The KVM infrastructure at the INFN Tier-1

Andrea Chierici, Guido Guizzunti,
Felice Rosso, Riccardo Veraldi

INFN-CNAF

Workshop CCR 2010

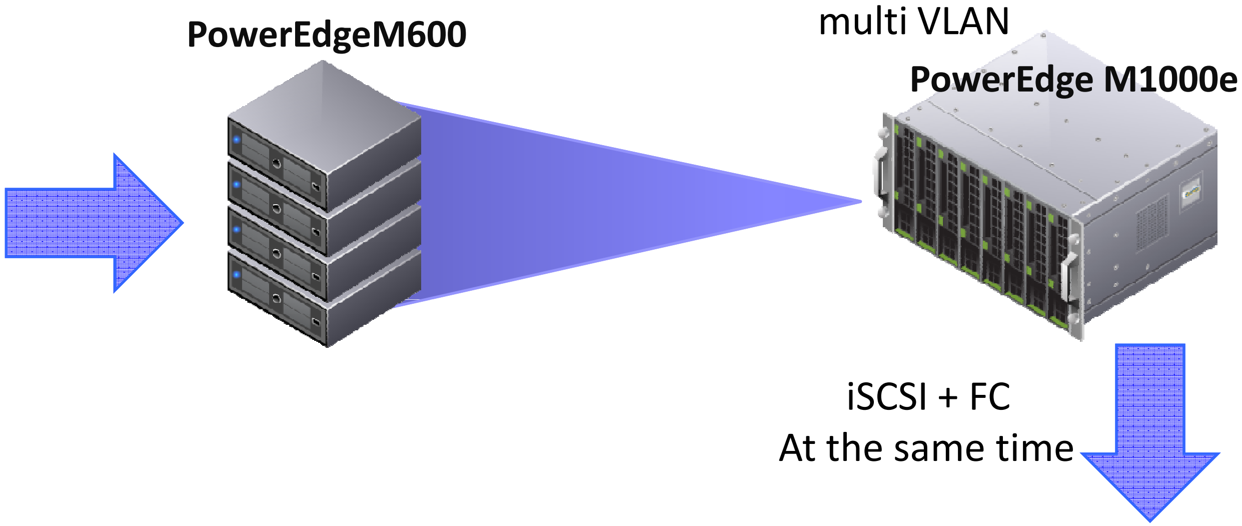
Outline

- State of the art
  - □ virtual services for CDF experiment
  - □ CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - □ migration from xen to kvm
  - □ virtio on sl4, sl5
  - □ libguestfs
- Developments
  - □ ksm
  - □ hugetlbfs

Outline

- State of the art
  - virtual services for CDF experiment
  - CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - migration from xen to kvm
  - virtio on sl4, sl5
  - libguestfs
- Developments
  - ksm
  - hugetlbfs

