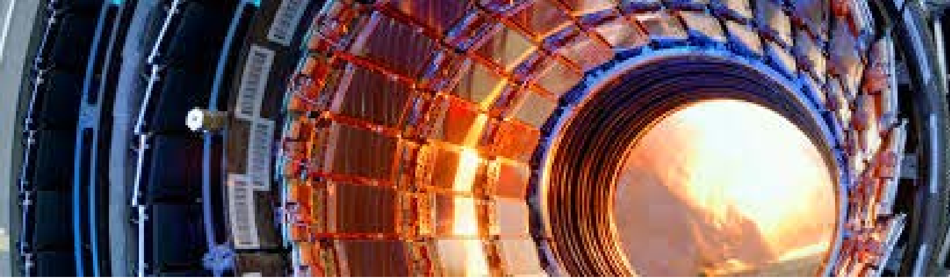
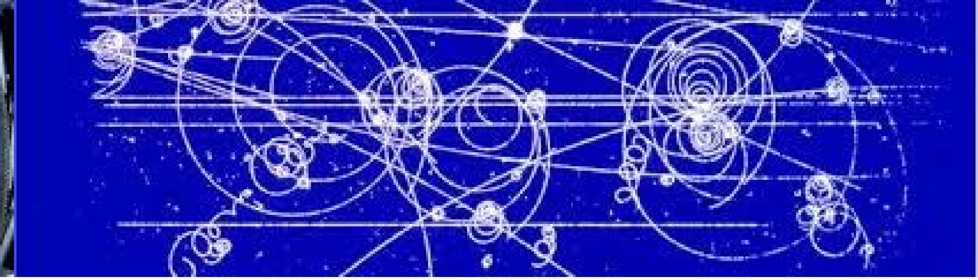




Building Virtual Scientific Computing Environment with Openstack

Yaodong Cheng, CC-IHEP, CAS

chyd@ihep.ac.cn

Contents

- Science facilities and computing requirements
- Cloud for scientific computing
- IHEP Cloud
- Conclusion

Contents

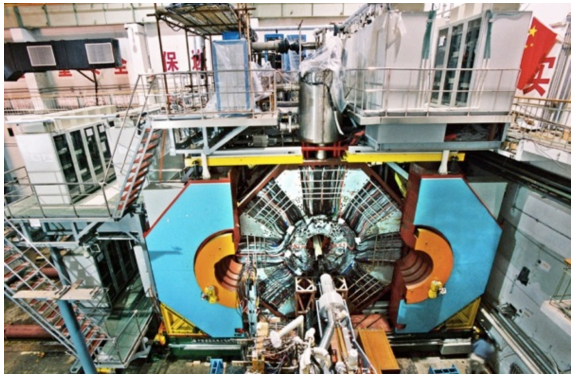
- Science facilities and computing requirements
- Cloud for scientific computing
- IHEP Cloud
- Conclusion

Large science facilities

- IHEP: The largest fundamental research center in China
- IHEP serves as the backbone of China's large science facilities
  - Beijing Electron Positron Collider BEPCII/BESIII
  - Yangbajing Cosmic Ray Observatory: ASg & ARGO
  - Daya Bay Neutrino Experiment
  - China Spallation Neutron Source (CSNS)
  - Hard X-ray Modulation Telescope(HXMT)
  - Jiangmen Neutrino Underground Observatory (JUNO)
  - Large High Altitude Air Shower Observatory (LHAASO)
  - Accelerator-driven Sub-critical System (ADS)
  - Under planning: BAPS, XTP, HERD, CEPC, …

BEPCII/BESIII

- 59 Institutions from China, US, Germany, Italy, Russian, Japan, …
- > 5PB in 5 years
- ~ 6000 CPU cores
  - simulation, reconstruction, analysis, …
- long-term data preservation
- data sharing between partners



Other experiments

- Daya Bay Neutrino Experiment
  - ~200TB per year
- JUNO: Jiangmen Neutrino Experiment
  - ~2PB per year
- LHAASO
  - ~ 2PB per year
- Atlas and CMS Tier2 site
  - 940TB disk, 1088 CPU cores
- CSNS, HXMT, …

100PB data in coming years!!

Computing resources status

- ~ 12000 CPU cores
  - ~ 50 queues, managed by Torque/PBS and HTCondor
  - difficult to share
- ~ 7PB disk
  - Lustre, Glustre, EOS, dCache/DPM, ...
- ~ 5PB LTO4 tape
  - two IBM 3584 tape libraries
  - modified CERN CASTOR 1.7

PC farm built with blades

Tape libraries

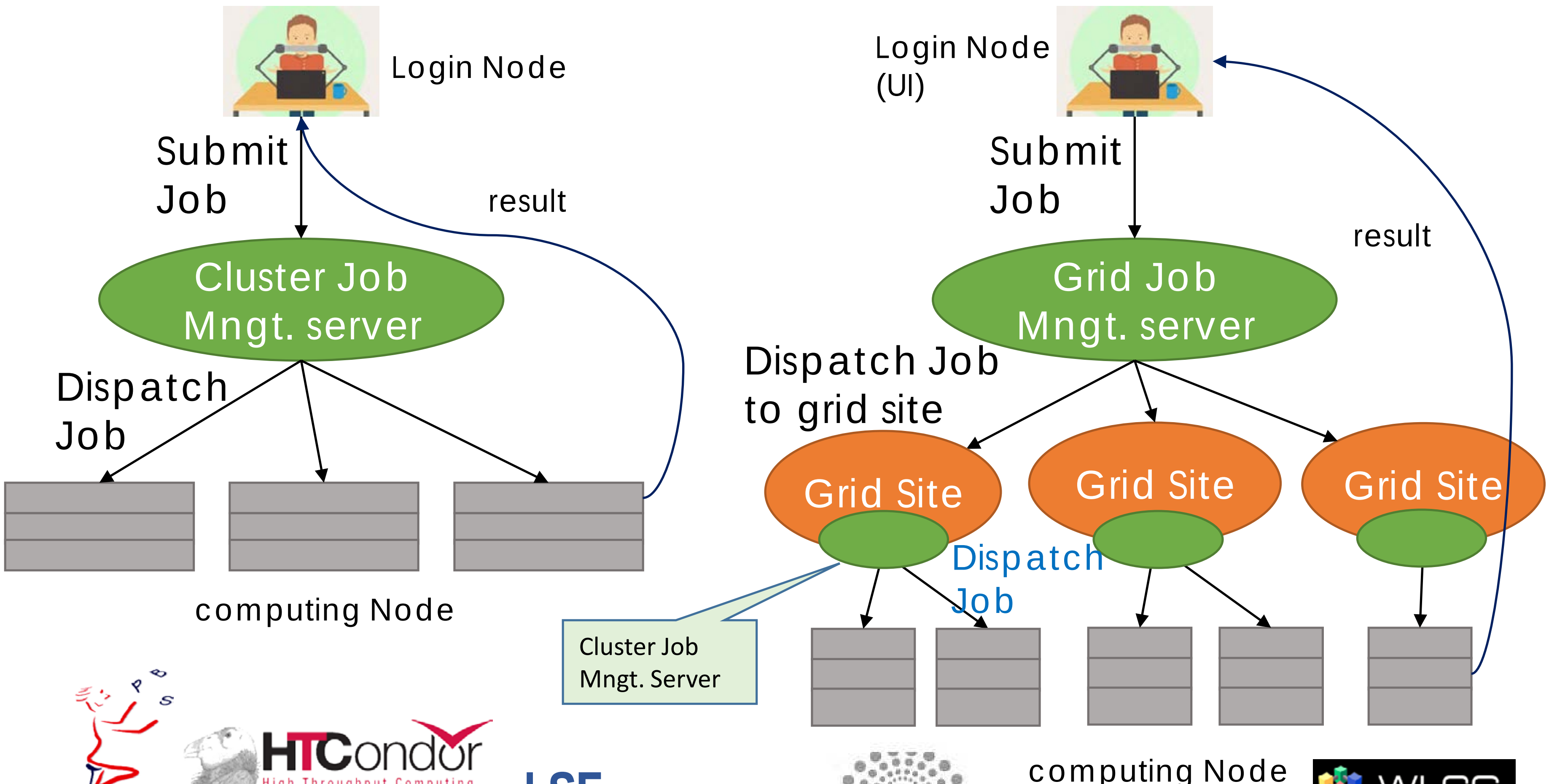
In the future, ...

