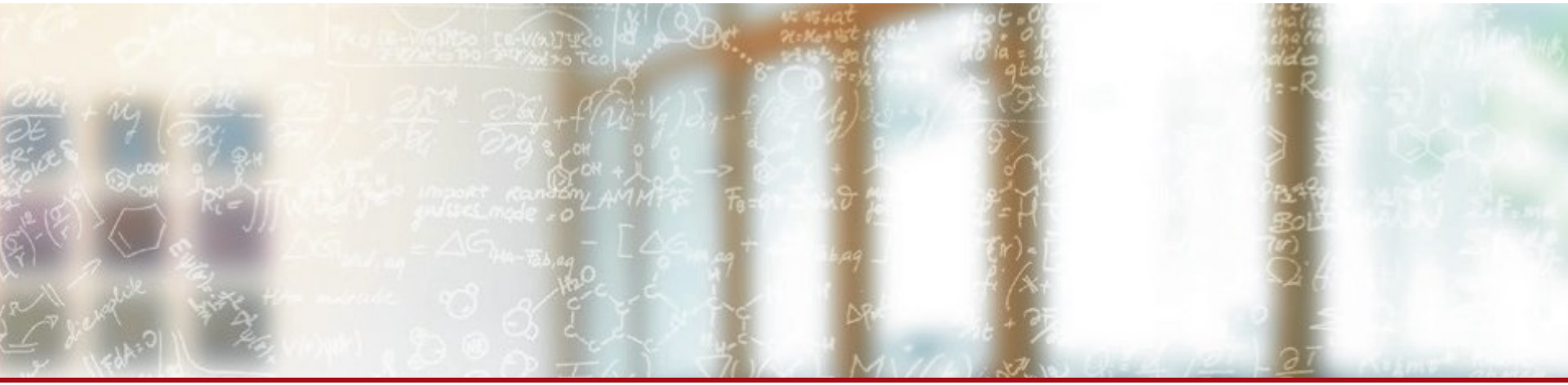
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



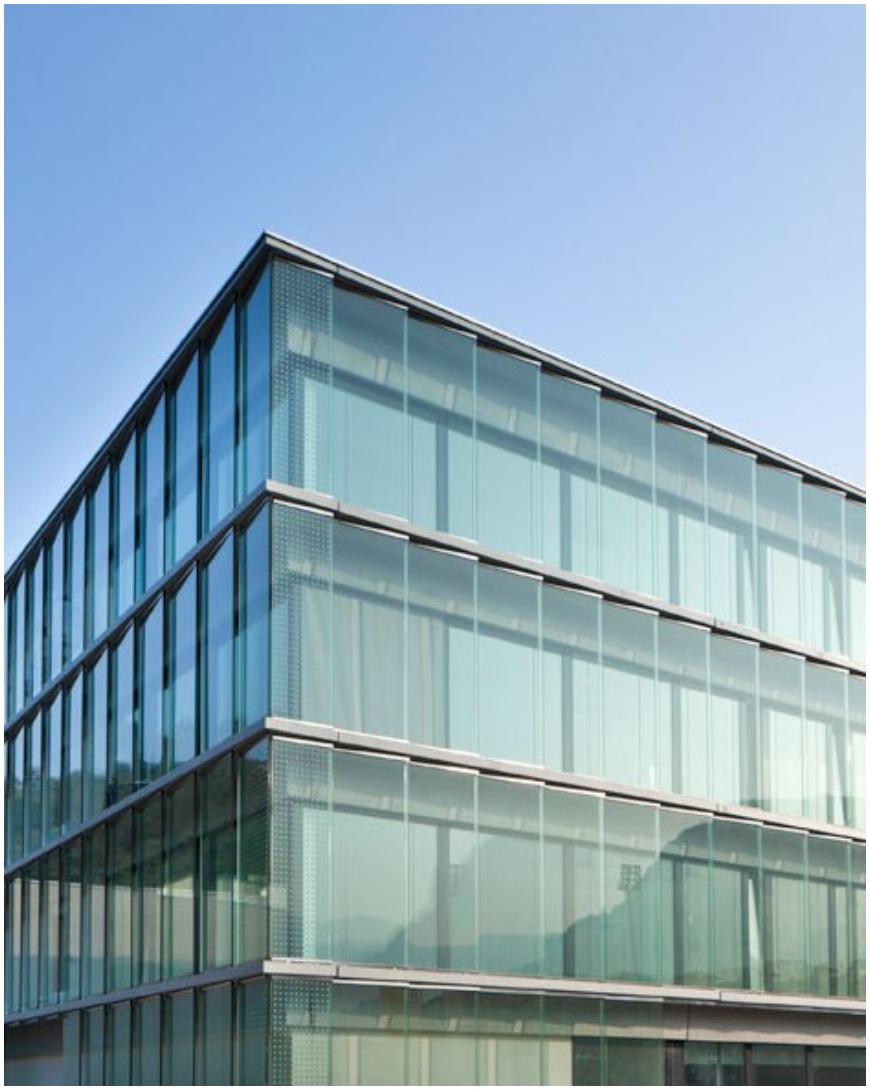
CSCS Site Status

Robinhood User Group 2015

Carmelo Ponti, CSCS

September 21st, 2015

Outline



- CSCS in a Nutshell
- Supercomputers at CSCS
- Robinhood at CSCS
- Robinhood Tuning
- Future Work

