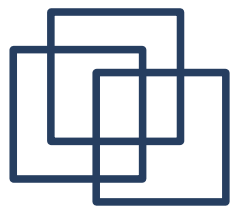


Virtualizzare con Xen

Bruno Bacci

*Responsabile Data Center Amministrazione Centrale
Universita' di Pisa*



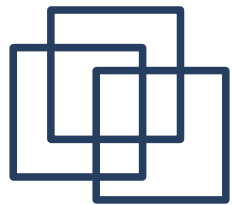
Xen: SO Supportati

- Paravirtualizzati

- Linux 2.4
- Linux 2.6
- NetBSD
- FreeBSD
- OpenSolaris

- HVM

- Windows Vista
- Windows XP
- Windows 2000
- Windows 2003
- Windows 2008
- Qualsiasi distro Linux



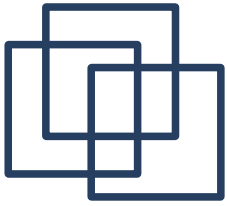
Il Mio Server Supporta HVM?

- Per processori AMD:

```
grep svm /proc/cpuinfo  
flags: fpu tsc msr svm extapic cr8_legacy
```

- Per processori Intel:

```
grep vmx /proc/cpuinfo  
flags: fpu tsc msr vmx extapic cr8_legacy
```



Xen Capabilities

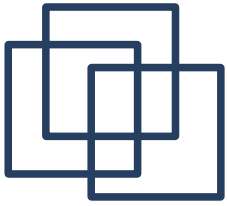
- Una volta installato Xen:

```
cat /sys/hypervisor/properties/capabilities
xen-3.0-x86_64 xen-3.0-x86_32p hvm-3.0-x86_32
hvm-3.0-x86_32p hvm-3.0-x86_64
```

- Alternativamente:

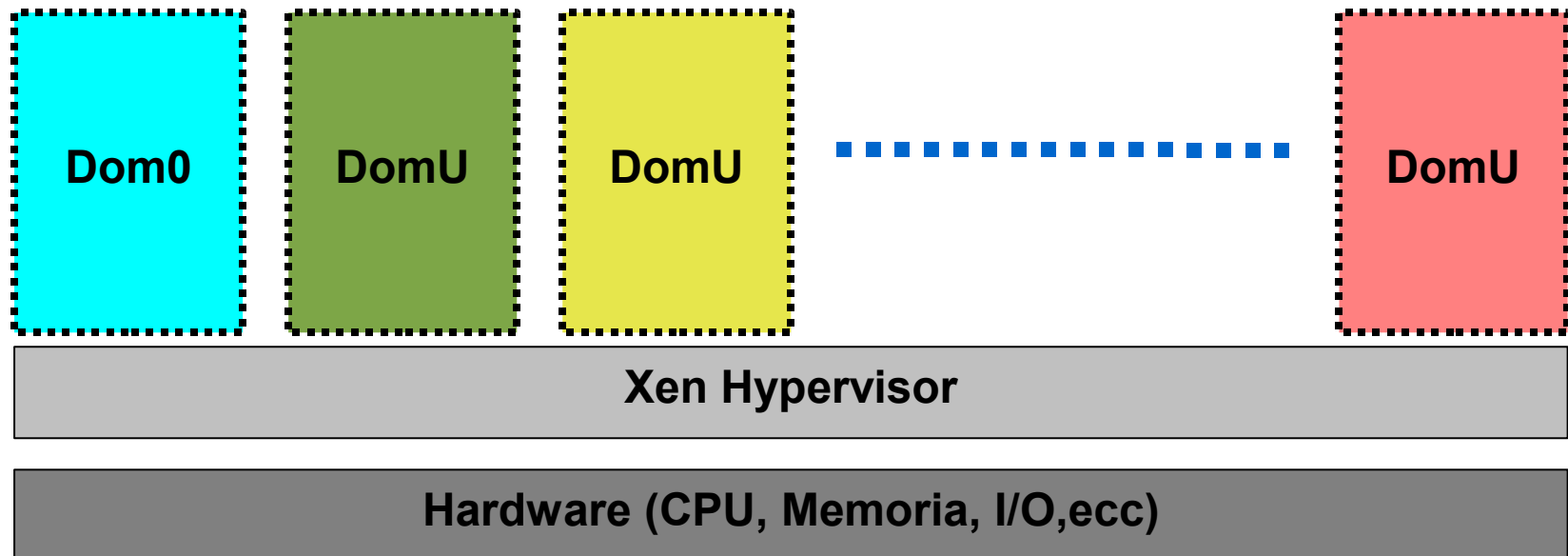
```
xm info
xen_caps      : xen-3.0-x86_64 xen-3.0-x86_32p hvm-3.0-x86_32
```

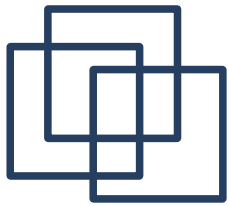
- **xen-3.0-x86_64**: paravirtualizzato 64 bit
- **xen-3.0-x86_32p**: paravirtualizzato 32 bit
- **hvm-3.0-x86_32**: HVM 32 bit
- **hvm-3.0-x86_32p**: paravirtualizzato con supporto HVM 32 bit
- **hvm-3.0-x86_64**: HVM 64 bit



Xen Domains

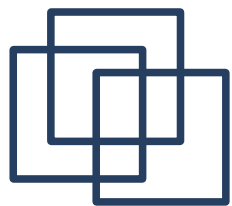
- Gli ambienti in cui sono eseguiti i sistemi operativi ospiti sono denominati *Domini*
- Esistono due tipi di domini:
 - *Domain 0* (*dom0*) dominio privilegiato
 - *Domain U* (*domU*) dominio non privilegiato





