Linux Lustre client state

A status update, Sept 2021

James Simmons
Storage Systems Engineer
Oak Ridge National Laboratory
History of this work

• 7 Years ago EMC pushed Lustre to staging tree.
  – Not supported by the Lustre developers.
  – Drop support after one year.
  – Oldest outstanding Lustre project?

• Oleg alone maintained tree for 2 years

• ORNL became involved for last 4.5 years

• SUSE started contributed 3.5 years ago

• Dropped out of staging tree in Linux kernel
  – Moved to private repo to continue the work

• Last year Linux client was sync to master branch of OpenSFS
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 (Ubuntu20 for example)
  – Support for latest MOFED stacks
  – Performance gains (LU-11089, LU-8130)

• At Linux 5.8 version with work to move to 5.10 – delayed due to fscrypt

• Lustre community effort
  – Neil Brown from SUSE
  – James Simmons from ORNL
  – Others –
    • Arshad Hussian from Aeon + Aurelien Degremont from Amazon
How healthy is the Linux client?

- Same testing as other community projects (ARM, Ubuntu)
  - Manually running test suite from OpenSFS master branch
  - Working on making Lustre’s utilities build against Linux client
    - LU-13903 – patches should be landed by the time of this presentation
    - Once done we can enable automatic testing
      - Need to work out test system.
  - sanity-Inet 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)
    - Handle severe UAPI breakage in OpenSFS tree.
The end is near !!!!

• 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)
  – Very big project which internal changes landing to Lustre 2.15

• 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

- Remove /proc usage (LU-8066)
  - Implement Netlink to replace complex debugfs (LU-9680)
  - Enforce proper sysfs naming (LU-13091)
  - Linux client already doesn’t use /proc

- Migration to rhashtable + Xarray (LU-8130)

- Rework LNet selftest (LU-8915)
  - Not in shape currently to push upstream
  - Doesn’t work well with newer kernels (RHEL8)

- Make sysfs file names ASLR compliant (LU-13118)

- Proper fid lookup cache (LU-9868 / LU-11501 / LU-8585)
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)
  – Whamcloud focus unknown with CentOS. Looks like rocky Linux

• Discuss having external testing / bug triage outside whamcloud.

• Goal is new developers will enter the community

• 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?
    • 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.
• Close to merging to Linus tree
• Requires community involvement for proper support
  – 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.