Virtualization & Clouds

2nd ASPERA Workshop
30-31 May 2011, Barcelona, Spain
P. Mato /CERN

Outline

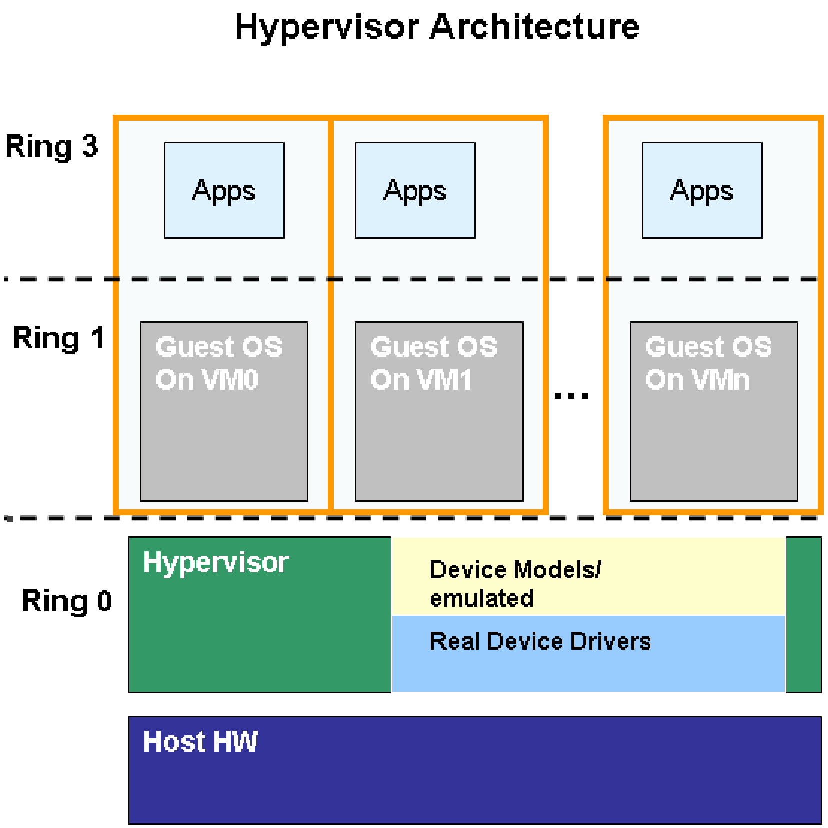
- Brief introduction to Virtualization
  - Enabling technology for the Cloud Computing revolution
- Usages of Virtualization technology
  - Use cases that are revolutionizing HEP Computing
- Main difficulties with Virtualization and Clouds
  - Limiting its generalized adoption
- Initial tests and success stories
  - Selected examples
- Summary

Virtualization

- The idea of abstracting computer resources is not new
  - Credit for bringing virtualization into computing goes to IBM with VM/370 (1972)
- In recent years a big inflation of “full virtualization” solutions making use of “hardware-assisted” virtualization
  - Parallels, VirtualBox, KVM, Xen, VirtualBox,Hyper-V, VMware, etc.
  - Hypervisors simulates enough hardware to allow an unmodified "guest" OS
- The key challenge for “full virtualization” is the interception and simulation of privileged operations
  - The effects of every operation performed within a given virtual machine must be kept within that virtual machine
  - The instructions that would "pierce the virtual machine" cannot be allowed to execute directly; they must instead be trapped and simulated.

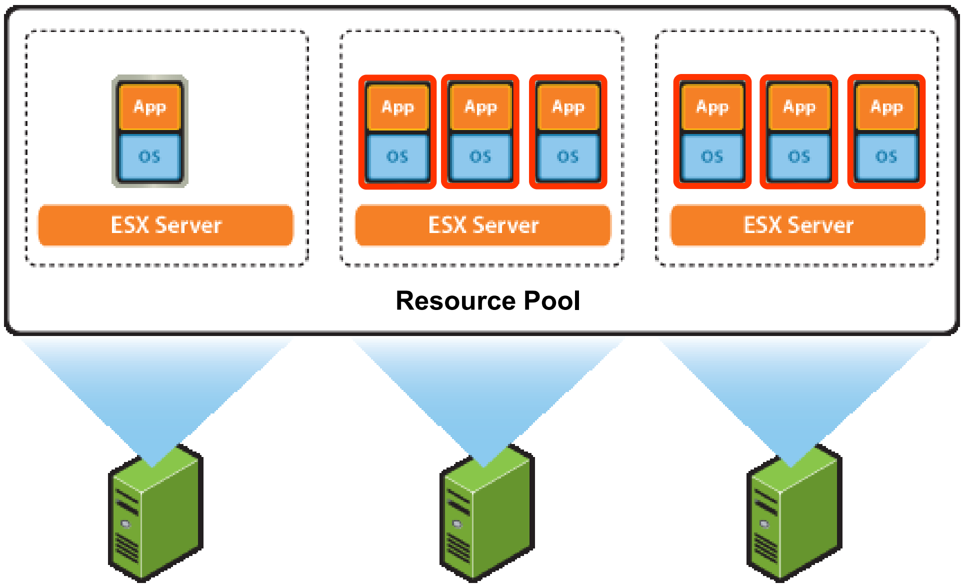
Hypervisor architecture

- A technique that all (software based) virtualization solutions use is ring deprivileging:
  - the operating system that runs originally on ring 0 is moved to another less privileged ring like ring 1.
  - This allows the VMM to control the guest OS access to resources.
  - It avoids one guest OS kicking another out of memory, or a guest OS controlling the hardware directly.



