Preparing the Linux native client submission

Settling the last details, May 2024

James Simmons

Storage Systems Engineer

Oak Ridge National Laboratory

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

Requirements and improvements for submission

- libcfs cleanup (in sync with upstream)
  - LU-9859 : cleanup of the code. Mostly done. Only crypto handlers left (hard requirement)
    - LU-17053 : Need to load LNet to use lctl debugging. Can use debugfs file instead of lctl markers.
    - LU-16796 : Remove LASSERT_ATOMIC_* debug macros. Cleanup libcfs_private.h. Done with extra work being done by Arshad.
  - LU-14428 : move tracefile to kernel ring_buffer
    - LU-16746 : live watching of debug messages like dmesg –w
    - Request to create crash utility plugin to filter out lustre logs
  - LU-14290 : Use fault-inject kernel API instead of a custom one. (not hard requirement)
  - LU-8130 : replace cfs_hash with rhashtable
    - Only ldlm cfs_hash table left for client side.
      - https://review.whamcloud.com/c/fs/lustre-release/+/45882

LNet requirements and improvements

- Main barrier to upstreaming is lack of IPv6 support (hard requirement)
  - LU-10391 foundation support is complete!!!!
  - Largest change required
- Simplify o2iblnd (LU-8874)
  - Nvidia GDS makes this harder ☹
  - Upstream hates o2iblnd
  - Disagreement on should we submit it with native client
- Backport upstream changes to OpenSFS branch
  - LU-12678 tracks this work upstream. Need to drop RHEL7 first.
    - LU-14633 tracks Al Viro’s iter_iov changes.

Lustre changes in the works for submission

- Resolving regressions in native client
  - LU-11085 replacing Lustre’s interval tree with kernels
    - Changes upstream introduce performance regressions
    - Massive development in this area recently to resolve this.
    - kunit test introduced
  - LU-11501 / LU-9868 dcache issues
    - Upstream changes broke .lustre/fid/”FID” handling
    - RHEL7 doesn’t work with these changes.
    - Real dcache bugs are showing up.
    - Resolution to issues are being worked on !!!! Also fixes some fileset issues.
      - https://www.review.whamcloud.com/c/fs/lustre-release/+/37013

Handling the death of /proc upstream

- Native Linux client has completely removed /proc
  - Move to debugfs prevents normal users some normal opertains (pool list, stats, target_obds)
  - OpenSFS has delayed this move.
    - Verify move doesn't break anything (LU-13091)
- Restore non root access (LU-11850)
  - Netlink solution for stats
    - https://www.review.whamcloud.com/fs/lustre-release/+/34256
    - Create wrappers to simplify this approach (LU-17472)
    - Enhancement to get sub fields of YAML output (LU-12841)
  - Need to do pool list and targets_obd as well. Have local early patches
    - https://www.review.whamcloud.com/fs/lustre-release/+/51959

Nice to have but not required

- Complete new server mount type (LU-12541)
  - mount –t lustre_tgt ….
  - Working in Native client but not server side for OpenSFS
- Make sysfs / debugfs files ALSR compliant (LU-13118)
  - Enable mounting with UUID
- Removed cached mount point (LU-10824)
  - https://www.review.whamcloud.com/c/fs/lustre-release/+/45608
- Use Netlink for HSM (LU-7659)

Minimize the difference between trees

- Meet Linux kernel code standards (LU-6142)
  - Many cleanups still underway
  - Checkpatch update
    - https://www.review.whamcloud.com/c/fs/lustre-release/+/54154
    - https://www.review.whamcloud.com/c/fs/lustre-release/+/54153
- Move to sphinx doc style (LU-9633)
  - Most neglected ☹
- OpenSFS branch support for newer gcc and kernels
  - Lacks supporting newer fscrypt API. Native client is lagging currently.
- Handle native clients with OpenSFS server stack (LU-8837 / LU-14291)

Special thanks

- Native client support is a true community effort
  - Neil Brown (SuSE)
  - Arshad Hussain (Aeon computing)
  - Shaun Tancheff (HPE)
  - Timothy Day (Amazon)
  - Chris Horn (HPE)
  - Whamcloud team

Conclusion

- Least amounts of slides for this project to date ☺
- Heavy development activity today
- After IPv6 I can focus on native client tree again
- Closest to completion we ever been.

Acknowledgments

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.