Virtualisation in CERN-IT: Overview and Service Consolidation

Ewan Roche, Ulrich Schwickerath,
Manuel Guijarro, Helge Meinhard et al.

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

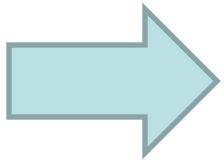
HEPiX Fall 2010 Workshop – Cornell University

Outline

- Virtualisation use cases at CERN-IT
- Details about virtualisation for service consolidation

Why virtualise?

- Better use existing resources and available power
- Simplify physical machine park
  - Far fewer (one?) OS on physical machines
- Ease intrusive interventions
  - E.g. emergency kernel upgrades
- Ease life-cycle management of physical machines



Remove strong coupling of services and physical machines

Virtualisation instances

- Service consolidation (PES and OIS)
  - Based on Hyper-V, mainly targeted at long-term (potentially critical) services such as VOboxes
  - Linux (SLC) guest OS, quattor managed
- Batch virtualisation (PES)
  - Based on KVM and ISF or OpenNebula, targeted at batch jobs, usually not critical
  - Linux (SLC) guest OS, quattor managed
  - (see dedicated talk)
- Self-service kiosk (OIS)
  - Based on Hyper-V, mainly targeted at short-term or development use
  - Windows or Linux guest OS
  - (see dedicated talk)
- Dedicated testbed (GT)
  - Based on XEN, targeted at developing Grid middleware
  - Various Linux guest OS
- …

Why different platforms?

- Service consolidation requires
  - Large-scale stable service
  - Associated “local” storage, shared across hypervisors
  - Transparent (live) migration of VMs from one hypervisor to another
- Batch virtualisation requires
  - Very large-scale service
  - Flexible VM provisioning and instantiation

Chosen platforms

- Service consolidation: Microsoft SCVMM on top of Hyper-V
  - Extended management layer for long-term provisioning of virtual machines and storage
    - Supports transparent migration of VMs
  - Alternative solution (industry preferred): VMware
    - Difficult to obtain affordable licensing schemes
- Batch virtualisation: Platform ISF or OpenNebula on top of KVM (or XEN)
  - Flexible provisioning schemes for short-term VMs
  - Provisioning on demand supported

Medium to long-term view

- Coexistence of two different platforms is (hopefully) temporary
- Keen in a platform that supports all different requirements
  - Must be affordable at our scale
  - Open-source solutions favoured by some people
- Vision: all CERN-IT services should be virtualised
  - They can run at CERN, at the current hosting centre, some remote T0 extension, …
  - All services use the same platform
    - One large pool of shared resources
    - Maximum decoupling of services from physical machines

Common short-term goals

- Seamlessly integrate virtual machines within the computing infrastructure in a sustainable manner
- Leverage the advantages of virtual machines in a manner transparent to service managers and end users
- Provide a level of service that is equal to or better than currently possible with physical resources
- Achieve complete equivalence between physical and virtual resources

Pain Points of Service Hosts

- Full physical machine often overkill
  - Does a Web server need >= 8 cores?
  - Sharing physical hosts saves power

Several VMs per physical machine

- Hardware failures/interventions cause service downtime
  - … unless two servers are configured in active-active or failover mode
    - But this aggravates the under-utilisation problem!

