Virtual Machine (VM)

“A VM is an efficient, isolated duplicate of a real machine”
[Popek&Goldberg 74]

- Duplicate: VM should behave identically to the real machine
  - Programs cannot distinguish between real or virtual hardware
  - Except for:
    - Fewer resources (and potentially different between executions)
    - Some timing differences (when dealing with devices)
- Isolated: Several VMs execute without interfering with each other
- Efficient: VM should execute at speed close to that of real hardware
  - Requires that most instruction are executed directly by real hardware

Hypervisor aka virtual-machine monitor: Software implementing the VM

"Real machine": Modern usage more general, “virtualise” any API

Types of Virtualisation

- Platform VM or System VM
  - Type-1 “Bare metal”
    - Requires hypervisor
  - Type-2 “Hosted”
    - Runs on top of existing OS

Programming Language

- Java
  - Requires Java VM

Operating System

- Hypervisor
  - Operating System
  - Process

- Java Program
  - Java VM

- OS API

Plus anything else you want to sound cool!
Why Virtual Machines?

• Historically used for easier sharing of expensive mainframes
  - Run several (even different) OSes on same machine
    o called guest operating system
  - Each on a subset of physical resources
  - Can run single-user single-tasked OS in time-sharing mode
    o legacy support
  - Gone out of fashion in 80’s
    - Time-sharing OSes common-place
    - Hardware too cheap to worry...

• Renaissance in recent years for improved isolation
  - Server/desktop virtual machines
    - Improved QoS and security
    - Uniform view of hardware
    - Complete encapsulation
      o replication
      o migration/consolidation
      o checkpointing
      o debugging
    - Different concurrent OSes
      o e.g. Linux + Windows
    - Total mediation
  - Would be mostly unnecessary
    - ... if OSes were doing their job!

Why Virtual Machines?

• Emerge driver today is Cloud computing
  - Increased utilisation by sharing hardware
  - Reduced maintenance cost through scale
  - On-demand provisioning
  - Dynamic load balancing though migration

Why Virtual machines

• Alternative to physical separation
  - low-overhead communication
  - size, weight and power (SWaP) reduction
  - consolidate complete components
    o including OS,
    o certified
    o supplied by different vendors
    o legacy support
  - “dual-persona” phone
  - secure domain on COTS device

Gernot's prediction of 2004: 2014 OS textbooks will be identical to 2004 version except for s/process/VM/g

Why Virtual Machines?
Hypervisor aka Virtual Machine Monitor

- Program that runs on real hardware to implement the virtual machine
- Controls resources
  - Partitions hardware
  - Schedules guests
    - "world switch"
  - Mediates access to shared resources
    - e.g. console
- Implications
  - Hypervisor executes in privileged mode
  - Guest software executes in unprivileged mode
  - Privileged instructions in guest cause a trap into hypervisor
  - Hypervisor interprets/emulates them
  - Can have extra instructions for hypercalls

Native vs. Hosted VMM

- Hosted VMM beside native apps
  - Sandbox untrusted apps
  - Convenient for running alternative OS on desktop
  - Leverage host drivers
- Less efficient
  - Double node switches
  - Double context switches
  - Host not optimised for exception forwarding

Virtualization Mechanics: Instruction Emulation

- Traditional trap-and-emulate (T&E) approach:
  - Guest attempts to access physical resource
  - Hardware raises exception (trap), invoking HV’s exception handler
  - Hypervisor emulates result, based on access to virtual resource
- Most instructions do not trap
  - Prerequisite for efficient virtualisation
  - Requires VM ISA (almost) same as processor ISA

Trap-and-Emulate Requirements

Definitions:
- Privileged instruction: traps when executed in user mode
  - Note: NO-OP is insufficient!
- Privileged state: determines resource allocation
  - Includes privilege mode, addressing context, exception vectors...
- Sensitive instruction: control- or behaviour-sensitive
  - Control sensitive: changes privileged state
  - Behaviour sensitive: exposes privileged state
    - Incl instructions which are NO-OPs in user but not privileged state
- Innocuous instruction: not sensitive
- Some instructions are inherently sensitive
  - Eg TLB load
- Others are context-dependent
  - Eg store to page table
Trap-and-Emulate Architectural Requirements

