

oVirt SR-IOV support

Barak Azulay
Senior Manager, Software Engineering

Credits: Alona Kaplan, Martin Polednik, Ido Barkan

Red Hat
21/08/15

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

specification that allows a PCIe device to appear to be multiple separate physical PCIe devices.

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.

oVirt SR-IOV basics - how to add VFs

Before

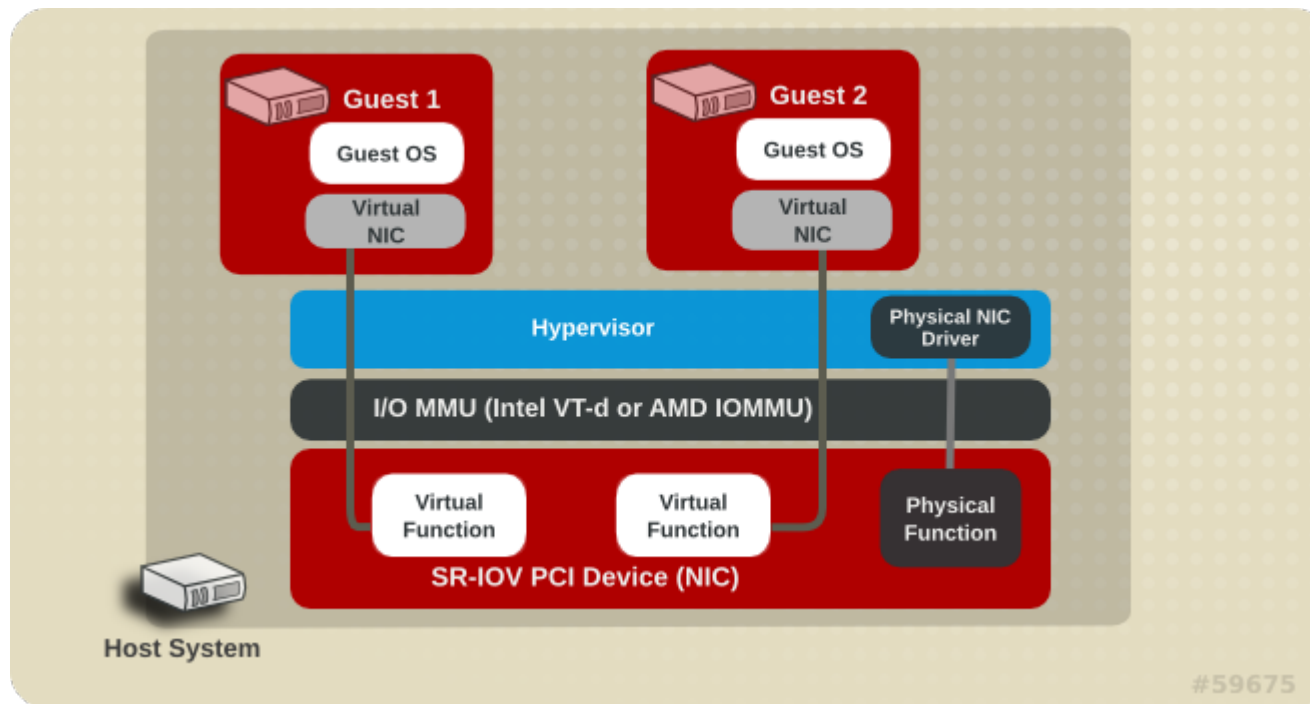
```
[root@nari04 ~]# ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
3: enp2s0f0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq master ovirtmgmt state UP mode DEFAULT qlen 1000
    link/ether 78:e7:d1:e4:8f:16 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: enp2s16: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether 4a:2f:20:98:fa:14 brd ff:ff:ff:ff:ff:ff
36: enp2s16f2: <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: enp2s16f4: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether 4a:c3:8f:6d:6e:40 brd ff:ff:ff:ff:ff:ff
38: enp2s16f6: <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
```



