The Diverse use of Clouds by CMS

2015 J. Phys.: Conf. Ser. 664 022012

(http://iopscience.iop.org/1742-6596/664/2/022012)

Journal of Physics: Conference Series **664** (2015) 022012 doi:10.1088/1742-6596/664/2/022012

The Diverse use of Clouds by CMS

Anastasios Andronis[1,3],Daniela Bauer[2], Olivier Chaze[3], David Colling[2*], Marc Dobson[3], Simon Fayer[2], Maria Girone[3], Claudio Grandi[4], Adam Huffman[2], Dirk Hufnagel[5], Farrukh Aftab Khan[6], Andrew Lahiff[7], Alison McCrae[3], Duncan Rand[2], Massimo Sgaravatto[8], Anthony Tiradani[5], Xiaomei Zhang[9] on behalf of the CMS collaboration.

[1] University of Ioannina, Greece
[2] Blackett Laboratory, Imperial College London, London, SW7 2AZ, UK
[3] CERN, CH-1211 Geneva 23, Switzerland
[4] INFN-Bologna, Viale Berti-Pichat 6/2, 40127 Bologna, Italy
[5] Fermilab, Batavia, IL 60510-5011, USA
[6] National Centre for Physics, Islamabad 45320, Pakistan
[7] Rutherford Appleton Laboratory, Harwell Oxford, Didcot, OX11 0QX, UK
[8] INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
[9] Institute of High Energy Physics, Chinese Academy of Sciences, 19B Yuquan Lu, Shijingshan District, Beijing, 100049, China

E-mail: *d.colling@imperial.ac.uk

Abstract. The resources CMS is using are increasingly being offered as clouds. In Run 2 of the LHC the majority of CMS CERN resources, both in Meyrin and at the Wigner Computing Centre, will be presented as cloud resources on which CMS will have to build its own infrastructure. This infrastructure will need to run all of the CMS workflows including: Tier 0, production and user analysis. In addition, the CMS High Level Trigger will provide a compute resource comparable in scale to the total offered by the CMS Tier 1 sites, when it is not running as part of the trigger system. During these periods a cloud infrastructure will be overlaid on this resource, making it accessible for general CMS use. Finally, CMS is starting to utilise cloud resources being offered by individual institutes and is gaining experience to facilitate the use of opportunistically available cloud resources.

We present a snap shot of this infrastructure and its operation at the time of the CHEP2015 conference.

1. Introduction

The flexibility, agility and reliability offered by the virtualisation of compute resources and the management of the Virtual Machines (VMs) by automated Cloud systems is becoming increasingly attractive to some of the sites that provide resources to CMS. In many instances the fact that a site is running its infrastructure as a Cloud is hidden from CMS as the site will continue of offer those resources to CMS via a traditional Grid interface, however there are an increasing number of sites that are exposing Cloud interfaces directly to CMS. In order to make use of these resources CMS has had to adapt its submission infrastructure and to understand how the operation of these resources differs from traditional Grid sites. In this conference report we describe three different scenarios in which CMS is using a Cloud infrastructure. These are:

- The modification to the CMS submission infrastructure and its use on the Cloud at CERN
- The use of the CMS High Level Trigger (HLT) farm as a Cloud
- The use of Cloud resources at CMS institutions.

Several of the Cloud systems with which CMS interacts rely on OpenStack [1], which is an open source Cloud software system.

2. The CMS Job Submission Infrastructure and use of the Cloud at CERN

2.1. The CMS job submission infrastructure with Grids and Clouds

The CMS job submission infrastructure is a pilot job based system called *glideinWMS* [2]. The glideinWMS is built upon several of the components of the HTCondor [3] workload management system. When submitting to a traditional Grid site, the glideinWMS submits pilot jobs to the Compute Element (CE) of the site. When each pilot job runs it starts an HTCondor *startd* daemon which allows it to join a distributed HTCondor pool. This node is then able to accept work from that pool. Authentication is performed using GSI and context switching is performed at runtime using glexec [4]. This is described in more detail in Ref. [5].

The glideinWMS was modified as early as 2011 to be able to submit to Amazon EC2 Clouds [6] as well as traditional Grid sites. A pre-built image is stored at the Cloud site. The glideinWMS is then able to use the EC2 interface to request that a VM be built from this image and to contextualise the VM once built. The VM then starts a HTCondor *startd* and so is able to join the distributed HTCondor pool.

