Xen 和 KVM 等四大虚拟化架构对比分析

前言

云计算如今已是一个相当热门的概念，各行各业包括政府，云建设都如火如荼地进行。华为正借助开源技术，向不同领域的客户提供多样化的云服务,包括提供全面的私有云、公有云和混合云。云简而言之就是把 IT 资源服务化。过去办公场景中我们每人一台 PC，拥有独立的 IT 资源，而云可以将 IT 资源按需分配给需要的租户，实现按需、弹性拓展。以前每个人一台 PC。现在大家共享一台超级 PC，按需访问，不用时资源自动释放，可供其他用户使用，这样资源得以最大化利用，并且可以按需扩展，及时满足使用需求。

个人 PC 机的操作系统向下管理和驱动底层硬件，如 CPU、内存、硬盘等，管理计算任务，调配资源， 向上为各类应用软件提供统一、标准的接口。

云操作系统功能有些类似，但功能更复杂。云操作系统它负责管理和调配一个或多个数据中心的硬件资源，这些资源可能由数量巨大的服务器、存储设备组成，并逻辑上把它们整合一台虚拟计算机系统，供上层云应用使用。为了提升硬件资源使用效率，一台硬件设备会首先被虚拟成多个具备独立功能的虚拟设备，以便同时供多个应用调用，这就涉及到虚拟化技术，这也是云操作系统的关键技术之一。而当前具备这种虚拟化功能的技术维基百科列举的就有超过 60 种，其中有四种虚拟化技术是当前最为成熟而且运用最广泛的，分别是：VMware 的 ESX、微软 Hyper-V、开源的 Xen 和 KVM 等，下面将针对这 4 种虚拟化技术架构进行分析。

一、 虚拟化架构分析

1、 虚拟化架构

从虚拟化的实现方式来看，虚拟化架构主要有三种形式：寄居虚拟化架构、裸金属虚拟化架构和操作系统虚拟化架构，其性能及主流产品如下

以虚拟化架构维度，分类如下

【寄居虚拟化架构】Hypervisor 运行在基础操作系统上，构建出一整套虚拟硬件平台，支持创建各种操作系统类型虚拟机。代表性产品：VMware WorkStation、Redhat KVM。优点：简单、易实现。缺点：上层 Guest OS 的处理需要逐层转换，发送到底层进行处理，依赖于 Host OS。

【裸金属虚拟化架构】Hypervisor 直接运行在硬件上，直接与硬件交互提升效率。代表性产品：VMware ESXServer、Citrix XenServer、Microsoft Hyper-V。优点：交互效率提升，不依赖操作系统。

【操作系统虚拟化架构】隔离性差，最后一种不常用，虚拟机运行在传统操作系统上，创建一个独立的虚拟化实例（容器 Container），指向底层托管操作系统，缺点是操作系统唯一，如果底层操作系统跑的是 Windows，那么 VPS/VE 就都得跑 Windows，

在宿主架构中的虚拟机作为主机操作系统的一个进程来调度和管理，裸金属架构下 则不存在主机操作系统，它是以 Hypervisor 直接运行在物理硬件之上，即使是有类似主机操作系统的父分区或 Domain 0，也是作为裸金属架构下的虚拟机存在的。宿主架构通常用于个人 PC 上的虚拟化，如 WindowsVirtual PC，VMware Workstation，Virtual Box，Qemu 等，而裸金属架构通常用于服务器的虚拟化。

