透传设备VMM热升级探索与实践

滴滴

滴滴一下 美好出行

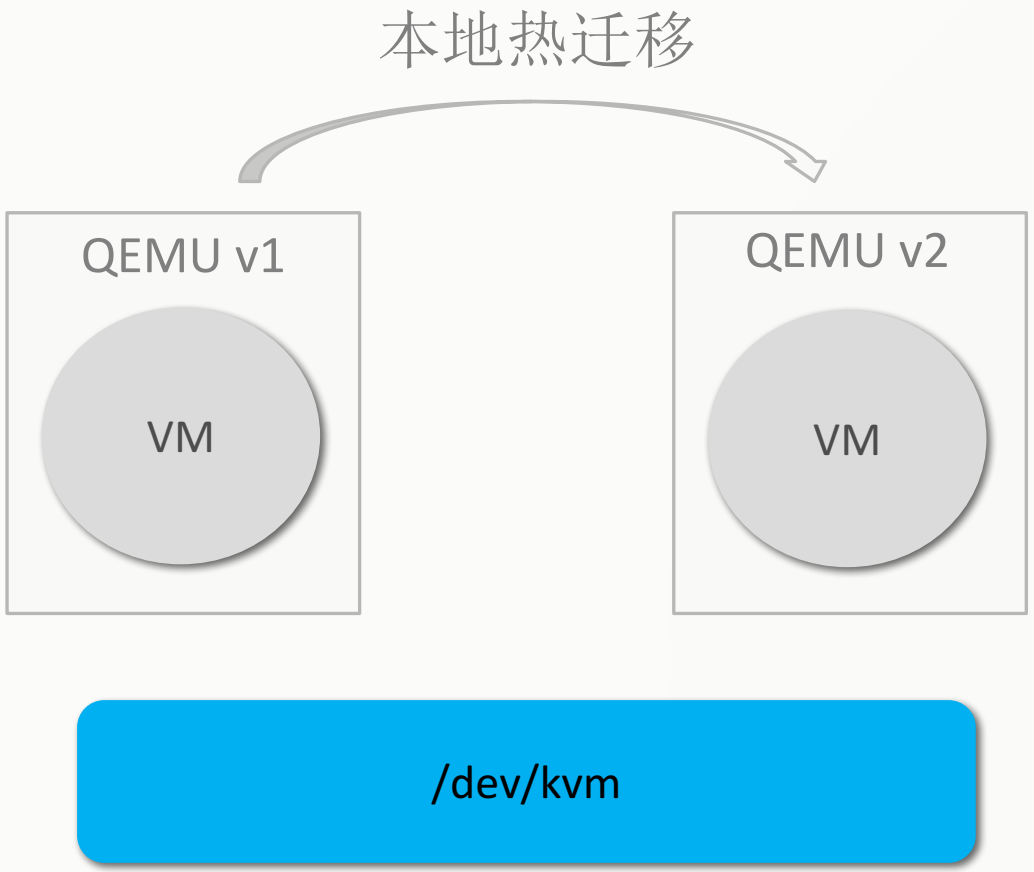
目录

- 背景介绍
- VMM热升级的整体方案
- 当前状态以及后续工作

背景介绍

目前随着公有云集群规模越来越大，如何在不间断虚拟机运行的情况下对虚拟化的各个组件进行热升级成为了一个非常关键的问题。CPU实例的VMM可以通过qemu 本地热迁移来解决热升级的问题，但是随着支持透传设备实例的出现，本地迁移已经无法满足该类型实例的热升级。

