energie atomique • energies alternatives

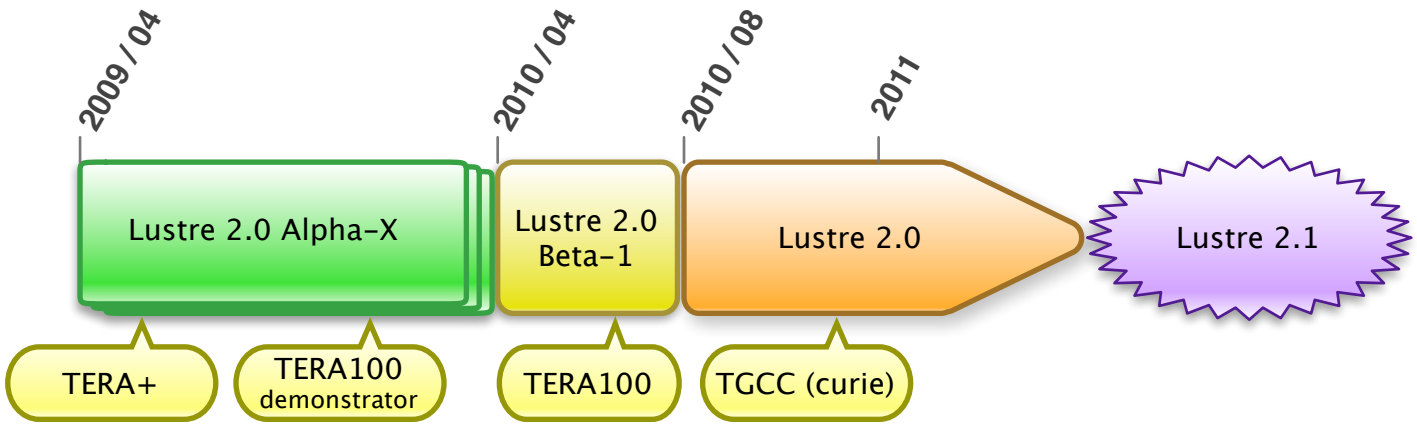
Administering Lustre 2.0 at CEA

European Lustre Workshop 2011

September 26-27, 2011

Stéphane Thiell – CEA/DAM
stephane.thiell@cea.fr

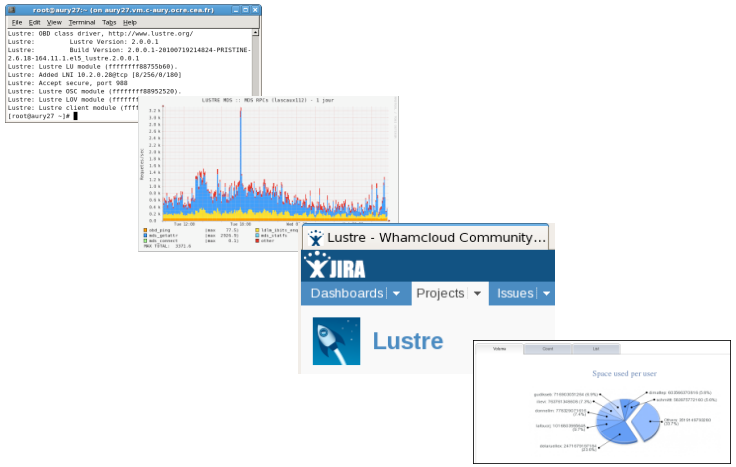
Lustre 2.0 timeline at CEA



- **Early adoption of Lustre 2.0**
  - Benefit from 2.0 features in new production machines
    - FIDs (eg. NFS-Ganesha FSAL/LUSTRE)
    - Changelogs
  - Ease future 2.x upgrades of production machines
    - Lustre HSM binding
  - Push forward Lustre technology

Administering Lustre 2.0 at CEA - Outline

- **Operations & Maintenance**
  - Shine
  - Lessons learned
  - Performance tunings
- **Monitoring**
- **Features**
- **Resolved issues**



