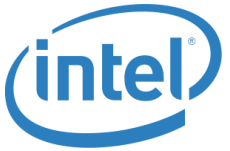
VFIO-based Mediated Pass Through – KVMGT as an example

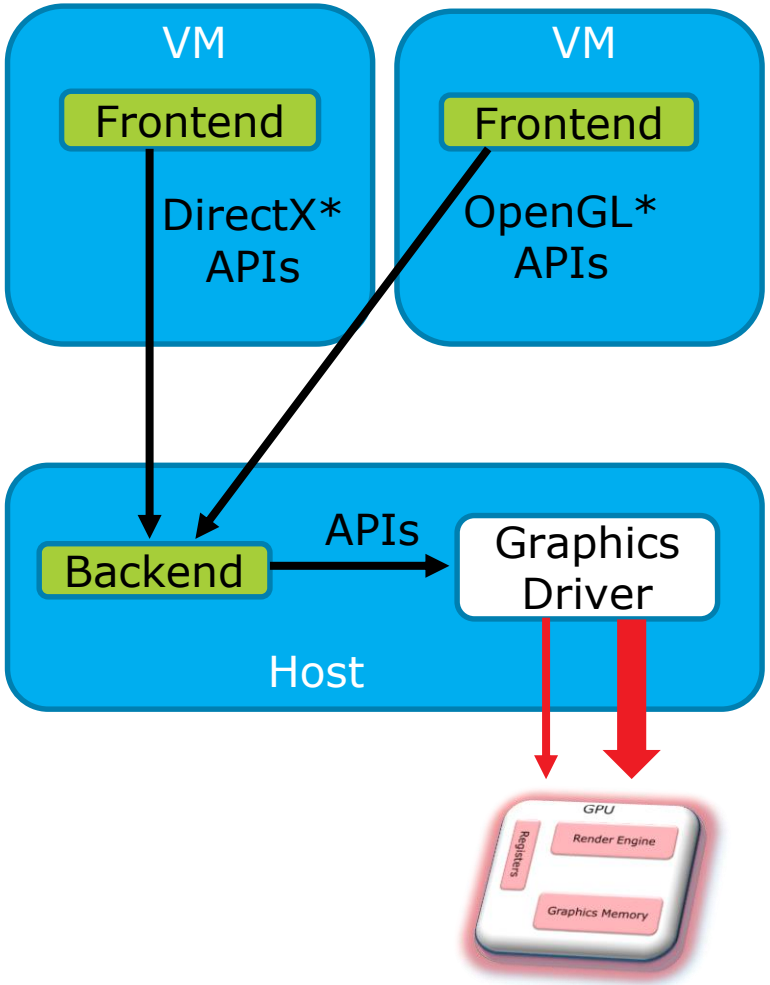
Jike Song jike.song@intel.com
October 2016

Agenda

- ❑ Graphics Virtualization: KVMGT and MPT
- ❑ VFIO
- ❑ VFIO-based KVMGT



Graphics Virtualization – Before MPT: API Forwarding



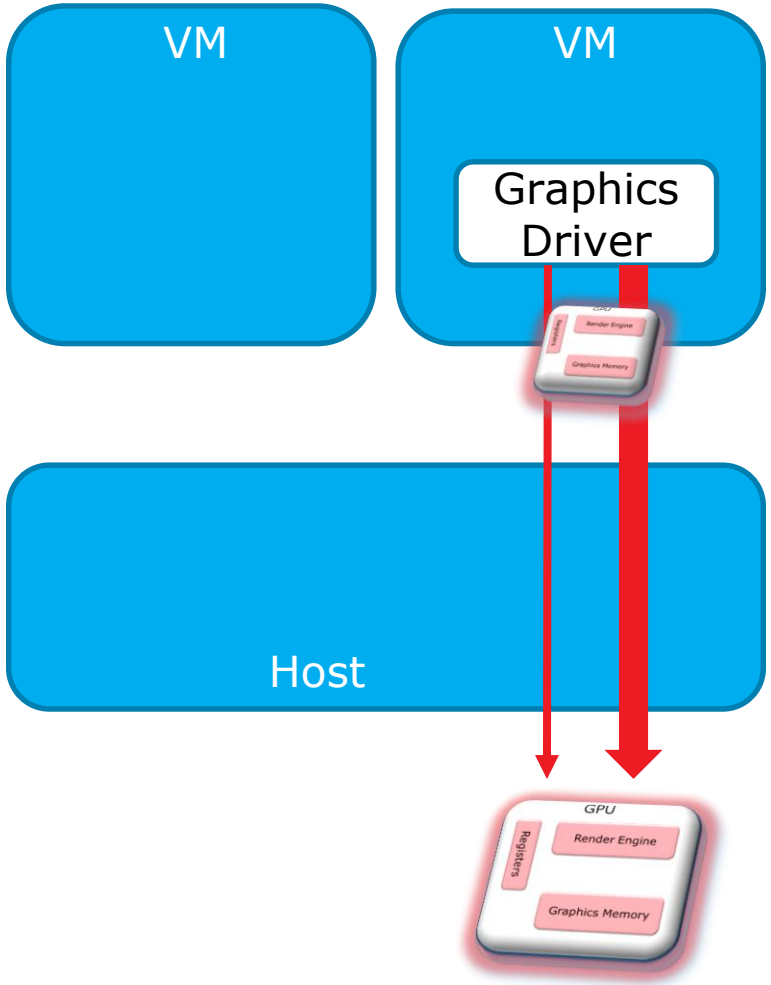
Pros

- Performance
- Scalability

Cons

- Lagging features
- Incompatible APIs
- Maintenance burden

Graphics Virtualization – Before MPT: direct Pass-Through



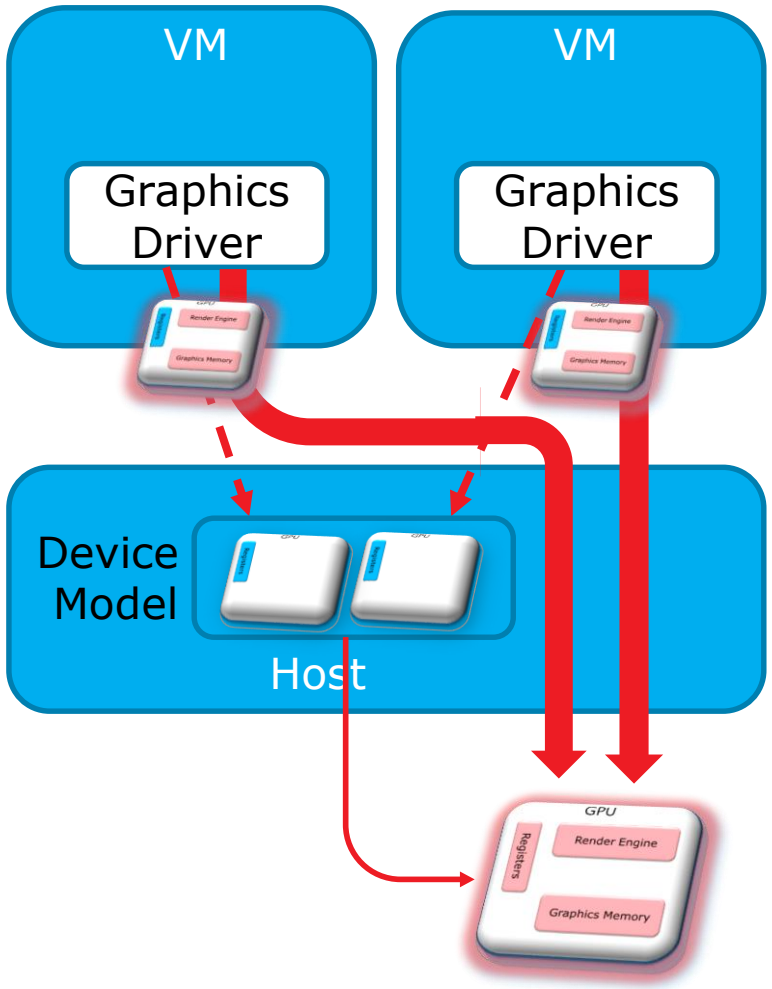
Pros

- Performance
- Full features

Cons

- No or limited sharing (w/o or w/ SR-IOV)

Graphics Virtualization – MPT: Mediated Pass-Through



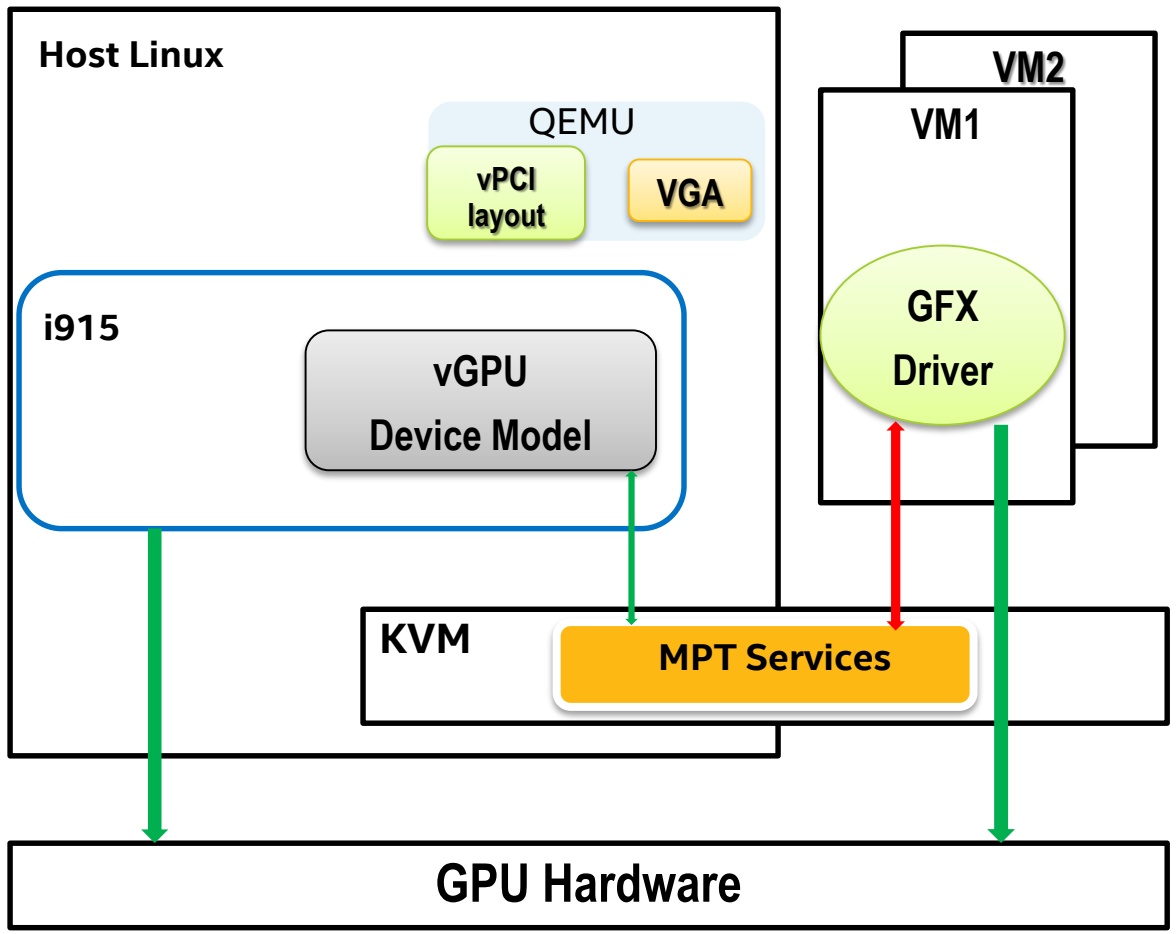
Pros

- Performance
- Full feature
- Scalability

Cons

- Vendor specific

Graphics Virtualization – MPT based KVMGT



Pros

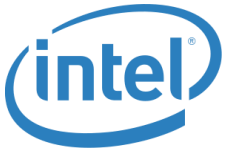
- Full Feature
- Performance
- Scalability

Cons

- Touched a lot: Kernel, KVM, i915, QEMU, SeaBIOS ...

Agenda

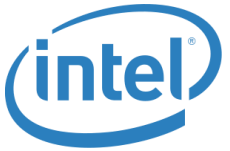
- Graphics Virtualization: KVMGT and MPT
- VFIO
- VFIO-based KVMGT



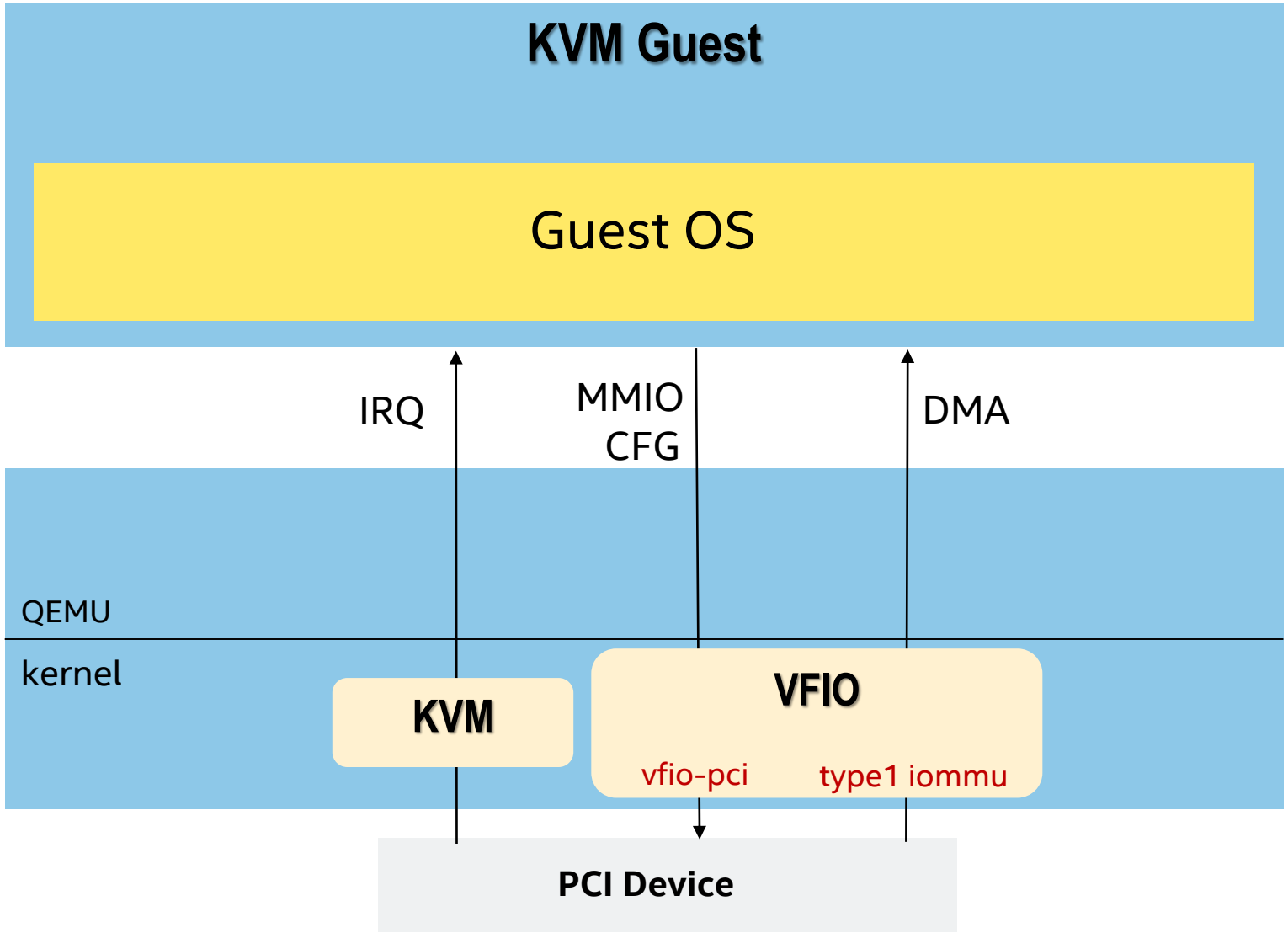
VFIO – Virtual Function I/O

- By Alex Williamson @ Redhat
- Used for: device Pass-Through in KVM
  - Replaced the legacy PCI Assignment in KVM
- Used for: Userspace Drivers
  - Replaced UIO

- Modular Bus drivers, Modular IOMMU backends
  - Available Bus drivers: PCI, platform
  - Available IOMMU backends: type1, SPARR



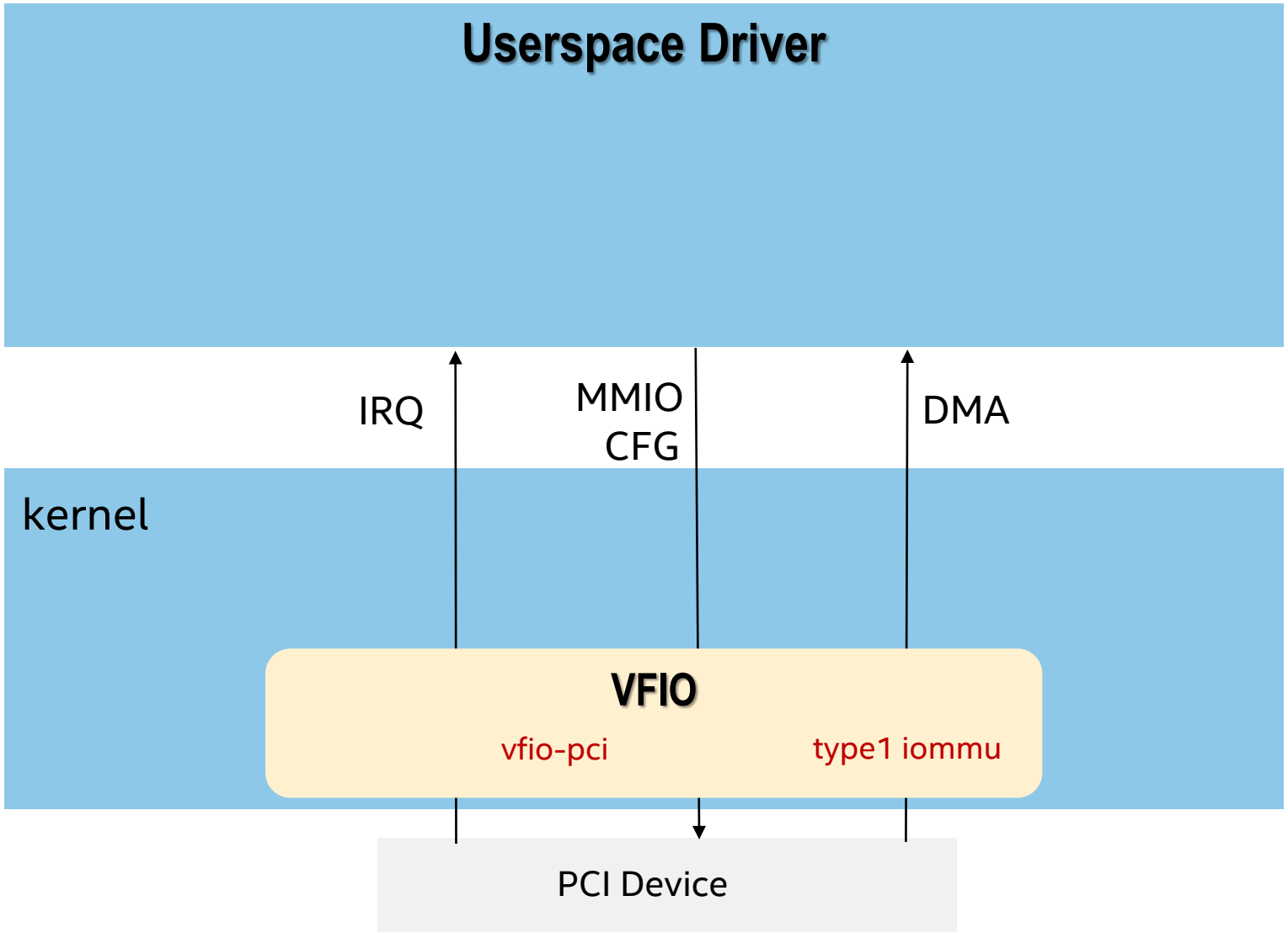
VFIO – PCI device Pass-Through to KVM Guest



VFIO – PCI device Pass-Through to KVM Guest

- A PCI Device or VF consists of:
  - PCI Configuration Registers
  - MMIO Registers
  - INTx/MSI/MSI-X IRQ
  - DMA
- VFIO passthrough it by:
  - vfio_pci bus driver
    - ✓ PCI CFG : proxying the access
    - ✓ MMIO : mmap to QEMU, thereby to guest
    - ✓ IRQ : eventfd to QEMU, ioctl to KVM & inject to guest
  - Type1 IOMMU backend
    - ✓ DMA: pin & map GPA(Guest Physical Address) to HPA(Host Physical Address)

VFIO – PCI device Userspace Driver

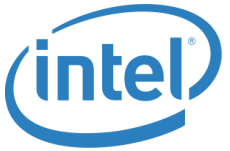


VFIO – PCI device Userspace Driver

- VFIO enables userspace driver by:
  - vfio_pci bus driver
    - ✓PCI CFG : porxying the access
    - ✓MMIO : mmap to userspace
    - ✓IRQ : eventfd to userspace
  - Type1 IOMMU backend
    - ✓DMA: pin & map userspace virtual address to physical address

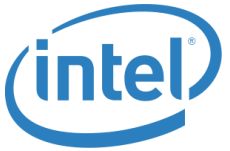
Agenda

- ❑ Graphics Virtualization: KVMGT and MPT
- ❑ VFIO
- ❑ VFIO-based KVMGT

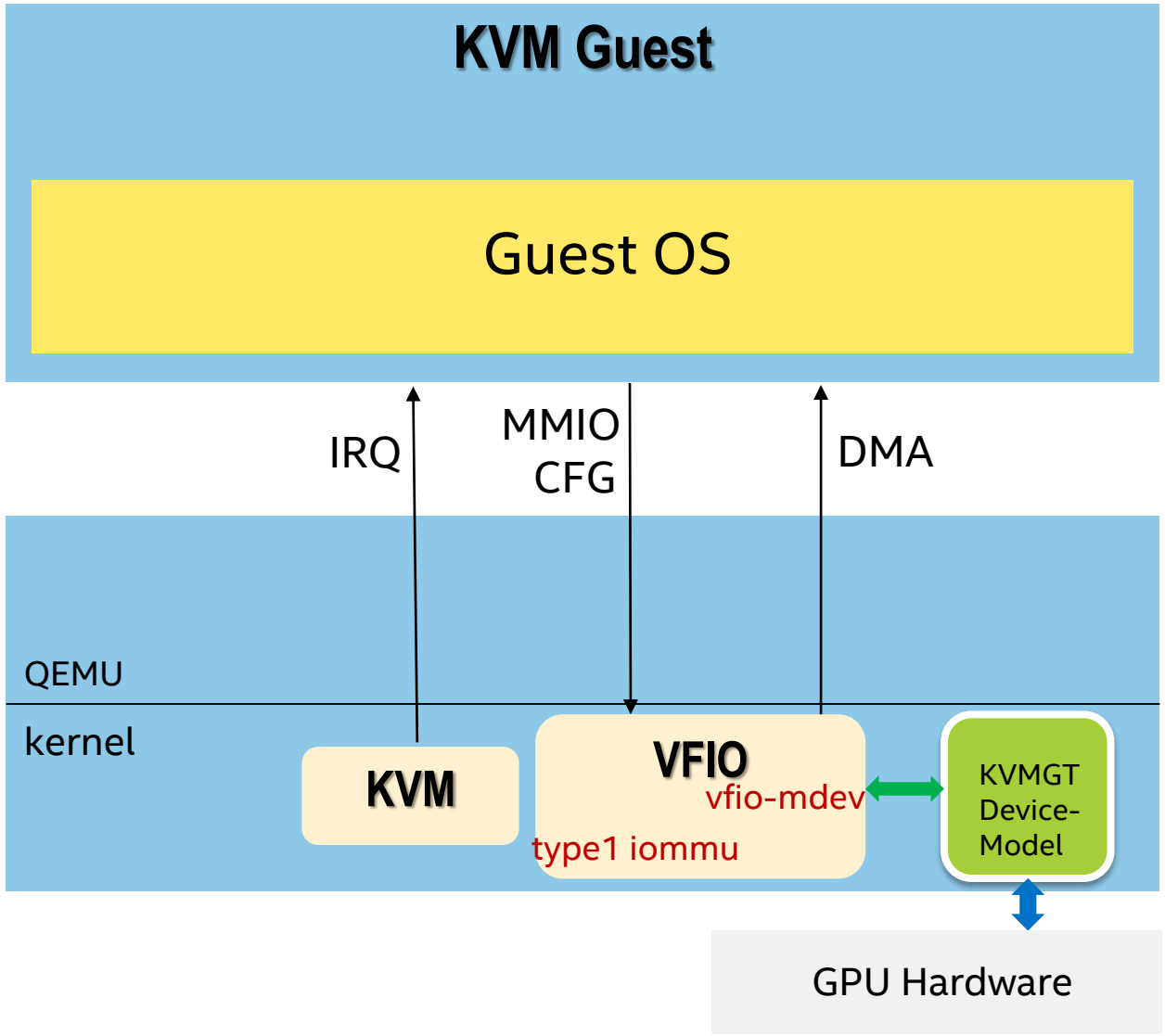


VFIO Based KVMGT – the VFIO MDEV support

- Framework first implemented by Nvidia
- Upstreaming in progress
- New Bus driver for Mediated Device
  - vfio-mdev
  - flexible configuration: trapping or passing-through
  - Capable of being compatible with the existing userspace API for PCIDEV
  - Yet not PCI-specific
- Extended type1 IOMMU backend
  - Pin guest pages on-demand
  - Without hardware IOMMU dependency
- Multiple Usage
  - vGPU Solution : Nvidia, Intel
  - CCW Pass-Through : IBM
  - Probably other mediated devices in the near future



VFIO Based KVMGT – the new KVMGT



Pros compared with old KVMGT

- API compatibility with vfio-pci and all vGPU vendors
- No QEMU/SeaBIOS changes

Cons compared with old KVMGT

- More difficult to support primary GPU mode
- A little performance drop: MMIO emulation is longer

Thank you!

Questions?