Batch virtualization and Cloud computing

Part 1: Batch virtualization

Tony Cass, Sebastien Goasguen, Belmiro Moreira, Ewan Roche,
Ulrich Schwickerath, Romain Wartel

2nd Workshop on adapting applications and computing services to multi-core and virtualization

See also related presentations:

- CloudViews2010 conference, Porto
- HEPIX spring and autumn meeting 2009, 2010
- Virtualization vision, Grid Deployment Board (GDB) 9/9/2009
- *Batch virtualization at CERN, EGEE09 conference, Barcelona*
- *1st multi-threading and virtualization workshop, CERN, 2009*

CERN IT Departmer
CH-1211 Genève 2
Switzerlan
www.cern.ch/i

Outline and disclaimer

- **Part 1: batch virtualization**
  - **What we are doing**
  - **Where we are right now**
  - **First experiences and results (preliminary)**
- **Part 2: cloud computing**
  - **Perspectives for new computing models**

Disclaimer: We are still in the testing and evaluation phase. No final decision has been taken yet on what we are going to use in the future.

All given numbers and figures are preliminary

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

Why virtualization and cloud computing ?

Service consolidation:

- Improve resource usage by squeezing mostly unused machines onto single big hypervisors
- Decouple hardware life cycle from applications running on the box
- Ease management by supporting life migration

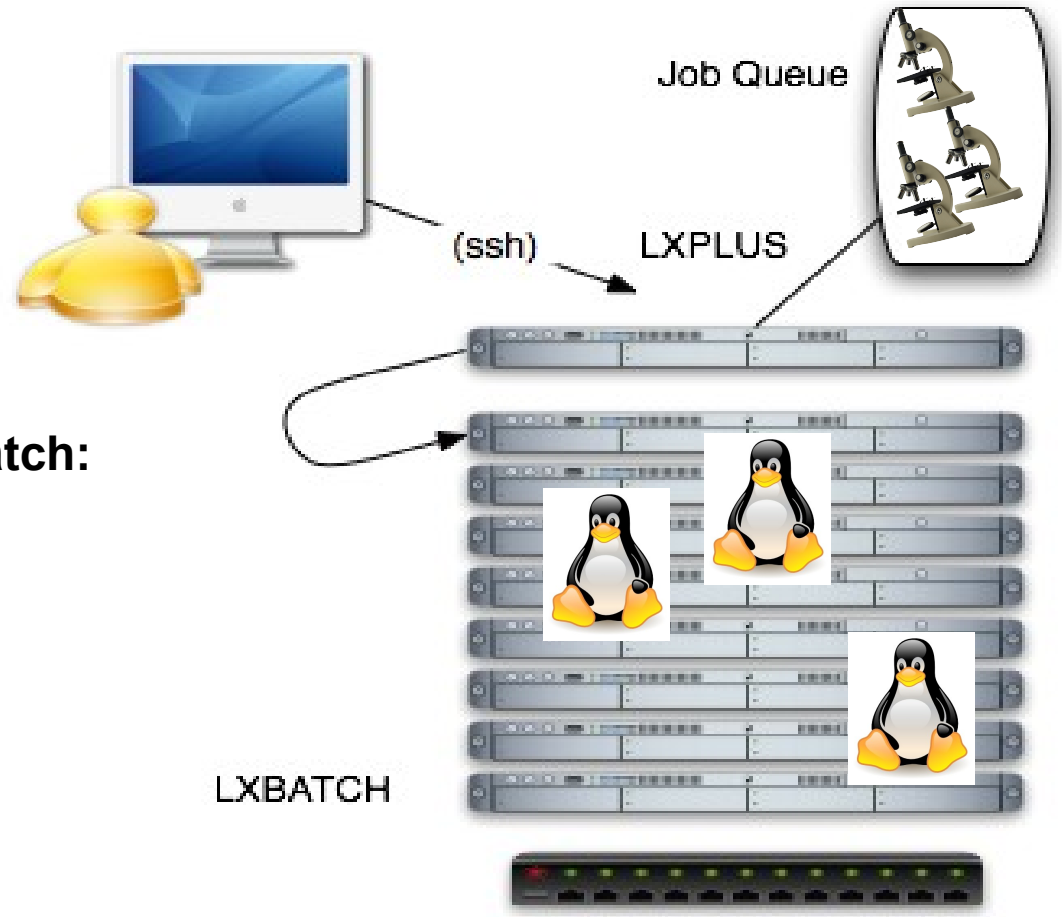
Virtualization of batch resources:

- Decouple jobs and physical resources
- Ease management of the batch farm resources
- Enable the computer center for new computing models

This presentation is about virtualization of batch resources only

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

Batch virtualization



CERN batch farm lxbatch:

~3500 physical hosts
~20000 CPU cores
>70 queues

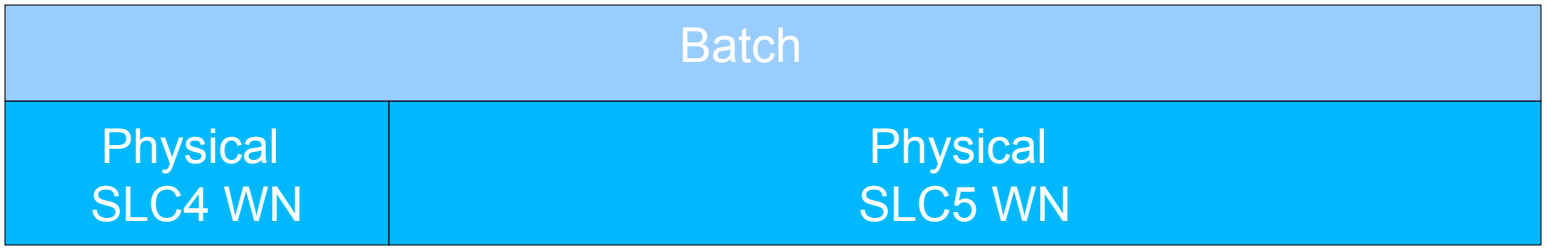
Current setup

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

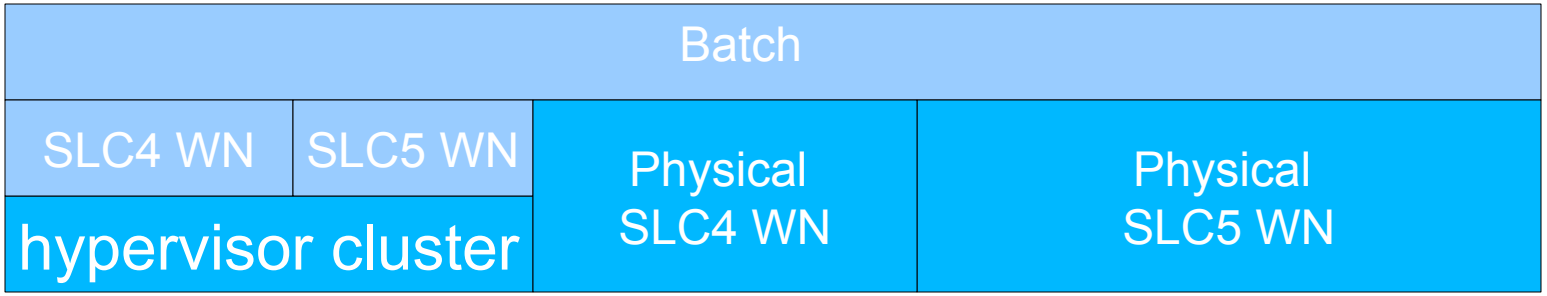
What we are doing: Philosophy

(SLC = Scientific Linux CERN)

Today



Near future:



(far) future ?

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

Virtual batch: basic ideas

Virtual batch worker nodes:

- Clones of real worker nodes, same setup
- Mix with physical resources
- Dynamically join the batch farm as normal worker nodes
- Limited lifetime: **stop accepting** jobs after 24h, and then destroy when em
- Only one user job per VM at a time

Note:
The limited lifetime allows for a fully automated system which dynamically adapts to the current needs, and automatically deploys intrusive updates.

