Virtualization at CERN: A status report

HEPIX2011, GSI Darmstadt

Outline:

- **Part 1: CERN Virtual Infrastructure (CVI) status report**
- **Part 2: Internal cloud status and road map**

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

CERN IT Department

CERN Virtual Infrastructure (CVI)

Status update

CVI team

CERN – IT/OIS

HEPiX 2011 Spring meeting

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

Reminder: what is CVI?

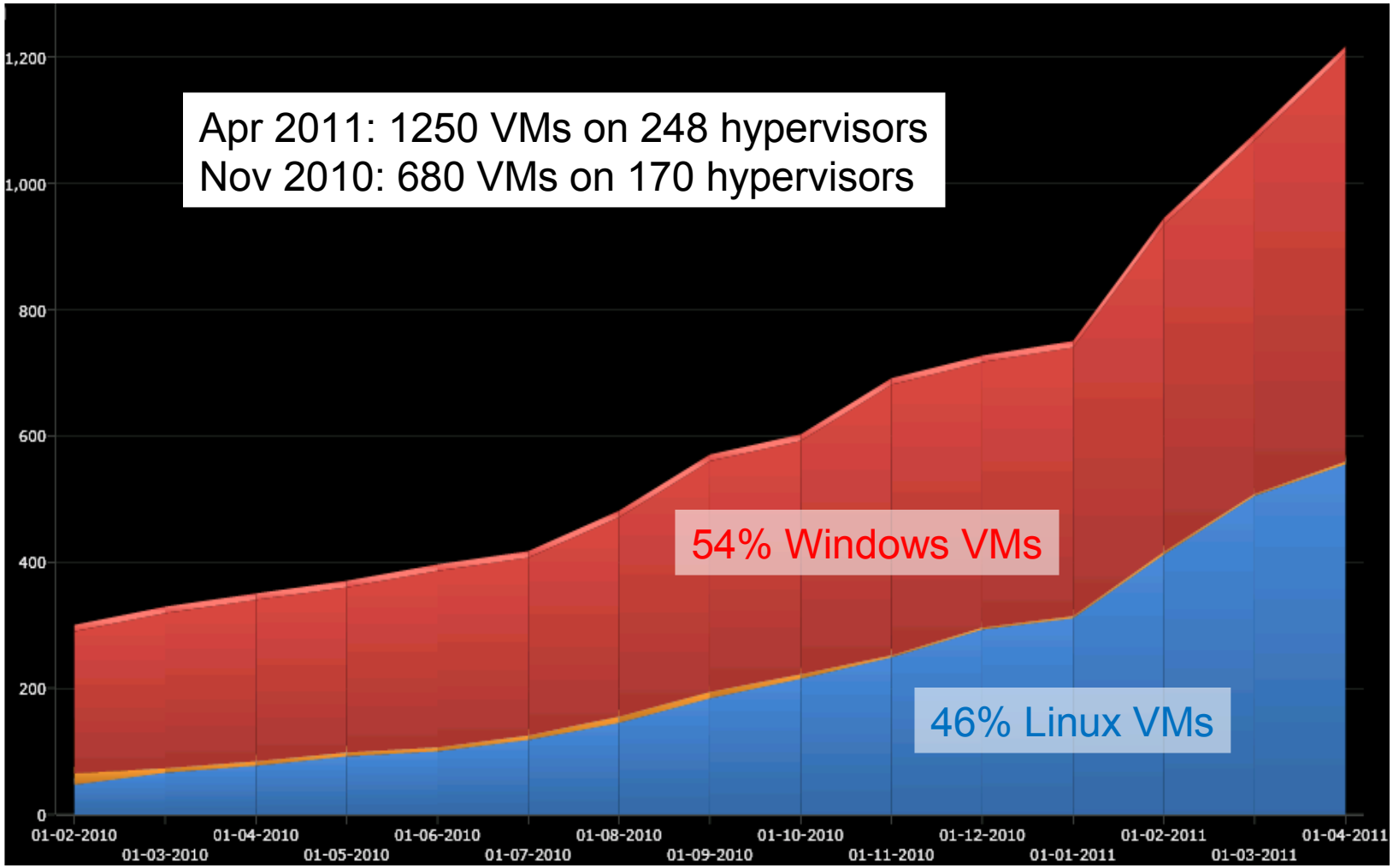
- The CERN Virtual Infrastructure custom virtual machines in the CERN computer centre
  - These VMs have a long-term lifetime of months/years
- User kiosk for requesting a VM in less than 30 mins
- Based on Microsoft's System Center Virtual Machine Manager (SCVMM)
  - Enterprise class centralized management
  - Rich feature set:
    - Allows grouping of hypervisors, with delegation of administrative privileges
    - VM migration, High availability
    - Checkpointing
    - PowerShell for administration / scripting

Updates since November 2010

- CVI 2.0 deployed in production
  - Improved stability
  - Added functionality (customer-supplied images, support for checkpointing)
- Hypervisors upgraded to Win 2008 R2 SP1
- Dynamic Memory Allocation
  - Allows for over-allocation of memory
  - Deployment for Windows VMs ongoing
- SLC6 templates made available
- Grow the service...

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

Spectacular growth continues!



