

The Status of Lustre on Arm64

Xinliang Liu, Senior Engineer, Server, Linaro
MAD24 2024.5



Agenda

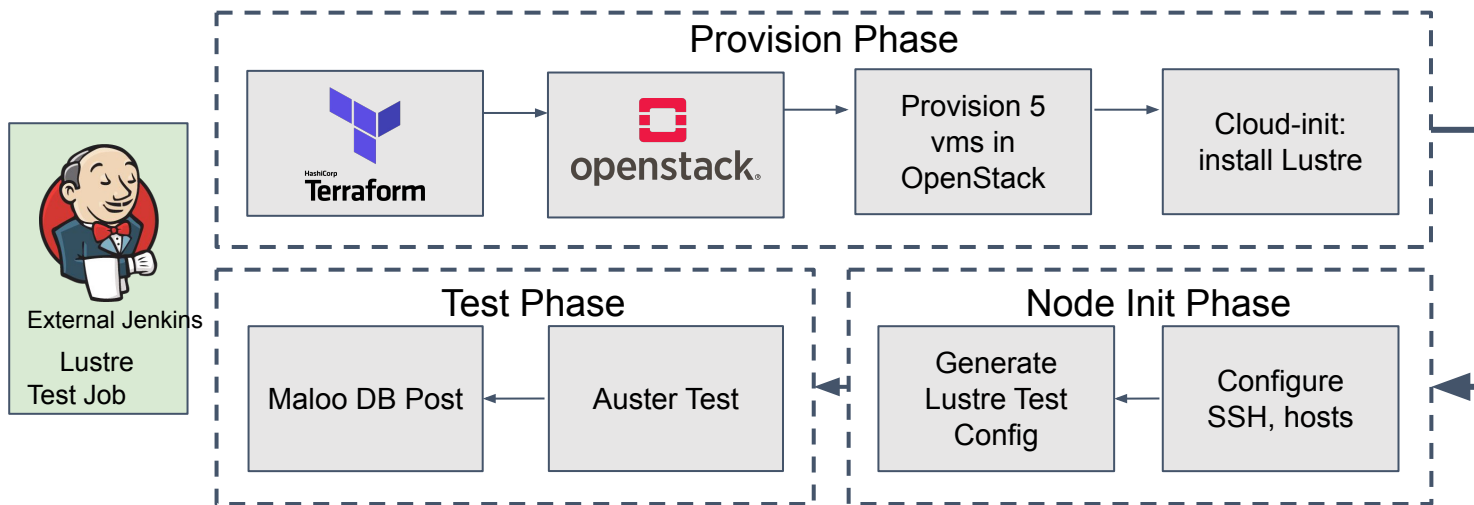
- Lustre on Arm64 Status
- IO500 Performance Test
- Future Plan

Lustre on Arm64 Status – Arm64 CI

Lustre upstream community Arm64 CI

- Only Arm64 client test
 - El8 x86_64 server end test with el8 Arm64 client.
 - Run basic sanity test suites per gerrit change.
- Need a complete Arm64 CI
 - Test both Arm64 server and client end.
 - Run full test suites.
 - Suggested an external Arm64 CI.