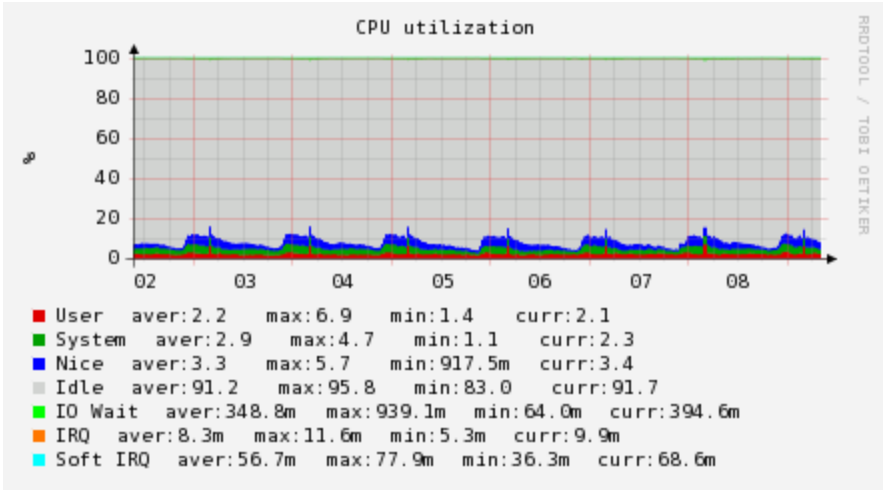
Live migration of VMs

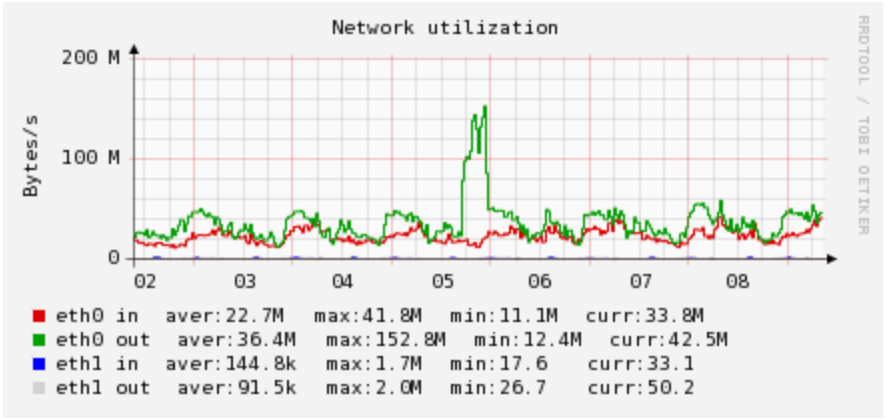
- Retiring machines and replacing them with new hardware is difficult
  - IT-PES responsible of hardware and basic OS, application software often under user control

Live migration of VMs

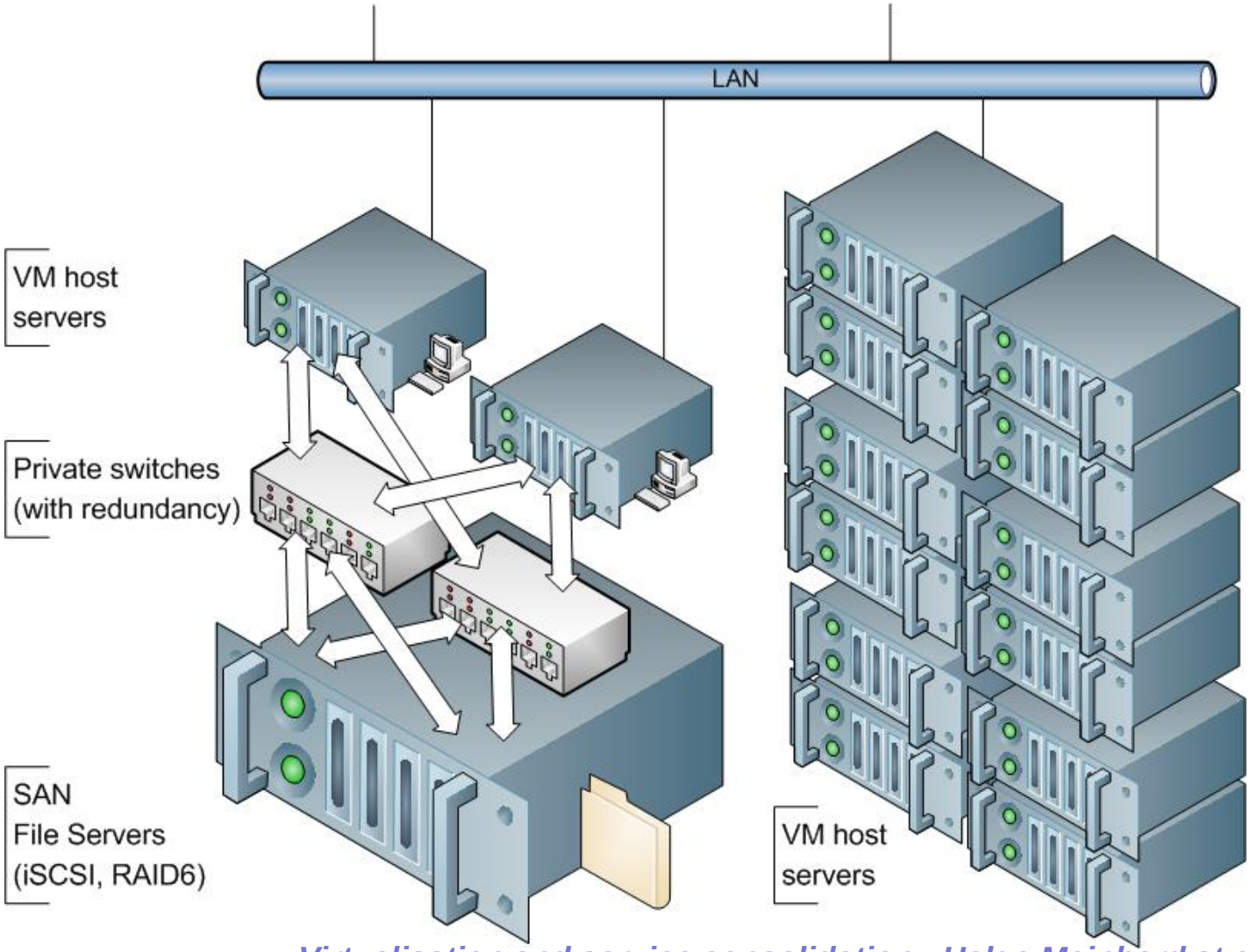
A perfect candidate?