Lustre 2.0 Operations & Maintenance

energie atomique • energies alternatives

- **Lustre administration software ecosystem at CEA**
  - Base suites
    - DM-Multipath (Linux native multipath I/O)
    - Shine (Lustre FS management and tuning)
    - Robinhood Policy Engine (Powerful FS content manager)
    - Puppet (Centralized configuration management)
    - Pacemaker (High availability)
  - Monitoring
    - Syslog-ng (Versatile replacement for syslogd)
    - SEC (Simple Event Correlator)
    - Nagios (Supervision)
    - RRDTool (Graphing engine)
  - Debugging
    - crash (dumps or live systems debugging)

Shine to manage Lustre 2.0 filesystems (1/5)

- **Shine Python Library and Tools for Lustre**
  - Setup and manage Lustre filesystem components in parallel
  - Open Source (GPL): http://lustre-shine.sf.net/
  - Human-readable configuration file to describe each FS

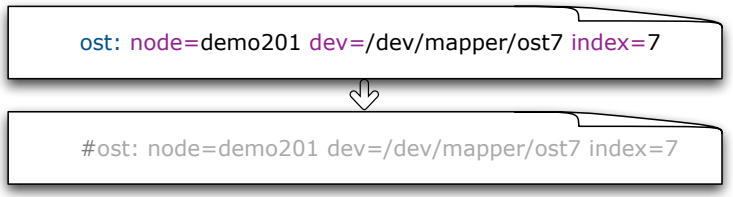
```
# file: /etc/shine/models/simple1.lmf - cluster: demo

fs_name: simple1
description: Simple Lustre filesystem model file example

nid_map: nodes=demo[0-999] nids=demo[0-999]-ib0@o2ib0

mgt: node=demo110 ha_node=demo111 dev=/dev/mapper/mgt

mdt: node=demo113 ha_node=demo114 dev=/dev/mapper/s1_mdt

ost: node=demo200 ha_node=demo201 dev=/dev/mapper/s1_ost[0-3] index=[0-3]
ost: node=demo201 ha_node=demo200 dev=/dev/mapper/s1_ost[4-7] index=[4-7]

client: node=demo[50-55,400-759]

mount_path: /demo/simple1

stripe_size: 1048576
stripe_count: 2
```

Shine to manage Lustre 2.0 filesystems (2/5)

```
# file: /etc/shine/models/more2.lmf - cluster: demo

fs_name: more2
description: More advanced Lustre filesystem model file example

nid_map: nodes=demo[0-999] nids=demo[0-999]-ib0@o2ib0

# Externally managed MGS
mgt: node=demo110 ha_node=demo111 mode=external

mdt: node=demo113 ha_node=demo114 dev=/dev/mapper/m2_mdt

ost: node=demo200 ha_node=demo201 dev=/dev/mapper/m2_ost[0-3] index=[0-3]
ost: node=demo201 ha_node=demo200 dev=/dev/mapper/m2_ost[4-7] index=[4-7]

client: node=demo[50-55,400-759]
mount_path: /demo/more2
mount_options: acl,user_xattr,flock
mdt_mount_options: acl,user_xattr
stripe_size: 1048576
stripe_count: 2
mdt_mkfs_options: -O dir_index,dirdata,uninit_bg,mmp,flex_bg -G 256
ost_mkfs_options: -m2 -O dir_index,extents,uninit_bg,mmp,flex_bg -G 256
quota: yes
```

Shine to manage Lustre 2.0 filesystems (3/5)

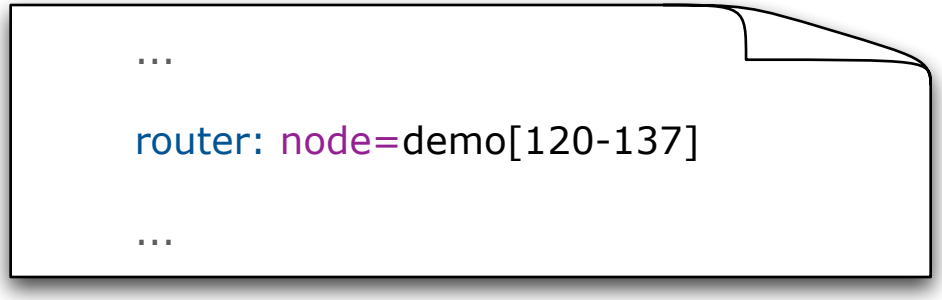
- **Use case: removing OST made easy with Shine (v0.910+)**
  - Purge OST #7 files (list and backup files before if needed)
    - `# robinhood --purge-ost=7,0 -i`
  - Stop all filesystem components
    - `# shine umount -f demofs`
      `# shine stop -f demofs`
  - Remove OST #7 definition from */etc/shine/models/demofs.lmf*



