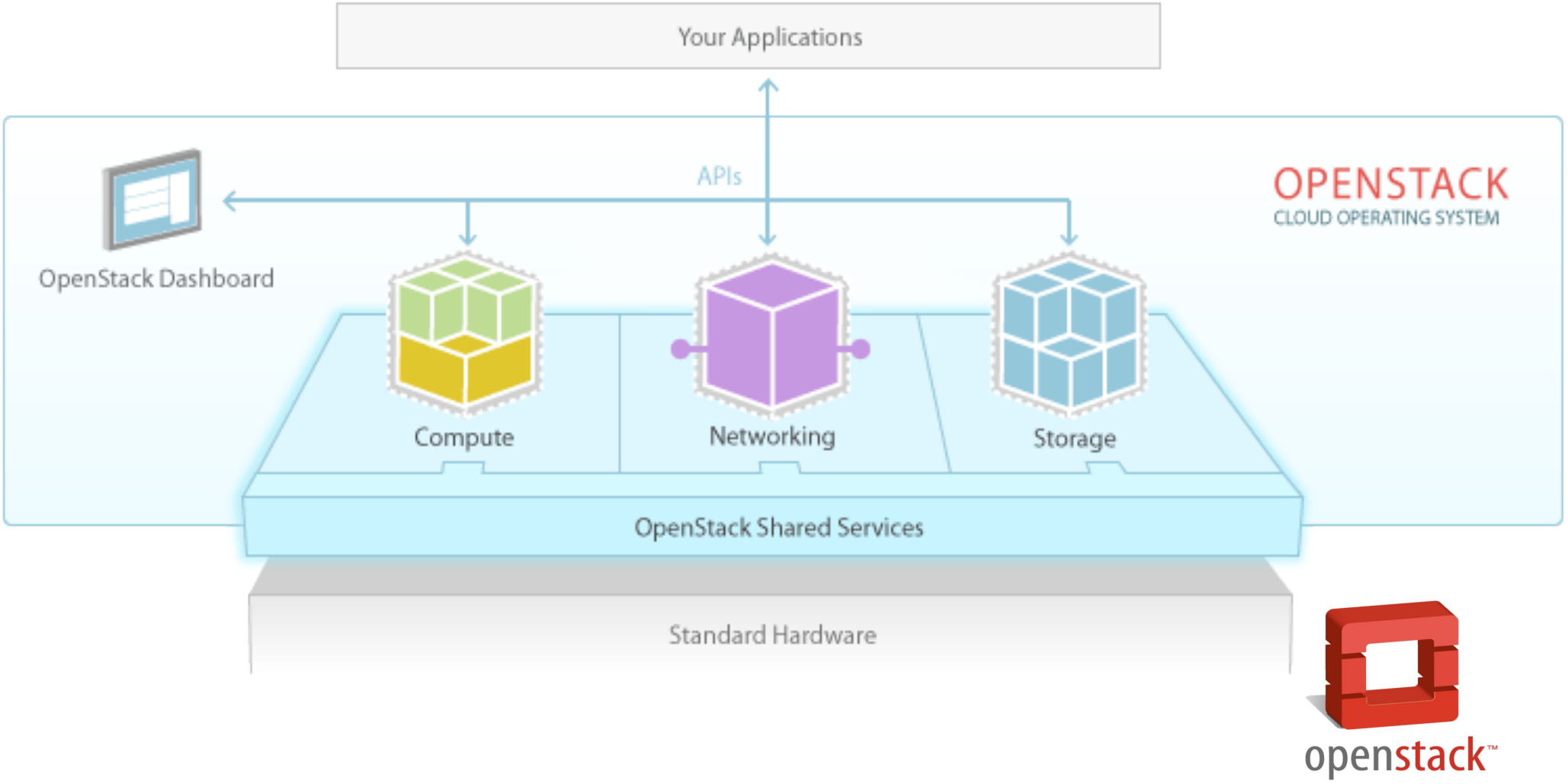
从 OpenStack 视角看 NFV

章宇
华为技术有限公司
云操作系统产品部

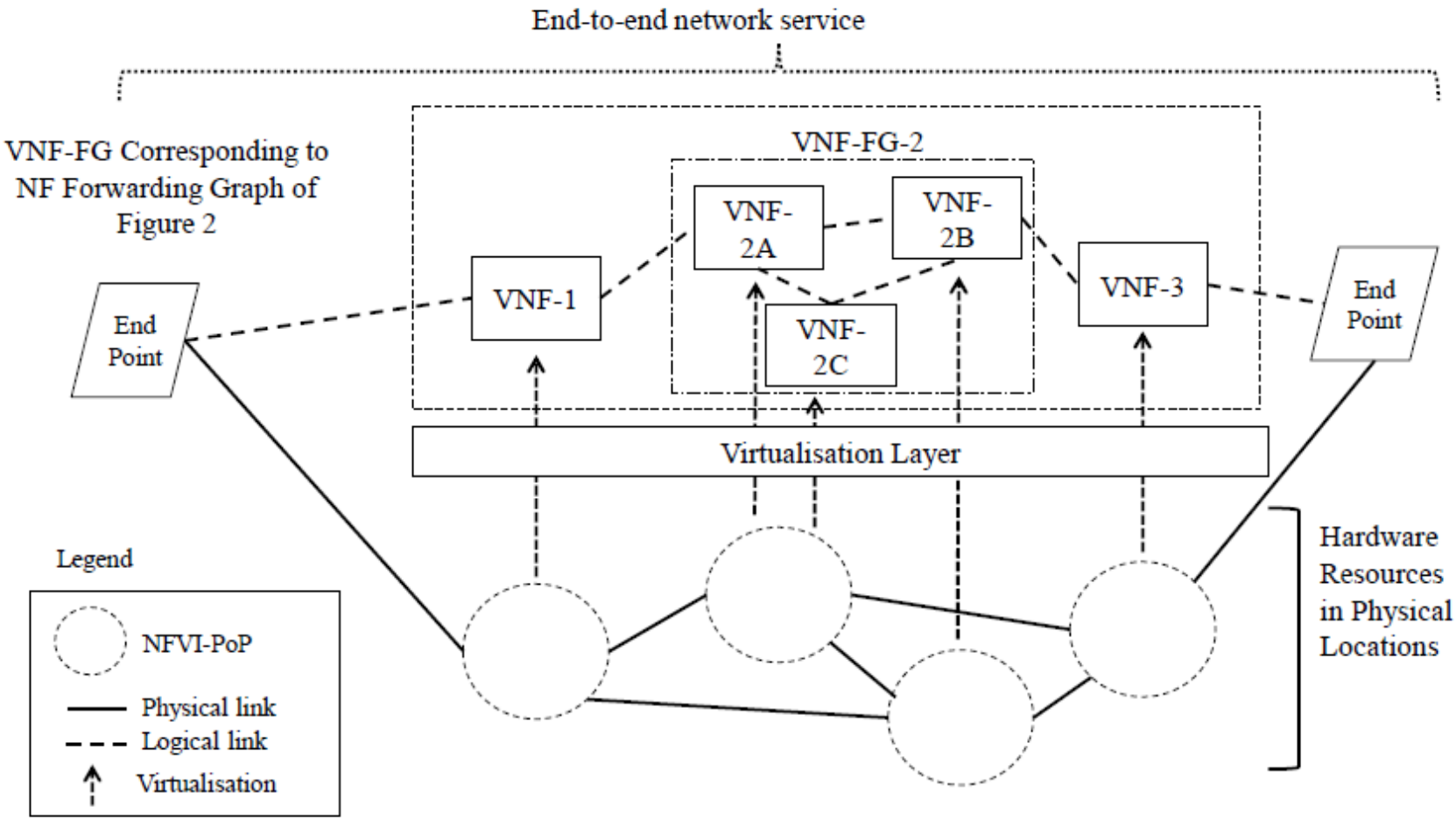
自报家门

- 章宇
- 2002 年和 2007 年先后于清华大学电子工程系获得工学学士及博士学位
- 目前于华为技术有限公司云操作系统产品部担任架构师，并参与 FusionSphere 5.0 & 5.1 产品研发
- 一棹凌烟 @ 微博

