Contribution ID: **153** Type: **Lightning talk**

Distribution of container images: From tiny deployments to massive analysis on the grid

Wednesday 27 January 2021 11:20 (10 minutes)

- Distribution of container images: From tiny deployments to massive analysis on the grid.

In recent years the use of containers proliferated to the point that they are now the de-facto standard to package and run modern applications. A crucial role in the successful distribution of container images is played by container registries, specialized repositories meant to store container images. Due to the popularity of containers, public registries had to constantly increase their storage and network capacity to withstand the huge demand from users. In August 2020, the Docker Registry has announced changes to the retention policy such that images hosted in free accounts would be deleted after 6 months of inactivity. While the enforcement of the new retention policy was postponed to mid-2021, many users of containers at CERN started to investigate alternative solutions to store their container images.

CERN IT offers a centralized GitLab Container Registry based on S3 storage. This registry is tightly integrated with code repositories hosted on CERN GitLab and allows for building and publishing images via CI/CD pipelines. Plans are to complement the GitLab Registry with Harbor (https://goharbor.io/), the Open Cloud Initiative container registry, which provides advanced capabilities including security scans of uploaded images, authentication and authorization (via LDAP, AD, OIDC, RBAC), non-blocking garbage collection of unreferenced blobs, and proxying/replication across other Harbor instances or Docker Hub.

Containers are also becoming more and more popular in the High Energy Physics community, where scientists encapsulate their analysis code and workflow inside a container image. The analysis is firstly validated on a small dataset to then run on the massive computing capacity provided by the Worldwide LHC Computing Grid. In this context, the typical approach of pulling a container image from a registry and extract it on the worker node show its limitations and results very inefficient. For this specific use-case, the CERN VM FileSystem (https://cernvm.cern.ch/fs/), a well-established service for the distribution of software at a global scale, comes to help. It features a dedicated ingestion engine for container images (based on per-file deduplication instead of per-layer) and an optimized distribution and caching mechanism that allows to greatly save on network bandwidth and local storage. The integration between CVMFS and the GitLab Registry (and Harbor) is being investigated to provide the end-users with a unified management portal for their container images (GitLab or Harbor) while supporting the large-scale analysis scenarios typical of the HEP world.

Author: BOCCHI, Enrico (CERN)

Presenter: BOCCHI, Enrico (CERN)

Session Classification: Tech Short Talks

Track Classification: Main session: Scalable Storage Backends for Cloud, HPC and Global Science