- More HEP experiments, need to manage twice or more servers as today
- but, no possibility of significant increase in staff numbers
- Is cloud a good solution ?
- Is cloud suitable for Scientific Computing?
- Time to change IT strategy!!

Contents

- Science facilities and computing requirements
- Cloud for scientific computing
- IHEP Cloud
- Conclusion

Traditional Scientific computing

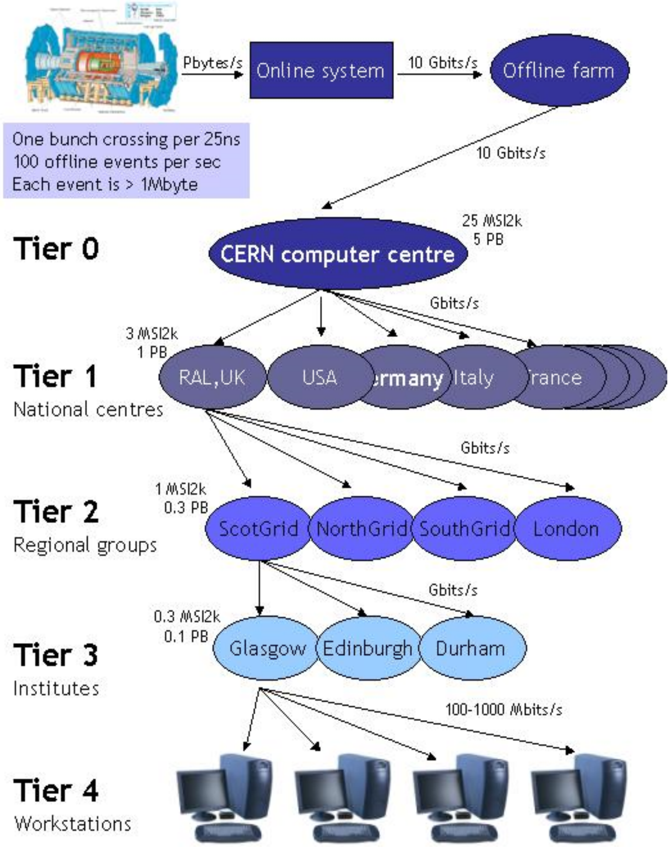


Cluster

Grid: Cluster of Cluster

WLCG Grid

- Worldwide Lhc Computing Grid
- 42 countries, 170 computing center's
- Tier0: CERN
  - First copy and First pass reconstruction
  - distribution of data to the Tier 1s
- Tier1: 15 large computer centers
  - sufficient storage and computing
  - distribution of data to Tier 2s
- Tier2: 150 sites
  - analysis, production and reconstruction
- Tier3: local computing resources
- Resource
  - 600k CPU core, 320PB disk, 300PB tape



Barriers for Adoption of Grid Model

Grid computing was never adopted outside (a part of) the scientific community

- a huge effort is needed to develop and maintain the non – industry standard middleware
- collaboration and tool sharing across experiments has always been difficult
- very difficult to use non-dedicated resources (all existing middleware is highly invasive) and the resource sharing issue still holds.

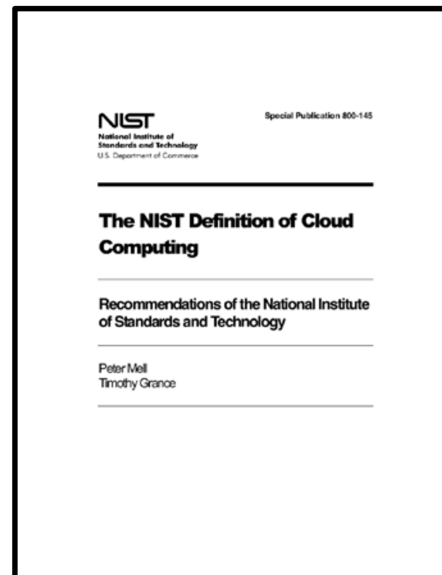
Grids are difficult to **maintain, operate and use**

↓

Virtualization and Cloud!!

What is Cloud Computing?

- Cloud computing
  - A technology to ease resource management, provisioning and sharing
  - An industrial standard technology
- NIST Definitions
  - Essential characteristics (5)
    - On-demand self-service
    - Broad network access
    - Resource pooling
    - Rapid elasticity
    - Measured service
  - Service models (3)
    - IaaS, PaaS, SaaS
  - Deployment models (4)
    - Public, private, community, hybrid


NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Special Publication 800-145

The NIST Definition of Cloud Computing

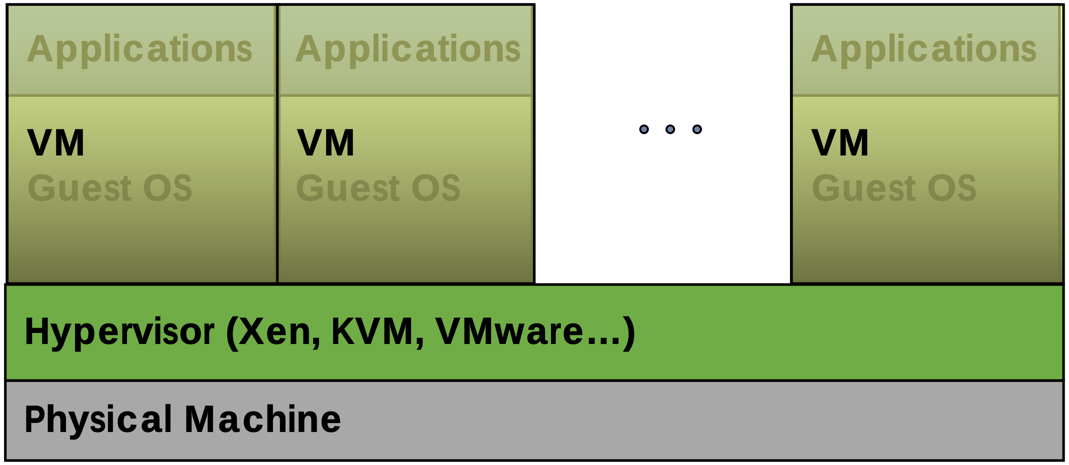
Recommendations of the National Institute of Standards and Technology

