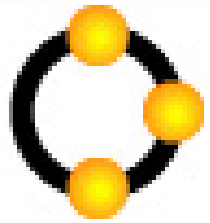
Lustre: the Intergalactic File System for the National Labs?

Peter J. Braam

braam@clusterfs.com

http://www.clusterfilesystems.com

Cluster File Systems, Inc



Cluster File Systems, Inc…

- Goals
  - Consulting & development
  - Storage and file systems
  - Open source
  - Extreme level of expertise
- Leading
  - InterMezzo – high availability file system
  - Lustre – next generation cluster file system
  - Important role in Coda, UDF and Ext3 for Linux

Partners...

- CMU
- National Labs
- Intel

Cluster File Systems, Inc

Talk overview

- Trends
- Next generation data centers
- Key issues in cluster file systems
- Lustre
- Discussion

Trends…

Cluster File Systems, Inc

Hot animals…

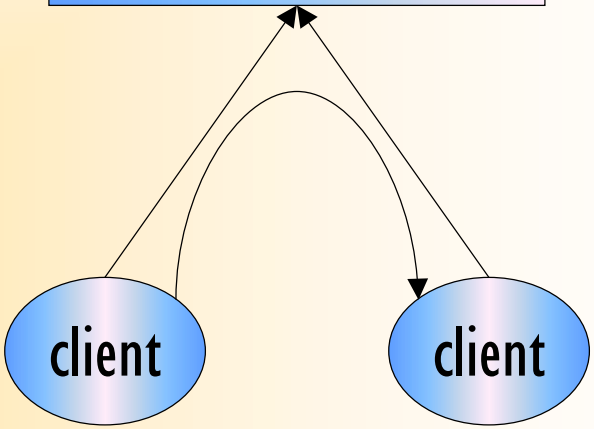
- NAS
  - Cheap servers deliver a lot of features
- NFS v4
  - Finally, NFS is getting it right…
  - Security, concurrency
- DAFS
  - High level storage protocol over VI storage network
- Open Source OS
  - Best of breed file systems (XFS, JFS, Reiser)

DAFS / NFS v4

DAFS server

high level
fast & efficient
storage protocol

NFS v4 server



concurrency control
with notifications to clients

Cluster File Systems, Inc

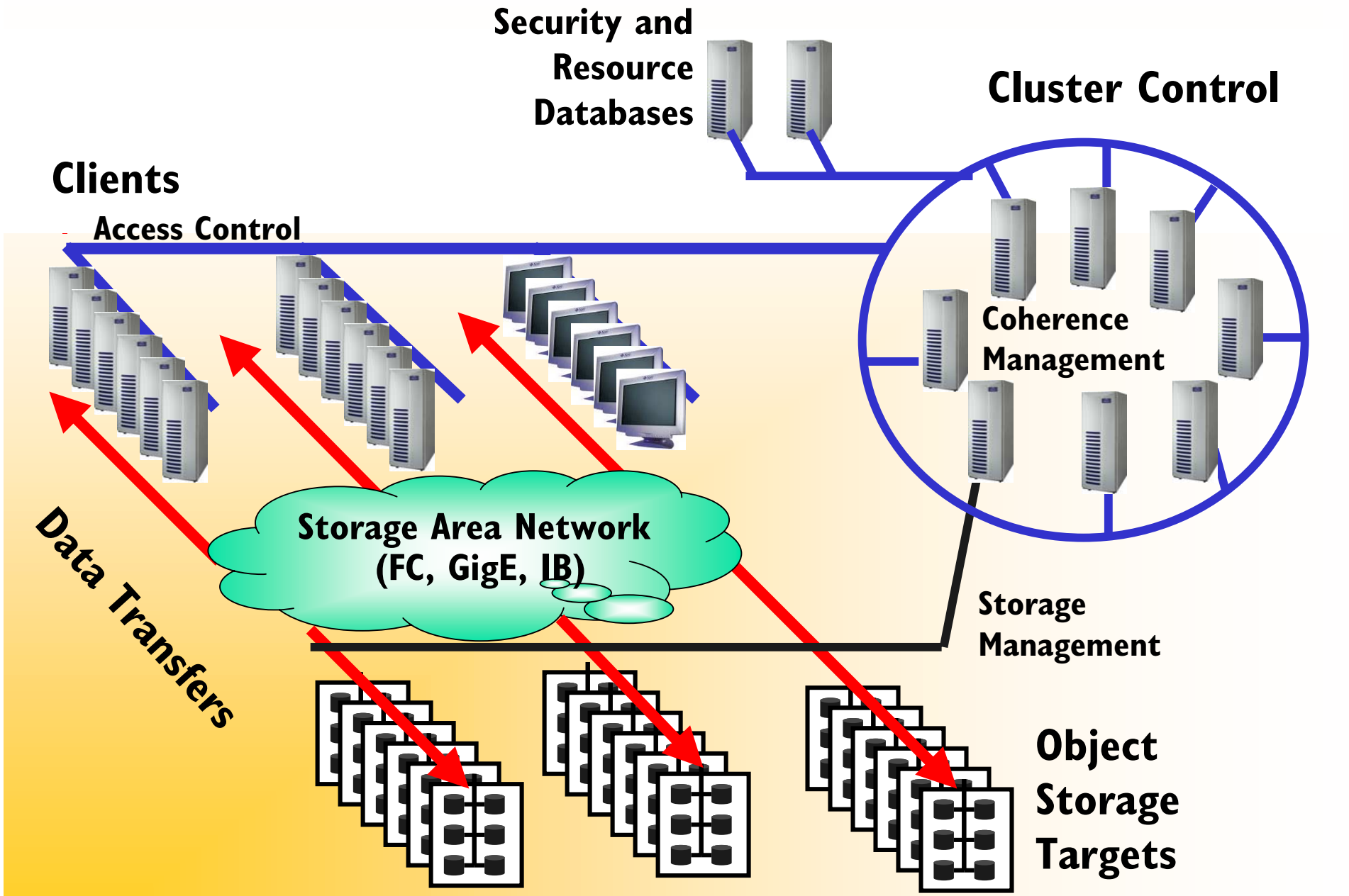
Key question

- How to use the parts…

Cluster File Systems, Inc

Next generation data centers

Cluster File Systems, Inc


Security and
Resource
Databases
Cluster Control
Clients
Access Control
Coherence
Management
Storage Area Network
(FC, GigE, IB)
Data Transfers
Storage
Management
Object
Storage
Targets

