CernVM-FS Turns 15

Jakob Blomer (CERN)

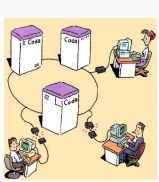
CernVM Workshop 2024

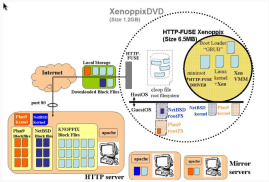
CERN, 16 September 2024

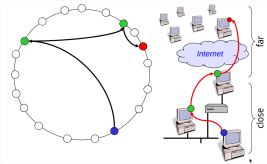
How CernVM-FS Came to Life

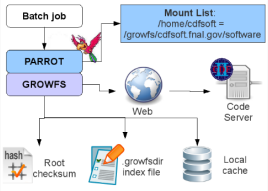
How CernVM-FS Came to Life

- Part of the CernVM R&D project on virtualization
  - Predrag's 2018 Talk: CernVM 10 years after
- Decouple the experiment software from the virtual machine image using a **global network file system**
- Looked into several existing options
  - Coda: AFS with offline mode
  - HTTP-Fuse: on-demand bootable Linux image
  - Igor-FS: file system with P2P transport
  - GROWFS: CernVM-FS pre-curser using the Parrot system call interception toolkit





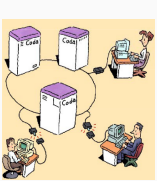


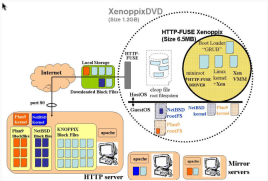


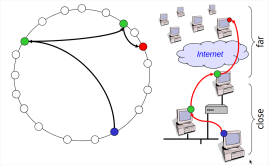
How CernVM-FS Came to Life

- Part of the CernVM R&D project on virtualization
  - Predrag's 2018 Talk: CernVM 10 years after
- Decouple the experiment software from the virtual machine image using a **global network file system**
- Looked into several existing options
  - Coda: AFS with offline mode
  - HTTP-Fuse: on-demand bootable Linux image
  - Igor-FS: file system w
  - GROWFS: CernVM-F the Parrot system call toolkit

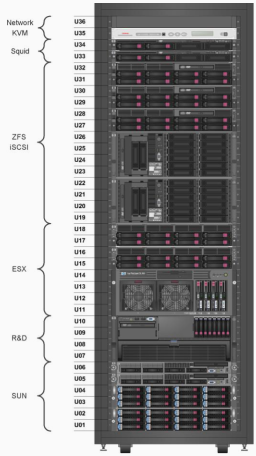
CVMFS v1 was a GROW-FS Fuse frontend written by Leandro Franco





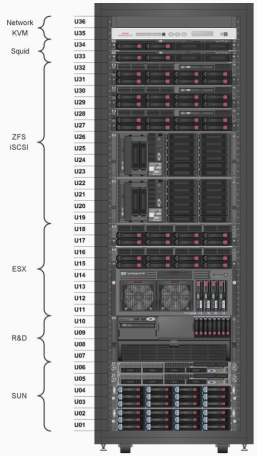


Proof of Concept: 2009-2012



- Initial technology choices leading to CVMFS 2.0:
  Fuse, C++, HTTP CDN, SQlite file catalogs, content-addressed storage
- CernVM infrastructure (including CVMFS storage, release managers etc.) operated from building 157
  - Fully virtualized with VMware ESX
  - Storage using Solaris/ZFS: our initial solution for snapshotting & replication
- Presented at CHEP 2010 in Taipei
  (15 million files under management)
- Growing interest in using CernVM-FS on the Grid outside the VM
  (virtualization came back later a few years later with OpenStack, Docker, k8s)
  - to address shortcomings of AFS, NFS, Grid installation jobs

