(Container)
Experience at RAL

Andrew Lahiff

GDB, 12th April 2017

Introduction

- Since we migrated from Torque+Maui to HTCondor in 2013, Linux kernel functionality has improved our ability to isolate jobs
  - cgroups for resource limits & monitoring, ensuring processes can’t escape the batch system
    - CPU, memory, ...
  - PID namespaces
    - processes in a job can’t see any other processes on the host
  - mount namespaces
    - /tmp, /var/tmp inside each job is unique
- This has mostly worked very well
- Still have a (big) limitation: all jobs share the same root filesystem as the host

HTCondor Docker universe

- Docker universe
  - Introduced in HTCondor 8.3.6 in June 2015
  - Successfully ran LHC jobs at RAL in 2015
    - jobs in SL6 containers on SL7 worker nodes
  - Lots of bug fixes & improvements made since then
  - Nebraska Tier 2 migrated fully to Docker universe in summer 2016
- Running jobs in containers on worker nodes now much more important for us
  - Migrating to Ceph-based storage system
  - Planning to run xrootd gateways on every worker node
    - this requires we use SL7 worker nodes as soon as possible

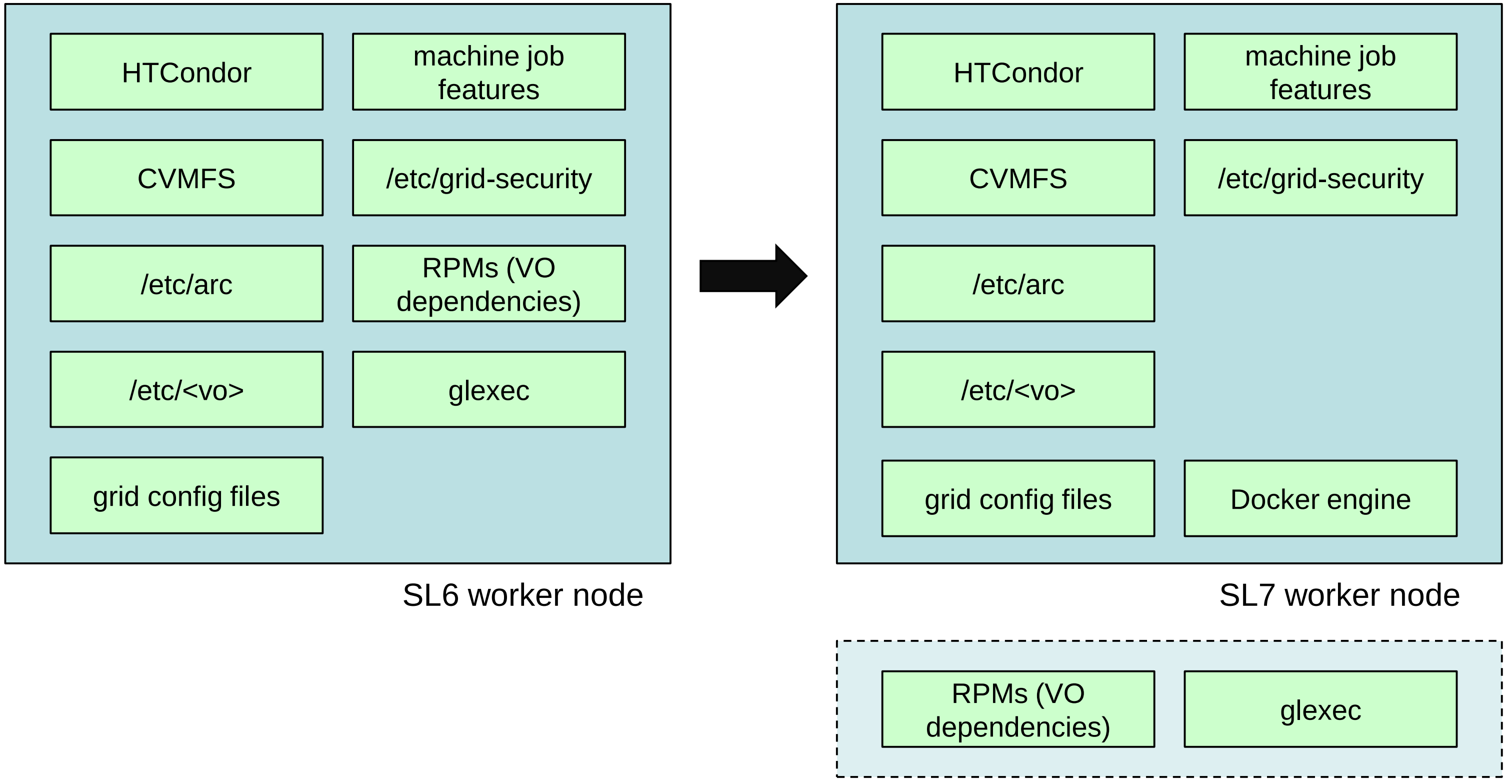
Worker nodes

- Versions in use at RAL
  - HTCondor 8.6.1
  - Docker 17.03 (CE)
- Storage driver
  - OverlayFS
  - XFS filesystem formatted correctly to be used as an overlay
- CVMFS
  - Installed on host, using autofs as usual
  - Bind mount /cvmfs into containers using shared mount propagation (only in recent versions of Docker)