Orders of magnitude

- Clients (aka compute servers)
  - 10,000's
- Storage controllers
  - 1000's to control PB's of storage (PB = 10**15 Bytes)
- Cluster control nodes
  - 10's
- Aggregate bandwidth
  - 100's GB/sec

Cluster File Systems, Inc

Applications

- Scientific computing
- Bio Informatics
- Rich media
- Entire ISP's

Key issues

Cluster File Systems, Inc

Scalability

- I/O throughput
  - How to avoid bottlenecks
- Meta data scalability
  - How can 10,000's of nodes work on files in same folder
- Cluster recovery
  - If something fails, how can transparent recovery happen
- Management
  - Adding, removing, replacing, systems; data migration & backup

Features

- The basics…
  - Recovery, management, security
- The desired…
  - Gene computations on storage controllers
  - Data mining for free
  - Content based security
  - …
- The obstacle…
  - An almost 30 year old pervasive block device protocol

Look back

- Andrew Project at CMU
  - 80's – file servers with 10,000 clients (CMU campus)
  - Key question: how to reduce foot print of client on server
  - By 1988 entire campus on AFS
- Lustre
  - Scalable clusters?
  - How to reduce cluster footprint of shared resources (scalability)
  - How to subdivide bottlenecked resources (parallelism)

Lustre

Intelligent Object Storage
http://www.lustre.org

What is Object Based Storage?

- Object Based Storage Device
  - More intelligent than block device
- Speak storage at "inode level"
  - create, unlink, read, write, getattr, setattr
  - iterators, security, almost arbitrary processing
- So…
  - Protocol allocates physical blocks, no names for files
- Requires
  - Management & security infrastructure

