OAK RIDGE National Laboratory

The return of Erasure Code

Initial implementation, Sept 2024

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

Lustre’s current solutions for resiliency

- Large systems have hardware failure even with best equipment
- Lustre was designed for high end hardware
  - Bare bones handling of failures
- Lustre present solutions
  - Did you turn it on and off ? :-)
  - Fail over pairs
  - File level replication (phase 1)

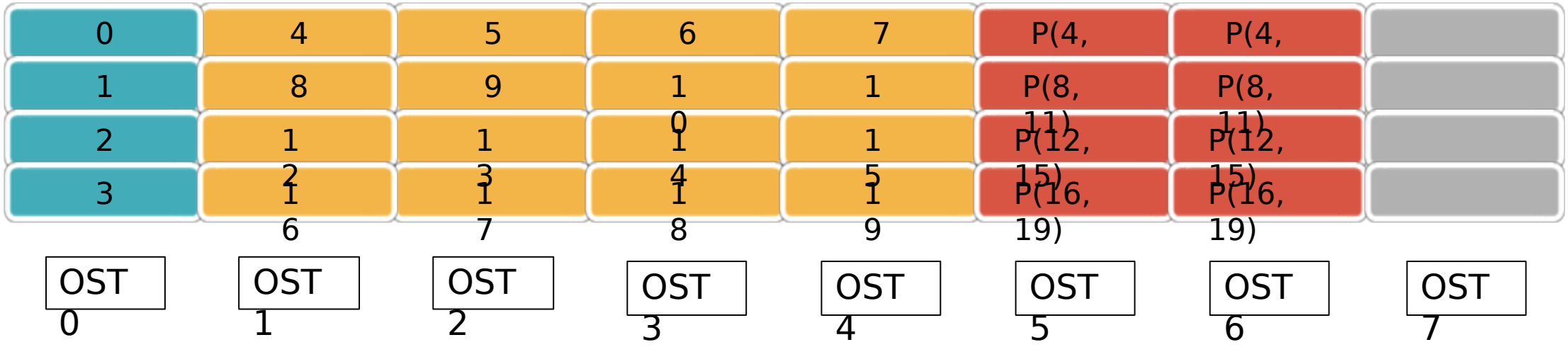
File level replication today and future

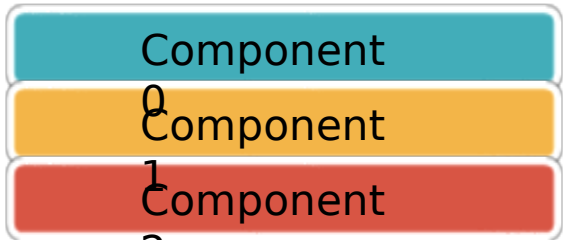
- FLR phase 1 supported today
  - Raid-1 format
  - Up to 16 mirrors
  - Manual resync. Data often stale.
  - Weakness of multiplying space usage
- FLR phase 4: erasure code (LU-10911)
  - Early work was started
  - M x N coding
  - Original patches written by Bobi Jam
    - LU-12186, LU-12187, LU-12188, LU-12189, LU-12688, LU-12669

What erasure code work is done

- Layout support in lod / lov layer
- Sending a write intent to the MDT
- Update user land tools to handle EC components
  - lfs setstripe –E eof –c 4 –E eof –L erasure_code –ec_data_count 4 –ec_parity_count 2 “file”
- Update EC parity value
- Current implementation has user manually update parity value
- Code to restore damaged file data
  - lfs mirror resync “filename”
- Done basic testing to validate above does work
  - Need to place in sanity-ec.sh

Erasure Coding to Lustre PFL mapping





Gaps in the implementation

- Inherits the issues of current FLR
  - No automatic updating of parity
  - No automatic data replacement
  - No checking of placement
- Interaction with fscrypt, compression, pools
- When do we use EC with server recovery
- Layout change, file migration
- Consider ER data size in soft quota limits?
- Implement foreign component for EC state
- No hardware acceleration

Conclusion

- Erasure coding is finally becoming a reality
- Initial patch set for base support is in pretty good shape
- Several gaps needs to resolved with the aim of 2.17 support]
  - Need feed back so we don’t miss anything
- Testers will be needed so surprises don’t show up after release

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.