The CERN internal Cloud status report

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

HEPIX2010, Cornell university

See also related presentations:

- Related presentations by Tony Cass, Romain Wartel, and myself
- Virtualization at CERN, HEPiX spring meeting 2010
- 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,*

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

Outline

- IaaS for CERN
  - Principles
  - Components
- Applications
  - Principles
  - Components
- Project status
- Summary

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

Principles

CERNs internal “cloud” prototype

… if we can call it a cloud

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

Requirements: the approach must

- be easy to manage
- smoothly integrate with the existing infrastructure
- be scalable
- be flexible for extensions

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

Principles

Design decisions:

- setup of an internal infrastructure for virtualization
- With the potential to become a public cloud

The initial deployment foresees that

- we deliver IaaS internally to **IT service managers**
- NOT (yet?) to end users
- the first and for now only customer is the **batch service**

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

The hypervisor cluster (lxcloud):

- Supports XEN and KVM (XEN being dropped now)
- Is **centrally managed** with a unique software setup
- Provides the infrastructure required for IaaS
- Runs the latest OS version to fully support the hardware

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

Components

The local image cache:

- Is a central place holding all approved images
- Is designed in collaboration with the HEPiX WG

See the presentation of Romain Wartel on VMIC

Internal image distribution

- Caching of images on the hypervisors
- Transfer using bit-torrent

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

The image provisioning system

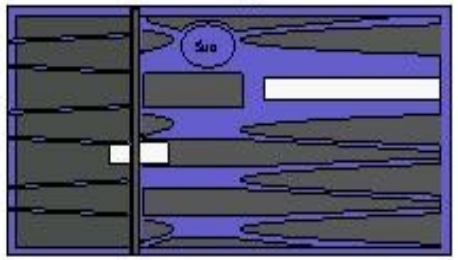
- Receives requests
- Selects a hypervisor
- Provisions the machine
- Controls and monitors it during its lifetime

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

Components

CERN IT Department

Incoming requests



VM provisioning system

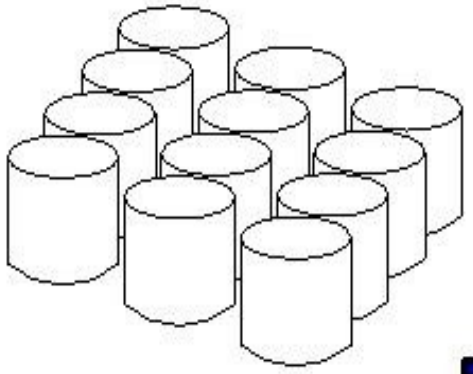
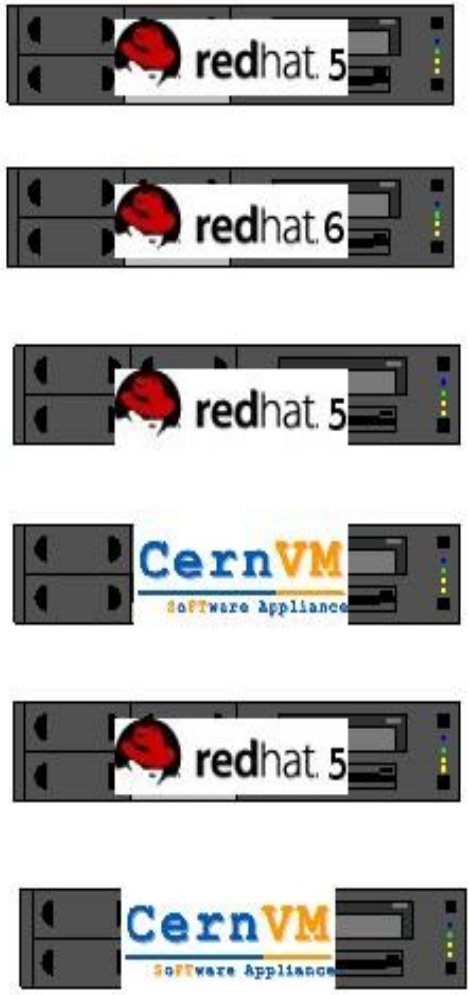


Image repository



VM instances

Components

ISF: InfraStructure sharing facility, Platform computing

- pVMO (physical virtual machine orchestrator)
- RIA (resource interface adapter)
- Native support modules (eg. for VmWare)
- AC (adaptive cluster) (requires LSF 8.0)

