Generic SR-IOV Device Support for Kernel Live Upgrade in Cloud Environment

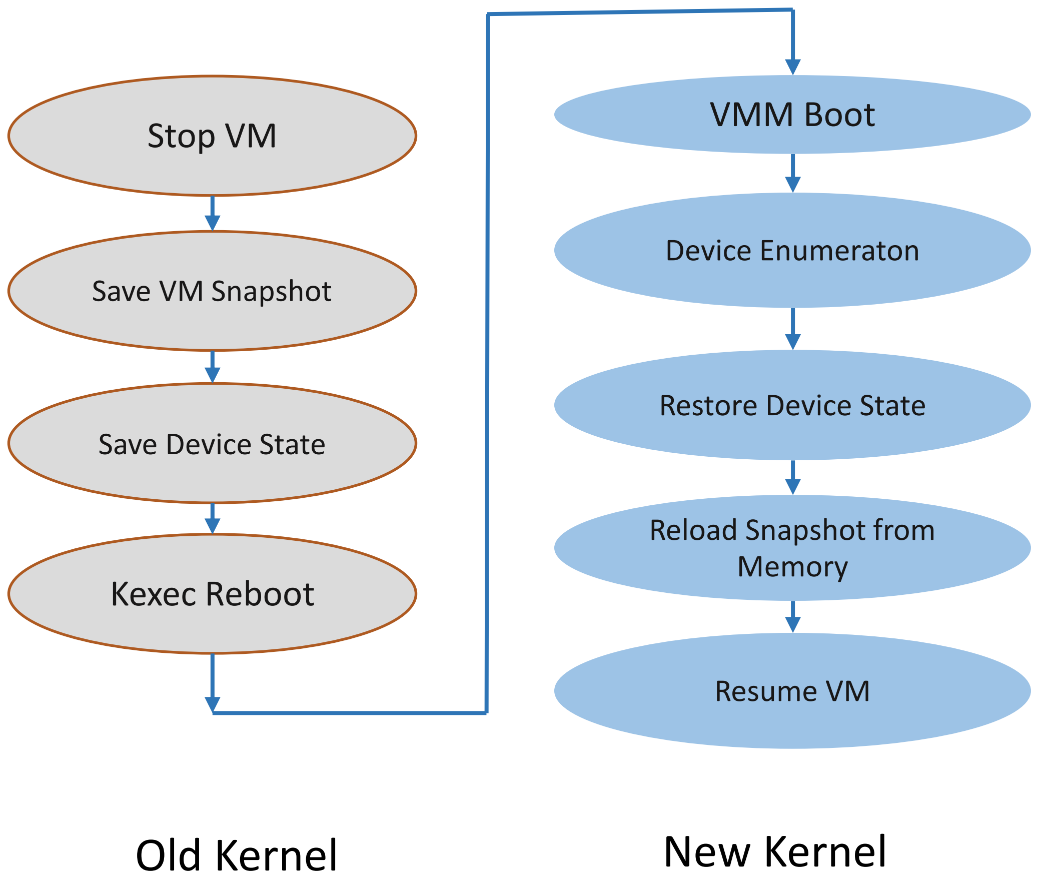
jason.zeng@intel.com

Disclaimers

Agenda

- Kernel Live Upgrade Recap
- SR-IOV Device Support
- Implementation

Kernel Live Upgrade Recap

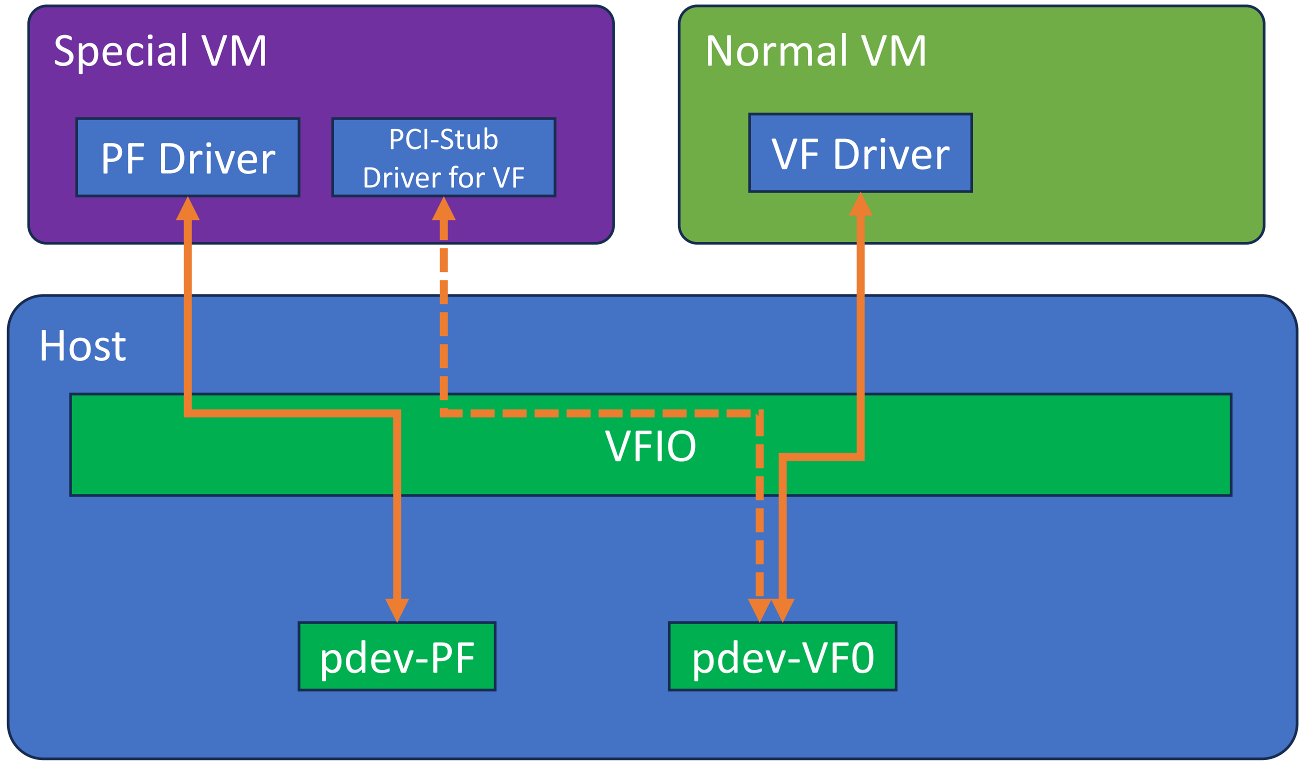


- Passthrough Device Support
  - Keep upstream devices alive
    - IOMMU
    - PCIe switch/root-port
  - Keep Device HW alive
    - DMA/Interrupt
    - Other HW states

