Lustre filesystem performance benchmark on build workload

# Architecture

There are many ways to configure Lustre filesystem depending on the characteristics of the workload patterns. The following configurations have been tried for this study. More specifically, two features related to small file performance (<http://193.62.125.70/CIUK-2016/AdamRoe.pdf>) have been tried (Configuration 2 and Configuration 3):

* DNE Phase II – Allow metadata to scale, available from Lustre 2.8+ (<http://wiki.lustre.org/Lustre_Metadata_Service_(MDS)>)
* Data-on-MDT – Allow small data to be written directly on metadata server, available from Lustre 2.11+ (<http://wiki.lustre.org/Data_on_MDT_Solution_Architecture>)

No data striping is enabled for files, because 99.7% of the files from our test workload (GCC4.9.2 build) are smaller than 1MB in size (following Lustre best practice guidance).

Configuration 1:

Single metadata server (MDS) backed by one metadata target (MDT), four object storage servers (OSS) each backed by an object storage target (OST). Both MDT and OST are general purpose SSD formatted with LDISKFS (a variant of EXT4)

gp2 SSD

LDISKFS

MGT/MDT0

OST0

OST1

OST2

OST3





MGS/MDS

m5.xlarge

OSS

m5.xlarge

OSS

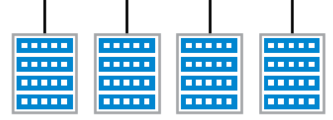
m5.xlarge

OSS

m5.xlarge

OSS

m5.xlarge



10GbE

Lustre Clients (1 - 40)

m5.xlarge

Configuration 2:

DNE II. Two or four MDS each backed by an MDT, four OSS each backed by an OST. All MDT and OST are general purpose SSD formatted with LDISKFS. File system directory metadata is distributed among two MDS with DNE II feature.

OST0

OST1

OST2

OST3

DNE MDT0 – MDT3

MGT/MDT0



…



MGS/MDS

m5.xlarge

MDS

m5.xlarge

MDS

m5.xlarge

OSS

m5.xlarge

OSS

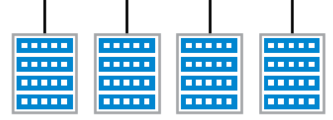
m5.xlarge

OSS

m5.xlarge

OSS

m5.xlarge



10GbE

Lustre Clients (1 - 40)

m5.xlarge

Configuration 3:

Data-on-MDT. Four MDS each backed by an MDT, four OSS each backed by an OST. All MDT and OST are general purpose SSD formatted with LDISKFS. File system directory metadata is distributed among four MDS. The difference between Configuration 3 and Configuration 2 is that with Configuration 3, the first 128KB of each file is configured to be placed on MDT.

## Software

|  |  |
| --- | --- |
| Client machine | CentOS 7.4.1708  3.10.0-693.21.1.el7.x86\_64  **Lustre Client 2.11.0**  **Disable network checksum**  *lctl set\_param osc./\*.checksums=0*  **Increase maximum RPCs in flight from 8 to 32**  *lctl set\_param osc./\*.max\_rpcs\_in\_flight=32*  **Increase maximum client side dirty page size from 32MB to 128MB**  *lctl set\_param osc./\*.max\_dirty\_mb=128* |
| Server machines | Metadata Server:  Each node has:  CentOS 7.4.1708  3.10.0-693.21.1.el7.x86\_64  **Lustre 2.11.0** |

# Application

|  |  |
| --- | --- |
| Workload  Characteristics | GCC4.9.2 build |
| Total number of files in the source tree | 84895 |
| Total storage space before build | 528MB |
| Total number of build tasks | 7851 |
| Average run time of each task | 0.43 seconds |
| Largest file generated after build | 225MB |
| Percentage of files smaller than 1MB after build | 99.7% |
| Percentage of files smaller than 64KB after build | 97% |
| Total storage space after build | 3.2GB |
| Level of sub-makes | 6 |

# Test Results

## Use Case 1: Copy

This use case tests the effect of multiple clients copy an entire GCC source code directory (528MB, 84895 files) from a local directory to the Lustre client mount at the same time under three different configurations:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Configurations  concurrent clients | Configuration 1:  Single MDS | | Configuration 2:  Multiple MDS with DNE.  No Data-on-MDT. | | | | Configuration 3:  Multiple MDS with DNE and Data-on-MDT | |
| Average copy time (seconds) | Max IOPS used | Average copy time  (seconds) | | Max IOPS used | | Average copy time | Max IOPS used |
| 2 MDS | 4 MDS | 2 MDS | 4 MDS | 4 MDS | 4 MDS |
| 1 | 105 | 300 | 134 | 151 | 286 | 644 | 135 | 450 |
| 10 | 135.3 | 2123 | 173.4 | 173.2 | 3900 | 4004 | 196.8 | 3700 |
| 20 | 155.1 | 5469 | 236.2 | 248.7 | 8602 | 9692 | 288.2 | 8018 |
| 30 | - | - | 354.6 | 353.3 | 11289 | 12449 | 416.1 | 9172 |
| 40 | - | - | 415.3 | 437.7 | 11700 | 11545 | 527.4 | 11400 |

## Use Case 2: Concurrent Build

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Configurations  concurrent clients | Configuration 1:  Single MDS | Configuration 2:  Multiple MDS with DNE.  No Data-on-MDT. | | Configuration 3:  Multiple MDS with DNE and Data-on-MDT |
| Average build time (seconds) | Average build time  (seconds) | | Average build time |
| 2 MDS | 4 MDS | 4 MDS |
| 1 | 1910 | 1504 | 1503 | 1487 |
| 10 | 1931 | 1480 | 1468.7 | 1494 |
| 20 | 2165.6 | 1558.1 | 1549.6 | 1520.6 |
| 30 | - | 1586.8 | 1582.1 | 1611.1 |
| 40 | - | 1691.5 | 1730.5 | 1706.7 |

## Use Case 3: Distributed Build with Elastic Make

20 build hosts with m5.xlarge instance. Each build host mounts the same Lustre volume.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Single MDS | 4 MDS with DNE.  No Data-on-MDT | 4 MDS with DNE and Data-on-MDT |
| 1 | 825 | 614 | 622 |
| 10 |  | 1284 |  |
| 20 |  | 2172\* |  |
|  |  |  |  |

\* 1 data corruption