SRIOV & POWER

Wei Yang

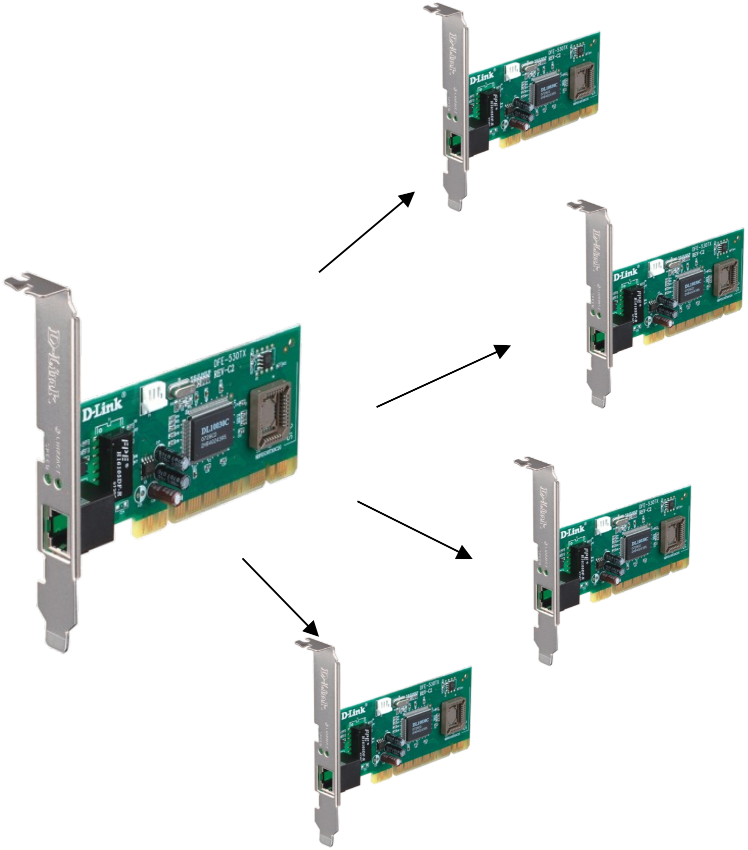
Linux Technology Center, IBM, China
ywywyang@cn.ibm.com

Oct, 19, 2014

Agenda

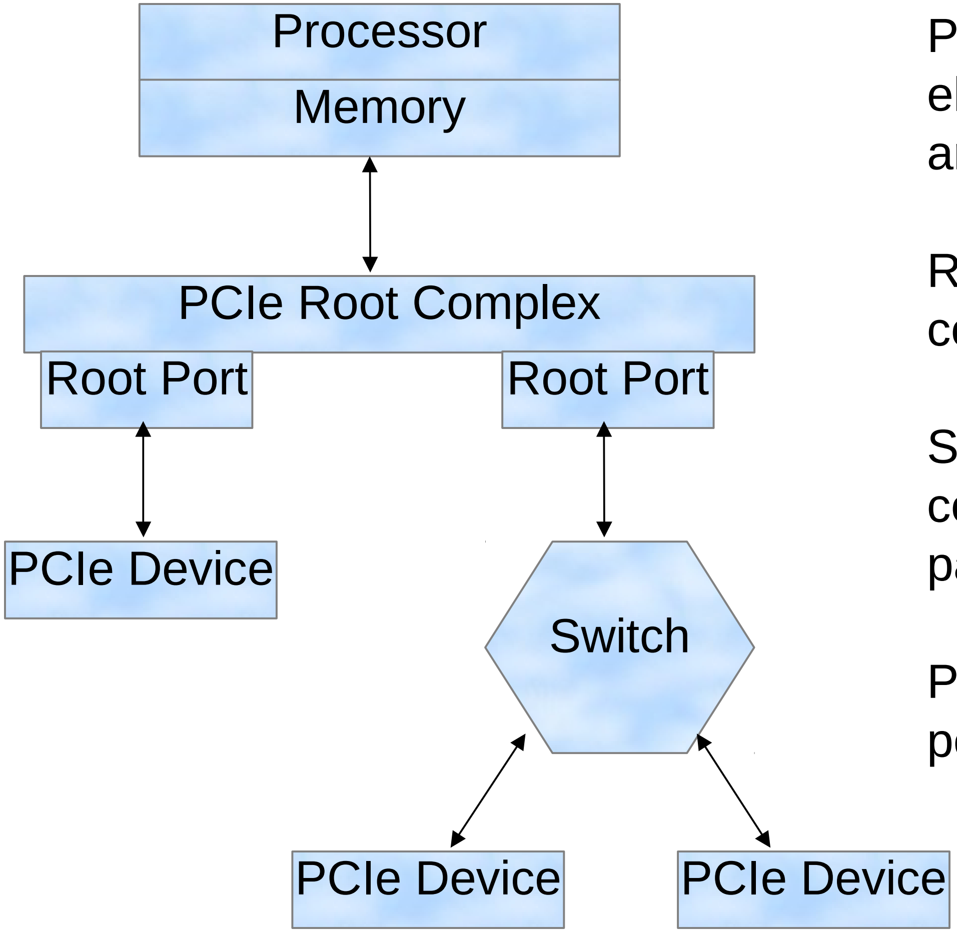
- **What's SRIOV?**
  - **PCIe Subsystem**
  - **PCIe device**
  - **SRIOV device**
- **I/O virtualization**
  - **Emulation in user space**
  - **Virtio**
  - **vhost-net**
  - **SRIOV**
- **SRIOV on POWER**
  - **Restrictions**
  - **Solution**
- **Future Work**

What's SRIOv?



