Deploying and managing containers on OSG

Marco Mambelli
HTCondor Week 2019
May 22, 2019
Where we left off

• CMS and OSG with their own wrapper
• GlideinWMS initial support of Singularity via USER_JOB_WRAPPER modeled on CMS and OSG
• HTCondor support of Singularity not flexible
Goal

- GlideinWMS would
  - implement all the features used by the VOs
  - Provide more structure
  - Migrate seamlessly the users from the different VO solutions to an HTCondor based solution
GlideinWMS Singularity support

• USER_JOB_WRAPPER (initial test & setup, re-invocation in singularity, final test & setup)
• Up-to-date with OSG and CMS wrappers (tests + setups)
  – Libraries support
  – GPU support
  – Auto-discovery of the Singularity binary, including the OSG-distributed unprivileged singularity
• Flexible support of bind mounts
Singularity use negotiation

- Singularity use is negotiated between jobs requestors and resource providers

<table>
<thead>
<tr>
<th></th>
<th>NEVER</th>
<th>OPTIONAL</th>
<th>PREFERRED</th>
<th>REQUIRED (REQUIRED_GWMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEVER</td>
<td>NEVER</td>
<td>NEVER</td>
<td>NEVER</td>
<td>FAIL</td>
</tr>
<tr>
<td>OPTIONAL</td>
<td>NEVER</td>
<td>NEVER</td>
<td>PREFERRED</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>PREFERRED</td>
<td>NEVER</td>
<td>PREFERRED</td>
<td>PREFERRED</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>REQUIRED</td>
<td>FAIL</td>
<td>REQUIRED</td>
<td>REQUIRED</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>DISABLE_GWMS</td>
<td>DISABLE</td>
<td>DISABLE</td>
<td>DISABLE</td>
<td>DISABLE (FAIL)</td>
</tr>
</tbody>
</table>

- Procedure to migrate from the VO solutions
- Evaluations of GlideinWMS wrappers in progress
- Plan to migrate to HTCondor starting Singularity
Images support

• Support for multiple images
  – Image dictionary
  – Mountables or CVMFS-Expanded images
  – Default SL6 and SL7 (OSG provided)
  – ~ 150 images pre defined by OSG staff
  – VO provided images
  – Researcher provided images

• Workflow for VO/User-defined images
  – [https://github.com/opensciencegrid/cvmfs-singularity-sync](https://github.com/opensciencegrid/cvmfs-singularity-sync)
  – Facilitate containers adoption
User-defined Container Workflow

From Mats Rynge’s presentation for OSG
Singularity in HTCondor

- Is much more flexible - thanks Greg Thain!

```plaintext
SINGULARITY = /cvmfs/oasis.opensciencegrid.org/misc/singularity/current/bin/singularity
SINGULARITY_IMAGE_EXPR = (isUndefined(TARGET.SingularityImage)) ? "/cvmfs/singularity.opensciencegrid.org/opensciencegrid/osgvo-el7:latest" : TARGET.SingularityImage
SINGULARITY_JOB = isUndefined(TARGET.NoSingularity)
#SINGULARITY_EXTRA_ARGUMENTS = "-c -i -p"
#SINGULARITY_EXTRA_ARGUMENTS = "--home "$PWD":/srv --pwd /srv --ipc --pid"
SINGULARITY_EXTRA_ARGUMENTS = "--ipc --pid --contain"
SINGULARITY_BIND_EXPR = "/cvmfs/opt/condor/dir2 /opt/condor/dir1"
SINGULARITY_TARGET_DIR = /srv
SINGULARITY_IS_SETUID = False
```

- Since HTCondor 8.8.2 it passes to the USER_JOB_WRAPPER the invocation line including the singularity invocation

- Supports condor_ssh_to_job
How does it looks like

- Command line passed at the wrapper

  Args: /usr/bin/singularity exec -S /tmp -S /var/tmp -B /cloud/login/marcom -B /opt/condor/v8_8_2/local.fermicloud047/execute/di731235:/srv -B /opt/condor/dir2 -B /opt/condor/dir1 --ipc --pid --contain -C /cvmfs/singularity.opensciencegrid.org/opensciencegrid/osgvo-el7:latest /bin/sleep 60

- Test script

  TEST stdout
  whoami: marcom
  uid=4667(marcom) gid=1752(cdadmin) groups=1752(cdadmin)
  context=system_u:system_r:condor_startd_t:s0
  Root user (/bin/bash -c echo "\{ read first _ < /proc/$$/uid_map ; echo "$first" ; \} 2>/dev/null" ): 4667

- And the processes

  marco    730783  730775  0 May21    ?       00:00:00  \_ condor_startd -f
  marco    742169  730783  0 00:02    ?       00:00:00  \_ condor_starter -f
  fermicloud000.fnal.gov
  marco    742178  742169  0 00:02    ?       00:00:00  \_ /cvmfs/oasis.opensciencegrid.org/mis/singularity/el7-x86_64/2.6.1/libexec/singularity/bin/action /bin/bash -c sleep 60
  marco    742192  742178  0 00:02    ?       00:00:00  \_ shim-init
  /bin/bash -c sleep 60
  marco    742193  742192  0 00:02    ?       00:00:00  \_ sleep 60
Nested containers?