背景介绍



热迁移技术进行热升级一些缺点：

1. 升级时会占用额外的内存和CPU资源。
2. 当前无法支持透传设备的VM的升级。
3. KVM模块是无法升级的。

面临的问题：

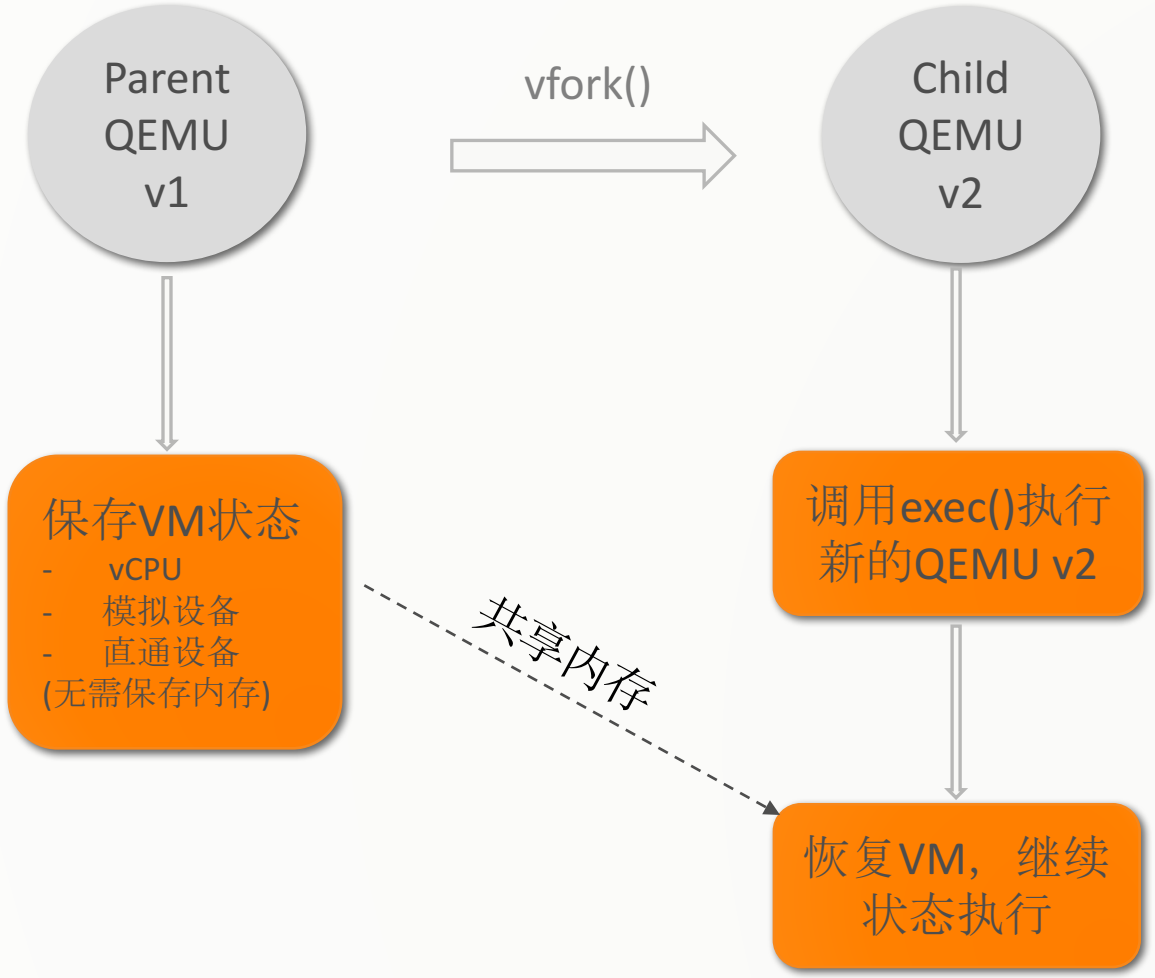
1. Qemu/KVM bugs 无法修复。
2. VMM相关的新feature 无法应用到线上。
3. ……

背景介绍

业界已有的一些解决方案或者讨论：

1. 基于vfio-mdev 设备热迁移（需要host driver 本身对热迁移的支持）
2. 基于vfio-pci 设备的热迁移 （ 需要host driver 本身对热迁移的支持）
3. 一些论文讨论热升级的方案

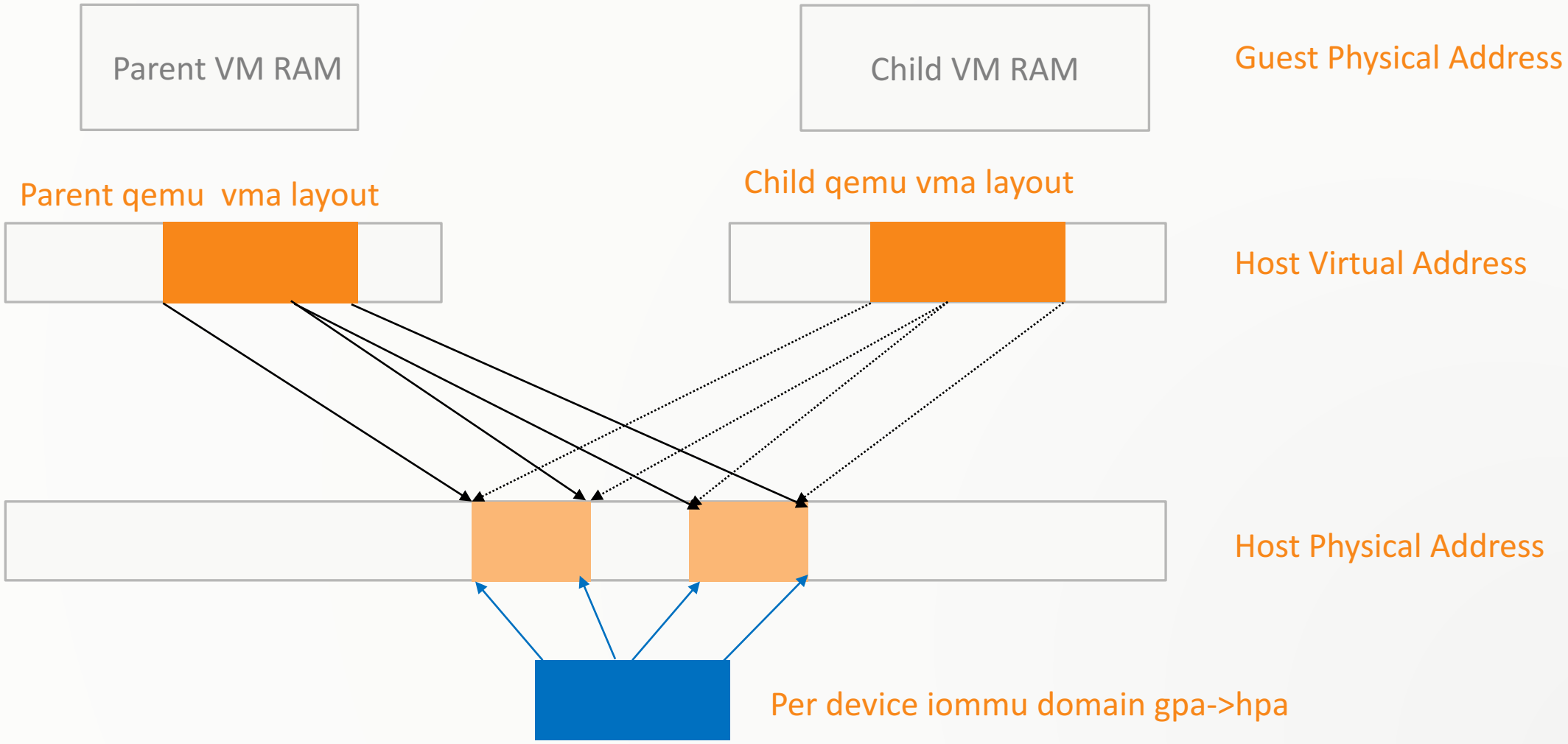
VMM热升级的整体方案 -父进程vfok-exec 子进程



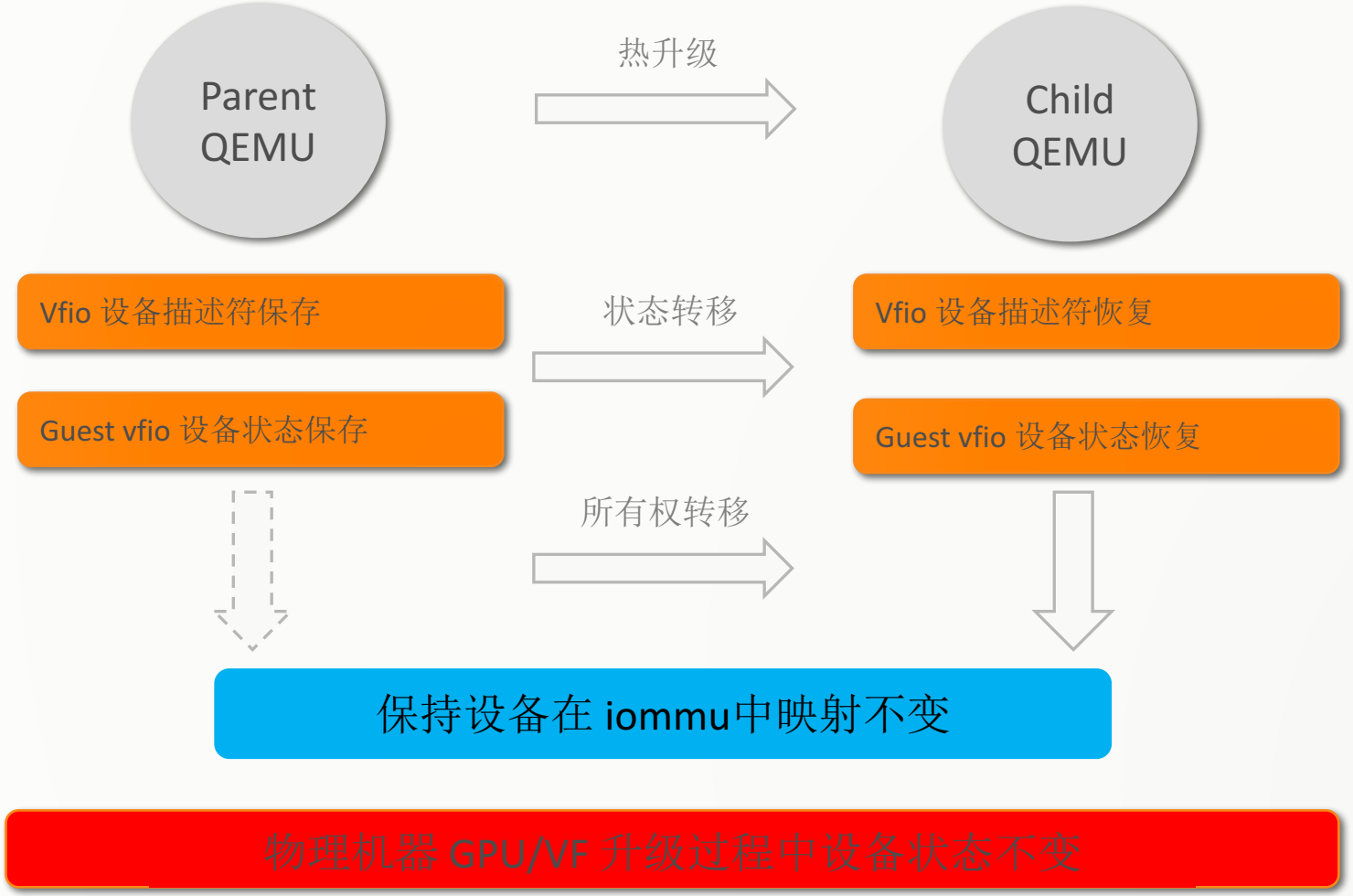
关键技术点：

1. 使用vfork 生成子qemu进程。
2. 保存和恢复设备以及vcpu的状态。

VMM热升级的整体方案-共享内存



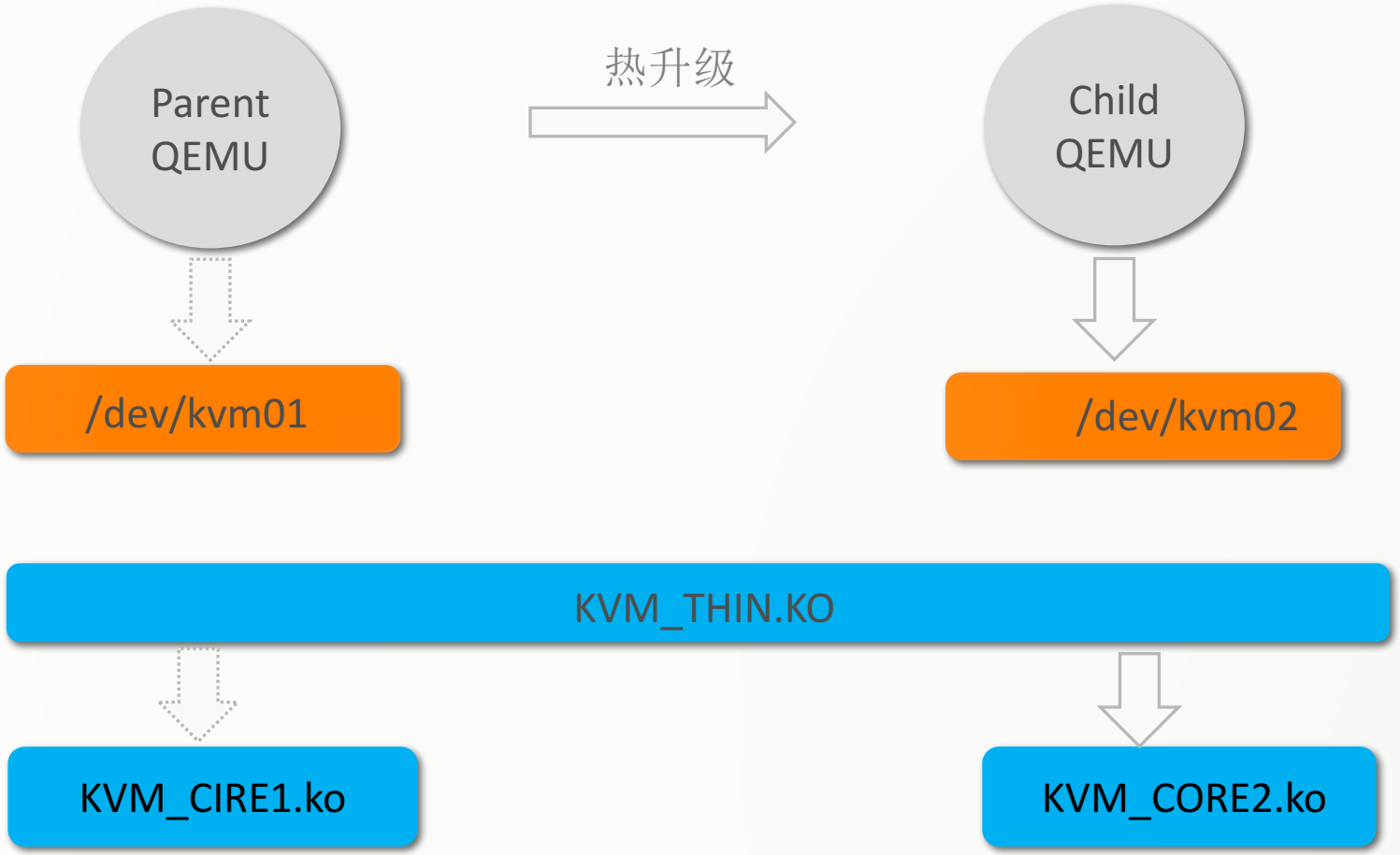
VMM热升级的整体方案-vfio状态保存及恢复



关键技术点：

1. 父子qemu 间 fd以及vfio设备状态的传递
2. 无需记录和传输DMA脏页。
3. 无需保存GPU硬件状态。

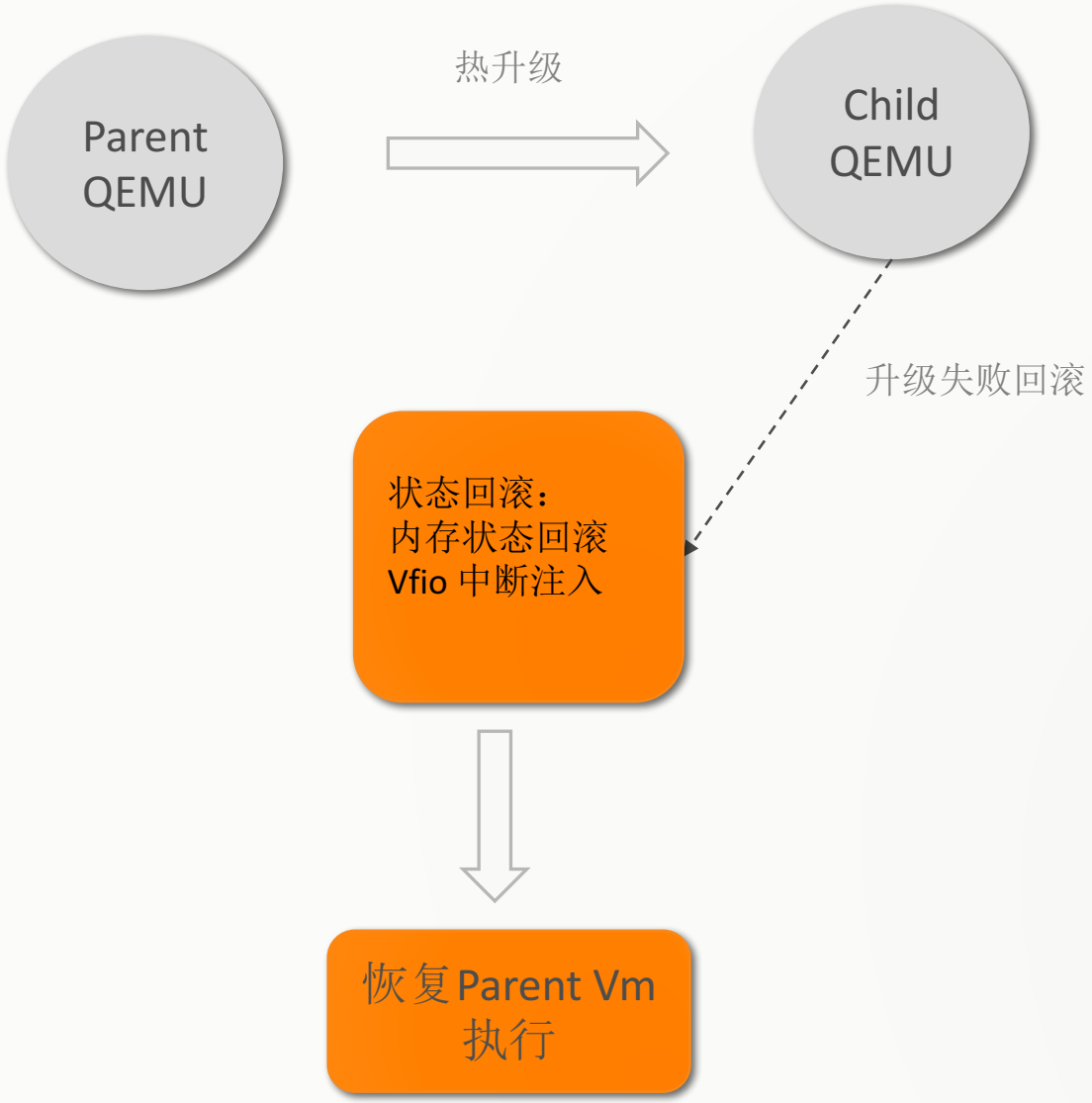
VMM热升级的整体方案- kvm 模块升级



关键技术点：

1. 修改kvm/kvm_intel 模块：
   kvm_core.ko：所有KVM相关核心逻辑
   kvm_thin.ko：全局符号的导出
2. 确保两个模块在热升级过程，同时运行VM。
3. Qemu展示出不同的设备。

VMM热升级的整体方案-状态回滚



VMM热升级的整体方案–数据

16 core 32G RAM 且负载较重的CPU

	基于热迁移的热升级	热升级
downtime (ms)	80~800	~20
升级时长(s)	300	0.3
资源消耗	内存： 32.1G CPU：0.65 core	内存：0.1G CPU: 0.1 core
透传设备以及KVM升级支持	不支持	支持

当前状态以及后续工作

当前状态：
全流程打通并且全量上线热升级方案。

后续工作：
更多透传设备的验证，以及downtime的优化。

THANK YOU

滴滴

滴滴一下 美好出行