Proof of Concept: 2009-2012



- Initial technology choi… Fuse, C++, HTTP… storage … dressed

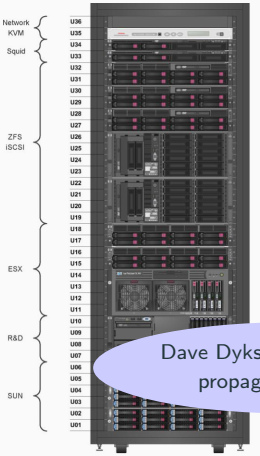
Meanwhile:

- Amazon EC2, S3, CloudFront
- iPhone, Android
- Higgs Discovery

- CernVM infrastruct… managers etc.) oper… [197]
  - Fully virtualized with VMware ESX
  - Storage using Solaris/ZFS: our initial solution for snapshotting & replication
- Presented at CHEP 2010 in Taipei
  (15 million files under management)
- Growing interest in using CernVM-FS on the Grid outside the VM

  (virtualization came back later a few years later with OpenStack, Docker, k8s)
  - to address shortcomings of AFS, NFS, Grid installation jobs

Proof of Concept: 2009-2012



- Initial technology choi
  Fuse, C++, HTTP
  storage
- CernVM infrastruct
  managers etc.) oper
  - Fully virtualized with VMware ESX
  - Storage using Solaris/ZFS: our initial solution for snapshotting & replication
- Presented at CHEP 2010 in Taipei
  (15 million files under
  interest
  - to address shortcomings of AFS, NFS, Grid installation jobs

Meanwhile:
- Amazon EC2, S3, CloudFront
- iPhone, Android
- Higgs Discovery

Dave Dykstra: "how will you propagate updates?"

Steve Traylen: "how about autofs for shared nodes?"

outside the VM

(virtualization came back later a few s later with OpenStack, Docker, k8s)

Adoption 2012-2014 (v2.0 $\longrightarrow$ v2.1)

With the adoption of experiments and a global infrastructure, we faced several challenges:

1. Robust and scalable writing / publishing
2. We split development and operations (CERN IT, RAL, BNL, FNAL) and needed to find "standard" replacements for the ZFS storage $\rightarrow$ "Stratum 0" to "Stratum 1" replication
3. File system clients would suddenly run on Grid worker nodes instead VMs
   - Client had to become more robust and gentle on resources; internally, we needed to switch from libfuse's path based interface to the low-level, inode based interface
   - No room for downtime for client updates $\rightarrow$ developed client hotpatch functionality
   - Eventually, we needed to split the Cern(VM) specific configuration from the core software, leading to the "cvmfs-config-xyz" packages and the config repository

Presented the roadmap to CVMFS 2.1 at CHEP 2012 in New York
(100 million files under management)

Adoption 2012-2014 (v2.0 ⟶ v2.1)

With the adoption of experiments and a global infrastructure, we faced several challenges:

1. Robus…blishing
2. We spl…ons (CERN IT, RAL, BNL, FNAL) and needed to find "standard" replace… → "Stratum 0" to "Stratum 1" replication
3. File system clients would suddenly run on Grid worker nodes instead VMs
   - Client had to become more robust and gentle on resources; internally, we needed to switch from libfuse's path based interface to the low-level, inode based interface
   - No room for downtime for client updates → developed client hotpatch functionality
   - Eventually, we needed to split the Cern(VM) specific configuration from the core software, leading to the "cvmfs-config-xyz" packages and the config repository

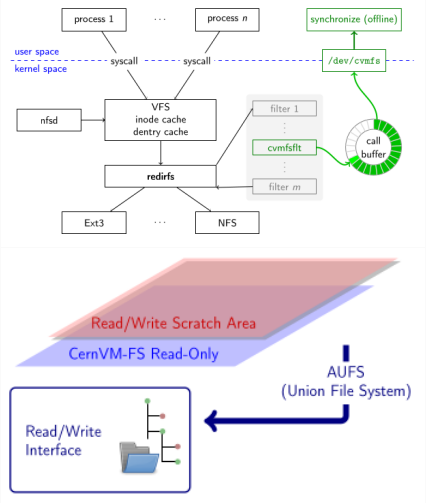
Among the early, crucial supporters were Doug Benjamin and Ian Collier

