Linux Kernel in Docker Containers

Leon Romanovsky

August, 2019

Short Bio

	Leon Romanovsky	Jason Gunthorpe
RDMA kernel maintainer	Internal (Mellanox)	External (upstream)
Unified RDMA user space library maintainer	External (upstream, co-founder)	External (upstream, founder)
Linux experience	> 20 years	> 20 years

Short Bio

	Leon Romanovsky	Jason Gunthorpe
RDMA kernel maintainer	Internal (Mellanox)	External (upstream)
Unified RDMA user space library maintainer	External (upstream, co-founder)	External (upstream, founder)
Linux experience	> 20 years	> 20 years
	Patch statistics in 2018	
Authored kernel patches	166	200
Authored rdma-core patches	60	220
Authored iproute2 patches	23	---
Handled (reviewed and successfully submitted) kernel patches	713	1164

Perfect Solution

- **Hide** operating system complexity from kernel and QEMU developers
- Give **latest** development and run environments
- **Seamless** integration with emulated and real hardware
- Run **real** operating system and **real** kernel
- **Fast** write-build-test loop
- **Source** and **patch** oriented flow
- Built-in **continuous integration**
- Work **anywhere**
- Provide **out-of-the box** experience
- **Easy** customization
- Ready for **cloud** orchestration software

Perfect Solution

- **Hide** operating system complexity from kernel and QEMU developers
- Give **latest** development and run environments
- **Seamless** integration with emulated and real hardware
- Run **real** operating system and **real** kernel
- **Fast** write-build-test loop
- **Source** and **patch** oriented flow
- Built-in **continuous integration**
- Work **anywhere**
- Provide **out-of-the box** experience
- **Easy** customization
- Ready for **cloud** orchestration software

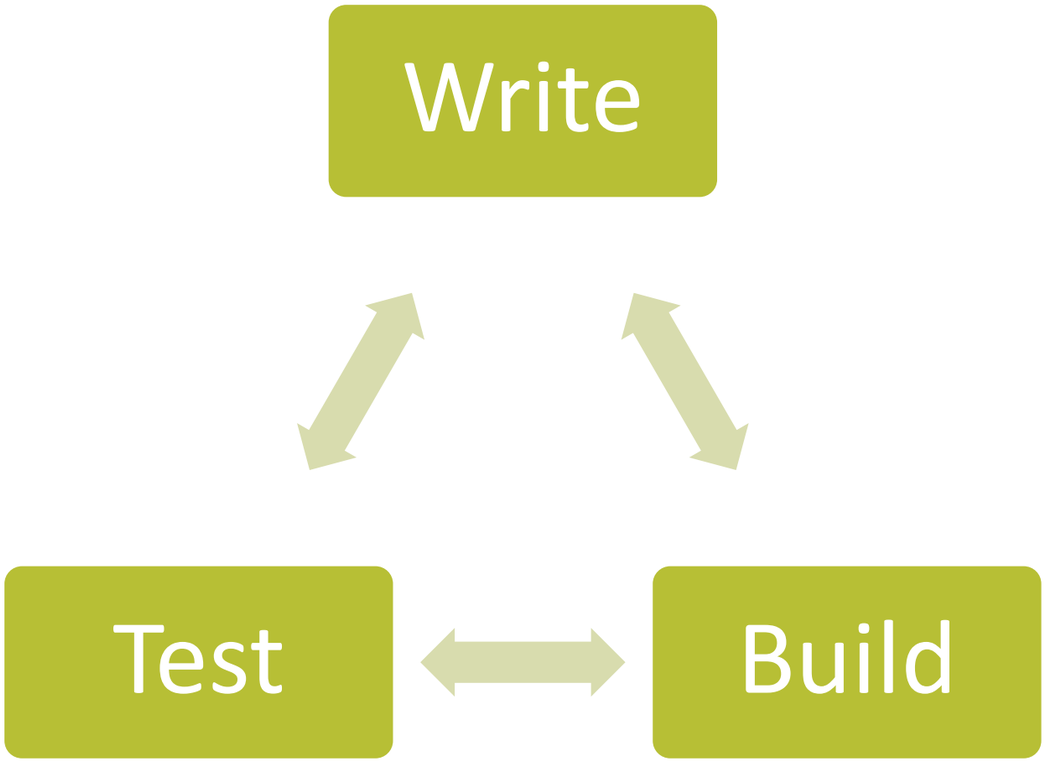
Development Flow

Ideal

Development Flow

Ideal

Reality



Existing Solutions

- Plenty of docker builders
- virtme
  - Relies on running kernel environment
  - Based on busybox and not on real OS
  - https://git.kernel.org/pub/scm/utils/kernel/virtme/virtme.git/
- docker-qemu
  - Run full VM
  - https://github.com/Ulexus/docker-qemu
- ….
  - Don’t have any option to use compiled kernel