  - Update distributed shine FS config files from new model
    - `# shine update -m /etc/shine/models/demofs.lmf`
  - Run needed tunefs.lustre in parallel
    - `# shine tunefs -f demofs` (perform *writeconf* in this case)
  - Start filesystem again
    - `# shine start -f demofs`
      `# shine mount -f demofs`

Shine to manage Lustre 2.0 filesystems (4/5)

- **Manage LNET routers easily with Shine**
  - One line in config file to define routers for a filesystem (.lmf):



  - Centralized routers management (start, stop, status)

```
# shine status -t router -f demofs
FILESYSTEM COMPONENTS STATUS (demofs)
+-----+---+--------------+-------+
|type |#  |    nodes     |status |
+-----+---+--------------+-------+
|ROU  |18 |demo[120-137] |online |
+-----+---+--------------+-------+
```

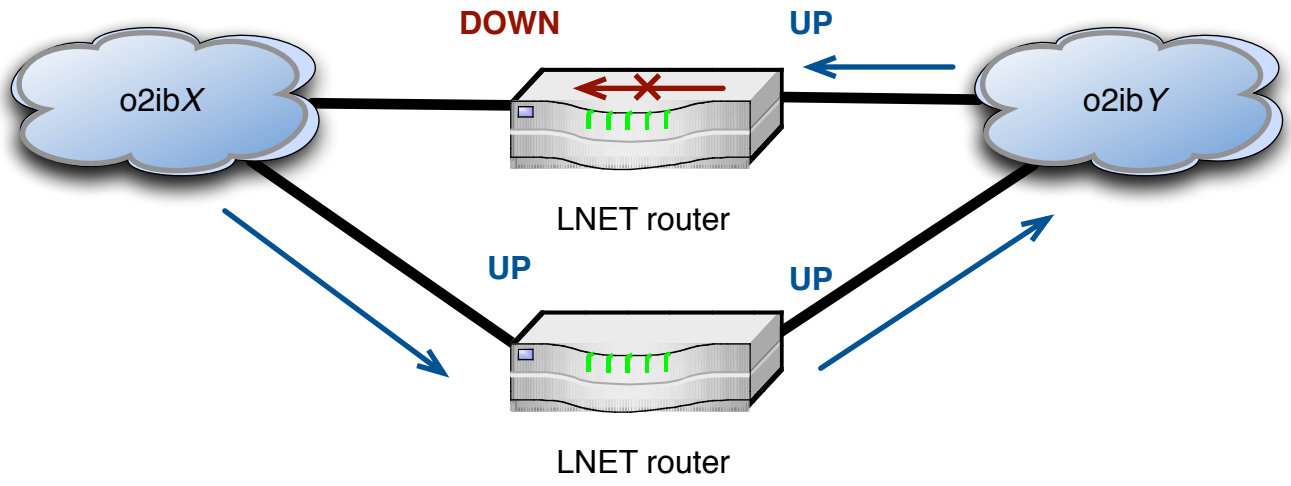
Shine to manage Lustre 2.0 filesystems (5/5)

- **Shine: advanced features**
  - Multi-NIDs support
    - nid_map rule can be repeated
    - Optional static network definition per target

```
nid_map: nodes=demo[0-999] nids=demo[0-999]-ib0@o2ib0
nid_map: nodes=demo[0-999] nids=demo[0-999]-ib1@o2ib1


ost: node=demo200 dev=/dev/mapper/ost[0-8/2] index=[0-8/2] network=o2ib0
ost: node=demo201 dev=/dev/mapper/ost[1-9/2] index=[1-9/2] network=o2ib1
```