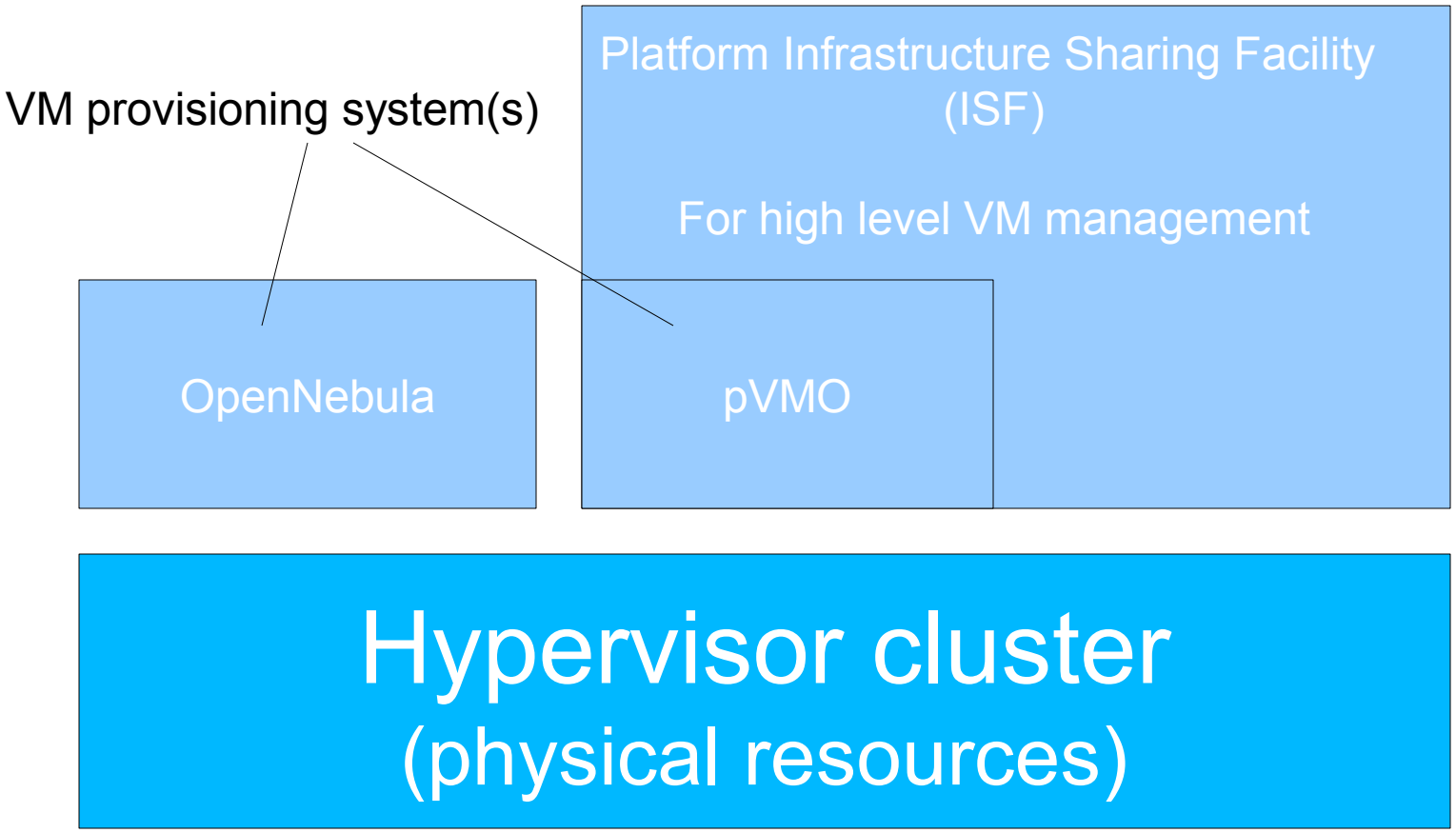
ONE: OpenNebula

- Modular OpenSource solution
- Now also professional support
- Cloud interfaces

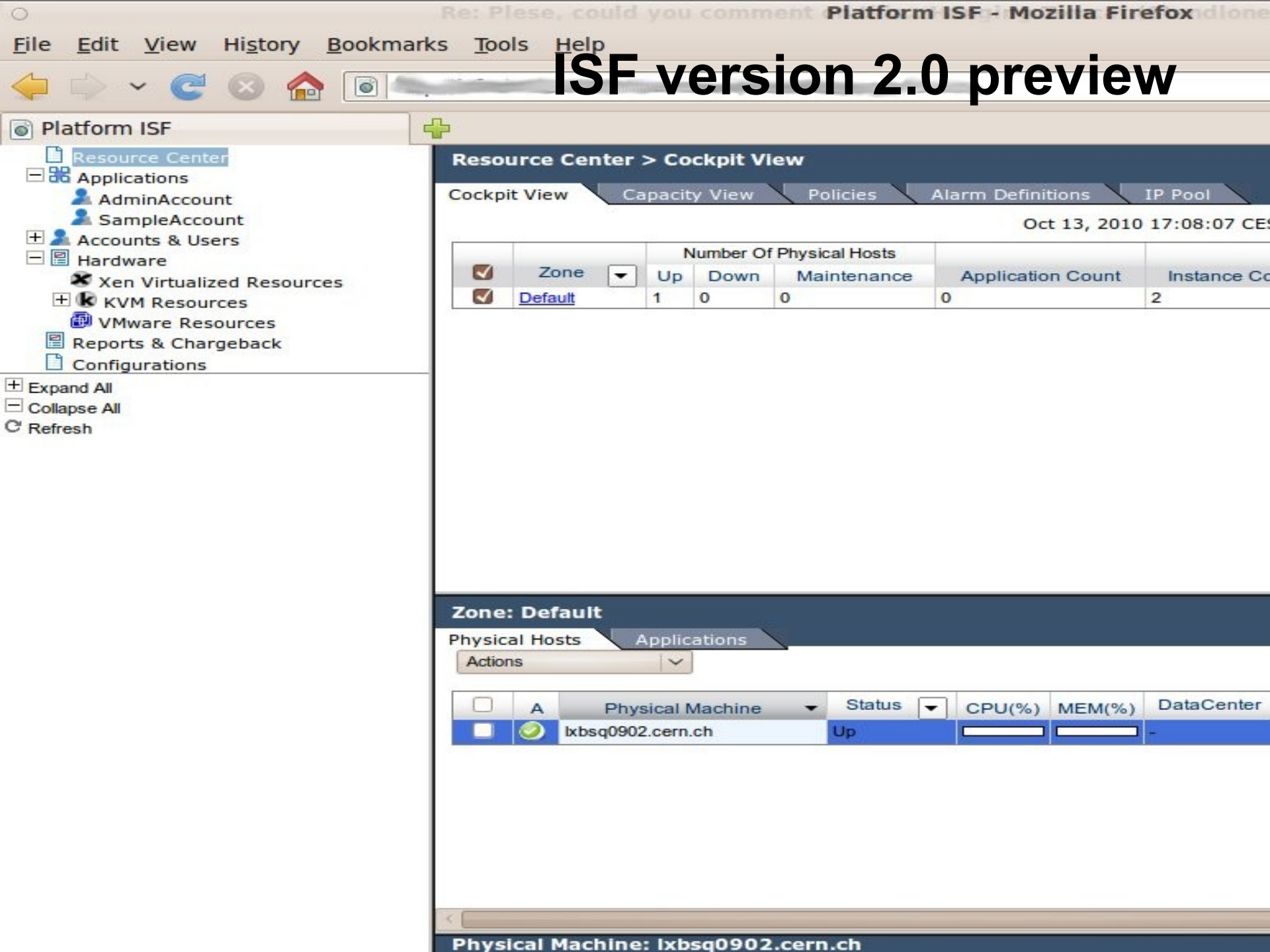
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

