开放原子开源基金会 OPENATOM FOUNDATION | OpenEuler

灵衢设备虚拟化关键技术和应用

叶镖翔 openEuler社区Virt SIG Committer

基于灵衢支持多种组件资源池化，灵活构建逻辑上的计算机

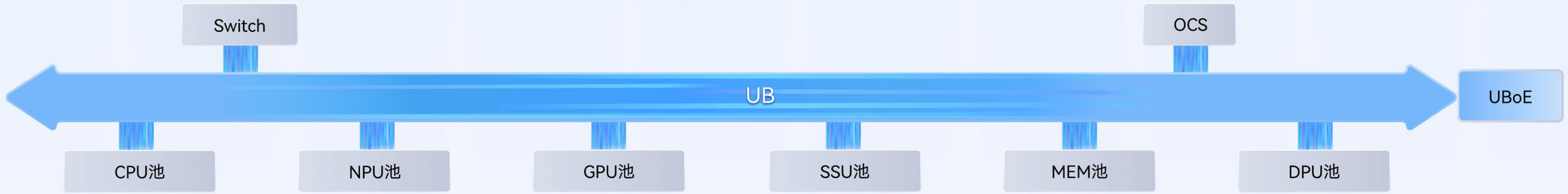
基于灵衢的超节点参考架构: 10 大场景,按需组合,构建各自场景参考架构

大模型预训练	中心推理	后训练与强化学习	多模态内容理解与生成	Agentic AI	虚拟化	大数据	数据库	分布式存储	高性能计算
超节点参考架构	超节点参考架构	超节点参考架构	超节点参考架构	超节点参考架构	超节点参考架构	超节点参考架构	超节点参考架构	超节点参考架构	超节点参考架构

六大共性特征

总线级互联 | 协议归一 | 平等协同 | 全量池化 | 大规模组网 | 高可用性

基本原则：基于UB统一互联



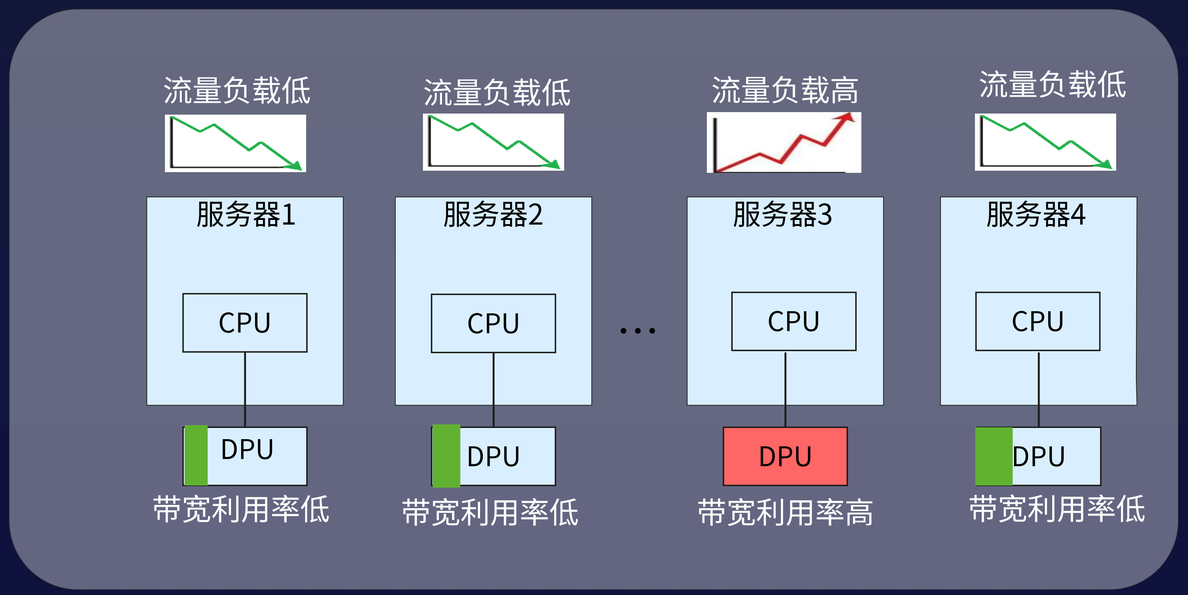
注：OCS：Optical Circuit Switch; UBoE: UB over Ethernet

