



QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Xen and the Art of Distributed Virtual Machine Management

Dr. Greg Lavender
Mr. Adam Zacharski

Department of Computer Sciences
The University of Texas at Austin

Session TS-1743

Talk Outline

Goals

Virtualization

Xen Hypervisor

System Configuration

Monitoring and Management System

Observations and Future Work

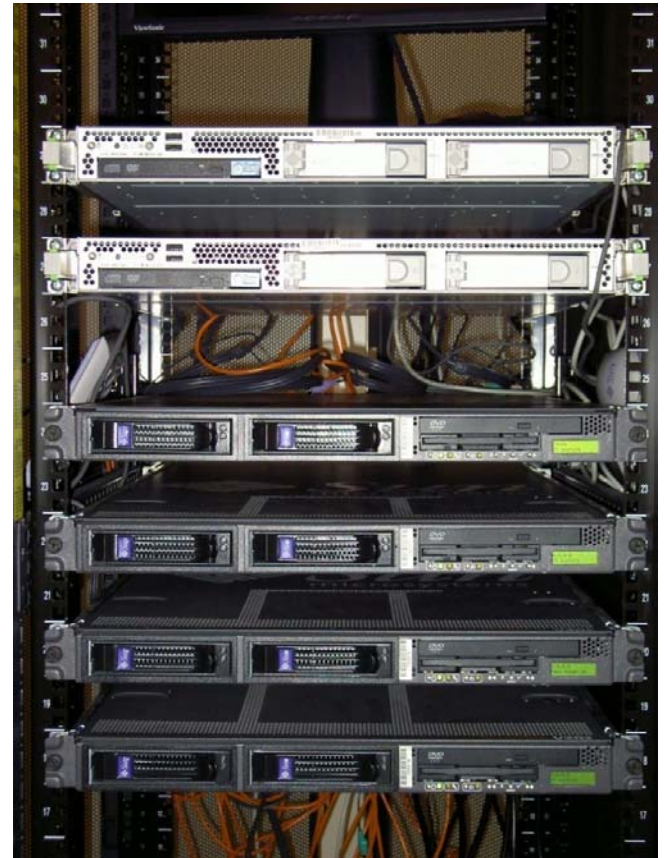
Overview

- Computers in instructional labs often sit idle except during peak times
- Replace lab machines with virtual machines
- Use Linux, OpenSolaris operating system (OS) and Xen for virtualization
- But virtualization requires intelligent management software
 - Management GUI
 - Monitoring/Management Agents
 - Pub/Sub using Java™ Message Service (JMS)/ Sun Java System Message Queue (JMQ)



Goals

- Replace fat desktops with multi-core servers
- Simplify management of virtual machines
 - A central interface for the creation and control of the virtual machines
 - An automated system that will keep virtual machines and their load as evenly distributed as possible

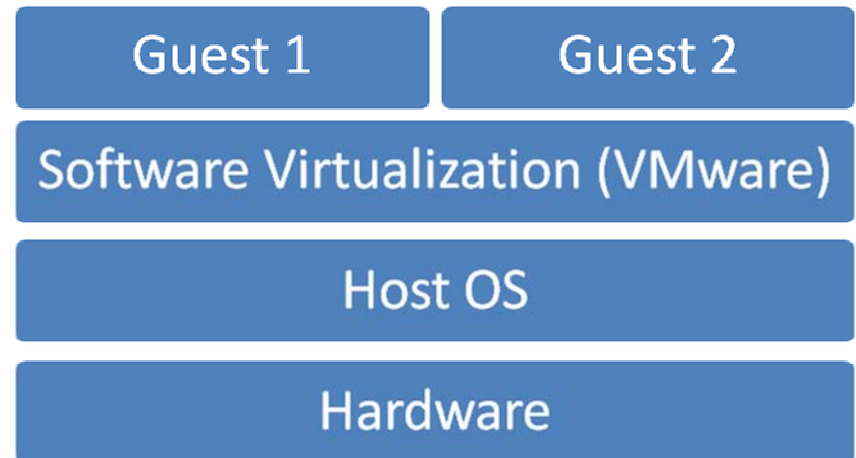


Virtualization

- Abstraction of physical machine resources
- Three kinds (broadly speaking)
 - Software virtualization
 - Paravirtualization
 - Hardware virtualization
- Reasons for virtualization
 - Better utilization
 - Sandboxing
 - Reliability
 - Manageability

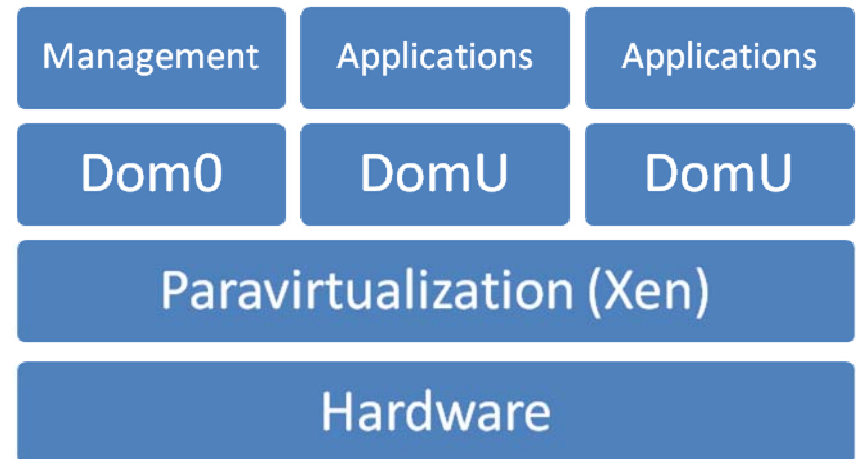
Software Virtualization

- Runs above the host OS
- Two methods
 - Binary translation
 - Trap-and-emulate
- VMWare and Parallels



Paravirtualization

- Software that runs directly on the hardware
- Requires host (Dom0) and guest (DomU) operating systems to be modified
- Performance benefits (e.g., IO)
- Can take direct advantage of hardware virtualization features
 - Xen, KVM, VMWare ESX



Hardware Virtualization

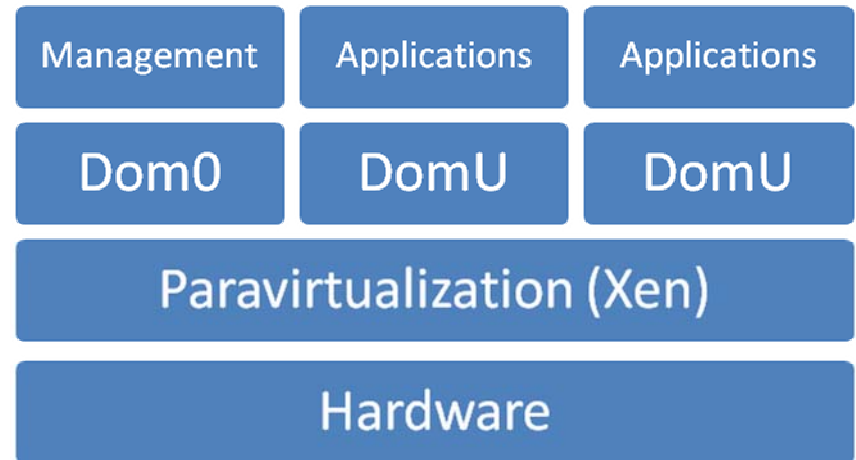
- Intel VT™ and AMD-V™ (“Pacifica”)
 - Both x86/x64 and AMD64
 - AMD Opteron™, Core 2 Duo, Woodcrest...
- Hardware isolation supports sandboxing
 - IOMMU
- Paravirtualization can directly benefit from hardware virtualization

What Is Xen?

