The Lustre Storage Architecture

Linux Clusters for Super Computing

Linköping 2003

Peter J. Braam
braam@clusterfs.com
http://www.clusterfs.com

Tim Reddin
tim.reddin@hp.com

Cluster File Systems, Inc

Topics

- History of project
- High level picture
- Networking
- Devices and fundamental API's
- File I/O
- Metadata & recovery
- Project status
- Cluster File Systems, Inc

Lustre's History

Project history

- 1999 CMU & Seagate
  - Worked with Seagate for one year
  - Storage management, clustering
  - Built prototypes, much design
  - Much survives today

2000-2002 File system challenge

- First put forward Sep 1999 Santa Fe
- New architecture for National Labs
- Characteristics:
  - 100's GB's/sec of I/O throughput
  - trillions of files
  - 10,000's of nodes
  - Petabytes
- From start Garth & Peter in the running

2002 – 2003 fast lane

- 3 year ASCI Path Forward contract
  - with HP and Intel

- MCR & ALC, 2x 1000 node Linux Clusters
- PNNL HP IA64, 1000 node Linux cluster
- Red Storm, Sandia (8000 nodes, Cray)
- Lustre Lite 1.0
- Many partnerships (HP, Dell, DDN, …)

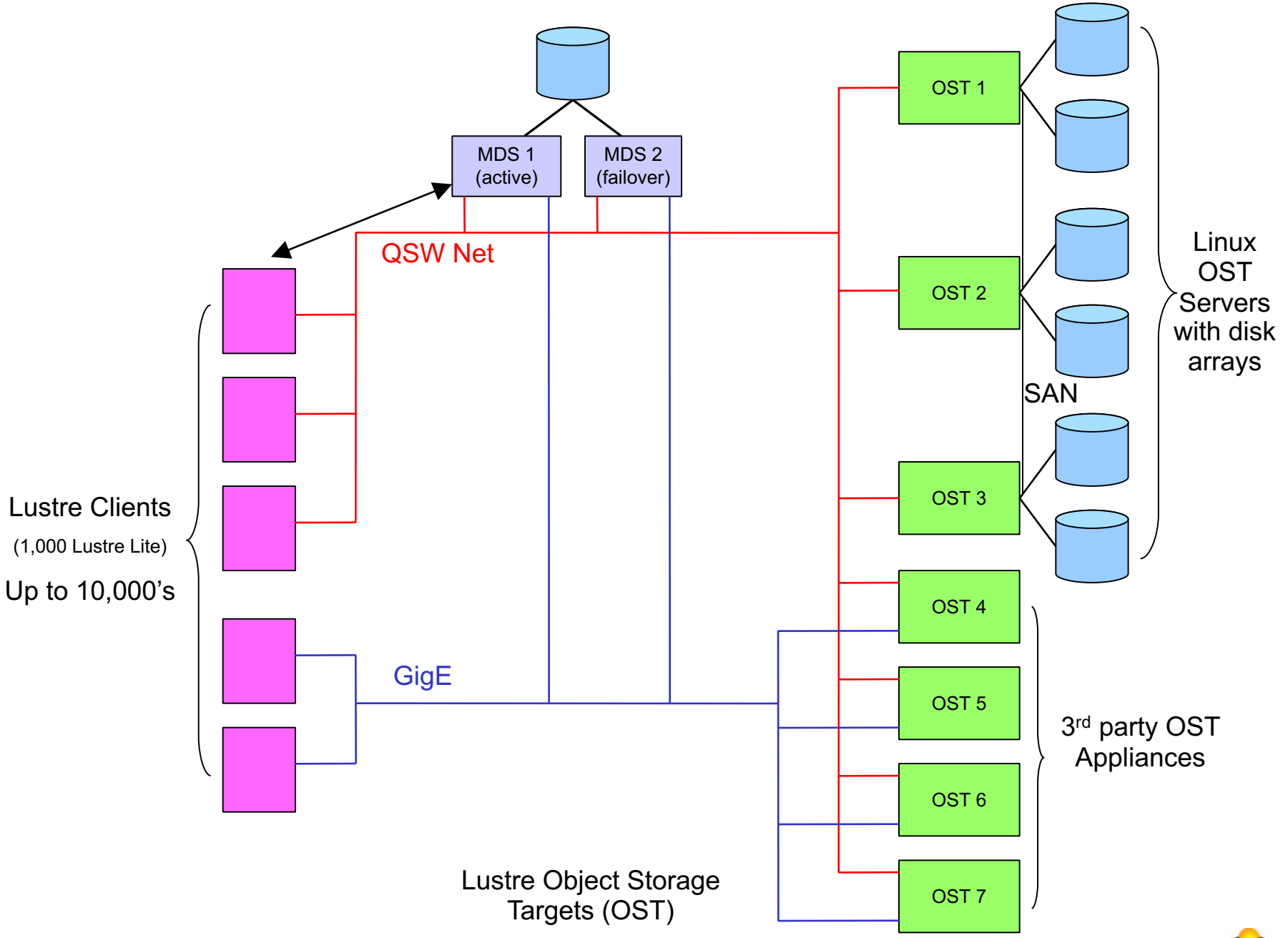
2003 – Production, perfomance

- Spring and summer
  - LLNL MCR from no, to partial, to full time use
  - PNNL similar
  - Stability much improved
- Performance
  - Summer 2003: I/O problems tackled
  - Metadata much faster
- Dec/Jan
  - Lustre 1.0

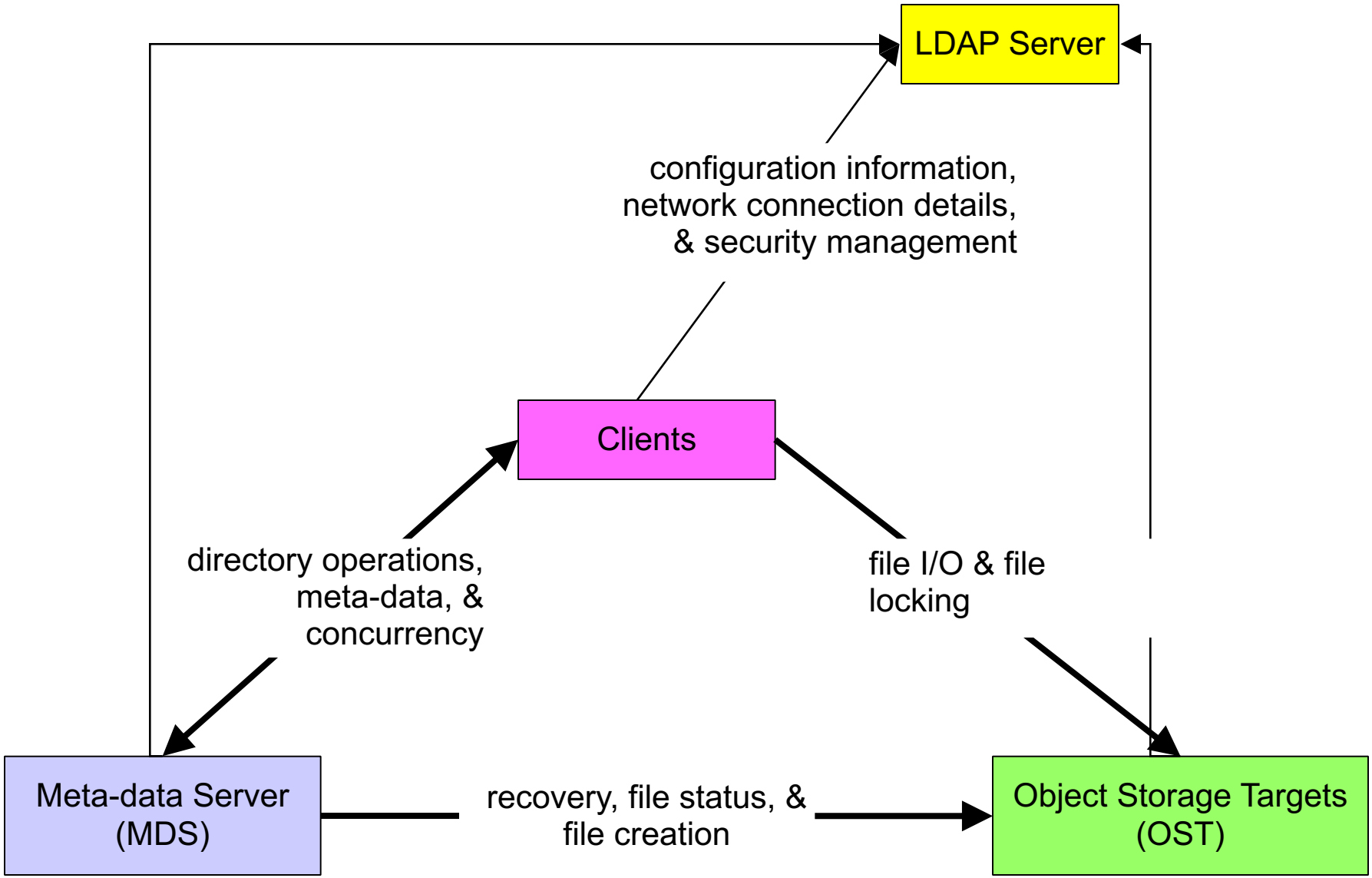
High level picture

Lustre Systems – Major Components

- Clients
  - Have access to file system
  - Typical role: compute server
- OST
  - Object storage targets
  - Handle (stripes of, references to) file data
- MDS
  - Metadata request transaction engine.
- Also: LDAP, Kerberos, routers etc.



