



Enterprise Grade Open Source Virtualization


Simon Crosby, XenSource Inc



www.getxen.org




**The Perfect Storm:
x86 Server Virtualization**



x86 server sprawl growing, and utilization rates low
 +
 x86 server spending 47% of total server market (i.e., cheap but not free)
 +
 Most applications are small, and Moore's law is outpacing single application growth
 +
 Technology to virtualize server resources becoming mainstream
 +
 CIO focus on infrastructure efficiency and agility growing (CIO survey = #2)
 =
The hottest trend in x86 servers is virtualization becoming a default by 2009

Gartner



The Analysts' Take

By 2009, there will be three competitive hypervisor architectures: VMware ESX Server, Xen and Microsoft's hypervisor (0.8 probability).

Gartner

1/26/06 Xen: Enterprise Grade Open Source Virtualization 3



We're Hiring!

- Kernel/systems engineers (deeeep x86)
 - Linux or Windows
 - Devices, storage, networking, VT, SMP optimization
- Packagers
- Sr QA
- Sales Engineers
- In Palo Alto, NYC and Cambridge, UK

1/26/06 Xen: Enterprise Grade Open Source Virtualization 4

What's All the Fuss About?

1/26/06
Xen: Enterprise Grade Open Source Virtualization
5

Problem: Success of Scale-out

- “OS+app per server” provisioning leads to server sprawl
- Server utilization rates <10%
- Expensive to maintain, house, power, and cool
- Slow to provision, inflexible to change or scale
- Poor resilience to failures

1/26/06
Xen: Enterprise Grade Open Source Virtualization
6

XenSource Delivers Virtualization Value

- Consolidation:** fewer servers slashes CapEx and OpEx
- “Instant on” provisioning:** any app on any server, any time
- Higher utilization:** make the most of existing investments
- Robustness** to failures and “auto-restart” of VMs on failure

1/26/06
Xen: Enterprise Grade Open Source Virtualization
7

Result: Lower CapEx and OpEx

“Xen and XenOptimizer from XenSource allow us to consolidate servers and truly enable utility computing.”
CTO, F50 Financial Services


Year	Servers
2005	15,600
2006	15,360
2007	12,000
2008	10,000
2009	8,000
2010	13,380



















Category	Value
Deployment	\$52M
Operators	\$52M
Power/Cooling	\$11M
H/W & S/W	\$21M
Total	\$136M



I don't know anyone involved with virtualization applications who are not taking Xen seriously.



Tony Iams
Senior Analyst, iDEAS International

1/26/06
Xen: Enterprise Grade Open Source Virtualization
8

	<h2>Who is XenSource?</h2>
<ul style="list-style-type: none"> Founded by Xen creators in 2005 <ul style="list-style-type: none"> Investors: Kleiner Perkins, Sevin Rosen, Accel, NEA First revenue in Q4 2005 Offices in New York, Palo Alto, Cambridge UK <p>Drive ubiquitous adoption of Xen</p> <ul style="list-style-type: none"> Lead community development of Xen, the fastest & most secure virtualization technology Deliver multi-OS virtualization solutions Trusted partner to leading OEMs and ISVs 	
1/26/06	Xen: Enterprise Grade Open Source Virtualization 9

	<h2>Our Community Partners</h2>
<div data-bbox="950 493 1339 535">    </div> <p>Operating System Vendors</p> <hr/> <div data-bbox="876 588 1380 682">       </div> <div data-bbox="885 651 1364 682">     </div> <p>Platforms / Storage</p> <hr/> <div data-bbox="885 735 1364 798">     </div> <p>Processors & I/O</p> <p><small>* Logos are registered trademarks of their owners</small></p>	
1/26/06	Xen: Enterprise Grade Open Source Virtualization 10

	<h2>Xen 3.0 Headline Features</h2>
<div data-bbox="235 1323 324 1375">  </div> <p>15K Downloads since 12/05 release</p> <ul style="list-style-type: none"> Up to 32-way SMP guest OSes Uses Intel® VT-x and AMD hardware virtualization to support all OSes PAE and x86/64 support Itanium (IA64) architecture and VT-i Superb performance - eg: 0.1% - 3.5% overhead for SPECjbb <p>Xen Projects Under Way:</p> <ul style="list-style-type: none"> Para-virtualized Solaris 10 on Xen 3.0 (Sun) Power5 (IBM) SPARC Port (Sun) DMTF CIM hypervisor management (XS, Novell, IBM) 	
1/26/06	Xen: Enterprise Grade Open Source Virtualization 11

