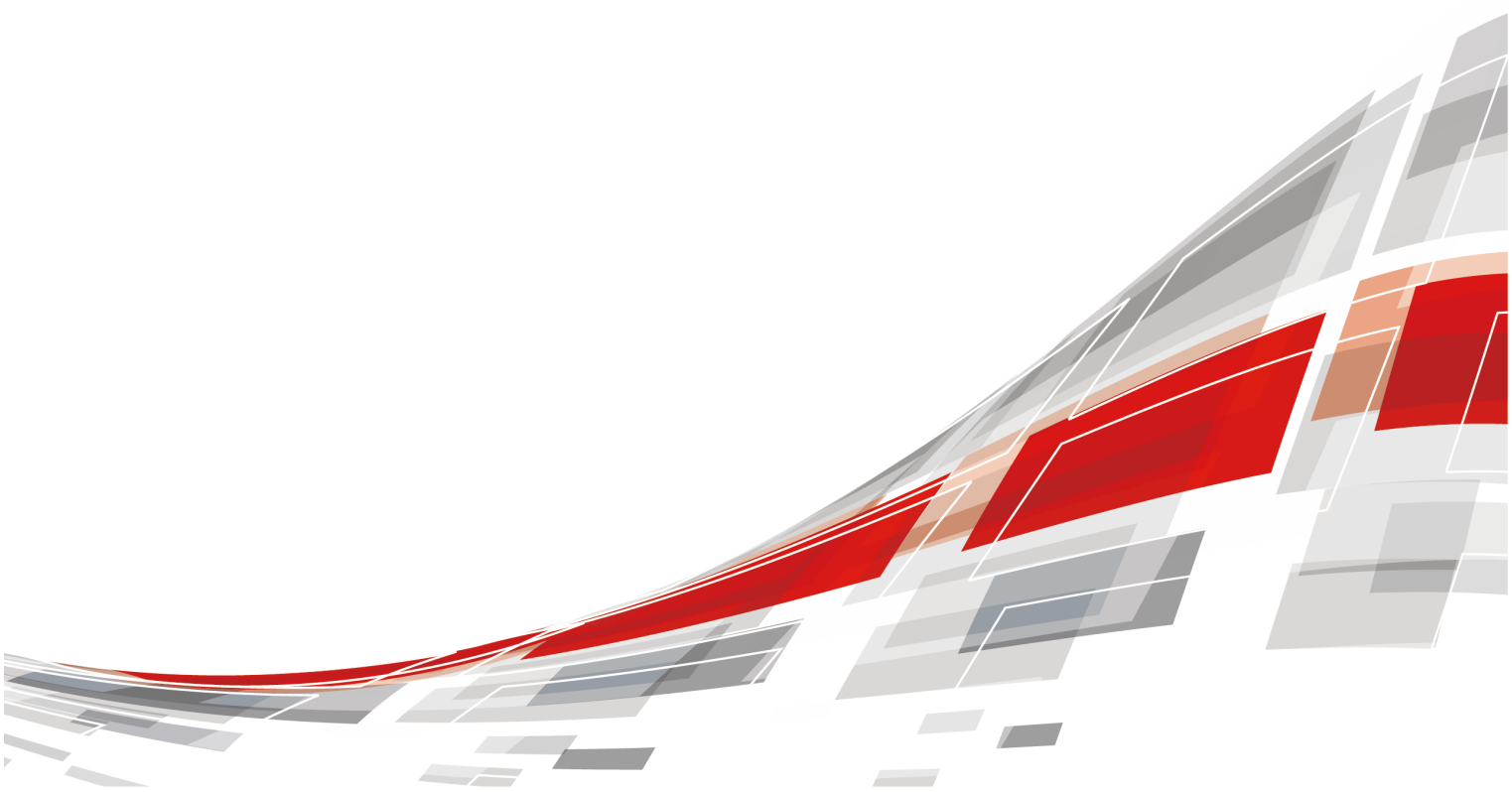
FusionOne HCI 23
23.6.1
技术白皮书

文档版本 01

发布日期 2026-07-30



超聚变数字技术股份有限公司

版权所有 © 超聚变数字技术股份有限公司 2026。 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股份有限公司的商标。
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，超聚变数字技术股份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股份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，超聚变数字技术股份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。

超聚变数字技术股份有限公司

地址： 河南自贸试验区郑州片区（郑东）龙湖内环北路99号 邮编：450000

网址： https://www.xfusion.com

前言

概述

本文档介绍了FusionOne HCI 23的产品价值、产品架构、主要特性、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。
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。

 版权所有 © 超聚变数字技术股份有限公司

修改记录

文档版本	发布日期	修改说明
01	2026-04-30	第一次正式发布。

 版权所有 © 超聚变数字技术股份有限公司

目 录

前言 ii
1 IT 基础架构变革 1
2 超聚变超融合介绍 4
2.1 技术架构 4
2.2 典型配置介绍 8
2.3 典型组网介绍 11
2.4 硬件平台介绍 13
3 计算虚拟化 15
3.1 概述 16
3.2 虚拟机资源 QoS 18
3.3 CPU 超线程 19
3.4 CPU 内存超分 20
3.5 虚拟机资源动态热添加(DRX) 23
3.6 虚拟机热迁移 24
3.7 虚拟机快速克隆 27
3.8 虚拟机回收站 27
3.9 第三方虚拟机模板导入 28
3.10 NUMA 调优 28
3.11 动态资源调度（DRS&DPM） 29
3.12 快照与快照策略 29
3.13 XPU 可视化管理与调度策略 30
3.14 GPU 虚拟化 32
3.15 NPU 虚拟化 33
3.16 沐曦 GPU 虚拟化 34
3.17 GPU 资源池 35
3.18 XPU 直通 44
3.19 网络 USB 45
3.20 SSH 密钥管理 46
3.21 PCIe 设备白名单管理 47
3.22 GuestOS 支持解耦 47
3.23 noVNC 文件上传 47
4 网络虚拟化 49

 版权所有 © 超聚变数字技术股份有限公司

4.1 概述.....50
4.2 所画即所得.....51
4.3 连通性检测.....51
4.4 分布式虚拟交换机.....52
4.5 网络安全组.....53
4.6 SDN 网络.....53
4.7 Overlay 网络.....54
4.8 VPC.....56
4.9 NAT 网关.....57
4.10 弹性 IP.....58
4.11 网络 ACL.....59
4.12 IP 地址预留.....60
4.13 内网域名解析 DNS.....61
4.14 网络 QoS.....61
4.15 SR-IOV 交换.....62
4.16 用户态 OVS 交换.....64
4.17 DHCP 服务.....65
4.18 虚拟路由器.....65
4.19 端口流量镜像.....66
4.20 在线修改虚拟机 IP 地址.....67
4.21 虚拟负载均衡器.....68
4.22 RDMA 技术.....70
5 存储虚拟化.....72
5.1 分布式存储.....72
5.2 虚拟机直通 SAN.....74
5.3 虚拟文件系统.....75
5.4 磁盘模式配置.....76
5.5 自动精简配置.....78
5.6 RBD 卷.....79
5.7 NoF 卷.....80
5.8 虚拟机镜像仓库.....81
5.9 统一存储.....82
5.10 块存储特性.....82
5.10.1 快照.....82
5.10.2 LUN Masking.....84
5.10.3 快速离线.....84
5.10.4 卷级 QoS.....84
5.10.5 精简配置.....85
5.10.6 卷克隆.....85
5.10.7 卷独占.....85
5.10.8 DHT 分布式哈希技术.....85
5.10.9 多硬盘池技术.....86

 版权所有 © 超聚变数字技术股份有限公司

5.11 NAS 文件存储特性 87
5.11.1 文件存储高可用技术 87
5.11.2 文件存储接口 87
5.11.3 文件配额 88
5.12 对象存储特性 89
5.12.1 小文件在线合并 89
5.12.2 对象多版本及 CDP 技术 90
5.12.3 对象生命周期管理 92
5.12.4 授权管理 93
5.12.5 对象压缩 94
5.12.6 对象重删 96
5.12.7 对象预览 98
5.12.8 对象检索 98
5.13 虚拟机备份 98
5.13.1 备份 99
5.13.1.1 定时备份和立即备份 99
5.13.1.2 分布式存储备份 99
5.13.1.3 虚拟化存储备份 99
5.13.1.4 新机恢复 100
5.13.1.5 数据重删 101
5.13.1.6 加密传输 101
5.13.2 CDP 备份 101
5.13.2.1 原理描述 102
5.13.2.2 使用场景 102
5.14 分布式存储内部机制 104
5.14.1 负载均衡 104
5.14.2 智能缓存 105
5.14.2.1 数据类型感知 105
5.14.2.2 多级缓存加速 106
5.14.2.3 分布式缓存 106
5.14.2.4 读缓存 107
5.14.2.5 写缓存 107
5.14.2.6 缓存资源回收 107
5.14.2.7 大块数据透传 107
5.15 存储高性能 107
5.15.1 端到端 NVMe over RDMA 108
5.15.2 分布式微控制器 108
5.15.3 并行流水线 109
5.15.4 虚拟内存管理 109
5.15.5 元数据高效索引 110
5.16 EC 块存储 111
6 智算资源管理扩展 113

 版权所有 © 超聚变数字技术股份有限公司

7 统一管理.....114
7.1 设备管理.....114
7.1.1 可视化超融合节点.....114
7.1.2 网络设备纳管.....115
7.1.3 操作系统纳管.....116
7.2 裸金属资源管理.....117
7.3 监控管理.....119
7.3.1 统一告警与事件.....119
7.3.2 可视化性能监控.....119
7.3.3 节点可视化运维.....119
7.3.4 硬件故障预测.....120
7.3.4.1 硬盘故障预测.....120
7.3.4.2 内存故障预测.....122
7.4 运维管理.....122
7.4.1 安装部署.....122
7.4.1.1 自动化一键部署.....122
7.4.1.2 自动发现原理.....123
7.4.1.3 出厂预安装，开箱即用.....124
7.4.2 初始配置.....124
7.4.3 日志管理.....125
7.4.4 一键式运维.....125
7.4.4.1 一键日志收集.....125
7.4.4.2 一键健康巡检.....125
7.4.4.3 一键系统升级.....126
7.4.4.4 一键系统扩容.....126
7.4.4.5 一键系统下电.....126
7.4.4.6 一键资源优化.....127
7.4.5 远程运维和移动运维.....127
7.4.6 自动化作业.....127
7.4.7 任务中心.....128
7.5 多云&多站点管理.....128
7.5.1 多云管理.....128
7.5.2 多站点管理.....130
8 云管.....132
8.1 概述.....132
8.2 产品架构.....132
8.2.1 功能架构.....132
8.2.2 技术架构.....133
8.3 关键特性.....134
8.3.1 虚拟机.....134
8.3.2 云网络.....134
8.3.2.1 VPC.....134

 版权所有 © 超聚变数字技术股份有限公司

8.3.2.2 弹性公网 IP....135
8.3.2.3 NAT 网关....135
8.3.2.4 网络 ACL....136
8.3.2.5 安全组....136
8.3.3 云运营....137
8.3.3.1 组织与项目管理....137
8.3.3.2 配额管理....138
8.3.3.3 流程订单....139
8.3.3.4 计量计费....139
8.3.4 云运维....140
8.3.4.1 审计日志....140
8.3.4.2 告警....140
8.3.4.3 历史 KPI 监控....140
9 系统可扩展性....142
10 系统可靠性....145
10.1 解决方案高可靠....146
10.1.1 基础备份....146
10.1.2 CDP 备份....148
10.1.3 CDP 持续保护....148
10.1.4 延展集群....150
10.2 管理服务高可靠....156
10.3 数据可靠性....159
10.3.1 多存储池....159
10.3.2 多副本....159
10.3.3 EC 纠删码....159
10.3.4 数据一致性....161
10.3.5 掉电保护....161
10.3.6 快速数据重建....162
10.3.7 数据恢复....162
10.4 业务高可靠....163
10.5 亚健康检测....164
10.6 硬件可靠性....167
10.7 故障可恢复....169
11 安全防护....171
11.1 安全威胁来源....171
11.2 平台安全....172
11.2.1 网络安全....173
11.2.1.1 网络隔离....173
11.2.1.2 平台访问控制....174
11.2.2 应用安全....174
11.2.2.1 权限管理....174

 版权所有 © 超聚变数字技术股份有限公司

11.2.2.2 Web 安全.....175
11.2.2.3 数据库加固.....175
11.2.2.4 日志管理.....176
11.2.3 主机安全.....176
11.2.4 数据安全.....176
11.3 业务安全.....176
11.3.1 虚拟机安全删除.....177
11.3.2 无代理防病毒&防火墙.....177
11.3.3 勒索病毒联动处置.....181
12 北向接口与开放兼容.....182
12.1 北向接口.....182
12.2 应用市场.....183
12.3 Oracle RAC 虚拟机部署.....184
12.4 信创超融合.....185
13 系统规格.....188

 版权所有 © 超聚变数字技术股份有限公司

1 IT 基础架构变革

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、业务数据呈现几何倍数增加，传统大型机、独立服务器、服务器+集中式存储的架构的数据中心已经无法很好满足业务发展需求。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。但在应用过程中，企业面临如下挑战：

- 虚拟平台部署和管理复杂，运维费用仍然维持增长趋势。
- 安装部署复杂，硬件来自多厂商，规划、部署、调优需要丰富的经验支撑。
- 多厂商设备，售后支持界面多，解决问题慢。

企业越来越关注成本控制、业务敏捷、风险管控，希望能拥有总成本低、新业务的上线时间快、资源可弹性伸缩、安全可靠、高性能的IT系统。新IT设施以虚拟化为基础，实现集中管理业务，对数据中心资源进行动态调整和分配；满足企业关键应用向X86系统迁移对于高性能、高可用和安全性的要求，同时提高基础架构的自动化管理水平，确保满足基础架构快速适应业务商业诉求。

超融合技术（Hyperconvergence）是一种集成计算、存储、网络和虚拟化等关键IT功能的架构。这种架构通过软件定义的方式，将这些组件融合到一起，形成了一个统一的资源池。超融合技术的一个核心特点是将这些组件整合到单个设备中，如x86服务器，并通过网络连接，实现模块化的无缝横向扩展（scale-out）。这种设计简化了数据中心的架构，降低了运营成本，并提高了系统的可用性和可扩展性。

超融合解决方案通常包括以下几个主要组件：

1. 计算虚拟化：超融合平台提供计算能力，通常由多个物理服务器节点组成。每个节点都包含处理器、内存和其他计算资源，可以承载虚拟机和应用程序。基于虚拟化技术，如VMware vSphere、Microsoft Hyper-V或KVM等，允许在硬件平台上创建和管理多个虚拟机。虚拟化提供了资源隔离、灵活性和资源优化的好处。
2. 存储虚拟化：超融合平台将存储资源集成到服务器节点中，形成一个分布式存储系统。这些存储资源可以通过软件定义的方式进行管理，提供高可用性、数据冗余和数据保护等功能。
3. 网络虚拟化：超融合平台提供网络连接，使节点之间可以进行通信和数据传输。网络资源通常通过内置的网络交换机或虚拟网络设备实现。
4. 统一管理：超融合提供统一的资源管理平台，可对超融合平台的日常运维、客户业务发放的统一管理。

超融合的主要优势包括：

 版权所有 © 超聚变数字技术股份有限公司

1. 简化部署和管理：超融合解决方案集成了多个关键组件，简化了数据中心的部署和管理。管理员可以通过一个统一的界面来管理整个基础架构，而不需要处理独立的服务器、存储和网络设备。
2. 可扩展性：超融合平台可以根据需求快速扩展，通过添加新的节点来增加计算和存储资源。这种灵活性使得企业能够更好地适应业务增长和变化。
3. 高可用性和数据冗余：超融合解决方案通常具有高可用性和数据冗余机制，保证了业务的连续性和数据的安全性。数据在多个节点之间进行复制和备份，以防止单点故障和数据丢失。
4. 资源利用率高：节省机柜空间和能源消耗，同时由于计算和存储融合，充分利用算力、存储。

以下是传统IT设备与超融合的对比：

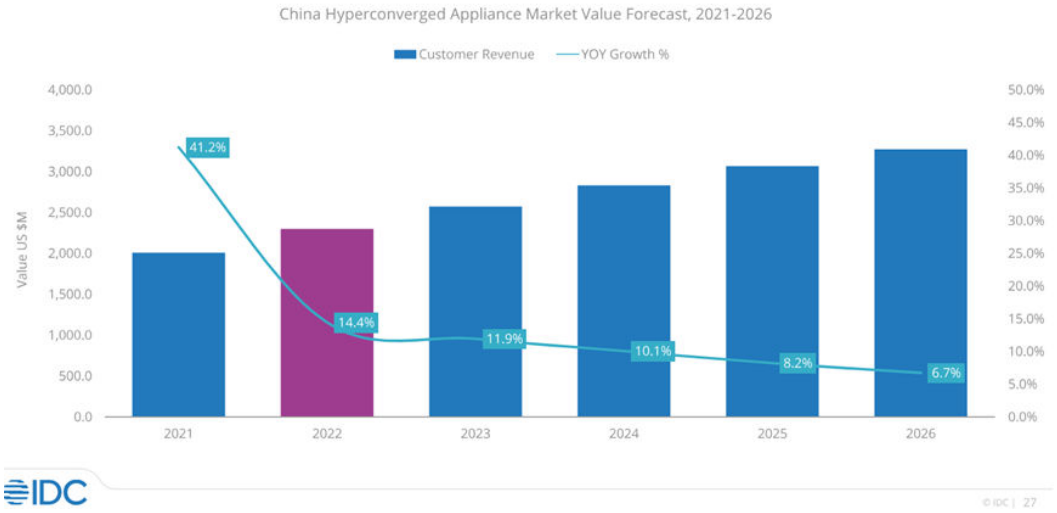
图 1-1 超融合与传统 IT 对比



超融合的主要应用场景：

在超融合架构模式下，用户所使用的虚拟机和存储空间是利用软件构建的，这样就使得底层物理设备与用户之间保持隔离的状态，实现了硬件资源与虚拟化平台的完整融合。越来越多的用户选择超融合的方式部署业务，以下是根据过去十年统计的IT设备份额组成，预计在相当长的时间，超融合依然会占据很大的比例。根据IDC报告统计数据，中国区的超融合应用一直呈现不断上涨的趋势：

图 1-2 IDC 报告超融合的发展趋势



 版权所有 © 超聚变数字技术股份有限公司

关键业务应用

关键应用是指那些对业务连续性和性能都要求非常高的应用，比如数据库、ERP和CRM等。越来越多的核心应用已经迁移到了虚拟化平台运行，而虚拟化与超融合是一种天然的组合，紧密集成，相得益彰。将核心应用部署在超融合上，可以提高系统的性能、可扩展性、可用性、可管理性。将应用从物理环境迁移到超融合，也可以降低采购和维护成本。

私有云

云计算是企业数字化转型的基石，越来越的应用开始向云迁移，新兴应用基于云环境开发设计，因此也驱动了私有云的快速普及。私有云以虚拟化和资源池为基础，要求底层基础设施简单、高效、弹性、稳健。基于超融合构建私有云可以为数据中心带来最优的效率、灵活性、规模、成本和可靠性。

开发测试

随着敏捷开发、微服务等“互联网+”思维方式和技术的引入，企业的开发测试工作，势必对测试平台在成本(降低)、效率(提升)、可用性(增加)和效果(显性化)等方面提出更高的要求。基于超融合构建开发测试云，可以实现快速的环境获取、能够模拟生产网络环境、运维自动化、更低的成本以及从容面对新技术的挑战。

VDI

超融合诞生之初的主要场景就是VDI（即，虚拟桌面基础架构），可以说，超融合与桌面云是一个完美的组合。通过超融合构建桌面云，不仅可以提供桌面云所需的性能、可用性、可扩展性，同时还可以降低架构的复杂性、管理难度，以及建设成本。利用超融合“随增长付费”的特性，用户可以按自己的节奏增长并控制成本。

分支机构

许多企业都拥有远程办公室和分支机构，它们依赖位于其他位置的IT员工管理的本地IT基础架构。面临管理效率低下以及可靠性和可用性差的挑战。HCI非常适合分支机构部署，它提供了一套低成本基础架构解决方案，其中集成了计算、存储和网络连接功能。另外，大幅简化IT运维管理，实现多分支机构统一管理，还可以轻松地横向或纵向扩展，非常灵活，足以适应不断变化的需求。

大数据

大数据具有4V特性（Volume、Velocity、Variety和Value），即来源和类型多样、海量，快速和时效性、价值发现。随着大数据在各行业中的广泛应用，用户对于大数据的时效性提出了更高的要求。基于HCI来构建大数据平台，在满足大数据系统对计算性能和存储容量需求的同时，也提供了更好的扩展性和伸缩性。

版权所有 © 超聚变数字技术股份有限公司

2 超聚变超融合介绍

FusionOne HCI是超聚变打造的一个开放的、可扩展的超融合系统，具有计算/存储/网络融合、高性能、高可靠、高安全、业务自动化快捷部署、统一管理、资源智能弹性伸缩、运维简单的特点，可帮助客户业务快速上线，快速实现不同云应用的部署，同时降低维护管理的难度，提升企业在业务上云的管理效率，最大化提升数字化转型的价值。

2.1 技术架构

2.2 典型配置介绍

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 HCI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（含分布式存储）、统一管理软件、云原生平台等组件组成，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。

- 超融合中的FusionOne Center提供超融合平台的管理平台，与分布式存储组件、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。
- FusionOne Compute提供虚拟化平台，提供计算资源、网络资源虚拟化，并通过分布式存储管理模块，对接FusionOne Storage。VIMS组件为支持第三方存储接入平台提供了标准接口，支持IP-SAN、FC-SAN存储接口。作为超融合的基础架构组件，为其他组件运行提供了基础平台。
- FusionOne Storage提供分布式存储服务，通过部署在基础架构的系统上，将集群内的磁盘，通过网络连接组成一个逻辑上的“大硬盘”，分布式的架构为集群提供海量的存储能力，并通过高可靠的技术，支持动态扩容均衡、故障快速重建。
- FusionOne Kubernetes Engine作为helm插件，可在HCI的基础上提供云原生的容器服务，可以部署敏态的业务，支持异构算力GPU、NPU以及其他异构卡资源池化管理，为新型的AI应用提供基础服务。

 版权所有 © 超聚变数字技术股份有限公司

图 2-1 FusionOne HCI 总体架构

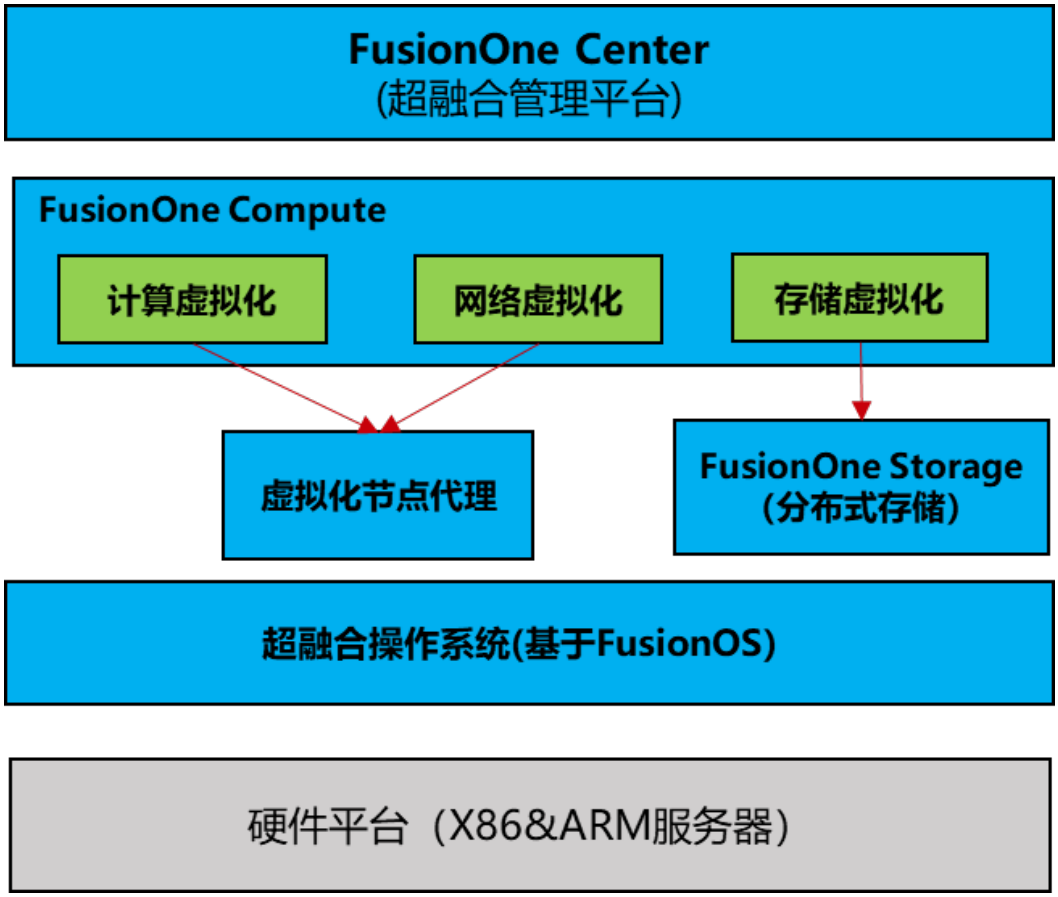


表 2-1 FusionOne HCI 总体架构组件说明

名称	说明
FusionOne Center	超融合管理软件，管理其中的虚拟化资源、硬件资源，提供系统监控管理和运维管理等功能。
分布式块存储(FusionOne Storage)	使用分布式存储技术，通过合理有序组织服务器的本地硬盘，提供高性能高可靠的统一存储业务（块、文件、对象）。
虚拟化平台(FusionOne Compute)	在自研FusionOS基础上，深度对KVM调优、自研虚拟化管理软件打造的虚拟化底座，为超融合系统提供虚拟化能力，包括计算虚拟化、网络虚拟化、存储虚拟化。
硬件平台	1. 使用超聚变自研服务器。 2. 支持计算、存储、交换、电源模块化设计，计算和存储节点按需混配，计算、存储都在服务器内部署完成。 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。

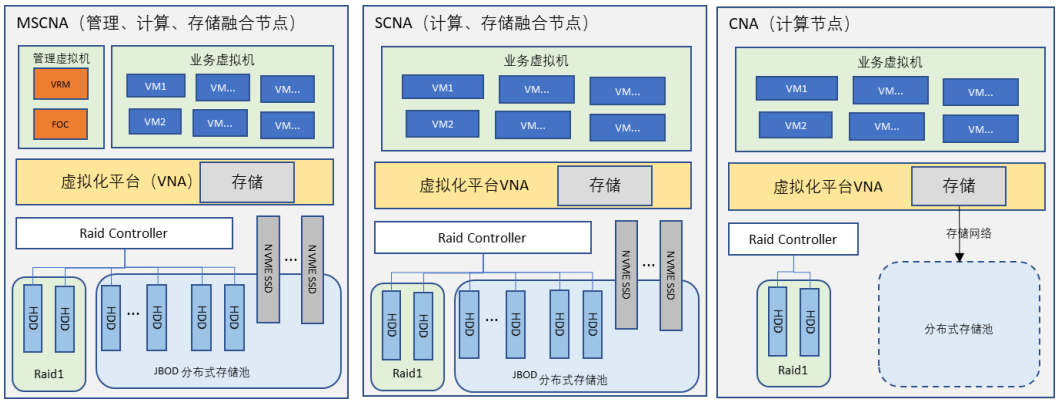
版权所有 © 超聚变数字技术股份有限公司

超融合部署形态

FusionOne HCI根据不同的业务场景与诉求，支持多种部署形态与组网，以下是各个场景的介绍：

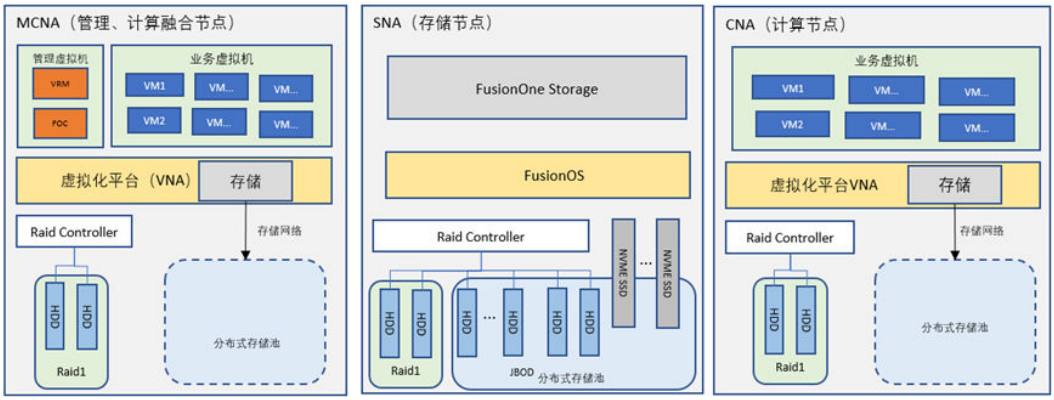
融合部署：常规超融合场景，可提供独立的计算节点，可按需为客户提供灵活的存储、计算资源组合。

图 2-2 融合部署场景



分离部署：针对高性能等客户特殊业务场景，可采用传统的分离架构，并提供硬件、资源、业务的统一纳管。

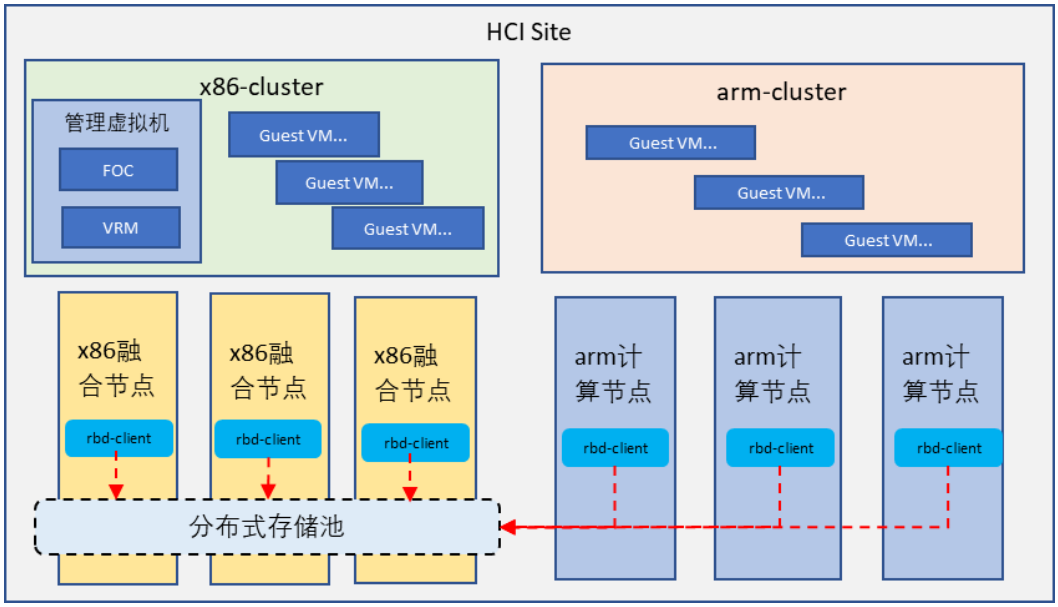
图 2-3 分离部署场景



混合部署：针对信创过渡场景，可支持x86节点+信创节点的共站点模式，为解决客户国产化平台切换提供过渡方案。

版权所有 © 超聚变数字技术股份有限公司

图 2-4 混合部署视示意图



轻量化集群（两节点部署）：针对小规模的边缘站点、分支站点场景，提供两节点起步的低成本方案，配合多套、云管，可在中心站点全局管理。

图 2-5 两节点部署图

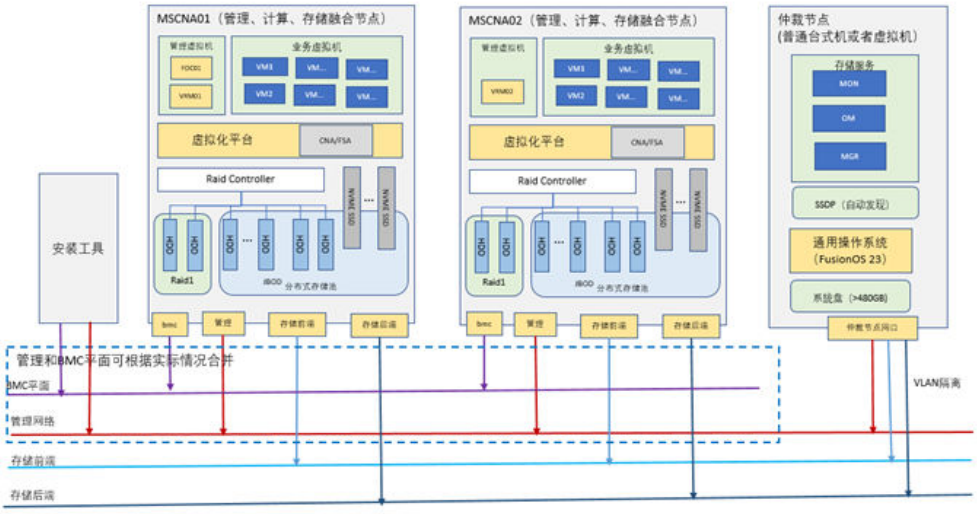


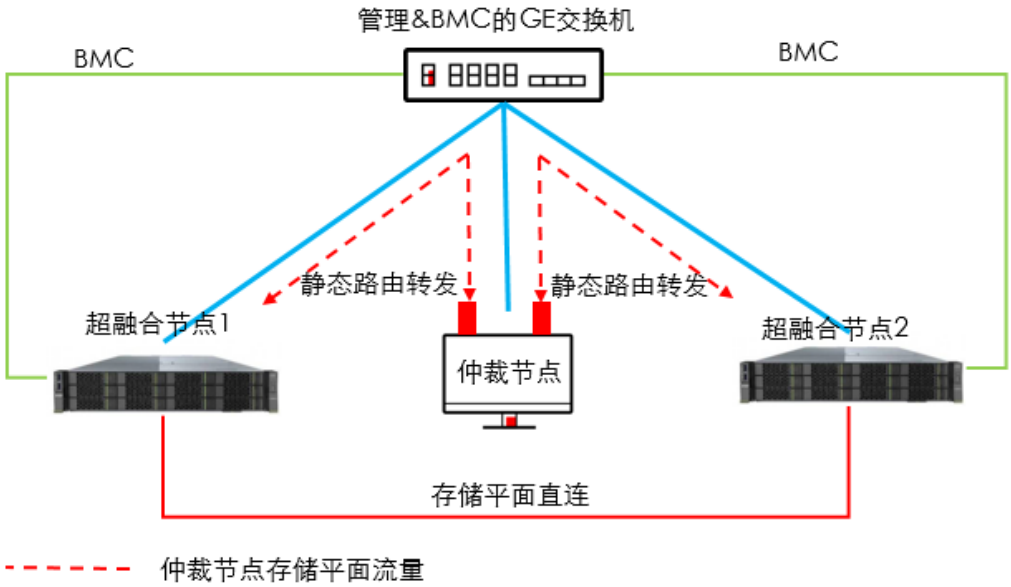
表 2-2 节点类型汇总说明

名称	说明	部署原则
MSCNA（管理节点）	具有管理功能的节点，其上部署了VRM（虚拟化资源管理）、FOC（FusionOne Center）管理虚拟机。同时也可提供存储和计算功能。	根据需要部署2个~多个。

 版权所有 © 超聚变数字技术股份有限公司

名称	说明	部署原则
MCNA（管理节点）	具有管理功能的节点，其上部署了VRM（虚拟化资源管理）、FOC（FusionOne Center）管理虚拟机。同时也可提供计算功能。	根据需要部署2个~多个。
SCNA（存储计算节点）	具有存储、计算功能的节点。提供分布式存储HDD磁盘以及SSD Cache存储资源以及虚拟化计算资源。	根据需要部署0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。
SNA（存储节点）	具有存储功能的节点，提供分布式存储HDD磁盘以及SSD Cache存储资源。	根据需要部署0个~多个。
SARB（仲裁节点）	超融合两节点部署下的仲裁服务，确保两节点模式下的集群可靠性。	两节点场景部署1个（可部署在虚拟机或MINI PC）。

针对边缘站点，支持免存储交换机部署，存储平面可通过节点存储网口直连，实现高性能且降低存储10GE交换机占用。



2.2 典型配置介绍

FusionOne HCI 支持灵活的硬件配置，可根据业务实际需求灵活配置，在典型机型基础上，可以对CPU、内存、网卡、主缓盘、外设灵活按需配置，贴合客户实际场景要求。FusionOne HCI面向不同场景，提供了各种节点类型，容量型节点、均衡型节点、高性能全SSD的性能型节点、纯计算节点、存储节点，具体的场景配置具体如下：

版权所有 © 超聚变数字技术股份有限公司

- **容量型、均衡型节点配置要求：**

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	SAS HDD：600GB、1.2TB、1.8TB、2.4TB SATA HDD：4TB、6TB、8TB、10TB、12TB、14TB、16TB SATA SSD：480GB、960GB、1.92TB、3.84TB 操作系统盘通常为2*600GB SAS盘，或2*480GB SATA SSD，或2*960GB SATA SSD	FusionOne Storage分布式存储要求SATA/SAS盘建议采用3副本冗余策略。
Cache	读密集型NVMe：960GB、1.92TB、3.84TB 读写混合型NVMe：1.6TB、3.2TB	系统的cache大小可根据客户业务压力灵活配置，一般默认配置为1.6TB/3.2TB NVME SSD盘，缓存主存比建议>=5% Cache类型可支持完成兼容性验证的NVMe盘。 建议使用读写混合型NVMe盘，读密集型擦写次数达到上限后存在故障的风险。
网卡	支持GE、10GE、25GE网卡，可灵活组合，推荐四网口起配。	默认推荐管理、业务共用2*10GE网口，存储网络平面独占2*10GE网口。
GPU、NPU	支持NVDIA GPU、华为Accend NPU的直通、虚拟化	x86：支持NVDIA GPU的直通、虚拟化。 arm：支持华为Accend NPU的直通、虚拟化。

 版权所有 © 超聚变数字技术股份有限公司

- **性能型节点配置要求：**

表 2-4 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	读密集型NVMe SSD：1.92TB、3.84TB 读写混合型NVMe SSD：1.6TB、3.2TB SATA SSD：480GB、960GB、1.92TB、3.84TB SAS HDD：600GB 操作系统盘通常为2*600GB SAS盘，或2*480GB SATA SSD，或2*960GB SATA SSD	分布式存储全闪存场景可采用2副本的冗余策略。 推荐使用读写混合型SSD盘，读密集型擦写次数达到上限后存在故障的风险。
网卡	支持GE、10GE、25GE网卡，可灵活组合，推荐四网口起配。	默认推荐管理、业务共用2*10GE网口，存储网络平面独占2*10GE网口。
GPU、NPU	支持NVDIA GPU、华为Accend NPU的直通、虚拟化	x86：支持NVDIA GPU的直通、虚拟化。 arm：支持华为Accend NPU的直通、虚拟化。

- **计算型节点配置要求：**

 版权所有 © 超聚变数字技术股份有限公司

表 2-5 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	默认推荐管理、业务共用2*10GE网口，存储网络平面独占2*10GE网口。

2.3 典型组网介绍

FusionOne HCI 系统组网包含：管理平面、BMC平面、存储平面（前端、后端）、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、存储前后端可根据实际硬件配置，也可以复用物理网口，但是实际生产环境交付中，考虑超融合的性能、安全隔离要求，存储前后端要求独占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：

版权所有 © 超聚变数字技术股份有限公司

图 2-6 存储融合场景组网图

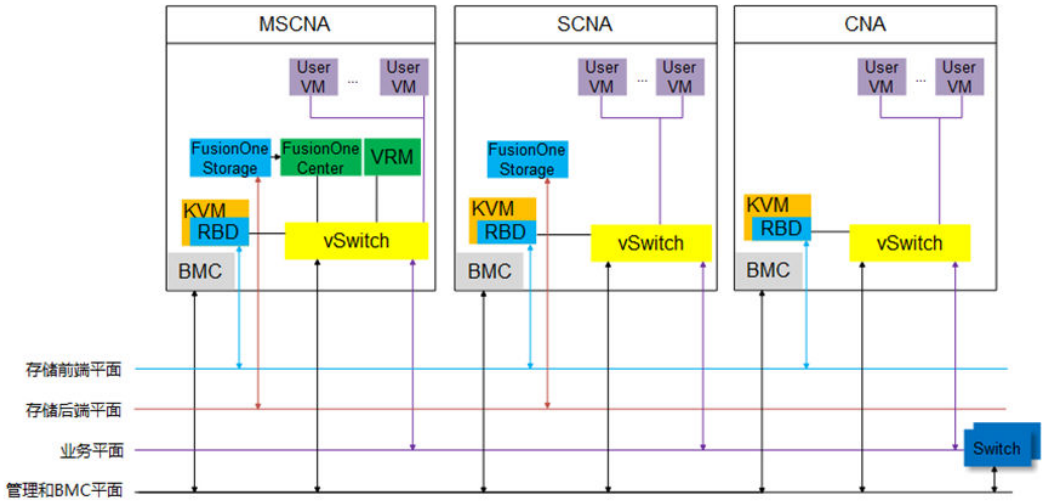
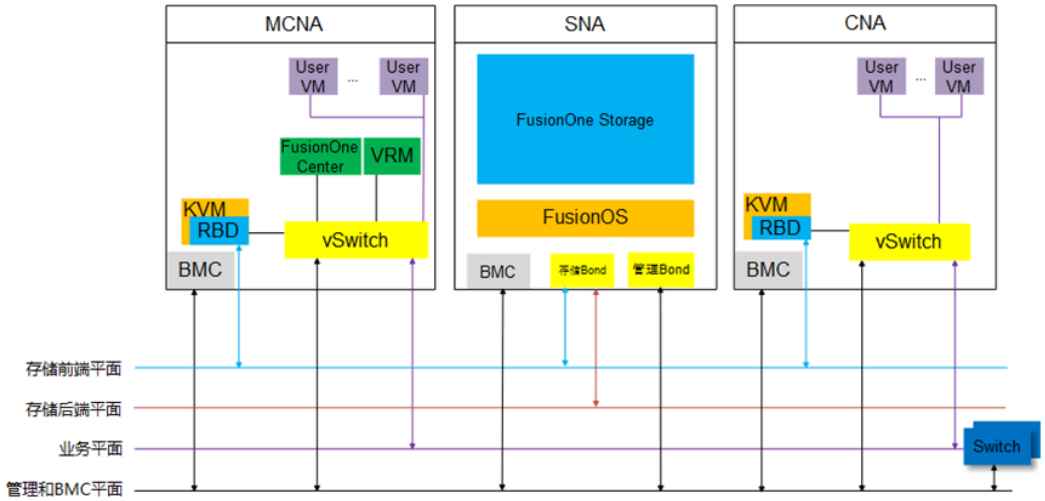


图 2-7 存储分离场景组网图



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。
- 存储前端平面: 存储业务数据读写的网络，对应对应FusionOne Storage网络的存储业务网络平面。
- 存储后端平面：存储节点之间通信的网络，对应FusionOne Storage网络的存储集群网络平面。
- 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。

 版权所有 © 超聚变数字技术股份有限公司

2.4 硬件平台介绍

FusionOne HCI基于超聚变FusionServer系列（Intel&AMD&Hygon）、KUNLUN（ARM-KUNPENG 920/920B）系列服务器，结合超融合业务形态深度调优，将服务器下的高可靠到超融合场景，主要是如下特色技术功能：

分类	特色技术
冗余设计	● 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

版权所有 © 超聚变数字技术股份有限公司

后续章节，针对超融合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（含分布式存储）
- 容器平台：新型敏捷的容器类应用编排平台。
- 统一管理（设备管理、监控管理、运维管理、多云&多站点管理、云管）
- 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 © 超聚变数字技术股份有限公司

3 计算虚拟化

3.1 概述
3.2 虚拟机资源QoS
3.3 CPU超线程
3.4 CPU内存超分
3.5 虚拟机资源动态热添加(DRX)
3.6 虚拟机热迁移
3.7 虚拟机快速克隆
3.8 虚拟机回收站
3.9 第三方虚拟机模板导入
3.10 NUMA调优
3.11 动态资源调度（DRS&DPM）
3.12 快照与快照策略
3.13 XPU可视化管理与调度策略
3.14 GPU虚拟化
3.15 NPU虚拟化
3.16 沐曦GPU虚拟化
3.17 GPU资源池
3.18 XPU直通
3.19 网络USB
3.20 SSH密钥管理
3.21 PCIe设备白名单管理
3.22 GuestOS支持解耦
3.23 noVNC文件上传

 版权所有 © 超聚变数字技术股份有限公司

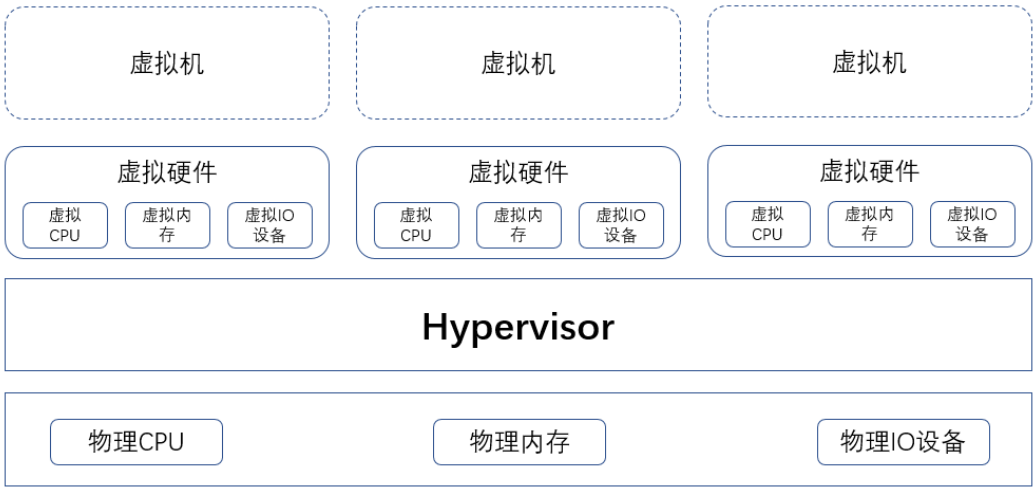
3.1 概述

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 HCI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。

图 3-1 虚拟化架构图



FusionOne Compute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

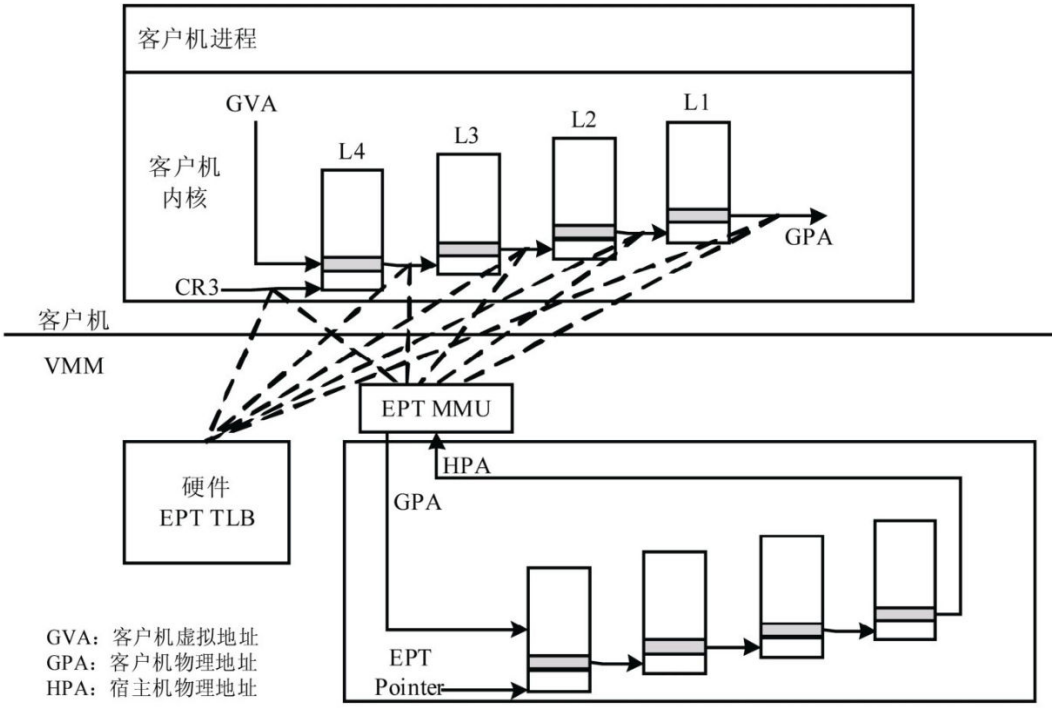
 版权所有 © 超聚变数字技术股份有限公司

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。

虚拟化深度调优、低损耗

 版权所有 © 超聚变数字技术股份有限公司

虚拟化平台通过多种调优方式，实现了虚拟化平台的低损耗，在X86和ARM场景的内存与CPU损耗小于5%，FusionOne HCI技术点如下：

1. 自适应资源QoS管理技术：根据不同的业务场景，精细化智能调度计算、存储、网络资源，实现单台服务器上可同时运行数百台高负载虚拟机。
2. 轻量化Hypervisor技术：通过对Hypervisor深度优化，实现Hypervisor轻量化，运行更快、占用资源更少、更安全、更可靠。
3. 自研FusionOS计算底座：通过对自研操作系统深度优化、软件与硬件的深度协同，实现高性能、低资源消耗、高可靠的计算底座。
4. QAT引擎加速算力释放：通过把加解密、压缩等操作卸载到CPU的QAT加速器中，释放更多的CPU算力给业务。

凭借出色的虚拟化性能，在SPECvirt_sc2013虚拟化基准测试中，通过搭载两颗英特尔®至强®8490H处理器的最新一代FusionServer 2288H V7两路服务器满载运行420台虚机，以7528分总成绩刷新SPECvirt世界纪录，领先榜单第二名超10%。

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。

 版权所有 © 超聚变数字技术股份有限公司

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：

- 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**

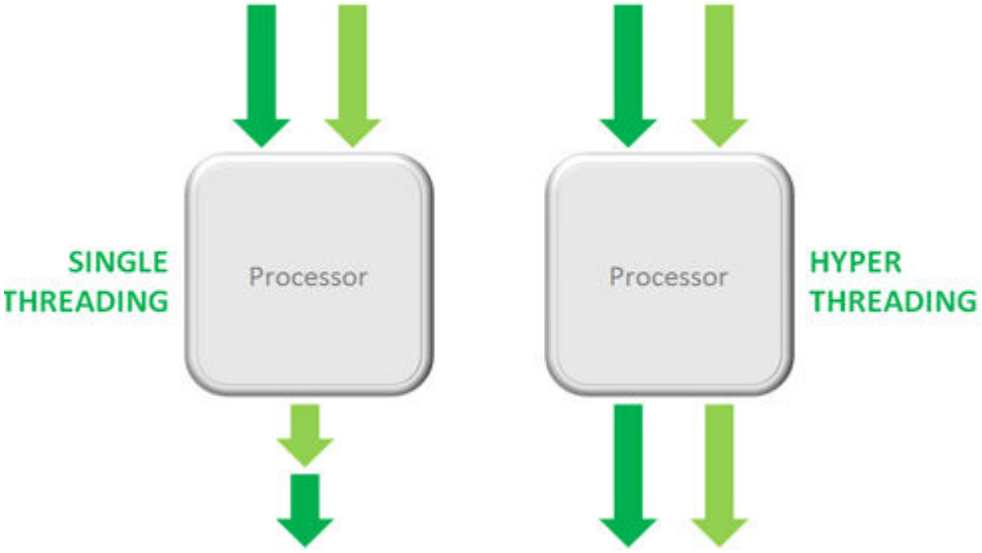
提供带宽配置控制能力，QoS功能不支持同一主机上虚拟机之间的流量限制。包含如下方面：

– 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 超线程

CPU超线程技术是一种允许单个物理处理器执行多个线程的技术。这项技术可以让每个物理核心模拟出更多的逻辑核心，从而在同一时间内处理更多的任务。超线程技术通过让处理器在等待某些操作完成（如内存访问）时，能够切换到另一个线程继续执行指令，从而提高处理器的利用率和系统的整体性能。如下图所示，从任务串行变成了任务并行处理。

 版权所有 © 超聚变数字技术股份有限公司



对于超融合来说，可以让平台可见的CPU资源翻倍。这样可以充分利用CPU的闲置周期，提高核心的利用率。在超融合环境中，这意味着可以更有效地利用有限的硬件资源，支持更多的虚拟机或容器实例。通过超线程技术，CPU可以在等待某些操作（如内存访问）完成时，切换到另一个线程继续执行指令，从而减少处理器空闲时间，提升整体性能。这有助于超融合系统在处理高并发请求时表现得更好，尤其是在运行I/O密集型任务时。

鲲鹏920B是华为公司基于ARM架构开发的服务器处理器，它继承了鲲鹏920的多项关键技术特性，其中包括超线程技术（Simultaneous Multi-Threading, SMT）。SMT技术允许每个物理核心同时执行多个线程，从而提高处理器的并发处理能力和资源利用率。在鲲鹏920B中，与X86 CPU类似，每个物理核心可以模拟出两个逻辑核心，这意味着一个物理核心可以同时处理两个线程的任务。这种技术特别适合于多任务处理环境，如超融合基础设施，因为它可以显著提高系统的整体性能和效率，特别是在处理I/O密集型和计算密集型任务时。FusionOne HCI支持新一代ARM架构的超线程能力，多核能力提升，集群用户可用CPU资源翻倍。

3.4 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。
- 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

 版权所有 © 超聚变数字技术股份有限公司

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	时间片5
vCPU1	running	wait	wait	wait	wait
vCPU2	wait	running	wait	wait	wait
vCPU3	wait	wait	running	wait	wait
vCPU4	wait	wait	wait	running	wait
vCPU5	wait	wait	wait	wait	running

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 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。

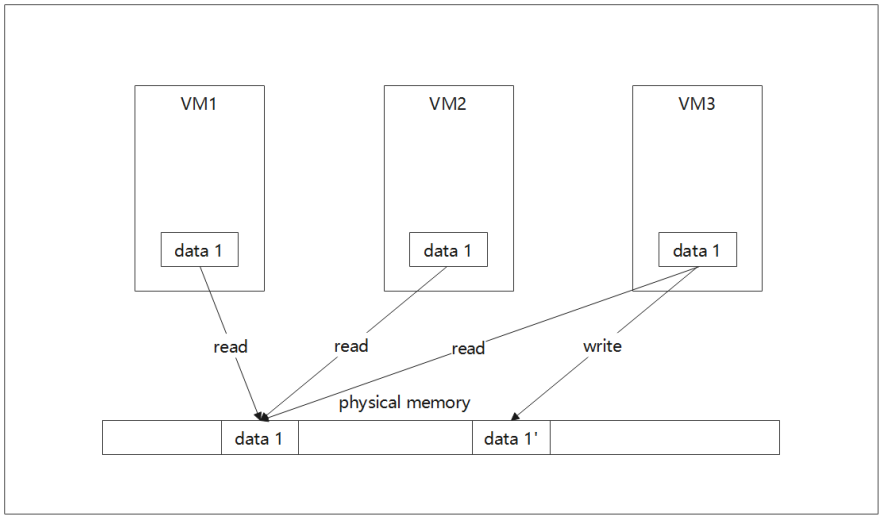
 版权所有 © 超聚变数字技术股份有限公司

内存超分

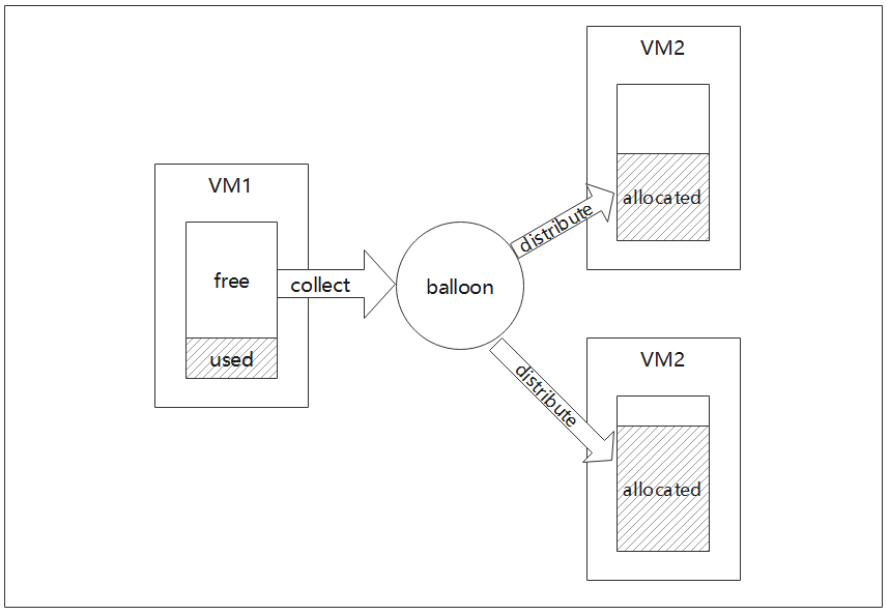
技术原理

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。

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。

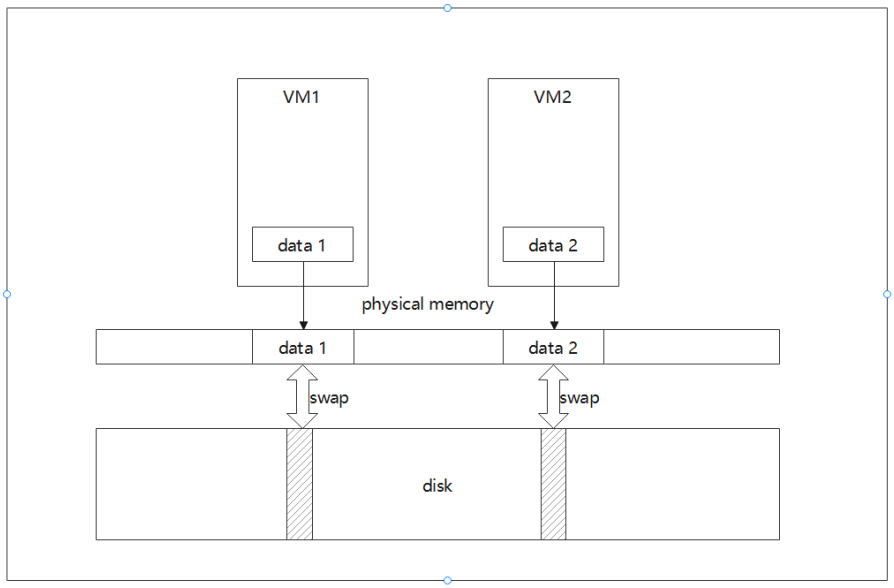


2. 内存气泡：系统主动回收虚拟机暂时不用的物理内存，分配给需要复用内存的虚拟机。内存的回收和分配均为系统动态执行，虚拟机上的应用无感知。



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 © 超聚变数字技术股份有限公司



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。

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。

3.5 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：

 版权所有 © 超聚变数字技术股份有限公司

系统实时监控指定虚拟机组的内存、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。

DRX策略：

用户可根据不同虚拟机设置不同的DRX策略与阈值，如某些业务波动较大的业务配置DRX，而某些相对固定的业务不启用，便于对不同业务的运维。

图 3-4 DRX 配置



3.6 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，可以在不同代次的Intel CPU跨主机热迁移。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。

 版权所有 © 超聚变数字技术股份有限公司

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

  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，如分布式存储卷、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。

