Virtualization and Cloud Computing in PANDA

Matthias Steinke
Ruhr-Universität Bochum

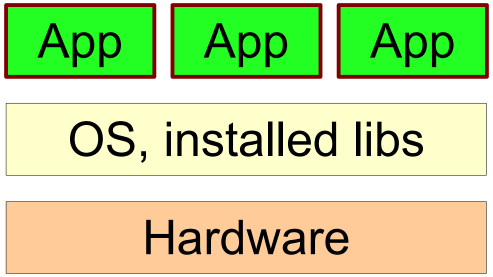
Some Points which have to be improved

- Usability of the offline software: Simulation and analysis work for the scrutiny process has shown that
  - Analysts need a lot of time to install PANDARoot and to bring it to work
    - Analysts want to run analysis jobs, they do not want to install external packages, apply bug fix patches, etc.
  - We waste person power b/c things have to be installed, compiled & linked, and administered at **each** site and on **each** desktop.

- Test coverage: The PANDA offline software does not always run in a controlled environment b/c of all the individually configured desktops/laptops/worker nodes
  - Running unit tests and integration tests for all the different setups is virtually impossible
    - → Can we publish results which are based on untested software??

- Distributed computing:
  - Alien is not supported anymore
  - Cloud computing is in some sense the successor of the “classical” grid computing
  - Clouds got invented for the LHC computing
    - → Should fit perfectly for PANDA needs

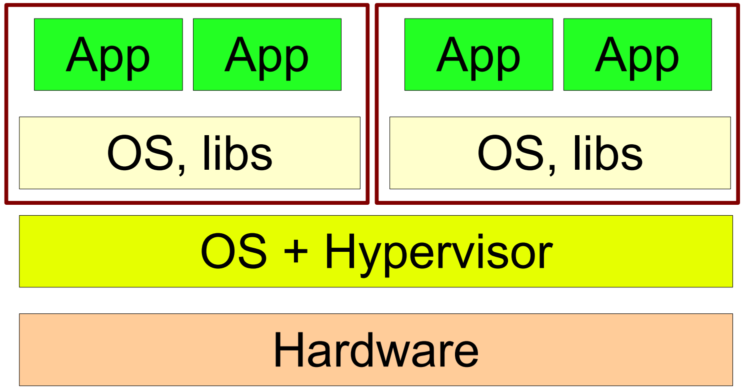
What is Virtualization?



The PANDA offline software runs the OS and uses the libs that are installed on a specific computer.

PANDA Computing provides an application, or even just the source code of an application.

The user has to maintain external packages and to compile and link applications.



A hypervisor virtualizes the hardware of a specific computer.

PANDA Computing provides appliances (VM images), bundles of applications and the libs and the OS they need to run.

The user starts a virtual machine and runs the applications which were compiled, linked and tested by the PANDA Computing experts, and which use the libs provided by PANDA computing.

Why does that help? What has to be done?

- Workload is removed from (many) users (who mostly are dummies in computer administration)
  - ➔ Physics analysis can be done more efficiently

- PANDA Computing gets the chance to test **all** applications that are used in simulation and reconstruction, and most of the code used in physics analysis

What has to be done:

- Administrators at computing centers or university institutes (or users for personal laptops) have to install a hypervisor like KVM/qemu or VirtualBox
  - These are available for free for all relevant platforms
  - Installation takes ~2 h and has to be done **once**
- PANDA Computing has to provide appliances for all production releases
  - Creating an appliance also takes ~2 h

- As a proof of principles I virtualized *Pawian,* a partial wave analysis software, and produced images for VirtualBox and for KVM
  - ➔ worked w/o problems

What is Cloud Computing?

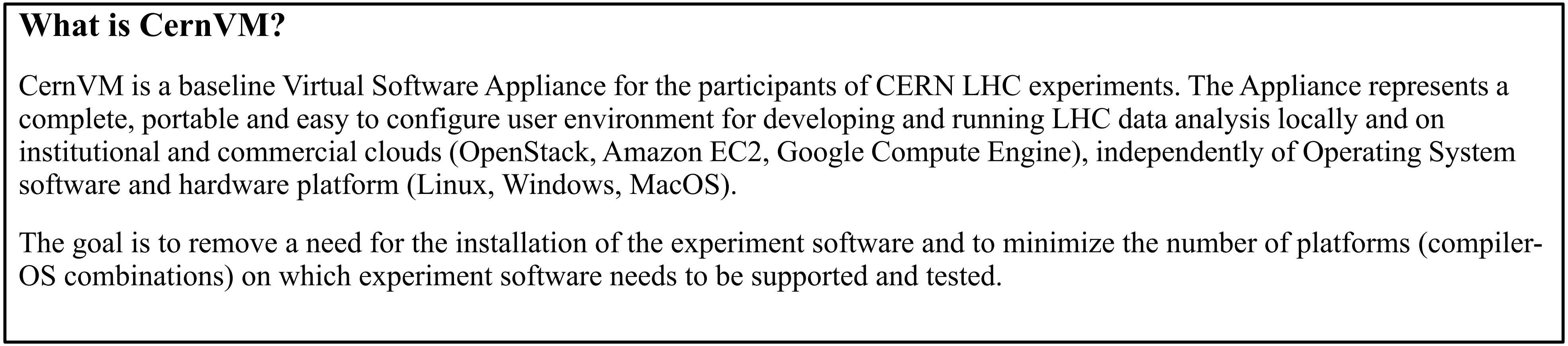
To run an analysis on a large data set or to run a Monte Carlo production, we use distributed computing farms

