Cloud computing at CERN

Experiences from lxcloud

"wlcg-teg-workload-mgmt F2F CERN, 3/11/2011

Ulrich Schwickerath

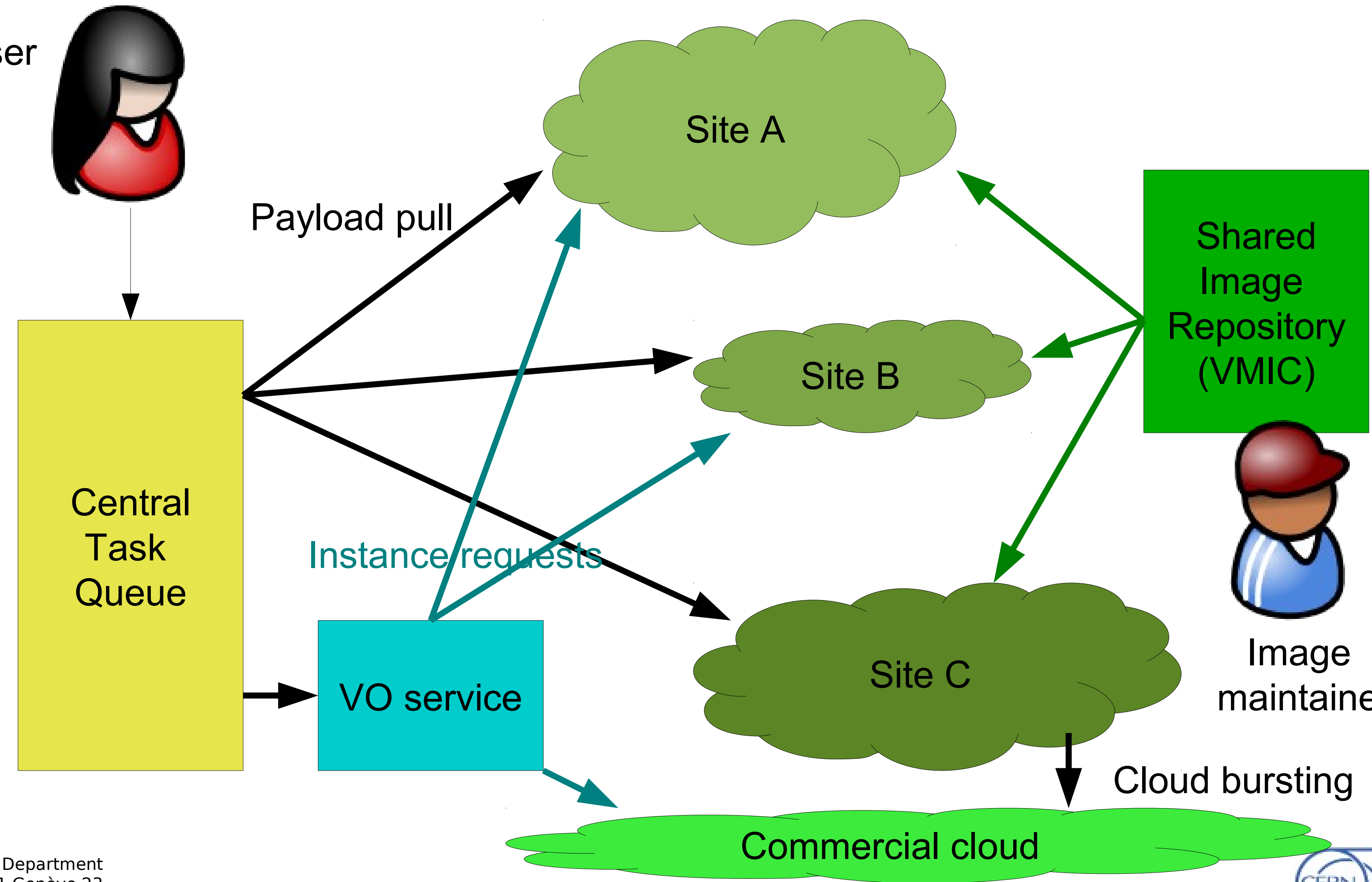
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Outline

- A new model ?
- Technology
  - History
  - Design principles
  - Status
- Applications and experiences
  - Virtualized batch systems
  - IaaS services for users
- General considerations
  - Performance
  - Strategy
  - ...

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

A new model ?


