Xen and the Art of Virtualization

Paul Barham, Boris Dragovic, Keir Fraser, Steven Hand, Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, Andrew Warfield

University of Cambridge Computer Laboratory

Kyle Schutt

CS 5204

Virtualization

- Abstraction of hardware resources
- Virtual Machine Monitors (Hypervisors)
- Key Players
  - Xen
  - VMWare
  - Hyper-V (Windows Server Virtualization)
  - KVM (Kernel Virtual Machine)

Virtualization Issues

- Isolation
- Reliability
- Security
- Scalability
- Performance
- Heterogeneous

Xen Virtualization

- Open source
- Paravirtualization and full virtualization
- Domain0 and DomainU
- Small footprint
- Direct hardware access
- Privilege control

Source: http://xen.org/

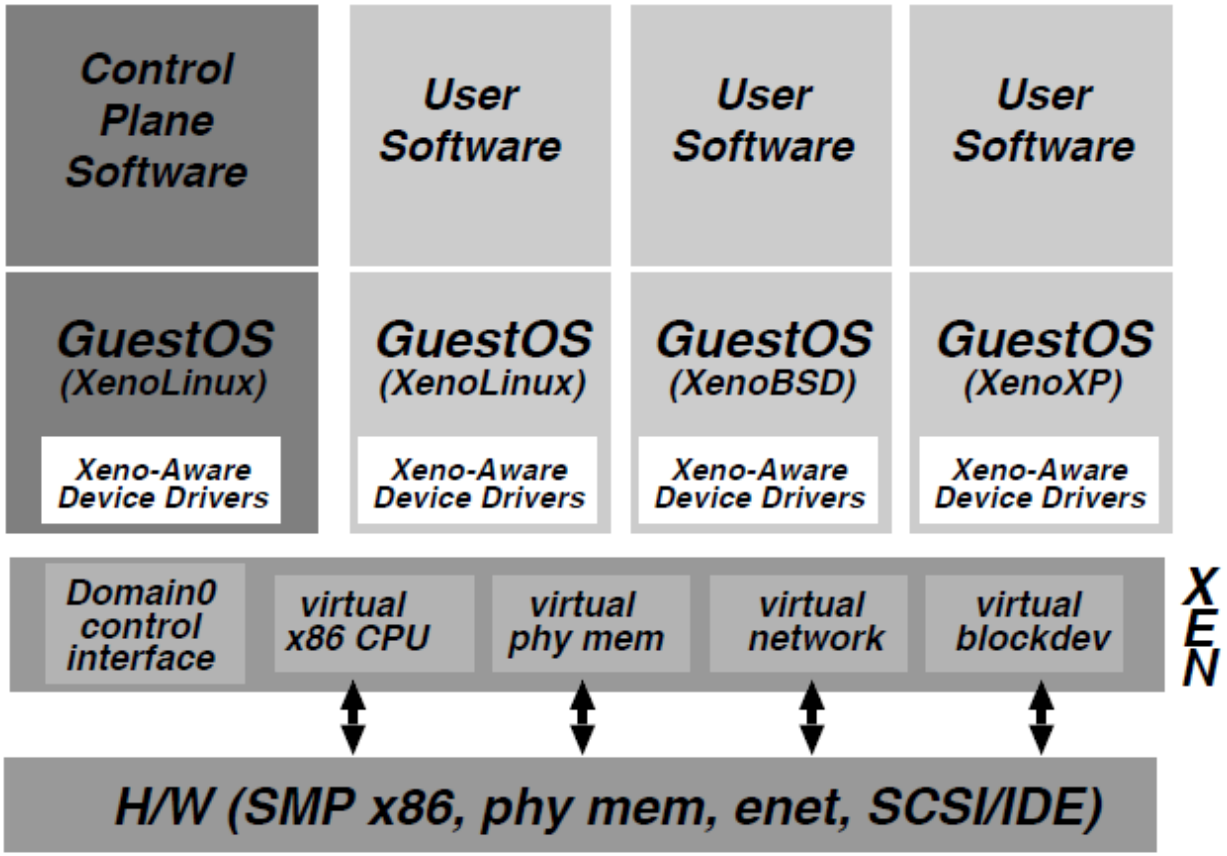
Overview

- Introduction
- Xen: Virtual Machine Monitor
- XenoLinux Evaluation
- Xen Current State
- Xen in Industry
- Xen Demo
- Discussion

Xen: Virtual Machine Monitor

- Hardware Layer
- x86 Paravirtualization
- Design Choices
  - Unmodified user application binaries
  - Full install of OSes
  - Paravirtualization – high performance and resource isolation
  - Transparent resource virtualization