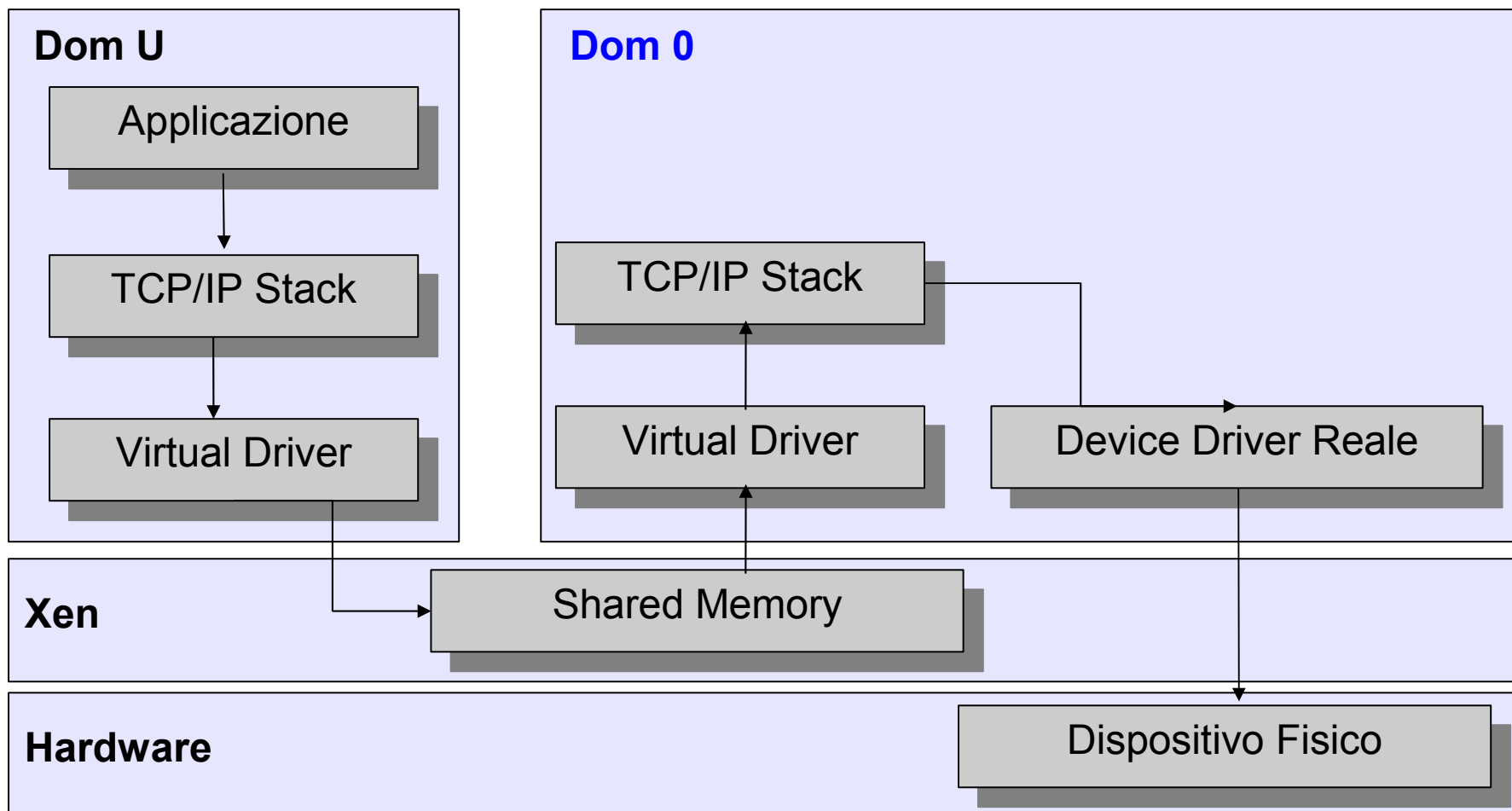
Il Ruolo di Dom 0

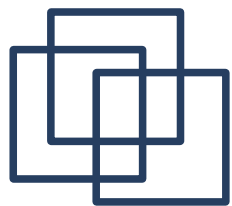
- Fornire i device driver e l'interfaccia grafica che non sono inclusi in Xen
- Mediare l'accesso all'HW che non supporta nativamente la capacita' di essere acceduto contemporaneamente da diversi sistemi operativi
- Fornire gli strumenti per gestire i Dom U (crearli, avviarli, distruggerli, ecc.)



Il Ruolo di Dom0 (2)

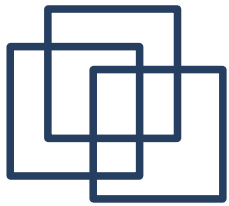
Esempio: invio di un pacchetto TCP/IP





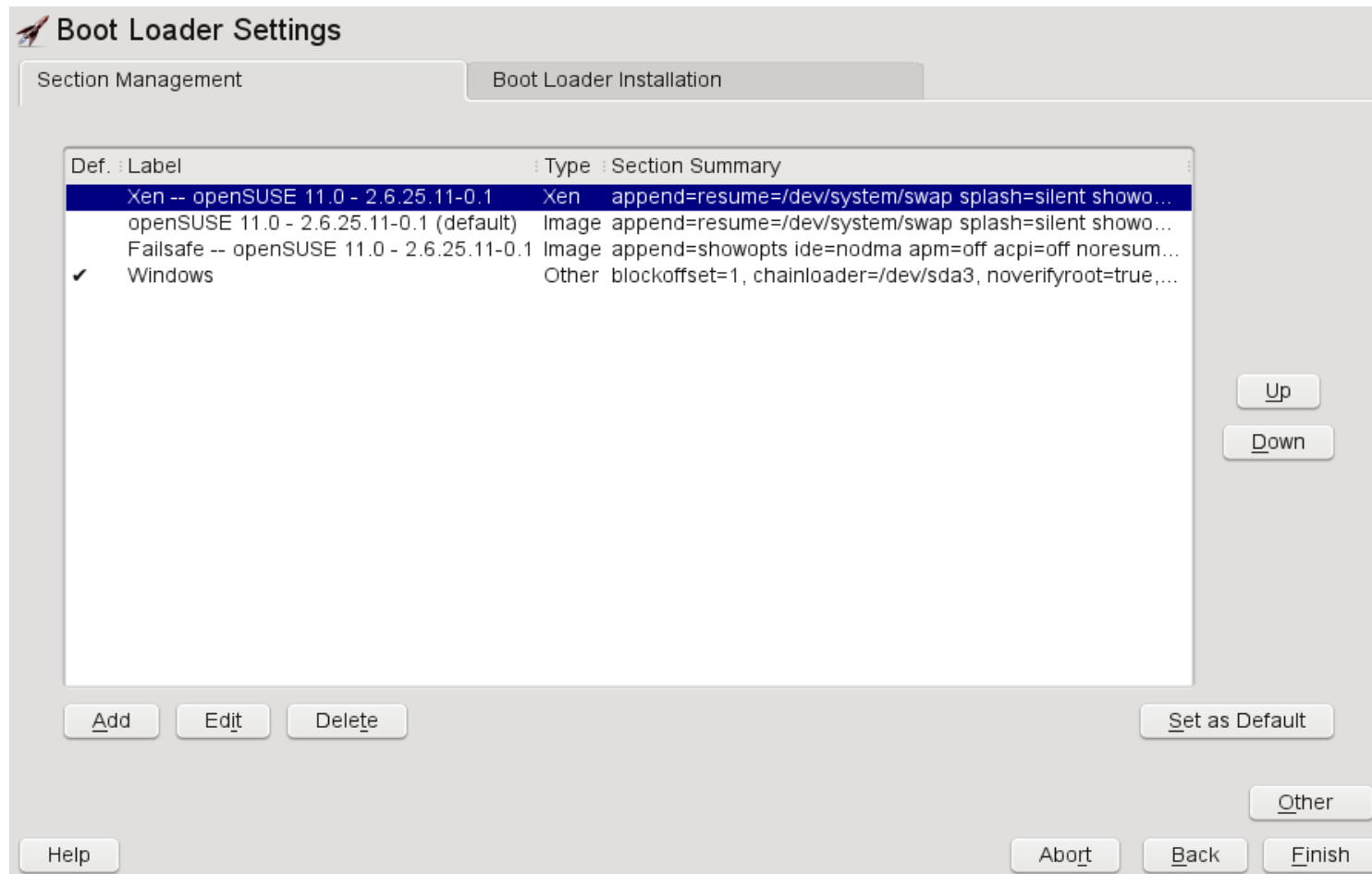
Installare Xen

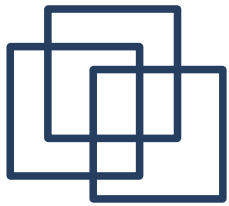
- Opzioni disponibili:
 - Scaricare, compilare ed installare manualmente
 - Utilizzare i pacchetti inclusi nelle varie distribuzioni Linux (OpenSuSe, Fedora, Ubuntu, Novell SuSE, Red Hat, ecc.)
- Il risultato e' una nuova entry nel boot loader che consente di effettuare il boot di Xen come se fosse un normale sistema operativo