- Open source software hypervisor that provides a paravirtualization layer
 - Xen 3.0.4 latest open source version
 - Created by Ian Pratt at the University of Cambridge
- Supported by Unix/Linux kernels
 - Ubuntu, Debian, SuSe, Red Hat...
 - OpenSolaris OS, FreeBSD, NetBSD...
 - Windows XP (currently unavailable)

Host and Guest VMs

- Dom0—Administrative
 - Controls all VMs
 - Admin tools for monitoring and management
- DomU—Guest
 - Users generally unaware they are on a virtual machine



Xen Live Migration

- Allows virtual machines to be migrated from one physical machine to another with minimal disruption of service
- Migrate entire VM state, including virtual memory pages and swap
- Phases of memory migration
 - Push
 - Stop and copy
 - Pull
- Remap network traffic to new machine
 - Remap IP traffic using ARP

Xen Migration Times

Job	Total Migration Time	Downtime
Quake 3 Server	7 Seconds	70 ms
SPECweb99	71 Seconds	210 ms
Stop&Copy	30 Seconds	3.5 Seconds

System Configuration

- Ubuntu Linux AMD-64 version 6.06
 - w/kernel extensions for Xen 3.0.3
- OpenSolaris operating system x64
 - Xen 3.0.4 support in progress
- Global Network Block Device (GNBD) for NAS
- JMQ 3.5 SP1
 - Pub/Sub agents

Management Platforms



Sun X2100—JMQ Server and Configuration Server

Processor	Dual Core AMD Opteron 175
RAM	4GB DDR2 ECC
Hard Drive	2x 250GB 7200RPM
Networking	2x 10/100/1000-Mbps Ethernet Ports
Operating System	Ubuntu 6.06 Dapper for the Configuration Server and Solaris™ 10 OS x64 for the JMQ Server

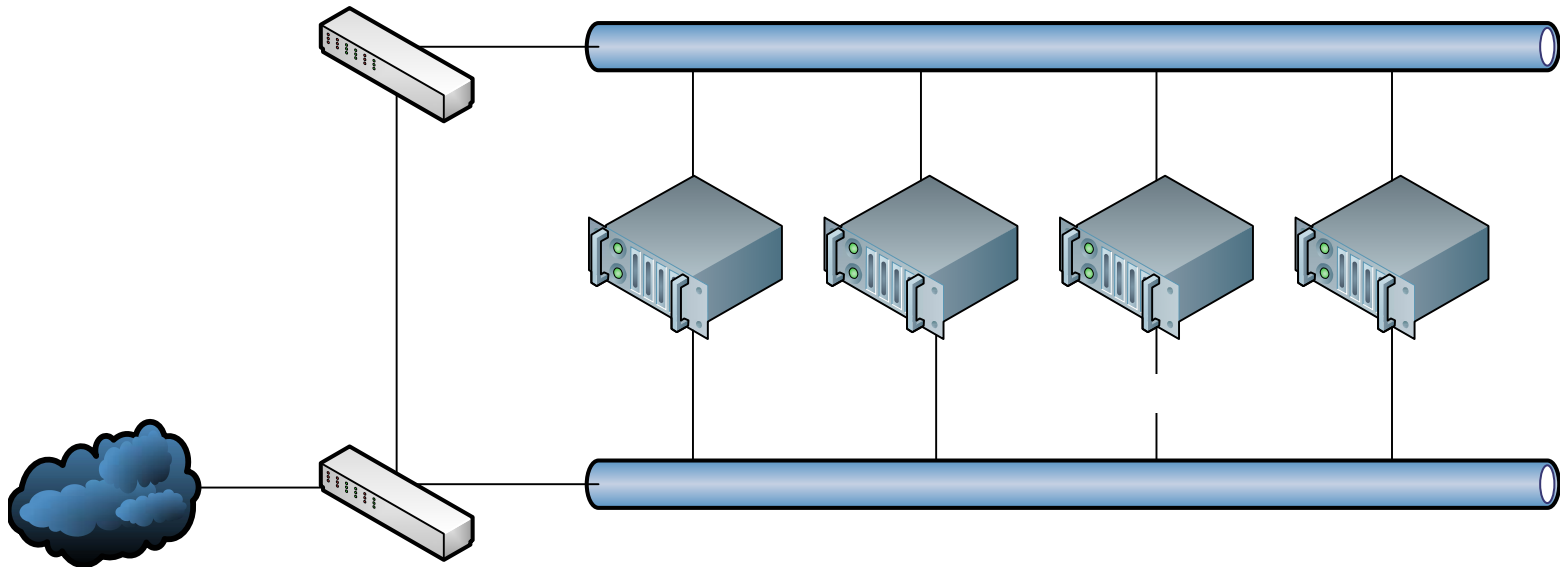
Virtual Machine Platforms



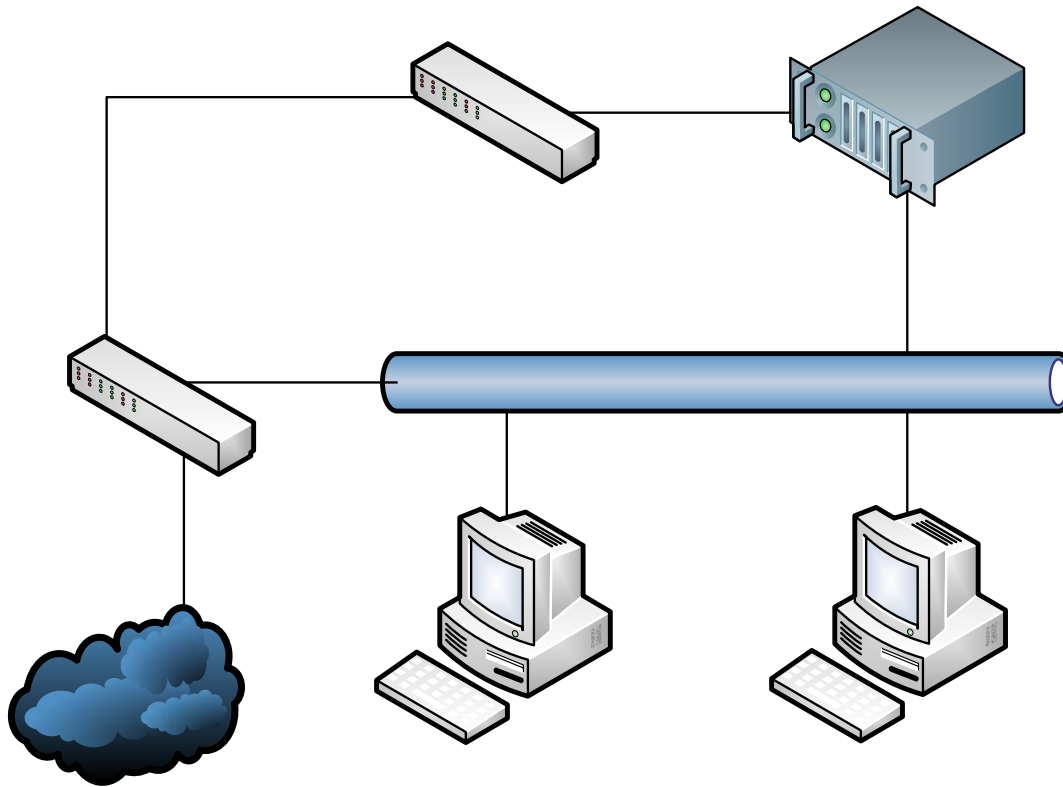
Sun V20z—Virtual Machine Hosts (x4)

Processor	2x Dual Core AMD Opteron 275
RAM	16GB DDR2 ECC
Hard Drive	2x 300GB 10K RPM
Networking	2x 10/100/1000-Mbps Ethernet Ports 10/100-Mbps Ethernet With Two External Ports
Operating System	Ubuntu 6.06 Dapper
Xen	3.0.3