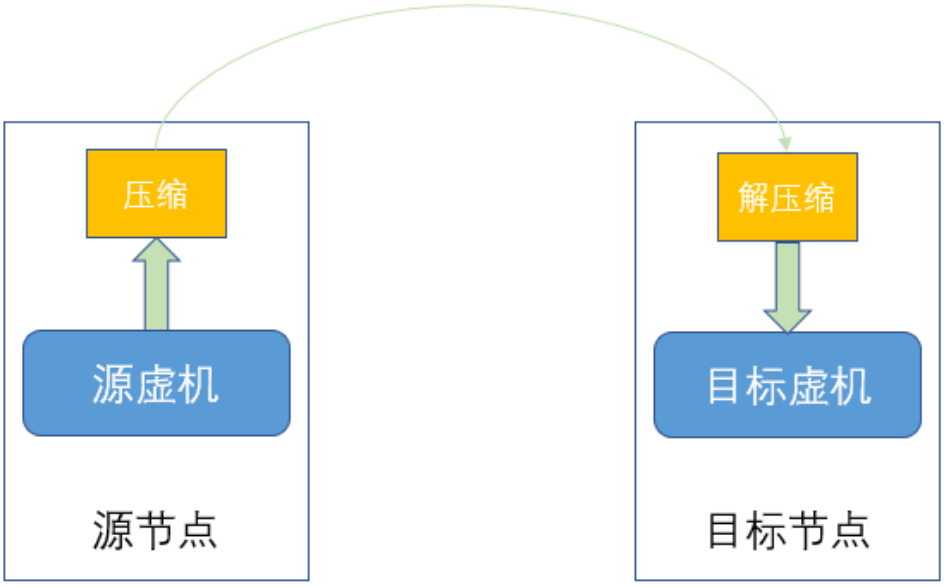
QAT迁移加速

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HCI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HCI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 © 超聚变数字技术股份有限公司



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

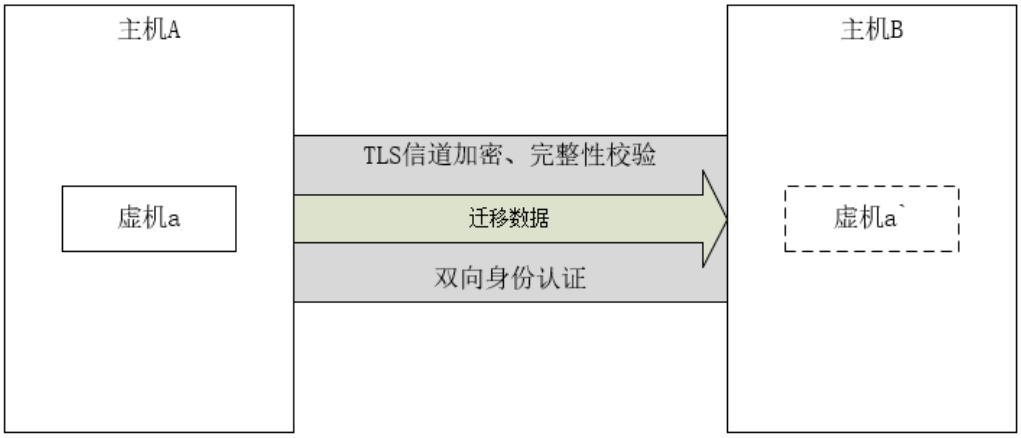
热迁移加密传输：

虚拟机热迁移过程中，由于虚拟机处于运行状态，内存中保存了当前业务的进程状态，特别是整个系统内核的数据结构以及用户空间应用程序的内存空间，如果恶意攻击者通过监听手段，获取迁移数据流的特征来捕获关键信息，例如，恶意篡改或者记录网络迁移中的内存数据，这个信息可能被攻击者用来针对某个特定虚拟机发起第二轮攻击。因此，在不可信网络环境下，虚拟机迁移数据加密对于虚拟化环境下的业务安全是至关重要的。

超聚变虚拟化平台支持配置迁移信道加密，

 版权所有 © 超聚变数字技术股份有限公司

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。



3.7 虚拟机快速克隆

超融合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。



3.8 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。

 版权所有 © 超聚变数字技术股份有限公司

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。

3.9 第三方虚拟机模板导入

FusionOne Compute支持导入Vmware、深信服、浪潮、新华三等第三方平台的OVF和OVA格式的虚拟机模板以及VMDK、qcow2虚拟磁盘导入，其原理如下：

- 按要求从Vmware等第三方平台中把虚拟机导出为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

  说明：支持从本地和共享目录导入ovf和qcow2格式

3.10 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。

 版权所有 © 超聚变数字技术股份有限公司

3.11 动态资源调度（DRS&DPM）

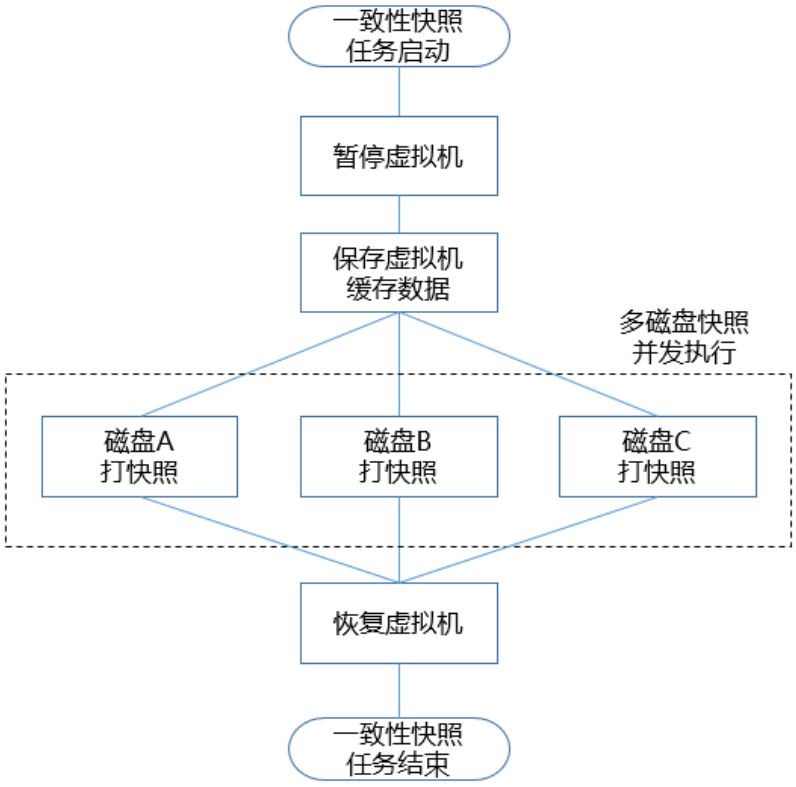
动态资源调度DRS（Dynamic Resource Scheduler），指采用智能负载均衡调度算法，并结合动态电源管理功能，通过周期性检查同一集群资源内各个主机的负载情况，在不同的主机间迁移虚拟机，从而实现同一集群内不同主机间的负载均衡，并最大程度降低系统的功耗。

- 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。

3.12 快照与快照策略

FusionOne Center 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：支持普通快照、内存快照、一致性快照。
  - 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 © 超聚变数字技术股份有限公司

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：

- 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。
- 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。
- 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。

3.13 XPU 可视化管理与调度策略

XPU可视化：

XPU可视化是利用XPU的管理工具，进行可视化数据采集，包括XPU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：

- 直通场景：XPU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：XPU虚拟化依赖于XPU的驱动能力，并受主机XPU驱动管理，因此主机上可获取主要的XPU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。

XPU调度：

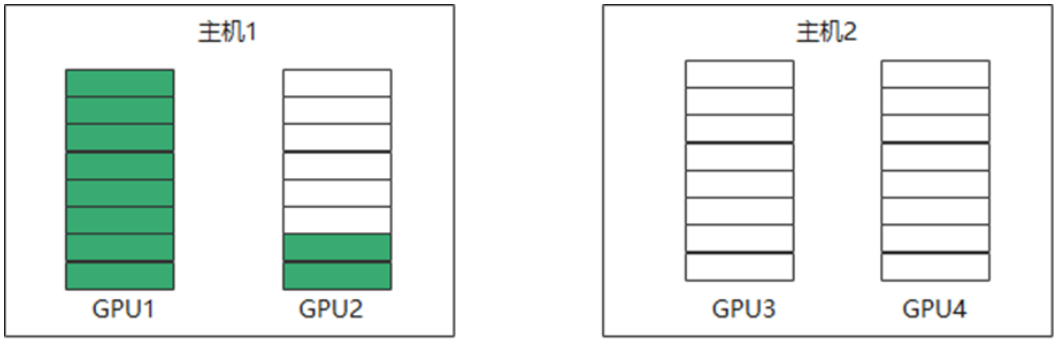
XPU调度分为虚拟机调度与XPU资源分配调度。虚拟机调度即按照一定策略在虚拟机开机或HA的过程中，调度到指定的主机上，如果虚拟机使用的是GPU虚拟化组，则还支持将对应虚拟机资源调度到指定的GPU卡上。目前GPU虚拟化支持以下两种调度模式：

- 密度模式：密度模式需要保证尽可能支持启动更多的带XPU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

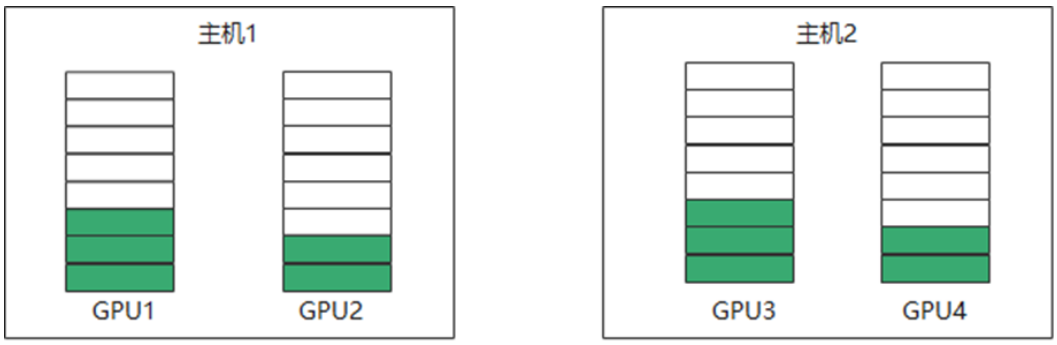
 版权所有 © 超聚变数字技术股份有限公司

到已使用直通XPU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、vNPU分配到已分配资源的XPU卡上。

- 性能模式：性能模式需要保证尽可能让XPU虚拟机均匀负载在所有主机或XPU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XPU计算虚机平均分配到XPU计算资源上。



密度模式



性能模式

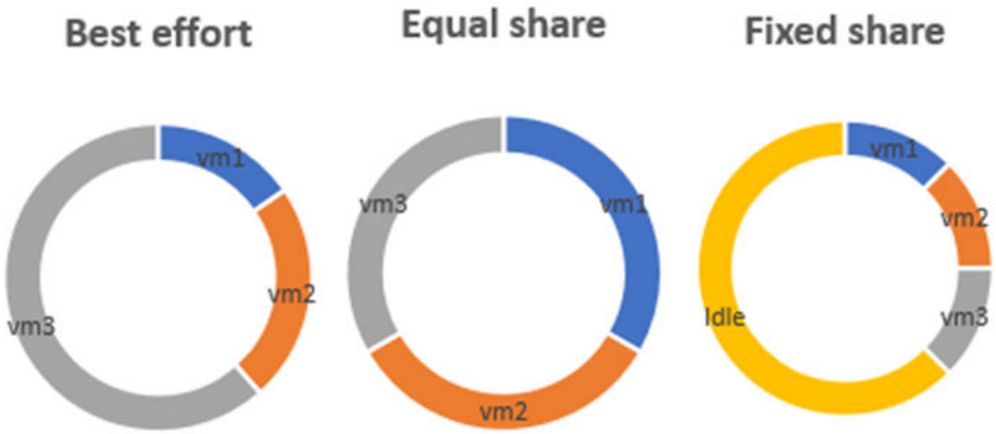
GPU资源分配策略：

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，性能最优，存在资源调度不公平。
- 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。该方式性能次之，GPU算力未能充分发挥性能。
- 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，调度最公平的方式，不会因为其他虚拟机造成性能波动，但是会造成运行的虚拟机少于切分的场景下的算力浪费。

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。抢占方式则根据业务需求尽量占用整个CPU的算力资源，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 © 超聚变数字技术股份有限公司



3.14 GPU 虚拟化

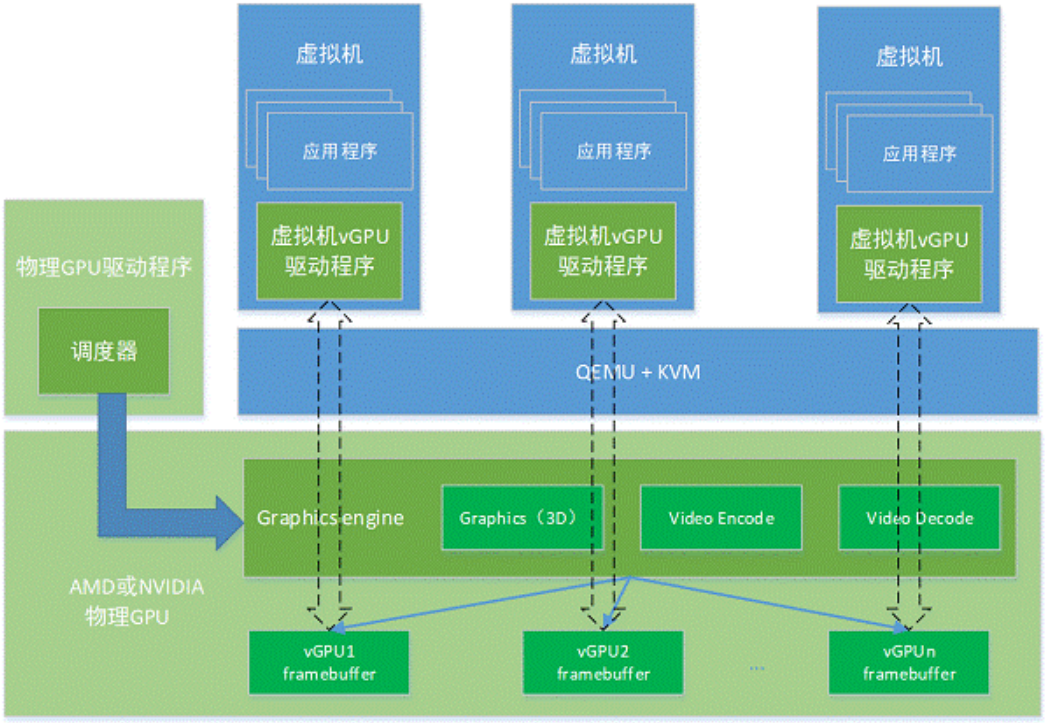
vGPU（Virtual GPU）是针对英伟达GPU的虚拟化环境中提供虚拟显卡的技术，通过在宿主机上运行虚拟化软件，在虚拟机中模拟出一个独立的显卡设备，让虚拟机可以直接访问显卡资源，从而实现高性能图形加速。通过vGPU技术，可分时复用GPU，充分利用GPU的算力。

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。

 版权所有 © 超聚变数字技术股份有限公司



FusionOne HCI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。

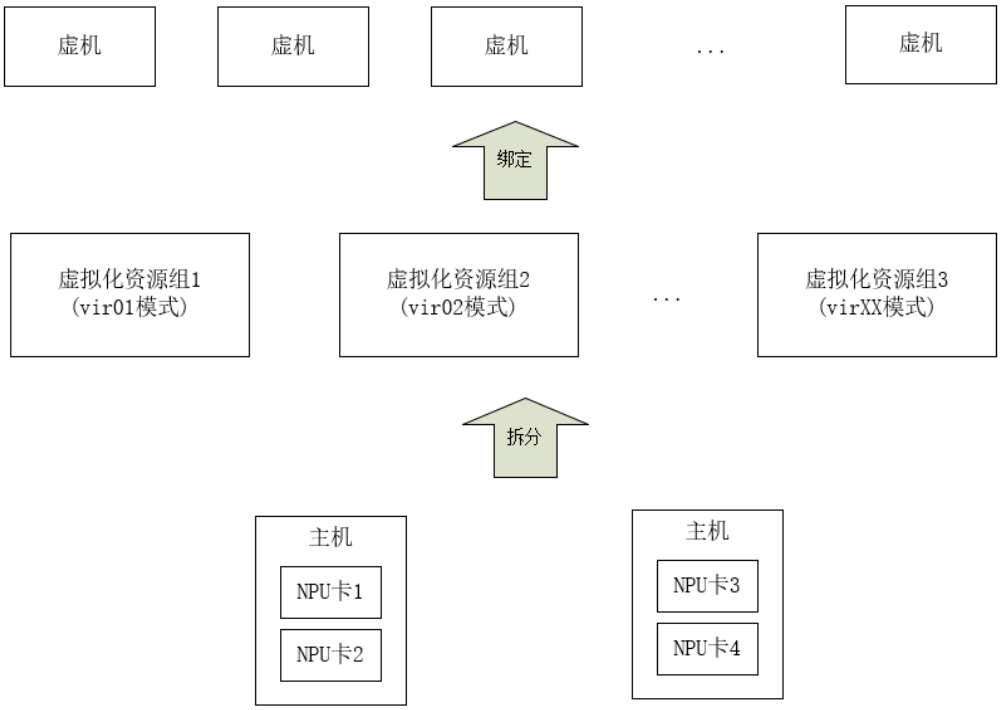
3.15 NPU 虚拟化

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支持以不同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。

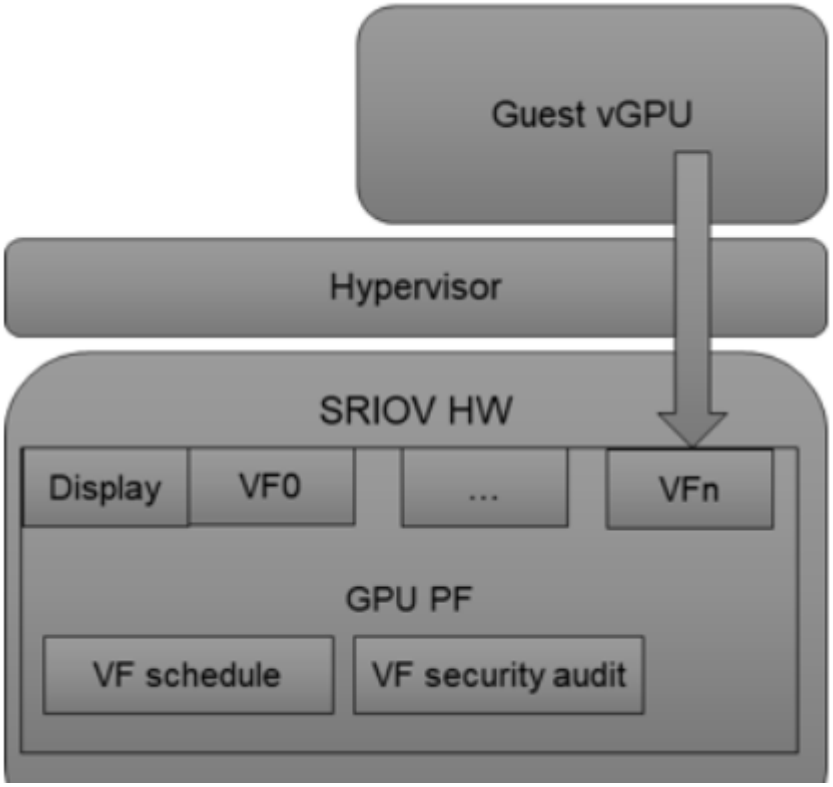
 版权所有 © 超聚变数字技术股份有限公司



3.16 沫曦 GPU 虚拟化

沫曦GPU卡的虚拟机采用SRIOV的PF和VF来实现，SRIOV可以将物理GPU虚拟为多个VF使用，每个VF拥有物理GPU的部分硬件资源。开启SRIOV硬件虚拟化时可以指定VF数量，以及对哪些GPU设备开启虚拟化。沫曦GPU卡可配置成1/2/4/8个VF。多个VF的资源是均等的。再将VF透传到虚拟机中使用，这种使用方式称为透传模式。

 版权所有 © 超聚变数字技术股份有限公司



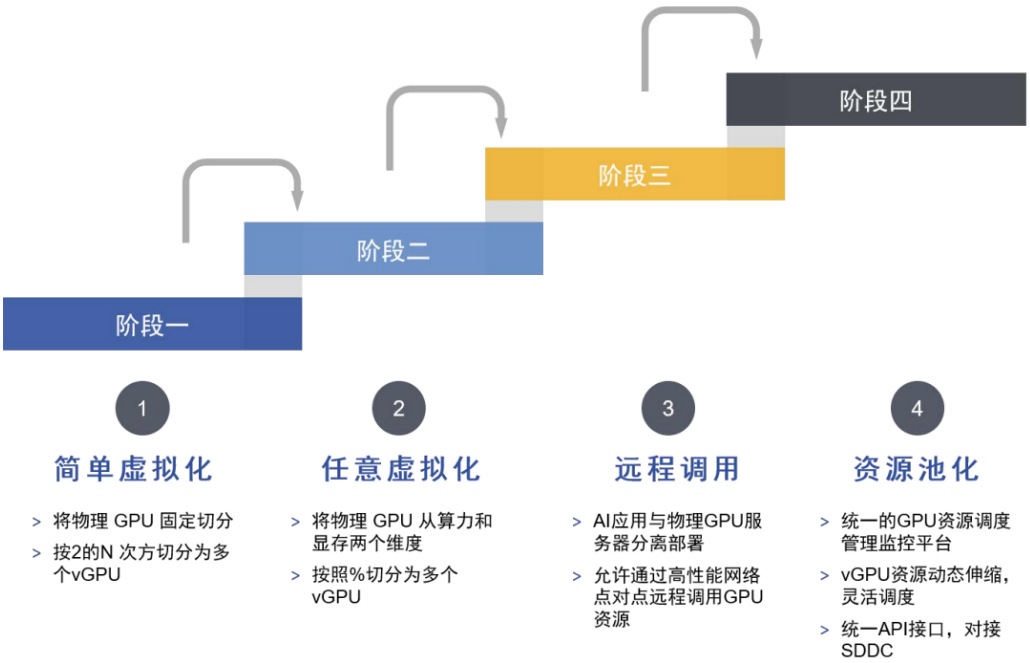
说明

1. 一旦开启SRIOV，物理GPU与vGPU将不支持MetaXLink互联
2. 由于厂商限制，节点上所有沫曦都切分成相同数量的vGPU，不支持不同vGPU数量
3. 切分成vGPU后，不用单独的license授权

3.17 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。

 版权所有 © 超聚变数字技术股份有限公司



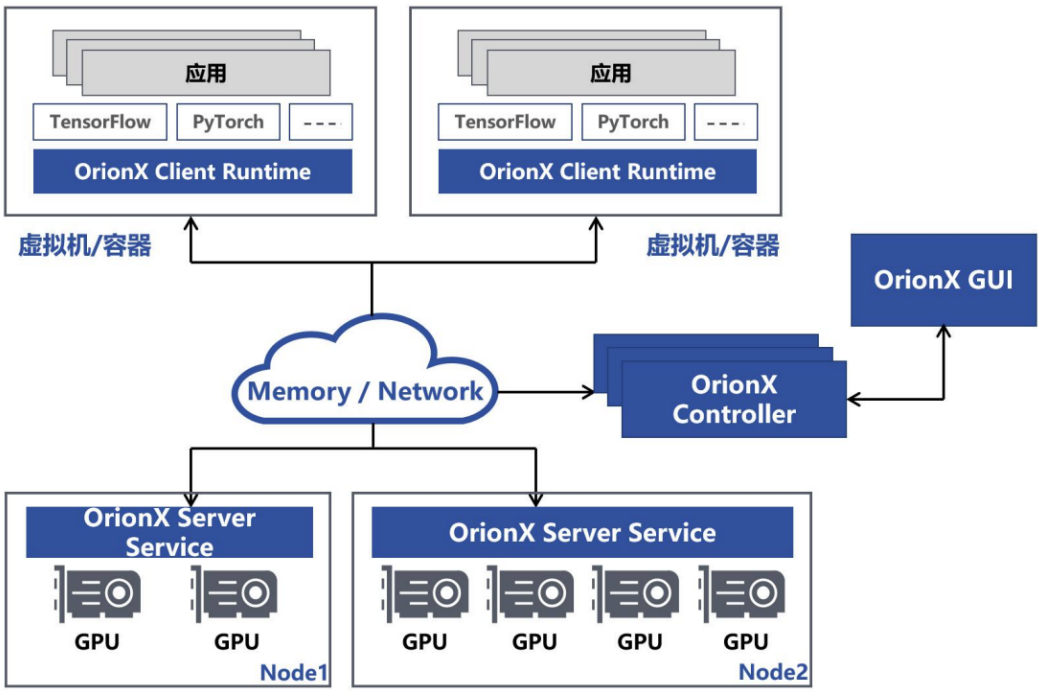
- **简单虚拟化。**将物理 GPU 按照 2 的 N 次方，切分成多个固定大小的 vGPU（VirtualGPU，虚拟 GPU），每个 vGPU 的算力和显存相等。实践证明，不同的 AI 模型对于算力、显存资源的需求是不同的。所以，这样的切分方式，并不能满足 AI 模型多样化的需求。
- **任意虚拟化。**将物理 GPU 按照算力和显存两个维度，自定义切分，获得满足 AI 应用个性化需求的 vGPU。
- **远程调用。**AI 应用与物理 GPU 服务器分离部署，允许通过高性能网络远程调用 GPU资源。这样可以实现 AI 应用与物理 GPU 资源剥离，AI 应用可以部署在私有云的任意位置，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。

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。

 版权所有 © 超聚变数字技术股份有限公司



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）

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）

 版权所有 © 超聚变数字技术股份有限公司

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：

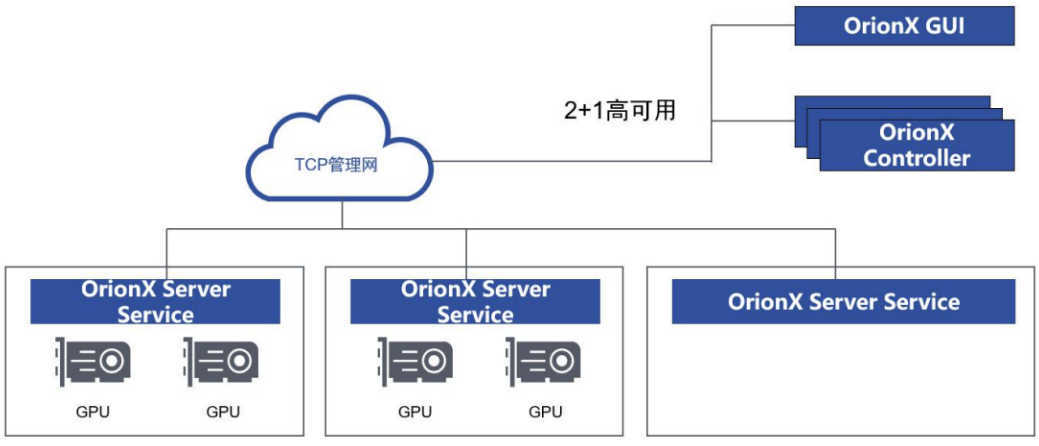
- 分级的运维账号登录与管理。
- 查看各组件的部署拓扑以及详情。
- 多维度、多角度查看资源池的资源使用情况。
- 细粒度管理资源池内的资源。
- 日志、监控及告警功能设置。

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。

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：

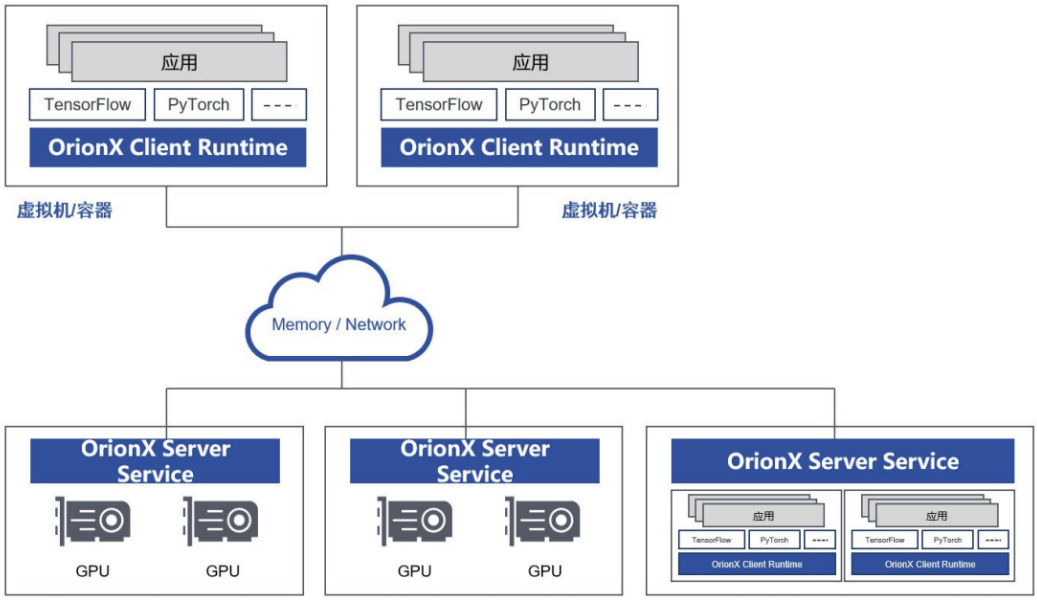
- OrionX Controller 支持多副本、高可用的部署模式。
- 各个功能组件启动的次序无要求。
- 当某一个功能组件从错误中恢复之后，可以自动同步到正确的状态。

 版权所有 © 超聚变数字技术股份有限公司

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。



部署形式

OrionX 支持在超融合的主机上部署。轻松实现GPU资源池化，同时支持CentOS、Ubuntu、Debian 等 Linux 发行版本。

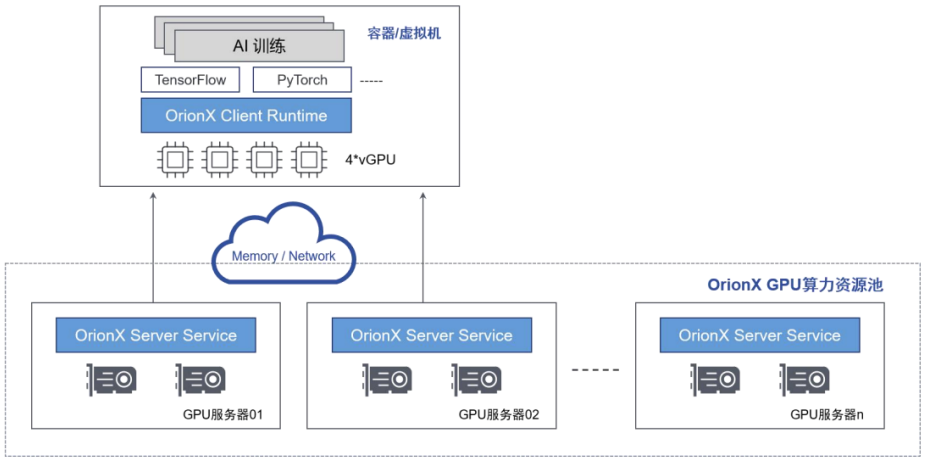
应用场景

- **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。

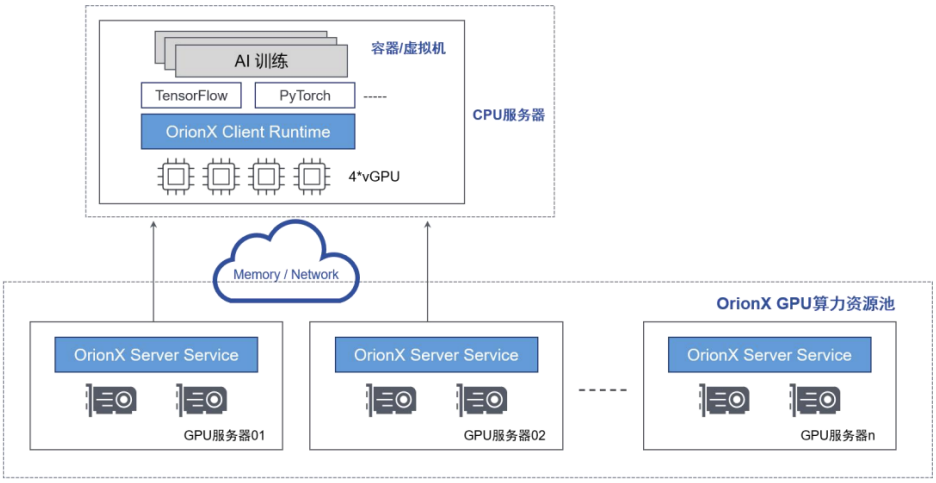
  – **“化零为整”功能支持训练。**OrionX 支持将多台服务器上的 GPU 提供给一个虚拟机或者容器使用，而该虚拟机或者容器内的基于分布式训练框架（Horovod 或 DistributedDataParallel）的 AI 应用无需修改代码。通过这个功能，用户可以将多台服务器的 GPU 资源聚合后提供给单一虚拟机或者容器使用。“化零为整”支持训练等大模型场景，为用户的 AI 应用提供数据中心级的海量算力。Horovod 是 Uber 开源的分布式深度学习框架，旨在使分布式深度学习变得快速且易于使用，使模型训练时间从几天和几周缩短到数分钟和数小时。使用 Horovod，可以将现有的训练脚本扩大规模，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 。

 版权所有 © 超聚变数字技术股份有限公司



– **“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。

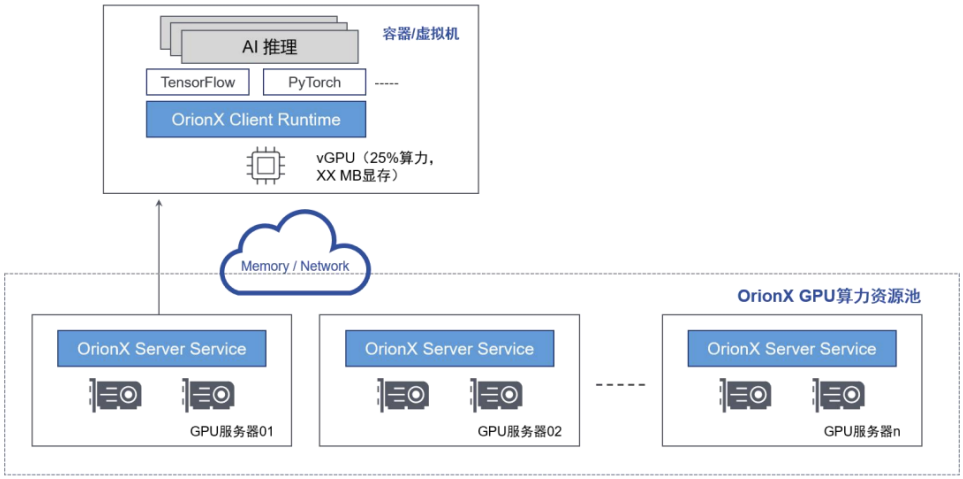


- **小模型场景的典型应用**

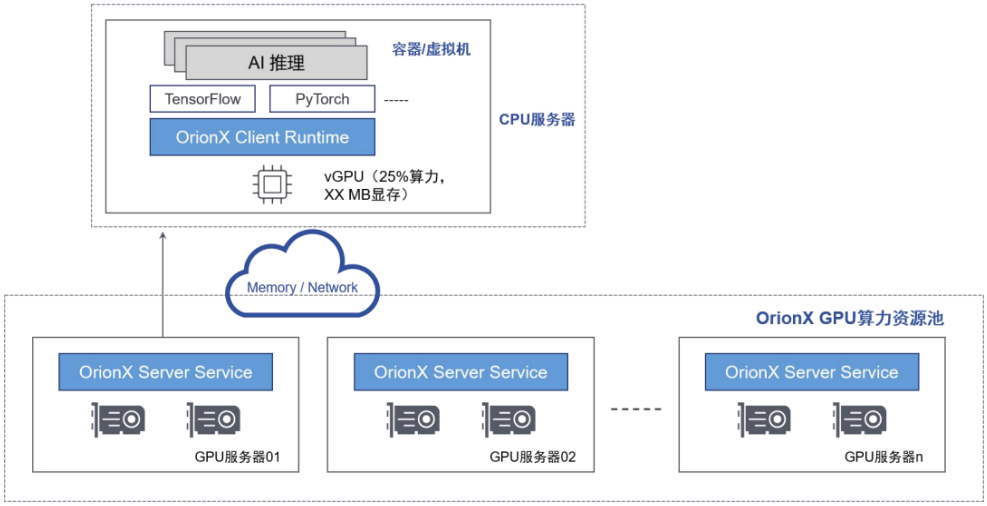
  小模型如推理、开发、教学实训等场景，对算力资源需求量小，通常不能占满一张 GPU 卡资源。作为 AI 算力资源池平台，OrionX 可以从算力和显存两个维度，切分 GPU。支持将多个小模型任务调度到一张卡，有效提高资源利用率。

  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。

 版权所有 © 超聚变数字技术股份有限公司



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。



- **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。

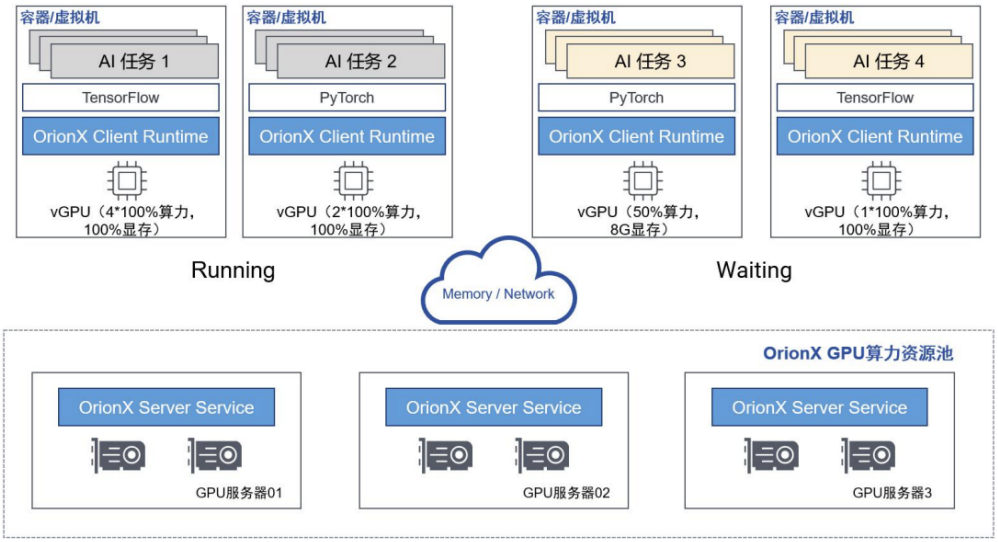
 版权所有 © 超聚变数字技术股份有限公司

– 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。



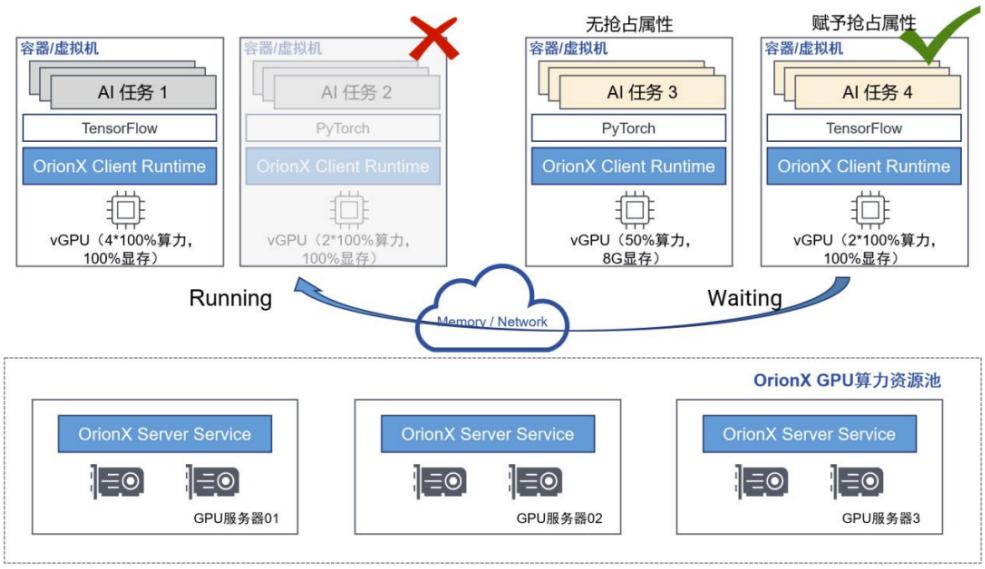
通过“任务队列”功能支持训练/推理任务自动排队。当请求 vGPU 资源的训练/推理任务遇到物理 GPU 剩余资源不足，无法分配的时候，OrionX支持任务排队能力。OrionX 会将任务放入等待队列中，直到队列中前面任务跑完，可调度等待任务所需资源时，任务进入运算状态。

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。

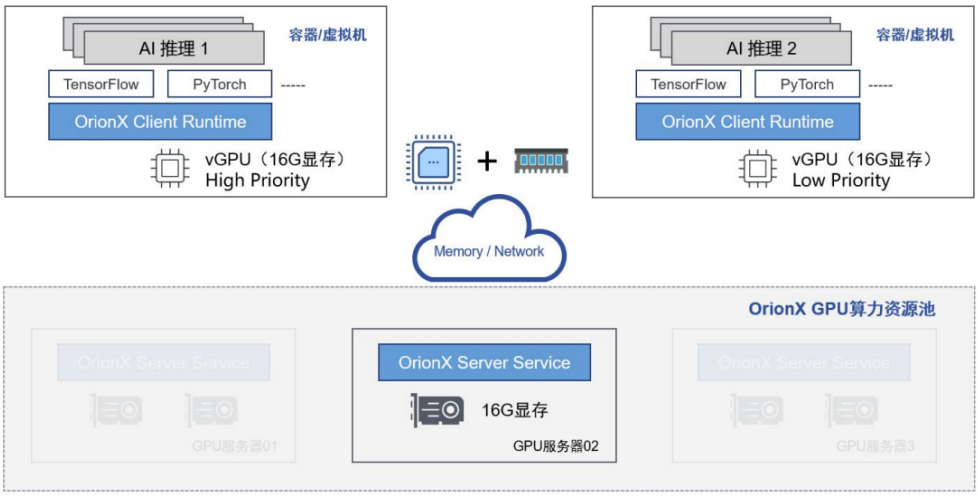
 版权所有 © 超聚变数字技术股份有限公司



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。

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

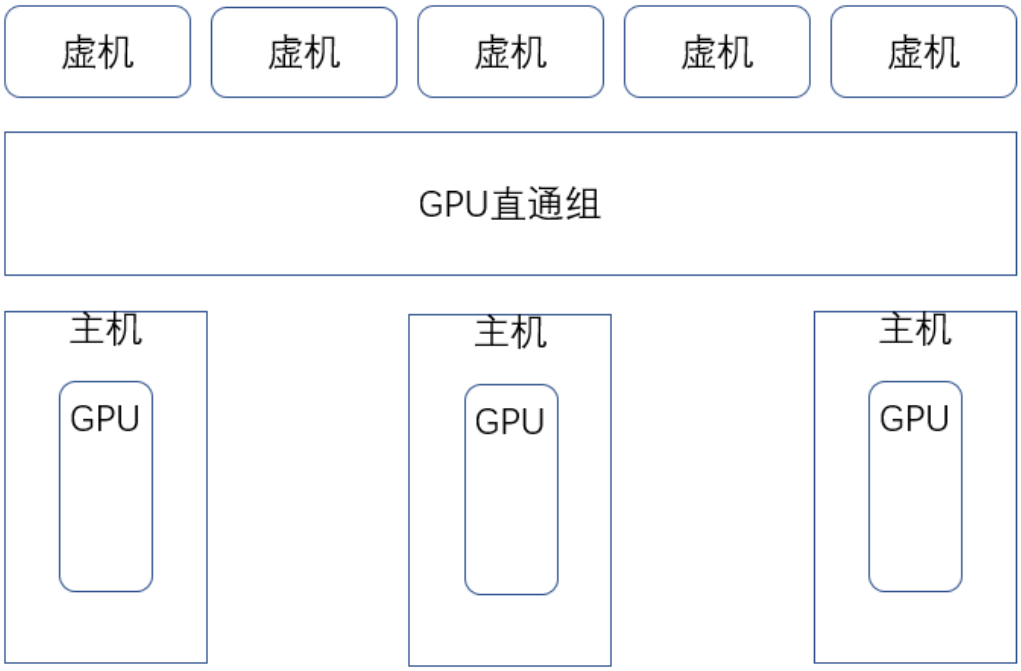
- **提高利用率**
  - 支持将 GPU 切片为任意大小的 vGPU，从而允许多 AI 负载并行运行，提高物理 GPU 利用率。
  - 提高 GPU 综合利用率多达 3-10 倍，1 张卡相当于起到 N 张卡的效果，真正做到昂贵算力平民化。
- **高性能**

 版权所有 © 超聚变数字技术股份有限公司

  - 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%。
  - vGPU 资源隔离，并行用户无资源互扰。
- **轻松弹性扩展**
  - 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。
  - 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**
  - 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。
  - 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**
  - 提供 GPU 资源管理调度策略。
  - 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。
  - 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。

3.18 XPU 直通

XPU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，实现与物理主机使用XPU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XPU设备添加到一个组中，虚拟机与XPU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）。超聚变虚拟化平台支持多种GPU卡的直通，包括主流的NVDIA、华为昇腾、数智铠50/100、沐曦 C500、N100、燧原S60、壁韧110E、铨兴AI扩容卡等。

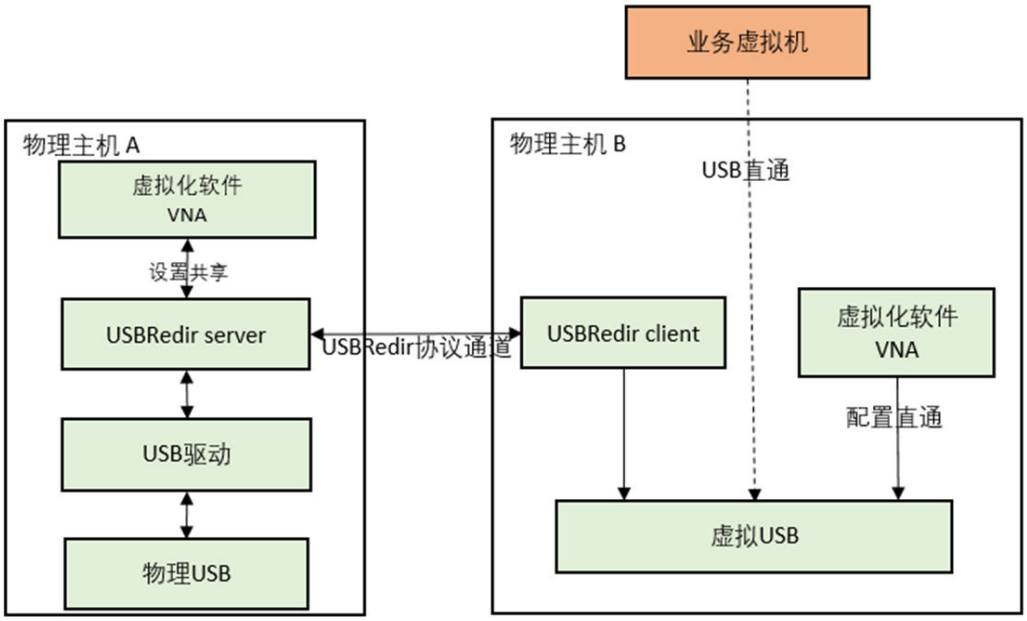


 版权所有 © 超聚变数字技术股份有限公司

3.19 网络 USB

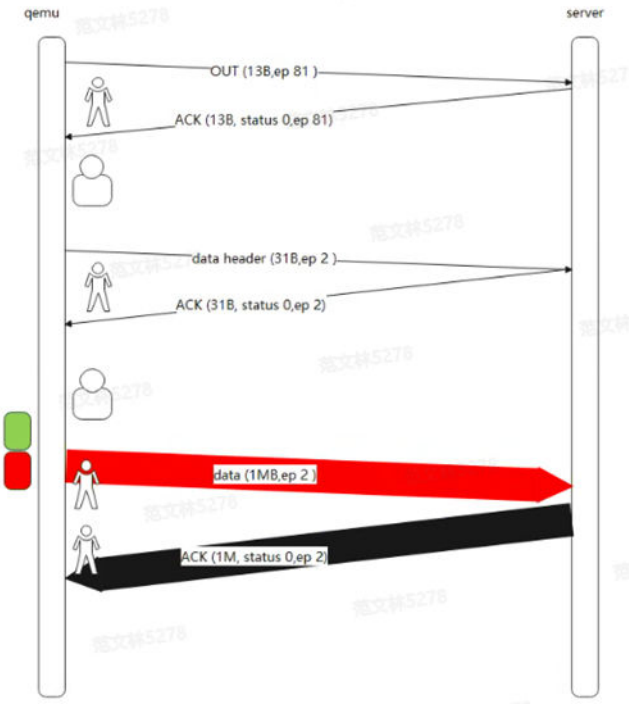
在虚拟机中，某些业务场景，如财务结算应用，需要提供USB支持，包括USB存储盘、USB key等。传统的方式采用将USB直通给虚拟机，这样会带来一个问题，虚拟机无法漂移给其他主机，可维护性差。

FusionOne HCI支持无代理跨物理主机虚拟机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。



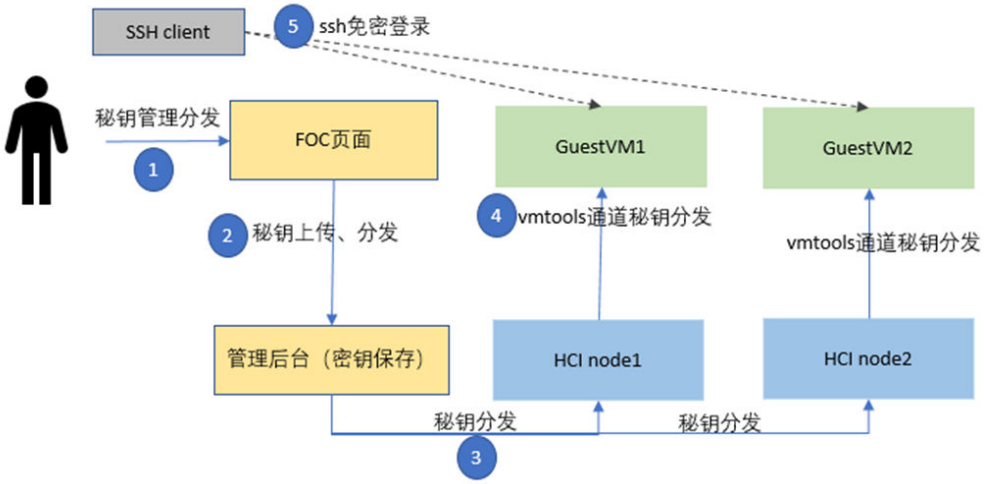
USB设备跨主机热迁移需要保证IO不中断，超聚变通过自研的IO冻结、IO缓存、重传等一系列的机制保证IO持续运行。虚拟机迁移时，当检测到有USB设备时，在虚拟机暂停期间，将IO冻结，并将以下时序处理中的消息缓存，当虚拟机迁移到目标主机后，将USB读写消息重传，从而保证业务IO的连续性。

版权所有 © 超聚变数字技术股份有限公司



3.20 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。

 版权所有 © 超聚变数字技术股份有限公司

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。

3.21 PCIe 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。

3.22 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。

3.23 noVNC 文件上传

noVNC 作为网页版 VNC 客户端，依托标准 RFB 协议连接虚拟机，原生仅支持文本剪贴板同步，不具备内置文件上传能力；为提升易用性，FusionOne HCI借助 VMtools配合虚拟串口完成宿主机与虚拟机的数据互通，用户可直接在 noVNC 页面选取本地小文件（<1MB》），实现文件向虚拟机内的传输。

 版权所有 © 超聚变数字技术股份有限公司



 版权所有 © 超聚变数字技术股份有限公司

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 内网域名解析DNS
4.14 网络QoS
4.15 SR-IOV交换
4.16 用户态OVS交换
4.17 DHCP服务
4.18 虚拟路由器
4.19 端口流量镜像
4.20 在线修改虚拟机IP地址
4.21 虚拟负载均衡器
4.22 RDMA技术

 版权所有 © 超聚变数字技术股份有限公司

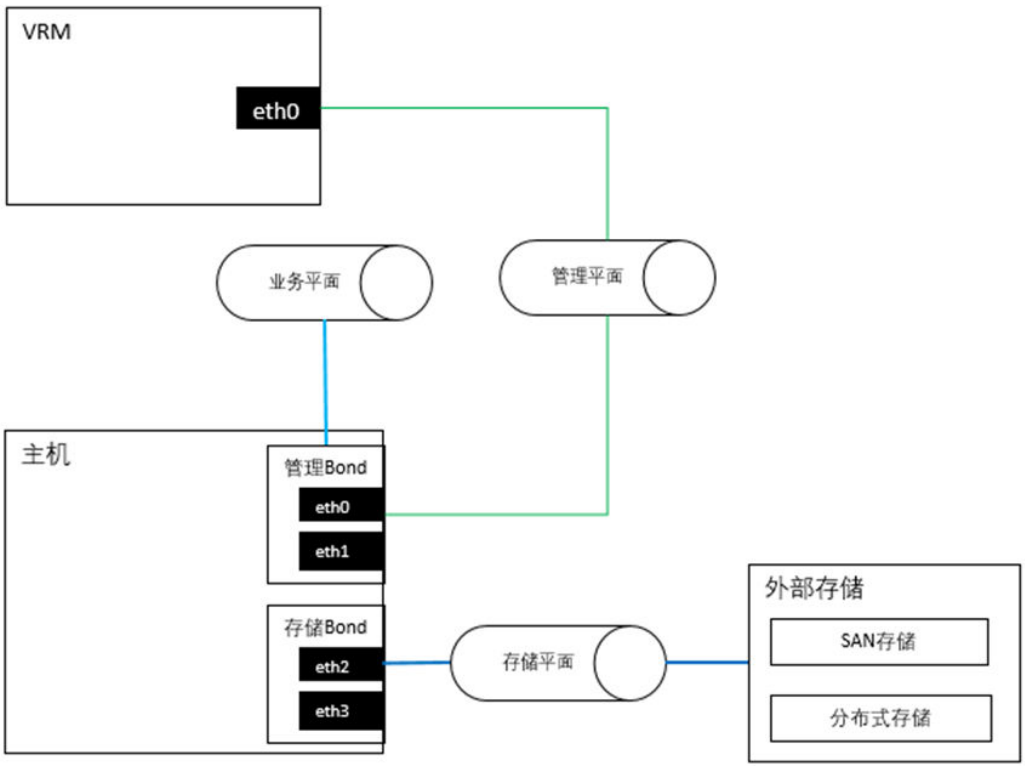
4.1 概述

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，超融合采用了业界成熟的软件定义网络SDN+虚拟化技术相结合的解决方案，从而实现企业级云架构中的网络虚拟化。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。超融合即支持简单易用的OVS二层网络，同时也支持SDN网络虚拟化，可为用户提供灵活的选择。

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。

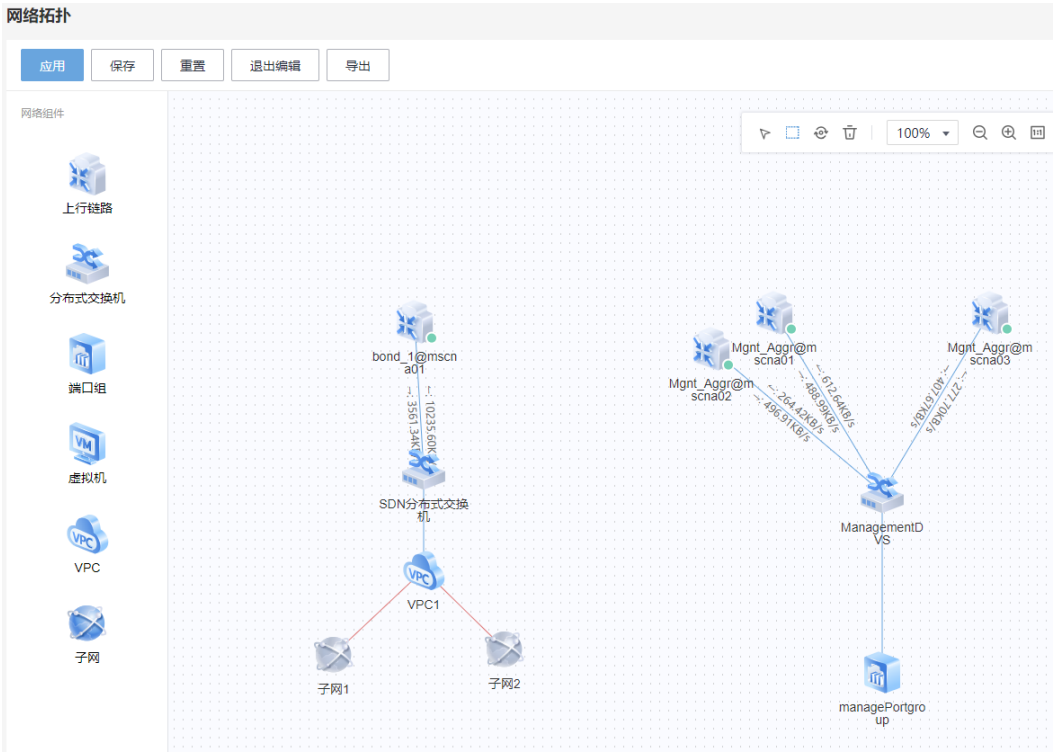


版权所有 © 超聚变数字技术股份有限公司

4.2 所画即所得

超融合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超融合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。

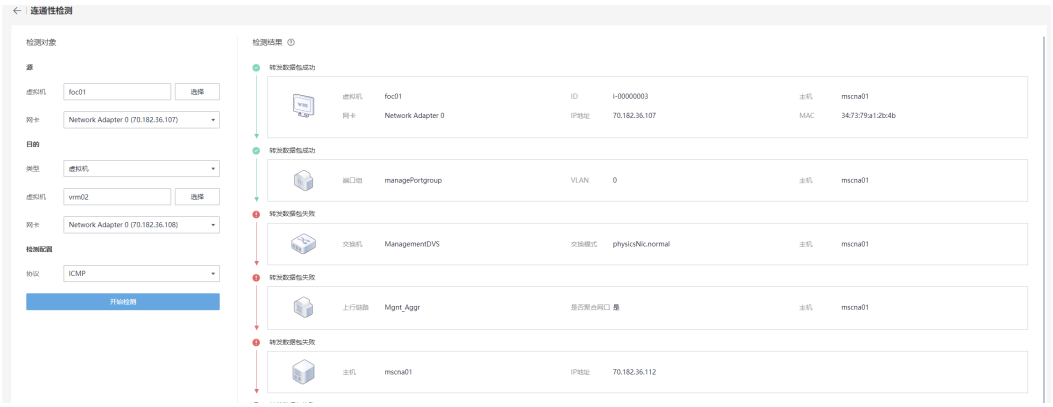


4.3 连通性检测

超融合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 © 超聚变数字技术股份有限公司



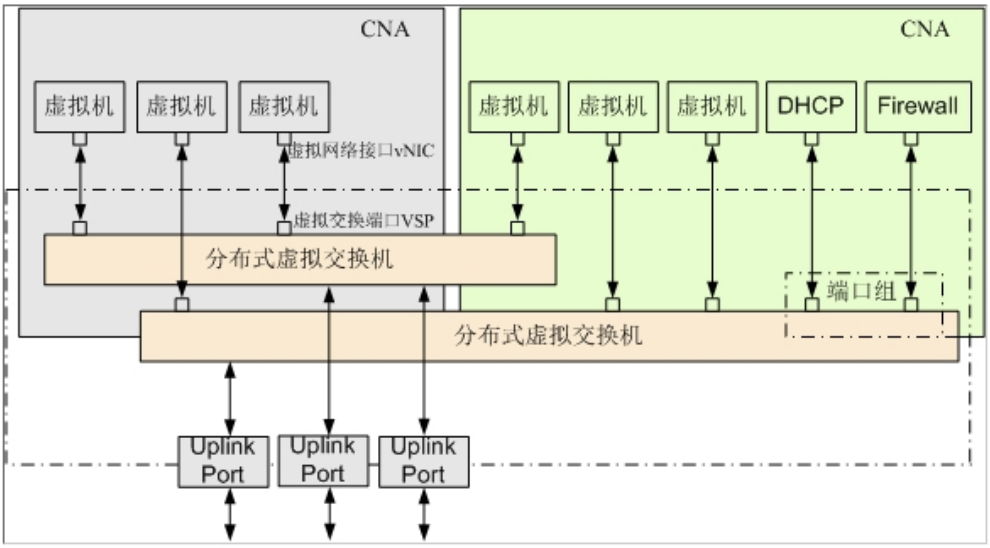
4.4 分布式虚拟交换机

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。

超融合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（Overlay、Underlay），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：



