Linux Virtualization

Why Virtualization

1- Consolidation

2- A typical (full) data center

3- Hardware isolation

4- Legacy operating systems

5- Testing

6- Maintenance

7- Power Savings

8- Security and performance isolation

CPU Modes

1- Kernel mode :

Unrestricted
M*aster mode*
Supervisor mode
Privileged mode
Supervisor state

2- user mode :

Restricted
slave mode
problem state

Virtualization technologies

1- Paravirtualization: guest kernel is changed
Xen, Lguest and UML(User Mode Linux)

2- Hardware assisted: Popek and Goldberg requirements
Xen and KVM

3- Coopvirt : cooperative virtualization
still in a research and prototyping phase

4- Containers : operating-system level with one kernel
Solaris Zones, Linux-VServer, OpenVZ

5- Binary rewriting / JIT
Qemu and Vmware

KVM : Kernel-based Virtual Machine

1- X86 CPU requirements :

    Intel VT
    AMD-V

2- Software requirements :

    2-1- At least kernel 2.6.30 is required for MSI-X support (Message Signaled Interrupts)
    2-2- qemu-kvm: http://sourceforge.net/projects/kvm/
    2-3- tunctl: http://tunctl.sourceforge.net/
    2-4- socat: http://www.dest-unreach.org/socat/

checking CPU for Intel VT or AMD-V Support

```
# egrep '^flags.*(vmx|svm)' /proc/cpuinfo

flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr
pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm
pbe nx lm constant_tsc arch_perfmon pebs bts aperfmperf pni dtes64
monitor ds_cpl vmx est tm2 ssse3 cx16 xtpr pdcm xsave lahf_lm
tpr_shadow vnmi flexpriority
flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr
pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm
pbe nx lm constant_tsc arch_perfmon pebs bts aperfmperf pni dtes64
monitor ds_cpl vmx est tm2 ssse3 cx16 xtpr pdcm xsave lahf_lm
tpr_shadow vnmi flexpriority
```

Kernel Config : (make menuconfig)

```
Virtualization  --->
  <M>   Kernel-based Virtual Machine (KVM) support
    <M>     KVM for Intel processors support
    <M>     KVM for AMD processors support
  <M>   Linux hypervisor example code
  <M>   PCI driver for virtio devices
  <M>   Virtio balloon driver
Device Drivers  --->
  [*] Block devices  --->
    <M>   Virtio block driver
  [*] Network device support  --->
    <M>   Universal TUN/TAP device driver support
    <M>   Virtio network driver
  Character devices  --->
    <M> Virtio console
    <M>   VirtIO Random Number Generator support
-*- Networking support  --->
  <M>   Plan 9 Resource Sharing Support (9P2000) --->
    <M>   9P Virtio Transport
```

Kernel version check and loading modules

```
# uname -r
2.6.32.5-smp

# modprobe tun
# modprobe kvm
# modprobe kvm_intel
# modprobe virtio_balloon
# modprobe virtio_ring
# modprobe virtio_pci
# modprobe virtio
# modprobe virtio_blk
# modprobe virtio-rng
# modprobe virtio_console
# modprobe virtio_net
# modprobe 9pnet_virtio
```

qemu-kvm installation (host)

```
/bin/cp qemu-kvm-0.12.5.tar.gz /usr/src/
cd /usr/src/
tar xf qemu-kvm-0.12.5.tar.gz
cd qemu-kvm-0.12.5
./configure --prefix=/usr/local/qemu
make && make install
```

tunctl installation (host)

```
/bin/cp tunctl-1.5.tar.gz /usr/src
cd /usr/src
tar xf tunctl-1.5.tar.gz
cd tunctl-1.5
make clean
make
/bin/cp tunctl /usr/local/sbin
```

socat installation (host)

```
/bin/cp socat-1.7.1.3.tar.bz2 /usr/src/
cd /usr/src/
tar xf socat-1.7.1.3.tar.bz2
cd socat-1.7.1.3
./configure --prefix=/usr/local/socat
make && make install
```

Partitioning schema (host)

```
# fdisk -l /dev/sda

Disk /dev/sda: 160.0 GB, 160041885696 bytes
255 heads, 63 sectors/track, 19457 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes
Disk identifier: 0x000c564e

   Device Boot   Start         End      Blocks   Id  System
/dev/sda1            1         132     1060258+  82  Linux swap[swap host]
/dev/sda2          133       13187   104864287+   5  Extended
/dev/sda5          133        2744    20980858+  83  Linux     [guest1]
/dev/sda6         2745        6661    31463271   83  Linux
/dev/sda7         6662        9094    19543041   83  Linux     [host]
/dev/sda8         9095       11007    15366141   83  Linux     [guest2]
/dev/sda9        11008       12538    12297726   83  Linux     [guest3]
/dev/sda10       12539       12670     1060258+  82  Linux swap[swap-guest1]
/dev/sda11       12671       12802     1060258+  82  Linux swap[swap-guest2]
/dev/sda12       12803       12934     1060258+  82  Linux swap[swap-guest3]
```

DO NOT forget to use separate swap partition for each operating system.

Installation script (guest1)

DO NOT install bootloader for guest machine, it will overwrite host machine's MBR.

```
#!/bin/bash
# Installation script for guest1
# Kernel options for ZenWalk installer : ata-vga noapic
# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Installer image
ISO="/storage/zenwalk-6.4.iso"

# Network interface
tunctl -t tap1
ifconfig tap1 192.168.11.1 netmask 255.255.255.0

# Start virtual machine and boot from CDROM
/usr/local/qemu/bin/qemu-system-x86_64 \
  -cpu host -m 512                     \
  -drive file=/dev/sda,cache=none      \
  -net nic,model=virtio                \
  -net tap,ifname=tap1,script=no       \
  -boot d -cdrom ${ISO}
```

qemu-kvm options

-cpu host	Guest CPU like host
-m 512	Guest RAM size in MB
-drive file=/dev/sda,cache=none	Use raw partition
-net nic,model=virtio	Guest network interface
-net tap,ifname=tap1,script=no,downscript=no	Host network interface
-boot d	Boot from first CDROM
-cdrom ${ISO}	CDROM image
-kernel	Use another kernel
-append	Kernel command line
-vnc	Redirect VGA over VNC
-daemonize	Detach from standard IO
-localtime	use host localtime
-nographic	Serial redirect to console
-monitor	Monitor redirect to host device

Copy guest's kernel to host

```
mkdir -p /mnt/guest1
mount /dev/sda5 /mnt/guest1
/bin/cp /mnt/guest1/boot/vmlinuz-2.6.33.4 /boot/guest1-vmlinuz-2.6.33.4
umount /mnt/guest1
```

Foreground startup script (guest1)

```
#!/bin/bash
# guest1 foreground startup script

# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Network interface
tunctl -t tap1
ifconfig tap1 192.168.11.1 netmask 255.255.255.0

# Start virtual machine
/usr/local/qemu/bin/qemu-system-x86_64 \
  -cpu host -m 512                     \
  -kernel /boot/guest1-vmlinuz-2.6.33.4 \
  -append "noapic root=/dev/sda5"      \
  -drive file=/dev/sda,cache=none      \
  -net nic,model=virtio                \
  -net tap,ifname=tap1,script=no
```

Background startup script (guest1)

```
#!/bin/bash
# guest1 background startup script

# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Network interface
tunctl -t tap1
ifconfig tap1 192.168.11.1 netmask 255.255.255.0

# Start virtual machine
/usr/local/qemu/bin/qemu-system-x86_64 \
  -cpu host -m 512                     \
  -kernel /boot/guest1-vmlinuz-kvm     \
  -append "noapic root=/dev/sda5"      \
  -drive file=/dev/sda,cache=none      \
  -net nic,model=virtio                \
  -net tap,ifname=tap1,script=no       \
  -vnc *:0 -daemonize -localtime       \
  -monitor unix:/root/socket/guest1.sock,server,nowait
```

VNC connection to guest1

```
vncviewer 127.0.0.1:5900
```

Installation script (guest2)

DO NOT install bootloader for guest machine, it will overwrite host machine's MBR.

```
#!/bin/bash
# Installation script for guest2
# Kernel options for Slackware13.1 installer : hugesmp.s noapic
# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Installer image
ISO="/storage/slackware-13.1-install-dvd.iso"

# Network interface
tunctl -t tap2
ifconfig tap2 192.168.12.1 netmask 255.255.255.0

# Start virtual machine and boot from CDROM
/usr/local/qemu/bin/qemu-system-x86_64 \
  -cpu host -m 512                     \
  -drive file=/dev/sda,cache=none      \
  -net nic,model=virtio                \
  -net tap,ifname=tap2,script=no       \
  -boot d -cdrom ${ISO}
```

Installation script (guest3)

DO NOT install bootloader for guest machine, it will overwrite host machine's MBR.

```
#!/bin/bash
# Installation script for guest3
# Kernel options for Slackware13.1 installer : hugesmp.s noapic
# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Installer image
ISO="/storage/slackware-13.1-install-dvd.iso"

# Network interface
tunctl -t tap3
ifconfig tap3 192.168.13.1 netmask 255.255.255.0

# Start virtual machine and boot from CDROM
/usr/local/qemu/bin/qemu-system-x86_64 \
  -cpu host -m 512                     \
  -drive file=/dev/sda,cache=none      \
  -net nic,model=virtio                \
  -net tap,ifname=tap3,script=no       \
  -boot d -cdrom ${ISO}
```

Copy guest's kernel to host

```
mkdir -p /mnt/guest2
mount /dev/sda8 /mnt/guest2
/bin/cp /mnt/guest2/boot/vmlinuz-huge-smp-2.6.33.4-smp /boot/guest2-vmlinuz-2.6.33.4
umount /mnt/guest2

mkdir -p /mnt/guest3
mount /dev/sda9 /mnt/guest3
/bin/cp /mnt/guest3/boot/vmlinuz-huge-smp-2.6.33.4-smp /boot/guest3-vmlinuz-2.6.33.4
umount /mnt/guest3
```

Foreground startup script (guest2)

```
#!/bin/bash
# guest2 foreground startup script

# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Network interface
tunctl -t tap2
ifconfig tap2 192.168.12.1 netmask 255.255.255.0

# Start virtual machine
/usr/local/qemu/bin/qemu-system-x86_64  \
  -cpu host -m 512                      \
  -kernel /boot/guest2-vmlinuz-2.6.33.4 \
  -append "noapic root=/dev/sda8"       \
  -drive file=/dev/sda,cache=none       \
  -net nic,model=virtio                 \
  -net tap,ifname=tap2,script=no
```

Foreground startup script (guest3)

```
#!/bin/bash
# guest3 foreground startup script

# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Network interface
tunctl -t tap3
ifconfig tap3 192.168.13.1 netmask 255.255.255.0

# Start virtual machine
/usr/local/qemu/bin/qemu-system-x86_64  \
  -cpu host -m 512                      \
  -kernel /boot/guest3-vmlinuz-2.6.33.4 \
  -append "noapic root=/dev/sda9"       \
  -drive file=/dev/sda,cache=none       \
  -net nic,model=virtio                 \
  -net tap,ifname=tap3,script=no
```

Using Serial port for login (guest)

```
echo “s0:2345:respawn:/sbin/agetty -L 115200 ttyS0 vt100 “ >> /etc/inittab
echo “ttyS0” >>  /etc/securetty
init q
```

Background startup script (guest2)

```
#!/bin/bash
# guest2 background startup script
# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Network interface
tunctl -t tap2
ifconfig tap2 192.168.12.1 netmask 255.255.255.0

# Start virtual machine inside GNU screen
/usr/bin/env SCREENDIR="/root/.screen"                   \
  /usr/bin/screen -h 4000 -dmS guest2                    \
  /usr/local/qemu/bin/qemu-system-x86_64                 \
    -cpu host -m 512                                     \
    -kernel /boot/guest2-vmlinuz-2.6.33.4                \
    -append "noapic root=/dev/sda8 console=ttyS0,115200" \
    -drive file=/dev/sda,cache=none                      \
    -net nic,model=virtio                                \
    -net tap,ifname=tap2,script=no                       \
    -localtime -nographic                                \
    -monitor unix:/root/socket/guest2.sock,server,nowait
```

Background startup script (guest3)

```
#!/bin/bash
# guest3 background startup script
# Load kernel modules
MODULES="
  tun kvm kvm_intel virtio_balloon virtio_ring virtio_pci virtio
  virtio_blk virtio-rng virtio_console virtio_net 9pnet_virtio
"
for MOD in $MODULES; do modprobe $MOD; done

# Network interface
tunctl -t tap3
ifconfig tap3 192.168.13.1 netmask 255.255.255.0

# Start virtual machine inside GNU screen
/usr/bin/env  SCREENDIR="/root/.screen"                  \
  /usr/bin/screen -h 4000 -dmS guest3                    \
  /usr/local/qemu/bin/qemu-system-x86_64                 \
    -cpu host -m 512                                     \
    -kernel /boot/guest3-vmlinuz-2.6.33.4                \
    -append "noapic root=/dev/sda9 console=ttyS0,115200" \
    -drive file=/dev/sda,cache=none                      \
    -net nic,model=virtio                                \
    -net tap,ifname=tap3,script=no                       \
    -localtime -nographic                                \
    -monitor unix:/root/socket/guest3.sock,server,nowait
```

Guest Shutdown script (host)

```
#!/bin/bash

# Send shutdown request to all virtual machines
for SOCKET in /root/socket/*.sock
  do echo 'system_powerdown' |\
    /usr/local/socat/bin/socat - unix-connect:${SOCKET} > /dev/null
done

# Wait for shutdown complete
echo -n "Wating for Virtual Machines shutdown"
for S in $(seq 1 300); do
  if ! pgrep '^qemu-system-x86$' > /dev/null; then break; fi
  echo -n .${S}
  sleep 1
done
echo

# Delete old sockets
rm -rf /root/socket/*.sock
```

More resources

Linux Virtualization Wiki: http://virt.kernelnewbies.org
CPU Modes: http://en.wikipedia.org/wiki/CPU_modes
MSI HowTo: http://lwn.net/Articles/44139/
Kernel-based Virtual Machine: http://www.linux-kvm.org/
Xen hypervisor: http://www.xen.org/
Lguest: http://lguest.ozlabs.org/
User-mode Linux: http://user-mode-linux.sourceforge.net/
Linux-VServer: http://linux-vserver.org/
OpenVZ: http://wiki.openvz.org/
QEMU: http://wiki.qemu.org/

Pejman Moghadam
pmoghadam@yahoo.com
Zanjan - 2010/08/05