CERN

Distribution of Container Images

From tiny deployments to massive analysis on the grid

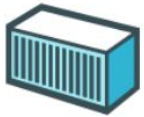
Enrico Bocchi
CERN, IT-Storage

HEPiX Online, March 2021

“Build, Ship, Run, Any App Anywhere”

Build

Develop an app using Docker containers with any language and any toolchain.



Ship

Ship the “Dockerized” app and dependencies anywhere - to QA, teammates, or the cloud - without breaking anything.

Run

Scale to 1000s of nodes, move between data centers and clouds, update with zero downtime and more.

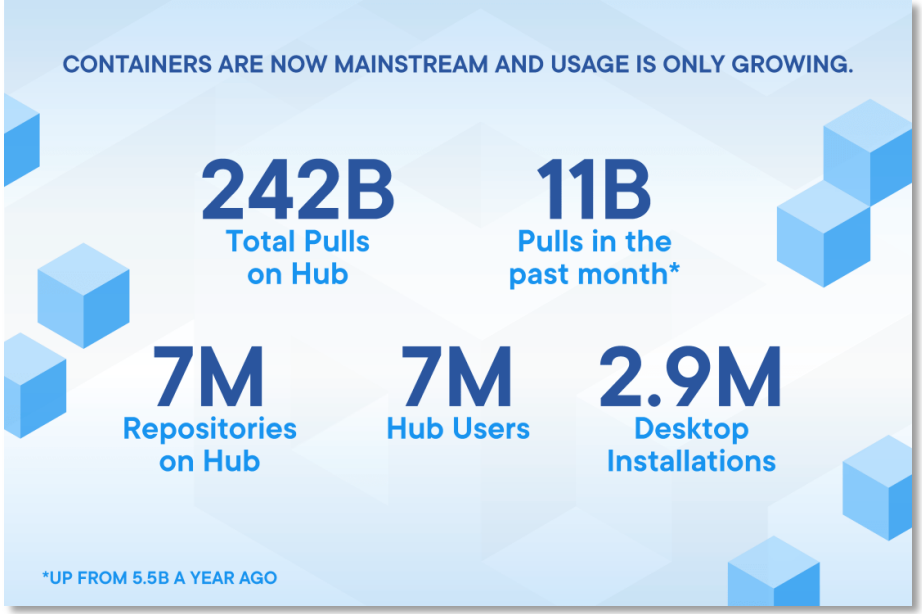
“Build, Ship, Run, Any App Anywhere”

- **Container Registry:** Specialized repository to store container images
  - Distribution of images by uploading (`docker push`) and downloading (`docker pull`)
  - **Public** – DockerHub
    Private – Amazon ECR, {MS Azure, Google, IBM, …} Container registry
    Self-hosted – Red Hat Quay, Docker `registry` container

The Docker Hub Registry

- Most popular public registry – Docker's default

Docker Index
30 July 2020



Docker Hub – The Free Lunch Is Over

docker blog

Why Docker? Products Developers Pricing Company

Scaling Docker's Business to Serve Millions More Developers: Storage

JEAN-LAURENT DE MORLHON

Aug 24 2020

- 150 M images
- 15 PB storage

- 4.5 PB idle images from free accounts

Starting on November 1, images […] not pushed or pulled in the last 6 months will be removed.

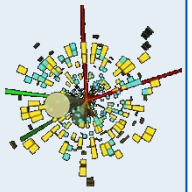
Containers at CERN

Images for Service Deployment

- Small images (< 1 GB)
- Run on few nodes
- Re-use from upstream

Images for Scientific Analysis

- Immutable unit for reproducibility
- Run on the Worldwide LHC Grid (potentially thousands of nodes)
- Very large in size (10+ GB)



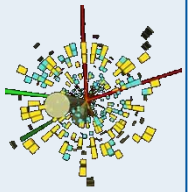
Containers at CERN

Images for Service Deployment

- Small images (< 1 GB)
- Run on few nodes
- Re-use from upstream

Images for Scientific Analysis

- Immutable unit for reproducibility
- Run on the Worldwide LHC Grid (potentially thousands of nodes)
- Very large in size (10+ GB)



- Currently using **GitLab Container Registry**
  - ✅ Tight integration with CI pipelines, Registry associated to GitLab project
  - ❌ No Garbage Collection of unreferenced blobs, No support for OCI artifacts

