LxCloud Infrastructure

status and lessons learned

Ulrich Schwickerath
Belmiro Moreira

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

What is LxCloud?

- Back 2009 CERN started to setup a virtualized infrastructure with the main aim to simplify the internal management of physical resources running Linux
- The current objective is to provide CERN services on top of this infrastructure

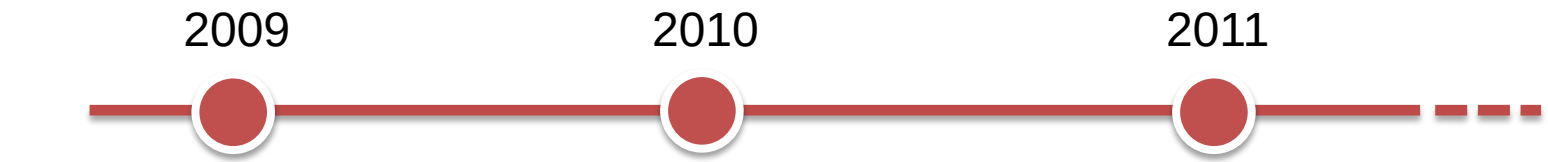
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

LxCloud – advantages...

- The usual cloud advantages:
  - Maintenance
    - Decoupling from the underlying hardware
    - Easy migration between OS / configurations
  - Dynamic Environment
    - The composition of compute nodes can be rapidly adapted to optimize resources utilization
  - Encapsulation
    - Virtualization allows the separation between the user environment and the hardware

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

LxCloud – brief history…



- XEN hypervisor;
- ONE;
- Only for Batch resources;

- XEN / KVM hypervisor;
- ONE / Platform ISF;
- IaaS concept;
- Scalability Test;
- VMIC;
- Image distribution based on Bittorrent;
- VirtualBatch with 96 VMs;

- KVM hypervisor;
- SLC6 migration;
- ONE / OpenStack;
- VMIC improvements;
- EC2 service evaluation;
- VirtualBatch with 432 VMs;

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

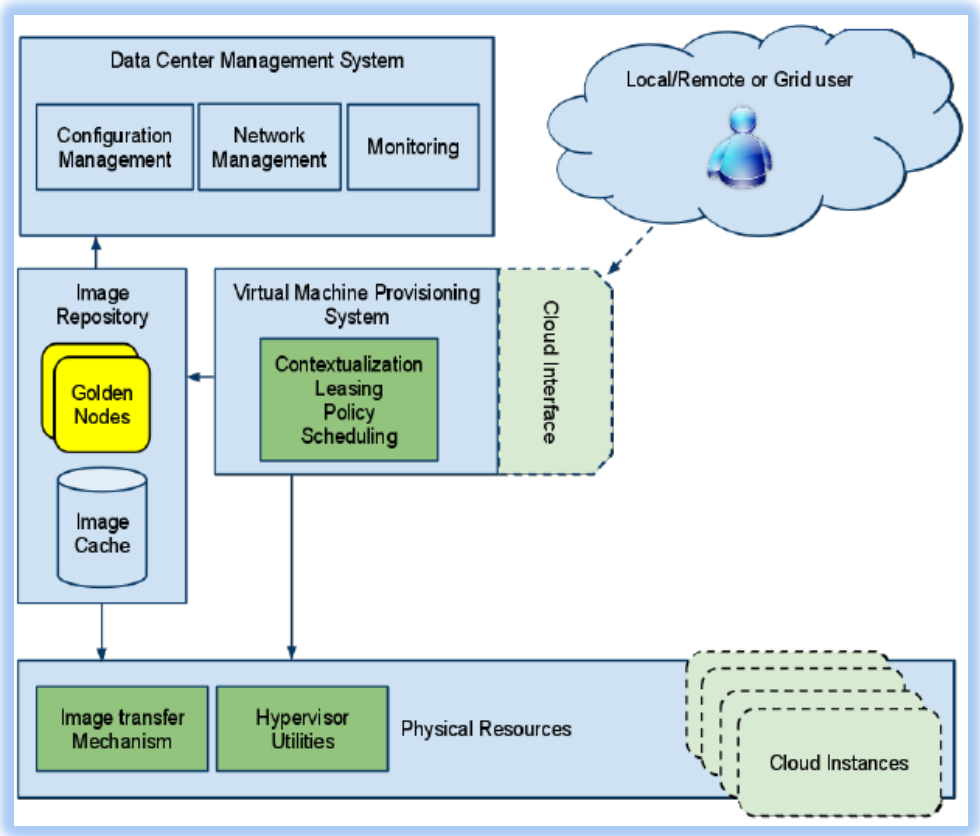
What is LxCloud?

