FROM RESEARCH TO INDUSTRY

cea

www.cea.fr

RobinHood Project Status

Robinhood User Group 2015

Thomas Leibovici <thomas.leibovici@cea.fr>

SEPTEMBER, 21st 2015

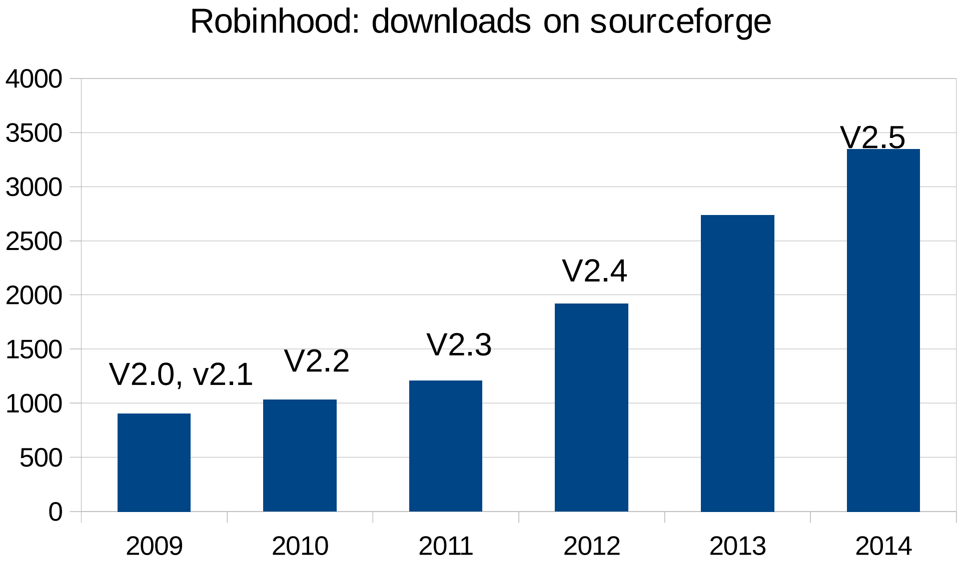
9/18/15

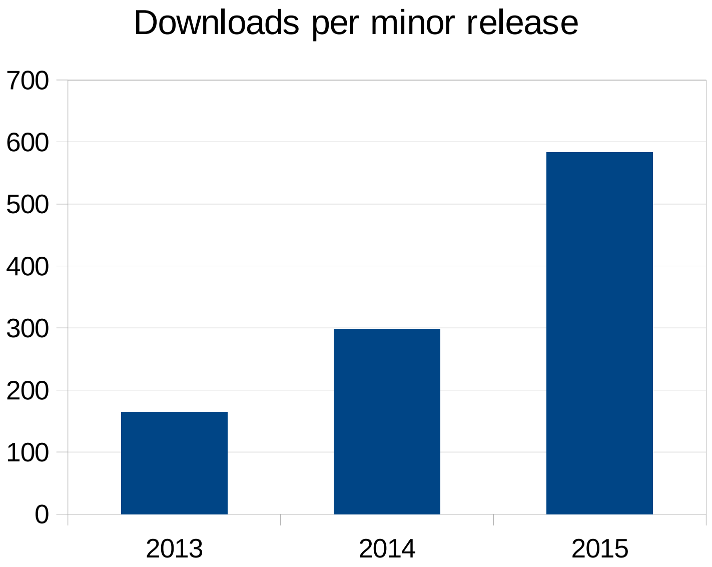
Project history...

- 1999: simple purge tool for HPC filesystems
  - 1 scan thread, 1 purge thread, entry list in memory
- 2005: massively multi-threaded (robinhood v1)
  - Multi-threaded scan algorithm
  - Purge queue with multiple worker threads
  - Basic policies (whitelist/blacklist of a user, group, directory...)
  - Aware of Lustre OSTs
  - List in memory
- Feb. 2009: Robinhood made open source (v1.0.6)
  - CeCILL-C (LGPL compatible)
- 2009: development of Robinhood v2
  - Based on a SQL database
    - Provides memory cache management, persistence, robustness, convenient query language, ...
  - Lustre v2 ready (fid, changelogs...)
  - Support complex expressions on entry attributes
    - e.g. `{(path == "/lustre/foo/bar*" or owner == foo) and size > 1GB)}`

...until today

- 2009-2015: a growing community
  - Continuous improvements and new features, support of latest Lustre versions and features, responsive support, vendor integration and contributions, ...





