云宏
WINHONG
自主安全的中国云

CNware WinSphere

产品白皮书

云宏服务器虚拟化产品

日期：2024-03

最新的技术文档可以从 Winhong 网站下载：http://www.winhong.com/

该网站还提供最近的产品更新信息。

您如果对本文档有任何意见或建议，请反馈至： support@winhong.com

版权声明

免责声明

本文档仅提供阶段性信息，所含内容可根据产品的实际情况随时更新，恕不另行通知。如因文档使用不当造成的直接或间接损失，本公司不承担任何责任。

云宏信息科技股份有限公司

电话：020-28260045

热线：400-6300-003

地址：广州市天河区元岗横路 33 号天河慧通产业广场 B2 栋

邮编：510650

目录

第1章 摘要

数字化转型浪潮席卷各行各业，以云计算、大数据、人工智能等新技术为首不断推动企业的 IT 架构与业务模式革新。新一代数据中心通过虚拟化、云计算技术对基础设施资源动态调度和分配，解决传统 IT 架构的各类问题，如资源利用和共享率低、采购及运营成本高居不下、业务上线周期长等，帮助企业迈向敏捷架构、获取新技术栈带来的价值收益，从而进一步降本增效，保障业务质量与安全。

近年来信息安全事件层出不穷，如“棱镜”项目的曝光，使得金融、大型政企越发重视 IT 安全、可靠；特别是中兴华为事件、中美贸易战深深震撼国人的内心，国产自主信息化发展不仅仅是改造浪潮，而是国家意志的体现。从核心的芯片、终端等硬件到软件及应用生态全面发展的信创领域，涌现了大批自主自强企业和技术。在服务器硬件侧，崛起了鲲鹏、飞腾、龙芯、海光、兆芯等国产芯片及整机，而开放生态的软件应用也正如火如荼地蓬勃发展；虚拟化作为关键的技术，实现软硬件的解耦，必然为国产自主软硬件生态注入充沛的活力，加速国产替代建设进程。

云宏自主知识产权的虚拟化云平台 CNware WinSphere，为客户提供业界领先、功能全面的企业级虚拟化云平台解决方案。CNware WinSphere 致力于构建稳定、高效、可靠的虚拟化 IT 基础架构，全面兼容 X86 体系、ARM 体系、MIPS、Alpha 体系芯片，提供虚拟化平台的整体解决方案，全面支撑传统和新型的企业服务，极大地提升资源使用和运营维护效率，帮助企业降低成本、创造更多价值。

本文档向您讲述云宏 CNware WinSphere 产品中的整体架构、功能特性和功能详解，通过阅读本文档，您能够了解到：

- CNware WinSphere 虚拟化云平台解决方案的整体软件架构和产品构成
- CNware WinSphere 是如何实现资源的虚拟化、自动化统一管理
- CNware WinSphere 中各产品组件的运行原理、技术要点和功能特性
- CNware WinSphere 部署方案和产品规格

第2章 产品介绍

2.1.产品定位

CNware WinSphere 是一款安全稳定、业界领先的虚拟化产品，致力于打磨可媲美国外虚拟化产品 VMware vSphere/vsan 的企业级虚拟化特性并无缝替代，实现计算存储基础设施资源池化、弹性调度及高可靠。

CNware WinSphere 构建了完全自主可控、软硬件生态开放的平台，广泛互认证适配信创云计算上下游产业生态，全面实现一云六芯（鲲鹏、飞腾、龙芯、海光、兆芯、申威），12家国产操作系统（统信、麒麟等），数十家主流数据库中间件（达梦、金蝶等）、十余家备份安全（奇安信、壹进制等）到数百行业应用（政务、金融等）的全方位兼容性和协同性优化，极致发挥硬件的性能，提供资源的全生命周期管理、资源策略调度、高可用保护、故障检测等能力。

与业界为数不多的专精于虚拟化的产品相比，CNware WinSphere 完全满足信创虚拟化标准规范（T/CESA 9163—2020），具有金融级可靠特性、内核级研发、软硬件充分解耦、全面兼容开放等优势，实现多芯片技术架构一致管控的体验，全面保证业务的安全性和连续性，拥有广泛的金融、政务、运营商、军工等行业的成功案例。

2.2.功能架构

图 2-1 产品架构图

产品核心包括：

WinServer：即 Hypervisor，提供虚拟化核心引擎能力

WinCenter：虚拟化管理平台

2.3.关键技术

2.3.1. 服务器虚拟化技术

完全内核级自主研发的服务器虚拟化技术 WinServer，采用裸金属虚拟化架构，直接在服务器硬件上加载运行，是一种高效可扩展的虚拟化系统，支持 VT-X、AMD-V、VE/VHE 等

硬件辅助虚拟化技术。

全面支持鲲鹏（kunpeng916/920）、飞腾(FT1500a/FT2000PLUS/S2500)、龙芯(3000/4000/5000 系列)、海光（5000/7000 系列）、兆芯（KH-30000 系列）、申威（SW3231）等国产芯片服务器，帮助客户快速搭建符合信创自主要求的虚拟化环境。

2.3.2. 集群自动化管理技术

集群管理技术主要实现对物理资源、虚拟资源的统一管理。通过集群，可以像管理单个实体一样轻松地管理多个主机和虚拟机，从而降低管理的复杂度。同时，系统将定时对集群内的主机和虚拟机状态进行监测，保证了数据中心业务的连续性。例如，当一台服务器主机出现故障时，运行于这台主机上的所有虚拟机都可以在集群中的其它主机上重新启动；在例如，当虚拟机发生 kernel panic、蓝屏崩溃等故障，虚拟机 qemu 进程将迅速被杀死并重新拉起，业务在数秒内重新恢复；以上是一种经济有效的保障业务连续性的方案；

2.3.3. 大规模智能资源调度技术

通过独有的基于门限的虚拟资源重配置优化算法、基于负载预测的资源调度算法，实现虚拟化数据中心资源的智能弹性调度和按需分配。当前的资源调度技术可应用于高达 5000 台以上的物理服务器和 10000 台以上的虚拟服务器组成的集群。主要应用于集群发生物理机高可用时虚拟机资源的重新分配，集群资源的负载均衡以避免物理服务器被过度使用，同时侦测并智能管理服务器能源从而达到绿色节能等场景。

2.3.4. 动态网络策略管理技术

采用业界主流的软件定义网络技术构建多层、高性能虚拟网络。通过支持可编程扩展实

现大规模的网络自动化和 VPC 隔离网络模型，动态管理和配置虚拟机网络；持续监控物理主机和虚拟机的网卡性能，通过策略管控虚拟机网卡流量出入方向、优先级、Qos 等，支持如 Netflow、sFlow、CLI 等多种标准管理接口，支持 vlan、vxlan、geneve 等隧道封装技术。

2.3.5. 多存储自适应技术

存储池提供虚拟机磁盘、配置文件、镜像文件的保存位置，主要分为本地存储、共享存储类型。WinSphere 研发的多存储自适应层以适应客户现有的基础设施为导向，包括本地目录、LVM、基于 iSCSI/FC LUN 的 LVM、NFS、共享文件系统等类型，满足绝大部分客户的存储需求；虚拟机选择存储池时，根据存储类型自动匹配总线、缓存、精简或厚置备等最佳参数配置，提供性能至优的磁盘存储。

2.3.6. 微服务高可用架构

管理平台采用先进的微服务架构进行设计开发，可拆分成多个解耦的模块并独立提供服务。系统中的各个微服务可被独立部署，各个微服务之间是松耦合的，可以独立开发演化、运行和扩展，实现平台组件化、松耦合、自治、去中心化，快速与企业现有系统集成以及迭代上线新特性。

2.4.产品特点

2.4.1. 生态开放全面

CNware 专注于持续加码虚拟化层的稳定、安全、好用，提高竞争壁垒，衔接云计算产业链上下游，通过开放整合生态伙伴来帮助客户构建健壮的虚拟化数据中心，降低兼容性风

