CERN-IT Plans on Virtualization

Ian Bird

On behalf of IT

WLCG Workshop,
9th July 2010

Multi-core jobs at CERN

- Multi-core jobs can be run on the CERN batch service:
  - Dedicated queue “sftcms”
  - dedicated resources
    - 2 nodes with 8 cores, one job slot each
    - Can be extended on request
  - access restricted to pre-defined users
    - Currently 12 registered users defined to use the queue
  - Service is in production since February 2010
    - But has hardly been used yet

Multi-core jobs Grid-wide

- EGEE-III MPI WG recommendations
  - Current draft at http://grid.ie/mpi/wiki/WorkingGroup
  - Final version to appear soon
- Next steps
  1. Validate new attributes in the CREAM-CE:
     - Using a temporary JDL attribute (CERequirements)
       - E.g. CERequirements = “WholeNodes = \”True\””;
     - Using direct job submission to CREAM
     - Validation starting with some user communities
  2. Modify CREAM and BLAH so that the new attributes can be used as first-level JDL attributes:
     - E.g. WholeNodes = True;
     - Coming with CREAM 1.7, Q3 2010
  3. Support for the new attributes in WMS
     - Coming with WMS 3.3, Q3 2010

Virtualisation Activities

- In many areas – not all directly relevant for LHC computing
- 3 main areas:
  - Service consolidation
    - “VOBoxes”, VMs on demand, LCG certification testbed
  - “LX” services
    - Worker nodes, pilots, etc (issue of “bare” WN tba)
  - Cloud interfaces

- Rationale:
  - Better use of resources – optimise cost, power, efficiency
  - Reduce dependencies esp between OS and applications (e.g. SL4 → SL5 migration), and between grid software
  - Long term sustainability/maintainability → can we move to something which is more “industry-standard” ?

+ don’t forget WLCG issues of how to expand to other sites
  - which may have many other constraints (e.g. may *require* virtualised WN)
  - Must address trust issue from the outset

Service consolidation

- VO Boxes (in the general sense of all user-managed services)
  - IT runs the OS and hypervisor; user runs the service and application
  - Clarifies distinction in responsibilities
  - Simplifies management for VOC – no need to understand system configuration tools
  - Allows to optimise between heavily used and lightly used services
  - (eventually) transparent migration between hardware: improve service availability
- VMs “on demand” (like requesting a web server today)
  - Request through a web interface
  - General service for relatively long-lived needs
  - The user can request a VM from among a set of standard images
  - E.g.:
    - ETICS multi-platform build and automated testing
    - LCG certification test bed. Today uses a different technology, but will migrate once live checkpointing of images is provided.

CVI: CERN Virtualisation Infrastructure

- Based on Microsoft’s Virtual Machine Manager
  - Multiple interfaces available
    - ‘Self-Service’ web interface at http://cern.ch/cvi
    - SOAP interface
    - Virtual Machine Manager console for Windows clients
- Integrated with LANdb network database
- 100 hosts running Hyper-V hypervisor
  - 10 distinct host groups, with delegated administration privileges
  - ‘Quick’ migration of VMs between hosts
    - ~1 minute, session survives migration
- Images for all supported Windows and Linux versions
  - Plus PXE boot images

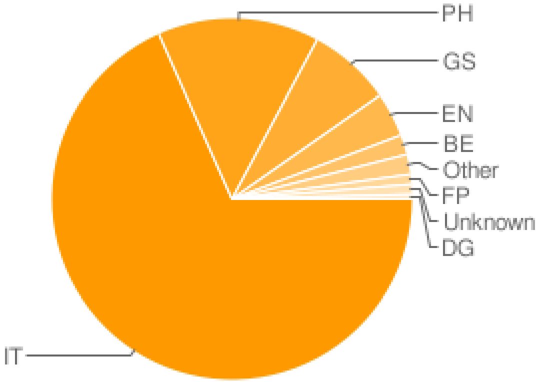
CVI: some usage statistics

Today, CVI provides 340 VMs…

- 70% Windows, 30% Linux

… for different communities

- Self-service portal
  - 230 VMs from 99 distinct users
  - Mixture of development and production machines
- Host groups for specific communities
  - Engineering Services: 85 VMs
  - Media streaming: 12 VMs
  - 6 Print servers, 8 Exchange servers
  - … etc



CVI: work in progress - II

- Consolidation of physics servers (IT/PES/PS)
  - a.k.a “VOBoxes”
  - ~300 Quattor managed Linux VM’s
    - IT servers and VOBoxes
  - 3 enclosures of 16 blades
    - Each with ISCSI shared storage
    - Failover cluster setup
  - Allows transparent <u>live</u> migration between hypervisors
  - Need to gain operational experience
    - Start with some IT services first (e.g. CAProxy, …)
    - Verify that procedures work as planned
  - Target date for first VOBoxes is ~July 2010
    - gain experience with non-critical VOBox services first
  - Next step: Virtual small diskservers (~ 5TB of storage)
    - By accessing iSCSI targets from VM’s