Presented the roadmap to CVMFS 2.1 at CHEP 2012 in New York
(100 million files under management)

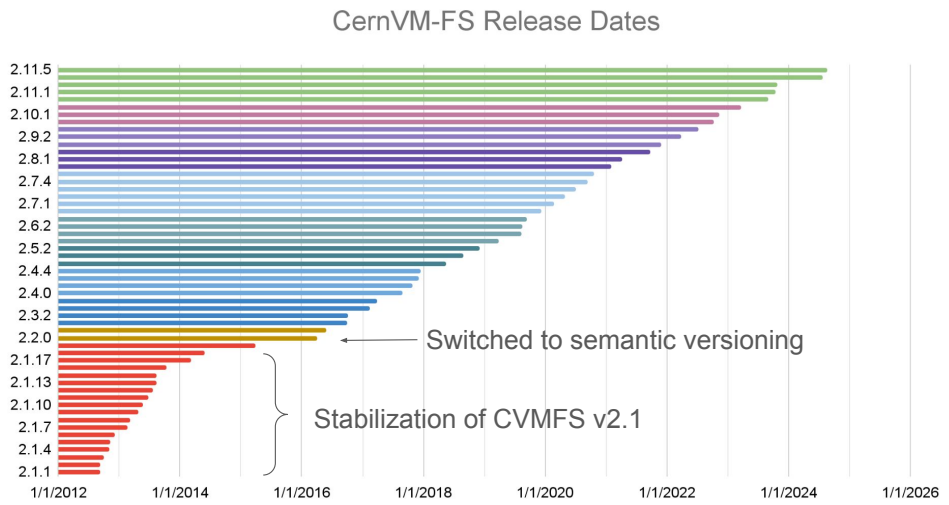
How to solve writing

Approaches in chronological order

1. Traverse entire repository: does not scale
2. Fuse module for writing: slow and intricate
3. Change log: worked, brittle (until ∼2011; propagation delay 0.5 days)
4. Union file system: AUFS for a while the only working kernel-level union file system, but not in the mainline kernel
   - Provided AUFS patched RHEL6 kernel
   - Moved to overlayfs in RHEL7 (initial contribution from Wellcome Sanger)
   - Propagation delay at <15 minutes
5. Special cases: grafting (since 2016), direct ingestion (since 2019)



Production Readiness Review

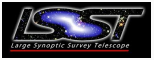



Off to new shores: 2015-2020

Off to new shores: 2015-2020

The production system stabilized, the environment changed

- Open source S3 compatible object stores became available (Ceph S3)
  fantastic alternative for CVMFS backend storage; took ∼5 years from prototype to full production
- 1 TB per night of integration builds to be distributed to the grid: repository garbage collection
- Distributed publishing: long envisaged (since 2012) and deployed around 6 years later
- Plus: containers, HPC, data distribution
- New developments were summarized at CHEP 2018 in Sofia



Off to new shores: 2015-2020

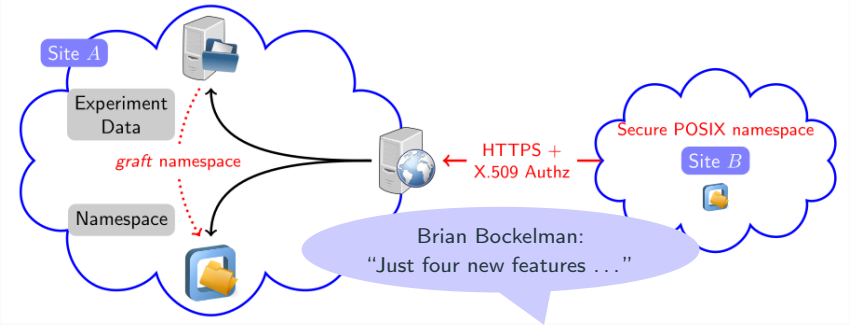
The production system stabilized, the environment changed

