Singularity: Ops Status and Issues

BONUS Slides: CMS Singularity Status and Open Questions

Brian Bockelman, pre-GDB, July 2017

Singularity: Ops Status

- Singularity 2.2.1 is distributed by EPEL and in use by CMS and ATLAS.
    - CMS and OSG worked closely together and adopted a 95%-common solution.
        - 2.2.x is the minimum version to have all the required features.
- Singularity has modest adoption amongst the WLCG sites. CMS sees about 12 sites using this.
    - Both OSG and EGI middleware are represented.
    - If both glexec and Singularity are present, Singularity is preferred.

Image Distribution

- Singularity supports several image formats.
    - Some formats (raw-image-like, squashfs) really needs a capable shared file system for efficient distribution.
    - Other formats (.tgz, Docker) require unpacking into a cache directory for each invocation -> not great for the high-throughput case.
- Given our heavy investment in CVMFS, it seems very natural to leverage it for image distribution.
    - Given CVMFS implementation details, images should be distributed as **flat directories**. Cache will work at the individual file level.
    - Flat directories work with non-setuid mode; single-image formats don't. Would *strongly* prefer non-setuid mode in the future.

Image Distribution - Future

- OSG maintains a DockerHub-to-CVMFS service. Used by several OSG users and CMS.
- “Flat directories in CVMFS” scales well at a coarse granularity.
    - Great for “one image per VO” or “one image per VO per OS.”
    - OK for “one image for science application.” CVMFS-based de-duplication saves significant cache space!
        - Preserving final version of an analysis for later reproduction.
- Likely *not* OK for:
    - Nightly builds on “regular” CVMFS infrastructure. Use same techniques as done for WLCG VO integration builds.
    - “One image per job” or “One image per user”. Highly dynamic files should be shipped with the job.
- Limitations due to:
    - Latency required to push updates through CVMFS.
    - CVMFS Stratum-1s must have a full copy of all unique files.

Non-technical Issues

- Overarching theme: new features and capabilities don't come for free.
- Young, fast-moving software project. Adding new features a good rate.
    - Downside of new features is new security bugs and unexpected backward compatibility breaks.
        - I expect big Singularity releases to result in validation work for VOs. OSG integration already has customizations for 2.2.x versus 2.3.x.
        - In the past year, there have been two security releases - one which potentially affected sites. Strongly suggest we aggressively coordinate with upstream.
- User-based isolation - defects and all - are well-understood and accepted.
    - Some aspects of container-based isolation are less-understood. **Example: how do you validate Singularity was invoked with the correct configuration and command line arguments?** Easier to determine two processes have distinct UIDs than determining whether they are "sufficiently" separated using namespaces.

Technical Issues - RHEL6

- Most technical issues encountered result from the RHEL6 kernel:
- **Problem**: OverlayFS is not available, meaning all bind mount destinations must exist in the image.
    - **Solution**: CMS rewrites environment variables and $PWD / $HOME for user jobs to a single known location (/srv).
- **Problem**: Automount-points do not propagate new mounts inside the container.
    - **Solution**: Prior to job startup, probe all needed CVMFS
- **Problem**: Kernel does not track mount point usage inside container. Autofs will unmount CVMFS during the job.
    - **Solution**: (OSG) Pilot runs periodic check of CVMFS mounts; side effect is autofs keeps the mount points active.
    - **Solution**: Have invoking process `cd` into the mount point, keeping a reference alive.

Non-RHEL6 Related

- **Problem**: Singularity is very restrictive about special characters in path names (HTCondor-CE/blahp generate paths with banned `#` character).
    - **Solution**: (Partial; works in practice) Change to a local scratch directory before startup.
- **Problem**: Environment variables set by sysadmin (or pilot!) may be invalid inside containers. E.g., add CVMFS directories to $PATH/$LD_LIBRARY_PATH with EL7-specific `gfal-copy` binaries.
    - **Solution**: Not clear if there is a generic solution.
    - CMS is looking at sourcing container-compatible grid UI environment from inside container.
    - CMS rewrites path names inside environment variables for whitelisted / known problems.

Why Singularity for CMS?

- Singularity offers CMS sites some massive simplifications:
    - Drop the dreaded “pool accounts” -> reduce a few thousand CMS-related accounts in your LDAP server.
    - A road to “no-setuid” for grid worker nodes. Hopefully, the light at the end of the tunnel is not an oncoming train.
    - Remove a config file you must customize to your site.
    - Removes need for site-level authorization server. Instead of a complex grid-specific service, authorization can be handled via a config file.

CMS Status

- Previously-mentioned issues currently affecting CMS sites:
  - `#` character in path names.
  - SL7 $PATH components inside SL6 image.
- Solutions being tested for both. Good support from the CMS site support team. Sites impatiently waiting on fixes.
- If Singularity is present and passes pilot startup verification, it is used for every payload job.
- SAM tests adopted so it can successfully land on either RHEL6 or RHEL7. Either glexec or singularity must be present.
- About 30M payload jobs run under Singularity.

CMS Open Issues

- Two things still stink:
    - CMS sites sometimes use CVMFS variant symlinks to have worker node config files locally instead of distributed by CVMFS. We now restrict the location they can put their config files!
    - POSIX storage elements must be mounted at a specific location (/cms) and the site config must reflect this.
        - Stageout now must go through GridFTP as the payload UID may not be the desired storage UID.
- Not clear either is “fixable”.
- Collection of site advice here: https://twiki.cern.ch/twiki/bin/view/Main/CmsSingularity