Usages of Virtualization

Server Consolidation

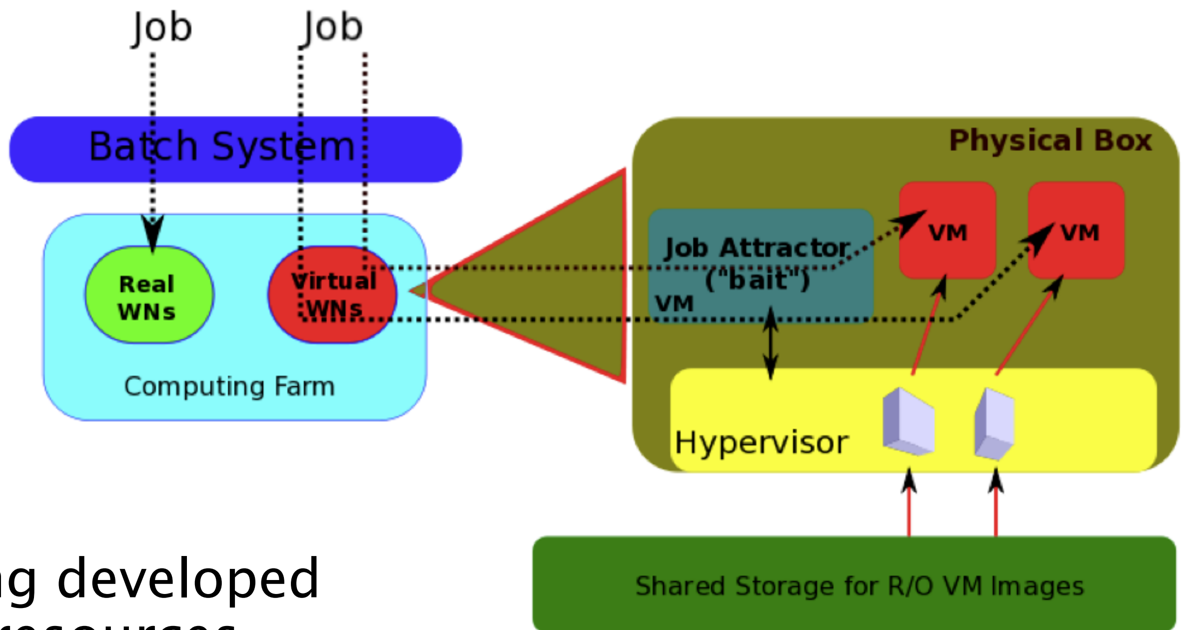


- Consolidates workloads onto fewer servers when the cluster needs fewer resources
- Places unneeded servers in standby mode
- Brings servers back online as workload needs increase

VMware vMotion technology

- Minimizes power consumption while guaranteeing service levels
- No disruption or downtime to virtual machines

Worker Node Virtualization



- Prototypes are being developed to virtualize batch resources
  - Decouple jobs and physical resources
  - Ease management of the batch farm resources
  - Enable the computer center for new computing model
- Examples: CERN virtual batch, CNAF worker nodes on demand

Software Testing

- Virtual machines can cut time and money out of the software development and testing process
- Great opportunity to test software in a large variety of 'platforms'
  - Each platform can be realized by a differently configured virtual machines
  - Easy to duplicate same environment in several virtual machines
  - Testing installation procedures from well defined 'state'
  - Etc.
- Example: Execution Infrastructure in ETICS (spin-off of the EGEE project)
  - Set of virtual machines that run a variety of platforms attached to an Execution Engine where Build and Test Jobs are executed on behalf of the submitting users

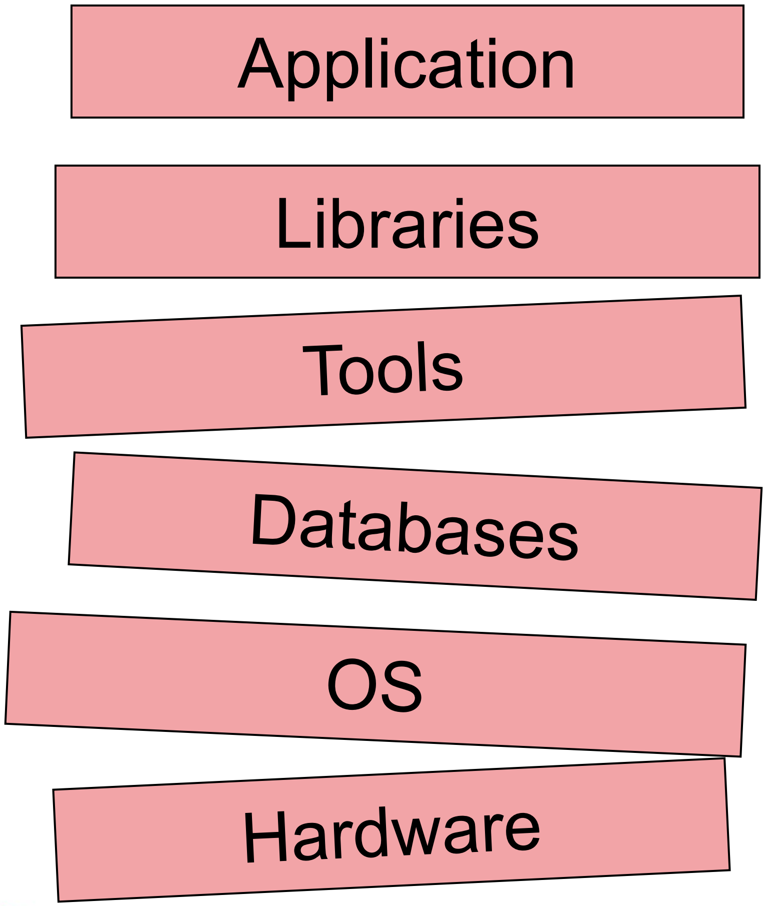
Training Platform

- Similar as for software testing infrastructure, virtualization helps to deploy rapidly dedicated software and workstations/servers for training
  - Need for many nodes rapidly and typically for a rather short period of time
  - Isolation with respect production servers
  - Disposable workstations/servers

