MIRANTIS
Pure Play OpenStack®

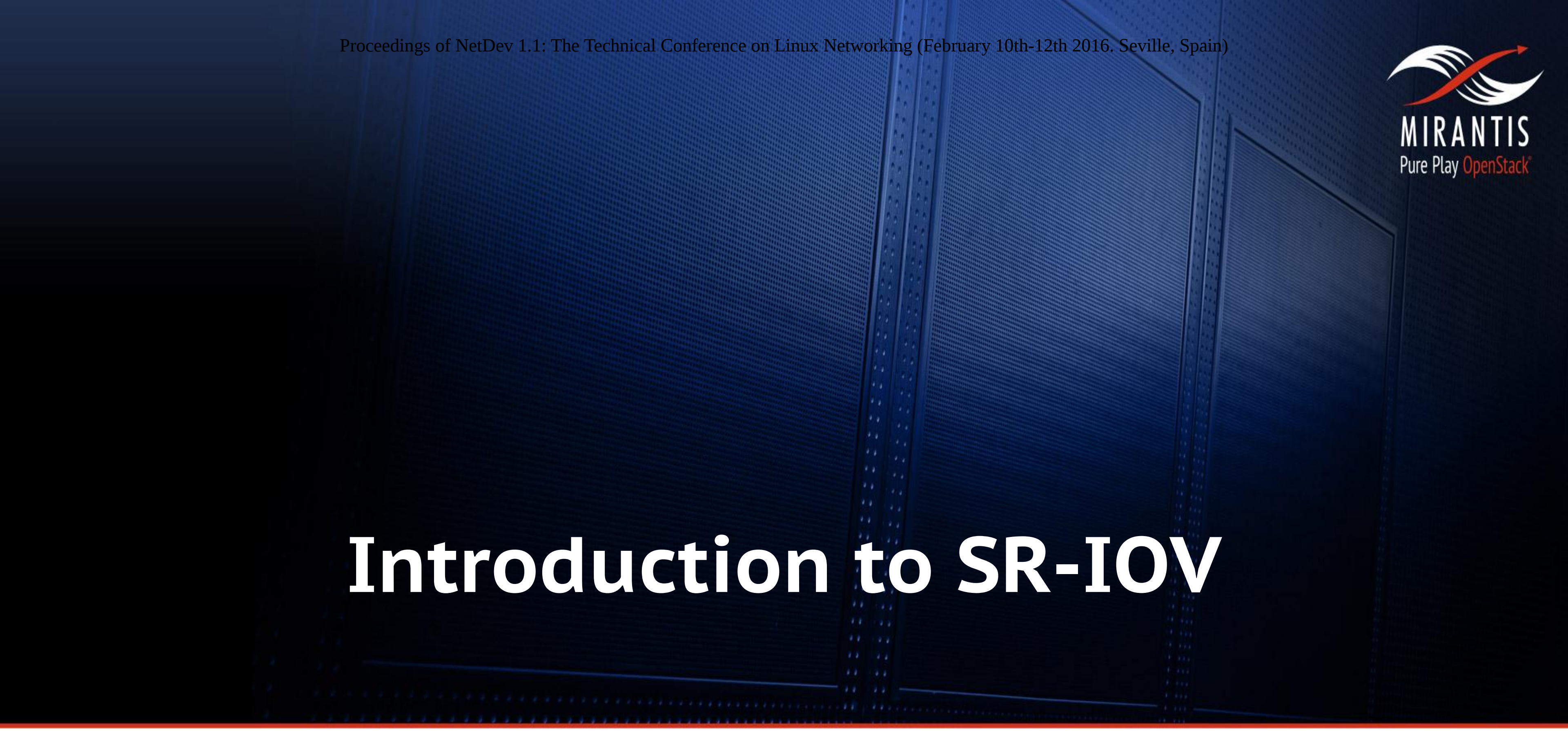
# Using SR-IOV on OpenStack

Alexander Duyck

www.mirantis.com

# Agenda

- Introduction to SR-IOV
  - What is SR-IOV
  - A History of SR-IOV on Linux
  - The Limitations of SR-IOV
- SR-IOV on OpenStack Demo
  - SR-IOV System Setup
  - SR-IOV DevStack setup
- The Future of SR-IOV
  - Hotplug
  - Live Migration
  - PF Promiscuous Mode



