CVMFS 2.11 - Improvements and outlook

Nov 8th, 2023, GDB
CVMFS Development Team@ CERN: Laura Promberger,
Valentin Völkl, Jakob Blomer

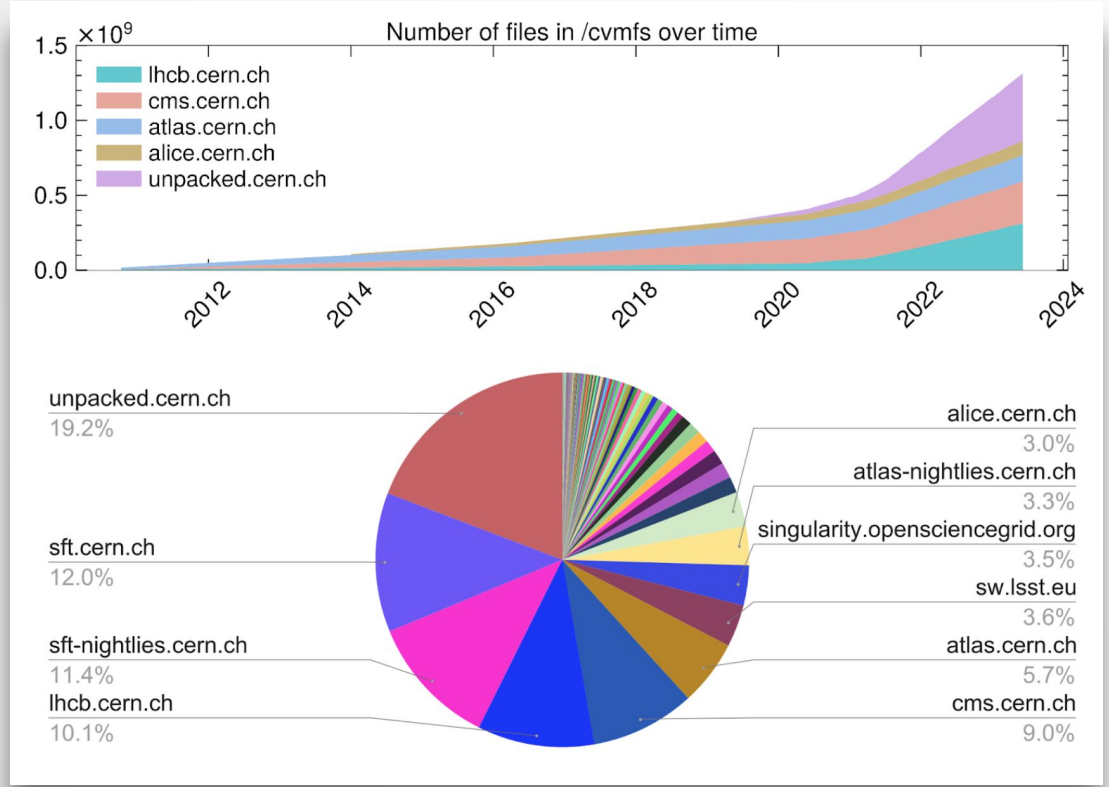
Introduction

Team:

- This presentation represents the perspective of the CernVM-FS Development Team in EP-SFT (Valentin Völkl, Laura Promberger, Jakob Blomer)
- Credit and thanks go to all the operators and admins
  - in particular in IT-SD for the CERN infrastructure
  - And our Knowledge-Transfer partner *Jump Trading*

CVMFS 2.11:

- 2.11 is a big release: many fixes and performance improvements,
  - few “new features”
  - In line with input from experiments, asking us to prioritize “bug fixes, client and Stratum 0 speedup rather than new features”
- Enables running certain workloads (see below)
  - dedicated update campaign?



Software stacks are getting larger, more complex, putting more load on CVMFS.