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

CSCS in a Nutshell

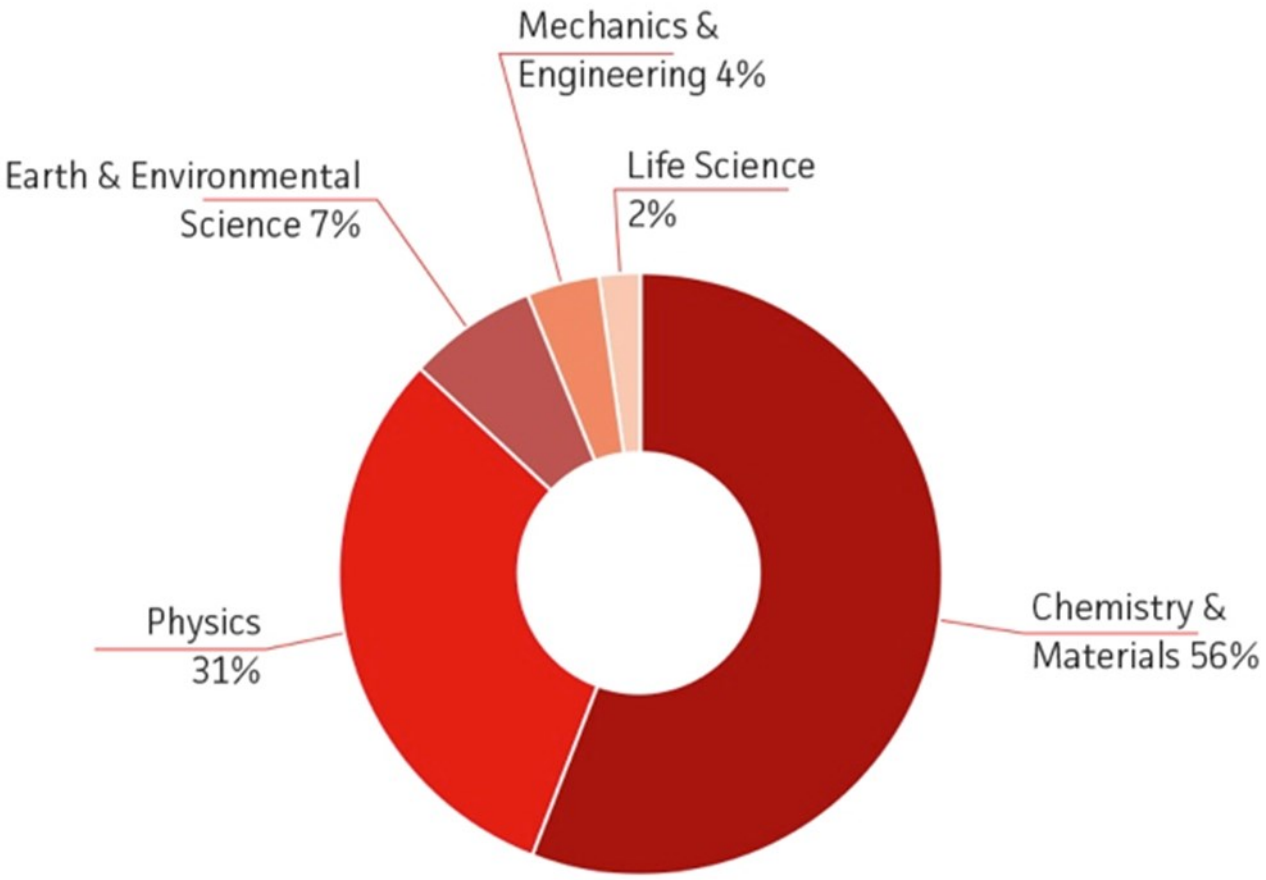
CSCS in a Nutshell

- A unit of the Swiss Federal Institute of Technology in Zurich (ETH Zurich)
  - founded in 1991 in Manno
  - relocated to Lugano in 2012
- Develops and promotes technical and scientific services
  - for the Swiss research community in the field of high-performance computing
- Enables world-class scientific research
  - by pioneering, operating and supporting leading-edge supercomputing technologies