# Introduction to SR-IOV

SR-IOV 101

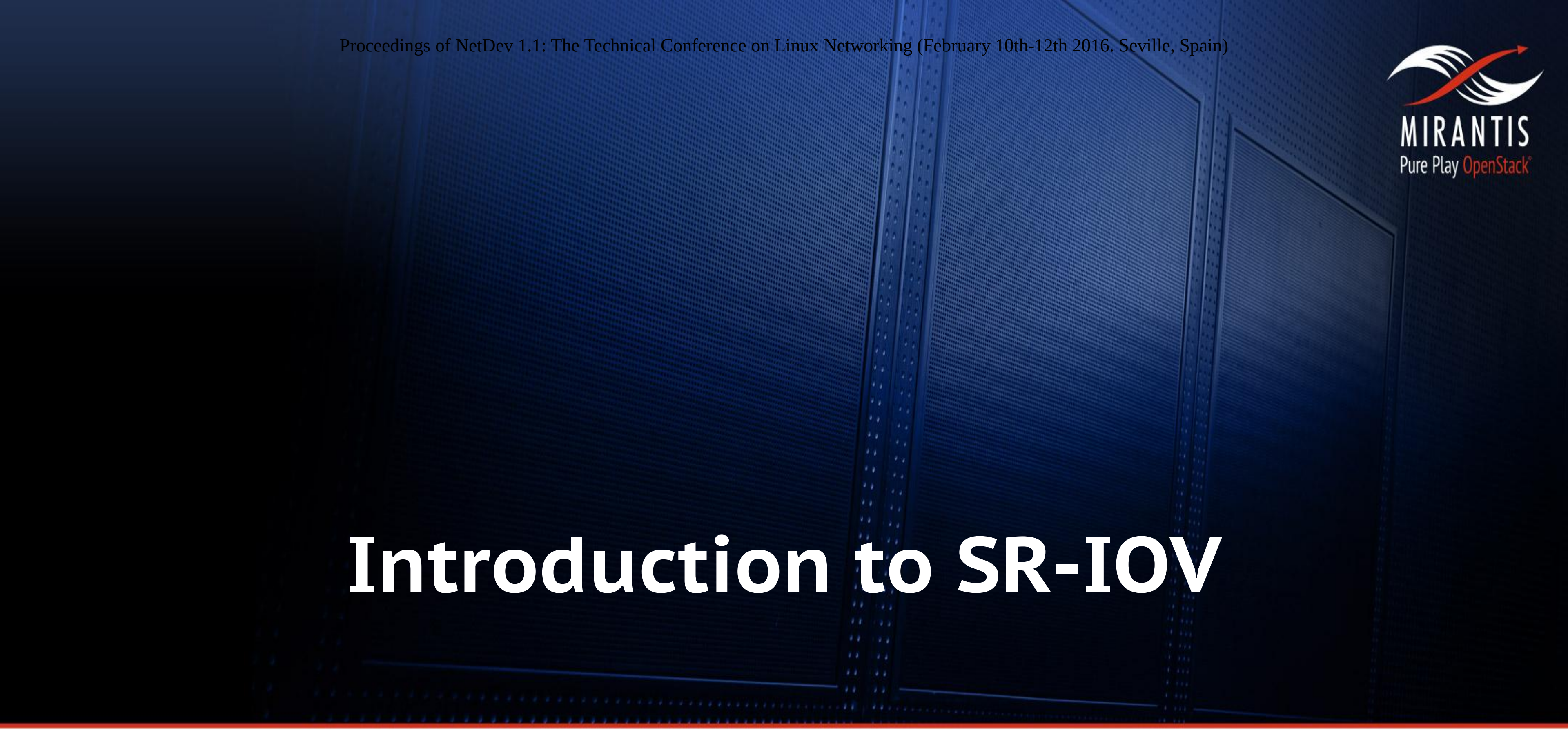
# SR-IOV 101

- What is SR-IOV?
  - In network virtualization, a single root input/output virtualization or SR-IOV is a network interface that allows the isolation of the PCI Express resources for manageability and performance reasons.

