基于MDEV的国产显卡虚拟化方案

邓林文
景嘉微虚拟化技术专家

目录

云桌面GPU虚拟化要解决的问题

业务的连续性，本地的性能，云上的灵活性

兼容性

应用兼容性（API）
平台兼容性（指令集）
硬件兼容性

性能

渲染性能（跑分）
编码推流性能（显示效果）

灵活性

资源灵活分配（套餐规格）
资源共享（实例密度）
方便云厂商对接（管理接口）

云桌面GPU虚拟化方案

Virtio

使用广泛，硬件要求低
性能差，应用兼容性差

QXL,VirtioGPU/Virgl

VFIO透传

性能好
不支持共享，可扩展性差

VFIO-PCI

SR-IOV

性能好，支持多实例
硬件实现复杂，灵活性差

AMD S7150

MDEV

性能好，支持多实例，配置灵活
软件复杂

NVIDIA GRID, Intel GVT

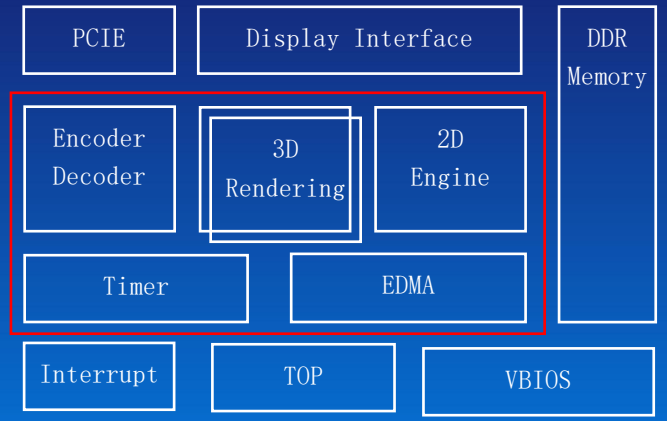
目录

基于MDEV的GPU虚拟化架构

GPU硬件

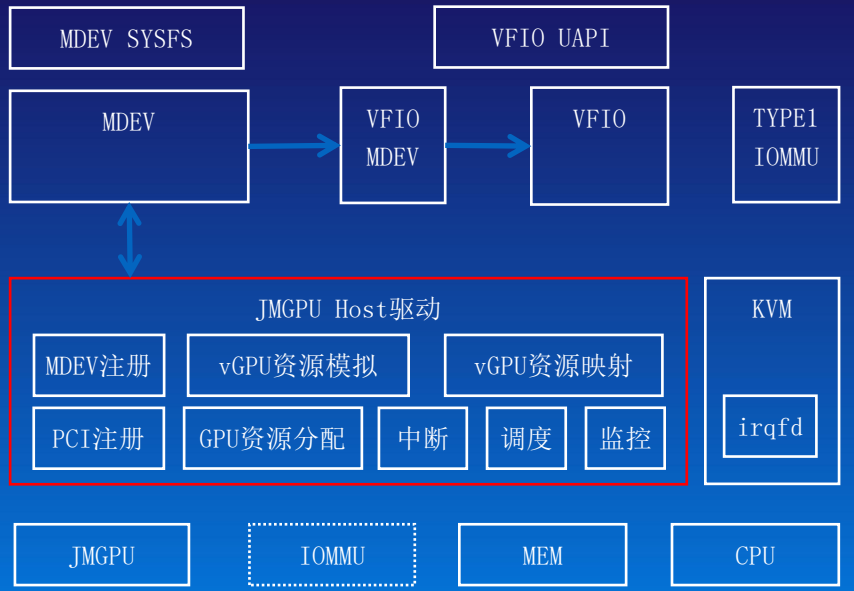
GPU是一个复杂的SoC系统

- 功能模块
  - 2D Engine、3D 、EDMA、编码器、解码器
- 控制模块
  - 中断、TOP...
- 接口模块
  - PCIE
  - 显示接口
- 帧存（DDR）