    ```
    -v /cvmfs:/cvmfs:shared
    ```
  - So far this has been working well (only issue was CVM-1200)

Worker nodes

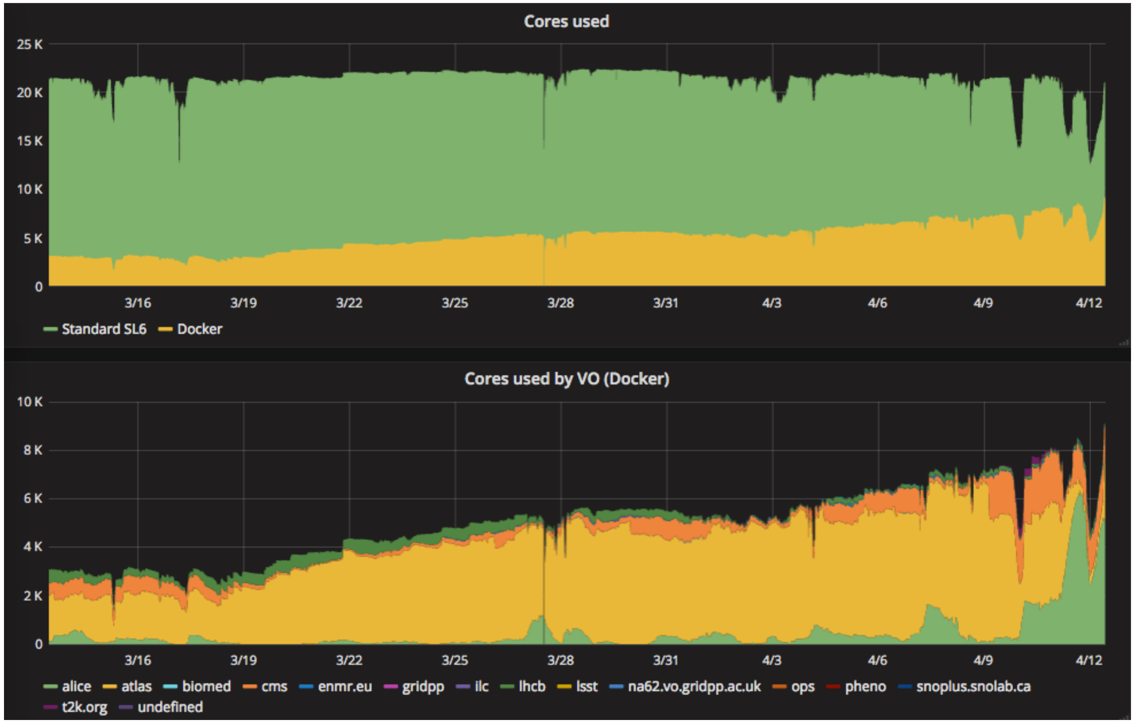
- As a first step, move to SL7 but with as few changes as possible
  - many things (CVMFS, config) bind-mounted into the containers



CentOS 6 image

Current status

- ~40% batch farm has been migrated to SL7
- Have run jobs from many VOs
  - ALICE, ATLAS, CMS, LHCb
  - biomed, eNMR, ILC, LSST, NA62, SNO+, T2K, ...



Plans (short/mid-term)

- Complete migration of the batch farm to Docker
- Start adding xrootd gateways to worker nodes
- Provide access to RHEL7 environments via CEs
  - easy for ATLAS & CMS (several options)
  - need to work out the best way to do this for DIRAC-based VOs
- Make Singularity available within RHEL7 environments
  - allow CMS to migrate from glexec to Singularity
  - useful for other experiments, e.g. ATLAS
- Get rid of pool accounts

Other activities

- Container cluster managers
  - Using Mesos as a platform for multiple computing activities & running services
  - Using Kubernetes as an abstraction across on-premises resources & multiple public clouds
    - deal with a single API rather than many different APIs
    - have run
      - CMS (CRAB3) jobs at RAL, Azure, GKE, AWS
      - LHCb jobs at RAL
      - ATLAS jobs at RAL, Azure
    - there will be an update at the Spring HEPiX

Thanks to Google & Amazon for credits & to Microsoft for an Azure Research Award

Any questions?