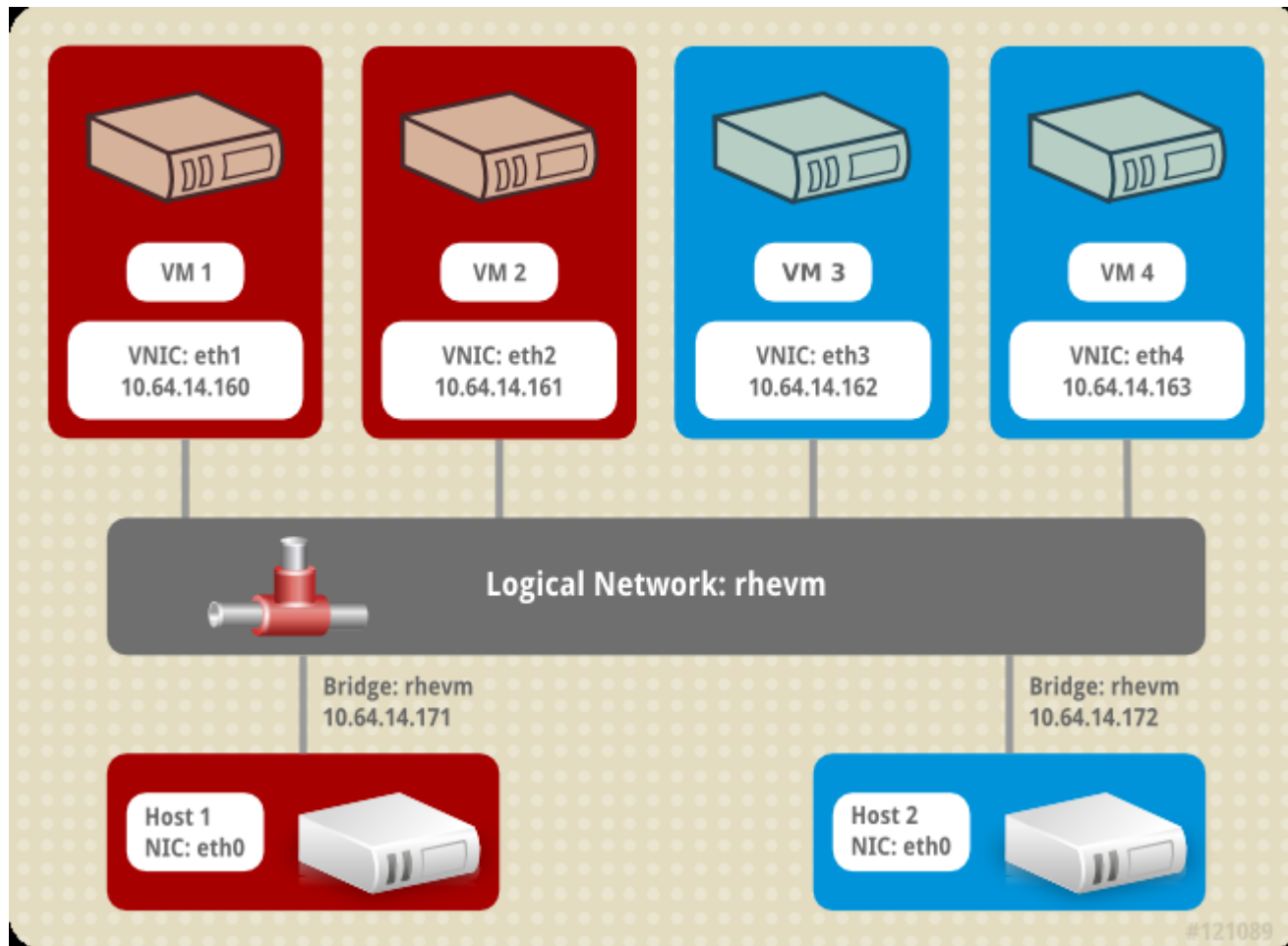
- ✓ VFs have near-native **performance**.
- ✓ low **latency**.
- ✓ **scalability** of the host is improved (more CPU available to apps in VMs).
- ✓ VM has **direct** access to the hardware.
- ✓ **Guest protection/isolation**
- ✓ VMs can **share** a single physical port.