Software Deployment Problem

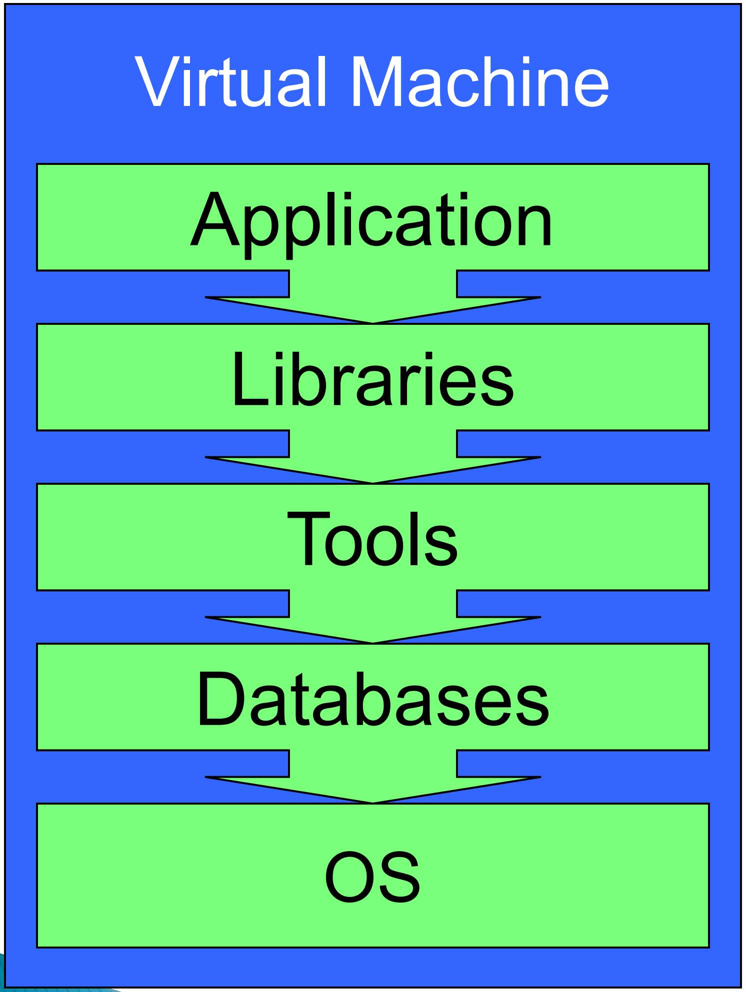
- Software @ LHC
  - Millions of lines of code
  - Different packaging and software distribution models
  - Complicated software installation/update/configuration procedure
  - Long and slow validation and certification process
  - Very difficult to roll out major OS upgrade (SLC4 -> SLC5)
  - Additional constraints imposed by the grid middleware development
    - Effectively locked on one Linux flavour
    - Whole process is focused on middleware and not on applications
- How to effectively harvest multi and many core CPU power of user laptops/desktops if LHC applications cannot run in such environment?
- Good news: We are not the only one with such problems…

Horizontal Integration



- Traditional model
  - Horizontal layers
  - Independently developed
  - Maintained by the different groups
  - Different lifecycle
- Application is deployed on top of the stack
  - Breaks if any layer changes
  - Needs to be certified every time when something changes
  - Results in deployment and support nightmare

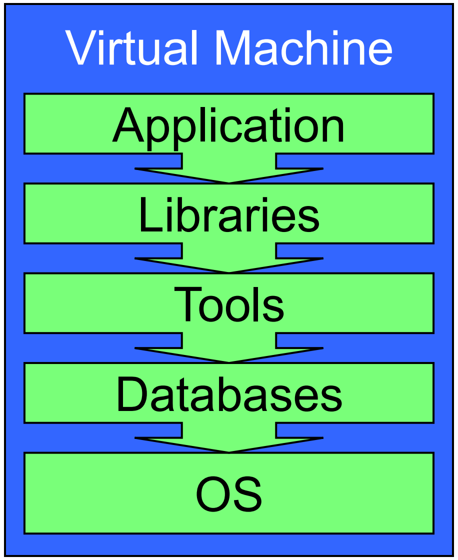
Vertical Integration



- Application driven approach
  - Analyzing application requirements and dependencies
  - Adding required tools and libraries
  - Building minimal OS
  - Bundling all this into Virtual Machine image
- Virtual Machine images should be versioned just like the applications
  - Assuring accountability to mitigate possible negative aspects of newly acquired application freedom

Rethinking Application Deployment

- Emphasis in the 'Application'
  - The application dictates the platform and not the contrary
- Application (e.g. simulation) is bundled with its libraries, services and bits of OS
  - Self-contained, self-describing, deployment ready
- What makes the Application ready to run in any target execution environment?
  - e.g. Traditional, Grid, Cloud
- ➔ Virtualization is the enabling technology



Cloud Computing

- Is the convergence of three major trends
  - Virtualization - Applications separated from infrastructure
  - Utility Computing – Capacity shared across the grid
  - Software as a Service – Applications available on demand
- Commercial Cloud offerings can be integrated for several types of work such as simulations or compute-bound applications
  - Pay-as-you-go model
  - Question remains in their data access capabilities to match our requirements
  - Good experience from pioneering experiments (e.g. STAR MC production on Amazon EC2, Belle2 production with DIRAC on Amazon EC2)
  - Ideal to absorb computing peak demands (e.g. before conferences)
- Science Clouds start to provide computer cycles for scientific communities

➔ More later

Volunteering Computing

- BOINC: potential of 300k volunteers, 5 PetaFLOPS
  - E.g. LHC@home: stability of proton orbits in CERN's LHC accelerator (40k volunteers, 70k PCs)
- Problems with "normal" BOINC used for LHC physics
  - Most of clients on Windows, LHC software runs on Scientific Linux
  - Physics code changes often
  - Keep the same job submission interface to physicists
  - Job management and monitoring is primitive
- Virtualization can clearly help as for the cloud use case

Data Preservation

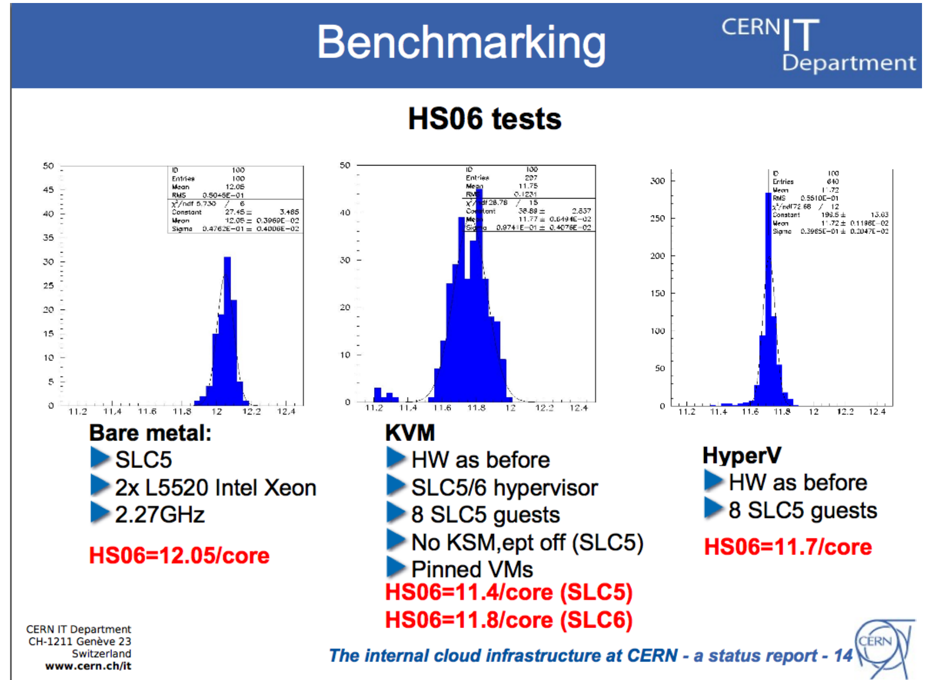
- Virtualization techniques are very attractive for data preservation and long-term analysis in HEP
  - The main problem is not to preserve the data itself
  - It is to be able to run the old software, which is needed to interpret the data and be able re-run analysis
- Virtual appliance can be created of the OS with the experiment software in a standard format
  - Care has to be taken on what protocols and interfaces the appliance will be communicating to the outside