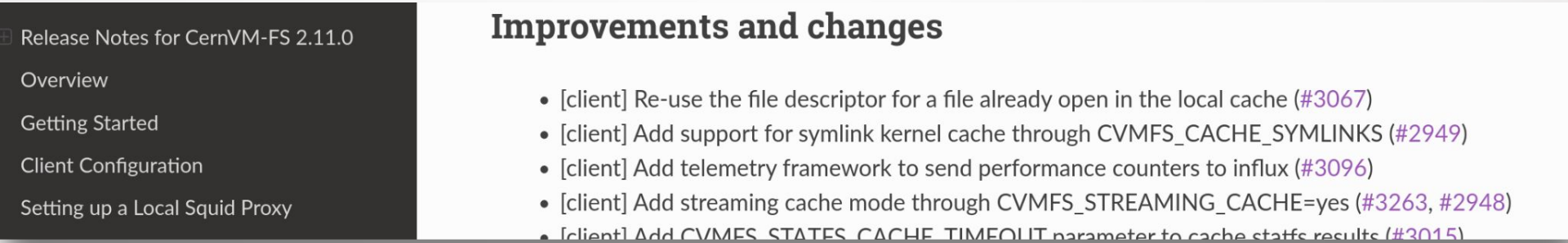
CVMFS can and does scale, but needs to continue efforts dedicated on performance

Release infrastructure

- Feature releases every 6-12 months, with patch releases as needed in between
- Upgradable without needing to stop processes / remount thanks to hotpatching capabilities
- Provide binaries in package repositories for many linux distributions, as well as MacOS and a service container
- Released packages are first staged in a testing repository for 3 / 7 days before being put into production
- Considering a “release-candidate” or “nightly/bleeding edge” repository to more quickly release fixes for specific problems

CVMFS 2.11

Will not link to individual issues, see full changelog for more details:



2.11 Improvements in **Logging**

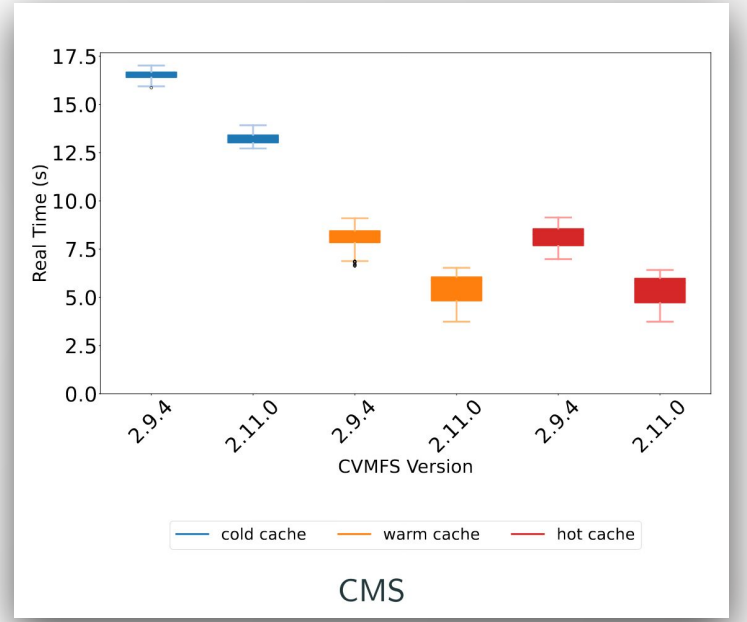
… are crucial, as many errors are hard to reproduce by the developers. Many small improvements were added in 2.11:

- Debug mode now preserved across cvmfs_config reload
- Debug output of internal Curl queries now available in the debug log
- Dedicated “*.mount” log files for the mount helper
- Re-store core-file generation after credentials drop
- Improve logging of FUSE I/O errors

2.11 also introduces client telemetry that can be used with InfluxDB (link) and custom http tracing headers

Performance improvements for caching

- Page Cache Tracker: Much better use of kernel page cache (already in 2.10)
- CVMFS_SYMLINK_CACHE possible on new enough FUSE/Kernel versions
  - Requires libfuse 3.10+
  - And kernel in rhel8+
- `Statfs` caching



