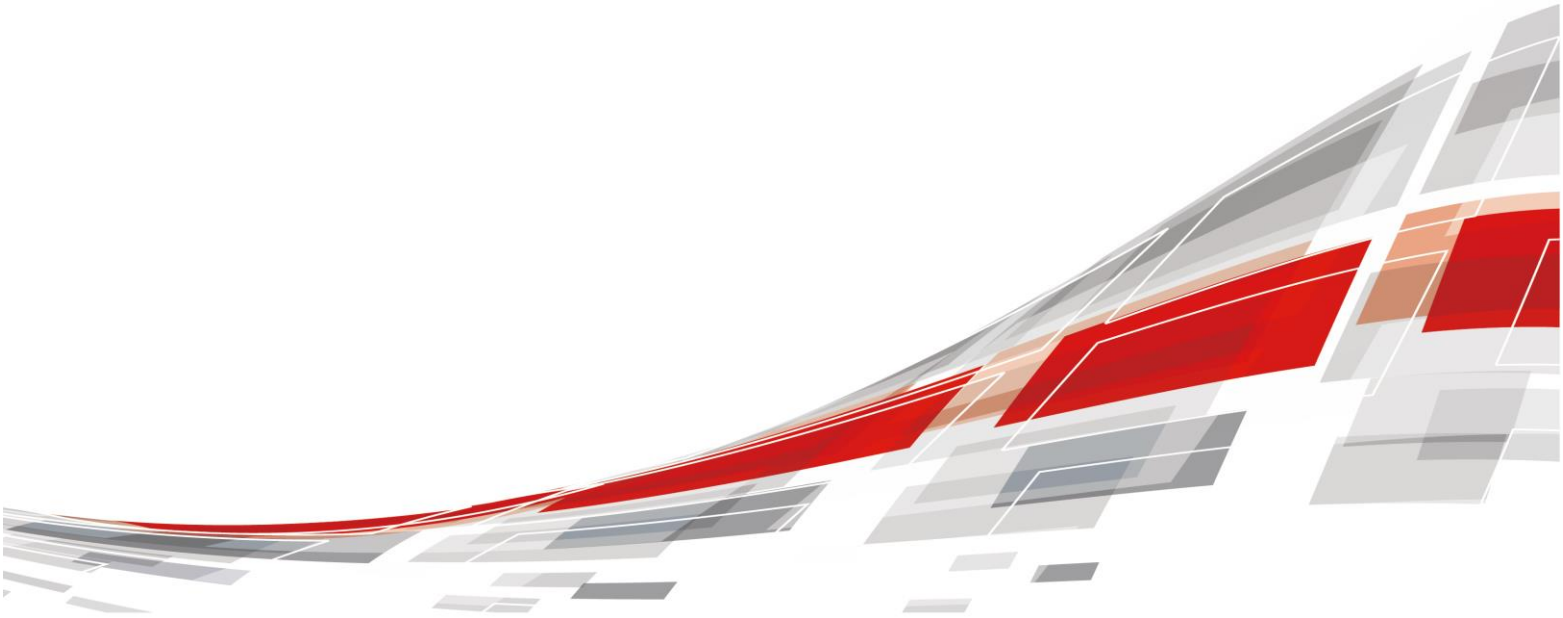
FusionOne Compute

23.3.0

技术白皮书

文档版本 01

发布日期 2025-04-30



超聚变数字技术有限公司

版权所有 © 超聚变数字技术有限公司 2025。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明

FUSION 和其他超聚变商标均为超聚变数字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

本文中，只是为了描述的简洁和方便理解，用“xFusion”指代“xFusion Digital Technologies Co., Ltd.”，这并不代表“xFusion”还可以具备其它含义。基于本文中单独提及或描述的“xFusion”，不能用于“xFusion Digital Technologies Co., Ltd.”之外的理解或表达，超聚变数字技术有限公司也不承担因单独使用“xFusion”所带来的其它任何法律责任。

您购买的产品、服务或特性等应受超聚变数字技术有限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，超聚变数字技术有限公司对本文档内容不做任何明示或默示的声明或保证。

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。

超聚变数字技术有限公司

地址： 河南省郑州市郑东新区龙子湖智慧岛正商博雅广场 1 号楼 9 层 邮编：450046

网址： https://www.xfusion.com

前言

概述

本文档介绍了 FusionOne Compute 的产品价值、产品架构、高性能、线性扩展、系统安全以及系统可靠性。

读者对象

本文档主要适用于以下工程师：

- 营销工程师
- 技术支持工程师
- 维护工程师

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
⚠ 危险	表示如不避免则将会导致死亡或严重伤害的具有高等级风险的危害。
⚠ 警告	表示如不避免则可能导致死亡或严重伤害的具有中等级风险的危害。
⚠ 注意	表示如不避免则可能导致轻微或中度伤害的具有低等级风险的危害。

 版权所有 © 超聚变数字技术有限公司

符号	说明
须知	用于传递设备或环境安全警示信息。如不避免则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。 “须知”不涉及人身伤害。
说明	对正文中重点信息的补充说明。 “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

文档版本	发布日期	修改说明
01	2025-04-30	第一次正式发布。

 版权所有 © 超聚变数字技术有限公司

目 录

前言.....ii

1 IT 设施虚拟化 1

2 FusionOne Compute..... 4

2.1 技术架构 5

2.2 配置介绍 8

2.3 典型组网介绍 9

2.4 硬件平台介绍 10

3 计算虚拟化..... 12

3.1 概述..... 13

3.2 虚拟机资源 QoS..... 16

3.3 CPU 内存超分..... 17

3.4 虚拟机资源动态热添加(DRX)..... 23

3.5 虚拟机热迁移..... 24

3.6 虚拟机快速克隆..... 27

3.7 虚拟机回收站..... 28

3.8 VMware 虚拟机模板导入 28

3.9 NUMA 调优..... 29

3.10 动态资源调度（DRS&DPM） 29

3.11 快照与快照策略 30

3.12 XPU 可视化管理与调度策略 31

3.13 GPU 虚拟化..... 34

3.14 NPU 虚拟化..... 35

3.15 GPU 资源池..... 36

 版权所有 © 超聚变数字技术有限公司

3.16 GPU 直通 46
3.17 网络 USB 47
3.18 SSH 密钥管理 48
3.19 PCI 设备白名单管理 49
3.20 GuestOS 支持解耦 49
4 网络虚拟化 50
4.1 概述 51
4.2 所画即所得 52
4.3 连通性检测 53
4.4 分布式虚拟交换机 54
4.5 网络安全组 55
4.6 SDN 网络 56
4.7 Overlay 网络 57
4.8 VPC 60
4.9 NAT 网关 61
4.10 弹性 IP 63
4.11 网络 ACL 63
4.12 IP 地址预留 65
4.13 网络 QoS 66
4.14 SR-IOV 交换 67
4.15 用户态 OVS 交换 69
4.16 DHCP 服务 70
4.17 虚拟路由器 71
4.18 端口流量镜像 72
4.19 在线修改虚拟机 IP 地址 73
4.20 虚拟负载均衡器 75
5 存储虚拟化 79
5.1 概述 79
5.2 虚拟机直通 SAN 80
5.3 虚拟文件系统（IP/FC SAN） 81
5.4 NAS 存储文件系统 83

 版权所有 © 超聚变数字技术有限公司

5.5 磁盘模式配置 83
5.6 自动精简配置 85
5.7 虚拟机备份 86
5.7.1 定时备份和立即备份 87
5.7.2 VIMS CBT 备份 87
5.7.3 CDP 备份 87
5.7.3.1 原理描述 87
5.7.3.2 使用场景 88
6 统一管理 91
6.1 设备管理 91
6.1.1 可视化节点 91
6.1.2 网络设备纳管 92
6.1.3 操作系统纳管 93
6.2 监控管理 94
6.2.1 统一告警与事件 94
6.2.2 可视化性能监控 95
6.2.3 节点可视化运维 95
6.2.4 硬件故障预测 96
6.2.4.1 硬盘故障预测 96
6.2.4.2 内存故障预测 98
6.2.5 性能监控 98
6.3 运维管理 99
6.3.1 一键部署 99
6.3.2 系统下电 100
6.3.3 日志收集 100
6.3.4 任务中心 100
6.3.5 日志管理 101
6.4 多云管理 101
7 容器平台 105
8 系统可扩展性 107

 版权所有 © 超聚变数字技术有限公司

9 系统可靠性 109
9.1 管理架构 109
9.2 业务虚拟机 112
9.3 硬件 114
9.4 故障恢复 117
10 安全防护 119
10.1 安全威胁来源 119
10.2 平台安全 121
10.2.1 网络安全 122
10.2.2 应用安全 123
10.2.2.1 权限管理 123
10.2.2.2 Web 安全 124
10.2.2.3 数据库加固 125
10.2.2.4 日志管理 125
10.2.3 主机安全 126
10.2.4 数据安全 126
10.3 业务安全 126
10.3.1 虚拟机安全删除 126
10.3.2 无代理防病毒&防火墙 127
11 北向接口与开放兼容 131
12 系统规格 133

 版权所有 © 超聚变数字技术有限公司

1 IT 设施虚拟化

IT 基础架构的时代变革，核心是数据中心的时代变革。随着数据不断增长以及互联网业务的兴起，新兴业务的激增、对 IT 平台的快速敏捷提出了更高的要求。分布式、云化技术应运而生，越来越多的企业采用虚拟化与云计算技术来构建 IT 系统，提升 IT 系统的资源利用率以及缩短业务上线周期。

服务器虚拟化技术是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，我们不再受限于物理上的界限，而是让 CPU、内存、磁盘、I/0 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，实现服务器资源整合，让 IT 对业务的变化更具适应力。

虚拟化的概念在 20 世纪 60 年代首次出现，利用它可以对属于稀有而昂贵资源的大型机硬件进行分区。随着时间的推移，微型计算机和 PC 可提供更有效、更经济的方法来分配处理能力，因此到 20 世纪 80 年代，虚拟技术已不再广泛使用。

但是到了 20 世纪 90 年代，研究人员开始探索如何利用虚拟化解决与廉价硬件激增相关的一些问题，例如，利用率不足、管理成本不断攀升和易受攻击等。虚拟化技术处于时代前沿，可以帮助企业升级和管理他们在世界各地的 IT 基础架构并确保其安全。虚拟化技术可以扩大硬件的容量，简化软件的重新配置过程。CPU 的虚拟化技术可以单 CPU 模拟多 CPU 并行，允许一个平台同时运行多个操作系统，并且应用程序都可以在相互独立的空间内运行而互不影响，从而显著提高计算机的工作效率。

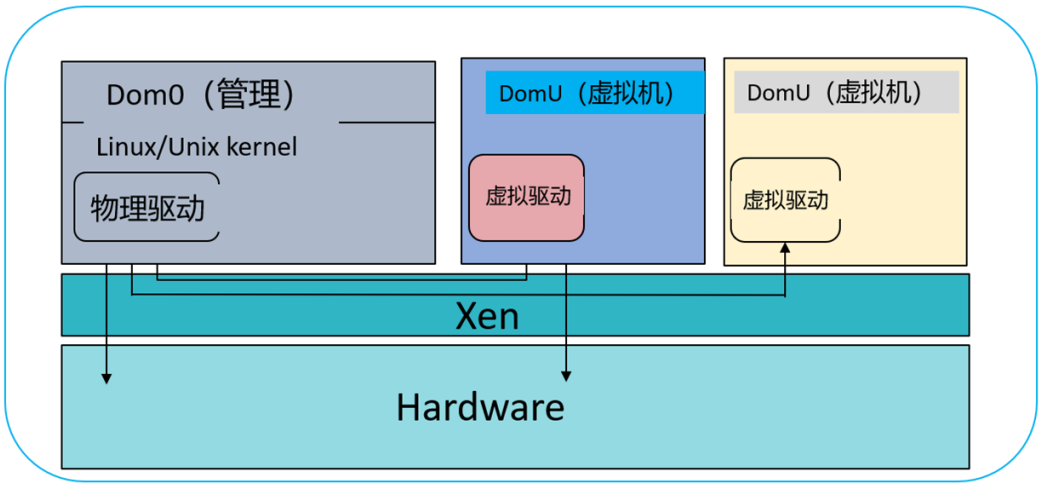
在物理环境下，直接在节点上安装操作系统，服务器被整个系统独占，资源利用率低。虚拟化则是在底层增加了一个 Hypervisor 层。可以在 Hypervisor 上创建多个虚拟机，虚拟机中可以安装不同的操作系统（如 Linux、Windows），提高了服务器资源使用率。

技术根据虚拟化程度，可以划分为全虚拟化、半虚拟化方式、硬件辅助虚拟化：

半虚拟化方式

 版权所有 © 超聚变数字技术有限公司

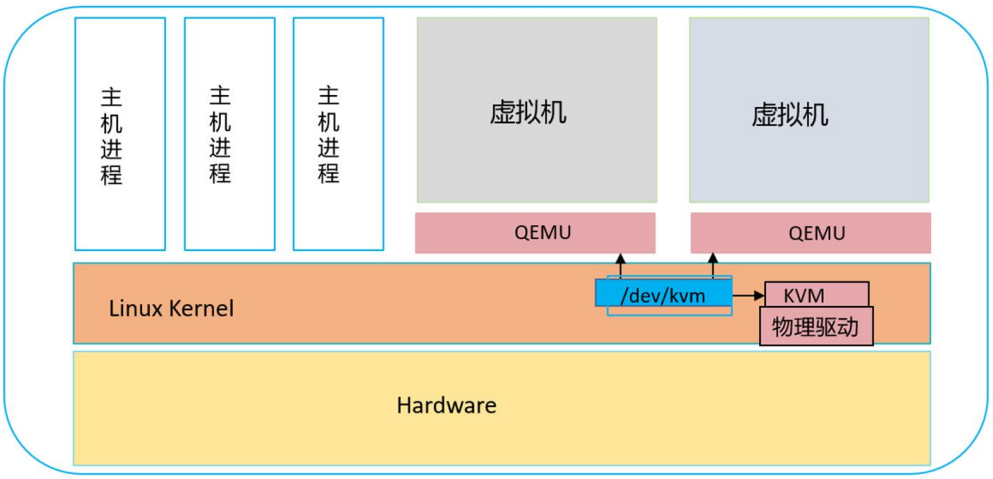
半虚拟化方式(Para—virtualization)是指虚拟管理层对底层硬件进行部分模拟，把底层硬件平台的低级 API 用一套高级 API 来代替并提供给上层虚拟机，客户操作系统需要支持底层硬件。这样，虚拟机在运行时可减少用户模式(User Mode)和特权模式(Supervisor Mode)之间的切换，从而降低运行时的开销。例如，虚拟化软件 XEN 就是使用了半虚拟化方式。



全虚拟化方式

全虚拟化方式(Full Virtualization)是指虚拟管理层对底层硬件进行完全模拟，把底层硬件平台的 API 完整拷贝并提供给上层虚拟机，虚拟机中的客户操作系统和应用无法感知真实机器的存在。全虚拟化方式的优点是兼容性好，可以在虚拟机上使用任何类型的操作系统，客户操作系统无需修改，所有软件都能在虚拟机中运行。

如下图，为 KVM 虚拟化技术原理：



硬件辅助虚拟化增强

 版权所有 © 超聚变数字技术有限公司

借助 CPU 针对虚拟化的指令，减少软件层面因为虚拟化的开销。将 GuestOS 运行在一种受控状态，原来的敏感指令可在受控下陷入 VMM，解决部分非特权敏感指令的“陷入-模拟”难题，而且模式切换上下文的保存和恢复均由硬件来完成，可大大提升切换的效率。以 Intel VT-x 硬件辅助虚拟化技术为例，Intel VT 和 AMD-V 创建一个新的 Ring-1 单独给 Hypervisor 使用。

 版权所有 © 超聚变数字技术有限公司

2 FusionOne Compute

FusionOne Compute 虚拟化平台，主要负责硬件资源的虚拟化，以及对虚拟资源、业务资源、用户资源的集中管理。它采用虚拟计算、虚拟存储、虚拟网络等技术，完成计算资源、存储资源、网络资源的虚拟化。同时通过统一的接口，对这些虚拟资源进行集中调度和管理，从而降低业务的运行成本，保证系统的安全性和可靠性。

- 统一运维

  FusionOne Center 提供整体的管理平台，与服务区 BMC、虚拟化组件对接，可统一进行虚拟机业务管理，同时提供自动化的运维工具，如一键运维、自动化作业、一体化监控告警等方式，降低平台运维难度，减少运维工作量。

- 统一虚拟化平台

  将计算资源划分为多个虚拟机资源，为用户提供高性能、可运营、可管理的虚拟机。支持虚拟机资源按需分配，支持多操作系统，QoS 保证资源分配，隔离用户间影响。

- 大集群

  单个计算集群最大可支持 256 个主机，5000 台虚拟机（单站点支持多个计算集群）。

- 自动化调度

  支持自定义的资源管理 SLA（Service-Level Agreement）策略、故障判断标准及恢复策略。通过 IT 资源调度、热管理、能耗管理等一体化拉通，降低维护成本。自动检测服务器或业务的负载情况，对资源进行智能调度，均衡各服务器及业务系统负载，保证系统良好的用户体验和业务系统的最佳响应。

- 完善的权限管理

 版权所有 © 超聚变数字技术有限公司

可根据不同的角色、权限等，提供完善的权限管理功能，授权用户对系统内的资源进行管理。

- 丰富的运维管理

  提供多种运维工具，实现业务的可控、可管，提高整个系统运营的效率。

  - 支持“黑匣子”快速故障定位

    系统通过获取异常日志和程序堆栈，缩短问题定位时间，快速解决异常问题。

  - 支持全 Web 化的界面

    通过 Web 浏览器对所有硬件资源、虚拟资源、用户业务发放等进行监控管理。

- 云安全

  采用多种安全措施和策略，并遵从信息安全法律法规，对用户接入、管理维护、数据、网络、虚拟化等提供端到端的业务保护。

2.1 技术架构

2.2 配置介绍

2.3 典型组网介绍

2.4 硬件平台介绍

2.1 技术架构

架构组成

FusionOne Compute 主要以虚拟化技术为核心，由计算虚拟化、网络虚拟化、存储虚拟化（VIMS），将计算、存储、网络等虚拟资源融合融合到标准的通用服务器中，形成基础架构平台，并通过网络聚合成一个集群，可实现积木式的平滑扩展，形成一个统一的基础设置资源池。FusionOne Compute 提供虚拟化平台，提供计算资源、网络资源虚拟化，并通过 VIMS 组件对接支持外部存储虚拟化，为虚拟化平台的基础架构组件，为其他组件运行提供了基础平台，同时支持部署 FusionOne Center 管理虚拟机（可选），实现硬件、平台共管。

 版权所有 © 超聚变数字技术有限公司

图2-1 总体架构

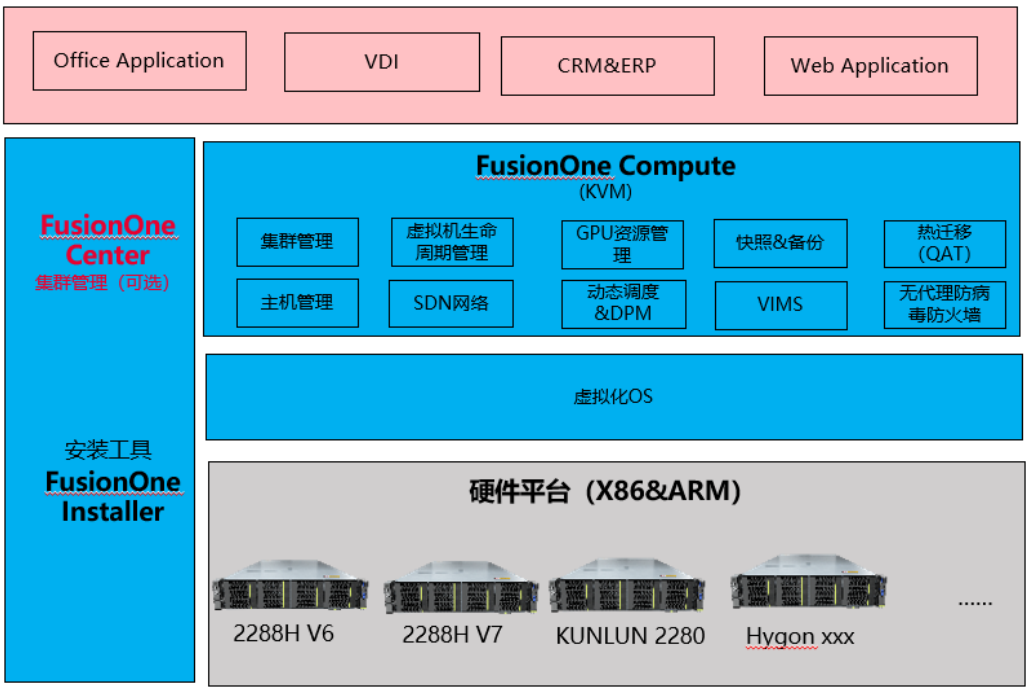


表2-1 节点类型说明

名称	说明
MCNA（管理节点）	具有管理功能的节点，其上部署了 VRM 管理虚拟机。同时也可提供计算功能。 根据需要部署 1 个～多个。
CNA（计算节点）	具有计算功能的节点，虚拟化计算资源。 根据需要部署 0 个～多个。

版权所有 © 超聚变数字技术有限公司

虚拟化管理与节点通信

图2-2 节点架构

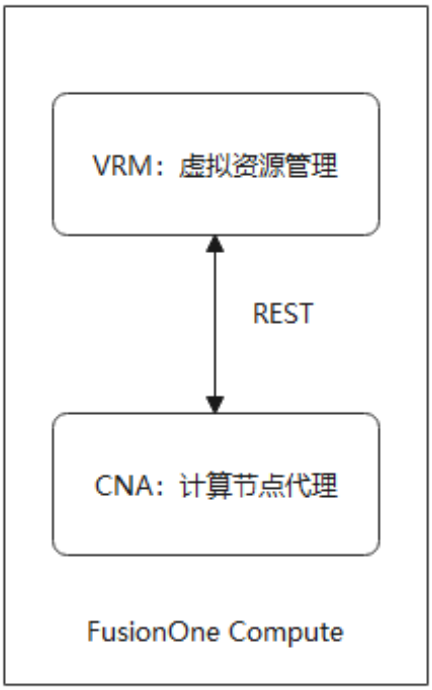


表2-2 虚拟化管理与代理的模块说明

名称	说明
VRM	VRM 主要提供以下功能： • 管理集群内的块存储资源。 • 管理集群内的网络资源(IP/VLAN)，为虚拟机分配 IP 地址。 • 管理集群内虚拟机的生命周期以及虚拟机在计算节点上的分布和迁移。 • 管理集群内资源的动态调整。 • 通过对虚拟资源、用户数据的统一管理，对外提供弹性计算、存储、IP 等服务。 • 通过提供统一的操作维护管理接口，操作维护人员通过 WebUI 远程访问 FusionOne Compute 对整个系统进行操作维护，包含资源管理、资源监控、资源报表等。

 版权所有 © 超聚变数字技术有限公司

名称	说明
CNA	具有计算功能的节点，虚拟化计算资源。 • 提供虚拟计算功能。 • 管理计算节点上的虚拟机。 • 管理计算节点上的计算、存储、网络资源。

2.2 配置介绍

节点配置要求：

表2-3 节点配置要求

配置项	典型配置	说明
服务器类型	机架服务器	根据客户对机柜空间、磁盘大小、密度、PCIE 网卡数量等选择合适的服务器类型。 机架服务器：优势为灵活，支持多种硬盘类型，预留多个 PCIE 槽位，支持 GPU 卡。
CPU/内存配置	支持 Intel® Xeon®系列 CPU 和 澜起®津逮®系列 CPU；支持鲲鹏 920 处理器 支持 DDR4 或 DDR5 内存	CPU/内存配置根据客户的业务规格和配置可以动态调整配置，提供更多的计算资源。 建议 CPU 最低主频 2.0GHz，最少 8 个物理核心。
磁盘	操作系统盘通常为 2*600GB SAS 盘，或 2*480GB SATA SSD，或 2*960GB SATA SSD	-
网卡	2*10GE+2*10GE	推荐管理、业务共用 2*10GE 网口，外部存储网络平面独占

 版权所有 © 超聚变数字技术有限公司

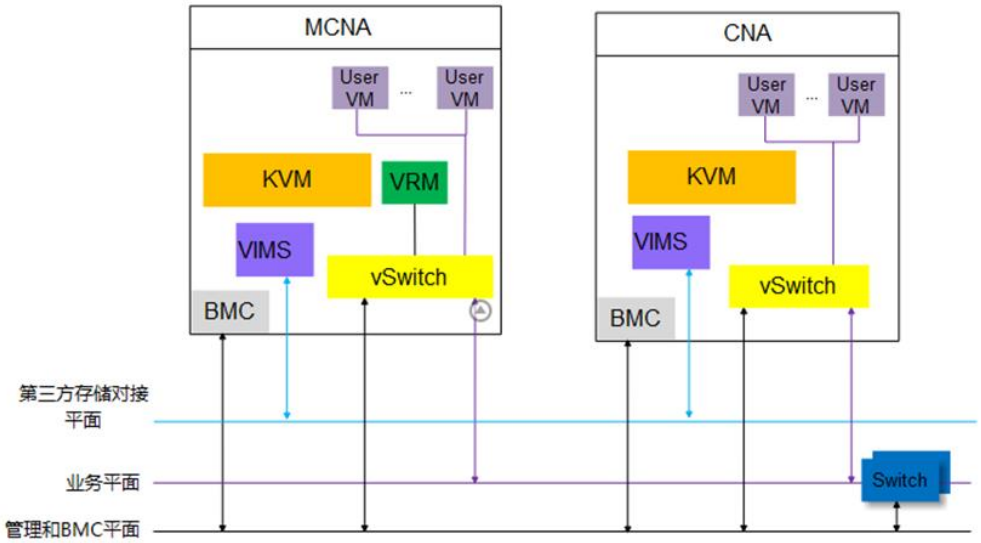
配置项	典型配置	说明
		2*10GE 网口（可选）。

2.3 典型组网介绍

FusionOne Compute 系统组网包含：管理平面、存储平面、业务平面。其中存储平面建议独占物理网口，管理与业务平面、存储平面推荐独占，但是实际生产环境交付中，考虑性能、安全隔离要求，存储要求独占 10GE 带宽以上的网口，不与管理或则业务平面复用。

详细的组网情况如下：

图2-3 组网图



通信平面类型说明介绍：

- 管理平面：用于系统的业务操作和运维管理，支持 TCP/IP 协议，支持 GE/10GE 组网，可以与业务平面共网卡，通过 VLAN 隔离。
- BMC 平面：服务器设备管理平面，访问服务器设备的运维管理平台。
- 第三方存储对接平面：存储业务数据读写的网络，对应存储网络的存储业务网络平面。

 版权所有 © 超聚变数字技术有限公司

- 业务平面：客户业务通信网络平面，支持 TCP/IP 协议，支持 GE/10GE 组网，可以与管理平面共网卡，通过 VLAN 隔离。

2.4 硬件平台介绍

FusionOne Compute 基于超聚变 FusionServer 系列（Intel&AMD）、KUNLUN（ARM-KUNPENG）系列服务器，结合虚拟化业务形态深度调优，将服务器下的高可靠到虚拟化场景，主要是如下特色技术功能：

分类	特色技术
架构设计	• 网络冗余、跨网卡绑定 • 电源 1+1 冗余 • 风扇 1+1 冗余 • 系统盘 Raid 冗余
电源	电源三大核心技术（无桥 PFC 技术、交错并联技术、多相磁集成技术），电源故障率比业界低 50%；
风扇	散热可靠：整机在 45℃环境温度下正常运行 • 热管拉远散热技术，较单体散热器能力提升 50%； • 仿生翼型扇叶对旋风扇，风量提升 30%；
CPU	CPU 温度、主频检测，异常告警
内存	• ECC 内存技术 • 自愈：业界独有智能内存故障自愈技术，服务器宕机风险较业界低 66%
网卡	链路告警、丢包检测、速率检测
Raid 卡	系统盘 Raid1、数据盘采用 Jbod 模式（良好的热插拔支持）
硬盘	• 硬盘可靠：降噪风扇、吸波减震、隔声防护、一体化减震支架等硬盘防震降躁设计，降低 60%硬盘故障率。 • 三重防振设计：振动源头-减振螺钉、振动传导-机框刚度加强、受体保护-弹片及垫片 • 降噪设计：定位敏感频率、悟空降噪风扇、均流板、吸音棉，减

 版权所有 © 超聚变数字技术有限公司

分类	特色技术
	少振动及避免运行中的敏感频率导致故障。 • 硬盘热插拔、Raid 技术 • SSD 寿命检测 • 坏道检测
AI 预测	AI 硬盘故障预测，提前 7-30 天预测出风险盘，故障有预防，数据最安全。
AI 诊断	AI 加持 FDM（Fault Diagnose Management）专利技术，支持基于部件的精准故障诊断（包括 CPU、内存、硬盘、电源等部件），方便部件故障定位和更换，准确率达 96%。
AI 监控	涵盖站点机箱、服务器、交换机等硬件设备监控，也包括计算、存储、网络等资源监控、告警

后续章节，针对虚拟化架构主要模块所涉及的主要产品技术做详细介绍，主要包括如下几个方面：

- 虚拟化：计算虚拟化、网络虚拟化、存储虚拟化
- 云原生(FusionOne Kubernetes Engine)
- 统一管理
- 系统可扩展性、系统可靠性、安全防护
- 北向接口与开放兼容
- 系统规格。

 版权所有 © 超聚变数字技术有限公司

3 计算虚拟化

3.1 概述
3.2 虚拟机资源 QoS
3.3 CPU 内存超分
3.4 虚拟机资源动态热添加(DRX)
3.5 虚拟机热迁移
3.6 虚拟机快速克隆
3.7 虚拟机回收站
3.8 VMware 虚拟机模板导入
3.9 NUMA 调优
3.10 动态资源调度（DRS&DPM）
3.11 快照与快照策略
3.12 XPU 可视化管理与调度策略
3.13 GPU 虚拟化
3.14 NPU 虚拟化
3.15 GPU 资源池
3.16 GPU 直通
3.17 网络 USB

 版权所有 © 超聚变数字技术有限公司

3.18 SSH 密钥管理

3.19 PCI 设备白名单管理

3.20 GuestOS 支持解耦

3.1 概述

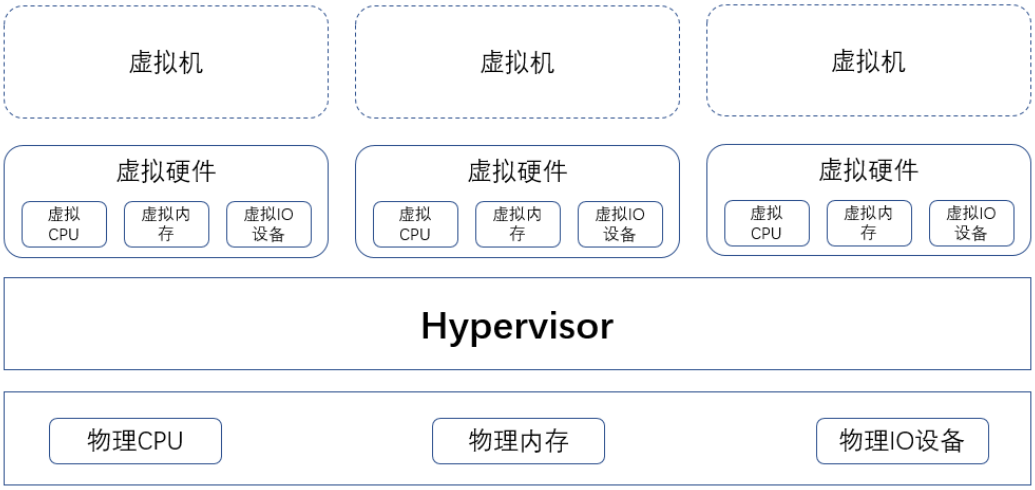
虚拟化是资源的逻辑表示，实现虚拟资源与物理硬件解耦。虚拟化的目标通常是提高物理资源利用率。虚拟化技术的实现形式是在系统中加入一个虚拟化层，即虚拟机监控器 Hypervisor，Hypervisor 有时也称为 VMM（Virtual Machine Monitor 虚拟机监控器）。通过直接对硬件资源进行抽象、模拟、调度，提供了虚拟硬件接口，将下层的资源抽象成另一形式的资源，提供给上层使用。并通过对硬件分区、分时复用等方法来管理硬件资源，完成硬件资源在多个虚拟机之间的共享和调度。

以 Fusionone Compute 为例，FusionOneCompute 向下管理硬件资源，向上提供硬件抽象。底层是物理硬件资源，包括三大主要硬件：CPU、内存和 I/O 设备。FusionOneCompute 运行在硬件资源层上，并为虚拟机提供虚拟的硬件资源。客户机操作系统(Guest OS)运行在虚拟硬件资源上，这与传统的操作系统功能一致，管理硬件资源并为上层应用提供统一的软件接口。总览整个架构，FusionOneCompute 具有两种功能：管理虚拟环境；管理物理资源。管理虚拟环境：为了完成用户对虚拟机的管理，管理虚拟环境需要提供虚拟机创建、修改、查询、删除、暂停等生命周期管理操作，以及虚拟机的迁移、快照等。

管理物理资源：为了提供硬件资源的调度和共享，物理资源的管理需要提供底层物理设备的管理、调度，并提供安全隔离机制以保证某个虚拟机中的恶意代码不会破坏整个系统的稳定性。

 版权所有 © 超聚变数字技术有限公司

图3-1 虚拟化架构图



FusionOneCompute 实现原理

虚拟化技术主要包括 CPU 虚拟化、内存虚拟化和 I/O 虚拟化。

CPU 虚拟化

在物理环境中，操作系统是所有物理资源的管理者，其拥有最高的权限级别，可以直接访问寄存器、内存和 I/O 设备等关键的物理资源。但是在虚拟环境下，存在多个虚拟机操作系统，因此需要 Hypervisor 来进行协调和管理。同时为了保证安全性和隔离性，虚拟机的操作系统不能拥有最高的访问权限。

因此虚拟机对物理资源的访问应当被“截获”并由 Hypervisor 中进行模拟和调度。这些访问物理资源的指令称为敏感指令。在 Intel VT-x 等硬件辅助虚拟化技术出现之前，软件虚拟化技术只能利用计算机体系中现有的特权级进行处理。这就要求所有的敏感指令都是特权指令，才能让 Hypervisor 截获所有的敏感指令。但系统中通常存在一些敏感非特权指令，它们称为“虚拟化漏洞”，CPU 虚拟化的关键就是消除虚拟化漏洞。此外，中断和异常的模拟及注入也是 CPU 虚拟化应当考虑的问题。虚拟设备产生的中断无法像物理中断一样直接传送到物理 CPU 上，而需要在注入等方式让虚拟机感知到中断并处理。FusionOneCompute 通过硬件辅助虚拟化等技术，实现虚拟机内运行指令与主机上运行几乎无差异。

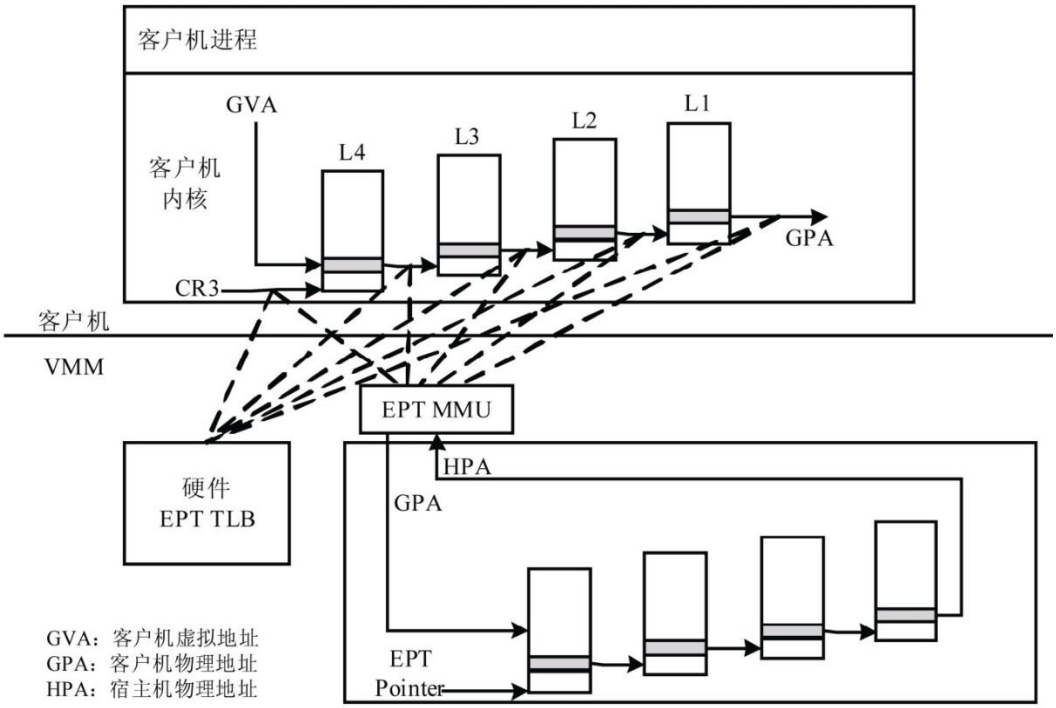
内存虚拟化

操作系统对于物理内存的使用基于两个假设：1、内存都是从物理地址 0 开始的；2、物理内存都是连续的。对于一台物理机上的多个虚拟机而言，它们共享物理内存资源，因此无法满足假设 1。对于此类问题通常的解决办法是增加一个中间转换层，这个转

 版权所有 © 超聚变数字技术有限公司

换也是通过硬件辅助实现的。FusionOneCompute 通过这类技术实现虚拟机高速访问内存。

图3-2 内存虚拟化原理图



I/O 虚拟化

在物理环境下，操作系统通过 I/O 端口访问特定的 I/O 设备，称为 PIO（Port I/O，端口 I/O），或者将 I/O 设备上的寄存器映射到预留的内存地址空间进行读写，称为 MMIO（Memory-Mapped I/O，内存映射 I/O）。因此 Hypervisor 需要截获所有的虚拟机的 PIO 和 MMIO 操作并对其进行模拟，再将结果告知虚拟机。当多个虚拟机运行在同一个物理机上时，Hypervisor 通过模拟，为每个虚拟机提供一个 IO 设备（可能是全虚拟化、半虚拟化、直通）。全虚拟化是指这个设备应当与现实设备具有完全相同的接口，从而允许用户无须修改客户机操作系统中原有的驱动程序就能使用这个虚拟设备。半虚拟化是需要虚拟机感知自己运行在虚拟机中，并通过使用特殊的前端驱动（virtio）与后端对接完成高性能的通信。直通是通过 PCIe 设备的 IOMMU、中断重映射等功能完成虚拟机与物理硬件的直接交互，无需 Hypervisor 参与，实现性能与物理主机上直接使用设备几乎无区别。这些类型设备 FusionOneCompute 都能支持。最常用的是半虚拟化，其对物理硬件无依赖并且性能很高。

 版权所有 © 超聚变数字技术有限公司

3.2 虚拟机资源 QoS

- **CPU QoS**

  虚拟机的 CPU QoS 用于保证虚拟机的计算资源分配，隔离虚拟机间由于业务不同而导致的计算能力相互影响，满足不同业务对虚拟机计算性能的要求，最大程度复用资源，降低成本。

  创建虚拟机时，可根据虚拟机预期部署业务对 CPU 的性能要求而指定相应的 CPU QoS。不同的 CPU QoS 代表了虚拟机不同的计算能力。指定 CPU QoS 的虚拟机，系统对其 CPU 的 QoS 保障，主要体现在计算能力的最低保障和资源分配的优先级。

  CPU QoS 包含如下三个参数：

  - CPU 资源份额

    CPU 份额定义多个虚拟机在竞争物理 CPU 资源的时候按比例分配计算资源。

    以一个主频为 2.8GHz 的单核物理主机为例，如果上面运行有三台单 CPU 的虚拟机。三个虚拟机 A，B，C，份额分别为 1000，2000，4000。当三个虚拟机 CPU 满负载运行时，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 份额只在各虚拟机竞争计算资源时发挥作用，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，例如，如果虚拟机 B 和 C 均处于空闲状态，虚拟机 A 可以获得整个物理核即 2.8GHz 的计算能力。

  - CPU 资源预留

    CPU 预留定义了多个虚拟机竞争物理 CPU 资源的时候分配的最低计算资源。

    如果虚拟机根据份额值计算出来的计算能力小于虚拟机预留值，调度算法会优先按照虚拟机预留值的能力把计算资源分配给虚拟机，对于预留值超出按份额分配的计算资源的部分，调度算法会从主机上其他虚拟机的 CPU 上按各自的份额比例扣除，因此虚拟机的计算能力会以预留值为准。

    如果虚拟机根据份额值计算出来的计算能力大于虚拟机预留值，那么虚拟机的计算能力会以份额值计算为准。

    以一个主频为 2.8GHz 的单核物理机为例，如果运行有三台单 CPU 的虚拟机 A、B、C，份额分别为 1000、2000、4000，预留值分别为 700MHz、0MHz、0MHz。当三个虚拟机满 CPU 负载运行时：

 版权所有 © 超聚变数字技术有限公司

  - 虚拟机 A 如果按照份额分配，本应得 400MHz，但由于其预留值大于 400MHz，因此最终计算能力按照预留值 700MHz 算。
  - 多出的(700-400)MHz 按照 B 和 C 各自的份额比例从 B 和 C 处扣除。
  - 虚拟机 B 获得的计算能力约为(800-100)MHz，虚拟机 C 获得的计算能力约为(1600-200)MHz。

  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。例如，如果虚拟机 B 和 C 均处于空闲状态，虚拟机 A 可以获得整个物理核即 2.8GHz 的计算能力。

- CPU 资源限额

  控制虚拟机占用物理 CPU 资源的上限。以一个两 CPU 的虚拟机为例，如果设置该虚拟机 CPU 上限为 3GHz，则该虚拟机的两个虚拟 CPU 计算能力被限制为 1.5GHz。

- 内存 QoS

  支持通过内存预留定义多个虚拟机竞争内存资源的时候分配的内存下限，能够确保虚拟机在实际使用过程中一定可使用的内存资源，保证关键虚拟机的内存足够。预留的内存被会虚拟机独占。即，一旦内存被某个虚拟机预留，即使虚拟机实际内存使用量不超过预留量，其他虚拟机也无法抢占该虚拟机的空闲内存资源。一般该功能配套内存超分使用，当主机内存紧张的时候，保证预留内存的虚拟机的内存使用。

- **网络 QoS**：策略提供带宽配置控制能力，QoS 功能不支持同一主机上虚拟机之间的流量限制。包含如下方面：
  - 基于端口组成员接口发送方向与接收方向的带宽控制。
  - 基于端口组的每个成员接口提供流量整形、带宽优先级的控制能力。

3.3 CPU 内存超分

在云计算和虚拟化技术中，超分配是一种资源管理策略，它允许在一个物理服务器上运行多个虚拟机，并将物理资源超出实际物理资源的总量分配给这些虚拟机。虚拟化超分可以给用户带来如下的好处：

- 提高资源利用率：超分可以将未被使用的资源分配给其他虚拟机，从而提高系统的资源利用率，这样可以节省硬件资源、降低硬件成本。

 版权所有 © 超聚变数字技术有限公司

- 提高灵活性和可扩展性：超分可根据实际需求动态调整资源分配，实现灵活的资源管理和扩展。当有新的虚拟机需要运行时，可以根据需要分配更多的资源，而不需要购买新硬件。
- 提高系统稳定性和可靠性：通过监控虚拟机的资源使用情况，可以预测并避免资源瓶颈和故障。当一个虚拟机发生故障，其他虚拟机可以继续运行，提高系统的稳定性和可靠性。

CPU 超分

技术原理

CPU 超分本质上是物理 1 个 CPU 通过虚拟化的方式呈现多个 vCPU，可以提升物理核 CPU 利用率。虚拟机的一个 vCPU 对应主机上的一个线程，一个主机上的虚拟机的总的 vCPU 数可以远远大于主机的物理 CPU 数量，但是当 vCPU 数量比物理 CPU 数量大到一定程度后，就可能出现物理 CPU 繁忙，虚拟机的 vCPU 较长时间得不到调度而出现卡顿等情况。当 CPU 复用比超过 3:1 后，就可能造成主机 CPU 利用率过高，造成性能影响。

图3-3 CPU 复用示意图

-	时间片1	时间片2	时间片3	时间片4
vCPU1	running	wait	wait	wait
vCPU2	wait	running	wait	wait
vCPU3	wait	wait	running	wait
vCPU4	wait	wait	wait	running
vCPU5	wait	wait	wait	wait

超分建议

- 资源复用比过高，进而造成主机 CPU、内存利用率过高，异常首先表现在虚拟机业务上。客户首先需要观察自己的业务有没有异常，例如部署的 Web 服务器有没有访问超时现象，或者办公电脑打开文档是否变慢无法忍受等。

  此外过高的 CPU、内存复用比，还可能造成虚拟机卡死，虚拟机 crash，虚拟机时钟变慢问题，虚拟机性能下降，虚拟机死锁等。可以观察 FusionOne

 版权所有 © 超聚变数字技术有限公司

Compute 界面上有没有“主机 CPU 占用率超过阈值”或“主机内存占用率超过阈值”等异常告警。

大部分故障可以通过迁移或关闭部分虚拟机，降低整体 CPU 利用率恢复。当前已知的不可恢复故障主要是虚拟机 crash 引起的虚拟机镜像损坏（类似物理主机掉电或蓝屏导致磁盘数据有异常），如果损坏后修复失败，可能引起客户数据丢失。不建议核心业务超过 CPU 复用比。

- 超过商用规格复用比后，各个厂商问题支撑策略如何?

