

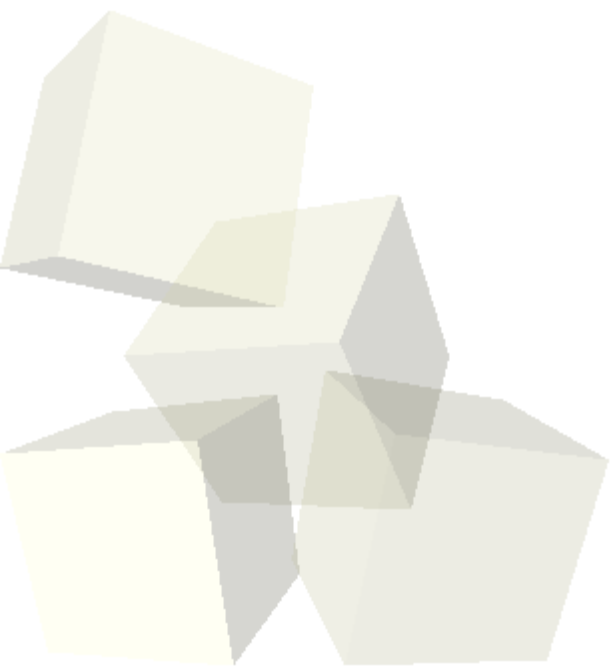
The Open Source Xen Hypervisor

Introduction to the Open Source Xen Hypervisor

Todd Deshane and Patrick Wilbur
(Clarkson University)

Stephen Spector (Citrix)

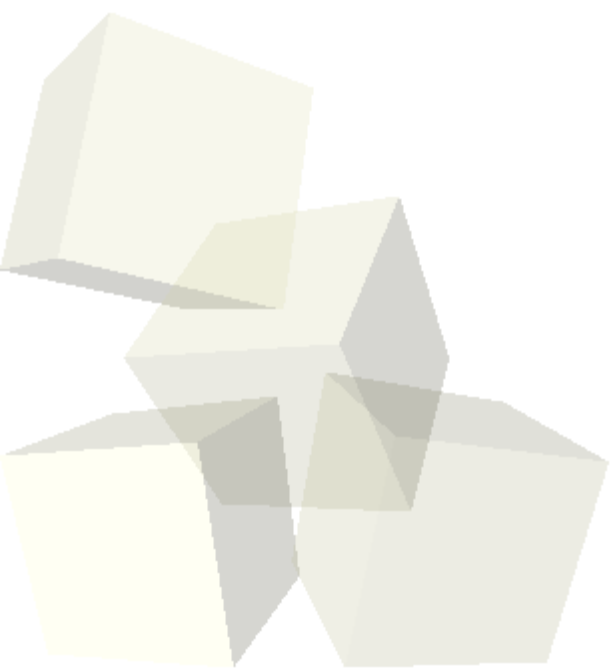
June 22, 2008



Overview of the Day and Schedule

- Unit 1: Virtualization and Xen Overview
- Morning Break
- Unit 2: Installing, Configuring, and Basic Usage
- Lunch
- Unit 3: Devices and Advanced Configuration
- Afternoon Break
- Unit 4: Security, Admin Tools, and Performance

Unit 1





Unit 1: Virtualization and Xen

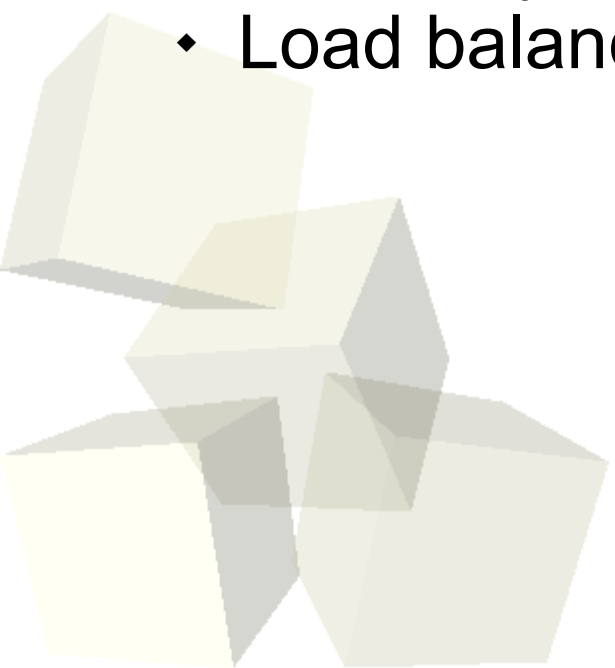
Overview

- Why Virtualize?
- Potential Pitfalls
- Types of Virtualization
- Xen Background
- Hardware-Assisted Virtualization Overview
- Xen Architecture
- Xen LiveCD Demonstration



Why Virtualize?

- Consolidation of servers
- Support heterogeneous and legacy OSes
- Rapid deployment and provisioning
- Advantages
 - ◆ Testing/debugging before going into production
 - ◆ Recovery and backup
 - ◆ Load balancing



Potential Pitfalls

- Hardware/software licensing complexities
- Added complexity per server
 - ◆ More to lose with a single server failure
- Lack of IT staff with virtualization skills
- Security challenges
- Loss of deterministic hardware performance
 - ◆ Mitigated with configuration

Types of Virtualization

■ No Virtualization

- ◆ Operating system running on native hardware

■ Emulation

- ◆ Fully-emulate a hardware architecture
 - Can be different than actual hardware architecture
- ◆ Unmodified guest OSes
- ◆ Examples: QEMU, Bochs

■ Full

- ◆ Simulate the base hardware architecture
 - Binary re-writing
 - Hardware-assisted virtualization
- ◆ Unmodified guest OSes
- ◆ Examples: VMware, VirtualBox, QEMU + kqemu, MS Virtual PC, Parallels, Xen/KVM (hardware-assisted)

Types of Virtualization

■ Para

- ◆ Abstracted base architecture
- ◆ Modified guest OSes
- ◆ Examples: Xen, UML, Lguest

■ Operating System Level

- ◆ Shared kernel (and architecture), separate user spaces
- ◆ Homogeneous guest OSes
- ◆ Examples: OpenVZ, Linux-VServer, Solaris Containers, FreeBSD Jails

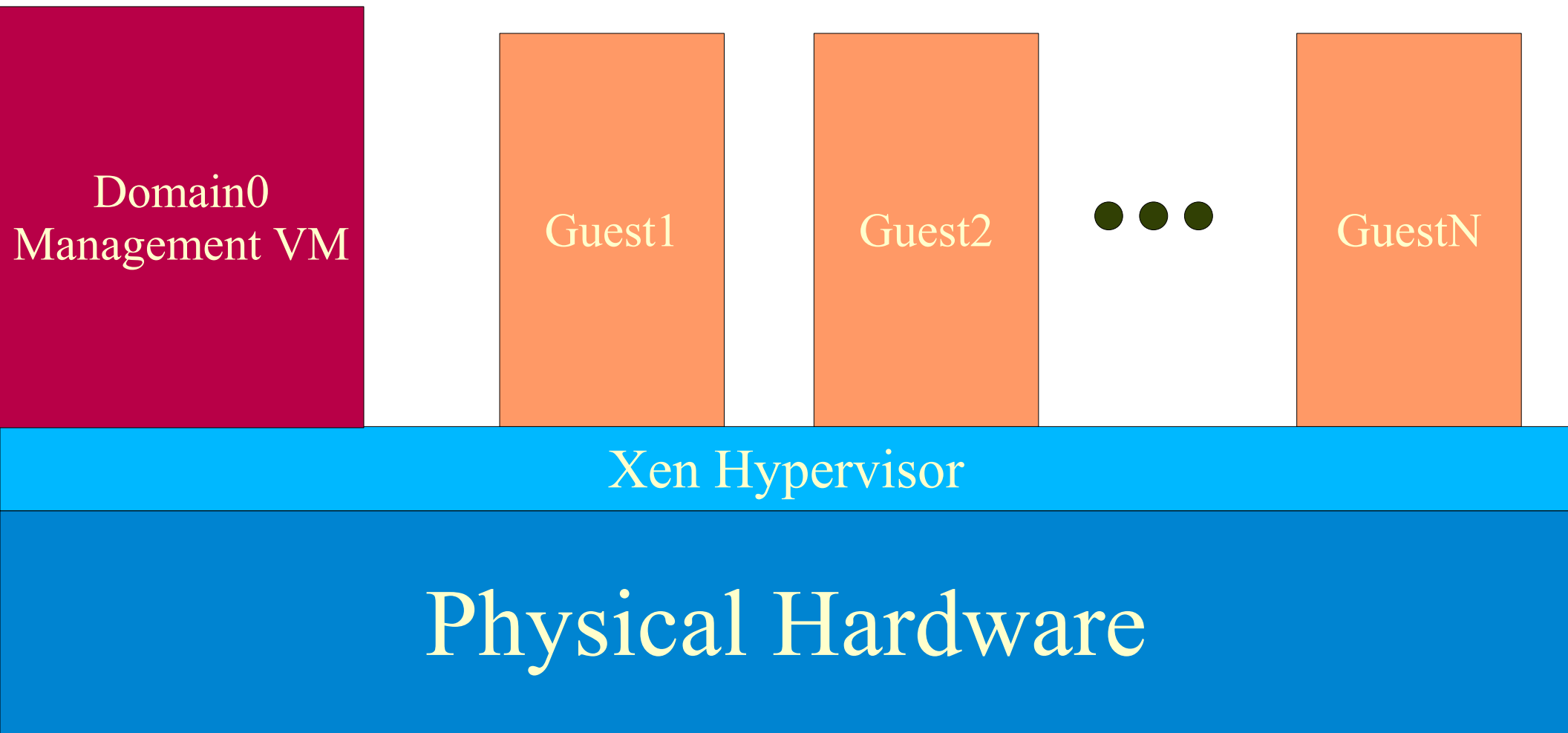
Xen Background

- Paravirtualization (PV)
 - ◆ High performance (claim to fame)
- Full Virtualization
 - ◆ With hardware support
 - ◆ Hardware Virtual Machine (HVM)
- Advantages
 - ◆ Open source
 - ◆ Standalone hypervisor
 - Citrix Xen Server
 - Virtual Iron
 - Solaris xVM
 - Oracle VM

Hardware-Assisted Virtualization

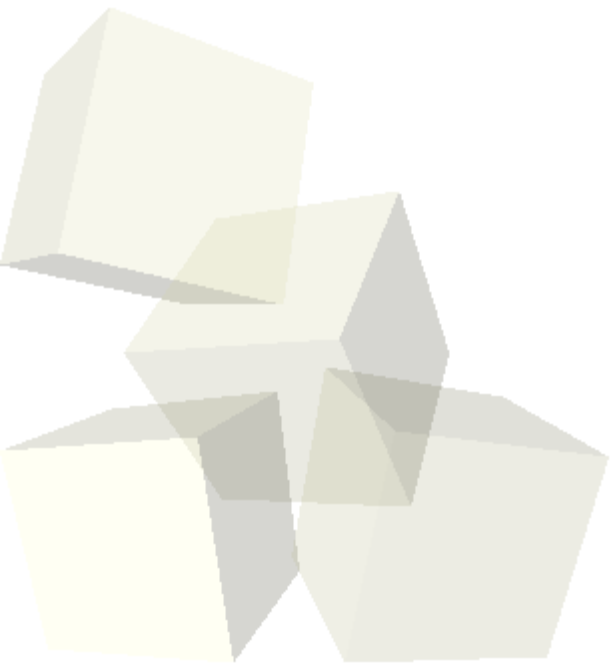
- IBM mainframe to commodity x86
 - ◆ Co-evolution of hardware and software
- x86 AMD/Intel
 - ◆ AMD-V, Nested Page Tables (NPT)
 - ◆ VT-x, VT-d, Extended Page Tables (EPT)
- Processor
 - ◆ AMD-V, VT-x
- Memory
 - ◆ Input/Output Memory Management Unit (IOMMU)
 - ◆ NPT, EPT
- I/O
 - ◆ Smart devices
 - Virtualization-aware NICs, graphics cards, etc.

Xen Architecture



Xen Hypervisor Role

- Thin, privileged abstraction layer
- Provides generic classes of devices
 - ◆ CPU, memory, disk, network, etc.
- Scheduling

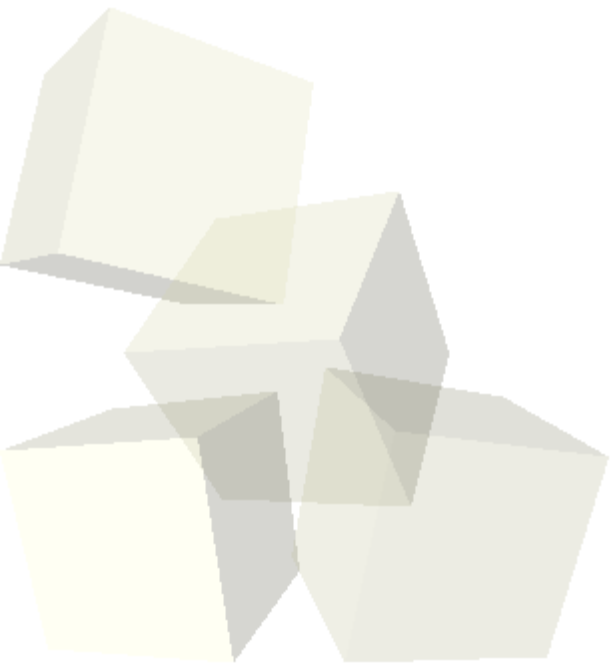


Domain0 Role

- Creates and manages guest VMs
- Interacts with the Xen hypervisor
 - ◆ Xend (Xen daemon)
 - ◆ xm commands
 - ◆ Xenstore
 - ◆ xenconsole
 - Abstraction for guest ttys
- Supplies device and I/O services
 - ◆ Runs backend drivers
 - ◆ Provides guest storage

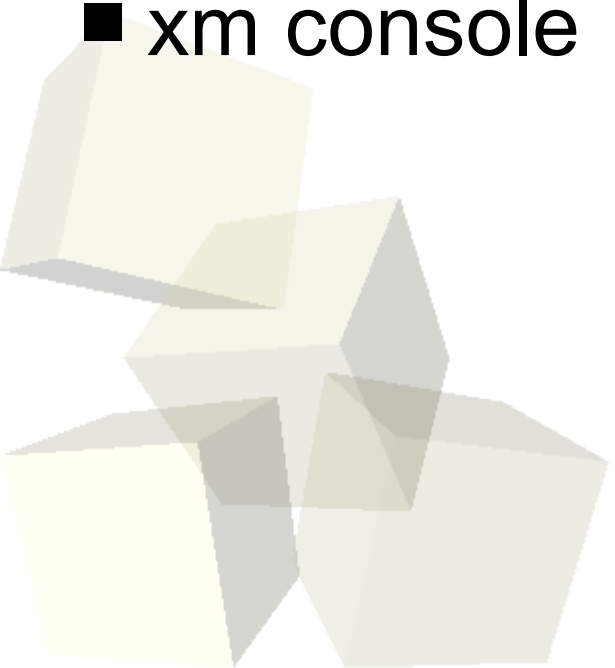
Xenstore Overview

- Database of configuration information
- Used by Domain0 to access guest state
- Guest domain drivers can write to xenstore
- Reference:
 - ◆ <http://wiki.xensource.com/xenwiki/XenStoreReference>