ISF and ONE relationship

pVMO and ONE do approximately the same thing



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


Platform ISF - Mozilla Firefox
File Edit View History Bookmarks Tools Help
ISF version 2.0 preview
Platform ISF
Resource Center
Applications
AdminAccount
SampleAccount
Accounts & Users
Hardware
Xen Virtualized Resources
KVM Resources
VMware Resources
Reports & Chargeback
Configurations
Expand All
Collapse All
Refresh
Resource Center > Cockpit View
Cockpit View
Capacity View
Policies
Alarm Definitions
IP Pool
Oct 13, 2010 17:08:07 CE
Number Of Physical Hosts
Zone
Up
Down
Maintenance
Application Count
Instance Co
Default
1
0
0
0
2
Zone: Default
Physical Hosts
Applications
Actions
A
Physical Machine
Status
CPU(%)
MEM(%)
DataCenter
lxbsq0902.cern.ch
Up
-
Physical Machine: lxbsq0902.cern.ch

Batch resources as application

Or: virtualization of batch resources

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

Requirements

- **transparent** to the end users
- possible to mix virtual machines and physical worker nodes
- possible to start small and grow
- scalable
- able to **solve the operational issues**

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

A Golden node:

- Is a centrally managed virtual machine
- Is a clone of physical machines running the same services
- Does not execute jobs by itself
- Serves as source for virtual machine creation

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

Adding automation and elasticity(*)

- **Limit the active life time of the virtual machines (to 24h)**
- Automatically Monitor the VM population and restart nodes as needed

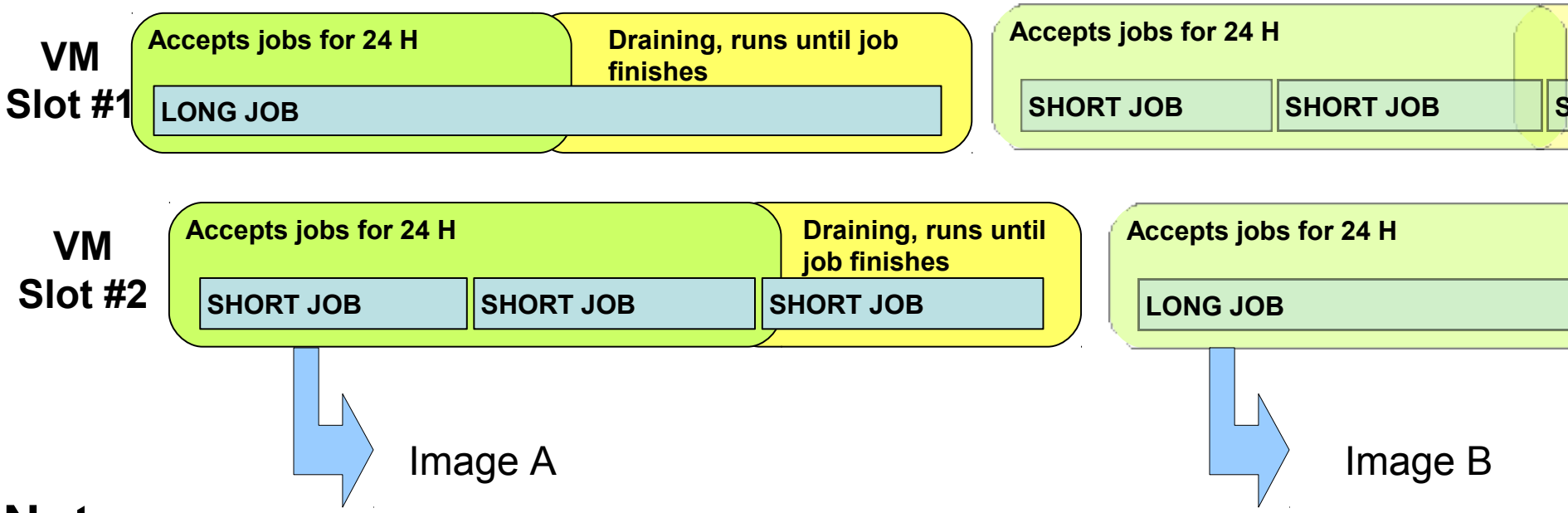
(*) Batch service managers point of view

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

Application components

CERN IT Department

How to manage **intrusive interventions** in a fragmented setup ?



Notes:

- NEW virtual machines always start with the latest image
- Image A and Image B can correspond to different OS versions

Feasibility studies

The question is:
will it work at the required scale ?

Or: will it ever fly ?