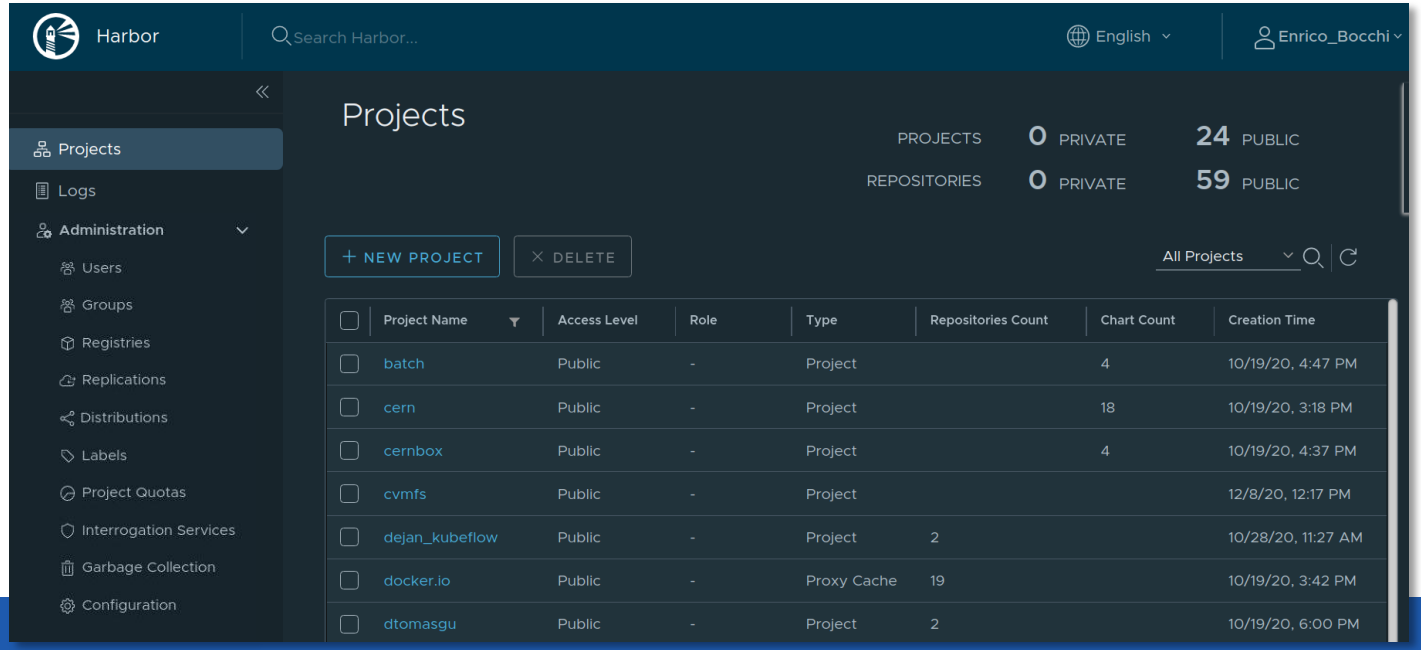
GitLab

HARBOR

CernVM File system

Harbor

- Active upstream project, CNCF Graduated
- Storage of images and OCI artifacts (e.g., Helm charts)



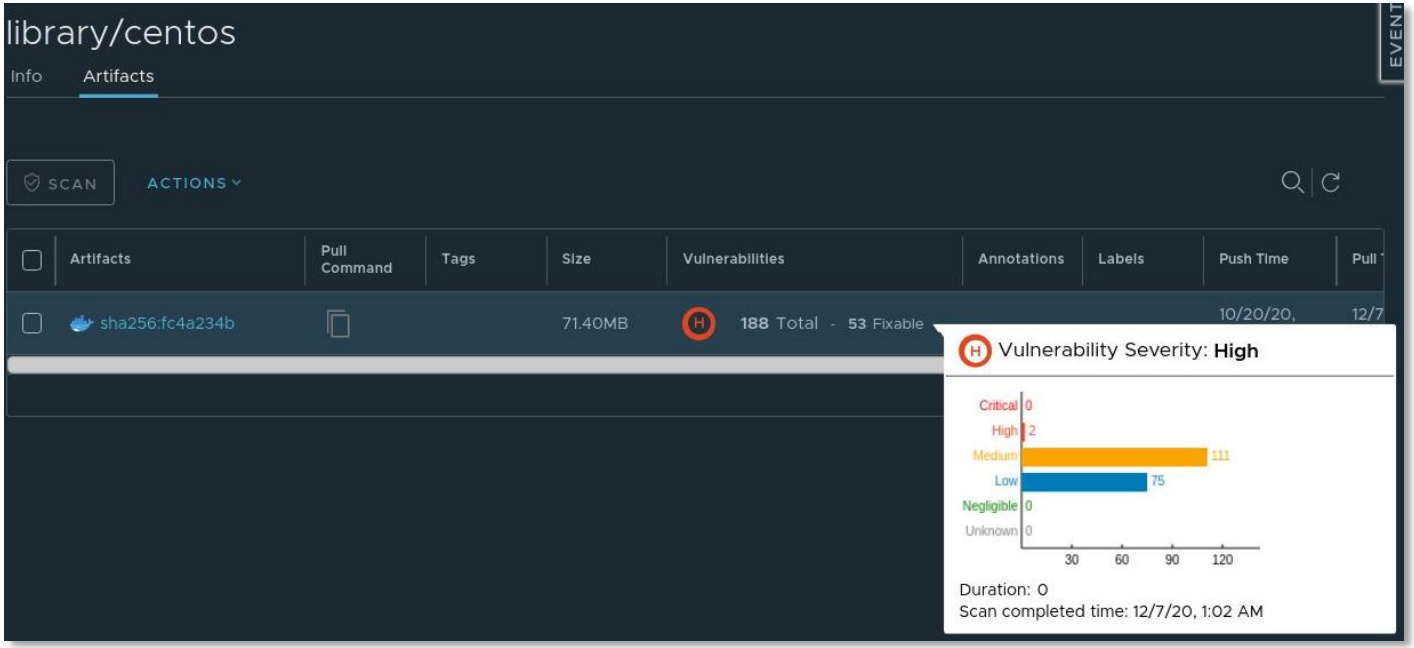
Harbor

- Available at CERN: https://registry.cern.ch
  - Deployed ~2 years ago as registry for Helm charts
  - Opened as general registry (images + OCI artifacts) in Q4 2020
  - Backed by S3 storage (others possible)

- Off-the-shelf functionalities matching our requirements
  - **Centralized User Management:** Quotas, authorization, authentication via OIDC, …
  - **Artifact Signing:** Ensure trusted source for artifacts being installed
  - **Garbage Collection**: Online deletion of unreferenced blobs on S3 storage
  - …

