支撑华为云计算的虚拟化关键技术

杨晓伟

www.huawei.com

HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

目录

- 虚拟化背景、热点与趋势
- UVP在云计算中的应用及关键技术

虚拟化的定义

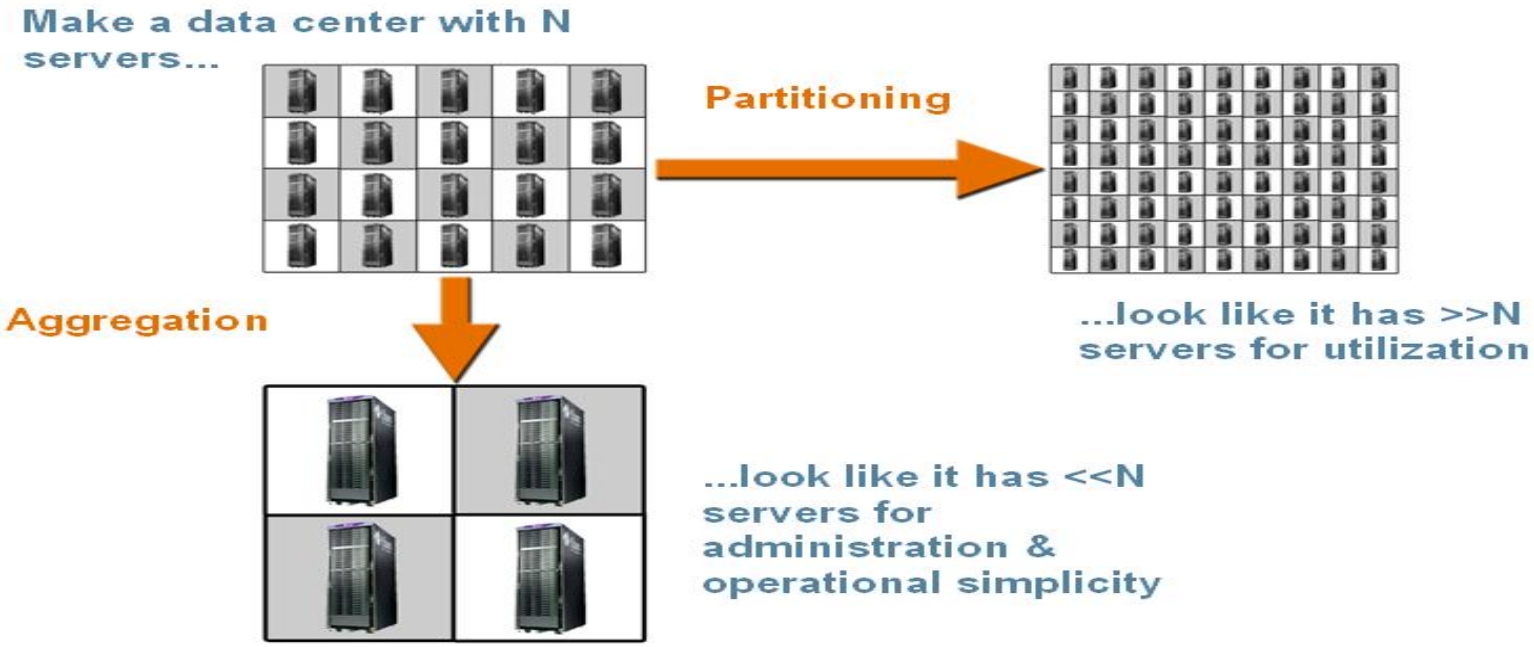
虚拟化一般意义上是指对计算机物理资源的进行抽象的手段；

虚拟化技术主要通过三种方式：

硬件仿真：指令集翻译，多用于开发环境中（如QEMU等）

分割：应用最为普遍（如VMware ESX、 XEN、 Hyper-V、KVM等）

聚合：应用相对较少（如ScaleMP/3Leaf等）



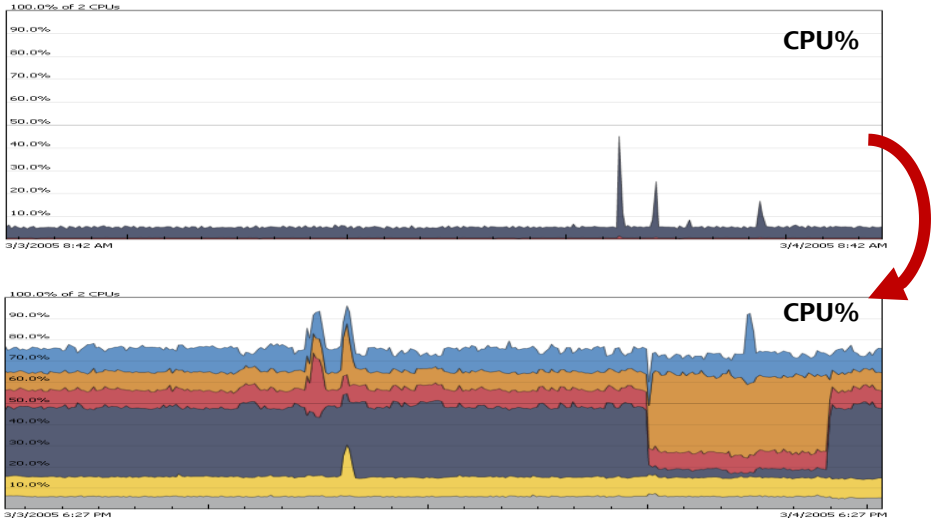
虚拟化1.0 = 整合

虚拟化1.0特征

抽象虚拟化硬件平台，**软硬件解耦**，减弱**OS**对硬件的敏感度：
1）CPU的前后兼容性；
2）IO设备的抽象模拟
异构应用之间**更彻底的隔离性**，支持多应用无修改地在同一硬件和平共处：
1）故障与安全隔离；
2）资源配额的隔离（**QoS**）

虚拟化1.0应用

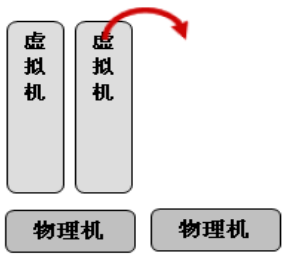
服务器整合
•35%--75% TCO 节省：降低**40**%软件硬件成本，降低**70-80%**运营成本
•提高运营效率，部署时间从小时级到分钟级
•提高服务水平：将所有服务器作为大的资源统一进行管理，自动态资源调配；无中断的按需扩容



2006以前：虚拟化1.0。虚拟化基本特征：抽象、解耦、隔离
主要应用：服务器整合--提高服务器的利用率，降低TCO

虚拟化2.0 = 敏捷

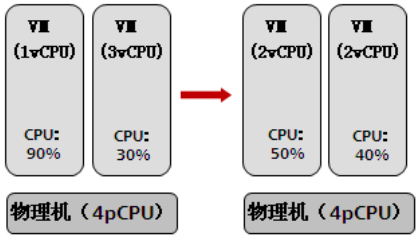
热迁移



虚拟机（应用）在多台物理服务器之间透明移动，**业务不中断**，实现**跨物理机的错峰削谷**

- **DRS**(动态负载均衡)
- **DPM**(动态节能管理)