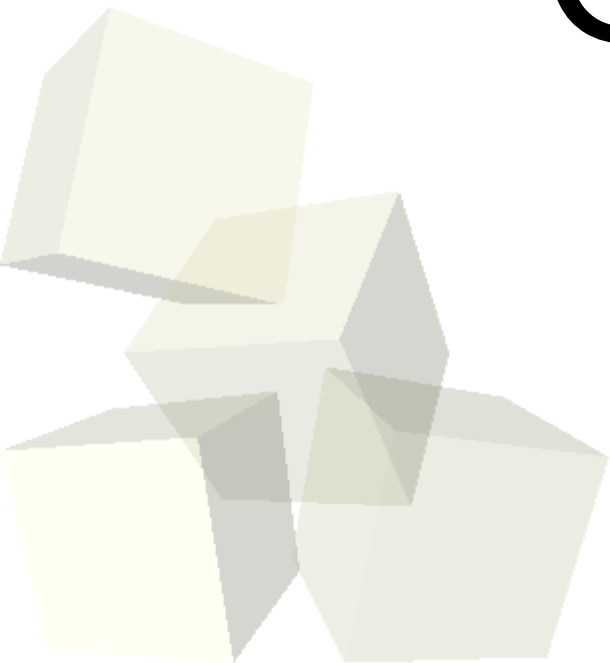


Xen LiveCD Demo

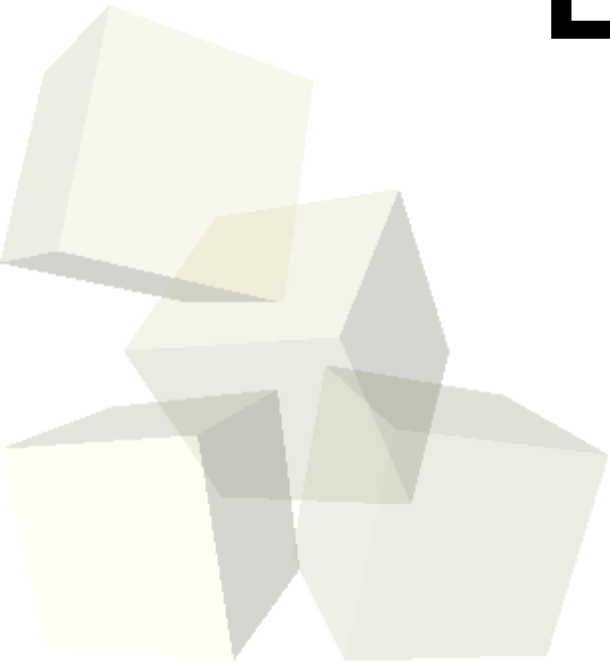
- xm list
- xm create
- xm shutdown
- xm reboot
- xm destroy
- xm pause
- xm unpause
- xm console



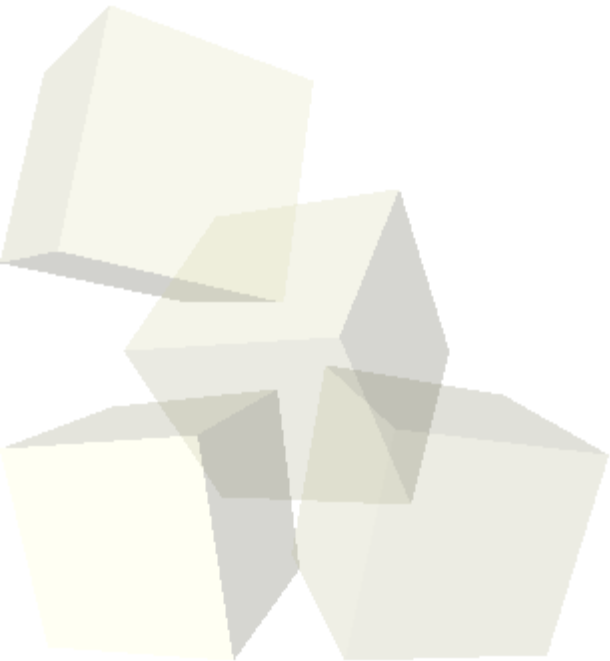
Questions?



Break Time

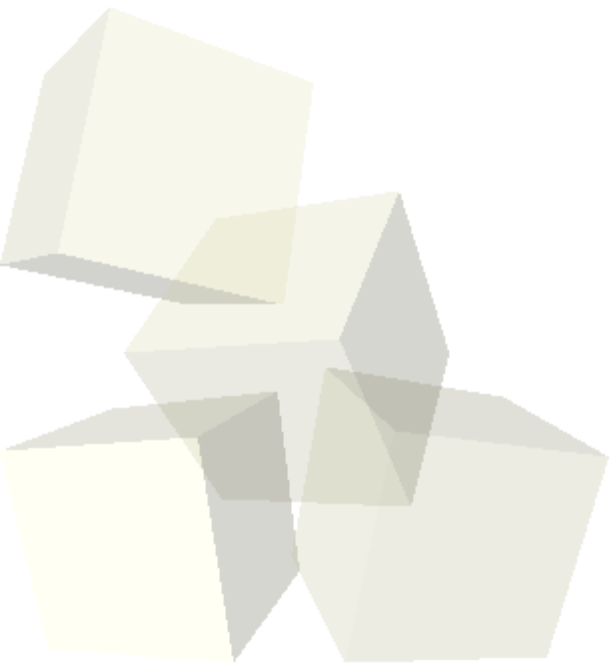


Unit 2



Unit 2: Installing, Configuring, Basic Use

- Hands-on Installation Demos
- Guest Installation
- Distro-specific Guest Installation Tools
- Interacting with Guests



Hands-on Installation Demos

■ Build from source

◆ Mercurial overview

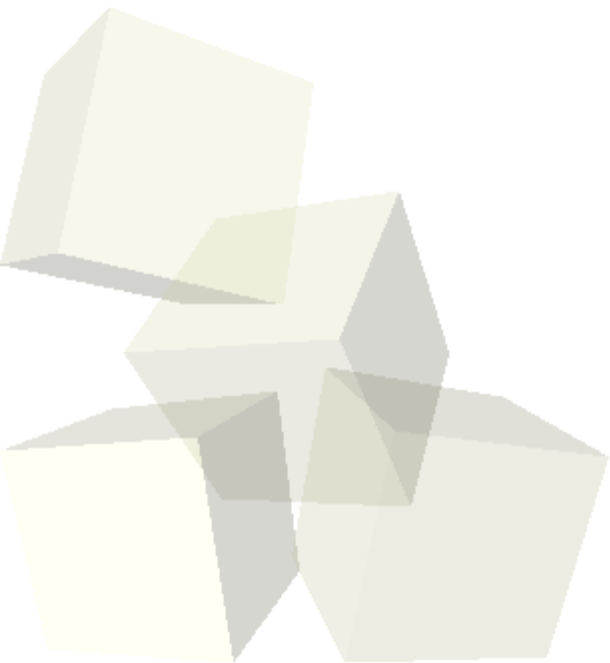
- Peer to peer version control
- hg serve, hg pull (-u), hg update, hg commit, etc.
- Supports tags, hg fetch extension in .hgrc
- As a user, hg clone is usually enough
- Example:
 - hg clone <http://xenbits.xensource.com/xen-unstable.hg>
 - hg clone <http://xenbits.xensource.com/linux-2.6.18-xen.hg>
 - cd xen-unstable.hg
 - make world
 - make install

◆ Source tarball of stable releases also available

- Distribution-supported packages preferred
 - Integrates better with distribution
 - Security and bug fixes from distribution maintainers

Open Source Distributions

- CentOS/Fedora
- Note: No Domain0 kernel in Fedora 9



During Installation: Customize Now

CentOS 5



The default installation of CentOS includes a set of software applicable for general internet usage. What additional tasks would you like your system to include support for?

- Desktop - GNOME
- Desktop - KDE
- Server
- Server - GUI

Please select any additional repositories that you want to use for software installation.

[+ Add additional software repositories](#)

You can further customize the software selection now, or after install via the software management application.

Customize later Customize now

[Release Notes](#)

[← Back](#)

[Next →](#)

■ Add Virtualization Group



CentOS 5

Virtualization Support.

3 of 3 optional packages selected

Optional packages

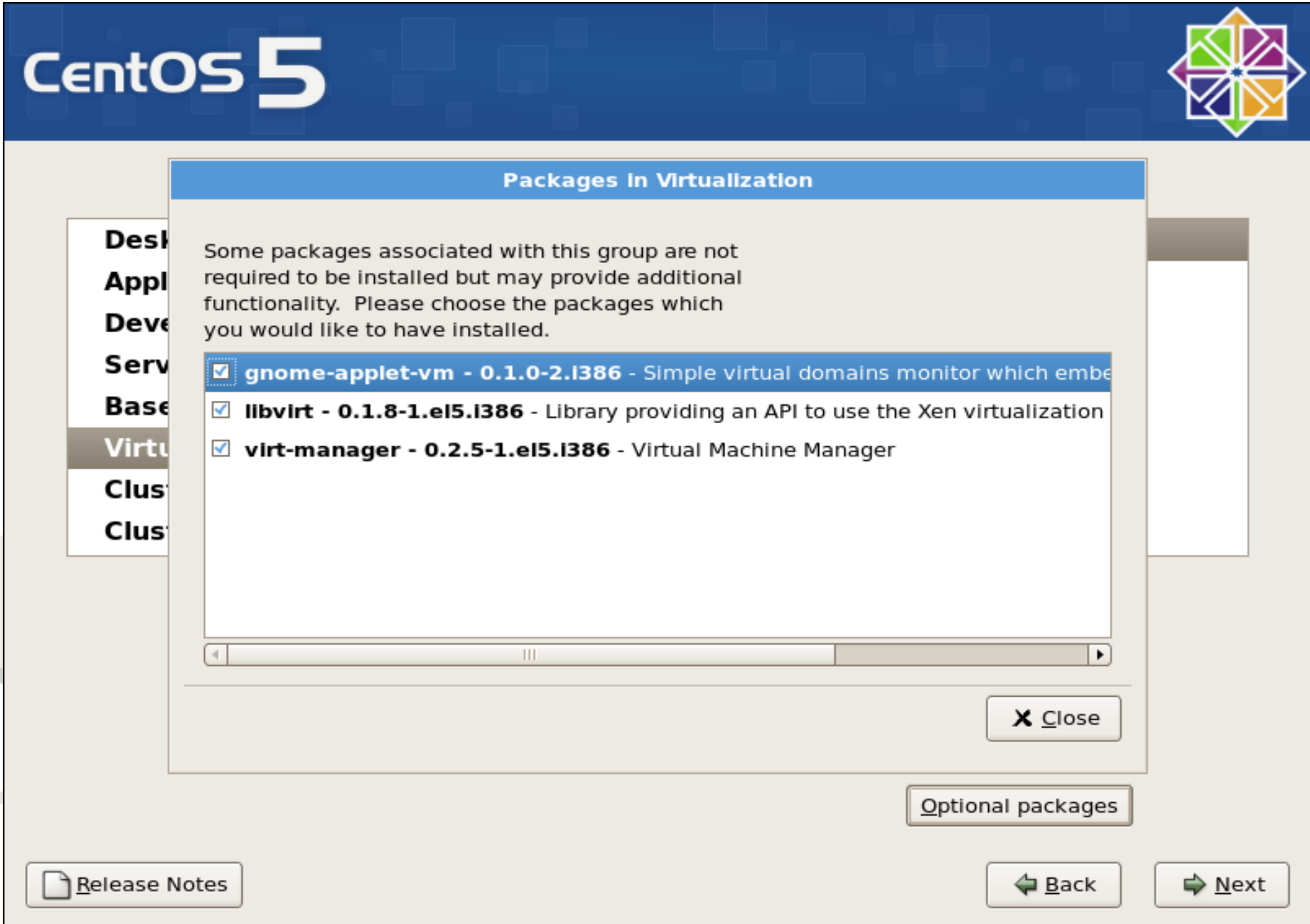
Release Notes

Back Next

■ 23

The image shows the CentOS 5 installation interface. At the top left is the 'CentOS 5' logo. At the top right is a colorful geometric logo. On the left side, there is a vertical list of categories: Desktop Environments, Applications, Development, Servers, Base System, Virtualization (highlighted), Clustering, and Cluster Storage. To the right of this list is a larger window titled 'Virtualization' with a checked checkbox. Below these is a text area containing 'Virtualization Support.' and a status message '3 of 3 optional packages selected'. At the bottom right, there is a button labeled 'Optional packages'. At the bottom left, there is a button labeled 'Release Notes'. At the bottom center, there are two buttons: 'Back' and 'Next'.

Optional Virtualization Packages



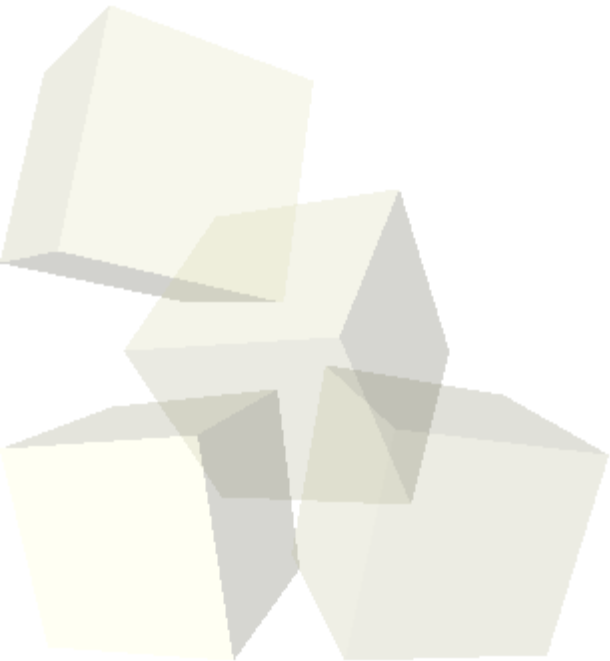
The image shows a CentOS 5 installation window titled "Optional Virtualization Packages". The window has a blue header with the CentOS 5 logo and a colorful geometric icon. The main content area contains a list of packages with checkboxes. The packages listed are:

- gnome-applet-vm - 0.1.0-2.i386** - Simple virtual domains monitor which embe
- libvirt - 0.1.8-1.el5.i386** - Library providing an API to use the Xen virtualization
- virt-manager - 0.2.5-1.el5.i386** - Virtual Machine Manager

At the bottom of the window, there are buttons for "Close", "Optional packages", "Release Notes", "Back", and "Next".

Open Source Distributions

- OpenSUSE





During Installation: Software Selection

This dialog allows you to define this system's tasks and what software to install. Available tasks and software for this system are shown by category in the left column. To view a description for an item, select it in the list.

Change the status of an item by clicking its status icon or right-click any icon for a context menu. With the context menu, you can also change the status of all items.

Details opens the detailed software package selection where you can view and select individual software packages.

The disk usage display in the lower right corner shows the remaining disk space after all requested changes will have been performed. Hard disk partitions that are full or nearly full can degrade system performance and in some cases even cause serious problems. The system needs some available disk space to run properly.

Software Selection and System Tasks

- Base Technologies**
 - openSUSE Base System
 - Novell AppArmor
 - Console Tools
 - Laptop
 - YaST System Administration
 - openSUSE Software Management
 - Enterprise Software Management (Z...)
- Graphical Environments**
 - GNOME Desktop Environment
 - GNOME Base System
 - KDE Desktop Environment
 - KDE Base System
 - X Window System
 - Fonts
- Desktop Functions**
 - Desktop Effects
 - Graphics
 - Games
 - Remote Desktop
 - Voice Over IP Clients
 - XML and LaTeX Editing Tools
- Server Functions**
 - File Server
 - Print Server
 - Network Administration
 - Misc. Server
 - Voice Over IP Server
 - Mail and News Server
 - Web and LAMP Server
 - Internet Gateway
 - DHCP and DNS Server
 - Directory Server (LDAP)
 - Xen Virtual Machine Host Server

openSUSE Base System

This is the base openSUSE runtime system.

Name	Disk Usage	Used	Free	Total
/	<div style="width: 64%; background-color: green; border: 1px solid black;"></div> 64%	2.2 GB	1.3 GB	3.5 GB

Xen Software Packages

YaST2@opensuse-xen

File Package Extras Help

Filter: Patterns

Pattern

- Graphics
- Games
- Remote Desktop
- XML and LaTeX Editing Tools
- Voice Over IP Clients
- Server Functions**
- File Server
- Network Administration
- Print Server
- Misc. Server
- Voice Over IP Server
- Mail and News Server
- Web and LAMP Server
- Internet Gateway
- DHCP and DNS Server
- Directory Server (LDAP)
- Xen Virtual Machine Host Server
- Development**
- Basis Development
- KDE Development

Package	Summary	Size
<input checked="" type="checkbox"/> bridge-utils	Utilities for Configuring the Linux Ethernet Bridge	80.5 K
<input checked="" type="checkbox"/> xen	Xen Virtualization: Hypervisor (aka VMM aka Microkernel)	13.3 M
<input checked="" type="checkbox"/> xen-doc-html	Xen Virtualization: HTML documentation	371.5 K
<input checked="" type="checkbox"/> xen-doc-pdf	Xen Virtualization: PDF documentation	558.6 K
<input checked="" type="checkbox"/> xen-libs	Xen Virtualization: Libraries	122.5 K
<input checked="" type="checkbox"/> xen-tools	Xen Virtualization: Control tools for domain 0	5.2 M
<input checked="" type="checkbox"/> xen-tools-ioemu	Xen Virtualization: BIOS and device emulation for unmodified guests	711.4 K
<input checked="" type="checkbox"/> xterm	The basic X terminal program	868.9 K
<input checked="" type="checkbox"/> yast2-vm	YaST2 Virtual Machine Installer	685.6 K

Description Technical Data Dependencies Versions File List Change Log


xen - Xen Virtualization: Hypervisor (aka VMM aka Microkernel)

Xen is a virtual machine monitor for x86 that supports execution of multiple guest operating systems with unprecedented levels of performance and resource isolation.