Number of Virtual Machines per Operating System

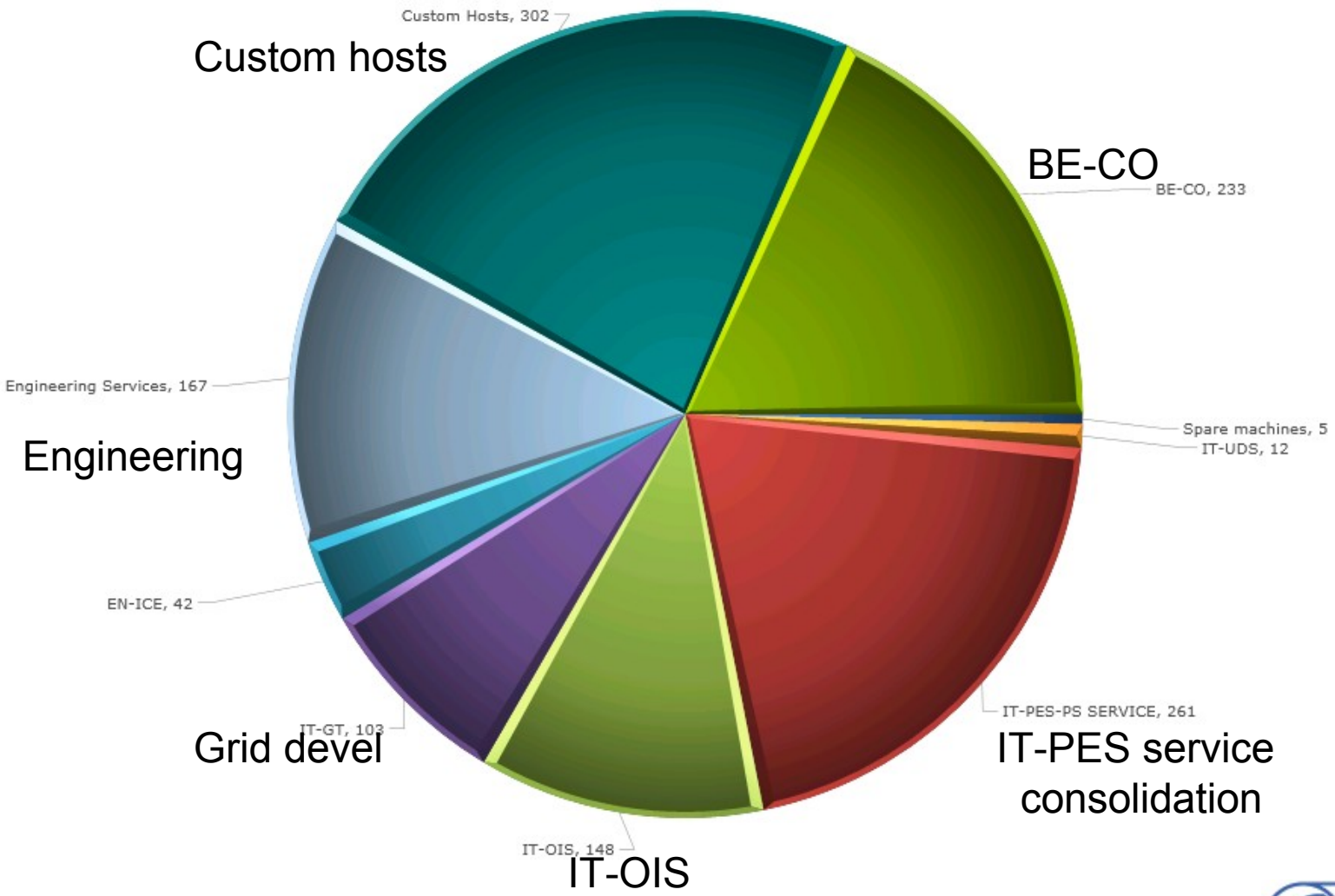
Customer groups

250 Hypervisors in **8 top-level** ‘hostgroups’:

- 7 dedicated hostgroups
  - For large, well-defined communities:
    - Physics Services, Engineering, Beams development, Grid developers, EN-ICE, Operating Systems Support, Conferencing
  - Admin privileges delegated:
    - To migrate VMs, modify virtual hardware, etc
- 1 Self-Service hostgroup
  - Shared ‘public’ resource
    - Many short-lived test/development VMs
    - But also production services

1250 VMs in 8 Customer groups

CERN IT Department



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

Linux VMs Integration Components

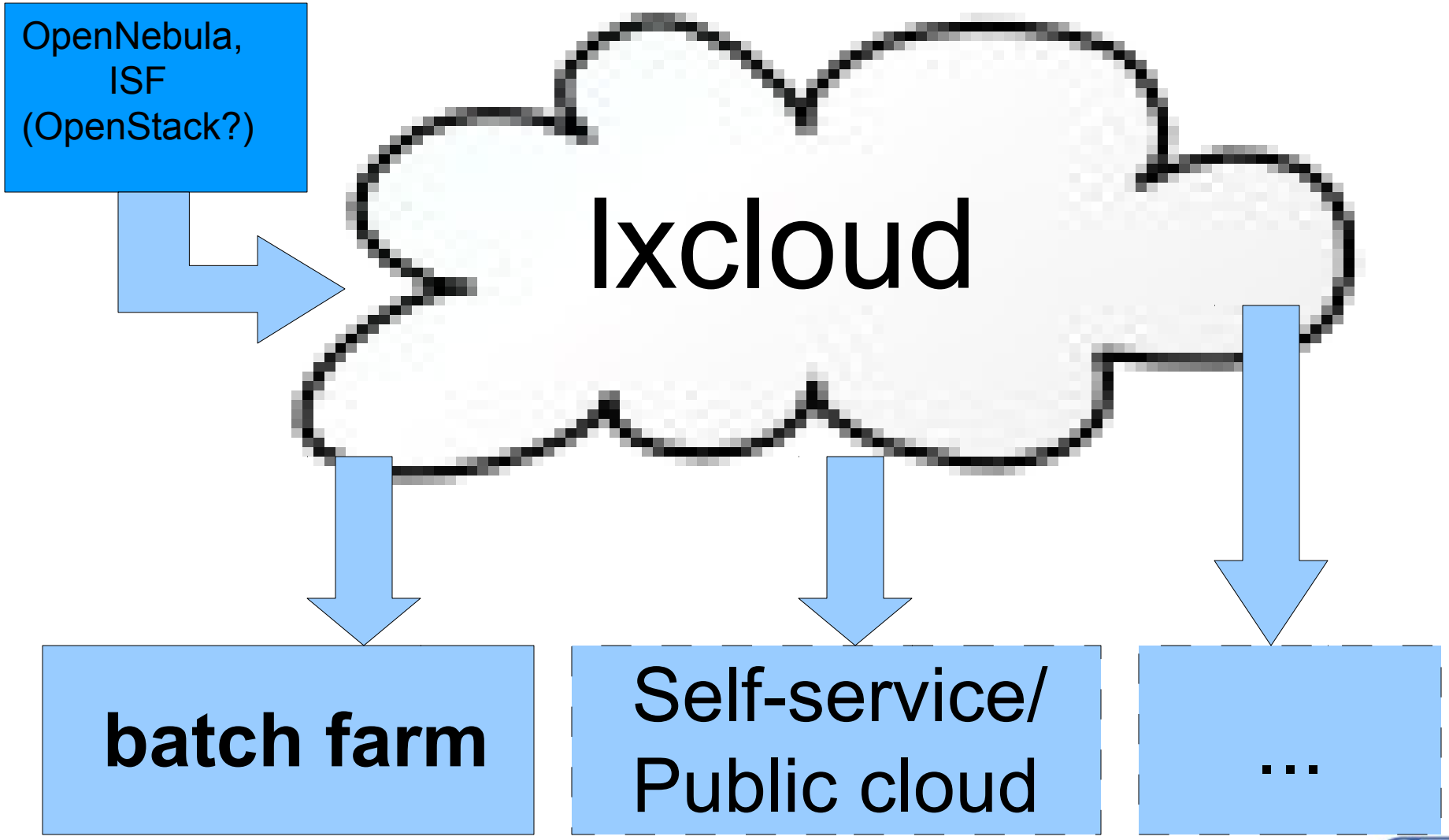
- In production for SLC5/RHEL5 since end-2011
- Very stable on 500 VMs
- Packaged RPMs by CERN Linux Team
  - yum install {kmod-,}mshvic
  - Includes kernel parameters to avoid clock drift
  - Time keeping problems are behind us!
  - Includes additional IC to enable synthetic mouse
- ‘Small disk IO’ performance to be understood
- Not available (yet) for RHEL6

Outlook for 2011

- Adding 3 blade enclosures + ISCSI storage
  - Capacity for 500 - 800 VMs
- Deploy Dynamic Memory Allocation
  - First for Windows 7 and Windows 2008 guests
  - Later for Linux guests
  - once Microsoft release this feature for Linux
- SLC6 support
  - … plus Integration Components
- CERN participate in VMM 2012 TAP program
  - Enhanced user interfaces
  - Many admin improvements

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

Part 2: CERN's internal cloud status



Part 2 : internal cloud

Reminder: What is it ?