- LxCloud is running on standard compute nodes used by the batch service
  - 2 x intel xeon L5520 @ 2.27 GHz (8 cores)
    - Nehalem architecture
  - 58 compute nodes
    - 48 nodes in production
    - 10 nodes for tests
  - 24 GB of RAM
  - 3 TB of local disk space

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

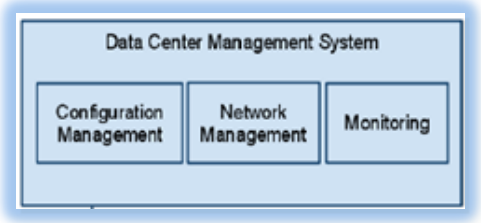
What is LxCloud?

- LxCloud architecture:



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

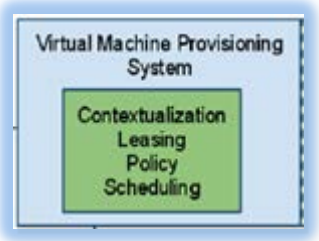
LxCloud – fabric managment tools



- Configuration Managment
  - LxCloud is integrated with the Fabric Management tools used by CERN
    - Quattor managed pool of resources
    - Alarming with LAS (Operator)
    - Hardware management by sys-admin team
    - “Draining” via sms state management
- Network Managment
  - Pre-allocation of VM “slots” in network DB
  - Compute node “knows” the available “name” of its guests
- Monitoring
  - All Compute nodes are monitored by Lemon

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

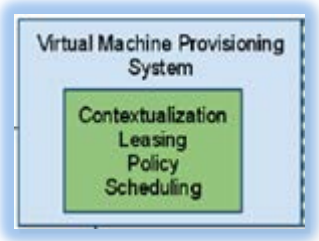
LxCloud – provision system



- Virtual Machine Provision System
  - OpenNebula
    - Migration from ONE 2.2 to ONE 3.0
    - Consolidation of different ONE instances in a single server
    - ONE server running on SLC6
  - ISF
    - Not using ISF
  - Evaluation of OpenStack

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

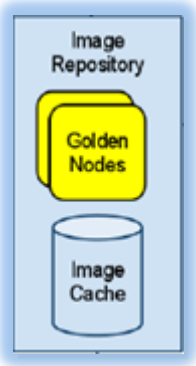
LxCloud – provision system



- Contextualization
  - ISO file with contextualization information is attached into the guest during the boot time
    - Contextualization scripts were defined inside the HEPiX virtualization WG
    - CernVM supports this contextualization model
- Scalability tests
  - In 2010 various scalability tests were made to evaluate the provision systems, LSF scheduler, and all LxCloud infrastructure
    - 16000 VMs running in about 500 compute nodes
    - LSF scalability concern

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

LxCloud – image repository



- Image Repository workflow
  - Golden Nodes
    - Images with the desired configuration are installed in a "Golden Node"
      - PXE installation / or not, using a slot on a compute node
    - This image is compressed and transfered to the "Image Cache"
      - If the image is quattor managed it needs to be "de-quattorized" and clean
  - Image Cache
    - It's managed by VMIC – Virtual Machine Image Catalogue

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

LxCloud – image management

- VMIC – virtual machine image catalogue
  - Using HEPiX virtualization WG specifications
    - Image sharing successful between Clemson University and CERN;
  - See “CloudMan and VMIC projects overview” presentation at HEPiX Fall 2011
- Image distribution management software
  - Responsible for managing images in each compute node
  - Images are pre-staged in the compute nodes;
  - BitTorrent image transfer
    - rtorrent client installed in all compute nodes

