Distributing container images with CernVM-FS

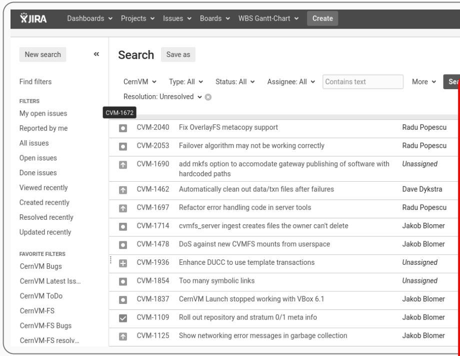
Jakob Blomer (CERN)

HSF Analysis Facilities Forum

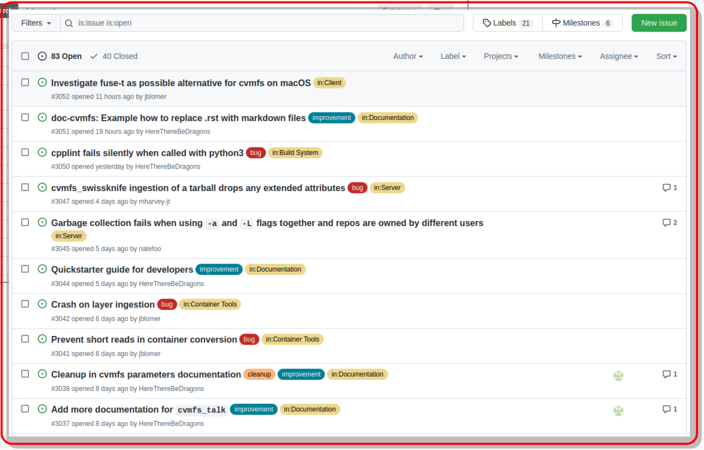
22 September 2022

Preface: Issue Tracking moved to GitHub

→ https://github.com/cvmfs/cvmfs/issues



CVM-2040	Fix OverlayFS metacopy support	Radu Popescu
CVM-2053	Failover algorithm may not be working correctly	Radu Popescu
CVM-1690	add mkfs option to accomodate gateway publishing of software with hardcoded paths	*Unassigned*
CVM-1462	Automatically clean out data/txn files after failures	Dave Dykstra
CVM-1697	Refactor error handling code in server tools	Radu Popescu
CVM-1714	cvmfs_server ingest creates files the owner can't delete	Jakob Blomer
CVM-1478	DoS against new CVMFS mounts from userspace	Jakob Blomer
CVM-1936	Enhance DUCC to use template transactions	*Unassigned*
CVM-1854	Too many symbolic links	*Unassigned*
CVM-1837	CernVM Launch stopped working with VBox 6.1	Jakob Blomer
CVM-1109	Roll out repository and stratum 0/1 meta info	Jakob Blomer
CVM-1125	Show networking error messages in garbage collection	Jakob Blomer



- Investigate fuse-t as possible alternative for cvmfs on macOS `in:Client`
  #3052 opened 11 hours ago by jblomer
- doc-cvmfs: Example how to replace .rst with markdown files `improvement` `in:Documentation`
  #3051 opened 19 hours ago by HereThereBeDragons
- cpplint fails silently when called with python3 `bug` `in:Build System`
  #3050 opened yesterday by HereThereBeDragons
- cvmfs_swissknife ingestion of a tarball drops any extended attributes `bug` `in:Server` 1
  #3047 opened 4 days ago by mharvey-jt
- Garbage collection fails when using `-a` and `-L` flags together and repos are owned by different users `in:Server` 2
  #3045 opened 5 days ago by natefoo
- Quickstarter guide for developers `improvement` `in:Documentation`
  #3044 opened 5 days ago by HereThereBeDragons
- Crash on layer ingestion `bug` `in:Container Tools`
  #3042 opened 6 days ago by jblomer
- Prevent short reads in container conversion `bug` `in:Container Tools`
  #3041 opened 6 days ago by jblomer
- Cleanup in cvmfs parameters documentation `cleanup` `improvement` `in:Documentation` 1
  #3038 opened 8 days ago by HereThereBeDragons
- Add more documentation for `cvmfs_talk` `improvement` `in:Documentation` 1
  #3037 opened 8 days ago by HereThereBeDragons

