Containers deployment

A. Forti

Jamboree

06 March 2019

Use Cases

- In the system most of these use cases are satisfied only if the pilot runs the container
  - In particular payload isolation

Id	Use case
1	Installation of different OS from SL/RHEL /CentOS
2	OS upgrades don't need coordination with experiments anymore
3	Minimal installation on the nodes if sites prefer
4	Allows experiment to run tests with specific software or setups
5	May offer another approach to software distribution to sites that don't support CVMFS
6	Reduces the impact of ATLAS software on large shared file systems on HPC resources
7	Payload isolation
8	User containers
9	GPUs
10	Benchmarking suite

Requirements

- At CentOS7 sites singularity is a **requirement**.
- Current baseline: **singularity 2.6.1**
  - **Default configuration works** for ATLAS
- **3.x** series is not yet functional for ATLAS
  - It's not even in EPEL for this reason yet
  - **Don't install it**

AGIS

- To enable the pilot containers functionality
  - container_type: singularity:pilot
    - In the future it will be possible to replace singularity with other rootless runtimes like podman or dockerd but for now we stick to it
  - container_options: <options>
    - This can be left empty. It might be useful for testing.or for sites that want the container to start with specific options

singularity:pilot

CVMFS containers

- Both production and users workflows
- Depend on pilot2 commissioning
  - Need container_type parameter set
  - Unpacked in /cvmfs/atlas.cern.ch/repo/containers/fs/singularity
- Doesn't need setuid, or overlay
- Limited to use software in cvmfs
- Containerization transparent to user and system administrator
  - Process tree similar to standard job

```
-runpilot2-wrapp ./runpilot2-wrapper.sh -q ANALY_MANC_TEST_SL7 -r ANALY_MANC_TEST_SL7 -s ANALY_MANC_TEST_SL7 -j user -d
  └─python pilot2/pilot.py -q ANALY_MANC_TEST_SL7 -r ANALY_MANC_TEST_SL7 -s ANALY_MANC_TEST_SL7 -i PR -j user --pilot-user=ATLAS -d
      ├─bash -c...
      │  └─startContainer. ...
      │      └─action-suid /alrb/.bashrc
      │          └─shim-init                                  /alrb/.bashrc
      │              └─.bashrc /alrb/.bashrc
      │                  └─python -u -Wignore ./runAthena-00-00-12 -a sources.20132189.derivation.tgz -r ./ --trf --useLocalIO --useCMake -
      │                      └─sh -c...
      │                          └─python /cvmfs/atlas.cern.ch/repo/sw/software/21.2/AthDerivation/21.2.33.0/InstallArea/x86_64-slc6-gcc62-
      │                              ├─MemoryMonitor --pid 2192 --filename mem.full.AODtoDAOD --json-summary mem.summary.AODtoDAOD.json --i
      │                              └─runwrapper.AODt ./runwrapper.AODtoDAOD.sh
      │                                  └─athena.py -tt/cvmfs/atlas.cern.ch/repo/sw/software/21.2/AthDerivation/21.2.33.0/InstallArea/
      │                                      └─{athena.py}
```

Standalone containers

- User workflows only
- Triggered by the user with a command line option
  - Independent from the pilot version
- Uses custom user images from a docker registry
  - Analysis images
  - Machine Learning images
  - Official docker images
- Doesn't need CVMFS to run

```
prun --containerImage docker://alpine --exec "echo 'Hello World\!'" --tmpDir /tmp --outDS user.aforti.test.20190306141519 --noBuild --site ANALY_MWT2_SL7
```

PanDA ID Attempt# of maxAttempts#	Owner Group	Request Task ID	Transformation	Status	Created	Time to start d:h:m:s	Duration d:h:m:s	Mod	Cloud Site
4267387837 Attempt 1 of 3	alessandra forti	1471 17334486	runcontainer	finished	2019-03-06 14:22:36	0:0:00:42	0:0:03:22	2019-03-06 14:31:22	US ANALY_MWT2_SL7 online no active blacklisting rules defined
	Job name: user.aforti.test.20190306141519/.4267387837 #1								
	Datasets: **Out:** user.aforti.test.20190306141519.log.235213854								

GPUs

- GPUs are becoming more popular
- Important to have GPUs available for the users to access to
  - Test algorithms
  - Test brokering options
  - Different GPU nodes setups
- With standalone images users can run code on GPU queues
  - Currently 4 sites with GPUs

cloud (1)	UK (20)
computingsite (2)	ANALY_MANC_GPU_TEST (10) ANALY_QMUL_GPU_TEST (10)
corecount (1)	1 (20)
eventservice (1)	ordinary (20)
gshare (1)	Analysis (20)

