Lustre High availability configuration in CETA-CIEMAT

Author: Alfonso Pardo Díaz
Event: LAD'2012
Place / Date: Paris, 25/09/2012

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS
CETA Ciemat

FEDER
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa

INDEX

Lustre HA in CETA-Ciemat

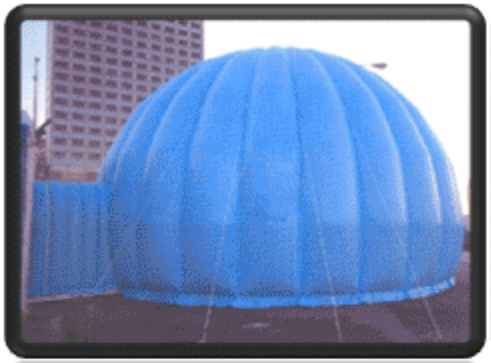
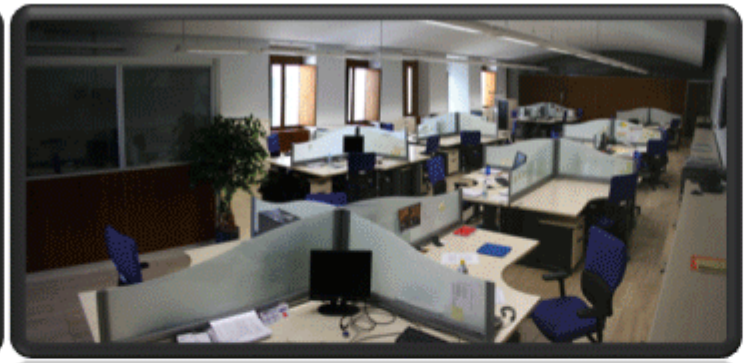
1 Who are we?

CIEMAT data center (MICINN), joint iniciative with regional government of Extremadura

- **Public** institution, financed by PGE & FEDER
- **Mission**: **Consolidate and disseminate eScience and ITs, specially Grid and eInfrastructures**
- Offer our resources: Grid, Cloud, and HPC (GPUs, clusters, …)
- Contribute to the effective expansion of eScience
- Facilitate usage of resources (new sites, new applications)

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

1 Who are we?







2 Lustre in CETA-CIEMAT

- Just now updating from version 1.8.4 to version 2.1.2!
- 3 different storage machines on 1 metadata server (heterogeneous environment)
- Separated MDS-MGS
- Tape library for backup or HSM (Tivoli Storage Manager? RobinHood?)
- Different clients: CentOS, Debian, RedHat, Scientific Linux,...
- Ethernet (at least for now!)

ÍNDICE

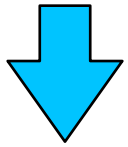
2 Lustre in CETA-CIEMAT

- **Design concerns: our MDS/MGS/MDT**
  - **MDS/MGS: 2 IBM x336 MDS/MGS in active/passive (HA).**
    - 16GB RAM (yes, I know, little ram!)
    - 2x3GHz Xeon CPU
    - 2GBit Ethernet lacp bonding
  - **MDT: IBM DS4100 for metadata target**
    - 2Gbit Fibre Channel connection from MDT to MDS
    - RAID 5, one LUN per filesystem
    - 1 Hot spare for RAID

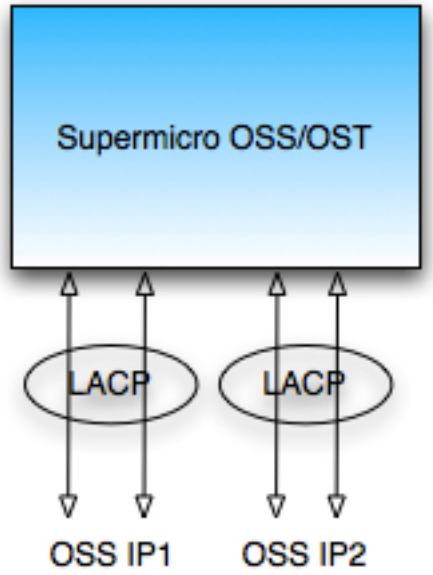
ÍNDICE

Design concerns: our OSS/OST

- 10 Supermicro as OSS/OST
- 24 TB RAW => 17 TB Lustre
- 2 RAID 6 per OST
- Hot spare for healthy RAID
- 8 GB RAM, 2x2,5Gz Xeon CPU
- 4x1Gb Ethernet interfaces = 2Gbit Ethernet lacp bonding and active/passive failover bonding
- How to reach 2 different bonding interfaces?