险且保护已有投资，最终提高云的弹性和业务上线效率。

云宏提前入局并极力拥抱信创生态，目前是信创互认证适配最全面且业界领先的云厂商之一。除了 Intel、AMD 等 X86 架构的芯片，CNware 在鲲鹏（kunpeng 916/920）、飞腾(FT1500a/FT2000PLUS/S2500)、龙芯(3000/4000/5000 系列)、海光（3000/5000/7000 系列）、兆芯（KH-30000 系列）、申威（SW3231）完成兼容互认证和深度调优。尤其与鲲鹏服务器在 BMC 整合、BoostKit 调优方面深度融合（2021 年 4 月作为首家虚拟化厂商获得鲲鹏 Validated 严选认证），充分发挥算力为应用加速；2020 年 11 月与飞腾最新发布的 S2500 芯片率先完成适配并入选某国有大行开放平台；2021 年 6 月与龙芯 5000 系列（LoongArch）率先适配且提供业内极少数的虚拟化方案；2021 年 9 月与海光 7000 系列适配方案因性能、适应性等方面业界领先荣获 2021 光合组织优秀解决方案大赛金融赛道一等奖。

CNware 的生态版图包括但不限于：

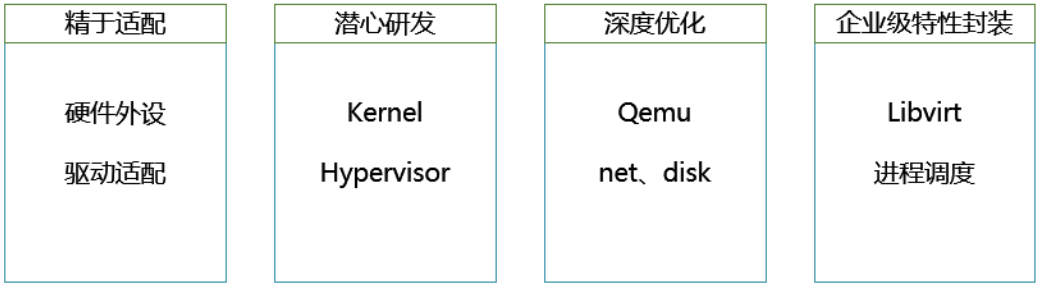
类型	适配范围
芯片	X86：Intel、AMD 信创： 鲲鹏（kunpeng 916/920） 飞腾(FT1500A/FT2000PLUS/S2500) 龙芯(3000/4000/5000 系列) 海光（3000/5000/7000 系列） 兆芯（ZX/KX/KH 系列） 申威（3231）
整机	华为泰山、广电鲲鹏、长江计算、黄河鲲鹏、四川长虹天宫、神州鲲泰、南京坤前、华诚金锐、北京计算机技术及应用研究所、联想、浪潮、中科曙光、紫光、清华同方、长城擎天、长城超云、天熠、天玥、柏科、五舟、宝德、北联国芯、华山、H3C、海康威视、百信等
操作系统	国内：统信、麒麟、深度、普华、新支点、中科方德、国心、同源、红旗、一铭、万里红、欧拉、拓林思、凝思等 国外&社区：Redhat、CentOS、Fedora、Debian、Ubuntu、Windows 等
数据库	国内：达梦、人大金仓、神舟通用、南大通用、瀚高、万里开源、爱可生、热璞、优炫、云和恩墨、虚谷伟业、海量数据、巨杉等 国外：Oracle、DB2、Mysql 等
中间件	国内：中创、金蝶天燕、东方通、华宇、普元、宝兰德 国外：IBM WebSphere、MQ、Tomcat、Apache 等

行业应用	政盟、华迪、太极、南威、中电福富、航天开元、致远互联、蓝凌、泛微、合明监控、久远银海、数腾、广东和诚、云上人和、coremail、云新、福昕鲲鹏、用友软件、壹石新科、中国高科、达烁高科、九思软件、相孚、博云、行云创新、中科汇联、数科网维等
存储设备	国内：华为、宏杉、同有、德拓、中软、百度、川源、浪潮、曙光、宁畅、XSky、SmartX、大道云行 国外：EMC、HDS、NetApp、IBM、HP、Dell
网络设备	Cisco、H3C、华为、迈普、迪普、锐捷、上海同悦等
备份安全	奇安信、信安世纪、三未信安、中孚信息、金城保密、壹进制、鼎甲、云祺、同创永益、中科热备、得安、科力锐、上海英方、上海相孚、精容数安、腾凌科技等

2.4.2. 内核深度自研

虚拟化产品最为核心的部分是 HostOS（Hypervisor），是否拥有内核的研发能力决定了产品的兼容性、稳定性、安全性以及技术兜底能力。如果 HostOS 使用第三方操作系统，虚拟化产品在持续迭代、性能调优、上下游适配的兼容性等需要严重依赖外部支持，自主可控能力受限，例如推出新的硬件（芯片、网卡等），没有内核源码很难快速适配。某些友商的产品虚拟化研发能力不足，HostOS 不得不借助 UOS、Kylin 等第三方[通用]操作系统（缺乏专注在虚拟化的积累），难以形成一体化的技术完整性验证，最终责任推诿、故障处理缓慢、技术断层的风险转嫁给了用户。

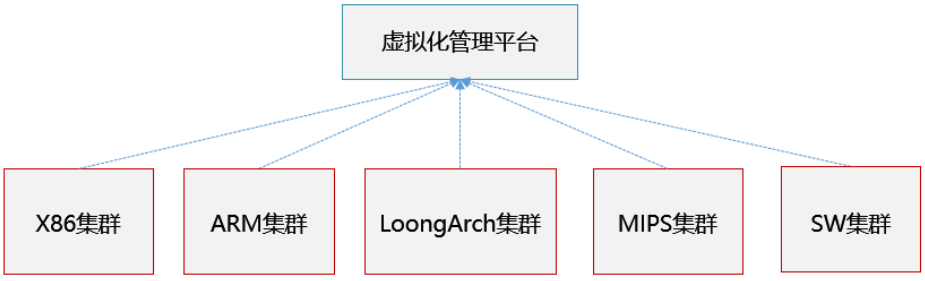
与此截然不同的是，CNware WinSphere 是云宏在虚拟化领域浸淫十年的技术积累，人才团队来自 SUSE、华为、IBM 等知名企业，专注于虚拟化内核在性能、兼容性、稳定性、可靠性、机密性的深度研究，并与国内多家操作系统厂商及社区保持联合攻关，形成自主研发、完整验证、一体化构建的产品，解决用户在技术可持续性上的忧虑。



虚拟化“硬核”研发能力

2.4.3. 一致管控视图

考虑技术可替代性，多芯片技术架构路线是主流建设趋势。用户环境中往往包含 x86、Arm、MIPS、LoongArch 等不同架构的芯片，为 IT 基础设施运维管理、信创应用适配增加了极高的复杂度。CNware 设计之初即考虑利用统一的模型屏蔽不同芯片的管理差异，实现架构的感知和按需分配，从而避免为每类芯片建立一套运维管理界面或工具，用户无需纠结选择多类芯片是否会带来倍增的运维工作难题，在 CNware 广泛的硬件兼容性和一致的视图下能够提升资源的容量感知和响应效率。



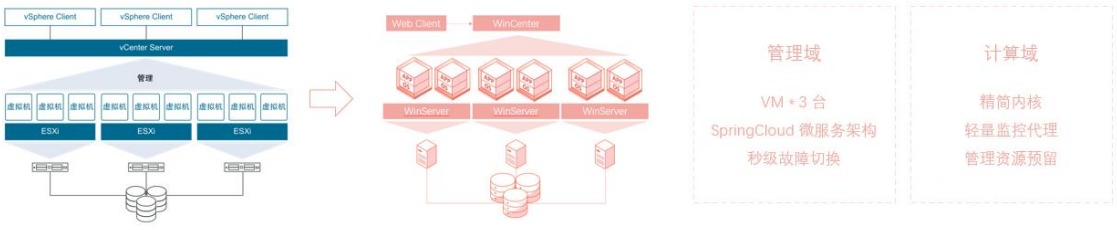
一致管理视图

2.4.4. 极致性能设计

性能永远是衡量数据中心承载能力的主旋律，CNware 从平台架构、软硬件协同的整体

提升角度，提供强大的性能与工作负载优化解决方案。

- 架构极简：产品仅含管理平台、计算节点两层架构，管理组件运行资源占用极低。管理平台支持以双机热备虚拟机方式承载，能够跨物理区域多点部署，灵活适应从大中型到小微型云数据中心的建设需求；计算节点上运行的轻量管理组件仅需 2C4G 的资源，允许精细控制资源预留和分配，这样则保证充足的计算资源提供给业务虚拟机使用，同时保证平台持续处于稳定的运行环境中；
- 内核精简：正如专业的事情应当交给专业的人去做，虚拟化的内核（HostOS）应当专注于虚拟化系统的性能、稳定、安全，无关的进程和组件不应该被加载。因此，CNware 的内核经过操刀可靠的裁剪和进程控制，剔除非必要的库文件、扩展包、开发套件等得到一个最小化的虚拟化系统，使计算性能近乎物理裸机损耗比例极低，同时稳定性、安全性跃升到非常高的层级。
- 软硬件协同：得益于内核的精简及可控，产品能够很好的适应各类服务器固件和外围存储、网络等设备，在硬件加速、IO 通道优化等方面充分发挥硬件的潜能。例如，基于鲲鹏 Boostkit 方案，产品能够感知硬件加速库，相比软件层面的优化提升应用加速效果数倍。