Xen: Virtual Machine Monitor



Xen: VMM Approach Overview

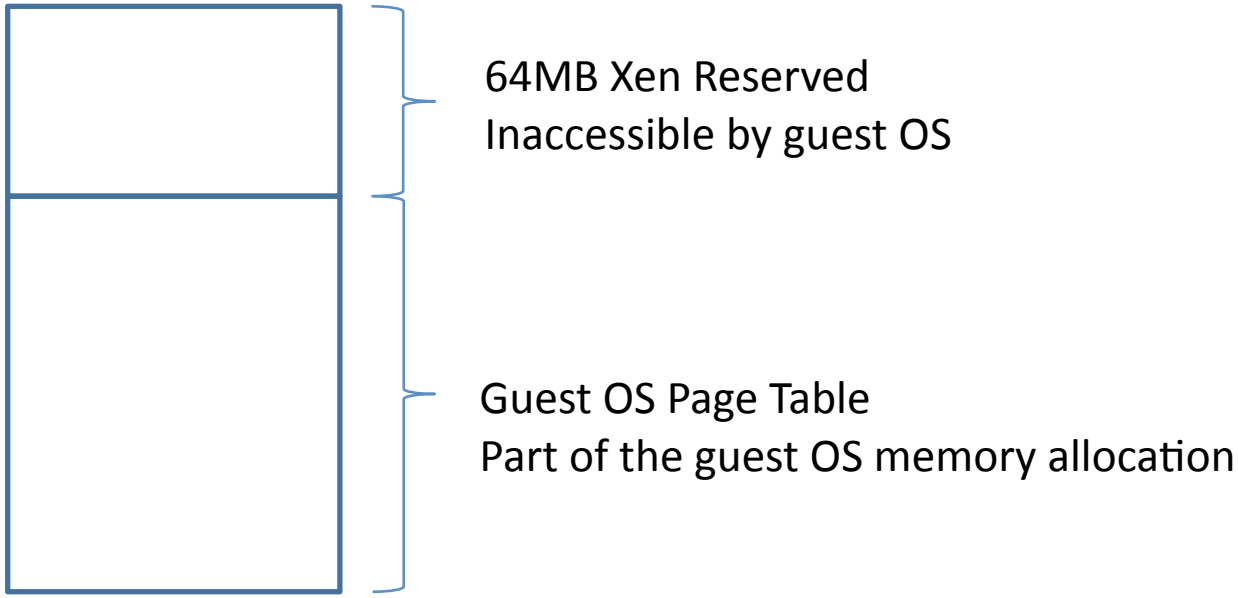
- x86 Specific Paravirtualization
- Data Transfers
- Intercommunication
- Porting Costs
- Control and Management
- Subsytems

Xen: VMM x86 Paravirtualization

- Memory Management
- CPU Scheduling
- Device I/O

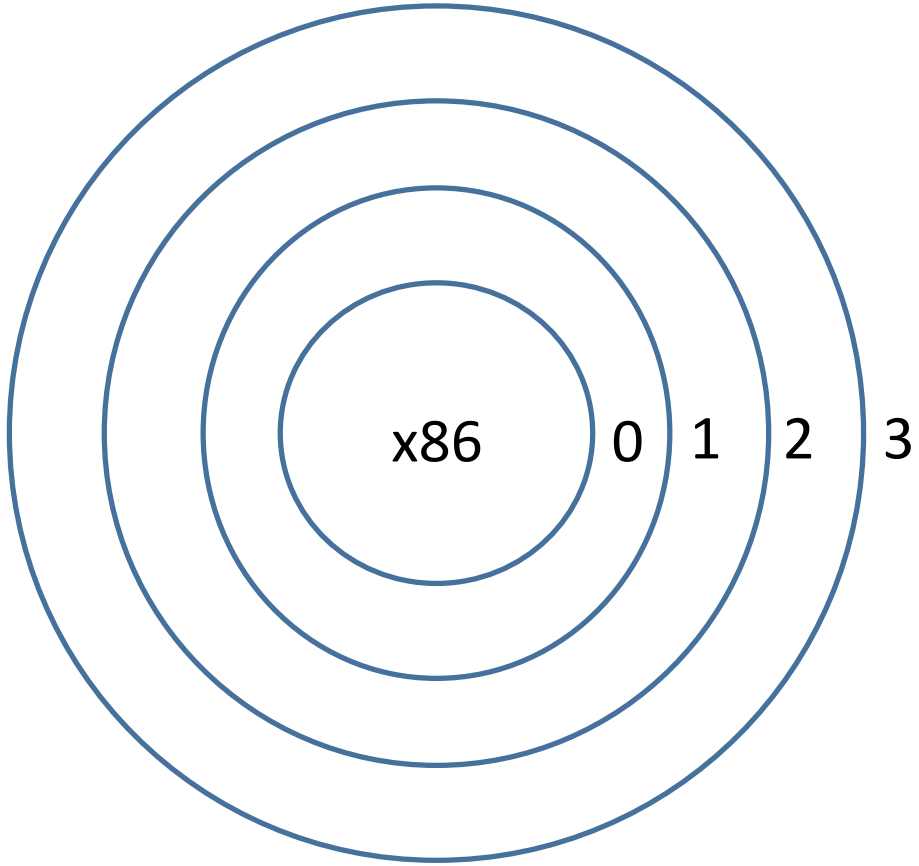
Xen: VMM x86 Memory Management

- Registers allocations with Xen
- Untagged vs. Software-managed TLB



Xen: VMM x86 CPU

- Privilege Levels
- Level 0
  - Typical OS
  - Xen Kernel
- Level 1
  - Guest OS w/ Xen
- Level 2
  - Unused
- Level 3
  - User Applications