https://en.wikipedia.org/wiki/Single-root_IOV

# SR-IOV 101

- What SR-IOV is not
  - A networking specification
    - Nothing in the PCI specification mentions networking
    - Specification only defines PCIe messaging and configuration space
  - Direct Assignment
    - SR-IOV is still usable without an IOMMU
    - Containers don’t require direct assignment



# Introduction to SR-IOV

A History of SR-IOV on Linux

# A History of SR-IOV on Linux

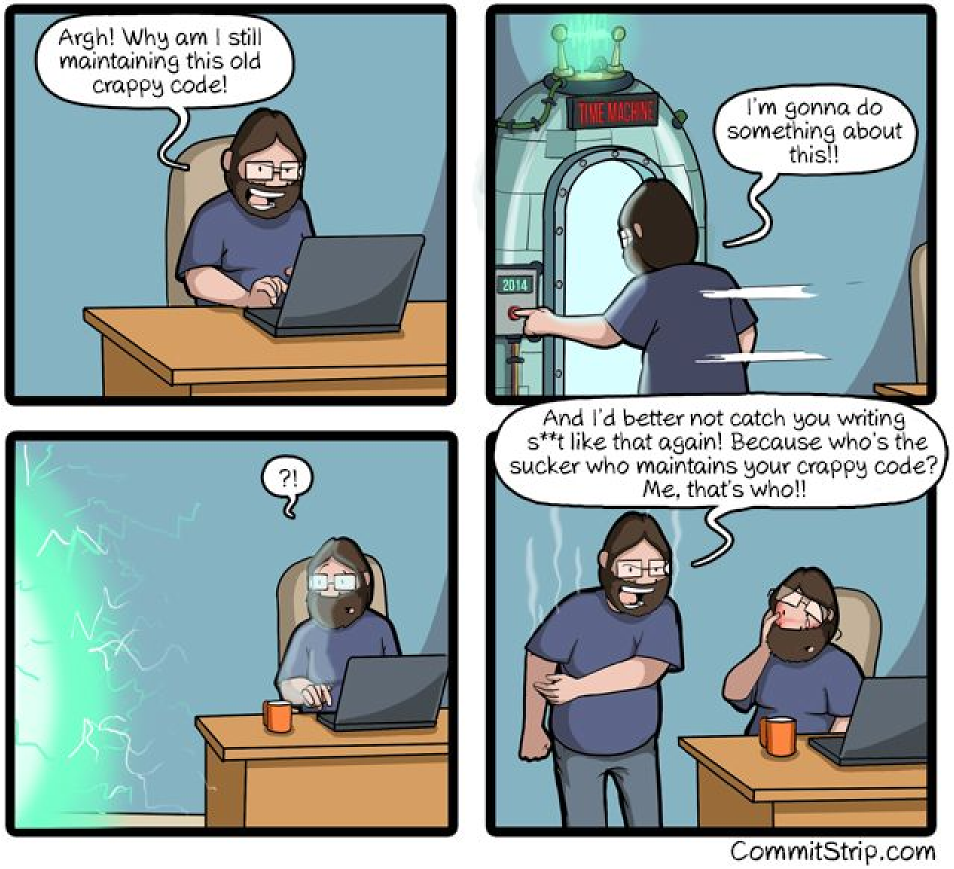
- First introduced in November 2008
  - "PCI: Linux kernel SR-IOV support"
    https://lkml.org/lkml/2008/11/21/357
  - Authored by Yu Zhao
  - Sysfs file for setting number of VFs
  - Example igb driver changes and igbvf tarball based on e1000e
- Version 12 finally accepted in March 2009
  - https://lkml.org/lkml/2009/3/19/509
  - Kernel version 2.6.30
  - sysfs file for num_vfs removed, left to driver to implement
  - igbvf and igb driver changes pushed off into separate patches
- igbvf driver and igb patches accepted April 2009
  - Added as a single patch with over 5000 lines
  - max_vfs module parameter added to igb

# A History of SR-IOV on Linux

- Other changes following the introduction of SR-IOV
  - 3.8
    - Support was added for using sysfs sriov_numvfs
  - 4.4
    - We have at least 18 drivers that support SR-IOV from 11 different vendors.