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

 版权所有 © 超聚变数字技术股份有限公司

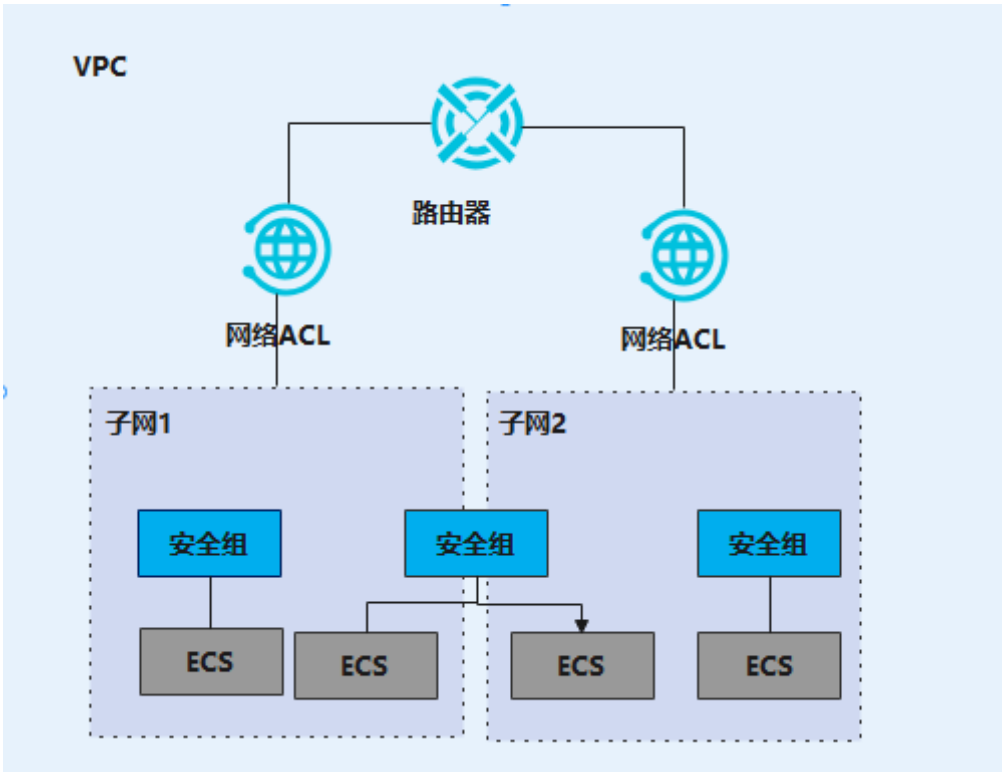
相同属性的一组端口，相同端口组的VLAN（Virtual Local Area Network）、QoS、安全属性相同。

- 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。
- 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

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。

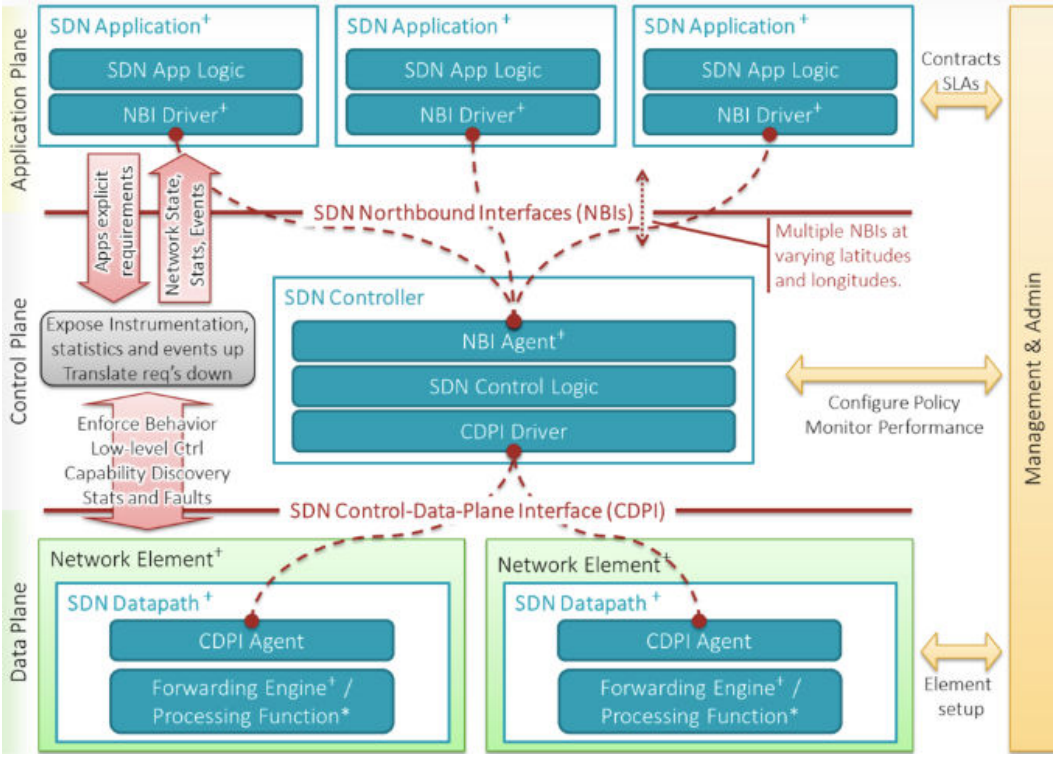
版权所有 © 超聚变数字技术股份有限公司

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、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

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。

 版权所有 © 超聚变数字技术股份有限公司

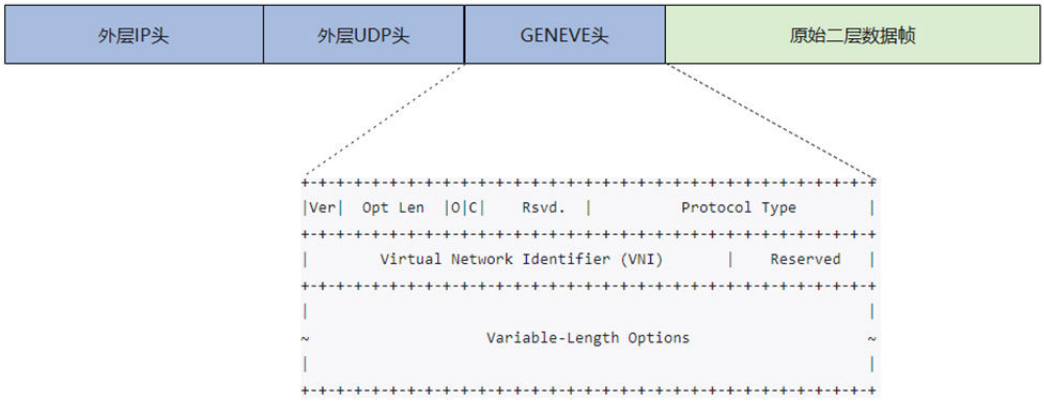
表 4-1 Underlay 网络 VS Overlay 网络：

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。

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：

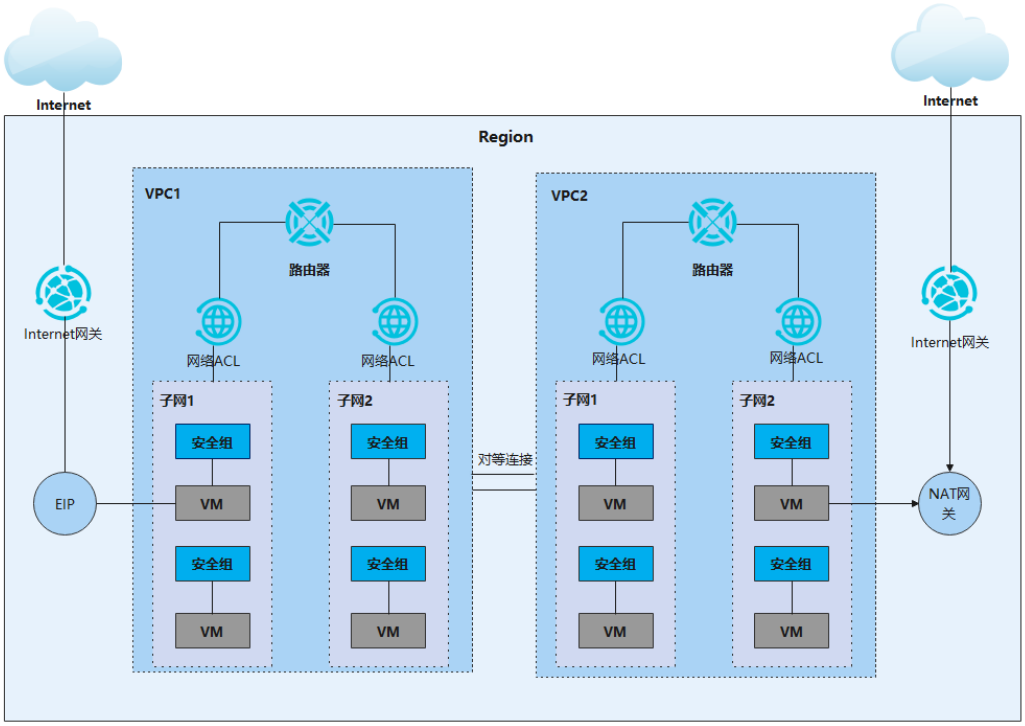


 版权所有 © 超聚变数字技术股份有限公司

- 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。

4.8 VPC

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：

 版权所有 © 超聚变数字技术股份有限公司

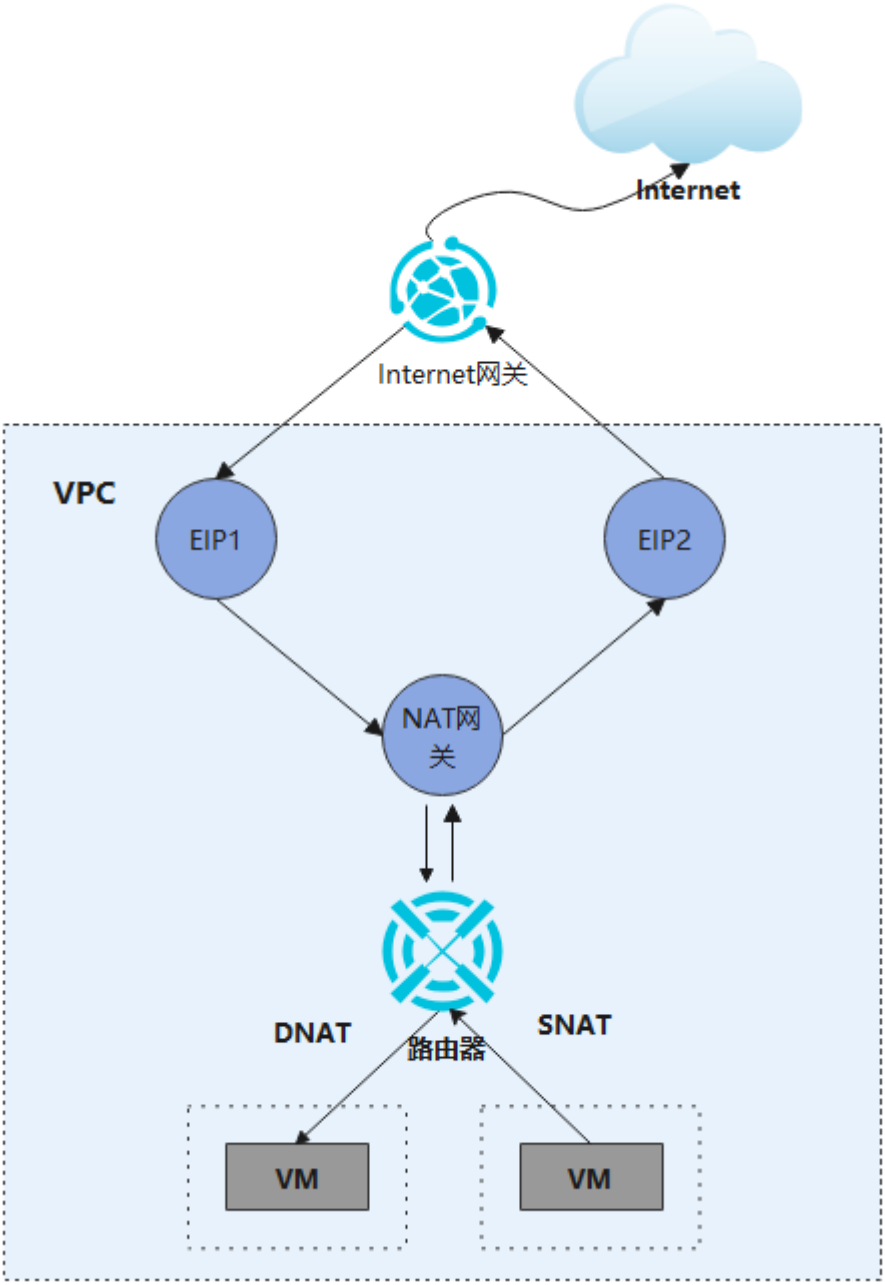
- 用户可以配置多个VPC，并管理VPC的路由表，每个VPC通过分布式路由器覆盖多个节点。
- 每个VPC内部可以通过IP网段划分出多个子网，子网内的虚拟机具备相同的网络和安全属性（网络ACL、VLAN、Underlay、QoS等），为方便IP地址管理，用户可以启用子网DHCP，对子网内虚拟机进行IP地址的动态分配。
- 子网内虚拟机内部交互和外部交互的基本方式由分布式路由器提供，同时可以根据业务需要为虚拟机分配EIP或通过NAT网关，VPN等方式对外通信。

4.9 NAT 网关

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可以用于实现多台虚拟机共享一个网关地址访问上层(路由的外层)网络。

 版权所有 © 超聚变数字技术股份有限公司



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。

4.10 弹性 IP

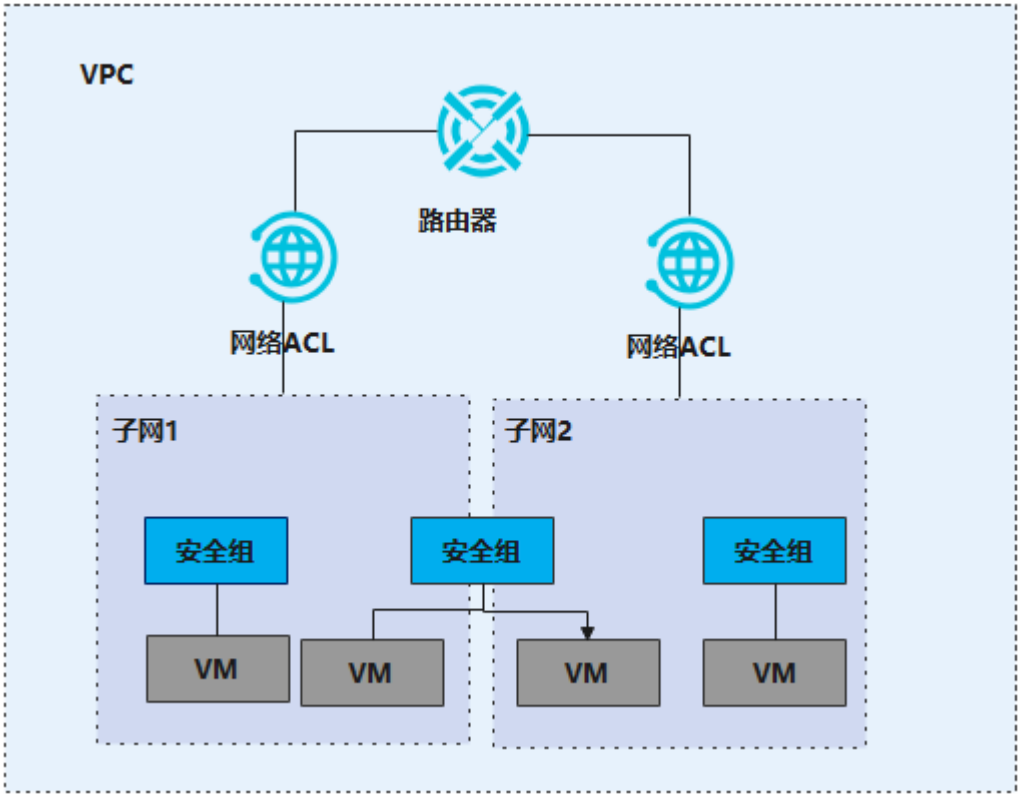
弹性公网IP（Elastic IP Address）提供独立的公网IP 资源，VPC子网中的云资源绑定EIP后可以与外界互相通信，其本质是一种NAT技术。

 版权所有 © 超聚变数字技术股份有限公司

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

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。



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。

表 4-2 网络 ACL 和安全组功能对比

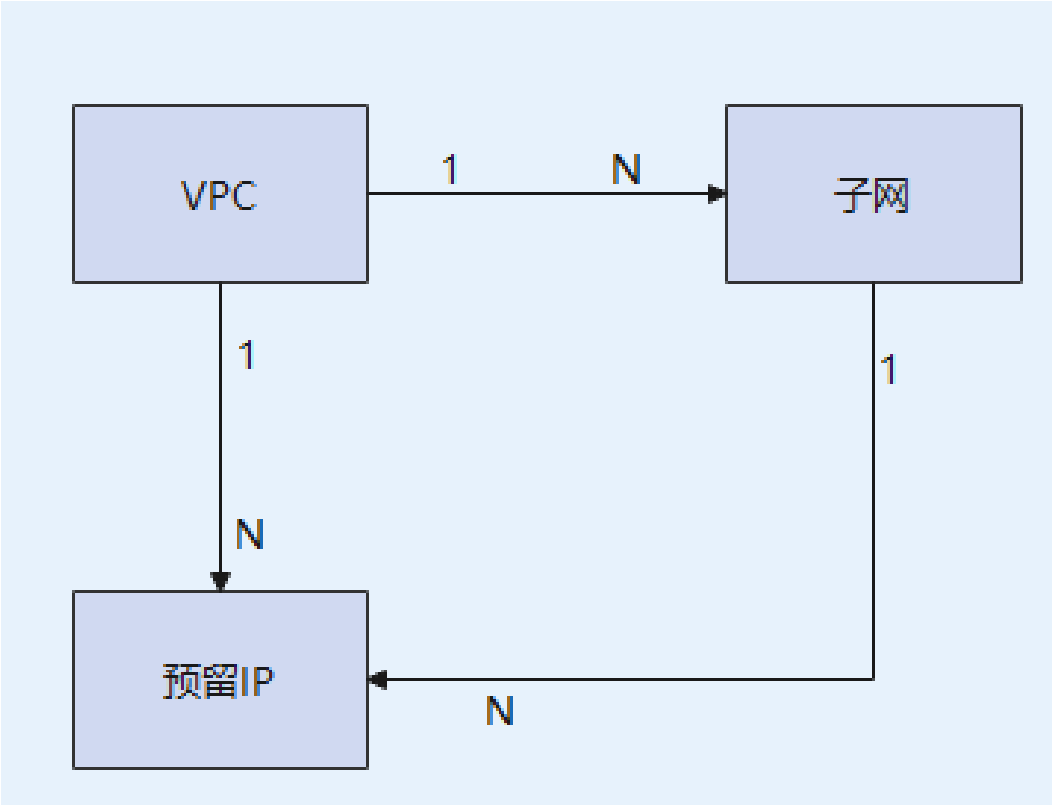
对比项	网络ACL	安全组
防护对象	子网级别：防护整个子网，子网内的所有虚拟机都会受到网络ACL的保护	虚拟机级别：防护安全组内的实例，比如虚拟机。
配置策略	网络ACL支持设置允许，拒绝。	安全组支持黑白名单，一个虚拟机不能同时关联黑白名单安全组

版权所有 © 超聚变数字技术股份有限公司

对比项	网络ACL	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。
应用操作	创建子网没有网络ACL选项，需要先创建网络ACL，添加出入规则，并在网络ACL内关联子网。	• 创建虚拟机时，可选择关联安全组，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

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，如当创建的多台虚拟机配置Keepalived作主备时，预留出的IP地址可以用作VIP管理多台虚拟机，这样可以防止预留的IP地址和创建的虚拟机IP地址冲突，同时也方便使用人员统一管理预留地址。

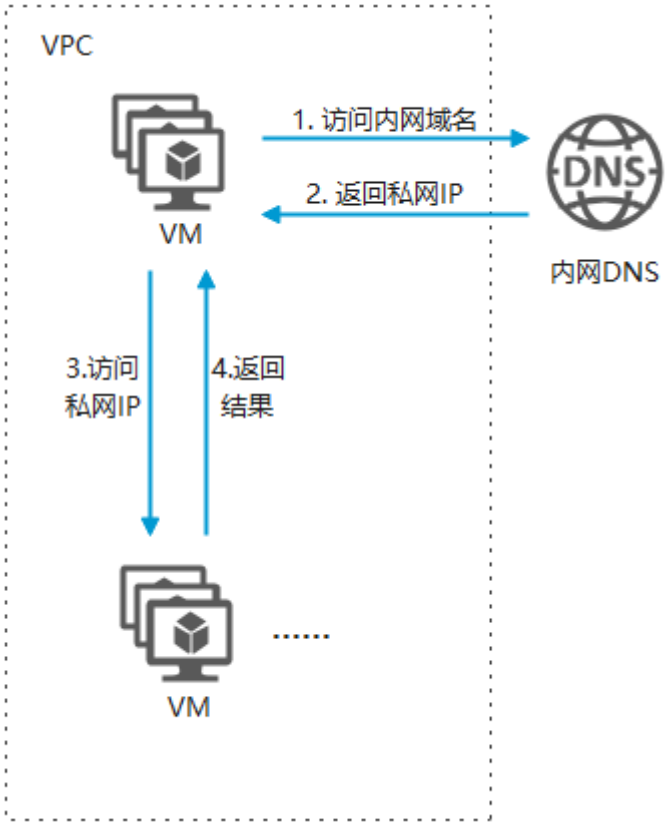


版权所有 © 超聚变数字技术股份有限公司

4.13 内网域名解析 DNS

内网域名解析是基于VPC网络的域名解析过程，通过内网DNS把域名（如xfusion.com）转换成私网IP地址（192.168.1.1）。内网域名解析实现虚拟机在VPC内直接通过内网域名互相访问。

当VPC内虚拟机访问内网域名时，内网DNS直接对内网域名进行解析，向虚拟机返回对应被访问的私网IP地址。



内网域名解析DNS提供的VPC内的内网域名解析服务，具有以下特点：

- 支持基于VPC任意定制内网域名，灵活自由。
- 一个域名可以关联多个VPC，方便统一管理部署。
- 提供VPC子网专用的内网DNS，直接响应内网域名，快速高效，有效防护劫持。

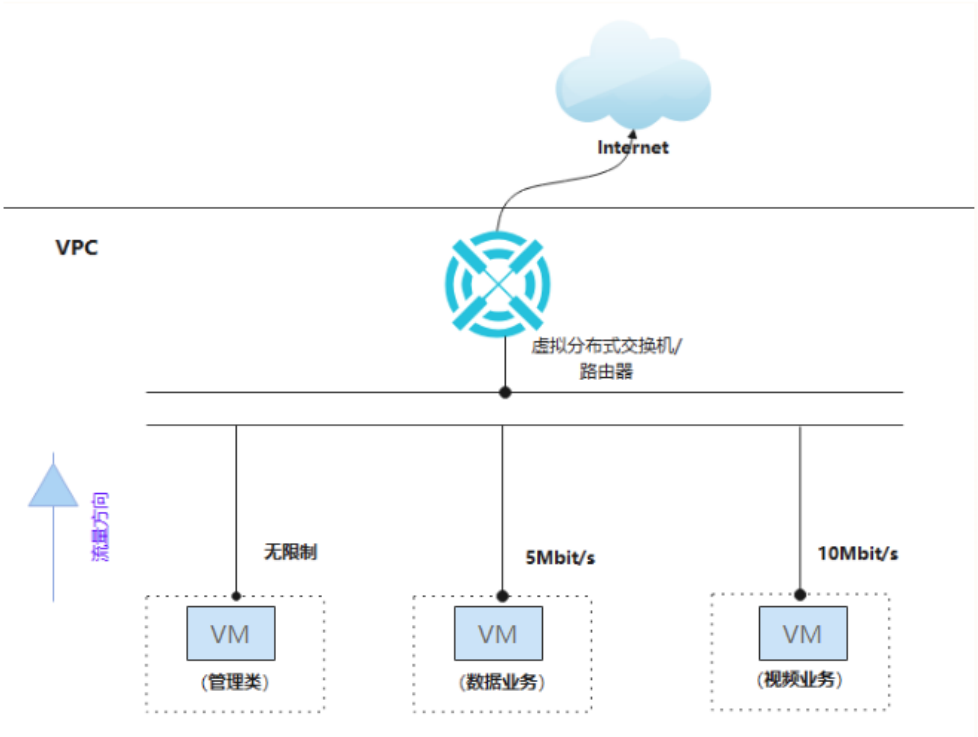
4.14 网络 QoS

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

版权所有 © 超聚变数字技术股份有限公司

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 。

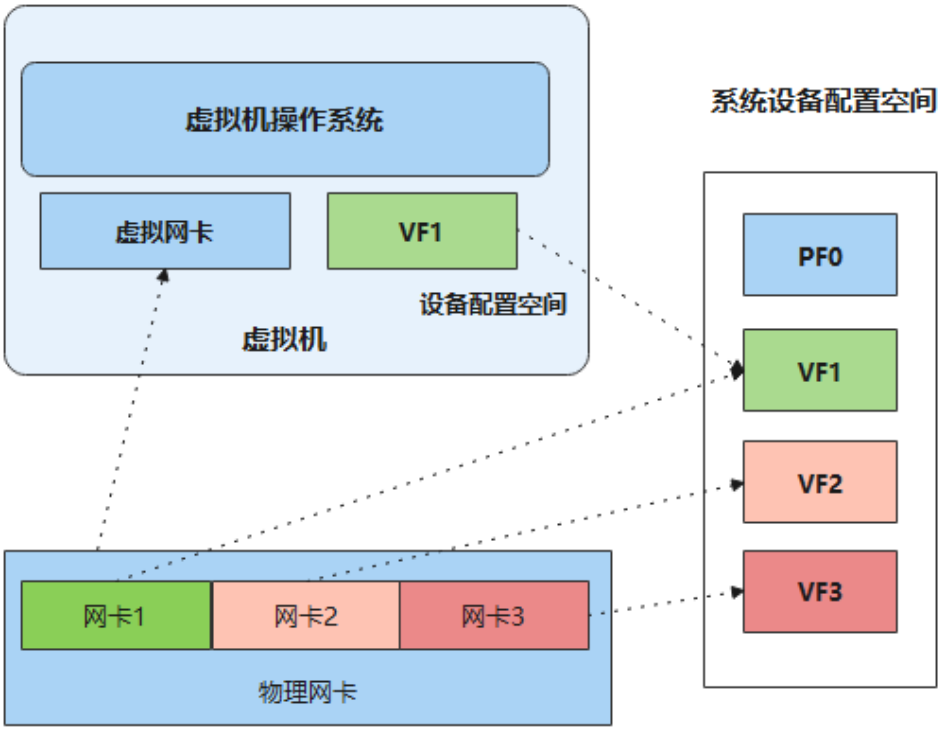


4.15 SR-IOV 交换

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用于分配给多个虚拟机，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 © 超聚变数字技术股份有限公司

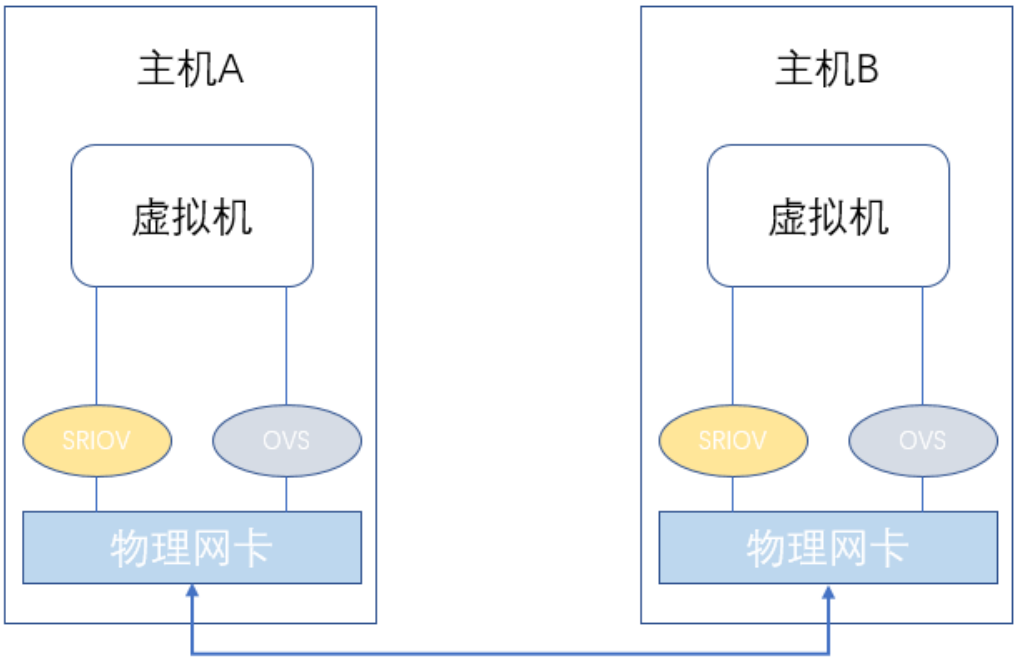


SR-IOV特性的使用需要硬件平台Intel VT-d、AMD IOMMU等技术的支持，在使用此功能时，相应特性必须在服务器的BIOS中被启用，如果BIOS中有SR-IOV的使能配置，那么也需要进行开启操作。

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，两个网卡组成fail over的主备网卡。

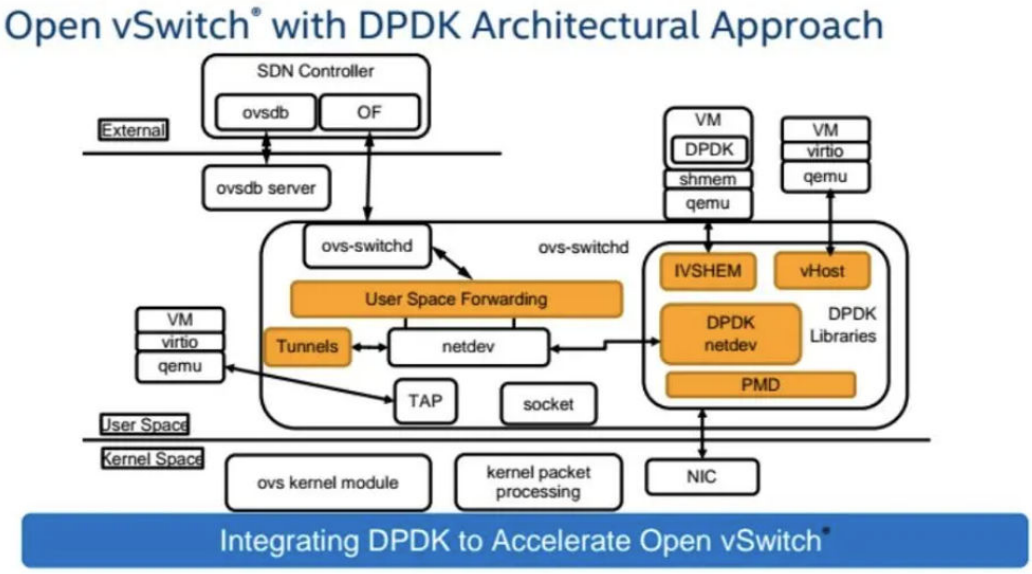
当触发启动热迁移时，虚拟化层会自动移除SRIOV网卡，此时业务网络流量自动切换到virtio网卡。虚拟机迁移到新的主机后，虚拟化平台自动还原SRIOV网卡，业务流重新切换到SRIOV网卡。该处理通过虚拟化平台自动把SRIOV网卡和virtio网卡规划到基于同一个物理网卡，实现底层自动配置，无需用户手动配置备用网络和选择备用网卡，提升简便性以及杜绝人为配置错误的问题。

 版权所有 © 超聚变数字技术股份有限公司



4.16 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，实现了高性能报文转发能力，提高虚拟机网络性能。



FusionOne Center 网络虚拟化支持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，包转发率及降低虚拟机之间的延迟和抖动。

 版权所有 © 超聚变数字技术股份有限公司

4.17 DHCP 服务

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。
- 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。

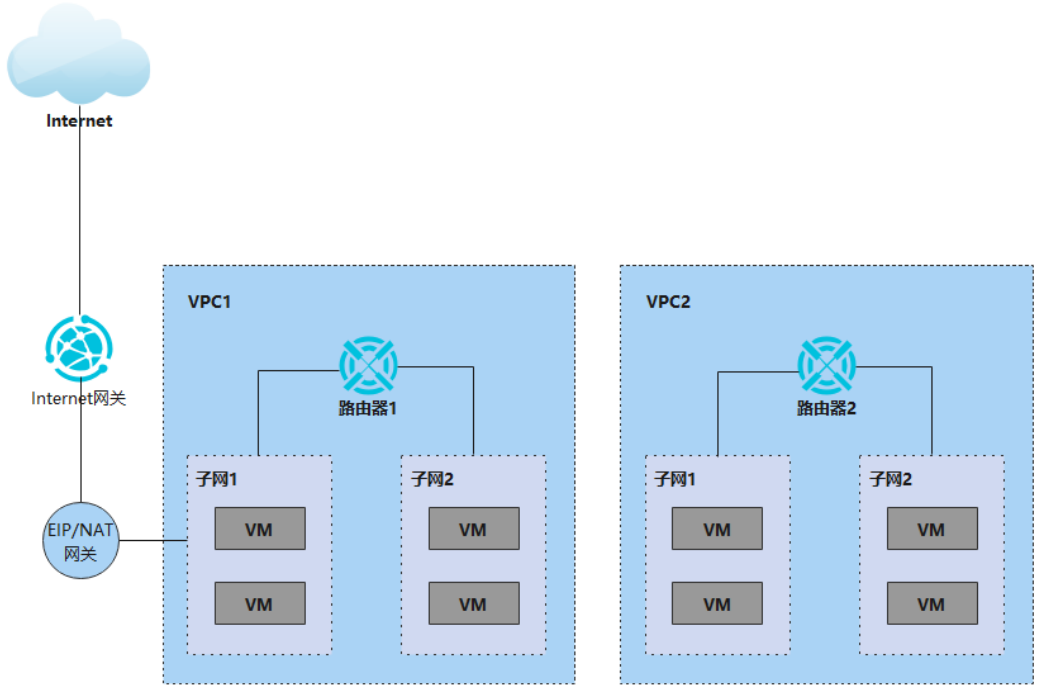
  相比手工静态配置IP地址，通过DHCP方式静态分配机制避免人工配置发生错误，方便管理员统一维护管理。

4.18 虚拟路由器

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虚拟机之间的流量转发，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，也提升了超融合虚拟化网络的故障备份的能力。

版权所有 © 超聚变数字技术股份有限公司



4.19 端口流量镜像

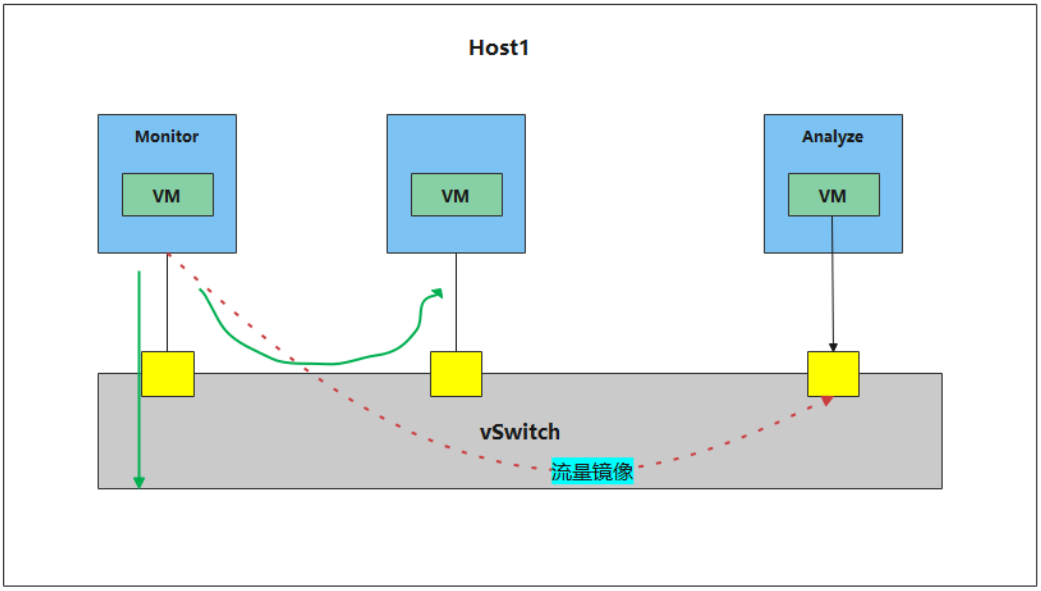
端口镜像是指通过将源端口的数据报文复制到与数据监测设备相连的目的端口，从而实现在目的端口上监测源端口的数据报文信息，用于数据分析和网络问题诊断。

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。

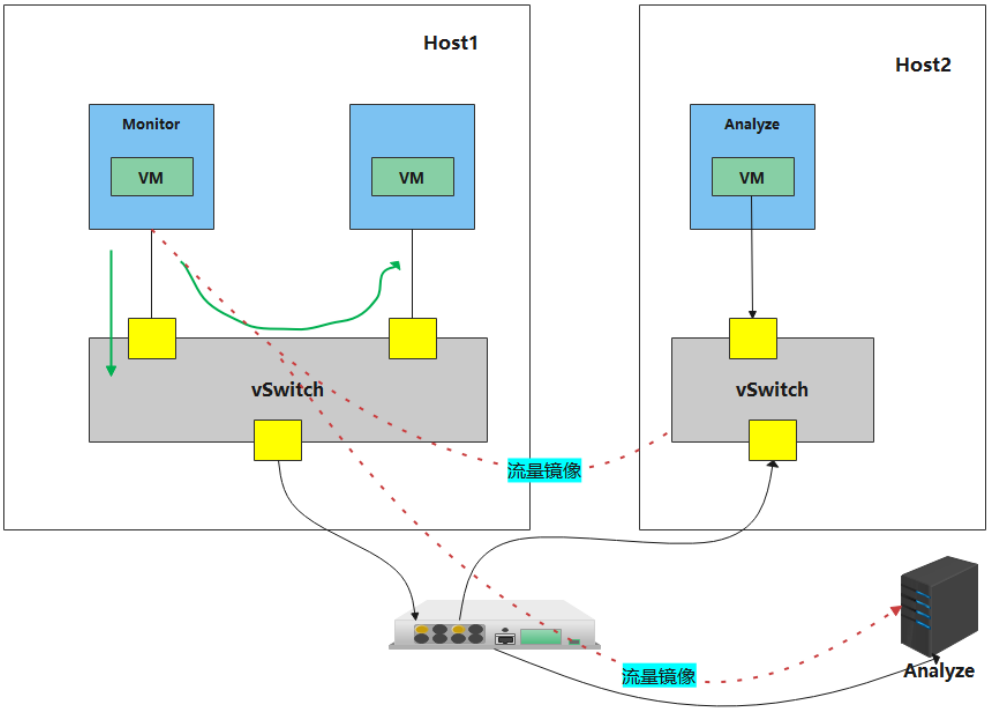
FusionOne Center提供的端口流量镜像功能主要包含本地端口镜像和远端端口镜像，镜像管理主要包括增加端口镜像，删除端口镜像，编辑端口镜像，显示端口镜像信息等功能。

本地镜像：

 版权所有 © 超聚变数字技术股份有限公司



远端镜像：



4.20 在线修改虚拟机 IP 地址

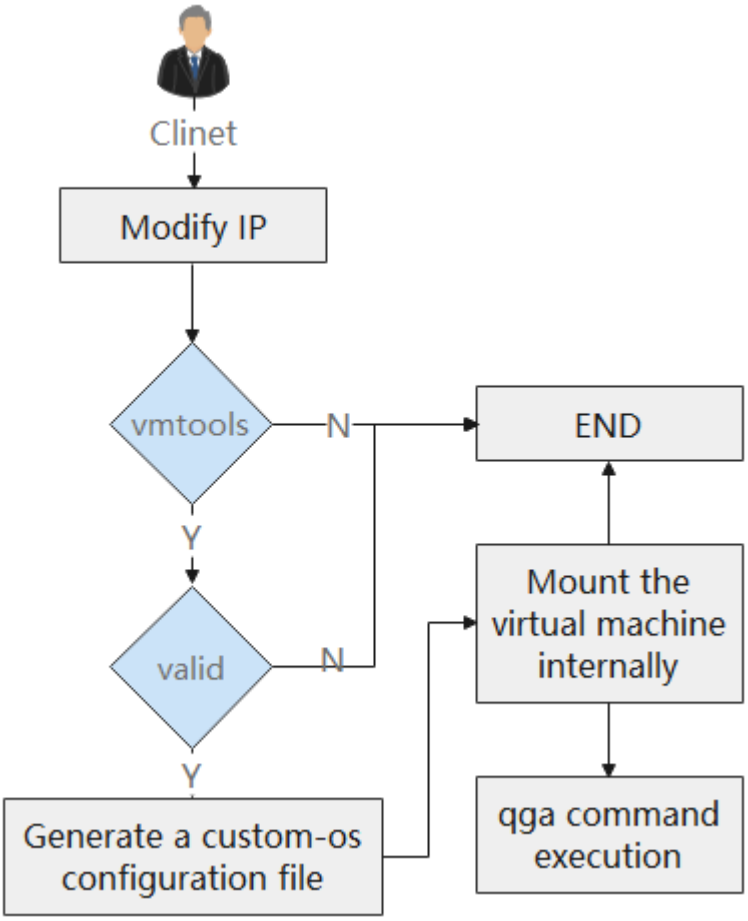
KVM虚拟机的获取的IP地址的方式目前主流的方式如cloud-init，vmtools等工具实现，但是如果要实现在线对虚拟机的网卡相关的配置进行配置，则需要一个在虚拟机开机状态下时，还能对虚拟机操作的工具，如vmtools，cloud-init由于在虚拟机开机完成后，服务程序会自动退出导致。

FusionOne HCI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。

 版权所有 © 超聚变数字技术股份有限公司

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。



4.21 虚拟负载均衡器

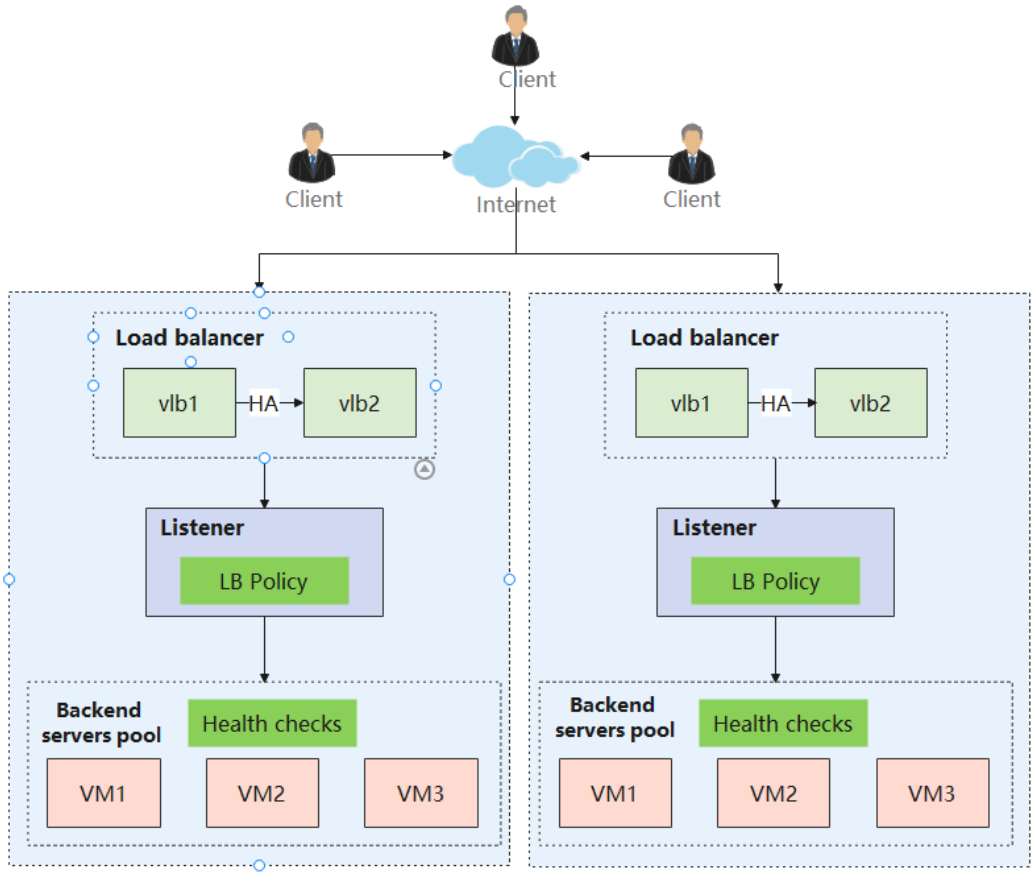
虚拟负载均衡器是将访问流量根据分配策略分发到后端多台服务器的流量分发控制服务。虚拟负载均衡可以通过流量分发扩展应用系统对外的服务能力，同时通过消除单点故障提升应用系统的可用性。

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。

 版权所有 © 超聚变数字技术股份有限公司



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

  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，用于检查后端虚拟机的运行状态，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。

 版权所有 © 超聚变数字技术股份有限公司

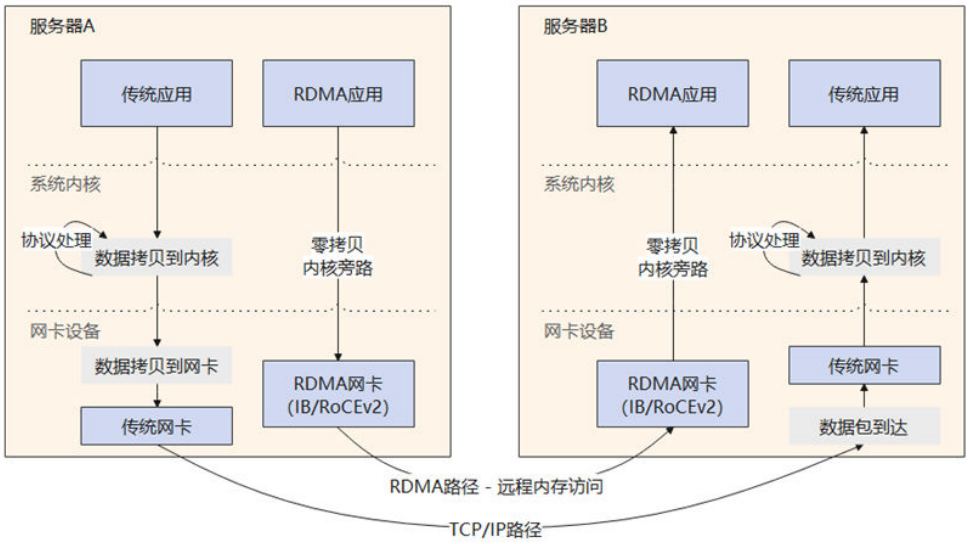
表 4-3 虚拟负载均衡的规格

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	4C20G
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）
后端服务器组	后端服务器组数量	30（个）
	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）

4.22 RDMA 技术

RDMA 是一种通过网络直接访问远程计算机内存的技术，它通过“零拷贝”、“内核旁路”和“无CPU参与”三大核心技术，实现了超低延迟、高吞吐量和极低CPU占用的远程数据通信。

 版权所有 © 超聚变数字技术股份有限公司



RDMA的典型应用场景：

- **AI训练&推理：** 在GPU服务器集群间高速同步模型参数和梯度，大幅缩短训练和推理时间。
- **分布式存储系统(NVMe-Of)：**通过RDMA网络提供接近本地NVMe SSD性能的远程块存储访问。
- **云计算与数据中心：** 使用RDMA来提升效率和降低延迟。
- **金融行业：** RDMA低延迟高性能，在高频交易系统中，微秒级的延迟差异都至关重要。

超融合通过网络功能虚拟化技术管理InfiniBand及RoCEv2的网卡设备，支持普通网卡切换为InfiniBand和RoCEv2模式，在RoCEv2模式下可配置pfc，ecn等参数，用户可以灵活的实现RDMA组网需求。

 版权所有 © 超聚变数字技术股份有限公司

5 存储虚拟化

5.1 分布式存储
5.2 虚拟机直通SAN
5.3 虚拟文件系统
5.4 磁盘模式配置
5.5 自动精简配置
5.6 RBD卷
5.7 NoF卷
5.8 虚拟机镜像仓库
5.9 统一存储
5.10 块存储特性
5.11 NAS文件存储特性
5.12 对象存储特性
5.13 虚拟机备份
5.14 分布式存储内部机制
5.15 存储高性能
5.16 EC块存储

5.1 分布式存储

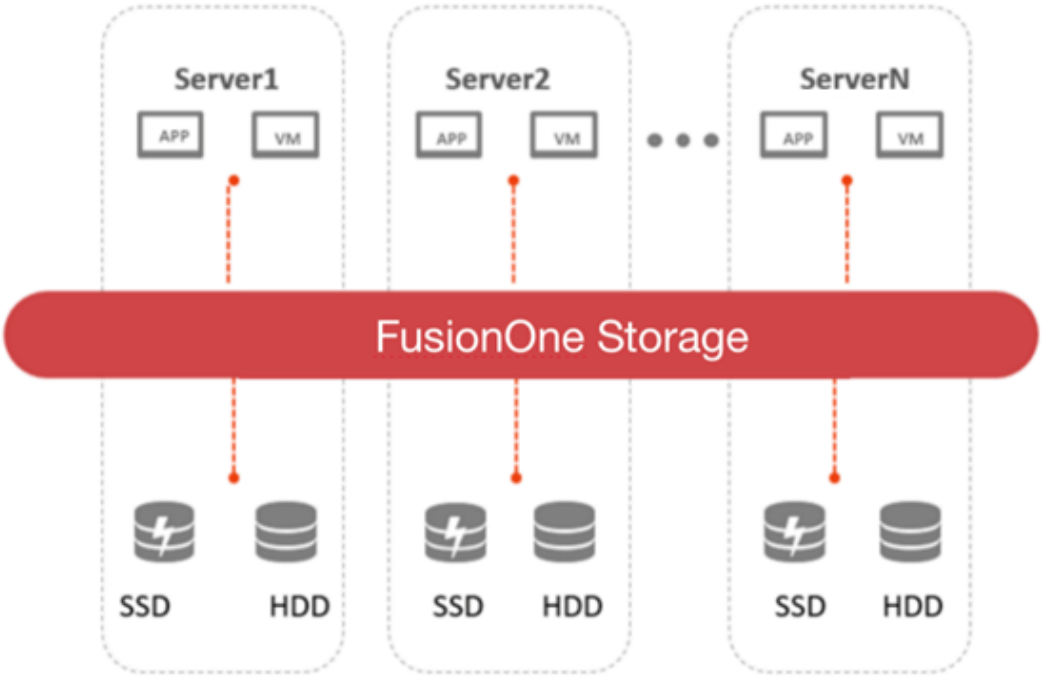
存储虚拟化，基于分布式存储技术，将HDD、SSD等存储资源组织成分布式存储资源池，为上层虚拟机和其它应用提供统一存储服务，消除资源组织模式形成资源孤岛、硬件资源利用不均、整体资源浪费等问题。

分布式存储设计理念

分布式存储组件FusionOne Storage，采用业界先进的全分布式和Share Nothing架构设计理念，采用基于策略的分布式哈希表数据路由算法，通过计算就能直接寻址到数

 版权所有 © 超聚变数字技术股份有限公司

据所在的存储节点，无需查找元数据节点，大大缩短了数据IO访问路径，提升了系统性能。同时，整个存储系统也无集中管理和控制节点，每一个数据节点都有能力承担任何数据节点的功能，节点之间通过内部高效的分布式协议完·成相互协作和通信。这种去中心化、无状态的全分布式数据处理架构是存储系统能实现水平、线性扩展能力的关键，有力保证了整个存储系统无单点故障、无性能瓶颈。这样的设计使得用户首次购买或者扩容时，不需要考虑独立的存储元数据节点或存储网管节点等问题，只需根据实际存储需求计算出需要的节点数即可。



- 块存储服务：提供RBD和iSCSI标准接口，支持广泛的虚拟化平台及数据库应用，提供高性能与高扩展能力，满足虚拟化、备份、视频监控、大数据等场景的存储需求。
- 对象存储服务：提供Amazon S3等标准API，支持主流云计算生态，满足内容存储、云备份、云归档及公有云存储服务运营场景需求。
- 文件存储服务：提供NFS和CIFS标准接口，支持视频、音频、高性能计算、视频监控等多业务场景。

分布式存储功能架构

分布式存储FusionOne Storage基于全分布式存储架构，通过软件层面的全分布式架构和数据冗余技术，来达到高可伸缩性和高可用性，性能、容量随节点数增加而线性增加。单集群最高可支持扩展至4096个节点。支持在单个集群内按磁盘维度划分多个硬盘池，支持图形化界面划分存储资源池。

版权所有 © 超聚变数字技术股份有限公司



- 存储接口层：提供标准iSCSI、SCSI协议的块设备接口或应用态块接口向应用提供卷设备访问功能。支持S3接口，实现文件的HTTP/HTTPS形式访问，也可以利用不同语言的SDK简化S3接口的访问。同时支持NFS、CIFS等传统文件访问接口。
- 存储服务层：提供各种存储高级特性，如块存储快照、精简配置、SSD加速、硬件感知框架，对象存储的生命周期、用户配额、负载均衡等功能，文件存储的共享配置、AD域、跨协议等功能。
- 存储引擎层：提供分布式存储基本功能，包括集群状态管理、分布式数据路由、强一致性复制协议，及在节点故障时集群故障自愈与并行数据重建子系统等。
- 存储硬件层：支持SAS、SATA、NVMe。为了对硬件进行控制和故障管理，实现了磁盘控制卡管理和磁盘错误管理功能。
- 存储管理层：实现存储可视化安装部署、可视化配置、在线升级、软硬件系统监控、告警日志等运维管理功能，为用户提供Web GUI管理界面和命令行管理系统。面向运维、运营、审计、业务等IT角色提供丰富的企业级特性，提供简单、易用、内容丰富的图形化管理界面，包括仪表盘、硬件监控、多数据中心管理、性能指标统计、软硬件告警及展示等功能特性，支持升级、扩容、配置等变更操作功能特性。对运营、业务、审计方提供面向角色的业务访问精细统计报表，运营数据等。

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。

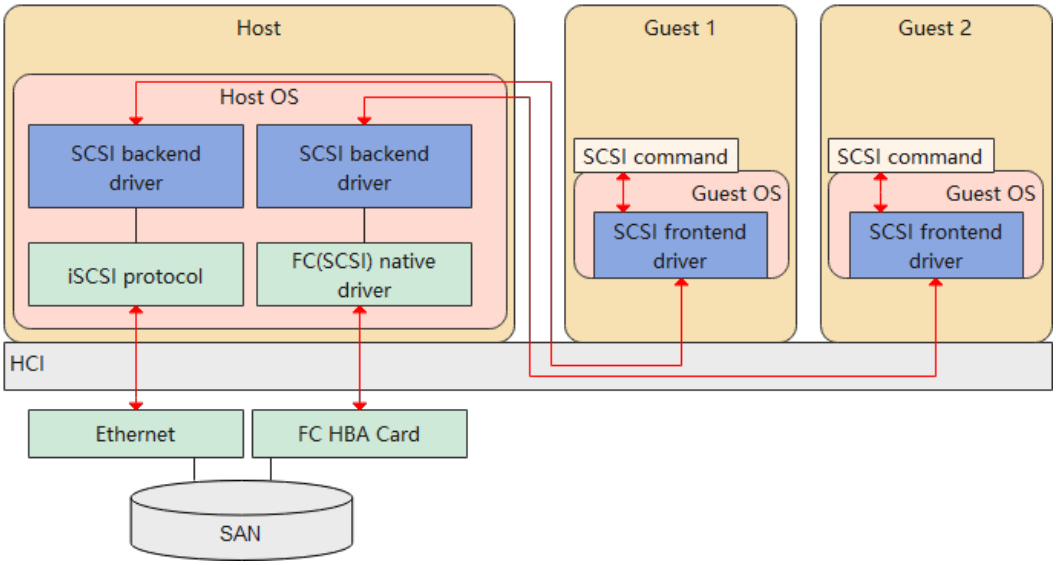
- 原理

  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

 版权所有 © 超聚变数字技术股份有限公司

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：



- 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

- 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 HCI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。

- 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：

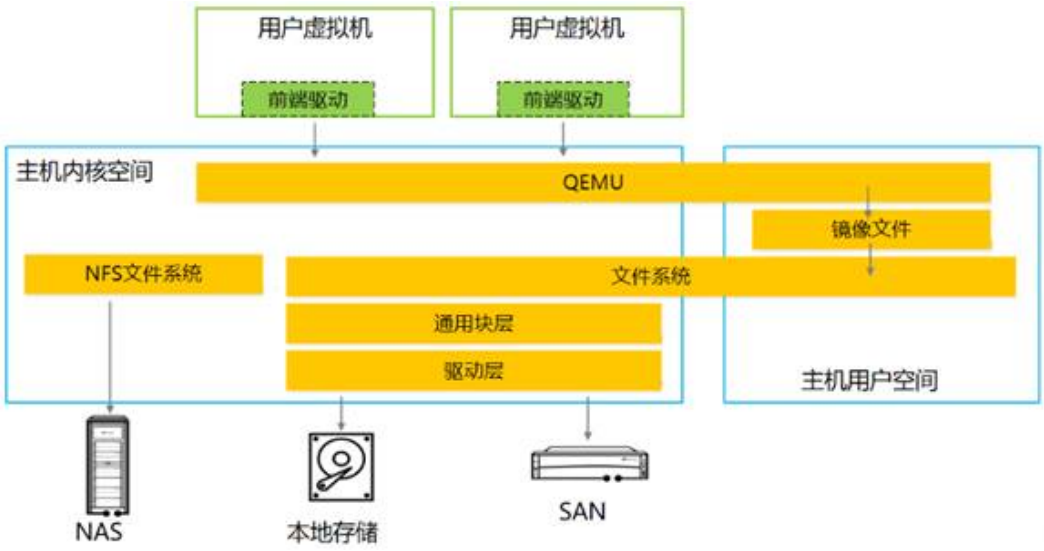
 版权所有 © 超聚变数字技术股份有限公司

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 HCI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

