The Lustre Storage Architecture

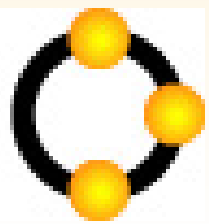
Peter J. Braam

President and Chief Technology Officer

braam@clusterfs.com

http://www.clusterfs.com

Cluster File Systems, Inc



Topics

- What is Lustre 1.0?
- How does Lustre work?
- Lustre 2.0/3.0

- Cluster File Systems, Inc

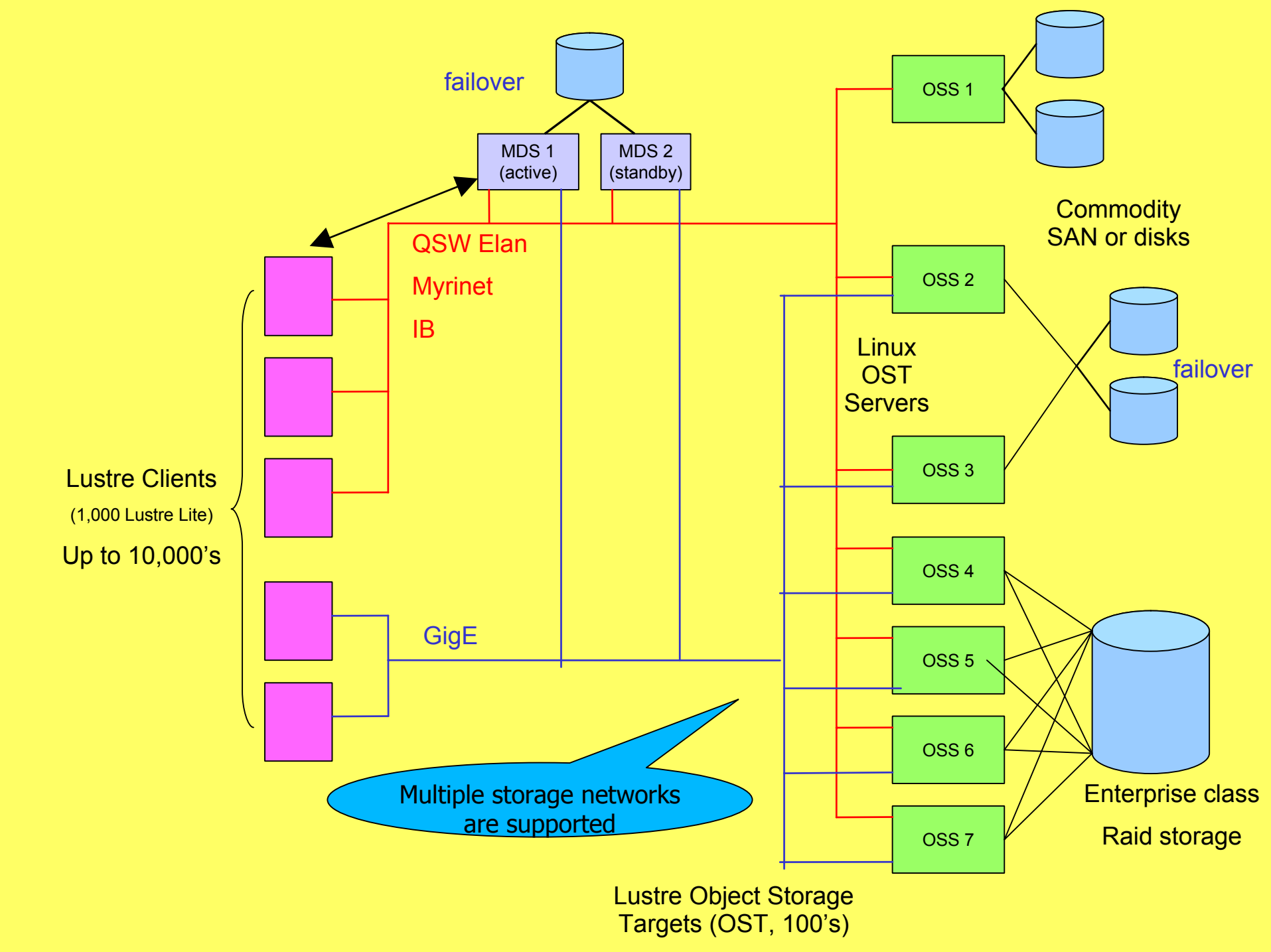
What is Lustre 1.0?