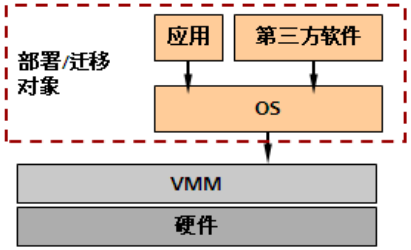
弹性伸缩



虚拟机（应用）之间可以动态共享资源，实现**物理机内的错峰削谷**

- **CPU**的热插拔
- 内存动态伸缩
- 磁盘空间动态伸缩

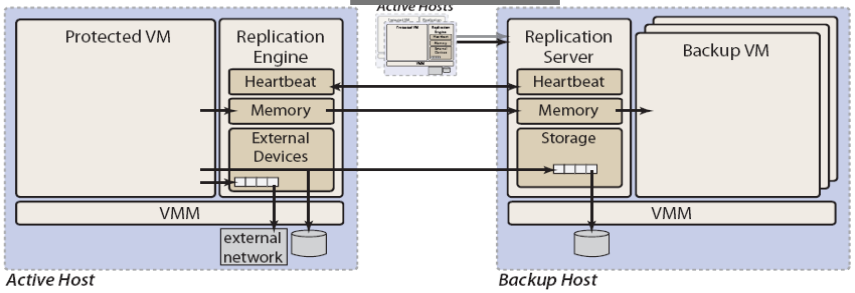
应用容器



虚拟化技术改变原有的应用打包/发布模式，实现应用对硬件更彻底的解耦

- **OS**专用，避免其他软件干扰应
- 安装/迁移效率大幅提升

虚拟机热备



- **VM**状态实时在主备**VM**间同步，应用无关的通用热备方案
- 物理机故障时，业务恢复时间在百**ms**内

虚拟机安全

- 提供更安全的入侵检测及防病毒方案

虚拟机快照

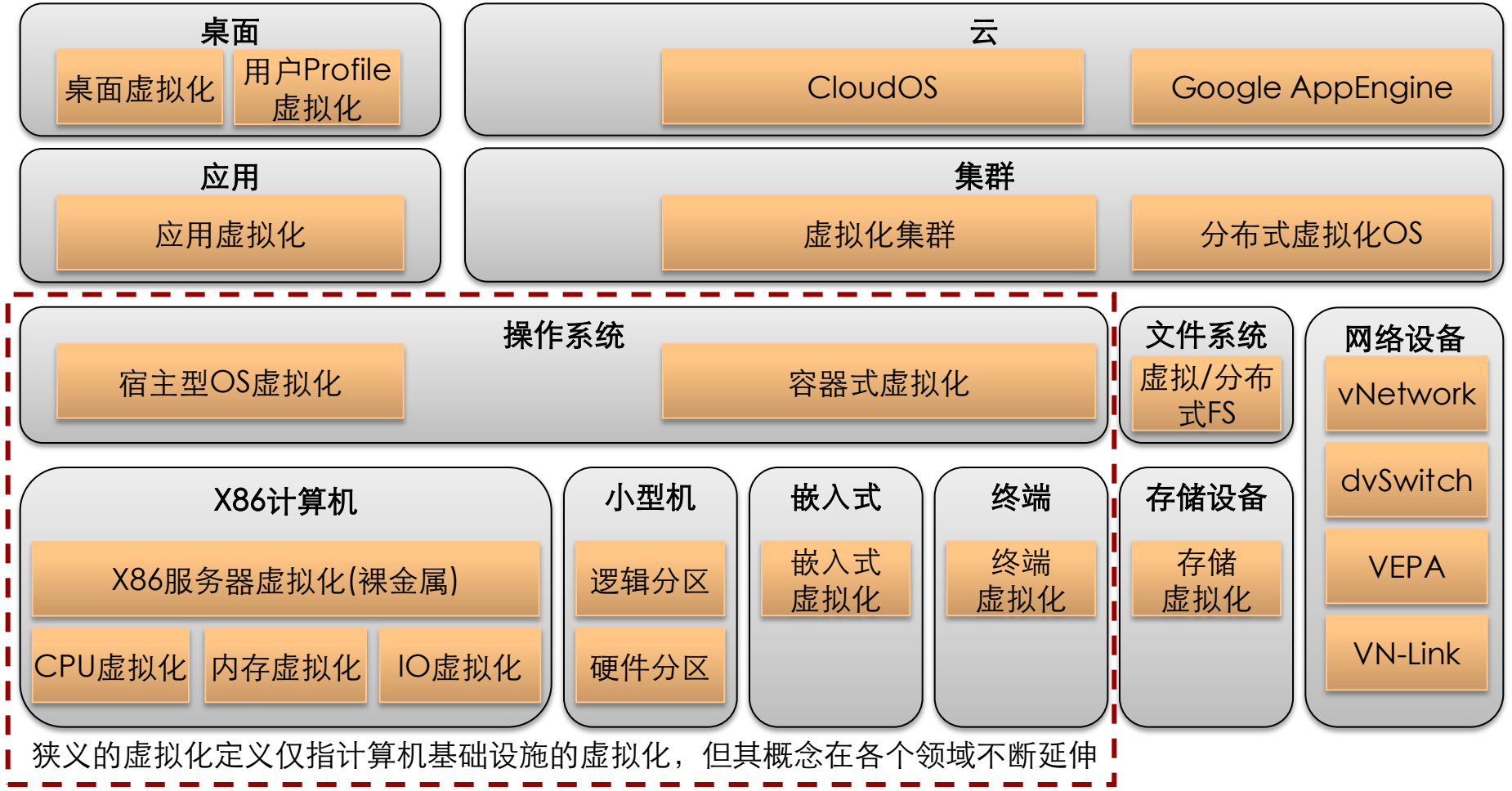
- 对系统的内存及存储进行快照保存
- 方便故障现场的重现

虚拟机录像

- 对虚拟机的事件进行记录
- 历史运行过程回放方便故障定位、黑客行为跟踪

2007-2010：虚拟化2.0。主要特点： 1) 灵活的资源管理：应用快速部署，OS绿色化；实现应用可移动性，虚拟机资源动态伸缩； 2) 透明的带外控制能力

虚拟化3.0：虚拟化手段不断进行纵横延伸



2010+： 虚拟化3.0。所有虚拟化的共同特征：抽象，隔离，封装，解耦
虚拟化3.0创新点在于在更细的层面及更广的范围内进行虚拟化抽象

云计算环境下的资源分层模型

应用层(SaaS)

服务对象：业务
目标：支持多租户的按需使用并计费的业务池
实例：Google Apps，Windows Office 365，saleforce

平台层(PaaS)

服务对象：应用平台/中间件
目标：提供云化的应用孵化、开发及运行平台
实例：Google AppEngine，Vmware Cloud Foundry

OS层(IaaS)