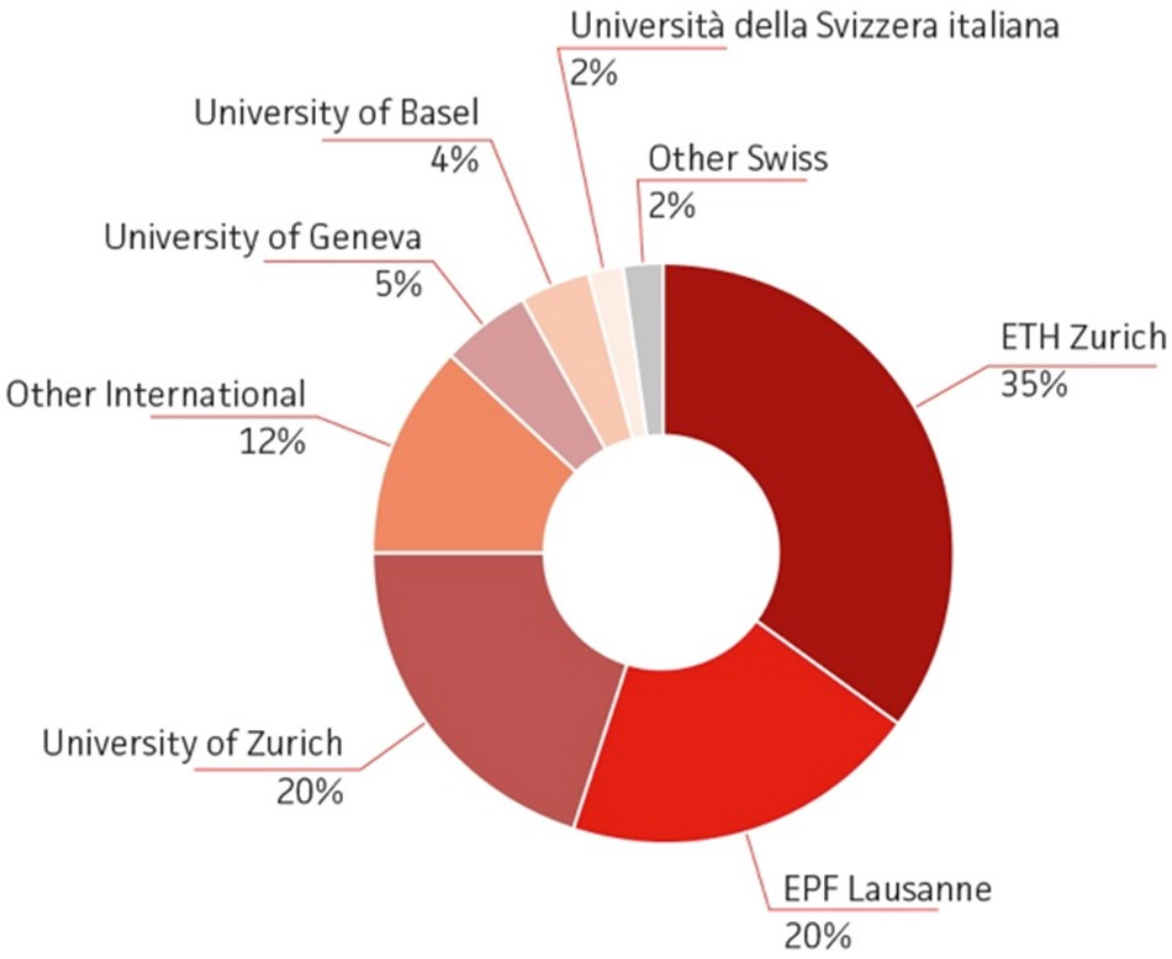
The users of CSCS

- Scientific users can access CSCS computing resources for free
  - They have to submit project requests that are assessed by international experts
  - 875 million computing hours have been used in 2014
  - 85 projects, 523 users
- CSCS operates third party systems for paying customers
  - MeteoSwiss to compute the numerical weather forecasts
  - The physicists of the Swiss universities to analyse the data from the LHC experiment at CERN
  - ETH Zurich Professors from various fields

Users by scientific field - 2014



Users by organisation - 2014



CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

Supercomputers at CSCS

A diverse pool of supercomputers

System	Supplier / Model	Installation / Upgrade	User	Theorical Peak Performance (Tflops)
Piz Daint	Cray XC30	2013	User Lab	7787
Piz Dora	Cray XC40	2014	User Lab	1246
BlueBrain 4	IBM BG/Q	2013	EPF Lausanne	839
Escha/Kesch	Cray CS-Storm	September 2015	New Meteo Swiss	2 x 190
Albis/Lema	Cray XE6	2012	Meteo Swiss	50
Phoenix	Cluster	2010 / 2011 / 2012 / 2014	CHIPP (LHC Grid)	22
Pilatus	Cluster	2012	User Lab	15

Piz Daint

- Cray XC30
- Operational since April 2013
- Extension and upgrade to hybrid in late 2013
- 5‘272 dual-socket nodes with Intel Xeon CPU and NVIDIA Tesla K20X GPU
- 168 TB RAM
- 2.7 PB local disk

Lustre file systems at CSCS

System	Lustre Supplier / Model	Capacity	Peak Performance (IOR)
Piz Daint	Sonexion 1600	2.7 PB / 192 OSTs	45 GB/s read - 120 GB/s write
Piz Dora	Sonexion 2000	1PB / 8 OSTs	9 GB/s read - 13 GB/s write
Mönch	NEC / NetApp	350 TB / 24 OSTs	18 GB/s read - 15 GB/s write
Escha/Kesch	CRAY / NetApp	2 x 73 TB / 5 OSTs	2 x 2 GB/s read - 2 x 2 GB/s write
Albis/Lema	Sonexion 1300	223 TB / 16 OSTs	4 GB/s read - 4 GB/s write

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

Robinhood at CSCS

Robinhood at CSCS

- Software
  - Robinhood 2.5.5
  - MySQL 5.1
  - Lustre client 2.5.3
  - CentOS 6.6
- Hardware
  - 16 x Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz (2 sockets, 8 cores per socket, 2 threads per core)
  - 132 Gbytes RAM (DDR3)
  - FDR (56 Gb/s)
  - SSD 250 GB (~500MB/s)
- E.g.: Daintrbh01 (daint Lustre Robinhood dedicated server)
  - 32 cores in hyper threading
  - Robinhood