  虚拟化厂商都规定了商用 vCPU 复用比，例如 VMware 承诺商用规格为 3:1，超过这个规格出问题后一般建议降低 CPU 复用比后复现。

- 当前环境物理 CPU 利用率很低，是否可以增加 CPU 复用比?

  当 vCPU 可以运行但是由于 CPU 复用的原因而得不到调度的时候，就会表现为调度时延。随着 CPU 复用比的增加，调度时延也会随之增加。这里有两个概念：调度时延增加的概率；调度时延增加的程度。调度时延增加的概率和 vCPU 利用率有关系，假设 CPU 复用比 3:1，三个 vCPU 利用率都是 10%; 则微观上，在 1 秒内，每个 vCPU 需要运行 100mS，很明显，vCPU 利用率越大，三个 vCPU 都需要运行而撞在一起的概率就会变大。当 CPU 复用比 3:1 时，最好的情况下，最差的那个 vCPU 需要等待 2 个时间片。当 CPU 复用比 5:1 时，最好的情况下，最差的那个 vCPU 需要等待 4 个时间片。很明显，随着复用比的增加，最小调度时延也会增加。这个调度时延和 CPU 利用率没有直接关系。

  如上讨论，调度时延的增加，会导致内核定时器不准而表现为各种异常。因此，建议不要随意增加 CPU 复用比，尽管物理 CPU 利用率较低。

- 超过 3:1 的复用比后，超聚变是否可以给出不同复用比问题发生概率?

  CPU 复用比大于 3:1,是否会对业务产生影响，取决于 Guest OS 本身及虚拟机内部业务特点。虚拟机内某些业务是时延不敏感的，例如文件读写，时延增加后会进入等待队列，最终仍然会成功。但是视频业务，则可能导致视频卡顿或花屏，客户感知。即使是某些内核 bug，也是在某些业务场景下很容易触发，另外的场景下很长时间也触发不到。

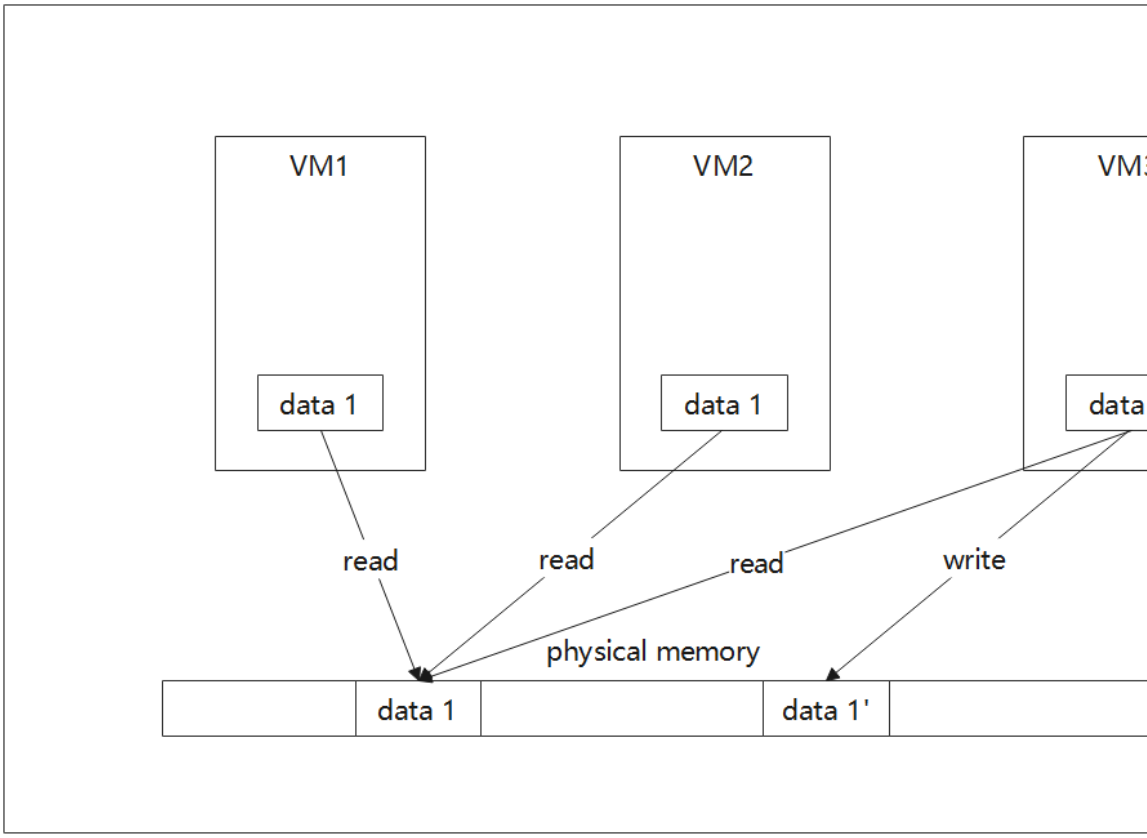
内存超分

技术原理

内存复用是指在服务器物理内存一定的情况下，通过综合运用内存复用技术（内存共享、内存气泡、内存置换）对内存进行复用。通过内存复用，使得虚拟机内存规格总和大于服务器规格内存总和，提高服务器中虚拟机密度。智能内存复用可提升内存资源的利用率，帮助用户节省内存采购成本，延长物理服务器升级内存的周期。

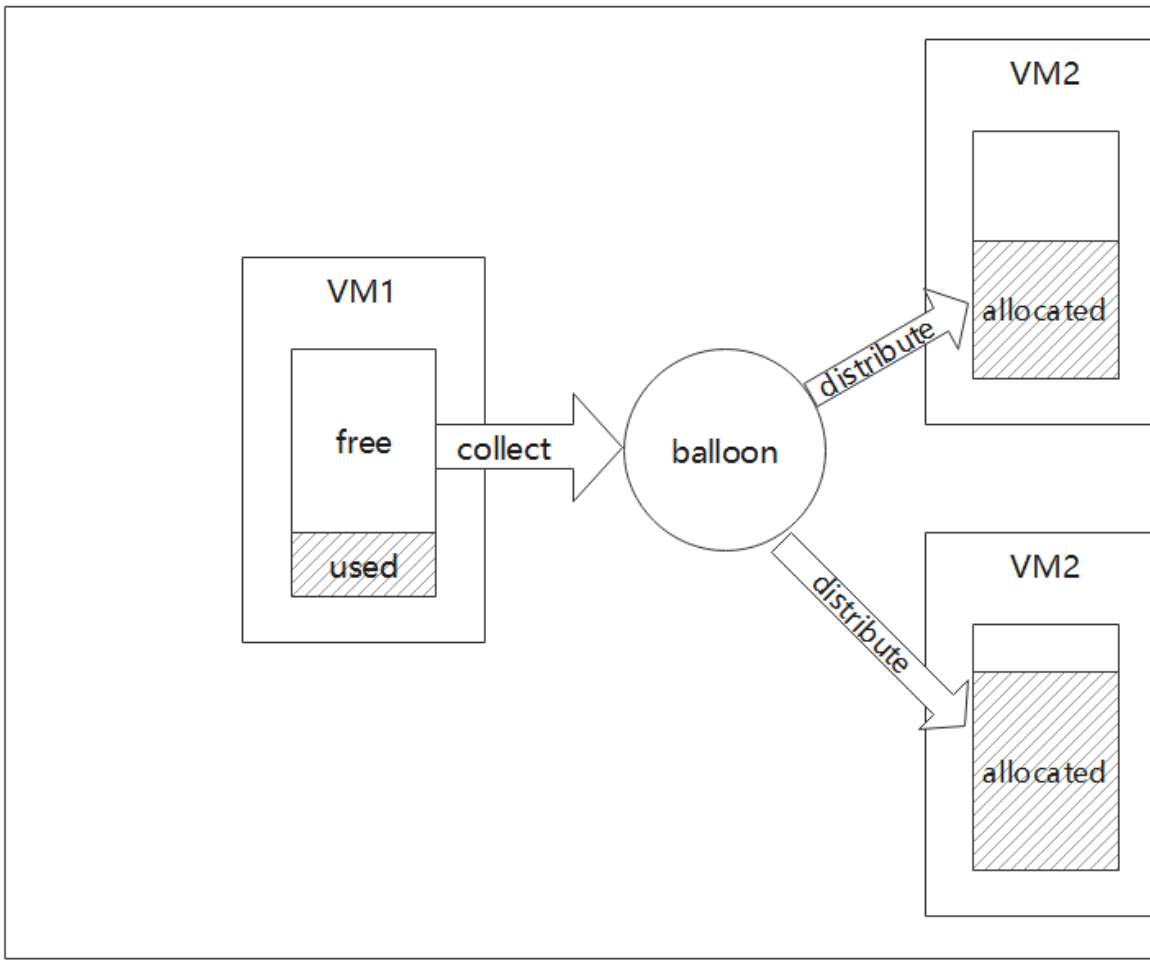
 版权所有 © 超聚变数字技术有限公司

1. 内存共享：多台虚拟机之间共享数据内容相同的内存页，此部分内存空间仅被虚拟机做只读操作，在需要写入时将通过写时复制操作，开辟新的内存空间，并修改内存映射，在 KVM 中通过 Linux 内核能力 KSM 实现；



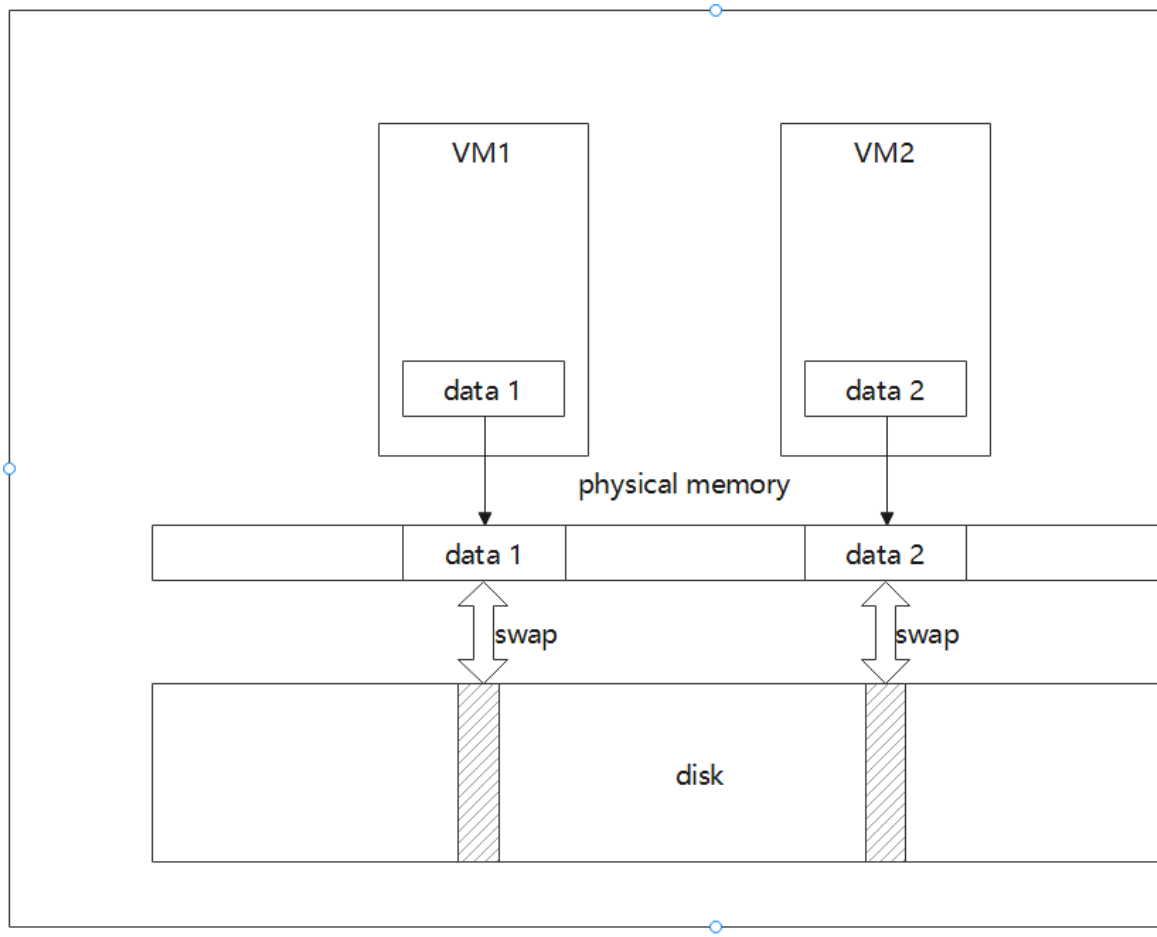
2. 内存气泡：系统主动回收虚拟机暂时不用的物理内存，分配给需要复用内存的虚拟机。内存的回收和分配均为系统动态执行，虚拟机上的应用无感知。

 版权所有 © 超聚变数字技术有限公司



3. 内存置换：将虚拟机长时间未访问的内存内容被置换到存储盘，即主机 Swap 分区；和 Swap 分区的标准使用场景类似，对冷数据会置换到磁盘存储中，系统需要使用这些数据时，再与预留在内存上的数据进行交换。

 版权所有 © 超聚变数字技术有限公司



超分建议

1、当内存已处于超配状态且主机内存占用率达到 70%，主机上承载的虚拟机业务是高内存消耗型，不建议开启内存复用。这种情况下如果开启内存复用，将大概率导致主机内存不足并通过内存交换策略产生空闲内存，进而导致非全内存预留的虚拟机性能较差。

2、为了均衡内存交换对本地磁盘的 IO 压力并使内存复用效果达到最佳，建议将 SSD 等性能较好的本地磁盘作为交换分区使用。

3、内存复用率建议控制在 150%以内。当内存复用率超过 120%会上报重要级别告警，当虚拟机实际使用内存总量大于虚拟化域内存和 Swap 空间之和的 90%时，会上报紧急级别告警。

4、开启内存复用时进行热迁移，若虚拟机部分内存位于内存交换磁盘，会导致迁移时间较长。

 版权所有 © 超聚变数字技术有限公司

资源复用上限限制

FusionOne HCI 支持对单个主机或集群配置复用率上限，CPU 支持复用率上限设置为 101% ~ 500%，内存支持复用率上限设置为 101%~150%，在不配置复用率上限情况下，CPU 将不限制复用率，内存则按照最高 150%的复用率限制。

根据业务积累的经验，当 CPU 复用率超过 300%，内存复用率超过 120%会触发告警，让用户及时了解到系统资源的不足的问题，提前准备扩容做好业务规划。

3.4 虚拟机资源动态热添加(DRX)

业务虚拟机如果前期规划的资源不充足，或者随着业务量的增加，导致规划资源无法满足业务需求，运维人员无法及时发现，无法及时添加资源，从而导致业务因为 CPU、内存资源不足导致业务受到影响。动态 CPU、内存的热添加可以保证业务在运行资源突发进展的情况下持续运行。当启用动态资源热添加功能后，系统根据实时获取的虚拟机 CPU、内存使用率，与预先配置的阈值以及扩容策略，动态添加 CPU、内存资源。动态热添加功能能够非常有效的利用主机资源，并通过自动配置策略，提升管理员的运维效率。

主要实现原理：

系统实时监控虚拟机的内存、CPU 使用率，当超过阈值一定时间后，通过虚拟机的 CPU、内存热添加的方式，自动扩容虚拟机的规格，临时将虚拟机的 CPU、内存使用率下降到阈值（默认为 80%）以下。

图3-4 DRX 配置



 版权所有 © 超聚变数字技术有限公司

3.5 虚拟机热迁移

虚拟机热迁移（在线迁移）是指在不中断业务的情况下，将同一个集群中虚拟机从一台物理服务器移动至另一台物理服务器。热迁移虚拟机需要完成虚拟机如下操作：

- 虚拟机的实时内存数据。
- 虚拟机的磁盘镜像迁移或共享切换。
- 虚拟机在新的主机激活运行，实现虚拟机的平滑运行，用户察觉不到任何变化。

虚拟机热迁移主要用于如下几个场景：

- 在进行服务器操作维护前，系统维护人员将该服务器上的虚拟机迁移到其他服务器，降低操作维护过程中业务中断的风险。
- 在进行服务器升级操作前，系统维护人员将该服务器上的虚拟机迁移到其他服务器，升级完成后将所有虚拟机迁回，降低服务器升级过程中业务中断的风险。
- 将空闲服务器上的虚拟机迁移到其他服务器，将没有负载的服务器关闭，降低业务运行成本。
- 某些主机上业务压力过大，需要均衡各主机的业务负载。

虚拟化平台支持 IMC 特性（仅 Intel 架构支持），主要原理是开启 IMC 特性后，集群内的虚拟机采用特定类型的 vCPU，使得虚拟机的指令、功能集在不同物理 CPU 型号上均可运行，即使这些主机的实际 CPU 不同，也不会因 CPU 不兼容而导致迁移虚拟机失败。

根据业务类型的不同，虚拟化平台提供手动迁移、自动迁移的能力，如下表所示。

表3-1 迁移说明

迁移类型	子类	说明
手动迁移	按目的迁移	系统维护人员通过 FusionOne Compute 的虚拟机迁移功能，手动迁移一台虚拟机到另一台服务器上。
自动迁移	虚拟机资源调度	在同一个集群内，系统根据预先设定的虚拟机调度策略，对虚拟机进行自动迁移。

针对热迁移的存在如下两个场景：

- 跨主机热迁移

 版权所有 © 超聚变数字技术有限公司

虚拟机热迁移过程仅仅改变虚拟机运行的主机，虚拟机的外存不变。主要应用于共享存储场景，如 VIMS 卷的虚拟机。

- 跨存储、跨主机热迁移

  存储热迁移可在存储虚拟化管理下的同一个存储设备内、不同存储设备之间进行迁移。热迁移使客户在业务无损的情况下动态调整虚拟机存储资源，以实现设备维护等操作。主要应用于不同主机之间无共享存储，如本地存储的虚拟机迁移。

超聚变虚拟化平台热迁移通过 vCPU 降频、QAT 加速技术，提升虚拟机的迁移效率。

虚拟机 vCPU 降频

通过虚拟机的 CPU QoS 调整，在迁移期间，通过不断的调整虚拟机的 vCPU 频率占用，减少虚拟机的调度次数，从而使得虚拟机的内存数据变化变慢。在迁移速度固定的情况下，通过减少虚拟机的内存变化数据，可以缩短虚拟机迁移的时长。

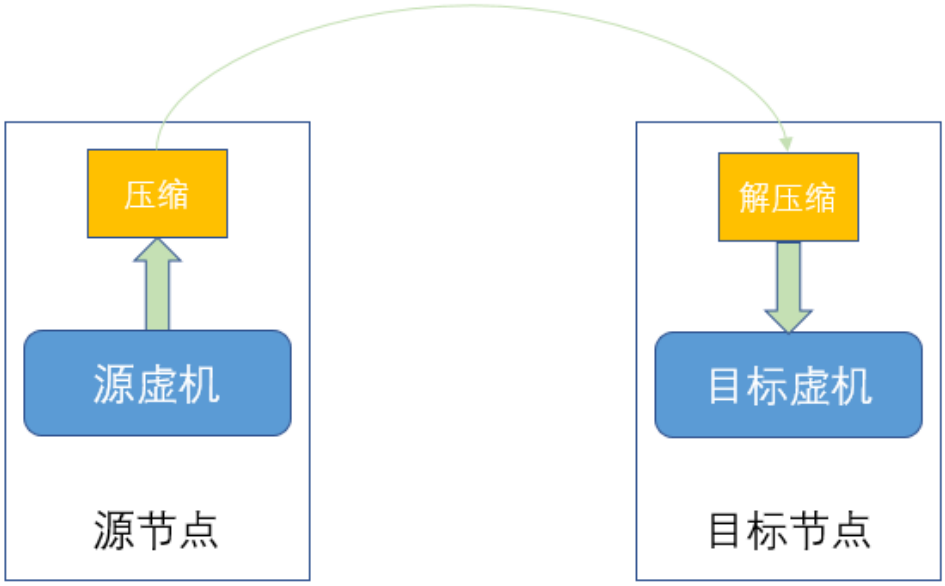
QAT 迁移加速

Intel 最新一代 CPU 集成了 QAT(Quick Assist Technology）设备，此设备可用于数据的压缩和解压缩，并且处理过程不会占用 CPU 的资源。对于 Compute 虚拟化平台而言，可以利用此“免费”的设备来对我们一些功能进行加速，充分释放硬件能力。典型的场景是虚拟机的热迁移。

虚拟机在热迁移过程中需要传输大量的内存数据以及迁移过程中产生的“脏页”，由于传输数据量大，并且要求传输速度比脏页产生速度快（否则会出现热迁移失败），这个过程对网络有较高要求，并且会与其他业务竞争带宽资源。某些业务场景，客户虚拟机内存页变化速度快的场景，传统迁移需要 1 小时以上或者一直无法迁移成功，启用 QAT 后，可以在 10s 内完成迁移，迁移效率明显提升。

传统的方案是在源端利用 CPU 对数据进行压缩，在目的端利用 CPU 对数据进行解压缩，从而大大减小需要传输的数据。但此方案需要使用较多的 CPU 资源。利用 CPU 集成的 QAT 设备来做迁移数据的压缩解压缩很好解决了对网络带宽占用，提高传输效率。同时又不占用 CPU 资源，解决了 Compute 平台中虚拟机热迁移占用资源多，迁移慢的问题，同时降低了虚拟机热迁移失败的概率。

 版权所有 © 超聚变数字技术有限公司



针对手动迁移虚拟机场景，虚拟化平台提供业务优先、迁移优先的两个迁移方式给用户选择，如下表所示。

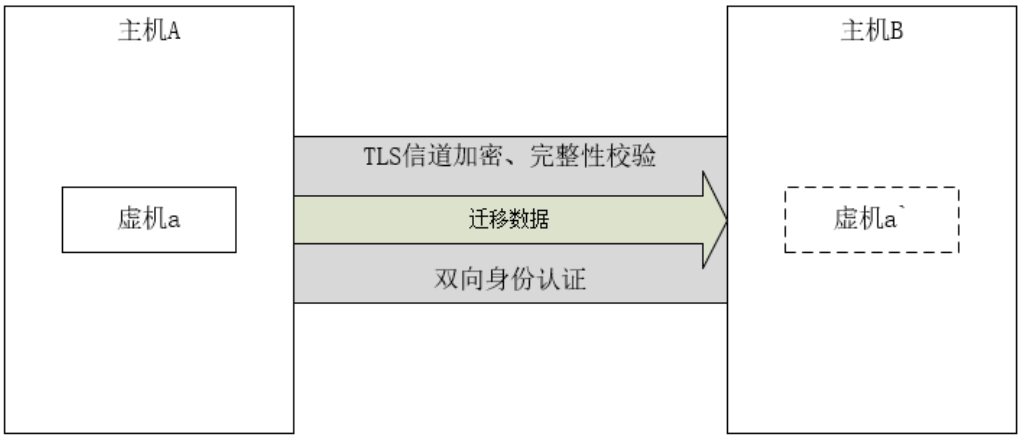
迁移方式	主要技术
业务优先	在全面保障业务的运行的情况下，尽可能进行迁移，但是可能会由于业务压力过大迁移失败。主要通过 QAT 加速，内存数据压缩提高迁移成功几率。 主要通过如下方式保证业务运行： • 虚拟机 vCPU 不降频，保证业务持续运行。 • 迁移带宽限制在一定百分比，不占用正常业务带宽。 • QAT 内存压缩加速，提升内存数据搬迁效率。
迁移优先	根据迁移过程中，虚拟机内存数据的变化率，实时调整业务虚拟机 vCPU 的资源，直到可迁移成功。 • QAT 内存压缩加速，提升内存数据搬迁效率。 • 虚拟机 vCPU 降频，减少虚拟机调度时长，减少待迁移的内存数据量。 • 迁移带宽不受限，加速内存数据迁移。

热迁移加密传输：

 版权所有 © 超聚变数字技术有限公司

普通迁移在数据传输过程为明文传输，由于普通迁移在用户组网条件较为复杂的情况下，可能存在中间人攻击或传输数据不完整等情况导致热迁移操作存在一定安全风险，超聚变虚拟化平台支持配置迁移信道加密，其内部使用了双向认证的 TLS 通信机制，实现了对传输数据机密性、完整性的保护，有效防止中间人攻击和窃听，在实际组网较为复杂的环境下能有效提高数据的安全性。

由于其内部实现基于 TLS 双向认证，因此在开启热迁移加密传输功能的时候，需要用户将证书进行手动上传根证书、客户端的证书与私钥、服务端的证书与私钥。



3.6 虚拟机快速克隆

FusionOne Compute 针对虚拟机模板转虚拟机，提供多种克隆方式，为虚拟机的复制根据需要提供更多的选择。

- 全量克隆：虚拟机完整的复制，将虚拟机磁盘通过完整复制数据卷，复制完成虚拟机才可以启动运行，虚拟机克隆耗时较长，但是完整复制不影响后续操作，如卷删除等。
- 快速全量克隆：虚拟机的卷通过快照方式复制，内部启动异步任务将卷进行卷的数据复制（用户不感知），虚拟机克隆速度快，但是在异步任务完成之前，不可以对虚拟机的卷进行删除等操作。
- 链接克隆：虚拟机的卷通过快照方式复制，虚拟机克隆速度快，但是虚拟机新增数据通过增量卷管理，母卷是共用的，主要用于空间紧张并且性能要求不高的场景，如呼叫中心办公、轻量办公、共用计算机等。

 版权所有 © 超聚变数字技术有限公司



3.7 虚拟机回收站

FusionOne Compute 支持临时存放用户通过安全删除/普通删除执行的放入回收站中的虚拟机，存放在回收站的虚拟机可以恢复，便于虚拟机的日常维护和使用，在误删除之后可以通过回收站找回。

在删除虚拟机的时候，可以选择 xx 天之后删除，虚拟机不会立即删除，而是在用户选择的天数（1~365 天，默认为 30 天）之后进行自动回收删除。用户也可以进入回收站页面，对待删除的虚拟机进行恢复，或者在回收站彻底删除。

虚拟机被删除到回收站后，虚拟机只是处于待删除状态，处于关机状态，无法被启动，实际虚拟机的卷、配置并没有被立即删除。

3.8 VMware 虚拟机模板导入

FusionOne Compute 支持导入 Vmware 的 OVF 和 OVA 格式的虚拟机模板，其原理如下：

- 按要求从 Vmware 中把虚拟机导出为 OVF 或者 OVA
- 在 FusionOne Compute 中选择此 OVF 或者 OVA 进行导入
- FusionOne Compute 对配置自动识别和转换
- 对于无法自动转换的配置，提示用户进行选择
- FusionOne Compute 执行磁盘格式转换、驱动替换和注入
- 在 FusionOne Compute 中启动此虚拟机进行验证

 版权所有 © 超聚变数字技术有限公司

3.9 NUMA 调优

NUMA 系统，它是一种多处理器环境下设计的计算机内存结构。NUMA 系统是从 SMP（Symmetric Multiple Processing，对称多处理器）系统演化而来。SMP 系统有如下特点：

- 所有的硬件资源都是共享的。即每个处理器都以相同的方式访问到任何内存、外设等。
- 所有的处理器都是平等的，没有主从关系。
- 内存是统一结构、统一寻址的（UMA, Uniform Memory Architecture）。
- 处理器和内存，处理器和处理器都通过一条总线连接起来。

SMP 的问题也很明显，因为所有的处理器都通过一条总线连接起来，因此随着处理器的增加，系统总线成为了系统瓶颈。为了克服以上的缺点，才应运而生了 NUMA 架构。在这个架构中，每个处理器都有本地的内存和总线，如 PCIE、SATA、USB 等。处理器和本地内存之间拥有更小的延迟和更大的带宽，而整个内存仍然可作为一个整体，任何处理器都能够访问，只不过跨处理器的内存访问的速度相对较慢一点。和内存一样，处理器访问本地的总线延迟低，吞吐率高；访问远程资源，则延迟高，并且要和其他处理器共享一条总线。

FusionOne Compute 两种技术来提升虚拟机的性能，Guest NUMA 和主机 NUMA。

Guest NUMA 是通过向虚拟机呈现物理 CPU NUMA 的结构，使 Guest OS 及其内部应用能正确识别 vCPU 对应的物理 NUMA 结构，从而实现 vCPU 尽量访问本地内存的目的。

Host NUMA 通过智能调度和绑定等技术，自动把 VM 的 CPU 和内存资源分配在同一个 node 上，并对主机 node 间的 CPU 资源做负载均衡。通过这两种技术可以极大提升虚拟机的性能。

3.10 动态资源调度（DRS&DPM）

动态资源调度 DRS（Dynamic Resource Scheduler），指采用智能负载均衡调度算法，并结合动态电源管理功能，通过周期性检查同一集群资源内各个主机的负载情况，在不同的主机间迁移虚拟机，从而实现同一集群内不同主机间的负载均衡，并最大程度降低系统的功耗。

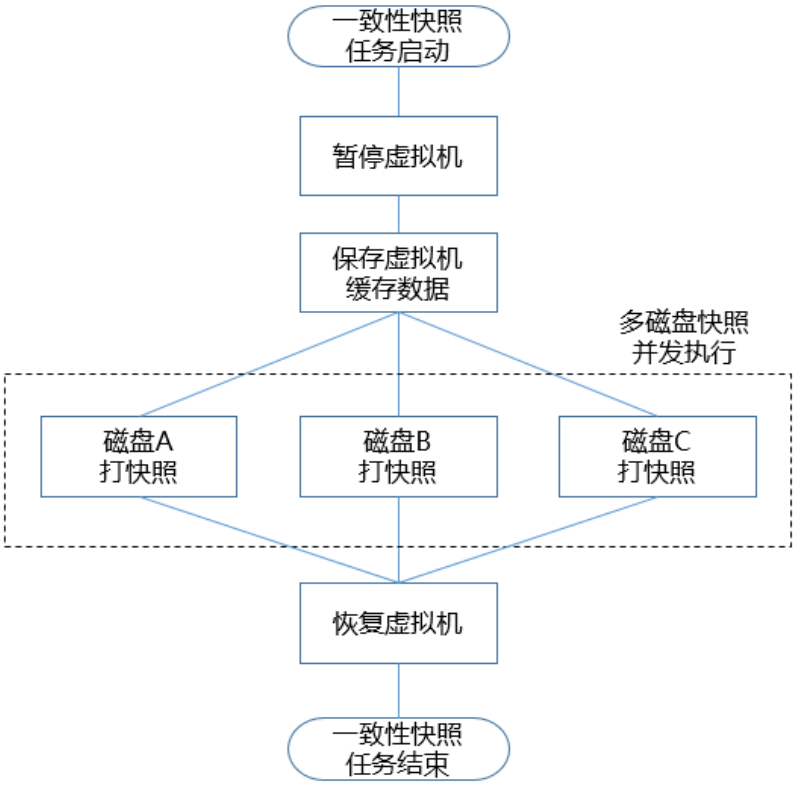
 版权所有 © 超聚变数字技术有限公司

- 系统轻载时，将迁移部分虚拟机，并将其集中在部分物理主机，随即将空闲主机下电。
- 系统重载时，将启动部分物理主机，并将虚拟机均衡分布在各主机中，以保证资源的供应和用户的体验。
- 通过计划任务，可根据系统运行的情况，分时段采取不同的资源调度策略，以满足不同场景的用户需求。

3.11 快照与快照策略

FusionOne Compute 支持创建快照策略，实现对业务虚拟机的定期的保护，在业务虚拟机出现逻辑性错误时，可以使用快照恢复虚拟机，降低业务损失。

- 快照类型：支持创建普通快照或者内存快照，默认是普通快照。
  - 普通快照：快照创建时，只保存虚拟机当前的磁盘数据，不保存缓存数据，也不保存虚拟机的内存状态及数据。
  - 内存快照：快照创建时，将虚拟机当前内存状态及数据保存至快照中。
  - 一致性快照：快照创建时，会先将虚拟机的缓存数据保存到磁盘中，然后再创建快照，可以保证虚拟机磁盘的数据一致性。



 版权所有 © 超聚变数字技术有限公司

一致性快照需要暂停虚拟机，进而可能对业务产生影响，所以暂停时间越短越好。为此采用了如下优化手段，缩短暂停时间：

1. 快照任务优先：普通任务需要进入任务队列，由任务框架依次调度执行，而快照任务可以抢占资源、优先执行。

2. 简化快照任务：一个快照任务完成端到端所有操作，避免打快照过程中的数据传递和任务调度。

3. 多磁盘并发：虚拟机有多个磁盘，对这些磁盘并发打快照。

单磁盘虚拟机快照的暂停时间约为 3s，每增加一个磁盘，暂停时间增加不超过 1s。具体暂停时间与其内部业务压力大小、虚拟机操作系统类型、磁盘个数等等因素都有关，建议在磁盘 IO 压力较小时创建快照。

- 快照策略功能：
  - 快照保留策略：可设置快照的最大保留个数，默认为 7。支持 1-32 的按需配置，超过保留个数的快照将自动删除(用户手动创建的快照不受快照保留策略控制)。
  - 定时创建快照：可设置定时自动创建快照的频率，支持每月、每周、每天的按需配置。
  - 立即执行：创建快照策略后，可支持用户按需立即执行创建快照。立即执行创建的快照，受快照保留策略控制。
  - 关联虚拟机：可为快照策略关联要作用的业务虚拟机。同一个策略支持关联多个业务虚拟机。
- 快照策略管理：支持修改、删除已创建的快照策略。

3.12 XPU 可视化管理与调度策略

须知

本小节特性需要配套 FusionOne Center。

GPU 可视化：

GPU 可视化是利用 nvidia-smi 或 npu-smi 管理工具，进行可视化数据采集，包括 GPU 使用率、内存使用率、温度、功耗等一系列常用的监控信息，并通过虚拟化引擎的监控系统形成 KPI；可视化数据采集来源分为直通与虚拟化两种场景：

 版权所有 © 超聚变数字技术有限公司

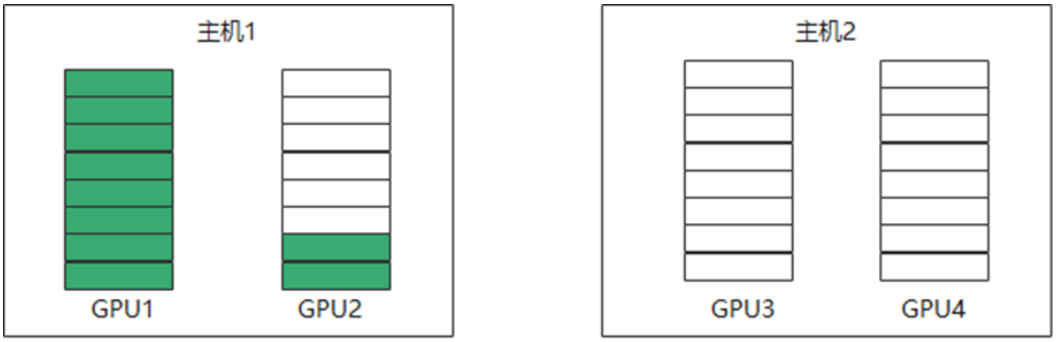
- 直通场景：GPU 直通的情况下，由于已直通到虚拟机的 PCI 设备无法在宿主机上被识别与管理，此时监控数据采集依赖于 GuestOS 带内采集获取，需要安装 vmtools 以进行主机与虚机的通信。
- 虚拟化场景：GPU 虚拟化依赖于 GPU 的驱动能力，并受主机 GPU 驱动管理，因此主机上可获取主要的 GPU 监控信息；部分在 vGPU/vNPU 运行时产生的性能指标，如 vGPU 的内存使用等信息则通过 GuestOS 带内通道进行采集。

可视化信息采集插件通过监控系统分发，定时按照直通/虚拟化、GPU/NPU 等不同维度以不同规则进行下发，采集到的数据上报至管理面，经过数据分析、聚合、持久化，最终在界面呈现。

GPU 调度：

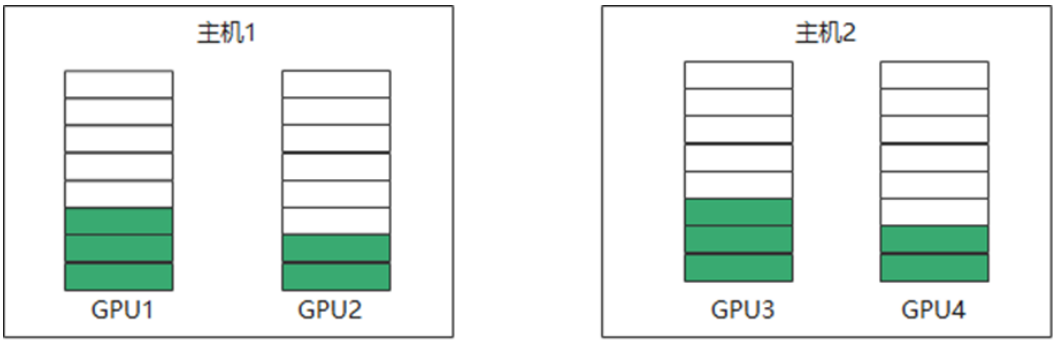
GPU 调度分为虚拟机调度与 GPU 资源分配调度。虚拟机调度即按照一定策略在虚拟机开机或 HA 的过程中，调度到指定的主机上，如果虚拟机使用的是 GPU 虚拟化组，则还支持将对应虚拟机资源调度到指定的 GPU 卡上。目前 GPU 虚拟化支持以下两种调度模式：

- 密度模式：密度模式需要保证尽可能支持启动更多的带 GPU 资源的虚拟机实例，要求虚拟机尽可能使用已在使用中的资源；在直通模式下，密度模式尽可能调度到已使用直通 GPU 的主机上，在虚拟化模式下，密度模式尽可能调度到分配 GPU 较多的主机时，进一步将 vGPU 分配到已有 vGPU 的卡上。
- 性能模式：性能模式需要保证尽可能让 GPU 虚拟机均匀负载在所有主机或 GPU 卡上，要求虚拟机尽可能使用未被分配的资源；在很多情况下比密度模式会有更好的性能表现；虚拟化模式下，性能模式尽可能将 GPU 计算虚机平均分配到 GPU 计算资源上。



密度模式

 版权所有 © 超聚变数字技术有限公司

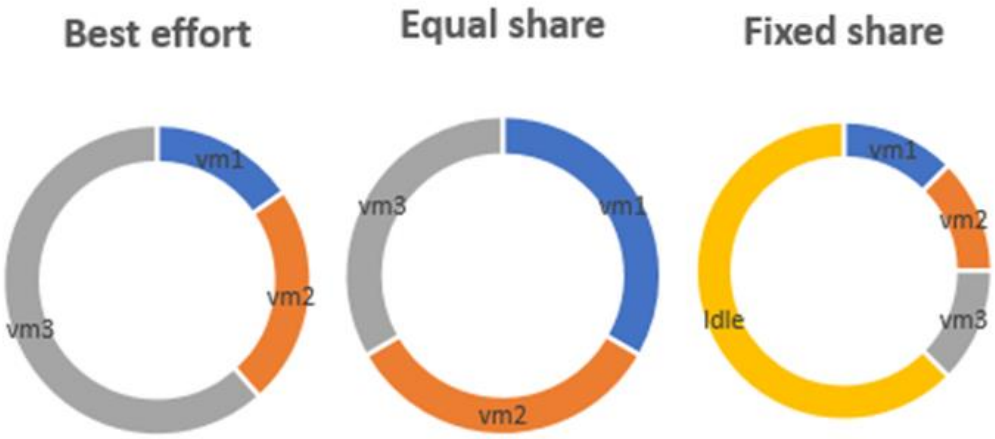


性能模式

GPU 资源分配调度则控制 GPU 在虚拟化模式下，资源分配按照一定的策略进行时间片的分配，仅 nvidia 的 GPU 支持。当前 GPU 资源分配调度支持配置以下策略：

- 抢占（Best Effort）：在抢占策略下，任务会尽可能地使用可用的 GPU 资源，但不保证一定能获得特定数量的资源。
- 等分（Equal Share）：在等分策略下，该 GPU 将会把资源平等分配给所有使用的 GPU 虚拟机。随 GPU 虚拟机启动的实例数量变多，每个虚机获得的算力配比会下降。
- 固定配比（Fixed Share）：在固定配比下，每个任务或用户在开始时就被分配了固定数量的 GPU 资源，每个 GPU 虚拟机分得的算力不会随 GPU 虚拟机启动的实例增加而动态变化。

