MLS in HPC Storage

Nathan Rutman, Seagate

Basic MLS

- MLS check is located in the client VFS
- Policy is set on the client(s)
- SEL Audit on client
- SEL label is stored in extended attribute (EA)

Challenges for a distributed FS

- Multiple clients
- Multiple server components
- FS split between clients and servers
  - Syscall and disk ops are on different nodes
- Disaggregated control, policy, audit
- Network is involved
- Performance

Lustre Overview

- Lustre is a high-end, scalable networked file system
  - Comparable to NFS or Samba in concept
  - Exploits multiple nodes and drives simultaneously for performance
- Used in a large number of large-scale US Government and commercial sites today
- Clients and servers are Linux loadable kernel modules
- Mainline tree assumes physical security + Unix

ClusterStor Product Line Overview

ClusterStor 6000
6GB/sec increments
Up to1TB/sec
Up to 25PB
Custom Rack

ClusterStor 1500
1GB/sec increments
1 to110GB/sec
Up to 7PB
Standard Rack

ClusterStor 9000
9GB/sec increments
Up to1TB/sec+
Up to 25PB+
Custom Rack

Performance

Metadata Servers

2U24 + quad node server
MDS, MGS, Management
Enterprise SAS dual-ported RAID10
HA redundancy with active/passive Lustre failover

- Disk Configuration
  - Qty 4 Lustre Management (MGS)
  - Qty 4 ClusterStor Management and NFS
  - Qty 2 Global Hot spares
  - Qty 14 Drives for MDT
- Disk & Inode
  - w/ 600 GB 10K RPM HDDs
    - **881 million inode entrie**
  - w/ 900 GB 10K RPM HDDs
    - **1,300 million inode entries**

Object Storage Servers

5U84 Dual ESM
Enterprise SAS dual-ported RAID6/PDRAID
HA Redundancy with active/active Lustre failover

Pair of H/A Embedded Application Servers
- **CS9000: = 9 GB/sec over Infiniband**
- **CS6000: = 6 GB/sec over Infiniband**

QSFP port Supports IB QDR/FDR or 10/40 GbE Network Link

Data Protection/Integrity
- 2 OSS's per SSU
- 1 OST per OSS (GridRAID)