Virtual services for CDF experiment

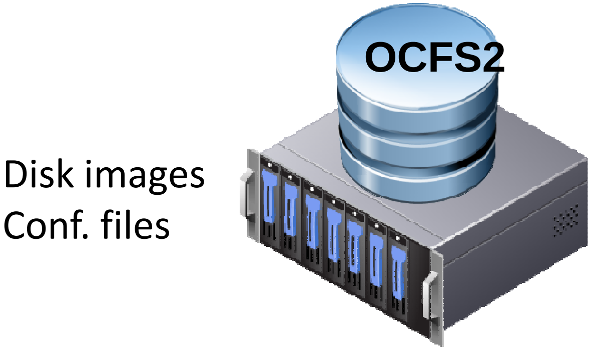


PowerEdgeM600

multi VLAN

PowerEdge M1000e

iSCSI + FC
At the same time

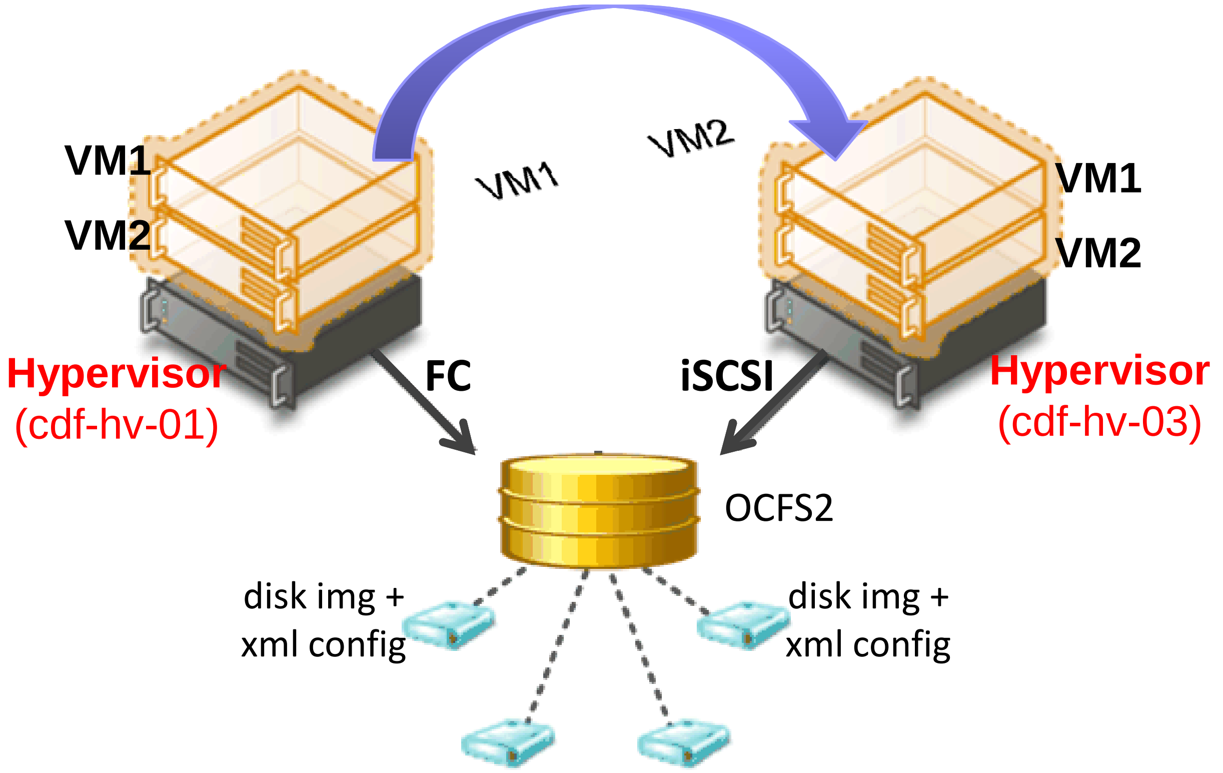


OCFS2

Disk images
Conf. files

- VM installed and maintained via Quattor
- Often part of the GPFS cluster

Live KVM migration with virsh



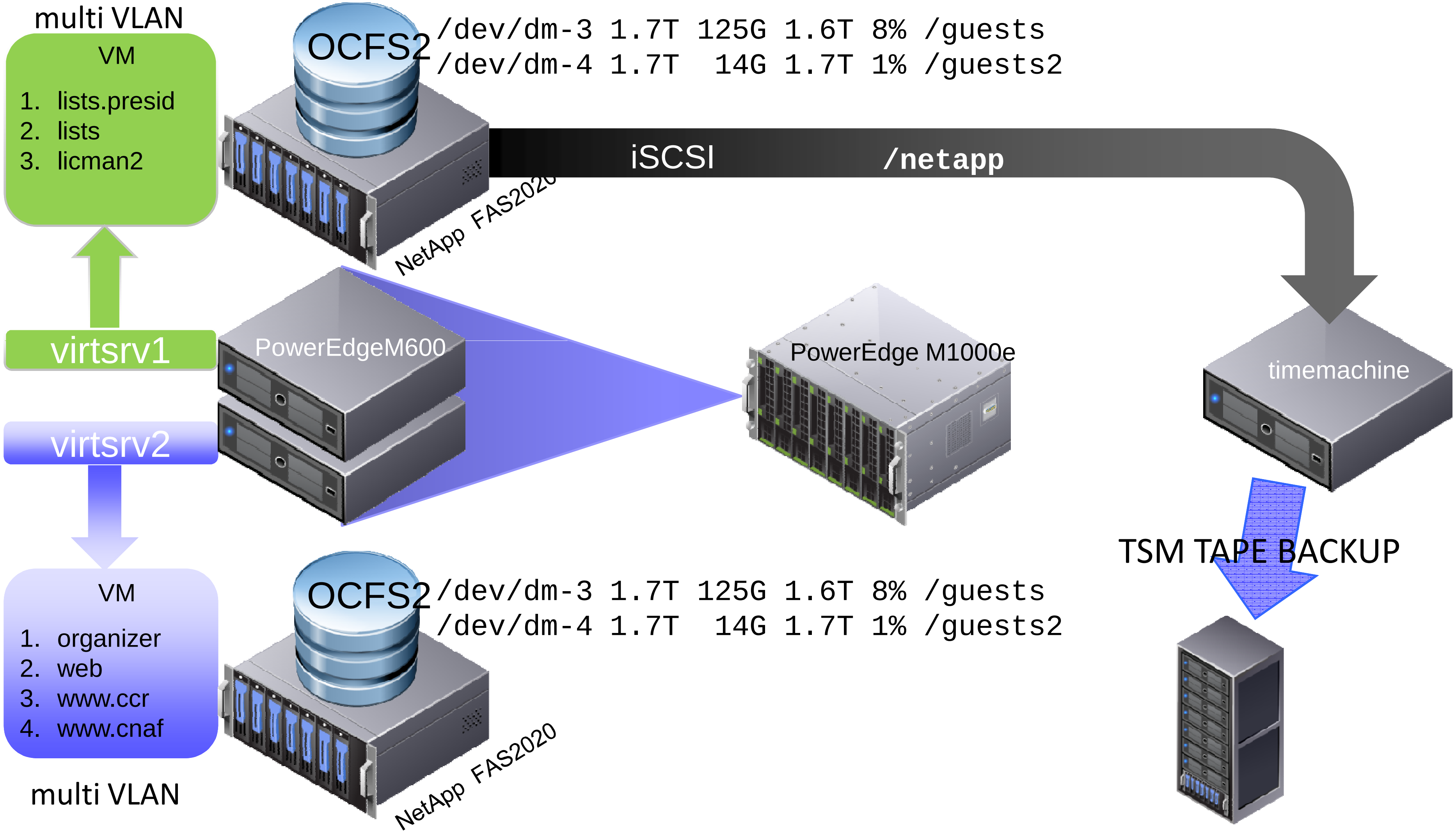
virsh migrate --live GuestName DestinationURL

- Load balancing
- Hardware failover
- Software upgrade

Outline

- State of the art
  - virtual services for CDF experiment
  - CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - migration from xen to kvm
  - virtio on sl4, sl5
  - libguestfs
- Developments
  - ksm
  - hugetlbfs

CNAF & INFN National Services on KVM



multi VLAN

VM

1. lists.presid
2. lists
3. licman2

OCFS2

```
/dev/dm-3 1.7T 125G 1.6T 8% /guests
/dev/dm-4 1.7T  14G 1.7T 1% /guests2
```

iSCSI **/netapp**

NetApp FAS2020

virtsrv1

PowerEdgeM600

virtsrv2

PowerEdge M1000e

timemachine

TSM TAPE BACKUP

VM

1. organizer
2. web
3. www.ccr
4. www.cnaf

OCFS2

```
/dev/dm-3 1.7T 125G 1.6T 8% /guests
/dev/dm-4 1.7T  14G 1.7T 1% /guests2
```

NetApp FAS2020

multi VLAN

Implementation

- 2x Server Dell PowerEdge M600
- 2x Netapp FAS2020 Head
  - 2.11 TB volume, 10% snapshot reserved
    - 1.8 TB effective volume space
      - 1.6 TB ocfs2 partition
  - Production LUN exported via FC
  - Snapshot LUN created on the fly exported via iSCSI
- 2x CentOS 5.4 (kvm enabled)
  - Using ocfs2 cluster FS as VM system storage
    - Support for multiple VM VLANs
  - Backup using NetApp snapshot feature and iSCSI LUN Export toward a backup server which mounts the snapshot partition and sends data to tape servers (TSM)
    - Snapshot on the fly using custom scripts (VM Sync, LUN Snapshot)

Advantages & Disadvantages

- KVM + NetApp storage is
  - □ Reliable
  - □ Robust
  - □ Opensource (KVM)