- ➔ If we want to use virtualization, we have to start VMs on the farm nodes
  - ➢ Cloud middleware exactly does that for you
    - ➔ We have to migrate from classical grid computing to cloud computing

PANDA is an experiment at FAIR, but it's also a CERN experiment

- ➔ Let's see what the CERN people did (for us) and what the LHC experiments are using

Virtualization @ CERN



What is CernVM?

CernVM is a baseline Virtual Software Appliance for the participants of CERN LHC experiments. The Appliance represents a complete, portable and easy to configure user environment for developing and running LHC data analysis locally and on institutional and commercial clouds (OpenStack, Amazon EC2, Google Compute Engine), independently of Operating System software and hardware platform (Linux, Windows, MacOS).

The goal is to remove a need for the installation of the experiment software and to minimize the number of platforms (compiler-OS combinations) on which experiment software needs to be supported and tested.

from cernvm.cern.ch/portal

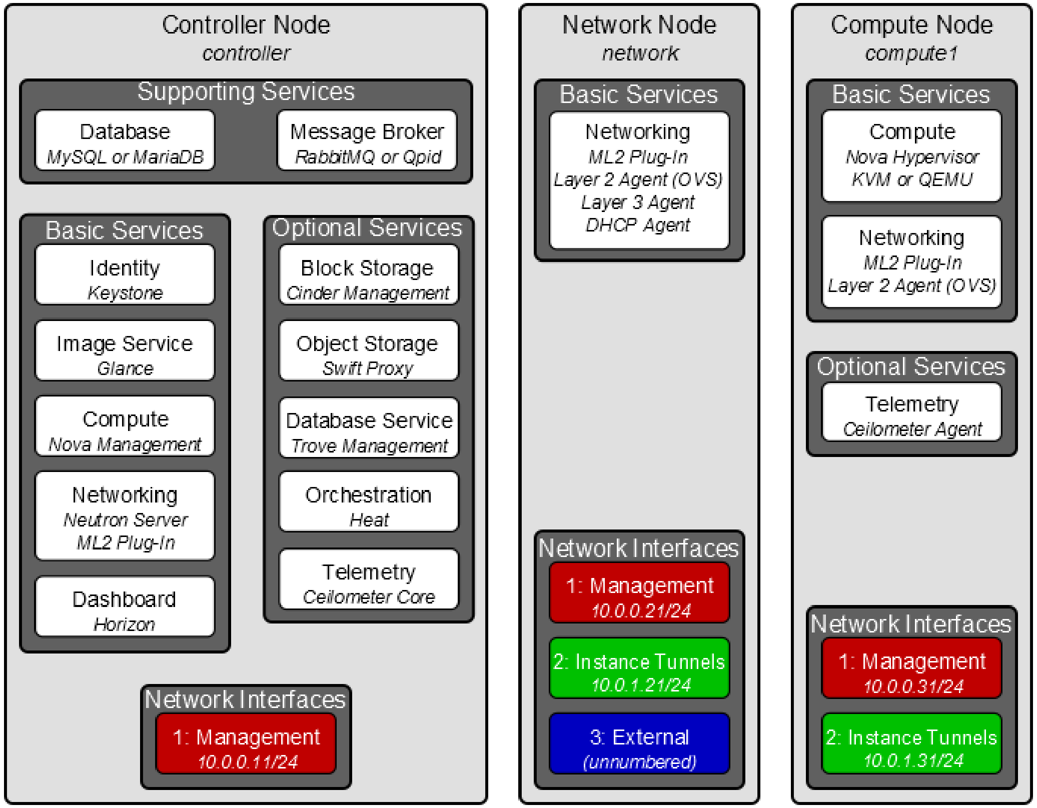
What has to be done?

CERN cloud software runs on OpenStack, which uses KVM

- OpenStack is the quasi-standard and is compatible with e.g. Amazon's ECC

➔ Admins of compute farms at universities and at computing centers have to set up OpenStack

➔ PANDA may make use of the CERN tools to monitor and administer of the “PANDA Cloud”



http://docs.openstack.org/icehouse/install-guide

Summary

By making use of virtualization we can

- improve the usability of the offline software
- improve the test coverage of the software

Cloud computing

- allows to use virtualization in distributed computing
  - ➔ the **same**, well **tested** code runs on all worker nodes and on all user desktops and laptops
- is in some sense an advancement of the “classical” grid computing

Using the CERN developments may be an easy path towards a PANDA Cloud.