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

Scalability at all levels

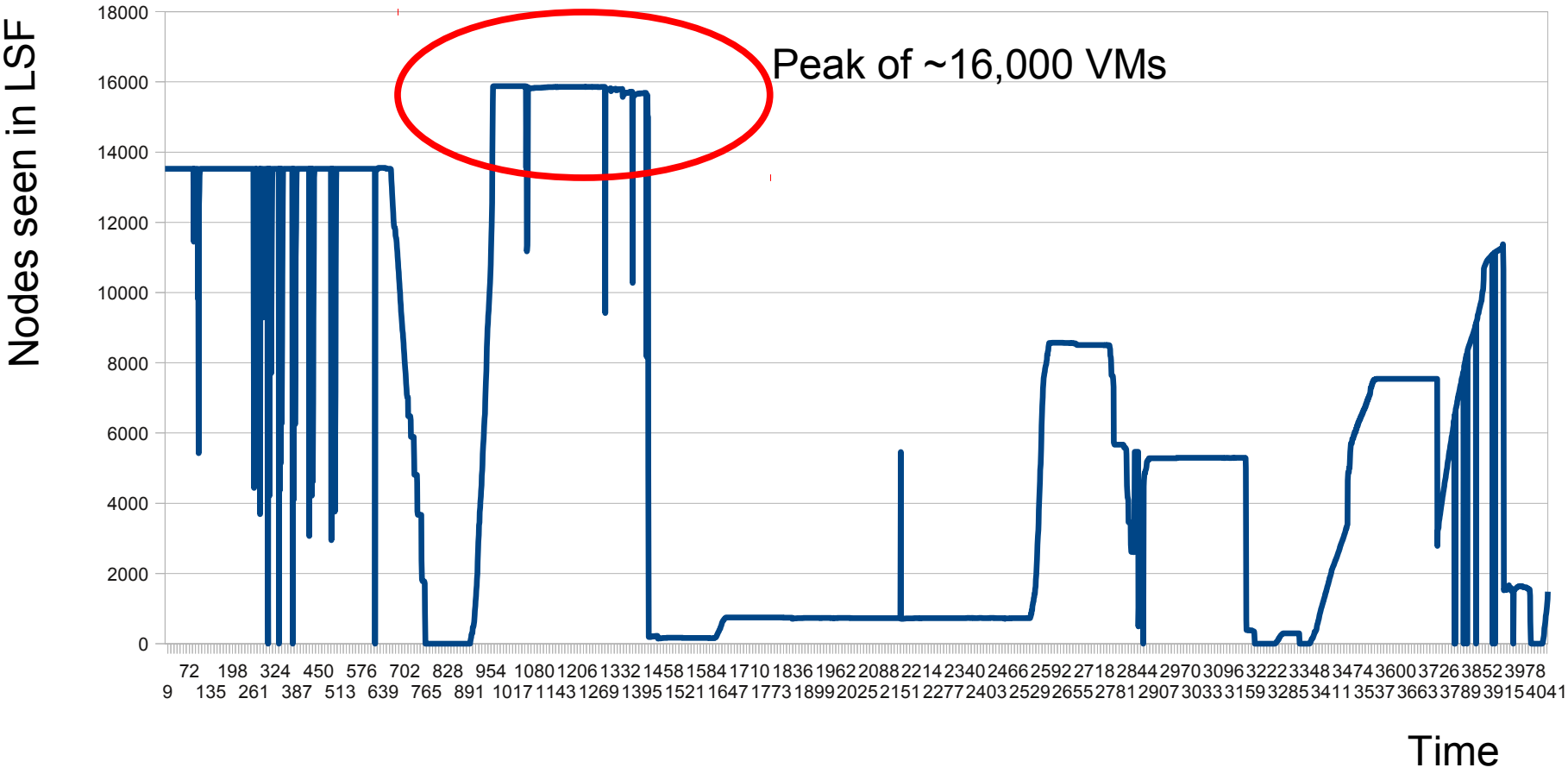
- General infrastructure (eg. ldap)
- Networking: number of IPs
- Virtual machine provisioning systems
- Image distribution efficiency
- Application level:
  - Virtual batch nodes
  - Batch system scalability → dedicated presentation

