Live Migration Support for GPU with SR-IOV

zhengxiao.zx@Alibaba-inc.com

Jerry.Jiang@amd.com

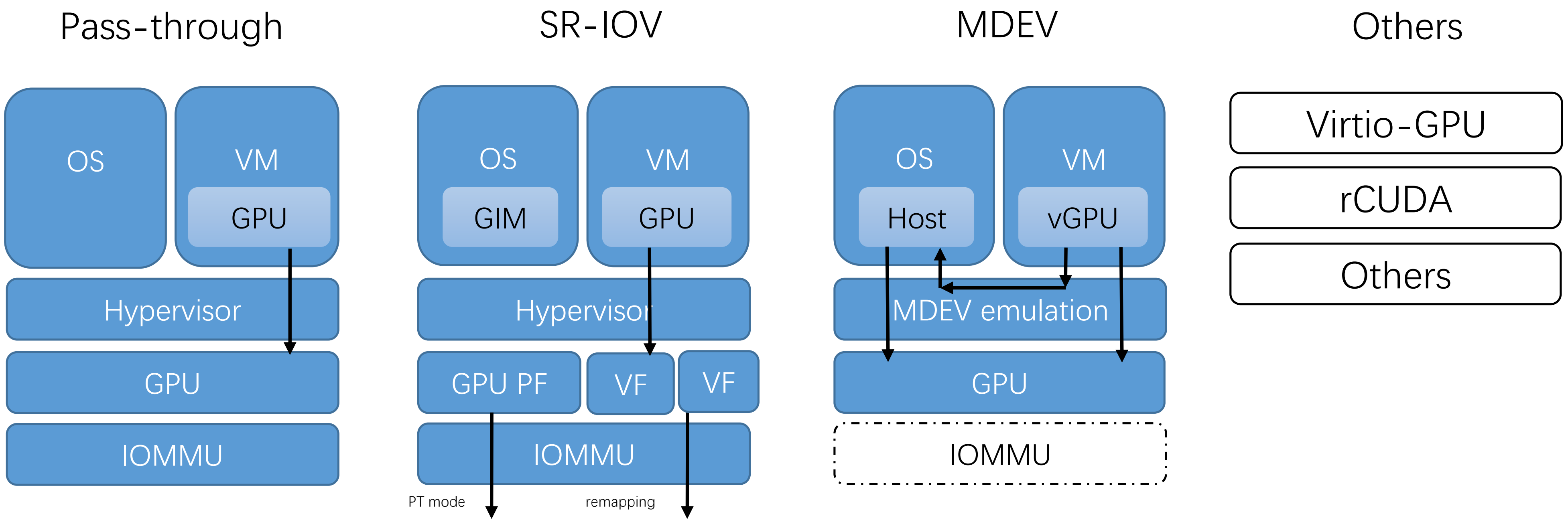
shuangtai.tst@alibaba-inc.com

Ken.xue@amd.com

Agenda – Hypervisor/QEMU

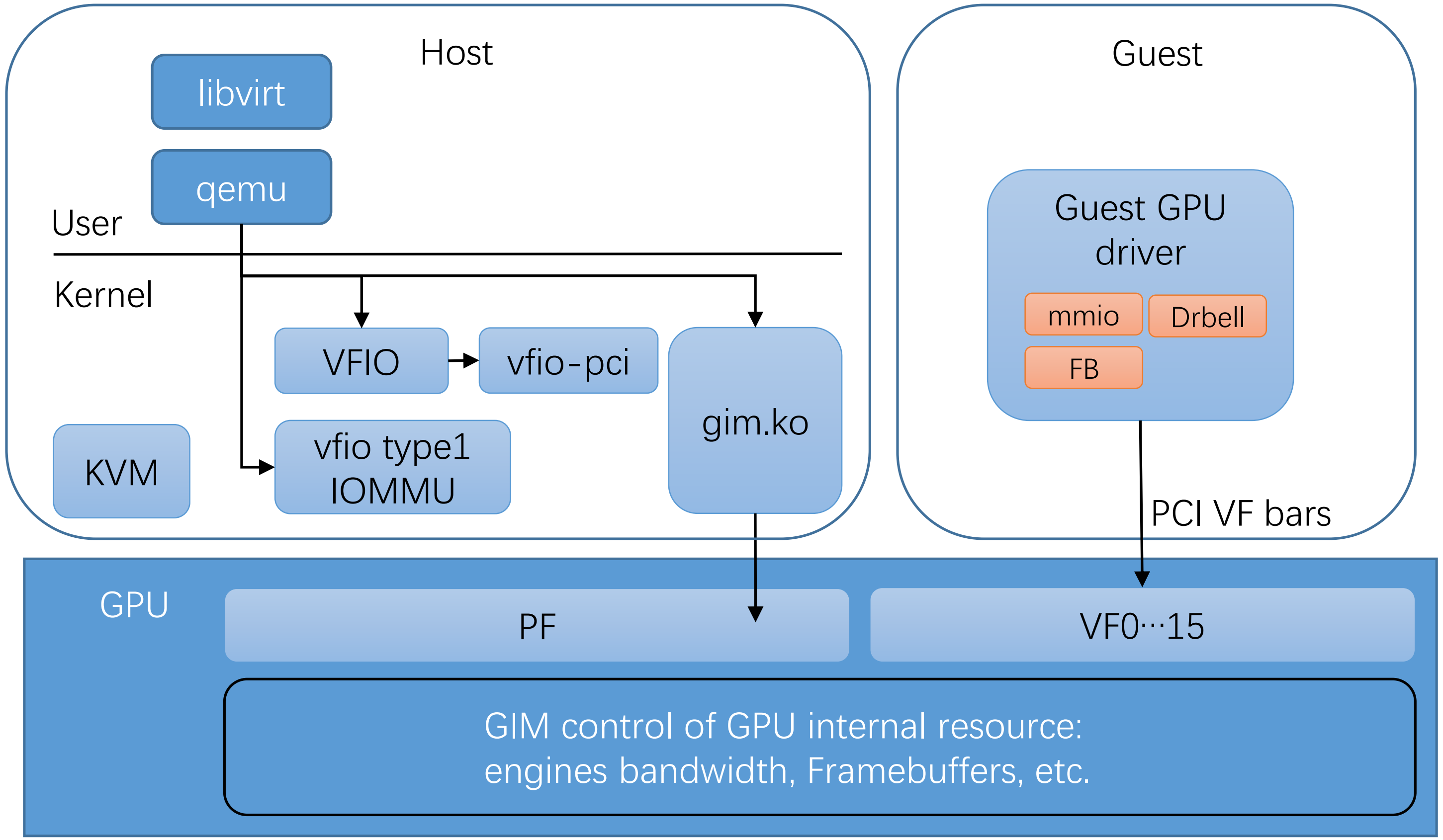
- GPU Virtualization Solutions
- SR-IOV GPU Virtualization Hypervisor View
- Current Migration Status
- Migration Sequence
- Challenge of Hypervisor's View

Common GPU Virtualization



- Full GPU capability, full featured
- High performance
- Production and commercialized
- Advance features: Live Migration support for SR-IOV and MDEV