This package contains the Xen Hypervisor. (tm)

Modern computers are sufficiently powerful to use virtualization to present the illusion of many

Check Autocheck Cancel Accept

Name	Disk Usage	Used	Free	Total	
/		29%	2.9 GB	6.8 GB	9.7 GB

Open Source Distributions

■ Debian/Ubuntu

- ◆ Install with apt or synaptic
- ◆ Meta-packages
 - Ubuntu
 - ubuntu-xen-server
 - ubuntu-xen-desktop

■ Gentoo

- ◆ Install with portage
 - May need to unmask the Xen packages

■ OpenSolaris

- ◆ xVM packages available

■ NetBSD

- ◆ Xen 3.X package support as of NetBSD 4.0

Commercial Distributions

- Red Hat Enterprise Linux (RHEL)
- SUSE Linux Enterprise Server (SLES)
- Virtual Iron
- Oracle VM
- Xen Server
 - ◆ Express Edition (demo)
 - ◆ Standard Edition
 - ◆ Enterprise Edition

Guest Installation

- Guest Configuration Files
- PV Guest Config Example
 - ◆ PV Guest Example (demo)
- HVM Guest Config Example
 - ◆ HVM Guest Example (demo)
- Pre-built Guest Images
- Converting VMware Images
- Distribution-specific Guest Installation Tools
- Interacting with Guests

■ Guest Configuration Files

■ Disk I/O and memory configuration parameters

◆ disk

→ Array of 3-tuples

- Device to export
- Device as seen by guest
- Access permission (r, w)

→ SCSI (sd) and IDE (hd) examples:

- disk = ['phy:sda, sda, w', 'phy:/dev/cdrom, cdrom:hdc, r']
- disk = ['tap:aio:hdb1, hdb1, w', 'phy:/dev/LV/disk1, sda1, w']

→ Xen Virtual Block Device (xvd) examples:

- disk = ['phy:sda, xvda, w', 'phy:/dev/cdrom, cdrom:hdc, r']
- disk = ['tap:aio:hdb1, xvdb1, w', 'phy:/dev/LV/disk1, xvda1, w']

◆ memory

→ Integer in megabytes (MB)

→ Example:

- memory = 512

Guest Configuration Files

■ Network configuration parameters

- ◆ vif
 - Array of virtual interface specifications
 - Zero or more name=value entries for each interface
 - Single interface examples
 - vif=[' ']
 - vif=['mac=00:16:3e:51:c2:b1']
 - vif=['mac=00:16:3e:36:a1:e9, bridge=eth0']
 - Multiple interface examples
 - vif=[' ', '']
 - vif= ['mac=00:16:3e:36:a1:e9, bridge=eth1', 'bridge=eth0']
- ◆ dhcp, gateway, hostname, netmask

Guest Configuration Files

- PV guest kernel configuration parameters
 - ◆ kernel
 - Use Domain0 kernel (external to guest)
 - ◆ ramdisk
 - Use Domain0 kernel (external to guest)
 - ◆ bootloader
 - pygrub
 - Security issues
 - Use Xen-compatible kernel and initrd (internal to guest)
 - ◆ root
 - Root file system device
 - ◆ extra
 - Append to kernel command line
 - extra="4" sets runlevel to 4

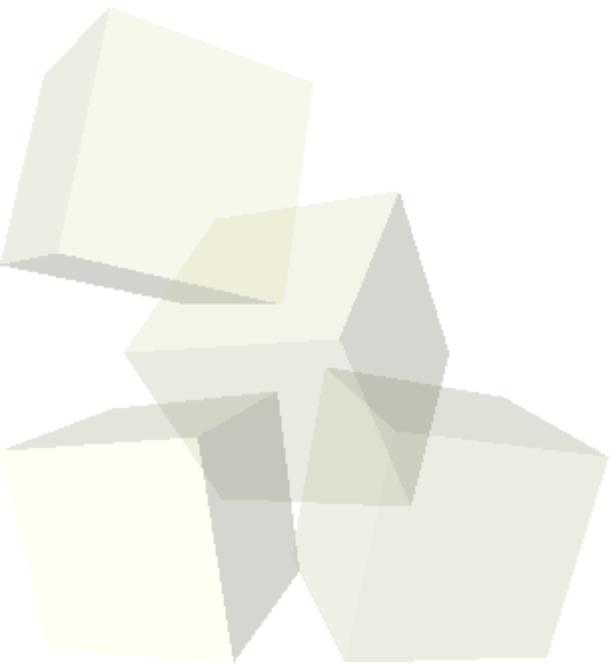
Guest Configuration Files

- HVM guest configuration parameters
 - ◆ kernel
 - Use hvmloader
 - ◆ builder
 - Use hvm
 - ◆ device_model
 - Usually QEMU device model (qemu-dm)
 - ◆ boot
 - Boot order
 - ◆ sdl
 - Simple DirectMedia Layer
 - Built-in to Xen virtual frame buffer
 - ◆ vnc
 - Virtual Network Computing

Guest Configuration Files

■ Deprecated parameters

- cdrom
 - Replaced by disk=['....cdrom:hdX...']
- file:/ in disk=['file:/...']
 - Replaced with disk=['tap:aio:/...']
- nics
 - Better to simply use the vif array

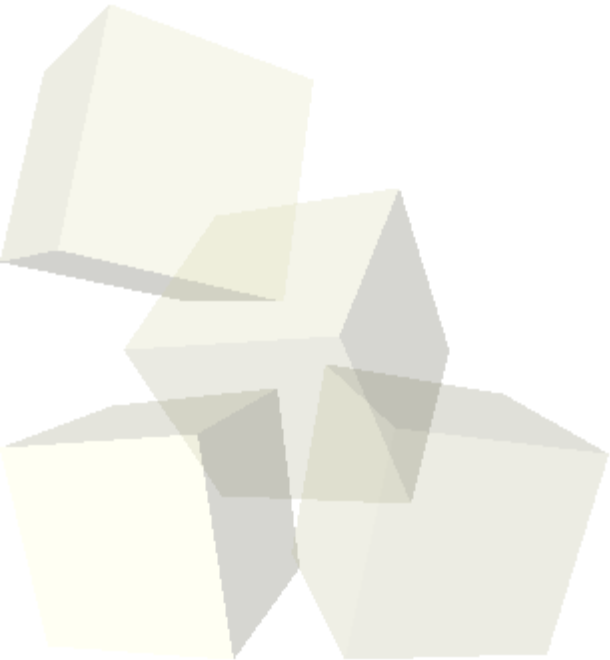


PV Guest Config Example

```
kernel="/boot/vmlinuz-2.6.24-16-xen"  
ramdisk="/boot/initrd.img-2.6.24-16-xen"  
disk=['tap:aio:/xen/para.partition,xvda1,w']  
memory=512  
vif=[' ']  
root="/dev/xvda1"
```

PV Guest Example

- Demo



HVM Guest Config Example

HVM guest example

```
kernel="/usr/lib64/xen/boot/hvmloader"
```

```
builder="hvm"
```

```
device_model = "/usr/lib64/xen/bin/qemu-dm"
```

```
disk=['phy:/xen/images/hvm.disk,hda,w',  
      'phy:/dev/cdrom,hdc:cdrom,r']
```

```
sdl=1
```

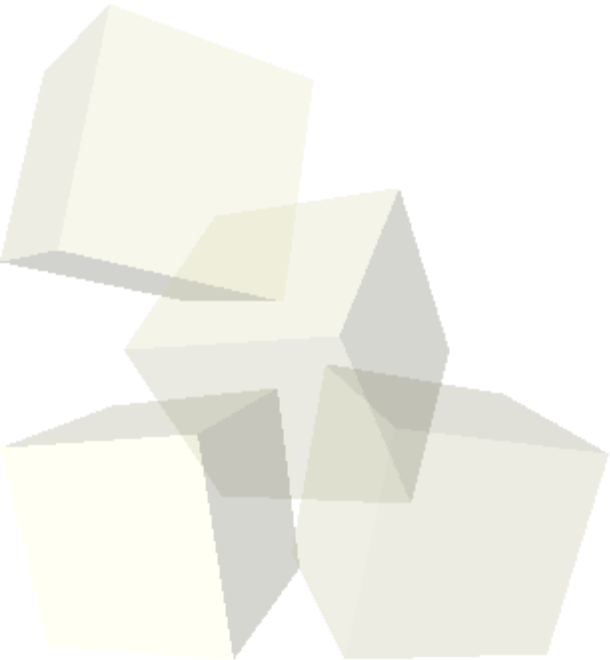
```
boot="dc"
```

```
memory=512
```

```
vif=['type=ioemu,bridge=eth0']
```

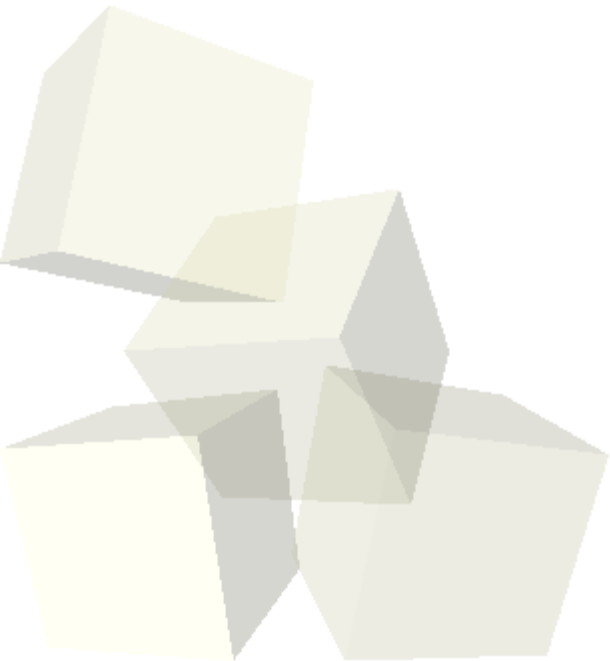
HVM Guest Example

- Demo



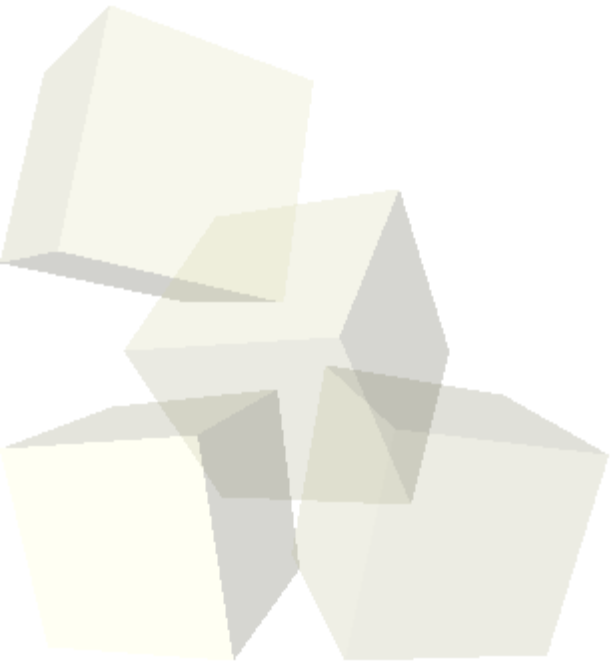
Pre-built Guest Images

- rpath.com
- jailtime.org
- virtualappliances.net
- jumpbox.com



Converting VMware Images

- qemu-img convert (demo)

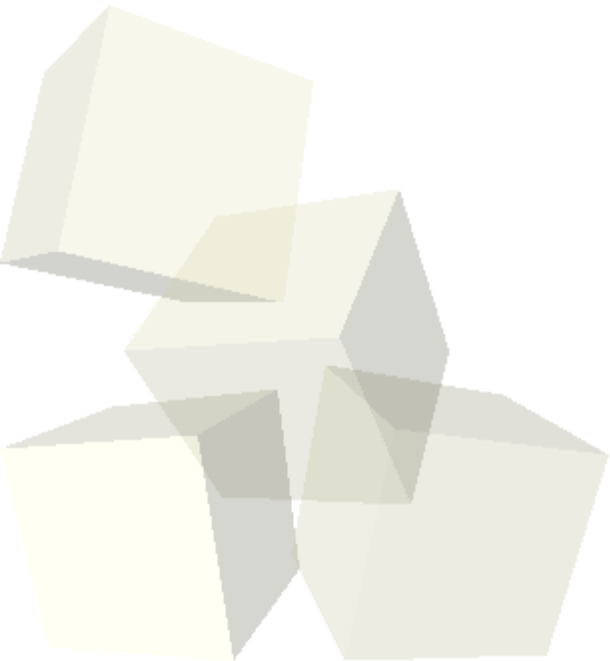


Distribution-specific Tools

- CentOS/Fedora
 - ◆ Virtual Machine Manager (virt-manager) and rpmstrap
- OpenSUSE
 - ◆ Yast Virtual Machine Management
- Debian/Ubuntu
 - ◆ debootstrap and xen-tools
- Gentoo
 - ◆ quickpkg and domi

Distribution-specific Tools

- Virtual Machine Manager example

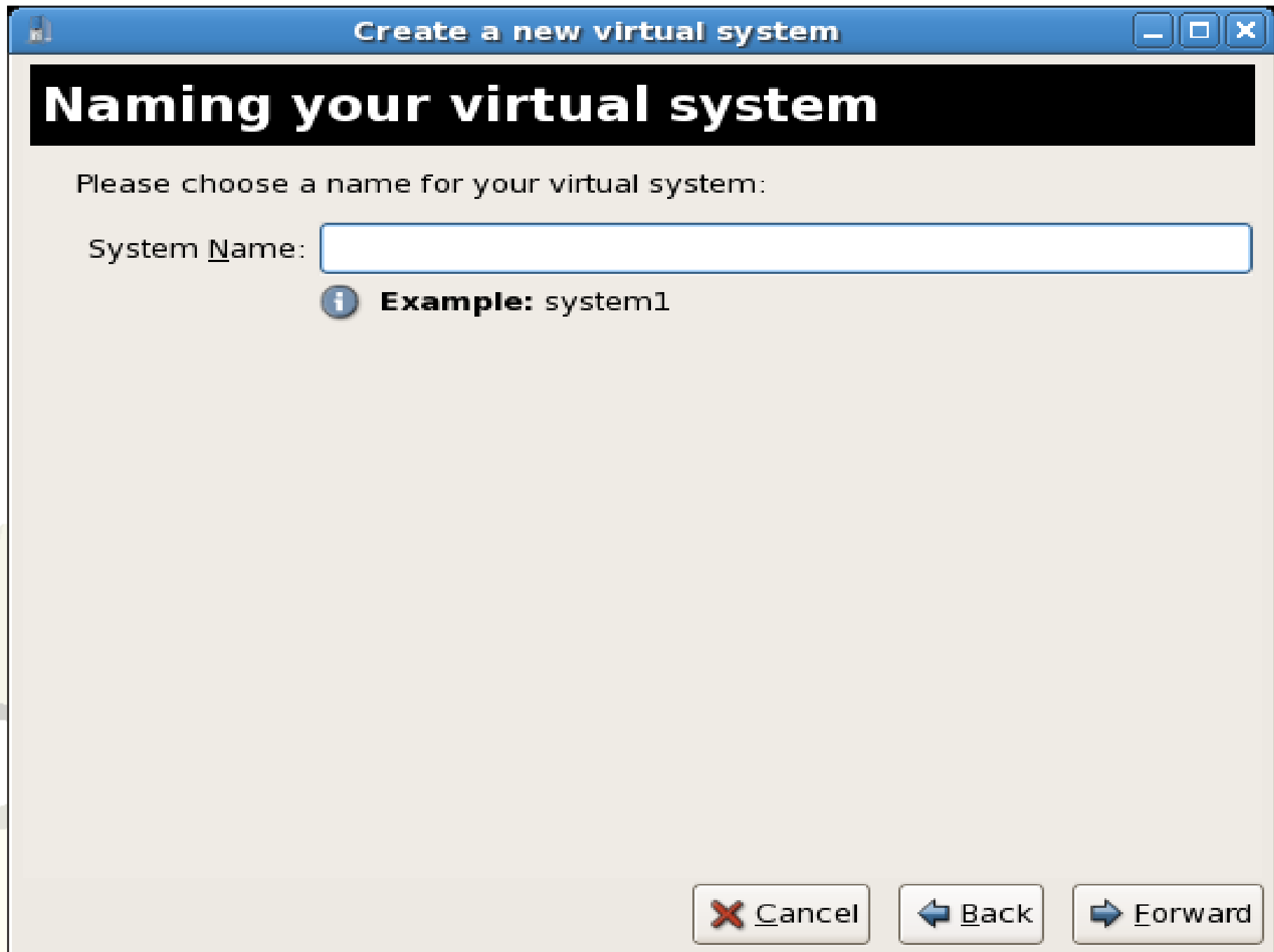


Main Window – Create New

The screenshot shows the main window of a Virtual Machine Manager. The window title is "Virtual Machine Manager". The menu bar includes "File", "Edit", "View", and "Help". A "View:" dropdown menu is set to "All virtual machines". Below the menu bar is a table with the following columns: "ID", "Name", "Status", "CPU usage", and "Memory usage". The table contains one entry: "0", "Domain-0", "Running", "5.57 %", and "937.32 MB (91.89%)". At the bottom of the window, there are four buttons: "Delete", "New", "Details", and "Open".