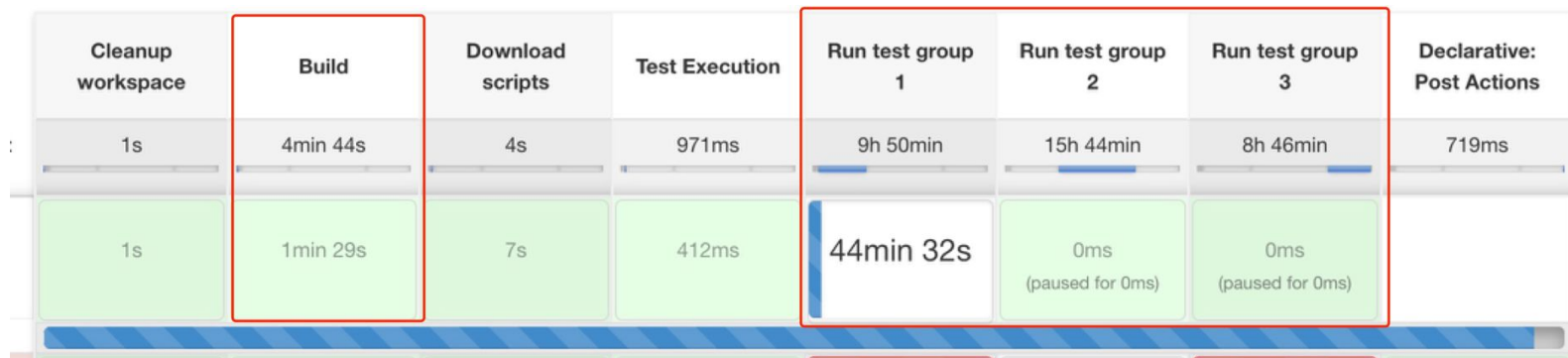
Lustre on Arm64 Status – Arm64 CI



CI Solution Inside

- **Provision Phase:** set up a 5 nodes cluster: 2 clients, 2 MDS, 1 OSS.
- **Node Init Phase:** set up test environment: ssh, multinode.sh.
- **Test Phase:** Run Auster and upload the data to Maloo DB.
- CI page: <http://213.146.155.72:8080/>
- Source code: <https://github.com/Linaro/lustretest/>

Lustre on Arm64 Status – Arm64 CI



CI Daily Test Pipeline

- One test pipeline per OS: RockyLinux el8, el9, openEuler 20.03, 22.03
- **Build:** Test master, v2.15 LTS.
- **Run test group**
 - Parallel run: linaro-full-part-1, linaro-full-part-2, linaro-full-part-3

Lustre on Arm64 Status – Arm64 CI

Test suites run

- Linaro-full-part-1
 - recovery-small, ost-pools, pjdfstest, sanity-lnet, mds-survey, sanity-dom, mmp, sanity-flr, replay-dual, runtests, metadata-updates, large-scale, replay-ost-single, obdfilter-survey, sanity-scrub
 - Total tests: **825**; Failed total: 2/825, **0.2%**; Skipped total: 4/825, 0.5%.
- Linaro-full-part-2
 - sanity, sanity-pcc, sanity-quota, parallel-scale-nfsv3, replay-vbr, sanity-sec, lustre-rsync-test, sanity-hsm, lnet-selftest
 - Total tests: **1283**; Failed total: 4/1283, **0.3%**; Skipped total: 14/1283, 1.1%.
- Linaro-full-part-3
 - sanityn, insanity, replay-single, sanity-benchmark, sanity-pfl, sanity-lfsck, parallel-scale-nfsv4
 - Total tests: **597**; Failed total: 3/597, **0.5%**; Skipped total: 3/597, 0.5%.

Defined: <https://github.com/Linaro/lustretest/blob/main/lustretest/cfg/test-groups.yaml>

Lustre on Arm64 Status – Arm64 CI

Download Arm64 rpms

- <https://www.lustre.org/download/>

Lustre RPM Packages

Lustre is available in RPM package form for a number of platforms.

- [Most Recent Lustre Release](#)
- [All Lustre Releases](#)

For servers, Lustre-specific e2fsprogs RPM packages are required.

- [Most recent Lustre-specific e2fsprogs](#)

Arm support

Linaro hosts builds for the latest [Lustre](#) and [e2fsprogs](#).