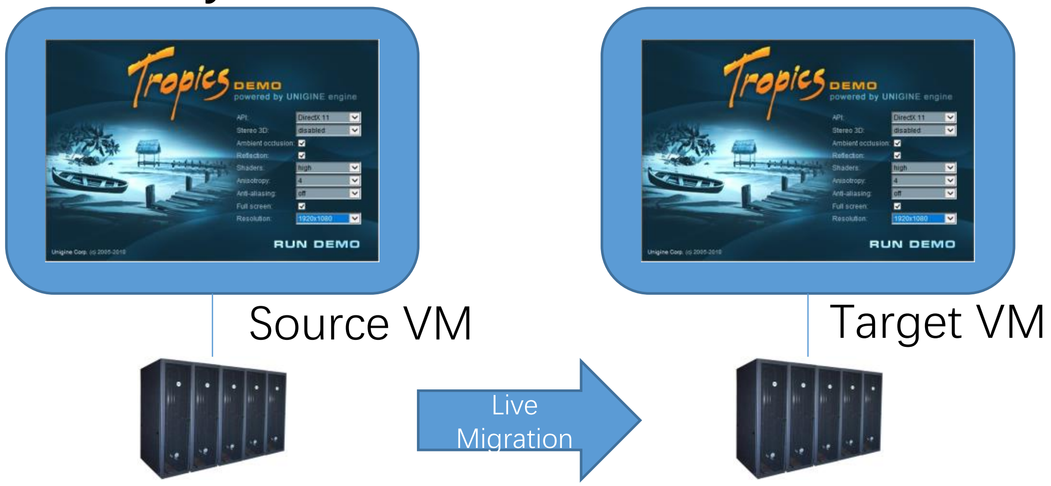
Virtualization of SR-IOV GPU



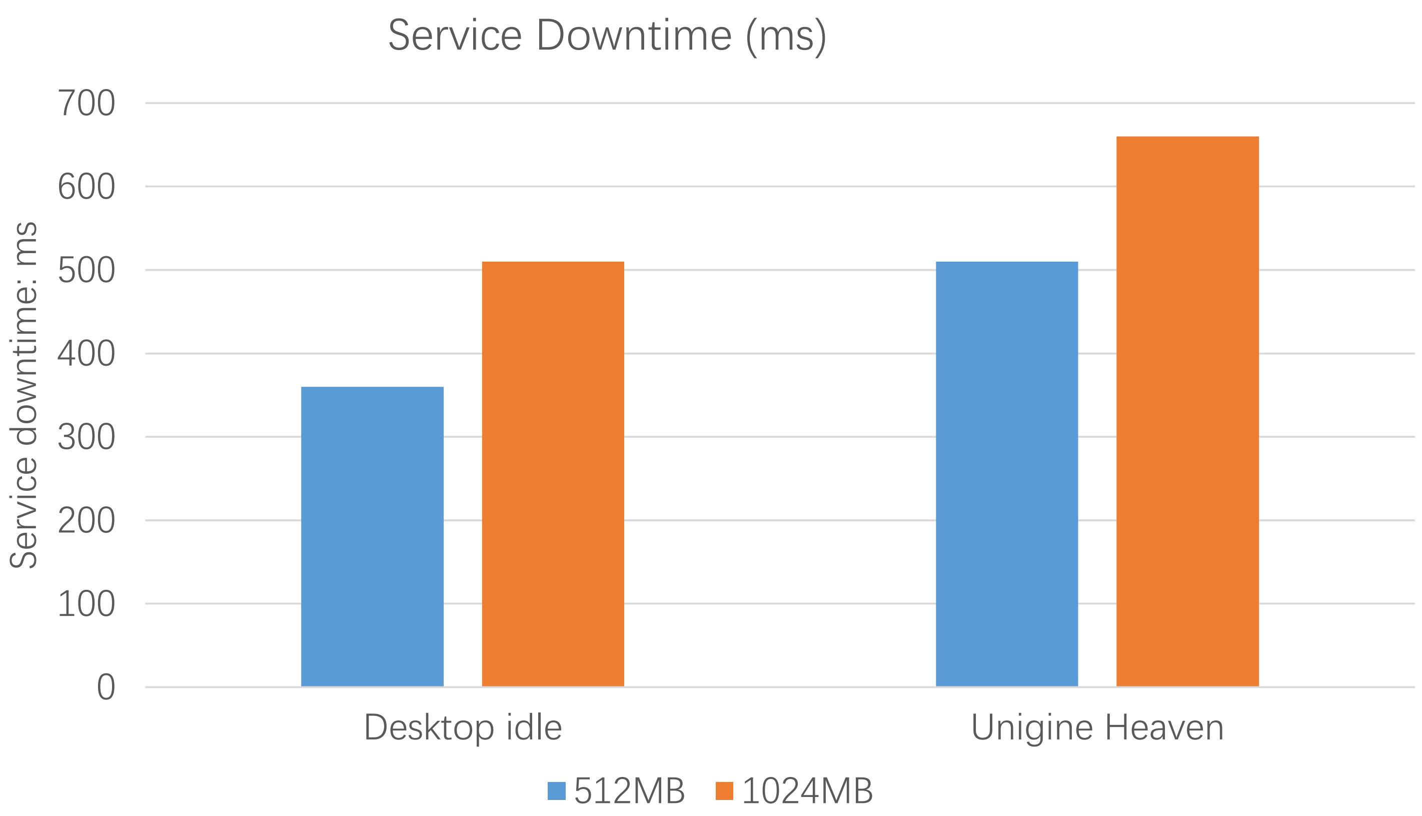
- PCIe PF/VF interface
- GPU graphics engine partitioned to support multiple VFs
- GPU video encoder engine partitioned to support multiple VFs
- Host driver (gim.ko) controls VF scheduling
- No display for Server GPU

