CernVM 2021 Virtual Workshop Summary

Jakob Blomer

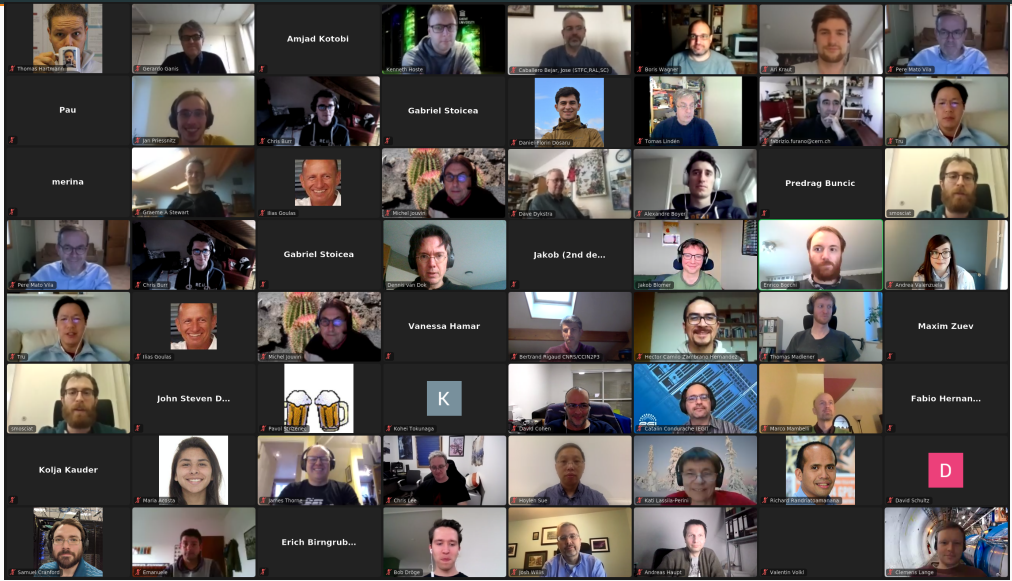
GDB, 10 February 2021

CernVM 2021 Virtual Workshop – Overview

- 2 days from 1-2 February
- Organizers: Dennis van Dok (NIKHEF) and Jakob Blomer, Enrico Bocchi, Simone Mosciatti, Andrea Valenzuela (CERN)
- Online format, afternoon sessions to allow for US participation
- 99 registered participants, 10 % to 20 % non-HEP
- Typically 45–65 participants connected
- Virtual drink on mibo
- Main topics this year: CernVM-FS as a **container hub**, new **publishing workflows**