下图是一个 GPU 切换为 8 个 vGPU 场景下不同调度策略的效果，假设实际分配运行的为 3 个 vGPU，分别给到 3 个虚拟机。等分方式获取的资源是实际运行 vGPU 划分时间片，而固定配比是根据切换数量划分的时间片。



 版权所有 © 超聚变数字技术有限公司

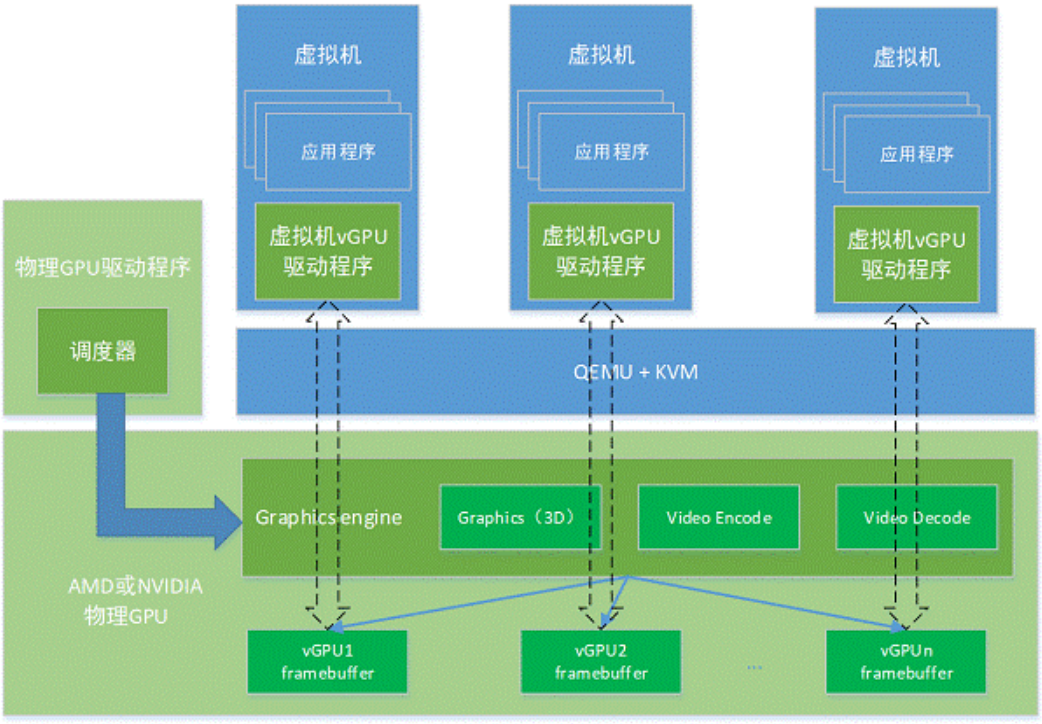
3.13 GPU 虚拟化

vGPU（Virtual GPU）是针对英伟达 GPU 的虚拟化环境中提供虚拟显卡的技术，通过在宿主机上运行虚拟化软件，在虚拟机中模拟出一个独立的显卡设备，让虚拟机可以直接访问显卡资源，从而实现高性能图形加速。通过 vGPU 技术，可分时复用 GPU，充分利用 GPU 的算力。

每个 vGPU 都类似于物理 GPU，有固定的显存大小，一个或者更多的虚拟显示输出。**vGPU 创建时就从物理 GPU 中显存分配出来，并且是独占的（GPU 引擎是共享的）**，每当轮询到某个 vGPU 的任务时，可占用整个 GPU 的算力资源。

vGPU 跟传统的 GPU 类似，具有固定数量的 GPU 帧缓冲器和一个或多个虚拟显示输出。在 vGPU 虚拟机被创建的时候，帧缓冲区就被分配到物理 GPU 的帧缓冲区中，vGPU 保留对该缓冲区的独占使用，直到它被销毁为止。留在物理 GPU 上的所有 vGPU 虚拟机共享对 GPU 引擎的访问，包括图形（3D）、视频编解码引擎等。

vGPU 技术解决了在虚拟化环境中无法满足图形处理需求的问题。在传统虚拟化环境中，虚拟机使用的是宿主机的物理显卡，虚拟机的图形性能和宿主机的性能有很大关系。如果宿主机的显卡性能较低，那么虚拟机的图形性能也会很差。而 vGPU 技术则让虚拟机可以独立使用虚拟显卡，不受宿主机的限制，从而提高了虚拟机的图形性能。



FusionOne Compute 提供对多种 GPU 卡的虚拟化支持，通过将一块 GPU 虚拟化为多块虚拟 GPU，在虚拟机中可以使用和直接硬件访问 GPU 一样的方式进行计算，并

 版权所有 © 超聚变数字技术有限公司

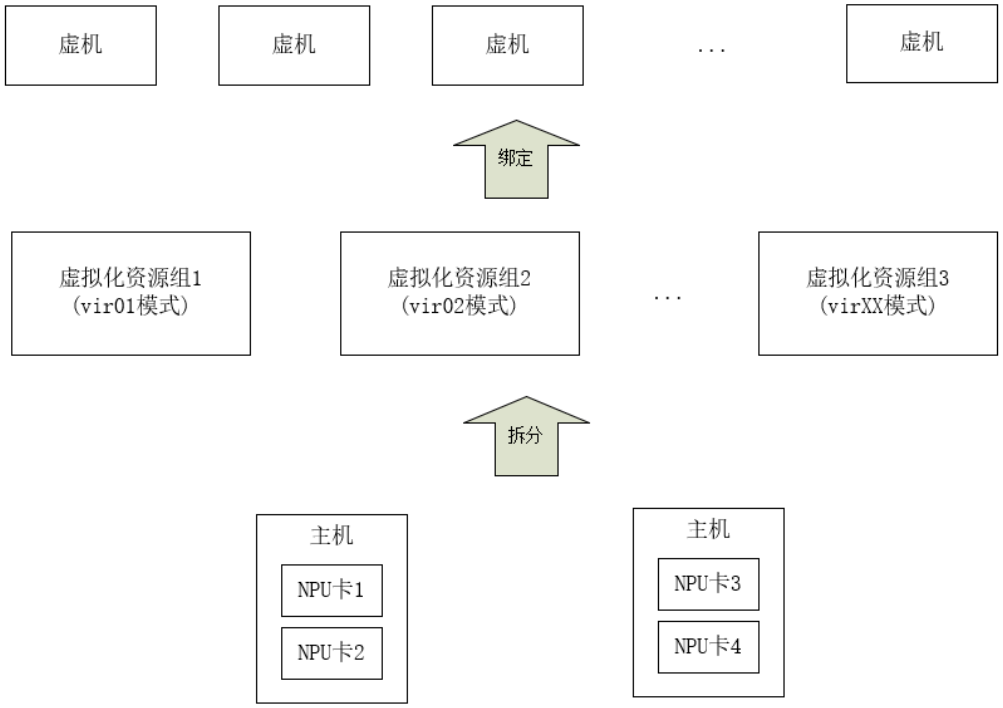
提供较高的隔离性、安全性和通用性，能够满足虚拟化环境下不同用户的需求，并充分利用宿主机上的 GPU 资源，提高整体的计算能力和性能。

3.14 NPU 虚拟化

NPU 虚拟化是针对华为 Ascend 系列 AI 加速卡的虚拟化方案，与 GPU 虚拟化不同，**通过空间划分的方式进行隔离的虚拟化，相比 vGPU 虚拟化，虚拟机之间的资源隔离性好。**NPU 虚拟化切分成若干份 vNPU（Virtual NPU），虚拟化管理方式能够实现统一不同规格资源的分配和回收处理，满足多用户反复申请/释放的资源操作请求。

NPU 虚拟化功能适用于多虚机多任务并行，且每个任务算力需求较小的场景。对算力需求较大的大模型任务，建议使用 NPU 直通。

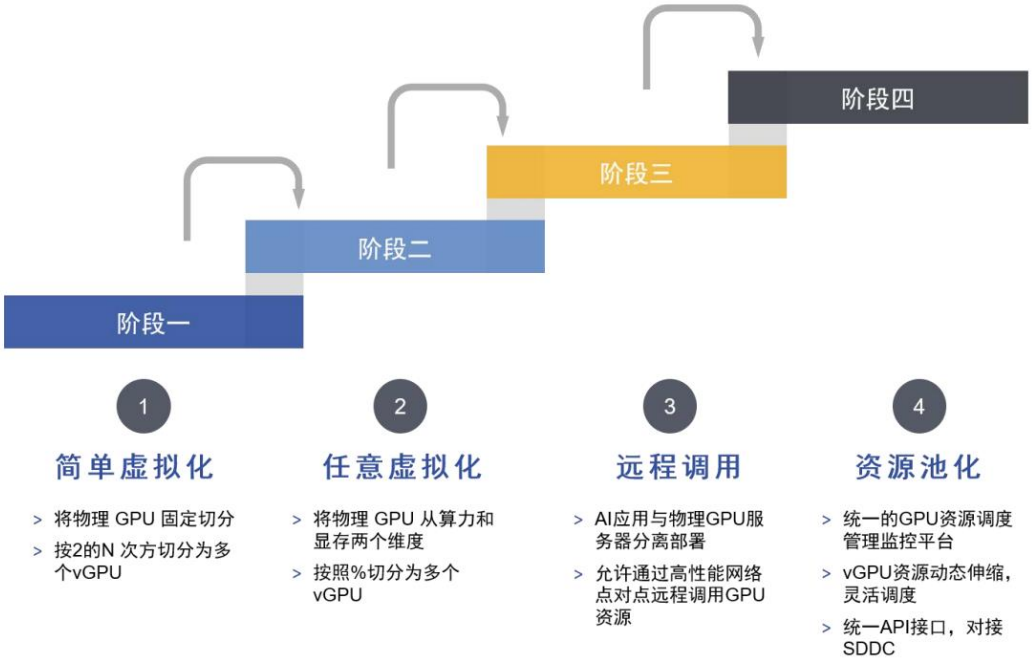
NPU 虚拟化开启需要在主机上安装 NPU 相关驱动，通过配置 NPU 资源组，将特定虚拟化模式的 vNPU 资源添加到该资源池组，可以实现虚拟化资源按需分配到虚机中。同一个 vNPU 支持以不通虚拟化模式关联到不同的资源组，因此在使用中可以支持 vNPU 的灵活拆分。



 版权所有 © 超聚变数字技术有限公司

3.15 GPU 资源池

GPU 资源池化技术从初期的简单虚拟化，到资源池化，经历了四个技术演进阶段。



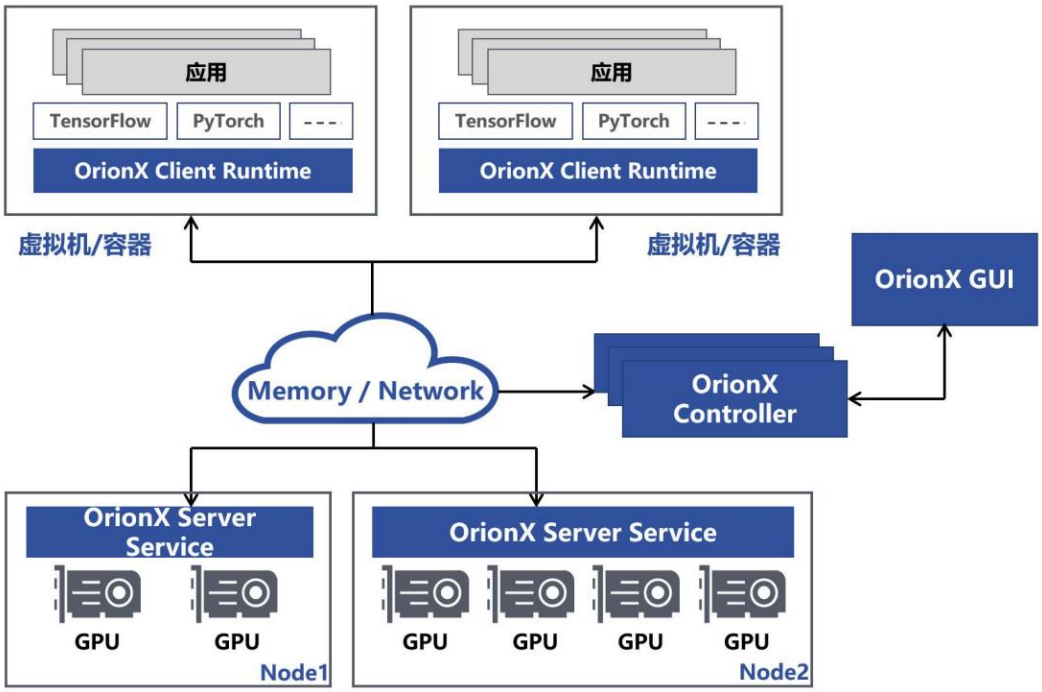
- **简单虚拟化。**将物理 GPU 按照 2 的 N 次方，切分成多个固定大小的 vGPU（VirtualGPU，虚拟 GPU），每个 vGPU 的算力和显存相等。实践证明，不同的 AI 模型对于算力、显存资源的需求是不同的。所以，这样的切分方式，并不能满足 AI 模型多样化的需求。
- **任意虚拟化。**将物理 GPU 按照算力和显存两个维度，自定义切分，获得满足 AI 应用个性化需求的 vGPU。
- **远程调用。**AI 应用与物理 GPU 服务器分离部署，允许通过高性能网络远程调用 GPU 资源。这样可以实现 AI 应用与物理 GPU 资源剥离，AI 应用可以部署在私有云的任意位置，只需要网络可达，即可调用 GPU 资源。
- **资源池化。**形成 GPU 资源池后，需要统一的管理面来实现管理、监控、资源调度和资源回收等功能。同时，也需要提供北向 API，与数据中心级的资源调度平台对接，让用户在单一界面，就可以调度包括 vGPU。

GPU 资源池化方案实现原理概述

通过软件定义算力，颠覆了原有的 应用直接调用物理 GPU 的架构，增加软件层，将应用与物理 GPU 解耦合。 应用调用逻辑的 vGPU，再由 OrionX 将 vGPU 需求匹配到具体的物理 GPU。OrionX 架构实现了 GPU 资源池化，让用户高效、智能、灵活地使用 GPU 资源，达到了降本增效的目的。

 版权所有 © 超聚变数字技术有限公司

一个典型的 GPU 资源池的逻辑架构中包含了 OrionX Controller（OC）、OrionXServer Service（OSS）、OrionX Client Runtime（OCRT）、和 OrionX GUI（OG）等功能组件。



OrionX Controller（OC）

OrionX Controller 是 GPU 资源池的核心管理调度模块，其他所有 OrionX 的功能组件都直接或者间接通过网络连接到 OrionX Controller，并与其保持信息同步。为了实现 OrionX GPU 资源池的统一管理以及资源调度，节点 IP 地址、物理 GPU 信息、虚拟 GPU 信息以及应用任务信息等都会汇总至该组件。

一个 OrionX GPU 资源池可以只部署一个 OrionX Controller。为了提高 OrionX 的可靠性，可以进行 2+1 冗余备份。OrionX Controller 提供如下功能：

- 各个分布式功能组件的服务注册、服务发现功能。
- 弹性虚拟 GPU 的调度分配功能。
- 多副本高可用的元数据存储和管理。
- License 管理。
- 提供运维所需要的各种 Rest API。

OrionX Server Service（OSS）

 版权所有 © 超聚变数字技术有限公司

OrionX Server Service 发现并管理物理节点上的 GPU 资源，同时把物理 GPU 的计算能力通过 OrionX 的高性能私有协议提供给数据中心内的各个物理节点，以及各个物理节点上的虚拟机、容器。

OrionX Server Service 部署在 OrionX 资源池内的每一个节点上，包括 GPU 节点和应用所在的节点。OrionX Server Service 提供如下功能：

- 发现和管理物理 GPU 资源。
- 把物理 GPU 资源抽象成弹性的 vGPU。
- 执行 AI 应用的 GPU 计算任务。
- 支持虚拟机、容器的网络隔离。

OrionX Client Runtime（OCRT）

OrionX Client Runtime 是一套兼容 Nvidia CUDA 编程环境的运行环境，模拟了 CUDA 的运行时接口。当 AI 应用在使用 Nvidia GPU 进行计算的时候，会自动调用 OrionX ClientRuntime。由于 OrionX Client Runtime 提供和 Nvidia GPU 兼容的 CUDA 接口，因此应用无需修改，可以透明无感知地运行在一个虚拟的 GPU 环境下。

OrionX Client Runtime 部署在每一个应用环境下，替代原有的 Nvidia CUDA SDK。OrionXClient Runtime 提供如下功能：

- 兼容 CUDA 接口。
- 自动完成虚拟 GPU 资源的申请、释放、弹性伸缩等功能。
- 支持虚拟机、容器和宿主机的网络隔离。

OrionX GUI（OG）

OrionX GUI 给运维提供一个友好的 GUI 界面，方便管理员对 OrionX 整体资源池进行全面管理。OrionX GUI 提供如下功能：

- 分级的运维账号登录与管理。
- 查看各组件的部署拓扑以及详情。
- 多维度、多角度查看资源池的资源使用情况。
- 细粒度管理资源池内的资源。
- 日志、监控及告警功能设置。

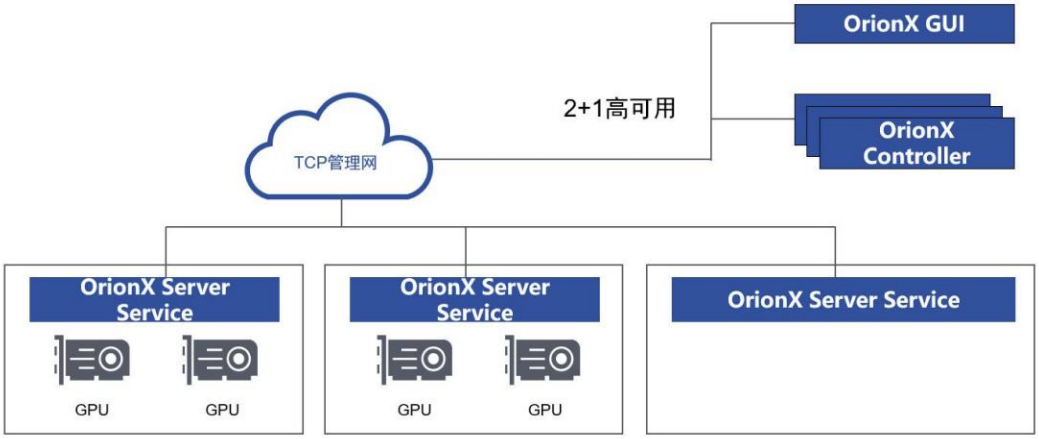
OrionX 组件间通信

OrionX 的各个功能组件通过管理平面网络和数据平面网络进行通信，共同完成 GPU 资源池的管理以及 GPU 资源的调度等功能。

 版权所有 © 超聚变数字技术有限公司

管理平面

在部署 OrionX 时，使用基于 TCP/IP 网络的管理平面，来承载整个系统的管理工作。通过管理网络，分布在各个节点的功能组件都保持和 OrionX Controller 同步。管理平面逻辑结构如下图所示。



通过私有的同步协议，分布式部署的各个功能组件具有如下特点：

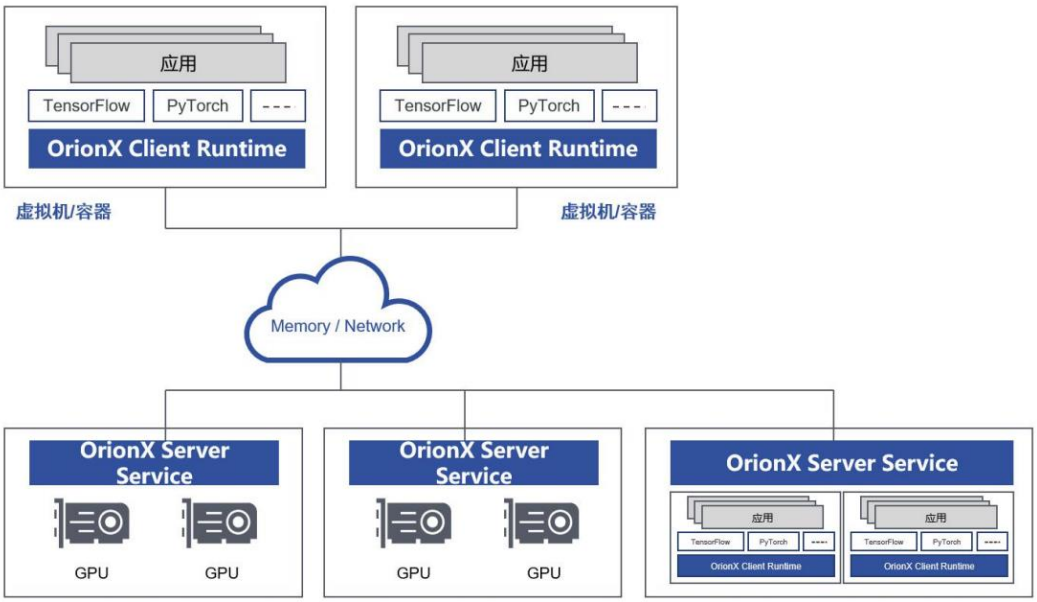
- OrionX Controller 支持多副本、高可用的部署模式。
- 各个功能组件启动的次序无要求。
- 当某一个功能组件从错误中恢复之后，可以自动同步到正确的状态。

数据平面

在应用运行的过程中，应用所在环境和 GPU 物理节点之间的数据传输使用的是 OrionX 的数据面。该数据面支持多种后端数据传输载体，包括 TCP/IP 以太网络、RoCE RDMA Infiniband RDMA、Share Memory 等。数据面具有如下的特点：

- 高带宽、低延迟。
- 同时支持多种传输协议，根据优先级自动使用高性能的传输方式。
- 支持虚拟机、容器和宿主机之间的 TCP/IP 网络隔离。

 版权所有 © 超聚变数字技术有限公司



部署形式

OrionX 支持 CentOS、Ubuntu、Debian 等 Linux 发行版本，同时支持基于 KVM 的虚拟机云平台和基于 Docker 的容器云平台。

OrionX 支持原生 KVM，各个组件都可以通 Binary 方式部署。在 KVM 环境，客户使用 OrionX 组件的安装脚本，就可以完成 OrionX 的基础部署，轻松实现 GPU 资源池化。

应用场景

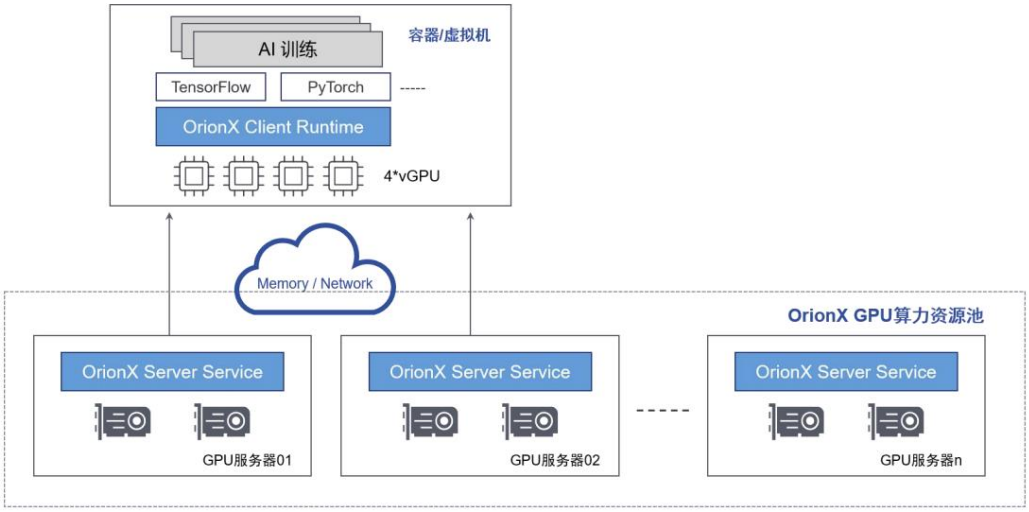
- **大模型场景的典型应用**

作为 AI 算力资源池平台，OrionX 既可以支持单台服务器上的单卡、多卡训练，也可以支持跨设备的多卡训练。

“化零为整”功能支持训练。OrionX 支持将多台服务器上的 GPU 提供给一个虚拟机或者容器使用，而该虚拟机或者容器内的基于分布式训练框架（Horovod 或 DistributedDataParallel）的 AI 应用无需修改代码。通过这个功能，用户可以将多台服务器的 GPU 资源聚合后提供给单一虚拟机或者容器使用。“化零为整”支持训练等大模型场景，为用户的 AI 应用提供数据中心级的海量算力。Horovod 是 Uber 开源的分布式深度学习框架，旨在使分布式深度学习变得快速且易于使用，使模型训练时间从几天和几周缩短到数分钟和数小时。使用 Horovod，可以将现有的训练脚本扩大规模，使其仅用几行 Python 代码就可以在跨设备的多个 GPU 上运行。一旦配置了 Horovod，就可以使用相同的基础结构来训练具有任何框架的模型，从而随着机器学习技术堆栈的不断发展，轻松地在 TensorFlow、PyTorch、MXNet 和将来的框架之间进行切换。DistributedDataParallel(简称 DDP)是 PyTorch 自带的分布式训

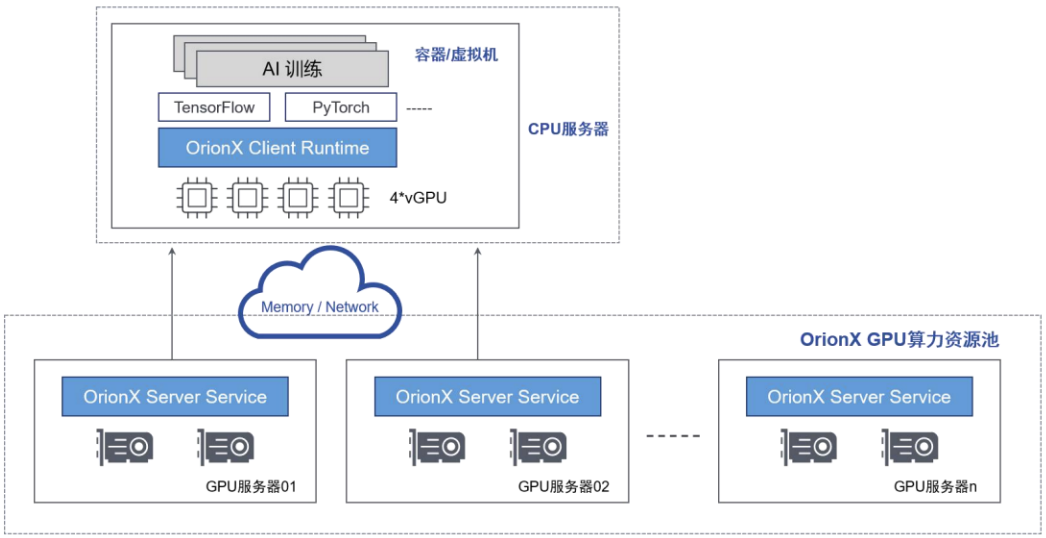
 版权所有 © 超聚变数字技术有限公司

练框架，支持多机多卡和单机多卡分布式训练。DDP 属于 Data Parallel，可以通过提高 batch size 来增加并行度。DDP 通过 Ring-Reduce 的数据交换方法提高了通讯效率，并通过启动多个进程的方式减轻 Python GIL 的限制，从而提高训练速度 。



“隔空取物”功能支持训练。

OrionX 支持将虚拟机或者容器运行在一台没有物理 GPU 的服务器上，通过计算机网络，透明地使用其他服务器上的 GPU 资源，“隔空取物”支持训练等大模型场景，既可以调取单台设备的多卡资源给容器或者虚拟机，实现类似单机多卡训练的场景；也支持调取多台设备的多卡资源给容器或者虚拟机，实现类似分布式多机多卡训练的场景。

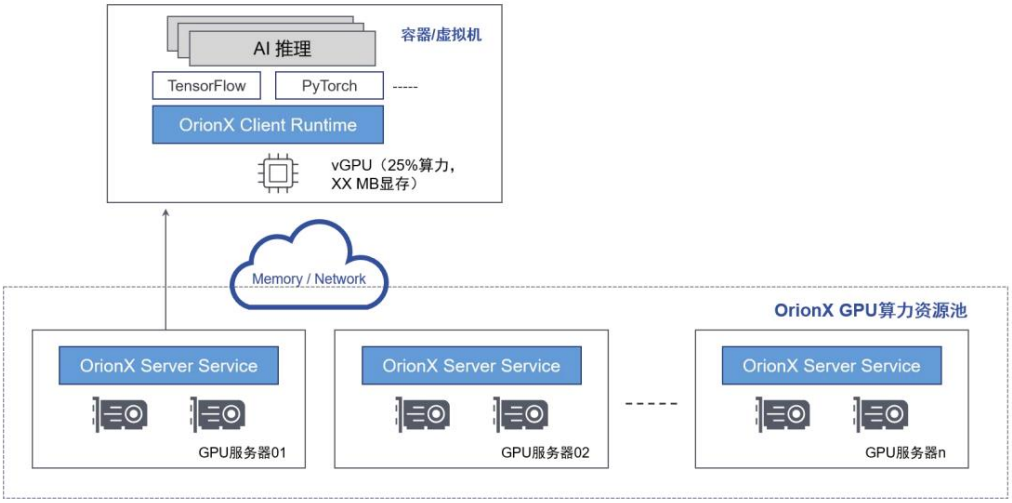


- **小模型场景的典型应用**

 版权所有 © 超聚变数字技术有限公司

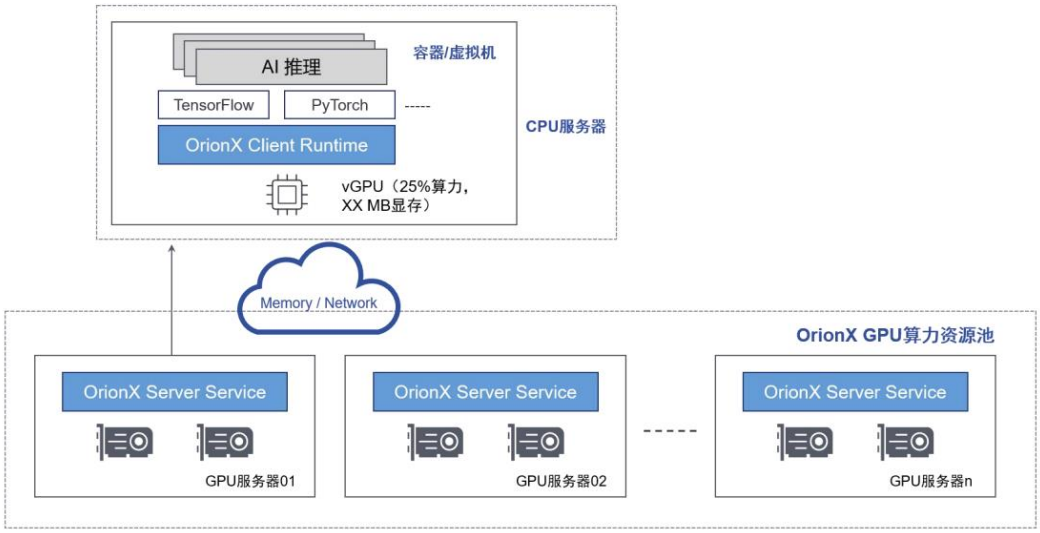
小模型如推理、开发、教学实训等场景，对算力资源需求量小，通常不能占满一张 GPU 卡资源。作为 AI 算力资源池平台，OrionX 可以从算力和显存两个维度，切分 GPU。支持将多个小模型任务调度到一张卡，有效提高资源利用率。

通过“化整为零”功能支持推理，将一块物理 GPU 细 粒度切分成多块 vGPU，每一块 vGPU 的显存和算力都能被独立设置和限制。算力切分的最小颗粒度为原物理 GPU 算力的 1%；显存切分的最小颗粒度为 1MB。



通过“隔空取物”功能支持推理。OrionX 支持将虚拟机或者容器运行在一台没有物理 GPU 的服务器上，通过计算机网络，透明地使用另一台服务器上的 GPU 资源，该虚拟机或者容器内的 AI 应用无需修改代码。通过这个功能，OrionX 帮助用户实现了 CPU 和 GPU 资源的解耦合。在一些 CPU 与 GPU 需要双向平衡的推理场景下，OrionX 可以更好的平衡 CPU 和 GPU 资源的分配，减少短板资源的发生，拉平数据中心的利用率。“隔空取物”支持推理、开发、教学实训等小模型场景，可以调取单台设备的细粒度卡资源给容器或者虚拟机，将多个小模型应用调用到一张物理 GPU 中，并严格限制 vGPU 资源间的隔离，实现资源利用率最大化。

 版权所有 © 超聚变数字技术有限公司



- **大/小模型场景的典型应用**

通过“随需应变”功能支持训练/推理， 支持用户在虚拟机或者容器的生命周期内，动态分配和释放所需要的 GPU 资源。通过这个功能，OrionX 帮助用户实现了真正的 GPU 资源动态伸缩，极大提升了 GPU 资源调度的灵活度。

OrionX 支持 vGPU 资源按需分配、随用随取，最大限度的利用算力资源。不论是大模型训练，还是小模型推理的环境中，用户都可以按照 AI 模型需求，动态的调整算力资源大小，而无需重启挂载 vGPU 的虚拟机/容器。OrionX 支持 vGPU 资源预留模式和获取模式：

预留模式：和使用物理 GPU 类似，客户申请的 vGPU 是独占的，不可被其他用户使用。

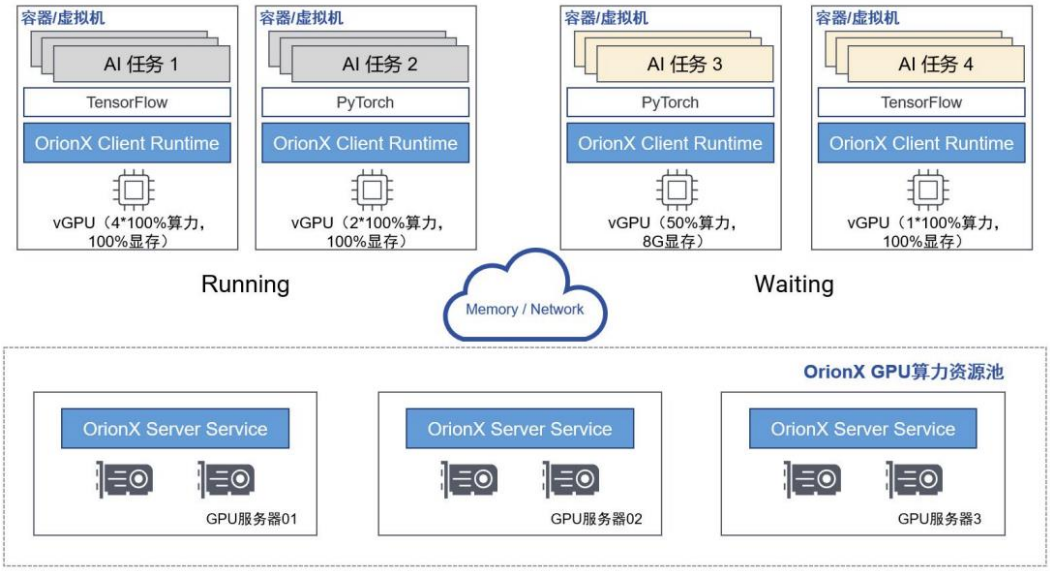
获取模式：客户申请的 vGPU 是动态的，只有在客户的 AI 应用运行时，vGPU 资源才锁定到具体的物理 GPU，一旦 AI 应用结束，物理 GPU 资源及时释放。



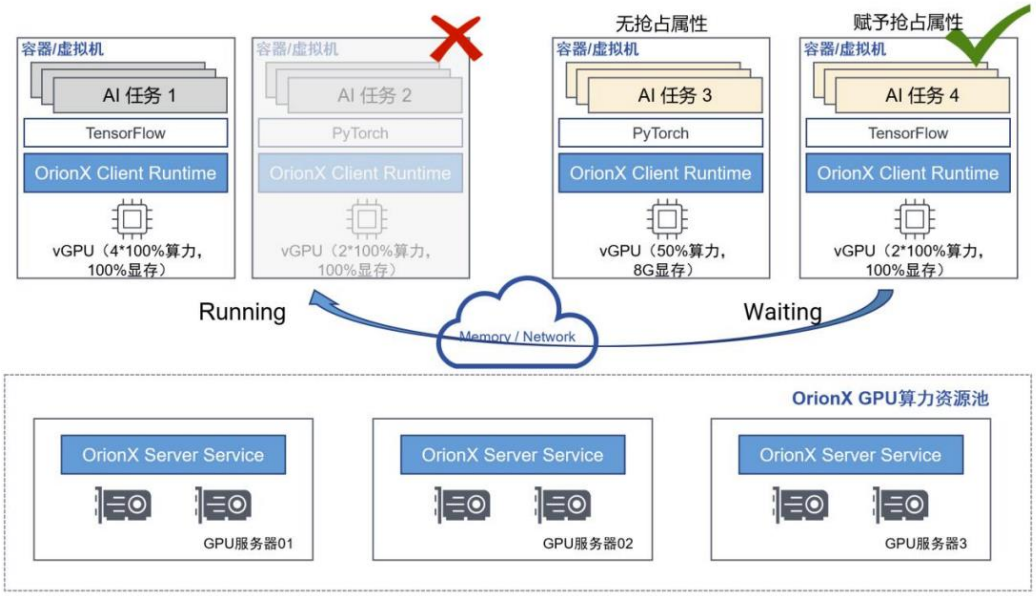
通过“任务队列”功能支持训练/推理任务自动排队。当请求 vGPU 资源的训练/推理任务遇到物理 GPU 剩余资源不足，无法分配的时候，OrionX 支持任务排队能力。OrionX 会将任务放入等待队列中，直到队列中前面任务跑完，可调度等待任务所需资源时，任务进入运算状态。

 版权所有 © 超聚变数字技术有限公司

OrionX 允许对任务进行队列优先级预设，一旦资源缺乏导致任务进入等待队列，按照优先级进行全局排列，优先匹配重要任务。



通过“抢占”功能支持任务抢占资源，上述“任务队列”中，OrionX 允许对某些任务赋予抢占属性，一旦资源不足导致任务进入等待队列，按照抢占属性的特点，该任务将立即分配到可用资源。赋予抢占属性的任务，在 OrionX 集群剩余资源不足而导致无法分配到资源时，可以通过抢占正在运行的任务，让一部分任务提前退出，这样集群可能就会空出足够的资源分配给该任务。

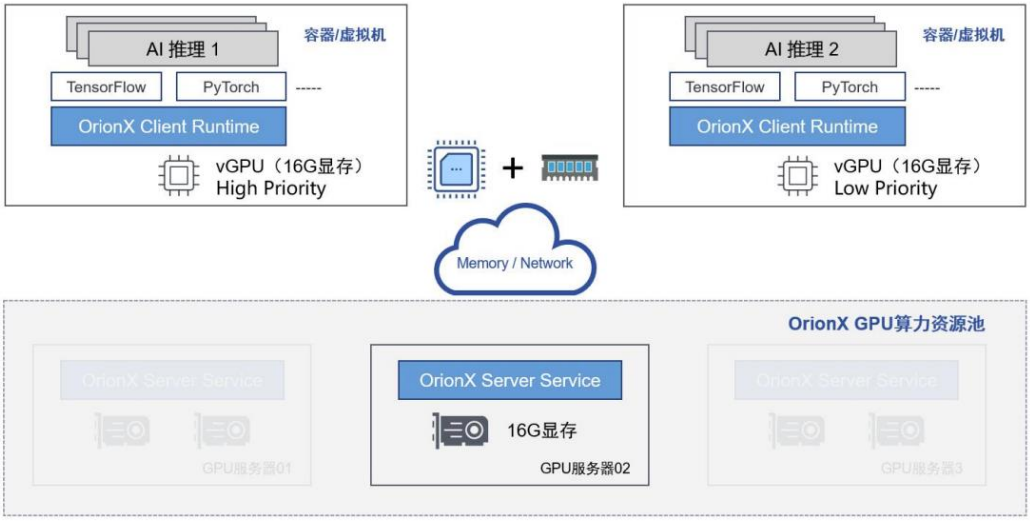


通过“显存超分”功能支持多任务叠加常驻。通常推理任务为满足最佳用户体验，会将推理模型常驻显存，24 小时不中断，以便拥有最快响应速度。但是这类常驻任务一般算力利用极低，而且潮汐效应明显。

 版权所有 © 超聚变数字技术有限公司

OrionX 支持多任务潮汐叠加。通过“显存超分”，OrionX 会调用系统内存补充 GPU 显存，在逻辑上扩大 GPU 显存的承载容量，从而支持多个常驻显存的长尾任务叠加在同一个物理 GPU 上，提高单个 GPU 的承载量，充分利用 GPU 闲置算力。

根据业务特点，OrionX 还支持不同任务设置不同优先级，从而保证突发高优先级任务的服务质量。



产品价值总结

提高利用率

o 支持将 GPU 切片为任意大小的 vGPU，从而允许多 AI 负载并行运行，提高物理 GPU 利用率。

o 提高 GPU 综合利用率多达 3-10 倍，1 张卡相当于起到 N 张卡的效果，真正做到昂贵算力平民化。

高性能

o 相比于物理 GPU，OrionX 本地 vGPU 性能损耗几乎为零，远程 vGPU 性能损耗小于 2%。

o vGPU 资源隔离，并行用户无资源互扰。

轻松弹性扩展

o 支持从单台到整个数据中心 GPU 服务器纳管，轻松实现 GPU 资源池的横向扩展。

o 全分布式部署，通过 RDMA（IB/RoCE）或 TCP/IP 网络连接各个节点，实现资源池弹性扩展。

灵活调度

 版权所有 © 超聚变数字技术有限公司

o 支持 应用负载与 GPU 资源分离部署，更加高效合理地使用 GPU 资源。

o CPU 与 GPU 资源解耦合，两种服务器分开购买、按需升级、灵活调度，有助于最大化数据中心基础设施价值。

全局管理

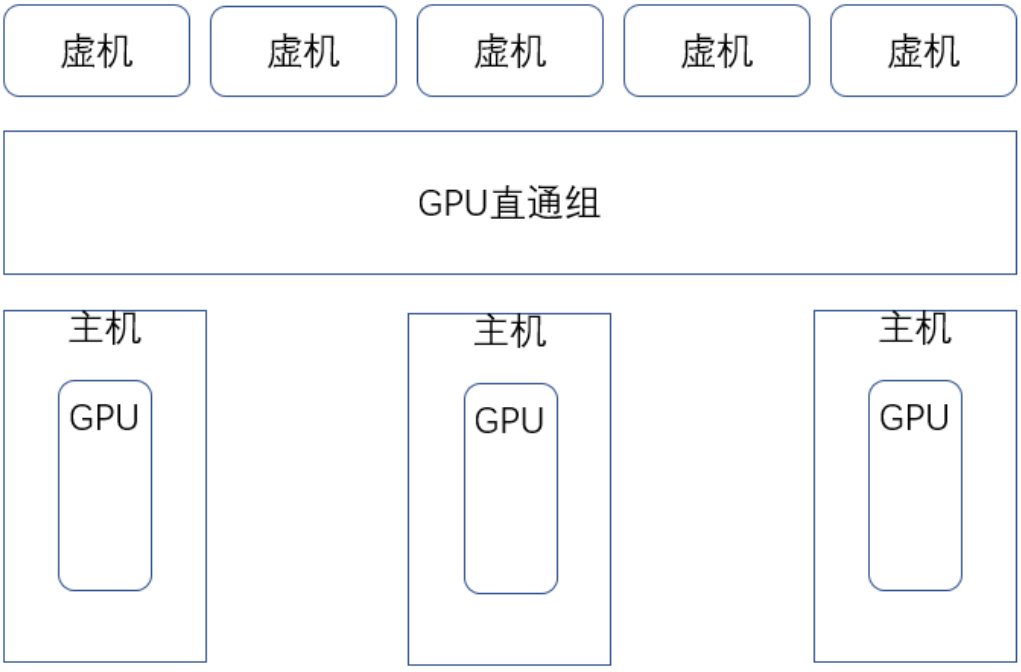
o 提供 GPU 资源管理调度策略。

o GPU 全局资源池性能监控，为运维人员提供直观的资源利用率等信息。

o 一键解决应用开发面临的训练模型中 GPU/CPU 配比和多机多卡模型拆分问题，为算法工程师节省大量宝贵时间。

3.16 GPU 直通

GPU 直通是通过硬件 IOMMU、中断重映射等技术，使虚拟机无需经过 Hypervisor 参与直接与硬件进行交互，实现与物理主机试用 GPU 几乎相同的性能，提升虚拟机的图形视频处理能力，以满足客户对于图形视频等高性能图形处理能力的需求。
FusionOne Compute 实现直通高性能的同时，通过把 GPU 设备添加到一个组中，虚拟机与 GPU 直通组进行关联，实现直通设备可以灵活的在更多的虚拟机中进行复用(不能同时使用，但是可以同时关联。虚拟机开机则占用资源，关机则释放资源，实现资源的合理利用)。