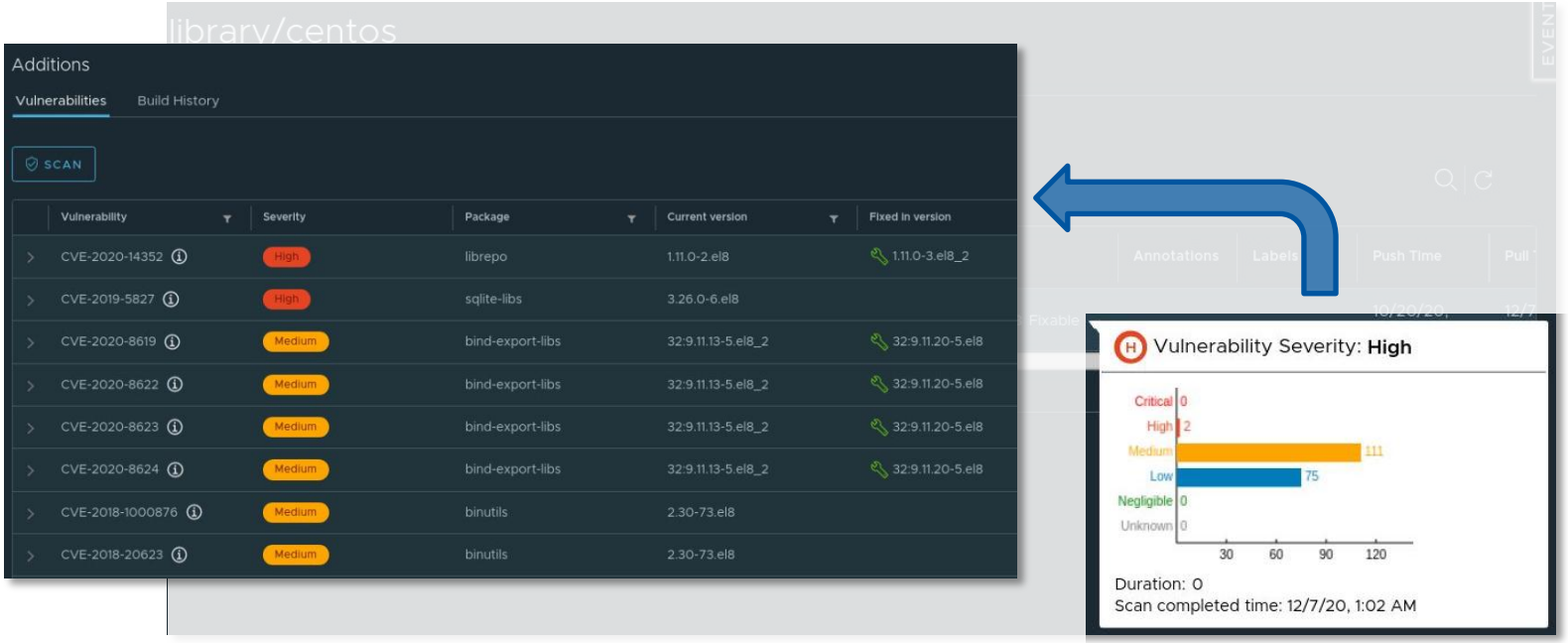
Harbor

- **Vulnerability Scanning:** Based on external plugins (Clair, Trivy, Sysdig)



Harbor

- **Vulnerability Scanning:** Based on external plugins (Clair, Trivy, Sysdig)

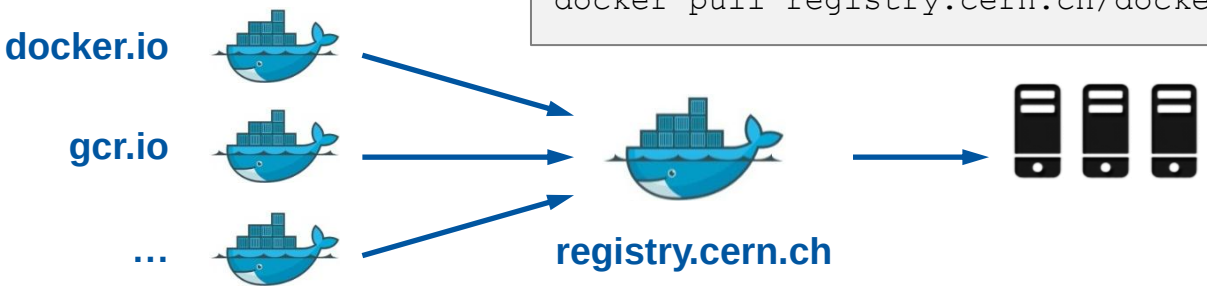


Harbor

- **Proxy Caching:** Pull-through cache for other registries
  - Enabled by administrators for specific registries (e.g., docker.io)
  - Vulnerability checks can be applied on top

- **Registry Replication:** Push/pull images to/from other registries
  - Based on regular expression matching on image tag

```
docker pull registry.cern.ch/docker.io/myrepo/myimage:mytag
```



Harbor

- **Proxy Caching:** Pull-through cache for other registries
  - Enabled by administrators for specific registries (e.g., docker.io)
  - Vulnerability checks can be applied on top
- **Registry Replication:** Push/pull images to/from other registries
  - Based on regular expression matching on image tag

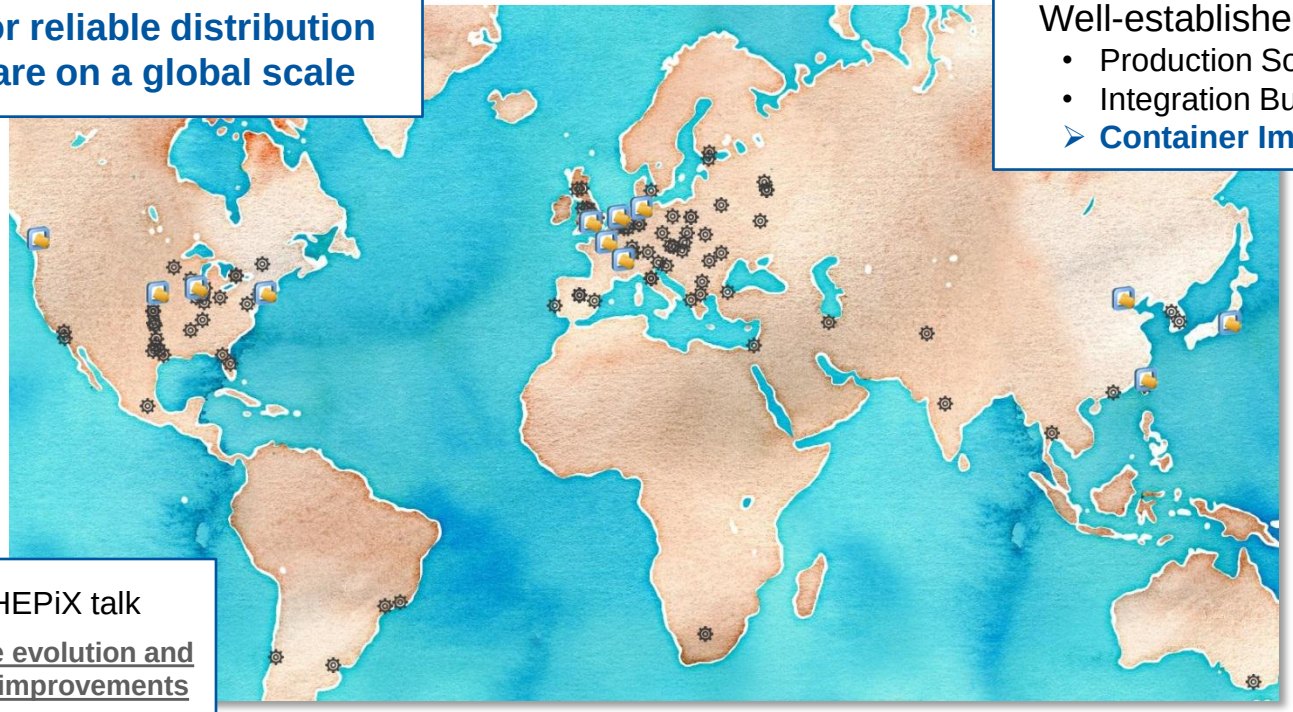
Harbor provides **traditional storage** for images
Advanced capabilities for **security**, management, replication

- How to distribute multi-GB images to thousands of nodes?