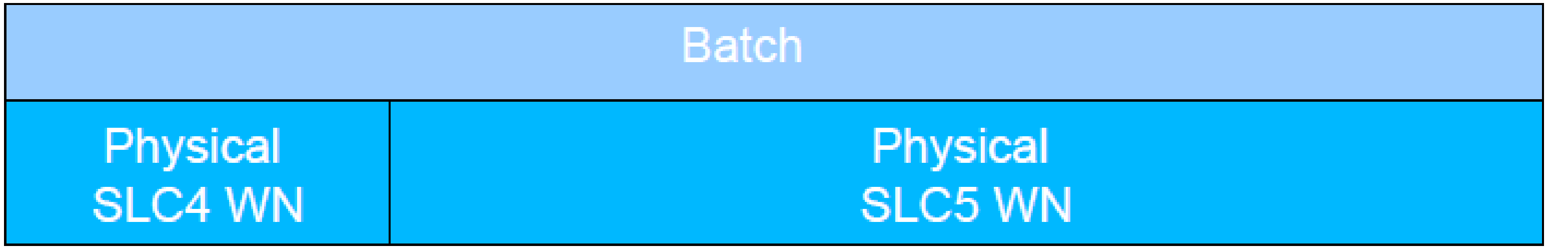
“LX” Services

- LXCloud
  - See presentation of Sebastian Goasguen + Ulrich Schwickerath at workshop
  - Management of a virtual infrastructure with cloud interfaces
  - Includes the capability to run CernVM images
  - Scalability test are ongoing
  - Tests ongoing with both Open Nebula and Platform ISF as potential solutions
- LXBatch
  - Standard OS worker node: as VM addresses dependency problem
  - WN with a full experiment software stack
    - user could choose from among a standard/certified set. These images could e.g. Be built using the experiment build servers.
  - As the previous case but with the pilot framework embedded
  - CERNVM images
- Eventually: LXBatch ➔ LXCloud

Evolution

(SLC = Scientific Linux CERN)

Today

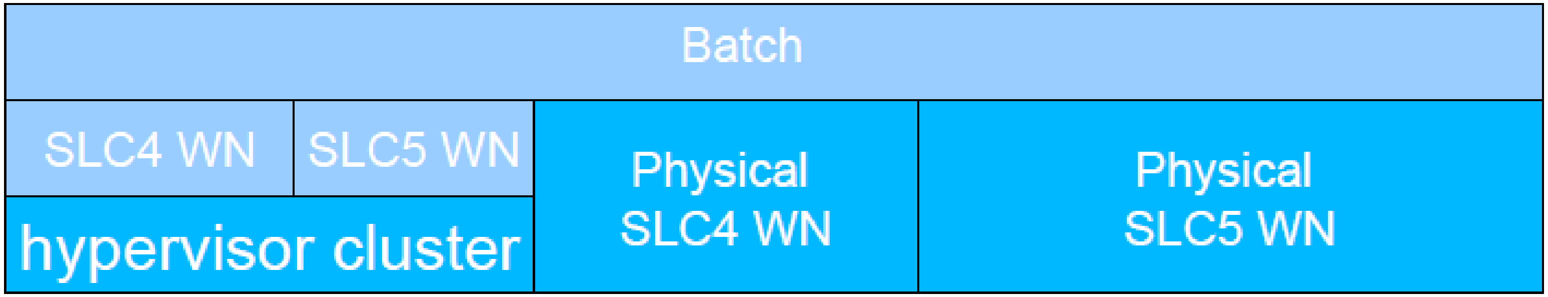


Batch

Physical SLC4 WN

Physical SLC5 WN

Near future:



Batch

SLC4 WN

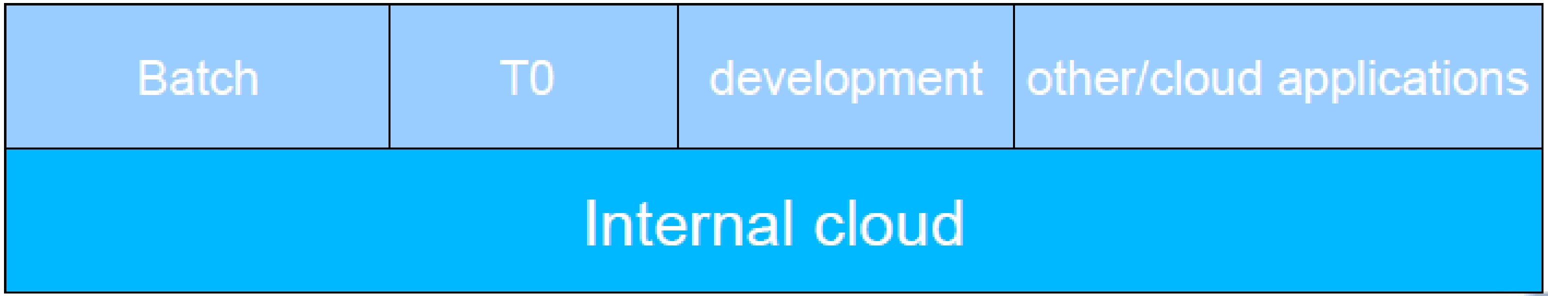
SLC5 WN

hypervisor cluster

Physical SLC4 WN

Physical SLC5 WN

(far) future ?