- **T&E virtualisable**: all sensitive instructions are privileged
  - Can achieve accurate, efficient guest execution
    - ... by simply running guest binary on hypervisor
  - VMM controls resources
  - Virtualized execution indistinguishable from native, except:
    - resources more limited (smaller machine)
    - timing differences (if there is access to real time clock)

- **Recursively virtualisable**:
  - run hypervisor in VM
  - possible if hypervisor not timing dependent, overheads low

Impure Virtualization

- Virtualise other than by T&E of unmodified binary
- Two reasons:
  - Architecture not T&E virtualisable
  - Reduce virtualisation overheads
- Change guest OS, replacing sensitive instructions
  - by trapping code ("hypercalls")
  - by in-line emulation code

Binary Translation

- Locate sensitive instructions in guest binary, replace on-the-fly by emulation or trap/hypercall
  - pioneered by VMware
  - detect/replace combination of sensitive instruction for performance
  - modifies binary at load time, no source access required
- Looks like pure virtualisation!
- Very tricky to get right (especially on x86!)
  - Assumptions needed about sane guest behaviour
  - "Heroic effort" [Orran Krieger, then IBM, later VMware]

Para-Virtualization

- New(ish) name, old technique
  - coined by Denai [Whitaker '02], popularised by Xen [Barham '03]
  - Mach Unix server [Golub '90], L4Linux [Härtig '97], Disco [Bugnion '97]
- Idea: manually port guest OS to modified (more high-level) ISA
  - Augmented by explicit hypervisor calls (hypercalls)
    - higher-level ISA to reduce number of traps
    - remove unvirtualisable instructions
    - remove “messy” ISA features which complicate
  - Generally outperforms pure virtualisation, binary re-writing
- Drawbacks:
  - Significant engineering effort
  - Needs to be repeated for each guest-ISA-hypervisor combination
  - Para-virtualised guests must be kept in sync with native evolution
  - Requires source
Virtualization Overheads

- VMM must maintain virtualised privileged machine state
  - processor status
  - addressing context
  - device state
- VMM needs to emulate privileged instructions
  - translate between virtual and real privileged state
  - eg guest ↔ real page tables
- Virtualisation traps are expensive
  - >1000 cycles on some Intel processors!
  - Better recently, Haswell has <500 cycle round-trip
- Some OS operations involve frequent traps
  - STI/CLI for mutual exclusion
  - frequent page table updates during fork()
  - MIPS KSEG addresses used for physical addressing in kernel

Virtualization and Address Translation

Two levels of address translation!

Virtual Memory

Virtual Memory

Virtual Memory

Guest Physical Memory

Virtual Page Table

Virtual Page Table

Virtual Page Table

Guest Physical Memory

Virtual Page Table

Guest Physical Memory

Page Table

Page Table

Guest Physical Memory

Virtual Memory

Virtual Memory

Virtual Memory

Physical address

Physical address

Physical address

Physical address

data

Virtualization Mechanics: Shadow Page Table

Hypervisor must shadow (virtualize) all PT updates by guest:
- trap guest writes to guest PT
- translate guest PA in guest (virtual) PTE using guest memory map
- insert translated PTE in shadow PT

Virtualization Mechanics: Shadow Page Table

Shadow PT has TLB semantics (i.e. weak consistency)
- Update at synchronisation points:
  - page faults
  - TLB flushes

Shadow PT as virtual TLB
- similar semantics
- can be incomplete: LRU translation cache

Used by VMware

须实施单一MMU转换！
Virtualisation Semantics: Lazy Shadow Update

User → Guest OS → Hypervisor

- User: access new page...
- Guest OS: add mapping to GPT
- Hypervisor: write-protect GPT, unprotect GPT & mark dirty
- update dirty shadow, write-protect GPT
- return to user

Virtualisation Semantics: Lazy Shadow Update

User → Guest OS → Hypervisor

- User: access new page...
- Guest OS: add mapping to GPT, add mappings...
- Hypervisor: write-protect GPT, unprotect GPT & mark dirty, update dirty shadow, write-protect GPT
- continue