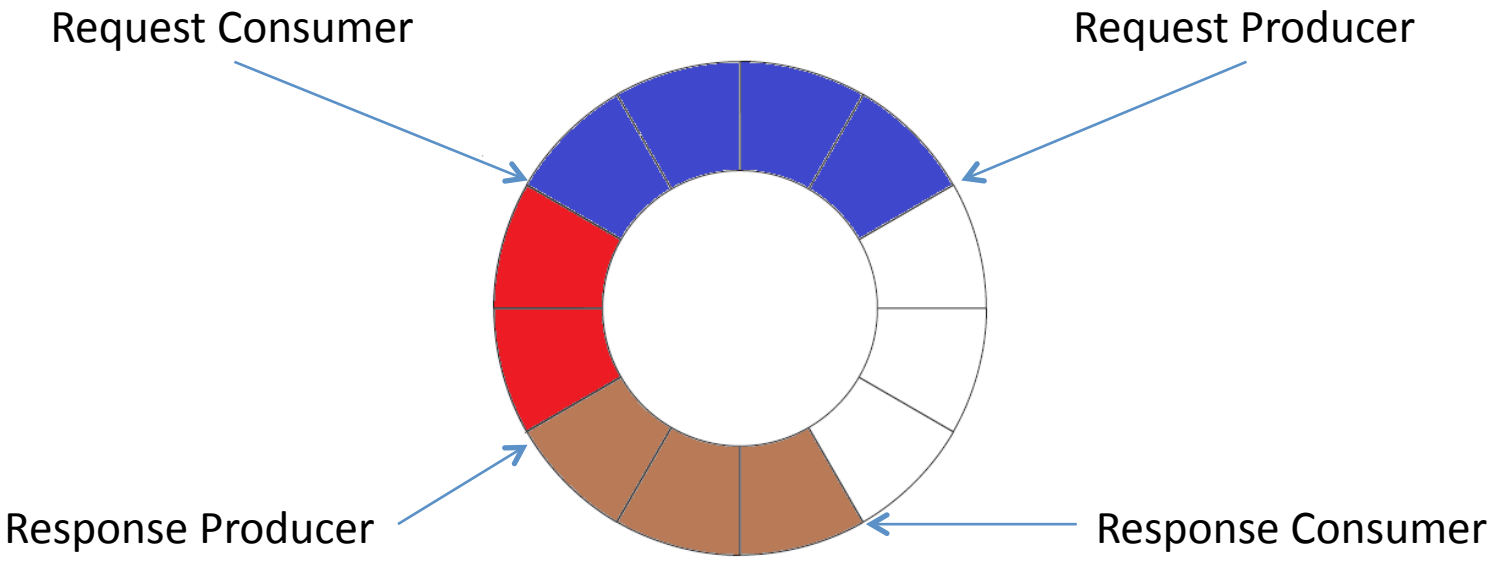


Xen: VMM x86 Device I/O

- Paravirtualize Devices
  - Abstraction
- Mediator
  - Validation
  - Channel links
- I/O Rings
  - Shared memory
  - Descriptor rings

Xen: VMM Data Transfers

- I/O Rings



Request queue - Descriptors queued by the VM but not yet accepted by Xen

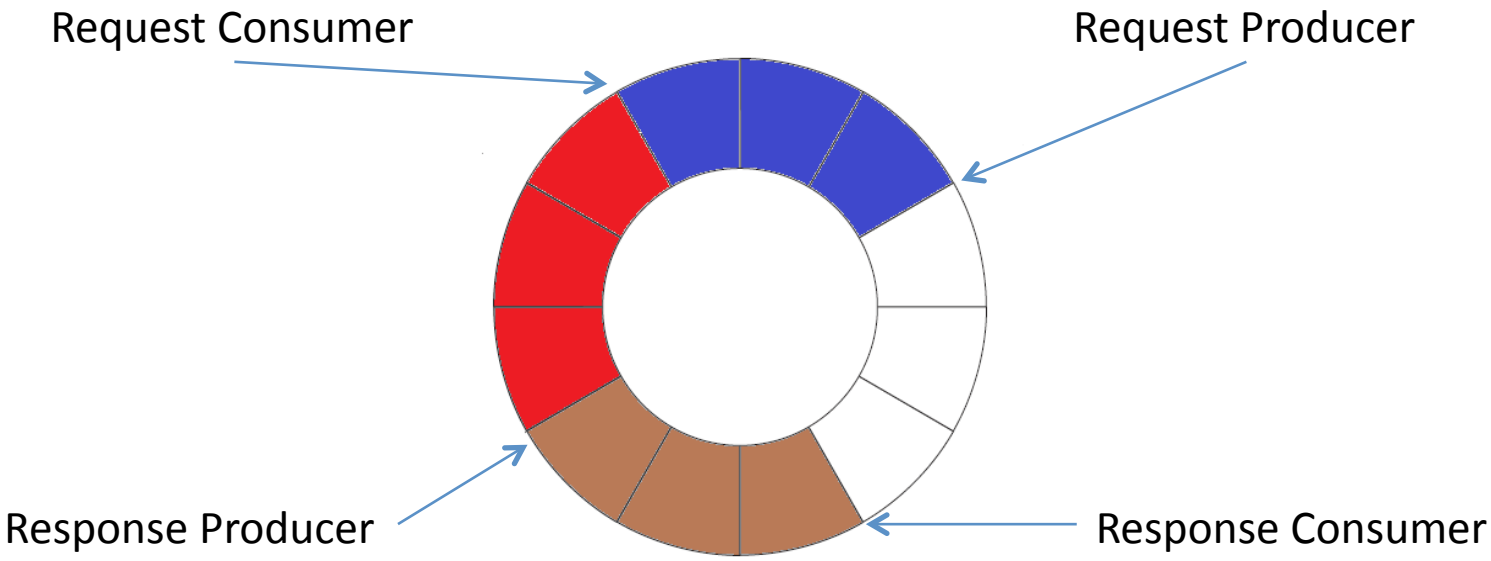
Outstanding descriptors - Descriptor slots awaiting a response from Xen

