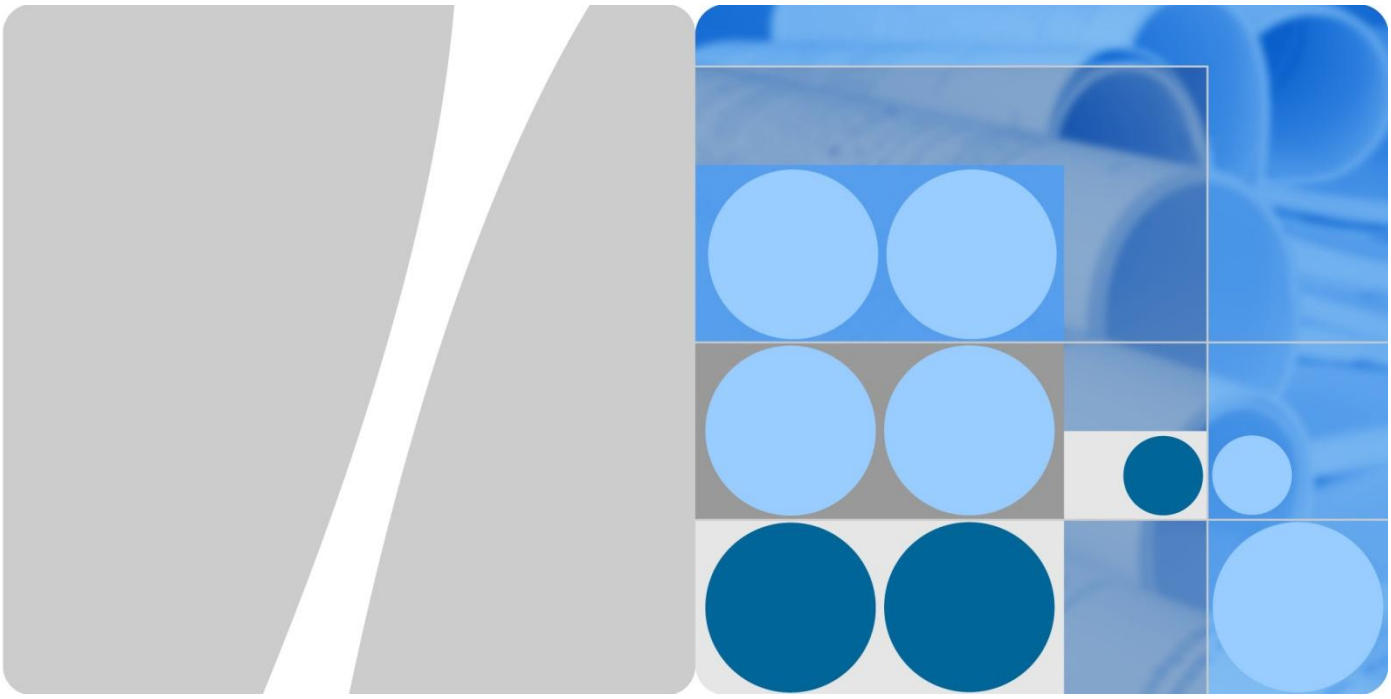


华为 FusionSphere 5.0
虚拟化技术白皮书

文档版本 V1.0

发布日期 2014-09-05

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

前　言

概述

本文档介绍 FusionSphere 产品的虚拟化技术。

读者对象

本文档主要适用于以下工程师：

- 公司 MKT、行销、渠道商在项目拓展中使用

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
	用于警示紧急的危险情形，若不避免，将会导致人员死亡或严重的人身伤害。
	用于警示潜在的危险情形，若不避免，可能会导致人员死亡或严重的人身伤害。
	用于警示潜在的危险情形，若不避免，可能会导致中度或轻微的人身伤害。
	用于传递设备或环境安全警示信息，若不避免，可能会导致设备损坏、数据丢失、设备性能降低或其它不可预知的结果。 “注意”不涉及人身伤害。
	用于突出重要/关键信息、最佳实践和小窍门等。 “说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。

修改记录

修改记录累积了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 V1.0 (2014-09-05)

第一次正式发布。

目 录

1 执行摘要/Executive Summary

华为公司过去二十几年一直在电信领域耕耘，随着电信业务的数据化和电信技术的 IT 化，由云计算等新技术掀起的 ICT 变革不可阻挡，为适应 ICT 行业正在发生的革命性变化与融合，华为做出面向客户的战略调整，华为的创新从电信运营商网络向企业业务、消费者领域延伸，协同发展“云-管-端”业务，而云计算正是其中重要的一环。作为华为 ICT 战略的核心，云计算战略的关键是打造最好的云计算平台，这个平台将支持“百万级服务器扩展、百万 T 的存储能力、百 T 级网络互连能力”，支撑海量信息的计算和存储，并通过“零参与的自动管控”，降低运维成本。

华为 FusionCompute 虚拟化平台（Unified Virtualization Platform）统一虚拟化平台是云计算基础平台的核心组成部分，通过服务器虚拟化技术将存储、网络连接和计算虚拟化有机地结合到一起，使整个 IT 环境比单独的物理硬件具有更高的适用性、可用性和效率，除了满足企业对于降低成本、简化管理、提高安全和敏捷度的诉求，主要为企业关键业务向云计算迁移、构建企业云数据中心提供核心的虚拟化技术和能力。

该技术白皮书从虚拟化技术的发展和原理入手，进而介绍华为虚拟化平台的技术架构、技术特点，重点内容放在对 FusionCompute 虚拟化平台技术特性的描述上，具体分计算虚拟化、存储虚拟化和网络虚拟化三个技术领域详细描述平台主要技术点的功能场景、技术原理和用户价值，突出 FusionCompute 虚拟化平台在性能、可靠性（业务连续性）、安全性以及可维护性等方面所具有的技术领先优势。最后从使用者的角度，描述虚拟化平台在服务器整合、互联网数据中心（IDC）和桌面云（VDI）三大解决方案中提供的核心技术能力，展现 FusionCompute 虚拟化平台及虚拟化技术给用户带来的使用价值。

2 产品简介/Introduction

20 世纪 90 年代，随着 Windows 的广泛使用及 Linux 服务器操作系统的出现奠定了 x86 服务器的行业标准地位，然而 x86 服务器部署的增长带来了新的 IT 基础架构和运作难题，包括：基础架构利用率低、物理基础架构成本日益攀升、IT 管理成本不断提高以及对关键应用故障和灾难保护不足等问题。X86 服务器虚拟化技术的出现，通过将 x86 系统转变成通用的共享硬件基础架构，充分挖掘硬件的潜力，提高硬件的利用效率，降低硬件和运营成本，并且简化运维降低管理成本，最终帮助用户把更多的时间和成本转移到对业务的投入上。

随着云计算和虚拟化技术向构建新一代数据中心方向发展，关键以虚拟化为基础，实现管理以及业务的集中，对数据中心资源进行动态调整和分配，重点满足企业关键应用向 X86 系统迁移对于资源高性能、高可靠、安全性和高可适应性上的要求，同时提高基础架构的自动化管理水平，确保满足基础设施快速适应业务的商业敏捷诉求，同时进一步减少企业的 IT 整体投入。

FusionCompute 虚拟化平台作为介于硬件和操作系统之间的软件层，采用裸金属架构的 X86 虚拟化技术，实现对服务器物理资源的抽象，将 CPU、内存、I/O 等服务器物理资源转化为一组可统一管理、调度和分配的逻辑资源，并基于这些逻辑资源在单个物理服务器上构建多个同时运行、相互隔离的虚拟机执行环境，实现更高的资源利用率，同时满足应用更加灵活的资源动态分配需求，譬如提供热迁移、DRS 等高可用特性，实现更低的运营成本、更高的灵活性和更快速的业务响应速度。

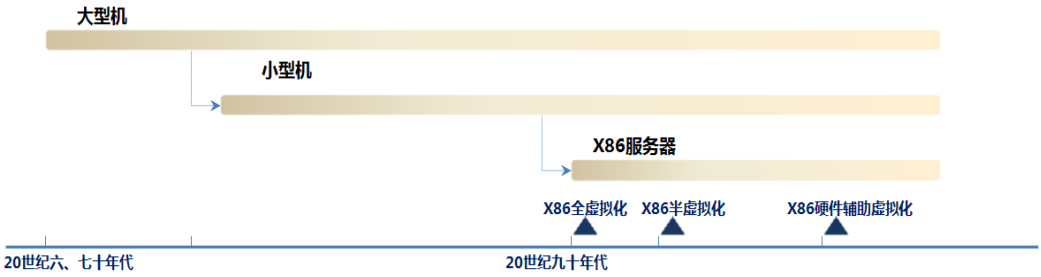
FusionCompute 虚拟化平台重点聚焦关键业务的虚拟化需求，满足 ERP、CRM、核心数据库、Web、电子商务、多应用整合等典型的关键业务应用对于性能、高可靠、高安全以及自动运维方面的业务诉求，提供相应的虚拟化技术特性，加快推动业务和应用的云化。

3 功能描述/Solution

3.1 技术原理

3.1.1 虚拟化历史

图3-1 虚拟化历史



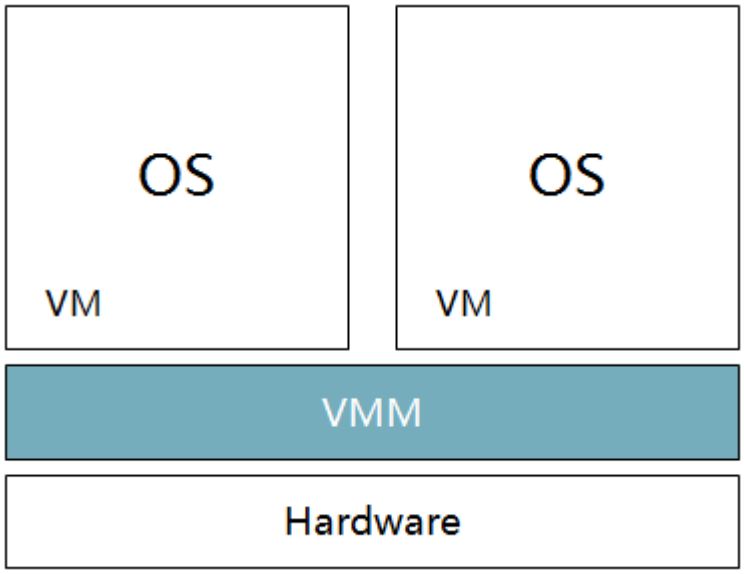
虚拟化技术源于大型机，最早可以追溯到上世纪六、七十年代大型机上的虚拟分区技术，即允许在一台主机上运行多个操作系统，让用户尽可能充分地利用昂贵的大型机资源。随着技术的发展和市场竞争的需要，虚拟化技术向小型机或 UNIX 服务器上移植，只是由于真正使用大型机和小型机的用户还是少数，加上各厂商产品和技术之间的不兼容，使得虚拟化技术不太被公众所关注。（注：由于 X86 架构在设计之初并没有考虑支持虚拟化技术，它本身的结构和复杂性使得在其之上进行虚拟化非常困难，早期的 X86 架构并没有成为虚拟化技术的受益者）

20 世纪 90 年代，以 VMware 为代表的部分虚拟化软件厂商采用一种软件解决方案，以 VMM(Virtual Machine Monitor, VMM 虚拟机监视器)为中心使 X86 服务器平台实现虚拟化。然而这种纯软件的“全虚拟化”模式，每个 Guest OS（客户操作系统）获得的关键平台资源都要由 VMM 控制和分配，需要利用二进制转换，而二进制转换带来的开销使得“完全虚拟化”的性能大打折扣。为解决性能问题，出现了一种新的虚拟化技术“半虚拟化”，即不需要二进制转换，而是通过对客户操作系统进行代码级修改，使定制的 Guest OS 获得额外的性能和高扩展性，但是修改 Guest OS 也带来了系统指令级的冲突及运行效率问题，需要投入大量优化的工作。当前，虚拟化技术已经发展到了硬件支持的阶段，“硬件虚拟化”技术就是把纯软件虚拟化技术的各项功能用硬件电路来实现，可减少 VMM 运行的系统开销，可同时满足 CPU 半虚拟化和二进制转换技术的需求，

使 VMM 的设计得到简化，进而使 VMM 能够按通用标准进行编写。硬件辅助虚拟化技术除了在处理器上集成硬件辅助虚拟化指令，同时提供 I/O 方面的虚拟化支持，最终将实现整个平台的虚拟化。X86 虚拟化技术的实现和发展，都向人们展示了虚拟化应用的广阔前景。

3.1.2 虚拟化原理

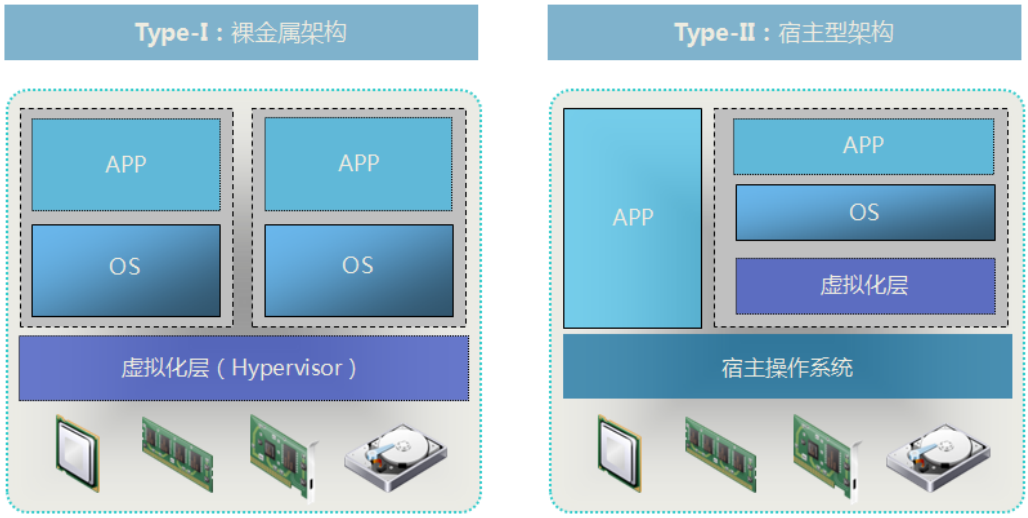
图3-2 虚拟化原理



虚拟化（Virtualization）是资源的逻辑表示，而不受物理限制的约束。虚拟化技术的实现形式是在系统中加入一个虚拟化层，将下层的资源抽象成另一形式的资源，提供给上层使用。

服务器虚拟化就是使软件和硬件相互分离，把软件从主要安装硬件中分离出来。它可以在服务器架构中的多个位置实施虚拟化，包括应用程序与操作系统之间（轻量级容器）或操作系统与硬件之间,后者指位于下层的虚拟化软件通过空间上的分割、时间上的分时以及模拟，抽象出一个虚拟的硬件接口，向上层操作系统提供一个与它原先期待一致的服务器硬件环境，使得上层操作系统可以直接运行在虚拟环境上，可允许多个操作系统同时运行在单个物理服务器上。

图3-3 虚拟化架构



服务器虚拟化的虚拟化软件层称为虚拟机监控器 (Virtual Machine Monitor, VMM)，也称 Hypervisor，常见的 Hypervisor 分两类：

Type-I（裸金属型）指 VMM 直接运作在裸机上,使用和管理底层的硬件资源，Guest OS 对真实硬件资源的访问都要通过 VMM 来完成，作为底层硬件的直接操作者，VMM 拥有硬件的驱动程序。

Type-II 型（宿主型）指 VMM 之下还有一层宿主操作系统，由于 Guest OS 对硬件的访问必须经过宿主操作系统，因而带来了额外的性能开销，但可充分利用宿主操作系统提供的设备驱动和底层服务来进行内存管理、进程调度和资源管理等。

服务器虚拟化前后的巨大差异，源于虚拟机与物理服务器的本质区别上：

虚拟机的定义：虚拟机 (Virtual Machine) 是由虚拟化层提供的高效、独立的虚拟计算机系统，每台虚拟机都是一个完整的系统，它具有处理器、内存、网络设备、存储设备和 BIOS，因此操作系统和应用程序在虚拟机中的运行方式与它们在物理服务器上的运行方式没有什么区别。

虚拟机的本质区别：与物理服务器相比，虚拟机不是由真实的电子元件组成，而是由一组虚拟组件（文件）组成，这些虚拟组件与物理服务器的硬件配置无关，关键与物理服务器相比，虚拟机具有以下优势：