- KVM + NetApp storage is
  - □ A little tricky to manage for snapshots
    - Requires customized scripts to sync all VMs and create the snapshot
    - Snapshots have to be managed manually
  - □ VM must be moved manually around hypervisors
    - No VM load balancing

Outline

- State of the art
  - □ virtual services for CDF experiment
  - □ CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - □ migration from xen to kvm
  - □ virtio on sl4, sl5
  - □ libguestfs
- Developments
  - □ ksm
  - □ hugetlbfs

Xen->kvm migration (1)

- CNAF migrated existing VMs from xen to kvm without any reinstallation
  - Xen phased out
  - Existing hosts rely on sl5.4 kvm distribution
  - Stable and "fast enough"
  - No more clock sync problems
    - Kernel options: notsc divider=10

Xen->kvm migration (2)

- Host is vanilla sl5.4
- Guest can be sl(c)4 or sl5
- We used disk-on-a-file but a partition should work too
- Procedure documented on INFN wiki
  - □ http://wiki.infn.it/cn/ccr/virtualizzazione/documentazione/xen_to_kvm
  - □ See Andrea Chierici’s poster
  - □ Basically only a small customization of the VM is required

Outline

- State of the art
  - □ virtual services for CDF experiment
  - □ CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - □ migration from xen to kvm
  - □ virtio on sl4, sl5
  - □ libguestfs
- Developments
  - □ ksm
  - □ hugetlbfs

Virtio

- Main platform for IO virtualization in KVM
- To use virtio drivers on guests:
  - □ sl4.x: kernel >= 2.6.9-89.0.3.EL
  - □ sl5.x: kernel >= 2.6.18-164.6.1.el5
- If you want to install a machine with virtio drivers add these lines to virt-install:

```
        --os-type=linux \
        --os-variant=virtio26 \
```

- Very stable, but performances are only fair
- It's possible to migrate from standard to virtio machine without re-installation, with custom initrd

Outline

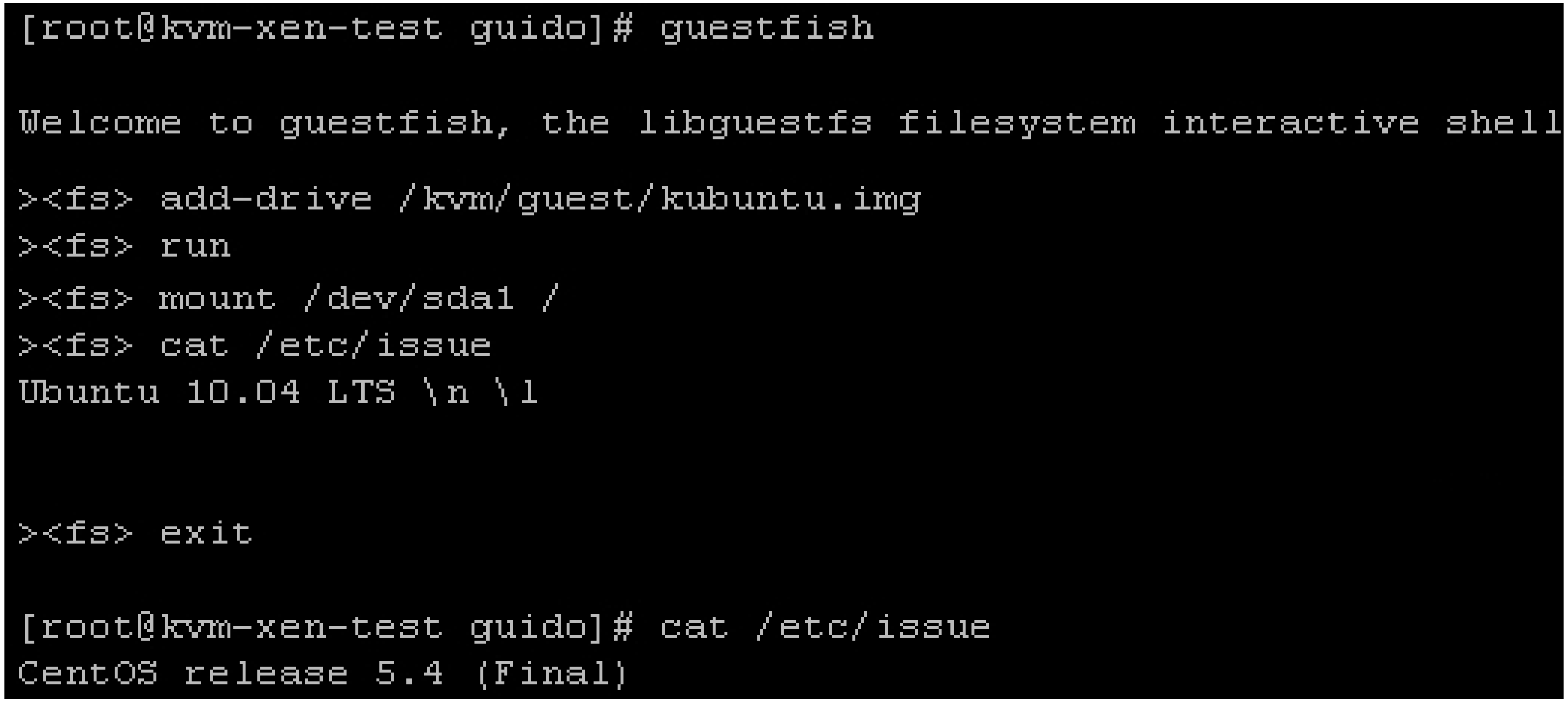
- State of the art
  - virtual services for CDF experiment
  - CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - migration from xen to kvm
  - virtio on sl4, sl5
  - libguestfs
- Developments
  - ksm
  - hugetlbfs