Installare Xen (2)

- Xen appare come un immagine *bootable* in un sistema *multi-boot*





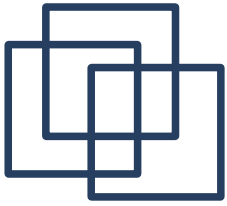
Xen Boot

- *Probing* ed inizializzazione dell'HW
- Mapping dei vari dispositivi e della memoria in modo che possa essere usata dai vari Domains
- Caricamento del kernel per Dom 0. Da questo punto il boot prosegue come un normale sistema Linux

```
XEN 3.1.1 (root@) (gcc version 4.1.2)
Latest ChangeSet: Thu Oct 11 10:12:07 2007
```

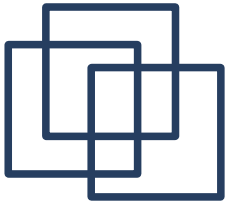
```
http://www.cl.cam.ac.uk/netos/xen
University of Cambridge Computer Laboratory
Xen version 3.1.1 (root@) (gcc version 4.1.2)
Latest ChangeSet: Thu Oct 11 10:12:07 2007
(XEN) Command line: noreboot dom0_mem=1G
(XEN) Video information:
(XEN) VGA is text mode 80x25, font 8x16
```

```
(XEN) *** LOADING DOMAIN 0 ***
(XEN) Xen kernel: 32-bit, PAE, lsb
(XEN) Dom0 kernel: 32-bit, PAE, lsb, paddr 0xc0100000
```



Usare Xen

- La gestione di macchine virtuali con Xen si basa su due strumenti:
 - Il programma `xm` (*command line*) con cui si controllano le macchine virtuali (creazione, shutdown, distruzione, ecc.)
 - Un file di configurazione con cui si specificano tutte le caratteristiche della macchina virtuale (dischi, cdrom, numero di CPU virtuali, RAM, configurazione della rete ecc.). Il programma `xm` utilizza questo file di configurazione per creare la macchina virtuale

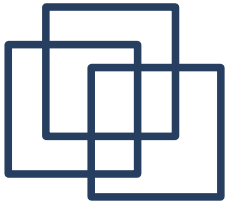


Xen Config File

- Contiene una serie di opzioni nella forma

value = name

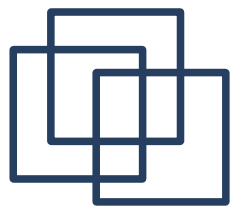
```
# Kernel image file.
kernel = "/usr/lib/xen/boot/hvmloader"
# The domain build function.
builder='hvm'
# Initial memory allocation (in megabytes) for the new
domain.
memory = 128
# A name for your domain. All domains must have different
names.
name = "ExampleHVMDomain"
# 128-bit UUID for the domain. The default behavior is to
generate a new UUID
# on each call to 'xm create'.
#uuid = "06ed00fe-1162-4fc4-b5d8-11993ee4a8b9"
#-----
# The number of cpus guest platform has, default=1
vcpus = 1
.....
.....
.....
```



Xen Config File: `name`

- *name*: definisce il nome del domino (macchina virtuale) che si sta creando.
 - Associa un nome mnemonico ad un dominio
 - Non e' legato in nessun modo con l'*hostname* eseguito all'interno del domino

```
name = "Mercurio"
```



Xen Config File: `builder`

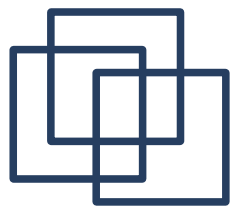
- *builder*: identifica la funzione che deve essere utilizzata per creare il nuovo dominio (domainU).

- Paravirtualizzato:

```
builder = "linux"
```

- Hardware Virtual Machine:

```
builder = "hvm"
```



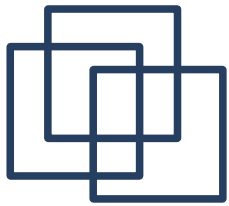
Xen Config File: `kernel`

- *kernel*: definisce il codice che deve essere caricato al momento del boot della macchina virtuale che si sta definendo.
 - Per macchine paravirtualizzate deve essere un kernel Xen-aware accessibile da Domain 0

```
kernel="/boot/vmlinuz-xen"
```

- Per macchine HVM e' il programma che emula il BIOS di una CPU x86

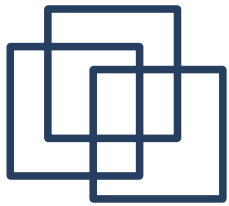
```
kernel="/usr/lib/xen/boot/hvmloader"
```



Xen Config File: `memory`

- `memory`: definisce in MB la quantita' di memoria fisica disponibile per il domainU che si sta definendo.
 - A ciascun domainU deve essere assegnata RAM sufficiente per effettuare il boot (ad esempio caricare il kernel ed i moduli)
 - La memoria totale utilizzata dai domainU e domain0 deve essere inferiore a quella fisicamente disponibile sul server
 - Se non e' disponibile sufficiente memoria fisica da assegnare ad un domainU il processo di creazione fallisce

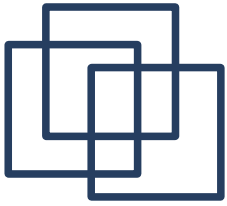
```
memory = 1024
```



Xen Config File: `vcpus`

- `vcpus`: definisce il numero di CPU virtuali associate ad un domainU
 - Xen ha avuto diversi scheduler. Al momento il piu' accreditato e' denominato *SMP Credit Scheduler*
 - Periodicamente lo scheduler esegue ogni Virtual CPU (VCPU) su una CPU reale
 - Sino a che e' possibile lo scheduler cerca di distribuire il piu' possibile le VCPU sulle CPU reali. In caso di *contention* lo scheduler esegue un arbitraggio *conservativo*.

```
vcpus = 3
```



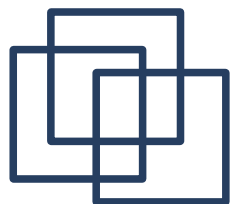
Xen Config File: `disk`

- *disk*: definisce i dischi a cui il dominio che si sta definendo ha accesso.
 - Espresso come un array di elementi ciascuno dei quali definisce un dispositivo di storage (disco, cdrom, ecc):

```
disk = ['elem1', 'elem2', ...]
```

- Ciascun elemento dell'array ha la forma:

```
backend-dev, frontend-dev, mode
```



Xen Config File: `disk` (2)

- `backend-dev`: indica il device come visto da Dom0. Sono supportati i seguenti formati:

- *Disk Image File*, il disco della macchina virtuale corrisponde ad un file:

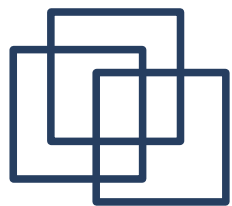
```
file:/Xen/Images/Mercurio_disk0.img
```

- *Physical Device*, il disco corrisponde ad una partizione dell'HD accessibile ma non usata da Dom0:

```
phy:/dev/sda7
```

Utilizzato anche per esportare CD-ROM:

```
phy:/dev/hda
```



Xen Config File: `disk` (3)