../	
e18/	13-Jun-2023 01:39
e19/	31-Jul-2023 08:12
oe2003sp3/	16-Aug-2023 09:06
oe2203sp1/	19-Jun-2023 15:47
oe2203sp2/	18-Jul-2023 17:04
oe2203sp3/	23-Jan-2024 04:50

Lustre on Arm64 Status – Bugfix

Server end critical issues

- Fixed and merged
 - [LU-15293](#): build: Add build support for arm64 centos8
 - [LU-15122](#): ASSERTION(iobuf->dr_rw == 0) crash
 - [LU-15412](#): sanity test 807 get stuck
 - [LU-15364](#): Kernel oops when stripe on Arm64 multiple MDTs
 - [LU-15978](#): fix striped directory deletion fails for 64K PAGE_SIZE
 - [LU-15722](#): fix write stuck for 64K PAGE_SIZE
- To be addressed(occur occasionally in io500 test)
 - [LU-16246](#): NULL pointer at lod_lookup+0x24/0x38
 - [LU-16245](#): ASSERTION(iobuf->dr_elapsed_valid == 0)

Lustre on Arm64 Status – Bugfix

Other test suites failed tests

- Fixed and merged: sanity, sanityn
 - [LU-16042](#): can not get cache size on Arm64
 - [LU-11787](#): Fix checkfilemap tests for 64K page
 - [LU-11597](#): Fix sanityn 16a failed on arm
 - [LU-11667](#): Fix sanity test 317 for 64K PAGE_SIZE OST
 - [LU-11596](#): Fix and re-enable sanity grant test for ARM
 - [LU-16275](#): Modify killall of replay-dual 25
 - [LU-11785](#): fix conf-sanity/98 mount check on 64K page
- To be addressed(Failed tests)
 - Conf-sanity(41c), sanity-lfsck(9a), sanity-pcc(7a, 7b)
 - Performance tests: performance-sanity.
 - Recovery tests: recovery-small, replay-single, replay-dual ...

Lustre on Arm64 Status – Summary

Arm64 status summary

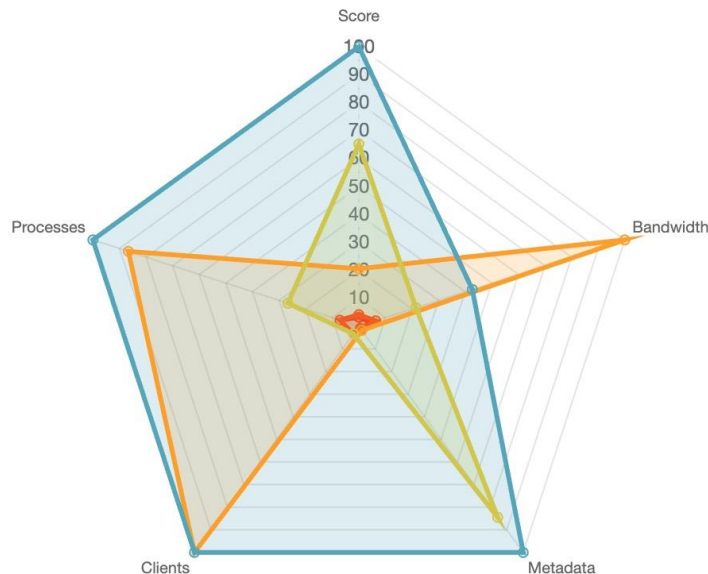
- Some vendors have already verified Arm Client years ago
- Released Lustre server/client rpms on openEuler
 - Client rpms on openEuler 22.03 SP2 2023-06
 - Server/Client rpms on openEuler 22.03 SP3 2023-12
- Arm64 CI pass ~95% tests on master/v2.15 (vm, ldiskfs, tcp)
 - Pass all the sanity, sanity test suites
 - Few failed tests to be fixed for other test suites
- Verified Lustre server/client on OSes:
 - el8, el9, openEuler 20.03, 22.03
- Time to verify/adopt Arm Lustre on production environment
 - Both Lustre server end and client end are functional on Arm64
 - Adopt Lustre client on compute nodes first is a good start