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

LxCloud – image format/storage

- Using RAW format image
  - Images are stored in LVMs
  - Using LVM snapshot functionality
- Evaluation of QEMU format image
  - Store images in files
  - Use “COW” functionality of QEMU image format
    - The link between the “main” image and the “snapshots” is hardcoded
      - If the “main” image is renamed the “snapshot” metadata is not updated

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

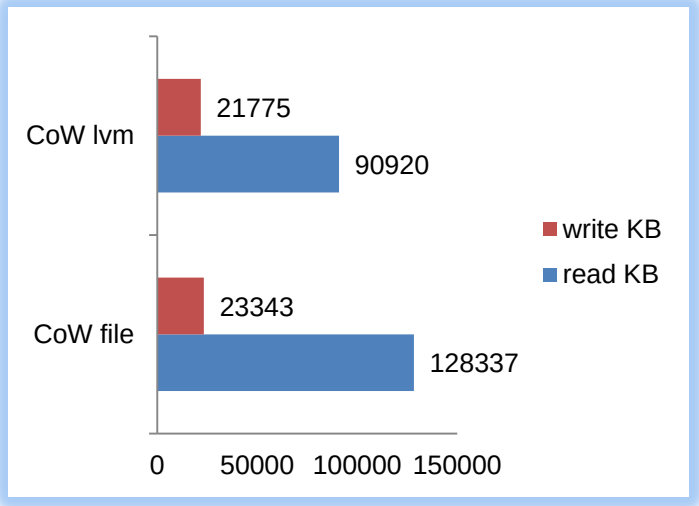
LxCloud – LVM/QEMU benchmark

- Methodology
  - `iozone -Mce -I -+r -r 256k -s 8g -f /external/iozone.dat -i0 -i1 -i2`
    - -i0=write/rewrite
    - -i1=read/re-read
    - -i2=random-read/write
  - All tests were executed in the same hardware/software
    - Scientific Linux CERN SLC release 6.1 beta (Carbon) – kernel 2.6.32
    - qemu-kvm-0.12.1.2 using virtIO drivers;
    - Guest – Scientific Linux CERN SLC release 5.6 (Boron) - 2.6.18
    - Processor: 2 x Intel xenon L5520 @ 2.27GHz
    - RAM: 24 GB
    - Disk: Hitachi HDE72101
  - 8 consecutive executions of each test
  - Average of all values

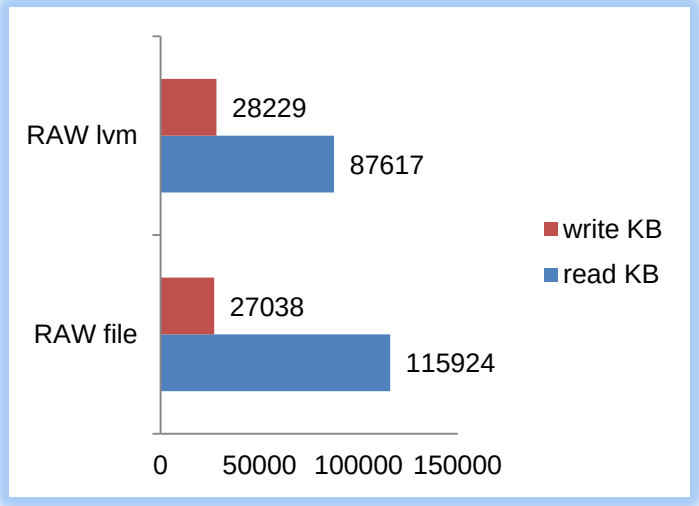
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

LxCloud – LVM/QEMU benchmark

1 virtual machine



A snapshot is created from the main image.
Snapshot IO performance evaluated – CoW.
(KB/sec)