Lustre 2.0 Operations: lessons learned

- **Enable *flex_bg* ext4 feature**
  - `ost_mkfs_options: -O dir_index,extents,mmp,uninit_bg,flex_bg -G 256`
  - Enabled by default in Lustre 2.1 (LU-255)
  - *fsck* time is not a problem anymore for us (as of today)
  - Tune backend filesystem according to hardware
- **Router status coherency**
  - Strictly avoid any incoherent status from both sides of router
  - Need to check `lctl route_list` output from both sides



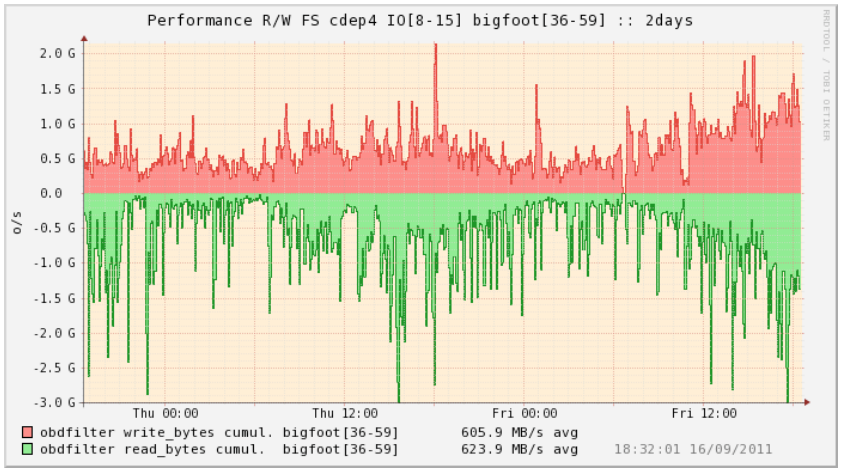
Lustre 2.0 Operations: performance tunings

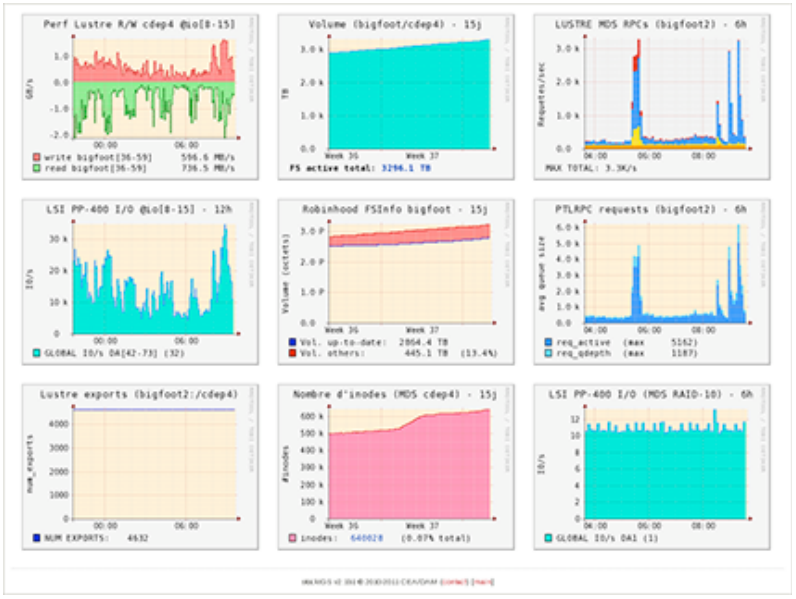
- **Lustre tunings (clients)**
  - `osc.*.max_pages_per_rpc=256`
  - `llite.*.max_cached_mb`
    - Lost feature as of Lustre 2.0 (LU-141)
- **Kernel tunings**
  - Increase `vm.min_free_kbytes` on clients
    - `vm.min_free_kbytes=262144`
  - `vm.zone_reclaim_mode` (depending on hardware)
    - `vm.zone_reclaim_mode=0`
      - pages reclaim (and allocation) satisfied from all zones
      - stable (but not best) performance with Lustre
    - `vm.zone_reclaim_mode=1`
      - pages reclaim from remote zones are avoided
      - variable, long-term inconsistent performance
      - not good until CPU-affinity thread pools in Lustre