- `frontend-dev`: indica come il device appare nel dominio che si sta definendo

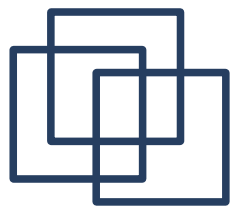
```
file: /Xen/Images/Mercurio_disk0.img, hda
```

```
phy: /dev/hda, hdc: cdrom
```

- `mode`: definisce le modalita' di accesso al device (**r** sola lettura, **w** lettura/scrittura)

```
'file: /Xen/Images/Mercurio_disk0.img, hda, w'
```

```
'phy: /dev/cdrom, hdc: cdrom, r'
```

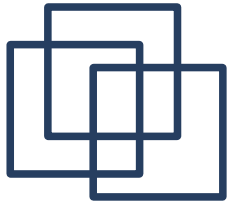


Xen Config File: disk (4)

- Per Hardware Virtual Machine l'immagine disk file è vuota. Lo possiamo creare con:

```
dd if=/dev/zero of=disk-image count=40960
```

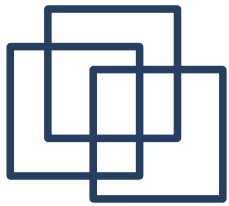
- Per macchine paravirtualizzate l'immagine disk file deve contenere un root filesystem. La maggior parte delle distribuzioni consente di creare macchine paravirtualizzate di se stesse
- L'immagine disk file può risiedere:
 - Local File System
 - Network File system (NFS, GFS, OCFS, ecc.)
 - Storage Area Network (iSCSI, Fiber Channel)



Xen Config File: boot

- *boot*: definisce la sequenza di boot del dominio che stiamo creando:
 - Boot on floppy: *a*
 - Boot from disk: *c*
 - Boot from network: *n*
 - Boot from CD: *d*

```
# default: hard disk, cd-rom, floppy  
boot="dc"
```



Xen & Networking

- Ad ogni dominio e' associato un identificatore intero (*ID*). Dom0 ha ID 0, il primo DomU ha ID 1, ecc.
- Ogni volta che viene creato un DomU per ogni interfaccia di rete vengono create una coppia di interfacce virtuali una associata con Dom0 ed una associata con DomU:

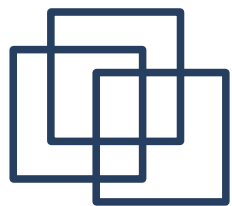
`<vifID.Int, ethInt>`

dove: ID e' l'intero che identifica il dominio ed Int e' 0 per la prima interfaccia di rete del dominio, 1 per la seconda ecc.

Ad esempio se Dom 1 ha due interfacce di rete:

`<vif1.0, eth0>`

`<vif1.1, eth1>`

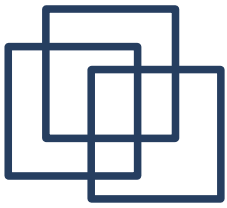


Xen & Networking (2)

- Ad ogni interfaccia virtuale vifX.Y viene associato un MAC Address
- Il pacchetto di indirizzi Ethernet riservato a Xen ha i primi tre gruppi esadecimali composti da 00:16:3e

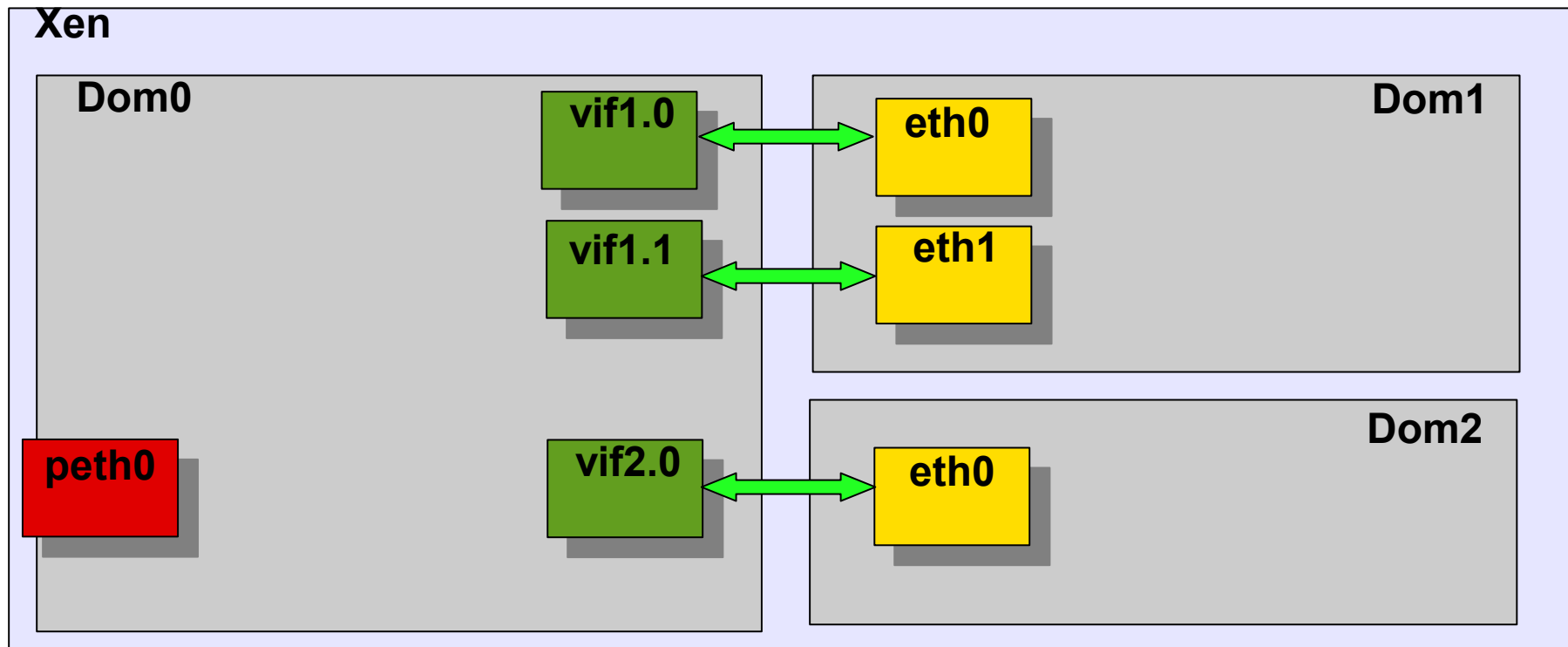
00:16:3e:xx:xx:xx

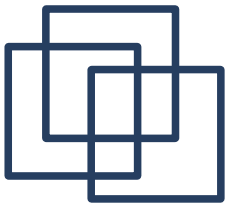
- Può essere assegnato in modo randomico o può essere deciso dall'amministratore



Xen & Networking (3)