Cluster File Systems, Inc


LDAP Server
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 Targets
(OST)

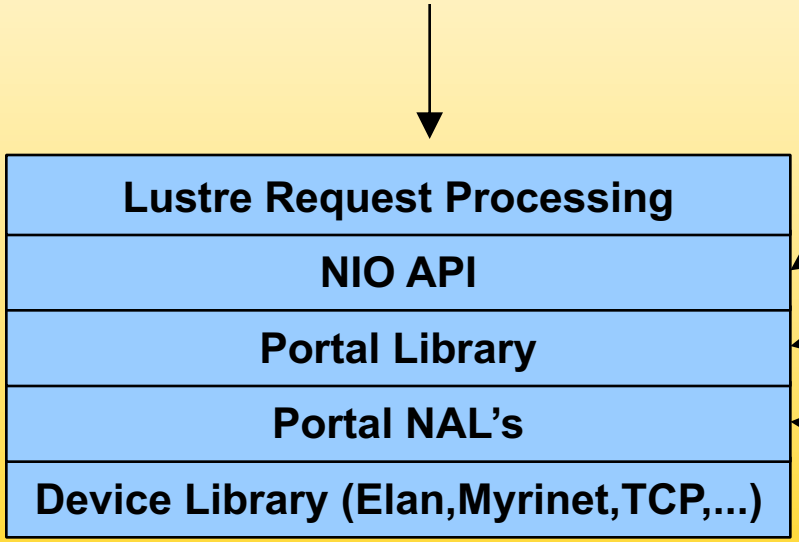
Networking

Lustre Networking

- Currently runs over:
  - TCP
  - Quadrics Elan 3 & 4
  - Lustre can route & can use heterogeneous nets
- Beta
  - Myrinet, SCI
- Under development
  - SAN (FC/iSCSI), I/B
- Planned:
  - SCTP, some special NUMA and other nets

Lustre Network Stack - Portals

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



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

Sandia's API,
CFS improved impl.

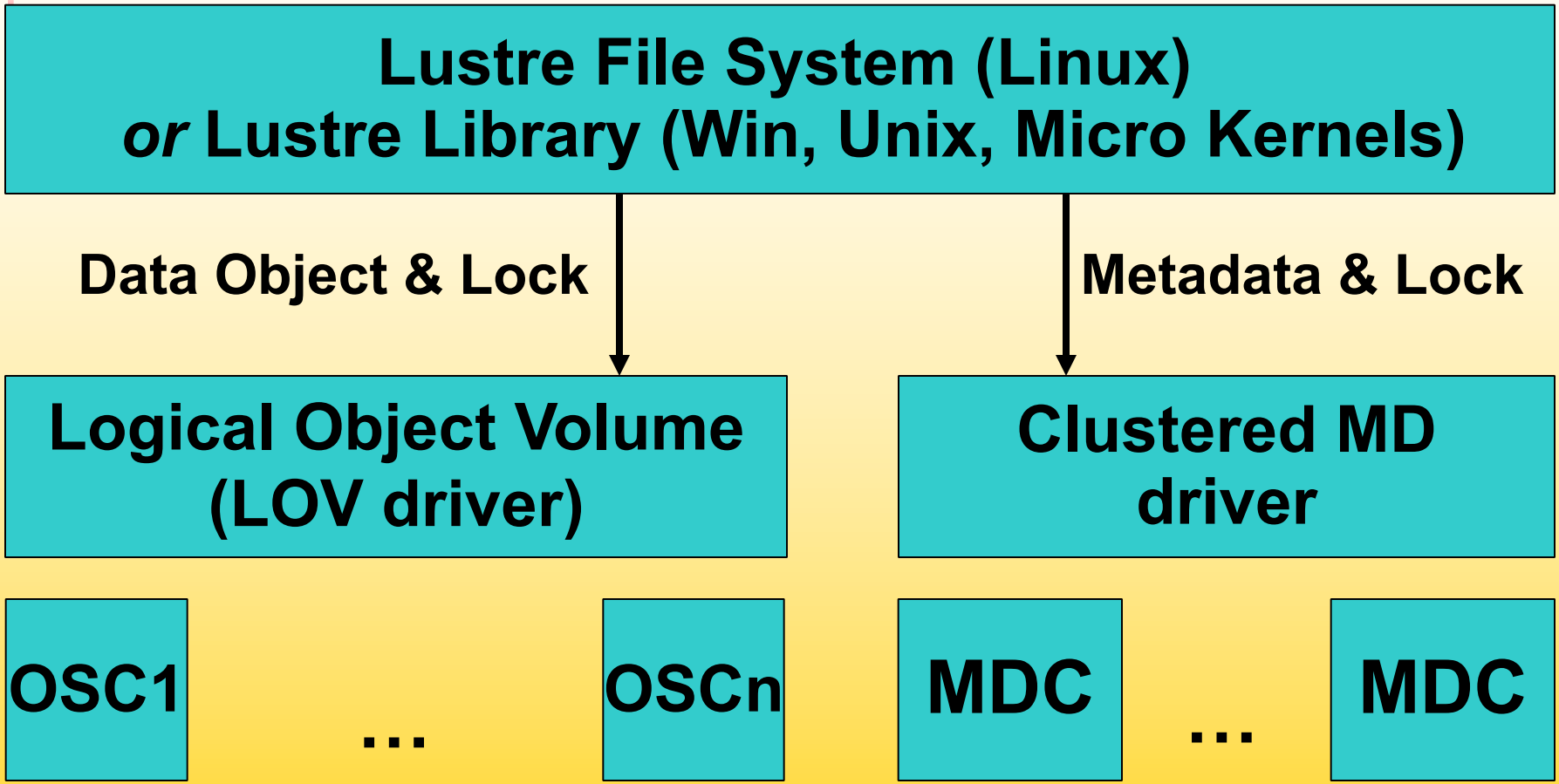
Network Abstraction Layer for
TCP, QSW, etc. Small & hard
Includes routing api.