Live Migration for SR-IOV GPU

- Collaborated between Alibaba Cloud Virtualization Team and AMD Virtualization Team
- Prototype solution based on AMD GPU MI25
- Support graphic 3D rendering migration
- Support planned for MM encoding engine migration in the future
- Support VM with SR-IOV VF checkpoint
- Service downtime: ~500ms with 1G graphic memory



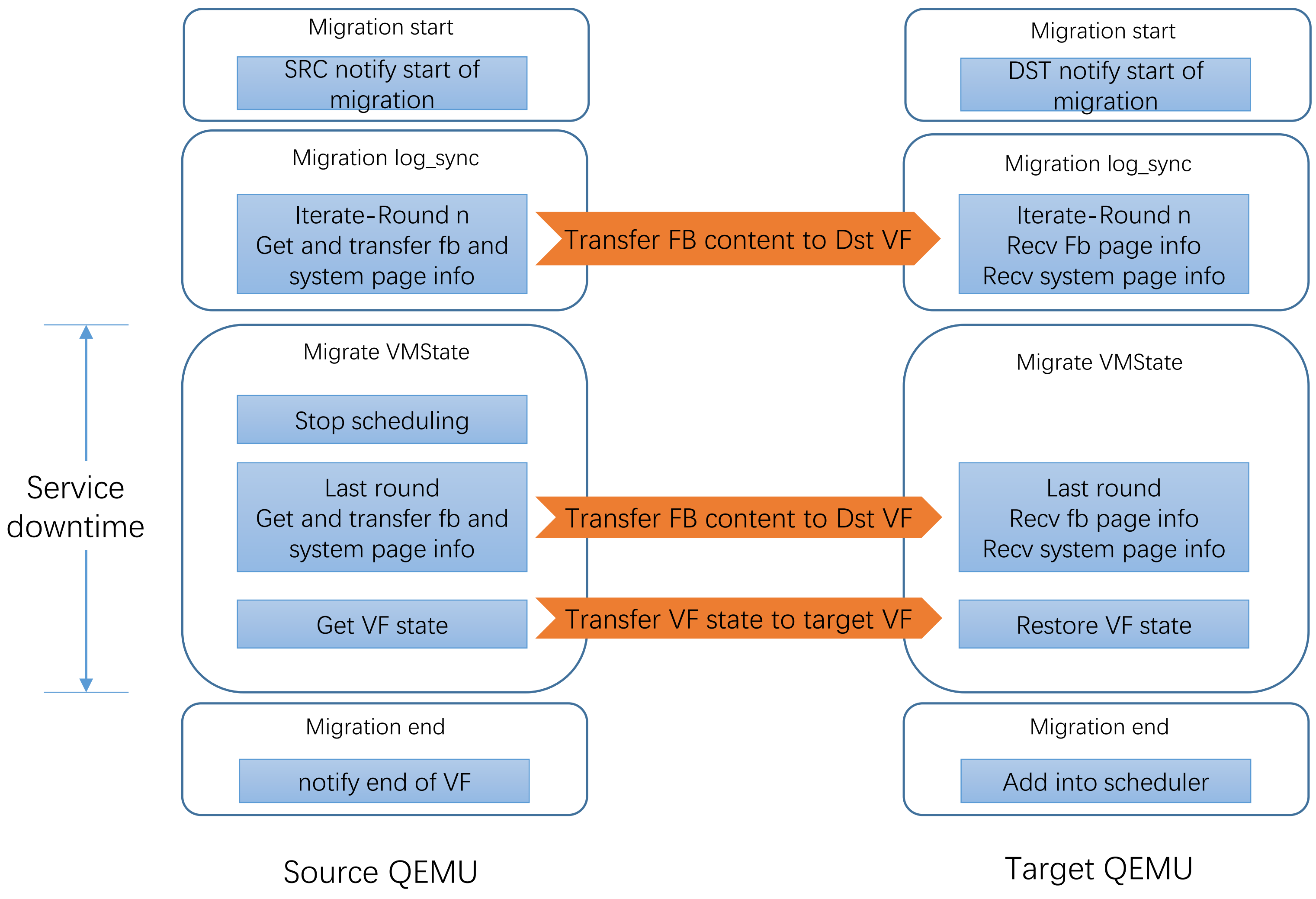
Evaluation Result



