Cloud^H^H^H^H^H

Virtualization Technology

Andrew Jones (drjones@redhat.com)

May 2011

Outline

- Promise to not use the word “Cloud” again

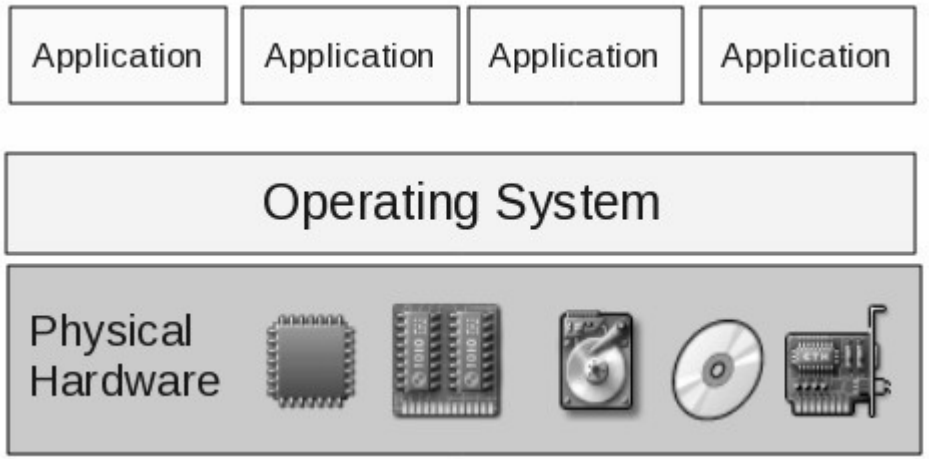
  ...but still give a couple use cases for Virtualization
- Emulation – it's not just for games
- The x86 arch
  - The Dark Ages (before Virt extensions)
  - The Age of Reason (after VT-x / AMD-V)
- x86 Hypervisors – aka why Red Hat likes KVM
- Quiz
- Q/A

Virtualization: a couple use cases

- Disk space – always too little or too much...
  - Hot plug (without the trip to the server room...)
  - Logical volumes, sparse files (allow over committing)
- Your mistakes never happened – snapshots
- The new “App” (Appliances vs. Applications)
- Debugging kernel code (even HV code? – nested virt)
- Primarily for a word that starts with 'C'
  - No, not “Cloud”. *Consolidation* and also for *Competitions* of uptime (err... RAS)
    - RAS - Reliability, Availability and Serviceability

Emulation

- Not another abstraction layer, but rather a layer of indirection
- Slipped between well defined abstraction layers (ABI)
- Some uses:
  - Hardware ahead of software
  - Software ahead of hardware
  - Hardware is just different
  - Support Virtualization
    - BIOS
    - Serial
    - PCI, USB busses

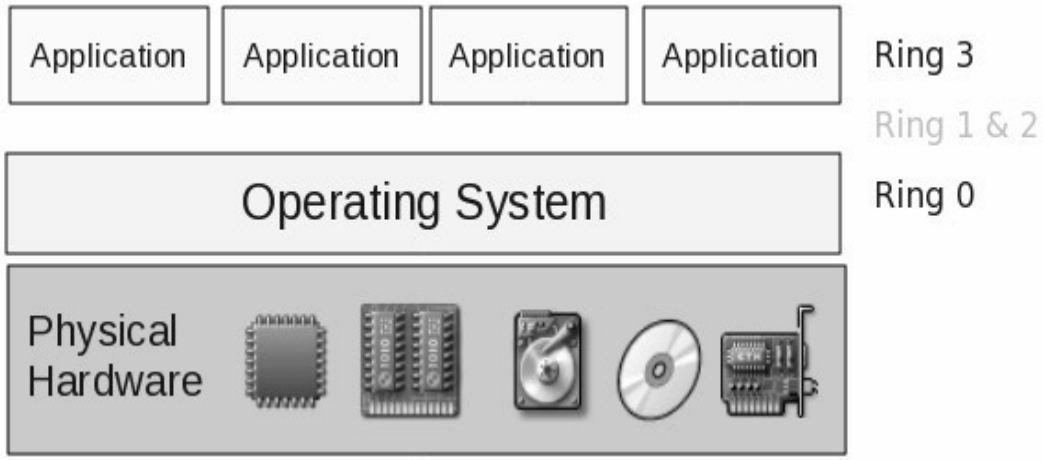


x86 Proliferation

- Maybe not the best instruction set or design, but...
  - Cheap home computers and notebooks
  - Cheap computer labs in schools
  - Cheap servers
  - Cheap clusters for parallel processing
- Comes with an excellent OS
  - Linux! Which is even free
- What could be better?
  - All that plus free Virtualization too, of course

x86 Virtualization

- The Dark Ages
  - 1998 – VMWare: Binary translation
    - Performance is limited (must read all VM binary code)
    - Not open source
  - 2003 – Xen: Paravirt
    - Open source
    - Requires modified guest kernel code (hypercalls)