服务对象：VM/OS
目标：提供安全、独立、按需应变的计算单元的资源池，即虚拟机资源池化
实例：Amazon EC2

VM VM VM VM 。。。。 VM VM VM VM

硬件层

服务对象：多个物理服务器的CPU/内存/IO设备
目标：OS跨多个物理服务器，以提供更多的计算能力
市场：替代低端的小型机

服务器/存储/网络资源聚合

服务对象：多个物理服务器及存储
目标：提供服务器集群管理能力，服务器及存储资源池化

虚拟化是云计算中承上启下的核心技术，没有虚拟化技术IaaS层的聚合将无从谈起

云计算环境下软件分层模型演进

传统桌面应用
软件分层模型
（微软）

OS
人机界面
应用
中间件/DB
应用环境
资源管理
硬件

Windows桌面
受到挑战；IE的
成功缓解了压力

应用部署与用户
界面的无关性使
Linux获得机会

互联网应用
软件分层模型
（Google）

云终端
Web浏览器

云/OS
应用
中间件/DB
应用环境
资源管理
硬件

应用虚拟化可以
在非Windows
操作系统获得
Windows应用
的体验

硬件管理被剥离
到虚拟化层，OS
只保留应用环境
的功能

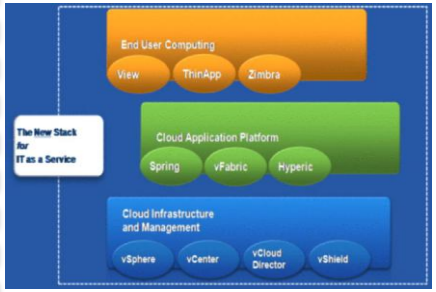
云计算应用
软件分层模型
（VMware）

云终端
桌面/应用虚拟化

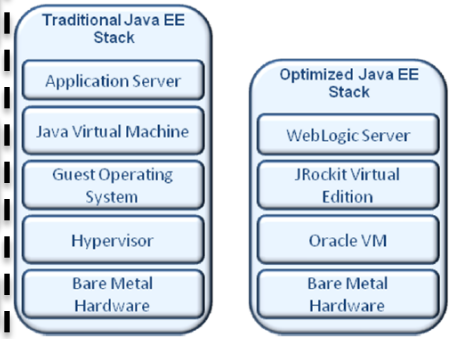
应用容器
应用
中间件/DB
应用环境

IaaS/VMM
资源管理
硬件

VMware提出的ITaaS
软件分层模型



Oracle发布不依赖OS
的JVM，性能提升30%



以虚拟化能力为核心的分布式操作系统将成为云计算基础设施新的OS形态

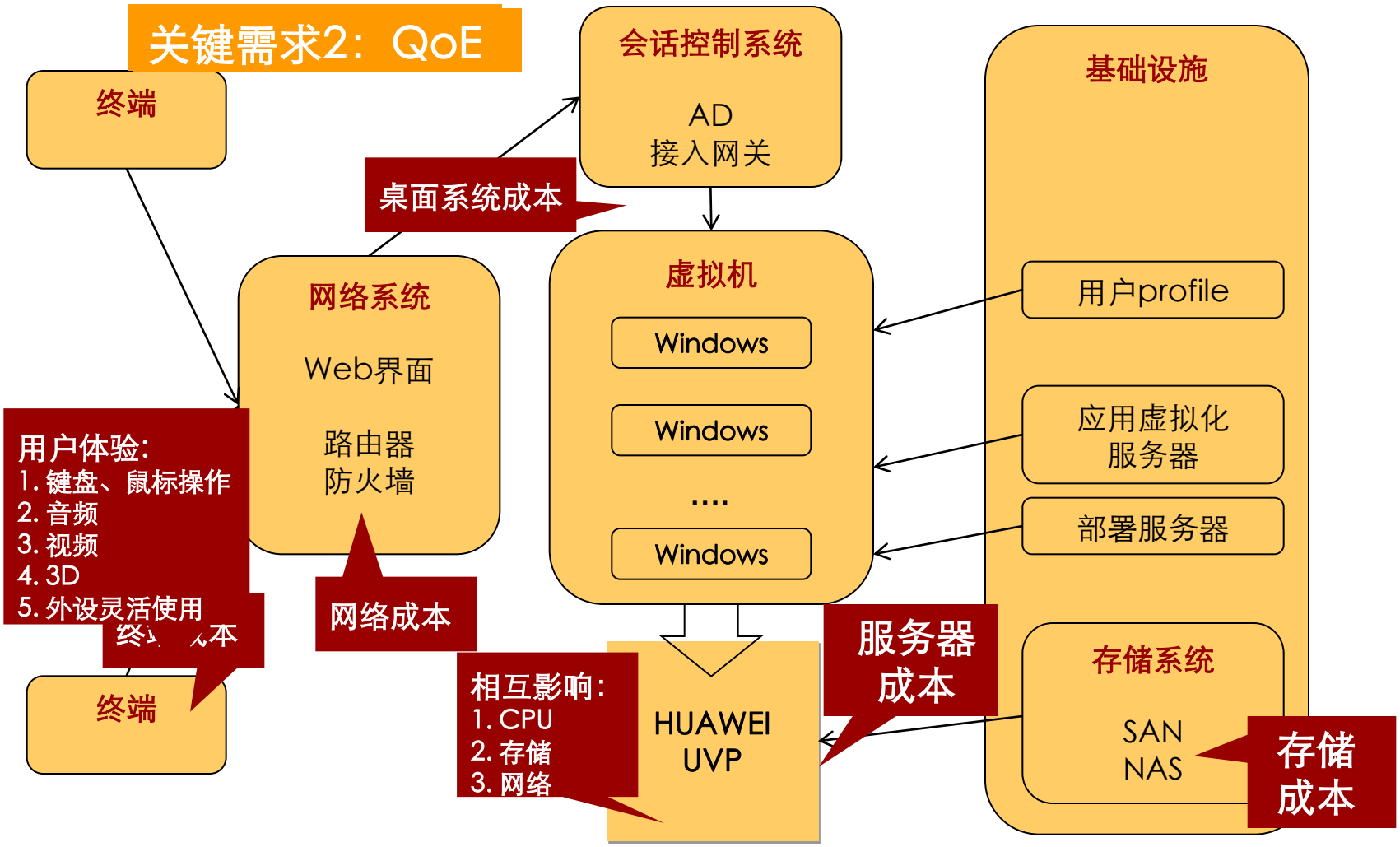
小结

- **虚拟化是对计算机物理资源的进行抽象的手段**
- **虚拟化1.0:**
  - 虚拟化基本特征：抽象、隔离、解耦
  - 应用：服务器整合--提高服务器的利用率，降低IT的TCO
- **虚拟化2.0:**
  - **灵活资源管理：**应用可移动/动态伸缩，资源最优化调配，应用快速部署
  - **透明带外控制能力：**热备，安全，可靠性，可维护性
- **虚拟化3.0从传统的服务器范围内向云计算各个领域渗透**
  - 存储/网络/桌面/应用虚拟化/…
- **虚拟化技术是云计算中承上启下的核心技术，在云计算时代将颠覆传统软件分层模型**

目录

- 虚拟化背景、热点与趋势
- UVP在云计算中的应用及关键技术

虚拟化应用一：桌面云


关键需求2：QoE
终端
会话控制系统
AD
接入网关
桌面系统成本
基础设施
网络系统
Web界面
路由器
防火墙
虚拟机
Windows
Windows
....
Windows
用户profile
应用虚拟化
服务器
部署服务器
用户体验:
1. 键盘、鼠标操作
2. 音频
3. 视频
4. 3D
5. 外设灵活使用
网络成本
终端
相互影响:
1. CPU
2. 存储
3. 网络
HUAWEI
UVP
服务器
成本
存储系统
SAN
NAS
存储
成本

虚拟化应用二：虚拟数据中心

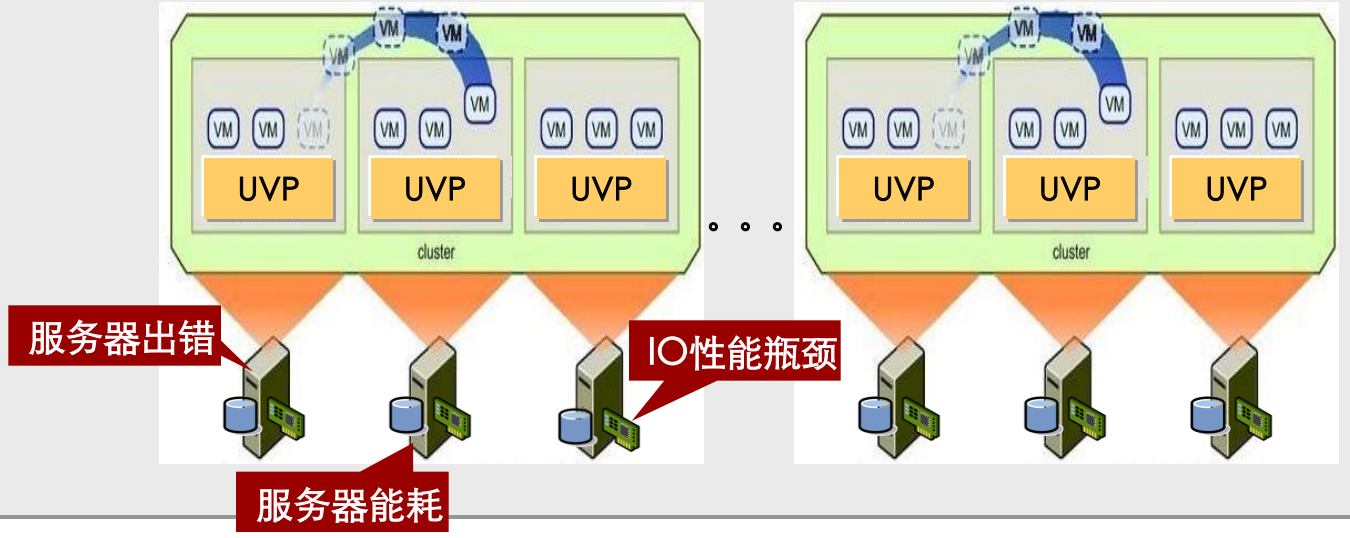
DC运营层

用户自助Portal	用户Porfile/计费/日志	用户Profile/鉴权	OMS/GMS/机房/散热管理
集群汇聚/应用Provisioning	分布式WEB应用基础设施	分布式大规模计算基础设施	桌面虚拟化基础设施