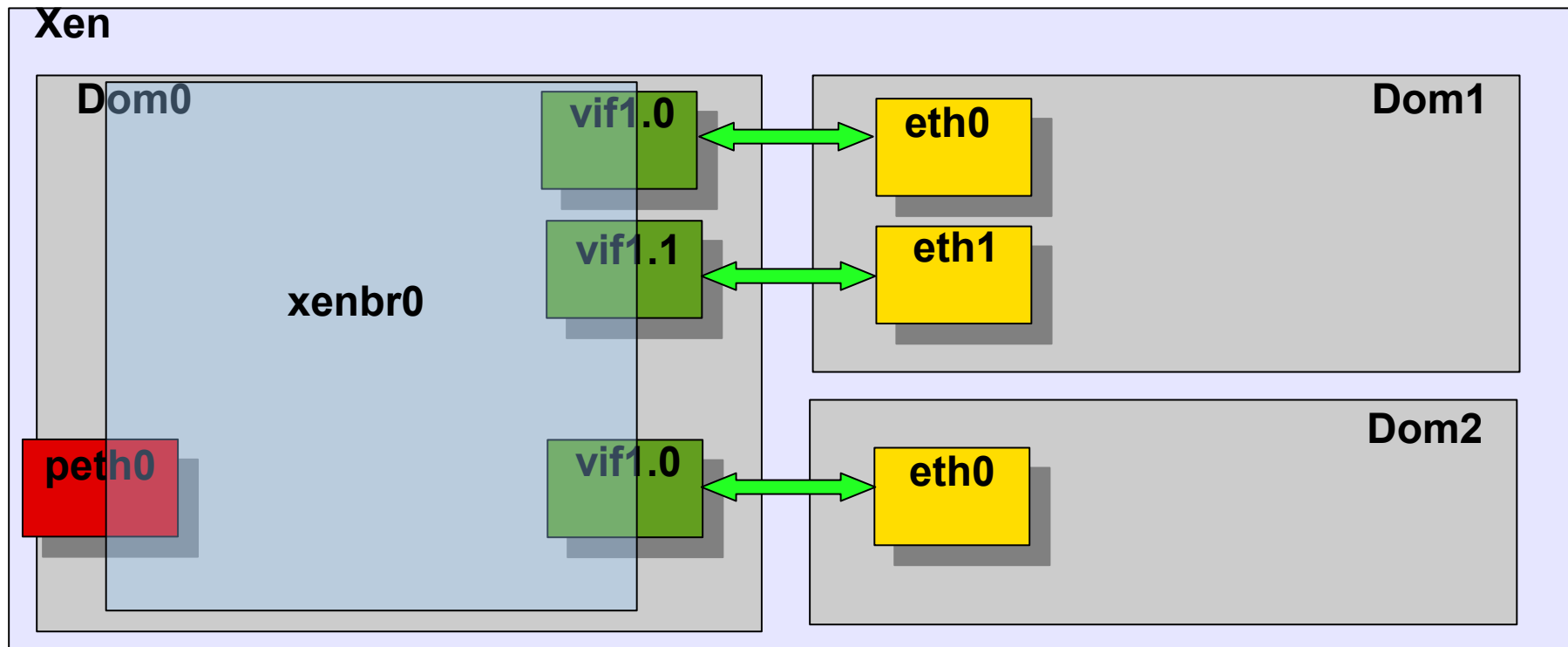
- L'interfaccia fisica e' gestita da Dom0 e denominata *peth0* (*peth1*, ecc.)





Xen & Networking (4)

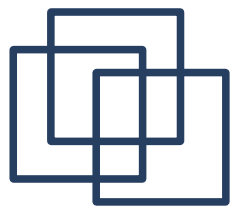
- L'interfaccia fisica e' connessa alle virtuali (vifX.Y) sfruttando il bridging di Linux (non e' il solo approccio possibile)





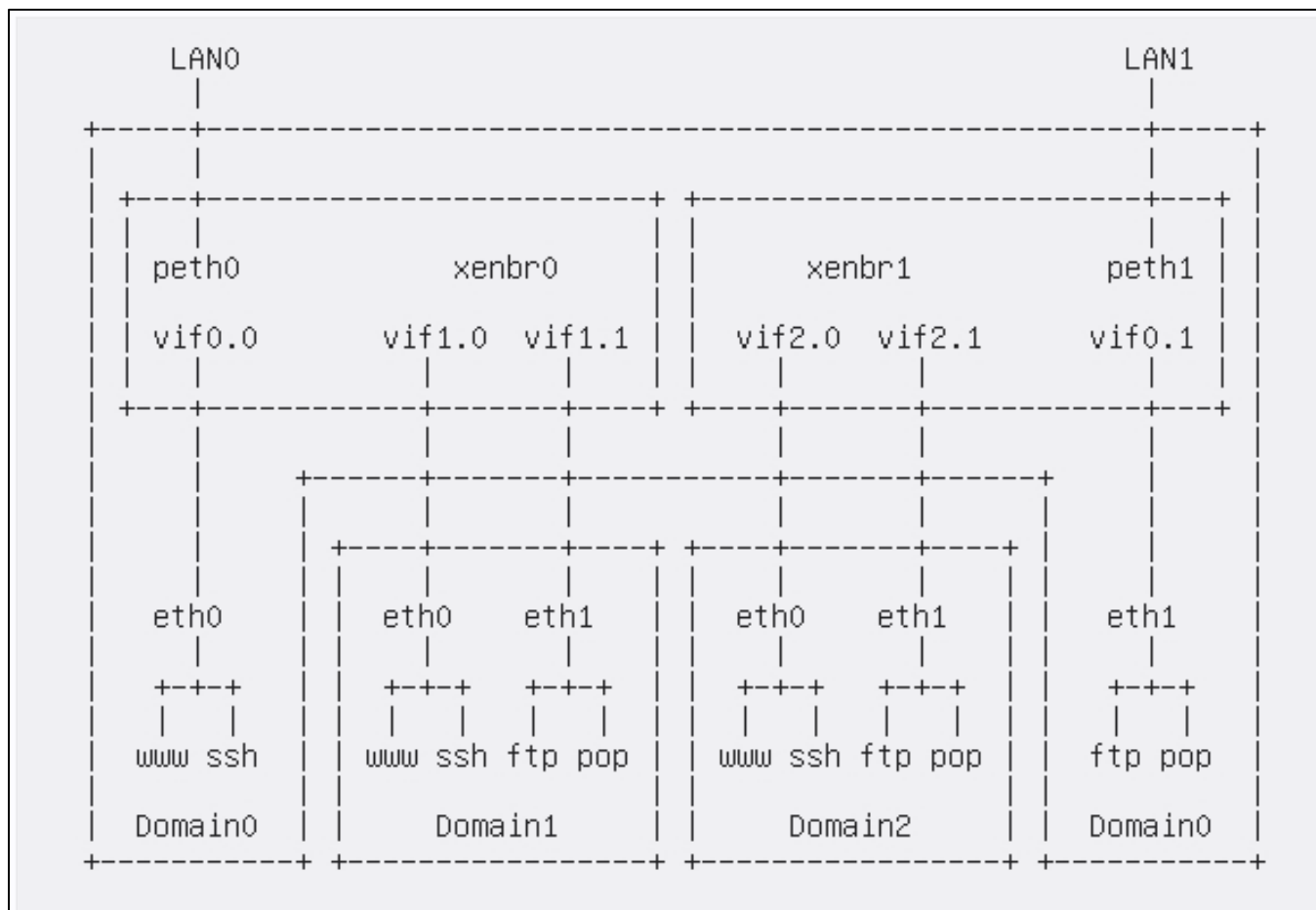
Xen & Networking (5)

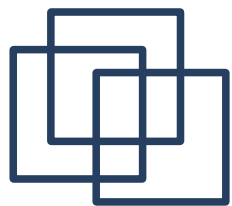
- Flusso dei pacchetti Ethernet:
 - Il pacchetto arriva sull'interfaccia fisica e gestito dal driver di peth0 di Dom0
 - Sulla base del MAC di destinazione il bridge instrada il pacchetto alla corretta interfaccia virtuale (vifX.Y) collegata al bridge
 - L'interfaccia virtuale (vifX.Y) che riceve il pacchetto lo passa a Xen che lo inserisce nell'interfaccia ethY del dominio con ID pari ad X



