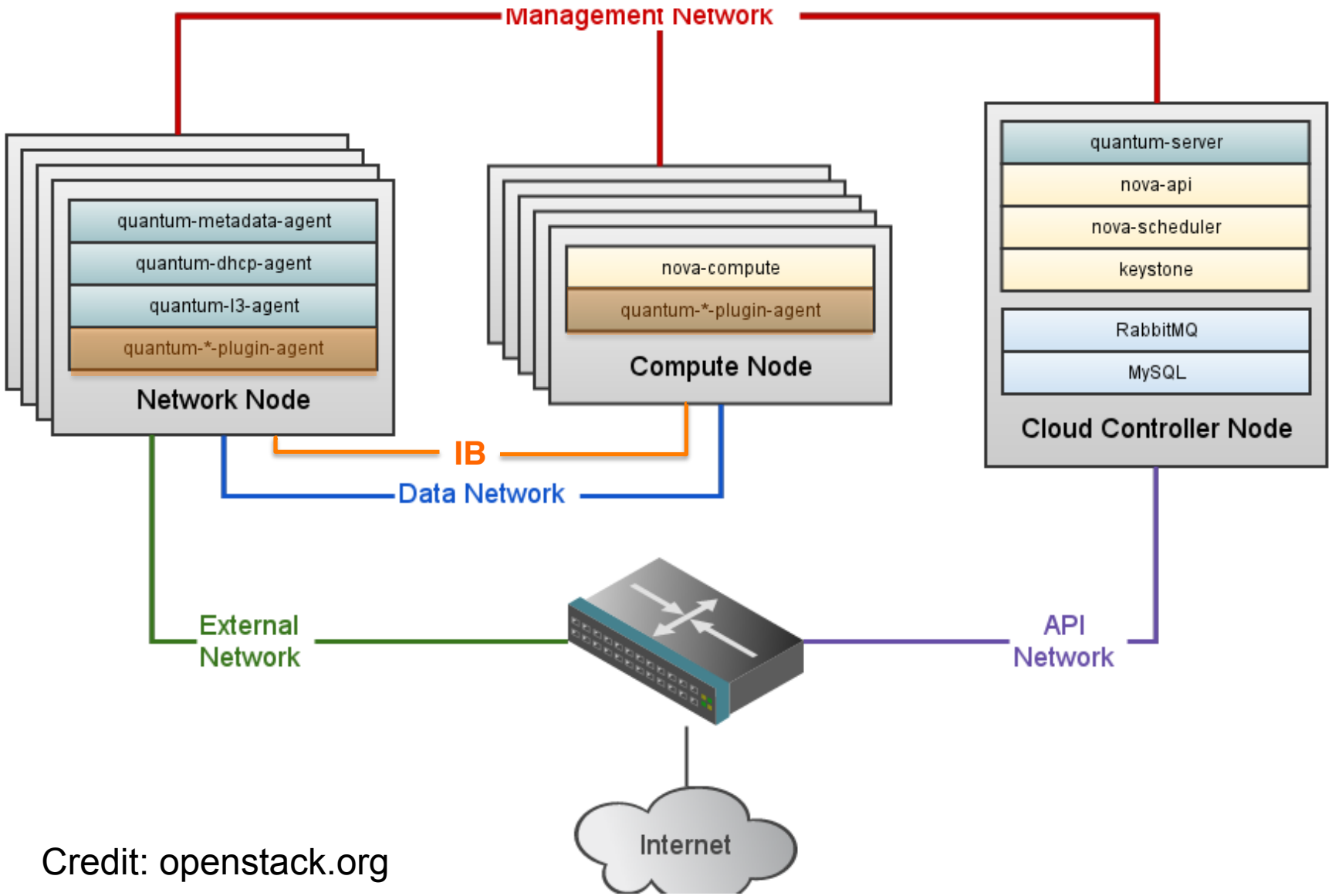
OpenStack and IB

Blake Caldwell
OFA Users Workshop
April 3, 2014

U.S. DEPARTMENT OF ENERGY
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

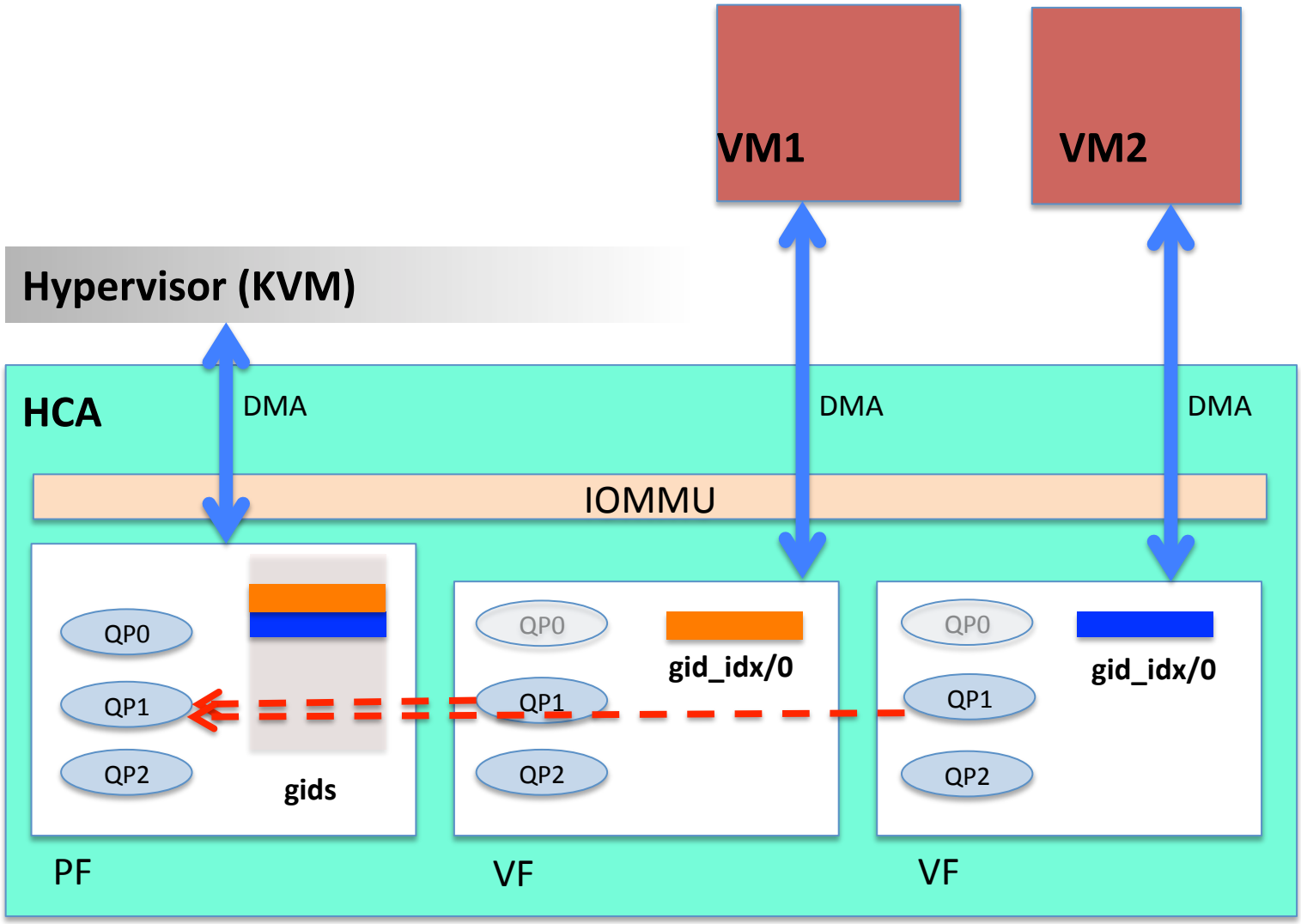
- Background
- Partitioning with P-keys
- SR-IOV complexities
- Configuration

Background: OpenStack Architecture



Credit: openstack.org

Background: SR-IOV



Background: SR-IOV

- QP0 on VF is non-functional, only on PF
- QP1 on VF is proxied through PF
- RID tags traffic for IOMMU translation (DMA)
- VF p-key and gid tables index into PF tables
- Configuration of P-keys through sysfs

P-Keys and VFs

PF (00:41:00.0)

/sys/class/infiniband/mlx4_0/iov/ports/2/pkeys

Index	Pkey
0	0xffff
1	0xb000
2	0xb030

VF1 (00:41:00.1)