基于MDEV的GPU虚拟化Host驱动

- VFIO（Virtual Function I/O）
  - 用户态直接访问设备的驱动框架
- MDEV（Mediated Device）
  - 复用VFIO框架，支持多个实例共用一个设备
- JMGPU Host驱动
  - GPU虚拟化具体实现



HOST虚拟化驱动实现

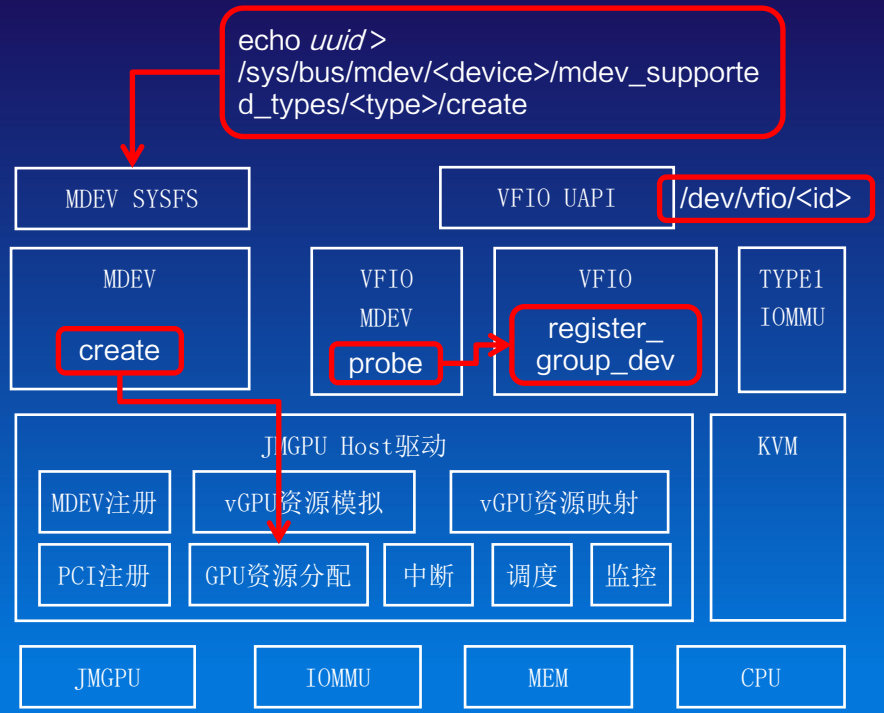
vGPU规格定义

- 每种vGPU规格对于着不用的资源分配方式
  - 帧存大小
  - 最大分辨率
  - 渲染核心（2D、3D）数量
  - 编解码核心数量
  - 显示模式
- 外部接口
  - /sys/bus/mdev/<device>/mdev_supported_types

HOST虚拟化驱动实现

vGPU创建

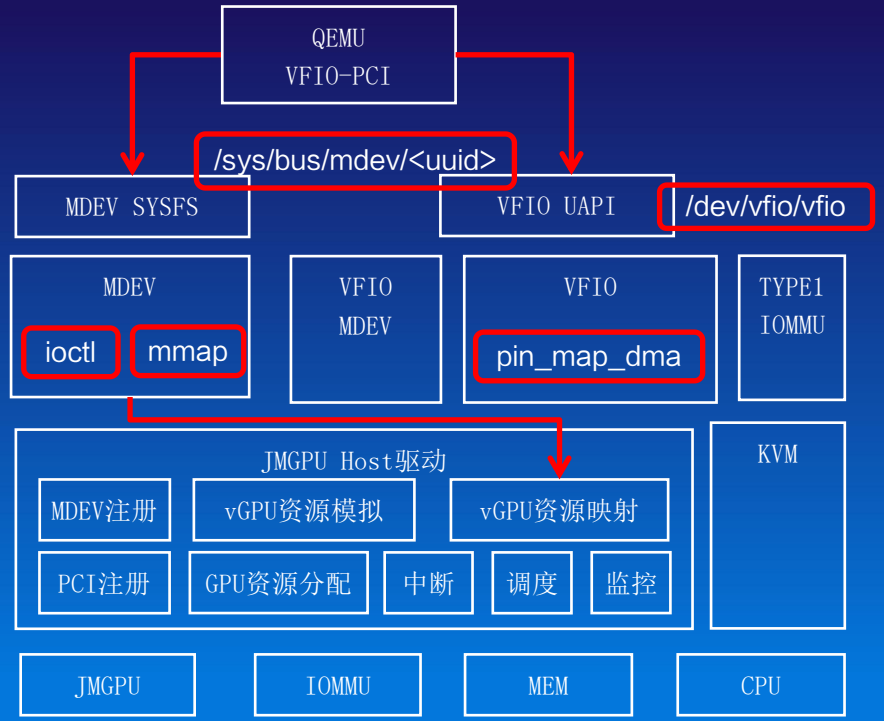
- 根据规格创建相应的vGPU资源
  - PCI配置空间
  - 帧存
  - 渲染资源（3D/2D）
  - 中断
  - 显示
  - …
- vGPU创建成功后,mdev设备节点可被Qemu访问
  - /sys/bus/mdev/<device>/<UUID>