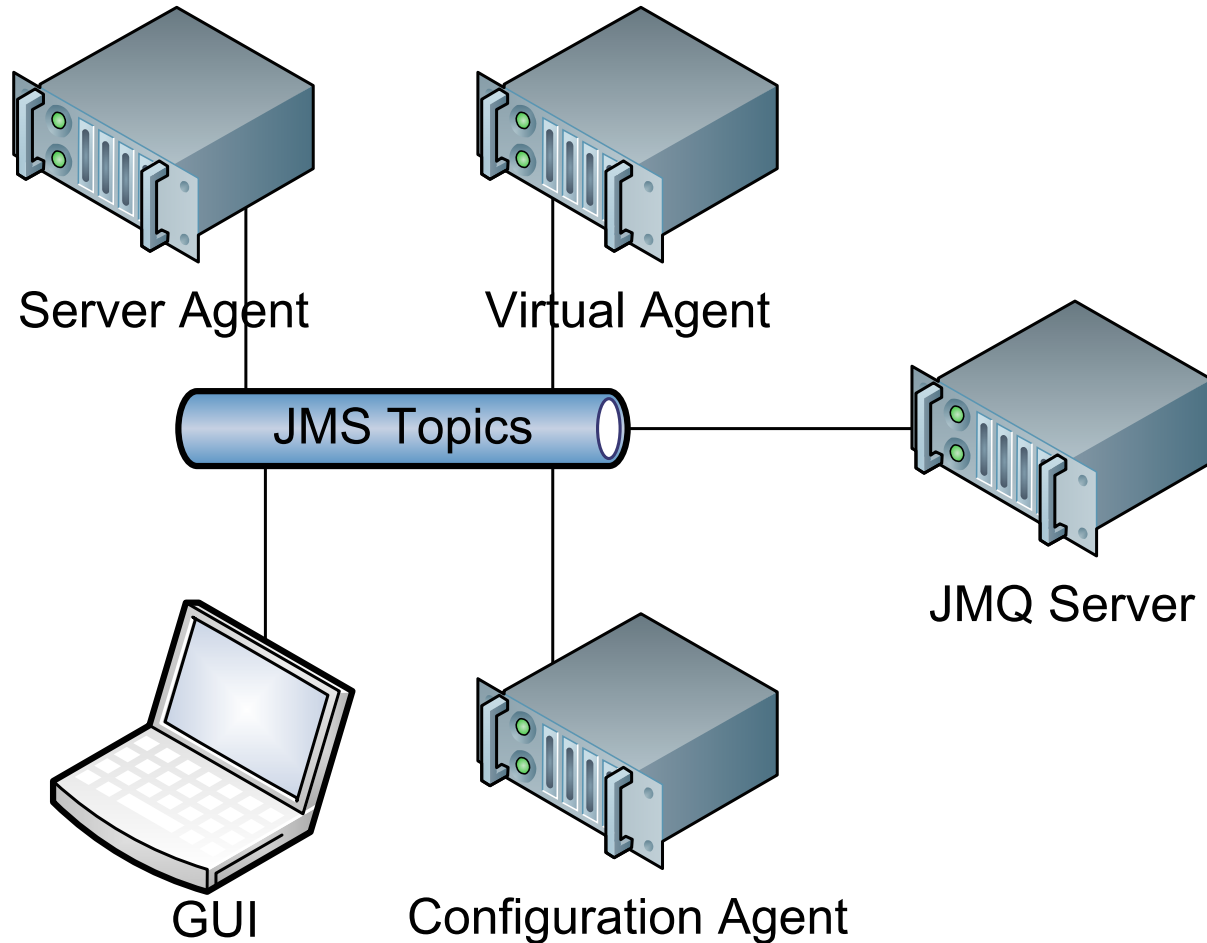
Physical Network Separation



VM Traffic on Separate VLAN



JMS/JMQ Technology Agent Configuration



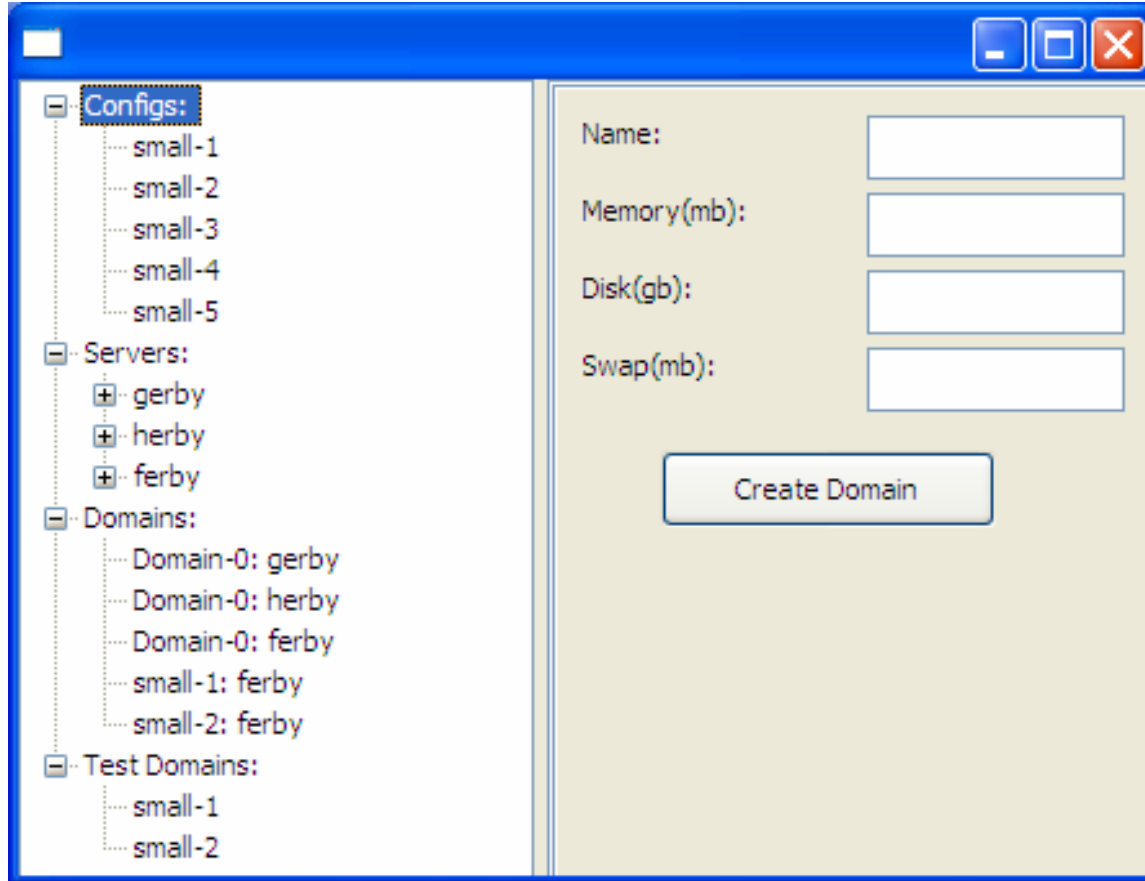
Management GUI and Agents

- Simplify VM Config, monitoring, and management
 - JMS/JMQ technology, SWT, JFreeChart
- Agents
 - Three types
 - System configuration agent
 - Physical server agents
 - Virtual machine agents
 - Collect information and execute actions on physical and virtual machines
 - Communicate using the JMS/JMQ technology message bus

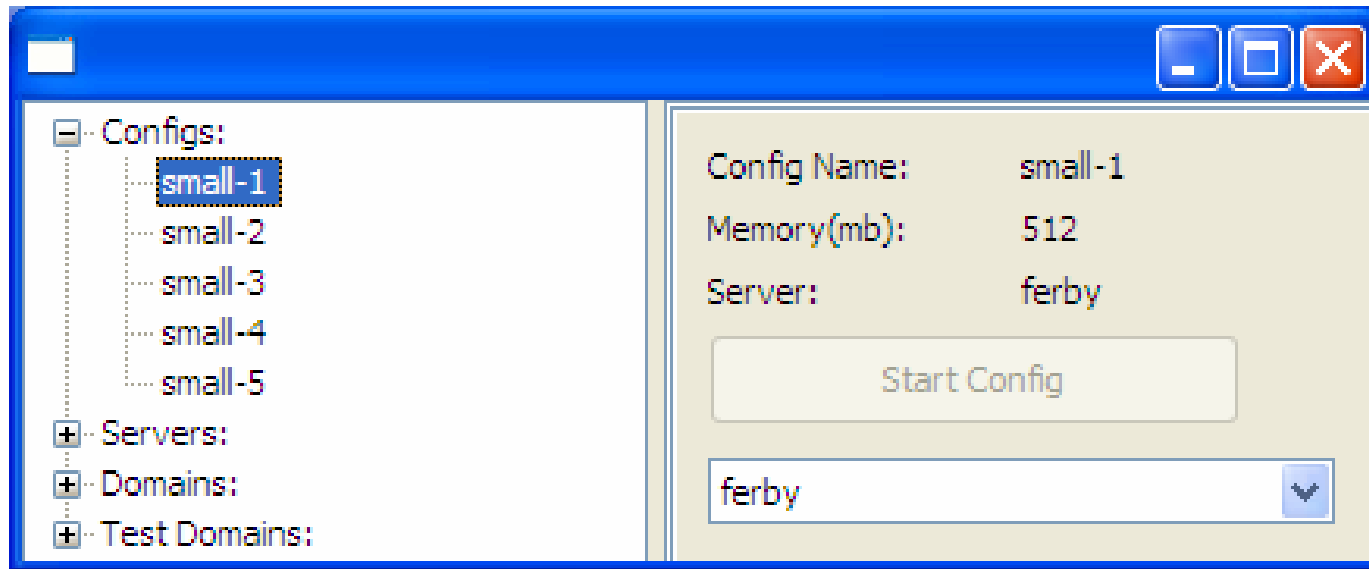
Configuration Agent

- Runs on the configuration server
 - Creates/manages VM configurations
- Publishes config messages with available configurations
- Management GUI subscribes to the config topics
- Management GUI publishes requests to create new VM DomUs
- Config agent subscribes to the config topics

