Virtualization at CERN – status report

Sebastien Goasguen, Belmiro Rodrigues Moreira, Ewan Roche,
Steve Traylen, Ulrich Schwickerath, Romain Wartel

HEPIX2010, Lisbon

See also related presentations:

- Batch virtualization at CERN, HEPIX autumn meeting 2009
- Virtualization, HEPIX spring meeting 2009
- Virtualization vision, GDB 9/9/2009 and HEPIX
- *Batch virtualization at CERN, EGEE09 conference, Barcelona*

Introduction and overview

CERN IT Department

Virtualization for **service consolidation**:

- Up to ~few 100 machines (today)
- Typically little CPU usage on these boxes
- Support for live-migration required
- Based on reliable hardware
- Critical services

Virtualization for **batch** (and similar) applications:

- Large scale, O(several 1000) machines
- High CPU usage, number crunching
- Limited life time is OK!
- Cheap batch hardware is OK!
- Individual machines are not critical
- Option for “cloud” like infrastructure

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

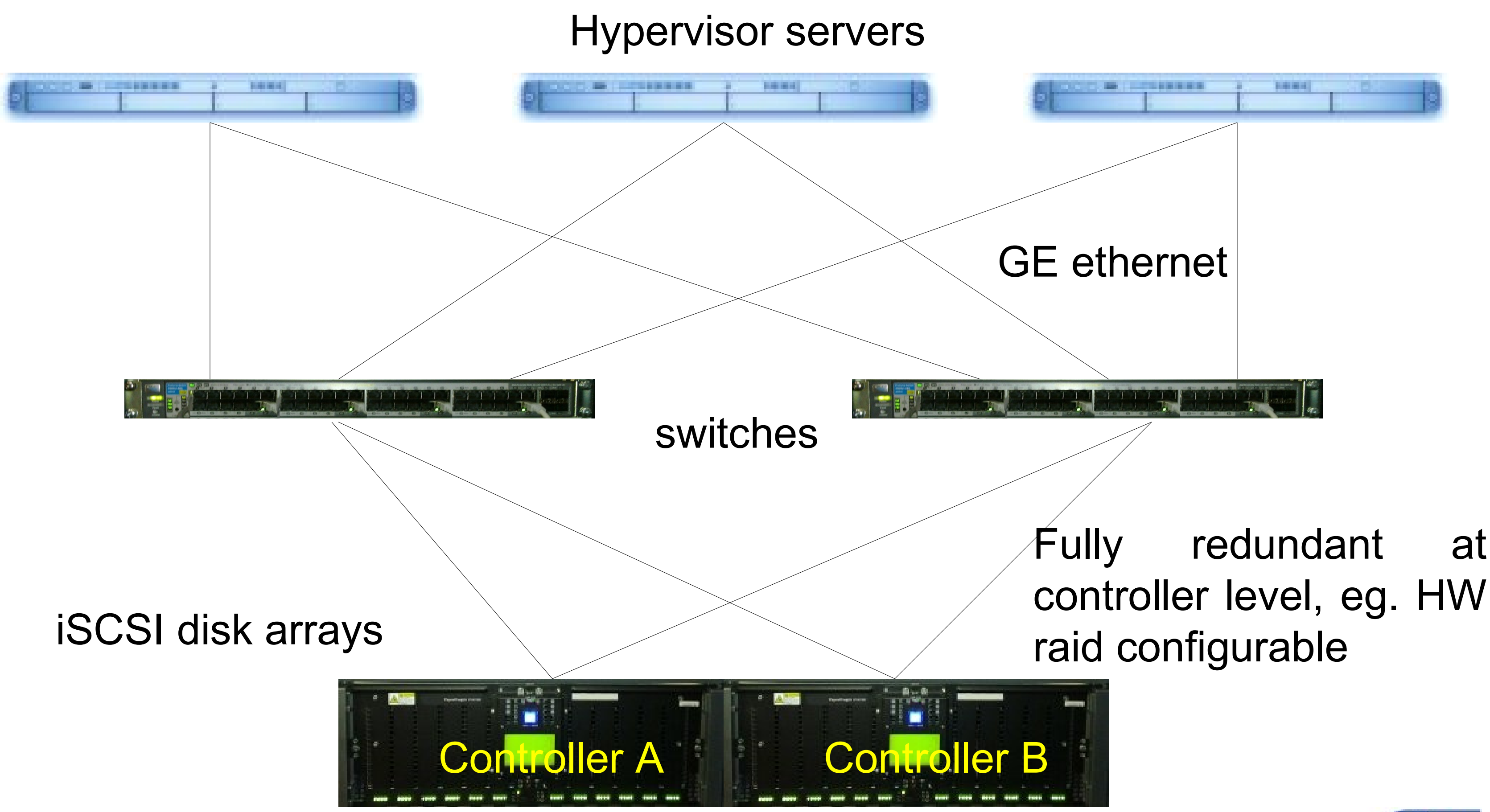
Outline

- Introduction and overview
- Service consolidation project
    - Philosophy
    - Resources
    - Status
- Batch virtualization
    - Philosophy and layout
    - Building block status:
        - ISF
        - OpenNebula (ONE)
        - Image distribution/virtualization kiosk
    - Use cases so far
        - Development
        - Scalability tests
    - Initial performance tests
    - Current issues