- Highly scalable, Linux (KVM) based cloud-like infrastructure
- Optimimized for efficiency/speed
- Main building blocks:
  - Resources with pre-allocated VM slots (lxcloud)
  - Efficient internal image distribution system with bit-torrent
  - Local pre-staged images and LV snapshotting
  - VM management and deployment with OpenNebula/ISF

Part 2: internal cloud status

Main changes since last meeting in Ithaca:

- Infrastructure:
  - VMIC and image distribution scripts updates
  - Upgrades and packaging of provisioning systems
  - First glance at SLC6
  - Benchmarking
- Applications:
  - 96 virtual batch nodes in production since 12/2010
  - Prototype of public cloud interface with ONE

Image distribution and VMIC

New internal image distribution tools

- environment check for new images
- **no parallel "deploy"**
- rtorrent doesn't trigger "deploy";
- SQLite db to keep track of images and history
- improved error recovery
- comprehensive logs;
- comprehensive CLI
- **"expired" images management**;
- new images states allows better image control;
- **publish available images for provision systems**;

VMIC:

- Updates for image exchange from Clempson University
- Image creation and upload needs some work

Add Virtual Machine Image

VMI endorsement

Endorser: ---------

VMI identification

VMI URI:

VMI hash:

Status of the VMI

☐ This VMI is APPROVED to be run locally ☐ This VMI can be shared with other sites

Metadata about the VMI

Production date: Date: Today Time: Now

Hypervisor:

Endorsement date: Date: Today Time: Now

Implementation status

CERN IT Department

	Batch application	Hypervisor cluster	VM kiosk and image distribution	VM management system
Initial deployment	OK	OK	OK	OK
Central management	OK	OK	OK	OK
Monitoring and alarming	OK	OK	OK	OK

SLC6: first glance

Features and motivation to go for it:

- Required for
  - ONE master with public cloud interface (ruby-gems)
  - OpenStack evaluation
- More recent version of libvirt and qemu-kvm
  - Improved CPU architecture support (native included)
  - VM processes can be run as non-root
- Better support for KSM

SLC6: first glance

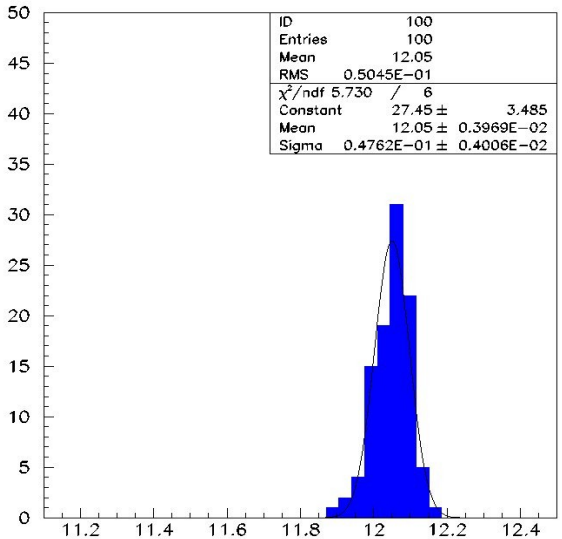
SLC6 for virtualization - status:

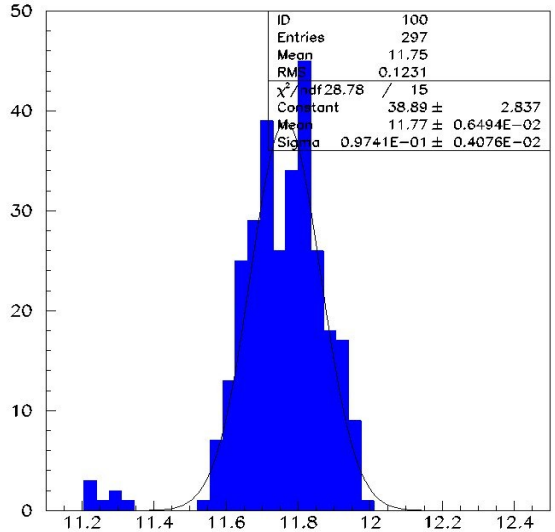
- Operating system is still a beta release only
- Basic stuff works
- Issues with multiple disks (disk “target” ignored)
  - Work around: XML file needs to be sorted by target
  - Recently spotted on SLC5 as well (libvirt update)
- Full quattorization still ongoing
  - Hardware monitoring
  - Serial console support

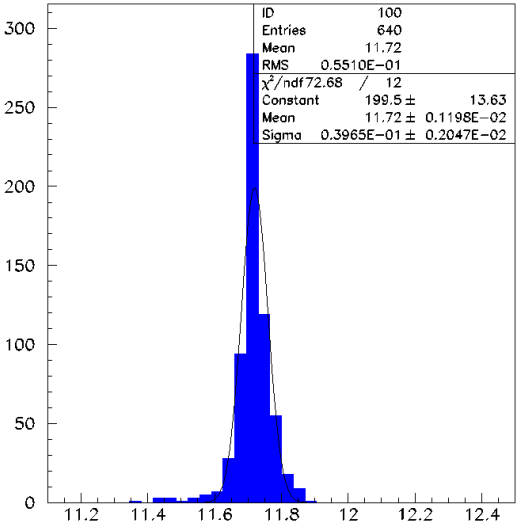
Benchmarking

CERN IT Department

HS06 tests







Bare metal:

- SLC5
- 2x L5520 Intel Xeon
- 2.27GHz

HS06=12.05/core

KVM

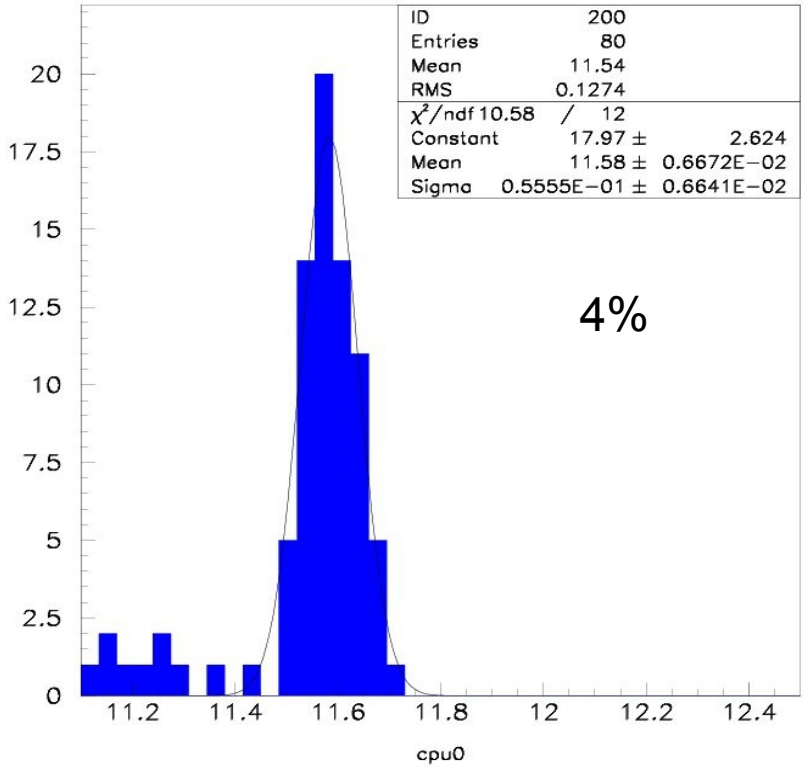
- HW as before
- SLC5/6 hypervisor
- 8 SLC5 guests
- No KSM,ept off (SLC5)
- Pinned VMs

HS06=11.4/core (SLC5)
HS06=11.8/core (SLC6)