HOST虚拟化驱动实现

vGPU 资源映射

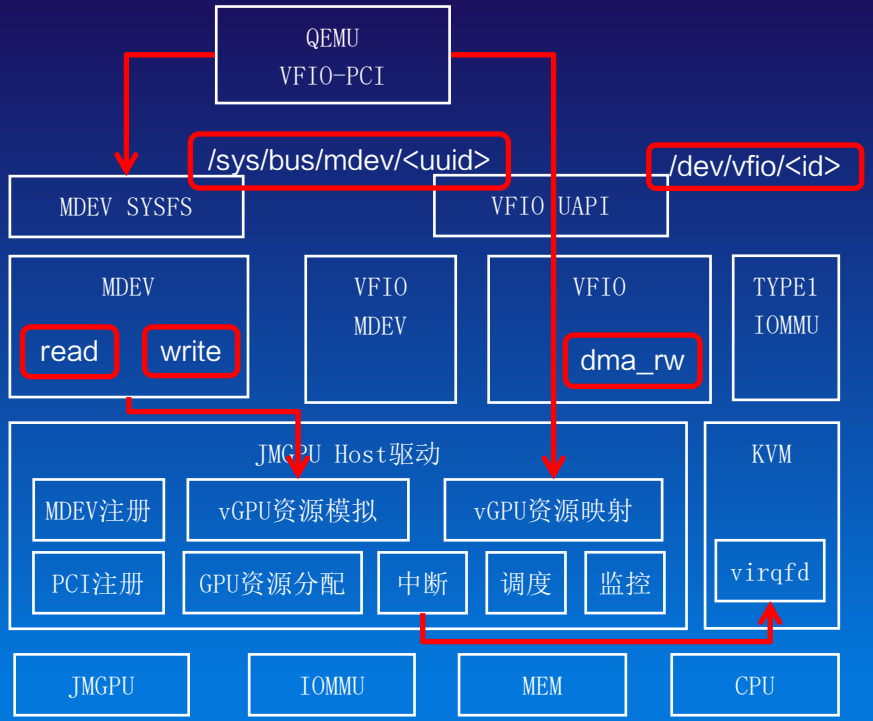
- 虚拟机创建时，通过MDEV IOCTL获取并配置vGPU基本信息
  - GET_INFO
  - GET_REGION_INFO
  - GET/SET_IRQ
  - QUERY_GFX_PLANE
- 通过VFIO IOCTL，将虚拟机的GPA与HPA建立映射
  - IOMMU_MAP



HOST虚拟化驱动实现

vGPU 使用

- 虚拟机访问vGPU资源
  - PCI配置空间读写
  - 寄存器读写
    - TRAP
    - MMIO 直接映射
    - 内存映射
  - 帧存访问
  - DMA
  - 中断