Guest Configuration:

- 8vCPU, 1GPU
- 2GB System RAM
- GPU FB: 512MB/1024MB

QEMU High Level Migration Sequence



Challenges

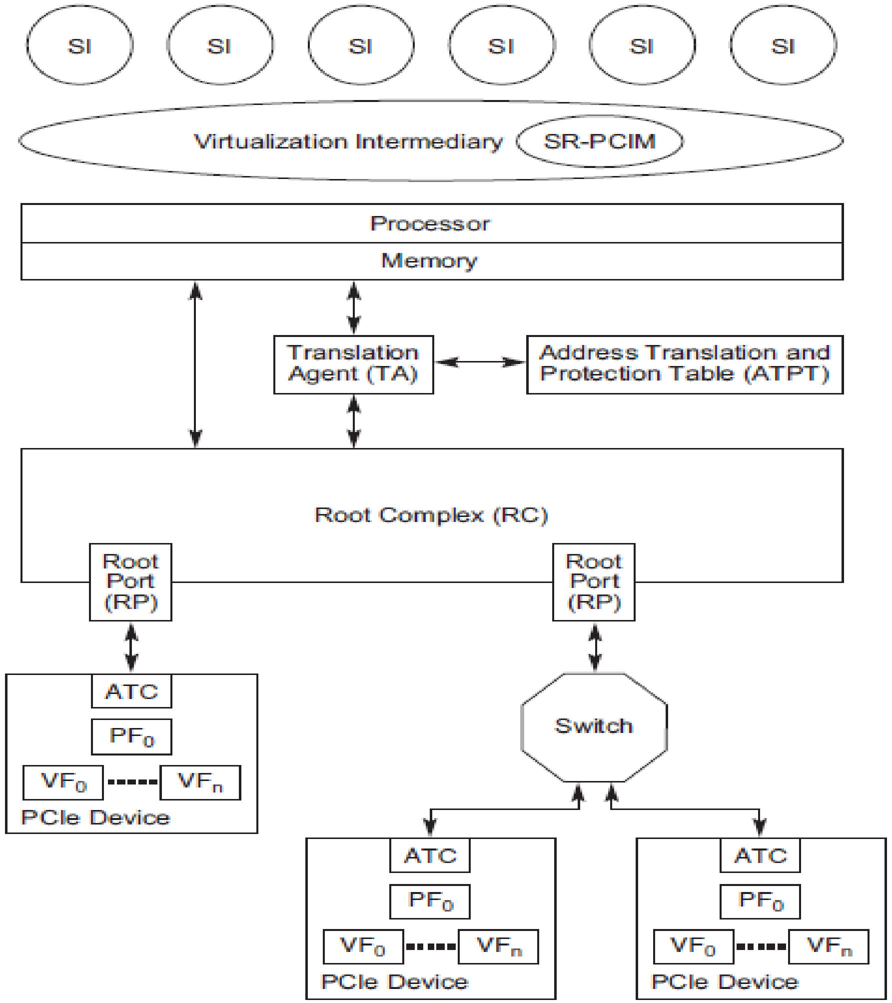
- Who should stop first: CPU or GPU
- Memory tracking
  - GPU -> system memory tracking
  - GPU -> Framebuffer tracking
- GPU workload preemption/World Switch
- GPU internal status migration
  - Page table
  - Interrupts
  - Context
  - Registers save/restore

