Fermilab U.S. DEPARTMENT OF ENERGY | Office of Science

Challenges of Containerization in the distributed computing environment

Marco Mambelli <marcom@fnal.gov>

HTCondor Week 2018

May 23, 2018

Especially for big scientific collaborations, the following is often true

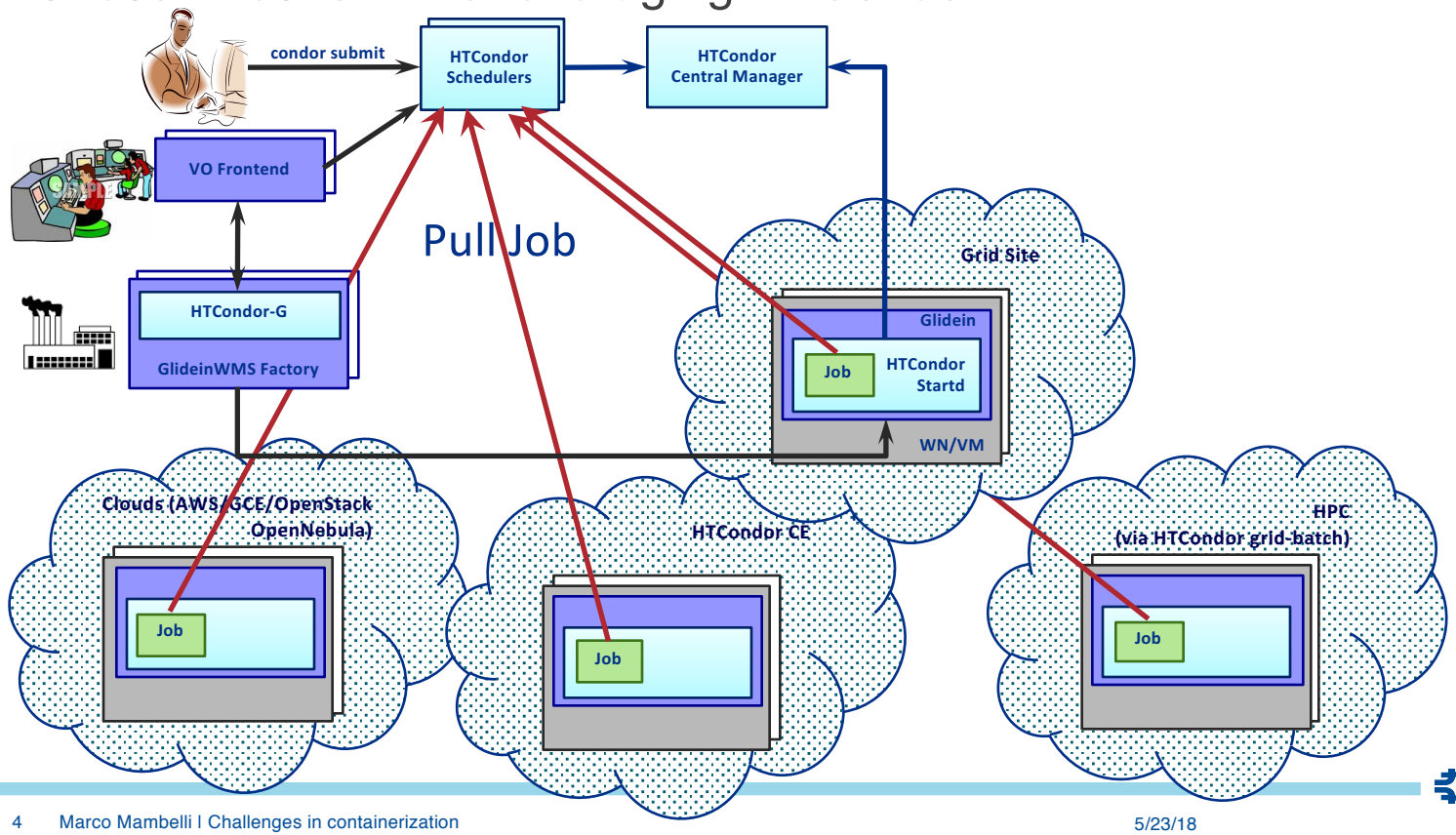
- Technologies change over the lifetime of the experiment
- Use of shared resources
- Owner and user of resources are different

Conflicting goals

- Performance
- Efficiency
- Reproducibility
- Mobility/Portability
- Reliability
- Uniformity
- Isolation
- Protection
- Traceability