SR-IOV Device Support – The problem

- SR-IOV PF usually managed by host
- PF HW state need to be preserved across kernel upgrade
- There are many PF vendors
    - Huge effort of vendor specific PF driver code change to preserve PF HW state

The Generic Way – Put PF in a Special VM

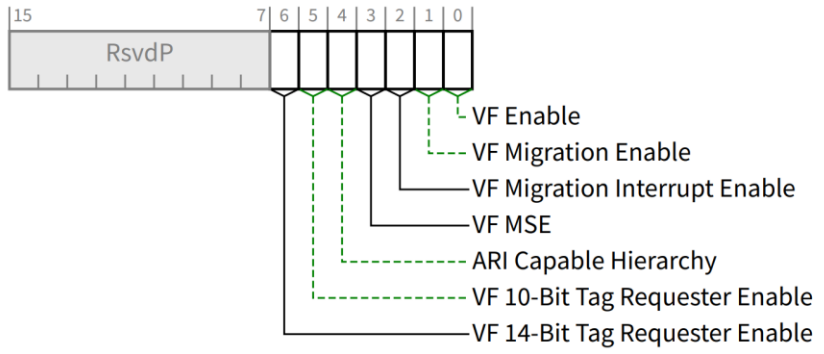


- Passthrough PF to Special VM
- When kernel live upgrade happens
  - PF HW states preserved together with the Special VM