灵衢

设备资源利用率挑战

问题场景

单服务器资源固定配比，业务资源静态配置，难以满足各业务对算力、IO等资源的使用诉求，DPU资源利用率低

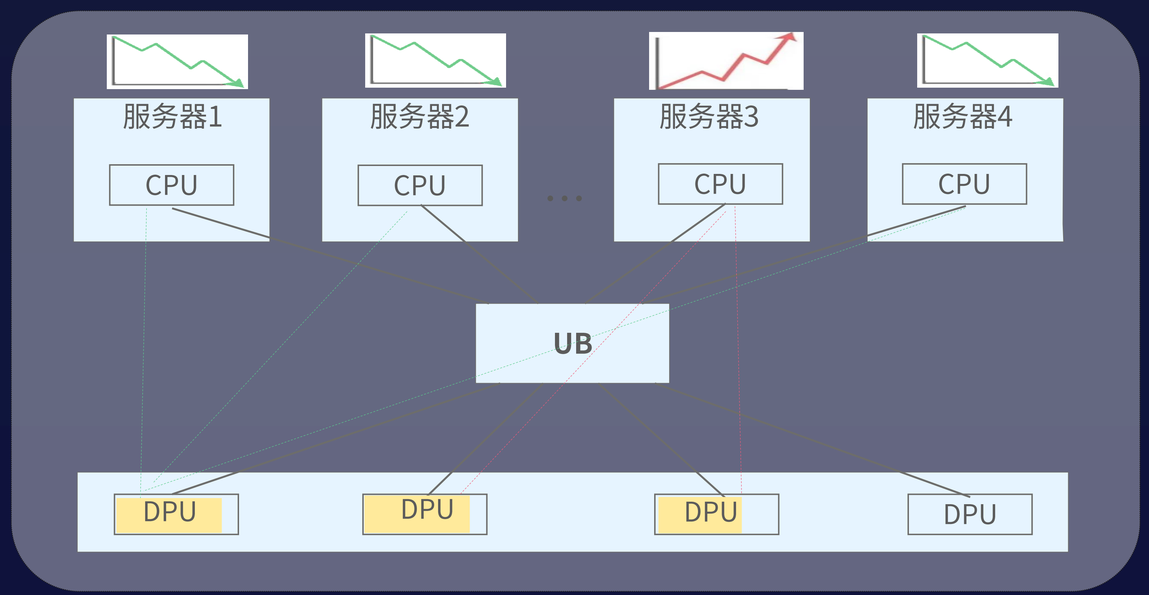


资源使用不均

- **DPU**平均带宽利用率**<20%**，算力利用率为**20%~30%**
- 部分业务(如DB类业务)对带宽诉求超过当前**DPU带宽上限**

解决方案

基于UB实现设备资源池化共享，按需动态共享使用DPU，融合使用提升DPU利用率

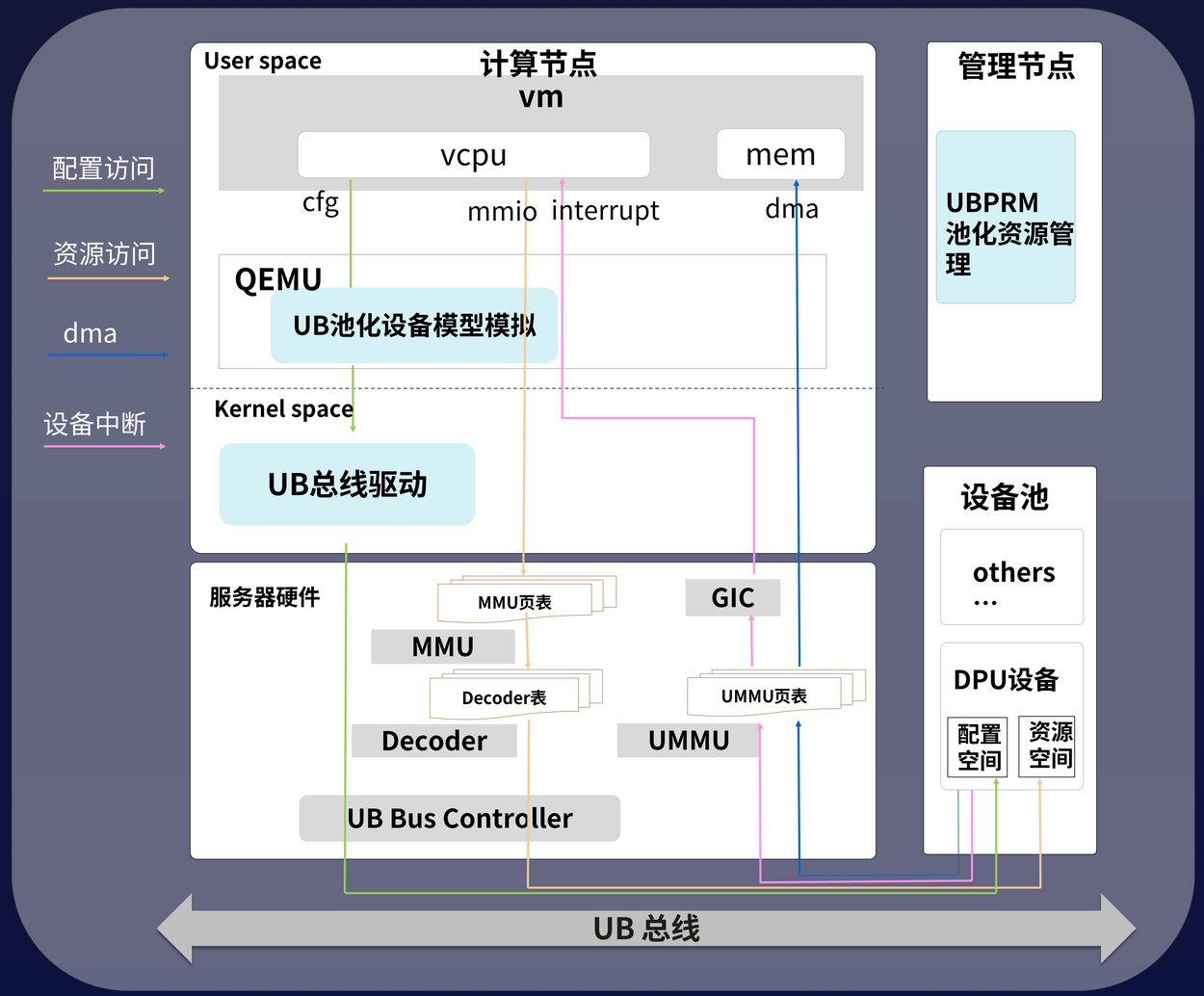


提升业务性能和整体资源效率

- 运行时对单一资源1:X抽象 ➔ **多资源M:N**融合抽象
- 按照峰值固定分配资源 ➔ 根据负载**动态调整**资源