2.2. The CERN Agile Infrastructure and the CMS Tier 0

CERN is in the process of moving all of its compute resources to run as VMs in an OpenStack Cloud Agile Infrastructure (AI). During Run 2 of the LHC, CMS will use the majority of its CERN resources as VMs within this infrastructure. Most significantly, all of the CMS Tier 0 resources running at CERN[1] will be run on this AI [7]. These resources will effectively be a large virtual cluster administrated by CMS. While this does bring some overhead to CMS it also gives CMS the flexibility to configure the resources as best suits CMS.

At the start of Run 2 the Tier 0 will consist of approximately 9000 cores and CMS has been carrying out considerable testing of the Tier 0 running on the CERN AI to ensure that it will run at the scale needed (see Figure 1). After considerable investigation the Tier 0 resources at CERN will be run on VMs with 8 cores and 16GB of memory. The image used for these VMs is built by CMS using the CERN IT automatic build system. During these tests all aspects of the CMS Tier 0 workflows were tested. The most CPU intensive activity performed at the Tier 0 is prompt reconstruction which is carried out using 4 core jobs. The Tier 0 activity is, by its very nature, bursty. When there are data to reconstruct the Tier 0 must reconstruct them with a high priority, but when there is no data to reconstruct the resources are not needed. However, because of the rate at which the CERN AI was able to instantiate VMs and the charging model used by CERN IT, it was also decided that the most efficient way of running these resources was through running long lived (currently 1 month) VMs. This means that other jobs use Tier 0 resources when they are not needed for Tier 0 workflows, however they are returned when needed by the Tier 0. This prioritisation is managed by setting priorities in the HTCondor pool to which the Tier 0 resources belong. This is known to work at the scale needed (see the example in Ref. [5]).

The Tier 0 resources will be split between the CERN computer centre at Meyrin and the Wigner computing centre. The data being reconstructed by the Tier 0 will also be split between

[1] CMS is moving to a model where some of the Tier 0 workflows run at CERN but some also run at the Tier 1 Centres [7]

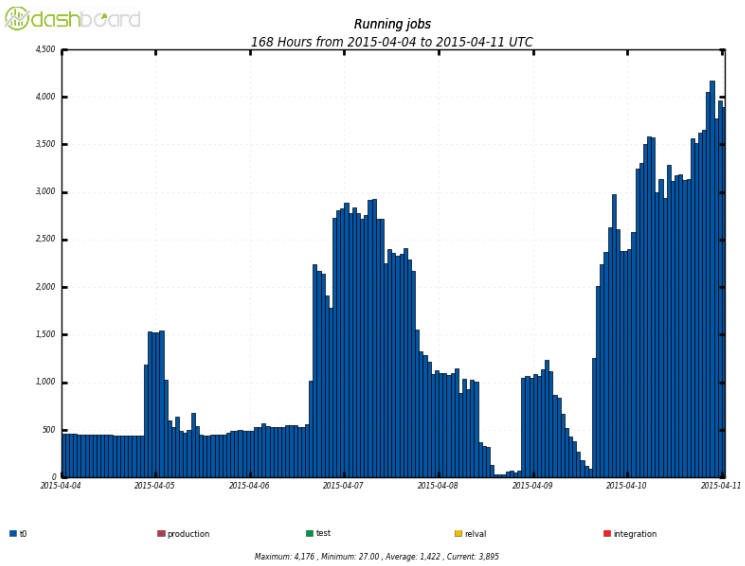

Figure 1. Number of jobs running during the testing of the Tier 0 running on the CERN AI immediately prior to the CHEP 2015 conference. These are predominately 4 core reconstruction jobs but with some single core jobs running different parts of the Tier 0 workflows. These tests used all 9000 cores of the CMS Tier 0 resources at CERN.

the two sites and so the reconstruction jobs run at the Tier 0 will be reconstructing data stored both on local disk resources and remote disk resources. This is only important for jobs that require low network latency such as the jobs that merge the output files. These merging jobs require very little CPU and so the glideinWMS has been patched to allow them to run alongside reconstruction jobs (in testing this has effectively meant a 4 core job running both a 4 core reconstruction job and a merging job) in a "lazy download" mode.

