Toward a Virtualization World Built on Mediated Pass-Through

Kevin Tian
Principal Engineer, Intel

Legal Disclaimer

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

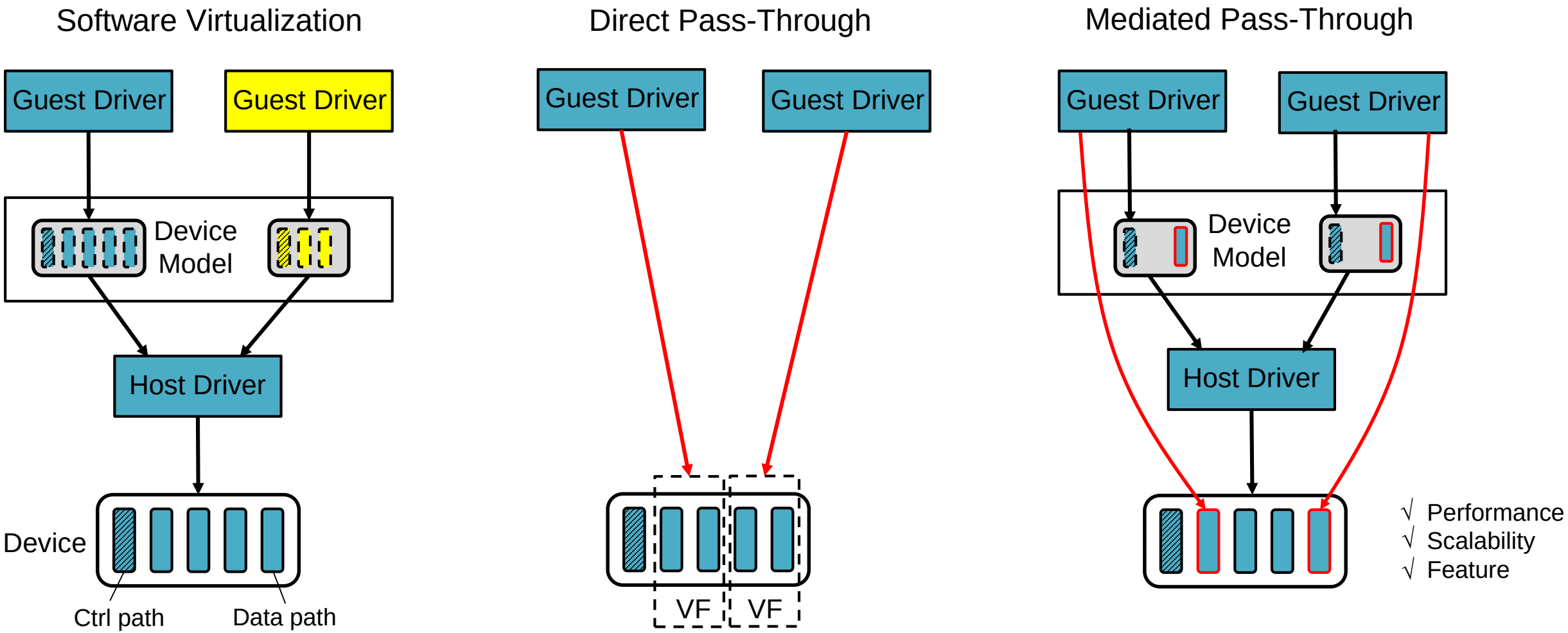
The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel and the Intel logo are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

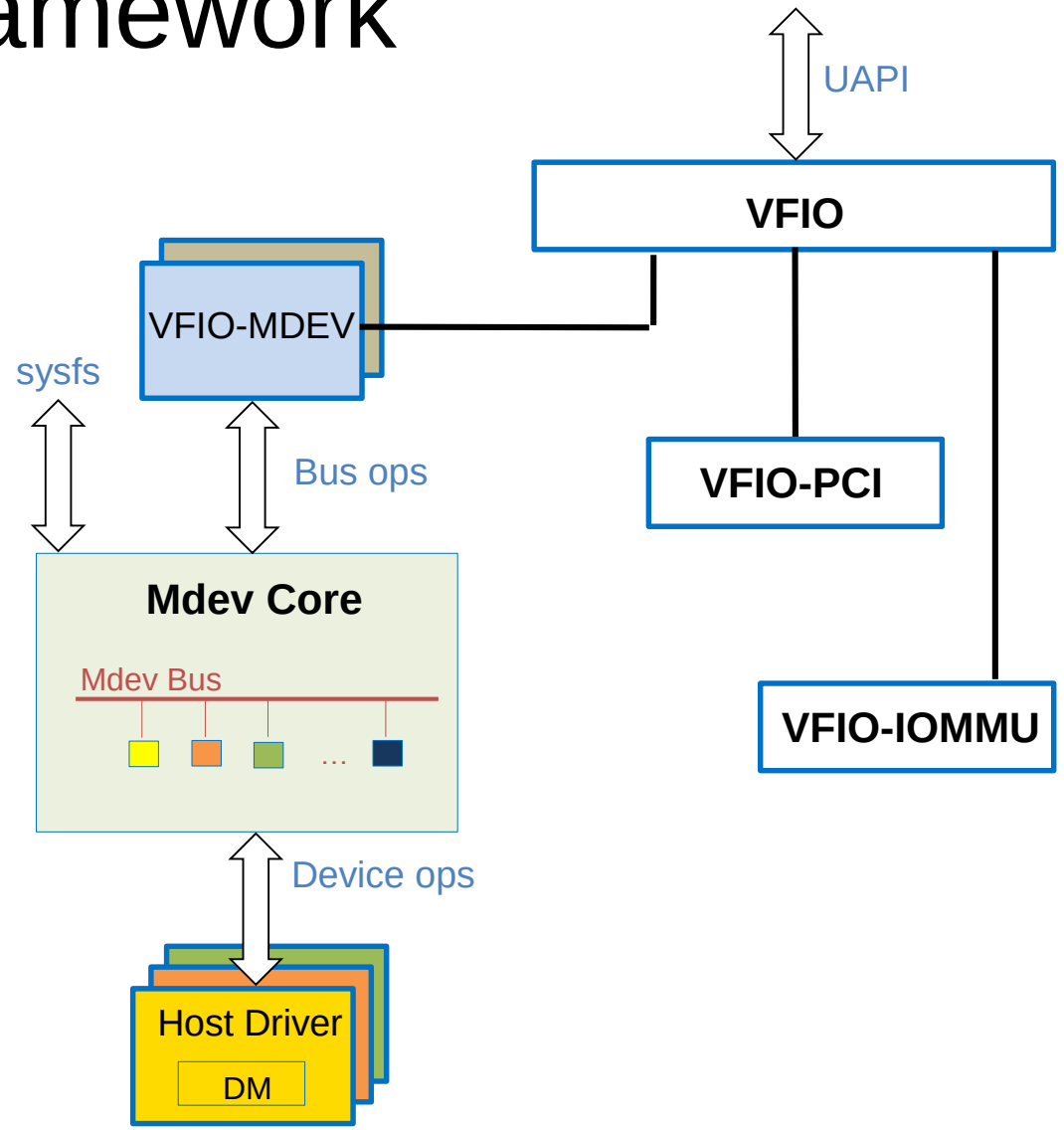
*Other names and brands may be claimed as the property of others

