Design and performance of the virtualization platform for offline computing on the ATLAS TDAQ Farm

S Ballestrero[1], S M Batraneanu[2], F Brasolin[3], C Contescu[4, 5], A Di Girolamo[4], C J Lee[1, 4], M E Pozo Astigarraga[4], D A Scannicchio[2], M S Twomey[6] and A Zaytsev[7]

[1] University of Johannesburg, South Africa
[2] University of California, Irvine, USA
[3] Istituto Nazionale di Fisica Nucleare Sezione di Bologna, Italy
[4] CERN, Switzerland
[5] Polytechnic University of Bucharest, Romania
[6] University of Washington Department of Physics, USA
[7] Brookhaven National Laboratory (BNL), USA

E-mail: atlas-tdaq-sysadmins@cern.ch, alezayt@bnl.gov

Abstract. With the LHC collider at CERN currently going through the period of Long Shutdown 1 there is an opportunity to use the computing resources of the experiments' large trigger farms for other data processing activities. In the case of the ATLAS experiment, the TDAQ farm, consisting of more than 1500 compute nodes, is suitable for running Monte Carlo (MC) production jobs that are mostly CPU and not I/O bound. This contribution gives a thorough review of the design and deployment of a virtualized platform running on this computing resource and of its use to run large groups of CernVM based virtual machines operating as a single CERN-P1 WLCG site. This platform has been designed to guarantee the security and the usability of the ATLAS private network, and to minimize interference with TDAQ's usage of the farm. Openstack has been chosen to provide a cloud management layer. The experience gained in the last 3.5 months shows that the use of the TDAQ farm for the MC simulation contributes to the ATLAS data processing at the level of a large Tier-1 WLCG site, despite the opportunistic nature of the underlying computing resources being used.

Introduction

During the Long Shutdown 1 (LS1) of the LHC at CERN there is an exceptional opportunity to use the computing resources of the experiments' large trigger farms for other data processing activities. In the case of the ATLAS experiment [1], the TDAQ High Level Trigger (HLT) farm [2], consisting of more than 1500 compute nodes [3] deployed in the SDX1 area of LHC Point 1 (also being referred to as P1), is suitable for running Monte Carlo production jobs that are mostly CPU and not I/O bound.

Since the very beginning of this project, named Sim@P1, two major constraints shaped the design of the system: existing ATLAS production jobs needed to run unmodified, without additional software development effort; and the security and reliability of the ATLAS Detector Control System needed to

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

be preserved at all times. We thus chose to isolate the entire system in a virtualized platform. A successful construction and operation of such virtualized systems, for the purposes of the offline computing in HEP based on custom built virtualization platform control solutions, had already been demonstrated on a smaller scale [4].

Each HLT farm node is connected both to the Control and Data Network. Sim@P1 uses a dedicated subnet which is isolated from the rest of the ATLAS Technical Network (ATCN). This is implemented as a VLAN between the host and its rack DATA switch, and from these via dedicated 1 Gbps links to the Castor data storage switch, which connects it to the CERN General Purpose Network (GPN), with an optional bandwidth limitation.

The physical nodes run Scientific Linux CERN 6, and the HLT software runs on the bare metal OS, for lowest network latency and best CPU performance. The KVM hypervisor is then used to run the virtual machines (VMs), which are only allowed access to the dedicated VLAN interface. This allows for fast switching between the execution of HLT and of Sim@P1, without requiring a reboot of the physical node. Considering the scale of the HLT farm (over 27k HT CPU cores in the current configuration), we also decided to use a cloud management layer. We chose Openstack [5] which is widely adopted, also at CERN [6]. The CernVM project [7] was selected as a primary source of the base images for the virtual machines.

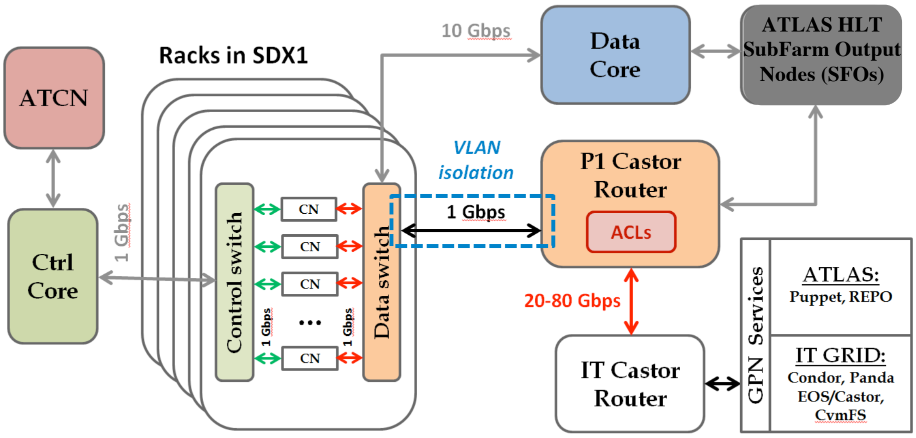


Figure 1. Networking layout of the Sim@P1 infrastructure deployed in LHC Point 1. VLAN isolation for the dedicated 1 Gbps uplinks available in every rack of compute nodes (CN) and the throughput limitation (20-80 Gbps depending on the mode of operation) on the Castor mass storage system uplink are highlighted.

