Cloud pre-GDB Introduction

Michel Jouvin
LAL, Orsay
jouvin@lal.in2p3.fr

GDB, July 2013

WG Goals

- WG is about exploring the possibility to use private/community clouds as a replacement for grid CEs
  - No intention to cover all the aspects of cloud usage by experiments
  - Focus on shared clouds rather than dedicated resources to one experiment
- Build on existing work in experiments: do not start a new huge R&D project
  - No manpower available to do it

Pre-GDB Goals

- Review progress/work about cloud usage in each experiment
  - Most of the work done with HLT farms… which are not shared clouds
  - An important milestone for integrating cloud backends in pilot factories
  - Not a review of cloud solutions/deployments at sites: may be a topic for another meeting
    - Also a lot of this has been covered already in HEPiX
- Review concrete work in the last months about implementing graceful VM termination as discussed in March
- Continue discussions on possible models for non static sharing of resources
  - "Fairshare-like approach" vs. economic (credit-based) models

Pre-GDB Attendance

- Well attended : ~40 people
  - Including remote North America participation
- 3 experiments present
  - ATLAS
  - CMS
  - LHCb
- Indico and presentations: https://indico.cern.ch/conferenceDisplay.py?confId=253858

Cloud Work @Experiments…

- Strong activities in all 3 experiments involving different parties
- Private clouds set up by ATLAS and CMS on HLT farms
  - Sandboxing of reconstruction/offline activities to prevent interferences with Trigger/DAQ operations
  - HLT clouds will remain in use after LS1, during data taking: dynamic sharing of resources between DAQ and cloud
    - LHCb: different strategy (deferred trigger) to ensure a 100% occupancy of HLT farm during data taking: no room for offline activities
  - Testbed for testing/integrating cloud infrastructures (OpenStack) at scale
    - (~6K jobs sustained for CMS, 12/16K jobs for ATLAS)

… Cloud Work @Experiments

- Lot of other private/community clouds used by all 3 experiments for daily production work
  - ATLAS: 15 clouds used (OpenStack, Nimbus)
    - Several sites providing only a cloud for compute resources
  - Cloud MW zoology is not a problem: open-source abstraction layers are available to provide a unified interface
    - LHCb: libcloud
    - CMS: DeltaCloud (integrated with Condor), will look at libcloud
- Clouds are seamlessly integrated into SW infrastructures and pilot factories
  - Same pilot infrastructure is used for clouds
  - Cloud is just another backend for the pilot factory: already supporting several anyway…

VM Image Life Cycle

- The critical part for sustainability
  - Build once, run everywhere: images must be agnostic to the infrastructure
  - CernVM is the base image: only a few additions by experiments
- CVMFS is a key component
  - Allow to reduce experiment specific SW installed in the image thus reducing the need to update the image
  - Positive impact on image size and thus instantiation time
  - Consistent with standard SW distribution
- User contextualization used/tested by every experiment
  - Mostly amiconfig (already integrated into CernVM) but works ongoing to move to CloudInit (the now standard solution)
  - Used mainly to pass nformation required to contact the pilot factory

VM Graceful Termination…

- On the wish list of every experiment…
  - LHCb is the most advanced in experimenting it (Vac, CERN)
- Based on the machinefeatures proposal already discussed in the past
  - In particular at the last GDB to streamline multi-core support
  - An Ops Coordination TF has been created to foster deployment by sites and testing by experiments
    - Led by Stefen Roiser
- Basic features
  - VM Life Time (shutdown time) is advertised to the job with a guaranteed advance notice
    - Can be updated by the site
  - A shutdown command is made available to the “user”
    - VM is killed if not shutdown before the termination time

… VM Graceful Termination

- Allow a VO to take advantage of the remaining time by scheduling an appropriate job
  - LHCb “job masonry” (~ back filling in batch systems)
  - LHCb: done by scheduling “flexible” MC
    - Number of events to simulate adjusted to the remaining time
  - Raise again the ability to advertize with enough precision the HS06 power of a (virtual) machine

- Allow a site to shutdown physical resources for scheduled operations without impacting the VO efficiency
  - Jobs will be completed before the shutdown rather than canceled
  - Resources can be used by the VO until the last moment, something difficult with usual downtimes
  - Sites ready to make it possible as it increases their credit
    - Seen as more dependable by the VO

Fair Use of Shared Clouds…

- The critical and difficult point with community clouds: no magic solution (yet!) but (quasi)-consensus on the problem
  - Resources are finite: cannot achieve full elasticity as in commercial clouds
    - The bigger the cloud, the easiest..
  - No site queuing: no pending requests that the site can arbitrate
    - Wide agreement that re-introducing queuing in clouds is not feasible/desirable
- Lively discussions on the economic (credit-based) approach: (virtual) coast advertised by sites used at the basis for decision by the VO
  - No implementation existing yet
  - Workflow pretty complex: not that far from brokering ideas that failed in WMS
  - Will have to rely on (accurate) information published by sites

… Fair Use of Shared Clouds

- Most people thinks desirable to find a simpler approach/solution
  - Avoid any negotiation: let site decide which VO has priority in accessing the resource based on some history (accounting?)
  - Basically 2 possible approaches
    - New request arbitration: for some time, refuse request from VO over quota to give a chance to other V0s to access the resources. Will result in free slots during a period of time.
    - Sites start VMs for VO under quota on behalf of the VO: requires an "instantiation" service at site acting a cloud client in IaaS clouds. Can be non trivial (authentication to the cloud in particular).
- Still work in progress but better understanding of the issue
  - LHCb planning to experiment the economic model
  - Vac demonstrating the second approach

Vac(uum): the IaaC Model

- (virtual) Machines created "spontaneously" by sites
  - Desktop grid/BOINC model
  - VM images configured to connect to a central queue operated by the VO (pilot factory, eg. DIRAC, Panda)
  - No way for a user to request instantiation of a VM: the big difference with IaaS
    - Less pervasive than IaaS, optimized for pilot factories
  - Same VM images used as in IaaS clouds
  - Deployed at 4 UK sites: main one is Manchester
- VM graceful termination implemented
  - Used successfully by LHCb
- Makes fair usage pretty easy to implement
  - Target shares influence decision on next VM to start
  - History could be based on accounting analysis