

Libvirt. Why should I care?



Michal Privozník

mprivozn@redhat.com

KVM Forum Düsseldorf 2014

QEMU/KVM

- Runs quickly
- Scalability
- Portable

QEMU/KVM

- No host management
- Multiple interfaces
- No language bindings
- No inter-VM perspective

Libvirt



<http://libvirt.org>

- C library, bindings
- Stable API
- Multiple hypervisors

Domain startup

- Query QEMU capabilities
 - Get list of supported devices, attributes, events, etc.
- Prepare host devices
 - PCI/USB/SCSI passthrough
- Reserve VNC/SPICE ports

Domain startup

- Consult `numad` for placement
 - Mystery since vCPU/memory can be hotplugged
- Build command line
 - Most of the host resources allocated
- `fork()`
 - Drop all unneeded capabilities

Domain startup (child)

- Report PID to the daemon
 - Needed later in the process
- Lock domain disks
 - `virtlockd`
- Honor NUMA settings
 - Place onto configured NUMA nodes
- Set process security labels, cwd, etc.
 - Drop the rest of unneeded admin capabilities

Domain startup

- Create cgroup hierarchy
 - Selectively allow devices, set blkio, etc.
- Set security labels on domain resources
 - Disks, host devices, chardevs, kernel, etc.
- Complete handshake to child
 - Child `execve()`