- SRIOV: Single Root I/O Virtualization.
- *PCI-SIG standard*
- Provides high throughput, low CPU utilization, high scalability
- Request platform support

PCIe Subsystem



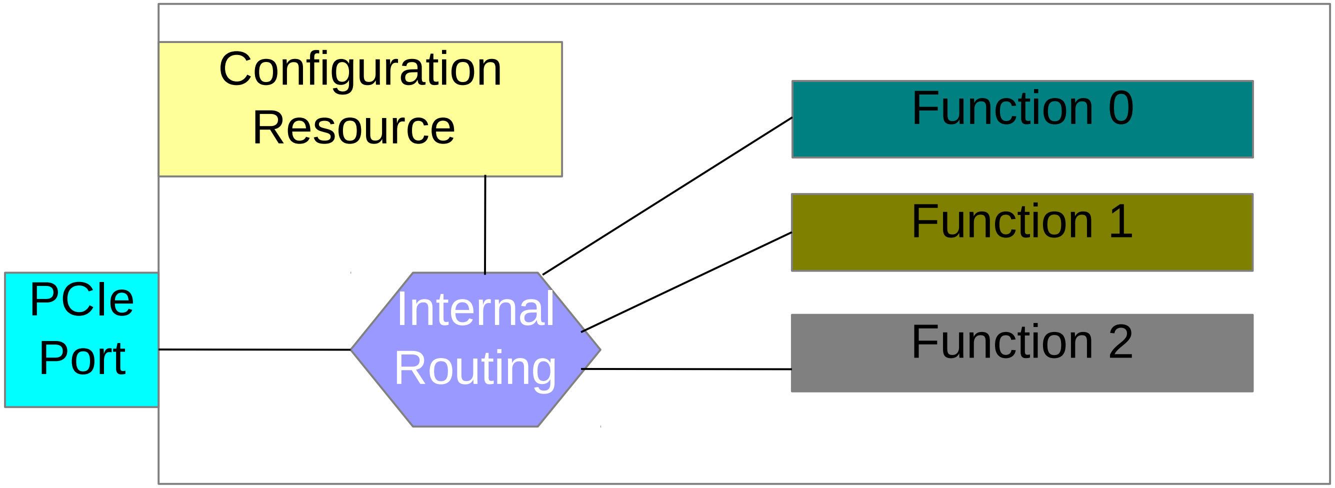
PCIe Root Complex: A system element that includes a Host Bridge and zero or more root ports.

Root Port: A PCIe port on root complex

Switch: A system element that connects two or more ports to allow packets to flow from one port to other.