HOST虚拟化驱动实现

vGPU 复位

- 虚拟机关机或销毁后，vGPU后端资源需要进行复位
  - PCI复位
  - 功能模块复位
- 基本要求
  - 不影响其它vGPU
  - 硬件状态能恢复

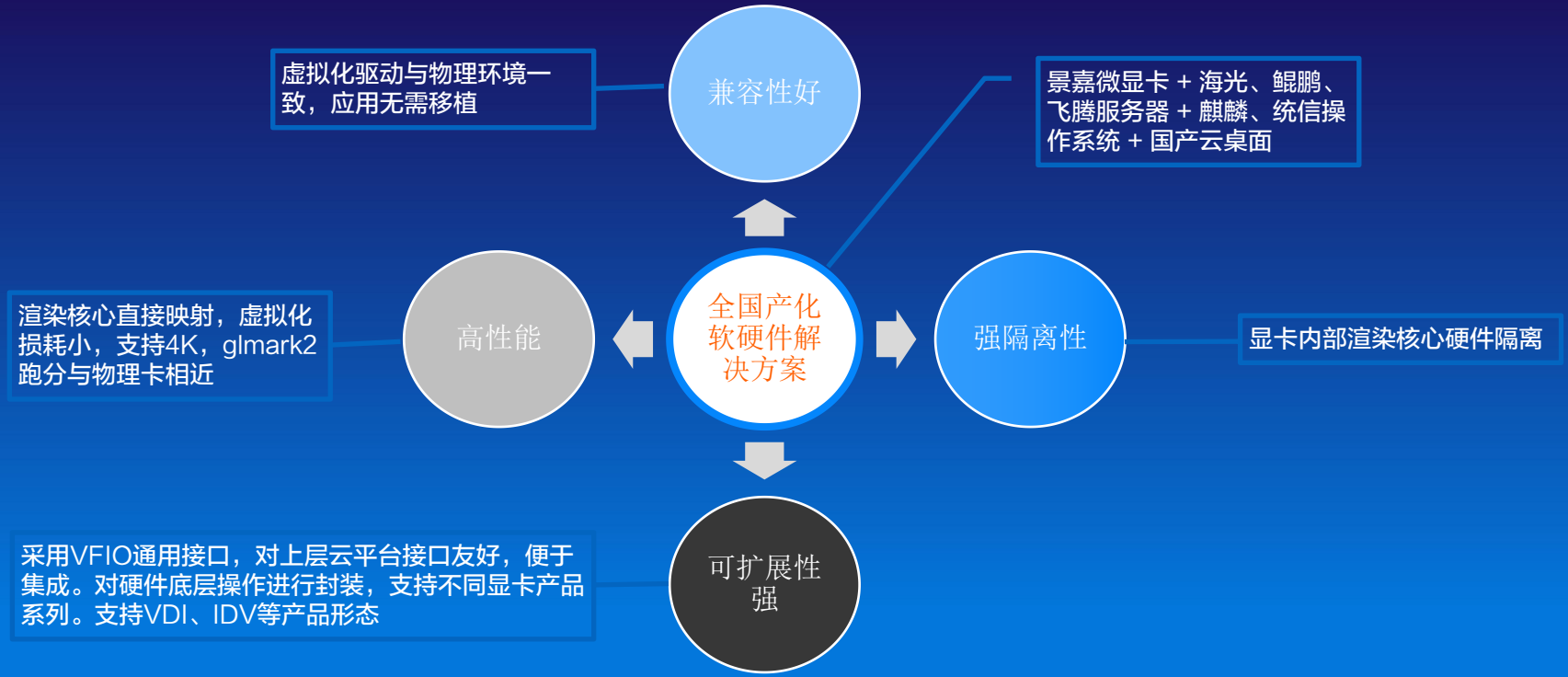
GUEST虚拟化驱动实现

驱动应与物理机保持“一致”，保证应用的兼容性

- 内核态驱动为保证性能，需要进行一些半虚拟化处理
  - 性能损耗：trap > memcpy > mmap
  - 独占的资源用mmap
  - 共享的控制模块用trap
  - 共享的功能模块用memcpy
- 虚拟显示
  - 不需要对物理显示控制器进行配置（KMS）
  - 需要独立接口传递EDID信息
- 监控
  - 多实例共享资源，需要对采样方法和数据进行额外处理
- 特权模块
  - 中断控制器的中断状态与功能模块内部的中断状态需要一致
  - 复位，

