NFV场景下虚拟机隔离部署

龚 磊

arei.gonglei@huawei.com

www.huawei.com

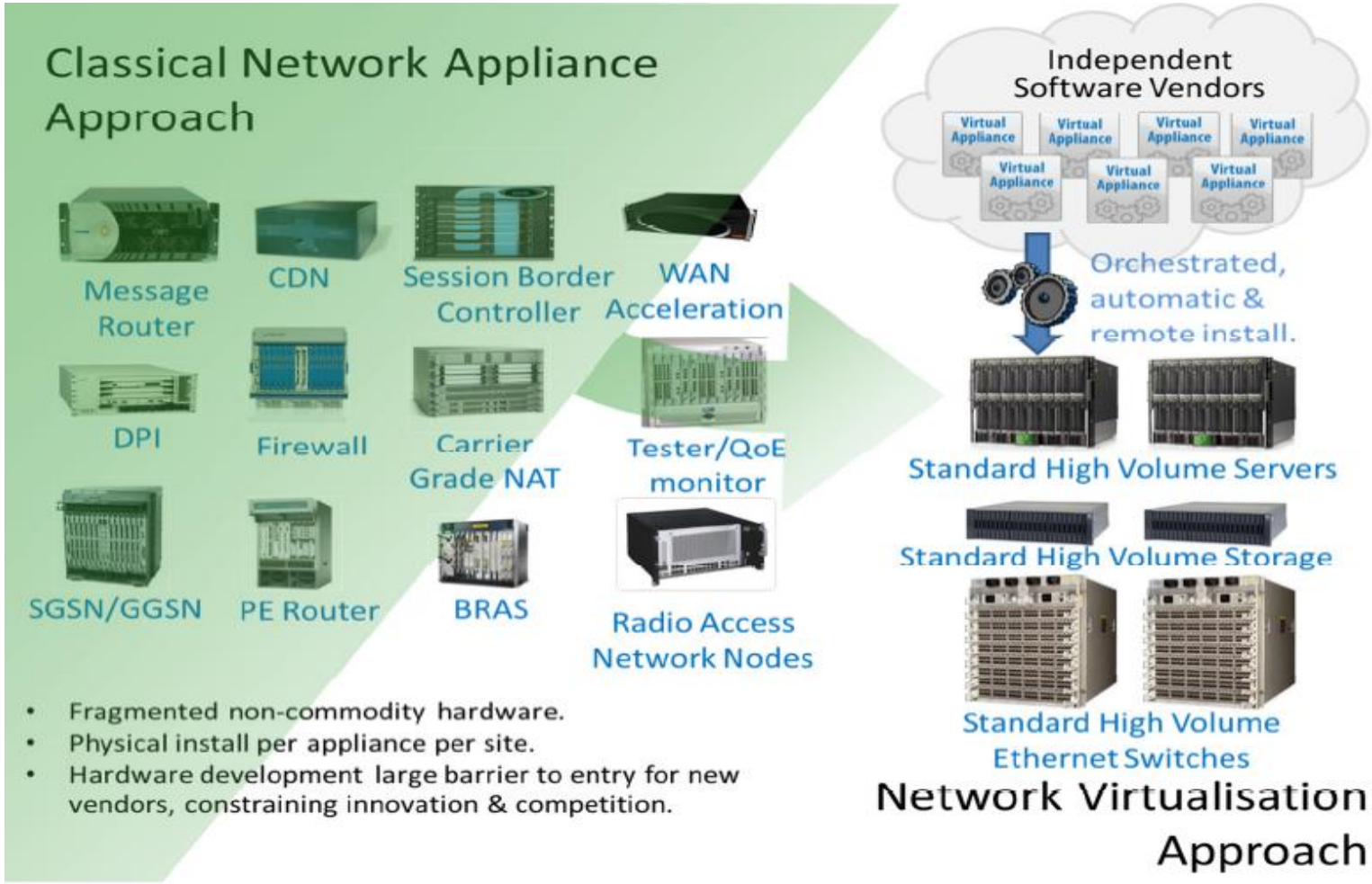
HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