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

Service Consolidation - Philosophy

Building blocks of 16 servers, 2 switches and 2 iSCSI attached disk arrays with private data network layout



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

PES

Service Consolidation - Resources

Three different layouts, reflecting different use cases:

1) Critical machines with little disk space requirements
- 40TB raw disk space per 16 hypervisors
- Full dual UPS coverage and Diesel backup

2) Less critical machines
- 192TB raw disk spare per 16 hypervisors
- For example for dedicated servers for experiments and similar

3) Small disk server consolidation
- 284 TB raw disk space per 16 servers
- Replacement for machines requiring O(2TB) secure storage

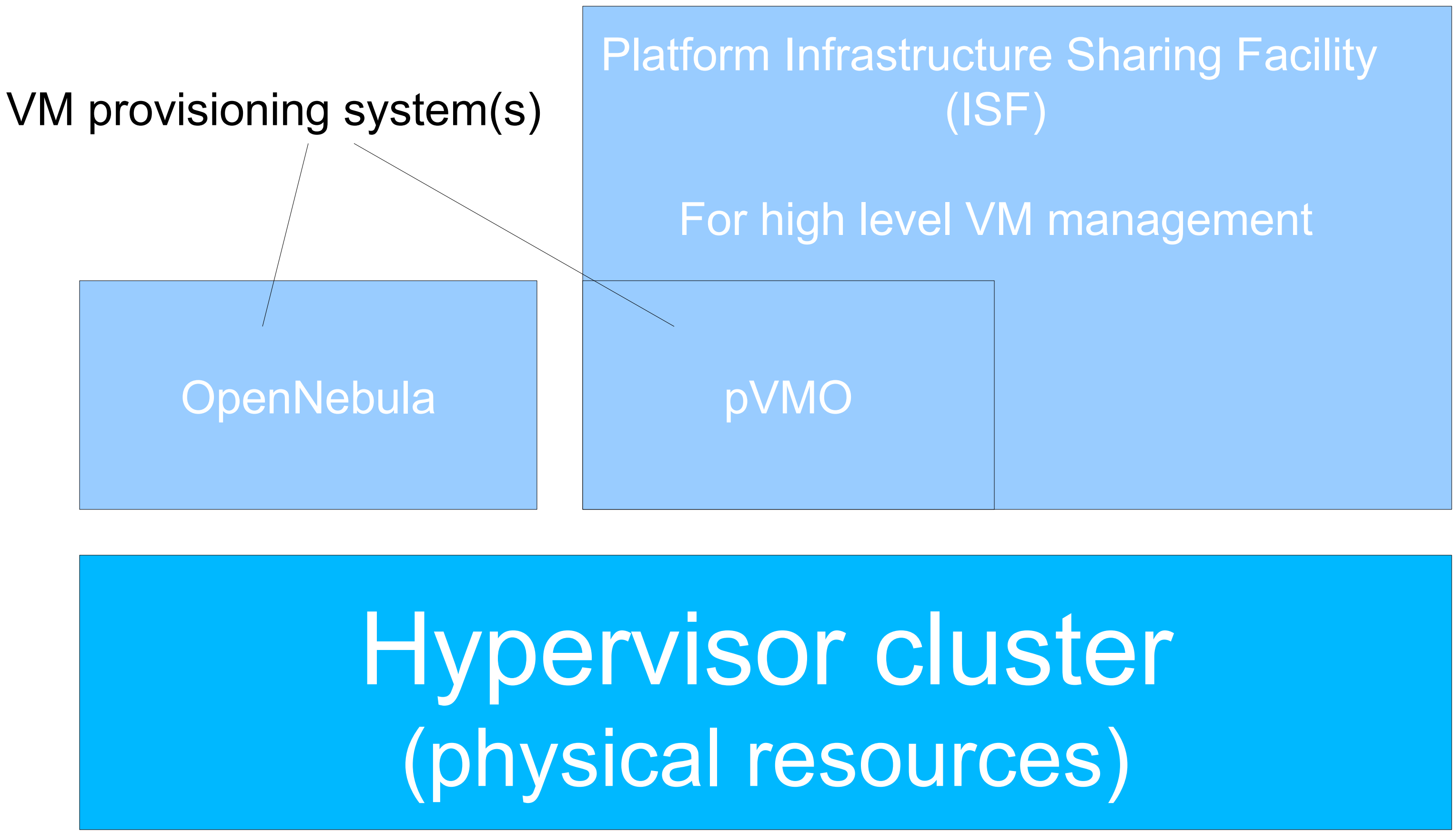
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Service consolidation: current status