  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 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。

 版权所有 © 超聚变数字技术股份有限公司

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

  writeback：数据写入主机缓存后即返回成功，异步下刷数据到磁盘。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。

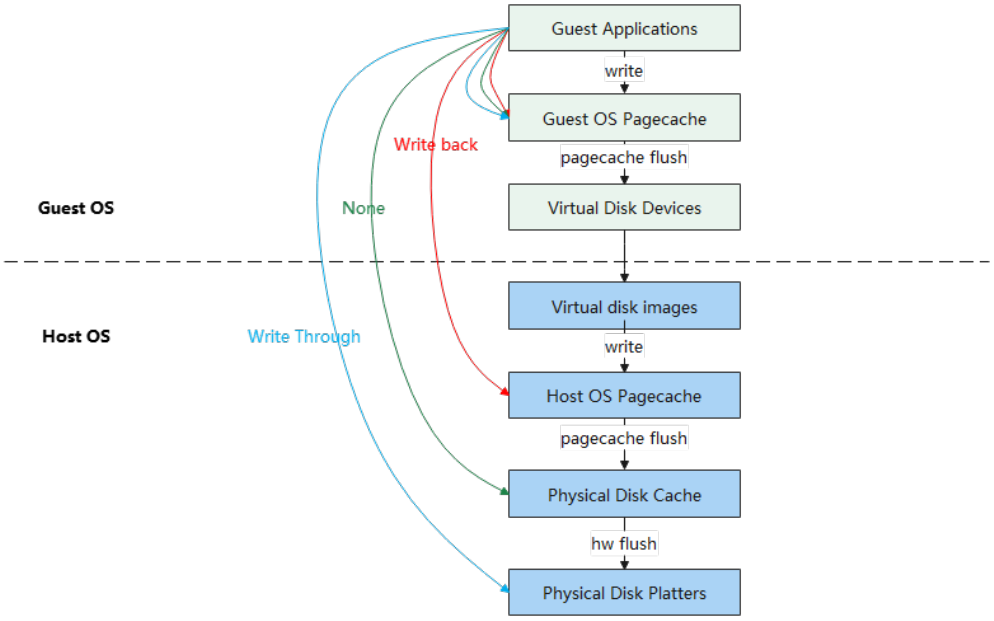
  none：关闭主机缓存，数据写入物理磁盘。如果物理磁盘开启了缓存，那么数据写入物理磁盘缓存。

  writethrough：关闭虚拟磁盘缓存，数据同时写入主机缓存和物理磁盘，确保数据安全。

  directsync：关闭主机和虚拟磁盘缓存，直接读/写物理磁盘，每次写操作都要等待数据完全写入物理磁盘后才返回成功，确保数据安全，但会导致性能下降。

  如下图所示：

 版权所有 © 超聚变数字技术股份有限公司



5.5 自动精简配置

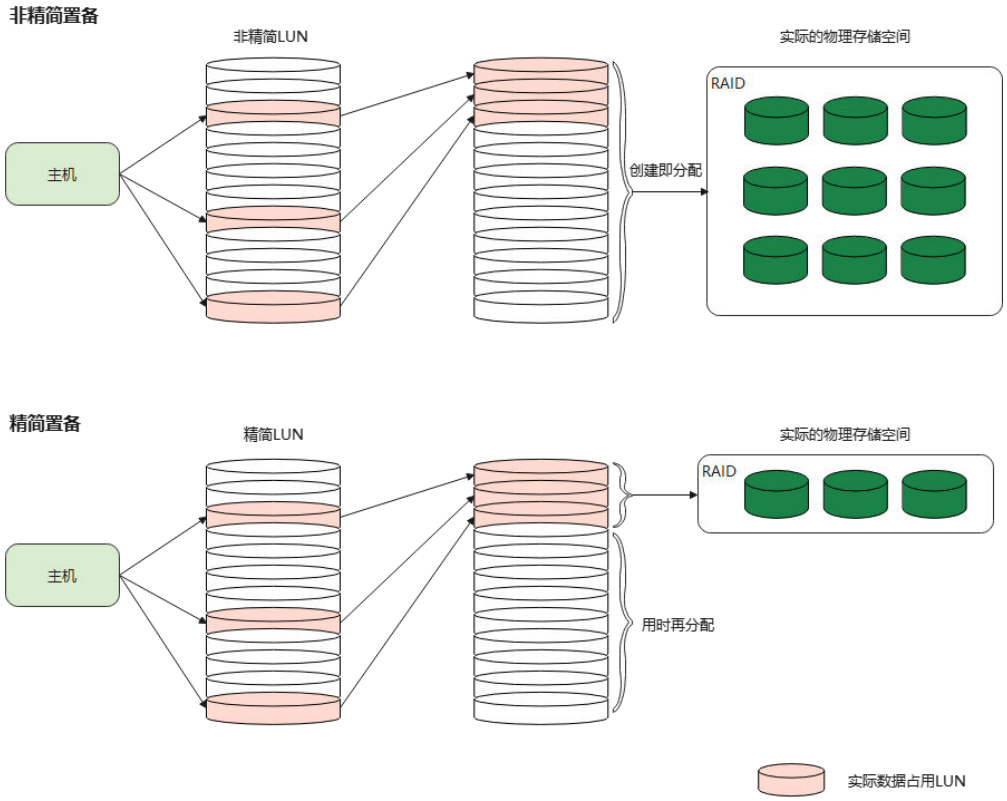
- 概述

  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虚拟化的存储自动精简配置不依赖于存储设备。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

  分布式块存储卷默认自动精简配置，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 © 超聚变数字技术股份有限公司



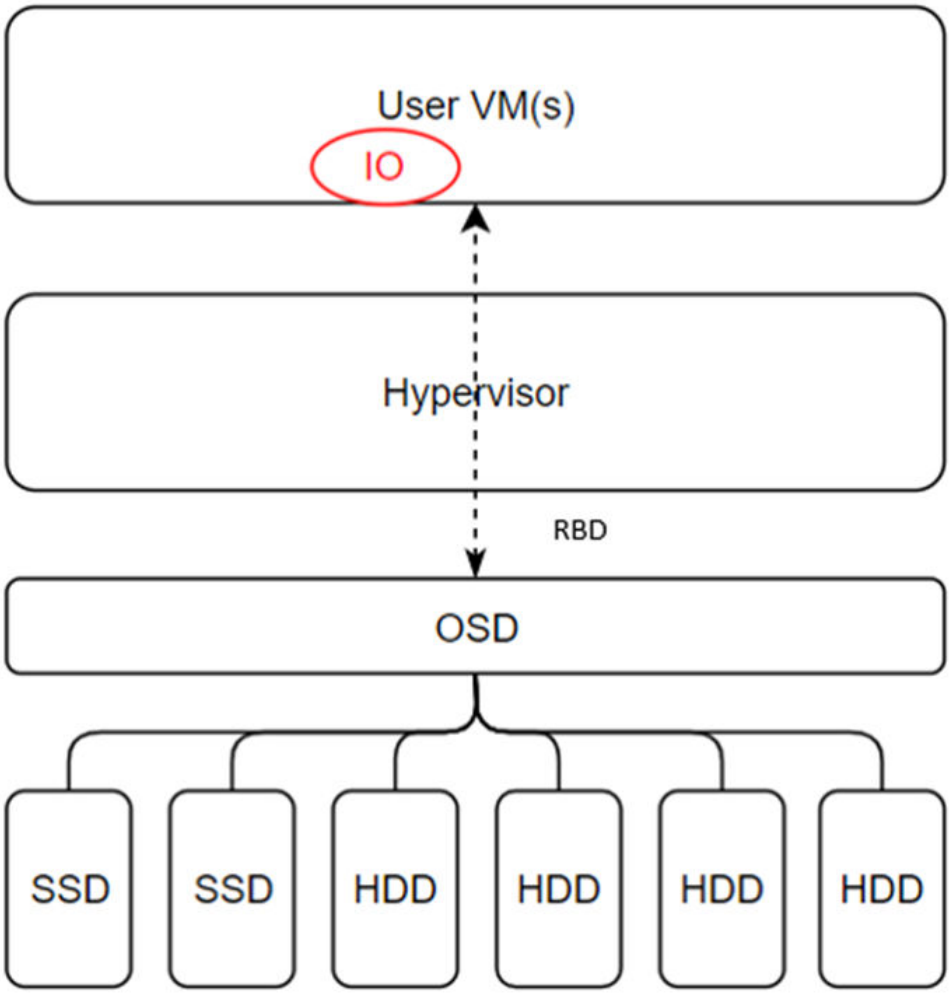
- 应用场景

  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。

5.6 RBD 卷

业务虚拟机部署在主机KVM虚拟化平台上，虚拟机所使用的分布式存储是通过RBD协议访问，虚拟机IO通过SCSI协议，直接与运行在Hypervisor内核中的分布式存储软件进行数据交互，详细业务IO流程参考下图所示：

 版权所有 © 超聚变数字技术股份有限公司



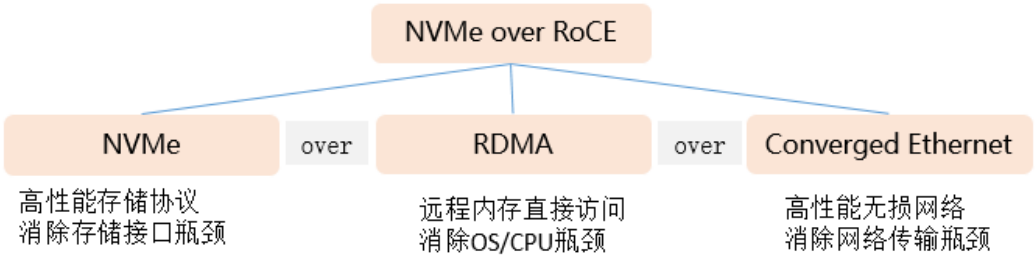
相比iSCSI对接，RBD本地卷访问的网络路径短，IO效率更高。

5.7 NoF 卷

NoF（NVMe over Fabrics）是一种将本地NVMe协议扩展到网络存储的技术，旨在提供高吞吐、低延迟的远程存储访问。

- NVMe是基于PCIe的存储协议，相比传统iSCSI协议，能够更好发挥NVMe硬盘性能。
- Fabrics是底层网络架构，支持FC、InfiniBand、RoCE v2、iWARP和TCP等各种网络协议。综合考虑性能和成本，RoCE v2使用比较广泛，即RDMA over Converged Ethernet，只需要以太网交换机和网卡就可以实现RDMA传输，性能接近InfiniBand。

 版权所有 © 超聚变数字技术股份有限公司



基于NVMe over RoCE实现的NoF卷，消除了IO路径中的存储瓶颈、网络瓶颈和OS/CPU瓶颈，为虚拟机提供高性能存储服务。

5.8 虚拟机镜像仓库

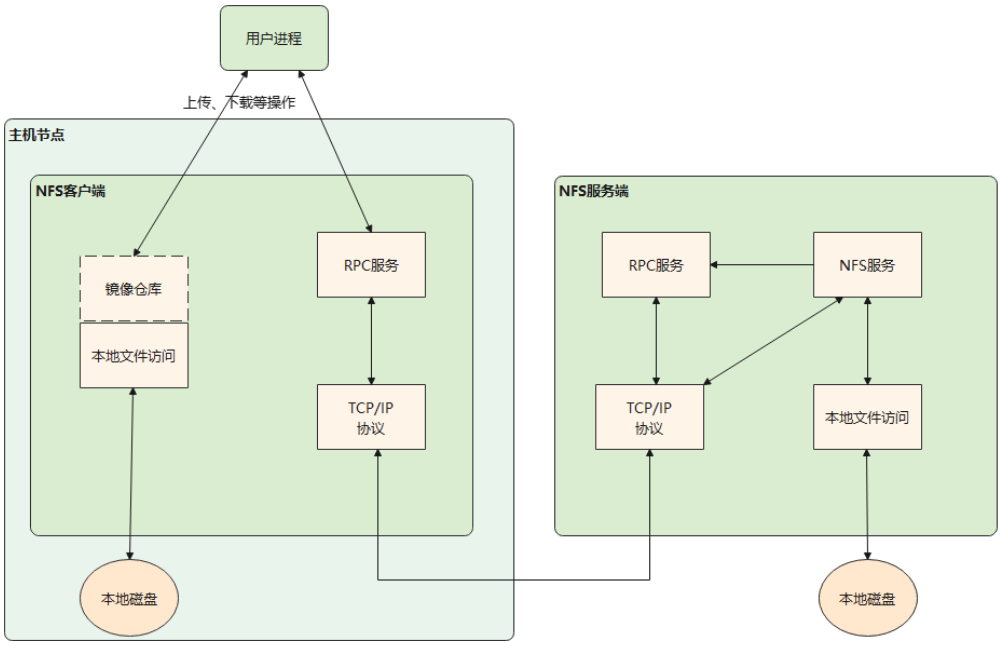
- 概述

  超融合默认提供了1TB容量的NAS共享空间用于保存用户的镜像文件，并提供了便捷的虚拟机镜像自助管理功能，用户可以灵活便捷的使用操作系统镜像来创建虚拟机，保证HCI虚拟化使用镜像的一致性。主要包含如下功能点：

  - 镜像文件生命周期管理：支持镜像文件的上传、删除、下载等，最大支持30GB镜像文件。
  - 镜像仓库文件夹管理：支持创建最多5级镜像文件夹，支持镜像跨文件夹目录移动、跨站点克隆和重命名。

- 原理

  镜像仓库是通过NFS协议共享空间来实现镜像文件的管理功能，虚拟机可以挂载镜像文件来进行使用，后台原理如下图所示：



- 应用场景

  用户需要将ISO文件挂载给虚拟机，用于安置虚拟机的操作系统或业务软件时，可以提前将文件上传到镜像仓库后挂载光驱时选择镜像仓库方式。

版权所有 © 超聚变数字技术股份有限公司

5.9 统一存储

- 概述

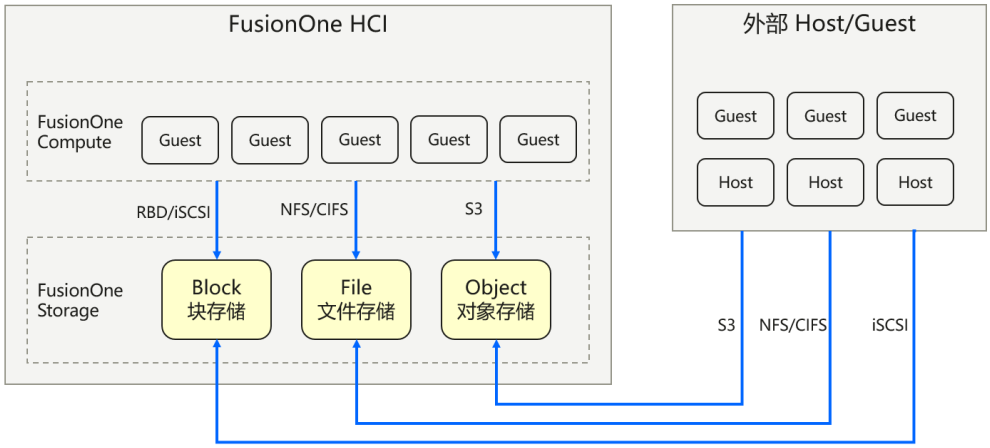
  超融合提供统一存储能力，即同时对外提供块存储、NAS文件存储和对象存储。

  不同的业务可以使用不同的存储，充分共享存储空间、提高空间利用率；同时可以减少部署、管理和运维工作量。

- 原理

  基于同一个物理存储资源池，FusionOne Storage可以同时创建不同类型的逻辑存储池，包括块存储、NAS文件存储、对象存储。

  块存储支持RBD/iSCSI接口，NAS文件存储支持NFS/CIFS接口，对象存储支持S3接口。



- 应用场景

  HCI为内部Guest添加硬盘，会通过RBD对接块存储。

  HCI内部Guest中的应用程序，可以通过iSCSI访问块存储、NFS/CIFS访问文件存储、S3访问对象存储。

  外部Host/Guest中的应用程序，可以通过iSCSI访问块存储、NFS/CIFS访问文件存储、S3访问对象存储。

5.10 块存储特性

5.10.1 快照

FusionOne Storage提供了快照机制，将用户的卷数据在某个时间点的状态保存下来，后续可以作为导出数据、恢复数据之用。FusionOne Storage快照数据在存储时采用ROW（Redirect-On-Write）机制，快照对主机业务性能影响小于5%。

ROW技术原理，请参见“秒级快照”。

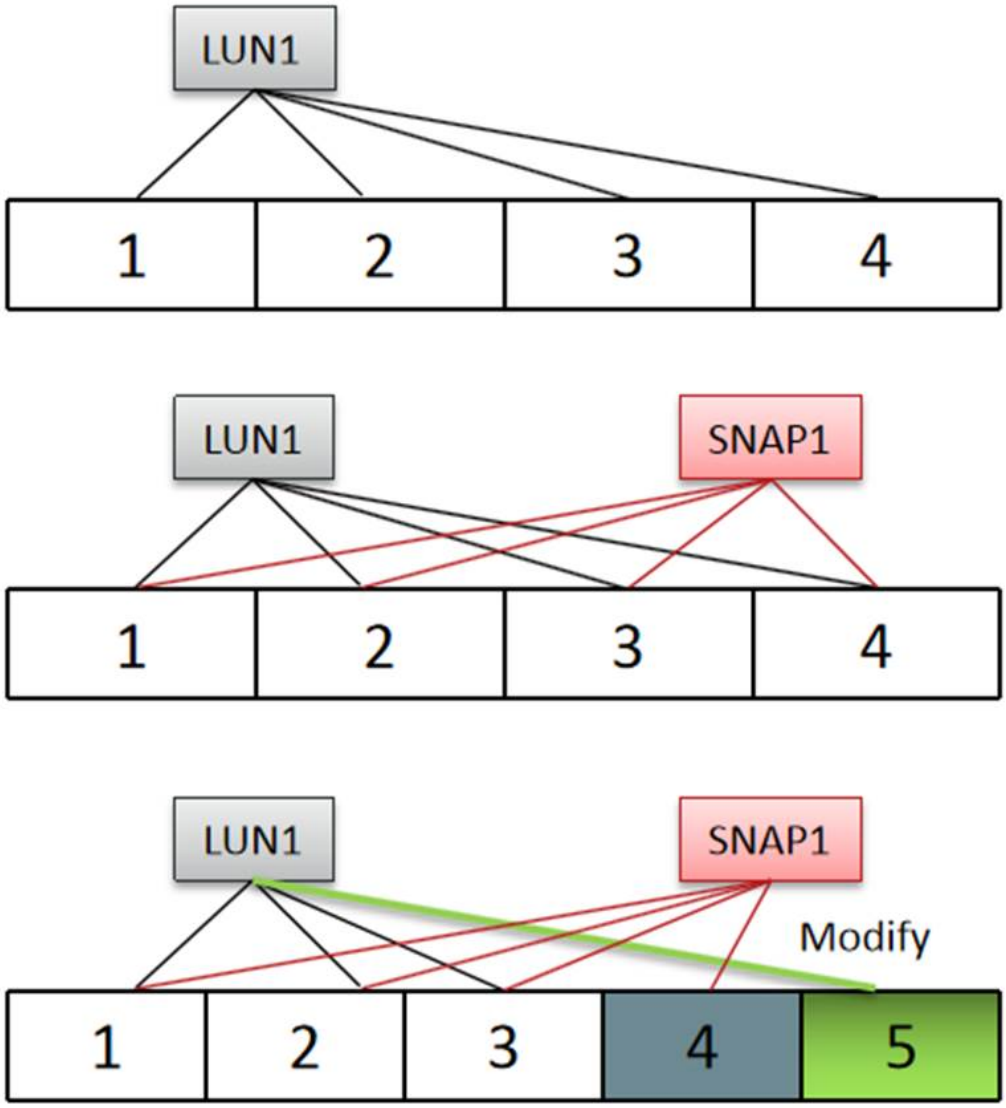
秒级快照

FusionOne Storage快照采用ROW（Redirect-On-Write）机制，做快照时生成对应快照数据块链接指向对应数据。当有数据写入时，把对应数据写入另外一个位置，然后修改原卷数据块链接到新位置。ROW写的时候性能基本没有损耗，只是修改指针链

版权所有 © 超聚变数字技术股份有限公司

接，实现效率很高，不会引起原卷性能下降。通过一次查找操作即可判断是否做过快照，以及最新快照的存储位置，保证业务效率。

图 5-1 快照生成示意图

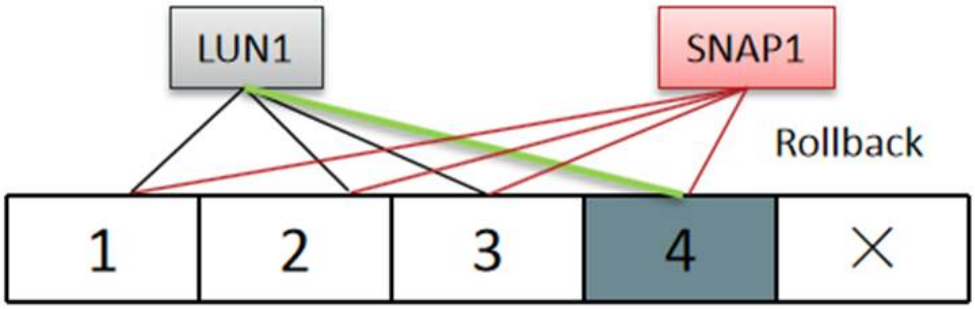


快照回滚

快照回滚操作在ROW机制中，只需要将当前数据块链接重新指回快照数据即可，回滚速度很快，用户可以短时间内将任何一个存储卷回滚到任何一个时间点的快照。回滚后当前数据不可恢复。

版权所有 © 超聚变数字技术股份有限公司

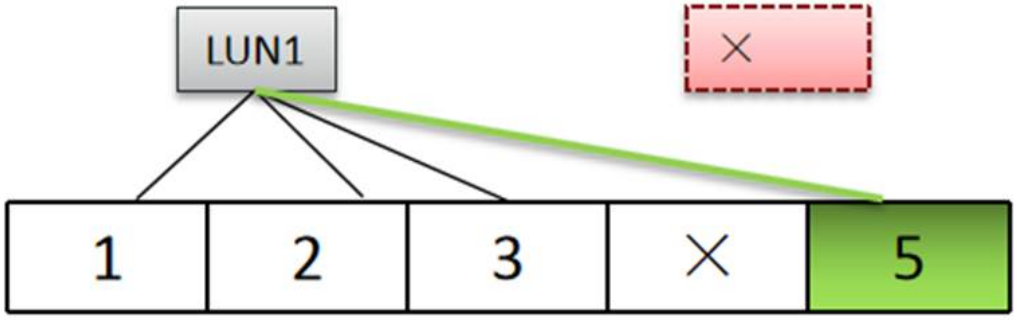
图 5-2 快照回滚示意图



快照删除

快照删除操作在ROW机制中，只需要删除对应快照的链接，相应数据块如果没有其他资源（原卷、快照、克隆卷等）引用才可清除，用户可以删除任何一个时间点的快照，而不影响后续快照，方便快照管理。

图 5-3 快照删除示意图



FusionOne Storage提供了秒级快照机制，将用户的逻辑卷数据在某个时间点的状态保存下来，作为备份恢复数据之用。单卷最大快照个数可以达到256，完全满足正常需求。

5.10.2 LUN Masking

FusionOne Storage通过LUN Masking功能将各个硬盘池和逻辑池的LUN资源隔离开，确保未授权的用户无法扫描到指定的LUN。

5.10.3 快速离线

FusionOne Storage提供非常强大的数据保护机制，当集群中有节点或者磁盘故障时，存储管理系统会及时将其踢出集群，并可根据业务需求设定重构和均衡时间，既减小对在线业务的性能影响，也可降低数据安全风险，同时兼顾了存储的性能和安全性。

5.10.4 卷级 QoS

FusionOne Storage块存储QoS提供了对卷的IOPS、带宽进行精细化控制，当卷的需求超出了基准IOPS或带宽时，允许在一定时间内使用超出基准性能的配额。

版权所有 © 超聚变数字技术股份有限公司

5.10.5 精简配置

FusionOne Storage块存储提供了精简配置功能，为应用提供比实际物理存储更多的虚拟存储资源。相比直接分配物理存储资源，可以显著提高存储空间利用率。

5.10.6 卷克隆

Fusion Storage块存储提供链接克隆机制，支持从卷快照创建链接克隆卷，支持将链接克隆的卷拍平操作，即从原始的快照信息进行分离。

5.10.7 卷独占

卷独占保证分布式锁的状态机不会异常、业务只能从一个客户端下发。保护逻辑卷数据一致性，对逻辑卷做修改时，需要持有此锁。

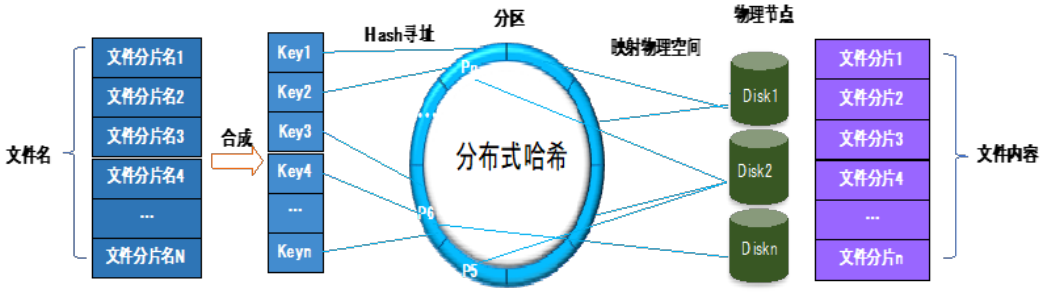
5.10.8 DHT 分布式哈希技术

技术原理

FusionOne Storage采用去中心化方式实现数据存储的管理。DHT分布式哈希算法建立一个虚拟的分布式哈希环，均匀地将整个哈希空间分成数千等分，然后将哈希空间的每一个分区均匀地映射到集群的所有磁盘上，确保每一个磁盘所负责的哈希分片数量是一致的。

数据按对象读写时，会将文件按照切片规则（默认4MB）大小生成每个切片的文件名列表（一般是Bucket+文件名+分片编号），然后将文件分片名通过哈希计算，生成一个哈希后的数值，利用哈希值在DHT表中进行查找，从而定位到该分片的磁盘位置获取文件分片的内容。通过组合每个分片，形成整个文件的最终内容。

从文件的定位过程可以发现，哈希计算和DHT表的查找时间是恒定不变的，不会因为文件数量的增加而影响效率，从而保证了在海量文件数量的情况下的访问性能。



- 通过DHT分布式哈希算法，分布式存储管理者可以通过配置信息指定集群数据分布方式（机架级别和服务器级别）。例如服务器级别下，数据默认2副本时，通过DHT将数据的主从副本分别存放在不同服务器的硬盘上，也就意味着当有一台服务器故障时，数据不会丢失。
- **使用DHT分布式哈希算法，数据均衡分布，能够充分利用磁盘的容量和性能，因此读取和写入都更加高效。**
- **抗故障性高。由于非中心化设计，不存在特定服务器异常而导致业务暂停。**
- **伸缩性好。在扩容或故障时，DHT会自动均衡变化的存储资源**，尽量不重构数据以保证业务性能。

版权所有 © 超聚变数字技术股份有限公司

客户价值

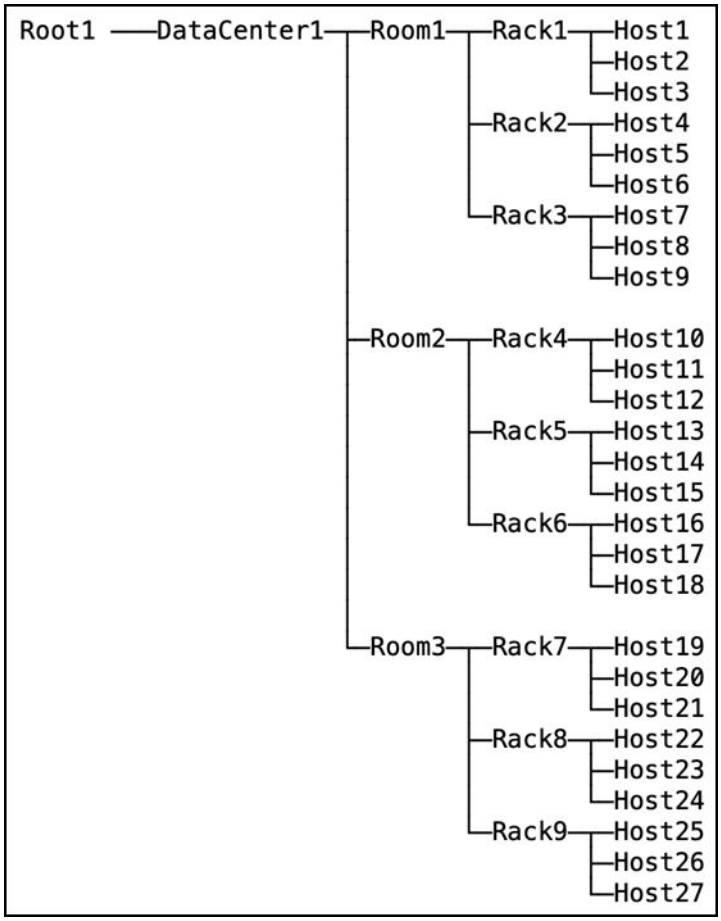
- 与集中式存储的raid机制不同，FusionOne Storage的DHT技术使得数据在数据盘均匀分布，提高存储性能和可用性。
- 支持机架级别和服务器级别的数据配置方式，在各类服务器集群部署下均能保证数据可靠性。
- 与集中式存储扩容涉及到PCIe插槽等硬件限制不同，FusionOne Storage支持线性扩容，易于服务器集群扩容或缩容，方便存储资源合理变更。

5.10.9 多硬盘池技术

FusionOne Storage存储系统，能够在同一套存储系统中，划分多个硬盘池，通过存储策略的控制，将应用业务分布到不同的硬盘池中，实现故障隔离，控制故障的影响范围。

技术原理

集群拓扑是多硬盘池技术的关键内容。集群拓扑在逻辑上维护了一份集群节点的分布拓扑图，其中包含了数据中心、机房、机架、服务器节点等从大到小的逻辑实体，通过树状图的方式逐层放置，从而描述出从机房到服务器节点的分布关系。其中，每一个树状图都有一个“根”，同时集群拓扑允许多个“根”的存在。



在维护了集群拓扑之后，通过制定数据的分布规则来实现多硬盘池技术。其中，数据分布规则有以下原则：

版权所有 © 超聚变数字技术股份有限公司

- 同一份数据只能存放在一个树状拓扑下，即数据只能在某个“根”内分布，数据的恢复、均衡等操作，也都限制在这个树状拓扑之内。这是多硬盘池技术的实现原则。
- 数据的分布存在安全级别，一份数据的多个副本不能放置在这个安全级别的同一逻辑实体上。例如，选择限定了某个存储池的安全级别为“机架”级别，则往这个存储池存放数据时，同一份数据的不同副本不能放置在同一个机架内的服务器上，必须跨机架分布。

多硬盘池技术，就是通过创建多个逻辑拓扑树状图，即创建多个“根”，任何的节点故障、数据恢复都在某个树状拓扑的范围内产生影响，而不会影响到其他树状拓扑，从而实现故障隔离。

客户价值

通过对故障进行隔离，将故障对业务的影响限制在了某个范围内，避免了“牵一发而动全身”的情况发生，提高了业务连续性。

在分布式架构中，节点的扩容往往意味着集群拓扑图的改变，也伴随着数据的重分布，重分布过程中，也会对集群性能产生影响。利用多硬盘池技术，能够将新扩容的节点独立成一个硬盘池，从而避免了数据的搬迁，大大降低了对业务的影响。

5.11 NAS 文件存储特性

5.11.1 文件存储高可用技术

FusionOne Storage文件存储构建于分布式引擎之上，关键模块是文件系统、协议和文件网关。

文件系统是提供文件存储接口语义的基础，实现文件形式的存储访问，同时提供文件级的特性，如配额、快照等。

协议对外提供NAS协议的访问，直接和客户端进行会话。

文件负载均衡及高可用技术则是依托于文件网关实现，用于保证文件存储业务的高可靠和高性能，同时支持不同部署形式，以匹配整体节点资源消耗。文件网关利用浮动IP实现高可用。文件网关由集群中的多个服务器节点组成。文件网关中的每个节点都可以对外提供文件共享服务，当其中一个节点故障时，同一网关中的其它节点接管该节点上的文件共享服务。

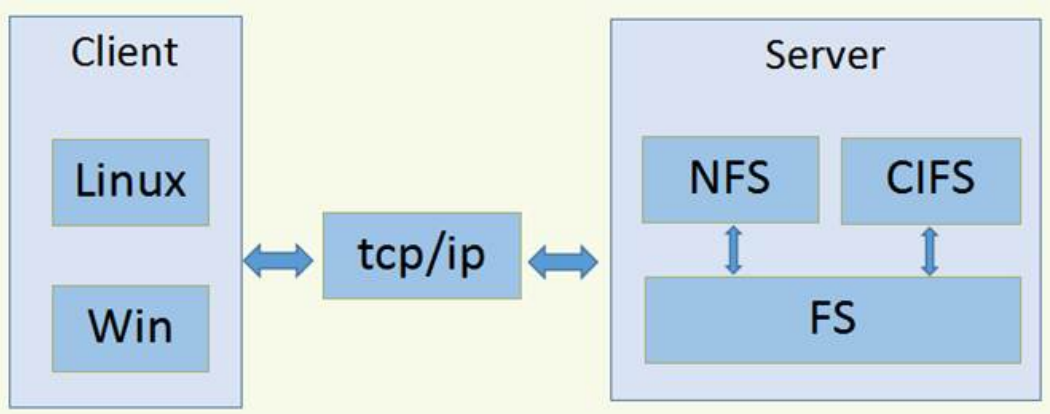
5.11.2 文件存储接口

FusionOne Storage的NFS和CIFS依托于tcp/ip协议，不同协议通过不同的网络端口进行会话，客户端会话通过网络层到达协议层，协议层解析协议请求然后转化为文件系统语义，然后调用FS层接口最终将请求发到文件系统。请求返回过程相反。

同一文件系统可以同时提供NFS和CIFS访问，不同客户端看到的同一个文件系统是完全一致的。文件系统按照协议标准保证元数据一致和数据持久化，NFS客户端写的文件可以通过CIFS客户端立即访问到，CIFS客户端写的文件也可以通过NFS客户端立即访问。为了保证业务行为的数据一致性，建议CIFS、NFS客户端不要同时写同一文件共享下的同一文件。

 版权所有 © 超聚变数字技术股份有限公司

图 5-4 文件业务流程



NFS共享采用IP认证权限，支持只读、读写方式访问共享，支持root_squash、all_squash开关配置，满足客户端将root或所有用户映射到匿名用户（即等于nfsnobody用户访问），便于权限控制。

CIFS共享采用用户认证，支持本地用户本地认证（默认）、AD认证和全局认证。本地用户即存储系统中创建的文件用户，注意不是系统管理员。AD认证支持与AD域对接，满足通用办公场景。支持配置是否允许来宾访问（即无需用户信息即可访问）。

技术特点

- 兼容性好：NFS支持最通用的NFSv3，CIFS支持SMB1、SMB2、SMB3，基本满足所有类型的客户端访问。文件系统基于POSIX语义，兼容所有的文件操作，满足各种业务需求。
- 跨协议访问：Linux客户端通常使用NFS，Windows客户端通常使用CIFS，不同客户端同时访问同一文件共享，可以实现数据跨协议的访问。
- 快速使用：不需要客户端额外安装软件，只要在存储侧配置好共享，Linux客户端或Windows客户端就可以直接访问使用，直接进行文件访问，如同本地文件系统一样操作。
- 共享方便：不同的客户端或不同类型协议的客户端均可以同时访问同一文件系统，数据完全共享，不存在难以互通的问题。
- 权限管控：基于客户端和用户的鉴权方式，支持对接AD域，精细化控制客户端或用户的读写访问权限，保证数据共享安全。

5.11.3 文件配额

FusionOne Storage文件配额包含用户配额和目录配额，支持容量和文件数限制。

用户配额

用户配额需要依赖用户，因此FusionOne Storage NAS存储实现了文件用户管理，包括用户和用户组。文件系统用户配额按照用户和用户组设置，单个用户可以属于多个用户组，而用户组也可以包含多个不同用户，满足群组使用的控制。

当文件系统产生文件和数据时，文件的owner和group表示数据的使用者，配额模块会将使用情况统计到存在的配额条目中，如果配额达到上限，将拒绝业务。

版权所有 © 超聚变数字技术股份有限公司

目录配额

目录配额依赖于文件系统目录，目前支持文件夹一级目录配置目录配额，目录需要通过系统API调用创建（通过协议客户端创建的目录暂不支持配置配额）。

当文件系统对应存在配额的目录写入文件或数据时，配额模块会统计此目录的使用情况，然后查看是否超过配额，如果超过配额则会拒绝业务。

技术特点

- 使用方便：文件存储用户配额和目录配额可以满足文件夹合理使用需求，防止单一用户或目录占用了太多存储空间。同时支持通过API调用进行统一化配置管理。
- 管理简单：配额拥有实时的资源占用统计情况，而且可以实时在线修改。当业务因为配额不足时，业务会产生配额不足错误，明确提示业务扩容配额或删除数据。

5.12 对象存储特性

5.12.1 小文件在线合并

存储集群在不同业务场景下需要面对不同数据类型的挑战。大文件（通常大于1MB）场景下，集群容量占用率达到70%时，文件数也不会太多，存储访问性能较稳定，受文件数量影响较小。但是在海量小文件（通常小于1MB）场景下，系统中的文件数会逐渐增加，相比大文件场景会有超过百倍以上的放大，系统面临着更高的海量数据存储管理的挑战。

虽然分布式对象存储系统相比传统文件系统可以承载更大的文件数量，但还是更加适用于大文件存储。如果不对大文件和小文件分开处理，存储性能可能会在海量小文件场景表现为性能随容量增加迅速下降或者空间浪费严重。例如70%已用容量相比空集群读写性能下降10倍以上，而集群空间占用也放大到实际文件总大小的2倍或者更多。

小文件合并技术应运而生，主要解决海量小文件带来的性能和空间浪费问题，常见的小文件合并有在线合并和离线合并。在线合并为文件写成功即合并成功，拥有较高的技术难度，效果最好。离线合并对文件第一次写入不做改动，而是等后期再单独对文件进行合并，存在读写压力放大，效果其次。

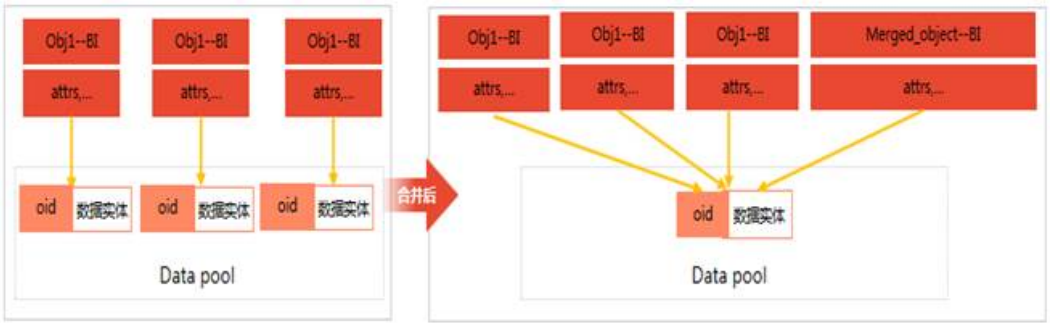
FusionOne Storage采用小文件在线合并技术解决海量小文件场景下的存储性能和空间占用问题。

技术原理

小文件合并即多个对象共享一个物理文件。在线小文件合并的原理是小文件写入之前，系统会分配适合大小的大物理文件用于存放文件数据实体，然后文件数据实体实际被通过追加写方式写入一个大文件中，作为大文件的一部分。这样多个文件写入后就自动合并成了一个大文件。

 版权所有 © 超聚变数字技术股份有限公司

图 5-5 小文件在线合并



文件写入过程中会映射和寻找适合的大对象，写入成功后更新大对象的状态属性，此时小文件的索引也记录了数据实体所属的大对象和对应的数据范围。之后需要获取小文件时，通过小文件索引的记录直接获取数据。

同理，对象文件的删除、同步、转移等特性也基于小文件合并的原理进行，删除后识别回收大文件，同步和转移继续保持合并状态，保证业务高效和正确。

技术特点

使用小文件在线合并技术，一套存储就可以支持不同业务类型，不管是大文件还是小文件场景或者混合模型，存储系统都可以应对自如。一方面不用担心集群使用率上升会带来性能严重下降问题，同时可以防止空间占用过高引起的资源浪费，降低后期扩容风险，降低采购成本。

小文件在线合并使存储拥有更大更好的规格体验，单桶10亿文件可以满足大部分的业务需求，业务无需人工划分数据分桶，降低业务规划难度和管理成本、降低数据需要迁移的工作量和风险和运维难度。

- **性能稳定**

  小文件合并直接解决存储在海量小文件下的性能问题，集群容量达到70%以上性能也不会下降超过20%。采用小文件在线合并，文件的写入和读取的方式发生了变化，但是读写依然是一次性完成，没有额外的耗时动作，时延相比没有变化。而且相比离线合并不会带来后续的读写放大，不会额外增加系统压力。

- **规格可靠**

  采用小文件在线合并，存储集群可以支撑更大体量的文件数量，可以达到单桶至少10亿、整集群1000亿或者更大规格。系统自动合并设定范围的小文件，除了默认的合并规则，系统还支持手动调整或优化小文件的识别界限、合并大文件的个数和大小，可以根据实际环境配置和业务保证更好的业务体验。

- **节省空间**

  一个大文件可以合并几十个小文件（默认80），数据实体的大小都是大文件级别，文件的实体数量得到降低，在海量文件数下可以降低空间占用。

5.12.2 对象多版本及 CDP 技术

在对象存储领域，最常用的数据保护方式为多版本技术。对象存储通过对数据的版本控制，能够有效防止用户数据的误删、误覆盖、误修改等逻辑错误，在存储系统本身提供了数据的备份-恢复能力。

除此之外，对象存储在多版本技术的基础上进行了增强，实现了持续数据保护特性（CDP），能够回滚到某个时间点的数据状态。

 版权所有 © 超聚变数字技术股份有限公司

技术原理

- **多版本控制**

  多版本技术中，实际上即是对对象版本的控制。开启多版本的情况下，每个对象都会携带版本号，在上传相同文件名的对象时，不会直接覆盖原有的对象，而是通过使用新的版本号的方式标记新的对象，新旧版本同时存在。与此同时，存储系统会将对象当前的版本指针指向最新的版本，在用户正常使用的情况下，只能看到当前最新版本的对象。

- **版本回滚**

  开启多版本后，可以通过特定的接口，查看对象的当前版本和所有历史版本，多个版本之间能够进行回滚。

版本回滚的操作，实际上就是将当前版本指针，指向某个历史版本，从而在用户访问时能够访问到老版本的历史版本。

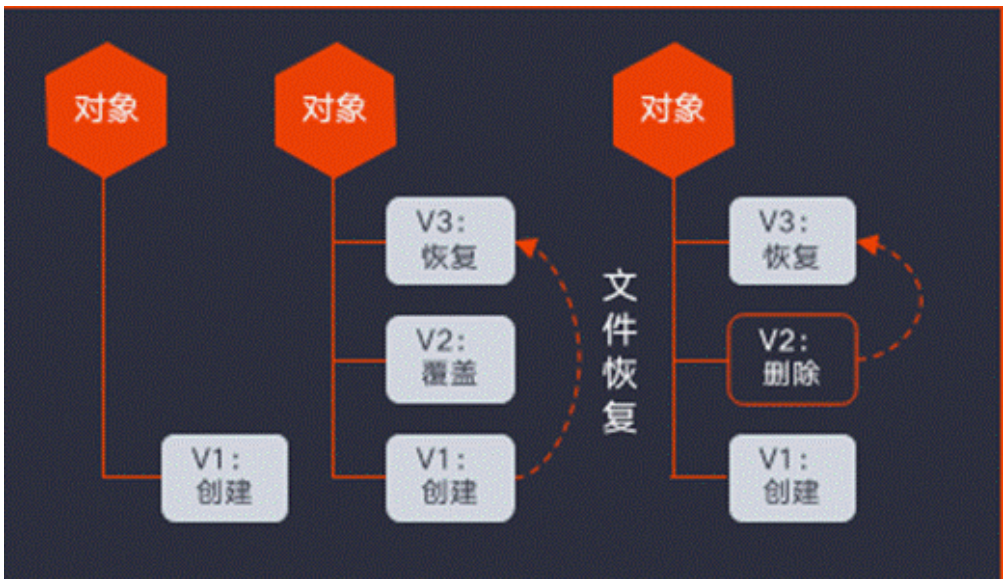
- **删除标记**

  使用多版本技术后，删除文件不会将实际的文件进行删除，而是通过创建一个特殊的空对象的方式来进行处理。这个空对象称为删除标记，历史的所有版本也会继续存在。

  当某个对象的最新版本是删除标记时，用户使用普通的接口无法找到这个对象，从表现上来看这个对象已经被“删除”了。

  带上删除标记的对象也能够通过版本回滚的机制，将历史版本恢复。

图 5-6 多版本示意图



- **版本日志**

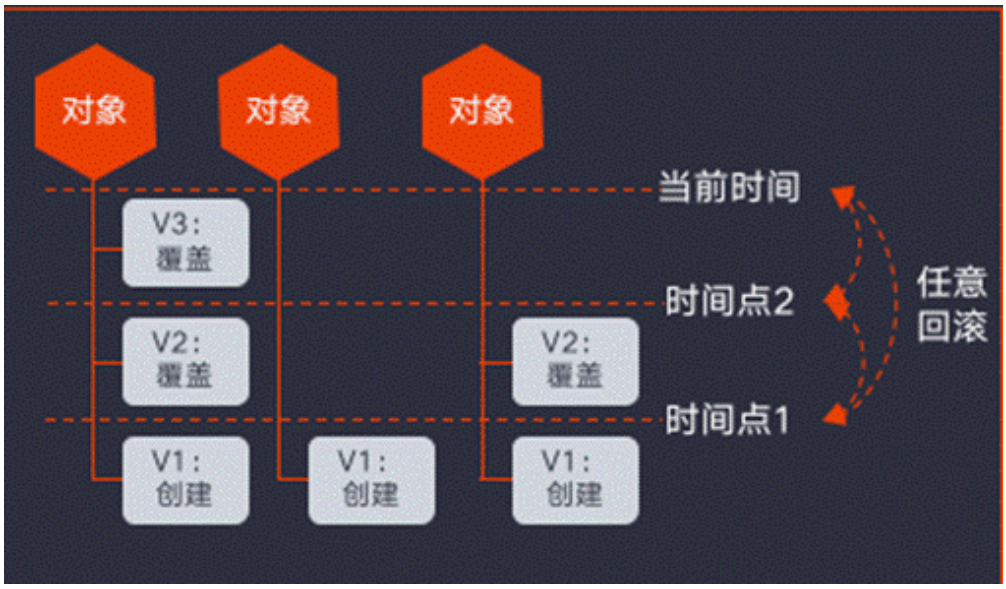
  基于多版本特性，对象存储系统额外开发了持续数据保护功能，用户能够直接将数据回滚到某个时间点的状态。

  每个版本变更时，会将版本变更的信息记录到版本日志，从而记录版本的变更信息。当用户回滚到某个时间点时，会将该时间点到最新时间的版本日志做一次扫描，将所有的更改进行回滚，新增的文件会打上删除标记，原有的文件会被回滚到指定时间点的版本。

  为了防止版本日志过于庞大，因此一般只保留最近7天的版本日志，超出7天的日志将会被覆盖。

版权所有 © 超聚变数字技术股份有限公司

图 5-7 CDP 示意图



技术特点

- 数据保护：多版本机制能够有效地保护用户的重要数据，在误删、误覆盖、被篡改的场景下，提供完备的数据恢复机制，大大提高了重要数据的安全性。
- 防业务逻辑错误：对象存储特有的持续数据保护特性，能够在业务出现逻辑错误，从而导致大量数据错误的场景下，快速、自动地对数据进行回滚恢复，极大的降低了逻辑错误场景下的数据恢复成本。
- 版本控制：多版本特性能够很好地保证数据的安全性，在防误删、防覆盖、防篡改方面应用广泛。对象存储中，每个对象最多能有256个版本，超出的版本将会覆盖最老的版本。
- 高性能：对象存储对多版本场景下进行了许多性能优化，保证开启多版本后的性能与未开启前持平。

5.12.3 对象生命周期管理

生命周期管理是企业管理中重要的一项工作，是对产品、数据等内容从产生到回收的一系列管理过程，它既是一门技术，又是一种制造理念。

对象存储中的生命周期管理技术，即客户根据不同业务数据的生命周期特点，自定义生命周期管理策略，由存储系统自动按照策略执行生命周期管理的一项技术。

良好的数据生命周期管理策略，能够更好的利用存储系统，降低存储成本，降低管理复杂度。

技术原理

对象存储的生命周期管理，主要从时间的维度对数据的生命周期进行度量，超出时间范围的对象，会触发相应的生命周期处理策略。

对象存储中，超出生命周期的处理策略是过期删除策略：针对超出生命周期时间的数据，采取删除操作，回收存储空间。

 版权所有 © 超聚变数字技术股份有限公司

技术特点

FusionOne Storage存储中的生命周期管理，能够大大简便企业数据的生命周期管理，并且支持多种管理粒度，更好的适应企业需要。

- 自动执行生命周期策略：生命周期策略设置完成后持续生效，管理无需人为干预，自动处理过期数据。
- 多粒度生命周期管理：支持针对整个桶设置生命周期，也支持对桶中部分子目录或携带特定前缀的对象设置生命周期。
- 生命周期的影响控制：生命周期的执行主要是数据的删除，会损耗一定的存储性能。FusionOne Storage能够限定生命周期执行的时间段，在业务空闲时间执行，控制生命周期执行过程中对业务的影响。
- 基于对象标签的精确控制：对象存储系统在标准的生命周期功能上，额外研发了基于对象标签的生命周期管理策略，针对拥有特定标签的对象制定单独的策略进行精确控制。

5.12.4 授权管理

对象存储相比传统的文件存储，在移动互联网时代有更好的适应性，在任何能够使用HTTP协议的终端，都能够访问对象存储，从而使得对象存储有着更广泛也更复杂的应用场景。

更多的应用场景，同时也意味着需要更完善的授权管理机制来保证访问安全及数据共享。

对象存储通过多用户机制来管理、隔离用户的数据，通过鉴权技术来保证访问安全，使用桶共享策略技术、外链技术来实现数据的共享。

技术原理

- 多用户

  多用户技术在对象存储中是最常用、最基础的权限管理技术。

  多用户技术支持存储管理员创建多个对象存储用户，每个用户的数据相互隔离，互不影响。每个用户上传对象时，都会携带owner信息，说明数据的属主，从而实现数据的隔离。

- 鉴权

  访问对象存储时，需要使用用户的鉴权信息来进行鉴权，鉴权通过后才能访问对象存储中的数据。

  完成鉴权所需最基础的信息为access key及secret key，客户端访问时需要使用access key及secret key使用加密算法生成签名，在向对象存储发送请求时，将access key及签名一同发送给存储系统，存储系统会根据access key找到对应用户的secret key，随后使用客户端相同的加密算法生成签名，通过对比计算出的签名与客户端发送过来的签名是否一致，从而判定鉴权是否通过。

  除了使用access key及secret key生成鉴权信息，也可以额外将HTTP Header中的字段加入到鉴权过程中，提高签名复杂度，进一步提高请求安全性。

- 桶共享策略

  由于多用户技术的存在，正常情况下，不同的用户无法访问到对方的数据。在实际的应用场景中，数据的共享非常常见，桶共享技术能够满足用户的数据共享需求。

  桶共享策略能够将桶或桶内的某个子目录下的对象以不同的操作权限共享给其他用户或全部用户。

 版权所有 © 超聚变数字技术股份有限公司

用户发出请求访问桶时，存储系统会先根据在这桶上的共享策略，判断用户是否具备访问权限，进而允许或拒绝用户的访问请求。

- 外链

  外链技术是一种预签名技术，通过外链技术共享出来的对象，对方不需要拥有access key及secret key就能够进行访问。

  外链技术共享对象时，存储系统需要将访问对象的URL进行预签名，预签名需要的信息有：请求的部分信息、过期时间戳，随后存储系统会根据这些信息生成一个外链（临时的URL）供用户访问，这个URL在超出时间后会失效。

  在用户访问外链时，请求的信息需要与生成预签名信息时的信息相匹配，否则也会访问失败。

技术特点

对象存储系统中的授权管理机制，从多个维度去保证数据的访问安全及共享安全，适应企业用户管理、数据隔离、共享管理的多方面需求。

桶共享的操作权限有以下几种：

- 只写：限制用户访问已共享的桶。
- 只读：以只读方式将桶共享给其他用户，用户只能读取，无法写入。
- 读写：以读写方式将桶共享给其他用户，用户能够随意读写数据，但无法更改桶的权限信息。
- 全部：授予其他用户完全的控制权限，包括读写、桶权限修改等。

另外，还能够添加共享的授权条件，进一步控制用户的共享访问。

- IP等于：指定用户只有使用特定的IP访问时，才能够访问共享的桶。
- IP不等于：指定用户无法使用某个IP来访问共享的桶。
- 时间点不超过：指定用户只能在某个时间点之前才能访问共享的桶。
- 基于标签的控制：只共享桶中带有指定标签的对象。

生成外链时使用的请求信息与访问外链时携带的请求信息需要保持一致，这些信息主要包括：

- Content-ty信息：请求中携带的content-type字段。
- Header信息：请求中携带的Header字段。

5.12.5 对象压缩

对象压缩与对象重删技术都是针对目前企业数据信息量的快速增长，两种技术都是为了缩减系统中所需要的存储容量，尽量提高空间利用率，降低数据存储成本，从而提升企业的投资回报率，但是两种技术具有不同层面的针对性，简单地说，重删就是将相同的对象只存储一次，而压缩则是通过一种算法来统计数据的规律从而改造数据排布，以减少数据存储空间占用。

对象数据压缩和对象数据重删对存储空间的节省效率与数据的类型强相关。在实际应用中，可以根据如下的原则选择对象数据压缩特性：

在某些归档备份场景，数据的价值密度比较低，客户希望尽量能够节约成本，可以考虑开启数据压缩技术，通过对对象文件进行压缩，降低存储空间成本，可以有效满足客户降低总体成本的需求，但是对于已经通过硬件设备或应用程序（包含备份和归档应用程序）进行压缩或者加密的数据，不推荐进行数据压缩。

 版权所有 © 超聚变数字技术股份有限公司

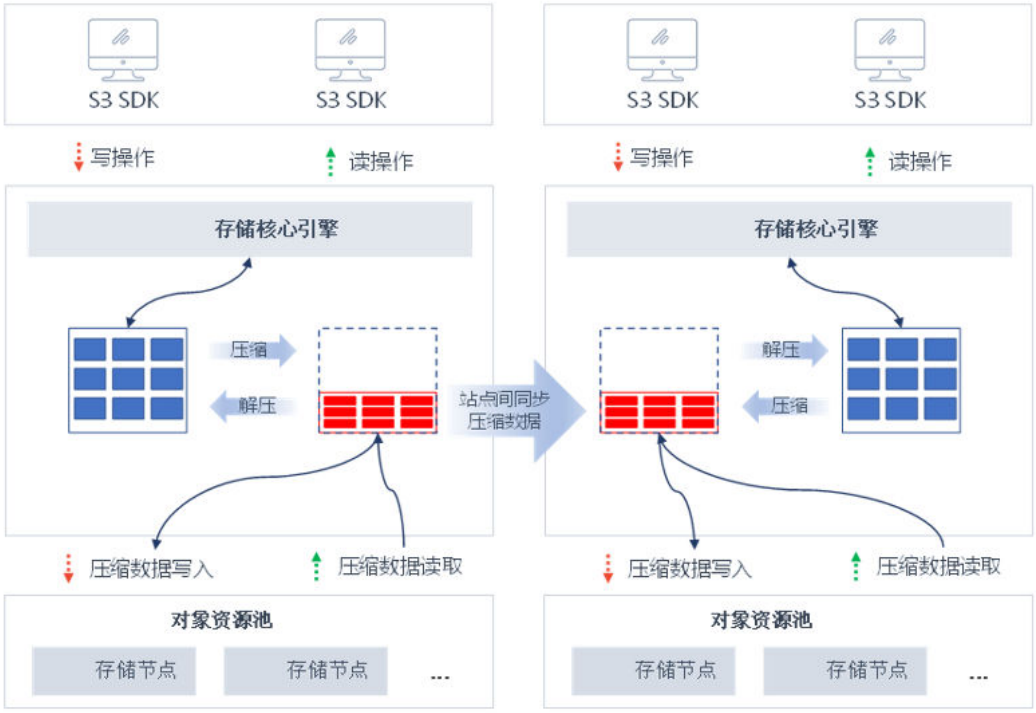
技术原理

对象压缩技术是一种减少数据的冗余和存储空间的技术，压缩技术实现的前提是信息的数据表达存在冗余，是指在不丢失有用信息的前提下，按照一定的算法通过对数据重新组织，即通过检索冗余数据并采用更短的指针表示来实现缩减数据容量。

数据压缩技术通过对单个文件或单个数据块消除数据的冗余，主要通过字符串匹配算法来检索相同数据块，简单来说，就是将一个大字符串中的子串用一个很简短的数字来标记，然后检索该字符串出现的位置，用个简单的字符来替代，从而来减少数据表达所需要的空间。另外数据压缩技术的冗余粒度会很小，可以到几个字节这样的小数据块，而且是自适应的，一般不需要事先指定一个粒度范围，越小的数据块粒度会获得越大的数据消冗效果，但计算消耗也更大。

对象压缩特性在存储架构上涉及上传数据、下载数据和复制对象等模块流程，在上传文件流程中，检查桶是否开启了压缩功能，若开启了压缩功能则先对数据进行压缩处理后，再下发至存储系统层，并将相关压缩信息记录在索引中。在对象下载、复制的流程中，检查对象的元数据记录是否有压缩信息，如果有则使用元数据记录的压缩方法进行解压，然后再将解压后数据发给客户端。

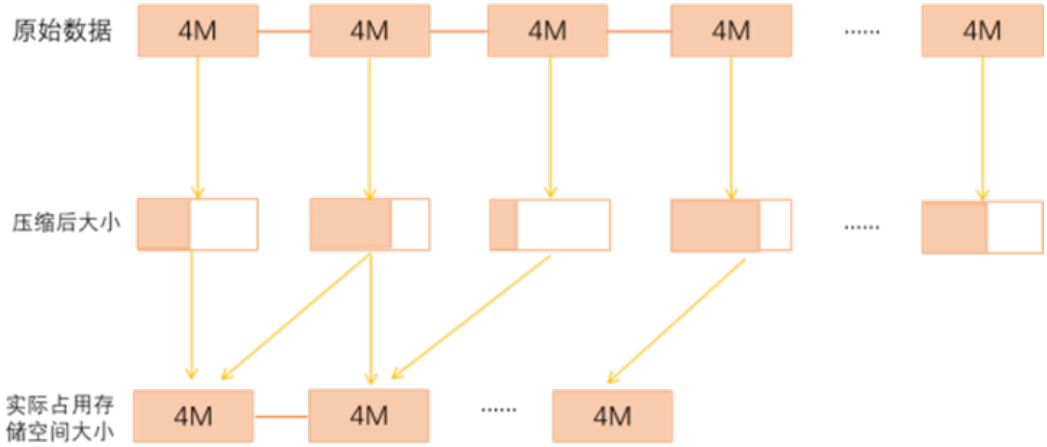
图 5-8 对象读写示意图



不同的文件或者数据片段压缩比率是不一样的，一个4M数据块压缩后可能是3M、2M，或者也有可能是几十至几百KB。对象压缩特性可以查询桶内压缩文件数量、原始数据容量大小及实际占用空间大小，从而也可以计算出数据的实际压缩比率。