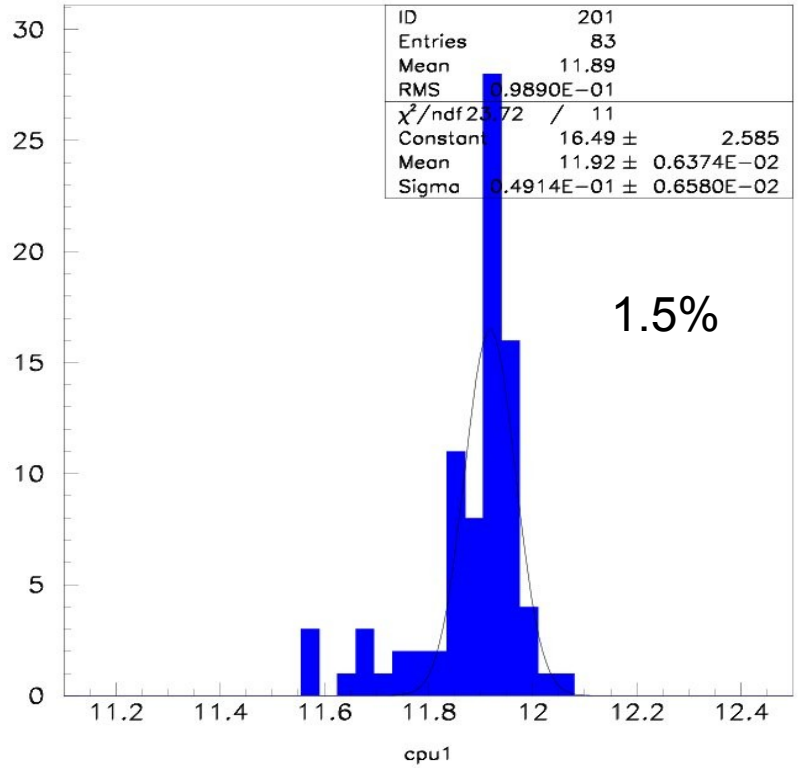
HyperV

- HW as before
- 8 SLC5 guests

HS06=11.7/core



CPU0 only

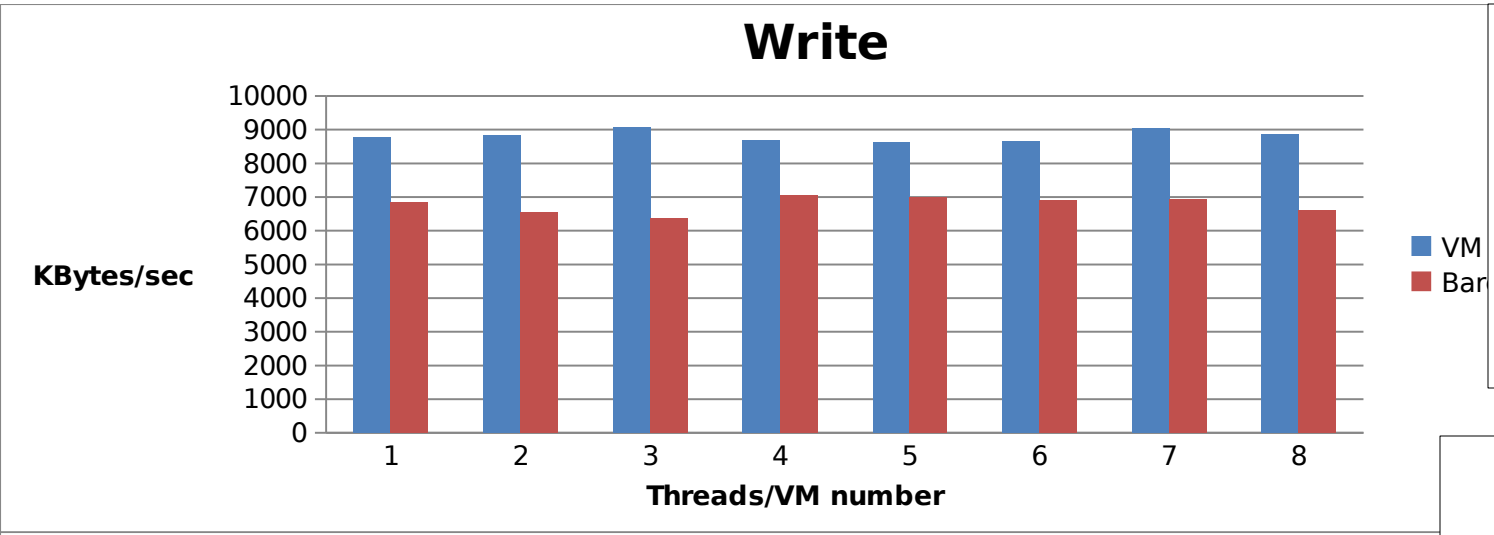


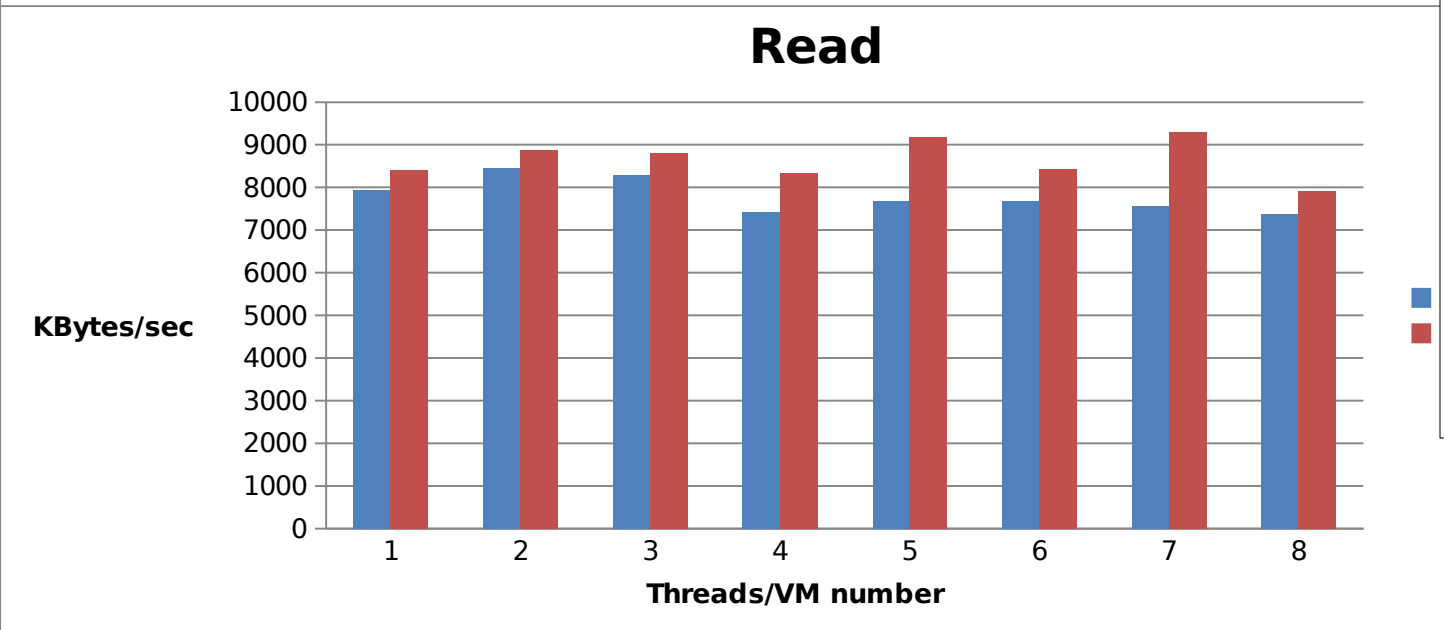
CPU1 only

I/O Benchmarking

CERN IT Department

-Mce -I -+r -r 256k -s 8g -f /pool/iozone_$i.dat$$ -i0 -i1 -i2





Analysis by Qiulan Huang (Chinese academy of science), December 2010

Notes:

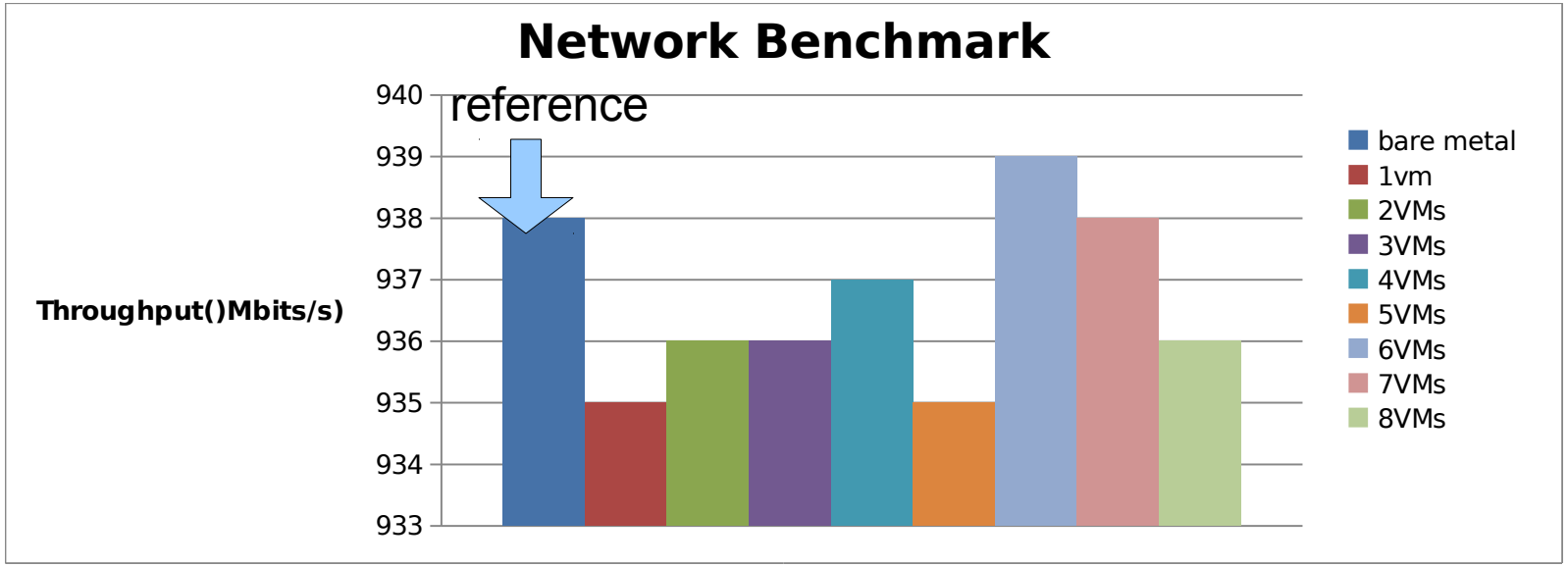
- Caching off
- SLC5 Hypervisor
- 20% penalty
- write worse
- block device disk on LV, exported to VM

Network Benchmarking

CERN IT Department

Analysis by
Qiulan Huang
(Chinese academy of science),
December 2010, CERN

Penalty <=1%



Iperf with TCP window size of 256k and 60s test time

Benchmarking conclusions

CERN IT Department

CPU benchmarking

- Best results of 2-3% requires tuning
- Ept=0 has fairly large effect on SLC5, less on SLC6
- Small effect by using the native CPU (SLC6)

I/O benchmarking house numbers (SLC5)

- Read/Write performance penalty 20%-30%

Network benchmarking (iperf, SLC5 GE)