引子： OpenStack



引子： NFV



ETSI

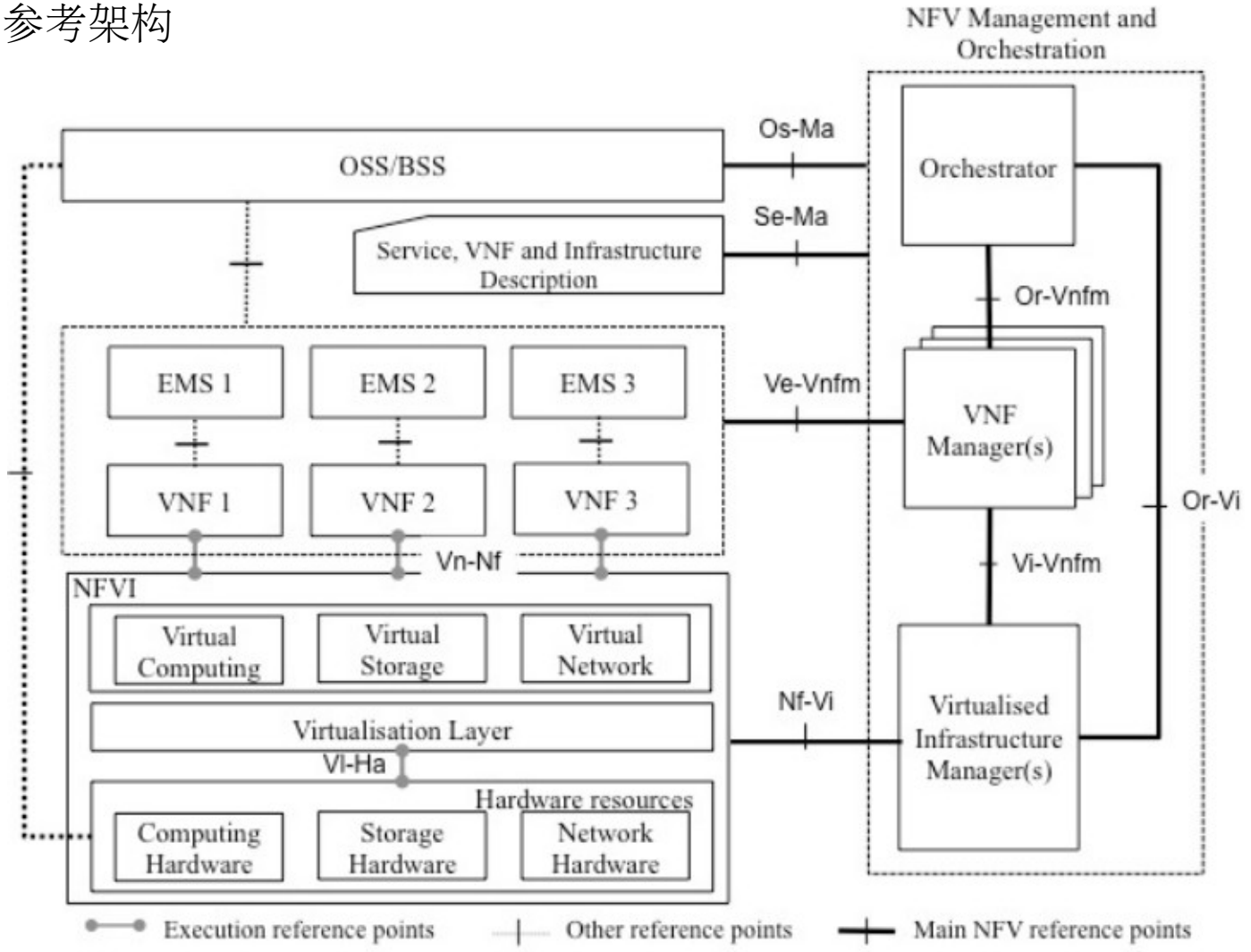
World Class Standards

Outline

- **NFV 与 OpenStack 的联系**
- NFV 对 OpenStack 的技术诉求
- OpenStack 针对 NFV 的增强
- 小结与展望

技术上的天然联系

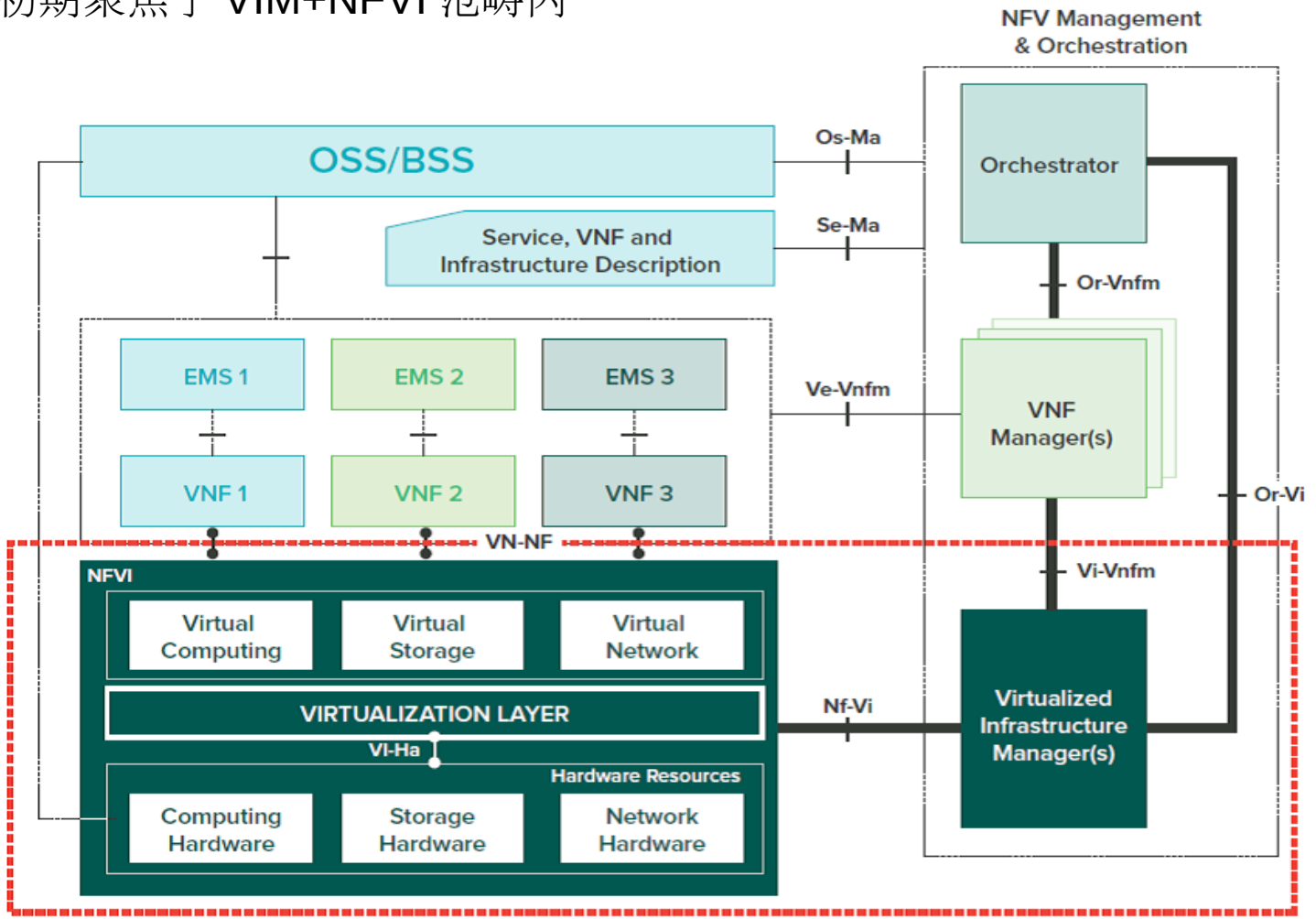
- ETSI NFV 参考架构



VIM + NFVI 基本等同于 OpenStack 管理下的 IaaS

NFV 生态系统的努力—— OPNFV

- Open Platform for NFV
- 2014 年正式成立， Linux 基金会所属项目
- 众多电信运营商、电信设备厂商和 IT 厂商共同发起
- 旨在为 NFV 提供基于开源软件的、电信级的 NFV 参考平台
- 初期聚焦于 VIM+NFVI 范畴内