- Two different IP interfaces
- Second IP bonding is an Active or Passive OSS of the first interface



2 Lustre in CETA-CIEMAT

- **Design concerns: our OSS/OST**
- **5 IBM x336 MDS in active/passive**
- 4GB RAM
- 2x3GHz Xeon CPU
- 2GBit Ethernet lacp bonding
- **10 IBM DS4100+exp110 as OST**
- 2Gbit Fibre Channel connections from OST to OSS
- RAID 5 per OST
- Hot spare for RAID

ÍNDICE

3 High availability issues

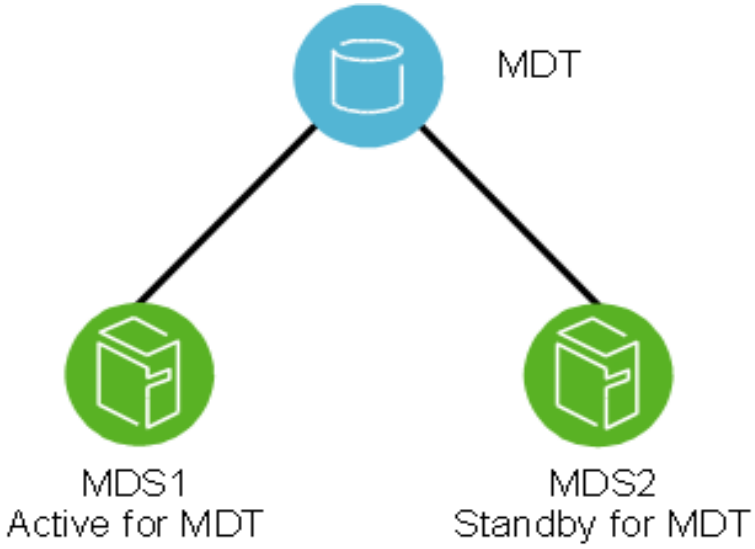
- **What happens if a MDS or MGS server fails?**
- Second server is in passive mode.
- Pacemaker: MDT mounted in only one MDS server. It manages where must be the active service.

- Clients are sensitive to MDS errors by timeout.

Clients have a list of possible MDS servers

MDS active server enter in “RECOVERY TIME”

FILESYSTEM STUCK DURING “RECOVERY TIME” !!



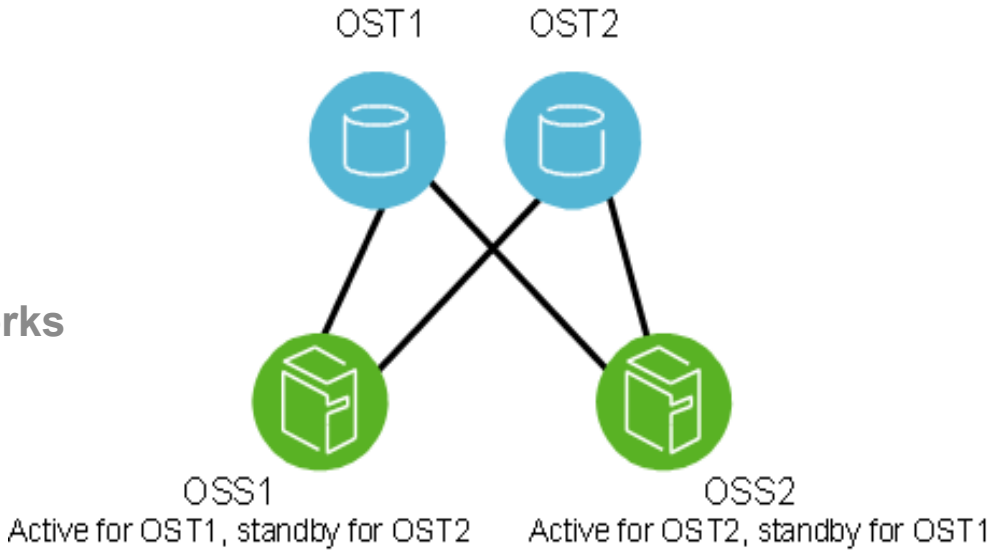
3 High availability issues

ÍNDICE

- **What happens if an OSS server fails?**
- **Second server is in active/active or active/passive mode.**

List of possible OSS

- **GOOD AND TRANSPARENT!! Filesystem still works without interruption**



3 High availability issues

ÍNDICE

But, what does it happens if an OST server fail?