Batch

T0

development

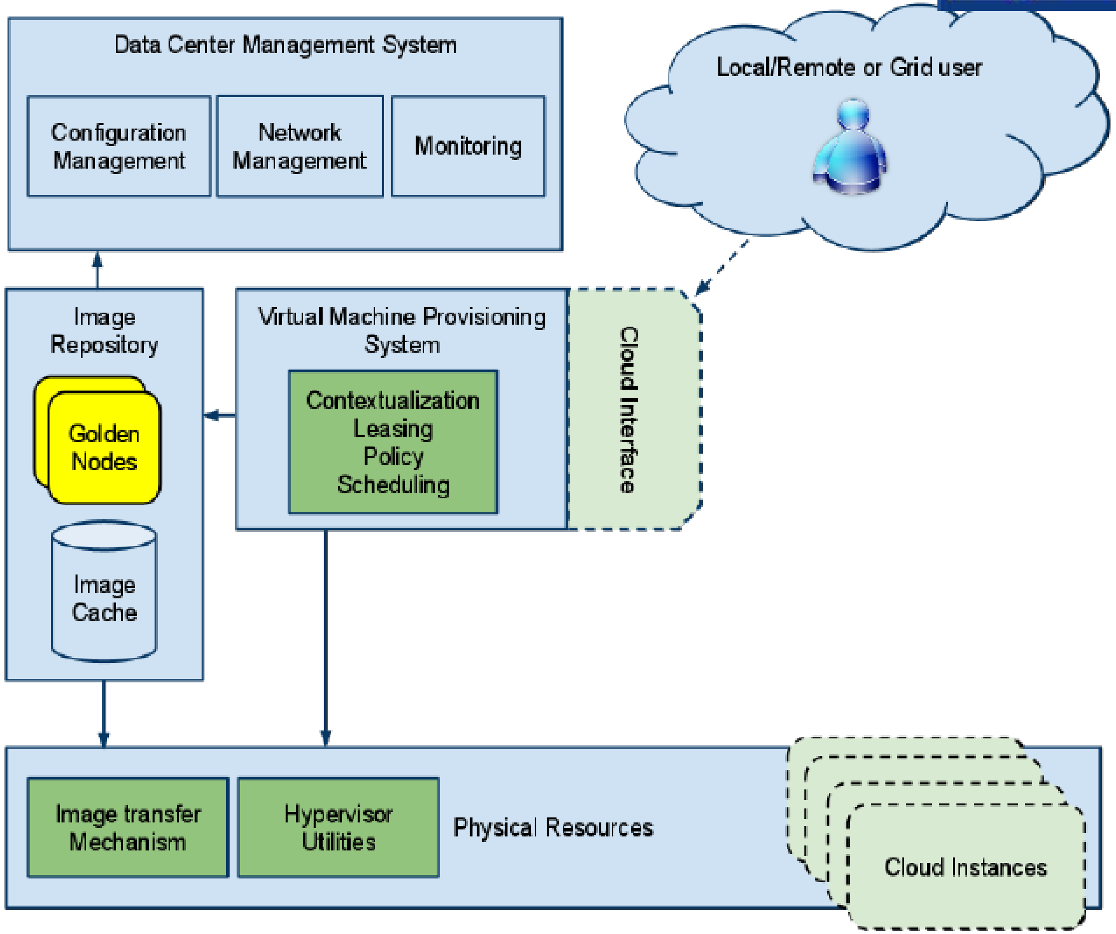
other/cloud applications

Internal cloud

LXCloud

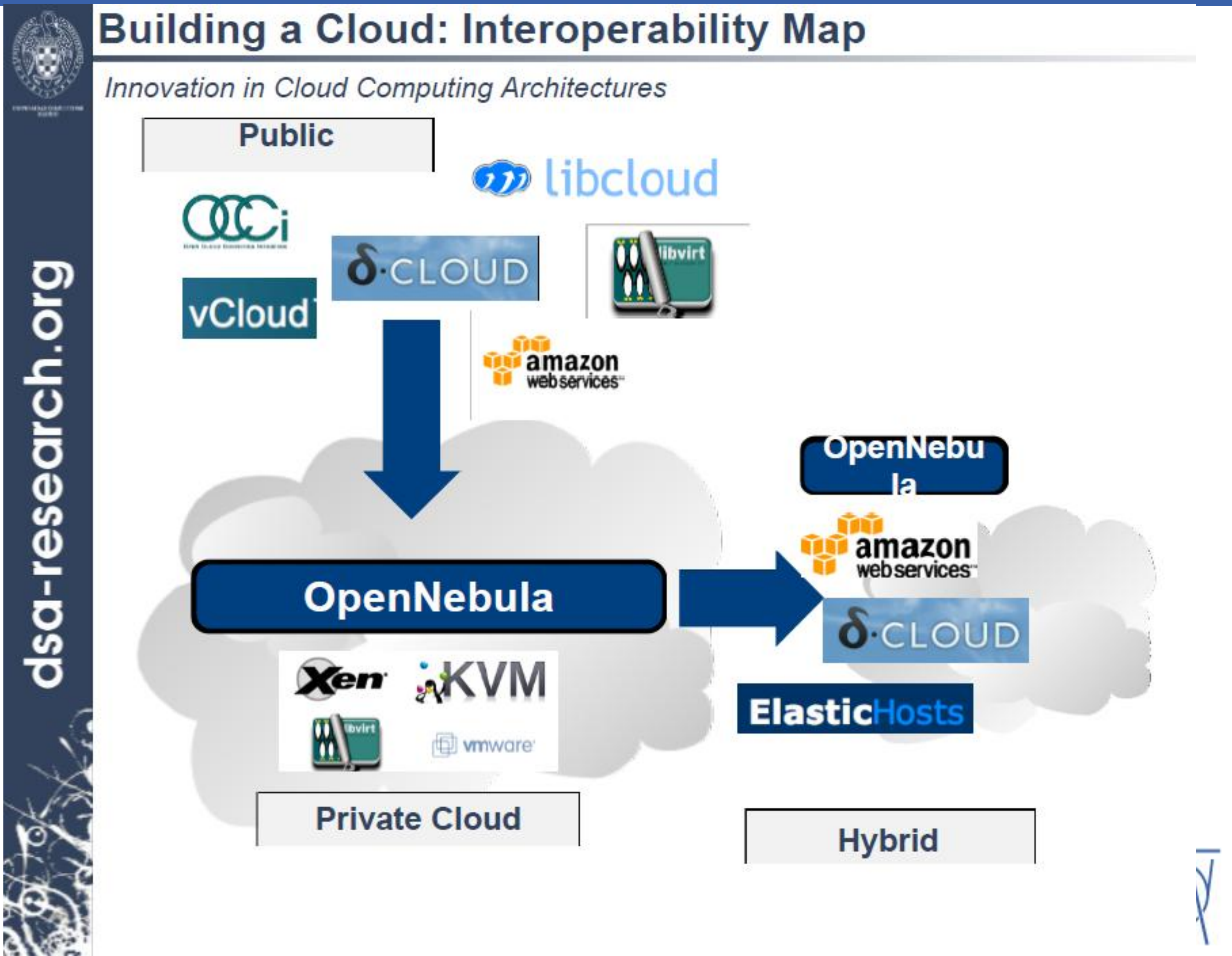
CERN's lxcloud architecture

- Image repository with Golden nodes.
- VM instances not quattor managed have finite lifetime
- Specific IP/MACs are pinned to hypervisors
- Currently testing two provisioning system: Opennebula and Platform ISF.



Ongoing work

- Integration with existing management infrastructure and tools
  - Including monitoring and alarm systems
- Evaluating VM provisioning systems
  - Opensource and commercial
- Image distribution mechanisms
  - With P2P tools
- Scalability tests
  - Batch system (how many VMs can be managed)
  - Infrastructure (network database, etc.)
  - Image distributions
  - VM performance
- To be understood:
  - I/O performance, particularly for analysis jobs
  - How to do accounting etc.
  - Virtualised infrastructure vs allocate “whole node” to application



Building a Cloud: Interoperability Map

Innovation in Cloud Computing Architectures

Public

Private Cloud

Hybrid

Summary

- Ongoing work in several areas
  - Will see benefits immediately (e.g. VOBoxes)
  - Will gain some experience in (e.g. LXCloud) test environment
    - Should be able to satisfy a wide range of resource requests – no longer limited to the model of a single job/cpu
- Broader scope of WLCG
  - Address issues of trust, VM management; integration with AA framework etc
  - Interoperability through cloud interfaces (in both directions) with other “grid” sites as well as public/commercial providers
  - Can be implemented in parallel with existing grid interfaces