- Hardware is installed and tested
- Hypervisor OS installation is proceeding (Windows HyperV)
- Performance testing ongoing, some issue currently under investigation

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

Philosophy



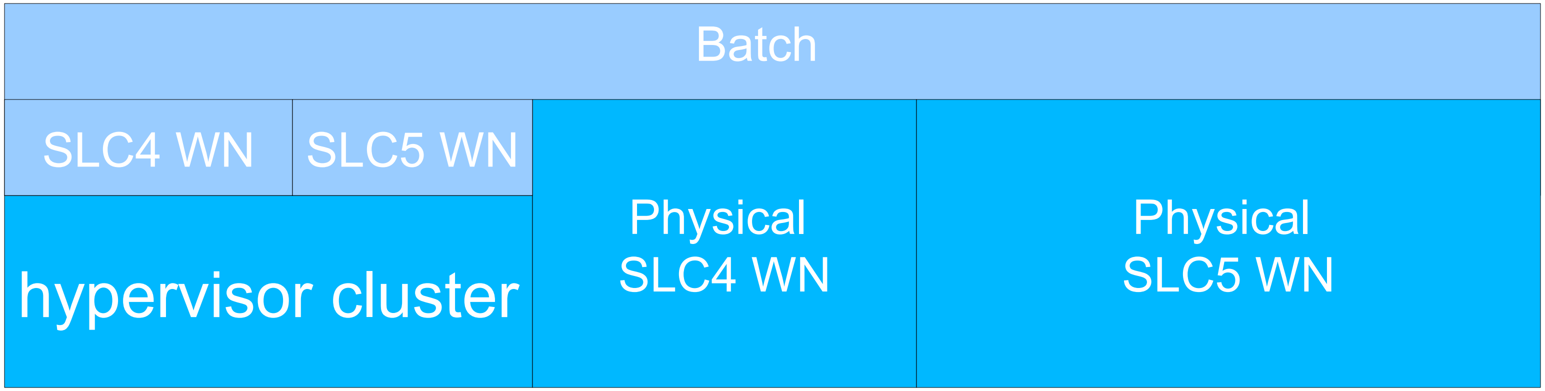
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Philosophy

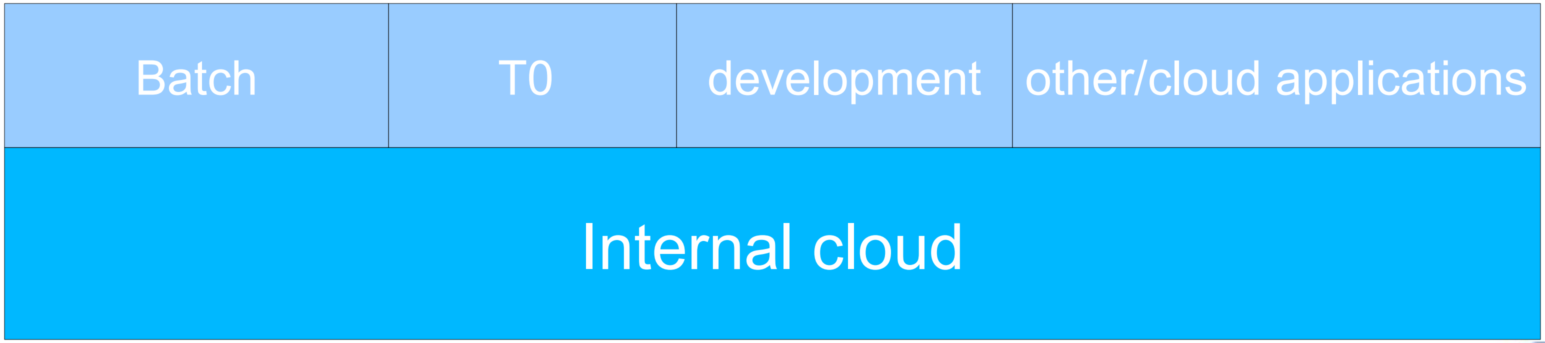
Vision: Transparent resource sharing by

dynamic re-allocation of automatically freed virtual machine slots

Near future:

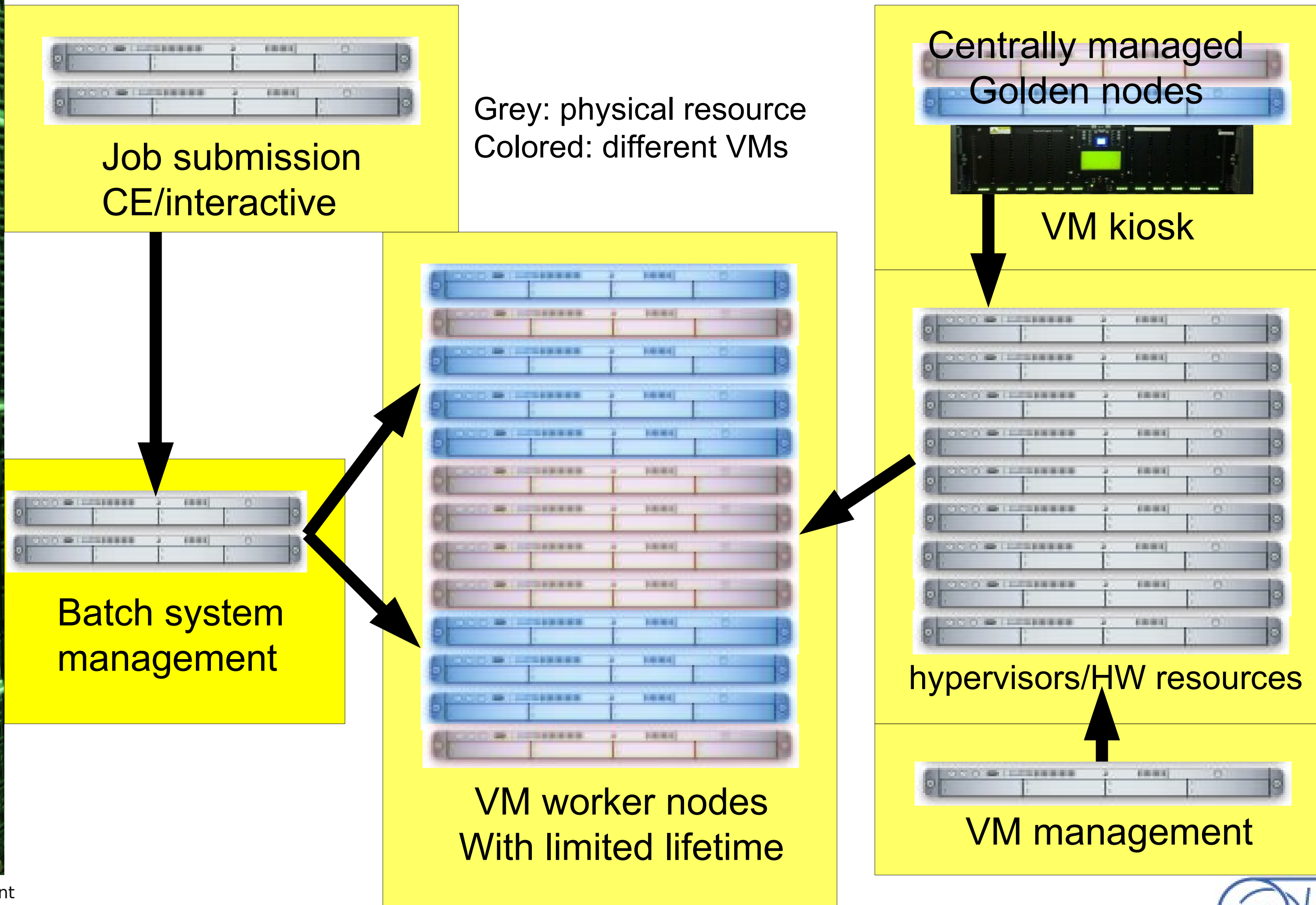


(far) Future ?



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

Batch virtualization: architecture



Some additional explanations ...

"Golden node":

A centrally managed (i.e. Quattor controlled) standard worker node which

- Is a virtual machine
- Does not accept jobs
- Receives regular updates

Purpose: creation of VM images

"Virtual machine worker node":

- A virtual machine derived from a golden node
- Not updated during their life time
- Dynamically adds itself to the batch farm
- Accepts jobs for only 24h
- Runs only one user job at a time
- Destroys itself when empty

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

Job submission and batch management

No change in job submission schema:

- Interactive job submission from lxplus and/or VOBoxes
- GRID integration via LCG and CREAM CEs
- Transparent for the users, WN look a bit different
- Initially, VMs will be put behind special queues for testing

Job processing:

- One job slot per VM only
- Stable software configuration during VM life time
- Limited VM life time allows for flexible reallocation of resources

Remarks:

- Submission hosts could become VM as well in the future
- Resource sharing via the internal cloud allows in principle a split of the batch farm into independent instances without loss of flexibility

Status of individual building blocks

Some Definitions and clarifications

Quattor managed:

- Full integration into Quattor toolkit
- Centrally managed and updated
- State management implemented (maintenance, production etc)

Lemon monitored:

- Monitoring sensors are present and configured
- Exceptions are configured were applicable
- The node can raise operator alarms

Auto registration:

- Hypervisors become visible and active to the provisioning system
- Batch nodes get automatically included in the batch farm

Central maintenance:

- Installation and deployment like any other box
- Procedures to handle alarms and exceptions are in place