轻量架构，极致性能

2.4.5. 原生安全可靠

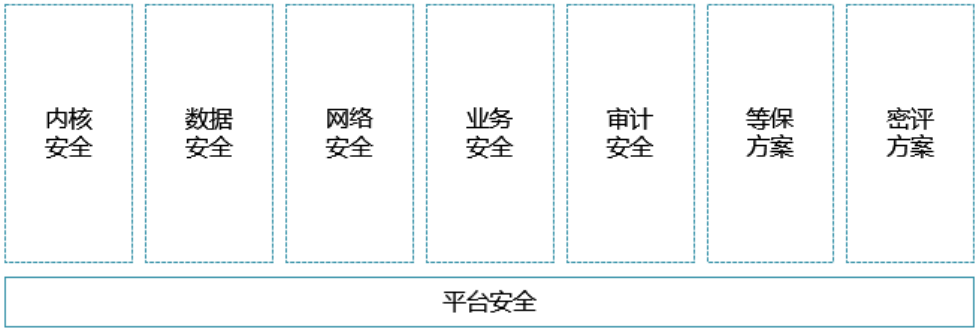
CNware 诞生之初则先用户对底层安全之忧而忧，在研发阶段即考虑云本身需要融合安全防护、稳健连续保护能力，提供全面的原生可靠解决方案。

- 内核安全：通过内核精简，持续监控和保护关键进程并尽最大程度杜绝安全事件；
- 虚拟化系统安全：通过关键文件完整性校验、违规外联检测、端口及漏洞加固、防虚拟机逃逸等技术确保宿主安全；同时，为承载的虚拟机提供 CPU 调度隔离、内存隔离、网络隔离、物理与虚拟环境隔离等控制特性。
- 网络安全：提供安全组、分布式防火墙、防 ARP 欺骗、抑制恶意 DHCP 广播报文、VLAN 隔离、虚拟子网隔离、VPC 隔离、Qos 带宽控制、传输加密、防 DDoS 攻击等技术手段进行管控，建立强大的阻断和防控壁垒；设计轻量 SDN 控制器实现集中式管控、安全下发的流表策略，避免网络黑盒。
- 传输安全：针对虚拟机迁移、平台访问、虚拟机控制台、卷上传下载等传输过程嵌入 SSL 隧道加密保护机制，防止数据窃取；
- 数据安全：提供虚拟机镜像完整性校验、用户数据安全加密、主机防篡改、虚拟磁盘数据擦除、磁盘数据软硬加密支持、容量实时监控等特性。
- 管理安全：提供用户登录与密码策略、三员管理、双因子身份验证等特性，也允许对接第三方的安全平台形成更完善的方案，如无代理杀毒、日志审计系统、加密机、USBKey、数据库审计系统、三合一等软硬件。
- 业务安全：内置杀毒安全防护驱动，同时提供基于 VPC 的东西向租户安全隔离与南北向业务安全防护方案，预留安全接口实现与奇安信、亚信等专业安全平台对

接，实现高级别的风险识别与防护，全面遵从等保要求。

- 密码学安全：原生支持国密算法对敏感信息、关键数据进行加解密防护，预留安全接口对接得安、渔翁、卫士通等密码机、SSL VPN 安全网关、PKI 认证服务、时间戳服务、签名验签服务等综合密码安全解决方案，完全符合信息化密评要求。

信息化安全的环境变化日趋复杂，作为承载关键业务的虚拟化系统只有通过内建式、基于负载和意图、更强的控制、更主动的防御体系，才能将安全威胁影响降至最低。



原生安全体系

2.4.6. 极简管理体验

CNware 为企业数字化基础设施带来极简的体验：

- 介质便携：针对同一芯片体系（例如 ARM），仅仅需要携带一套介质即可具备计算、存储、网络虚拟化的完整部署能力，无需多套介质反复传输拷贝。
- 轻量架构：对标 VMware vSphere 的简洁架构，具有轻量、易管理的明显优势；
- 部署便捷：计算节点、存储节点和 SDN 网络控制器封装一键式的脚本；
- 扩容迅速：借助裸机管理平台实现自动化、任意规模的横向扩展；
- 业务交付简单：支持批量复制、基于策略等部署方式，结合负载均衡器实现超高的弹性支撑能力，轻松应对各类集群、分布式架构应用，满足多版本共存、快速

升级回退的部署要求。

- 运维轻松：不需要维护像 OpenStack 庞大的管理组件（nova、cinder、neutron...）；基于 springcloud 微服务架构开发，得益于微服务清晰健全的调用机制，服务的配置文件和日志简洁明了；所有运维场景均提供页面实时日志及详细步骤反馈，帮助自行定位快速解决问题无需寻求服务商支持。

2.4.7. 裸机入云管理

物理裸机是云数据中心重要的组成部分，因为在部分特殊的业务场景下，非虚拟化架构是客户的选择之一。除了虚拟化云平台服务能力，CNware WinSphere 还提供基于带外的物理裸机管理功能，帮助管理员清晰掌握数据中心所有计算资源的服务状态。

- 资产视图：陈列在机架上的服务器享有云上的孪生数字化信息，管理员轻松知悉其品牌型号、序列号、带外管理 IP、机架位置、负责人、维保信息；
- 电源控制：基于带外远程控制，查看/控制其上下电状态，结合 DRS/DPM 技术实现绿色、低碳、低成本的数据中心；
- 硬件级监控：硬件健康度直接影响业务可靠性，包括硬件温度、硬件功耗、风扇转速、电源功率等实时监控信息；
- 一键带外：一键链接到服务器 BMC 带外管理平台，方便进行操作系统部署、硬件维护等操作。

2.4.8. 企业级管理特性

CNware WinSphere 提供的不仅是虚拟化云平台技术，而是可替代 VMware vSphere 的企业级虚拟化产品。即，CNware 本身融入虚拟化技术且实现面向企业非常关注的功能特性、

增强了高安全高稳定特性，与仅提供虚拟化底层技术而非注重虚拟化场景特性的云产品（例如 OpenStack）截然不同。譬如：

集群层面，CNware WinSphere 的弹性计算功能模块支持集群 HA、虚拟机 HA、计算 DRS、存储 DRS、DPM 动态能耗管理等特性。支持按 CPU、内存、存储利用率比例设置集群动态调度策略，系统自动根据负载均衡各计算节点及业务，保证系统稳定性、资源最佳响应和良好的用户体验。

宿主机层面，提供查看物理 NUMA 拓扑结构和 NUMA 各个节点的 CPU、内存使用信息；支持查看 NUMA 节点中 LCPU 和 VCPU 的绑定关系；支持物理 NUMA、vNUNA；设备方面，可以集中查看宿主机的网卡、硬盘、PCI 设备、USB 设备、CPU、内存等设备，支持 SR-IOV 等特性；性能优化方面，支持配置 IOMMU、配置内存大页规格给虚拟机使用、配置主机关键进程的内存资源预留等；