- Works!

whoami: marcomuid=4667(marcom) gid=1752(cdadmin) groups=1752(cdadmin)
context=unconfined_u:unconfined_r:unconfined_t:s0-s0:c0.c1023
/opt/condor/v8_8_2/local.fermicloud047/execute/dir_794998:/srv -B /cvmfs -B
/opt/condor/dir2 -B /opt/condor/dir1 --ipc --pid --contain
/cvmfs/singularity.opensciencegrid.org/opensciencegrid/osgvo-el7:latest
/cvmfs/oasis.opensciencegrid.org/mis/singularity/current/bin/singularity exec -S /tmp -S
/var/tmp -B /cloud/login/marcom/condor-test -B
/opt/condor/v8_8_2/local.fermicloud047/execute/dir_742460:/srv -B /opt/condor/dir2 -B
/opt/condor/dir1 --ipc --pid --contain
/cvmfs/singularity.opensciencegrid.org/opensciencegrid/osgvo-el7:latest /bin/bash -c echo "`{ read first _ < /proc/$$/uid_map ; echo "$first" ; } 2>/dev/null`"

TEST stdout
whoami: marcom
uid=4667(marcom) gid=1752(cdadmin) groups=1752(cdadmin)
context=system_u:system_r:condor_startd_t:s0
Root user (/bin/bash -c echo "`{ read first _ < /proc/$$/uid_map ; echo "$first" ; } 2>/dev/null`" ): 4667
Nested containers? Not really

- Needed some manual tweaking

```
whoami: marcomuid=4667(marcom) gid=1752(cdadmin) groups=1752(cdadmin)
context=unconfined_u:unconfined_r:unconfined_t:s0−s0:c0.c1023
Args: /cvmfs/oasis.opensciencegrid.org/mis/singularity/3.1.1/bin/singularity
extec −S /tmp −S /var/tmp −B /cloud/login/marcom −B
/opt/condor/v8_8_2/local.fermicloud047/execute/dir_794998:/srv −B /cvmfs −B
/opt/condor/dir2 −B /opt/condor/dir1 --ipc --pid --contain -C
/cvmfs/singularity.opensciencegrid.org.opensciencegrid/osgvo-el7:latest
/cvmfs/oasis.opensciencegrid.org/mis/singularity/current/bin/singularity exec
−S /tmp −S /var/tmp −B /cloud/login/marcom/condor-test −B
/opt/condor/v8_8_2/local.fermicloud047/execute/dir_742460:/srv −B
/opt/condor/dir2 −B /opt/condor/dir1 --ipc --pid --contain
/cvmfs/singularity.opensciencegrid.org.opensciencegrid/osgvo-el7:latest
/bin/bash -c echo "\{ read first _ < /proc/$$/uid_map ; echo "$first" ; } 2>/dev/null"
ERROR : Multiple devpts instances unsupported and "mount devpts" configured
ABORT : Retval = 255
```

- Need to remove the “containall” (-C) option
Condor_ssh_to_job

• Works with system condor

```bash
$ condor_ssh_to_job 21Welcome to fermicloud047.fnal.gov!
Your condor job is running with pid(s) 707226.
-sh: cannot set terminal process group (-1): Inappropriate ioctl for device
-sh: no job control in this shell-sh-4.2
$ ps PID TTY          TIME CMD
1 ?   00:00:00 shim-init
2 ?   00:00:00 sleep
3 ?   00:00:00 sh
26 ?   00:00:00 ps
```

• Something missing in unprivileged mode

```bash
bash-4.2$ condor_ssh_to_job 6
Welcome to fermicloud047.fnal.gov!
Your condor job is running with pid(s) 740832.
senter: cannot open reserve-credentials: No such file or directory
```
HTCondor Desiderata

• More flexibility
  – Handling of bind-mounts (not simple path w/o spaces)
  – Invocation options (-C)
• Some things could be automatic (SINGULARITY_IS_SETUID)
• Singularity support is mostly documented
• More clear errors in condor_ssh_to_job
• Mechanism to select the platform: most VOs are used to request for a specific OS an, Singularity allows to morph
• Resource requirements depending form the available container (loose border)
  – You can choose an Image
  – You can match on Machine attributes