 版权所有 © 超聚变数字技术股份有限公司

图 5-9 对象压缩示意图



对象压缩特性可以按照桶粒度设置数据的压缩开关，压缩及解压过程在文件上传和下载等具体流程中，由对象存储服务网关层自行处理，对用户无感知。

技术特点

- 用户开启对象压缩以后可以直观的查看存储文件数量、压缩文件数量、原始数据总存储量及实际占用存储空间大小，简单高效地计算出数据的整体压缩比率，可在管理界面直接展示节省空间。
- 开启对象压缩特性后，在多站点之间进行数据同步时直接传输压缩后的数据，从而节省传输带宽，提高数据同步效率，同时在执行生命周期转移功能时同样也是直接转移压缩数据，实现数据的快速转移，提高转移效率。
- 对象压缩技术可以通过手动设置数据压缩文件格式的黑白名单，根据对象的格式决定是否进行压缩。

5.12.6 对象重删

随着对象存储的应用领域越来越广泛，企业数据中心存储需求量越来越庞大，数据量呈指数级增长，已从之前的TB级上升到PB级甚至EB级。随着企业数据信息量的不断增加，企业需要不断购置大量的存储设备去解决“数据无休止增长”带来的数据存储问题，使得空间要求、散热、能耗等问题日益突出，造成数据中心的运营成本、数据管理成本不断上升。同时，容易造成存储资源浪费、空间利用效率低等问题。

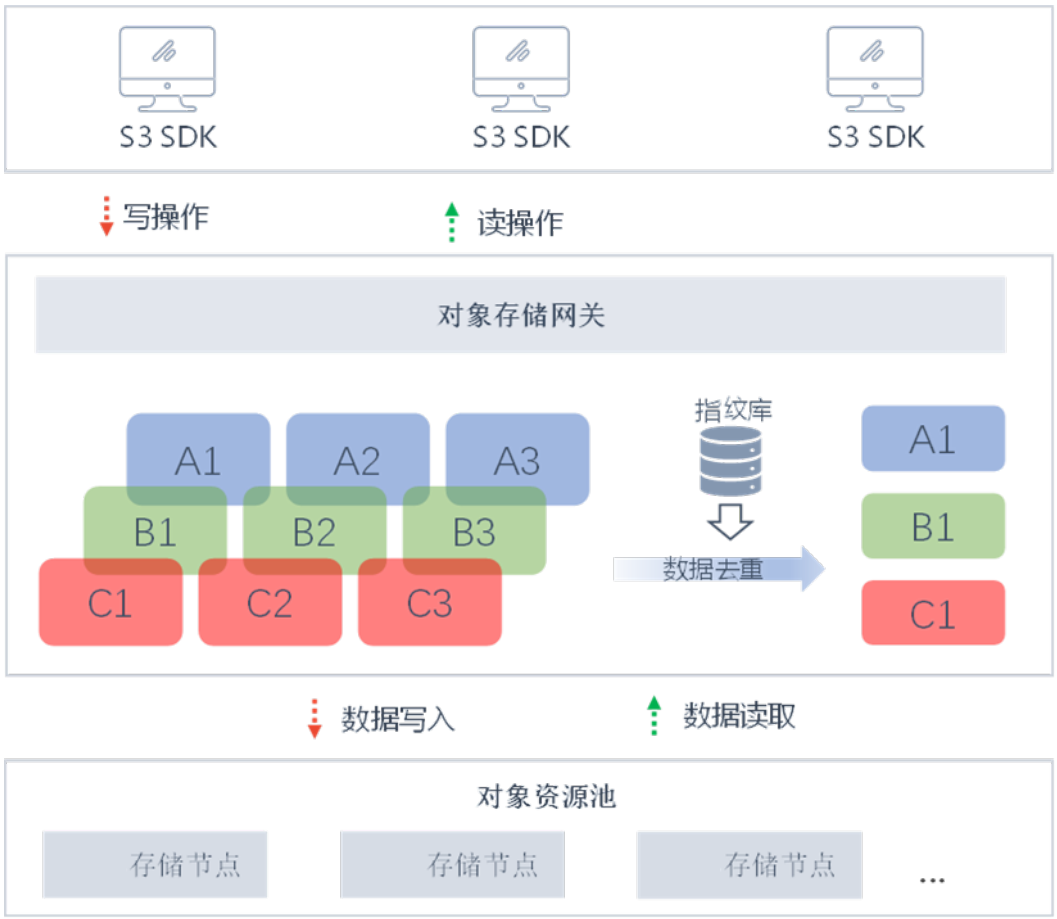
随着应用系统所保存的数据时间周期越来越长，企业各部门对数据的调取和存储，会出现相同文件的多次存储情况，因此文件冗余比例也上升。FusionOne Storage对象存储重删功能解决了重复数据空间占用问题，缩减系统中所需要的存储容量，提高存储空间利用率，降低了数据存储成本，大幅提升了投资回报率。

技术原理

对象重删是指对象存储中的重复数据进行删除，旨在减少存储系统中数据占用的存储容量，以达到节省存储空间的目的。重删功能通过将对象文件指纹数据跟系统指纹库对比分析，可判断对象是否相同。如果是相同对象则删除多余的对象副本，仅保留一份对象数据，重复对象更新自己的元数据，指向同一对象数据实体；如果不存在相同对象，则不需要重删操作，将对象指纹更新至指纹库，即保证了重复的数据仅在磁盘介质上写一份，其它重复的副本建立“索引”，这样可以有效地节省存储空间。

版权所有 © 超聚变数字技术股份有限公司

图 5-10 对象重删示意图



对象重删的主要流程如下：

1. 上传对象文件。
2. 对象写入成功后添加到重删队列。
3. 获取对象的MD5值，查找指纹库，判断对象是否需要进行重删，如果已存在，说明需要重删。
4. 对于需要去重的对象，删除重复数据，更新元数据指向同一对象实体数据，则对象文件写完成。
5. 当业务需要删除对象时，如果对象数据存在其他对象引用，则只删除当前对象的相关引用；如果对象数据只剩当前对象引用，则进行更新指纹库，同时删除数据，保证一致性。
6. 如果不需要重删操作，将对象MD5值更新至指纹库。

技术特点

FusionOne Storage数据对象重删功能可以实现从存储系统的多个管理维度实现不同粒度、不同方式的对象重删，满足客户多层次数据管理需求。对象重删技术可以消除冗余数据，因此采用对象重删技术可以在给定数量的磁盘设备上存储更多的数据，极大节省存储空间，提升空间利用率，减少数据中心设备数量及电力消耗，降低运营成本。

同时可以防止空间占用过高引起的资源浪费，降低采购成本。

 版权所有 © 超聚变数字技术股份有限公司

- 重删范围粒度策略

  对象重删功能基于存储池范围可以按照全局、用户、桶粒度设置重删开关，即可以按照站点、用户以及桶粒度开启或关闭重删特性。桶的配置优先级高于用户，用户的配置优先级高于全局配置。默认情况下重删特性是未启用的，通过设置对象重删特性开启后，系统就会按照开启的重删范围扫描新写入的对象并进行重复数据删除。

- 重删对象粒度

  支持对象级、分片级重删，分段上传的对象可以选择按照分片来重删还是按照整体来重删，按整体重删需要重新计算整个对象的MD5值，去重率比较高，但会增加部分性能消耗，按分片重删则直接使用现在保存分片的MD5值，采用这种方式可以避免性能的额外消耗。

- 重删方式

  FusionOne Storage对象存储采用异步重删方式，对象写入成功后才会进行重删处理，因此不会影响业务写性能。

5.12.7 对象预览

图片预览功能

FusionOne Storage分布式对象存储系统在预览jpg和bmp格式的图片时，支持默认最佳比例全屏预览，且允许用户通过键盘上下键切换上一张、下一张图片。

图片在线编辑支持输入文字

FusionOne Storage分布式对象存储系统在线编辑图片时，支持输入文字，并可以定义字体（微软雅黑及免费字体）、颜色、字体大小。

5.12.8 对象检索

支持批量检索

FusionOne Storage分布式对象存储系统不仅能够支持百亿级对象数量的存储，还支持检索功能，即使数据量达到百亿级，仍能够提供快速批量检索

- 支持定义一批value进行搜索，value最多支持1w条。
- 通过导入csv格式的方式检索相关数据。

DataExplorer 支持按照时间标签的时间范围检索数据

FusionOne Storage分布式对象存储系统在检索模板中，能够定义需要检索的标签为时间类型，并在数据管理系统中允许根据某个指定标签的时间范围检索到相应的数据。

5.13 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

 版权所有 © 超聚变数字技术股份有限公司

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）。

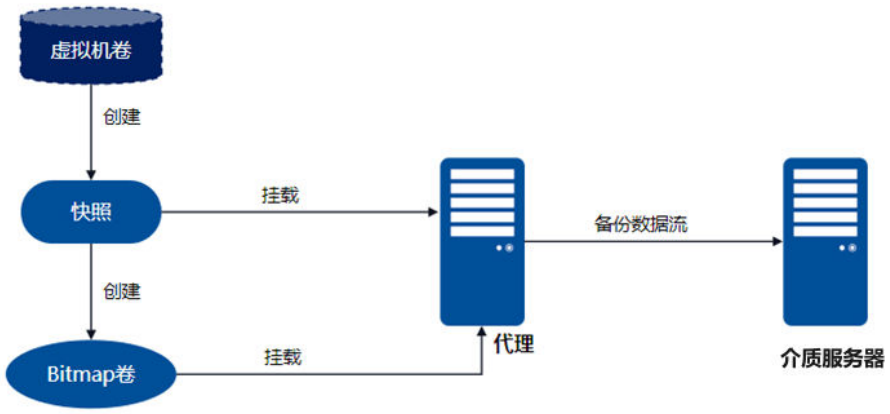
5.13.1 备份

5.13.1.1 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。

5.13.1.2 分布式存储备份

分布式存储备份，先对虚拟机的卷创建快照，然后通过快照做对比（快照与原卷做对比，或者快照之间最对比），将快照之间的数据做导出，并将导出的数据拷贝到存储介质中。



5.13.1.3 虚拟化存储备份

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。

通过如下机制，解决备份恢复性能低的问题：

1. 单卷多队列并发：

   对于同一个卷使用多链接与多并发的写，因此可以让多个写操作同时进行，减少了写操作的总等待时间。

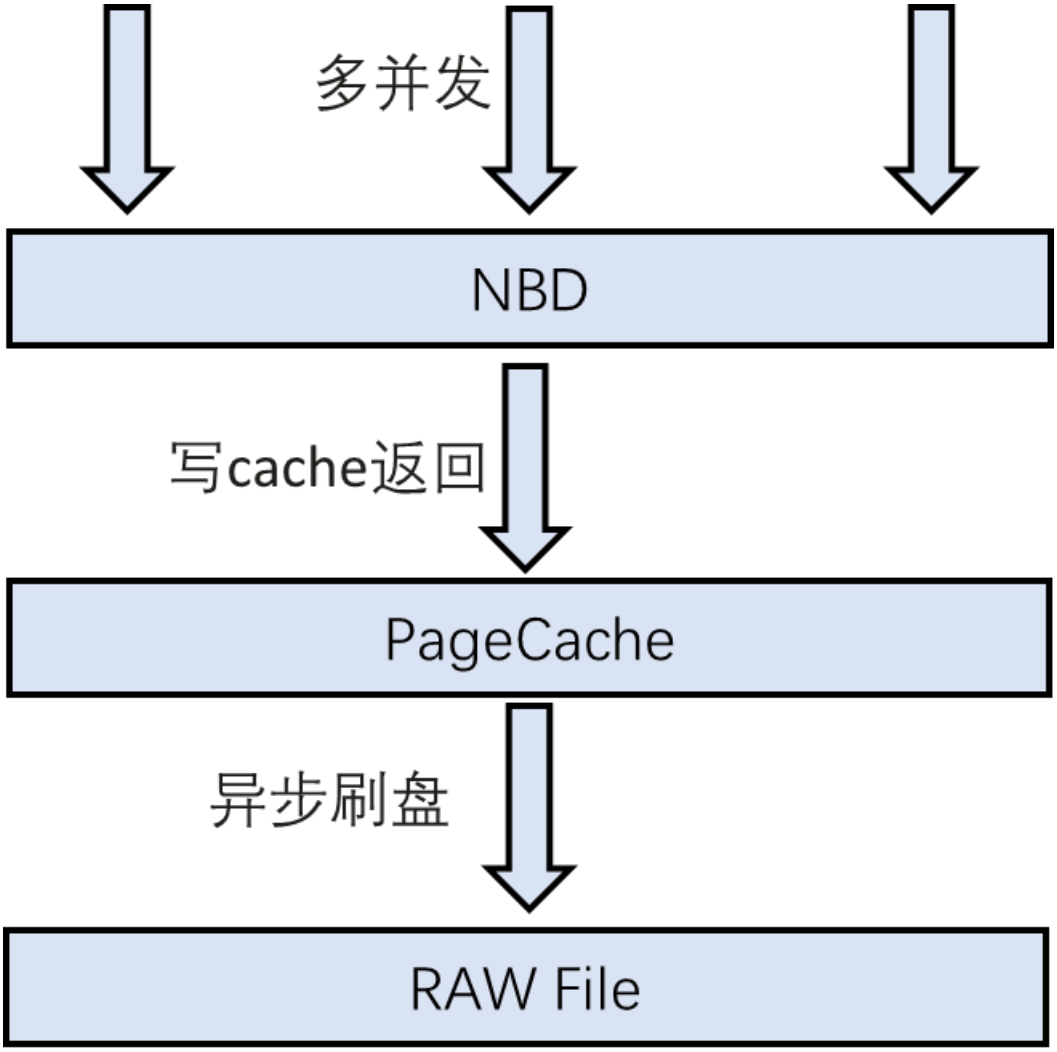
2. 写cahce返回

   使用WriteBack的模式挂载NBD，降低时延，提高写带宽的吞吐量

3. 异步刷盘

版权所有 © 超聚变数字技术股份有限公司

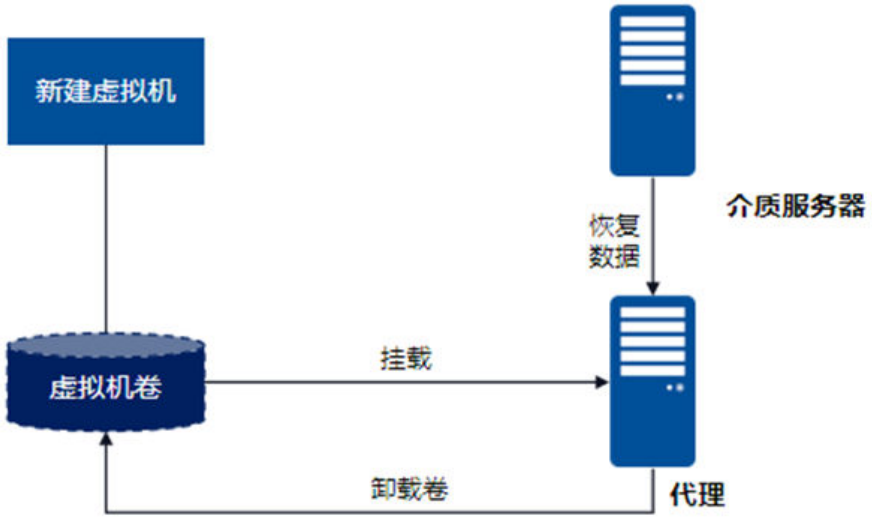
后台线程定时flush块设备和sync文件，定期触发缓存刷盘，在缓存内部合并写请求，将随机写改为顺序写，可以提升刷盘效率。



5.13.1.4 新机恢复

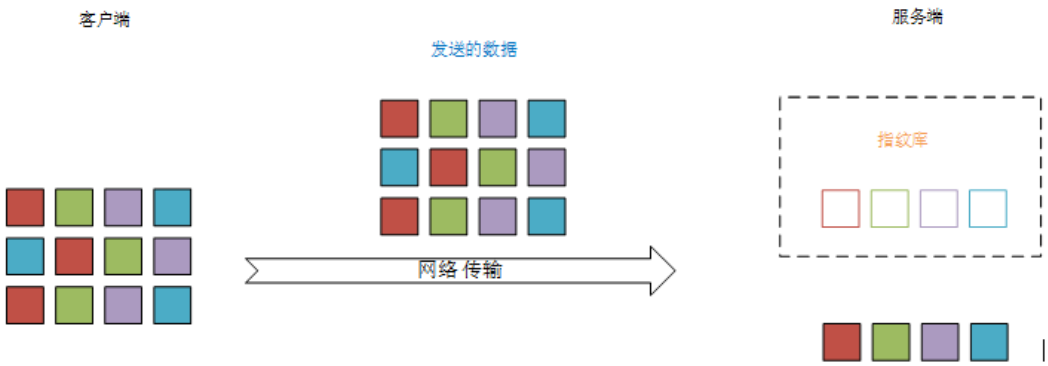
新建虚拟机恢复，先创建新虚拟机，然后将虚拟机的卷挂载到代理上，再将介质服务器上的数据拷贝到卷中，最后在代理上卸载卷，将卷挂载到虚拟机上。

 版权所有 © 超聚变数字技术股份有限公司



5.13.1.5 数据重删

对原数据分块，并计算出每个数据块的指纹（Hash值），相同的指纹说明数据块的内容是相同的，将每个数据块的指纹与指纹库里指纹比对，如果不同，说明是新的数据块，需要存储数据块并将其指纹存入指纹库，如果相同说明数据块是重复的，则不需要存储数据块。对于一次备份，如果相同的指纹越多，那么重删率越高，需要存储的数据块越少，占用的空间越少。数据重删包括源端重删和目标端重删。本系统支持目标端重删，将完整的原数据发送到服务端，在服务端定长切分数据块和计算指纹，和指纹库比对，只存储新的数据块。



5.13.1.6 加密传输

将原数据流在客户端按固定长度分块，再做加密操作（其中加密采用AES256、SM4算法），数据处理完通过网路传输到服务端，达到降低传输数据量及保证数据安全的效果。

5.13.2 CDP 备份

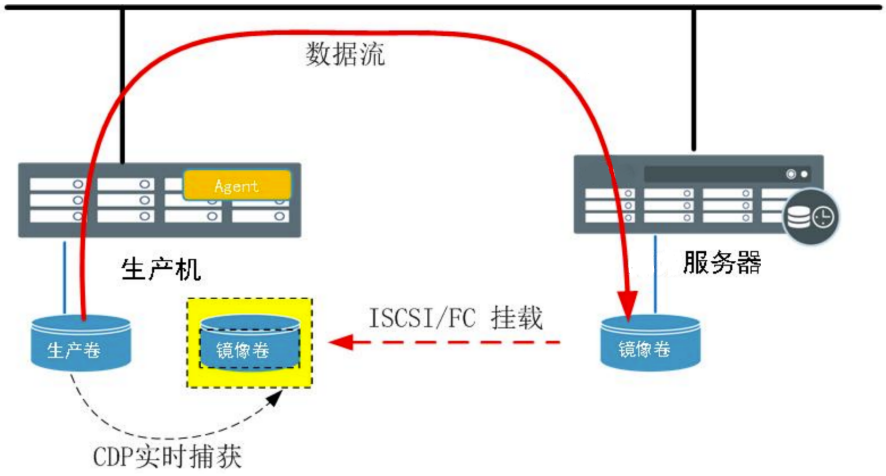
CDP（Continuous Data Protection）是持续数据保护的简称，对生产机数据进行一次初始化备份，并持续监控数据变化，将初始化备份数据和实时捕获到的变化数据存储到备份服务器中，最终实现持续数据保护的目的。其中实时捕获到的变化数据块都会加上时间戳，将变化数据的元数据信息记录到日志中。根据日志可以实现任意时间点恢复、挂载、细粒度恢复等，满足企业更高要求的恢复时间目标（RTO）和恢复点目

版权所有 © 超聚变数字技术股份有限公司

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。

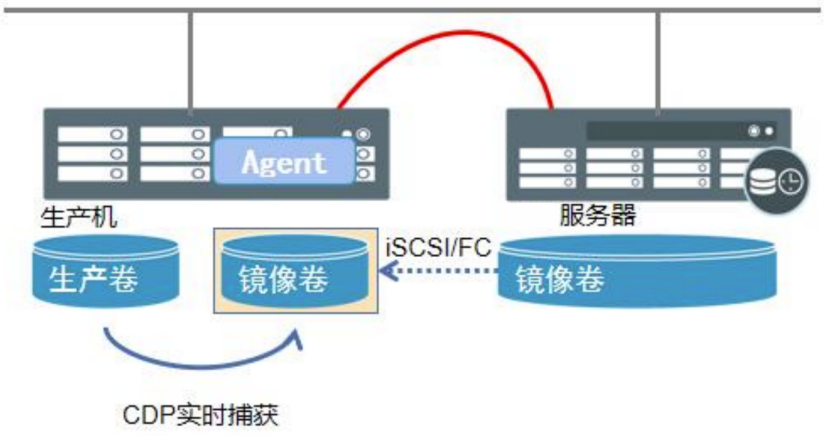
5.13.2.1 原理描述

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：



5.13.2.2 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。

版权所有 © 超聚变数字技术股份有限公司

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。

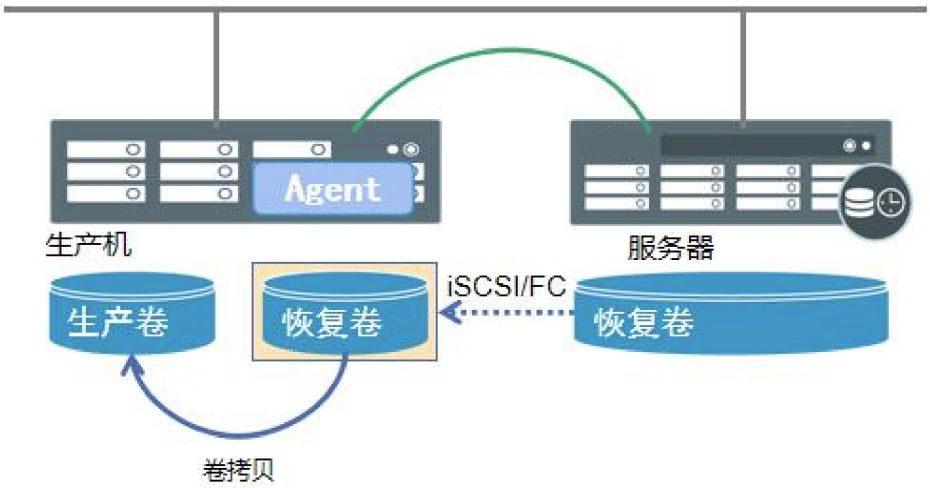
步骤7 服务器将镜像卷写操作的数据块打上时间戳并记录块序号，写入备份文件和日志文件中。

----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。

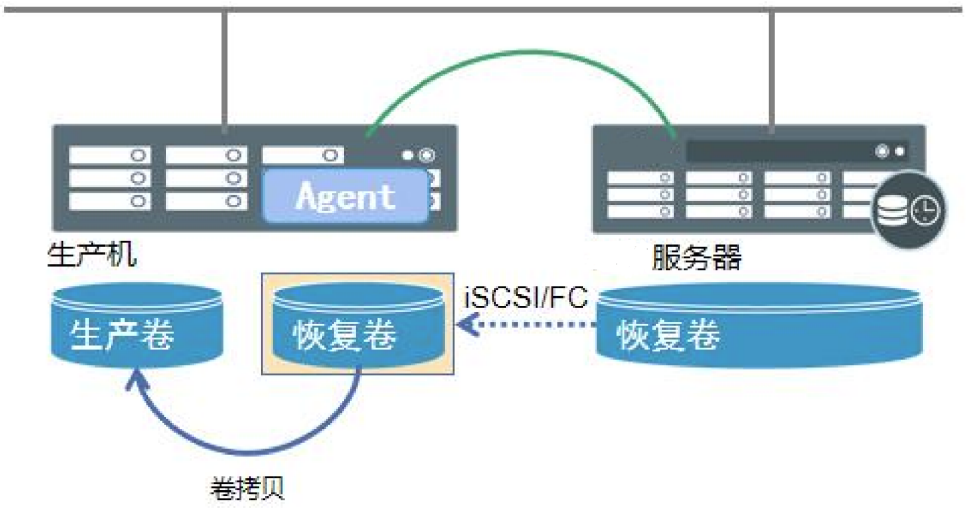
----结束

 版权所有 © 超聚变数字技术股份有限公司

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。

5.14 分布式存储内部机制

5.14.1 负载均衡

FusionOne Storage存储系统是一套分布式存储系统，由多个存储节点组成，存储集群中每个存储节点都能够对外提供存储服务。负载均衡及高可用技术，能够在现有的网络结构上扩展网络设备及存储节点的带宽，增加吞吐量及网络处理能力，并提供访问入口高可用，保障系统的可靠性及业务的连续性。

 版权所有 © 超聚变数字技术股份有限公司

多路径

多路径是由块业务客户端和存储设备一起配合实现，每个主机可以通过不同的网络路径连接到存储不同存储节点。FusionOne Storage支持块客户端多路径自动切换存储服务的机制，当某个节点或者某条IO路径出现故障时，FusionOne Storage客户端会自动切换到其他正常节点和IO路径，以保证业务的持续可用性。

多路径下支持块业务负载均衡，客户端业务通过不同路径均衡的下发业务，保证整体业务性能达到最优。

高效负载均衡

块业务通过多路径可以实现业务高可用和负载均衡，多路径下的每条路径都是随时可用，保证在故障场景下能很快切换其他路径，极大地降低业务感知时间。

高性能

存储每个节点都可以对外提供可靠的业务，不同节点可以配置或通过客户端分配业务压力，这样可以大大增加分布式存储的优势，提高整集群性能和吞吐量。

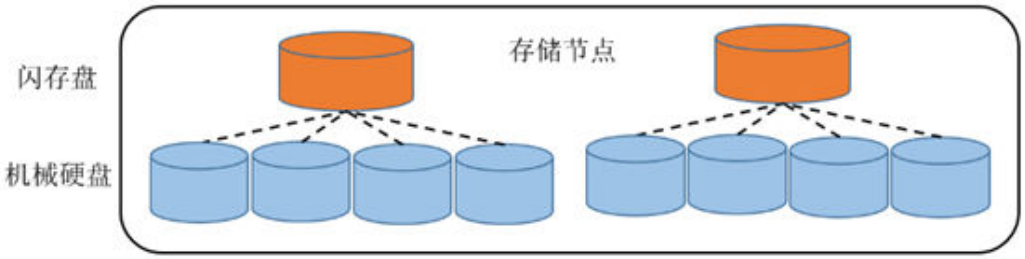
快速故障切换

在节点故障的场景下，故障切换能够保证在10s内完成，大大提高业务连续性。

FusionOne Storage系统提供块存储卷的负载均衡及高可用方案，具备高性能、高可靠性等优点，能够在实际生产使用时发挥出存储集群的整体性能。快速的故障切换能够大大缩短节点故障所带来的切换时间，保障业务连续性。

5.14.2 智能缓存

FusionOne Storage的每个存储节点上配置少量闪存盘，一块闪存盘与多块硬盘组成绑定关系。每块闪存盘可给多块HDD做缓存加速。一般情况下，建议每个存储节点至少配置两块或以上闪存盘。



5.14.2.1 数据类型感知

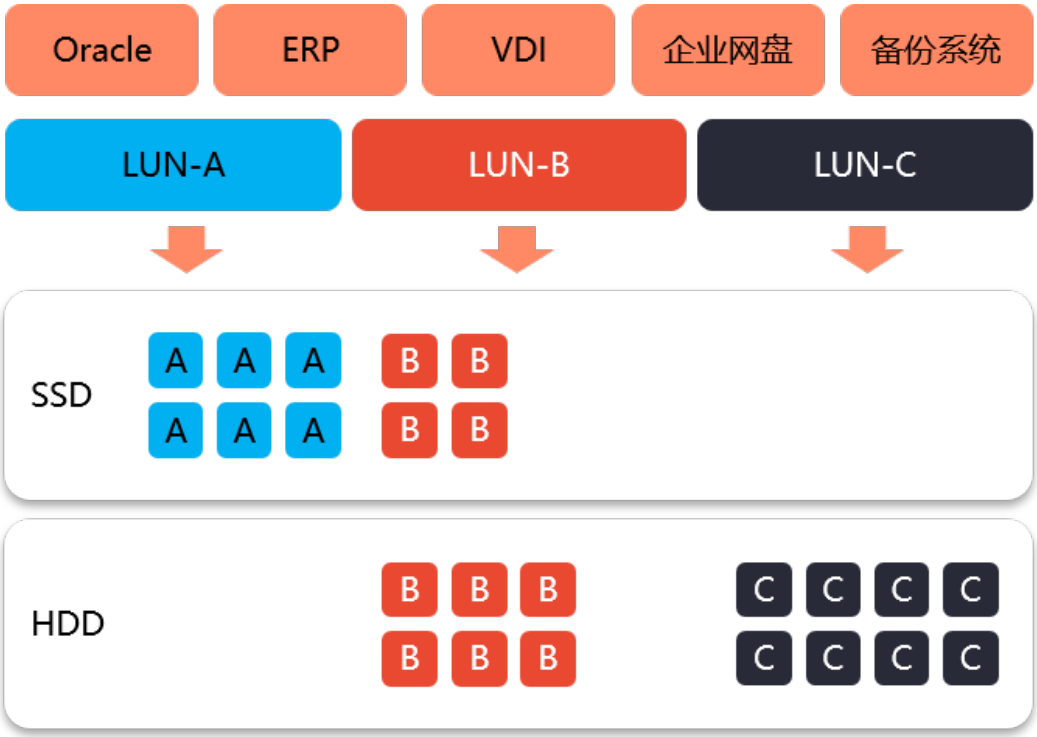
FusionOne Storage的智能缓存模块可自动识别存储系统不同类型的数据，针对不同的数据类型自动采用不同的加速策略，优化缓存资源利用率与业务IO性能。

对于系统元数据、高优先级的数据（如设置了高优先级的桶数据），智能缓存模块不会将其淘汰出闪存，保证IO路径上的元数据访问性能以及较为繁忙业务的数据访问性能。

对于恢复过程产生的数据流量，智能缓存模块同样可以自动识别，这部分数据会透过缓存模块，直接写入硬盘，保证集群不会因为一时的恢复导致缓存资源被临时占用。

版权所有 © 超聚变数字技术股份有限公司

图 5-11 数据感知的智能缓存



5.14.2.2 多级缓存加速

FusionOne Storage的智能缓存模块采用两级缓存技术，第一级缓存为内存缓存，缓存更为频繁访问的数据，加速数据的读取。第二级为闪存缓存，可同时加速数据读写，由于闪存盘的保电特性，当新写入的数据到达每个存储节点，且数据写到闪存盘后，便可立即返回客户端写操作成功。数据进入闪存盘就表示数据已经安全，因此不需要立即刷到对应的硬盘中，存储系统会以异步模式将闪存盘中的数据进行去重和合并处理后刷到后端的硬盘中，即如果对一个数据进行多次修改，则只需将最新修改的数据刷到硬盘中，如果多个数据块是属于一个并且数据在逻辑上是连续的，则可以将这些逻辑上连续的数据写入到物理连续的硬盘上，这样数据被访问时即使未命中缓存，也会因为其在硬盘上的顺序性和连续性而提高数据的访问性能。

5.14.2.3 分布式缓存

FusionOne Storage块存储缓存和带宽都均匀分布到各个节点上。

FusionOne Storage块存储集群内各节点节点的硬盘使用独立的I/O带宽，不存在独立存储系统中大量磁盘共享计算设备和存储设备之间有限带宽的问题。

FusionOne Storage块存储支持将节点部分内存用作读缓存，SSD用作写缓存，数据缓存均匀分布到各个节点上，所有节点的缓存总容量远大于采用外置独立存储的方案。即使采用大容量低成本的SATA硬盘，FusionOne Storage块存储仍然可以发挥很高的I/O性能，整体性能提升1~3倍。

FusionOne Storage块存储支持SSD用作数据缓存，除具备通常的写缓存外，增加热点数据统计和缓存功能，加上其大容量的优势，进一步提升了系统性能。

版权所有 © 超聚变数字技术股份有限公司

5.14.2.4 读缓存

FusionOne Storage使用SSD作为读Cache介质以加速存储访问。FusionOne Storage的读缓存采用分层机制：

第一层为内存Cache，内存Cache采用LRU机制缓存数据；

第二层为SSD Cache，采用热点读机制，系统会统计每个读取的数据，并统计热点访问因子，当达到阈值时，系统会自动缓存数据到SSD中，同时会将长时间未被访问的数据移出SSD。

同时FusionOne Storage支持预读机制，统计读数据的相关性，读取某块数据时自动将相关性高的块读出并缓存到SSD中。

5.14.2.5 写缓存

分布式存储在收到客户端发送的写IO操作时，会将写IO缓存在SSD缓存后完成本节点写操作。

副本场景：SSD缓存中的内容为IO的完整数据

EC场景：SSD缓存中的内容为数据或者冗余条带

同时，分布式存储会周期将缓存在SSD缓存中的写IO数据批量写入到硬盘，写缓存有一个水位值，未到刷盘周期超过设定水位值也会将缓存中数据写入到硬盘中。

5.14.2.6 缓存资源回收

数据下盘

新写入闪存盘的数据，会被智能缓存模块刷到后端硬盘上，这个过程称为刷盘。刷到硬盘之后，原来驻留在闪存盘中的数据仍可以提供读命中，不会立即失效。除非智能缓存模块判定其为不经常被访问的冷数据，才会被回收。

数据老化

当闪存盘中的数据达到老化阈值时，会按照LRU来回收最长时间未被访问的缓存数据，以便预留出足够的缓存空间给新业务数据。

5.14.2.7 大块数据透传

FusionOne Storage的智能缓存模块可根据预设的数据块大小缓存策略来按业务写请求的大小进行是否缓存的决策。默认情况下，当数据块大小大于512KB，并且此次写请求与上次写请求的逻辑地址连续时，即使闪存中有空间也不会写入闪存，而是直接写入硬盘，充分发挥机械硬盘的顺序数据访问性能。

5.15 存储高性能

软件定义的超高性能分布式存储系统FusionOne Storage HBS，实现了用户态端到端NVMe over RDMA（RoCEv2）；通过私有CPU调度器与虚拟内存管理，基于微控制器实现高效的并发流水线作业，从而从根源上解决存储系统的性能问题；同时考虑SSD的地址映射、多队列、垃圾回收等一系列特性，充分发挥SSD真实性能。使用典配机型组建的3节点全闪高性能存储集群，3副本冗余策略，4K随机读IOPS可以达到200万以上、4K随机写IOPS可达60万以上。

 版权所有 © 超聚变数字技术股份有限公司

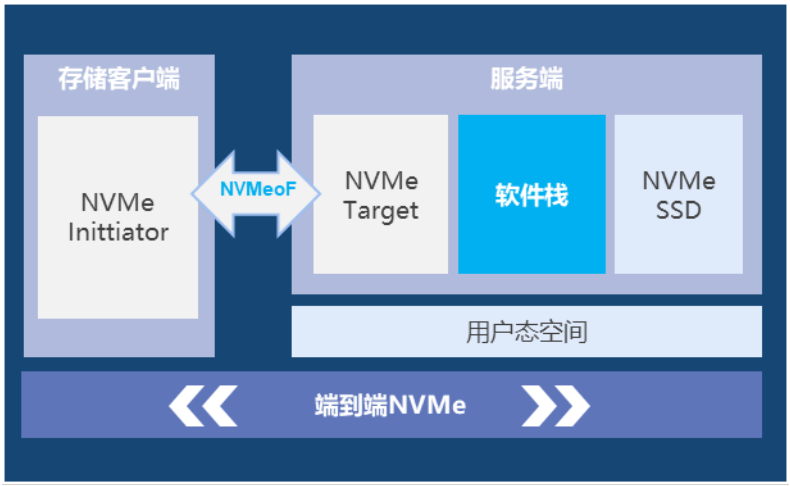
5.15.1 端到端 NVMe over RDMA

存储系统采用RDMA Transport（RoCE v2）实现NVMeoF协议访问。

NVMeoF访问链路分为Host（Initiator）端和Target端，Host端采用NVMe-CLI工具，Target端采用SPDK的NVMeoF Target组件，全用户态实现端到端NVMe。

NVMeoF访问时，完成协议解析后会通过子卷控制器连接到目标存储卷。

三节点全闪集群，采用25GE网卡，4K随机读写性能可以达到读220万IOPS/写100万IOPS。



全用户态实现；端到端NVME

5.15.2 分布式微控制器

FusionOne Storage HBS将数据读写I/O分解为协议解析、I/O处理、元数据访问、落盘等阶段，每个阶段由不同的微控制器来专门处理。

主要有两类微控制器：I/O控制器和元数据控制器。I/O控制器包括TGctl、FRctl、RGctl、BActl等，元数据控制器有MDctl。

存储集群每个节点上存在多种微控制器，每种微控制器数量也各不相同，各微控制器按最优配比关系分配CPU核心。随着I/O负载特性的变化，存储系统会根据实际情况进行智能调整，灵活根据负载情况为微控制器分配合理的CPU核心与内存资源。每类微控制器服务专注于特定的任务，独享充足的计算资源，完全避免了NUMA多核架构下内存跨核访问的性能问题。

各微控制器的功能见下表：

TGctl	目标端控制器，转发I/O请求到本节点内同一NUMA节点内的前端控制器（FRctl），并进行下一步I/O处理
FRctl	前端控制器，负责I/O路径管理与选择，并将I/O转发到后端控制器（BActl）进行读写落盘
RGctl	子卷控制器，向元数据控制器（MDctl）查询并缓存存储子卷的元数据信息，并负责子卷的数据分布管理，协调子卷的I/O读写、修复、平衡等任务

版权所有 © 超聚变数字技术股份有限公司

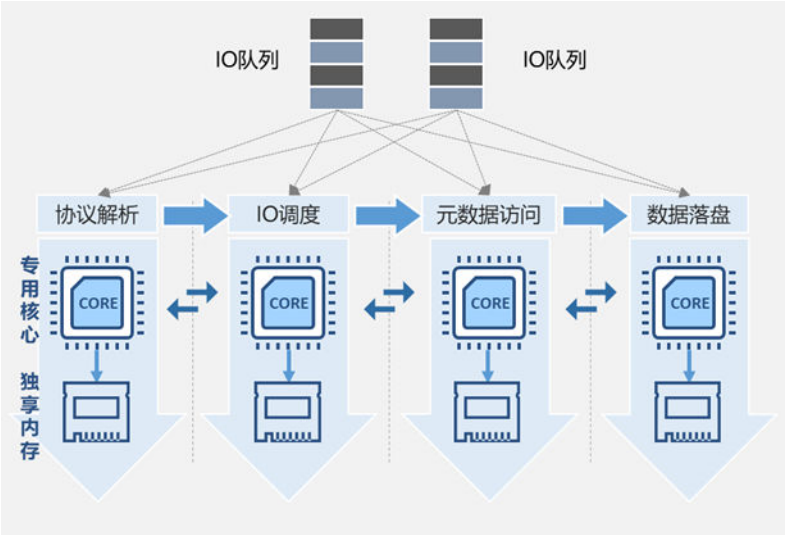
BActl	后端控制器，负载管理本地硬盘的数据读写，提供数据的落盘写入和读取服务
MDctl	元数据控制器，运行K/V数据库，提供元数据查询与持久化写入管理，由集群中某节点上的元数据控制器（MDctl）充当Master角色。

5.15.3 并行流水线

在存储集群中，每个节点的角色都是完全对等的，读写I/O从应用端流入后，被拆分成多个阶段，分别对应着各个微控制器进程。

每个I/O被转发到本节点内的其他微控制器上进行下一步处理，逐步流水线推进直到数据最终落盘。

各个微控制器间由于相对固定的资源锁定，几乎不存在CPU核心资源竞争与跨NUMA节点的访问，也不存在因为CPU上下文切换与分支预测错误带来的巨大延迟开销，多个读写I/O宏观上被均匀地分布在整个存储集群，由各个节点的多个CPU核心进行高效的并行协作处理，从而实现了在分布式体系下微秒级的极低时延响应。



5.15.4 虚拟内存管理

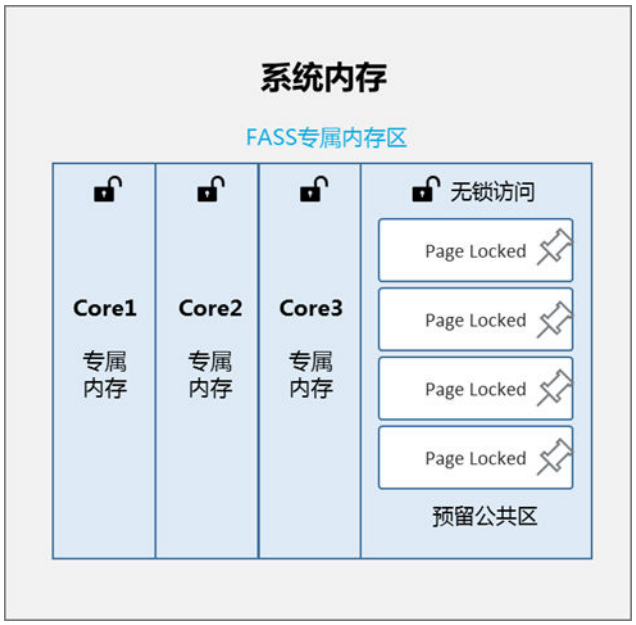
Linux操作系统提供的内存分配器，存在CPU亲和性、碎片化、同步等关键问题，会导致极大的访存开销。

FusionOne Storage HBS采用HugePage虚拟内存管理机制，一次性从操作系统申请出所需内存，降低页表查询开销。同时将页面锁定到内存，避免发生页面交换。

存储系统为其每个CPU核心分配专属内存空间，实现真正的专核专用，避免内存锁带来的访问开销。同时针对每个微控制器服务（如FRctl、RGctl等），可自动根据I/O负载情况，灵活为不同微控制器分配不同数量的CPU核心，同时也保障了每个微控制器都拥有充足的专属内存空间。

为防止极端情况下出现内存资源不足的问题，存储系统会预留足够的公共内存区，在内存资源不足时，各微控制器可以从公共区申请使用。存储系统的虚拟内存管理器可以维护不同尺寸的页表粒度，实现灵活保证高效的内存资源分配，并避免Page交换的情况发生。

版权所有 © 超聚变数字技术股份有限公司



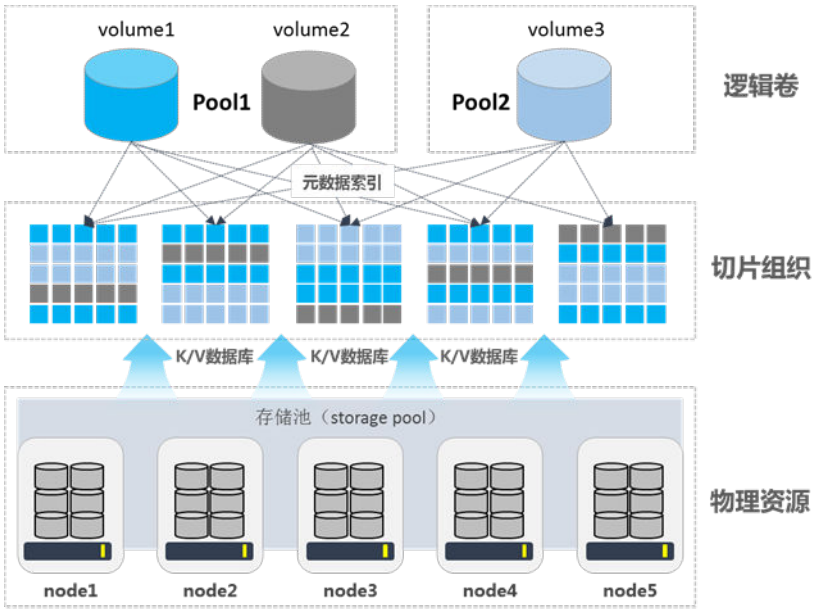
5.15.5 元数据高效索引

通常情况下分布式存储系统依赖各节点操作系统提供的本地文件系统，业务数据被打散写入到不同节点，由各节点的本地文件系统管理，上层再进行逻辑整合以实现统一管理。在本地文件系统基础上加入一层封装与索引管理，毫无疑问会加大了系统复杂度、降低系统的整体性能。如再进行块协议导出，则又多出一层块存储的映射索引，使整个系统运行效率低下。

为发挥出硬件的最大性能，FusionOne Storage HBS摒弃了各节点的本地文件系统，自主开发了轻量级的二级块索引服务，通过对裸盘的直接管理，以Slice切片为单位进行逻辑映射重组，实现远高于其他分布式存储系统的I/O性能。

FusionOne Storage HBS通过多层元数据索引进行数据切片高效定位与管理，主元数据服务管理数据到各个节点的映射关系，通过元数据查询可以迅速定位到数据的存储节点； 同时每个节点上都维护有一份本地数据块索引，通过K/V数据库（Key-Value数据库）记录了本节点所有数据盘上Slice分布信息，并将其缓存在内存中；每个数据盘都由后端控制器（BActl）管理其具体读写操作；每块硬盘都对应一份Diskmap信息，记录着该硬盘上的Slice分配情况，新建存储卷时，子卷控制器（RGctl）根据Diskmap信息与数据分布策略，为新存储卷申请分配相应的存储空间。

 版权所有 © 超聚变数字技术股份有限公司



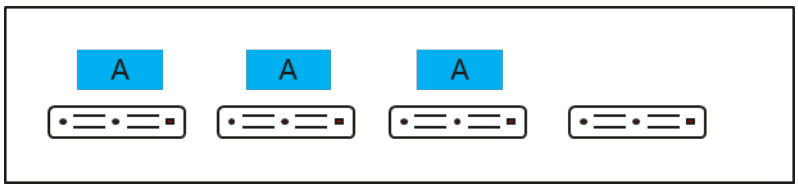
5.16 EC 块存储

FusionOne Storage HBS支持两种主流的数据冗余技术：副本和纠删码（EC，Erasure Coding）。它们的核心区别是用空间换计算（副本），还是用计算换空间（EC）。

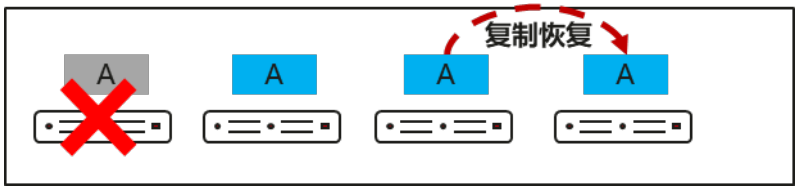
EC N+M 纠删码：

1. 将数据切为 N 个数据分片，通过计算生成 M 个校验分片，共 N+M 个分片。
2. 任意丢失 ≤ M 个分片都能恢复。
3. 存储空间利用率为N / (N+M)，例如 EC 4+2的空间利用率66.7%。

副本模式下（三副本为例），每个节点都有一个完整的数据副本。数据放置方式如下：

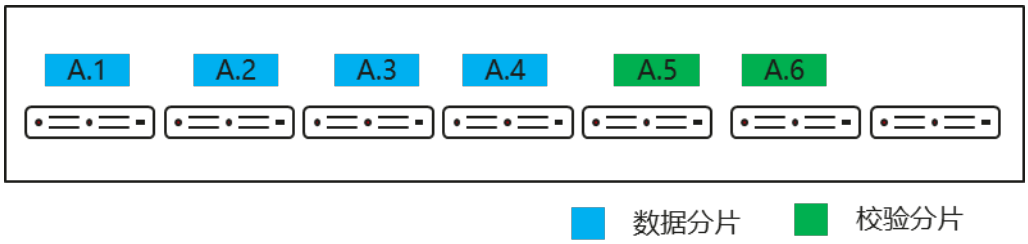


如果节点发生故障，数据副本会通过复制的方式，恢复到其它正常节点：

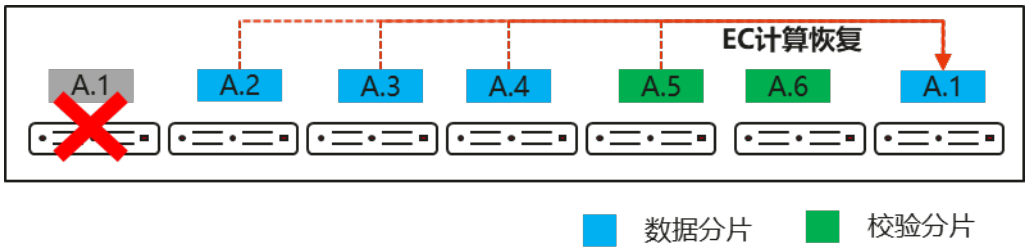


EC模式下（以EC4+2为例），数据被切分成4个数据分片，然后通过EC算法计算出2个校验分片，总共6个分片，每个节点上只有一个分片、没有完整数据。数据放置方式如下：

 版权所有 © 超聚变数字技术股份有限公司



如果节点发生故障，数据分片会通过EC计算的方式，恢复到其它正常节点：



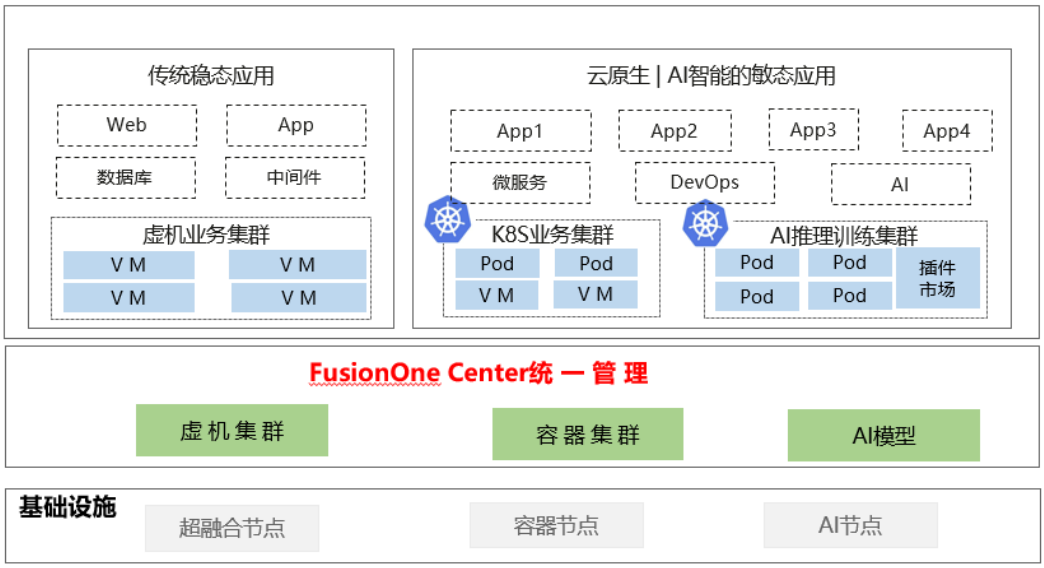
FusionOne Storage HBS的EC冗余，具有如下特点和优势：

1. 支持EC 2+2、4+2、6+2、8+2，根据节点数量配置。最少3节点即可部署。
2. 支持EC折叠N+M:1，节点数不足时自动折叠、节点数增加时自动展开。例如：EC 4+2通常需要6个节点、每个节点放置1个数据分片；而EC 4+2:1可以在1个节点上放置2个数据分片，总共只需要3个节点即可。
3. 相比三副本33%的空间利用率，EC 4+2可以达到66%、EC 8+2可以达到80%。

 版权所有 © 超聚变数字技术股份有限公司

6 智算资源管理扩展

FusionOne HCI支持在一个管理平台支持通算（虚机）和智算（容器与AI）双栈管理，支持在原通算平台平滑扩展智算业务



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 技术白皮书 01》。

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

 版权所有 © 超聚变数字技术股份有限公司

7 统一管理

7.1 设备管理
7.2 裸金属资源管理
7.3 监控管理
7.4 运维管理
7.5 多云&多站点管理

7.1 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。

7.1.1 可视化超融合节点

支持超融合节点自动纳管，支持超融合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。
- 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。

 版权所有 © 超聚变数字技术股份有限公司

- 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。

图 7-1 服务器信息



图 7-2 真实视图



7.1.2 网络设备纳管

支持纳管超融合网络设备，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。
- 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。

版权所有 © 超聚变数字技术股份有限公司

图 7-3 部件信息



图 7-4 真实面板



7.1.3 操作系统纳管

支持纳管超融合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1：

表 7-1 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 © 超聚变数字技术股份有限公司

图 7-5 基本信息

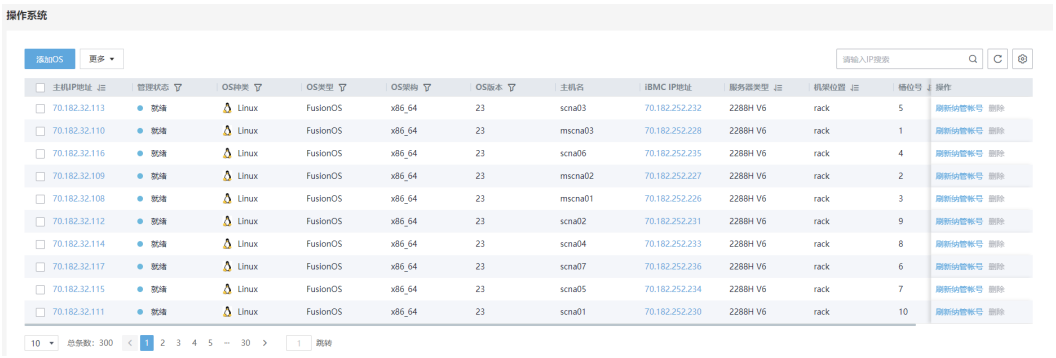
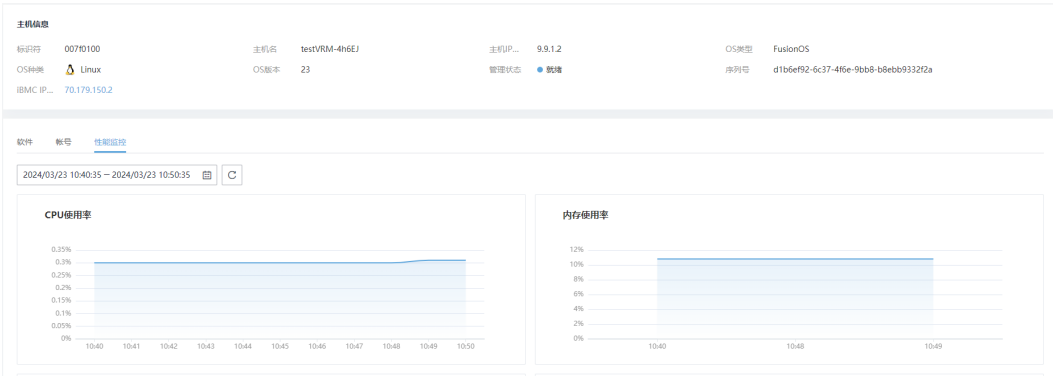


图 7-6 性能监控



7.2 裸金属资源管理

裸金属服务器是用户专属的计算资源，继承物理服务器特征，无虚拟化开销和性能损失，100%释放算力资源，为用户上线核心数据库、关键应用系统、高性能计算、大数据、AI等业务提供卓越的计算性能以及数据安全。

版权所有 © 超聚变数字技术股份有限公司



主要功能如下：

- 资源池管理
  - 资源池管理：支持创建资源池、删除资源池、修改资源池， 资源池是一个逻辑的分组概念，不区分服务器厂商和架构。
  - 资源池统计：支持以列表展示资源池所含的裸金属总数量、CPU总容量、内存总容量、磁盘总容量、显卡总数量、显卡总容量，以及资源可用情况，如可用CPU、可用内存、可用磁盘等。
  - 资源池监控：支持查看资源池的CPU使用率、内存使用率、磁盘使用率等指标的历史趋势。如果裸金属存在GPU算力卡，则支持查看GPU或NPU的使用率指标历史趋势。
- 裸金属管理
  - 裸金属生命周期管理：支持添加裸金属、删除裸金属、占用/释放裸金属、设置裸金属标签、刷新纳管账号密码等。暂时不支持为裸金属自动安装操作系统，因此在添加裸金属前，需要确保已完成在裸金属上手动安装了操作系统。
  - 裸金属监控：在安装node-exporter监控组件后，支持查看裸金属的CPU使用率、内存使用率、磁盘使用等指标的历史趋势；如果裸金属上有英伟达GPU

 版权所有 © 超聚变数字技术股份有限公司

卡且在安装了dcgm-exporter后，支持查看GPU使用率和GPU显存使用率等指标的历史趋势；如果裸金属有NPU卡且在安装安装了npu-exporter后，支持查看NPU显存使用率、NPU AI CPU使用率、NPU AI Core使用率等指标的历史趋势。

- 指标保存周期：默认14天，最多可保存90天，超过14天，需要扩大FOC虚拟机的内存和磁盘空间，请联系技术人员评估。
- 兼容操作系统：FusionOS、Ubuntu。
- 兼容硬件范围：超聚变服务器、KunLun服务器、三方服务器如华为、浪潮、H3C等品牌。

7.3 监控管理

7.3.1 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HCI软硬件各子系统生成的告警，在FusionOne Center界面统一展示。
- 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。

7.3.2 可视化性能监控

监控计算集群、计算节点、虚拟机、存储集群、存储节点、存储池、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HCI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HCI监控仪表盘，从多维度展示HCI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。

7.3.3 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。

 版权所有 © 超聚变数字技术股份有限公司

- 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）、2288 V8、2188 V8等
- 常用操作：电源上下电、UID灯开关、硬盘灯开关

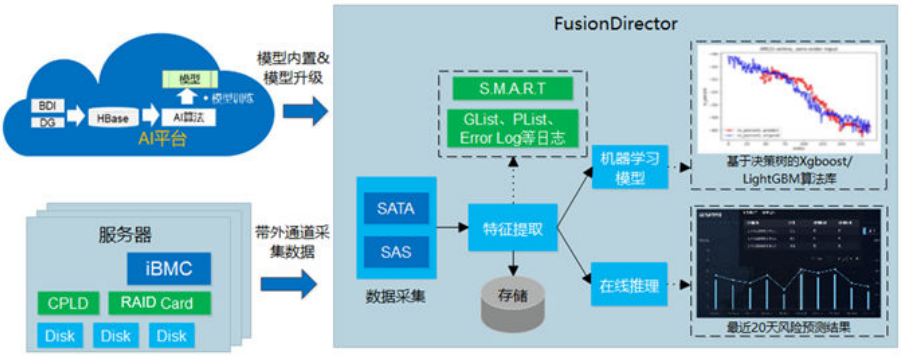


7.3.4 硬件故障预测

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。

7.3.4.1 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，超融合平台支持对超聚变自研品牌服务器SAS/SATA机械盘、SSD盘的故障预测。



 版权所有 © 超聚变数字技术股份有限公司

1. 硬盘数据采集
   - RAID带外采集通道。

     数据流向为“Disk > RAID > iBMC > FusionOne Center”，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