Agenda – GPU

- SR-IOV Architecture
- SR-IOV SW Stack
- SR-IOV Advantage for VF Migration
- SR-IOV VF Migration
- Demo Video
- Challenge

Single-Root I/O-Virtualization (SR-IOV)

- Defines hierarchy of Physical Functions (PF) / Virtual Functions (VF) with a single root complex
  - Mix of PFs and VFs
- SR-IOV Capability Structure defines VF Capabilities associated with each PF
  - Each VF is uniquely addressable with RID
  - VFs have their own Configuration Spaces and Capability Structures
- PCIe endpoint is responsible for VF-PF scheduling and HW resource sharing
- SR-IOV is built on PCIe base spec v1.1 or later



❖ Source from SR-IOV spec 1.1
❖ SI: system image / virtual machine

Enhancement with AMD GPU SR-IOV

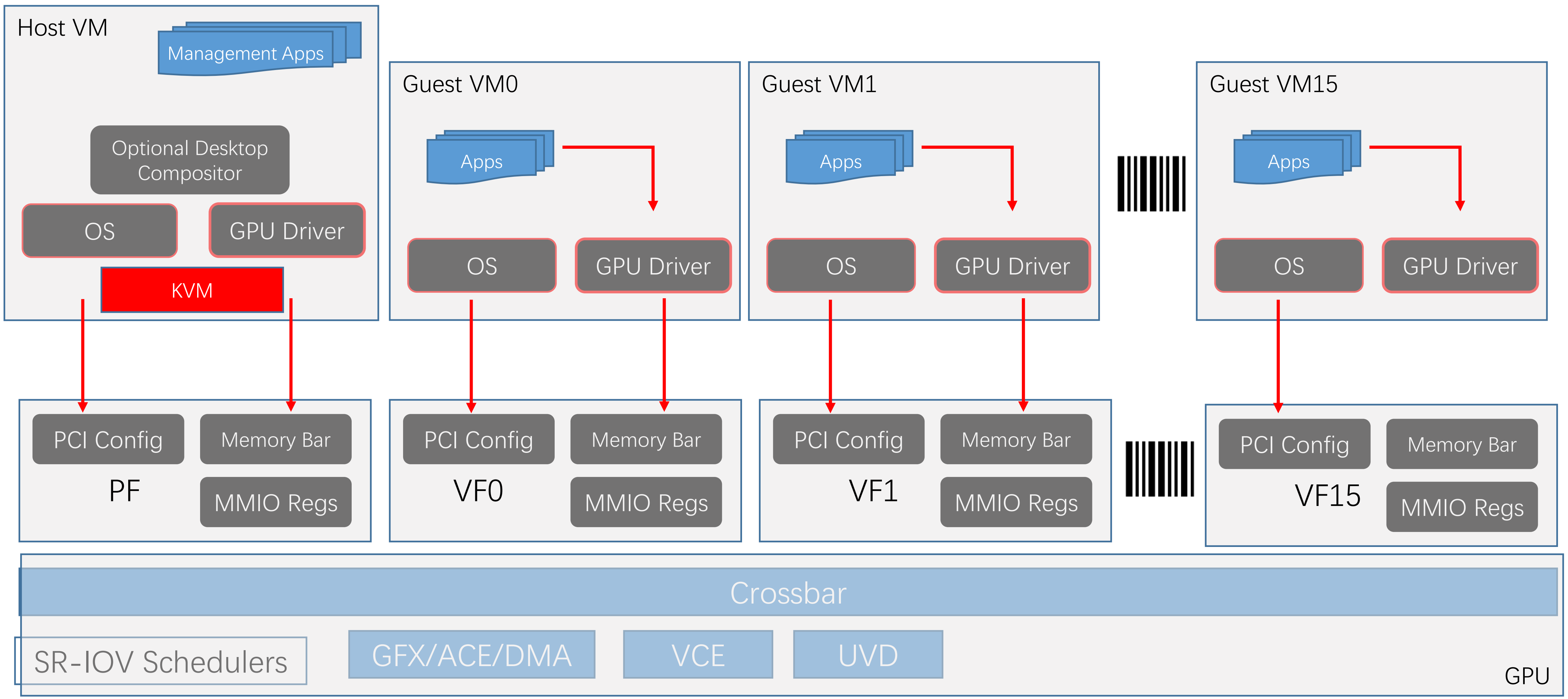
Support Existing virtualization architectures

- Uses standard PCI-SIG programming interface (with extension) to manage the GPU
- Light-weight vendor-specific driver in hypervisor/host
- VM uses unmodified OS and applications
- VM uses standard GPU driver
- Support of a host VM as a privileged managing partition

Hardware based virtualization functionality