IO500 Performance Test - IO500



IO500 benchmark

- A distributed file system benchmark
- Test phases
 - IOEasy: Applications with well optimized I/O patterns
 - IOHard: Applications that require a random workload
 - MDEasy: Metadata/small objects
 - MDHard: Small files (3901 bytes) in a shared directory
 - Find: Finding relevant objects based on patterns
- Utilize test tools
 - lor, mdtest, pfind
- See more: <https://io500.org/about>



IO500 Performance Test - Testbed

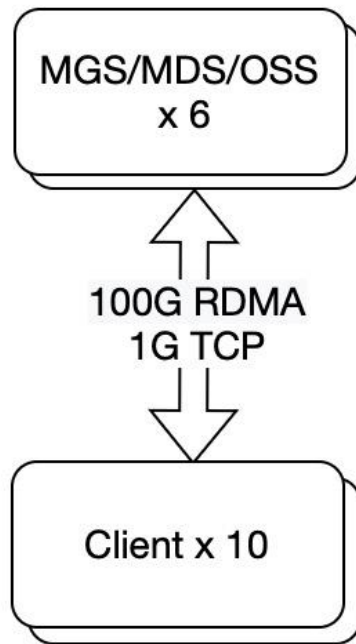
Hardware

6 servers : TaiShan 2280 V2

- CPU: Arm64 Kunpeng 920
- Disk(server): 4 x ES3000 V6 NVMe SSD 3.2T
 - 4k randread: IOPS=1527k, BW=5963MiB/s
 - 4k randwrite: IOPS=1174k, BW=4586MiB/s
 - 5 partitions per disk, one for MDT other for OSTs
- Network: 1x MLNX ConnectX-5 100Gb IB, 1x1G tcp

10 clients

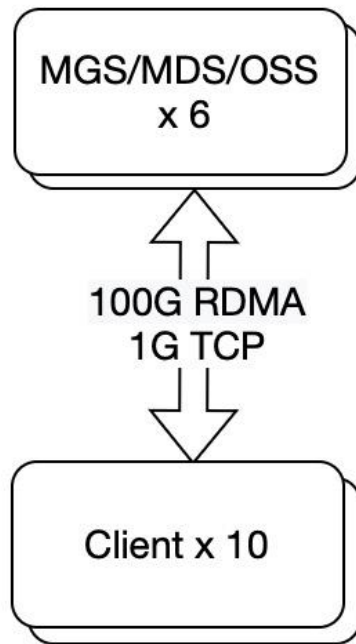
- CPU:
 - Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz, 96 cpus, 2 numa nodes
 - Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz, 112 cpus, 2 numa nodes
 - AMD EPYC 7642 48-Core Processor, 192 cpus, 2 numa nodes
- Network: 1x MLNX ConnectX-5 100Gb IB, 1x1G tcp



IO500 Performance Test - Testbed

Software

- OS: openEuler 22.03 LTS SP3, kernel 5.10.0-192.0.0.105.oe2203sp3
- Lustre: 2.15,
- io500: io500-isc24_v3, master
- openMPI: v4.1.x branch 4.1.7a1
- UCX: 1.16.0



IO500 Performance Test - 1 server

1 server + 9 clients (best score 21.7)

- cores_per_node: 4, np: 39
- Client(s): client 1-9
- Server(s): server1 Mdt: 4, ost: 16
- With MPIIO API for ior-hard-write/read test phases

```
IO500 version io500-isc24_v3 (standard)
[RESULT]      ior-easy-write      9.972334 GiB/s : time 437.266 seconds
[RESULT]      mdtest-easy-write   61.036971 kIOPS : time 462.976 seconds
[ ]           timestamp           0.000000 kIOPS : time 0.000 seconds
[RESULT]      ior-hard-write      1.643983 GiB/s : time 300.126 seconds
[RESULT]      mdtest-hard-write   20.270423 kIOPS : time 428.935 seconds
[RESULT]      find                3676.975451 kIOPS : time 10.027 seconds
[RESULT]      ior-easy-read       10.041841 GiB/s : time 433.831 seconds
[RESULT]      mdtest-easy-stat     125.393302 kIOPS : time 225.854 seconds
[RESULT]      ior-hard-read       5.287970 GiB/s : time 59.101 seconds
[RESULT]      mdtest-hard-stat     104.526170 kIOPS : time 83.973 seconds
[RESULT]      mdtest-easy-delete   62.883500 kIOPS : time 449.604 seconds
[RESULT]      mdtest-hard-read     48.954865 kIOPS : time 178.159 seconds
[RESULT]      mdtest-hard-delete   17.456735 kIOPS : time 498.136 seconds
[SCORE ]      Bandwidth 5.431867 GiB/s : IOPS 86.739654 kiops : TOTAL 21.706180
```