Response queue - Descriptors returned by Xen in response to serviced requests

Unused descriptors

Xen: VMM Data Transfers

- I/O Rings



Request queue - Descriptors queued by the VM but not yet accepted by Xen

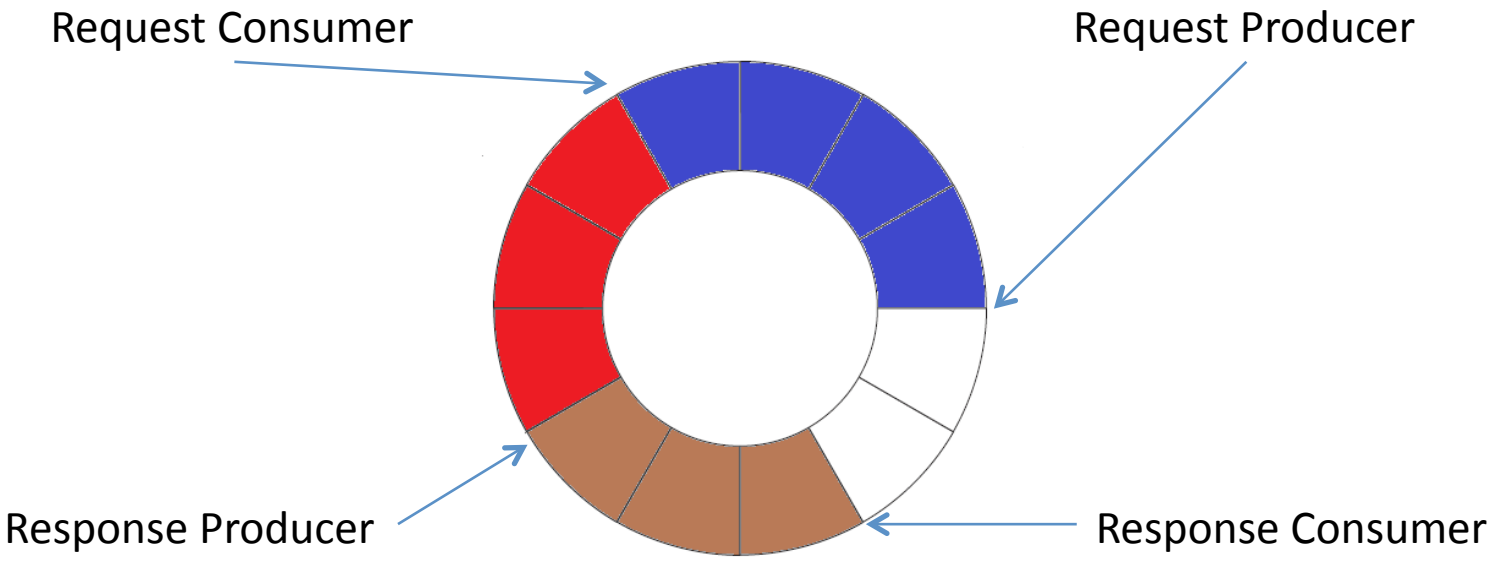
Outstanding descriptors - Descriptor slots awaiting a response from Xen

Response queue - Descriptors returned by Xen in response to serviced requests

Unused descriptors

Xen: VMM Data Transfers

- I/O Rings



Request queue - Descriptors queued by the VM but not yet accepted by Xen

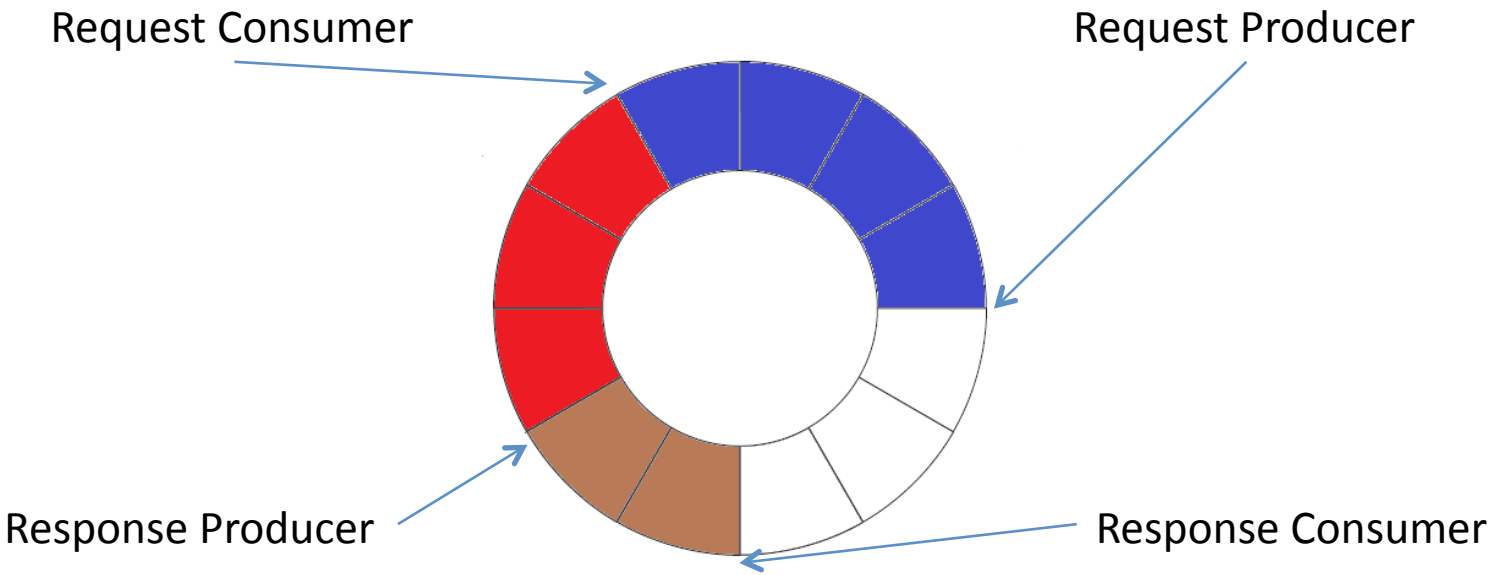
Outstanding descriptors - Descriptor slots awaiting a response from Xen