设备池化：基于高性能互连计算架构，打造互联池化架构底座，提升通用计算超节点性价比

架构设计



关键技术

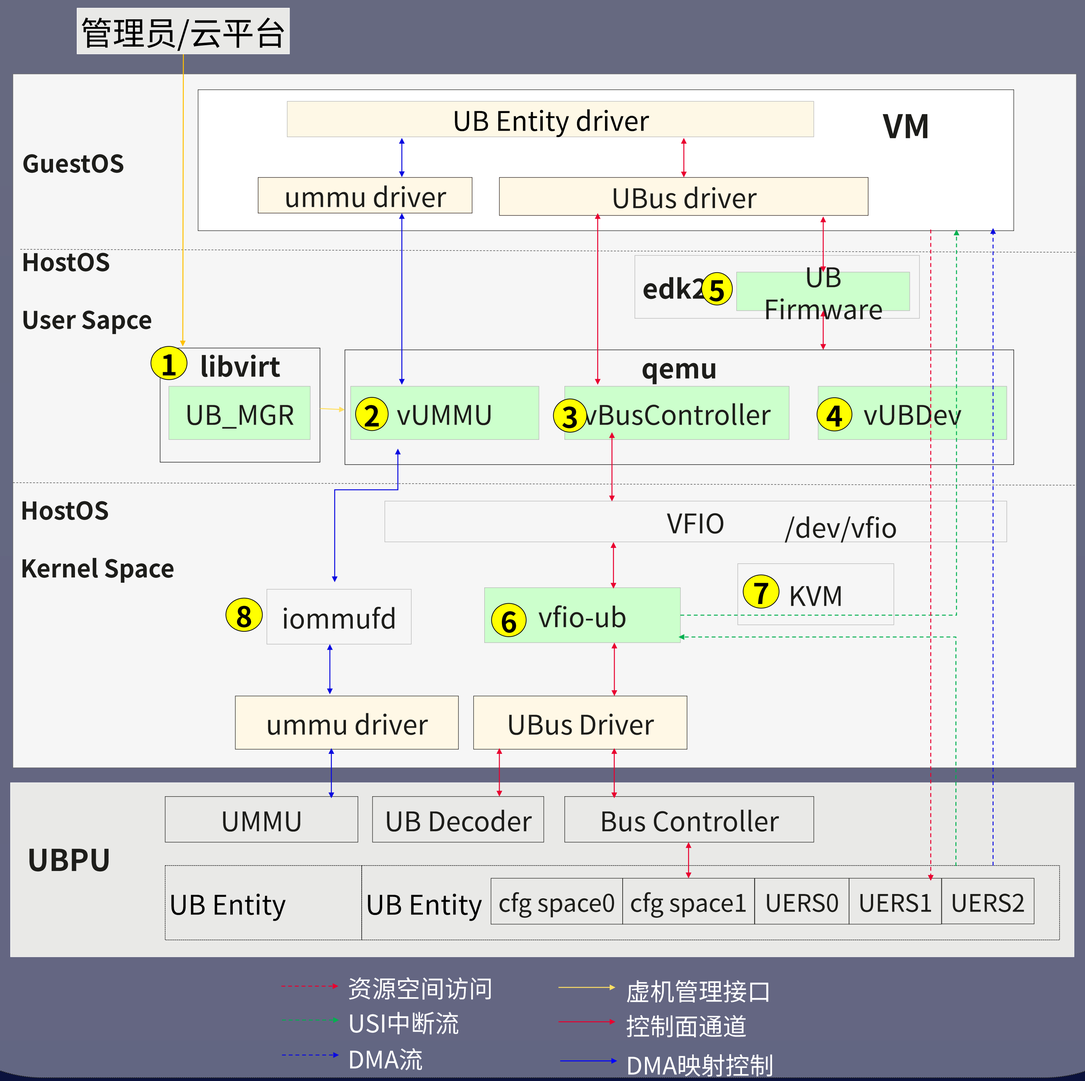
①UB池化设备模型模拟： 虚机支持UB总线模拟，数据面实现UB设备的直通访问mmio、interrupt以及设备dma等，支持虚机使用远端UB设备池按需弹性扩展。

②UB总线驱动：提供UB设备模型管理及相关接口，通过UB芯片编程接口接入UB总线，识别UB硬件设备信息，支持系统接入UB总线。支持用户态访问UB设备。

③UBPRM池化资源管理：管理域内池化设备资源，和Ubus总线驱动配合实现池化设备资源的动态注册等功能

设备池化：基于高性能互连计算架构，打造互联池化架构底座，提升通用计算超节点性价比

软件架构分层



软件分层介绍

① UB_MGR：支持通过libvirt xml配置文件创建、管理UB虚拟机。

② vUMMU：为虚拟机提供UMMU模拟，支持ummu driver建立stage1页表，同时完成物理UMMU 页表配置。

③ vBusController：为虚拟机模拟Bus Controller，提供虚拟UB总线入口，虚拟机内UB总线驱动 通过Bus Controller 完成 UB 设备拓扑扫描发现及访问配置等。

④ vUBDev：为虚拟机模拟UB设备模型，提供完整的UB配置空间、资源空间、虚拟拓扑等，完成 资源空间映射。

⑤ UB Firmware：负责将Bus Controller信息、UMMU信息、可控分配的UB资源以及与UMMU和 中断控制器的映射关系等上报给GuestOS。