See CHEP 2023 for more details

Reference-counting Cache Manager

- `CVMFS_CACHE_REFCOUNT`: fixes a long-standing issue with many processes concurrently reading the same files; impacted ALICE in particular
  - cvmfs would open new file descriptors for the same files, sometimes reaching the system limit
  - Can be worked around, but requires effort on sys-admin side
- New cache manager mode keeps a table with references to file descriptors in memory and no longer duplicates them
  - Small overhead, but should not exceed a few MB of memory
- Already enabled for ALICE in the cvmfs-config.cern.ch repository, will become default in 2.12
  - But needs 2.11.2 in order to avoid having to increase the cvmfs file descriptor limit

Improvements for external data

External Data: used by LIGO, osgstorage/stashcache, and (privately) by Jump Trading

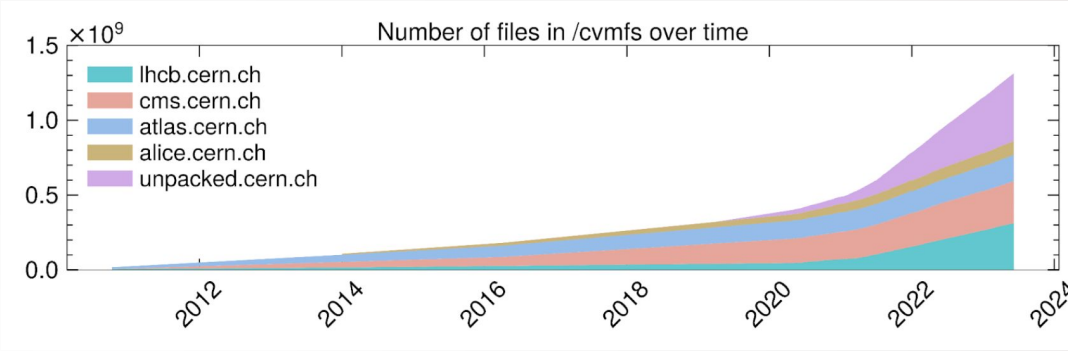
- `CVMFS_CACHE_STREAMING`: 2.11 introduces a new “streaming” cache manager mode. This bypasses the cache completely, except for catalogs. Useful in a very special set of circumstances, mostly not for the software distribution usecase

- Protected extended attributes: allows to restrict xattrs by uid

Patch Release 2.11.1 + 2.11.2

- Fixes three important problems (mostly race conditions)
  - Fix for error in client reload if config repository is unavailable
  - Mount helper race condition that could spawn additional processes
  - Error handling when trying to auto-mount non-existing repositories, avoiding spurious directories in /cvmfs
- Changes and improvements in packaging
  - Watch out for changes if not using the cvmfs-release package!

Outlook: unpacked.cern.ch



- Very useful bridge to container deployment model
  - And lower-barrier entry to cvmfs publishing
- Many improvements that will be included in 2.12, following successful summer student project
  - REST API
  - Major refactor
- Can possibly free up some space by garbage-collection campaign

Outlook on possible new features (2.12)