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

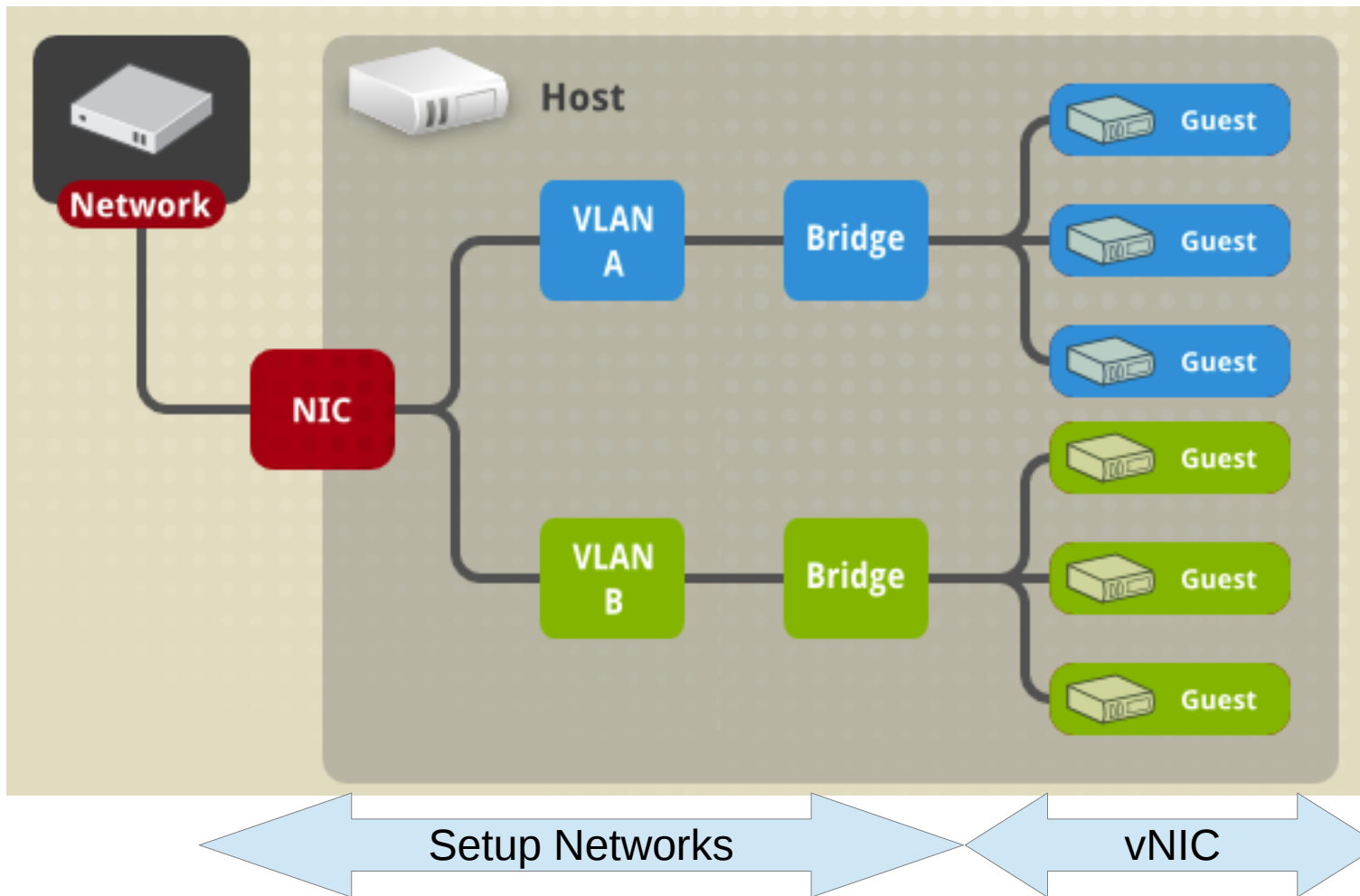
- hypervisor
 - hardware IOMMU support (AMD-Vi, Intel VT-d enabled in BIOS) .
 - kernel enabled IOMMU support (intel_iommu=on for Intel, amd_iommu=on for AMD in kernel cmdline) .
 - SR-IOV capable hardware.
 - RHEL7 or newer (kernel \geq 3.6).
- SR-IOV support in the guest (driver).

oVirt Networking

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



Host Networks setup



oVirt Attaching VM to a new network demo

oVirt OPEN VIRTUALIZATION MANAGER

Network:

Data Centers Clusters Hosts **Networks** Storage Disks Virtual Machines Pools Templates Volumes Users

New Import Edit Remove

Name	Comment	Data Center	Description	Role	VLAN tag	Label	Provider
ovirtmgmt		Default	Management Network	✓	-	-	
net-1		mb		✓	-	-	
net-10		mb		✓	-	-	
net111		mb		✓	-	-	
net-2		mb		✓	-	mb	
net-3		mb		✓	-	-	
net-4		mb		✓	-	-	
net-5		mb		✓	-	-	
net-6		mb		✓	-	-	
net-7		mb		✓	-	-	
net-8		mb		✓	-	-	
net-9		mb		✓	-	-	
ovirtmgmt		mb	Management Network	✓	-	-	

Last Message: ✓ 2015-Aug-06, 12:19 Network sr_iov_net1 was removed from Data Center: mb

- ✓ 2015-Aug-06, 12:19 ✗ Network sr_iov_net1 was removed from Data Center: mb
- ✓ 2015-Aug-06, 12:18 ✗ Interface nic1 (VirtIO) was removed from VM vm_6_... (User: admin@internal)
- ✓ 2015-Aug-06, 12:17 ✗ Network changes were saved on host puma22.scl.ltv.redhat.com
- ✓ 2015-Aug-06, 12:15 ✗ Interface nic1 (VirtIO) was added to VM vm_6_... (User: admin@internal)

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

oVirt Setup Networks

Setup Host 10.1.64.53 Networks

Drag to make changes

Interfaces	Assigned Logical Networks
 em1	ovirtmgmt
 em2	sr_net1
em3	no network assigned
em4	k label_net5 lb_net1 (VLAN 145)

Unassigned Logical Networks

Networks Labels

Required

Non Required

External Logical Networks ?

