Technical Support Bulletin

EXAScaler Directory Striping Limitations and Workaround

ALERT! On EXAScaler file systems where directory striping is enabled on directories, ***also enable large directories on all MDTs.***

Issue Summary

With the Lustre Distributed NamespacE (DNE) feature, the metadata workload for a single file system can be distributed across multiple metadata targets (MDTs) to improve throughput. Two possible implementations for distributed metadata are supported: remote directories (known as DNE phase 1 or DNE1), and striped directories (known as DNE phase 2 or DNE2). Remote directories distribute different subdirectories across MDTs and are the preferred load balancing method for nearly all use cases. Striped directories stripe the contents of each subdirectory across multiple MDTs and are sometimes chosen for very large subdirectories that have millions of files created directly in them, or for the top-level directory in a very large file system.

Directory stripes are created in the internal REMOTE_PARENT_DIR directory on each MDT volume. The capacity of this internal directory is limited to 10 million entries by default, which may not be sufficient for the very large file systems with which striped directories are used. This problem is magnified when striped directories are enabled by default.

ALERT! ***Directory striping should NOT be enabled on directories by default.*** This can cause excessive growth in REMOTE_PARENT_DIR and severe issues with performance, function, or system stability.

If large directories (the `large_dir` feature) are not enabled on the MDTs when directory striping is enabled on a directory (that is, when the directory stripe count is set to -1 or to some positive number), significant issues with performance, function, or system stability may result. EXAScaler server logs may report error messages like the following:

```
kernel: LDISKFS-fs warning (device dm-14): ldiskfs_dx_add_entry:2629: Large
directory feature is not enabled on this filesystem
```

```
kernel: LDISKFS-fs warning (device dm-21): ldiskfs_dx_add_entry:2618: inode
626163226: comm mdt01_071: index 2: reach max htree level 2
```

File system clients may report "no data available" when copying or moving files, such as:

```
mv: cannot move 'make.log' to 'a/make.log': No data available
```

Affected Products

EXAScaler 5.x systems that support striped directories are affected by this issue.

Resolution

EXAScaler 5.2.3 and higher enable the large directory feature by default.

In addition, customers are advised to disable default directory striping after upgrade. (See “Workaround” for instructions.)

Workaround

NOTE This section has been updated in revision B0.

To work around this issue, enable large directories on all MDTs and disable default directory striping.

1. Stop the filesystem with the line command:

   ```
   esctl cluster --action stop
   ```

2. Add the large directory feature to each MDT individually with the command:

   ```
   tune2fs -O large_dir <MDT device>
   ```

   where *<MDT device>* is the name of the MDT device on which large directories should be enabled.

3. The MDT device will be deactivated when the filesystem is stopped and must be reactivated. Use the `vgchange` command to activate all logical volumes on the device. The syntax for the command is as follows:

   ```
   vgchange -ay vg_<MDT device>_<filesystem name> --config
      'activation{volume_list=[" vg_<MDT device>_<filesystem name>"]}'
   ```

   where *<MDT device>* is the name of the MDT device on which large directories should be enabled and *<filesystem name>* is the name of the stopped filesystem.

   For example, to reactivate all logical volumes of `mdt 0` on filesystem `tfs`, use the command:

   ```
   vgchange -ay vg_mdt0000_tfs --config 'activation{volume_list=["vg_mdt0000_tfs"]}'
   ```

4. Confirm that the large directory feature was enabled on the target MDT with the command:

   ```
   dumpe2fs -h <MDT device> | egrep "Filesystem features:"
   ```

5. Repeat Steps 2 through 4 on all MDTs in the storage cluster.

6. After large directories have been enabled on all MDTs, the logical volumes on the MDTs must be deactivated again to prevent errors on restart of the filesystem. Use the `vgchange` command with the `-an` option to do this. For example, to deactivate all logical volumes of `mdt 0` on filesystem `tfs`, use the command:

   ```
   vgchange -an vg_mdt0000_tfs --config 'activation{volume_list=["vg_mdt0000_tfs"]}'
   ```

7. Repeat Step 6 for all MDTs on the filesystem.

8. Restart the filesystem:

   ```
   esctl cluster --action start
   ```

9. On the client, or on the MDS if it has a local client mount point, disable default directory striping with the command:

```
# lfs find <mountpoint> -type d -print0 | xargs -0 lfs setdirstripe -d
```

This command can be run on a live system and should not impact normal usage. If interrupted, it is safe to re-run the command.

ALERT! Repeat Step 9 after upgrade to EXAScaler version 5.2.3 or later.

Contacting DDN Support

Please contact DDN Technical Support at any time if you have questions or need assistance. Support can be reached online, by email, or by phone as listed below.

Web

DDN Community Support Portal	https://community.ddn.com/login
Portal Assistance	webportal.support@ddn.com

Email

Support Email	support@ddn.com

Telephone

DDN Support Worldwide Directory	https://www.ddn.com/support/global-services-overview/

Bulletins & Notices

Support Bulletins	http://www.ddn.com/support/technical-support-bulletins
End-of-Life Notices	http://www.ddn.com/support/end-of-life-notices
Release Notes	https://community.ddn.com/login
Subscription Requests	support-tsb@ddn.com