Devices and API's

Lustre Devices & API's

- Lustre has numerous driver modules
  - One API - very different implementations
  - Driver binds to named device
  - Stacking devices is key
  - Generalized "object devices"
- Drivers currently export several API's
  - Infrastructure - a mandatory API
  - Object Storage
  - Metadata Handling
  - Locking
  - Recovery

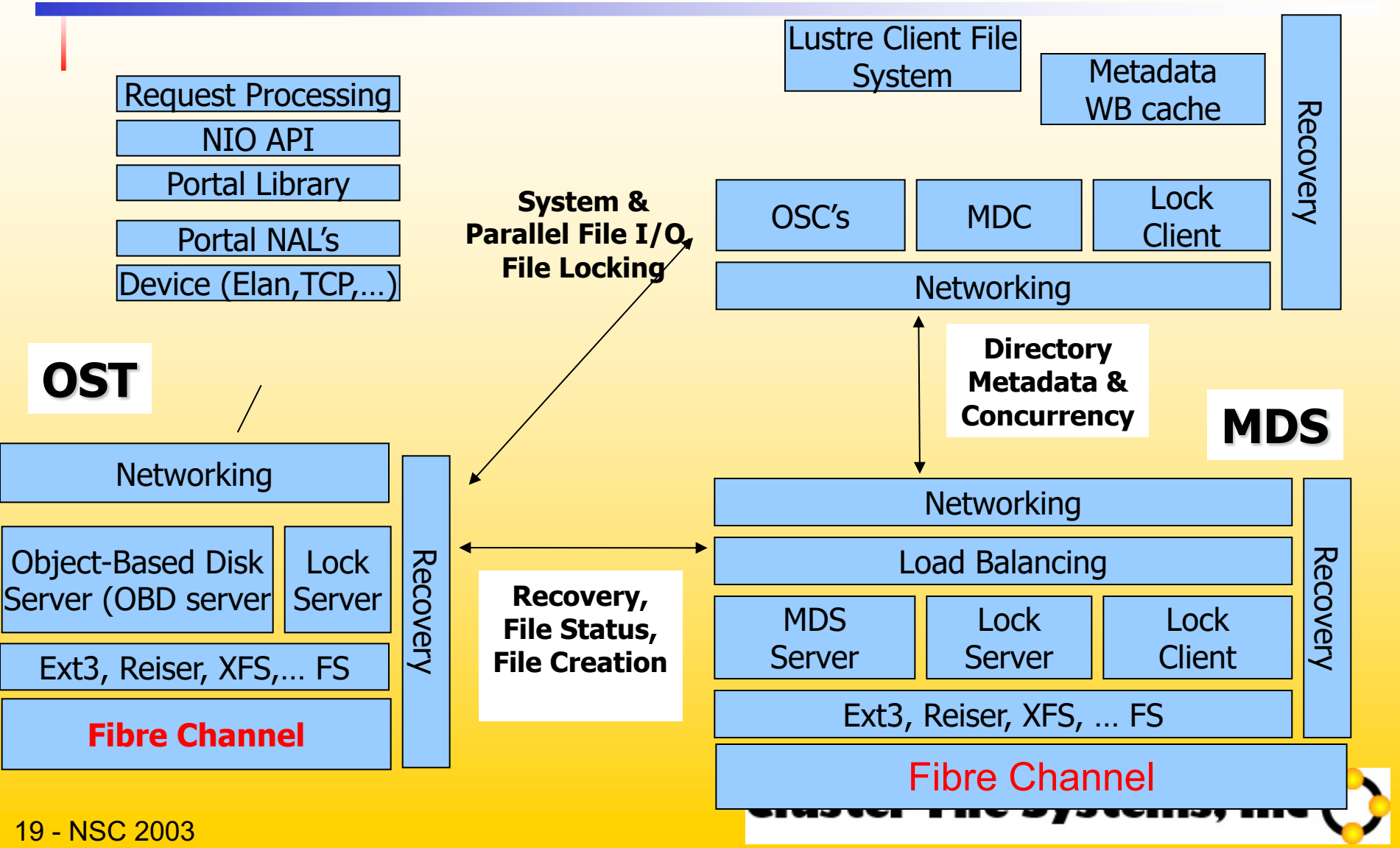
Lustre Clients & API's



Object Storage Api

- Objects are (usually) unnamed files
- Improves on the block device api
  - create, destroy, setattr, getattr, read, write
- OBD driver does block/extent allocation
- Implementation:
  - Linux drivers, using a file system backend