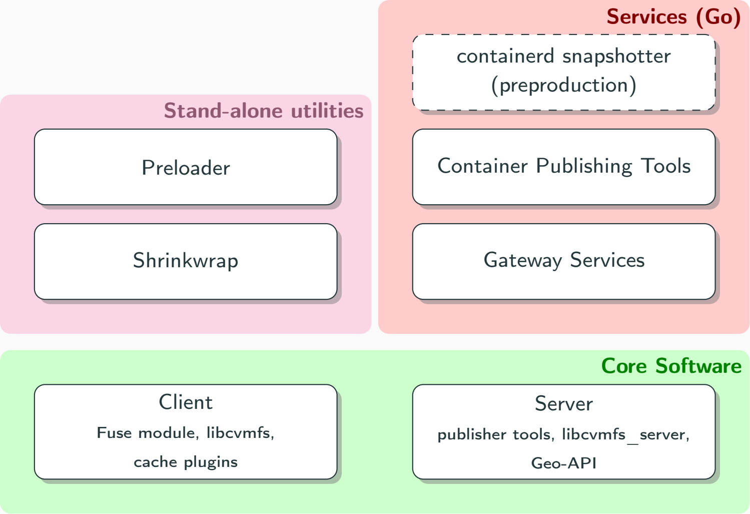
- Low barrier for submitting issues on GitHub
- Close integration of issues with pull requests
- JIRA tracker stays online for reference
- Updating existing tickets still possible

→ Mattermost channel for unpacked.cern.ch

CernVM-FS Components

Extras:

- cvmfsexec
- cvmfs-csi
- cvmfs-servermon
- github-action-cvmfs
- cvmfs-x509-helper
- repository monitor
- ...



Stand-alone utilities

Preloader

Shrinkwrap

Services (Go)

containerd snapshotter (preproduction)

Container Publishing Tools

Gateway Services

Core Software

Client

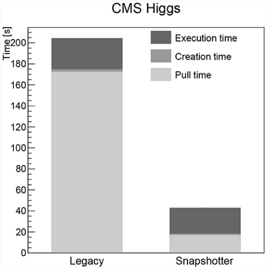
Fuse module, libcvmfs, cache plugins

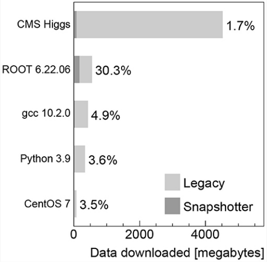
Server

publisher tools, libcvmfs_server, Geo-API

Motivation: unpacked image distribution through CernVM-FS

- CernVM-FS provides scalabale image storage and distribution through file-by-file approach
- Only tiny fraction of images used at runtime → fast container startup through on-demand loading
- Automatic cache management and sharing
- Well-established service in WLCG, piggy-backs on standard software distribution





Front. Big Data

CernVM-FS as a Container Hub

/cvmfs/unpacked.cern.ch

- > 2200 images
- > 10 TB
- > 250 M files

/cvmfs/singularity.opensciencegrid.org

- > 900 images
- > 3.5 TB
- > 75 M files

Images are readily available to run with apptainer (singularity), including **base operating systems**, **experiment software stacks**, **explorative tools (ML etc.)**, **user analyses**, and special-purpose containers such as **folding@home**

```
$ /cvmfs/oasis.opensciencegrid.org/mis/apptainer/current/bin/apptainer \
  exec '/cvmfs/unpacked.cern.ch/registry.hub.docker.com/library/debian:stable' \
  cat /etc/issue
Debian GNU/Linux 11 \n \l
```

CernVM-FS as a Container Hub

/cvmfs/unpacked.cern.ch

- > 2200 images
- > 10 TB
- > 250 M files

/cvmfs/singularity.opensciencegrid.org

- > 900 images
- > 3.5 TB
- > 75 M files

> 2× growth in the last 18 months

Images are readily available to run with apptainer (singularity), including **base operating systems**, **experiment software stacks**, **explorative tools (ML etc.)**, **user analyses**, and special-purpose containers such as **folding@home**

```
$ /cvmfs/oasis.opensciencegrid.org/mis/apptainer/current/bin/apptainer \
  exec '/cvmfs/unpacked.cern.ch/registry.hub.docker.com/library/debian:stable' \
  cat /etc/issue
Debian GNU/Linux 11 \n \l
```

Container Runtime Integration

Runtime	CernVM-FS Support
Apptainer	native
podman	native / pre-production (use image storage from /cvmfs)
containerd / k8s	plugin / pre-production (through cvmfs snapshotter) → CernVM'22 Workshop
docker	*"graph driver"* image storage plugin – deprecated[1] through containerd in the future

Documentation chapter on containers & CernVM-FS:

→ `https://cvmfs.readthedocs.io/en/latest/cpt-containers.html`

[1] Soon replaced by containerd ▸ Docker's announcement

unpacked.cern.ch

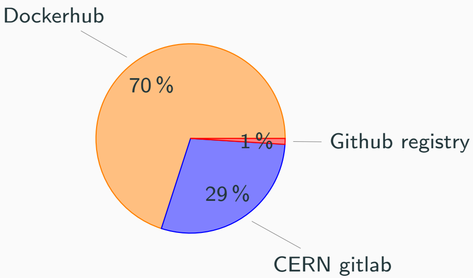
- Image wishlists on ▸ CERN GitLab and ▸ GitHub
- Editable by merge/pull request

```yaml
version: 1
user: cvmfsunpacker
cvmfs_repo: 'unpacked.cern.ch'
output_format: >
  https://gitlab-registry.cern.ch/unpacked/sync/$(image)
input:
   - 'https://gitlab-registry.cern.ch/sft/docker/ubuntu20:latest'
   - 'https://registry.hub.docker.com/library/centos:*'
   ...
```

Semi-automatic procedure:

- Added images automatically kept in sync
- Globbing support for tags
- Sync delay ∼20 minutes

Origin of images on unpacked.cern.ch



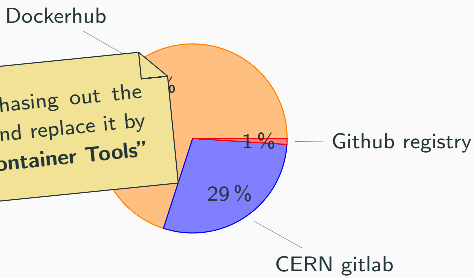
new Images from Docker Hub and GitHub are proxied through registry.cern.ch → CernVM'22 workshop

unpacked.cern.ch

- Image wishlists on ▸ CERN GitLab and ▸ GitHub
- Editable by merge/pull request

```
version: 1
user: cvmfsunpacker
cvmfs_repo: 'unpacked.cern.ch'
output_format: >
  https://gitlab-registry.cern.ch/unpacked/sync/$(im
input:
   - 'https://gitlab-registry.cern.ch/sft/docker/ubu
   - 'https://registry.hub.docker.com/library/centos:
   ...
```

Origin of images on unpacked.cern.ch



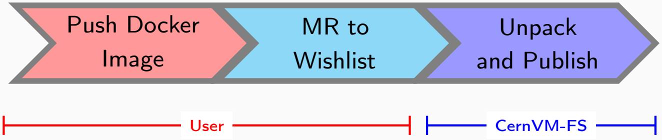
Note: we are phasing out the name "DUCC" and replace it by **"CernVM-FS Container Tools"**

Semi-automatic procedure:

- Added images automatically kept in sync
- Globbing support for tags
- Sync delay ~20 minutes

new Images from Docker Hub and GitHub are proxied through registry.cern.ch → CernVM'22 workshop

Container Conversion Service



Wishlist https://gitlab.cern.ch/unpacked/sync

```
version: 1
user: cvmfsunpacker
cvmfs_repo: 'unpacked.cern.ch'
output_format: >
  https://gitlab-registry.cern.ch/unpacked/sync/$(image)
input:
   - 'https://registry.hub.docker.com/library/fedora:latest'
   - 'https://registry.hub.docker.com/library/debian:stable'
   - 'https://registry.hub.docker.com/library/centos:*'
```

Multiple wishlists possible, e.g. experiment specific

/cvmfs/unpacked.cern.ch

```
# Singularity/Apptainer
/registry.hub.docker.com/fedora:latest -> \
  /cvmfs/unpacked.cern.ch/.flat/d0/d0932...
# containerd, k8s, podman
/.layers/f0/1af7...
# Support for incremental publishing
/.chains/e7/6af9...
```

Important Client Bits

Notable client features and fixes in support of containers:

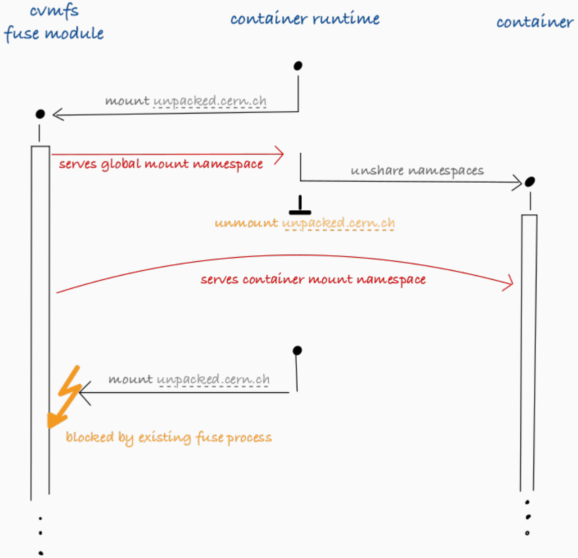
- CernVM-FS synthetic xattrs are hidden by default to reduce cost of overlayfs copy-up (available since 2.9.1, default as of 2.10)
- Client available as container (helps to create, for instance, k8s daemon set)
- Fixed "zombie mountpoint" issue (see next slide)

