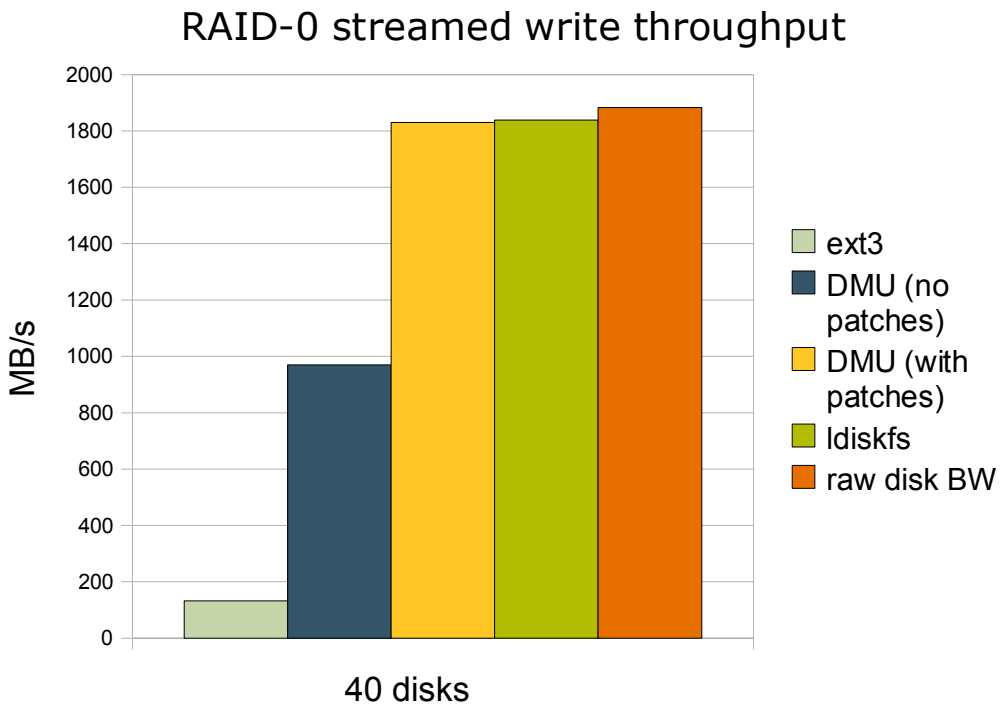
- Capacity
  - 1 trillion files per file system
  - 10 billion files per directory
  - 100 PB system capacity
  - 1 PB single file size
  - >30k client nodes
  - 100,000 open files
- Reliability
  - End-to-end data integrity
  - No performance impact during RAID build
- Performance
  - 40,000 file creates/sec from a single client node
  - 10,000 directory listings/sec aggregate
  - 30GB/sec streaming data capture from a single client node
  - 240GB/sec aggregate I/O – file per process and shared file