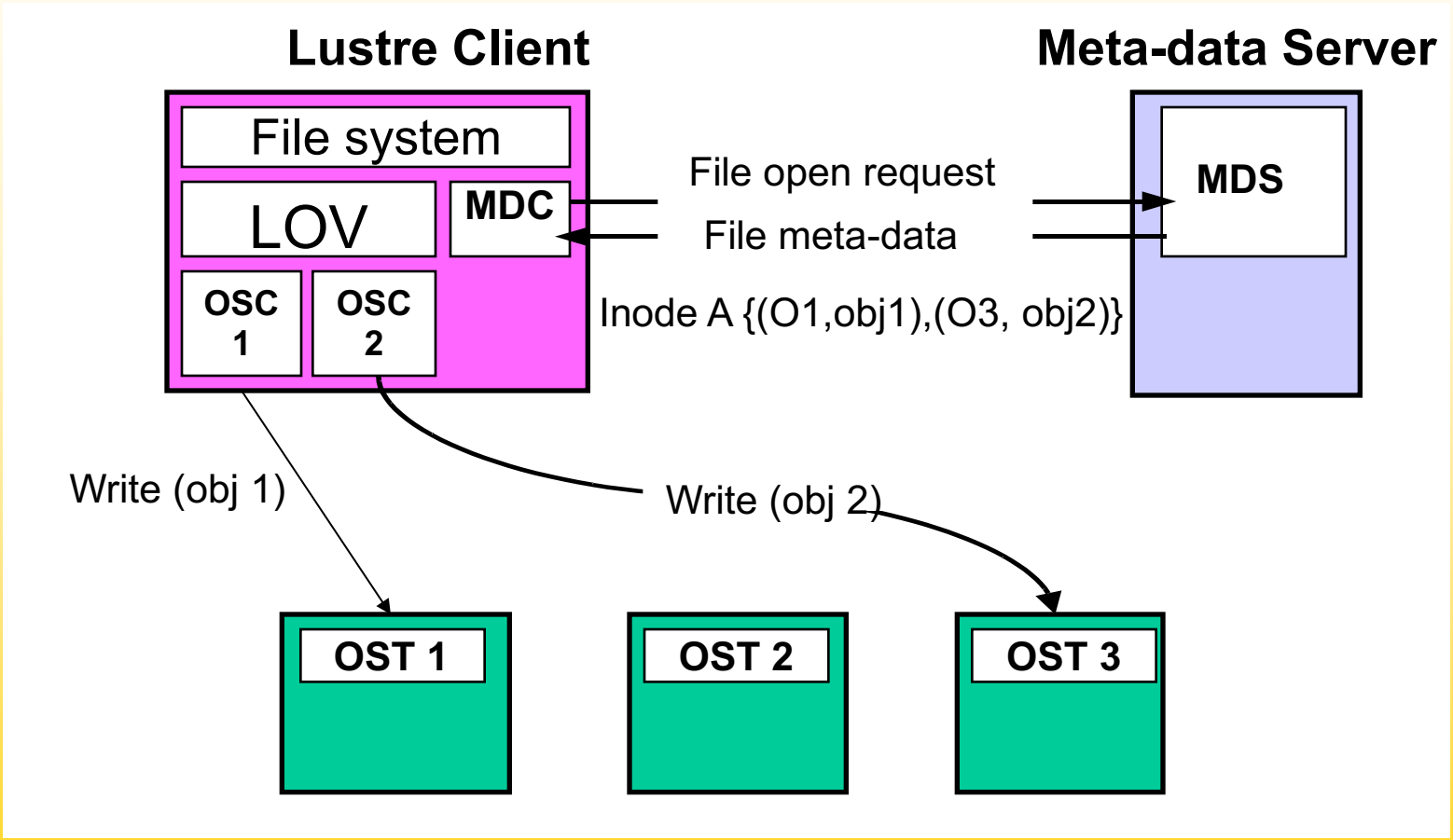
Bringing it all together



File I/O

File I/O – Write Operation

- Open file on meta-data server
- Get information on all objects that are part of file:
  - Objects id's
  - What storage controllers (OST)
  - What part of the file (offset)
  - Striping pattern
- Create LOV, OSC drivers
- Use connection to OST
  - Object writes to OST