	<h2>Catalyzes Adoption of Virtualization</h2>
<div data-bbox="885 1365 1071 1512">  </div> <ul style="list-style-type: none"> Superb performance means it can go into heavy duty production Xen Open Source means it can be universally improved and adopted Affordable, so customers can deploy virtualization broadly Backed by all major IT vendors New opportunities for savings in systems management, provisioning, fault management 	
1/26/06	Xen: Enterprise Grade Open Source Virtualization 12

Unlocking Platform Innovation

Security: Intel LT & AMT, AMD SEM, IBM TPM

- Building blocks for Trusted Computing infrastructure
- Supports "virtual" TPM 1.1 & 1.2 for secure boot & OS services
- Integrated IDS & security features

Multi-core Processors & Hyper-threading

- Load balances parallel execution units capable of running SMP workloads
- Hides complexity from guests

Virtualization: Intel VT, AMD Pacifica, IBM Power

- Hardware support accelerates virtualization
- XE is the industry's first supported product for Intel & AMD virtualization
- Virtualization "on the bare metal"

1/26/06
Xen: Enterprise Grade Open Source Virtualization
13

Enhances Enterprise Security

Security features insulate OSes and applications from attack

- DoS proof VMs due to superb resource partitioning
- Support for (virtualized) TPM 1.1, 1.2 and trusted boot capability
- Implements security policy outside the guests
- Core Xen is under 50K LOC - scrutinized by the security community
- Foundation for XenSource, IBM, Intel Multi-Level-Secure Architecture

1/26/06
Xen: Enterprise Grade Open Source Virtualization
14

Virtualization Before Xen

Existing Virtualization Products

- A (typically proprietary) microkernel / OS under your OS
- Full virtualization requires binary patching of the OS at runtime
- Microkernel contains device drivers
- Emulates native chipset, so significant performance overhead
- Separate maintenance schedule for microkernel (& drivers) and the virtualized OS (& drivers)
- Vulnerable to driver failure
- Large code base
- But it runs unmodified OS images

1/26/06
Xen: Enterprise Grade Open Source Virtualization
15

Inside Xen, and Why It's so Cool

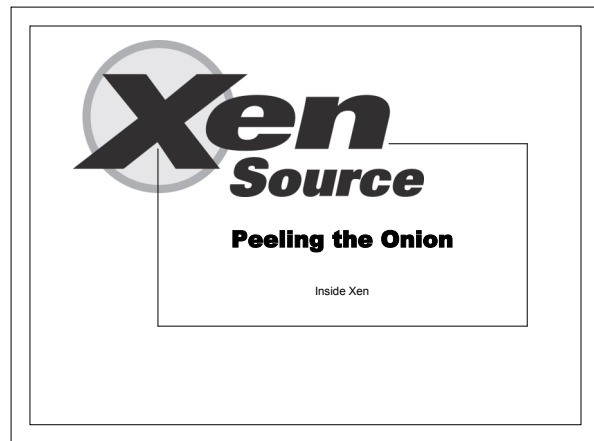
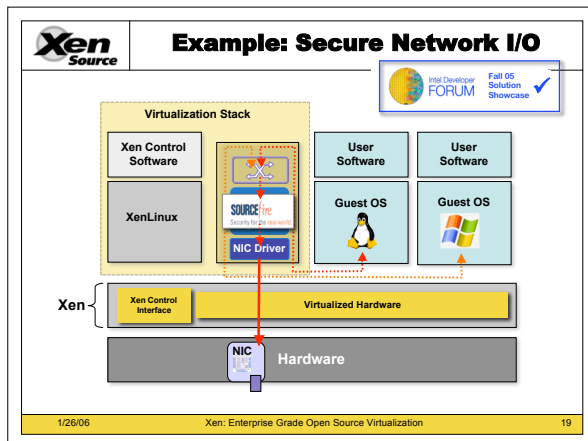
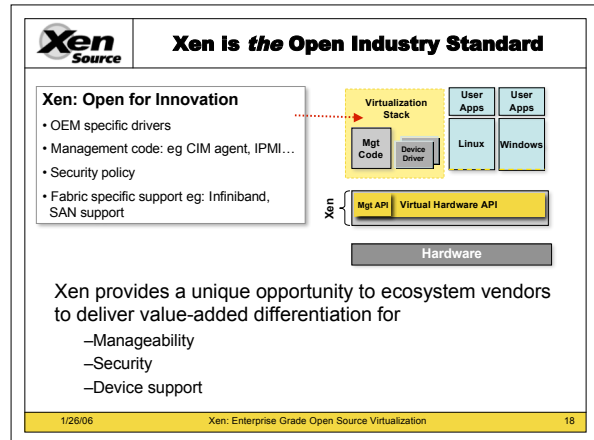
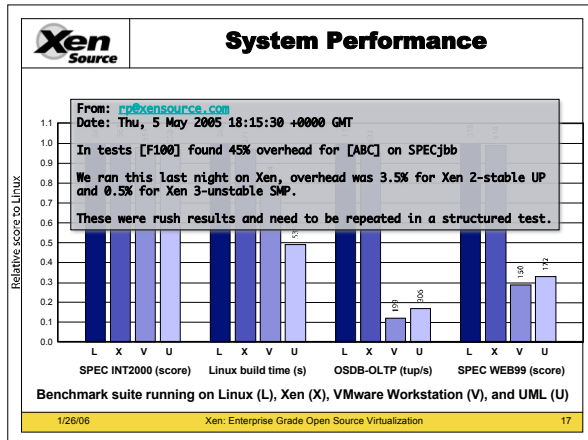
Xen: A Para-virtualizing Hypervisor