虚拟机层面，支持创建、启动、关闭、关闭电源、重启、休眠、暂停、恢复、删除、在线迁移、备份、克隆等基本生命周期管理功能，以及虚拟机重置密码，虚拟磁盘还原至镜像状态，查看虚拟机运行日志，虚拟机网卡多队列，虚拟机 MTU 设置，虚拟机回收站，虚拟机与虚拟机亲和性规则，虚拟机与宿主机亲和性规则，删除磁盘时允许擦除数据，防止数据泄密。

2.5.接口协议

CNware WinSphere 对外提供标准、开放的 RESTful 接口，可供应用系统、自动化系统、云管平台集成调用，实现对虚拟化资源和的管理和调度。CNware WinSphere 接口描述请参考配套 API 文档。WinSphere 管理平台自身各组件间的接口和调用关系如下图所示：

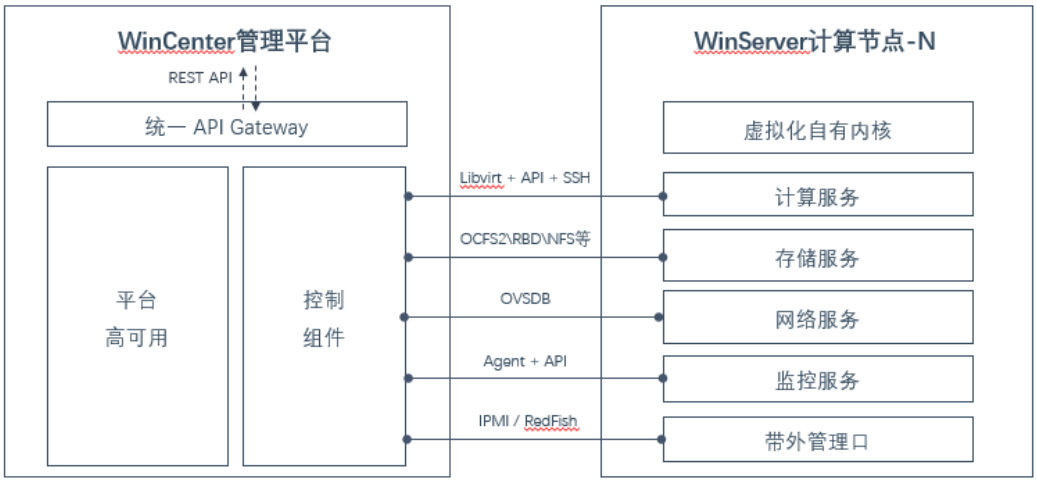


图 2-2 接口协议规范

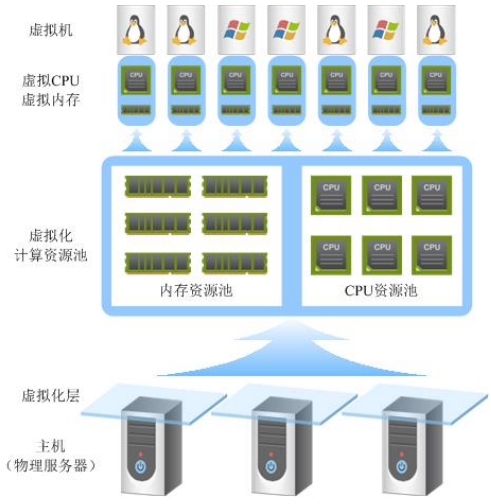
第3章 产品功能

3.1.技术概览

3.1.1. 计算虚拟化介绍

虚拟化技术正以前所未有的脚步通过提高服务器的整合能力对数据中心产生重要的影响。各大企业机构都已经开始削减他们所需要的物理机的数量，并通过这样的方式来达到降低成本、减少能耗和空间的目的需求。

CNware WinSphere 服务器虚拟化是基于 KVM 底层技术实现，将服务器物理资源抽象成逻辑资源，让一台服务器变成几台甚至上百台相互隔离的虚拟服务器，不再受限于物理上的界限，而是让 CPU、内存、磁盘、I/O 等硬件变成可以动态管理的“资源池”，从而提高资源的利用率，简化系统管理。



虚拟资源池化技术

裸金属架构

WinServer 使用裸金属架构，架设在计算机硬件和操作系统之间的虚拟化，直接在硬件上安装虚拟化软件，将硬件资源虚拟化，通过裸金属架构的虚拟化，计算机硬件可以直接被切割成若干的虚拟机，然后再这些虚拟机上面再进行各自的系统和应用程序的安装。由于使用了裸金属架构， WinServer 可为用户带来接近服务器性能、可靠性和可扩展性的虚拟机。

CPU 虚拟化

WinServer 将物理服务器的 CPU 虚拟成虚拟 CPU（VCPU），供虚拟机运行时使用。为了获取接近物理 CPU 的计算能力，可以将性能有敏感要求的业务虚拟机的 vCPU 直接绑定到物理 CPU 核（或超线程）上；

内存虚拟化

WinServer 以进程的方式管理虚拟机，在 host 上通过正常的内存分配进行内存的分配与使用；通过 qemu 规划虚拟机的内存地址空间。虚拟机集中访问内存时通过其内部操作系统页表访问虚拟物理地址 GPA，在 hypervisor 通过页表转换技术将 GPA 转换为主机虚拟地址 HVA，并进一步通过主机的页表技术转换为主机物理地址 HPA，从而完成内存的访问。目前内存地址转换技术得益于硬件辅助虚拟化的发展，通常采用 CPU 厂商的扩展页表技术，将虚拟物理地址直接转换为主机物理地址。WinServer 支持 DMC（动态内存控制），自动调整正在运行的虚拟机内存，支持虚拟机总配置内存超过物理主机实际运行内存，从而使每个服务器具有更大的虚拟机密度，降低企业成本。

GPU 虚拟化

基于 GPU 生产商的硬件虚拟化技术，将物理 GPU 划分为多个虚拟 vGPU 单元，并基于 SR-IOV 技术将虚拟 vGPU 单元通过主机设备的方式分配给虚拟机进行使用。

NUMA 架构感知技术

利用对物理机内存与 cpu 的 NUMA 架构的识别，将虚拟机进行内存与 cpu 的资源划分的方式将虚拟机分配到特定的 NUMA 节点之上；并根据虚拟机的配置，将配置的 NUMA 信息通过虚拟硬件信息的方式传递给虚拟机操作系统，让虚拟机操作系统能够识别到虚拟出来的 NUMA 架构。通常虚拟机的 vNUMA 的配置都遵循实际的物理 NUMA 架构完成。

3.1.2. 业务连续性保护

基于虚拟化架构，能够带来先进的业务连续性解决方案：

虚拟机在线迁移

在业务不中断的前提下，可自动或手动的把在线的虚拟机迁移到其它物理主机，满足计划内维护、均衡负载等需求场景。

存储在线迁移

WinServer 支持把在线的虚拟机的存储进行迁移，不限制本地和共享存储，可以虚拟机和存储同时迁移，以满足用户把虚拟机从开发环境迁移到生产环境，在基础设施维护和升级时可以不中断业务。

主机故障保护

当主机发生计划外故障的时候，WinServer 支持自动的把故障主机上的虚拟机迁移至其它可用的主机上，实现资源池内物理主机的高可用。实现主机故障保护需配置集群 HA。

虚拟机故障保护

当虚拟机发生计划外故障的时候（如关键进程崩溃），CNware WinSphere 通过持续保护

检测机制获取不到虚拟机正常运行的状态信息，将会迅速响应重启该虚拟机。实现虚拟机故障保护同样需要配置集群 HA。

网络冗余

WinServer 支持通过物理网卡绑定冗余，来提高网络可靠性和吞吐量。

存储多路径

存储多路径是指存储设备通过一条或多条链路与主机连接，通过存储设备的控制器控制数据流的路径，实现数据流的负荷分担，保证存储设备与主机连接的可靠性。WinServer 支持存储多路径配置，通过容错、I/O 流量负载均衡甚至更细粒度的 I/O 调度策略调校，实现更高的可用性和性能。

虚拟机快照和恢复

虚拟机快照可保存虚拟机配置和虚拟机磁盘的数据，用于虚拟机数据的还原和恢复；如果是内存快照，还可以保持和恢复虚拟机内存状况。

WinSphere 支持磁盘快照和内存快照，当虚拟机出现故障或系统奔溃时可通过恢复快照回到之前的可用点。

3.2.计算虚拟化

3. 2. 1.　集群管理

利用组建虚拟化集群，操作员可以像管理单个实体一样轻松地管理多个宿主机和虚拟机，工作负载能够在集群内实现弹性的算力汲取和移动，同时借助集群的高级特性能够持续保护业务的连续性。