Materials and recordings: https://indico.cern.ch/e/cvm21

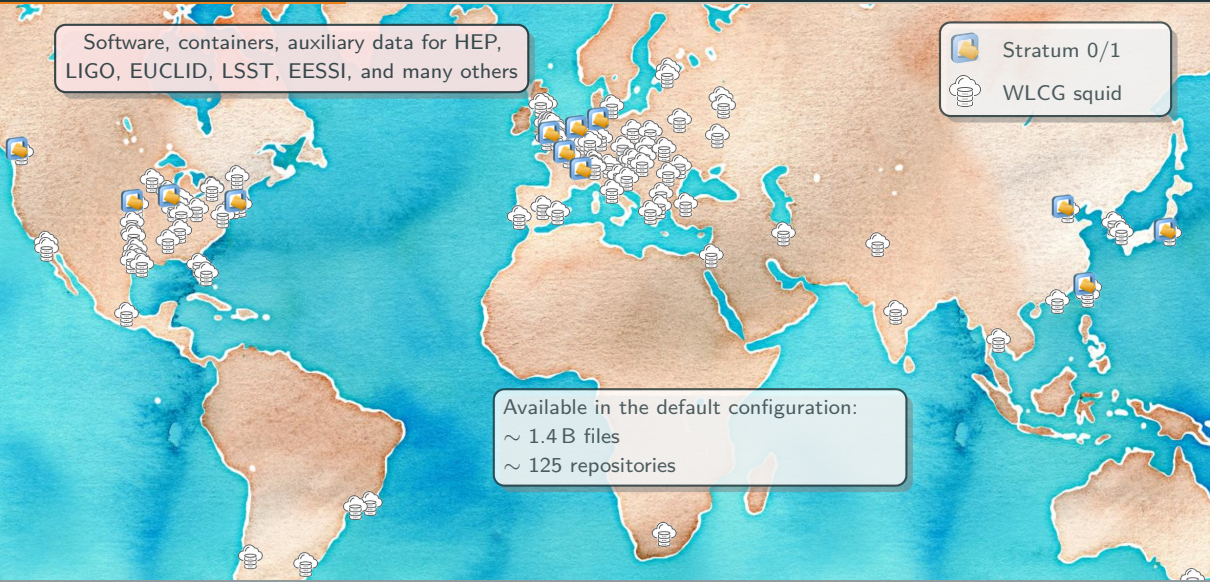
Conference Picture



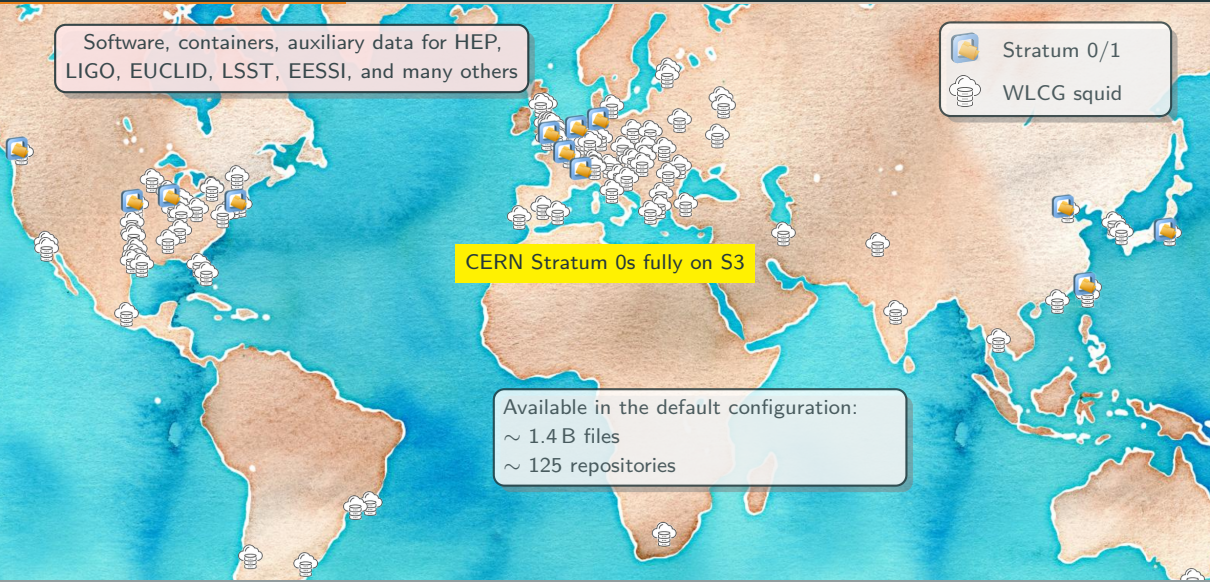
Highlights Since the 2019 Workshop

- Two CernVM-FS feature releases: 2.7 and 2.8
- New presence: web site, monitoring site, forum: https://cernvm-forum.cern.ch
- Leap in container management
  - /cvmfs/unpacked.cern.ch established as container hub with 750+ images
  - Scalable image distribution in WLCG (e.g., folding@home)
  - Integration with major container runtimes: singularity, containerd/k8s, podman
- New and faster publishing workflows: “template transactions”, “ephemeral writable shell”
- CernVM 5: EL8 based prototype
- Presentation links: ▸ CernVM-FS ▸ CernVM

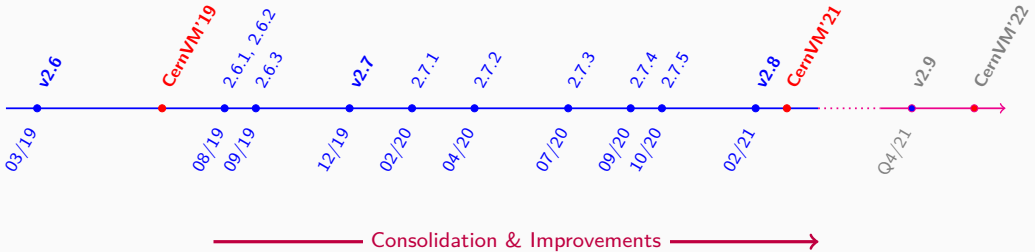
Infrastructure: WLCG



Infrastructure: WLCG



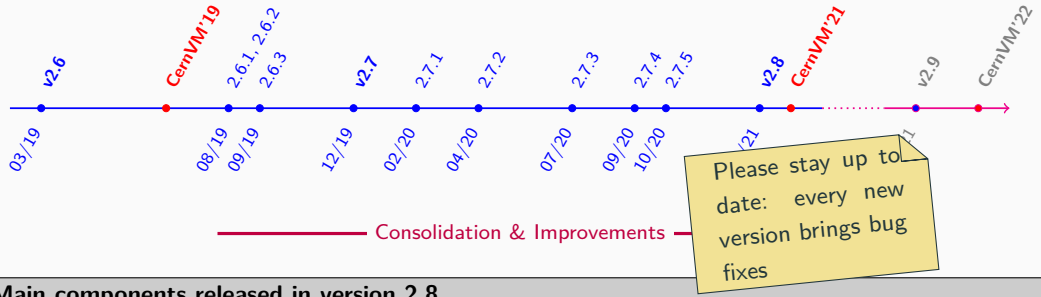
Software: Release Timeline



Main components released in version 2.8

1. Container integration: containerd, podman, k8s
2. Ephemeral writable shell: `cvmfs_server enter`
3. Template transactions: fast directory tree clones
4. Publish statistics dashboard