- **抽象解耦：**1. 可在任何 X86 架构的服务器上运行；2. 上层应用操作系统不需修改即可运行；
- **分区隔离：**1. 可与其他虚拟机同时运行；2. 实现数据处理、网络连接和数据存储的安全隔离；
- **封装移动：**

  1.可封装于文件之中，通过简单的文件复制实现快速部署、备份及还原；

  2.可便捷地将整个系统（包括虚拟硬件、操作系统和配置好的应用程序）在不同的物理服务器之间进行迁移，甚至可以在虚拟机正在运行的情况下进行迁移；
- **弹性扩展：**

1.可对单个物理服务器上的虚拟资源（VCPU、VNIC 等）进行按需动态扩展（不停机）；

2.可作为即插即用的虚拟工具进行构建和分发，按集群弹性资源分配机制实现动态扩展；

3.1.3 虚拟化实现

VMM (Virtual Machine Monitor)对物理资源的虚拟可以划分为三个部分：CPU 虚拟化、内存虚拟化和 I/O 设备虚拟化,其中以 CPU 的虚拟化最为关键。

CPU 虚拟化

经典的虚拟化方法：

现代计算机体系结构一般至少有两个特权级（即用户态和核心态，x86 有四个特权级 Ring0~Ring3）用来分隔系统软件和应用软件。那些只能在处理器的最高特权级（内核态）执行的指令称之为特权指令，一般可读写系统关键资源的指令（**即敏感指令**）决大多数都是特权指令（X86 存在若干敏感指令是非特权指令的情况）。如果执行特权指令时处理器的状态不在内核态，通常会引发一个异常而交由系统软件来处理这个非法访问（**陷入**）。经典的虚拟化方法就是使用“特权解除”和“陷入-模拟”的方式，即将 Guest OS 运行在非特权级，而将 VMM 运行于最高特权级（完全控制系统资源）。解除了 Guest OS 的特权级后，Guest OS 的大部分指令仍可以在硬件上直接运行，只有执行到特权指令时，才会陷入到 VMM 模拟执行（陷入-模拟）。“陷入-模拟” 的本质是保证可能影响 VMM 正确运行的指令由 VMM 模拟执行，大部分的非敏感指令还是照常运行。

X86 的虚拟化漏洞：

因为 X86 指令集中有若干条指令是需要被 VMM 捕获的敏感指令,但是却不是特权指令(称为**临界指令**)，因此“特权解除”并不能导致他们发生陷入模拟，执行它们不会发生自动的“陷入”而被 VMM 捕获，从而阻碍了指令的虚拟化。具体 X86 下的敏感指令分类大致如下：

1、访问或修改机器状态或虚拟机状态的指令。

2、访问或修改敏感寄存器或存储单元的指令，比如访问时钟寄存器和中断寄存器。

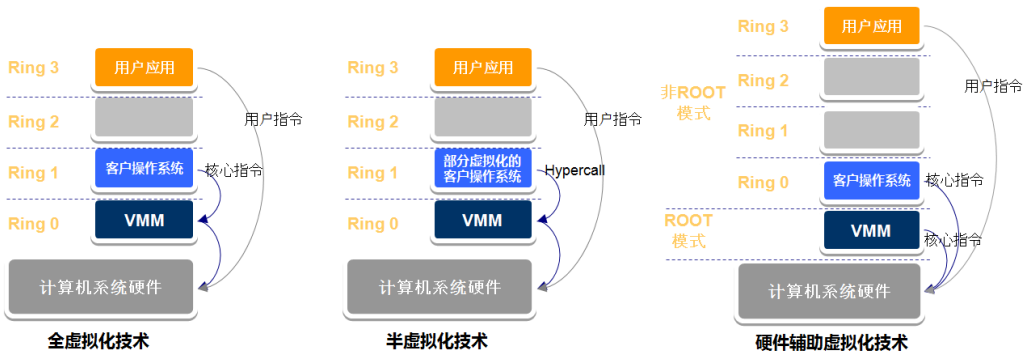
3、访问存储保护系统或内存、地址分配系统的指令。（段页之类）

4、所有 I/O 指令。

其中的(1)和(4)都是特权指令，在内核态下执行时会自动产生陷阱被 VMM 捕获，但是(2)和(3)不是特权指令，而是临界指令。部分临界指令会因为 Guest OS 的权限解除执行失败，但是却不会抛出异常，所以不能被捕获，譬如：(3)中的 VERW 指令。

X86 的虚拟化方法：

图3-4 X86 的虚拟化



由于 x86 指令集中有十多条敏感指令不是特权指令，因此 x86 无法使用经典的虚拟化技术完全虚拟化。鉴于 x86 指令集本身的局限，长期以来针对 x86 的虚拟化实现大致分为两派，即以 VMWare 为代表的 Full virtualization 派和以 Xen 为代表的 Para virtualization 派。两派区别主要在对非特权敏感指令的处理上，Full 派采用的是动态的方法，即：运行时监测，捕捉后在 VMM 中模拟；而 Para 派则主动进攻，将所有用到的非特权敏感指令全部替换，这样就少掉了大量的陷入-> 上下文切换 -> 模拟 -> 上下文切换过程，获得了大幅的性能提升。

1、X86“全虚拟化”（指所抽象的 VM 具有完全的物理机特性，OS 在其上运行不需要任何修改）

Full 派秉承无需修改直接运行的理念，对“运行时监测，捕捉后模拟”的过程进行优化。该派内部之实现又有些差别，其中以 VMWare 为代表的基于二进制翻译（BT）的全虚拟化为代表，其主要思想是在执行时将 VM 上执行的 Guest OS 指令，翻译成 x86 指令集的一个子集，其中的敏感指令被替换成陷入指令。翻译过程与指令执行交叉进行，不含敏感指令的用户态程序可以不经翻译直接执行。

2、X86“半虚拟化”（指需 OS 协助的虚拟化，在其上运行的 OS 需要修改）

Para 派的基本思想是通过修改 Guest OS 的代码，将含有敏感指令的操作，替换为对 VMM 的超调用 Hypercall，类似 OS 的系统调用，将控制权转移到 VMM，该技术因 Xen 项目而广为人知。该技术的优势在于 VM 的性能能接近于物理机，缺点在于需要修改 Guest OS（如：Windows 不支持修改）及增加的维护成本，关键修改 Guest OS 会导致操作系统对特定 hypervisor 的依赖性，因此很多虚拟化厂商基于 Xen 开发的虚拟化产品部分已经放弃了 Linux 半虚拟化，而专注基于硬件辅助的全虚拟化开发，来支持未经修改的操作系统。

3、X86“硬件辅助虚拟化”：

其基本思想就是引入新的处理器运行模式和新的指令，使得 VMM 和 Guest OS 运行于不同的模式下，Guest OS 运行于受控模式，原来的一些敏感指令在受控模式下全部会陷入 VMM，这样就解决了部分非特权的敏感指令的“陷入-模拟”难题，而且模式切换时上下文的保存恢复由硬件来完成，这样就大大提高了“陷入-模拟”时上下文切换的效率。

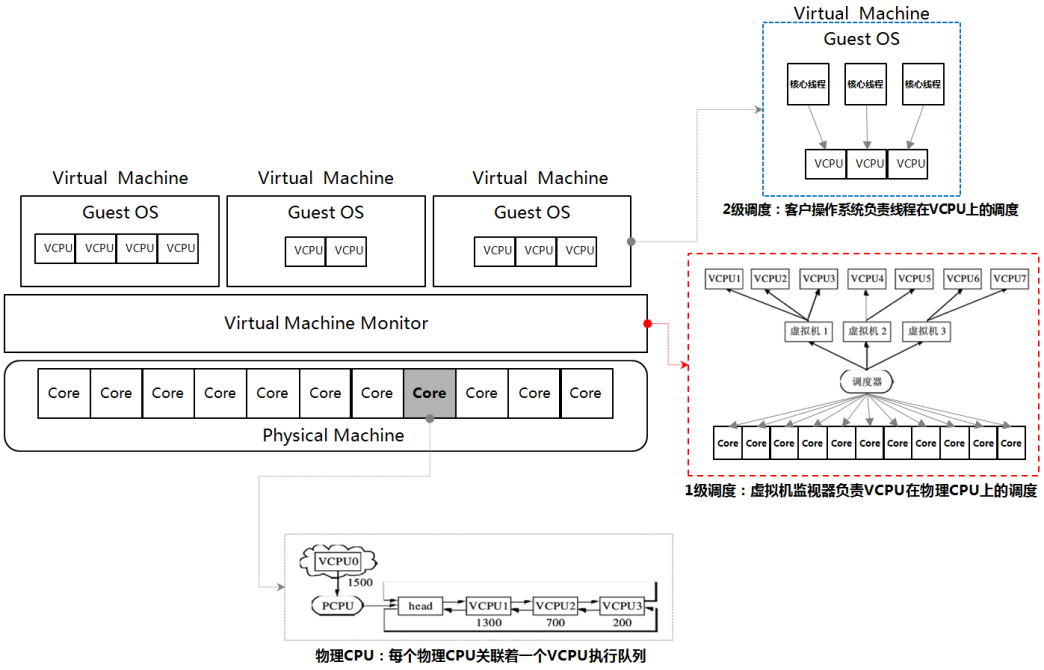
以 Intel VT-x 硬件辅助虚拟化技术为例，该技术增加了在虚拟状态下的两种处理器工作模式：根（Root）操作模式和非根（Non-root）操作模式。Xen 运作在 Root 操作模式下，而 Guest OS 运行在 Non-root 操作模式下。这两个操作模式分别拥有自己的特权级环，

Xen 和未经修改内核的 Guest OS 运行在这两个操作模式的 0 环。这样，既能使 Xen 运行在 0 环，也能使 Guest OS 运行在 0 环，避免了修改 Guest OS。Root 操作模式和 Non-root 操作模式的切换是通过新增的 CPU 指令（VMXON,VMXOFF 等）来完成。

硬件辅助虚拟化技术消除了操作系统的 ring 转换问题，降低了虚拟化门槛，支持任何操作系统的虚拟化而无须修改 OS 内核，得到了虚拟化软件厂商的支持。硬件辅助虚拟化技术已经逐渐消除软件虚拟化技术之间的差别，并成为未来的发展趋势。

核心技术原理： vCPU 调度分配机制

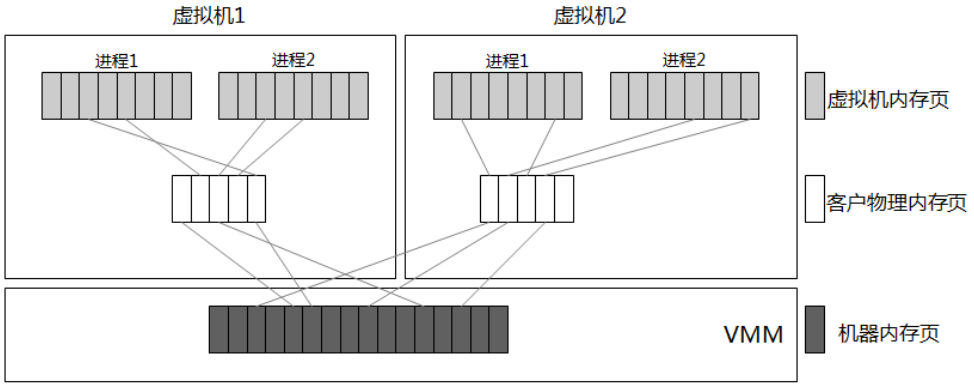
图3-5 vCPU 调度分配机制



从虚拟机系统的结构与功能划分可以看出，客户操作系统与虚拟机监视器共同构成了虚拟机系统的两级调度框架，如图所示是一个多核环境下虚拟机系统的两级调度框架。客户操作系统负责第 2 级调度,即线程或进程在 VCPU 上的调度（将核心线程映射到相应的虚拟 CPU 上）。虚拟机监视器负责第 1 级调度，即 VCPU 在物理处理单元上的调度。两级调度的调度策略和机制不存在依赖关系。VCPU 调度器负责物理处理器资源在各个虚拟机之间的分配与调度,本质上即把各个虚拟机中的 VCPU 按照一定的策略和机制调度在物理处理单元上可以采用任意的策略来分配物理资源，满足虚拟机的不同需求。VCPU 可以调度在一个或多个物理处理单元执行（分时复用或空间复用物理处理单元），也可以与物理处理单元建立一对一固定的映射关系（限制访问指定的物理处理单元）。

内存虚拟化

图3-6 内存虚拟化三层模型



因为 VMM (Virtual Machine Monitor) 掌控所有系统资源，因此 VMM 握有整个内存资源，其负责页式内存管理，维护虚拟地址到机器地址的映射关系。因 Guest OS 本身亦有页式内存管理机制，则有 VMM 的整个系统就比正常系统多了一层映射：

A. 虚拟地址(VA)，指 Guest OS 提供给其应用程序使用的线性地址空间；

B. 物理地址(PA)，经 VMM 抽象的、虚拟机看到的伪物理地址；

C. 机器地址(MA)，真实的机器地址，即地址总线上出现的地址信号；

映射关系如下：Guest OS: PA = f(VA)、VMM: MA = g(PA)

VMM 维护一套页表，负责 PA 到 MA 的映射。Guest OS 维护一套页表，负责 VA 到 PA 的映射。实际运行时，用户程序访问 VA1，经 Guest OS 的页表转换得到 PA1，再由 VMM 介入，使用 VMM 的页表将 PA1 转换为 MA1。

页表虚拟化技术原理：

普通 MMU 只能完成一次虚拟地址到物理地址的映射，在虚拟机环境下，经过 MMU 转换所得到的“物理地址”并不是真正的机器地址。若需得到真正的机器地址，必须由 VMM 介入，再经过一次映射才能得到总线上使用的机器地址。如果虚拟机的每个内存访问都需要 VMM 介入，并由软件模拟地址转换的效率是很低下的，几乎不具有实际可用性，为实现虚拟地址到机器地址的高效转换，现普遍采用的思想是：由 VMM 根据映射 f 和 g 生成复合的映射 fg，并直接将这个映射关系写入 MMU。当前采用的页表虚拟化方法主要是 MMU 类虚拟化（MMU Paravirtualization）和影子页表，后者已被内存的硬件辅助虚拟化技术所替代。

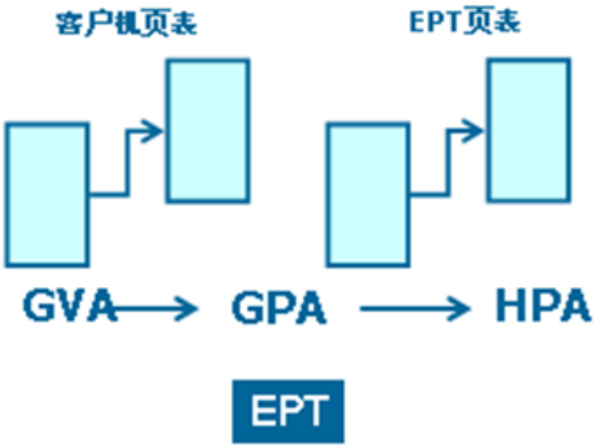
1、MMU Paravirtualization

其基本原理是：当 Guest OS 创建一个新的页表时，会从它所维护的空闲内存中分配一个页面，并向 Xen 注册该页面，Xen 会剥夺 Guest OS 对该页表的写权限，之后 Guest OS 对该页表的写操作都会陷入到 Xen 加以验证和转换。Xen 会检查页表中的每一项，确保他们只映射了属于该虚拟机的机器页面，而且不得包含对页表页面的可写映射。后 Xen 会根据自己所维护的映射关系，将页表项中的物理地址替换为相应的机器地址，最后再

把修改过的页表载入 MMU。如此，MMU 就可以根据修改过页表直接完成虚拟地址到机器地址的转换。

2、内存硬件辅助虚拟化

图3-7 内存硬件辅助虚拟化技术原理图



内存的硬件辅助虚拟化技术是用于替代虚拟化技术中软件实现的“影子页表”的一种硬件辅助虚拟化技术，其基本原理是：GVA（客户操作系统的虚拟地址）-> GPA（客户操作系统的物理地址）-> HPA（宿主操作系统的物理地址）两次地址转换都由 CPU 硬件自动完成（软件实现内存开销大、性能差）。以 VT-x 技术的页表扩充技术 Extended Page Table（EPT）为例，首先 VMM 预先把客户机物理地址转换到机器地址的 EPT 页表设置到 CPU 中；其次客户机修改客户机页表无需 VMM 干预；最后，地址转换时，CPU 自动查找两张页表完成客户机虚拟地址到机器地址的转换。使用内存的硬件辅助虚拟化技术，客户机运行过程中无需 VMM 干预，去除了大量软件开销，内存访问性能接近物理机。