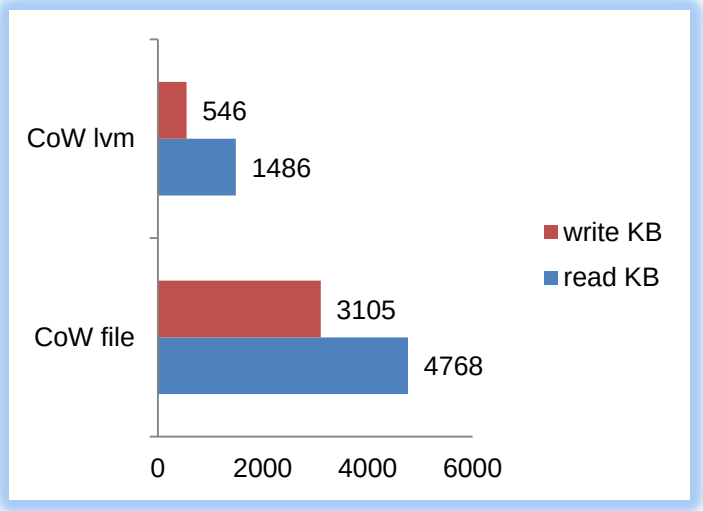


A snapshot is created from the main image.
An RAW disk (LVM/raw q-emu) with 10G is attached.
Attached RAW disk IO performance evaluated.
(KB/sec)

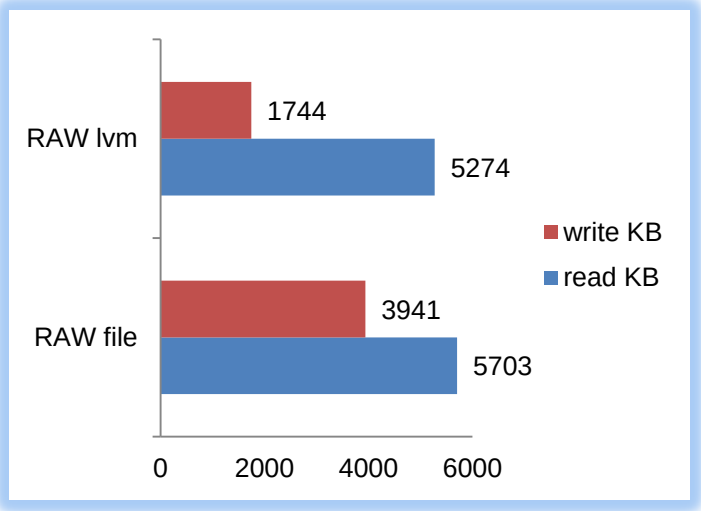
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

LxCloud – LVM/QEMU benchmark

8 virtual machines



Comparation between LVM and q-emu IO performance. Main image tested. (KB/sec)



Comparation between LVM and q-emu IO performance. Test volume externally attached. (KB/sec)

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

LxCloud – Physical Resources

- Compute nodes are running SLC6
  - Updated versions of libvirt and kvm
  - KSM enabled
  - Problem with disks in VM XML description
    - Expected fix in SLC6.2

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

LxCloud - services

- Services running on top LxCloud
  - virtualBatch
    - Production service since 2010
    - Gradually increasing capacity
      - 96 VMs (2010)
      - 432 VMs (2011);
  - EC2
    - First tests in 2011

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

LxCloud - virtualBatch

- virtualBatch characteristics
  - Instances are always derived from the newest available golden image
    - VM TTL is set to 48 hours
  - Customized at boot time (contextualization)
  - Instances are manageable by Quattor

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

LxCloud - virtualBatch

- Running batch jobs:
  - 48 compute nodes (384 cores)
    - 432 VMs (432 job slots)
    - 9 VMs per node
      - CPU pinning
  - 2.6 GB per VM
    - Memory overcommitted
    - Large swap but rarely used

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

LxCloud – EC2 service

- Access for restricted users - only on request
- Predefinied set of images
  - Users can’t upload their own images
- Using EC2 driver for OpenNebula
  - econe driver
  - Amazon EC2 API is not totally suported by ONE

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

Summary

- LxCloud proven flexible and scalable
- Seamless integration into the existing fabric management tools
- Open infrastructure to new tools – OpenStack;
- Running one production service
  - virtualBatch
    - better flexibility and maintenance operations when compared with “lxbatch”
  - New services in test

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

LxCloud - References

R. Wartel, T. Cass, B. Moreira, E. Roche, U. Schwickerath and S. Goasguen; “Image Distribution in Large Scale Cloud Providers”; 2nd IEEE Cloud Computing Conference; Indianapolis; 2010.