Response queue - Descriptors returned by Xen in response to serviced requests

Unused descriptors

Xen: VMM Data Transfers

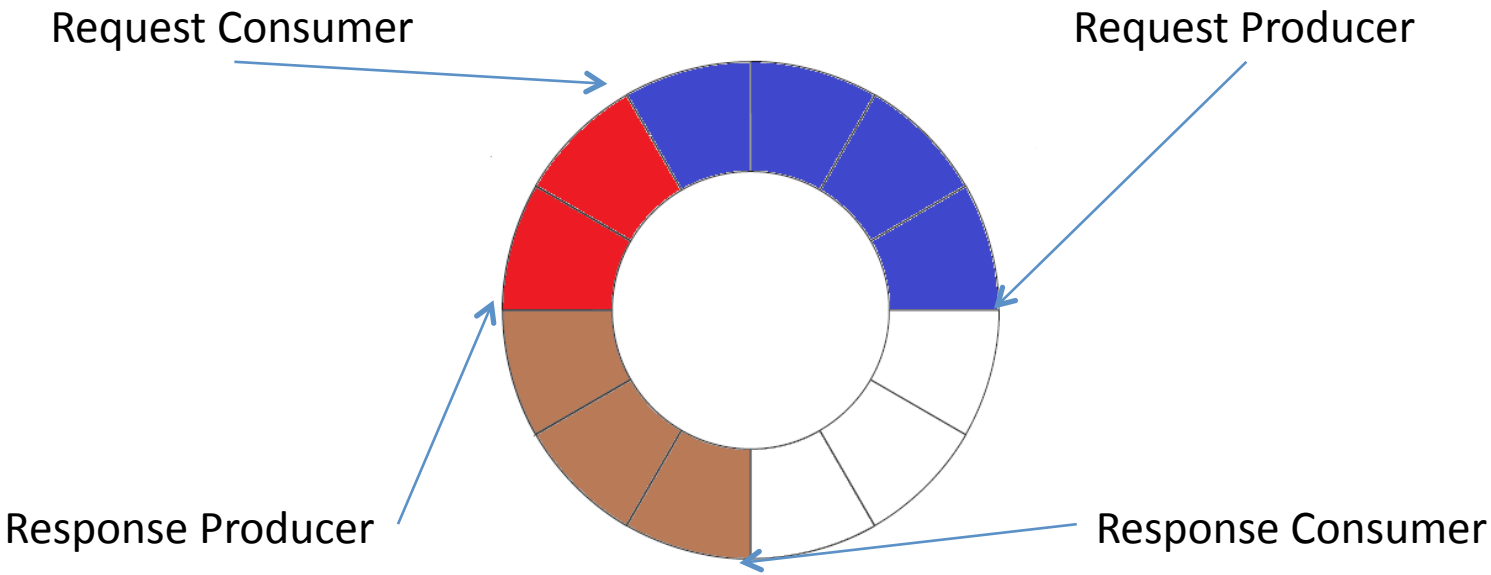
- I/O Rings



Request queue - Descriptors queued by the VM but not yet accepted by Xen
Outstanding descriptors - Descriptor slots awaiting a response from Xen
Response queue - Descriptors returned by Xen in response to serviced requests
Unused descriptors

Xen: VMM Data Transfers

- I/O Rings



Request queue - Descriptors queued by the VM but not yet accepted by Xen

Outstanding descriptors - Descriptor slots awaiting a response from Xen

Response queue - Descriptors returned by Xen in response to serviced requests

Unused descriptors

Xen: VMM Intercommunication

- Hypercalls
  - Domain to Xen communication
  - Synchronous
  - Batched
- Events
  - Xen to Domain communication
  - Asynchronous
- Direct link through the hypervisor

Xen: VMM Pass-through

- New feature
- Performance increase
- Direct access to hardware resources
- No need for Domain0

Xen: Port Costs

- Idealized abstraction
- Linux and Windows
  - XenoLinux
- Paravirtualization port of x86 code base
- Device drivers
- Page-table entries
- Privileged subroutines

Xen: Control and Management

- Daemons
- XML RPC
- Xm
- Xend
- Libxenctrl
- Xenstored
- Qemu-dm
  - full virtualization daemon for disk/network I/O

