Virtualizzazione da infarto

cluster di virtualizzazione in alta affidabilità

Contenuti

Il contesto

Cluster di alta affidabilità

Un cluster di virtualizzazione

Altre soluzioni

Domande?

...un po' di contesto #1

Centro Servizi Informatici della Facoltà di Ingegneria

Gli utenti
~10000 studenti (~1600 immatricolati nel 2010)
~900 tra docenti e personale tecnico amministrativo

La rete
30 apparati di rete tra switch e router
12 subnet per un totale di 16384 indirizzi
VLAN - Network Access – Routing IPV4 e IPV6 – Monitoring

I servizi
Radius – LDAP – DNS – Posta elettronica – Web
Gestione corsi (moodle) – VPN - Videoconferenza – SVN/GIT – etc.
Gestione aule informatiche (200 pc + software)

...un po' di contesto #2

Centro Servizi Informatici della Facoltà di Ingegneria

Pre virtualizzazione (2009)
- decine di servizi = decine di server
- hardware disomogeneo
- *N* versioni di Linux e FreeBSD

Post virtualizzazione (2010)
- 1 cluster di virtualizzazione composto da 3 server
- 2 template di distribuzioni linux
- 1 server di backup

Perché virtualizzare

Consolidamento server

Disaccopiamento hardware

1 server = 1 servizio

Flessibilità

Accentramento gestione

Perché NON virtualizzare

Hardware dedicato
- **dipende dall'hardware**

Sistemi operativi non supportati
- **dipende dalla tecnologia di virtualizzazione**

