Data life cycle monitoring using RoBinHood at scale

Gabriele Paciucci – Solution Architect

Bruno Faccini – Senior Support Engineer

September 2015 - LAD

Agenda

- Motivations
- Hardware and software setup
- The first scan problem
- Changelog injection
- DU, FIND
- Conclusion

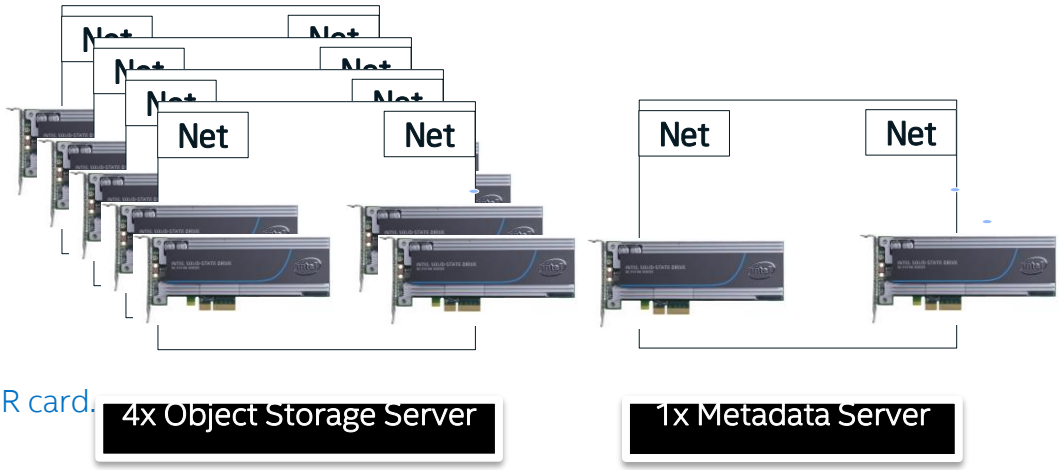
Motivations

During the last few years several customers started to adopt RoBinHood (RBH) to monitor and manage (HSM) the life of the data into Lustre*.

Intel currently supports RBH, included in the Intel® Enterprise Edition for Lustre* software.

The objective of this presentation is to guide the audience on how to size, test and troubleshoot RBH to handle a very large number of files.

Hardware Setup

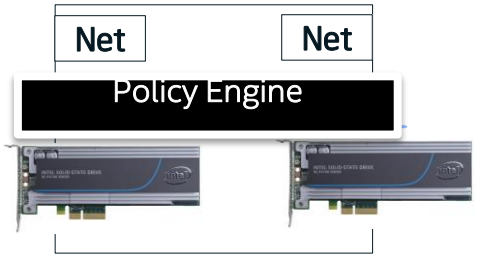


4x Object Storage Server:

- 2x Intel® Xeon® E5-2643v2 CPU
- 64GB of RAM.
- 4x Intel® SSD DC P3700 Series 400GB
- Mellanox* FDR card and Intel® True Scale Fabric QDR card.

1x Metadata Server:

- 2x Intel® Xeon® E5-2643v2 CPU
- 64GB of RAM.
- 2x Intel® SSD DC P3700 Series 400GB
- Mellanox* FDR card and Intel® True Scale Fabric QDR card.



1x Policy Engine

- 2x Intel® Xeon® E5-2643v2 CPU
- 64GB of RAM.
- 2x Intel® SSD DC P3700 Series 400GB
- Mellanox* FDR card and Intel® True Scale Fabric QDR card.