/sys/class/infiniband/mlx4_0/iov/**0000:41:00.1**/ports/2/pkey_idx

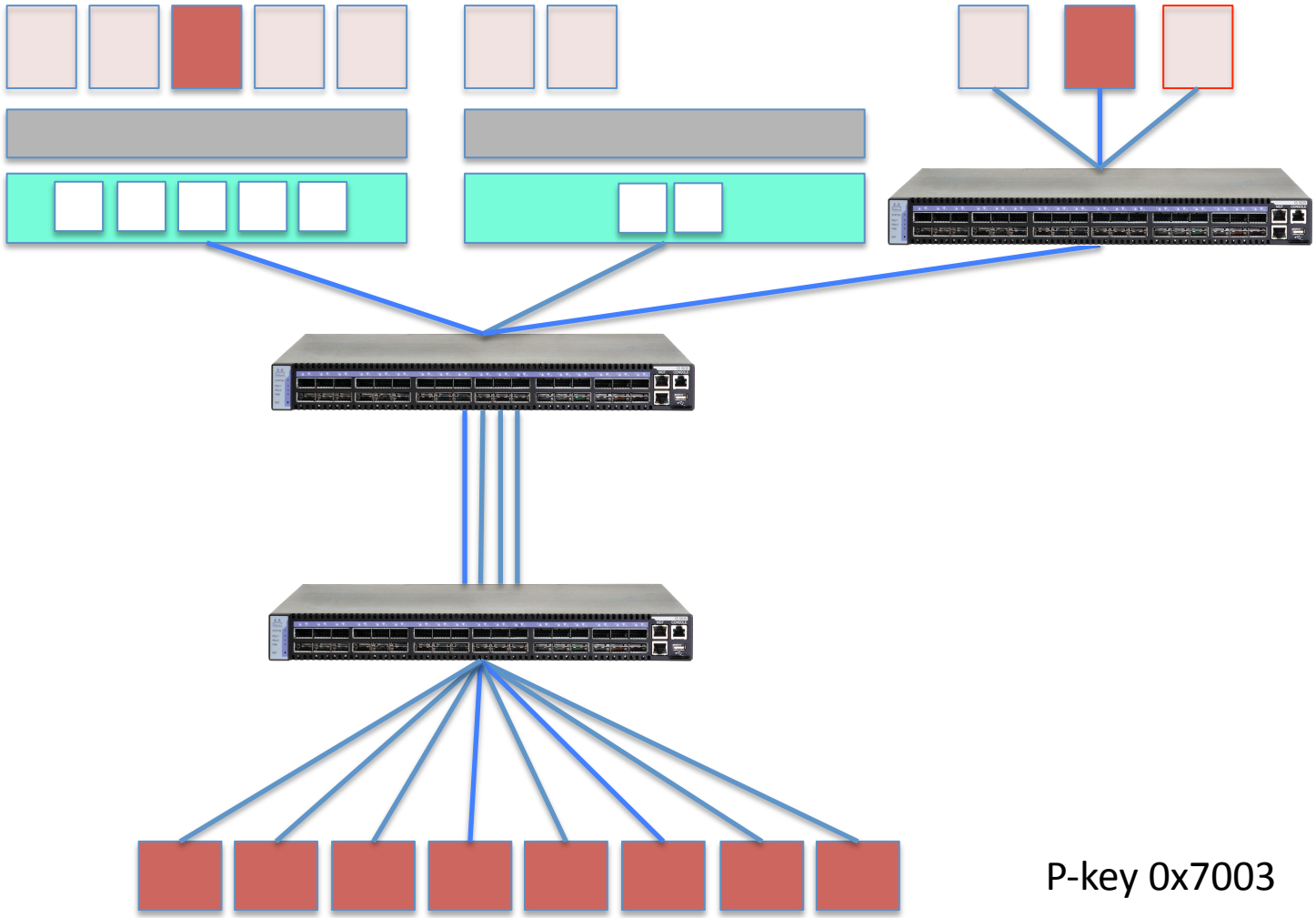
Index	Pkey
0	1
1	0

VF2 (00:41:00.2)