HEPiX meeting Spring and Fall 2009, 2010, 2011

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

CERN Virtual Infrastructure

Status update

CVI team
CERN – IT/OIS
Hepix 2011 Fall meeting

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

Reminder: what is CVI?

- The CERN Virtual Infrastructure custom virtual machines in the CERN computer center
  - 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
  - Enterprise class centralized management
  - Rich feature set:
    - Allows grouping of hypervisors, with delegation of administrative privileges
    - VM migration, High availability
    - Checkpointing
    - PowerShell Snap-In for administration / scripting

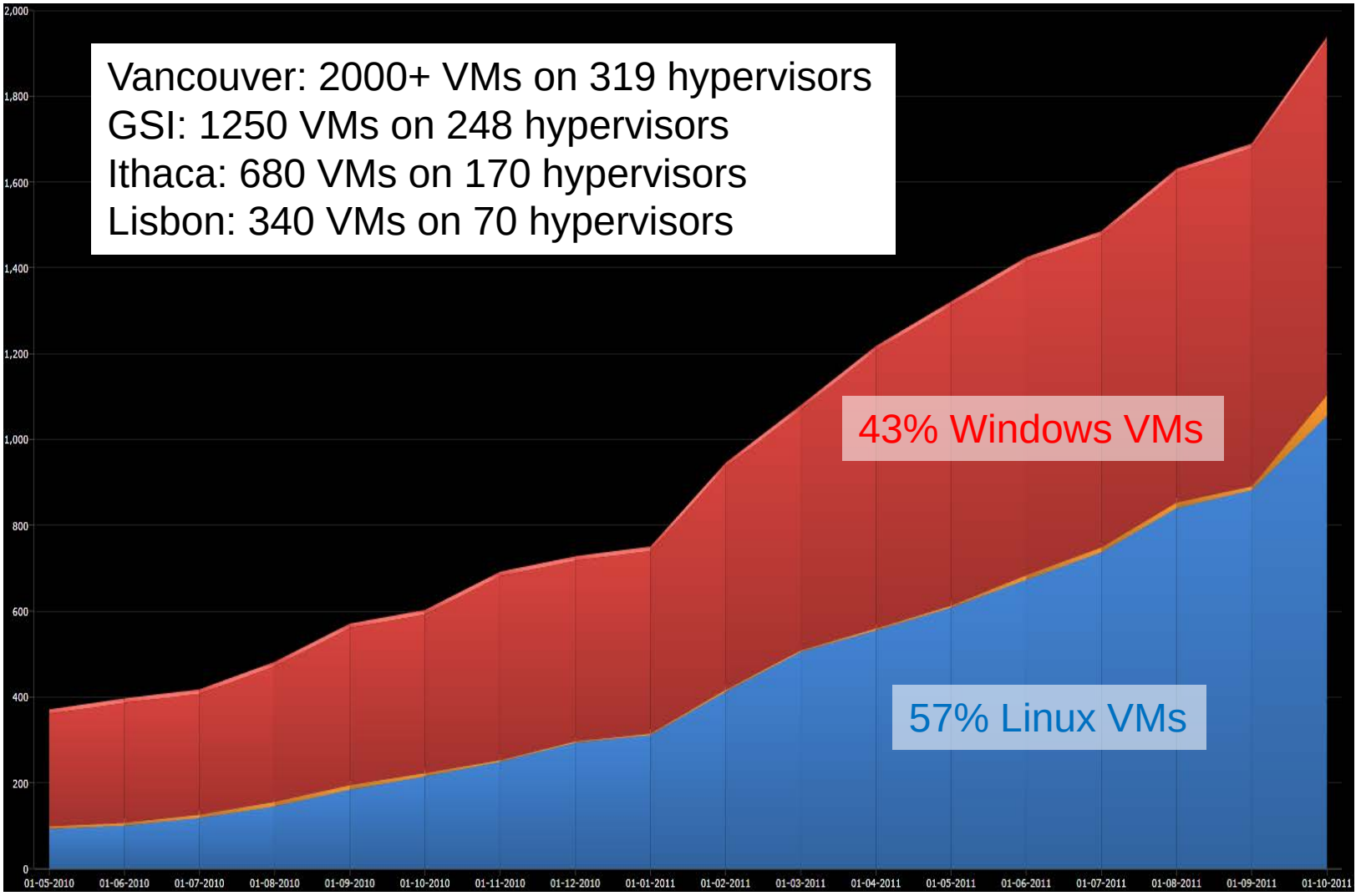
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Updates since Spring 2011