- Virtualizes (only) the base platform
 - CPU
 - MMU & Memory
 - Low level interrupts
- Small, reliable, efficient, trusted base-platform personality
- Guest OS co-operates with Xen
- Near native performance
- Lean and getting leaner (50 KLOC)
- Supports native Linux device drivers
- Separates the driver from the guest
- No separate maintenance schedule
- Runs on x86_64, IA64, Power 5

Xen

Free from your favorite Linux distro and www.getxen.org/

1/26/06
Xen: Enterprise Grade Open Source Virtualization
16





Para-Virtualization in Linux

Arch xen_x86 : like x86, but Xen hypercalls required for privileged operations

- Avoids binary rewriting
- Minimize number of privilege transitions into Xen
- Modifications relatively simple and self-contained

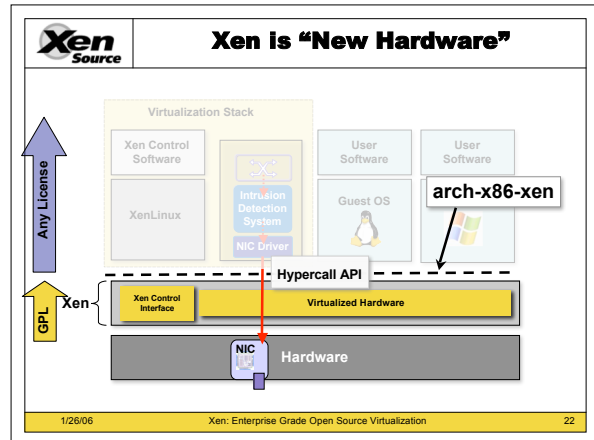
Modify kernel to understand virtualised environment


- Wall-clock time vs. virtual processor time
 - Xen provides both types of alarm timer
- Expose real resource availability
 - Enables OS to optimise behaviour