2、 虚拟化技术

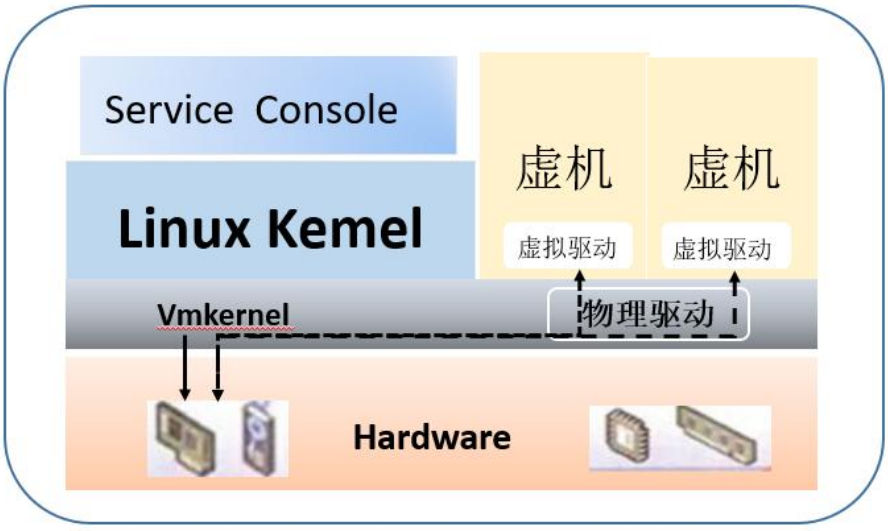
以虚拟化技术维度，分类如下

【全虚拟化】也称为原始虚拟化技术，运行在虚拟机上的操作系统通过 Hypervisor 来最终分享硬件，所以虚拟机发出的指令需经过 Hypervisor 捕获并处理。

【半虚拟化】半虚拟化技术是后来才出现的技术，它就是在全虚拟化的基础上，把客户操作系统进行了修改，增加了一个专门的 API，这个 API 可以将客户操作系统发出的指令进行最优化，即不需要 Hypervisor 耗费一定的资源进行翻译操作，因此 Hypervisor 的工作负担变得非常的小，因此整体的性能也有很大的提高。

【硬盘辅助虚拟化】Hypervisor 可以在部分功能上与硬件直接交互，提升性能。比如在 CPU 性能较差的网络 IO 方面与硬件直接交互。

二、 ESX 虚拟化架构



ESX 虚拟化架构示意图

ESX 是 VMware 的企业级虚拟化产品，ESX 服务器启动时，首先启动 Linux Kernel，通过

这个操作系统加载虚拟化组件，最重要的是 ESX 的 Hypervisor 组件，称之为 VMkernel，VMkernel 会从 LinuxKernel 完全接管对硬件的控制权，而该 Linux Kernel 作为 VMkernel 的首个虚拟机，用于承载 ESX 的 serviceConsole，实现本地的一些管理功能。

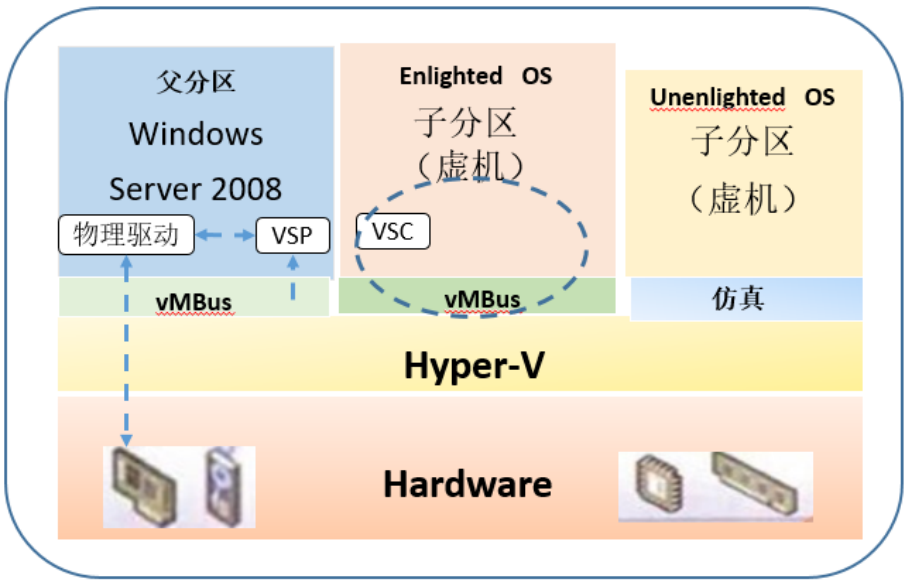
VMkernel 负责为所承载的虚拟机调度所有的硬件资源，但不同类型的硬件会有些区别。