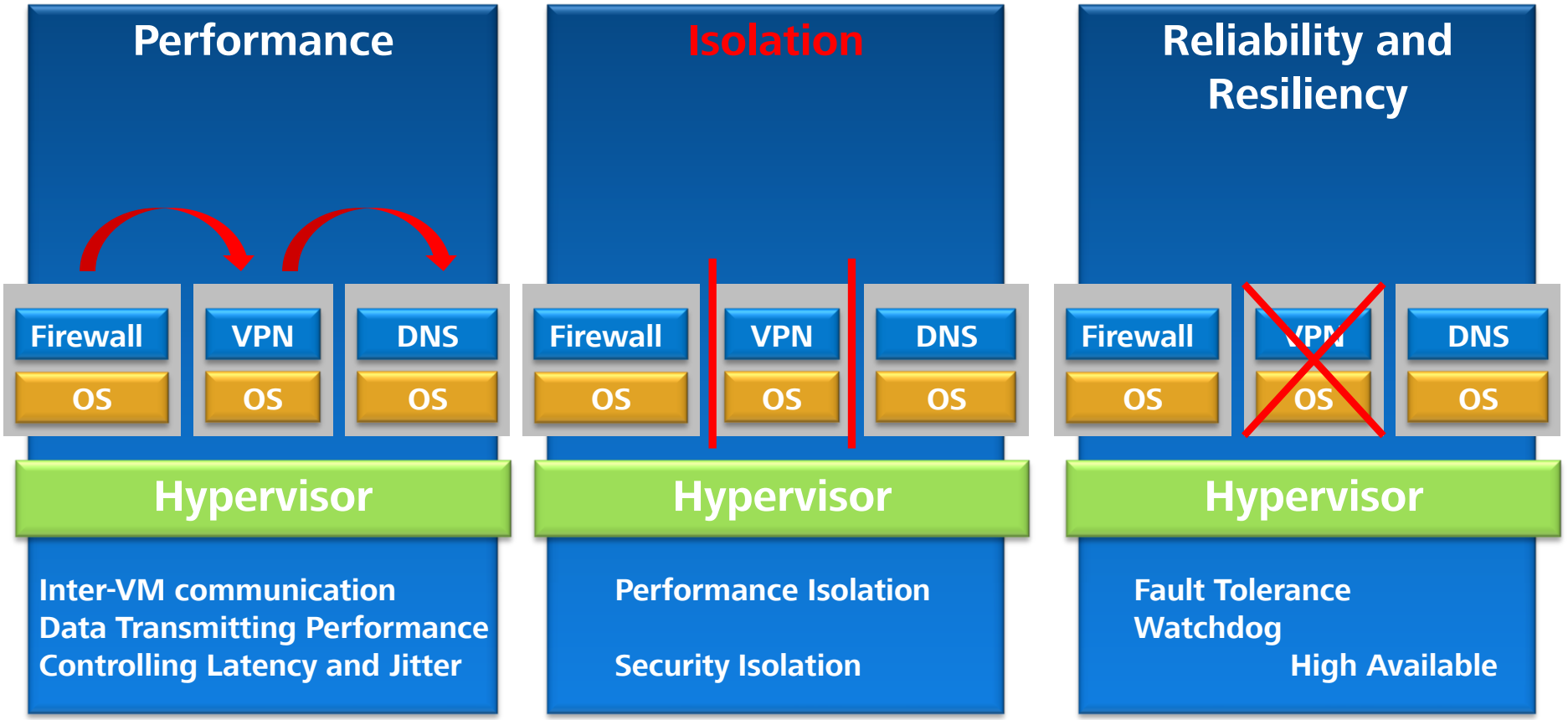
目录

背景：NFV是什么？

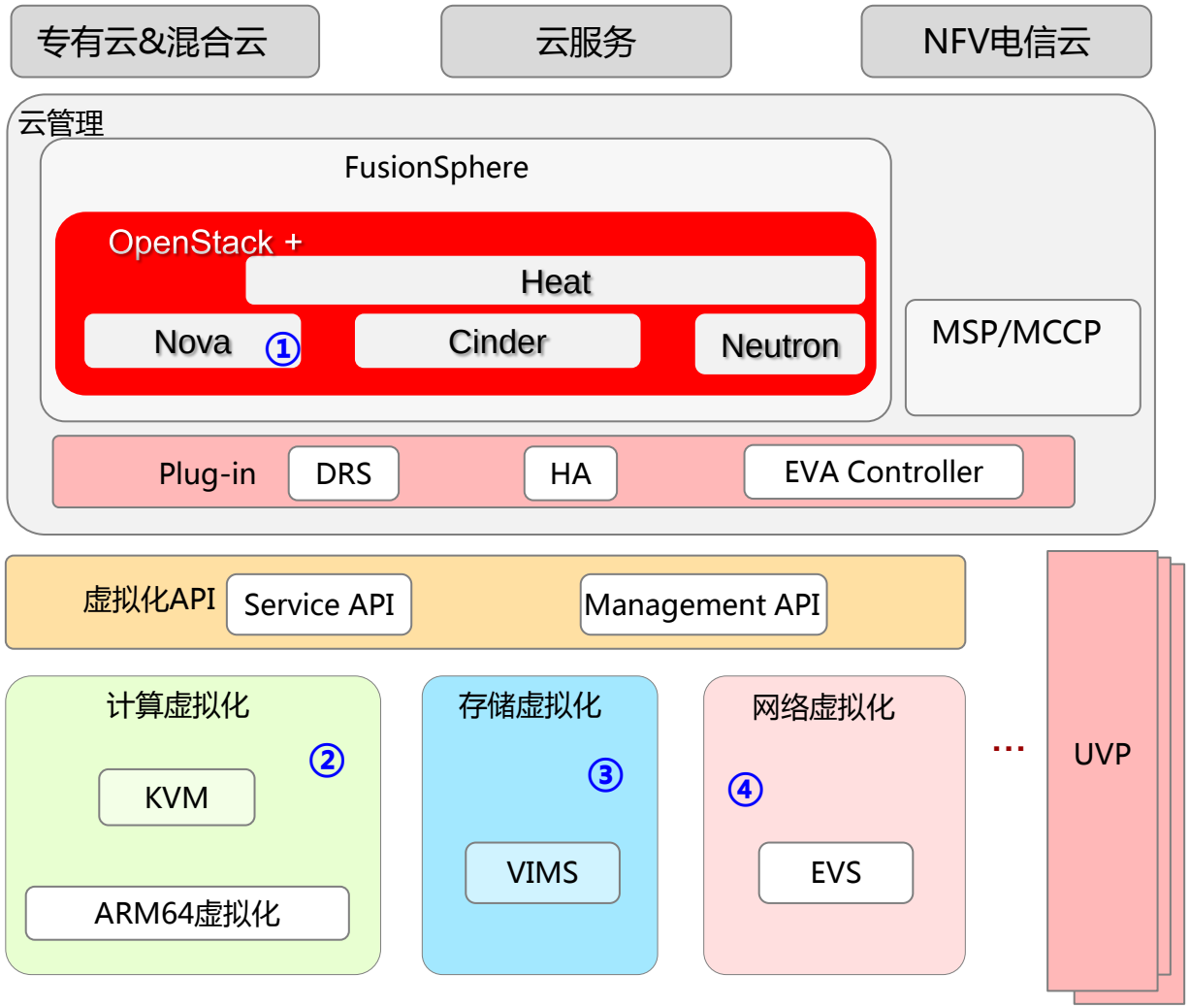
NFV： Network Functions Virtualization 网络功能虚拟化