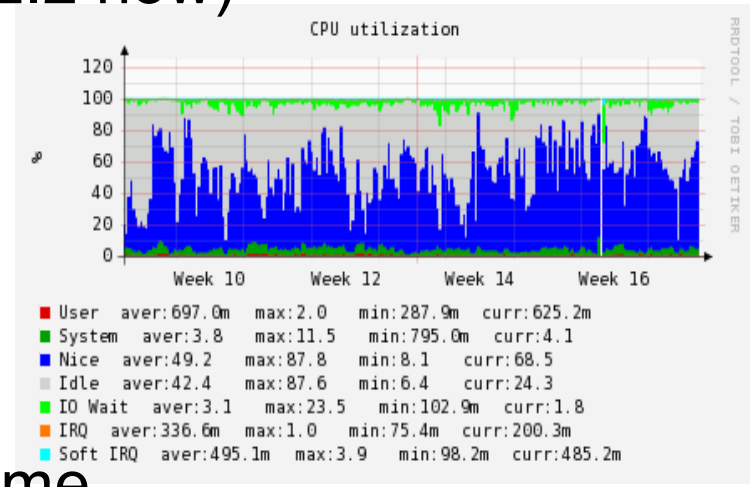
- Very small performance loss
- Possibly not significant within statistics

Virtual batch: production

CERN IT Department

96 nodes (12 hypervisors) in full production in public batch

- 6 hypervisors controlled by ISF (2.0)
- 6 hypervisors controlled by ONE (2.2 now)
- Short public and GRID jobs
- 1 VM / core and 1 job per VM
- Statically defined in LSF



Notes:

- Originally unmanaged during life time
- Now updates possible via Quattor (Golden Node)
- Image change only required for intrusive updates
- 12 identical physical nodes for job throughput comparison

Virtual batch: observations

CERN IT Department

Observation: More short jobs scheduled to virtual batch nodes

Job success rate:

Virtual nodes : 88%
Physical nodes: 82%

Delivered wall clock time:
(Ratio of time and total wall clock time seen by jobs)

Virtual nodes : 76%
Physical nodes: 81%

Possible improvements:

- Life time of VMs can be increased now
- Several improvements in boot sequence
- Machine renewal

Virtual batch: production

CERN IT Department

Impressions from 4 months of operations

- Roll out of intrusive interventions very convenient
  - Enforcement of updates possible via golden node
  - Change of lifetime for urgent easily updates possible
- Full integration into infrastructure is very convenient
  - Krb5 root access
  - Lemon monitoring, serial console access, ...
- Stable running over the whole period
- Job throughput measurements not entirely conclusive

Further plans

CERN IT Department

- Improve and increase virtual batch resources
- Prepare other applications to be run on the internal cloud
- Work on SLC6 deployment
- Migrate hypervisors to SLC6
- Make use of public cloud interface:
  - Offer default images via EC2 interface as a pilot
  - Support CERNVM images (proof of concept done)
- Work on image exchange between sites
- Evaluation of OpenStack

OpenStack

Why is it interesting ?

- Lots of interest from industry (60 companies)
- Big players: NASA and Rackspace
- Covers both CPU (NOVA) and storage (SWIFT)
- Comes with full EC2 compatible interface

Status:

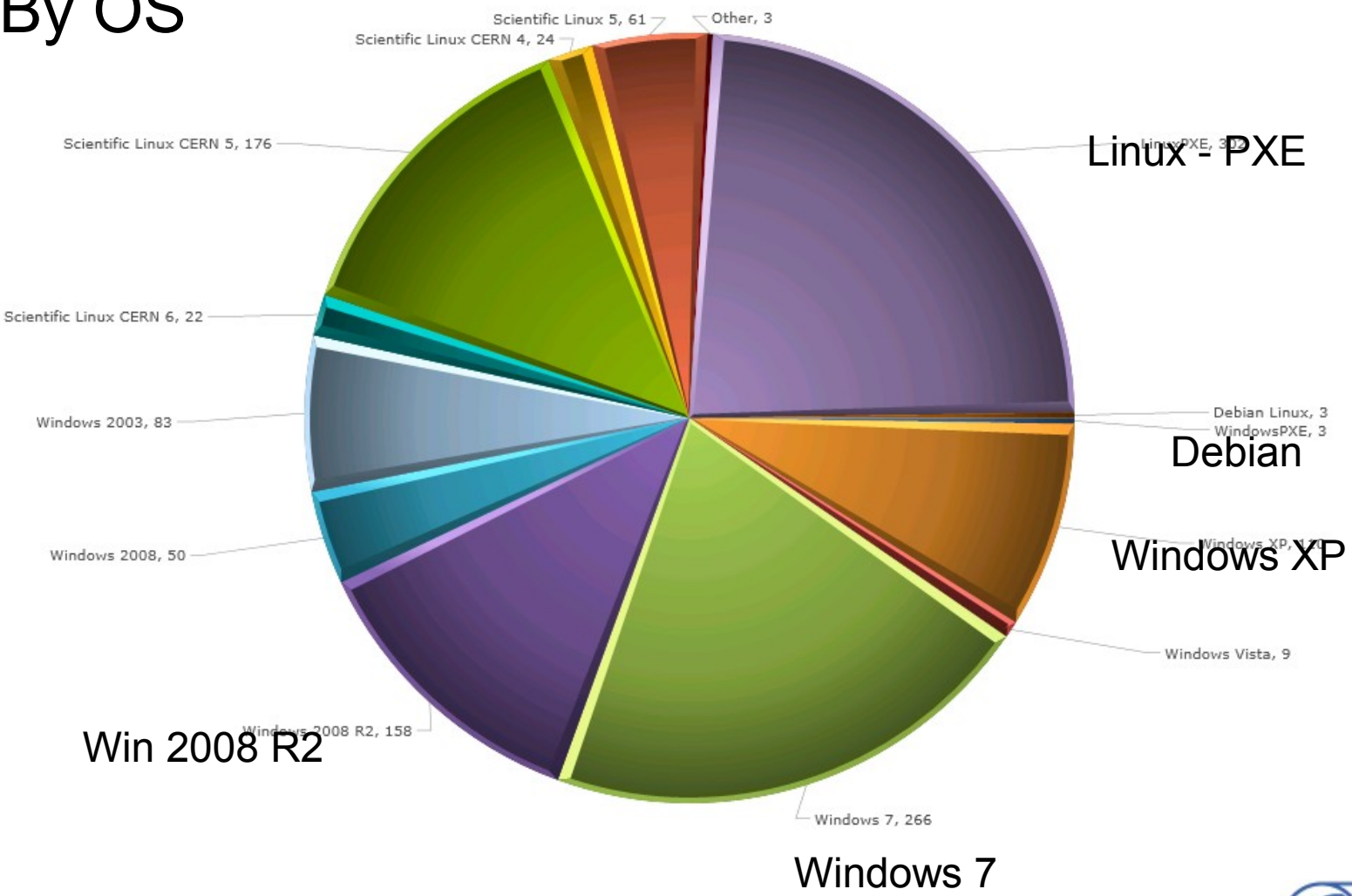
- Test nodes (server, hypervisors) available
- Considering SLC6 only (rpms available upstream
- Manual installation
- Just started with latest release (Cactus release)

