HPC Storage Operations
from experience to new tools

Redmond, CUG2017

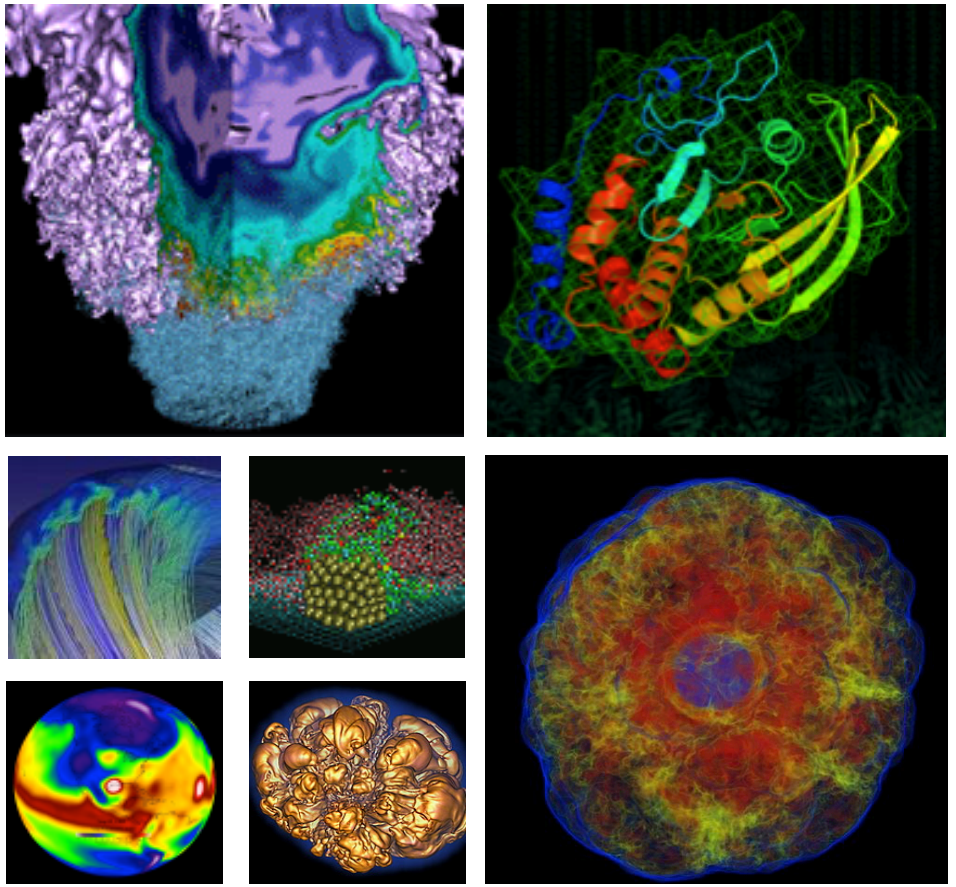
Matteo Chesi (CSCS), Tina Declerck (NERSC),
Maciej L. Olchowik (KAUST), Oliver Treiber (ECMWF)

May 9th, 2017

Since one year ago…

- “Jobs I/O monitoring for Lustre at scale BoF” in London
- Cray Caribou project
- Sonexion User experience improved, but still waiting for key features.
- A new BoF on HPC Storage Operations from Cray Storage Administrators, let’s discuss!

Lustre Purge



Jack Deslippe,
November, 2016

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

What is purging? Why Purge?

- **What is purging?**
- Purging is generally a process for identifying and removing files a system.
  - The most common means of identifying these files is based on age
  - Site policy states how long files can expect to live so users are not surprised
- Considerations
  - Queue length – shouldn't purge files for jobs that are waiting to run
  - Shortest time that allows user jobs to complete and generated data to be stored
  - Can possibly be longer if quotas help keep the file system under control
  - Are there 'special' users who need longer aging or are not subject to purging
- **Why purge?**
- Most file systems do NOT like being full
- If your users are like ours they don't necessarily do a good job of cleaning up files

How is purging done?