背景：NFV面临的挑战


Performance
Firewall
VPN
DNS
OS
OS
OS
Hypervisor
Inter-VM communication
Data Transmitting Performance
Controlling Latency and Jitter
Isolation
Firewall
VPN
DNS
OS
OS
OS
Hypervisor
Performance Isolation
Security Isolation
Reliability and Resiliency
Firewall
VPN
DNS
OS
OS
OS
Hypervisor
Fault Tolerance
Watchdog
High Available

虚拟机隔离部署



■ **1、I层管理隔离**
智能主机选取

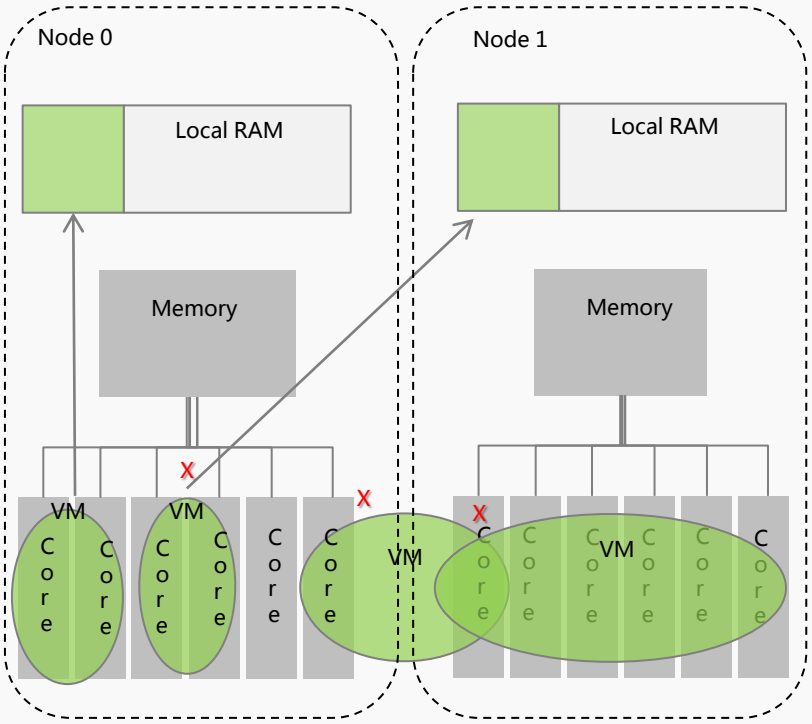
■ **2、计算资源隔离**
包括**VCPU**隔离、中断绑定、主进程隔离、虚拟机内存预占、**Qemu**内存隔离等