Prestazioni
- **I/O prima di tutto – CPU ~95-98%**

Aumento della complessità
- **tradeoff con flessibilità e funzionalità**

Single point of failure * N
- **alta affidabilità**

Alta affidabilità e bilanciamento

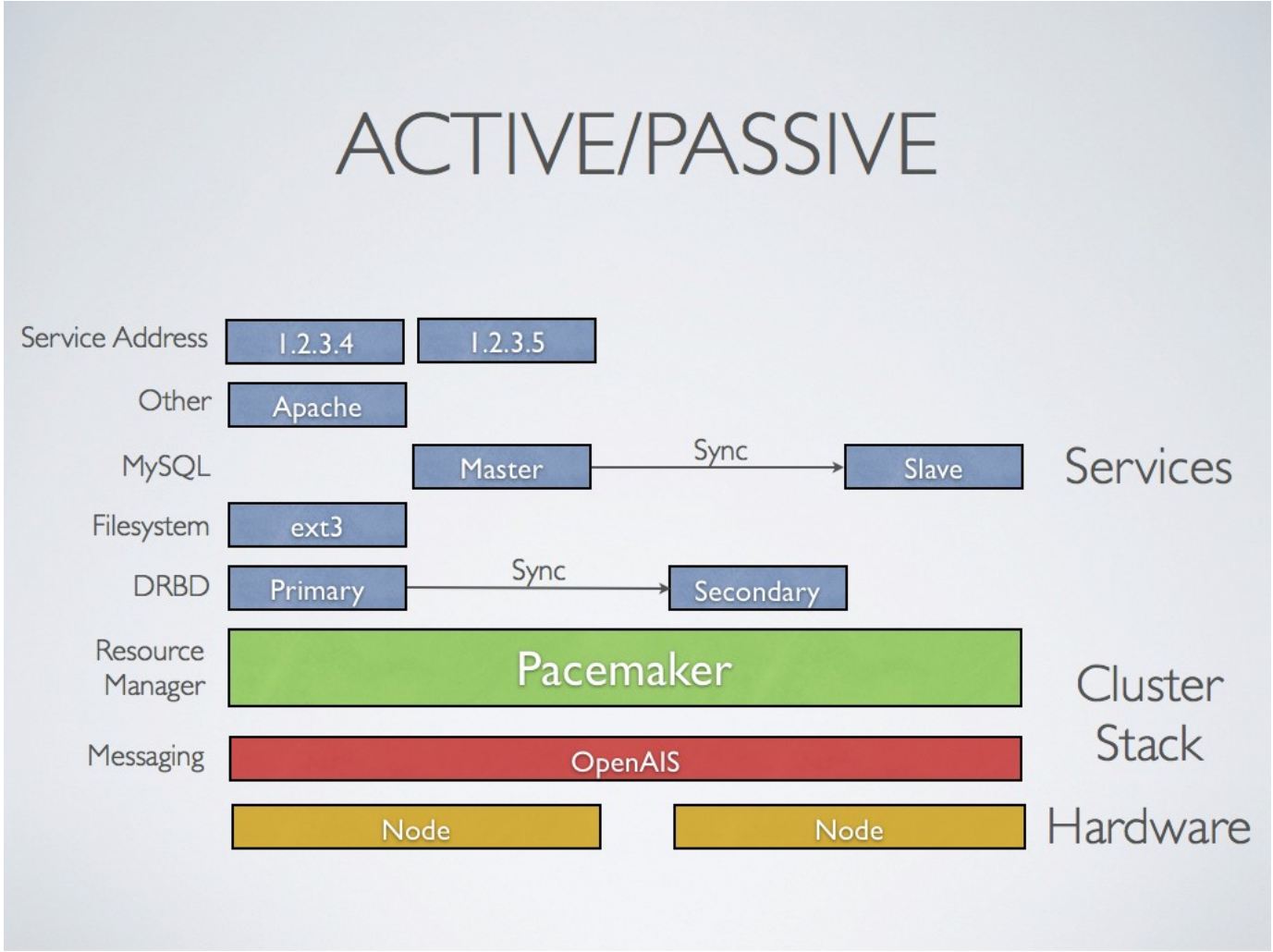
- **Alta affidabilità**

  Servizi e server in coppie (o N+1)
  active/passive o *active/standby*

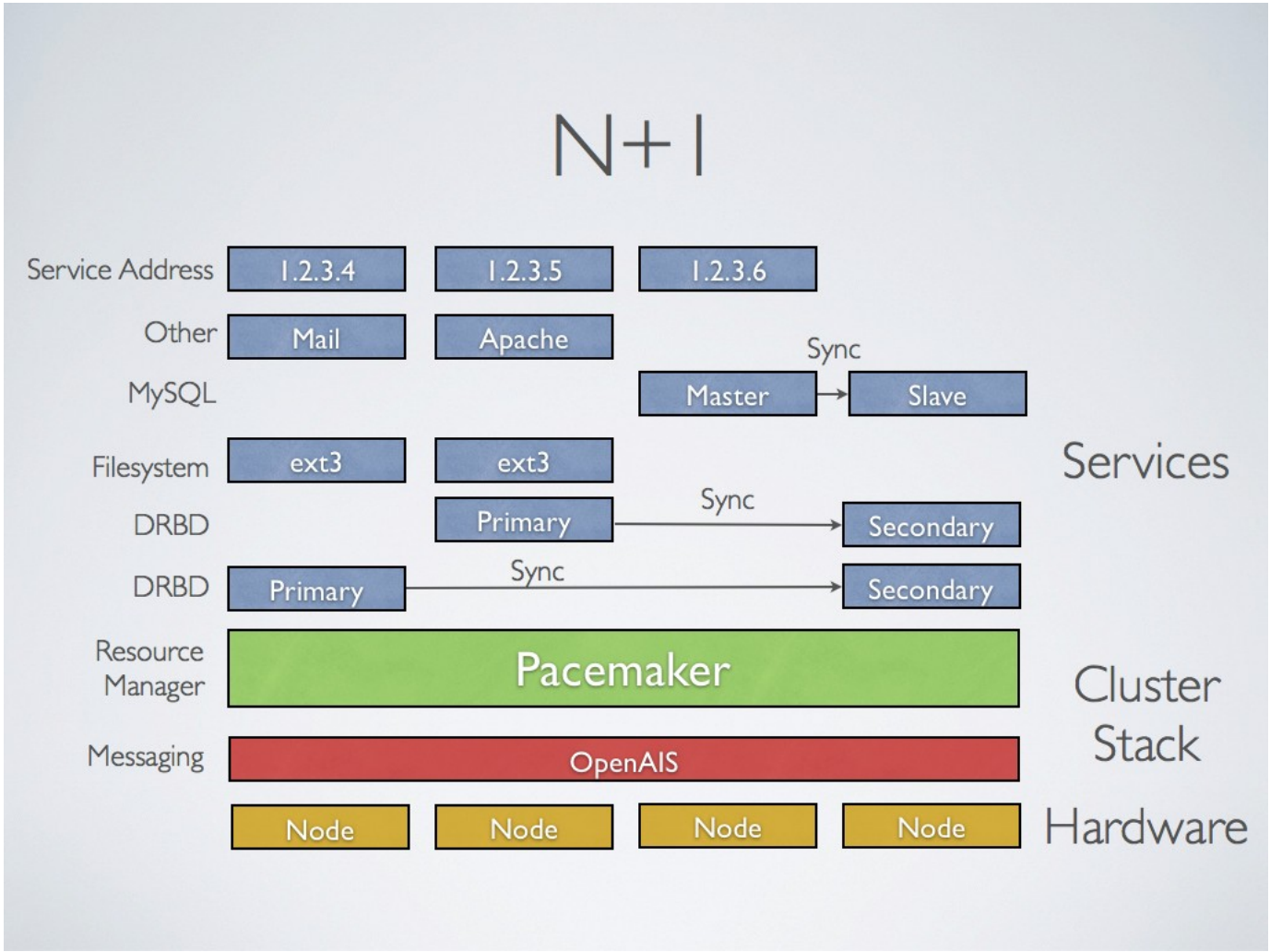
- **Bilanciamento**

  Servizi e server in *pool* – richieste raccolte tramite VIP e indirizzate ai membri del *pool* con algoritmi diversi

Pacemaker: cluster in alta affidabilità



Pacemaker: cluster in alta affidabilità



HA Cluster lingo

- Fail-over

  rilocazione delle risorse in caso di malfunzionamento

- Fail-back

  una volta riparato il malfunzionamento le risorse possono tornare ad essere ospitate sul nodo

- Simmetria

  determina se le risorse possono essere ospitate su ogni nodo del cluster

HA Cluster lingo

- Stickiness

  proprietà che determina la migrazione automatica di un servizio (ad es. fail-back)