VM Config Creation



View Existing Configurations



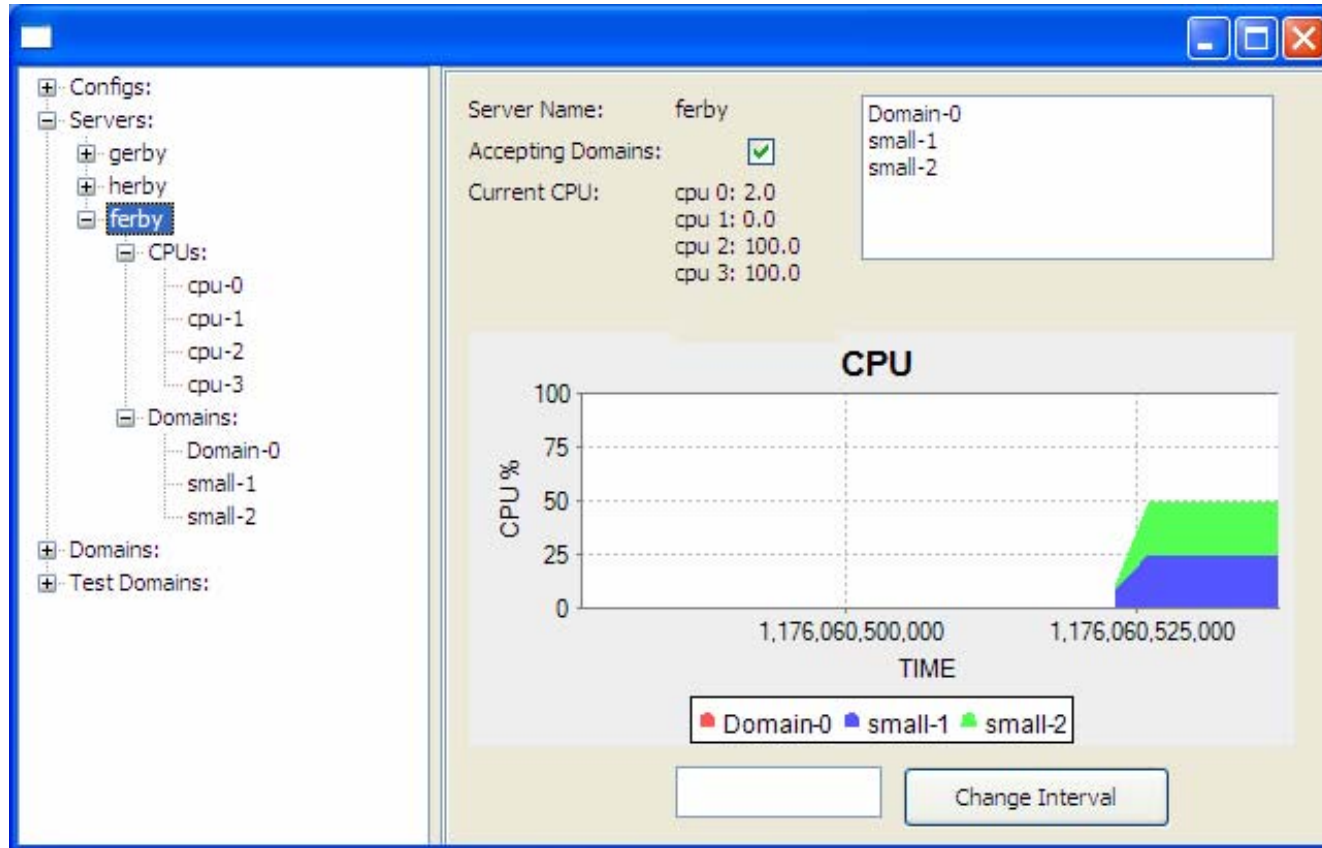
Server Agents

- Run in Dom0 of each physical machine
 - Provide all VM control functions per host
 - Negotiates and initiates live migration of DomU guests
 - Also respond to explicit migration commands from Management GUI
- Monitor server resources and local VMs
 - Collect same information as xentop
 - CPU, memory, network, IO
 - Publishes status messages every 5 seconds
 - Management GUI subscribes to the status topic

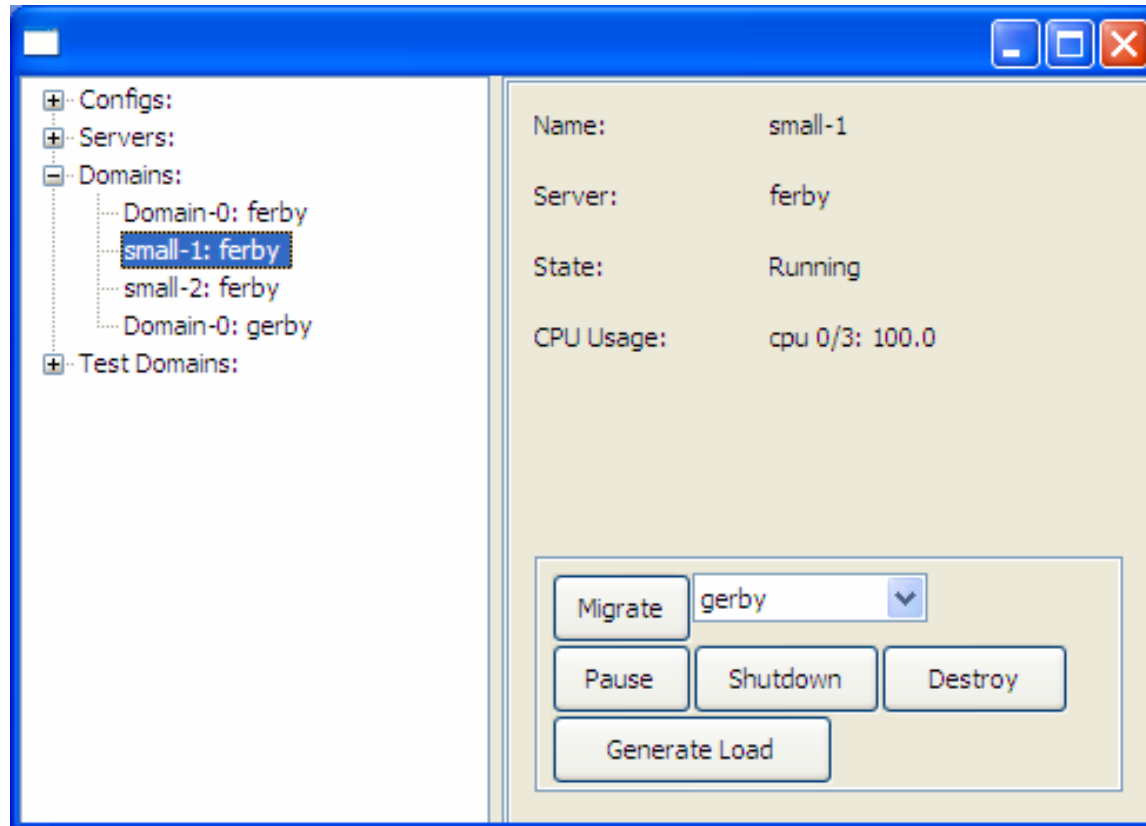
Server Agents

- Customizable monitoring algorithm per host
 - Knows about local machine resources
 - Sampling frequency, thresholds, history sensitive
 - CPU, memory, IO, network
- Heuristically determines when to migrate a VM
 - Broadcasts a “help” message to the message bus
 - Currently chooses “first responder”
 - As a responder, doesn’t offer help if its own system is overutilized or near thresholds
- Can migrate all VMs on command
 - To offload all VMs for machine maintenance
 - Good idea to keep 1–2 machines lightly loaded

Server View



Domain View



The screenshot shows a window titled "Domain View" with a blue title bar. On the left is a tree view with the following structure:

- Configurations: Configs: (expanded)
- Servers: (expanded)
- Domains: (expanded)
 - Domain-0: ferby
 - small-1: ferby** (selected)
 - small-2: ferby
 - Domain-0: gerby
- Test Domains: (expanded)

The right pane displays the following details for the selected domain:

Name:	small-1
Server:	ferby
State:	Running
CPU Usage:	cpu 0/3: 100.0

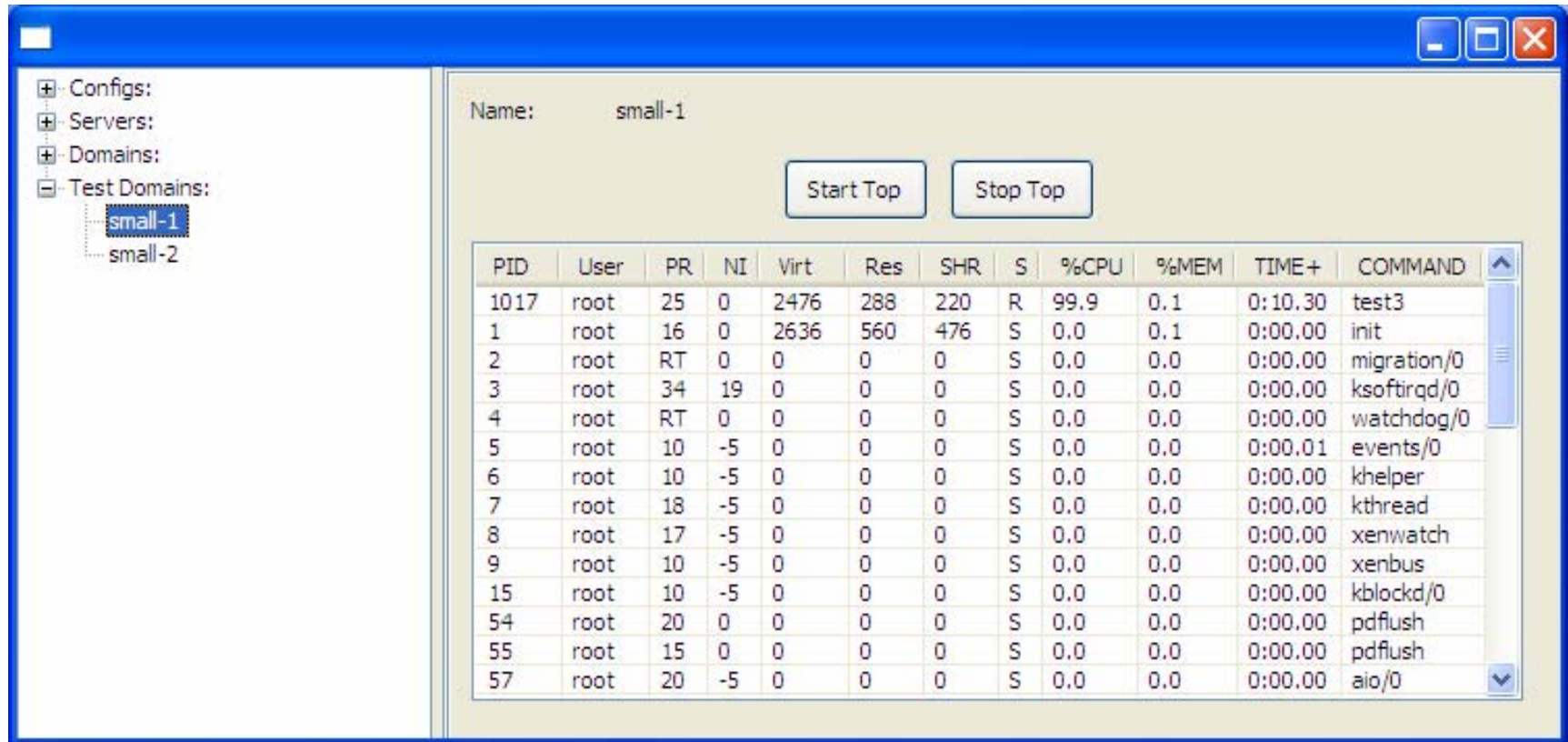
At the bottom of the right pane, there are several control buttons:

- Migrate: with a dropdown menu currently showing "gerby"
- Pause
- Shutdown
- Destroy
- Generate Load

Virtual Agent

- Runs in every DomU VM instance
- Publishes information about local processes on request from the management console
- Can start/stop programs within a VM
 - Basically acts as virtual superuser
- Used to control special benchmarking VM on each physical node for testing

DomU Process View



The screenshot shows a window titled 'DomU Process View' with a tree view on the left and a process list on the right. The tree view shows a hierarchy: Configs, Servers, Domains, and Test Domains. Under Test Domains, 'small-1' is selected. The process list on the right shows the following data:

PID	User	PR	NI	Virt	Res	SHR	S	%CPU	%MEM	TIME+	COMMAND
1017	root	25	0	2476	288	220	R	99.9	0.1	0:10.30	test3
1	root	16	0	2636	560	476	S	0.0	0.1	0:00.00	init
2	root	RT	0	0	0	0	S	0.0	0.0	0:00.00	migration/0
3	root	34	19	0	0	0	S	0.0	0.0	0:00.00	ksoftirqd/0
4	root	RT	0	0	0	0	S	0.0	0.0	0:00.00	watchdog/0
5	root	10	-5	0	0	0	S	0.0	0.0	0:00.01	events/0
6	root	10	-5	0	0	0	S	0.0	0.0	0:00.00	khelper
7	root	18	-5	0	0	0	S	0.0	0.0	0:00.00	kthread
8	root	17	-5	0	0	0	S	0.0	0.0	0:00.00	xenwatch
9	root	10	-5	0	0	0	S	0.0	0.0	0:00.00	xenbus
15	root	10	-5	0	0	0	S	0.0	0.0	0:00.00	kblockd/0
54	root	20	0	0	0	0	S	0.0	0.0	0:00.00	pdflush
55	root	15	0	0	0	0	S	0.0	0.0	0:00.00	pdflush
57	root	20	-5	0	0	0	S	0.0	0.0	0:00.00	aio/0

Automatic Migration Scenario

- 1 Dom0, 5 DomUs on 4 CPU machine
- Artificially drive up load on each VM
 - CPU intensive programs are started on each
 - Sample load over window of time (configurable)
 - Make heuristic decision to migrate a VM
- Server Agent requests help from neighbors
 - Neighbors willing/able to help respond, or not
 - Negotiates which VM to migrate
 - Invokes Xen live migration via Xen API
 - Remap network, reroute file system traffic

Xentop View of Domains

```

root@ferby: ~
xentop - 13:03:18 Xen 3.0.3-0
6 domains: 1 running, 5 blocked, 0 paused, 0 crashed, 0 dying, 0 shutdown
Mem: 16710696k total, 3874204k used, 12836492k free CPUs: 4 @ 2190MHz