■ **3、存储资源隔离**
存储**QoS**、**Virtio-blk dataplane**

■ **4、网络资源隔离**
使用**VLAN/VXLAN**、**ACL**、**Openflow**流表规则或不同的转发平面设置，网络**QoS**

主机智能选取

亲和性调度



NUMA

- ✓ NUMA亲和性调度，禁止跨节点访问远端内存导致的业务时延
- ✓ 支持核绑定，禁止超分配
- ✓ 避免跨Node分配，减少切换消耗，提升性能

Nova中虚拟机信息记录数据表（示意）：

虚拟机	vmem	vcpu	cpuset	Node	主机id
vm1	5G	4	0~3	0	host1
…					

VCPU隔离

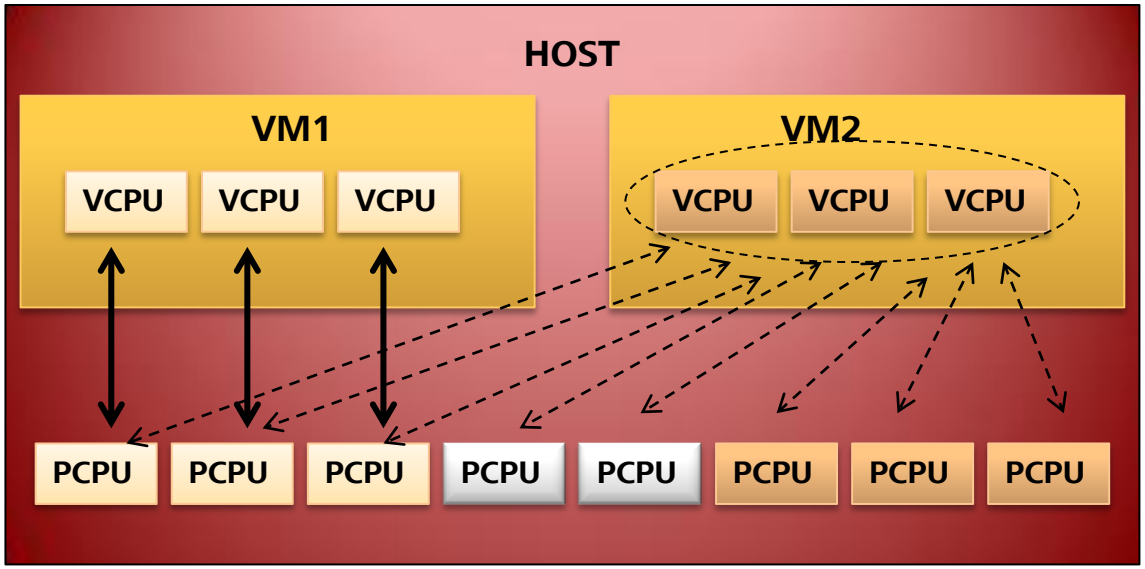
- **应用场景**

  用户希望把虚拟机的**VCPU**绑定在特定物理**CPU**上，**VCPU**只在绑定的物理**CPU**上调度，达到隔离**VCPU**并提升虚拟机性能的目的。如果没有作**VCPU**绑定，则虚拟机的**VCPU**可以在所有物理**CPU**上调度。

- **实现方式**

```
<cputune>
  <vcpupin vcpu='0' cpuset='0'/>
  <vcpupin vcpu='1' cpuset='1'/>
 <vcpupin vcpu='2' cpuset='2'/>
</cputune>
```

具体的绑定策略由用户来设定。



中断隔离

- **应用场景**

  用户需要对虚拟机的网卡中断做到相互隔离，互不影响，同时又可以提高网络收发包性能。包括两种场景：

  - 直通网卡 （**SR-IOV**，**PCI**直通）
  - 前后端网卡（**vhost-net/tap** & **virtio-net**）

如何做到网卡中断自动隔离绑定？

中断隔离

- **直通网卡 （SR-IOV，PCI直通）**

  两个概念：

  virq: 虚拟机内部看到的直通网卡对应的中断号

  pirq: 主机上看到的直通网卡对应的中断号

  主机上直通网卡中断的亲和性默认为全绑定。我们必须保证**pirq**能够自动绑定到**vcpu**对应的**pcpu**上面。

- **实现方式:**

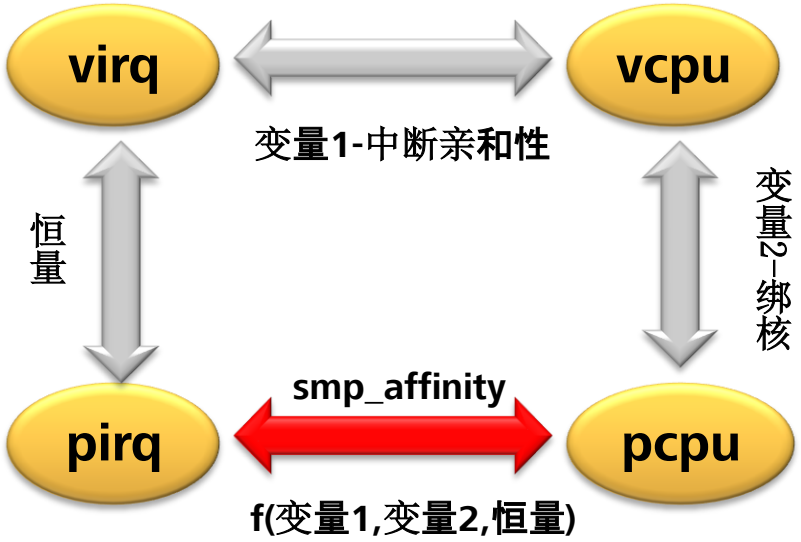
输入：用户设置**virq**的中断亲和性到**vcpuX**

输出：**pirq**绑定到**vcpuX**对应的**pcup**上面

如何自动获取**vcpu**绑定的**pcpu?**

如何自动获取**virq**与**vcpu**的亲和性关系**?**

如何自动获取**virq**与**pirq**的对应关系?