ID	Name	Status	CPU usage	Memory usage
0	Domain-0	Running	5.57 %	937.32 MB (91.89%)

Name the Guest VM





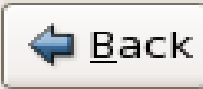

Create a new virtual system

Naming your virtual system

Please choose a name for your virtual system:

System Name:

 **Example:** system1


PV Installation Media Choices

Create a new virtual system

Locating installation media


Please indicate where installation media is available for the operating system you would like to install on this **paravirtualized** virtual system. Optionally you can provide the URL for a kickstart file that describes your system:

Install Media URL: ▼

 **Example:** `http://servername.example.com/distro/i386/tree`

Kickstart URL: ▼

 **Example:** `ftp://hostname.example.com/ks/ks.cfg`

 Help  Cancel  Back  Forward

HVM Installation Media Choices

Create a new virtual system

Locating installation media

Please indicate where installation media is available for the operating system you would like to install on this **fully virtualized** virtual system:

ISO Image Location:

ISO Location:

CD-ROM or DVD:

Path to install media:

Please choose the type of guest operating system you will be installing:

OS Type:

OS Variant:

Choose Disk Storage

Create a new virtual system

Assigning storage space

Please indicate how you'd like to assign space on this physical host system for your new virtual system. This space will be used to install the virtual system's operating system.

Normal Disk Partition:

Partition:

i **Example:** /dev/hdc2

Simple File:

File Location:

File Size: MB

i **Note:** File size parameter is only relevant for new files

i **Tip:** You may add additional storage, including network-mounted storage, to your virtual system after it has been created using the same tools you would on a physical system.

Memory and CPU Settings

Create a new virtual system

Allocate memory and CPU

Memory:

Please enter the memory configuration for this VM. You can specify the maximum amount of memory the VM should be able to use, and optionally a lower amount to grab on startup.

Total memory on host machine: 1020 GB

VM Max Memory:

VM Startup Memory:

CPUs:

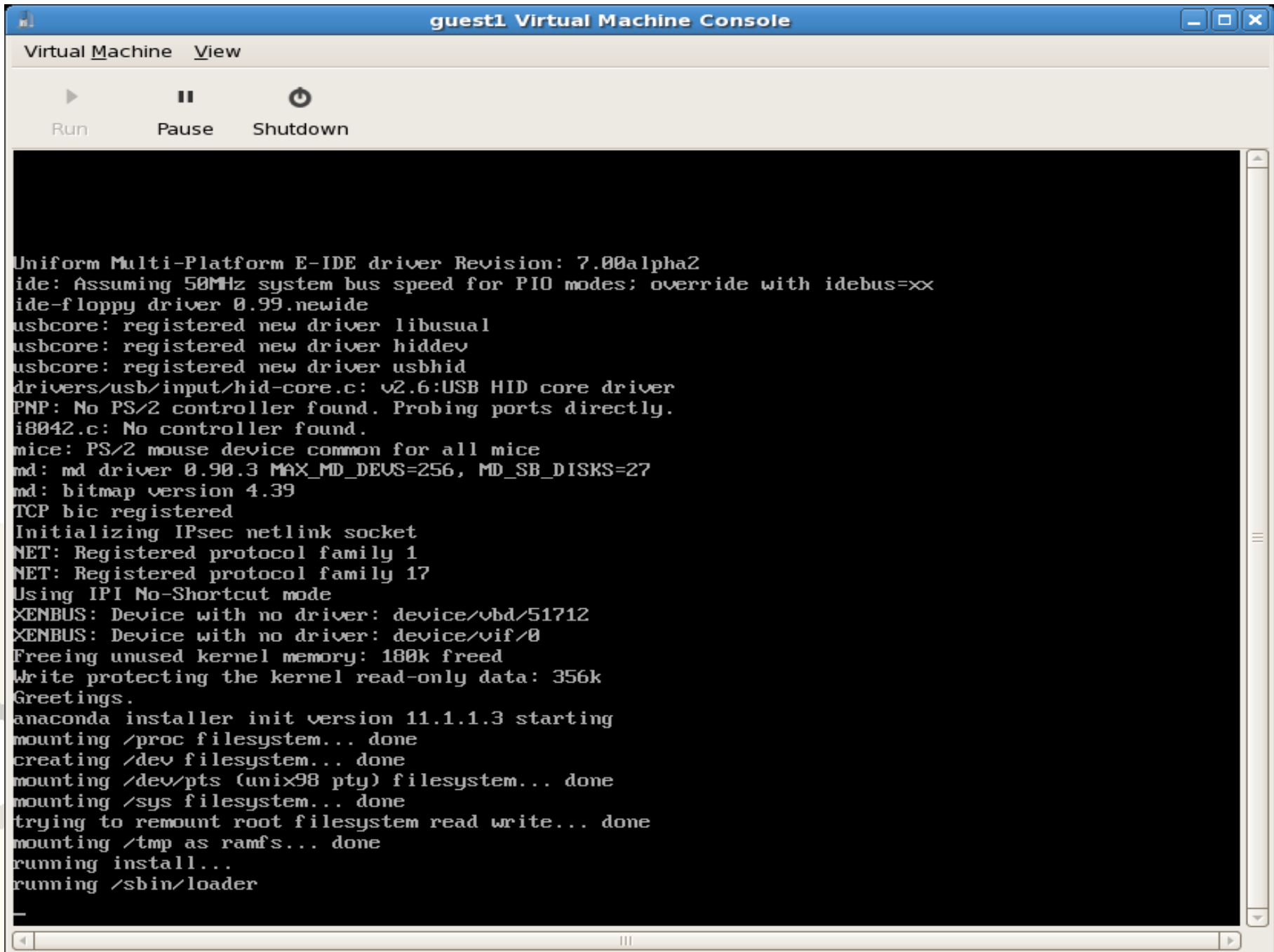
Please enter the number of virtual CPUs this VM should start up with.

Physical host CPUs: 2

VCPUs:

i Tip: For best performance, the number of virtual CPUs should be less than (or equal to) the number of physical CPUs on the host system.

Begin the Installation

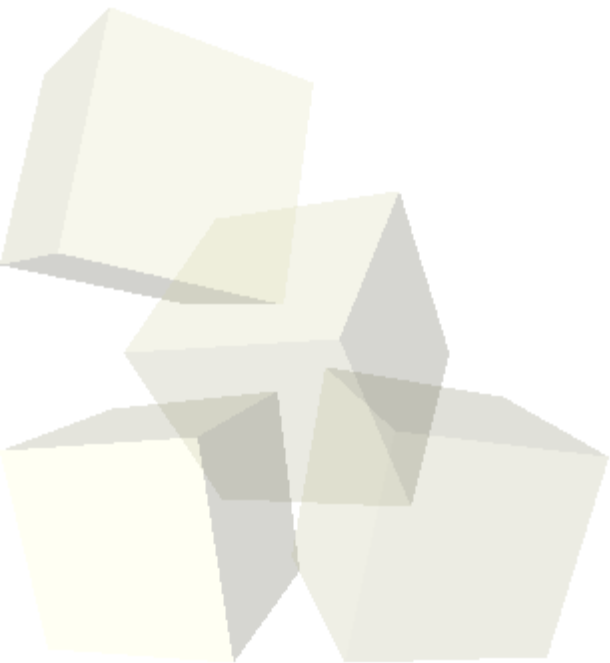


```
Virtual Machine View
Run Pause Shutdown

Uniform Multi-Platform E-IDE driver Revision: 7.00alpha2
ide: Assuming 50MHz system bus speed for PIO modes; override with idebus=xx
ide-floppy driver 0.99.newide
usbcore: registered new driver libusual
usbcore: registered new driver hiddev
usbcore: registered new driver usbhid
drivers/usb/input/hid-core.c: v2.6:USB HID core driver
PNP: No PS/2 controller found. Probing ports directly.
i8042.c: No controller found.
mice: PS/2 mouse device common for all mice
md: md driver 0.90.3 MAX_MD_DEVS=256, MD_SB_DISKS=27
md: bitmap version 4.39
TCP bic registered
Initializing IPsec netlink socket
NET: Registered protocol family 1
NET: Registered protocol family 17
Using IPI No-Shortcut mode
XENBUS: Device with no driver: device/vbd/51712
XENBUS: Device with no driver: device/vif/0
Freeing unused kernel memory: 180k freed
Write protecting the kernel read-only data: 356k
Greetings.
anaconda installer init version 11.1.1.3 starting
mounting /proc filesystem... done
creating /dev filesystem... done
mounting /dev/pts (unix98 pty) filesystem... done
mounting /sys filesystem... done
trying to remount root filesystem read write... done
mounting /tmp as ramfs... done
running install...
running /sbin/loader
```

Distribution-specific Tools

- Yast Virtual Machine Management example



Main Window – Add Guest VM

Manage Virtual Machines

A virtual machine (VM) is a defined instance of virtual hardware, such as CPU, memory, network card, and block devices, and the operating system that runs on it.

The number of VMs you can create depends on the requirements for each VM and the available hardware resources.

For the most current information on Novell VM server technology, see www.novell.com/documenta

Name	Virtualization Mode	Status	Memory (MB)	Console	
------	---------------------	--------	-------------	---------	--

Manage Virtual Machines

Add Refresh Delete

Start View Shutdown Terminate

Close

Guest Installation Choices

Create a Virtual Machine

Creating a VM requires that you install the VM's operating system by either running an OS installation program or specifying a disk image that already contains an operating system.

Run an OS Installation Program: You can install a VM's operating system by running an OS installation program from a YaST Network Installation Source, a CD / DVD device, or an ISO image file.

Use a Disk Image: You can specify that the VM boots an already-installed operating system from boot files located on a disk image or a physical disk.

Create a Virtual Machine

Method for Installing the VM's Operating System

- Run an OS installation program
- Use a disk image or a physical disk that contains OS boot files

Back Abort Next

Start Installation

Virtual Machine (Installation Settings)

These settings define an initial VM environment to be used for the installation of the VM's operating system.

Clicking Next launches the OS installation program in a separate window. Follow the on-screen instructions to install the OS.

After completing the OS installation program, you will be prompted to finalize the VM settings.

Virtual Machine (Installation Settings)

Click any headline to make changes or use the "Change..." menu below.

AutoYaST

- ◆ AutoYaST Profile: *none*

Virtualization Mode

- ◆ Paravirtualization

VM Properties

- ◆ Name of Virtual Machine: vm1

Hardware

- ◆ Memory Size: 256 MB
- ◆ Number of Virtual CPUs: 1
- ◆ Hardware Clock: UTC

Disks

- ◆ Disk hda: Create 4096 MB Image (Sparse File)

Network

- ◆ Number of Virtual Network Cards: 1

Operating System Installation

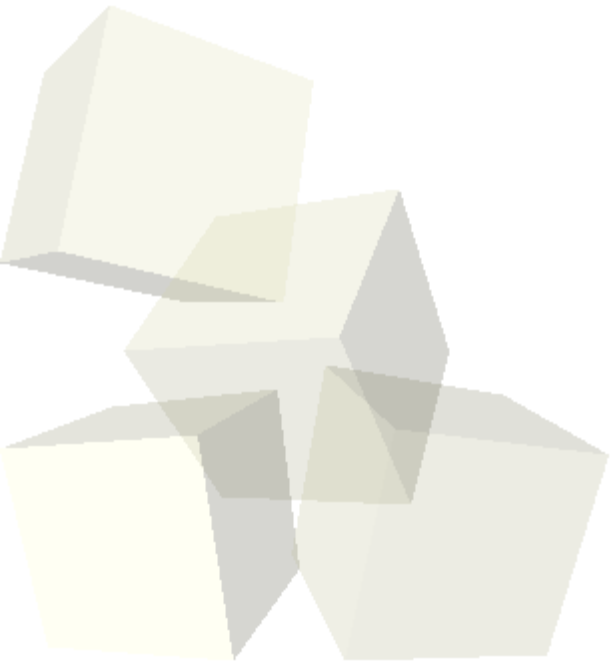
- ◆ SUSE Installation Source: unknown (<http://download.opensuse.org/distribution/10.2/repo/oss/>)
- ◆ Installation Options: TERM=xterm textmode=1 vnc=0

Change...

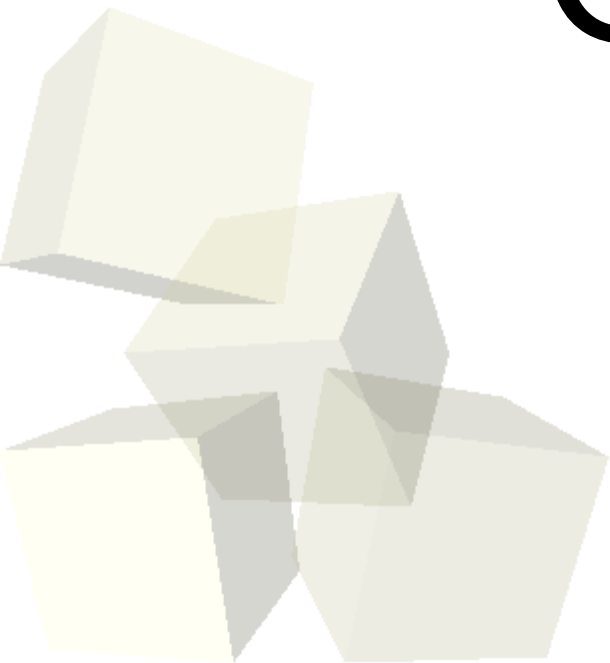
Back **Abort** **Next**

Interacting with Guests

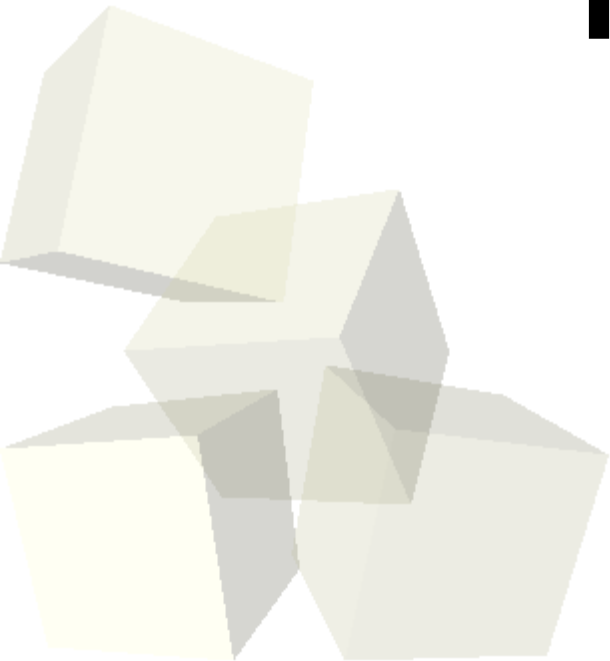
- SSH
- VNC
- FreeNX
- Xen console
- Xen virtual frame buffer (sdl)
- Remote Desktop (demo)



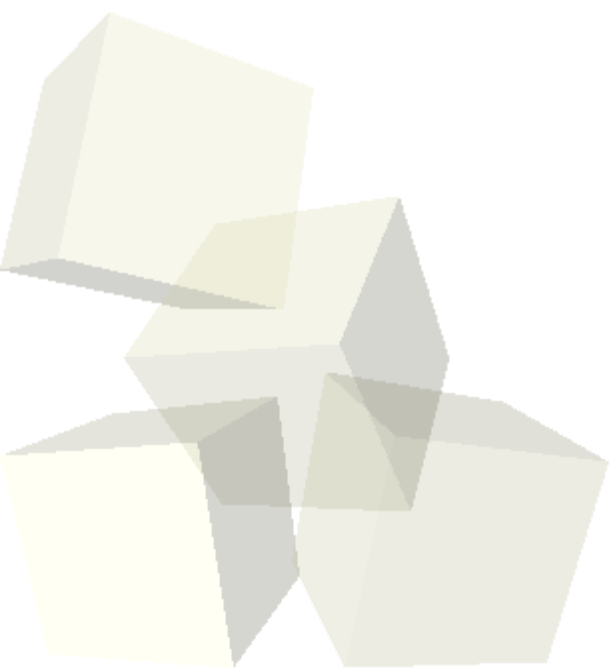
Questions?



