Virtualization in Embedded Systems
Lecture for the Embedded Systems Course
CSD, University of Crete (April 24, 2023)

- Manolis Marazakis (maraz@ics.forth.gr)

FORTH-ICS Institute of Computer Science (ICS)
Foundation for Research and Technology – Hellas (FORTH)

Today's lecture

CS-428 focus shift in remainder of lectures: from "simple" to "complex" embedded

- Introduction of concepts
  - Virtualization (ISA/ABI/API, VM, VMM/Hypervisor)
  - Taxonomies of virtualization approaches
- Motivation in the context of embedded systems
  - H/W + S/W co-design
  - Use-cases (mostly from mobile)
- Virtualization techniques (& overheads)
  - Dynamic Binary Translation
  - (De-)Privileged execution, Traps (instr. & trace faults)
  - Memory management, Primary vs. Shadow structures

Architecture viewpoint x86, ARM (+ RISC-V)

OS viewpoint: Focusing on Linux, in embedded systems

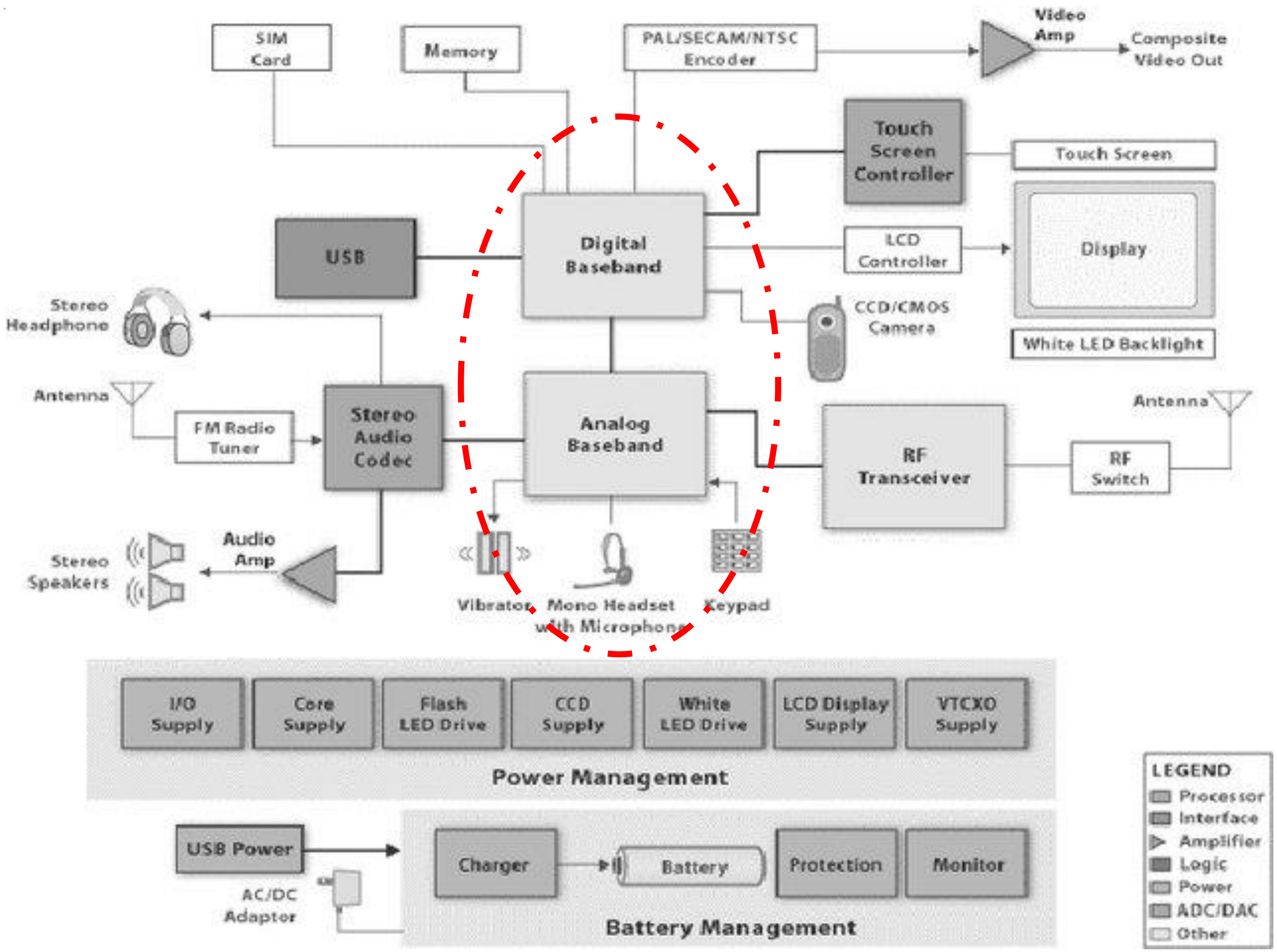
Virtualization Definitions

- Virtualization
  - A layer mapping its visible interface and resources onto the underlying layer or system on which it is implemented
  - Purposes: abstraction, replication, isolation
- Virtual Machine (VM)
  - An efficient, isolated duplicate of a real machine
    - Programs should not be able to distinguish between execution on real or virtual H/W (except for: fewer/variable resources, and device timing)
    - VMs should execute without interfering with each other
    - Efficiency requires that most instructions execute directly on real H/W
- Hypervisor / Virtual Machine Monitor (VMM)
  - Partitions a physical machine into multiple "virtual machines"
    - Host : machine and / or software on which the VMM is implemented
    - Guest : the OS which executes under the control of the VMM

Uses of virtual machines

- Multiple (identical) OS’es on same platform
  - The original *raison d'être*
  - These days mostly driven by server consolidation
- Interesting variants of this:
  - Different OSes (e.g. Linux + Windows)
  - Old version of same OS
  - OS debugging (most likely uses Type-II VMM)
  - Checkpoint-restart
    - minimize lost work in case of crash
    - useful for debugging, incl. going backwards in time
    - re-run from last checkpoint to crash, collect traces, invert trace from crash
- Live system migration
  - Load balancing, Environment take-home
- Ship application with complete OS
  - Reduce dependency on environment
- **What about embedded systems?**

Block diagram of a basic mobile phone



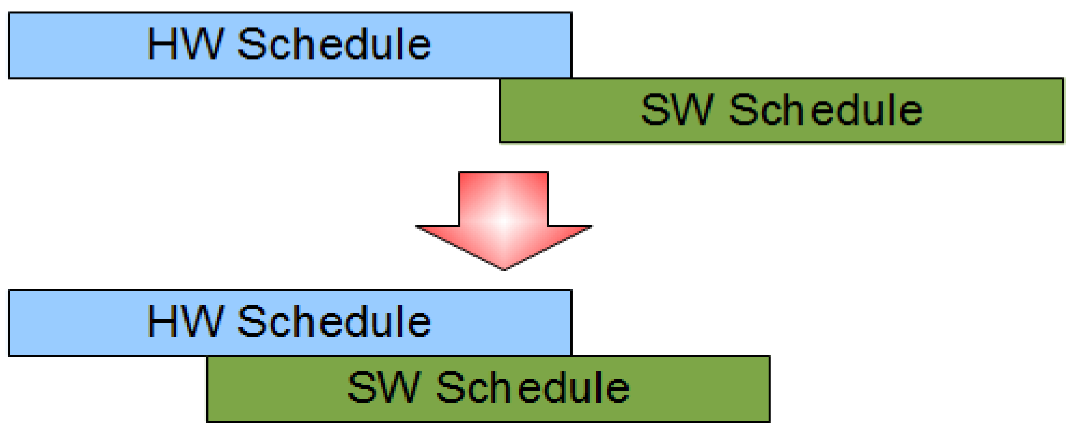
Virtualization to enable H/W-S/W co-design

- How to **co-design/co-develop H/W + S/W** for a system ?
  - Limited availability
  - Bugs in the production environment cannot be reproduced in the laboratory
  - Difficult to debug on-site
  - Narrow time windows
  - Sometimes in a dangerous environment …
- **Debugging** challenges
  - Is it a problem in the driver or in the device?
  - Is the firmware faulty? Is it wrongly loaded/configured?
  - Is the hardware damaged?
  - How can we reproduce the bug?
  - Do we have easy access to the environment?
  - Is it remotely located?

Writing (and testing) device drivers … without hardware

Shift Left

- Hardware + Software = Complete product
- Feature-complete software by A-0 silicon
- Software needs to happen earlier



[source: PJ Waskiewicz & Shannon Nelson - Linux Plumbers Conference, 2011]

OS vs Hypervisor (VMM)

- Hypervisor / Virtual Machine Monitor (VMM)
  - Software that supports virtual machines on a physical machine
  - Determines how to map VM resources to physical ones
  - Physical resources may be time-shared, partitioned, or emulated
- The OS has complete control of the (physical) system
  - Impossible for >1 operating systems to be executing on the same platform
  - OS provides execution environment for processes
- Hypervisor (VMM) "virtualizes" the hardware interface
  - GuestOS's do not have complete control of the system
  - VMM provides execution environment for OS
    - "virtual hardware"

What needs to be emulated for a VM? [Hardware]

- CPU and memory hierarchy
  - ISA, Register state, Memory state
  - Privilege levels, Exceptions/Traps, Interrupts
- Memory Management Unit (MMU)
  - Page tables, segments → virtual memory support
  - Controlled via special registers, and via page tables
- Platform
  - Interrupt controller, timers, peripheral buses
- Firmware (BIOS)
- Peripheral devices
  - Disk, network interface, serial line
  - Programmed I/O, Direct Memory Access (DMA)
  - Events delivered to software via polling or interrupts

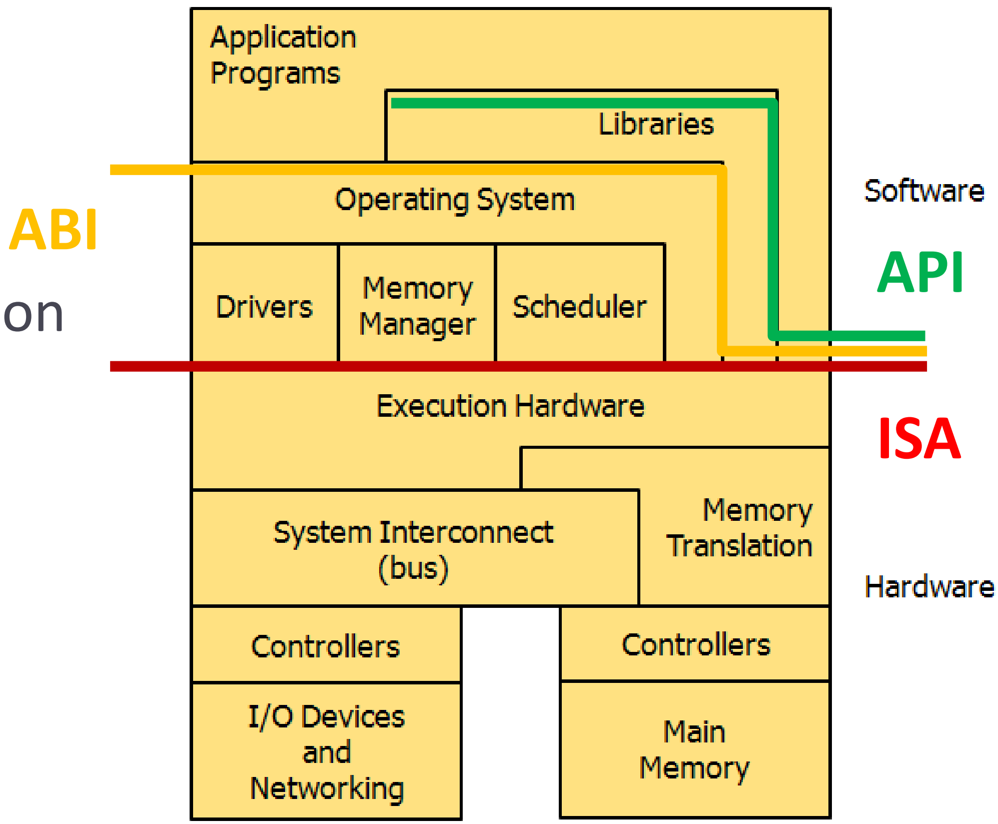
Hardware is not (commonly) designed be multiplexed → Loss of isolation

What needs to be emulated for a VM? [OS, App]

- OS
  - OS issues instructions to control hardware devices
  - … interacts with hardware devices using "sensitive" instructions
  - Allocate and manage hardware resources on behalf of programs
  - … OS runs at higher privilege level than applications
  - Expose system call interface to applications
  - … implemented using low-level H/W interfaces
- Application
  - Relies on the system call interface, runs in unprivileged mode
  - Special instruction(s) to call into OS code
  - OS provides a program with the illusion of its own memory
    - Virtual address spaces (implemented via MMU) → isolation
      - from OS and other App's
  - Most instructions run directly on the CPU
    - Sensitive instructions cause the CPU to throw an exception to the OS