例如，系统定时对集群内的主机和虚拟机状态进行监测，保证了数据中心业务的连续性。例如，当一台服务器主机出现故障时，运行于这台主机上的所有虚拟机都可以在集群中的其它主机上重新启动。此外还包括其他的高级特性，例如动态资源负载调度、电源能耗管理、亲和性规则等，提升弹性、高效、智能的资源响应能力。

3.2.1.1. DRS 动态资源调度

在虚拟化环境中，每个计算节点均可能承载各类形色的业务类型，且业务负载占用的资源可能瞬息万变，因而虚拟机的计算资源往往极为容易成为瓶颈，超出计算节点能够提供服务的算力资源，如此反映业务运行缓慢、效率低下、稳定性难以保障等现象。虚拟化平台提供的 DRS 动态资源调度特性由此应运而生。

DRS 动态资源调度是集群的高级特性，在虚拟机部署和运行过程引入自动化优化的机制，动态地自动优化和平衡资源池内负载的分配，不断侦测判断集群内的承载情况，确保虚拟机及时被调度到冗余资源的计算节点；同时，结合用户自定义的亲和性规则，保障关键业务能够持续被正确的调度，保障关键业务持续获得充沛的算力和处于长期稳定健康的运行环境。

CNware DRS 特性的实现原理主要是通过下发策略至各个计算节点运行的守护进程，守护进程持续侦测节点的 CPU、内存、存储的使用率，若满足策略的条件则立即向上报告和执行计算调度。

计算 DRS

计算 DRS 策略配置包括衡量依据、条件判断、触发规则。

- 衡量依据：CPU、内存的利用率，需配置阈值；

- 条件判断：CPU 或内存其一或两者同时满足阈值要求；
- 触发规则：持续时间、间隔时间，两者组合形成敏感度。

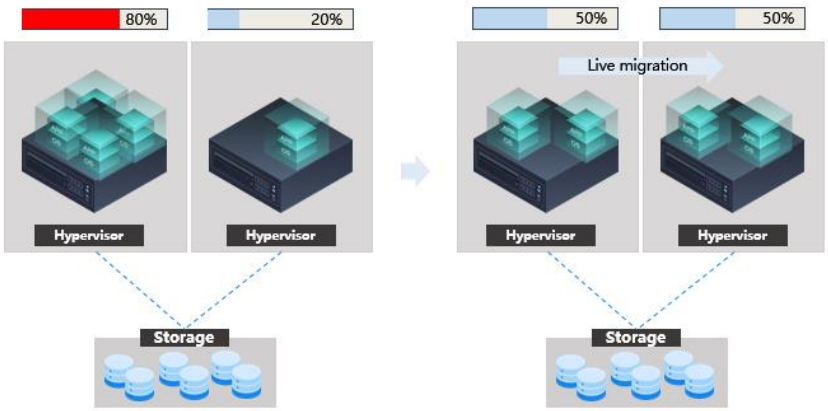
当集群中的某台宿主机的计算资源利用率达到策略触发条件，则其上的虚拟机自动按预定的策略调度分配到其他“符合预期”的宿主机上运行，防止过度负载导致不稳定的风险；

存储 DRS

存储 DRS 策略配置包括衡量依据、触发规则。

- 衡量依据：存储资源利用率，需配置阈值；
- 触发规则：持续时间、间隔时间，两者组合形成敏感度。

当集群内的某个存储池利用率达到触发条件，则存储池内的虚拟磁盘按预定的策略调度分配到其他“符合预期”的存储池，防止存储池过度使用导致容量性能不足。



集群DRS原理

3.2.1.2. DPM 电源智能管理

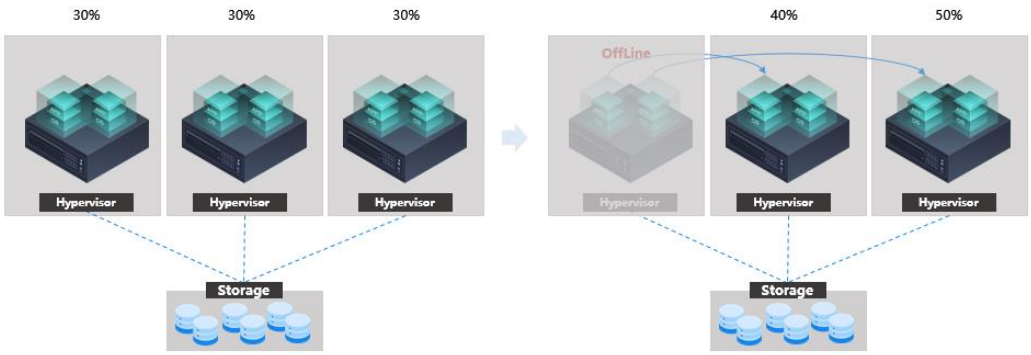
DPM 动态电源管理（或智能能耗管理）是云数据中心建设实现绿色节能、降本增效的一项特性，非繁忙时刻能够使部分主机退出服务并优化工作负载的位置，实现最大功率节省

数据中心的电力；一旦业务再次繁忙起来，主机又能够立即再次上线投入服务。最终平衡业务投入与成本控制的关系，通常与 DRS 动态资源调度搭配使用。

DPM 策略配置包括衡量依据、触发条件。

- 衡量依据：CPU、内存的最低和最高利用率阈值；
- 触发条件：持续时间、间隔时间，两者组合形成敏感度。

系统持续巡检集群整体负载状态，达到低负载触发条件后则将低负载的计算节点智能回收并置于睡眠状态以减少电力消耗，达到高负载触发条件时则智能唤醒处于睡眠状态的计算节点并快速恢复提供服务。



集群DPM原理

3.2.1.3. 亲和性调度规则

由于某些业务的特殊性或特定要求，例如性能、可用性等方面的考量，有时需要对特定业务虚拟机的运行位置作出干预和影响。

举个例子，构成数据库集群的两个主备节点不应该同时运行于同一宿主机，又或者高并发处理的业务虚拟机必须位于指定的高性能宿主机。因此，包括虚拟机与虚拟机之间、虚拟机与主机之间实际上是需要亲和性和反亲和性的规则来约束的，且必须由系统来完成智能的

调度。

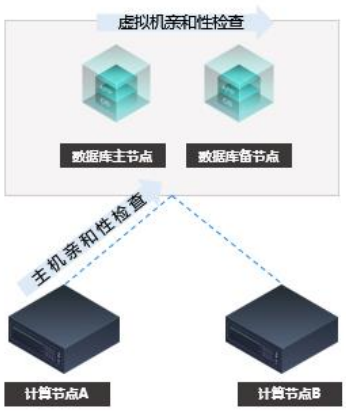
虚拟机与虚拟机亲和性规则

虚拟机与虚拟机亲和规则，定义了一组虚拟机集群并设置聚集或互斥规则，约束指定的虚拟机集群必须一致运行于同一宿主机或不能在同一宿主机上；

虚拟机与宿主机亲和性规则

虚拟机与主机亲和规则，定义了一组虚拟机集群并与指定的宿主机关联，即约束同一虚拟机集群必须或尽可能运行于指定宿主机或不能运行于指定宿主机。约束规则包括：

- 禁止在主机上运行
- 必须在主机上运行
- 不应在主机上运行
- 应该在主机上运行



亲和性规则检查

3.2.2. 主机管理

主机即虚拟化主机、物理主机或者计算节点，相对于虚拟机而言，是指实体的、部署虚

拟化的物理服务器。宿主机的作用是给虚拟机提供模拟的、隔离的硬件环境，达到同一物理服务器可以安装多个操作系统的目的，并且多个操作系统间还可以互相通信，体验如同运行在真实的物理机硬件。

3.2.2.1. 主机生命周期管理

主机管理

管理平台通过 SSH 协议安全纳管计算节点，可以按 CPU 架构类型加入对应集群。

迁移主机

根据业务的动态调整，基础架构也不是一成不变的，这就需要对计算节点作平滑的规划调整。计算节点能够在主机池内改变其位置，例如集群外的主机迁移到集群内、集群内的主机迁移到其他集群或者集群外。