Peter Mell
Timothy Grance

http://dx.doi.org/10.6028/NIST.SP.800-145

Virtualization

- Virtualization – the key technology to improve the utilization of server
- Separation of Virtual Machines from Physical Infrastructure
- A VM is an isolated runtime environment (guest OS and applications)



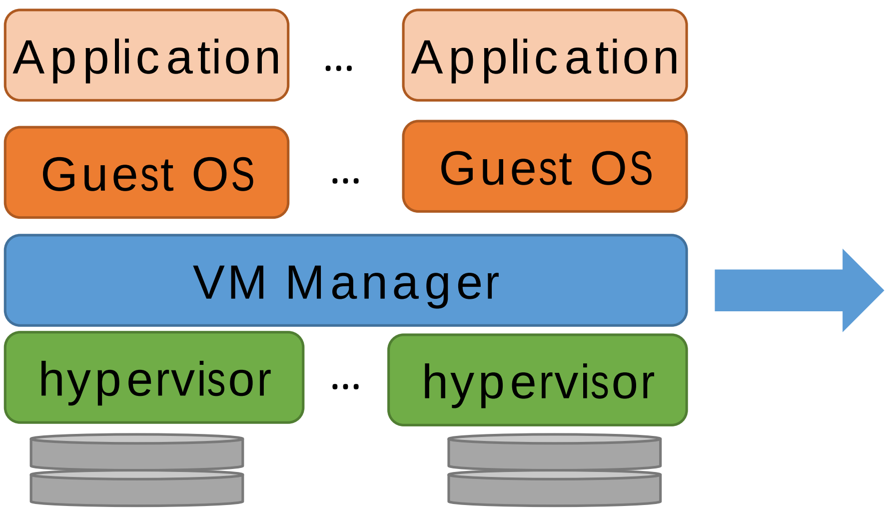
One Physical machine

Benefits of Virtualization Platforms

- Natural way to deal with the heterogeneity of the infrastructure
- Allow partitioning and isolating of physical resources
- Execution of legacy applications

VM Manager

- VM Manager creates a distributed virtualization layer
  - decouple the VM from the physical location
- Transform a distributed physical infrastructure into a flexible and elastic virtual infrastructure



Many physical machines

- Centralized management
- Balance of workload
- Server consolidation
- Dynamic resizing of the infrastructure
- Dynamic cluster partitioning
- Support for heterogeneous workloads
- On-demand provision of VMs

Openstack

- a popular VM manger, also called cloud operating system
- Controls large pools of compute, storage, and networking resources throughout a data center
- Hundreds of the world's largest brands rely on OpenStack to run their businesses every day

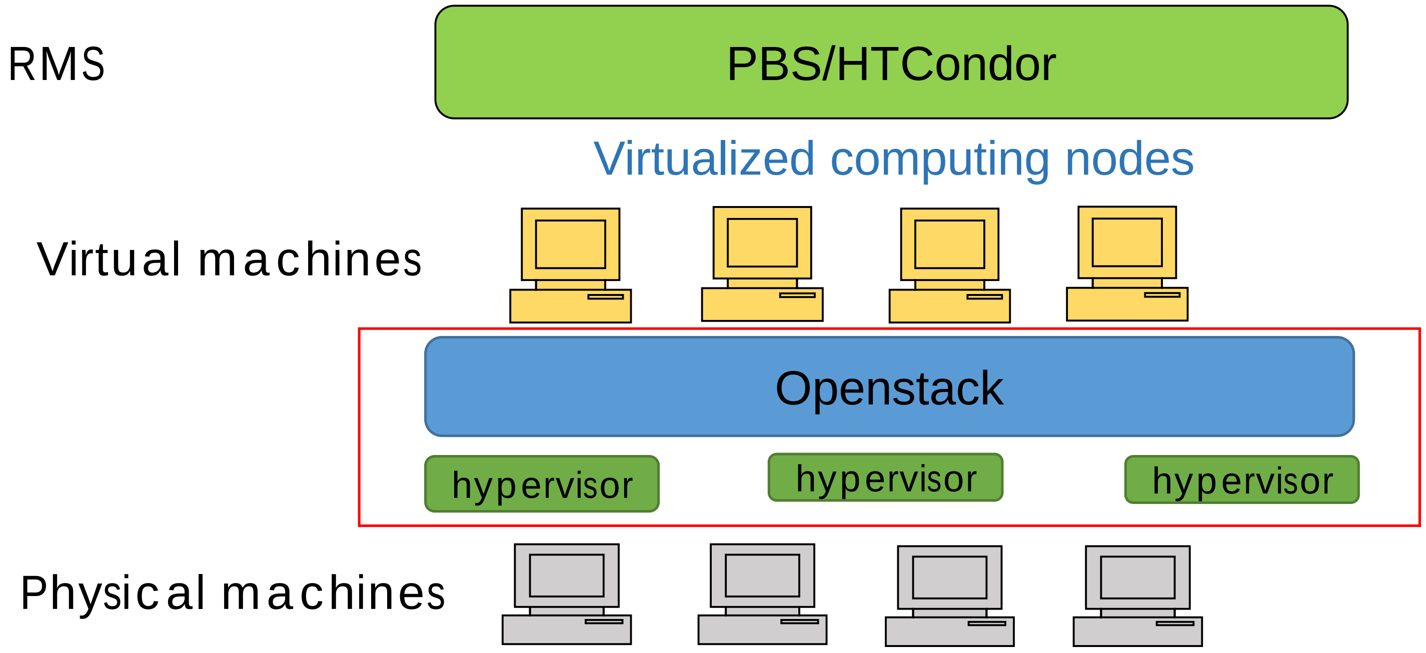
Austin (2010.1) → Bexar (2011.1) → Cactus (2011.2) → Diablo (2011.3) → Essex (2012.2) → Folsom (2012.2) →
Grizzly (2013.1) → Havana (2013.2) → Icehouse (2014.1) → Juno (2014.2) → Kilo (2015.4) → Liberty (2015.10) →
Mitaka (2016.4) → Newton (2016.10) → Ocata (Future)

Openstack release

Neutron heat openstack DEVSTACK Nova

Virtual Computing Cluster

- Computing nodes are installed in virtual machines
- Seamless integration with the existing middleware stacks.
- Completely transparent to the computing service and end users