     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。

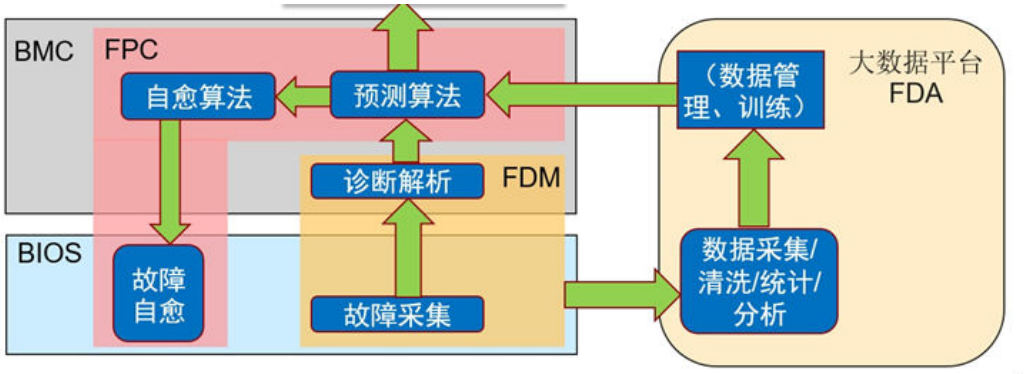
       为了兼顾准确率和维护可用性，FusionOne Center采用提前20天对风险盘进行预测，预测准确率可以达到80%以上（业界标准算法，FDR：提前识别的故障盘/总故障盘数目），误报率小于0.2%（业界标准算法，FAR：误报的盘/总好盘个数）。
   - 离线故障诊断

     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 © 超聚变数字技术股份有限公司

7.3.4.2 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。

7.4 运维管理

FusionOne HCI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 HCI提供自动化安装部署工具FusionOne Installer；运维管理方面FusionOne HCI提供了统一的运维管理平台FusionOne Center管理工具。

7.4.1 安装部署

7.4.1.1 自动化一键部署

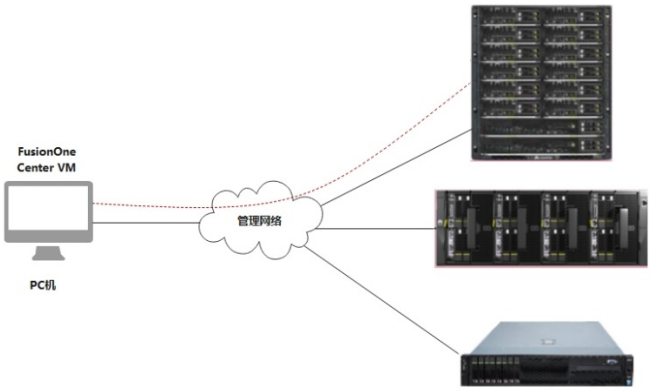
FusionOne HCI提供工具安装部署，挂载镜像安装两种方式，用户可灵活选择。

1. 挂载镜像部署是通过将超融合的软件包ALL IN ONE，可以像安装普通操作系统一样部署超融合节点，可大大简化超融合的部署难度。

 版权所有 © 超聚变数字技术股份有限公司

2. 批量安装使用FusionOne Installer可批量完成系统软件的安装。同时配套初始化配置向导功能，只需要完成系统基本参数配置后，就自动完成各节点网路配置，管理集群和存储集群的创建。

图 7-7 批量部署系统示意图



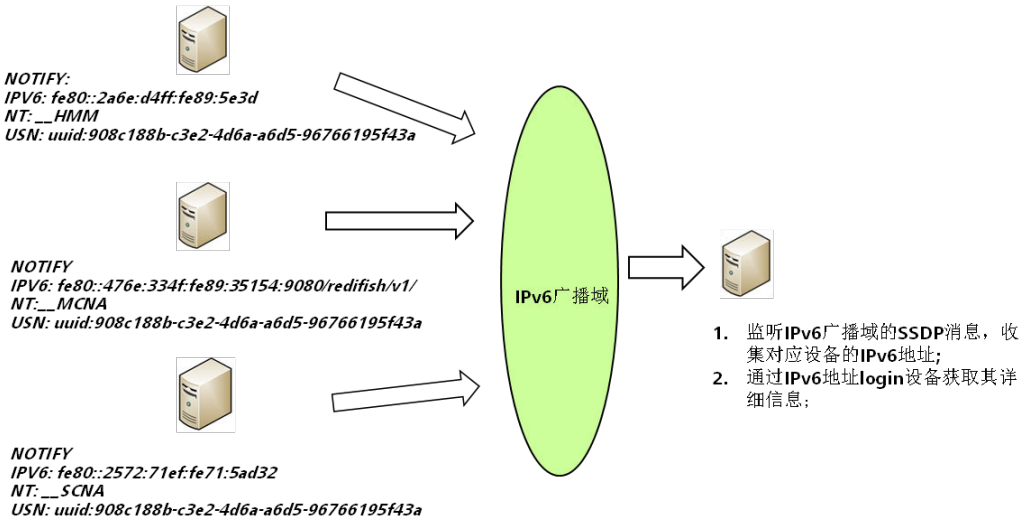
- FusionOne Installer安装工具可部署在PC机或者虚拟机中。
- FusionOne Installer通过Simple Service Discovery Protocol（SSDP）简单服务发现协议或者扫描IP方式发现服务器，读取服务器信息。
- FusionOne Installer连接服务器BMC，使用KVM挂载光盘功能进行引导，启动安装。最大支持8个节点并行安装。
- 安装过程中使用NFS共享传送安装软件包和配置。

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。

7.4.1.2 自动发现原理

FusionOne HCI系统在系统安装、初始化和扩容过程中均支持设备自动发现。设备自动发现通过Simple Service Discovery Protocol（SSDP）服务实现。

图 7-8 SSDP 设备自动发现示意图



 版权所有 © 超聚变数字技术股份有限公司

SSDP设备自动发现流程如下：

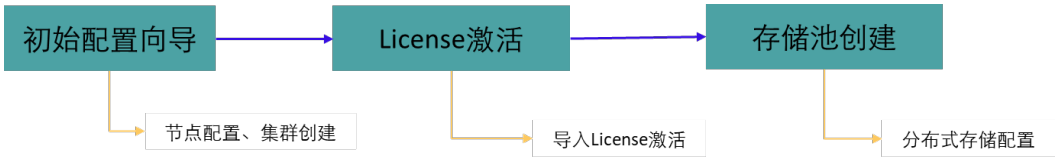
1. 超融合节点的服务器BMC中内置SSDP服务，设备上电后会自动通过SSDP服务广播主机的链路IPv6地址消息。
2. Installer中部署了SSDP服务端，负责监听IPv6本地广播域中的SSDP消息，收集相应设备的IPv6地址；
3. 通过IPv6地址登录设备，获取相应服务器的详细信息。

初始化和扩容过程中的SSDP设备自动发现流程如下：

1. 系统安装后在管理虚拟机、主机操作系统中已经内嵌有SSDP客户端，SSDP客户端自动通过IPv6地址广播SSDP消息。
2. FusionOne HCI初始配置向导中部署了SSDP服务端，负责监听管理平面IPv6广播域中的SSDP消息，收集相应设备的IPv6地址。
3. 通过IPv6本地地址登录设备，获取相应设备的详细信息。

7.4.1.3 出厂预安装，开箱即用

支持出厂灌装超融合软件，出厂预安装完全覆盖了使用FusionOne Installer的效果，支持主力发货机型2288H V6、2288H V7、KUNLUN机型的融合节点、计算节点的出厂预安装软件。节点发货到客户现场后，无需进行软件的安装部署，只需要将服务器上电、开箱即可根据现网配置业务。以下是预安装交付流程：



服务器上架后，选择任意主机配置IP地址后，可直接页面向导式完成业务网络、集群参数，省去了在现场使用FusionOne Installer部署软件的过程，无需携带任何软件工具现场安装，仅仅浏览页向导即可完成集群的配置。

7.4.2 初始配置

为了方便用户部署，超融合节点每个节点均启动了初始配置向导服务，可以登录任意节点即可进行初始化。

系统初始化流程如下：

1. 进入系统初始化界面，系统自动发现设备，并在初始化参数配置页面显示节点相关信息。
2. 完成初始化参数（包括网络参数和存储池参数）配置，参数校验通过后则可进行系统初始化构建，整个初始配置构建约为50分钟做，期间无人值守，可自动完成。
3. 系统自动完成初始化，包括各节点的网络配置，管理集群和存储集群、计算集群创建，并将主机加入计算集群。

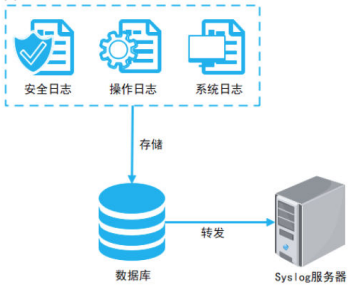
系统完成初始化后即具备VM发放功能，只需要完成License导入，存储配置、业务网络配置、NTP等配置即可交付使用。

 版权所有 © 超聚变数字技术股份有限公司

7.4.3 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 Center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HCI各子系统日志的一键收集，包括：硬件（BMC、SSD、NIC等组件）、分布式存储、FusionOne Compute、OS、FusionOne Center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。

7.4.4 一键式运维

一键式运维功能作为FusionOne HCI产品的核心功能，提供用户更为高效、自动化的运维管理功能特性，主要提供了一键式扩容、一键式巡检、一键式日志收集等功能。需要注意，由于第三方服务器自身能力不足的原因，导致在一键式扩容、一键式巡检、一键式日志收集以及升级等一键式运维操作中，无法对第三方服务器硬件设备进行运维操作。

7.4.4.1 一键日志收集

在FusionOne Center管理界面上集成日志收集功能，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：硬件（BMC、SSD、NIC等组件）、分布式存储、FusionOne Compute、OS、FusionOne Center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 Center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。

7.4.4.2 一键健康巡检

在FusionOne Center管理界面上集成系统健康检查功能，能一键式对于系统的各个部件以及节点进行相应的系统排查，检查系统的健康状态，是否存在健康风险或故障，提供一定的排查建议，最终输出相应的巡检报告。为用户高效运维提供指导。

版权所有 © 超聚变数字技术股份有限公司

- 巡检策略：支持手动巡检和定时巡检（周期可自定义，精确到秒），支持配置巡检模板。
- 灵活选择巡检项：支持面向FusionOne Center管理面、FusionOne Compute、FusionOne Storage、计算节点、存储节点设置不同巡检项指标。
- 健康打分：支持根据巡检结果智能评分，异常巡检项给出可能原因和处理建议
- 巡检报告：支持在线查看巡检结果报告和通过邮件给客户发送巡检结果报告。

7.4.4.3 一键系统升级

FusionOne Center提供整系统在线升级能力，升级过程中，充分利用节点滚动和虚拟机热迁移特性，实现客户业务不中断的在线升级，最大化降低升级过程中对客户业务的影响。

- **升级包管理**：支持对升级包进行管理，如本地上传，删除，查询等。
- **智能版本管理**：基于发布的产品基线，FuisonOne Center能够智能管理升级包配套关系，避免出现升级版本不配套造成的升级事故。
- **向导式升级**：对FusionOne Center管理面、FusionOne Compute(VRM、CNA、VMTools) 、FusionOne Storage子系统的升级，提供页面向导式配置升级的能力，为用户提供快捷高效、安全可靠的升级体验。
- **升级模式**：为应对不同升级场景需求，超融合提供在线升级、离线升级两种模式。在线升级可保证业务不中断下进行滚动式升级，而离线升级则需要停止业务虚拟机，支持并发升级，升级速度快，最大支持10节点并发升级。
- **融合升级：**支持CNA和FusionOne Storage的融合升级，虚拟机最多迁移1次，缩短整体升级时长。
- **支持热补丁升级**，某些升级版本如果是热补丁，支持不重启快速升级。

7.4.4.4 一键系统扩容

FusionOne HCI系统扩容节点，将待扩容节点上架，摆放和插线，启动待扩容节点，将节点网络配置与已有系统网络对接好。节点扩容前准备好后，在FusionOne Center扩容界面可通过SSDP扫描将待扩容节点发现，完成相应的系统配置，包括：IP地址、主机名、网关、存储池等参数，单击进行**校验**是否待扩容节点符合系统要求，网络配置是否正常等。

完成校验后，单击**扩容**按钮，进行系统扩容操作，将待扩容节点加入系统集群中。

7.4.4.5 一键系统下电

在FusionOne Center管理界面上，能一键式对FusionOne HCI 各个子系统执行安全下电操作，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HCI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HCI各子系统自动执行安全下电操作，用户无需额外干预。

 版权所有 © 超聚变数字技术股份有限公司

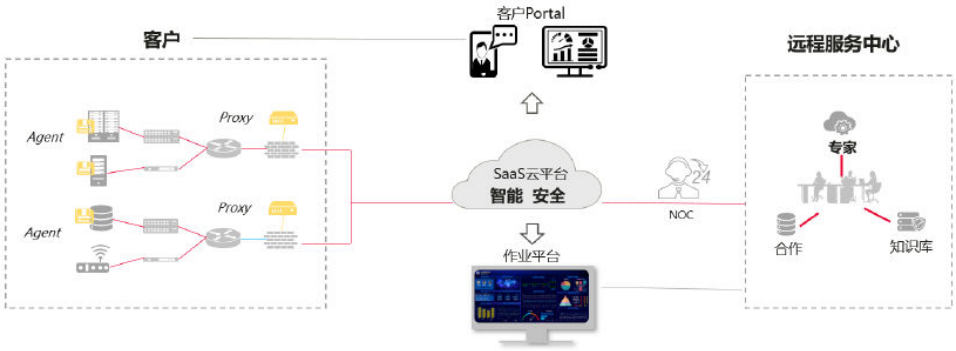
7.4.4.6 一键资源优化

在FusionOne Center管理界面上，能一键式对FusionOne HCI 关键资源（虚拟机、模板、镜像）进行主动、被动的扫描，对闲置长期不使用的、使用效率低的资源自动扫描，便于管理员快速找到闲置资源。管理员可根据闲置情况，一键清理掉占用的资源。

7.4.5 远程运维和移动运维

借助超聚变FusionOnline云端智能运维平台及其FusionMobile移动端，客户可实现对HCI整系统运维的远程化、自动化、智能化。支持客户7*24小时，随时随地看护HCI整系统运行状况，有效缩短故障处理时间，全面提升运维效率。

- Call HOME服务：在取得客户授权的情况下，FusionOne Center支持将超融合软件系统的日志及告警、服务器设备告警等自动上报到超聚变云端智能运维平台FusionOnline。
- 故障处理可视化：FusionOnline告警和工单联动，在FusionOnline上，用户可直接查看工单处理进展。客户可以任何时间、任何地点、任何PC或手机访问FusionOnline。

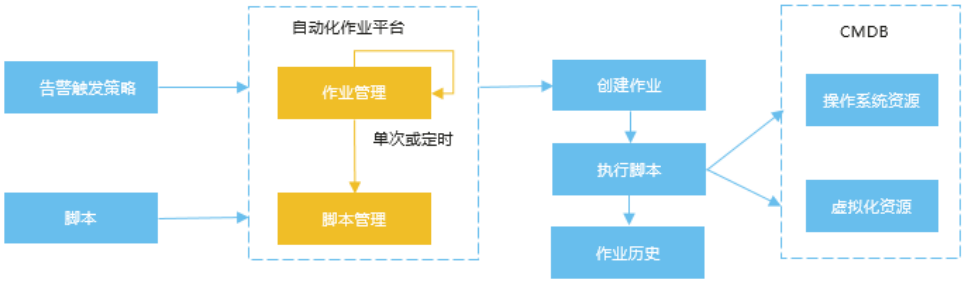


7.4.6 自动化作业

FusionOne HCI的运维自动化平台，提供从HCI的基础设施到虚拟化资源的全栈自动化运维能力。旨在帮助企业运维团队大幅提高运维效率。

- 丰富场景化：内置了丰富的场景化运维操作库，满足日常运维需求。如故障信息收集、主机诊断命令、文件分发、虚拟机创建、虚拟机扩资源等。
- 高效运维：支持对资源批量执行脚本完成日常运维操作，提高运维效率。
- 触发策略：支持手动触发、定时触发、基于规则的告警触发。
- 安全可控：提供完善的安全控制机制，事前有安全提醒，事后可审计操作日志，避免人为的操作安全风险。

版权所有 © 超聚变数字技术股份有限公司



7.4.7 任务中心

FusionOne HCI 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。

7.5 多云&多站点管理

7.5.1 多云管理

在现代IT架构中，组织经常需要同时管理包括本地各种品牌的超融合环境，如VMware、FusionCube、深信服HCI等多种异构超融合平台。为了提升管理效率，FusionOne HCI内置集成了多云管理功能。实现对自身站点及VMware等不同异构超融合平台的高效管理。

- 统一管理界面：FusionOne HCI23支持对VMware VCenter、FusionCube、FusionOne HCI22、SANGFOR HCI等异构超融合平台的纳管，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个异构超融合站点。
- 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 HCI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 HCI 平台上运行。

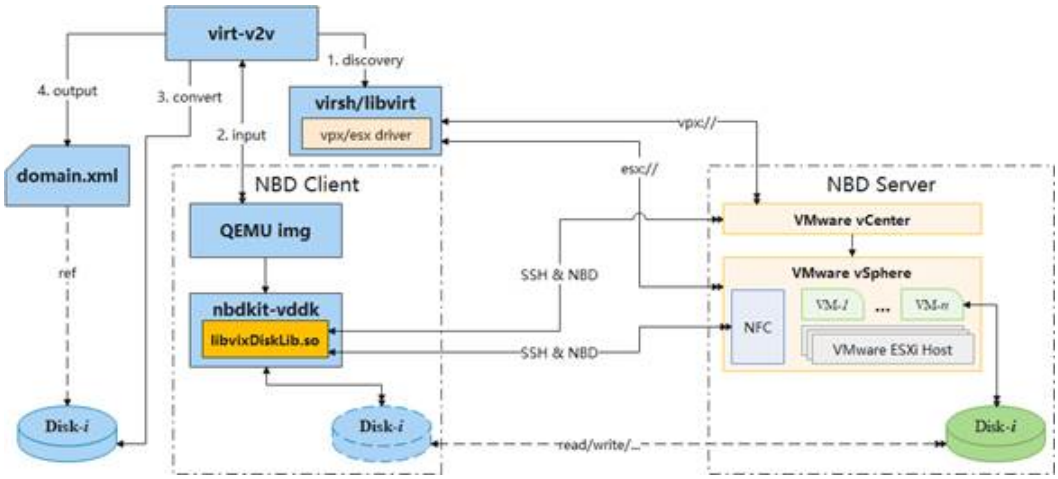
- V2V温迁移：支持V2V迁移，迁移过程中，无需经历导出/导入虚拟机的过程。迁移过程中不停机，迁移完成后自动关闭VMware上的虚拟机，实现虚拟机温迁移。

 版权所有 © 超聚变数字技术股份有限公司

- 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：

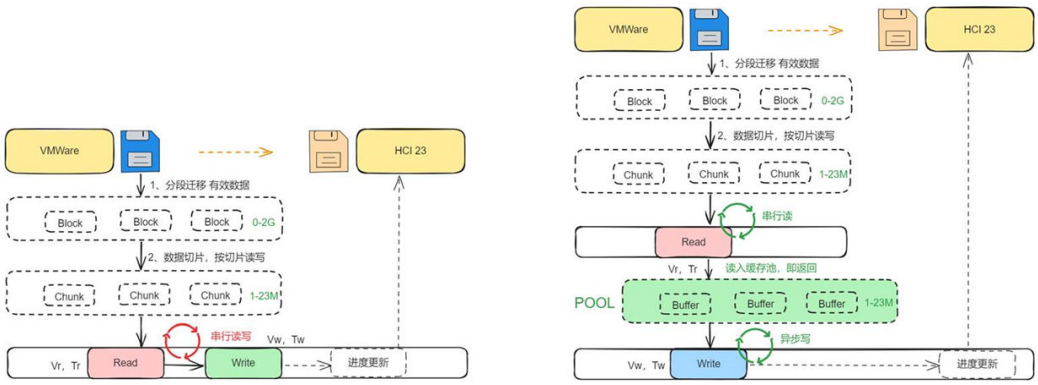


Virt-v2v通过virsh连接VMware服务（vCenter或ESXi），获取源虚拟机信息（配置文件、磁盘、快照等）；通过QEMU NBD客户端，创建源网络磁盘；然后通过磁盘复制、转换过程，把源磁盘/快照复制到本地；把源虚拟机的配置文件转换为目标格式；通过libguestfs转换磁盘数据，同时替换磁盘路径、网络映射。磁盘数据转换主要包括：安装VirtIO驱动程序、更新/etc/fstab和xorg.conf文件、重建initrd文件等。

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

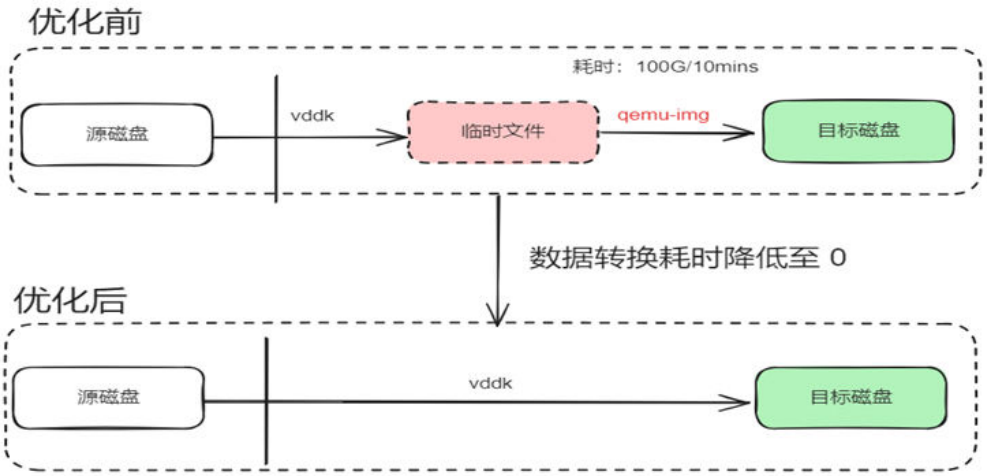
原理如下图所示：



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。

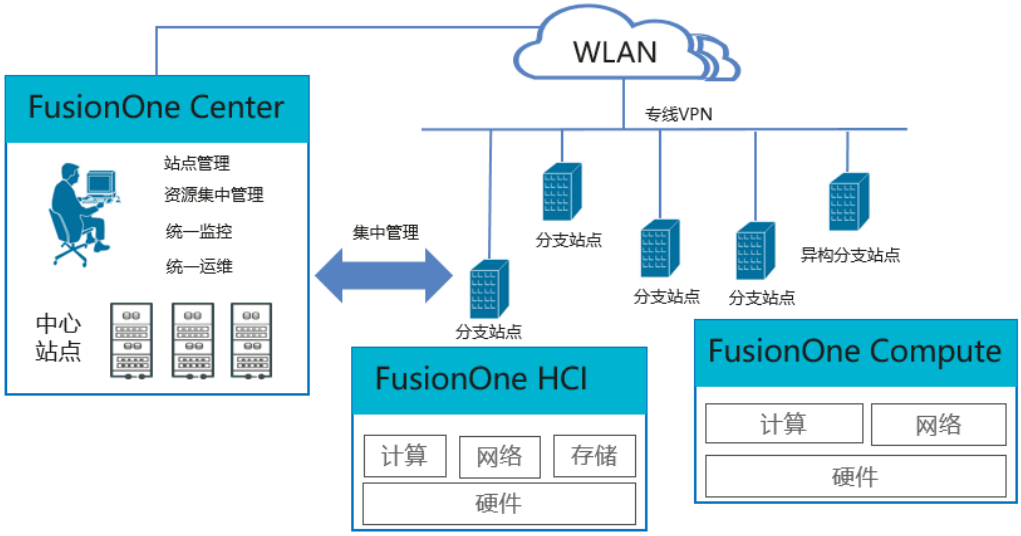
 版权所有 © 超聚变数字技术股份有限公司



7.5.2 多站点管理

FusionOne HCI 支持同构站点的多站点管理能力，无需额外部署管理平台。客户可根据实际需求，将任意FusionOne HCI站点转换为中心站点，添加纳管下级FusionOne HCI的站点，形成一个中心站点统一管理下属的站点的形态。方便运维人员，在中心站点对所有的下级站点进行管理运维。

- 精简中心站点：中心站点只有FusionOne Center 管理系统，无计算虚拟化子系统和存储子系统。只具备多站点管理功能，本级站点不具备HCI的其他功能。
- 高配中心站点：中心站点自身是完整的FusionOne HCI，具备计算虚拟化子系统和存储子系统。支持在中本级站点使用全套的HCI功能。
- 规格： 最多支持500个下级站点、最多管理20000个虚拟机。

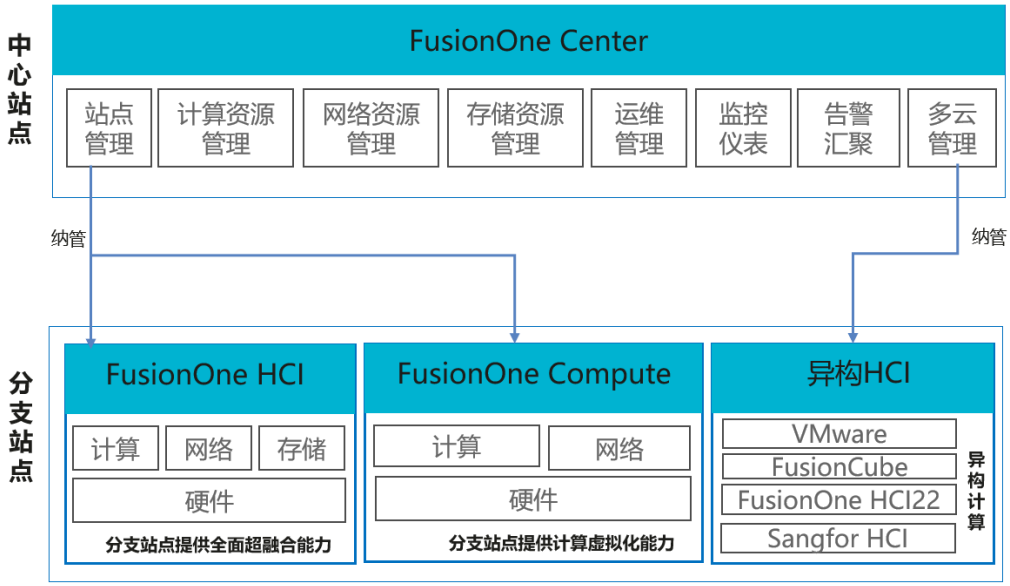


中心站点纳管分支站点后，支持通过以站点切换的方式，对下属分支站点进行管理和运维。

- 计算资源管理：集群管理、主机管理、虚拟机生命周期管理、虚拟机快照和备份、虚拟机模板管理、GPU资源组管理等
- 网络资源管理：分布式交换机管理、虚拟私有云、安全组、MAC地址池、网络拓扑等

 版权所有 © 超聚变数字技术股份有限公司

- 存储资源管理：数据存储资源管理、磁盘管理、镜像仓库管理等
- 运维管理：日志收集、资源优化、并且可以对镜像文件、虚拟机模板、应用模板跨站点同步等
- 告警数据汇聚：在中心站点告警列表，支持查看所有站点的告警数据，并支持以站点进行告警检索。
- 审计日志汇聚：在中心站点所有对下级站点的操作，记录审计日志，方便运维审计。
- 站点统计仪表：在中心站点首页提供以站点维度的统计仪表，包括：站点告警TOP5、站点CPU占用率TOP5、站点内存利用率TOP5、站点存储容量占用率TOP5、站点告警列表、站点告警趋势等。



 版权所有 © 超聚变数字技术股份有限公司

8 云管

8.1 概述
8.2 产品架构
8.3 关键特性

8.1 概述

FusionOne云管平台是一个综合基础设施资源运维管理平台，帮助企业用户更有效率地管理和使用分布在混合云、多云场景下的IT基础设施资源。通过集成可自定义的业务审批流程、多租、计费等特性，企业客户能够对来自不同云平台的各种资源进行合理、有效地分配与使用。极大的规避了各种虚拟化管理平台差异性带来的使用不便，实现多云统一融合，简化使用多云的复杂度，提高多云管理效率，为企业客户提供全方位的云资源供给、运维和运营管理能力。

8.2 产品架构

8.2.1 功能架构

FusionOne HCI 云平台是一款面向政企数据中心的云管理平台，构建“云服务、云运维、云运营”一体化的管理体系，为企业用户提供按需、自助、敏捷、弹性地云服务使用体验。

 版权所有 © 超聚变数字技术股份有限公司



易安装

- 超融合和云两种软件形态原生共存，云是超融合软件的一个安装形态，共用超融合安装工具一键安装部署。

多角色管理视图

- 1个软件实例提供运维视图+运营视图两个管理面，系统管理员、运营管理员、组织管理员、项目管理员、普通租户各司其职，实现权限可控、资源隔离、流程协同和数据分视角呈现。

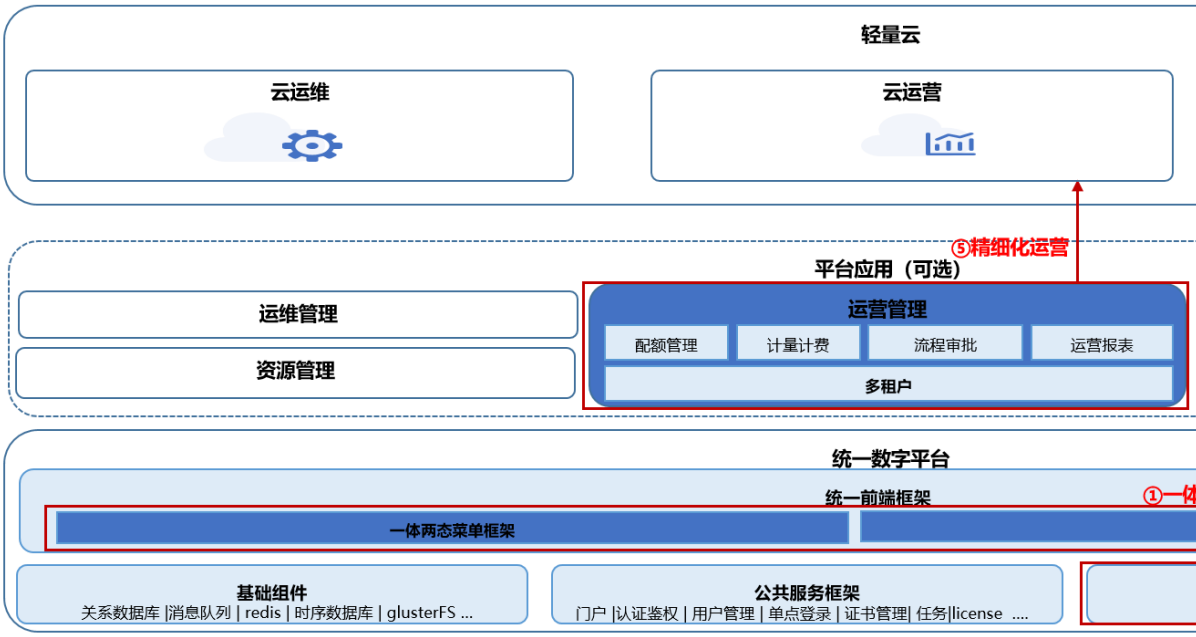
自服务和精细化运营

- 提供基础 IaaS 资源的自助申请和受控审批，结合多租户隔离、资源配额、计量计费等能力，实现云资源按需供给、规范使用与精细化运营。

8.2.2 技术架构

- **一体两态架构**：底层与超融合共用1个管理软件，超融合和云交互形态全兼容，提供多角色管理视图。

版权所有 © 超聚变数字技术股份有限公司



8.3 关键特性

8.3.1 虚拟机

虚拟机是采用虚拟化技术构建的运行在宿主机上的实例。

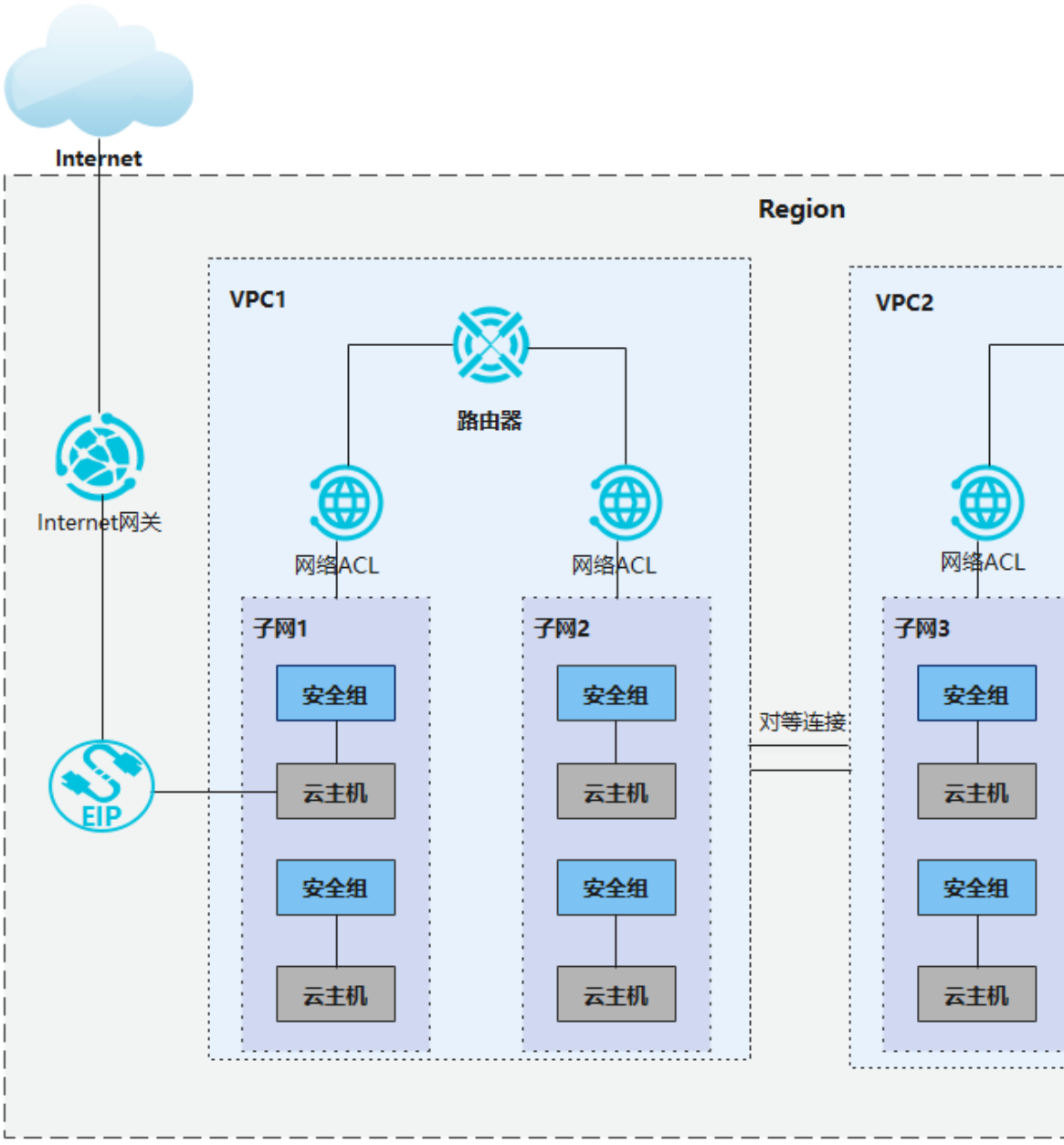
- 支持对虚拟机基本的生命周期管理动作，包括创建、删除、开机、关机、重启、挂起、恢复、更改项目、设置共享、调整资源配置规格、挂载ISO、卸载ISO、迁移、查看详情等。
  - GPU虚拟机：如果云管所纳管的FusionOne HCI 23平台的计算节点上有GPU设备且创建了GPU资源组，则可在云管平台上使用这个超融合站点的资源创建GPU虚拟机。
- 支持对虚拟机进行快照管理。
- 虚拟机规格：用户可以创建Flavor的规格，用于定义一种虚拟机类型（内存、CPU、GPU等），由系统管理员创建，供普通用户在创建虚拟机时使用。

8.3.2 云网络

8.3.2.1 VPC

虚拟私有云（Virtual Private Cloud）是一个逻辑隔离的网络空间，可以为租户虚拟机提供私密的虚拟网络环境。每个租户可拥有一个或多个VPC，VPC之间网络隔离，VPC支持丰富的网络管理功能，包括子网、路由表、安全组、网络ACL等。同时可以通过弹性公网IP和NAT网关连通VPC和公网网络。

 版权所有 © 超聚变数字技术股份有限公司



8.3.2.2 弹性公网 IP

弹性公网IP（Elastic IP，简称EIP）提供独立的公网IP资源，包括公网IP地址与公网出口带宽服务。可以与虚拟机、NAT网关等资源灵活地绑定及解绑。弹性公网IP需要从弹性公网IP池中申请使用，弹性公网IP池是应对批量弹性公网IP的管理难题，提升弹性IP管理效率的解决方案。弹性公网IP池可以记录和管理弹性公网IP的地址段、范围及使用情况，拥有IP子网的管理能力，因此弹性公网IP池是一种特殊的IP子网。

8.3.2.3 NAT 网关

NAT网关为虚拟私有云（VPC）内的虚拟机提供网络地址转换服务，使多个虚拟机可以共享使用弹性公网IP访问Internet，节省租户的弹性公网IP资源。

NAT 网关支持配置SNAT 和DNAT 规则。SNAT用于虚拟机访问Internet时，将虚拟机的私有网络地址转化为弹性公网IP地址以支持Internet访问；DNAT用于租户通过

 版权所有 © 超聚变数字技术股份有限公司

Internet访问虚拟私有云内部虚拟机是，将弹性公网IP地址和端口转化为VPC的私有网络地址和端口以支持Internet访问虚拟机。

8.3.2.4 网络 ACL

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。

网络ACL支持与子网绑定，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。

表 8-1 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。
应用操作	创建子网没有网络ACL选项，需要先创建网络ACL，添加出入规则，并在网络ACL内关联子网。	• 创建虚拟机时，可选择关联安全组，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。

8.3.2.5 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 © 超聚变数字技术股份有限公司

- 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。

8.3.3 云运营

8.3.3.1 组织与项目管理

FusionOne 云平台支持组织+项目的矩阵式管理方案，更贴近客户业务运作的实际情况。

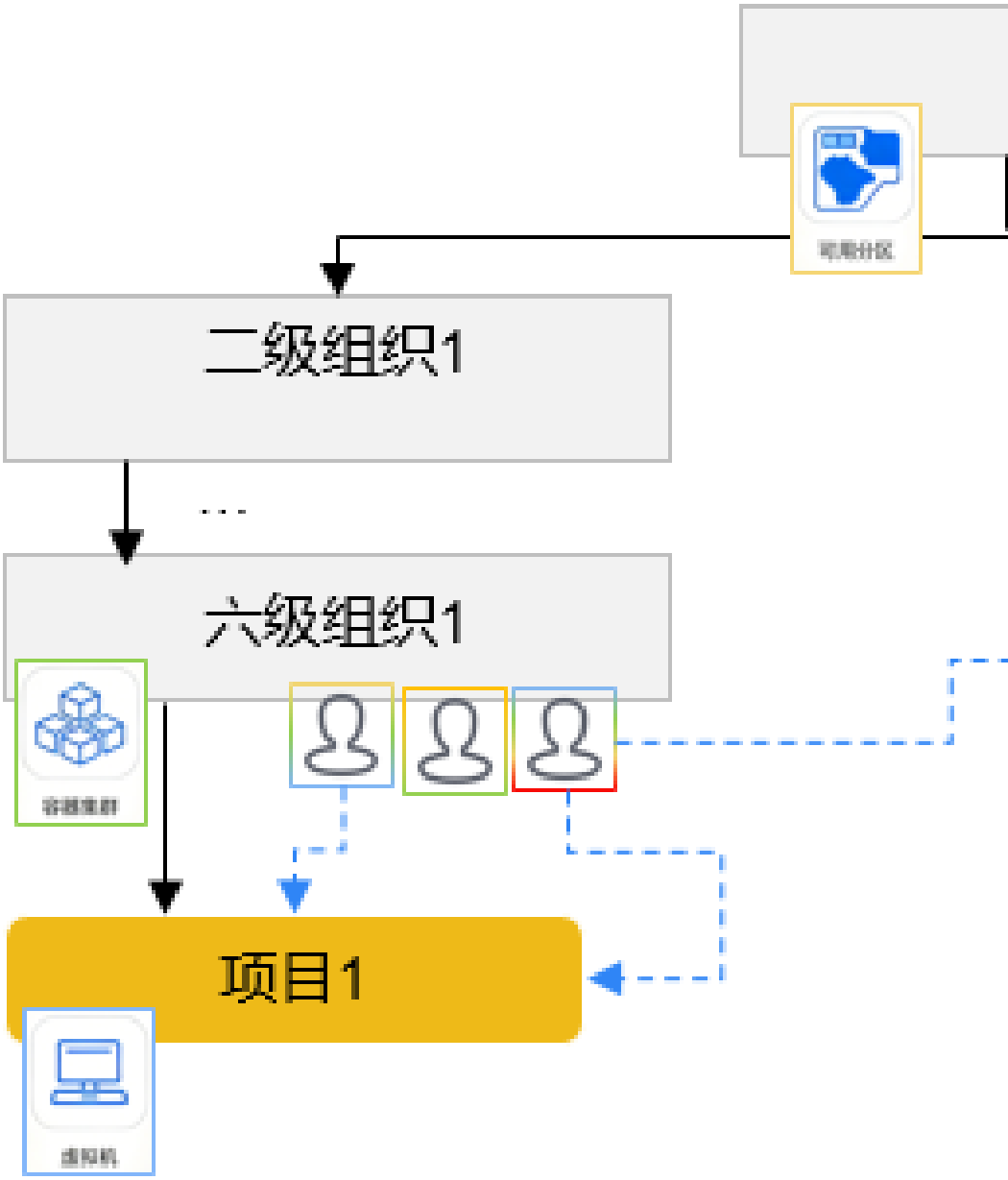
组织管理

- 分为三级：租户、组织、项目
- 支持多级组织创建，最大支持6级
- 用户只能属于一个组织，支持变更组织
- 默认创建根组织，作为系统用户的组织归属

项目管理

- 项目为一组资源的载体
- 一个项目属于且仅能属于一个组织
- 云资源归属于项目

 版权所有 © 超聚变数字技术股份有限公司

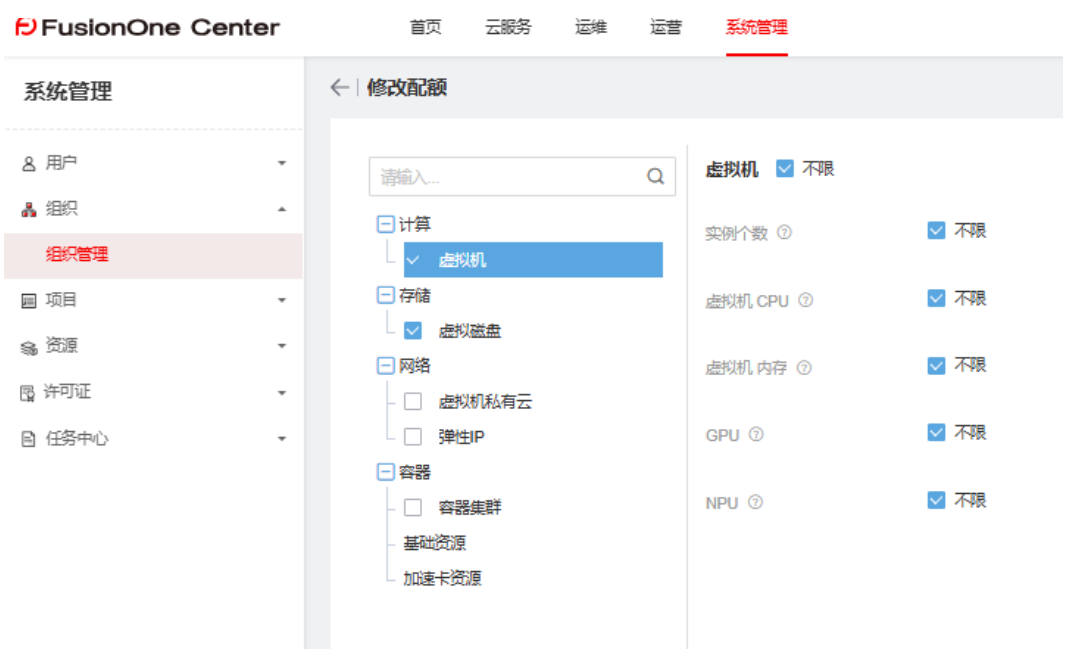


8.3.3.2 配额管理

在使用云管系统时，为了对有限的资源进行合理地规划和使用，需要适当限制每个组织、项目可使用的资源数量。额项管理提供配置某个组织、项目可操作的某种资源（CPU、内存、硬盘、网络等虚拟化相关的资源；用户、用户组、角色等系统资源）的数量的能力。

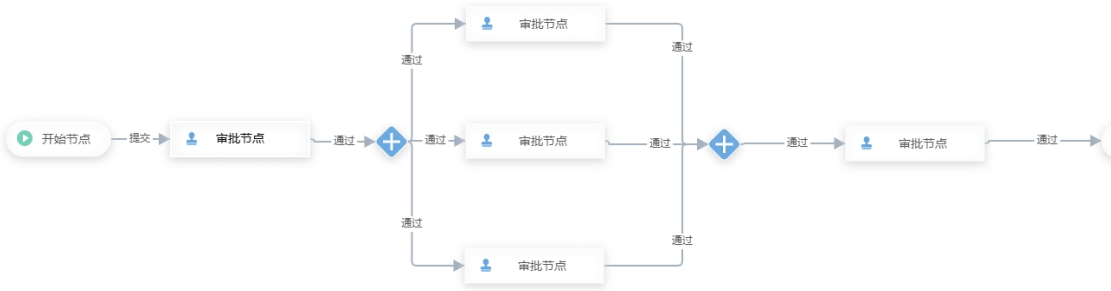
- 定义了一个工作空间（组织或项目）能获取资源的上限。
- 按照组织和项目的层级关系，对配额进行逐级控制。

版权所有 © 超聚变数字技术股份有限公司



8.3.3.3 流程订单

用户申请虚拟机、硬盘后，自动生成订单。订单提交给相应的审批人员进行审批。审批通过后，系统自动执行订单任务。订单的审批流程也可根据实际的组织架构来定义，方便客户的个性化使用。



8.3.3.4 计量计费

随着云计算的广泛应用，企业和组织在使用云资源（如 虚拟机、虚拟磁盘、GPU/NPU等）过程中，需要清晰了解资源的使用成本。不同类型的资源具有不同的计费模式，这要求有一个专门的计量计费系统来准确计算费用。企业内部通常存在多种组织架构、多个项目并行开展以及众多用户使用云资源的情况。不同组织、项目和用户对资源的需求和使用方式各不相同，这就需要计量计费系统能够支持基于这些维度的账单统计、查看和导出功能。

费率设置

 版权所有 © 超聚变数字技术股份有限公司

- 针对虚拟机，按照flavor规格设置单价，支持设置开机价格和关机价格

  1）默认费率推导公式：

  (cpu核心数 * 0.2 + 内存容量GiB * 0.05) + (fp8算力 * 0.25 + 显存GiB * 0.10) * 定位系数 [* GPU个数] [/ GPU最大实例数]

  2）定位系数：模组1.5 消费级0.8 标卡1.0
- 针对虚拟磁盘，按照存储容量，灵活设置单价
- 针对容器集群，按照资源个数，灵活设置单价
- 针对NAT网关、公网IP，按照资源格式，灵活设置单价

费用明细

- 支持从组织、项目、用户三个维度查看账单费用概览
- 在组织层面，运营管理员或者租户管理员管理员可以了解不同组织或业务单元在云资源使用上的成本分布
- 项目维度的账单展示有助于项目负责人实时监控项目成本进
- 用户维度的账单则使每个云服务使用者能够清楚知晓自己的资源使用费用
- 支持最多查询一年前至今的账单概览

账单明细

- 支持从组织、项目、用户、资源名称、资源类型筛选账单明细
- 账单明细包含资源名称、资源类型、ID、消费金额、组织、项目、用户、开始结束时间等详细信息。
- 支持最多查询最近半年的账单明细。
- 系统支持将账单明细导出为 Excel，方便用户进行进一步的数据分析和处理。

账单核销

- 支持运营管理员按照月度维度，核销某个租户的账单。

8.3.4 云运维

8.3.4.1 审计日志

提供高效的审计日志记录、审计日志查询、审计日志导出等功能，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。

支持对所有页面操作记录审计日志，以及对审计日志的查询、导出。

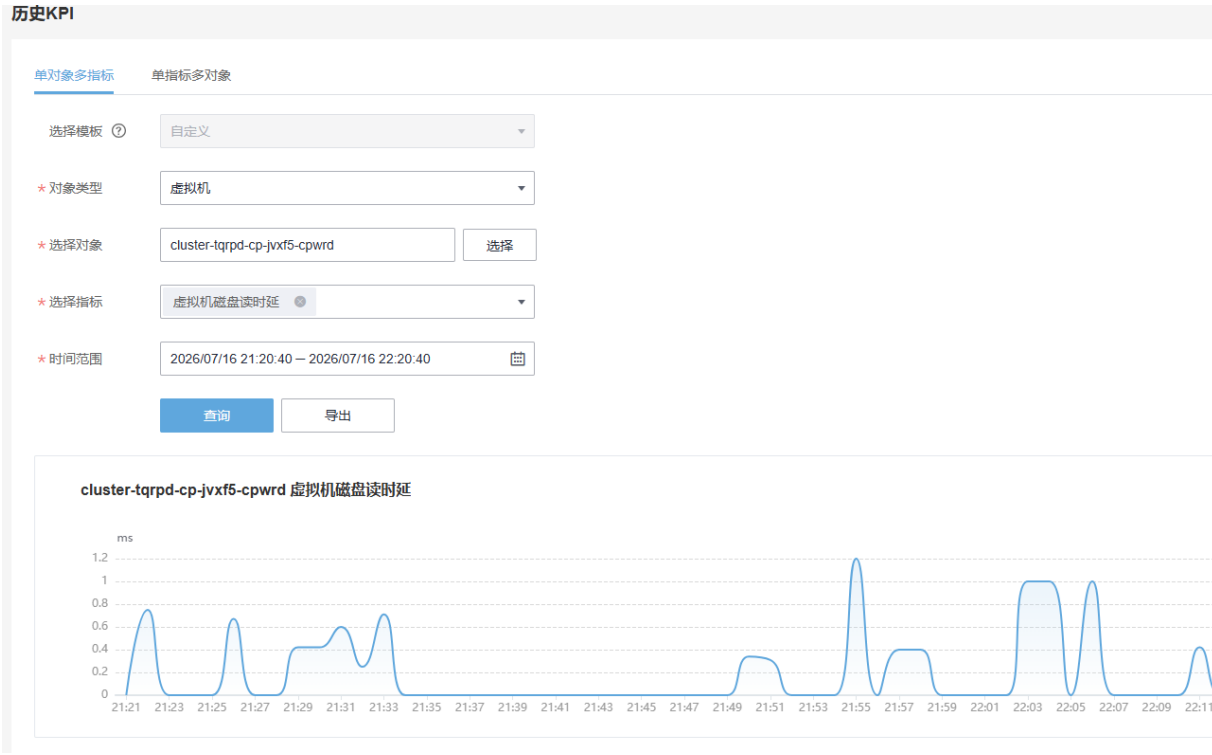
8.3.4.2 告警

运营管理员、组织管理员、项目管理员、普通租户，分视角呈现所能看到的云资源生成的告警。

8.3.4.3 历史 KPI 监控

支持按照用户所在的组织和项目，展示虚拟机资源的性能指标历史KPI趋势图，帮助用户掌握资源对象的运行状态和变化趋势。

 版权所有 © 超聚变数字技术股份有限公司



 版权所有 © 超聚变数字技术股份有限公司

9 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。而传统非超融合的架构下，扩容往往难度高且风险大，主要受制于集中式存储。

超融合（HCI）将计算虚拟化和分布式存储进行一体化融合部署，与传统的“虚拟化+集中式存储”模式相比，不仅精简了设备层级和数量，简化了配置管理步骤，良好的弹性带来的按需投资能力也成为用户选择超融合的重要因素之一。相比于业内大部分超融合产品，FusionOne HCI超融合为用户提供了更加弹性、灵活的扩容选择，扩容过程中不影响业务的运行，主要表现在如下几个方面：

- **在线线性扩容：**存储扩容的颗粒度细化，支持集群级、节点级、磁盘级的在线扩容，即便只增加 1 块 HDD 或 SSD，也能被顺利融入整体资源池，扩容后，集群的性能可线性增加，大大减少传统模式下性能调优效率。

· **全局自动负载均衡：**在集群内在线添加节点扩展计算和存储资源。扩容后的数据存储在不影响业务的情况下自动均衡，无需人工干预。

同时通过多集群管理平台 FusionOne Center实现更大范围的资源池建设和统一管理。运维方面，FusionOne HCI提供了一键扩容功能，支持自动扫描待扩容主机，一键完成超融合集群集群额扩展，省时省力。

全局自动负载均衡

传统文件系统存储采用B/B+树索引方式对文件进行管理，以提高文件的查找访问速度。由于是采用查找法进行数据的访问，当文件个数越多，则查找的性能就越慢，到达一定规模时，查找的时间就无法满足业务的实时性要求。根据传统存储的实现方式，一般索引树会尽量优先放置在高速缓存中，当索引的数据超过高速缓存后，访问的效率问题会变得更加严重。

为了解决海量小文件的访问效率问题，分布式存储组件采用了业界先进的DHT（动态哈希表）技术，该技术的特点是利用一次哈希计算来定位文件，算法的复杂度为O(1)，即使已经存储了海量文件，定位文件的时间永远的固定的，从而大大优化了海量文件的访问性能。

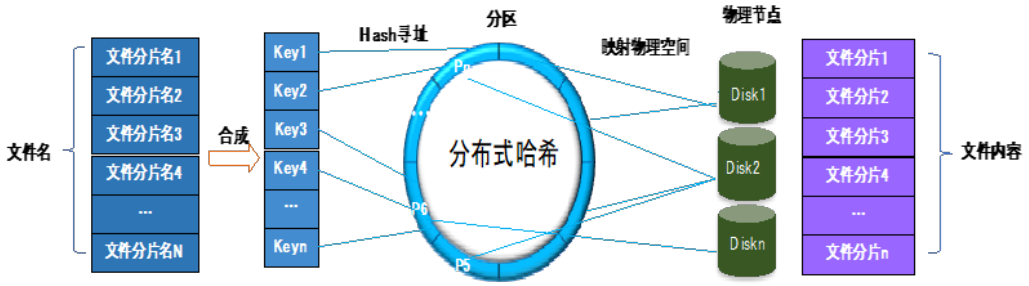
 版权所有 © 超聚变数字技术股份有限公司

分布式存储采用去中心化方式实现数据存储的管理。DHT分布式哈希算法建立一个虚拟的分布式哈希环，均匀地将整个哈希空间分成数千等分，然后将哈希空间的每一个分区均匀地映射到集群的所有磁盘上，确保每一个磁盘所负责的哈希分片数量是一致的。

数据按对象读写时，会将文件按照切片规则（默认4MB）大小生成每个切片的文件名列表（一般是Bucket+文件名+分片编号），然后将文件分片名通过哈希计算，生成一个哈希后的数值，利用哈希值在DHT表中进行查找，从而定位到该分片的磁盘位置获取文件分片的内容。通过组合每个分片，形成整个文件的最终内容。

从文件的定位过程可以发现，哈希计算和DHT表的查找时间是恒定不变的，不会因为文件数量的增加，而影响效率，从而保证了在海量文件数量的情况下的访问性能。

图 9-1 DHT 分布式哈希



- 通过DHT分布式哈希算法，分布式存储管理者可以通过配置信息指定集群数据分布方式（机架级别和服务器级别）。例如服务器级别下，数据默认2副本时，通过DHT将数据的主从副本分别存放在不同服务器的硬盘上，也就意味着当有一台服务器故障时，数据不会丢失。
- 使用DHT分布式哈希算法，数据均衡分布，能够充分利用磁盘的容量和性能，因此读取和写入都更加高效。
- 抗故障性高。由于非中心化设计，不存在特定服务器异常而导致业务暂停。
- 伸缩性好。在扩容或故障时，DHT会自动均衡变化的存储资源，尽量不重构数据以保证业务性能。

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。超融合集群在一定规模内（40节点以内），超融合集群的吞吐量随着节点数量的增加呈线性增加，自动调节超融合集群内配置，简单高效，自动将超融合集群配置为最优状态。

版权所有 © 超聚变数字技术股份有限公司

表 9-1 扩容方式

扩容方式	适用场景	备注
添加硬盘	存储系统可用容量无法满足当前业务需求，且当前集群有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。 • 扩容前需要调整主机内存资源预留 • 可在存储管理页面添加硬盘完成磁盘扩容。
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。 • 可在超融合工具安装后，一键扩容。

数据副本在线转换

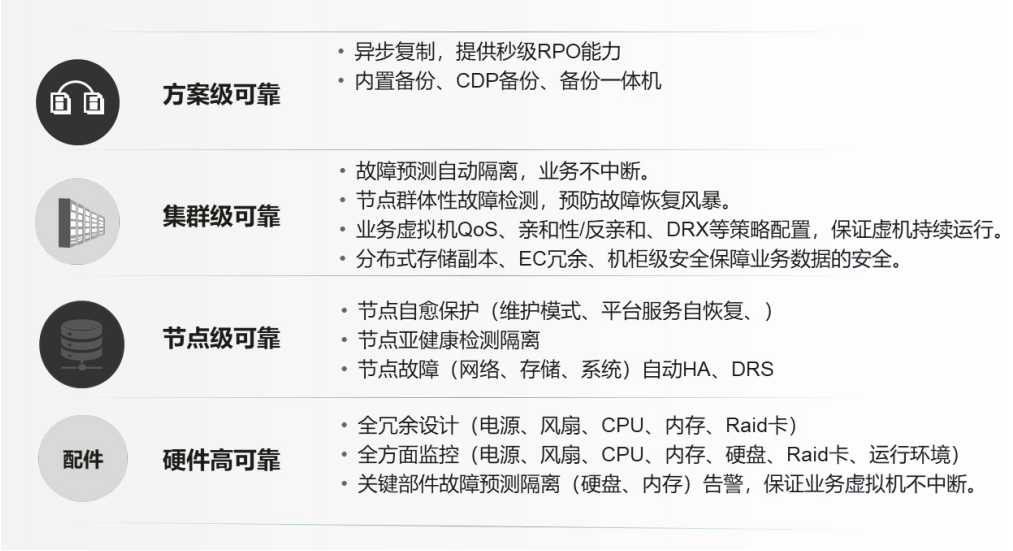
分布式存储提供了2~3副本的存储冗余策略，并实现了在线副本策略转换，在适应应用数据安全性需求，提高存储资源利用率，保护用户存储资源投资方面具有非常好的实际应用意义，例如：

- 根据业务系统不同时期的重要性和数据安全要求变化，可以转换为不同的副本策略，副本数越多，数据安全性越高；
- 存储空间资源不足时，可动态降低相对次要业务系统的副本数，提高存储空间利用率；
- 在线执行副本策略转换，无缝提高业务系统的数据安全保护等级，避免传统设备停机进行数据迁的方式，保证业务连续性。

 版权所有 © 超聚变数字技术股份有限公司

10 系统可靠性

FusionOne HCI系统提供了不同层级的可靠性能力，包括方案级可靠、集群级可靠、节点级可靠、硬件级可靠。



下面将从四个方面介绍FusionOne HCI的可靠性：

方案级可靠：基础备份、CDP备份、容灾、延展集群

集群级可靠：主要从业务方面介绍平台的可靠性，重点聚焦于管理架构、业务数据、业务虚拟机级的可靠性。

节点级可靠：基于超融合内针对节点、节点内的服务、器件的亚健康检测。

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

10.1 解决方案高可靠
10.2 管理服务高可靠
10.3 数据可靠性
10.4 业务高可靠
10.5 亚健康检测
10.6 硬件可靠性

 版权所有 © 超聚变数字技术股份有限公司

10.7 故障可恢复

10.1 解决方案高可靠

随着IT的不断发展，数据资产已经成为企业生存的根本，企业的生产、经营、战略，都离不开数据的支撑。FusionOne HCI 提供了基础备份、CDP备份、平台级容灾方案分级分层保护企业数据资产。