中断隔离

- 前后端网卡 （**vhost-net + virtio-net**）

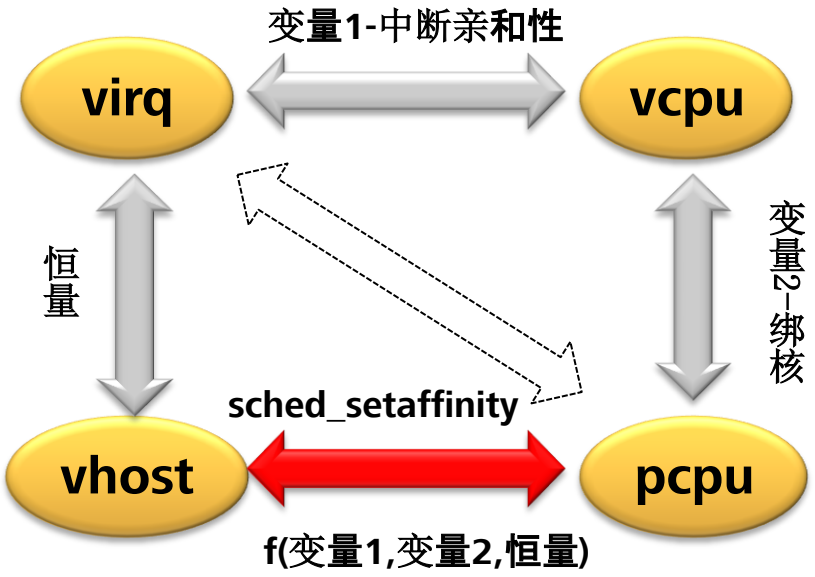
  在前后端场景下，主机没有**pirq**的概念，但是每个虚拟网卡**virtio-net**会对应一个**vhost**线程，该线程默认在主机上也是全绑定的，为了隔离虚拟机资源，我们如何保证**vhost**线程也能够自动绑定到**vcpu**对应的**pcpu**上面呢?

- **实现方式:**

输入：用户设置**virq**的中断亲和性到**vcpuX**

输出：**vhost**线程绑定到**vcpuX**对应的**pcup**上面

如何自动获取**virq**与**vhost**线程的对应关系?

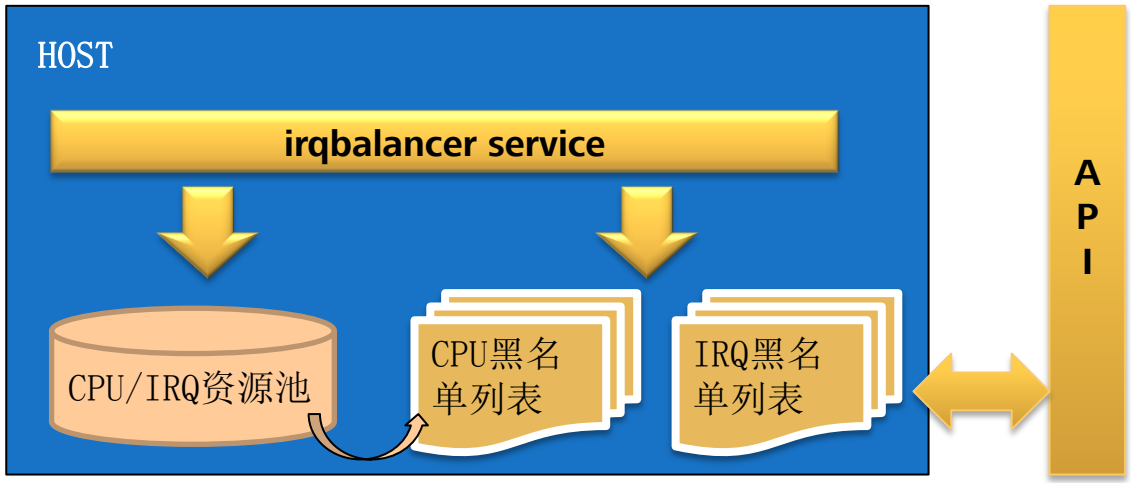


中断隔离

- 中断均衡优化

  irqbalance可以动态地平衡多个**cpu**上面的中断负载，来做到全部中断能均衡绑定到各个**cpu**，可以更好的防止网络带宽抖动。但这个是与虚拟机隔离部署相冲突的，所以如果要同时使用**irqbalance**功能，又能够实现对**CPU**和中断的隔离，就需要对**irqbalance**做优化，引入黑名单功能，主动脱离**irqbalance**的管理。