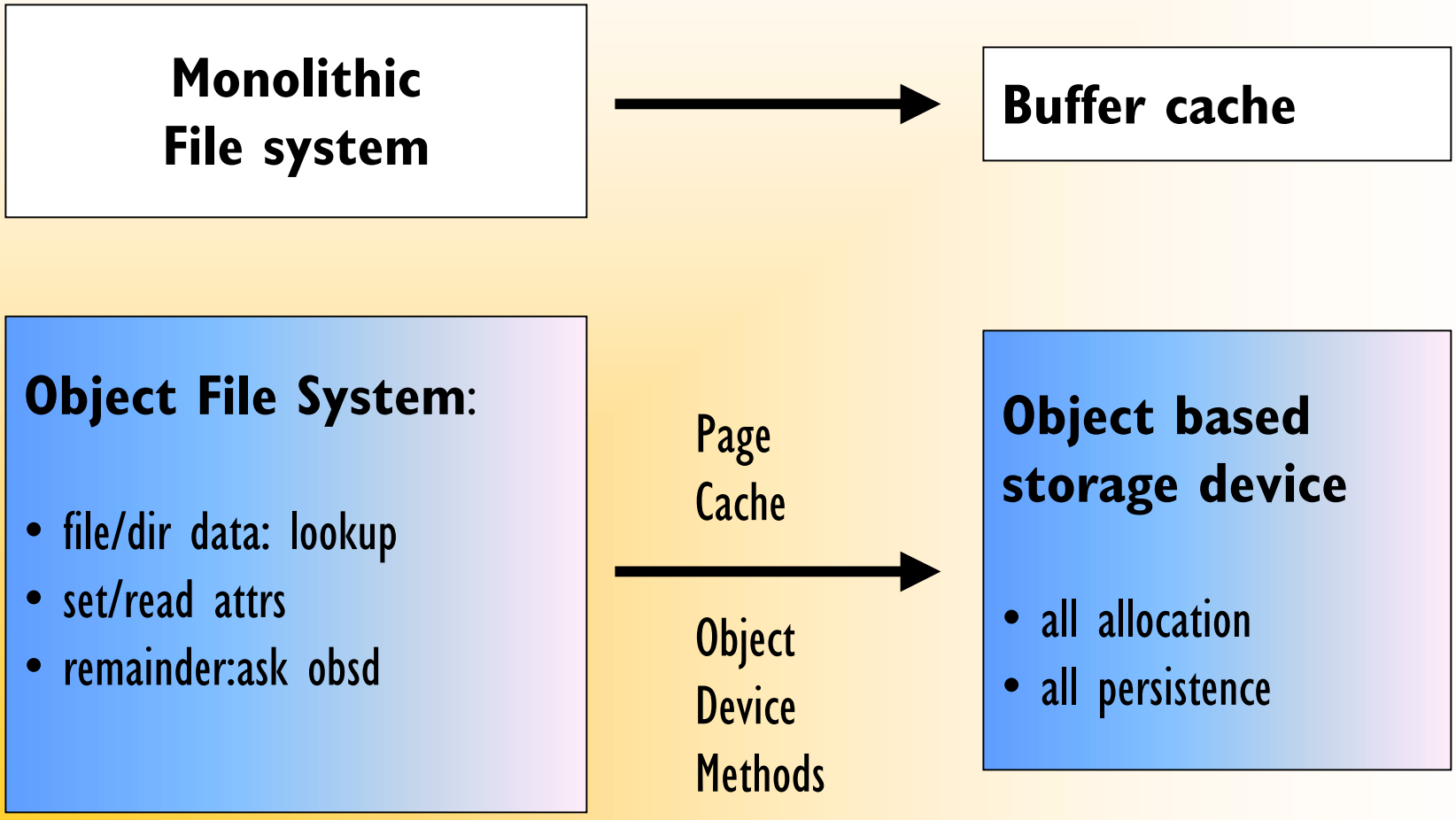
Project history

- Started between CMU – Seagate – Stelias Computing
  - Another road to NASD style storage
  - NASD now at Panasas – originated many ideas
- Los Alamos
  - More research
  - Nearly built little object storage controllers
  - Currently looking at genomics applications
- Sandia, Tri-Labs
  - Can Lustre meet the SGS-FS requirements?

Components of OB Storage

- Storage Object Device Drivers
  - class drivers – attach driver to interface
    - **Targets, clients** – remote access
    - **Direct drivers** – to manage physical storage
    - **Logical drivers** – for intelligence & storage management
- Object storage applications:
  - (cluster) file systems
  - Advanced storage: parallel I/O, snapshots
  - Specialized apps: caches, db's, filesrv

Object File System



Cluster File Systems, Inc

Accessing objects

- Session
  - connect to the object storage target, present security token
- Mention object id
  - Objects have a unique (group,id)
- Perform operation
- So that's what the object file system does!

Objects may be files, or not…

- Common case:
  - Object, like inode, represents a file
- Object can also:
  - represent a stripe (RAID)
  - bind an (MPI) File_View
  - redirect to other objects