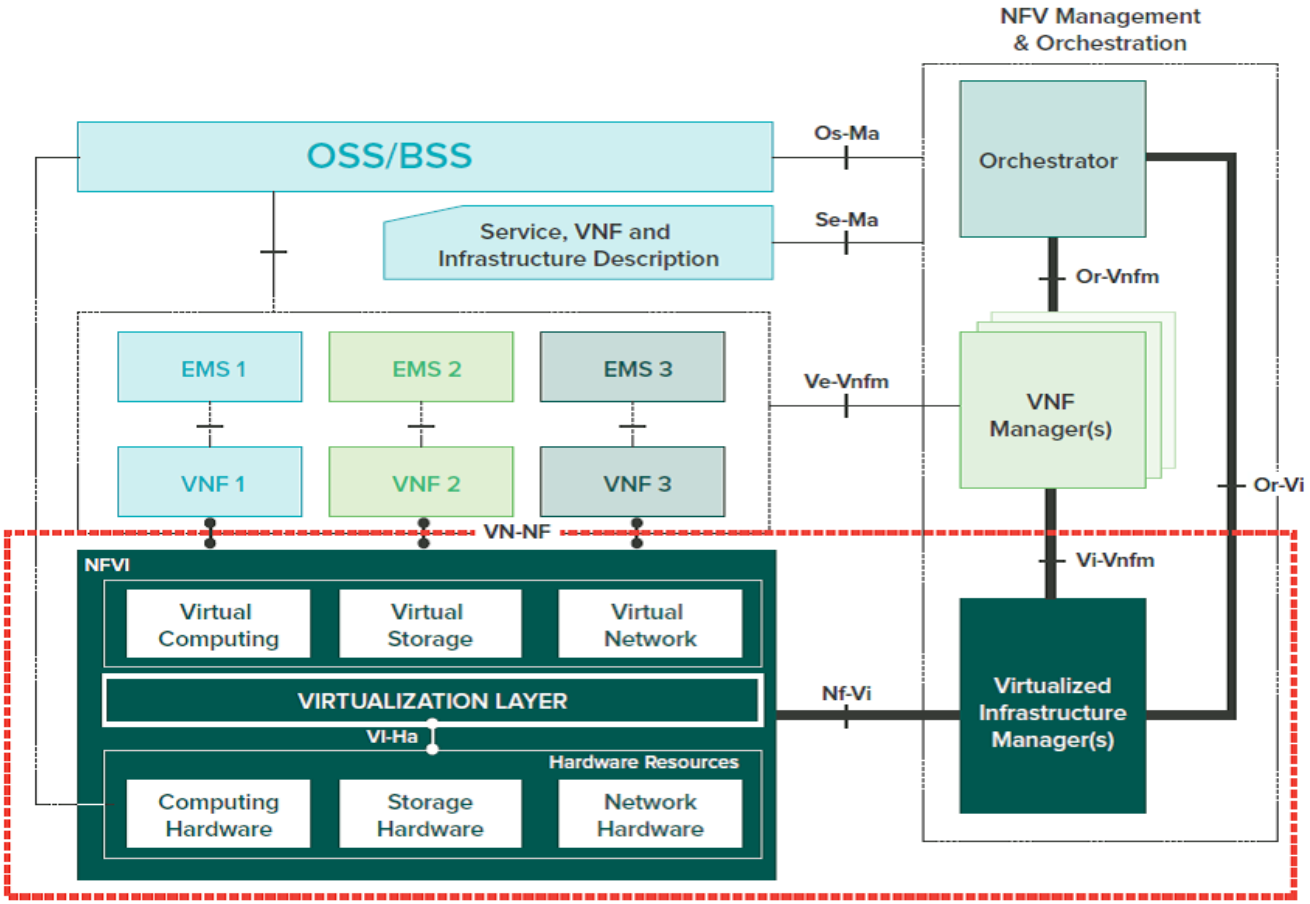
OpenStack 是 OPNFV 在 VIM 模块实现中，事实上的唯一选择

Outline

- NFV 与 OpenStack 的联系
- **NFV 对 OpenStack 的技术诉求**
- OpenStack 针对 NFV 的增强
- 小结与展望

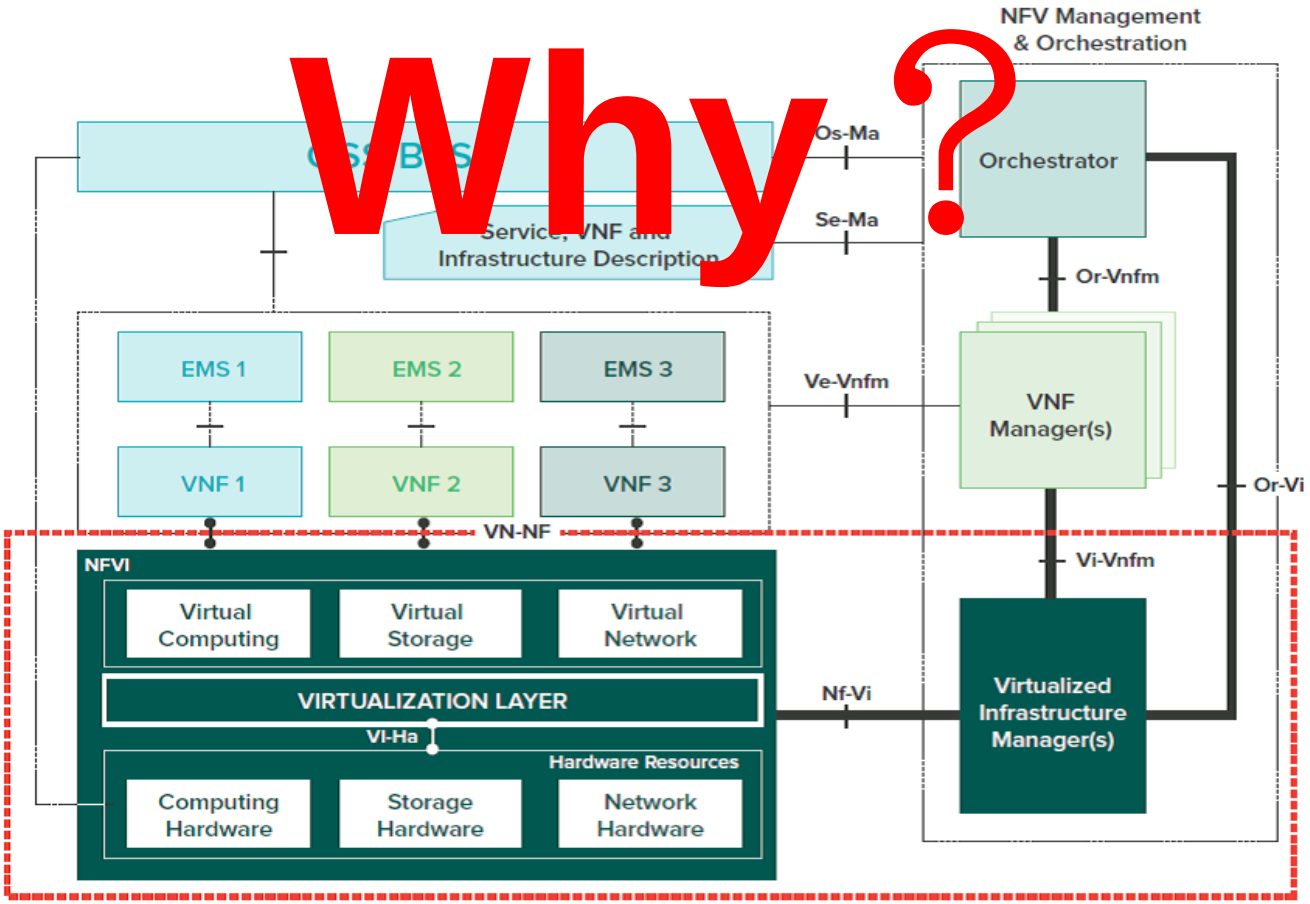
NFV 对 OpenStack 的三大基本诉求

- 业务面高性能
  - 虚拟机
  - 网络转发
- 高可用、高可靠
  - 控制面
  - 业务面
- 易运维、易管理
  - 自动化



NFV 对 OpenStack 的三大基本诉求

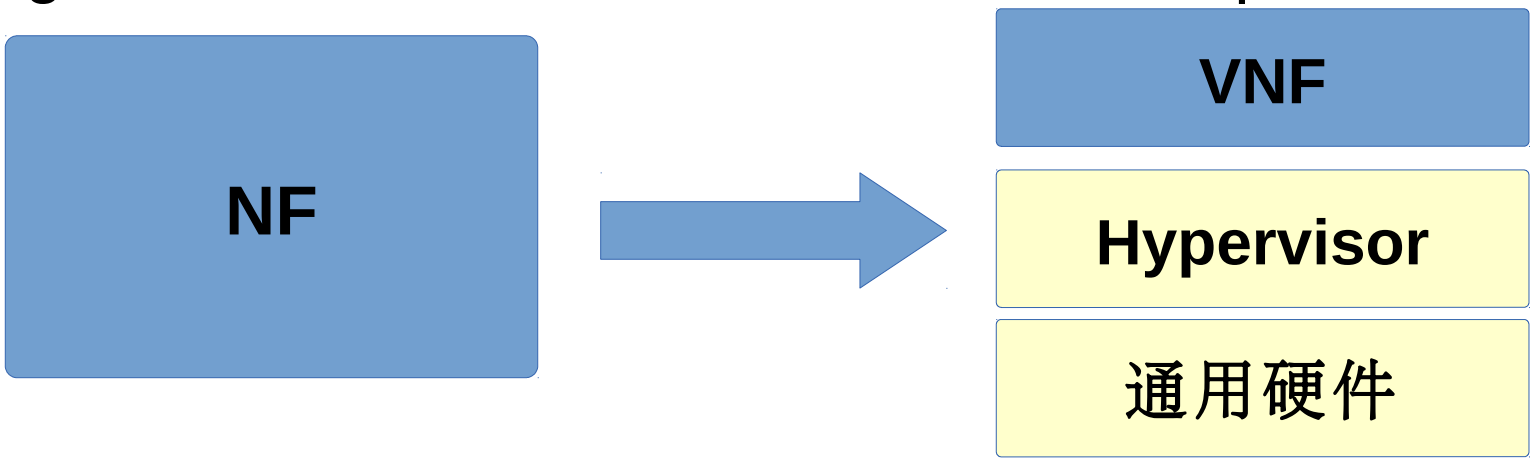
- 业务面高性能
  - 虚拟机
  - 网络转发
- 高可用、高可靠
  - 控制面
  - 业务面
- 易运维、易管理
  - 自动化



从 NFV 的产生说起

- 电信业务自身的特点
  - 高性能指标要求
  - 高可靠性可用性要求
    - 至少 5 个 9
    - E.g. 控制节点全部故障不影响业务
  - 易运维易管理要求

- 已有技术体制的演进过程
  - 既有架构、代码的平滑迁移和复用
    - E.g. 单一 VM 在同一 network 内接入两个 port