  - No MDS involvement at all



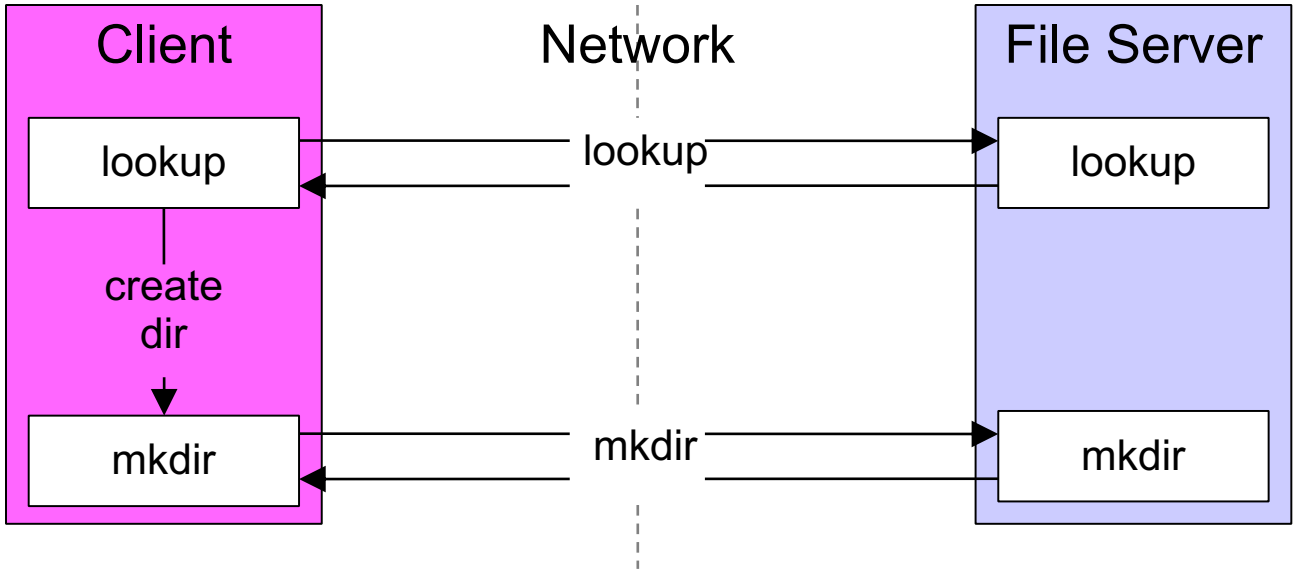
I/O bandwidth

- 100's GB/sec => saturate many100's OSTs
- OST's:
  - Do ext3 extent allocation, non-caching direct I/O
  - Lock management spread over cluster
- Achieve 90-95% of network throughput
  - Single client, single thread Elan3: W 269MB/sec
  - OST's handle up to 260MB/sec
  - W/O extent code, on 2 way 2.4GHz Xeon