1/26/06

Xen: Enterprise Grade Open Source Virtualization

21





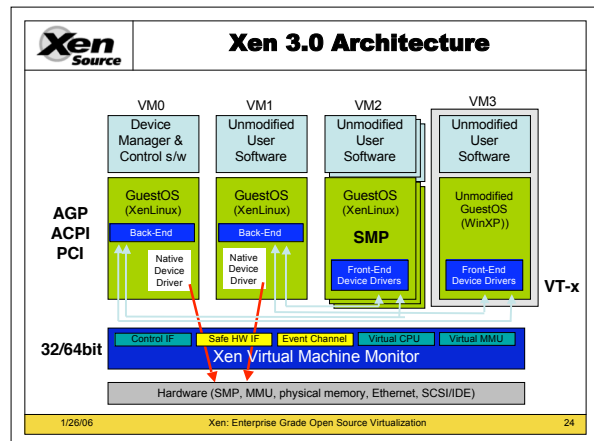
x86 CPU virtualization

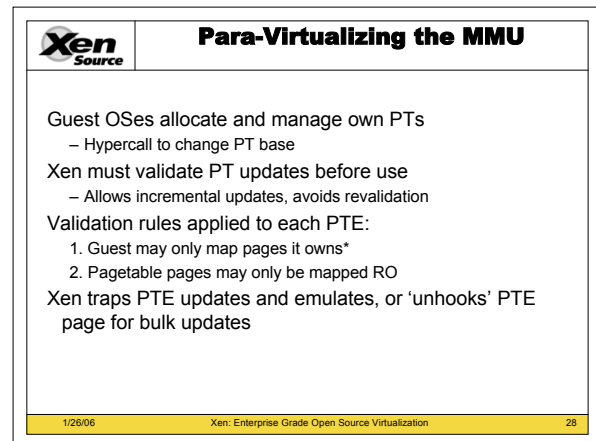
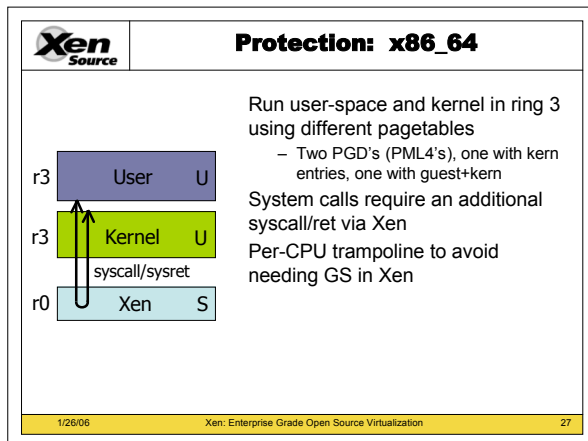
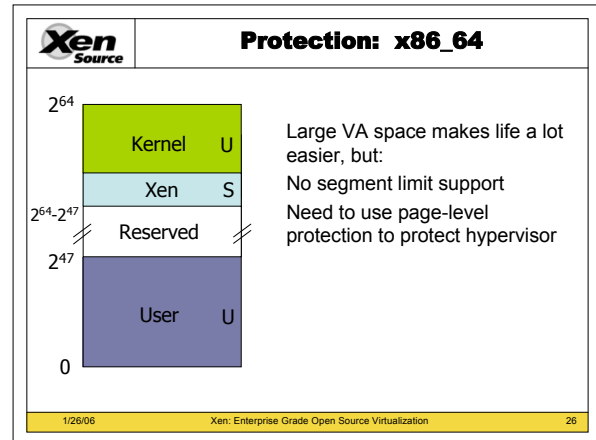
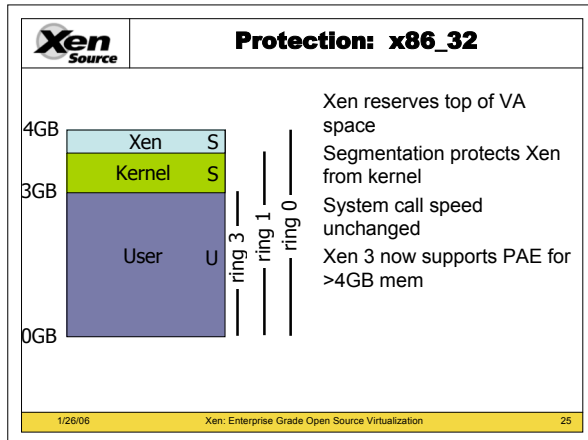
- Xen runs in ring 0 (most privileged)
- Ring 1/2 for guest OS, 3 for user-space
 - GPF if guest attempts to use privileged instr
- Xen lives in top 64MB of linear addr space
 - Segmentation used to protect Xen as switching page tables too slow on standard x86
- Hypercalls jump to Xen in ring 0
- Guest OS may install 'fast trap' handler
 - Direct user-space to guest OS system calls
- MMU virtualisation: shadow vs. direct-mode


1/26/06

Xen: Enterprise Grade Open Source Virtualization

23







SMP Guest Kernels

Xen extended to support multiple VCPUs

- Virtual IPI's sent via Xen event channels
- Currently up to 32 VCPUs supported


Simple hotplug/unplug of VCPUs

- From within VM or via control tools
- Optimize one active VCPU case by binary patching spinlocks

1/26/06

Xen: Enterprise Grade Open Source Virtualization

29



I/O Architecture

Xen *IO-Spaces* delegate guest OSes protected access to specified h/w devices

- Virtual PCI configuration space
- Virtual interrupts
- (Need IOMMU for full DMA protection)