从 NFV 的产生说起

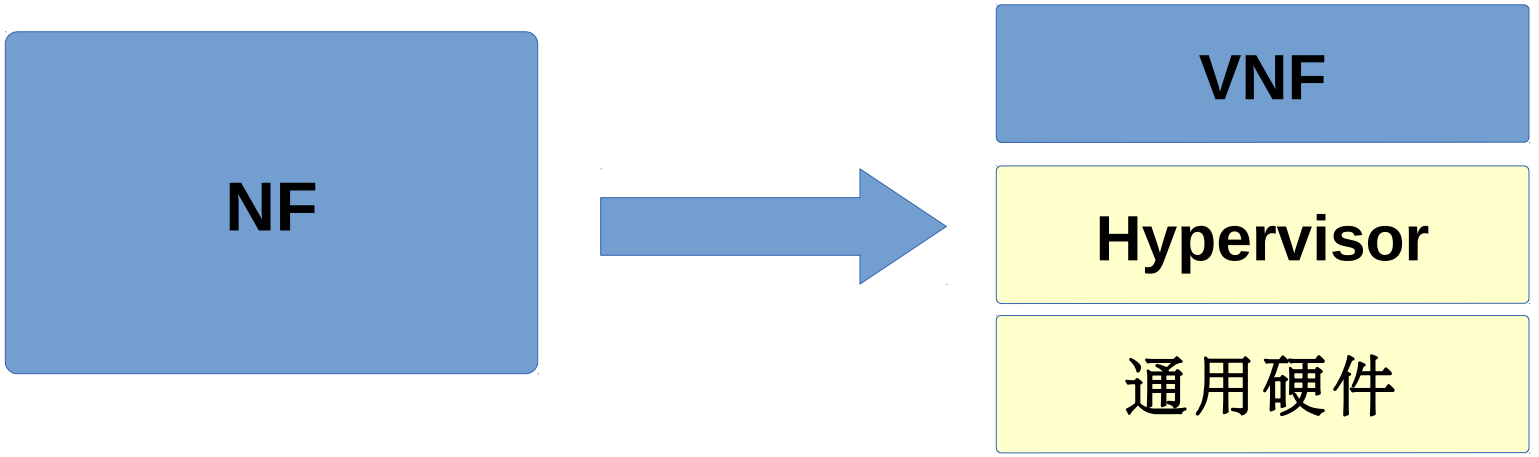
- 电信业务自身的特点
  - 高性能指标要求
  - 高可靠性可用性要求
    - 至少 5 个 9
    - E.g. 控制节点全部故障不影
  - 易运维易管理要求

从 OpenStack 的视角看：

1. “上什么山，唱什么歌”

2. “罗马不是一天建成的”

- 已有技术体制的演进过程
  - 既有架构、代码的平滑迁移和复用
    - E.g. 单一 VM 在同一 network 内接入两个 port



问题： OpenStack 适合 NFV 吗？

云计算 ⇨

- IT 能力服务化
- 按需使用，按量计费
- 多租户隔离
- 。。。

Vs

- 环境隔离，资源复用
- 降低隔离损耗，提升运行效率
- 提供高级虚拟化特性
- 。。。

⇦ **虚拟化**

问题： OpenStack 适合 NFV 吗？

IT 与 CT 的融合

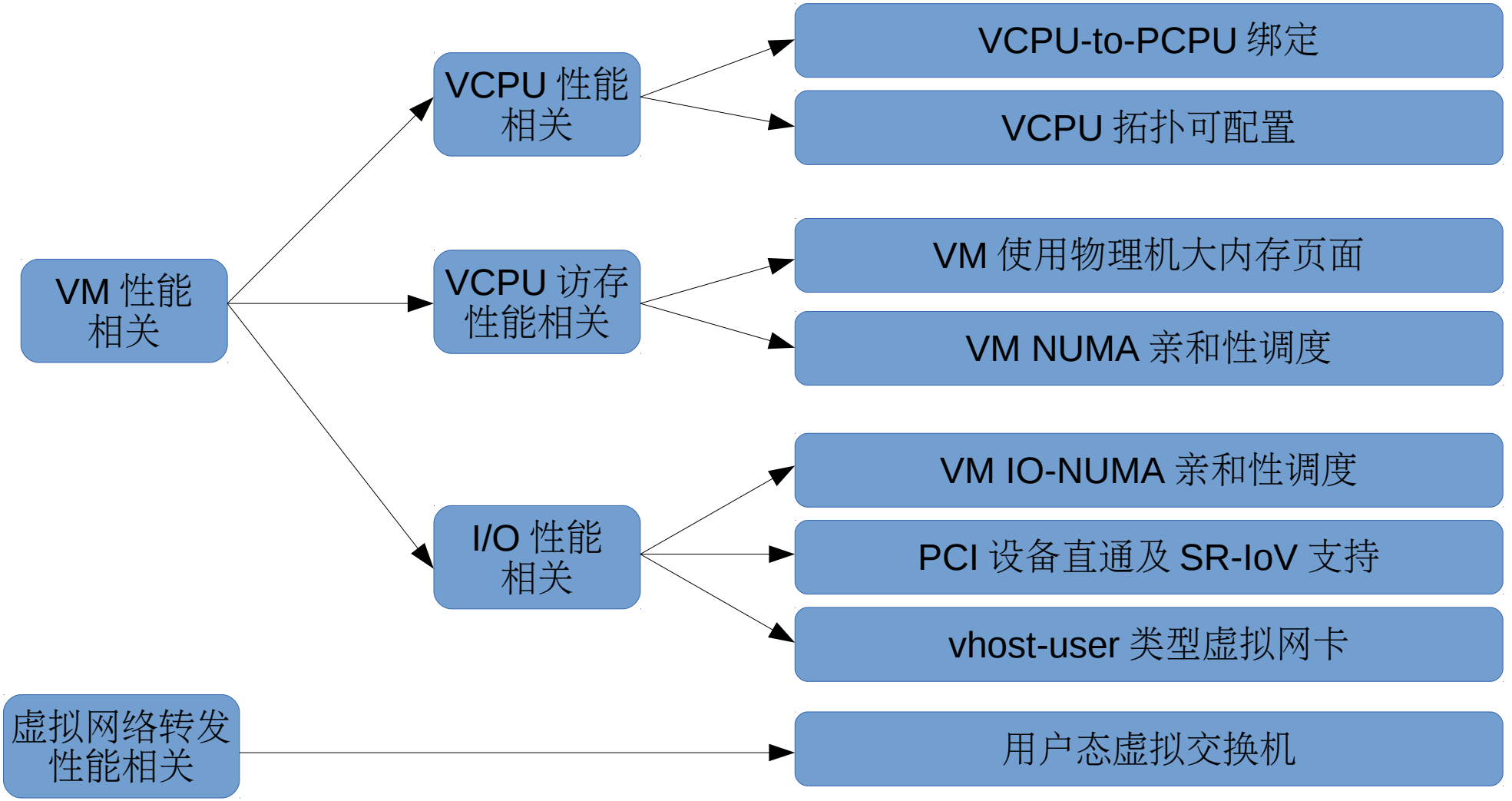
NFV 对 OpenStack 的其他诉求

- 高级网络功能
  - L4-L7 的网络能力
  - Service chain
- 更复杂精确的 VM 调度能力
  - 在资源池规模小、资源使用率高、约束条件多的情况下，对 VM 进行调度
- ...

