DISRUPTIVE STORAGE WORKSHOP

Lustre-over-ZFS: Theory & Architecture

Rick Wagner
rick@globus.org

Baltimore, MD — July 14, 2017

Topics

- Lustre Fundamentals
  - From Intel Lustre training portal
  - https://www.intel.com/content/www/us/en/lustre/web-based-training.html#curriculum
- Lustre Roadmap & Resources
- Lessons from the Trenches
  - Systems perspective
  - User support perspective

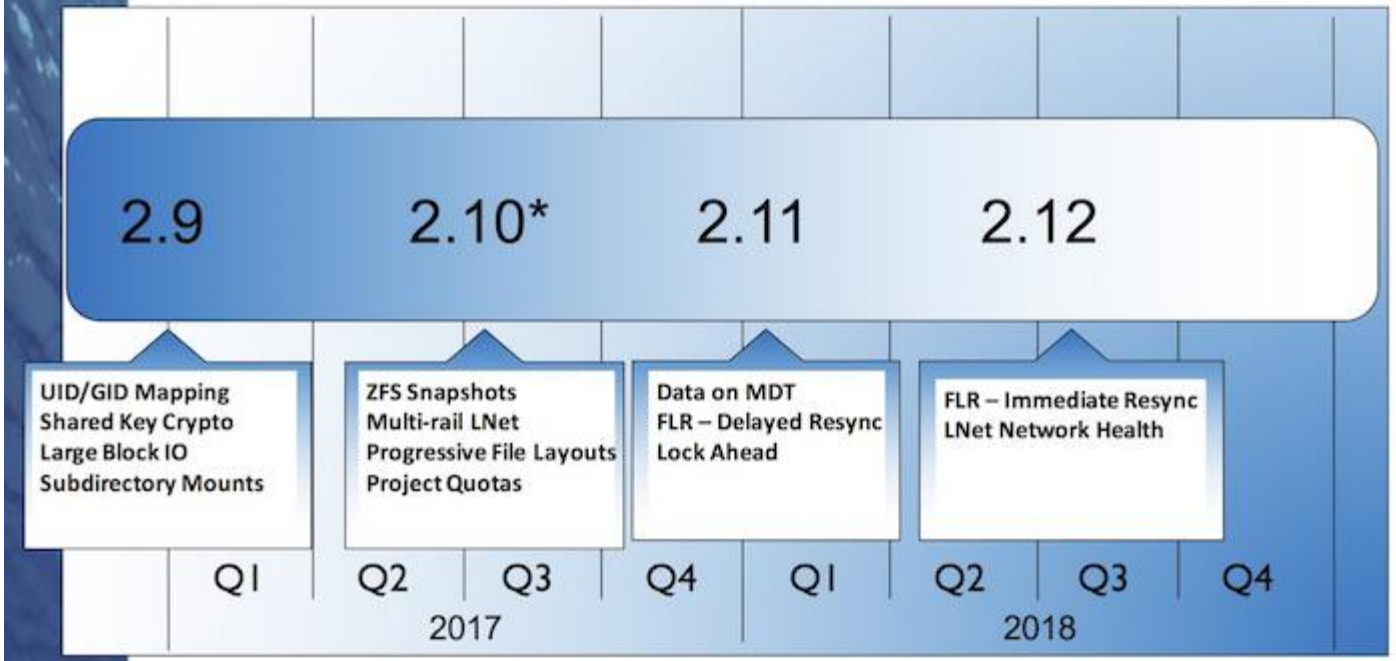
Warning:

- Slides are an amalgam of various talks
- There's going to be some visual whiplash
- We're not going to get through all of it
- Please read on your own for background info

Lustre Status & Roadmap

- Active & Healthy
- ~80% of TOP100 supercomputers use Lustre
- Intel HPDD recently refocused on supporting the community release only
- Big win for those running the community release without support
- Check out Peter Jones slides from LUG 2017 on the Community Release

Community Release Roadmap



*LTS Release with maintenance releases provided

Estimates are not commitments and are provided for informational purposes only

Fuller details of features in development are available at http://wiki.lustre.org/Projects
Last updated: April 20th 2017

http://lustre.org/roadmap/

OpenSFS

- http://opensfs.org/
- Nonprofit organization dedicated to the success of Lustre
- Runs Lustre User Group meeting
- Provides support for the Lustre Working Group
- Provides support for lustre.org along with EOFS
- Provides forum for frank and direct contact between vendors and users
- Community elected Board of Directors
- Please join!
  - Members (user organizations): $1,000/year
  - Participants (vendors): $5,000/year

Lustre Resources

- http://lustre.org
- Wiki (http://wiki.lustre.org/Main_Page)
- Lustre Working Group (http://wiki.opensfs.org/Lustre_Working_Group)
- Past LUG presentations (http://opensfs.org/lug-2017/)
- Intel Lustre training portal
  - https://www.intel.com/content/www/us/en/lustre/web-based-training.html#curriculum

Events

- Check out http://opensfs.org/events/ (Post yours!)
- Come hang out (near) Chicago for LUG 2018!
  - Week of April 23, 2018; hosted by Argonne, sponsored by Globus

Systems Perspective (1)

- Very helpful community
  - Posting to lustre-discuss works well
  - Access to lustre jira instance to browse and search open tickets
    - Good to know if bugs are already reported
- ZFS introduction eases dependence on hardware RAID
- Keeping up with current OS versions
- Backporting bugs led PSC to IEEL
- Intel support
  - Intel ZFS install guide
  - Omni-Path works well
- lnetctl is a big improvement

System Perspective (2)

- Lustre User Group
  - Talks on YouTube
- Performance tools
  - Intel tool (open source?)
  - InfluxDB/Grafana
- Documentation could improve
  - 2.x manual
- Error messages
  - Better than they used to be
  - Mailing list and jira is a help
- Open Source

User and Support Perspective (Performance)

- Codes can achieve scalable performance with the right I/O approach
  - large block I/O, control number of simultaneous writers, align I/O with filesystem parameters (striping)
  - Collective buffering, alignment options - MPI implementations are Lustre aware
  - I/O libraries like NETCDF, HDF5,ADIOS
- Most community codes have good I/O implementations - several codes achieve a good fraction of peak even on shared filesystem under load.
- Peak (measured) aggregated performance on Comet - 200 GB/s, on Gordon - 100GB/s

User and Support Perspective (Common Issues)

- Most issues can be traced to metadata load
  - typically caused by non-scalable I/O approaches
  - what works on a few nodes or cores creating a few hundred files doesn't quite scale to thousands of cores and millions of files.
  - cannot limit metadata load from particular users or nodes
  - leads to system wide response issues (why is my ls so slow!)
- Rapid increase in the number/type of applications/codes using HPC systems.
  - I/O requirements dictated by the problem being studied
  - cannot easily be modified to fit performance guidelines

User and Support Perspective (Wish List!)

- Increased metadata performance - limited at the moment, DNE (distributed namespace) helps but need more.
- Tools for finding source of metadata loads, identifying problem areas, and provide visibility into performance.
- Ability to control I/O from particular sources (users/nodes).
- UID/GID mapping, subdirectory mounts.