CVMFS

CernVM File system

Service for reliable distribution of software on a global scale



Well-established CDN for

- Production Software
- Integration Builds
- ➢ **Container Images**

Previous HEPiX talk

CVMFS service evolution and infrastructure improvements

CVMFS for Container Images

CernVM File system

Server: Ingestion of existing images

- Extraction of layers into flat root filesystem
- Efficient file-based deduplication
- Publication into CVMFS repository

Client: Efficient pulling and caching

- No need to store the entire image locally
- On-demand fetching of required files
- Self-managed local cache

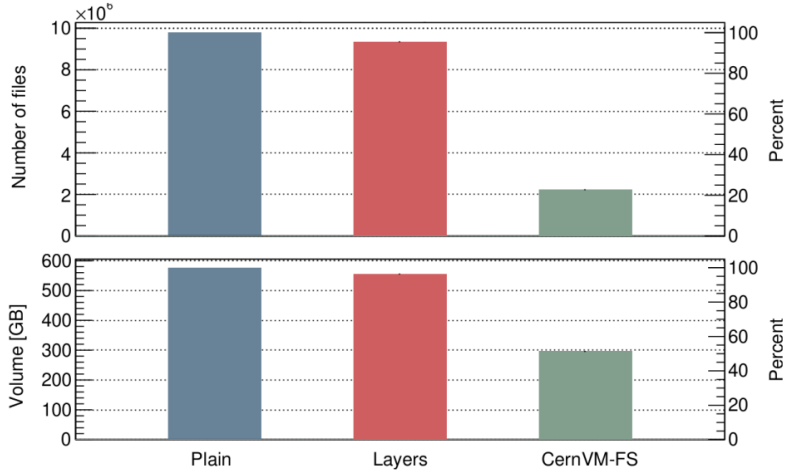
Leverage on existing Content Distribution Network

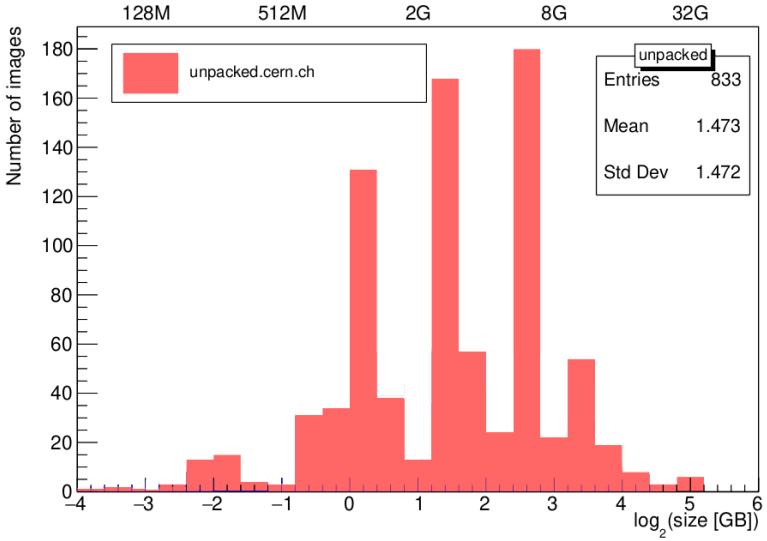
- Specialized *DUCC* daemon
  - Regulates and triggers ingestion
- Ingestion based on
  - Whish-list
  - Integration with traditional registries
- Integration with Container runtimes
  - Containerd (Docker, k8s)
  - Singularity
  - Podman

CVMFS for Container Images

CernVM File system

- **unpacked.cern.ch** – First CVMFS-powered container hub
  - 750+ container images, 3.5 TB, 50 M files
  - Distribution of container image size
  - Efficiency of file-based deduplication





CVMFS for Container Images

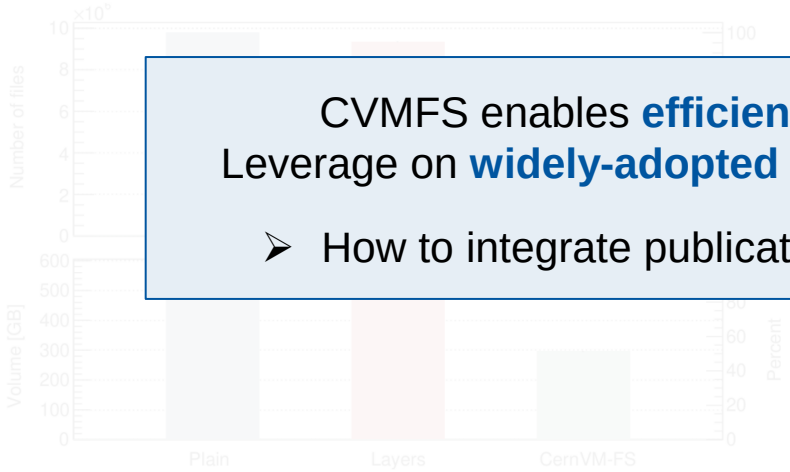
CernVM File system

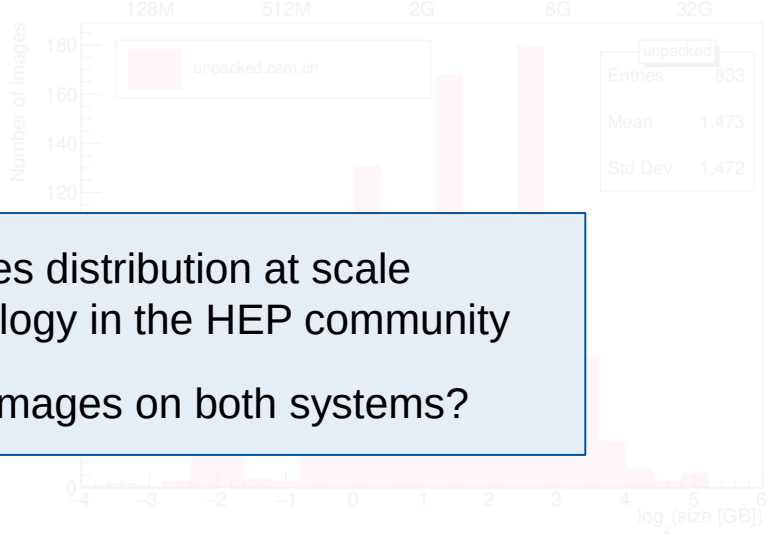
- Example of Large-Scale Deployment: **Folding@Home**
  - Runs on the grid off containers served from `/cvmfs`

CVMFS for Container Images

CernVM File system

- **unpacked.cern.ch** – First CVMFS-powered container hub
  - 700+ container images from major experiments
  - Distribution of container image size
  - Efficiency of file-based deduplication





CVMFS enables **efficient** images distribution at scale
Leverage on **widely-adopted** technology in the HEP community

- How to integrate publication of images on both systems?