Outline

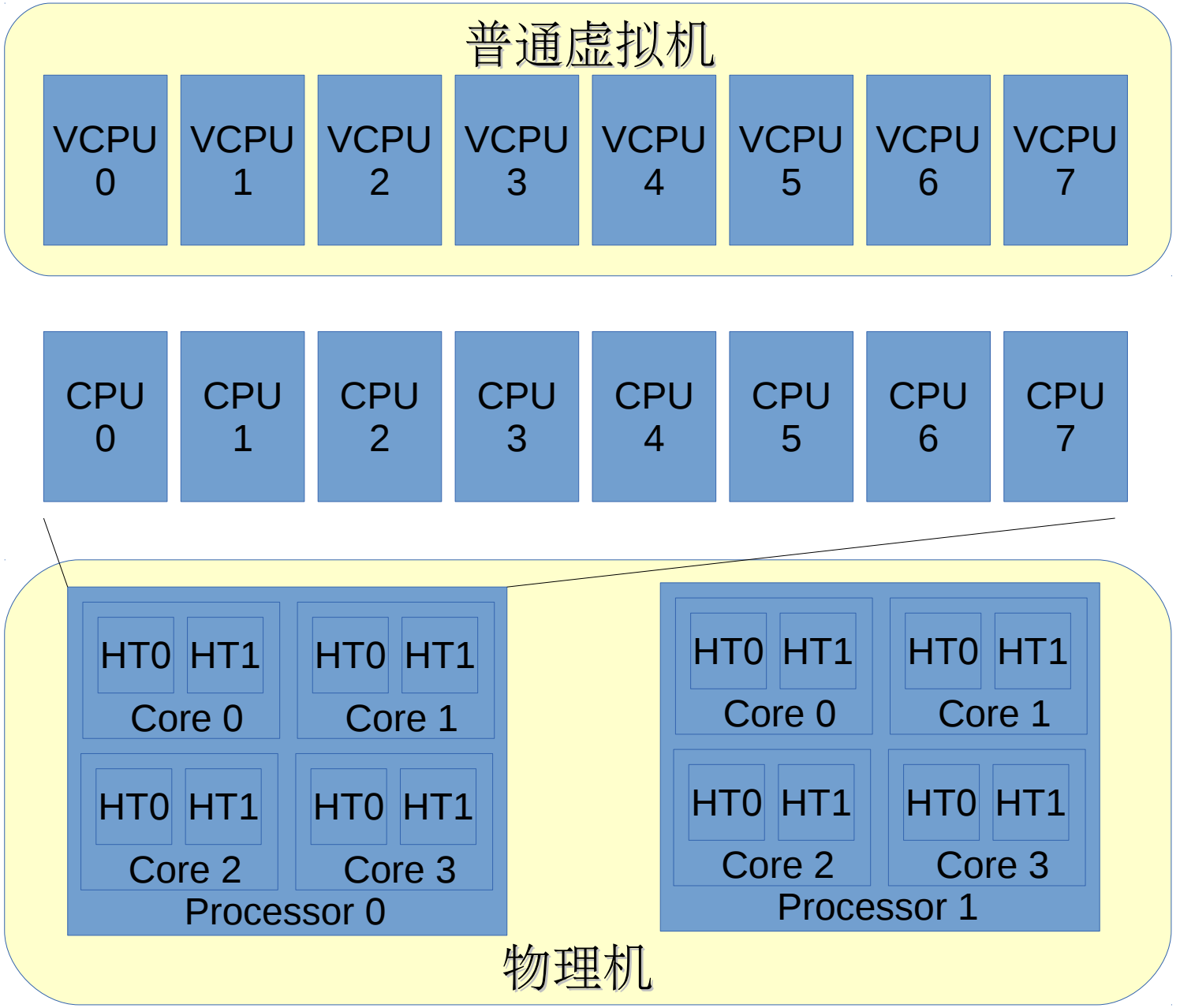
- NFV 与 OpenStack 的联系
- NFV 对 OpenStack 的技术诉求
- **OpenStack 针对 NFV 的增强**
- 小结与展望

OpenStack 针对业务面性能的主要增强



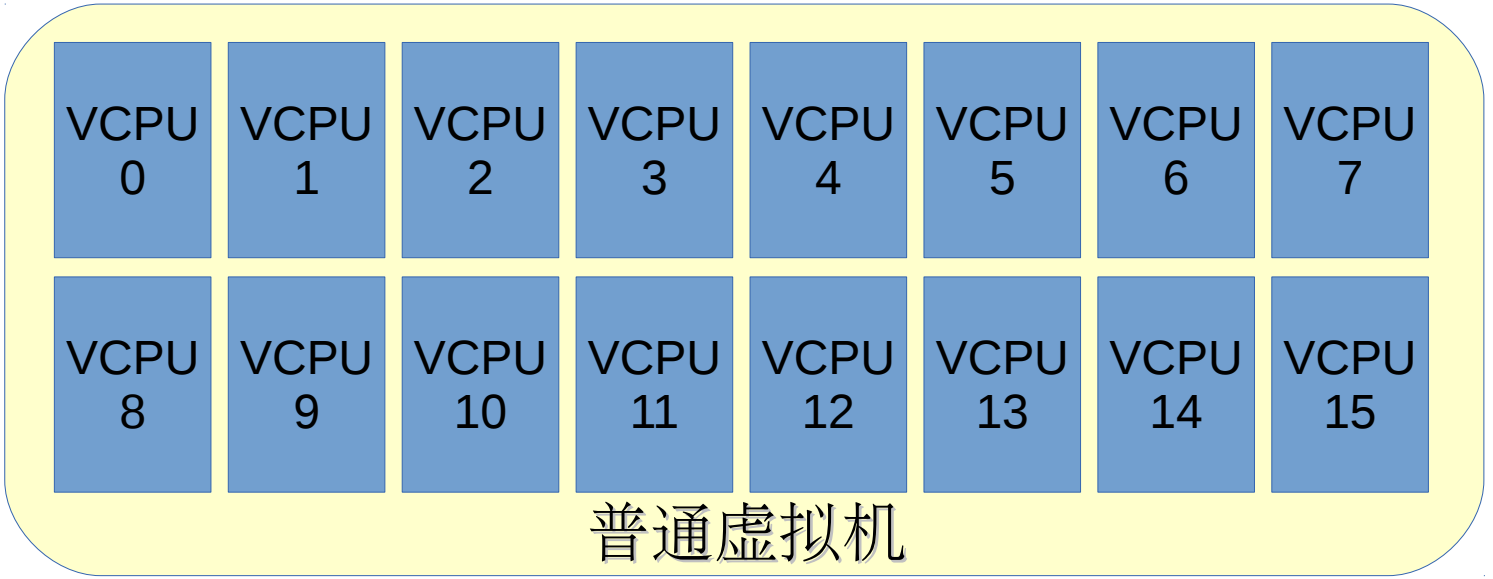
- 主要针对 OpenStack + KVM 的场景
- 集中于 OpenStack Juno 和 Kilo 版本周期内在上游代码中实现，主要在 Nova 和 Neutron 中
- 本质上是对虚拟化性能优化技术的充分利用