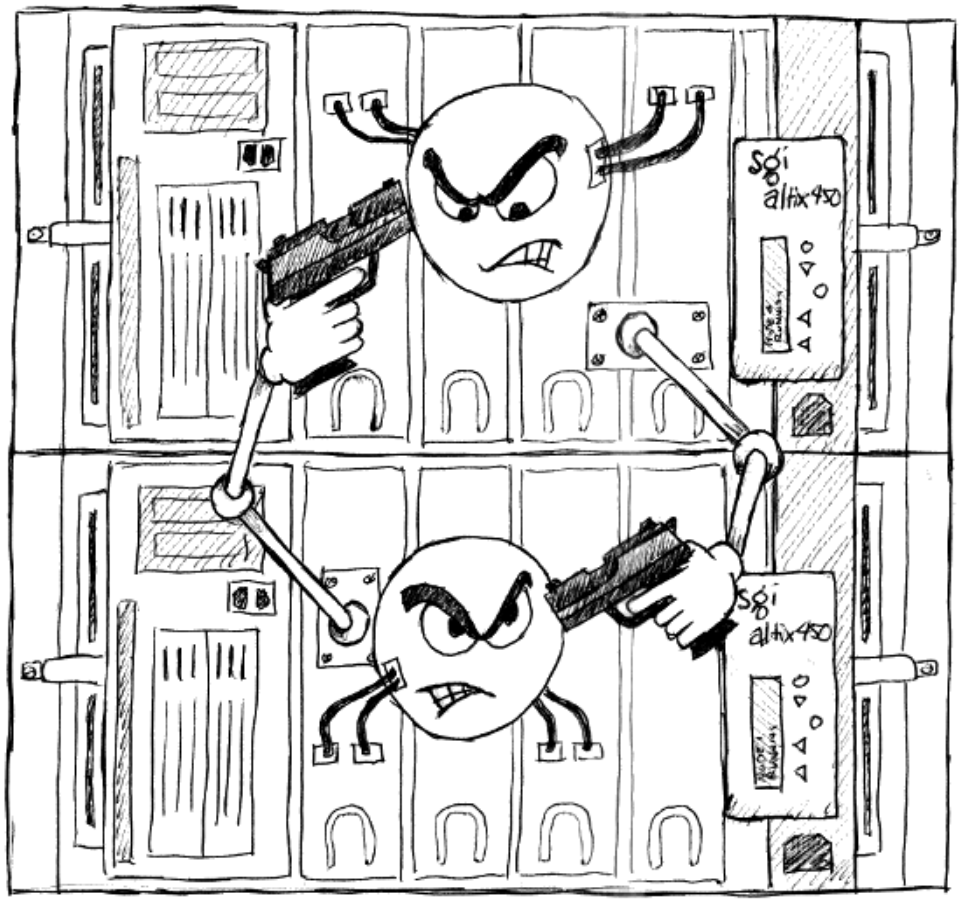
- Split-brain

  situazione in cui 2 o più nodi del cluster ritengono di essere gli unici attivi

- STONITH

  Shoot The Other Node In The Head – strumento per prevenire lo split-brain

STONITH DEATHMATCH

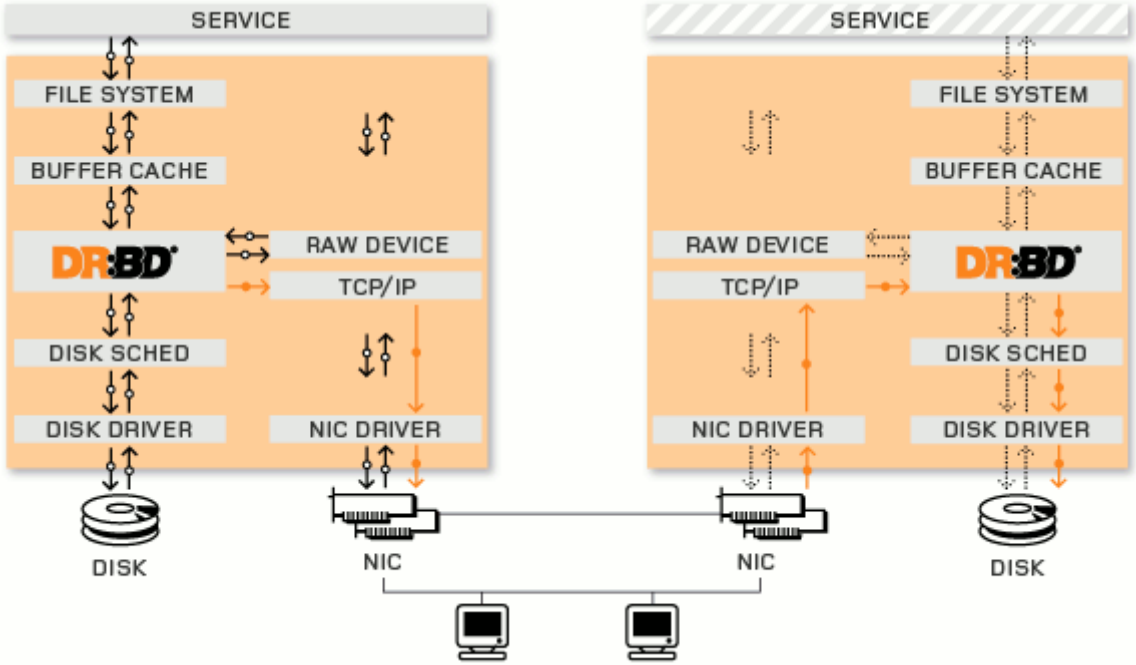


DON'T ANYBODY MOVE...

