Data-on-MDT (DoM) User Guide

1. Introduction

The Lustre Data-on-MDT (DoM) feature improve a small IO by placing small files on MDT and improve large IO as well by avoiding OST being polluted with small random IO. So that users can expect reasonable performance as for small file IO as for a mixed IO patterns.

The layout of a DoM file is stored on disk as *composite layout* and is special case of a PFL file. For DoM files, the file layout is composed of the first special component which is placed on MDT and rest of components on OSTs. The first component has MDT layout which is new layout type. It is placed on MDT in the MDT object data blocks. This component has always single stripe with size equal to the component size. Such component with MDT layout can be only the first component in composite layout. The rest of components are placed over OSTs as usual with RAID0 layout. They are not instantiated initially and that happens upon file grow over the limit of MDT layout.

In the remaining of this document, some user commands for user to operate DoM files are introduced and some examples are illustrated as well.

2. User Commands

Lustre provides `lfs setstripe` command for users to create DoM files. Also, as usual, `lfs getstripe` commands can be used to list the striping/component information for a given file, and `lfs find` commands can be used to search the directory tree rooted at the given directory or file name for the files that match the given DoM component parameters, e.g. layout type.

2.1 lfs setstripe

`lfs setstripe` command is used to create DoM files.

Command

`lfs setstripe <--component-end|-E end1> <--layout|-L> mdt [<--component-end|-E end2> [STRIPE_OPTIONS] ...] <filename>`

Create a file with the special composite layout which defines the first component as MDT component. The MDT component must start from offset 0 and its end '**end1**' is the stripe size at the same time. No other options are required.

The rest of components uses just usual syntax for composite files creation.

Example

```
$ lfs setstripe -E 1M -L mdt -E -1 -S 4M -c -1 /mnt/lustre/domfile
```

This command creates a file with DoM layout. The first component has 'mdt' layout and is placed on MDT, it covers [0, 1M). The second component covers [1M, EOF) and is striped over all available OSTs.

That layout is illustrated on Figure 1.

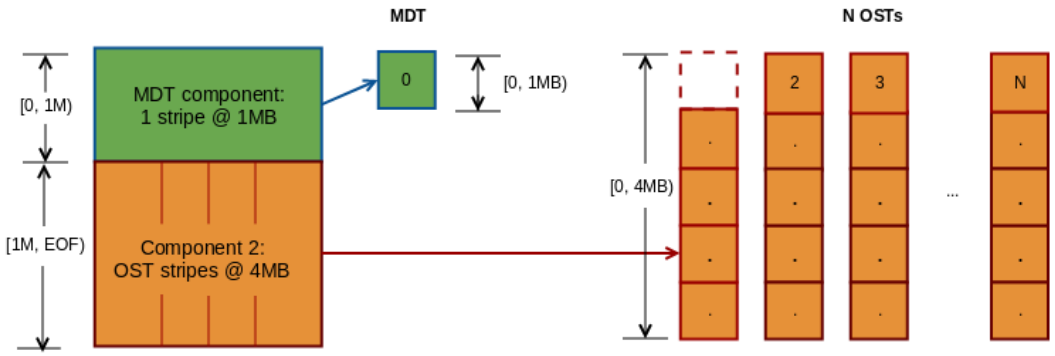

Figure 1. Example of DoM layout

Resulting layout can be checked with `lfs getstripe`:

```
$ lfs getstripe /mnt/lustre/domfile
/mnt/lustre/domfile
  lcm_layout_gen:  2
  lcm_entry_count: 2
    lcme_id:             1
    lcme_flags:          init
    lcme_extent.e_start: 0
    lcme_extent.e_end:   1048576
      lmm_stripe_count:  0
      lmm_stripe_size:   1048576
      lmm_pattern:       100
      lmm_layout_gen:    0
      lmm_stripe_offset: 0
      lmm_objects:

    lcme_id:             2
    lcme_flags:          0
    lcme_extent.e_start: 1048576
    lcme_extent.e_end:   EOF
      lmm_stripe_count:  -1
      lmm_stripe_size:   4194304
      lmm_pattern:       1
      lmm_layout_gen:    65535
      lmm_stripe_offset: -1
```

It shows that the first component has size 1MB and pattern '`100`' which means '`mdt`', the second component is not instantiated yet which is seen by `lcme_flags: 0`.

If we will write more than 1MB of data to the file then `lfs getstripe` output is changed:

```
# lfs getstripe /mnt/lustre/domfile
/mnt/lustre/domfile
  lcm_layout_gen:  3
  lcm_entry_count: 2
    lcme_id:             1
    lcme_flags:          init
    lcme_extent.e_start: 0
    lcme_extent.e_end:   1048576
      lmm_stripe_count:  0
      lmm_stripe_size:   1048576
      lmm_pattern:       100
      lmm_layout_gen:    0
      lmm_stripe_offset: 2
      lmm_objects:

    lcme_id:             2
    lcme_flags:          init
    lcme_extent.e_start: 1048576
    lcme_extent.e_end:   EOF
      lmm_stripe_count:  2
      lmm_stripe_size:   4194304
      lmm_pattern:       1
      lmm_layout_gen:    0
      lmm_stripe_offset: 0
      lmm_objects:
      - 0: { l_ost_idx: 0, l_fid: [0x100000000:0x2:0x0] }
      - 1: { l_ost_idx: 1, l_fid: [0x100010000:0x2:0x0] }
```

The second component has objects on OSTs now with 4MB stripe.