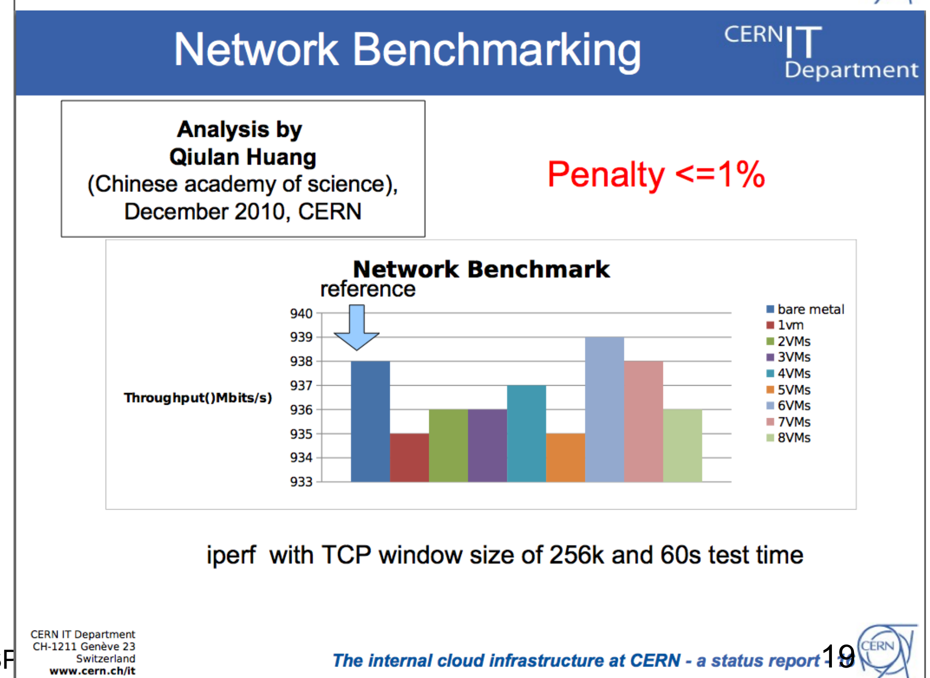
Main Difficulties with Virtualization

- Performance
  - Fear that the performance will much lower than bare-metal
- Management tools and standard interfaces
  - Many open source and commercial solutions
  - EC2 becoming de-facto standard API
- Managing large VM images
  - Distributing large images to many centers can be a problem
- Trusting VM images
  - Many sites do not trust user provided images
- Contextualization
  - Need to customize images to specific function
- Adequate storage architecture for HEP
  - Large data access requirements
- Costs
  - Cost comparisons between public and private clouds

Performance

- CPU benchmarking
  - Requires some tuning !!
  - Differences between hypervisors
  - Penalty < 2-3%
- Disk I/O benchmarking
  - Penalty 20%-30%
- Network benchmarking
  - Insignificant lost of performance



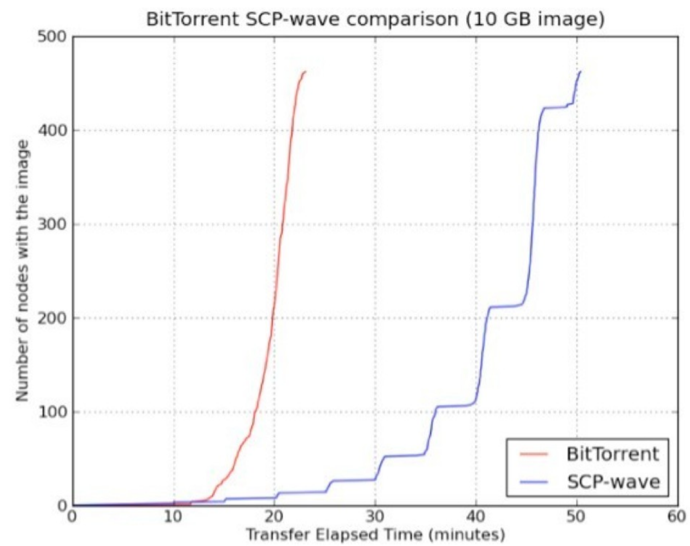


Management tools and API

- Many vendors offer Cloud (IaaS) management tools
  - VMware vSphere, vShield, vCloud familiy
  - Platform ISF
- Open source solutions being evaluated at various sites
  - Eucalyptus
  - Nimbus
  - OpenNebula
  - OpenStack
- OCCI from **Open Grid Forum** (OGF) is a protocol and API for all kinds of management tasks
- Amazon/EC2 API becoming de factor standard
  - Open source systems like OpenNebula, Nimbus implements a sub-set of the interface
  - Allows to start/stop virtual machines and monitor them