I/O 设备虚拟化

VMM 通过 I/O 虚拟化来复用有限的外设资源，其通过截获 Guest OS 对 I/O 设备的访问请求，然后通过软件模拟真实的硬件，目前 I/O 设备的虚拟化方式主要有三种：设备接口完全模拟、前端 / 后端模拟、直接划分。

1、设备接口完全模拟：

即软件精确模拟与物理设备完全一样的接口，Guest OS 驱动无须修改就能驱动这个虚拟设备，Vmware 即使用该方法。

优点：没有额外的硬件开销，可重用现有驱动程序；

缺点：为完成一次操作要涉及到多个寄存器的操作，使得 VMM 要截获每个寄存器访问并进行相应的模拟，这就导致多次上下文切换；由于是软件模拟，性能较低。

2、前端 / 后端模拟：

VMM 提供一个简化的驱动程序（后端, Back-End），Guest OS 中的驱动程序为前端(Front-End, FE)，前端驱动将来自其他模块的请求通过与 Guest OS 间的特殊通信机制直接发送给 Guest OS 的后端驱动，后端驱动在处理完请求后再发回通知给前端，Xen 即采用该方法。

优点：基于事务的通信机制，能在很大程度上减少上下文切换开销，没有额外的硬件开销；

缺点：需要 VMM 实现前端驱动，后端驱动可能成为瓶颈。

3、直接划分：

即直接将物理设备分配给某个 Guest OS，由 Guest OS 直接访问 I/O 设备（不经 VMM），目前与此相关的技术有 IOMMU（Intel VT-d, PCI-SIG 之 SR-IOV 等），旨在建立高效的 I/O 虚拟化直通道。

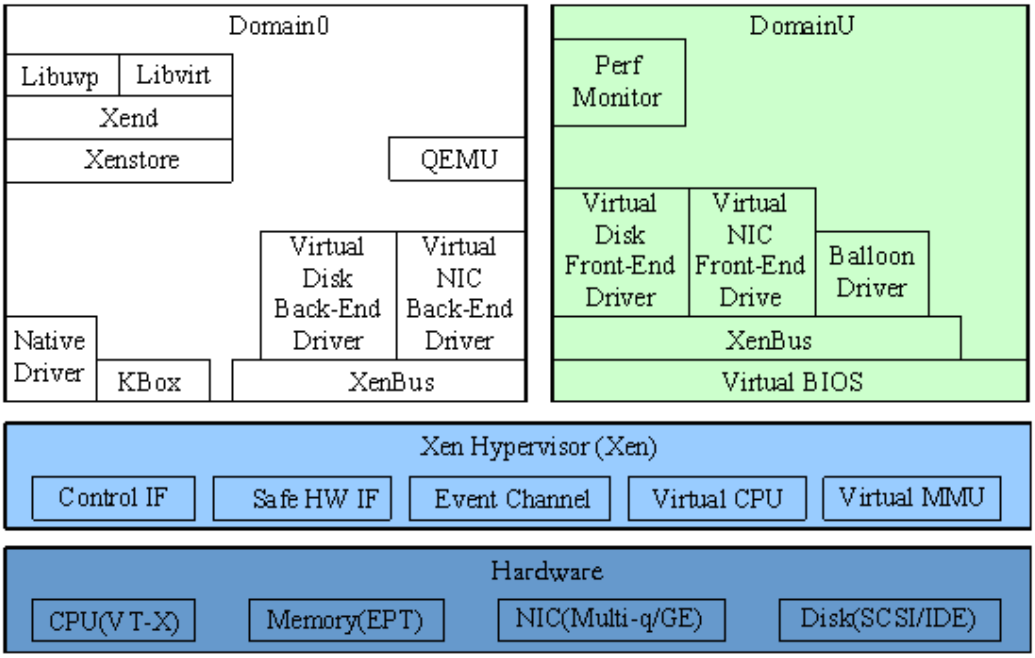
优点：可重用已有驱动，直接访问减少了虚拟化开销；

缺点：需要购买较多额外的硬件。

3.2 功能架构

3.2.1 FusionCompute 虚拟化平台技术架构

图3-8 FusionCompute 虚拟化平台架构图



FusionCompute 虚拟化平台采用的是裸金属化架构，主要由以下部分组成：

- Xen Hypervisor 层（图中深蓝色部分）

Xen Hypervisor 是一个介于硬件和操作系统之间的软件层，它负责在各虚拟机之间进行 CPU 调度和内存分配（partitioning）。Xen Hypervisor 不仅抽象出硬件层，同时控制虚拟机的执行，因为这些虚拟机共享同一个处理环境。Xen Hypervisor 不会处理网络、存储设备、视频以及其他 I/O。

- Domain0（图中白色部分）

Domain 0 是一个修改过的 Linux kernel，是运行在 Xen Hypervisor 之上的特权虚拟机，它拥有访问物理 I/O 资源的权限，同时和系统上运行的其他虚拟机（DomainU）进行交互。Domain 0 需要在其它 Domain 启动之前启动。

- DomainU（图中绿色部分）

DomainU 分为两种：

1、**DomainU PV Guest**：运行在 Xen Hypervisor 上的所有半虚拟化（paravirtualized）虚拟机被称为“Domain U PV Guests”，其上运行着被修改过内核的操作系统，如 Linux、Solaris、FreeBSD 等其它 UNIX 操作系统。

2、**DomainU HVM Guest**：所有的全虚拟化虚拟机被称为“Domain U HVM Guests”，其上运行着不用修改内核的操作系统，如 Windows 等。完全虚拟化由于不需要修改客户机操作系统，因此具有很好的兼容性和同时支持异种操作系统或不同版本操作系统的能力。相反，半虚拟化技术则通常具有比完全虚拟化技术更好的性能。**FUSIONCOMPUTE 虚拟化平台只支持全虚拟化虚拟机。**

其中 **Xen Hypervisor** 是系统的核心，负责为上层运行的操作系统提供虚拟化的硬件资源，负责管理和分配这些资源，并确保上层虚拟机之间的相互隔离。**Domain0** 是一个特权虚拟机，内部包含了真实的设备驱动（原生设备驱动），可直接访问物理硬件，负责与 Hypervisor 提供的管理 API 交互，并通过 Agent 接受管理系统的管理指令，实现对其它虚拟机（DomainU）的管理。

为了提升 I/O 虚拟化性能，子系统采用**分离设备驱动模型**实现 I/O 的虚拟化。该模型将设备驱动划分为前端驱动程序、后端驱动程序和原生驱动三个部分，其中前端驱动在 DomainU 中运行，而后端驱动和原生驱动则在 Domain0 中运行。前端驱动负责将 DomainU 的 I/O 请求传递到 Domain0 中的后端驱动，后端驱动解析 I/O 请求并映射到物理设备，提交给相应的设备驱动程序控制硬件完成 I/O 操作。

3.2.2 FusionCompute 虚拟化平台功能

图3-9 FUSIONCOMPUTE 虚拟化平台功能特性

FUSIONCOMPUTE 虚拟化平台主要定位企业关键应用领域，采用业界领先的 Xen 技术，实现在开源基础上持续增强和优化，提供完整的虚拟机生命周期管理功能，充分发挥 Xen 的性能和安全方面的技术优势，并利用 Intel 和 AMD 的硬件辅助虚拟化技术，提供关键应用对于高性能、高可靠、安全性和高可适应性上的各种虚拟化功能要求。

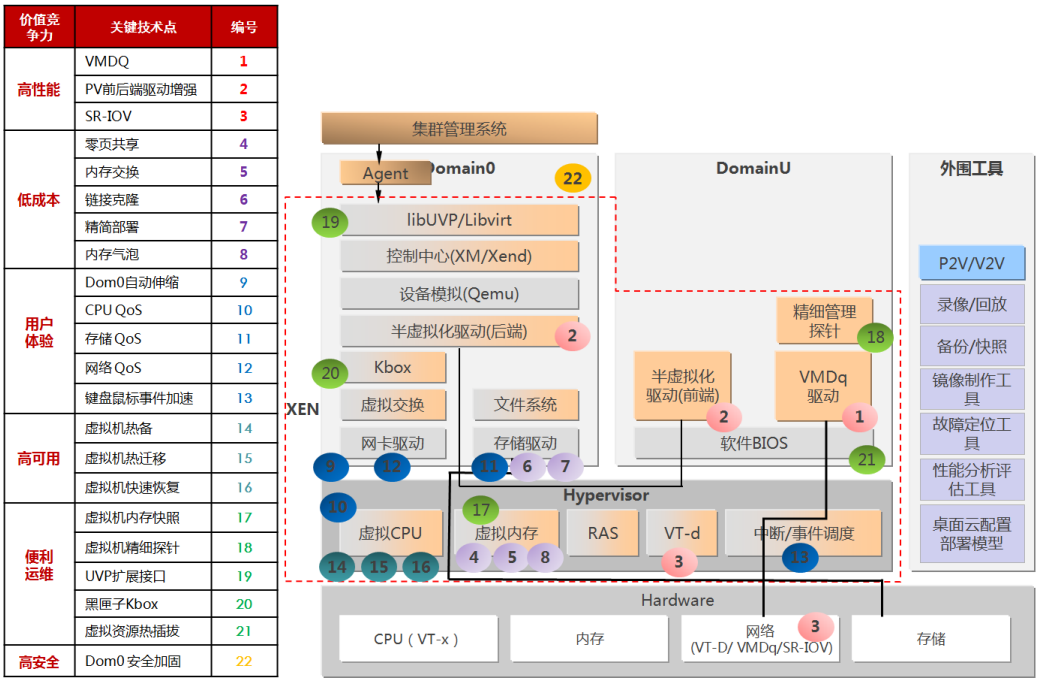
首先，在基础架构服务层，提供持续的性能优化和增强功能，在计算虚拟化上的内存复用技术、GPU 虚拟化技术等满足用户对于性能和体验方面的要求，在存储虚拟化上的链接克隆技术、快照备份技术可有效降低硬件采购成本，在网络虚拟化结合硬件辅助虚拟化技术提供 SR-IOV 网卡直通等技术，可满足应用对于高 I/O 性能的要求。

其次，在应用程序服务层，提供可用性、可维护性和安全性方面的功能支持，包括提供虚拟机热迁移、虚拟机热备份和虚拟资源的热插拔技术，降低系统计划内/外宕机实际，提高业务的连续性；提供 Kbox、Guest OS 故障检测功能，提高系统的可维护性；提供虚拟机安全加固、VLAN 和安全组特性，提高企业应用的安全性保障；

以上 FUSIONCOMPUTE 虚拟化平台所有虚拟化特性，可以为企业关键应用整体运营带来显著的改善。

3.2.3 FusionCompute 虚拟化平台技术特点

图3-10 FUSIONCOMPUTE 虚拟化平台技术特点



华为 FUSIONCOMPUTE 虚拟化平台通过对开源 xen 进行安全加固、功能扩展、性能优化和可靠性增强，着力打造安全、高效、稳定、开放的虚拟化平台，主要具备如下特点：

1. 高性能

在计算虚拟化上，提供 CPU 调度优化，实现软实时调度，降低 CPU 响应延时和 CPU cache 失效，提高任务实时性，满足电信级需求；在 I/O 虚拟化上，采用高效的“前后驱”通信技术，减少 CPU 模式切换和内存拷贝带来的开销，同时充分利用硬件辅助虚拟化技术，提供 VMDQ 和 SR-IOV 特性，减少中断次数和内存拷贝，提高虚拟机 IO 性能。

2. 低成本

在计算虚拟化上，提供内存气泡、内存零页共享和内存交换技术，并通过智能复用以上三种技术提升内存复用比，在同等内存资源条件下可提升虚拟机密度，降低硬件(内存)采购成本。在存储虚拟化上，采用存储链接克隆、存储瘦分配技术，减少对虚拟磁盘的过度调配，可节省或延迟存储设备采购时间，降低硬件(存储)采购成本。

3. 高可用性

提供虚拟机热迁移技术，支持最高 8 个虚拟机的并发迁移任务，可将业务无中断的迁移到其他物理机上，支持 VCPU、VMemory、VDisk、VNIC 的热插拔功能，减少系统计划内宕机时间，同时提供虚拟机的热备份技术，确保业务数据的运行时安全和灾难时可恢复。

4. 高安全性

提供自研 OS 作为管理域 Domain0，实现严格的操作权限控制、服务裁剪、网络端口扫描和访问控制、病毒入侵检测和防护、系统风险扫描和预警等；内置的 SuseFireWall 虚拟防火墙提供灵活的安全访问策略配置，结合提供虚拟机安全组和 VLAN 技术，实现多层次安全纵深防御。

5. 可管理性

提供虚拟机运行状态查询能力、虚拟机动态调整能力及虚拟机远程安装部署能力，支撑虚拟机大规模运维管理；提供电信级“黑匣子”技术，在系统出现异常或宕机时自动存储 VMM 内核日志、系统快照、内核诊断信息及临终遗言，并保存至非易失性存储设备或自动传送至网络服务器；提供 CPU/存储/网络的 QOS 功能，支持进程级的资源优先级控制，确保关键应用或虚拟机获得所需的服务器资源，提高用户的使用体验。

6. 开放性

扩充业界标准接口 Libvirt 作为 VMM 管理接口，提供对外开放接口，开放部分平台 VMM 功能代码，并与 Xen 开源社区和其它商业公司共同维护，具有广泛的兼容和生态链支持。

3.3 计算虚拟化

鉴于关键业务连续性对企业业务的重要性、需要保证虚拟机的性能满足业务的需求、保证关键应用的响应速度。FUSIONCOMPUTE 虚拟化平台针对高并发，高负载应用场景下的计算性能优化和调度算法优化，发挥多核处理的性能优势，提供多种内存复用技术，并结合 CPU 硬件辅助虚拟化技术充分发挥平台的性能优势。

3.3.1 内存复用技术

FUSIONCOMPUTE 虚拟化平台提供多种内存复用技术和灵活自动的内存复用策略。对于某些物理内存资源比较紧张的场景，如果用户希望运行超过物理内存能力的虚拟机，以达到节省成本的目的，就需要有内存复用策略来动态地对内存资源进行分配和复用。内存复用策略通过内存复用技术，提升物理内存利用率的同时，尽可能减少对虚拟机性能的影响。客户无需关心何时调用和怎么调用几种复用技术，只需简单配置和开启复用策略后就能达到提升虚拟机密度的目的。

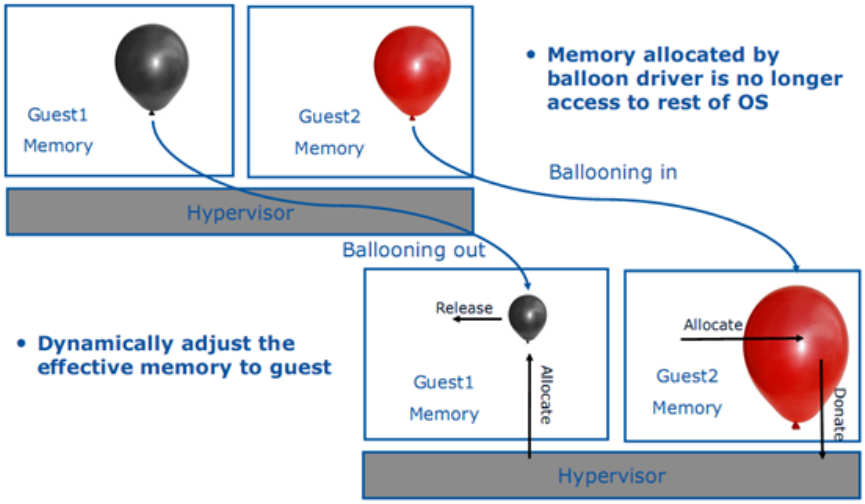
FUSIONCOMPUTE 虚拟化平台的内存复用技术有以下三种：内存气泡、内存零页共享和内存交换技术。

1. **内存气泡技术**(Ballooning)

