University of Salzburg
Computational Systems Seminar
SS 2004

# 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
15 JJ Thomson Avenue, Cambridge, UK, CB3 0FD
{firstname.lastname}@cl.cam.ac.uk

Presented by
Marcus Harringer
mharring@cosy.sbg.ac.at

# XEN Virtual Machine Monitor

1. Introduction
2. Overview
3. Detailed Design
4. Evaluation
5. Conclusion

# XEN Virtual Machine Monitor

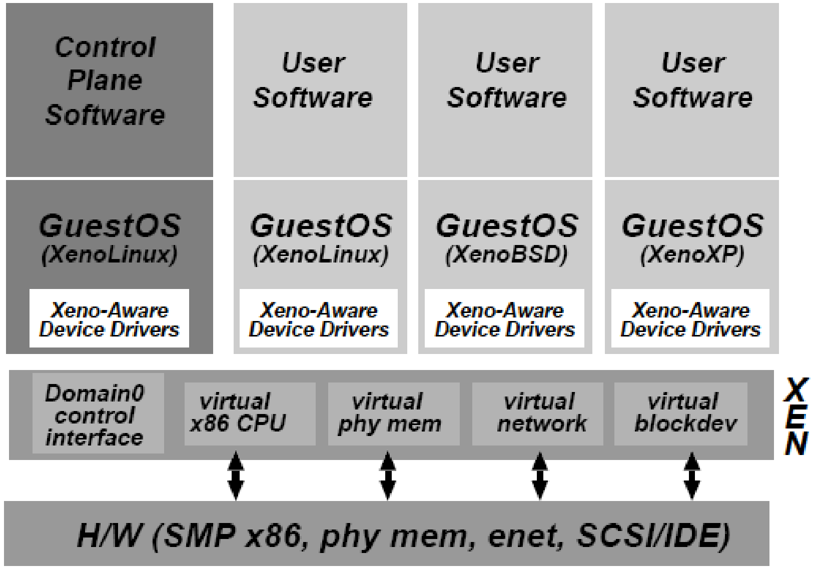
- University of Cambridge Computer Lab.
- High performance virtual machine monitor.
- Challenges:
  - Virtual machines have to be isolated from another.
  - Support a variaty of operating systems.
  - Small performance overhead.
- XenoLinux(2.4), Xenoserver

# XEN Virtual Machine Monitor

- Approaches:
  - Deploy hosts to running a standard OS.
  - Retrofit support for *performance isolation* to the OS.
    - Resource containers, Linux/RK, QLinux.
    - Problem of „QoS crosstalk“.
  - Multiplexing at low level.
    - Exokernel, Nemisis

# XEN Virtual Machine Monitor

- **Xen:**
  - Multiplex physical resources at the granularity of an *entire* OS.
  - Performance isolation.
  - Very flexible.
  - 100 hosted OS instances.
  - Run unmodified binaries.
  - But: GuestOs has to be modified !



# XEN Virtual Machine Monitor

- Outline:
  - Design principles.
  - The virtual machine interface.
    - Memory Management.
    - CPU.
    - Device I/O.
  - The cost of porting an OS to Xen.
  - Control Management.

# XEN Virtual Machine Monitor

- Design Principles:
  - Full virtualization vs. Paravirtualization.
    - X86 Hardware problems.
    - ESX Server uses shadow page tables.
    - Real & Virtual resources are desireable.
      - GuestOS can improve performance using **superpages** or **colored pages**.
    - Support unmodified application binaries.
    - GuestOs has to be modified.

# XEN Virtual Machine Monitor

- Design Principles:
  - Multi-application OS, unlike DenaliOs.
  - GuestOs perform its own paging.
    - Self-paging (NemesisOs).
  - Physical resources directly visible to GuestOs.

# XEN Virtual Machine Monitor

- The Virtual Machine Interface:
  - Memory Management(paging):
    - Most difficult part (mechanism and porting Guest).
    - Software managed TLB, tagged TLB.
    - X86: Hardware TLB
    - GuestOS is responsible for managing hardware page tables.
    - Xen at top of every address space, avoiding TLB flush .

# XEN Virtual Machine Monitor

- The Virtual Machine Interface:
  - Memory Management(paging):
    - Example:
      - Os requires a new page table.
      - Allocates page from own memory reservation.
      - Registers with Xen.
      - Os has to relinquish direct write privileges to PT memory.
      - Updates validated by Xen.
      - May update batch requests.