End-to-End Data Integrity

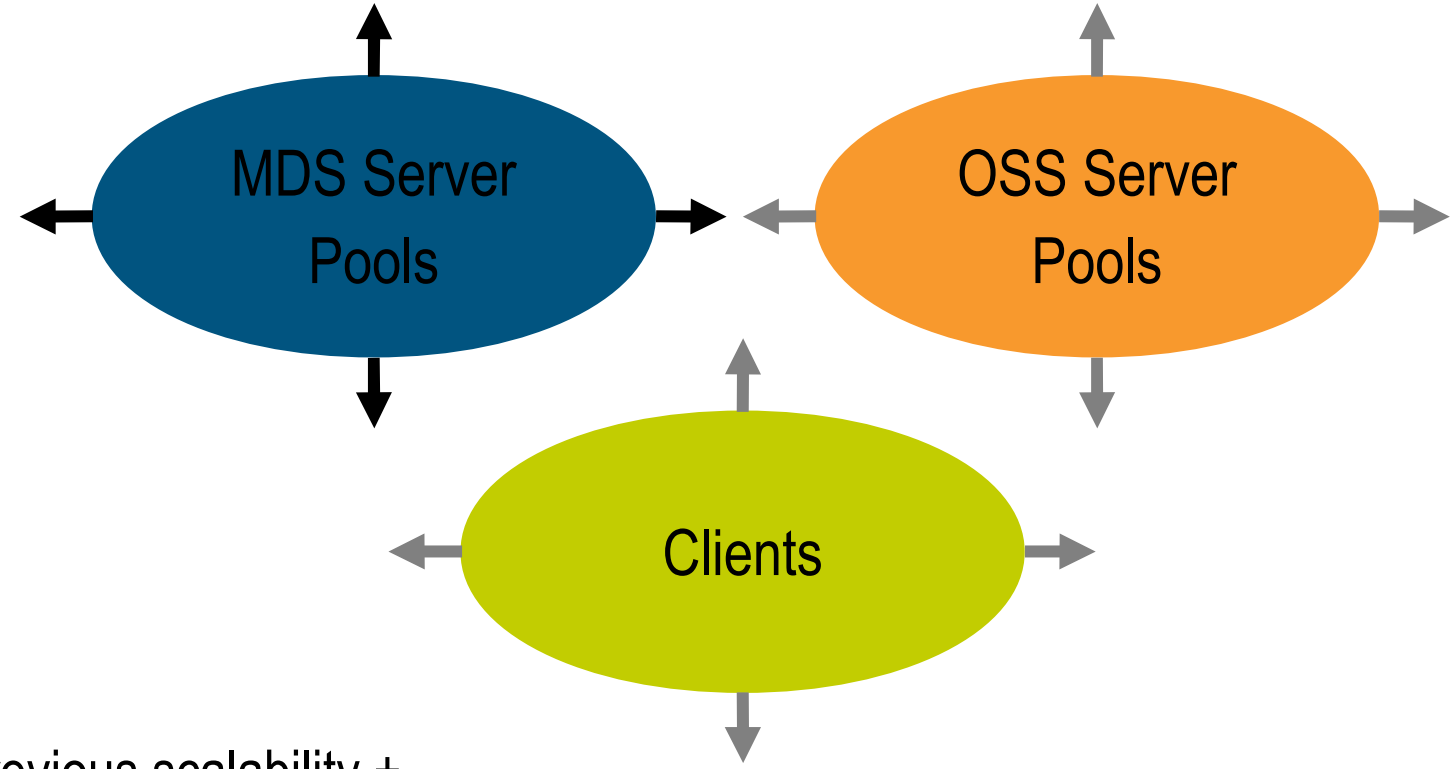
- ZFS Data Integrity
  - Copy-on-write, transactional design
  - Everything is checksummed
  - RAID-Z/Mirroring protection
  - Disk Scrubbing
- Lustre Data Integrity
  - Data is checksummed before and after network transport
  - Protects against silent data corruption anywhere along the data path, including network HBA/HCAs

ZFS Performance

- ldiskfs delivers 90% of raw disk bandwidth on Linux today
- DMU can reach par performance with ldiskfs through implementation of a zero-copy API



Clustered Metadata



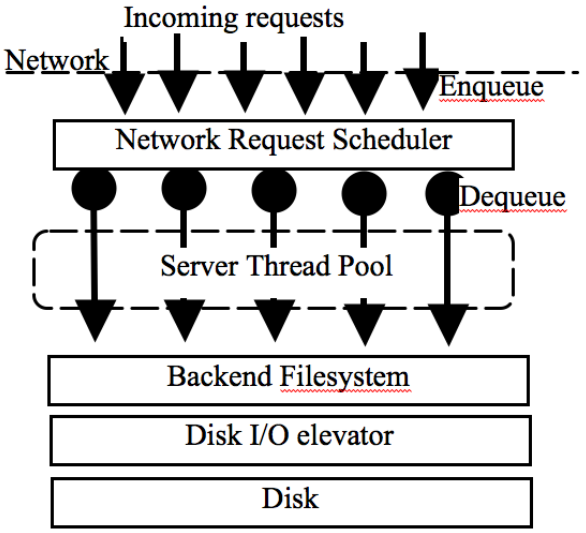
Previous scalability +
Enlarge MD Pool: enhance NetBench/SpecFS, client scalability
Limits: 100's of billions of files, millions of metadata operations / second
Load Balanced

