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.

SR-IOV basics - Virtual Functions (VFs)

‘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
```

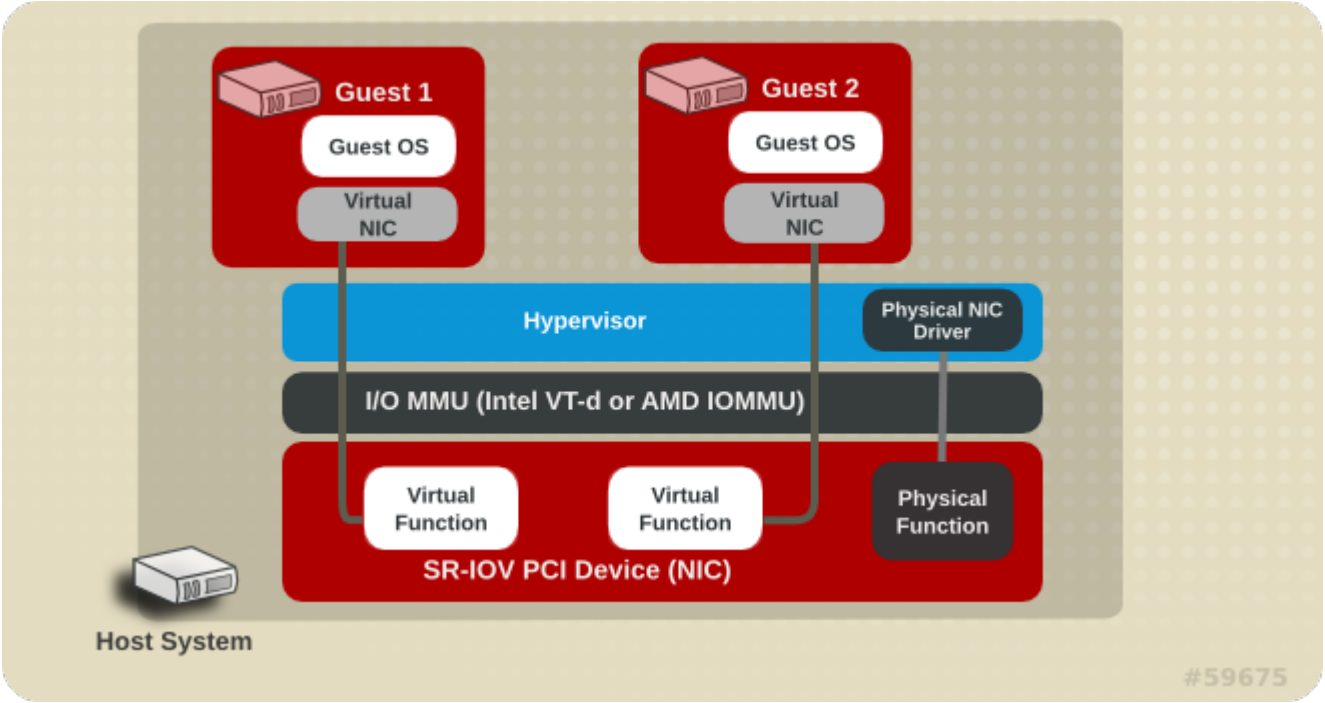
Add VFs

```
[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: 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
```

SR-IOV basics – Hypervisor view



oVirt

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.