PCIe device: a physical entity that performs a specific I/O function

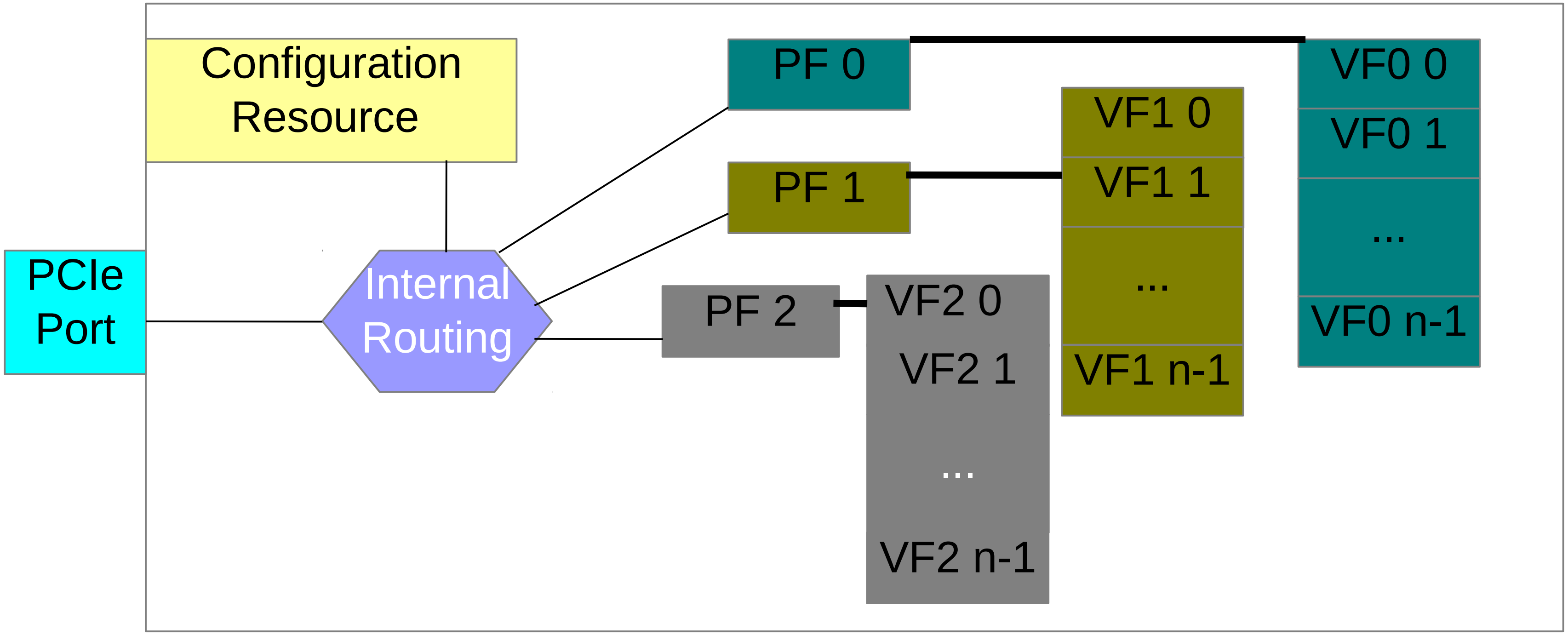
PCIe device



MMIO Space

Function 0	Function 1	Function 2

SRIOV device



PCIe Port

Configuration Resource

Internal Routing

PF 0

PF 1

PF 2

VF0 0

VF0 1

...

VF0 n-1

VF1 0

VF1 1

...

VF1 n-1

VF2 0

VF2 1

...