User
Payload pull
Site A
Site B
Site C
Shared Image Repository (VMIC)
Central Task Queue
Instance requests
VO service
Image maintainer
Cloud bursting
Commercial cloud

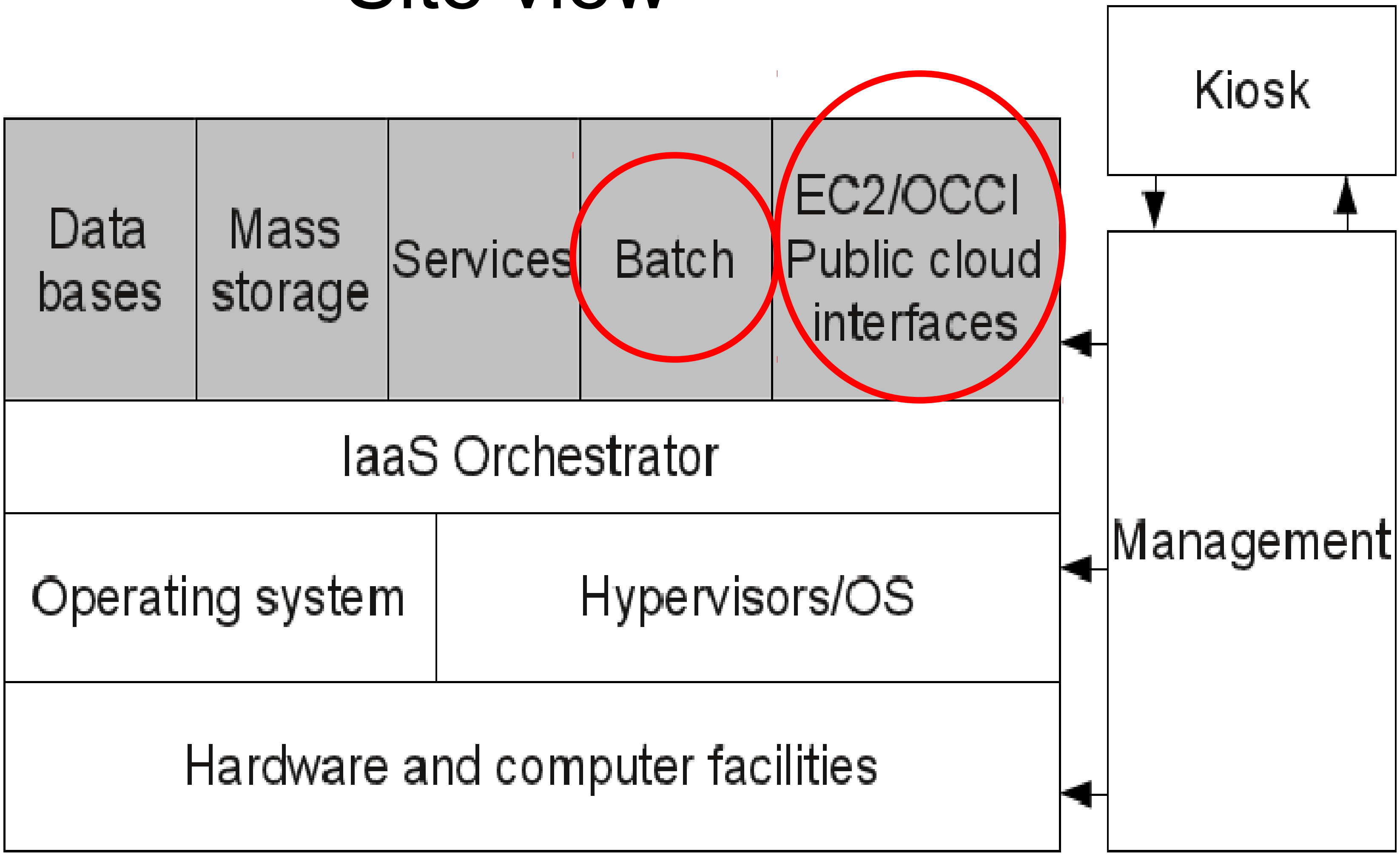
A new model ?

Not that different from current pilot job frame works

!

A new model ?

Site view



CERNs internal cloud

Technology

From the idea …

… to a working prototype

What is lxcloud ?

CERN IT Department

History

2008
Discussions with Platform Computing on batch efficieny issues

2009
Feasibility studies, first proof of concept, focus on batch

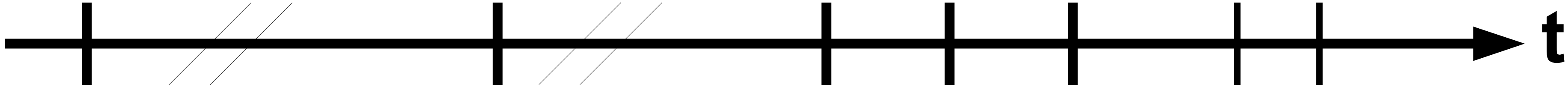
8/2010
Large scale tests on 480 machines, 16000 VMs for LSF testing

