Contribution ID: **135** Type: **Poster**

Long-term preservation of analysis software environment

Tuesday 22 May 2012 13:30 (4h 45m)

Long-term preservation of scientific data represents a challenge to all experiments. Even after an experiment has reached its end of life, it may be necessary to reprocess the data. There are two aspects of long-term data preservation: “data” and “software”. While data can be preserved by migration, it is more complicated for the software. Preserving source code and binaries is not enough; the full software and hardware environment is needed. Virtual machines (VMs) may offer a solution by “freezing” a virtual hardware platform “in software”, where the legacy software can run in the original environment.

A complete infrastructure package is developed for easy deployment and management of such VMs. It is based on a dedicated distribution of Linux, CERNVM. Updated versions will be made available for new software, while older versions will still be available for legacy analysis software. Further, a HTTP-based file system, CVMFS, is used for the distribution of the software. Since multiple versions of both software and VMs are available, it is possible to process data with any software version, and a matching VM version. OpenNebula is used to deploy the VMs. Traditionally, there are many tools for managing clouds from a VM point-of-view. However, for experiments, it can be more useful to have a tool which is mainly centred around the data, but also allows for management of VMs. Therefore, a point-and-click web user interface is being developed that can (a) keep track of the processing status of all data; (b) select data to be processed and which type of processing, also selecting the version of software and matching VM; and (c) the configuration of the processing nodes, e.g. memory and number of nodes. It is preferable that the interface has an experiment-dependent module which will allow for easy adoption to various experiments. The complete package is designed to be easy to replicate on any processing site, and to scale well. Besides data preservation, this paradigm also allows for distributed cloud-computing on private and public clouds through the EC2 interface, for both legacy and contemporary experiments, e.g. NA61 and the LHC experiments.

Summary

Long-term preservation of scientific data represents a challenge to experiments, especially with regard to the analysis software. Preserving source code and binaries is not enough; the full software and hardware environment is needed. Virtual machines (VMs) make it possible to preserve hardware “in software”. A complete infrastructure package is developed for easy deployment and management of VMs, based on CERNVM Linux. Older CERNVM versions will still be available for legacy software. Further, a HTTP-based file system, CVMFS, is used for the distribution of the software. It is possible to process data with any software version, and a matching VM version. Most importantly, a point-and-click web user interface is being developed for setting up the complete processing chain, including VM/software versions, number/type of processing nodes, and the particular type of analysis and data. This paradigm also allows for distributed cloud-computing on private and public clouds, for both legacy and contemporary experiments.

Author: LARSEN, Dag (University of Bergen (NO))

Co-authors: HARUTYUNYAN, Artem (CERN); CHARALAMPIDIS, Ioannis (Aristotle Univ. of Thessaloniki (GR)); BLOMER, Jakob (Ludwig-Maximilians-Univ. Muenchen (DE)); BUNCIC, Predrag (CERN)

Presenters: HARUTYUNYAN, Artem (CERN); LARSEN, Dag (University of Bergen (NO))

Session Classification: Poster Session

Track Classification: Distributed Processing and Analysis on Grids and Clouds (track 3)