IO500 Performance Test - 6 servers

6 server + 9 clients (best score 58.8)

- cores_per_node: 31, np: 279
- Client(s): client1-9
- Server(s): server1-6 Mdt: 24, ost: 96

```
IO500 version io500-isc24_v3 (standard)
[RESULT]      ior-easy-write      31.782807 GiB/s : time 567.100 seconds
[RESULT]      mdtest-easy-write   277.729283 kIOPS : time 651.809 seconds
[RESULT]      timestamp          0.000000 kIOPS : time 0.000 seconds
[RESULT]      ior-hard-write      3.491108 GiB/s : time 2081.294 seconds
[RESULT]      mdtest-hard-write   18.728041 kIOPS : time 460.250 seconds
[RESULT]      find                7637.925323 kIOPS : time 24.780 seconds
[RESULT]      ior-easy-read       59.112669 GiB/s : time 290.268 seconds
[RESULT]      mdtest-easy-stat     594.959982 kIOPS : time 304.680 seconds
[RESULT]      ior-hard-read       11.486420 GiB/s : time 610.366 seconds
[RESULT]      mdtest-hard-stat    230.424709 kIOPS : time 38.294 seconds
[RESULT]      mdtest-easy-delete   307.762698 kIOPS : time 589.000 seconds
[RESULT]      mdtest-hard-read     136.657849 kIOPS : time 63.875 seconds
[RESULT]      mdtest-hard-delete   15.791519 kIOPS : time 545.331 seconds
[SCORE ] Bandwidth 16.567432 GiB/s : IOPS 208.832260 kiops : TOTAL 58.820185
```

IO500 Performance Test - Issues

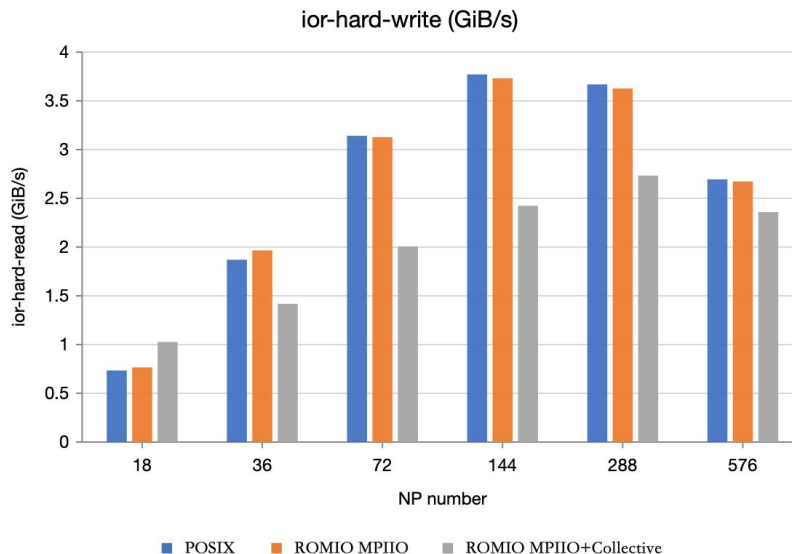
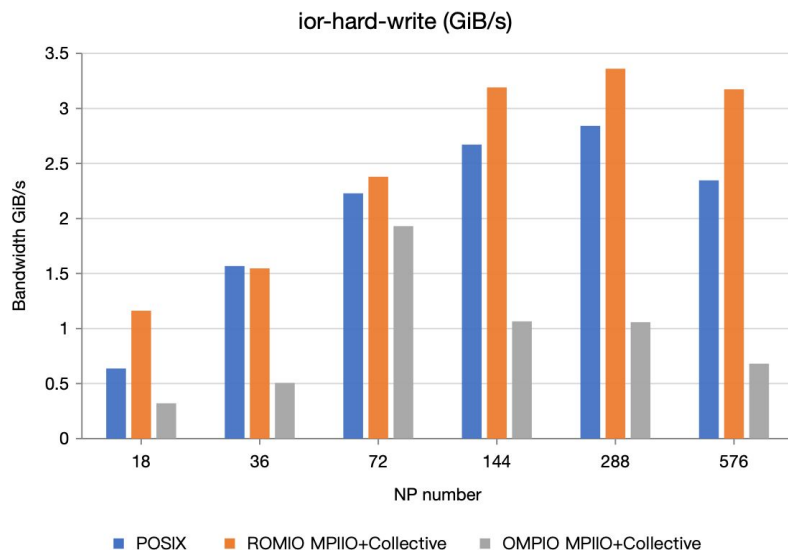
IO500 running issues