# XEN Virtual Machine Monitor

- The Virtual Machine Interface:
  - Memory Management(segmentation):
    - Like paging scheme.
    - Validating updates to segment descriptor tables.
      - Lower priviledged than Xen.
      - May not allowed to access to the Xen reserved portion of address space.

# XEN Virtual Machine Monitor

- The Virtual Machine Interface:
  - CPU:
    - Application- Level, OS- Level, Hypervisor- Level.
    - X86 supports 4 priviledge levels !
    - Requires modification of GuestOs.
    - Piviledged instructions → hypervisor call
      - Installing new page tables, yield(),...
      - Exceptions: memory faults and software traps

# XEN Virtual Machine Monitor

- The Virtual Machine Interface:
  - CPU (Exception Handling):
    - Page Fault Handler:
      - Normally read faulting address from a privildged register.
      - Copy exception stack frame on the guestOs stack.
    - Register „fast“ exception handler.
    - What about safety ?
    - Double Faults.

# XEN Virtual Machine Monitor

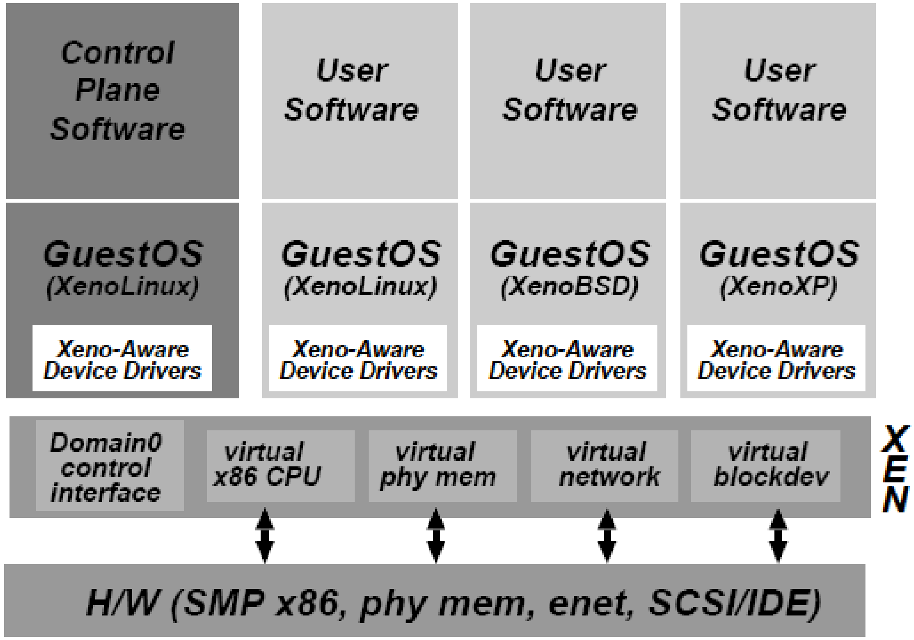
- The Virtual Machine Interface:
  - Device I/O:
    - Clean and simple device abstractions.
    - Data transfer:
      - shared memory.
      - asynchronous buffer desciption rings.
      - Lightweight event-delivery.

# XEN Virtual Machine Monitor

- The Cost of Porting an OS to Xen:
  - Linux 2.4:
    - Total: 2995 LOC.
    - Portion of total x86 code base: 1,36%.
  - Windows XP:
    - Total: 4620 LOC.
    - Portion of total x86 code base: 0.04%.
    - Not yet finished.

# XEN Virtual Machine Monitor

- **Control & Management:**
  - Create domains
  - Terminate domains
  - Scheduling parameters
  - Physical memory allocation
  - Access to devices