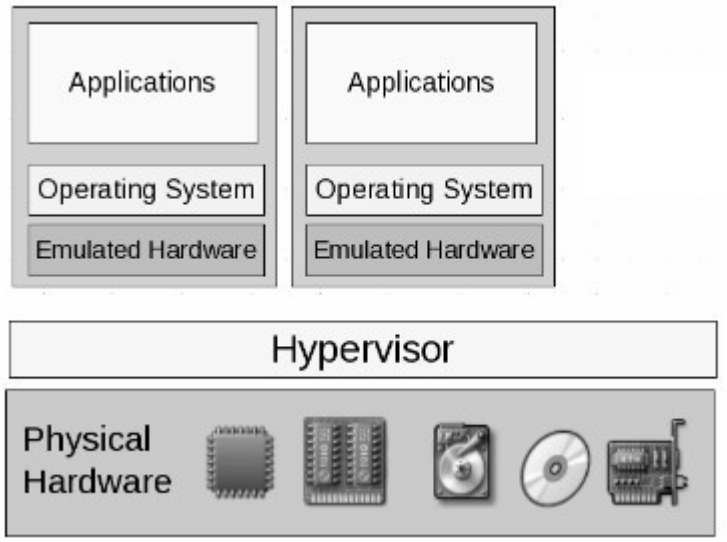


x86 Virtualization

- The Age of Reason (2005)
  - VT-x and AMD-V
  - New Guest mode – VMLAUNCH/VMRESUME
  - Instructions that should trap (privops), now do trap – VMEXIT
- Guest page tables
  - Round 1: Shadowed in the hypervisor
  - Round 2: vMMU (HAP – Hardware Assisted Paging)
    - Intel – EPT (Extended Page Tables), AMD – RVI (Rapid Virt Indexing)
    - ASID (Address Space ID) for TLB sharing
- Guest Device Access
  - IOMMU (DMA), Device Assignment, SR-IOV
- x2apic – Virt interrupt controller

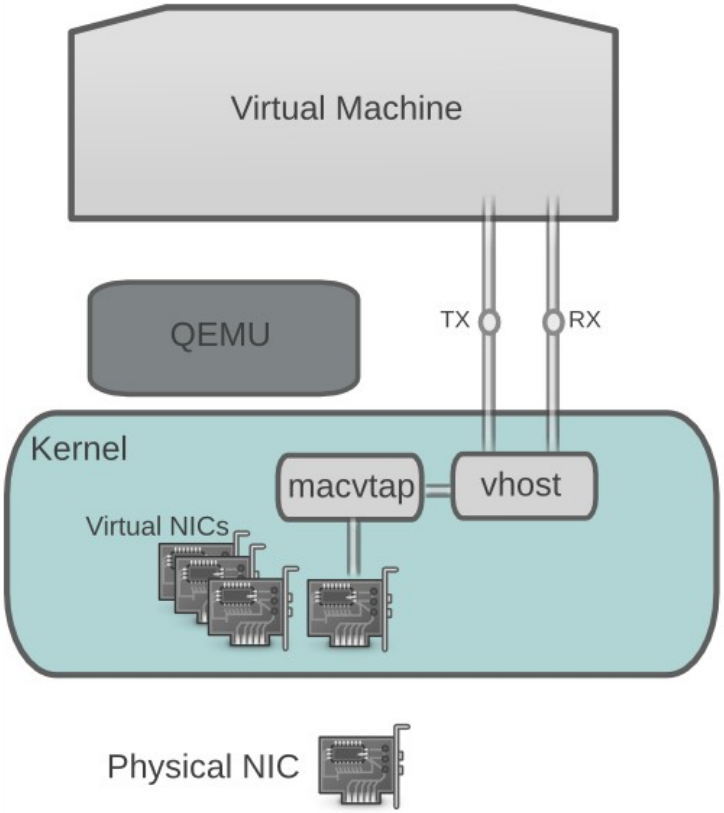
Hypervisors

- Deal with guest privops
  - Binary translation
  - Paravirt
  - VMEXITs
- Schedule VMs
- Manage guest memory
  - Both shadow and HPA need support
- Grant access to I/O resources
- Implement core VM management (e.g. launch a VM)
  - Extended management done in userspace (libvirt)



Hypervisors

- What they don't generally do
  - Implement a console
  - Implement I/O and network stacks
- No I/O? A useless guest?
  - Bring back the emulator
- But emulation has poor performance...
  - Bring back paravirt
  - Device assignment (IOMMU, SR-IOV)
  - Even both at the same time