- **My disk crashed!**
  - Control parity → Data redundancy --> degraded RAID
  - More RAID level, more reliability
  - More RAID level, less space

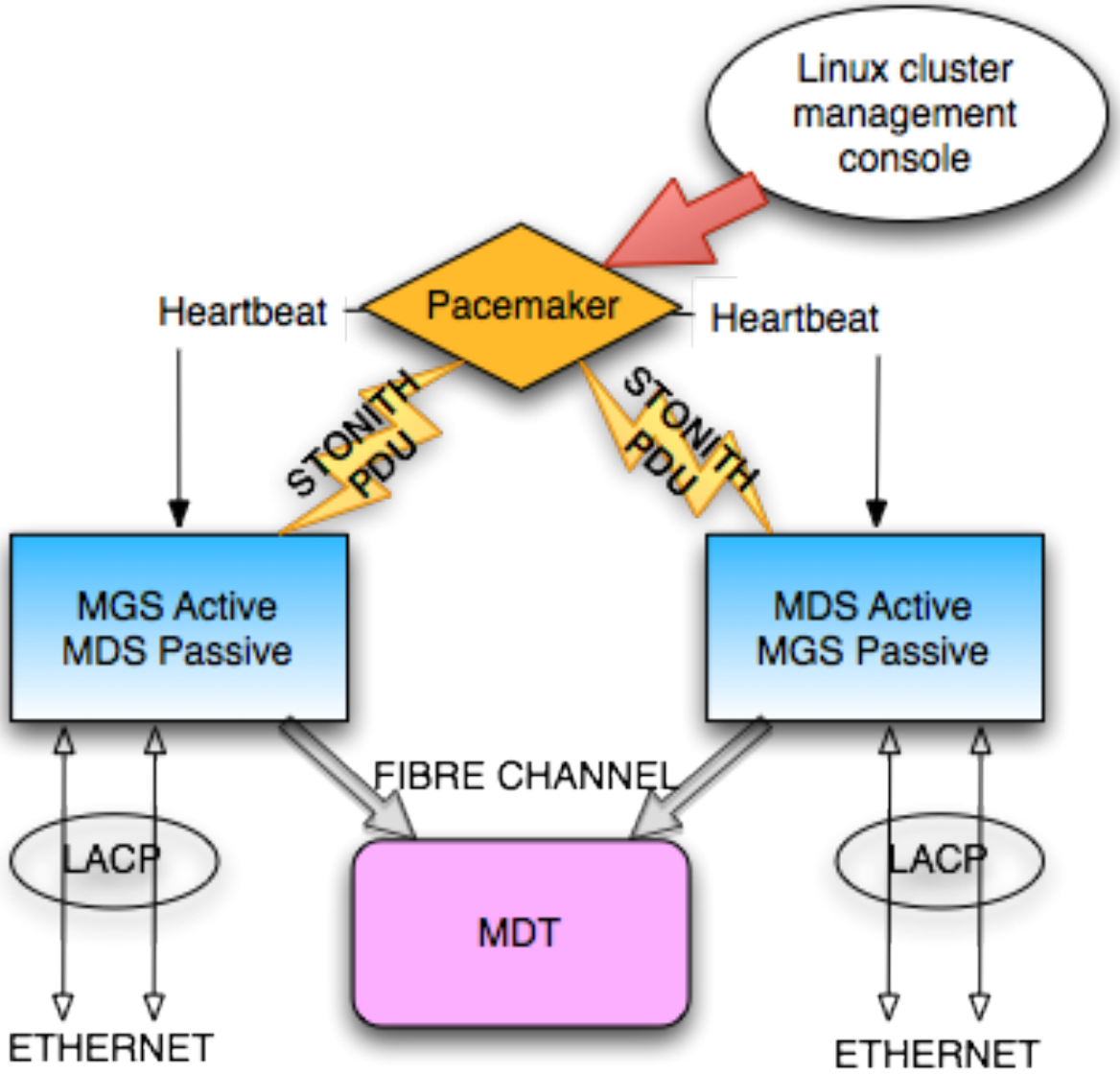
- But, If my disk controller fails or if 2,3 or more disks crash!
  - RAID failure → lost data
  - Filesystem stucks during RAID reconstruction
  - Filesystem stucks during disk controller substitution