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

Feasibility studies

CERN IT Department

Virtual machine provisioning system tests

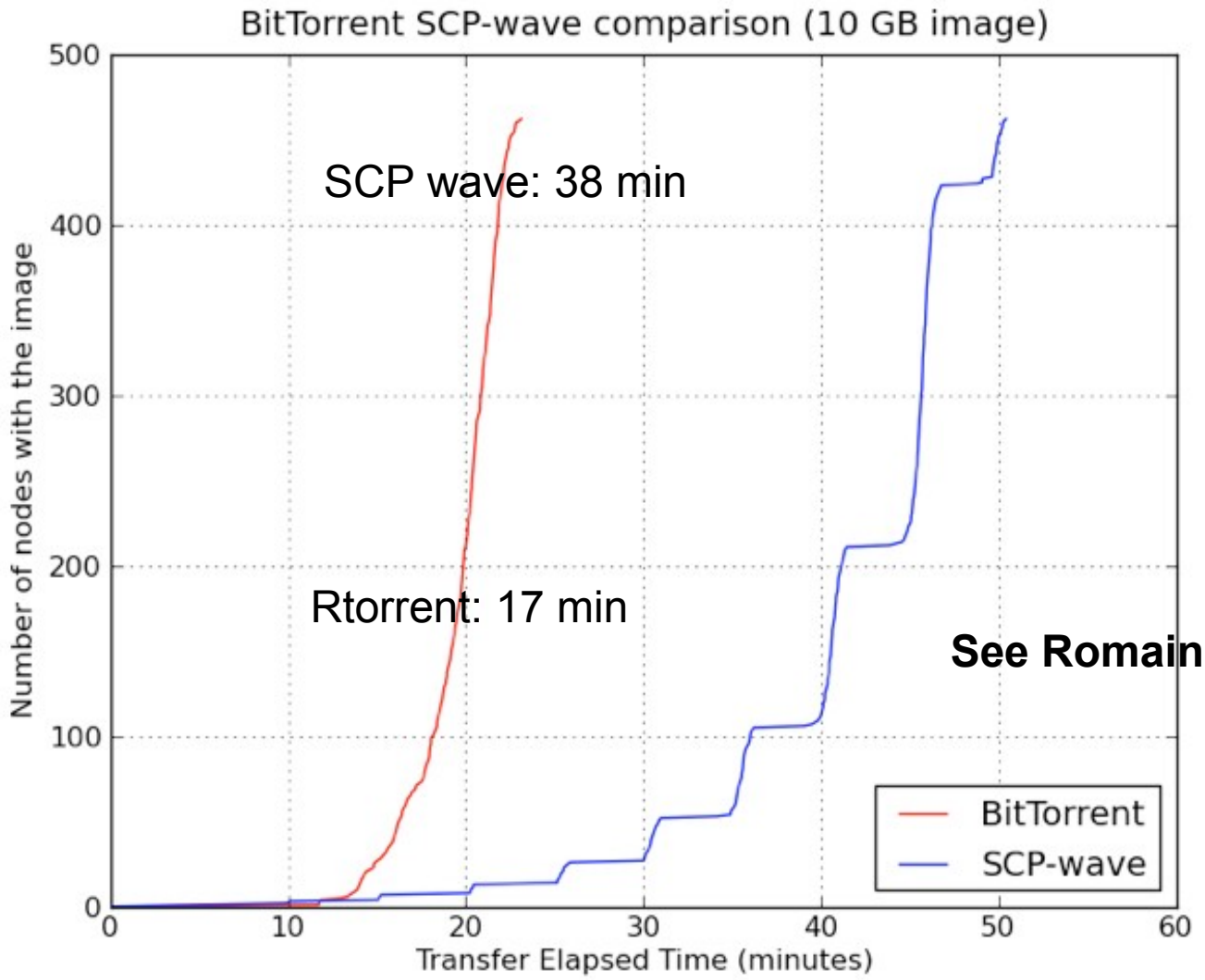


Feasibility studies

- OpenNebula:
  - Up to 16,000 VMs started, all available slots filled
- ISF/pVMO:
  - Up to ~10,000 VMs started
  - Limitation is understood and no longer present

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

Image Distribution



See Romain's presentation

Feasibility studies

Virtual batch nodes with real userjobs

Currently we have:

- 12 KVM hypervisors ready for production
- hypervisors with private IPs
- And 8 VM slots with public IPs
- 2 KVM hypervisors deployed into production
- 16 KVM VMs running as public batch nodes for ~4 weeks

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

Virtual machines in public batch

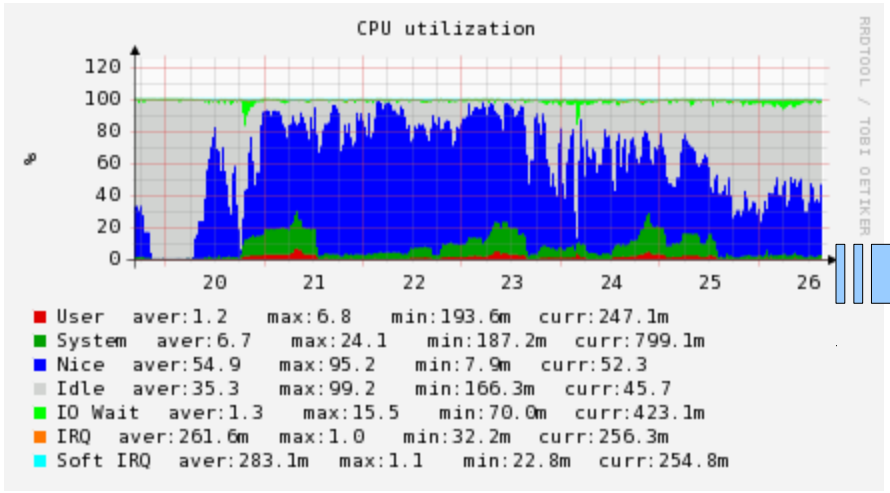
New feature which is:

- not seen on physical worker nodes
- caused by network monitoring tools used at CERN
- under investigation

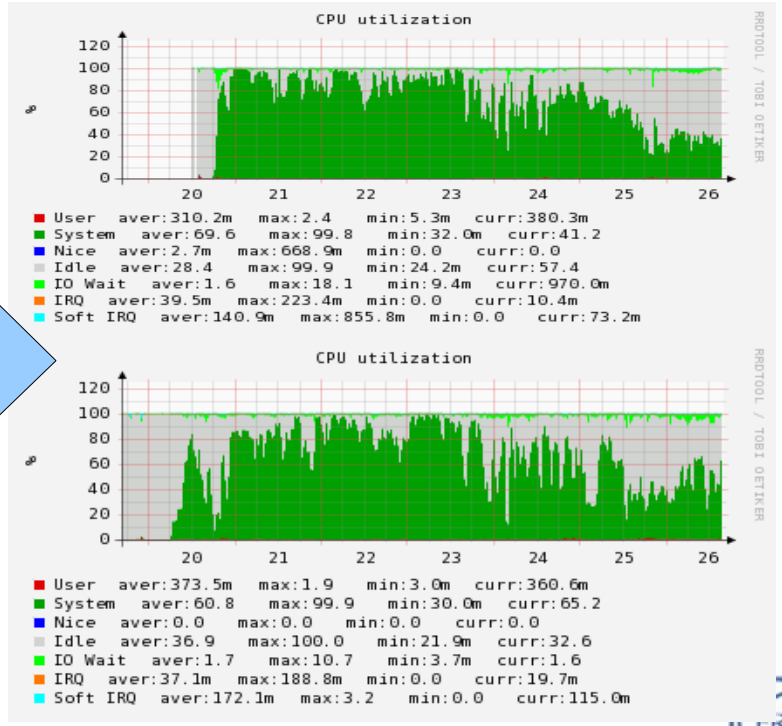
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Changes since last HEPiX