+ other channels (vendor distributions...)

- September 2015: first Robinhood User Group
  - Thanks for attending!

Robinhood Modes
and Supported Filesystems

Robinhood flavors (v2)

Scratch filesystem management (tmpfs)

- Monitoring/accounting, purge (=unlink), rmdir (=*rmdir* or *rm -rf*)
- Package: `robinhood-tmpfs`
- Supported:
  - All POSIX filesystems
  - All Lustre versions from 1.8 to 2.7

Backup

- Monitoring/accounting, archiving, undelete
- Package: `robinhood-backup`
- Supported:
  - All Lustre versions from 2.0 to 2.7

Lustre HSM (lhsm)

- Monitoring/accounting, migration (=archive), purge (=release), HSM rm
- Package: `robinhood-lhsm`
- Supported:
  - All Lustre versions from 2.5 to 2.7

[spoiler] Robinhood v3

1 single instance for all purposes

- Available for POSIX filesystems and all Lustre versions from 1.8 to 2.7 and later

Plugins for specific feature support

- “backup” plugin available for Lustre >= 2.0
- “lhsm” plugin available for Lustre >= 2.5
- [teaser] Other plugins... or your own!

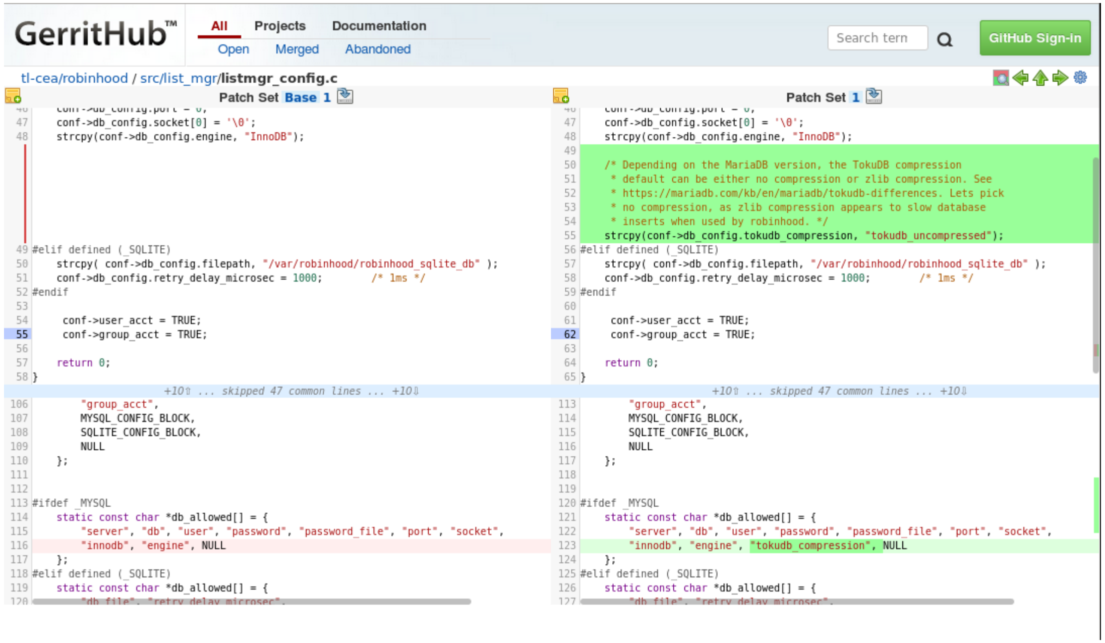
- A single robinhood instance can handle multiple plugins
  e.g. manage Lustre/HSM policies + delete old files + ...