内存气泡技术是一种 VMM 通过“诱导”客户机操作系统来回收或分配客户机所拥有的宿主机物理内存的技术。当客户机物理内存足够时，客户机操作系统从其闲置客户机物理机内存链表中返回客户机物理内存给气球；当客户机物理内存资源稀缺时，客户机操作系统必须回收一部分客户机物理内存，以满足气球申请客户机物理内存的需要。通过 Balloon Driver 模块，从源虚拟机申请可用内存页面，通过 Grant Table 授权给目标虚拟机，并更新虚拟机物理地址和机器地址映射关系表。

通过使用 Ballooning 技术，可以提升内存使用效率。

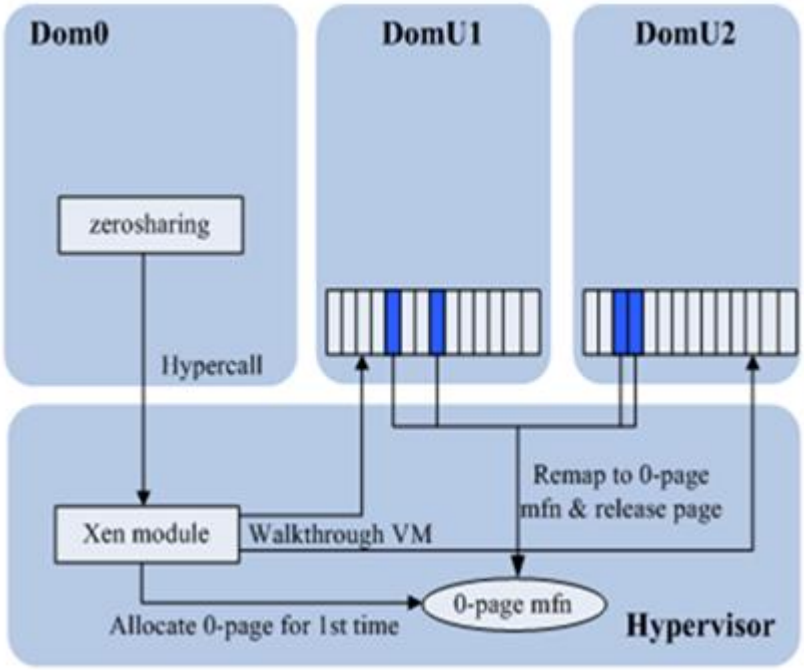
图3-11 Balloon 技术原理图



2. 零页共享技术

内存零页共享技术作为内存复用技术的一种，能有效地识别和释放虚拟机内未分配使用的零页，以达到提高内存复用率的目的。客户开启零页共享技术后，能实时从虚拟机内部把零页进行共享，从而把其占用的内存资源释放出来给其他虚拟机使用，以创建更多的虚拟机，实现提高虚拟机密度的目的。与内存气泡技术不同，零页共享后的内存页对于虚拟机来说还是可用的，虚拟机可以随时根据需要再收回这部分内存，用户体验相对来说更加友好。

图3-12 内存零页共享原理

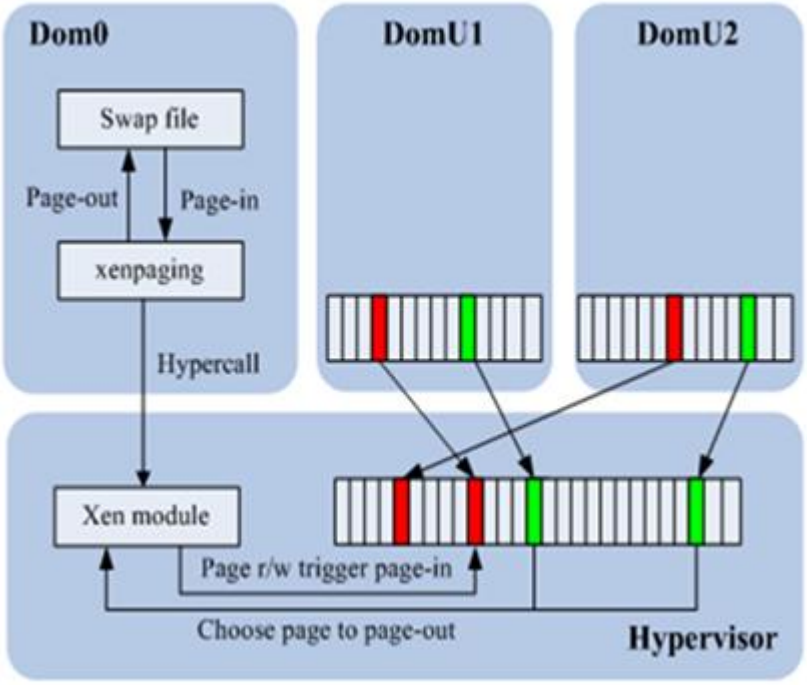


用户进程定时扫描虚拟机的内存数据，如果发现其数据内容全为零，则通过修改 P2M 映射的形式把其指向一个特定的零页。从而做到在物理内存中仅保留一份零页拷贝，虚拟机的所有零页均指向该页，从而达到节省内存资源的目的。当零页数据发生变动时，由 Xen 动态地分配一页内存出来给虚拟机，使修改后的数据有内存页进行存放，因此对于 GuestOS 来说，整个零页共享过程是完全不感知的。

3. 内存交换技术

内存交换技术作为内存复用技术的一种，能通过 Xen 把虚拟机内存数据换出到存储介质上的交换文件中，从而释放内存资源，以达到提高内存复用率的目的。由于内存气泡和零页共享的数量与虚拟机本身的内存使用情况强相关，因此其效果不是很稳定，用户使用内存交换技术，可以弥补上述不足，即可以保证释放出一定量的内存空间出来（理论上所有虚拟机内存都能交换出来），但同时也会带来一定程度的虚拟机性能下降。

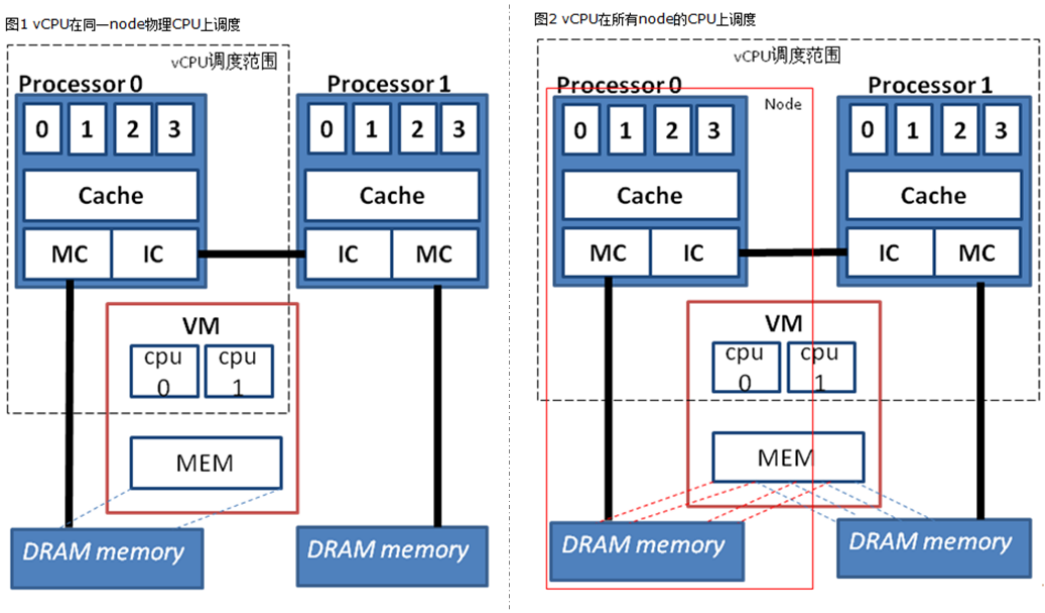
图3-13 内存交换技术原理图



内存交换触发时，根据用户需要告知 Xen 需要向某个虚拟机交换出一定量的内存页出来，Xen 按一定的选页策略从虚拟机中选择相应数量的页后，把页数据保存到存储介质上的交换文件中，同时释放原先存放数据的那些页供其他虚拟机使用。当虚拟机读写的页正好是被换出的页时，在缺页处理时 Xen 会重新为其分配一页内存，然后从存储介质上的交换文件中把相应的页交换回新分配的内存页中，同时再选择另外一页内存交换出去，从而保证虚拟机对页的正常读写的同时，稳定交换页的数量。这个过程与零页共享一样，对 GuestOS 都是不感知的。

3.3.2 Host NUMA

图3-14 Host NUMA 原理图



技术原理：

FUSIONCOMPUTE 虚拟化平台实现的 Host NUMA 主要提供 CPU 负载均衡机制，解决 CPU 资源分配不平衡引起的 VM 性能瓶颈问题，当启动 VM 时，Host NUMA 根据当时主机内存和 CPU 负载，选择一个负载较轻的 node 放置该 VM，使 VM 的 CPU 和内存资源分配在同一个 node 上。如图 1 所示，Host NUMA 把 VM 的物理内存放置在一个 node 上，对 VM 的 vCPU 调度范围限制在同一个 node 的物理 CPU 上，并将 VM 的 vCPU 亲和性绑定在该 node 的物理 CPU 上。考虑到 VM 的 CPU 负载是动态变化，在初始放置的 node 上，node 的 CPU 资源负载也会随之变化，这会导致某个 node 的 CPU 资源不足，而另一个 node 的 CPU 资源充足，在此情况下，Host NUMA 会从 CPU 资源不足的 node 上选择 VM，把 VM 的 CPU 资源分配在 CPU 资源充足的 node 上，从而动态实现 node 间的 CPU 负载均衡。

对于 VM 的 vCPU 个数超过 node 中 CPU 的核数的 VM，如图 2 所示，Host NUMA 把该 VM 的内存均匀的放置在每个 node 上，vCPU 的调度范围为所有 node 的 CPU。用户绑定了 VM 的 vCPU 亲和性，Host NUMA 特性根据用户的 vCPU 亲和性设置决定 VM 的放置，若绑定在一个 node 的 CPU 上，Host NUMA 把 VM 的内存和 CPU 放置在一个 node 上，若绑定在多个 node 的 CPU 上，Host NUMA 把 VM 的内存均匀分布在多个 node 上，VM 的 vCPU 在多个 node 的 CPU 上均衡调度。

特性描述：

FUSIONCOMPUTE 虚拟化平台提供复杂的 NUMA 调度程序来动态平衡处理器负载，根据当时主机内存和 CPU 负载优先把 VM 的 CPU 和内存资源分配在同一个 node 上，并随着资源负载的动态变化对主机 node 间的 CPU 资源做负载均衡。

用户场景：

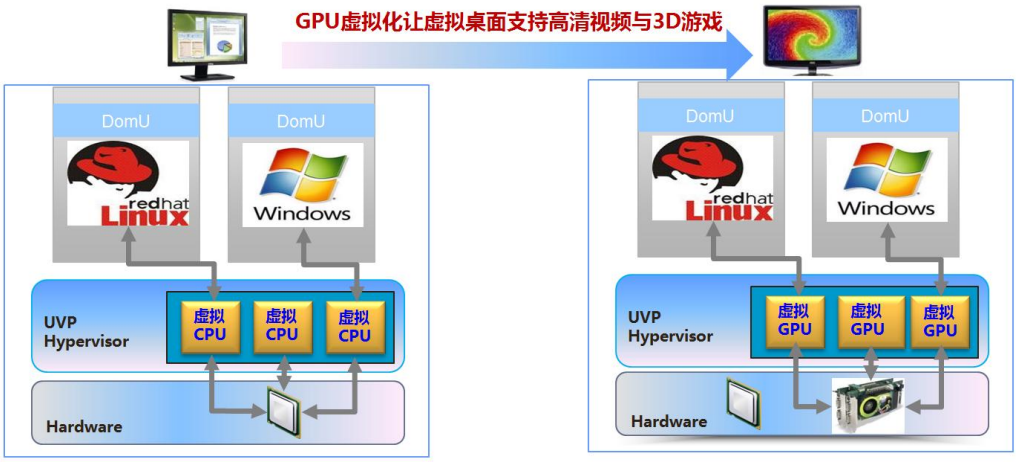
此特性为 FUSIONCOMPUTE 虚拟化平台基本特性，不需要管理员明确处理节点之间的虚拟机平衡，在各种场景都可使用。

客户价值：

Host NUMA 保证 VM 访问本地物理内存，减少了内存访问延迟，可以提升 VM 性能，性能提升的幅度与 VM 虚拟机访问内存大小和频率相关。

3.3.3 GPU 虚拟化

图3-15 GPU 虚拟化示意图



功能描述：

随着虚拟化应用的普及，用户场景也变得更加多样和广泛,特别是在桌面虚拟化应用环境，高性能专业制图和普通用户对视觉体验要求的不断提高，GPU 虚拟化已经成为加速视频和图形应用的关键技术，通过将图形处理工作从 CPU 移交给 GPU 处理能实现更高的虚拟桌面密度，动态地分配资源以满足不断变化的企业需求，保持用户所需的、完整而丰富的 GPU 加速体验。当前 FUSIONCOMPUTE 虚拟化平台支持两种 GPU 虚拟化技术实现：

1、GPU 直通技术：通过 GPU 资源被单路 VM 独享，确保性能，满足高端应用场景需求；

2、GPU 软件虚拟化技术：通过对 GPU 资源虚拟和共享，支持多路 VM 共享 GPU 加速；

通过灵活配置，可以让运行在数据中心服务器上的 GPU 处理器被一个虚拟机独占或多个虚拟机共享进行图形运算，实现 GPU 的处理能力的按需分配，提供用户的视觉体验。

技术原理：

GPU（图形处理器单元）主要进行浮点运算和并行运算，其浮点运算和并行运算速度比 CPU 更为强大，使用 GPU 虚拟化技术之后，可以让运行在数据中心服务器上的虚拟机实例共享使用同一块或多块 GPU 处理器进行图形运算。**GPU 直通技术**，通过 VT-d 技术把物理 GPU 直通给虚拟机，使虚拟机能够完全拥有物理 GPU 的资源和性能 ，但是一个物理 GPU 只能给一个虚拟机使用。**GPU 软件虚拟化技术**，采用图形命令重定向架构，在虚拟机的虚拟 GPU 驱动里截获图形命令调用，并转发到主机端，在主机端的物理 GPU 上处理图形命令，主机上对多个虚拟机的图形命令管理及渲染处理，最后把渲染好的图像传回给虚拟机，达到一个 GPU 加速多个虚拟机的目的，实现资源共享。

使用场景：

1. VDI 专业制图：在 VDI 中支持专业制图软件，满足企业客户对 VDI 产品的门槛要求；
2. VDI 办公与高清视频播放：加速办公软件和视频，提升用户体验；

用户价值：

用户体验是 VDI 的核心竞争力，通过 GPU 虚拟化加速 VDI 中的图形应用是提升用户体验的关键。

3.3.4 CPU QoS

技术原理：

Hypervisor 层根据分时复用的原理实现对虚拟 CPU 的调度，CPU Qos 的原理是定期给各虚拟 CPU 分配运行时间片，并对各虚拟 CPU 在物理 CPU 上运行的时间进行记账，对于消耗完时间片的虚拟 CPU 将被限制到物理 CPU 上运行，直到获得时间片。以此控制虚拟机获得物理计算资源的比例。以上分配时间片和记账的时间周期很短，对虚拟机用户来说会感觉一直在运行。

CPU 份额和 CPU 预留只在各虚拟机竞争计算资源的时候才发挥作用，如果没有竞争情况发生，有需求的虚拟机可以独占物理 CPU 资源。

特性描述：

FusionCompute 虚拟化平台提供 CPU QoS（Quality of Service）特性，对虚拟机可获得的物理 CPU 资源进行限制，平衡 VCPU 之间的调度。利用公平调度算法分配物理 CPU 资源，提高物理资源利用率，防止某些虚拟机占用过多物理 CPU 资源而影响其他虚拟机。支持设置虚拟机的 CPU 的 QoS 权重及绝对值上限，确保资源分配合规与可控，隔离用户行为之间的相互影响，同时保证关键业务虚拟机获得所需的计算资源，确保关键客户的用户体验。CPU Qos 功能包括三个方面的特性：上限、份额和预留。

CPU 上限 ：控制虚拟机占用物理资源的上限。以一个双核虚拟机为例，如果该虚拟机 CPU 上限为 3GHz，则该虚拟机的两个虚拟核计算能力被限制为 1.5GHz。