Lustre 2.0 Monitoring (1/3)

- **Lustre activity monitoring**
  - Track user creativeness with Robinhood alerts
  - Home-made Lustre activity monitoring solution
    - ClusterShell scripts to gather Lustre metrics
      - http://clustershell.sf.net/
    - RRDTool graphs

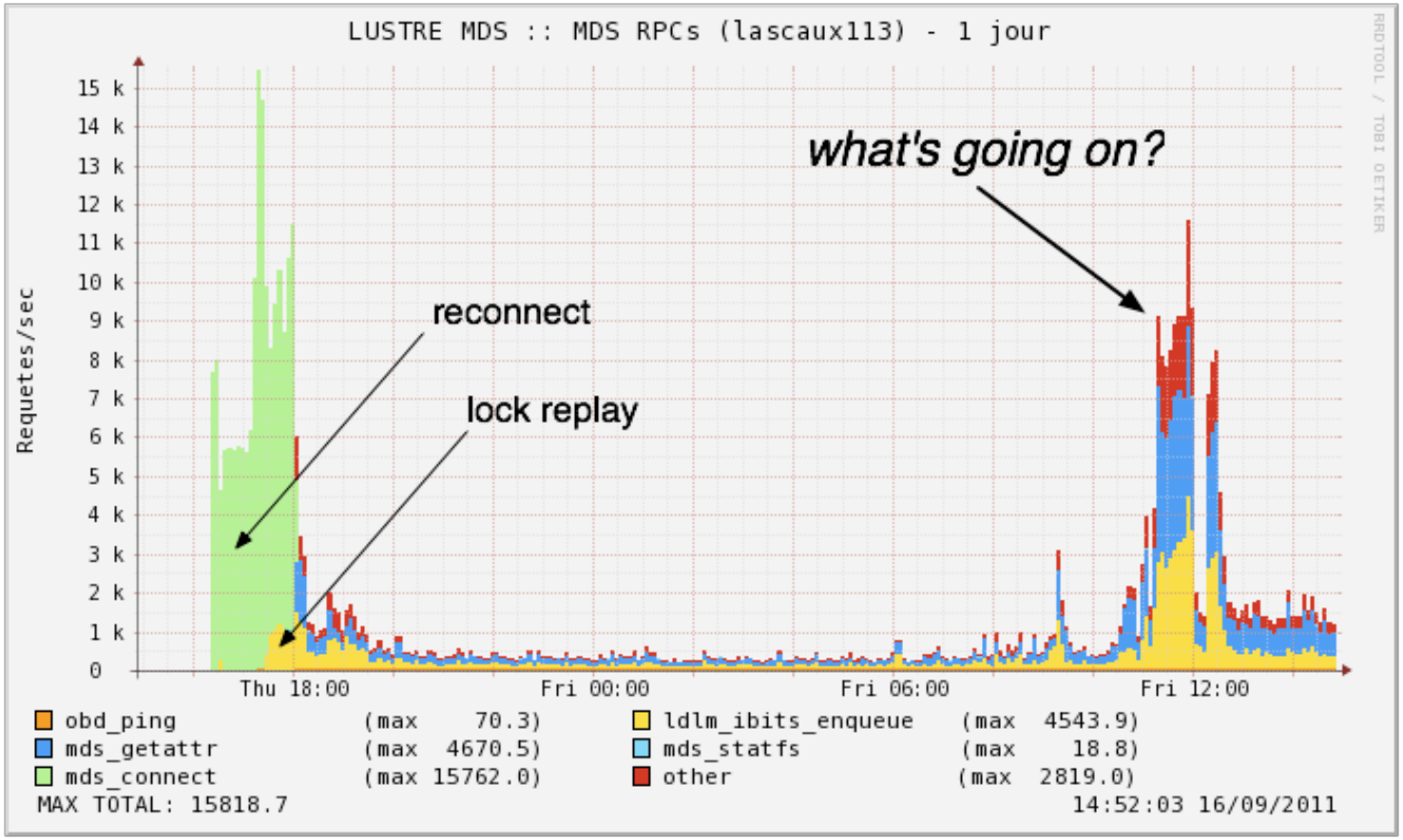
« Overlook »