Snapshots as logical module

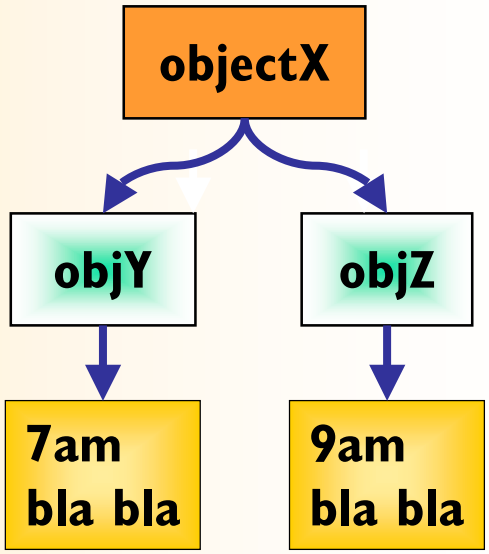
Present multiple views of file systems

Object File System

/current/
/yesterday/

Versioned objects: follow redirectors

Snapshot / versioning logical driver

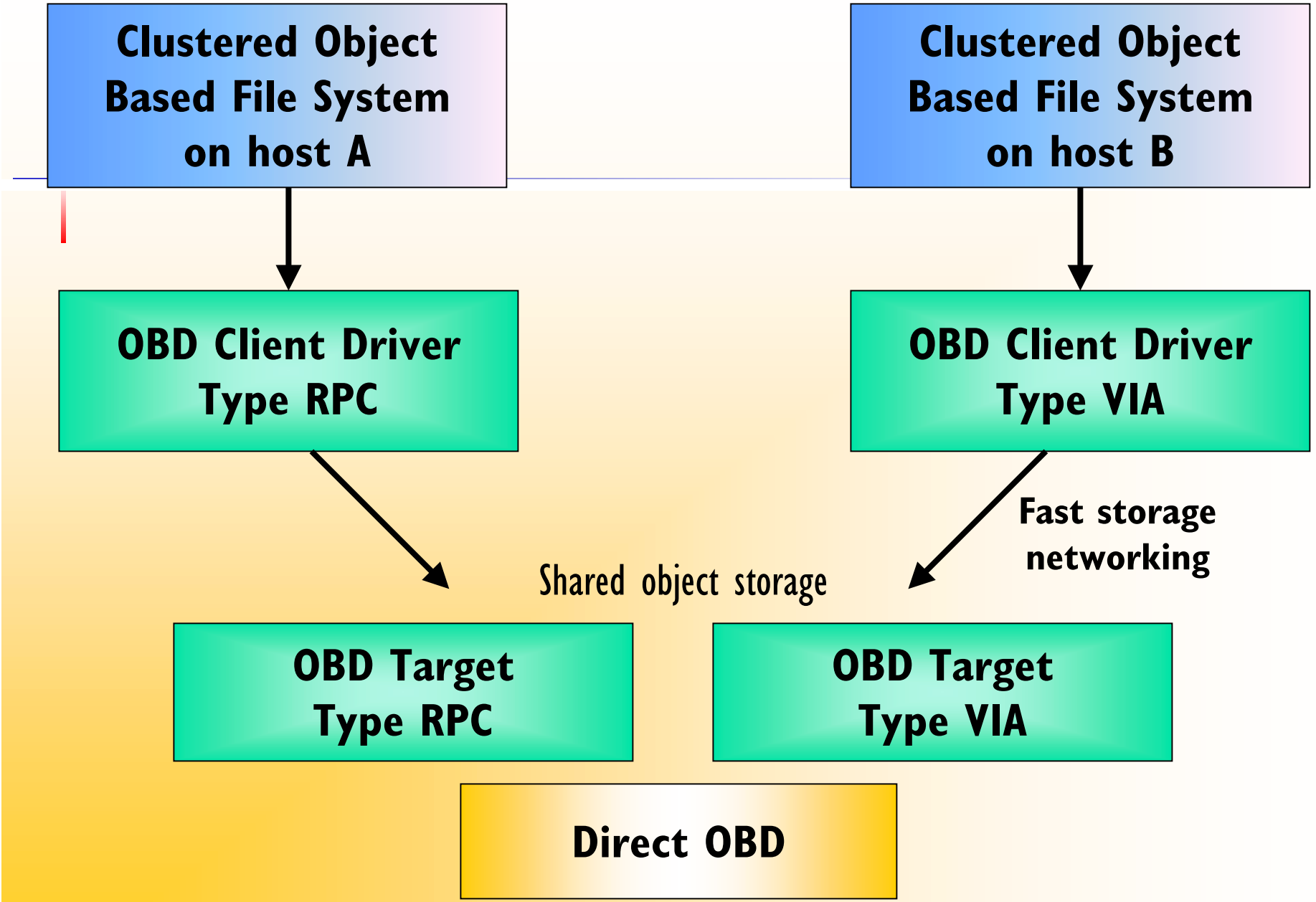


Access to raw objects

Direct object driver

System Interface

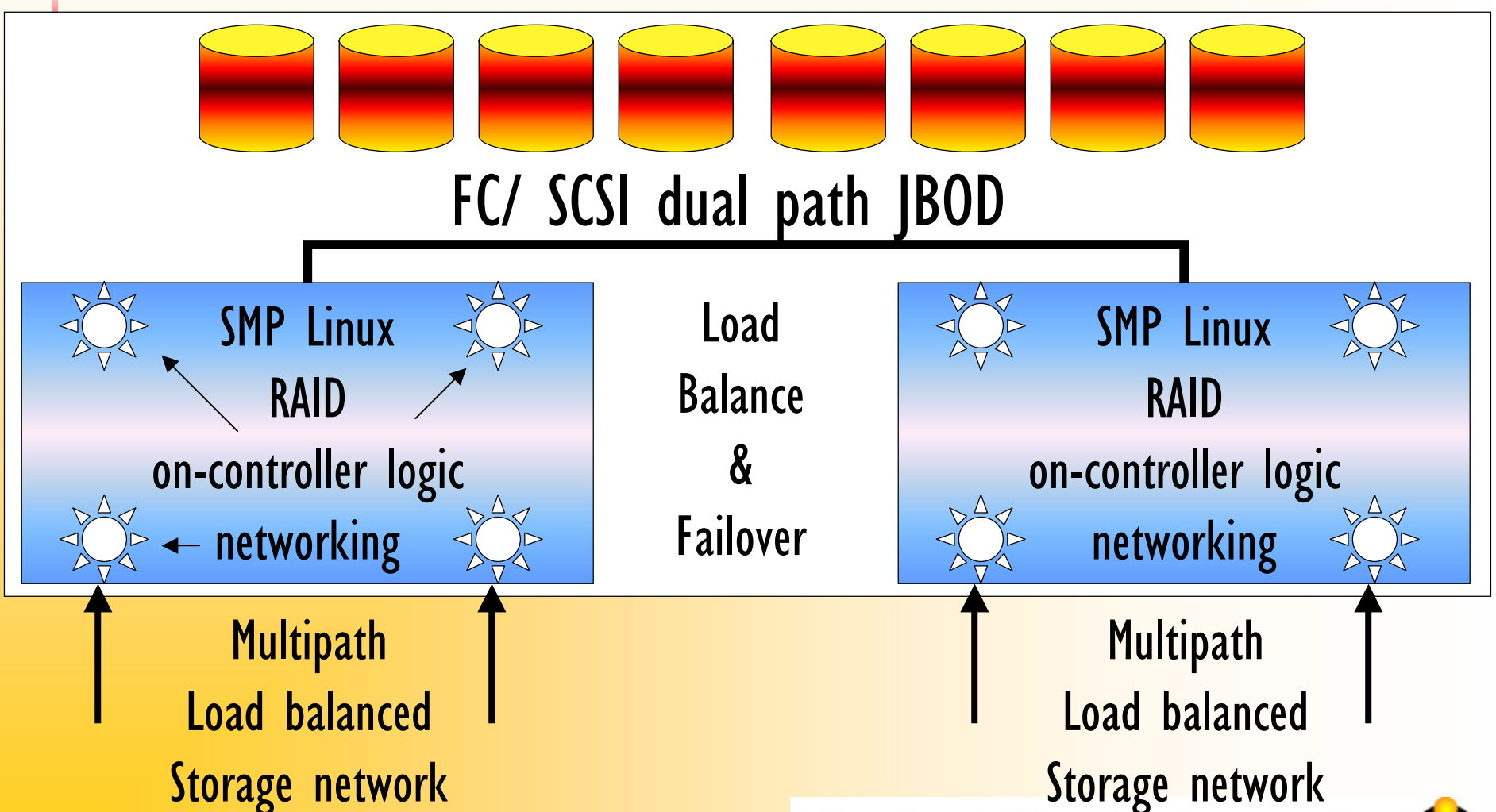
- Modules
  - Load the kernel modules to get **drivers** of a certain **type**
  - Name **devices** to be of a certain type
  - Build **stacks** of devices with assigned types
- For example:
  - insmod obd_xfs ; obdcontrol dev=obd1,type=xfs
  - insmod obd_snap ; obdcontrol current=obd2,old=obd3,driver=obd1
  - insmod obdfs ; mount –t obdfs –o dev=obd3 /mnt/old


Clustered Object Based File System on host A
Clustered Object Based File System on host B
OBD Client Driver Type RPC
OBD Client Driver Type VIA
Fast storage networking
Shared object storage
OBD Target Type RPC
OBD Target Type VIA
Direct OBD