2x SSD OSS journal disks for increased performance
- 2x Hot Spare HDD's

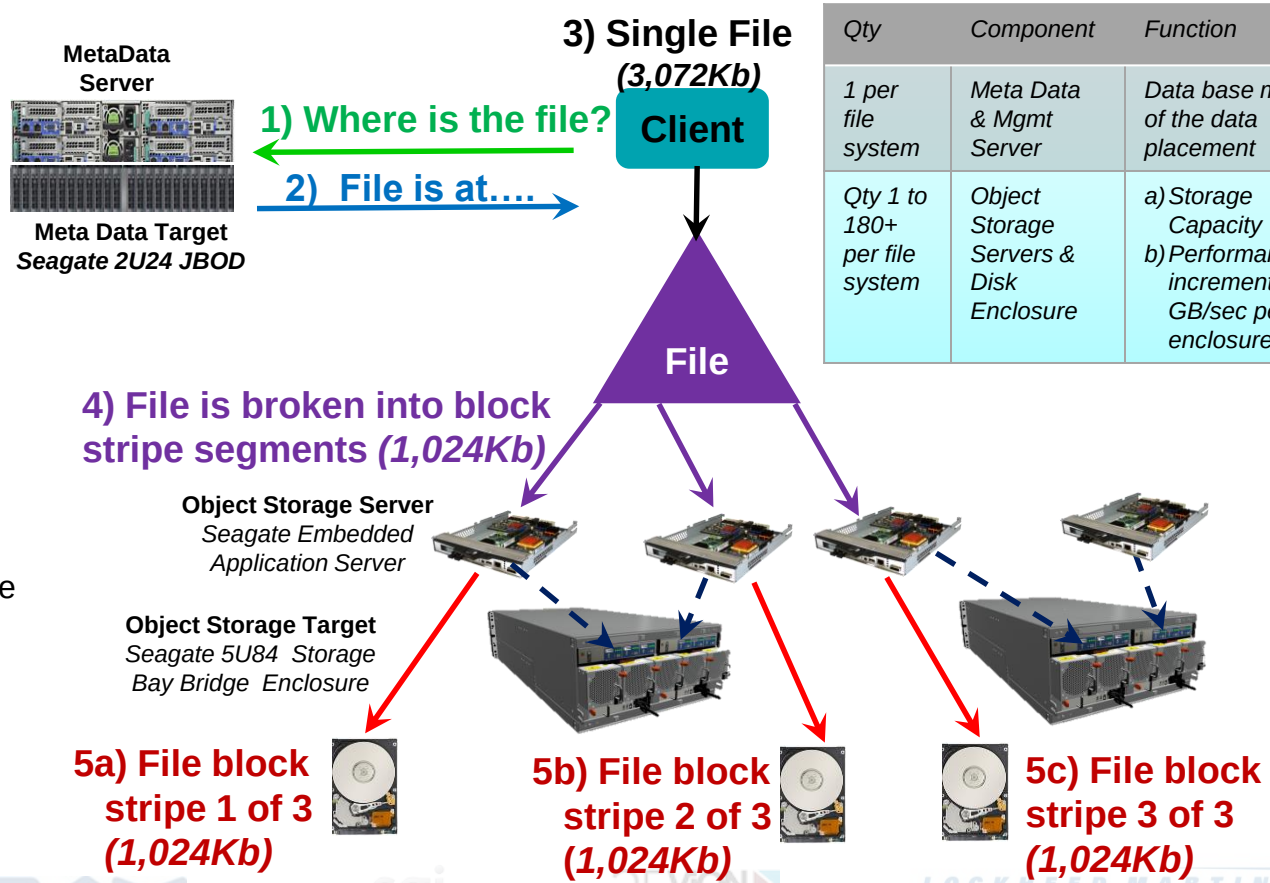
82 Usable Data Disks per SSU
- **2TB HDD x 82 - 128TB usable per SSU**
- **3TB HDD x 82 - 192TB usable per SSU**
- **4TB HDD x 82 - 256TB usable per SSU**
- **6TB HDD x 82 - 362TB usable per SSU**

Lustre file component mapping

The Lustre File System is the best file system for large file sequential write and read I/O operation.