github.com/mellanox/mkt

Note: It contains Mellanox specific code

Layers

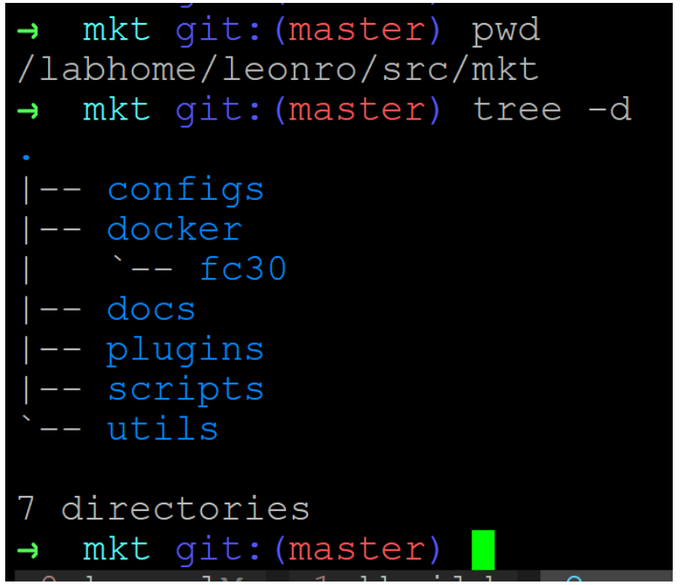
Hypervisor	• source code • build artefacts • logs
Runners	• build • ci • run • images
Containers	• qemu runner • builders support

Initial Setup

- Pre-requirements
  - Modern distribution, tested on Fedora 26, Ubuntu 16.04 and RedHat 8
  - Python 3.5 or above

Initial Setup

- Pre-requirements
  - Modern distribution, tested on Fedora 26, Ubuntu 16.04 and RedHat 8
  - Python 3.5 or above
- Download and setup MKT to be in the PATH
  - **cd**
  - **git clone https://github.com/Mellanox/mkt.git**
  - **mkdir ~/bin**
  - **ln -s $HOME/mkt/mkt ~/bin/**
  - **export PATH=$HOME/bin:$PATH**



Initial Setup

- Pre-requirements
  - Modern distribution, tested on Fedora 26, Ubuntu 16.04 and RedHat 8
  - Python 3.5 or above
- Download and setup MKT to be in the PATH
  - **cd**
  - **git clone https://github.com/Mellanox/mkt.git**
  - **mkdir ~/bin**
  - **ln -s $HOME/mkt/mkt ~/bin/**
  - **export PATH=$HOME/bin:$PATH**
- Install Docker CE, git and bring source code from gerrit
  - **mkt setup**
  - **mkt setup-master** – for multi-machine setups
  - **mkt setup-slave MASTER_IP** – for multi-machine setups

```
→  mkt pwd
/labhome/leonro/.config/mellanox/mkt
→  mkt hostname
nps-server-14
→  mkt l hv-nps-server-14.mkt
-rw-r--r-- 1 leonro mtl 916 Jul 14 15:21 hv-nps-server-14.mkt
```

```
→  mkt git:(master) pwd
/labhome/leonro/src/mkt
→  mkt git:(master) tree -d
.
|-- configs
|-- docker
|   `-- fc30
|-- docs
|-- plugins
|-- scripts
`-- utils

7 directories
→  mkt git:(master)
```

```
→  leonro pwd
/images/leonro
→  leonro tree -d -L 2 --noreport
.
|-- ccache
|-- logs
|   |-- build
|   |-- ci
|   |-- images
|   |-- run
|   |-- setup
|   `-- setup-master
`-- src
    |-- iproute2
    |-- kernel
    `-- rdma-core
→  leonro
```