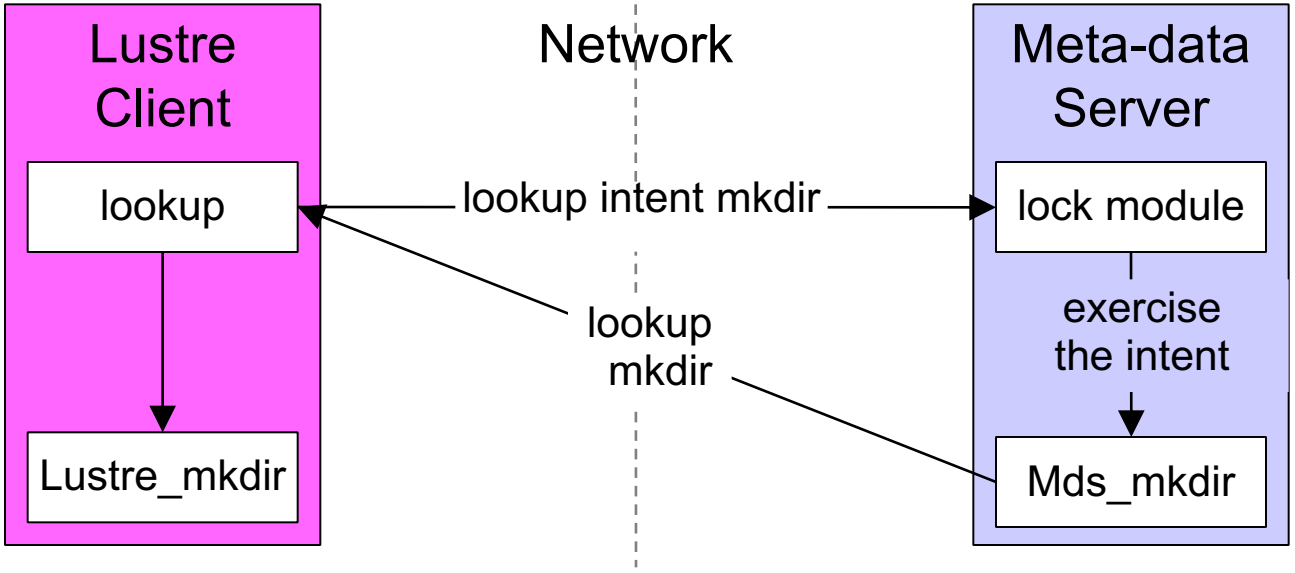
Metadata

Intent locks & Write Back caching

- Clients – MDS: protocol adaptation
- Low concurrency - write back caching
  - Client in memory updates
  - delayed replay to MDS
- High concurrency (mostly merged in 2.6)
  - Single network request per transaction
  - No lock revocations to clients
  - Intent based lock includes complete request



a) Conventional mkdir



b) Lustre mkdir

Lustre 1.0

- Only has high concurrency model
- Aggregate throughput (1,000 clients):
  - Achieve ~5000 file creations (open/close) /sec
  - Achieve ~7800 stat's in 10 x1M file directories
- Single client:
  - Around 1500 creations or stat's /sec
- Handling 10M file directories is effortless
- Many changes to ext3 (all merged in 2.6)