12/2010
First 96 virtual batch nodes in full production
KVM, SLC5, ISF and ONE

3/2011
Increase to 384 virtual batch nodes EC2 system tests start

8/2011
SLC6 migration
Increase to 432 VMs

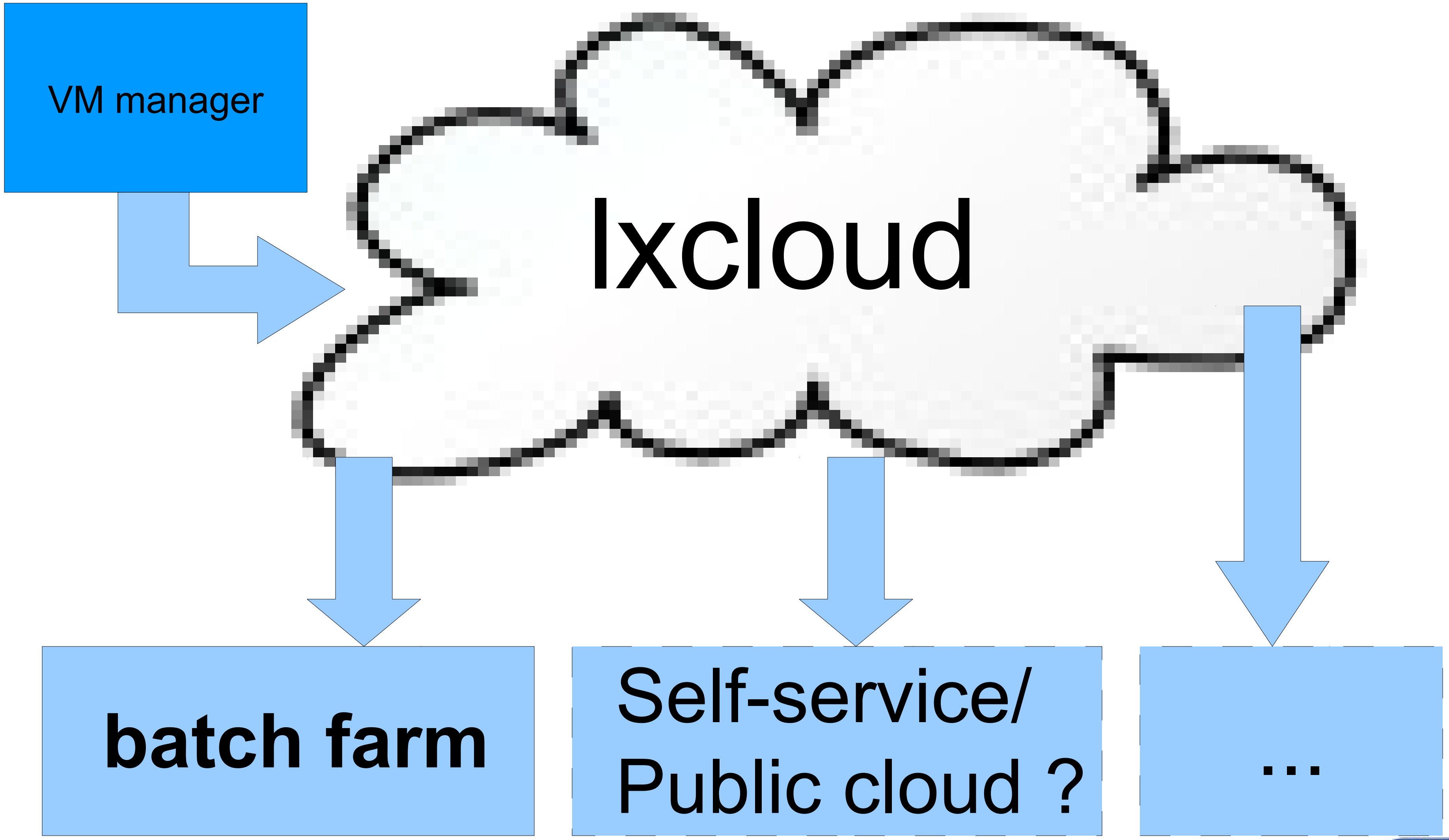
10/2011
ONE3.0
Unification with test system



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

What is lxcloud ?

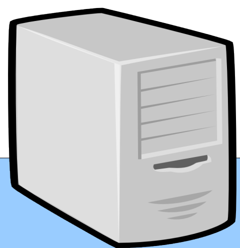
CERN IT Department



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

What is lxcloud ?