4 HA for MDS/MGS

- **Design concerns:**
- **LUSTRE Metadata server only permits one active server**
- Active/Passive architecture
- MGS/MDS in separated servers → More efficient for high I/O request
- Server 1: MGS active and MDS passive
- Server 2: MDS active and MGS passive
- Pacemaker “kills” the failure server → STONITH service
- STONITH with PDU management (APC7000)

ÍNDICE

4 HA for MDS/MGS



4 HA for MDS/MGS

<primitive class="ocf" id="resMGS" provider="heartbeat" type="Filesystem"> <instance_attributes id="resMGS-instance_attributes"> <nvpair id="resMGS-instance_attributes-device" name="device" value="/dev/mapper/MGS"/> <nvpair id="resMGS-instance_attributes-directory" name="directory" value="/mgs"/> <nvpair id="resMGS-instance_attributes-fstype" name="fstype" value="lustre"/> </instance_attributes> <operations id="resMGS-operations"> <op id="resMGS-start-0" interval="0" name="start" timeout="30"/> <op id="resMGS-stop-0" interval="0" name="stop" timeout="30"/> <op id="resMGS-monitor-60" interval="15" name="monitor" start-delay="0" timeout="30"/> </operations> <meta_attributes id="resMGS-meta_attributes"> <nvpair id="resMGS-meta_attributes-target-role" name="target-role" value="started"/> </meta_attributes> </primitive> <primitive id="resMDS" class="ocf" provider="heartbeat" type="Filesystem"> <instance_attributes id="resMDS-instance_attributes"> <nvpair id="resMDS-instance_attributes-device" name="device" value="/dev/mapper/MDT"/> <nvpair id="resMDS-instance_attributes-directory" name="directory" value="/cetafs"/> <nvpair id="resMDS-instance_attributes-fstype" name="fstype" value="lustre"/> </instance_attributes> <operations id="resMDS-operations"> <op interval="0" id="resMDS-start-0" name="start" timeout="120"/> <op interval="0" id="resMDS-stop-0" name="stop" timeout="120"/> <op id="resMDS-monitor-60" name="monitor" interval="15" timeout="30" start-delay="0"/> </operations> <meta_attributes id="resMDS-meta_attributes"> <nvpair id="resMDS-meta_attributes-target-role" name="target-role" value="started"/> </meta_attributes> </primitive>

- **Resources definition**

HA for MDS/MGS

▪Fence definition

<primitive id="stonith_fence_apc_snmp_fence_apc2" class="stonith" type="fence_apc_snmp"> <instance_attributes id="stonith_fence_apc_snmp_fence_apc2-instance_attributes"> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-action" name="action" value="off"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-ipaddr" name="ipaddr" value="192.168.9.112"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-login" name="login" value="apc"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-passwd" name="passwd" value="apc"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-port" name="port" value="2"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-pcmk_host_check" name="pcmk_host_check" value="static-list"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-pcmk_host_list" name="pcmk_host_list" value="ic-d1-01 ic-d1-02"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc2-pcmk_host_map" name="pcmk_host_map" value="ic-d1-01:1 ic-d1-02:2"/> </instance_attributes> <meta_attributes id="stonith_fence_apc_snmp_fence_apc2-meta_attributes"/> </primitive> <primitive id="stonith_fence_apc_snmp_fence_apc1" class="stonith" type="fence_apc_snmp"> <instance_attributes id="stonith_fence_apc_snmp_fence_apc1-instance_attributes"> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-action" name="action" value="off"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-ipaddr" name="ipaddr" value="192.168.9.114"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-login" name="login" value="apc"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-passwd" name="passwd" value="apc"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-port" name="port" value="2"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-pcmk_host_check" name="pcmk_host_check" value="static-list"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-pcmk_host_list" name="pcmk_host_list" value="ic-d1-01 ic-d1-02"/> <nvpair id="nvpair-stonith_fence_apc_snmp_fence_apc1-pcmk_host_map" name="pcmk_host_map" value="ic-d1-01:1 ic-d1-02:2"/> </instance_attributes> <meta_attributes id="stonith_fence_apc_snmp_fence_apc1-meta_attributes"/> </primitive>