以下是相关的名词解释：

1.快照技术：

快照是指定数据集合的一个完全可用拷贝，该拷贝包括相应数据在某个时间点的映像。快照可以是其所表示数据的一个副本，也可以是数据的一个复制品。快照可用于在线数据备份与恢复。目前实现快照的主流技术包括：

写时拷贝（Copy On Write）：执行快照操作后，数据第一次写入到某个存储位置时，首先将原数据读出来，写到另外为快照保留的存储空间，然后再将数据写入到存储设备中。而下次针对这一位置的写操作将不再执行写时复制操作。

写时重定向（Redirect On Write）：执行快照操作后，写操作会进行重定向，所有的写操作都被重定向到新卷中，所有旧数据均保留在只读的源卷中。因此，每次生成的快照文件都是放在连续的存储区域中，同时解决了COW写两次的性能问题。因其存储性能较好，目前虚拟化平台主要采用写时重定向快照技术。

2.虚拟机CBT

数据块修改跟踪技术(Changed Block Tracking)是虚拟化平台简化和提高虚拟机备份效率的重要组成部分，它可以实现只备份变化块数据，而不需要备份全部数据，从而减少备份数据，提高备份效率。

3.完全备份

完全备份是对目标数据，比如虚拟机、磁盘、逻辑磁盘、文件系统等所有数据进行完全拷贝。这种备份方式的特点是数据最全面、最完整，数据一致性得到完全保护。当发生灾难时，只要用最新完全备份点，就可以恢复全部数据。

由于完全备份的数据量非常大，占用备份存储空间较多，因此，备份时间较长。而且，如果两次完全备份操作间隔较短，则存在大量的重复数据，实际上只有一小部分数据发生了变化，因此完全备份频率一般较低。

4.增量备份

增量备份只拷贝上次备份以来发生的数据变化。在大多数情况下，连续两次备份之间只有小部分虚拟机数据发生变化。增量备份可以利用此特性，减少备份资源需求，提高备份效率，增加备份频率。

10.1.1 基础备份

FusionOne HCI 内置备份功能，支持无代理定时周期性备份，可以根据现场对业务数据的RPO和RTO要求制定具体备份方案。

技术特性

- 无代理备份恢复：不需要在环境中安装任何代理程序即可实现虚拟机的备份恢复功能。

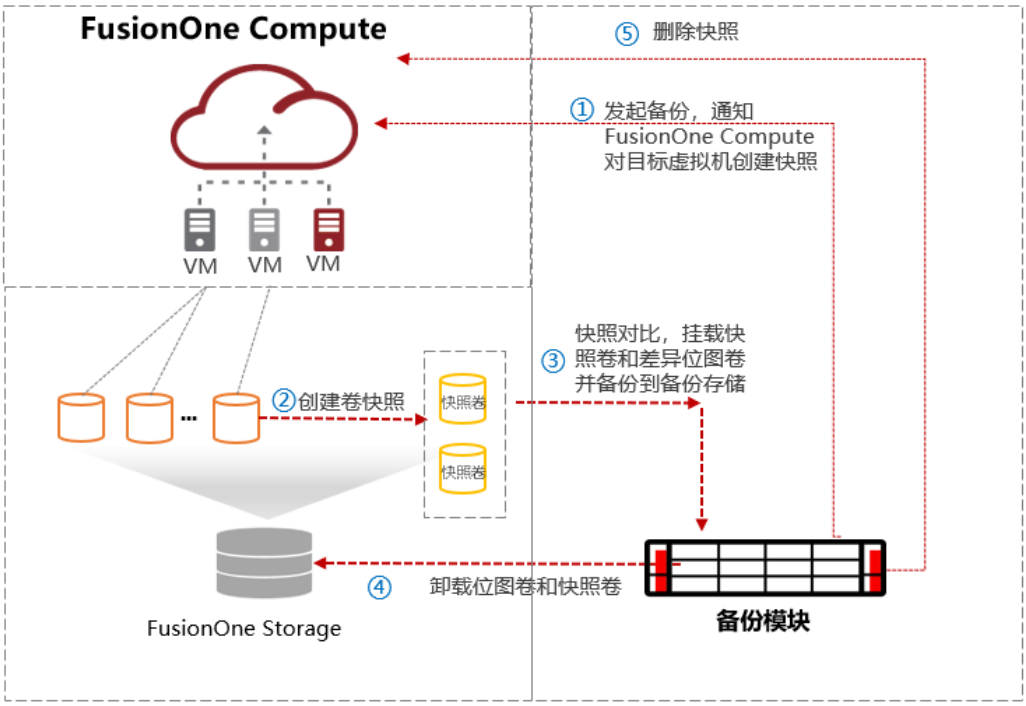
 版权所有 © 超聚变数字技术股份有限公司

- 备份时自动发现虚拟机：支持对整个宿主机进行备份，支持自动发现新增虚拟机并对其进行备份。
- 多种数据保护方式：支持对虚拟机进行完全备份、增量备份，灵活数据保护。
- 无保留快照差异备份：不需要保留快照即可实现差异备份，避免保留快照占用生产存储空间。
- 快速备份恢复：获取虚拟机磁盘的有效数据和变化数据进行读写，无需读写全盘数据，大幅提升备份恢复速率。
- 通过REST接口兼容ARM架构等国产平台。

备份恢复原理

LAN-Free备份流程：

图 10-1 FusionOne HCI Lan-Free 备份



1. 备份模块通知FusionOne Compute对目标虚拟机创建快照。
2. FusionOne Storage存储上产生对应快照卷。
3. 通过快照差异比对，创建差异位图卷，挂载快照卷和位图卷，然后读取快照卷的有效数据块，并备份至备份存储。
4. 卸载位图卷和快照卷。
5. 通知FusionOne Compute删除位图卷和历史快照卷。

恢复流程：

1. 根据备份的配置信息创建虚拟机。
2. 写入数据到虚拟机磁盘中。
3. 恢复完成。

 版权所有 © 超聚变数字技术股份有限公司

10.1.2 CDP 备份

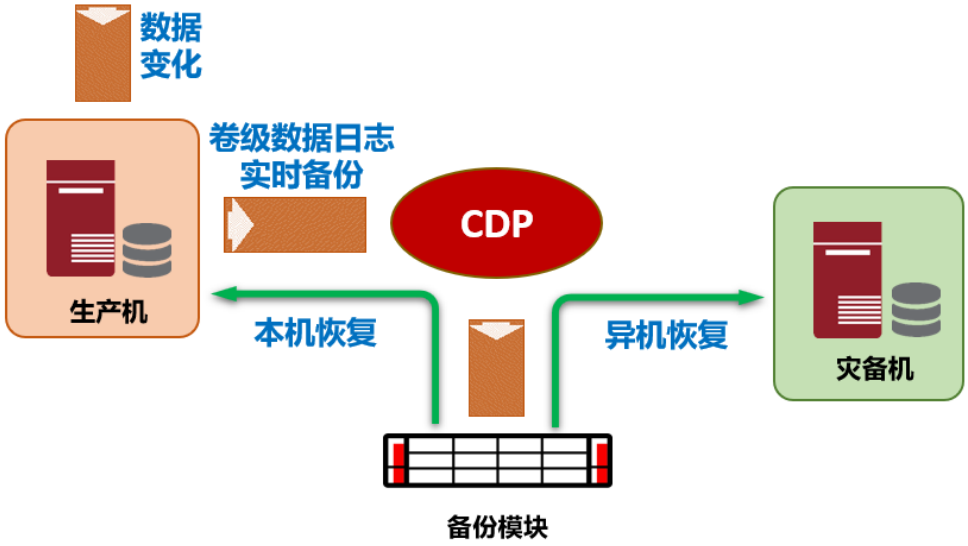
持续数据保护（Continuous Data Protection，简称CDP）是一种在不影响主要数据运行的前提下，可以实现持续捕捉或跟踪目标数据所发生的任何改变，并且能够恢复到此前任意时间点的方法。

功能简介

概述

CDP（Continuous Data Protection，连续数据保护）支持对卷级数据的持续性保护，通过持续监控卷数据的变化，深入系统层获取数据的变化日志，实时存储到备份服务器中。当发生数据灾难时，可基于数据日志实现任意时间点回退，并通过数据挂载方式即时恢复，满足企业更高要求的恢复时间目标（RTO）和恢复点目标（RPO）。

图 10-2 CDP 流程



功能特点

- 实时监控：建立系统级的监控体系，以卷级数据I/O的消息触发为驱动，实现对变化数据的实时监控。保护业务系统运行环境的完整一致性。
- 秒级备份：借助监控的实时性，对卷级变化数据的抓捕可达到秒级细颗粒度效果，备份窗口降至卷级同步可以实现的最小值，极少消耗主机资源，RPO趋近于0。
- 任意时间点恢复：用户配置恢复任务时，支持以秒为单位选择备份时间窗口内的任意历史时间点。

10.1.3 CDP 持续保护

通过结合持续数据保护技术和高可用技术的原理，针对RTO和RPO敏感度较高的用户推出了容灾接管方案。该方案包含实时备份和应急接管两种模式。

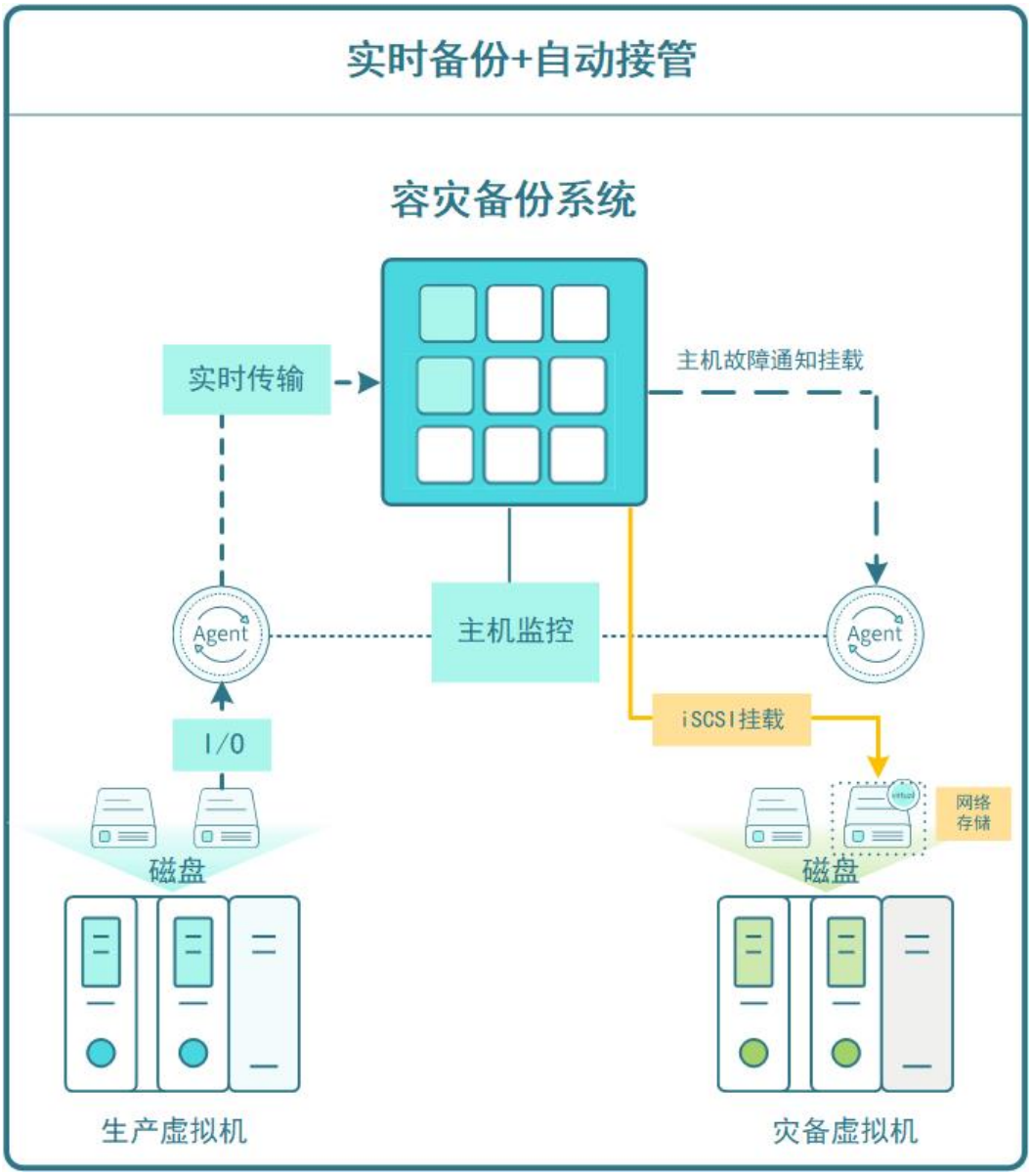
实时备份模式：只需要在用户环境中部署一套容灾备份系统并在用户主机/虚拟机上安装主机保护代理即可实现，实时备份任务启动后，在首次同步之后就会进入实时备份阶段，一但生产主机/虚拟机发生故障，用户可通过恢复向导将实时备份任意时间点环

 版权所有 © 超聚变数字技术股份有限公司

境恢复到原主机或者新主机/虚拟机上。该模式RPO≈0，RTO为恢复数据大小除恢复链路带宽，适用于对RTO要求不高，但是需要在故障发生后尽量少丢失数据的用户。

实时备份+应急接管模式：该模式在实时备份模式的基础上还需要配置额外的备用主机/虚拟机，当生产主机/虚拟机发生故障时，可以直接由备机/虚拟机进行接管，而无需进行恢复操作，相比实时备份模式RTO时间更短，RTO为分钟级。实时监控阶段的数据到达备份系统后，不会将实时监控数据发送至备机/虚拟机，当生产主机/虚拟机发生故障时，可通过自动挂载的方式将最新时间点的映射卷挂载至备机/虚拟机，并自动将数据和应用无缝切换到备机/虚拟机。由于接管数据实际上在备份服务器上，此容灾模式对备机/虚拟机的要求较低，只需能够运行对应版本的操作系统及应用即可，无需配置与生产主机/虚拟机对等的数据卷，但另一方面，接管后由于数据不在备机/虚拟机本地，性能会有所下降。

图 10-3 实时备份+应急接管原理图

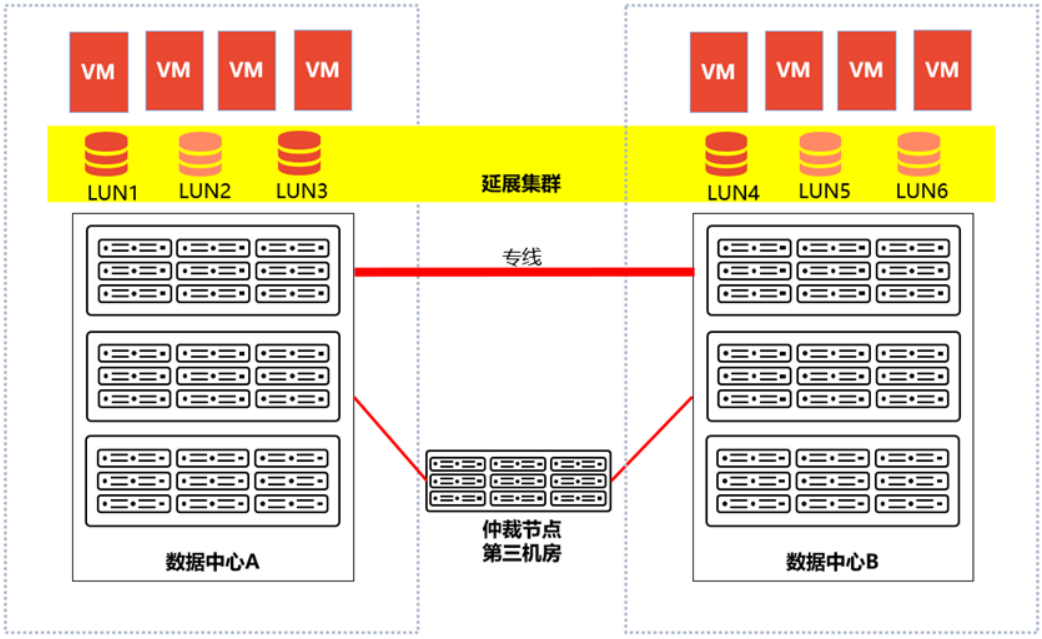


 版权所有 © 超聚变数字技术股份有限公司

10.1.4 延展集群

延展集群主要用于跨数据中心的容灾方案（一般为同城网络），两个数据中心通过裸光纤等方式链接，实现低时延高带宽的二层网络。在延展集群的配置中，两个数据站点均为活动状态。一旦其中一站出现故障，集群将自动切换至另一站点的存储资源，并由虚拟化平台重新启动故障的虚拟机，主要用于关键应用或者行业的场景，如医疗、制造、金融等场景，结合业务层面的业务负载均衡能力，可实现业务双活能力。

FusionOne HCI双活中心容灾方案以超融合集群延展技术为核心，提供跨数据中心的计算、存储高可用能力， 以极低成本为客户提供中心级容灾，为关键业务永续提供有力保障。双中心最低4节点起配，并支持灵活水平扩展。整体方案如下图所示：



延展集群主要技术原理

1、延展副本，数据中心级可靠

将数据复制成多份，按照故障域类型存储在分布式存储中不同的节点、机架、数据中心中。延展集群的存储策略支持配置2、3、4副本，即当副本数配置为2或4时，数据的2个或4个副本分别平均写入到A、B两个数据中心，当副本数配置为3时，部分数据的1个副本在A数据中心，2个副本在B数据中心，其他的数据则相反，从而使数据均衡分布到2个数据中心，因此数据中心之间的容量不会相差太大。

2、跨站点通信调优，提升延展集群性能

在延展集群中，站点之间的带宽往往成为延展集群的性能瓶颈，FusionOne HCI深度调优延展集群的IO处理策略，集群在正常运行、故障、故障恢复等场景支持本地IO策略，尽可能减少跨站点带宽的占用，并提升了业务的性能。

- 读本地优先：

在延展集群中定义逻辑池的“读优先数据中心”后，客户端会从“读优先数据中心”读取数据，从而确保IO读性能接近普通HCI的性能，并且节约站点之间的裸光纤带宽。

- 重构本地优先：

故障重构期间，为了减少站点间通信，FusionOne HCI支持重构时，优先通过读取本地已有副本来恢复数据，减少数据中心之间的专线带宽的占用，降低带宽成本。

版权所有 © 超聚变数字技术股份有限公司

- 故障恢复自动回切：

在本地站点故障后，可切换到远端读取，业务不中断。当本地站点故障恢复后，FusionOne HCI支持自动回切，将主副本自动回切到本地，减少跨站点通信，提升业务的性能。

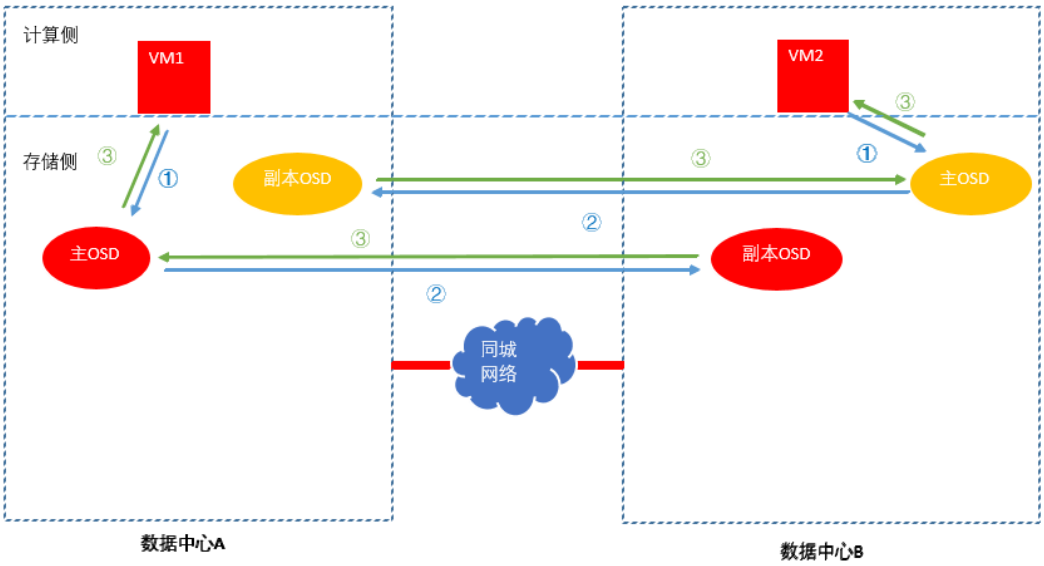
3、主站点仲裁优先

支持设置优先数据中心，“优先数据中心”具有仲裁获胜的优先权，数据中心间的链路故障时，“优先数据中心”将优先获得仲裁胜利，优先为主机提供访问。

4、高可靠的IO处理机制

支持同步写、本地读策略，为RPO=0提供了可靠的保证，兼具超融合整体的性能与高可靠。

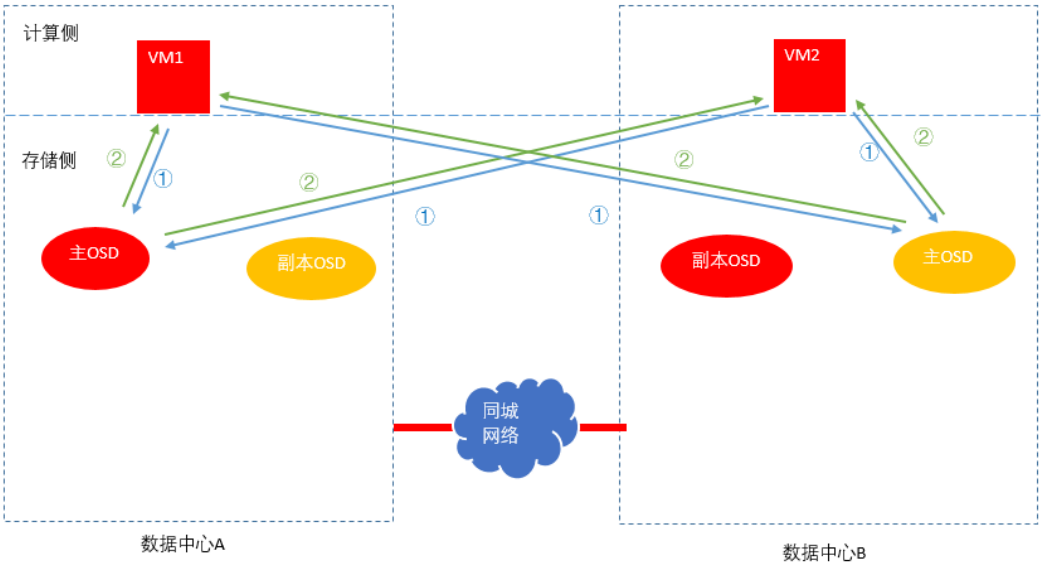
数据写I/O流程。



步骤：

1. 主机下发写I/O到存储池中多副本的主OSD，即数据的副本。
2. 数据中心A的主OSD同时将数据写入数据中心B的副本OSD。
3. 等待主OSD及副本OSD的数据写入的状态成功后，再返回给业务侧，I/O写入完成。

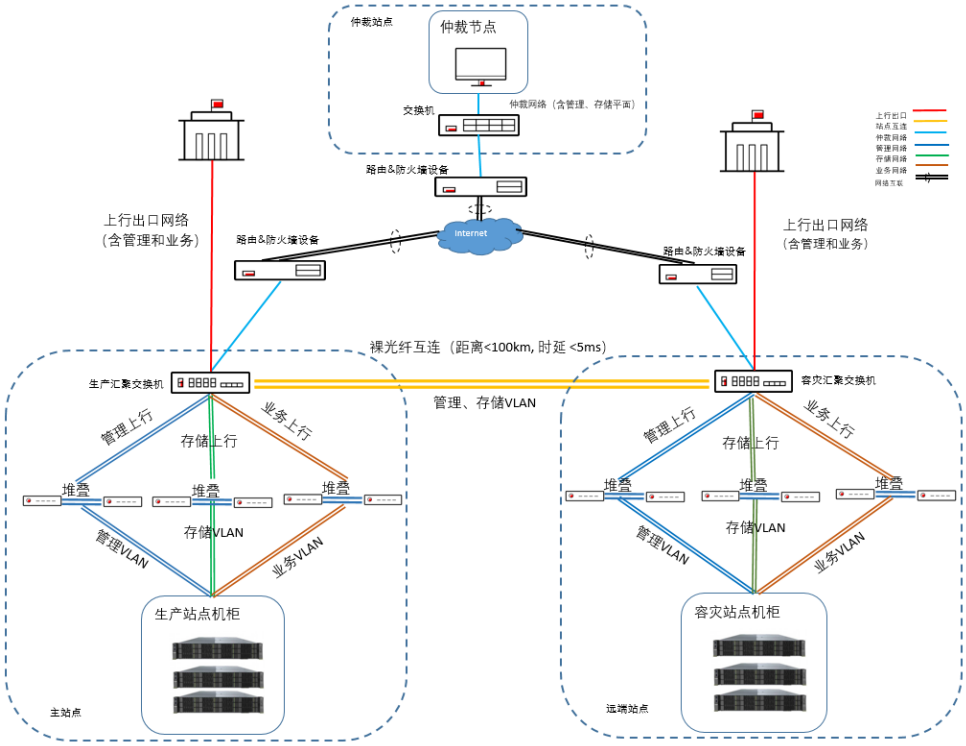
 版权所有 © 超聚变数字技术股份有限公司



步骤：

1. 业务侧下发的读请求发到主OSD。
2. 主OSD返回数据。

延展集群组网拓扑：



网络上行要求：

根据用户实际情况可部署网关，在生产站点、容灾站点汇聚交换机配置路由策略，将不同的上行流量分别路由到不同的网关处理。

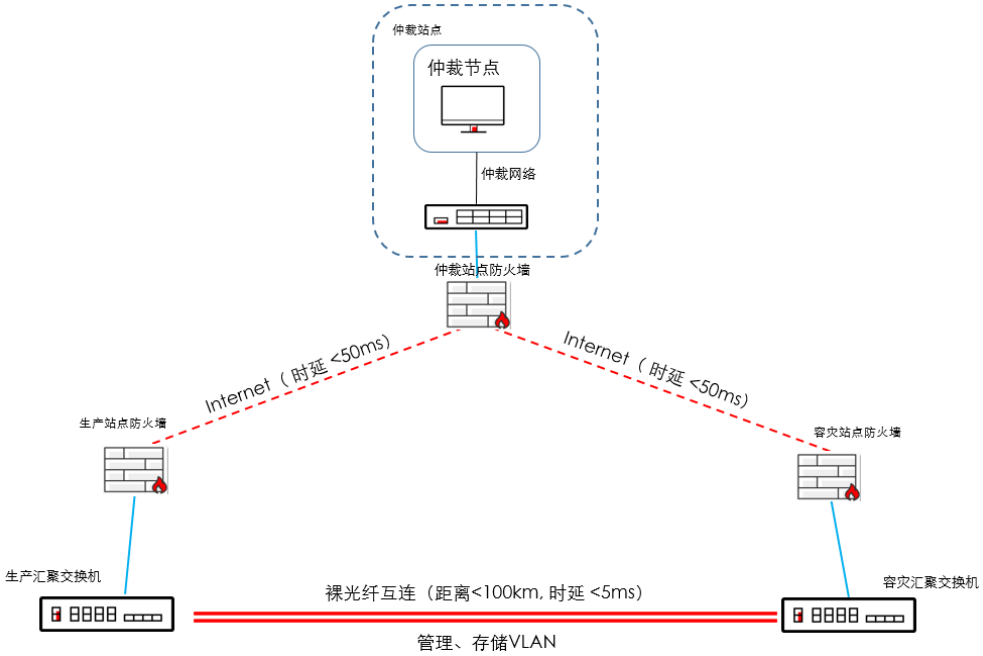
站点间网络要求：

 版权所有 © 超聚变数字技术股份有限公司

两个站点通过汇聚交换机相连，堆叠或者M-LAG方式，或通过网络四层负载均衡等方式。两个站点通过裸光纤物理相连。网络要求参考业界的规范约束，裸光纤网络要求网络时延小于5ms，考虑时延要求，一般要求两个站点的距离不超过100KM。

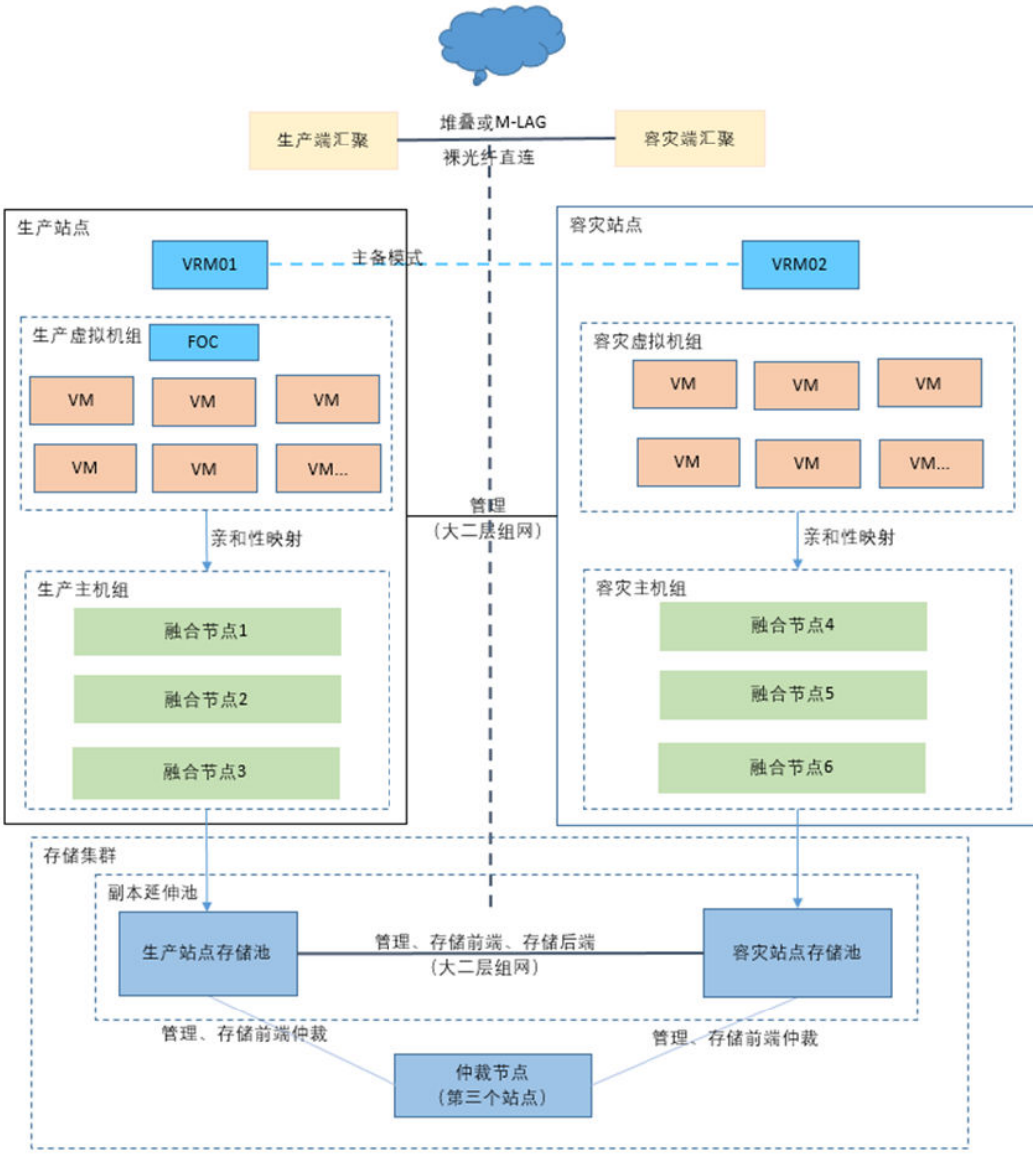
仲裁网络要求：

仲裁站点与生产站点和容灾站点要求同网段的互联，一般通过企业专线互联，因此一般会通过IP Sec等VPN互连，最低带宽不低于100Mbps，时延小于50ms，示例组网如下：



部署图：

 版权所有 © 超聚变数字技术股份有限公司



硬件层面：

生产与容灾站点部署相同规格的融合节点。仲裁服务部署在第三个数据中心。要求融合节点四节点起步，且两个站点的融合节点个数相同（考虑副本的均衡性，不建议有节点数差异）。另外在计算节点规划上也建议保持一致，防止HA的时候资源不足。

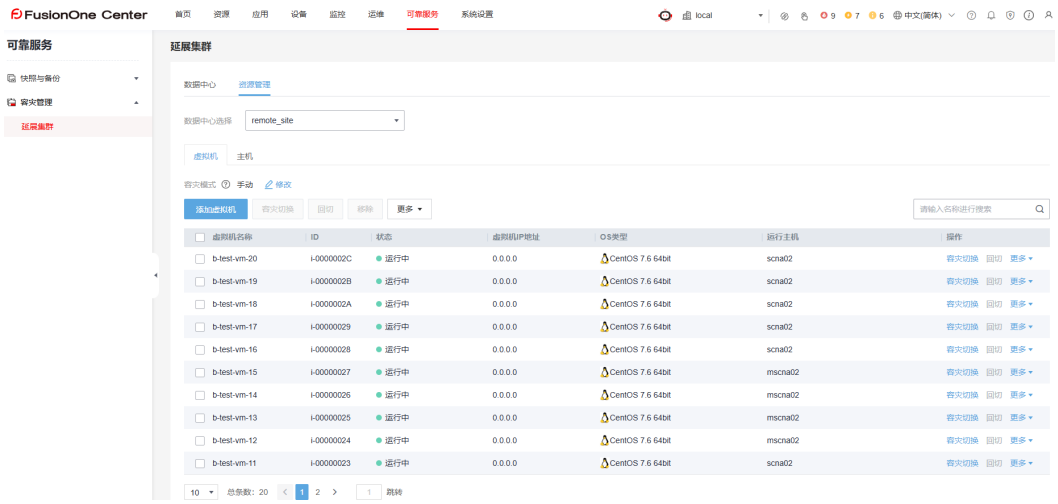
软件层面：

1. 生产站点部署FOC管理虚拟机、VRM01虚拟机。容灾站点部署VRM02管理虚拟机。
2. VRM01和VRM02形成主备集群关系，生产站点故障后可接管。
3. FOC虚拟机部署在分布式存储上，通过HA高可用实现运行在生产站点，可HA到容灾站点。
4. 业务虚拟机通过主机组与虚拟机组的亲和性，实现虚拟机就近的亲和性运行。
5. 为充分利用容灾站点资源，存储池采用双池方案，形成生产与容灾站点互为备份关系。

延展集群站点管理

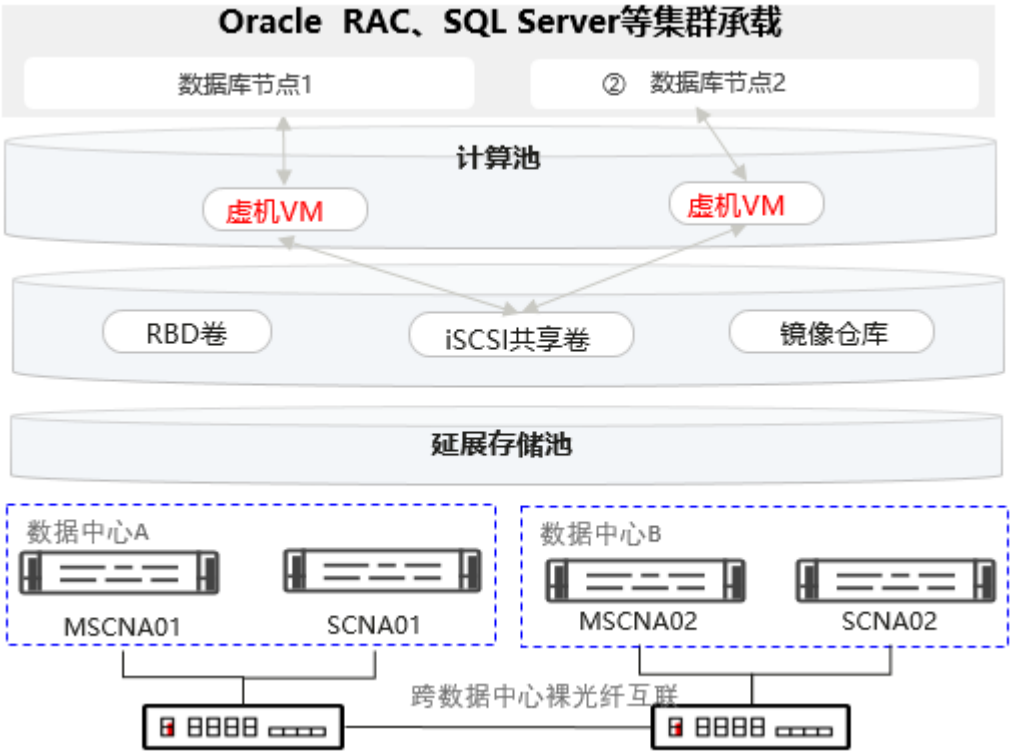
 版权所有 © 超聚变数字技术股份有限公司

通过可视化的站点管理，实现对数据中心、容灾虚拟机全局管理。通过追踪容灾虚拟机的状态，可以让用户清晰的了解到容灾虚拟机的运行状态，并通过提供一键在线的容灾切换，可以快速将虚拟机切换到容灾数据中心，或者回切到本端数据中心。



延展集群iSCSI双活

常规的虚拟机发放的磁盘为RBD接口的卷，无法运行依赖iSCSI预留锁机制的共享卷业务，如Oracle RAC、SQL Server集群等。为了解决该问题，FusionOne HCI支持iSCSI卷，通过iSCSI的ALUA机制，结合延展集群的主机物理位置，自动实现本地优先的选路径机制，保证关键业务可以在容灾虚拟机中稳定运行。



延展集群副本的可靠性基础上，同时提供librbd、iSCSI两种延展集群模式下的存储卷访问。iSCSI延展卷提供端到端的就近访问功能，主要通过如下策略实现：

 版权所有 © 超聚变数字技术股份有限公司

1. iSCSI目标器（Target）本地节点组（FusionOne Storage）
2. iSCSI启动器（Initiator）选择本地节点（FusionOne Compute）
3. 主机的亲和性绑定主机，实现存储的本地数据中心访问。

10.2 管理服务高可靠

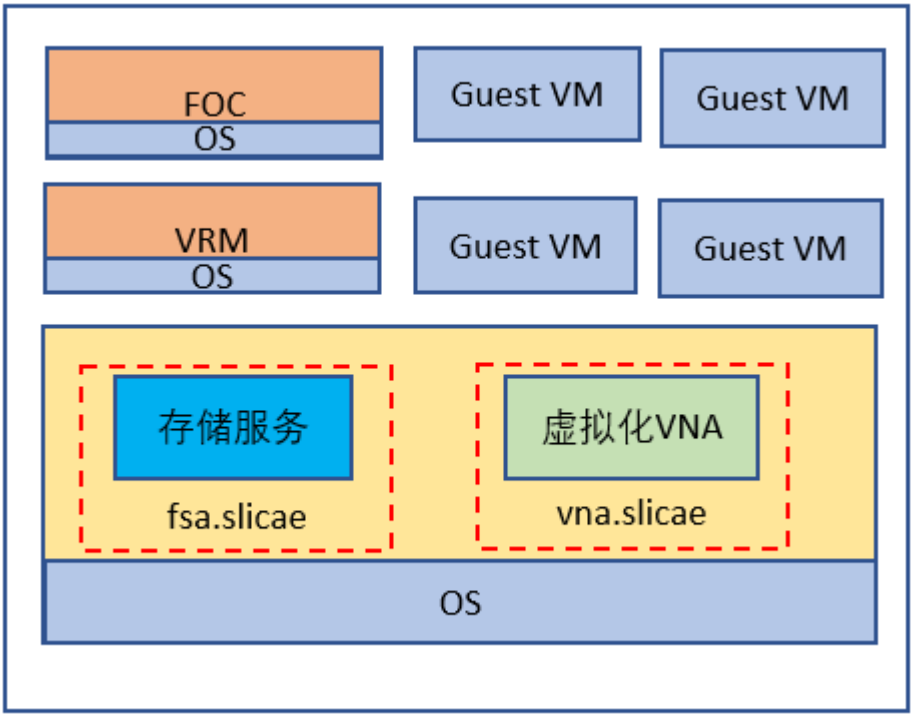
服务安全隔离

为了提升各个子系统的可靠性，业界常见的隔离方式有进程级隔离（cgroup）、容器隔离、系统隔离三种方式，其中系统隔离方式最为安全。FusionOne HCI结合超融合实际业务场景的性能、安全场景，针对不同服务采用不同的隔离方式。

管理服务采用系统隔离的方式，将管理服务通过虚拟机方式隔离。

计算服务：运行在主机系统，与存储服务通过cgroup资源隔离。

存储服务：运作在主机系统，与计算服务通过cgroup资源隔离。



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。

FOC管理集群采用基于K8S的集群方案，一旦发现任一管理节点故障，其他管理节点将立刻拉起对应的管理服务，持续对外提供服务。针对管理节点上POD，对运行的服务POD进行实时检测，如发现POD异常，自动触发相关POD终止重新创建。

 版权所有 © 超聚变数字技术股份有限公司

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

超融合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。

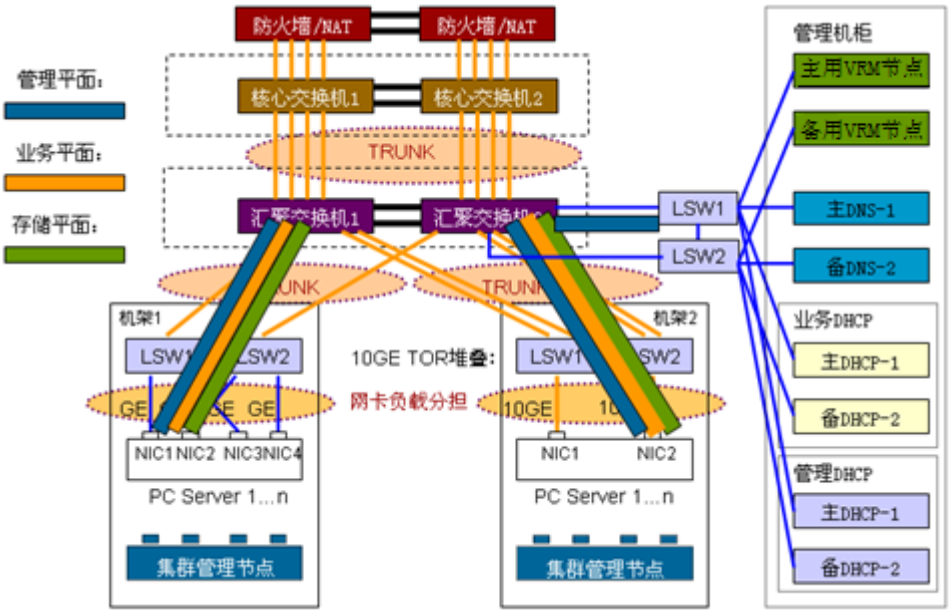
网络平面隔离

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，超融合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。

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：

 版权所有 © 超聚变数字技术股份有限公司

图 10-4 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。

版权所有 © 超聚变数字技术股份有限公司

- 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。

10.3 数据可靠性

10.3.1 多存储池

超融合的分布式存储集群支持多个存储池，如全闪存储池、混配存储池，可以自有组合，最大支持128个存储池。分布式存储同时还支持节点共存储池，即一个节点同时属于两个存储池。

例如，当服务器配置了SATA SSD、SATA HDD的硬盘时，在初始配置可以将其中SATA SSD作为一个全闪存储池，SATA HDD+NVMe 缓存盘作为一个混闪存储池。每个服务器节点单个存储池内的主存盘数量推荐不少于4个。

无感知扩容

在分布式架构中，节点的扩容往往意味着集群拓扑图的改变，也伴随着数据的重分布，重分布过程会对集群性能产生影响。

利用多存储池技术，能够将新扩容的节点独立成一个存储池，从而避免了数据的搬迁，大大降低了对业务的影响。

10.3.2 多副本

存储系统中硬盘的增多除了带来容量和性能的提升外，也增大了硬盘故障的发生概率。将数据复制成多份，放在不同的硬盘上，这就是副本冗余技术，是最简单直观的数据可靠性技术。

分布式存储的块存储，采用数据多副本备份机制来保证数据的可靠性，即同一份数据可以复制保存为2~3个副本。针对系统中的每1个卷，分片后的数据按照DHT算法保存集群节点上。在副本冗余策略下，各副本数据最终写入不同的硬盘。而且DHT算法可以保证各个硬盘的容量使用均衡。副本策略可以按节点或机架级感知，应用层不感知副本存在什么硬盘上。默认存储节点级时，两个硬盘属于两个不同的存储节点。此时即使一个节点硬盘损坏，数据也不会丢失。

副本冗余技术具有以下特点：

- 可靠性高：数据副本跨硬盘或机架，只要不是全部副本损坏，数据都不会丢失。
- 可配置性好：副本策略可以按节点或机架级感知配置，副本可以选择2~3，便于空间上的管理。
- 节点间容量均衡：使用DHT算法可以保证各硬盘使用率分布均衡，避免单个硬盘写满而影响整个业务。
- 故障恢复：自动恢复数据效率高，保证数据可靠性恢复到设置的副本状态。

10.3.3 EC 纠删码

超融合的分布式存储支持Erasure Code（EC，纠删码）方式来保证数据的可靠性。相对三副本，EC数据冗余保护机制在提供高可靠性的同时也能够提供更高的磁盘利用率，最高可使磁盘空间利用率最高达80%，相同可用容量需求条件下，采用EC方式可以节省30%~59%左右的硬件，也更省空间、更节能。

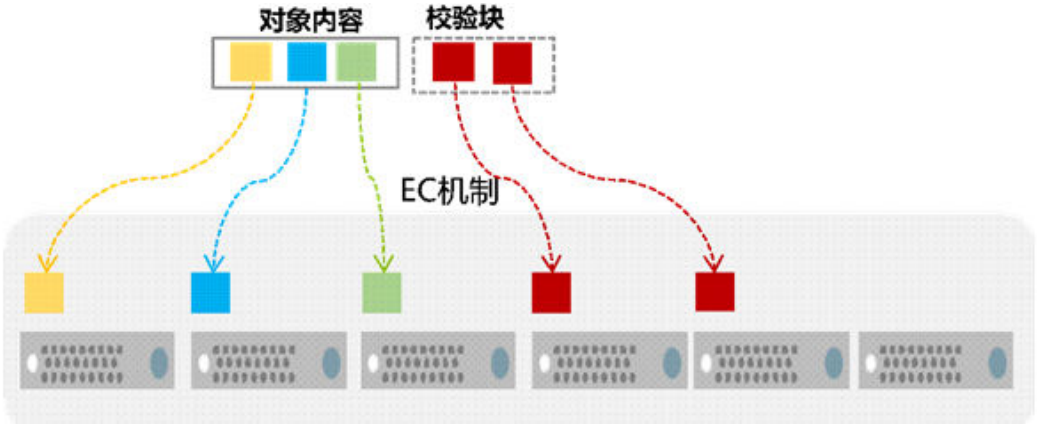
纠删码技术主要原理如下：

 版权所有 © 超聚变数字技术股份有限公司

写入的数据分为数据被拆分为K个数据块，然后编码生成M个校验块，总共K+M份数据通过DHT算法分别存入不同硬盘中。当系统出现故障，丢失了某些块时，只要丢失的块数目不超过M，就可通过数据恢复算法，将丢失的数据块从剩余的块中计算出来。在这种方式下，空间的利用率约为K/(K+M)，数据的可靠性由M值的大小决定，M越大可靠性越高。

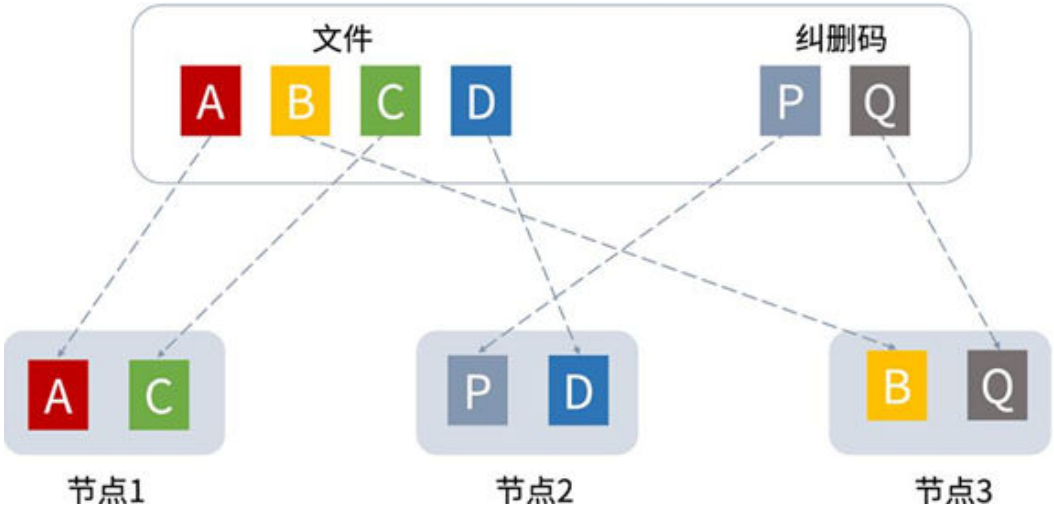
例如3+2 EC配置下一个对象写入时，对象会生成5个块，每个块均存入不同的硬盘，如下图所示：

图 10-5 EC 机制示例



默认的纠删码是按照节点来分配数据的，例如EC（4+2）需要配置6个节点，但亚节点纠删可按照硬盘来分配数据，例如EC（4+2）:1，则是把3个节点当6个节点用，每个节点选择2块硬盘，整个集群选择6块不同的硬盘来存放4+2总计6个分片数据。

“4+2：1”EC策略示意图：



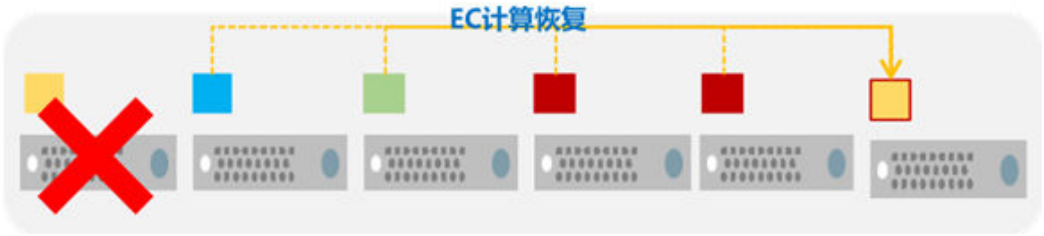
使用EC策略后，如果读取数据时disk2和disk4异常无法返回数据，则可以通过其他硬盘数据计算出完整的对象数据。

如果写K+M的块时，有超过M个写入失败，FusionOne Storage纠删码会将对象还原到上一个完整版本中，以保证数据的一致性。

同副本机制类似，当某个节点恢复时超过一定时间（可配置），该故障节点依然无法恢复时，系统将会自动启动数据自愈，即利用现有的数据和EC校验块恢复出原有数据，保证数据的可靠性。

 版权所有 © 超聚变数字技术股份有限公司

EC计算恢复示意图：



10.3.4 数据一致性

分布式存储采用多种技术手段，包括一致性复制协议、读修复技术和数据完整性保护技术设计，保证了数据的一致性和完整性。

一致性复制协议

强一致性复制协议可保证多个副本数据的一致性，即只有当所有副本都写成功，才返回写入磁盘成功。正常情况下分布式存储可保证每个副本上的数据都是完全一致，从任一副本读到的数据都是相同的。如果某个副本中的某个磁盘短暂故障，则暂时不写这个副本，等恢复后再恢复该副本上的数据；如果磁盘长时间或者永久故障，则会把这个磁盘从群集中移除掉，并为副本寻找新的副本磁盘，再通过“rebalance”机制使得数据在各个磁盘上的分布均匀。

数据恢复技术

- 临时故障

  在一定时间内可以重新上线OSD的故障，如升级维护、重启和掉电等。在故障期间降级存储数据，待故障修复后，自动将降级的数据恢复。提升数据存储的冗余度，降低了数据丢失的风险。

- 永久故障

  无法重新上线OSD的故障，如坏盘。在一定时间后，自动剔除故障OSD上的数据，将数据恢复到其他OSD，恢复数据的冗余度，提高数据存储的安全性。

数据完整性校验技术

在线实时读写数据时，做CRC（Cyclic Redundancy Check）校验，防止静默数据错误。

提供集群副本间数据扫描的操作，检测副本间数据的一致性，包括元数据校验和深度校验。元数据校验只对元数据信息进行扫描，相对比较快。深度校验不仅对元数据进行扫描，还会对存储的数据进行扫描，相对比较慢。用户可以根据需求进行配置，如果发现有不一致的数据，可以及时人工干预，恢复正确的数据，提高数据存储的安全性。

10.3.5 掉电保护

系统运行过程中可能会出现服务器掉电可能，超融合采用使用SSD作为保电介质保存元数据和缓存数据，以防掉电而丢失元数据和用户数据。

分布式存储支持SSD作为保电介质。程序运行过程中会把元数据和缓存数据写入保电介质中，节点异常掉电并重启后，系统自动恢复保电介质中的元数据和缓存数据，并把数据自动恢复到一致性状态，保证业务能正常访问。

 版权所有 © 超聚变数字技术股份有限公司

10.3.6 快速数据重建

分布式存储节点具备强大的数据保护和恢复机制。系统因节点、磁盘或存储节点发生故障导致数据不一致，可通过内部自检机制，自动发现数据故障并启动数据恢复机制，多个节点并行工作，有效避免单个节点的性能瓶颈，对上层业务的影响做到最小化。

每个硬盘都保存了多个数据块，这些数据块的副本按照策略分散在系统中的其他节点。当检测到硬盘或者节点硬件发生故障时，自动在后台启动数据修复。由于数据块的副本被分散到多个不同的存储节点上，数据修复时，将会在不同的节点上同时启动数据重建，每个节点上只需重建一小部分数据，多个节点并行工作，有效避免单个节点重建大量数据所产生的性能瓶颈，对上层业务的影响做到最小化。

支持并行、快速故障处理和重建：

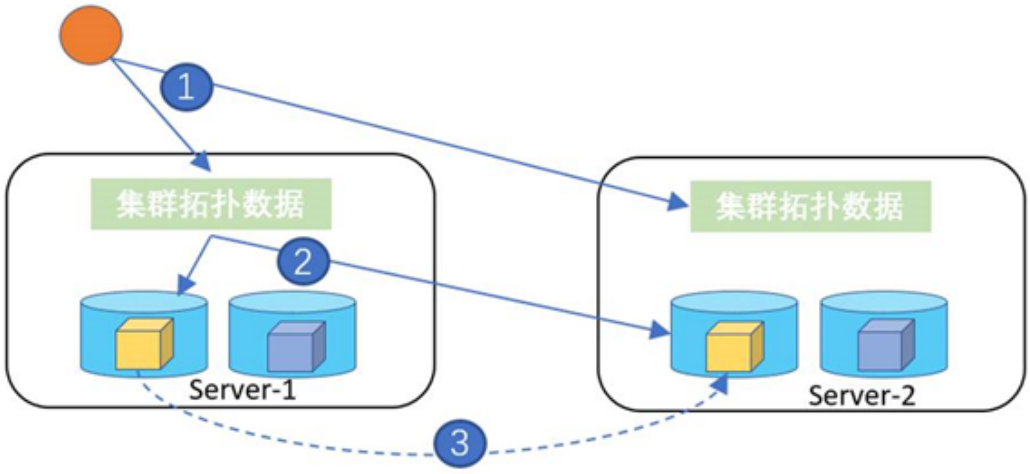
- 数据块及其副本分散在整个资源池内，当磁盘或存储节点故障时，系统能自动进行数据重建，全NVME SSD数据盘配置，数据重建速度需能满足每TB≤10分钟（空闲时），全HDD数据盘配置，数据重建速度需能满足每TB≤20分钟。
- 数据分布上支持跨节点，不会因某个节点故障导致的数据不可访问和不可重建。

故障或者扩容时可以自动进行负载均衡，应用无需调整即可获得更大的容量和性能。

10.3.7 数据恢复

副本冗余技术与纠删码冗余技术可以保证在一定的硬件故障场景下，存储系统不间断地对外提供业务。但是仍有可能发生某些不可逆的故障，如多台存储节点操作系统崩溃、网络全部瘫痪故障、系统内部故障等发生时，存储系统受到致命影响无法对外提供服务的场景。这种情况下，存储数据的硬盘并未全部损坏，数据仍可通过一定的技术实现抢救恢复。

FusionOne Storage提供的集群数据恢复工具，可充分应对这种极限情况。在存储系统不能对外提供服务时，将数据从发生故障的系统中完整导出，抢救客户的全部数据。恢复工具的核心原理为根据系统数据与元数据的在集群中的分布特征，定位出数据的位置，读取数据信息，当有数据损坏时，尝试根据副本或纠删码冗余信息恢复出有效数据。具体工作流程如下图所示。



- 首先获取集群拓扑信息，检验拓扑信息的可用性。
- 然后进一步根据集群拓扑信息定位检查所有数据盘的可用性，并扫描数据盘上的数据块。

 版权所有 © 超聚变数字技术股份有限公司

- 最后根据DHT数据分布式路由算法的原理，按照数据在系统中各块硬盘上的分布，恢复出所有数据。

FusionOne Storage集群数据恢复工具，具有以下技术特点：

- 必须在业务停止运行时执行恢复，恢复过程中完整校验所有数据，保证数据一致性。如果遇到数据不一致的情况，可根据选择自动修复数据或跳过不一致的数据。
- 支持按照一定的策略恢复部分数据
- 支持深度恢复模式与一般恢复模式，深度恢复模式下，元数据以及标签数据全部得以保留，一般恢复模式仅支持恢复出数据及元数据。

10.4 业务高可靠

FusionOne HCI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。

主机外设故障 HA

支持对虚拟机关键外设的故障检测，并主动触发HA，包括管理、业务网络上行链路故障、GPU设备故障等，一旦故障，支持配置主动触发HA，快速恢复业务。

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

 版权所有 © 超聚变数字技术股份有限公司

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。

10.5 亚健康检测

亚健康是一种状态。处于亚健康的组件是指性能严重低于预期的组件，包括网卡，硬盘，内存，CPU等。系统中亚健康组件影响系统的整体性能，包括直接影响和间接（扩散）影响。系统当前具备的亚健康检测与处理机制，按照资源分类如下：

- 节点亚健康检测与处理。
- 网络亚健康检测与处理。
- 介质亚健康检测与处理。

节点亚健康检测与处理

OSD集群节点和复制集群节点由于节点软硬件问题导致节点进入亚健康状态，比如CPU降速，内存反复纠错导致访问降速等；在这种场景下，系统服务时延受到影响，系统通过检测时延信息定位处于亚健康状态的节点，对节点（或OSD）进行隔离，恢复服务时延。