```

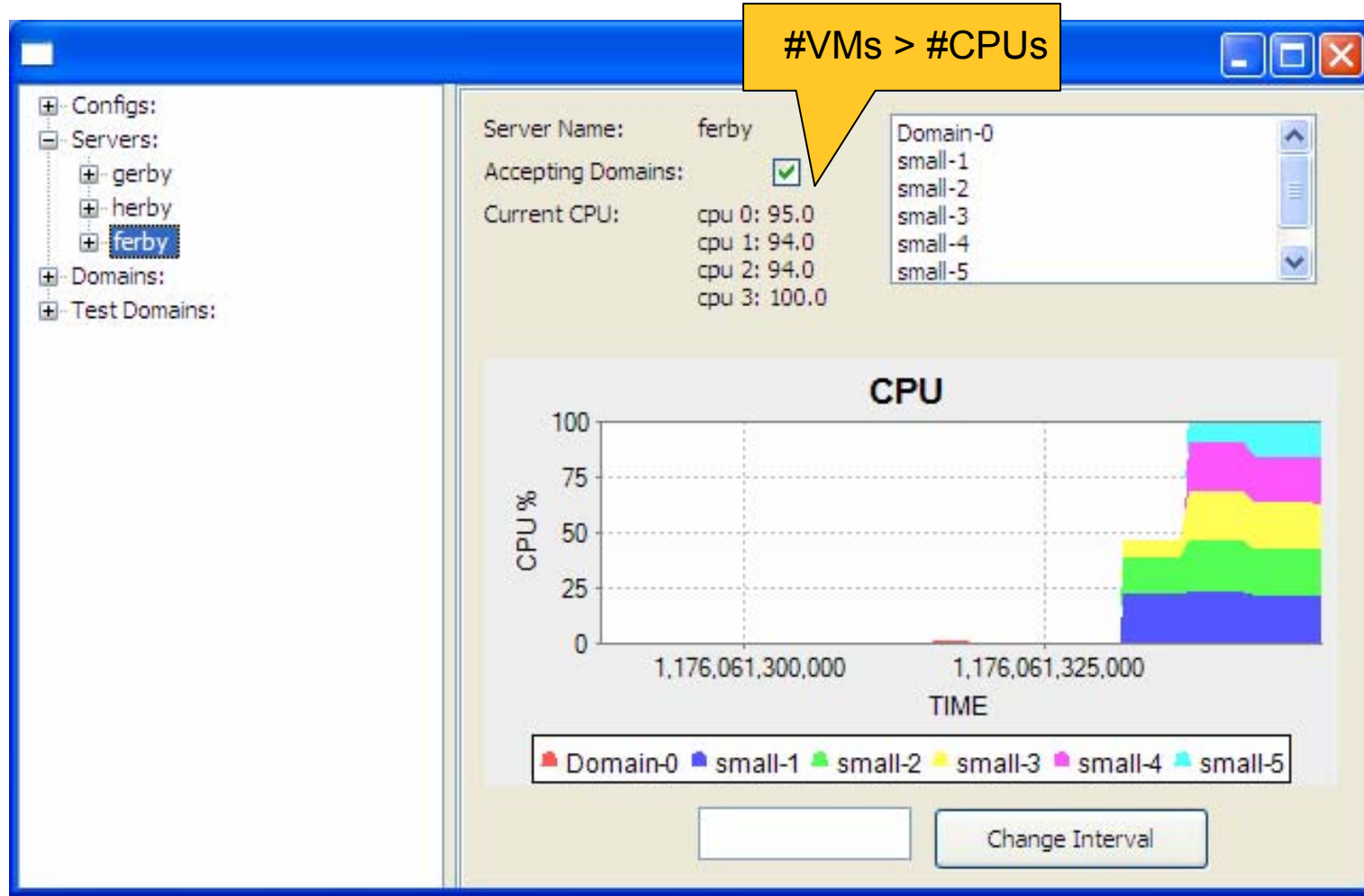
NAME	STATE	CPU(sec)	CPU(%)	MEM(k)	MEM(%)	MAXMEM(k)	MAXMEM(%)	VCPUS	NETS
Domain-0	-----r	14034	1.6	1047428	6.3	no limit	n/a	4	4
0	51319	0	0	0	0				
VCPUs(sec):		0:	2131s	1:	5431s	2:	6149s	3:	321s
small-1	--b---	195	0.0	522836	3.1	532480	3.2	1	1
53	44249	2	0	53	3846	0			
VCPUs(sec):		0:	195s						
small-2	--b---	41	0.0	522860	3.1	532480	3.2	1	1
28	23404	2	0	0	1588	0			
VCPUs(sec):		0:	41s						
small-3	--b---	278	0.0	522768	3.1	524288	3.1	1	1
91	48928	2	0	1111	2739	0			
VCPUs(sec):		0:	278s						
small-4	--b---	280	0.0	522784	3.1	524288	3.1	1	1
91	48907	2	0	1149	2841	0			
VCPUs(sec):		0:	280s						
small-5	--b---	319	0.0	522856	3.1	524288	3.1	1	1
90	48904	2	0	1125	4566	0			
VCPUs(sec):		0:	319s						

```

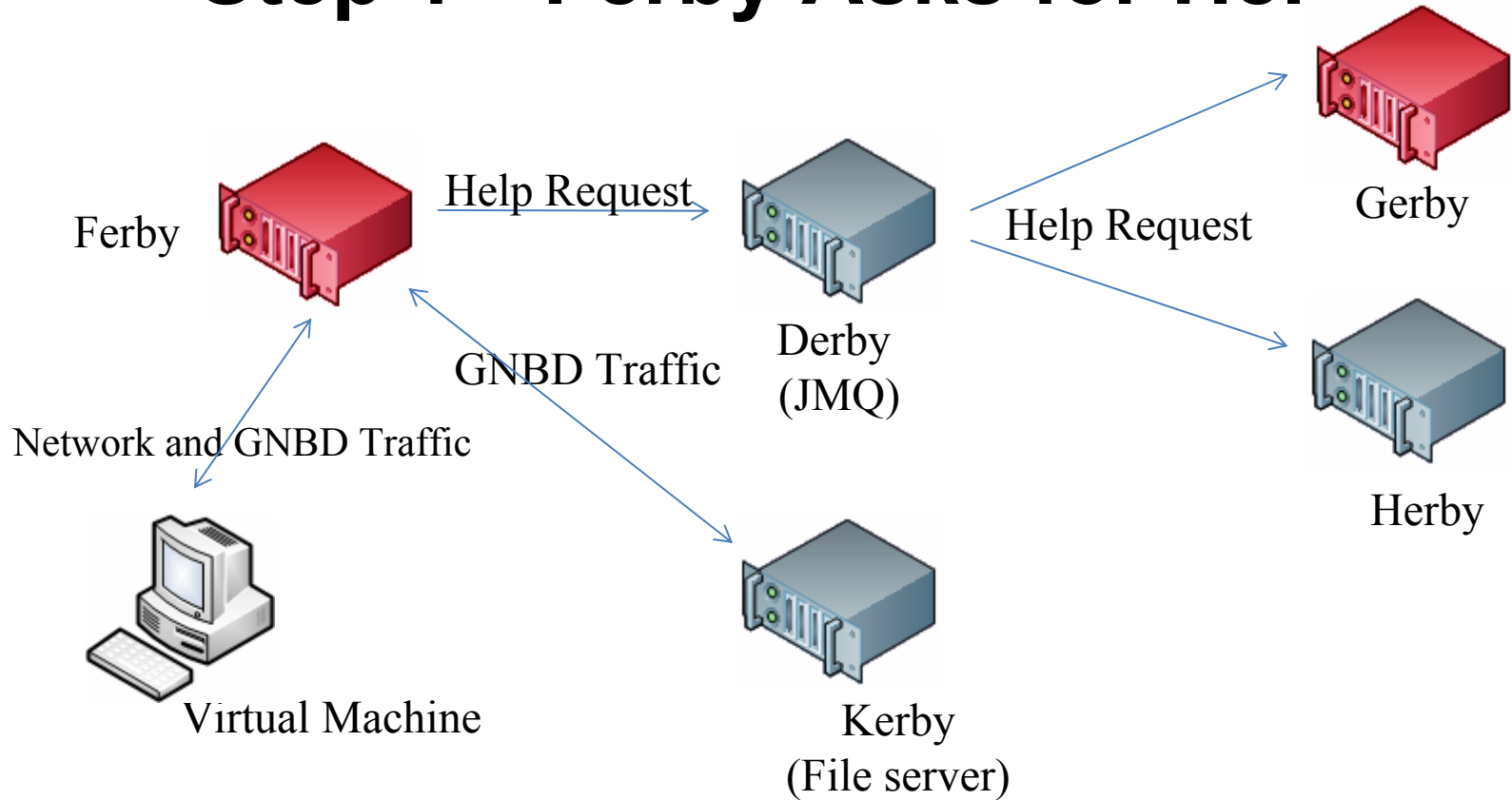
Delay Networks vBds VCPUs Repeat header Sort order Quit