- **实现方式**



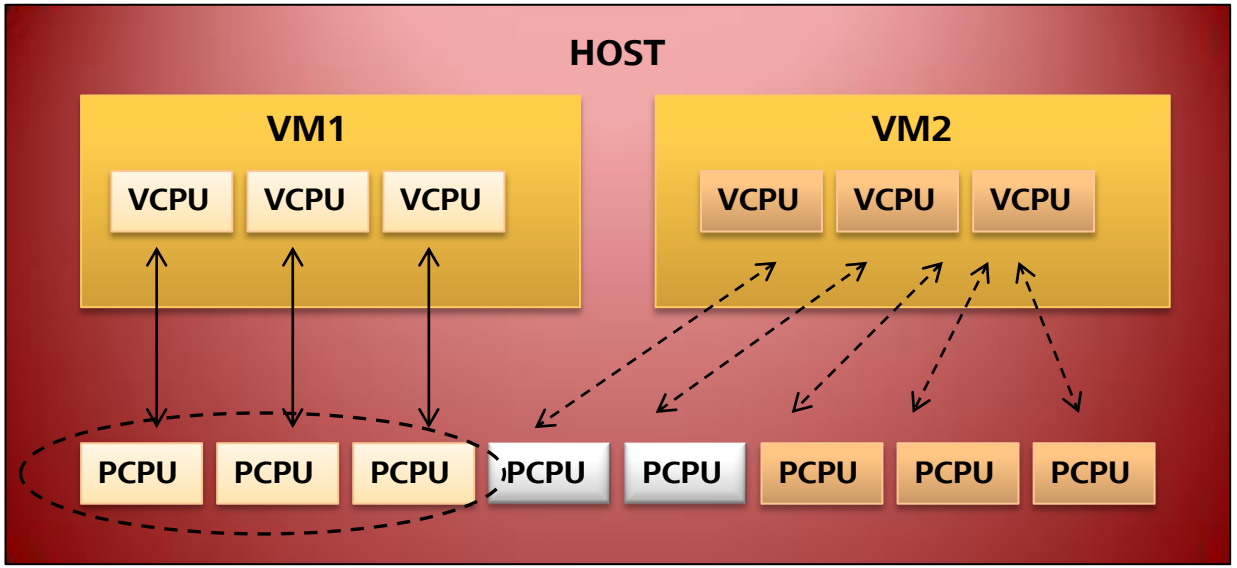
主机CPU隔离

- **应用场景**

  当业务上层要创建一台对实时性和性能要求高的虚拟机时（比如高精度时钟），需要将该虚拟机的**vcpu**所绑定的**pcpu**隔离出来，不运行**I**层程序且不进行物理中断处理，使得**vcpu**可以完全独占该**pcpu**，以达到实时并具有极高的调度处理能力。

- **实现方式**

cgroup隔离



QEMU进程隔离

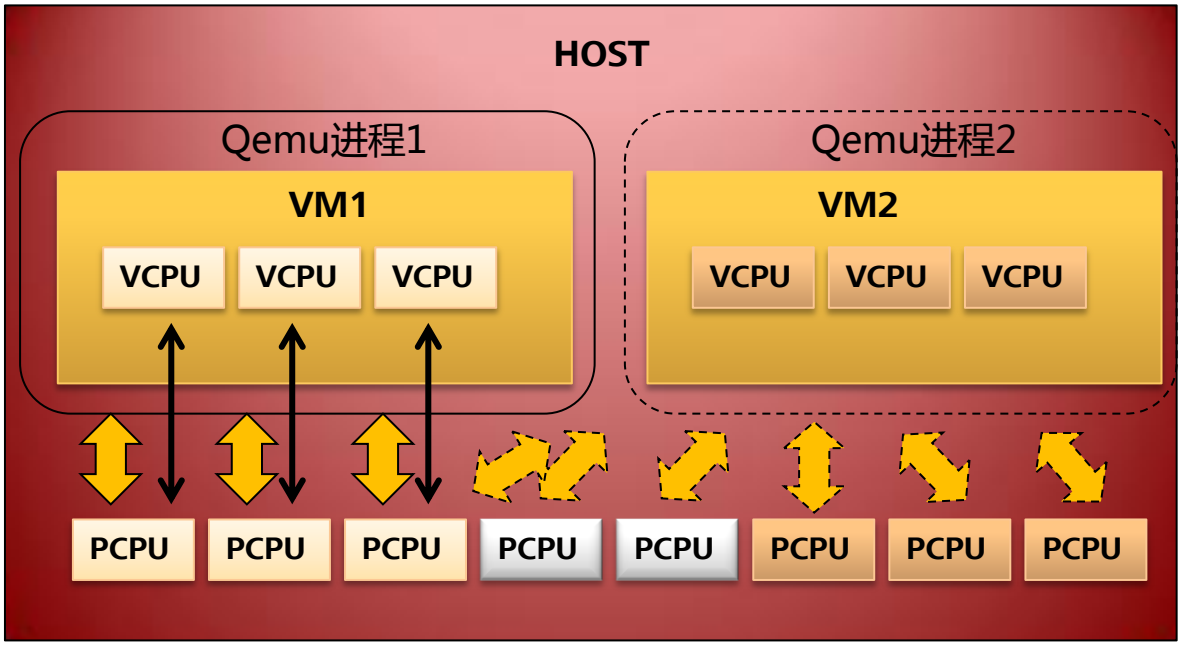
- **应用场景**

  对于运行计算密集业务的虚拟机，会利用较高的**CPU**资源。管理员可以静态配置虚拟机**Qemu**进程绑定到特定的**CPU**上，保证不会干扰到邻位的虚拟机。