Virtualization Mechanics: Real Guest PT

- Hypervisor maintains guest PT
- Guest OS
- Hypervisor
- Guest PT
- HV PT
- Physical address
- Memory
data

- User: ld r0, addr
- Guest virtual address
- Guest PT
- HV PT
- Physical address

- Hypervisor maintains guest PT
- Guest translates PTEs itself when reading from PT
- supported by Linux PT-access wrappers
- Guest batches PT updates using hypercalls
- reduced overhead

Virtualization Mechanics: Optimised Guest PT

- User: ld r0, addr
- Guest virtual address
- Guest PT
- HV PT
- Physical address
- data

- Hypervisor
- Guest PT
- HV PT
- Physical address

- User: ld r0, addr
- Guest virtual address
- Guest PT
- HV PT
- Physical address
- data

- Hypervisor
- Guest PT
- HV PT
- Physical address

- Guest OS: invalidate mapping in GPT
- Hypervisor: invalidate mapping, flush TLB
- write-protect GPT, unprotect GPT & mark dirty
- update dirty shadow, write-protect GPT, flush TLB
Virtualization Techniques

- Impure virtualisation methods enable new optimisations
  - avoid traps through ability to control the ISA
  - changed contract between guest and hypervisor
- Example: virtualised guest page table
  - lazy update of virtual state (TLB semantics)
- Example: virtual interrupt-enable bit (in virtual PSR)
  - requires changing guest’s idea of where this bit lives
  - hypervisor knows about VM-local virtual state
  - eg queue virtual interrupt until guest enables in virtual PSR

```
VPSR            PSR
paid Trap       Trap
  0          0
  1          1
```

```
mov r1,#VPSR
ldr r0,[r1]
orr r0,r0,#VPSR_ID
sto r0,[r1]
```

Virtualization Mechanics: Emulated Device

- Each device access must be trapped and emulated
  - unmodified native driver
  - high overhead!

Virtualization Mechanics: Split Driver (Xen speak)

- Simplified, high-level device interface
  - small number of hypercalls
  - new (but very simple) driver
  - low overhead
  - must port drivers to hypervisor

```
"Para-virtualized driver"
```
Virtualization Mechanics: Driver OS (Xen Dom0)

- Leverage Driver-OS native drivers
  - no driver porting
  - must trust complete Driver OS guest!
  - huge TCB!

Virtualization Mechanics: Pass-Through Driver

- Unmodified native driver
- Can’t share device between VMs
- Must trust driver (and guest)
  - unless have hardware support (I/O MMU)

Available on modern x86, latest ARM

Modern Architectures Not T&E Virtualisable

- Examples:
  - x86: many non-virtualizable features
    - e.g. sensitive PUSH of PSW is not privileged
    - segment and interrupt descriptor tables in virtual memory
    - segment description expose privileged level
  - MIPS: mostly ok, but
    - kernel registers k0, k1 (for save/restore state) user-accessible
    - performance issue with virtualising KSEG addresses
  - ARM: mostly ok, but
    - some instructions undefined in user mode (banked registers, CPSR)
    - PC is a GPR, exception return is MOVIS to PC, doesn’t trap
- Addressed by virtualization extensions to ISA
  - x86, Itanium since ~2006 (VT-x, VT-i, AMD-V), ARM since ‘12
  - additional processor modes and other features
  - all sensitive ops trap into hypervisor or made innocuous (shadow state)
    - e.g. guest copy of PSW

x86 Virtualization Extensions (VT-x)

- New processor mode: VT-x root mode
  - orthogonal to protection rings
  - entered on virtualisation trap

Non-Root
- Ring 3
- Ring 2
- Ring 1
- Ring 0

Kernel entry

Root
- Ring 3
- Ring 2
- Ring 1
- Ring 0

VM exit

Guest Kernel

Hypervisor
**ARM Virtualization Extensions (1)**

**Hyp mode**

New privilege level
- Strictly higher than kernel
- Virtualizes or traps all sensitive instructions
- Only available in ARM TrustZone “normal world”