Support Container

- Has all dependencies to build
  - Kernel
  - QEMU
  - rdma-core
  - iproute2
- In use for
  - Source code build
  - Local CI
  - Local KVM images
- Files generated with user ownership
- Allows installation of any software
  - git tree url
  - git commit SHA
  - spec file
  - extra patches

```
#!/bin/bash
# ---
# git_url: git://repo.or.cz/smatch.git
# git_commit: f0092daff69d4b06b174122d301d8e3d7cdf3825
# other_files:
#   - 0001-Explicitly-use-python2-to-solve-rpmbuild-error.patch

patch -p1 < /opt/00*.patch

cat <<EOF > smatch.spec
Name: smatch
Version: 1
Release: 1%{?dist}
Summary:    A semantic parser of source files
Group:      Development/Tools
License:    MIT
URL:        http://smatch.sourceforge.net/

%description
Smatch is a semantic parser of source files.

%build
make %{?_smp_mflags}

%install
make INSTALL_PREFIX="/opt/smatch" DESTDIR="%{buildroot}" PREFIX="/opt/smatch" install
mkdir -p %{buildroot}/opt/smatch/share/smatch/smatch_data/
cp -r /opt/src/smatch_data/db %{buildroot}/opt/smatch/share/smatch/smatch_data/

%clean
make clean

%files
/opt/smatch/share/man/man1/*
/opt/smatch/bin/*
/opt/smatch/include/*
/opt/smatch/share/smatch/*
/opt/smatch/share/smatch/smatch_data/db/*
/opt/smatch/lib/*
/opt/smatch/lib/pkgconfig/*
EOF

rpmbuild --build-in-place -bb smatch.spec
```

Build Code

- Silent and smart project build discovery
  - **mkt build** <project_to_build>
- Preconfigured CCACHE to speed up recompilation
- Proper compilation flags
- Correct understanding of number of available CPUs for build
- Build from recipe file for custom builds
- Able to build user space applications against new kernel headers, useful for user space development
- **Minimal** kernel .config
  - virtio-* drivers
  - Pre-configured to boot from 9pfs filesystem
  - Only Mellanox drivers are enabled

```
→  kernel git:(rdma-next) pwd
/images/leonro/src/kernel
→  kernel git:(rdma-next) time mkt build
Start kernel compilation in silent mode
mkt build  0.16s user 0.06s system 0% cpu 46.612 total
```

CI Testing

- Focused on code static analyzers
  - smatch from the git
  - sparse from the git
  - Latest gcc with extra warnings
  - checkpatch
  - clang-9
  - Various compilation tests
- Reuse support container and build runner
  - Common CCACHE
  - Deep patch inspection to compile only minimal part
- Non-blocking asynchronic compilation
- Executed with **mkt ci**

```
➜  kernel git:(rdma-next) time mkt ci
2467425b0b34 (HEAD -> build) IB/mlx5: Add CREATE_PSV/DESTROY_PSV for devx interface
In file included from ./include/rdma/ib_verbs.h:64,
                 from drivers/infiniband/hw/mlx5/mlx5_ib.h:38,
                 from drivers/infiniband/hw/mlx5/gsi.c:33:
./include/linux/dim.h:378:1: warning: ‘rdma_dim_prof’ defined but not used [-Wunused-const-variable=]
  378 | rdma_dim_prof[RDMA_DIM_PARAMS_NUM_PROFILES] = {
      | ^~~~~~~~~~~~~
./include/linux/dim.h:326:1: warning: ‘tx_profile’ defined but not used [-Wunused-const-variable=]
  326 | tx_profile[DIM_CQ_PERIOD_NUM_MODES][NET_DIM_PARAMS_NUM_PROFILES] = {
      | ^~~~~~~~~~
./include/linux/dim.h:320:1: warning: ‘rx_profile’ defined but not used [-Wunused-const-variable=]
  320 | rx_profile[DIM_CQ_PERIOD_NUM_MODES][NET_DIM_PARAMS_NUM_PROFILES] = {
      | ^~~~~~~~~~
mkt ci  0.17s user 0.05s system 0% cpu 1:13.80 total
```

Run Flow

- Rich CLI and configuration file
  - **mkt run** <subsection>
- Fast boot into VM
- No need to generate VM image for QEMU
- No need to copy/install kernel and modules
- Includes working network and SSH connection
- Ctrl-A X closes QEMU and kills container

```
[  OK  ] Started Update UTMP about System Runlevel Changes.
[   19.831047] IPv6: ADDRCONF(NETDEV_CHANGE): ib0: link becomes ready
[leonro@nps-server-14-015 ~]$ pwd
/labhome/leonro
[leonro@nps-server-14-015 ~]$ uname -a
Linux nps-server-14-015 5.2.0-rc6+ #205 SMP Wed Jul 17 12:09:58 UTC 2019 x86_64 x86_64 x86_64 GNU/Linux
[leonro@nps-server-14-015 ~]$ lspci |grep nox
00:0d.0 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4]
[leonro@nps-server-14-015 ~]$
[leonro@nps-server-14-015 ~]$ ls -l /lib/modules/5.2.0-rc6+/modules
total 68
lrwxrwxrwx 1 root root 49 Jul 17 13:00 crc32_generic.ko -> /images/leonro/src/kernel/crypto/crc32_generic.ko
lrwxrwxrwx 1 root root 44 Jul 17 13:00 echainiv.ko -> /images/leonro/src/kernel/crypto/echainiv.ko
lrwxrwxrwx 1 root root 58 Jul 17 13:00 ib_cm.ko -> /images/leonro/src/kernel/drivers/infiniband/core/ib_cm.ko


→  ~ ssh root@nps-server-14-015
root@nps-server-14-015's password:
[root@nps-server-14-015 ~]# pwd
/root
```