Cluster di virtualizzazione

- **DRBD** *duplicazione dati*
- **Xen** *virtualizzazione*
- **Pacemaker** *gestione cluster*

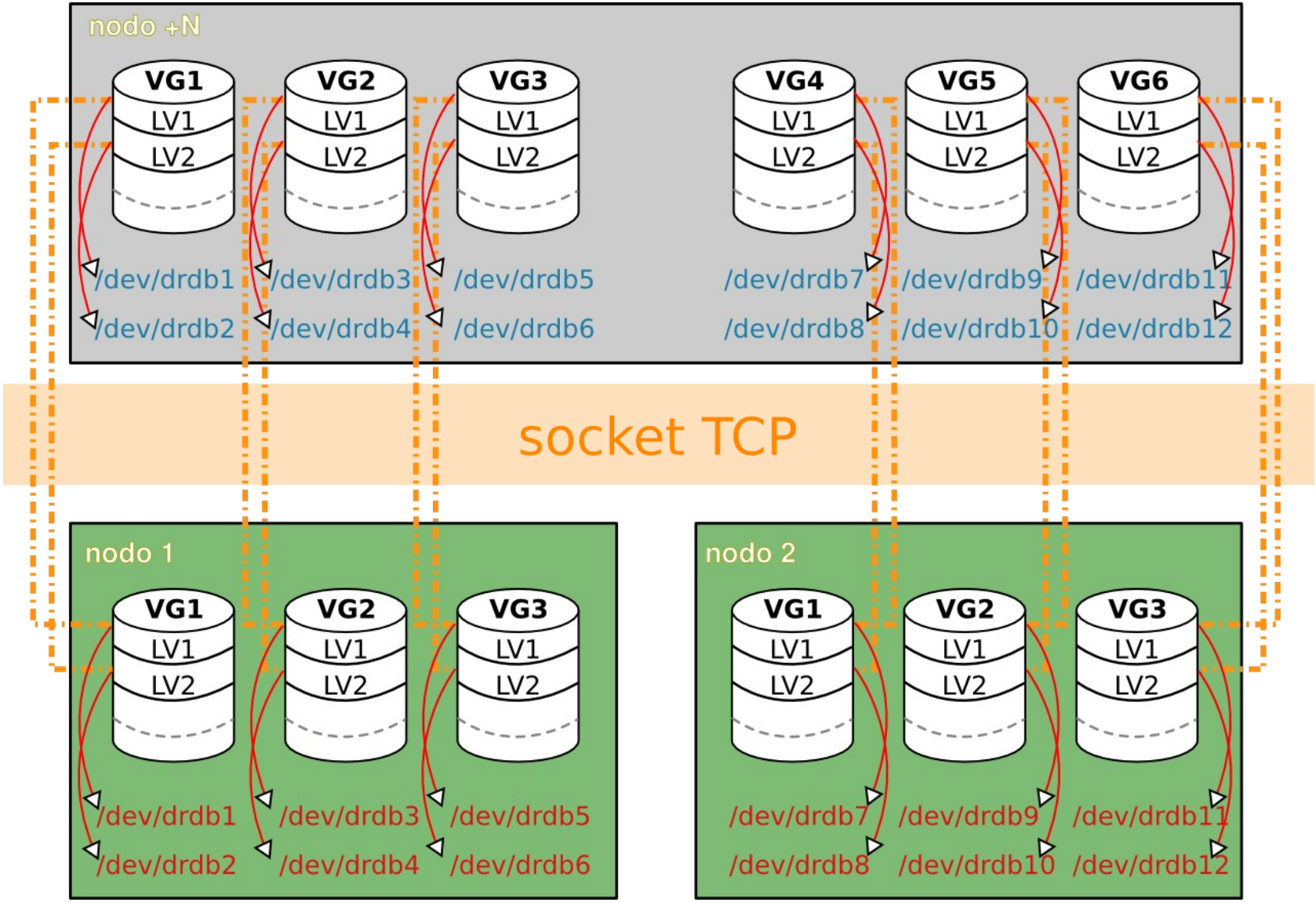
DRBD (Distributed Replicated Block Device)



http://www.drbd.org/ - Copyright © © 2008: LINBIT HA-Solutions GmbH

- Standard Raw Block Device
- Kernel mainline 2.6.33
- Network RAID-1
- Cluster aware
- Xen ready
- Compatibile con LVM
- Semplice da configurare

DRBD e LVM


nodo +N
VG1
LV1
LV2
VG2
LV1
LV2
VG3
LV1
LV2
VG4
LV1
LV2
VG5
LV1
LV2
VG6
LV1
LV2
/dev/drdb1
/dev/drdb2
/dev/drdb3
/dev/drdb4
/dev/drdb5
/dev/drdb6
/dev/drdb7
/dev/drdb8
/dev/drdb9
/dev/drdb10
/dev/drdb11
/dev/drdb12
socket TCP
nodo 1
VG1
LV1
LV2
VG2
LV1
LV2
VG3
LV1
LV2
/dev/drdb1
/dev/drdb2
/dev/drdb3
/dev/drdb4
/dev/drdb5
/dev/drdb6
nodo 2
VG1
LV1
LV2
VG2
LV1
LV2
VG3
LV1
LV2
/dev/drdb7
/dev/drdb8
/dev/drdb9
/dev/drdb10
/dev/drdb11
/dev/drdb12