- CVI 2.0 running stably
- Integration Components for RHEL6/SLC6
  - Released by Microsoft in July
  - Working fine
  - Much improved packaging wrt RHEL5
    - Provided as RPMs
    - Source code is snapshot of upstream kernel drivers
      - Still in staging area ☹
- Service grows by ~100 VMs per month
  - New customers: IPv6 testing,3D rendering farm

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

Spectacular growth continues!



Number of Virtual Machines per Operating System

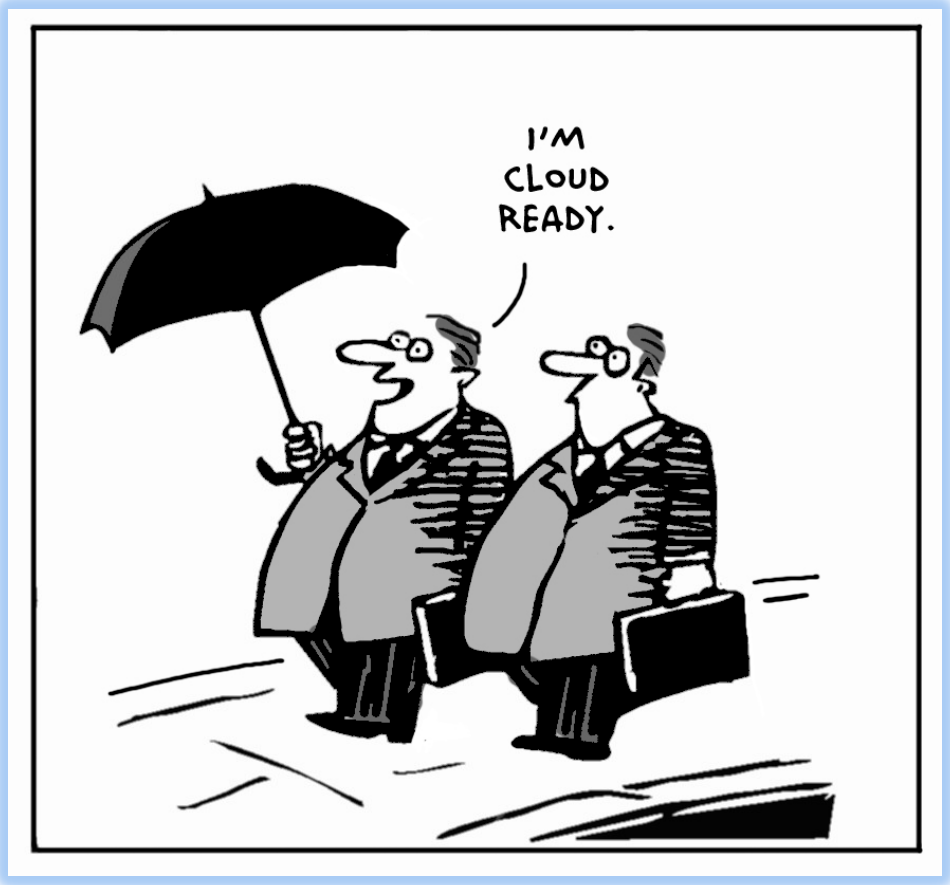
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Outlook for 2012

- Upgrade VMM 2012
  - CERN participated in VMM 2012 TAP program
- Progressively deploy new features
  - Customize properties
    - VM expiration date, backup policy, …
  - Introduce VMM private Clouds for different user communities
  - Provide enhanced user interface
- Integrate with Operation Manager

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

Thank you...



www.cloudtweaks.com – David Fletcher

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