Steps：

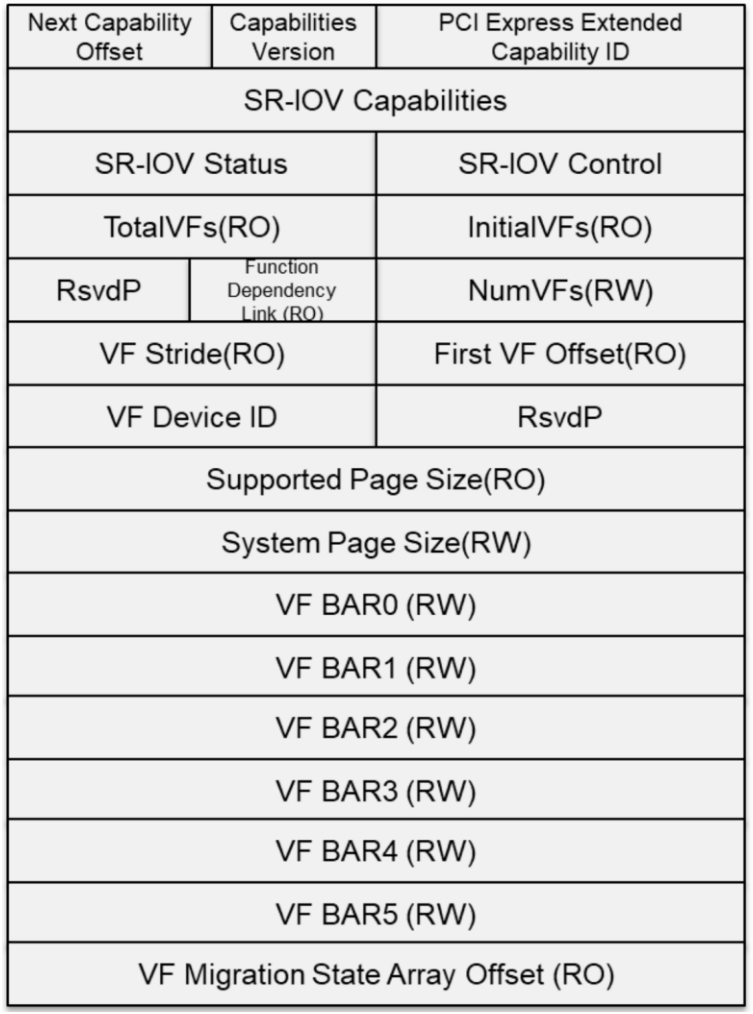
1. From Host
   - Passthrough PF to Special VM
2. From Special VM
   - Create VFs
   - Bind VF driver to pci-stub driver
3. From Host
   - Assign VF to Normal VM
4. From Normal VM
   - Bind VF driver to vendor’s VF driver

Implementation - Passthrough SR-IOV PF


15
7
6
5
4
3
2
1
0
RsvdP
VF Enable
VF Migration Enable
VF Migration Interrupt Enable
VF MSE
ARI Capable Hierarchy
VF 10-Bit Tag Requester Enable
VF 14-Bit Tag Requester Enable