- The voatlas cluster – 138 nodes
- Most CPU, network and disk loads appear to be from nightly builds
- Otherwise very low IO rates
- Even a consolidation factor of 2 saves a large amount of resources
- In industry a consolidation factor of at least 4 is normally possible for servers with no impact on performance





Hardware (1)


LAN
VM host servers
Private switches (with redundancy)
SAN File Servers (iSCSI, RAID6)
VM host servers

Hardware (2)

- Servers: Industry-grade blade servers with good levels of redundancy (PSU, disk drives)
- Storage: Resilient iSCSI devices connected to servers via redundant private network
- Storage devices configured as SAN islands with no interconnection between islands
- Hardware is high quality and provides protection against a range of failures
- Each server/storage unit still a single point of failure
  - must be taken into account when deploying critical services

Hardware (3)

- 5 separate islands (servers + storage):

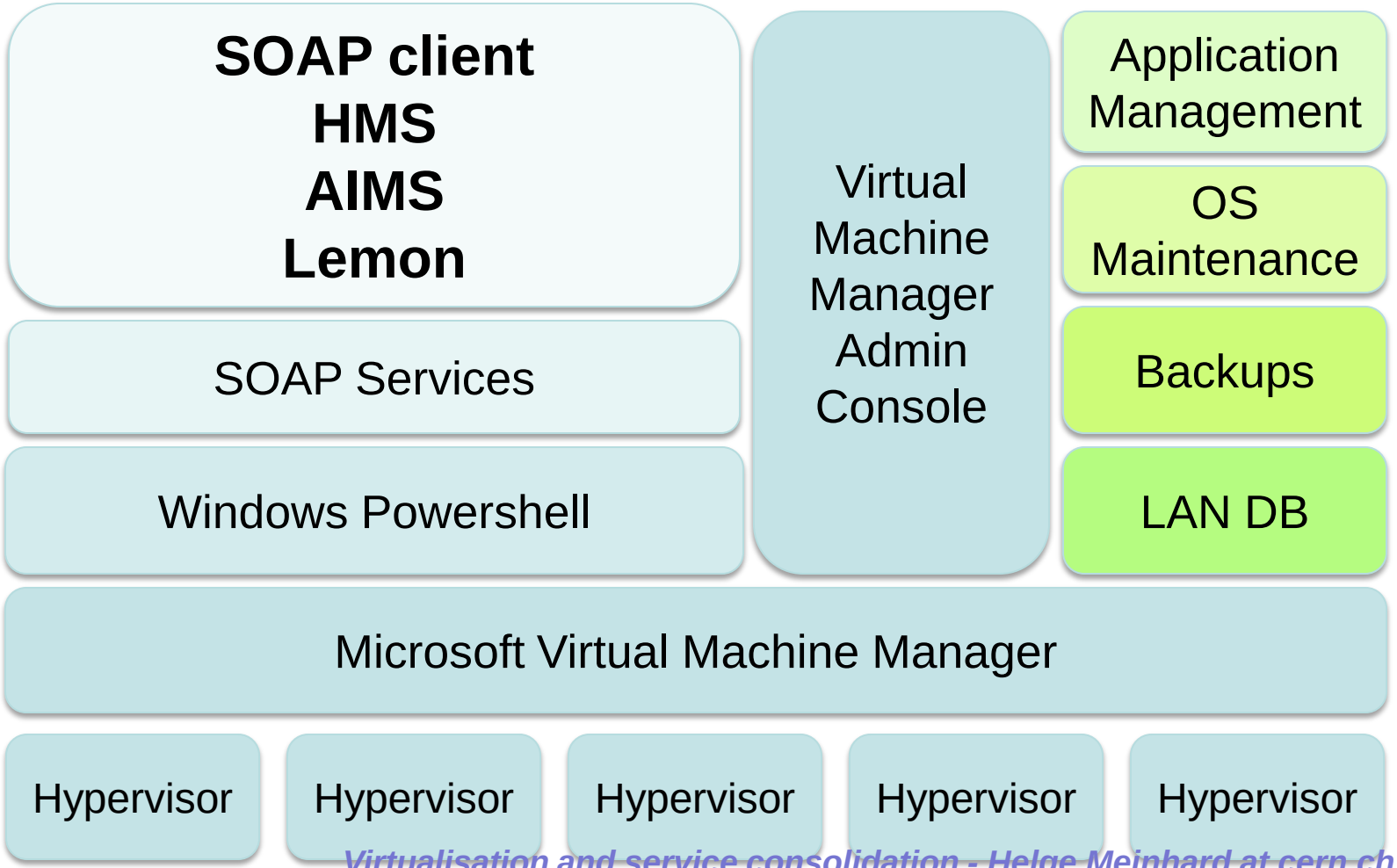
Servers		Netw	Storage		Location
16	8 cores, 24 GB	LCG	4	48 x 1 TB	Hosting centre
16	8 cores, 48 GB	LCG	4	48 x 1 TB	Hosting centre
16	4 cores, 12 GB	LCG	8	48 x 1 TB	513 Physics
16	4 cores, 12 GB	LCG	8	48 x 1 TB	513 Physics
16	4 cores, 12 GB	GPN	2	16 x 1 TB	513 critical