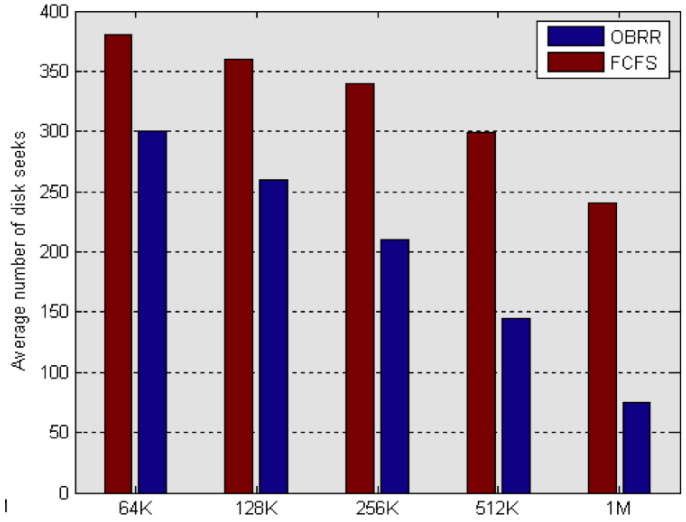
CMD Resilience / Recovery

- Scalable Pinger
  - Scalable health monitoring
  - Immune to congestion
  - Prompt cluster-wide notifications
- Epochs
  - Distributed rollback / rollforward
  - Asynchronous distributed operations
  - Client eviction
  - Server failover
  - Cluster poweroff

Network Request Scheduler

- File servers today process a request queue as FIFO
- The NRS will re-order requests
  - > Allow clients to make fair progress
  - > Re-order I/O to make it sequential on the disk
  - > Pre-fetch metadata to avoid blocking
- Estimate 30% IO performance improvement for some workloads





Metadata Writeback Cache

- Problem
  - Disk file systems make updates in memory
  - Network file systems require RPCs for metadata operations
- Goal
  - Deliver Lustre metadata performance similar to local disk file system
  - 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

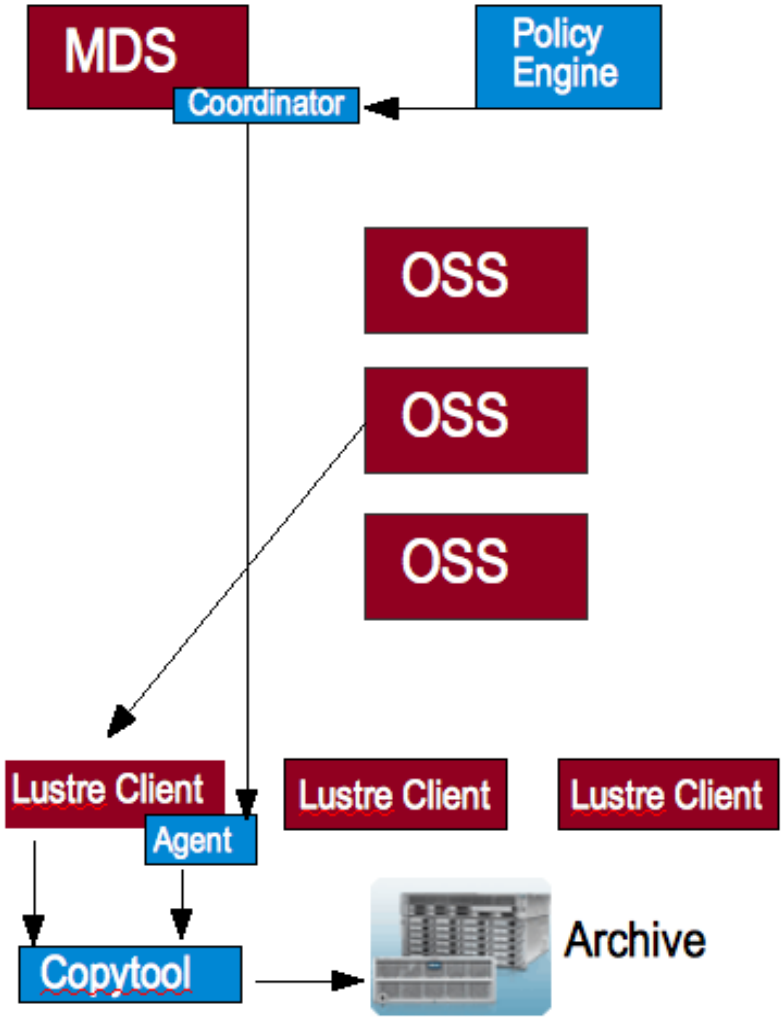
RAS

- Database providing real-time information about cluster configuration and status
  - Redundant and highly available
- Tools for monitoring and alerts, data mining, problem prediction, and control
- Lustre Fault Monitor Collector (LFMc)
  - Runs on all client, server, and router nodes
  - Reports problems to RAS database associated with the problem node, including logs and diagnostic information