- Solutions generally
  - Scan the file system to generate a ”hit list”
  - Eliminate any files identified for the ‘special’ cases
  - Remove the specified files
    - Large file systems consider re-verify the data prior to removal
- Robinhood – The French Alternative Energies and Atomic Energy Commission (CEA)
  - Policy engine for managing large file systems
  - Provides scan capability with policies that can be used to purge
    - Also can use the Lustre changelog
    - Supports DNE with a changelog reader per MDS

Issues & Benefits

- **Issues**
- Scale
  - Time to scan the file system can be measured in days
  - Changelogs have been problematic
    - Can't keep up
    - Filling the changelog space makes the file system unusable
  - Various problems with stripe width – currently resolved
- **Benefits**
  - Robinhood database provides access without impacting the file system
    - du
    - File system data size reports
    - Top user reports

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Data Migration Policies issues and TAS

Maciej Olchowik

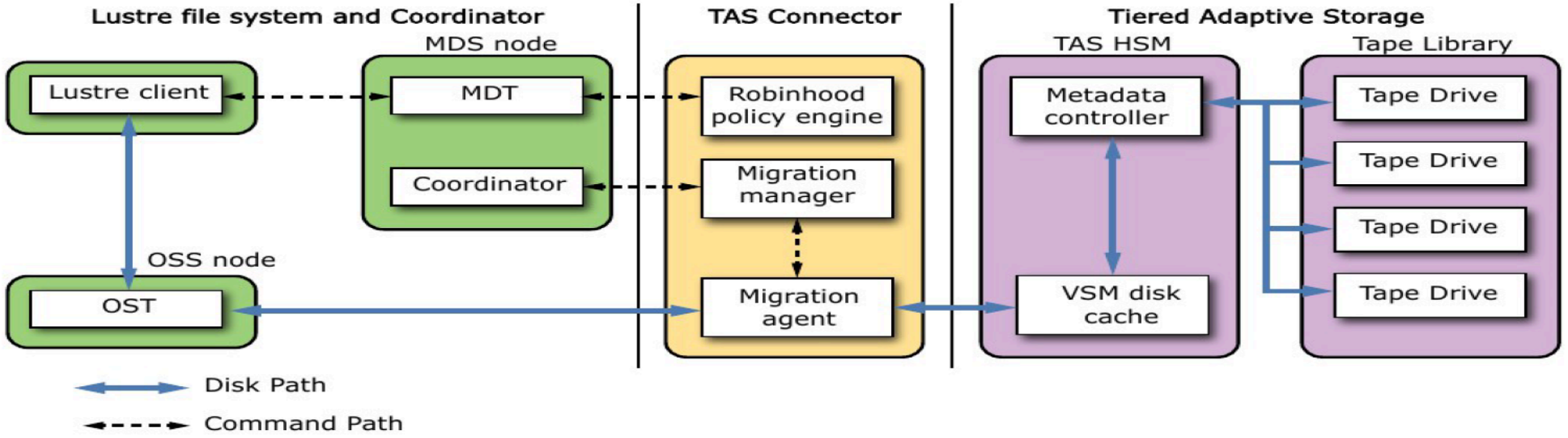
System Administrator, Supercomputing Core Lab
King Abdullah University of Science and Technology

Storage Overview

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

- 30PB Spectra Logic Tfinity tape library (20 drives)
- 16 PB single lustre filesystem divided into:
  /project - 20TB limit per project, anything above that is migrated to tape
  /scratch - no limit, but purge policy to remove files older than 60 days