intel Software

Software stack used

Intel® Enterprise Edition for Lustre* 2.3

- Lustre* version 2.5.37.7
- Kernel version 2.6.32-504.30.3.el6_lustre.x86_64

Patched version of RBH 2.5.5

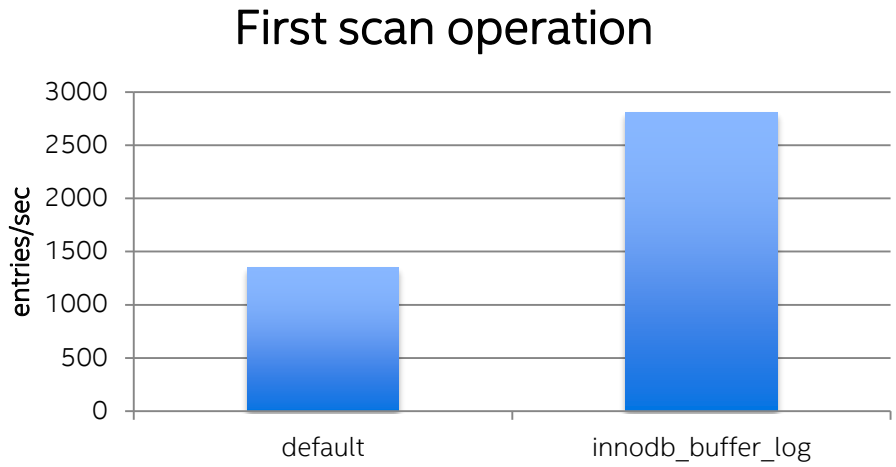
- RBH accounting has been disabled for performance and to benefit recent commit bd6aa4f (multi-threading DB batch operations)

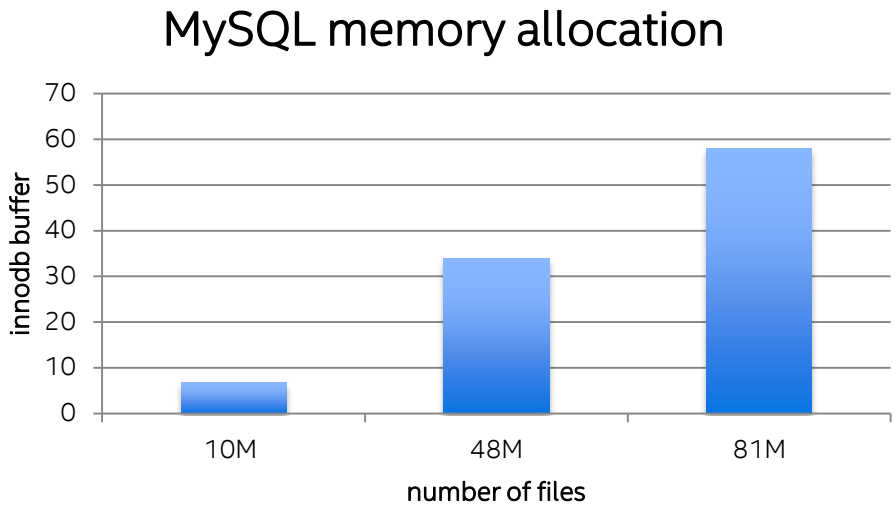
MySQL

- Version 5.1.73 included in the CentOS* distro
- Community version 5.6.26 from Oracle
  - Processor hw crc32 checksuming to reduce the buf_calc_page_new_checksum()

