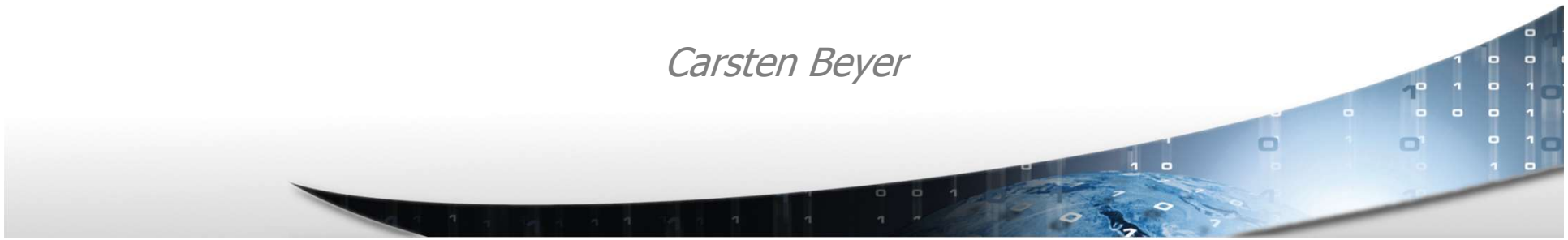
DKRZ

DEUTSCHES
KLIMARECHENZENTRUM

Robin Hood 2.5 on Lustre 2.5 with DNE

Site status and experience at German Climate Computing Centre in Hamburg

Carsten Beyer

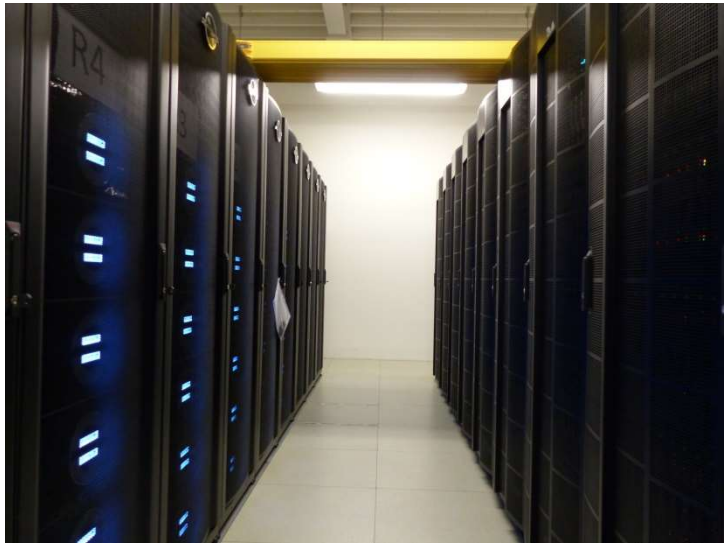


About DKRZ

- High Performance Computing Center
  - Exclusively for the German Climate Research
  - Limited Company, non-profit
- Staff: ~ 70
- Services for Climate Research:
  - Support for Scientific Computing and Simulation, Model Optimization, Parallelization
  - Data Management and Archiving
  - Data Visualization (3D Graphics and Video)
- University Research Group: HPC (Prof. Dr. Ludwig)