Any questions

54%Windows, 46% Linux VMs

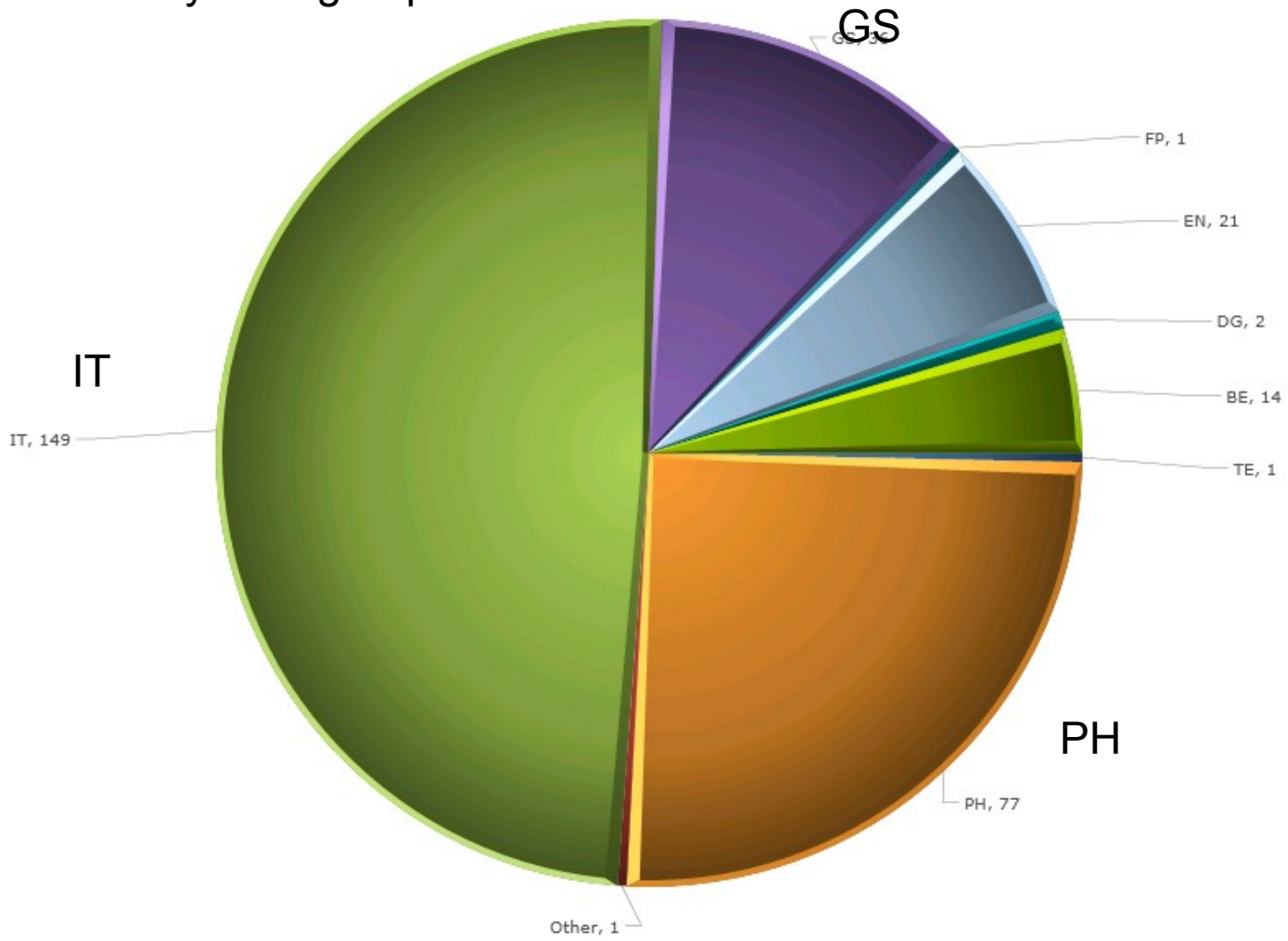
CERN IT Department

By OS



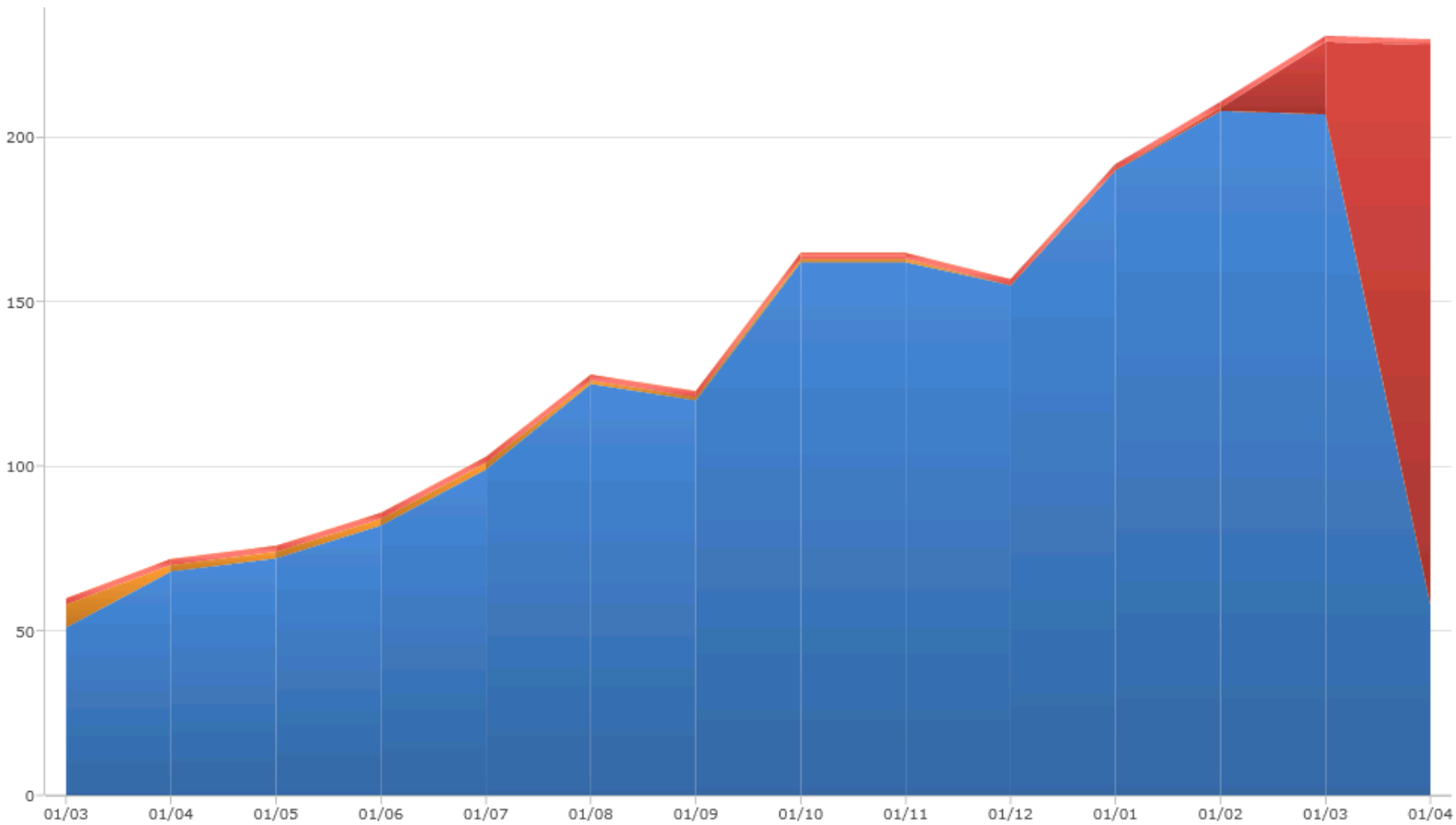
302 VMs in the Self-Service

Self-Service by user group



Number of Hypervisors

CERN IT Department

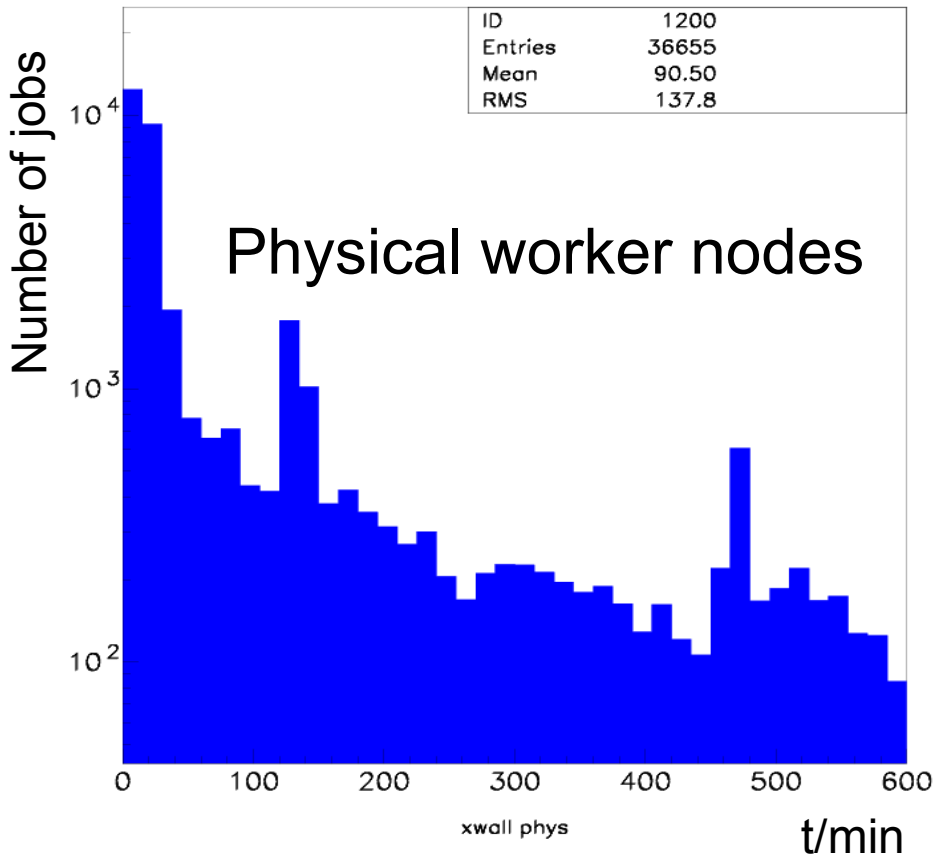