Setuid/user namespaces

- From CetnOS7.6 user names spaces can be enabled without tweaking the kernel
- You can enable them with sysctl
  - echo "user.max_user_namespaces = 15000" > /etc/sysctl.d/90-max_user_namespaces.conf
  - sysctl -p /etc/sysctl.d/90-max_user_namespaces.conf
- To use them in singularity you need to switch off setuid or remove the setuid binaries
  - If you remove setuid some funcrtionality may not work anymore
    - Version 3.x isn't working yet.

Overlay/underlay

- Both mechanisms to bind a directory that doesn't exist in the image.
  - We cannot guarantuee user images will have all the directories
  - Sites with caches will need them
- ATLAS needs one of the two enabled or both
- Overlay needs setuid enabled to work
- Underlay works also in user namspaces

Tests

2.6.1 EPEL	CentOS 7.6.1810	3.10.0-957.1.3	No (conf file)	Yes	Underlay	singularity --debug exec --nv docker://lukasheinrich/atlasml:latest python /btagging/DL1_c_vs_b_slim.py trainingfile.h5 10 gpu 50000	docker	OK	
2.6.1 EPEL	CentOS 7.6.1810	3.10.0-957.1.3	No (conf file)	Yes	Underlay	singularity --verbose exec -C -B $PWD:/data --nv docker://lukasheinrich/atlasml:latest python /btagging/DL1_c_vs_b_slim.py /data/trainingfile.h5 10 gpu 50000	docker	OK	
3.0.2 OSG	CentOS 7.6.1810	3.10.0-957.1.3	No (no setuid binaries)	Yes	Underlay	/cvmfs/oasis.opensciencegrid.org/mis/singularity/3.0.2/bin/singularity exec -C -B $PWD:/data docker://alpine cat /data/pluto	docker	Failed	FATAL: container creation failed: mount error: can't mount image /proc/self/fd/9: failed to find loop device: could not attach image file too loop device: permission denied
3.0.2 OSG	CentOS 7.6.1810	3.10.0-957.1.3	No (no setuid binaries)	Yes	Underlay	singularity exec -C -B $PWD:/data /cvmfs/atlas.cern.ch/repo/containers/fs/singularity/x86_64-centos6 cat /data/pluto	unpacked	Failed	FATAL: container creation failed: mount error: can't mount devpts filesystem to /cvmfs/oasis.opensciencegrid.org/mis/singularity/x86_64/3.0.2/var/singularity/mnt/session/dev/pts: invalid argument
3.0.3rc2 tarball rpm	CentOS 7.6.1810	3.10.0-957.1.3	Yes	Yes	Overlay	singularity exec -C -B $PWD:/data docker://alpine cat /data/pluto	docker	OK	
3.0.3rc2 tarball rpm	CentOS 7.6.1810	3.10.0-957.1.3	No (conf file)	Yes	Underlay	/cvmfs/oasis.opensciencegrid.org/mis/singularity/3.0.2/bin/singularity exec -C -B $PWD:/data docker://alpine cat /data/pluto	docker	Failed	FATAL: container creation failed: mount error: can't mount image /proc/self/fd/9: failed to find loop device: could not attach image file too loop device: permission denied

- Example of tests with different setups
  - Spreadsheet results

Traceability and Isolation

- Containers don't offer any on site traceability
  - It will still require to query ATLAS services
- Isolation requirements
  - Payloads have to be isolated from the pilot environment
  - Payloads have to be isolated from each other
- Requirements are for user payloads only
- There is a draft WLCG policy document which is now open for comment
  - Traceability and Isolation policy document (draft)

Benchmarking

- WLCG/Hepix benchmarking working group is working on replacing HS06 with custom WLCG benchmark built with experiments workloads using docker containers
- Another case of standalone containers
  - Built to be used by sys admins, vendors, or run in jobs
- Working to make these containers as small as possible
  - Built around a combination of options that will not change
- WLCG/Hepix Benchmark Twiki
- WLCG/Hepix Benchmark JIRA

Conclusions

- Containers are being enabled
- On CentOS7 WNs singularity is mandatory
  - No HC test yet, but we can run standalone container to test minimal functionality
  - Install 2.6.1, 3.x will be reviewed when opened issues solved
- If you have GPUs that you can put online let us know we will add them to the current tests
- Containers Deployment twiki