- **OSD集群节点亚健康检测与处理**

  检测机制：

  系统在访问OSD的节点部署有检测模块，检测节点访问OSD的路径时延，对访问时延超过阈值的OSD上报控制节点。

  OSD集群控制节点收集各访问节点上报的OSD信息，按照大多数原则（访问某OSD的大多数节点上报该OSD亚健康）对OSD亚健康状态进行判断。

  隔离机制：

  OSD集群控制节点对判断出的亚健康OSD，在满足数据冗余的前提下进行隔离处理。

  恢复机制：

  维护人员在解决亚健康问题原因后，通过命令将被隔离的OSD重新加入集群。

- **副本IO路径亚健康快速检测与服务切换（Hint可用性增强）**

  检测机制：

  – 系统在访问OSD的节点部署有检测模块，在短时间内检测节点访问OSD的路径时延，对访问时延超过阈值的OSD上报控制节点。

    OSD集群控制节点收集各访问节点上报的OSD信息，按照大多数原则（访问某OSD的大多数节点上报该OSD亚健康）对OSD亚健康状态进行判断。

  – 当单个IO时延过大时，为保证业务IO快速恢复，判断此OSD处于亚健康状态。

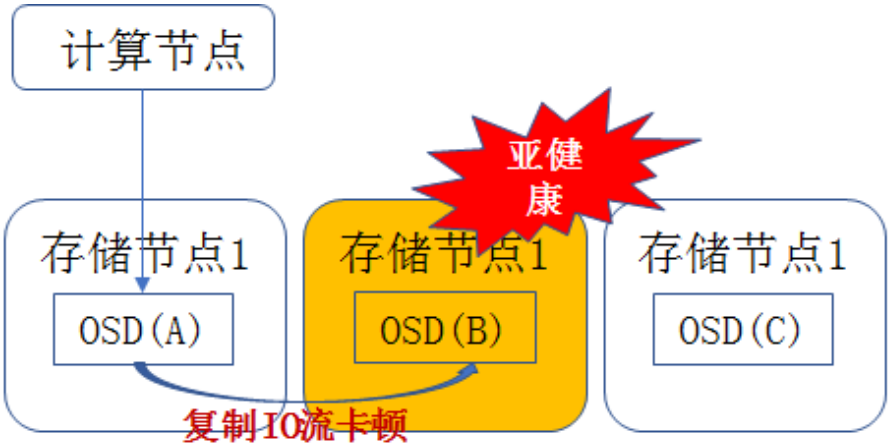
  切换机制：

 版权所有 © 超聚变数字技术股份有限公司

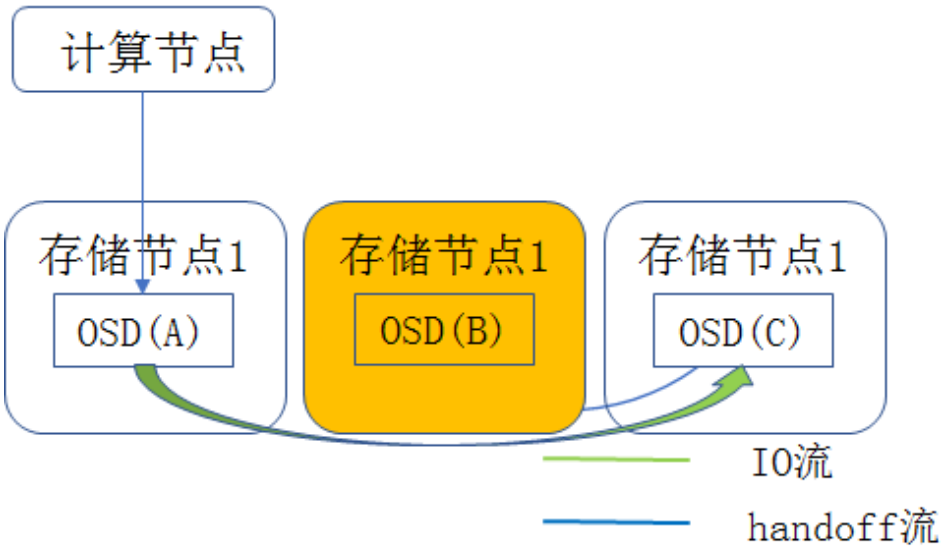
OSD集群控制节点对判断出的亚健康OSD，生成临时视图将其服务切换到其他OSD（Hint节点），主节点的复制IO写到Hint节点后即可返回。

恢复机制：

系统在一定时间后探测亚健康OSD访问时延是否恢复；如果恢复，则将临时数据通过后台Handoff流程推送给原来的亚健康节点，数据传输完成后删除临时视图，将服务切换回原OSD；如果没有恢复，则进行隔离处理。



单点亚健康场景IO流



单点亚健康场景IO流（hinted）

网络亚健康检测与处理

集群网络性能降级，进入亚健康状态，比如网卡降速，丢包/错包率增加等；系统通过检测网络资源状态的变化，定位受到网络亚健康影响的节点，进行bond主备切换或者节点隔离。

检测机制：

系统网络亚健康检测方法：

 版权所有 © 超聚变数字技术股份有限公司

场景		检测方法
节点本地检测	网口闪断	节点本地检测单位时间内网口闪断次数是否超过阈值
	协议栈丢包/错包	节点本地检测单位时间内丢包/错包比例是否超过阈值
	网卡降速	节点本地检测网卡传输速率是否降低
	PCIe降速	节点本地检测PCIe传输速率是否降低
节点间相互检测	节点网络亚健康	通过节点间互ping机制定位网络亚健康节点

对于节点本地检测，节点对本地检测到的网络亚健康状态上报控制节点，控制节点下发网口切换命令或进行节点隔离。

对于节点网络亚健康，系统通过节点间互ping机制进行检测。各节点ping集群内的其他节点，上报时延超过阈值的节点信息到控制节点。控制节点基于收集到的各节点上报的网络亚健康节点信息，按照大多数原则（ping某节点的大多数节点上报该节点网络亚健康）对节点的网络亚健康状态进行判断。控制节点对判断出的网络亚健康节点进行网口切换或节点隔离处理。

切换和隔离机制：

对于网络亚健康节点，系统首先会尝试切换bond主备网口；在非主备bond配置或者切换bond无法恢复的情况下，系统对网络亚健康节点在满足数据冗余的前提下进行节点隔离。

恢复机制：

对于通过互ping机制检测并隔离出的节点，控制节点在一段时间后会尝试将节点重新加入集群。

存储介质亚健康检测与处理

集群节点内出现存储介质的亚健康状态，比如固件问题、机械问题等导致的HDD/SSD访问降速；系统检测受到影响的存储介质，进行隔离。

检测机制：

存储介质亚健康的检测机制包括基于阈值的检测机制和基于同质比较的检测机制：

- 节点在本地对IO时延进行统计，基于阈值判断亚健康OSD，支持定期检测磁盘SMART信息，判断磁盘亚健康情况(硬盘扇区重映射数超过门限、读错误率统计超标、慢盘)。
- 节点上报IO时延统计信息给控制节点，由控制节点进行比较，识别出存储池中比其他OSD慢的 OSD。

隔离机制：

控制节点对判断出的亚健康OSD，在满足数据冗余的前提下进行隔离处理。

恢复机制：

维护人员在解决亚健康问题原因后，通过命令将被隔离的OSD重新加入集群。

 版权所有 © 超聚变数字技术股份有限公司

10.6 硬件可靠性

FusionOne HCI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。

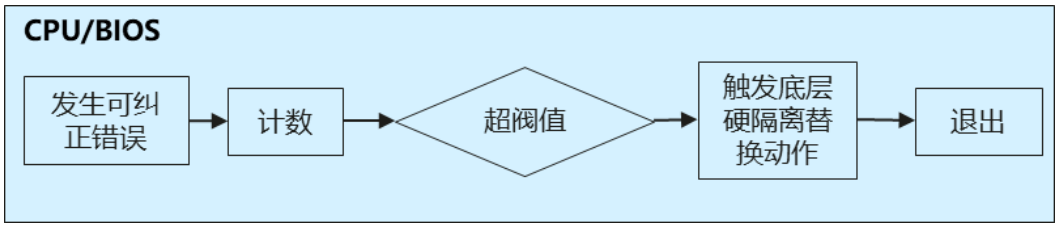
内存可靠性

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：

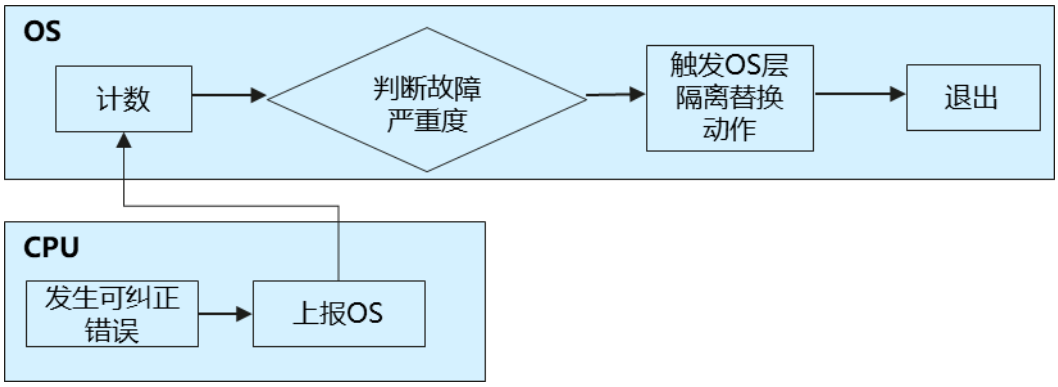
方案1：由CPU内部的纠错机制识别出内存可纠正错误后，BIOS对其计数并与初始设定阈值对比，达到阈值门限后，实施底层隔离，简称硬隔离。



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。

版权所有 © 超聚变数字技术股份有限公司

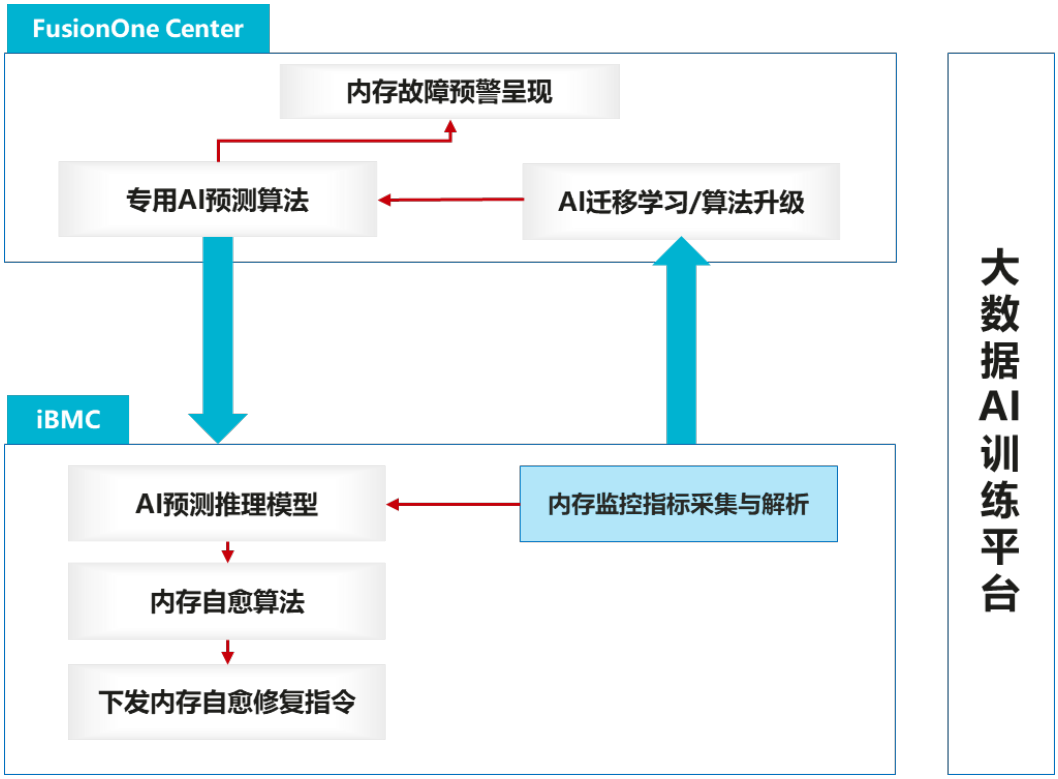


方案缺点：

- 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：

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

 版权所有 © 超聚变数字技术股份有限公司

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。

超融合采用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
- 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

超融合节点配置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。

1. 支持ECC显存，提升GPU的内存数据的可靠性
2. 支持GPU可视化监控
3. 支持GPU关键状态的告警：GPU资源不足、vGPU License失效、GPU内存ECC可纠正、不可纠正错误、GPU丢失、风扇异常、电源异常的告警。

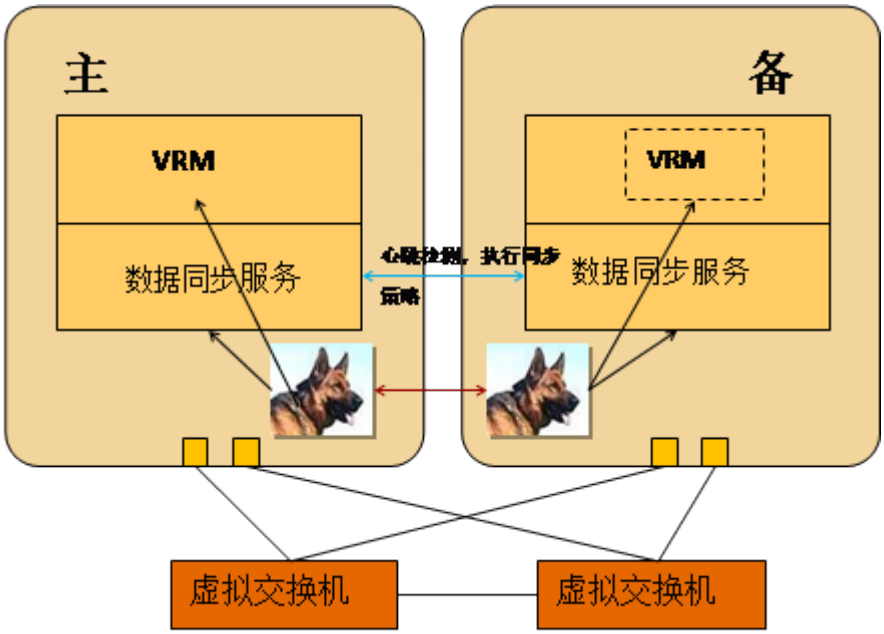
10.7 故障可恢复

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 © 超聚变数字技术股份有限公司

图 10-6 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、分布式存储、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、GPU、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. 分布式存储：支持冗余范围内故障100%无损恢复。
4. 节点操作系统故障：可基于管理平台一键同步下发配置还原主机配置。
5. 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 © 超聚变数字技术股份有限公司

11 安全防护

11.1 安全威胁来源
11.2 平台安全
11.3 业务安全

11.1 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。
- 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。

 版权所有 © 超聚变数字技术股份有限公司

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

  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。

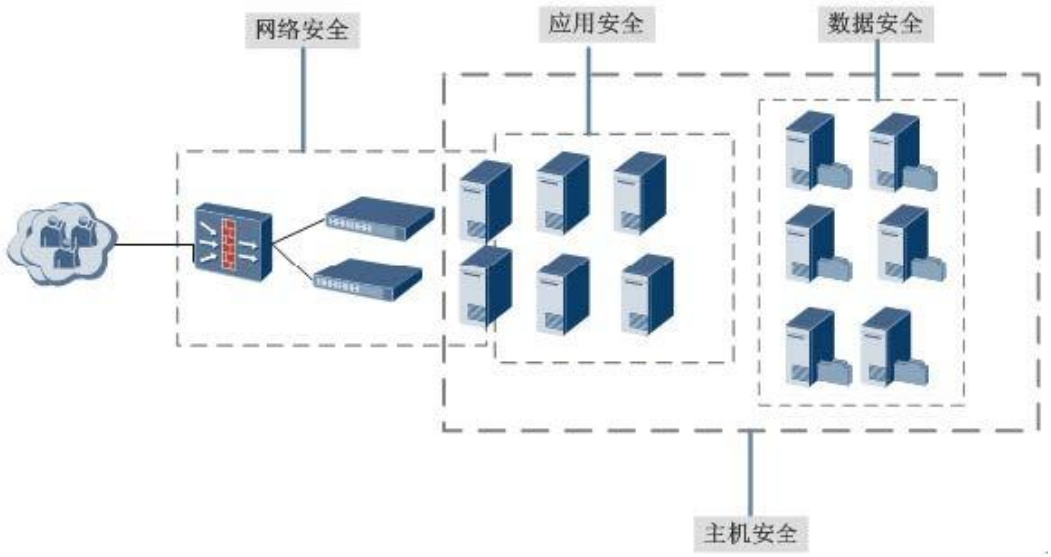
- 管理制度缺乏技术依据，安全策略无法有效落实。

11.2 平台安全

依据系统面临的安全威胁和风险，FusionOne HCI产品提供安全解决方案，如**图11-1**所示。FusionOne HCI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。

 版权所有 © 超聚变数字技术股份有限公司

图 11-1 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

  从身份认证、权限控制、审计控制等方面介绍FusionOne HCI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。

11.2.1 网络安全

11.2.1.1 网络隔离

FusionOne HCI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。
- 存储前端/后端平面

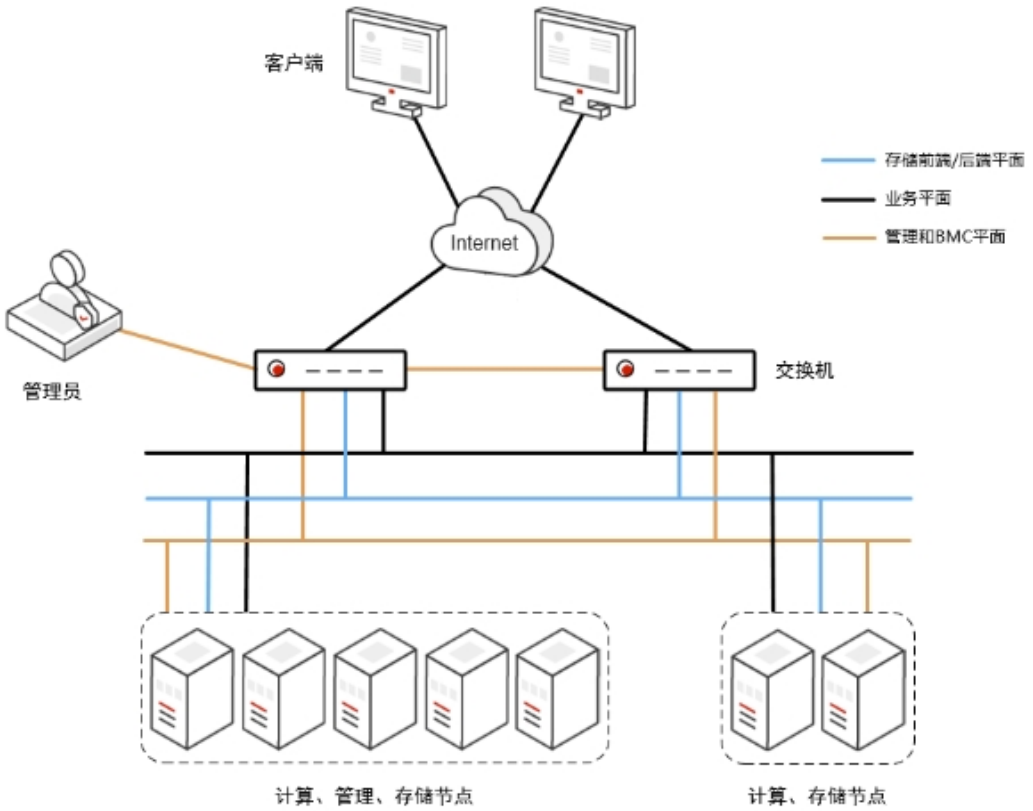
  为虚拟机提供访问存储资源的通信平面，但不直接与虚拟机通信，而通过虚拟化平台转化。
- 管理平面

  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11-2所示。

 版权所有 © 超聚变数字技术股份有限公司

图 11-2 平面隔离示意图



11.2.1.2 平台访问控制

FusionOne HCI平台提供基于管理/存储网段和端口的访问控制策略，确保外部客户端访问HCI平台的安全性，例如，任意客户端都无法访问内部管理服务，只能访问HCI的web平台80/443端口；只允许存储后端网段访问存储后端等。

11.2.2 应用安全

11.2.2.1 权限管理

FusionOne HCI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。

三员模式下可手动创建的用户分配的角色包括：

 版权所有 © 超聚变数字技术股份有限公司

- 只读用户
- 自定义角色
- VMManager（虚拟机管理员）

11.2.2.2 Web 安全

FusionOne HCI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。

11.2.2.3 数据库加固

FusionOne HCI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。

 版权所有 © 超聚变数字技术股份有限公司

- 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。

11.2.2.4 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。

11.2.3 主机安全

FusionOne HCI中计算节点、存储节点、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。

11.2.4 数据安全

FusionOne HCI通过多种存储安全技术保证存储的用户数据安全可靠。

- 数据分片存储

  FusionOne HCI存储节点上的数据会自动保存多份，每一个分片的不同副本也被分散保存到不同的存储节点上，恶意用户无法利用单个存储节点或物理磁盘获取用户数据。

- 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。

11.3 业务安全

FusionOne HCI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，主要包括基础备份、CDP备份、

 版权所有 © 超聚变数字技术股份有限公司

11.3.1 虚拟机安全删除

安全删除功能是为了保护超融合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。

FusionOne HCI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，由于采用分布式存储打散的方式存储，风险可控。

11.3.2 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。

1. **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。
2. **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：

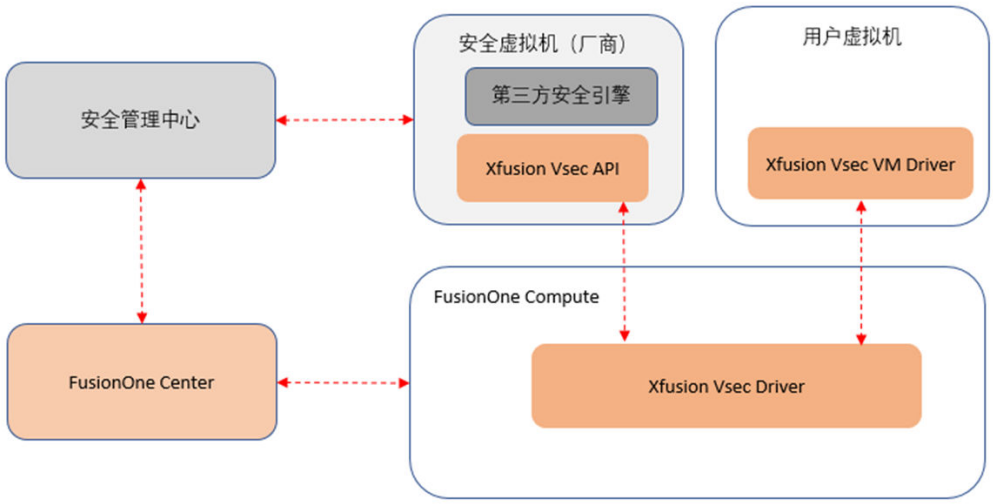
 版权所有 © 超聚变数字技术股份有限公司

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）。

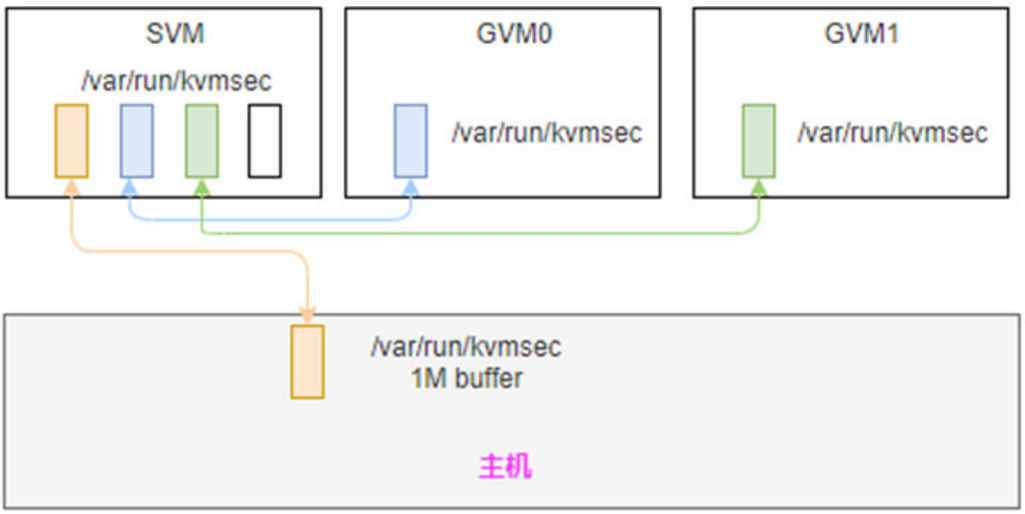
FusionOne HCI 方案

FusionOne HCI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 HCI无代理原理图，FusionOne HCI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。

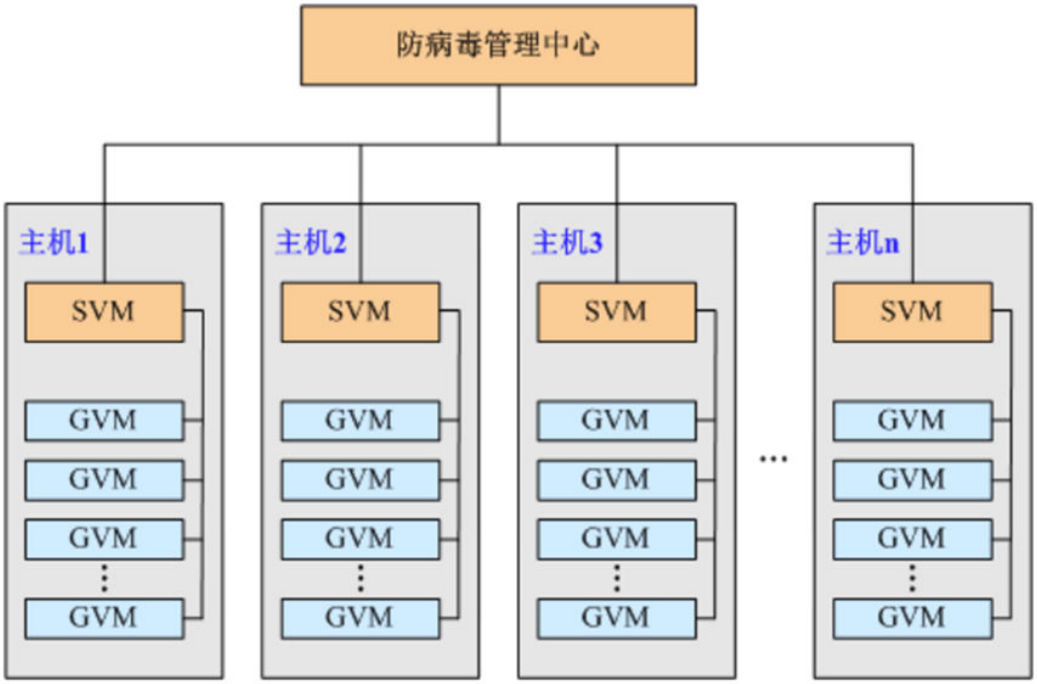


在SVM与GVM、主机之间的通信主要通过内存共享（ivshmem）设备实现。



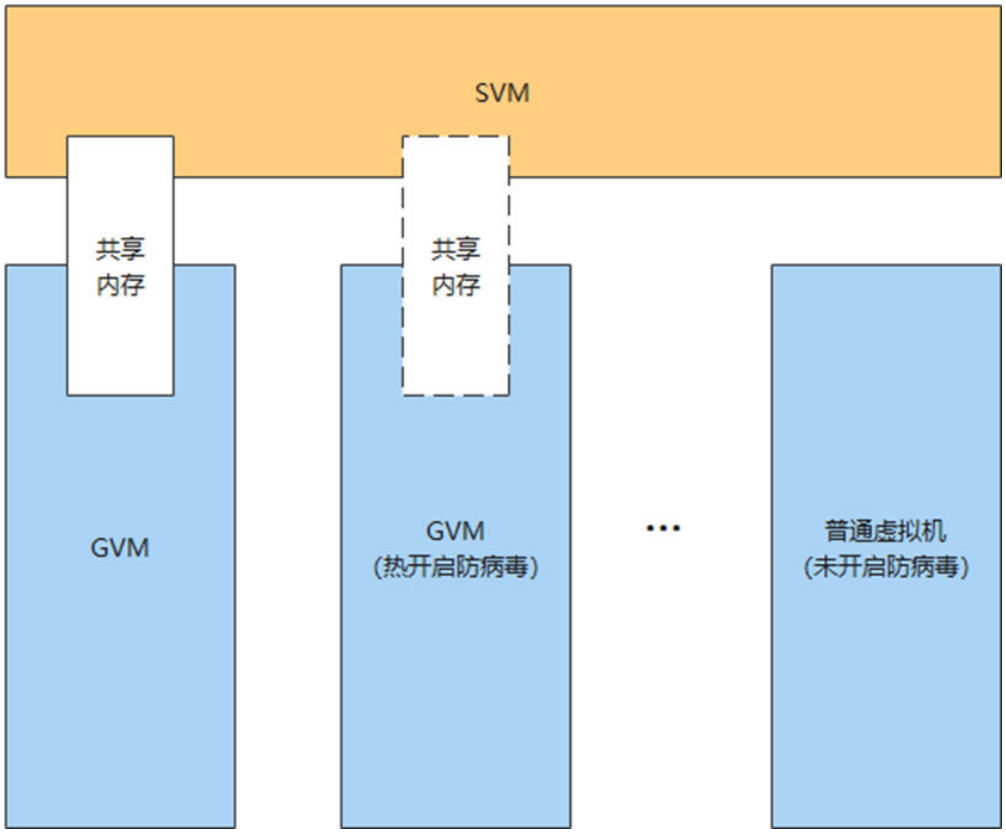
以下是以防病毒部署部署结构图：

版权所有 © 超聚变数字技术股份有限公司

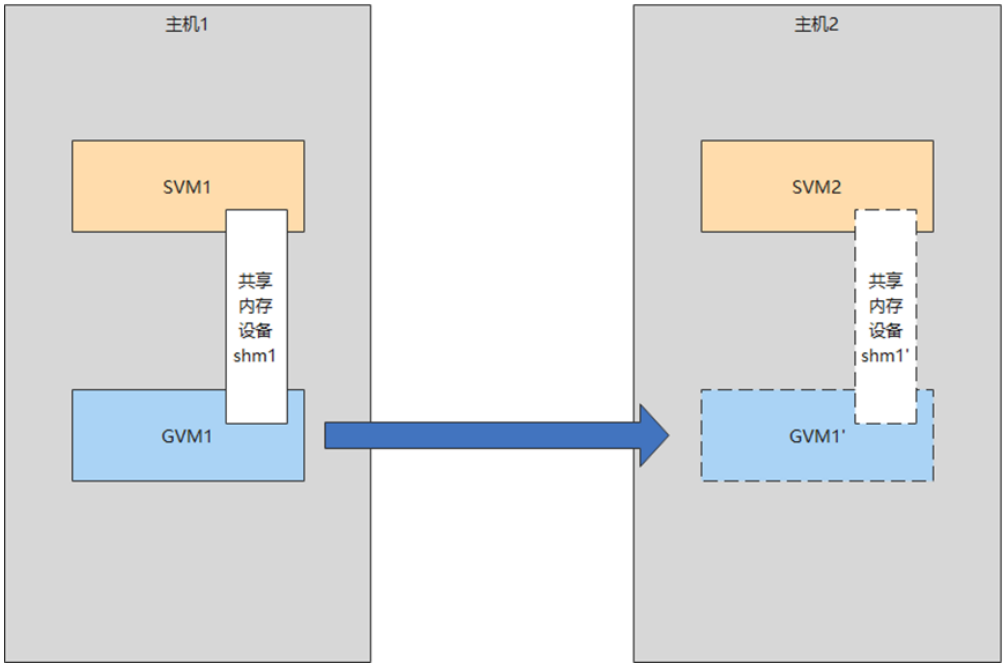


- SVM: 安全服务虚拟机由防病毒厂商提供。
- GVM: 安全用户虚拟机所安装的防病毒驱动由防病毒厂商提供。
- GVM：仅支持部分Windows的虚拟机作为安全用户虚拟机。
- 虚拟化防病毒功能所提供的具体防病毒服务及性能由防病毒厂商提供的防病毒引擎的能力决定。
- 安全虚拟机热开启，可以在不中断业务的情况下开启虚拟机防病毒功能；SVM通过共享内存与GVM进行通信，对普通虚机不停机状态下开启需要热添加一个共享内存设备。

 版权所有 © 超聚变数字技术股份有限公司



- **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。**



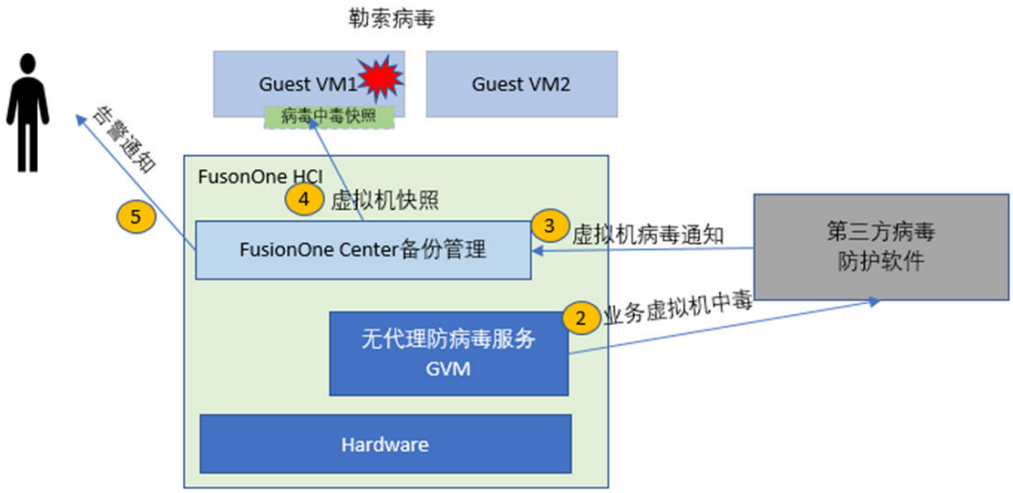
 版权所有 © 超聚变数字技术股份有限公司

11.3.3 勒索病毒联动处置

当前勒索病毒对业务的危害越来越多，并且无法第一时间感知，FusionOne HCI通过与杀毒软件联创，实现检测到中毒立刻采取措施，保护数据安全，并及时告警通知管理员。

通过历史未中毒、以及中毒后的快照结合进行文件级手动恢复。

工作流程如下图：



1. 第三方防病毒或者防火墙检测到业务虚拟机中毒等事件后，自动提醒平台中毒的虚拟机。
2. 平台自动对病毒的虚拟机进行自动快照，记录虚拟机中毒的现场，并自动触发告警。

使用约束：

1. 第三方病毒软件需要适配调用超融合平台接口中毒通知，目前支持瑞星杀毒。
2. 快照虚拟机中的内容已中毒的快照，该快照不会被定时快照覆盖。
3. 如果需要还原到未中毒状态，只能通过定时快照解决。

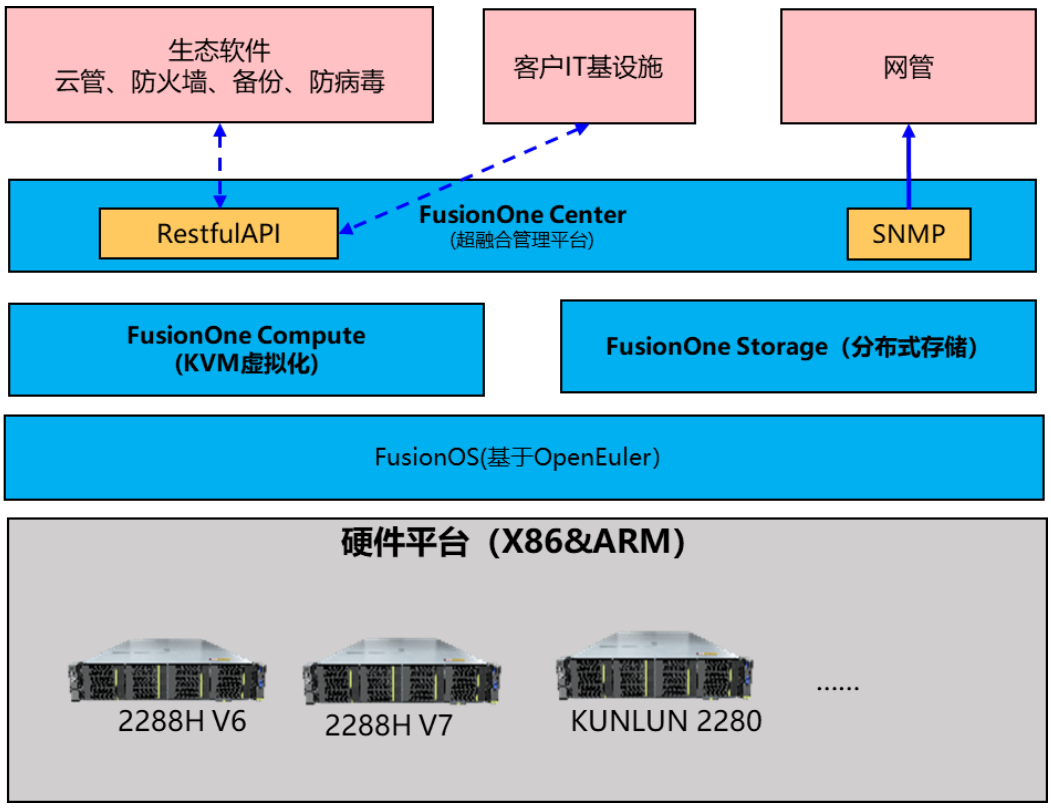
 版权所有 © 超聚变数字技术股份有限公司

12 北向接口与开放兼容

12.1 北向接口
12.2 应用市场
12.3 Oracle RAC虚拟机部署
12.4 信创超融合

12.1 北向接口

FusionOne HCI生态开放，为生态伙伴、最终客户提供了全面的接口支持，包括超融合平台的RestAPI接口、SNMP告警对接接口。



 版权所有 © 超聚变数字技术股份有限公司

主要在如下应用场景：

- 客户定制化业务发放

  可根据业务需求，将超融合平台融入到客户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，将超融合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

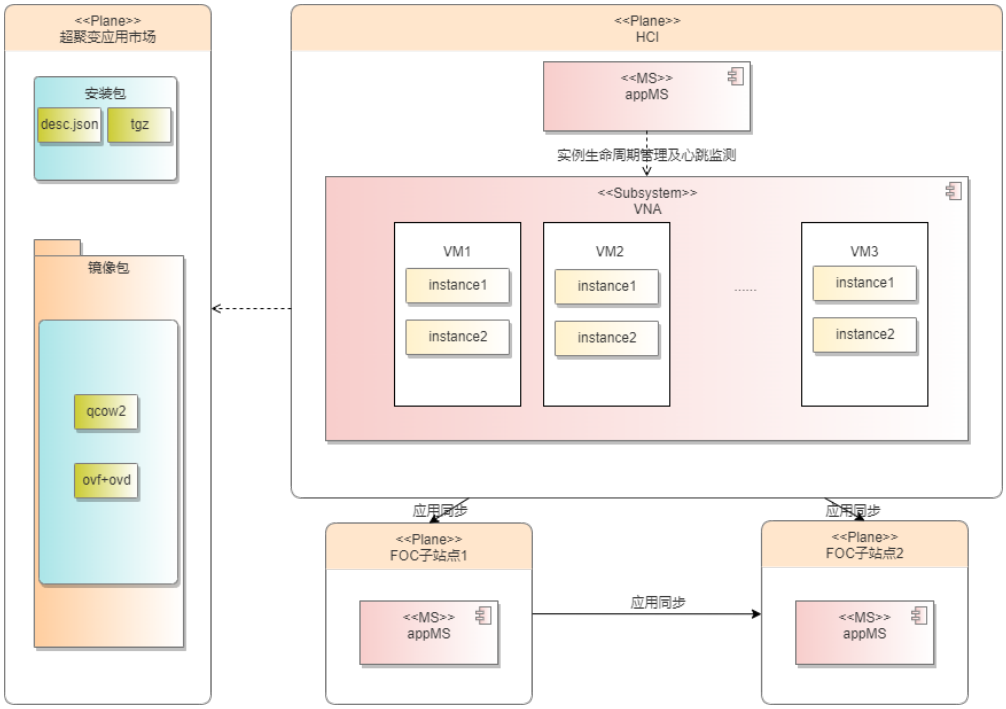
  超融合平台相关的无代理防火墙、无代理防病毒、备份、云管等主流软件已经与FusionOne HCI对接兼容，可为客户提供一整套的IT解决方案。

12.2 应用市场

FusionOne Center提供本地应用市场管理功能，完成HCI基础设施与生态伙伴企业应用的深度集成，可以让用户如同使用手机应用市场一般方便的使用生态伙伴的企业应用，从而帮助生态伙伴用好超融合基础设施，服务好客户。主要支持如下功能：

- 应用软件包管理：从超聚变应用市场获取生态应用软件包和签名文件后，支持在FusionOne Center上传，删除，查询等。
- 应用快速上线：支持在FusionOne Center界面通过页面向导式快速部署应用。
- 应用实例监控管理：支持对已部署的所有生态应用进行资源和状态监控以及实例的生命周期管理。
- 应用多站点间同步：支持任意站点之间的应用包同步。

  以下是应用市场架构：



 版权所有 © 超聚变数字技术股份有限公司

12.3 Oracle RAC 虚拟机部署

Oracle RAC是数据库集群技术的一种实现方式，通过在多台服务器间进行共享存储和通信，实现多台服务器共同管理一个数据库的目的。业界通常将Oracle RAC部署在物理机上，物理机部署Oracle RAC虽然可以带来高性能、高可用等特性，但是由于部署在物理机，物理机的可靠性对数据库的稳定运行是一个考验，服务器的部件故障、维护更换都会影响到数据库的持续运行。

虚拟化部署数据库可以很好的帮助客户解决了物理机带来的稳定性、可靠性挑战，让企业用更轻松的方式去解决数据库的使用需求。客户可以节省投入，将原先用来做硬件平台稳定性、可靠性的人员，包括研发、DBA和相关技术人员，可以更加聚焦数据库业务本身，提高了数据库的上线效率、省时省力。

FusionOne HCI平台具备灵活扩展、高可用、高可靠等优势，在超融合上部署Oracle RAC，则可以很好地解决物理机部署的问题，同时结合超融合的DRS、HA、亲和性、故障检测恢复、反亲和性、网络虚拟化等机制，让Oracle节点的可靠性更上一层楼。

FusionOne HCI通过如下技术优化，给Oracle RAC带来接近物理机的性能，同时具备更高的可靠性。

- 弹性资源

  支持Oracle RAC中的虚拟机规格动态调整，如CPU、内存、磁盘、网卡均可以动态扩容，大大减少物理机部署带来的。

- 动态调度DRS

  基于CPU、内存负载动态选择负载低的主机运行虚拟机，确保Oracle RAC虚拟机运行在负载最优最优的主机上。

- 故障HA切换

  可配置虚拟机故障策略，如虚拟机内死机，可自动检测重启虚拟机，防止Oracle节点持续挂死，大大缩短数据库宕机的时间。

- 数据库节点互斥

  Oracle RAC集群的节点，可以通过配置反亲和性，将数据库集群的各个节点分别运行在不同的主机，确保业务的可靠性。

- 虚拟共享存储

提供基于多路径的存储，单个存储路径故障，可以动态切换到其他路径，保证共享卷数据持续在线提供卷服务。

- 网络调优：

  支持配置网卡的MTU，对数据库的私有网络提供高效的巨帧通信通道。

  支持对网络流量QoS，可以对业务网口的数据通信优先级进行配置，确保Oracle RAC的数据优先通行。

- I/O高性能：

  支持NVME SSD缓存加速技术，大大提升Oracle RAC节点的IO读写性能。全闪NVME存储池可提供百万级IOPS性能，IO吞吐量可达GB级的IO带宽。

- 大页内存优化

  支持配置大页内存分配给数据库节点，减少操作系统页表的切换，可大大提升事物处理性能，开启大页后，性能可提升30%以上。

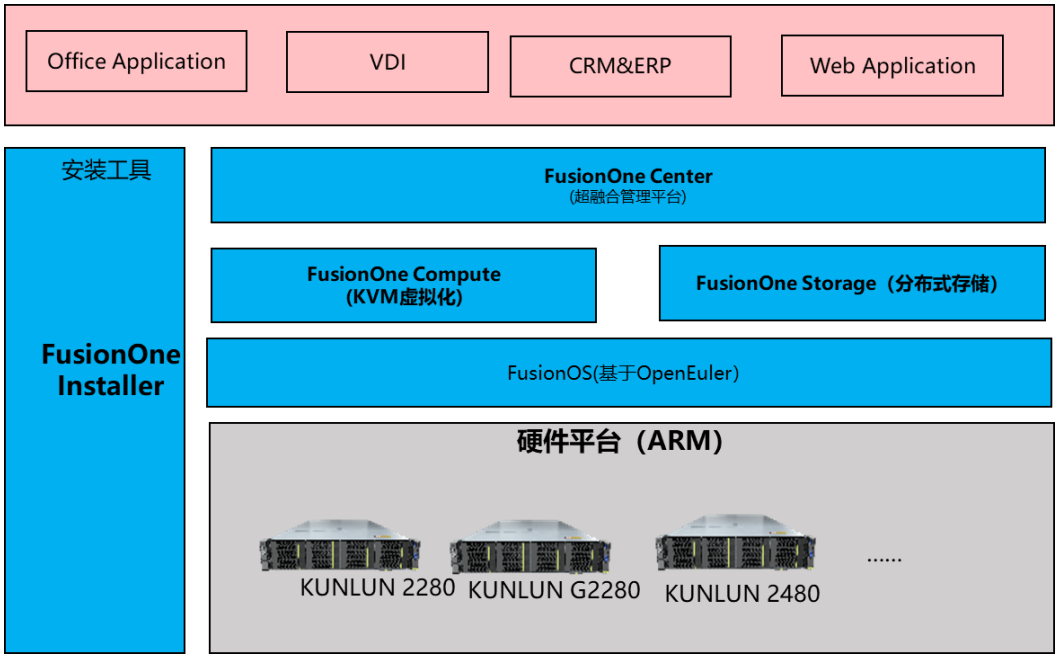
 版权所有 © 超聚变数字技术股份有限公司

12.4 信创超融合

国产化平台

FusionOne HCI支持支持信创领域的商用，通过从硬件平台到超融合平台，均使用自研的全国产化组件，为政府、金融等信创关键领域保驾护航。

以下是信创超融合的架构：



整体平台采用全自研的国产化方案，服务器基于KUNPENG芯片的ARM服务器，操作系统为超聚变基于FusionOS专门为超融合优化的虚拟化OS。

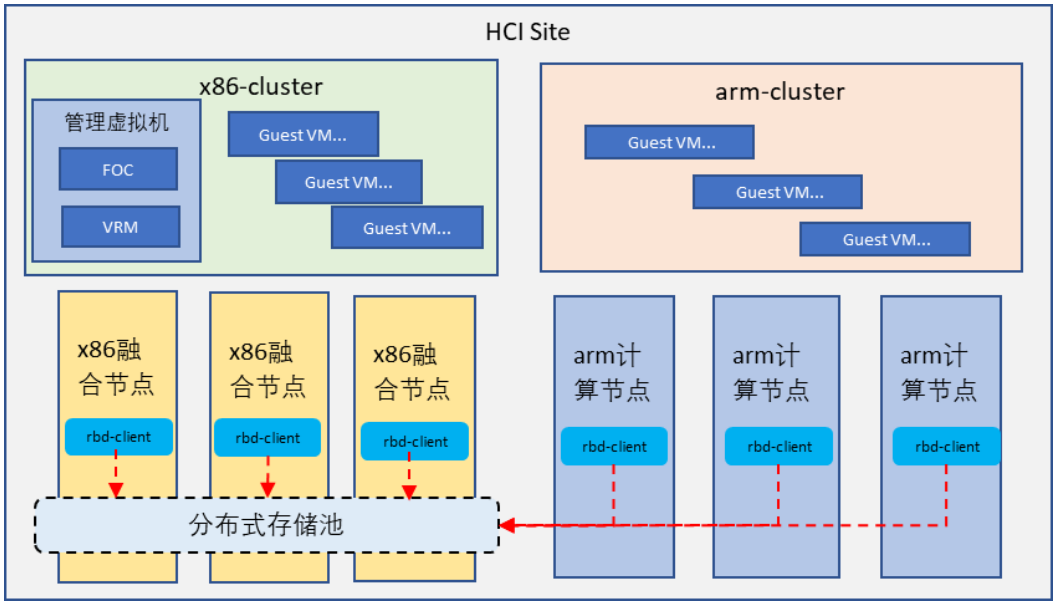
分布式存储、虚拟化平台、FusionOne Center均为国产自研软件，并已经通过了《Q/KXY HC004面向信创的超融合平台技术》标准认证。

平滑演进

FusionOne HCI支持混合部署，可根据业务实际情况，选择部分在X86集群运行，部分在ARM上运行，统一管理保护客户投资，让用户可以渐进式从X86生态演进到ARM生态。

版权所有 © 超聚变数字技术股份有限公司

图 12-1 混合配置示意图

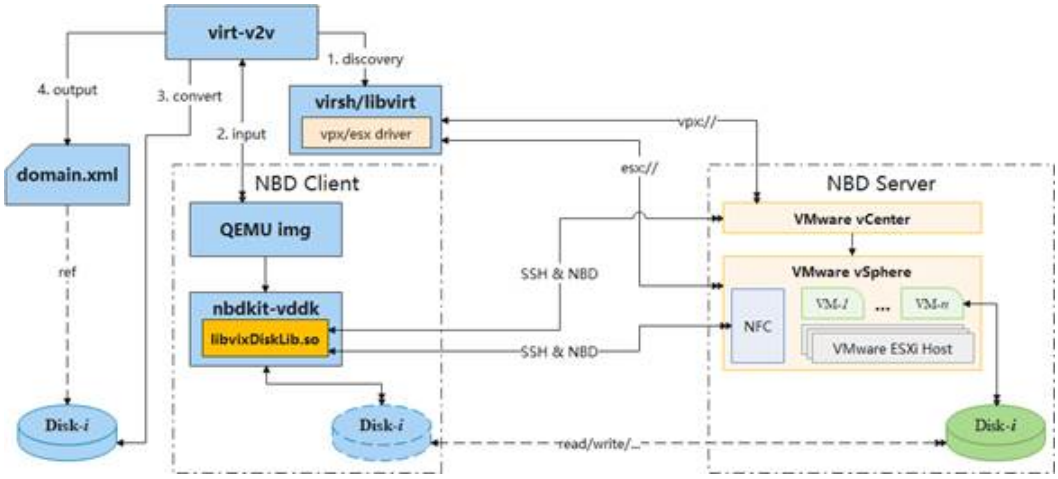


- ARM架构部署：HCI支持部署在ARM架构的物理机上，整体功能与X86一致。
- X86与ARM混部：管理组件部署在同一个CPU架构，分布式存储部署在同一个CPU架构。
- 异构集群管理：支持在一个超融合站点内创建ARM、X86两种不同架构的计算集群，实现硬件资源高度集约的同时，保证集群操作的一致性。

VMware 纳管

FusionOne HCI内置多站点管理，支持对VMware平台的纳管，并可通过统一的界面管理，包括告警监控、业务虚拟机生命周期管理，同时内置的迁移功能，客户可根据业务运行情况，将VMware平台的虚拟机温迁移到FusionOne HCI。

如下是FusionOne HCI对VMware虚拟机迁移的技术原理介绍：



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；

 版权所有 © 超聚变数字技术股份有限公司

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。

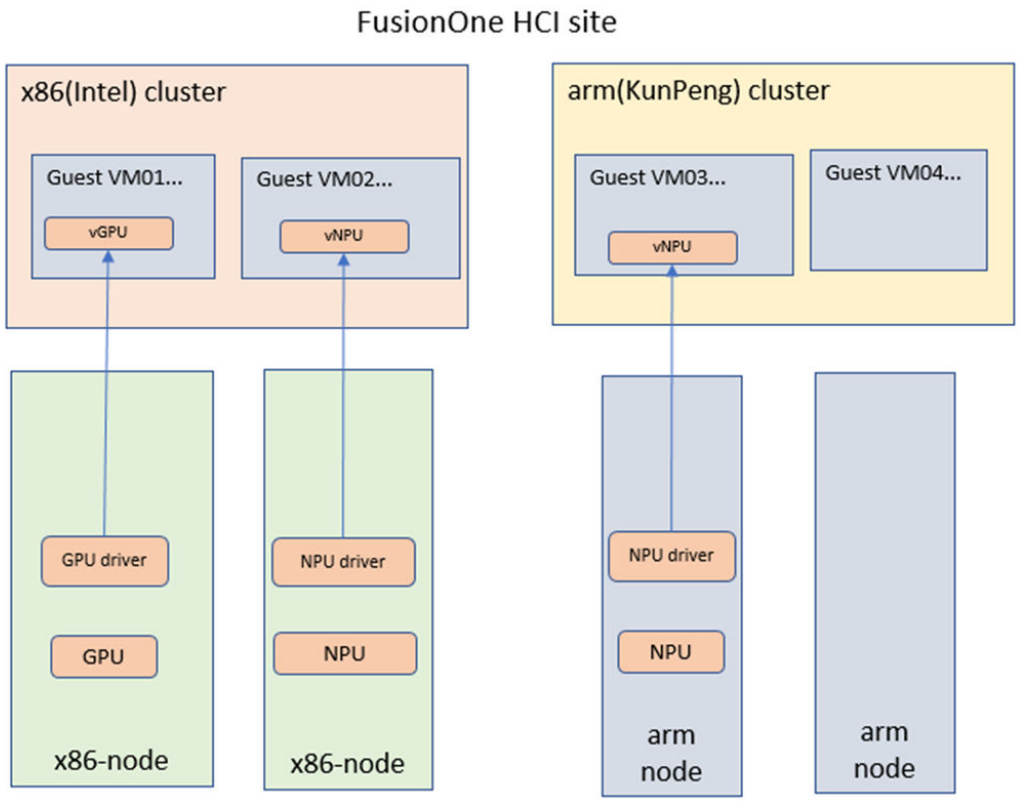
NPU、GPU 混合加速

FusionOne HCI将GPU资源作为全局资源，可根据所在节点配置的NPU、GPU池化，可在对应主机灵活配置NPU、GPU的虚拟化。可在x86、arm的节点上配置NPU，并虚拟化给客户虚拟机使用，可让客户图形计算加速等业务平滑运行。

表 12-1 各种组合对比

组合方式	业务场景
x86节点+GPU	切换信创前的状态，用于部分业务软件无法切换的场景
x86节点+NPU	用于图形加速可切换，但是业务软件无法切换的场景
arm节点+NPU	全面信创场景，信创的最终形态。

如下图为GPU、NPU混插的示例图：



 版权所有 © 超聚变数字技术股份有限公司

13 系统规格

规格项	规格小类	X86(Intel)	X86(AMD)	X86(Hygon)	ARM(鲲鹏)
系统支持的节点规模	支持的节点数	2~256	2~256	2~256	3~256
镜像仓库存储大小	-	1TB	1TB	1TB	1TB
管理虚拟机规格	推荐规格（3~100节点）	FOC：4U12GB VRM：8U8GB	FOC：4U12GB VRM：8U8GB	FOC：4U12GB VRM：8U8GB	FOC：8U16GB VRM：8U8GB
	大规模（>100节点）	FOC：8U16GB VRM：12U12GB	FOC：8U16GB VRM：12U12GB	FOC：8U16GB VRM：12U12GB	FOC：8U16GB VRM：12U12GB
多站点管理	最大纳管的分支站点数	500	500	500	500
单个VRM集群支持的最大虚拟机数量	-	5000	5000	5000	5000
单个虚拟化逻辑集群支持的主机数量	-	64	64	64	32
VIMS集群支持的最大主机数	-	32	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	768	128

版权所有 © 超聚变数字技术股份有限公司

	每物理主机支持的最大物理内存	16TB	16TB	16TB	4TB
	每物理主机支持的最大虚拟机数量	1024	1024	1024	1024
	每物理主机最大挂载LUN数量	1024	1024	1024	511
	每物理主机最大vCPU数	4096	4096	4096	384
	每物理主机最大虚拟网卡数	2048	2048	2048	2048
	每物理主机支持的最大虚拟磁盘数	2048	2048	2048	2048
	每物理主机的NUMA节点数	16	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	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	50
	单个分布式交换机管理的最大主机数量	256	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	10000
	SDN交换弹性IP大地址池数	20	20	20	20
	SDN交换每个地址池最大段数	10	10	10	10
	SDN交换网络ACL规则最大数	200	200	200	200

 版权所有 © 超聚变数字技术股份有限公司

	SDN交换网络ACL每个规则最大关联子网	150	150	150	150
	SDN交换NAT网关最大个数	10	10	10	10
	SDN交换每个NAT网关支持最大规则数	200	200	200	200
虚拟机规格需求	单个虚拟机支持的vCPU数量	255	255	255	128
	单个虚拟机支持的网卡数量	16	16	16	16
	单个虚拟机支持的磁盘数量	60	60	60	60
	单个虚拟机支持的内存容量	6TB	6TB	6TB	256GB
	虚拟机支持的分布式存储单磁盘规格	32TB	32TB	32TB	32TB
	虚拟机支持的本地磁盘单磁盘规格	16TB	16TB	16TB	16TB
	虚拟机支持的VIMS单磁盘规格	64TB	64TB	64TB	64TB
	虚拟机支持的vGPU数量	1	1	1	1
	单个虚拟机支持的快照数量	32	32	32	32
分布式存储规格	系统支持的存储池个数	系统最大支持128个物理存储池，单个存储池推荐服务器节点数≤40个	系统最大支持128个物理存储池，单个存储池推荐服务器节点数≤40个	系统最大支持128个物理存储池，单个存储池推荐服务器节点数≤40个	系统最大支持128个物理存储池，单个存储池推荐服务器节点数≤40个

 版权所有 © 超聚变数字技术股份有限公司

	MON监控节点部署个数	1）融合节点数量≤10台，推荐部署3个MON 2）10 < 融合节点数量≤40台，推荐部署5个MON 3）融合节点数量 > 41~128台，推荐部署7个MON	1）融合节点数量≤10台，推荐部署3个MON 2）10 < 融合节点数量≤40台，推荐部署5个MON 3）融合节点数量 > 41~128台，推荐部署7个MON	1）融合节点数量≤10台，推荐部署3个MON 2）10 < 融合节点数量≤40台，推荐部署5个MON 3）融合节点数量 > 41~128台，推荐部署7个MON	1）融合节点数量≤10台，推荐部署3个MON 2）10 < 融合节点数量≤40台，推荐部署5个MON 3）融合节点数量 > 41~128台，推荐部署7个MON
	数据冗余保护机制	2副本、3副本（推荐）、EC	2副本、3副本（推荐）	2副本、3副本（推荐）	2副本、3副本（推荐）、EC
网络负载均衡器	负载均衡器数量	50（个）	50（个）	50（个）	50（个）
	单负载均衡器实例数量	2（个）	2（个）	2（个）	2（个）
	单负载均衡器实例虚拟机默认规格	4C50G	4C50G	4C50G	4C50G
	监听器数量	100（个）	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）	30（个）
	监听器转发策略数量	100（个）	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）	5（条）
	单转发策略下可添加转发规则数量	10（条）	10（条）	10（条）	10（条）
	单转发规则可后端服务器组数量	1（条）	1（条）	1（条）	1（条）

版权所有 © 超聚变数字技术股份有限公司

	后端服务器组数量	30（个）	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）	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）	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）	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）	1-20（秒），默认2（秒）

 版权所有 © 超聚变数字技术股份有限公司