Rich Configuration Syntax

```
[defaults]
src = /images/leonro/src/
kernel = /images/leonro/src/kernel/
rdma-core = /images/leonro/src/rdma-core/
iproute2 = /images/leonro/src/iproute2/
simx = /images/leonro/src/simx/
logs = /images/leonro/logs/
ccache = /images/leonro/ccache/
image = simx
dir = /images/leonro/src/rdma-core /images/leonro/src/iproute2

[cx5-ib]
pci = 0000:05:00.0 0000:05:00.1
boot_script = /labhome/leonro/scripts/opensm

[cx4-ib]
pci = 0000:0b:00.0 0000:0b:00.1
boot_script = /labhome/leonro/scripts/opensm

[cx5-roce]
pci = 0000:88:00.0 0000:88:00.1

[cx4-roce]
pci = 0000:84:00.0 0000:84:00.1

[cx3]
pci = 0000:81:00.0
num_of_vfs = 3
boot_script = /labhome/leonro/scripts/opensm

[cxib]
pci = 0000:08:00.0
boot_script = /labhome/leonro/scripts/opensm

[simx]
pci = cx4-ib

[simx-sriov]
pci = cx4-eth cx6-eth
num_of_vfs = 6
custom_qemu = true
```

QEMU Image

- Don’t create VM images – use container layout
- Built on the fly as docker entrypoint
- Everything is virtio-9p-pci
- Mount with systemd

QEMU Image

- Don’t create VM images – use container layout
- Built on the fly as docker entrypoint
- Everything is virtio-9p-pci
- Mount with systemd

Mount root fs inside docker

QEMU Image

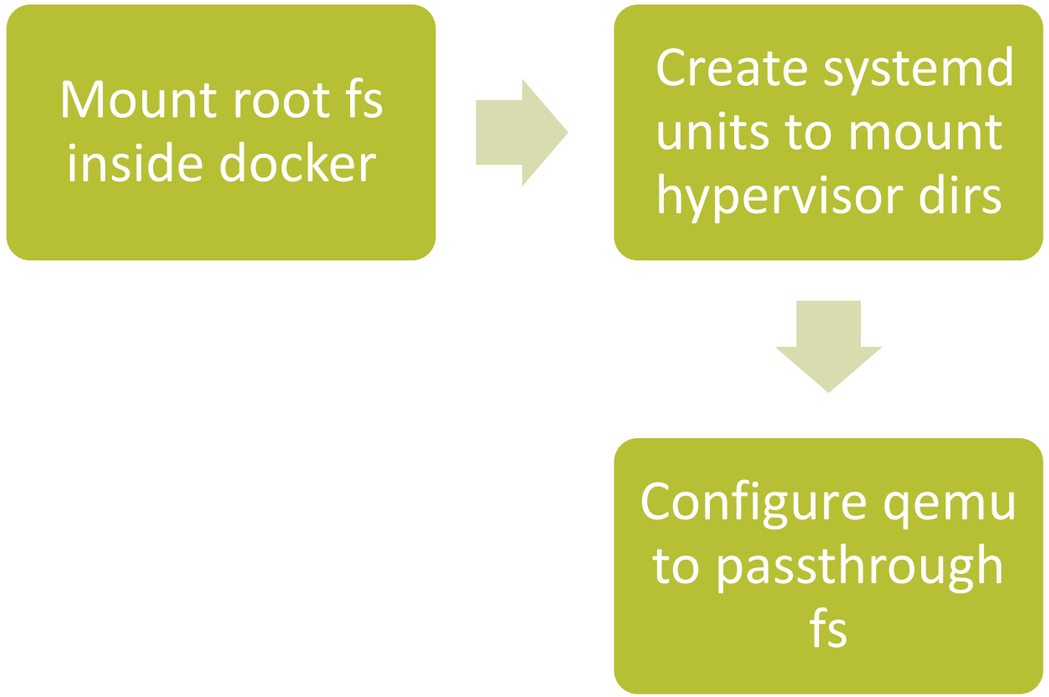
- Don’t create VM images – use container layout
- Built on the fly as docker entrypoint
- Everything is virtio-9p-pci
- Mount with systemd