CPU 份额 ：CPU 份额是在多个虚拟机竞争物理 CPU 的时候按比例分配计算资源。以一个主频为 2.8GHz 的单核物理机为例，如果上面运行有三台单核的虚拟机。三个虚拟机 A，B，C，他们的份额分别为 1000，2000，4000。当三个虚拟机内部都运行满 CPU 负载的应用时。Hypervisor 会根据三个虚拟机的份额按比例分配计算资源。份额为 1000 的虚拟机 A 的计算能力约为 400MHz 的，份额为 2000 的虚拟机 B 获得的计算能力约为 800MHz，份额为 4000 的虚拟机 C 获得的计算能力约为 1600MHz。（以上举例仅为说明 CPU 份额的概念，实际应用过程中情况会更复杂）。

CPU 预留 ：CPU 预留是在多个虚拟机竞争物理 CPU 的时候最低分配的计算资源。 以上面 CPU 份额的例子进行说明，如果设置份额为 1000 的虚拟机 A 的 CPU 预留值为 700MHz，那么在虚拟机内部都运行满 CPU 负载的应用时，虚拟机 A 的计算能力约为 700MHz，这样比按份额计算出来的 400MHz 的计算能力多 300MHz，这 300MHz 会按份额比例从其他虚拟机里扣除，这样份额为 2000 的虚拟机 B 获得的计算能力约为 700MHz，份额为 4000 的虚拟机 C 获得的计算能力约为 1400MHz。（以上举例仅为说明 CPU 预留的概念，实际应用过程中情况会更复杂）。

用户场景：

1. 对于类似亚马逊的租户场景，管理员可以通过控制虚拟机 CPU 上限实现不同等级的付费用户的 SLA 等级。
2. 对于桌面虚拟机与编译虚拟机混合部署的场景，管理员可以对不同类型的虚拟机设定不同的 CPU 份额。

客户价值：

CPU Qos 功能为客户提供了控制虚拟机计算能力的手段：

1. CPU 上限可以供客户实现虚拟机资源隔离的功能，通过控制某个虚拟机的 CPU 上限，可以有效避免该虚拟机负载过大时影响同一服务器上的其他虚拟机的性能。
2. CPU 份额可以控制一个服务器上的各个虚拟机在竞争 CPU 资源时的资源占用率，可以针对不同等级的用户提供 SLA 服务。
3. CPU 预留可以控制一个服务器上的各个虚拟机在竞争 CPU 资源时的占有资源的最低值，在竞争不充分时又可以将资源按需分配给虚拟机使用，既达到了资源的高效复用，又实现了在资源竞争情况下根据虚拟机优先级分配计算资源。

3.3.5 Guest NUMA

技术原理：

1）Hypervisor 层为虚拟机模拟 vCPU NUMA 拓扑结构，使虚拟机能够根据 vCPU 配置，将内存分布为不同的 vNode 节点上，在虚拟机内部展示 vCPU 与 内存对应关系。

2）Hypervisor 层能够根据 Guest Numa 配置，解决 CPU 资源分配不平衡引起的 VM 性能瓶颈问题，当启动 VM 时，Guest NUMA 根据当时主机内存和 CPU 负载，选择负载较轻的 node 放置虚拟机的 vNode，使虚拟机同一 vNode 的 CPU 和内存资源分配在同一个物理 node 上。

客户价值：

FusionCompute 虚拟化平台对向客户机操作系统公开虚拟 NUMA 拓扑的支持，这样便于客户机操作系统和应用程序 NUMA 优化，从而可提高性能。

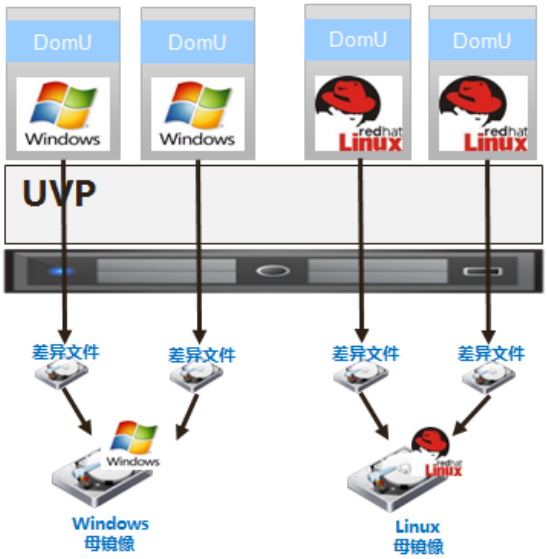
3.4 存储虚拟化

云计算带来的另一个挑战就是解决大规模虚拟机部署所面临的存储方面的问题，首先是存储 I/O 性能瓶颈问题，因为存储性能增长速度相比于计算能力的增长要慢，因此对于虚拟化而言，I/O 瓶颈和缓慢的存储性能成为主要瓶颈，FusionCompute 虚拟化平台通过提供不同 I/O 性能优化手段，有效缓解存储瓶颈。其次是存储利用率低的问题，FusionCompute 虚拟化平台提供存储瘦分配技术，提高存储利用率降低存储硬件的采购成本。再次是大规模部署虚拟机的效率问题，FusionCompute 虚拟化平台提供链接克隆技术，可以缩短规模部署虚拟化的时间。

3.4.1 链接克隆

FusionCompute 虚拟化平台提供的虚拟机链接克隆技术就是根据一个源虚拟机克隆出一个或多个克隆虚拟机，且克隆虚拟机拥有与源虚拟机完全相同的操作系统、应用系统乃至数据和文档。克隆技术可分为完整克隆和链接克隆两种。完整克隆方式下，克隆虚拟机和源虚拟机是两个完全独立的实体，源虚拟机的修改乃至删除不会影响到克隆虚拟机

的运行，但缺点是 2 个虚拟机运行时需要占用 2 份内存和 2 份磁盘空间;与之相对应的是链接克隆方式，克隆虚拟机必须在源虚拟机存在的情况下才能运行，但优点是多个克隆虚拟机之间的公共部分(共同来自源虚拟机的部分)可以共用同一份内存空间和同一份磁盘空间，因此在服务器主机资源相同的情况下，采用链接克隆的方式可以支持更多的虚拟机，运行更多的业务，或者运行更多的虚拟桌面，从而使企业的 IT 成本更低。



（图示： 链接克隆原理图）

如图所示，链接克隆技术使相同 Guest OS 多个虚拟机之间共享母镜像，不再使用一对一的镜像存储机制，而一对一保存虚拟化镜像差异化部分，对应虚拟机都是链接到源虚拟机的副本虚拟机上，用户数据盘只创建差异文件,如果改变主虚拟机,变化也会被复制到每一个与之相连接的克隆虚拟机中，取代传统的使用模板进程来创建虚拟机的方式，在节省存储空间降低存储成本的同时，实现虚拟机的快速部署，节省大规模系统部署的时间。

客户价值：

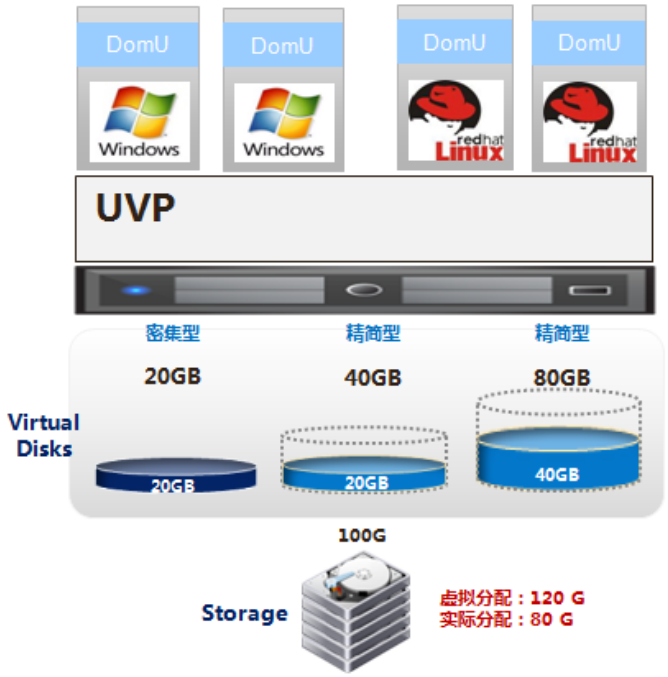
可节省存储空间；

可减少大规模部署虚拟机的时间。

3.4.2 存储瘦分配

在传统的存储系统中，当某项应用需要一部分存储空间的时候，往往是预先从后端存储系统中划分出一部分足够大的空间预先分配给该项应用，即使这项应用暂时不需要使用这么大的存储空间，但由于这部分存储空间已经被预留了出来，其它应用程序无法利用这些已经部署但闲置的存储容量。这种分配模式一方面使闲置的存储数量不断增加，系统总体拥有成本升高；另一方面用户不得不购买更大的存储容量，才能适应环境，成本进一步加大。解决存储过量供给的最有效的方式，是通过使应用程序只消耗必要的存储资源来将块或块组写入特定卷，优化存储利用，不必再购买或维持超过实际所需的存储。

图3-16 存储瘦分配原理图



如图所示，FusionCompute 虚拟化平台存储瘦分配技术的核心原理是“欺骗”操作系统，让操作系统认为存储设备中有很大的存储空间，而实际上的物理存储空间则没有那么大，虚拟机始终可以看到完整的逻辑磁盘大小，实际物理磁盘仅占用正在使用的物理磁盘空间。通过使用该技术，可减少对虚拟磁盘的过度调配，降低存储成本。

客户价值：

1. 减少对虚拟磁盘的过度调配，实现存储资源的按需分配，通过降低或延迟存储设备采购的方式节省硬件投资成本；

3.4.3 存储 QOS

FusionCompute 虚拟化平台提供存储 QoS（Quality of Service）特性，对远程存储设备（IPSAN、FCSAN 和 VBS）挂载到 Dom0 上的卷进行 IO 上限控制，意味着虚拟机存储后端对该卷的任何 IO 操作都无法超过设置的上限值。存储卷 IOPS 上限设置涉及卷挂载、卷卸载、卷审计场景，用户需要在挂卷的同时设置对应的 IOPS 上限，从而保证该卷在分配给虚拟机使用后 IO 能力得到控制。

客户价值：

通过对虚拟机磁盘的 IO 上限控制，控制虚拟机获得存储 IO 的能力，使得 IO 密集型虚拟机或存储 IO 异常的虚拟机不会影响其他虚拟机的用户体验。

3.4.4 存储在线扩容

FusionCompute 虚拟化平台提供存储在线扩容特性，使一个数据存储可以管理多个物理存储设备空间，通过添加另外的物理存储设备至数据存储或者对物理存储设备进行扩容再扩容数据存储，从而实现对数据存储灵活地空间扩容，有效提高数据存储的扩展性。

客户价值：

1. 将几个具有相同作用的物理存储设备添加至同一个数据存储，管理更方便，视图更清晰。
2. 当一个数据存储空间不足时，可以将其它物理存储设备添加至数据存储，达到扩容数据存储的目的。

3.4.5 PVSCSI 支持

传统的部署 Oracle RAC 和 MSCS 群集业务的方式是直接在物理服务器进行部署，需要使用大量的物理服务器，成本较高。

FusionCompute 虚拟化平台提供 PVSCSI 特性，通过使用 PVSCSI，可以让虚拟机识别 scsi 磁盘，实现在虚拟机内部下发 scsi 命令、交给主机然后透传给存储设备进行处理、最后将应答返回，能够很好地支撑 Oracle RAC 和 MSCS 群集业务在虚拟机上的正常运行。

客户价值：

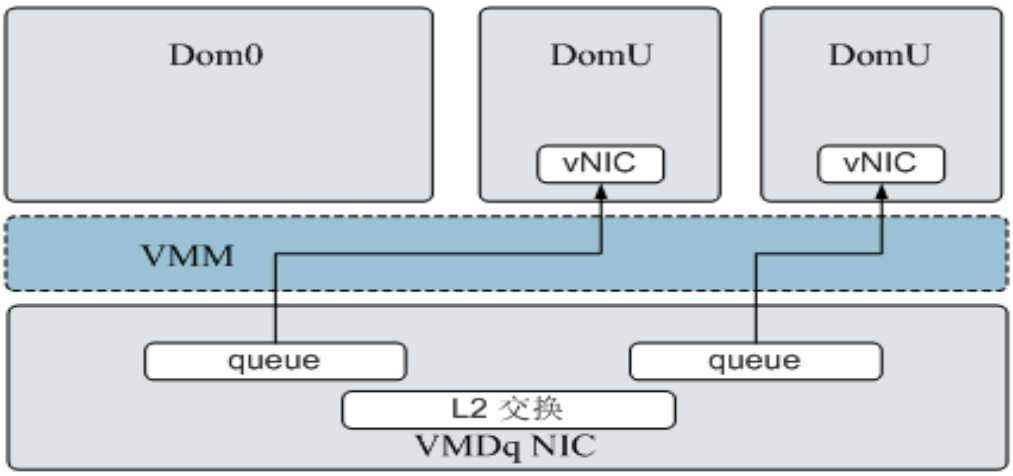
通过对 PVSCSI 特性的支持，可以支撑虚拟化环境下 Oracle RAC 和 MSCS 群集业务在虚拟机上的正常运行，降低了大量使用物理服务器的成本。

3.5 网络虚拟化

虚拟化环境下服务器整合了更多的应用服务，工作负载更加依赖于网络 I/O，同时随着处理器多核技术的发展，需要充分提高资源利用率，而当前外部 IO 性能已经跟不上处理器等的发展，需要在系统性能与网络能力之间达到一种平衡，从整合中实现最理想的应用服务。在 IO 设备上，频繁的 VMM 切换以及对中断的处理是导致虚拟化效率低下的两个重点因素，FusionCompute 虚拟化平台结合硬件辅助虚拟化技术提供 SR-IOV 网卡直通等技术，确保企业关键应用的性能体验。

3.5.1 VMDQ

图3-17 VMDQ 特性描述



特性描述：

在虚拟环境中，hypervisor 管理网络 I/O 活动，随着平台中的虚拟机和传输量增加，hypervisor 要求更多的 CPU 周期以分类数据包，并将其路由到适合的虚拟机中，减少对应用可用的 CPU 空间。FUSIONCOMPUTE 虚拟化平台利用 VMDq（Virtual Machine Device Queues 虚拟机设备队列）技术，针对对虚拟机网络性能有极高要求的场景，在支持 VMDq 的网卡上，用硬件实现了一个 Layer 2 分类/排序器，根据 MAC 地址和 VLAN 信息将数据包发送到指定的网卡队列中去，这样虚拟机收发包时就不需要 Dom0 的参与，这种模式极大地提升了虚拟化网络效率。

技术原理：

图3-18 Intel VMDQ 技术原理

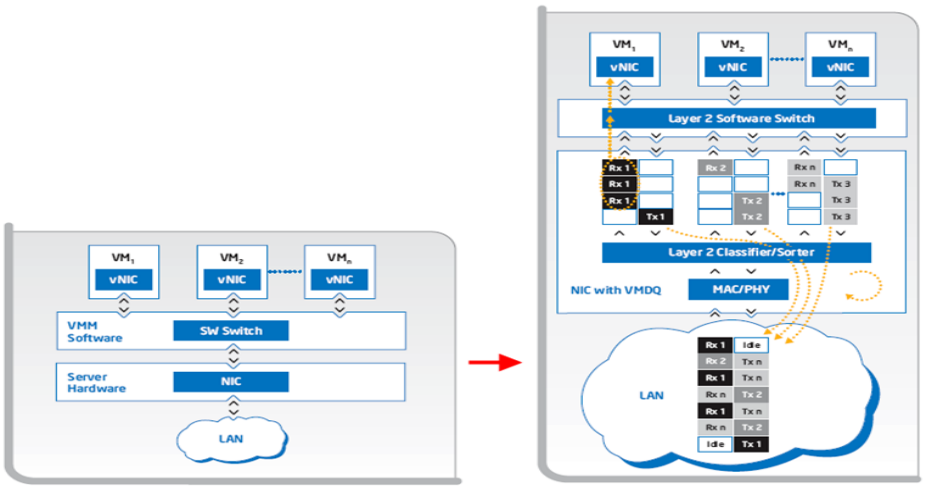


Figure 1. In virtual environments today, the hypervisor manages network I/O.

Figure 2. VMDq offloads network I/O management to the network silicon.