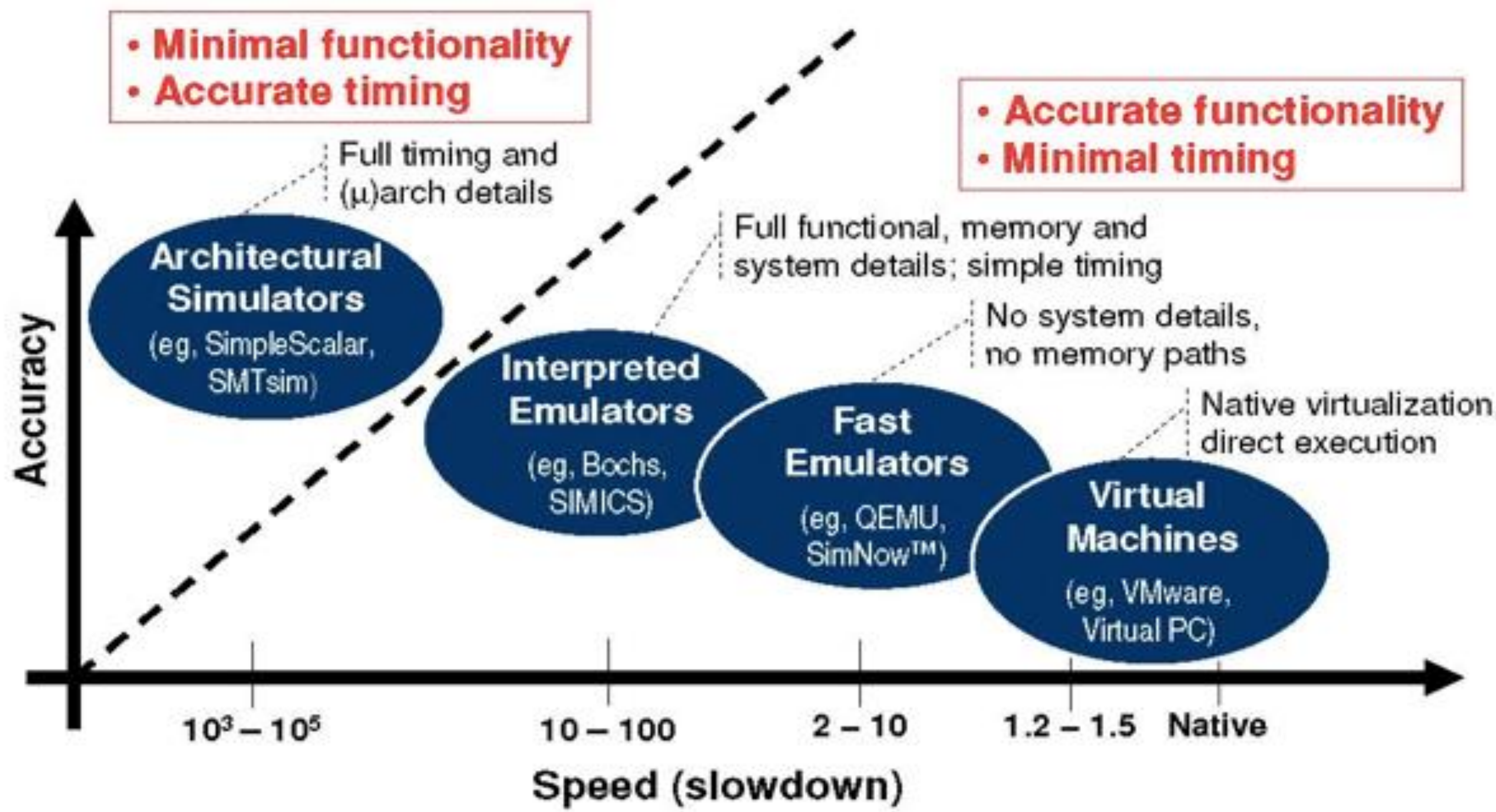
Computing systems are built on levels of abstraction

- Different perspectives on what a “machine” is
  - OS → **ISA**: Instruction Set Architecture
    - h/w – s/w interface
  - Compiler → **ABI**: Application Binary Interface
    - User ISA + OS calls
    - Calling conventions
  - Application → **API**: Application Programming Interface
    - User ISA + Library calls



By Glenford Myers (1982)

Virtualization alternatives & their performance



Embedded Linux System : Outline

Custom Application(s)

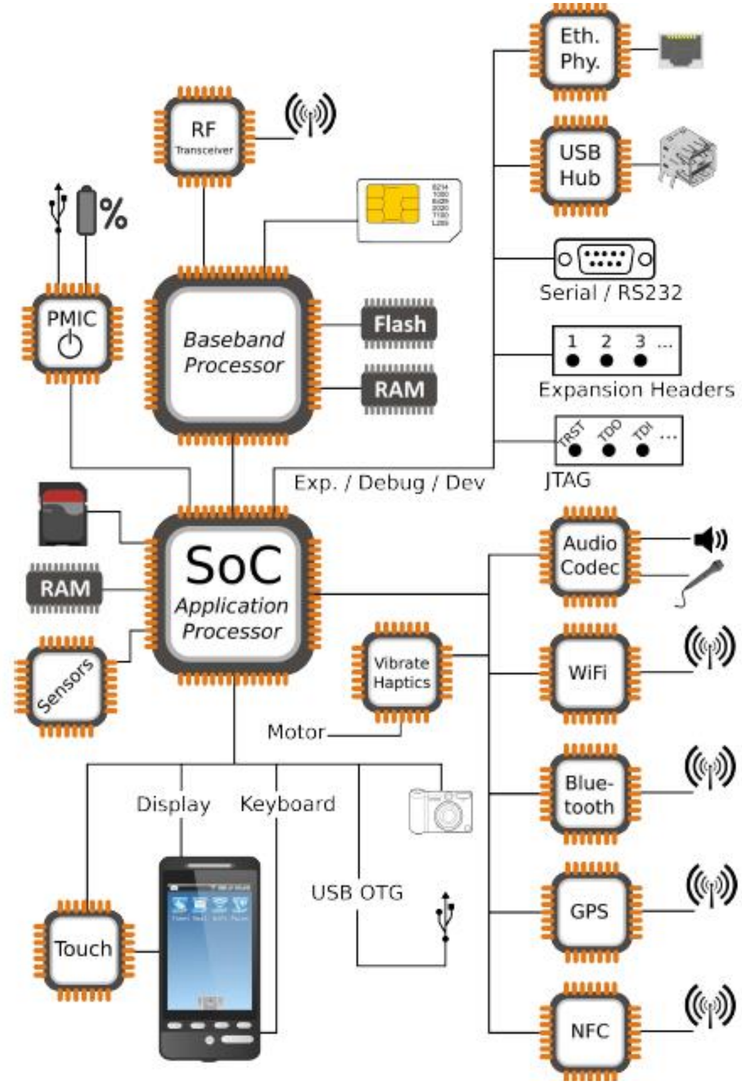
BusyBox

Libc

uClibc / eglibc / musl / glibc

Linux kernel

Process & Memory mgmt, HW support, …



Embedded Linux system development

+ Build Host to create cross-chain:
binutils, kernel headers, C/C++ libraries, gcc, gdb

Development Host

- **Cross-compilation toolchain**
- **Debugger**
- **Misc. tools**

JTAG, Serial

Ethernet

Embedded System Target

Application

Application

Library

Library **Library**

C Library

Linux OS kernel + drivers

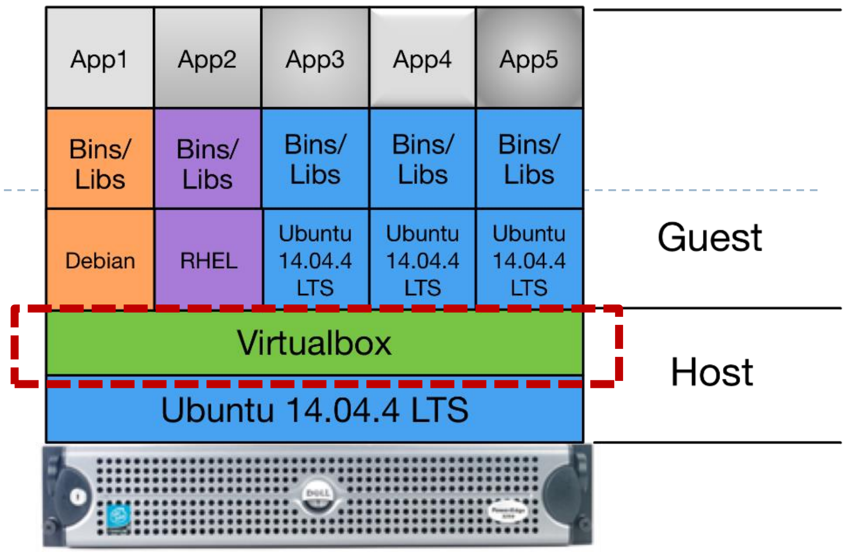
Bootloader

- Board support package (BSP) development
- System integration
- Development of applications

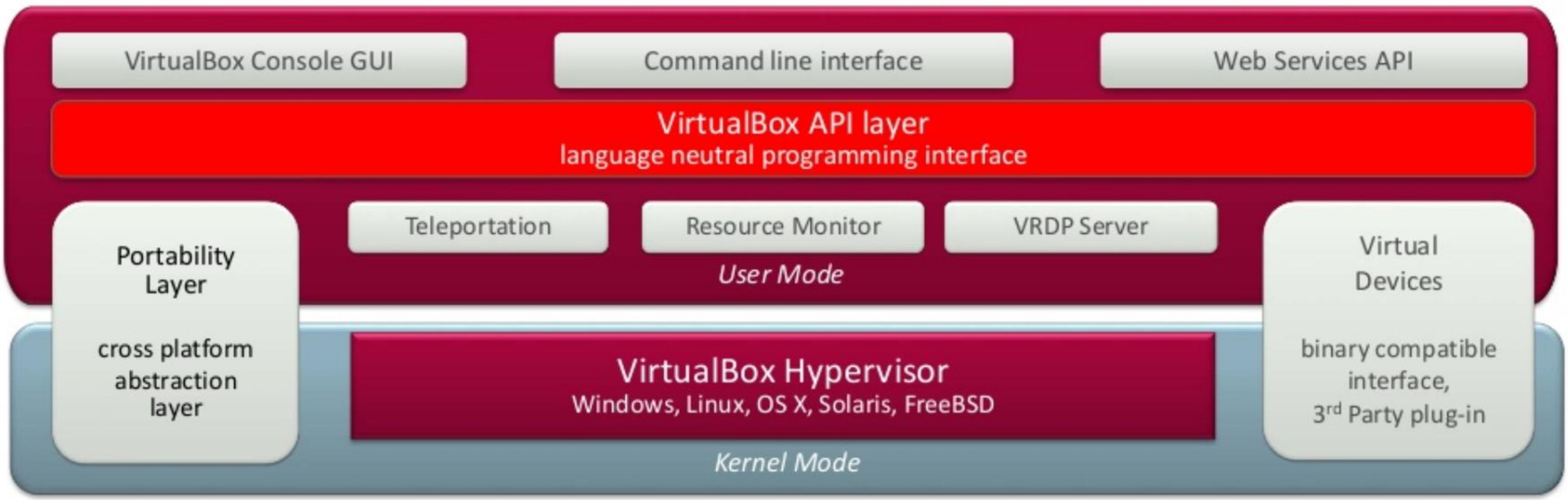
Embedded Linux := the usage of the Linux kernel and various open-source components in embedded systems

Hypervisor “mysteries”

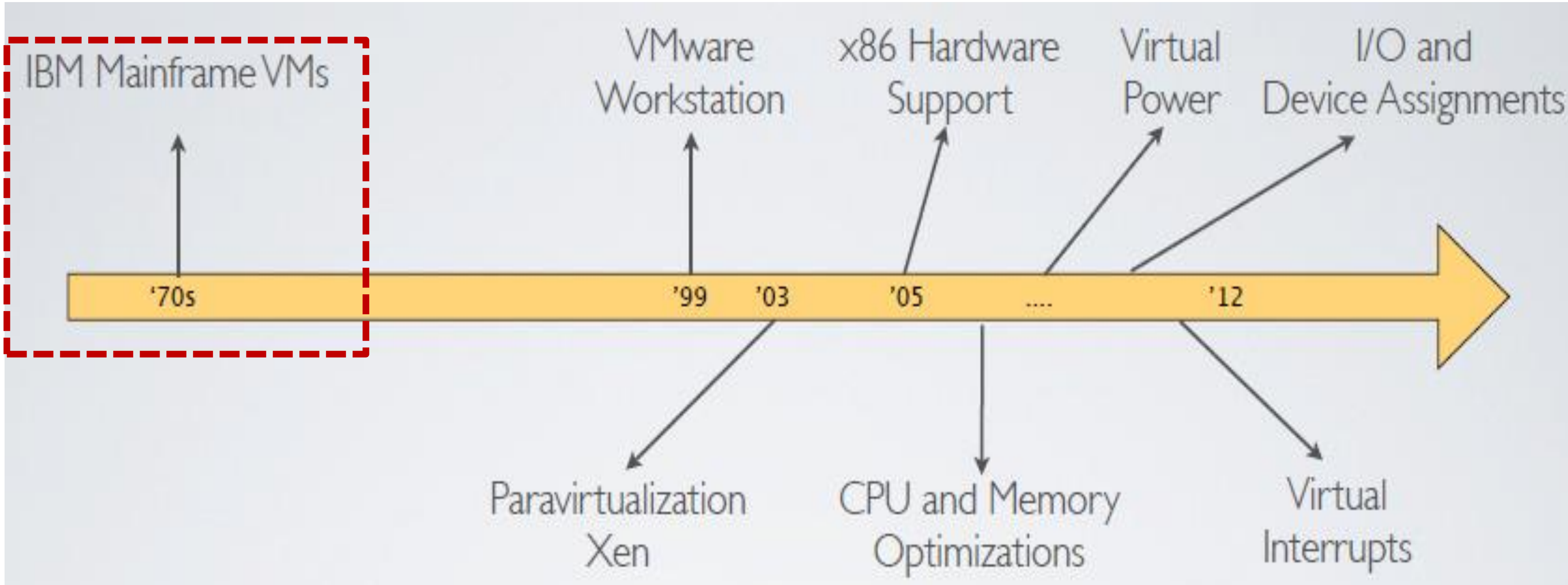
- Virtual Machine Taxonomy
  - Process virtual machines
    - Multiprogrammed systems
    - Emulators and dynamic binary translation
    - High-level-language virtual machines
  - **System virtual machines**
    - “Classic” virtual machines
    - Hosted virtual machines
    - Whole-system virtual machines
- Key virtualization techniques