 版权所有 © 超聚变数字技术有限公司

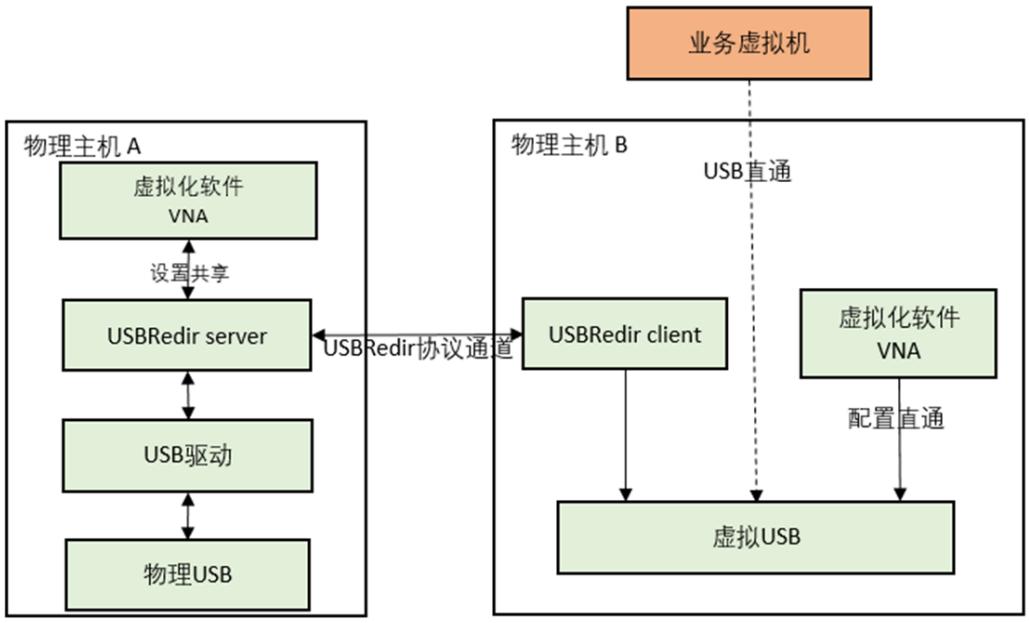
3.17 网络 USB

须知

本小节特性需要配套 FusionOne Center。

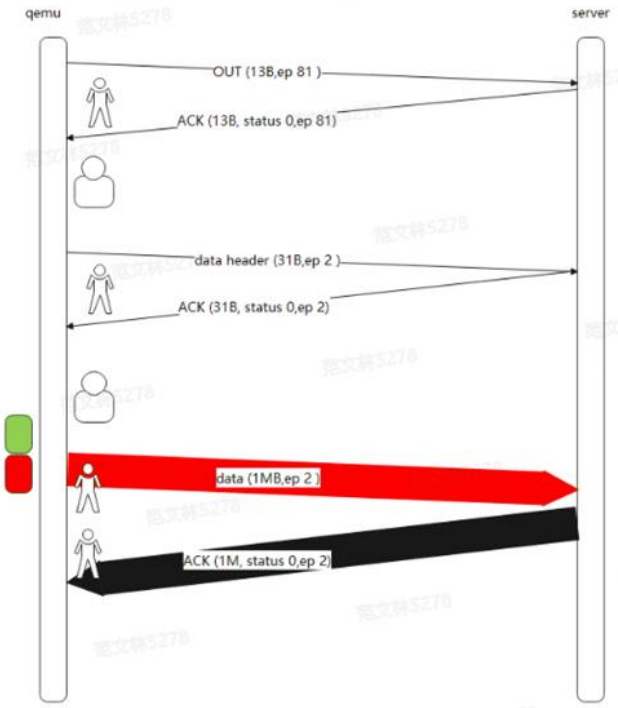
支持无代理跨物理主机的虚拟机 USB 映射，需要使用 USB 设备时，无需再虚拟机上安装客户端插件，且虚拟机漂移到在其它物理主机后，仍能正常使用迁移前所在物理主机上的 USB 资源。网络 USB 相比 USB 直通，对于业务的自适应能力、使用便捷性更佳。

绑定 USB 的虚拟机跨主机可以启动运行，虚拟机故障后也支持 HA 到其他任意在集群内的主机上，也可以支持 DRS 调度或手动热迁移虚拟机到其他主机。



USB 设备跨主机热迁移需要保证 IO 不中断，超聚变通过自研的 IO 冻结、IO 缓存、重传等一系列的机制保证 IO 持续运行。虚拟机迁移时，当检测到有 USB 设备时，在虚拟机暂停期间，将 IO 冻结，并将以下时序处理中的消息缓存，当虚拟机迁移到目标主机后，将 USB 读写消息重传，从而保证业务 IO 的连续性。

 版权所有 © 超聚变数字技术有限公司



3.18 SSH 密钥管理

超聚变虚拟化平台支持 SSH 密钥管理，包括密钥生成、上传、查询、删除与分发等能力，以支持用户在平台直接配置虚拟机的免密登录。

生成：

在虚拟化平台上直接触发生成一组随机的 SSH 密钥对，并支持配置算法与密钥长度，生成后平台将保存 SSH 公钥，私钥进行一次性返回到前端界面且后续无法通过其他方式再次查询。

上传：

用户可将自己预先生成的 SSH 公钥上传到平台，上传完成后该密钥与平台生成的密钥会统一管理。

查询：

用户可以列表查询已生成或上传的密钥。

分发：

用户可将 SSH 公钥分发到虚机内部，分发完成后用户可在已配置 SSH 私钥的客户端对已分发密钥的虚机机进行免密登录。

 版权所有 © 超聚变数字技术有限公司

3.19 PCI 设备白名单管理

物理主机上有可能会安装各种的 PCIe 设备，有些 PCIe 设备只能主机上使用，如系统盘的 raid 卡；有些 PCIe 设备是需要直通到虚拟机中使用，如 NVMe 数据盘、GPU 卡、加密卡、网卡等。为了防止误操作，虚拟化平台中对当前兼容性列表中 PCIe 设备都有记录对应的白名单，通过白名单来识别 PCIe 设备的具体类型。

为了方便引入了根据业务诉求，需要添加新的 PCIe 设备时，可通过 PCIe 白名单自定义添加新的 PCIe 设备让虚拟机可以识别到，这样就可将这个设备在虚拟机使用。

3.20 GuestOS 支持解耦

面对客户不断变化的 GuestOS 的需求，传统做法是需要升级平台版本才可以支持，升级需要规划时间窗口，费时费力，无法及时响应客户需求。

为了提升效率，FusionOne Compute 将 GuestOS 的兼容性与版本解耦。为应对客户新的 GuestOS 诉求，发布独立的 SIA 升级包。SIA 补丁包更新相比平台的更新，仅仅需要在主机进行简单的文件替换，无需重启主机、也无需迁移虚拟机，可快速在现网平台业务无感的情况下，快速完成对新的 GuestOS 的兼容。

。

 版权所有 © 超聚变数字技术有限公司

4 网络虚拟化

4.1 概述
4.2 所画即所得
4.3 连通性检测
4.4 分布式虚拟交换机
4.5 网络安全组
4.6 SDN 网络
4.7 Overlay 网络
4.8 VPC
4.9 NAT 网关
4.10 弹性 IP
4.11 网络 ACL
4.12 IP 地址预留
4.13 网络 QoS
4.14 SR-IOV 交换
4.15 用户态 OVS 交换
4.16 DHCP 服务
4.17 虚拟路由器

 版权所有 © 超聚变数字技术有限公司

4.18 端口流量镜像

4.19 在线修改虚拟机 IP 地址

4.20 虚拟负载均衡器

4.1 概述

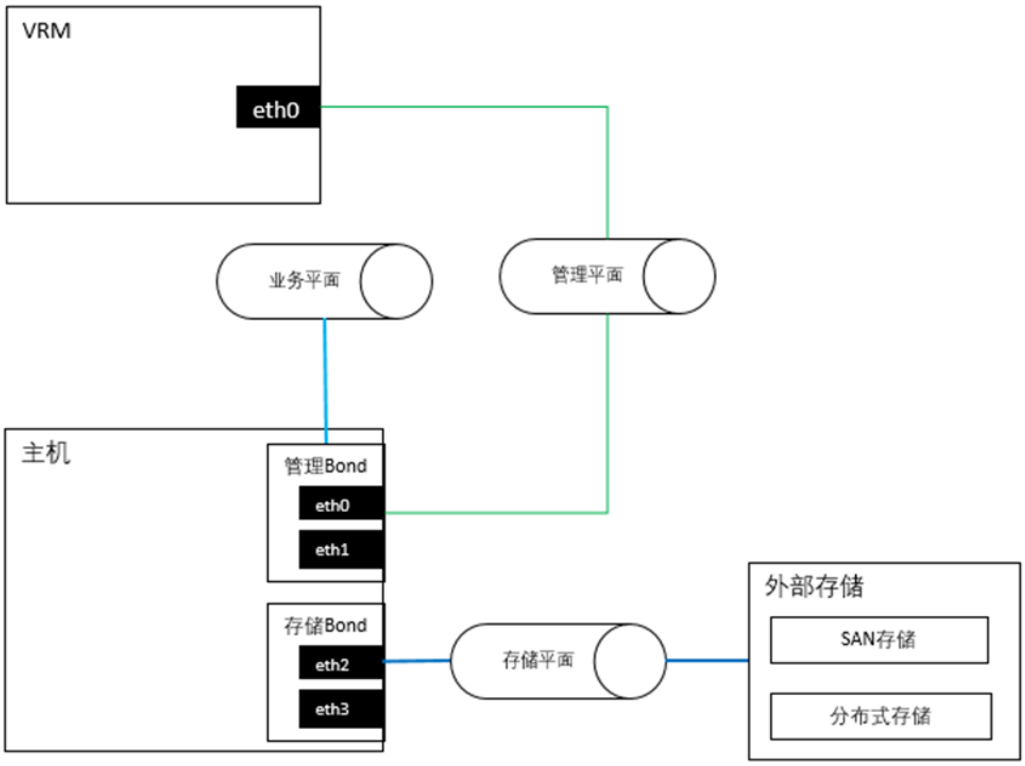
网络虚拟化是构建企业级云架构中非常重要的一部分，如果在云计算虚拟化环境中，我们的网络如果继续采用传统的 IT 架构中硬件方式定义网络的话，就会存在很多问题：

1. 如何保障虚拟机在保持相应的网络策略不变的情况下进行虚拟机迁移。
2. 虚拟化后的数据中心涉及业务众多，对外部提供云接入服务时，传统的 Vlan 技术已经无法满足业务隔离的需求，解决大规模租户和租户之间、业务和业务之间的安全隔离也是面临的首要问题。
3. 虚拟化后的数据中心的业务系统的构建和上线对网络功能的快速部署、灵活弹性甚至成本，提出了更高的要求。
4. 传统网络中，不论底层的 IT 基础设施还是上层的应用，都由专属设备来完成，这些设备成本高昂，能力和位置僵化，难以快速响应新业务对网络快速、灵活自动化部署的需求。

基于上述问题，FusionOne Compute 采用了业界成熟的 OVS 网络技术方案，从而实现企业网络虚拟化。通过软件实现类似物理交换机的功能，将物理网口（上行链路）、虚拟机的网口连接，实现网络共享互联。

虚拟化平台由两个部分组成，虚拟化管理平台（VRM）以及主机软件组成。主机软件为基于共内核版本的 FusionOS 优化裁剪的虚拟化 OS，内置计算虚拟化、网络虚拟化、存储虚拟化，每个节点部署计算代理（VNA），负责与 VRM 通信。

 版权所有 © 超聚变数字技术有限公司



4.2 所画即所得

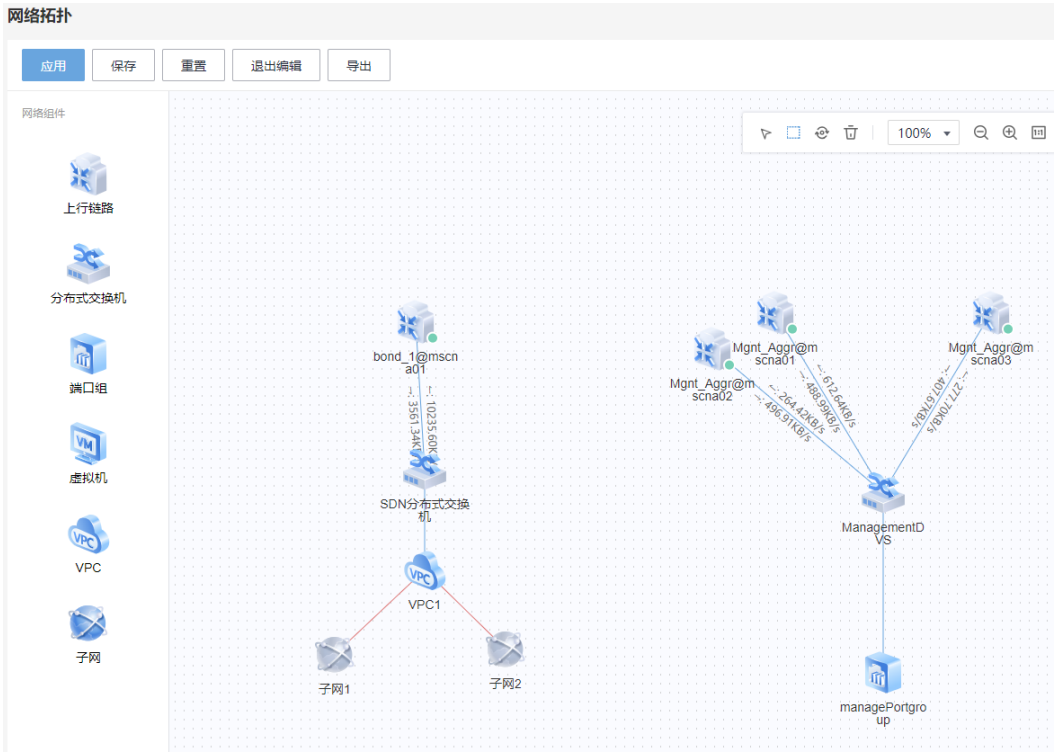
须知

本小节特性需要配套 FusionOne Center。

平台提供数据中心网络拓扑可视化界面，通过管理平面便可从资源池中调取相应的资源，即可呈现出不同的拓扑架构，在管理界面，支持通过拖拽式快速搭建业务所需的网络拓扑。在网络拓扑提交应用后，平台会自动根据网络拓扑部署业务网络，屏蔽了底层虚拟网络实现的复杂性，良好的人机交互使得数据中心网络运维更加容易。

- 网络资源组件：支持使用上行链路、分布式交换机、端口组、虚拟机等网元搭建网络拓扑。
- 网元信息可视化：支持选中网元后，展示该网元的基本信息和状态。
- 网络负载可视化：支持在上行链路与分布式交换机之间的网络拓扑连线上展示网络负载信息，如上下行包速、上下行流速。

 版权所有 © 超聚变数字技术有限公司



4.3 连通性检测

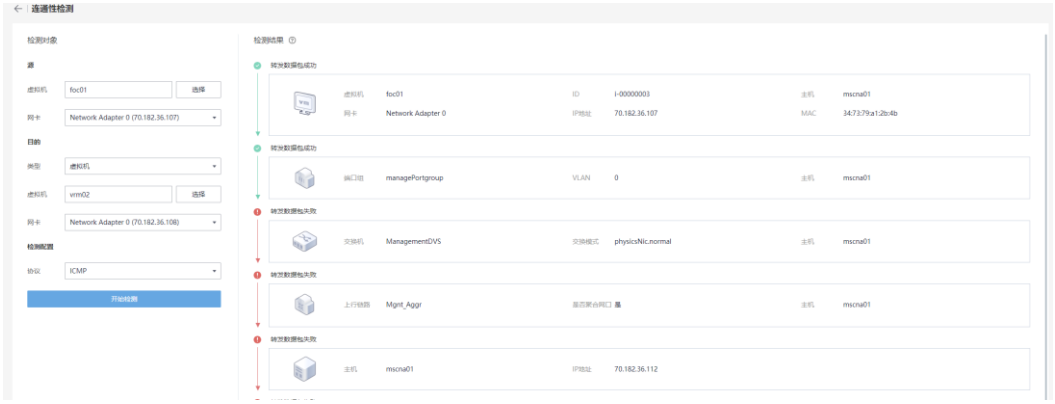
须知

本小节特性需要配套 FusionOne Center。

虚拟化平台提供虚拟机连通性监测功能，用于检测和展示源虚拟机到目的虚拟机或者目的 IP 地址之间的连通情况，帮助用户排查网络通信故障原因，为用户节约大量的运维工作。

- 检测对象：支持源虚拟机到目的虚拟机或者目的 IP 地址的连接检测。
- 检测协议：支持 ICMP、TCP、UDP
- 数据包转发路径展示：支持展示源虚拟机到目的虚拟机或者目的 IP 之间的数据包在安全组、分布式交换机、主机、目的端口等网元间的转发路径，绿色线代表数据包转发成功，红色线代表数据包转发失败。

 版权所有 © 超聚变数字技术有限公司



4.4 分布式虚拟交换机

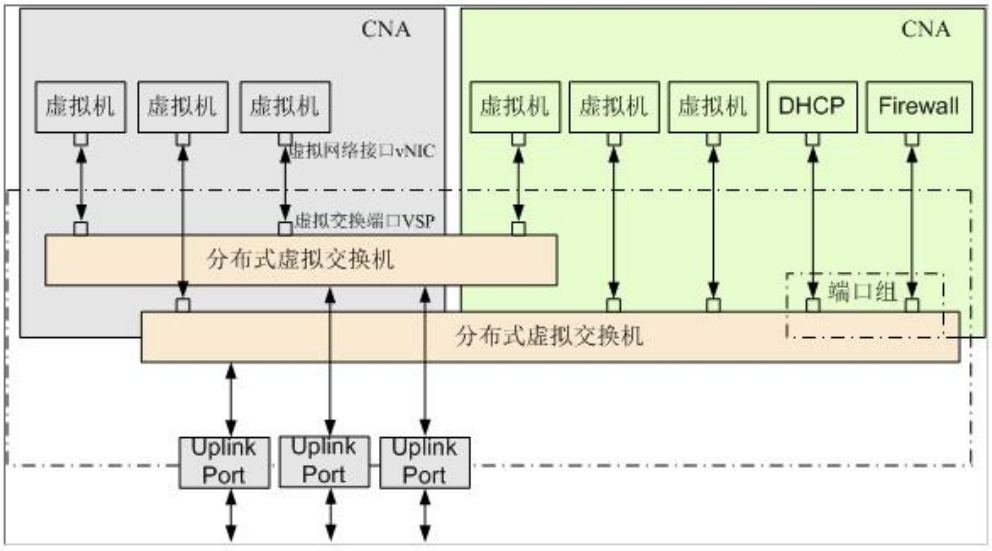
分布式虚拟交换机是管理一至多台 CNA 服务器主机上的虚拟交换机（基于软件的虚拟交换机或智能网卡虚拟交换机）的虚拟网络管理方式，包括对服务器主机上的物理端口和虚拟机虚拟端口进行配置/维护。

虚拟化平台提供多种类型的分布式虚拟交换机，如：普通模式（端口组），SDN 模式，SRIOV 直通模式，用户态交换模式（DPDK），通过虚拟分布式交换机把分布在集群中的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，简化虚拟机网络的部署、管理和监控，适合于大规模的网络部署。

分布式虚拟交换机可以保证虚拟机在主机之间迁移时网络配置的一致性，同时提供丰富的网络配置管理功能，实现网络资源统一管理，实时化网络监控。

分布式虚拟交换机的模型如下图所示：

 版权所有 © 超聚变数字技术有限公司



分布式虚拟交换机模型的基本特征：

- 用户可以配置多个分布式交换机，每个分布式交换机可以覆盖集群中的多个 CNA 节点。
- 每个分布式交换机具有多个分布式的虚拟端口 VSP（Virtual Switch Port），每个 VSP 具有各自的属性（速率、统计和安全组等），为了管理方便采用端口组管理相同属性的一组端口，相同端口组的 VLAN（Virtual Local Area Network）、QoS、安全属性相同。
- 每个分布式交换机可以配置一个 Uplink 上行链路组，用于虚拟机对外的通信，Uplink 上行链路组可以包含多个物理网卡，这些物理网卡可以配置网口聚合和负载均衡策略。
- 每个虚拟机可以具有多个 vNIC（Virtual Network Interface Card）接口，vNIC 可以和交换机的 VSP 一一对接。

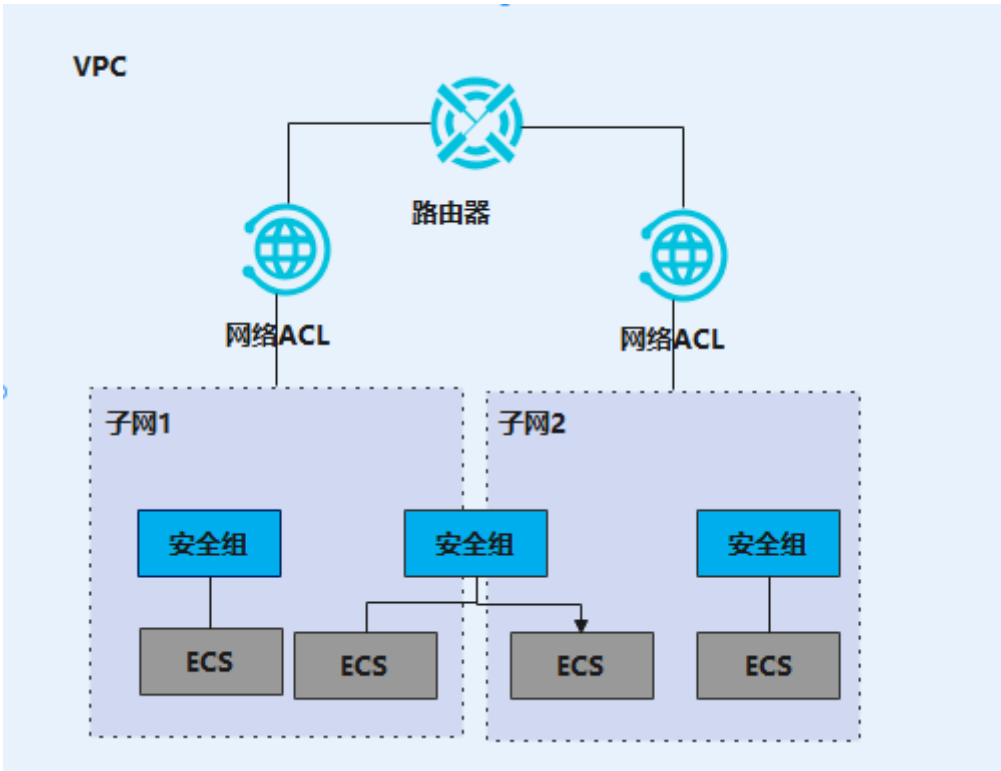
4.5 网络安全组

安全组服务(Security Group) 是对进出虚拟机端口的网络报文进行限制的安全过滤规则。关联虚拟机端口与安全组后，该安全组的规则会对进出该虚拟机端口的网络报文进行过滤，便于管理员对业务虚拟机的网络安全进行控制。用户可以根据实际需求，对出入口流量，根据配置规则，配置白名单、黑名单进行网络访问控制。

- 白名单：添加安全组后，默认禁止该安全组的虚拟机与外界互通，需要在安全组中添加规则后才可使网卡在该安全组的虚拟机与外界互通，默认为白名单。

 版权所有 © 超聚变数字技术有限公司

- 黑名单：添加安全组后，默认该安全组的虚拟机可以与外界互通。需要在安全组中添加规则后才可拒绝网卡在该安全组的虚拟机与外界某些地址互通。



4.6 SDN 网络

须知

本小节特性需要配套 FusionOne Center。

软件定义网络（SDN，Software Define Network），即通过软件的方式控制和调度网络。SDN 是一种创新性的网络架构，它通过标准化技术（如 openflow）实现网络设备的控制层和数据层面的分离，进而实现对网络流量的灵活化，集中化，细粒度的控制，从而为网络的集中管理和应用加速创新提供了良好的平台，由此可获得对网络的可编程性，自动化和控制能力，使网络很容易适应变化的业务需求，从而建立高度可扩展的弹性网络。

传统网络架构下，增加新业务需要手动配置多种设备，各厂商设备配置命令千差万别，协议繁复，配置极其复杂；没有统一的配置管理工具，维护成本高，而随着社交网络、

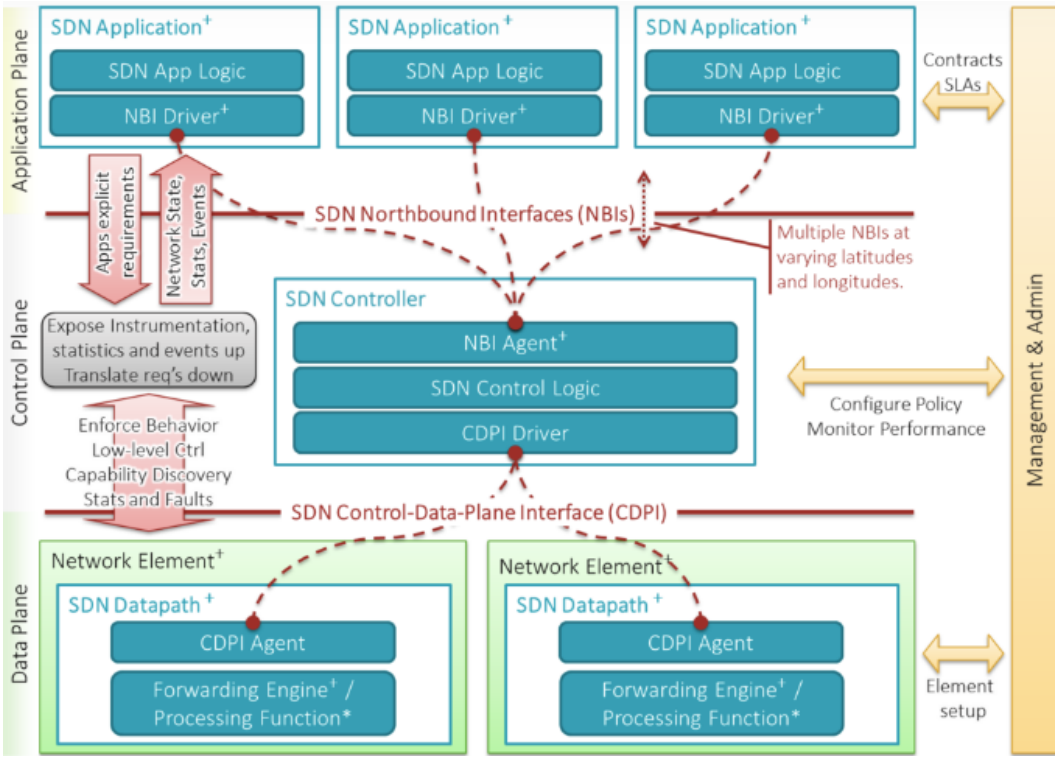
 版权所有 © 超聚变数字技术有限公司

移动互联网、物联网等业务领域的快速发展，传统网络已经无法满足数据中心业务弹性伸缩、快速部署、频繁变化的要求。

如例：某政企云项目客户内部有 20 个部门及子单位，要求根据部门/子单位的需求（访问其他政务外网的特定服务）按需快速提供网络、按需配置规则访问策略，实际的组网过程中可能会有多个部门/子单位公用或者单独一台设备交换机设备，传统的方式需要一台一台的修改配置 ACL 策略访问规则，这种繁复的工作耗时又容易出错。

SDN 初期就是为了改善上述的高昂的时间及错误成本，通过一个集中式的控制面管理多个交换机，实现一个指令同步更新配置 ACL 规则到所有的交换机设备，简化了人工连线和重复操作的错误风险。随着云计算虚拟化的高速发展，发现 SDN 功能融合虚拟化技术之后，友好支持大规模集群，通过 overlay 隧道网络实现支持多租户虚拟网络，基于 OpenFlow 标准协议，配置 OpenVswitch 虚拟交换机更好的实现控制面和数据面隔离。

.



如 SDN 架构图所示，控制平面和数据平面分离解耦合，俩者之间不再相互依赖。

4.7 Overlay 网络

 版权所有 © 超聚变数字技术有限公司

须知

本小节特性需要配套 FusionOne Center。

Overlay 网络是通过网络虚拟化技术，在同一张 Underlay 网络上构建出的一张或者多张虚拟的逻辑网络。不同的 Overlay 网络虽然共享 Underlay 网络中的设备和线路，但是 Overlay 网络中的业务与 Underlay 网络中的物理组网和互联技术相互解耦。

表4-1 Underlay 网络 VS Overlay 网络:

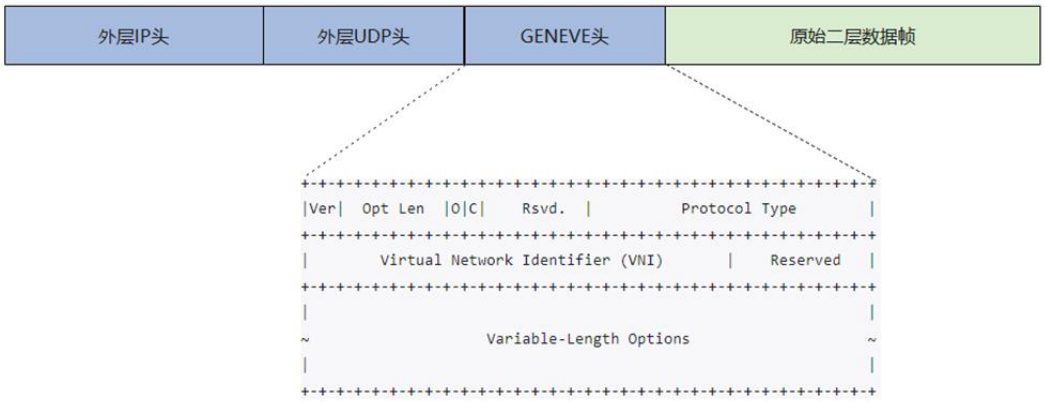
对比项	Overlay 网络	Underlay 网络
数据传输	沿着节点间的虚拟链路进行传输	通过网络设备例如路由器、交换机进行传输
包封装和开销	需要跨源和目的封装数据包，产生额外的开销	发生在网络的二层和三层
报文控制	面向软件	面向硬件
部署时间	只需更改虚拟网络中的拓扑结构，可快速部署	上线新服务涉及大量配置，耗时多
多路径转发	支持虚拟网络内的多路径转发	因为可扩展性低，所以需要使用多路径转发，而这会产生更多的开销和网络复杂度
扩展性	扩展性强，例如 VLAN 最多可支持 4096 个标识符，而 Geneve 则提供多达 1600 万个标识符	底层网络一旦搭建好，新增设备较为困难，可扩展性差
协议	VXLAN、GENEVE，NVGRE、SST、GRE、EVPN	以太网交换、VLAN、路由协议(OSPF、IS-IS、BGP 等)
多租户管理	能够管理多个租户之间的重叠 IP 地址	需要使用基于 NAT 或者 VRF 的隔离，大型网络中存在巨大的挑战

在网络虚拟化的场景中，主流协议如 VXLAN，GENEVE，STT 等，都各具特色，具体的解决方案场景会根据协议的自身的灵活性以及可扩展性选择。

 版权所有 © 超聚变数字技术有限公司

FusionOne Center 中的 Overlay 网络采用了 Geneve 协议的格式封装报文，当然这也是 OVS&OVN 官方社区主推的隧道报文协议。

GENEVE 与 VXLAN 类似，仍然是 Ethernet over UDP，也就是用 UDP 封装 Ethernet。VXLAN header 是固定长度的（8 个字节，其中包含 24bit VNI），与 VXLAN 不同的是，GENEVE header 中增加了 TLV（Type-Length-Value），由 8 个字节的固定长度和 0~252 个字节变长的 TLV 组成。GENEVE header 中的 TLV 代表了可扩展的元数据。我们来看一下 GENEVE header：



- Version（2bit）：目前是 0
- Opt Len（6bit）：以 4 字节为单位，表明 Variable Length Options 的长度。因为只有 6bit，所以 Variable Length Options 最多是 252（63*4）字节。
- O（1bit）：表明这是一个 OAM 包，包含了控制信息，而非数据。Endpoint 可以根据这个 bit 来优先处理这个包
- C（1bit）：表明在 Variable Length Options 里面，存在一个或者多个 Critical 的 option。当 C 被置位时，Variable Length Options 必须被解析，如果当前 Endpoint 不支持 GENEVE 解析，那么应该丢弃数据包。如果 C 没有被置位，那么 Endpoint 可以根据 Opt Len 直接丢弃所有的 Variable Length Options。
- Rsvd.（6bit）：保留字段。
- Protocol Type（16bit）：被封装的协议类型，例如 Ethernet 就是 0x6558。
- VNI（24bit）：网络标识，24 比特，用于区分 GENEVE 段。
- Reserved（8bit）：保留字段。
- Variable Length Options：由 TLV 构成，包含了可扩展的元数据。

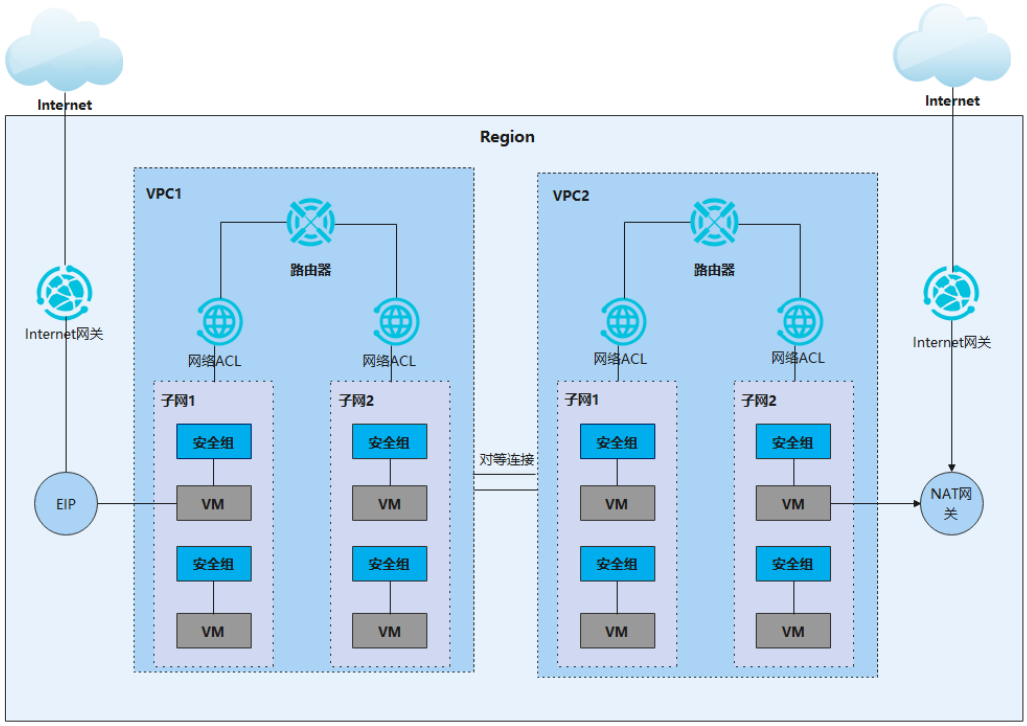
 版权所有 © 超聚变数字技术有限公司

4.8 VPC

须知

本小节特性需要配套 FusionOne Center。

虚拟私有云（Virtual Private Cloud）是一个逻辑隔离的网络空间。租户可拥有一个或多个 VPC，VPC 之间网络隔离，用户在使用的时候不受其他用户的影响。用户可使用 VPC 自主构建管理隔离的虚拟网络环境，提升用户在私有云中资源的安全性。VPC 能简化用户的网络部署，在灵活部署、安全隔离、丰富接入、访问控制等诸多方面具备独到的优势。通过 VPC，用户可以完全掌控自己的虚拟网络，包括创建子网、创建路由、网络 ACL、配置 DHCP，NAT 网关等网络特性，使用户能够在保证网络隔离的同时，更加科学的划分网络范围，部署灵活、管理方便，满足不同的业务需求。



虚拟私有云的基本特征包括：

1. 用户可以配置多个 VPC，并管理 VPC 的路由表，每个 VPC 通过分布式路由器覆盖多个节点。

2. 每个 VPC 内部可以通过 IP 网段划分出多个子网，子网内的虚拟机具备相同的网络和安全属性（网络 ACL、VLAN、Underlay、QoS 等），为方便 IP 地址管理，用户可以启用子网 DHCP，对子网内虚拟机进行 IP 地址的动态分配。

 版权所有 © 超聚变数字技术有限公司

3. 子网内虚拟机内部交互和外部交互的基本方式由分布式路由器提供，同时可以根据业务需要为虚拟机分配 EIP 或通过 NAT 网关，VPN 等方式对外通信。

4.9 NAT 网关

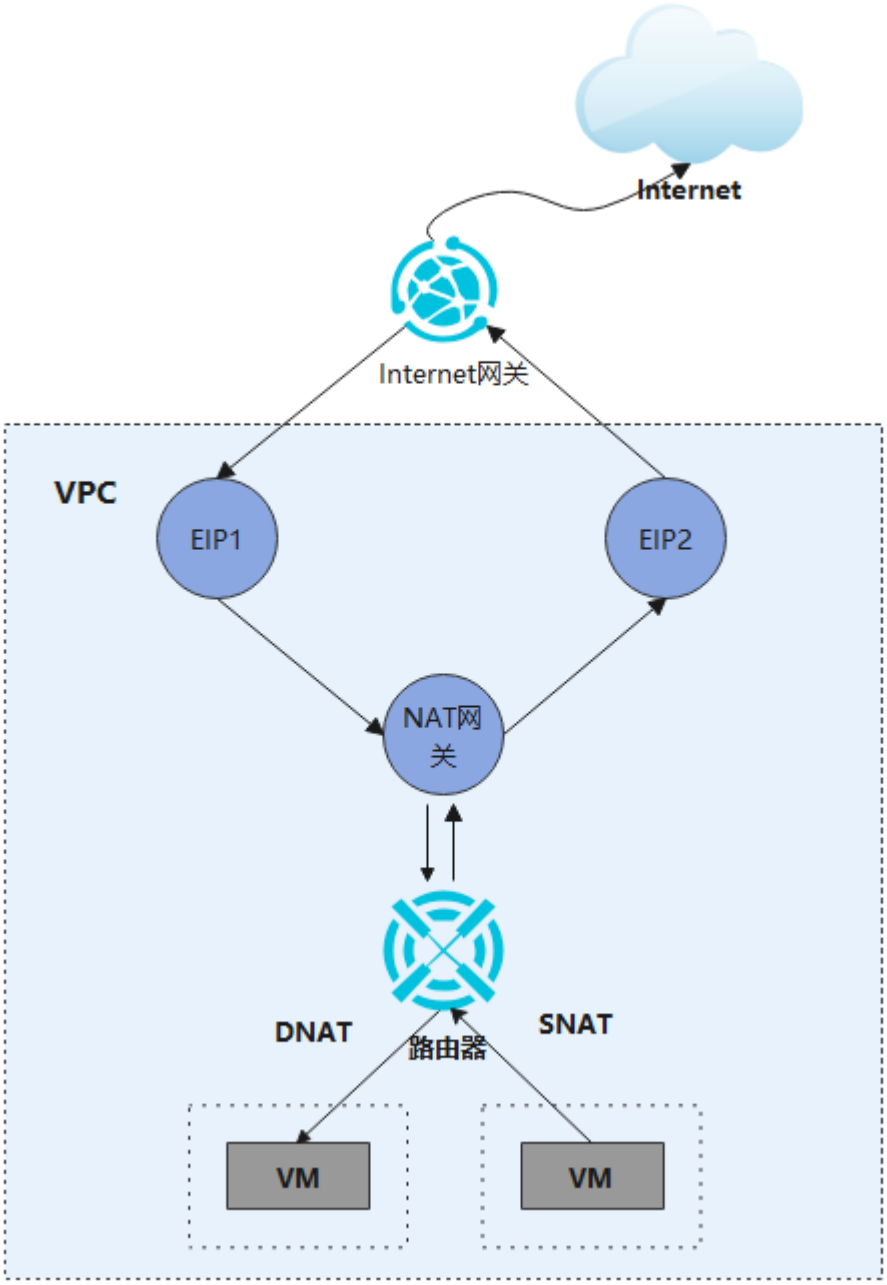
须知

本小节特性需要配套 FusionOne Center。

NAT 网关用于虚拟私有云内的虚拟机提供网络地址转换服务，使多个弹性云服务器可以共享使用弹性 IP 访问 Internet，节省客户公网 IP 资源。

NAT 网关分为 SNAT 和 DNAT 两个功能。DNAT 可以用于实现一个上层（路由的外层）网络地址的端口转发到下层（路由的内层）网络某个地址的某端口；SNAT 可以用于实现多台云主机共享一个网关地址访问上层(路由的外层)网络。

 版权所有 © 超聚变数字技术有限公司



虚拟私有云中存在一个默认路由器，用于将云内网络联通起来。通过在 VPC 内创建 NAT 网关，并配置 NAT 规则，使得云内虚拟机可以将地址转换为网关 IP，为上层网络提供服务。通过添加 DNAT 规则，使得外部网络的 IP 的指定端口可以映射为云内 IP 的指定端口，实现上层网络访问云内虚拟机。