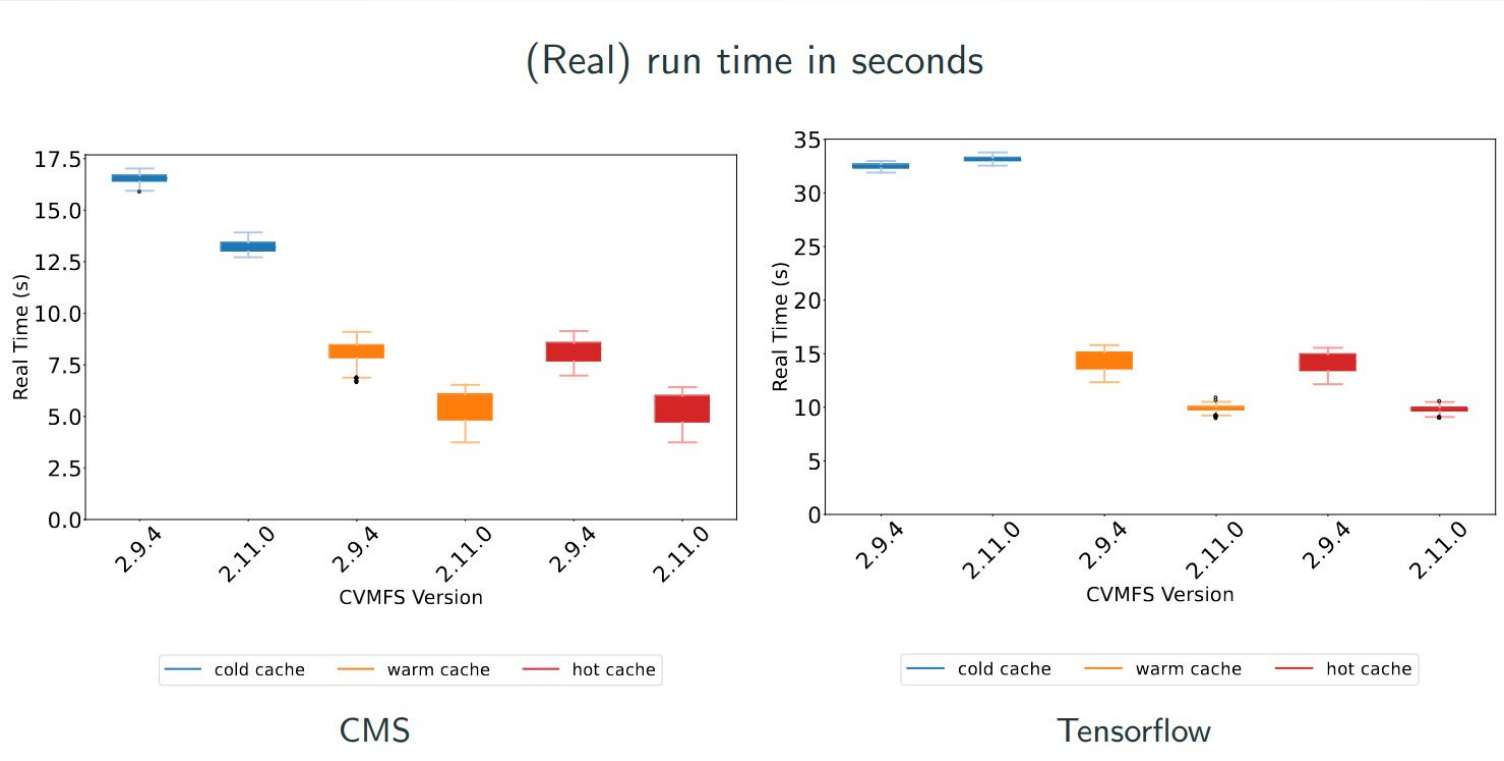
- File Bundles
  - Groups downloads of files that are accessed together
  - Can improve interactive access
- Container tools and ephemeral write shell
  - Helm charts
- Zstd compression