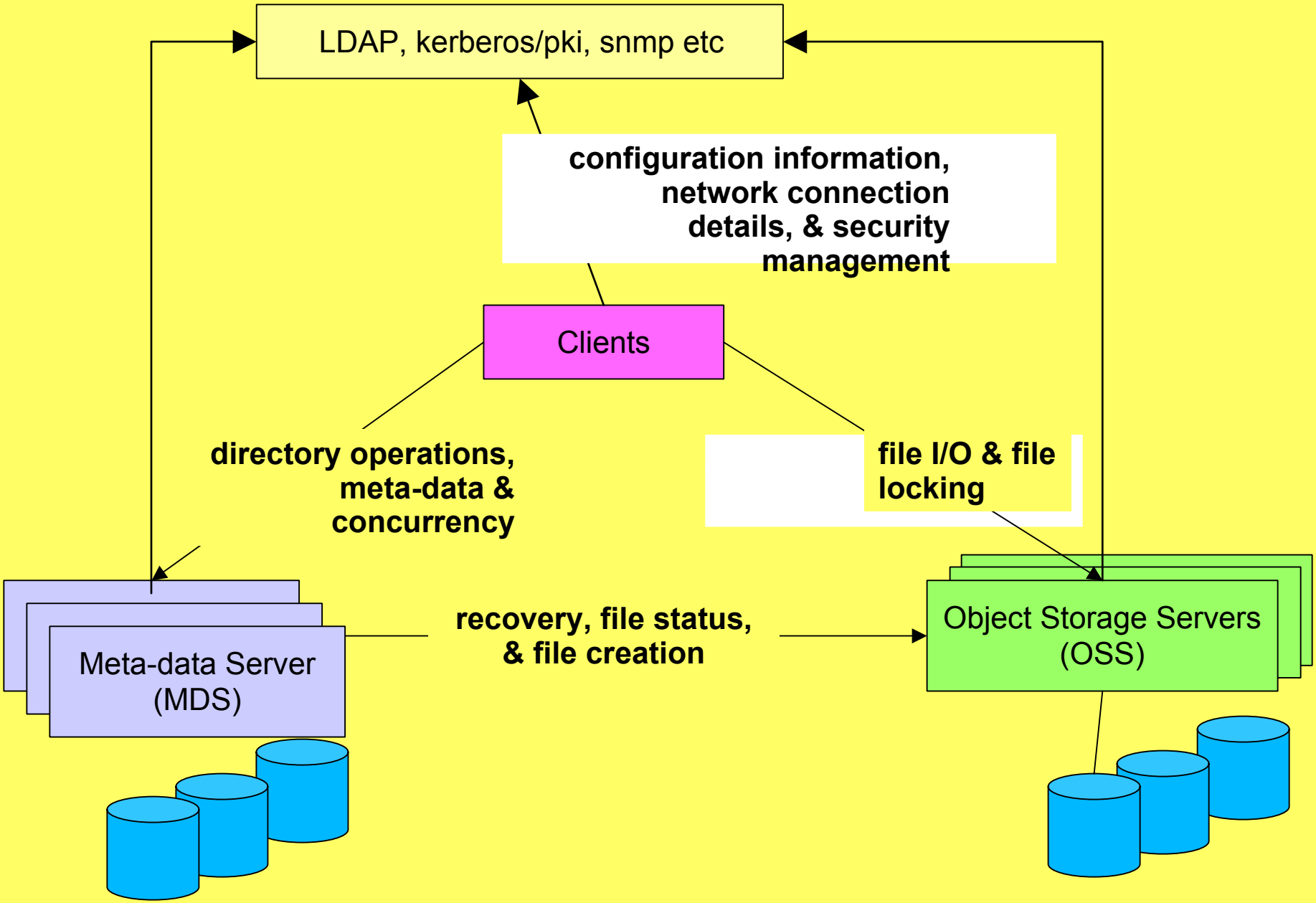
Lustre 1.0

- A shared file system for HPC clusters
  - Open Source software (GPL)
  - Commodity servers, storage, networking
- Very high metadata and I/O performance
  - 5,000 file creations/sec in 1 dir, 1,000 nodes
  - Client application: up to 288 MB/sec,
  - Aggregate: up to 11 GB/sec (B=bytes, writes)
- Scalable to 1,000's of nodes
  - In production now on such clusters
- Completely POSIX compliant