1. Early stages of the project

A dedicated group consisting of members of the ATLAS TDAQ SysAdmins and NetAdmins teams, the CERN IT-SDC group and the RHIC and ATLAS Computing Facility at BNL was organized in order to conduct the design and implementation activities. Its early stages, which took place in February – April 2013, were devoted to the following activities:

- Identifying and addressing the design restrictions, such as the limited connectivity to GPN, the limited external network connectivity available for each rack and the limited amount of RAM on the new high density compute nodes (1 GB of RAM per single HT CPU core);
- Building a fully functional prototype in the TDAQ Test Laboratory and demonstrating, by using the HammerCloud [8] stress tests, its capabilities of running up to 0.9k ATLAS MC production jobs simultaneously with only about 0.05% of inefficiency;
- Defining the exact networking configuration, and in particular the security layout, for the full scale system;

- Defining the layout of servers needed to support the centralized Openstack infrastructure in the production environment and creating their profiles in the Puppet configuration management system [9].

During these early stages of the project a dedicated Grid site name CERN-P1 was registered and configured in order to serve as a representation of the Sim@P1 resources within the Worldwide LHC Computing Grid (WLCG) [10] and the distributed offline computing environment of the ATLAS experiment.

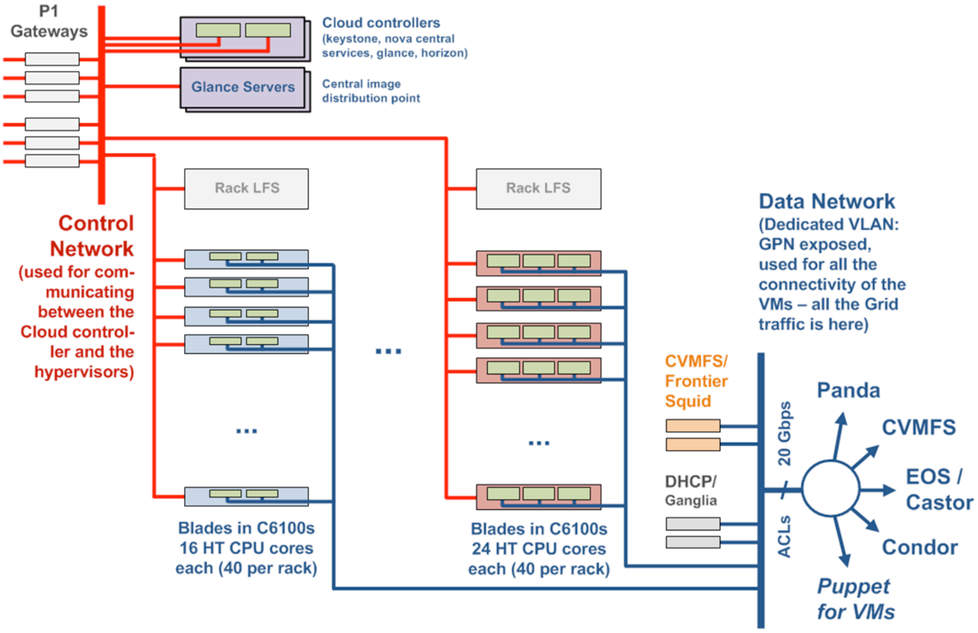

P1 Gateways
Cloud controllers
(keystone, nova central services, glance, horizon)
Glance Servers
Central image distribution point
Rack LFS
Rack LFS
Control Network
(used for communicating between the Cloud controller and the hypervisors)
Data Network
(Dedicated VLAN: GPN exposed, used for all the connectivity of the VMs – all the Grid traffic is here)
...
...
...
CVMFS/ Frontier Squid
DHCP/ Ganglia
20 Gbps
ACLs
Panda
CVMFS
EOS / Castor
Condor
Puppet for VMs
Blades in C6100s 16 HT CPU cores each (40 per rack)
Blades in C6100s 24 HT CPU cores each (40 per rack)

Figure 2. Openstack services deployment layout and their network connectivity. External services and storage systems required for the normal operations of the CERN-P1 site are also shown.