  (Cf. next talk about robinhood v3)

Validation & Release Process

Code review: Gerrithub

- https://review.gerrithub.io
- Project: cea-hpc/robinhood



```
    conf->db_config.port = 0;
    conf->db_config.socket[0] = '\0';
    strcpy(conf->db_config.engine, "InnoDB");
#elif defined (_SQLITE)
    strcpy( conf->db_config.filepath, "/var/robinhood/robinhood_sqlite_db" );
    conf->db_config.retry_delay_microsec = 1000;        /* 1ms */
#endif

     conf->user_acct = TRUE;
     conf->group_acct = TRUE;

    return 0;
}
                +10⇧ ... skipped 47 common lines ... +10⇩
        "group_acct",
        MYSQL_CONFIG_BLOCK,
        SQLITE_CONFIG_BLOCK,
        NULL
    };


#ifdef _MYSQL
    static const char *db_allowed[] = {
        "server", "db", "user", "password", "password_file", "port", "socket",
        "innodb", "engine", NULL
    };
#elif defined (_SQLITE)
    static const char *db_allowed[] = {
        "db_file", "retry_delay_microsec",
```

```
    conf->db_config.port = 0;
    conf->db_config.socket[0] = '\0';
    strcpy(conf->db_config.engine, "InnoDB");

    /* Depending on the MariaDB version, the TokuDB compression
     * default can be either no compression or zlib compression. See
     * https://mariadb.com/kb/en/mariadb/tokudb-differences. Lets pick
     * no compression, as zlib compression appears to slow database
     * inserts when used by robinhood. */
    strcpy(conf->db_config.tokudb_compression, "tokudb_uncompressed");
#elif defined (_SQLITE)
    strcpy( conf->db_config.filepath, "/var/robinhood/robinhood_sqlite_db" );
    conf->db_config.retry_delay_microsec = 1000;        /* 1ms */
#endif

     conf->user_acct = TRUE;
     conf->group_acct = TRUE;

    return 0;
}
                +10⇧ ... skipped 47 common lines ... +10⇩
        "group_acct",
        MYSQL_CONFIG_BLOCK,
        SQLITE_CONFIG_BLOCK,
        NULL
    };


#ifdef _MYSQL
    static const char *db_allowed[] = {
        "server", "db", "user", "password", "password_file", "port", "socket",
        "innodb", "engine", "tokudb_compression", NULL
    };
#elif defined (_SQLITE)
    static const char *db_allowed[] = {
        "db_file", "retry_delay_microsec",
```

Test Suite & Test Platform

Test suite

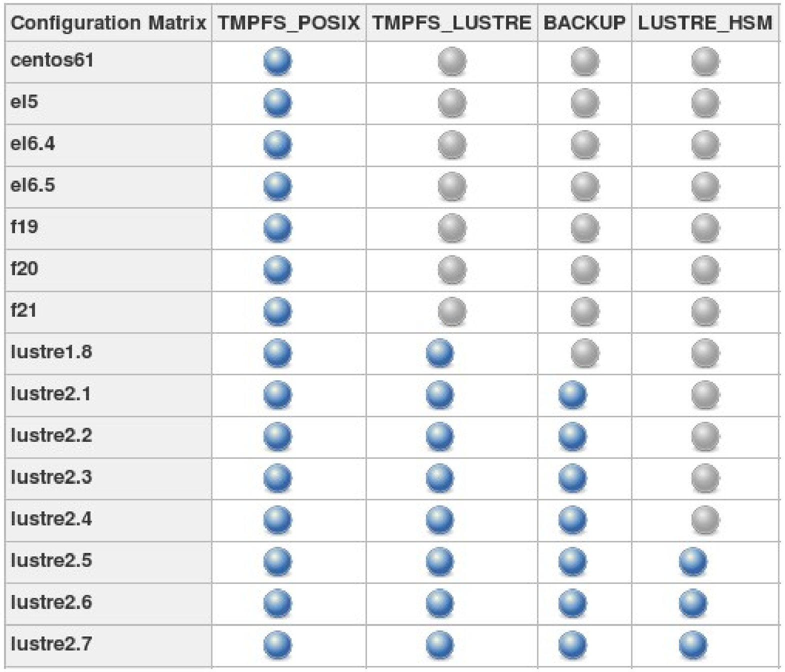
- Bash scripts + related config files
- 171 tests for Lustre filesystems
- 89 tests for POSIX filesystems
- Each test includes multiple cases and operations (up to 1min per test)