DRBD: /etc/drbd.conf

```
common {
  syncer { rate 40M; }
}

# VM web
resource web-root {
  protocol C;
  on nodo1 {
        device  /dev/drbd1;
        disk    /dev/vg2/web-root;
        address 192.168.0.1:7789;
        meta-disk       internal;
        }
  on nodo2 {
        device  /dev/drbd1;
        disk    /dev/vg2/web-root;
        address 192.168.0.2:7789;
        meta-disk       internal;
        }
}
[...]
```

Banda per la sincronizzazione in background

Affidabilità sincronizzazione

- C – write ok quando altro write ok
- B – write ok quando pacchetto arrivato
- A – write ok quando pacchetto nel buffer tcp

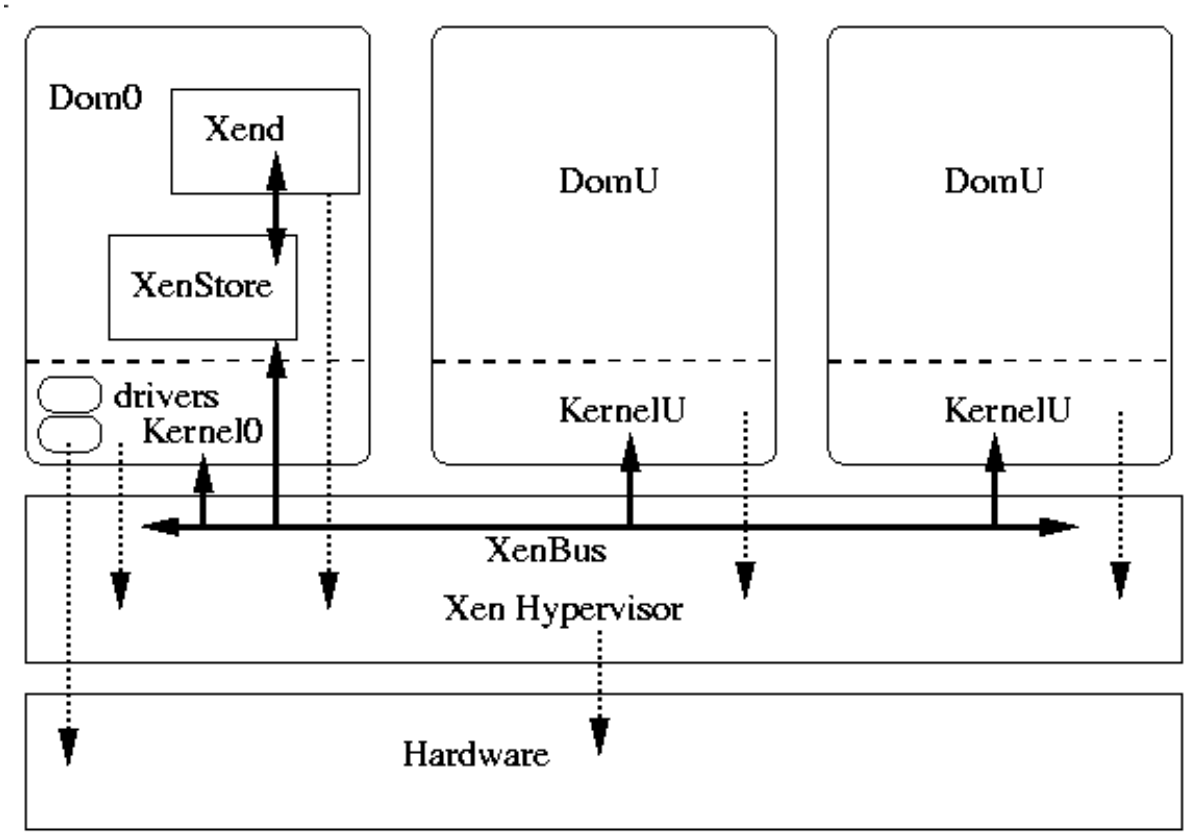
Storage dei metadati (grandezza, identificativo, activity log, quick-sync bitmap)

- Internal (preferito): stesso disco
- External: altro disco

Xen

Linux Kernel support:

- 2.6.23 domU
- 2.6.37/3.0 dom0/domU



Xen: xm

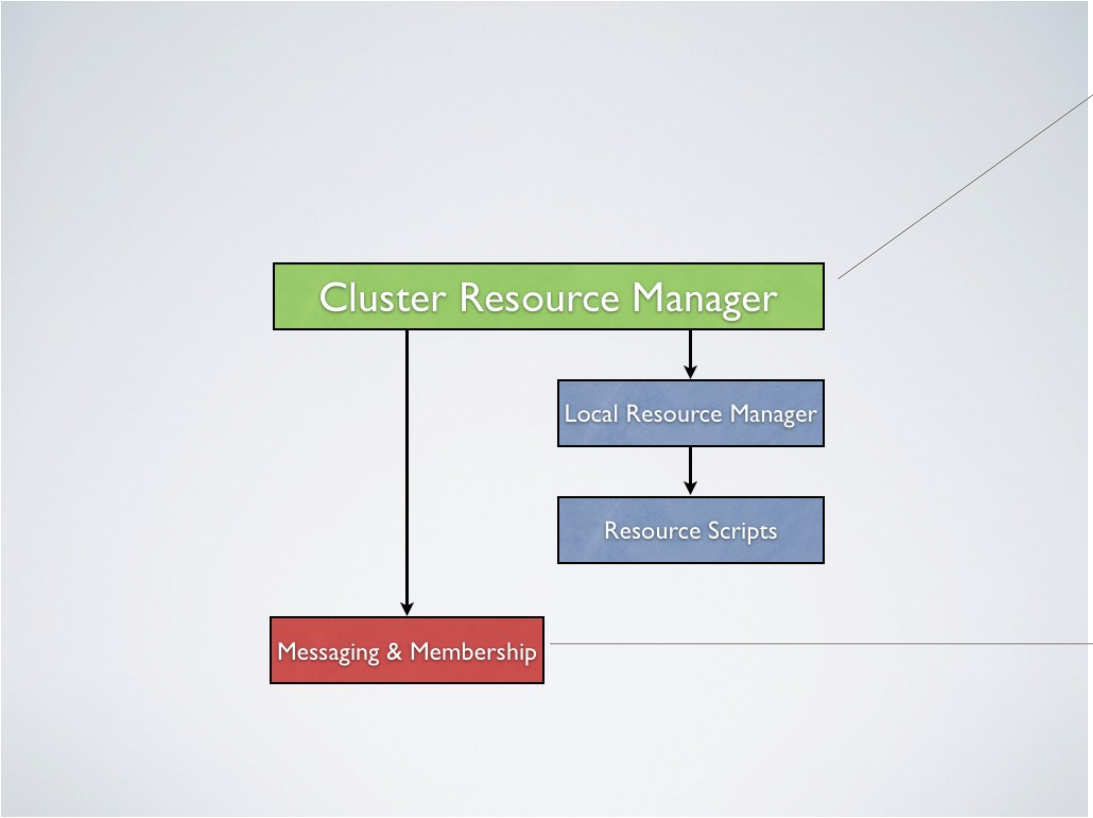
xm tool command line per la gestione delle istanze virtuali

avvio	**xm create** *file_di_configurazione*
arresto	**xm shutdown** *<Domain>*
poweroff	**xm destroy** *<Domain>*
migrazione	**xm migrate** *<Domain>*
lista vm	**xm list**
top hv	**xm top**
scheduler	**xm sched-credit** -d *<Domain>* -w weight
block-attach	**xm block-attach** *<Domain> <BackDev> <FrontDev> <Mode>*

Xen: file di configurazione di un'istanza

```
bootloader  = '/usr/lib/xen-default/bin/pygrub'
vcpus       = '2'
memory      = '1024'
root        = '/dev/xvda2 ro'
disk        = [
                  'drbd:web-root,xvda2,w',
                  'drbd:web-swap,xvda1,w',
              ]
name        = 'web'
vif         = [ 'mac=00:16:3E:61:01:03' ]
on_poweroff = 'destroy'
on_shutdown = 'destroy'
on_reboot   = 'destroy'
on_crash    = 'destroy'
extra       = 'clocksource=hpet'
```

Pacemaker



http://www.clusterlabs.org/wiki- Copyright © 2009-2011 Andrew Beekhof

crm: tool command line gestione cluster

crm node online/standby *<nodo>*

crm resource stop/start *<risorsa>*

crm resource migrate *<risorsa>*

crm configure

corosync: cluster engine (Service Availability Forum)

- closed process group
- simple availability manager
- quorum system

Pacemaker: virtualizzazione come risorsa

```
$ crm configure show
node nodo1 \
 attributes standby="off"
node nodo2 \
 attributes standby="off"
node nodo+N \
 attributes standby="off"
[...]
primitive web ocf:heartbeat:Xen \
 params xmfile="/etc/xen/ha/web.cfg" shutdown_timeout="20" \
 op monitor interval="10s" \
 op start interval="0s" timeout="60" \
 op stop interval="0s" timeout="40s" \
 meta target-role="Started"
[...]
location web-loc1 web 100: nodo1
location web-loc2 web 1: nodo+N
property $id="cib-bootstrap-options" \
 dc-version="1.0.9-74392a28b7f31d7ddc86689598bd23114f58978b" \
 cluster-infrastructure="openais" \
 expected-quorum-votes="3" \
 stonith-enabled="false" \
 last-lrm-refresh="1285163787" \
 symmetric-cluster="false"
rsc_defaults $id="rsc-options" \
```