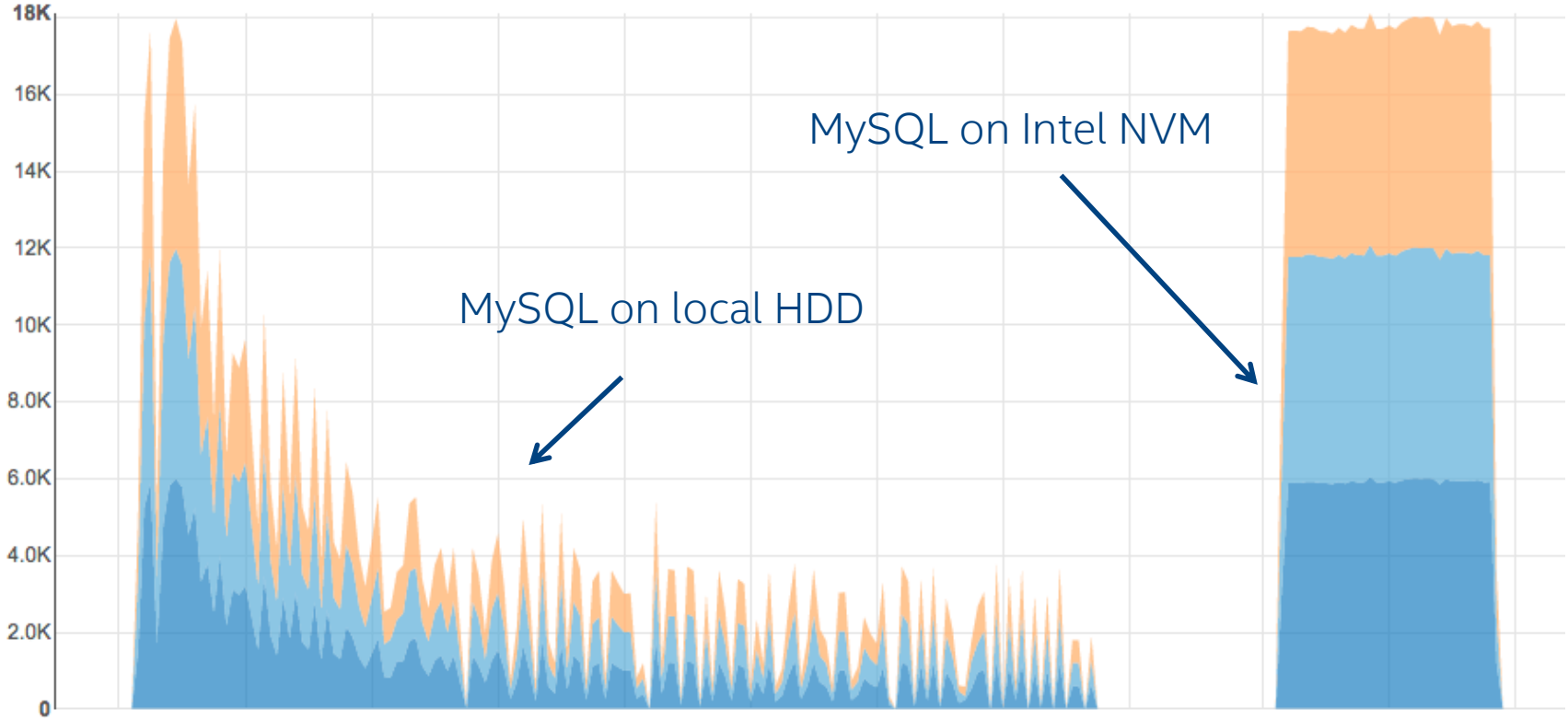
How to design the RBH server

- High frequency CPU to increase metadata queries
- As much memory as possible
- 80% of available RAM for innodb_buffer_log





SSD devices are essential



First scan operation with RBH on the same server. In the chart metadata operations per second collected by Intel Manager for Lustre*

MySQL software tuning

Maintain the configuration simple

Use mysqltuner for advice

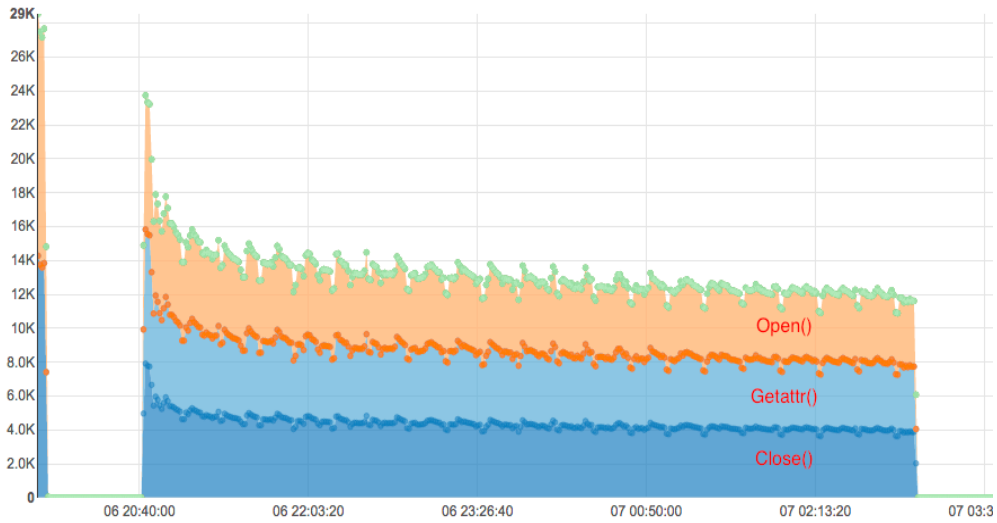
- slow_query_log=1
- query_cache_size=8M
- thread_cache_size=4
- innodb_buffer_pool_size= < 80% of the
- tmp_table_size=64M
- max_heap_table_size=64M
- innodb_flush_neighbors=0
- innodb_flush_log_at_trx_commit=2