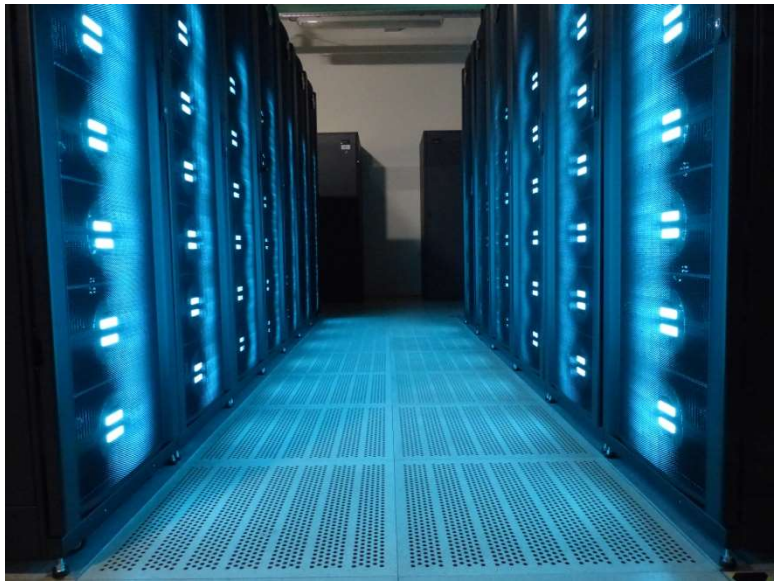
Mistral

- First phase 2015 (second phase 2016) , total cost: 41 Mio Euro
- Bull Supercomputer: 26 Mio Euro
  - Bullx B700 DLC-System
  - 37.000 (+57.000) cores (Intel Haswell/Intel Broadwell)
  - 1.500 nodes (2x 12 Cores) (+1600 nodes 2x 18 Cores)
  - 1,4 (3,1+) PetaFLOPS
  - 75 TB (200 TB) main memory
  - Infiniband FDR
- Parallel file system:
  - Lustre, ca. 21 (+33) PetaByte
  - Throughput > 0.5 TeraByte/s



Lustre - ClusterStor

- Seagate ClusterStor Setup (Phase 1 / Phase 2)
  - 62 OSS / 124 OST's / 6TB disks
  - 5 MDT
  - 21 PB / max. 6 Billion files
  - Lustre 2.5.1
  - Infiniband FDR



  - + 74 OSS / 148 OST's / 8TB disks
  - + 7 MDT
  - + 33 PB / max. 8 Billion files
  - Lustre 2.5.1
  - Infiniband FDR

Robinhood - Usage

- Report generation (rbh-du)
  - HOME-directories (‚user'-Quota) -> 4 times a day
  - SCRATCH-directories (‚user'-Quota) -> 4 times a day
  - WORK-directories (‚project'-Quota) -> once per day
    - Total usage per project
    - Usage per user in a project
      - up to 150 user in a project
      - Currently 220 ongoing projects
- Deletion of files and empty directories in SCRATCH area (older 14 days)
  - configured but not enabled yet

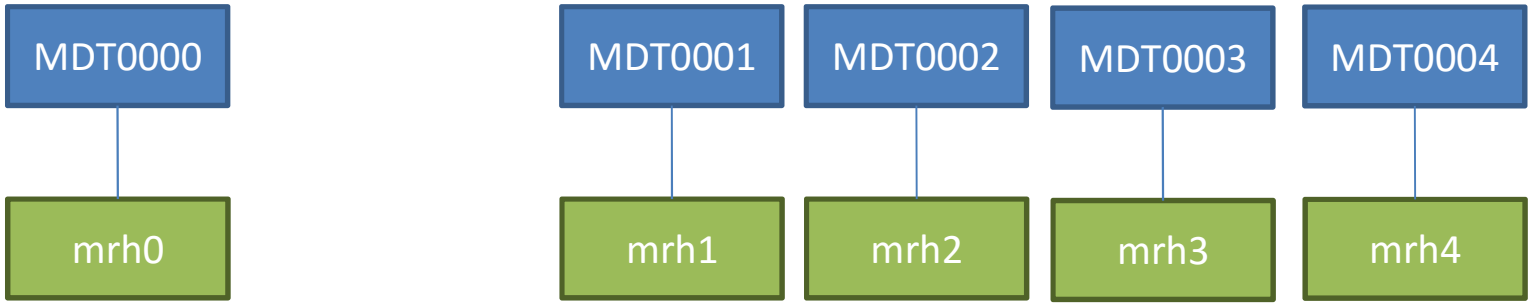
Robinhood Setup – Phase 1

- Robinhood Hardware Setup (5 Server, 1 per MDT)
  - 256 GB Memory
  - 4x SSD (500 GB) in two RAID 1 sets
    - 1x OS and logs (ext4)
    - 1x MariaDB /var/lib/mysql (xfs)
  - 2x Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz (12 Cores Haswell)
  - Infiniband FDR