- [LU-16246](#): NULL pointer at lod_lookup+0x24/0x38
 - Occasionally, when run find/mdtest-hard-write phases
- [LU-16245](#): ASSERTION(iobuf->dr_elapsed_valid == 0)
 - Occasionally, when run mdtest-hard-write phase
- [LU-12832](#): watchdog: BUG: soft lockup - CPU#45 stuck for 22s!
 - Workaround
 - `lctl set_param ldlm.namespaces.*.lru_max_age=30000`
- `mpiexec "Fatal error in PMPI_Bcast: Unknown error class, error stack:"`
 - Fixed by stopping firewalld
 - `systemctl disable firewalld.service`
 - `systemctl stop firewalld.service`

IO500 Performance Test - Issues

ior-hard MPI-IO not perform as expected

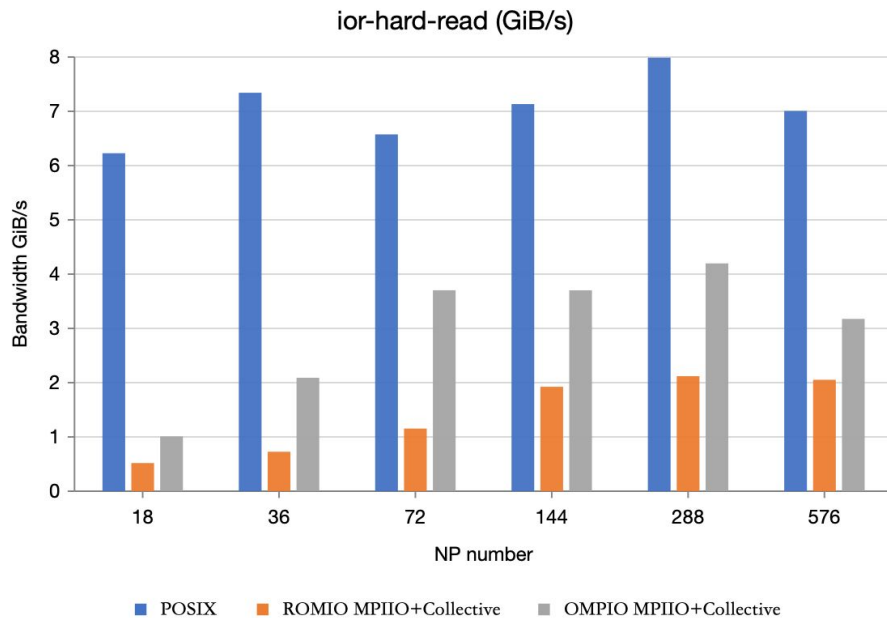
- Left figure 1 server test : ior-hard-write has some improvement for ROMIO not OMPIO
- Right figure 6 server test : ior-hard-write drop a lot compared to POSIX API
- Upstream discuss: <https://github.com/IO500/io500/issues/68>