- **实现方式**

```
<cputune>
  <emulatorpin cpuset="0-2"/>
</cputune>
```

cpu亲和性绑定

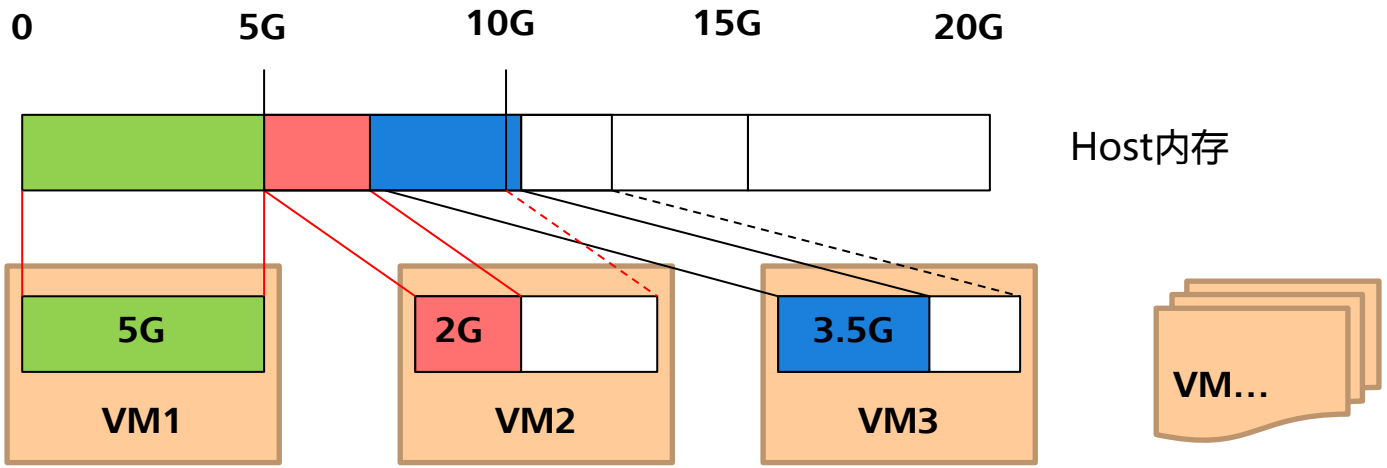


内存预占

- **概念**

  内存预占是预先占住并锁住分配给虚拟机的物理内存资源，以达到提高虚拟机的早期性能和提供给用户精确规划内存分配的能力、降低或者避免**VM**间的内存争用的特性，达到隔离虚拟机内存的目的。

- **实现方式**



存储QoS

- **概念**

  存储**QoS**指的是针对存储资源的服务质量保证。在虚拟化环境下，远端存储资源的读写能力和带宽都是有限的，**I/O**密集型虚拟机会抢占有限的存储资源。通过存储**QoS**功能设置对存储资源的**IO**上限，实现虚拟机存储资源的隔离，使得**IO**密集型虚拟机不会影响同一环境下其他虚拟机的**IO**性能。

- **实现方式**

  - Cgroup blkio controller

  - Qemu I/O throttle (file, LVM, NFS, Ceph)