Storage target Implementations

- Classical storage targets
  - Controller — expensive, redundant, proprietary
    - EMC: as sophisticated & feature rich as block storage can get
  - A bunch of disks
- Lustre target
  - Bunch of disks
  - Powerful (SMP, multiple busses) commodity PC
  - **Programmable/Secure**
- Could be done on disk drives but…

Inside the storage controller…

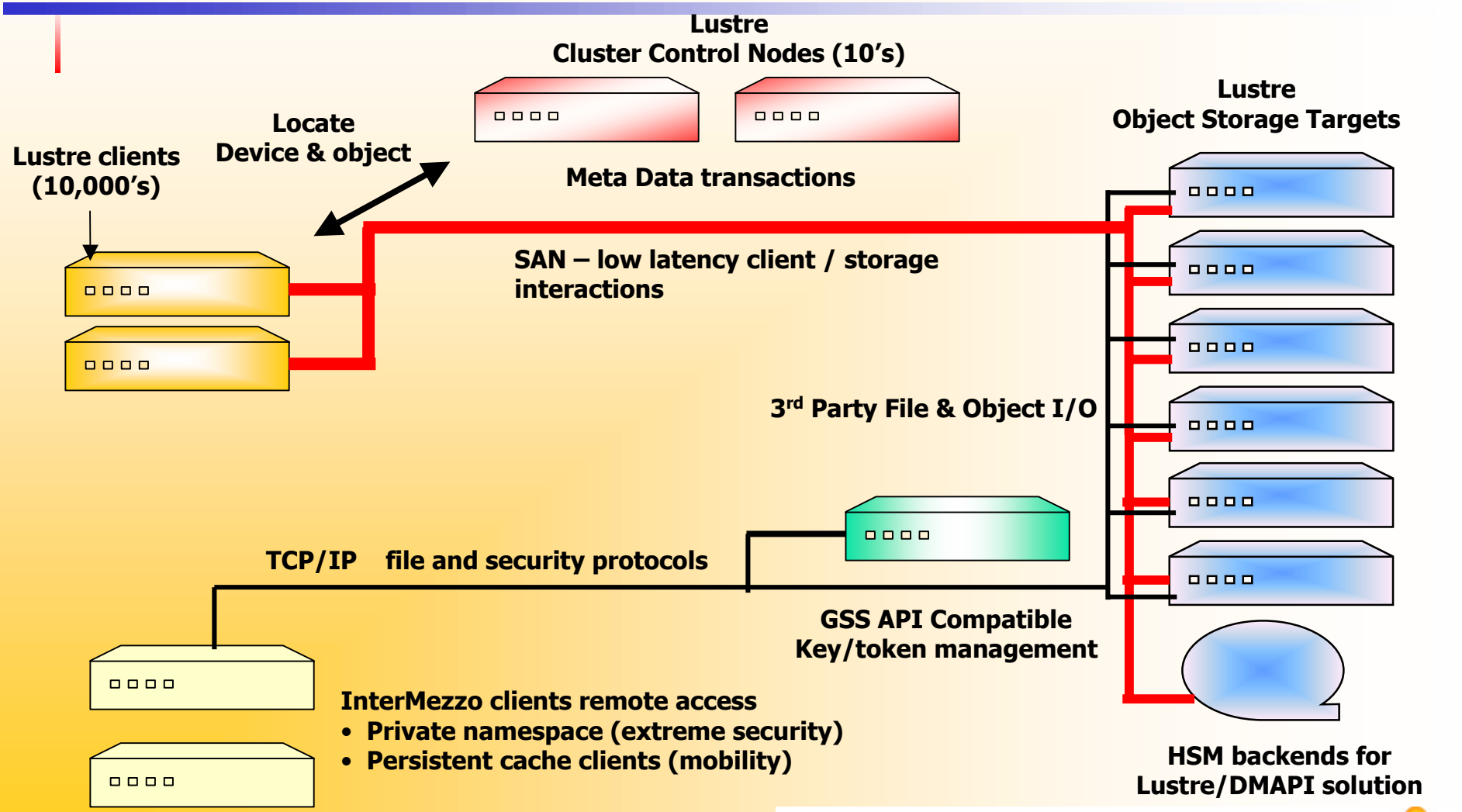


Cluster File Systems, Inc

Objects in clusters…

Cluster File Systems, Inc

Lustre clusters

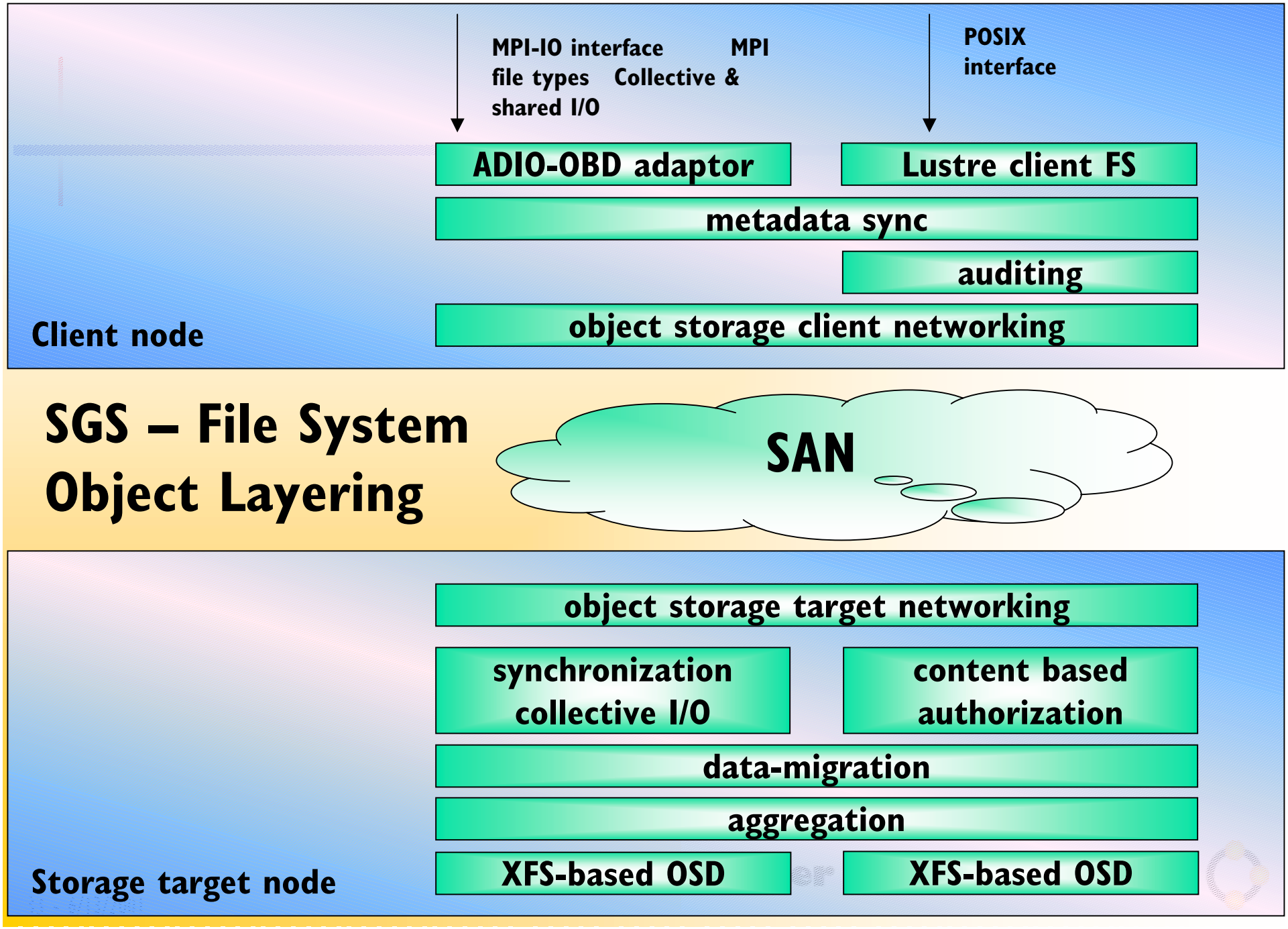


Cluster control nodes

- Database of references to objects
- E.g. Lustre File System
  - Directory data
  - Points to objects that contain stripes/extents of files
- More generally
  - Use a database of references to objects
  - Write object applications that access the objects directly
    - LANL asked for gene processing on the controllers

Examples of logical modules

- Tri-Lab/NSA: SGS – File system (see next slide)
  - Storage management, security
  - Parallel I/O for scientific computation
- Other requests:
  - Data mining while target is idle
  - LANL: gene sequencing in object modules
  - Rich media industry: prioritize video streams



SGS – File System Object Layering

Other applications…

- Genomics
  - Can we reproduce the previous slide?
- Data mining
  - Can we exploit idle cycles and disk geometry?
- Rich media
  - What storage networking helps streaming?

Why Lustre…

- It's fun, it's new