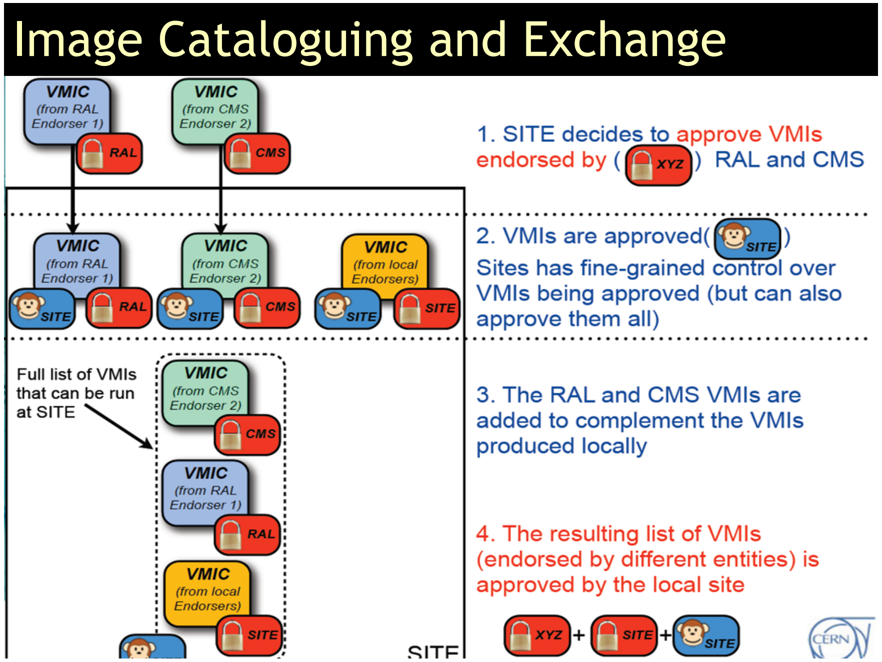
Managing Large VM images

- Images including ‘standard batch node’ [2GB] and all experiment software [10GB] will be large
- Distributing them is not a simple problem
  - CERN-IT is developing the transfer of the images using Bittorrent
- Experiment software changes often
  - Often every week a new version
- Image migration and managing several formats
  - Each hypervisor basically requires a different format



Trusting VM Images

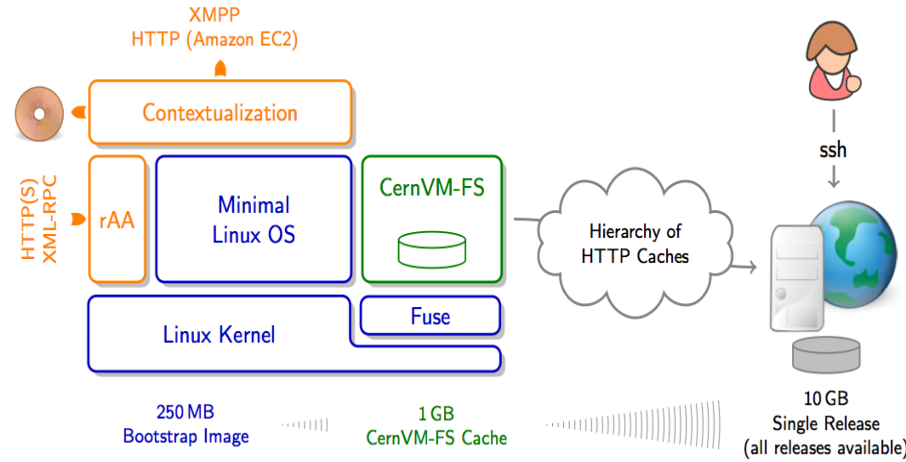
- The EC2 model not acceptable for many HEP sites
- HEPiX working group established to enable exchange of trusted virtual machine images between sites
  - Defining image generation policies
- Cataloguing endorsed images and to track images approved for instantiation at a given site
- Contextualization is needed so that sites can configure images (add to the them) to interface to local infrastructure



Contextualization

- The process of customizing a VM template to its deployment context is called **contextualization**
- This is needed to give configuration parameters to a newly started virtual machine
  - Defining what will be the VM function
  - Configuring its own IP address and make known others
  - Adding additional services for monitoring and accounting
  - Installing user credentials, public keys, etc.
- Several methods exists
  - Mounting a CDROM image and executing some scripts (HEPiX recommended)
  - EC2 API user_data method
  - Web UI, Contextualization Agent, etc.

CernVM
SoFTware Appliance



- CernVM is a R&D project started 3 years ago on Virtualization
- The CernVM image is an attempt to mitigate the mentioned difficulties (performance, image distribution, trust, contextualization, etc.)
  - Tuned for best performance of HEP applications
  - Single image fits all [LHC] experiments
  - Very small is size (only 250MB) with just-enough OS
  - Experiment software is factorized out (dedicated File System)
  - Flexible configuration and contextualization mechanisms

➔ See presentation this afternoon for details

Storage Architecture

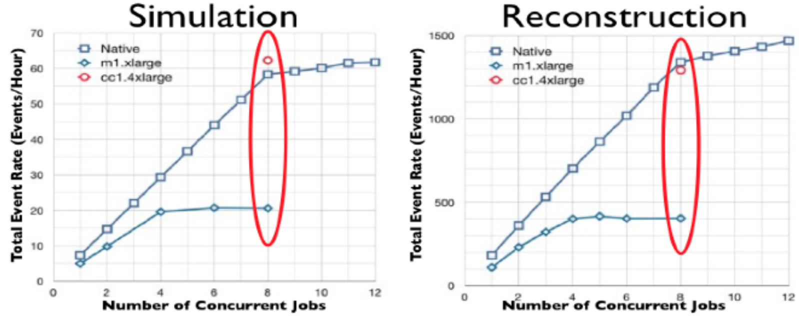
- General consensus that adequate [for HEP] storage solutions for Cloud Computing is a real challenge
  - Storage such as Amazon S3 is probably insufficient in terms of access performance [and cost]
- What is probably required is a global file system able to store many new Petabytes every year with sufficient redundancy and tactical local caches
  - Secured access with good performance from anywhere
  - Interfaced to data access package such as XRootd

➔ See talk from Dirk Duellmann later in the morning

Cost

- Public vs. Private Clouds
- Recent estimates shows that Amazon EC2 can be several times more expensive than dedicated HEP centers
- Amazon EC2 Spot Instances can reduce the cost of running HEP jobs
  - Making it competitive to Tier 3 sized centers
- Storage cost has not been included in these estimates

Cost Estimate (Cost per 1K Evt)



Assume we run 8 concurrent jobs for all cases, the cost per 1k event is calculated.