Conclusion

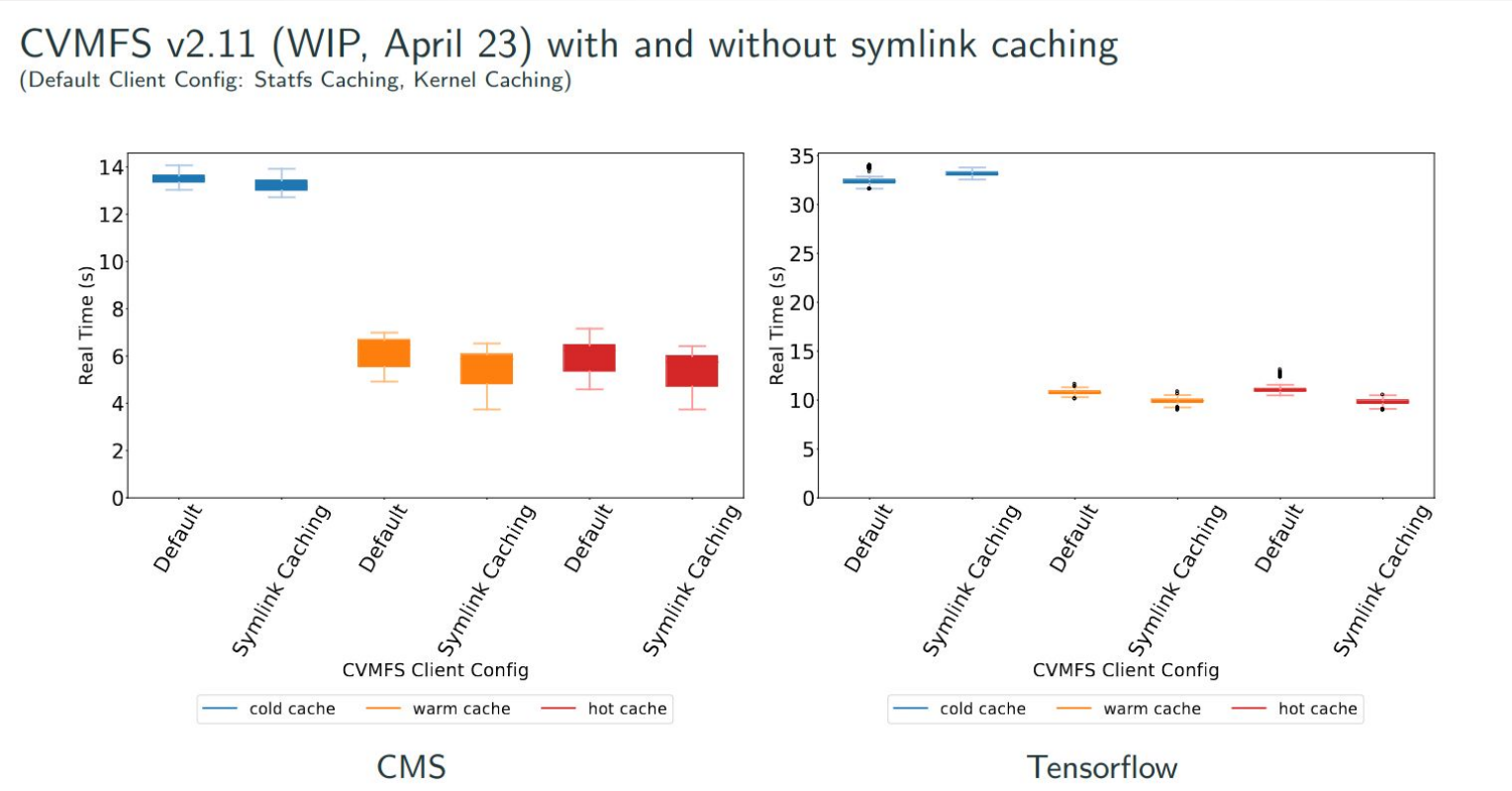
- **CVMFS 2.11** has a long list of improvements and fixes
  - Upgrading highly recommended, as it also enables certain concurrent workloads to run
- Continue developments to scale to HL-LHC and beyond

Thank you ! Questions?

Backup


(Real) run time in seconds
Real Time (s)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2.9.4
2.11.0
2.9.4
2.11.0
2.9.4
2.11.0
CVMFS Version
cold cache
warm cache
hot cache
CMS
Real Time (s)
35
30
25
20
15
10
5
0
2.9.4
2.11.0
2.9.4
2.11.0
2.9.4
2.11.0
CVMFS Version
cold cache
warm cache
hot cache
Tensorflow

Backup


CVMFS v2.11 (WIP, April 23) with and without symlink caching
(Default Client Config: Statfs Caching, Kernel Caching)
Real Time (s)
0
2
4
6
8
10
12
14
Default
Symlink Caching
Default
Symlink Caching
Default
Symlink Caching
CVMFS Client Config
cold cache
warm cache
hot cache
CMS
Real Time (s)
0
5
10
15
20
25
30
35
Default
Symlink Caching
Default
Symlink Caching
Default
Symlink Caching
CVMFS Client Config
cold cache
warm cache
hot cache
Tensorflow