- TAS connector



Issues with Lustre and TAS

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

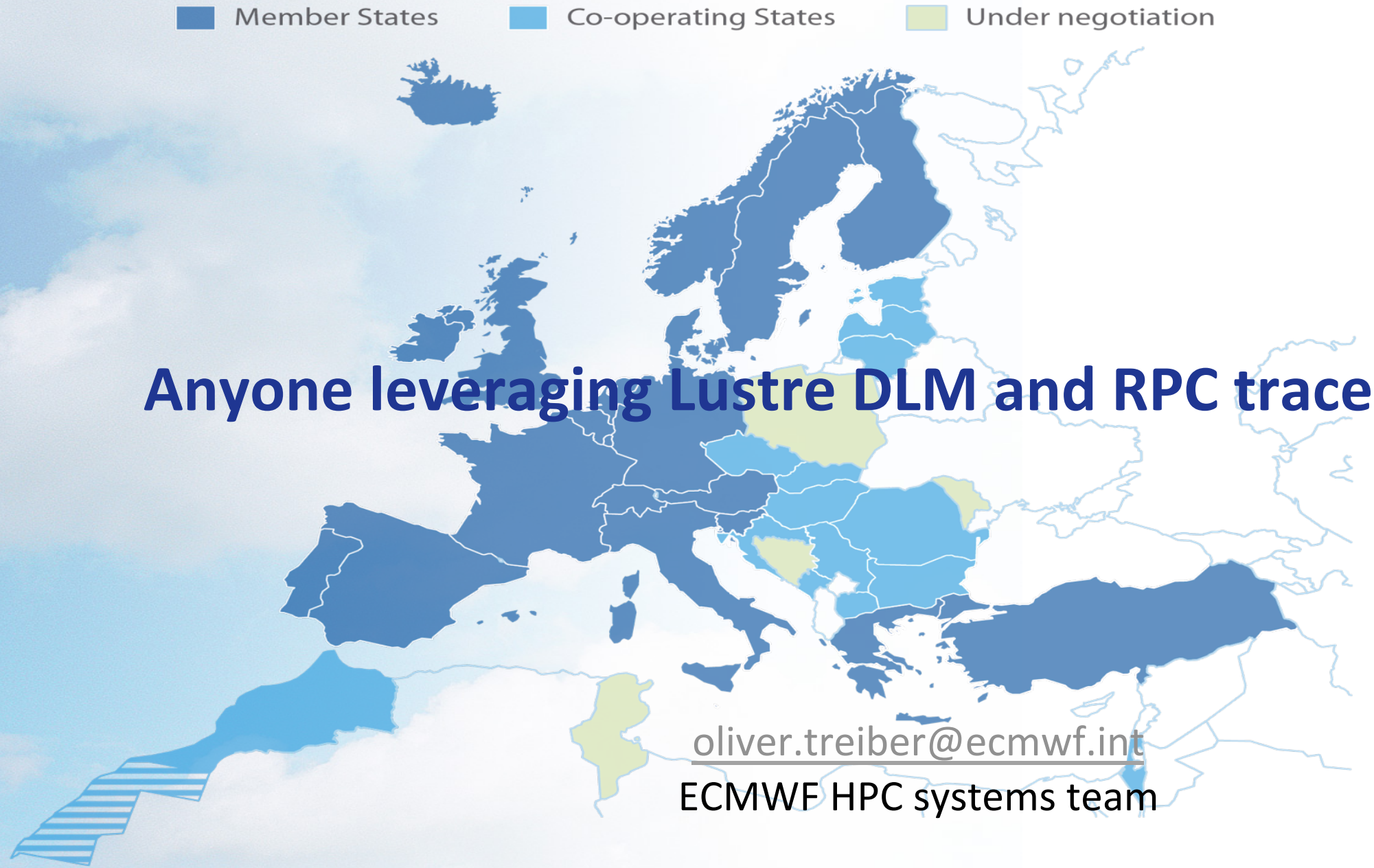
- Robinhood policy engine not keeping up with lustre changelogs at our scale. Manual scans are required.

- File recall (from TAS) causing lustre deadlock (LU-7988) which affects all clients
  Temporary workaround to manually failover Sonexion MDS failover.
  Issue apparently fixed in SU23A.

- Lustre HSM coordinator not providing candidate files to archive at the expected rate (LU-8626). Lustre issue.