oVirt

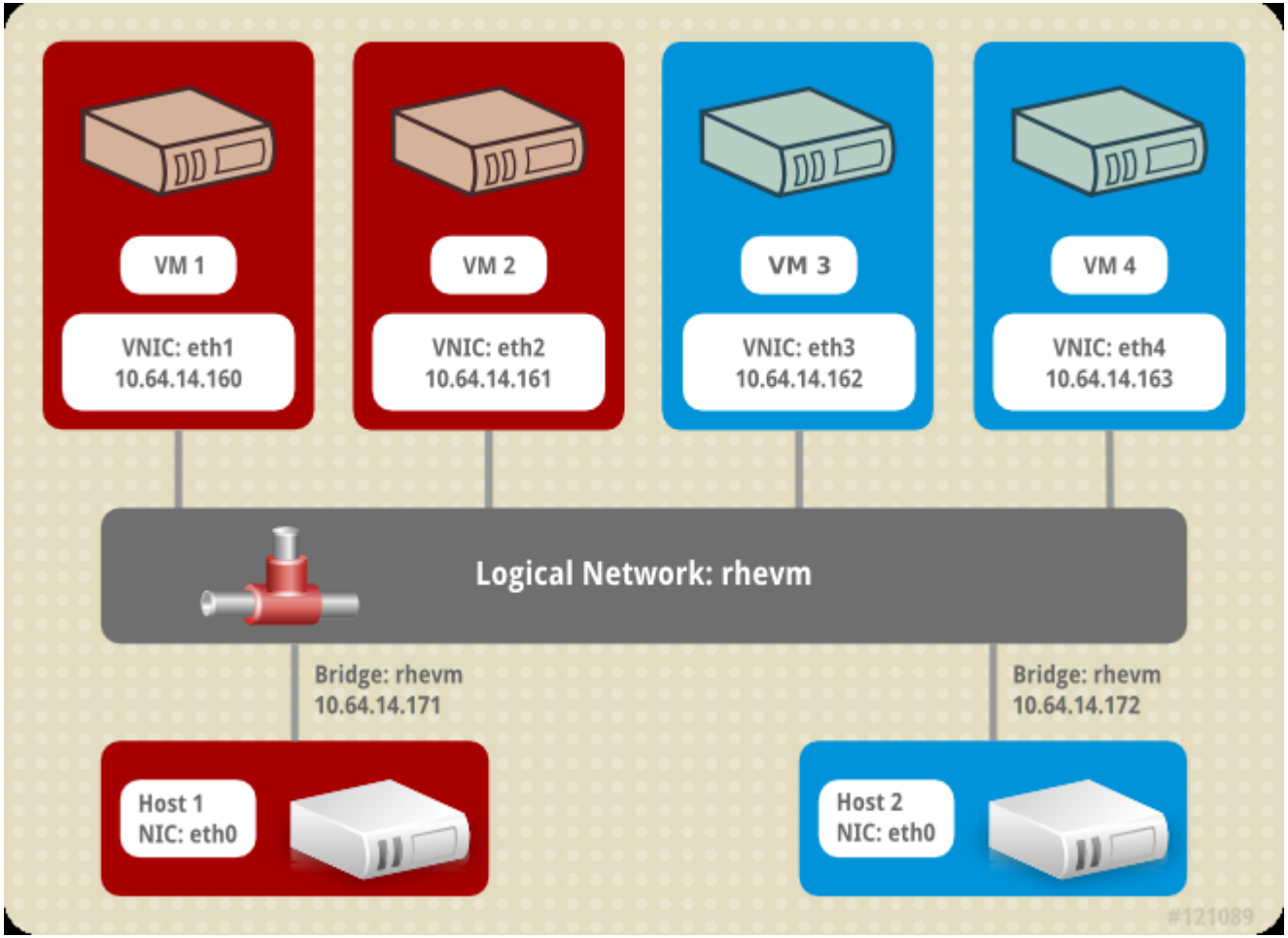
SR-IOV basics - Requirements

- 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).