HSM

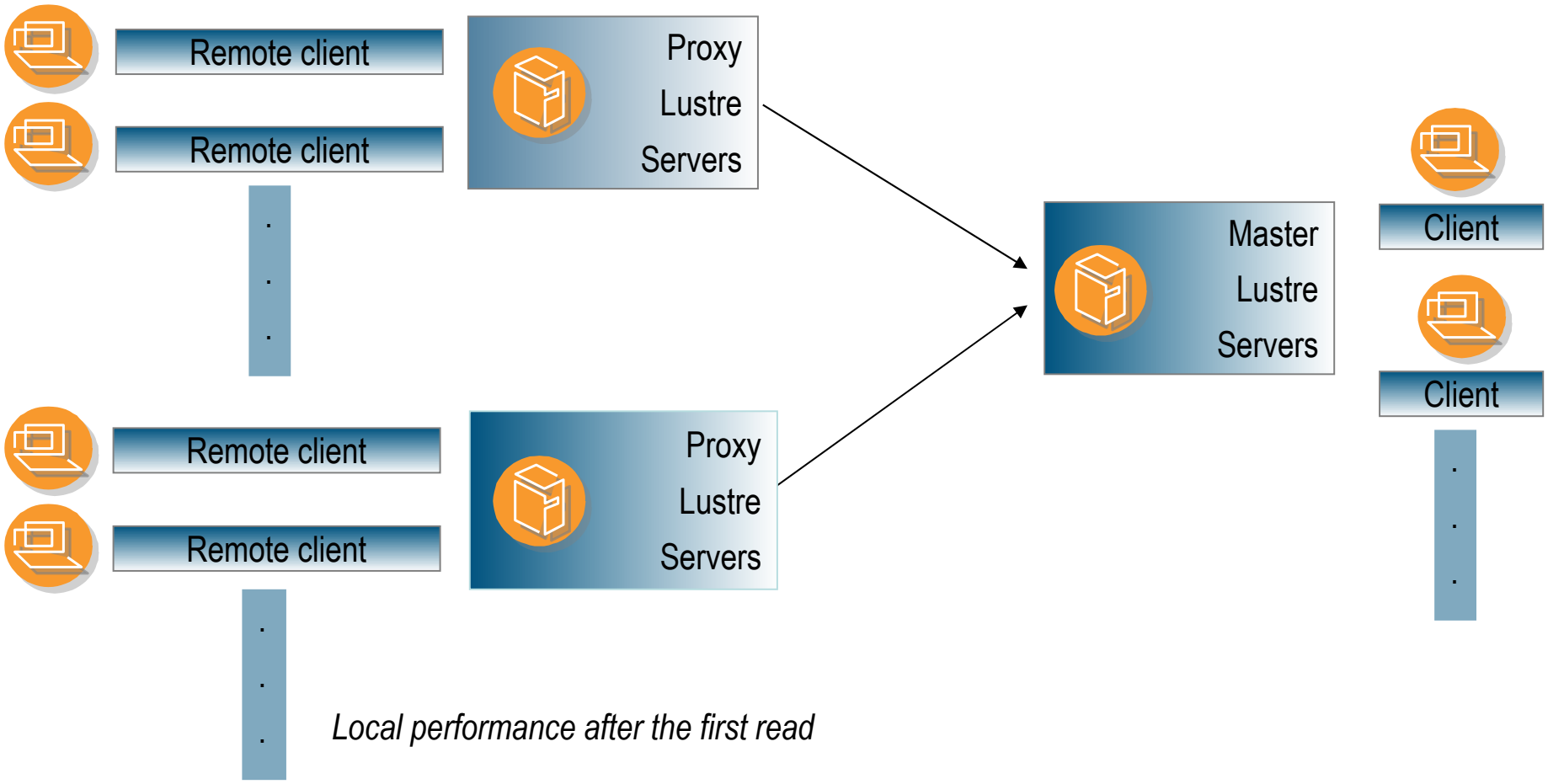
- Introducing HSM integration with Lustre
  - Archive, restore, or delete disk/flash files to lower cost archives per site policies
  - Lustre retains visibility to archived files
  - Automatic or command driven archives and recalls
  - Flexible, reliable, and performance-centric
- Open Systems Approach to Archives
  - Policy engine drives integrated MDS and OSS interfaces
  - Separate module interfaces to HSM engines
  - HPSS, Sun Storage Archive Manager (SAM) are initial HSM's to be supported
- Community and Sun Partnership
  - CEA driving overall design and implementation, as well as HPSS interface
  - Sun doing SAM, and later ADM, interfaces