- Robinhood Software Setup
  - RHEL 6
  - MariaDB 10.0.21
  - Robinhood 2.5.5-5

Robinhood Setup – Phase 1

One Robinhood server (changelog reader) per Metadata Server



/mnt/lustre01

../pf (HOME dirs)

../pool (Common data)

../sw (Software tree)

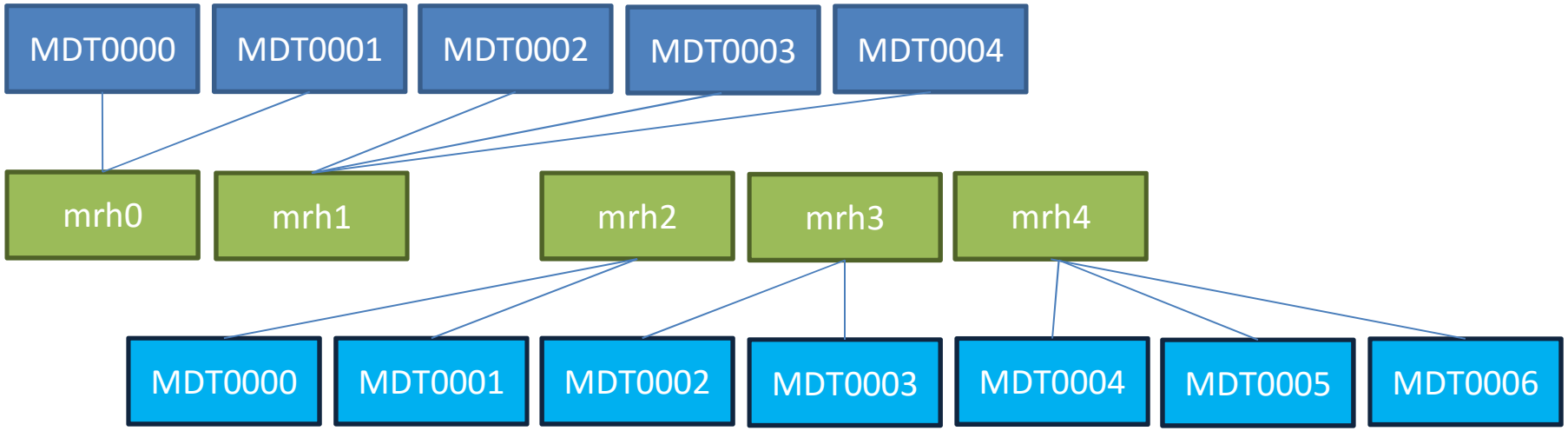
../scratch/[a,b,g,k,m,u]

../work/[project1,…]

Needs ignore list for each Robinhood server for the FS scan

Robinhood Setup – Phase 2 Plan

Robinhood Server / Metadata Server (Plan)



- Current plan: consolidate RBH server and metadata server
  - Phase 2 storage is about 33 PB / max. 8 billion files
  - Extension for the /work area

RBH issues – At beginning

- Crash of Lustre client on RBH-Server
  - Up to 30 times per day per server (reboot + crash dump)
  - Fixed by newer kernel and newer Lustre-client provided by Seagate

- Database Performance (Fixed by a few lines in /etc/my.cnf.d/server.conf)

  innodb_buffer_pool_size=128GB

  innodb_buffer_pool_instances=128

  innodb_thread_concurrency=48

  innodb_max_dirty_pages_pct=15

  innodb_log_buffer_size=32M

  innodb_log_file_size=500M

  max_connections=512

Robinhood and DNE

- Setting up an ignore list for each RBH server for FS scan
  - Each server needs it‘s own ignore list if you have separate RBH-server for each MDS