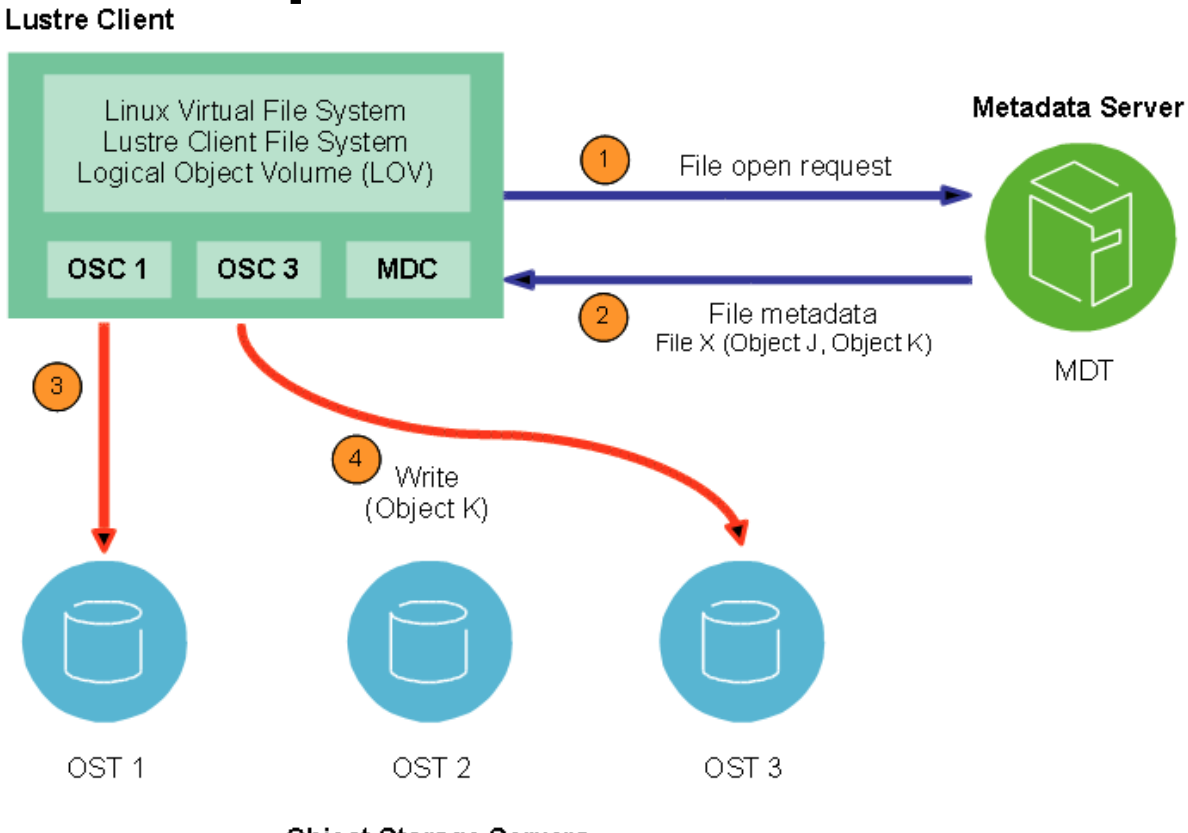
The Lustre File System breaks down large files sizes into 1MB block stripe sizes. The 1MB Block Stripes are written in a simultaneous operation across multiple servers providing the “parallel” operation eliminating traditional serialization performance bottleneck.

The largest ClusterStor installation (NSCA) supports 25,000 clients writing and reading across 360 Object Storage Servers at over 1TB/s



Qty	Component	Function
1 per file system	Meta Data & Mgmt Server	Data base map of the data placement
Qty 1 to 180+ per file system	Object Storage Servers & Disk Enclosure	a) Storage Capacity b) Performance increments of 9 GB/sec per enclosure

Lustre Open RPC



SEL on Lustre clients

- SEL, enforcing mode, MLS
- Special version of Lustre includes context with RPC
- Servers verify clients are running SEL
- Clients do local MLS policy enforcement
- Label caching / handling for performance
- Minimize number of extra RPCs

SEL on Lustre servers

- SEL, enforcing mode, MLS
- No user access to servers
- No admin/root access to Lustre data
- Only allow Lustre kernel processes to access storage
- Full audit of admin ops

SEL on Lustre servers, cont

- Modified Lustre protocol sends context label with each RPC
- Servers use ext4 variant (ldiskfs)
- MDS stores security label as local EA at create
- Distribute label with file stripes at first write
- OSTs store security label as well
- Every local disk component is MLS protected