Lustre HSM Overview



- Policy engine generates file list and actions
- Passes this info to Coordinator
- Coordinator knows which clients are HSM connected
- Coordinator passes action info to an Agent
- Agent initiates Copytool
- Copytool reads file and passes to archive manager
- Copytool updates MDS
- On open() with no file, Coordinator initiates retrieval

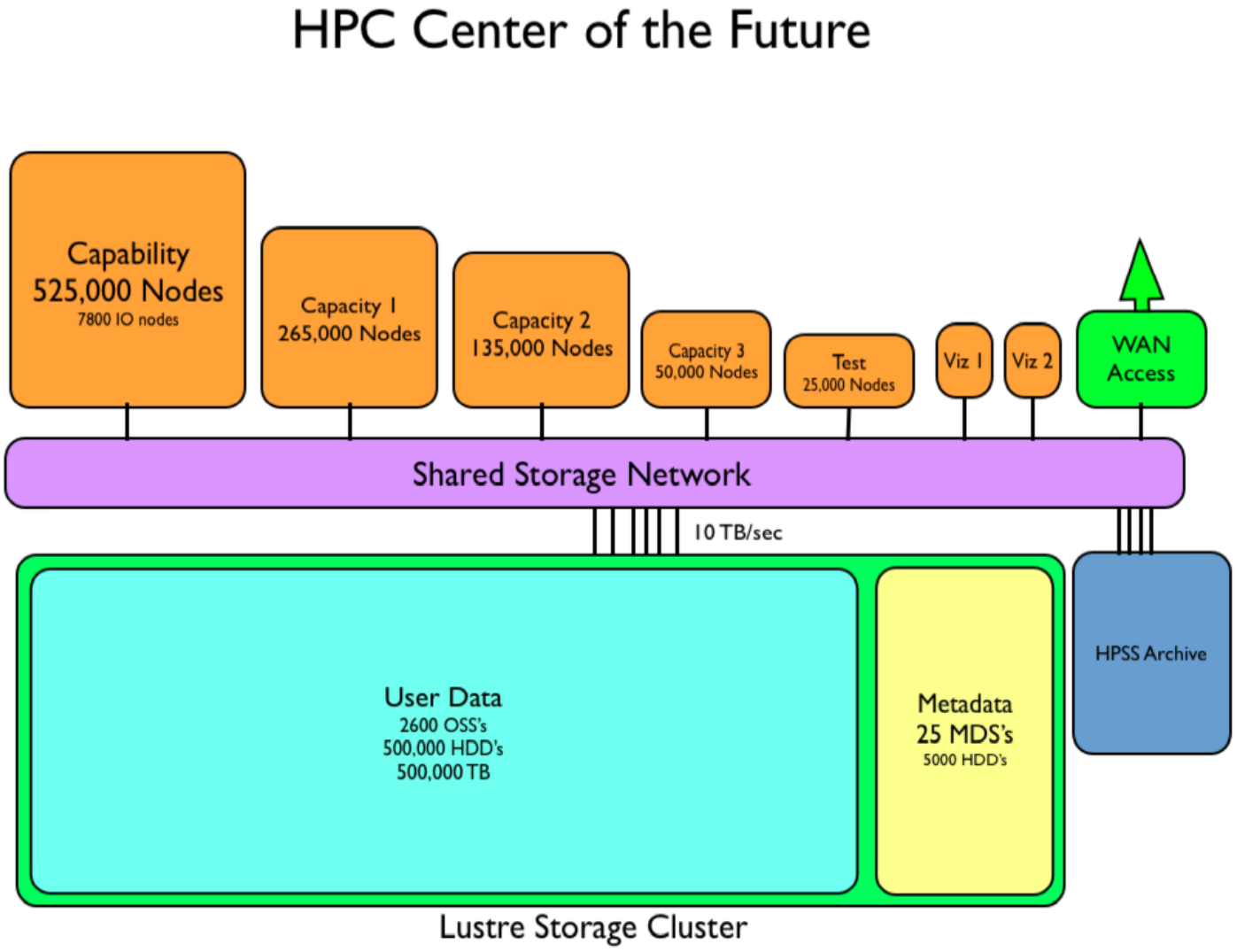
Proxy Clusters



Local performance after the first read

THANK YOU

Future Multi-PF Systems at ORNL



Lustre Scaling Requirements