Tests of Cloud Computing and Storage System features for use in H1 Collaboration Data Preservation model

Bogdan Łobodziński

H1 Collaboration, DESY Notkestrasse 85, 22607 Hamburg Germany

E-mail: `bogdan.lobodzinski@desy.de`

Abstract. Based on the currently developed strategy for data preservation and long-term analysis in HEP and in particular on the features for use in a data preservation model for the H1 Collaboration we made a tests of possible future Cloud Computing based on the Eucalyptus Private Cloud platform and the petabyte scale storage open source system CEPH. Improvement of the computing power and strong development of storage systems allows the assumption that a single Cloud Computing supported on a given site will be efficient source of computing and storage resources requested by analysis requirements beyond the end-date of experiments. This work describes our test-bed architecture which could be applied to fulfill the requirements of the physics program of the H1 Collaboration after the end date of the Collaboration. We discuss the reasons why we choose the Eucalyptus platform and storage infrastructure - open source system CEPH for tests systems as well as our experience with installations and support of these infrastructures.

Using our first test results we will examine performance characteristics, noticed failure states, deficiencies, bottlenecks and scaling boundaries.

1. Introduction

The H1 Collaboration first started tests of computing platforms for a data preservation model and Long Term Analysis intensively developed in cooperation with other HEP experiments [1]. The fast progress in the hardware and software development opens up possibilities of creation and practical realization of a new computing paradigm, as well as creating a unique computing opportunities for science. However, each progress creates difficulties as well. It has become more complicated to archive current data and results and make them possible for analysis in the future. A general strategy for archiving projects could be the exploration of operational models allowed by virtualization techniques. Using definitions created for Cloud computing concepts we may see the computing requirements for a data preservation model and Long Term Analysis project as a mixture of Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). For a given PaaS we will need a SaaS in parallel. The requested infrastructure should support:

- a non constant exploration of the computing resources,
- different operational systems and virtual hypervisors,
- well localized storage centers which means support of heterogeneous storage systems,
- common storage area for all virtual instances,

- a simple base to maintain a possible migration of the experimental software to new platforms if necessary,
- access to the external infrastructures such as the Grid and another Cloud installations.

All points listed above move us into the direction of optimized and virtualized internal infrastructure, including the storage layer with the open possibility to be connected to external Cloud or Grid infrastructures.

The data storage part is as well important in our assumptions as the Cloud Middleware. The proper choice of filesystem is an integrated part of the test-bed installation which we would like to check. We have to consider not small capacities of a storage area, therefore, it is not enough to use a local solution of a filesystem. For data of the H1 Collaboration we need about 0.5 PB of storage area, which, located only on a single site could become a bottleneck for the future computing solutions. Therefore, in case of a storage system we put a great importance into the scalability (in terms of an easy expansion or reduction) of the storage solution.

These two parts together: Cloud Middleware and Storage System are the subject of our tests.

In this note we will not consider non-technological concerns, like access aspects and cost models for the computing infrastructure.

2. Middleware selection criteria

Having a variety of available cloud managers, we had to decide which kind of software should we use for our tests. The criteria for a choice of a cloud middleware were following:

- open source,
- possible use of a created Virtual Machine on Amazon AWS without any additional modifications,
- support of multiple clusters,
- possibility to start Scientific Linux 5.x images,
- access to the common storage area,
- easy way to build the interface between involved batch system and Cloud Middleware based on the dynamical execution of virtual instances recognition.

The last point in the list of requirements includes the hidden assumption that we will need an internal, scalable batch system running on the cluster of virtual machines. Available solutions of Cloud Computing managers are not supporting batch scheduling.

In terms of the storage part we used similar criteria as above:

- open source,
- data safety based on file replication (as a manageable level),
- capabilities of support of data scale from Tera to Petabytes,
- seamless scalability: non disturbing and easy way to add or remove an additional storage unit,
- availability of usual operations on the filesystem (editing, writing, reading, updating and saving) directly on the storage area.

Looking into available solutions, we decided to use for our pilot tests: Eucalyptus 2.0 [2] as a Cloud Management System and CEPH Petabyte File System as a storage system, version 0.21 [3]. The CEPH filesystem is an official part of the Linux ernel 2.6.34 and beyond.

3. Description of the test installation

It is not the goal of the note to overwrite the manual of the Eucalyptus 2.0 and the CEPH Petabyte Filesystem, therefore below, we review shortly the main points which lead us to the first results of our tests of the infrastructure.

For tests we used the test installation, schematically shown on the figure 1.