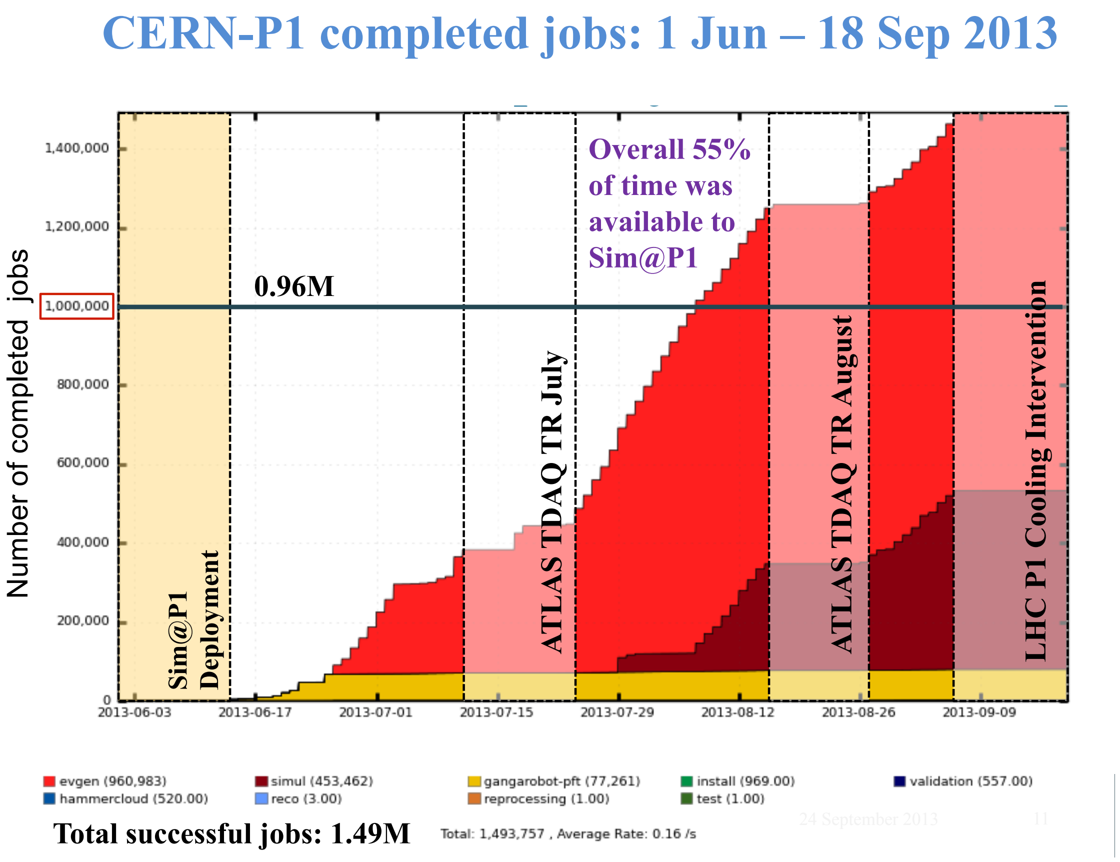
2. Full scale deployment and production runs

Once the design and prototyping phases of the project were finished, full-scale deployment was performed in May – July 2013:

- Installation and configuration of the VLAN, dedicated cabling, and upgrade of the Castor data storage uplink, according to the layout shown in figure 1;
- Deployment of the components of the centralized Openstack infrastructure (Folsom release) as shown in figure 2;
- Scaling up of operations to reach full utilization of all the resources allocated for Sim@P1 activities, finally providing the CERN-P1 site with more than 16.5k single CPU core job slots.

Overall, the deployment and scaling did not encounter major issues or fundamental limitations, and most effort went into tuning and debugging the many resources and services involved.

Following this period, three successful production runs were carried out until September 2013 with single core ATLAS MC production jobs; they are summarized in figure 3. The results of the quantitative analysis of the contribution to ATLAS MC production activities within the entire WLCG over the last 3.5 months are shown in figure 4.



CERN-P1 completed jobs: 1 Jun – 18 Sep 2013

Overall 55% of time was available to Sim@P1

Total successful jobs: 1.49M

Figure 3. Summary of the cumulative number of successfully completed jobs for the CERN-P1 site during the period of the first 3 production Sim@P1 runs.