Mount root fs inside docker

Create systemd units to mount hypervisor dirs

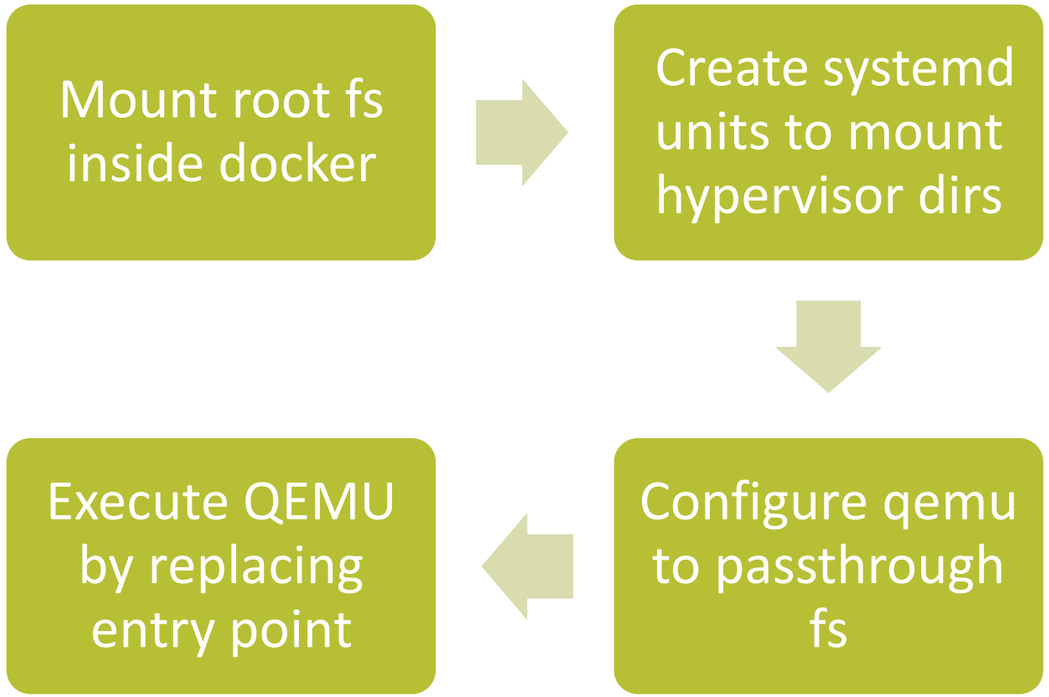
QEMU Image

- Don’t create VM images – use container layout
- Built on the fly as docker entrypoint
- Everything is virtio-9p-pci
- Mount with systemd