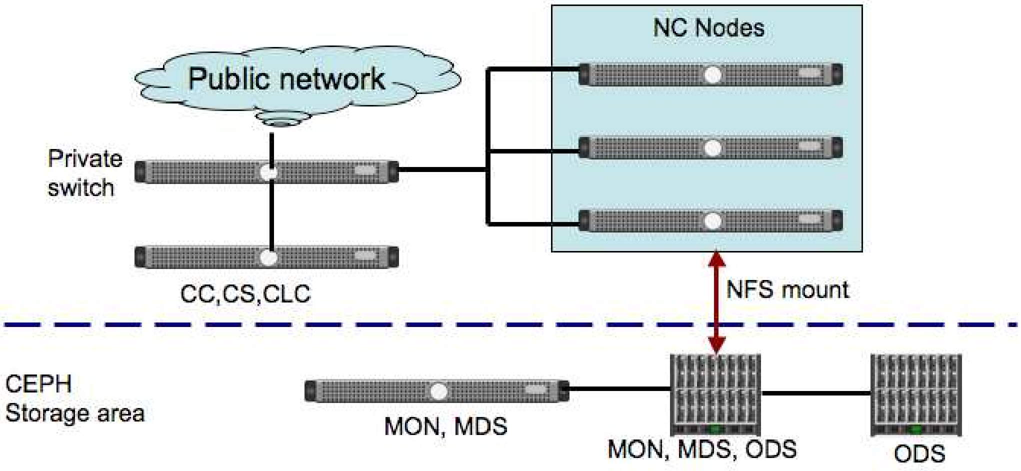

Figure 1. The general scheme of the hardware configuration of described tests. The data storage layer (CEPH Petabyte Filesystem) is attached to the running virtual machines on the node controllers (NC nodes)) using NFS mount point. Description of the acronyms is moved to the text of the note.

As a base for Eucalyptus 2.0 (called later CM) we used 4 nodes with Ubuntu 10.04 OS (4 core Intel 3 GHz CPU, 8 GB RAM). On a single machine we started 3 main services

- Cloud Controller (CC) - which exposes and manages underlying virtualized instances (e.g. node controllers management and access control policies) and storage area,
- Cluster Controller (CLC) - which schedules virtual machines execution to node controllers,
- Storage Controller (SC)- which is responsible for block-accessed network storage.

The remaining 3 nodes are used as a node controllers (NC). The nodes belonging to the cloud manager are configured as a private subnet. For the tests of the CEPH Petabyte Filesystem is used a set of 3 nodes with Ubuntu 10.04 OS. Monitoring service (MON) and Metadata Server (MDS) are running on 2 nodes, Object Storage Device (OSD) which acts as a storage unit is installed on 2 nodes. Details of distribution of the CEPH services are available Figure 1.

Installation of Eucalyptus is well described [4]. For description details of the configuration files and main difficulties during installation process we would like to point to our transparencies [5]. As a network model we used MANAGED mode [6]. It allows to explore all features supported by Eucalyptus 2.0 as security groups and Virtual instances isolation. In the MANAGED mode all virtual instances are controlled by DHCP service fully controlled by Eucalyptus. The initial network configuration has to be created on all nodes before start-up of Eucalyptus services.

For the creation of the system image we used virtual hypervisors KVM [7]. Proper building of the system image is a subject of IT knowledge. We will stress only a general scheme and constraints requested by Eucalyptus middleware. The system image used as a base of virtual machine have to be created as an independent set of filesystem, ramdisk and kernel units. During creation of the virtual machine one must consider the following items:

- use only single partition as small as possible, without any graphical components,
- disable SELinux, remove hardware MAC address from the network configuration files,
- add all necessary libraries requested by future applications,
- a ramdisk has to created with virtio, mptspi and acpiphp modules.

After uploading of all components into the Eucalyptus storage controller we can start to use the image system as a virtual machine. For a future application one can create more kernels and use different combinations of available kernel, ramdisks and filesystems.

After start-up of the virtual machine we can build as an independent task the logical volume with requested amount of space and attach it to the running virtual instance. In addition, we can attach a common storage area (CEPH) using an NFS mount point or create directly the ODS on the running virtual machine with attached logical volume. According to the Eucalyptus manual it should be possible to create a snapshot of the logical volume and attach such an volume instance to the subsequently started virtual machines. However, we did not here success working with this solution.

As a working example we were able to start the H1 Monte Carlo application on the virtual machines using the scheme shown on Figure 2.