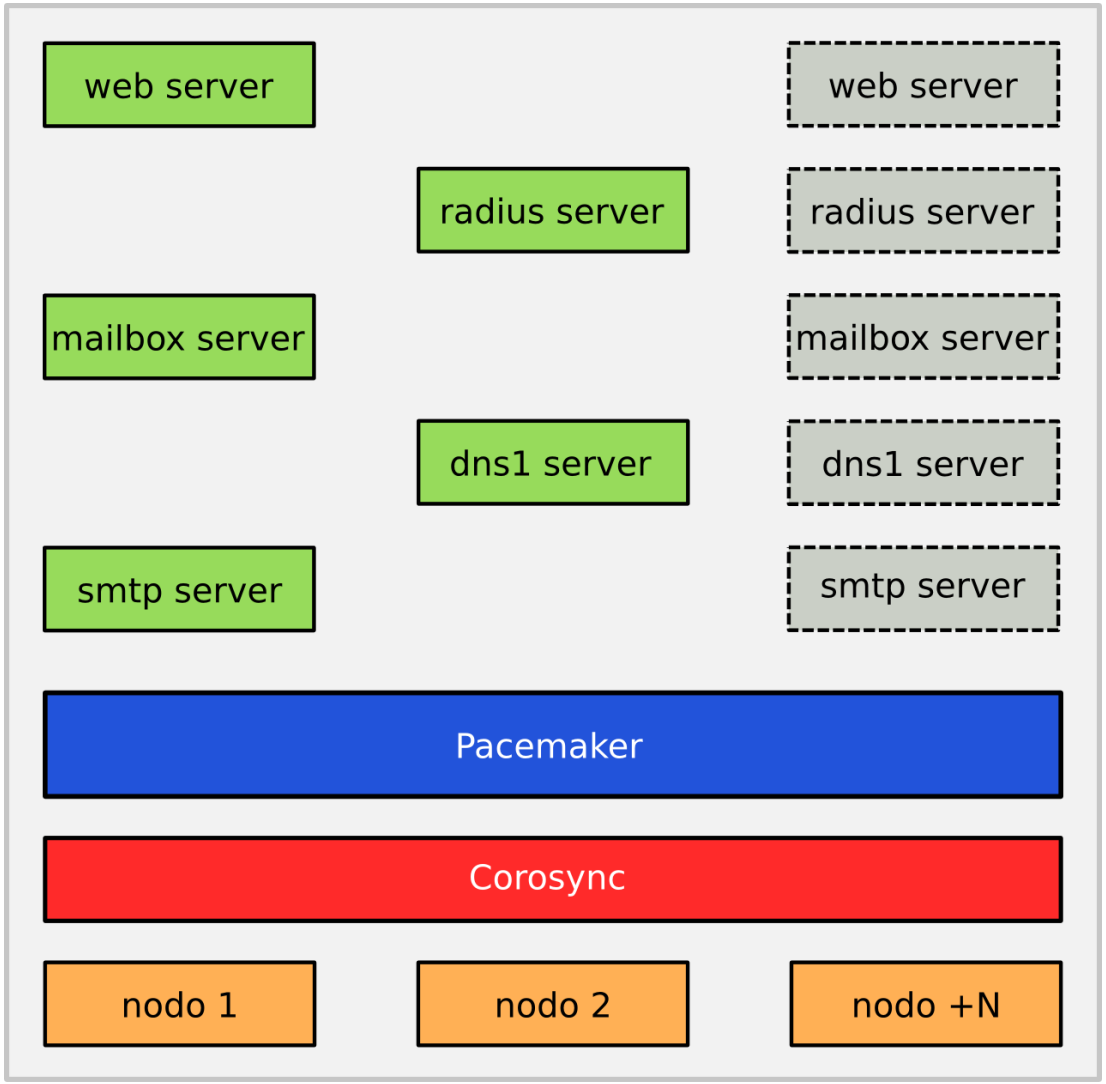
OCF Open Cluster Framework

ocf:heartbeat:Xen
/usr/lib/ocf/resource.d/heartbeat/Xen

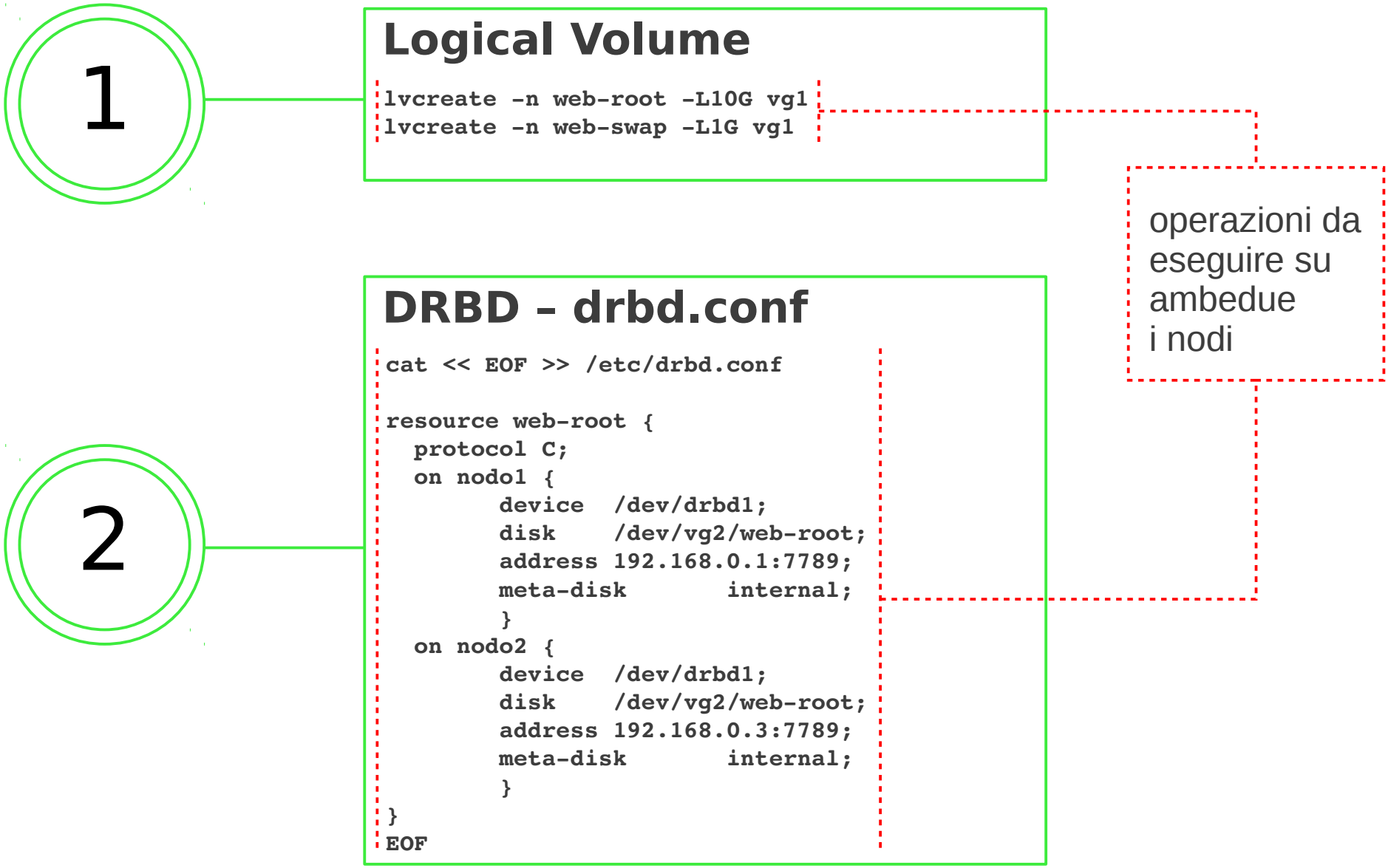
```
#!/bin/sh
#
#
# Support:      linux-ha@lists.linux-ha.org
# License:      GNU General Public License (GPL)
#
#         Resource Agent for the Xen Hypervisor.
#         Manages Xen virtual machine instances by
#         mapping cluster resource start and stop,
#         to Xen create and shutdown, respectively.
#
#         usage: $0  {start|stop|status|monitor|meta-data}
#
#           OCF parameters are as below:
#                   OCF_RESKEY_xmfile
#                           Absolute path to the Xen control file,
#                           for this virtual machine.
#                   OCF_RESKEY_allow_mem_management
#                           Change memory usage on start/stop/migration
#                           of virtual machine
#                   OCF_RESKEY_reserved_Dom0_memory
#                           minimum memory reserved for domain 0
#                   OCF_RESKEY_monitor_scripts
#                           scripts to monitor services within the
#                           virtual domain
[...]
```

location
<nome_location> <risorsa> <peso> <nodo>

Pacemaker: visione d'insieme



Aggiungere un'istanza passo passo



Aggiungere un'istanza passo passo

3

DRBD - metadevice

```
drbdadm create-md web-root
drbdadm up web-root

drbdadm -- --overwrite-data-of-peer primary web-root