HA for MDS/MGS

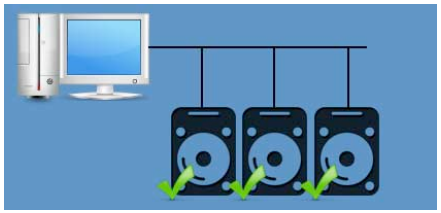
▪Resources and fence allocation

<rsc_location id="loc_resMGS_ic-d1-01" node="ic-d1-01" rsc="resMGS" score="2"/> <rsc_location id="loc_resMDS_ic-d1-02" node="ic-d1-02" rsc="resMDS" score="2"/> <rsc_location id="loc_resMDS_ic-d1-01" node="ic-d1-01" rsc="resMDS" score="0"/> <rsc_location id="loc_resMGS_ic-d1-02" node="ic-d1-02" rsc="resMGS" score="0"/> <rsc_location id="loc_stonith_fence_apc_snmp_fence_apc2_ic-d1-01" rsc="stonith_fence_apc_snmp_fence_apc2" node="ic-d1-01" score="2"/> <rsc_location id="loc_stonith_fence_apc_snmp_fence_apc2_ic-d1-02" rsc="stonith_fence_apc_snmp_fence_apc2" node="ic-d1-02" score="0"/> <rsc_location id="loc_stonith_fence_apc_snmp_fence_apc1_ic-d1-01" rsc="stonith_fence_apc_snmp_fence_apc1" node="ic-d1-01" score="2"/> <rsc_location id="loc_stonith_fence_apc_snmp_fence_apc1_ic-d1-02" rsc="stonith_fence_apc_snmp_fence_apc1" node="ic-d1-02" score="0"/>

5 Proposal for data HA

ÍNDICE

- **When a OST fail:**
- RAID reconstruction can take a long time.



- A failed RAID can lost a lot of files in a splitted filesystem

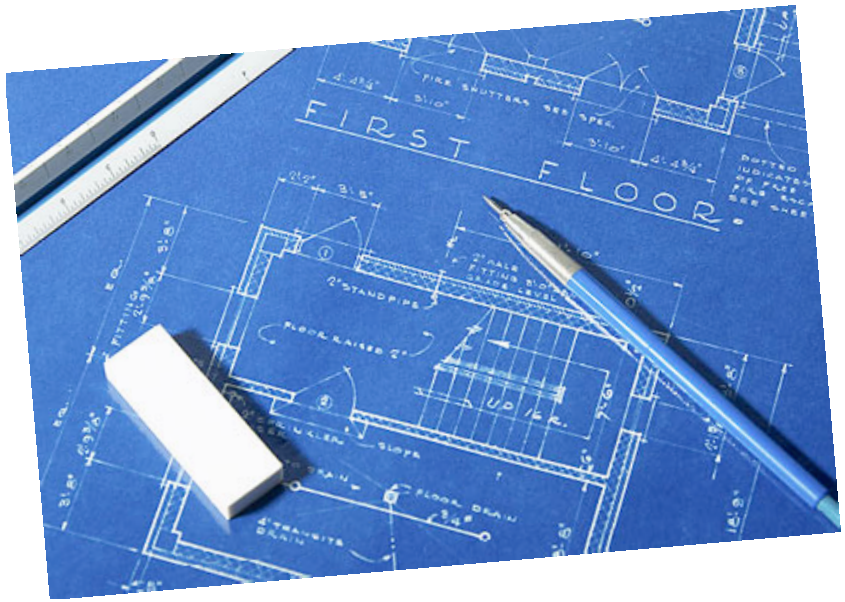
- A failure in a disk controller could stuck a entire filesystem

