SLAC NATIONAL ACCELERATOR LABORATORY

Containers for HPCs

Wei Yang

ATLAS Site Jamboree and HPC Strategy
06-Mar-2019

Why do we need special containers on HPCs

No CVMFS on most US HPCs

- Install software on shared POSIX file systems
- Poor performance when running many concurrent jobs
  - due to small files (.py, .so) loading during jobs startup

If we put ATLAS software in a “fat” container, we may be able to solve

- Software distribution issue
- IO issue
- We use EXT3 image or SquashFS image for Singularity and Shifter

1st generation HPC containers

1. OS: came from ATLAS CentOS 6 singularity image in CVMFS
2. /cvmfs/atlas.cern.ch: extract from CVMFS using “uncvmfs” (or “Stratum-R”)
   - Incremental update, deduplicate
3. Mix them in an EXT3 file system (in a file)
   - “rsync” and its fine-grain filtering (IO intensive)
   - Incremental process to add new files from step 2.
   - Final image size 400-600 GB, built in 12 - 16 hours

   Or

3. **Mix 1 and 2 to a SquashFS image**
   - Using mksquashfs and its (weak) filter rules (CPU intensive)
   - Deduplicate, ~200 GB, built in 8 - 10 hours

Challenges:

1. Need to know what can be filtered,
2. and what can not.
3. Difficult to move the image around, or delivery to its users.

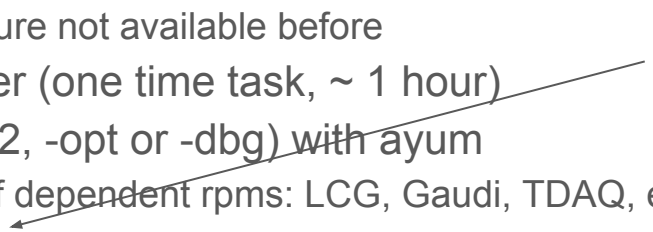
Built at BNL. Works well at NERSC and ALCF

2nd Try: Installing ATLAS software in a container

Several people tried this idea with different approaches. SLAC method below:

- Install under /cvmfs/atlas.cern.ch inside the container, same as real CVMFS
- To be able to run under Grid production, for all workflows

1. OS: came from ATLAS C6 singularity image in CVMFS (one time task)
   - Put it in a Singularity sandbox - a feature not available before
2. Install manageTier3SW in the container (one time task, ~ 1 hour)
3. Install a 21.0.x release (gcc49 or gcc62, -opt or -dbg) with ayum
   - First release brings in large number of dependent rpms: LCG, Gaudi, TDAQ, etc. ~1 hour
   - Additional 21.0.x release, ~15 minutes. Final size depend on how many release to be included
4. Copy missing things from CVMFS into the container
   - DBReleases, Calibration data, sw/local, sw/ddm
5. Compress into a SquashFS image (with some obvious filtering)



Testing at SLAC

We run jobs at SLACXRD_MP8 to test the SquashFS Image:

Have to install a large number of releases:

- 21.0.{15,16}-gcc**49**-opt, 21.0.{22,31,37,53,54,72,77,81,89,90}-gcc**62**-opt
- and any other releases been used at the site
- Image size so far: 29 GB

We can repeat the same rpm installation process for 21.0.x at here

Still missing:

- Release 21.2.x; gcc6.2.0-2bc78 from stf.cern.ch; condDB repo
- Identified missing files/packages through debugging
- Included them through bind mounts from real CVMFS

Lesson Learned, Questions

1. A single release image will likely be small enough (< 20 GB) to run on a desktop/laptop
2. Is it possible to further trim down manageTier3SW?
   a. Remove the i386 stuffs
   b. Remove old versions of software (e.g. xrootd 3.x?)
3. Can we remove the dependence on gcc6.2 in sft.cern.ch?
   a. Why not just install it in atlas.cern.ch ?
4. Maintenance cost is high if we constantly add releases
   a. And **have to debug what is wrong with newly added releases**
      i. Just saw something wrong with SLACXRD_MP8: release 21.2.55, AODtoDAOD
   b. If HPC operations used just a few releases for long period, maintenance will be lower

Acknowledgement

Many thanks to Asoka De Silva, Doug Benjamin, Taylor Childers, Lukas Heinrich, Vakho Tsulaia for sharing knowledges and debugging problems