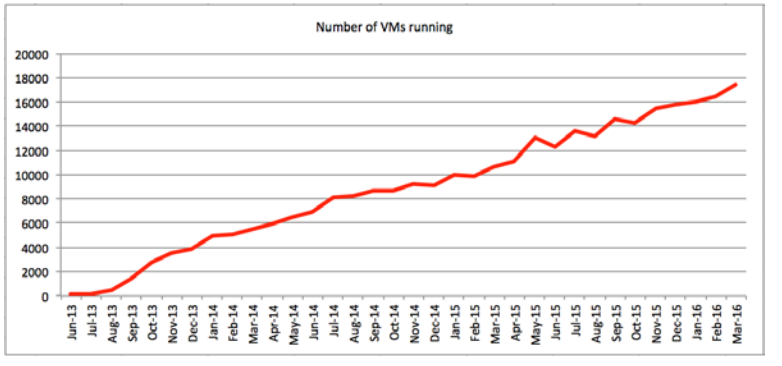


CERN Cloud

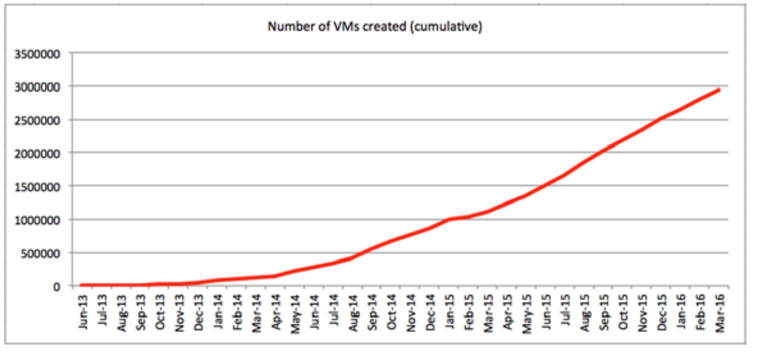
- CERN Cloud is one of largest virtual computing cluster
- CERN Cloud Service is one of the three major components in CERN IT's AI project
  - Policy: Servers in CERN IT shall be virtual
  - It will be a milestone for scientific cloud
- Based on OpenStack
  - Production service since July 2013
  - Already transition from Juno to Kilo
  - Nova, Glance, Keystone, Horizon, Cinder, Ceilometer, Heat

CERN Cloud in Numbers (1)

- 5’800 hypervisors in production (6m ago: +25%)
  - Majority qemu/kvm now on CC7 (~150 Hyper-V hosts)
  - ~2’100 HVs at Wigner in Hungary (batch, compute, services)
  - 370 HVs on critical power (+50%)

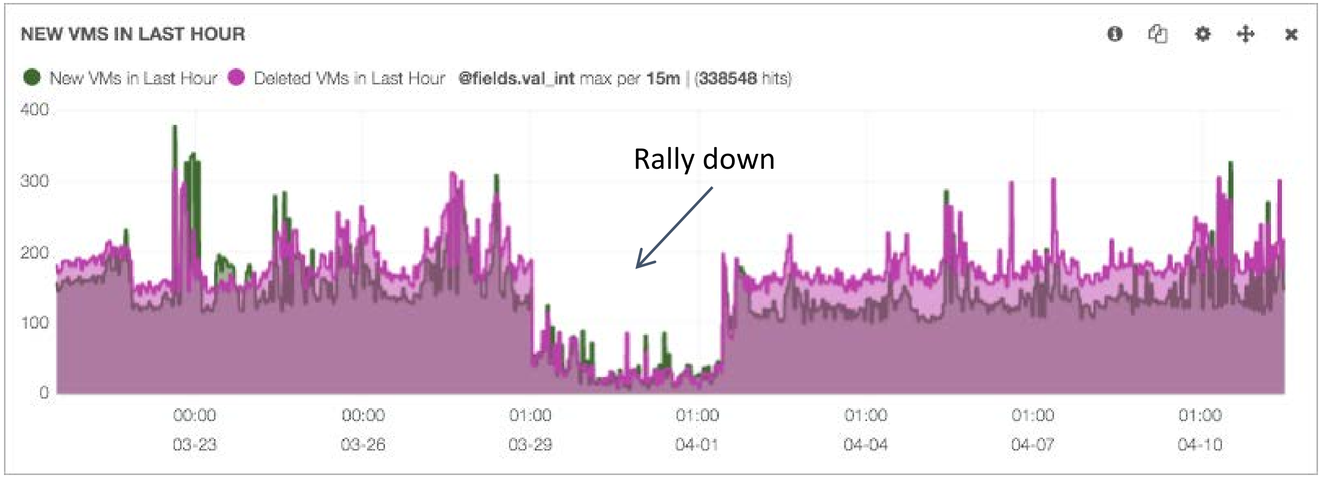


- 155k Cores (+30k)
- ~350 TB RAM (+100TB)
- ~18’000 VMs (+3’000)



- To be increased in 2016!
  - +57k cores in spring
  - +400kHS06 in autumn

CERN Cloud in Numbers (2)



Every 10s a VM gets created or deleted in CERN cloud!

- 2’700 images/snapshots (+700)
  - Glance on Ceph
- 2’200 volumes (+700, uptake doubled)
  - Cinder on Ceph (& NetApp) in GVA & Wigner

ceph

Performance problem

- HS06: HEP-wide benchmark for measuring CPU performance
- CERN: The “20% overhead” problem
  - HS06 rating of full node VMs was ~20% lower than on the underlying host
    - Full node VMs are needed to the limit of the total number of hosts in LSF
  - Smaller VMs behaved much better: ~8%
    - The sum of simultaneous HS06 runs on 4x 8-core VMs on a 32-core host
    - Better, but still pretty high
- IN2P3 reported significant performance penalties for ATLAS MC jobs when EPT* was switched on
  - 26% vs. 6% in wall clock time for EPT on vs. EPT off compared to bare metal
  - Surprising as EPT is supposed to make things faster

***EPT:** Extended Page Tables is Intel’s implementation of a hardware-assisted virtualization technology for page table management (secondary address translation or nested pages). AMD’s implementation is called RVI (Rapid Virtualization Indexing).

Virtual machine performance test (1)

- BES simulation job
  - Same number of jobs running on physical vs VM, each VM runs one job.
  - The number of VM on physical machine(24 cores):1,12,24
- Test environment
  - Virtual machine:
    1 CPU cores , 2GB memory
  - Physical machine:
    24 CPU cores , 16GB memory
- Test Result:
  - 1 job :Running time penalty on VM is about 3%
  - 24 job: ~2%

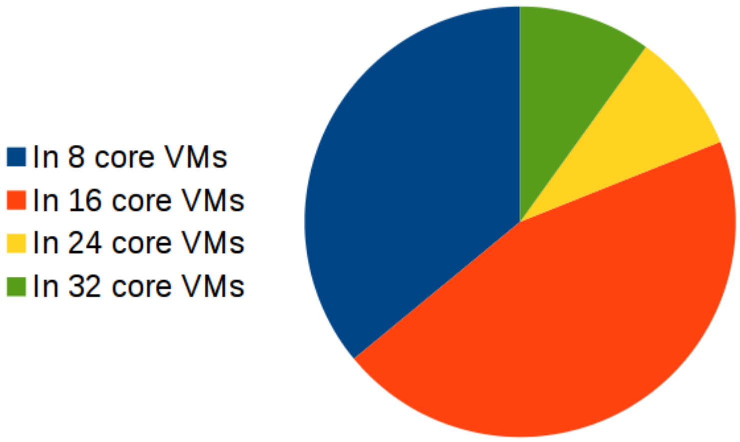
Job	alltime	usertime	CPU	slow
1-pm	3318.51	3303.13	99.5%	
1-vm	3427.12	3391.56	98.9%	3.3%
12-pm	3761.75	3740.76	99.5%	
12-vm	3862.58	3828.31	99.1%	2.7%
24-pm	3786.45	3750.01	99.5%	
24-vm	3870.08	3829.19	98.9%	2.2%

Virtual machine performance test (2)

- BES reconstruction job
  - Same number of jobs running on physical vs VM, each VM runs one job.
  - The number of VM on physical machine(24 cores):1,12,24
- Test environment
  - Virtual machine:
    1CPU cores , 2GB memory
  - Physical machine:
    24CPU cores , 16GB memory
- Test Result:
  - 1 job :Running time penalty on VM is about 3%
  - 24 job: about 6%