Virtual Machines



Virtualization Timeline (C. Dall – 2013)

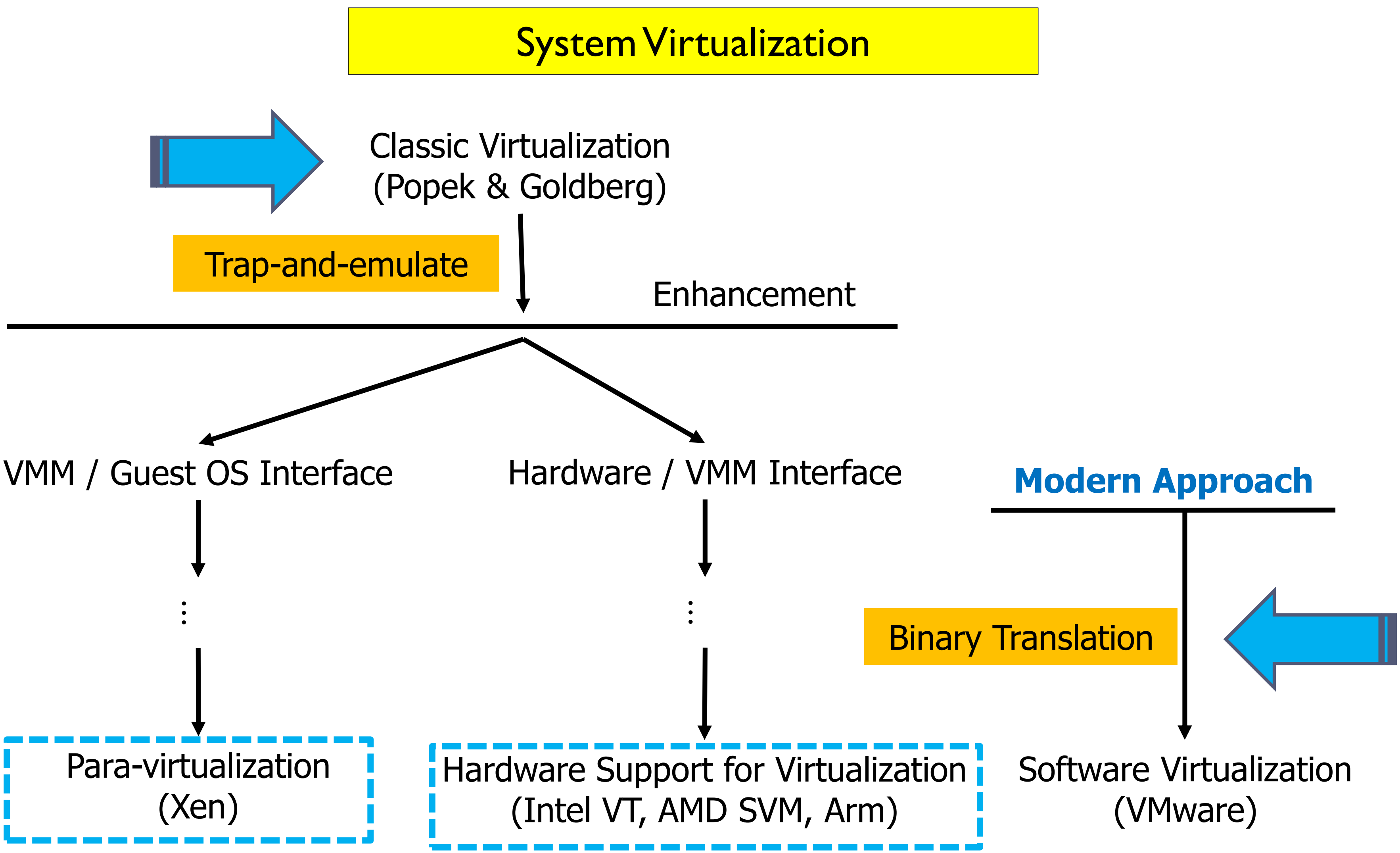


Virtual machines were popular in 60s-70s : IBM OS/370

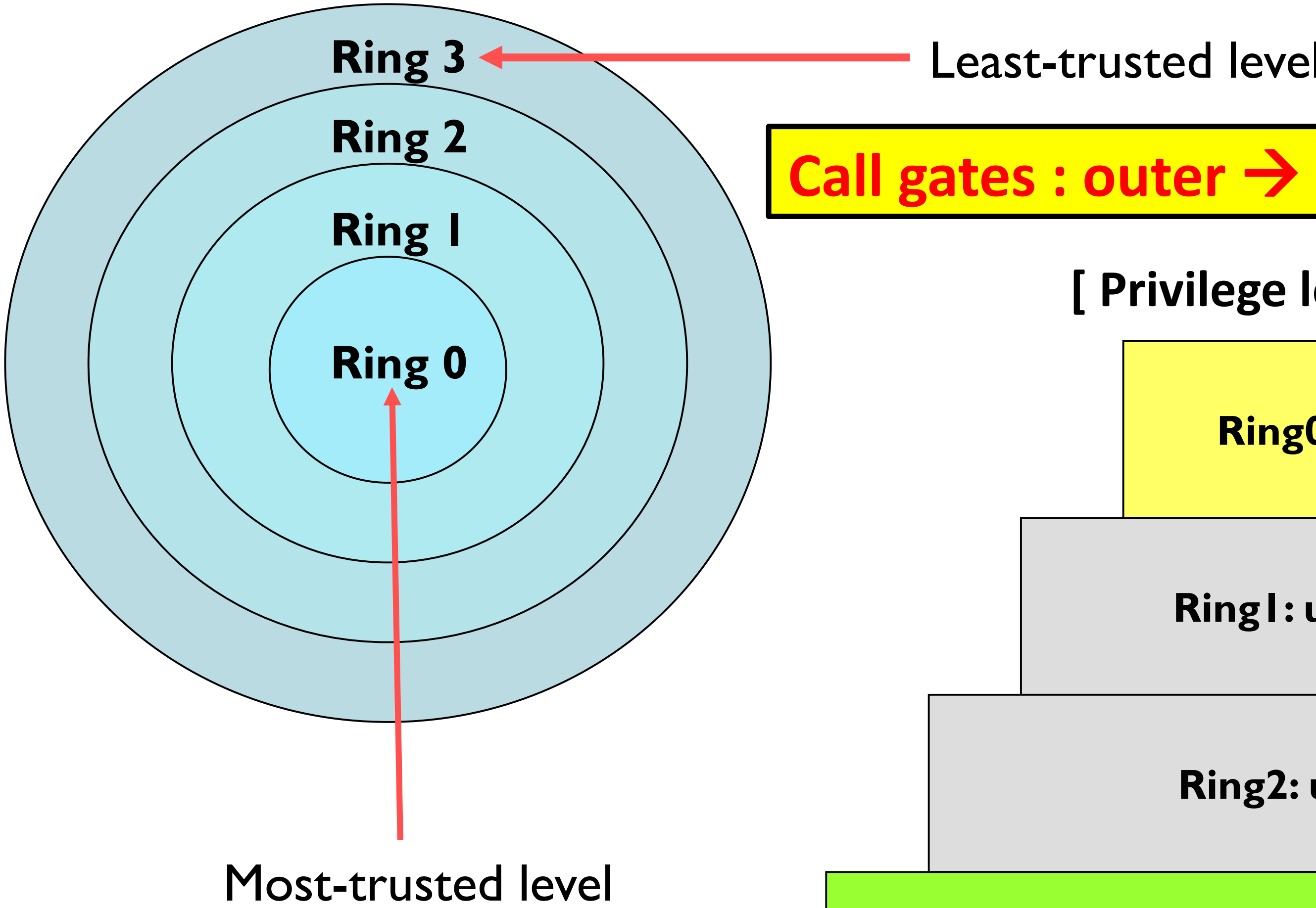
CP/CMS, 1967 → zSeries, 1972 → PR/SM (LPAR), 1985

- Share resources of mainframe computers to run multiple single-user OSs
- Interest is lost by 80s-90s: development of multi-user OS, rapid drop in H/W cost
- Hardware support for virtualization is "lost" … until the late 90s (VMware)

Evolution of System Virtualization

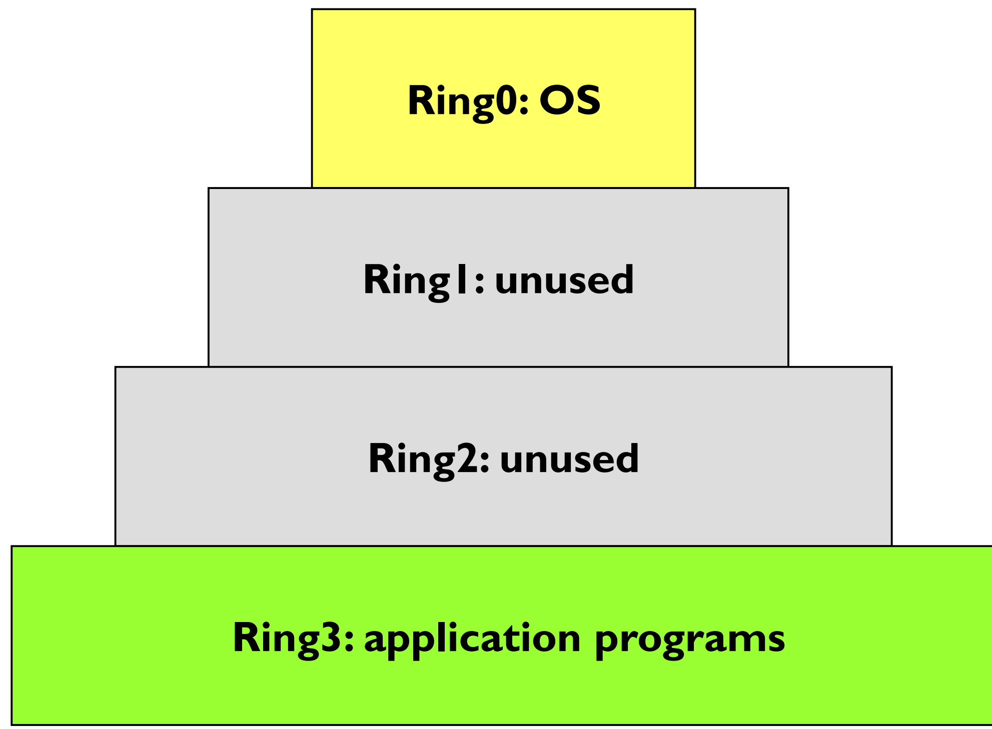


Processor privilege levels (x86) : Ring Transitions



Call gates : outer → inner ring transition

[Privilege levels in Linux]

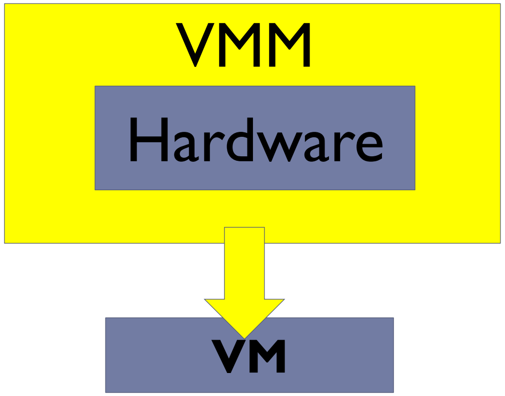


Processor privilege levels (x86) : Sharing vs Isolation

- Procedure-calls typically require that two separate routines share data-values (e.g., parameter-values get passed from the caller to the callee)
  - To support reentrancy and recursion, the processor's stack is frequently used as a shared-access storage-area
  - Among routines with different levels of privilege, this would create a "security hole" !
- To guard against unintentional sharing of privileged information, different stacks are provided for each "ring"
- Transition from one ring to another must necessarily be accompanied by a "<u>stack-switch</u>" operation
  - The CPU provides for automatic switching of stacks and copying of parameter-values
  - Special instructions ("far calls") and "call gates" (control data structures, in protected memory)

“Classic” VM (Popek & Goldberg, 1974) (1/4)

- Essentials of a Virtual Machine Monitor (VMM)
  - An efficient, isolated duplicate of the real machine.
  - **Equivalence**
    - Software on the VMM executes identically to its execution on hardware, barring timing effects. i.e. Running on VMM == Running directly on HW



  - **Performance**
    - Non –Privileged instructions can be executed directly by the real processor, with no software intervention by the VMM.