**EL0**
- User mode

**EL1**
- Kernel modes

**EL2**
- Kernel modes

**EL3**
- Monitor mode

**ARM Virtualization Extensions (2)**

**Configurable Traps**

- x86 similar
- User mode
- Kernel mode

- User mode
- Native syscall

- User mode
- Virtual syscall

- Kernel mode
- Virtual syscall

**Trap to guest**

**Big performance boost!**

**New privilege level**
- Strictly higher than kernel
- Virtualizes or traps all sensitive instructions
- Only available in ARM TrustZone “normal world”

**ARM Virtualization Extensions (3)**

**Emulation Support**

1) HW decodes instruction
   - No L1 miss
   - No software decode

2) SW emulates instruction
   - Usually straightforward

**No x86 equivalent**

**Emulation**

1) Load faulting instruction
   - Compulsory L1-D miss!

2) Decode instruction
   - Complex logic

3) Emulate instruction
   - Usually straightforward

**L1 I-Cache**

- `ld r1, (r0, ASID)`
- `ld sp, (r1, kern_stk)`

**L1 D-Cache**

- `...`
- `mv CPU ASID, r1`
- `...`
- `mv CPU ASID, r1`

**L2 Cache**

- `ld r1, (r0, ASID)`
- `ld sp, (r1, kern_stk)`
- `...`
- `...`
- `...`

**R2**
- `mv CPU ASID, r1`

**R3**
- `...`

**R1**
- `...`
- `...`

**L1 I-Cache**

- `ld r1, (r0, ASID)`
- `ld sp, (r1, kern_stk)`

**L1 D-Cache**

- `...`
- `...`
- `...`
- `...`

**L2 Cache**

- `ld r1, (r0, ASID)`
- `ld sp, (r1, kern_stk)`

**33** COMP9242 S2/2016 W03 © 2016 Gernot Heiser. Distributed under CC Attribution License

**34** COMP9242 S2/2016 W03 © 2016 Gernot Heiser. Distributed under CC Attribution License

**35** COMP9242 S2/2016 W03 © 2016 Gernot Heiser. Distributed under CC Attribution License

**36** COMP9242 S2/2016 W03 © 2016 Gernot Heiser. Distributed under CC Attribution License
ARM Virtualization Extensions (4)

2-stage translation
- Hardware PT walker traverses both PTs
- Loads combined (guest-virtual to physical) mapping into TLB
- Eliminates "virtual TLB"

1st PT ptr (Hardware)
Guest virtual address

2nd PT ptr (Hardware)
Guest physical address

Hypervisor's guest memory map

Physical address

data

ARM Virtualization Extensions (5)

Virtual Interrupts
- ARM has 2-part IRQ controller
  - Global "distributor"
  - Per-CPU "interface"
- New H/W "virt. CPU interface"
  - Mapped to guest
  - Used by HV to forward IRQ
  - Used by guest to acknowledge
- Halves hypervisor invocations for interrupt virtualization

x86: issue only for legacy level-triggered IRQs

ARM Virtualization Extensions (6)

System MMU (I/O MMU)
- Devices use virtual addresses
- Translated by system MMU
  - Elsewhere called I/O MMU
  - Translation cache, like TLB
  - Reloaded from same page table
- Can do pass-through I/O safely
  - Guest accesses device registers
  - No hypervisor invocation

x86 different (VT-d)

Many ARM SoCs different

TLB

Physical Address

System MMU

Physical Memory

Guest Physical Address

VM
World Switch

**x86**
- VM state is $\leq 4$ KiB
- Save/restore done by hardware on VMexit/VMentry
- Fast and simple

**ARM**
- VM state is 488 B
- Save/restore done by software (hypervisor)
- Selective save/restore
  - Eg traps w/o world switch

```
    Guest 1 state
    \downarrow
    VM 1 control block
    \downarrow
    Save

    Guest 2 state
    \downarrow
    VM 2 control block
    \downarrow
    Restore
```