- Impact of lustre problems on the TAS connector

- Occasional lustre locks for not fully understood reasons

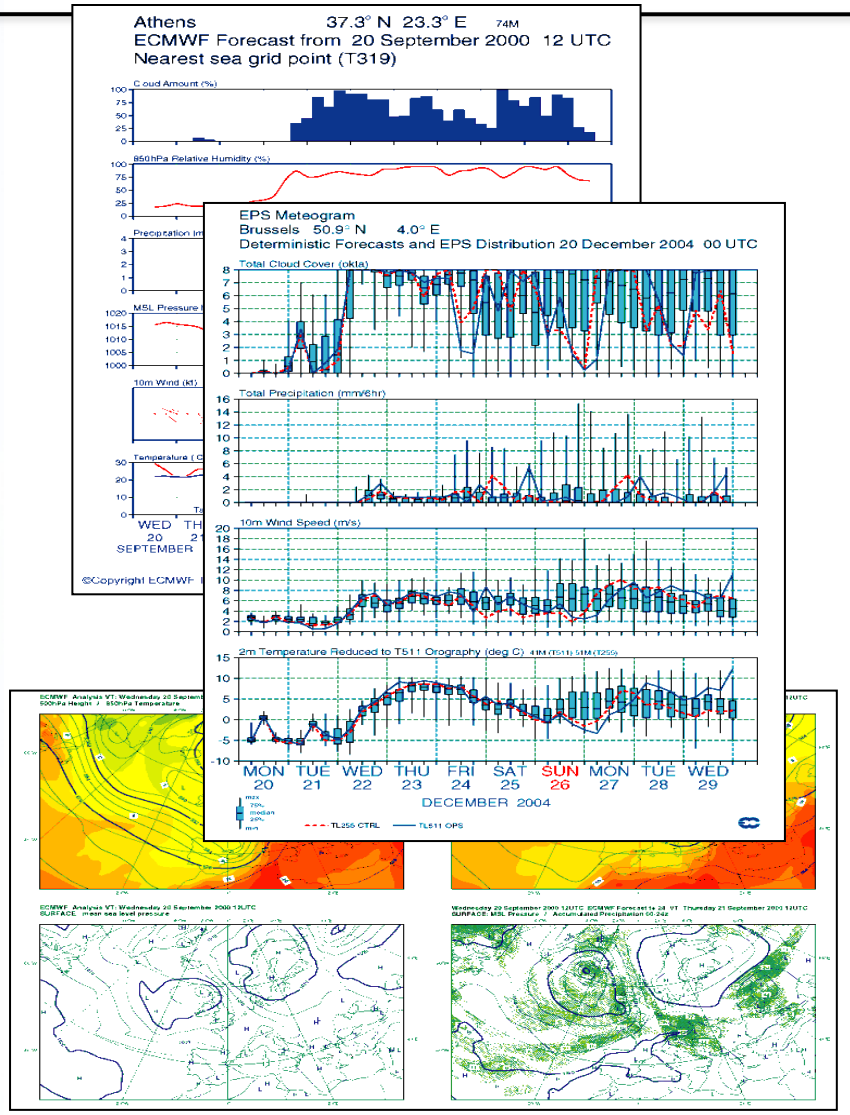


Anyone leveraging Lustre DLM and RPC traces?

oliver.treiber@ecmwf.int
ECMWF HPC systems team

ECMWF: European Centre for Medium-Range Weather Forecasts

European Centre	independent international organisation funded by 34 States
Medium-Range	forecasts up to fifteen days ahead; also monthly and seasonal forecasts, collection/store of meteorological data.
Weather Forecasts	global weather forecasts
Copernicus (EU)	ECMWF implements CAMS and C3S
People	~300 staff, specialists and contractors
Computer	2 XC40 (each ~3600n) 26 cabs SNX 1600+2000



lustre “distributed” lock management (LDLM) and RPCs: Lustre traces

- MDT/namespace vs. OST/file-range locking
- ISSUE at EC recently: need to underst latencies for metadata ops like open, access, stat for specific paths

- I am just fishing: any experiences/expertise with leveraging Lustre traces out there that can be shared?
  - then, let’s talk and collect…