# XEN Virtual Machine Monitor

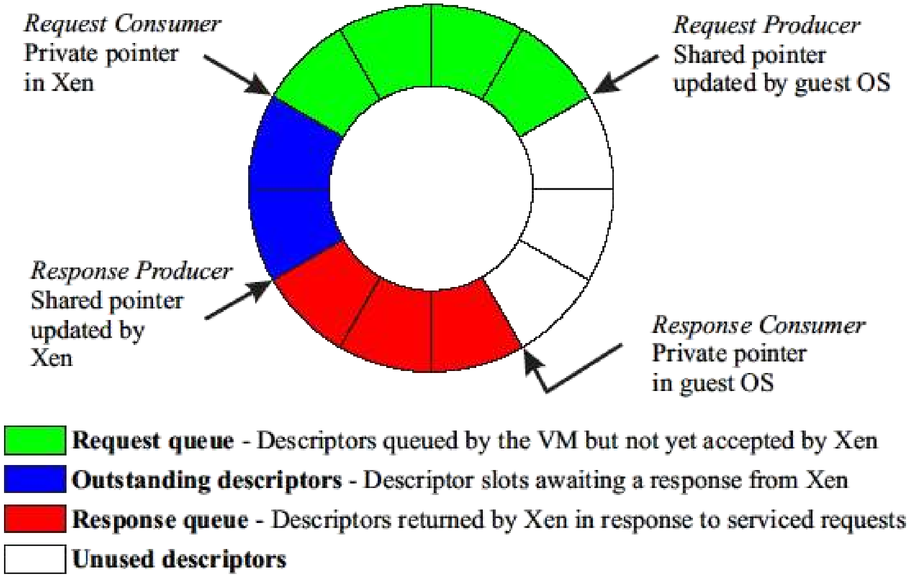
- Outline:
  - Control Transfer: hypercalls, events.
  - Data Transfer: I/O Rings.
  - Subsystem Virtualization
    - Domain scheduling.
    - Virtual address translation.
    - Physical memory.
    - Network.
    - Disk.

# XEN Virtual Machine Monitor

- Control Transfer:
  - – Synchronous calls: hypercalls.
    - Priviledged operations: page-table updates.
  - – Asynchronous calls: events.
    - Device interrupts, lightweight notification.
    - Event-callback handler defined by GuestOS.
    - Per domain bitmask for pending events.
    - Domain defers notofications → disabling interrupts.

# XEN Virtual Machine Monitor

- **Data Transfer: I/O Rings.**
  - Goal: little overhead.
  - I/O descriptor ring.
  - Allocated by GuestOS
  - Accessible within Xen.
  - Not directly contain data.
  - Have not to be ordered.



🟩 **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 Virtual Machine Monitor

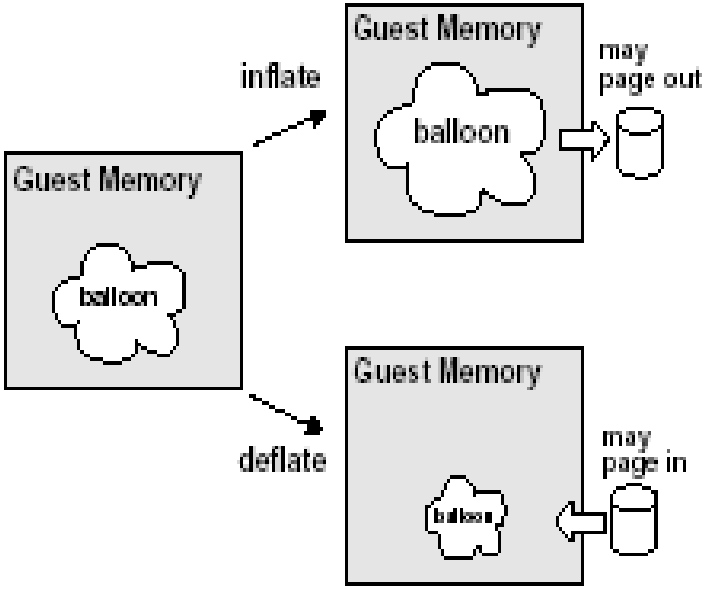
- Domain Scheduling:
  - – Borrowed Virtual Time (BVT) algorithm
    - Universal Scheduler.
    - Thread execution is monitored.
    - Latency-sensitive thread is allowed to „warp“ back in virtual time, for earlier dispatch.
  - – Fast dispatch: minimize the effect of virtualization.
  - – Low latency dispatch.

# XEN Virtual Machine Monitor

- Virtual address translation:
  - VMWare: virtual page tables, costly.
  - Xen: only involved by updates.
  - Hypercall Validation:
    - Type per machine page frame.
      - PD = Page Directory.
      - PT = Page Table.
      - LDT = Local Descriptor Table.
      - GDT = Global Descriptor Table.
      - RW = writeable.

# XEN Virtual Machine Monitor

- Physical memory:
  - Statically partitioned due to initial memory reservation.
  - Reservation Limit.
  - XenoLinux implements a balloon driver.
  - Illusion of continous memory by GuestOS.
  - Shared translation array.



# XEN Virtual Machine Monitor

- Network:
  - Virtual network interface (VIF):
    - I/O descriptor buffer rings (receive, transmit).
    - Transmit:
      - Scatter-gather DMA: only packet-header is copied.
      - Round-Robin scheme.
    - Receive:
      - Exchange an unused page frame for each packet.
      - Page-aligned receive buffers.