Source: http://xen.org

Xen: VMM Subsystems

Xen: VMM Subsystems Overview

- CPU and Scheduling
- Timing
- Virtual Address Translation
- Physical Memory
- Device I/O
- Network

Xen: VMM CPU and Scheduling

- Privileges
- Scheduling
  - Borrowed Virtual Time
  - Low-latency
  - Favors new domains
- Exceptions
  - Guest OS registers handlers
  - Stack copied from guest OS

Xen: VMM Timing

- Real Time
  - Time since domain boot
  - Utilizes the clock speed of the processor
- Virtual Time
  - Execution time of the guest OS
- Wall-Clock Time
  - Current real time offset
- Timer Queues
  - Guest OS

Xen: VMM Virtual Addresses

- Page Tables
  - Guest OS allocates directly with Xen
  - Read-only
  - Updates are handle by hypercalls
- Validation
  - Manage page frame types
  - Reference counts
  - Updates based on types

Xen: VMM Virtual Addresses

- Frame Types
  - Page Directory
  - Page Table
  - Local Descriptor Table
  - Global Descriptor Table
  - Writable
- Batch updates in a single hypercall

Xen: VMM Physical Memory

- Reservations
- Balloon driver
  - Existing OS instructions
- Illusion of contiguous
- Mapping by guest OS
- Shared Translation Array
  - Accessible to all
  - Xen validated

Xen: VMM Device I/O

- Device abstractions
- Virtual Block Devices
  - Reordering
  - Uses I/O Ring
- Domain0
  - Disk
  - Network
- Round-robin scheduling

Xen: VMM Network Communication

- Asynchronous I/O Rings
  - Transmit
  - Receive
- Virtual Firewall-Router
- Virtual Network Interfaces
- Direct Memory Access
- Round-robin scheduling for packets

XenoLinux Evaluation

XenoLinux Evaluation

- Comparison
  - VMWare Workstation (without ESX Server)
  - User-Mode Linux (UML)
  - Native Linux
  - XenoLinux (Linux 2.4.21)
- RedHat 7.2 distribution

XenoLinux Performance Evaluation

- SPEC INT2000
- Build Linux 2.4.21 with GCC 2.96
- Open Source Database Benchmark
  - Information Retrieval
  - On-Line Transaction Processing
- dbench
- SPEC WEB99

XenoLinux Performance Evaluation

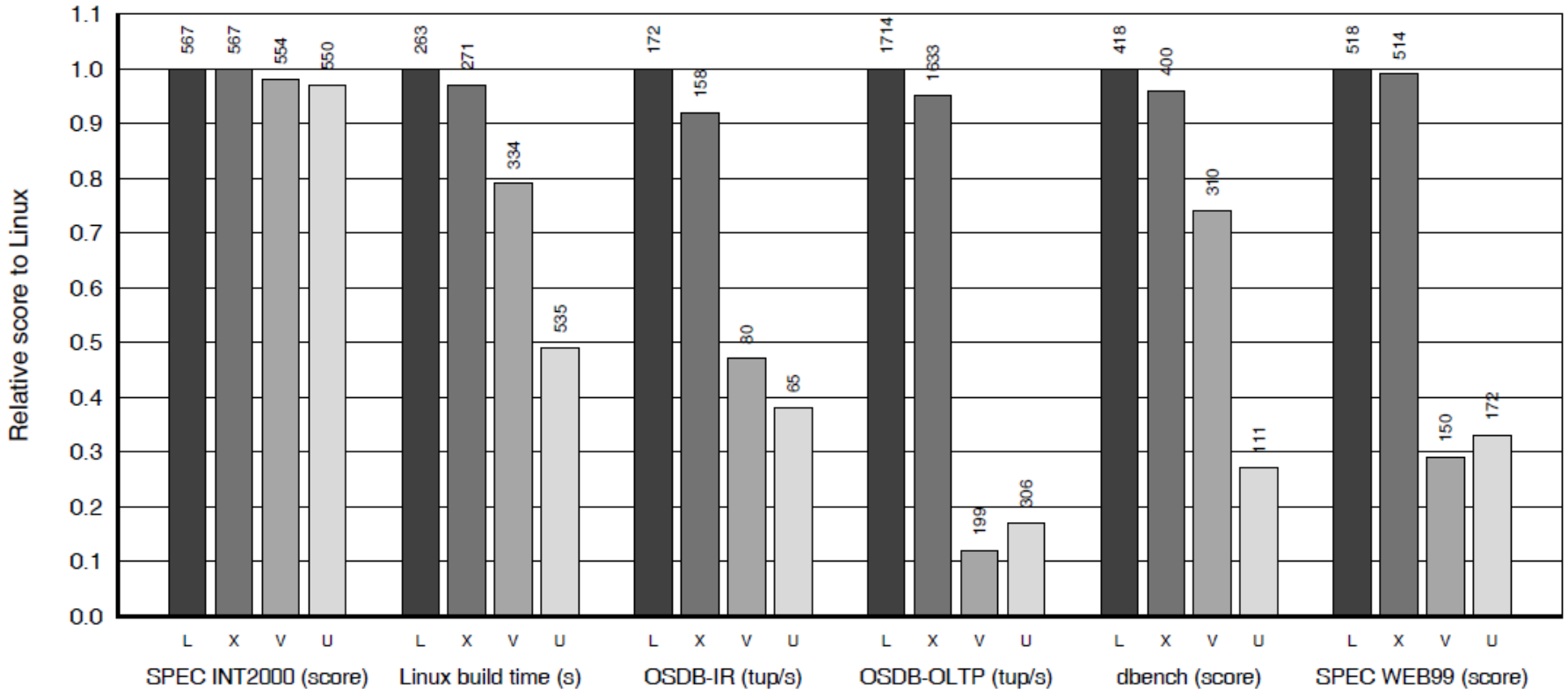


Figure 3: Relative performance of native Linux (L), XenoLinux (X), VMware workstation 3.2 (V) and User-Mode Linux (U).