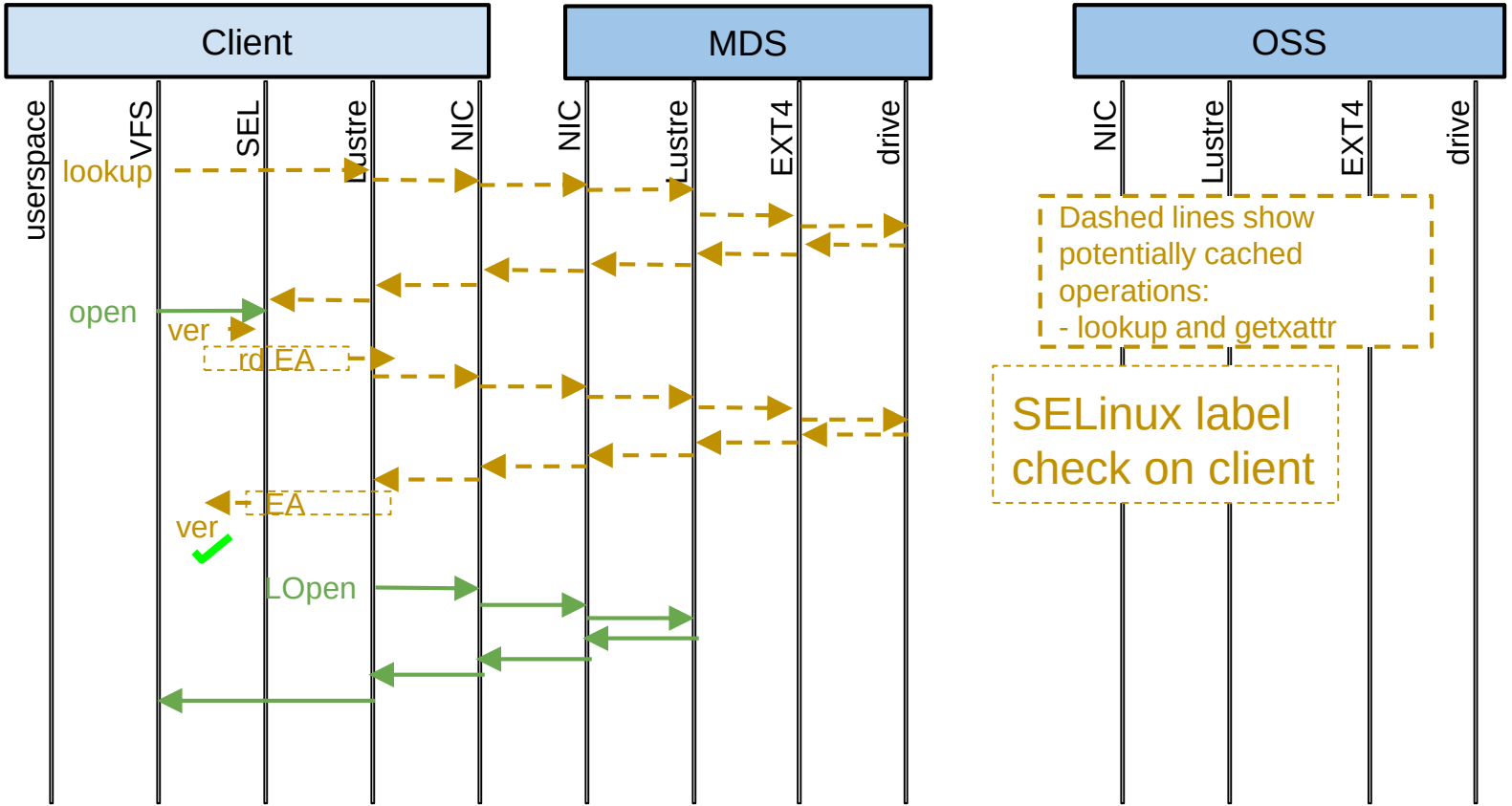
MDS MLS recheck

- Context sent with RPC can be re-verified on server
- Insure uniform/minimal policy
- Centralized audit
  - Includes client ID