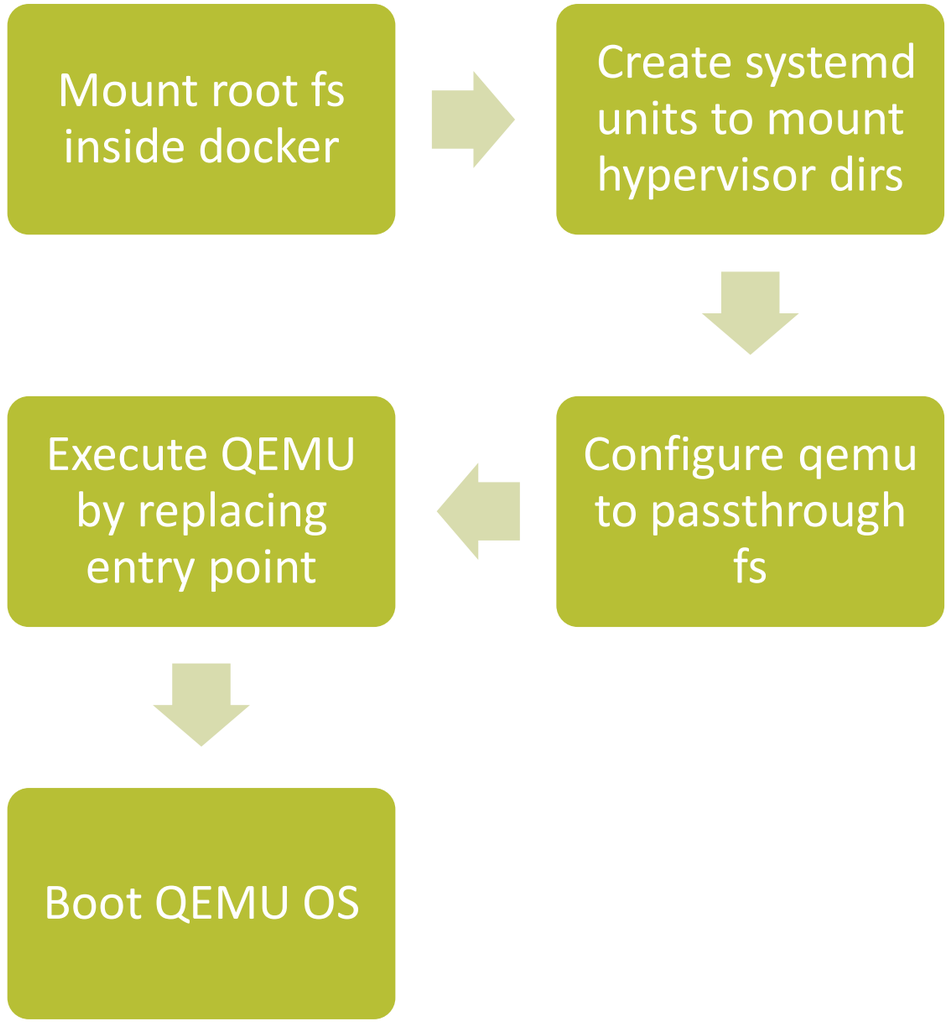
QEMU Image

- Don’t create VM images – use container layout
- Built on the fly as docker entrypoint
- Everything is virtio-9p-pci
- Mount with systemd



QEMU Image

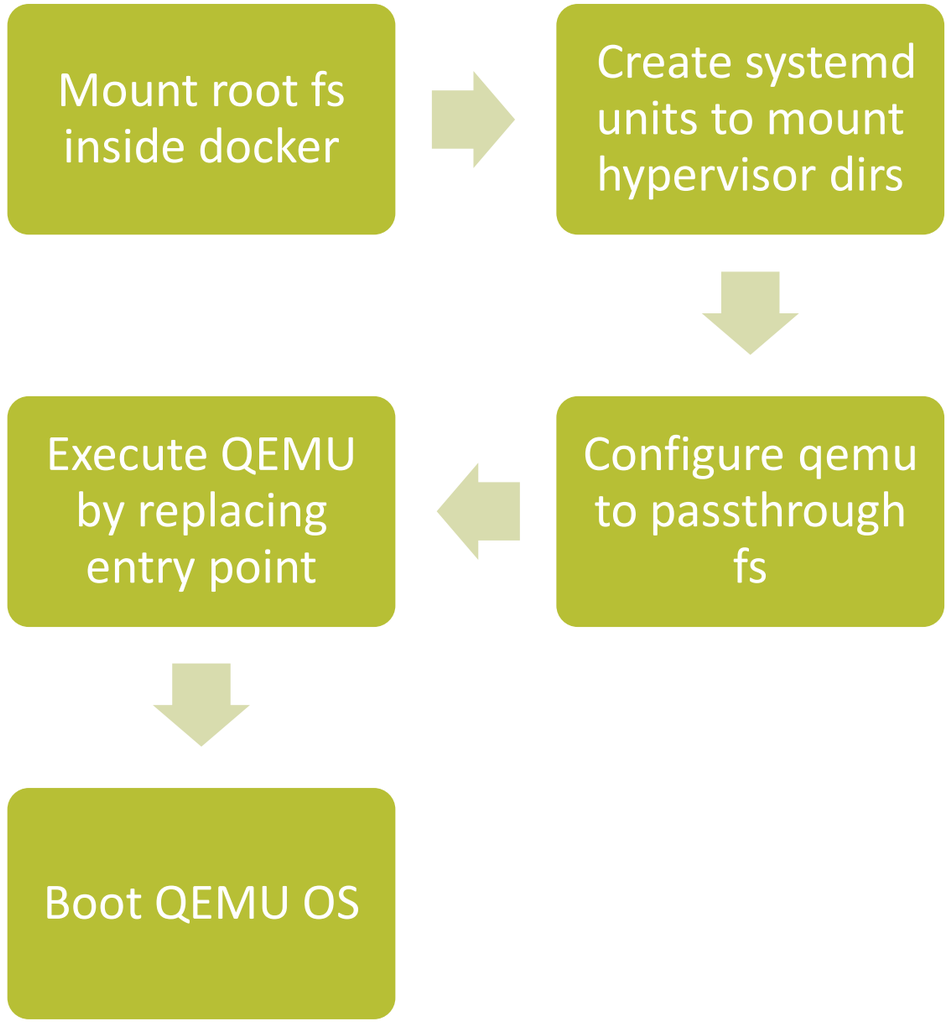
- Don’t create VM images – use container layout
- Built on the fly as docker entrypoint
- Everything is virtio-9p-pci
- Mount with systemd