Use case: intrusive updates on worker nodes (eg. Kernel upgrades)

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

Virtual batch: basic ideas, technical

Images:

- staged on hypervisors
- master images, instances use LVM snapshots
- Start with few different flavors only

Image creation:

- Derived from a centrally managed “golden node”
- Regularly updated and distributed to get updates “in”

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

Virtual batch: basic ideas

Images distribution:

- Only shared file system available is AFS
- Prefer peer to peer methods
  - SCP wave
  - Rtorrent

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

Virtual batch: basic ideas

VM placement and management system

- Use existing solutions
- Testing both a free and a commercial solution
  - OpenNebula (ONE)
  - Platform's Infrastructure Sharing Facility (ISF)
- Both implement a request queue and a very basic scheduler to place VMs

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

Status of building blocks (test system)

	Submission and batch managemen	Hypervisor cluster	VM kiosk and image distribution	VM management system
Initial deployment	OK	OK	OK	OK
Central management	OK	OK	Mostly implemented	ISF OK, ONE missing
Monitoring and alarming	OK	(OK) Switched off for tests	missing	missing

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

Test setup for scalability tests

Main goal: test **batch system scalability** using VMs

Side effect tests:

- General infrastructure (eg network database)
- Image distribution
- VM provisioning systems (ONE, ISF)
- VM performance

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

Test setup for scalability tests

Temporary access to **500** new worker nodes

- 2 x XEON L5520 @ 2.27GHz CPUs (8 cores)
- 24GB RAM
- ~1TB local disk space
- HS06 rating ~97

Software: SLC5/64bit with XEN

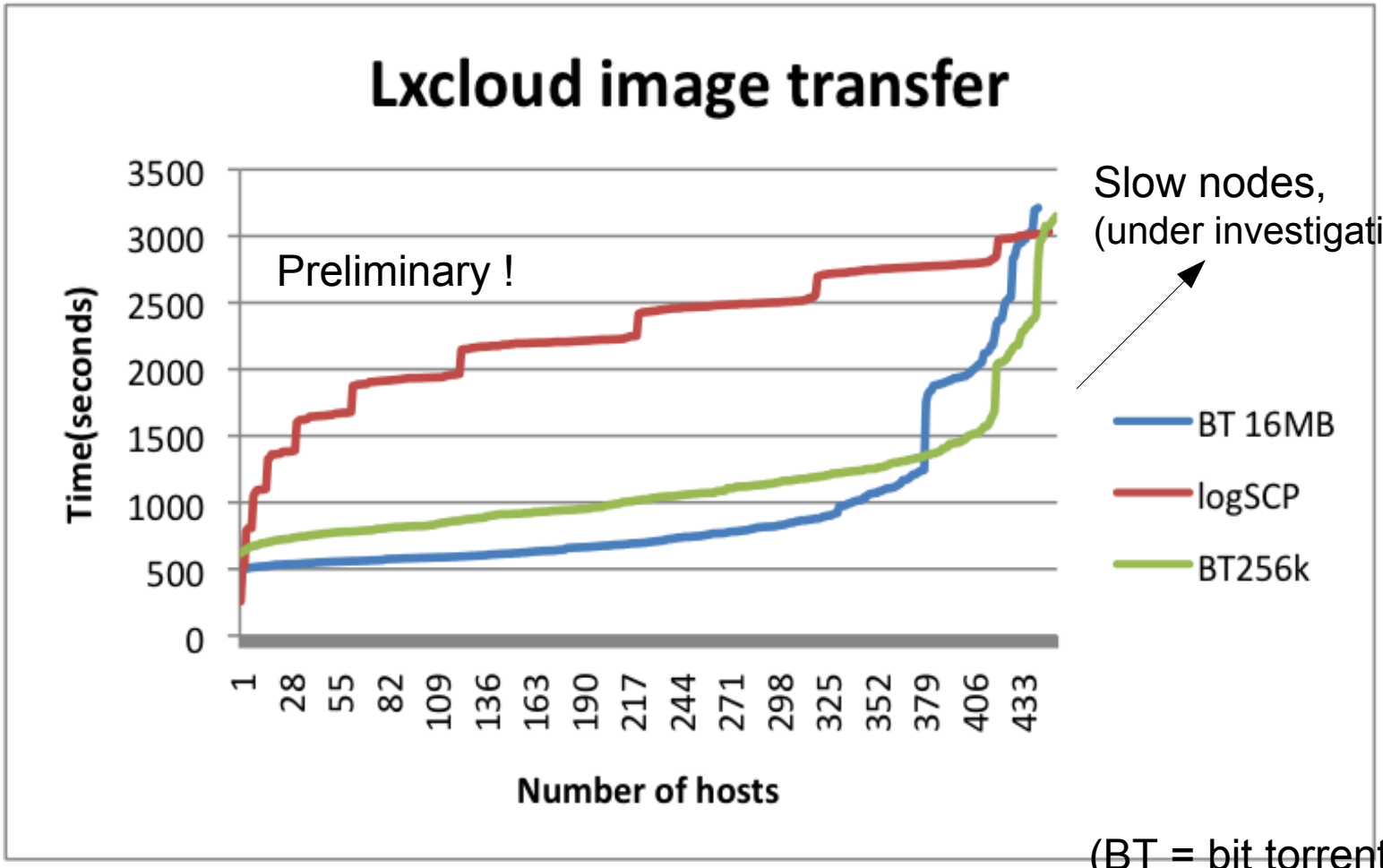
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Performance tests: overview