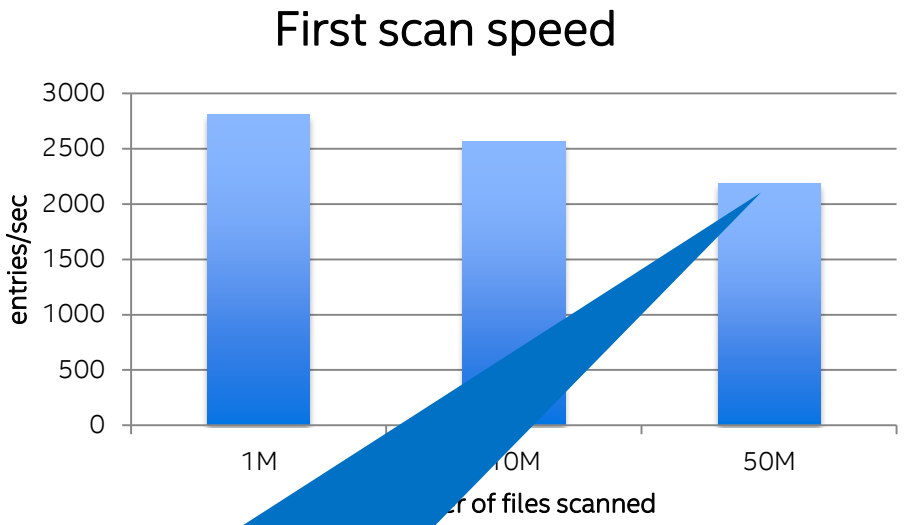
Disable sequential optimization for MySQL and save CPU

Disable transaction protection in MySQL if your server is safe enough

First scan operation

- RBH calls path2fid, lstat, getstripe, hsm_state to collect information about entries
- The metadata server must be fast enough
- Many operation modes are possible:
  - Single scan server
  - Single scan server multiple threads
  - Multiple scan servers



