Lustre: Striping by clients / Templated striping

riaux.jb@intel.com || jb.riaux@intel.com

Hello World

- First time at the dev summit
- First contribution *from scratch* (which is not a backport)
- Very enjoyable and best way to step-up.

Concept

- Be able to specify targets (stripe layout) for a group of clients (1-N).

  => Nodemap was perfect match

- Nodemap can be queried from an export (mti_exp):
  - Easy to get the template on the MDS

- “Just” needed to identify the correct code section and best way to implement

- You can find the PoC here : LU-9982

Impacted areas

- For configuration : nodemap / ptlrpc / lproc / mgs / lctl / obd
- For striping : mdt / mdd / lod / osd

No impact on client side

Nodemap

- Added a FID template in structure lu_nodemap:
  - For the PoC : char* used but will be a lu_fid at term

-
```
[root@client testfs2]# echo "test" > testfile
[root@client testfs2]# lfs getstripe -c testfile
1
[root@client testfs2]# lfs getstripe -c my_template
2
[root@client testfs2]# lfs path2fid my_template
[0x200000402:0x1:0x0]

[root@mds1]# lctl nodemap_add  SG1
[root@mds1]# lctl nodemap_add_range --name SG1 --range 192.168.10.[130-140]@tcp
[root@mds1]# lctl set_param -P nodemap.SG1.template=0x200000401:0x1:0x0
or
[root@mds1]# lctl nodemap_set_template --name SG1 --template 0x200000401:0x1:0x0
or
[root@mds1 ~]# echo 0x200000401:0x1:0x0 > /proc/fs/lustre/nodemap/SG1/template
[root@mds1]# lctl nodemap_modify --name SG1 --property admin --value 1 # working as root
[root@mds1]# lctl nodemap_activate 1
[root@mds1]# cat /proc/fs/lustre/nodemap/SG1/template
0x200000402:0x1:0x0

[root@client testfs2]# echo "test" > testfile2
[root@client testfs2]# lfs getstripe -c testfile2
2
```

MDS integration (current code)

- At file and directory creation call to « nodemap_get_from_exp » and open the object matching the template FID.
- To respect Lustre striping:
  - Template is applied only if the directory has default striping (not set by any admin before)
- Overload existing function do_create and do_ah_init and underlaying layer (MDD/LOD/OSD-ldiskfs/zfs) with the template object.
- Template object is locked internally on the MDT (lock handle MDT_LH_LOCAL)

Limits / issues

- Strange behavior observed with admin nodemap commands (admin param)

- Current patch works for files only and will be exploded in 3 parts:
  - nodemap, mdt, tests

- New (not uploaded) patch works for directories but hitting LU-9766 (wrong directory inheritance with DNE).

⇒ Still working on this

Other possible approach ?

- Use a dt_allocation_hint structure:
  - Create the hint before the do_create

    ⇒ Internal lock on template could be released earlier (now it's released after the creation as the object is passed to underlaying layers)

  - Need to change mdd_object_make_hint function which currently memset 0 the existing hint argument.