Intel VMDq（Virtual Machine Device Queue 虚拟机设备队列）技术，是专门用于提升网卡的虚拟化 IO 性能的硬件辅助 I/O 虚拟化技术，主要解决 IO 设备上频繁的 VMM 切换以及对中断的处理是导致虚拟化效率低下的问题，可以减轻 hypervisor 负担、同时提高虚拟化平台网络 I/O 性能。

VMDq 技术可以将网络 I/O 管理负担从 hypervisor 上卸载掉，多个队列和芯片中的分类智能支持虚拟环境中增强的网络传输流，从应用任务中释放处理器周期，提高向虚拟机的数据处理效率及整体系统性能。

用户场景：

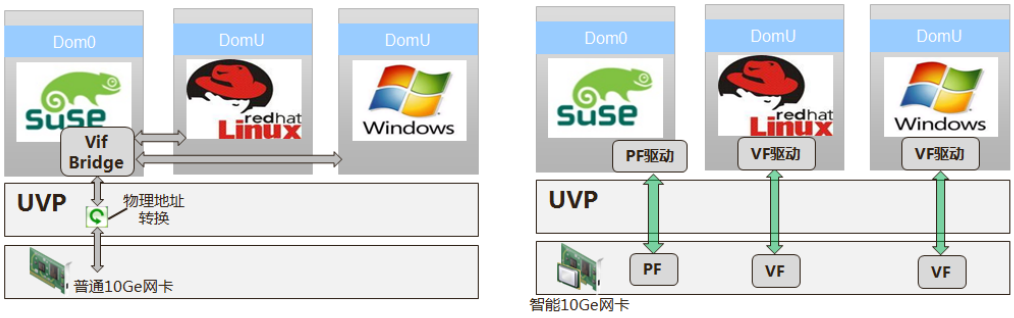
应用于装载了华为智能网卡的硬件环境，提供高效的虚拟机网络通信性能。

客户价值：

为虚拟机提供接近物理机的网络通信性能。兼容部分虚拟化高级特性，比如在线迁移，虚拟机快照等。

3.5.2 SR-IOV

图3-19 普通网卡和基于 FUSIONCOMPUTE 虚拟化平台 SR-IOV 方案比较



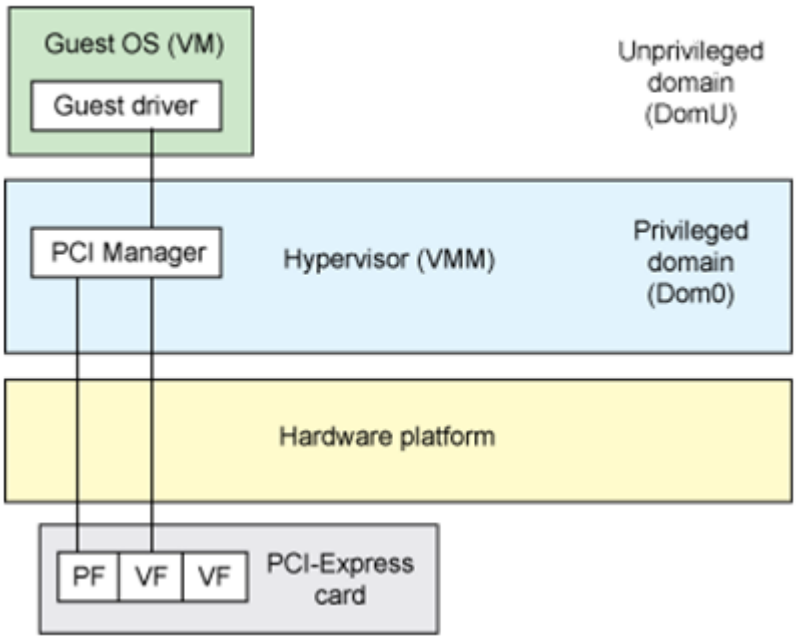
	普通网卡虚拟化方案	UVP SR-IOV智能网卡方案
差异点	虚拟网桥队列Vif Bridge转发数据	每个虚拟网卡拥有单独队列，直通物理网卡
	需要Hypervisor完成地址转换	Hypervisor无需进行地址转换，减少计算损耗
竞争力	Dom0，DomU VCPU调度时延成为瓶颈 无法满足高网络IO应用要求	满足高网络IO应用要求，无损虚拟化特性(热迁移、内存复用、虚拟机网络管控等），技术领先业界
	吞吐量：3.6 Gbit/s	吞吐量：9.5 Gbit/s

功能描述：

通常针对虚拟化服务器的技术是通过软件模拟共享和虚拟化网络适配器的一个物理端口，以满足虚拟机的 I/O 需求，模拟软件的多个层为虚拟机作了 I/O 决策，因此导致环境中出现瓶颈并影响 I/O 性能。FUSIONCOMPUTE 虚拟化平台提供的 SR-IOV 是一种不需要软件模拟就可以共享 I/O 设备 I/O 端口的物理功能的方法，主要利用 iNIC 实现网桥卸载虚拟网卡，允许将物理网络适配器的 SR-IOV 虚拟功能直接分配给虚拟机，可以提高网络吞吐量，并缩短网络延迟，同时减少处理网络流量所需的主机 CPU 开销。

技术原理：

图3-20 通过 SR-IOV 实现透传



SR-IOV（Single Root I/O Virtualization）是 PCI-SIG 推出的一项标准，是虚拟通道（在物理网卡上对上层软件系统虚拟出多个物理通道，每个通道具备独立的 I/O 功能）的一个技术实现，用于将一个 PCIe 设备虚拟成多个 PCIe 设备，每个虚拟 PCIe 设备如同物理 PCIe 设备一样向上层软件提供服务。通过 SR-IOV 一个 PCIe 设备不仅可以导出多个 PCI 物理功能，还可以导出共享该 I/O 设备上的资源的一组虚拟功能，每个虚拟功能都可以被直接分配到一个虚拟机，能够让网络传输绕过软件模拟层，直接分配到虚拟机，实现了将 PCI 功能分配到多个虚拟接口以在虚拟化环境中共享一个 PCI 设备的目的，并且降低了软加模拟层中的 I/O 开销，因此实现了接近本机的性能。如图所示，在这个模型中，不需要任何透传，因为虚拟化在终端设备上发生，允许管理程序简单地将虚拟功能映射到 VM 上以实现本机设备性能和隔离安全。SR-IOV 虚拟出的通道分为两个类型：

1、PF(Physical Function) 是完整的 PCIe 设备，包含了全面的管理、配置功能， Hypervisor 通过 PF 来管理和配置网卡的所有 I/O 资源。

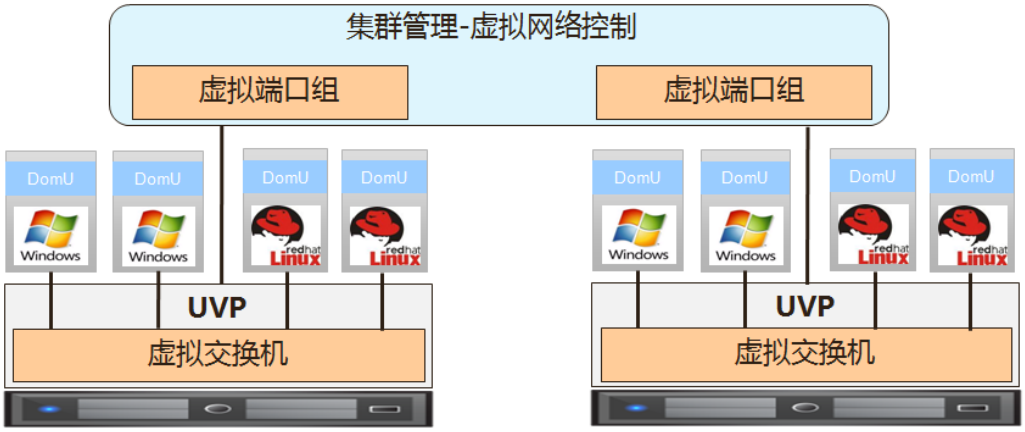
2、VF(Virtual Funciton)是一个简化的 PCIe 设备，仅仅包含了 I/O 功能，通过 PF 衍生而来好象物理网卡硬件资源的一个切片，对于 Hypervisor 来说，这个 VF 同一块普通的 PCIe 网卡一模一样。

客户价值：

可满足高网络 IO 应用要求，无需特别安装驱动，且无损热迁移、内存复用、虚拟机网络管控等虚拟化特性。

3.5.3 分布式交换机

图3-21 分布式虚拟交换机



FUSIONCOMPUTE 虚拟化平台提供的分布式虚拟交换机就是把分布在集群中多台主机的单一交换机逻辑上组成一个大的集中式交换机，减少每台虚拟交换机需要单独分别配置过程，同时为集群级别的网络连接提供一个集中控制点，使虚拟环境中的网络配置不再以主机为单位，简化虚拟机网络连接的部署、管理和监控，适合于大规模的网络部署。

客户价值：

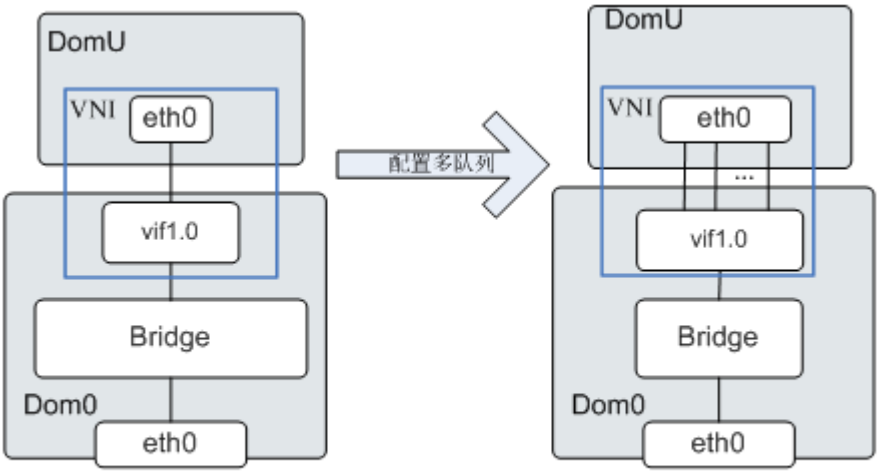
通过提供丰富的网络配置管理功能，端口动态绑定，静态绑定，IP 接入控制、虚拟机网络 Qos，实现网络资源统一管理，实时化网络监控。

3.5.4 VNI 网卡

FusionCompute 虚拟化平台提供 VNI（Virtual Network Interface，虚拟网络接口）特性，基于 virtio_net 的一套前后端网络方案，旨在为虚拟化环境提供一套统一通用的虚拟化网络接口，为用户提供稳定、高性能的前后端网络能力。

由于虚拟机的虚拟网卡只能使用一个收发包队列对外通信，仅能利用单个 CPU 的处理能力，无法满足高网络带宽需要求。VNI（Virtual Network Interface）提供对虚拟网卡配置多队列的能力（见下图），借助多个 CPU 资源带来网络带宽的大幅度提升。

图3-22 VNI（Virtual Network Interface）虚拟网卡多队列配置



eth0：DomU中代表虚拟网卡/Dom0中代表物理网卡
Bridge：Dom0中的弹性虚拟交换机（EVS）

客户价值：

大幅提升虚拟网络带宽，为虚拟机提供高宽能力。

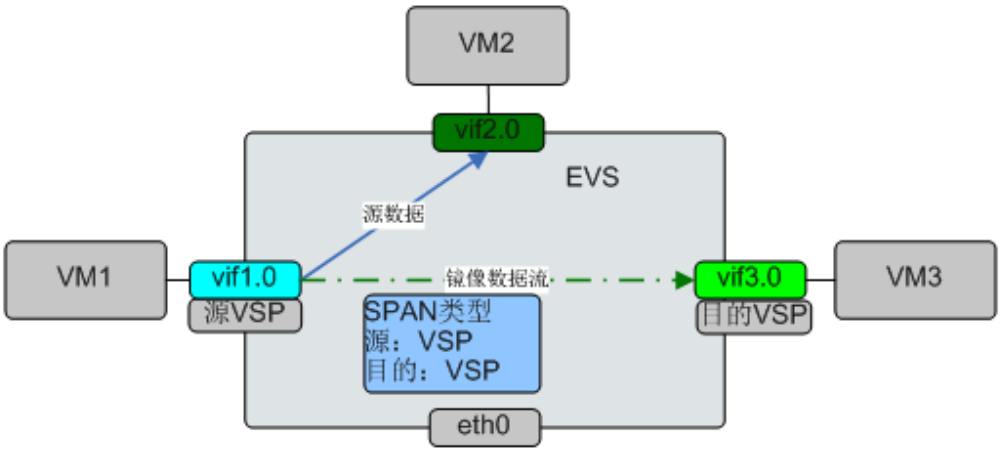
3.5.5 端口镜像

FusionCompute 虚拟化平台提供端口镜像功能，管理员通过配置端口镜像，将虚拟交换机上受控端口数据镜像一份给目的端口，通过目的端口上配置的流量分析仪，可实现对网络数据的监控。

目前支持 3 种端口镜像类型：SPAN、RSPAN、ERSPAN。

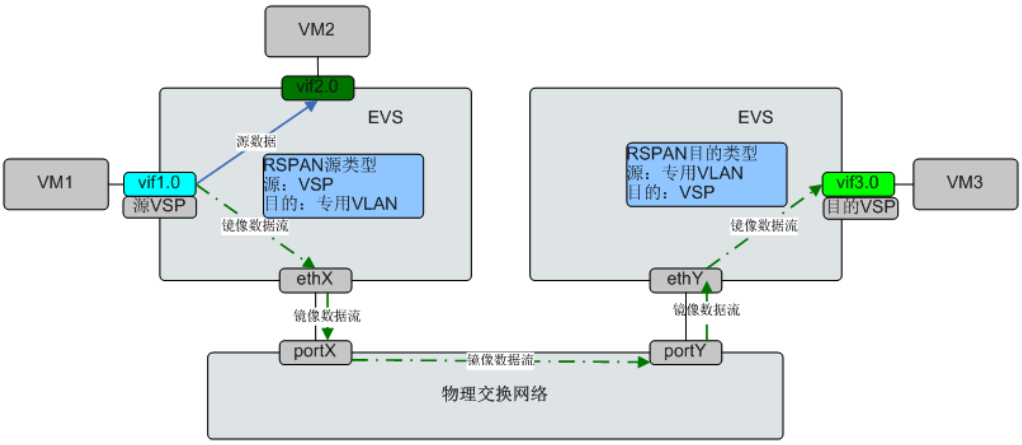
- SPAN(本地端口镜像)，端口镜像的源 VSP 和目的 VSP 在同一个 EVS 内（如下图所示）。

图3-23 SPAN 类型端口镜像



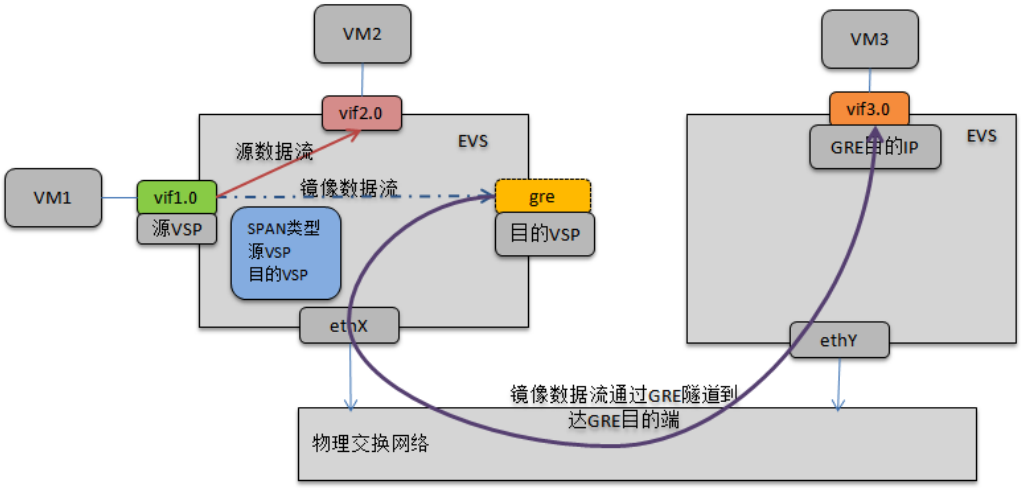
- RSPAN（远程端口镜像），端口镜像的源 VSP 和目的 VSP 不在同一个 EVS 内，镜像数据在专用 VLAN 域内进行广播。源 VSP 所在的 EVS 配置的端口镜像为 RSPAN 源类型，目的 VSP 所在的 EVS 中配置的端口镜像类型为 RSPAN 目的类型（如下图所示）。