XenoLinux Other Evaluations

- *lmbench* suite – microbenchmarks
  - 65%
  - Page Table Updates
- Network
  - *ttcp* benchmark
  - Negligible bandwidth differences

XenoLinux Other Evaluations

Config	null call	null I/O	stat	open close	slct TCP	sig inst	sig hndl	fork proc	exec proc	sh proc
L-SMP	0.53	0.81	2.10	3.51	23.2	0.83	2.94	143	601	4k2
L-UP	0.45	0.50	1.28	1.92	5.70	0.68	2.49	110	530	4k0
Xen	0.46	0.50	1.22	1.88	5.69	0.69	1.75	**198**	**768**	**4k8**
VMW	0.73	0.83	1.88	2.99	11.1	1.02	4.63	874	2k3	10k
UML	24.7	25.1	36.1	62.8	39.9	26.0	46.0	21k	33k	58k

Table 3: `lmbench`: Processes - times in μs

	TCP MTU 1500		TCP MTU 500	
	TX	RX	TX	RX
Linux	897	897	602	544
Xen	897 (-0%)	897 (-0%)	516 (-14%)	467 (-14%)
VMW	291 (-68%)	615 (-31%)	101 (-83%)	137 (-75%)
UML	165 (-82%)	203 (-77%)	61.1 (-90%)	91.4 (-83%)

Table 6: `ttcp`: Bandwidth in Mb/s

Further Evaluations

- Concurrency
  - SPEC WEB99
- Isolation
  - Fork Bomb
  - Intensive Disk Access
- Scalability
  - 1 to 128 domains
  - SPEC CINT2000

SPEC WEB99

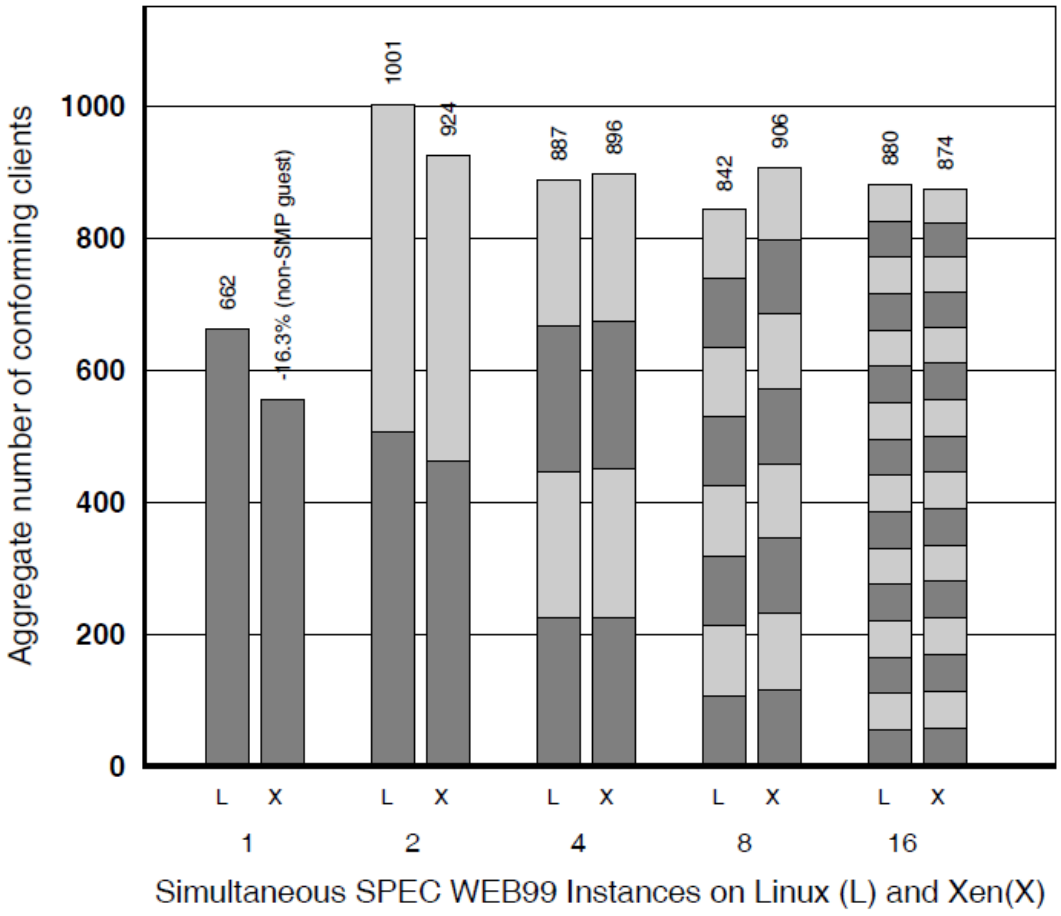


Figure 4: SPEC WEB99 for 1, 2, 4, 8 and 16 concurrent Apache servers: higher values are better.