- To cope with growing demand for VMs, all servers will be upgraded to 8 cores and 48 GB before end of 2010

Software (1)

- Base hypervisor technology provided by IT-OIS, based upon
  - Hyper-V 2.0
  - System Centre Virtual Machine Manager 2008 R2 (SCVMM)
  - Custom SOAP interface
- Main features:
  - SCVMM is an enterprise class management suite
  - Live migration of VMs between hypervisors
  - High availability in the case of hypervisor failure
  - Integration with CERN network database via the SOAP interface
  - Full support for Linux guests with up to 4 cores
  - Paravirtualised I/O drivers integrated into SLC5

Software (2)


SOAP client
HMS
AIMS
Lemon
SOAP Services
Windows Powershell
Virtual Machine Manager Admin Console
Application Management
OS Maintenance
Backups
LAN DB
Microsoft Virtual Machine Manager
Hypervisor
Hypervisor
Hypervisor
Hypervisor
Hypervisor

Integrating Hyper-V and ELFMS


```
[lxadm05] /afs/cern.ch/user/d/dg > CDBDump wmsmon01

#######################################
#
# Hostname "wmsmon01"
#
# Cluster         "gridadm" - subcluster "undef"
# Function        "undefined"  (Comment: "")
# State           "maintenance", importance "50"
# Contracttype    "D"
# IT-section      "PES-PS"
# IT-contact      "ccservice.manageronduty@cern.ch"
# User-contact    "Ricardo.Silva@cern.ch"
# ITCM cc-email   ""
#
# Hardware        "vm_01_02_100" (microsoft)
# Serial Number   "000001"
# Warranty until  ""
#
# Operating System "slc5" - kernel "2.6.18-194.el5" - tpl "" - stage "prod" - osdate "20100503"
#    Architecture "x86_64" - svcclass "gridadm" - resource "undef" - customization "undef"
#
# Enclosed systems  "No enclosed object found"
# Location        "hyperv/"
#
#       information retrieved by CDBDump at Wed Jun 30 11:44:07 2010
#
#######################################
```

Project status

- Initially deployed services owned by IT-PES
- Then more IT services
- Started meanwhile to hand out VOboxes controlled by users outside IT
- Larger scale deployment awaiting server upgrades
- Currently about 130 machines, 4 machines per week, of which 100 in production
- Coming next: virtual small disk servers

Conclusions

- From the service manager (or VOC) perspective there is no difference between a real and a virtual machine
- Hardware requests still go via the usual channels but at the end you may be allocated virtual hardware
- Virtual machines make lifecycle management much easier by breaking the link between system image and hardware lifetime
- Virtualisation does not change the need for fabric management tools like Quattor
- Currently hosting more than 130 VMs. Growing rate: 4 VMs/month