5 Proposal for data HA (Blueprint)

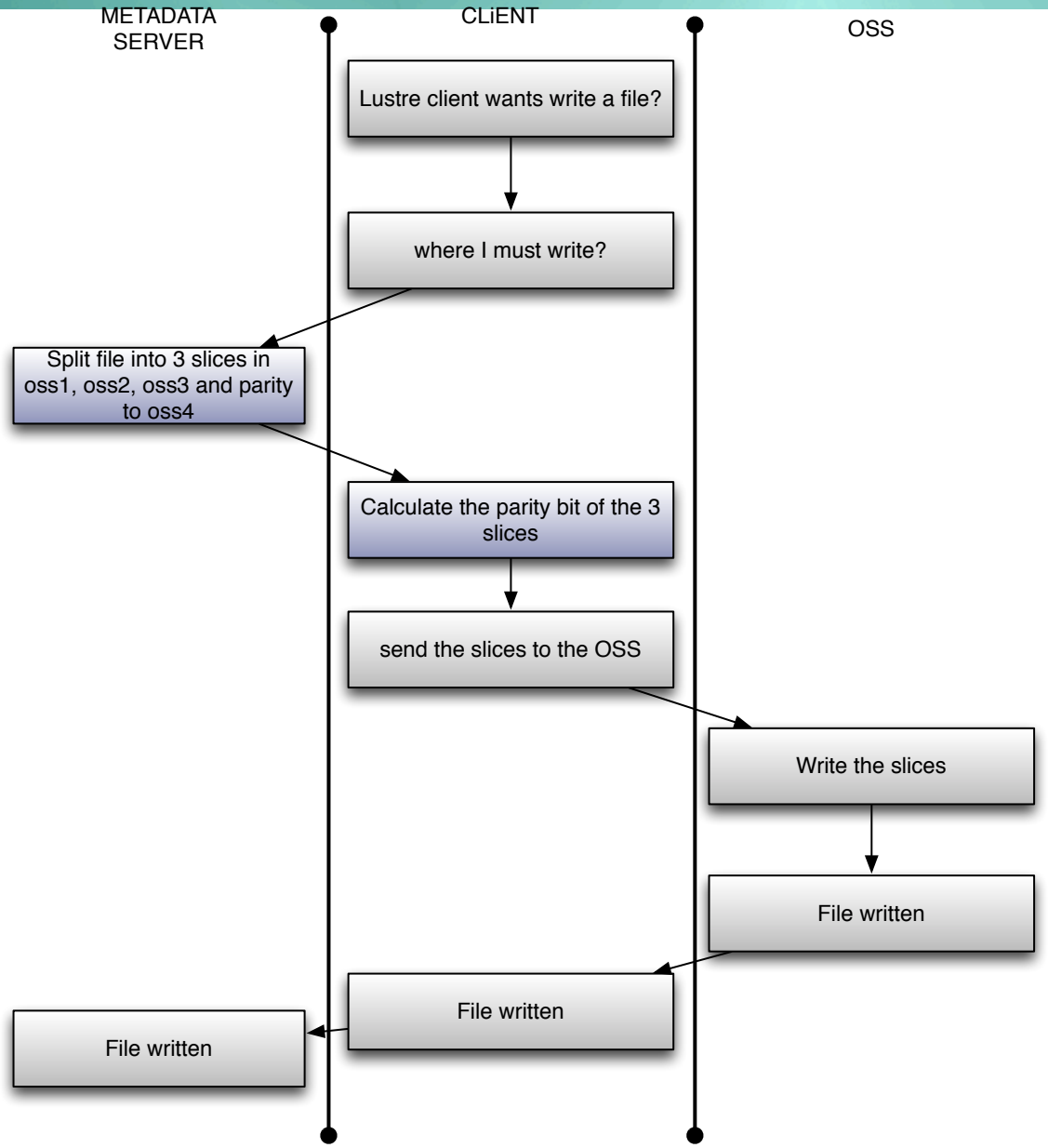
- Basically brings the RAID parity bit to a splitted filesystem
- Clients calculate the parity bit and this bit is written in an OST
- If OSS/OST fails, the client can reconstruct the file in runtime
- Increased client CPU usage.
- Are you a security data paranoid or your company need it?