3. Using the CMS High Level Trigger as a Cloud Resource

The CMS High Level Trigger (HLT) is a considerable compute resource. At the start of Run 2 the HLT will consist of approximately 15000 cores and approximately 490kHEPSpec06, which is larger than the total 2015 Tier 1 CMS resource request of 300kHEPSpec06. This resource is completely owned and managed by CMS. The HLT has no local mass disk storage and is currently connected to CERN IT via a 60Gb/s network connection. During the first long shutdown of the LHC (LS1) CMS overlaid an OpenStack infrastructure over the HLT hardware and used it as a general Cloud resource when it was needed and when it was not being used for development of the HLT system itself [8]. A great deal of experience was gained during this period and a number of minor but temporarily "show-stopping" problems were overcome and eventually this was seen as a great success. The number of running jobs in the overlaid Cloud infrastructure during LS1 can be seen in Figure 2. During this time the VMs and the jobs were managed by the glideinWMS exactly as described above. Data needed by the running jobs was read from the EOS disk system at CERN and any data produced was written back out to EOS.

3.1. Plans for running the HLT as a Cloud resource During Run 2

Following the success of operating the HLT as a Cloud resource during LS1 it was decided to try to use it during Run 2. However, this parasitic use of the HLT must never interfere with its primary function as part of the CMS trigger system. LHC technical stops are planned and often

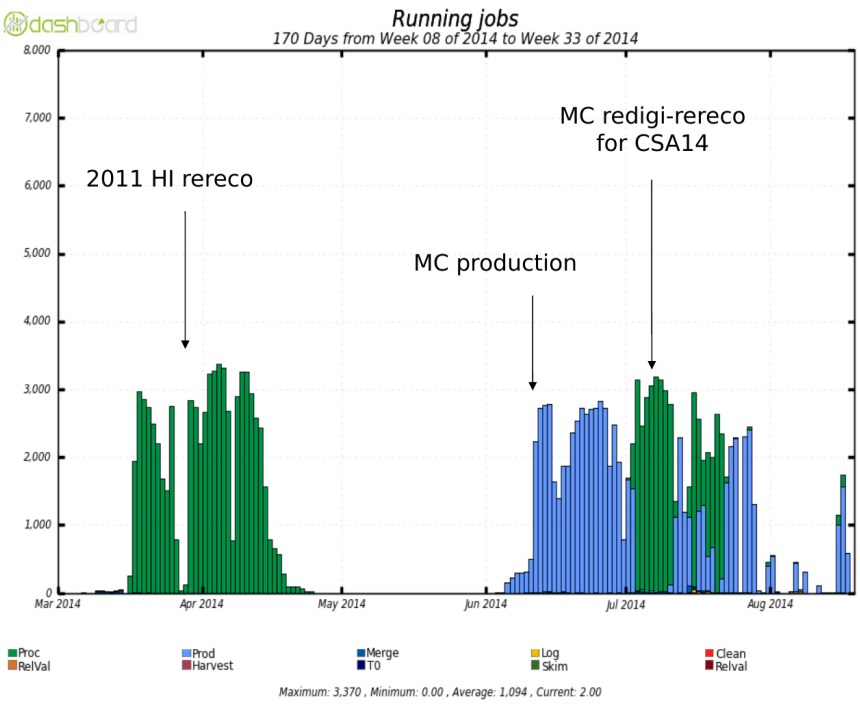

Figure 2. The number of jobs running in the Cloud overlaying the HLT during LS1. The labels refer to the Heavy Ion (HI) data reprocessing, the production of some high priority Monte Carlo samples and participation in the CSA14 exercise.

last a week or even longer so these are obvious times in which to operate the HLT as a Cloud. If Run 2 follows a similar pattern to the 2012 running of the LHC this would be approximately 1 week in every 10.

During normal LHC operation in 2012 the average duration of stable beams was about 6.8 hours and the average duration between stable beams (but excluding the planned technical stops) was about 9.3 hours. The distributions of durations can be seen in Figure 3 & 4.