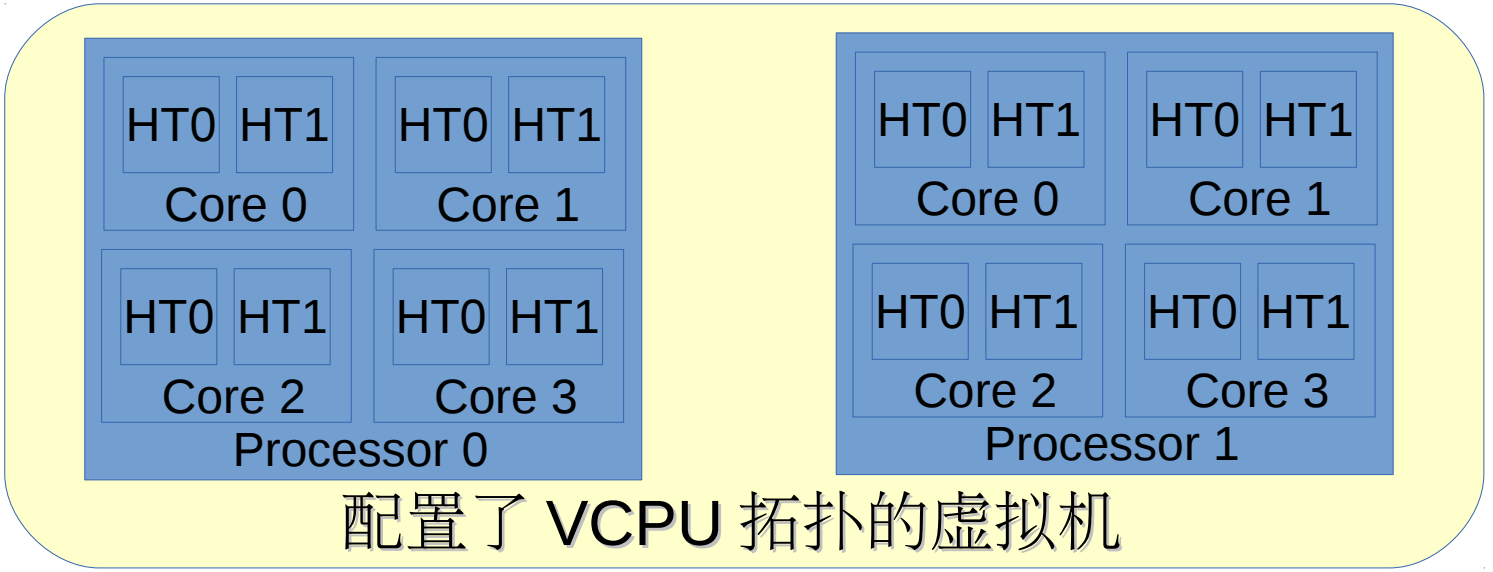
VCPU 到 PCPU 绑定



- 将 VM 的 VCPU 对应的逻辑线程绑定在物理处理器 / 硬件线程上，避免 VCPU 线程的漂移和资源争抢，同时提升 cache 命中率
- 需统筹考虑一台物理机上的所有 VM 的情况，避免非绑定的 VM 漂移过来争抢资源
- QEMU 占用的资源还需单独处理

