Singularity Update

1 March 2017

Brian Bockelman

Singularity Integration Progress

- OSG VO has made significant progress in integrating / using Singularity.
    - If a pilot detects the binary & it passes validation, then *all* payload jobs are started inside the container.
    - Very quick ramp-up of site support. Only need to **install a single RPM**. There is no daemon to run, no config files to tweak, no UID switching: admins are OK with installing this.
- **About 15 sites total**. Variety of sites have helped to shake out the bugs in the integration.
- We passed **1M singularity-based jobs** total this week and peak at **250k singularity-based jobs per day**.
    - Usage has **increased 2 orders of magnitude** in the last two weeks.
- **40-60% of the OSG VO** pool utilizes singularity.

CMS Integration

- CMS integration started this week.
    - Mostly copy/paste of scripts from OSG to CMS instance of glideinWMS.
    - Some customization required:
        - Instead of users specifying an image name, CMS users specify `rhel6` or `rhel7`. Must translate that to an official image.
        - Must figure out how to detect which bind mounts to perform if site mounts storage via POSIX.
- **Deploying to CMS testbed ~ today**. Hopefully production & analysis test jobs are successful by the end of the week.

Considerations for Integration

- RHEL6-only: new CVMFS mounts will not be forwarded into the container. Must at least touch them from pilot before executing singularity.
- Bind mount job working directory into a fixed place (CMS uses `/srv`).
    - Transform paths in command line arguments and environment variables to reflect new job working directory name.
- Making `$PWD` and `$HOME` point at the new working directory.
- Must whitelist any writable / accessible directory outside the container.
    - CMS had to copy a few files from the pilot directory into the job directory.
    - RHEL6-only: no overlayfs; bind mount targets must exist inside container. *Trickier than it sounds*: POSIX mount points (say, Lustre) must exist inside the container.
- For CMS, this was 200 lines of bash.
- Since there is no UID switching, no other parts of the pilot need to be refactored besides job startup.

Example Invocation

```
singularity exec --home $PWD:/srv \
                 --bind $PWD:/srv \
                 --pwd /srv \
                 --scratch /var/tmp \
                 --scratch /tmp \
                 --containall \
                 /cvmfs/singularity.opensciencegrid.org/bbockelm/cms:rhel6 \
                 /srv/.osgvo-user-job-wrapper.sh $CMD
```

From:
https://github.com/jamesletts/CMSglideinWMSValidation/blob/master/singularity_wrapper.sh

Image Distribution

- If a user wants a custom Docker image available, they **only need to send a PR** to https://github.com/opensciencegrid/cvmfs-singularity-sync/blob/master/docker_images.txt
    - Flat file format; just add the Docker Hub image name to the file.
- **Within 30 minutes**, the requested Docker image is synchronized to CVMFS and ready to be used with Singularity.
    - `docker pull opensciencegrid/osg-wn:3.3-el7` becomes `/cvmfs/singularity.opensciencegrid.org/opensciencegrid/osg-wn:3.3-el7`
    - Once whitelisted, any subsequent pushes to the image in Docker Hub are also synchronized to CVMFS in about 30 minutes. **No human involvement for updates**.
- **Currently about 40 images**. Ranges from generic worker nodes to VO-specific images to application stacks.
- OSG defaults to `rynge/osgvo:el6`. CMS defaults to `bbockelm/cms:rhel6`.

CMS Policy Update

- CMS has agreed on a policy:
  - Starting 1 April 2017, a site can provide a RHEL7 environment to the pilot ***only if*** **it also provides `singularity`**.
  - Otherwise, pilot must start with a RHEL6 environment.
- If singularity is not provided on RHEL7, CMS may have very few jobs that can run on the site.
- Policy was carefully crafted to allow other batch-system-level container experiments to proceed! E.g., Nebraska still launches everything inside Docker.
- Policy allows sites to proceed with upgrades at their own pace. **Does not provide a good "drop dead" deadline for RHEL6 hosts**. We have full control of transition.