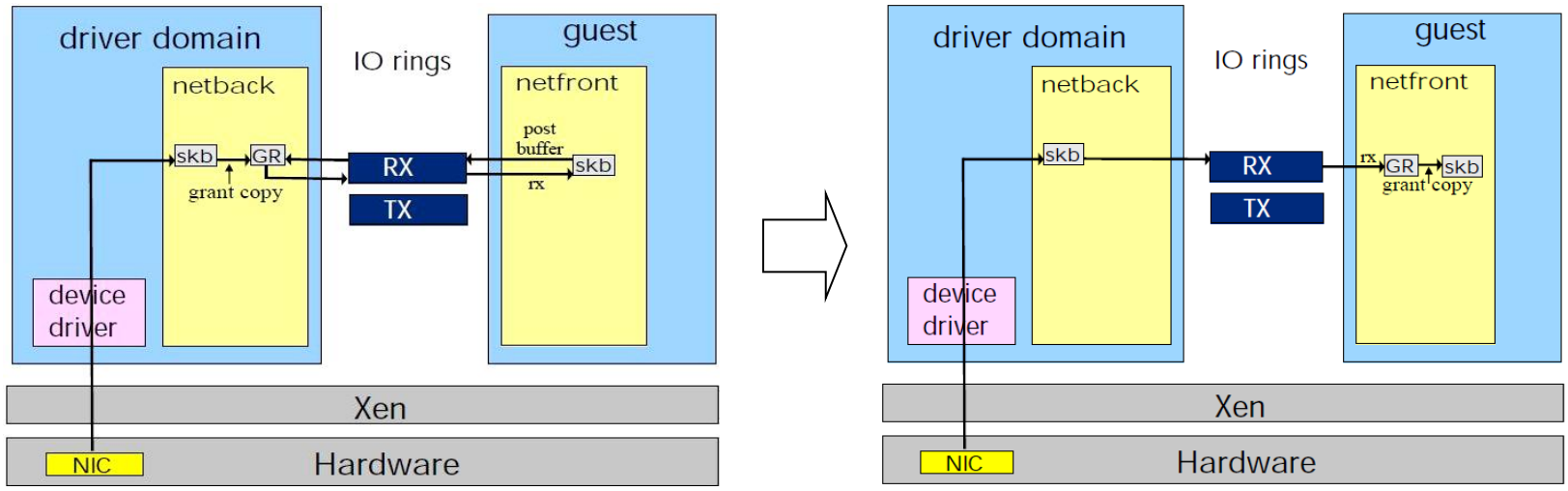
集群管理层

关键需求：高性能、高可靠、节能

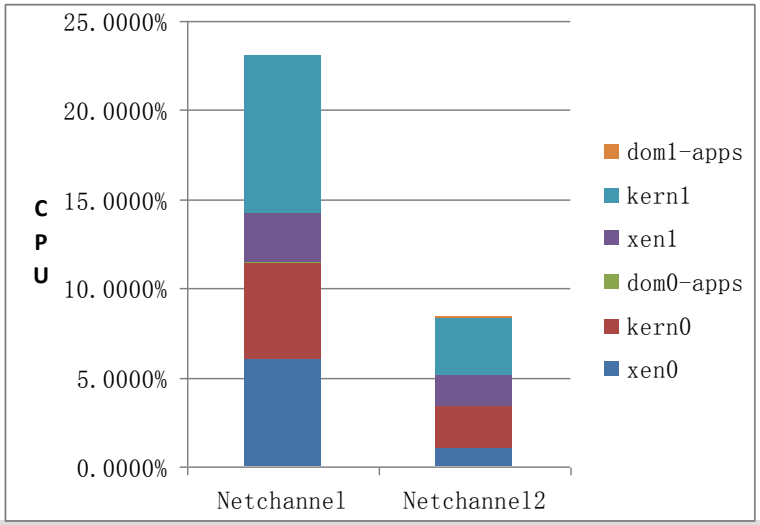
UVP平台



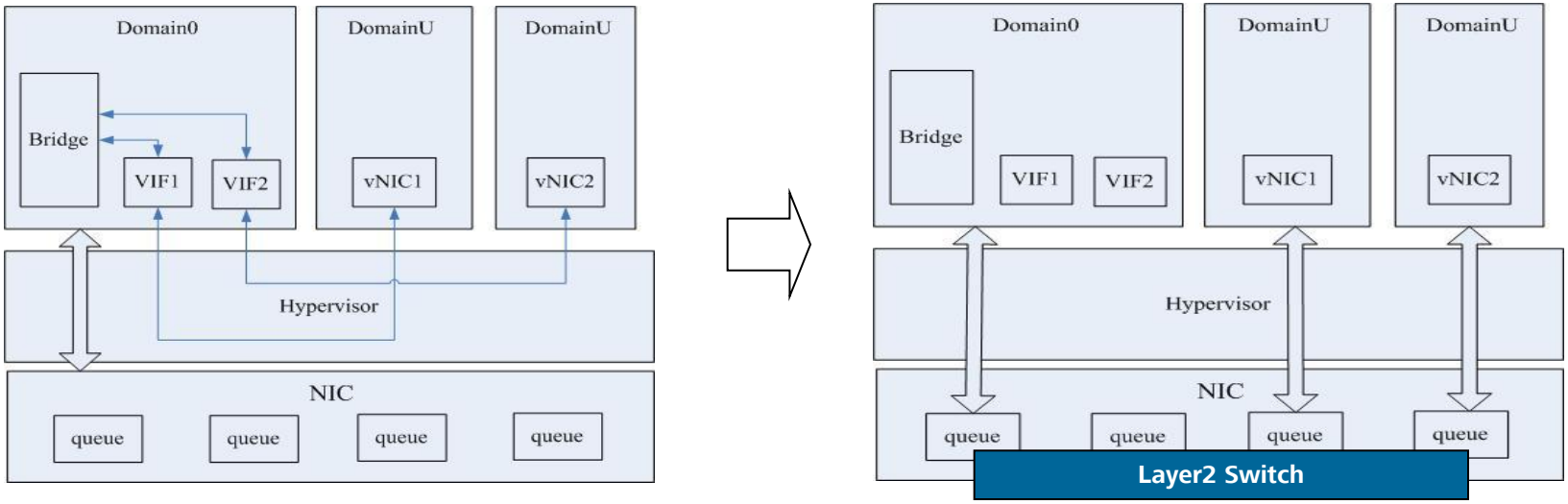
PV模式增强—netchannel2



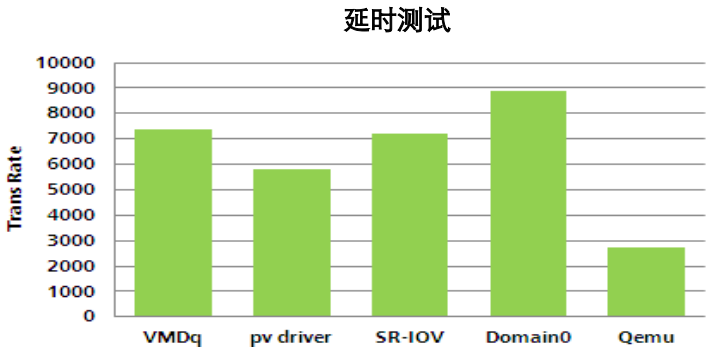
优化后效果

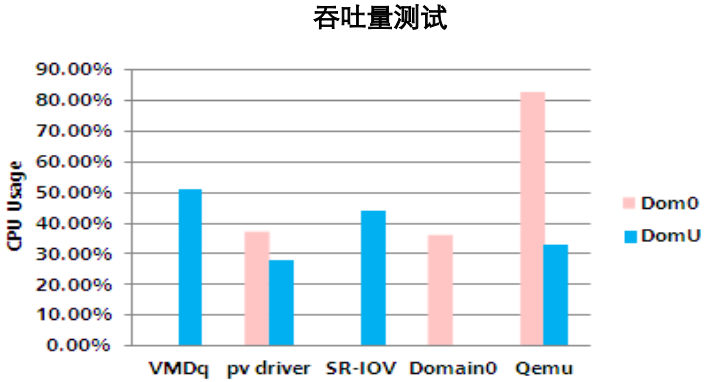


基于VMDq/iNIC的网卡队列直通

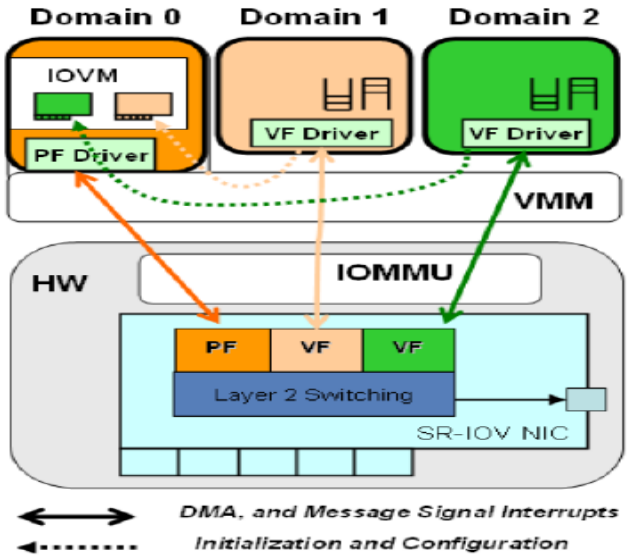


优化后效果



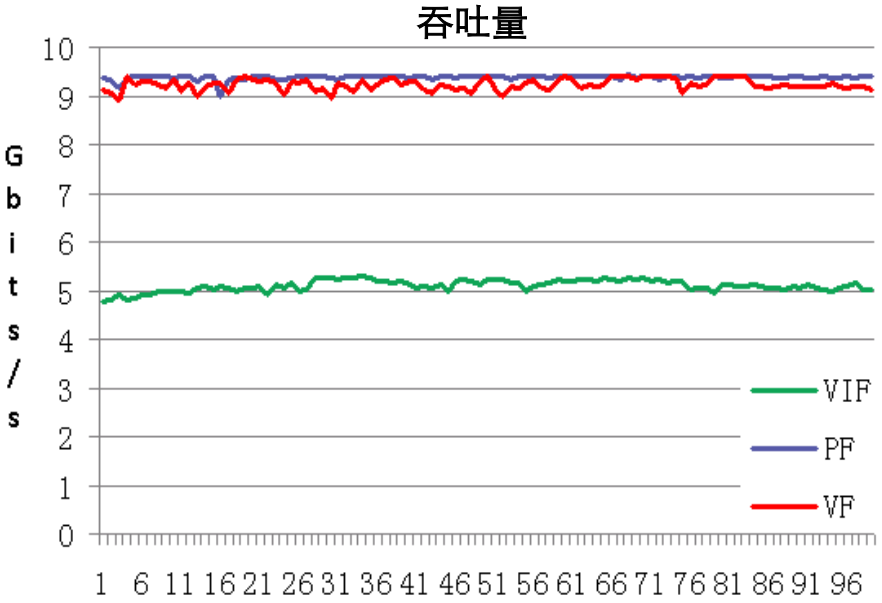


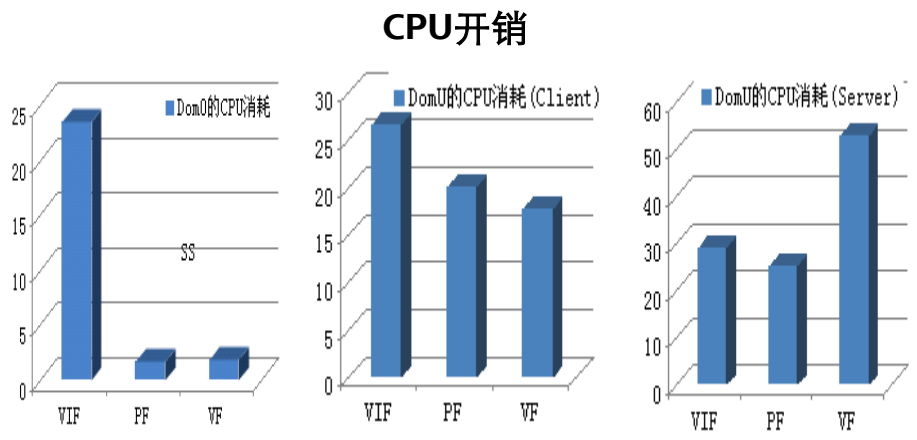
VT-d/SR-IOV技术



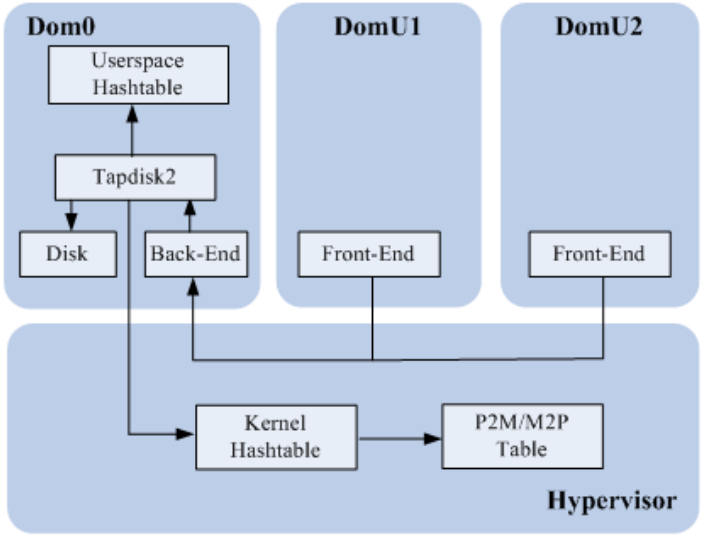
- VT-d：设备直接分配给VM，DMA直接到VM，不通过VMM/dom0，性能与native持平
- SR-IOV：解决了可分配设备数量问题

