OAK RIDGE National Laboratory

Lustre support for ARM

A status update, September 2019

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

The Lustre ecosystem

- What is Lustre?
  - Open source parallel file system for HPC systems
  - Nearly 20 years of development
  - Most deployed HPC file system in the Top500
- Who is involved?
  - OpenSFS (similar to Linaro) - http://opensfs.org/
    - Lustre Working Group -http://wiki.opensfs.org/Lustre_Working_Group
  - Developers
    - OpenSFS source tree is hosted by DDN's Whamcloud division
      - git://git.whamcloud.com/fs/lustre-release
    - Cray, ORNL, SUSE major contributors to Lustre

Non x86 platforms Lustre been tested on

ORNL

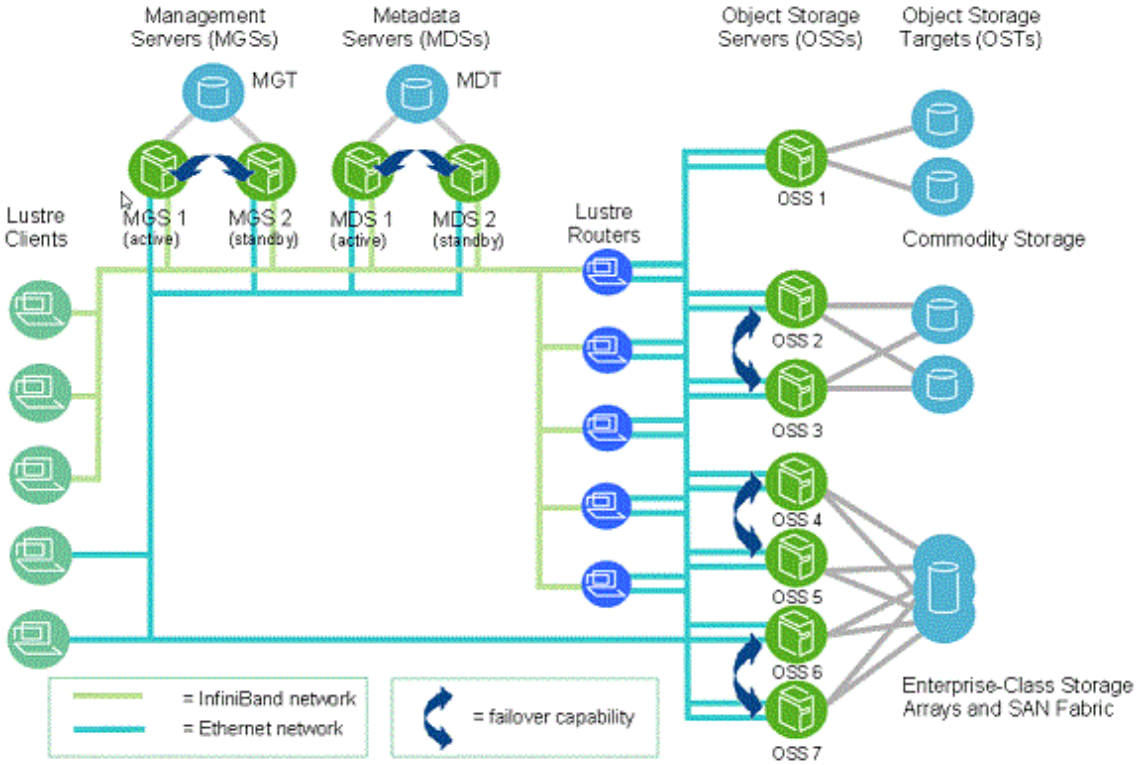
- SUSE 12 (4.4 kernel)
  - Cray prototype System with ethernet only
  - 4K page size
- RHEL7 alt (4.14 kernels)
  - Version Whamcloud test for its ARM clients
  - Our system is IB based
  - 64K page size
- Native Linux client (PowerPC)
  - https://github.com/neilbrown/linux (branch lustre)
  - 64K page support since 4.20-rc1
  - 2.10 LTS version