Security EA caching

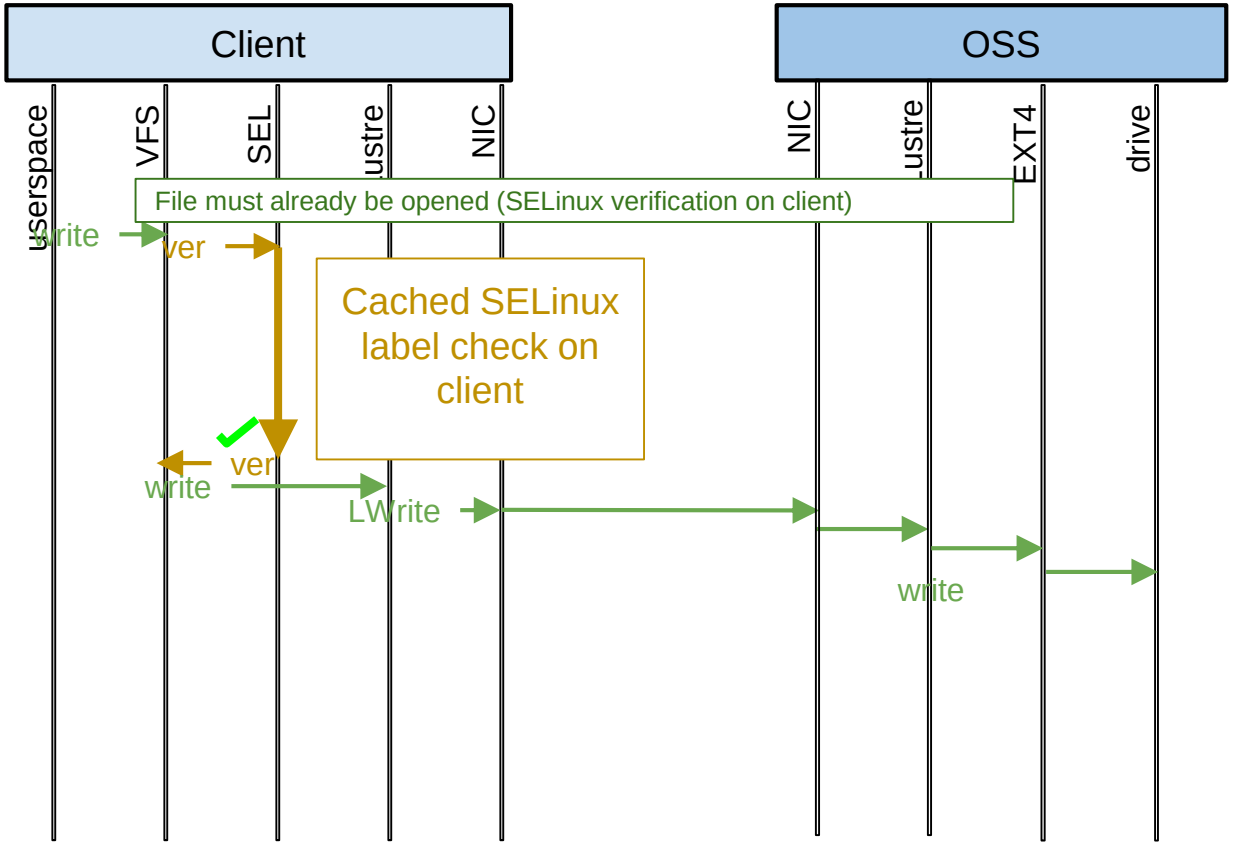
- SEL check at every syscall
- Don’t want to go back to MDS for every one
- Cache it at first access
- But someone else may change label
- Shared reader lock on cached EA
- SEL local cache must also be invalidated

SDA flow diagram: open(2) (w/o MDS recheck)



