Lustre on openEuler

Xinliang Liu, Senior Engineer, Server, Linaro – SIG SDS
openEuler SIG Gathering 2024 July

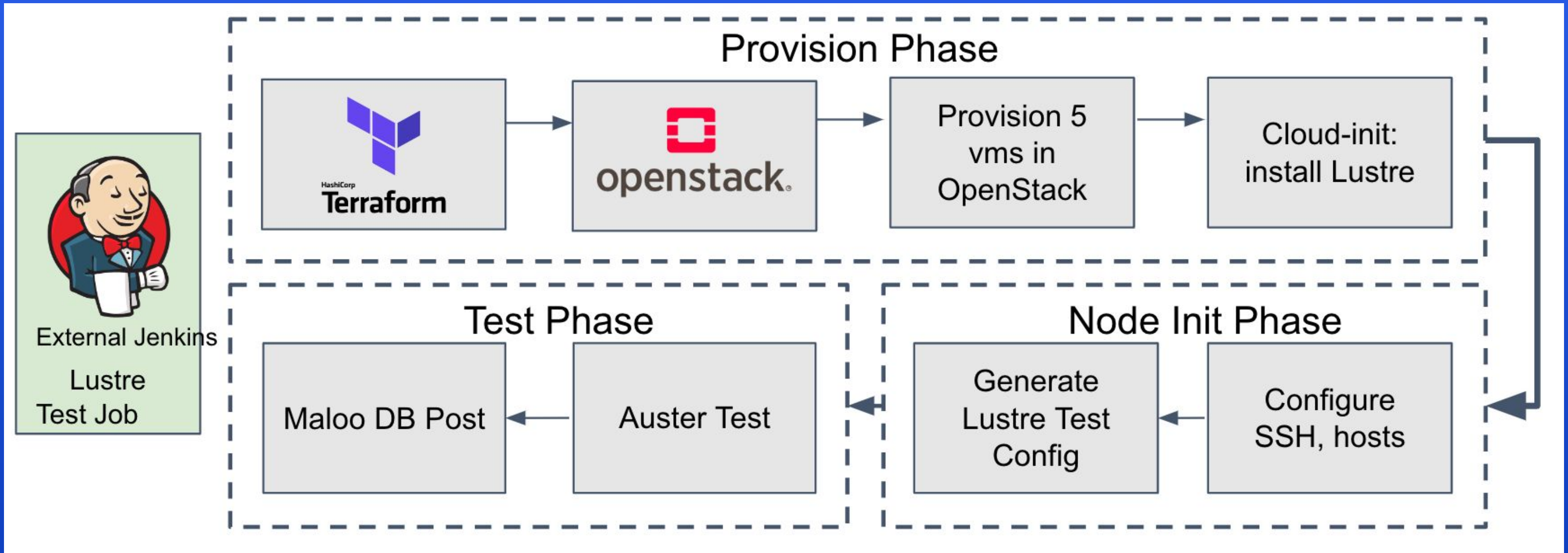
开放原子开源基金会
OPENATOM FOUNDATION | openEuler

Agenda

- Lustre on openEuler上游支持
- Lustre on openEuler版本发布
- Lustre on openEuler性能测试
- Lustre on openEuler总结

Lustre on openEuler上游支持

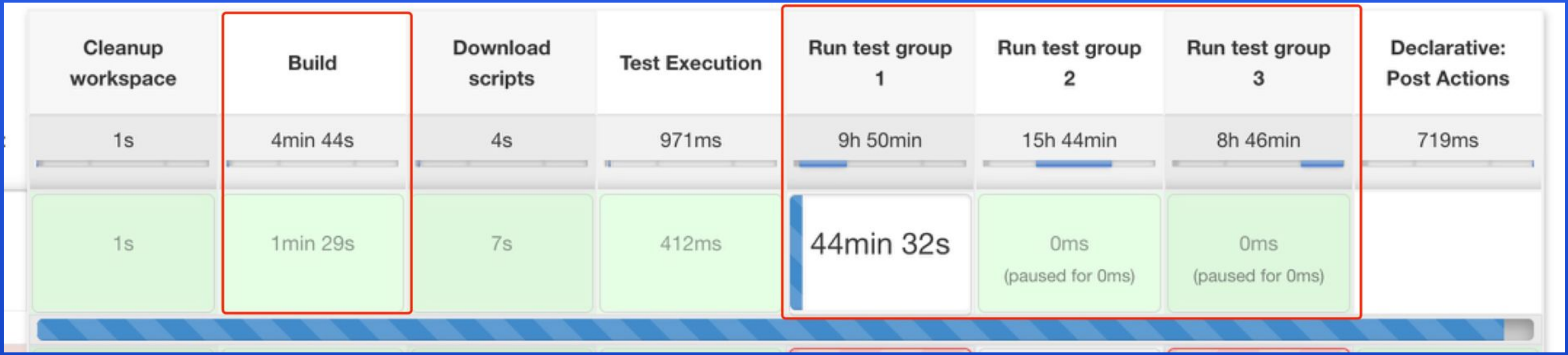
Lustre on openEuler上游支持 - External Arm CI