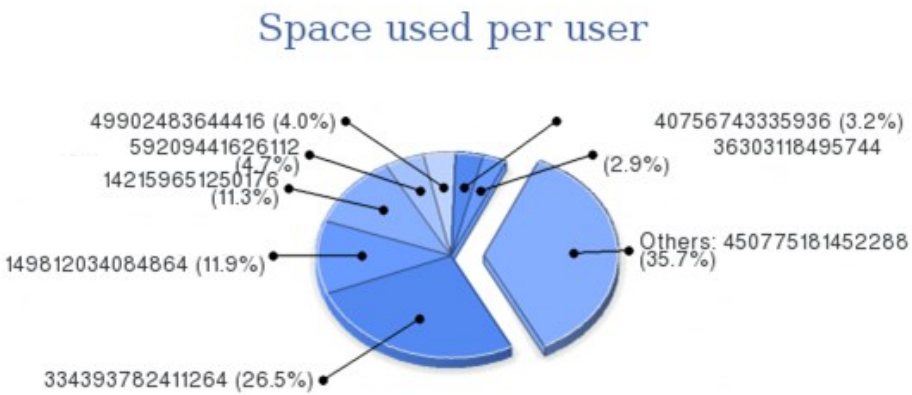
```
nb_threads = 8
nb_threads_scan = 16
nb_threads_purge = 8
nb_threads_rmdir = 8
```

  - MySQL

```
innodb_write_io_threads = 32
innodb_read_io_threads = 32
innodb_buffer_pool_size= 64G
```

Why do we need Robinhood

- CSCS first Robinhood installation: 2012
- Cleaning policies
  - All files older than 30 days are daily deleted
  - If possible keep Lustre file systems under 40% usage
- Statistics
  - We are interested on space and inodes used by groups and users
  - Some time the 30 days cleaning policy is not enough. In this cases we contact directly our user by mail (rbh-report --top-user)



Robinhood Monitoring

- Robinhood Web Site
- Ganglia
  - Number of changelog open

```
# lfs changelog xxxx-MDT0000 | wc -l
# /proc/fs/lustre/mdd/xxxx-MDT0000/changelog_users (MDS only)
```

  - Changelog speed read

```
# grep "read speed" in /var/log/robinhood.log
```

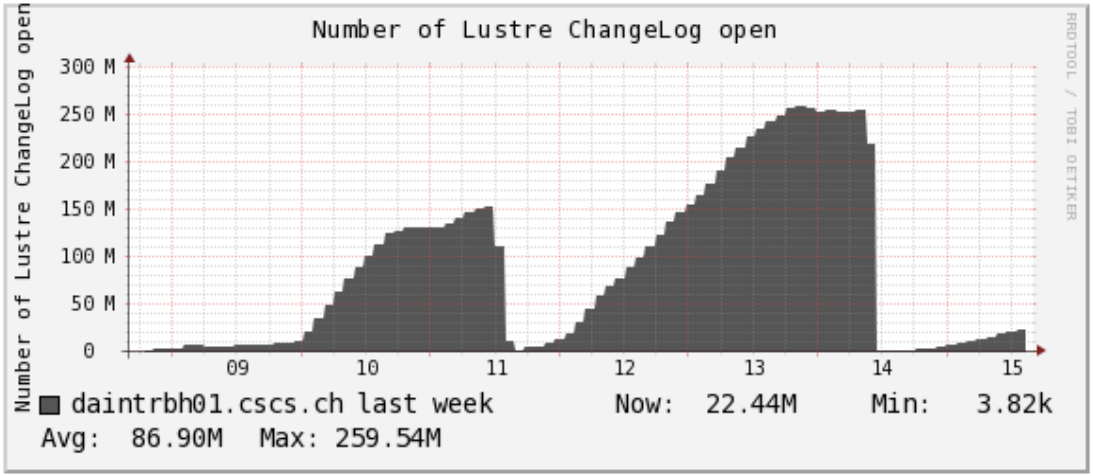
- Robinhood log file
  - GET_FID, GET_INFO_DB, GET_INFO_FS, DB_APPLY
- Memory usage

```
# watch 'cat /proc/meminfo | grep -e Buffer -e Cached -e Slab'
```

- top, ltop, slabtop, lltop (github.com/jhammond/lltop), iotop, sar, etc…

Most Common Problems

- Lustre users usage (number of Changelog entries)