Streamlined Management and Publication of Images

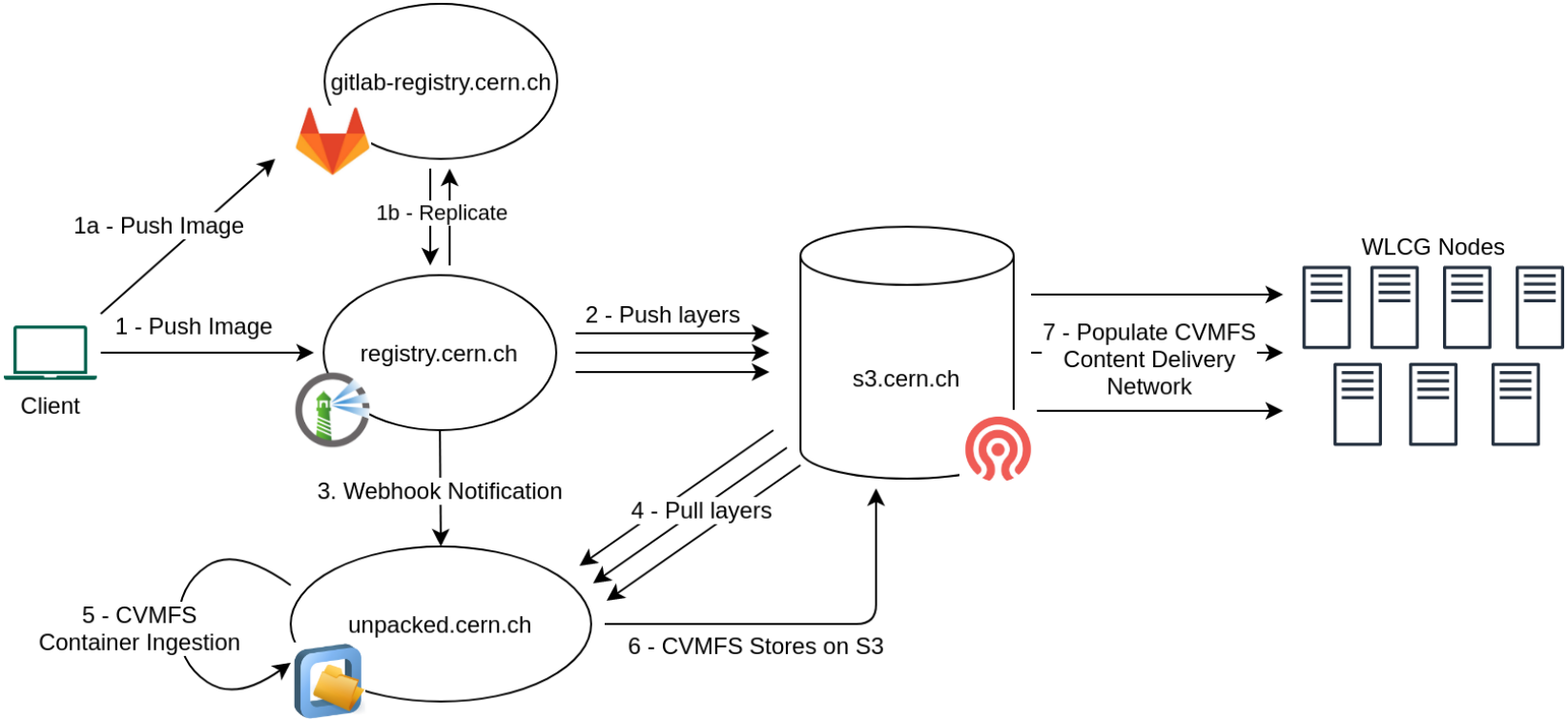
docker push

- Centralized management of images
- Vulnerability scans + OCI artifacts signing
- Supports service deployment use-case

Notification via webhooks

- Improved storage and distribution efficiency
- Enables large-scale analysis on the grid
- Supports scientific analysis use-case

Streamlined Management and Publication of Images



Conclusions

- Containers are mainstream and pervasive
  - For general service deployment in IT
  - For scientific and HEP use-cases (large-scale analysis, reproducibility)
- Storage and distribution of containers can be challenging
  - HEP images can grow big and should run on thousands of nodes
  - Traditional push – pull model and layer-based deduplication become inefficient
- Combine existing technologies to best support end-users
  - No one-fits-all solution exist at the moment
  - Harbor + CVMFS provide advanced features and efficient distribution
  - Integration prototyped and running, now validating scalability

Thank you!

Questions? || Comments?

Enrico Bocchi

enrico.bocchi@cern.ch

CERN

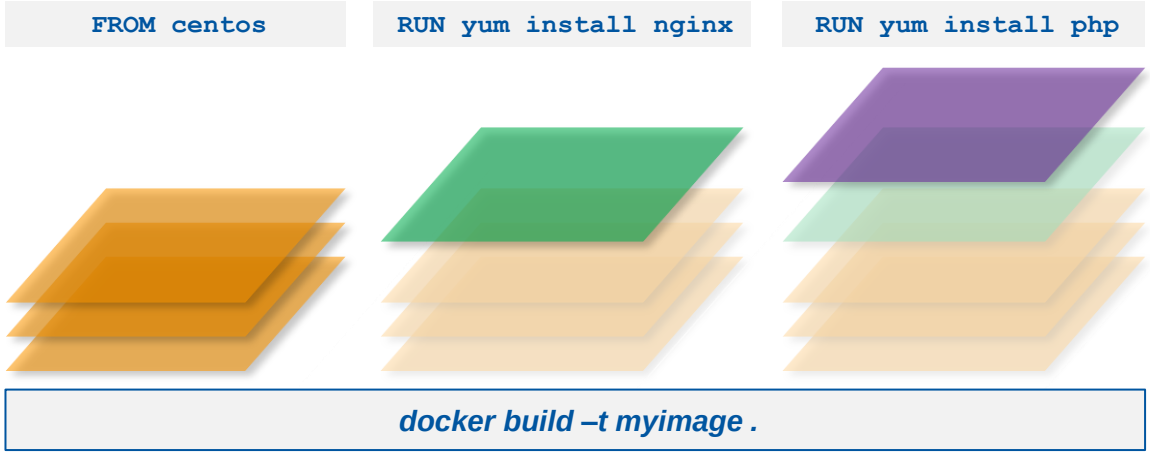
Backup

Recap on Containers Nomenclature

- **Container:** Runtime instance of an image and its execution environment
  - Provides isolation from the host environment (and from other containers)
  - Can access external resources – Network, volumes, host devices, …