性能测试结果





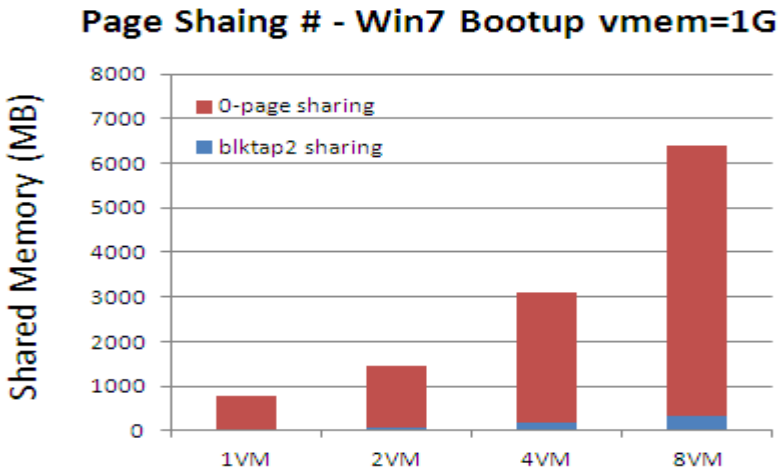
内存复用技术--内存共享

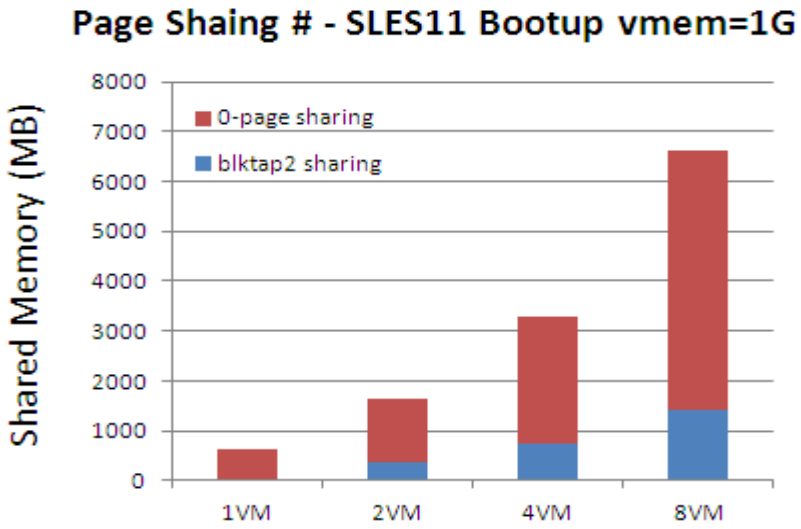


桌面云场景中，用户VM环境单一，存在大量重复内存可共享

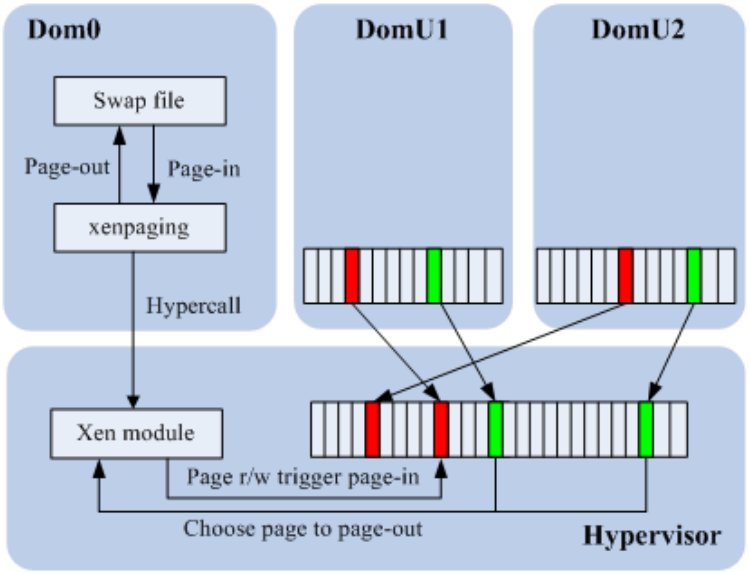
VM中存在大量零页空闲内存可共享

性能测试结果



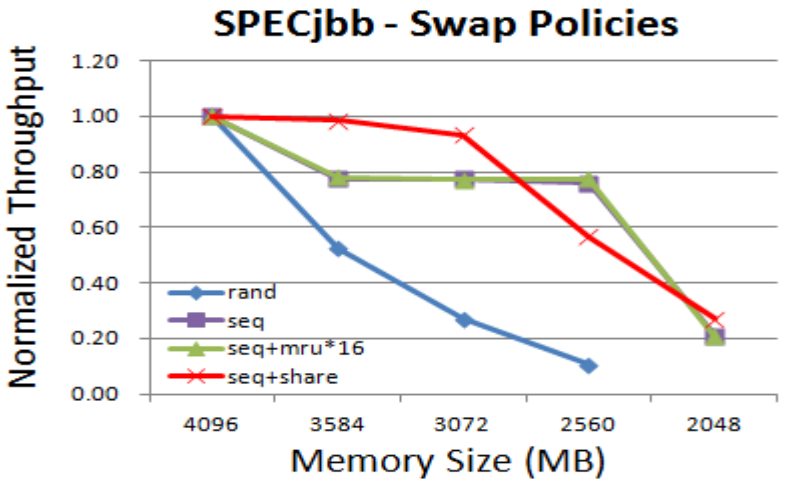


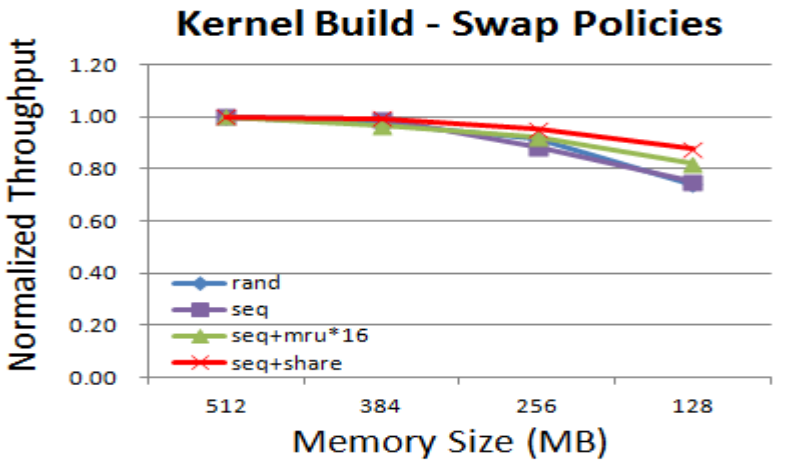
内存复用技术--内存交换



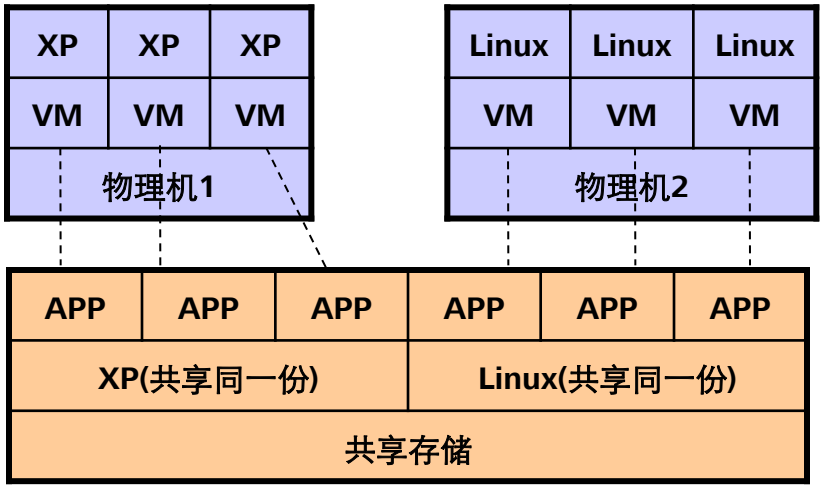
NC场景中内存大小成为限制VM密度的一个重要瓶颈，当内存压力高时通过内存交换应急措施。