What is libguestfs?

- An API for creating, accessing, manipulating and modifying filesystems and disk images.
- Gives access from many different programming languages, or the command line.
- A set of useful tools and applications
  - guestfish, virt-cat, virt-inspector, virt-df, virt-resize

guestfish

- guestfish is the “guest filesystem interactive shell”
- you can just run it on any disk image you happen to find.
- You don’t need to be root



```
[root@kvm-xen-test guido]# guestfish

Welcome to guestfish, the libguestfs filesystem interactive shell

><fs> add-drive /kvm/guest/kubuntu.img
><fs> run
><fs> mount /dev/sda1 /
><fs> cat /etc/issue
Ubuntu 10.04 LTS \n \l


><fs> exit

[root@kvm-xen-test guido]# cat /etc/issue
CentOS release 5.4 (Final)
```

Binding for Python and other languages

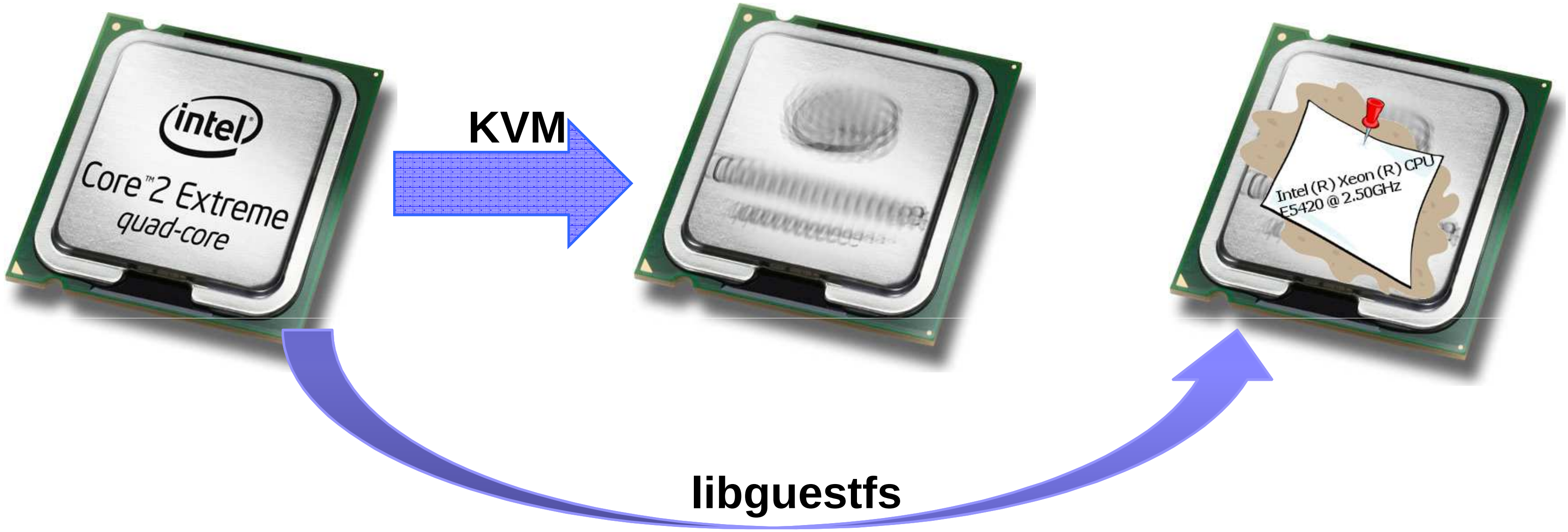
- Language bindings for many common programming languages (Perl, OCaml, C, C++ and shell script)
- Example:

```
#!/usr/bin/python

import guestfs
g = guestfs.GuestFS ()
g.add_drive_ro (“/kvm/guest/kubuntu.img”)
g.lunch ()

parts = g.list_partitions ()
print “disk partitions: %s” % (“, “.join (parts))
```