I/O Virtualization



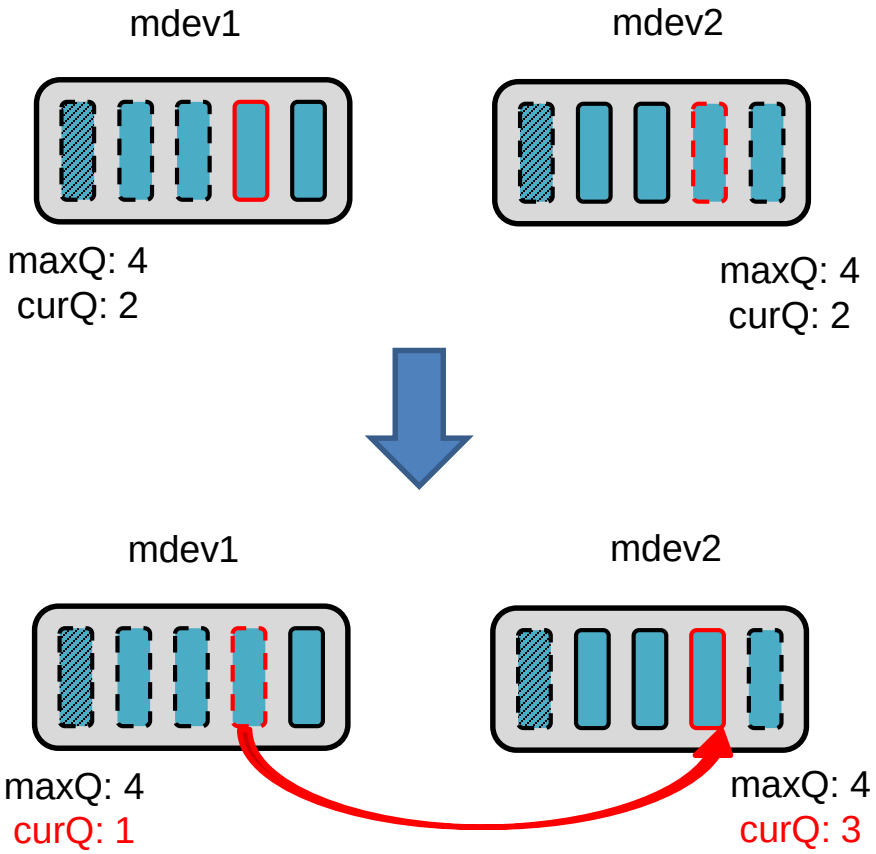
Linux* Mediated Device Framework

- Introduced in kernel 4.10
- Device ops to connect all kinds of mediated devices
  - GPUs, NICs, platform devices, etc.
- Bus ops as the bridge to various UAPIs
  - VFIO, virtio, vhost, etc.
- Life-cycle management through sysfs