主机概要

主机概要反映了主机整体的状态及性能，包括基础信息、配置信息、策略配置。

- 基础信息：包括主机的名称、IP、IQN、IOMMU 状态、BMC 链接、维护状态、连接状态、NUMA 节点及拓扑、CPU 信息（插槽/内核/线程/型号/主频）、内存容量、服务器型号、序列号、虚拟化版本及补丁记录、虚拟机数量及运行分布、服务器时间、运行时间、备注等信息。
- 配置信息：包括 CPU 总频率和利用率、内存总量和分配量及利用率、存储总量和分配量及利用率；
- 策略配置：主机是否为启用 HA 状态。

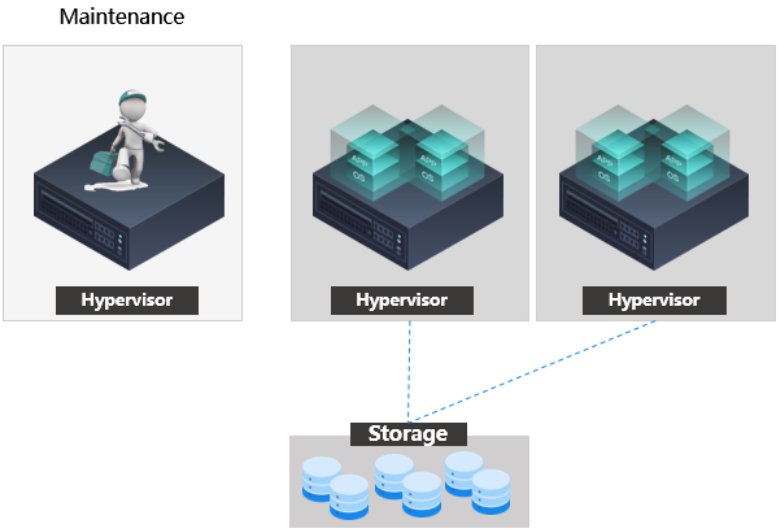
- 资源分配比：主机的 CPU、内存等资源的容量及分配比率。

3.2.2.2. TopN 利用率分析

由于每个计算节点承载的业务类型包含各类形色，因此业务工作负载的资源消耗情况总是参差不齐的。虚拟化系统应当持续监控宿主机上的虚拟机运行情况，按 CPU、内存、磁盘使用率等维度自动计算、推举、降序排列利用率 TopN 的实例，快速发现业务的性能瓶颈，帮助用户建立对单个业务的事前预防和分析管理。

3.2.2.3. 维护模式管理

维护模式是执行计划内维护动作的一项功能。例如，管理员期望对宿主机做硬件维护、设备升级、存储脱机等操作，避免业务调度到该节点上受到影响时，可将节点置于维护模式；处于维护模式的宿主机，不允许在其上部署新的虚拟机，同时所有虚拟机要求必须关闭电源或迁移到其他节点，其他节点的虚拟机也无法调度到该节点，确保所有的运行态和存储态数据撤出；当宿主机恢复正常可稳定提供计算服务时，才可将主机退出维护模式。



维护模式

3.2.2.4. NUMA 拓扑感知

NUMA（Non-Uniform Memory Access，非一致性内存访问）是一种关于多个 CPU 如何访问物理内存的架构模型，现代服务器大部分支持 NUMA 架构。在 NUMA 出现之前，所有 CPU 对内存的访问基于共享的总线模型（SMP）保证访问的一致性，即每个处理器核心共享相同的内存地址空间；然后随着 CPU 朝高频率的方向发展遇到了天花板，转而向着多核心的方向发展，这样的架构难免遇到问题，比如对总线的带宽带来挑战、访问同一块内存的冲突和瓶颈。因此，NUMA 出现了：在 NUMA 架构下，CPU 被分成了多个节点 Node。每个节点有自己的内存 Controller，不再受内存总线带宽的限制。每个 NUMA 节点上面有自己的内存控制器、有自己的内存。

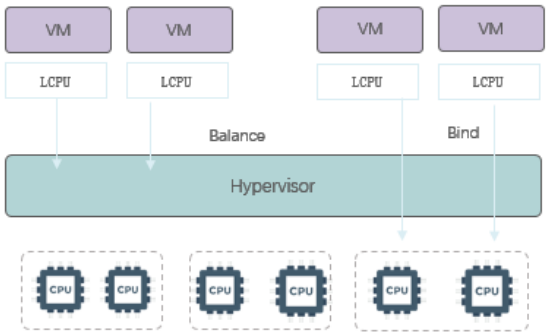
NUMA 架构使应用程序能够得到更好的优化，由于该架构将内存划分为相对于 CPU 核心的本地和远程内存，因此围绕 NUMA 的性能优化，核心的目标就是使处理计算的部分（特别是需要频繁访问内存的应用）访问最短距离的内存，同时提升并发处理的能力。

NUMA 架构的优势天然吻合虚拟化领域的需求，两者实际上都在做计算系统的调度优化。每个虚拟机在虚拟化层看来实际上属于一个 qemu 进程，占用的 vCPU 也会实际抢占 pCPU 的时间片资源，令人遗憾的是虚拟化层对这种调度是随机的（尽管有一定的调度算法，但总体对虚拟机应用的影响是消极的），除了频繁的切换的同时也不符合 NUMA 架构的本地最短距离访问优化目标，对内存敏感的应用来说就是性能是不可靠的。因此，NUMA 架构的感知和绑定技术，能够将虚拟机所在的 Qemu 进程进行 pCPU/vCPU 绑定，减少 CPU 时间片资源的争抢挤占、虚拟化层调度开销，内存同样绑定在对应的本地物理内存，大幅度提升虚拟机性能和稳定性。

CNware 产品设计了直观的图形化界面，一键获取服务器的物理 NUMA 拓扑结构，能够

帮助管理员优先分配资源，从而实现对虚拟机计算性能进行调优。包括：

- NUMA node 分布拓扑结构图
- NUMA node 的物理 CPU 核心分布
- NUMA node 内存总量及剩余可分配容量
- NUMA node 的 pCPU 核与 vCPU 的绑定关系



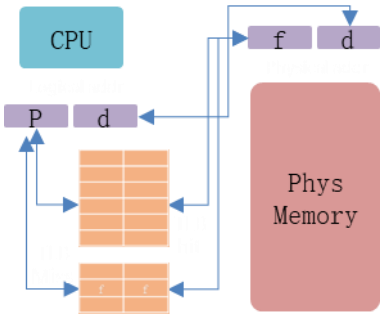
NUMA 感知与绑定

3.2.2.5. 大页内存管理

大页内存(HugePage)也是一项性能优化技术。在虚拟内存管理中，内核维护一个将虚拟内存地址映射到物理地址的表，对于每个页面操作，内核都需要加载相关的映射；如果你的内存页很小，那么你需要加载的页就会很多，导致内核会加载更多的映射表，而这会降低性能。使用“大内存页”，意味着所需要的页变少了，从而大大减少由内核加载的映射表的数量，这提高了内核级别的性能最终有利于应用程序的性能。简而言之，最终的目的是减少页表映射数目的处理，提供 TLB 缓存命中率，从而提高系统整体的访问性能。

虚拟化系统使用大页内存特性辅助调优，。虚拟机配置使用内存大页，需要操作系统内核支持，对某些应用来说能够达到较大幅度的性能优化提升。针对部分应用（例如 Oracle AMM）可能不支持大页内存，请谨慎评估采用。

支持配置大页内存规格（如 2M、32M、512M、1G）、大页数量，系统将自动计算启用大页内存的总容量。按虚拟化场景最佳实践，建议至少预留总内存的 15%给主机使用。



内存大页性能提升原理

3.2.2.6. 资源超分策略

物理资源的超分能带来一些资源充分利用方面的好处，比如在一台性能强大的硬件服务器正作为虚拟客户机运行着不同业务类型的服务，但是它们并非在同一时刻都会负载很高，有的业务服务器在白天工作时间负载较重，而有的业务服务器则主要在晚上工作，所以如果对物理资源进行合理的超分超配，给这几个业务服务器分配的系统资源总数大于实际拥有的物理资源，就可能在白天和夜晚都充分利用物理硬件资源。