⑥ vfio-ub：提供用户态设备管理功能，支持把UB设备暴露给用户态qemu使用。

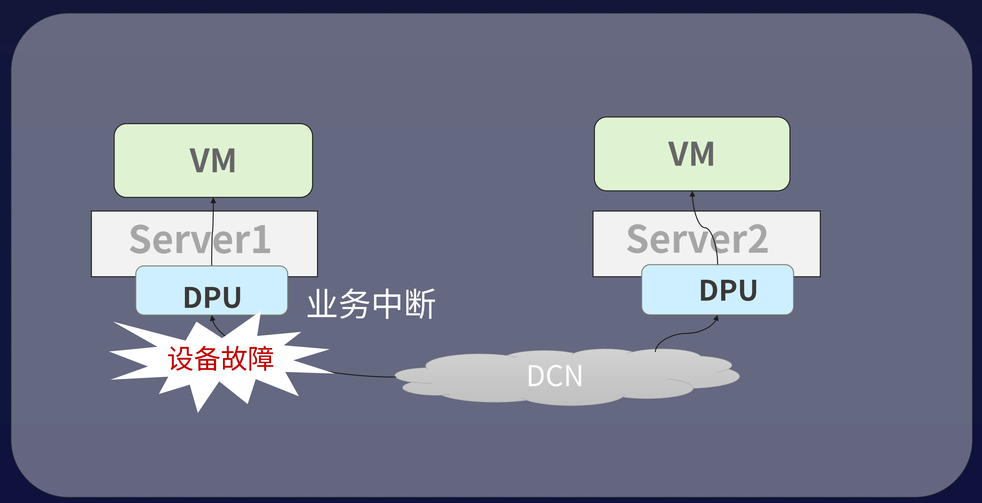
⑦ KVM: USI中断重映射, 提供虚拟USI中断注入或者中断透传。

⑧ iommufd：提供用户空间管理I/O 页表的API，供虚拟化完成UMMU页表配置

灵衢池化架构对可靠性的提升

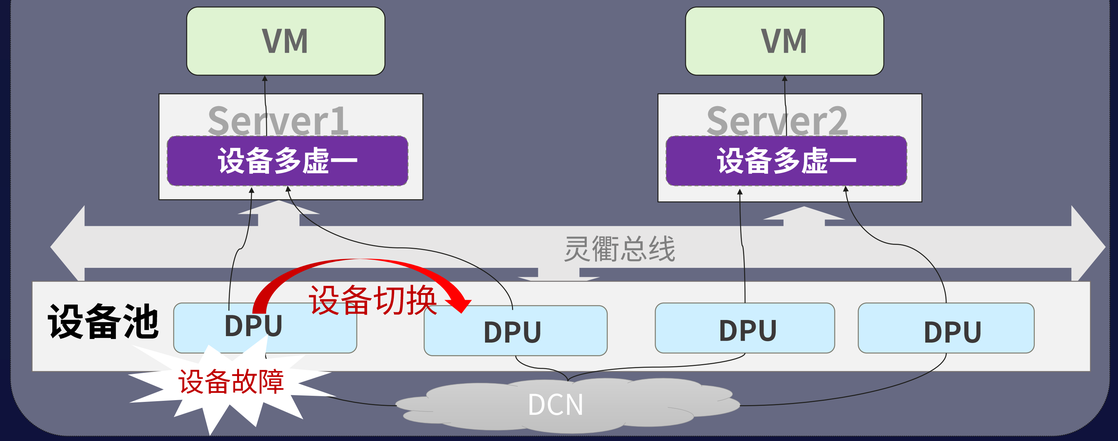
问题场景

单服务器资源固定配比，服务器设备故障，导致业务中断



解决方案

统一聚合设备模型抽象，支持设备多虚一，支持虚机直通访问



传统高可靠技术存问题

- 传统设备高可靠技术需用户感知，例如网卡bond聚合，需要虚机guest感知配置
- 仅支持单一设备类型，例如网卡bond，无法支持磁盘类设备