Status of individual building blocks

	Hypervisor cluster	SLC5 virtual batch nodes	SLC4 virtual batch nodes
Quattor managed	OK	OK, via golden node	OK, via golden node
Lemon monitored	OK	NO	NO
Auto-registration	OK	OK, in LSF	OK, in LSF
Central maintenance	OK, some bits missing	Not required	Not required
ISF support	OK, testing	OK, testing	NO
ONE support	OK	OK	NO

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

VM management systems - status

	ISF server	ONE server	ISF Agent	ONE client
Quattorized	Work in progress	NO	YES	Not (yet) relevant
Lemon monitored	Only OS and hardware	Only OS and hardware	Service is not yet monitored	Not (yet) relevant

ISF: still work in progress
OpenNebula (ONE): used for scalability tests

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

VM kiosk and Image distribution

Image distribution constraints for CERN:

- Network infrastructure with a single 1GE connection
- No dedicated fast network for transfers that could be used (eg 10GE, IB or similar)

Image distribution with scp:

- ~1h for 7GB to 500 nodes

Using ***rtorrent:*** under test now

- *.torrent* files are created centrally and then distributed
- One central tracker (using opentracker), option to remove it later on
- Compliant with current JSPG security policy draft for image distribution

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

VM kiosk and Image distribution:status

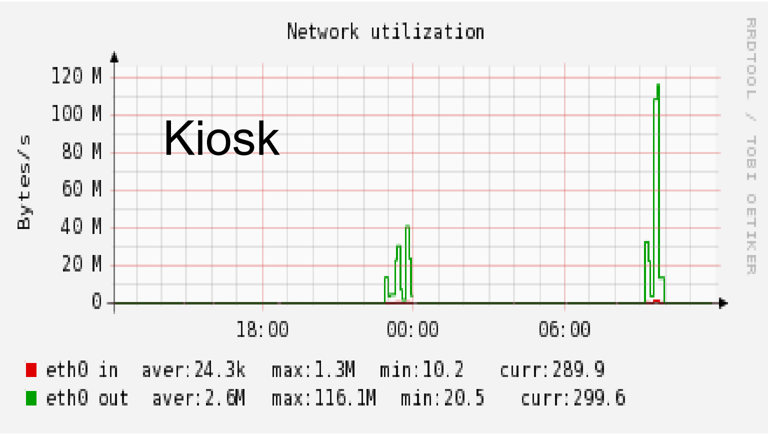
	Repository server	Image reception	Image deployment
Initial development and testing	OK	OK	OK
Quattorization	OK	OK	OK
Lemon monitoring	Hardware and OS only	Process not monitored yet	NO

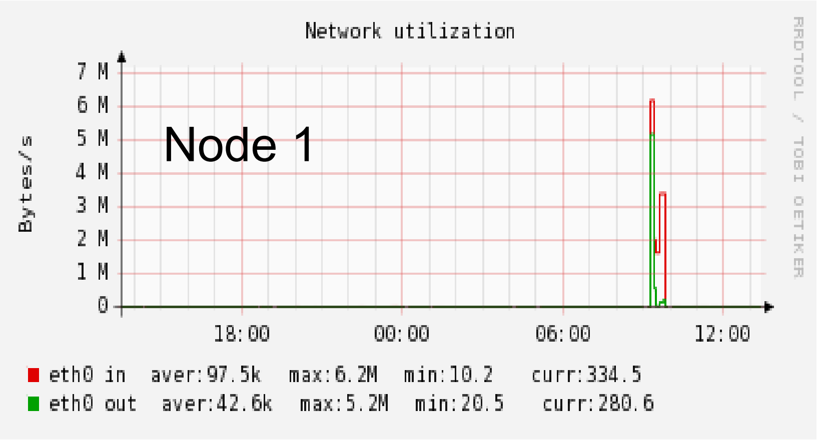
Images deployment: create logical volume and dd image into it

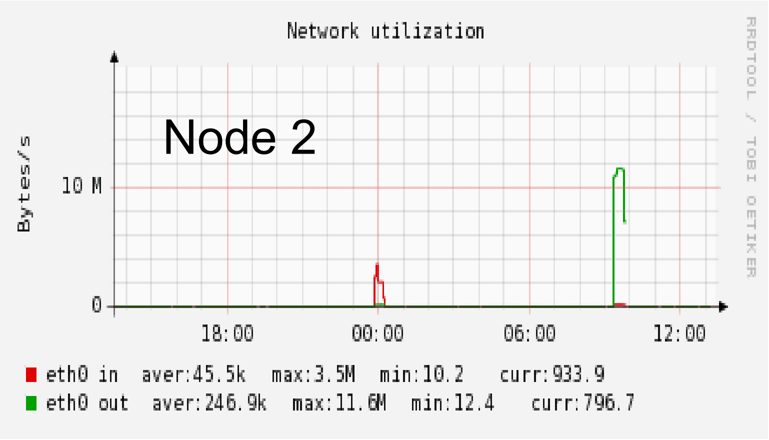
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