IO500 Performance Test - Issues

lor-hard MPI-IO not perform as expected

- `ior-hard-write` drop too much on 1/6 server test compared to POSIX API



IO500 Performance Test - Tuning Tips

How to use MPI-IO API

- Build openMPI/mpich with Lustre FS support

OpenMPI: `./configure --with-lustre --with-io-romio-flags=--with-file-system=lustre`

Mpich: `./configure --with-file-system=lustre`

- io500 config

```
[ior-hard]
```

```
API = MPIIO
```

```
collective = TRUE
```

- io500.sh running

ROMIO:

```
io500_mpiargs="-hostfile /root/io500test/mpi-hosts --map-by node -np $np \
```

```
-mca pml ucx -mca btl ^openib \
```

```
-mca io romio321 -x ROMIO_FSTYPE_FORCE=lustre: \
```

```
"
```

OMPIO:

```
io500_mpiargs=" ... -mca fs lustre -mca fcoll dynamic_gen2 "
```

IO500 Performance Test - Tuning Tips

Lustre parameters tuning

- References
 - [Lustre and IO-500 Experiences with the Cambridge Data Accelerator](#)
 - [IO-500 A Storage Benchmark for HPC](#)
- Server end setting
 - Increase the RPC and inflight number for 100 Gib IB network
 - DoM lock tuning

obdfilter.*.brw_size=16

obdfilter.*.precreate_batch=1024

osp.*.max_rpcs_in_flight=128

mdt.*.dom_lock=trylock

debug=0

IO500 Performance Test - Tuning Tips

Lustre parameters tuning

- Client end setting
 - More aggressive RPCs to server
 - Readahead tuning
 - Avoid soft lockup
 - Disable checksum and debug

llite.*.max_read_ahead_mb=2048

llite.*.max_read_ahead_per_file_mb=32

llite.*.max_cached_mb=8192

mdc.*.max_rpcs_in_flight=128

osc.*.max_pages_per_rpc=4M

osc.*.max_rpcs_in_flight=256

osc.*.max_dirty_mb=2000

ldlm.namespaces.*.lru_size=4000000

ldlm.namespaces.*.lru_max_age=30000

osc.*.checksums=0 debug=0"

IO500 Performance Test - Tuning Tips

Lustre parameters tuning

- lo500 test dirs setting

```
# Dir stripping
```

```
if (( $(lfs df $workdir | grep -c MDT) > 1 )); then
```

```
    lfs setdirstripe -D -c -1 $workdir
```

```
fi
```

```
lfs setstripe -c 1 $workdir
```

```
...
```

```
# Try overstriping for ior-hard to improve scaling, or use wide striping
```

```
lfs setstripe -C $((osts * 4)) $workdir/ior-hard ||
```

```
    lfs setstripe -c -1 $workdir/ior-hard
```

```
# Try to use DoM if available, otherwise use default for small files
```

```
lfs setstripe -E 64k -L mdt $workdir/mdtest-easy || true #DoM?
```

```
lfs setstripe -E 64k -L mdt $workdir/mdtest-hard || true #DoM?
```

```
lfs setstripe -E 64k -L mdt $workdir/mdtest-rnd
```

Future Plan

- Keep maintaining the Arm64 CI
 - Fixing new found building issues and remain failed tests
- Address the io500 test issues
- Try advanced features on Arm
 - Caching: Lustre Persistent Client Cache, Metadata Write Back Cache
 - Erasure Coding
 - Compression and Encryption
- Support partners to adopt Lustre on Arm



Linaro Connect
MADRID 2024 | MAY 12-17 2024

Thank you

