CS 447/647

Virtualization

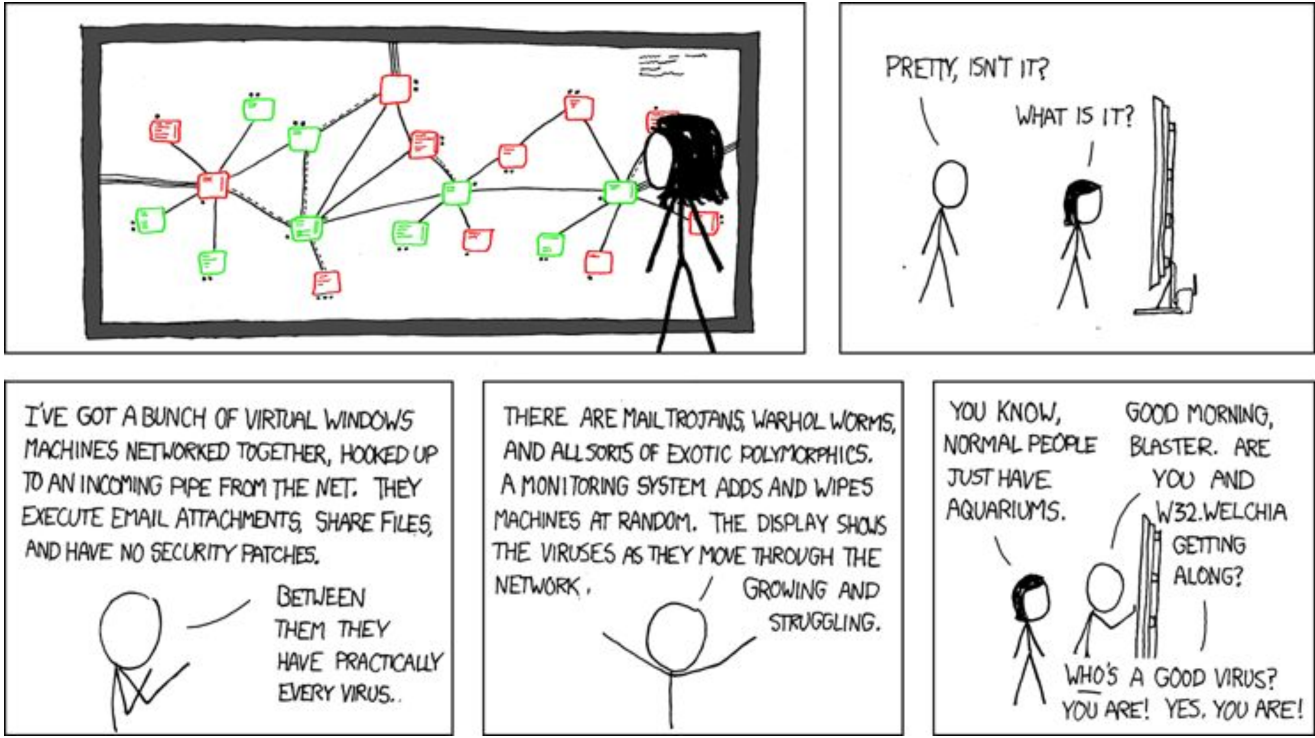
Overview

What is Virtualization?

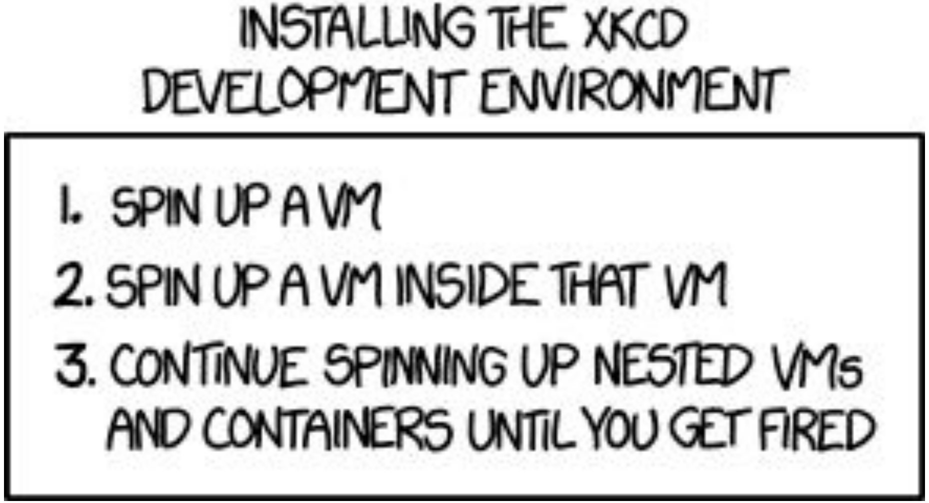
What is Xen and KVM?

Proxmox

Apptainer (Singularity)


PRETTY, ISN'T IT?
WHAT IS IT?
I'VE GOT A BUNCH OF VIRTUAL WINDOWS MACHINES NETWORKED TOGETHER, HOOKED UP TO AN INCOMING PIPE FROM THE NET. THEY EXECUTE EMAIL ATTACHMENTS, SHARE FILES, AND HAVE NO SECURITY PATCHES.
BETWEEN THEM THEY HAVE PRACTICALLY EVERY VIRUS.
THERE ARE MAIL TROJANS, WARHOL WORMS, AND ALL SORTS OF EXOTIC POLYMORPHICS. A MONITORING SYSTEM ADDS AND WIPES MACHINES AT RANDOM. THE DISPLAY SHOWS THE VIRUSES AS THEY MOVE THROUGH THE NETWORK,
GROWING AND STRUGGLING.
YOU KNOW, NORMAL PEOPLE JUST HAVE AQUARIUMS.
GOOD MORNING, BLASTER. ARE YOU AND W32.WELCHIA GETTING ALONG?
WHO'S A GOOD VIRUS? YOU ARE! YES, YOU ARE!

https://xkcd.com/350/



INSTALLING THE XKCD DEVELOPMENT ENVIRONMENT

1. SPIN UP A VM
2. SPIN UP A VM INSIDE THAT VM
3. CONTINUE SPINNING UP NESTED VMs AND CONTAINERS UNTIL YOU GET FIRED

Virtualization

- Multiple operating systems concurrently on the same hardware
- One Host OS to many Guest OS’s
- Big Industry - VMWare Revenue ~$10B
  - Purchased by Dell
- Why?
  - Flexibility
  - Efficiency
  - Backups
  - Resilience
  - Security

Virtual Vernacular

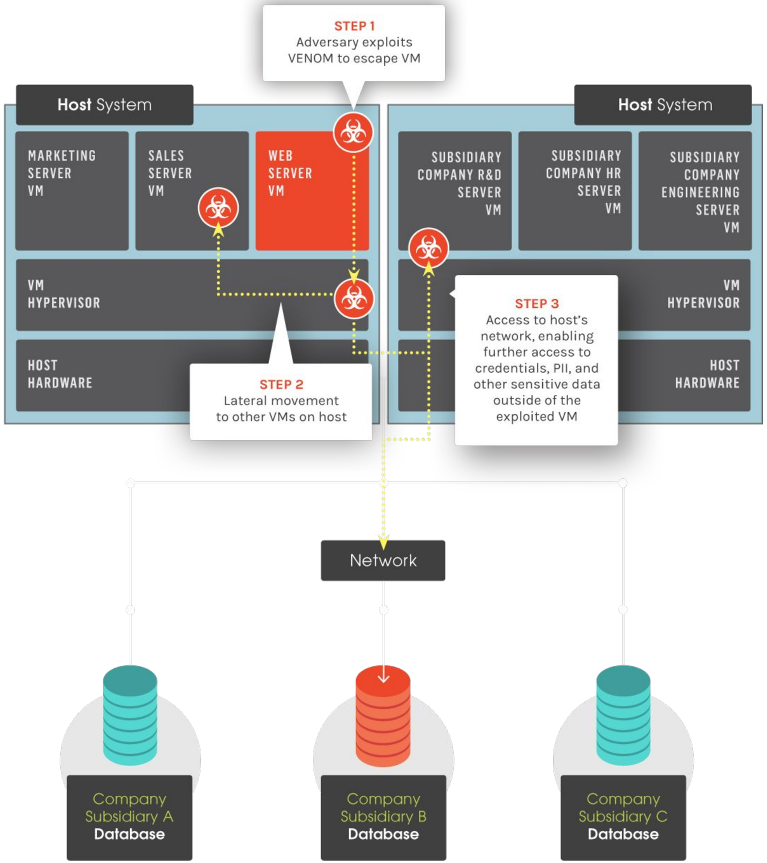
- Virtualized Hardware
    - Hypervisors
- OS-Level Virtualization
    - Docker - commercial
    - lxc - Free
        - lxd extends it
    - Singularity - HPC
    - Windows - Host OS must match
        - Desktop
        - Server
        - Nano Server

Hypervisors

- Virtual Machine Manager/Monitor
- Software layer between the Guest OS and Hardware
- Shares system resources
  - CPU, RAM, Disk and Network
- Isolates* Guest OS’s
  - https://www.crowdstrike.com/blog/venom-vulnerability-details/
- Agnostic
  - Ubuntu
  - Windows
  - FreeBSD
  - Debian

CROWDSTRIKE™

VENOM Vulnerability



STEP 1
Adversary exploits VENOM to escape VM

STEP 2
Lateral movement to other VMs on host

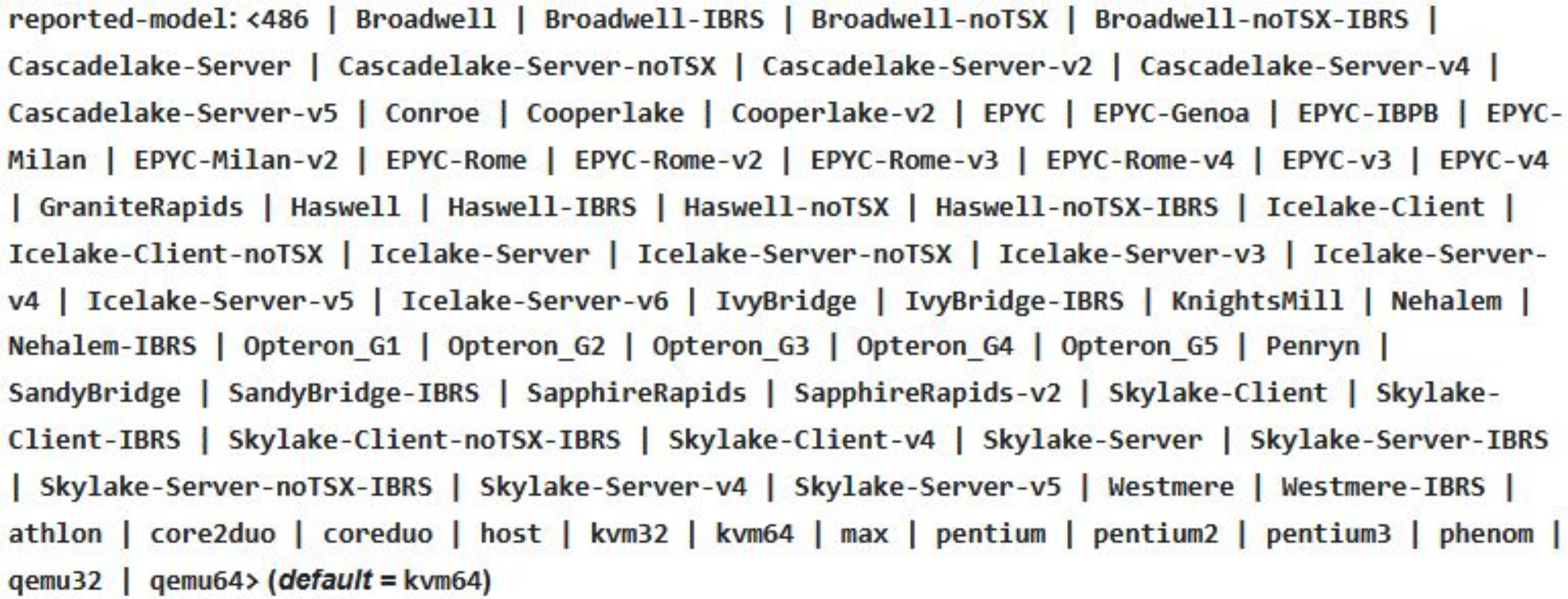
STEP 3
Access to host's network, enabling further access to credentials, PII, and other sensitive data outside of the exploited VM

Host System
MARKETING SERVER VM
SALES SERVER VM
WEB SERVER VM
VM HYPERVISOR
HOST HARDWARE
Host System
SUBSIDIARY COMPANY R&D SERVER VM
SUBSIDIARY COMPANY HR SERVER VM
SUBSIDIARY COMPANY ENGINEERING SERVER VM
VM HYPERVISOR
HOST HARDWARE
Network
Company Subsidiary A Database
Company Subsidiary B Database
Company Subsidiary C Database

Full Virtualization

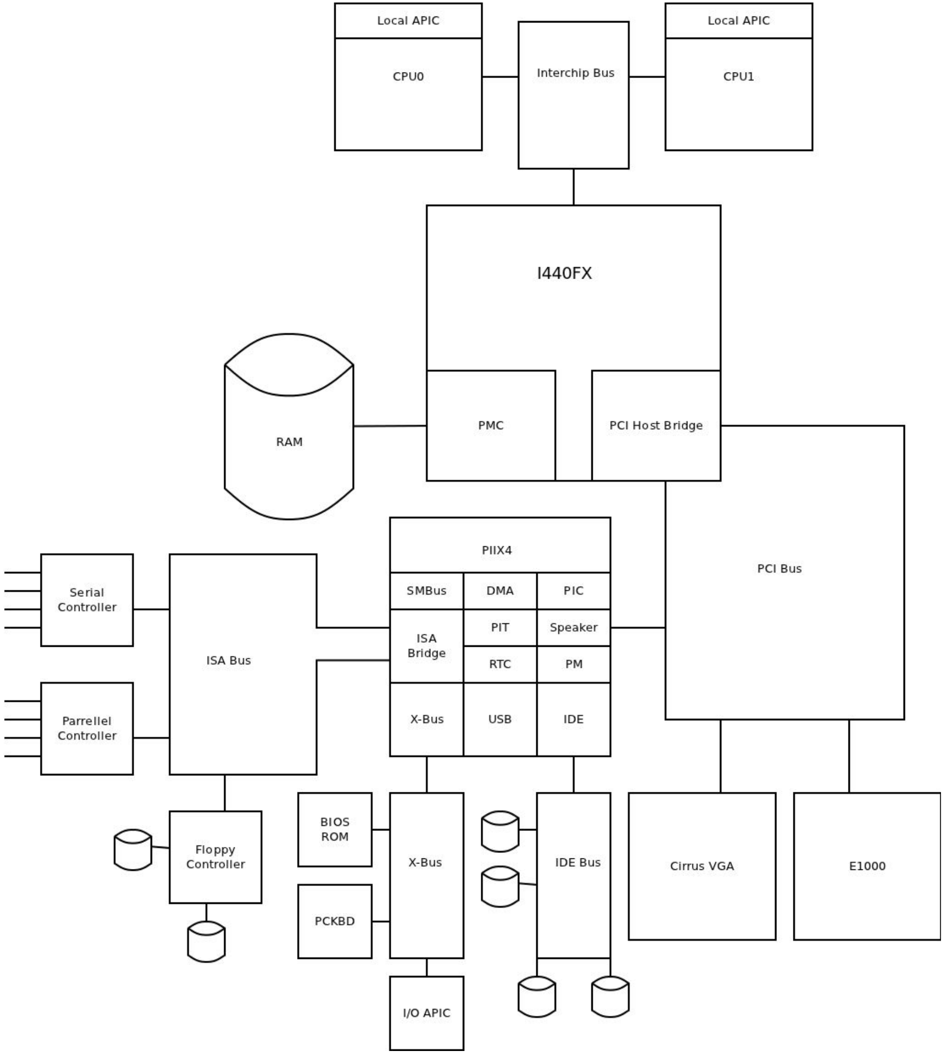
- Emulates the underlying hardware
  - `qemu-system-x86_64 -machine help`
  - `qemu-system-x86_64 -cpu help`
    - Broadwell, Skylake and Pentium I (1993)
- Virtual Hardware components
  - CPU
  - Hard disks
  - Ethernet
  - Interrupts
  - Motherboard Hardware
- Quick EMUlator (QEMU)
  - Best known Linux Full Virtualization Software
  - X86, ARM, MIPS, PowerPC

QEMU CPU Types

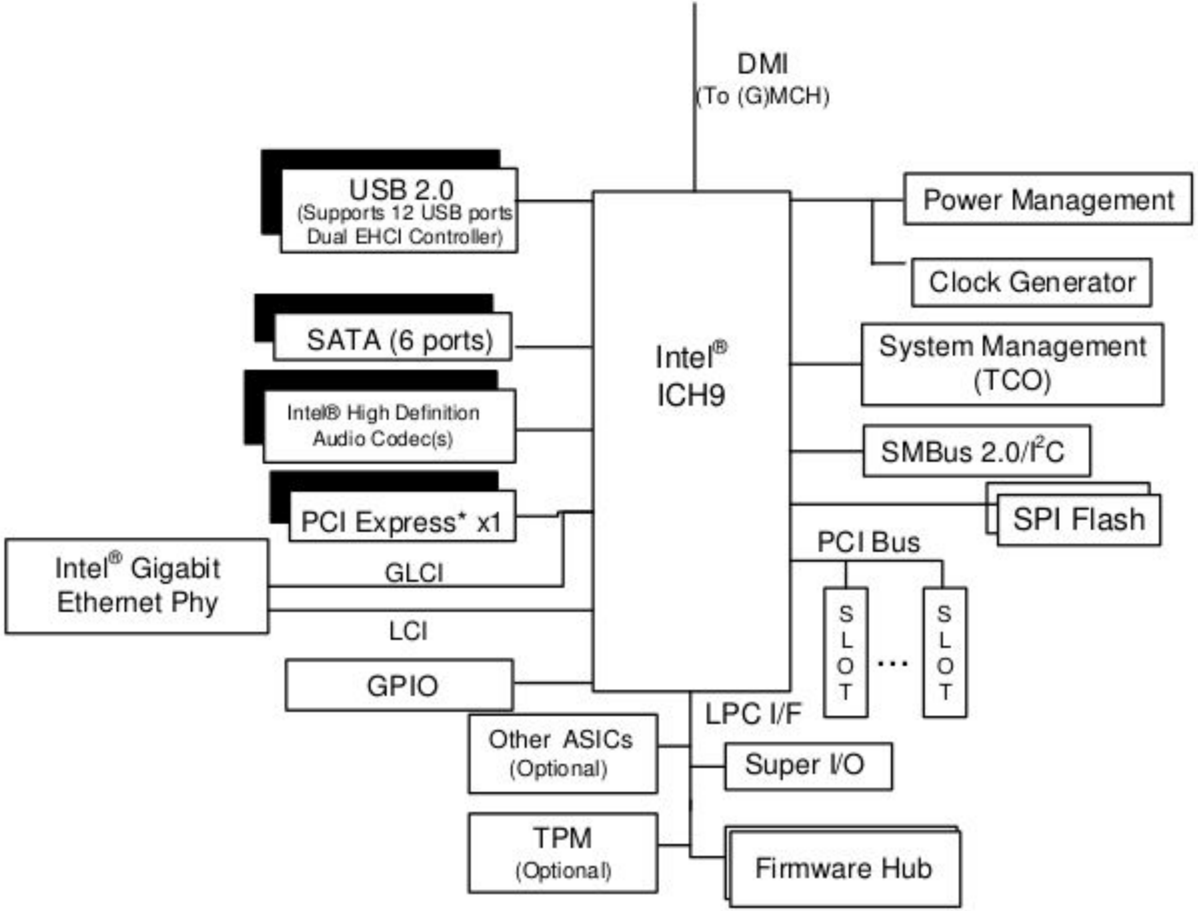


reported-model: <486 | Broadwell | Broadwell-IBRS | Broadwell-noTSX | Broadwell-noTSX-IBRS | Cascadelake-Server | Cascadelake-Server-noTSX | Cascadelake-Server-v2 | Cascadelake-Server-v4 | Cascadelake-Server-v5 | Conroe | Cooperlake | Cooperlake-v2 | EPYC | EPYC-Genoa | EPYC-IBPB | EPYC-Milan | EPYC-Milan-v2 | EPYC-Rome | EPYC-Rome-v2 | EPYC-Rome-v3 | EPYC-Rome-v4 | EPYC-v3 | EPYC-v4 | GraniteRapids | Haswell | Haswell-IBRS | Haswell-noTSX | Haswell-noTSX-IBRS | Icelake-Client | Icelake-Client-noTSX | Icelake-Server | Icelake-Server-noTSX | Icelake-Server-v3 | Icelake-Server-v4 | Icelake-Server-v5 | Icelake-Server-v6 | IvyBridge | IvyBridge-IBRS | KnightsMill | Nehalem | Nehalem-IBRS | Opteron_G1 | Opteron_G2 | Opteron_G3 | Opteron_G4 | Opteron_G5 | Penryn | SandyBridge | SandyBridge-IBRS | SapphireRapids | SapphireRapids-v2 | Skylake-Client | Skylake-Client-IBRS | Skylake-Client-noTSX-IBRS | Skylake-Client-v4 | Skylake-Server | Skylake-Server-IBRS | Skylake-Server-noTSX-IBRS | Skylake-Server-v4 | Skylake-Server-v5 | Westmere | Westmere-IBRS | athlon | core2duo | coreduo | host | kvm32 | kvm64 | max | pentium | pentium2 | pentium3 | phenom | qemu32 | qemu64> (*default* = kvm64)

I440FX


Local APIC
CPU0
Interchip Bus
Local APIC
CPU1
I440FX
RAM
PMC
PCI Host Bridge
PIIX4
SMBus
DMA
PIC
ISA Bridge
PIT
Speaker
RTC
PM
X-Bus
USB
IDE
PCI Bus
Serial Controller
ISA Bus
Parrellel Controller
Floppy Controller
BIOS ROM
PCKBD
X-Bus
I/O APIC
IDE Bus
Cirrus VGA
E1000

ICH9


DMI
(To (G)MCH)
USB 2.0
(Supports 12 USB ports
Dual EHCI Controller)
SATA (6 ports)
Intel® High Definition
Audio Codec(s)
PCI Express* x1
Intel® Gigabit
Ethernet Phy
GLCI
LCI
GPIO
Intel®
ICH9
Power Management
Clock Generator
System Management
(TCO)
SMBus 2.0/I^{2}C
SPI Flash
PCI Bus
SLOT
...
SLOT
LPC I/F
Other ASICs
(Optional)
Super I/O
TPM
(Optional)
Firmware Hub

Paravirtualization (Xen)

- Guest OS detects virtualized state
- Improved performance (1% - 3% overhead)
- Guest OS requires various drivers or kernel modules
- No Windows
  - Needs full hardware virtualization

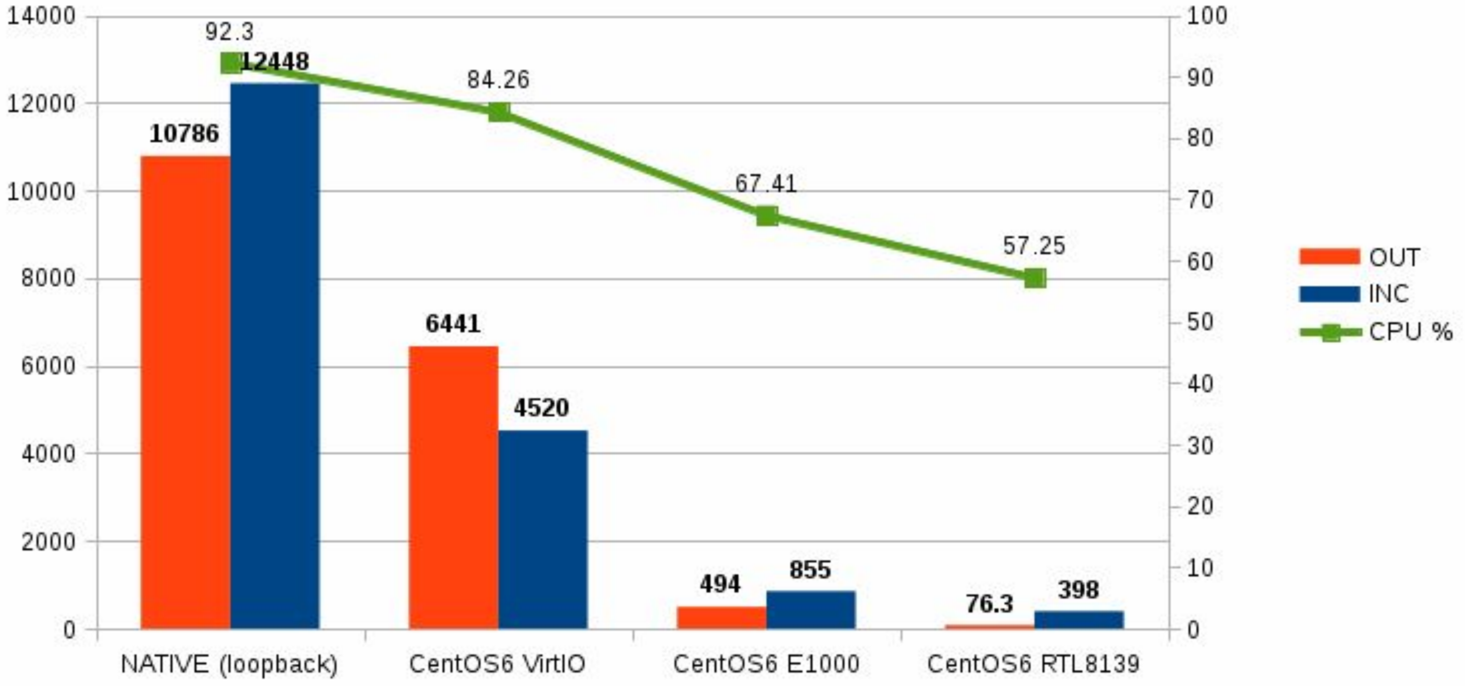
Hardware-Assisted Virtualization

- CPU Feature Based
    - Intel-V
    - AMD-V
- Accelerated Virtualization
    - CPU and memory virtualized by hardware
- Works with full virtualization and paravirtualization
- GPU Virtualization
    - Nvidia vGPU
        - Paid licensing
        - Formerly time-sliced, now SR-IOV
    - AMD MxGPU
        - Free, open-source
        - SR-IOV
        - No drivers/documentation available to the general public

Paravirtualized Drivers

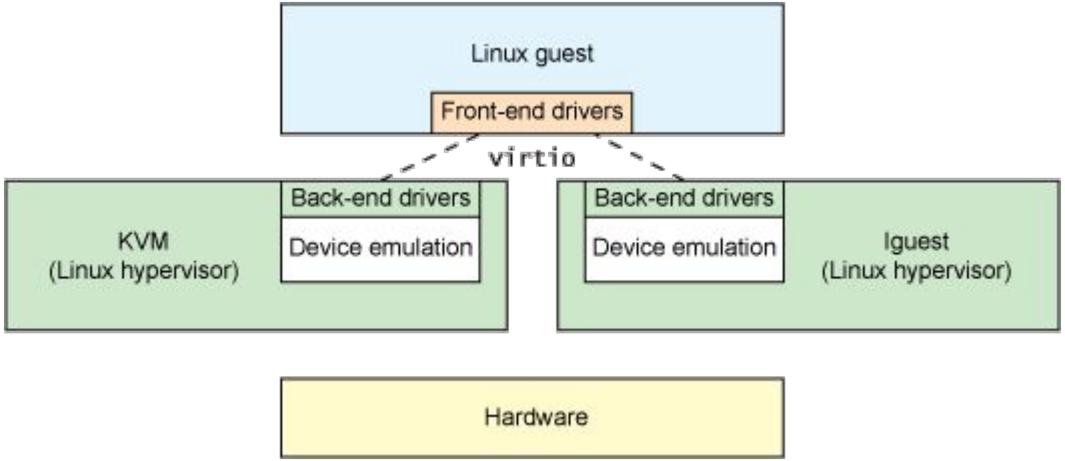
- Enabled by Hardware Assisted Virtualization
- Paravirtualized components
  - Disks
  - Networking
  - Graphics Cards
  - Filesystems - (9p)
- Greatly reduces the amount of full virtualization
- Performance
  - e1000 - 600Mb/s
  - virtio-net - little to no overhead

TCP transfer speed, Mbit/s and CPU load

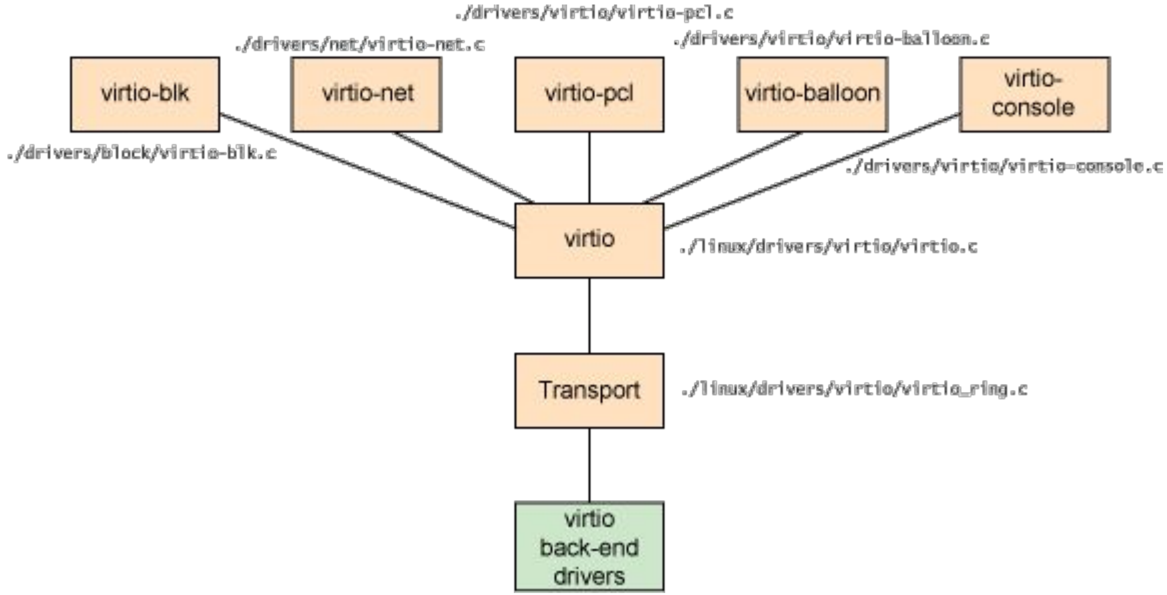


VirtIO

Paravirtualized Drivers



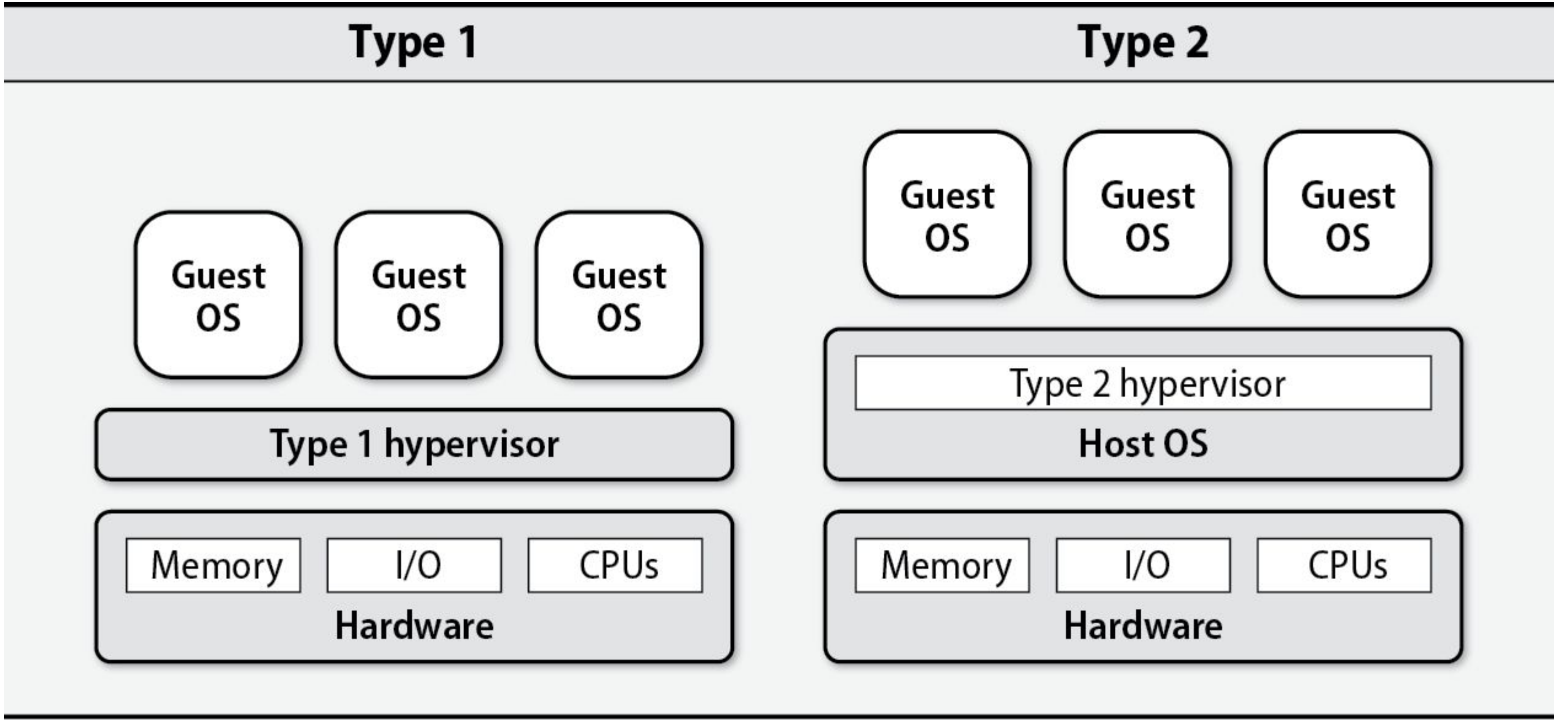
virtio framework



./drivers/virtio/virtio-pcl.c
./drivers/net/virtio-net.c
./drivers/virtio/virtio-balloon.c
virtio-blk
virtio-net
virtio-pcl
virtio-balloon
virtio-console
./drivers/block/virtio-blk.c
./drivers/virtio/virtio-console.c
virtio
./linux/drivers/virtio/virtio.c
Transport
./linux/drivers/virtio/virtio_ring.c
virtio back-end drivers

Hypervisors

- Type 1 - Runs on hardware without supporting OS
  - ESXI (Commercial)
  - XenServer (Commercial)
  - Proxmox (Free)
- Type 2 - User-space applications within an OS
  - QEMU
  - Virtualbox
  - VMWare Player + Workstation


Type 1
Type 2
Guest OS
Guest OS
Guest OS
Type 1 hypervisor
Memory
I/O
CPUs
Hardware
Guest OS
Guest OS
Guest OS
Type 2 hypervisor
Host OS
Memory
I/O
CPUs
Hardware

Live Migration

- Move a VM between Hypervisors
  - Real time
  - Little to no service interruption
- Copies memory
- Copies disk
  - DRBD Dual-Master
- High-availability, disaster recovery, maintenance

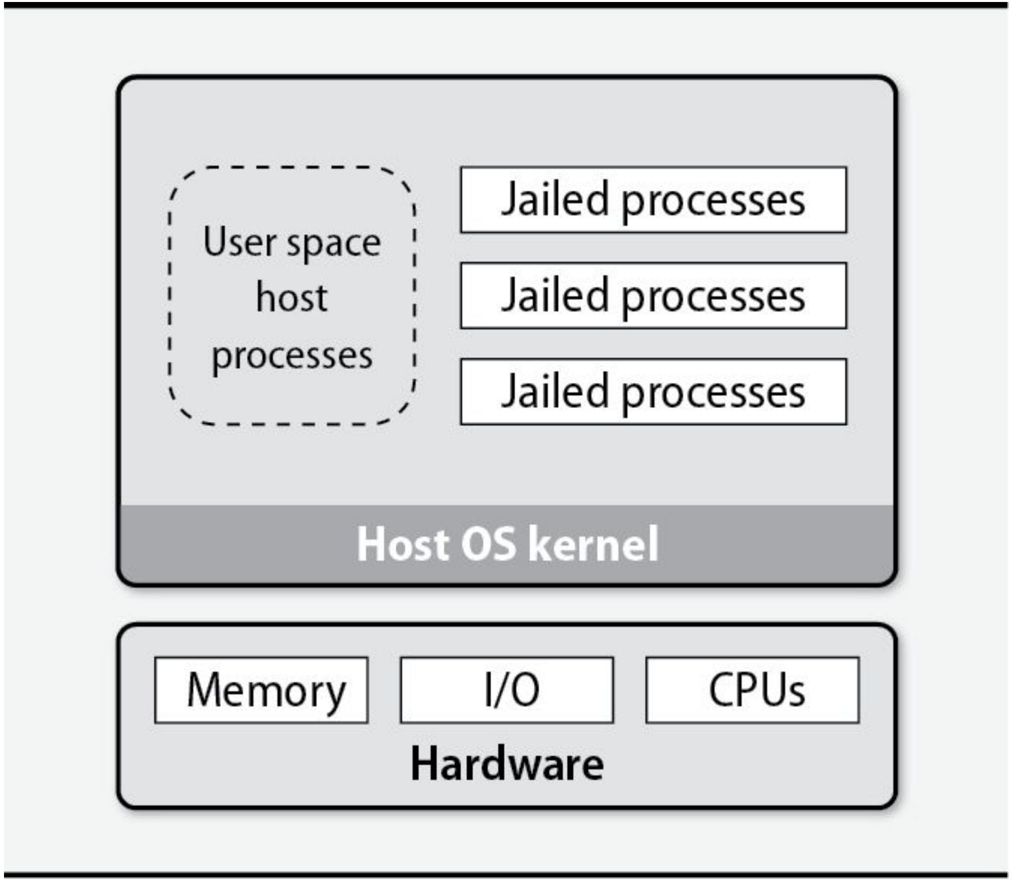
Virtual machine images (disks)