FusionOne Center 中的 NAT 网关支持分布式，支持 NAT 网关高可用 HA。

 版权所有 © 超聚变数字技术有限公司

4.10 弹性 IP

须知

本小节特性需要配套 FusionOne Center。

弹性公网 IP（Elastic IP）提供独立的公网 IP 资源，VPC 子网中的云资源绑定 EIP 后可以与外界互相通信，其本质是一种 NAT 技术。

基于软件 SDN 实现 EIP 时，创建 OVN 的 NAT 资源，NAT 的类型为 dnat_and_snat。OVN 按照 NAT 信息转换成相应的 OVS 流表。最终 OVS 流表实现 NAT 的转换。对于目标地址是 EIP 的报文，更换目标 IP 为虚拟机 IP，继续转发。对于源地址是虚拟机 IP，出去外部的流量，更换源地址为 EIP。

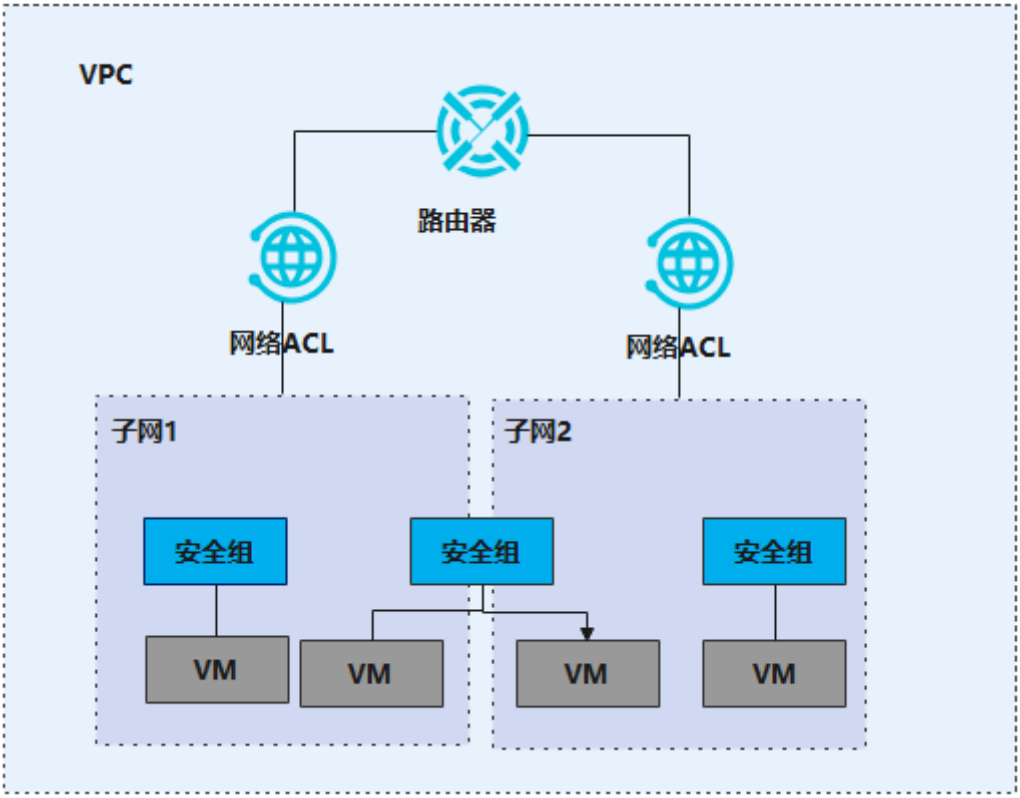
4.11 网络 ACL

须知

本小节特性需要配套 FusionOne Center。

网络 ACL 是一个子网级别的可选安全层，通过与子网关联的出方向 /入方向规则控制出入子网的数据流。网络 ACL 与安全组类似，都是安全防护策略，当您想增加额外的安全防护层时，就可以启用网络 ACL。安全组对云内虚拟机进行防护，网络 ACL 对子网进行防护，两者结合起来，可以实现更精细、更复杂的安全访问控制。

 版权所有 © 超聚变数字技术有限公司



网络 ACL 支持与子网绑定，支持基于网络 ACL 模板创建网络 ACL，一个 网络 ACL 支持和多个子网绑定，一个子网不能同时加入多个 网络 ACL。

网卡 ACL 与子网关联后，会相应下发默认拒绝策略，如果进/出子网的流量未匹配上自定义规则，则匹配默认规则，无法流入子网。

网络 ACL 和安全组功能对比：

表4-2 网络 ACL 和安全组功能对比

对比项	网络 ACL	安全组
防护对象	子网级别：防护整个子网，子网内的所有虚拟机都会受到网络 ACL 的保护	虚拟机级别：防护安全组内的实例，比如虚拟机。
配置策略	网络 ACL 支持设置允许，拒绝。	安全组支持黑白名单，一个虚拟机不能同时关联黑白名单安全组
优先级	多个规则按照优先级的排序生效，优先级数字越小，优先级越高。	多个规则按照规则结果的并集生效。
应用	创建子网没有网络 ACL 选项，需要先	• 创建虚拟机时，可选择关联安全

 版权所有 © 超聚变数字技术有限公司

对比项	网络 ACL	安全组
操作	创建网络 ACL，添加出入规则，并在网络 ACL 内关联子网。	组，如果当前没有安全组，则可直接创建安全组用于关联 • 虚拟机创建时未关联安全组，创建完成后再次关联安全组，则需要重启虚拟机才可生效。 • 虚拟机开机状态下，不支持关联，去关联安全组操作。
报文组	支持报文五元组（即源、目的 IP 地址，协议号、源、目的端口号）过滤。	支持报文三元组（即协议号、端口号、源/目的 IP 地址）过滤。

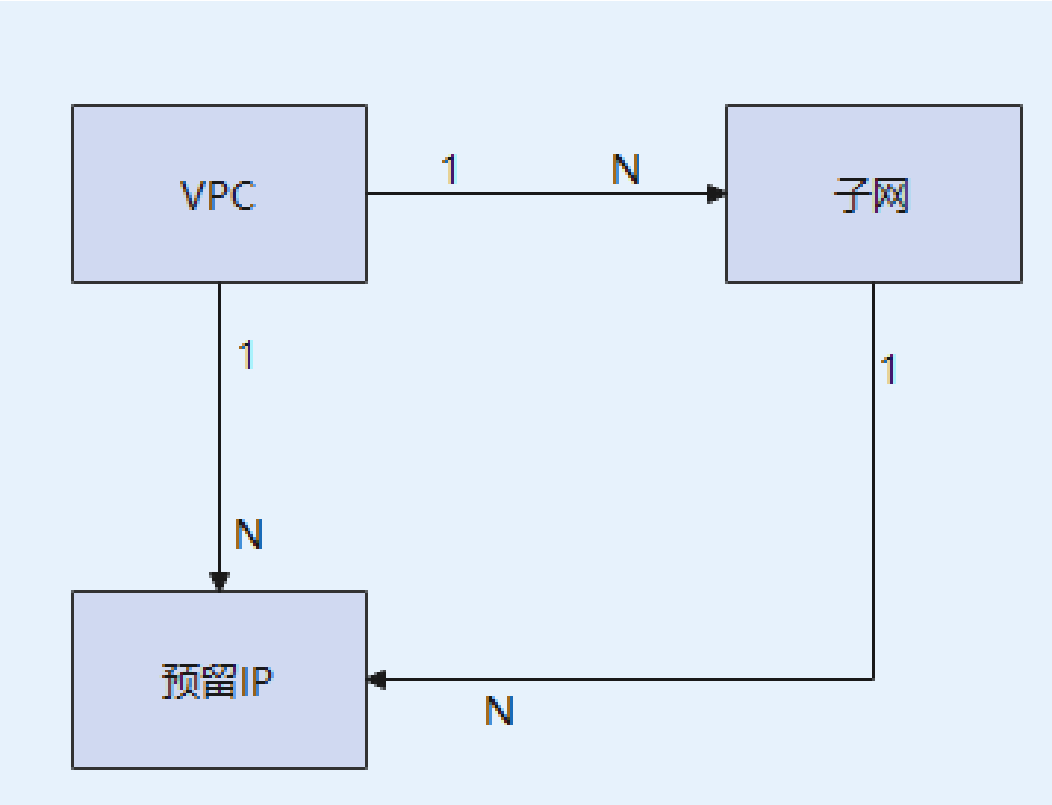
4.12 IP 地址预留

须知

本小节特性需要配套 FusionOne Center。

平台通过在子网中对部分 IP 地址的预留，这样预留的 IP 地址不会和后续其他服务创建时所使用的 IP 产生冲突，如当创建的多台云主机配置 Keepalived 作主备时，预留出的 IP 地址可以用作 VIP 管理多台云主机，这样可以防止预留的 IP 地址和创建的云主机 IP 地址冲突，同时也方便使用人员统一管理预留地址。

 版权所有 © 超聚变数字技术有限公司

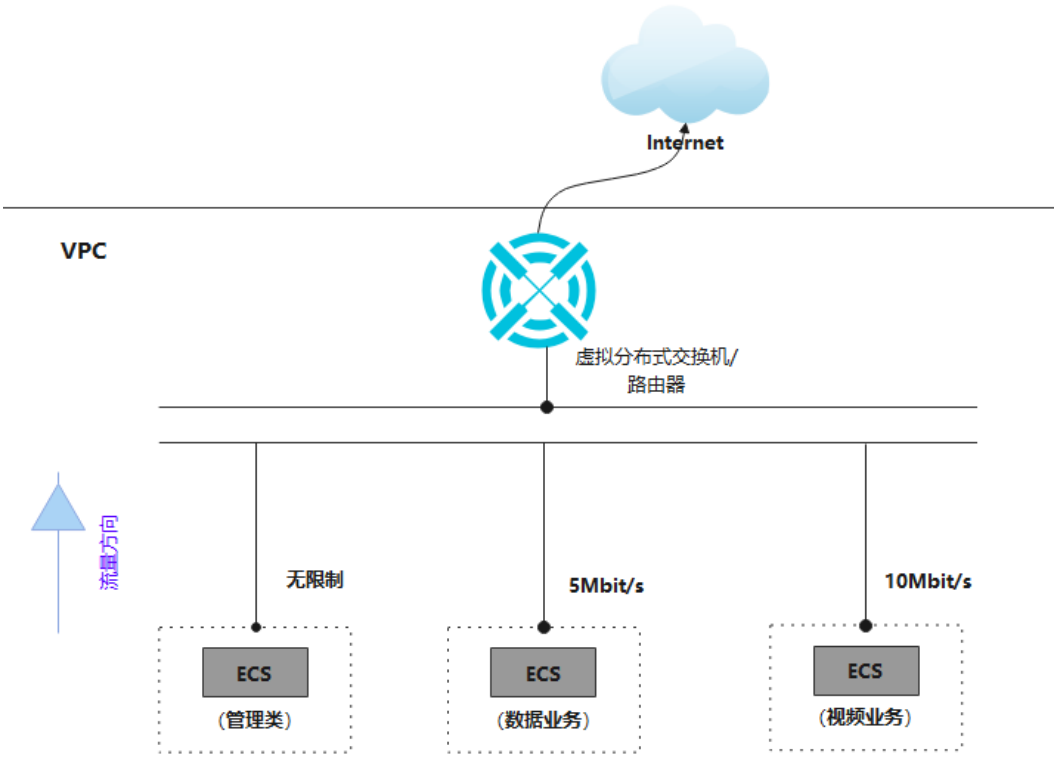


4.13 网络 QoS

在虚拟私有云中，租户虚拟机 上可能部署了不同类型的业务，比如管理、视频和数据等多种不同的业务，当大量的业务流量进入网络侧时，可能会因为带宽不足产生拥塞，需要对三种业务提供不同的带宽，优先保证其中一种业务报文的转发。因此可以针对不同业务的通信端口设置带宽限速 ，进行流量整形 。

流量整型在配制流量的平均带宽（等价于此前的上限带宽），同时可以设置突发的流量的大小。允许业务流量在系统存在可用资源的时候，最大带宽可以短时间内达到峰值速率，时间的长短取决于 Burst 的大小，即超过平均速率的流量的大小不能超过 Burst 的大小 。

 版权所有 © 超聚变数字技术有限公司

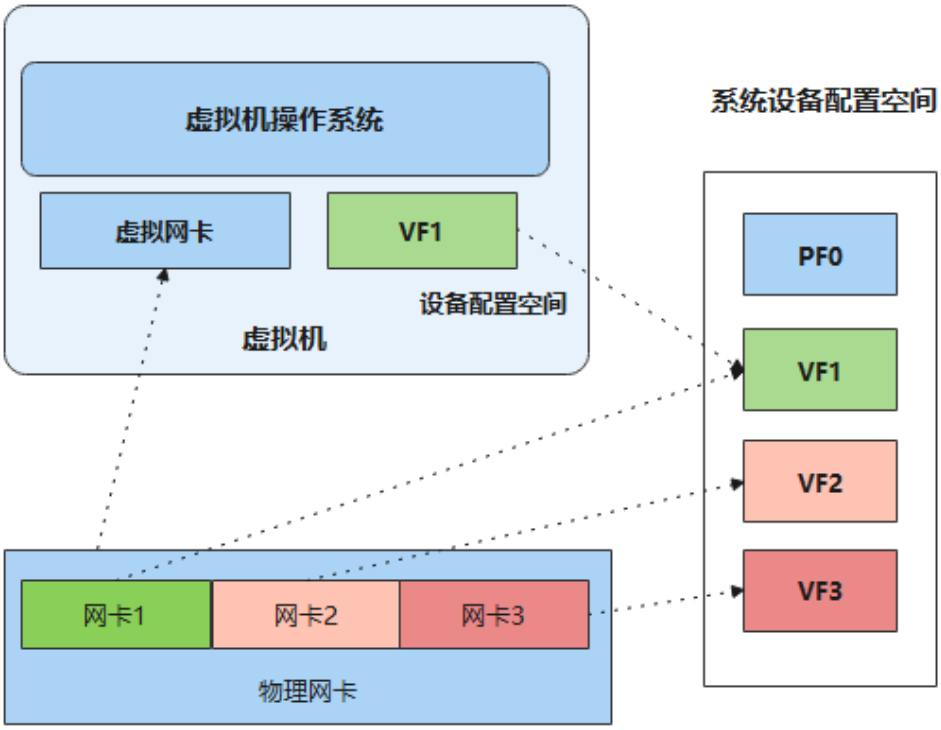


4.14 SR-IOV 交换

支持 SR-IOV 特性，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 设备的资源。SR-IOV 能够让网络传输绕过软件模拟层，直接分配到虚拟机。这样就降低了软件模拟层中的 I/O 开销。

SR-IOV 技术是一种基于硬件的虚拟化解决方案，SR-IOV 使用两种功能（function）：物理功能（Physical Functions，PF），PF 是完整的带有 SR-IOV 能力的 PCIe 设备，PF 能像普通 PCI 设备那样被发现、管理和配置；虚拟功能（Virtual Functions，VF）：简单的 PCIe 功能，VF 只能处理 I/O，每个 VF 都是从 PF 中分离出来的。每个物理硬件都有一定数目的 VF 可以配置，一个 PF 能被虚拟成多个 VF 用于分配给多个云主机，VMM 能将一个或者多个 VF 分配给一个虚机。在某一时刻，一个 VF 只能被分配给一个虚机，一个虚机可以拥有多个 VF。在虚机的操作系统看来，一个 VF 网卡看起来和一个普通网卡没有区别。

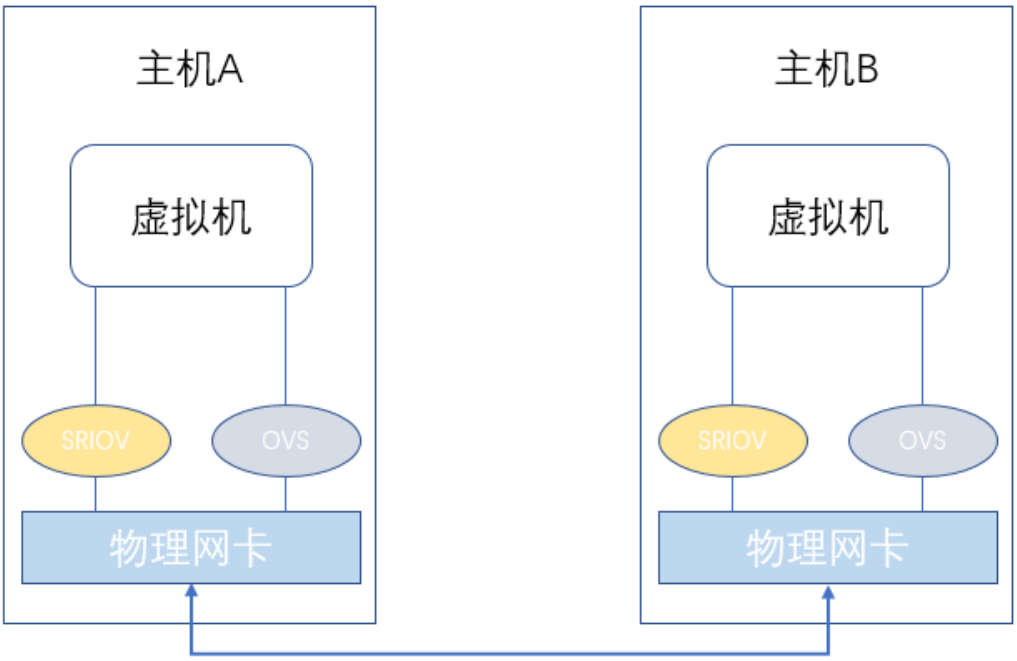
 版权所有 © 超聚变数字技术有限公司



SR-IOV 特性的使用需要硬件平台 Intel VT-d、AMD IOMMU 等技术的支持，在使用此功能时，相应特性必须在服务器的 BIOS 中被启用，如果 BIOS 中有 SR-IOV 的使能配置，那么也需要进行开启操作。

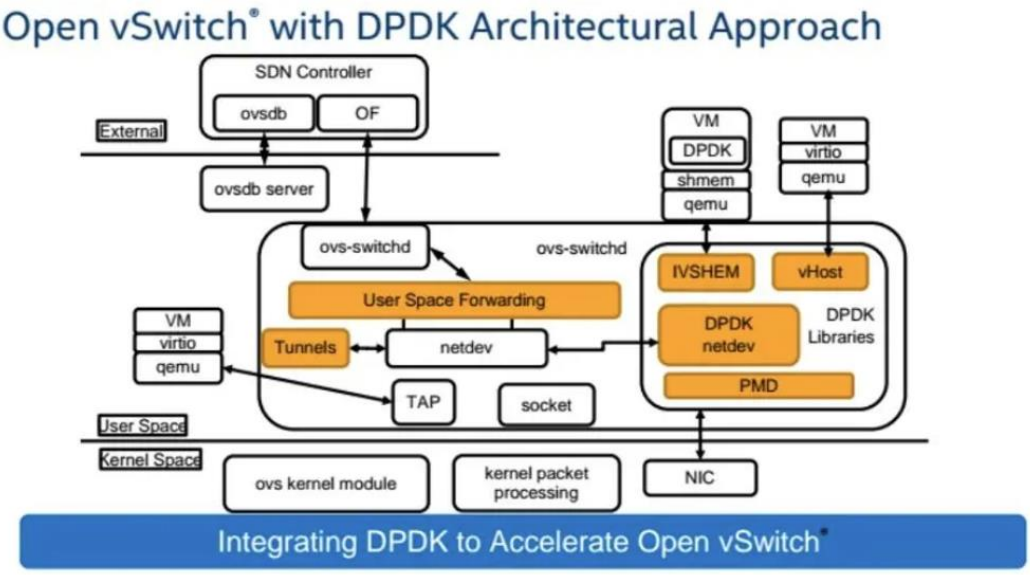
SRIOV 网卡热迁移特性，实现了虚拟机使用高速直通 SRIOV 网卡的同时，满足虚拟机热迁移的诉求。为了实现此功能，在给虚拟机添加 SRIOV 网卡时，底层添加了两张网卡，分别是正常业务流使用的 SRIOV 网卡，以及迁移过程中使用的 virtio 网卡，这两个网卡在虚拟机中被识别为 fail over 的主备网卡。当启动热迁移时，底层会自动移除 SRIOV 网卡，此时业务流切换到 virtio 网卡，迁移完成后，重新插入 SRIOV 网卡，业务流重新切换到此直通网卡上。通过把 SRIOV 网卡和 virtio 网卡规划到基于同一个物理网卡，实现底层自动配置，无需用户手动配置备用网络和选择备用网卡，提升简便性以及杜绝人为配置错误的问题。

 版权所有 © 超聚变数字技术有限公司



4.15 用户态 OVS 交换

通过使用 DPDK（Data Plane Development Kit，DPDK 是一系列库和驱动的集合）技术，在数据平面开发套件用来在 x86 平台进行快速的数据包处理。它通过环境抽象层旁路内核协议栈、轮询模式的报文无中断收发、优化内存/缓冲区/队列管理、基于网卡多队列和流识别的负载均衡等多项技术，实现了在 x86 处理器架构下的高性能报文转发能力，提高虚拟机网络性能。



 版权所有 © 超聚变数字技术有限公司

虚拟化虚拟化支持 OVS+DPDK 的用户态 OVS 的模式，通过 dpdk 的加速技术减少路径开销，提升网络性能，如提高网络吞吐率，包转发率及降低云主机之间的延迟和抖动。

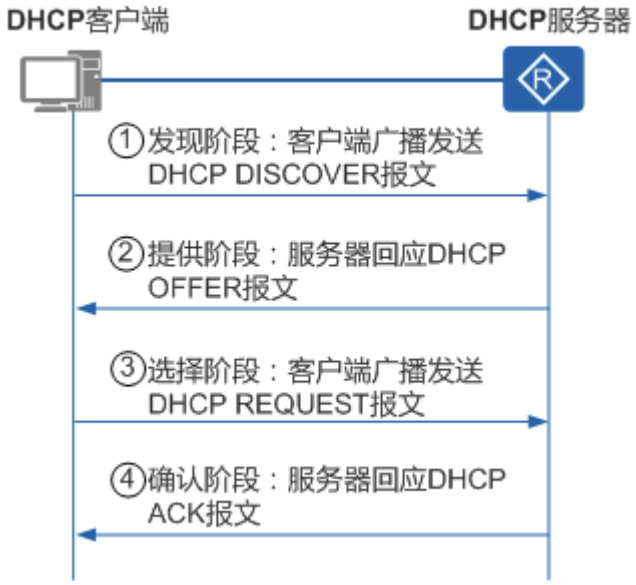
4.16 DHCP 服务

须知

本小节特性需要配套 FusionOne Center。

动态主机配置协议 DHCP（Dynamic Host Configuration Protocol）是一种网络管理协议，用于集中对用户 IP 地址进行动态管理和配置。

DHCP 协议采用 UDP 作为传输协议，DHCP 客户端发送请求消息到 DHCP 服务器的 68 号端口，DHCP 服务器回应应答消息给 DHCP 客户端的 67 号端口。



FusionOne Center 中的 DHCP 服务提供了两种地址分配机制，租户可以根据网络需求为不同的虚拟机选择不同的分配策略。

- 动态分配机制：通过 DHCP 为虚拟机分配一个有使用期限的 IP 地址。

  DHCP 使用了租期的概念，或称为设备 IP 地址的有效期。租用时间是不定的，主要取决于用户在某地连接 Internet 需要多久，这种分配机制适用于虚拟机需要临时接入网络或者空闲地址数小于网络主机总数且主机不需要永久连接网络的场景。

 版权所有 © 超聚变数字技术有限公司

- 静态分配机制：租户通过 DHCP 为指定的主机分配固定的 IP 地址。

  相比手工静态配置 IP 地址，通过 DHCP 方式静态分配机制避免人工配置发生错误，方便管理员统一维护管理。

4.17 虚拟路由器

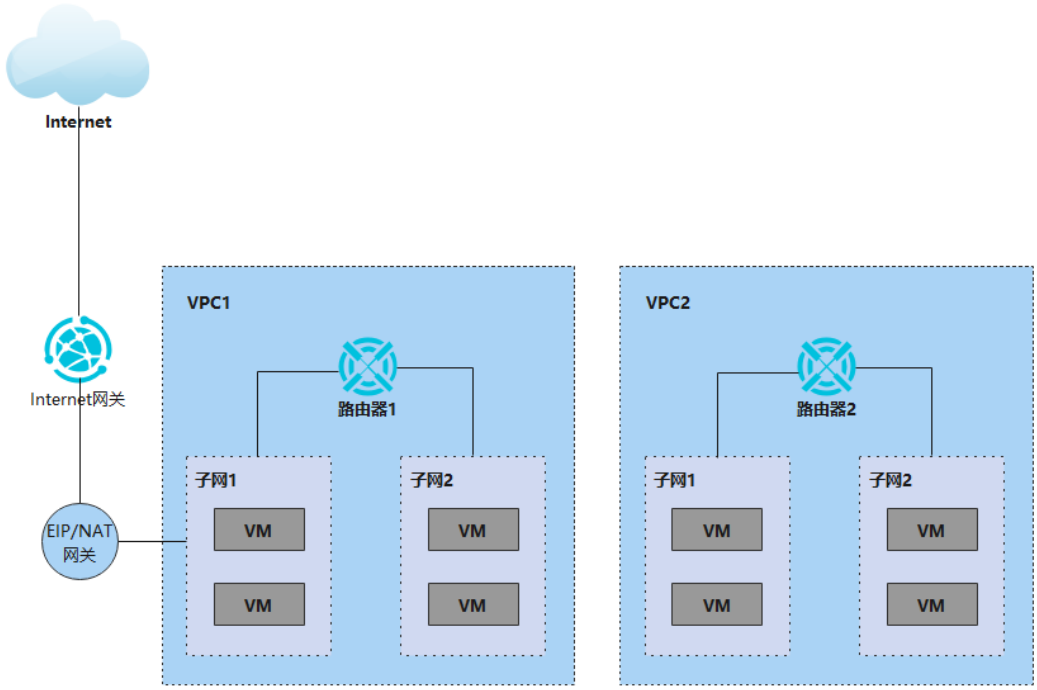
须知

本小节特性需要配套 FusionOne Center。

虚拟路由器用于子网与子网或子网与外部网络之间互通的虚拟服务。虚拟路由器通过关联子网实现租户不同子网的云主机之间的流量转发，通过关联外部网关实现虚拟机和外网之间网络流量的路由和转发。

FusionOne Center 提供的虚拟路由器是 VPC 内的不同子网互通的一种解决方案，每个 VPC 内一个虚拟路由器，通过逻辑流表实现网络流量的转发，逻辑流表采用了分布式的形态实现分布式和高可用。东西方向（E-W）上的流量不再需要将数据包发送到在单一网络节点后再转发，而是由本地的 DVR Router 直接进行跨子网的转发；南北方向（S-N）上，需要送到网络网关的主节点，然后通过主外部网关节点送出。外部网关节点之前存在主备备份机制，当主外部网关节点故障后，备外部网关节点立即接管南北流量的路由功能，也提升了虚拟化网络的故障备份的能力。

 版权所有 © 超聚变数字技术有限公司



4.18 端口流量镜像

须知

本小节特性需要配套 FusionOne Center。

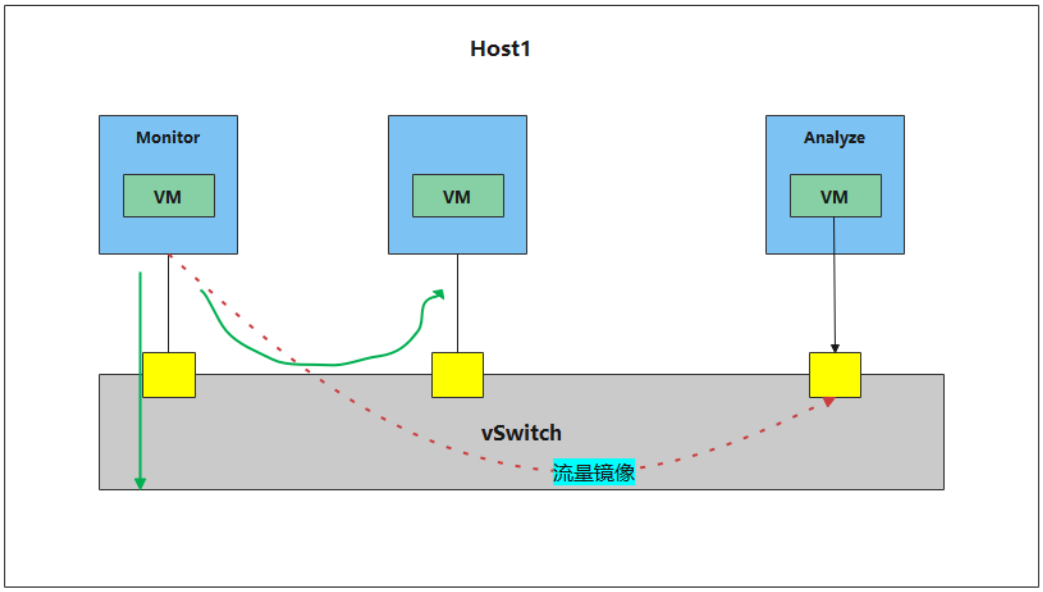
端口镜像是指通过将源端口的数据报文复制到与数据监测设备相连的目的端口，从而实现在目的端口上监测源端口的数据报文信息，用于数据分析和网络问题诊断。

虚拟交换机开启端口镜像功能，可以将该虚拟交换机上一个或多个源端口的数据流量转发到某一个指定端口或者指定 IP 来实现对网络流量的监听，在不严重影响源端口正常吞吐流量的情况下，通过镜像端口对网络的流量进行监控分析。

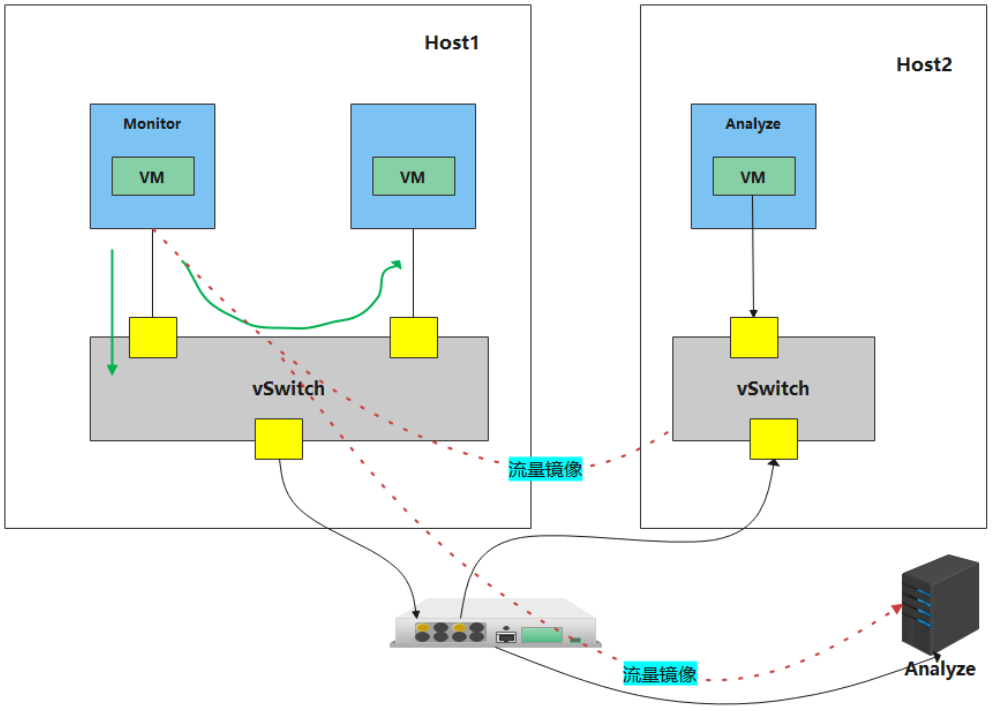
FusionOne Center 提供的端口流量镜像功能主要包含本地端口镜像和远端端口镜像，镜像管理主要包括增加端口镜像，删除端口镜像，编辑端口镜像，显示端口镜像信息等功能。

本地镜像：

 版权所有 © 超聚变数字技术有限公司



远端镜像：



4.19 在线修改虚拟机 IP 地址

 版权所有 © 超聚变数字技术有限公司

须知

本小节特性需要配套 FusionOne Center。

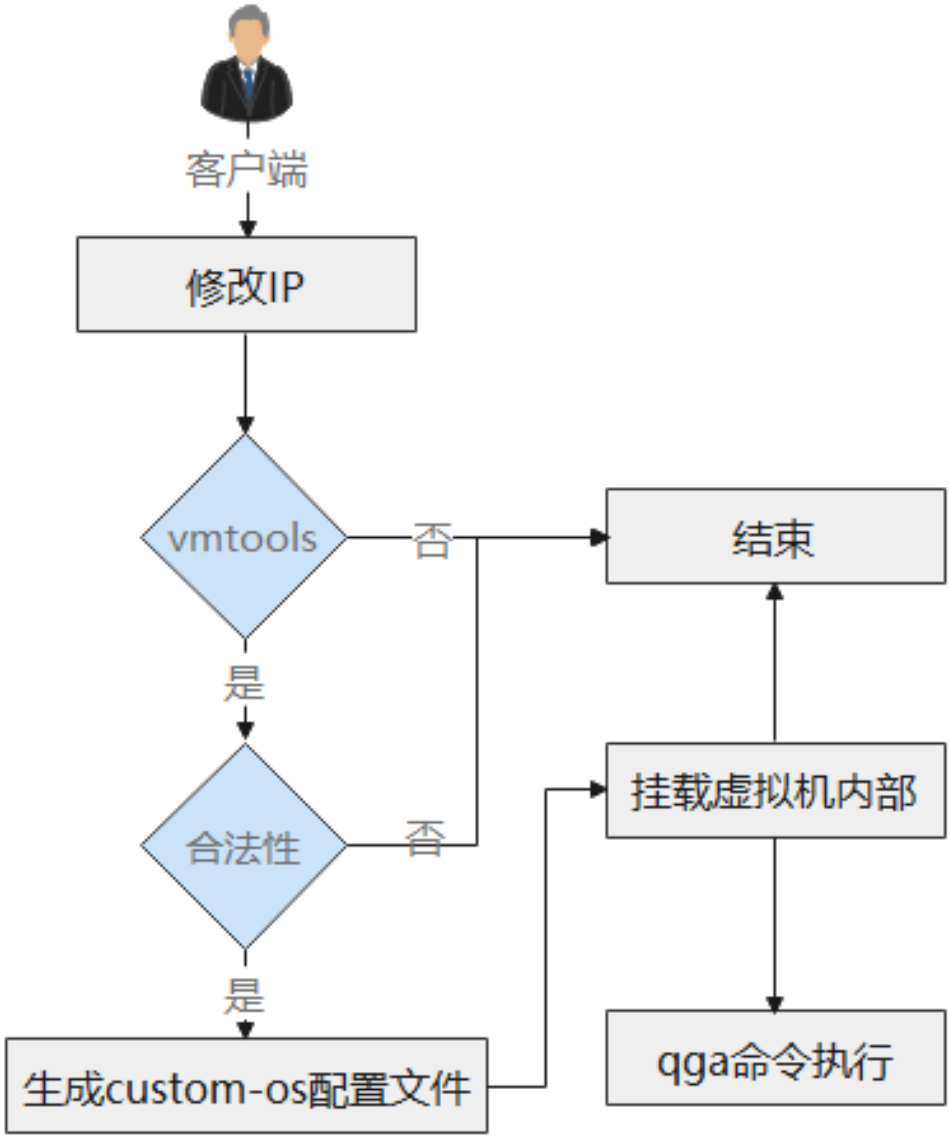
KVM 虚拟机的获取的 IP 地址的方式目前主流的方式如 cloud-init，vmtools 等工具实现，但是如果要实现在线对虚拟机的网卡相关的配置进行配置，则需要一个在虚拟机开机状态下时，还能对虚拟机操作的工具，如 vmtools，cloud-init 由于在虚拟机开机完成后，服务程序会自动退出导致。

FusionOne Compute 针对该场景进行优化，覆盖全场景的 IP 地址的修改，因此新增支持发放后的虚拟机在线修改 IP。

主要技术原理：

将内置修改 IP 服务内置到 VMtools，且虚拟机内部已运行 qga（qemu-guest-agent）的进程。前端在线修改 IP 后，将修改的参数通过 ISO 方式挂载传递到虚拟机内。然后调用 vmtools 的对虚拟机内置的命令如 qga，调用 VMtools 内置的修改 IP 地址的服务，执行后可完成对虚拟机网络的热配置。

 版权所有 © 超聚变数字技术有限公司



4.20 虚拟负载均衡器

须知

本小节特性需要配套 FusionOne Center。

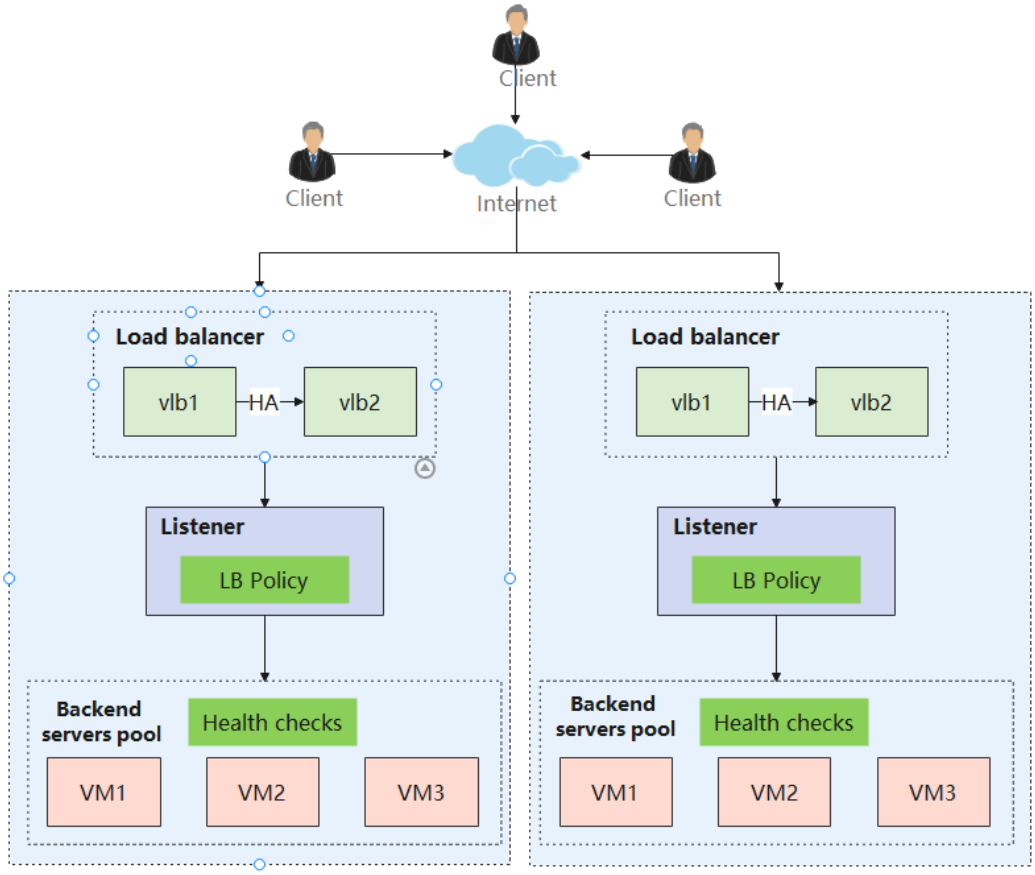
虚拟负载均衡器是将访问流量根据分配策略分发到后端多台服务器的流量分发控制服务。虚拟负载均衡可以通过流量分发扩展应用系统对外的服务能力，同时通过消除单点故障提升应用系统的可用性。

 版权所有 © 超聚变数字技术有限公司

虚拟负载均衡的主要应用场景如：

- 使用虚拟均衡服务为高访问量业务进行流量分发
- 使用虚拟均衡服务和 DRX 为潮汐业务弹性分发流量
- 使用虚拟均衡服务消除单点故障
- 使用虚拟均衡服务跨可用区特性实现业务容灾部署

虚拟负载均衡是基于 Nginx 软件开发实现，Nginx 是一个高性能的 HTTP 和反向代理服务器，通过对其中的健康检查和监听器部分进行了部分增强，用 nginx 作为负载均衡器通过 nginx 的反向代理功能，实现负载均衡的效果，它将客户端的请求分发到后端的多个服务器上，以实现负载均衡和高可用性。



虚拟负载服务提供如下功能：

- 负载均衡管理

用户可通过普通分布式交换机网络或者 SDN 网络创建虚拟负载均衡器，客户端可直接通过公网 IP 或者公网弹性 IP 直接访问。

- 监听器管理

 版权所有 © 超聚变数字技术有限公司

在已创建的虚拟负载均衡器上管理监听器，包括名称，描述，负载均衡协议（TCP/UDP/HTTP/HTTPS）及端口，负载均衡算法（加权轮询，加权最小连接，源 IP 等），会话保持等，并同时创建健康检查。

- 后端服务器组管理

一个或多个后端服务器的逻辑集合，用于将客户端的流量转发到一个或多个后端服务器，满足用户同时处理海量并发业务的需求，用户将提供应用服务的虚拟机添加到指定的监听器后面，设置成员的权重。

- 后端服务器管理

负载均衡器会将客户端的请求转发给后端服务器处理，后端处理应用业务的虚拟机

- 健康检查管理

用户为监听器配置健康检查，用于检查后端云主机的运行状态，包括健康检查协议((TCP/HTTP)，检查周期，超时时间，最大轮询次数等。

表4-3 虚拟负载均衡的规格

类别	子项	规格
负载均衡器	负载均衡器数量	50（个）
	单负载均衡器实例数量。	2（个）
	单负载均衡器实例虚拟机默认规格	4C50G
监听器	监听器数量	100（个）
	单负载均衡器下可添加监听器数量	30（个）
	转发策略数量	100（个）
	单监听器下可添加转发策略数量	5（条）
	单转发策略下可添加转发规则数量	10（条）
	单转发规则可后端服务器组数量	1（条）
后端服	后端服务器组数量	30（个）

 版权所有 © 超聚变数字技术有限公司

类别	子项	规格
务器组	单后端服务器组下可添加后端服务器数量	50（个）
健康检查器	检查间隔	5-50(秒)，默认 10（秒）
	重试次数	1-5（次），默认 3（次）
	超时时间	1-20（秒），默认 2（秒）