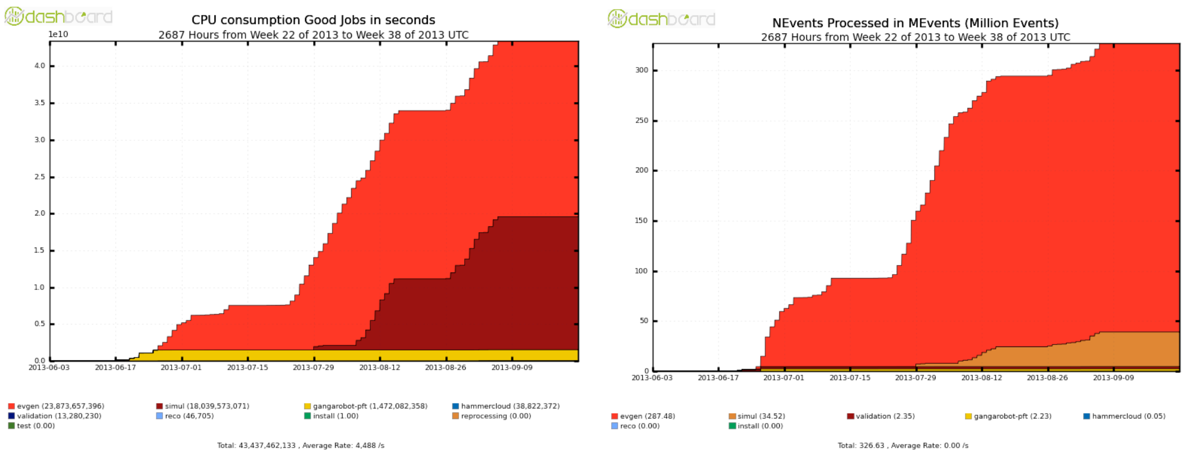


Figure 4. Sim@P1 contributed for the 7.4% of the ATLAS production in the period June-September 2013

Summary and future developments

The Sim@P1 project is dedicated to the design and deployment of a virtualized platform running on the ATLAS TDAQ computing resources. It is used to run large groups of CernVM based virtual machines operating as a single CERN-P1 WLCG site. This platform has been designed to guarantee the security and the usability of the ATLAS private network and to minimize interference with TDAQ usage of the farm; Openstack has been chosen to provide a cloud management layer. The project is a result of the combined effort of the ATLAS TDAQ SysAdmins and NetAdmins teams, CERN IT-SDC Group and RHIC and ATLAS Computing Facility at BNL. The setup phase required a total of approximately 6 person months, and support for the production require about 0.7 FTE.

During the period of February–September 2013 the project has gone through many stages, from prototyping up to the full scale deployment in the environment of ATLAS TDAQ HLT farm. The full scale deployment was finalized in June – July 2013 demonstrating operation of the virtualized infrastructure at the level of 17.1k single CPU cores exported to the CERN-P1 site by 2.1k virtual machines. It was also shown that it is feasible to operate such a group of VMs with a single instance of centralized Openstack services, with 45 minutes needed to switch from the TDAQ mode to Sim@P1, and about 10 minutes for switching back. Both operations are typically performed via Puppet without having to reboot the farm.

The first production operations for Sim@P1 project were carried out over the follow up period of July – September 2013, successfully completing more than 1.4 million ATLAS MC production jobs and delivering more than 1.4k CPU-years. This corresponds to 7.4% of the resources used for ATLAS MC production on all of the WLCG sites during the period of June 1 – September 18.

Overall, the Sim@P1 is capable of contributing to the ATLAS MC production on a level of computing power comparable to that of a large Tier-1 WLCG site, despite the opportunistic nature of the underlying resources, which have been available for approximately 55% of the time. Similar conditions should persist until the expected end of the LHC Long Shutdown 1 in the end of 2014. The possibility of operating the Sim@P1 infrastructure even beyond that date, during the periods of ATLAS data taking, is also being investigated. Finally, there is a potential for increasing beyond 20k the total number of HT-enabled CPU core count exported to the CERN-P1 site, with the same physical resources, by using the multi-core job slots on all the virtual machines deployed on the high density compute nodes instead of multiple single core job slots.

References

[1] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 (2008) S08003

[2] ATLAS Collaboration, ATLAS High Level Trigger, Data Acquisition and Controls: Technical Design Report, CERN/LHCC/2003-022 (2003), ISBN 92-9083-205-3

[3] C. Lee *et al.*, ATLAS TDAQ System Administration: an overview and evolution. Submitted to Proceedings of Science The International Symposium on Grids and Clouds (ISGC) 2013, March 17-22, 2013 Academia Sinica, Taipei, Taiwan

[4] A. Adakin *et al.*, Virtualized High Performance Computing Infrastructure of Novosibirsk Scientific Center. Physical Review Special Topics – Accelerators and Beams, Special Edition, proceedings of ICALEPCS 2011, Grenoble, France, 10-14 October 2011, pp.630-633. Edited by N. Neufeld. ISSN 2226-0358

[5] OpenStack Open Source Cloud Computing Software: http://www.openstack.org

[6] P. Andrade *et al.*, Review of CERN Data Centre Infrastructure, J. Phys.: Conf. Ser. 396 (2012) 042002

[7] CernVM Project: http://cernvm.cern.ch

[8] HammerCloud Portal: http://hammercloud.cern.ch

[9] Puppet Configuration Management System: http://puppetlabs.com

[10] Worldwide LHC Computing GRID: http://wlcg.web.cern.ch