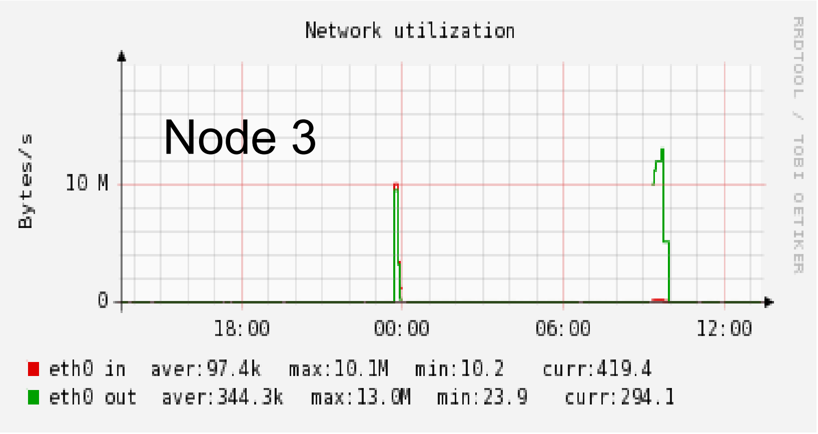
Initial tests (preliminary!)

Image distribution: 7GB size compressed image to 48 nodes



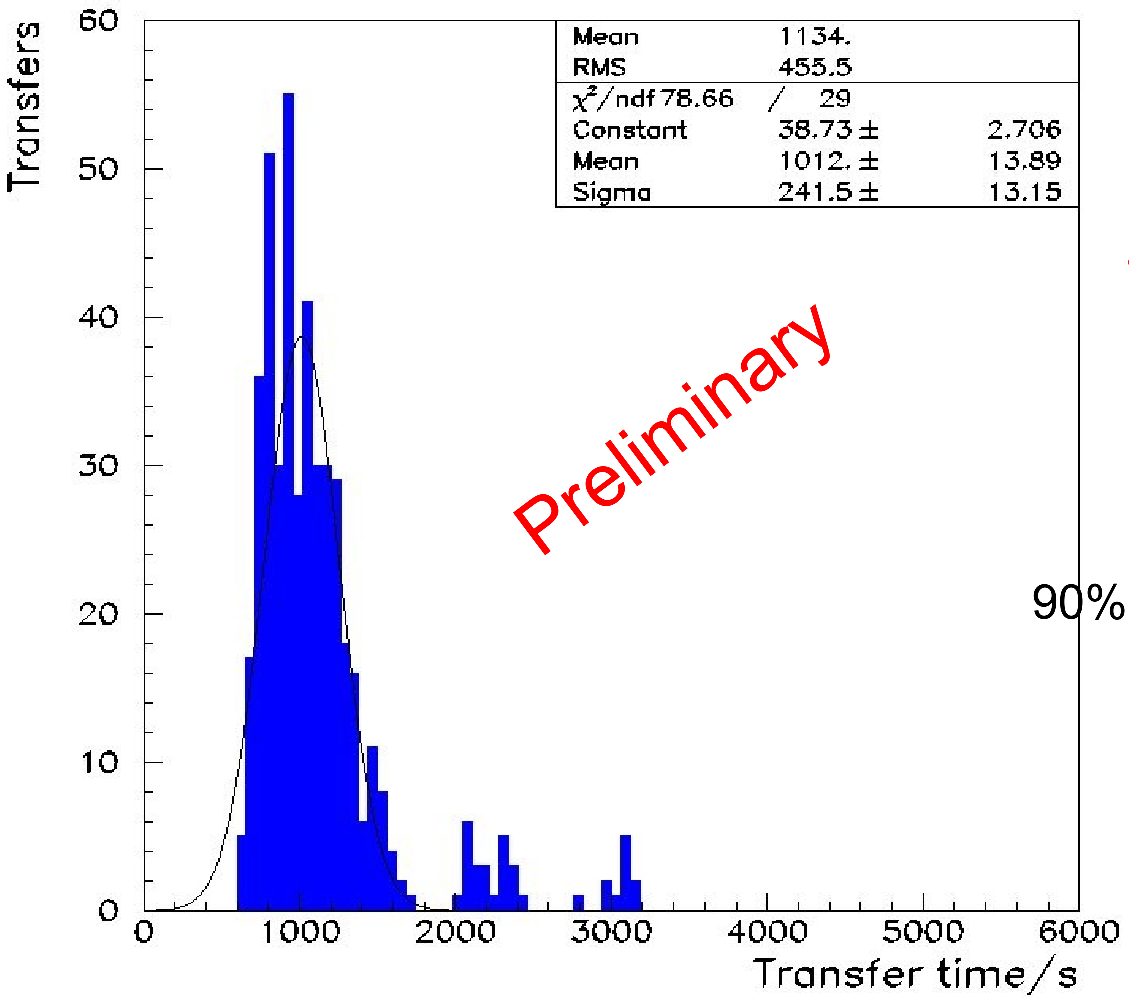






CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Image distribution: transfer speed