虚拟机对于 CPU 和内存资源是通过 VMkernel 直接访问，最大程度地减少了开销，CPU 的直接访问得益于 CPU 硬件辅助虚拟化(Intel VT-x 和 AMD AMD-V，第一代虚拟化技术)，内存的直接访问得益于 MMU(内存管理单元)硬件辅助虚拟化。

虚拟机对于 I/O 设备的访问则有多种方式，以网卡为例，有两种方式可供选择：一是利用 I/O MMU 硬件辅助虚拟化的 VMDirectPath I/O，使得虚拟机可以直接访问硬件设备，从而减少对 CPU 的开销；二是利用半虚拟化的设备 VMXNETx，网卡的物理驱动在 VMkernel 中，在虚拟机中装载网卡的虚拟驱动，通过这二者的配对来访问网卡，与仿真式网卡相比有着较高的效率。半虚拟化设备的安装是由虚拟机中 VMware tool 来实现的，可以在 Windows 虚拟机的右下角找到它。网卡的这两种方式，前者有着显著的先进性，但后者用得更为普遍，因为 VMDirectPath I/O 与 VMware 虚拟化的一些核心功能不兼容，如：热迁移、快照、容错、内存过量使用等。

ESX 的物理驱动是内置在 Hypervisor 中，所有设备驱动均是由 VMware 预植入的。因此，ESX 对硬件有严格的兼容性列表，不在列表中的硬件，ESX 将拒绝在其上面安装。

三、 Hyper-V 虚拟化架构



Hyper-V 虚拟化架构示意图

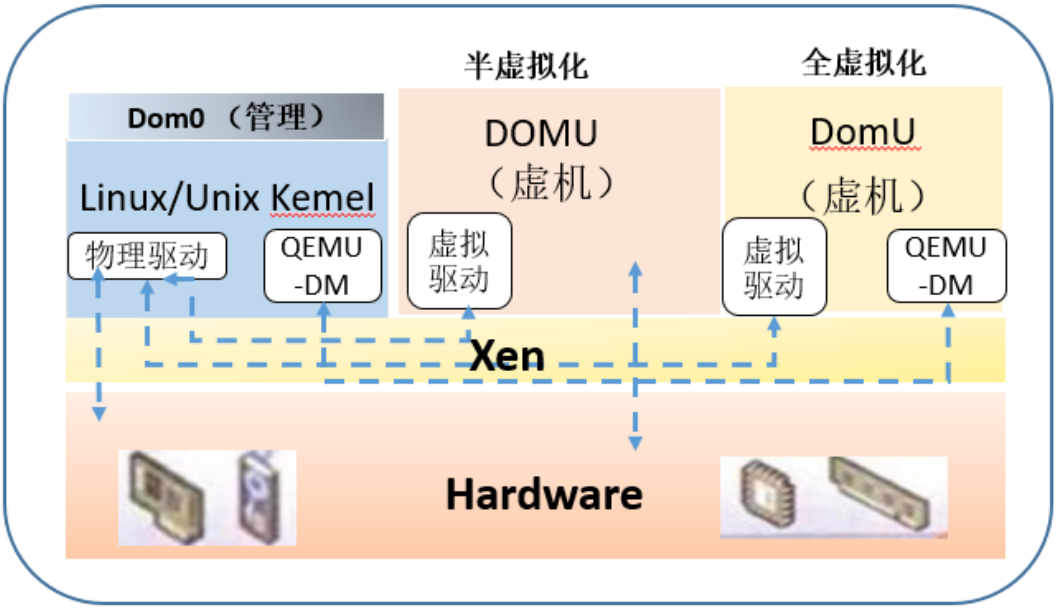
Hyper-V 是微软新一代的服务器虚拟化技术，首个版本于 2008 年 7 月发布，Hyper-V 有两种发布版本：一是独立版，如 Hyper-V Server 2008，以命令行界面实现操作控制，是一个免费的版本；二是内嵌版，如 Windows Server 2008，Hyper-V 作为一个可选开启的角色。