Recommended to update to EL9 and CernVM-FS 2.10 (released soon)

Fixed Zombie Mountpoints

Fixed in version 2.9 + Kernel 5.15 (EL 9.1)

- Depending on the container engine (use of `unshare`), mounting a repository could hang
- Fixed by allowing new mounts to attach to existing fuse module
- Usually not a problem with singularity/apptainer



Summary and Next Steps

1. CernVM-FS distributes HEP containers efficiently at scale
2. Two main publisher workflows
   - guarded by software & dataset librarians
   - container ingestion open to a broader community

Next steps

- Production release of the cvmfs-snapshotter planned for Q4/2022
- Production podman support of container conversion planned for Q1/2023
- Support for multi-arch images in container conversion planned for Q1/2023
- Release of webhook-based conversion if required
- Mid-term goals:
  - Encourage generic containerd snapshotter that supports (some) layers unpacked on a file system
  - Add missing functionality to the gateway to use together with container conversion
  - Agree on procedure for image lifecycle / retention

Backup Slides

Simple Case: CernVM-FS Available on the Host

```
$ docker run -v /cvmfs:/cvmfs:shared busybox ls /cvmfs/sft.cern.ch
README.md lcg
```

```
$ singularity exec -B /cvmfs docker://busybox ls /cvmfs/sft.cern.ch
README.md lcg
```

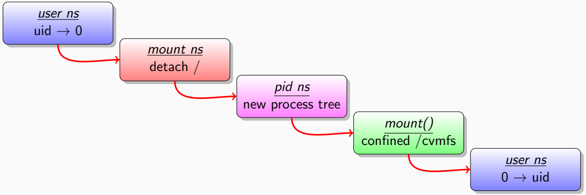
Important: use *shared* bind mount with docker so that that repositories can be mounted on demand from inside the container

Unprivileged Mounting with `cvmfsexec`

```
$ cvmfsexec grid.cern.ch atlas.cern.ch -- ls /cvmfs
atlas.cern.ch cvmfs-config.cern.ch grid.cern.ch
```

Technical foundations

- User namespaces completing container support
- As of Linux kernel version 4.18 (EL8, but also EL 7.8),
  fuse mounts are unprivileged in user name spaces
- Overlay-FS implementation available as a fuse module



user ns
uid → 0

mount ns
detach /

pid ns
new process tree

mount()
confined /cvmfs

user ns
0 → uid

For HPCs: Pre-mounted by Singularity

- With the new Fuse3 libraries, mounting can be handed off to a trusted, external helper.
- Fuse3 libraries have been backported to EL6 and EL7 platforms.
- Gives access to /cvmfs in containers started by singularity (`singularity --fusemount`)
- **Required cvmfs client to be installed and prepared in the container**

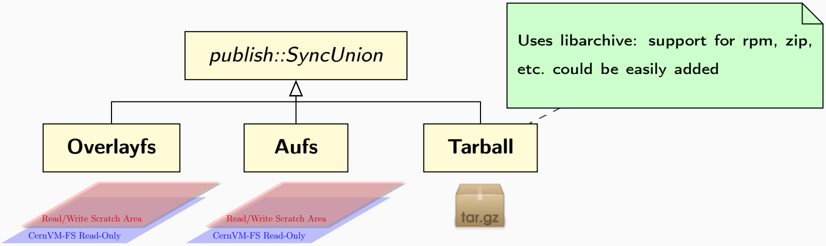
```
$ CONFIGREPO=config-osg.opensciencegrid.org
$ mkdir -p $HOME/cvmfs_cache
$ singularity exec -S /var/run/cvmfs -B $HOME/cvmfs_cache:/var/lib/cvmfs \
  --fusemount "container:cvmfs2 $CONFIGREPO /cvmfs/$CONFIGREPO" \
  --fusemount "container:cvmfs2 sft.cern.ch /cvmfs/sft.cern.ch" \
  docker://davedykstra/cvmfs-fuse3 ls /cvmfs/sft.cern.ch
README.md lcg
```

Enabling Feature for Container Publishing: Tarball Ingestion

Direct path for the common pattern of publishing tarball contents

```
$ cvmfs_server transaction
$ tar -xf ubuntu.tar.gz
$ cvmfs_server publish
```

```
$ cat ubuntu.tar.gz | \
    cvmfs_server ingest -t -
```



Uses libarchive: support for rpm, zip, etc. could be easily added

Performance Example

Ubuntu 18.04 container – 4 GB in 250 k files: **56 s untar + 1 min publish vs. 74s ingest**