Double parity

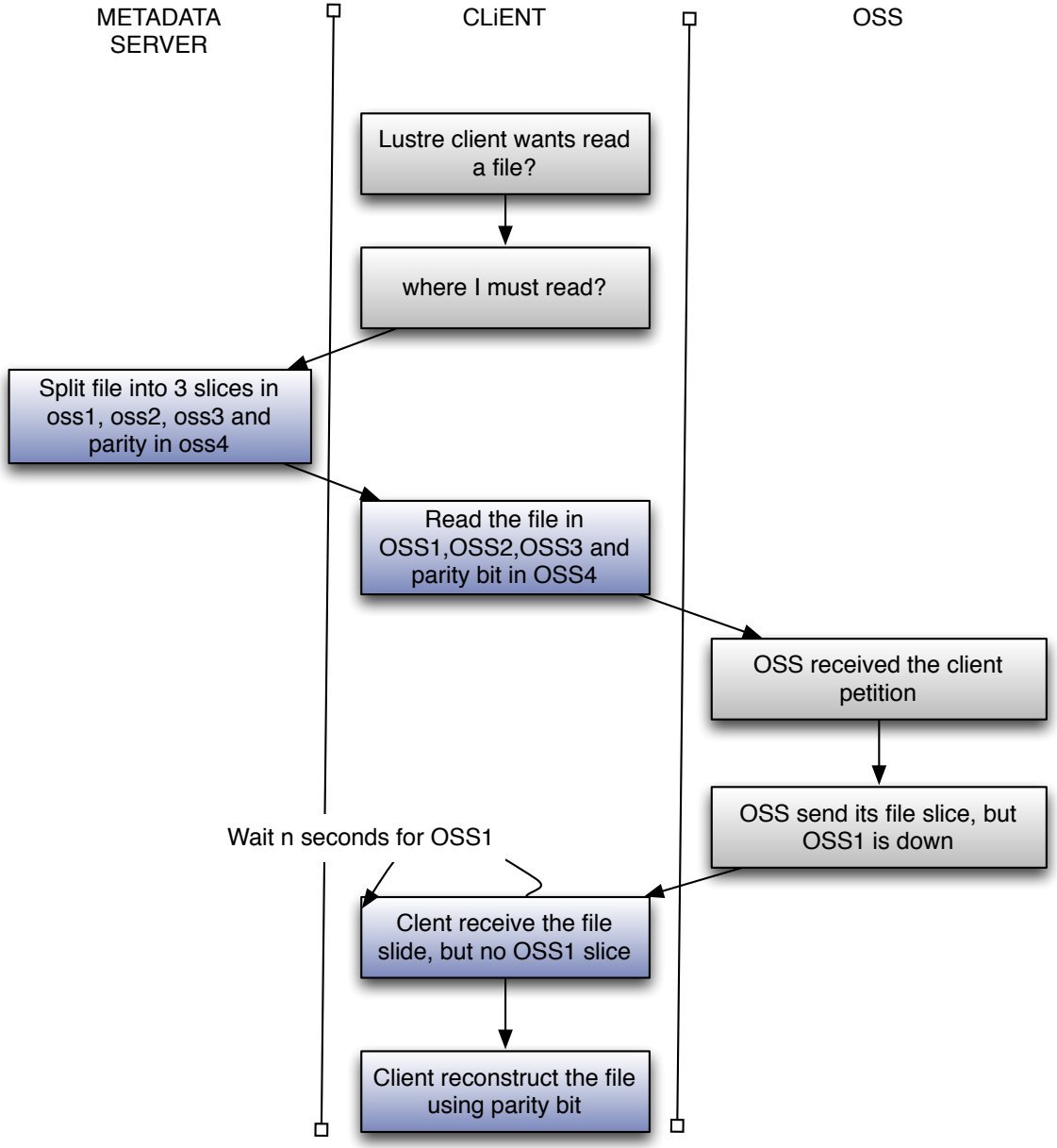
We are working on it!



5 Proposal for data HA



5 Proposal for data HA



A Appendix: Grid software incompatibilities

ÍNDICE

- **Grid middleware troubles:**
- Quotas:

When user quota finishes, the grid software (middleware) still tries to write.

- Free disk space report:

Grid software “see” the full file system. This space is limited by quota.

Grid middleware must-to-see the free quota space.

gLite
Lightweight Middleware for Grid Computing

XRootD

THANKS!!!

AND THANKS TO:

PROYECTO COFINANCIADO POR LA UNIÓN EUROPEA

FONDO EUROPEO DE DESARROLLO REGIONAL

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

Conventual de San Francisco, Sola 1, 10200 Trujillo
Telephone: 927 65 93 17 Fax: 927 32 32 37
www.ceta-ciemat.es

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA E INNOVACIÓN

CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS

CETA Ciemat

FEDER

Fondo Europeo de Desarrollo Regional

Una manera de hacer Europa