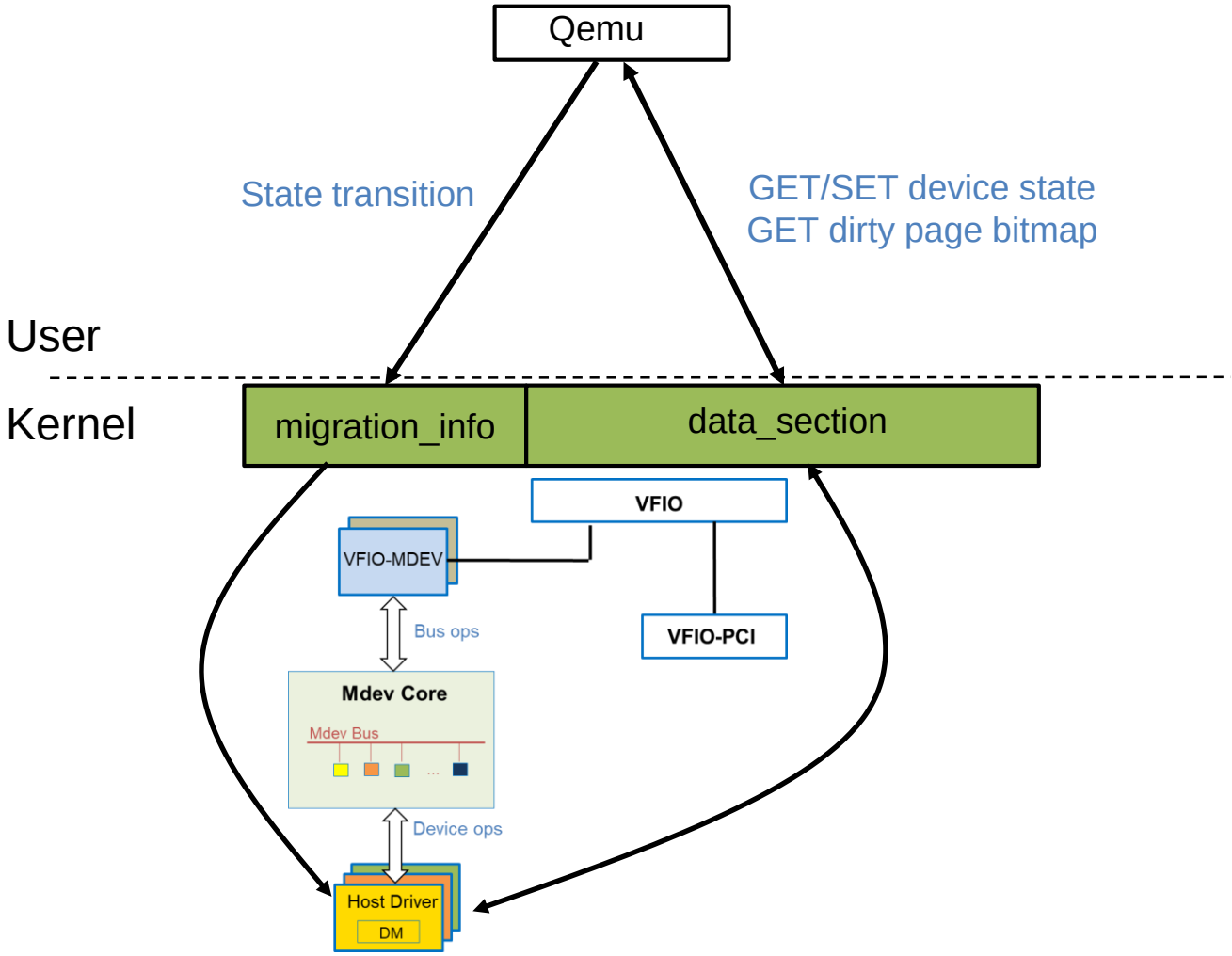
Explore More Values beyond Mediate-for-Sharing

Resource Scaling



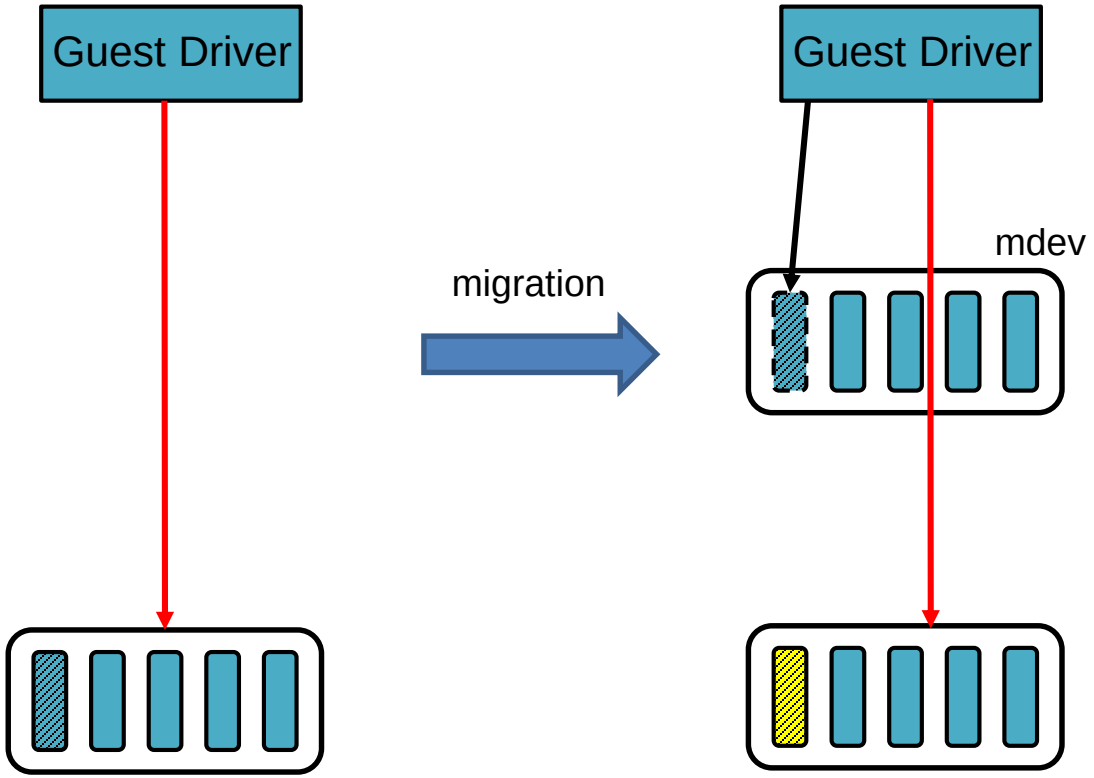
- Dynamic queue re-allocation
  - Load-balancing, queue over-commitment, etc.
  - maxQ vs. curQ
- Mdev device driver registers VMA fault handler for re-allocation (example)
  - Guest-transparent way (only see #maxQ)
  - Guest-cooperated way (see both #maxQ and #curQ)
- Same technique could also be used for failover

Live Migration



- Mediate for composing device state
- A new migration region on mdev
  - State transition (running, stopped, etc.)
  - GET/SET device state
  - GET dirty page bitmap
- Currently in v8