# A History of SR-IOV on Linux

- I was the one who submitted the patch for igbvf


Argh! Why am I still maintaining this old crappy code!
TIME MACHINE
2014
I'm gonna do something about this!!
?!
And I'd better not catch you writing s**t like that again! Because who's the sucker who maintains your crappy code? Me, that's who!!
CommitStrip.com

http://www.commitstrip.com/en/2016/01/18/what-idiot-wrote-this-code/

MIRANTIS
Pure Play OpenStack®

# Introduction to SR-IOV

## The Limitations of SR-IOV

# The Limitations of SR-IOV

- SR-IOV requires VF to rely on PF
  - Changes on PF can force VF driver to need to be reloaded

```
i40e 0000:02:00.0: Reload the VF driver to make this change effective.
ixgbe 0000:03:00.0: Reload the VF driver to make this change effective.
igb 0000:04:00.0: Reload the VF driver to make this change effective.
```

  - Prior to sysfs option max_vfs would affect all PFs

```
modprobe igb max_vfs=7
```

- VFs are allocated all at once
  - Cannot add/remove one VF at a time without removing all VFs
  - Changes to number of VFs can affect function number of all VFs
- In many cases VF cannot do anything unless PF is up
  - If PF is not up then VF cannot pass traffic
  - If PF is not up VFs cannot determine their own MAC address

# The Limitations of SR-IOV

MIRANTIS
Pure Play OpenStack

- Most NICs don't support a true L2 switch
  - Promiscuous mode is either disabled or mirrors outgoing traffic
  - Instead of learning MAC addresses they must be told

```
bridge fdb add 00:de:ad:be:ef:00 eth2
```

  - Limited support for adding MAC addresses
    - Intel I350 only supports 16 unicast addresses
    - Intel 82599 and X540 only support 128
  - Limited support for VLAN filtering
    - igb and ixgbe promiscuous mode didn't support VLAN trunking
  - Promiscuous mode cannot be supported on VFs
    - Replication bandwidth could easily exceed PCIe bandwidth

MIRANTIS
Pure Play OpenStack®

# SR-IOV on OpenStack Demo

## SR-IOV System Setup

# SR-IOV System Setup

- System contains
  - Single socket Core i7 4930K
  - Intel X540 dual port 10Gb Ethernet NIC (8086:1528)
  - Stock CentOS 7 system with latest updates
- Enable IOMMU
  - IOMMU(VTd)

```
# dmesg | grep IOMMU | grep enabled
[    0.000000] Intel-IOMMU: enabled
```

    - Enabled with kernel parameter “intel_iommu=on”
    - Use kernel parameter “iommu=pt” for better performance on host

# SR-IOV System Setup

- Enable Support for SR-IOV
  - SR-IOV Resource Allocation

    ```
    # lspci -vvv -d 8086:1528 | grep Region | grep ")$"
             Region 0: Memory at 00000000d2200000 (64-bit, non-prefetchable)
             Region 3: Memory at 00000000d2300000 (64-bit, non-prefetchable)
             Region 0: Memory at 00000000d2400000 (64-bit, non-prefetchable)
             Region 3: Memory at 00000000d2500000 (64-bit, non-prefetchable)
    ```

    - Resolved with kernel parameter “pci=realloc”
  - ARI

    ```
    # lspci -vvv -d 8086:1528 | grep ARIHierarchy
             IOVCtl:  Enable- Migration- Interrupt- MSE- ARIHierarchy+
             IOVCtl:  Enable- Migration- Interrupt- MSE- ARIHierarchy-
    ```