- Dropped support for SLC4
- Favor KVM over XEN
  - KVM is the emerging technology
  - XEN support for in RHES/SL6 is questionable
  - CPU usage monitoring works out of the box with KVM



16 VMs



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

KVM in SLC5, impressions

- Fairly stable
- Some performance issues
  - CPU performance O(10%) low
  - I/O performance even more
- Requires tuning, and eventually a newer version
- KVM processes stay in S mode
- Utilization turns up as system CPU, not user CPU

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

Impementation status

CERN IT Department

	Hypervisor cluster	SLC5 virtual batch nodes
Quattor managed	OK	OK, via golden node
Lemon monitored	OK	OK
Auto-registration	OK	OK
Central maintenance	OK	OK
ISF support	OK	OK
ONE support	OK	OK

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	ISF OK, ONE missing
Monitoring and alarming	OK	OK	OK	missing

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

Plans and remaining technical issues:

- VM CPU **accounting** / CPU factors
- ISF adaptive cluster may come late
- Still some development needed for VM renewal
- Performance considerations
- For scaling up:
  - Krb5 authentication for root access may not scale
  - May need our own lemon instance ?
  - LSF doesn't scale up to 15k nodes, redesign is needed

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

Summary

CERN IT Department

Quite some progress during summer 2010

- Only one application so far (lxbatch)
- 16 VMs under test with real payload
- Up to 96 by Christmas
- Scale up
  - Expected for 2011
  - Depends on experiences


Any questions
?

Benchmarking

CERN IT Department

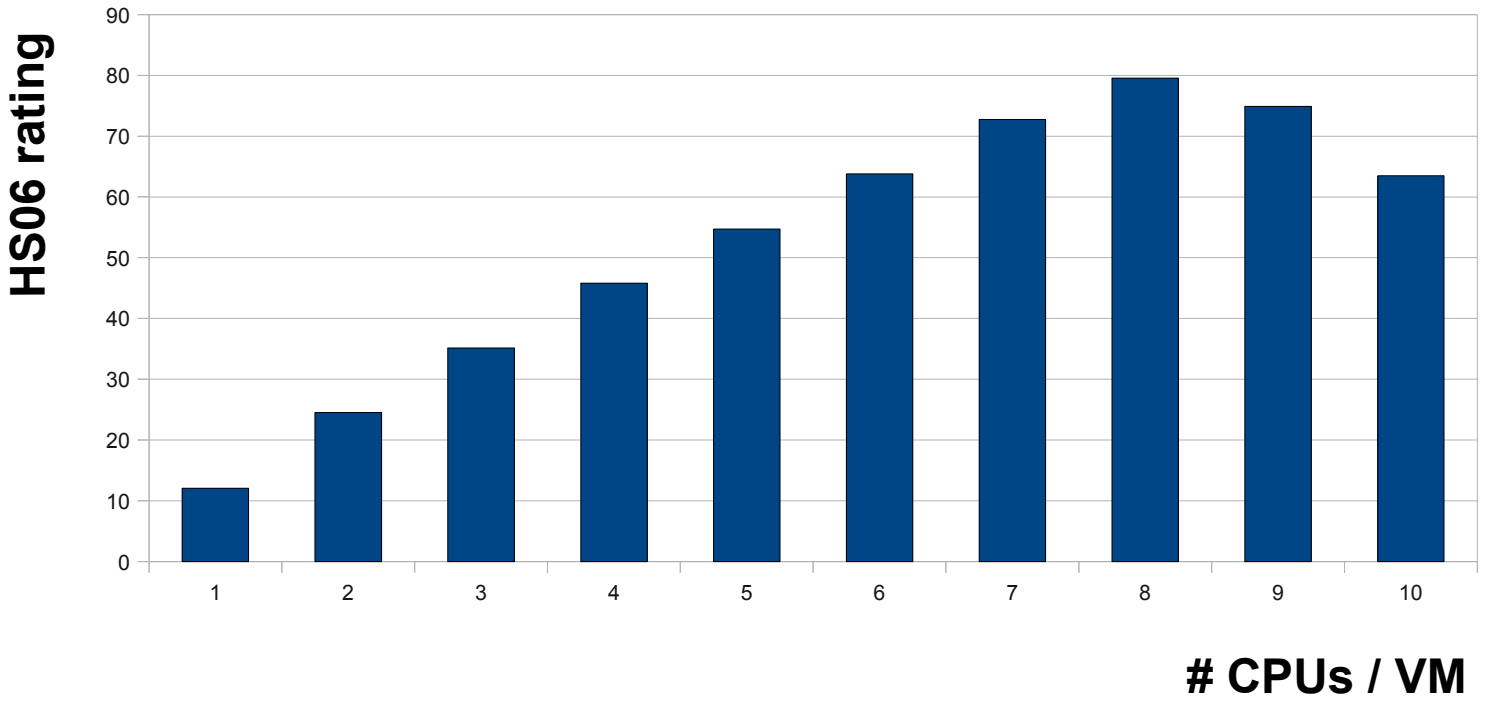
CPU benchmarks: HS06
I/O : iozone

- No tunings applied
- Up to 17% penalty in CPU with current KVM
- Same order or worse for IOzone
- Needs further investigation and tuning