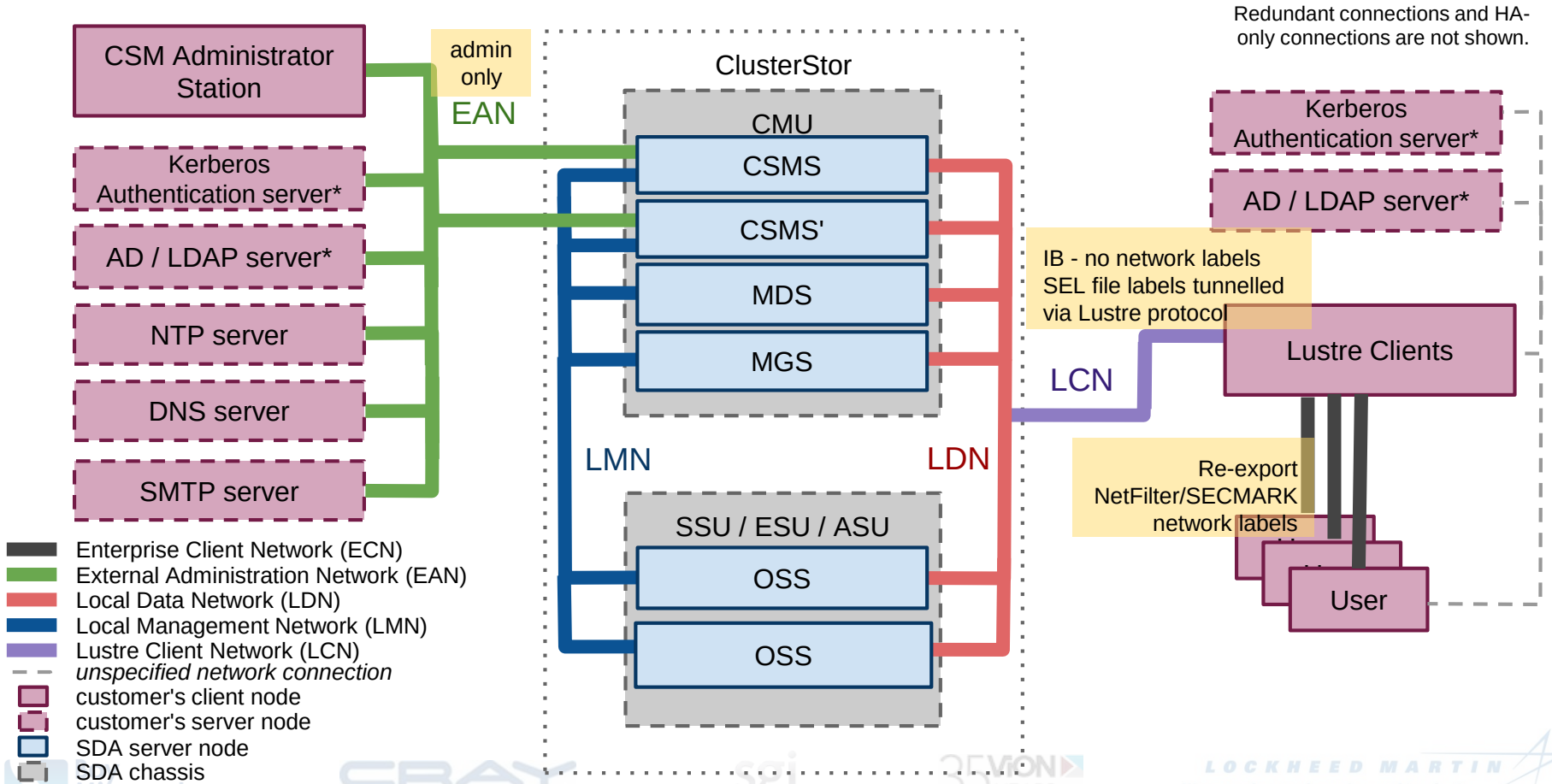
✓ SELinux module verification of process context with security label EA

SDA flow diagram: write(2) (w/o MDS recheck)