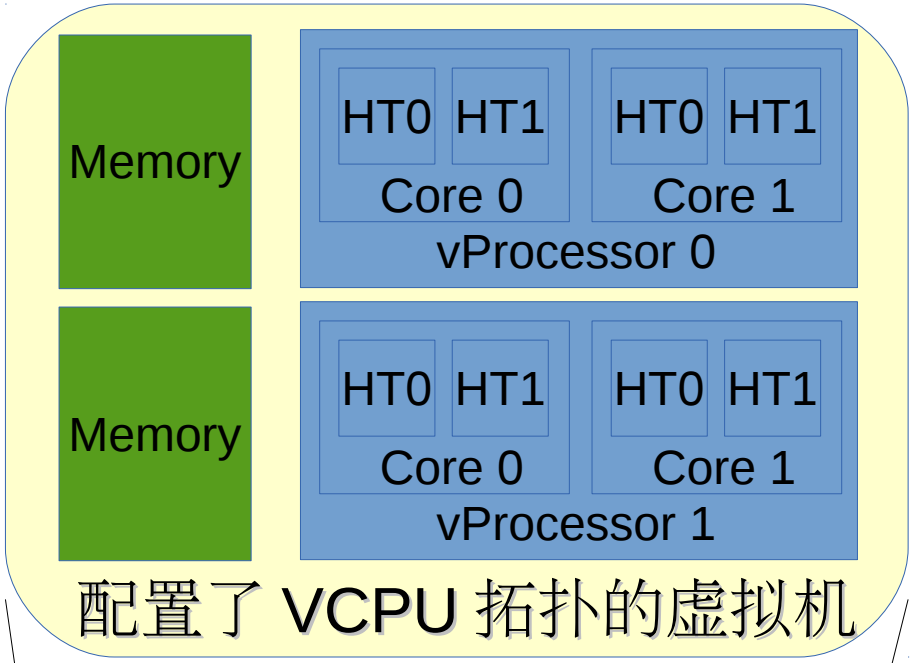
VCPU 拓扑配置



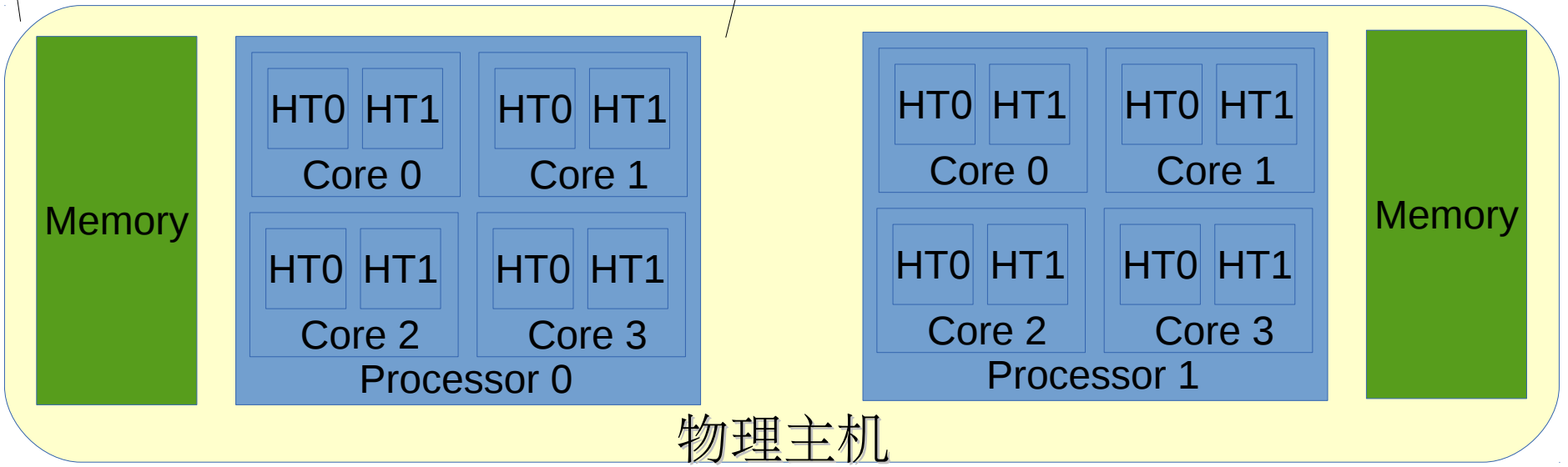


- 让 VM 虚拟出处理器具备的拓扑结构，以便应用软件感知与利用
- 可与其他特性（NUMA 亲和性，物理 CPU 绑定）结合使用，共同提升 VCPU 性能

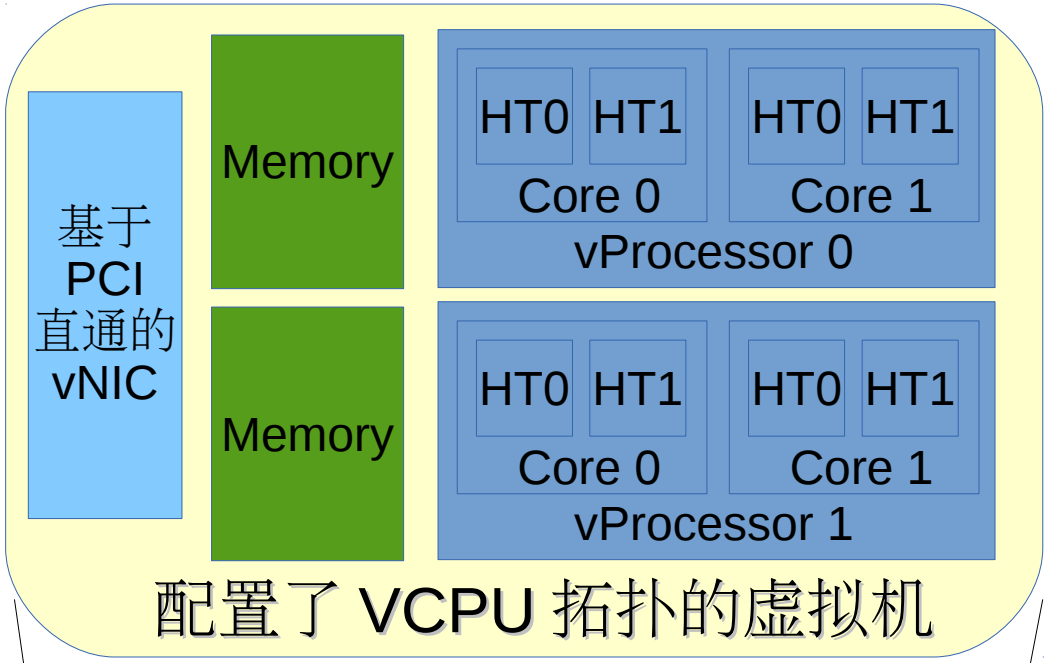
VM NUMA 亲和性调度



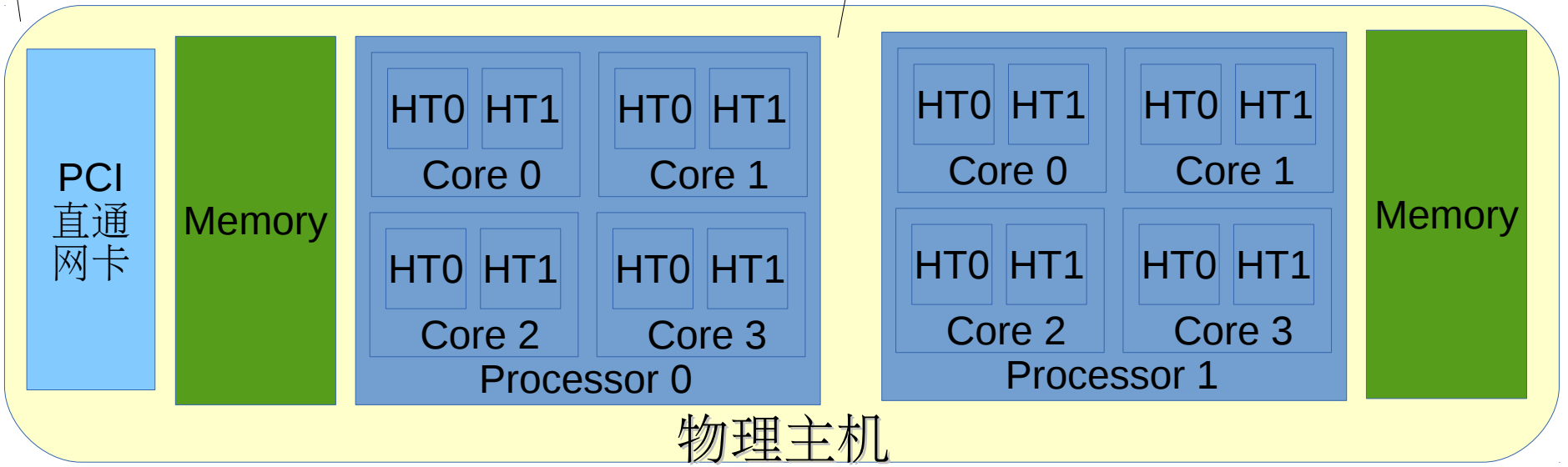
- 常用 case： 将 VM 的多个虚拟处理器及相关内存（即多个虚拟 NUMA 节点）映射到物理主机的同一个物理处理器及相关内存（一个物理 NUMA 节点）上，以避免跨物理 NUMA 节点的数据传输延迟和内存访问延迟
- 如有必要，也可以将多个虚拟 NUMA 节点分别映射到多个物理 NUMA 节点上，以保持 VM 和物理机的特性一致
- 应考虑和 VCPU 拓扑配置、 VCPU-to-PCPU 绑定特性一起使用



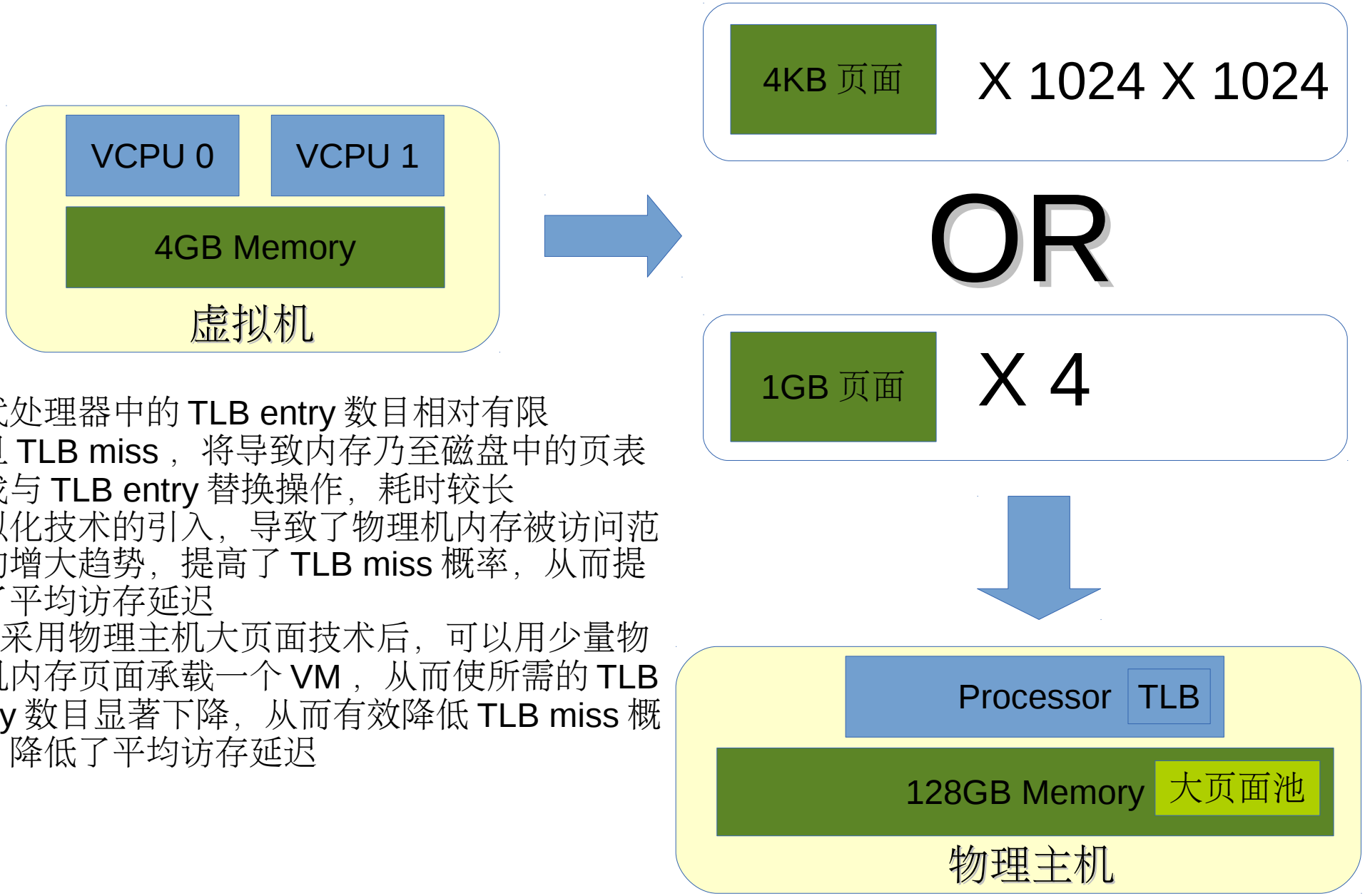
VM IO-NUMA 亲和性调度



- 常用 case：将使用了 PCI 直通网卡的 VM，调度到“靠近”该网卡的物理主机 NUMA 节点上，以尽可能减小 VM 上应用的网络通信延迟



VM 使用物理主机大页面内存



- 现代处理器中的 TLB entry 数目相对有限
- 一旦 TLB miss，将导致内存乃至磁盘中的页表查找与 TLB entry 替换操作，耗时较长
- 虚拟化技术的引入，导致了物理机内存被访问范围的增大趋势，提高了 TLB miss 概率，从而提升了平均访存延迟
- VM 采用物理主机大页面技术后，可以用少量物理机内存页面承载一个 VM，从而使所需的 TLB entry 数目显著下降，从而有效降低 TLB miss 概率，降低了平均访存延迟

