Containers in the Open Science Pool

Mátyás ("Mat") Selmeci

HTC 23 - July 12, 2023

PATh PARTNERSHIP to ADVANCE THROUGHPUT COMPUTING
iris hep
OSG
CHTC Center For High Throughput Computing

Containers for jobs in the OSPool

- How they've worked in the past
- What major changes we've made in the last two years
- What our plans are

Containers in the OSPool (2017-2021)

- Only Singularity is available – pilot EPs don't have the privileges to use Docker
- SIF files (single-file Singularity images) were too large to transfer for tens of thousands of jobs, and not all sites supported them
- Singularity "sandbox" (i.e., directory tree) images used instead
  - we distribute them via CVMFS so only the files that jobs actually made use of get transferred

Singularity in the OSPool (2017-2021)

- Users upload (or find) an image in a public Docker registry (e.g., Docker Hub), and tell OSG Staff the location (in a PR to cvmfs-singularity-sync)

- A cron job pulls each listed image, converts it to a Singularity sandbox, and adds it to CVMFS

- Users specify the CVMFS paths in their submit files, e.g.:
  +SingularityImage="/cvmfs/singularity.opensciencegrid.org/htc/el8:latest"

Singularity in the OSPool without CVMFS (2021-)

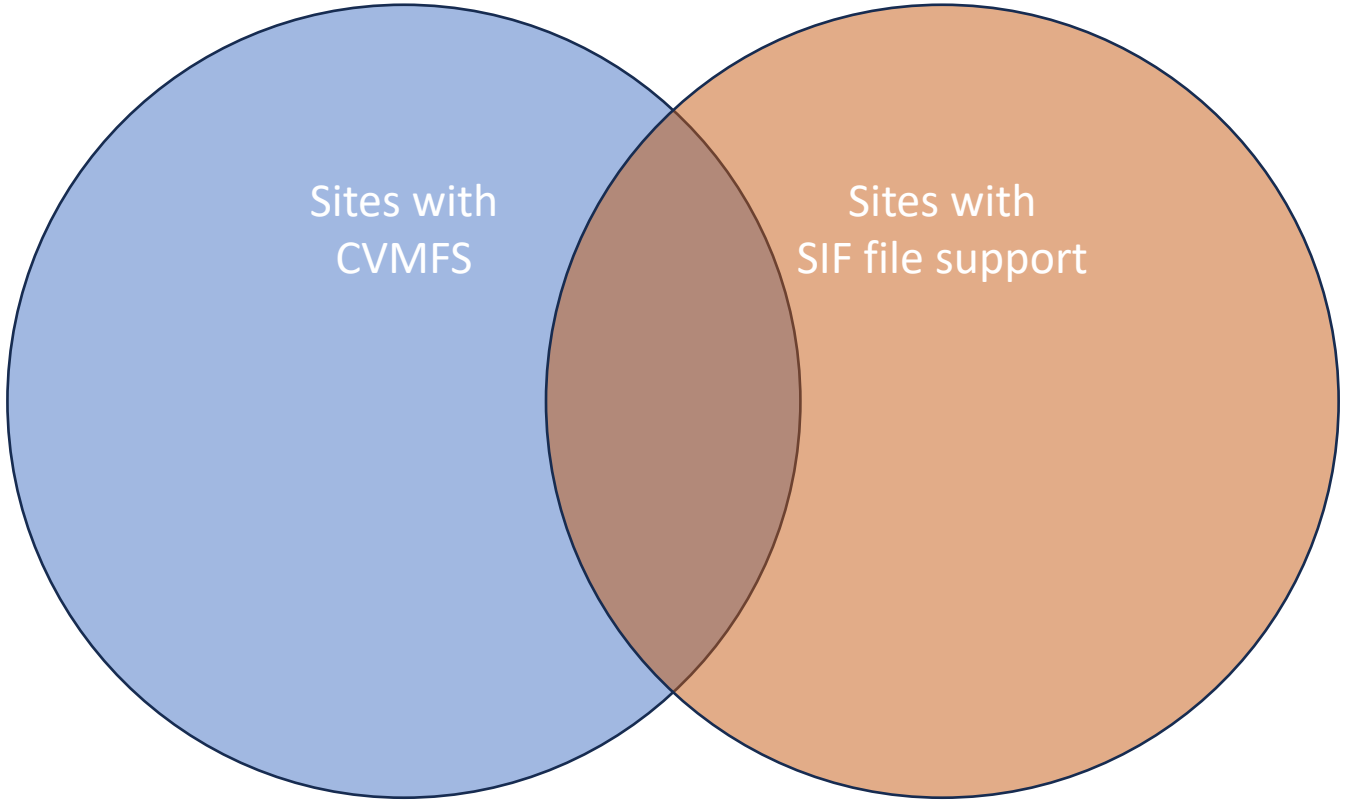
- Many sites had Singularity installed but not CVMFS
- How do we get our images to these sites?

