Containers WG status

GDB 13 February 2019

Gavin McCance

Recent meetings

- Last GDB update was in September last year
- Only two meetings since then (one of which Vincent B. kindly agreed to chair in my absence)
- Still very much focusing on Singularity together with how we might distribute container images efficiently
    - Containers w/Singularity very much in prod (CMS since Q1 2018)

Job containers

- Simple command line to run your job inside a containerised environment

  - Isolation for multiple payloads of the same pilot

  - Separation of job's OS from system one. We typically mount the container OS from an unpacked directory in CVMFS

  - Currently one tool, "Singularity", is in use, providing a common experience across SLC6 / CC7

Singularity status

- "Underlay" feature was added in 2.6 and tested
    - Bind mount tricks to allow mount of CVMFS directories unprivileged - lots of testing by experiments - seems to be working well for standard sites
- Sylabs (upstream) have reimplemented Singularity in Go (3.x branch)
    - 2.6 LTS security patches will be released
    - 3.0.3 in test but has some breaking issues still
    - Security review request with trustedci.org accepted (summer 2019)
- Recommendation is to stick on 2.6.x for now

Unprivileged

- General agreement that unprivileged (RH7.6 user namespaces) is the desirable end-goal modulo testing with experiments
    - Makes Singularity (with "underlay") a standard process (can run directly out of CMVFS)
    - ...and admits the use of other tools (containerd and friends) which probably have longer shelf-life
- Note some big areas (notably many HPC sites) where privileges are still needed (overlay, loop mounts of image files)
- Testing under-way with sites and experiments
    - CERN has enabled user namespaces on CC7 and will work with experiments for testing unprivileged-only
    - Note: unprivileged doesn't yet work properly for 3.x!

Container distribution

- General agreement that CVMFS distribution is most efficient and where we want to go
    - unpacked.cern.ch (new service, in test) / singularity.opensciencegrid.org (in prod) convert docker images to unpacked CVMFS directories
    - unpacked.cern.ch additionally converts docker to CVMFS-hosted docker layers, usable directly by docker/containerd (with driver)
    - CVMFS caching granularity for both is at file level - and most files in "image" are never actually accessed by running jobs
    - Other solutions needed for some HPC (e.g. packed images)

Container distribution

- Both repos use similar mechanism (PR referencing docker image -> unpacked directories/layers in CMVFS)

- Less obvious need for "common inter-experiment base image" but important within experiment to build images as a hierarchy of docker layers, to maximise cacheability

    - Experiments are doing this, e.g. building user-analysis containers off a common analysis-base

    - Important, notably for ATLAS who are working towards a model with (potentially many) user analysis containers

- Expiry / cleanup policies (in CMVFS) being discussed

Summary

- Singularity 2.6 branch in production now with underlay feature
    - New implementation by upstream on 3.x on the horizon and being tested
- Unprivileged (RH7.6, user namespaces with underlay) looks attainable in the future and some sites would like experiments to be able to support it
    - Testing with experiments and sites is needed, and is ongoing - works with Singularity 2.6 but not yet 3.x
- Standardising on container distribution via CVMFS for common case, which is by far the most efficient way
    - Works for both Singularity and docker/containerd, ...
- HPC add extra considerations (both on container distribution and tooling)