Xen & Networking (6)

- Possibile gestire piu' interfacce fisiche





Xen & Networking (7)

- La definizione dei bridge e' controllata da un insieme di script su Dom0

```
virago2:~ # /etc/xen/scripts/network-bridge status
-----
bridge name      bridge id          interfaces
xenbr0           8000.feffffffffff vif0.0
                  peth0
                  vif1.0
                  vif2.0
                  vif3.0
                  vif5.0
                  vif6.0

xenbr1           8000.feffffffffff vif0.1
                  peth2
                  vif1.1
                  vif5.1
                  vif12.1
                  vif19.1
                  vif17.1

xenbr2           8000.feffffffffff vif0.2
                  peth3
                  vif5.2

xenbr3           8000.feffffffffff vif0.3
                  peth4
                  vif4.0
```



Xen Config File: *vif*

- *vif*: definisce le interfacce di rete a cui il dominio che si sta definendo ha accesso.
 - Espresso come un array di elementi ciascuno dei quali definisce un'interfaccia di rete:

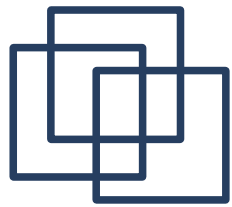
```
vif = ['interf1', 'interf2', ...]
```

- Ciascun elemento dell'array ha la forma:

```
mac, bridge
```

Esempio:

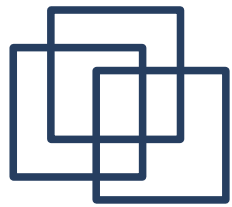
```
mac=00:16:3e:27:53:d4,bridge=xenbr0
```



Xen & Command Line (1)

- `xm` e' la *command line interface* per gestire i domini.
- Deve essere in esecuzione lo Xen Control Daemon (`xend`)
- Molte delle operazioni realizzate da `xm` sono asincrone
- Forma generale:

```
xm subcommand domain-id [OPTIONS]
```



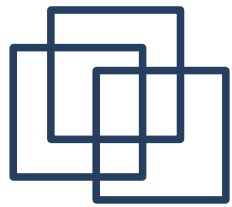
Xen & Command Line (2)

- Creazione di un dominio:

```
xm create [-c] configfile
```

- `-c` collega la console al dominio appena esso va in esecuzione
 - `configfile` il path assoluto del file di configurazione del dominio che si vuole creare
- Collega la console ad un dominio già creato

```
xm console domain-id
```



Xen & Command Line (3)

- Shutdown di un dominio:

```
xm shutdown domain-id
```

- Reboot di un dominio

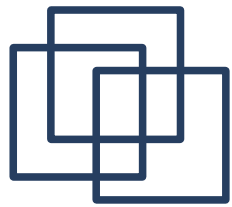
```
xm reboot domain-id
```

- Distruzione di un dominio:

```
xm destroy domain-id
```

- Informazioni sullo stato dei domini:

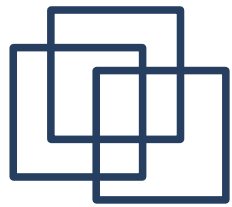
```
xm list [domain-id]
```



Xen & Command Line (4)

- Inoltre `xm` consente operazioni quali:
 - Live Migration
 - Aggiungere dispositivi virtuali quali interfacce di rete o dischi mentre un dominio e' in esecuzione
 - Salvare lo stato di un dominio mentre e' in esecuzione e riprestinarlo successivamente
- Informazioni esaustive:

```
man xm
```



Xen: creare HVM "a mano"

- Creazione di un image disk file vuoto

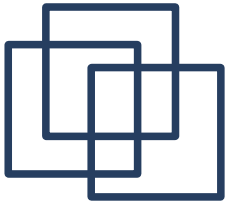
```
dd if=/dev/zero of=disk-image count=40960
```

- Creazione di un file di configurazione (win2003.cfg)

```
name="win2003VM"  
Memory=3072  
Vcpus=1  
builder="hvm"  
kernel="/usr/lib/xen/boot/hvmloader"  
boot="dc"  
disk=[ 'file: /VM_XEN_02/Satirol/satirol_disk0,hda,w',  
       'phy:/dev/cdrom,hdc:cdrom,r', ]
```

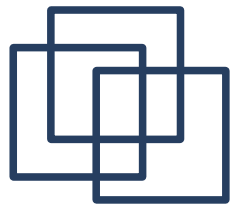
- Creazione del dominio tramite `xm`

```
xm create -c win2003.cfg
```



Xen & GUI

- Le maggiori distribuzioni includono strumenti grafici per la creazione e gestione di domini (es. Virtual Machine Manager)
- Vantaggi:
 - Supporto per l'installazione di diverse distribuzioni Linux paravirtualizzate
 - Supporto per l'installazione di HVM (le principali versioni di Windows)
 - Collegamento alla console dei vari domini
 - Monitoraggio dei domini in esecuzione (CPU, Memoria, ecc.)



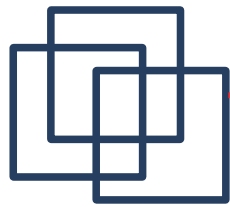
Virtual Machine Manager

Type of Operating System

Please specify the type of operating system that will run within the virtual machine. This defines many defaults, and helps decide how to start paravirtualized operating systems.

- ▶ NetWare
- ▶ Other
- ▶ RedHat
- ▼ SUSE
 - Novell Open Enterprise Server 2 (Linux)
 - SUSE (other)
 - SUSE Linux Enterprise Desktop 10
 - SUSE Linux Enterprise Server 8
 - SUSE Linux Enterprise Server 9
 - SUSE Linux Enterprise Server 10**
 - openSUSE
- ▼ Solaris
 - Solaris 9 and older
 - Solaris 10
- ▼ Windows
 - Windows (other)
 - Windows (other, x64)
 - Windows NT
 - Windows Server 2008
 - Windows Server 2008 (x64)
 - Windows Vista
 - Windows Vista (x64)
 - Windows XP, 2000, 2003
 - Windows XP, 2003 (x64)

Cancel Back Forward



Virtual Machine Manager (2)

Summary

Click any headline to make changes. When the settings are correct, click **OK** to create the VM.

Virtualization Method
Paravirtualized

Name of Virtual Machine
sles10

Hardware
Initial Memory: 512 MB
Maximum Memory: 131072 MB
Virtual Processors: 1

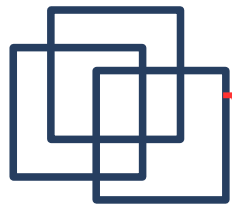
Graphics
Paravirtualized Graphics Adapter

Disks
1: 4.0 GB Hard Disk (file:/var/lib/xen/images/sles10/disk0)

Network Adapters
1: Paravirtualized; Randomly generated MAC address

Operating System Installation
Operating System: SUSE Linux Enterprise Server 10
Installation Source:
Automated Installation:
Additional Arguments:

Cancel **Back** **OK**




Virtual Machine Manager (3)


Virtualization Method

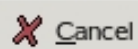
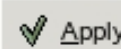
Virtual machines can use paravirtualization or full virtualization. Paravirtualization is faster but requires operating system support. Full virtualization runs a broader range of operating systems but requires hardware support. Which do you prefer?

