OAK RIDGE
National Laboratory

Submission of the Native Client

It’s really happening, April 2025

James Simmons

Storage Systems Engineer

Oak Ridge National Laboratory

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

U.S. DEPARTMENT OF
ENERGY

OAK RIDGE
National Laboratory | LEADERSHIP
COMPUTING
FACILITY

Progress over the years

- Decade long project
  - Was once in staging tree
    - Removed 5 years ago
    - Very bad fit
- Much larger community support
  - Amazon, Aeon, ORNL, SuSE, HPE
- Work done for two trees
  - git clone git://git.whamcloud.com/fs/lustre-release.git
  - git clone git://git.github.com/jasimmons1973/linux.git
- Meet with Linux file system maintainers week of March 24

How healthy is the native Linux client

- Mostly works out of the box
- Lagging behind supporting newer kernels
- Lagging behind OpenSFS tree
- No proc, all sysfs and debugfs
  - debugfs is root only and cloud environments disable it
- Missing GSS code. Working on restoring
- FID magic path (.lustre/fid/X) is broken
  - LU-17625, LU-11501, LU-8585
- Possible interval tree issues
  - LU-16917

OpenSFS tree nearly represents the final state

- IPv6 support is largely done
- libcfs is only debugging. Move into LNet soon
- Special hash table are almost gone on the client
- Proc is nearly gone. Especially on the client side
- Working on dropping RHEL7 handling
- Need better fscrypt support
  - Impacts kernel version support of Native client
- Removal of write handler
- Implement flio support

Game plan for upstreaming

- Lustre 2.17 will be Lustre 3.0 !!!!
- Rework OpenSFS tree to mirror Linux kernel tree
- Compact module for older kernel support.
    - Separate from libcfs
- Reduce module count
- Separate mount targets. mount -t lustre_tgt.
- Allow building Lustre against Linus tree snapshots
- Sync native kernel client with OpenSFS tree
- Change in development process
- Eventually splitting of the tree into two work spaces

Purposed development process

- Differences for creating patches
  - Break patches up more (kernel, utility code, tests)
  - Don’t ignore checkpatch style warnings. Mostly ‘*/’ on its own line
    - Lustre is stricter about column count
  - Stop using __u32 for kernel only code
  - Please use sphinix docs style for new kernel code
- Handle patch flow with gerrit and fsdevel / lustre-devel mailing list
  - Auto push of gerrit patches to fsdevel / Lustre-devel
    - Limit which patches to send. Don’t send patches for server or utility / test code. Avoid non code change rebase / retest sending as well.
    - Can be done before merging upstream after OpenSFS tree reorg
  - Once client is upstream we collect patches from fsdevel to gerrit
    - Can test against native client tree on github before merger

Testing and gate keeping policy

- What is the final place for our git tree?
- Will our gatekeeper be the final say?
- Strick patch landing
  - Patch can not land directly to outside git tree that merge with Linus
  - Must always test each patch no matter how simple
    - Maloo can fail due to unrelated issues. Perhaps after first failure only send after Maloo pass all test to not flood mailing list.
    - Maloo must pass before reviews are considered by gatekeeper
  - Must have two or more positive reviews to consider for landing
  - Only select people can land patch to final place.

Special thanks

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

Conclusion

- Least amounts of slides for this project to date
- Heavy development activity today
- Closest to completion we ever been. Hitting mile stones
- In discussions with Linux file system maintainers

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.