  - SR-IOV Bus Allocation

    ```
    # lspci -vt -n | grep 8086:1528
                 +-03.2-[03]--+-00.0  8086:1528
                 |            \-00.1  8086:1528
    ```

    - Resolved with kernel parameter “pci=assign-busses”

# SR-IOV System Setup

- Blacklist ixgbevf (optional)

```
# echo blacklist ixgbevf >> /etc/modprobe.d/blacklist.conf
```

- Load driver and configure interfaces

```
# modprobe ixgbe
# ip link set dev enp3s0f0 up
# ip link set dev enp3s0f1 up
# echo 7 > /sys/class/net/enp3s0f1/device/sriov_numvfs
# lspci -vt -n | grep 8086:1515
             |            +-10.1  8086:1515
             |            +-10.3  8086:1515
             |            +-10.5  8086:1515
             |            +-10.7  8086:1515
             |            +-11.1  8086:1515
             |            +-11.3  8086:1515
             |            \-11.5  8086:1515
```

# SR-IOV on OpenStack Demo

SR-IOV DevStack Setup

# DevStack SR-IOV Setup

- DevStack pull
  - https://git.openstack.org/openstack-dev/devstack
  - Ran “easy_install requests” to resolve install issue
- Create user for devstack

```
# cd devstack
# DEST=/opt/stack/
# tools/create-stack-user.sh
# sudo -i -u stack
$ git clone -l /srv/git/devstack
$ cd devstack
```

# DevStack SR-IOV Setup

- Edit local.conf
  - Disable Nova networking and enable Neutron w/ SR-IOV

```
## Services
disable_service n-net
## Neutron
ENABLED_SERVICES+=,q-svc,q-dhcp,q-meta,q-agt,q-sriov-agt
```

  - Setup tenant VLANs to support VFs

```
## Neutron Options
ENABLE_TENANT_VLANS=True
TENANT_VLAN_RANGE=3001:4000
PHYSICAL_NETWORK=physnet1
OVS_PHYSICAL_BRIDGE=br-enp3s0f0
PUBLIC_INTERFACE=enp3s0f0
Q_USE_PROVIDER_NETWORKING=True
Q_L3_ENABLED=False

## Neutron Networking options used to create Neutron Subnets
PROVIDER_NETWORK_TYPE="vlan"
SEGMENTATION_ID=2010
```

# DevStack SR-IOV Setup

- Edit local.conf (continued)
  - Enable ML2 plugin for Neutron

```
# ML2 Configuration
Q_PLUGIN=ml2
Q_ML2_PLUGIN_MECHANISM_DRIVERS=openvswitch,sriovnicswitch
Q_ML2_PLUGIN_TYPE_DRIVERS=vlan,flat,local
Q_ML2_TENANT_NETWORK_TYPE=vlan

# ML2 SR-IOV agent configuration
enable_plugin neutron git://git.openstack.org/openstack/neutron.git
PHYSICAL_DEVICE_MAPPINGS=physnet1:enp3s0f1

# ML2 plugin bits for SR-IOV enablement of Intel x540 NIC
[[post-config|/$Q_PLUGIN_CONF_FILE]]
[ml2_sriov]
supported_pci_vendor_devs = 8086:1528, 8086:1515
```

# DevStack SR-IOV Setup

- Edit local.conf (continued)
  - Add PCI passthru configuration for NOVA

```
# Add PCI Passthru filter, add alias, add all ports on PF
[[post-config|$NOVA_CONF]]
[DEFAULT]
scheduler_default_filters=RamFilter,ComputeFilter,AvailabilityZoneFilter,
ComputeCapabilitiesFilter,ImagePropertiesFilter,PciPassthroughFilter
pci_alias={\\"name\\":\\"x540vf\\",\\"product_id\\":\\"1515\\",\\"
vendor_id\\":\\"8086\\"}
pci_passthrough_whitelist={\\"devname\\":\\"enp3s0f1\\",\\"
physical_network\\":\\"physnet1\\"}
```