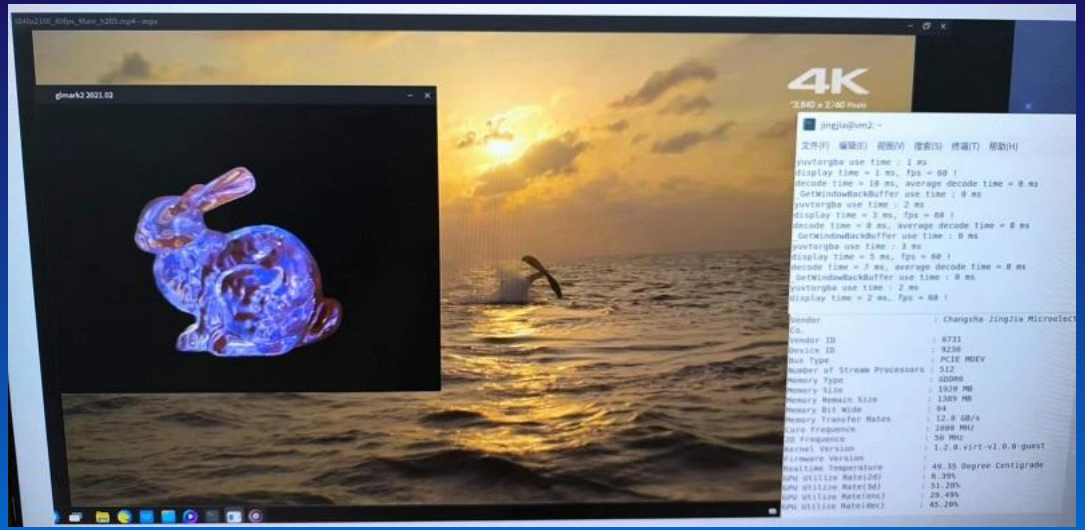
目录

基于MDEV的GPU虚拟化产品优势

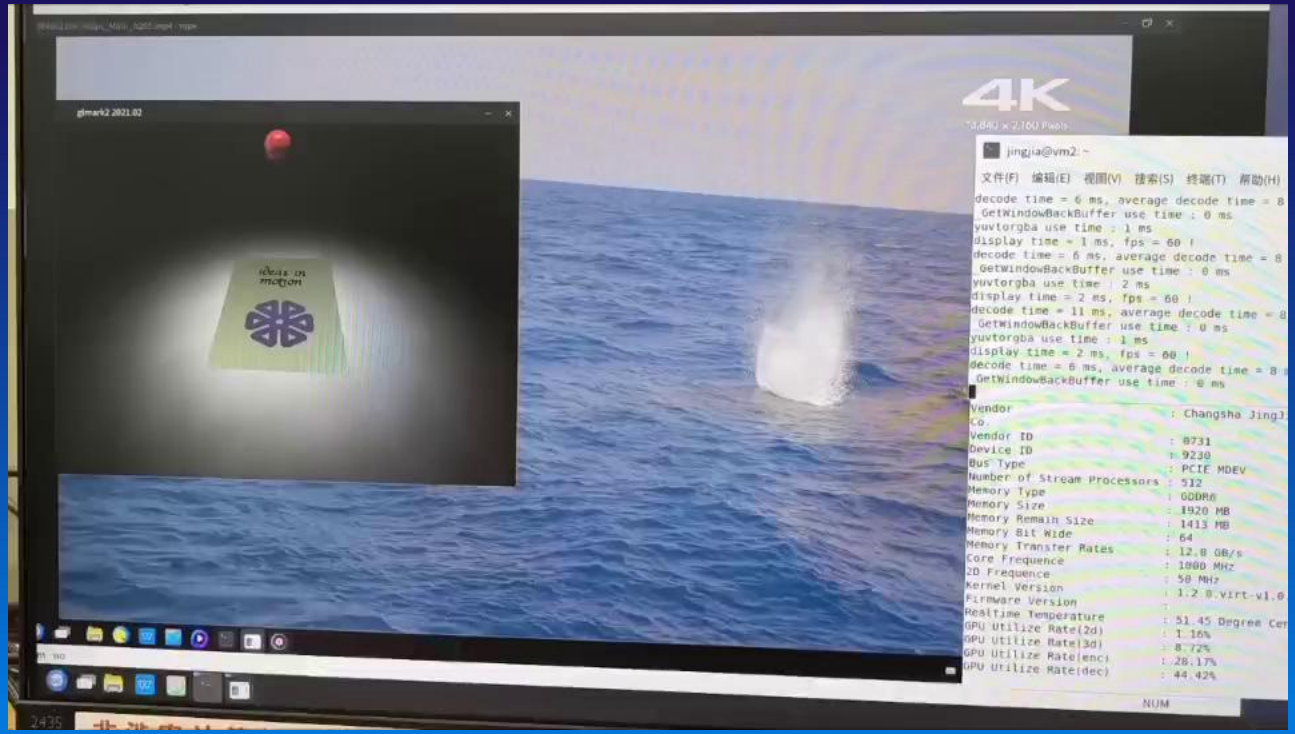


基于MDEV的GPU虚拟化效果

- 测试环境
- 服务器
  - 鲲鹏920 + 景云1号 + KylinV10 Server
  - Qemu + Spice-server
- 虚拟机
  - 4C8G + 1/2 vGPU + KylinV10 Desktop
  - 虚拟机内抓屏编码（SDK）
  - 单卡最多支持8路虚拟化
- 客户端
  - 飞腾D2000 + JM9100
  - Spice-client
- 用例
  - 3D测试：glmark2
  - 视频播放：4K@60 FPS H.265