VF2 n-1

PF: physical function, VF: virtual function

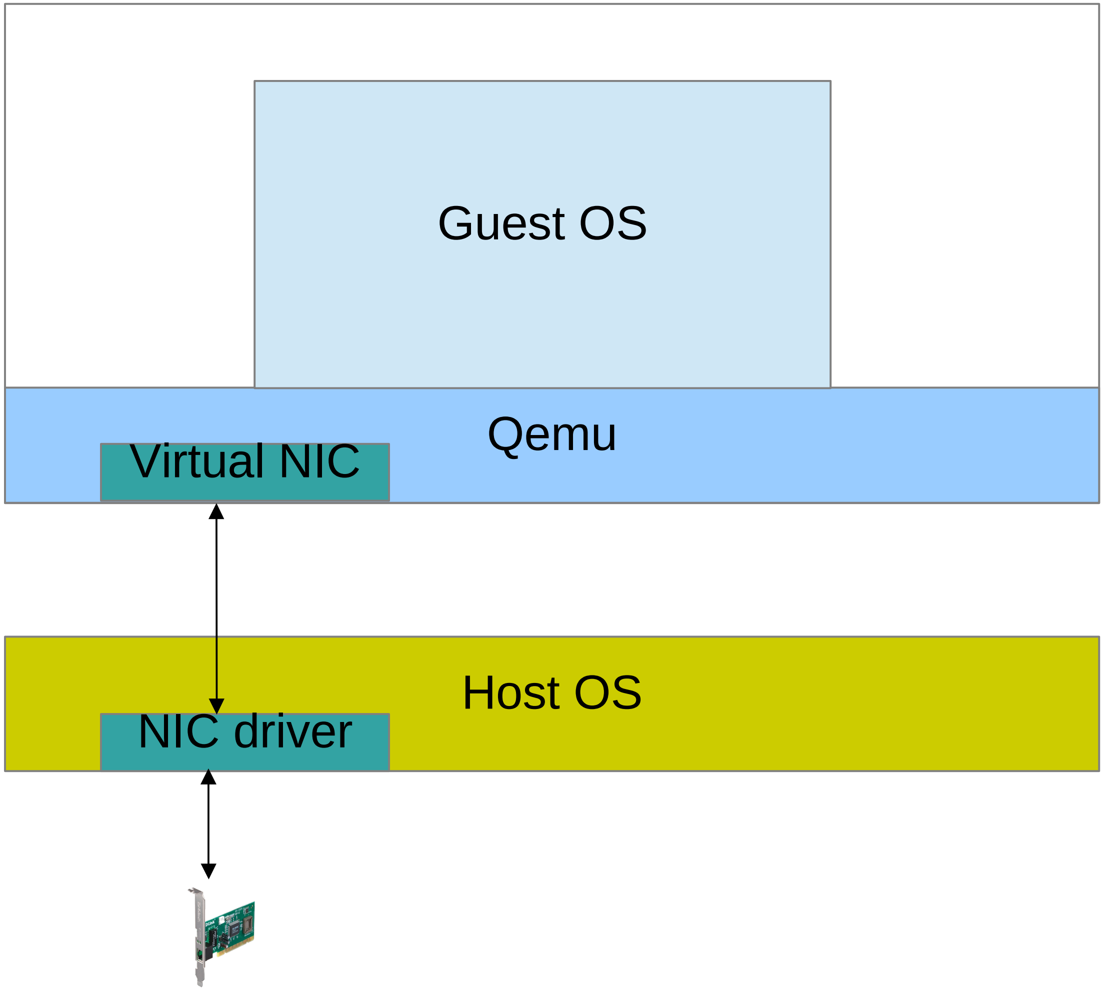
MMIO Space

PF 0	VF 0	VF ..	VF n-1	PF 1	VF 0	VF ...	VF n-1

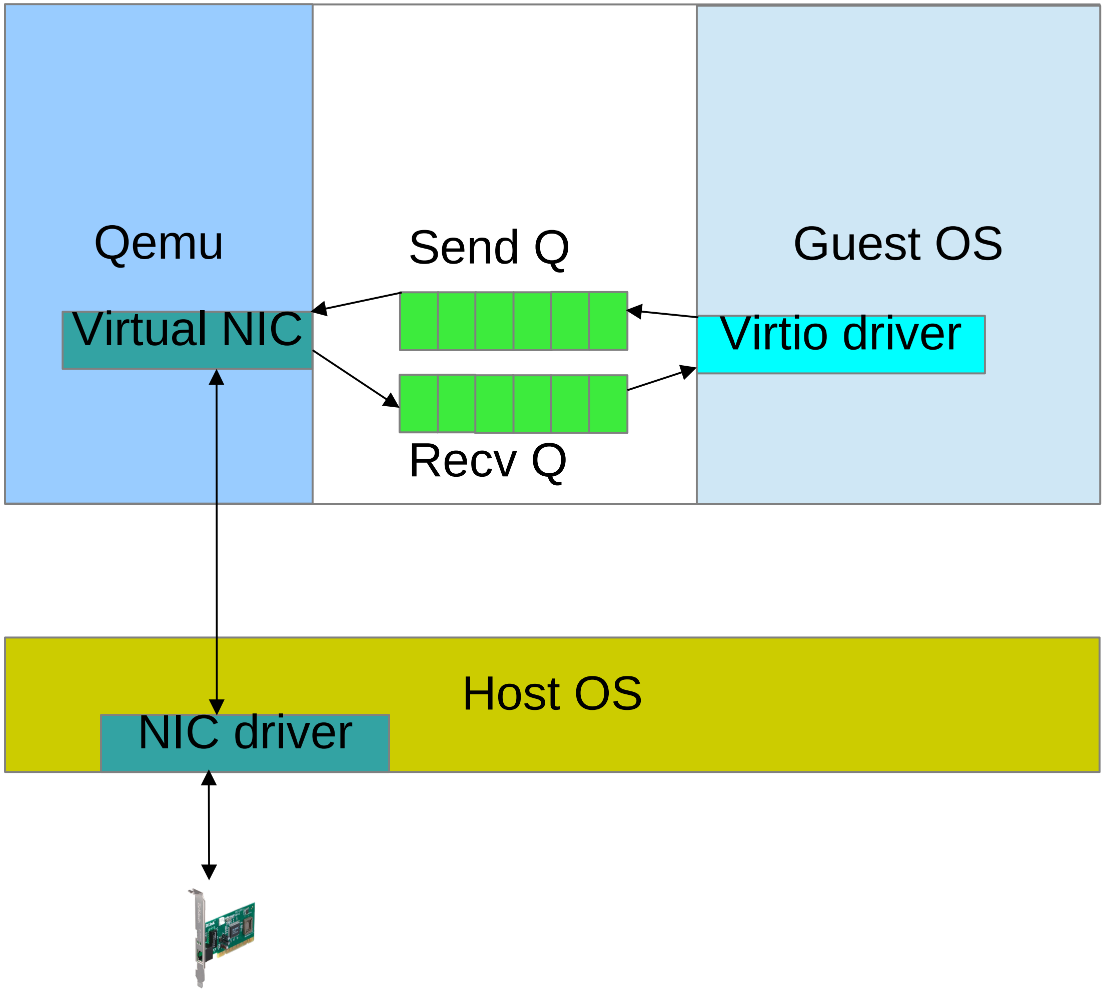
Agenda

- What's SRIOV?
  - PCIe Subsystem
  - PCIe device
  - SRIOV device
- **I/O virtualization**
  - **Emulation in user space**
  - **Virtio**
  - **vhost-net**
  - **SRIOV**
- SRIOV on POWER
  - Restrictions
  - Solution
- Future Work