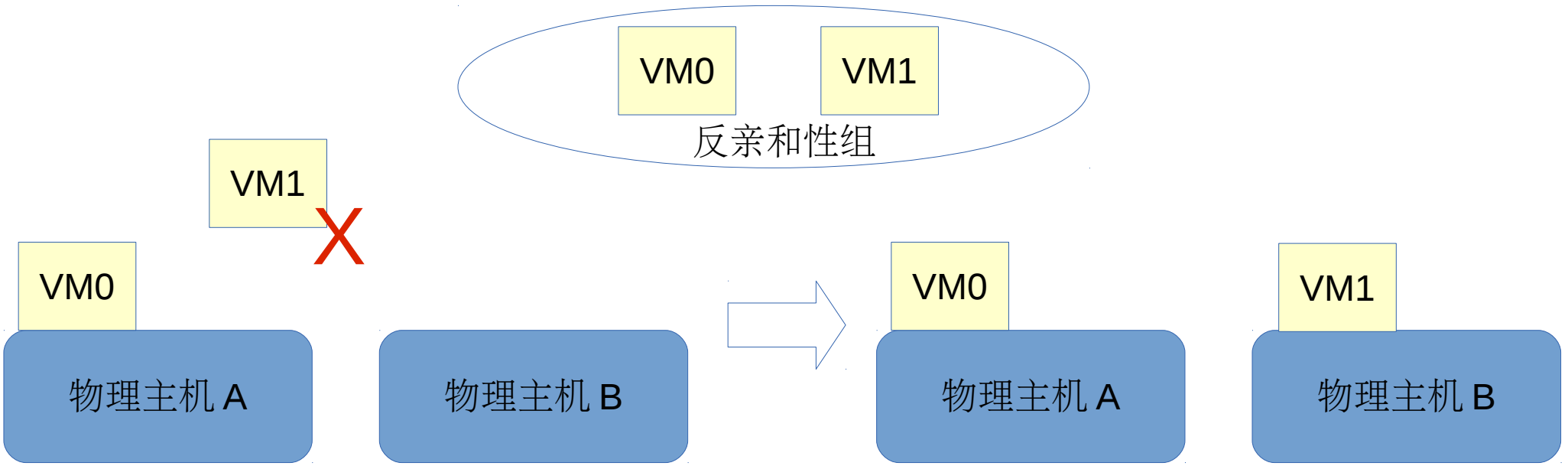
性能相关扩展特性的核心思路

榨干 KVM 的每一滴血 ...

NFV 对 OpenStack 的业务面可靠性需求举例——VM 反亲和性调度

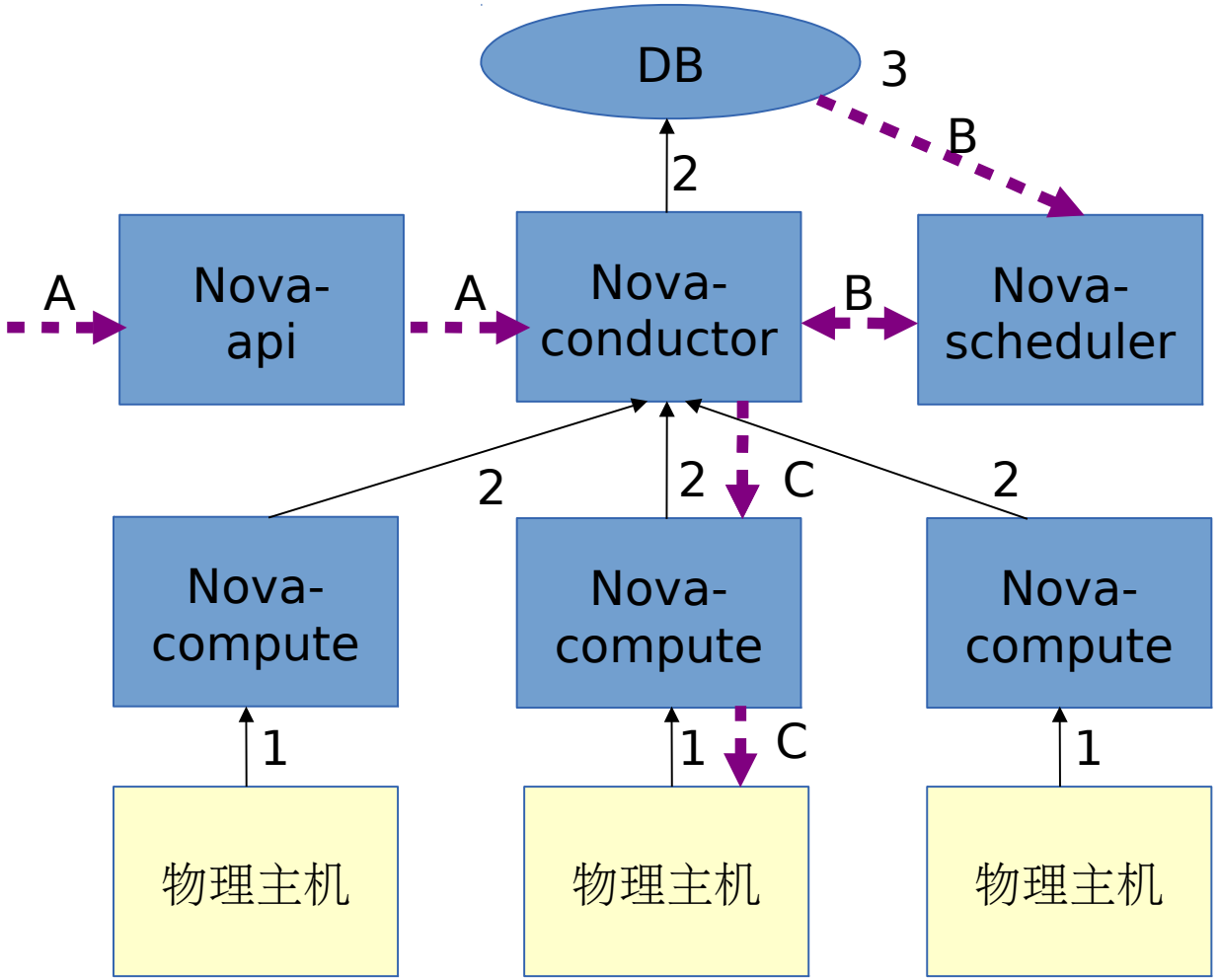
- 提出背景
  - 大量 VNF 通过双 VM 主备部署方式保证业务面 HA
  - 如果主备 VM 被部署到同一个物理主机上，则一旦该主机故障，会导致业务中断
  - 非 NFV 需求实例：在 VM 上运行 Hadoop，同一数据的三副本不应在同一物理主机上

- 实现方式
  - 在 Nova 中实现了计算节点本地的冲突检测与调度重试机制



由于 Nova 原生调度能力的限制，在并发创建一个组内的多台 VM 时，部分 VM 可能多次重试。

Nova 的资源管理与调度机制



资源刷新流程

- 1： 资源发现
- 2： 资源上报与刷新
- 3： 资源记录

资源申请流程

- A： 资源申请
- B： 资源调度
- C： 资源占用

- NFV 对 OpenStack 的大量诉求均牵扯到对 Nova 的资源管理和调度机制的扩展
  - NUMA 亲和性、IO-NUMA 亲和性、大页面、位置反亲和性等
- 本质原因：引入了新的资源类性 / 属性，以及新的调度需求，需要 Nova 相应扩展资源管理与调度能力

NFV 在 OpenStack 中引入特性的问题

- 若干原生特性还不成熟，不能直接应用于产品化场景中
  - 例如：内存大页面

- 部分原生特性之间可能相互制约
  - 例 1： SR-IoV 网卡与 VM 迁移
  - 例 2： IO-NUMA 亲和性与 CPU 绑定

- 造成了不同节点间的资源异构特性，为安装、运维、使用带来了限制与不便
  - 不同节点上的资源不同，多个异构资源池
    - 有 / 无大页面资源池
    - 有 / 无 SR-IoV 网卡
    - 有 / 无用户态虚拟交换机
  - 复杂的网络组网方案
    - 可能出现同一物理节点上出现多于一种网络资源

- 整体方案上需要精心设计，审慎组合使用各类原生特性

Outline

- NFV 与 OpenStack 的联系
- NFV 对 OpenStack 的技术诉求
- OpenStack 针对 NFV 的增强
- 小结与展望

在一起，为明天

- 技术演进和商业利益双重驱动下的必然产物
- 实际落地的速度，最终取决于主要客户群体的决心和技术成熟的速度

谢谢各位！