- Open source S3 compatible object stores became available (Cep… fantastic alternative for CVMFS backend storage; took ∼5 years … duction
- 1 TB per night of integration builds to be distributed to the grid … tion
- Distributed publishing: long envisaged (since 2012) and deployed around 6 years later
- Plus: containers, HPC, data distribution
- New developments were summarized at CHEP 2018 in Sofia

Close collaboration with Ben Couturier and Enrico Bocchi

POSIX namespace for data

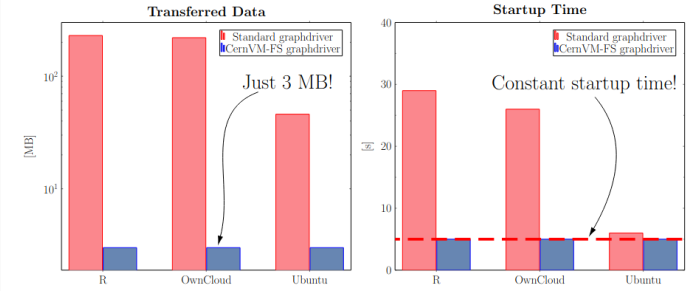
1. Namespace grafting
2. External data (split squid and XRootD HTTP traffic)
3. Uncompressed files
4. Authorization plugins



Virtualization strikes back: Containers

- With the advent of container virtualization (docker etc.), the platform isolation problem was finally solved without performance overhead.
- However, unlike full virtualization, we could (not yet) mount our file system in a container.
- The software distribution problem very much stayed the same; the docker "layers" offered modularization for image construction, but little advantage for distribution.

1. Tricked Docker into reading from CVMFS
2. Singularity/apptainer allowed for direct use of unpacked images
3. containerd snapshotter interface for file-based transfer



CVMFS on HPC

So close, yet so far:
no fuse, *no* internet connectivity, *no* local disk (cache), *no* site caches

Required development of a set of deployment options à la carte:

- preloader
- shrinkwrap
- parrot connector (now: cvmfsexec)
- cache plugins

CVMFS on HPC

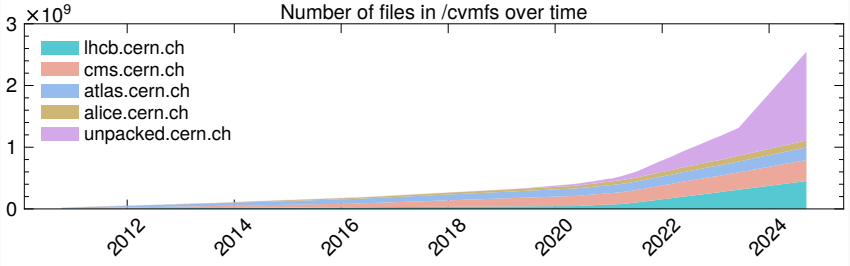
So close, yet so far:
no fuse, *no* internet connectivity, *no* local disk (cache), *no* site caches

Required development of a set of deployment options à la carte:

- preloader
- shrinkwrap
- parrot connector (now: c
- cache plugins

Close collaboration with Douglas Thain's team at Notre Dame

To the Limits and Beyond: 2020-2024



- Accelerating repository size: more platforms, more CI builds, containers, Python, ...
- New, exciting use cases: $\mathcal{O}(100\,\mathrm{PB})$ data archive distributed at **jumptrading**
- **EESSI**: a new approach to compute environment on European HPCs
- Continuous improvement in Linux and fuse; most importantly: unprivileged fuse mounts (cvmfsexec)

→ **Subject of this workshop**


Internal Affairs

Tech Stack

2008 · · · 

2011 · · · GitHub

2012 · · · JIRA Software
ECOMMANDER

2013 · · · googletest

2014 · · · coverity™

2015 · · · Jenkins docker

2016 · · · ERLANG

2017 · · · Read the Docs

2019 · · · Golang

2021 · · · Discourse 

- git replaced svn
- JIRA replaced Savannah
- Perl was phased out in 2018
- Jenkins replaced Electric Commander
- golang replaced erlang
- Github issues replaced JIRA in 2020
- clang-tidy replaced Coverity in 2021