Emulation in user space



Virtio



Qemu

Virtual NIC

Send Q

Recv Q

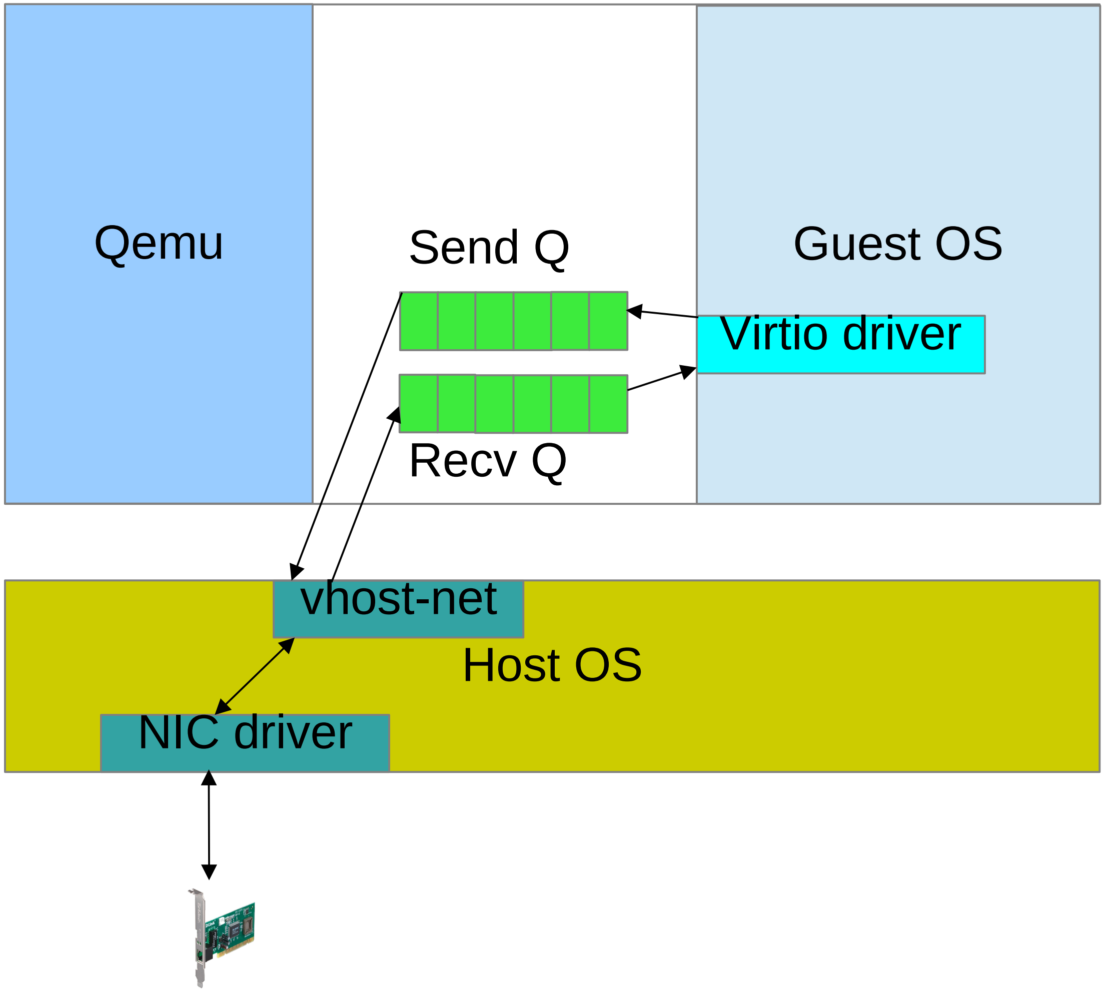
Guest OS

Virtio driver

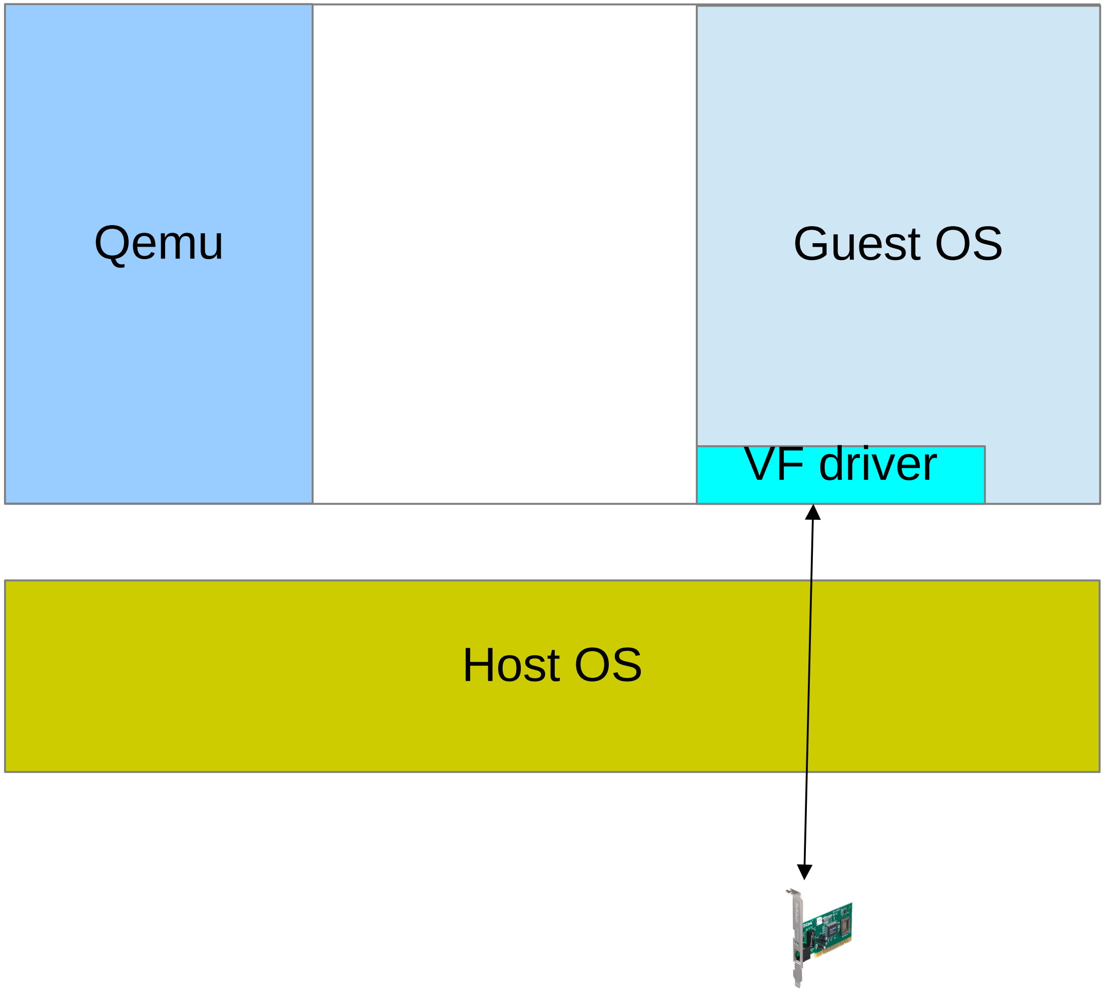
Host OS

NIC driver

host-net


Qemu
Send Q
Guest OS
Virtio driver
Recv Q
vhost-net
Host OS
NIC driver

SRIOV



Agenda

- What's SRIOV?
  - PCIe Subsystem
  - PCIe device
  - SRIOV device
- I/O virtualization
  - Emulation in user space
  - Virtio
  - vhost-net
  - SRIOV