性能测试结果

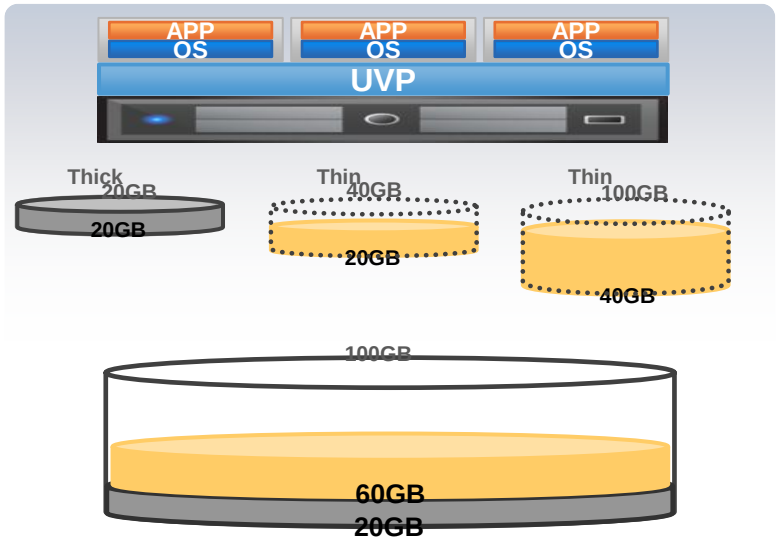




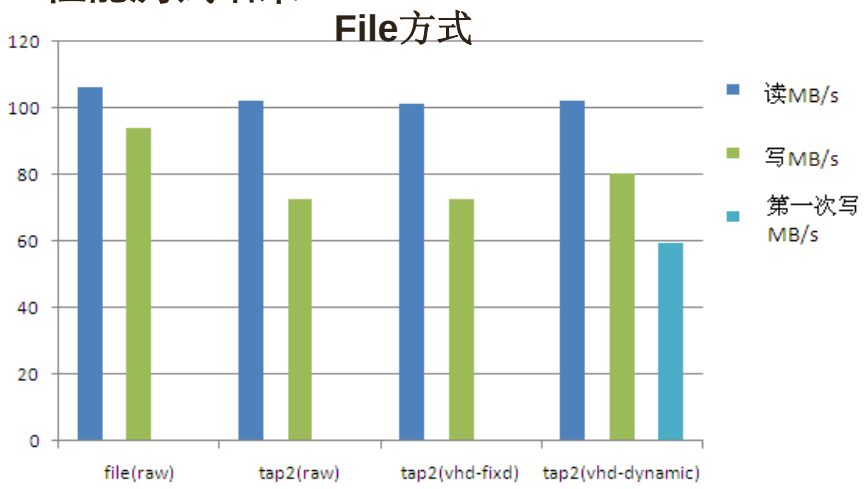
链接克隆与精简部署



VM之间存在大量的重复数据，通过链接克隆提升VM虚拟磁盘利用率

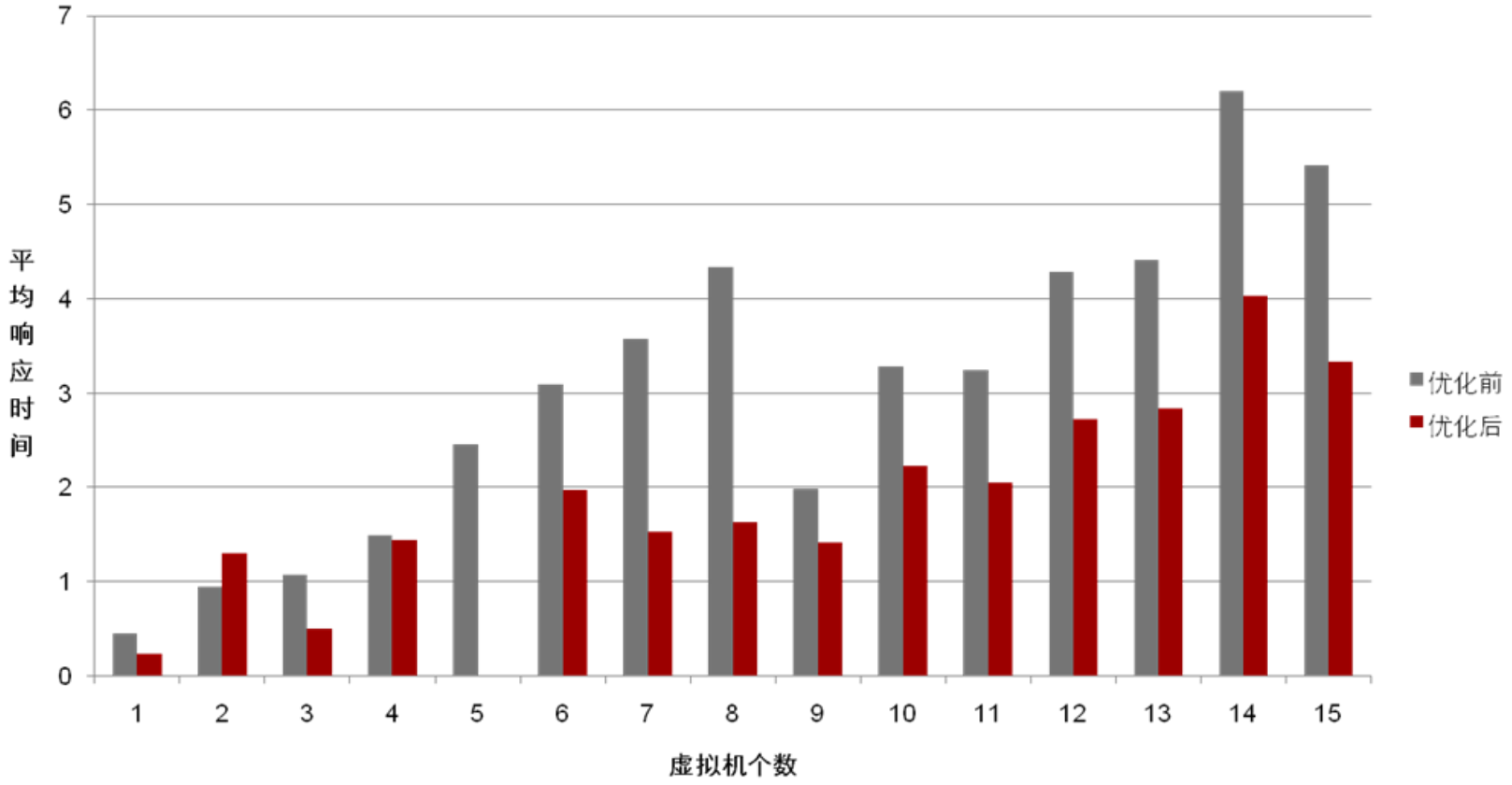


性能测试结果



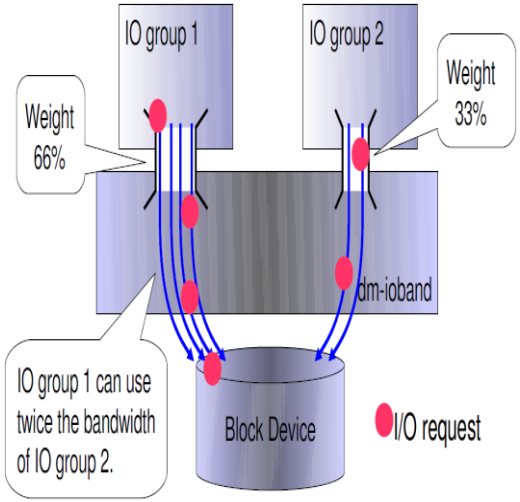
CPU调度优化（二）

桌面云应用中，用户键盘、鼠标优化，减少延时，改善用户体验

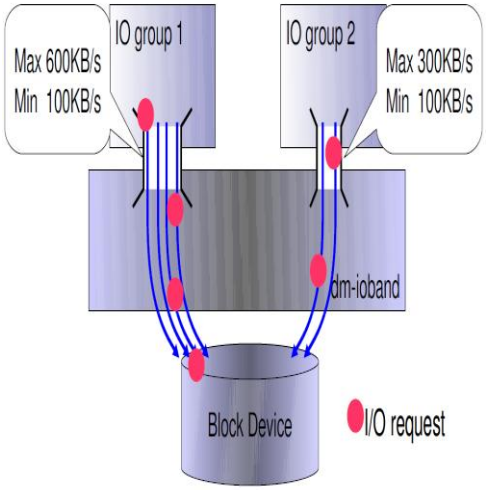


存储QoS

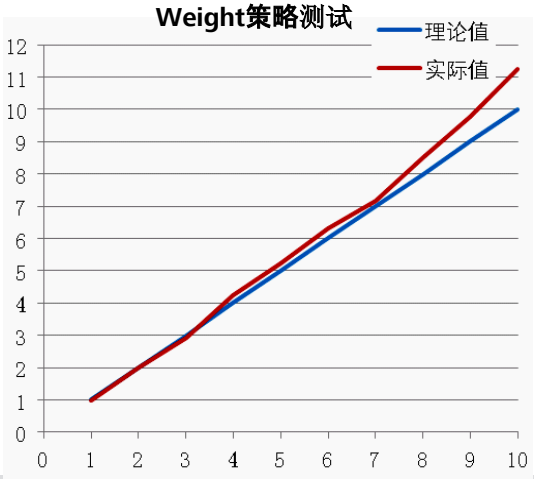
基于Weight的QoS控制

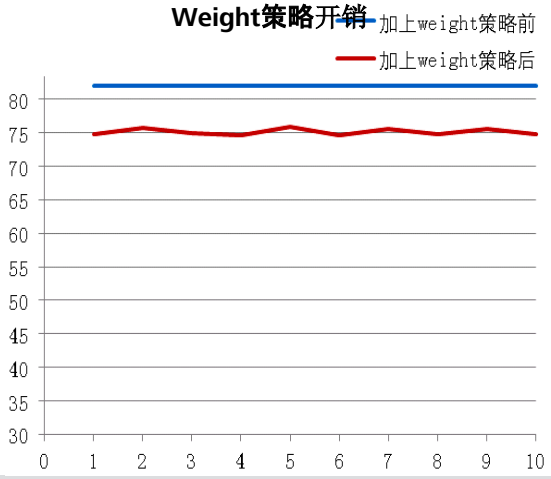


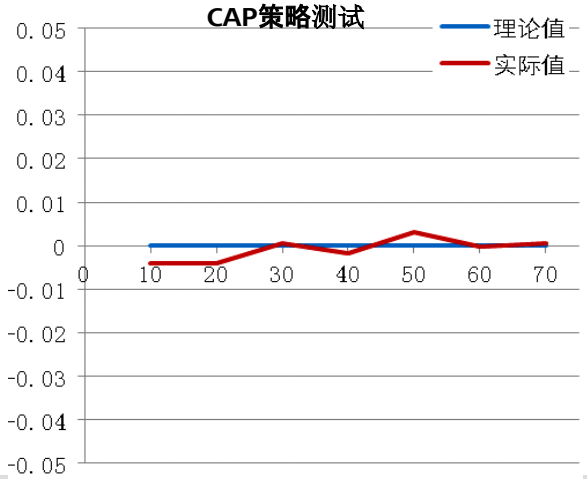