Lustre systems in a cluster

- Clients
  - 1000's now, 10,000's future
  - Obtain access to Lustre file system
  - Typical role: Linux compute server
- OSS – transparent failover
  - 100's now, 1000's future
  - Object storage servers (formerly OST's)
  - Linux servers handling (stripes of) file data
- MDS – transparent failover
  - 2 now, 10's future
  - Metadata request transaction engine.
  - Linux server handling metadata requests
- Also: LDAP, Kerberos, SNMP, routers etc.


failover
MDS 1
(active)
MDS 2
(standby)
OSS 1
Commodity
SAN or disks
QSW Elan
Myrinet
IB
OSS 2
Linux
OST
Servers
failover
OSS 3
Lustre Clients
(1,000 Lustre Lite)
Up to 10,000's
OSS 4
GigE
OSS 5
OSS 6
Enterprise class
Raid storage
Multiple storage networks
are supported
OSS 7
Lustre Object Storage
Targets (OST, 100's)

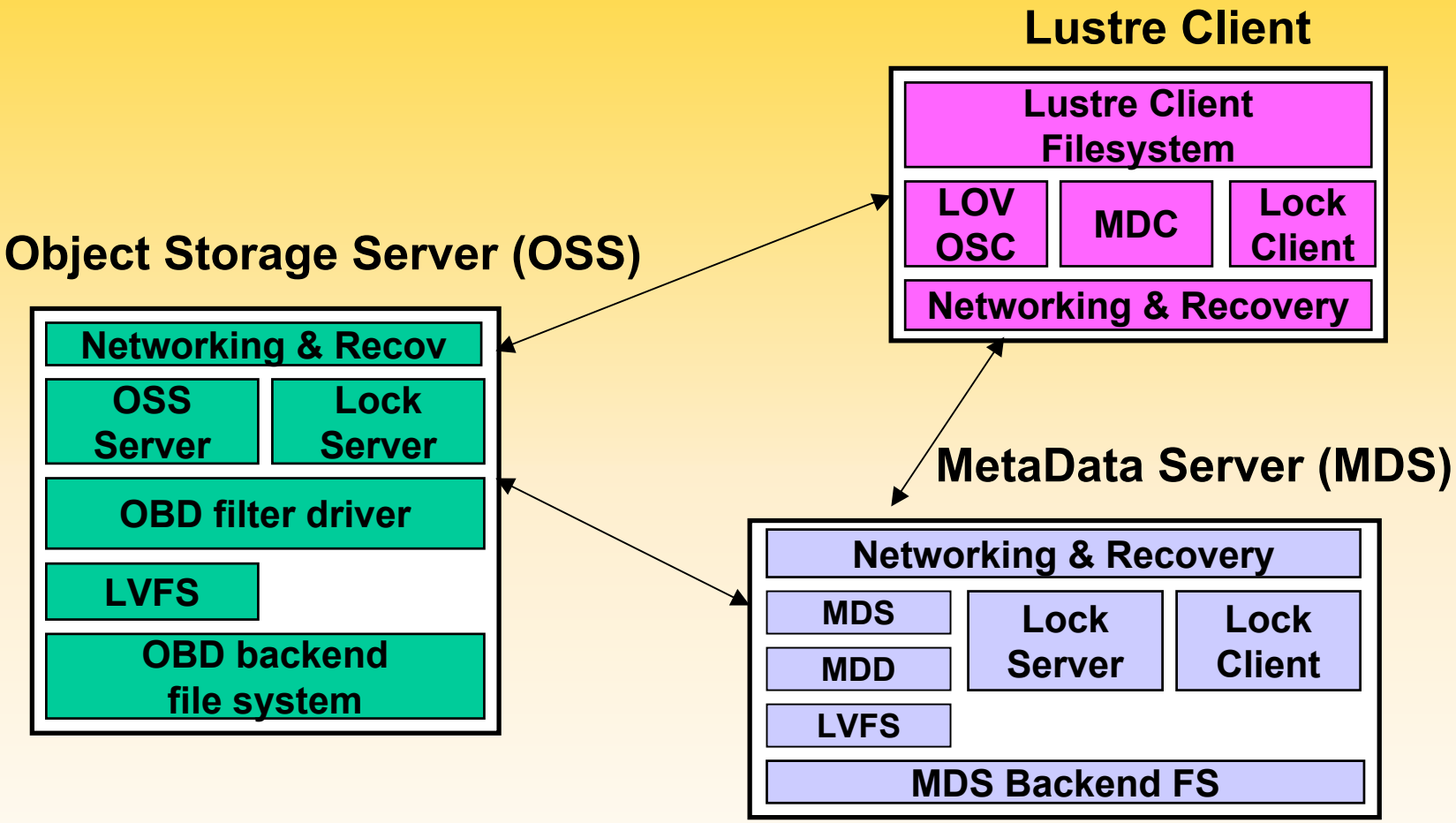

LDAP, kerberos/pki, snmp etc
configuration information, network connection details, & security management
Clients
directory operations, meta-data & concurrency
file I/O & file locking
Meta-data Server (MDS)
recovery, file status, & file creation
Object Storage Servers (OSS)

Lustre 1.X, X=1,2,3…

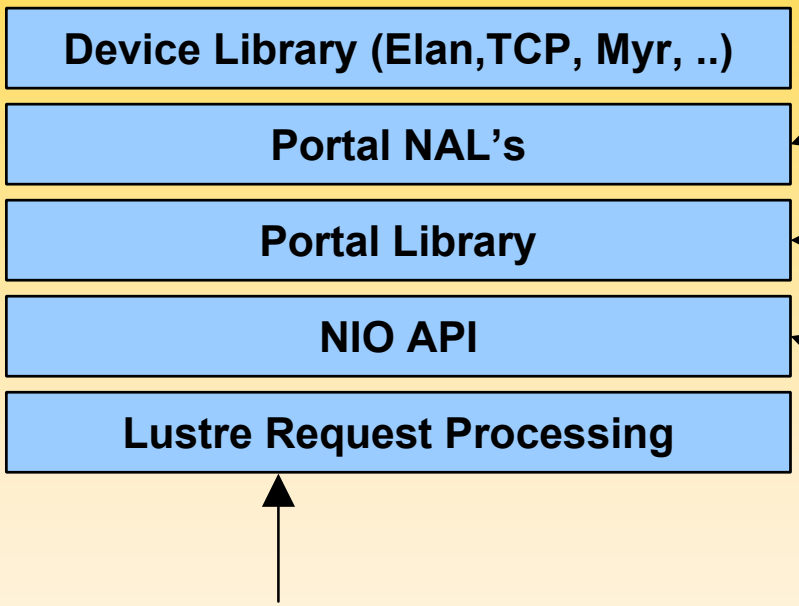
- Small additions, e.g.
  - RAID1 OSS servers
  - New network types (I/B)
  - Red Storm client fs
  - Improved management, recovery
    - Online removal/addition of OST's
- All currently planned deployments
  - will run or run 1.X version

How does Lustre work?

Building blocks



Lustre Network Stack - Portals



Network Abstraction Layer for TCP, QSW, etc. Small & hard Includes routing api.

Sandia's API, CFS improved impl.