- **SRIOV on POWER**
  - **Restrictions**
  - **Solution**
- Future Work

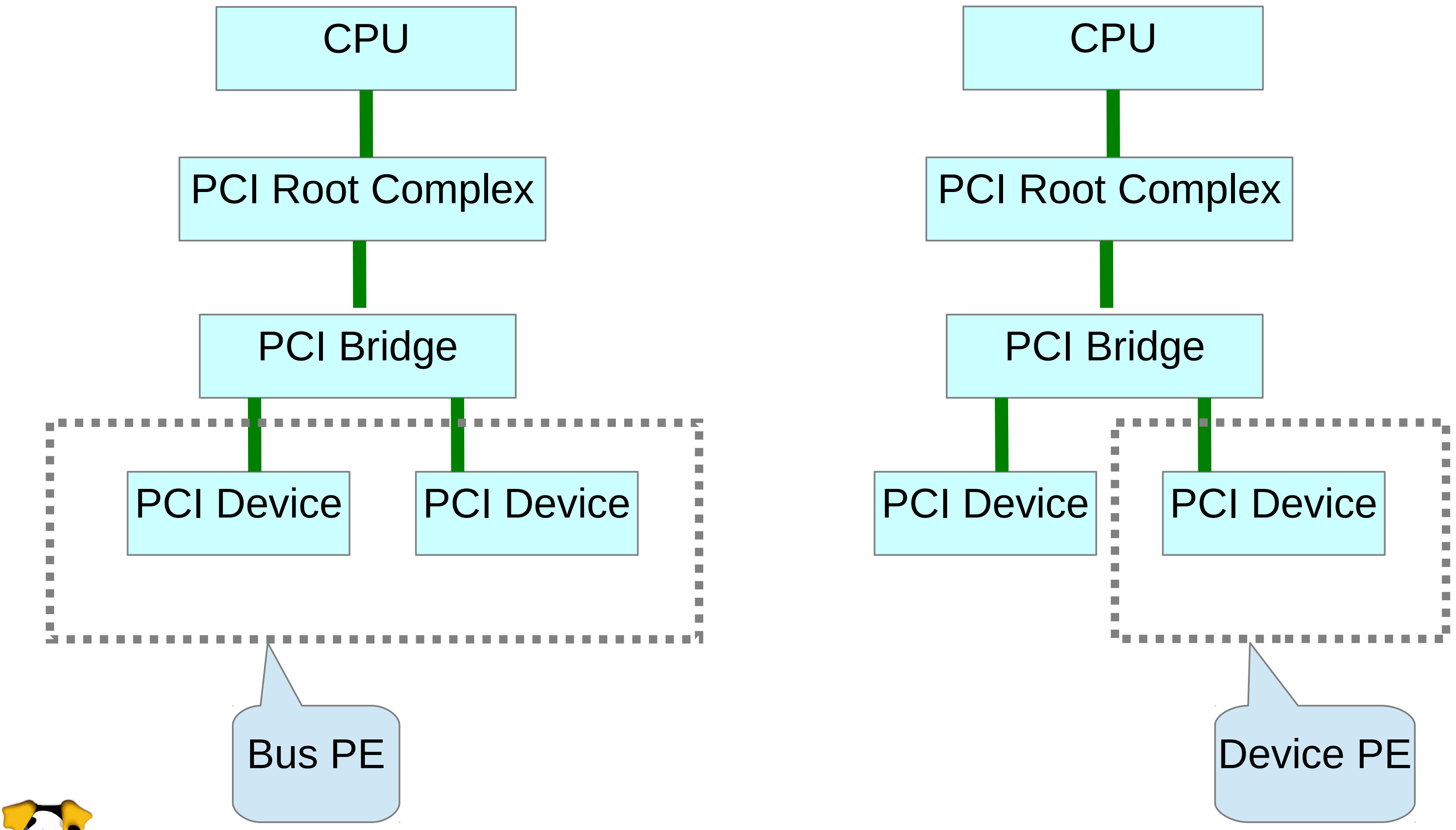
Partitionable Endpoint (PE)

- PE is an I/O error and recovery domain made up of
  - A single or multi-function IO Adapter or
  - A function of a multi-function IO Adapter or
  - Multiple IOAs, possibly includes upstream switches and bridges

- Partitionable Endpoint (PE) is defined in PAPR (Power Architecture Platform Requirements).