网络隔离

- 网络规划

  VLAN/VxLAN，三层网络不同转发平面

- 防火墙

  ACL规则、**iptables**规则

- **Openflow**

  流表规则

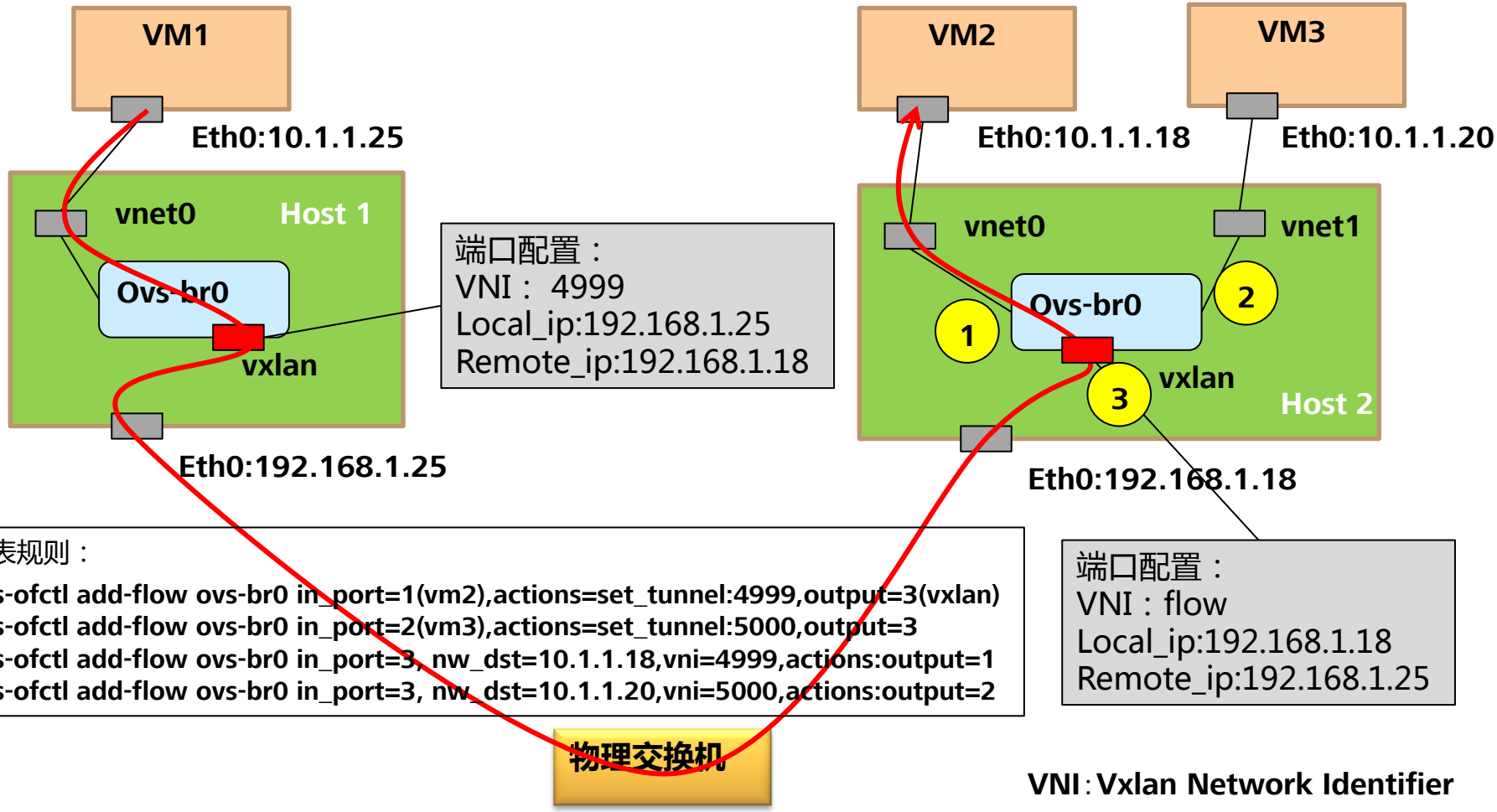
- **网络QoS**

  - Linux Traffic Control

  - Open vSwitch QoS

VxLAN+OpenFlow

VxLAN(Virtual Extensible LAN) 虚拟可扩展局域网



网络QoS

Why?

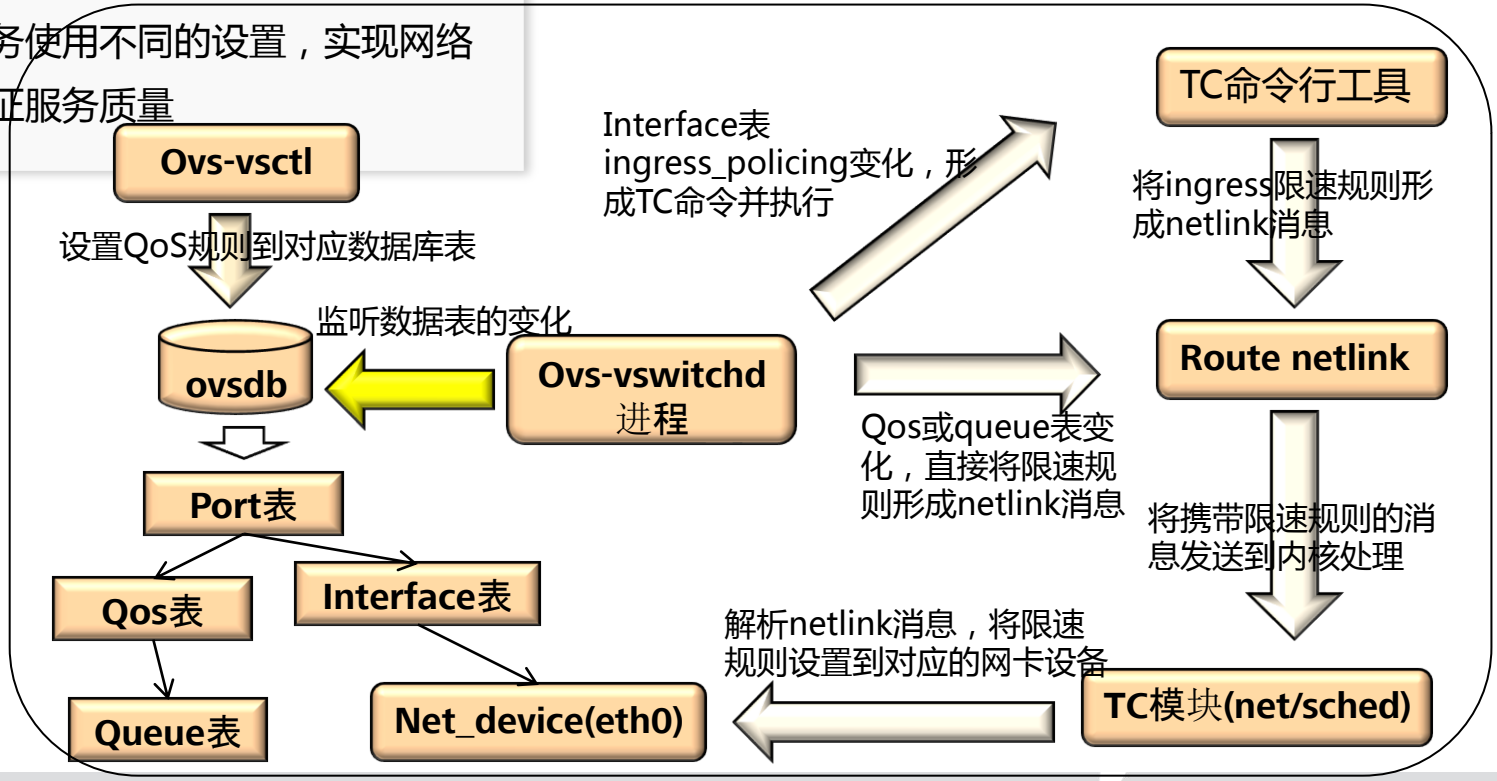
✓很多网络业务，如VoIP、视频点播、VPN、远程数据库访问等，对网络带宽、延迟、抖动比较敏感。

✓在虚拟化场景下，必须保证业务能够正常满足需求，这就需要通过QoS手段对网络数据流加以控制，对不同服务需求的业务使用不同的设置，实现网络带宽资源的隔离，保证服务质量

Open vSwitch QoS

✓端口限速

✓借助Linux的TC功能



Thank you

www.huawei.com