CI Solution Inside

- Jenkins test engine: run as containers
- 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/
- CI source code: https://github.com/Linaro/lustretest/

Lustre on openEuler上游支持 - External Arm CI



Cleanup workspace	Build	Download scripts	Test Execution	Run test group 1	Run test group 2	Run test group 3	Declarative: Post Actions
1s	4min 44s	4s	971ms	9h 50min	15h 44min	8h 46min	719ms
1s	1min 29s	7s	412ms	44min 32s	0ms (paused for 0ms)	0ms (paused for 0ms)	

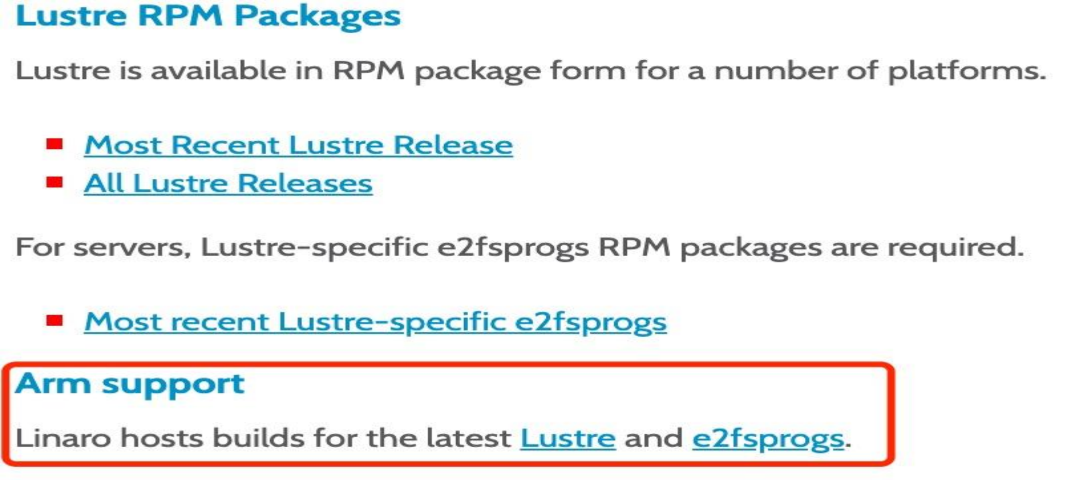
CI Daily Test Pipeline

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

Lustre on openEuler上游支持 - External Arm CI

Download upstream latest rpms

Link: https://www.lustre.org/download/



Lustre RPM Packages

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

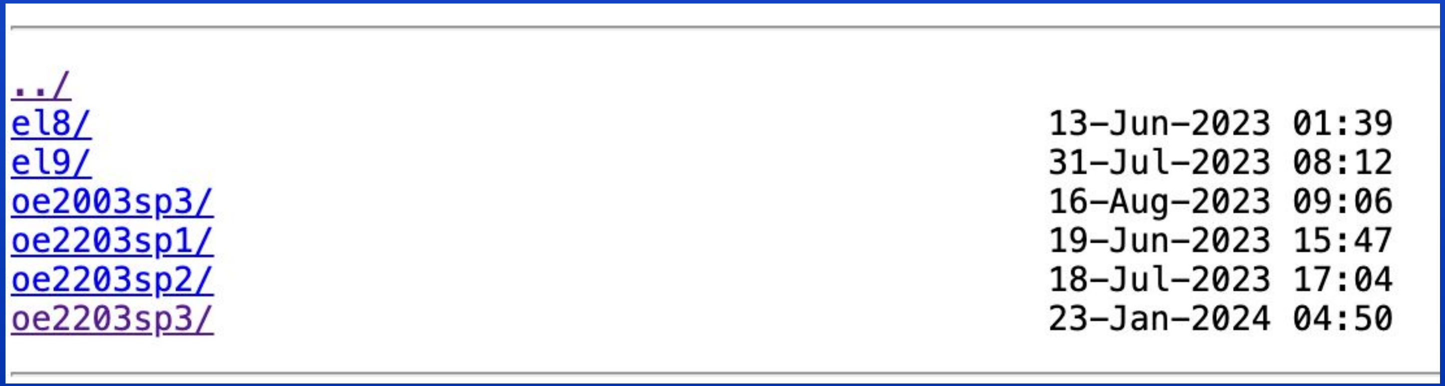
- Most Recent Lustre Release
- All Lustre Releases