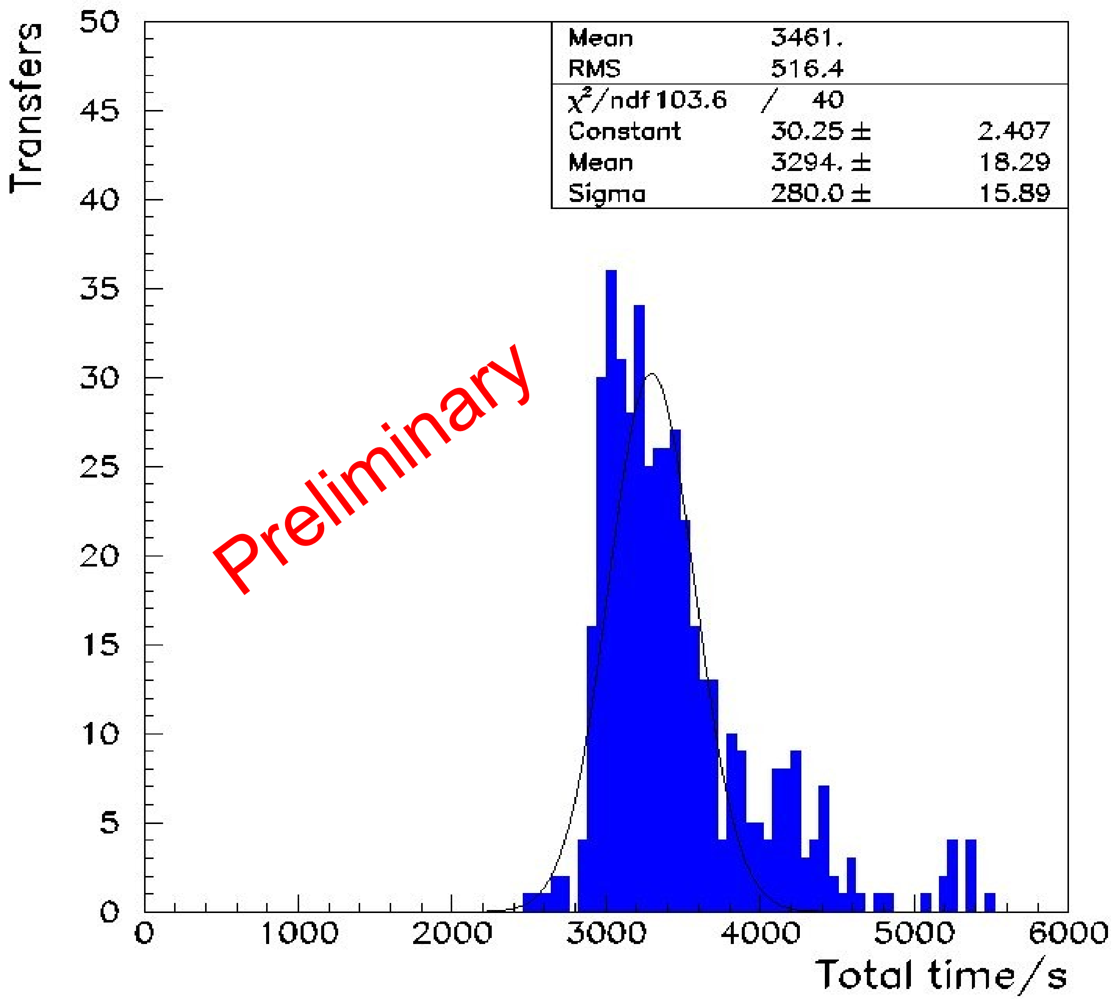
7GB file compressed
452 target nodes

90% finished after 25min

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

Image distribution: **total** distribution speed

→ Unpacking is very expensive !



7GB file compressed
452 target nodes

All done after 1.5h

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

Use cases (so far)

PES

CERN IT Department

Development for production services

- Testing and debugging of new worker node releases
- Testing and debugging of the glExec worker node
- Testing and debugging of CREAM

Scalability tests (ongoing)

- Testing Platform LSF in terms of number of nodes
- Testing OpenNebula (and ISF)
- Insertion and deletion of entries in LANDB at large scale

Still too early for true production !

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

Scalability tests

Why large scale scalability tests ?

- Context of batch virtualization
- Need to know the limits of the production system as well!