Food Time

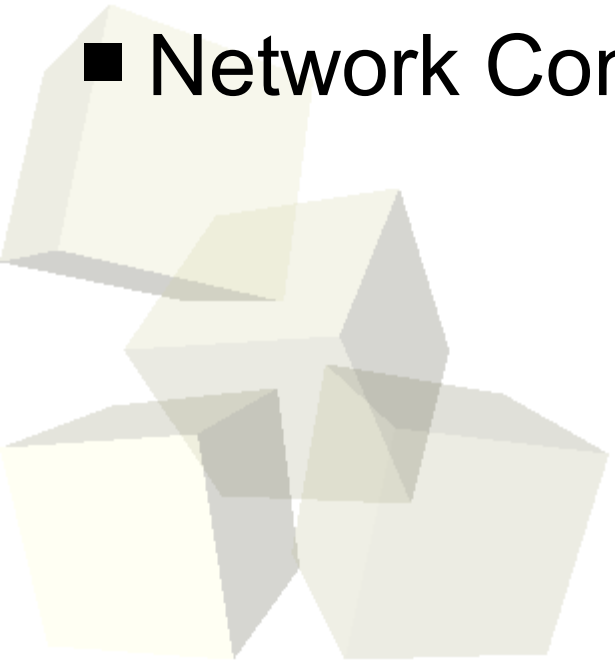


Unit 3





- Quick LVM Refresher
- Network Storage
- Guest Image Files from Scratch
- Guest Save, Restore, and Live Migration
- Xen Device Models
- Network Configurations

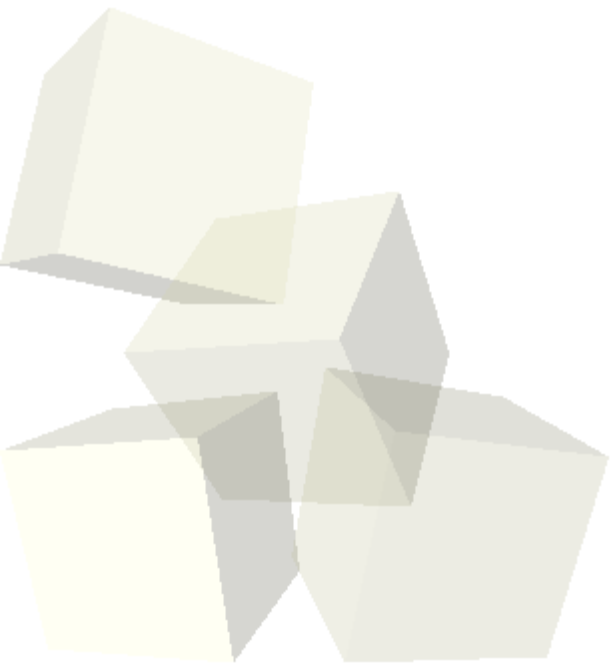


Quick LVM Refresher

- Logical Volume Manager (LVM)
 - ◆ Provides abstraction above block devices
 - ◆ Allows the system administrator to:
 - Span logical volumes across physical volumes
 - Grow and shrink logical volumes
 - Use various RAID levels
 - Use copy-on-write
- Why LVM?
 - ◆ Flexible
 - Resize partitions (guest images)
 - Density of logical volumes (guests)
 - ◆ Feature-rich
 - Copy-on-write snapshots
 - Backup and recovery

Quick LVM Refresher

- Key terms
 - ◆ **physical volume (PV)**
 - ◆ **volume group (VG)**
 - ◆ **logical volume (LV)**



Quick LVM Refresher

■ LVM commands

- ♦ pvcreate – initializes a physical volume for LVM use
 - # pvcreate /dev/sda3 /dev/sda4 /dev/hda1
- ♦ vgcreate – creates a new volume group
 - # vgcreate xen_vg /dev/sda3 /dev/sda4 /dev/hda1
- ♦ lvcreate – creates a new logical volume
 - # lvcreate -L 4G -n guest_partition xen_vg
- ♦ lvextend – grows a logical volume
 - # lvextend -L 5G /dev/xen_vg/guest_partition
- ♦ lvreduce – shrinks a logical volume
 - # lvreduce -L 3G /dev/xen_vg/guest_partition

Quick LVM Refresher

■ LVM caveats

- ◆ When extending a logical volume
 - Resize the underlying file system after the lvextend
- ◆ When reducing a logical volume
 - Resize the underlying file system before the lvreduce
- ◆ Underlying file system support
 - Resize larger or smaller
 - Resize online (while the file system is mounted)

■ File system-specific commands

- ◆ ext2/3
 - resize2fs
 - e2fsck
 - tune2fs

Network Storage

- ATA over Ethernet (AoE)
 - ◆ Exports block devices over the network
 - ◆ Lightweight Ethernet layer protocol
 - ◆ No built-in security
- Internet Small Computer System Interface (iSCSI)
 - ◆ Exports block devices over the network
 - ◆ Scales with network bandwidth
 - ◆ Network layer protocol
 - ◆ Client and user-level security
- Network File System (NFS)
 - ◆ Exports file system over the network
 - ◆ Well-known and widely used
 - ◆ Network layer protocol
 - ◆ Known performance issues as root file system

Network Storage

- Network Block Device (NBD)
 - ◆ Exports block device over the network
 - ◆ Scales with network bandwidth
 - ◆ Network layer protocol
 - ◆ Not recommended as root file system
- Cluster file systems
 - ◆ Advantages of block devices and file servers
 - ◆ More difficult to setup and configure
 - ◆ Examples:
 - Global Network Block Device (GNBD)
 - Distributed Replicated Block Device (DRBD)

Guest Image Files from Scratch

- Compressed tar image files
 - ◆ Smallest guest image file
 - ◆ Best for backup and sharing
 - ◆ More difficult to setup guest
 - Need existing place to extract to
- Disk image files
 - ◆ Single file containing root and swap partitions
 - ◆ Largest guest image file
 - ◆ More commands needed
- Partition image files
 - ◆ Separate files for root and swap partitions
 - ◆ Easiest to work with (standard commands)
 - ◆ Slightly more image files to work with (separate swap)

Compressed Tar Image Files

■ Creation

- ◆ `tar -czpf /linux-root.tgz \`
`--exclude /proc \`
`--exclude /linux-root.tgz \`
`/`

■ Using

- ◆ Setup a partition, local volume, disk or partition image
- ◆ Extract with tar command
- ◆ Customize partition setup
 - `/etc/fstab`
 - Kernel command line
- ◆ Configure network-specific details
 - `/etc/hosts` and `/etc/hostname`
 - IP address setup

Disk Image Files

■ Creation

- ◆ Use dd to make a sparse or pre-allocated file
- ◆ Partition the image (fdisk)
- ◆ Make partitions available to the system (kpartx)
- ◆ Format the file systems (mkfs/mkswap)
- ◆ Populate root file system

■ Using

- ◆ If no customization is needed, use directly
- ◆ If customization is needed
 - Associate disk image with block device (losetup)
 - Make partitions available to system (kpartx)
 - Mount partitions in /dev/mapper (mount)
 - Do customizations
 - Unmount (umount), release partitions (kpartx -d), un-associate with block device (losetup -d)

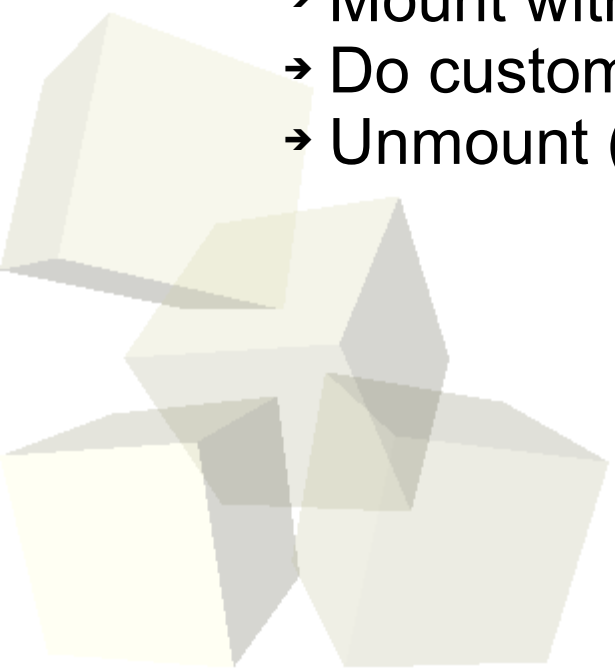
Partition Image Files

■ Creation

- ◆ Use dd to make a sparse or pre-allocated file
- ◆ Format each file system (mkfs/mkswap)
- ◆ Populate file system (if root partition)

■ Using

- ◆ If no customization is needed, use directly
- ◆ If customization is needed
 - Mount with loop option (mount -o loop)
 - Do customizations
 - Unmount (umount)



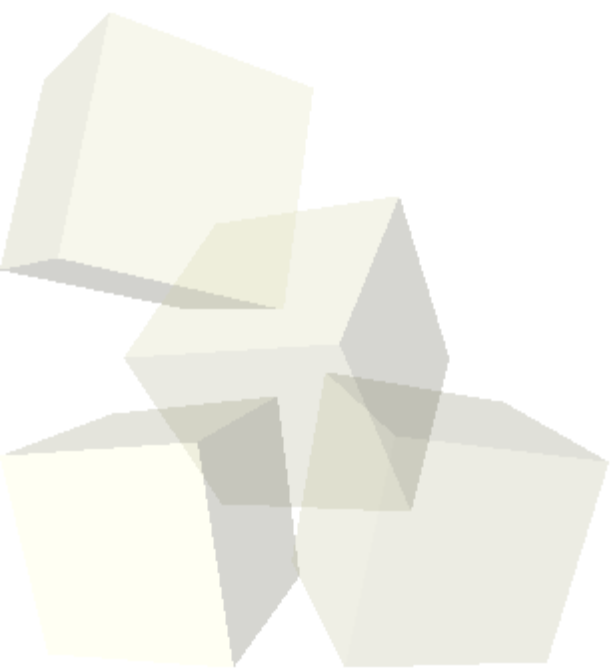


Guest Save, Restore, and LIVE

Migration

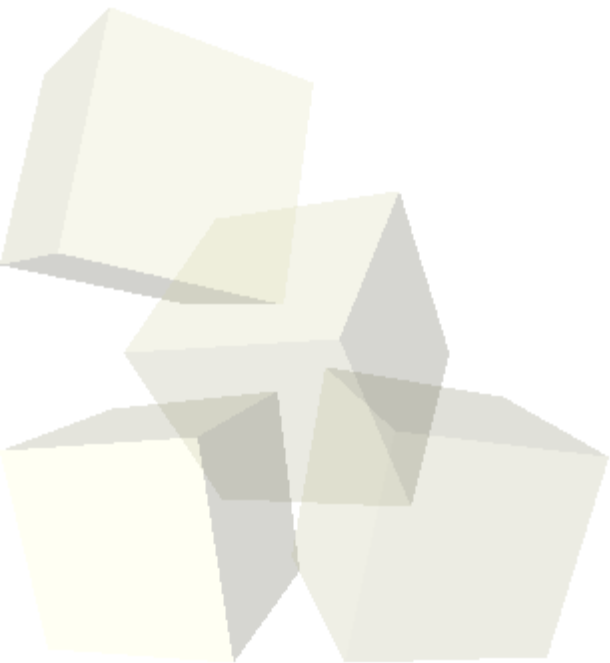
- Guest save and restore
 - ◆ Similar to hibernate and resume of physical PC
 - ◆ Built-in Xen functionality and xm commands

- Guest relocation and migration
 - ◆ Cold static relocation
 - ◆ Warm static (regular) migration
 - ◆ Live migration



Guest Save and Restore

- Save and restore commands
 - ♦ xm save
 - Pauses and suspends (hibernates) a guest
 - Saves guest state to a file on disk
 - ♦ xm restore
 - Restores a guest's state from disk
 - Resumes execution of guest



Guest Relocation

■ Cold static relocation

- ◆ Image and config files need to be manually copied from source to target Domain0

■ Benefits

- ◆ Hardware maintenance with less downtime
- ◆ Backup of guest images
- ◆ Shared storage not required

■ Limitations

- ◆ More manual process
- ◆ Guests should be shut down during copy

Guest Migration

■ Warm static (regular) migration

- ◆ **xm migrate**
 - Pauses a guest
 - Transfers guest state across network to a new Domain0
 - Resumes guest on destination host
- ◆ Network connections to and from guest are interrupted and probably will timeout

■ Live migration

- ◆ **xm migrate --live**
 - Copies a guest's state to a new Domain0
 - Repeatedly copies dirtied memory until transfer is complete
 - Re-routes network connections
- ◆ Network connections to and from guest remain active and uninterrupted

Guest Migration

■ Benefits

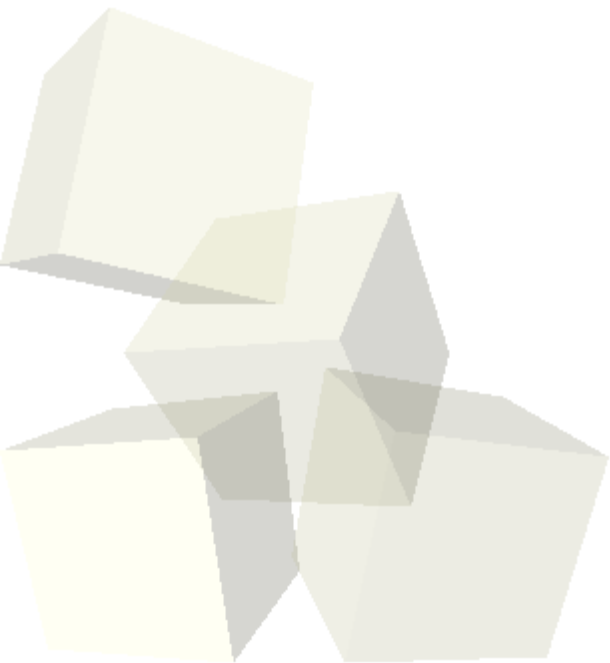
- ◆ Load balancing
- ◆ Hardware maintenance with little or no downtime
- ◆ Relocation for various reasons

■ Limitations

- ◆ Shared storage required
- ◆ Guests on same layer 2 network
- ◆ Sufficient resources needed on target machine
- ◆ CPU architectures need to match
- ◆ Some constraints on hypervisor version