 版权所有 © 超聚变数字技术有限公司

5 存储虚拟化

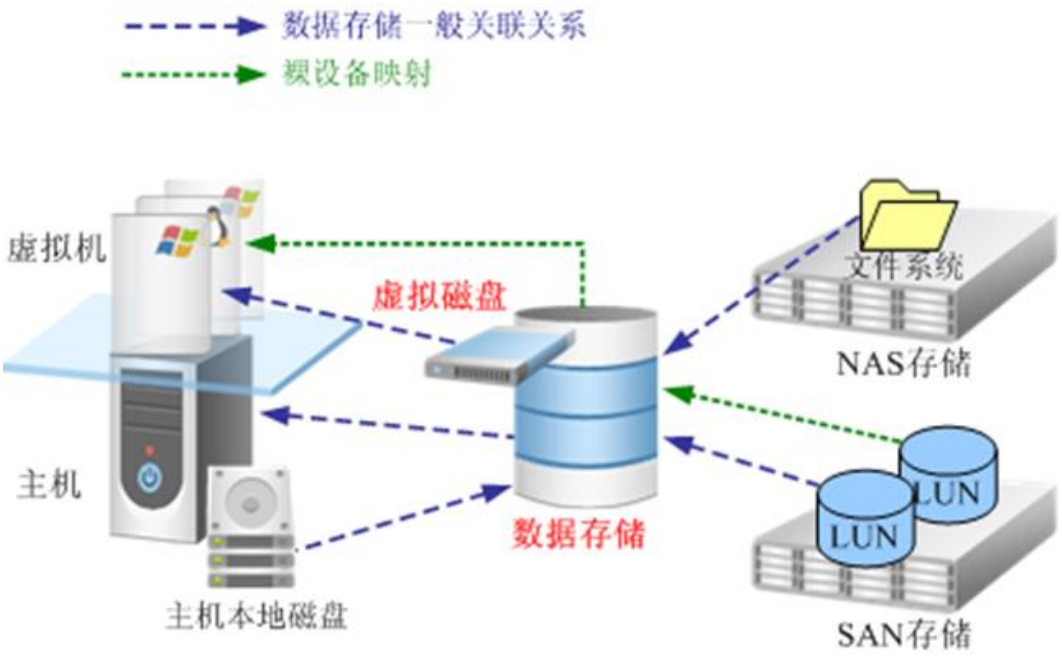
5.1 概述
5.2 虚拟机直通 SAN
5.3 虚拟文件系统（IP/FC SAN）
5.4 NAS 存储文件系统
5.5 磁盘模式配置
5.6 自动精简配置
5.7 虚拟机备份

5.1 概述

存储虚拟化为上层虚拟机和其它应用提供统一存储服务，消除资源组织模式形成资源孤岛、硬件资源利用不均、整体资源浪费等问题。

- SAN 存储：根据不同接口，支持 IP SAN、FC SAN 存储对接。
- NAS 存储服务：提供 NFS 和 CIFS 标准接口，对接外部 NAS 存储，用于文件共享。
- 本地存储：本地 SAS、SATA 磁盘对接，格式化为 LVM 管理的存储。

 版权所有 © 超聚变数字技术有限公司



5.2 虚拟机直通 SAN

- 概述

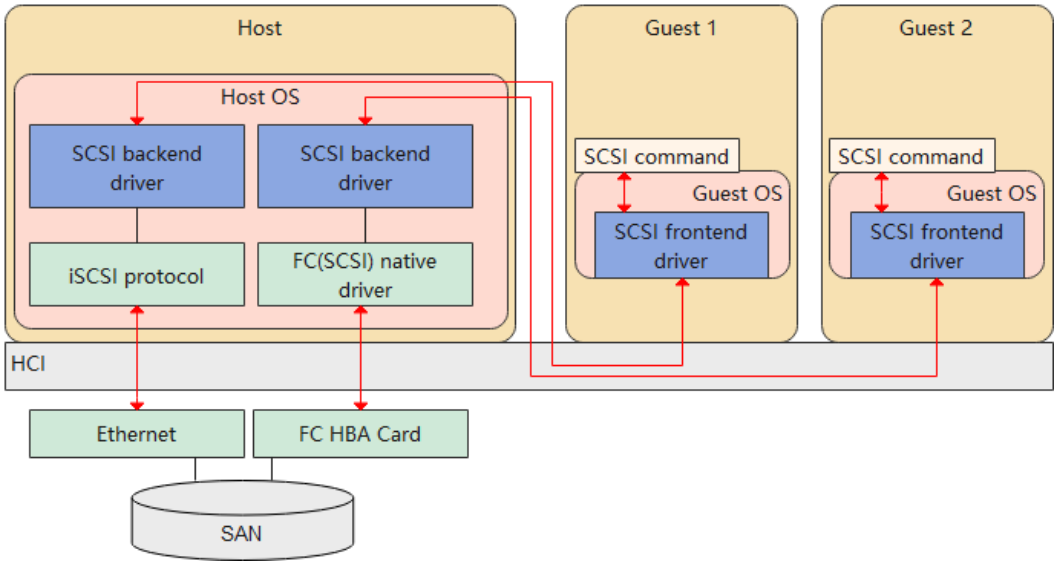
  存储直通技术(另称裸设备映射，RDM)为虚拟机提供了一种机制来直接访问物理存储上的 LUN，这样虚拟机中的业务能够直接访问存储设备或对存储设备下发控制命令。通过使用裸设备映射，可以让虚拟机识别 SCSI 磁盘，实现在虚拟机内部下发 SCSI 命令，然后发给主机再透传给存储设备进行处理，能够很好的支撑如 Oracle RAC 等业务在虚拟机上运行，降低大量使用物理服务器的成本。

- 原理

  裸设备映射将整个 LUN 映射到虚拟机中，技术实现上分两部分。一部分为 SCSI 前端驱动，通过 PV Driver(准虚拟化驱动)安装在虚拟机内部；另一部分是 SCSI 后端驱动，作为 Linux 内核模块运行在主机系统中。当虚拟机部署相关业务需使用裸设备映射，对映射到虚拟机内部的 SCSI 磁盘下发 SCSI 命令时，SCSI 前端驱动将 SCSI 命令交给后端，后端 SCSI 驱动再将 SCSI 命令透传给实际的物理存储设备，并将物理存储设备的应答返回。

  原理流程图如下所示:

 版权所有 © 超聚变数字技术有限公司



- 应用场景

  应用于集群中一个或多个 CNA 节点上运行了如 Oracle RAC 业务的虚拟机，用户划分好独立的 LUN，然后配置好带 SCSI 磁盘的配置文件，在虚拟机启动或者挂载磁盘的过程中，给虚拟机配置 SCSI 磁盘，让虚拟机可以使用 SCSI 磁盘。SCSI 磁盘可以配置成虚拟机的系统盘或数据盘。当虚拟机内部安装 PV Driver，并处于运行状态，在虚拟机的 SCSI 磁盘上部署应用，就可以正常使用这些业务。

5.3 虚拟文件系统（IP/FC SAN）

- 概述

  虚拟镜像管理系统 VIMS (Virtual Image Management System) 是一种高性能的集群文件系统，通过连接 SAN 存储，将 LUN 格式化成 ocfs2 文件系统格式，然后挂载到主机节点上进行使用。VIMS 文件系统使虚拟化技术的应用超出了单个存储系统的限制，可让多个虚拟机共同访问一个整合的集群式存储池，从而显著提高了资源利用率。

  存储区域网络 SAN，是一种高速的、专门用于存储操作的网络，通常独立于计算机局域网（LAN）。提供主机和存储系统之间的数据传输，网络内部数据传输的效率快。常见的架构有 FC SAN、IP SAN，具有存储容量利用率高、兼容性高、传输距离远、高带宽、主机、存储设备可以独立扩展等优点。

  FusionOne Compute 通过在 IPSan 或 FCSan 的卷上创建一个 VIMS 共享文件系统，并在多个节点上挂载，实现了多个节点共享一个存储卷。可以在此共享文件系统上创建虚拟机，由于数据在多个节点间共享，可以实现虚拟机迁移时免迁移存储数据。

 版权所有 © 超聚变数字技术有限公司

- 原理

  VIMS 文件系统的实现依赖 ocfs2 组件，在 CNA 节点上已经预置了对应的工具和驱动。

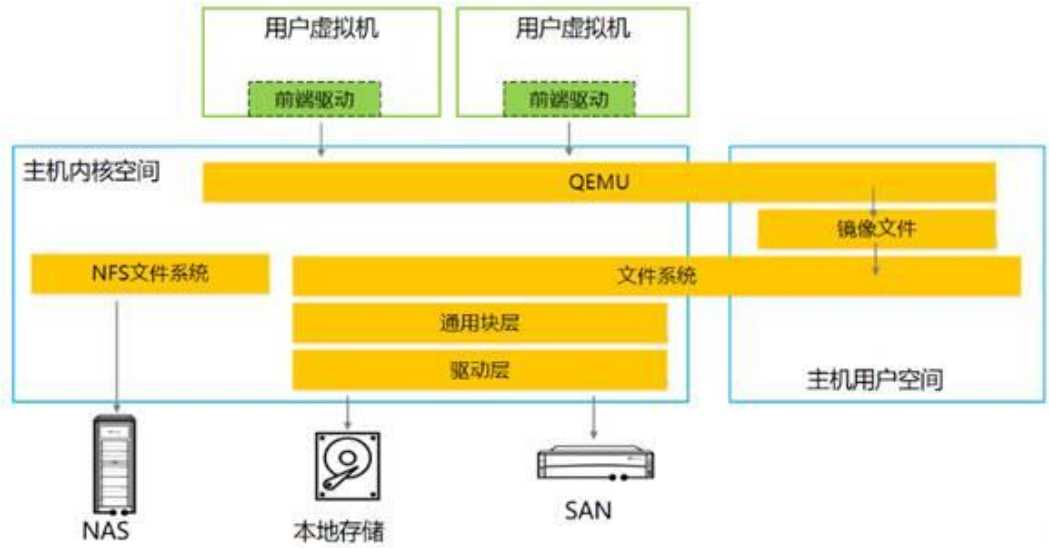
  一个节点添加 VIMS 数据存储分为如下几个步骤：

  a. 用 ocfs2 组件的工具来创建 VIMS 集群(只有第一个加入集群的节点需要创建集群)，并将该节点加入到集群中。集群中的节点通过分布式锁管理器（DLM）来实现对共享存储卷的访问控制（一个 LUN 对应一个存储设备，一个存储设备对应一个 VIMS 数据存储，对应一个 DLM 域，每个 DLM 域对应一个 DLM 锁）。

  b. 将存储设备格式化成 ocfs2 文件系统（只有第一个加入集群的节点需要执行格式化操作）。

  c. 使用 mount.ocfs2 工具挂载到节点上。

  每个 VIMS 数据存储在节点上都有对应的目录，可以通过访问该目录来访问 SAN 存储。ocfs2 文件系统具有分层的目录结构，每个用户虚拟机的镜像文件可以存放在 VIMS 数据存储对应目录的单个文件夹中，便于管理。对应的架构图如下所示：



VIMS 集群还提供了 fence 机制来保证集群高可用，当网络心跳或磁盘心跳异常时，都可能触发 fence 机制（网络心跳是指 VIMS 集群中各节点之间的心跳，磁盘心跳是指各节点访问磁盘的心跳）。FusionOne Compute 的 VIMS fence 处理方案是故障节点在内核态释放自身锁资源，并通知用户态解挂载异常的 VIMS 数据存储。结合主机集群的 HA 配置，可以让故障节点上使用了异常的 VIMS 数据存储的虚拟机 HA 到其它正常节点上运行。

- 应用场景

 版权所有 © 超聚变数字技术有限公司

VIMS 数据存储兼容 IPSAN，FCSAN，支持数据存储动态扩容，支持创建固定空间磁盘，动态空间磁盘和差分磁盘。主要应用于需要使用存储迁移，快照，链接克隆等高级特性的用户虚拟机。

5.4 NAS 存储文件系统

- 支持对接 NAS 存储，NAS 存储设备通过 NFS 协议，在网络中以共享文件目录的形式提供存储资源。主机与 NAS 存储设备之间通过普通网线连接，以 TCP/IP 协议访问存储设备。
- 支持在 NAS 存储上分配虚拟磁盘。

5.5 磁盘模式配置

- 磁盘总线类型

  VIRTIO： 是一种高性能的虚拟化驱动技术，旨在提供直接 I/O 访问，以实现更好的虚拟机磁盘性能。在 KVM 虚拟化环境中，VIRTIO 通常被认为是性能最好的选择，因为它通过内核态驱动程序与虚拟机磁盘进行直接通信，避免了对数据的多次拷贝，从而提高了性能。

  IDE（Integrated Drive Electronics）：是一种传统的计算机磁盘接口标准，用于连接硬盘驱动器到计算机系统。在虚拟化环境中，IDE 通常被用于连接虚拟机的虚拟硬盘，它是一种简单且易于使用的接口，适用于一些基本的虚拟化场景。

  SCSI（Small Computer System Interface）：是一种通用的磁盘接口标准，具有较强的扩展性和性能。在虚拟化环境中，SCSI 接口通常用于连接高性能、高需求的虚拟机磁盘，例如需要更好的并发性能或支持高级存储特性的场景。

  创建虚拟机时，数据存储上创建的磁盘可以挂载在 VIRTIO、IDE、SCSI 总线上，VIRTIO 性能较好，IDE 性能较差，仅用于虚拟机镜像制作及虚拟机光驱。选择合适的磁盘接口取决于你的虚拟化环境需求。一般来说，如果性能是首要考量因素，VIRTIO 是一个很好的选择；而对于一般用途的虚拟机，IDE 和 SCSI 也都是常见的选择。

- 磁盘 IO 模式

  native：性能优先，可靠性有所下降。在 KVM 中的本地磁盘 I/O 模式通常指的是直接将宿主机上的物理磁盘直通给虚拟机，也称为裸设备映射。这样虚拟机

 版权所有 © 超聚变数字技术有限公司

可以绕过宿主机文件系统和虚拟磁盘镜像，直接与物理磁盘进行读写操作，从而获得更低的延迟和更高的性能。

theads：通过起线程方式实现异步 IO，可靠性高。在 KVM 中的线程磁盘 I/O 模式可能指的是通过多线程并发执行磁盘 I/O 操作来提高虚拟机的磁盘访问性能。通过并行化处理磁盘请求，可以加快数据传输速度和提高系统的吞吐量。

dataplane：开启单独的 IO 线程，提升虚拟机多磁盘性能，但是对宿主的资源占用会增加。在 KVM 中的数据平面磁盘 I/O 模式通常涉及使用专门的数据处理路径来优化磁盘访问。这可以包括使用专用的设备驱动程序、虚拟化技术或者硬件加速功能，以降低磁盘 I/O 的延迟并提高性能。

在 KVM 虚拟化环境中，如上是用于优化虚拟机磁盘 I/O 性能的三种常见模式。选择适当的模式取决于具体的应用需求和性能目标，以确保虚拟机系统能够获得最佳的磁盘访问性能。

- 磁盘缓存模式

  none：关闭主机缓存，数据从虚拟磁盘缓存中读写，这样可以确保数据的实时性，但通常会导致较低的性能，因为每次读写操作都需要直接与磁盘进行通信。

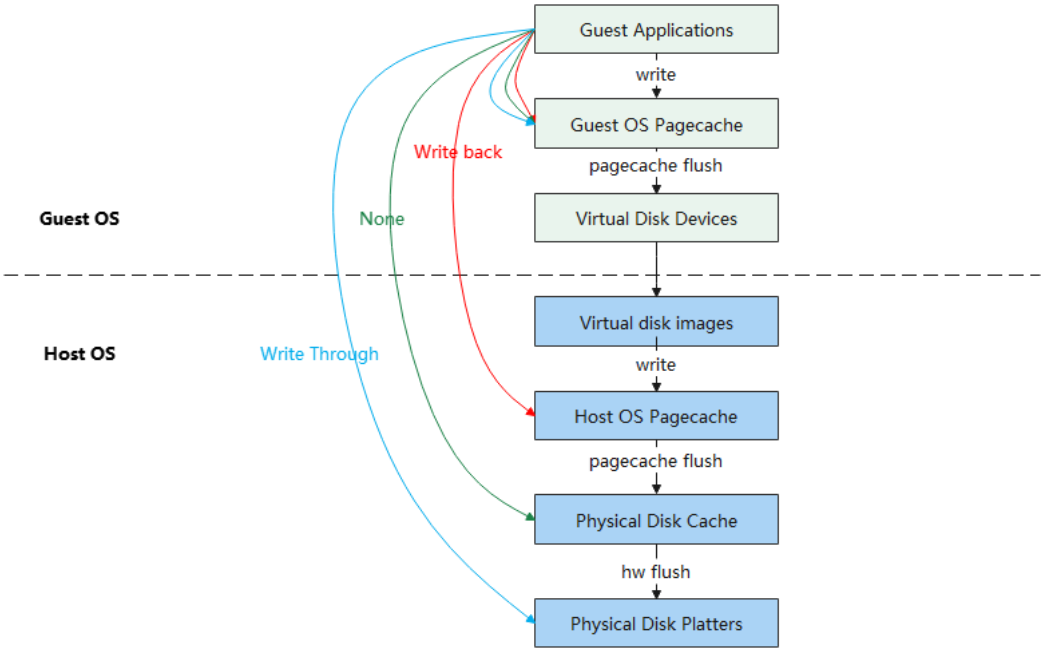
  directsync：关闭主机和虚拟磁盘缓存，直接读写物理磁盘，这意味着每次写操作都要等待数据完全写入磁盘后才返回成功，这样可以确保数据的持久性，但可能会导致较低的性能。

  writethrough：关闭虚拟磁盘写缓存，数据直接被写入磁盘，有读缓存，这样可以提高写入性能，但会牺牲一定的数据持久性，因为在数据写入缓存后，如果系统崩溃，则可能会丢失部分数据。

  writeback：数据在写入磁盘前，首先写入缓存，写缓存成功之后，写操作即返回，数据并没有立即同步到磁盘。这种方式可以提高写入性能，因为不需要等待数据同步到磁盘上，但也增加了数据丢失的风险，因为在数据写入磁盘之前发生系统故障时可能会丢失缓存中的数据。

  如下图所示：

 版权所有 © 超聚变数字技术有限公司



从性能对比：none>writeback>writethrough；

从数据一致性对比：writethrough>none>writeback；

directsync 类似于 writethrough，属于新增缓存模式，性能应该略强于 writethrough，一致性很高。

5.6 自动精简配置

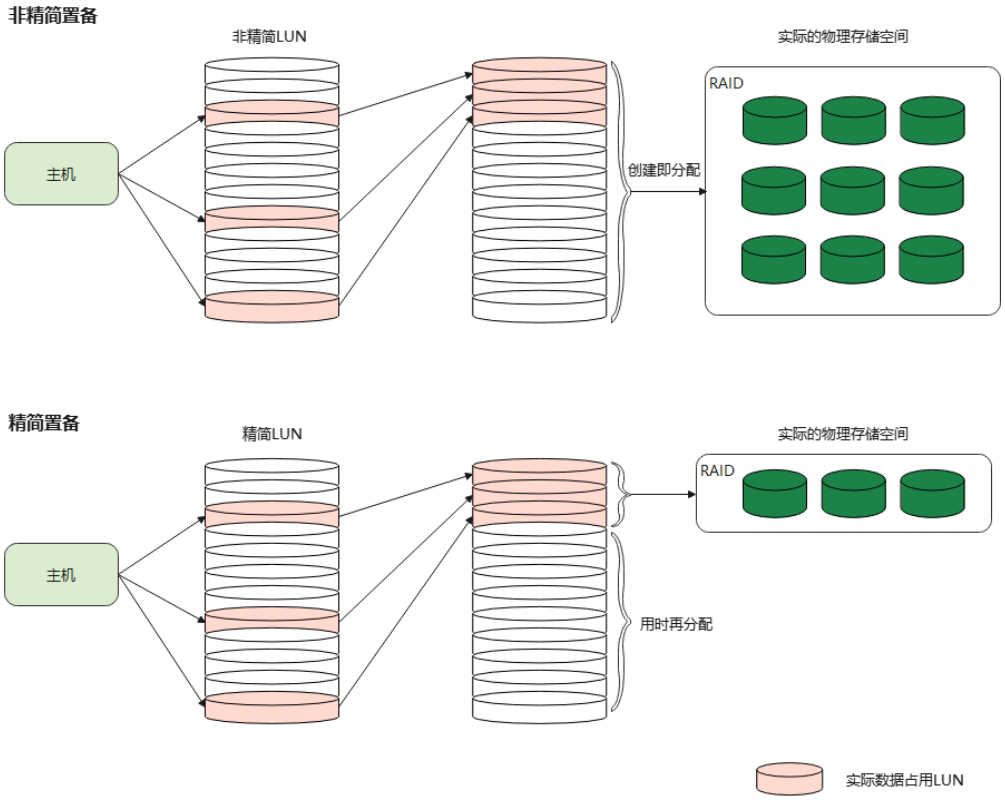
- 概述

  存储自动精简配置（Thin Provisioning），是一种智能高效的容量分配和管理技术，可以为客户虚拟出比实际物理存储更大的虚拟存储空间，为用户提供存储超分配的能力。只有写入数据的虚拟存储空间才能真正分配到物理存储，未写入的虚拟存储空间不占用物理存储资源。FusionOne Compute 虚拟化存储提供多种模式选择，普通模式、精简模式、厚置备延迟分配等模式供用户选择。自动精简配置不依赖于远端存储设备。存储自动精简配置的应用场景主要针对虚拟机的用户数据卷.。

- 原理

  如下图所示普通非精简和精简两种配置下同样的实际数据占用大小所对应的物理存储空间情况，精简配置下明显比非精简所占用的物理存储资源少，用多少分配多少。

 版权所有 © 超聚变数字技术有限公司



- 应用场景

  系统不会一次性划分过大的空间给某项应用，而是根据应用实际所需要的容量，多次分量地分配给该应用。当精简 LUN 容量不能满足需要时，系统会从精简池中补充分配存储空间。适用于当企业或运营商宣称提供的容量较大，但用户实际却用不完时，可通过存储自动精简配置帮助企业或运营商大幅降低存储的初始投资成本。

5.7 虚拟机备份

随着虚拟化趋势的日渐普及，企业中越来越多的核心数据开始存放于虚拟化环境中。因而现代企业对于虚拟化数据的保护需求更加迫切，而虚拟化平台中的数据备份和恢复效率也成为企业进行数据保护所关注的核心问题。FusionOne HCI 支持定时备份、立即备份、原机恢复、异机恢复、磁盘恢复等功能，备份时会备份虚拟机配置信息、位图文件以及磁盘文件，其中虚拟机配置信息包含 CPU、内存、硬盘、卷位置、存储位置、系统信息、网络信息等。同时支持 CDP 实时备份、磁盘挂载等功能，对生产机数据进行一次初始化备份，并持续监控数据变化，将初始化备份数据和实时捕获到的变化数据存储到备份服务器中，最终实现持续数据保护的目的。其中实时捕获到的变化数据块都会加上时间戳，将变化数据的元数据信息记录到日志中。根据日志可以实

 版权所有 © 超聚变数字技术有限公司

现任意时间点恢复、挂载、细粒度恢复等，满足企业更高要求的恢复时间目标 （RTO）和恢复点目标（RPO）。

5.7.1 定时备份和立即备份

定时备份，指的是一种根据备份策略定时执行数据备份的技术，即在指定时间发起备份任务，可自定义备份计划；立即备份，指的是对虚拟机发起即时的一次备份任务。备份任务发起后，将指定的数据备份到介质服务器中，当数据损坏时，使用备份数据能够进行数据的恢复。备份方式支持永久增量备份。备份过程支持加密、目标端重删。支持对备份集进行校验。

5.7.2 VIMS CBT 备份

虚拟化存储（包括 VIMS 存储和本地存储）使用了 CBT（Changed Block Tracking）技术。CBT 能够识别和跟踪自上次备份后虚拟机被修改过的块，这些修改会存放到文件中。当启动了 CBT 备份之后，备份软件通过获取 CBT 跟踪的变化的数据块，可以进行全量或者增量备份。

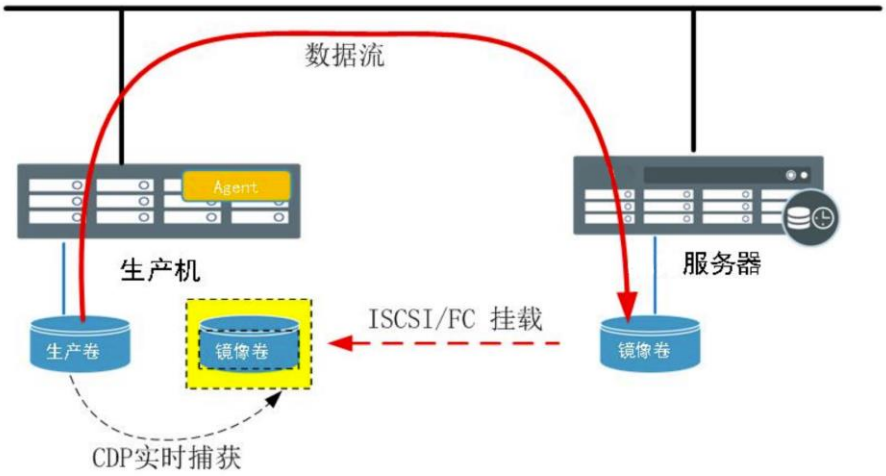
5.7.3 CDP 备份

CDP（Continuous Data Protection）是持续数据保护的简称，对生产机数据进行一次初始化备份，并持续监控数据变化，将初始化备份数据和实时捕获到的变化数据存储到备份服务器中，最终实现持续数据保护的目的。其中实时捕获到的变化数据块都会加上时间戳，将变化数据的元数据信息记录到日志中。根据日志可以实现任意时间点恢复、挂载、细粒度恢复等，满足企业更高要求的恢复时间目标（RTO）和恢复点目标（RPO）。持续数据保护（CDP）可分为文件级 CDP 和卷级 CDP 两种方式，文件级 CDP 监控的对象是文件，卷级 CDP 监控的对象是磁盘、卷或者分区。卷级 CDP 是基于扇区层面的，系统有数据变化，当写入磁盘时，变化的数据在扇区层将被捕获到。本系统采用的是卷级 CDP 技术，通过主机安装代理的实现方式，在系统运行过程中，每当数据块有新的生产数据写入时，数据的拷贝都会被捕获，并存放在另外的独立存储设备中实现镜像保护。同时按客户设定的 CDP 任务计划持续产生有一致性保障的 CDP 快照点，以供后续快速恢复使用。

5.7.3.1 原理描述

在生产机上安装 CDP 实时捕获代理，通过磁盘过滤驱动实时捕获生产机磁盘上的 I/O 数据变化，并将变化的数据块按原生产卷当时写入的 I/O 大小复制并实时同步到镜像卷上。CDP 采用的是异地同步方式，对原生产机影响极低。持续数据保护原理图如下：

 版权所有 © 超聚变数字技术有限公司



5.7.3.2 使用场景

本地数据实时保护



步骤 1 生产机安装 CDP 实时捕获代理 Agent。

步骤 2 服务器创建镜像卷，镜像卷跟生产机的生产卷大小一样。

步骤 3 服务器通过 iSCSI 协议将镜像卷远程挂载到生产机上。

步骤 4 生产机首先做一次初始化备份，解析磁盘有效数据分布，只备份有效数据，数据写到镜像卷同样的位置。初始化备份期间，驱动只记录变化位置。

步骤 5 初始化完成后，通过驱动记录的变化位置，从磁盘上读取数据，并同步到镜像卷上。

步骤 6 直到驱动记录的变化位置的数据都同步完成后，驱动切换到实时同步模式，将捕获的数据直接写入镜像卷的对应位置。

 版权所有 © 超聚变数字技术有限公司

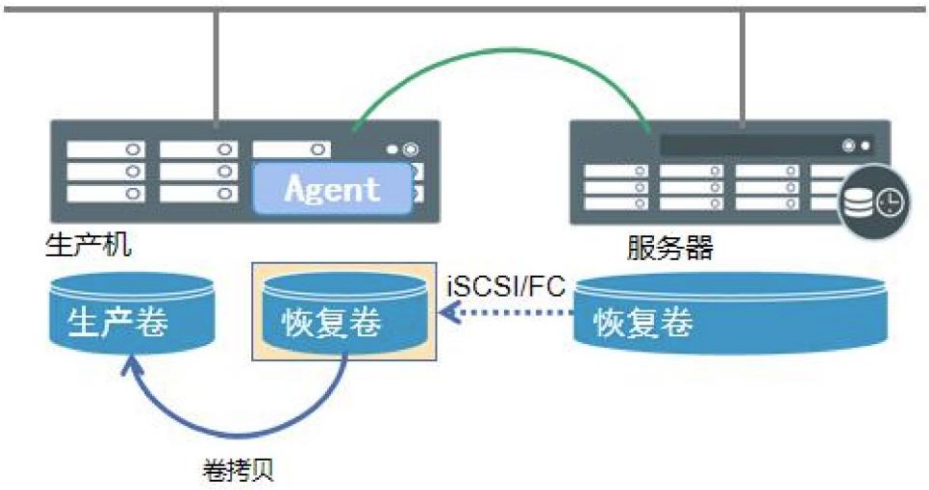
步骤 7 服务器将镜像卷写操作的数据块打上时间戳并记录块序号，写入备份文件和日志文件中。

----结束

技术特点如下：

1. 只备份有效数据，大大减少备份数据量。
2. 生产机正常重启后，驱动可继续有效捕获。不需要增量初始化备份来校准数据，大大缩短续传时间窗口。
3. 异步同步方式，对系统性能影响小。
4. 基于卷级 CDP，面向的是扇区数据，不关心具体的应用变化内容和应用相关的文件，兼容性广。
5. 初始化备份支持传输通道中数据的加密，保证数据安全传输。
6. 初始化备份支持压缩、重删，适用于窄带宽。
7. 支持定时快照，创建快照会刷新缓存数据，确保数据一致性。

任意时间点恢复



步骤 1 服务器通过 CDP 备份时记录 IO 日志，快速创建指定时间点的恢复卷。

步骤 2 服务器通过 iSCSI 协议将恢复卷映射出去。

步骤 3 恢复代理将恢复卷挂载到生产机上。

步骤 4 通过卷拷贝程序，将恢复卷数据拷贝到生产卷。

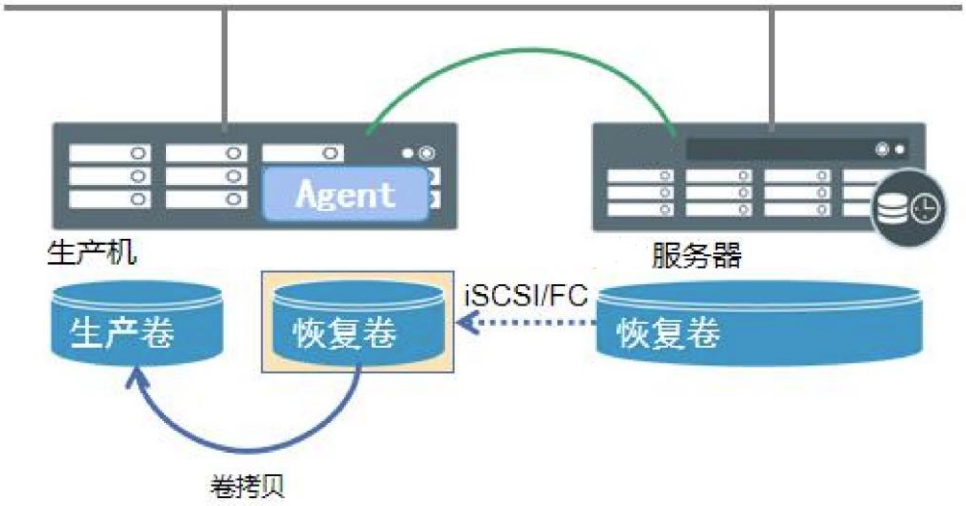
----结束

 版权所有 © 超聚变数字技术有限公司

技术特点如下：

1. 只恢复有效数据，大大缩短恢复窗口。
2. 传输通道支持数据加密，保证数据传输安全。
3. 传输通道支持数据压缩，适用于窄带宽。
4. 支持一键恢复，原机恢复自动匹配对应磁盘，减少人工匹配工程。
5. 支持在线热恢复，生产机只恢复数据盘时，可以不进入 LiveCD 恢复，操作简单方便。

- 细粒度恢复



步骤 1 服务器通过 CDP 备份时记录 IO 日志，快速创建指定时间点的恢复卷。

步骤 2 服务器通过 iSCSI 协议将恢复卷映射出去。

步骤 3 恢复代理将恢复卷挂载到生产机上。

步骤 4 从恢复卷中，基于文件系统层面，将指定文件恢复到生产卷。

----结束

- 技术特点如下：

1. 支持只恢复单个文件，缩短恢复周期。
2. 支持恢复到异机。
3. 方便文件测试对比。

 版权所有 © 超聚变数字技术有限公司

6 统一管理

6.1 设备管理
6.2 监控管理
6.3 运维管理
6.4 多云管理

6.1 设备管理

设备纳管功能，支持服务器、第三方服务器、网络多种设备及 OS 的管理，以满足各种业务诉求。支持对纳管后的设备进行管理、数据采集。只要管理网络连通且满足网络要求（时延≤100ms、丢包≤0.1%，带宽≥50Mbit/s，设备无 IP 冲突），就可以纳管跨地域的设备。

须知

本小节特性需要配套 FusionOne Center。

6.1.1 可视化节点

支持服务器节点自动纳管，支持节点带内带外信息统一展示，纳管后支持对已接入的服务器设备部件信息进行真实前后面板的呈现和管理：

- 电源信息：名称、状态、序列号、额定功率、固件版本、协议、制造商、丝印号、工作模式、额定功率。

 版权所有 © 超聚变数字技术有限公司

- 风扇信息：名称、状态、部件编码、型号、速率比和转速。
- 处理器信息：名称、状态、厂商、型号、当前频率、温度、位置、核数/线程数、缓存、部件编码、启用状态、序列号。
- 内存信息：名称、厂商、部件编码、当前频率、主频、容量、序列号、类型、位宽、最小电压、状态。
- 存储信息：支持查看 Raid 卡名称、类型、固件版本、健康状态、模式、配置版本、内存大小等信息；逻辑卷名称、状态、Raid 级别、容量、度策略、写策略等信息；物理磁盘接口类型、制造商、型号、健康状态、系列号、容量、通电时长等信息。
- 网卡信息：名称、芯片型号、状态、网口数目、厂商、所在单板位置、芯片厂商、PCB 版本。
- 端口信息：端口、网口名称、连接状态、MAC 地址、介质类型。
- 整机信息：iBMC 主机名、设备序列号、设备型号、资产标签、主机名称、设备别名、所属组、设备标签、告警信息、健康状态、电源状态和 UID 灯状态。
- PCIe 卡：序列号、名称、型号、类型、固件版本、状态、厂商、所在单板位置和状态。
- 统计数据：CPU 利用率、内存利用率逻辑盘容量、逻辑盘占用率、物理盘容量、输入功率、输出功率、当前使用功耗、风扇转速、风扇转速百分比、入风口温度、CPU 温度、GPU 利用率、GPU 显存利用率。

6.1.2 网络设备纳管

支持纳管平台关联的网络设备，纳管后支持对已接入的网络设备部件信息进行真实前后面板的呈现和管理：

- 机框信息：框号、名称、管理状态、运行状态、型号、固件版本号、软件版本、备份状态、序列号、CPU 名称、CPU 利用率、内存名称、内存利用率、温度、功率。
- 板卡信息：父设备名称、板号、名称、管理状态、运行状态、型号、固件版本号、软件版本、备份状态、序列号、CPU 名称、CPU 利用率、内存名称、内存利用率、温度、功率。
- 电源信息：名称、板号、管理状态、运行状态、型号、固件版本号、软件版本号、备用状态、序列号、温度、电压、电流。
- 风扇信息：名称、板号、管理状态、运行状态、型号、固件版本号、软件版本号、备用状态、序列号、温度、转速。

 版权所有 © 超聚变数字技术有限公司

- 接口信息：名称、类型、管理状态、运行状态、速率、光模块在为状态、MAC 地址、MTU、光模块信息、端口邻居信息。
- 统计数据：CPU 利用率、内存利用率、系统实时功率、温度、流量统计。

图6-1 部件信息



6.1.3 操作系统纳管

支持平台内的主机、管理虚拟机的 OS 自动纳管，支持手动单个或批量纳管业务虚拟机，支持纳管时自动安装性能监控组件。纳管速度快具体性能规格如下：

表6-1 性能规格

功能	操作耗时（分钟/台）	操作耗时（分钟/20 台）	操作耗时（分钟/100 台）
接入操作系统耗时	0.5	1	5

支持添加业务 OS 账号并对账号进行管理，该操作是 8.7 章节自动化作业的前置条件，OS 添加后还具备管理账号的刷新、OS 密码的修改功能。

针对 OS 的管理详细支持以下功能：

- 基本信息：主机 IP 地址、管理状态、OS 种类及类型、OS 架构和版本、主机名、iBMC IP 地址、服务器类型、机架号、槽位号、节点类型。
- 账号信息：账号录入、管理账号刷新、纳管 OS 账号修改（支持同步修改 OS 的账号）。
- 性能监控：CPU 使用率、内存使用率、网卡下载速度、网卡上传速度、磁盘读写速度、文件系统使用率。

 版权所有 © 超聚变数字技术有限公司

图6-2 基本信息

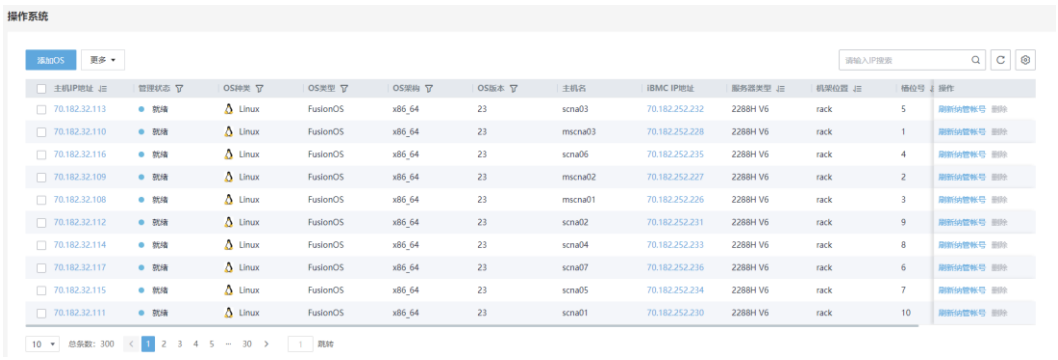
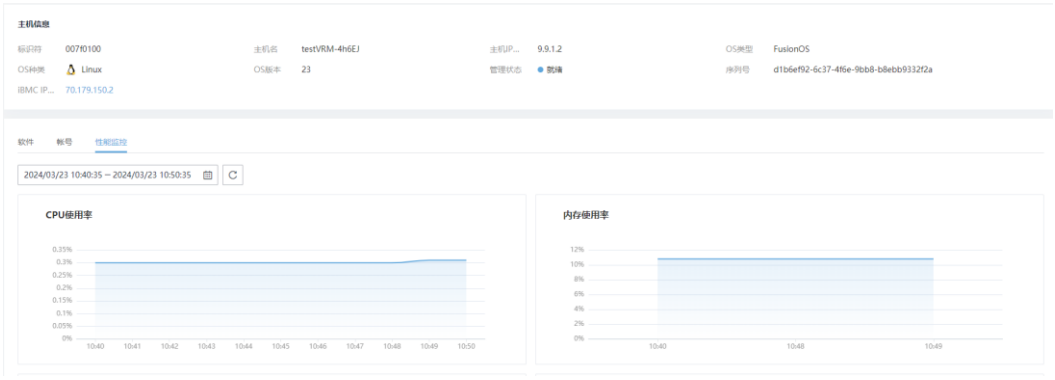


图6-3 性能监控



6.2 监控管理

须知

本小节特性需要配套 FusionOne Center。

6.2.1 统一告警与事件

告警管理提供了告警数据接收、处理、通知等功能，支持告警的全生命周期管理，帮助运维人员根据告警信息快速排除故障。

- 告警统一：汇聚软硬件各子系统生成的告警，在 FusionOne Center 界面统一展示。

 版权所有 © 超聚变数字技术有限公司

- 告警管理：支持查看实时、历史告警，支持设置告警屏蔽规则，支持设置告警保存周期、支持设置事件转告警规则
- 告警通知：支持通过 Email、短信、企业微信、钉钉，按照客户制定的规则，通知到运维人员。
- 告警上报：支持通过 SnmpTrap 方式，将 HCI 告警统一上报到第三方运维管理系统。

6.2.2 可视化性能监控

监控计算集群、计算节点、虚拟机、服务器、交换机、OS 等资源，展示相关资源的性能指标历史 KPI 趋势图和 TOP 统计，帮助管理员评估系统中不同资源对象的运行状态和变化趋势，进而全方位了解资源运行状况。

- 计算集群、计算节点、虚拟机的性能指标数据，可查询最大 6 个月时间范围。
- 存储集群、存储节点、存储池，可查询最大 3 个月时间范围。
- 服务器、交换机、OS，默认保存 14 天，如果需要查询更长时间范围，需要扩大 FOC 虚拟机的内存和硬盘资源。
- 支持自定义指标，进行历史 KPI 查询和 TOP 统计。
- 内置监控仪表盘，从多维度展示各资源的性能指标实时统计和历史趋势展示。
- 可灵活启停要监控的指标，可动态设置指标数据采集周期。

6.2.3 节点可视化运维

FusionOne Center 使用服务器前后面板真实物理视图叠加器件状态指示灯、高亮效果等展示服务器的信息和状态。相比传统展示手段，展示效果更加形象、直观，方便用户查看设备、器件的信息和状态，了解服务器故障器件发生位置，给用户带来良好的交互体验。

- 展示的信息和状态：支持展示整机、硬盘、网口、电源的信息和状态。
- 物理真实视图兼容型号：2288H V6（8 盘直通 8*2.5）-32DIMM、2288H V6（12 盘直通 12*3.5）-32DIMM、2288H V6（12 盘直通 12*3.5）-16DIMM、2288H V6（25 盘 EXP）-32DIMM、2288H V6（24NVMe）-32DIMM、2288H V7（8 盘直通 8*2.5)、2288H V7（12 盘直通 12*3.5)、2488H V6（ 4xSAS/SATA+16xNVMe)、2488H V6（4P+8 盘 8x2.5 英寸硬盘直通配置）
- 常用操作：电源上下电、UID 灯开关、硬盘灯开关。

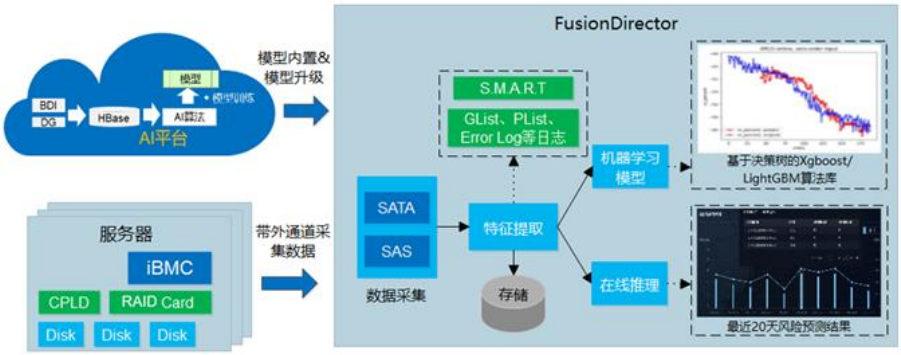
 版权所有 © 超聚变数字技术有限公司

6.2.4 硬件故障预测

FusionOne Center 支持硬盘和内存故障预测，主动预防，提升硬件可靠性，多维分析，全面监控硬件状态。

6.2.4.1 硬盘故障预测

结合超聚变硬件管理的经验积累，结合 BMC 软件，云平台支持对超聚变品牌服务器 SAS/SATA 机械盘、SSD 盘的故障预测。



1. 硬盘数据采集
   - RAID 带外采集通道。

     数据流向为“Disk > RAID > iBMC > FusionDirector”，带外通道采集不影响用户业务和带内网络带宽，且监控数据会在 iBMC 持久化一段时间。极端情况时，OS 宕机数据丢失后，不影响 FusionOne Center 通过带外通道获取这部分数据。
   - S.M.A.R.T（Self-Monitoring Analysis and Reporting Technology）