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.



```
../
el8/                                    13-Jun-2023 01:39
el9/                                    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 openEuler上游支持 - Enablement

OpenEuler 22.03 LTS upstream support (Merged)

- LU-16322: add client build support for openEuler
- LU-16481: add server support for openEuler
- LU-16824: add server support for openEuler 22.03 LTS SP1
- LU-16976 ldiskfs: add support for openEuler 22.03 SP2
- LU-16862 rpm: set kmod-lustre-tests requires kmod-lustre explicitly
- LU-16610 ldiskfs: fix directory corruption on openeuler 22.03
- LU-16662 autoconf: fix configure test compile for CONFIG_KEYS
- Backports to b2_15 for openEuler 5.10 kernel
    - https://review.whamcloud.com/q/project:fs/lustre-release+branch:b2_15+owner:linaro.org
- OpenZFS build support for openEuler
    - https://github.com/openzfs/zfs/pulls?q=is%3Apr++is%3Aclosed+openeuler
- e2fsprogs
    - LU-16337: build rpms support for openEuler

Lustre on openEuler上游支持 - Enablement

OpenEuler 20.03 LTS upstream support

- LU-17028 ldiskfs: add support for openEuler 20.03 LTS (Reviewing)
- LU-17052 libcfs: fix build for old kernel (Merged)