对于一台没有开启 Hyper-V 角色的 Windows Server 2008 来说，这个操作系统将直接操作硬件设备，一旦在其中开启了 Hyper-V 角色，系统会要求重新启动服务器。在这次重启动过程中，Hyper-V 的 Hypervisor 接管了硬件设备的控制权，先前的 Windows Server 2008 则成为 Hyper-V 的首个虚拟机，称之为父分区，负责其他虚拟机(称为子分区)以及 I/O 设备的管理。Hyper-V 要求 CPU 必须具备硬件辅助虚拟化，但对 MMU 硬件辅助虚拟化则是一个增强选项。

其实 Hypervisor 仅实现了 CPU 的调度和内存的分配，而父分区控制着 I/O 设备，它通过物理驱动直接访问网卡、存储等。子分区要访问 I/O 设备需要通过子分区操作系统内的 VSC(虚拟化服务客户端)，对 VSC 的请求由 VMBUS（虚拟机总线）传递到父分区操作系统内的 VSP(虚拟化服务提供者)，再由 VSP 重定向到父分区内的物理驱动，每种 I/O 设备均有各自的 VSC 和 VSP 配对，如存储、网络、视频和输入设备等，整个 I/O 设备访问过程对于子分区的操作系统是透明的。其实在子分区操作系统内，VSC 和 VMBUS 就是作为 I/O 设备的虚拟驱动，它是子分区操作系统首次启动时由 Hyper-V 提供的集成服务包安装，这也算是一种半虚拟化的设备，使得虚拟机与物理 I/O 设备无关。如果子分区的操作系统没有安装 Hyper-V 集成服务包或者不支持 Hyper-V 集成服务包(对于这种操作系统，微软称之为 Unenlightened OS，如未经认证支持的 Linux 版本和旧的 Windows 版本)，则这个子分区只能运行在仿真状态。其实微软所宣称的启蒙式 (Enlightenment)操作系统，就是支持半虚拟化驱动的操作系统。

Hyper-V 的 Hypervisor 是一个非常精简的软件层，不包含任何物理驱动，物理服务器的设备驱动均是驻留在父分区的 Windows Server 2008 中，驱动程序的安装和加载方式与传统 Windows 系统没有任何区别。因此，只要是 Windows 支持的硬件，也都能被 Hyper-V 所兼容。

四、 Xen 虚拟化架构



XEN 的虚拟化架构示意图

XEN 最初是剑桥大学 Xensource 的一个开源研究项目，2003 年 9 月发布了首个版本 XEN 1.0，2007 年 Xensource 被 Citrix 公司收购，开源 XEN 转由 www.xen.org 继续推进，该组织成员包括个人和公司(如 Citrix、Oracle 等)。该组织在 2011 年 3 月发布了版本 XEN 4.1。

相对于 ESX 和 Hyper-V 来说，XEN 支持更广泛的 CPU 架构，前两者只支持 CISC 的 X86/X86_64 CPU 架构，XEN 除此之外还支持 RISC CPU 架构，如 IA64、ARM 等。