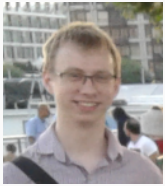
Former Influential CernVM-FS Contributors

Predrag Buncic
CernVM Founder

Artem Harutyunyan
CernVM Original Team
→ Qualys, Mesosphere, Bardeen

Carlos Aguado Sanchez
CernVM Original Team
→ AWS

René Meusel
Union-FS based server
→ Rohde & Schwarz



Nikola Hardi
Docker graph driver
→ Swisscom

Simone Mosciatti
unpacked.cern.ch
→ AWS

Radu Popescu
Gateway
→ Logitech

Jan Priessnitz
Parallel GC
→ MPI

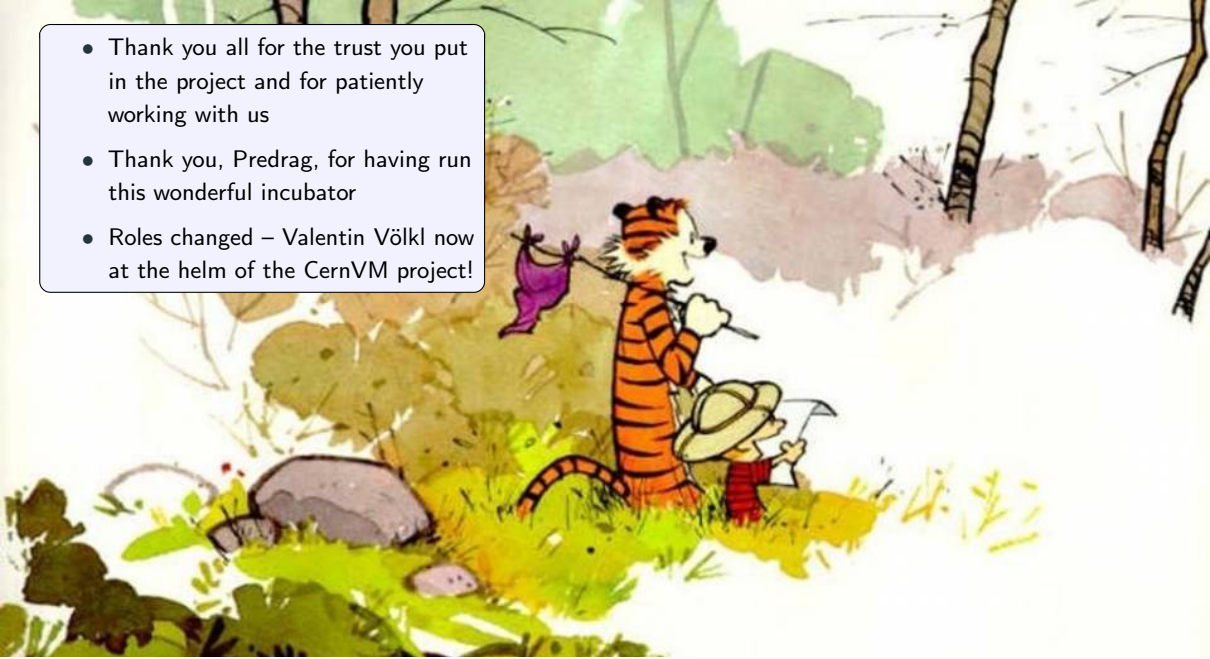
Andrea Valenzuela
Containerized publisher
→ CMS

On Tour 2016 @ RAL

Thanks to Catalin Condurache

On Tour 2022 @ NIKHEF

Thanks to Dennis van Dok



- Thank you all for the trust you put in the project and for patiently working with us
- Thank you, Predrag, for having run this wonderful incubator
- Roles changed – Valentin Völkl now at the helm of the CernVM project!