- googl’ing does not seem to produce a lot of detail
  - not much in lustre pubs; source code maybe not best entry point
  - some background material
    - http://people.redhat.com/ccaulfie/docs/rhdlmbook.pdf (DLM)
    - http://wiki.old.lustre.org/images/d/da/Understanding_Lustre_Filesystem_Internals.pdf
    - http://cdn.opensfs.org/wp-content/uploads/2014/04/D3_S32_LustreLogAnalyzer.pdf

lustre “distributed” lock management (LDLM) and RPCs

- why interested?
- e.g., at EC, recently struggled a lot with this (MDT context) b/c of “suboptimal application config”
  - example pathology: struggled with suddenly appearing 10000x latency increases in timecrit apps…
    - jobs open/access/stat files through specific intermediate dir X: /lus/snx?/X/…/…./… (ops need protection by “shared lock”)
    - as it turns out: some other apps also issuing high rate of “gratuitous”, but NOT-HARMLESS “non-op” syscalls
      - rmdir(/lus/snx?/X/z), where z is not-empty
      - mkdir(/lus/snx/X/z), where z already exists
    - despite resulting in non-ops, on MDT these still trigger global lock revocation
      - latency for this global lock revocation depends on processing speed on MDS and latency of ldlm_cancel responses by lock holding client after receiving blocking ASTs
        - even when situation was bad, MDS had low CPU load and loads of free memory
        - suspicion was there can be sluggish Lustre clients
    - BTW: as non-ops, these rmdir/mkdir calls are not seen in lustre changelog
    - but those ops do (“anonymously”, without identifying resoucrce) increment counters in MDT’s exports’ stats and ldlm_stats

lustre “distributed” lock management, LDLM and RPC tracing

- suboptimal approach in tracking down: try to find contending syscalls through cluster-wide client-side strace and ftrace snooping
- instead try to gain insight (e.g., contended FIDs, latencies, nids,…) from “scripted” analysis of LDLM/RPC trace data
  - gather on demand: +rpctrace and +dlmtrace on /proc/sys/lnet/debug, and “lctl dk <outfile> 1”
  - nice events sequences visible- when looking at test system, but messier on production system
  - is this a per-core cyclic logbuffer? (to assess completeness/time covered)
  - what are the relevant patterns to correlate in output?

00010000:00010000:18.0:1487922524.103132:0:43815:0:(ldlm_lock.c:638:ldlm_add_bl_work_item()) ### lock incompatible; sending blocking AST. ns: mdt-snx11057-MDT0000_UUID lock: ffff8805ef351c00/0xe852b1841c75eb18 lrc: 2/0,0 mode: PR/PR res: [0x200047111:0x2:0x0].0 bits 0 x3 rrc: 8 type: IBT flags: 0x4200000000000 nid: 6@gni remote: 0x6726e5786c2d0790 expref: 494 pid: 43826 timeout: 0

[...]

00010000:00010000:0.0:1487922524.103613:0:86788:0:(ldlm_lockd.c:2252:ldlm_cancel_hpreq_check()) ### hpreq cancel loc k ns: mdt-snx11057-MDT0000_UUID lock: ffff8805ef351c00/0xe852b1841c75eb18 lrc: 4/0,0 mode: PR/PR res: [0x200047111:0x2:0x0].0 bits 0x3 rrc: 8 type: IBT flags: 0x4200000000020 nid: 6@gni remote: 0x6726e5786c2d0790 expref: 495 pid: 43826 timeout: 45055250172

lustre “distributed” lock management, LDLM and RPC tracing

- such lock contention on individual resources seems not exposed in usual Lustre monitoring tools
  - pay some attention by inclusion of global LDLM event rates in Caribou ?

- where are we now with this at EC? we first defused situation by cleaning up applications...
- but then entered next episode: MD latencies in aforementioned workflow increased again by ~1000x
  - we failed over MDS service, then latencies dropped!!! why did it clear, where did latency increase come from?
  - what mattered here ito the MDS failover? server or clients recovery?
    - we dlm/drop_caches on MAMUs once per hour anyway
  - currently investigating if related to leak in “granted” count (vs lock_count) (LU-8246) (not clear if applies)

The Big Picture

XC system

Sonexion

CRAY® TIERED ADAPTIVE STORAGE

HSM

HPSS High Performance Storage System

lustre™ CRAY intel SEAGATE OpenSFS EOFS

Parallel File System

Locks and RPCs

Cleaning Policies

Robinhood

cea

Policy Engine

POSIX