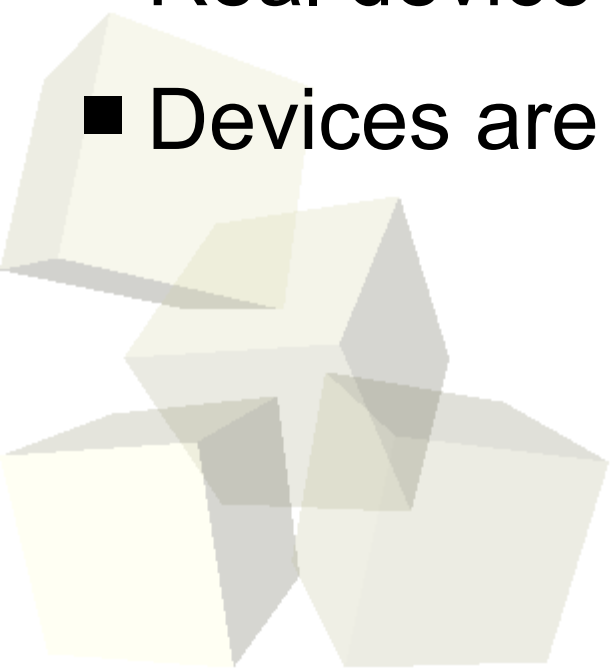
Xen Device Models

- Split driver model
- QEMU device model
- No virtualization device support
 - ◆ PCI passthrough demo

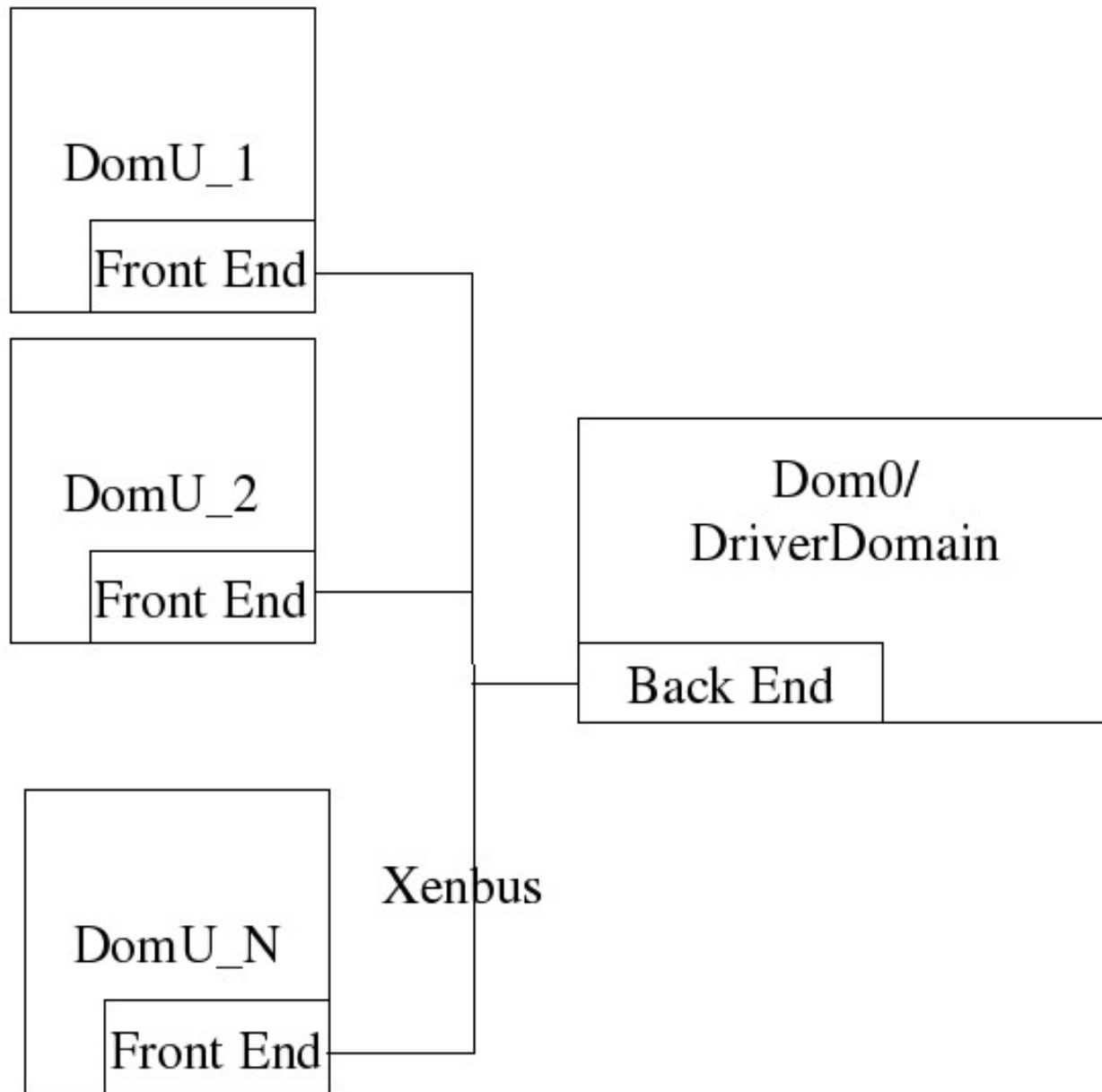


Split Driver Model

- Generic backends
 - ◆ Loaded in DriverDomain (often Domain0)
- Generic frontends (virtual devices)
 - ◆ Loaded in guest domain
 - ◆ Connects to corresponding backend driver
 - ◆ Guests use standard Xen virtual device drivers
- Real device-specific drivers are in DriverDomain
- Devices are multiplexed to the guests

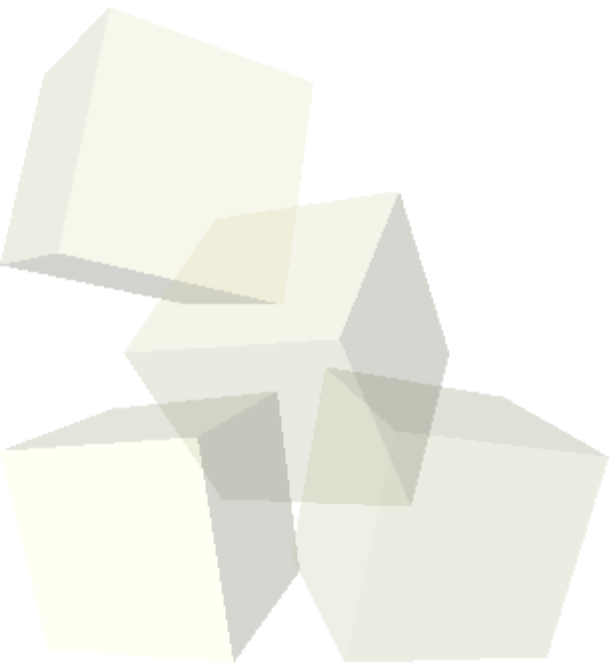


Split Driver Model



QEMU Device Model

- Provides emulation of devices
- Provides exclusive access illusion to the guests
- Used primarily for HVM guests
 - ◆ `device_model` set to `qemu-dm` binary in guest config

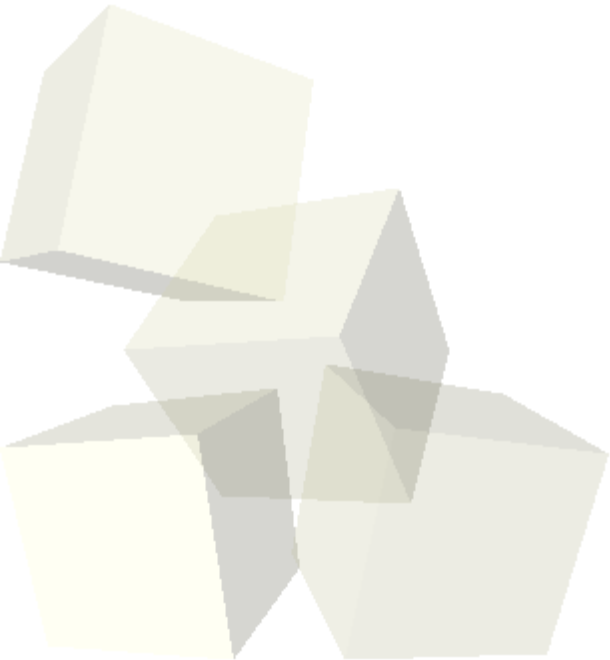


No Virtualization Device Support

- Guests granted full access to specific PCI devices
- The actual device driver runs in the guest
- Benefits
 - ◆ Highest performance for a device
 - ◆ Useful when virtualization doesn't support a device
 - ◆ System stability with buggy driver
- Limitations
 - ◆ Not (yet) well-tested
 - ◆ DriverDomain as backend can be tricky
 - ◆ HVM guest support still limited
 - ◆ Security considerations
 - Without an IOMMU, guests can (direct memory access) DMA into main memory

PCI Passthrough

■ Demo



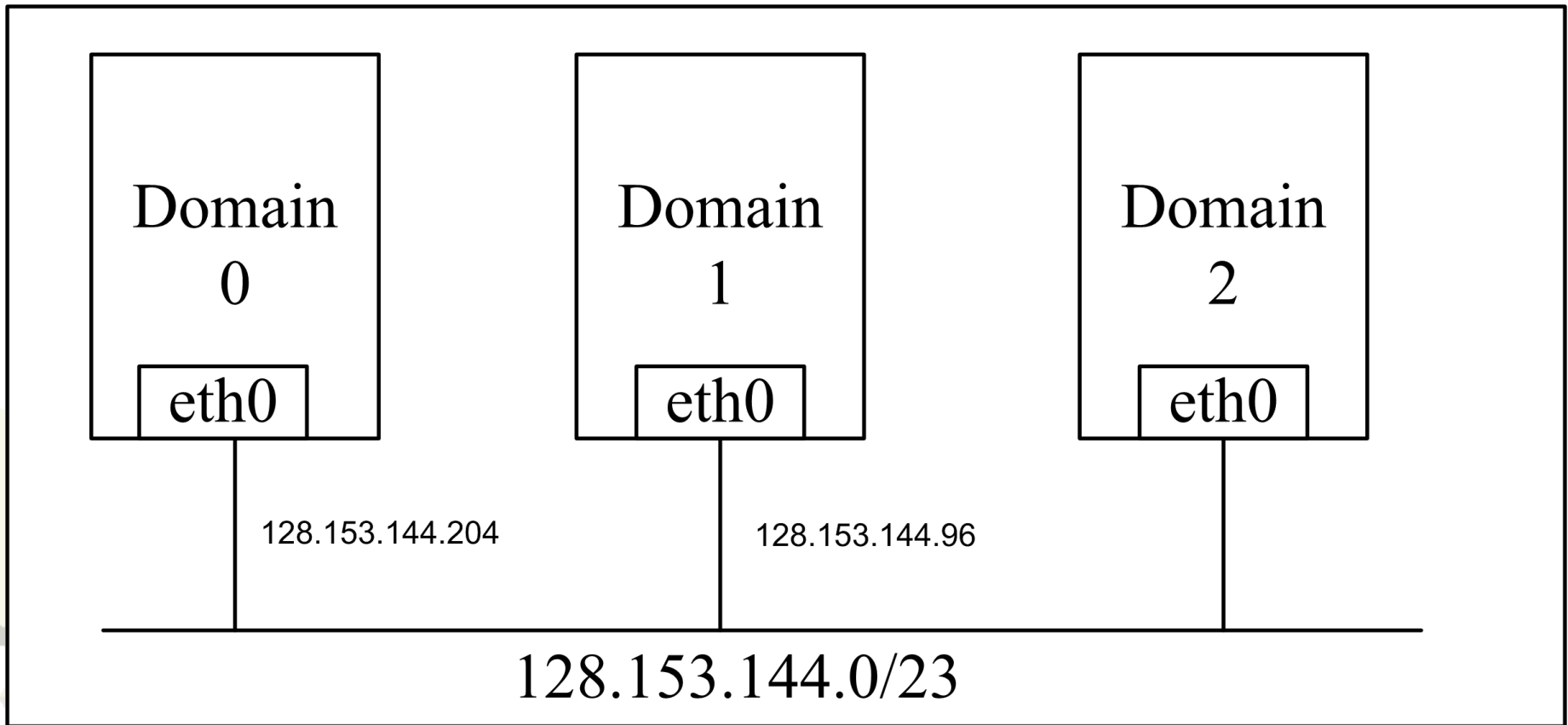
Network Configurations

- Network Bridge Configuration
- Network Route Configuration
- Network NAT Configuration
- Host-only Networking
- Handling Multiple Interfaces
- Virtual Private Network (VPN)
 - ◆ Xen's vnet configuration

Network Bridge Configuration

- A bridge relays traffic based on MAC address
- Behavior of guests in bridge mode
 - ◆ Appear transparently on the DriverDomain's network
 - ◆ Access the network directly (through software bridge)
 - ◆ Obtain IP address on the local Ethernet
- Configuration details
 - ◆ Set network-bridge and vif-bridge in xend config
 - ◆ Bridge is default guest configuration
 - Can set bridge option manually
- Secure with ebtables
 - ◆ MAC layer filtering (similar to iptables)
- Troubleshoot with brctl (bridge-utils package)
 - ◆ brctl show

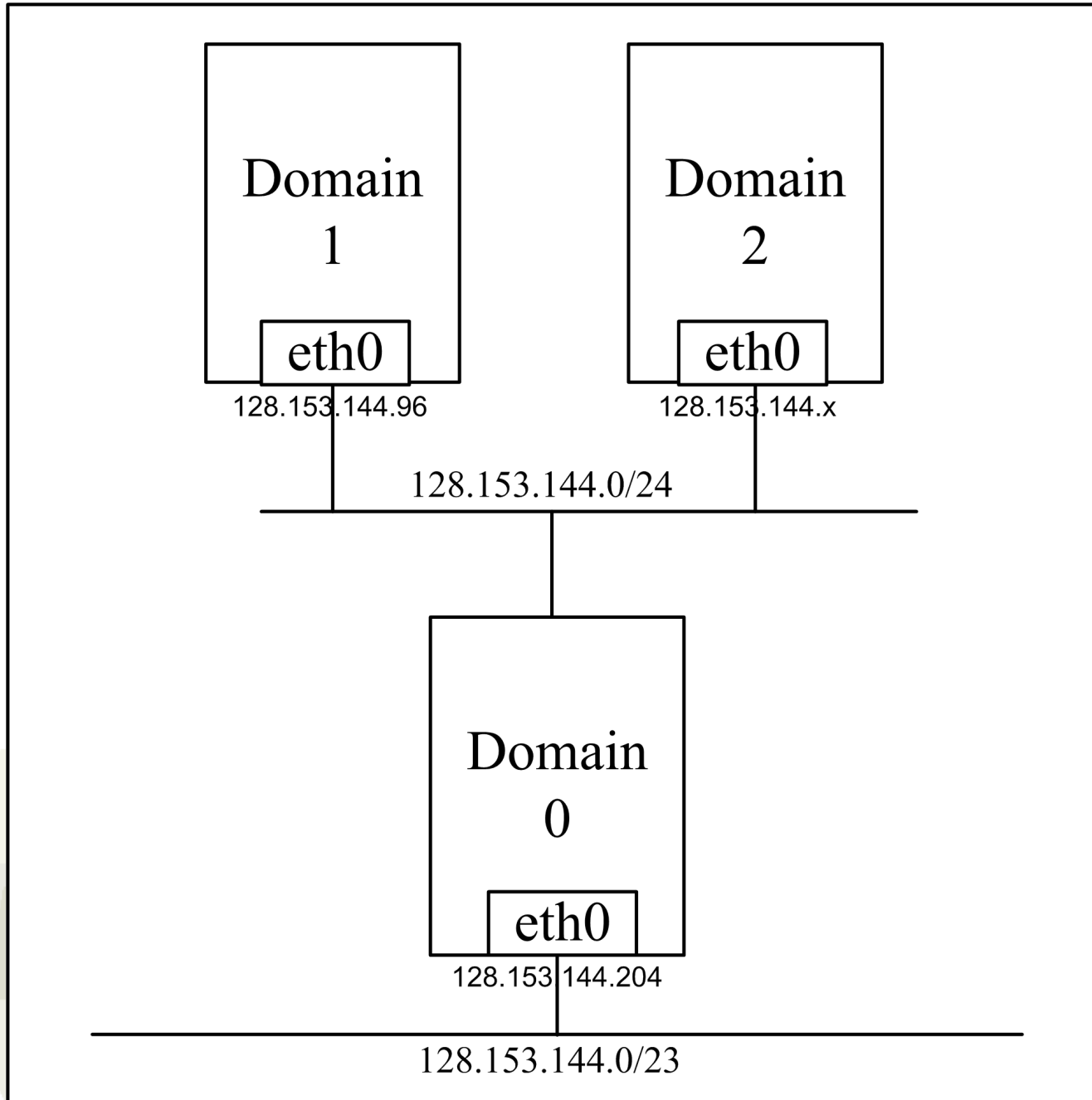
Network Bridge Configuration



Network Route Configuration

- A router relays traffic based on IP address
- Behavior of guests in route mode
 - ◆ Get routed to the Ethernet through the DriverDomain
 - ◆ Access the network via DriverDomain
 - ◆ Obtain IP address from DriverDomain
 - Manually or with DHCP relay
- Configuration details
 - ◆ Set network-route and vif-route in xend config
 - ◆ Set IP, netmask, and gateway in guest config
- Secure with iptables
 - ◆ ip_forwarding in DriverDomain
- Troubleshoot at IP layer
 - ◆ Network traces and iptables configuration