     作为硬盘监控规范，发展至今已 15 年，业界对其认可度较高，例行硬盘巡检时通常也是依赖 S.M.A.R.T 信息。SMART 功能不断从硬盘上的各个传感器收集信息，并把信息保存在硬盘的系统保留区（service area）内。S.M.A.R.T 监控项共提供 200 余项指标，但并非所有的指标都合适于预测，数据特征的选取/预处理是算法能够收到良好效果的关键。FusionOne Center 联合第三方 AI 训练平台收集 30 万台服务器硬盘数据，结合专家经验对坏盘数据进行分析，提取十余项数据作为算法模型的输入。
   - 硬盘 Log 日志

 版权所有 © 超聚变数字技术有限公司

业界硬盘故障预测绝大部分实践都是通过 S.M.A.R.T 来做的，而 SAS 接口的机械硬盘或者 SSD 硬盘无法提供 S.M.A.R.T 监控数据。SAS 接口和 SSD 硬盘提供了 SCSI Log Page 信息，这些 Log Page 中的信息详细记录当前硬盘的状态和当前性能指标，例如 grown defect list（GList）、non-medium error、read/write/verify uncorrected error 等日志。FusionOne Center 通过带外采集通道获取上述日志，提取关键数据特征，覆盖 SAS 硬盘的故障预测。

2. 机器学习模型调优

   基于模型算法进行训练，通过有限次迭代，优化模型算法，输出为推理算法模型。FusionOne Center 采用离线推理的方式，在第三方 AI 训练平台对数据进行分析、特征提取以及模型调优。

3. 推理模型选择

   FusionOne Center 通过对百万 SAS 硬盘、数十万 SATA 硬盘进行长达两年的持续数据收集、测试验证以及结合专家分析经验，最终采用 Xgboost、LightGBM 等算法构建出最佳硬盘失效模型。

4. 故障预测支持两种场景：

   - 在线故障预测

     对于在线且 BMC/BIOS 接口已支持故障采集的设备，实时对设备进行故障信息采集后，使用训练模型对硬盘进行失效预测，输出为预测结果。

     从经验和数据验证结果来看：

     - 预测时间窗越短，准确率越高，但收益越小，留给用户迁移数据和备件准备的时间越少。
     - 预测时间窗越长，准确率越低，客户页面上硬盘故障率越高，停机维护次数越多。

       为了兼顾准确率和维护可用性，FusionOne Center 采用提前 20 天对风险盘进行预测，预测准确率可以达到 80%以上（业界标准算法，FDR：提前识别的故障盘/总故障盘数目），误报率小于 0.2%（业界标准算法，FAR：误报的盘/总好盘个数）。

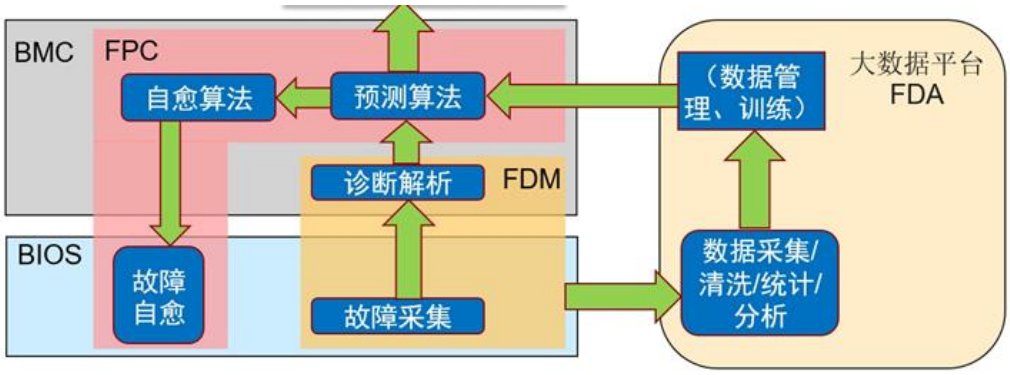
   - 离线故障诊断

对于离线的设备或 BMC/BIOS 接口未支持故障采集的设备，支持在 FusionOne Center 上面手动导入 BMC 日志包后，对设备硬盘进行故障诊断。

 版权所有 © 超聚变数字技术有限公司

6.2.4.2 内存故障预测

配套超聚变自研服务器，可支持服务器的内存故障预测。



1. 内存故障预测各组件介绍
   - FDM：Failure Diagnostic Management 故障诊断管理系统，实现故障采集和诊断解析，获取设备运行错误数据。
   - FDA：Failure Data Analysis Platform 故障大数据分析平台，开展故障大数据分析和 AI 训练，得到故障预测算法。目前已有 10+万套服务器的运行日志数据，涵盖了云服务、金融、大企业等主流应用客户数据。
   - FPC：Failure Prediction and Correction 故障预测自愈系统，在 BMC 中植入故障预测算法和自愈算法，实施隔离修复或上报预警。
2. 故障预测支持两种场景：
   - 在线故障预测

     对于在线且 BMC/BIOS 接口已支持故障采集的设备，支持在 FusionOne Center 实时对设备进行故障信息采集后，对设备内存进行在线故障预测。
   - 离线故障诊断

对于离线的设备或 BMC/BIOS 接口未支持故障采集的设备，支持在 FusionOne Center 上面手动导入 BMC 日志包后，对设备内存进行离线故障诊断。

6.2.5 性能监控

监控计算集群、计算节点、虚拟机、服务器、交换机、OS 等资源，展示相关资源的性能指标历史 KPI 趋势图和 TOP 统计，帮助管理员评估 Compute 系统中不同资源对象的运行状态和变化趋势，进而全方位了解资源运行状况。

- 计算集群、计算节点、虚拟机的性能指标数据，可查询最大 6 个月时间范围。
- 支持自定义指标，进行历史 KPI 查询和 TOP 统计。

 版权所有 © 超聚变数字技术有限公司

- 内置监控仪表盘，从多维度展示平台资源的性能指标实时统计和历史趋势展示。
- 可灵活启停要监控的指标，可动态设置指标数据采集周期。

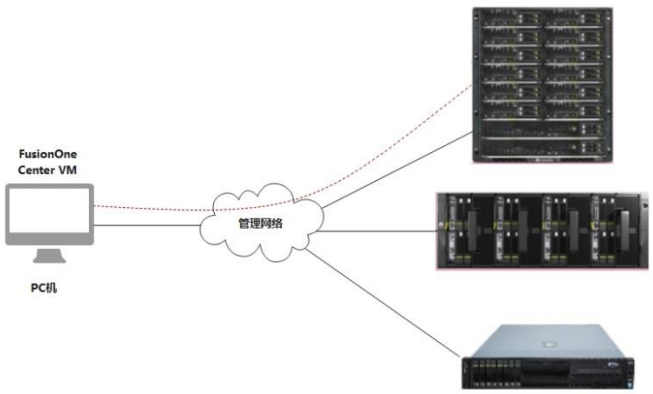
6.3 运维管理

FusionOne Compute 系统提供部署简单、运维便捷的体现，实现一站式交付、系统部署当天业务上线，一键运维，降低对 IT 管理人员的技能要求。其中安装部署方面，FusionOne Compute 提供自动化安装部署工具 FusionOne Installer；运维管理方面 FusionOne Compute 提供了统一的运维管理平台 FusionOne Center 管理工具。

6.3.1 一键部署

FusionOne Installer 快速安装部署工具完成系统软件的安装，自动完成管理集群的创建。

图6-4 部署系统示意图



- FusionOne Installer 安装工具可部署在 PC 机或者虚拟机中。
- FusionOne Installer 通过 Simple Service Discovery Protocol（SSDP）简单服务发现协议或者扫描 IP 方式发现服务器，读取服务器信息。
- FusionOne Installer 连接服务器 BMC，使用 KVM 挂载光盘功能进行引导，启动安装。最大支持 8 个节点并行安装。
- 安装过程中使用 NFS 共享传送安装软件包和配置。

FusionOne Installer 提供安装向导和统一的安装配置界面，可协助用户快速完成参数设置，FusionOne Installer 根据相关的安装配置快速自动完成系统软件的安装。

 版权所有 © 超聚变数字技术有限公司

6.3.2 系统下电

在 FusionOne Compute 管理界面上，能一键式对节点安全下电操作，满足用户计划性上下电的诉求。

安全下电集群：

- 向导式配置：提供向导式的配置页面，引导用户一步步执行一键安全下电的操作。
- 下电前系统检查：支持在实际下电前对各子系统产生的告警、子系统状态等进行检查，如有异常，提示用户在下电前处理。避免恢复供电后出现 Compute 不可用的情况。
- 系统备份： 引导用户在下电前，自动执行一次系统本地备份，在出现系统异常后，可使用备份数据快速还原管理面业务。
- 安全下电：对虚拟化平台自动执行安全下电操作，用户无需额外干预。

6.3.3 日志收集

FusionOne Compute 集成日志收集功能，一键式收集系统故障时各个组件的相关日志，支持一键式收集系统所有日志，也支持针对性收集部件日志，为客户日常运维，问题定位，提供有效帮助。

- 可支持收集日志项包括：VRM、CNA、OS 等系统组件的相关日志项。
- 日志收集一次收集的时间段暂只支持 2 天时间的日志文件，支持并发收集节点日志。

在 FusionOne Compute 日志收集页面，选择待收集的日志时间段，节点类型，日志类型以及需要收集的节点，即可进行收集日志。

日志收集完成后，可将相应的收集日志下载分析。

6.3.4 任务中心

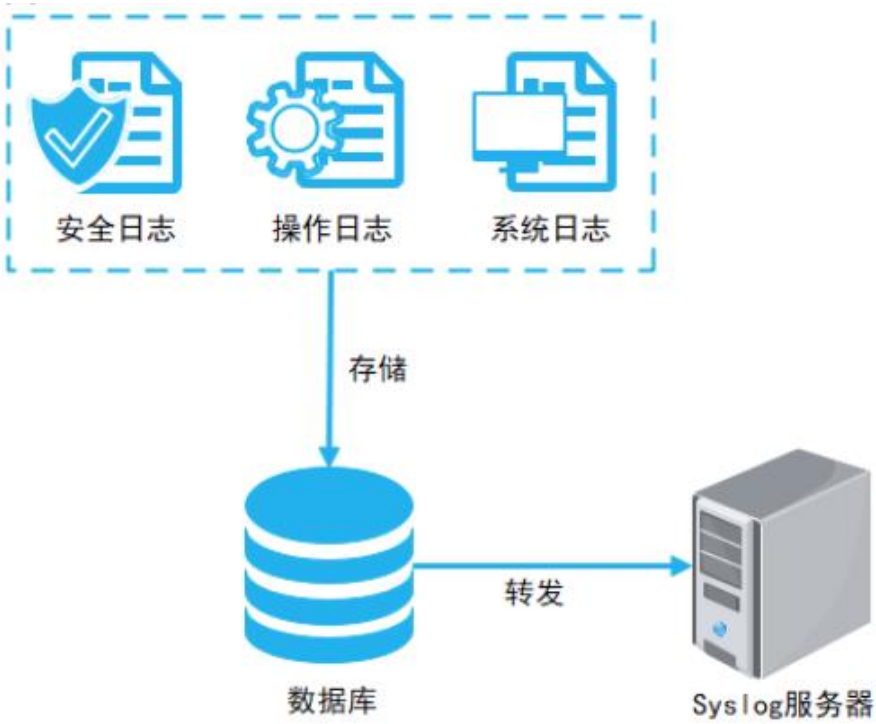
FusionOne Compute 支持对系统运行中启动的所有任务信息进行显示和管理，帮助客户查看任务当前进度和历史任务结果。主要展示的任务信息包括：任务名称、子任务数目、任务状态、任务进度、持续时长，用户可以对历史任务进行管理，删除已经完成的任务。

 版权所有 © 超聚变数字技术有限公司

6.3.5 日志管理

日志管理功能，提供高效的审计日志记录、审计日志查询、审计日志上报第三方、运行日志收集、下载等功能，帮助运维人员帮助用户发现安全风险、支撑问题定位和故障排除，降低运营、运维成本。

- 支持对 FusionOne Compute 的所有页面操作记录审计日志，以及对审计日志的查询、导出。
- 支持配置对接的 Syslog 服务器地址和收集的日志类型，将日志实时上传到第三方日志管理服务器上。



- 支持对虚拟化平台日志的一键收集，包括：FusionOne Compute、OS 等系统组件的相关日志项。 日志收集完成后，可将相应的收集日志下载分析。

6.4 多云管理

须知

本小节特性需要配套 FusionOne Center。

 版权所有 © 超聚变数字技术有限公司

在现代 IT 架构中，组织经常需要管理包括本地各种品牌的云平台环境，如 VMware VCenter、FusionCube、深信服 HCI 等多种异构平台。为了提升管理效率，FusionOne Compute 内置集成了异构云管理功能，实现对自身站点及 VMware 等不同异构云平台的高效管理。

- 统一管理界面：支持对 VMware VCenter、FusionCube、FusionOne HCI 22、SANGFOR HCI 等异构云平台的纳管，实现多云平台的异构虚拟化资源融合，实现集中管理。
- 异构站点管理：支持纳管最多 16 个异构云平台站点。
- 异构资源管理：支持对异构虚拟机的创建、开机、关机、重启、创建快照、删除快照、虚拟机详情、虚拟机任务列表等常用操作。
- 性能监控：支持查看异构虚拟机的性能监控数据，让用户可以实时监控本地站点和异构站点资源的性能指标。
- 跨平台迁移：支持将虚拟机从 VMware、FusionCube 温迁移到 FusionOne Compute 上。

异构平台版本兼容以下平台，具体版本信息查询兼容性网站：

- 支持的 VMware 版本
- 支持的 FusionCube 版本
- 支持的 FusionOne HCI 22 版本
- 支持的 SANGFOR HCI 版本(不支持温迁移)

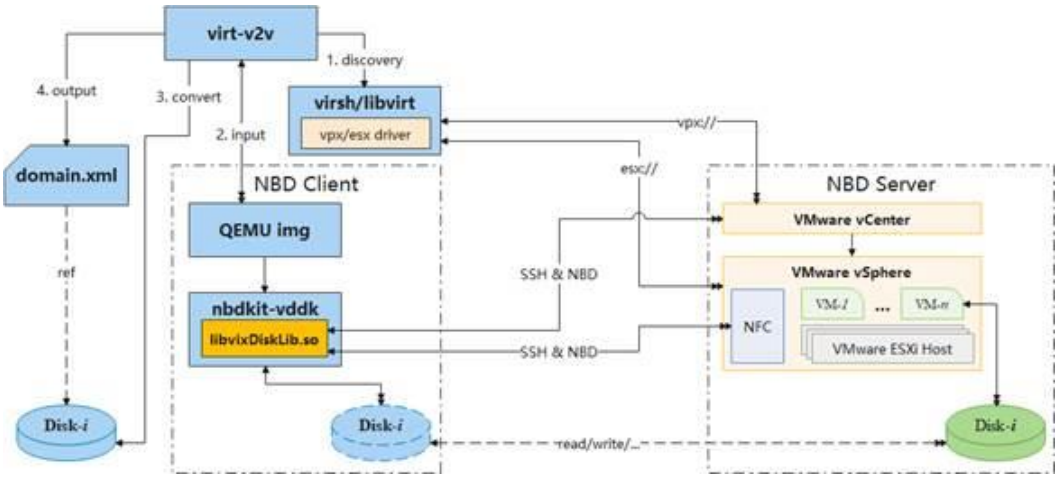
在纳管 FusionCube、VMware 平台基础上，可以支持对上述平台的虚拟机进行温迁移，将纳管的业务虚拟机，迁移到 FusionOne Compute 平台上运行。

- V2V 温迁移：支持 V2V 迁移，迁移过程中，无需经历导出/导入虚拟机的过程。迁移过程中不停机，迁移完成后自动关闭 VMware 上的虚拟机，实现虚拟机温迁移。
- 支持版本可查询兼容性网站
- 批量迁移：单次任务支持最多 10 个虚拟机的迁移。
- 迁移自动化：提供异构虚拟机迁移向导，引导用户一步步完成迁移配置，配置完后，实现虚拟机虚拟机的自动迁移，人工干预少，自动化效率高。
- 迁移 QoS，支持对虚拟机迁移网络自定义选择，并支持迁移速度设置。

如下是 VMware 迁移的技术原理介绍：

使用 Virt-v2v 实现虚拟机迁移：

 版权所有 © 超聚变数字技术有限公司

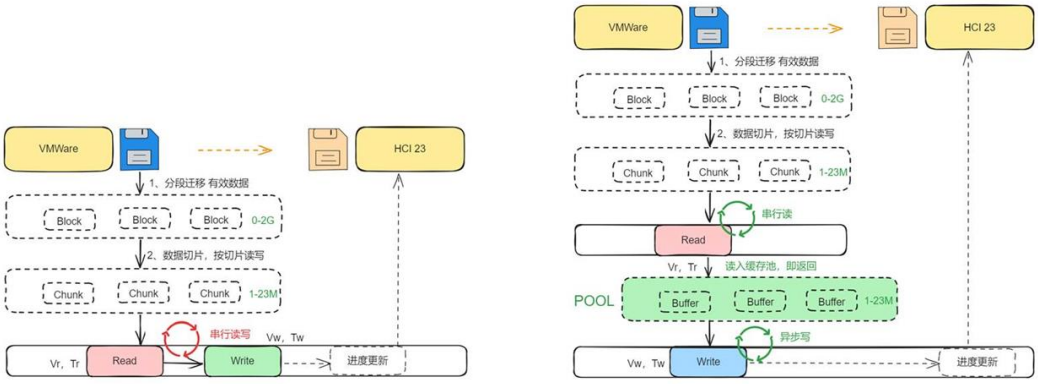


Virt-v2v 通过 virsh 连接 VMware 服务（vCenter 或 ESXi），获取源虚拟机信息（配置文件、磁盘、快照等）；通过 QEMU NBD 客户端，创建源网络磁盘；然后通过磁盘复制、转换过程，把源磁盘/快照复制到本地；把源虚拟机的配置文件转换为目标格式；通过 libguestfs 转换磁盘数据，同时替换磁盘路径、网络映射。磁盘数据转换主要包括：安装 VirtIO 驱动程序、更新/etc/fstab 和 xorg.conf 文件、重建 initrd 文件等。

优化点 1：数据迁移异步并发：

同步读+异步写操作，读写之间通过高效的管道传输，提高了数据迁移的并行性和效率，迁移速率最大提升 500%。单卷 FusionCube 120MB/s 以上，Vmware 单卷 GE 网络下可达 80MB/s 以上)。

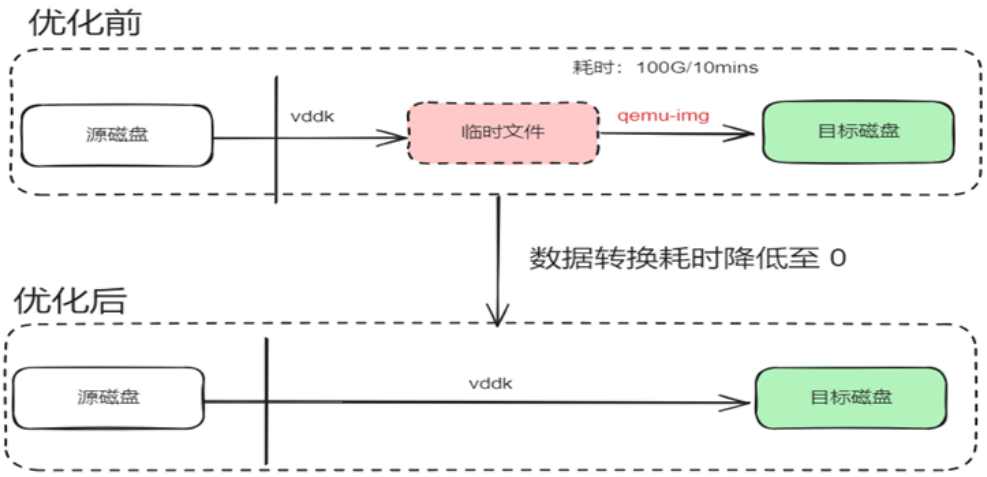
原理如下图所示：



优化点 2：virt-v2v，提升 VMware 磁盘转换效率

相比上个版本，省去临时文件转换，直接从源数据盘到目标磁盘，减少无效 IO 处理。

 版权所有 © 超聚变数字技术有限公司



 版权所有 © 超聚变数字技术有限公司

7 容器平台

FusionOneCompute 虚拟化平台支持部署虚拟机上 FusionOne Kubernetes Engine，或者在物理机上部署 FusionOne Kubernetes Engine。

容器架构如下：



 版权所有 © 超聚变数字技术有限公司

物理机容器部署架构如下:

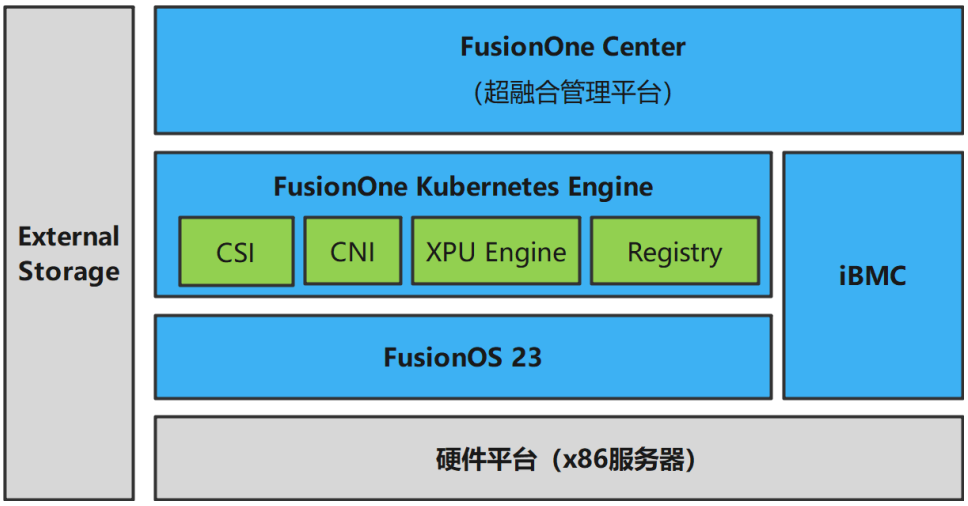


表7-1 架构组件说明

名称	说明
FusionOne Center	超融合管理软件，管理其中的容器化化资源、硬件资源，提供系统监控管理和运维管理等功能。
FusionOne Kubernetes Engine	基于 Kubernetes 的容器调度引擎，提供丰富、灵活、可靠的容器服务
硬件平台	1、使用超聚变自研服务器 X86 服务器以及国产化 KunLun 服务器。 2、支持计算、存储、交换、电源模块化设计，节点计算和存储按需搭配。 3、支持 NVME SSD 等 IO 加速扩展，支持丰富的交换模块 GE、10GE、25GE，可根据业务要求灵活配置。

容器平台技术进一步的介绍请参考容器平台技术白皮书《 FusionOne Kubernetes Engine 23.3.0. 技术白皮书 01》。

 版权所有 © 超聚变数字技术有限公司

8 系统可扩展性

随着业务量的不断增长，必定会面临资源需要扩容的场景，否则业务增长后平台超负荷运行，一定程度上影响到业务稳定运行。如:

- 硬件资源不足，整个集群 CPU、内存、存储 IO 资源使用率过高，影响业务性能。
- 高负载下，如果集群中某个主机故障，其他主机没有足够冗余资源承接该主机的业务，导致部分业务中断，无法快速 HA 恢复。

因此，基础设施不仅仅要求扩展硬件来解决集群业务压力，同时需要保证在扩容过程中保证简单、快捷、平滑的操作，业务不中断。

FusionOne Compute 提供了更加弹性、灵活的扩容选择，扩容过程中通过虚拟机热迁移等功能，可以保证节点硬件变更不影响业务的运行。

在线线性扩容

只需将新节点和原有集群置于同一网络，通过图形控制台将新节点添加到集群中即可。扩展后，新增节点的计算和存储资源会自动立即生效，整个扩展过程都是在线完成，不会影响正在运行的业务。虚拟化集群在一定规模内（40 节点以内），虚拟化集群的吞吐量随着节点数量的增加呈线性增加，自动调节虚拟化集群内配置，简单高效，自动将虚拟化集群配置为最优状态。

表8-1 扩容方式

扩容方式	适用场景	备注
添加 CPU、内存	虚拟化节点资源无法满足需求时，可更换或增加 CPU、内存、	• 建议选择当前集群业务较少或无业务时间段操作，

 版权所有 © 超聚变数字技术有限公司

扩容方式	适用场景	备注
	网卡资源。	以防影响业务性能。 • 节点停机前将业务虚拟机热迁移到其他主机。
添加节点	存储系统可用容量无法满足当前业务需求，且当前集群内节点无可用空闲硬盘或槽位。	• 建议选择当前集群业务较少或无业务时间段操作，以防影响业务性能。 • 可在虚拟化工具安装后，添加节点。

 版权所有 © 超聚变数字技术有限公司

9 系统可靠性

下面将从 2 个方面介绍 FusionOne Compute 的可靠性：

集群级可靠：主要从业务方面介绍虚拟化的可靠性，重点聚焦于管理架构、业务虚拟机可靠性。

硬件级可靠：介绍基于 FusionServer、KunLun 服务器的可靠性技术。

9.1 管理架构

9.2 业务虚拟机

9.3 硬件

9.4 故障恢复

9.1 管理架构

管理节点高可靠

VRM 主备管理节点采用管理平面的心跳检测，备用节点实时检测主用节点的健康状态，一旦发现主用管理节点故障，备用管理节点将立刻接管主用节点业务，持续对外提供服务。针对管理节点上的应用进程，通过采用软件狗的方式对运行在管理节点上的进程进行实时检测，如发现进程吊死或进入死循环，软件狗将会检测到相关进程的异常状态，并触发相关进程的重启恢复；如果发现进程重启后仍不能恢复正常，则进行业务管理节点的主备倒换并出主备心跳异常告警以保证应用进程的可靠性。

 版权所有 © 超聚变数字技术有限公司

黑匣子

内置黑匣子，黑匣子用于收集当前系统的信息，当系统出现故障的时候，黑匣子会保存系统的故障现场信息。借鉴于黑匣子，可以方便地进行故障定位。

黑匣子保存如下信息：

- 存储内核日志
- 系统快照
- 异常退出前屏幕输出信息
- 诊断工具的诊断信息

网络全路径冗余

虚拟化解决方案的网络按照层次划分，可以分为核心层、汇聚层、接入层和虚拟网络层。

核心层交换设备主要完成各数据中心之间的通信互联，同时提供虚拟化对外网络出口。可通过使用交换机集群，保证对外与防火墙/NAT 和对内各数据中心汇聚交换机连接的冗余。

汇聚层交换设备位于各个数据中心机房内部，完成本数据中心内各接入层交换机的流量汇聚，对外与核心层交换机通过三层互通，同时对接入层交换机提供二层接入功能。通过使用交换机集群，保证了对外与核心层交换设备和数据中心内接入层交换机连接的冗余。

接入交换机位负责本机柜内部的服务器接入。可通过使用交换机堆叠，保证对外与汇聚层交换设备和对内虚拟网络层连接的冗余。

虚拟网络层位于服务器内部，负责服务器内部的虚拟机之间以及对外通信功能。通过采用多网卡绑定，避免单个网卡故障引发的业务中断。

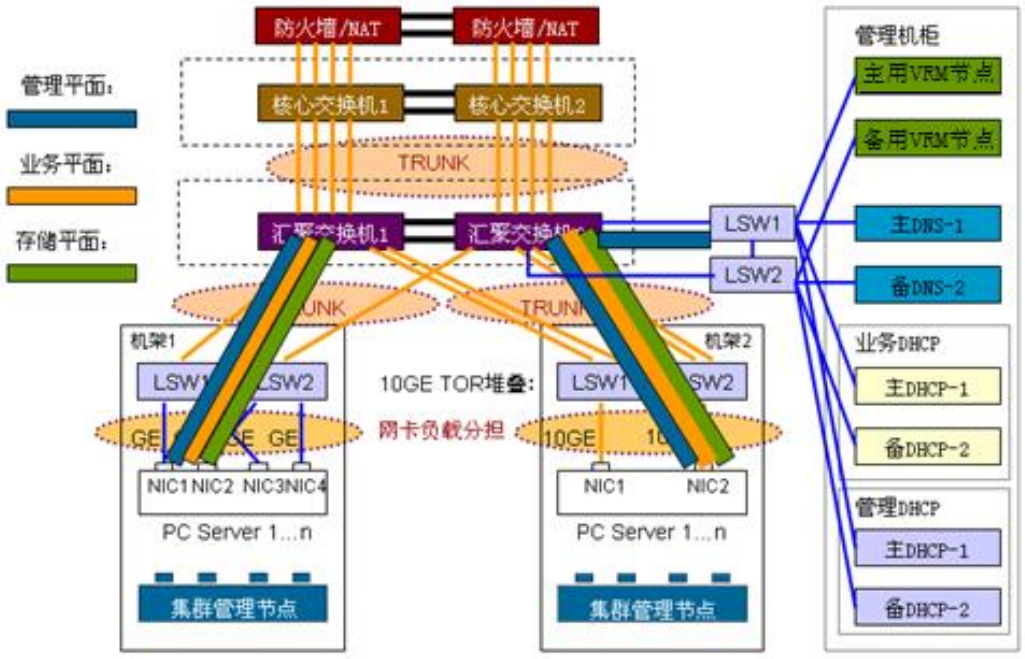
网络平面隔离

整个系统逻辑上可以分为三个平面：管理平面、存储前端、存储后端平面和业务平面。为了保证各种网络平面数据的可靠和安全，虚拟化采用分网络平面的架构方案，不同平面间采用 VLAN 进行隔离，单个平面的故障不影响其他平面继续工作。例如当管理平面暂时故障时，业务平面还能够用于继续访问虚拟机。此外，系统还支持基于 VLAN 的优先级设定，使得内部的管理/控制报文具备最高的权限，从而使得在任何时候，管理员和用户均可以管控系统。

 版权所有 © 超聚变数字技术有限公司

下图给出了从服务器 - 接入层交换设备 - 汇聚层交换设备间的网络连接图：

图9-1 网络分平面通信隔离示意图



在服务器内部，可通过对多个网卡的合理绑定和分类，允许将管理、业务和存储平面部署在不同物理网卡上，并将其连接到不同的接入层交换设备接口上，从而实现物理层面的网络隔离。

故障检测

系统提供了故障检测和告警的功能，同时它包括了在 Web 浏览器中显示故障信息的工具。一旦集群进入正常状态，系统提供使用数据可视化工具观察集群管理和分配负载的功能，可以帮助用户确定是否有负载均衡问题、失控进程或硬件性能下降的趋势，将对合理调整、分配系统资源，提高系统整体性能起到重要作用。

通过在每个被检测的节点包括定制化的虚拟机上运行监控程序，OM 系统可以收集被检测节点或者虚拟机的核心指标如 CPU 使用情况、基础网络流量和内存数据等，检测到诸如进程崩溃、管理和存储链路异常，节点宕机、系统资源过载等各种异常，使系统具备完善的故障检测能力。

支持解决方案一键巡检功能，为技术支持工程师和维护工程师提供的一套日常检查，并能输出各部件健康检查报告，方便技术支持工程师和维护工程师快速了解系统的健

 版权所有 © 超聚变数字技术有限公司

康状况。通过检查系统当前信息和运行状态，反映系统健康或亚健康状态，在开局、巡检、升级等维护场景中使用。

管理数据备份

系统提供配置数据和业务数据定期在本地和异地备份的能力。当管理节点服务异常无法自动修复时，通过本地备份数据立即恢复；当由于灾难性的故障导致管理节点双节点同时故障且不能通过重启等操作进行恢复时，可使用异地备份数据立即恢复，减少故障恢复时间。

快照

支持对虚拟机进行快照功能，当虚拟机损坏时，可以使用备份的快照恢复虚拟机。系统对虚拟机进行快照，把某一时刻虚拟机的状态（例如：硬盘信息）像照片一样保存下来。

- 当用户虚拟机损坏时，用户可以使用已经备份的虚拟机快照快速创建虚拟机。
- 用户可以根据需要，使用虚拟机快照将虚拟机恢复到创建快照的时刻。

9.2 业务虚拟机

FusionOne Compute 支持节点故障、虚拟机故障快速 HA 切换，保证业务在突发故障时可快速恢复业务。

虚拟机 HA(High Available)机制，可提升虚拟机的可用度，允许虚拟机出现故障后能够重新在资源池中基于优先级自动启动虚拟机。

主机故障 HA

当主机发生故障时，故障恢复的一般做法是通过集群检测主机的状态，如果检测到主机故障，超融合平台会主动触发虚拟机 HA，将业务虚拟机快速迁移到其他主机上。超聚变 FusionOne HCI 也采用相同的方案，在主机突发故障时，主动 HA 虚拟机到其他主机，包括含资源池外设如 GPU、网络 USB、SRIOV 直通网卡的虚拟机，均支持 HA 到其他节点

为了应对大规模集群故障风险，如单个机柜掉电，若干个主机同时故障，此时大量的虚拟机需要 HA 到其他主机上，虚拟机 HA 过程触发大量的计算消耗、IO 消耗，造成若干主机故障影响到整个集群其他主机上正在运行的虚拟机，FusionOne HCI 专门设

 版权所有 © 超聚变数字技术有限公司

计为此设计了群体性故障检测机制，防止多个主机同时故障，爆发大面积的虚拟机 HA，影响到其他主机上的业务可用性，提升业务的恢复效率。

群体性故障控制需要在“主机相关故障和响应策略”开启时生效，通过配置“故障控制时间”内(默认 2 小时)“允许处理虚拟机 HA 的主机数”(默认 2)，达到控制群体性故障时出现大规模的虚拟机 HA。被群体性策略控制的主机，其上虚拟机只能等待主机恢复后重新启动虚拟机。举例说明：2 小时内连续故障了 5 个主机，系统只会对前 2 个主机上的虚拟机按照“主机故障处理策略”和“虚拟机替代项”的配置进行 HA，其他 3 个主机上的虚拟机只能等待主机恢复后重新启动虚拟机，有效防止 HA 风暴。

虚拟机故障 HA

虚拟机高可用性是当计算节点上的虚拟机出现故障时，系统自动将故障的虚拟机在正常的计算节点上重新创建，使故障虚拟机快速恢复。

当虚拟机本身发生故障或者虚拟机所在物理服务器发生故障导致虚拟机故障时，系统能够根据用户预先设置的故障处理策略，决定在本地或异地重新启动虚拟机，以尽快恢复业务的运行。用户也可以设置为虚拟机发生故障后不作处理，在这种故障处理策略下，系统即使检测到虚拟机发生故障，也不会做处理。对于虚拟机 OS 内部故障，如 Windows 虚拟机的蓝屏故障（x86），或 Linux 虚拟机的 panic 状态，这类故障系统都能检测到并处理。

- 计算节点掉电恢复或重启

  当计算节点掉电恢复或重启时，系统将计算节点上具有 HA 属性的虚拟机重新创建至其他计算节点。

- 虚拟机蓝屏

  当系统检测到虚拟机蓝屏故障且该虚拟机蓝屏处理策略配置为 HA 时，系统选择其他正常的计算节点重新创建虚拟机。

资源预留

FusionOne Compute 提供丰富的资源预留策略，可根据业务可靠性需求配置，包括 HA 资源预留、专用故障切换主机、集群允许主机故障设置、群体性故障控制等。

- HA 资源预留：在整个集群内按照配置的值预留 CPU 与内存资源，该资源仅用于虚拟机 HA 功能使用。
- 使用专用故障切换主机：预留指定主机作为专用的故障切换主机，当主机作为故障切换主机时，普通虚拟机禁止在该主机上启动、迁入、唤醒（X86 架构）和快照恢复。仅当虚拟机 HA 时，虚拟机将优先选择在故障切换主机上启动，待故障

 版权所有 © 超聚变数字技术有限公司

切换主机资源不足时，才选择在普通主机上启动。当勾选“开启自动迁空”时，系统会定期将故障切换主机中的虚拟机迁移到其他有合适资源的普通主机上，以确保为故障切换主机上预留资源。

- 集群允许主机故障设置：设置集群内允许指定数目的主机发生故障，系统定期检查集群内是否留有足够的资源来对这些主机上的虚拟机进行故障切换。当资源不足时，系统会自动上报告警，对用户提出预警，以确保剩余主机的资源足够故障主机上的虚拟机 HA。
- 设置集群的群体性故障控制：通过配置“故障控制时间”内（默认 2 小时）“允许处理虚拟机 HA 的主机数”（默认 2 个），达到控制主机批量故障时出现大规模的虚拟机 HA 现象。被群体性策略控制的主机，其上虚拟机只能等待主机恢复后才能重新启动虚拟机。以默认配置举例说明：2 小时内共有 5 个主机顺序出现故障，系统只会对前 2 个主机上的虚拟机按照“主机故障处理策略”和“虚拟机替代项”的配置进行 HA，其余 3 个主机上的虚拟机只能等待主机恢复后重新启动虚拟机。

9.3 硬件

FusionOne Compute 选用高可靠的自研硬件，通过系统冗余设计保证系统可靠性，具有如下特点：

- 单板硬件采用电信级器件及加工工艺流程，可显著提高系统可靠性。
- 支持热插拔的 SAS/SATA/NVMe 硬盘，支持 RAID1 保护。使用 SSD 后的可靠性远远高于传统机械硬盘，从而能够延长系统运行时间。
- 整机提供 2 个热插拔电源模块，支持 1+1 冗余；提供 4 个热插拔风扇模块，支持 N+1。
- 提供 iBMC 直连管理接口，支持 iBMC 近端运维，提升运维效率。
- 支持 FPC Failure Prediction and Correction ）功能，对内存故障进行预测，并利用多种自愈技术，做出自愈隔离，以免影响业务正常运行。
- 网络双平面设计。
- 支持节点内磁盘漫游，磁盘节点换位插拔可自动识别，5 分钟内互换槽位插入，系统不发生数据重构。

 版权所有 © 超聚变数字技术有限公司

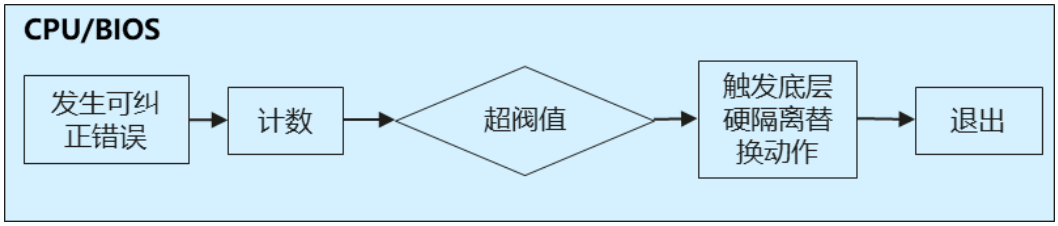
内存可靠性

内存错误主要包括硬错误和软件错误，其中硬件错误是由于失效或者损坏的硬件造成的，器件会不断返回不正确的数据，硬件错误可以通过 FusionServer 等自研服务器启动时 BIOS 的内存自检发现。

内存使用中经常碰到的为软件错误，软件错误不能通过内存自检发现，只有通过一些内存检错和纠错的算法来保护内存中的数据。FusionServer 等自研服务器在内存软件错误纠正上采用内存 ECC（Error Checking and Correction）技术，采用工业标准的纠错算法，能够检测内存 2bit 错误，并修复内存单 bit 错误。

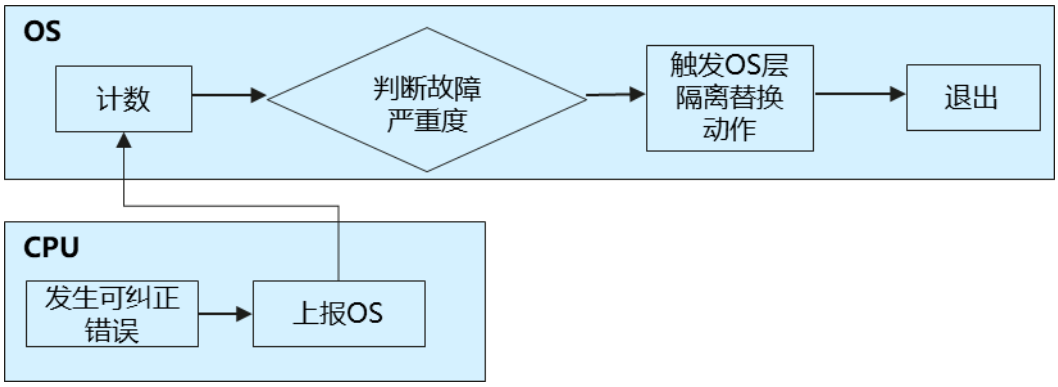
内存错误业界主流方案如下：

方案 1：由 CPU 内部的纠错机制识别出内存可纠正错误后，BIOS 对其计数并与初始设定阈值对比，达到阈值门限后，实施底层隔离，简称硬隔离。



方案缺点：简单的计数算法，无法准确判断内存故障的严重程度和精确位置。

方案 2：OS 对 CE 计数并判断故障严重度，达到设定门限后，由 OS 下发命令触发 OS 中的错误隔离机制，简称软隔离。



方案缺点：

- OS 中获得的 CE 信息少，预测准确性低。
- OS 只能做到上层动作，无法使用底层的硬隔离能力。

超聚变故障预测自愈系统：

 版权所有 © 超聚变数字技术有限公司

在 ECC 校验的基础上，超聚变自研了故障预测自愈系统（Failure Prediction and Correction，FPC），主要通过超聚变服务器日志大数据系统，机器学习算法得到内存故障特征模型，利用 BMC 芯片 AI 算法准确识别每个内存故障特征，对影响系统可靠性的故障特征进行预测，并利用多种自愈能力，做出针对性的自愈隔离措施或预告警，防止业务受损。

硬盘可靠性

- 硬盘热插拔：服务器支持系统运行时的硬盘（SAS/SATA）热插拔。
- 硬盘 RAID：服务器支持 RAID0、1、5 等多种 RAID 方式，支持 RAID 下另加热备盘的配置，保证了硬盘数据的高可靠性，在 RAID 组的某颗硬盘坏掉后，支持数据恢复、RAID 组恢复和在线更换硬盘。其中 RAID 卡支持电池，可以对 Cache 数据进行保护，既可以提高对硬盘的访问性能，又可以防止意外掉电时数据的丢失。