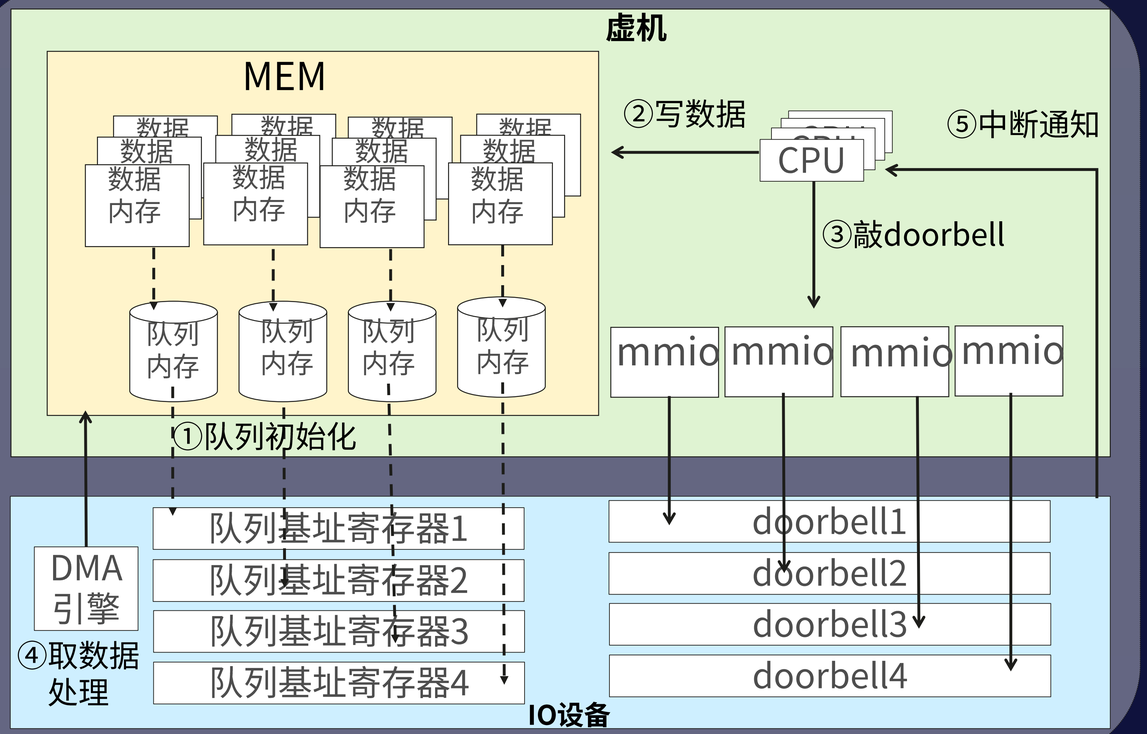
设备多虚一无感提升设备可靠性

- 统一设备模型，支持virtio-net、virtio-blk、virtio-scsi多设备类型
- 支持AA、AP双模式，提升IO吞吐和可靠性

统一聚合设备模型：支持IO队列任务的跨卡调度

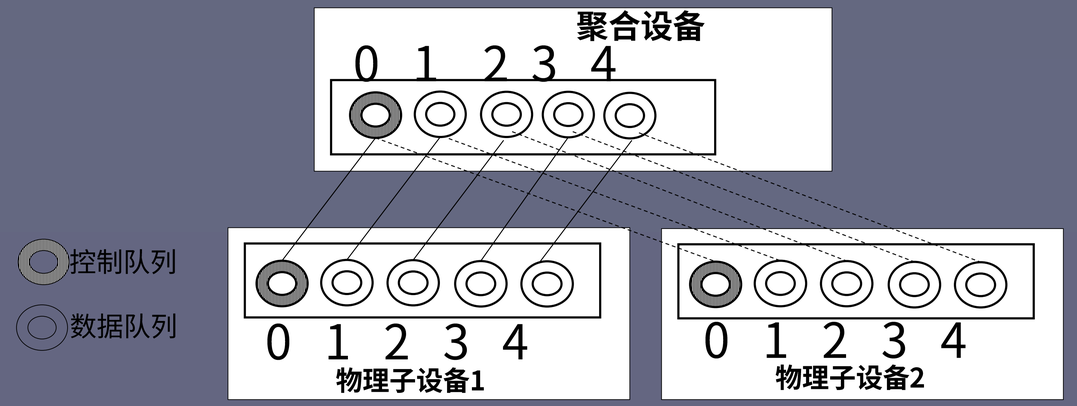
IO设备工作原理

队列是处理IO任务的最小工作单元



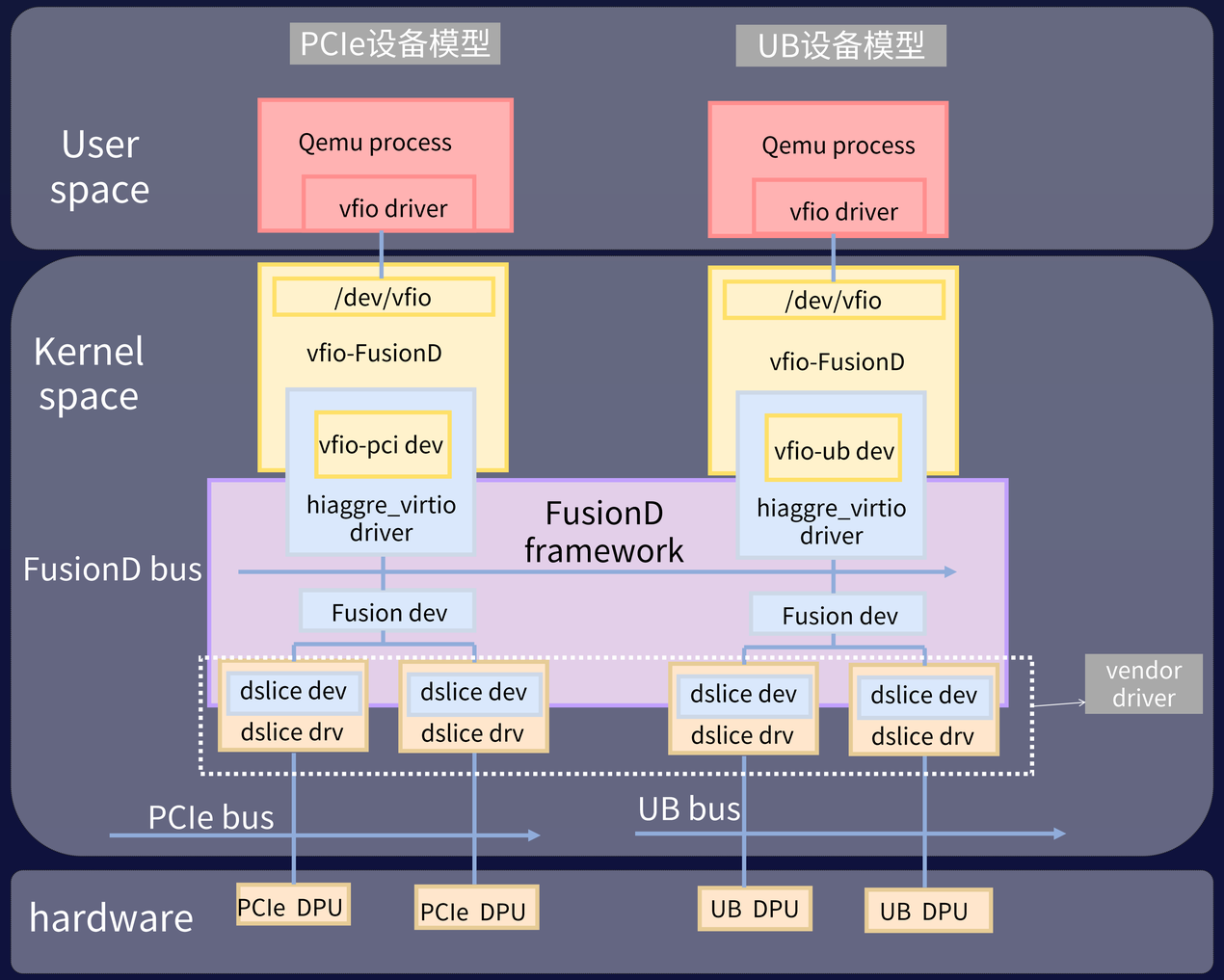
解决方案

通过队列级的聚合映射控制，实现聚合设备IO的跨卡调度



统一聚合设备模型：支持将多个物理设备虚拟成单一设备，提升IO带宽&设备冗余可靠性

软件架构分层



软件分层介绍

- QEMU通过标准的PCI直通虚拟化或者UB直通虚拟化来打开内核的聚合设备，呈现给GuestOS；
- vfio_fusiond基于聚合框架模拟出的聚合设备，附加上PCI设备模型或者UB设备模型的模拟，并通过标准VFIO接口对qemu暴露出聚合设备；
- FusionD框架基于驱动注册的ops获取物理设备信息，模拟出聚合设备，并提供聚合设备数据面直通映射所需的公共接口。
- hiaggre_virtio处理vfio_fusiond下发的设备空间访问，翻译为virtio协议的行为，并根据子设备映射关系通过驱动转发给对应的物理子设备的操作。
- hiaggre_dslice物理子设备驱动，直接操作物理设备，向上提供框架定义的ops接口

Thank You

分批开源计划

- 首批UB总线模拟、设备直通虚拟化等基础能力在25年11月开源；
- 统一聚合设备模型预计26年H2逐步开源；

加入我们

openEuler VirtSig

Q&A

?