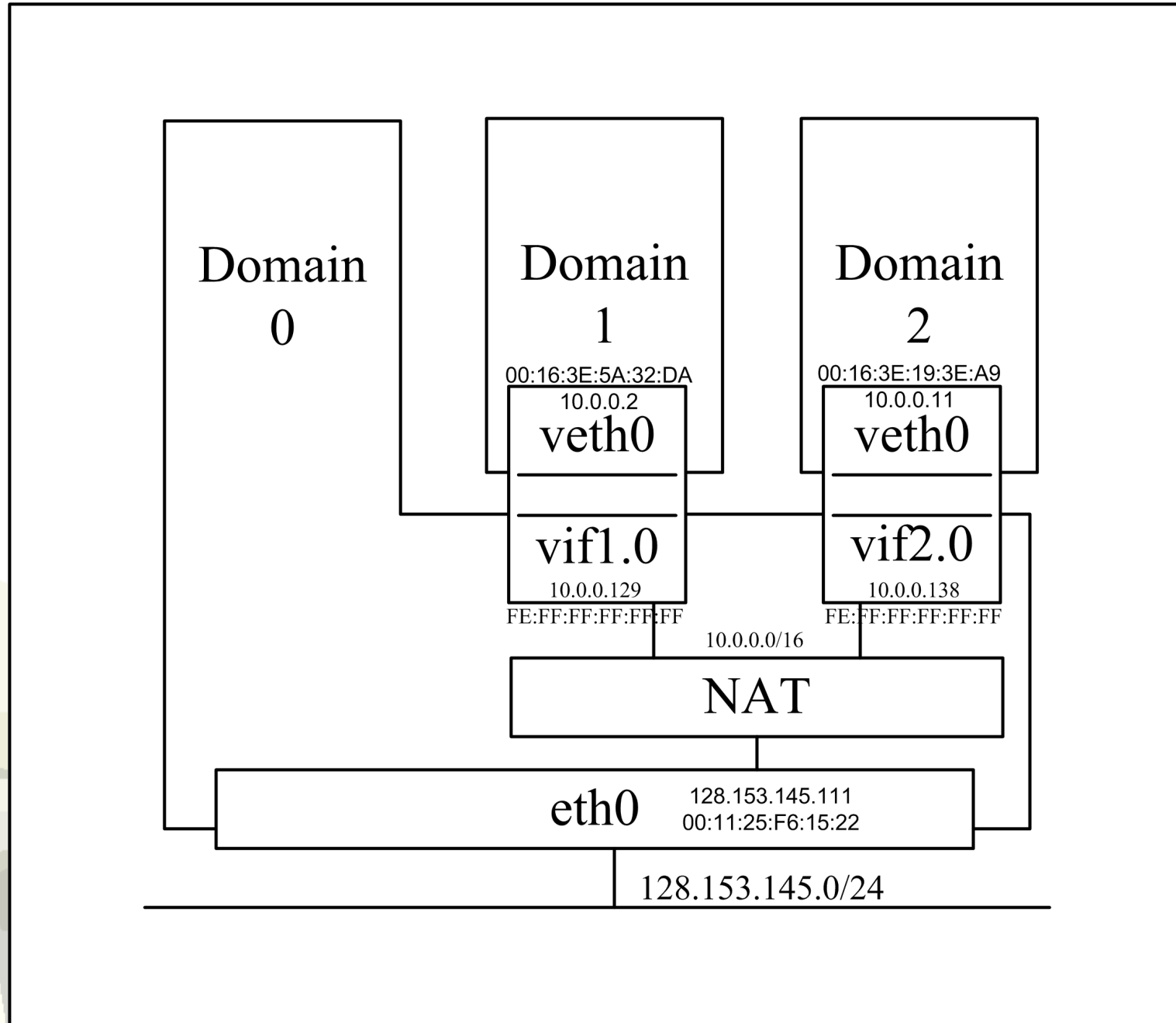
Network Route Configuration



Network NAT Configuration

- A NAT gateway translates between a single globally routable IP to many internal IP addresses
- Behavior of guests in NAT mode
 - ◆ Get NATed through the DriverDomain
 - ◆ Access outgoing traffic transparently through NAT
 - ◆ Are not visible from the outside (behind NAT router)
 - ◆ Obtain internal IP from software NAT router
- Configuration details
 - ◆ Set network-nat and vif-nat in xend config
 - ◆ Set IP, netmask, and gateway in guest config
- Secured with iptables
 - ◆ Uses MASQUERADE chain
- Troubleshoot with internal and external traces

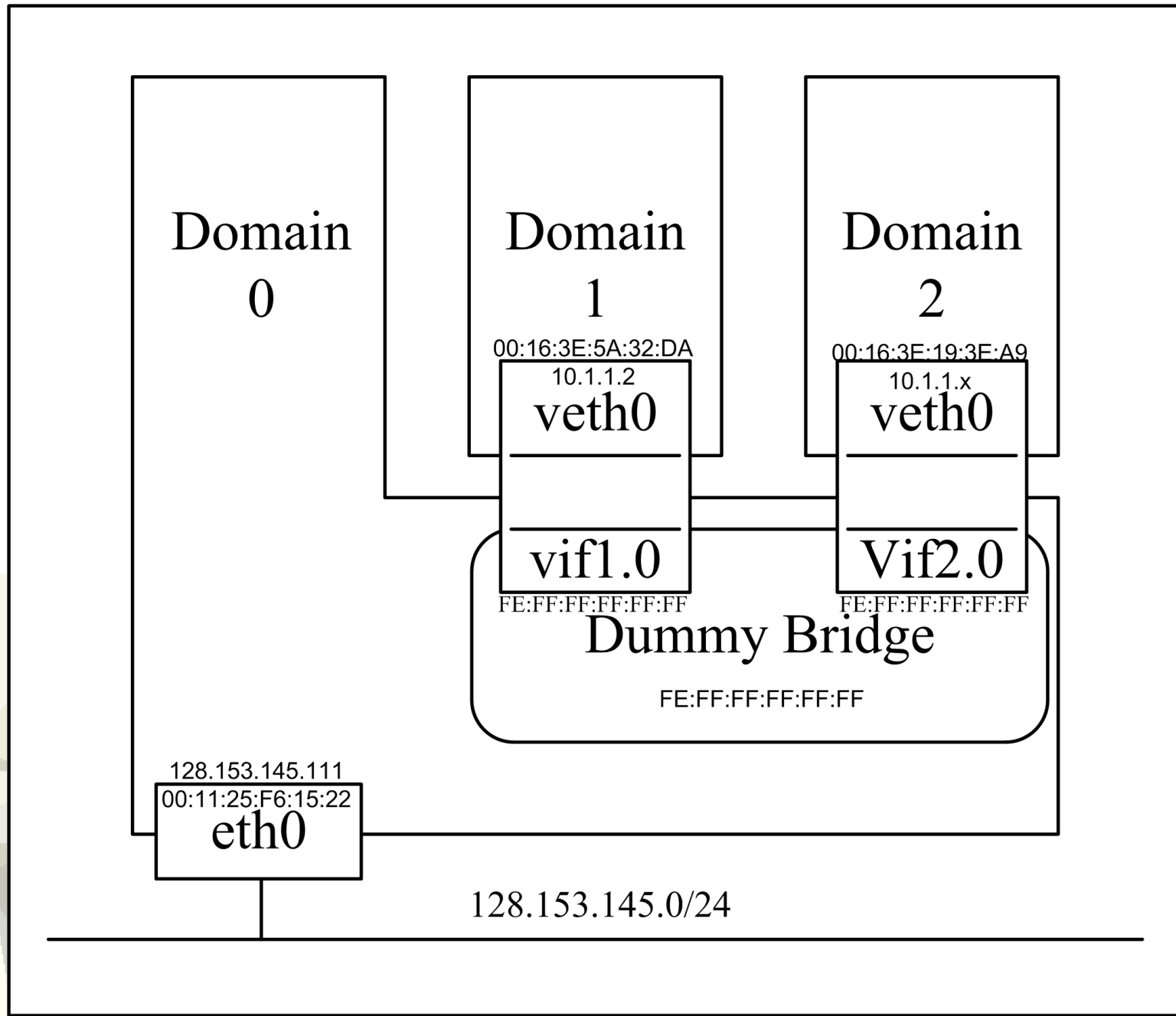
Network NAT Configuration



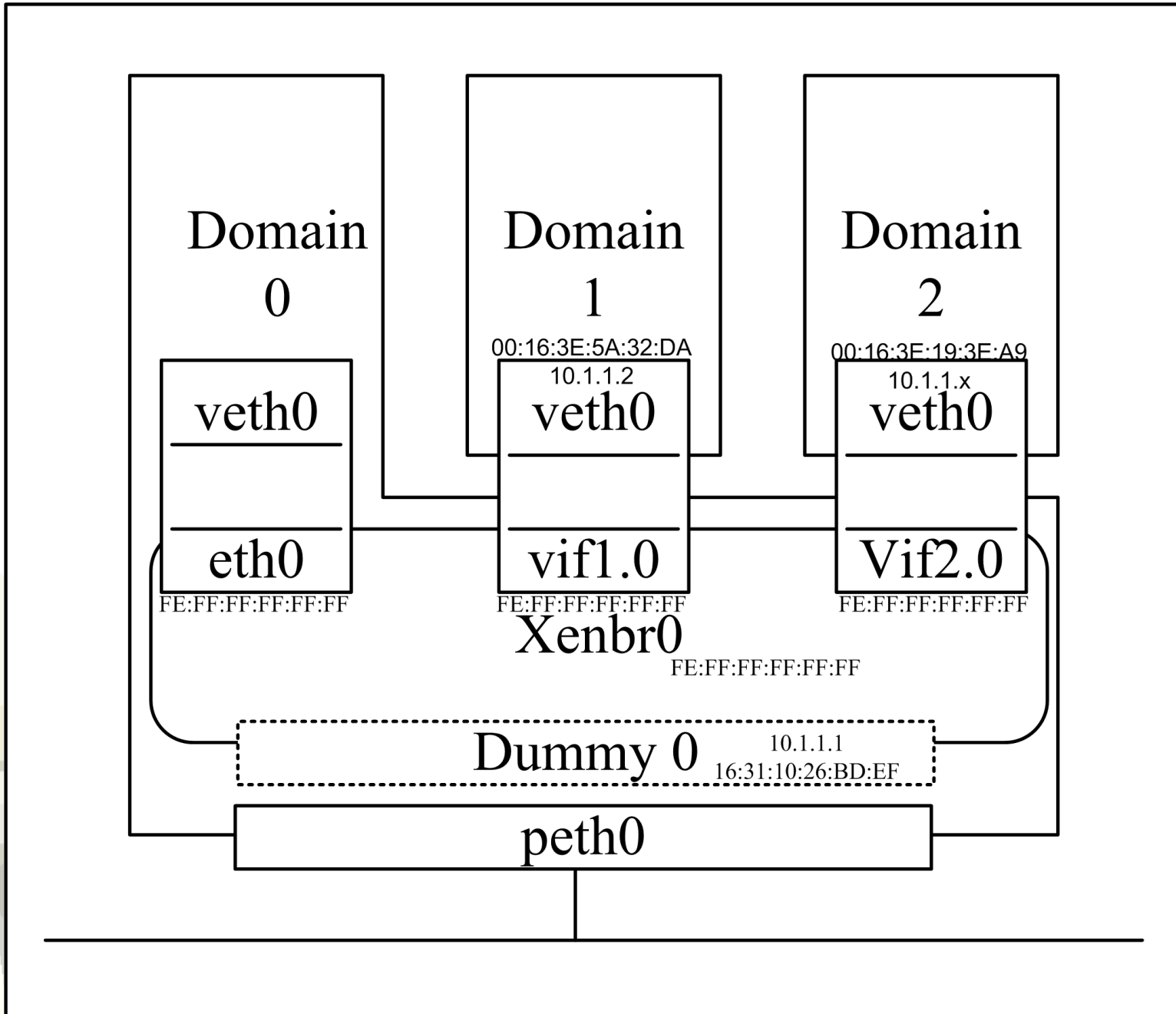
Host-only Networking

- Supports local-only network traffic
 - ◆ Share data on private network
- Dummy bridge connection options
 - ◆ Guests to other guests only
 - ◆ Host (DriverDomain) to guests only
- Configuration details
 - ◆ Set up dummy bridge in DriverDomain
 - ◆ Set bridge to dummy bridge in guest config
- Secure with ebtables (if needed)
- Troubleshoot with brctl and iptables

Host-only Networking (Guest to Guest)



Host-only Networking (Host to Guests)



■ Handling Multiple Interfaces

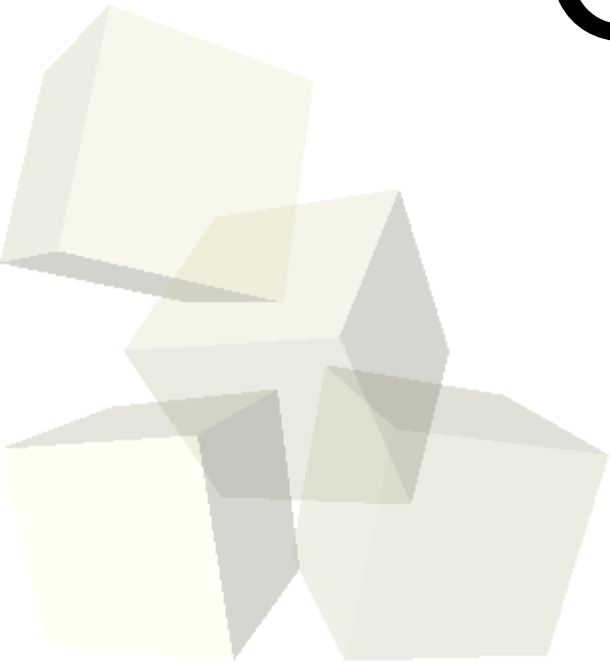
- Configure the purpose of NICs in xend config
- Specify bridge device in guest config
- Create custom scripts that call network-* scripts
 - ◆ Set virtual device number (vifnum)
 - ◆ Set network device (netdev)
 - ◆ Set bridge device (bridge)
- Add custom network script in xend config
- Add vif entries and specify bridge in guest config

Virtual Private Network (VPN)

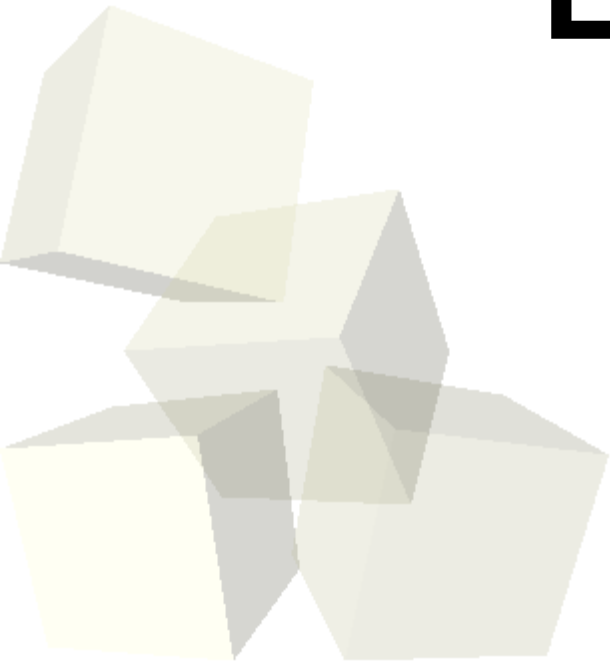
- Provides confidential communication on public network
 - ◆ Illusion of VPN client on VPN server's local network
- Xen's vnet configuration
 - ◆ VPN implementation based on bridged network
 - ◆ Installation methods
 - Kernel module
 - Userspace daemon in DriverDomain
 - ◆ Not well-tested (yet)



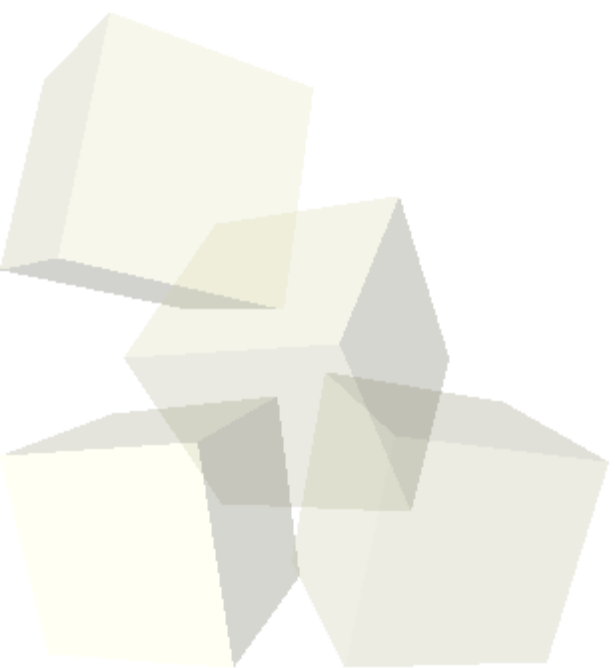
Questions?



