Lustre* Tuning Parameters
High Performance Data Division

Bobbie Lind, Systems Engineer
April 16, 2013

* Other names and brands may be claimed as the property of others.
The mystery of tuning Lustre

• Everyone has their own secret sauce
• What parameters are tunable?
• What are the most common tuning parameters?
• How do I monitor those parameters while tuning?
• Where can I find this information?
  • Lustre Operations Manual
  • Online – Web searches
  • Phone a friend
  • HPDD SE’s are working on a tuning class
Multiple Layers to Tune

- Underlying OS
- Network (IB, eth, etc.)
- LNet
- Backend Storage
- Lustre Software
- Application I/O
How to find YOUR tuning sweet spot

• Determine what your data looks like
  • Large Files
  • Small Files
  • Video files
  • Other

• Benchmark your current system setup

• Monitor/Change Parameters

• Re-benchmark your system
Baseline Benchmark of Lustre 2.1.5

• Scenario
  • Small Lustre set up
    • 1 MDS
    • 4 OSS (1 OST per)
    • 2 Client
    • Ethernet
  • Sample 300MB files

• Benchmark tools used
  • IOZone
  • DD
  • MDTes
Baseline Benchmark of Lustre 2.1.5

**IOZone Baseline Run:**

```shell
./iozone -w -M -t 1 -s 300m -r 1m -i 0 -i 1 -F /lustre/scratch/iozone/test1 -R
```

- **Write:** 14184.41 KB/s
- **Re-Write:** 9725.66 KB/s
- **Read:** 2084858.62 KB/s
- **Re-Read:** 2188099

**DD Baseline Run:**

```shell
dd if=/dev/zero of=/lustre/scratch/ddtest/dd1 bs=1M count=300 oflag=direct conv=fdatasync
```

- 300+0 records in
- 300+0 records out
- 314572800 bytes (315 MB) copied, 29.3717 s, 10.7 MB/s

**MDTest Baseline Run:**

```shell
./mdtest -l 10 -i 5 -z 5 -b 2 -d /lustre/scratch/mdtest
```

**SUMMARY:** (of 5 iterations)

<table>
<thead>
<tr>
<th>Operation</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directory creation</td>
<td>472.990</td>
<td>368.132</td>
<td>447.905</td>
<td>40.049</td>
</tr>
<tr>
<td>Directory stat</td>
<td>510.034</td>
<td>425.078</td>
<td>491.362</td>
<td>33.180</td>
</tr>
<tr>
<td>Directory removal</td>
<td>241.112</td>
<td>145.006</td>
<td>219.004</td>
<td>37.092</td>
</tr>
<tr>
<td>File creation</td>
<td>248.578</td>
<td>194.355</td>
<td>235.320</td>
<td>20.586</td>
</tr>
<tr>
<td>File stat</td>
<td>253.255</td>
<td>202.930</td>
<td>239.983</td>
<td>18.683</td>
</tr>
<tr>
<td>File read</td>
<td>260.604</td>
<td>223.014</td>
<td>246.949</td>
<td>14.482</td>
</tr>
<tr>
<td>File removal</td>
<td>458.793</td>
<td>413.149</td>
<td>441.495</td>
<td>16.709</td>
</tr>
<tr>
<td>Tree creation</td>
<td>339.951</td>
<td>325.089</td>
<td>332.853</td>
<td>5.046</td>
</tr>
<tr>
<td>Tree removal</td>
<td>165.150</td>
<td>139.698</td>
<td>156.381</td>
<td>8.903</td>
</tr>
</tbody>
</table>
Common “Lustre Specific” Tuning Parameters

- **stripe size**
  - `Ifs getstripe <file or directory>`

- **max_rpcs_in_flight**
  - `/proc/fs/lustre/osc/*/max_rpcs_in_flight`

- **readcache_max_filesize**
  - `/proc/fs/lustre/obdfileter/*/readcache_max_filesize`

- **max_dirty_mb**
  - `/proc/fs/lustre/osc/*/max_dirty_mb`

- **max_read_ahead_mb**
  - `/proc/fs/lustre/llite/*/max_read_ahead_mb`
Monitoring Those Parameters

• RPCs
  • lctl get_param osc.*.import – realtime
  • lctl get_param osc.*.rpc_stats

• Stripe Size
  • dd if=/dev/zero of=/lustre/test/1stripe/testfile bs=1M count=300 oflag=direct conv=fdatasync

• Monitoring I/O Patterns (for readcache_max_filesize)
  • brw_stats (oss)
  • strace (application)

• Intel Manager for Lustre
Second Benchmark of Lustre 2.1.5

**IOZone Second Run:**

```
./iozone -w -M -t 1 -s 300m -r 1m -i 0 -i 1 -F /lustre/scratch/iozone/test1 -R
```

Write: 19218.00 KB/s  
Re-Write: 12180.47 KB/s  
Read: 2505728.50 KB/s

**DD Second Run:**

```/dd if=/dev/zero of=/lustre/scratch/ddtest/dd1 bs=1M count=300 oflag=direct conv=fdatasync
```

300+0 records in  
300+0 records out  
314572800 bytes (315 MB) copied, 28.5367 s, 11.0 MB/s

**MDTest Baseline Run:**

```
./mdtest -l 10 -i 5 -z 5 -b 2 -d /lustre/scratch/mdtest
```

**SUMMARY:** (of 5 iterations)

<table>
<thead>
<tr>
<th>Operation</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directory creation:</td>
<td>502.790</td>
<td>378.232</td>
<td>483.737</td>
<td>37.049</td>
</tr>
<tr>
<td>Directory stat</td>
<td>510.034</td>
<td>495.063</td>
<td>501.189</td>
<td>20.530</td>
</tr>
<tr>
<td>File creation</td>
<td>248.478</td>
<td>230.355</td>
<td>240.026</td>
<td>5.786</td>
</tr>
<tr>
<td>File stat</td>
<td>252.257</td>
<td>239.930</td>
<td>244.783</td>
<td>6.698</td>
</tr>
<tr>
<td>File read</td>
<td>258.444</td>
<td>240.054</td>
<td>249.418</td>
<td>12.483</td>
</tr>
<tr>
<td>File removal</td>
<td>458.623</td>
<td>441.161</td>
<td>450.324</td>
<td>8.506</td>
</tr>
<tr>
<td>Tree creation</td>
<td>346.499</td>
<td>339.089</td>
<td>341.174</td>
<td>3.821</td>
</tr>
<tr>
<td>Tree removal</td>
<td>168.150</td>
<td>152.234</td>
<td>159.508</td>
<td>9.904</td>
</tr>
</tbody>
</table>
Changes Between Benchmarks - IOZone

**Writes**

- First (light blue)
- Second (dark blue)

** Reads**

- First (light blue)
- Second (dark blue)
Changes Between Benchmarks - DD

Writes

Reads

First  Second

MB/s

11
10.9
10.8
10.7
10.6
10.5

12.2
12.15
12.1
12.05
12.0
11.95
11.9

MB/s
Changes Between Benchmarks - MDTest

Ops Per Second

- Dir Creation
- Dir Stat
- Dir Removal
- File Creation
- File Stat
- File Read
- File Removal
- Tree Creation
- Tree Removal

First
Second