CERN batch farm in numbers:

- ~2600 physical nodes known to a single LSF instance right now
- 1500 new nodes are on the floor and almost ready to go

Scalability tests

“borrow” new batch resources before they enter production

- About 500 recent worker nodes (10 racks)
  - SuperMicro twin2 systems
  - Dual Intel XEON 5520 based (“Nehalem”) running at 2.27GHz
  - 2-3 TB local disk space
  - 24 GB RAM (few nodes with up to 48GB)
- About 10,000 registered virtual machine slots
  - On average 19 private IP adresses per hypervisor
- Dedicated LSF test instance
  - Master setup clones of production masters
  - Nodes join this cluster dynamically
  - Expected scalability limit : O(5k) machines

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

Scalability tests: launching VMs ...

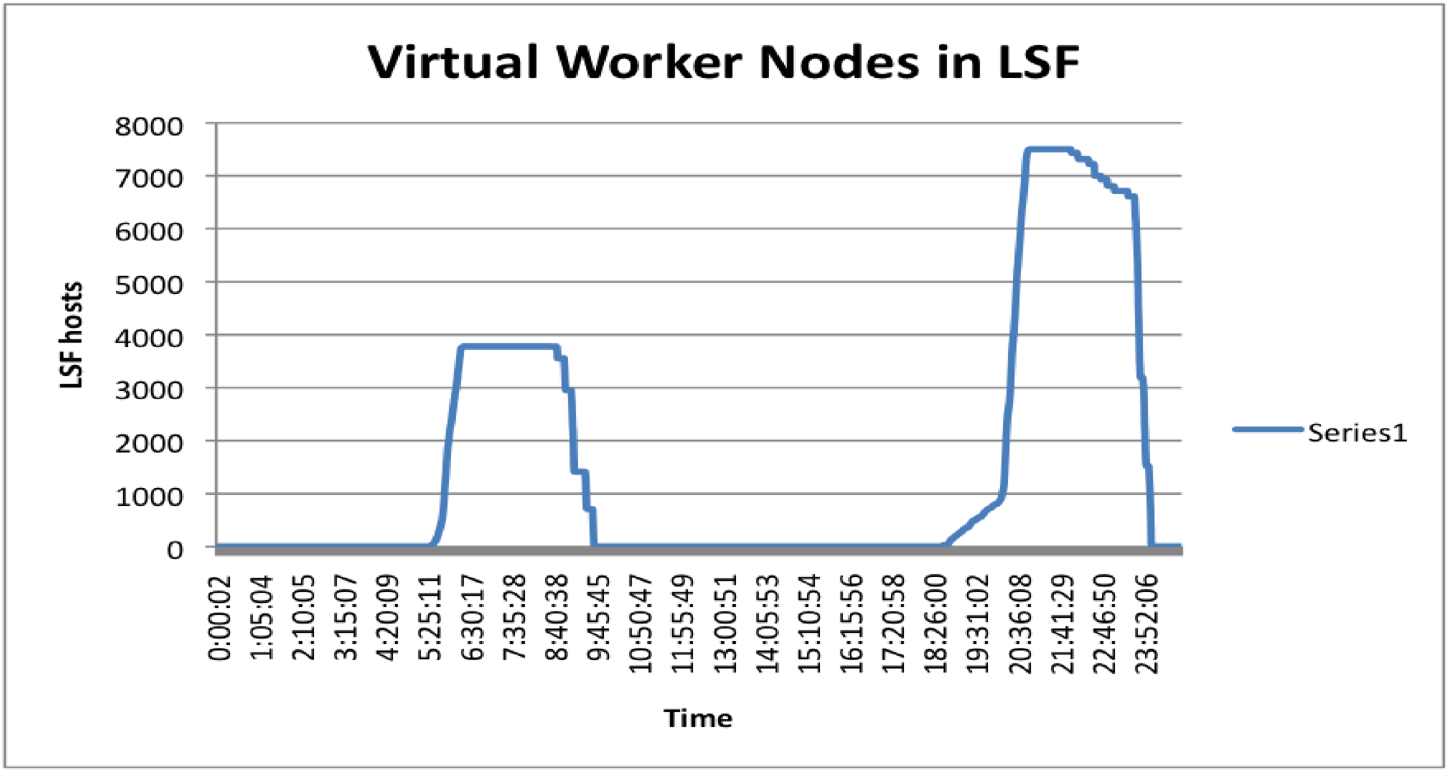
- OpenNebula:
  - Close collaboration with developers, addressing issues as they turn up
  - Up to 7,000 VMs started
- ISF/pVMO:
  - Initial deployment issues
  - Equivalent tests are in preparation

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

Scalability tests: some first numbers

One shot test with OpenNebula:

- Inject virtual machine requests
- And let them die
- Record the number of alive machines seen by LSF every 30s



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

Summary

Virtualization efforts at CERN are proceeding.
Still some work to be done.

Main challenges include

- Scalability of provisioning system(s)
- Batch system scalability
- Networking (public IPs)
- Performance of image distribution and VMs
- Seamless integration into the existing infrastructure
- Monitoring

Plan:

make production quality VMs available to our customers within the next months

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