- LU-17027 target: include linux/file.h (Merged)
- E2fsprogs (Merged)
  - LU-17028 build: add support for openEuler 20.03 LTS
- Derived OS support, like KylinOS (Merged)
  - LU-17029 lustre.spec.in: match rpm macro openEuler for openEuler Linux

Lustre on openEuler上游支持 - Enablement

OpenEuler 24.03 LTS upstream support (WIP)

- Upstream kernel 6.6 support status (master merged)
  - Client support: https://review.whamcloud.com/c/fs/lustre-release/+/52908/
  - Server support: https://review.whamcloud.com/c/fs/lustre-release/+/52919/9
  - Lustre 2.16 release soon, should contain kernel 6.6 support.
- openEuler 24.03 LTS support (WIP)
  - Ldiskfs patch set
  - gcc 12.03 warning fix

Lustre on openEuler版本发布

Lustre on openEuler颁布发布

OpenEuler 22.03 LTS SP2 cycle

- Gcc compile error
  - Issue #I5XMD0: stringop-overflow error
  - Issue #I5T8DL: "multiple definition of 'enum fsconfig_command'" error
- PR-123: Kmod build fixed
- STOR-180: zfs rpm pkgs on openEuler
  - openEuler Master only
  - V2.1.10 stable release
  - Backport openEuler build support patches
- STOR-179: e2fsprogs rpm pkgs on openEuler
  - openEuler Master only
  - V1.46.5.wc1 with backport openEuler build support patches