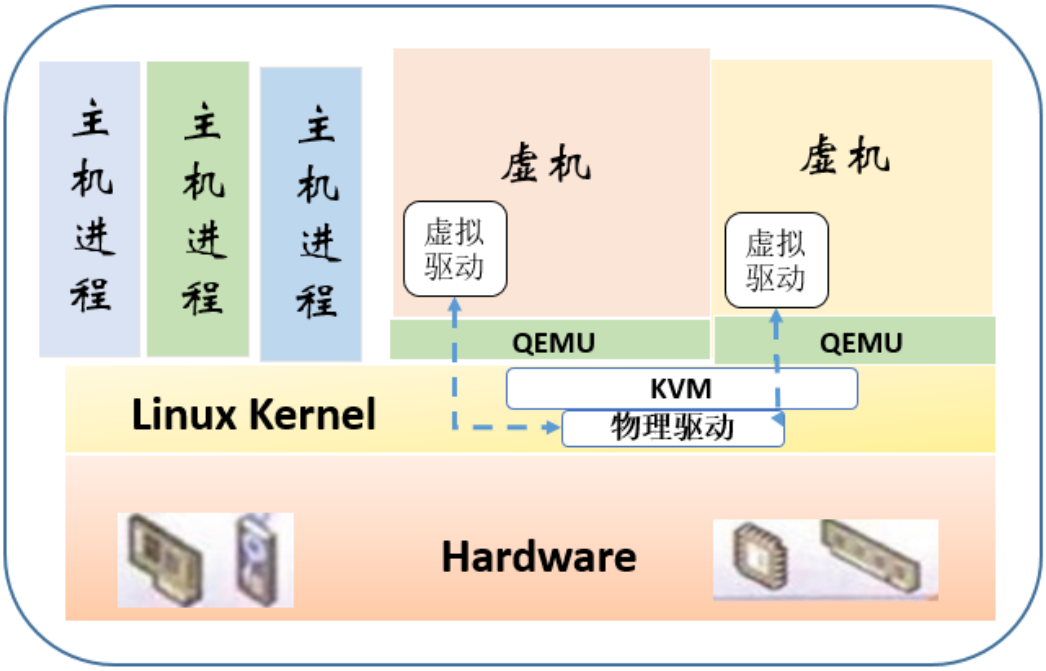
XEN 的 Hypervisor 是服务器经过 BIOS 启动之后载入的首个程序，然后启动一个具有特定权限的虚拟机，称之为 Domain 0(简称 Dom 0)。Dom 0 的操作系统可以是 Linux 或 Unix，Domain 0 实现对 Hypervisor 控制和管理功能。在所承载的虚拟机中，Dom 0 是唯一可以直接访问物理硬件(如存储和网卡)的虚拟机，它通过本身加载的物理驱动，为其它虚拟机(Domain U，简称 DomU)提供访问存储和网卡的桥梁。

XEN 支持两种类型的虚拟机，一类是半虚拟化(PV，Paravirtualization)，另一类是全虚拟化(XEN 称其为 HVM，Hardware Virtual Machine)。半虚拟化需要特定内核的操作系统，如基于 Linux paravirt_ops(Linux 内核的一套编译选项)框架的 Linux 内核，而 Windows 操作系统由于其封闭性则不能被 XEN 的半虚拟化所支持，XEN 的半虚拟化有个特别之处就是不要求 CPU 具备硬件辅助虚拟化，这非常适用于 2007 年之前的旧服务器虚拟化改造。全虚拟化支持原生的操作系统， 特别是针对 Windows 这类操作系统，XEN 的全虚拟化要求 CPU 具备硬件辅助虚拟化，它修改的 Qemu 仿真所有硬件，包括：BIOS、IDE 控制器、 VGA 显示卡、USB 控制器和网卡等。为了提升 I/O 性能，全虚拟化特别针对磁盘和网卡采用半虚拟化设备来代替

仿真设备，这些设备驱动称之为 PV on HVM，为了使 PV on HVM 有最佳性能。CPU 应具备 MMU 硬件辅助虚拟化。

XEN 的 Hypervisor 层非常薄，少于 15 万行的代码量，不包含任何物理设备驱动，这一点与 Hyper-V 是非常类似的，物理设备的驱动均是驻留在 Dom 0 中，可以重用现有的 Linux 设备驱动程序。因此，XEN 对硬件兼容性也是非常广泛的，Linux 支持的，它就支持。

五、 KVM 虚拟化架构



KVM 虚拟化架构示意图

KVM 的全称是 Kernel-based Virtual Machine，字面意思是基于内核虚拟机。其最初是由 Qumranet 公司开发的一个开源项目，2008 年，Qumranet 被 RedHat 所收购，但 KVM 本身仍是一个开源项目，由 RedHat、IBM 等厂商支持。

与 XEN 类似，KVM 支持广泛的 CPU 架构，除了 X86/X86_64 CPU 架构之外，还将会支持大型机(S/390)、小型机(PowerPC、IA64)及 ARM 等。