- Image distribution: from repository to the hypervisors (ongoing)
- VM provisioning systems (ongoing)
  - OpenNebula (ONE) (Open source solution)
  - Infrastructure Sharing Facility (ISF), Platform computing
- HS06 rating of VMs
- Disk performance tests (ongoing)
- Network performance tests (todo)

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

Image distribution: SCP wave versus rtorrent



Slow nodes,
(under investigation)

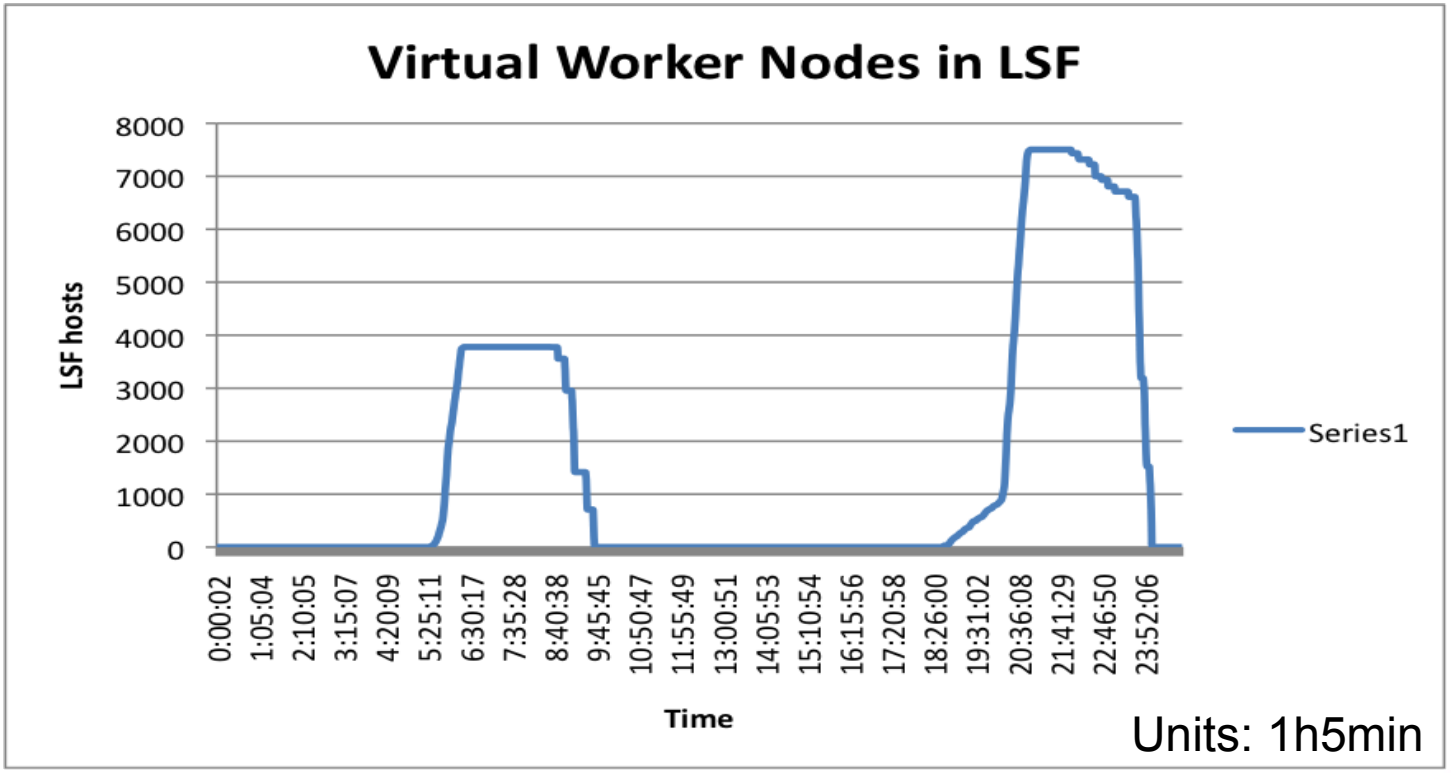
(BT = bit torrent)