These distributions suggest that there could be considerable benefit in using the HLT as a Cloud between files. Post fill operations mean that it is likely the HLT may only be available for approximately 6 of the 9 hours between fills, however this would increase the usable period to about 50% of the total available time. In this mode jobs would be killed and VMs shut down as stable beam approached. Equally, to utilise this time efficiently, it must be possible to start VMs quickly and start jobs running as soon as the resources become available. This requires the pre-caching of images on the HLT nodes. The ability to start up and shut down VMs has been extensively tested and Figures 5 and 6 show results from just two of these tests.

Running in this mode can, potentially, be very wasteful as work that has been carried out by a job that is then killed is lost. If the period between stable beams is modelled by a falling exponential with a mean of 6 hours, the fraction of time used and the fraction of time wasted can be simulated and the results can be seen in Figure 7. CMS will initially use 2 hour jobs when running in this mode, but this will be modified in light of the experience that we gain.

The requirement that the parasitic Cloud use of the HLT *never* interferes with its role in the trigger system means if we are to run between LHC fills it must be the online team who have control for starting and stopping the VMs rather than the glideinWMS. A new tool, called

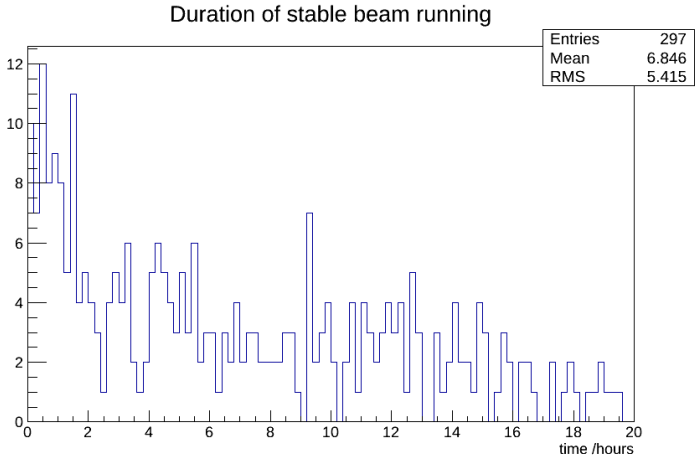

Figure 3. The duration of periods of stable beams during the 2012 LHC running.

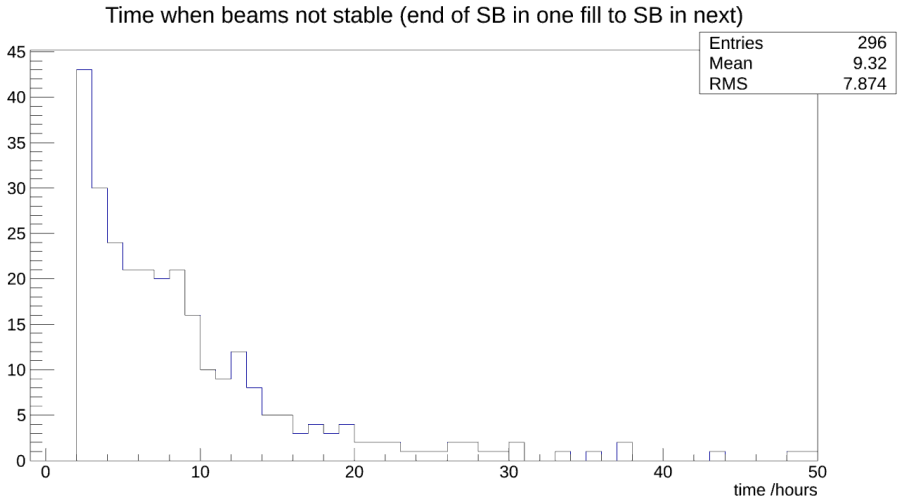

Figure 4. The duration of periods between stable beams during the 2012 LHC running.

Cloud-igniter, has been developed and deployed in order to do this.

Even during LHC fills the load on the HLT is not constant. Figure 8 shows the luminosity during the 2013 Heavy Ion running. As the instantaneous luminosity drops it is hoped that the online team will be able to use Cloud-igniter to start up some fraction of the HLT as a Cloud resource.

This means that CMS has a three stage plan for overlaying a Cloud infrastructure on the HLT:

- **Stage 1.** Run during the technical stops in the LHC programme. If that works ...
- **Stage 2.** Run during the periods between LHC fills. If that works ...
- **Stage 3.** Start to run as the load on the HLT reduces at the end of the fill.