Hybrid Hypervisor OSes

- Idea: turn standard OS into hypervisor
  - … by running in VT-x root mode
  - eg: KVM (“kernel-based virtual machine”)
- Can re-use Linux drivers etc
  - Huge trusted computing base!
- Often falsely called a Type-2 hypervisor

```
Non-Root

VM
Guest apps
Guest kernel

\rightarrow

Root

VM
Guest apps
Guest kernel

\rightarrow

Hypervisor

\rightarrow

Ring 0

Linux kernel

Native Linux apps

Linux demons

```

ARM: seL4 vs KVM [Dall&Nieh ’14]
Virtualisation Cost (KVM)

<table>
<thead>
<tr>
<th>Component</th>
<th>ARM A15 cycles</th>
<th>x86 Sandybridge cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>VM exit+entry</td>
<td>27</td>
<td>821</td>
</tr>
<tr>
<td>World Switch</td>
<td>1,135</td>
<td>814</td>
</tr>
<tr>
<td>I/O Kernel</td>
<td>2,850</td>
<td>3,291</td>
</tr>
<tr>
<td>I/O User</td>
<td>6,704</td>
<td>12,218</td>
</tr>
<tr>
<td>EOI+ACK</td>
<td>13,726</td>
<td>2,305</td>
</tr>
</tbody>
</table>

KVM needs WS for any hypercall!

Source: [Dall&Nieh, ASPLOS’14]

Fun and Games with Hypervisors

- Time-travelling virtual machines [King ‘05]
  - debug backwards by replay VM from checkpoint, log state changes
- SecVisor: kernel integrity by virtualisation [Seshadri ‘07]
  - controls modifications to kernel (guest) memory
- Overshadow: protect apps from OS [Chen ‘08]
  - make user memory opaque to OS by transparently encrypting
- Turtles: Recursive virtualisation [Ben-Yehuda ‘10]
  - virtualize VT-x to run hypervisor in VM
- CloudVisor: mini-hypervisor underneath Xen [Zhang ‘11]
  - isolates co-hosted VMs belonging to different users
  - leverages remote attestation (TPM) and Turtles ideas
  … and many more!

Hypervisors vs Microkernels

- Both contain all code executing at highest privilege level
  - Although hypervisor may contain user-mode code as well
    - privileged part usually called “hypervisor”
    - user-mode part often called “VMM”
- Both need to abstract hardware resources
  - Hypervisor: abstraction closely models hardware
  - Microkernel: abstraction designed to support wide range of systems
- What must be abstracted?
  - Memory
  - CPU
  - I/O
  - Communication

What’s the difference?

<table>
<thead>
<tr>
<th>Resource</th>
<th>Hypervisor</th>
<th>Microkernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>Virtual MMU (vMMU)</td>
<td>Address space</td>
</tr>
<tr>
<td>CPU</td>
<td>Virtual CPU (vCPU)</td>
<td>Thread or scheduler activation</td>
</tr>
</tbody>
</table>
| I/O           | • Simplified virtual device  
               • Driver in hypervisor  
               • Virtual IRQ (vIRQ)  
               • I/O interface to user-mode driver  
               • Interrupt IPC |
| Communication | Virtual NIC, with driver and network stack | High-performance message-passing IPC |

• Similar abstractions
• Optimised for different use cases

Just page tables in disguise
Just kernel-scheduled activities
Real Difference?

Modelled on HW, Re-uses SW
Minimal overhead, Custom API
Closer Look at I/O and Communication

- Communication is critical for I/O
  - Microkernel IPC is highly optimised
  - Hypervisor inter-VM communication is frequently a bottleneck

Hypervisors vs Microkernels: Drawbacks

**Hypervisors:**
- Communication is Achilles heel
  - more important than expected
  - critical for I/O
  - plenty improvement attempts in Xen
- Most hypervisors have big TCBs
  - infeasible to achieve high assurance of security/safety
  - in contrast, microkernel implementations can be proved correct

**Microkernels:**
- Not ideal for virtualization
  - API not very effective
  - L4 virtualization performance close to hypervisor
  - effort much higher
  - Needed for legacy support
  - No issue with H/W support?
- L4 model uses kernel-scheduled threads for more than exploiting parallelism
  - Kernel imposes policy
  - Alternatives exist, eg. K42 uses scheduler activations