Test platform

- Jenkins on a private cluster at CEA
- Test all Lustre major versions
  - from 1.8 to 2.7+ (master)
- Test on POSIX filesystems with various OS
- Test all modes (tmpfs, backup, lhsm)

Test duration

- Run for 1 mode on a lustre config: 30min to 1h



Configuration Matrix	TMPFS_POSIX	TMPFS_LUSTRE	BACKUP	LUSTRE_HSM
centos61	🔵	⚪	⚪	⚪
el5	🔵	⚪	⚪	⚪
el6.4	🔵	⚪	⚪	⚪
el6.5	🔵	⚪	⚪	⚪
f19	🔵	⚪	⚪	⚪
f20	🔵	⚪	⚪	⚪
f21	🔵	⚪	⚪	⚪
lustre1.8	🔵	🔵	⚪	⚪
lustre2.1	🔵	🔵	🔵	⚪
lustre2.2	🔵	🔵	🔵	⚪
lustre2.3	🔵	🔵	🔵	⚪
lustre2.4	🔵	🔵	🔵	⚪
lustre2.5	🔵	🔵	🔵	🔵
lustre2.6	🔵	🔵	🔵	🔵
lustre2.7	🔵	🔵	🔵	🔵

Pre-production Platform

Validating release candidate

- Release candidate is installed on a mid-range pre-production system at CEA

- 1 tmpfs instance to manage a scratch Lustre filesystem
  - ClusterStor 9000
  - 339TB, 420M inodes
  - Lustre 2.5 servers
  - Both Lustre 2.5 and 2.7 clients (300 clients)

- 1 lhsm instance to manage a Lustre/HPSS data migration
  - SFA12K-E (4 OSS) + 1MDS
  - 684TB, 125M inodes
  - Lustre 2.5 servers
  - Both Lustre 2.5 and 2.7 clients (300 clients)
  - 3 HSM copy agents (Lustre 2.5 and 2.7)

- Last but not least: real bad users!!!
  - High metadata load

Last Step Before Release

Last step before GA: production!

- Installed on TERA100 global Lustre file system
  - 11Po filesystem
  - 200GB/s throughput
  - 4500 clients
  - Lustre 2.5

- Manage Lustre to HPSS data migration
  - Archive 100TB per day

If everything is OK: the version is officially released

Versioning policy

Versioning (X.Y.Z)

- Z: minor versions
  - Minor features, bug fixes
  - Upgrade is straight forward
    - No DB schema change (unless minor & compatible)
    - Config file compatibility

- Y: major versions
  - Major features, DB changes
  - DB schema may change (may require to rescan)
  - Config file compatibility as much as possible

- X: in-depth changes
  - Architectural changes
  - Compatibility: best effort
    - Command line may change
  - Conversion helpers can be provided

Git branches

Master: “maintenance” branch

- master = last minor release + minor fixes
- master = next minor release
- Currently, master is v2.5.x (v2.5.5+)

➔ You can safely use 'master', it is stable

Branches: old versions and next major release

- b_2.4: last 2.4.x version
- b_3.0: development branch of v3.0

Release status

Recent Robinhood versions

Previous release: 2.5.4 (Dec, 2014)

- Update stripe info on layout change (Lustre 2.4+)
- Eviction-resilent scanning
- Improved DB request batching
- Configurable DB engine

Last release: 2.5.5 (Jun, 2015)

- Lustre 2.7 support
- Unleashed DB performance with accounting off
- Policies: set default parameters for performance

Next versions

Next 2.5.x versions

- Already scheduled for 2.5.6:
  - Improved management of archive_id for Lustre/HSM
- Minor patches will still be integrated to branch 2.5 until v3 is widely adopted
  - Possible next v2.5.x

Development version: v3.0

- In development since mid-2014
- Main contributors: CEA and Cray
- Current status: stabilization, documentation, ...

- More details in next presentations...

Summary

We do our best to deliver a robust, efficient and featured software

Community continuously grows (users and contributors)

Your feedback is important

Thanks for attending Robinhood User Group!

Thanks for your attention!

Questions?

Commissariat à l'énergie atomique et aux énergies alternatives
CEA / DAM Ile-de-France| Bruyères-le-Châtel - 91297 Arpajon Cedex
T. +33 (0)1 69 26 40 00

DAM Île-de-France

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019