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 live and die, reduced life time 4h
- Record the number of alive machines seen by LSF every 30s

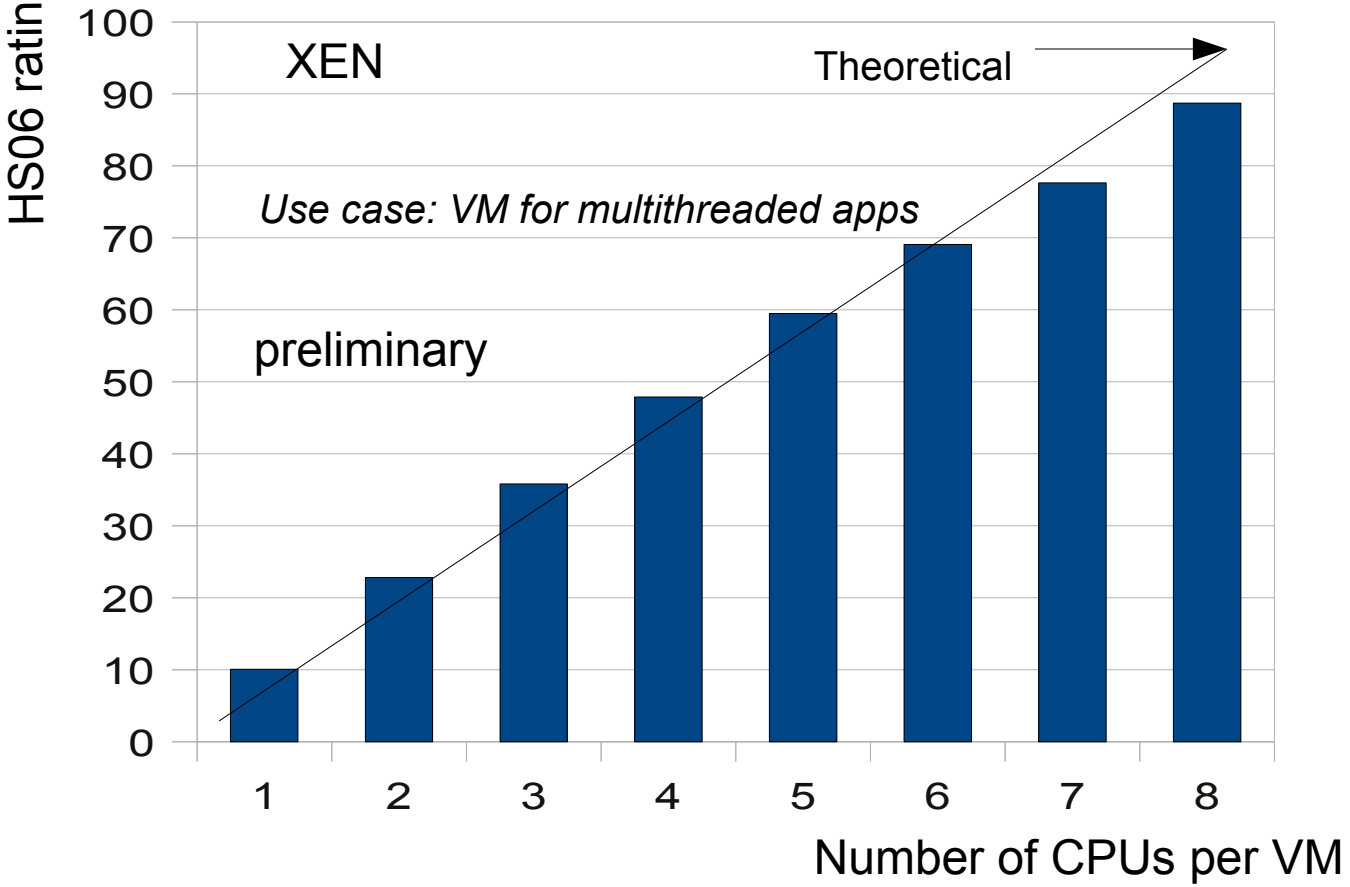


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CPU performance: HS06 benchmark results (preliminary)