图3-24 RSPAN 类型端口镜像



- ERSPAN（增强型远程端口镜像），实质上是一种 SPAN 类型的端口镜像，其目的 VSP 必须是 GRE 端口，镜像到 GRE 端口的数据流通过 GRE 隧道传输到 GRE 隧道的目的端口，在 GRE 隧道的目的端口则可获取到源 VSP 的数据流（如下图所示）。

图3-25 ERSPAN 类型端口镜像

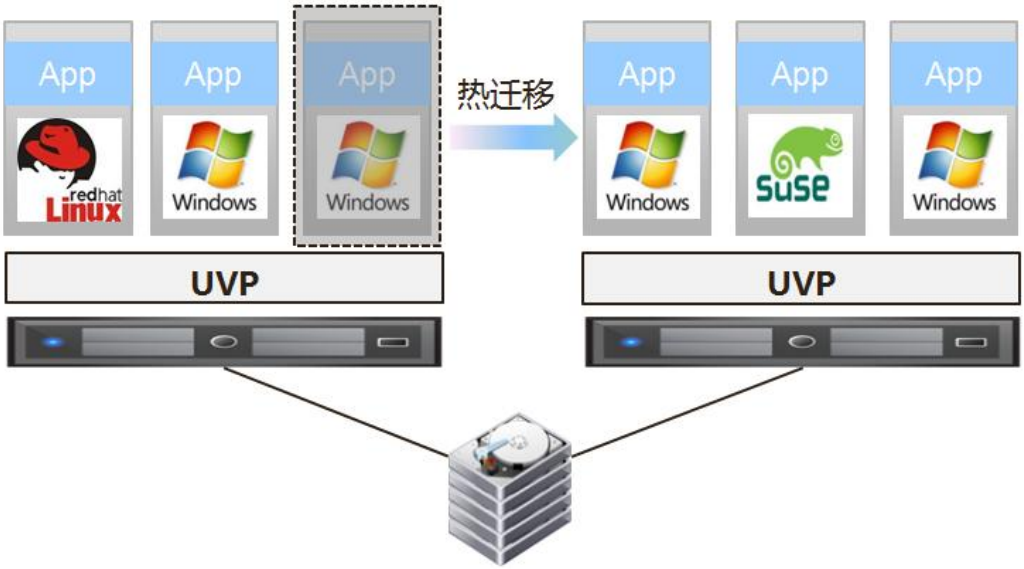


3.6 高可用性

虚拟化环境中，物理服务器和存储上承载更多的业务和数据，设备故障时造成的影响更大。FUSIONCOMPUTE 虚拟化平台提供虚拟机热迁移和虚拟机热备份技术，降低宕机带来的风险、减少业务中断的时间。

3.6.1 虚拟机热迁移

图3-26 虚拟机热迁移



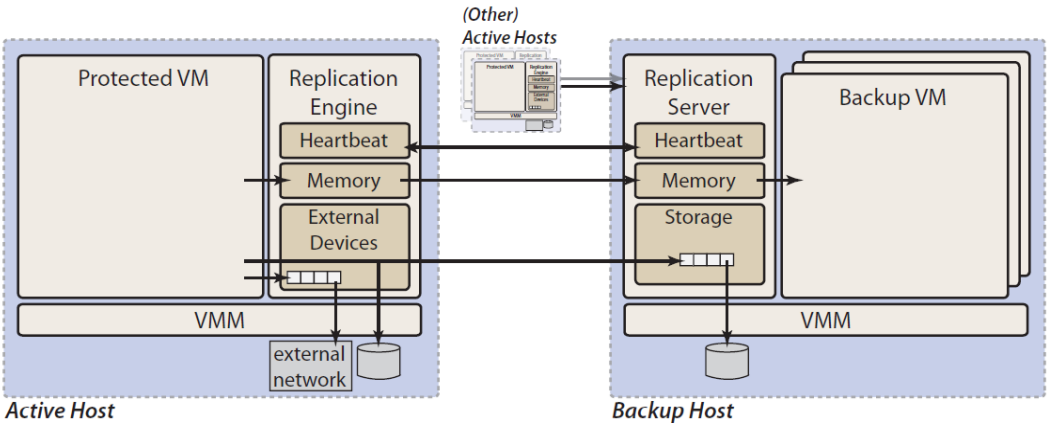
FUSIONCOMPUTE 虚拟化平台热迁移技术是指把一个虚拟机从一台物理服务器迁移到另一台物理服务器上，即虚拟机保存/恢复(Save/Restore)。首先将整个虚拟机的运行状态完整保存下来，同时可以快速的恢复到目标硬件平台上，恢复以后虚拟机仍旧平滑运行，用户不会察觉到任何差异。虚拟机的热迁移技术主要被用于双机容错、负载均衡和节能降耗等应用场景。FUSIONCOMPUTE 虚拟化平台热迁移提供内存压缩技术，使热迁移效率提升一倍，可支持并发 8 台虚拟机同时迁移。

场景价值：

1. 在设备维护过程中，通过热迁移手动将应用迁移至另一台服务器，维护结束后再迁回来，中间应用不停机，减少计划内宕机时间。
2. 可结合资源动态调度策略，例如在夜晚虚拟机负荷减少时，通过预先配置自动将虚拟机迁移集中至部分服务器，减少服务器的运行数量，从而降低设备运营能耗上的支出。

3.6.2 虚拟机热备份

图3-27 FT 技术原理



技术原理：

在虚拟环境下设置主备虚拟机，在备节点上创建主虚拟机的完整拷贝，包括 CPU 状态、内存、磁盘操作、QEMU 等都进行低延迟的定时同步。备节点虚拟机处于 P 状态定时检测主节点虚拟机心跳，在指定时间内收不到心跳即认为异常发生，备虚拟机切换到正常运行状态。优势是主备节点可以保持状态完全同步，数据完全一致，缺点是会带来一些性能开销。

特性描述：

FUSIONCOMPUTE 虚拟化平台支持虚拟机热备份（FT）技术，可支持多次硬件故障实时切换，当故障发生主备虚拟机倒换后，自动再次形成主备状态，并可支持多核虚拟机。

客户价值：

提供关键业务虚拟机实时备份机制，能够在硬件发生问题时，做到零停机（业务不中断）和零数据丢失，最大限度保护用户利益。

3.6.3 虚拟资源热插

虚拟资源热插功能作为在虚拟化场景下的一个高级特性，在不影响用户业务的情况下，支持动态在线给虚拟机增加或减少 VCPU 等虚拟资源数量，来实现虚拟机计算资源的动态分配。当前支持 VCPU、虚拟内存、虚拟网卡和虚拟磁盘的热插功能。（注：具体支持须由虚拟化平台和 Guest OS 两者同时支持才能生效）

用户场景

在虚拟化环境中，虚拟机管理程序会监控虚拟机上的资源压力情况，当出现计算资源 CPU 压力达到上限阈值时，通过调用 VCPU 热插拔来增加一个 VCPU 个数，从而缓解系统计算资源的压力。

客户价值

虚拟资源热插功能提供给客户计算资源的按需分配，当客户业务压力增加时可以获得更多的计算资源，而到业务压力减少时可以释放计算资源，从而把多于的计算资源让给高业务的用户，以提升计算资源的利用率。

3.6.4 虚拟机内存快照

特性描述：

FUSIONCOMPUTE 虚拟化平台提供虚拟机内存快照，即在虚拟机的运行状态下，不中断用户的业务，实现虚拟机内存状态的备份，同时，通过该备份文件，可以方便地恢复虚拟机，保证恢复后的虚拟机状态与快照点完全一致，包括：打开了哪些应用程序、窗口，编写一半的文档等，都能如实还原。

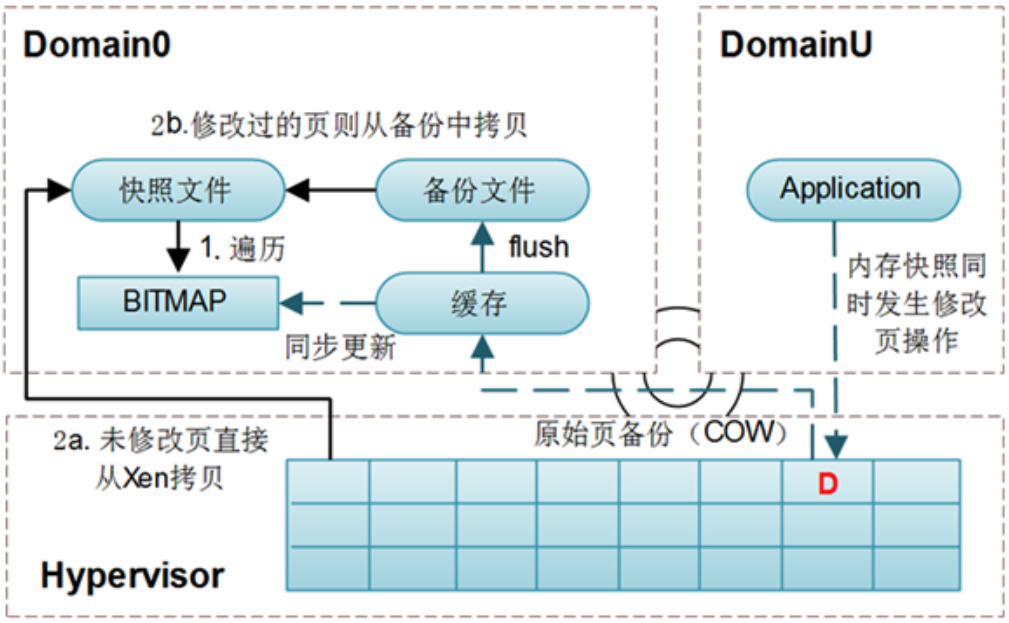
虚拟机内存快照必须配合存储快照一起使用，保证数据一致性，主要用户场景为：创建快照和还原快照。用户通过创建快照可以实时保存虚拟机的状态，包括虚拟机的内存、磁盘等数据信息。创建快照后，会生成一个内存快照文件和存储快照卷。当虚拟机发生故障，或者用户想恢复到之前做快照时刻的虚拟机状态，可以选择相应的内存快照文件和存储快照卷来进行快照还原。快照还原后，虚拟机恢复到快照时间点的状态。

内存快照分为两个主要流程：

1. 流程为顺序备份虚拟机内存，按照虚拟机内存块数据如实保存当前虚拟机的内存状态。
2. 另一个流程则是监控虚拟机内存变化，在变化写入之前把原始数据拷贝出来进行备份。

最终通过把两个流程的数据合并到一起，实现快照点内存数据的准备保存，从而在恢复快照时能够如实还原虚拟机的内存状态。

图3-28 内存快照流程图



使用场景：

1. 内存快照较常用于用户需要快速恢复虚拟机状态的场景（业务启动、加载过程复杂繁琐）；
2. 也用于用户在某些特定场景下希望保留现场使用（比如某长时间运行的用例执行到一半，需要视情况进行分支选择时）；
3. 以及快速批量部署的情况（通过快照的内存数据快速批量创建和启动虚拟机）。

客户价值：

虚拟机内存快照必须配合存储快照一起使用，共同保证数据一致性。

约束：

虚拟机 64G 内存，4 块网卡以内；或 1G 以下内存，12 块网卡，Dom0 默认配置（2U3G）。

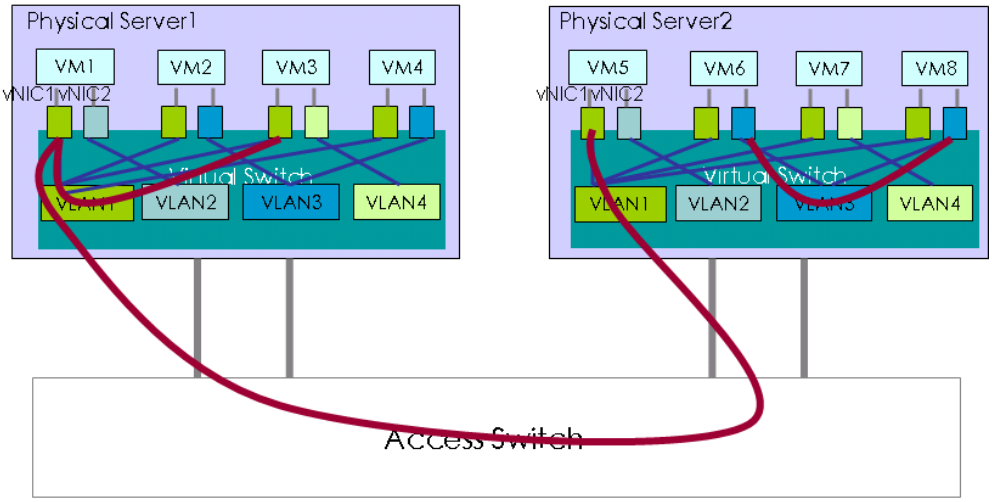
虚拟机 64G 内存，5 到 10 块网卡，Dom0 配置 6U3G。

虚拟机 64G 内存，11 或 12 块网卡，建议不要在 IO 繁忙时进行虚拟机内存快照。

3.7 高安全性

3.7.1 VLAN

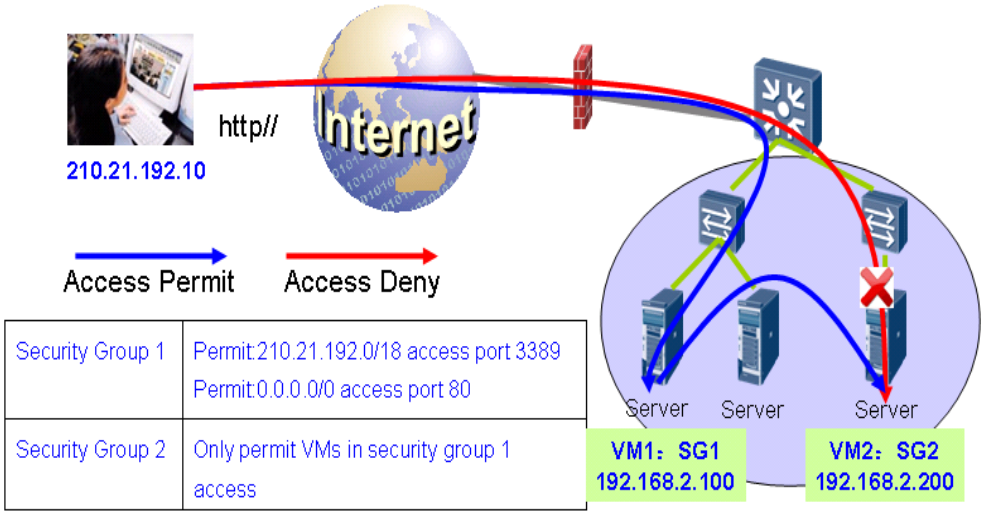
图3-29 VLAN 组网图



FUSIONCOMPUTE 虚拟化平台通过虚拟网桥（支持 VLAN 功能）实现虚拟交换功能，实现 VLAN 隔离，确保虚拟机之间的安全隔离。如图所示，处于不同物理服务器上的虚拟机通过 VLAN 技术可以划分在同一个局域网内，同一个服务器上的同一个 VLAN 内的虚拟机之间通过虚拟交换机进行通信，而不同服务器上的同一 VLAN 内的虚拟机之间通过交换机进行通信，确保不同局域网的虚拟机之间的网络是隔离的，不能进行数据交换。

3.7.2 安全组

图3-30 安全组示意图



Security Group 1	Permit:210.21.192.0/18 access port 3389 Permit:0.0.0.0/0 access port 80
Security Group 2	Only permit VMs in security group 1 access

FUSIONCOMPUTE 虚拟化平台提供的虚拟机安全组是一组虚拟机的集合，也是关于这组虚拟机的网络安全规则的集合。同一个虚拟机安全组中的虚拟机可能分布在多个物理位置分散的物理机上，一个安全组内的虚拟机之间是可以相互通信，而不同的安全组之间的虚拟机是不允许进行通信的。因此虚拟机安全组的作用是在一个物理网络中，划分出相互隔离的逻辑虚拟局域网，提高网络安全性。

3.7.3 安全加固

FUSIONCOMPUTE 虚拟化平台选取业界通用安全扫描工具 CIS-CAT 对 Linux 系统的安全建议，在不影响正常业务的前提下实施了对系统的安全加固措施。从文件权限、账户控制、密码保护等方面提升系统安全性，有效地保护系统免离非法操作和恶意攻击，为用户提供了更为安全的系统环境。