KVM - Kernel-based Virtual Machine

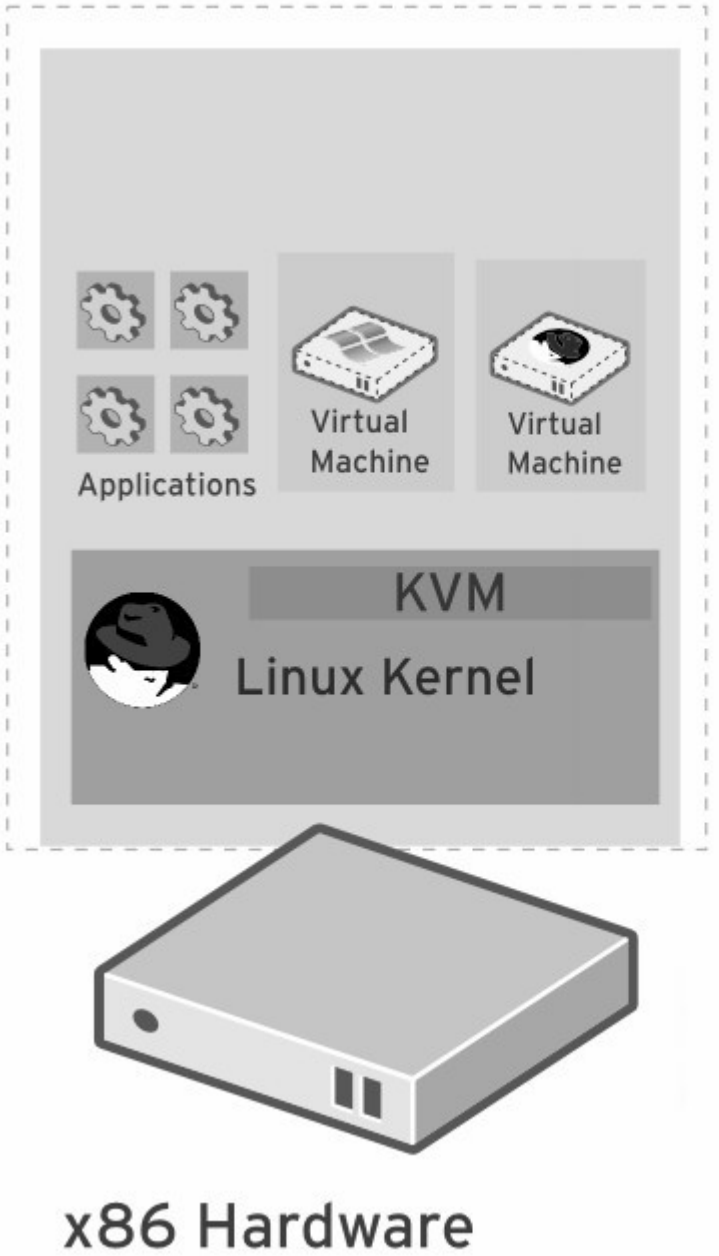
- Recall HVs need a scheduler and a memory manager
- They also need to boot (surprise, surprise...) and enable/drive all the hardware
  - And not just the little box you have at home
  - Also machines with hundreds of cores and Terabytes of memory (even NUMA)
- KVM (released 2007) does all this already
  - Busy Engineers? Yes, but not *that* busy
  - Linux already does all the above – KVM adds the support for Virt extensions

KVM

- Kernel module that can be loaded on your Linux box
- Launches/manages VMs (vcpus – VMCS structures)
  - VM is a Linux QEMU process (Q emulator)
  - qemu-kvm runs the guest image (kernel + userspace)
  - Guest image can use paravirt drivers (VirtIO)
- Linux memory manager
  - Swap, shared memory, THP (Transparent Huge Pages)
  - KSM (Kernel Samepage Merging)
  - Ballooning
  - Over committing for fun and profit...

KVM

- kvmclock
- Current work
  - Guest NUMA awareness
  - Always looking for speedup opportunities
  - More security
  - Nested Virt (looks fun)
  - Keep up with the hardware



Quiz

The name of the game is

'The answer is C'

Quiz

What language is the kernel and KVM written in?

a) PHP

b) Lisp

c) C

d) What's the kernel? What's KVM?

Quiz

Virtualization is

a) the same as abstraction

b) abstraction, but also a multiplexer

c) a layer of indirection

d) Who care's? If it's just virtual, then it's not real anyway...

Quiz

Virtualization is good for

a) Allowing one VM to move around to other machines

b) Allowing multiple VMs to run on the same machine

c) Both (a) & (b)

d) Nothing, why are we talking about it?

Quiz

If OS is to syscall, Hypervisor is to _____?

a) event

b) libcall

c) hypercall (hcall)

d) call home

Quiz

When a guest tries to issue a privileged instruction (privop) what happens?

a) runs without the privileges

b) traps to the OS

c) traps to the Hypervisor

d) crashes the system, starting a big fire...

Quiz

Paravirtualization means

a) running an unmodified guest

b) running a system without virtualization

c) running a guest that implements at least some parts with hypervisor-aware mechanisms

d) A virtual parachute

...Ouch

Q/A

Thank you!

Further Questions?
drjones@redhat.com

Type 1 vs. Type 2