```

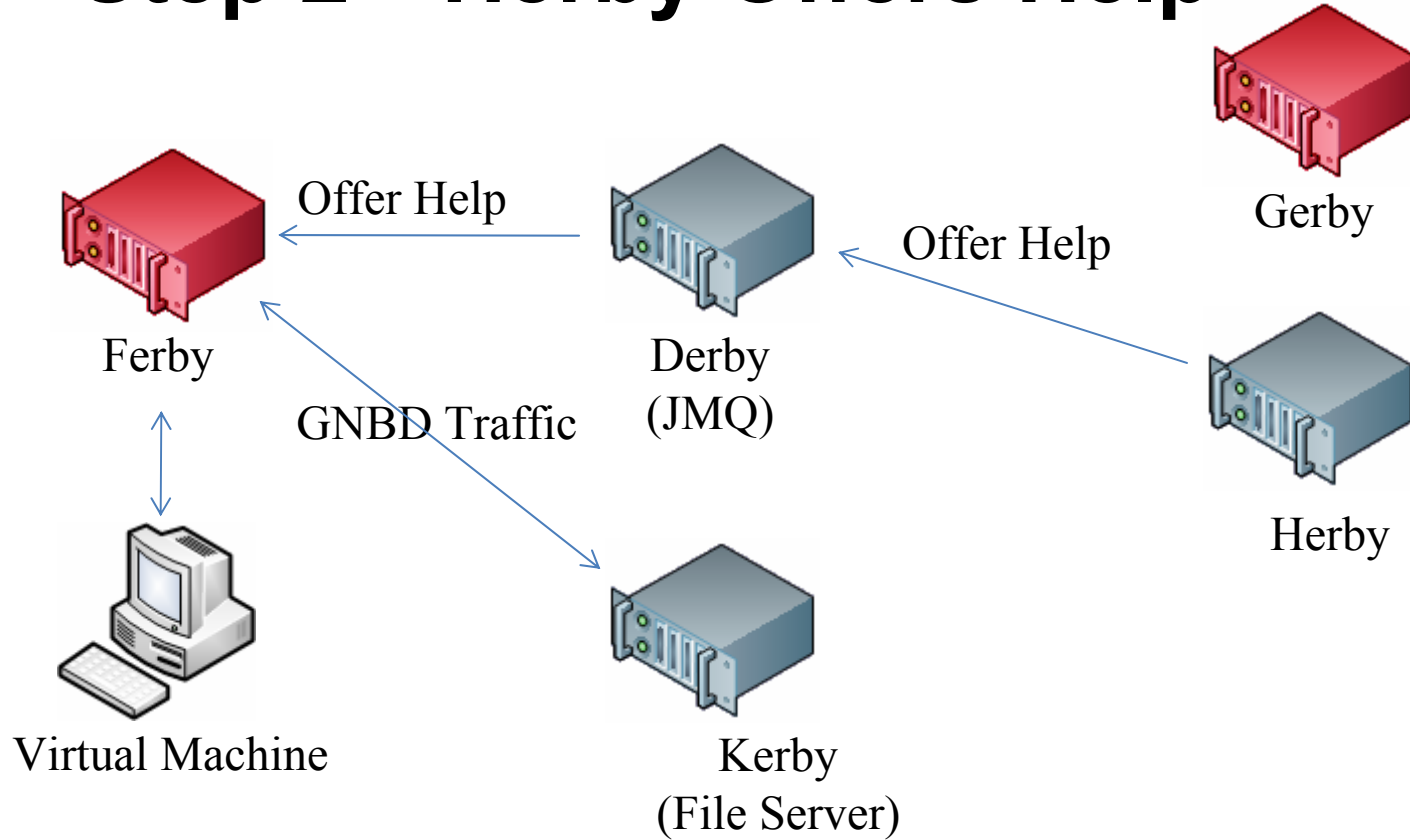
Ferby: 4 CPUs, 5 DomU VMs



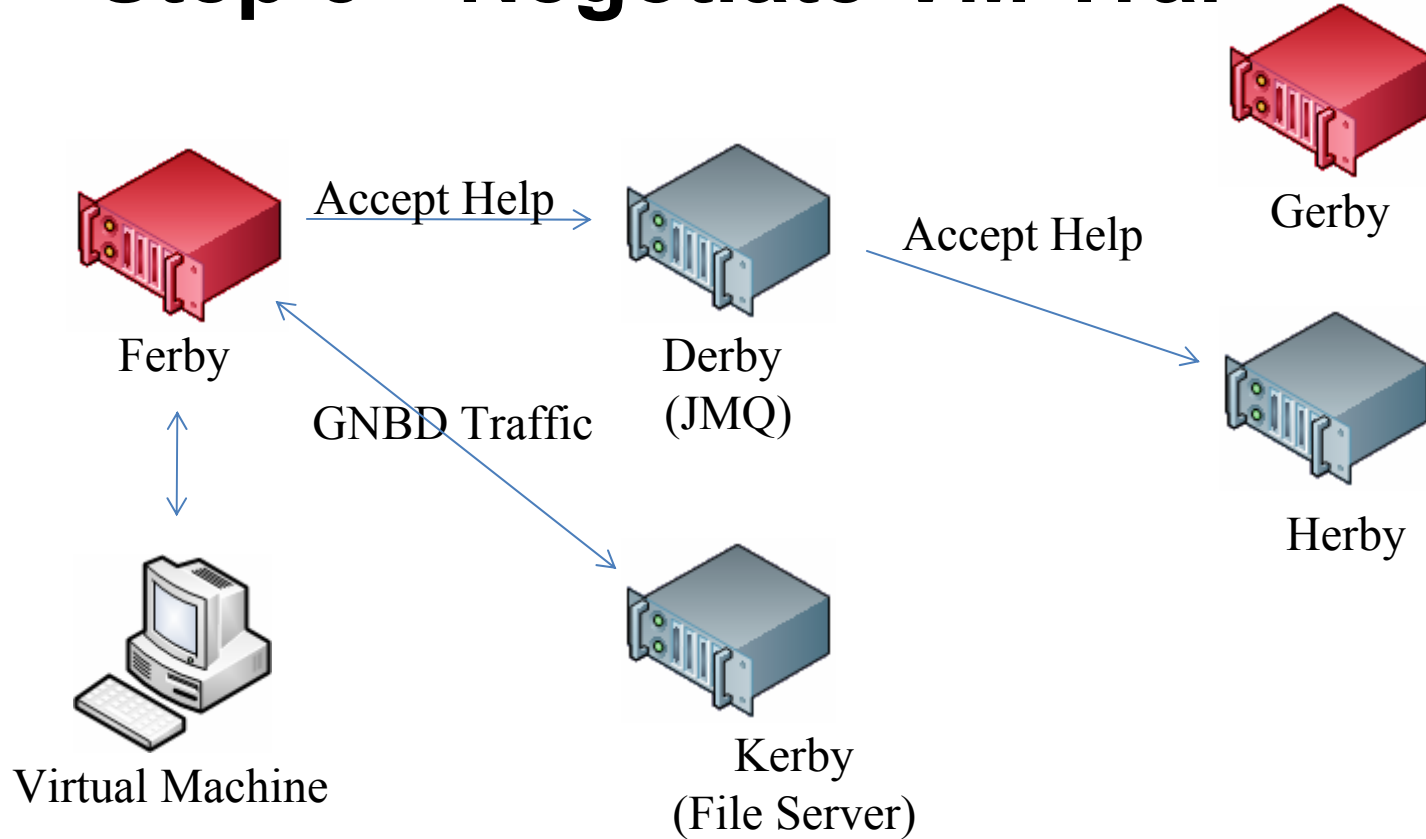
Step 1—Ferby Asks for Help



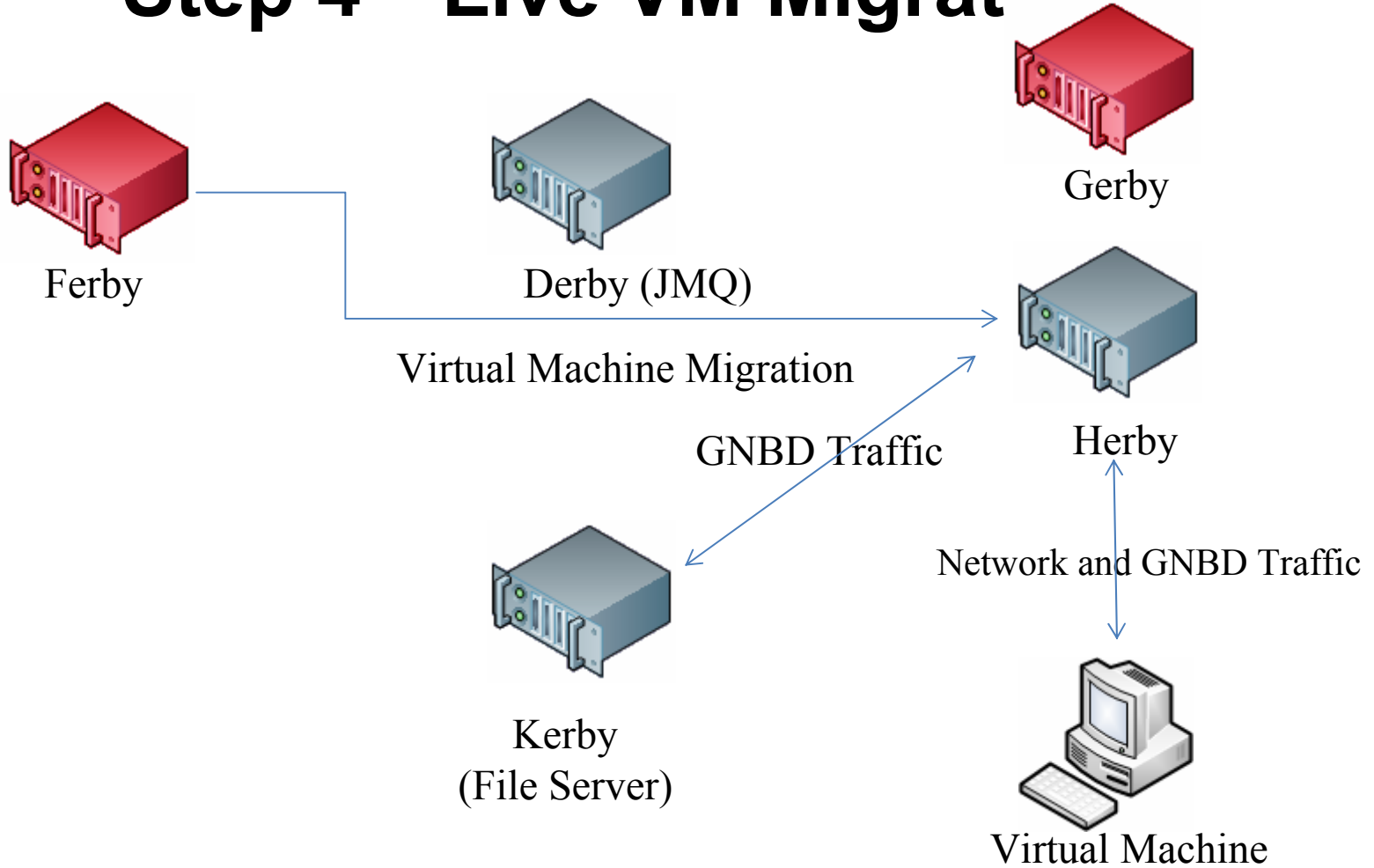
Step 2—Herby Offers Help



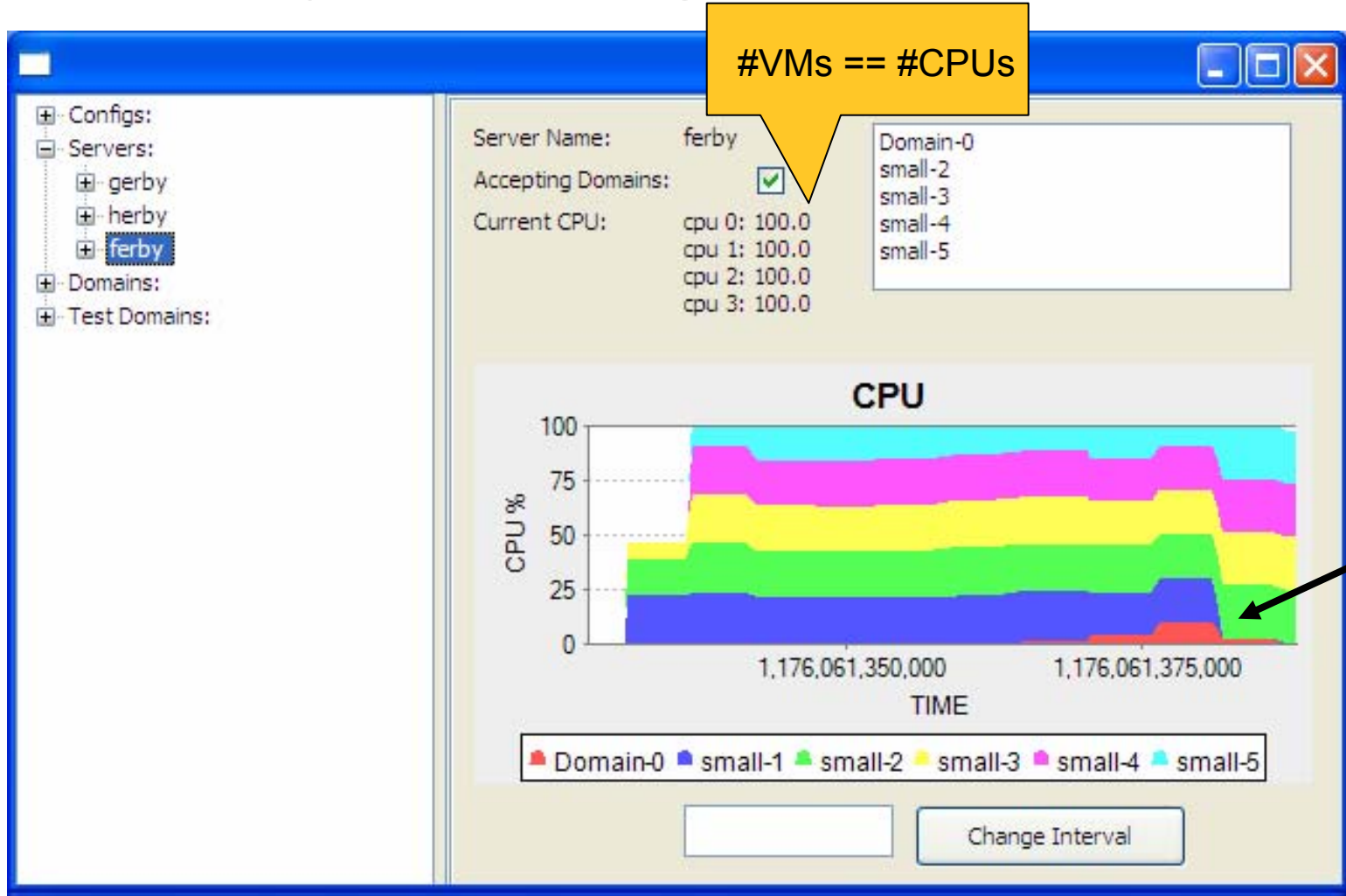
Step 3—Negotiate VM Transfer



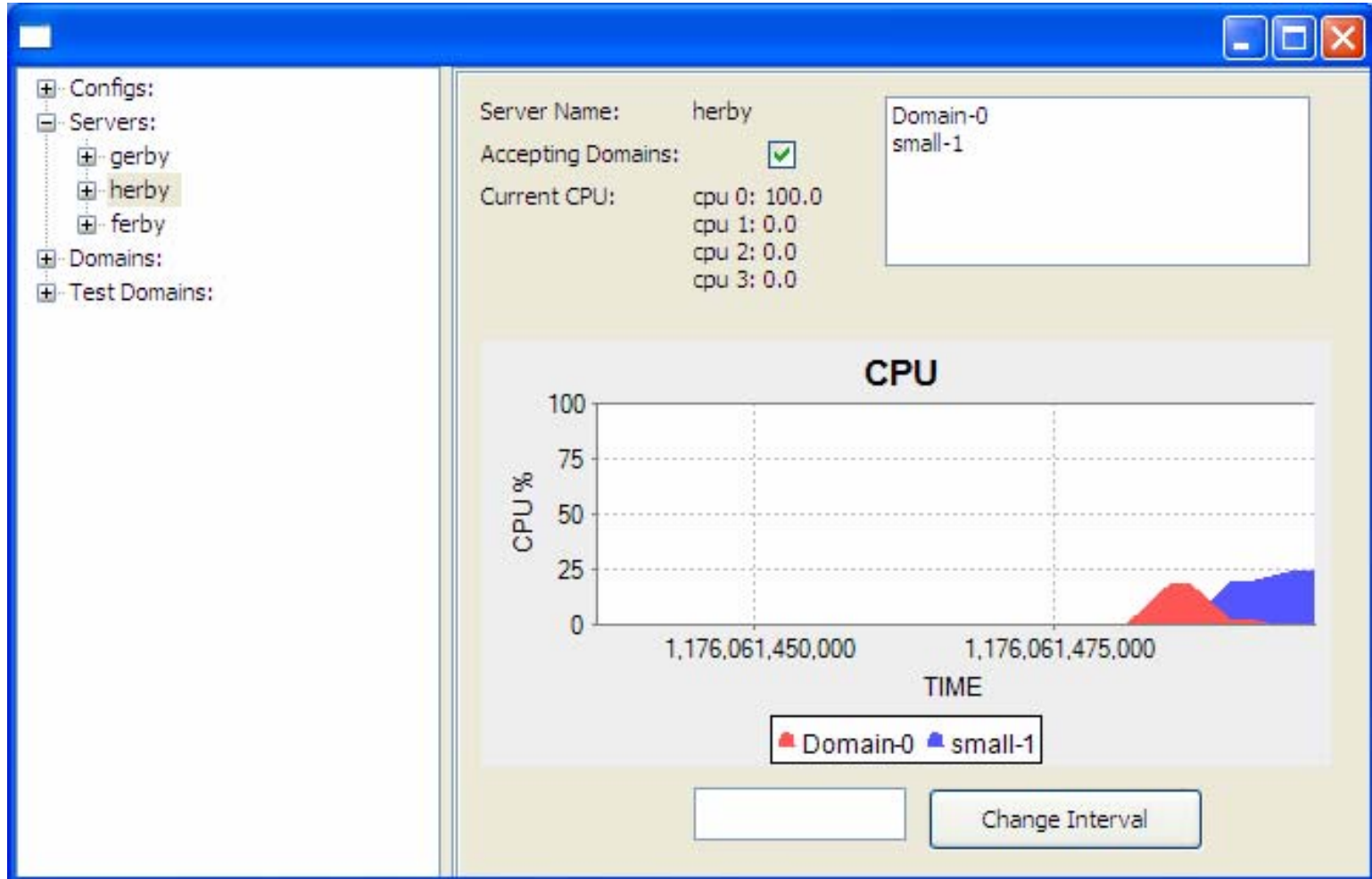
Step 4—Live VM Migration



Ferby After Migration



Herby After Migration



Observations

- We can virtualize our computing labs and significantly reduce administration time and cost
 - Software can scale to manage a few hundred virtual machines, just add more boxes
 - Enable VM migration to student laptops/USB sticks
 - Security issues are paramount
 - More work is required to ensure VMs are not “ganged” into DDoS attack bots
 - Message signing and encryption for agent messaging to prevent rogue agent messaging
- OpenSolaris operating system will enable more sophisticated resource management, QoS, and network virtualization

Future Work

- Add additional guest operating systems to automatic configuration
 - Allow web-based user selection of packages
- Extend migration algorithm to allow for priorities
 - Take advantage of enhanced API in Xen 3.0.4
