Xen and the Art of Virtualization

Ian Pratt
VP, Citrix Systems and
Chairman of Xen.org
Outline

• A brief overview of Xen and Xen.org
• Why virtualization is important
• Virtualization frontiers:
  – Virtualization Security
  – IO Virtualization
  – High-Availability
  – Client Device Virtualization
  – Multi-tenancy for Cloud
Xen History

- Mar 1999 XenoServers HotOS paper
- Apr 2002 Xen hypervisor development starts
- Apr 2003 First public Xen Release
- Oct 2003 Xen SOSP paper
- Apr 2004 Xen 1.0 released
- Jun 2004 First Xen developer’s summit
- 2004 Hardware vendors start taking Xen seriously
- 2005 RedHat, Novell, Sun and others adopt Xen
- 2006 VMware and Microsoft adopt paravirtualization
- Sep 2006 First XenEnterprise released
- May 2008 Xen embedded in Flash on HP/Dell servers
Xen Project Mission

• Build the industry standard open source hypervisor
  – Core "engine" that is incorporated into multiple vendors’ products

• Maintain Xen’s industry-leading performance

• Maintain Xen’s reputation for stability and quality
  – Security must now be paramount

• Support multiple CPU types; big and small systems
  – From server to client to mobile phone

• Foster innovation

• Drive interoperability
First and Best to support new CPU, chipset, and Smart IO Technologies

Pioneers of OS Para-virtualization
Virtualization Benefits
Why Virtualization is ‘Hot’

• Clearing up the mess created by the success of ‘scale-out’
  – One Application per commodity x86 server
  – Leads to ‘server sprawl’
  – 5-15% CPU utilization typical

• Failure of popular OSes to provide
  – Full configuration isolation
  – Temporal isolation for performance predictability
  – Strong spatial isolation for security and reliability
  – True backward app compatibility
First Virtualization Benefits

- **Server consolidation**
  - Consolidate scale-out success
  - Exploit multi-core CPUs

- **Manageability**
  - Secure remote console
  - Reboot / power control
  - Performance monitoring

- **Ease of deployment**
  - Rapid provisioning

- **VM image portability**
  - Move image between different hardware
  - Disaster Recovery
2nd Generation Virtualization Benefits

- Avoid planned downtime with VM Relocation
- Dynamically re-balance workload to meet app SLAs or to save power
Virtualization Security
Hypervisors and Security

• Exploitation of a hypervisor is a real threat
• Hypervisors add more software and thus increase the attack surface
  – Network-facing control stack
  – VM containment
• Xen smaller and defensible than an OS
  – Need a “strength in depth” approach
    • Disaggregate, De-privilege, narrow interfaces
    • Xen Security Modules
  – Secure Boot
Improving Security with Hypervisors

• Hypervisors allow administrative policy enforcement from outside of the OS
  – Firewalls, IDS, malware scanning etc
    • More robust as not so easily disabled
    • Provides protection within a network rather than just at borders
  – Backup policy, multi-path IO, HA, FT etc
    • Availability and Reliability
  – Hardening OSes with immutable memory, taint tracking, logging and replay
  – Introspection is an active research area
Network IO Virtualization
Network Interface Virtualization

• Network IO is tough to virtualize
  – High packet rate
    • Batches often small
  – Data must typically be copied to VM on Receive
  – Some apps are latency sensitive

• Xen’s network IO virtualization has evolved significantly over time
  – Need to take advantage of new NIC features
I/O Architecture

Virtual MMU & Virtual CPU

Hardware (SMP, MMU, physical memory, Ethernet, SCSI/IDE)
Xen Driver Domains

VM0
- Device Manager & Control s/w
- GuestOS
  - Back-End
    - Native Device Driver
- Control IF
- Safe HW IF
- Event Channel
- Virtual CPU
- Virtual MMU
- Xen Virtual Machine Monitor
- Hardware (SMP, MMU, physical memory, Ethernet, SCSI/IDE)

VM1
- GuestOS
  - Back-End
    - Native Device Driver

VM2
- GuestOS
  - Front-End Device Drivers

VM3
- GuestOS
  - Applications
  - Front-End Device Drivers
  - Device Emulation
Isolated Driver VMs for High Availability

- Run device drivers in separate domains
- Detect failure e.g.
  - Illegal access
  - Timeout
- Kill domain, restart
- E.g. 275ms outage from failed Ethernet driver
Multiple RX Queues

• NIC supports multiple free and RX buffer Q’s
  – Choose queue based on destination MAC, VLAN
  – Default queue used for multicast/broadcast

• Great opportunity for avoiding data copy for high-throughput VMs
  – Try to allocate free buffers from buffers the guest is offering
  – Still need to worry about broadcast, inter-domain etc

• Multiple TX queues with traffic shaping
IOMMU: Direct Device Assignment

- **Device Manager & Control s/w**
  - GuestOS
  - Back-End
  - Native Device Driver

- **Applications**
  - Event Channel
  - Virtual CPU
  - Safe HW IF
  - IOMMU
  - Virtual MMU

- **Device Emulation**
  - Front-End Device Drivers
  - Applications

Hardware (SMP, MMU, physical memory, Ethernet, SCSI/IDE)
SR-IOV: Hardware IO Virtualization

- NIC presents itself as multiple PCI devices, one per guest
  - Relies on IOMMU for protection
  - Still need to deal with the case when there are more VMs than virtual h/w NICs
  - h/w-specific driver in guest, loses some of the hardware abstraction benefits of virtualization

- Full Ethernet switch functionality on NIC
  - Inter-domain traffic can go via NIC
    - But data goes over PCIe bus twice, may be slow
SR-IOV NIC Demonstration