✓ SELinux module verification of process context with security label EA

Networks Overview

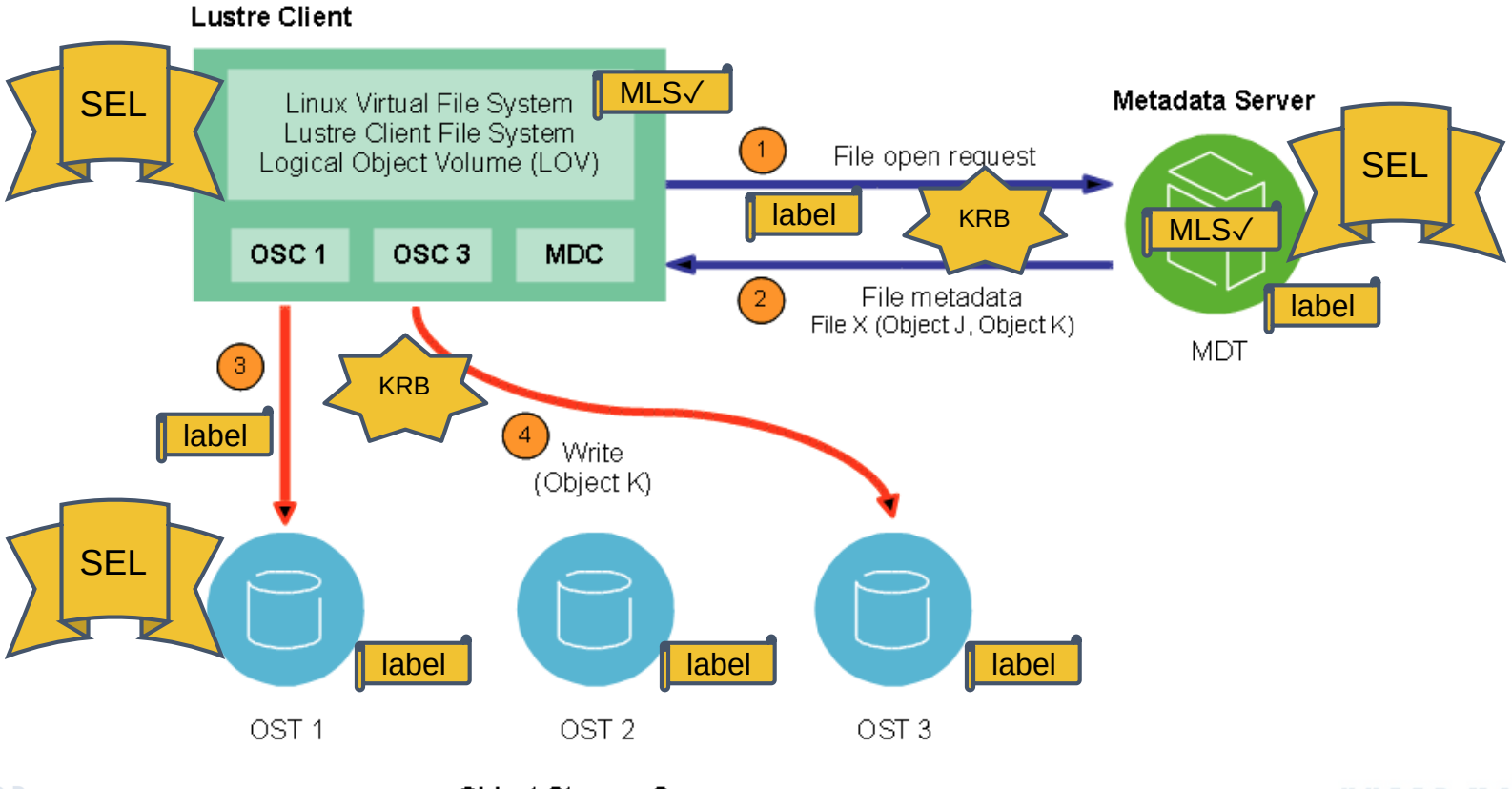


* Kerberos and AD/LDAP servers, if present, must be common to the SDA and Lustre clients

Network Security

- Lustre Secure RPC protocol
- Kerberized communication client-servers
  - header and/or payload
  - checksum, integrity, or privacy
- Options with different performance impacts
- Works over IB and TCP

Lustre Open RPC



Misc

- Approvals process - ICD-503
- Manufacturing process - secure US factory site