2.1.1 Set default DoM layout to an existing directory

DoM layout can be set to an existing directory as well. Then all the files created after that will inherit this layout by default.

Command

lfs setstripe <--component-end|-E end1> <--layout|-L> mdt [<--component-end|-E end2> [STRIPE_OPTIONS] ...] <dirname>

Example

```
$ mkdir /mnt/lustre/domdir
$ touch /mnt/lustre/domdir/normfile
$ lfs setstripe -E 1M -L mdt -E -1 /mnt/lustre/domdir/
$ lfs getstripe -d /mnt/lustre/domdir
  lcm_layout_gen:  0
  lcm_entry_count: 2
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 0
    lcme_extent.e_end:   1048576
      stripe_count:  0       stripe_size:   1048576       pattern:
100       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 1048576
    lcme_extent.e_end:   EOF
      stripe_count:  1       stripe_size:   1048576       pattern:       1
stripe_offset: -1
```

We can see that directory has default layout with DoM component. Let's check layouts of files in that directory now

```
$ touch /mnt/lustre/domdir/domfile
$ lfs getstripe /mnt/lustre/domdir/normfile
/mnt/lustre/domdir/normfile
lmm_stripe_count:  2
lmm_stripe_size:   1048576
lmm_pattern:       1
lmm_layout_gen:    0
lmm_stripe_offset: 1
 obdidx   objid   objid   group
      1             3          0x3             0
      0             3          0x3             0

$ lfs getstripe /mnt/lustre/domdir/domfile
/mnt/lustre/domdir/domfile
  lcm_layout_gen:  2
  lcm_entry_count: 2
    lcme_id:             1
    lcme_flags:          init
    lcme_extent.e_start: 0
    lcme_extent.e_end:   1048576
      lmm_stripe_count:  0
      lmm_stripe_size:   1048576
      lmm_pattern:       100
      lmm_layout_gen:    0
      lmm_stripe_offset: 2
      lmm_objects:

    lcme_id:             2
    lcme_flags:          0
    lcme_extent.e_start: 1048576
    lcme_extent.e_end:   EOF
      lmm_stripe_count:  1
      lmm_stripe_size:   1048576
      lmm_pattern:       1
      lmm_layout_gen:    65535
      lmm_stripe_offset: -1
```

We can see that first file **normfile** in that directory has just ordinary layout, but file **domfile** inherits directory default layout and is DoM file.

Note about size limit and default directory layout

Directory default layout setting will be inherited by new files even if server DoM size limit will be set to a lower value.

2.1.2 DoM stripe size restrictions

The maximum size of DoM component is restricted in several way to protect MDT from being filled with large files eventually.

LFS limits for DoM component size

A `lfs setstripe` allows to set component size for MDT layout up to 1GB, size must be also aligned by 64KB. This is maximum possible size on MDT. Meanwhile there is another limit which is checked by `lfs setstripe` is provided by MDT server itself.

MDT server limits

The LOD parameter **`dom_stripesize`** is used to control per-server maximum size for DoM component. It is 1MB by default and can be changed with `lctl` tool. Check section **2.4 lctl and dom_stripesize parameter** in this guide.

2.2 lfs getstripe

`lfs getstripe` commands can be used to list the striping/component information for a given file. For DoM files that can be used to check its layout and size. The component ID for a DoM layout is always 1.

Commands

```
lfs getstripe [--component-id|-I [comp_id]] [--layout|-L] [--stripe-size|-S] <dirname|filename>
```

Examples

```
$ lfs getstripe -I1 <directroy|file>
```

```
$ lfs getstripe -I1 /mnt/lustre/domfile
/mnt/lustre/domfile
  lcm_layout_gen:  3
  lcm_entry_count: 2
    lcme_id:             1
    lcme_flags:          init
    lcme_extent.e_start: 0
    lcme_extent.e_end:   1048576
      lmm_stripe_count:  0
      lmm_stripe_size:   1048576
      lmm_pattern:       100
      lmm_layout_gen:    0
      lmm_stripe_offset: 2
      lmm_objects:
```

Quick check for DoM file:

```
$ lfs getstripe -I1 -L <directroy|file>
```

Below three files were created: **normfile** is ordinary file without composite layout, **compfile** is composite layout file without DoM component and **domfile** is file with DoM layout.

```
$ lfs getstripe -I1 -L /mnt/lustre/normfile
$ lfs getstripe -I1 -L /mnt/lustre/compfile
1
$ lfs getstripe -I1 -L /mnt/lustre/domfile
100
```