KVM 充分利用了 CPU 的硬件辅助虚拟化能力，并重用了 Linux 内核的诸多功能，使得 KVM 本身是非常瘦小的，KVM 的创始者 Avi Kivity 声称 KVM 模块仅有约 10000 行代码，但我们不能认为 KVM 的 Hypervisor 就是这个代码量，因为从严格意义来说，KVM 本身并不是 Hypervisor，它仅是 Linux 内核中的一个可装载模块，其功能是将 Linux 内核转换成一个裸金

属的 Hypervisor。

通过 KVM 模块的加载将 Linux 内核转变成 Hypervisor，KVM 在 Linux 内核的用户(User)模式和内核(Kernel)模式基础上增加了客户(Guest)模式。Linux 本身运行于内核模式，主机进程运行于用户模式，虚拟机则运行于客户模式，使得转变后的 Linux 内核可以将主机进程和虚拟机进行统一的管理和调度，这也是 KVM 名称的由来。

KVM 用来模拟 CPU 的运行，但缺少了对 Network 和 I/O 的支持。QEMU-KVM 是一个完整的模拟器，它基于 KVM 上，提供了完整的 Network 和 I/O 支持。其中 Openstack 为了跨 VM 性，所以不会直接控制 QEMU-KVM，而是通过 libvit 的库去间接控制 QEMU-KVM 。

KVM 利用修改的 QEMU 提供 BIOS、显卡、网络、磁盘控制器等的仿真，但对于 I/O 设备(主要指网卡和磁盘控制器)来说，则必然带来性能低下的问题。因此，KVM 也引入了半虚拟化的设备驱动，通过虚拟机操作系统中的虚拟驱动与主机 Linux 内核中的物理驱动相配合，提供近似原生设备的性能。从此可 以看出，KVM 支持的物理设备也即是 Linux 所支持的物理设备。

六、 总结

当前具备虚拟化功能的技术有多种，如微软 Hyper-V、VMware Vsphere、KVM、Xen 等，为了方便管理多种虚拟化技术并向上提供统一、标准的接口，业界开始推出开源云管理平台项目，当前影响最大的当属 OpenStack，当前已有 500+厂商参与此开源项目，不乏 IBM、AMD、Intel 当前已成为事实云平台标准。华为一直扮演着 OpenStack 的倡导者和推动者，积极贡献，目前已成为亚洲唯一的 OpenStack 开源社区白金会员。华为已在全球建设了众多基于 OpenStack 的多种部署模式的云，推动了 OpenStack 作为企业级平台的快速成熟。

传统概念下的半虚拟化和全虚拟化的界线越来越模糊了，而且半虚拟化和全虚拟化得到了有机的整合，如半虚拟化的设备驱动和全虚拟化的虚拟机在上述四种虚拟化架构中得到了统一，很多虚拟化厂商也不再明确自己的虚拟化产品归类(如 VMware 和微软)。

随着 CPU 硬件辅助虚拟化技术发展到了二代，而且新版的操作系统对虚拟化技术的原生支持(如 Windows7 的 Natively Enlightened，Linux 的 paravirt_ops 内核选项)，以及 Hypervisor 对虚拟机的 CPU 调度和内存管理越来越少的干预。则软件做得越少而硬件做得越多，如虚拟机之间内存管理所需用到的地址翻译由软件的影式分页(Shadow Paging)转变为由 CPU 硬

件加速的嵌套分页(Nested Paging)，各种虚拟化技术既有全虚拟化技术对操作系统的兼容性，又有半虚拟化技术所带来的性能优势。

从架构上来看，各种虚拟化技术没有明显的性能差距，稳定性也在逐渐逼近中，各自有着自身的优势场景和市场群体。因此，我们在进行虚拟化技术选型时，不应局限于某一种虚拟化技术，而应该有一套综合管理平台实现对各种虚拟化技术的兼容并蓄，实现不同技术架构的统一管理及跨技术架构的资源调度，最终达到云计算可运营的目。众望所归我司扮演重要角色的 OpenStack 就是这样一个平台。