基于绝对带宽的QoS控制



- 保证高优先的VM得到更高的磁盘带宽
- VM进行高吞吐量的磁盘操作时不干扰其他VM的磁盘响应



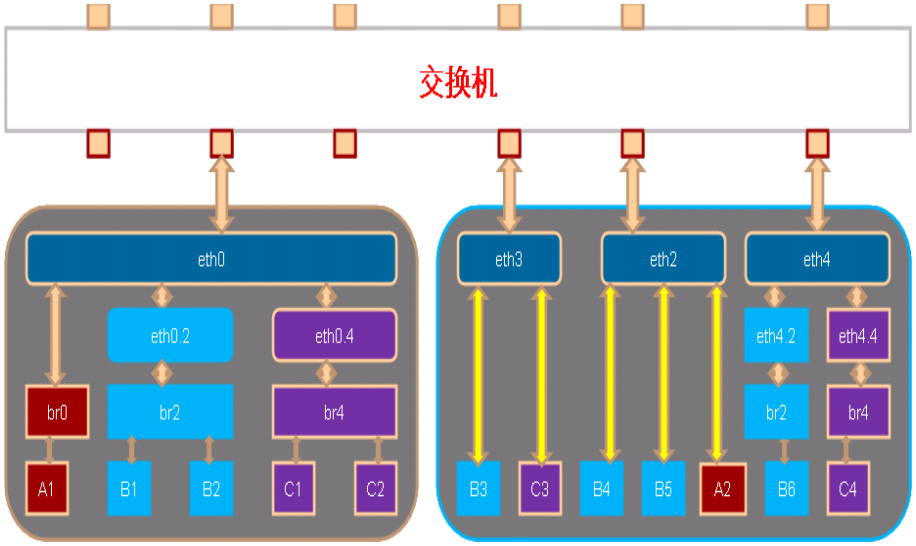




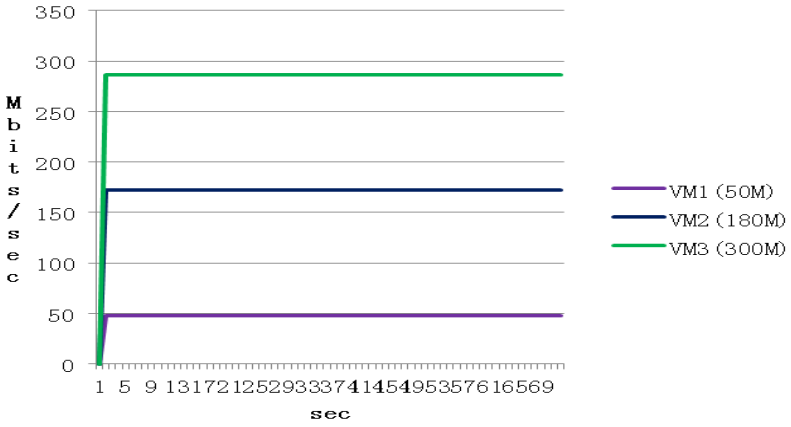
网络隔离、QoS

保证高优先的VM得到更多的网络吞吐量
隔离不同租户的VM之间的网络数据包

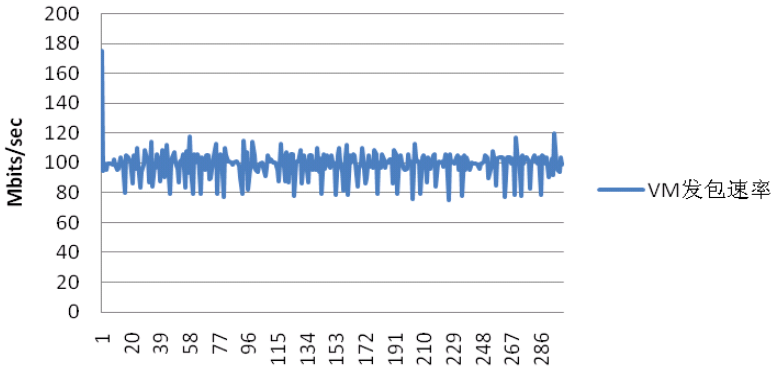
通过Vlan隔离网络数据包



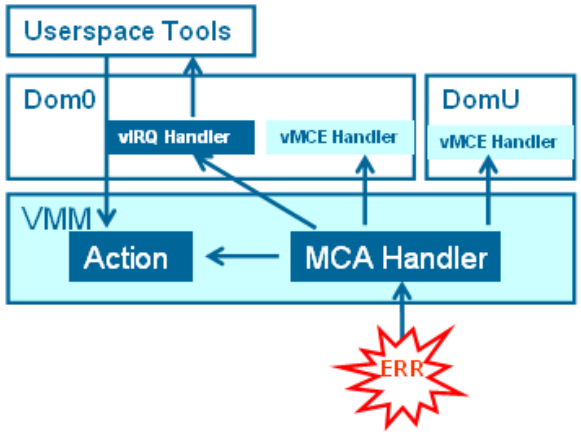
基于SR-IOV的流控

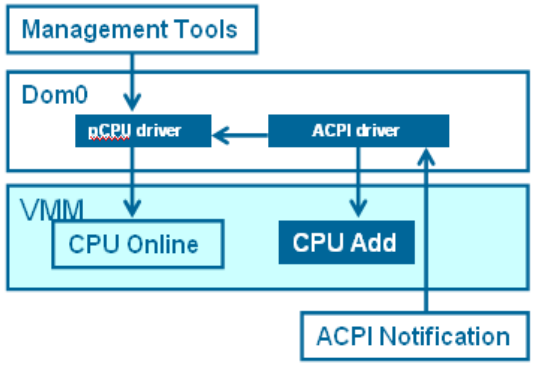


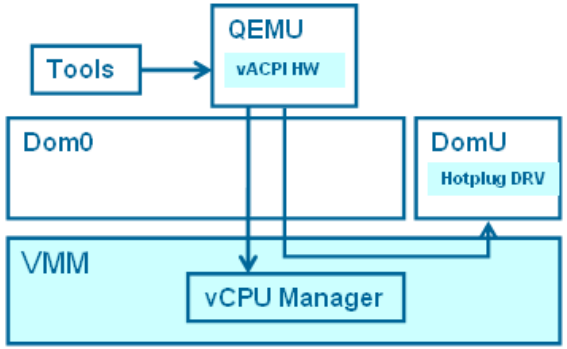
基于软件方案的流控



硬件RAS







MCA

一台服务器上聚合虚拟机的能力也在增加，服务器部件的故障对应用的影响加大，**MCA**限制故障范围