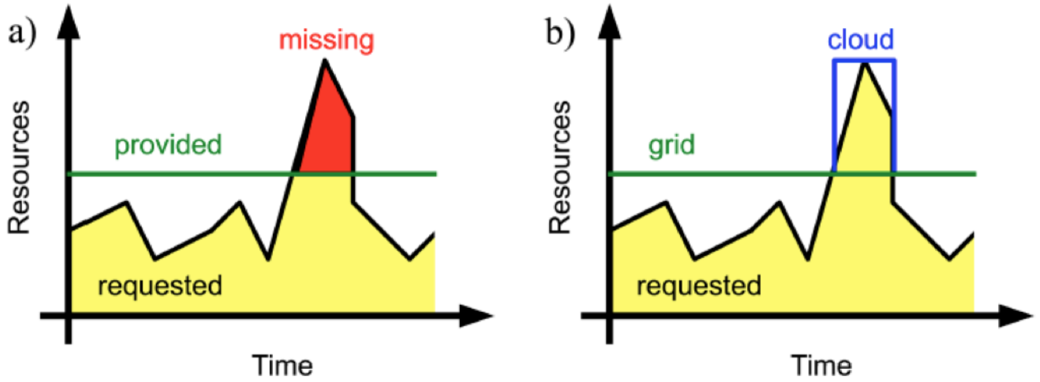
Cost / 1K Event (USD)	EC2		Tier3 Size Center	Large Center (hundreds K cores)
	m1.xlarge	cc1.4xlarge		
Simulation	37	26	11	5
Reconstruction	1.88+Storage	1.24+Storage	0.48	0.24

EC2 cost doesn't include storage (in/out/store), which is very significant as well!!!!

Yushu Yao, *Performance of ATLAS Jobs in the EC2 Cloud*, April 4, 2011.

Absorbing peak demands

- General consensus that peak demands on computational requirements (mainly simulation due to its low I/O demands) could be satisfied using public cloud capacity
- This implies that there is a to common interface (API) between private and public clouds



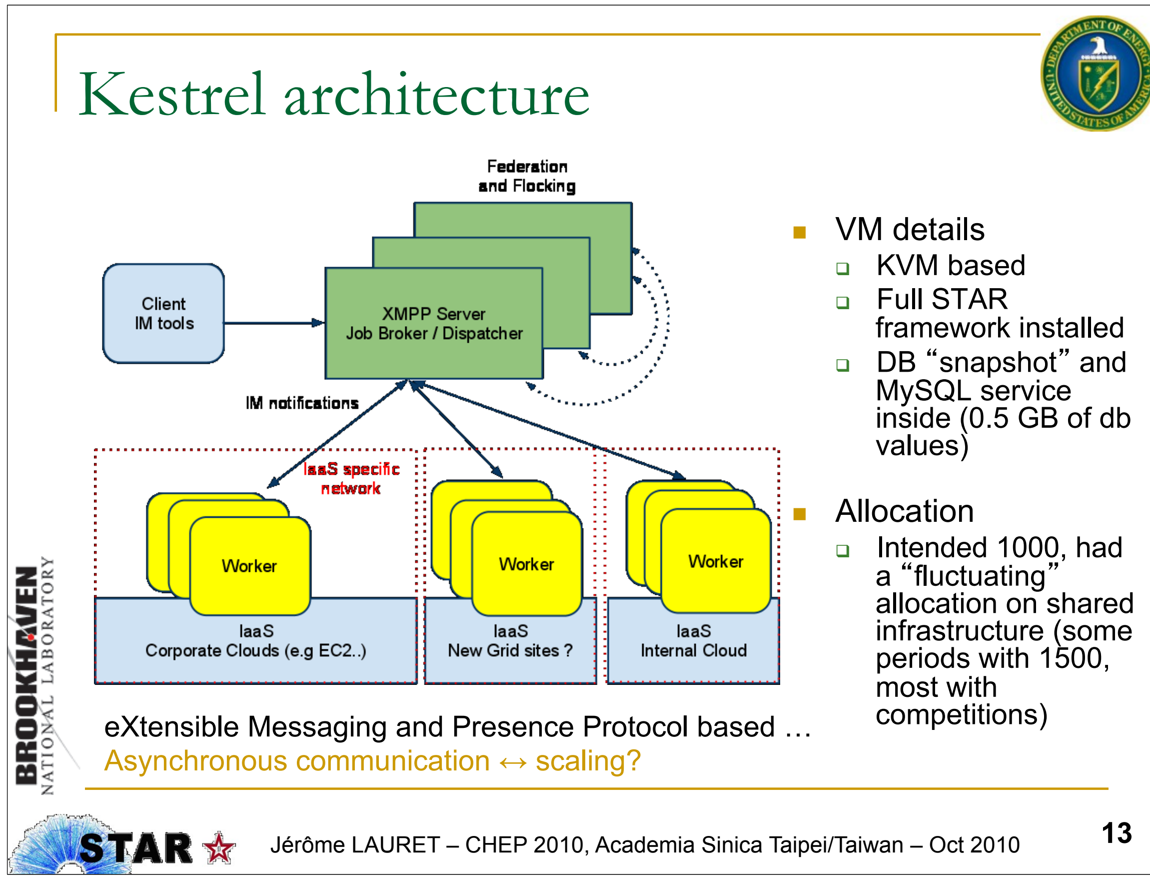
Virtual Clusters

- Cloud computing should enable us to 'instantiate' all sort of virtual clusters effortless
  - PROOF clusters for individuals or for small groups
  - Dedicated Batch clusters with specialized services
  - Etc.
- Turnkey, tightly-coupled cluster
  - Shared trust/security context
  - Shared configuration/context information
- IaaS tools such as Nimbus or OpenNebula would allow one-click deployment of virtual clusters
  - E.g. the OSG STAR cluster: OSG head-node (gridmapfiles, host certificates, NFS, Torque), worker nodes: SL4 + STAR

Initial Tests and Success Stories

STAR at BNL

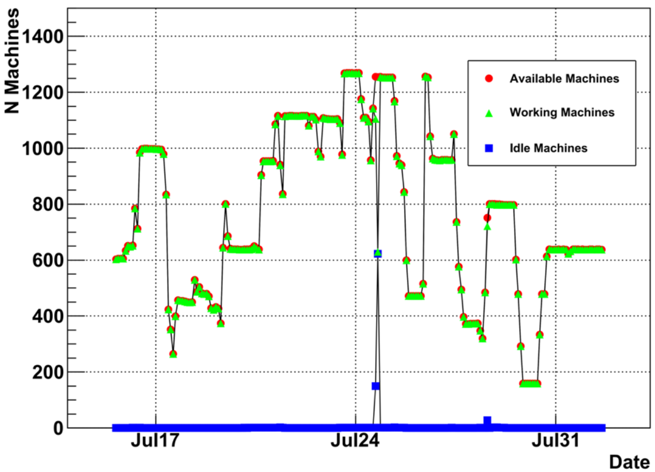
- STAR has been pioneering in testing cloud solutions in HEP
- The adopted architecture combines the use of private and public IaaS