Test case: 1 VM / hypervisor, raising the number of CPUs/hypervisor

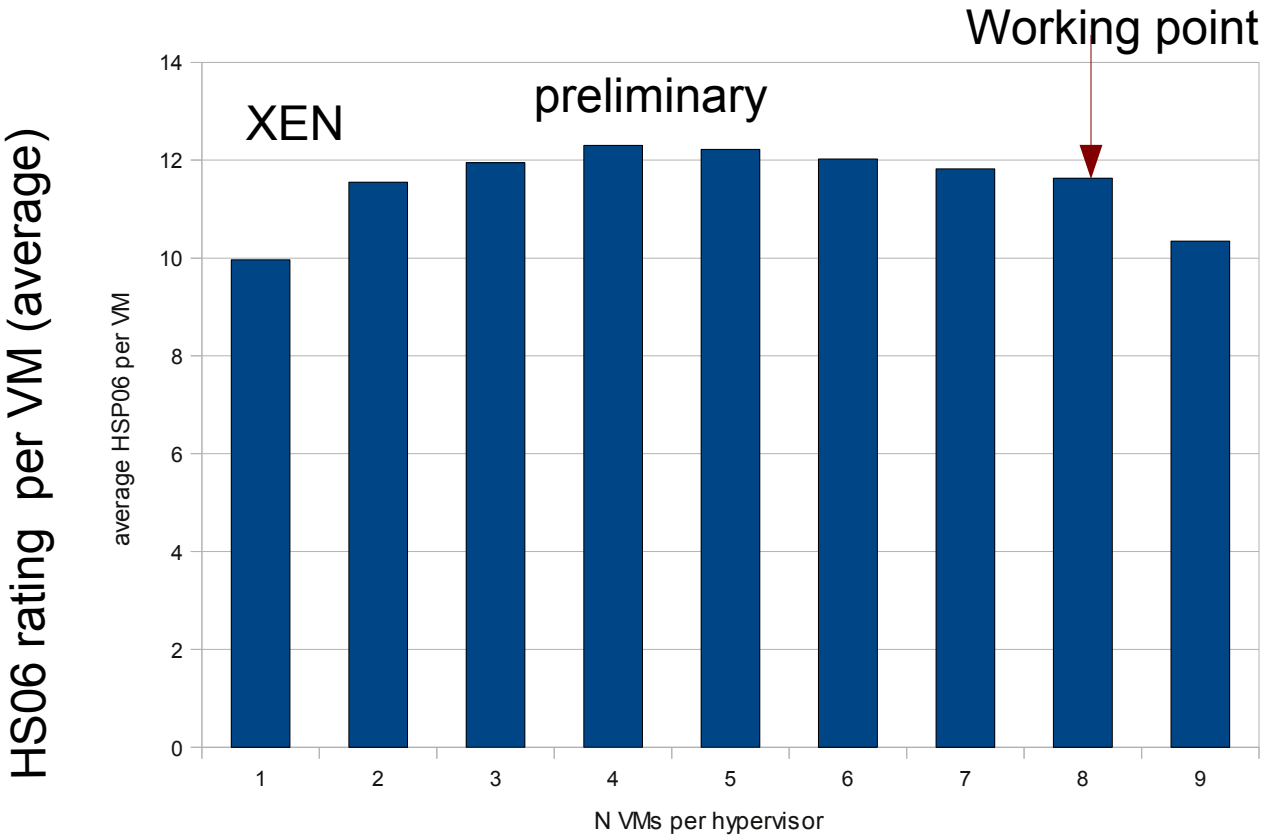
Physical box: 97 HS06



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

CPU performance: HS06 benchmark results (preliminary)

Test case : raise number of 1core VMs on a single hypervisor



Number of VMs per hypervisor

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

Disk performance: iozone tests

Test case:

- Raise the number of VMs running iozone on a single hypervisor
- Compare with running N iozone processes in parallel on a single physical server
  - Write performance OK
  - Read performance requires tuning