- SIF files are big but are frequently reused
- By 2021 we had a mature caching infrastructure: OSDF
- We could use all our images from there instead

- ... if only all sites supported SIF files

A mix of site support

- Can't just keep going with CVMFS – there are more sites we want to reach
- Can't completely switch to SIF-via-OSDF – we'd lose sites that can't use SIFs
- How do we use both without making the user worry about it?



Adding a layer of indirection

- Every image added to /cvmfs/singularity.opensciencegrid.org now also gets added to OSDF as a SIF
- Users keep specifying the CVMFS paths in their submit files
- If the EP decides SIF-via-OSDF is better, then the EP will download the SIF and use it instead of CVMFS
- The SIF is saved in a per-pilot cache outside the execute directory:
  - The SIF could be reused by multiple jobs
  - Users shouldn't need to request more disk for the SIF if it wasn't their decision to use it
- In production since mid-2021

Implementation

Where does all this code run? (2017-2022)

- Until this year, the code for running OSPool jobs in Singularity was in the USER_JOB_WRAPPER

- A USER_JOB_WRAPPER is a script on an EP that "wraps around" each job's executable:
  - HTCondor calls the wrapper *instead of* the job executable
  - Wrapper maybe does some setup
  - Wrapper runs the original executable
  - Wrapper maybe does some teardown

- The OSPool job wrapper, not HTCondor, was launching Singularity

What's wrong with job wrappers?

- Wrappers are a "black box" to HTCondor:
  - HTCondor doesn't know if the exit code is from the job or the wrapper
  - HTCondor doesn't know if stdout/stderr is from the job or the wrapper

What was wrong with the OSPool job wrapper?

- HTCondor most definitely didn't know that the job wrapper launched the real job in Singularity!
- Users' stderr files got polluted with messages from Singularity
- condor_ssh_to_job landed *outside* the container, not inside
- 800+ lines of Bash

How did we fix the job wrapper?

- By 2023, HTCondor supported all the Singularity features we needed
- Almost all of what the job wrapper did, we could move into either:
  - HTCondor config for the EP (for things that only needed to be run once per pilot):
    - Downloading the default image
    - Configuring Singularity extra arguments
    - Configuring volume mounts
  - A "prepare-job hook" (for things that needed to be run once per job)
    - Obtaining the job's image from CVMFS or OSDF

What's a prepare-job hook?

- A script on an EP that can run before each job
- Runs in a separate stage so HTCondor can distinguish it from the real job
- Hooks can modify job ads which may change how HTCondor launches the job
- Errors in the hook can be distinguished from errors in the job; specific error codes and messages can be returned for better debugging

- The download code was moved from the user job wrapper to a prepare-job hook
- After the download is complete, the hook sets the image for HTCondor to use for launching the job

What do OSPool users see? (2023-)

- OSPool switched to the job hook in February 2023
- No more spam about "/lizard not found or not a directory" in your stderr
- You can condor_ssh_to_job into Singularity jobs
- Otherwise... everything should work as before

What's next?

- Container Universe – works from the OSG Staff-managed APs, would like to get it working for flocking APs too
- Management of EP image cache – image cleanup, cap on space, etc.
- Management of SIF image transfer – no running stashcp in a Bash script

- Making all this code portable and useable by others

Thank you!
Questions?

This material is based upon work supported by the National Science Foundation under Grant Nos. 1836650 and 2030508. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.