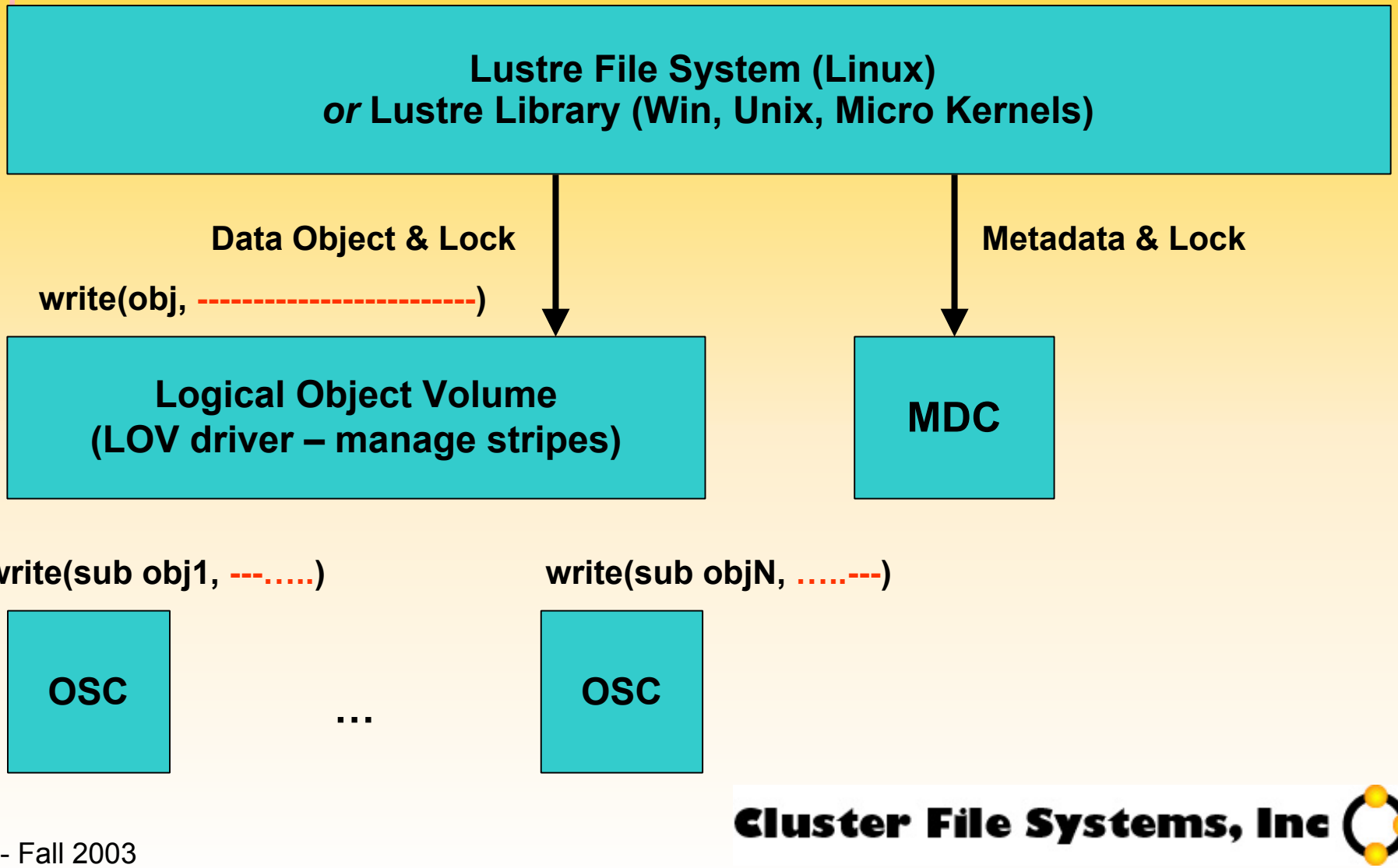
Move small & large buffers, Remote DMA handling, Generate events

0-copy marshalling libraries,
Service framework,
Client request dispatch,
Connection & address naming,
Generic recovery infrastructure

Result: File I/O goes at close to wire efficiency (>90%)

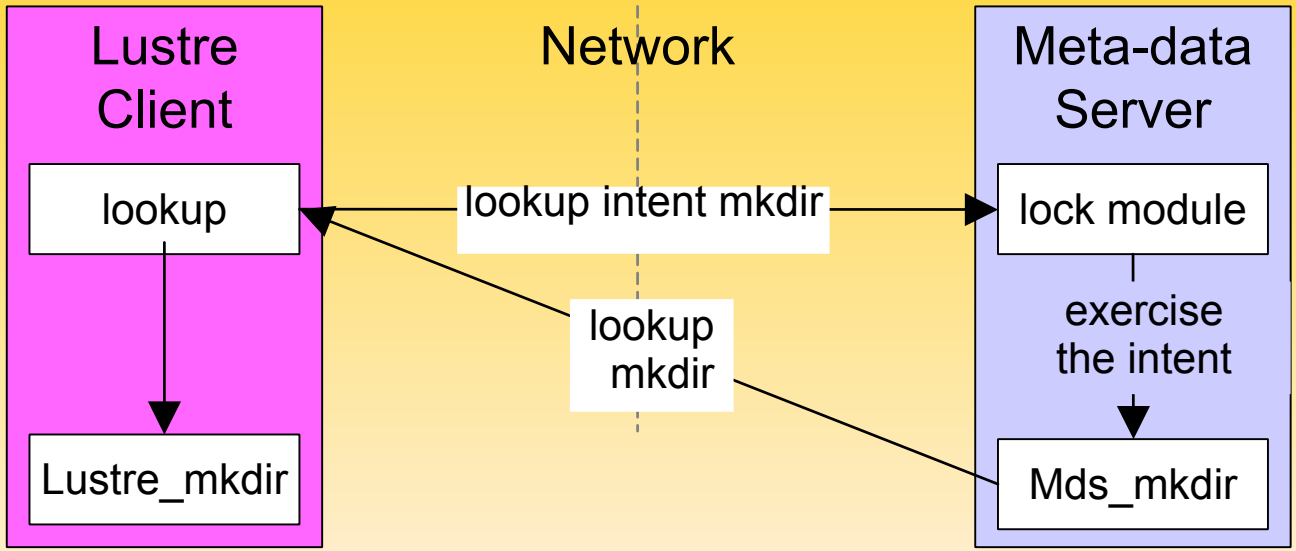
Gige: 110-118 MB/sec
Dual gige: 210 MB/sec
Quadrics Elan3: 288 MB/sec