Job	alltime	usertime	CPU	slow
1-pm	6409.75	6394.53	99.7%	
1-vm	6642.33	6632.84	99.3%	3.6%
12-pm	7333.58	7305.78	99.7%	
12-vm	7639.41	7583.24	99.4%	4.2%
24-pm	7366.25	7333.02	99.7%	
24-vm	8564.37	8286.49	97%	5.7%

Core Distribution & Effective Impact



Take into account the core distribution over VM flavors to determine real loss

Small VMs (8 or 16 cores)	**Large VMs (24 or 32 cores)**
Closer to bare metal performance	Better memory flexibility
Closer to LSF* limitations (less of a problem after instance split)	Higher performance penalty

*LSF is a commercial cluster job management system used at CERN

Contents

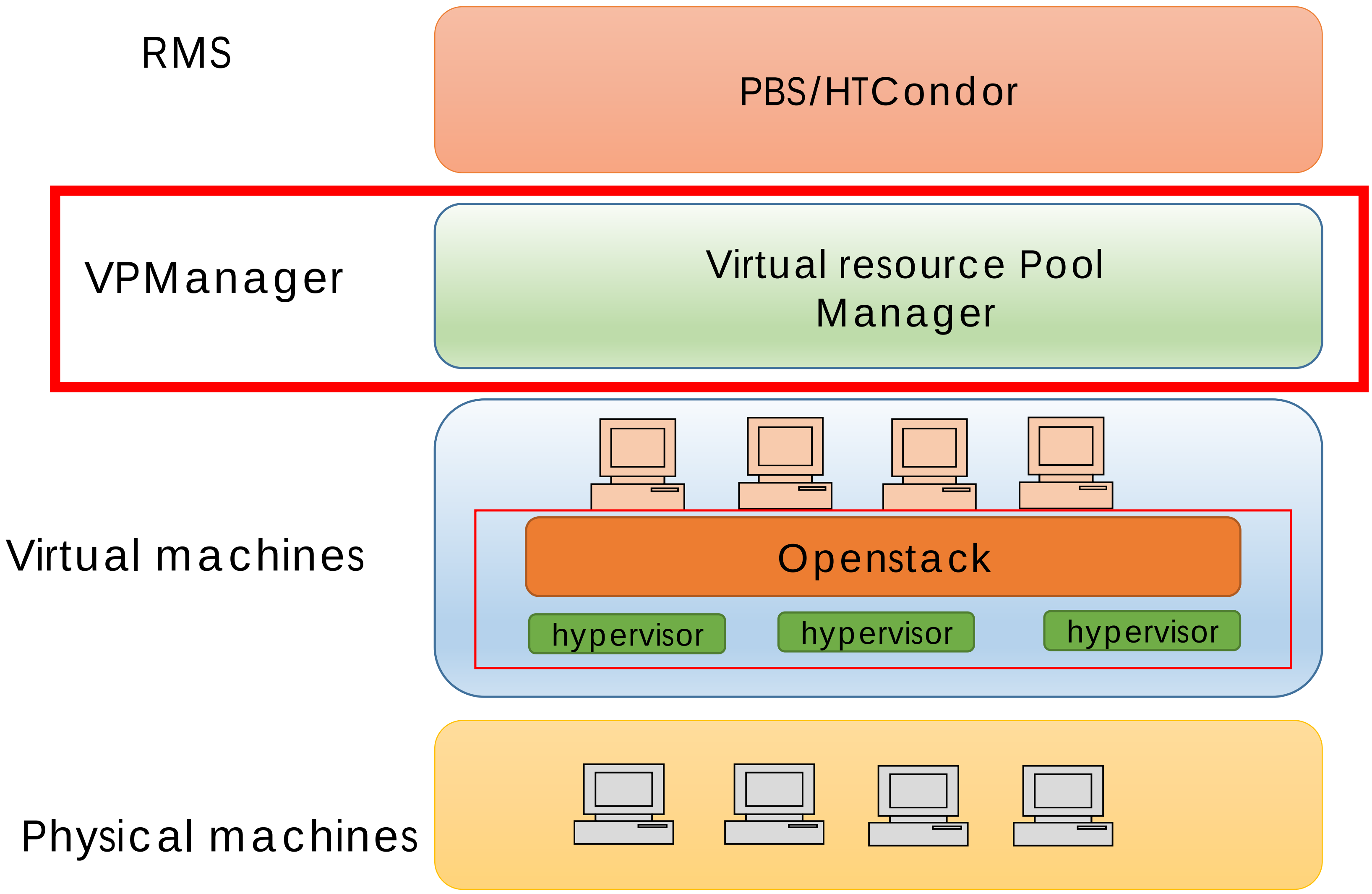
Motivation

- More HEP experiments, need to manage twice or more servers as today
- Low resources utilization
  - Annual Utilization of computing resources is less than 50% on average
- Computing resources are non-shared
  - Every experiment such as BESIII,YBJ, has its own computing machines
- Cloud widely accepted in scientific and industrial domain
  - Improve operational efficiency
  - Improve resource efficiency
  - Improve responsiveness