- RTAS compliant firmware supports EEH related operations at PE granularity.

Partitionable Endpoint (PE)

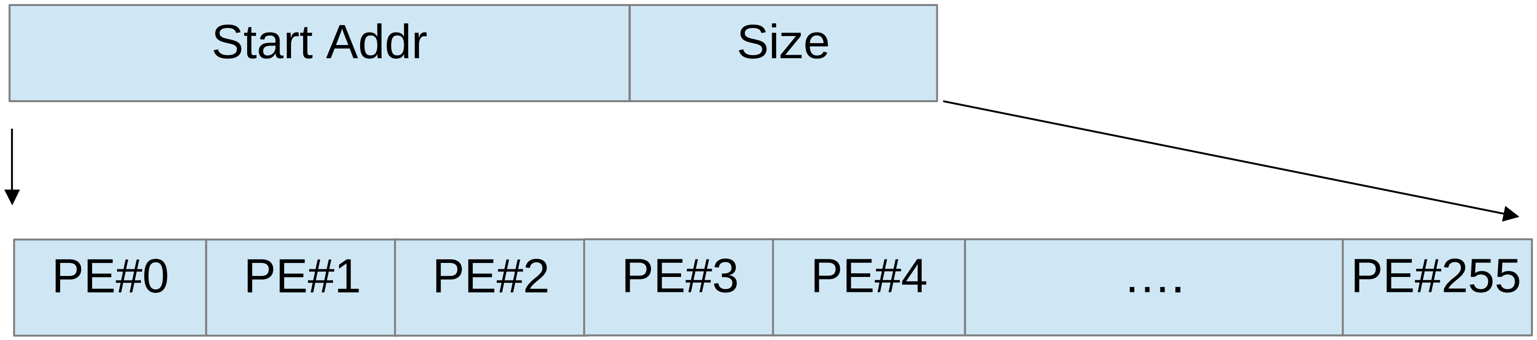


How to find the exact PE?

BDF range → PE number
MMIO Address → PE number
DMA Address → PE number
MSI Address → PE number