```

operazioni da eseguire su ambedue i nodi

4

Untar

```
mkfs.ext4 /dev/drbd1
mkswap /dev/vg1/web-swap
mount /dev/drbd1 /mnt
cd /mnt
tar xpf /var/xen/image-archive/squeeze.tar
cd /
umount /dev/drbd1
```

Aggiungere un'istanza passo passo

5

Xen

```
 cat << EOF >> /etc/xen/ha/web.cfg
bootloader = '/usr/lib/xen-default/bin/pygrub'
vcpus       = '2'
memory      = '1024'
root        = '/dev/xvda2 ro'
disk        = [
                  'drbd:idm1-root,xvda2,w',
                  'phy:/dev/vg1/web-swap,xvda1,w',
                ]
name        = 'idm1'
vif         = [ 'mac=00:16:3E:61:01:13' ]
on_poweroff = 'destroy'
on_shutdown = 'destroy'
on_reboot   = 'destroy'
on_crash    = 'destroy'
extra       = 'clocksource=hpet'
EOF
```

6

Pacemaker

```
crm configure
primitive web ocf:heartbeat:Xen \
params xmfile="/etc/xen/ha/web.cfg" shutdown_timeout="15" \
op monitor interval="10s" \
op start interval="0s" timeout="60" \
op stop interval="0s" timeout="40s"
location web-loc1 web 100: nodo1
location web-loc2 web 1: nodo2
commit
bye
```

HA Cluster: un'architettura aperta

- Nodi completamente indipendenti
  - hardware
  - distribuzione/kernel (upgrade differenziate)
- Plain Old Xen
  - utilizzo altri tool di gestione
  - avvio indipendente da pacemaker
- Flessibilità
  - un unico tool per HA server e applicazioni

Altre soluzioni

ganeti

http://code.google.com/p/ganeti/

openQRM

http://www.openqrm.com

OpenStack Cloud Manager

http://www.openstack.org

Domande ?