- Support up to 16 Virtual Functions (VF) and 1 Physical Function (PF) on a single GPU
- Support up to 16 guests on a single GPU; more GPUs support more guests
- Exposes complete graphics and compute feature set of the GPU, e.g. D3D9/10/11/12, OpenGL®, OpenCL™, Vulkan®
- Support of H.264 and HEVC video encoders
- Performance enhancements
  - Guest GPU driver interfaces the GPU directly, eliminating software data copy overhead

AMD SR-IOV Solution on KVM Overview



SR-IOV Driver Stack

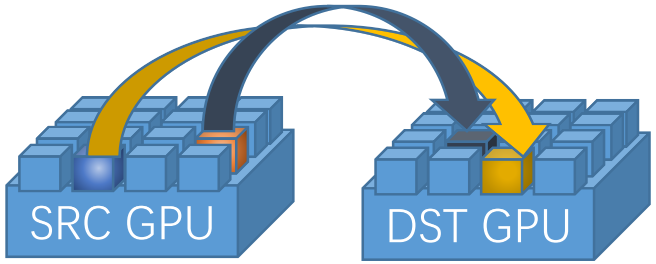
- SR-IOV GPU driver on PF
  - GPU PF initialization
  - Error detection and virtual function functional level reset (FLR)
  - Live migration operation
- Guest GPU Driver on each VF:
  - Owns independent engine ring buffers, interrupt vector, dedicated FB, doorbell, GPU state, GPU VM, etc.
  - Interacts with GPU hardware through VF GPU resources
  - Guest driver is unaware of world switch during run time
  - Supports Direct3D, OpenGL®, OpenCL™, Vulkan®, Rocm APIs
  - Support of remote display software such as Windows® RDP, Horizon View, Teradici, HDX3D, etc.