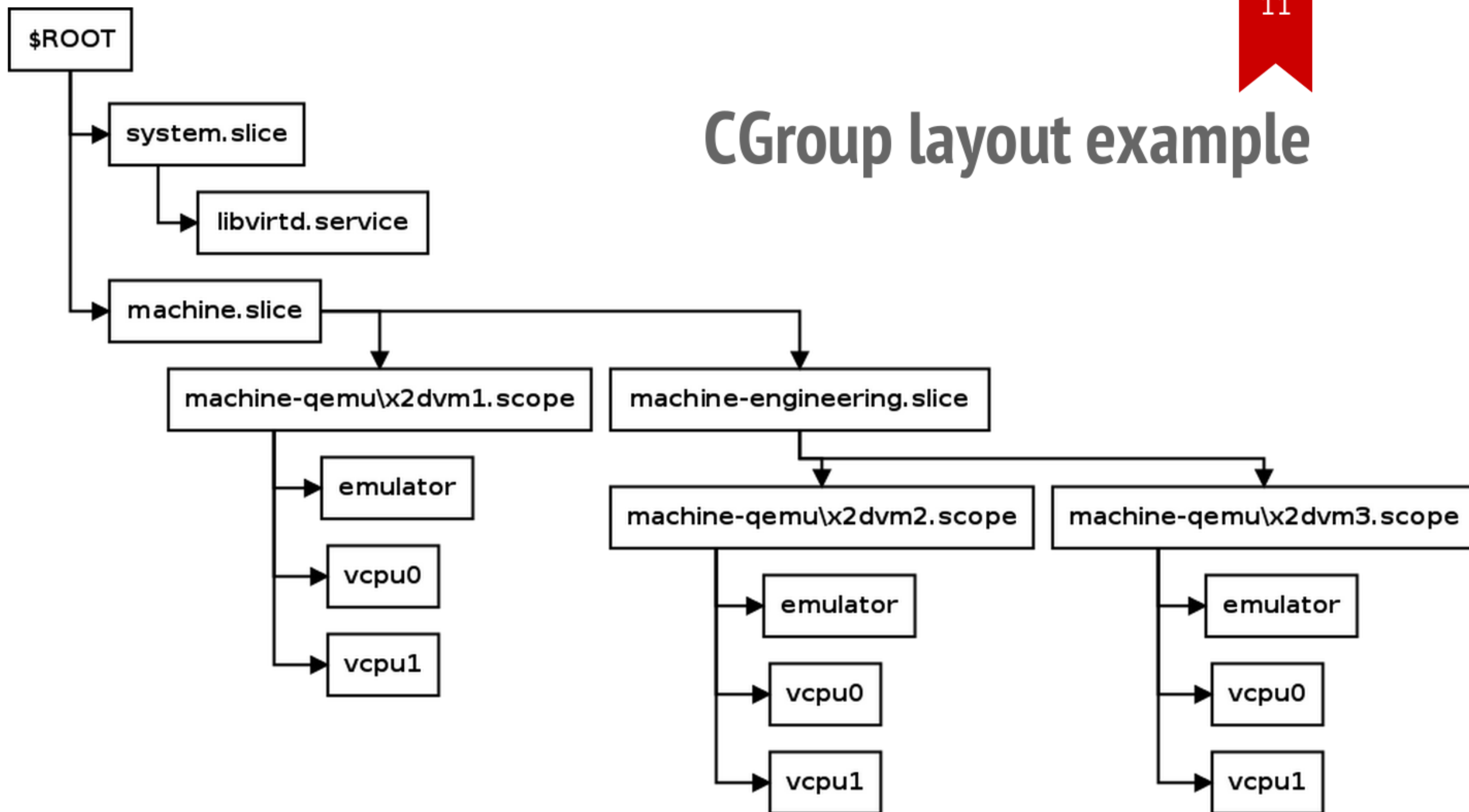
Domain startup

- Connect to the monitor
 - Finish setting cgroup, set runtime values
- Start domain vCPUs
- Run post-exec hook script

CGroup layout

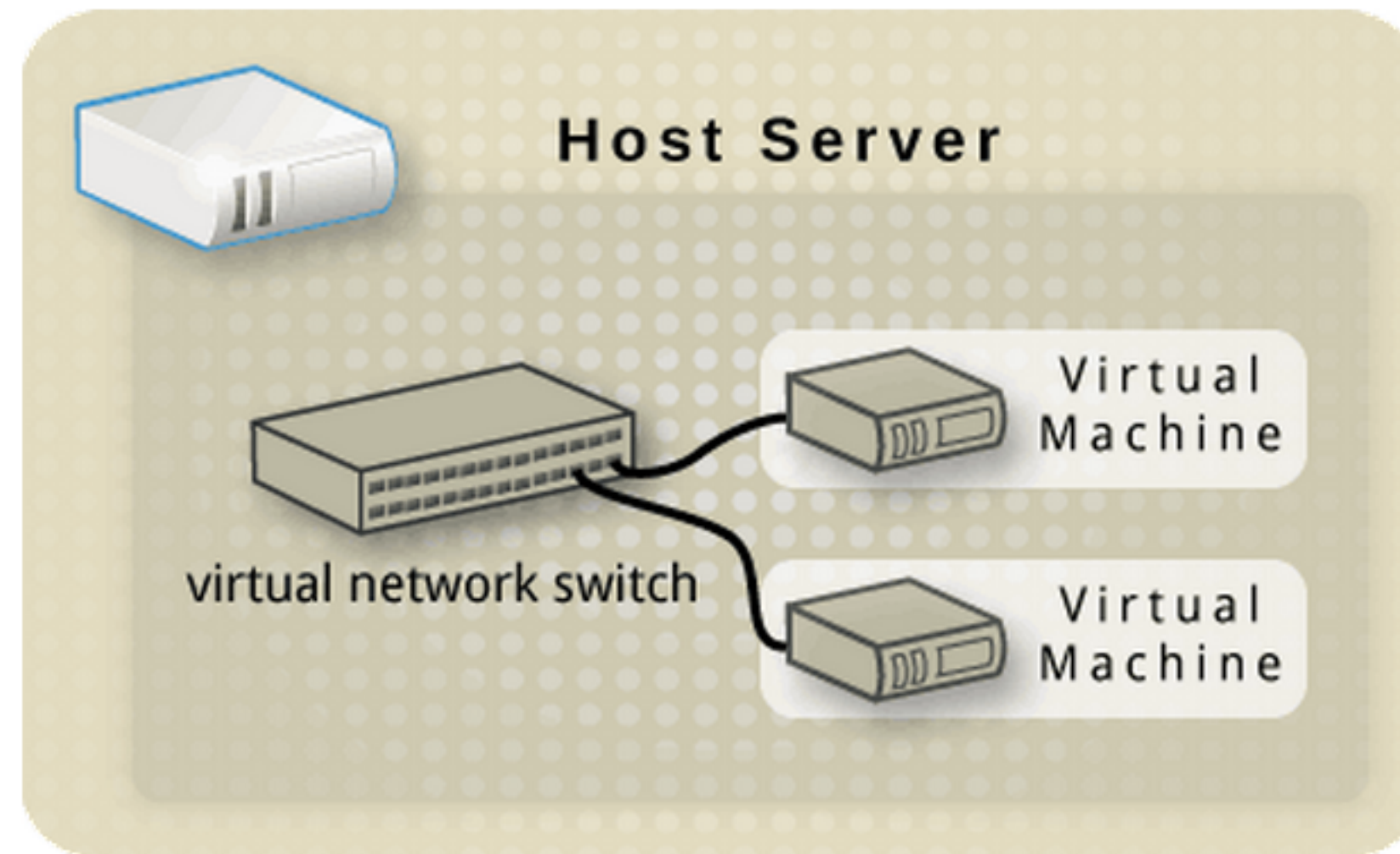
- Keep directory as flat as possible
 - `$ROOT/system/libvirtd.service/libvirt/qemu/dom1`
 - `$ROOT/machine/dom1.libvirt-qemu`
- SystemD integration
 - Idea is to have one manager
 - A domain (scope), a group (slice)
- The path is configurable in XML