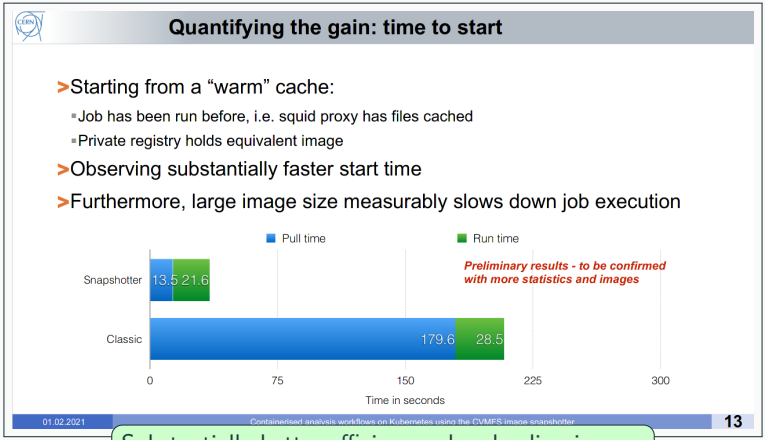
Software: Release Timeline



Main components released in version 2.8

1. Container integration: containerd, podman, k8s
2. Ephemeral writable shell: `cvmfs_server enter`
3. Template transactions: fast directory tree clones
4. Publish statistics dashboard

Highlights I

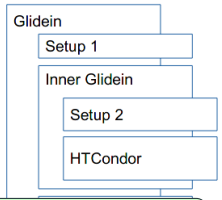


Substentially better efficiency when loading images from CernVM-FS in kubernetes ▸ Link

Fermilab

Glidein structure (planned)

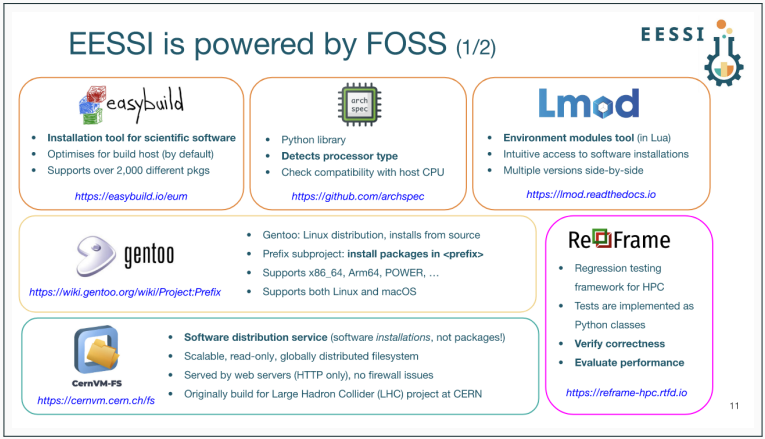
- Spawns multiple glideins if needed (MPI, multi-glidein, ...)
- Initial checks
- Download of scripts and setup
  - includes CVMFS mounting and Singularity testing
- Exec the inner part of the Glidein
- Complete setup
- Start HTCondor startd and join the pool
  - Start Singularity for each job individually (via wrapper)
- Cleanup
  - included modular scripts



GlideinWMS' use of unprivileged CernVM-FS to build a universal pilot ▸ Link 1 ▸ Link 2

14

Highlights III



EESSI – a turnkey software stack for HPC ▸ Link

CVMFS on Cori experiences

24 repositories - multiple diverse projects

Holds up fine at scale

But *many* issues:

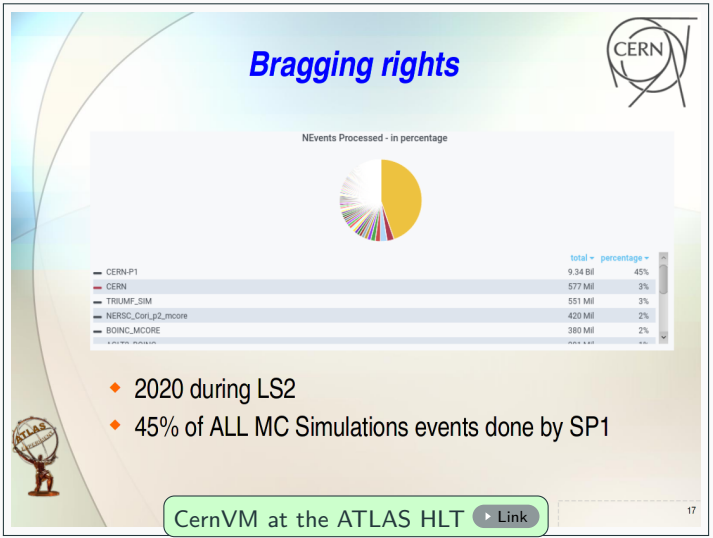
- E.g. returning the wrong file due to inode clash - occurred 3 separate times for 3 different reasons. (The last one was because ld.so uses its own caching of inodes to not load the same library (not the VFS layer))