GlideinWMS works in a diverse environment

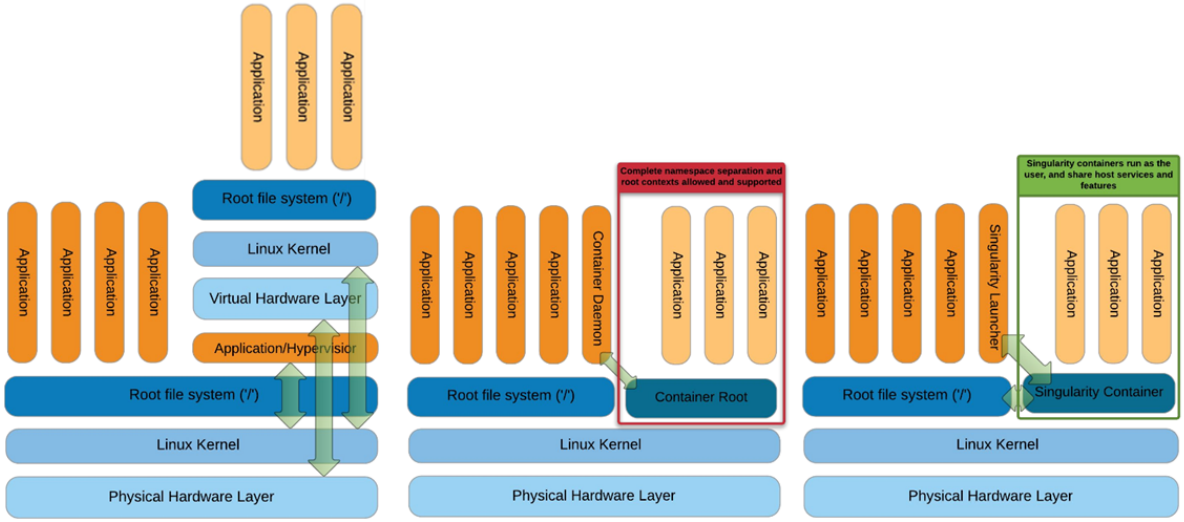
- Pilot based system (Glideins)
- Provisions resources for HTC leveraging HTCondor



Fermilab

Some solutions

- Virtual Machines (Clouds, AWS, GCE, Azure)
- Docker (microservices)
- Shifter
- Singularity
- gLExec



Greg Kurtzer keynote at HPC Advisory Council 2017 @ Stanford

Fermilab

Singularity

- Singularity runs a container as a system user
- Allows to select the namespace isolation (mnt, pid, ipc, net, or user)
- Allows to run a different OS
- Is accepted on HPC resources and on more clusters than Docker
- Can run as unprivileged user on RHEL 7.4 and above, by enabling unprivileged user namespaces (there is a version distributed in OASIS)

Singularity in GlideinWMS

- Resources define if Singularity is required, optional or not available. If Singularity is available, must be configured in the path or via module
- Job submitters define if Singularity is required, optional or not desired. Optionally they can specify an image in the job
- A default image can be chosen by the VO or the resource via configuration or via a script. OSG is providing EL6 and EL7 Singularity images in OASIS
- The Glidein runs tests outside and inside Singularity to reduce the possibility of jobs failures
- Jobs run in a separate session to allow secure CVMFS

- A VO can run using Singularity with minimal extra effort

Some open issues

- Mechanism to select the platform: most VOs are used to request for a specific OS and Singularity allows to morph
  - CMS allow to choose between RHEL6 and RHEL7
  - OSG pre-allocates between RHEL6 and RHEL7 with assigned probability
  - Should there be a vetted image repository and a mechanism to match all possible platforms and select the desired one?
  - Should tests in the Glidein be on all the images?
- What tests can we perform during the Glidein initialization and lifetime to ensure that jobs matching a Glidein will successfully run on it?

Some open issues (cont)

- How can we do exhaustive tests on all resources?
  - E.g. Recently we enabled Linux sessions and processes groups to allow secure CVMFS. In condor manual: “This can cause problems when HTCondor is run under another job execution system”. Are Linux sessions actually interfering with Local Resource Managers?

Singularity in HTCondor

- Can start a job in Singularity job

```
# Only set if singularity is not in $PATH.
#SINGULARITY = /opt/singularity/bin/singularity
# Forces _all_ jobs to run inside singularity.
SINGULARITY_JOB = true
# Forces all jobs to use the CernVM-based image.
SINGULARITY_IMAGE_EXPR = "/cvmfs/cernvm-prod.cern.ch/cvm3"
# Maps $_CONDOR_SCRATCH_DIR on the host to /srv inside the image.
SINGULARITY_TARGET_DIR = /srv
# Writable scratch directories inside the image. Auto-deleted after the job exits.
MOUNT_UNDER_SCRATCH = /tmp, /var/tmp
```

- Advertises Singularity's availability

```
HasSingularity = true
StarterAbilityList = "HasFileTransfer,HasTDP,HasSingularity, …"
```

- Using expressions instead of set values, you can use job attributes to control whether to use singularity or not and to select the image to use

Singularity in HTCondor – desiderata…

- More flexibility in the invocation options. Singularity is still in very active development and may change quickly
  - OSG currently uses --ipc --pid --contain
- Ease support for GPUs
- Allow setup/teardown scripts outside and inside Singularity
  - Job Hooks, PreCmd and PostCmd
- Running tests also in Singularity via HTCondor hooks
  - Benchmark jobs and startd cron jobs