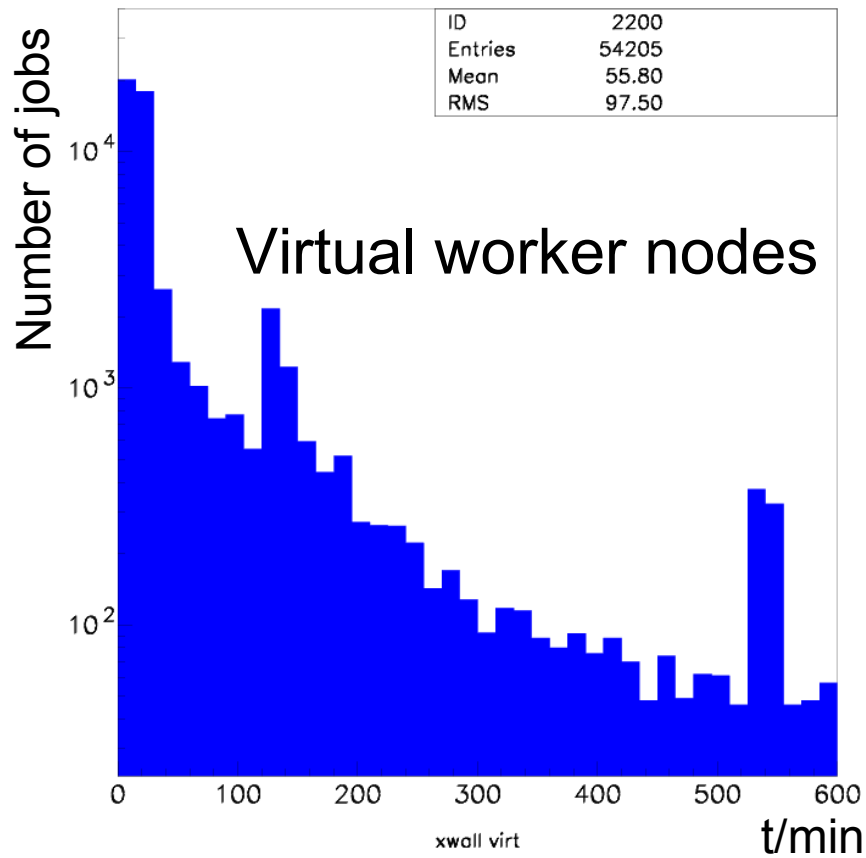
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Virtual batch: production

CERN IT Department

Wall clock time distributions by job





Virtual batch: production

CERN IT Department

CPU time distributions by job