- Highly scalable, Linux (KVM) based cloud-like infrastructure
- Optimized for efficiency/speed



Physical resources

Local images on LV

LV snapshotting

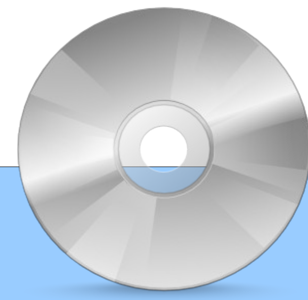


Image repository
and
image distribution
System

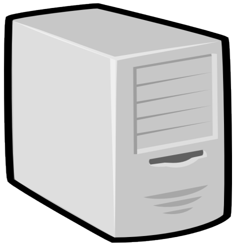
VM managing
System

(OpenNebula 3.0)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Resource pool details

- **Quattor managed** pool of resources (lxcloud)
- **Hardware**: (cheap) CPU server type, local disks
- **LANDB integration**
  - Pre-allocation of VM “slots” in landb
  - Hypervisor “knows” the name of its guests
- **Disk management**
  - Use of LVM snapshots
  - All free disk space in one big LV
  - Pre-stage raw images on LV on the hypervisors
  - Fast instantiation of VMs using LV snapshots



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Image management

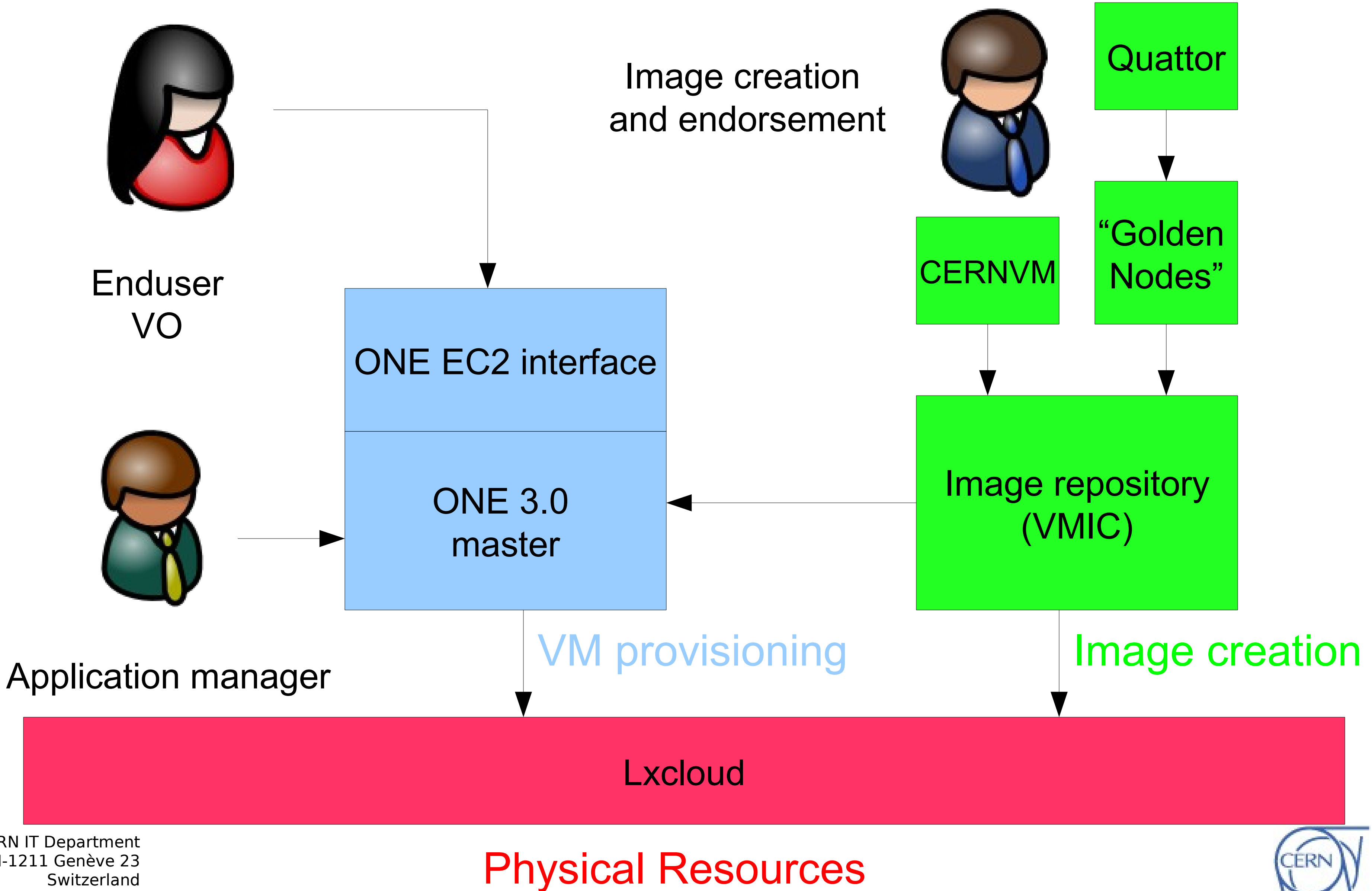
- **Central image catalogue (VMIC)**
  - Close collaboration with HEPiX
  - No direct user access/user images
  - Images require endorsement by IT
- **Image distribution system**
  - Central image repository of <u>trusted</u> images
  - Fast distribution using Bit-torrent (rtorrent)
  - Pull model: Hypervisors ask if there are updates
  - Transparent update of images using LV tools
  - Hypervisors advertise existing images