Tunings: kvm kernel module options, CPU affinity

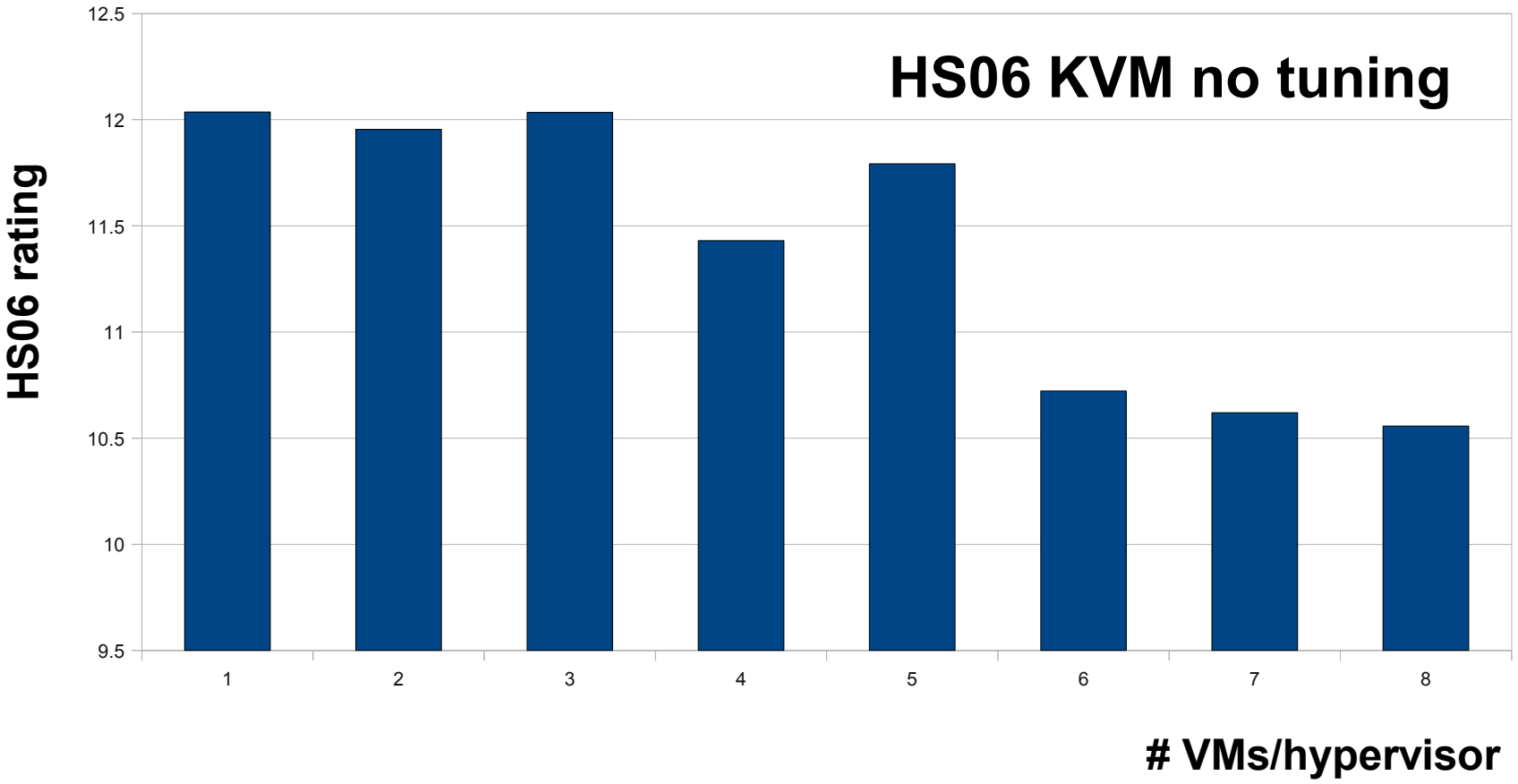
Benchmarking

HS06 KVM, no tuning applied



Bare metal: 8 Cores, 8 processes: ~96 HS06, so 17% performance penalty

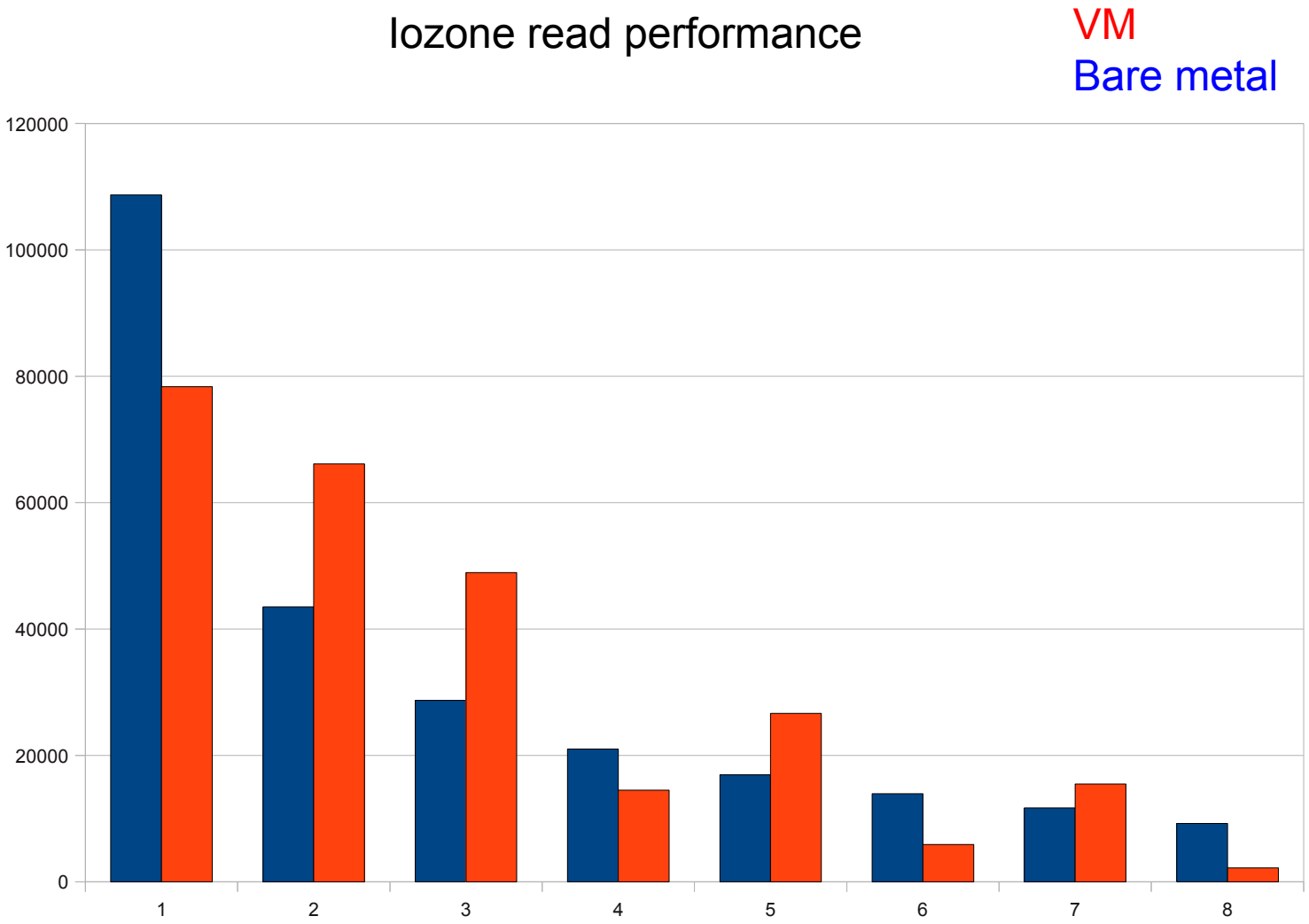