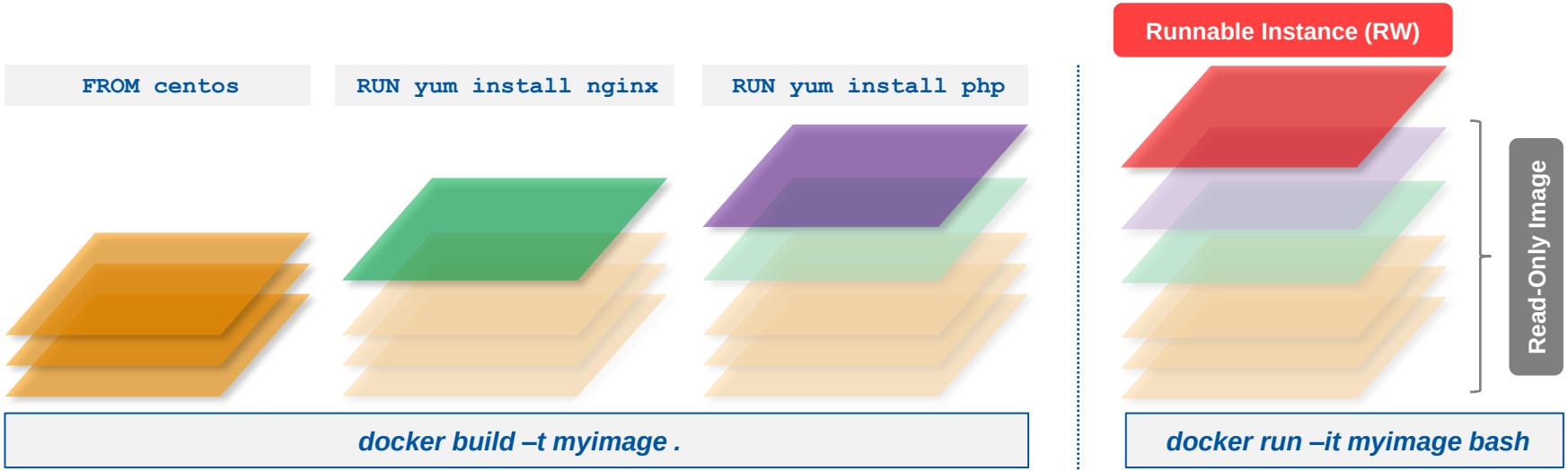
Recap on Containers Nomenclature

- **Container:** Runtime instance of an image and its execution environment
- **Image:** Self-standing portable package of software
  - Embeds all is needed to run an application (software, dependencies, settings, …)
  - Union of several layers (tar files) stacked together



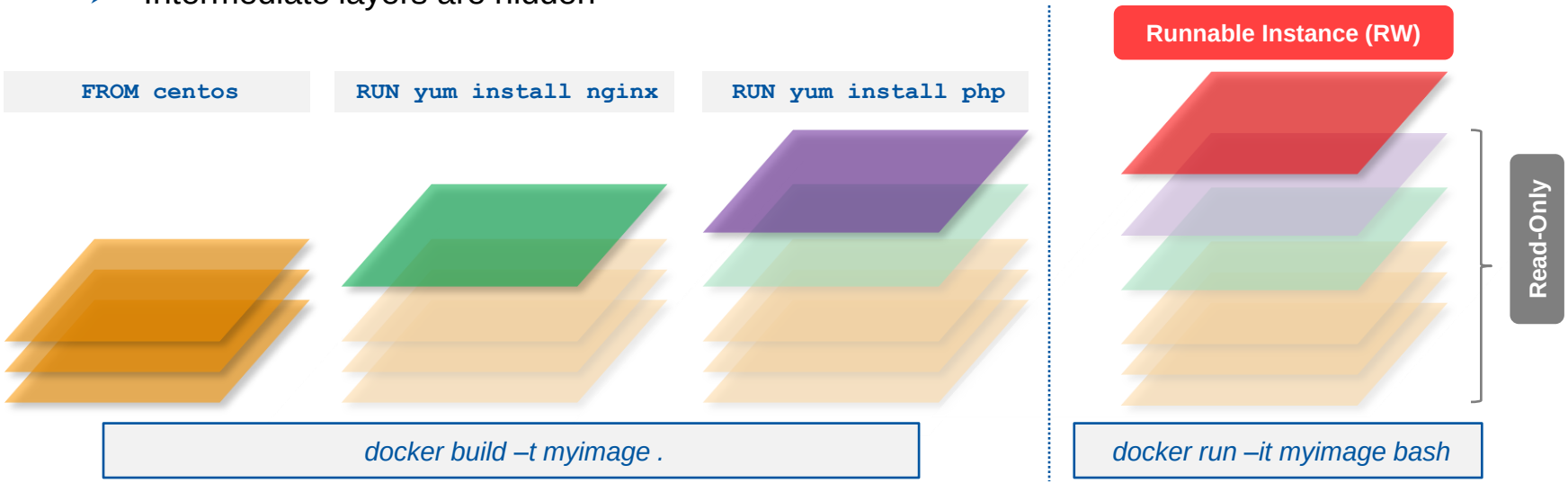
Recap on Containers Nomenclature

- **Container:** Runtime instance of an image and its execution environment
- **Image:** Self-standing portable package of software