  - Multi-version rpm separated from existing e2fsprogs
- Ticket #I6T8OP: Lustre client rpm
  - V2.15.2 stable release
  - With backport openEuler support patches
  - https://repo.openeuler.org/openEuler-22.03-LTS-SP2/EPOL/main/aarch64/Packages

Lustre on openEuler颁布发布

OpenEuler 22.03 LTS SP3 cycle

- Multi-version solution due to e2fsprogs pkg
  - CI project: openEuler_22.03_LTS_SP3_Epol_Multi-Version_lustre_2.15
  - Install: dnf install lustre-release && dnf install lustre
- E2fsprogs
  - V1.47.0 wc5 with Lustre patches
  - Multi-version rpm to original e2fsprogs
- Ticket #I80F3Y: both Lustre server and client rpm
  - V2.15.3 stable release
  - With backport openEuler support patches
  - https://repo.openeuler.org/openEuler-22.03-LTS-SP3/EPOL/multi_version/lustre/

Lustre on openEuler颁布发布

OpenEuler 22.03 LTS SP4 cycle

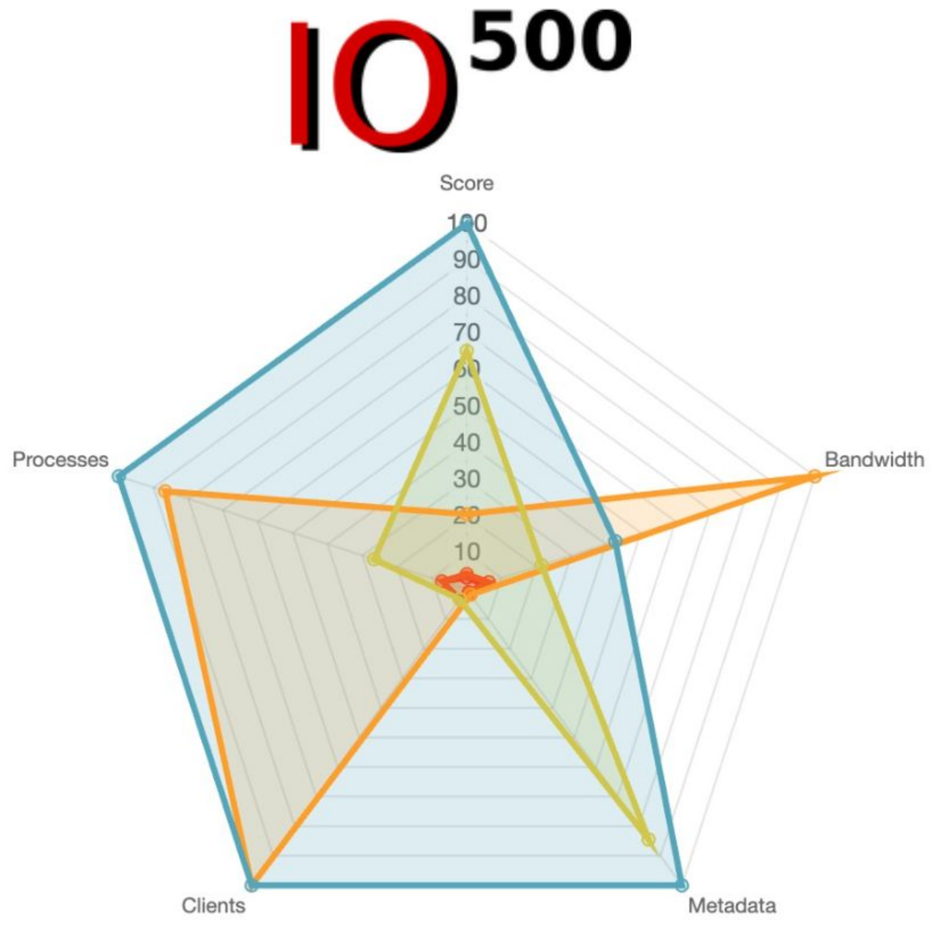
- Multi-version solution due to e2fsprogs pkg
  - CI project: openEuler_22.03_LTS_SP4_Epol_Multi-Version_lustre_2.15
  - Install: dnf install lustre-release && dnf install lustre
- E2fsprogs
  - V1.47.0 wc6 with Lustre patches
  - Multi-version rpm to original e2fsprogs
- Ticket #I9RSL5: both Lustre server and client rpm
  - V2.15.4 stable release
  - With backport openEuler support patches
  - https://repo.openeuler.org/openEuler-22.03-LTS-SP4/EPOL/multi_version/lustre/

Lustre on openEuler性能测试

Lustre on openEuler性能测试 - 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
  - Ior, mdtest, pfind
- See more: https://io500.org/about



Lustre on openEuler性能测试 - Testbed

Hardware

6 servers： TaiShan 2280 V2

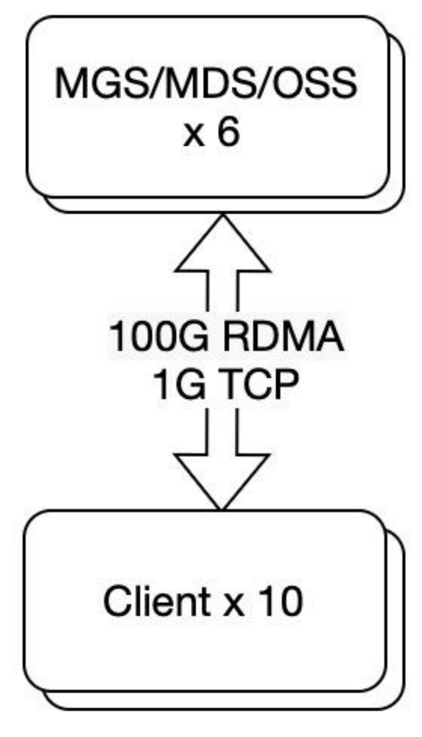
- CPU: Arm64 Kunpeng 920
- Disk(server): 4 x ES3000 V6 NVMe SSD 3.2T
  - 4k randread: IOPS=1527k, BW=5963MiB/s
  - 4k randwrite: IOPS=1174k, BW=4586MiB/s
  - 5 partitions per disk, one for MDT other for OSTs
- Network: 1x MLNX ConnectX-5 100Gb IB，1x1G tcp

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

Software

- OS: openEuler 22.03 LTS SP3， kernel 5.10.0-192.0.0.105.oe2203sp3
- Lustre: 2.15.4,
- io500: io500-isc24_v3, master
- openMPI: v4.1.x branch 4.1.7a1
- UCX: 1.16.0



Lustre on openEuler性能测试 - Issues

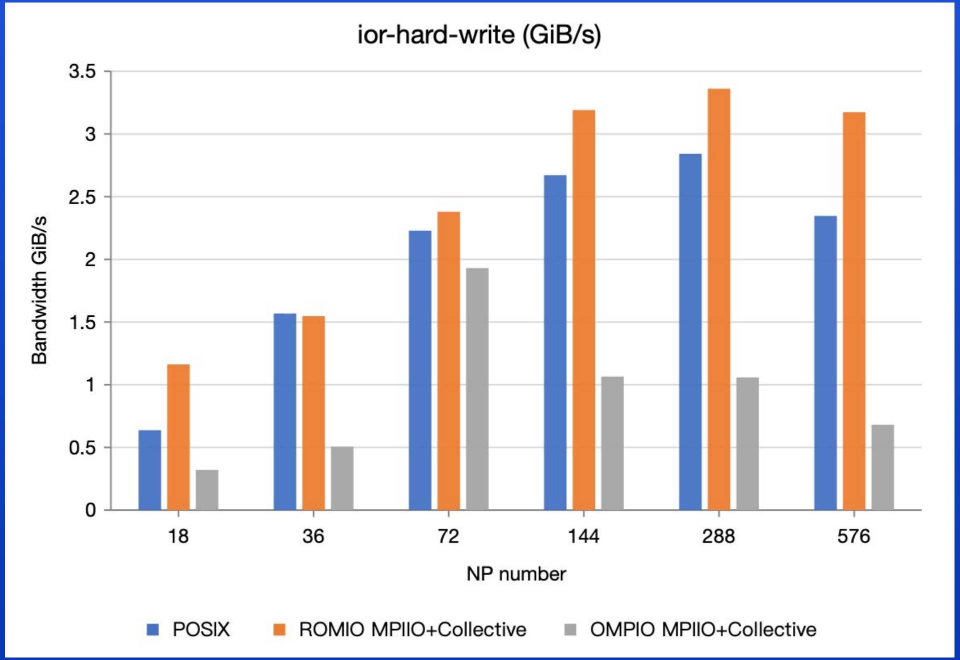
IO500 running issues

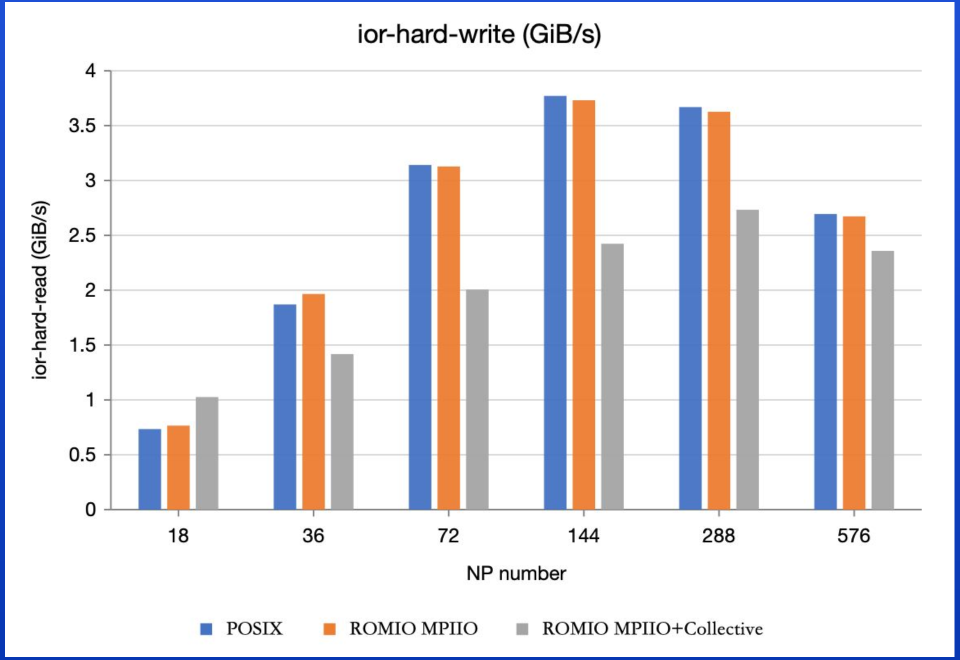
- LU-16246: NULL pointer at lod_lookup+0x24/0x38
  - Occasionally, when run find/mdtest-hard-write phases
- LU-16245: ASSERTION(iobuf->dr_elapsed_valid == 0)
  - Occasionally, when run mdtest-hard-write phase
- LU-12832: watchdog: BUG: soft lockup - CPU#45 stuck for 22s!
  - Workaround
    - lctl set_param 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

Lustre on openEuler性能测试 - Issues

Ior-hard MPI-IO not perform as expected – ior-hard-write

- 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
  - Has improvement when using lower speed tcp and disk
  - But not for high speed IB and NVMe disk.

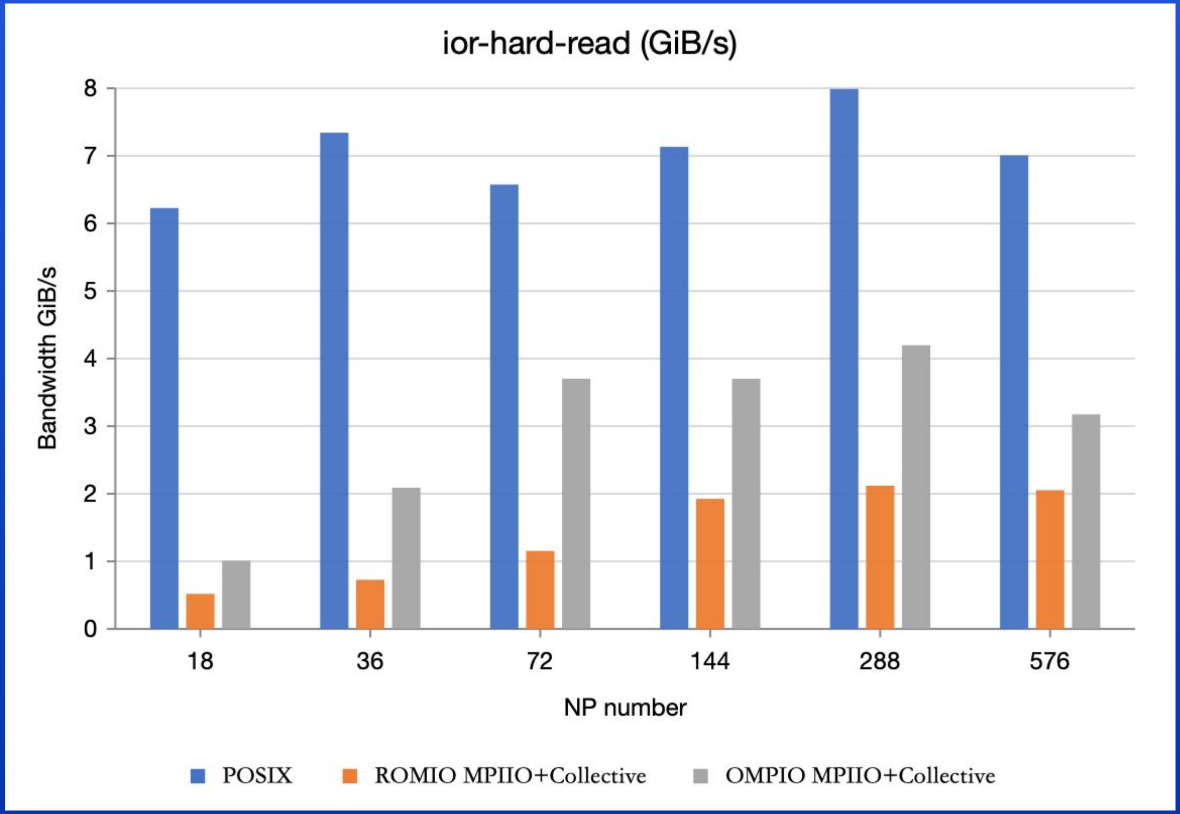




Lustre on openEuler性能测试 - Issues

Ior-hard MPI-IO not perform as expected – ior-hard-read

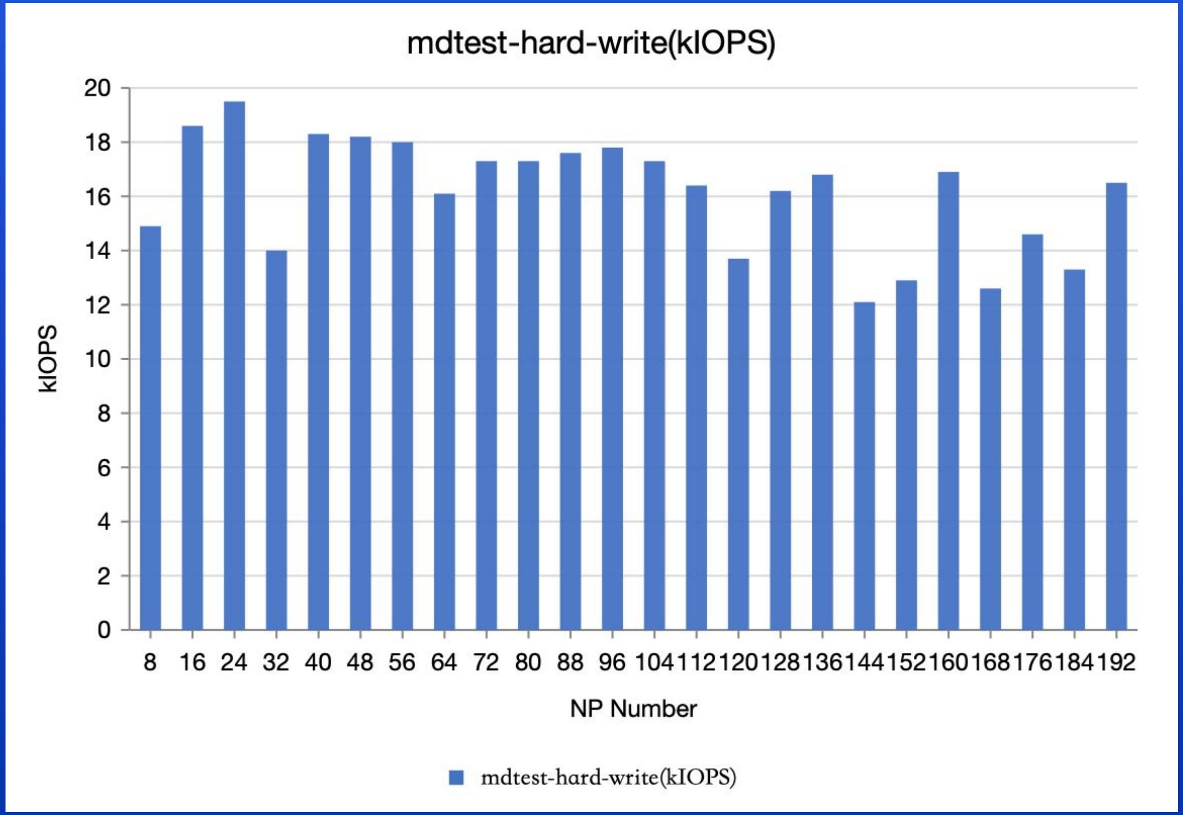
- ior-hard-read drop too much on 1/6 server test compared to POSIX API



Lustre on openEuler性能测试 - Issues

mdtest-hard-write/delete performance is poor and does not scale

- The mdtest-hard-write IOPS perfmances is poor less than 20 KIOPS.
- It does not scale, it won’t increase as the threads and server nodes increase.
- Mdtest-hard-delete test has the same issue.



Lustre on openEuler性能测试 - Tips

How to use MPI-IO API tips

- Build opemMPI/mpich with Lustre FS support

  OpemMPI: ./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 “
  ```

Lustre on openEuler性能测试 - Tips

Lustre parameters tuning tips – server end

- 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
```

Lustre on openEuler性能测试 - Tips

Lustre parameters tuning tips – client end

- 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=16M

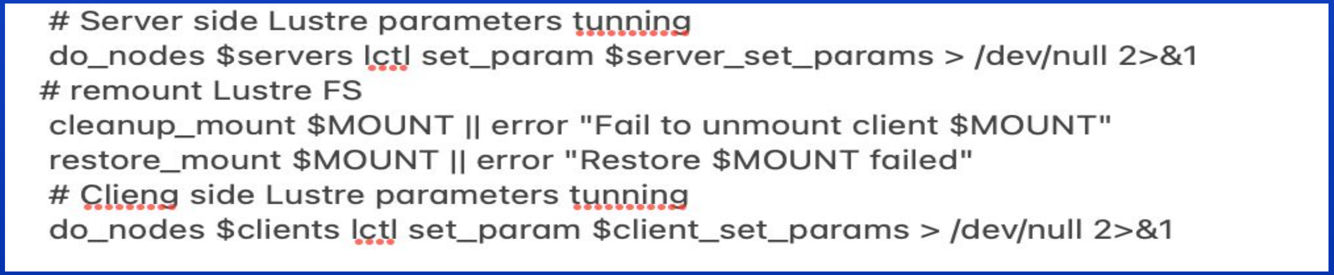
  osc.*.max_rpcs_in_flight=256

  osc.*.max_dirty_mb=2000

  ldlm.namespaces.*.lru_size=4000000

  ldlm.namespaces.*.lru_max_age=30000

  osc.*.checksums=0 debug=0"
- Need to remount clients for setting “osc.*.max_pages_per_rpc=16M”



```
 # Server side Lustre parameters tunning
 do_nodes $servers lctl set_param $server_set_params > /dev/null 2>&1
# remount Lustre FS
 cleanup_mount $MOUNT || error "Fail to unmount client $MOUNT"
 restore_mount $MOUNT || error "Restore $MOUNT failed"
 # Clieng side Lustre parameters tunning
 do_nodes $clients lctl set_param $client_set_params > /dev/null 2>&1
```

Lustre on openEuler性能测试 - Tips

Lustre parameters tuning tips – test dirs

- Io500 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
```

Lustre on openEuler总结

总结

- Ldiskfs patchset 基于 ext4, 需要跟随 ext4更新
- Arm64 CI pass ~90% tests on master/v2.15 (vm, ldiskfs, tcp)
  - 基本的 sanity, sanityn test suites 全pass
  - 少数 failed tests to be fixed for other test suites
  - 已验证 OSes: openEuler LTS 20.03, 22.03, el8, el9
- 部署指导：https://docs.openeuler.org/zh/docs/22.03_LTS_SP4/docs/lustre/user-guide.html
- Lustre介绍文档
- Lustre编译指导文档
- Lustre性能测试文档
- Contact me: xinliang.liu@linaro.org

THANKS

开放原子开源基金会
OPENATOM FOUNDATION | openEuler