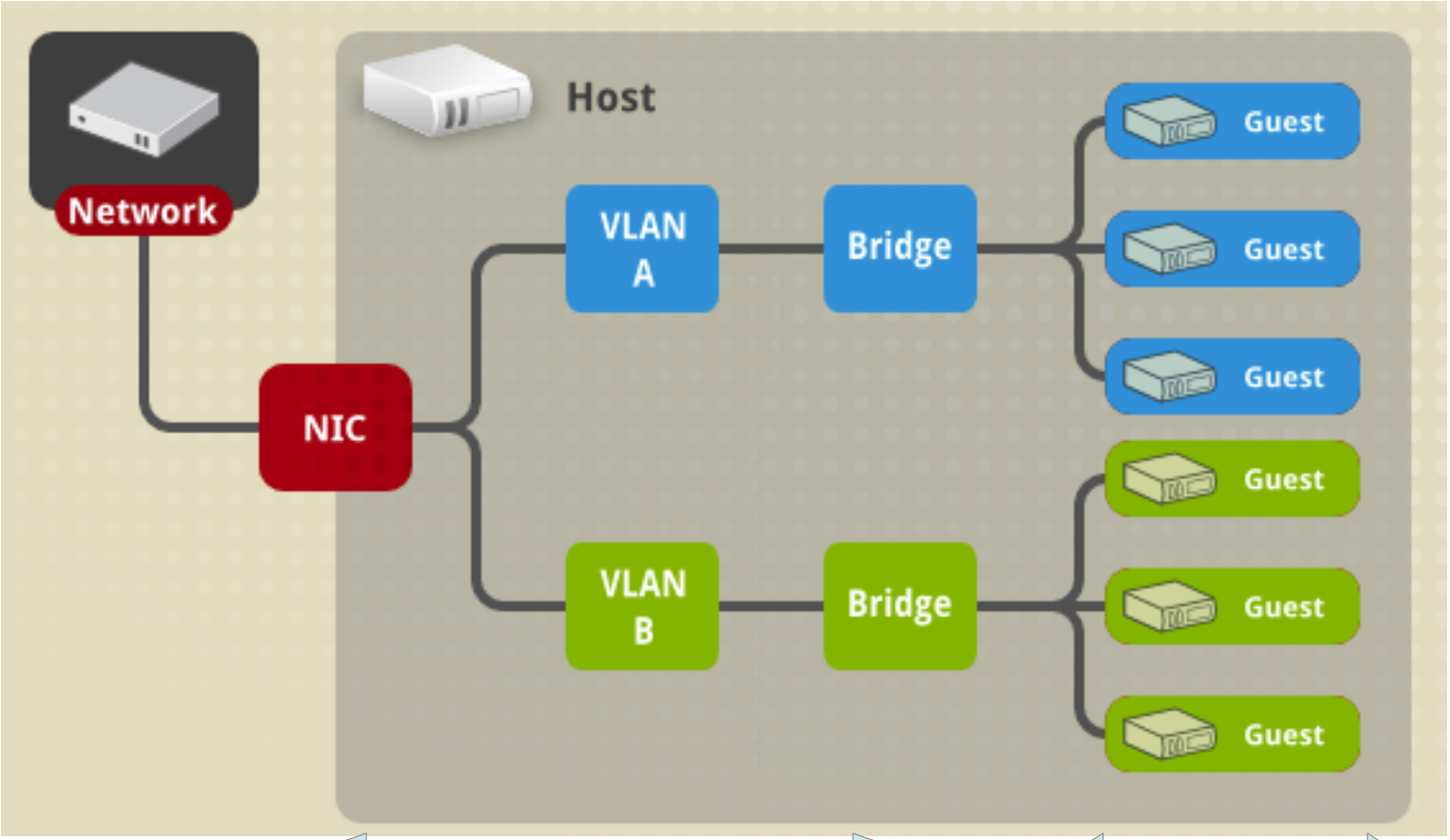
oVirt Networking

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

Logical Network



Host Networks setup



Setup Networks

vNIC

Attaching VM to a new network demo

https://10.35.161.15/ovirt-engine/webadmin/WebAdmin.html?locale=en_US#networks