Usage of libguestfs within WNoD



- making batch configuration changes to guests
- viewing and editing files inside guests

Outline

- State of the art
  - virtual services for CDF experiment
  - CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - migration from xen to kvm
  - virtio on sl4, sl5
  - libguestfs
- Developments
  - ksm
  - hugetlbfs

KSM (1)

- Kernel Samepage Merging
- New feature allowing to share “common memory pages” between VMs
  - Still a work in progress under SL, working well on fedora 12
- We made some preliminary tests that showed good performances and stable functionality
- Linux services: **ksm, ksmtuned**

KSM (2)

- In-kernel values related to ksm under /sys/kernel/mm/ksm
  - full_scans max_kernel_pages pages_shared pages_sharing pages_to_scan pages_unshared pages_volatile run sleep_millisecs

KSM (3)

- kvm machine not running:
  - □ Full_scans: 0
  - □ Max_kernel_pages: 2058369
  - □ Pages_shared: 0
  - □ Pages_sharing: 0
  - □ Pages_to_scan: 100
  - □ Pages_unshared: 0
  - □ Pages volatile: 49000
  - □ Run: 1
  - □ Sleep_millisecs: 20

- 3 kvm machines 8GB:
  - □ Full_scans: 47
  - □ Max_kernel_pages: 2058369
  - □ Pages_shared: 69186
  - □ Pages_sharing: 186555
  - □ Pages_to_scan: 64
  - □ Pages_unshared: 46593
  - □ Pages volatile: 948
  - □ Run: 1
  - □ Sleep_millisecs: 10

Outline

- State of the art
  - virtual services for CDF experiment
  - CNAF & INFN national services
    - back-up solutions / snapshot (experiences with netapp)
  - migration from xen to kvm
  - virtio on sl4, sl5
  - libguestfs
- Developments
  - ksm
  - hugetlbfs

Hugetlbfs (1)

- Huge Translation Lookaside Buffer FS
  - small cache used for storing virtual-to-physical mapping information
  - to keep translations as fast as possible, the TLB is usually small
  - It is not uncommon for large memory applications to exceed the mapping capacity of the TLB
- Backing a KVM host with hugepages can give your guest machine a performance boost anywhere up to 10%

Hugetlbfs (2)

- How to check if your kernel supports hugepages:

```
$ grep -i huge /proc/meminfo
  HugePages_Total: 0
  HugePages_Free: 0
  HugePages_Rsvd: 0
  Hugepagesize: 2048 kB
```

- Enable hugetlbfs on VMs:
  - mount –t hugetlbfs hugetlbfs /dev/hugepages
  - Command line: append `–mem-path /hugepages`
  - Via libvirt: add these lines to xml:

    ```
    <memoryBacking> <hugepages/> </memoryBacking>
    ```

References

- http://www.linux-kvm.com/content/using-ksm-kernel-samepage-merging-kvm
- http://fedoraproject.org/wiki/Features/KVM_Huge_Page_Backed_Memory

Any questions?

Thanks!

E-mail: guido.guizzunti@cnaf.infn.it