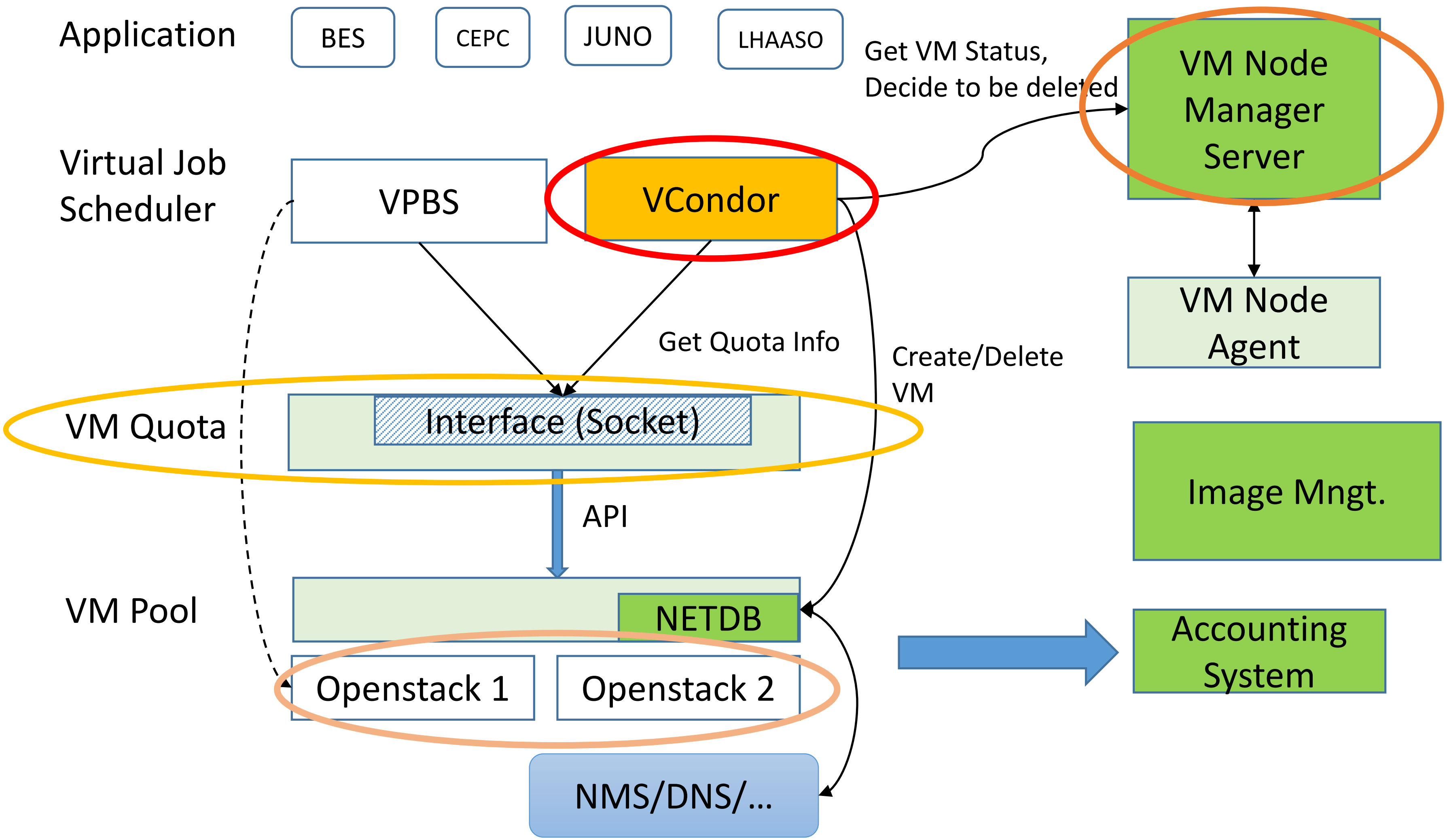
IHEPCloud: a Private IaaS platform

- Launched in May 2014
- Three use scenario
  - User self-Service virtual machine platform (IaaS)
    - User register and destroy VM on-demand
  - dynamic Virtual Computing Cluster
    - Job will be allocated to virtual queue automatically when physical queue is busy
  - Distributed computing system
    - Work as a cloud site: Dirac or other applications call cloud interface to start or stop virtual work nodes

Dynamic virtual computing cluster



VPManager(Virtual resource Pool Manager)

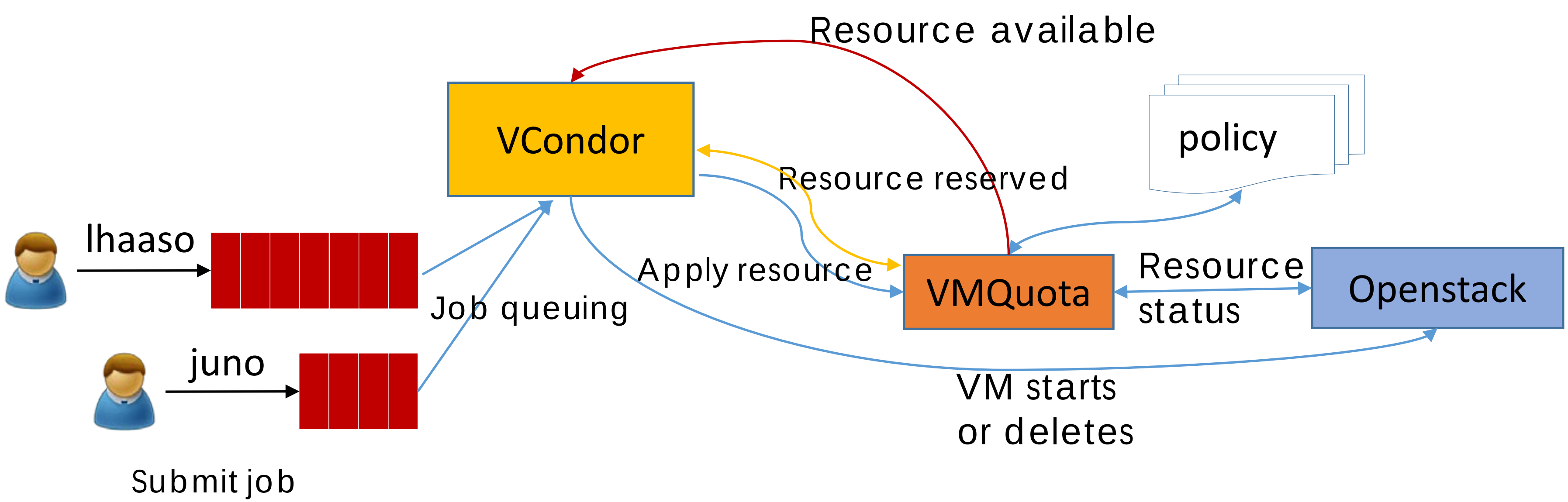


VPManager components

- VM Pool
  - manages one or more openstack deployments, which hides the detailed information of openstack from upper applications
  - makes it possible to deploy multiple and different versions of openstack
- VM Quota
  - checks the information of VM Pool and requirements of different applications to allocate or reserve resources.
- Virtual job manager, VPBS and VCondor
  - checks the status of different queue and get the available VM number and create new VMs or destroy existing VMs.
- VM node manager
  - checks and controls all the VM run environment such network status, affiliated job queue by an agent running in the virtual machine
- Accounting system
  - keeps all the usage information of each virtual machine and generate bills to user

VCondor

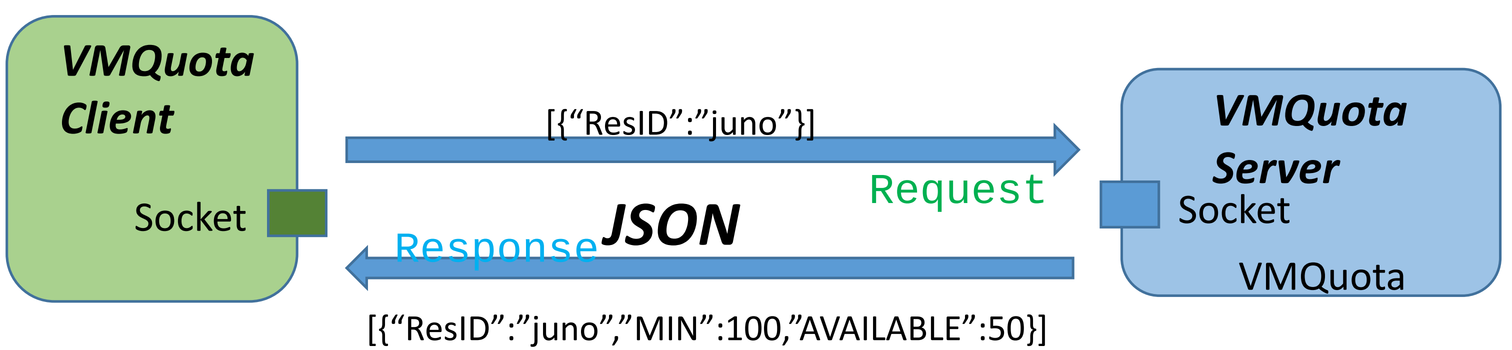
- VCondor
  - Resource allocation as demand
  - Using HTCondor
- VMQuota
  - Resource quota management system