Virtual machine orchestration

- OpenNebula.org

  - Many new features and improvements in 3.0
  - Group management etc
  - Allows to merge test and production instance
- OpenStack
  - is a new interesting product worth to be checked

Putting it all together



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

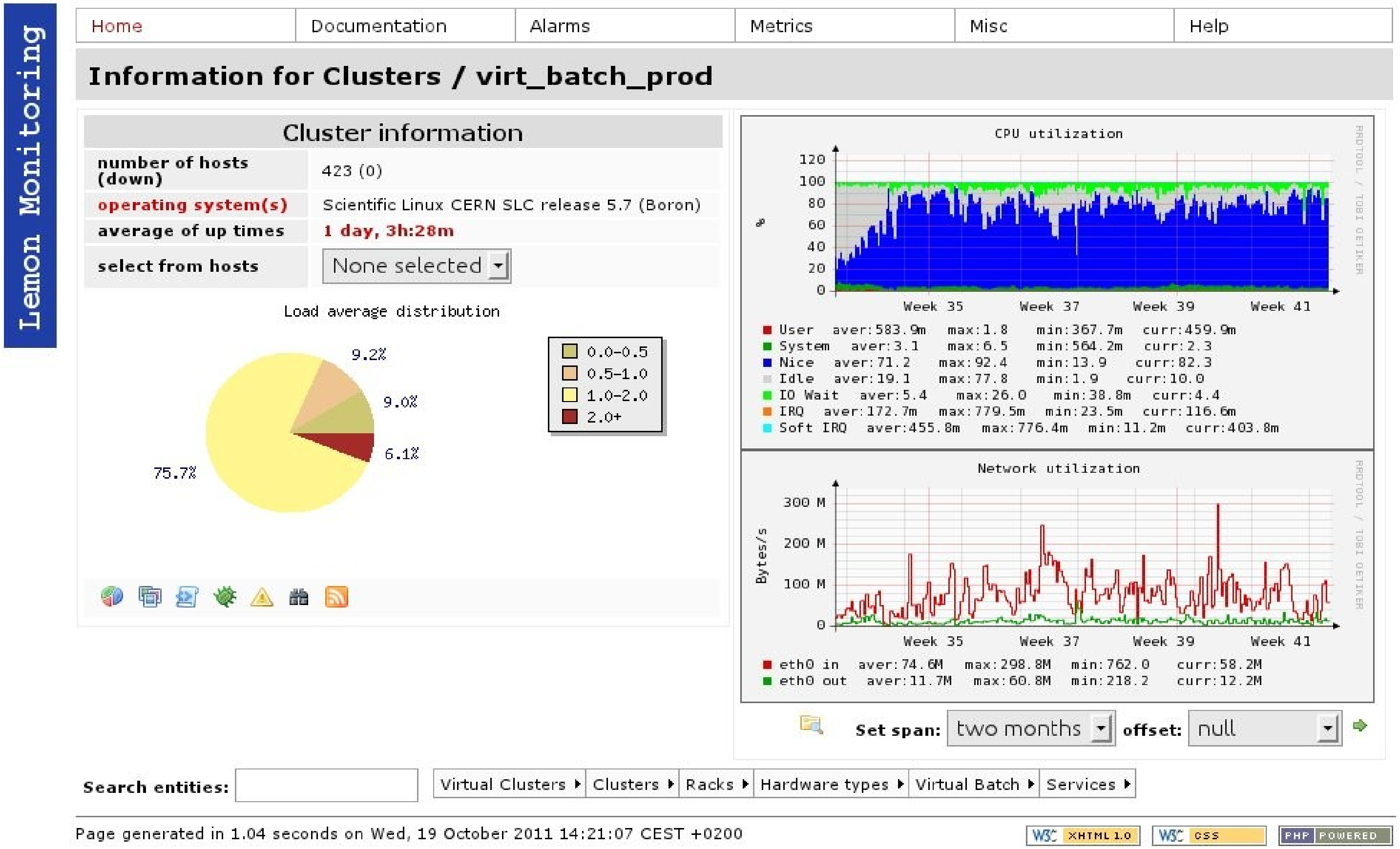
Applications

What can you do with it...