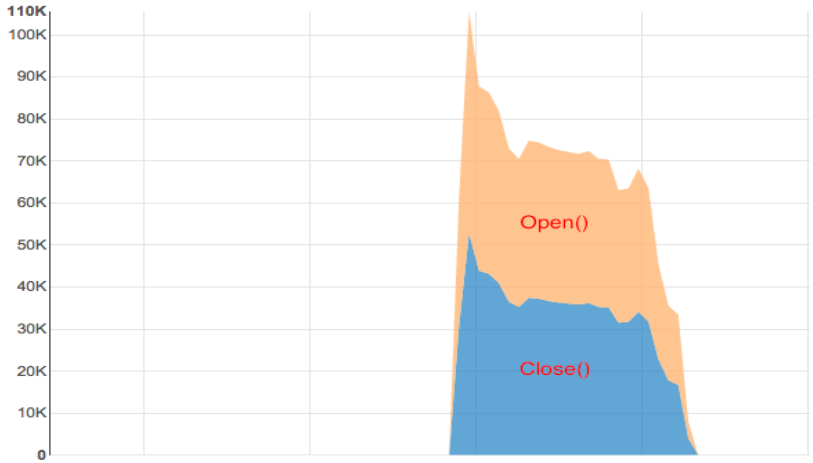


We run out of innodb_buffer_log

Changelog operations

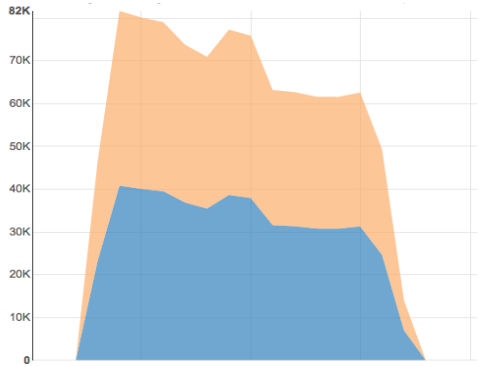
- RBH uses the Lustre* changelog to identify changes in the file system after the first scan.
- We need (again) a fast metadata server