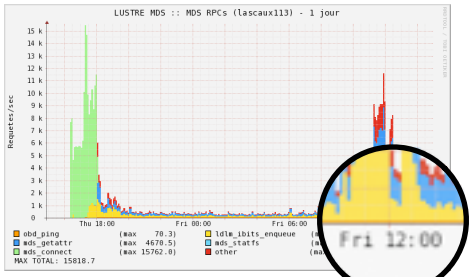
Lustre 2.0 Monitoring (2/3)

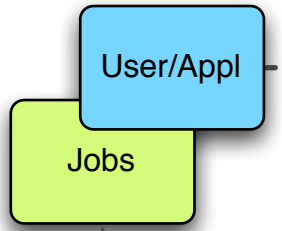
- **Lustre RPC monitoring on MDS**
  - Track nasty behavior



Lustre 2.0 Monitoring (3/3)

- **Lustre RPC tracing**





slurm (sacct)

Lustre clients (NIDs) / Nodes

RPC types

```
# ./lustre_rpc_sample.py analyze –f \
        lascaux113_11-09-16_11:40:07.log.gz -D
Reading sample file 'lascaux113_11-09-16_11:40:07.log.gz'
GENERAL
Duration:                5.0 sec
Total RPC:             53776     (10775.1 RPC/sec)
Ignored ping RPC:        582

NODENAME     COUNT        %   CUMUL   DETAILS
lascaux3283   1536     2.86    2.86   mds_getattr: 1536
lascaux3191   1536     2.86    5.71   mds_getattr: 1536
lascaux3181   1536     2.86    8.57   mds_getattr: 1536
lascaux3188   1536     2.86   11.43   mds_getattr: 1536
lascaux3189   1536     2.86   14.28   mds_getattr: 1536
lascaux3185   1536     2.86   17.14   mds_getattr: 1536
lascaux3187   1536     2.86   19.99   mds_getattr: 1536
lascaux3146   1536     2.86   22.85   mds_getattr: 1536
lascaux3190   1536     2.86   25.70   mds_getattr: 1532
```

Lustre 2 at CEA: Enabled features

- **Quotas**
  - Quotas with RHEL6 issues (LU-91, LU-369, LU-484)
  - Inodes and volume quotas enabled on TGCC (curie)
- **Changelogs**
  - Patchs applied (bz23035, bz23298, bz23120, LU-81, LU-542)
  - Enabled on TGCC (curie) for faster Robinhood
  - `mdd.*.changelog_mask`
    - `"MARK CREAT MKDIR HLINK SLINK UNLNK RMDIR RNMFM RNMTO TRUNC SATTR MTIME"`
- **Large OSTs support (>16TB)**
  - Added by LU-136 (available in Lustre 2.1)

Lustre 2 at CEA: Major issues resolved since 2.0 GA

- **Client hang issues (resulting to lots of evictions)**
  - LU-416, LU-437 (LU-394)
- **Client crash on special I/O transfert nodes**
  - LU-185
- **I/O errors when using DM-multipath (RHEL6)**
  - LU-275
- **OSS pseudo-hang after OSTs stop/start**
  - LU-328
- **OSS frequent crashes due to LBUG/[ASSERTION(last_rcvd >= le64_to_cpu(lcd->lcd_last_transno))**
  - bz24420
- **Binary execve() over Lustre FS issues**
  - LU-300, LU-613

Conclusion

- **Lustre 2 stability**
  - Patched Lustre 2.0 is now very stable at CEA
    - So will be Lustre 2.1 as all major patches have been integrated
  - Thanks to Whamcloud and Bull Lustre teams
    - Both very responsive
- **Admin tools**
  - Good, scalable Lustre admin tools save a lot of time
  - We will continue to improve our Open Source tools
- **Lustre activity profiling and tracing**
  - Things are already possible today
  - Improvements needed for the future

energie atomique • energies alternatives

Questions ?

Shine
http://lustre-shine.sf.net/

ClusterShell
http://clustershell.sf.net/

Robinhood Policy Engine
http://robinhood.sf.net/

NFS-Ganesha FSAL/LUSTRE
http://nfs-ganesha.sf.net/