QEMU Image

- Don’t create VM images – use container layout
- Built on the fly as docker entrypoint
- Everything is virtio-9p-pci
- Mount with systemd

- mount –bind / /mnt/self
- qemu … -fsdev local,id=host_fs,security_model=**passthrough**,path=/mnt/self -device virtio-9p-pci,fsdev=host_fs,mount_tag=/dev/root …
- os.execvp(….)



QEMU Network

- Managed interface
  - Containers run in privileged mode
    - --net=host --privileged
  - Senses br0 interface
    - + full external in/out network
    - - NAT access for localhost port 4444 connected to SSH

QEMU Network

- Managed interface
  - Containers run in privileged mode
    - --net=host --privileged
  - Senses br0 interface
    - + full external in/out network
    - - NAT access for localhost port 4444 connected to SSH
- Tested interface (External routing)
  - Disable reverse proxy and ARP filtering
  - Configure routing table
  - Increase priority of output port
  - Clean routing cache

External Routing (2 NICs)

echo 0 > /proc/sys/net/ipv4/conf/all/rp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/accept_local
echo 1 > /proc/sys/net/ipv4/conf/all/arp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/arp_ignore
echo 2 > /proc/sys/net/ipv4/conf/all/arp_announce

Disable filtering

External Routing (2 NICs)

```
echo 0 > /proc/sys/net/ipv4/conf/all/rp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/accept_local
echo 1 > /proc/sys/net/ipv4/conf/all/arp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/arp_ignore
echo 2 > /proc/sys/net/ipv4/conf/all/arp_announce

ip rule del pref 0
ip rule add from all lookup local pref 100

ip rule add iif eth1 lookup local pref 0
ip rule add from 192.168.122.76 table 10 pref 10
ip route add 192.168.122.0/24 dev eth1 src 192.168.122.76 table 10
ip route add local 192.168.122.76 dev eth1 src 192.168.122.76 table 10

ip rule add iif eth2 lookup local pref 0
ip rule add from 192.168.122.77 table 11 pref 10
ip route add 192.168.122.0/24 dev eth2 src 192.168.122.77 table 11
ip route add local 192.168.122.77 dev eth1 src 192.168.122.77 table 11
```

Disable filtering

Configure routing table

External Routing (2 NICs)

```
echo 0 > /proc/sys/net/ipv4/conf/all/rp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/accept_local
echo 1 > /proc/sys/net/ipv4/conf/all/arp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/arp_ignore
echo 2 > /proc/sys/net/ipv4/conf/all/arp_announce

ip rule del pref 0
ip rule add from all lookup local pref 100

ip rule add iif eth1 lookup local pref 0
ip rule add from 192.168.122.76 table 10 pref 10
ip route add 192.168.122.0/24 dev eth1 src 192.168.122.76 table 10
ip route add local 192.168.122.76 dev eth1 src 192.168.122.76 table 10

ip rule add iif eth2 lookup local pref 0
ip rule add from 192.168.122.77 table 11 pref 10
ip route add 192.168.122.0/24 dev eth2 src 192.168.122.77 table 11
ip route add local 192.168.122.77 dev eth1 src 192.168.122.77 table 11

ip rule add to 192.168.122.77 table 10 pref 10
ip rule add to 192.168.122.76 table 11 pref 10
```

Disable filtering

Configure routing table

Increase priority

External Routing (2 NICs)

```
echo 0 > /proc/sys/net/ipv4/conf/all/rp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/accept_local
echo 1 > /proc/sys/net/ipv4/conf/all/arp_filter
echo 1 > /proc/sys/net/ipv4/conf/all/arp_ignore
echo 2 > /proc/sys/net/ipv4/conf/all/arp_announce

ip rule del pref 0
ip rule add from all lookup local pref 100

ip rule add iif eth1 lookup local pref 0
ip rule add from 192.168.122.76 table 10 pref 10
ip route add 192.168.122.0/24 dev eth1 src 192.168.122.76 table 10
ip route add local 192.168.122.76 dev eth1 src 192.168.122.76 table 10

ip rule add iif eth2 lookup local pref 0
ip rule add from 192.168.122.77 table 11 pref 10
ip route add 192.168.122.0/24 dev eth2 src 192.168.122.77 table 11
ip route add local 192.168.122.77 dev eth1 src 192.168.122.77 table 11

ip rule add to 192.168.122.77 table 10 pref 10
ip rule add to 192.168.122.76 table 11 pref 10

ip route flush cache
```

