Linux Lustre client state

A status update, Sept 2022

James Simmons

Storage Systems Engineer

Oak Ridge National Laboratory

ORNL is managed by UT-Battelle LLC for the US Department of Energy

The project that wouldn’t die

- One of the the oldest project
  - Pushed 8+ years ago upstream.
  - Removed from upstream due to lack of involvement.
    - Also staging was wrong fit.
- Limited resources, limit support
  - Only RHEL x86 supported by whamcloud.
  - Community involvement has kept it alive.
  - Rotating support.
- Close to the final lap.
  - Synced to OpenSFS tree.
  - Will submit upstream once IPv6 work is complete

Progress over the last year.

- Kept in sync with tip of OpenSFS master branch
- Flow of work from Linux client to OpenSFS branch
  - Faster support of newer kernels
  - Rapid support of newer distros (Ubuntu22 / RHEL9 for example)
  - Support for latest MOFED stacks
  - Performance gains (LU-11089, LU-8130)
- At Linux 5.9 version with work to move to 5.15 – delayed due to fscrypt
  - Patch in the works
- Lustre community effort
  - Neil Brown from SUSE
  - James Simmons from ORNL
  - Others

How healthy is the Linux client ?

- Same testing as other community projects (ARM, Ubuntu)
  - Manually running test suite from OpenSFS master branch
  - Can build Lustre’s utilities for native Linux client
    - configure –disable-server –disable-modules
    - We can enable automatic testing
      - Need to work out test system. Work already done for external ARM support (LUG 2021).
  - sanity-lnet and sanity test
    - Mostly same bugs between both trees.
    - Largest source of failures in Linux client is FID lookup cache (LU-9868 / LU-11501)
      - Patch in the works at https://review.whamcloud.com/#/c/44846
  - Resolving other failures in the test suite (bug squashing mode, unique bugs)
  - Handle occasional UAPI breakage in OpenSFS tree.

The final touches !!!!

- What is left - https://jira.whamcloud.com/projects/LU/versions/12991
  - Some things are big changes
  - LU-12511 also tracks this work
- Last barrier to pushing to Linus tree
  - LNet IPv6 support (LU-10391). Large chuck done.
- IB support is a must have
  - ko2iblnd is disliked by infiniband developers (LU-8874)
- Squash as many bug as possible as testing expands
  - Linux client exposes unique bugs

Big ticket items left for OpenSFS tree.

- Remove /proc usage (LU-8066)
  - Implement Netlink to replace complex debugfs (LU-9680)
  - Enforce proper sysfs naming (LU-13091)
  - Native linux client already doesn’t use /proc
- Migration to rhashtable + Xarray (LU-8130)
- Make sysfs file names ASLR compliant (LU-13118)
- Rework mount code (LU-12541)
- Proper fid lookup cache (LU-9868 / LU-11501 / LU-8585)

Visible Benefits

- Udev rules (sysfs) for tunables
  - Today you can do :
    ```
    SUBSYSTEM=="lustre", ACTION=="add", DEVPATH=="*MDT*", ATTR{max_rpcs_in_flight}="64"
    SUBSYSTEM=="lustre", ACTION=="add", DEVPATH=="*OST*", ATTR{max_dirty_mb}="2000", ATTR{max_rpcs_in_flight}="64",
    ATTR{checksums}="0", RUN+="/bin/bash -c 'sleep 1; /usr/sbin/lctl set_param *.*.max_dirty_mb=2000'"
    SUBSYSTEM=="lustre", ACTION=="add", DEVPATH=="*llite*", ATTR{max_read_ahead_mb}="512",
    ATTR{max_read_ahead_per_file_mb}="512"
    ```
  - Client eviction reporting (LU-10756 for Lustre 2.16)
  - LNet health events
- Unified sysfs naming using UUID. Currently varies between nodes and across reboots. (LU-13118)
- mount –t lustre_target /dev/sda /mnt/OST
  - Will start up and shutdown LNet when mounting server disks
- Working on fhandle and filesets
- Using genradix tree to allocate large data sets (LU-15058)
- Fix filesets and fhandle API.
- Use Netlink for Lustre stats (LU-11085).

LNet changes coming

- IPv6 + IB hardware address support

  -
    ```
    lctl list_nids
    fe80::a242:3fff:fe38:abfe@tcp
    ```
- New Netlink YAML API means no more backwards compatibility issues.
- LNet selftest using YAML (Netlink + IPv6 support)
- Use LNet discovery when mounting (LU-10360).

What the future holds

- Once merged into Linus tree it will show up in newer distros
  - SUSE will provide good support
  - Ubuntu is an unknown (closest to upstream). Heavy demand
- Discuss having external testing / bug triage outside whamcloud.
- Goal is new developers will enter the community
- Kernel improvement needed by Lustre can be accepted. (fscrypt)
- Entire Lustre OpenSFS tree will be moved to Linux kernel
  - Remove the need to patch ext4 (LU-6202)
    - https://patchwork.kernel.org/patch/10695037
  - All backport changes from Upstream are applied to entire OpenSFS tree.
  - Move to Linux kernel will be much smaller leap

Lustre community involvement

- Prepare for upstream merge in 2.16 time frame
- We need greater scope of Lustre testing
  - testing exposes unique bugs
- How do you test?
  - https://github.com/jasimmons1973/lustre
  - http://wiki.lustre.org/Testing
  - Report bugs at https://jira.whamcloud.com/secure/Dashboard.jspa
    - Add upstream label so we can see it
- Questions ?
  - http://lists.lustre.org/listinfo.cgi/lustre-devel-lustre.org
- Company Involvement
  - http://wiki.opensfs.org/Lustre_Working_Group
- Lustre conferences [LAD (conference), LUG (US and/or China)]

Conclusions

- Lustre Linux client mostly works
- Lustre Linux client is kept up to date.
- Very very close to merging to Linus tree (should be last LAD talk)
- Requires community involvement for proper support
  - Join OpenSFS ▯ - http://opensfs.org/
  - Don’t be afraid to ask questions or report problems
  - LWG calls
  - Lustre-devel mailing list
  - Report on Whamcloud JIRA
  - Contact me directly   jsimmons@infradead.org

Acknowledgements

This work was performed under the auspices of the U.S. DOE by Oak Ridge Leadership Computing Facility at ORNL under contract DE-AC05-00OR22725.