- Use OpenSolaris OS x64 as Dom0 on all machines
 - Use the new Sun multithreaded 10 GbE networking cards
 - Take advantage of Solaris OS Virtual NIC (VNIC)
 - Manage network bandwidth allocation and QoS per virtual machine: Project Crossbow
 - <http://www.opensolaris.org/os/project/crossbow/>

Special Thanks

- John Fowler, EVP, Sun Systems Group
 - For personally donating the hardware
- Jeff Jackson, VP and Bill Franklin, Senior Director, both in Solaris OS Engineering
 - For providing research funding
- Tim Marsland, Sun Distinguished Engineer
 - For driving OpenSolaris OS x64 and supporting Xen



Q&A

lavender@cs.utexas.edu





QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Xen and the Art of Distributed Virtual Machine Management

Dr. Greg Lavender
Mr. Adam Zacharski

Department of Computer Sciences
The University of Texas at Austin

Session TS-1743



Extra Slides



Simple DomU Configuration

```
1:    kernel = "/boot/vmlinuz-2.6-xen"  
2:    disk = ['phy:/dev/gnbd/small-1-disk,sda1,w',  
            'phy:/dev/gnbd/small-1-swap,sda2,w' ]  
3:    memory = 512  
4:    name = 'small-1'  
5:    vif = [ 'bridge=xenbr0' ]  
6:    root = "/dev/sda1 ro"
```