MMIO → PE#

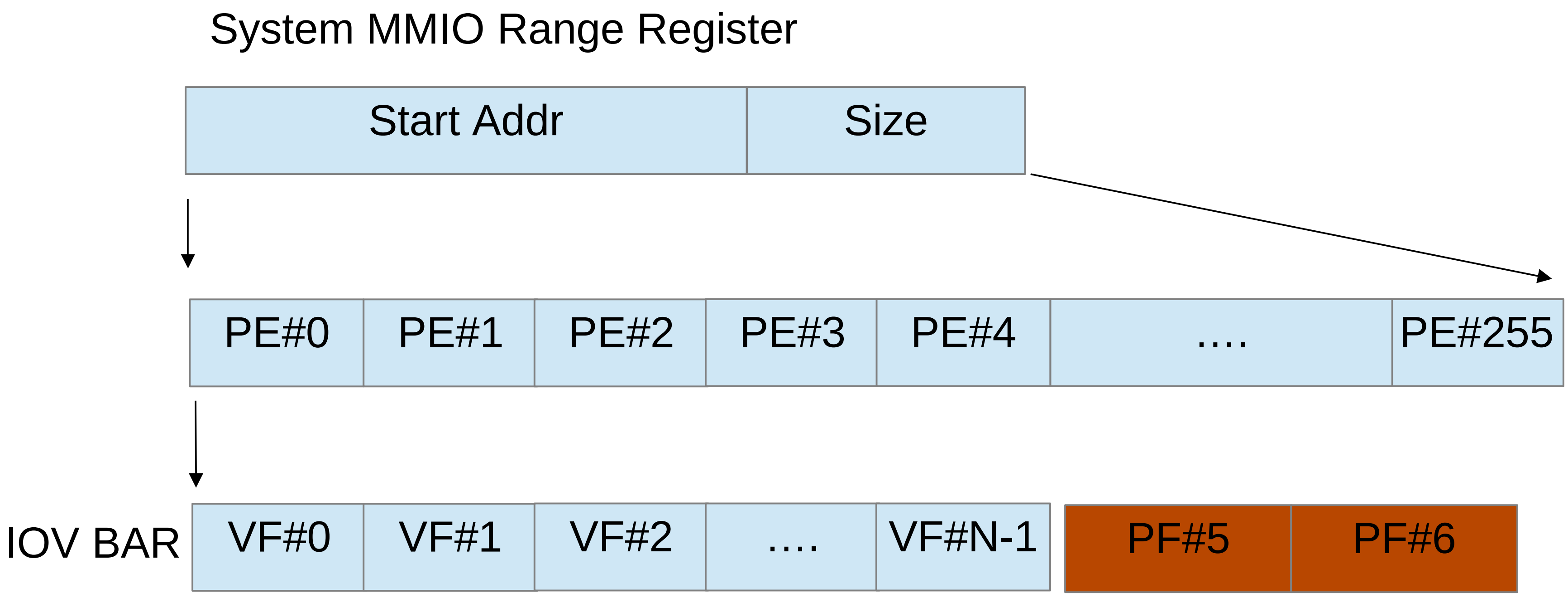
System MMIO Range Register



Start Addr | Size

PE#0 | PE#1 | PE#2 | PE#3 | PE#4 | …. | PE#255

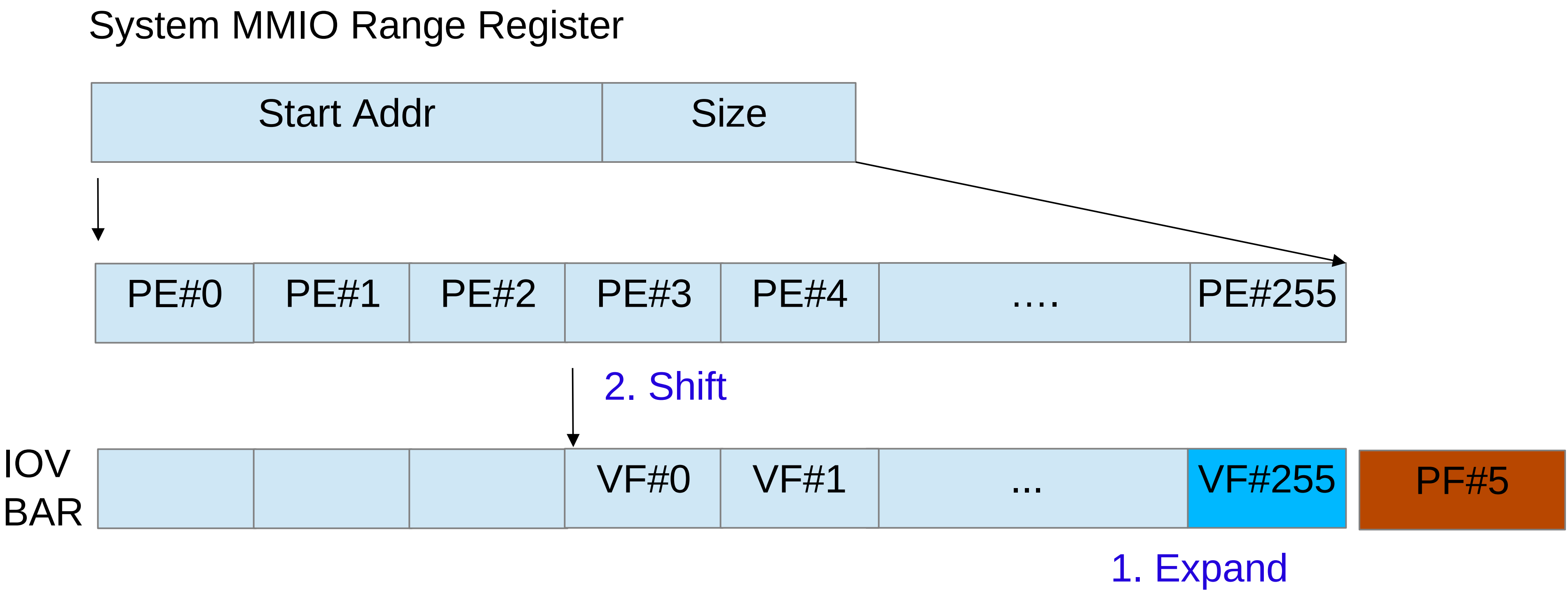
Restrctions on MMIO



Conflict:

1. PF#5 and PF6 may belong to other PE
2. PE#0 may already allocated

Our Solution



Solution of MMIO

Size: M64 must cover 256 x (seg size) MMIO range

1. Expand the IOV BAR to 256xVF BAR

IOV BAR shift: M64 will cover the whole PE# space

2. Jump the PE# which has been used, otherwise will be conflict

Alignment: M64 must be size aligned

3. Count in the IOV BAR alignment in PCI MMIO sizing/assignment

Agenda

- What's SRIOV?
  - PCIe Subsystem
  - PCIe device
  - SRIOV device
- I/O virtualization
  - Emulation in user space
  - Virtio
  - vhost-net
  - SRIOV
- SRIOV on POWER
  - Restrictions
  - Solution
- **Future Work**

Future Work

1. Virtual Bus support
   For some sriov devices, VF could sits on a virtual bus.
   This needs supports from both firmware and kernel.

2. EEH for Vfs
   EEH stands for Enhanced Error Handling.
   After VF introduced, the we need some special steps to handle it well.

Legal Statement

This work represents the view of the author and does not necessarily represent the view of IBM.

IBM, IBM (logo), AIX, POWER, POWER6, POWER7 and PowerVM are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries.

Linux is a registered trademark of Linus Torvalds.

Other company, product and service names may be trademarks or service marks of others.

Thanks & Questions