Verify connectivity between Host and Engine ?

Save network configuration ?

OK Cancel

The screenshot displays the oVirt management interface. On the left, a sidebar titled "Setup Host 10.1.64.53 Net" shows a list of interfaces: em1, em2, em3, and em4. The main window is titled "Edit Virtual Functions (SR-IOV) configuration of em1".

Edit Virtual Functions (SR-IOV) configuration of em1

- Number of VFs setting:** A text input field contains "0", with a range of "0 - 63" below it.
- Allowed Networks:** Two radio buttons are present: "All networks" (unselected) and "Specific networks" (selected).
- Select Network(s):** A table with columns "Network" and "Via Label".

Network	Via Label
<input type="checkbox"/> label_net5	k
<input checked="" type="checkbox"/> lb_net1	k
<input checked="" type="checkbox"/> net1	k
<input checked="" type="checkbox"/> newlb	
<input type="checkbox"/> ovirtmgmt	
<input type="checkbox"/> sr_net1	
- Labels:** A text input field contains "k", with minus and plus buttons to its right.

At the bottom of the dialog, there are "OK" and "Cancel" buttons. The background interface is partially visible, showing "Logical Networks" and "Required" sections.

oVirt Passthrough VM Interface profile

The image shows a dialog box titled "VM Interface Profile" with a close button (X) in the top right corner. The dialog contains several fields and checkboxes:

- Network:** A dropdown menu showing "sr_net1".
- Name:** A text input field containing "sr_net1".
- Description:** An empty text input field.
- QoS:** A dropdown menu showing "[Unlimited]".
- Passthrough:** A checked checkbox.
- Port Mirroring:** An unchecked checkbox.
- Key:** A dropdown menu showing "Please select a key..." with minus and plus buttons to its right.

At the bottom right of the dialog are "OK" and "Cancel" buttons.



Edit Network Interface ✕



Name

Profile

Type

Custom MAC address
Example: 00:14:4a:23:67:55

Link State  Up  Down

Card Status  Plugged  Unplugged

oVirt Run VM with passthrough vNic

The screenshot displays the oVirt Open Virtualization Manager interface. At the top, the navigation bar includes 'oVirt OPEN VIRTUALIZATION MANAGER', a user dropdown for 'admin', and links for 'Configure', 'Guide', 'About', and 'Feedback'. Below this is a search bar labeled 'Network:' with 'x', 'star', and 'Q' icons. A secondary navigation bar contains tabs for 'Data Centers', 'Clusters', 'Hosts', 'Networks', 'Storage', 'Disks', 'Virtual Machines', 'Pools', 'Templates', 'Volumes', and 'Users'. The 'Networks' tab is active, showing a table of network configurations. The table has columns for Name, Comment, Data Center, Description, Role, VLAN tag, Label, and Provider. The selected network is 'net_sr_iov1' with a VLAN tag of 162. Below the table, the 'General' tab is selected, showing details for the network: Name: net_sr_iov1, Id: 819dd182-3e79-446f-ad80-efd7487e0208, Description: (empty), VM Network: true, VLAN tag: 162, and MTU: Default (1500). A 'Quit' button is visible in the bottom left corner. The status bar at the bottom shows a last message from 2015-Aug-06, 16:49, 8 alerts, and 10 tasks.

Name	Comment	Data Center	Description	Role	VLAN tag	Label	Provider
ovirtmgmt		Default	Management Network	✓	-	-	
net-1		mb		✓	-	-	
net-10		mb		✓	-	-	
net111		mb		✓	-	-	
net-2		mb		✓	-	mb	
net-3		mb		✓	-	-	
net-4		mb		✓	-	-	
net-5		mb		✓	-	-	
net-6		mb		✓	-	-	
net-7		mb		✓	-	-	
net-8		mb		✓	-	-	
net-9		mb		✓	-	-	
net_sr_iov1		mb		✓	162	-	
ovirtmgmt		mb	Management Network	✓	-	-	

General | vNIC Profiles | Clusters | Hosts | Virtual Machines | Templates | Permissions

Name: net_sr_iov1 VM Network: true
Id: 819dd182-3e79-446f-ad80-efd7487e0208 VLAN tag: 162
Description: MTU: Default (1500)

Quit

Last Message: ✓ 2015-Aug-06, 16:49 VM vm_sriov was powered off ungracefully by admin@internal (Host: puma22.scl.lab.tlv.redhat.com) (Reason: Not Specified) Alerts (8) Events Tasks (10)

- 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.

- Mixed mode- bridged PF with VFs.
- Specifying boot order on Vfs (enableing booting VM with passthrough vNics from pxe).

- 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).

Hardware 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