磁盘在线定时故障检测和预警

硬盘的故障一般分为两种：可预测的（predictable）和不可预测的（unpredictable）。后者偶而会发生，也没有办法去预防它，例如芯片突然失效，机械撞击等。但像电机轴承磨损、盘片磁介质性能下降等都属于可预测的情况，可以在在几天甚至几星期前就发现这种不正常的现象。

对于可预测的情况，如果能通过磁盘监控技术，通过测量硬盘的几个重要的安全参数和评估他们的情况，然后由监控软件得出两种结果：“硬盘安全”或“不久后会发生故障”。那么在发生故障前，至少有足够的时间让使用者把重要资料转移到其它储存设备上。

虚拟化采用 S.M.A.R.T.（Self-Monitoring Analysis and Reporting）技术对基于 ATA 和 iSCSI 接口的硬盘进行检测和可靠性管理，检查其可靠性并预测磁盘错误。其技术原理是主要时通过侦测硬盘各属性，如数据吞吐性能、马达起动时间、寻道错误率等属性值和标准值进行比较分析，推断硬盘的故障情况并给出提示信息，帮助用户避免数据损失。

支持 S.M.A.R.T.技术的硬盘可以通过硬盘上的监测指令和主机上的监测软件实现对磁头、盘片、马达以及电路的运行情况、历史记录及预设的安全值进行分析和比较，当出现安全值之外的情况时，自动向用户进行报警。

一般情况下 S.M.A.R.T.的几个主要的关键检测属性包含如下：

- Read Error Rate 错误读取率

 版权所有 © 超聚变数字技术有限公司

- Start/Stop Count 启动/停止次数(又称加电次数)
- Relocated Sector Count 重新分配扇区数
- Spin up Retry Count 旋转重试次数(即硬盘启动重试次数)
- Drive Calibration Retry Count 磁盘校准重试次数
- ULTRA DMA CRC Error Rate (ULTRA DMA 奇偶校验错误率)
- Multi-zone Error Rate 多区域错误率

电源可靠性

虚拟化节点配置 2 组电源（PSU)，提供电源故障告警，支持电源 1+1 冗余和热插拔，可以在一组电源故障后，系统持续运行而不影响业务；并且可以在线更换故障电源。

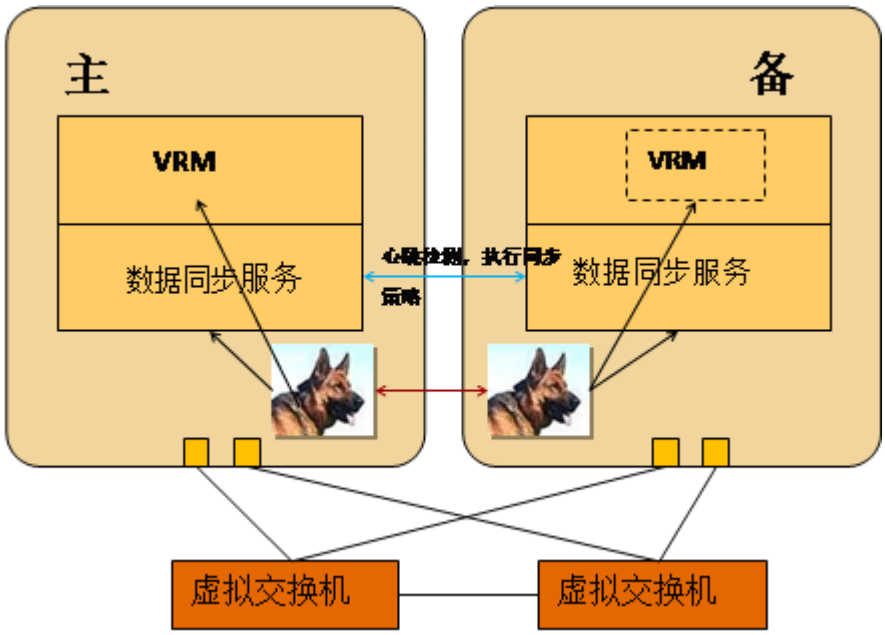
9.4 故障恢复

进程僵死自愈

由于系统原因会出现进程运行状态正常，但是不提供服务的情况，这种状态叫进程僵死，增加了关键进程僵死保护的机制，可以检查出进程处于僵死状态， 并自动将出于僵死状态的进程杀死重新启动，从而让进程正常提供服务。

 版权所有 © 超聚变数字技术有限公司

图9-2 管理节点 HA 示意图



管理节点负责对全系统的业务进行管理，采用主备高可靠性的工作方式，如果主备管理节点同时故障，相关的新增业务会受影响，例如虚拟机的创建和删除等，但对于已经存在并运行中的虚拟机无影响，用户继续使用虚拟机上的应用程序，不会有任何感知。

全场景故障可恢复

支持全场景故障可恢复、可更换，包括管理虚拟机、硬件、节点操作系统、超冗余故障的恢复预案，一旦遇到故障可遵循故障恢复步骤快速恢复，主要分为如下几个方面：

1. 节点硬件故障：包括 CPU、内存、网卡、磁盘、主板、整机更换。
2. 管理虚拟机故障：支持单个虚拟机基于冗余恢复、基于备份数据重建全部管理虚拟机。
3. 节点操作系统：可基于管理平台一键同步下发配置还原主机配置。
4. 超冗余故障：支持超冗余故障抢救预案。

通过上述的全场景覆盖，可支撑全场景的故障恢复。

 版权所有 © 超聚变数字技术有限公司

10 安全防护

10.1 安全威胁来源
10.2 平台安全
10.3 业务安全

10.1 安全威胁来源

来自外部网络的安全威胁

- 传统的网络 IP 攻击

  如端口扫描、IP 地址欺骗、Land 攻击、IP 选项攻击、IP 路由攻击、IP 分片报文攻击等。

- 操作系统与软件的漏洞

  在计算机软件（包括来自第三方的软件，商业的和免费的软件）中已经发现了不计其数能够削弱安全性的缺陷。黑客利用编程中的细微错误或者上下文依赖关系，已经能够控制操作系统。常见的操作系统与软件的漏洞有：缓冲区溢出、滥用特权操作、下载未经完整性检查的代码等。

- 病毒、木马、蠕虫等。

- SQL 注入攻击

  攻击者把 SQL 命令插入 Web 表单的输入域或者页面请求的查询字符串中，欺骗节点执行恶意的 SQL 命令，在某些表单中，用户输入的内容直接用来构造（或者影响）动态 SQL 命令，或作为存储过程的输入参数，这类表单特别容易受到 SQL 注入攻击。

 版权所有 © 超聚变数字技术有限公司

- 钓鱼攻击

  钓鱼攻击是一种企图从电子通讯中，通过伪装成信誉卓著的法人媒体以获取如用户名、密码和信用卡明细等个人敏感信息的犯罪诈骗过程。这些通信都声称来自于著名的社交网站，拍卖网站，网络银行，电子支付网站或网络管理者，以此来诱骗受害者的轻信。钓鱼攻击通常是通过 email 或者即时通讯进行。

- 零日攻击

  “零日漏洞”通常指还没有打补丁的安全漏洞，而“零日攻击”则是指利用这种漏洞进行的攻击。由于安全漏洞出现后，厂商需要时间确认、验证、评估、修补漏洞，很难当日拿出补丁。因此，零日漏洞的利用程序对网络安全具有巨大威胁。

来自内部网络的安全威胁

- 攻击方法日新月异，内部安全难以防范

  内网 ARP 欺骗与恶意插件滥用问题等将产生新的安全威胁。被攻破的内网主机，容易被攻击者做为“肉鸡”进行内网的渗透攻击，导致重要数据泄露，或者将其作为 DDOS 工具向外发送大量的攻击包，占用网络带宽。员工滥用恶意插件或浏览被植入病毒或木马的网页，也易受到攻击。

- 补丁升级与病毒库更新不及时、蠕虫病毒利用漏洞传播危害大

  由于网络内主机和设备的操作系统、数据库、应用软件存在安全漏洞，没有及时安装最新的安全补丁，主机杀毒软件病毒库没有及时更新，给恶意的入侵者提供了可乘之机，使病毒和蠕虫的泛滥成为可能。大规模的蠕虫爆发可能导致企业内网全部陷于瘫痪，业务无法正常进行。

- 非法外联难以控制、内部重要机密信息泄露频繁发生

  企业员工通过电话、VPN、GPRS 无线等拨号方式绕过防火墙的监控直接连接外网，使得企业内网 IT 资料暴露在外，易导致重要机密信息泄露。

- 移动设备随意接入、网络边界安全形同虚设

  员工或临时外来人员的笔记本电脑、掌上电脑等移动设备，由于经常接入各种网络环境，很可能携带有病毒或木马等恶意软件，一旦未经审查就接入企业内网，将对内网安全构成巨大的威胁。

- 软硬件设备滥用、资产安全无法保障

  内网资产（CPU、内存、硬盘等）被随意更换与修改，缺乏有效的技术跟踪手段和统一管理，一旦出现攻击行为或者安全事故，责任定位非常困难。

- 应用软件缺乏监控，产生新的安全隐患

 版权所有 © 超聚变数字技术有限公司

随着 QQ、MSN、微博等社交应用的普及，通过这些工具传播病毒、蠕虫、木马已成为新威胁的流行趋势；使用 BitTorrent、电驴等网络工具下载电影、游戏、和软件，可导致关键业务应用系统带宽无法保证。

- 缺乏外设管理手段，数据泄密、病毒传播无法控制

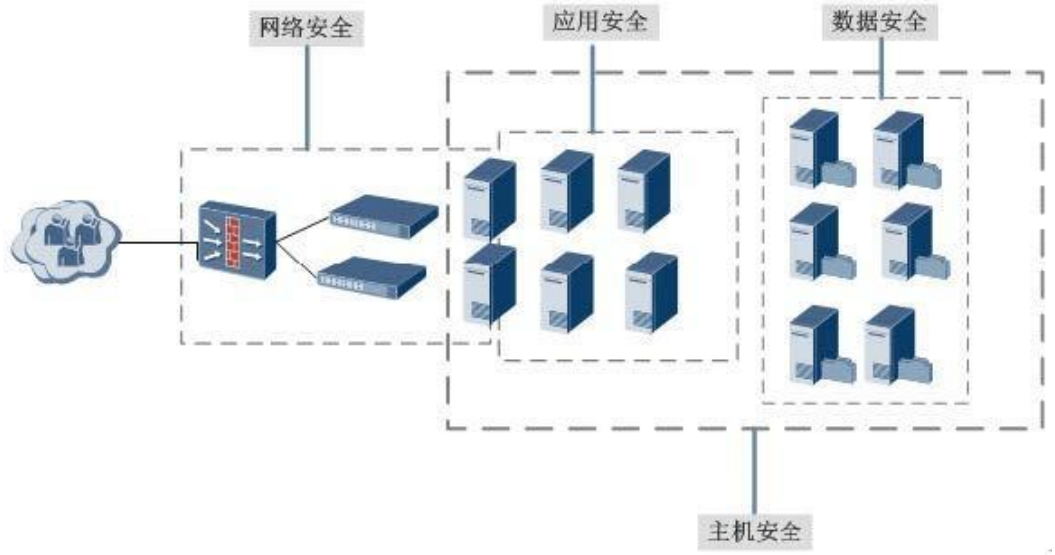
  U 盘、光驱、打印、红外、串口、并口等外设，由于使用方便，已成为数据泄密、病毒感染的出入口。通过封贴端口、制度要求等方式无法灵活对外设进行管理，特别是对 USB 接口的管理，因此，需通过其它技术手段解决存在的问题。

- 管理制度缺乏技术依据，安全策略无法有效落实。

10.2 平台安全

依据系统面临的安全威胁和风险，FusionOne Compute 产品提供安全解决方案，如图 10-1 所示。FusionOne Compute 安全框架通过网络、主机、应用以及数据四个维度上来保证系统的安全性。

图10-1 安全解决方案框架



简要介绍如下：

- 网络安全

  通过网络隔离，保证数据处理、存储安全和维护正常运行。

- 应用安全

 版权所有 © 超聚变数字技术有限公司

从身份认证、权限控制、审计控制等方面介绍 FusionOne Compute 目前已经具备的安全措施。

- 主机安全

  通过对系统内节点的操作系统安全加固等手段保证节点正常运行。

- 数据安全

  从数据备份、数据完整性、数据包密性等方面保证用户数据的安全。

10.2.1 网络安全

FusionOne Compute 的网络通信平面划分为业务平面、存储前端/后端平面和管理平面：

- 业务平面

  为用户提供业务通道，为虚拟机虚拟网卡的通信平面，对外提供业务应用。

- 存储平面（对接存储外部）

  为虚拟机提供访问存储资源的通信平面，但不直接与虚拟机通信，而通过虚拟化平台转化。

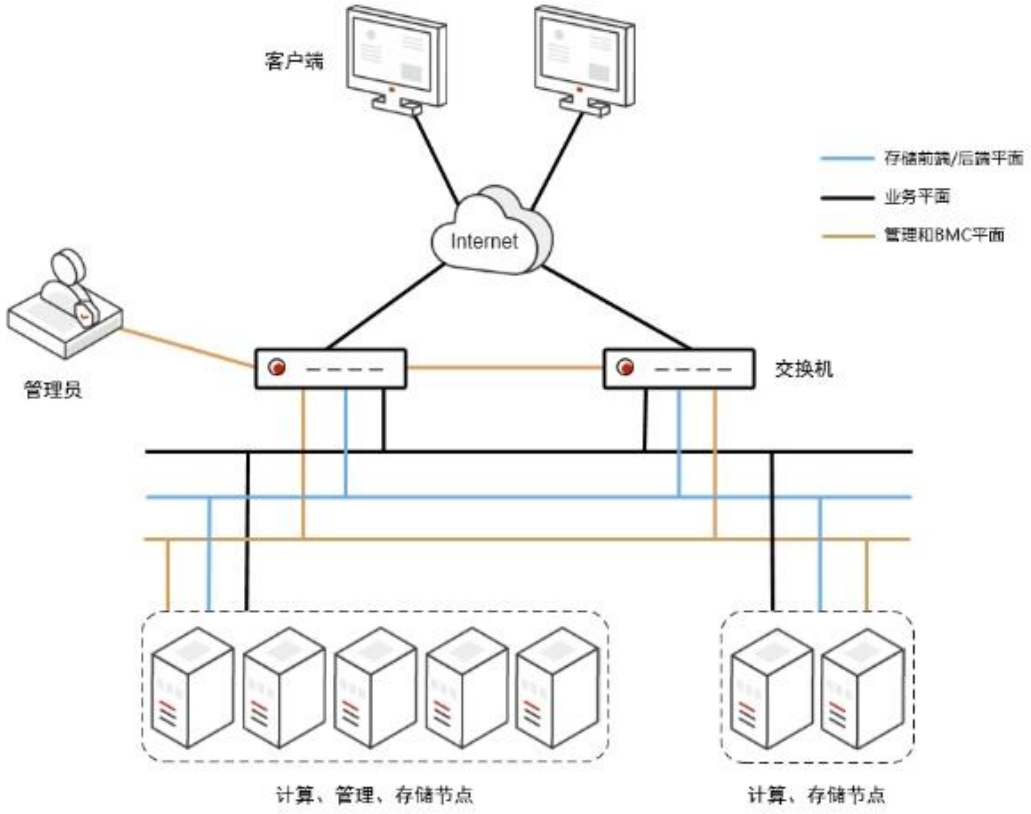
- 管理平面

  负责整个系统的管理、日常维护、业务配置、系统加载等功能的通信。

强烈建议三个平面之间相互隔离，平面隔离原理如图 10-2 所示。

 版权所有 © 超聚变数字技术有限公司

图10-2 平面隔离示意图



10.2.2 应用安全

10.2.2.1 权限管理

FusionOne Compute 支持“三员分立”特性，通过角色对帐户进行权限控制，系统安装成功后会预置角色，分别是系统管理员、安全管理员（账户激活、安全策略管理）、安全审计员（安全审计）、使得不同用户具有不同的权限，从而保证系统的安全。单个帐户只能拥有系统管理员、安全管理员和安全审计员三者中的一种身份，便于管理员间的权限隔离和相互监督。安装完成之后默认帐户如下：

- 系统管理员（sysadmin）：仅具有系统业务的操作维护权限，以及创建、删除用户的权限。创建的用户不属于任何角色，处于锁定状态。
- 安全管理员（secadmin）：仅具有用户、角色的权限管理权，但不能创建用户。系统管理员创建的用户需要安全管理员赋予角色并解除锁定。
- 安全审计员（AuditAdmin）：仅有日志查看和日志导出权限，用于对其他用户的操作进行审查。

 版权所有 © 超聚变数字技术有限公司

三员模式下可手动创建的用户分配的角色包括：

- 只读用户
- 自定义角色
- VMManager（虚拟机管理员）

10.2.2.2 Web 安全

FusionOne Compute 各 Web 服务具有的安全功能如下：

- 自动将客户请求转换成 HTTPS

  Web 服务平台能够自动把客户的请求转向到 HTTPS 连接。当用户使用 HTTP 访问 Web 服务平台时，Web 服务平台能自动将用户的访问方式转向为 HTTPS，以增强 Web 服务平台访问安全性。

- 防止跨站脚本攻击

  跨站点脚本攻击是指攻击者利用不安全的网站作为平台，对访问本网站的用户进行攻击。

- 防止 SQL 注入式攻击

  SQL 注入式攻击是指，攻击者把 SQL 命令插入到 Web 表单的输入域或页面请求的查询字符串，欺骗服务器执行恶意的 SQL 命令。

- 防止跨站请求伪造

  跨站请求伪造是指欺骗一个已登录的被攻击者装载一个包含恶意请求的页面，该请求利用浏览器自动发送鉴别凭证的功能，继承了被攻击者的身份和特权，执行一个对攻击者有益的恶意操作，如更改被攻击者的口令、地址等个人信息。

- 隐藏敏感信息

  隐藏敏感信息防止攻击者获取此类信息攻击系统。

- 限制上传和下载文件

  限制用户随意上传和下载文件，防止高安全文件泄漏，以及非安全文件被上传。

- 防止 URL 越权

  每类用户都会有特定的权限，越权指用户对系统执行超越自己权限的操作。

- 登录页面支持图片验证码

  在 Web 系统的登录页面，系统随机生成验证码；只有当用户名、密码和随机验证码全部验证通过时，用户才能登录。

 版权所有 © 超聚变数字技术有限公司

10.2.2.3 数据库加固

FusionOne Compute 管理节点的数据库类型为 GaussDB 数据库。

数据库必须进行基础的安全的配置，保证数据库运行安全，GaussDB 数据库的主要安全配置如下：

- 访问源控制

  基于访问的实际业务需求与安全标准，只对本地开放访问。所有跨机访问数据库的连接请求都被拒绝，避免受到系统外部的攻击。

- 最小授权原则

  数据库超级管理员之外的其他用户，均按照最小权限的需求设定角色。

- 目录保护

  数据安装目录与其数据区目录属主为安装用户，且其以及其子目录权限控制为读写执行。

- 敏感文件保护

  对于数据库的核心配置文件，属主为安装用户，权限控制为读写。

- 连接数限制

  系统默认的最大连接数是 300，用户可根据实际需要修改配置文件中的最大连接数来防止超大连接数的恶意尝试攻击。

为保证数据安全，必须对数据库进行定期的备份，防止重要数据丢失。数据库支持本地在线备份方式和异地备份方式：

- 本地备份：数据库定时执行备份脚本进行备份。
- 异地备份：数据异地备份到第三方备份服务器。

10.2.2.4 日志管理

日志查看时采取的安全措施如下：

- 任何人员不能在界面上修改或删除日志。
- 有查询权限的人才能导出日志。

 版权所有 © 超聚变数字技术有限公司

10.2.3 主机安全

FusionOne Compute 中计算节点、管理节点均使用 Linux 操作系统，为保证此类设备的安全，必须对 Linux 操作系统进行基础的安全配置，基础安全配置的主要内容如下：

- 关闭不必要的服务，如屏蔽 Telnet 服务和 FTP 服务。
- 加固 SSH 的服务。
- 控制文件和目录的访问权限。
- 限制系统访问权限。
- 管理用户密码。
- 记录操作日志。
- 检测系统异常。

10.2.4 数据安全

FusionOne Compute 通过多种存储安全技术保证存储的用户数据安全可靠。

- 用户敏感数据加密存储

  用户敏感数据（如鉴权信息）采用 AES-256 算法或者 SHA-256 算法加密后存储。

10.3 业务安全

FusionOne Compute 提供面向业务的安全防护能力，包括内置、生态合作伙伴的安全能力。

10.3.1 虚拟机安全删除

安全删除功能是为了保护虚拟化内用户的个人隐私和敏感信息落入他人手中。当选择普通删除文件或格式化硬盘时，这些数据实际上并没有被完全删除，而是被标记为可被覆盖。如果不对这些数据进行彻底擦除，那么有可能会被恶意人士使用数据恢复软件来恢复这些数据。这可能会导致个人信息、银行账户、社交媒体账户、电子邮件以及其他敏感信息泄露。擦除可能存在关键数据的磁盘是非常重要的，可以帮助保护个人隐私和敏感信息，防止未经授权的人访问自己的数据，确保信息安全。

 版权所有 © 超聚变数字技术有限公司

FusionOne Compute 兼顾安全与性能，提供支持两种虚拟机的删除方式，安全删除与普通删除，安全删除虚拟机可以针对安全隐私比较高的虚拟机的数据销毁，将对删除的磁盘内部的数据进行安全低格处理，确保关键信息资产不泄漏。

- 安全删除：通过覆盖性擦写对磁盘空间进行 BIT 位删除，彻底将数据从虚拟磁盘中删除，通过该方式，即使被删除的虚拟卷被恢复，也只能得到一些随机写的无用数据，而无法获取到原始的数据内容，这样可以避免数据被恶意恢复，安全性高，但删除速度较慢，且删除时会占用少量系统资源，一般用于存在敏感数据的虚拟机的删除操作。
- 普通删除：通过破坏磁盘文件系统对磁盘空间进行删除，主要对磁盘的格式元数据清理，删除速度快，存在通过残余信息恢复数据的可能性。

10.3.2 无代理防病毒&防火墙

传统的安全防护需要在每个业务操作系统部署防病毒、防火墙软件，随着云平台的发展，一些企业网络环境呈现出多样性和复杂性。有的企业是私有云环境，依附 VM 等虚拟品台建构自己企业内部的生产环境，直接将业务转移到虚拟化平台，还有一些企业是私有、公有云都有的状态。无论哪种，对于服务器主机的安全，都应该做到全面性。

传统安全防护弊端

1. **杀毒风暴：**当物理服务器上的众多虚拟机同时更新病毒特征库或按需全盘扫描时，可能导致物理服务器性能消耗激增，CPU、内存和磁盘 I/O 可能出现拥堵，造成虚拟机无法正常提供服务。
2. **维护复杂：**传统防病毒方案中，每台虚拟机都需要单独安装升级、独立查毒，部署运维成本较高，且需要对每台虚机单独执行防病毒策略、安全加固与补丁管理等工作。
3. **存在防护间歇：**虚拟机处于关闭或开启过程中等不同状态时，安全策略不统一，导致了防护间歇问题。如某台一直处于关闭状态的虚拟机在业务需要时会自动启动，但在这台虚拟机启动时，其包括防病毒也还在启动，存在防控的间歇期，虚拟机短时间内处于不安全状态，造成安全隐患。

常用解决方案

- 轻代理

轻代理就是在每一台主机上安装一个轻量化的代理 agent。许多安全厂商就是利用这个代理进行自有安装产品的推发，如安全防护、杀毒、入侵防御、虚拟防火墙、安全

 版权所有 © 超聚变数字技术有限公司

基线、完整性监控。轻代理的任务是有控制中心进行统一下发，例如对客户端进行扫描任务的下发，客户端程序完成病毒扫描任务后，将任务执行状态、发现病毒文件详情日志发送给控制中心。

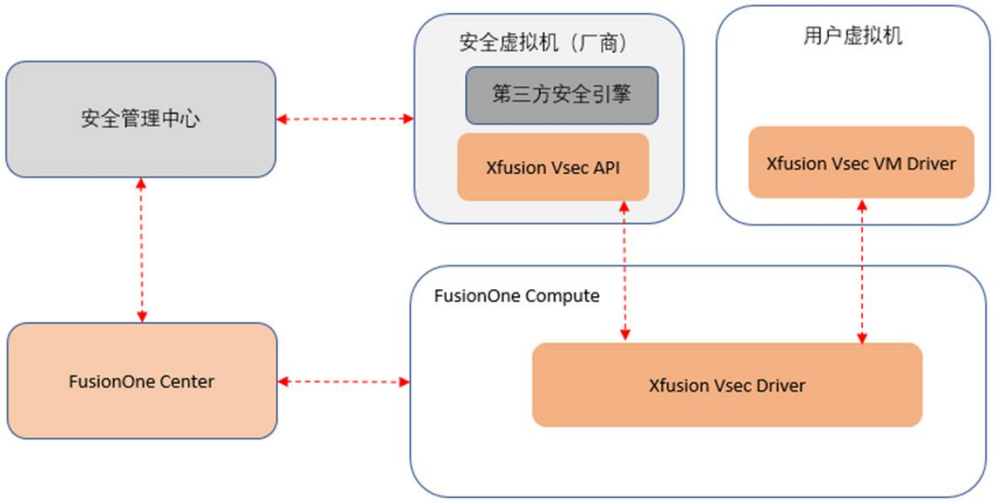
- 无代理

无代理是不需要再在每个虚拟机上安装 agent。此时防护功能由单独 1 台安全虚机（或虚拟化系统上的一个进程实现）。

FusionOne Compute 方案

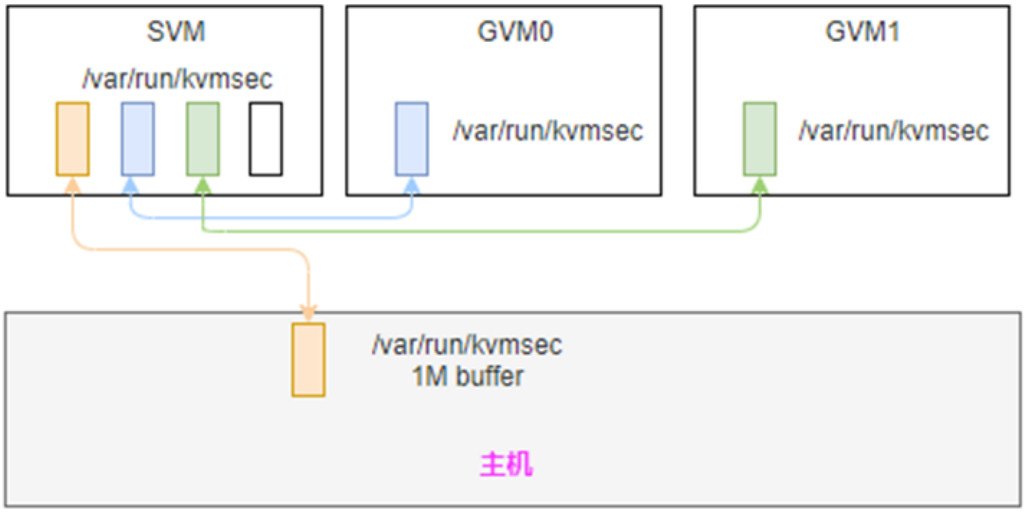
FusionOne Compute 选择无代理的方式实现，提供开放的 Rest 接口，安全厂商将安全软件部署在一个轻量化的 GVM 虚拟机即可对该主机上对业务虚拟机进行安全防控，让用户的虚拟机安全部署变得更加简单，且可大大节约平台资源损耗。

如下是超聚变 FusionOne Compute 无代理原理图，FusionOne Compute 提供安全 API 以及对应的驱动，安全厂商可调用接口对被主机内的虚拟机进行安全防控。

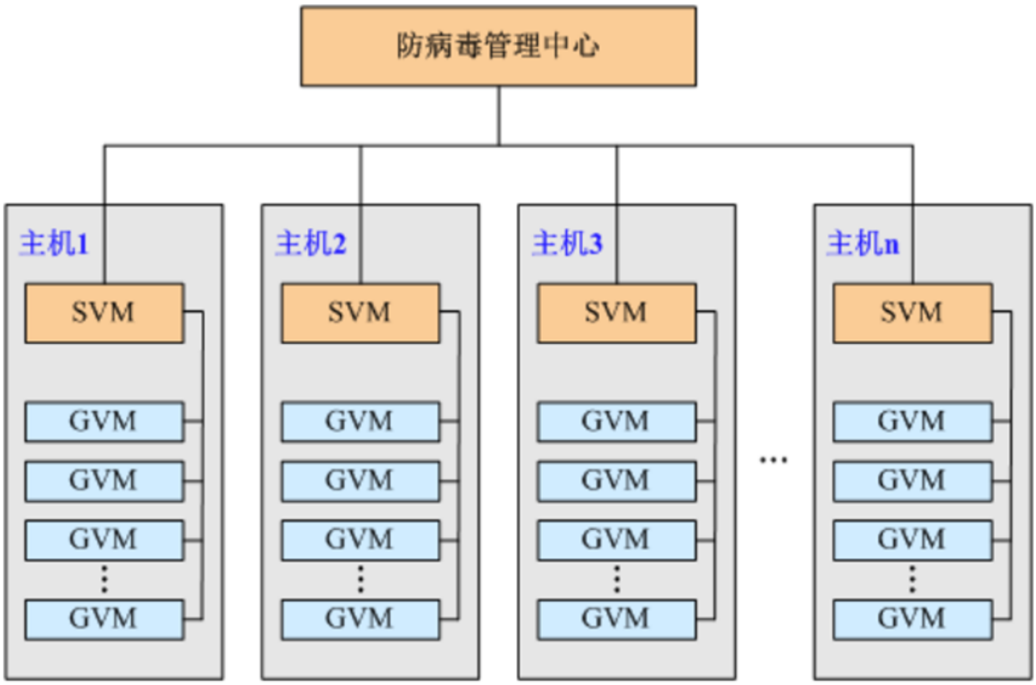


在 SVM 与 GVM、主机之间的通信主要通过 ivshmem 设备实现内存共享。

 版权所有 © 超聚变数字技术有限公司

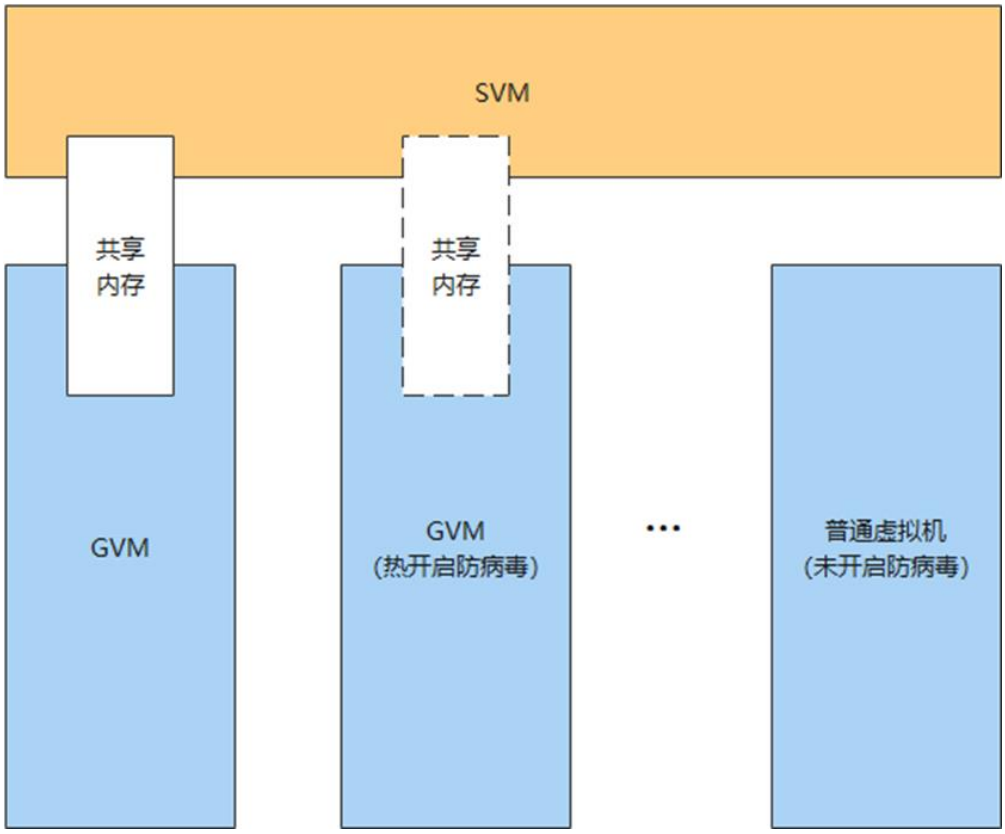


以下是以防病毒部署部署结构图：

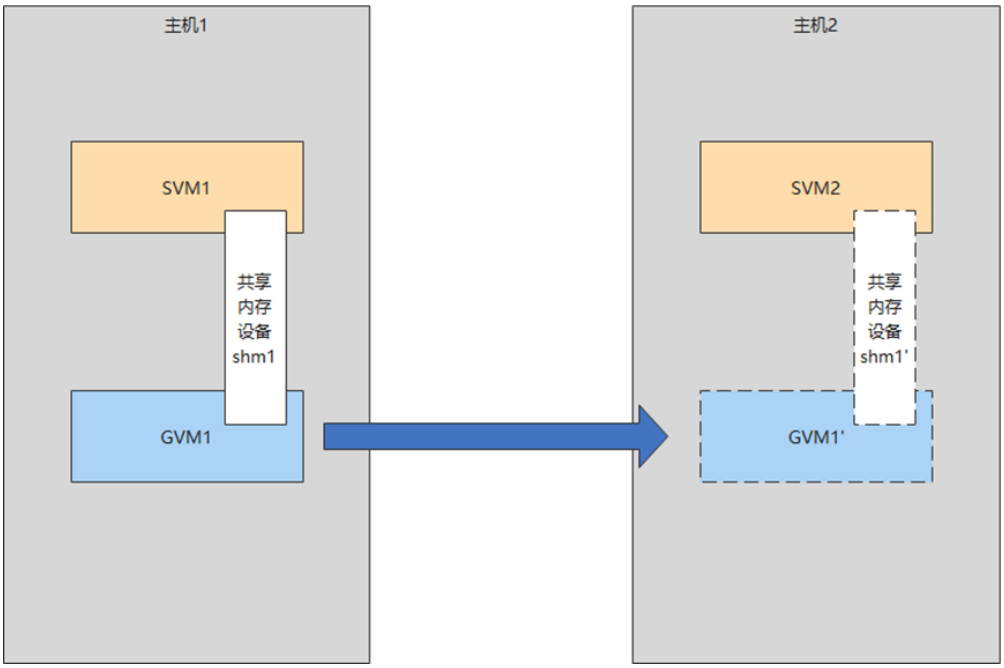


- SVM：安全服务虚拟机由防病毒厂商提供。
- GVM：安全用户虚拟机所安装的防病毒驱动由防病毒厂商提供。
- GVM：仅支持部分 Windows 的虚拟机作为安全用户虚拟机。
- 虚拟化防病毒功能所提供的具体防病毒服务及性能由防病毒厂商提供的防病毒引擎的能力决定。
- 安全虚拟机热开启，可以在不中断业务的情况下开启虚拟机防病毒功能；SVM 通过共享内存与 GVM 进行通信，对普通虚机不停机状态下开启需要热添加一个共享内存设备。

 版权所有 © 超聚变数字技术有限公司



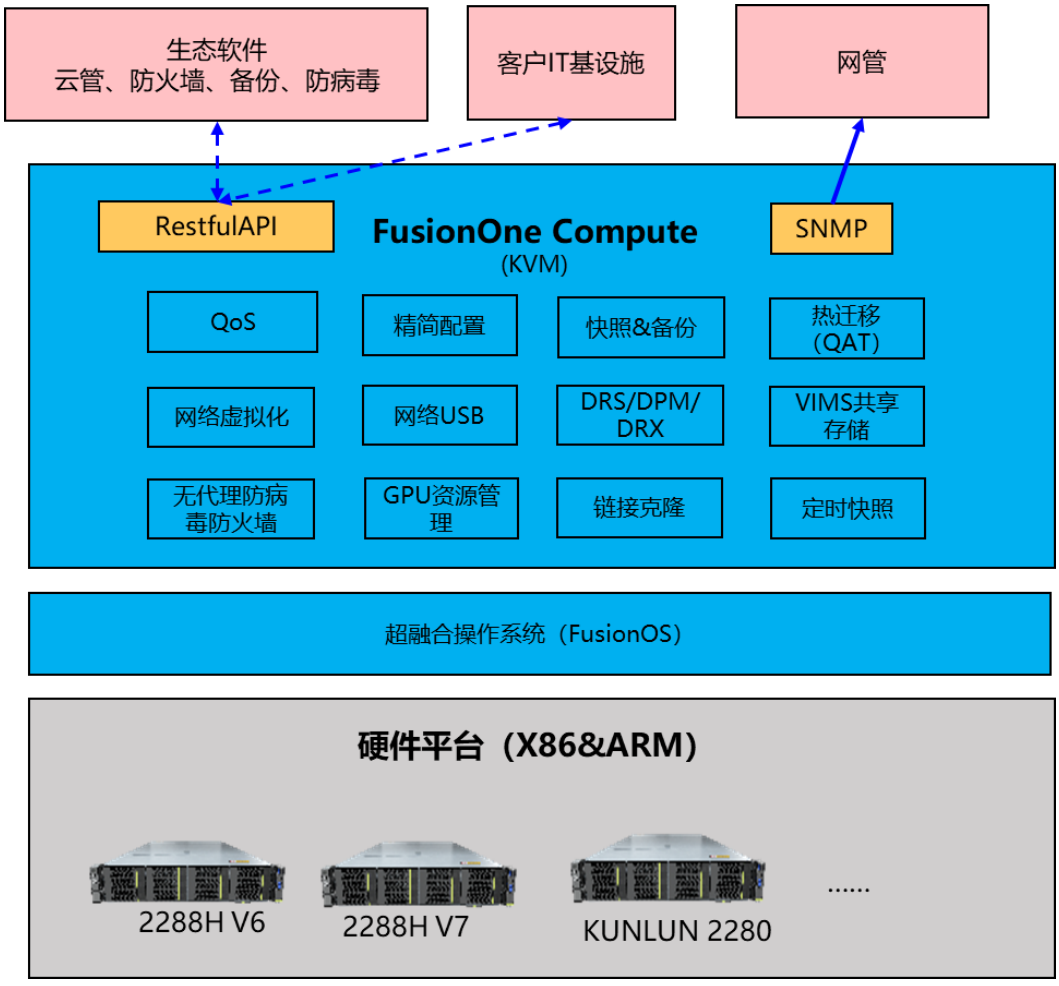
- **安全虚拟机热迁移，可以在满足安全的同时，支持虚拟机的在线迁移到其他主机，在平台升级、节点维护场景，可业务安全持续在线。迁移过程将共享内存设备以及其内部的数据一同迁移到目的端主机。**



 版权所有 © 超聚变数字技术有限公司

11 北向接口与开放兼容

FusionOne Compute 生态开放，为生态伙伴、最终客户提供了全面的接口支持，包括虚拟化平台的 RestAPI 接口、SNMP 告警对接接口。



主要在如下应用场景：

版权所有 © 超聚变数字技术有限公司

- 客户定制化业务发放

  可根据业务需求，将虚拟化平台融入到客户 IT 平台，为客户的运维提供更加便捷的运维。如通过客户内部管理平台申请虚拟机资源，自动化发放业务。

- 客户自动化运维

  可通过标准的 SNMP 接口，将虚拟化平台的告警信息同步到客户内部的网管软件，为客户统一运维提供便利的接口。

- 生态软件对接

  虚拟化平台相关的无代理防火墙、无代理防病毒、CBT 备份、云管等主流软件已经与 FusionOne Compute 对接兼容，可为客户提供一整套的 IT 解决方案。

 版权所有 © 超聚变数字技术有限公司

12 系统规格

规格项	规格小类	X86(Intel)	X86(AMD)	ARM(鲲鹏)
系统支持的节点规模	支持的节点数	1~256	1~256	1~256
VRM 管理虚拟机规格	最小配置（2~32 节点）	4U6GB	4U6GB	4U6GB
	推荐规格（32~100 节点）	8U8GB	8U8GB	8U8GB
	大规模（>100 节点）	12U12GB	12U12GB	12U12GB
单个 VRM 集群支持的最大虚拟机数量	-	5000	5000	5000
单个虚拟化逻辑集群支持的主机数量	-	64	64	32
VIMS 集群支持的最大主机数	-	32	32	32
虚拟化主机规格	每物理主机支持的最大逻辑 CPU	768	768	128
	每物理主机支持	16TB	16TB	4TB

 版权所有 © 超聚变数字技术有限公司

	的最大物理内存			
	每物理主机支持的最大虚拟机数量	1024	1024	1024
	每物理主机最大挂载 LUN 数量	1024	1024	256
	每物理主机最大 vCPU 数	4096	4096	384
	每物理主机最大虚拟网卡数	2048	2048	2048
	每物理主机支持的最大虚拟磁盘数	2048	2048	2048
	每物理主机的 NUMA 节点数	16	16	4
	每物理主机支持的在线迁移虚拟机并发数量	8	8	8
分布式交换机规格	系统支持的最大分布式交换机数量	50	50	50
	单个分布式交换机管理的最大主机数量	256	256	256
	单个分布式交换机支持的最大虚拟交换端口数量	10000	10000	10000
虚拟机规格	单个虚拟机支持的 vCPU 数量	255	255	128
	单个虚拟机支持	16	16	16

 版权所有 © 超聚变数字技术有限公司

	的网卡数量			
	单个虚拟机支持的磁盘数量	601	60	60
	单个虚拟机支持的内存容量	6TB	6TB	256GB
	虚拟机支持的本地磁盘单磁盘规格	16TB	16TB	16TB
	虚拟机支持的VIMS单磁盘规格	64TB	64TB	64TB
	单个虚拟机支持的快照数量	32	32	32
网络负载均衡器	负载均衡器数量	50（个）	50（个）	50（个）
	单负载均衡器实例数量	2（个）	2（个）	2（个）
	单负载均衡器实例虚拟机默认规格	4C50G	4C50G	4C50G
	监听器数量	100（个）	100（个）	100（个）
	单负载均衡器下可添加监听器数量	30（个）	30（个）	30（个）
	监听器转发策略数量	100（个）	100（个）	100（个）
	单监听器下可添加转发策略数量	5（条）	5（条）	5（条）
	单转发策略下可添加转发规则数量	10（条）	10（条）	10（条）

 版权所有 © 超聚变数字技术有限公司

	单转发规则可后端服务器组数量	1（条）	1（条）	1（条）
	后端服务器组数量	30（个）	30（个）	30（个）
	单后端服务器组下可添加后端服务器数量	50（个）	50（个）	50（个）
	健康检查器检查间隔	5-50(秒)，默认 10（秒）	5-50(秒)，默认 10（秒）	5-50(秒)，默认 10（秒）
	健康检查器重试次数	1-5（次），默认 3（次）	1-5（次），默认 3（次）	1-5（次），默认 3（次）
	健康检查器超时时间	1-20（秒），默认 2（秒）	1-20（秒），默认 2（秒）	1-20（秒），默认 2（秒）

 版权所有 © 超聚变数字技术有限公司