Advantage of SR-IOV VF Live Migration

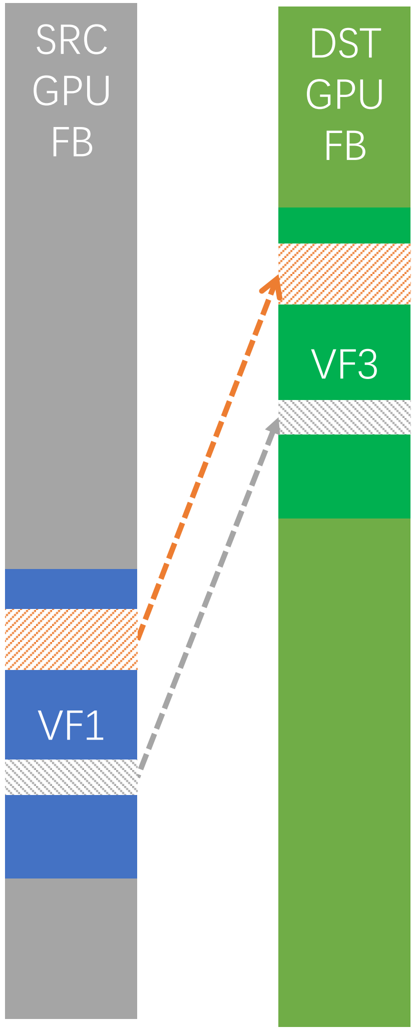
- Normal passthrough
  - Host/Hypervisor has difficulties to
    - Stop passthrough device
    - Get source device state
    - Copy source device dedicate memory content
    - Restore – reinitialize target device state after migration
    - Restore target device dedicate memory content
    - Notify guest device driver to perform all above actions
- SR-IOV GPU driver on PF is able to
  - Take snapshot of VF’s FB
  - Take snapshot of VF’s GPU state
  - Control (stop) VF’s running time slice
  - Restore snapshot of target VF FB content
  - Restore snapshot of target VF GPU state
  - Guest VM seamless migration of 3D rendering services
    - Pre-emptied GPU command will resume after migration
    - Suspended interrupt will resume

VF Migration

- On source GPU, GPUV driver
  - Stop scheduling any time slice to source VF
  - Copy source VF FB content to system memory
  - Save source VF GPU state to system memory
  - Copy source VF non-local memory to system memory
- On target GPU, GPUV driver
  - Copy to VF FB from system memory
  - Restore VF GPU state from system memory
  - Restore VF non-local memory from system memory
- After migration, guest VM
  - No re-initialization
  - Pre-empted commands continue to run



VF GPU State Migration



VF FB Migration

Demo show or Video

One system with two Mi25

One VM with VF migrate between VF2 on GPU0 and VF2 on GPU1

Guest VM continue to run Unigine Heaven

An app shows the device BDF info

VF Migration Challenge

- Source VF FB dirty page tracking
  - GPU HW tracks all the dirty page between two sequential queries
- Source VF non-local memory dirty page tracking
  - IOMMU tracking – current CPU doesn't support; future CPU will add support
  - QEMU – tracks the no-local location – no accurate
  - Driver tracking – driver go through all VF page table - takes time or additional shadow page list
- GPU compatibility checking
  - Different GPU generation – not supported – provide a compatibility API to check
  - Different FW version might not be compatible
  - Different guest GFX driver version might not be compatible
  - Hypervisor/QEMU migrates the VM between compatible source VF – target VF

Summary - SR-IOV GPU Virtualization

- PCIe compliance - natively fit into existing KVM architecture
- Enhanced Security – VF resources and VF GPU states are isolated by GPU hardware
- Low TCO (total cost of ownership) – better GPU HW resource utilization by partitioned GPU
- Live migration friendly – APIs on host for QEMU to manage the migration
- Complete graphics and compute feature set in guest VM

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION