面向内核态vDPA的通用设备模型及热迁移支持

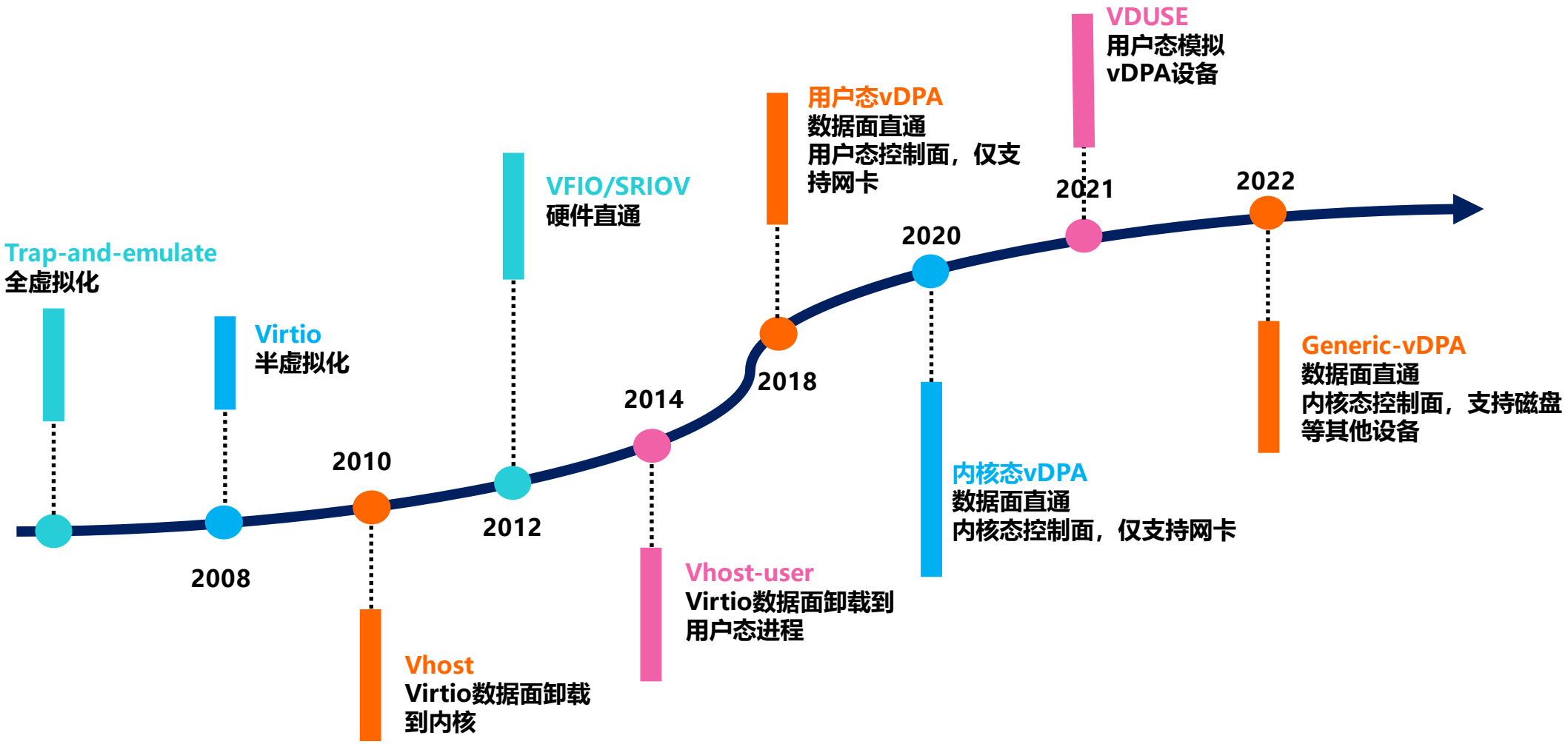
方　毅　　华为技术有限公司

OpenEuler

目录

- 方案背景
- 内核态vDPA面临的问题及方案探索
- openEuler对内核态vDPA的支持计划及下一步工作

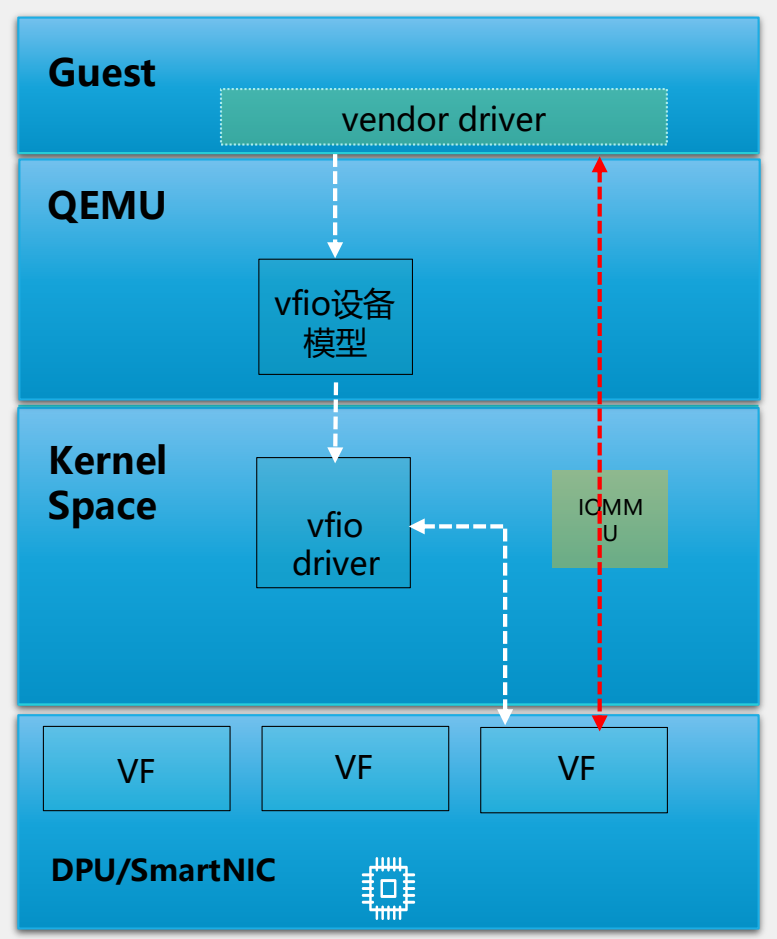
方案背景 - 设备虚拟化演进



整体发展趋势：优化数据面提升IO性能，优化控制面提升设备管理灵活性

vDPA方案实现软硬结合、控制面与数据面分离，性能与灵活性的兼容

方案背景 – SRIOV的优点与不足



SRIOV优点

性能持平物理硬件

DMA remapping

Interrupt Posting

技术成熟稳定

发展10+年

大量应用场景检验

方案通用性强

不感知具体设备类型

统一API，适用范围广

SRIOV不足

控制面暴露

不利于产品内硬件多样性

硬件要求高，部分兼容性问题

直面硬件，安全性稍差

热迁移支持复杂

开源社区尚无成熟热迁移方案

不同厂商之间硬件无法热迁移

为实现性能不下降的同时，解决SRIOV的不足，引入vDPA方案

方案背景 – vDPA介绍

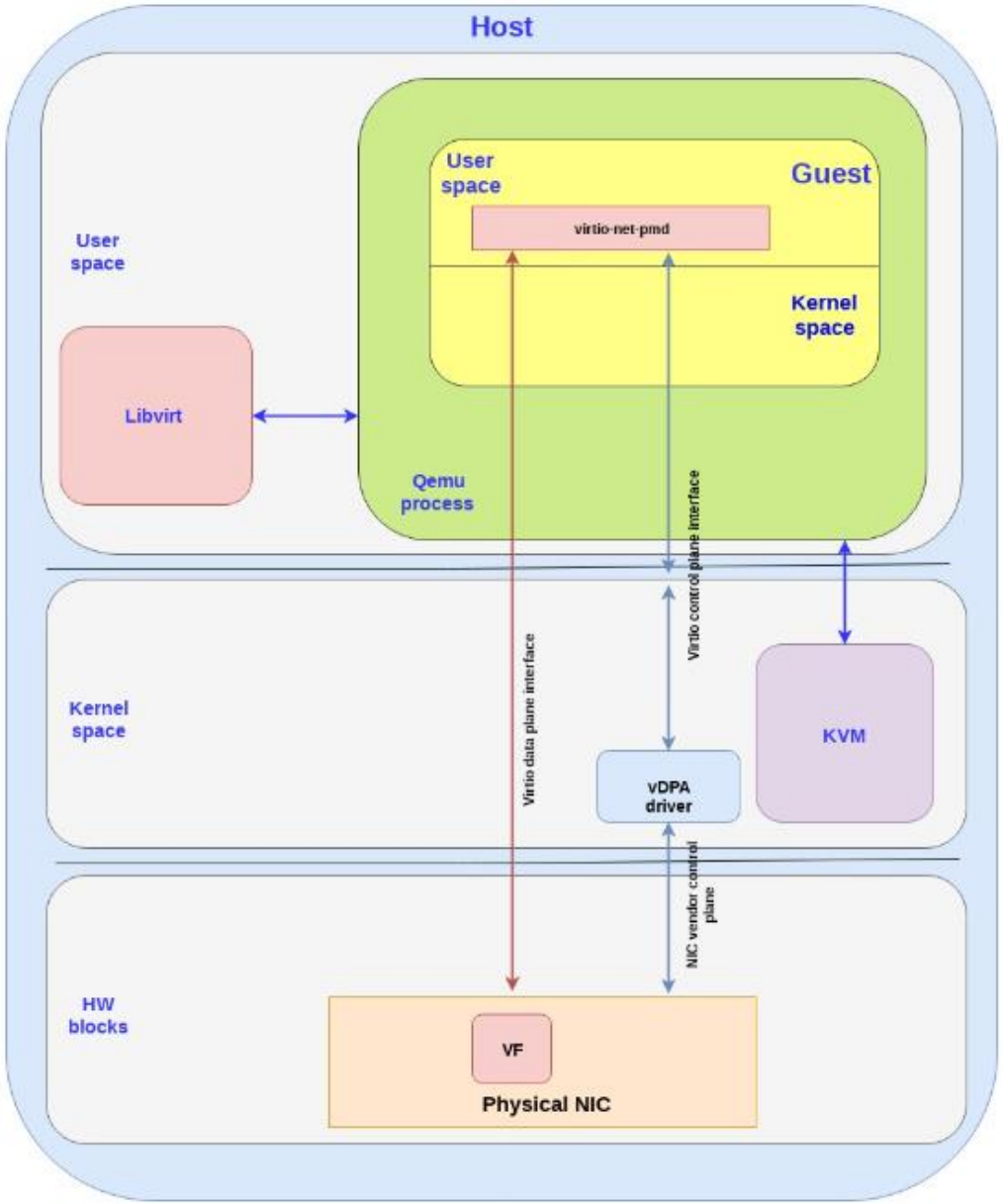
vDPA全称virtio data path acceleration，即virtio数据路径加速。其目的是分离设备的控制面与数据面，控制面由qemu、vDPA框架及厂商vDPA驱动共同实现，数据面遵循virtio vring布局，实现直通，性能可以达到硬件线速。这样做的好处如下：

优点

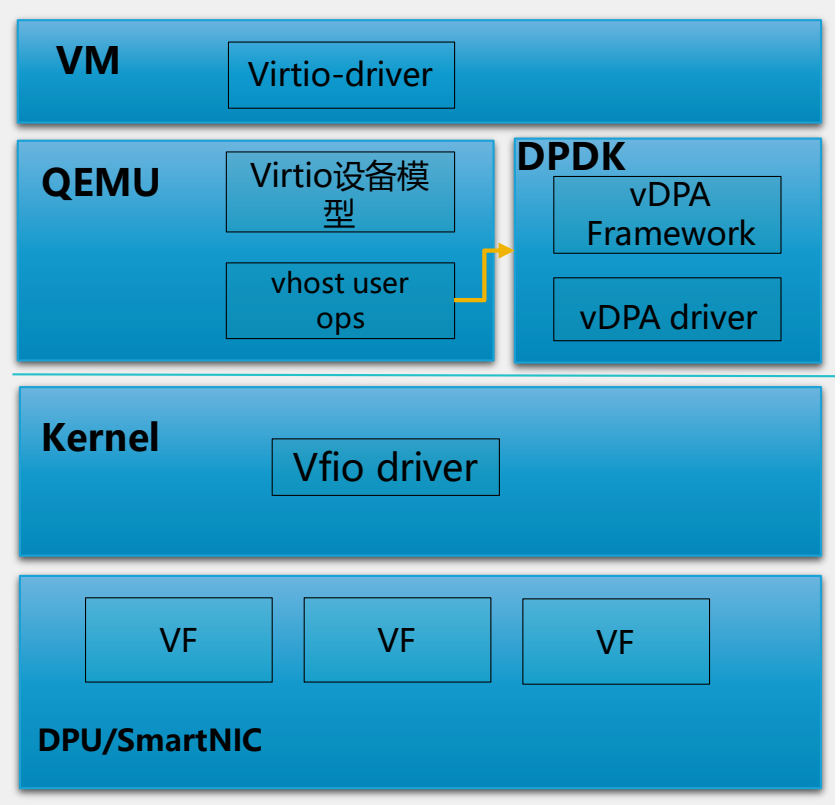
- 性能优：数据面直通，无中间层，达到硬件线速
- 控制面灵活：由厂商灵活定制，硬件设计更简单
- 支持Guest免额外驱动：遵循virtio规范，虚拟机内部直接可以使用virtio驱动，无需安装厂商定制驱动
- 跨硬件热迁移：支持不同厂商硬件之间虚拟机热迁移

根据vDPA的实现位置，可分为：

- **内核态vDPA**：qemu新增一套vDPA设备模拟框架，vDPA框架和驱动在内核中实现。
- **用户态vDPA**：设备模拟复用qemu中vhost-user方案，vDPA框架和驱动在DPDK中实现。



方案背景 – 用户态vDPA的优点与不足



用户态vDPA优点

复用vhost-user框架，改动量小

Qemu vhost-user client

DPDK vhost-user server

技术更成熟

18年开始支持

支持的厂商多，Mellanox、Intel等支持

用户态vDPA不足

只支持网卡

磁盘、fs等其他设备不支持

DPU场景卸载不彻底

Host侧需保留dpdk进程，占用额外的CPU、内存资源

不能充分利用kernel能力

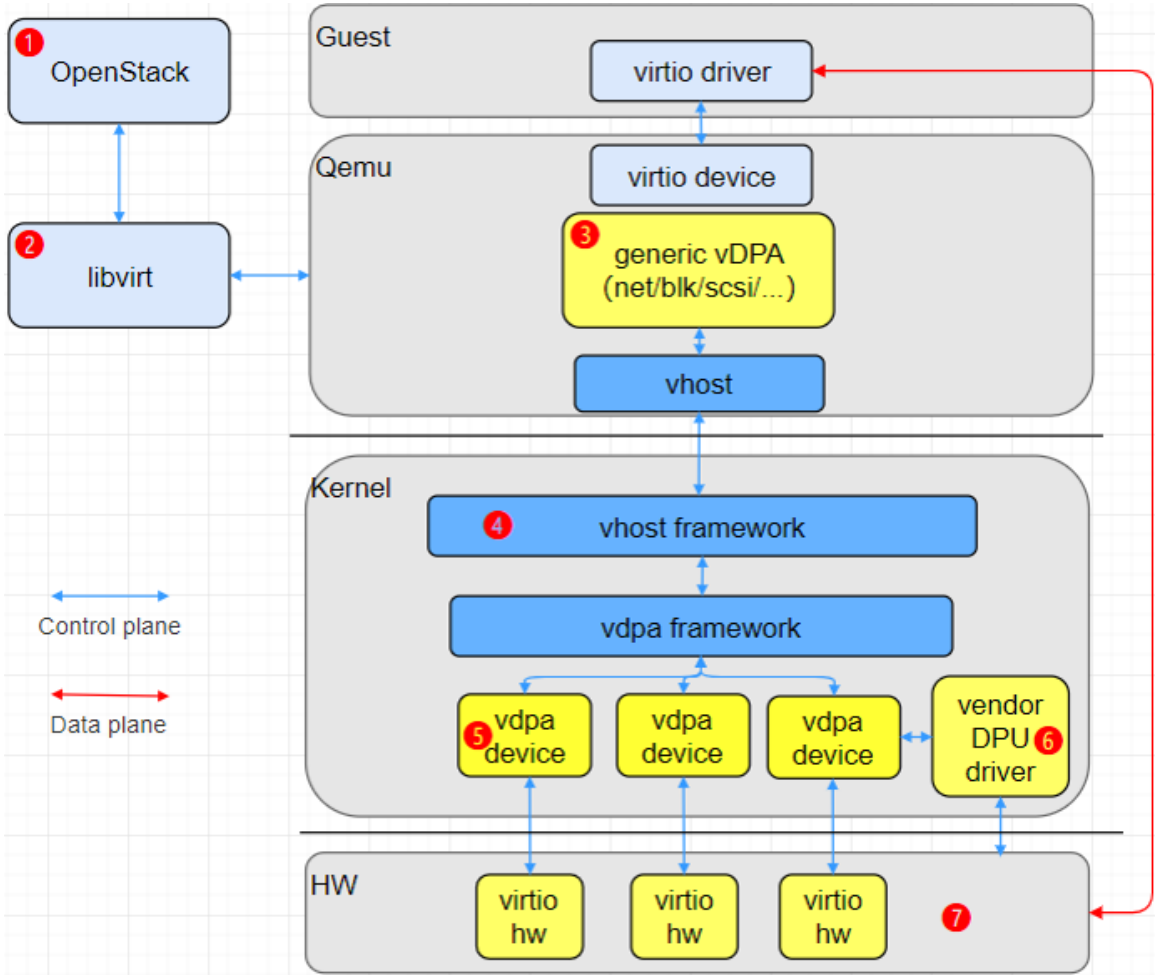
Vhost-user仅提供用户态接口，无法充分利用kernel能力，如eBPF

DPDK专注数据面，不提供配置和控制硬件的工具

方案背景 – 内核态vDPA

内核态vDPA涉及的组件大致如图所示：

1	OpenStack	OpenStack 通过nova服务支持纳管vdpa设备（生命周期管理、热迁移、热插拔等）
2	Libvirt	Libvirt实现vDPA设备的XML配置及解析，对外提供vDPA设备管理接口，支持管理vdpa设备（生命周期管理、热迁移、热插拔等）
3	generic vDPA virtio device	实现vDPA通用设备模块框架，及生命周期管理、热迁移、热插拔等功能
4	vhost-vdpa framework	打通vdpa管理路径和实现部分热迁移逻辑
5	vdpa device	对接vdpa框架实现virtio设备热迁移、热插拔、生命周期管理
6	Vendor DPU driver	支持DPU virtio设备和vmstate热迁移、生命周期管理、热插拔
7	virtio hw	微码支持DPU virtio设备和vmstate热迁移、生命周期管理、热插拔



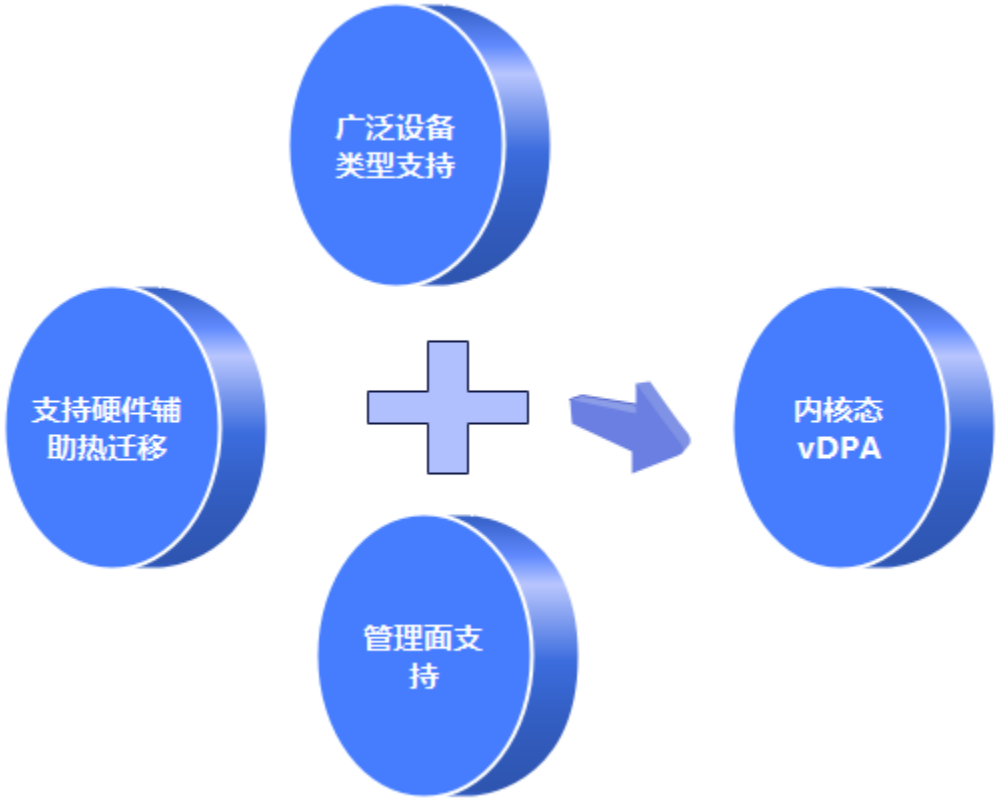
OpenEuler

目录

- 方案背景
- 内核态vDPA面临的问题及方案探索
- openEuler对内核态vDPA的支持计划及下一步工作

内核态vDPA面临的问题

- 问题一：仅支持网卡，不支持磁盘等其他设备类型，限制了DPU卸载场景vDPA的使用
- 问题二：管理组件如OpenStack、Libvirt不支持内核态vDPA，
- 问题三：内核态vDPA不支持硬件辅助的热迁移，无法实现vDPA最初的设计目标



方案探索 – Generic vDPA

OpenEuler

问题

- 社区只支持网卡vDPA，其余设备类型发展缓慢
- 每个类型设备有单独一套设备模型代码，通用及扩展性低

解决方案

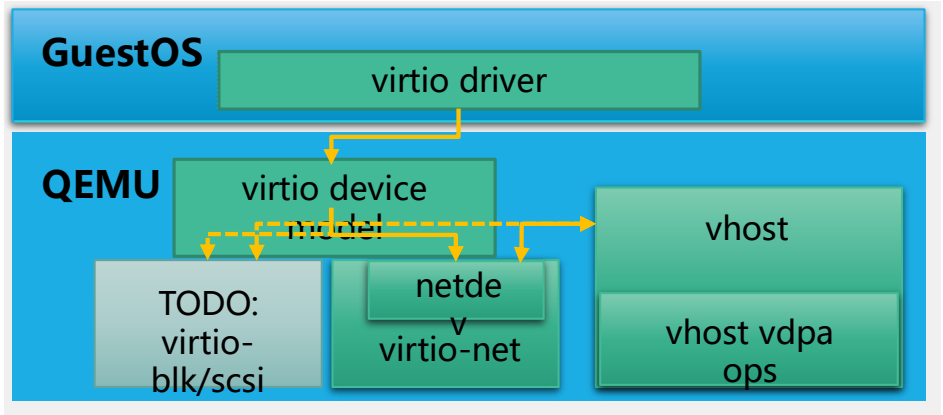
- 通用设备模型，不区分具体virtio设备类型，**一套框架支持virtio-net/blk/scsi/fs等所有virtio设备**
- 使用方式：

virtio-pci

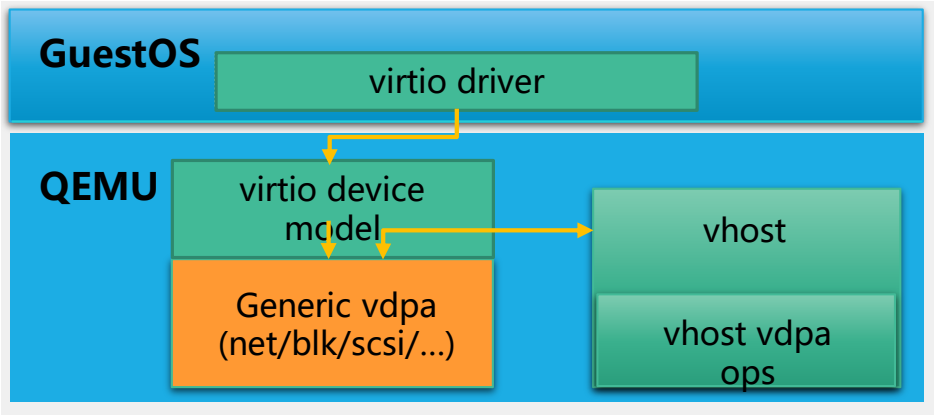
-device vhost-vdpa-device-pci,vdpa-dev=/dev/vhost-vdpa-x

当前Qemu支持的网卡vDPA方案

-netdev **type=vhost-vdpa**,vhostdev=/dev/vhost-vdpa-0,id=vhostvdpa1
-device **virtio-net-pci**,netdev=vhostvdpa1



华为向社区推送的Generic vDPA方案



方案探索 – Libvirt支持Generic vDPA

OpenEuler

问题

- 社区发展缓慢，不支持配置Generic vDPA设备

解决方案

- 推荐方案一
- 社区patch：

https://listman.redhat.com/archives/libvir-list/2023-March/239038.html

方案一：Libvirt新增vDPA subsystem，vDPA与具体设备类型解耦

```
<devices>
    <hostdev mode='subsystem' type='vdpa'>
      <source dev='/dev/vhost-vdpa-0'/>
    </hostdev>
</devices>
```

方案二：参照当前社区网卡vDPA方案，新增磁盘等其他类型vDPA的配置

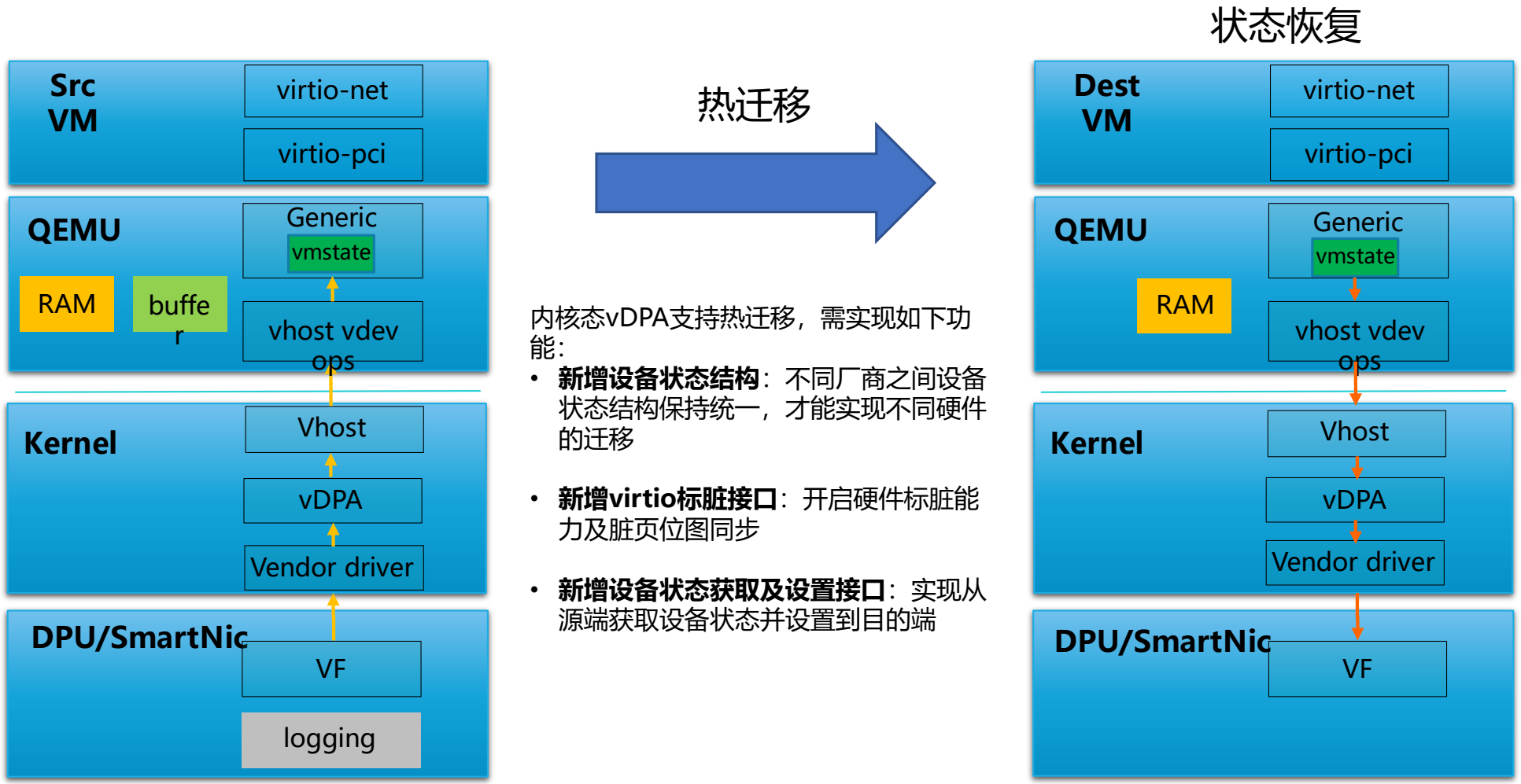
```
net:
 <devices>
  <interface type='generic-vdpa'>
   <source dev='/dev/vhost-vdpa-0'/>
  </interface>
 </devices>

 disk:
 <devices>
   <disk type='generic-vdpa’ >
    <source dev='/dev/vhost-vdpa-0'/>
   </disk>
 </devices>
```

方案三：参照PCI直通进行配置，该方案需Libvirt完成驱动绑定等功能，与现在社区已有的vdpa工具功能重复。

```
<devices>
    <hostdev mode='subsystem' type='pci' managed='yes'>
      <driver name='vdpa'/>
      <source>
        <address domain='0x0000' bus='0x04' slot='0x01' function='0x01'/>
      </source>
    </hostdev>
</devices>
```

方案探索 – 支持vDPA热迁移



内核态vDPA支持热迁移，需实现如下功能：

- **新增设备状态结构**：不同厂商之间设备状态结构保持统一，才能实现不同硬件的迁移
- **新增virtio标脏接口**：开启硬件标脏能力及脏页位图同步
- **新增设备状态获取及设置接口**：实现从源端获取设备状态并设置到目的端

社区交流：

Upstream：https://lists.linuxfoundation.org/pipermail/virtualization/2023-September/068152.html

openEuler：https://gitee.com/openeuler/kernel/pulls/2372

OpenEuler

目录

- 方案背景
- 内核态vDPA面临的问题及方案探索
- openEuler对内核态vDPA的支持计划及下一步工作

openEuler对内核态vDPA的支持计划及下一步工作

OpenEuler

2023.7：基于文中方案完成vDPA基本能力及迁移原型验证

2023.12： openEuler 22.03 SP3对外发布内核态vDPA基本能力及迁移支持

2024：完善vDPA方案并推送Upstream社区

内核态vDPA后续重点工作：

- 内核态vDPA高级特性（磁盘扩容等）
- 完善内核态vDPA对不同DPU&SmartNic的支持
- 内核vDPA性能优化（启停时间、迁移中断时间、数据面等）
- Generic vDPA支持Shadow VQ迁移方案


OpenEuler

方案背景 - 设备虚拟化演进

OpenEuler

	控制面方案	数据面方案	性能	资源占用
Trap-and-emulate	纯软实现	纯软实现	低	高
virtio	Qemu	Qemu	中	高
Vhost-kernel	Qemu&内核态vhost	Kernel	中	高
SRIOV	硬件直通	硬件直通	高	低
Vhost-user	Qemu&用户态vhost	用户态vhost	高	高
用户态vDPA	Qemu&用户态vhost	硬件直通	高	低
内核态vDPA	Qemu&内核态vhost	硬件直通	高	低
VDUSE	Qemu&内核态vhost	用户态	高	高
Generic vDPA	Qemu&内核态vhost	硬件直通	高	低