# XEN Virtual Machine Monitor

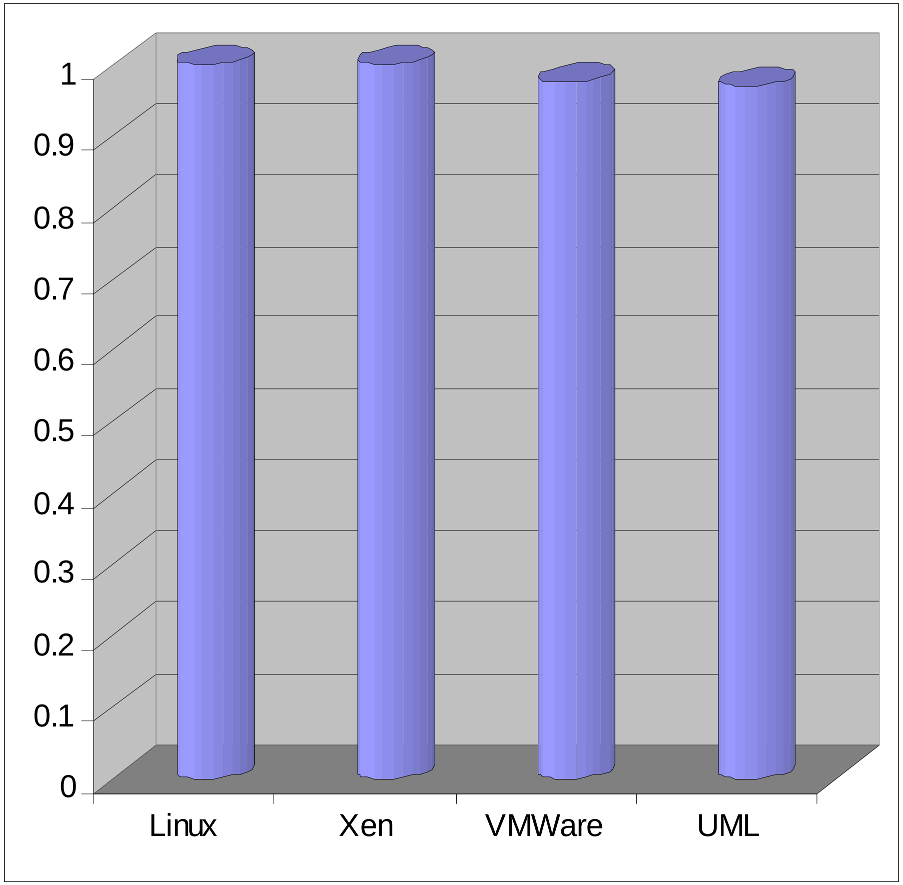
- Disk:
  - Virtual Block Devices:
    - Domain0 has unchecked access.
    - Accessed via I/O Ring mechanism.
    - Ownership, access-control.
    - Translation table managed by Domain0.
    - Reorder requests.
    - Round-Robin.
    - Reorder barriers: write ahead logs.

# XEN Virtual Machine Monitor

- Outline:
  - Relative performance.
  - Concurrent Virtual Machines.
  - Scalability.