CVMFS will be supported on Perlmutter:

- We are evaluating 'normal' FUSE-based solutions
- Don't have local disk so exploring either tiered-cache or loopback fs (similar to shifter per-node-cache but needs to persist across jobs)

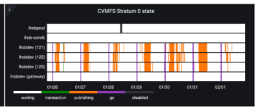
NERSC 11 BERKELEY LAB U.S. DEPARTMENT OF ENERGY Office of Science

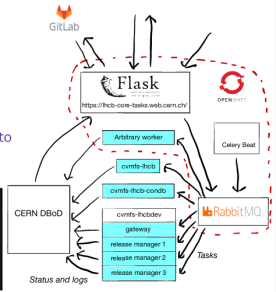
CernVM-FS support at NERSC improving with the next machine *Perlmutter* ▸ Link


Bragging rights
CERN
NEvents Processed - in percentage
total
percentage
CERN-P1 9.34 Bil 45%
CERN 577 Mil 3%
TRIUMF_SIM 551 Mil 3%
NERSC_Cori_p2_mcore 420 Mil 2%
BOINC_MCORE 380 Mil 2%
2020 during LS2
45% of ALL MC Simulations events done by SP1
ATLAS
CernVM at the ATLAS HLT
Link
17

Garbage collection

- Garbage collection must be run on the gateway
  - Added as a separate node with a dedicated message queue
  - Create a “lock” file in $HOME which disables all nodes
  - Run garbage collection once active leases are closed
  - Re-enable transactions once done
- Much faster than the old node
  - Biggest difference is likely the new multithreaded algorithm
  - Typically takes 1-2 hours, never runs into the morning like it used to
  - Area of concern if the repository continues growing

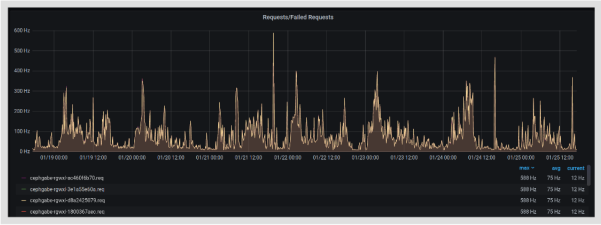




Use of parallel CernVM-FS publishers to sustain the nightly build load ▸ Link

S3 Storage for CVMFS

- CVMFS is the top S3 user for number of IOPS
  - Average number IOPS is moderate (~300 Hz)
  - Can be very spiky – Observed peaks over 4KHz



Infrastructure experience: from S3 and Stratum 1 operations at CERN to serving the long tail of science at NIKHEF Link 1 Link 2

Highlights VIII

Usage - Last 1 year

	Directory	Data Read	Access
1	/pnfs/fnal.gov/usr/nova	7.4PB	CVMFS
2	/user/ligo	7.4PB	CVMFS
3	/gwdata/O2	1.4PB	CVMFS
4	/gwdata/O1	529.1TB	CVMFS
5	/osgconnect/public/cgomes02	492.5TB	
6	/pnfs/fnal.gov/usr/uboone	459.5TB	CVMFS
7	/pnfs/fnal.gov/usr/minerva	386.5TB	CVMFS
8	/osgconnect/public/odgerk	369.3TB	
9	/chtc	362.5TB	CVMFS
10	/osgconnect/public/gziegler	283.5TB	

OSG storage repositories serve petabytes of data through StashCache and CernVM-FS ▸ Link

Summary and Future Directions

- **Workshop Mechanics**
  - Very fruitful workshop atmosphere that helped to expose key priorities for the coming months
  - Nevertheless, I missed the intense creativity of a presence event

- **Upcoming Focus of Work**
  - Regular maintenance: bug fixes, new platforms, etc.
  - Further scaling up the **CernVM-FS container hub** (/cvmfs/unpacked.cern.ch)
  - Improvements to the **CernVM-FS Gateway Services** to reach feature parity with the local publishers
  - Fully exploit **ephemeral publish container** for flexible publisher deployment
  - Client performance review and improvement through **pre-caching**

Many thanks to all the participants!