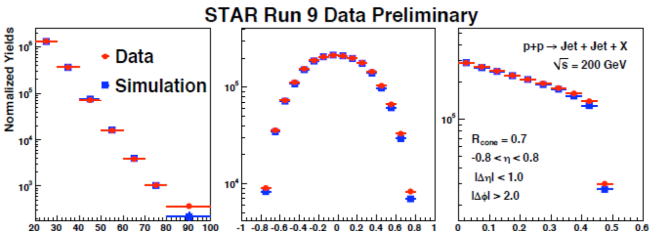


STAR at BNL (2)

Clemson/Kestrel model

- **Generation**
  - **12 Billion PYTHIA events were generated** – LARGEST sample produced we know off
  - **Used over 400,000 CPU hours on 1,000 CPUs** at Clemson (+CERN) over the course of a month
- **Comparison to normal operation**
  - Cloud allowed STAR to expand its computing resources by 25%. Student thesis work possible
    - Available #of CPU per users ~ 50
  - A year long science wait time.
- **Achievement for this analysis**
  - 4 orders of magnitude increase in number of events used in similar analysis in STAR
  - Near elimination of all uncertainties caused by statistics
  - Un-ambiguously demonstrated good agreement between our data sample and simulation
  - **Results presented at Spin 2010 conference (October)**



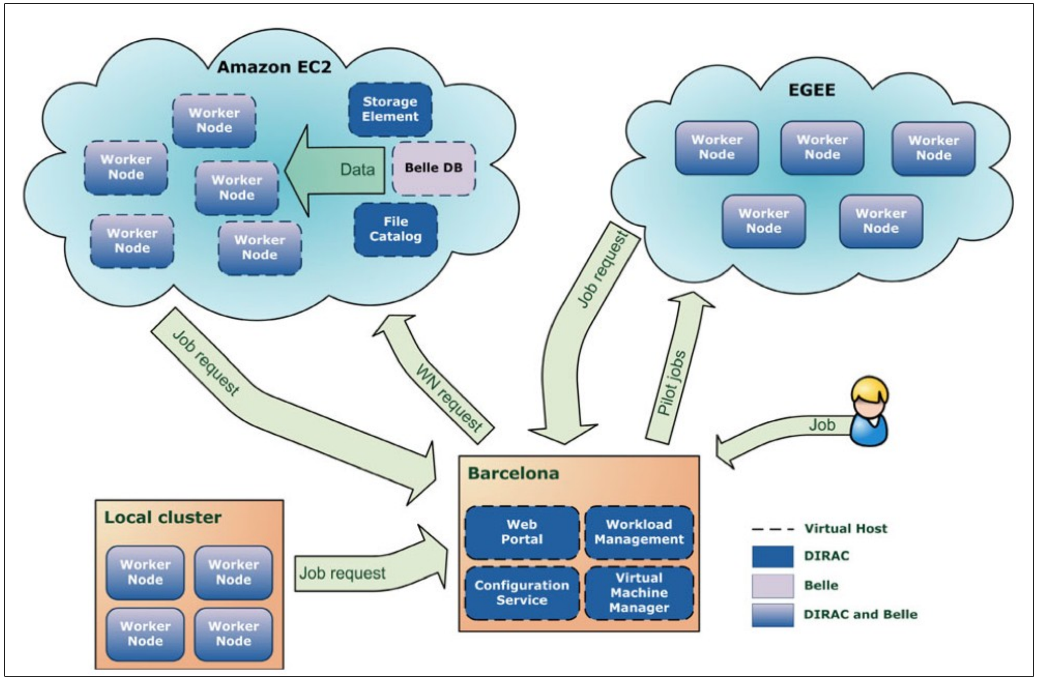


BROOKHAVEN NATIONAL LABORATORY

STAR

Belle II with DIRAC

- Extended the DIRAC distributed "Agents" to run on Amazon EC2
  - Offering exactly the same interface to "end-users"
- The results of a first test using over 2000 days of CPU time showed over 90% efficiency



➔ See *Automating Data Pipelines with DIRAC* presentation this afternoon

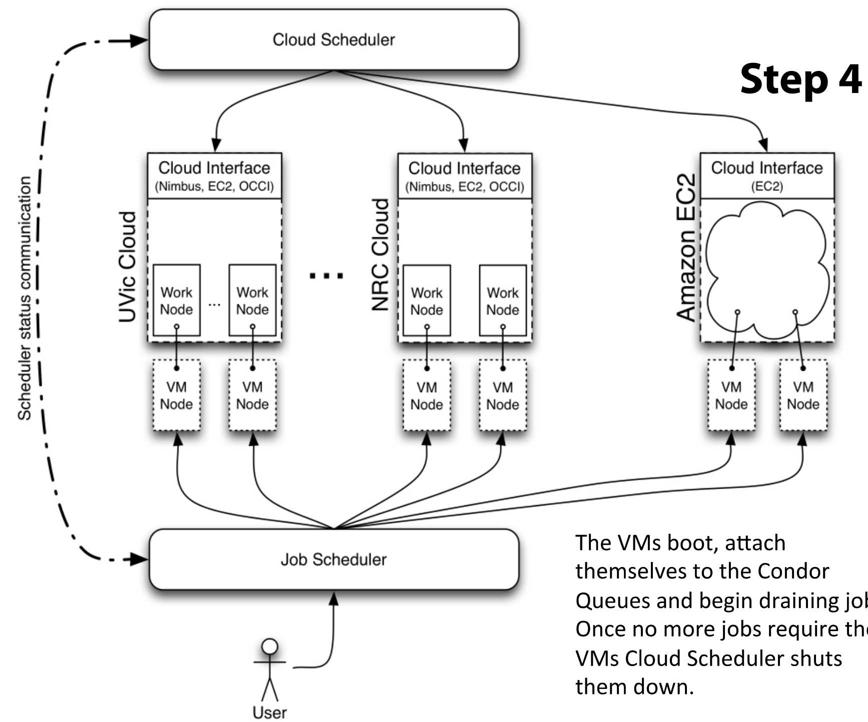
Cloud for BaBar

- Combining resources from different 'clouds' in a transparent way
  - Condor queues are the user interface
- Biggest challenge is data access
  - Using Xrootd with some lost of efficiency