SPEC CINT200

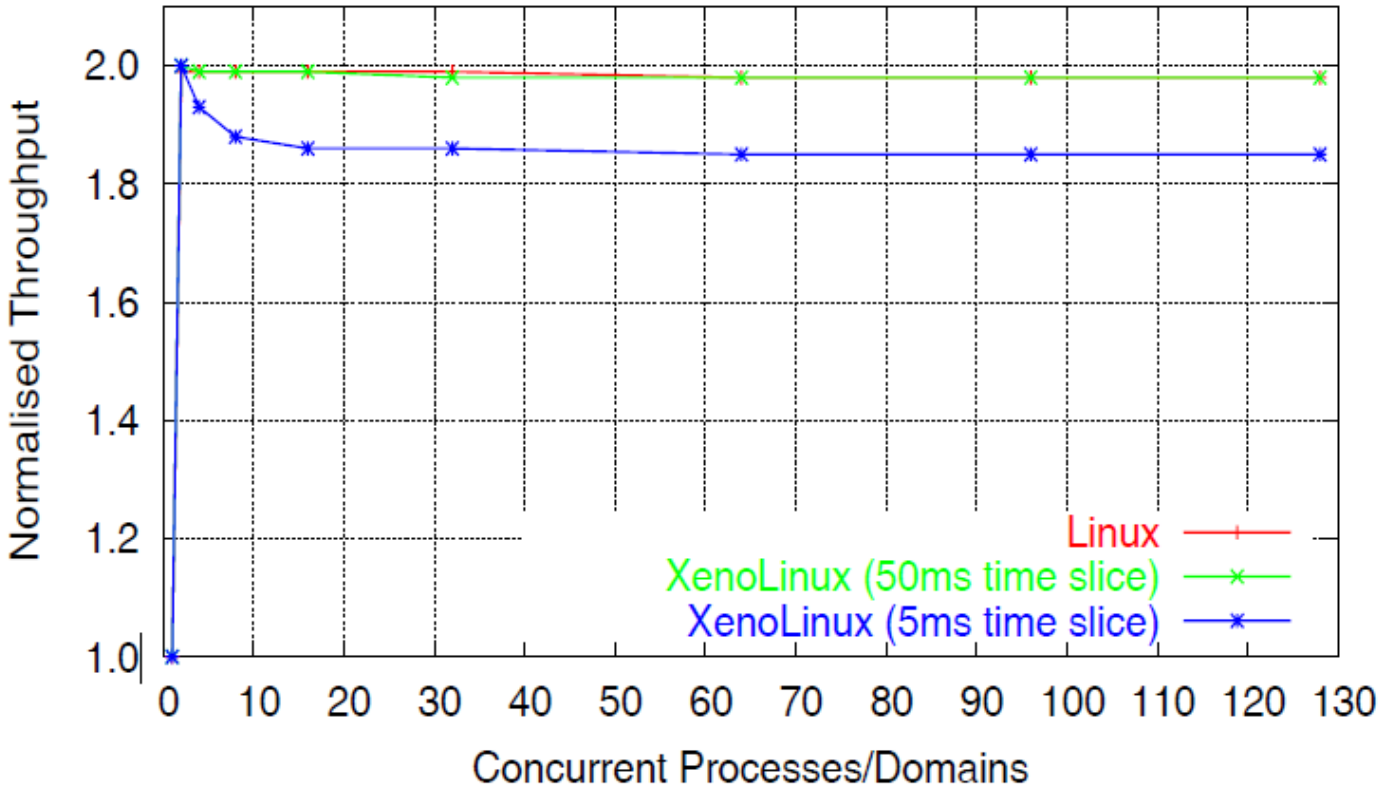


Figure 6: Normalized aggregate performance of a subset of SPEC CINT2000 running concurrently on 1-128 domains

Xen Evaluation

- Isolation
- Reliability
- Security
- Scalability
- Performance
- Heterogeneous

Xen Current State

- Supported Architectures
  - x86
  - x86_64
  - PowerPC
  - IA64
  - ARM (in progress)

Source: http://en.wikipedia.org/wiki/Xen

Xen Current State

- Host OSes
  - Ubuntu, CentOS, RedHat, etc.
  - Linux releases between 2009 and early 2011
    - Not in mainline kernel until 2.6.37
    - Some do not have domain0 support

Source: http://en.wikipedia.org/wiki/Xen

Xen Current State

- Guest OSes
  - Patched Linux 2.6.23 with paravirtualization
  - OpenSolaris
  - Modified WindowsXP
  - Unmodified Windows
    - Intel VT-x
    - AMD-V

Source: http://en.wikipedia.org/wiki/Xen

Xen in Industry

- Amazon Web Services
- Rackspace
- Other Commercial Applications
  - Citrix XenServer, XenDesktop, XenApp, XenClient
  - Oracle VM
  - Sun xVM

Sources: http://xen.org/, http://www.citrix.com

Xen Demo

- Recursive VMs
  - Win7 with VMWare Workstation 7.1.5
  - CentOS 5 with Xen 2.6
  - Fedora 7
- "Russian Doll Effect"

Discussions