  - 抓屏编码：1920x1080@60 FPS H.264



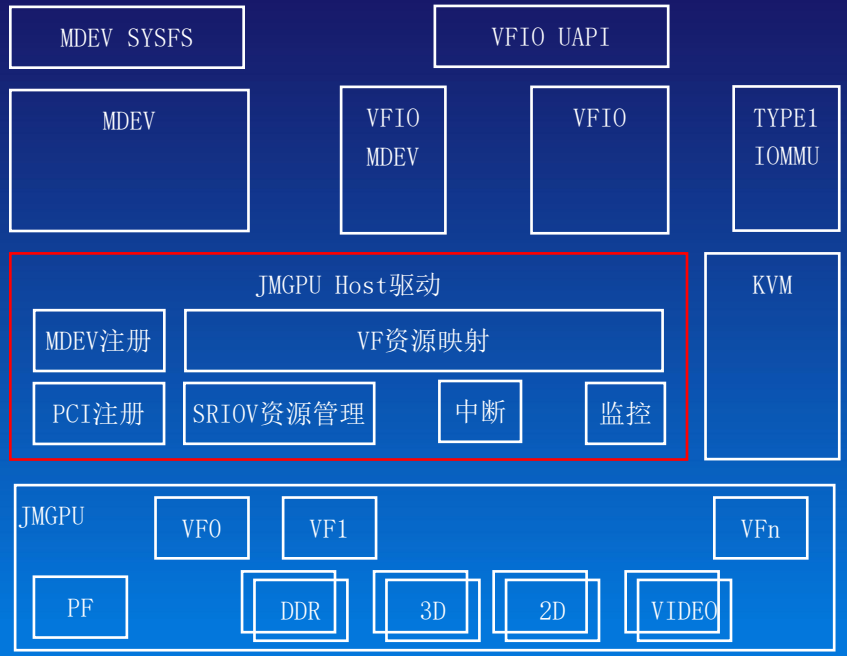
基于MDEV的GPU虚拟化效果



目录

后续规划

MDEV + SRIOV，MDEV设备与VF 1v1绑定，性能与灵活性兼顾



主机接口	PCIE 4.0 x16
核心频率	1.2 Ghz
显存容量	32GB DDR4/LPDDR4
高清编码	8路4K@60fps编码，32路1080@60fps编码
高清解码	16路4K@60fps解码，64路1080@60fps解码
API接口	OpenGL4.6/OpenCL3.0/Vulkan1.2/DX11
硬件虚拟化	SR-IOV, 最大支持32路
典型功耗	150W

Thanks_

景美公众号