Generational Compatibility



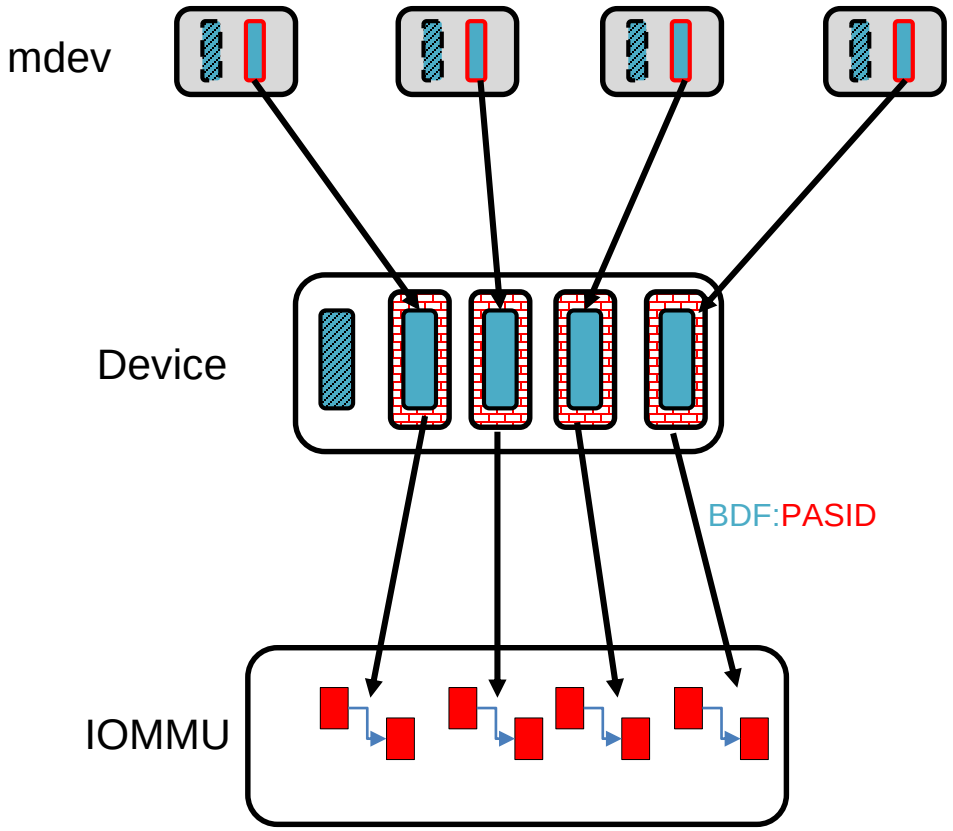
- Underlying device is incompatible to guest driver
  - Post migration, legacy OS, etc.
- Mediate for compatible device interface
- Useful for small generation jump
  - Incompatible control interface
  - Compatible data path

Memory Footprint

- Mediate for pinning guest DMA pages
  - For efficient memory utilization
- Mdev device driver tracks the status of guest DMA pages
  - E.g. scanning ring descriptors or device mmu page tables, etc.
- vfio_pin_pages for selectively pinning a set of guest PFNs

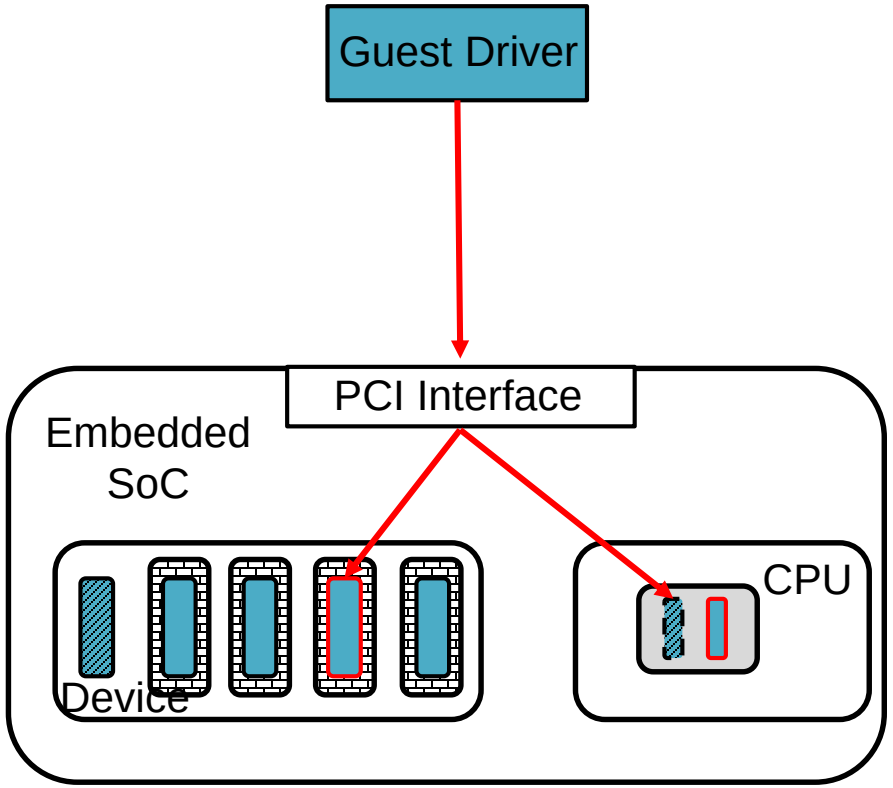
Innovate the Mediated Device framework

Hardware Assistance



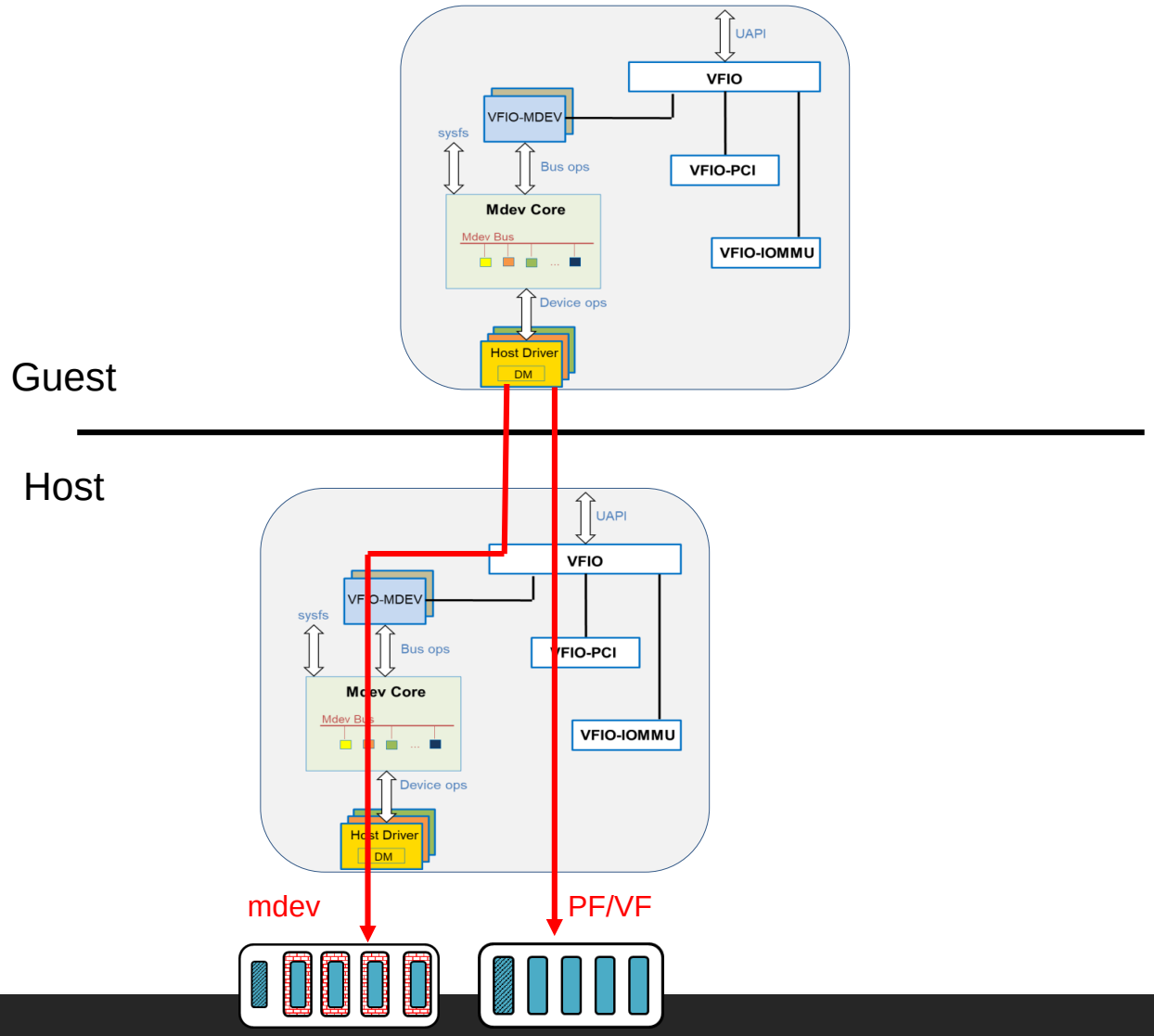
- Example: Intel® Scalable I/O Virtualization
  - For higher density and security
- Device: finer-grained resource isolation
  - ADI: queue, queue pair, or context
  - 4K aligned MMIO ranges
  - Scalable interrupt message storage
  - Independent reset
  - …
- IOMMU: finer-grained DMA isolation
  - PASID-granular address translation
  - Primary domain vs. Auxiliary (AUX) domain
- VFIO: iommu-capable mdev

Hardware-offloaded Mediation



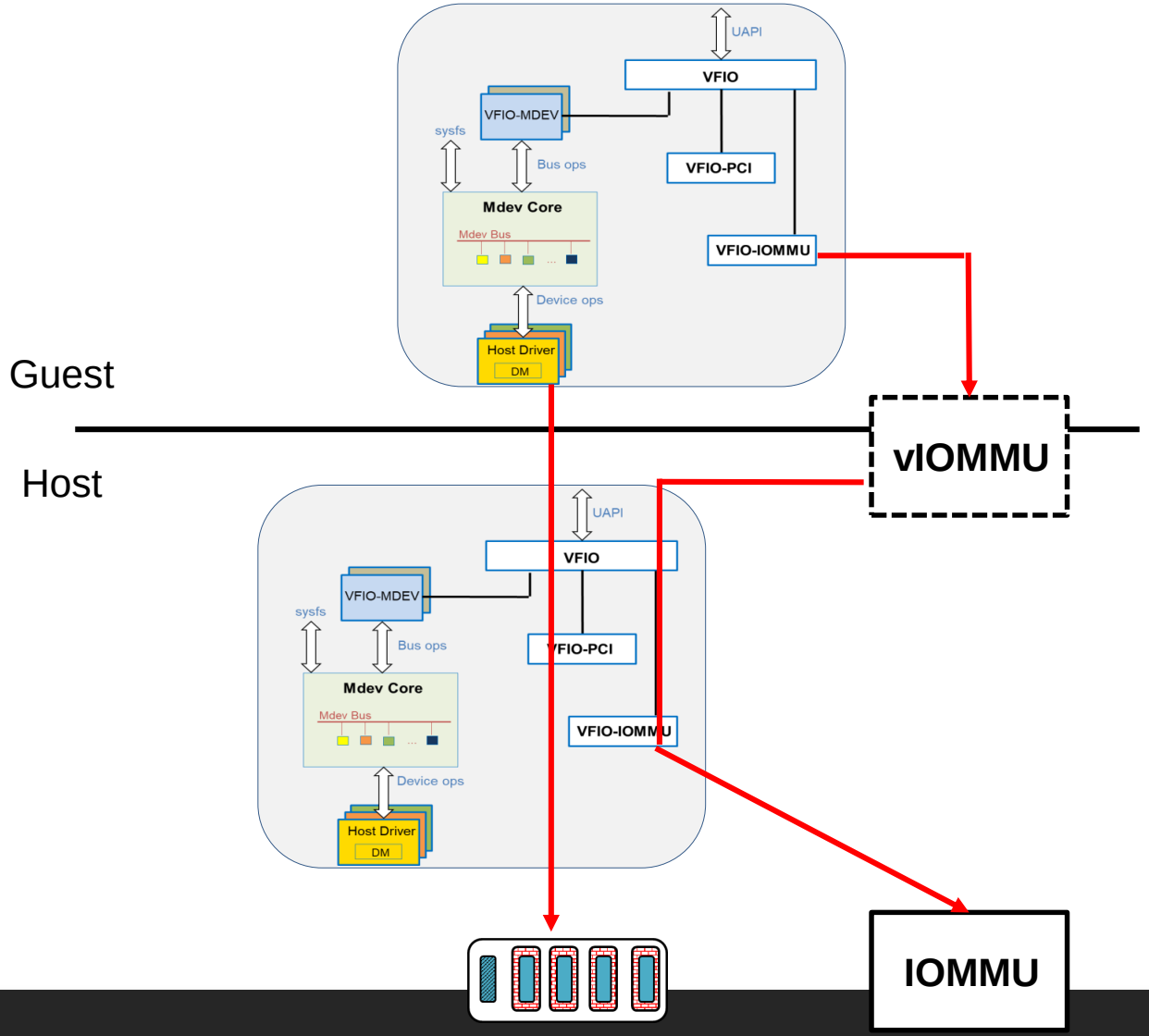
- Offloading device model to embedded CPU
  - Data path directly routed to embedded device
  - Control path mediated by embedded CPU
- Simplified host software stack
  - State/resource management through the embedded controller