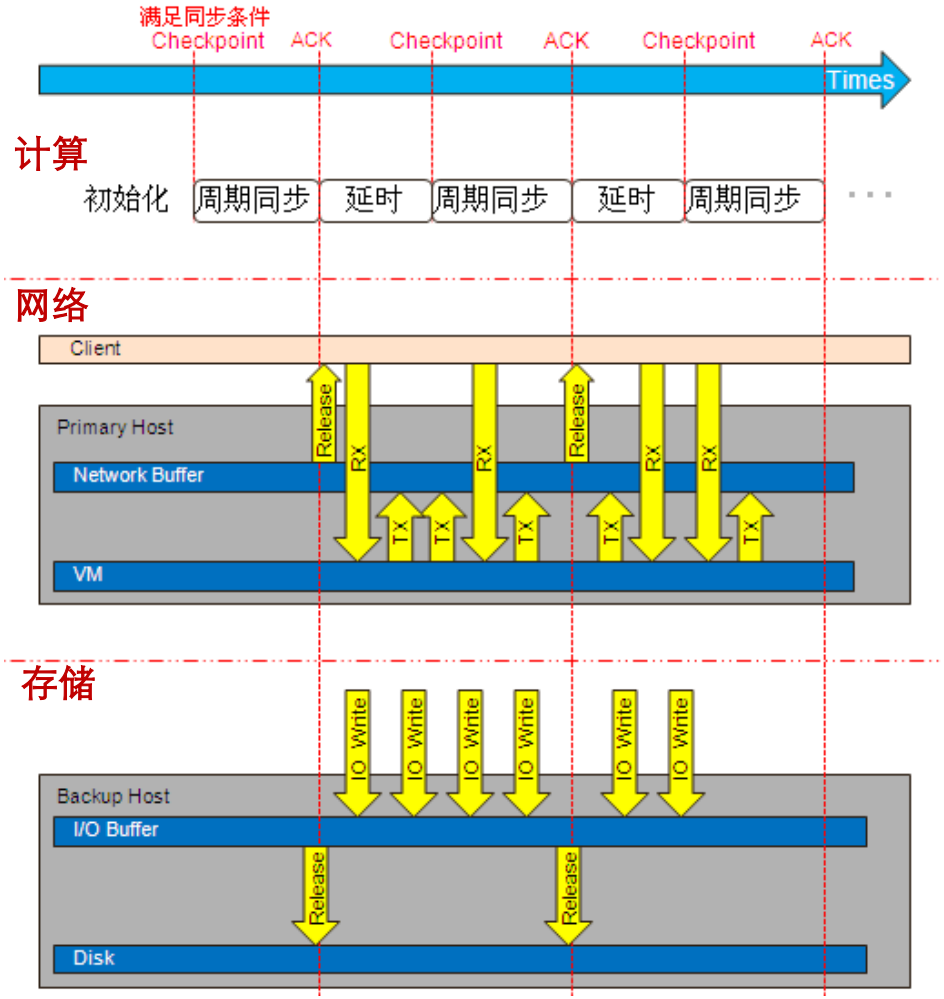
宿主机CPU/内存热插拔

- 宿主机CPU/内存硬件出错时，通过hotplug 替换出错部件的需求
- 通过hotplug动态调整宿主机CPU/内存能力

客户机CPU热插拔

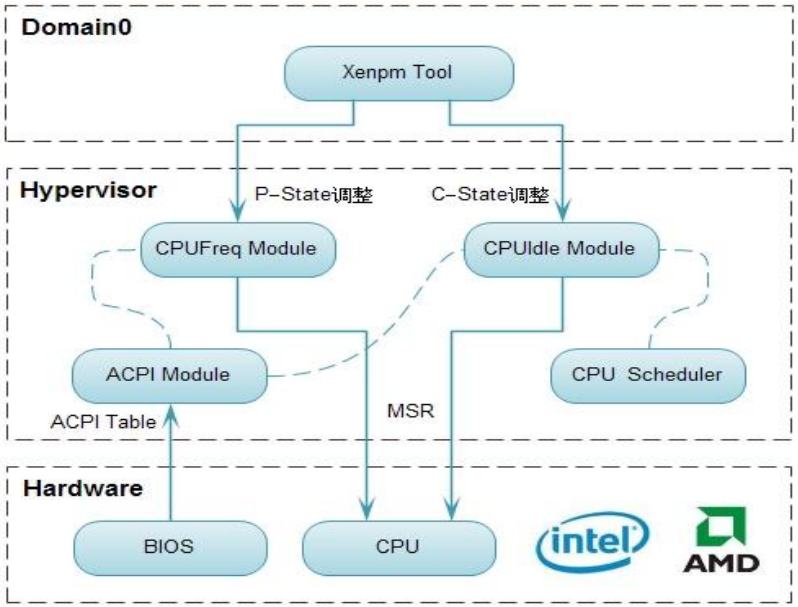
- 通过vHotplug动态调整客户机CPU能力

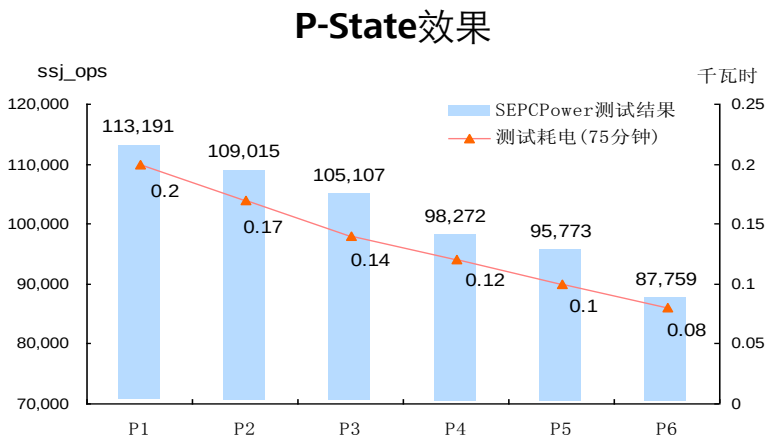
虚拟机热备

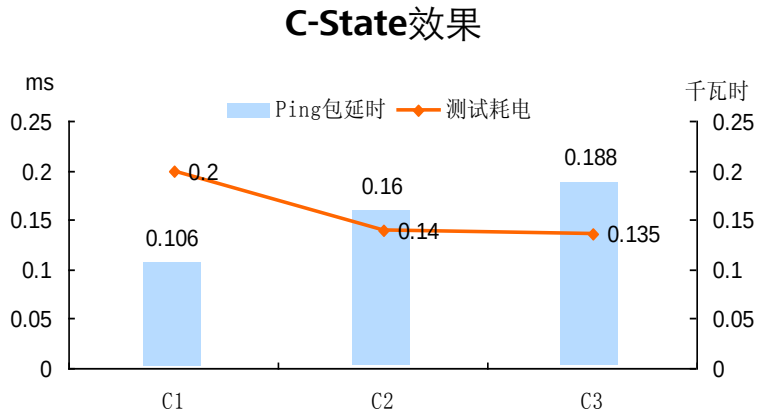


单台服务器宕机后，其上运行的VM提供的服务中断，传统HA方案仍然会造成服务短期中断。虚拟机热备提供了**百ms内主备倒换**，TCP连接不会中断，用户几乎感知不到异常的发生。

处理器节能







物理机在低负载情况下，通过P-state调节CPU频率；当CPU空闲时，让CPU进入C-State，实现服务器节能。

Thank you

www.huawei.com