系统支持从界面配置主机的 CPU、内存、存储超分策略，让用户可以根据主机承载的虚拟机资源使用情况，量身定制化超分比例。

CPU 的超分

用户可以根据业务的负载，将 CPU 按照 1:1、1:2、1:4 等不同的超配比进行超分。比如之前物理 CPU 核心是 20 核，开启超线程为 40 核；如果按照 1:2 进行超分之后，可分配的资源就变成 80 核，这样就可以创建更多的虚机资源出来。

内存的超分

由于虚拟机操作系统及其运行的应用程序并非一直 100%地使用其分配到的内存，而且宿主机上的多个虚拟机一般也不会同时达到 100%的内存使用率，所以内存也可以进行超分。内存的超分配比跟 CPU 类似，可以根据 1:1、1:2 等不同的配比超配，比如物理内存是 8G，超分 1:2 则可以给虚机分配的资源为 16G。一般为了保障业务性能，内存超配比较低。

存储的超配

系统在创建虚拟磁盘时并不占用物理存储的空间，只有当虚拟机写入数据时,再根据写入数据量动态分配物理存储空间。因此虚拟机分配的磁盘大小总和不受物理存储总容量的限制，即虚拟机磁盘容量可以超配。

因为虚拟机运行过程中会持续写入数据，当虚拟机动态分配的存储空间接近物理存储的实际容量时，虚拟机无法继续分配到存储空间从而导致运行异常。所以配置虚拟机的磁盘大小总和不建议超配太多，比如超配比为 1:1.2。为了保障整个存储集群的稳定性可靠性，原则上要求实际存储量一般不能超过集群总容量的 80%-90%，如果超过这个阈值就要进行存储集群的扩容了。

3.2.3. 虚拟机管理

虚拟机与主机类似，它们主要的区别在于虚拟机并不是由电子元器件组成的，而是由一组文件构成的。每台虚拟机都是一个完整的系统，它具有 CPU、内存、网络设备、存储设备和 BIOS，因此操作系统和应用程序在虚拟机中的运行方式与它们在普通物理机上的运行方式没有任何区别。

3.2.3.1. 虚拟机生命周期管理

虚拟机基本的生命周期管理能力，包括创建、启动、关闭、关闭电源、暂停、休眠、恢复、重启、删除等。

虚拟机概要

虚拟机概要反映了虚拟机整体的状态及性能，包括基础信息、配置信息、策略配置、安全配置。

- 基础信息：包括虚拟机名称、所属宿主机、系统类型、系统版本、状态、Hostname、UUID、IP 地址、VNC 端口、SPICE 端口、时钟源、创建时间、运行时间、tools 版本、关联镜像等；
- 配置信息：包括 CPU 分配情况（分配值、最大值、利用率）、内存分配情况（容量、预留百分比、利用率）、存储分配情况（虚拟机硬盘设备名、容量、类型、格式、置备类型、存储池及路径）、网卡分配情况（所属虚拟交换机、网络策略、状态、MAC 地址、IP 地址）。
- 策略配置：是否启用高可用、是否启用自动迁移；
- 安全配置：是否为安全状态、是否加密、是否安装杀毒组件。

创建虚拟机

即创建虚拟裸机（空白虚拟机），再完成系统从无到有的安装灌入。虚拟裸机包含虚拟机运行所需的完整虚拟硬件设备，包括 CPU、内存、磁盘、虚拟网卡、光驱等，不包含操作系统和磁盘数据；特别的，虚拟裸机创建之初即要求指定与实际部署的 OS 一致或尽可能相近的系统类型，这是由于虚拟机配置文件需要针对不同的 OS 作兼容和优化，以更好辅助 OS

的装载和运行。

虚拟裸机装载操作系统的两种方式：

- 指定已有的、包含操作系统的虚拟磁盘作为系统盘；
- 通过 ISO 系统镜像引导，向不含任何数据的虚拟磁盘完成操作系统全新安装。

部署虚拟机

即利用虚拟机镜像（模板）生成新的虚拟机。对于管理者来说，以镜像为基础的虚拟化环境能够大幅度提升系统的连贯性，虚拟机的部署速度和规模能够以一种前所未有的速度不断扩展，高效、快速、准确成为标准的部署体验；更为重要的是，镜像能够保持虚拟机和应用版本的一致性。

通过镜像批量部署生成虚拟机，支持两种方式：通过表单式引导、Excel 模板导入。前者表单式引导即通过镜像管理界面，选择单一镜像并根据逐步的引导完成表单填充，该方式适用于一次性生成相同规格的虚拟机；后者 Excel 模板导入允许管理员导出模板并按一定规则完成填充后，上传至平台自动解析，该方式适用于多次重复、生成不同规格的虚拟机。

启动虚拟机

为虚拟机执行加电操作，引导虚拟机内的操作系统；

修改虚拟机配置

在线或离线状态修改虚拟机的概要信息（名称、Hostname、时钟源、IO 优先级、是否自动迁移）、虚拟硬件（CPU、内存、磁盘、控制台、光驱、网卡、显卡、鼠标、键盘、写字板）、设备引导顺序、vNUMA 架构等。

重启虚拟机

向虚拟机操作系统发送重启指令，操作系统收到信号并进行正常重新引导；

强制重启虚拟机

向 Hypervisor 发送强制重启指令，Hypervisor 立即对虚拟机进行重新引导，而不考虑 GuestOS 正在执行的工作。

关闭虚拟机

向虚拟机操作系统发送关机指令，操作系统收到信号并进行正常关机；

关闭虚拟机电源

向 Hypervisor 发送强制关机指令，Hypervisor 立即为虚拟机执行下电操作，而不考虑 GuestOS 正在执行的工作。

暂停虚拟机

暂停虚拟机当前所有进行的 IO 活动。在某些场景下，如果虚拟机持续处理密集型任务导致节点性能飞速下降，无法再执行优先级更高的管理任务，考虑按下暂停键。

休眠虚拟机

保存虚拟机当前的状态并进入休眠状态，直到恢复时可以快速进入系统，包括已启动的程序也处于休眠前的状态。

恢复虚拟机

把虚拟机从休眠状态恢复为运行状态，且在休眠之前运行的应用程序将恢复运行状态，而不更改其内容；

删除虚拟机

将虚拟机从管理平台中删除。包含几种可选的删除方式：

a. 只删除虚拟机，保留存储卷数据，虚拟机移入回收站；虚拟机可以从回收站中还原，并恢复与虚拟磁盘的关系；

b. 删除虚拟机的同时，不保留存储卷数据，虚拟机和虚拟磁盘同时移入回收站；虚拟机和虚拟磁盘可以从回收站中还原；

c. 移除虚拟机，虚拟机从管理平台移出，但底层不执行任何操作，虚拟机仍然保持原有状态；同步主机数据时，虚拟机再次恢复到管理平台；

d. 销毁虚拟机，虚拟机不经过回收站即彻底删除所有数据。请谨慎配置，云宏不承担任何由策略引发的数据损失责任。

批量管理虚拟机

集群或单计算节点往往承载数十到数百个虚拟机，如果仅能作用单个虚拟机的管理，可想而知操作工作量是非常恐怖的线性增长水平。因而，常用的启动、关闭、关闭电源、迁移、重启、恢复、暂停、休眠、删除、导出、同步数据等操作支持批量执行，极大地释放运维人员的精力。

3.2.3.2. 虚拟机 Tools

虚拟机 tools 是 CNware 提供的虚拟化增强工具，包含了性能优化、驱动适配、企业级虚拟化功能扩展组件，能够提高虚拟机的 I/O 处理性能，实现可用性侦测和改善其他高级管理功能。虚拟化管理平台提供 tools 一键挂载的便捷设计，并就如何为虚拟机安装 tools 自动弹出提示。