Lustre 1.0 clients API's

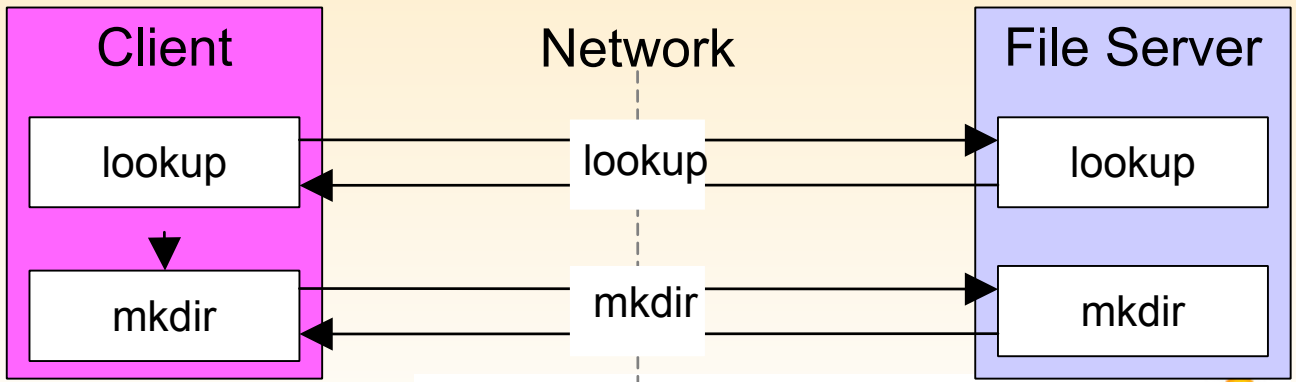


Single RPC metadata processing

Lustre
1 rpc



Conventional
multiple RPCs



Cluster File Systems, Inc

Lustre 2.0 & 3.0

Ingredients

- Core system largely similar
- Radical extension are possible with modules
  - Cache & proxy servers, replication
  - Clustered Metadata
  - Security
  - Small file optimizations
  - Management: hot migration etc.
  - Snapshots
  - HSM
  - File sets