Benchmarking



Bare metal: ~12 HS06/core, so 12% performance penalty

I/O Benchmarking

KVM

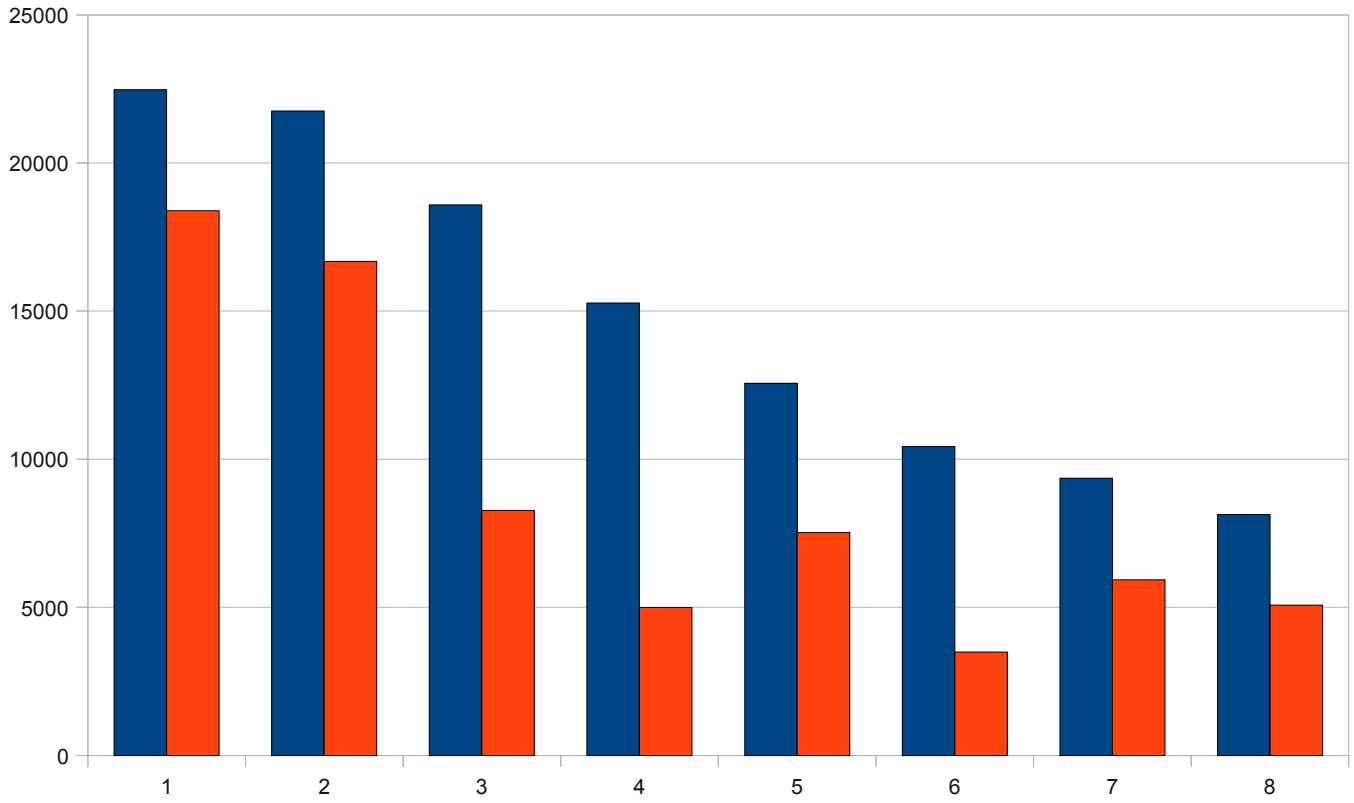


I/O Benchmarking

KVM

Iozone write performance

VM
Bare metal



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