… applications and experiences

Virtual batch at CERN

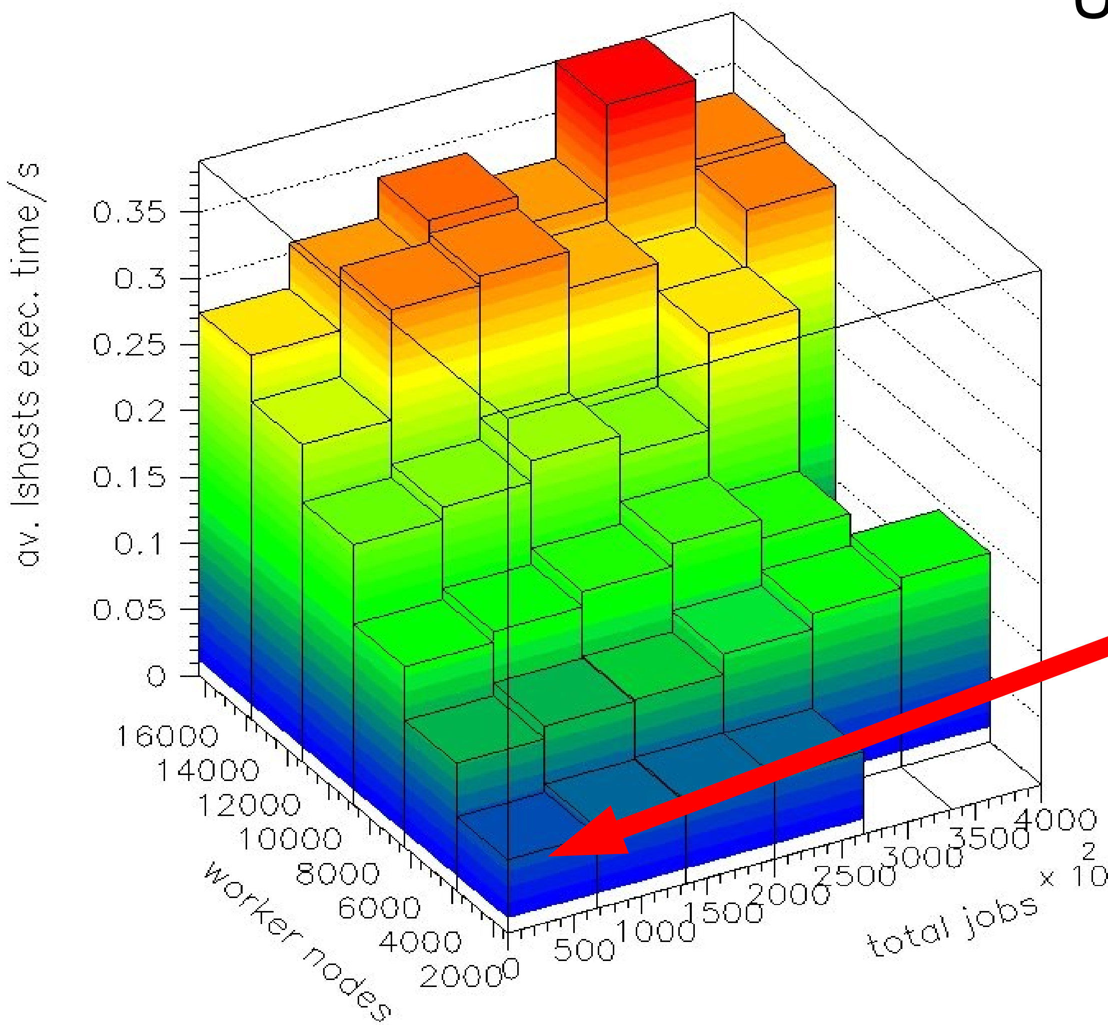
In production since December 2010 at CERN



48 hypervisors, 432 VMs in total, integrated in lxbatch

Scalability studies of 2010

Batch system tests: resource layer



Up to **15000** nodes
Up to **400000** jobs

Probed up to more than **3x** of what is officially supported by LSF

Current production system

A single batch system instance works up to 5-10k worker nodes only

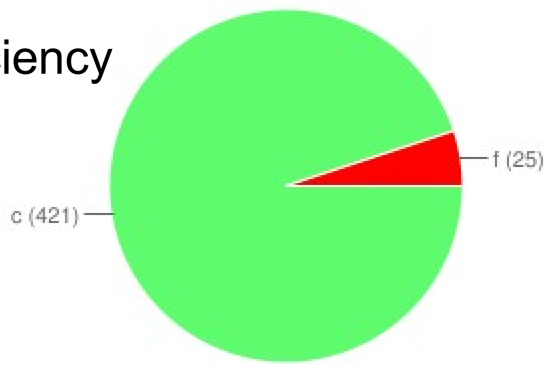
Public cloud access tests (ATLAS)

CERN IT Department

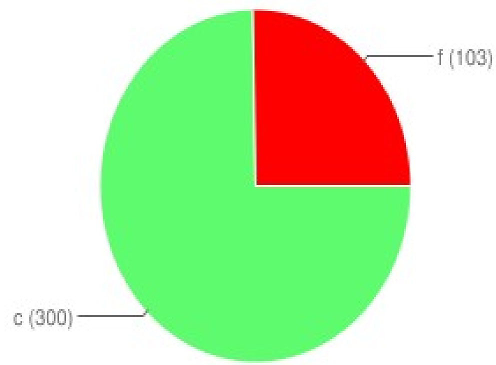
EC2 lxcloud	Direct submission	Regular GRID jobs