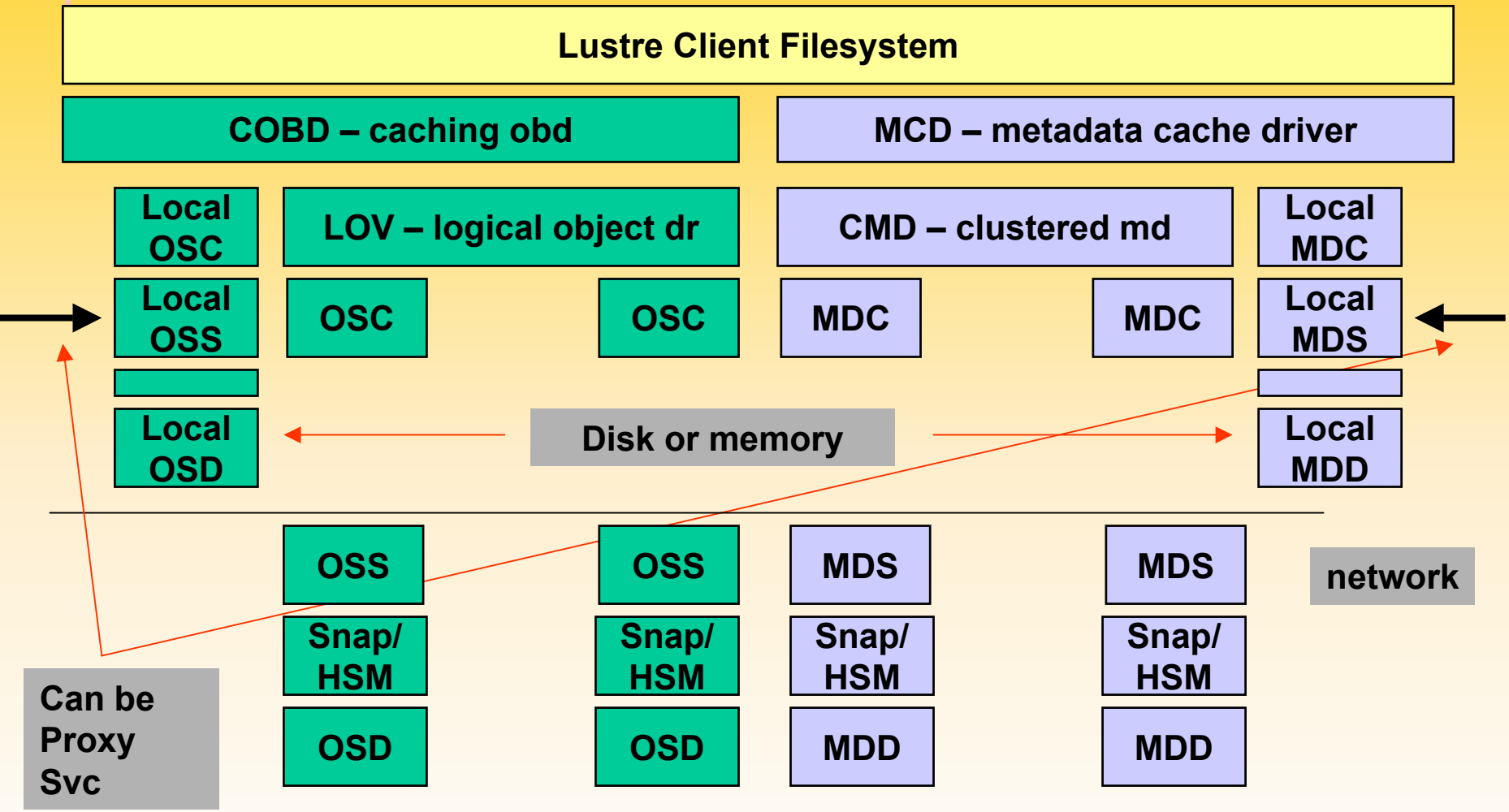
Metadata roadmap

- Subtree locks
  - New, fast validation routine
  - Concurrency causes transition between
    - Subtree lock
    - Node lock – lock one point in namespace
- Clustered Metadata
  - Like the LOV, now CMD
    - Stripes of large metadata
    - Locate metadata on other servers
    - Common operations no extra overhead