Devices are virtualised and exported to other VMs via *Device Channels*


- Safe asynchronous shared memory transport
- 'Backend' drivers export to 'frontend' drivers
- Net: use normal bridging, routing, iptables
- Block: export any blk dev e.g. sda4, loop0, vg3

(Infiniband / Smart NICs for direct guest IO)

1/26/06

Xen: Enterprise Grade Open Source Virtualization

30



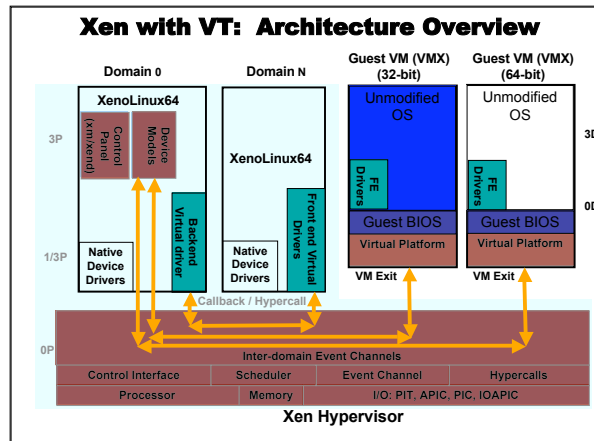
VT-x / Pacifca

- Enable Guest OSes to be run without paravirtualization modifications
- CPU provides traps for certain privileged instrs
- Shadow page tables used to provide MMU virtualization
- Xen provides simple platform emulation
 - BIOS, Ethernet, IDE and SCSI emulation
- Install paravirtualized drivers after booting for high-performance IO

1/26/06

Xen: Enterprise Grade Open Source Virtualization

31





x86_64

- Intel EM64T and AMD Opteron
- Requires different approach to x86 32 bit:
 - Can't use segmentation to protect Xen from guest OS kernels because there are no segment limits
 - Switch page tables between kernel and user
 - Not too painful thanks to Opteron TLB flush filter
 - Large VA space offers other optimisations
- Current design supports up to 8TB mem

1/26/06

Xen: Enterprise Grade Open Source Virtualization

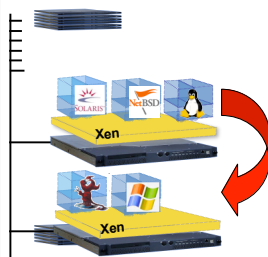
33



Xen's Live Relocation



VM Relocation : Motivation



VM relocation enables:

- High-availability
 - Machine maintenance
- Load balancing
 - Statistical multiplexing gain

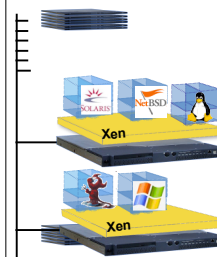
1/26/06

Xen: Enterprise Grade Open Source Virtualization

35



Assumptions



Networked storage

- NAS: NFS, CIFS
- SAN: Fibre Channel
- iSCSI, network block dev
- drdb network RAID

Good connectivity

- common L2 network
- L3 re-routing

1/26/06

Xen: Enterprise Grade Open Source Virtualization

36

Challenges

- VMs have lots of state in memory
- Some VMs have soft real-time requirements
 - E.g. web servers, databases, game servers
 - May be members of a cluster quorum
- **Minimize down-time**
- Performing relocation requires resources
 - **Bound and control resources used**

1/26/06
Xen: Enterprise Grade Open Source Virtualization
37

Relocation Strategy

Stage 0: pre-migration

Stage 1: reservation

Stage 2: iterative pre-copy

Stage 3: stop-and-copy

Stage 4: commitment

VM active on host A
Destination host selected
(Block devices mirrored)