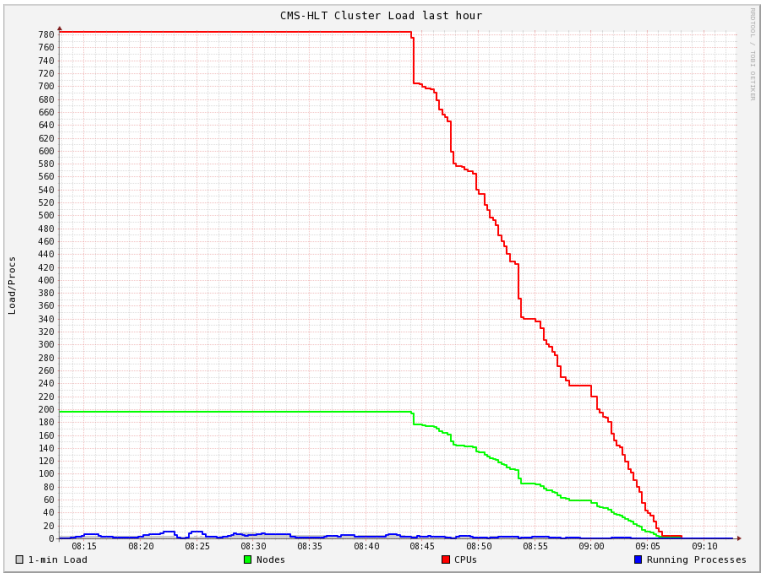

Figure 5. The load on the HLT during a test to shut down 700 running VMs. All VMs are shut down within 20 minutes.

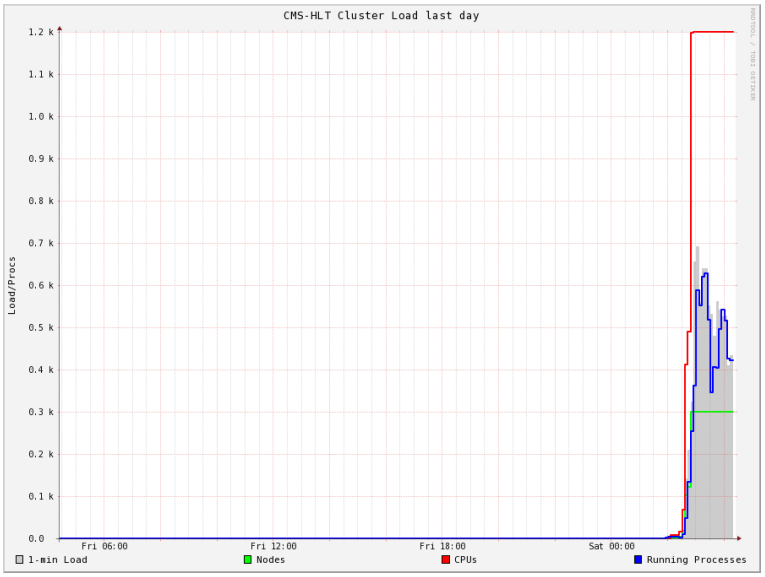

Figure 6. The load on the HLT during a test to start 1200 VMs on the HLT. All VMs are running jobs with 10 minutes.

4. The use of Cloud resources at CMS institutions.

The main focus of the cloud activities within CMS has been at CERN (both on the AI and on the CMS HLT), however in parallel to this, if at a much lower level, CMS has also been investigating the use of Clouds at individual institutions in China, Italy and the UK. In these tests user analysis jobs have been run at all sites using glideinWMSes hosted in Italy and the UK. These have run successfully. This activity will become more central to our objectives once Run 2 has started and the Tier 0 and HLT Cloud have been proven to work. We are currently