Changelog OFF



35282 file creation/sec

Changelog ON



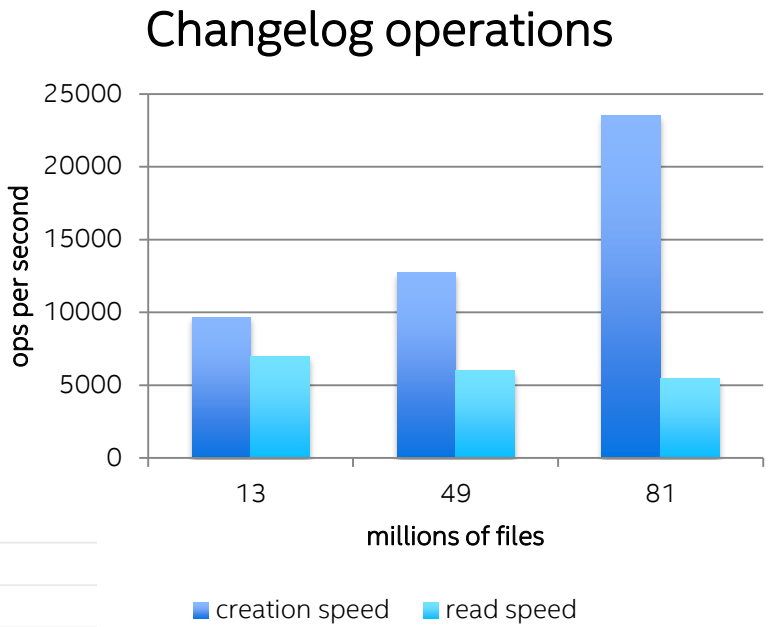
33730 file creation/sec

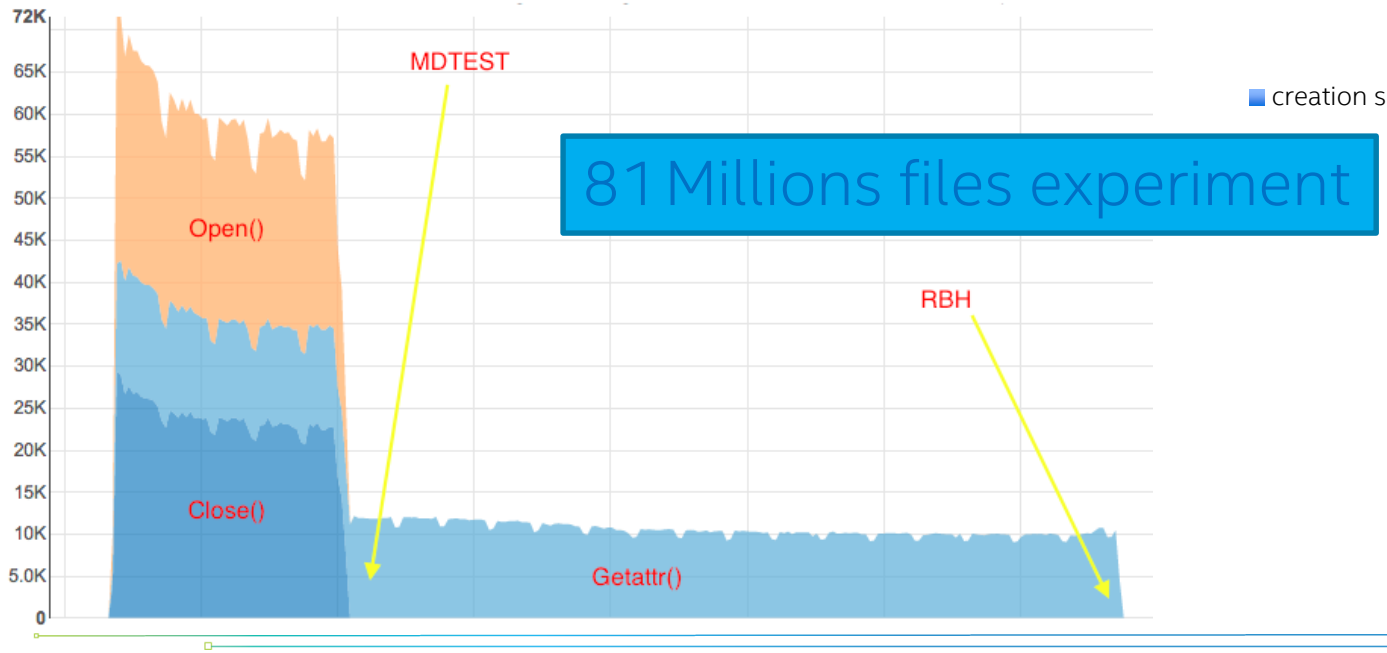
Changelog injection test bed

- Dataset 10M/50M/80M, 1M files per directory
- Real small files (64/32K) created by mdtest
  - $MDTEST_BIN -I 1000000 -w 64000 -F –C -u -d $DIR
  - Used 4 to 16 clients
- Tested only CREATE operations on the file system
- Tools used to troubleshoot:
  - perf
  - mysqltuner
  - Intel Manager for Lustre*
  - iostat, vmstat and other Linux tools

Speed measured during the experiments

- Increasing the number of clients file creation speed is increasing
- The RBH speed remain stable but below the speed of the file system