Caches & proxies

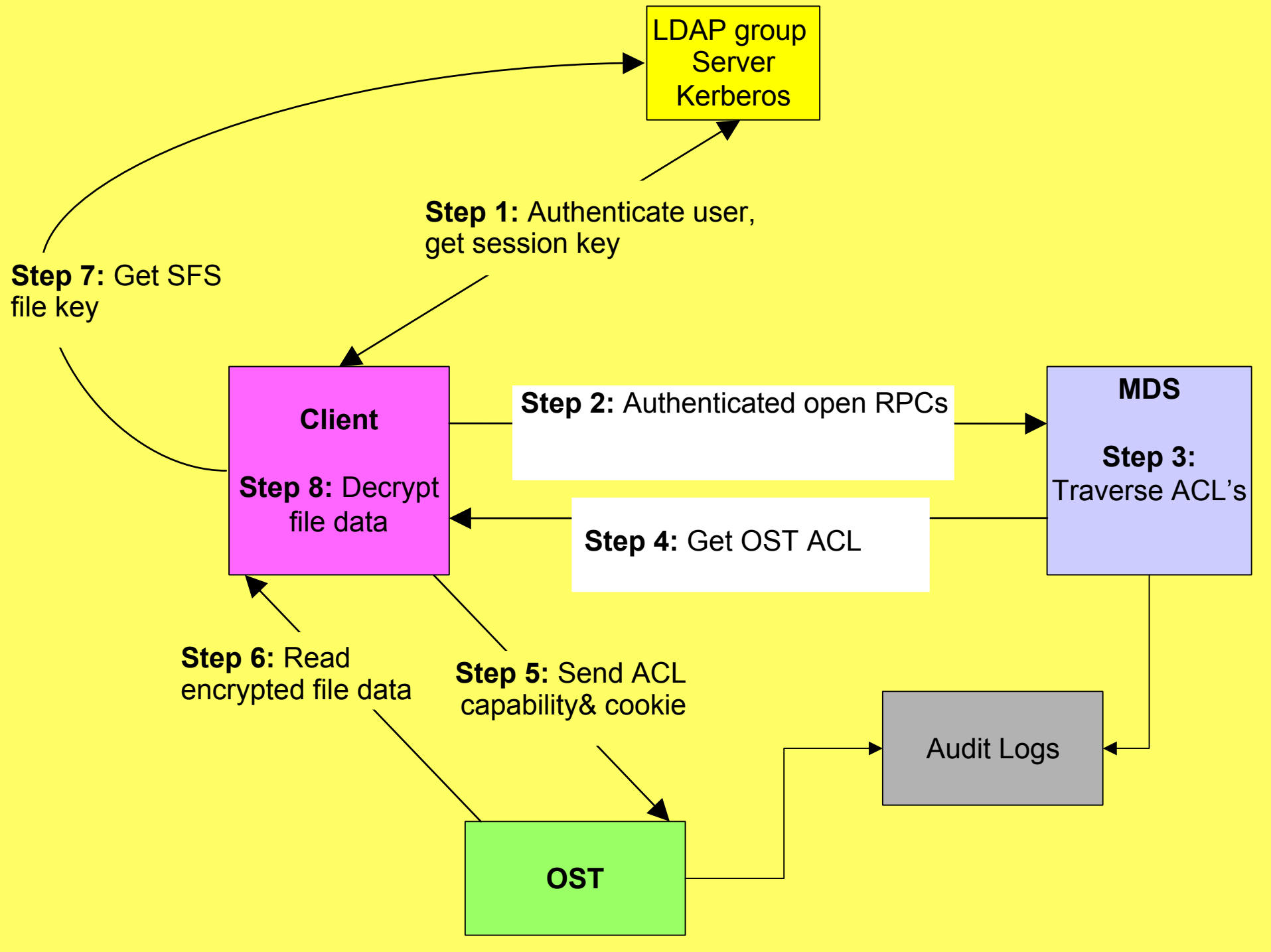
- MCD – metadata cache driver
  - Offers Metadata service
    - Local (WB cache) or networked (proxy)
  - choose:
    - Locally running MDS
    - Remote MDS
- Simple to build
  - Component and protocol re-use
- Similar, ro caching obd (cobd) exists today

Lustre 2.0/3.0 components



Security

- Authentication: kerberos & PKI
- POSIX style authorization
- NASD style OST authorization
  - Refinement: use OST ACL's and cookies
- File crypting with group key service
  - STK secure file system
- Audit logs
  - Failed and successful operations


LDAP group
Server
Kerberos
Step 1: Authenticate user,
get session key
Step 7: Get SFS
file key
Client
Step 8: Decrypt
file data
Step 2: Authenticated open RPCs
MDS
Step 3:
Traverse ACL's
Step 4: Get OST ACL
Step 6: Read
encrypted file data
Step 5: Send ACL
capability& cookie
Audit Logs
OST

Cluster File Systems, Inc.

Cluster File Systems, Inc.

- Principal maintainer of Lustre
- Lustre Engineering
- Lustre Support
  - Standard 8x5 to Mission Critical 24x7
  - Through Partners
  - Direct
- Lustre Training
  - On-site deployment and administration training

Lustre Retrospective

- 1999 Initial ideas @CMU
- Seagate: management aspects, prototypes
  - Much survives today
- 2000 National Labs
  - Can Lustre be next generation FS?
    - 100 GB/sec, trillion files, 10,000's clients, secure, PBs
- 2002 – 2003 Fast lane
  - MCR, PNNL, ASCI Pathforward, Cray Redstorm, NCSA
  - Many partners: Dell, HP, Cray, LNXI, DDN others
  - Production use, 1.0 released