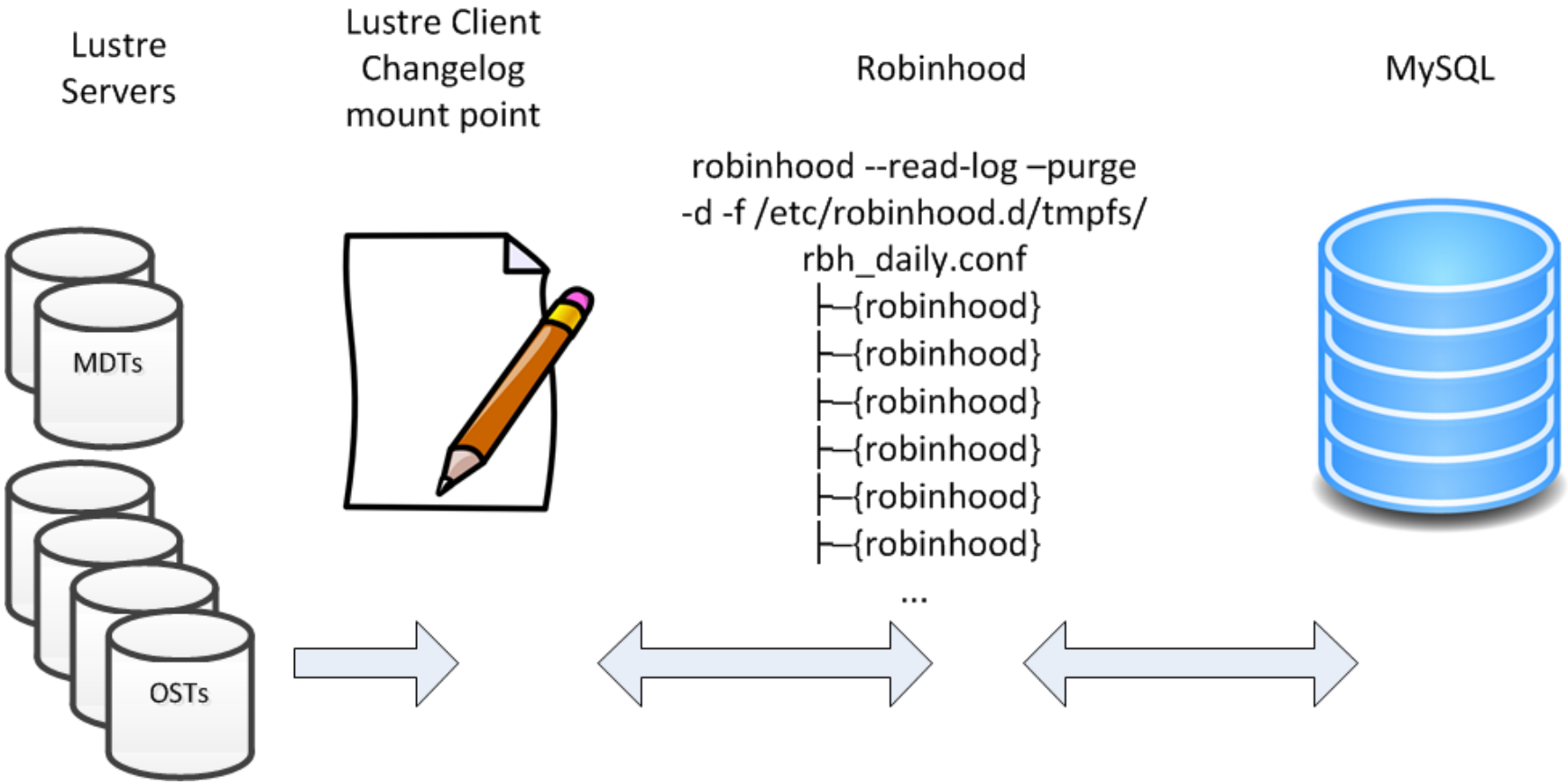
- Lustre inodes used (millions of small files)
- Changelog corruption (Lustre 2.1.6 jira.hpdd.intel.com/browse/LU-4481)

```
# umount /lustre/lnec-mdt
# mount -t ldiskfs /dev/mapper/lnec-mdt /mnt
# mv /mnt/CONFIGS/changelog_[catalog,users] /tmp
# umount /mnt
# mount -t lustre /dev/mapper/lnec-mdt /lustre/lnec-mdt
# lctl --device lnec-MDT0000 changelog_register
```

- Scan too slow (MDS overloaded or space usage > 60%)

Robinhood Tuning

Robinhood Tuning



Lustre and Changelog

- If possible we don't want to interfere with users file system usage. Anyway we are constantly monitoring the usage behavior (lltop) to prevent that only one user impact the work of all community
- Some time we have a very huge number of Changelog entries in a very short time
- Changelog speed reading must be the higher possible
  - Lustre client (after every mount)

```
# lctl set_param ldlm.namespaces.*.lru_size=400
# echo 32 > /proc/fs/lustre/mdc/lustre-MDT0000-mdc-*/max_rpcs_in_flight
(in case of many small files)
# lctl set_param ldlm.namespaces.*.lru_size=clear
# lctl set_param ldlm.namespaces.*.lru_max_age=3600
```

  - Umount and mount Lustre it speeds up the changelog reading (for a while only)
    - The Lustre client cache reaches the threshold
    - Monitor buffer, cached & slab memory usage (/proc/meminfo)
    - Tune vfs_cache_pressure and lru_max_age to reclaim the memory
    - Alternatively (to free pagecache, dentries and inodes)

```
# echo 3 > /proc/sys/vm/drop_caches
```

Robinhood Tuning

- Mainly we follow the Robinhood Manual recommendation
  - After a first initial scan we use Robinhood as following:

```
RBH_OPT="--read-log --purge –rmdir”
```

  - Robinhood parameters

```
commit_behavior = autocommit;

user_acct  = enabled;
group_acct  = enabled;

nb_threads = 8 ;
max_pending_operations = 100000 ;

batch_ack_count = 1024 ;
force_polling    = ON ;
polling_interval = 1s ;

nb_threads_purge = 8 ;
post_purge_df_latency = 1min ;
trigger_on          = OST_usage ;

nb_threads_rmdir = 8 ;
```