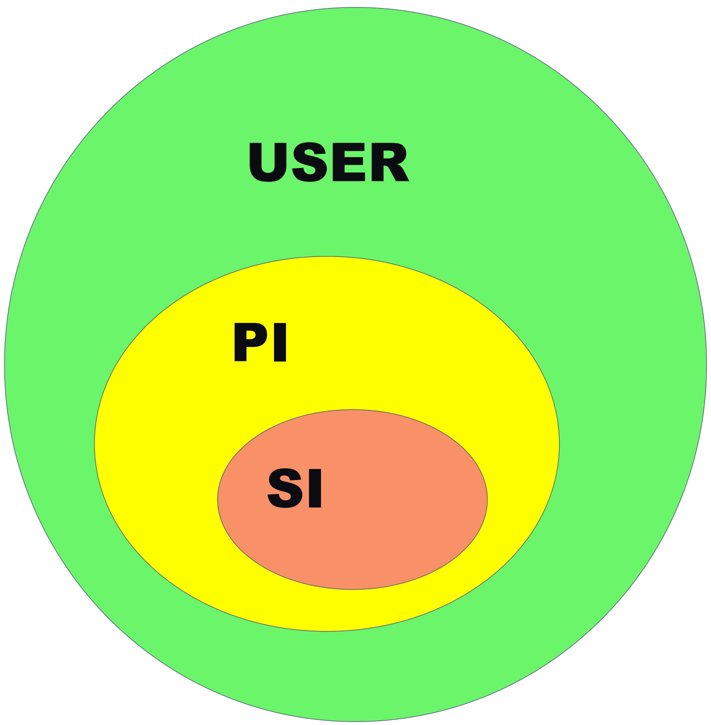
      i.e. Performance on VMM == Performance on HW
  - **Resource control**
    - The VMM must have complete control of the virtualized resources.

“Classic” VM (Popek & Goldberg, 1974) (2/4)

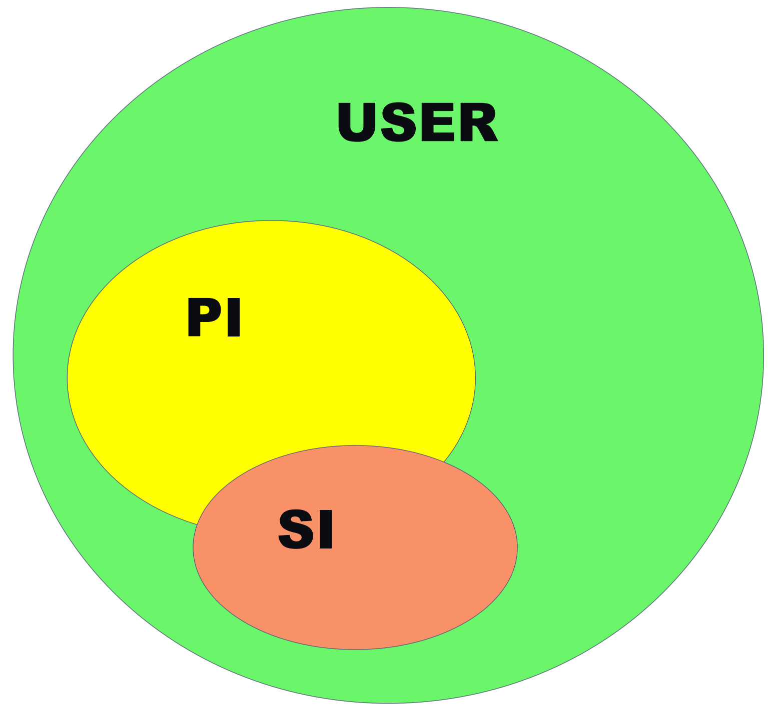
- Instruction types
  - Privileged instructions: generate trap when executed in any but the most-privileged level
    - Execute in privileged mode, trap in user mode
    - E.g. x86 LIDT : load interrupt descriptor table address
  - Privileged state: determines resource allocation
    - Privilege mode, addressing context, exception vectors, …
  - Sensitive instructions: instructions whose behavior depends on the current privilege level, or modify H/W state
    - Control sensitive: change privileged state
    - Behavior sensitive: exposes privileged state
    - E.g. x86 POPF : pop stack to EFLAGS (in user-mode, the ‘interrupt enable’ bit is not over-written)

"Classic" VM (Popek & Goldberg, 1974) (3/4)

Theorem 1: A VMM may be constructed if the set of SI's is a subset of the set of PI's



ISA is Virtualizable



ISA is NOT Virtualizable

“Classic” VM (Popek & Goldberg, 1974) (4/4)

- To build a VMM, it is sufficient for all instructions that affect the correct functioning of the VMM (SI’s) always trap and pass control to the VMM.
  - This guarantees the “resource control property”
  - Non-privileged instructions are executed without VMM intervention
  - Equivalence property: We are not changing the original code, so the output will be the same.

Mostly-virtualizable Architectures ☹

- x86
  - Sensitive push/pop instructions are not privileged
  - Segment and interrupt descriptor tables in virtual memory
- Itanium
  - Interrupt vectors table in virtual memory
- MIPS
  - User-accessible kernel registers k0, k1 (save/restore state)
- ARM
  - PC is a general-purpose register
  - Exception returns to PC (no trap)

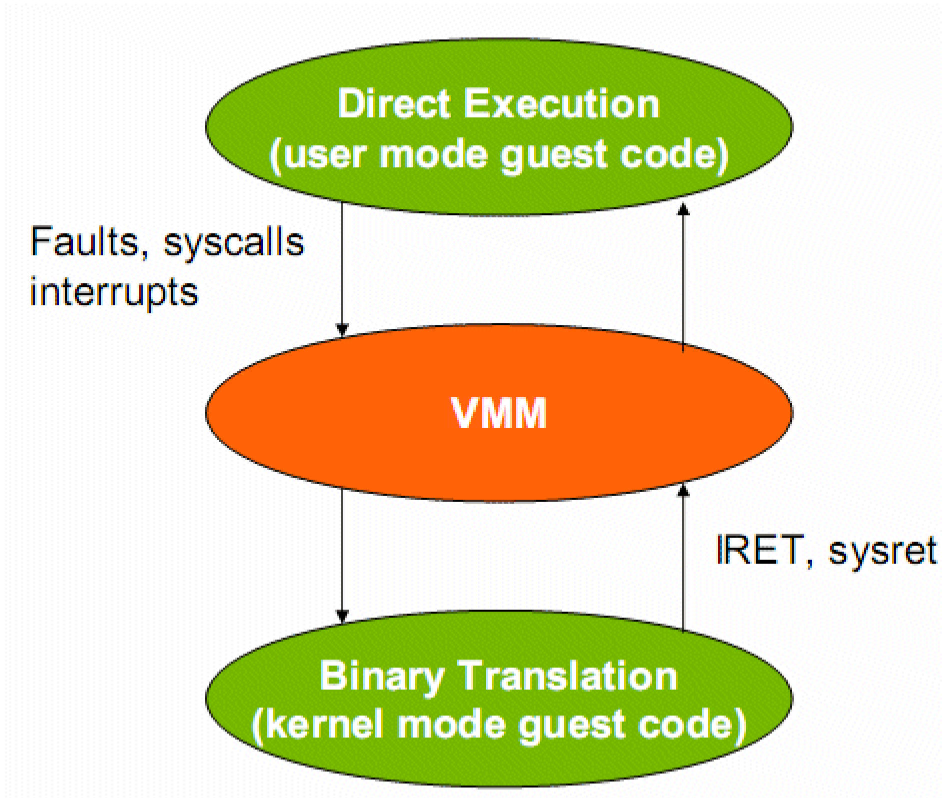
Virtualization overheads

- VMM maintains virtualized privileged machine state
  - Processor status, addressing context, device state, …
- VMM emulates privileged instructions
  - Translation between virtual and real privileged state
    - E.g. guest-to-real page tables
- Traps are expensive
  - Several 100s cycles (for x86)
- Certain important OS operations involve several traps
  - Interrupt enable/disable for mutual exclusion
  - Page table setup/updates for fork()

How to achieve safe –and- fast virtualization?

- Emulation
  - Interpret each instruction
- Paravirtualize
  - Modify the guest OS to avoid non-virtualizable instructions
- Binary translation (instead of trap-and-emulate)
  - Static vs Dynamic
- Change processor architecture
  - Intel VT , AMD Pacifica → extend x86 to make "Classic Virtualization" possible [VM/370 origins !]
  - Add a new CPU mode to distinguish VMM from guest/app

Binary Translation

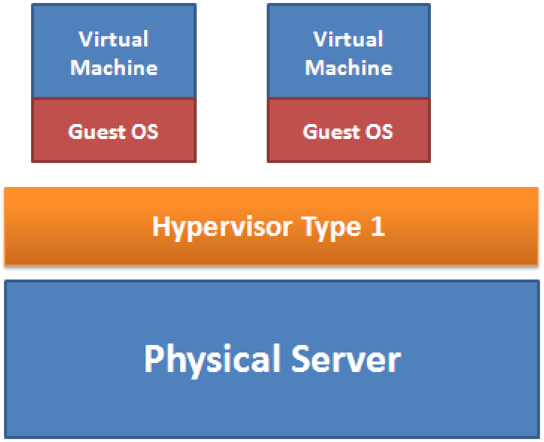


+ translator cache
+ trace cache

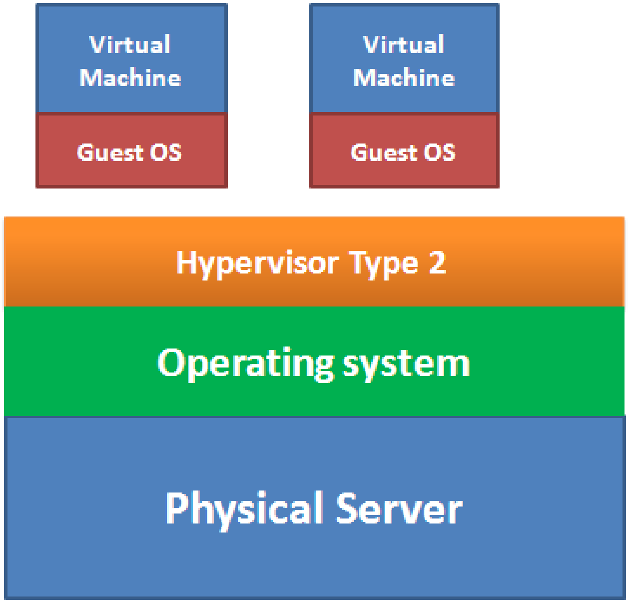
- User applications are not translated, but run directly.
- Binary Translation only happens when the guest OS kernel gets called.