  - Start DevStack setup

```
$ ./stack.sh
```

# DevStack SR-IOV Setup

- Post-setup
  - Setup OpenStack credentials

    ```
    $ . ./openrc admin demo
    ```

  - Generate and assign login ssh key

    ```
    $ ssh-keygen -q -N "" -f ~/.ssh/id_rsa
    $ nova keypair-add --pub_key=/opt/stack/.ssh/id_rsa.pub stack-ssh
    ```

  - Update security rules to allow ssh and ping

    ```
    $ nova secgroup-add-rule default tcp 22 22 0.0.0.0/0
    $ nova secgroup-add-rule default icmp -1 -1 0.0.0.0/0
    ```

  - Create VF port

    ```
    $ neutron port-create physnet1 --vnic-type direct --name physnet1-vf1
    ```

# DevStack SR-IOV Setup

- Launch VMs with VF attached
  - Identify Neutron port ID of VF port

```
$ port_id=`neutron port-show physnet1-vf1 -F id -f value`
```

  - Start VM specifying that we want to use VF

```
$  nova boot --flavor m1.small --key_name stack-ssh \
>             --image Fedora-Cloud-Base-23-20151030.x86_64 \
>             --nic port-id=$port_id Instance-1
```

  - Check status of VMs

```
$ nova list
+--------------------------------------+------------+--------+------------+-------------+-------------------+
| ID                                   | Name       | Status | Task State | Power State | Networks          |
+--------------------------------------+------------+--------+------------+-------------+-------------------+
| fe972329-0666-4160-9950-e1f6f79146c5 | Instance-1 | ACTIVE | -          | Running     | physnet1=10.0.0.3 |
+--------------------------------------+------------+--------+------------+-------------+-------------------+
```

# DevStack SR-IOV Setup

- Log into VMs
  - Get namespace name

    ```
    $ ip netns
    qdhcp-19c44e2f-6abf-4f5e-bb93-5665fc9d0d9e
    ```

  - Change namespace

    ```
    $ sudo ip netns exec `ip netns | grep qdhcp` sudo -i -u stack
    ```

  - Log into VM

    ```
    $ ssh fedora@10.0.0.3
    Warning: Permanently added '10.0.0.3' (ECDSA) to the list of known hosts.
    [fedora@instance-1 ~]$
    ```

# The Future of SR-IOV

## Hotplug, Live Migration, and PF Promiscuous Mode

# Hotplug

- Currently supported by QEMU and Linux Kernel
- OpenStack has yet to implement
  - Requires cooperation between Nova and Neutron to create bindings

# Live Migration

- Live migration with a VF is still a work in progress
  - Currently not supported by QEMU or Kernel
  - Best solution so far consists of bonding or team w/ virtio interface
    - Requires VF be evicted before warm-up phase
    - Requires significant guest cooperation
    - Slows down VM during migration
  - DMA dirty page tracking
    - Requires significant guest cooperation
    - Yet to be implemented
  - Quiescing the device
    - Requires significant guest cooperation
    - Yet to be implemented

# PF Promiscuous Mode

- PF cannot support full promiscuous mode
  - bridge fdb add
    - Works as long as VLANs are setup correctly
      - Fixes submitted for igb and ixgbe
    - Resources limited as devices only have so many MAC filters
  - VLAN Trunking
    - Default behavior should be to support VLAN
    - Intel parts use bit array, all should support VLAN Trunking
    - More research needed for other parts