  - Infrastructure for storage target based computing
- Storage management: components – not monolithic
  - File system snapshots, raid, backup, hot migration, resizing
  - Much simpler
- File System:
  - Clustering FS considerably simpler, more scalable
  - But: close to NFS v4 and DAFS in several critical ways

And finally – the reality, what exists…

- At http://www.lustre.org (everything GPL'd)
  - Prototypes:
    - Direct driver for ext2 objects, Snapshot logical driver,
    - Management infrastructure, Object file system
- Current happenings:
  - Collaboration with Intel Enterprise Architecture LAB:
    - They are building Lustre storage networking (DAFS, RDMA, TUX)
  - The grand scheme of things has been planned and is moving
- Also on the WWW:
  - OBD storage specification
  - Lustre SGS – File System implementation plan.

Linux clusters

Cluster File Systems, Inc

Clusters - purpose

Require:

- A scalable almost single system image
- Fail-over capability
- Load-balanced redundant services
- Smooth administration

Cluster File Systems, Inc

Ultimate Goal

- Provide generic components
- OPEN SOURCE
- Inspiration: VMS VAX Clusters
- New:
  - Scalable (100,000's nodes)
  - Modular
- Need distributed, cluster & parallel FS's
  - InterMezzo, GFS/Lustre, POBIO-FS

Technology Overview

Modularized VAX cluster architecture (Tweedie)

Core	Support	Clients
Transition	Cluster db	Distr. Computing
Integrity	Quorum	Cluster Admin/Apps
Link Layer	Barrier Svc	Cluster FS & LVM
Channel Layer	Event system	DLM