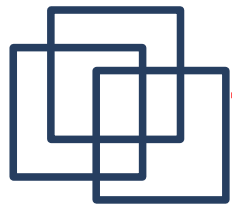
Paravirtualization

Full virtualization

 The operating system does not support full virtualization.

 The processor(s) in this machine do not support full virtualization.



Virtual Machine Manager (4)

Hardware

Specify the amount of memory and number of processors to allocate for the VM.

Memory:

Available Memory: 14395 MB


Initial Memory: MB

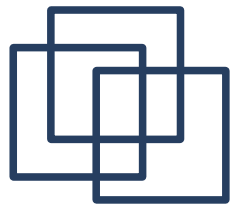
Maximum Memory: MB

Processors:

Available Processors: 8

Virtual Processors:

 For best performance, the number of virtual processors should be less than or equal to the number of physical processors.



Virtual Machine Manager (5)

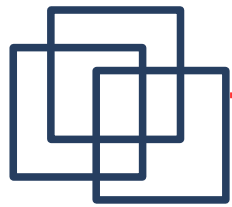
Disks

Name	Type	Source	Size (GB)
1	Hard Disk	file:/var/lib/xen/images/sles10/disk0	4.0
2	Hard Disk	file:/var/lib/xen/images/sles10/disk1	30.0
3	CD-ROM or DVD	phy:/dev/cdrom	?

Up Down

CD-Rom Harddisk Edit Remove

Cancel Apply



Virtual Machine Manager (6)

Virtual Network Adapter

Please specify the settings for the virtual network adapter.

Name: 1

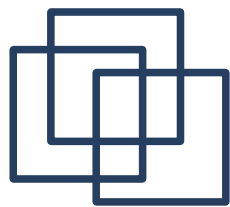
Type:

Source:

MAC Address:

Randomly generated MAC address

Specified MAC address 00:16:3e:



Virtual Machine Manager (7)

Create a Virtual Machine (on afrodite1)

Operating System Installation

Specify the bootable virtual disk (often labeled as Disk 1) or the network installation source URL. Each CD, DVD, or ISO image required for installation must be added as a virtual disk.

Virtual Disk:

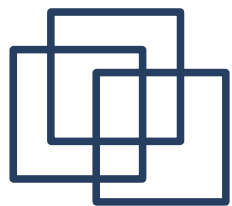
Network URL:

Some operating systems support automating the installation by specifying a URL or file(s). Select a directory to include multiple files.

AutoYaST file:

Some operating systems accept additional arguments, used to customize the installation or boot process.

Additional Arguments:

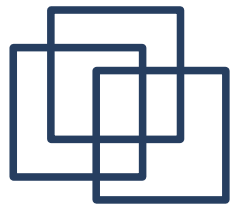


Windows 2003 su Xen

The screenshot displays a Xen Virtual Machine Manager interface. The main window shows the boot screen of a Windows Server 2003 Enterprise Edition virtual machine. The boot screen includes the Microsoft logo, the text "Windows Server 2003 Enterprise Edition", and the instruction "Press Ctrl-Alt-Delete to begin." Below this, it states "Requiring this key combination at startup helps keep your computer secure. For more information, click Help." The taskbar at the bottom of the VM window shows icons for Home and Trash, and the system tray displays the date and time as "Tue Oct 16, 22:18".

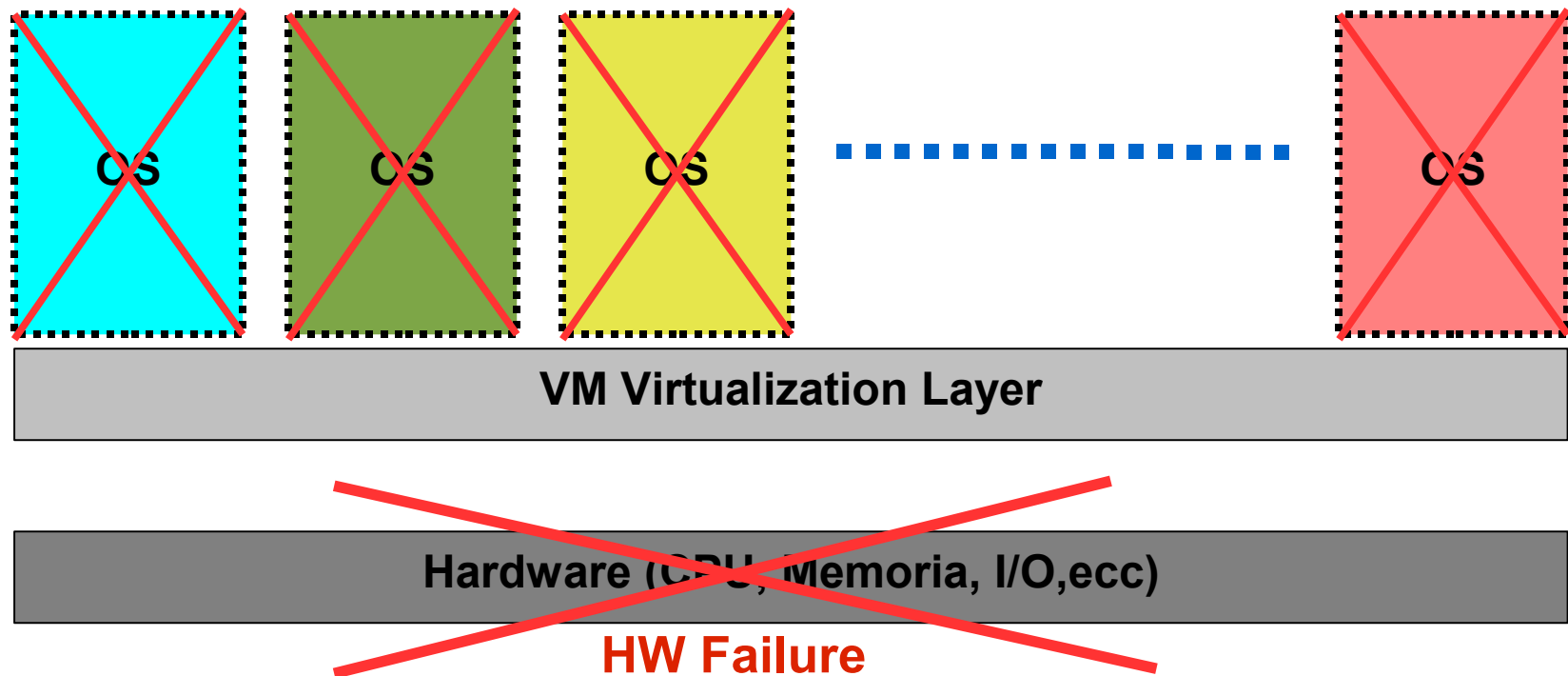
On the right side of the interface, the Virtual Machine Manager (Xen: afroditel1.adm.unipi.it) displays a table of running virtual machines:

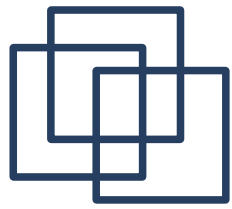
ID	Name	Status	CPU usage	VCPUs	Memory usage
1	Cupido	Running	0.12 %	1	2.01 GB 6 %
0	Domain-0	Running	6.08 %	4	2.00 GB 6 %
2	Satiro1	Running	0.17 %	1	3.01 GB 9 %
3	Satiro2	Running	0.18 %	1	4.01 GB 12 %
4	Satiro6	Running	0.18 %	1	3.01 GB 9 %