MySQL Tuning

- MySQL with InnoDB engine

```
[mysqld]
…
max_connections= 128
…
innodb_buffer_pool_size= 64G
…
innodb_additional_mem_pool_size = 16M
innodb_file_per_table = 1
innodb_flush_method=O_DIRECT
innodb_write_io_threads = 32
innodb_read_io_threads = 32
innodb_io_capacity=50000
innodb_log_files_in_group = 4
```

- Future work: test MariaDB with Tokudb engine (not compressed)

Numactl

- numactl - Control NUMA policy for processes or shared memory
- numactl --hardware

```
available: 2 nodes (0-1)
node 0 size: 65381 MB
node 0 free: 2098 MB
node 1 size: 65536 MB
node 1 free: 3357 MB
```

- Robinhood

```
# /etc/init.d/robinhood
…
DAEMON="/usr/bin/numactl --cpunodebind=1 --membind=1
        /usr/sbin/robinhood"
```

- MySQL

```
# /etc/init.d/mysql
…
/usr/bin/numactl --cpunodebind=0 --membind=0 $exec
                 --datadir="$datadir" --socket="$socketfile"
                 --pid-file="$mypidfile"
                 --basedir=/usr --user=mysql >/dev/null 2>&1 &
```

Future Work

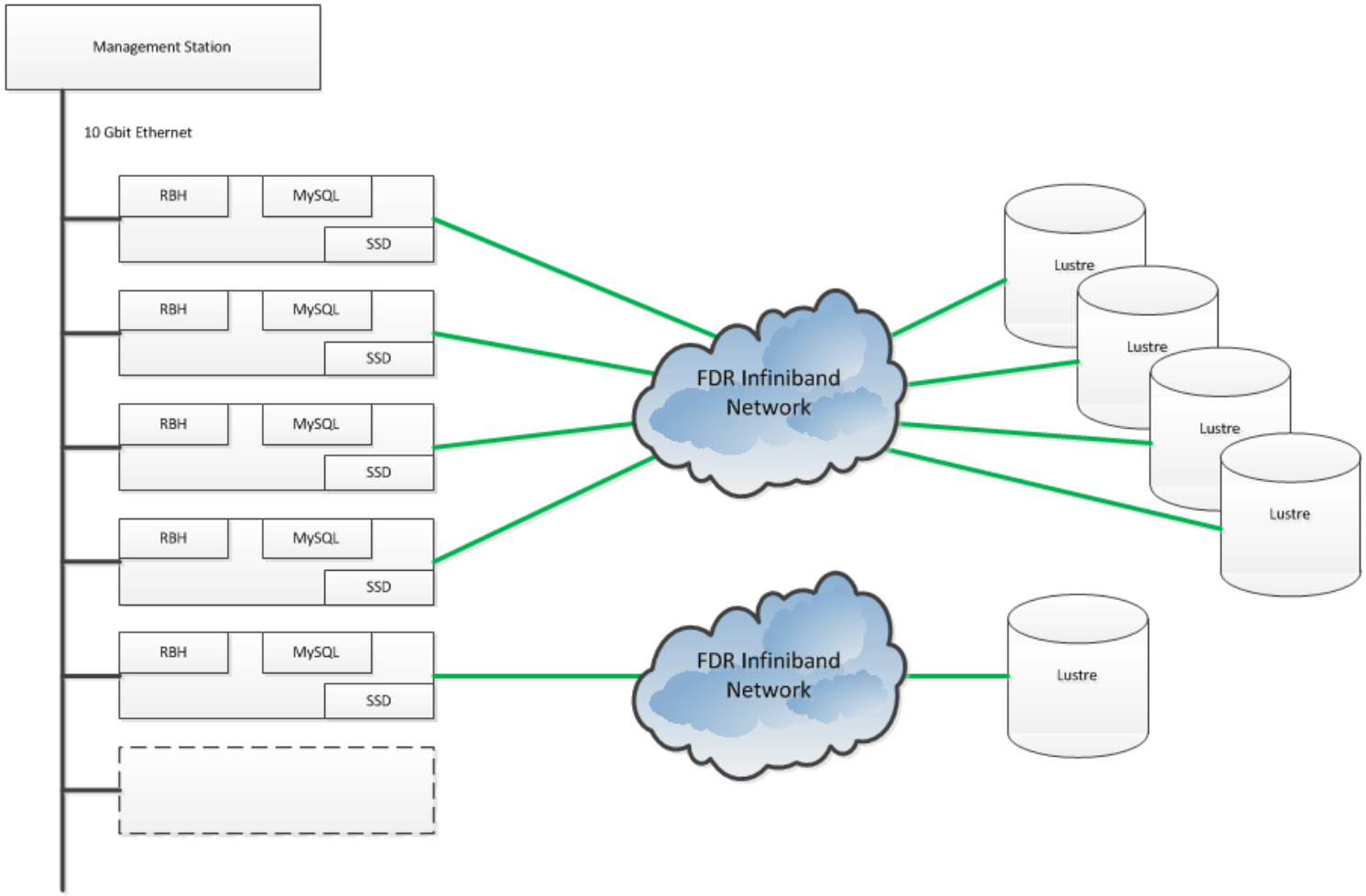
Future work - Contact the Users

- Automatically contact the users (by mail) in case they exceed some thresholds

```
Dear Carmelo Ponti,

Your user 'cponti' has exceeded quota limits on
filesystem /scratch/daint since 2015-09-16 09:10:01
(1 days 0 secs).

Your Occupation is:


      Used   Soft Limit   Hard Limit
 56.71 TiB    80.00 TiB   500.00 TiB

     Iused   Soft Limit   Hard Limit
   1.41 M    1000.00 K      30.00 M

Please free the resource as soon as possible,

Best Regards,
CSCS Storage Team
```