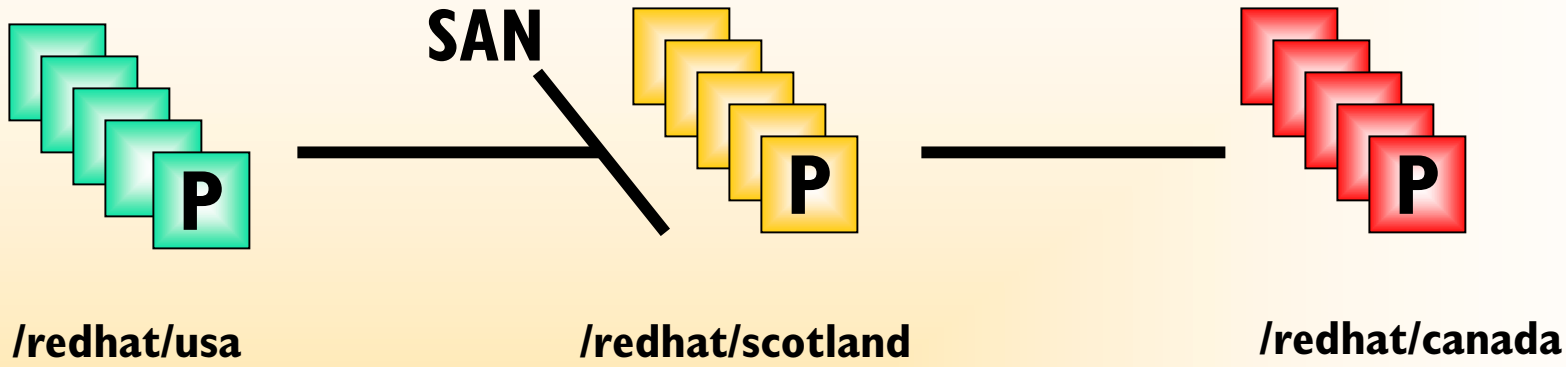
Cluster File Systems, Inc

Events

- Cluster transition:
  - Whenever connectivity changes
  - Start by electing "cluster controller"
  - Only merge fully connected sub-clusters
  - Cluster id: counts "incarnations"
- Barriers:
  - Distributed synchronization points
- Partial implementations available:
  - Ensemble, KimberLite, IBM-DLM, Compaq Cluster Mgr

Cluster File Systems, Inc

Scalability – e.g. Red Hat cluster



- **P = peer**
  - Proxy for remote core cluster
  - Involved in recovery
- **Communication**
  - Point to point within core clusters
  - Routable within cluster
  - Hierarchical flood-fill
- **File Service**
  - Cluster FS within cluster
  - Clustered Samba/Coda etc
- **Other stuff**
  - Membership / recovery
  - DLM / barrier service
  - Cluster admin tools

InterMezzo

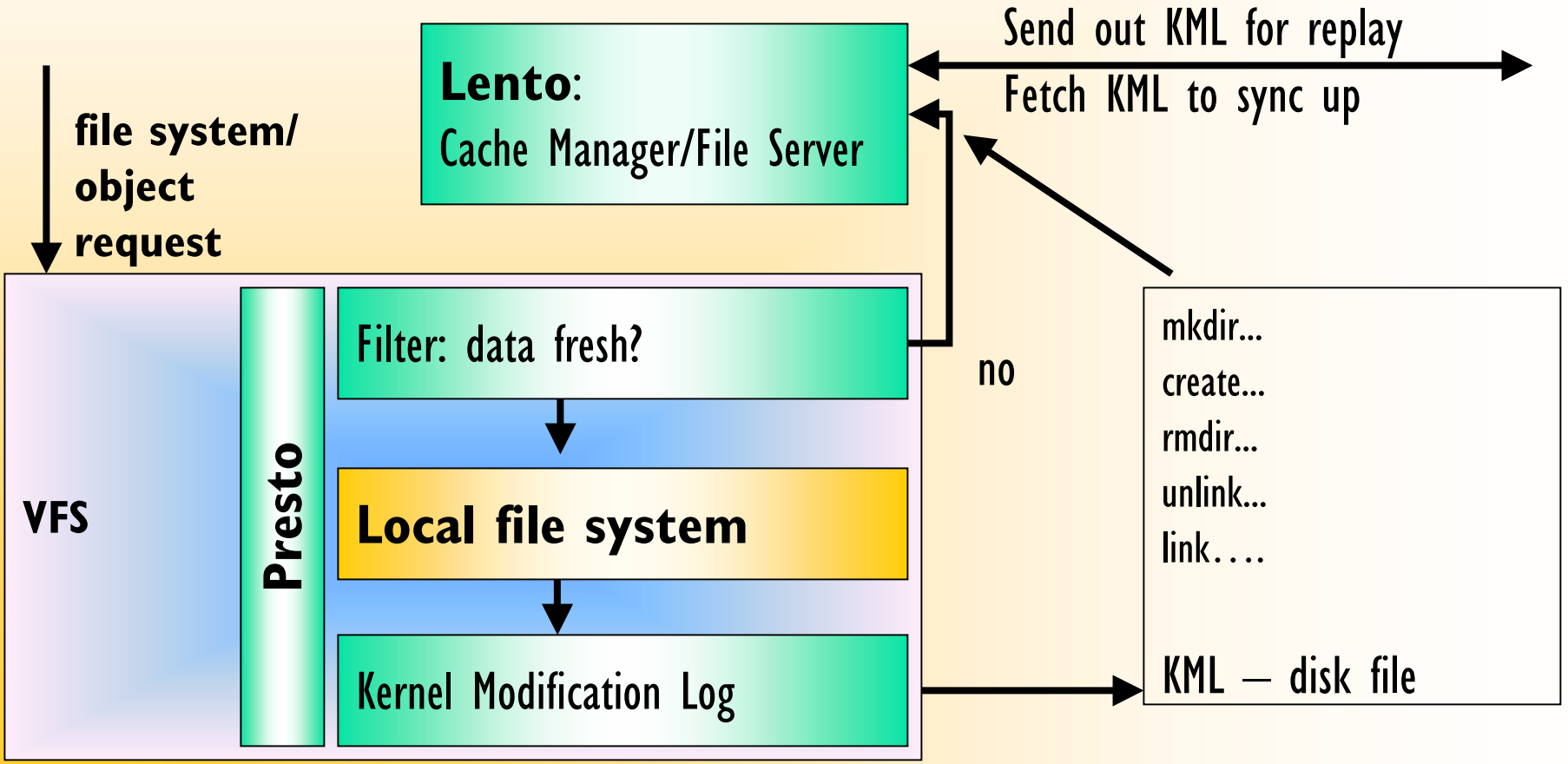
http://www.inter-mezzo.org

Cluster File Systems, Inc

Target

- Replicate or cache directories
  - Automatic synchronization
  - Disconnected operation
  - Proxy servers
  - Scalable
- Purpose
  - Entire System Binaries, laptop/desktop
  - Redundant object storage controllers
- Very simple
  - Coda style protocols
  - Wrap around local file systems as cache