Xen & High Availability

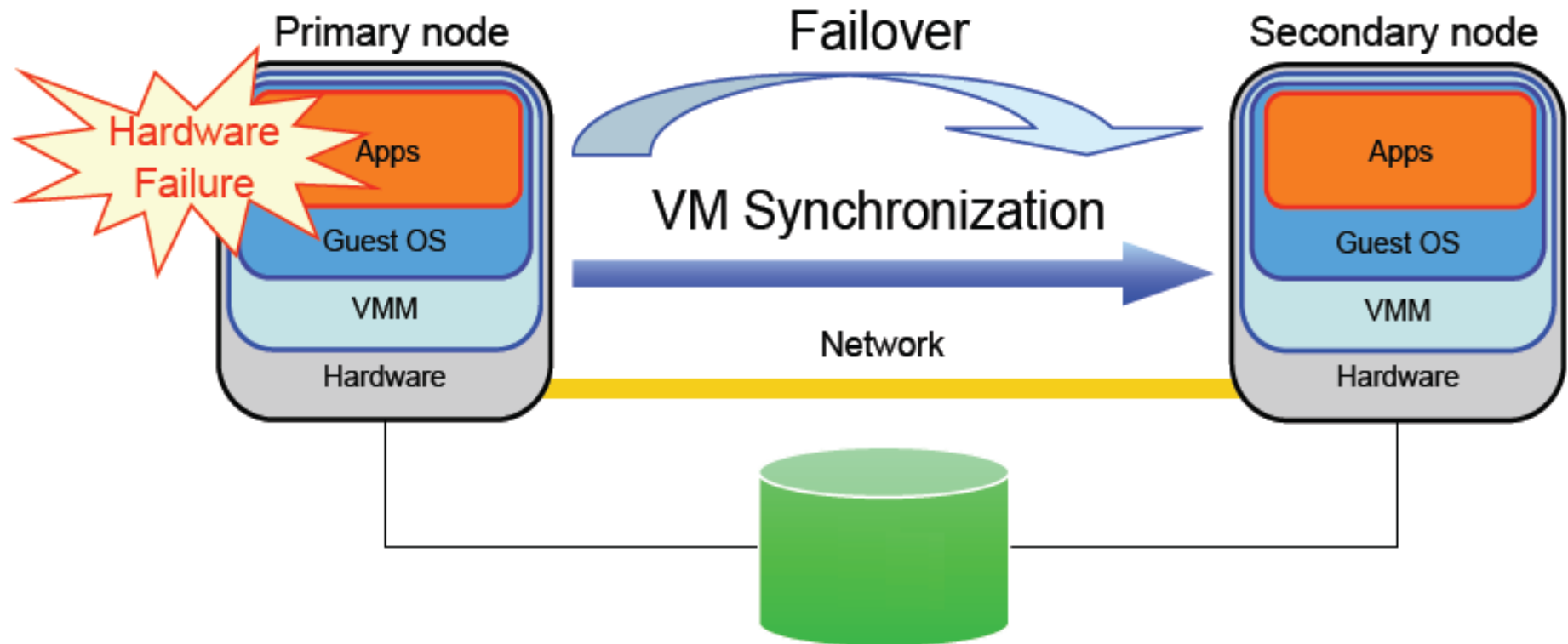
Single Point of Failure

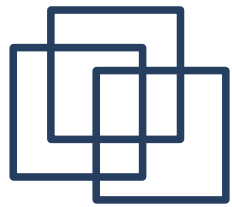




Xen & High Availability(2)

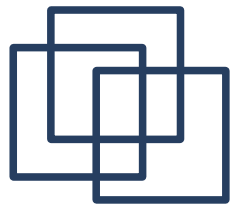
Immagini dei domini condivisi tra macchine fisiche





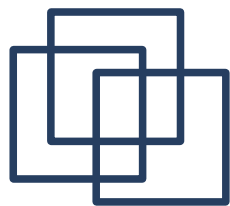
Xen & High Availability(3)

- La condivisione dello storage puo' essere ottenuta con soluzioni che hanno costi, performance e complessita' differente:
 - NFS
 - Distributed Replicated Block Device (DRDB) consente un *realtime mirror*
 - iSCSI
 - SAN (Fiber Channel)



Xen & High Availability⁽⁴⁾

- Xen non include software per HA
- Si utilizza *Heartbeat* un software per il *clustering* che viene installato sul Dom0 delle macchine fisiche che appartengono al cluster
- Heartbeat vede i servizi come delle risorse, li monitora e in caso di fallimento li riavvia
- Heartbeat tratta i domini di Xen come risorse
- Ottima interfaccia grafica (`hb_gui`)



Xen & High Availability(5)

Inserire un domnio come una risorsa di Heartbeat

Add Native Resource (on afrodite1)

Resource ID: Belong to group: (type for new one)

Type(double click for detail):

Name	Class/Provider	Description
WinPopup	ocf/heartbeat	WinPopup resource agent
Xen	ocf/heartbeat	Manages Xen DomUs
Xinetd	ocf/heartbeat	Xinetd resource agent

Parameters:

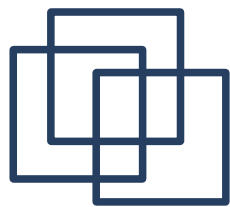
Name	Value	Description
target_role	stopped	press "Default" or "Start" button in toolbar/menu to start the resource
xmfile		Xen control file

If belong to a clone or master/slave:

Clone Master/Slave Clone or Master/Slave ID:

clone_max: clone_node_max:

master_max: master_node_max:



Xen & High Availability(6)

La console di Heartbeat

The screenshot displays the Heartbeat console interface. The left pane shows a tree view of the cluster configuration:

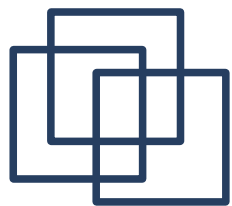
Name	Status
linux-ha	with quorum
Nodes	
afrodite1	running
Cupido	running on [afrodite1]
Satiro1	running on [afrodite1]
Satiro2	running on [afrodite1]
Satiro6	running on [afrodite1]
afrodite2	running(dc)
Satiro7	running on [afrodite2]
Satiro-Coll	running on [afrodite2]
Satiro3	running on [afrodite2]
Satiro4	running on [afrodite2]
afrodite3	running
Satriro8	running on [afrodite3]
Artemide	running on [afrodite3]
Arianna	running on [afrodite3]
Ate	running on [afrodite3]
Hostingw	running on [afrodite3]
afrodite4	running
Satiro5	running on [afrodite4]
Hostingw2	running on [afrodite4]

The right pane shows configuration parameters:

- Version: 2.0.8
- Debug Level: 0
- UDP Port: 694
- Keep Alive: 1000ms
- Warning Alive: 15000ms
- Dead Time: 30000ms
- Initial Dead Time: 30000ms
- Symmetric Cluster
- Stonith Enabled
- Transition Timeout: 20s
- Resource Stickiness: 0
- No Quorum Policy: stop
- Resource Failure Stickiness: 0

Buttons: Apply, Reset

Connected to 127.0.0.1



Xen & Live Migration

- Lo storage condiviso consente di migrare un dominio tra due macchine fisiche senza effettuare un restart dei domini