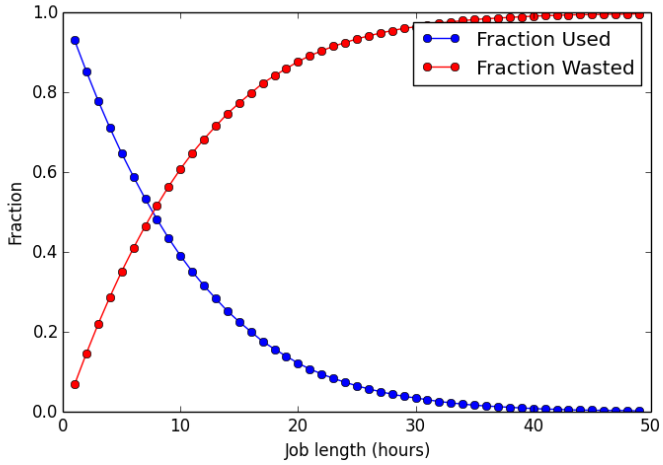

Figure 7. The fraction of resource used and the fraction of resource wasted as a function of job length assuming an average available time of 6 hours.

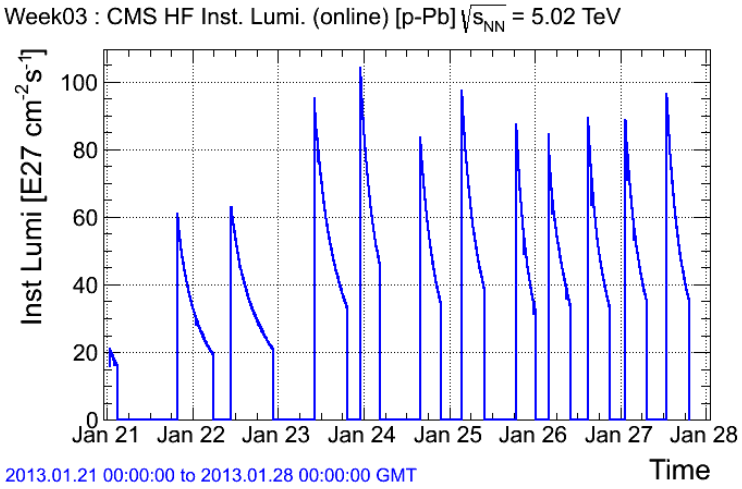

Figure 8. Instantaneous luminosity during a period of the 2013 Heavy Ion running.

stetting a regular testing framework for these sites with a view to them becoming part of the Global CMS HTCondor pool in the near future.

While most of the testing of institution based Cloud resources have been using the glideinWMS we have also tested more novel approaches (for CMS) such using VAC and VCycle [9].

CMS is also investigating the possibility of bursting onto commercial Clouds.

5. Conclusions

CMS is increasingly making direct use of Cloud infrastructures. The main focus of our activity over the past 2 years has been on the resources of the CERN AI and the CMS HLT. Both of these are vital to the success of CMS computing. However, as these are shown to work well during Run 2 of the LHC our activity we will concentrate more on the Cloud resources offered

by individual institutaion and even on commercial providers.

References

[1] OpenStack 2015 Openstack–open source cloud computing software ONLINE: Available at `http://www.openstack.org/` [Accessed: 2010-05-16]
[2] Sfiligoi I, Bradley D C, Holzman B, Mhashilkar P, Padhi S and Wüerthwein F 2009 *Computer Science and Information Engineering, World Congress on* **2** 428–432
[3] Thain D, Tannenbaum T and Livny M 2005 *Concurrency - Practice and Experience* **17** 323–356
[4] Sfiligoi I, Bradley D, Miller Z, Holzman B, Wurthwein F *et al.* 2012 *J.Phys.Conf.Ser.* **396** 032101
[5] Letts J *et al* 2015 Using the glideinWMS system as a common resource provisioning layer in CMS. This Conference
[6] Andrews W, Bockelman B, Bradley D, Dost J, Evans D, Fisk I, Frey J, Holzman B, Livny M, Martin T *et al.* 2011 *Journal of Physics: Conference Series* vol 331 (IOP Publishing) p 062014
[7] Hufnagel D, Gutsche O, Contreras L and Tiradani A 2015 The CMS Tier-0 goes Cloud and Grid for LHC Run 2. This Conference
[8] Colling D, Huffman A, McCrae A, Lahiff A, Grandi C *et al.* 2014 *J.Phys.Conf.Ser.* **513** 032019
[9] McNab A, Love P and MacMahon E 2015 Managing virtual machines with vac and vcycle. This Conference