Hypervisor (VMM) types

- Type I: run directly on hardware (minimal OS)
  - Bare-metal (minimal OS)
  - e.g. XEN (Citrix XenServer), Microsoft Hyper-V, IBM LPAR, VMware ESXi (vmkernel)
  - Monolithic (kernel + device drivers+ I/O stack)
  - Microkernel –based: I/O stack and HW-specific device drivers in “parent” partition

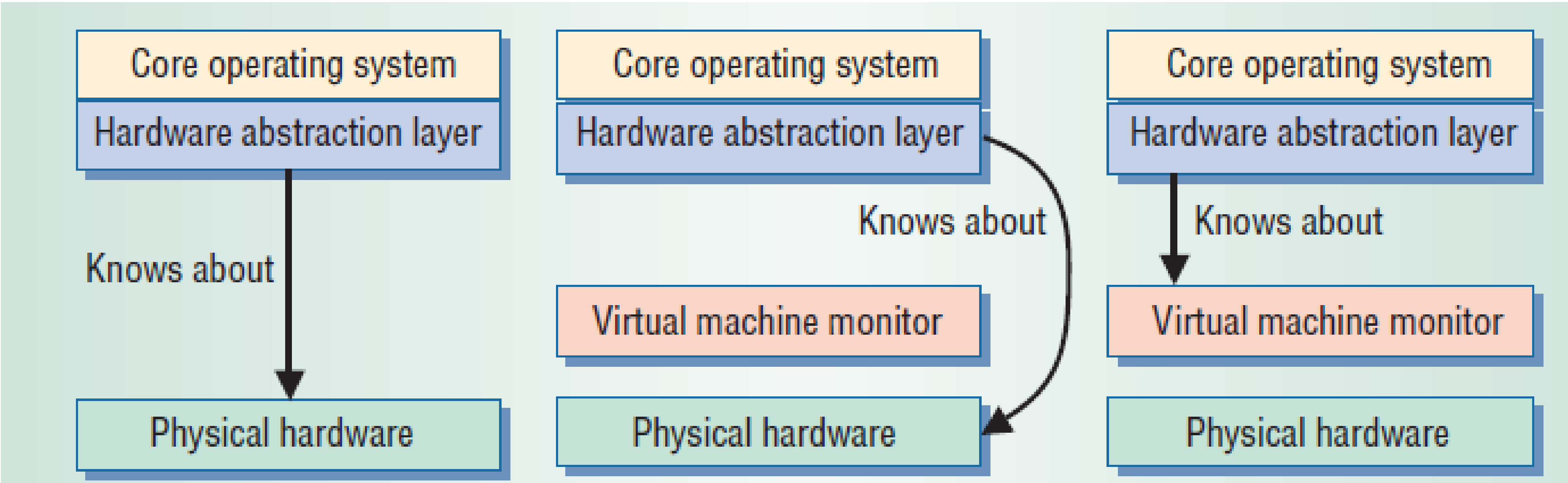


- Type II: run on host OS
  - Hosted
  - one user-space process, or one user-space process per VM
  - e.g. VMware Workstation, VirtualBox, KVM (Linux), QEMU



FORTH-ICS Institute of Computer Science

VMM architectures



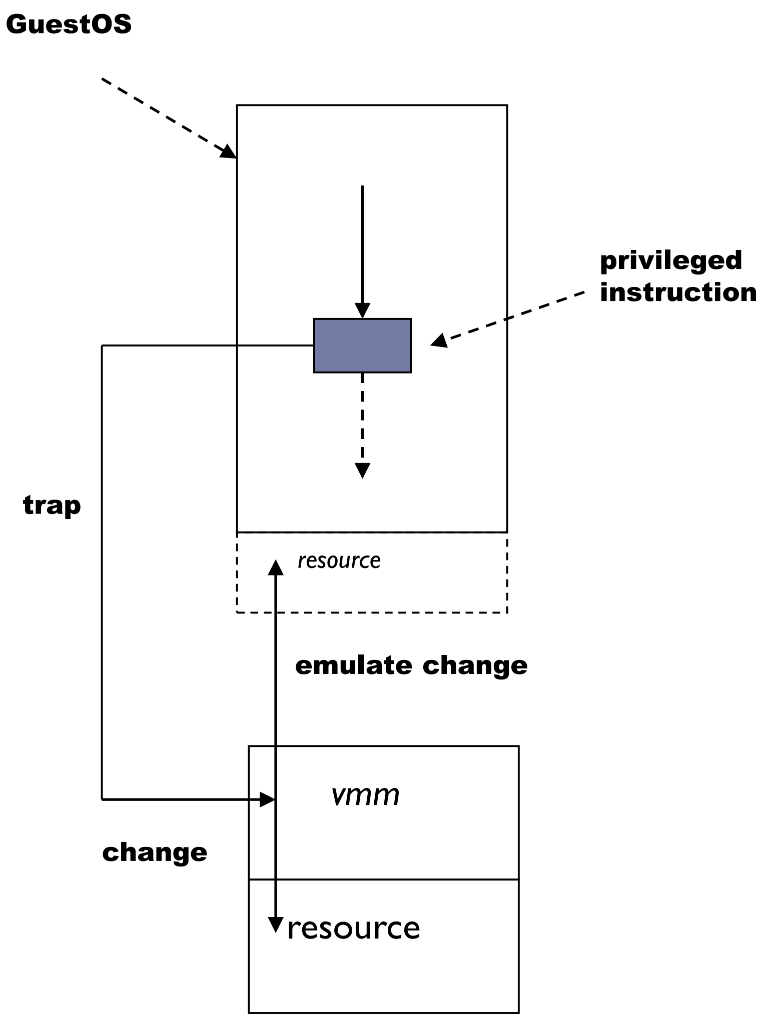
Only OS knows about H/W | **Unmodified view of H/W** | **Modified view of H/W**

Paravirtualized VMM

VMM provides a HW/SW interface to guest OSs :

- Full virtualization: trapping & emulating sensitive instructions
- Para-virtualization: OS-assisted ("hyper-calls")
- HW-accelerated virtualization (unmodified Guest OS)

Key Techniques (1/3): De-privileging

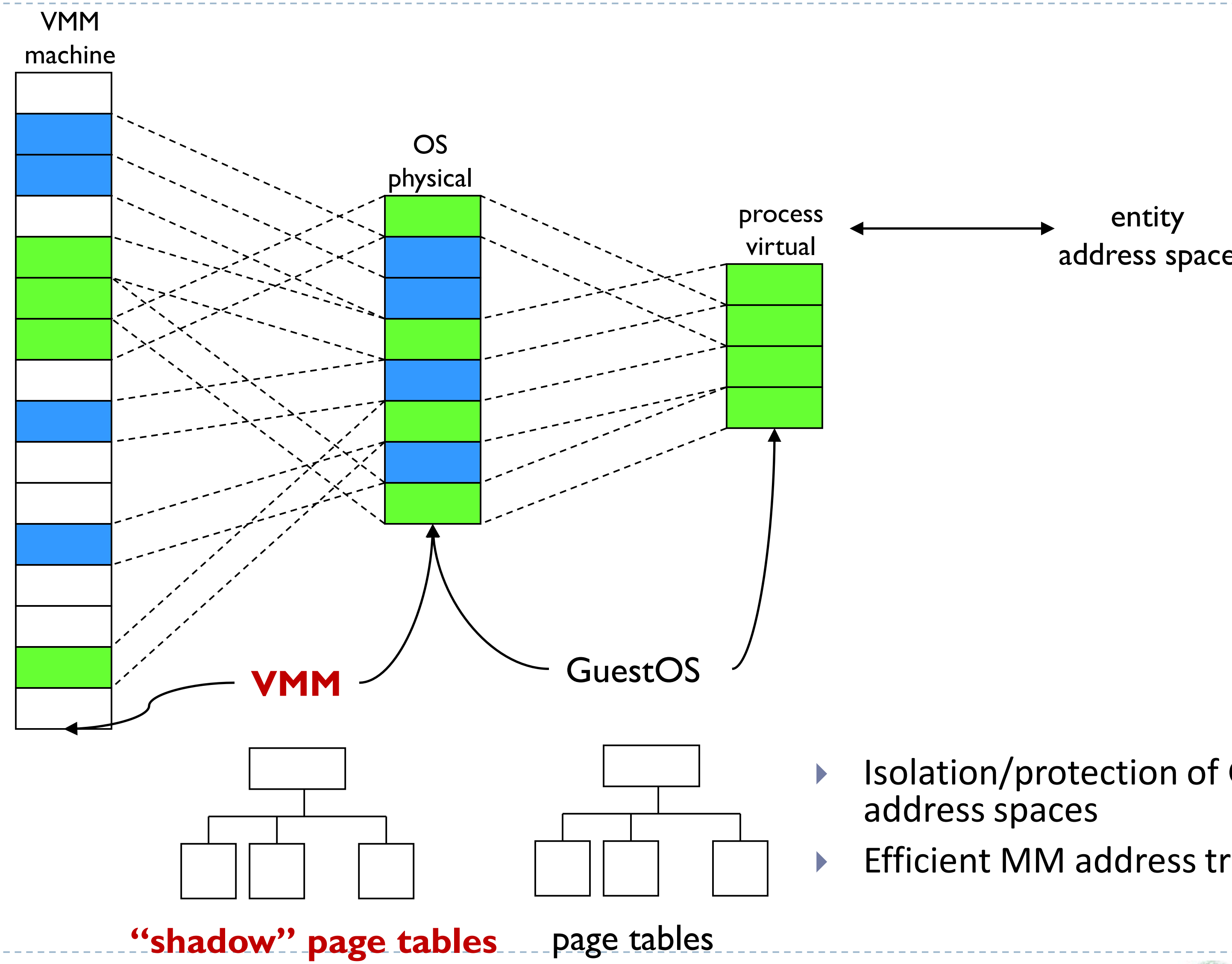


- VMM emulates the effect on system/hardware resources of privileged instructions whose execution traps into the VMM
  - aka **trap-and-emulate**
- Typically achieved by running GuestOS at a lower hardware priority level than the VMM
  - "Normal" instructions run directly on processor
  - "Privileged" instructions trap into VMM (for safe emulation)
- Problematic on architectures where privileged instructions do not trap when executed at deprivileged priority!

Key Techniques (2/3): Primary vs Shadow Structures

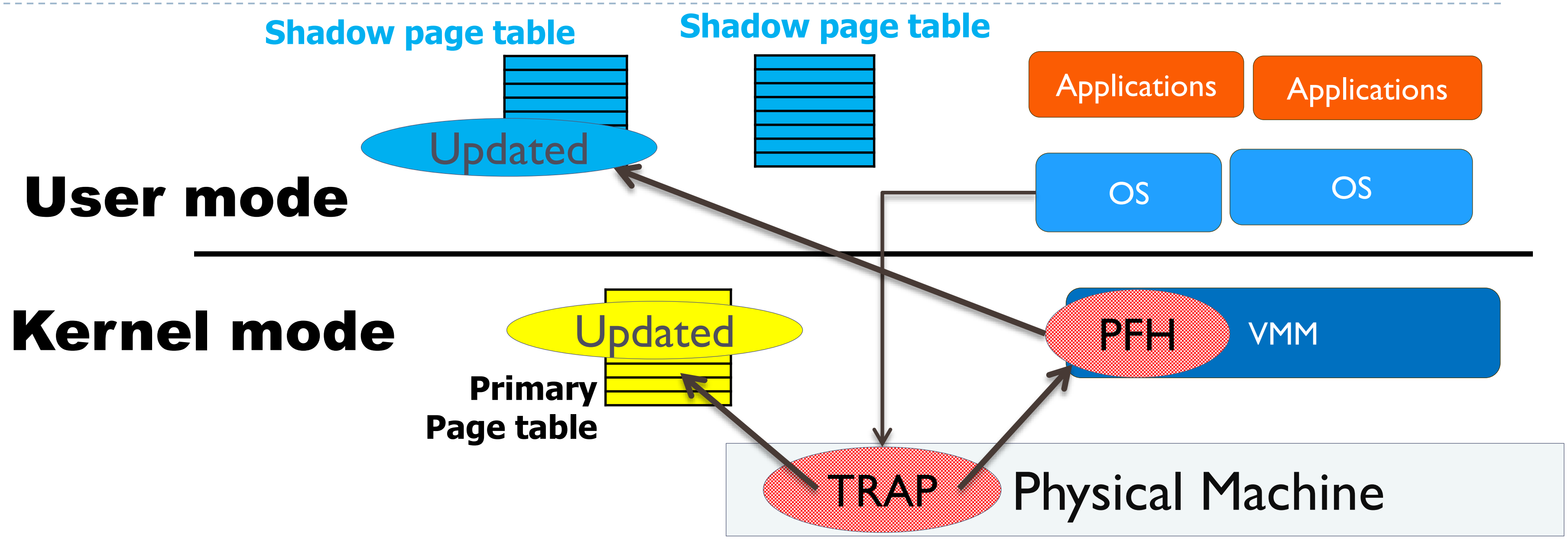
- VMM maintains “shadow” copies of critical structures whose “primary” versions are manipulated by GuestOS
  - e.g., page tables
- Primary copies needed to insure correct environment visible to GuestOS

Memory Management by the VMM



- Isolation/protection of Guest OS address spaces
- Efficient MM address translation

Key Techniques (3/3): Memory Tracing (Trace faults)

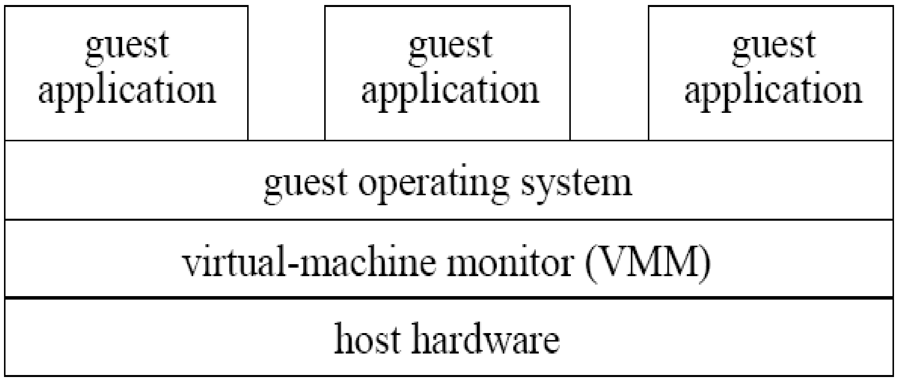


- Control access to memory so that the shadow and primary structures remain coherent
  - Write-protect primary structure so that update operations cause page faults → caught, interpreted, emulated by the VMM
  - VMM typically use hardware page protection mechanisms to trap accesses to in-memory primary structures