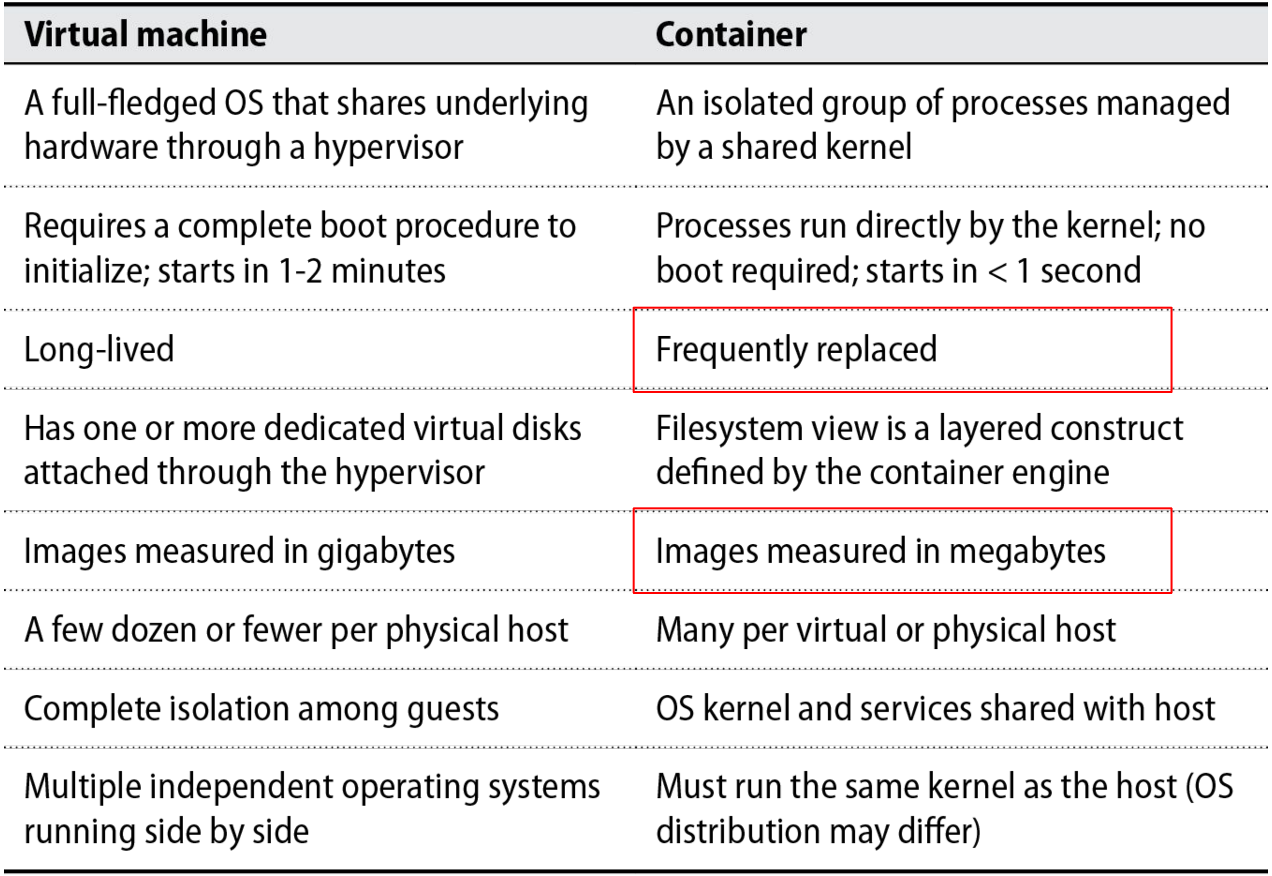
- Images
- Canned Operating System
- File Formats - ova, img, qcow2, raw
- Snapshotting
- Portable
- Metadata

```
qemu-img create -f qcow2 disk1.img 10G
#info, resize
```

Containerization

- OS-level Virtualization
    - Portable
    - Isolated
- Relies on Kernel Features for isolation
    - cgroups
    - tap networking
    - process namespaces
- Cannot access files or resources outside of container.*
- Benefits
    - Performance - low latency IO, near native


User space host processes
Jailed processes
Jailed processes
Jailed processes
Host OS kernel
Memory
I/O
CPUs
Hardware



Virtual machine	Container
A full-fledged OS that shares underlying hardware through a hypervisor	An isolated group of processes managed by a shared kernel
Requires a complete boot procedure to initialize; starts in 1-2 minutes	Processes run directly by the kernel; no boot required; starts in < 1 second
Long-lived	Frequently replaced
Has one or more dedicated virtual disks attached through the hypervisor	Filesystem view is a layered construct defined by the container engine
Images measured in gigabytes	Images measured in megabytes
A few dozen or fewer per physical host	Many per virtual or physical host
Complete isolation among guests	OS kernel and services shared with host
Multiple independent operating systems running side by side	Must run the same kernel as the host (OS distribution may differ)

Virtualization with Linux

- Xen
    - Paravirtual Hypervisor
        - Performance overhead of 0.1%-3.5%
    - Domains aka Virtual Machine
        - dom0 = Host
- Kernel-based Virtual Machine
    - Full Virtualization
    - Paired with QEMU

Proxmox

- Open-source hypervisor based on Debian
- Server cluster management
  - Centralized backups
  - Live migration
  - High availability
- Underlying technologies
  - Storage - LVM, ZFS, Ceph, Directories
  - VMs - QEMU/KVM
  - Containers - LXC
  - Networking - firewalls, SDNs (software-defined networking)
  - User management - PAM, PVE database, LDAP, Microsoft AD, OIDC
  - Web server
- Proxmox Backup Server
- Proxmox Datacenter Manager

Ganeti

- Originally from Google Switzerland
  - Now maintained by GRNET (Greek Research Network) on Debian
- Virtual machine cluster management tool
  - Disk creation management
  - Operating system installation
  - Startup, shutdown, and failover
- Uses:
  - debootstrap
  - chroot
  - lvm2
  - drbd

Singularity / Apptainer

- Container platform
- Portable
  - Filesystem stored in an .simg file
- Built for HPC
  - Reproducible Science
- Legacy code or systems
- https://sylabs.io/