Disable filtering

Configure routing table

Increase priority

Flush

QEMU Hardware Support

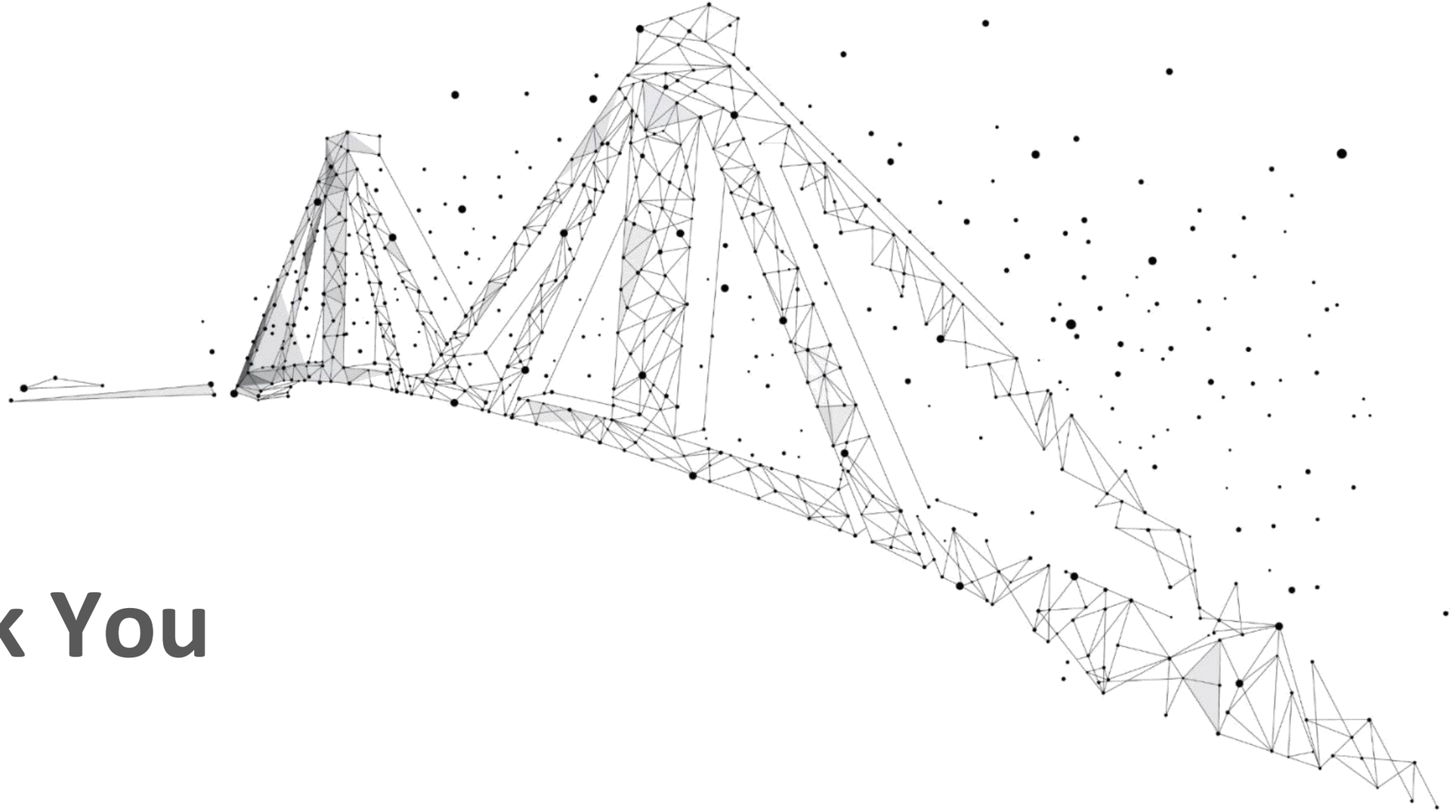
- Based on VFIO PCI
- qemu … -device vfio-pci,host=PCI_BOF …

QEMU Hardware Support

- Based on VFIO PCI
- qemu … -device vfio-pci,host=PCI_BOF …

github.com/mellanox/mkt

Join us and make MKT generic



Thank You