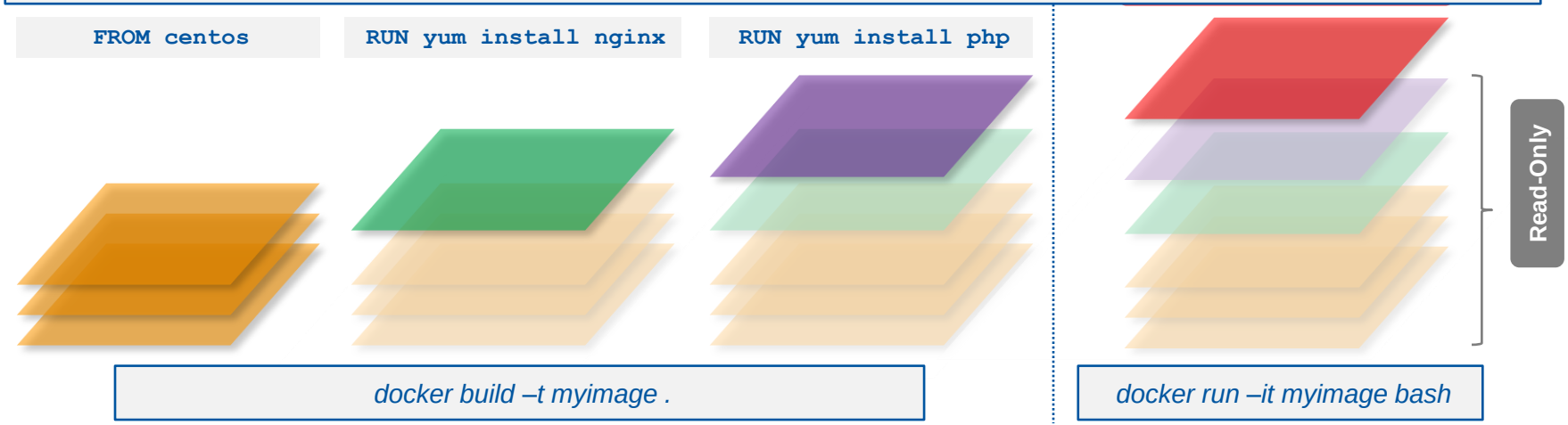
Quick Recap on Containers Images

- Image: Read-only template with instructions for creating a container
  - Produced as several layers (tar files) stacked together
  - Layering is used to improve storage utilization (can be reused)
  - Intermediate layers are hidden



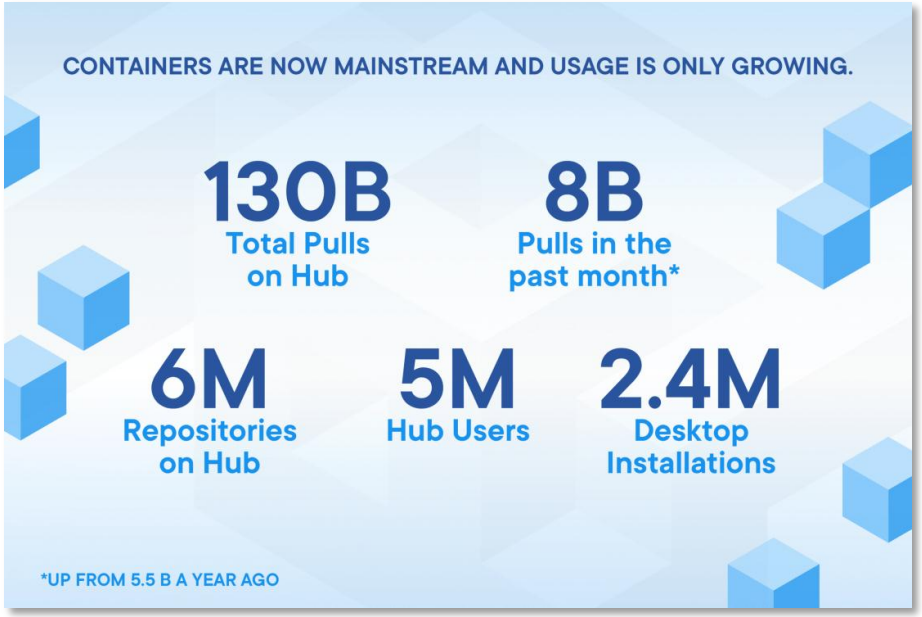
Quick Recap on Containers Images

```
[root@ThinkPad-X1]# docker history myimage
IMAGE               CREATED             CREATED BY                                      SIZE
75cc2375258a        4 seconds ago       /bin/sh -c yum -y install php                   66.9MB
e779b8a4024f        9 seconds ago       /bin/sh -c yum -y install nginx                 77.8MB
470671670cac        4 days ago          /bin/sh -c #(nop)  CMD ["/bin/bash"]            0B
<missing>           4 days ago          /bin/sh -c #(nop)  LABEL org.label-schema.sc…   0B
<missing>           7 days ago          /bin/sh -c #(nop) ADD file:aa54047c80ba30064…   237MB
```



The Docker Hub Registry

- Most popular public registry – Docker's default



Docker Index
04 February 2020

https://www.docker.com/blog/introducing-the-docker-index/

The Free Lunch Is Over

docker blog

Why Docker? Products Developers Pricing Company

Rate Limiting Questions? We have answers

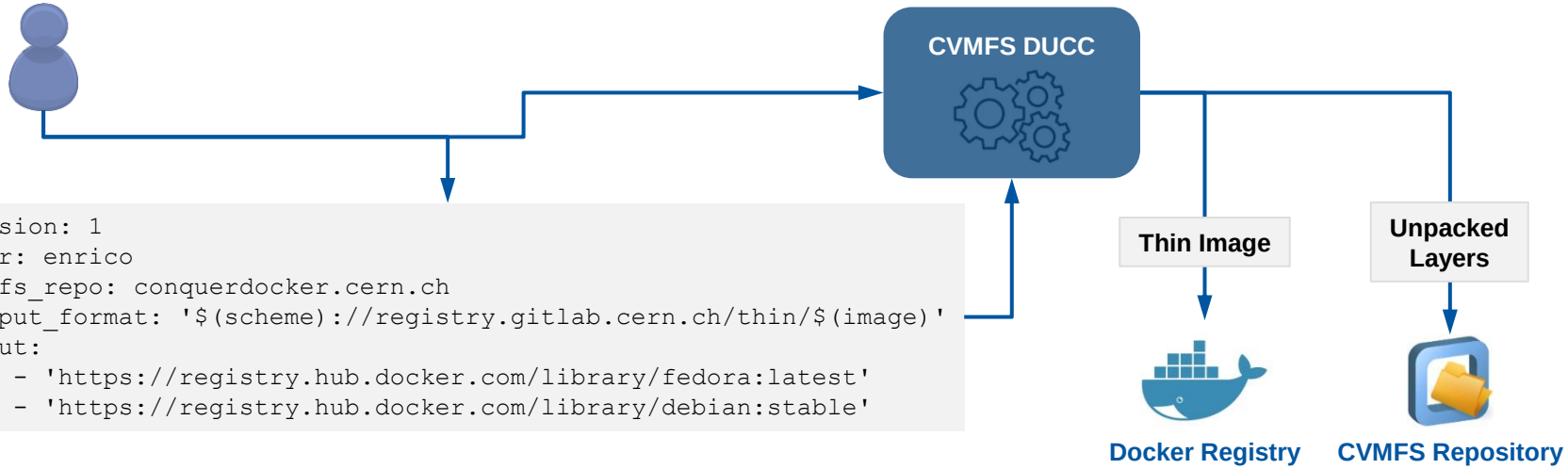
MATT CARTER
Nov 12 2020

As we have been implementing rate limiting on Docker Hub for free anonymous and authenticated image pulls, we've heard a lot of questions from our users about how this will affect them. And we've also heard a number of statements that are inaccurate or misleading about the potential impacts of the change. We want