CGroup layout example



Virtual Networks

- Create a virtual bridge
- Operating modes: NAT, Isolated, Routed
- Run DHCP/DNS server



Network Filters

Enforce network traffic filtering on vNIC basis:

```
01. <devices>
02.   <interface type='bridge'>
03.     <mac address='00:16:3e:5d:c7:9e' />
04.     <filterref filter='clean-traffic'>
05.       <parameter name='IP' value='10.0.0.1' />
06.     </filterref>
07.   </interface>
08. </devices>
```

Network Filters

Filters written in XML:

```
01. <filter name='no-ip-spoofing' chain='ipv4-ip' priority='-710'>
02.   <uuid>2b308492-52d3-4bda-8f0c-1dedbcf58e04
03.   <rule action='return' direction='out' priority='100'>
04.     <ip srcipaddr='0.0.0.0' protocol='udp' />
05.   </rule>
06.   <rule action='return' direction='out' priority='500'>
07.     <ip srcipaddr='$IP' />
08.   </rule>
09.   <rule action='drop' direction='out' priority='1000' />
10. </filter>
```


Network Filters

Automatic IP address detection:

- DHCP snooping
 - Multiple IPs per interface
 - Combine with filtering untrusted DHCP server
- IP packed snooping
 - Single IP per interface

Secrets

- Used to store passphrases for QCOW2/Ceph/iSCSI disks
- Libvirt provides stored passphrase to auth mechanism
- Passphrases can be stored on disk, or in memory, and set to be private

Storage management

- Pools
 - Local: directory, LVM VG, disk
 - Shared: NFS, iSCSI, Gluster, RBD
- Volumes
 - Local: file, LVM LV, partition
 - Shared: file, LUNs

sVirt

- DAC is not enough.
- Malicious guest is threat to others running under the same user.
- Aim is MAC policy enforced by the host kernel.
- Libvirt generates dynamic SELinux labels.
- Set label on disk images, sockets, devices, etc.

Snapshots

	disk	memory	checkpoint
internal	No	N/A	Yes
external	Yes	Yes	Yes

- Create, revert, merge (pull/commit), delete
- Libvirt keeps metadata

Questions?