oVirt SR-IOV support

Barak Azulay
Senior Manager, Software Engineering

Credits: Alona Kaplan, Martin Polednik, Ido Barkan

Red Hat
21/08/15
Agenda

- SR-IOV basics (what, how, limitations)
- Ovirt Networking basics
- Ovirt Implementation of SR-IOV support
- Future improvements
SR-IOV basics - What is SR-IOV?

specification that allows a PCIe device to appear to be multiple separate physical PCIe devices.
SR-IOV basics - Physical Function (PF)

Full PCIe device that includes the SR-IOV capabilities.
‘lightweight’ PCIe functions that contain the resources necessary for data movement but have a carefully minimized set of configuration resources.
SR-IOV basics - how to add VFs

Before

```
[root@nari04 ~]# ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
3: enp2s0f0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq master ovirtmgmt state UP mode DEFAULT qlen 1000
    link/ether 78:e7:d1:e4:8f:16 brd ff:ff:ff:ff:ff:ff
4: enp2s0f1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master bond0 state UP mode DEFAULT qlen 1000
    link/ether 78:e7:d1:e4:8f:17 brd ff:ff:ff:ff:ff:ff
```

Add VFs

```
[root@nari04 ~]# echo 4 > /sys/class/net/enp2s0f0/device/sriov_numvfs
```

After

```
[root@nari04 ~]# ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
3: enp2s0f0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq master ovirtmgmt state UP mode DEFAULT qlen 1000
    link/ether 78:e7:d1:e4:8f:16 brd ff:ff:ff:ff:ff:ff
    vf 0 MAC 00:00:00:00:00:00, spoof checking on, link-state auto
    vf 1 MAC 00:00:00:00:00:00, spoof checking on, link-state auto
    vf 2 MAC 00:00:00:00:00:00, spoof checking on, link-state auto
    vf 3 MAC 00:00:00:00:00:00, spoof checking on, link-state auto
4: enp2s0f1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master bond0 state UP mode DEFAULT qlen 1000
    link/ether 78:e7:d1:e4:8f:17 brd ff:ff:ff:ff:ff:ff
35: enp2s1f0: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether 4a:2f:29:9b:fa:14 brd ff:ff:ff:ff:ff:ff
36: enp2s1f2: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether fe:0c:29:cc:b5:fa brd ff:ff:ff:ff:ff:ff
37: enp2s1f4: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether 4a:c3:8f:6d:0e:40 brd ff:ff:ff:ff:ff:ff
38: enp2s1f6: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether b2:32:2a:82:4d:fd brd ff:ff:ff:ff:ff:ff
```
SR-IOV basics - Advantages

✔ VFs have near-native **performance**.

✔ **low latency**.

✔ **scalability** of the host is improved (more CPU available to apps in VMs).

✔ VM has **direct** access to the hardware.

✔ **Guest protection/isolation**

✔ VMs can **share** a single physical port.
SR-IOV basics - Limitations

- Vfs number is limited by the device hardware.
- realistic 'num of VFs' should be set manually.
- VFs have limited configuration functions.
- live migration.
SR-IOV basics - Requirements

- hypervisor
  - hardware IOMMU support (AMD-Vi, Intel VT-d enabled in BIOS).
  - kernel enabled IOMMU support (intel_iommu=on for Intel, amd_iommu=on for AMD in kernel cmdline).
  - SR-IOV capable hardware.
  - RHEL7 or newer (kernel >= 3.6).

- SR-IOV support in the guest (driver).
oVirt Networking

- Logical Network (VM, non-VM).
- Setup networks - Configuring the logical networks on the hypervisor
- VM Interface Profile (vNic profile).
- VM Interface (vNic).
Logical Network

VM 1
VNIC: eth1
10.64.14.160

VM 2
VNIC: eth2
10.64.14.161

VM 3
VNIC: eth3
10.64.14.162

VM 4
VNIC: eth4
10.64.14.163

Logical Network: rhevm

Bridge: rhevm
10.64.14.171

Host 1
NIC: eth0

Bridge: rhevm
10.64.14.172

Host 2
NIC: eth0

KVM forum, Aug 2015
Host Networks setup

[Diagram showing Host, VLAN A, Bridge, VLAN B, and multiple Guests connected through vNICs and Setup Networks.]