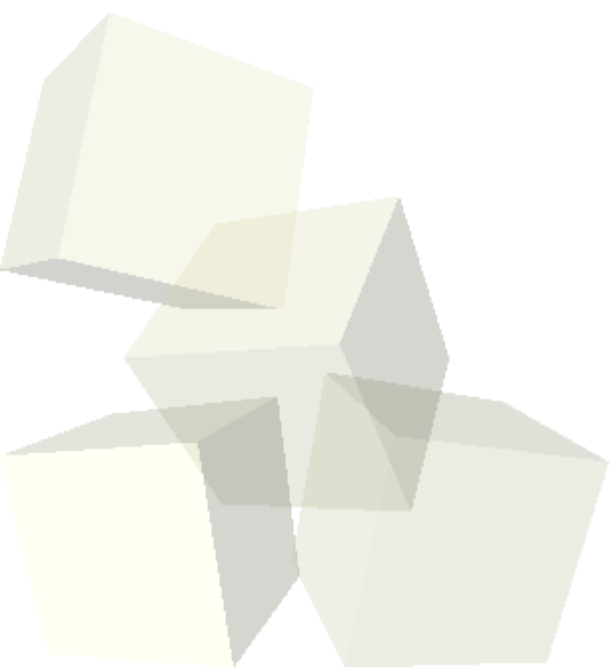
Break Time



Unit 4

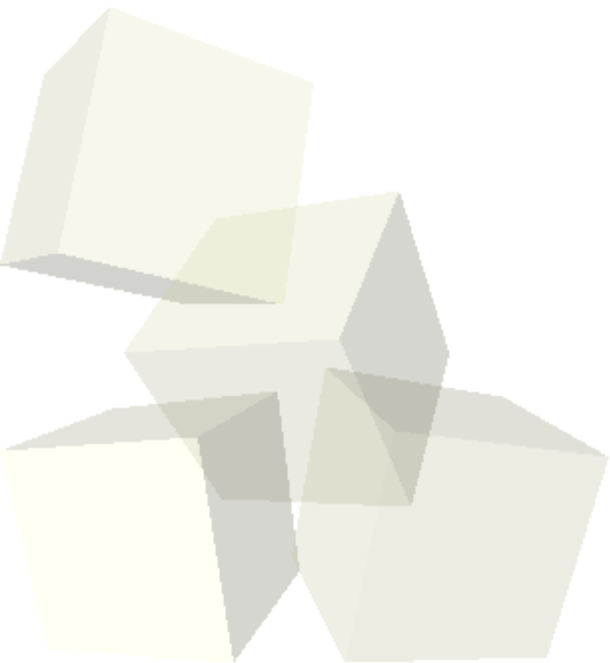


- Security Basics
- Advanced Security: sHype and XSM
- Management APIs and Utilities
- Resources and Performance
- Future Directions



Security Basics

- Standard system security practices apply
- Secure Domain0 and Xen
- Secure guest domains normally



Security Basics

- Secure Xen Hypervisor
 - ◆ Secure Hypervisor (sHype)
 - ◆ Xen Security Modules (XSM)
- Secure Domain0
 - ◆ Minimize software packages
 - ◆ Minimize running services and open ports
 - ◆ Use firewall and intrusion detection systems
- Secure guests
 - ◆ Similar to Domain0 – software, services, and ports
 - ◆ IOMMU for direct PCI device pass-through
 - Prevent insecure memory access through driver DMA

Advanced Security: sHype and XSM

■ Secure Hypervisor (sHype)

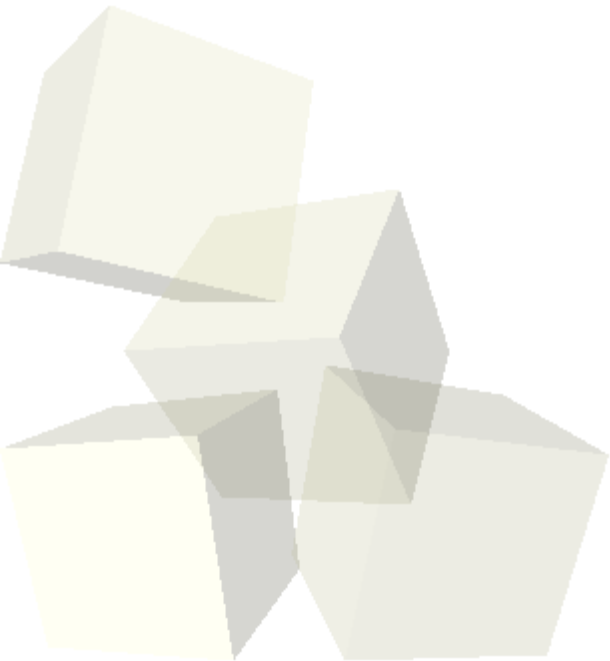
- ◆ IBM Research project
- ◆ Role-Based Access Control (RBAC)
 - Rights-based on privilege group
- ◆ Mandatory Access Control (MAC)
 - Labeled objects
 - Privilege groups given access to specific object labels
- ◆ Virtual Trusted Platform Module (vTPM)
 - Enables remote attestation by digitally signing cryptographic hashes of software components
 - “Attestation” means to affirm that some software or hardware is genuine or correct

■ Xen Security Modules (XSM)

- ◆ National Security Agency (NSA) security framework
- ◆ Based on SELinux

Advanced Security: sHype

- **Demo**



Management APIs and Utilities

■ APIs

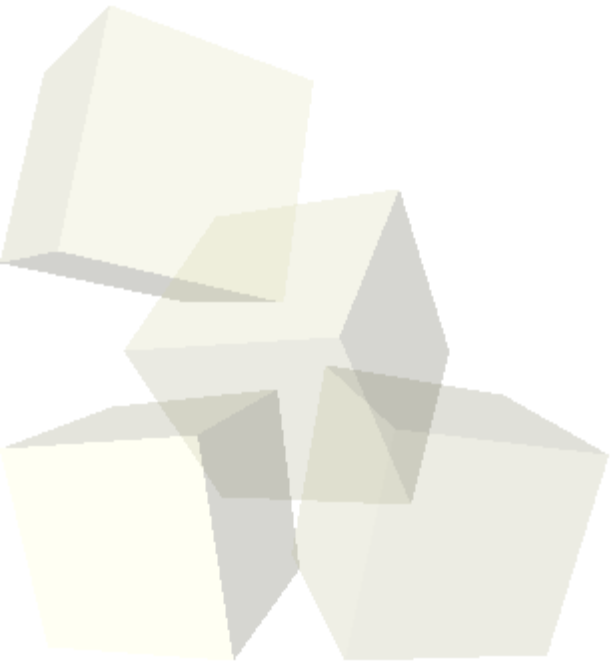
- ◆ libvirt
- ◆ Xen Application Programming Interface (API)
- ◆ Xen Common Information Model (CIM)

■ Management utilities

- ◆ virt-manager
- ◆ virsh
- ◆ XenMan (ConVirt)
- ◆ Enomalism
- ◆ Citrix XenServer
- ◆ Virtual Iron
- ◆ IBM Director (IBM Virtualization Manager extension)

Management APIs and Utilities

- Demo



Resources and Performance

- Gauging performance
 - ◆ xm top (demo)
 - ◆ xentop
- Memory management
 - ◆ xm mem-set
 - ◆ xm mem-max
- Virtual CPU management
 - ◆ xm vcpu-list
 - ◆ xm vcpu-set
 - ◆ xm vcpu-pin

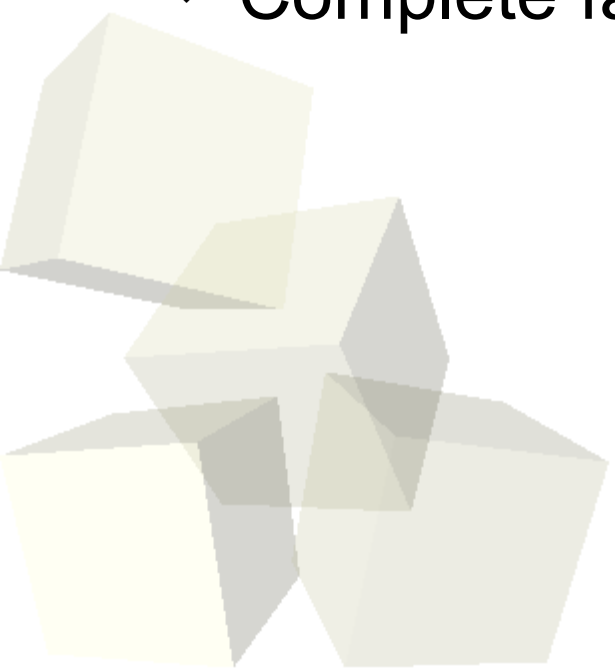
Resources and Performance

■ Xen scheduler

- ◆ Credit scheduler
 - Weight and cap

■ I/O schedulers

- ◆ Noop scheduler
- ◆ Deadline scheduler
- ◆ Anticipatory scheduler (as)
- ◆ Complete fair queuing scheduler (cfq)



Resources and Performance

■ Macro benchmarks

- ◆ Disk I/O
 - iozone, iometer, bonnie++
- ◆ Network
 - netperf, iperf, ttcp
- ◆ CPU
 - SPECint, kernbench
- ◆ Web and database servers
 - SPECweb, Hammerhead, Apache bench (ab)
 - Database benchmarks

■ Micro benchmarks and profiling

- ◆ xenoprof, xenperf, xentrace, xenmon
- ◆ virtbench

■ Performance isolation suite (from Clarkson)

Future Directions

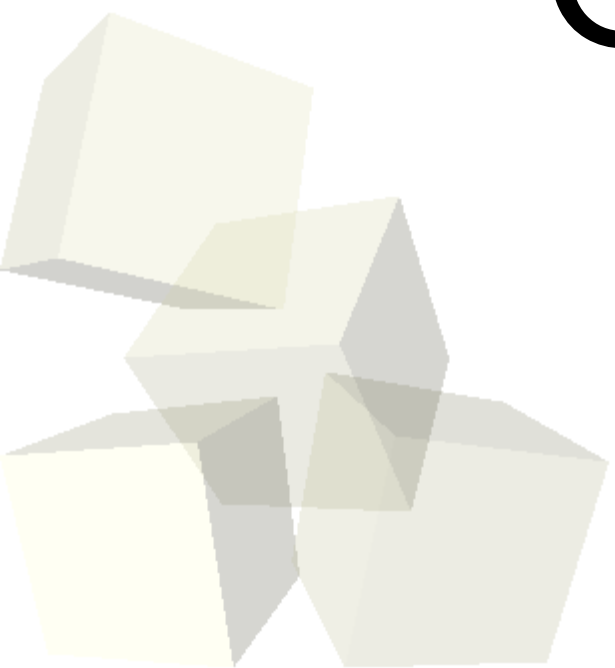
- Smart devices
 - ◆ Virtualization-aware NICs, video cards
- Next generation hardware support
 - ◆ AMD NPT, Intel EPT, etc.
- IOMMUs
- Trusted Platform Module (TPM) support
- Xen Domain0 inclusion in mainline Linux
- PV drivers for HVM guests
 - ◆ PVGPL
 - ◆ Citrix
 - ◆ Novell
 - ◆ Various others
- Microsoft enlightenments support

Further Information

■ Useful resources

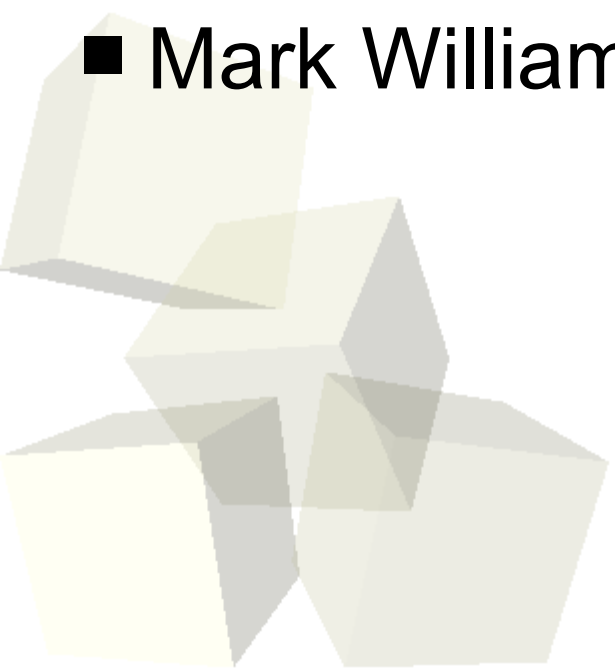
- ◆ Xen Community
- ◆ XenWiki
- ◆ Xen Mailing Lists
- ◆ Xen Bugzilla
- ◆ Xen Summit
- ◆ Xen source code
- ◆ Academic papers and conferences
- ◆ ***The Definitive Guide to the Xen Hypervisor*** (book)
- ◆ ***Running Xen: A Hands-On Guide to the Art of Virtualization*** (book)

Questions?



Acknowledgments

- “Running Xen” book co-authors
 - ◆ Jeanna Matthews
 - ◆ Eli M. Dow
 - ◆ Wenjin Hu
 - ◆ Jeremy Bongio
 - ◆ Brendan Johnson
- Padmashree Apparao (Intel Research)
- Mark Williamson (University of Cambridge)



References

- <http://runningxen.com>
- http://runningxen.com/mailman/listinfo/readers_runningxen.com
- <http://www.usenix.org/publications/login/2007-02/pdfs/hand.pdf>
- <http://wiki.xensource.com/xenwiki/XenStoreReference>
- <http://portal.acm.org/citation.cfm?id=1281700.1281706&coll=&dl=ACM>
- <http://www.usenix.org/publications/login/2007-02/pdfs/griffin.pdf>
- <http://passat.crhc.uiuc.edu/dasCMP/papers/dasCMP07/paper01.pdf>
- <http://ieeexplore.ieee.org/iel5/4299339/4299340/04299347.pdf?isnumber=4299340&prod=CNF&arnumber=4299347&arSt=2&ared=2&arAuthor=Apparao%2C+Padma%3B+Makineni%2C+Srihari%3B+Newell%2C+Don>
- http://workspace.globus.org/vtdc06/VTDC_files/programdraft.htm
- http://xen.org/files/xensummit_fall07/28_PadmaApparao.pdf
- <http://opensolaris.org/os/project/libmicro/>
- <http://www.computerworld.com/action/article.do?command=viewArticleBasic&taxonomyName=Storage&articleId=9081798&taxonomyId=19>
- http://weblog.infoworld.com/daily/archives/2008/04/top_3_gotchas_o.html
- http://news.zdnet.com/2100-3513_22-6191965.html
- <http://www.cl.cam.ac.uk/research/srg/netos/xen/readmes/hg-cheatsheet.txt>
- <http://www.cuddletech.com/blog/pivot/entry.php?id=469>

Thank you for coming.

Questions/Comments?