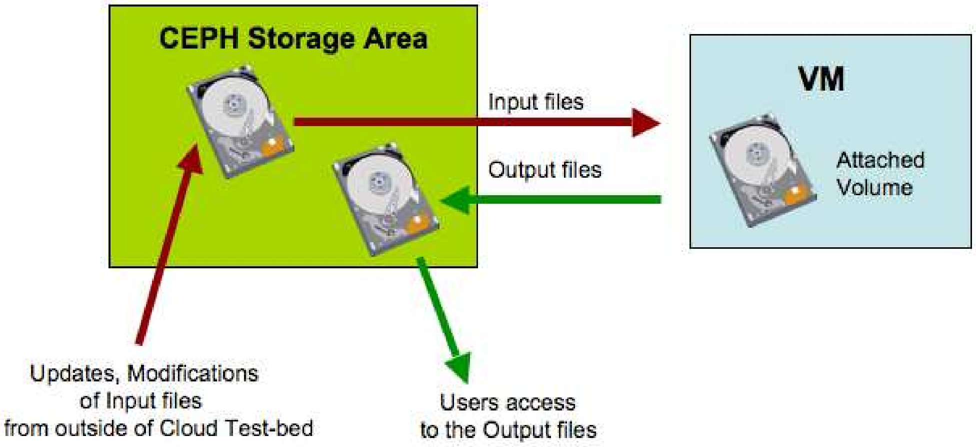
CEPH Storage Area
Input files
Output files
VM
Attached Volume
Updates, Modifications of Input files from outside of Cloud Test-bed
Users access to the Output files

Figure 2. The scheme of successful architecture for pilot tests of the H1 Monte Carlo simulations on the tested configuration. The user starts the virtual machine (VM) with the Monte Carlo application. The processes started on the VM copy input files from the CEPH storage area and move the generated output files back. The CEPH storage part can be accessed from outside.

The second kind of test runs were performed on the installed CEPH Filesystem. As in the case of the Cloud Manager, for details of the configuration files we refer the reader to the ref. [5]. The installation procedure can be found in [8]. The important point is to use btrfs [9] as a base file system. The CEPH system presents very transparent administration level of control. The file replication level can be adjusted independently for the Metadata and Data pools with a single command. The adding of additional storage unit or removing them can also be done seamlessly by the administrator. However, we found that the CEPH Filesystem (version 0.21) is not able to work with too many files. Therefore, the usage of this release of the CEPH software is very limited.

4. Summary

Our tests are a first step towards the future computing platform dedicated to the H1 HEP Collaboration Data Preservation model. We decided to adopt the Private Cloud Computing paradigm with on-demand processing power and storage, operating systems and virtual hypervisors.

Working with Eucalyptus 2.0 as a Cloud Manager and CEPH Petabyte Filesystem (version 0.21) as a storage unit we found a few aspects which show clearly that this kind of software may match well with our general assumptions. However, it cannot be used for more advanced tests due to the poor stability and reliability of both solutions. They are still in the test phase of development. In case of Eucalyptus 2.0 We found such misbehaviors like:

- not availability of the Metadata Service,
- snapshots of the volumes cannot be properly created and attached to the running virtual machines,
- necessity of too frequent manual intervention in case of problems with removing of volumes, snapshots of volumes and registering of a new node controllers,
- non-modular structure - which creates a problem for potential future modifications,
- very complicated and large log files.

For the CEPH Filesystem we have only one comment, which is nevertheless very vital: the system has serious problems with support of a large number of files. However we did not test the last available releases and patches.

Taking into account the early stage of development of tested solutions, out results are very encouraging for continuation of our tests later and shows that our general assumptions about the future computing platform for the data preservation and long-term analysis in HEP and for the H1 Collaboration model in particular are well determined.

4.1. Acknowledgments

The author would like to thank to Yves Kemp from the DESY-IT Department at DESY Hamburg for preparation of the cluster of nodes for our tests.

References

[1] https://www.dphep.org/, http://arxiv.org/pdf/1101.3186
[2] http://open.eucalyptus.com/
[3] http://ceph.newdream.net
[4] http://open.eucalyptus.com/wiki/EucalyptusInstalltionDebian_v2
[5] http://117.103.105.177/MaKaC/getFile.py/access?contribId=268&sessionId=73&resId=0&materialId=slides&confId=3
[6] http://open.eucalyptus.com/wiki/EucalyptusNetworkConfiguration_v2.0
[7] short review of image creation for Eucalyptus 2.0 using KVM Hypervisor: http://open.eucalyptus.com/participate/wiki/creating-images-iso-kvm
[8] http://ceph.newdream.net/wiki/Debian
[9] https://btrfs.wiki.kernel.org/index.php/Main_Page