Guest Mediation



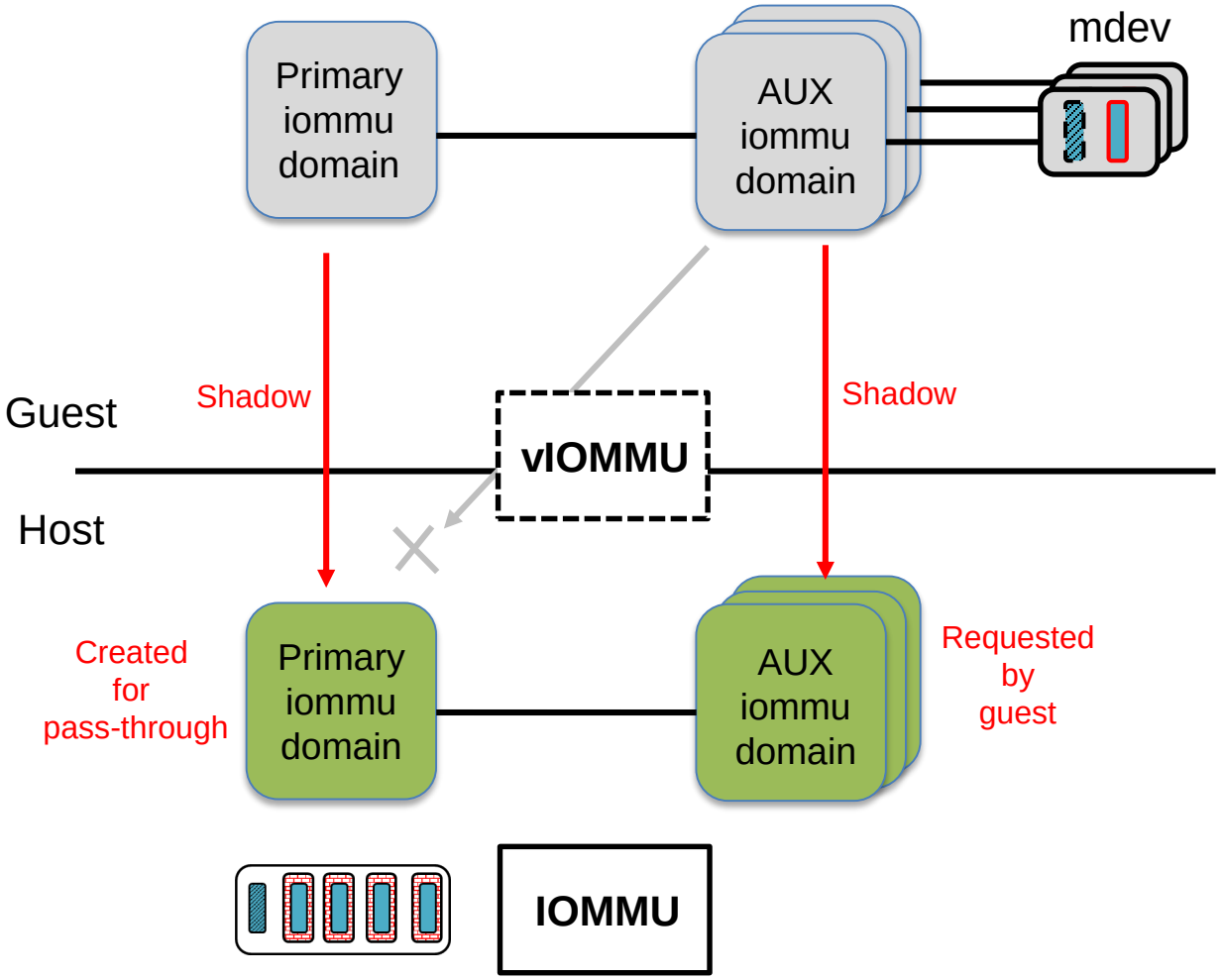
- Mediation on assigned PF/VF
  - One-level mediation
- Mediation on assigned mdev
  - Nested mediation
- Additional host support
  - Not required for software-based mediation
  - Required for hardware-assisted mediation