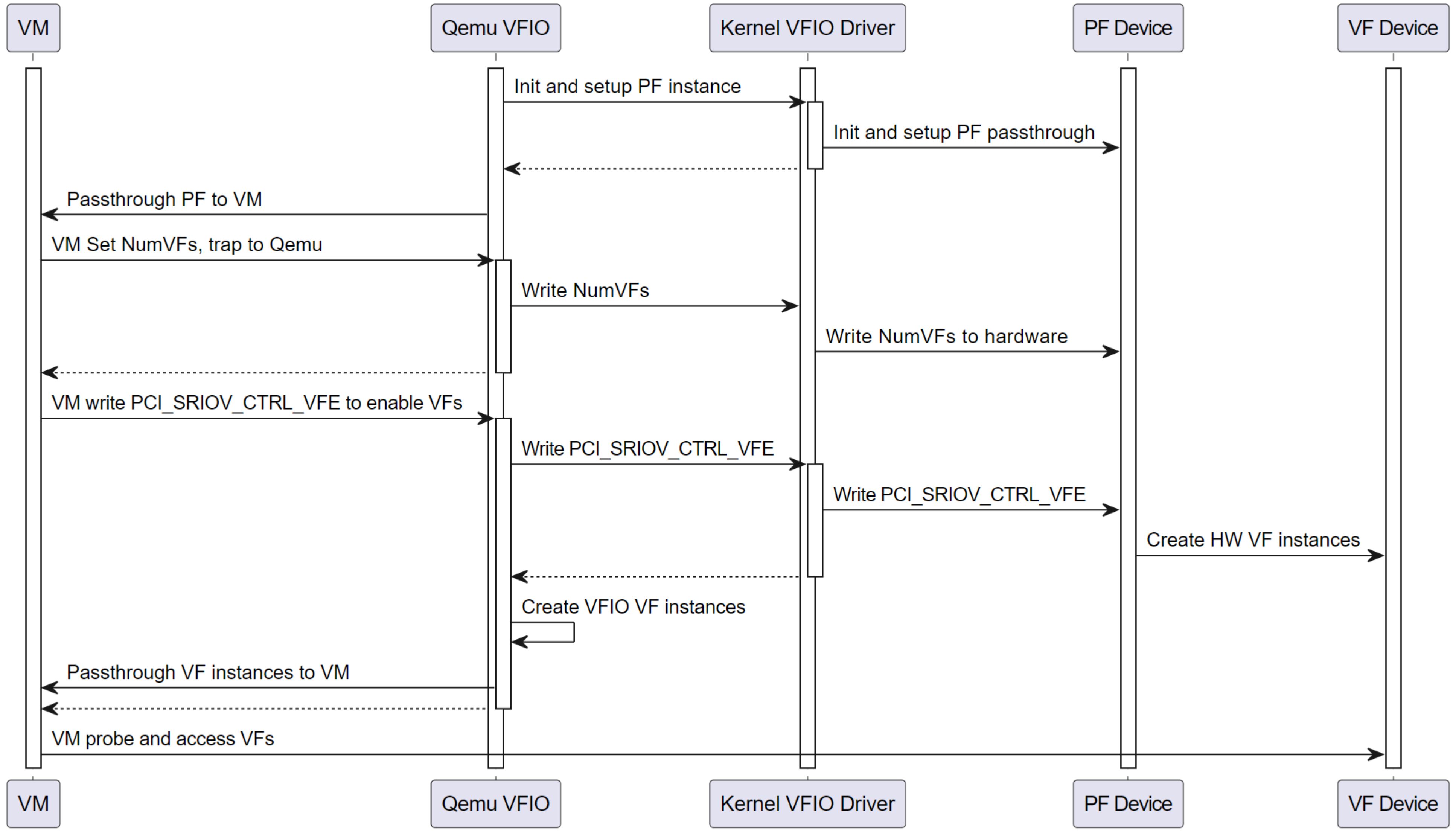
§ *Figure 9-13 SR-IOV Control Register*

- Kernel VFIO driver
  - Intercept NumVFs
  - Intercept VF_Enable & VF_MSE in Control Register
  - SR-IOV VF BAR support
    - VFIO device region
- Qemu
  - Enable SR-IOV capability handling in VFIO driver
  - SR-IOV VF BAR support
  - VF creation/destroy upon VM SR-IOV cap read/write



Next Capability Offset	Capabilities Version	PCI Express Extended Capability ID
SR-IOV Capabilities		
SR-IOV Status		SR-IOV Control
TotalVFs(RO)		InitialVFs(RO)
RsvdP	Function Dependency Link (RO)	NumVFs(RW)
VF Stride(RO)		First VF Offset(RO)
VF Device ID		RsvdP
Supported Page Size(RO)		
System Page Size(RW)		
VF BAR0 (RW)		
VF BAR1 (RW)		
VF BAR2 (RW)		
VF BAR3 (RW)		
VF BAR4 (RW)		
VF BAR5 (RW)		
VF Migration State Array Offset (RO)		

Flow: Create VFs from VM



Status & Plan

- PF Passthrough POC
  - https://gitee.com/x56Jason/vmmfr-linux/tree/pfpt/
  - https://gitee.com/x56Jason/vmmfr-qemu/tree/pfpt/

- Next
  - ARI capability support?
  - Target openEuler-22.03-LTS-SP3?