Metadata Future

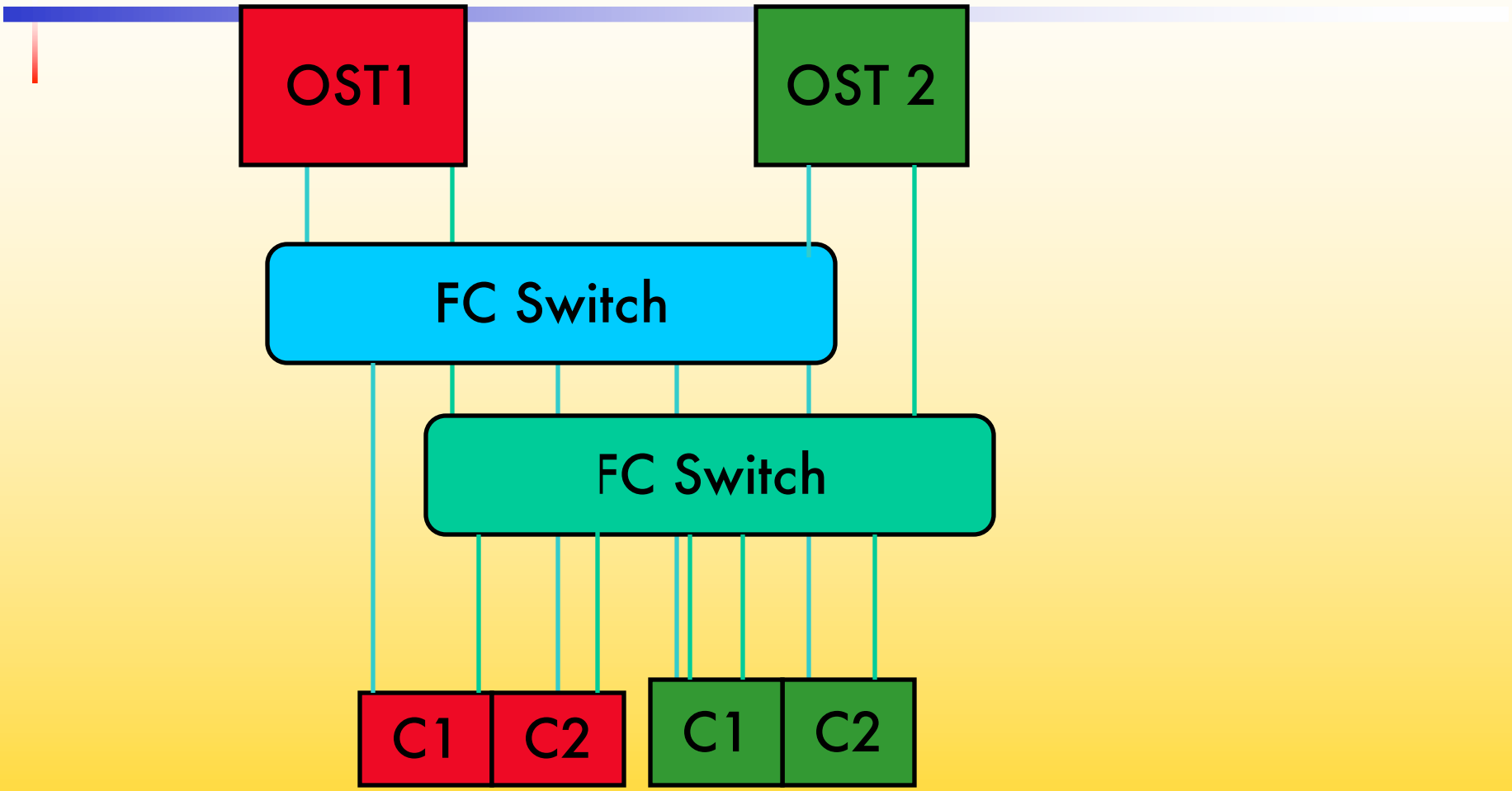
- Lustre 2.0 – 2004
- Metadata clustering
  - Common operations will parallelize
- 100% WB caching in memory or on disk
  - Like AFS

Recovery

Recovery approach

- Keep it simple!
- Based on failover circles:
  - Use existing failover software
  - Left working neighbor is failover node for you
- At HP we use failover pairs
  - Simplify storage connectivity
- I/O failure triggers
  - Peer node serves failed OST
  - Retry from client routed to new OST node

OST Server – redundant pair



Configuration

Lustre 1.0

- Good tools to build configuration
- Configuration is recorded on MDS
  - Or on dedicated management server
  - Configuration can be changed,
    - 1.0 requires downtime
- Clients auto configure
  - mount –t lustre –o … mds://fileset/sub/dir /mnt/pt
- SNMP support

Futures

Advanced Management

- Snapshots
  - All features you might expect
- Global namespace
  - Combine best of AFS & autofs4
- HSM, hot migration
  - Driven by customer demand (we plan XDSM)
- Online 0-downtime re-configuration
  - Part of Lustre 2.0

Security

- Authentication
- POSIX style authorization
- NASD style OST authorization
  - Refinement: use OST ACL's and cookies
- File crypting with group key service
  - STK secure file system

Project status

Lustre Feature Roadmap

Lustre (Lite) 1.0 (Linux 2.4 & 2.6)	Lustre 2.0 (2.6)	Lustre 3.0
2003	2004	2005
Failover MDS	Metadata cluster	Metadata cluster
Basic Unix security	Basic Unix security	Advanced Security
File I/O very fast (~100's OST's)	Collaborative read cache	Storage management
Intent based scalable metadata	Write back metadata	Load balanced MD
POSIX compliant	Parallel I/O	Global namespace

Cluster File Systems, Inc.

Cluster File Systems

- Small service company: 20-30 people
  - Software development & service (95% Lustre)
  - contract work for Government labs
  - OSS but defense contracts
- Extremely specialized and extreme expertise
  - we only do file systems and storage
- Investments - not needed. Profitable.
- Partners: HP, Dell, DDN, Cray

Cluster File Systems, Inc

Lustre – conclusions

- Great vehicle for advanced storage software
  - Things are done differently
  - Protocols & design from Coda & InterMezzo
  - Stacking & DB recovery theory applied
- Leverage existing components
- Initial signs promising

HP & Lustre

- Two projects
  - ASCI PathForward – Hendrix
  - Lustre Storage product
    - Field trial in Q1 of 04

Questions?