Efficiency

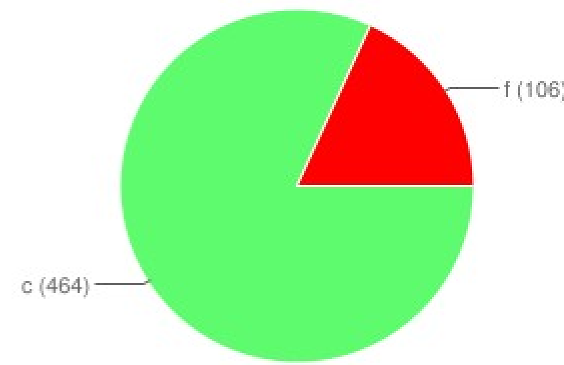
ANALY_CERNVM_test



ANALY_CERNVM_test

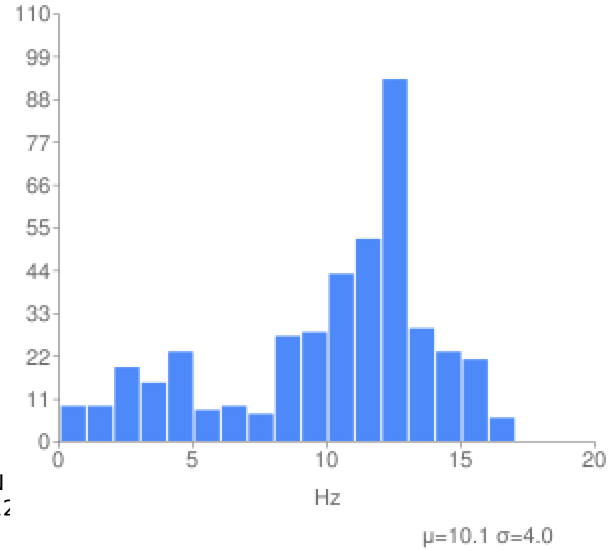


ANALY_CERN

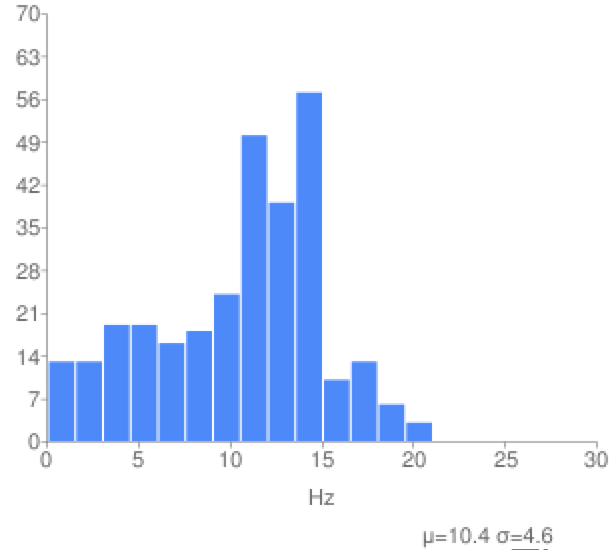


Event rate

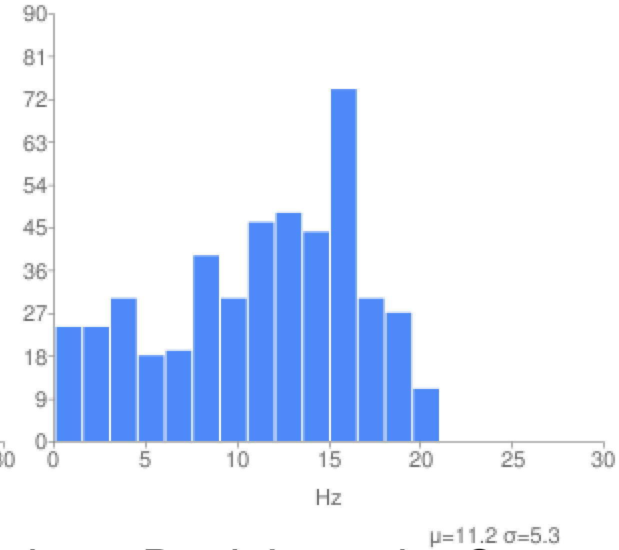
ANALY_CERNVM_test Events/Wallclock(s)



ANALY_CERNVM_test Events/Wallclock(s)



ANALY_CERN Events/Wallclock(s)



Thanks to Daniel van der Ster

Lxcloud resources

Batch virtualization:

- Works nicely on 48 hypervisors, automating many operational tasks
- Further grow limited by LSF scalability (see: next slide)
- Requires better master hardware, and a review of the current design

Public cloud:

- Low usage for the time being
- No point in increasing resources

Current plans for lxcloud

CERN IT Department

Status: mission accomplished

Next steps:

- Continue service consolidation (merge, almost done)
- Exploit some new features of ONE 3.0 for EC2 access?
- Learn how to use the EC2 interface effectively
- Currently **no plans** to increase resources

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

General considerations

Stuff that is missing...