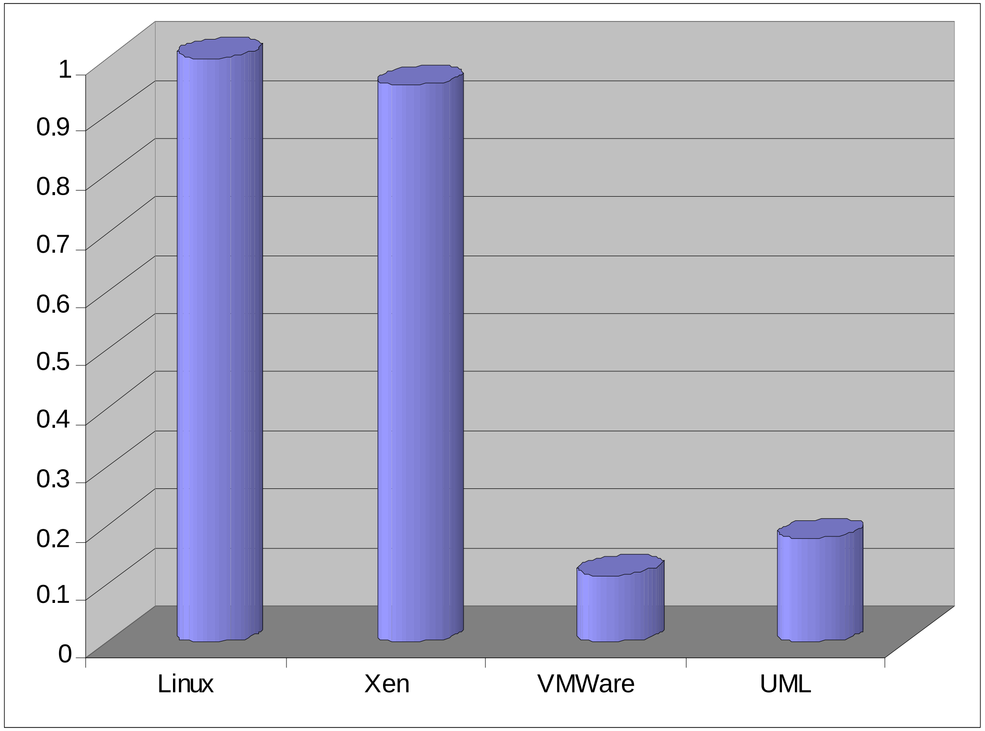
# XEN Virtual Machine Monitor

- SPEC INT2000:
  - Long running computationally-intensive applications.
  - Almost time spent in user- space code.