Hardware-assisted Guest Mediation



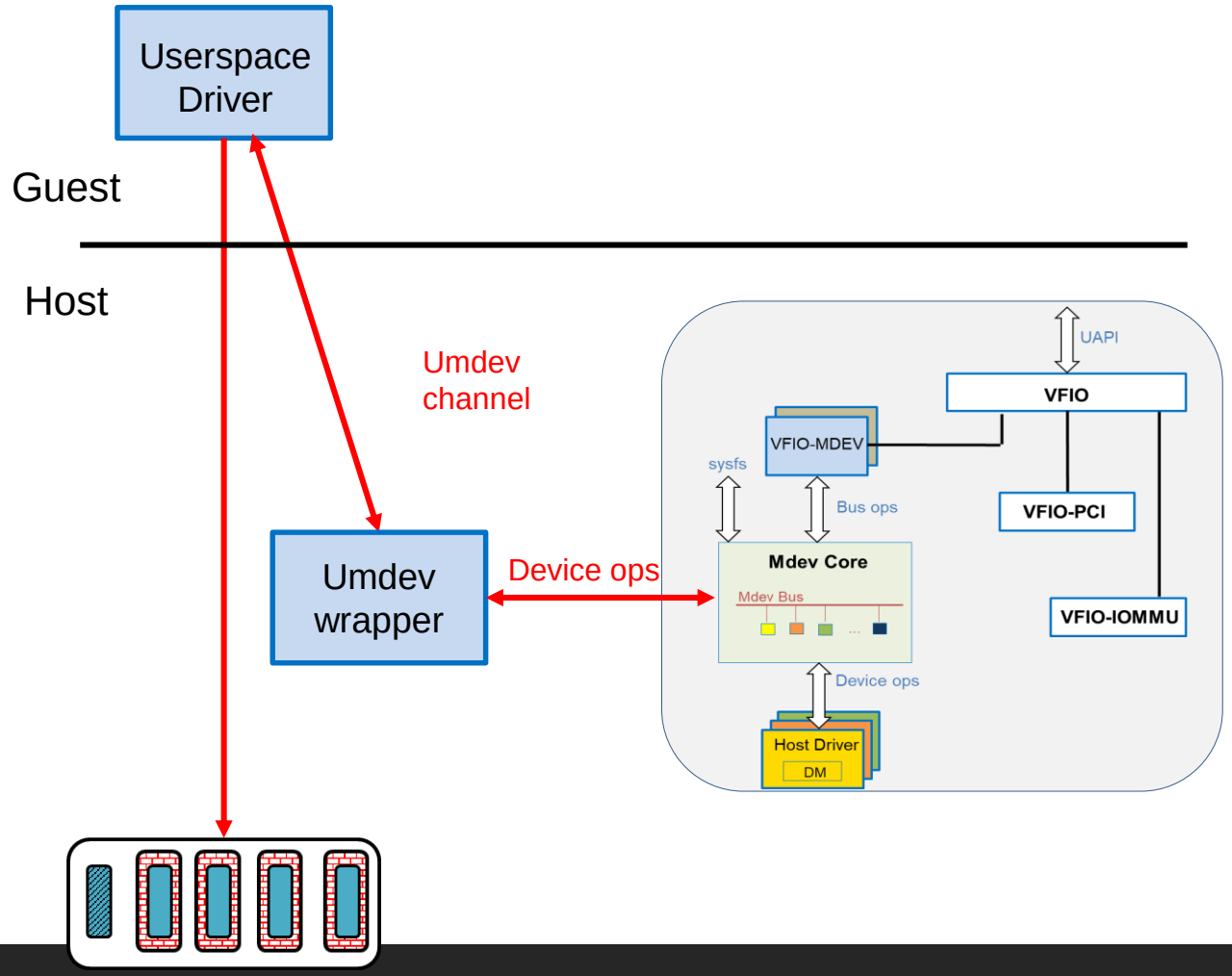
- Assign a capable PF to the guest
  - E.g. Intel ® Scalable IOV
- Expose a capable vIOMMU
  - E.g. Intel ® VT-d rev3.0
  - PASID-granular DMA isolation
- Sync vIOMMU to physical IOMMU
  - Nested translation, PASID management, page fault, etc.
  - Part of vSVA effort

Hardware-assisted Guest Mediation (Cont.)



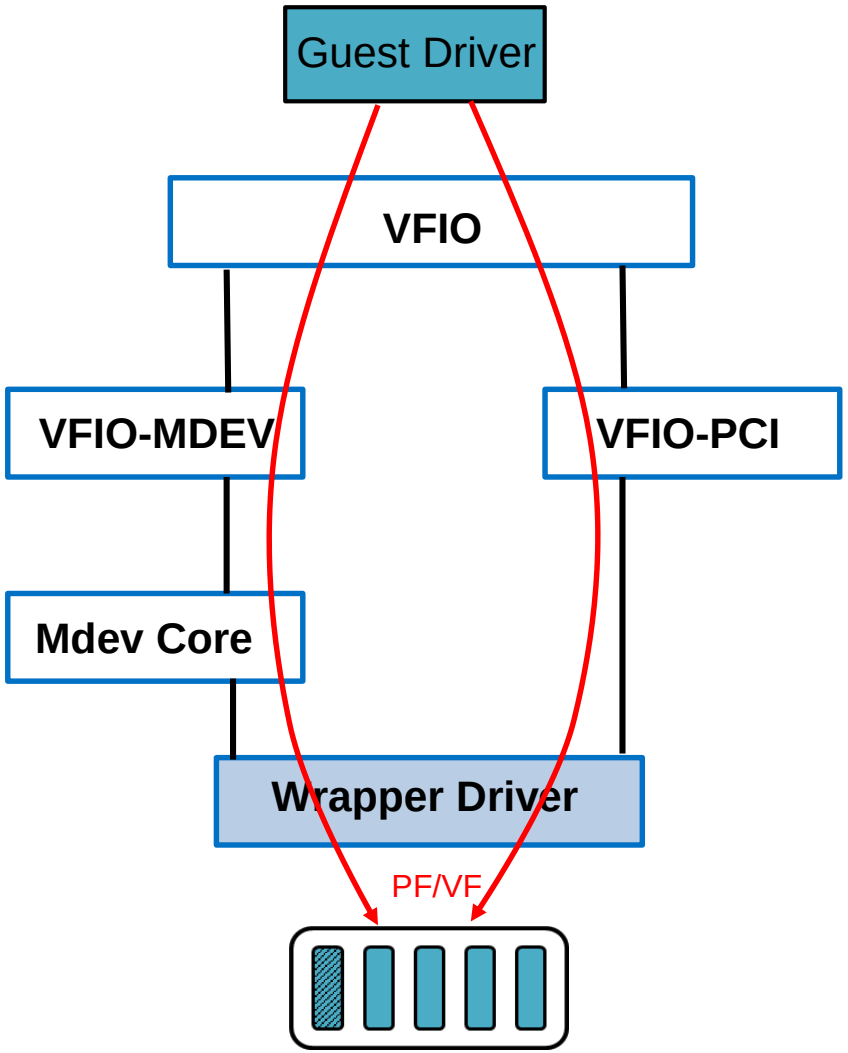
- Remains a gap on IOMMU domain
  - Host maintains only one domain for PF pass-through
  - However, guest creates multiple AUX domains on assigned PF
- Shadow guest AUX domains
  - Guest-initiated AUX domain management
  - Currently under internal exploration