KVM forum, Aug 2015
### Attaching VM to a new network demo

#### Open Virtualization Manager

<table>
<thead>
<tr>
<th>Name</th>
<th>Comment</th>
<th>Data Center</th>
<th>Description</th>
<th>Role</th>
<th>VLAIR Tag</th>
<th>Label</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>netmgmt</td>
<td>Default</td>
<td>mb</td>
<td>Management network</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-1</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-10</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net11</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-2</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-3</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-4</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-5</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-6</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-7</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-8</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net-9</td>
<td>mb</td>
<td>mb</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>netmgmt</td>
<td>mb</td>
<td>mb</td>
<td>Management network</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

---

KVM forum, Aug 2015
OVirt & SR-IOV

The problem:

- SR-IOV passthrough belongs to the physical layer of Network
- It is not associated with logical network

The solution:

- Define in advance the networks list that could be used by the SR-IOV device (PF)
- Add specific vNIC profile type of passthrough
- Associate the vNIC to the passthrough vNIC profile
Setup Networks

Drag to make changes

Interfaces

Assigned Logical Networks

- `em1`
- `em2`
- `em3`
- `em4`

Unassigned Logical Networks

- `ovirtmgmt`
- `sr_net1`
- `no network assigned`

External Logical Networks

- `label_net5`
- `lb_net1 (VLAN 145)`

- Verify connectivity between Host and Engine
- Save network configuration

KVM forum, Aug 2015
VF configuration

Edit Virtual Functions (SR-IOV) configuration of em1

- **Number of VF setting**
  - Number of VFs: 0

- **Allowed Networks**
  - Specific networks
  - Select Network(s):
    - label_net5: k
    - lb_net1: k
    - net1: k
    - newlb
  - Labels:
    - k

KVM forum, Aug 2015
Passthrough VM Interface profile

VM Interface Profile

Network: sr_net1
Name: sr_net1
Description:
QoS: [Unlimited]
Passthrough: checked
Port Mirroring: unchecked
Passthrough VM Interface

Edit Network Interface

- Name: nic1
- Profile: sr_net1/sr_net1
- Type: PCI Passthrough
- Custom MAC address: 00:1a:4a:16:01:51
- Link State: Up
- Card Status: Plugged
Run VM with passthrough vNic

<table>
<thead>
<tr>
<th>Name</th>
<th>Comment</th>
<th>DataCenter</th>
<th>Description</th>
<th>Role</th>
<th>VLAN Tag</th>
<th>Label</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>ovirtmgnt</td>
<td></td>
<td>mb</td>
<td>Management Network</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net1</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net10</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net11</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net2</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mb</td>
</tr>
<tr>
<td>net3</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net4</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net5</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net6</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net7</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net8</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net9</td>
<td>mb</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>net_sr_int</td>
<td>mb</td>
<td></td>
<td></td>
<td>162</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ovirtmgnt</td>
<td>mb</td>
<td></td>
<td>Management Network</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General</th>
<th>vBMC Profiles</th>
<th>Clusters</th>
<th>Hosts</th>
<th>Virtual Machines</th>
<th>Templates</th>
<th>Permissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>net_sr_int</td>
<td></td>
<td></td>
<td></td>
<td>VM network</td>
<td>true</td>
</tr>
<tr>
<td>ID</td>
<td>01051562-067b-4e4b-a3d0-d07f4e070228</td>
<td></td>
<td></td>
<td></td>
<td>VLAN tag</td>
<td>112</td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mtu:</td>
<td>Default (1500)</td>
</tr>
</tbody>
</table>
oVirt SR-IOV capabilities

- Change & persist number of VFs (sysfs) via ui.
- Managing PFs network connectivity white-list.
- Scheduling – no need to pin to a host
- Setting VLAN and MAC address on a VF.
SR-IOV capabilities- cont

- Mixed mode- bridged PF with VFs.
- Specifying boot order on Vfs (enabling booting VM with passthrough vNics from pxe).
oVirt SR-IOV future capabilities

- Hot plug/unplug passthrough vnics.
- Live Migration
- Opportunistic passthrough vnic.
VF missing functionality

- MTU (not supported)
- QoS (in/out- average link share, average upper limit, average real time).
Harware issues

- VFs share the IOMMU group.
- IOMMU is not supported (under sysfs - the devices doesn’t get iommu-group number).
- Hacks are needed
  - pci=realloc - ‘igb <0000:02:00.1>: not enough MMIO resources for SR-IOV’
  - pci=assign-busses - ‘igb <0000:06:10.0>: SR-IOV: bus number out of range’
  - vfio_iommu_type1.allow_unsafe_interrupts=1 - On systems with broken interrupt remapping (problematic chipset)
Questions ?
THANK YOU!

http://www.ovirt.org
bazulay@redhat.com
bazulay@irc.oftc.net#ovirt