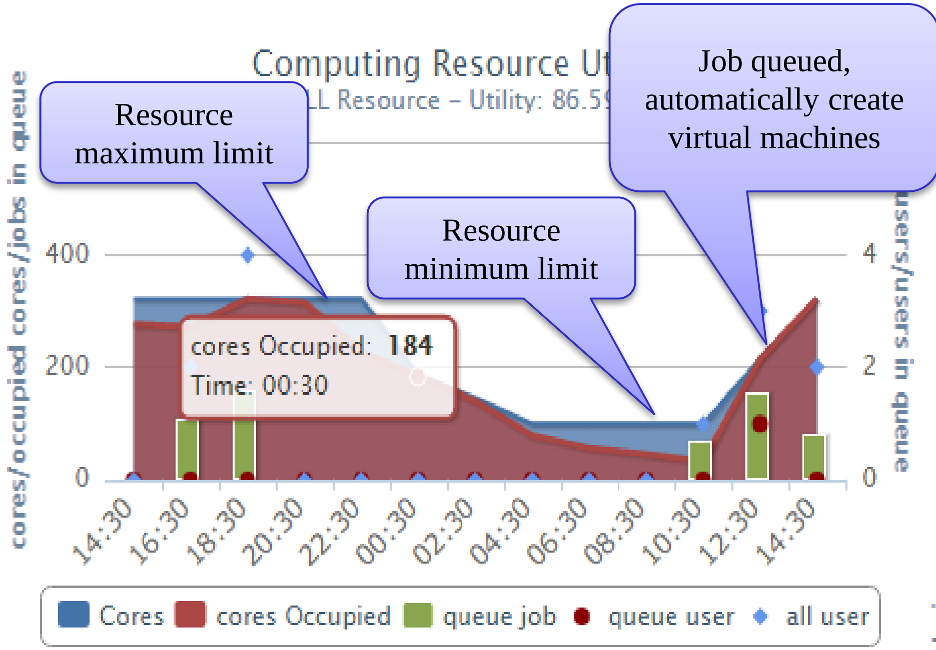
VMQuota

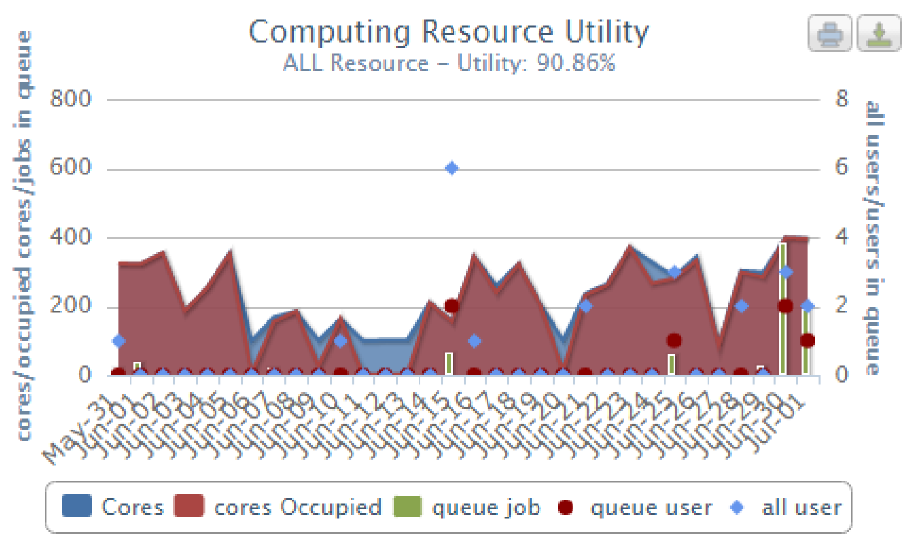
- Resource Quota management for different experiments
- Different experiments have different resource queues
- Allocate and reserve resources for different queues

Queue Name	Min	Max	Available	Reserve time(s)
BES	100	400	200	600
JUNO	100	300	200	600