Mediation in User Space



- Need a channel to connect userspace DM to the mdev core
  - When parent device driver is in user space
- Hardware-assisted userspace mediation
  - User-initiated AUX domain management
  - Verify ownership of the parent device
- Currently in prototyping

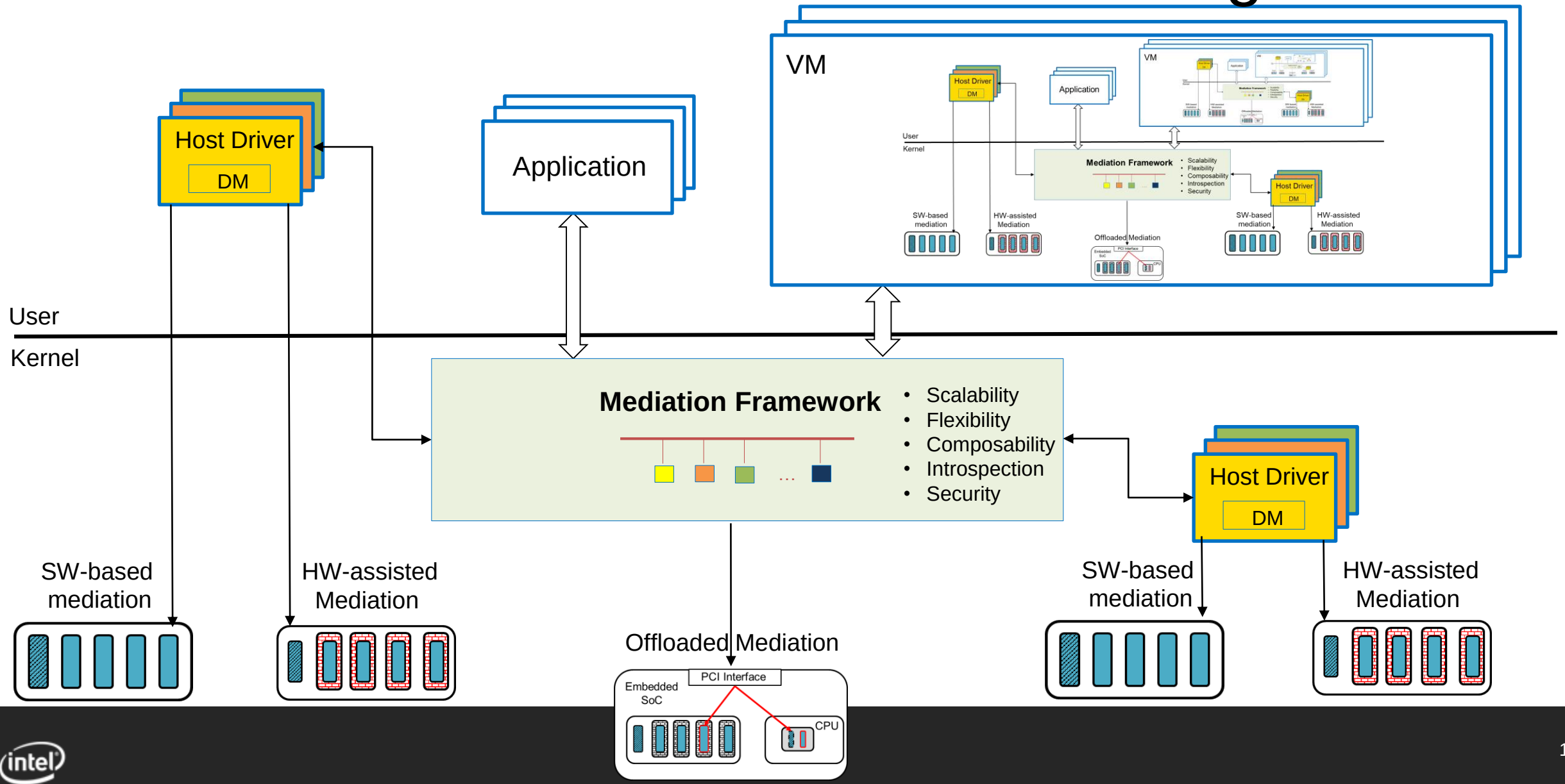
‘Mediated’ Direct Pass-Through



- Mediation wrapper driver for fixing limitations in direct pass-through
  - Live migration, generational compatibility, etc.
- Based on vfio-mdev
  - Wrap the device into a single mdev
  - Implement mdev_parent_ops
- Based on vfio-pci
  - Directly hook to vfio_pci_ops
  - Under discussion in mailing list

The Future

A World Built on Mediated Pass-Through



Q/A