Others in the community

- RHEL8
  - Clients done, servers work just landed for 2.13
  - No testing for ARM
  - Main focus of Whamcloud team for ARM future
- Ubuntu 18
  - Not tested by Whamcloud or ORNL
  - Rasberry PI people show up from time to time

Lustre file system architecture

- client <-> server model
  - Lustre servers have an OSD abstraction top of real file systems
    - ZFS file system backend is supported
    - Other option is ldiskfs which is modified ext4 file system.
- Lowest level is LNet
  - Network abstraction layer based on portals
  - YAML based configuration with lnetctl tool
  - Supports routers



Lustre ARM client status

- It works!
- Latest long term release (LTS) 2.12
  - b2_12 branch of git://git.whamcloud.com/fs/lustre-release
  - Officially only RHEL x86 platforms are supported, rest are community efforts
    - http://wiki.lustre.org/Development#Kernel_Policy
    - Currently I'm only person working on ARM support for Lustre
    - First Lustre version 'lightly' tested for ARM by Whamcloud division at DDN
  - First Lustre version that works out of the box for ARM clients
    - Not 100% perfect but its pretty good.
    - Lustre ARM client deployments
      - Isambard system at University of Bristol discussed deploying Lustre at SC18
      - Astra system at Sandia National Laboratories
      - Oak Ridge National Laboratories Research Center ARM client cluster
        - https://www.olcf.ornl.gov/olcf-resources/compute-systems/wombat/

Lustre ARM server status

- Not tested by Whamcloud / DDN
- ZFS works with no additional regression
  - No patches needed!!!
  - Can use 2.12 LTS version
- ldiskfs work left
  - Need the latest Lustre OpenSFS tree
    - git://git.whamcloud.com/fs/lustre-release
  - LU-12137 / LU-11832
    - inode_lock() migrated up the VFS stack
    - Not a simple fix. Lots of open coding of VFS functions
  - LU-11200
    - Support for RHEL7 alt kernels
  - Local testing shows no new regressions

Outstanding Lustre ARM client issues

- Client issues can be tracked with
  - https://jira.whamcloud.com/issues/?filter=15555
  - https://jira.whamcloud.com/browse/LU-10300
- 64K page size issues
  - LU-11596 : grant code is assuming page size is 4K
    - Not yet examined. Most failures are due to this
  - LU-11671 : sanity test 45 – write not cached.
    - Reason for failure is unknown. I/O page accounting so it could be page related
- Test defects
  - LU-11667 : sanity test 317
    - dd write chunks not aligned if block size != PAGE_SIZE

Newer kernel challenges for ARM clients

- Overlaps with x86 running newer kernels
- Functionality
  - LU-12362 : do not call blocking ops when !TASK_RUNNING
    - Shows up when lockdep is enabled on newer kernels
  - LU-11501 : FID handling is broken for newer kernels
    - A bunch of issues, hard links in dcache, fileset don’t work etc
  - LU-11803 : security violations for sysfs file naming
    - Need to migrate to UUID for sysfs file naming
- Test infrastructure issues
  - LU-10073 : LNet selftest failures to run
    - Appears to be VM related with newer kernels
    - Plan to rewrite LNet selftest : LU-8915
- Performance enhancement
  - LU-9019 : kernel time rework to reduce noise

Lustre ARM community involvement

- Resolve remaining known ARM issues in 2.14 time frame
- We need greater scope of Lustre testing
  - ARM testing exposes very hidden bugs.
- Make native Linux Lustre client an option
- How do you test?
  - http://wiki.lustre.org/Testing
  - Report bugs at https://jira.whamcloud.com/secure/Dashboard.jspa
- 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

- This year saw Lustre ARM client deployments
- Lustre ARM client mostly works
- Lustre ARM server work is nearly complete
- Requires community involvement for Lustre ARM support
  - Join OpenSFS - http://opensfs.org/
  - Don’t be afraid to ask questions or report problems
  - LWG calls
  - Lustre mailing list
  - Report on Whamcloud JIRA
  - Contact me directly - simmonsja@ornl.gov

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.