- At DKRZ:
  - Automatic generation of ignore lists and Robinhood configuration file per server 4 times a day
    - ‚lfs getstripe –M /mnt/lustre01/work/k20200‘ to get info for each highlevel dir
  - Md5sum running config vs. new config and reload if new config differs

```
FS_Scan {
    Ignore {
        tree == /mnt/lustre01/pf
        or tree == /mnt/lustre01/pool
```

Rare cases - changelog

- ‚Changelog stuck‘
  - Repeating disconnect/connect from/to MDT errors in /var/log/syslog
  - Lustre changelog not read anymore
  - STATS section in robinhood.log remains same, see e.g. changelog recored/fid in
    GET_INFO_DB : first/last
    GET_INFO_FS : first/last
  - Trying ‘lfs fid2path’ manually for these four fid’s, one hangs and could not be killed
    - burst of activity -> read changelog -> process fid2path for a record in that read changelog -> something triggers disconnect
    - disconnect is temporary -> fid2path is now hung

Rare cases - changelog

- ‚Repair' - action
  - Stop Robinhood service (remains some minutes in [defunct])
  - Reboot the node
  - Empty DB with: rbh-config empty_db robinhood_lustre01'
  - Clear changelog: lfs changelog_clear lustre01-MDT0001 cl1 0
  - Start Robinhood service (--readlog)
  - Start initial scan with: /usr/sbin/robinhood --scan –once
    - Scan speed is about 5 to 6 million entries per hour (9 to 12 hours)

Rare cases

- MySQL/MariaDB growing ibdata1 file
  - On one server this file was growing to about 49 GB
  - Filesize does not shrink even if db is emptied (known bug)
  - On other server the filesize ranges from 60 MB to 650 MB

- Proposed repair procedure (currently not done)
  - Do a mysqldump of all databases, procedures, triggers etc except the mysql and performance_schema databases
  - Drop all databases except the above 2 databases
  - Stop mysql
  - Delete ibdata1 and ib_log files
  - Start mysql
  - Restore from dump

Report for each user in a project

```
[root@mrh1 ~]# rbh-report --fs-info
Using config file '/etc/robinhood.d/tmpfs/mrh1.conf'.
type    ,      count,     volume,   avg_size
symlink ,     465800,   33.93 MB,         76
dir     ,    2170829,   16.07 GB,    7.76 KB
file    ,   66976609,    2.99 PB,   47.87 MB

Total: 69613238 entries, 3361826583341465 bytes (2.99 PB)

[root@mrh1 ~]# time rbh-du -d -k /mnt/lustre01/work/bmx825
Using config file '/etc/robinhood.d/tmpfs/mrh1.conf'.
/mnt/lustre01/work/bmx825
        symlink count:39625, size:5046, spc_used:126984
        dir count:1047724, size:8880796, spc_used:8927156
        file count:32402106, size:77091372434, spc_used:77155346111

real    14m32.417s
user    0m29.672s
sys     0m52.994s
```

Report for each user in a project

```
[root@mrh1 ~]# time /usr/bin/rbh-du -d -k -u b324031
/mnt/lustre01/work/bmx825
Using config file '/etc/robinhood.d/tmpfs/mrh1.conf'.
/mnt/lustre01/work/bmx825
        symlink count:1597, size:256, spc_used:5856
        dir count:79905, size:625776, spc_used:627720
        file count:2161979, size:1123752735, spc_used:1125492068

real    15m27.082s
user    0m38.584s
Sys     0m15.304s
```

Problem: 150 users in this project, report for each user takes 15 min.

With 5 rbh-du in parallel, time increases, but could be done in some hours

Things to do / test

- Tuning to be tested
  - rbh-config optimize_db
  - Set ‚innodb_file_per_table = 1‘ (growing ibdata1 file)
  - mysqltuner.pl
  - Number of nb_threads (currently 16)
  - Lustre parameter on RBH server

- Robinhood v3.0
  - Installed on testsystem with MariaDB 10.1.16

Thank you for your Attention!

http://www.dkrz.de

Carsten Beyer

beyer@dkrz.de

Questions ?