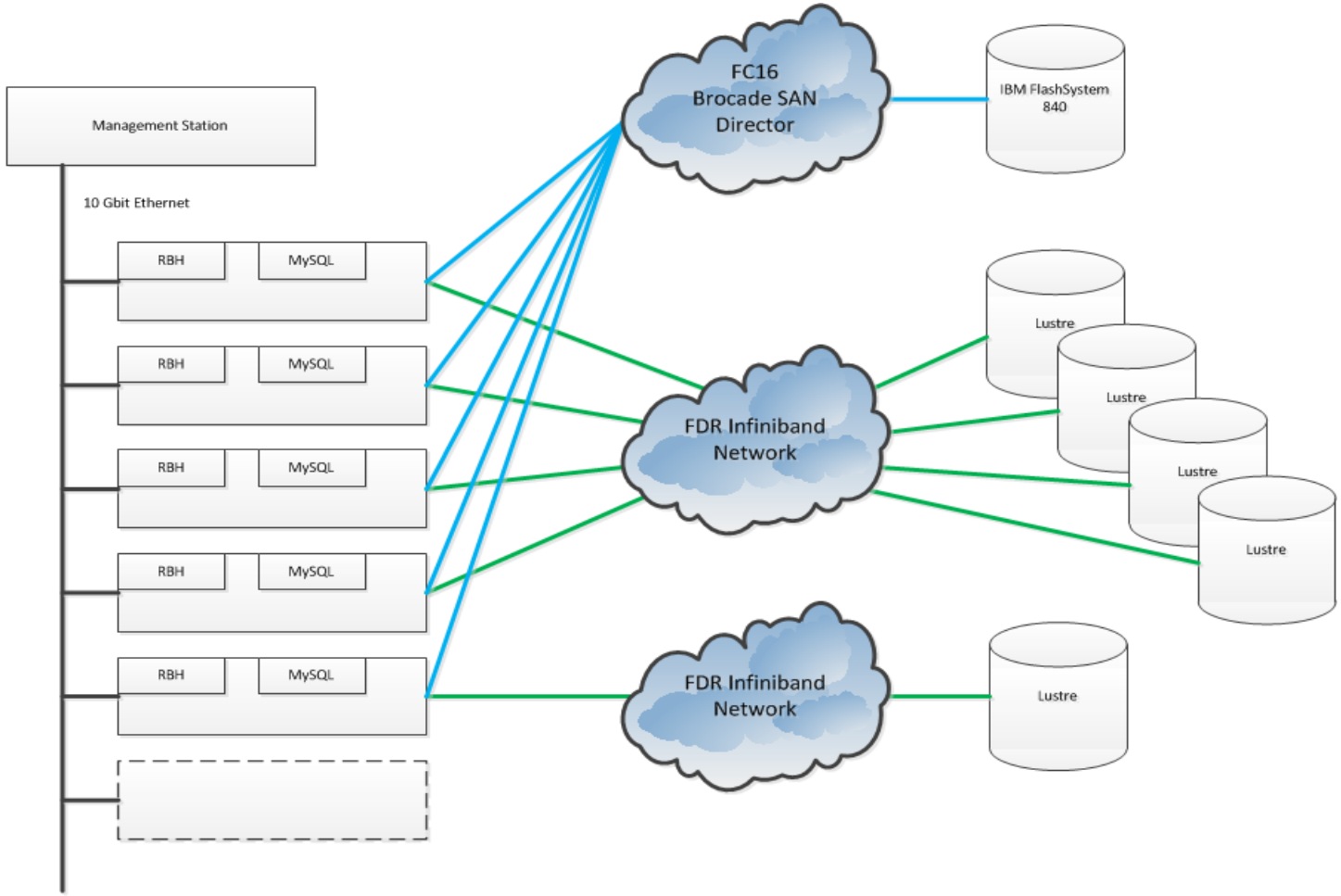
Future Work - Consolidation

- Consolidate all Robinhood servers in one cluster



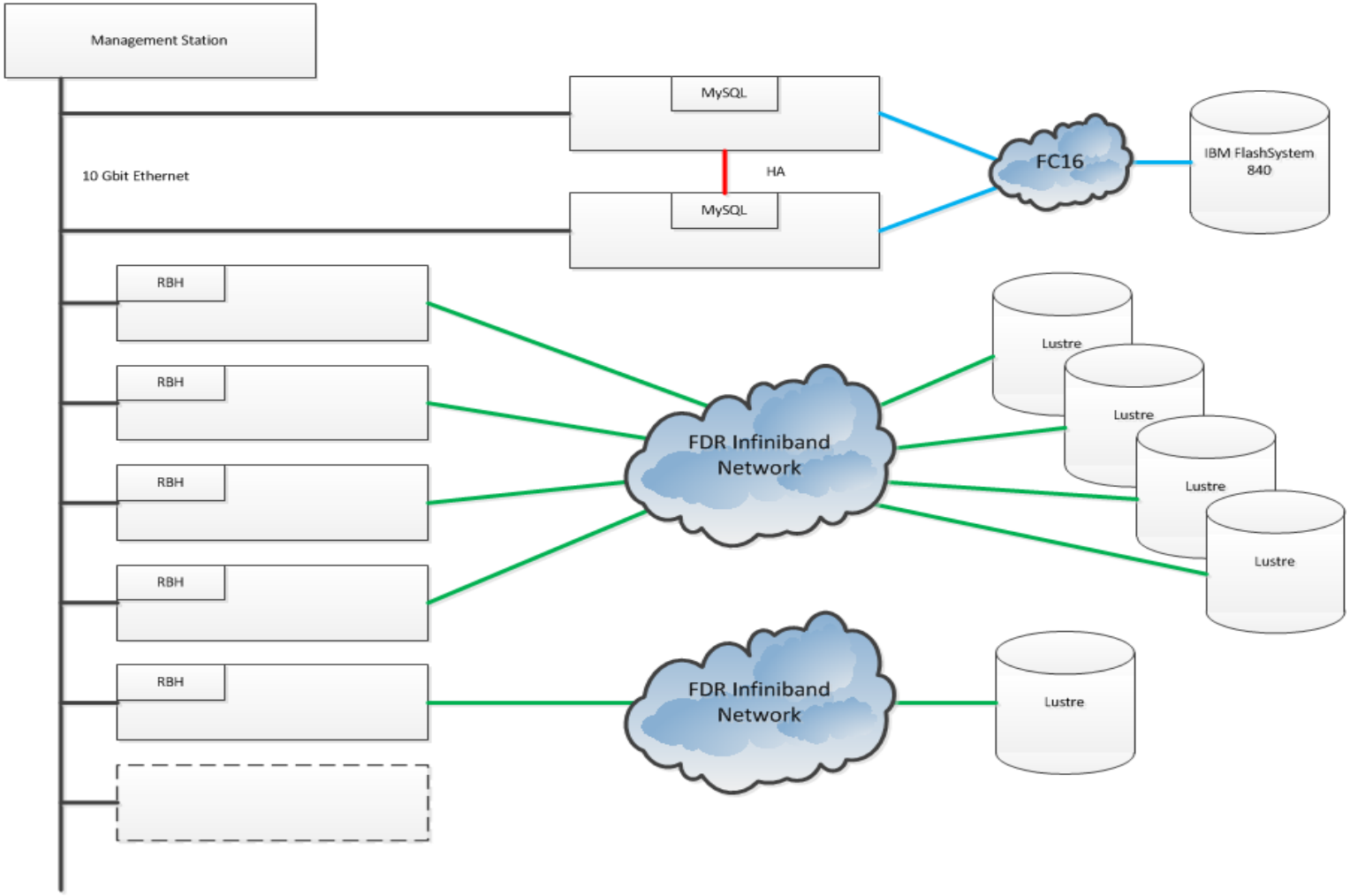
Future Work – Scenario 1

- Substitute all local SSD with a centralized FlashSystem