3.7.4 防火墙

FUSIONCOMPUTE 虚拟化平台集成 SuSEfirewall 组件，能提供全方位的网络安全保护。用户可以根据自己的业务状况，配置不同的规则策略，按需阻断网络攻击，提升产品整体网络安全能力。

3.8 可管理性

3.8.1 Kbox 黑匣子

黑匣子作为一个内核模块存在于 FUSIONCOMPUTE 虚拟化平台 Dom0 的内核，其主要功能是在 Dom0 或 Xen 发生异常时候能捕获异常信息，并且把这些异常信息通过网络输出、本地存储，或直接输出到系统日志中，为系统异常定位提供强有力支撑。

客户价值

黑匣子可获取异常时的上下文，如：系统死机、异常重启等。通过这些信息有助于管理员迅速定位导致系统异常的原因，解决系统异常等严重问题，提高系统的稳定性。

3.8.2 一键式收集工具

一键式收集工具，用户可以在 FUSIONCOMPUTE 虚拟化平台出现问题时利用 FusionCompute 虚拟化平台 log 命令收集系统的日志信息，发给 FUSIONCOMPUTE 虚拟化平台开发人员协助问题定位。用户也可以收集特定的日志信息。

收集的信息主要包括如下内容：

1. 收集系统运行环境信息
2. 收集系统日志
3. 收集虚拟化相关信息

客户价值

帮助用户保留问题现场，高效的抓取问题定位信息，为后续的问题定位提供保障。可以收集特定的信息，提高可服务性。

3.8.3 GuestOS 故障检测

FUSIONCOMPUTE 虚拟化平台提供 Guest OS 故障检测功能，当客户机发生严重故障时（例如 Windows 系统蓝屏），虚拟机管理程序会监控到客户机故障。虚拟机管理程序可以重启或关闭客户机，从而避免有故障的客户机持续占用计算资源。

客户价值

主动上报客户机的故障状态，虚拟机管理程序可以及时重启或关闭客户机的运行，避免计算资源空转浪费，提升 RAS 特性。

3.8.4 系统故障告警

FUSIONCOMPUTE 虚拟化平台提供基本的系统故障告警能力，能定期扫描预设的告警项。调用外部接口进行上报并且记录日志，同时对告警项进行跟踪，对已经恢复的告警项进行清理。

目前 FUSIONCOMPUTE 虚拟化平台预设的告警项分为三类：

1. 虚拟机运行类：主要对虚拟机运行过程中异常状况进行扫描并告警；
2. 虚拟机管理类：主要对 FUSIONCOMPUTE 虚拟化平台虚拟机管理相关的模块进行扫描并告警；
3. 资源类：主要对内存、网络、存储等外部资源进行监控并告警。

客户价值

FUSIONCOMPUTE 虚拟化平台主动上报系统故障，及时向用户传递系统警告信息以便及时暴露问题。用户可以结合业务模块对告警信息进行汇总和维护，方便对故障进行及时的修复、跟踪和记录。

3.9 可节能性

3.9.1 CPU 节能管理

FUSIONCOMPUTE 虚拟化平台提供了休眠唤醒的功能来进行节能降耗。当判断虚拟机无人使用时，主动休眠虚拟机可以节省大量的计算资源给其他需要运行的虚拟机使用。由于休眠是将当前的业务保留起来，当唤醒时，当前业务得以继续运行而不需要重新启动，节省了客户的时间。

用户场景：

当用户下班后（夜晚）不需要操作客户机时，虚拟机管理程序休眠客户机以节省计算资源。当用户上班前（凌晨），虚拟机管理程序唤醒客户机。

客户价值：

主动休眠虚拟机可以节省大量的计算资源给其他需要运行的虚拟机使用，提高节能降耗效果。由于休眠是将当前的业务保留起来，当唤醒时，当前业务得以继续运行而不需要重新启动，节省了客户的时间。

4 平台应用/Experience

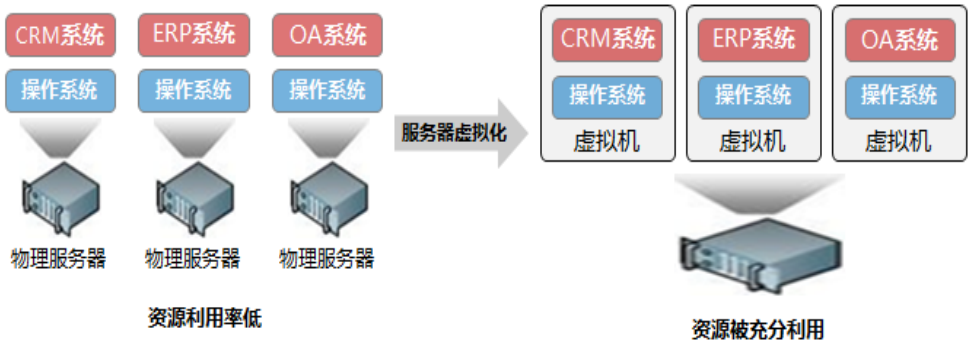
4.1 应用概述：

华为的云计算注重业务和应用的云化，强调为客户带来实际的价值，作为云计算平台的核心技术平台，FUSIONCOMPUTE 虚拟化平台重点提供服务器整合、企业虚拟桌面和互联网数据中心三大应用解决方案所需的虚拟化技术能力，围绕用户的核心诉求，充分发挥虚拟化的实际使用价值。

4.2 应用一：服务器整合

场景技术原理：

图4-1 服务器整合前与整合后



服务器整合（server consolidation）是一种典型的集中式部署，可以细化为地理整合、物理整合、数据整合和应用整合，简单地说，就是把分散的计算能力集中到一个或几个大的计算中心中，其基本特点是削减特定类型服务器的数量，达到以少胜多的目的。其利用的核心技术就是服务器虚拟化技术，通过将服务器物理资源抽象成逻辑资源，让一台服务器变成几台甚至上百台相互隔离的虚拟服务器，让 CPU、内存、磁盘、I/O 等硬件变成可以动态管理的“资源池”，从而提高资源的利用率，简化系统管理，实现服务器整合，让 IT 对业务的变化更具适应力。服务器整合带来的潜在 ROI（Return On Investment 投资回报率）优势，如下：

1. 降低系统有形成本：减少服务器、软件许可、服务器空间，降低能耗，从而降低 IT 总成本；
2. 改善系统可管理性：IT 标准化有助于更好地管理服务器环境，确保高可用性和高安全性；
3. 提高服务水平和运行效率：提高产品的可靠性和性能，使维护人员专注于更高价值的任务；

- 核心业务诉求：

图4-2 服务器整合场景诉求

a） 提升资源利用率，减少硬件投资成本和运营支出

典型问题：服务器数量不断增加，服务器资源利用率停留在 5%~25%左右，低使用率的服务器长期运行，提高总体能耗。

b） 提高可靠性，减少为保障业务连续性和提供系统故障容灾所需的成本

典型问题：采用单机时，硬件故障、软件故障、系统单点故障、自然灾害，甚至计划维护所导致的停机时间，都有可能影响到业务，采用双机时，成本较高。

c） 提升运维效率和业务体验，降低系统运维成本，提高业务支撑响应能力

典型问题：服务器、存储、网络各自分层管理，无统一管理，硬件故障后，硬件更换和业务恢复时间较长，各部件升级，中断业务，业务上线效率低。

虚拟化价值体现：

图4-3 服务器整合价值

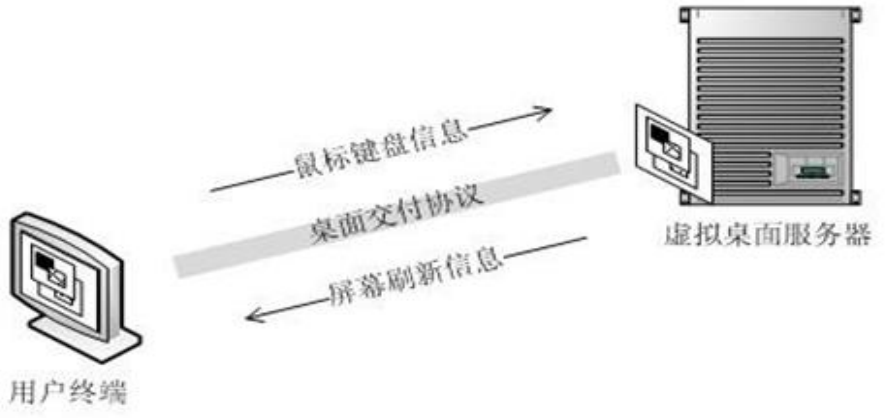
基于虚拟化的优势，服务器整合帮助客户：

1. 实现服务器虚拟化，硬件资源的利用率得到大幅度提升，CPU 的利用率从原来的不足 10%提升到 60%～80%，硬件采购成本大幅降低，同时节省了设备能耗；
2. 提供虚拟化热迁移、资源热插拔等特性，减少系统计划内宕机时间，同时实现零宕机硬件维护和升级，确保业务连续性；
3. 提供虚拟机快照备份和恢复技术，实现系统故障自动切换，减少系统计划外宕机时间；
4. 提供一键式部署，简化业务部署，实现快速业务上线，提升业务敏捷度。

4.3 应用二：企业虚拟桌面

场景技术原理：

图4-4 虚拟桌面基本工作原理



虚拟桌面是典型的云计算应用，可以为用户提供部署在云端的远程计算机桌面服务，通过在虚拟桌面平台服务器的虚拟机上运行用户所需的个人操作系统（Windows XP 或 Windwos7）和应用软件，采用桌面交付协议将操作系统桌面视图以图像的方式传送到用户端设备上显示。同时，用户端的输入通过网络传递至服务侧进行处理，并更新桌面视图内容，是一种将计算机用户使用的个人计算机桌面与物理计算机相隔离的技术。理想情况下，计算机桌面均由网络中的服务器提供而非用户本地计算机，所有程序的执行和数据的存取都在远程服务器中完成，用户可以通过网络访问虚拟桌面（虚拟机）并获得与使用本地计算机桌面相近的体验（如同访问传统的本地安装桌面一样）。

虚拟桌面的优势主要体现在如下方面：对用户而言，不再需要携带笔记本电脑上下班，用户应用统一运行在云端的数据中心中，在安全性方面更有保障，而且无须担心数据的同步和一致性问题。对运维人员而言，虚拟桌面的部署和应用使他们不再需要下到广泛分布的个人计算机使用现场进行系统运维，而只是在集中部署的数据中心进行统一的软/硬件管理。

核心业务诉求：

图4-5 桌面虚拟化诉求

1. 提高虚拟桌面用户体验：

   用户体验是决定了桌面产品生命力的关键，需要提供最佳的用户使用体验，譬如：网络性能体验、高清视频体验等；

2. 确保数据安全性：

   用户应用和数据保存在服务侧，需要提供更高级别的安全防护和访问控制，避免数据泄密。

3. 提高资源利用率：

   用户应用集中在服务器侧运行，数据统一保存在服务器侧的存储设备中。需要提供更高的性能、桌面虚拟机密度和存储设备使用效率，在确保用户使用体验前提下，实现节能减排。

4. 降低软件维护成本：

   提供相关的部署、管理和维护管理功能，提高了工作效率，减少因管理带来的维护成本。

虚拟化价值体现：

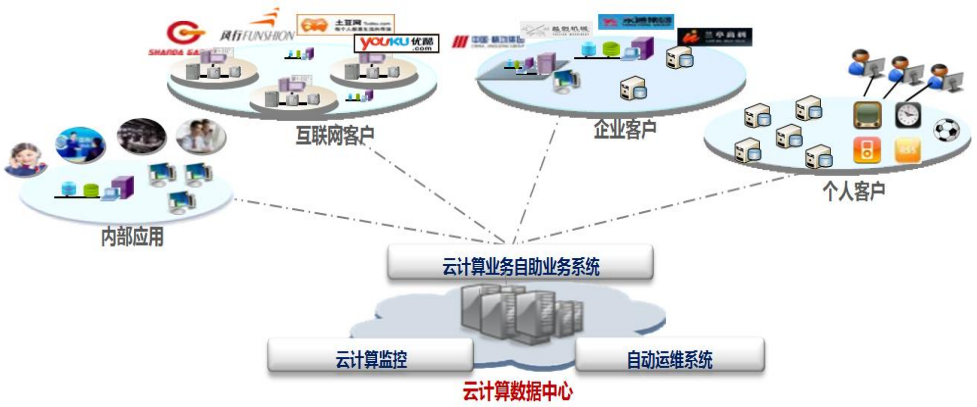
基于虚拟化的优势，服务器整合帮助客户：

1. 提供内存复用技术，提升虚拟机密度；
2. 提供链接克隆和存储瘦分配技术，降低硬件投资成本；
3. 提供按需自动动态资源调配，降低人工管理成本，提升运营效率；
4. 提供 SR-IOV 技术，支撑网络 IO 密集业务，改善用户桌面使用体验；
5. 提供 GPU 虚拟化技术，支撑高性能专业制图和视觉体验要求较高的业务场景；

4.4 应用三：互联网数据中心

场景技术原理：

图4-6 IDC 应用场景



与传统机房不同，对 IDC（互联网数据中心）业务而言，除了满足企业内部需要，其主要是服务于不同企业的互联网服务器托管数据中心，利用已有的互联网通信线路、带宽资源，为企业、政府提供服务器托管、租用以及相关增值等方面的全方位服务，同时 IDC 服务需要对客户提供有保障的维护和管理工作。IDC 服务的市场需求虽然巨大，但是服务商的 IDC 业务所面对的竞争压力不容忽视，需要利用新技术和新理念，开拓新的业务空间，降低运维和消耗的成本，提升 IDC 机房整体的运转效率，强调性价比，体现 IDC 机房更大的价值。**（注：目前国内 IDC 业务经营企业以电信运营商为主）**

核心业务诉求：

图4-7 传统 IDC 面临的问题与挑战

1. 更高的效率(性能)，实现更低的成本

   提供高性能的虚拟机和资源弹性伸缩能力，提高资源利用率，降低机房硬件投资成本；

2. 系统可靠性

   提供应用可靠性保障，快速感知虚拟机运行状态，支持虚拟机级 HA（可靠性）；

3. 联动绿色节能，实现更低的能耗

   提供更多绿色节能技术，实现基础设施与 IT 设备联动节能、负荷均衡，降低机房能耗；

4. 更优的运维服务，实现精细化管理功能

   提供高效、智能管理维护能力，缩短业务上线周期，提供企业用户更优的服务；同时提供精细化管理能力，进一步提升用户体验；

虚拟化价值体现：

基于虚拟化的优势，服务器整合帮助客户：

1. 提供CPU核动态调度能力和GuestOS应用状态感知，结合资源按需DRS调度功能，实现基础设施与 IT 设备联动节能，降低服务器资源空闲时间，有效降低能耗；
2. 提供多种弹性主机模板和一键式部署能力，缩短业务上线时间，加快 IDC 业务提供；
3. 提供虚拟机资源 QoS 能力，提供客户可理解的资源的 QoS 控制，保障资源合理分配，保护关键应用和关键客户，并满足计费要求；

5 结 论/Conclution

FUSIONCOMPUTE 虚拟化平台统一虚拟化平台作为华为云计算解决方案的关键技术平台，主要定位企业关键应用领域，采用裸金属架构的 X86 虚拟化技术，是一个基于开源 XEN 技术增强的全面的服务器虚拟化平台。通过实现对服务器物理资源的抽象，将 CPU、内存、I/O 等服务器物理资源转化为可统一管理和分配的逻辑资源，为应用提供安全隔离的虚拟机运行环境，可充分利用硬件辅助虚拟化技术，具有更高的性能、可用性和安全性特点，结合上层云操作系统管理功能，可实现更低的运营成本、更高的自动化管理水平和更快速的业务响应速度。

6 缩略语/Acronyms and Abbreviations

表6-1 缩略语清单

英文缩写	英文全称	中文全称
FUSIONCOMPUTE 虚拟化平台	Unified Virtualization Platform	统一虚拟化平台
VM	Virtual Machine	虚拟机
VMM	Virtual Machine Monitor	虚拟机管理器
KBox	Kernel black BOX	黑匣子
VDI	Virtual Desktop Infrastructure	虚拟桌面基础结构
IDC	Internet Data Center	互联网数据中心
ISA	Instruction Set Architecture	指令集
NUMA	Non Uniform Memory Access	非一致性内存访问
VLAN	Virtual Local Area Network	虚拟局域网
MMU	Memory Management Unit	存储器管理单元