Basic InterMezzo



Cluster File Systems, Inc

Distributed Lock Manager

IBM released HACMP DLM

Open Source/ VAX style

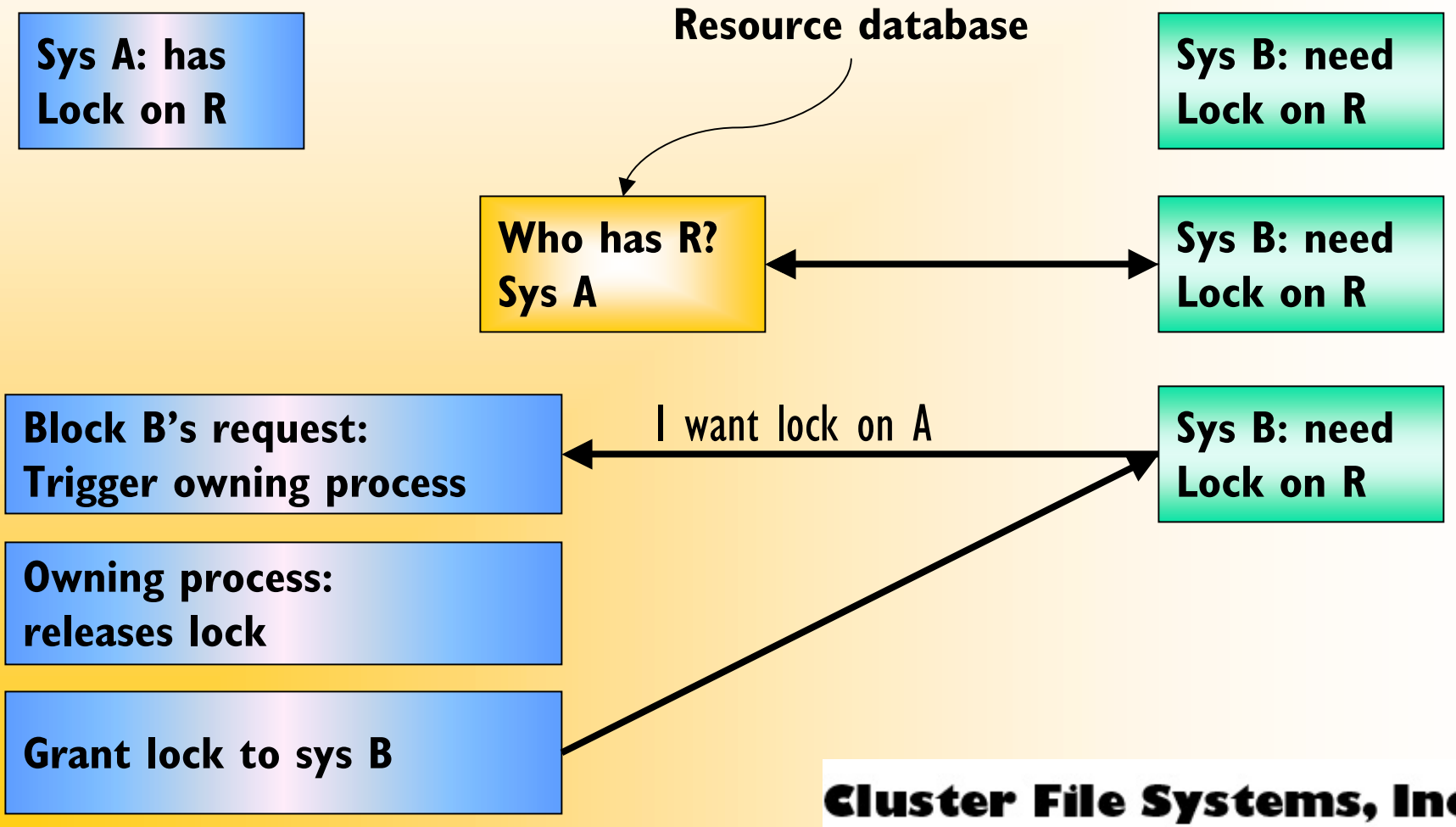
http://www.ibm.com/developerworks/open/source

Locks & resources

- Purpose: generic, rich lock service
- Will subsume “callbacks”, “leases” etc.

- Lock resources: resource database
  - Organize resources in trees
  - Most lock traffic is local
- High performance
  - node that acquires resource manages tree

Typical simple lock sequence



Cluster File Systems, Inc

A few details…

- Six lock modes
  - Acquisition of locks
  - Promotion of locks
  - Compatibility of locks
- First lock acquisition
  - Holder will manage resource tree
- Remotely managed
  - Keep copy at owner
- Callbacks:
  - On blocking requests
  - On release, acquisition
- Recovery (simplified):
  - Dead node was:
    - Mastering resources
    - Owning locks
  - Re-master rsrc
  - Drop zombie locks

Cluster File Systems, Inc