Initialize container on target host

Copy dirty pages in successive rounds

Suspend VM on host A
Redirect network traffic
Synch remaining state

Activate on host B
VM state on host A released

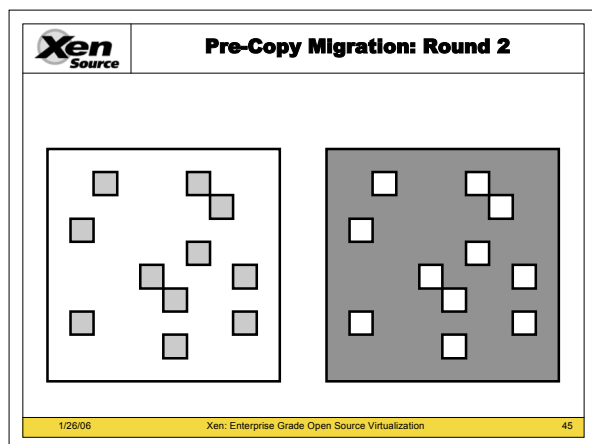
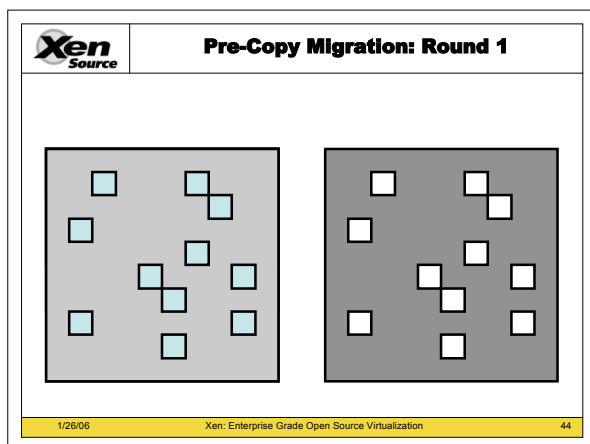
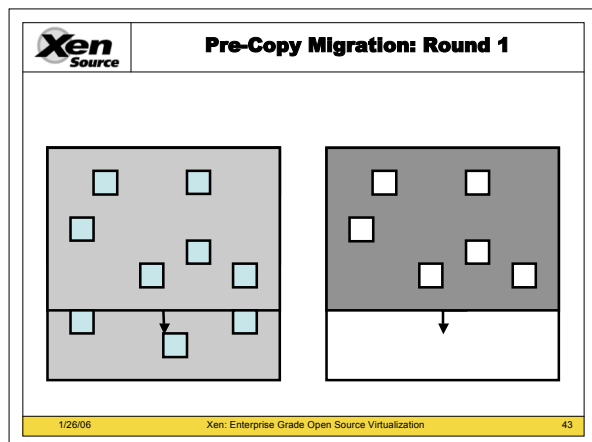
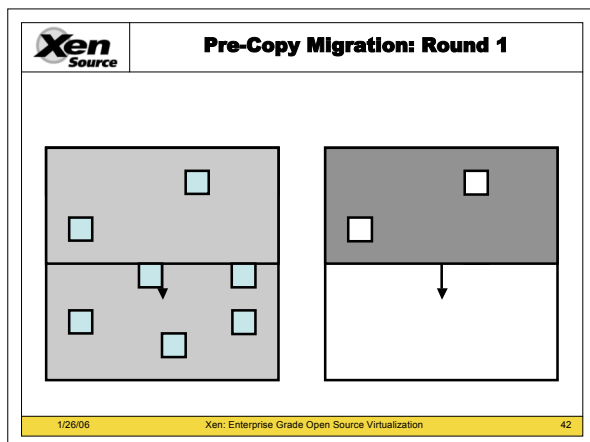
1/26/06
Xen: Enterprise Grade Open Source Virtualization
39