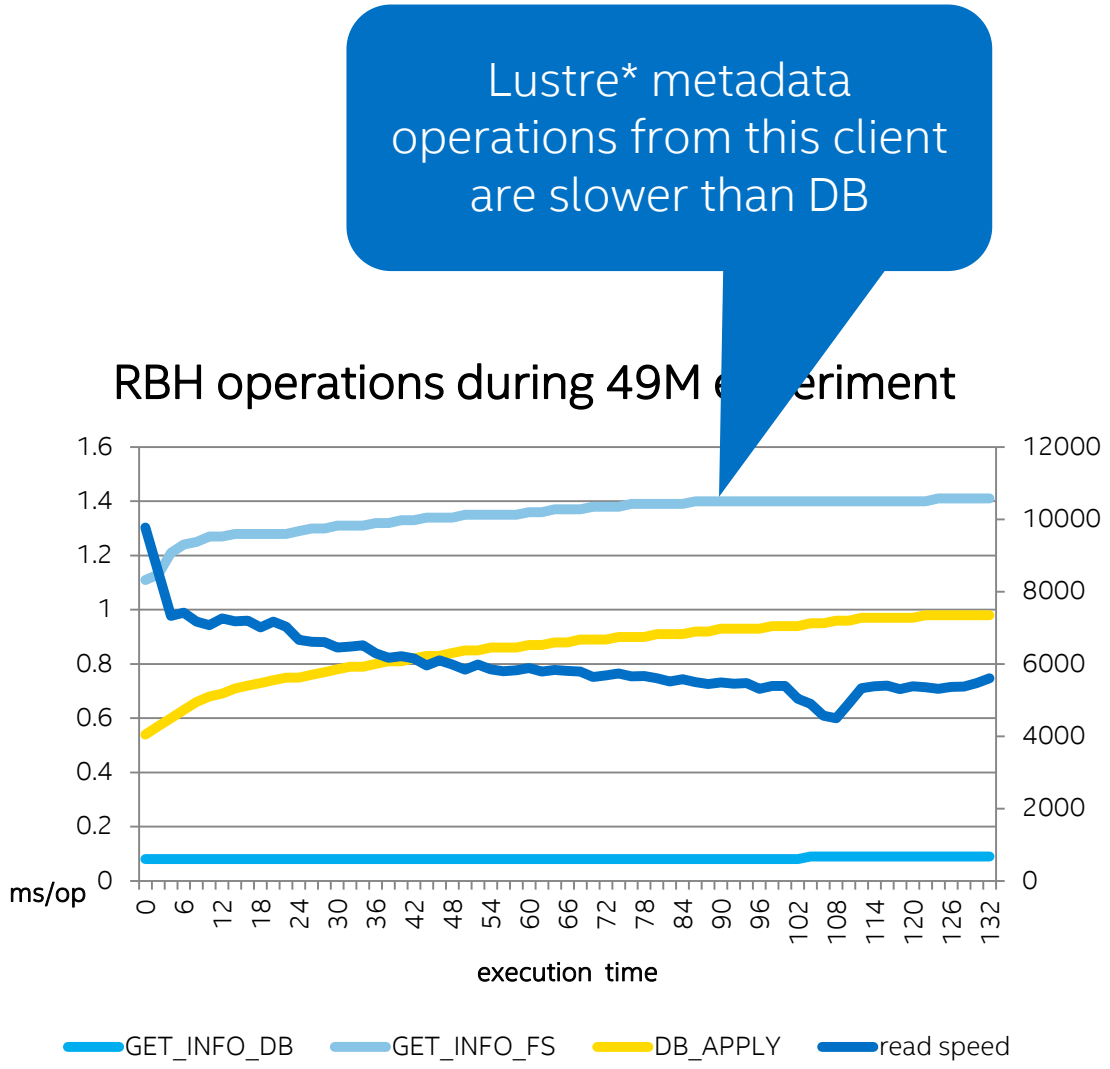


How to troubleshoot RBH

- Activate the stats on RBH
- EntryProcessor Pipeline Stats
  - GET_INFO_DB
  - GET_INFO_FS
  - DB_APPLY
  - Idle threads
- We want to maintain the pipe full without overloading
  - Increasing the n. threads and the objects to process is not always a good idea

Troubleshooting

- The latency measured in the RBH's report file give us great insights
- In our environment the latency of RBH to access to the file system is bigger than the DB operations
- A Lustre* tuning is necessary to decrease the latency



Lustre* metrics and tunables

Increasing the capacity of RBH to perform metadata operations:

- lctl set_param llite.*.statahead_max
- lctl set_param ldlm.namespaces.*.lru_size
- lctl set_param ldlm.namespaces.*.lru_max_age
- sysctl -w lnet.debug=0
- lctl set_param mdc.*.max_rpcs_in_flight (remember to increse peer_credits if necessary)

RobinHood helps SysAdmin

	49 millions	81 millions
du -h	5h 40m 12sec	**Still working !!!**
rbsh-lhsm-du -H -d	1m 39sec	1m 46sec
lfs find --ost 0	6h 49m 16sec	**Still working !!!**
rbh-lhsm-find --ost 0	33m 54sec	36m 52sec

Conclusion

- We successfully tested the capabilities of RBH to report the utilization of file system busy with millions of files
- MySQL must be very well designed and tuned
- Manage a RBH instance for more than 100M of files could be a problem
- Next steps:
  - Verify the impact of deep tree layouts
  - Try to mimic a more realistic workload including different metadata operation
  - Verify the scalability of RBH in a HSM environment

Notices and Disclaimers

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 intel.com.

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.

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.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

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.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, and Intel Xeon® trademarks of Intel Corporation in the United States and other countries.

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


intel
Software