- Full 20Gb/s bi-directional throughput to VMs
- Low latency, High CPU efficiency
- Live relocation between hosts - Even hosts with different NICs
Network Performance

- New Smart NICs reduce CPU overhead substantially
- Care must be taken with SR-IOV NICs to ensure benefits of VM portability and live relocation are not lost
High Availability and Fault Tolerance
Hardware Fault Tolerance

- Restart-HA monitors hosts and VMs to keep apps running

- Hardware Fault Tolerance with deterministic replay or checkpointing

Xen’s Software-Implemented Hardware Fault Tolerance enables true High Availability for unmodified applications and operating systems
Hardware Fault Tolerance

• E.g. University of British Columbia’s “Remus”
• Smart checkpointing approach yields excellent performance
  – VM executes in parallel with checkpoint transmission, with all externally visible state changes suppressed until checkpoint receipt acknowledged
  – Checkpoints delta compressed
• Checkpointing possible across wide-area, even for multi-vCPU guests
Virtualization on Client Devices
The Xen Client Initiative

• Formed in 2007 to develop Xen for desktop and laptop
• Develop enhanced power management, USB, WiFi, WWAN, 3D Graphics, fingerprint reader, multi-touch, etc
• Support for latest hardware technologies
• Tiny footprint hypervisor, Embeddable in Flash memory or small disk partition
• Aiming to make virtualization ubiquitous on client devices...
Client Hypervisor Benefits

• Security, Manageability, Supportability, Auditability

• Building Multi-Level Secure systems
  – Run multiple VMs with policy controlled information flow
    • E.g. Personal VM; Corporate VM; VM for web browsing; VM for banking
  – Trusted hypervisor provides secure isolation

• Enables “out-of-band” management and policy enforcement
  – Malware detection, remote access, image update, backup, VPN, etc.

⇒ Requires a true type-1 hypervisor architecture
⇒ Xen is ideally suited to this!
Types of Virtual Machine Monitor

**Type 1 Hypervisor**
- personal image
- corporate image

**Type 2 VMM**
- corporate image
- Type-2 hypervisor
- Personal Image
- hardware
Xen Client Architecture

- Control Domain
- Service VM
- VM1
- VM2
- Xen Hypervisor
- x86 Hardware
  - Audio
  - USB
  - GPU
  - Disk
  - ACPI
  - NIC
  - TXT
  - TPM
“Business” & “Personal” Environments

- Locked Down
- No Local App Installs
- Tightly Managed
  - Self-Service Corporate App Installs
- Allows Local App Installs
- Minimal Management
  - Virus Scanner
  - Security Patches
- No SLA
  - Self-Service Wipe
The XenClient Solution

• **High Performance Client Virtualization**
  • Provides a High Quality User Experience
  • Securely Run Multiple Hardware Independent Images
  • Provide Ability to get under the client OS and manage it

• **A New Way to Deliver Desktops**
  • Use a Single Image for Initial Deployment and Ongoing Management
  • Efficient Two-way Data Synchronization
  • Flexible policy controls
  • Integrated Encryption and Backup
From Laptops to Mobiles

• Smart phones and PDAs
  – Xen ARM
  – Smart phones now suffer from many of the same problems as PCs

• Simple restricted use cases:
  – Three VMs running on one CPU:
    • Real time VM for controlling the radio
    • VM for vendor/operator -supplied s/w
    • VM for user-downloaded software
Virtualization in the Cloud
XenoServers: University Project from 1999

- Incremental rollout
- Flexible platform
- Unified management

- Global services and apps
- Exploit network topology
- Open commercial platform
Amazon has thousands of servers running Xen
- Server consolidation and workload management
- EC2 (Elastic Computing Cloud) “Rent a VM”
Xen Cloud Platform (XCP)

• XCP Expands Xen.org’s remit beyond the core hypervisor, to create a full virtual infrastructure layer for Cloud deployments
  – Simplify and streamline use of Xen by Cloud providers and vendors
  – Promote greater standardisation of components between vendors

• Advanced virtual infrastructure to enable Virtual Private Datacenters rather than just Virtual Private Servers
  – Multi-tenant hosts, networking, storage, etc
  – Promote interoperability between xen-based clouds and other clouds
  – Drive standards activities via DMTF
New Open vSwitch

Isolation • Resource control • Multi-tenancy • Visibility • Security

- Open Source Virtual Switch maintained at www.openvswitch.org
- Rich layer 2 feature set
Distributed vSwitch

Built-in policy-based ACLs move with VMs

Virtual Interface (VIF) {MAC, IP} ACLs
- permit tcp 10.0.0.0 0.0.0.255 10.20.0.0 0.0.0.255 eq domain
- permit tcp 192.168.0.0 0.0.0.255 10.20.0.0 0.0.0.255 eq domain
- permit tcp 172.16.0.0 0.0.0.255 10.20.0.0 0.0.0.255 eq domain
- permit udp 10.0.0.0 0.0.0.255 10.20.0.0 0.0.0.255 eq domain
- permit udp 192.168.0.0 0.0.0.255 10.20.0.0 0.0.0.255 eq domain
- permit udp 172.16.0.0 0.0.0.255 10.20.0.0 0.0.0.255 eq domain
- permit tcp 10.0.0.0 0.0.0.255 10.20.0.0 0.0.0.255 eq 123
Distributed vSwitch

Isolation · Resource control · Multi-tenancy · Visibility · Security

Distributed Virtual Switch

Tenant A
Tenant B
Conclusions

• Open Source is a great way to get impact from University research projects!
• Hypervisors will become ubiquitous, near zero overhead, embedded in the hardware
• Virtualization may enable a new "golden age" of operating system diversity
• Virtualization is a really fun area to be working in!

ian.pratt@xen.org