Resource	Cores	Notes	Collaboration Certified
FutureGrid @Argonne Lab	100 Cores Allocated	Resources allocation to support BaBar	★
Elephant Cluster @UVic	88 Cores	Experimental cloud cluster hosts (xrootd for cloud)	★
NRC Cloud in Ottawa	68 Cores	Hosts VM image repository (repoman)	★
Amazon EC2	Proportional to $	Grant funding from Amazon	★
Hermes Cluster @Uvic	Variable (280 max)	Occasional Backfill access	

★ Certified for Monto Carlo Production by Babar Collaboration

University of Victoria NRC·CNRC Ian Gable 17

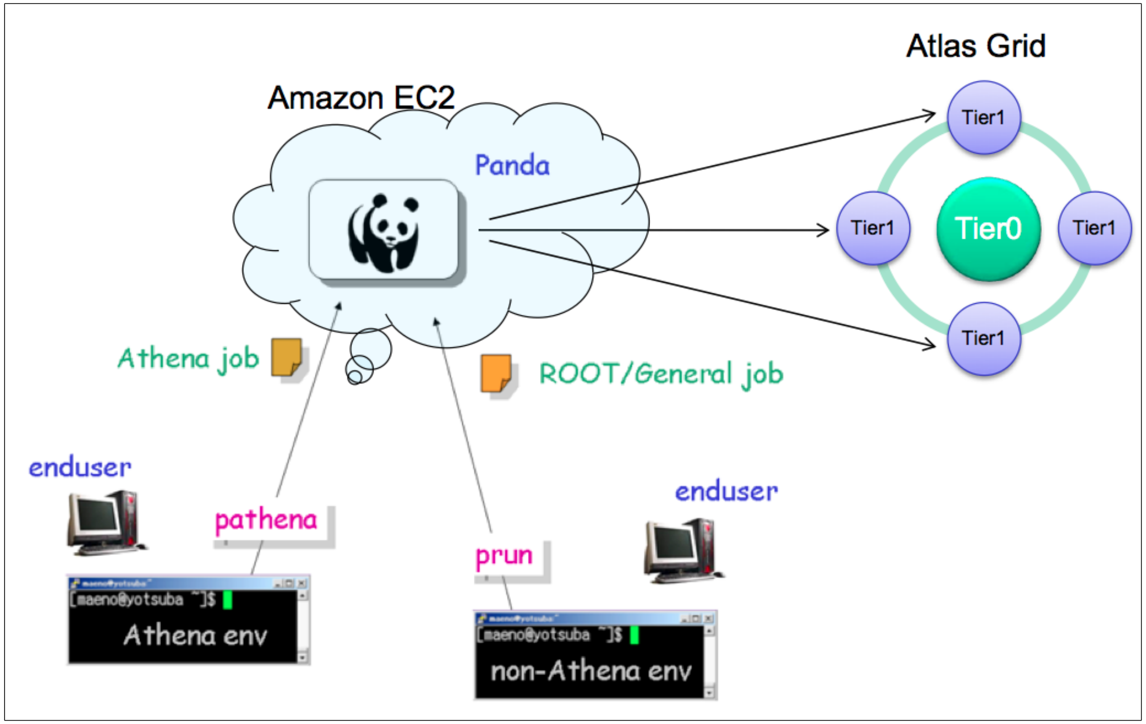


The VMs boot, attach themselves to the Condor Queues and begin draining jobs. Once no more jobs require the VMs Cloud Scheduler shuts them down.

University of Victoria NRC·CNRC Ian Gable 8

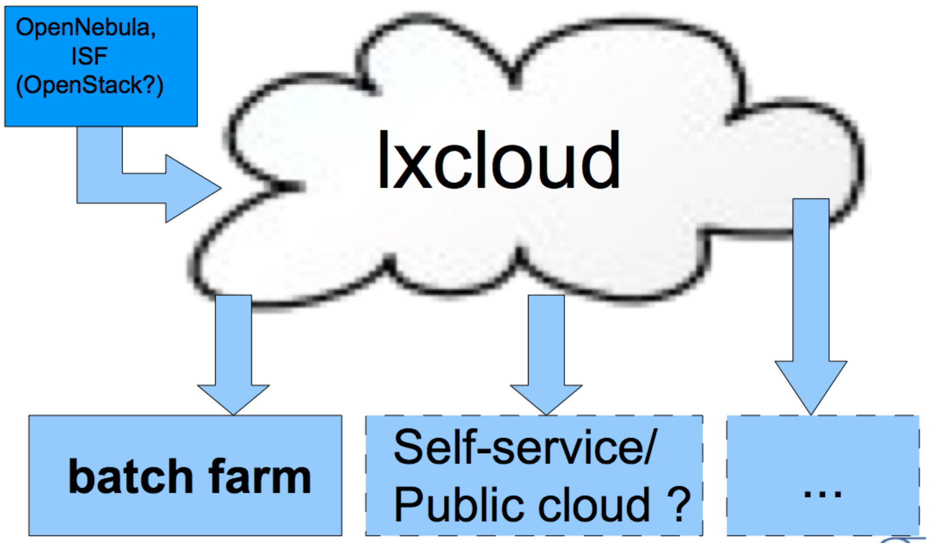
PanDA in the Clouds

- PanDA – Production and Distributed Analysis system for ATLAS
- Different approach: what was run in the Cloud was the PanDA server
  - Until it was installed at CERN



LXcloud

- CERN long-term strategy is to fully virtualize the computer center
  - All IT services will run on virtual servers
- Implementing also an internal cloud (LXcloud)
- Starting to evaluate the test cloud with experiment job schedulers such as PanDA
  - Using CernVM image



Summary

- Virtualization is a broad term that refers to the abstraction of computer resources
  - Enabling vertical software integration
  - Enabling technology of Cloud computing and making Volunteering computing useful for HEP
  - Virtualization is here to stay for a foreseeable future
- Reviewed the main difficulties with the virtualization and cloud technology
  - CernVM image to mitigate some of them
- Selected a number a pioneering tests and success stories
  - Towards fully virtualized computers centers