MIRANTIS
Pure Play OpenStack®

# Thank You

AlexanderDuyck@gmail.com

# Backup

# local.conf

```
[[local|localrc]]
HOST_IP=192.168.1.116
ADMIN_PASSWORD=nova
DATABASE_PASSWORD=$ADMIN_PASSWORD
RABBIT_PASSWORD=$ADMIN_PASSWORD
SERVICE_PASSWORD=$ADMIN_PASSWORD
SERVICE_TOKEN=$ADMIN_PASSWORD

# Update project repos
PIP_UPGRADE=True

# Services
disable_service n-net
disable_service zookeeper
# Neutron
ENABLED_SERVICES+=,q-svc,q-dhcp,q-meta,q-agt,q-sriov-agt

## Neutron Options
ENABLE_TENANT_VLANS=True
TENANT_VLAN_RANGE=3001:4000
PHYSICAL_NETWORK=physnet1
OVS_PHYSICAL_BRIDGE=br-enp3s0f0
PUBLIC_INTERFACE=enp3s0f0
Q_USE_PROVIDER_NETWORKING=True
Q_L3_ENABLED=False
IP_VERSION=4

## Neutron Networking options used to create Neutron Subnets
PROVIDER_NETWORK_TYPE="vlan"
SEGMENTATION_ID=2010
```

# local.conf (continued)

```
# ML2 Configuration
Q_PLUGIN=ml2
Q_ML2_PLUGIN_MECHANISM_DRIVERS=openvswitch,sriovnicswitch
Q_ML2_PLUGIN_TYPE_DRIVERS=vlan,flat,local
Q_ML2_TENANT_NETWORK_TYPE=vlan

# ML2 SR-IOV agent configuration
enable_plugin neutron git://git.openstack.org/openstack/neutron.git
PHYSICAL_DEVICE_MAPPINGS=physnet1:enp3s0f1

# Default Fedora 23 image
IMAGE_URLS+="https://download.fedoraproject.org/pub/fedora/linux/releases/23/Cloud/x86_64/Images/Fedora-Cloud-Base-23-20151030.x86_64.qcow2"

# Add PCI Passthru filter, add alias, add all ports on PF
[[post-config|$NOVA_CONF]]
[DEFAULT]
scheduler_default_filters=RamFilter,ComputeFilter,AvailabilityZoneFilter,ComputeCapabilitiesFilter,ImagePropertiesFilter,PciPassthroughFilter
pci_alias={\\"name\\":\\"x540vf\\",\\"product_id\\":\\"1515\\",\\"vendor_id\\":\\"8086\\"}
pci_passthrough_whitelist={\\"devname\\":\\"enp3s0f1\\",\\"physical_network\\":\\"physnet1\\"}

# ML2 plugin bits for SR-IOV enablement of Intel x540 NIC
[[post-config|/$Q_PLUGIN_CONF_FILE]]
[ml2_sriov]
supported_pci_vendor_devs = 8086:1528, 8086:1515
```

# local.sh

```
#!/bin/bash

# Add default key for admin and demo
nova keypair-add --pub_key=/opt/stack/.ssh/id_rsa.pub stack-ssh

# Enable ping and ssh
for i in admin demo
do
  nova --os-project-name $i secgroup-add-rule default \
       tcp 22 22 0.0.0.0/0
  nova --os-project-name $i secgroup-add-rule default \
       icmp -1 -1 0.0.0.0/0
done

# Add host nameserver to provider_net
ns=`grep nameserver /etc/resolv.conf | head -n1 | awk '{print $2}'`
neutron --os-project-name demo subnet-update \
        --dns-nameserver $ns provider_net

# Create 7 VF ports
for i in `seq 1 7`
do
  neutron --os-project-name demo port-create physnet1 \
          --vnic-type direct --name physnet1-vf$i
done
```

# start-vms.sh

```
#!/bin/bash
set -x

. /opt/stack/devstack/openrc admin demo

# Get ID for physnet to later create OVS ports
net_id=`neutron net-show physnet1 -F id -f value`

# One instance, with one vNIC port on OVS w/ VLAN
nova boot --flavor m1.small --key_name stack-ssh \
          --image Fedora-Cloud-Base-23-20151030.x86_64 \
          --nic net-id=$net_id Instance-0

# Two instances, each with one VF port
for i in 1 2
do
  port_id=`neutron port-show physnet1-vf$i -F id -f value`
  nova boot --flavor m1.small --key_name stack-ssh \
            --image Fedora-Cloud-Base-23-20151030.x86_64 \
            --nic port-id=$port_id Instance-$i
done
```