Robinhood and Lustre*-HSM: a return of experience

RUG 2015

Sept 21, 2015

Bruno Faccini, High Performance Data Division

Story points

- CEA and CFS joint design specifications for Lustre-HSM (start in early 2008)
- Early decision to rely on an optional and external Policy Engine (during 2008)
- RBH early interfacing with Lustre-HSM API by CEA team (completed late 2009)
- RBH selected for policy engine in Intel EE for Lustre Software (mid 2014)
- Production work-loads exposure of both Lustre-HSM and RBH working together at EE sites (late 2014)

Families of issues encountered

- RBH platform HW configuration/sizing
- Configuration (CDT, RBH, MySQL/DB)
- Performance/Scalability
- Stability/Reliability
- Cleanup
- Limitations and potential enhancements
- Bugs!

RBH platform HW configuration/sizing

- Some sites have started running RBH on non-dedicated nodes (on any Lustre-Client, on IML management node, …)
- RBH (logs) and MySQL (DB) directories to be put on non-shared devices/disks
- SSD usage is a must
- Experience from large deployments with RBH clearly indicates the need of a strong (CPUs/Cores, Memory size/ bandwidth, DB disks speed, …) node to support production work-load

Configuration (CDT, RBH, MySQL/DB)

- On Lustre-HSM/CDT side :
  - max_requests to match with copytool's capability to handle requests in //
  - max_rpcs_in_flight=64
  - CL_CLOSE in ChangeLog mask
- On RBH side :
  - Disable user/group accounting
  - Change inappropriate/out-dated settings in config
- On MySQL side :
  - innodb_buffer_pool_size setting is critical
  - innodb_flush_log_at_trx_commit=2
  - "http://mysqltuner.pl" advices

Performance/Scalability

- On Lustre-HSM/CDT side :
  - ChangeLog clear scan order (LU-5405)
  - Cookie/FID indexed request hash
  - cdt_llog_lock from mutex to RW-lock
  - Update HSM_Actions LLOG records in-place
- On RBH side :
  - Multi-threading of batched DB operations (commit bd6aa4f)
  - set default parameters for performance (commits 4d2ec05, 025078c, 00bf69d)

Stability/Reliability

- Strengthen Lustre-HSM (and related) code
  - CDT Error handling
  - HSM State/Flags/archive_id validity (LU-5757)
  - Protection
  - Policy-Engine <-> CDT <-> Copytool flow control

Cleanup

- Avoid to leak space/entries on archive/copytool side
  - RBH patch to save archive_id in DB (commit 1a3de27)
  - CDT to broadcast hsm_remove requests
  - RAoLU policy when no RBH/Policy-Engine (LU-4640)

Limitations and potential enhancements

- Layout_swap handling in RBH (commit 7df5f3f)
- DNE support (commits a9818b7, c5e9dbc)
- Copytool and RBH handling of sparse files (LU-3833)
- Migration does not generate a ChangeLog to report FID change (LU-6868)
- RBH/DB HA config evaluations

Bugs!

- RBH daemon coredumps during log switch (commit 3a94cf7)
- RBH wrongly relied on old pathname during pre-match of policies phase for a Rename ChangeLog (commit 65a2c45)
- LLOG regression unveiled by RBH and Lustre-HSM operations (LU-6556)

Legal Information

- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.
- Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.
- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.
- For more complete information about performance and benchmark results, visit http://www.intel.com/performance.
- Intel, the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others

intel Software


intel
Software