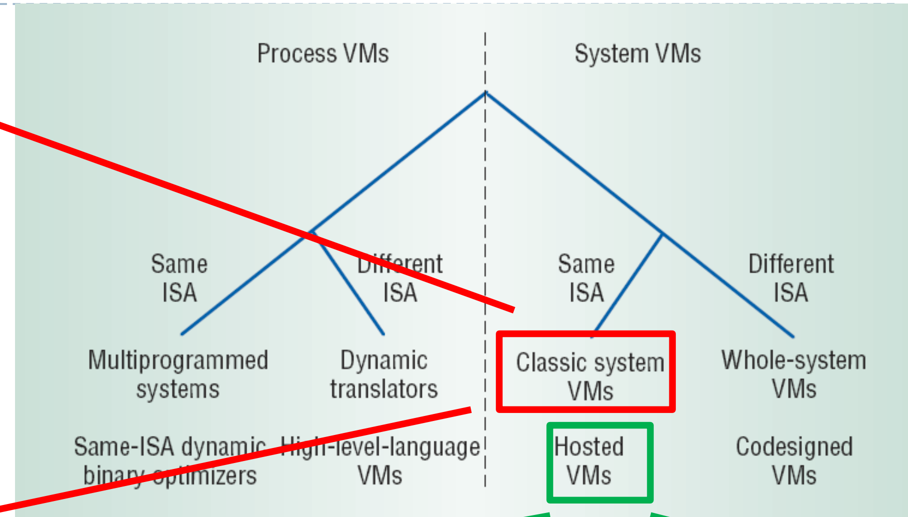
Sources

- James E. Smith, Ravi Nair, **The Architecture of Virtual Machines**, IEEE Computer, vol.38, no.5, May 2005
- Mendel Rosenblum, Tal Garfinkel, **Virtual Machine Monitors: Current Technology and Future Trends**, IEEE Computer, May 2005.
- A. Whitaker, R.S. Cox, M. Shaw, S.D. Gribble, **Rethinking the Design of Virtual Machine Monitors**, IEEE Computer, vol.38, no.5, May 2005.
- Kirk L. Kroeker, **The Evolution of Virtualization**, CACM, vol.52, no. 3, March 2009
- G.J. Popek, and R.P. Goldberg, **Formal Requirements for Virtualizable Third Generation Architectures**, CACM, vol. 17 no. 7, 1974.
- Jim Smith and Ravi Nair, **Virtual Machines: Versatile Platforms for Systems and Processes**, ISBN-10: 1558609105, Elsevier, 2005

System VMMs



Type 1

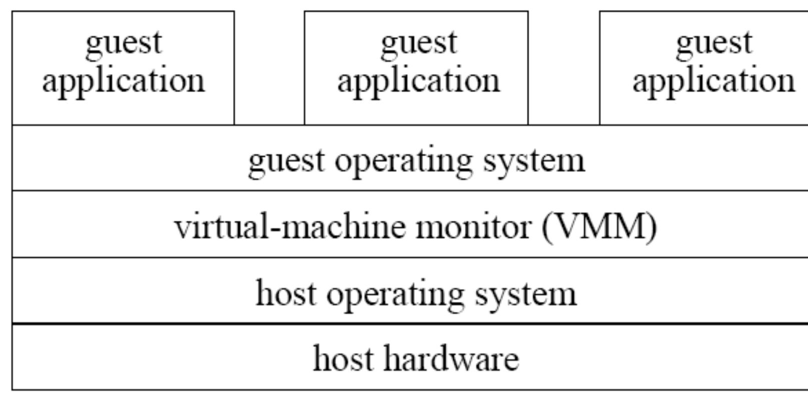


Type 1: runs directly on hardware

- primary goal: performance
- Examples: OS/370, VMware ESXi

Type 2: runs on host OS

- primary goal: ease of installation
- Example: User-Mode Linux, VMware Workstation

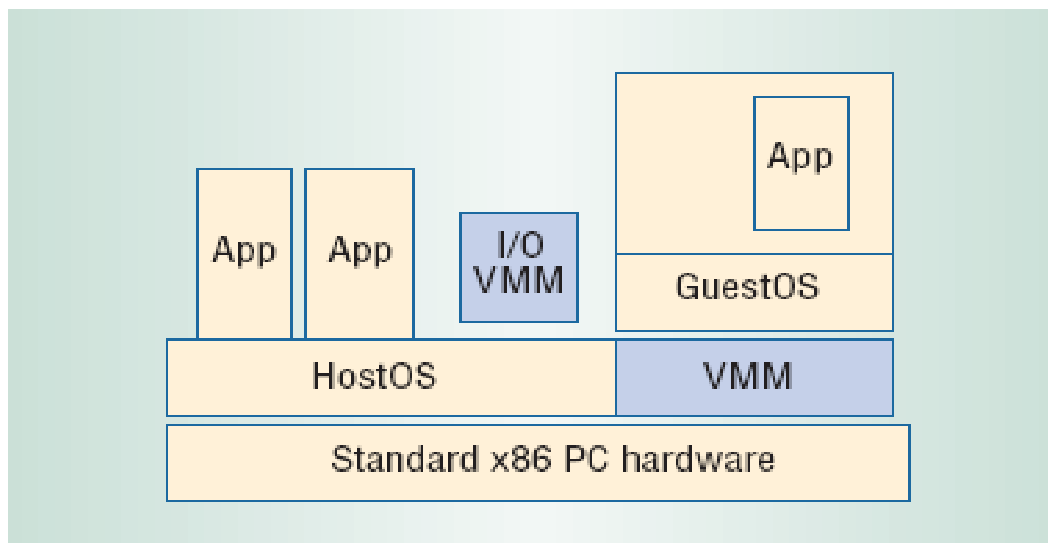


Type 2

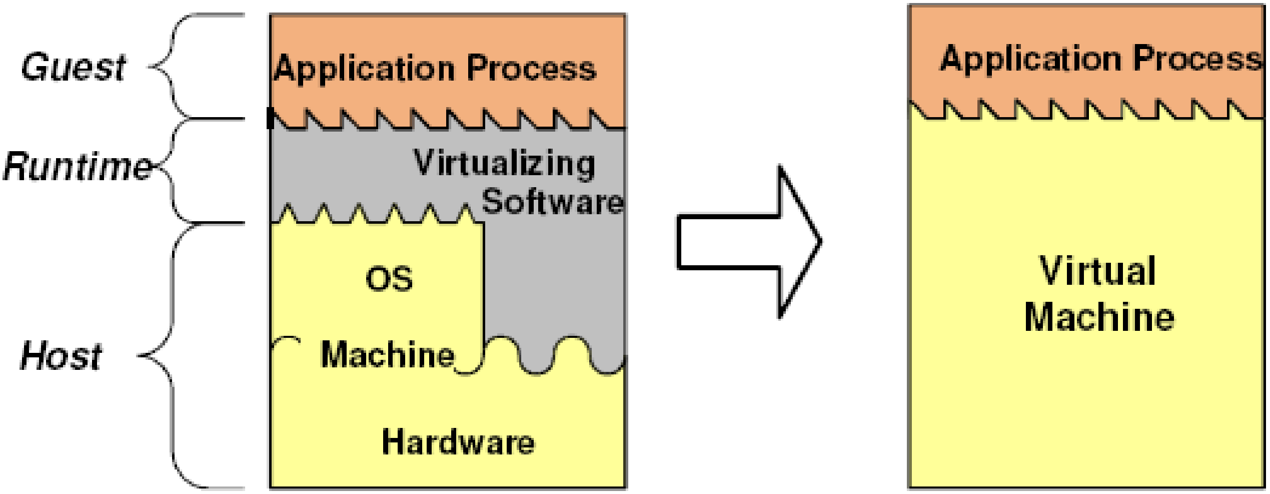
Hosted VMMs

- Hybrid between Type 1 and Type 2
  - “Core VMM” runs directly on hardware
    - Improved performance as compared to “pure Type 2”
    - Leverage s/w engineering investment in host OS for I/O device support
  - I/O services provided by host OS
    - Overhead for I/O operations, reduced performance isolation

Example: VMware Workstation



Process vs System VM



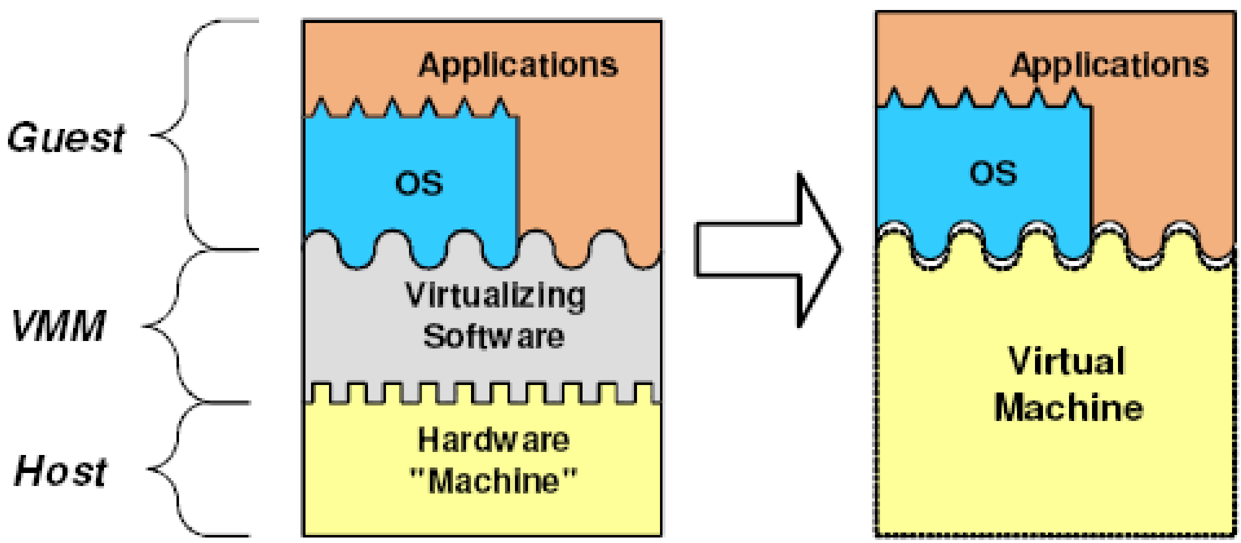
Process:

Provides API interface

+ Easier to install

+ Leverages OS services – e.g. device drivers

- Execution overhead



System:

Provides ABI interface

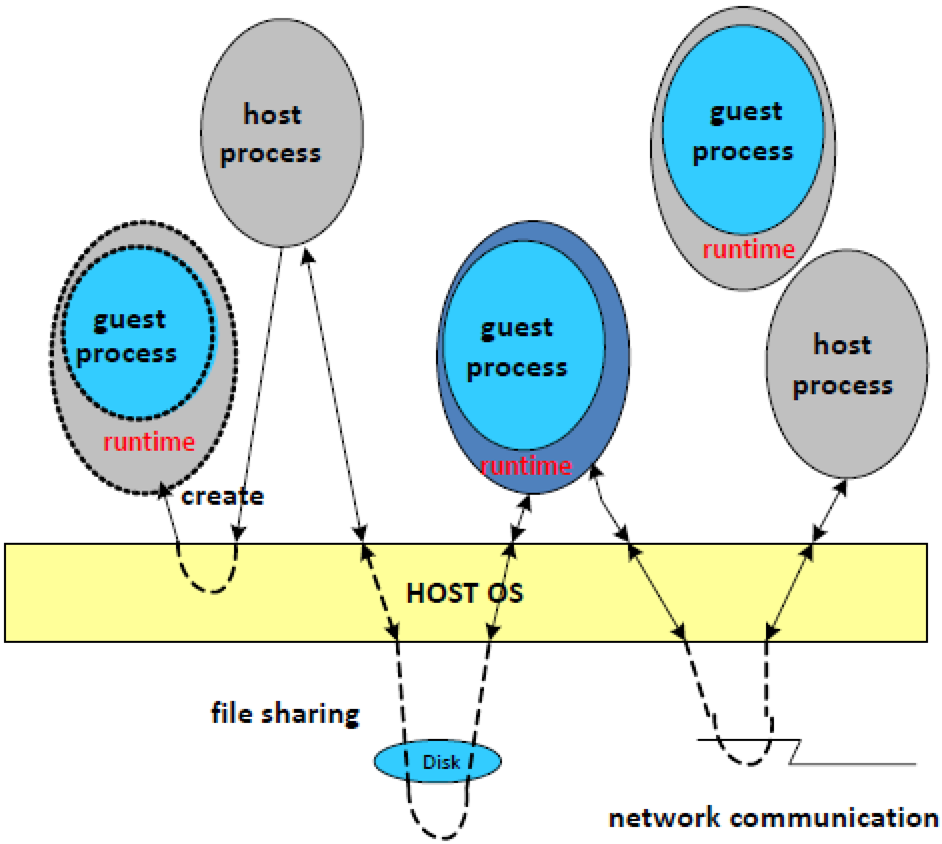
+ Efficient execution

+ Can add OS-independent services – e.g. migration, checkpointing, sandbox

FORTH-ICS
Institute of Computer Science

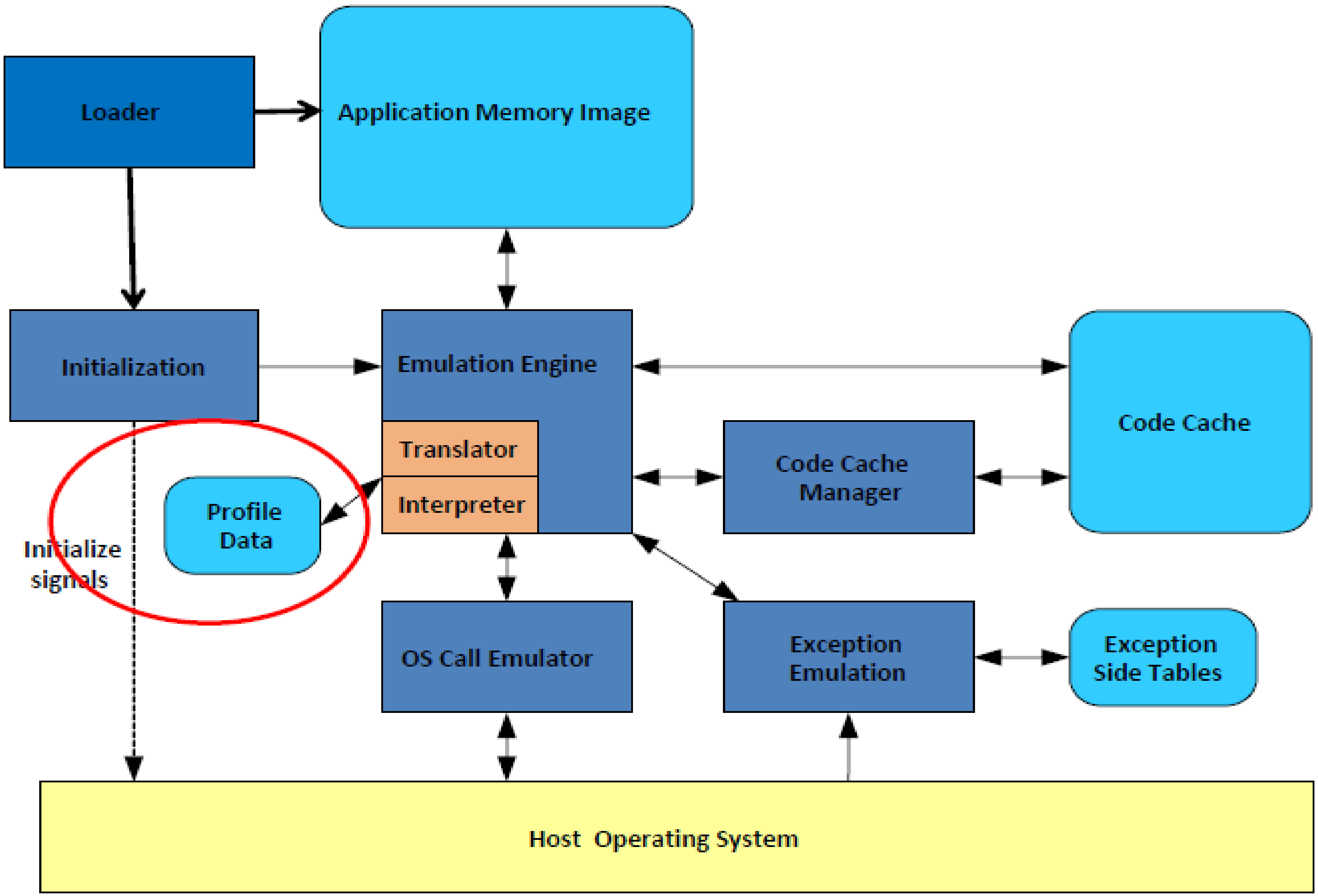
Process VM concept

- A guest program developed for a machine (ISA and OS) other than the user's host system can be used in the same way as all other programs in the host system
- **Runtime system**
  - Encapsulates an individual guest process giving it the same appearance as a native host process
  - All host processes appear to conform to the guest's worldview



FORTH-ICS
Institute of Computer Science

Process VM architecture



Acceleration techniques

- Binary translation
  - locate sensitive instructions in guest binary and replace on-the-fly with emulation code or hypercall
    - VMware, QEMU
- Para-virtualization
  - Port the GuestOS to modified ISA
    - Xen, L4, Denali, Hyper-V
    - Reduce number of traps
    - Remove un-virtualizable instructions
- Hardware support for virtual machines
  - Make all sensitive instructions privileged (!)
  - Intel VT-x, AMD SVM
    - Xen, VMware, kvm
  - Nested page tables
  - Direct device assignment, IOMMU, Virtual interrupts

Virtualization use-cases (mobile, media, automotive)

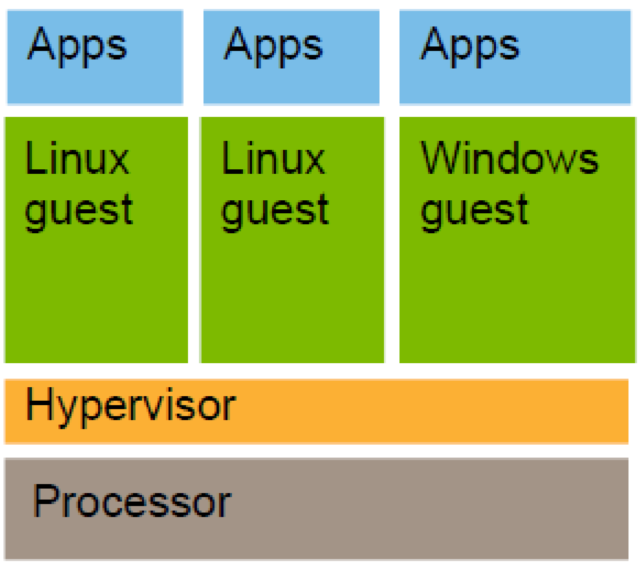
- Processor consolidation & dynamic allocation (eg. mobile)
- Software architecture abstraction (esp. for product series)
- Certification re-use
- License separation
- User-configured OS, Personal vs Enterprise environment
- Separation of systems code from applications
- IP protection and secure payments (eg. set-top box)
- Digital Rights Management … on open device
- Processor consolidation: control + infotainment
- Componentization
  - for IP blocks
  - For security

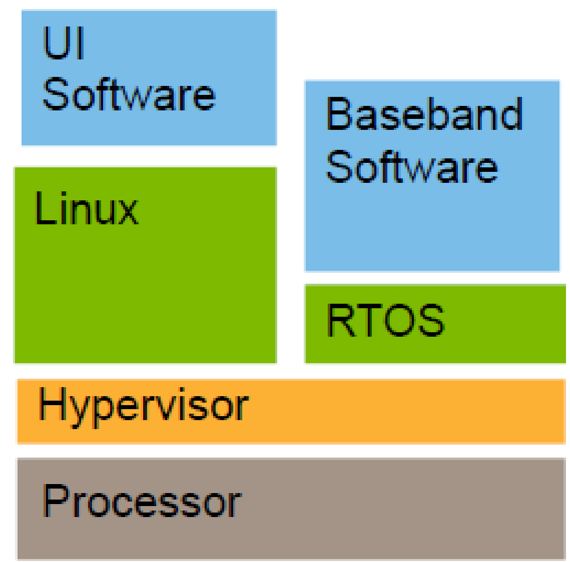
Enterprise vs Embedded Systems VMs

Homogenous vs heterogenous guests

→ Enterprise: many similar guests
- hypervisor size irrelevant
- VMs scheduled round-robin

→ Embedded: 1 HLOS + 1 RTOS
- hypervisor resource-constrained
- interrupt latencies matter





Isolation vs Cooperation

Enterprise

- → Independent services
- → Emphasis on isolation
- → Inter-VM communication is secondary
  - performance secondary
- → VMs connected to Internet (and thus to each other)

Embedded

- → Integrated system
- → Cooperation with protection
- → Inter-VM communication is critically important
  - performance crucial
- → VMs are subsystems accessing shared (but restricted) resources

Isolation vs Cooperation : Scheduling

Enterprise

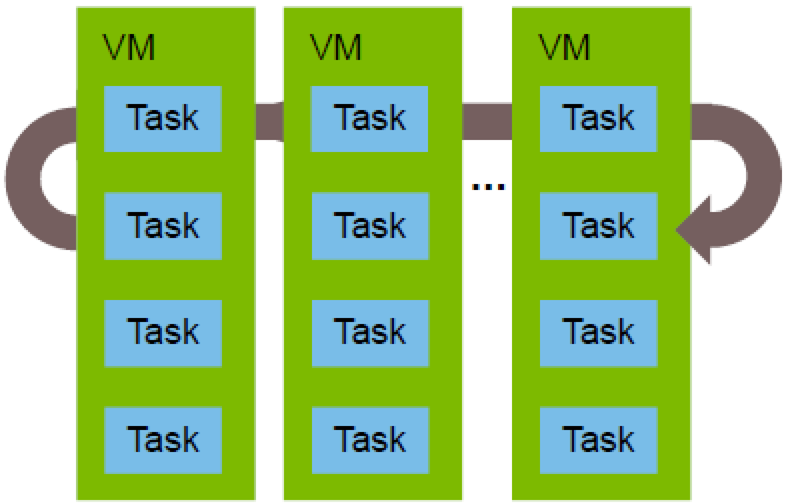
→ Round-robin scheduling of VMs

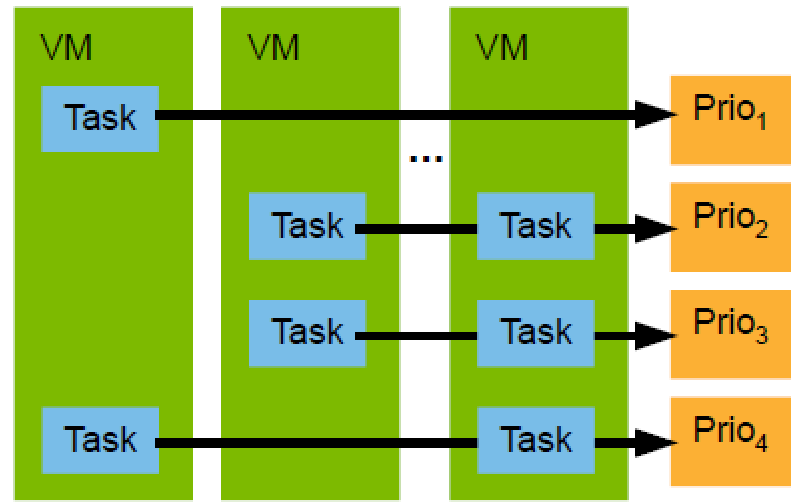
→ Guest OS schedules its apps

Embedded

→ Global view of scheduling

→ Schedule threads, not VMs



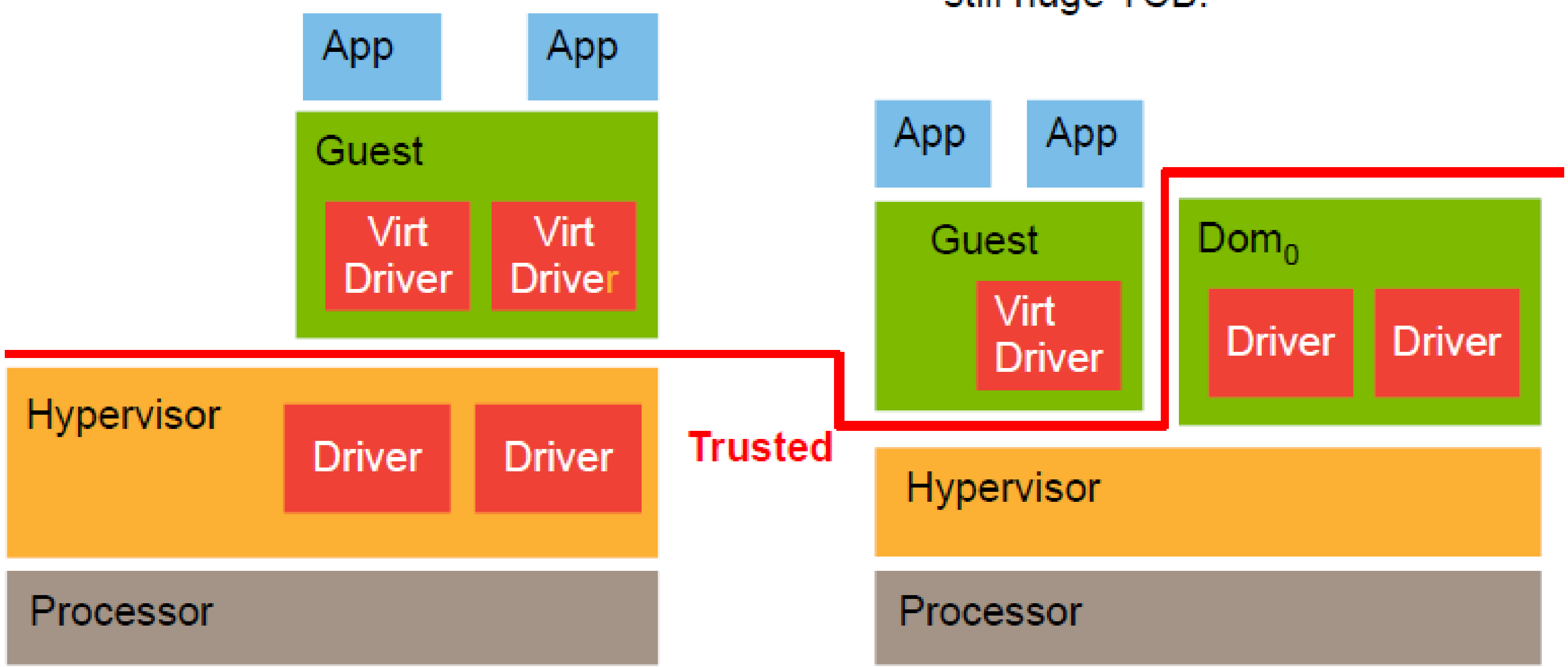


→ Similar for *energy management*:

- energy is a global resource
- optimal per-VM energy policies are not globally optimal

Devices in enterprise virtual machines

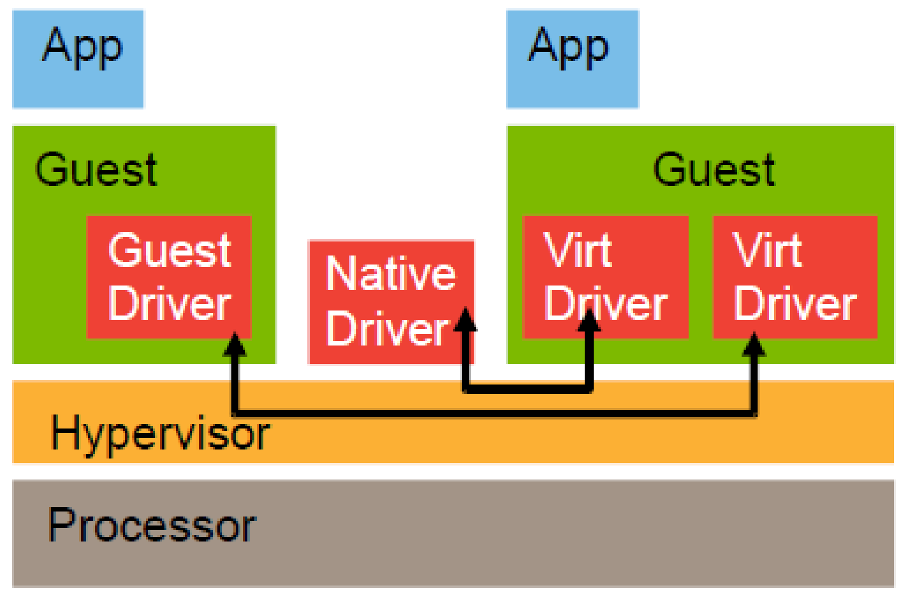
- → Hypervisor owns all devices
- → Drivers in hypervisor
  - need to port all drivers
  - huge TCB
- → Drivers in privileged guest OS
  - can leverage guest's driver support
  - need to trust driver OS
  - still huge TCB!



Devices in embedded virtual machines

- Some devices owned by particular VM
- Some devices shared
- Some devices too sensitive to trust any guest
- Driver OS too resource hungry

- Use isolated drivers
  - protected from other drivers
  - protected from guest OSes

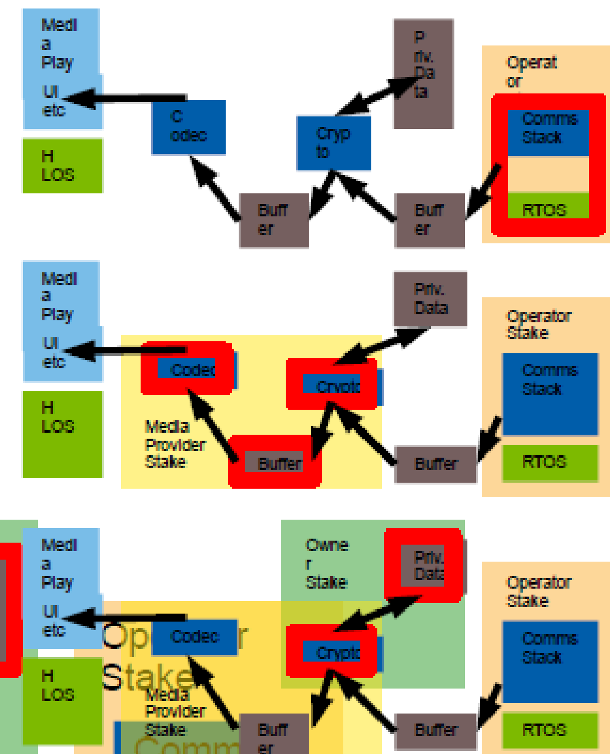


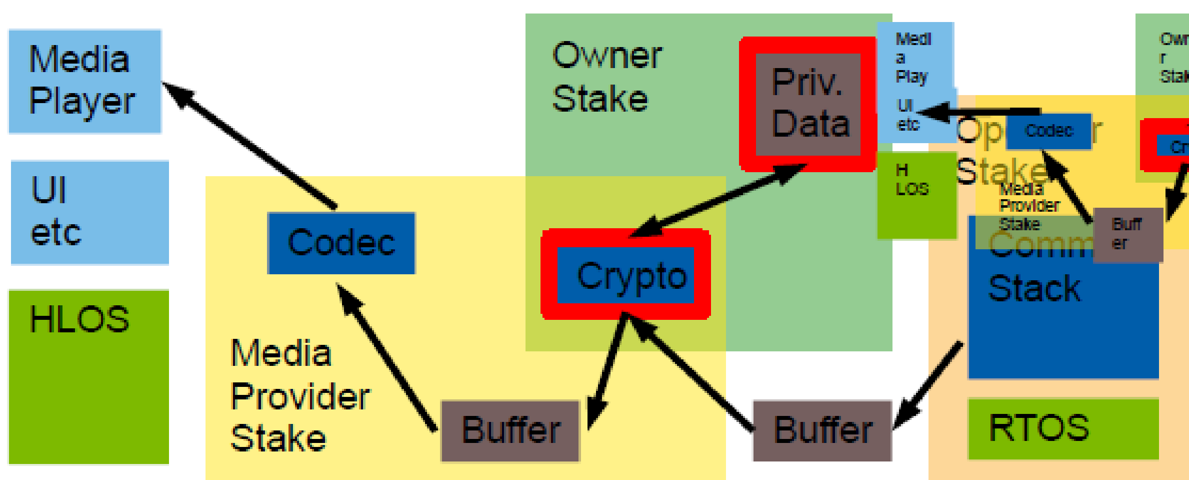
Inter-VM Communication

Modern embedded systems are multi-user devices!

→ Eg a phone has three *classes* of “users”:

- the network operator(s)
  - assets: cellular network
- content providers
  - media content
- the owner of the physical device
  - assets: private data, access keys



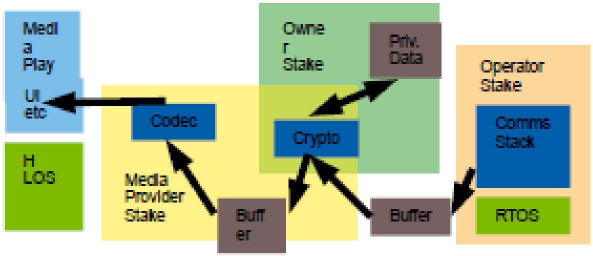


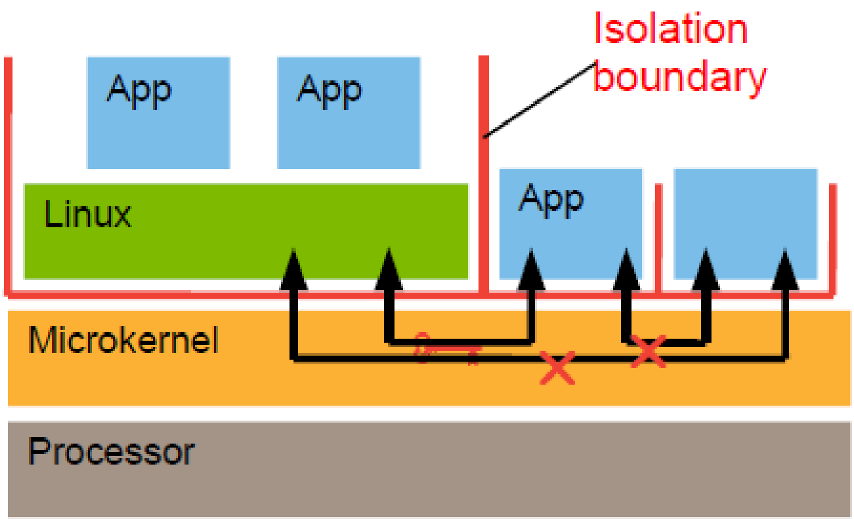
FORTH-ICS Institute of Computer Science

Inter-VM Communication

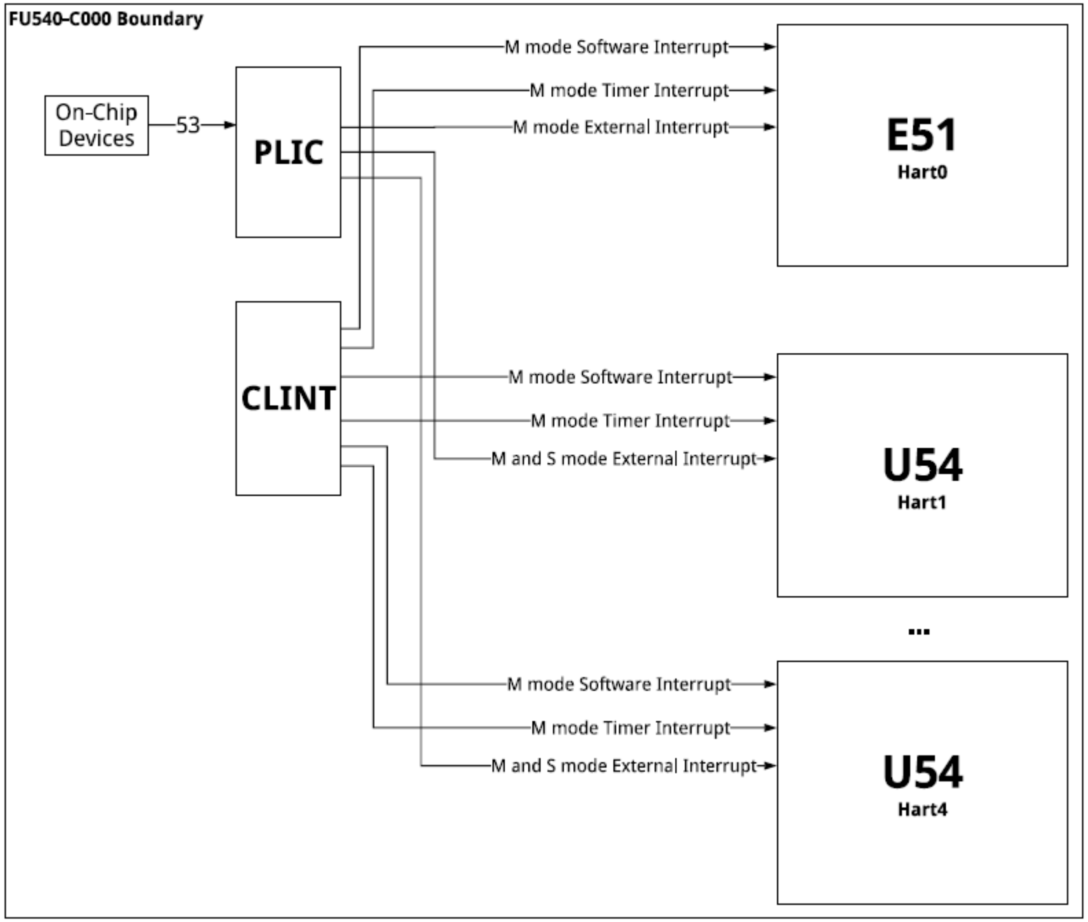
Need to protect integrity and confidentiality against *internal* exploits
Need control over *information flow*
- strict control over who has access to what
- strict control over communication channels

- → Different "users" are mutually distrusting
- → Need strong protection / information-flow control between them
- → Isolation boundaries ≠ VM boundaries
  - some are much smaller than VMs
    - individual buffers, programs
  - some contain VMs
  - some overlap VMs
- → Need to define information flow between isolation domains





RISC-V: interrupts (on SiFive Unleashed platform)



PLIC: platform-level interrupt controller

→ routes all signals through the EI (external interrupt) pin

CLINT: core-local interrupter → implements Software, Timer, and External interrupts.