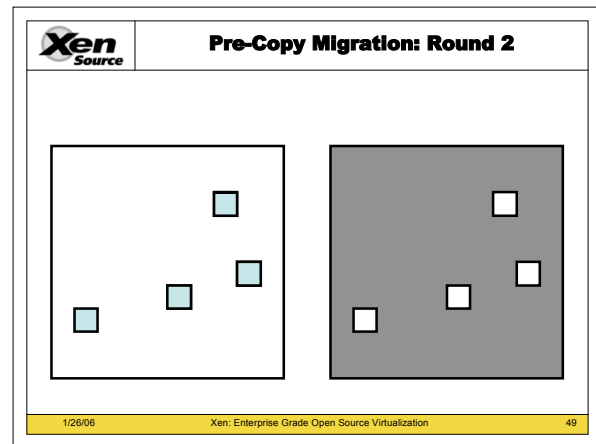
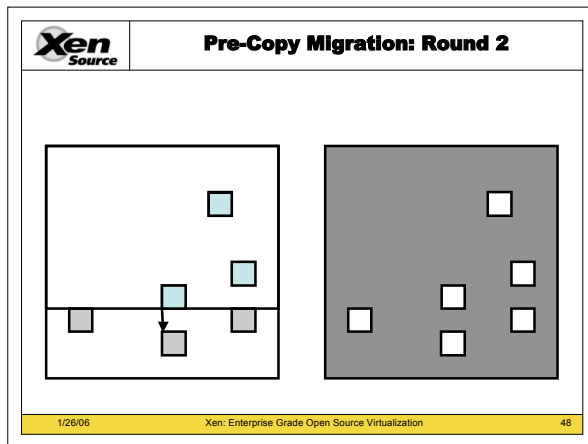
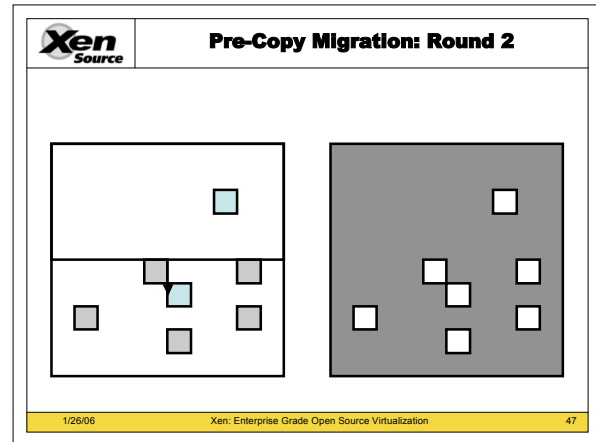
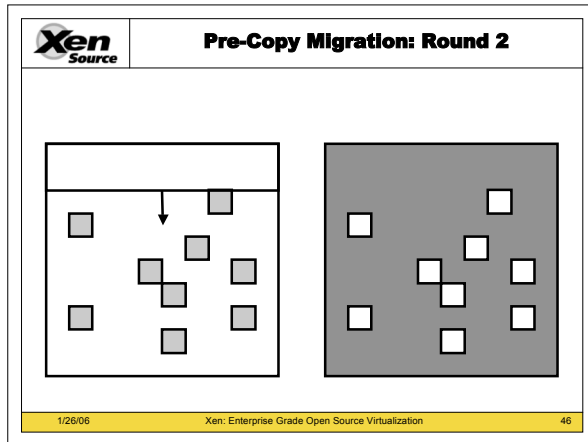
Pre-Copy Migration: Round 1


1/26/06
Xen: Enterprise Grade Open Source Virtualization
40

Pre-Copy Migration: Round 1

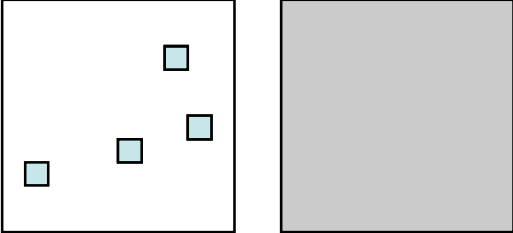
1/26/06
Xen: Enterprise Grade Open Source Virtualization
41








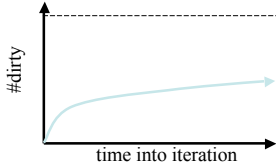
Pre-Copy Migration: Final



1/26/06
Xen: Enterprise Grade Open Source Virtualization
50




Page Dirtying Rate



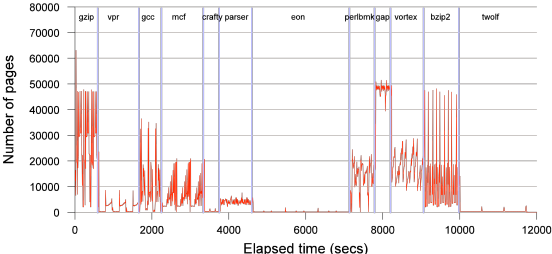
- Dirtying rate determines VM down-time
 - Shorter iters → less dirtying → shorter iters
 - Stop and copy final pages
- Application 'phase changes' create spikes

1/26/06
Xen: Enterprise Grade Open Source Virtualization
53




Writable Working Set

Tracking the Writable Working Set of SPEC CINT2000



1/26/06
Xen: Enterprise Grade Open Source Virtualization
54



Rate Limited Relocation

- Dynamically adjust resources committed to performing page transfer
 - Dirty logging costs VM ~2-3%
 - CPU and network usage closely linked
- E.g. first copy iteration at 100Mb/s, then increase based on observed dirtying rate
 - Minimize impact of relocation on server while minimizing down-time

1/26/06
Xen: Enterprise Grade Open Source Virtualization
57