oVirt OPEN VIRTUALIZATION MANAGER

admin | Configure | Guide | About | Feedback

Network:

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

New Import Edit Remove

1-13

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

Alerts (8) | Events | Tasks (10)

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.lab.tlv.redhat.com

2015-Aug-06, 12:15 Interface nic1 (VirtIO) was added to VM vm_6_. (User: admin@internal)

OVirt & SR-IOV

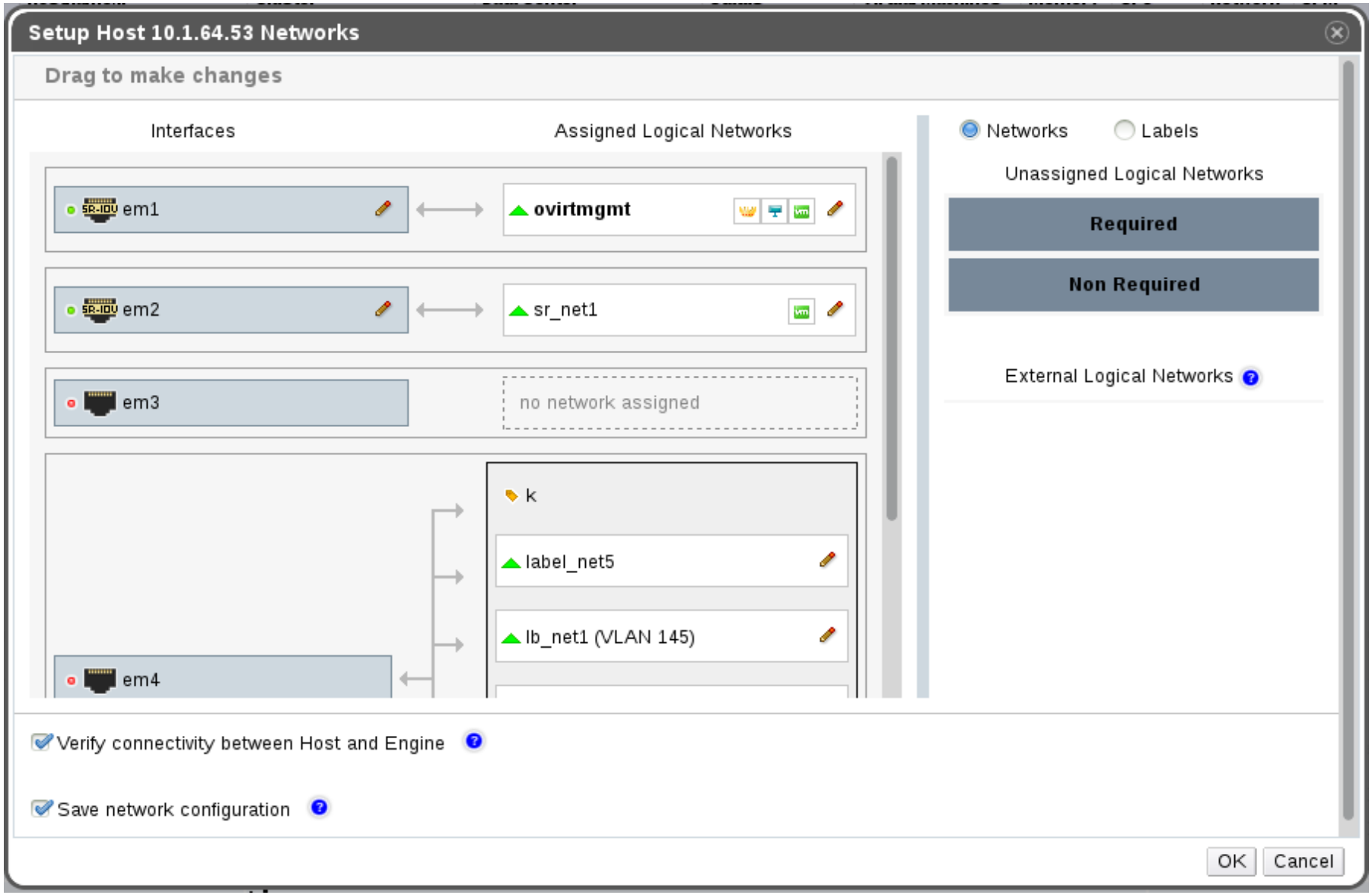
The problem:

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

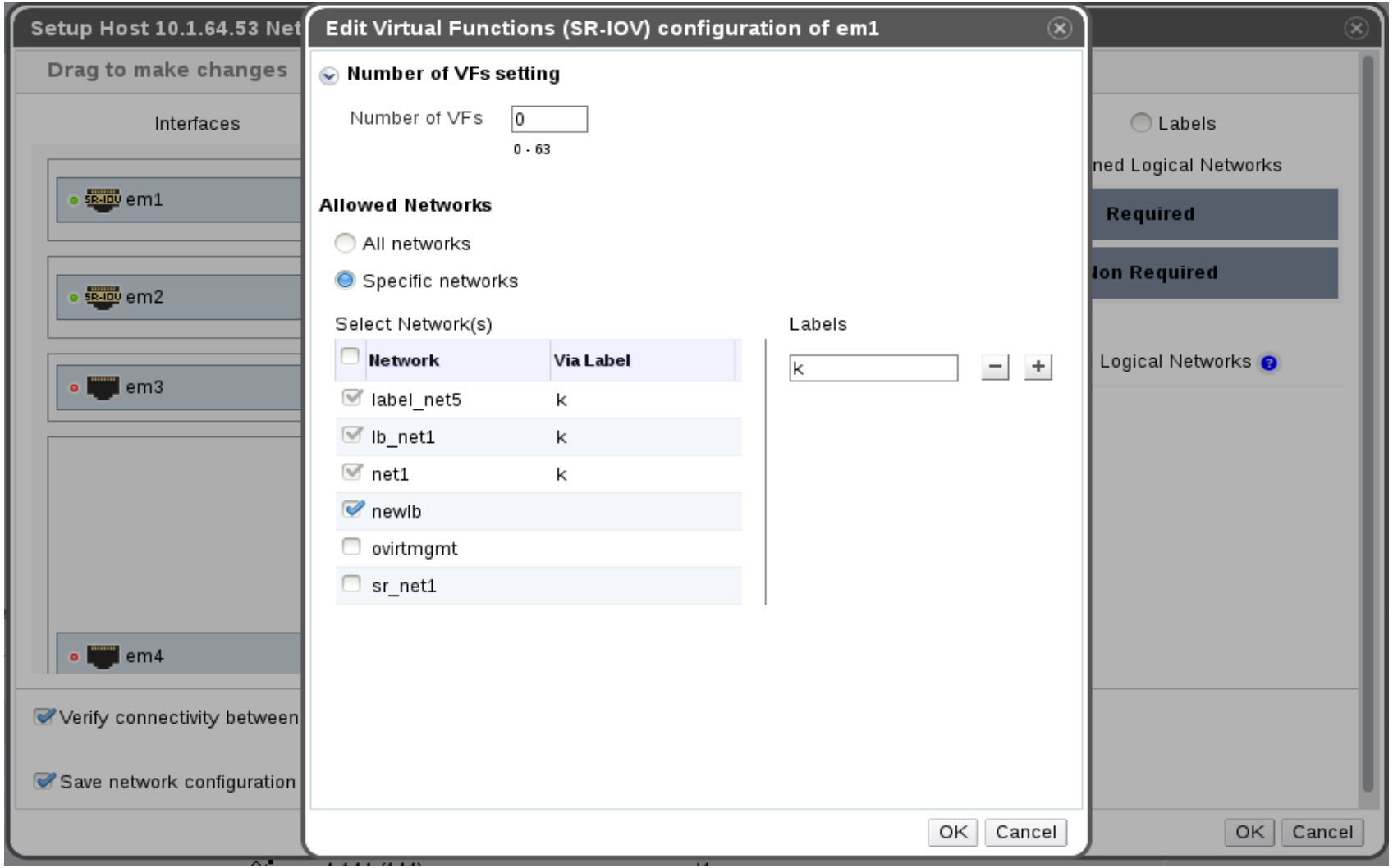
The solution:

- 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

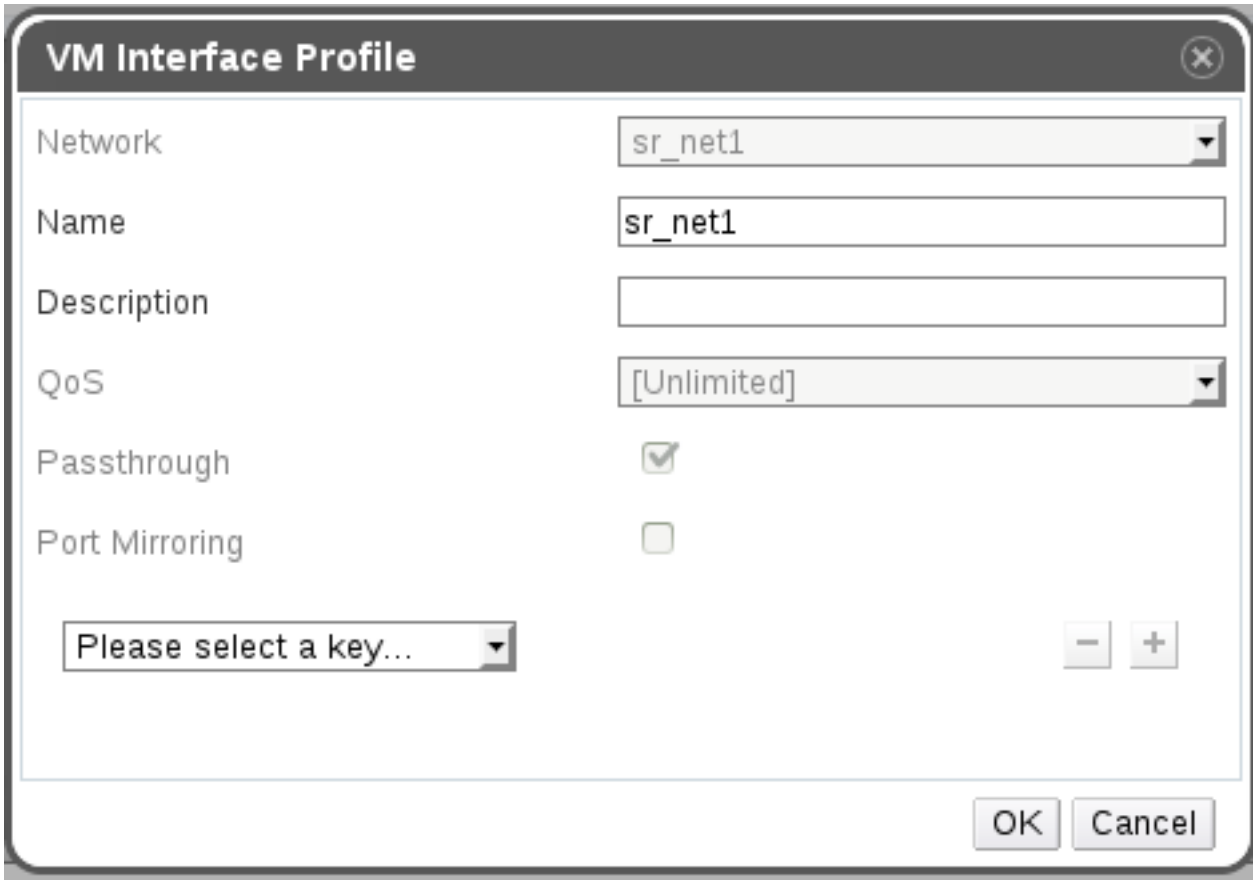
Setup Networks



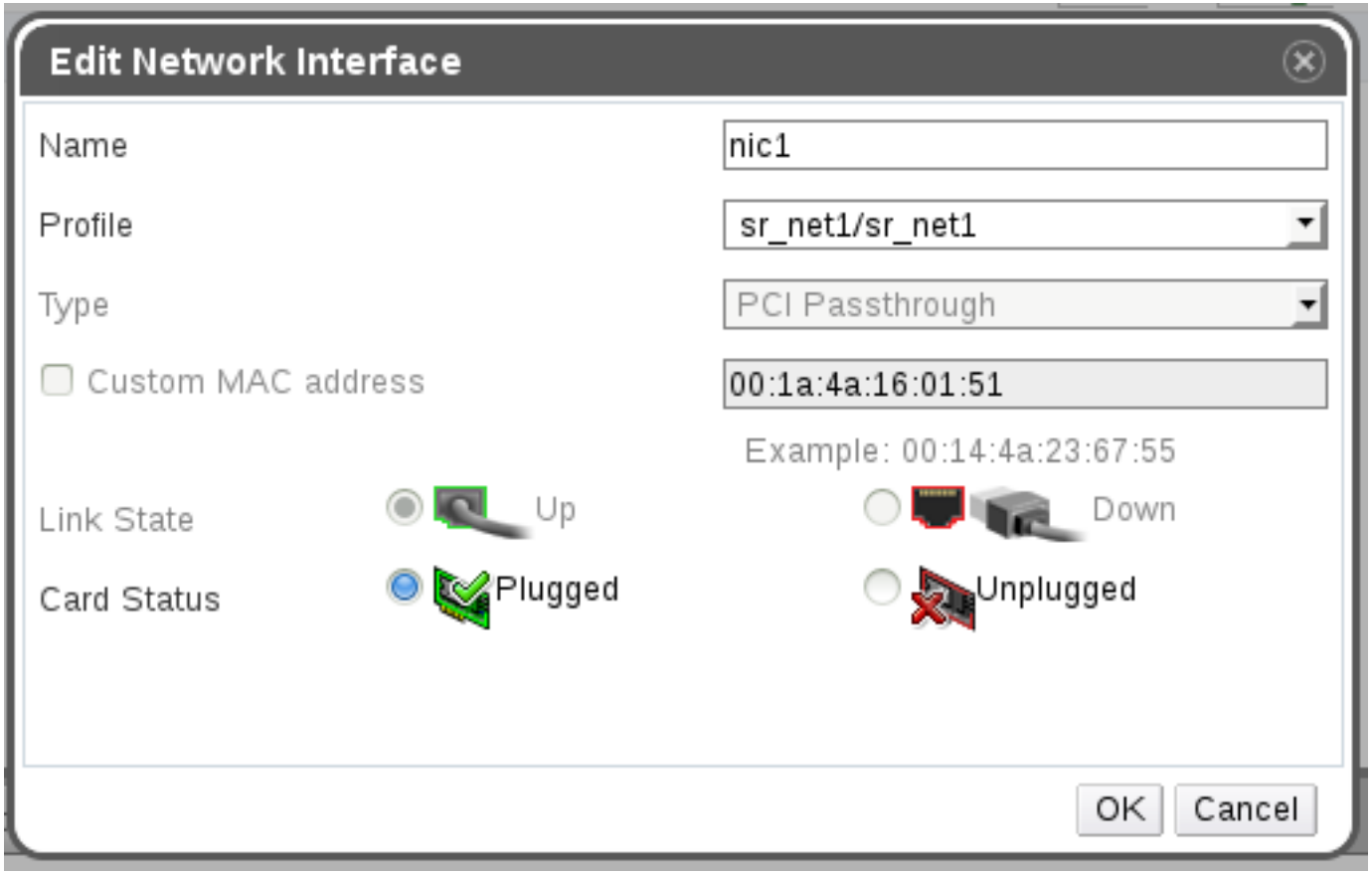
VFs configuration



Passthrough VM Interface profile



Passthrough VM Interface



Run VM with passthrough vNic

oVirt OPEN VIRTUALIZATION MANAGER

admin | Configure | Guide | About | Feedback

Network:

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

New Import Edit Remove

1-14

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)

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 (enableing 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