VCondor monitoring

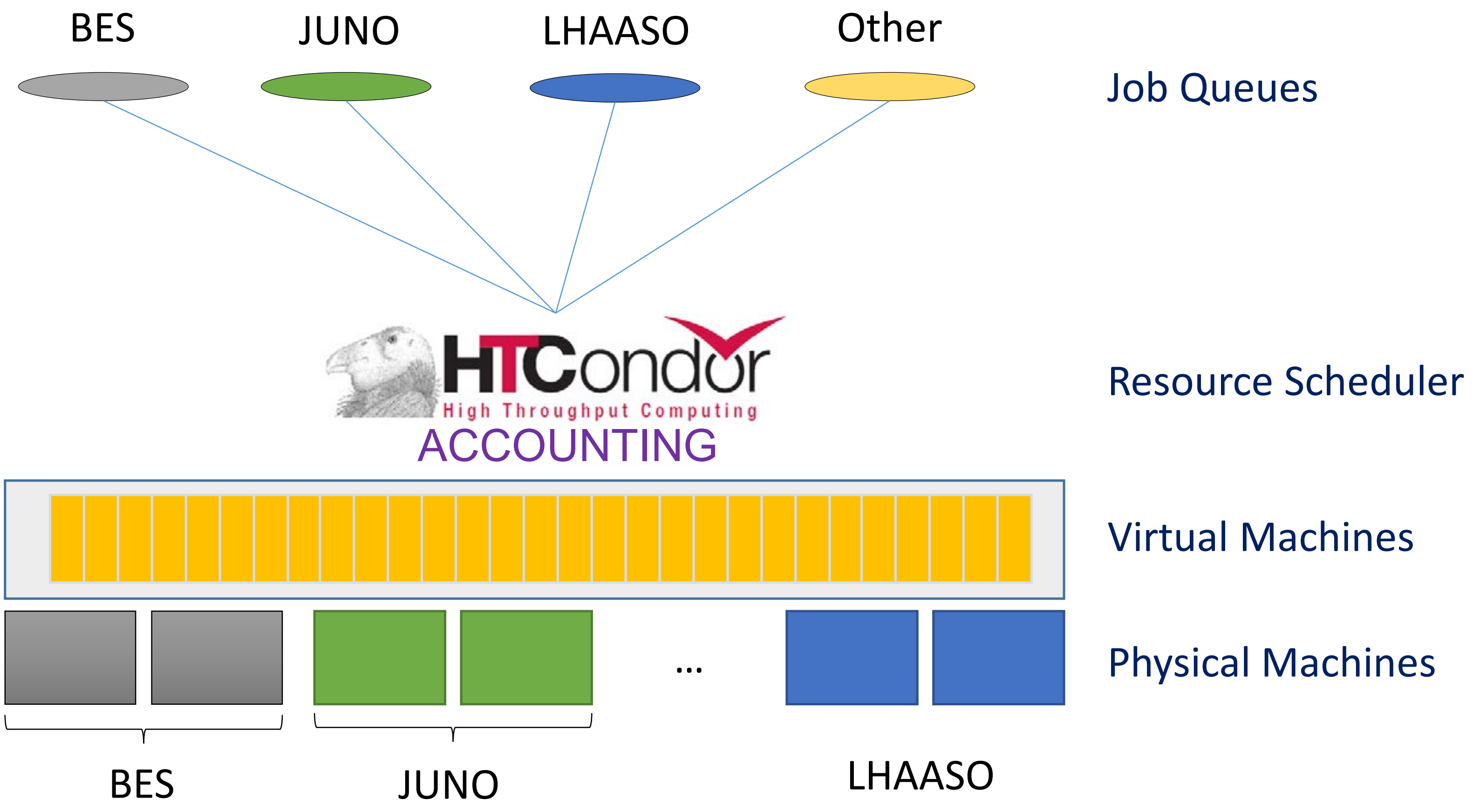




Future deployment plan

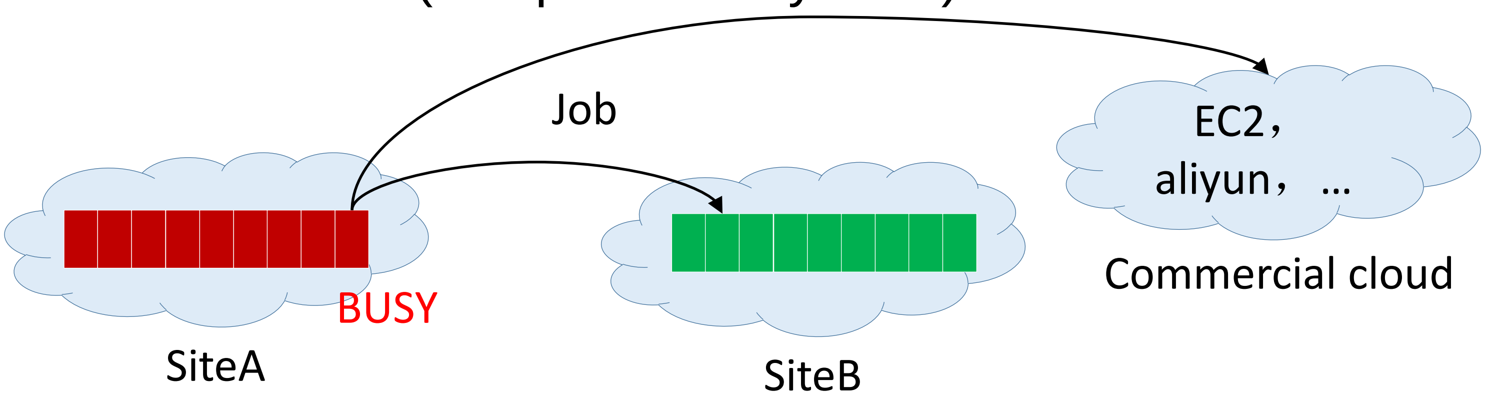
- Four layers
  - 1st layer: Physical machines
    - bought and owned by different experiments
  - 2nd layer: Virtual machines
    - Shared resource pools, not belong to any experiments
  - 3rd layer: Resource scheduler
    - Dynamically allocate resources to different experiments depending on the task list
    - Resource allocation policies to balance the resource sharing and physical machine invest
    - Detailed accounting for resource use
  - 4th layer: job queues
    - Different job queues for end users of different experiments
    - Same way to use as traditional cluster

Deployment architecture

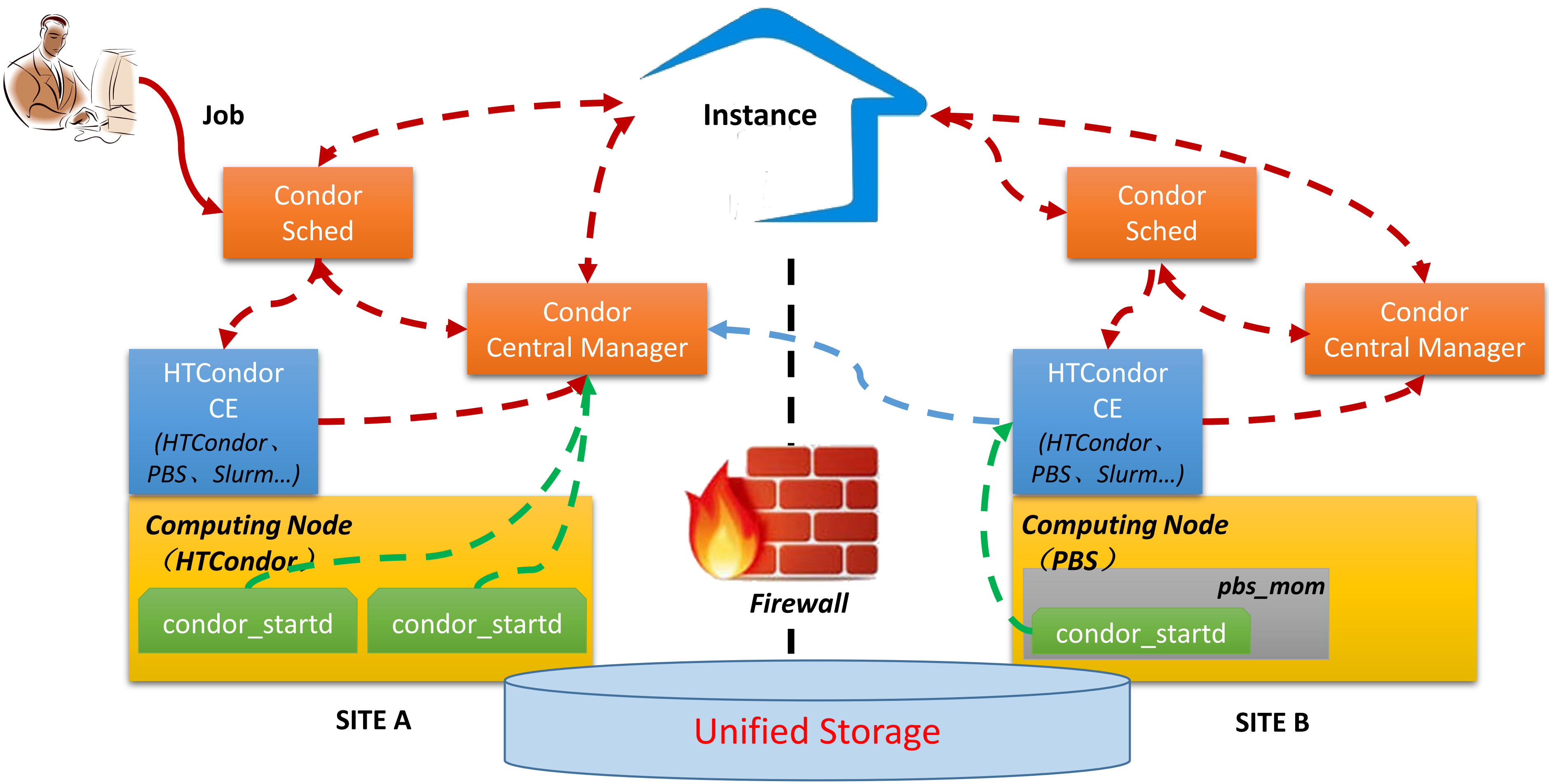


China HEP Community Cloud Plan

- Sharing resource across different sites
  - improve resource utilization
- Different data management solution from grid
  - Job dispatched to site where data is located in grid
  - Impossible to subscribe data in dynamic cloud
  - Same storage / file system view across different sites
  - Streaming and cache data in cloud
  - EOS or LEAF (our planned system)



Distributed Cloud deployment



Conclusion

- Cloud computing is widely accepted by industrial and scientific domain
- Scientific computing are preparing the move to cloud
- The performance penalties is acceptable
- IHEPCloud aims at providing self-service virtual machine platform and virtual computing environment
- More resources will be added to IHEPCloud

Thank you!

Any Questions?