/sys/class/infiniband/mlx4_0/iov/**0000:41:00.2**/ports/2/pkey_idx

Index	Pkey_idex
0	2
1	0

Fabric Partitioning

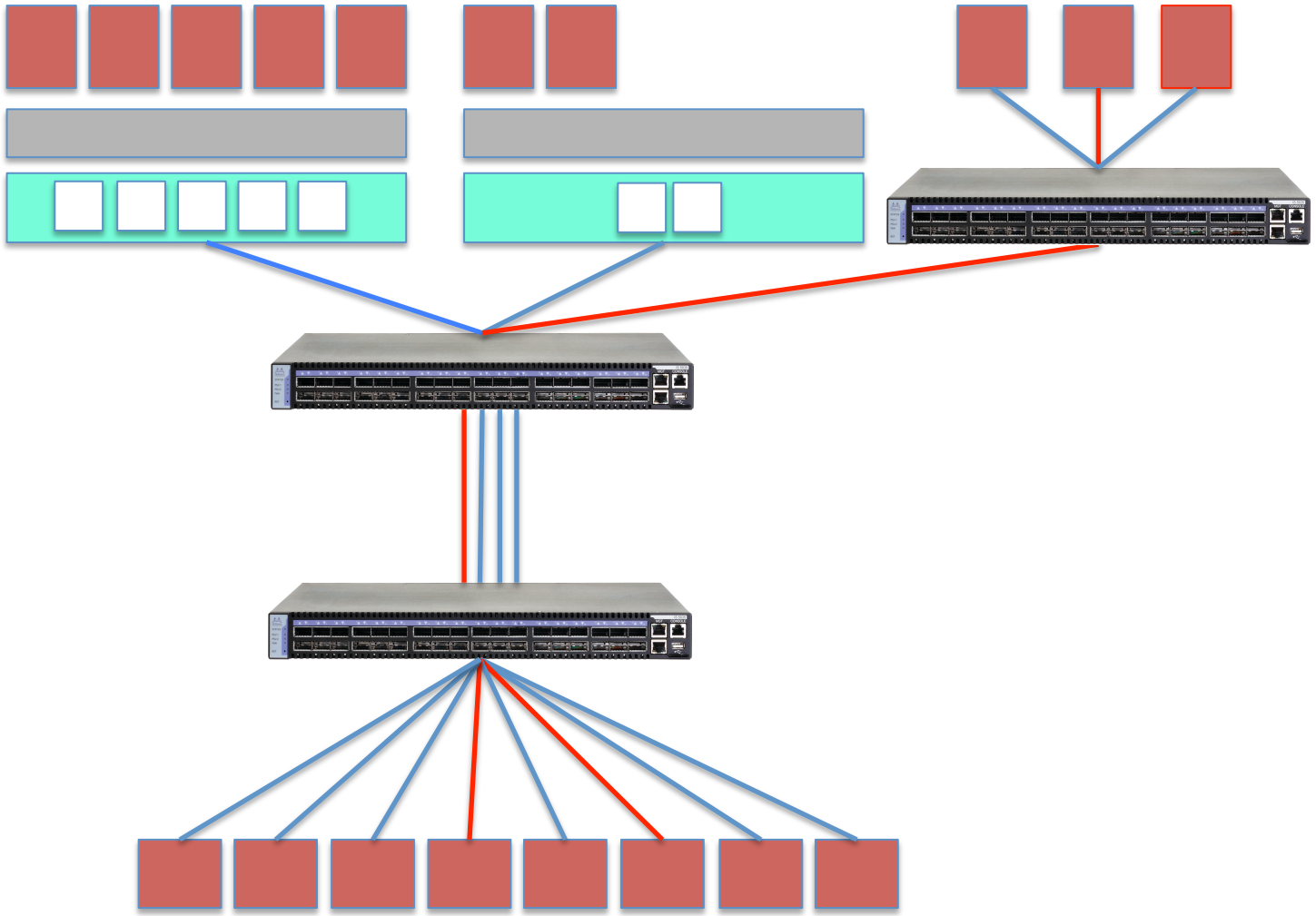


P-key 0x7003

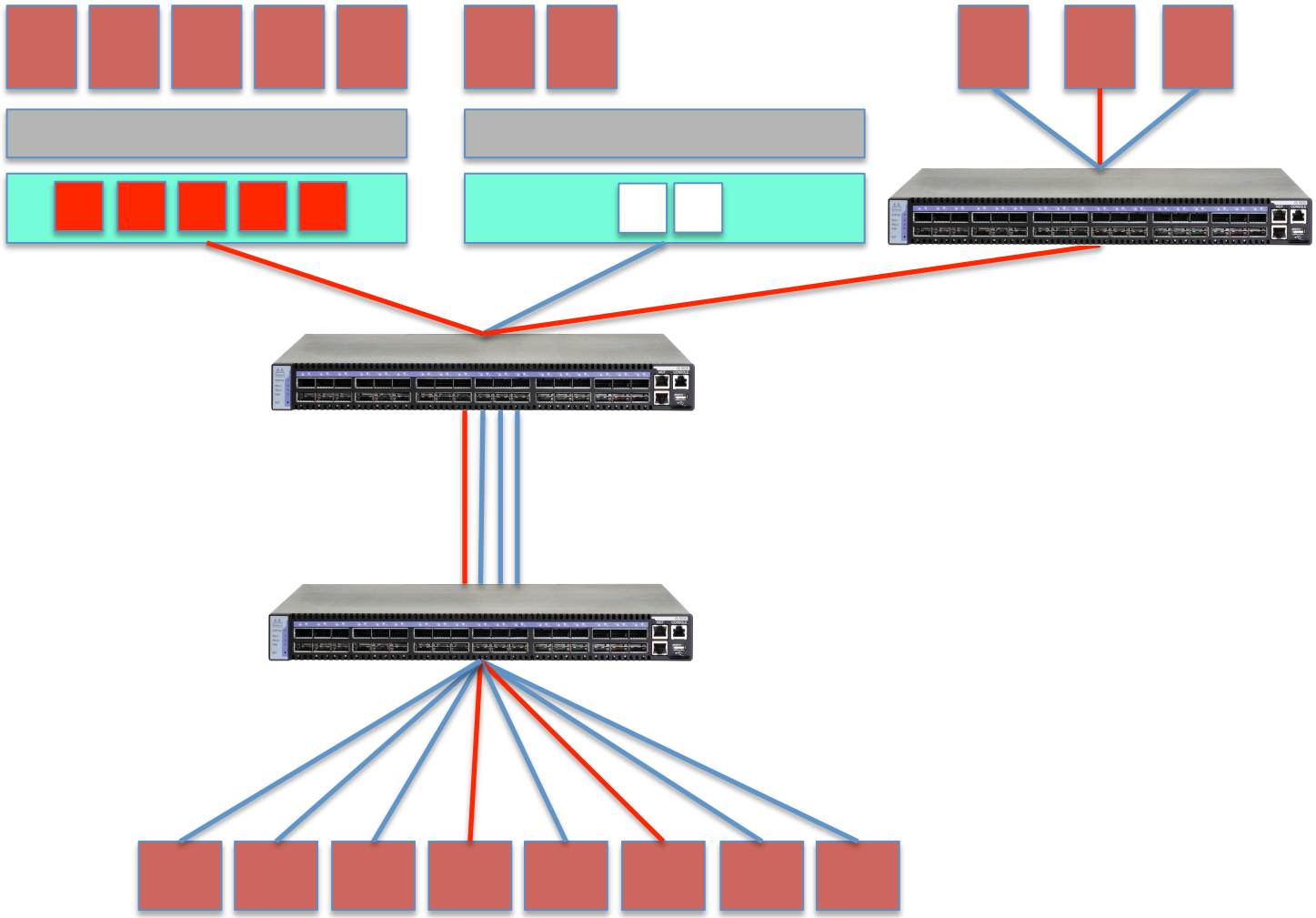
Complexities with SR-IOV

- Still have shared resources
- How to administer vHCAs (tools don't work)
- Increasing functionality embedded within HCAs
- Routing virtualized topologies

Routing with Virtualization



Routing with Virtualization



Base SR-IOV Configuration

- **Add SR-IOV config options in firmware**
  - **ConnectX-2 (2.9.1200 to get bug fix for FLR)**
  - **ConnectX-3**

```
# mstflint -dev 82:00.0 dc
[HCA]
num_pfs = 1
total_vfs = <0-63>
sriov_en = true
```

- **Check BIOS settings**
- **Kernel**
  - `CONFIG_DMAR_DEFAULT_ON=y` OR Intel/AMD specific kernel cmdline options
- **Modprobe parameters**

```
options mlx4_core port_type_array=2,1 num_vfs=16 probe_vf=1

Options mlx4_ib sm_guid_assign=0
```

OpenSM Configuration

- partitions.conf

```
management=0x7fff,ipoib, sl=0, defmember=full : ALL, ALL_SWITCHES=full,SELF=full;
vlan1=0x1, ipoib, sl=0, defmember=full : ALL;
vlan2=0x2, ipoib, sl=0, defmember=full : ALL;
vlan3=0x3, ipoib, sl=0, defmember=full : ALL;
```

- opensm.conf

```
allow_both_pkeys TRUE
```

OpenStack Configuration

- Compute node
  - Select Mellanox VIF driver
  - Optionally add PCI device to pci_passthrough_whitelist
- Configure plugin (compute and network nodes)
  - Add plugin to network node and compute node
  - Define vlan range (see partitions.conf)
  - **vnic-type:** hostdev | macvtap | virtio | bridge
- Define neutron port for SR-IOV device
- Launch instances with newly created nic port

```
$ nova boot --flavor m1.large --image rh6.5_mlnx_ofed \
   --nic port-id=a43d35f3-3870-4ae1-9a9d-d2d341b693d6 sriov_instance
```

Other Features

- Expose different interface types to VMs
  - With kernel modules: EoIB/IPoIB/RoCE
  - Paravirtualized interface (eIPoIB bridge)
- QoS at VM granularity
- Storage plugins (Cinder service)
  - iSER plugin from Mellanox

Questions?

blakec@ornl.gov