It returns numeric value where '1' is raid0 for a **compfile**, '100' is mdt for a **domfile** and returns nothing if file has ordinary layout like a **normfile**

Short info about layout and size of DoM component:

```
$ lfs getstripe -I1 -L -S <directroy|file>
```

```
$ lfs getstripe -I1 -L -E <directroy|file>
```

```
$ lfs getstripe -I1 -L -S /mnt/lustre/domfile
      lmm_stripe_size:   1048576
      lmm_pattern:       100
$ lfs getstripe -I1 -L -E /mnt/lustre/domfile
    lcme_extent.e_end:   1048576
      lmm_pattern:       100
```

Both command will return layout type and its size. Stripe size is equal to extent size of component in case of DoM files, so both can be used to get size on MDT.

2.3 lfs find

`lfs find` commands can be used to search the directory tree rooted at the given directory or file name for the files that match the given parameters. Here, only those parameters new for DoM files are showed and their usages are similar to `lfs getstripe` commands.

Commands

```
lfs find <directory|filename> [--layout|-L] [--stripe-size|-S]
```

Examples

1. Find all files with DoM layout under directory /mnt/lustre:

```
$ lfs find -L mdt /mnt/lustre
/mnt/lustre/domfile
/mnt/lustre/domdir
/mnt/lustre/domdir/domfile

$ lfs find -L mdt -type f /mnt/lustre
/mnt/lustre/domfile
/mnt/lustre/domdir/domfile

$ lfs find -L mdt -type d /mnt/lustre
/mnt/lustre/domdir
```

By using this command you can find all DoM objects, only DoM files or only directories with default DoM layout.

2. Find the DoM files/dirs with particular stripe size:

```
$ lfs find -L mdt -S -1200K -type f /mnt/lustre
/mnt/lustre/domfile
/mnt/lustre/domdir/domfile

$ lfs find -L mdt -S +200K -type f /mnt/lustre
/mnt/lustre/domfile
/mnt/lustre/domdir/domfile
```

First command finds all DoM files with stripe size less than 1200K, the second one does the same for files with stripe size greater than 200K. In both cases all our DoM files were found because their DoM size is 1MB.

2.4 lctl tool to manage `dom_stripesize` parameter

MDT controls default maximum DoM size on server via parameter **`dom_stripesize`** in LOD device. Its default value is 1MB and can be changed with `lclt` tool.

2.4.1 lctl get_param

Command

```
lctl get_param lod.*MDT<index>*.dom_stripesize
```

Examples

```
$ lctl get_param lod.*MDT0000*.dom_stripesize
lod.lustre-MDT0000-mdtlov.dom_stripesize=1048576


$ lctl get_param -n lod.*MDT0000*.dom_stripesize
1048576

$ lfs setstripe -E 2M -L mdt /mnt/lustre/dom2mb
Create composite file /mnt/lustre/dom2mb failed. Invalid argument
error: setstripe: create composite file '/mnt/lustre/dom2mb' failed:
Invalid argument
```

Command gets maximum allowed DoM size on that server. And an attempt to create file with bigger size on MDT is failed.

2.4.1 lctl set_param

Command is used to change default maximum value to DoM files on particular server.

Command

```
lctl set_param lod.*MDT<index>*.dom_stripesize=<value>
```

Examples

```
$ lctl set_param -n lod.*MDT0000*.dom_stripesize=$((1048576*4))
$ lctl get_param -n lod.*MDT0000*.dom_stripesize
4194304
$ lfs setstripe -E 2M -L mdt /mnt/lustre/dom2mb
$ lfs getstripe -I1 /mnt/lustre/dom2mb
/mnt/lustre/dom2mb
  lcm_layout_gen:  1
  lcm_entry_count: 1
    lcme_id:             1
    lcme_flags:          init
    lcme_extent.e_start: 0
    lcme_extent.e_end:   2097152
      lmm_stripe_count:  0
      lmm_stripe_size:   2097152
      lmm_pattern:       100
      lmm_layout_gen:    0
      lmm_stripe_offset: 0
      lmm_objects:
```

Here we change default DoM limit on server to 4MB and created file with 2MB DoM size successfully.

2.4.1 lctl conf_param

When default size limit on server should changed permanently then it is saved as config parameter with `lctl conf_param`

Command

`lctl conf_param <fsname>-MDT<index>.lod.dom_stripesize=<value>`

Examples

```
$ lctl conf_param lustre-MDT0000.lod.dom_stripesize=524288
$ lctl get_param -n lod.*MDT0000*.dom_stripesize
524288

--- remount Lustre server ---

$ lctl get_param -n lod.*MDT0000*.dom_stripesize
524288
```

In that case parameter is saved in config log permanently.

2.4.2 Disable DoM on server

When `lclt set_param` or `lctl conf_param` sets dom_stripesize to **`0`** value the DoM files creation will be prohibited on the selected server if needed.

Note

Nevertheless DoM files still can be created in existing directories with DoM default layout.