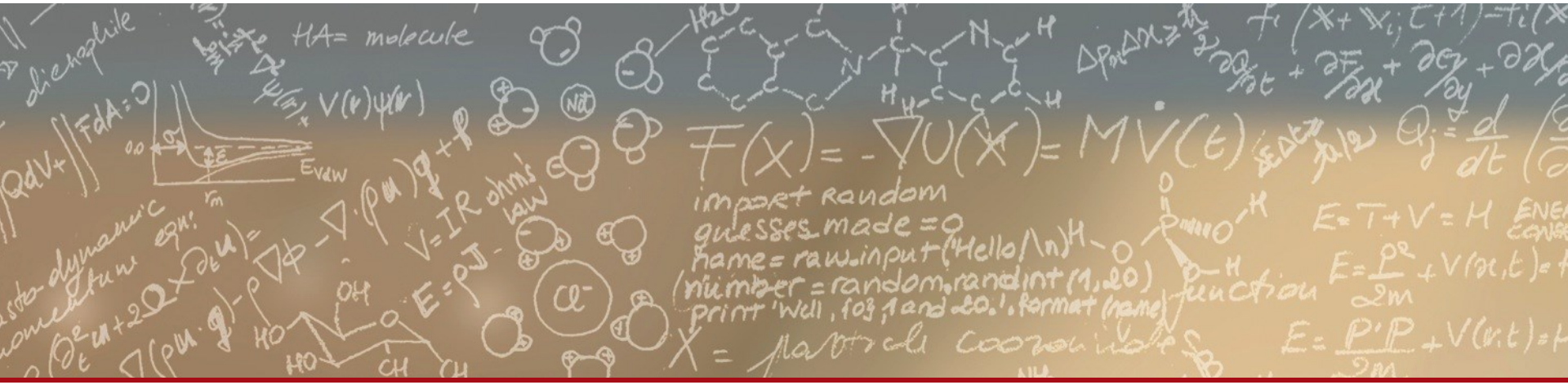
Future Work – Scenario 2

- Centralize MySQL in HA Cluster



Acknowledges

- Thomas Leibovici (for all mailing list support)
- Kilian Cavalotti (LU-4481 problem)
- Chris Hunter (Changelog tuning advices)



Thank you for your attention.