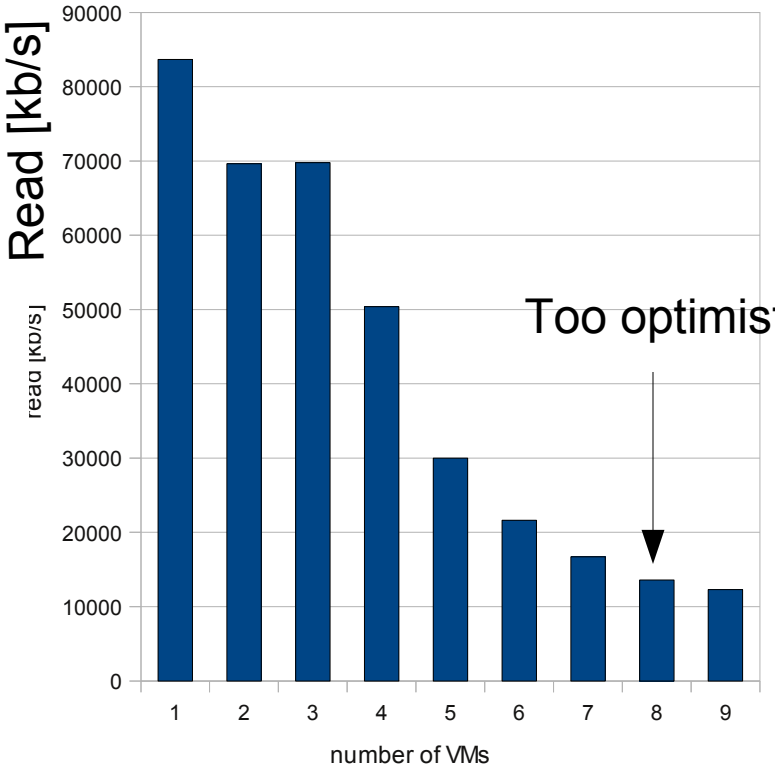
Tuning: eg. change read-ahead size in /sys/block/svdX/queue/read_ahead_kb

Command line parameters: iozone -Mce -r 256k -s 8g -f /root/iozone.dat -i0 -i1 -i

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

Disk performance: iozone tests

read performance



Compare with physical node:
~100MB/s read
~82MB/s write

~20% performance penalty

Needs further investigation !

Command line parameters: iozone -Mce -r 256k -s 8g -f /root/iozone.dat -i0 -i1 -i

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

Network performance

Still to do ...

… no big worries here though

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

Technical issues and challenges

- Scalability
  - General infrastructure
  - Image distribution
  - Provisioning system
  - Batch system
- Networking / public IPs

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

Technical issues and challenges

- VM CPU **accounting**/ CPU factors
- VM monitoring
- VM read performance under load
- VM space requirements
  - Scratch space
  - Swap space
  - AFS cache size

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

Batch virtualization

CERN IT Departmer
CH-1211 Genève 2
Switzerlan
www.cern.ch/i

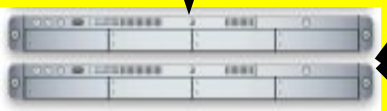
Batch virtualization: architecture

Job submission
CE/interactive

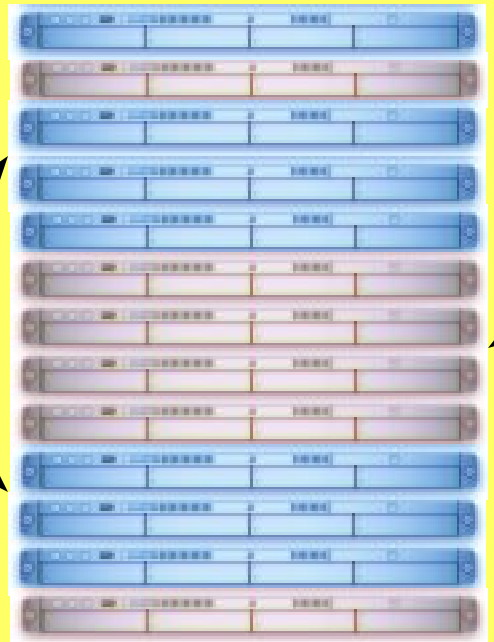
Grey: physical resource
Colored: different VMs

Centrally managed
Golden nodes

VM kiosk



Batch system
management



VM worker nodes
With limited lifetime

Hypervisors / HW resources

VM management system

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