... or incomplete

HEPiX virtualization working group

The HEPiX community has already been very active in the field of virtualization. The results can be very relevant for us.

Fields of activity include:

- Policies and security
- Image exchange (between sites)
- Establishing trust relationships
- Contextualization of images

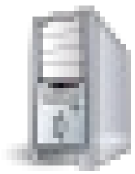
General considerations (and caveats)

- Performance
- Limitations
- Policies and Security
- Resource sharing and management
- Accounting and billing

Performance ...

CERN IT Department

Cost for virtualization (KVM, SLC5/6)



CPU : 2%-5%

Network : small

I/O : depends/difficult
20%-30% is realistic

See also:

https://indico.cern.ch/contributionDisplay.py?sessionId=6&contribId=6&confId=138424

Technical limitations

- Reasonable I/O performance requires local disks
  - Files better than LV
  - LV have nice features which files don't have ...
- AFS inside guests
  - Good I/O requires locally attached disk
  - IP/MAC binding
  - Preserve AFS cache during interventions
  - Minimize downtime of guests

Note: this sets constraints on live migrations

Policies and security

Adapt **HEPiX** policies for virtual machine images and security

- Only trusted images (no user images)
- No secrets in the image
- No root access for users by policy
- Contextualization is used to put in secrets and site dependencies

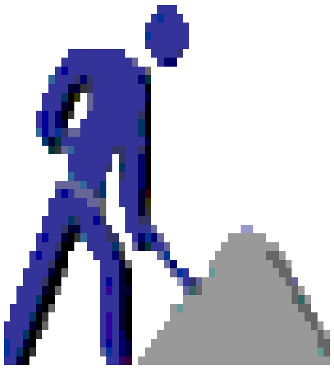
https://documents.egi.eu/document/771

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Policies and security

IaaS (EC2 self-service) access

- User/password access, as offered by OpenNebula
  - X509 authentication possible
  - Based on user DNs
  - Good enough ?
- Selected users only
  - proof of concept



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Policies and security

IaaS self-service access protocols:

EC2 :

- not a standard but invented by Amazon
- Works for their commercial clouds as well
- Many clients available

OCCI(*) :

- Being developed (http://occi-wg.org/)
- offered by most VM orchestrators
- May not work with commercial providers

(*) Open Cloud Computing Interface

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Resource sharing

Resource sharing happens on the resource layer (hypervisors)

- Requires sophisticated schedulers for virtual machines
  - Current implementations are rather basic
  - No concept of fair share but hard quota
- Requires that VMs are given back when no longer needed
  - Ensure elasticity
  - Example batch: adapt worker node composition to needs

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Resource sharing

IaaS clouds and (non) infinite resources:

- IaaS Clouds will reject new requests if no resources are available
- Queue up new incoming requests ?
  - Could do fair share and accounting on these queues
  - Solutions exist at other sites who schedule VMs via a batch system but this couples the infrastructure to an application and is site specific
- Or do it the cloud way
  - Pay for what you get, and if you don't pay ...

Resource allocations

Need a way to allocate resources to different areas

- Services for Experiments
- IaaS access authorization and quota
- Traditional batch system shares

Idea: develop a high level resource allocation framework. We call it CloudMan:

https://indico.cern.ch/contributionDisplay.py?sessionId=6&contribId=5&confId=138424

(work in progress)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Accounting and billing

To be decided ...

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Accounting and billing

CERN IT Department

Options 1: let the batch system do that:

- Model: launch jobs which lauch VMs
- Use the batch system accounting mechanism

Pros:

- Reuse existing infrastructure
- Get the accounting “for free”
- Grid infrastructure untouched
- Get benefits from virtualization

Cons:

- Couples traditional batch with virtualization
- Site specific implementation
- Site specific accounting
- Difficult to integrate services (eg voboxes)
- Island solution: special for HEP sites

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Accounting and billing

Options 2: a “cloud” way approach

- Model: IaaS infrastructure with self-service
- Use accounting/monitoring info from orchestrators

Pros:

- Allows for a generic infrastructure
- No (or soft) couping with applications
- 3rd party software and support
- Possible interface to a billing schema

Cons:

- More deployment work initially
- 3rd party software may not be ready
- How to do resource sharing ?

Data access

… how do current pilot frameworks address this issue?

Efficient data access within the site, good network performance and good local I/O are crucial

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Thank you !