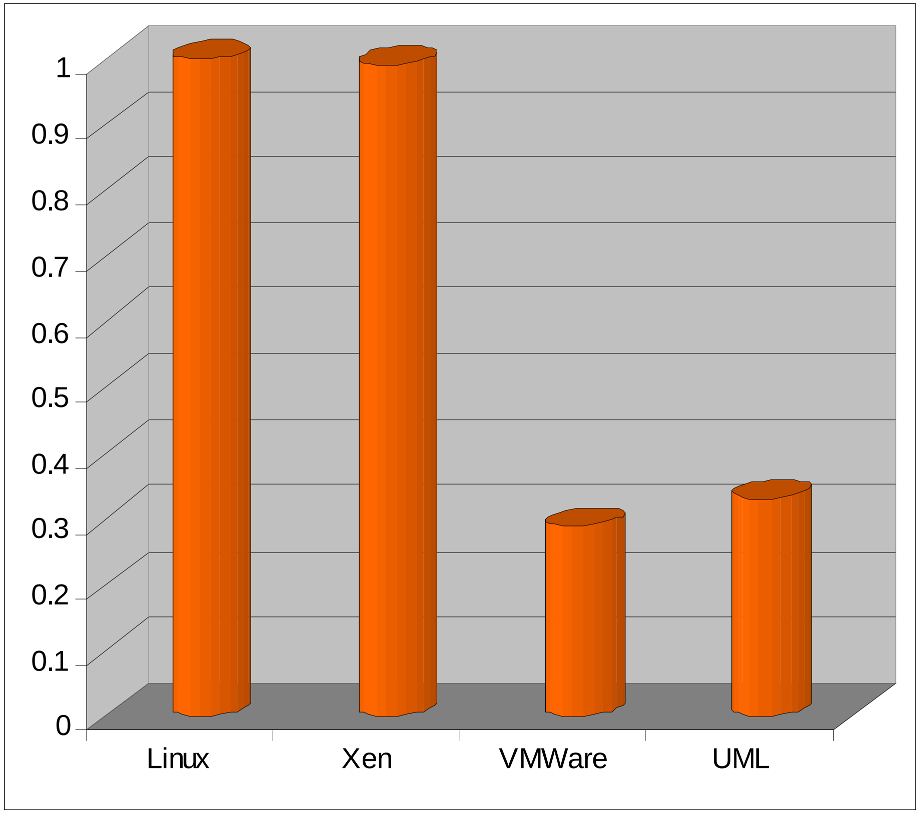
# XEN Virtual Machine Monitor

- OSDB- OLTP(tup/ s):
  - PostgreSQL 7.1.3
  - Online transaction processing



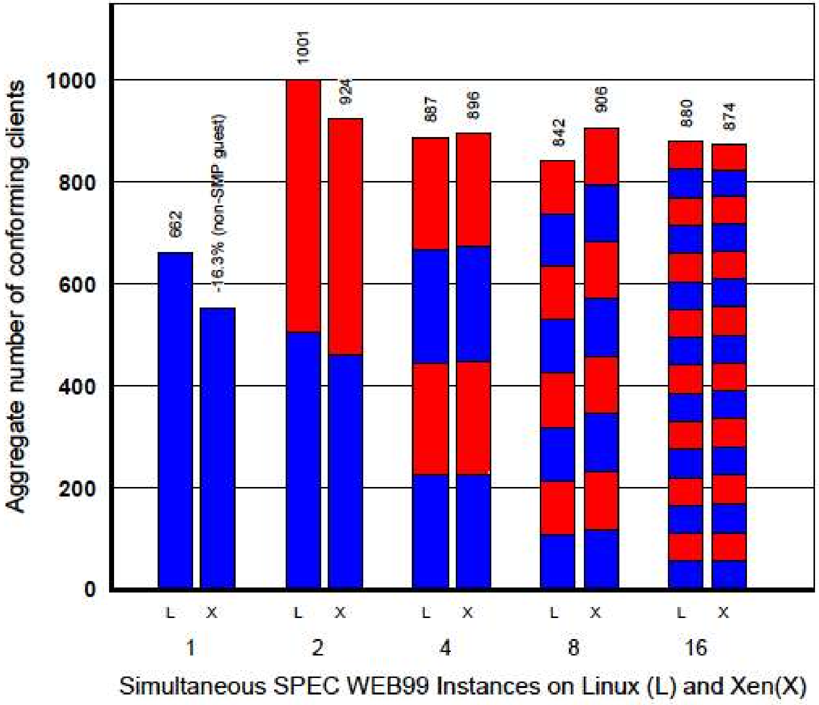
# XEN Virtual Machine Monitor

- SPEC WEB99:
  - Complex application-level benchmark for evaluating webservers and systems that host them.
  - Determine max. users.
  - Apache 1.3
  - CPU-bound
  - Most time spent in GuestOS.



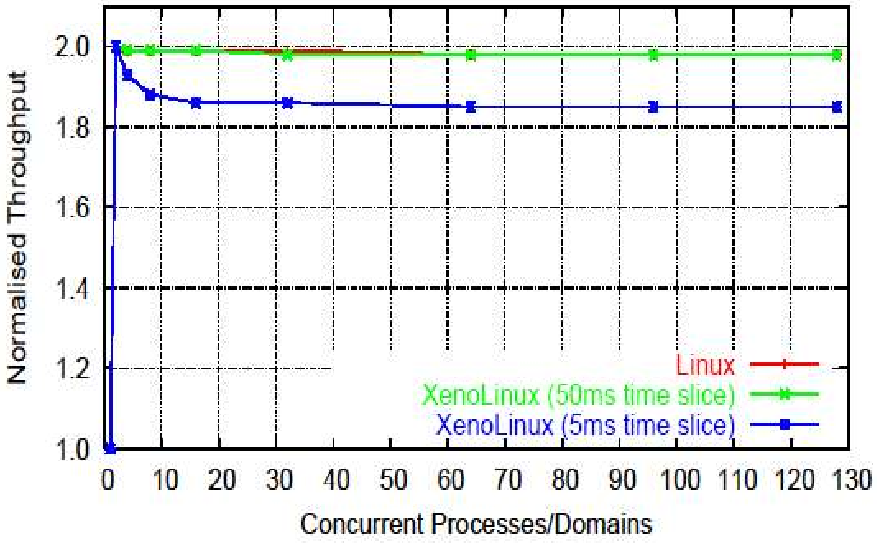
# XEN Virtual Machine Monitor

- Concurrent Virtual Machines:
  - SPEC WEB99.
  - Multiple applications running in GuestOS.
  - 2 CPU machine.
  - Native Linux with SMP.
  - XenoLinux on uniprocessor.



# XEN Virtual Machine Monitor

- Scalability:
  - SPEC CINT2000:
  - Linux 50ms timeslice.
  - XenoLinux, 5 and 50 ms.



**Normalized aggregate performance of a subset of SPEC CINT2000 running concurrently on 1-128 domains**

# XEN Virtual Machine Monitor

- Summary:
  - Virtual machine monitor.
  - Paravirtualization.
  - 100% binary compatibility.
  - GuestOS has to be modified.
  - Support for general purpose OSes (Linux,XP,BSD).
  - Scales up to 100 instances.
  - Very efficient.
  - Easy configuration.
  - Transient servers for short periods of time and with low instantiation costs.
  - http://www.cl.cam.ac.uk/Research/SRG/netos/xen/

# XEN Virtual Machine Monitor

Thanks for your attention !