- Unauthenticated:
  100 pulls / 6 hrs
- Free accounts:
  200 pulls / 6 hrs
- Mirroring to private registries recommended

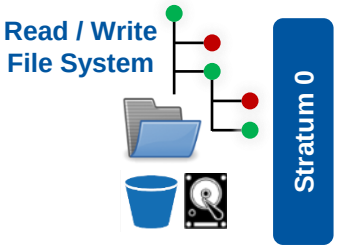
CVMFS ingesting Docker Layers

- DUCC: Daemon to convert and publish unpacked layers
  - Based on wishlist of Docker images to be ingested
  - Automatic generation and publication of thin image and unpacked layers

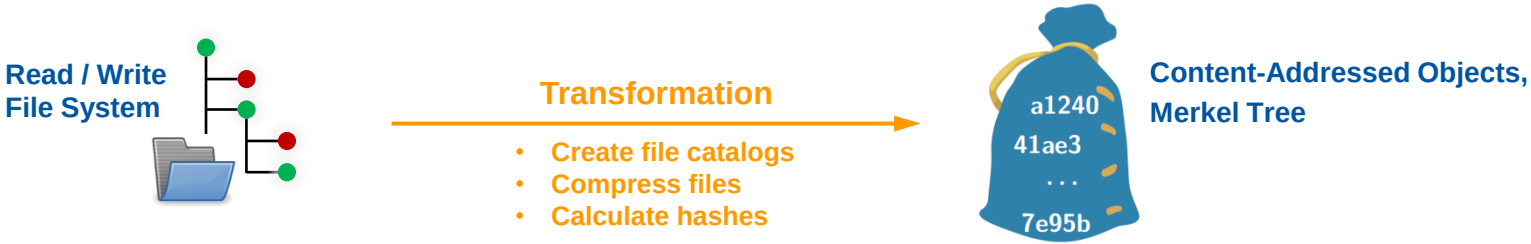


CVMFS Stratum 0s

- `cvmfs_server` package for repository management



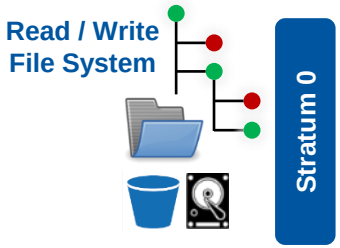

```
# cvmfs_server transaction myrepo.cern.ch
# cd /cvmfs/myrepo.cern.ch && tar xvf myarchive.tar.gz
# cvmfs_server publish myrepo.cern.ch
```

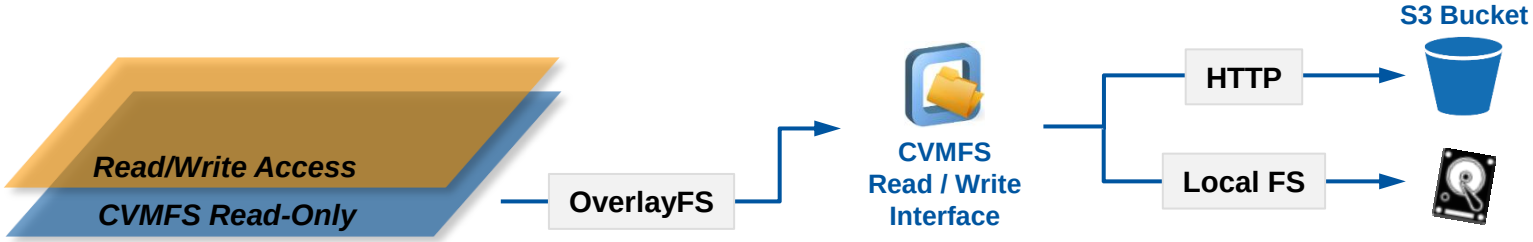


- Implicit file de-duplication via content-addressable objects
- Directory structure and file metadata stored in file catalogs

CVMFS Stratum 0s

- `cvmfs_server` package for repository management
- Authoritative storage for repository content
  - Local file system
  - S3 compatible storage system (e.g., Amazon, Ceph)

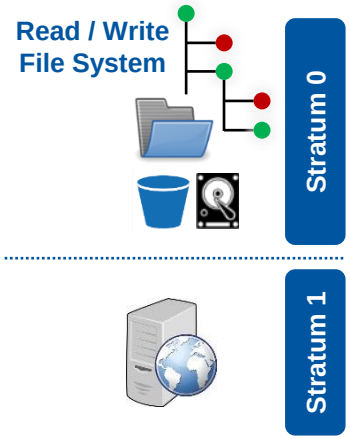




  - Updates applied by overlaying a copy-on-write union file system volume
  - Changes are accumulated in the volume and synchronized afterwards

CVMFS Stratum 1s

- Stratum 1 servers in Europe, US, Asia
  - Reduced RTT to caches and clients
  - Improved availability in case of Stratum 0 failure



- RESTful CVMFS GeoAPI service
  - Clients submit request with desired resource and Stratum 1s list
  - Stratum 1 returns sorted list of Stratum 1s
  - Based on MaxMind IP database

```
HTTP GET
http://s1.cs3.org/cvmfs/<desired_resource>/api/v1.0/geo/<list_of_known_stratum1s>
```

Site Caches

- Off-the-shelf HTTP caching software
- Squid-cache as forward proxy
  - Recommended for clusters of clients
  - Reduced latency to clients and load on Stratum 1s
- Take advantage of cloud based CDNs
  - OpenHTC on CloudFlare
  - Helix Nebula Cloud
    (RHEA, T-Systems, IBM Cloud)