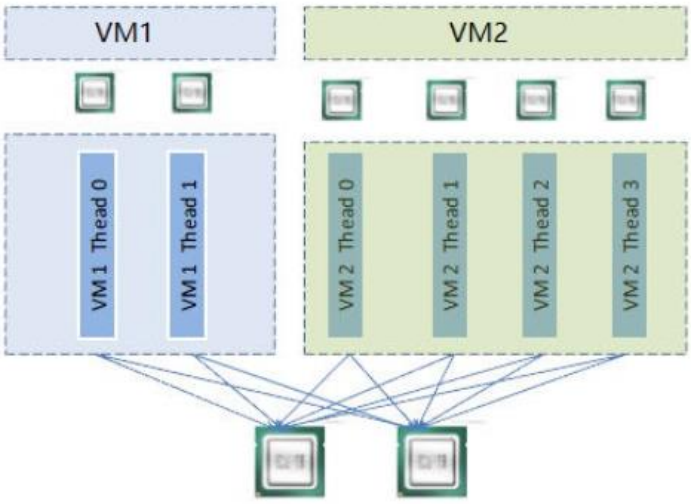
CNware Tools 进程经过全面的安全检测和控制，性能资源占用几乎可以忽略不计。兼容

性方面，已支持主流的 linux/windows 操作系统（包括最新的 Kylin、UOS 等发布版本第一时间获得支持）及多架构支持。为了支持完整的虚拟化功能集，强烈建议为虚拟机安装 tools。

3.2.3.3. vCPU 管理

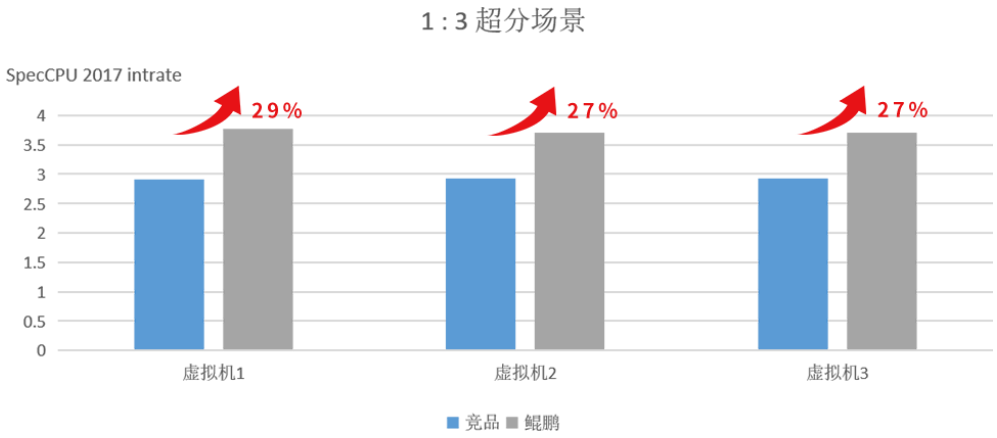
虚拟机不直接感知物理 CPU，而是由 Hypervisor 抽象出相对物理 CPU 而言的 vCPU 作为虚拟机的计算单元。

从虚拟化结构分析，虚拟化层提供两级的调度框架，两级调度的调度策略和机制不存在依赖关系。虚拟机的操作系统负责第 2 级的调度，虚拟机内的进程和线程在 vCPU 上调度；Hypervisor 负责第 1 级的调度，即 vCPU 在物理 CPU 的分配和调度。虚拟机仅能看到虚拟化层呈现的 vCPU，每个 vCPU 对应一个 VMCS（Virtual-Machine Control Structure）结构，当 vCPU 被从物理 CPU 上切换下来的时候，其运行上下文会被保存在其对应的 VMCS 结构中；当 vCPU 被切换到 pCPU 上运行时，其运行上下文会从对应的 VMCS 结构中导入到物理 CPU 上。vCPU 可以调度在一个或多个物理处理单元执行（分时复用或空间复用物理处理单元），也可以与物理处理单元建立一对一固定的映射关系（限制访问指定的物理处理单元）。



CPU 虚拟化技术

虚拟机的 CPU 管理实际是 vCPU 的分配和调度，那么每个计算节点实际提供的 vCPU 数目是多少呢（不考虑超分）？业界有相对认可的公式：系统可用的 vCPU 总数(逻辑处理器) = Socket 数（CPU 个数）x Core 数（内核）x Thread 数（超线程）。需要注意，超线程是 Intel 的 SMT 技术，在 ARM 等架构可能不支持该技术。尽管如此，每个计算节点可以创建和运行的虚拟机的 vCPU 总数可以远远大于实际可提供的 VCPU 数目。



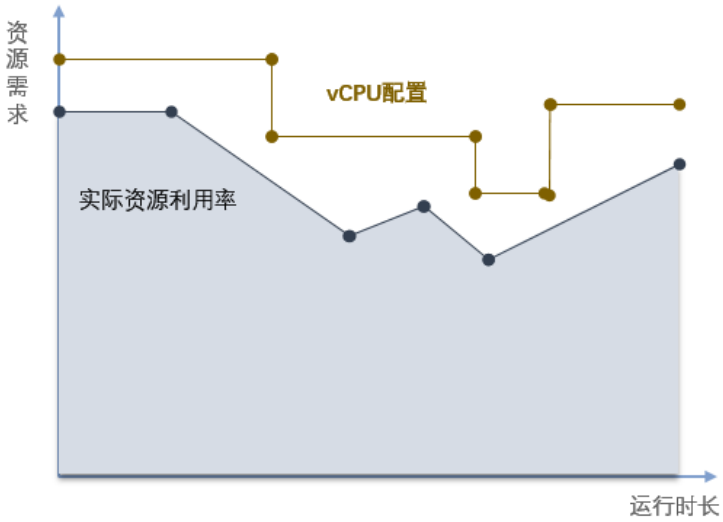
Kunpeng Validated 认证-X86&kunpeng 超分性能对比

vCPU 扩缩容

vCPU 提供虚拟机核心的计算能力，最佳实践是使性能尽可能贴合工作负载的消耗变化范围。虚拟机 vCPU 并不是越多性能就越好，因为线程切换会耗费大量的时间，应当尽可能贴合业务的消耗并设置最小值。

无论初始规划得如何完美，工作负载的变化总有可能与预期有出入。因此，当计算资源分配过于闲余时，虚拟机应当减少 vCPU 数目；当计算资源分配不足以支撑负载时，虚拟机应当增加 vCPU 数目。

扩缩容调整选择在较为稳妥的时间窗口进行，并明确可能造成的业务影响，一般建议虚拟机处于离线状态。特别的，某些业务苛刻要求支持在线完成；在线调整需要满足一定的前提条件，包括物理 CPU 的芯片是否支持、虚拟机操作系统内核是否支持热添加特性等。CNware 产品领先业界实现同时支持在 X86、ARM 架构下的在线调整特性。

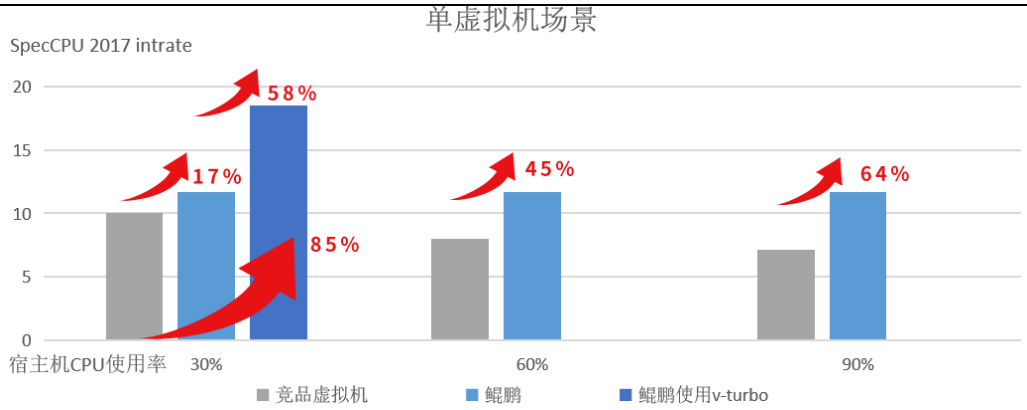


vcpu 扩缩容

vCPU 拓扑

对物理 CPU 来说，CPU 有 socket、core、thread 的概念，形成固定的物理拓扑。Socket 即主板上 CPU 的插槽数量；Core 即每个 CPU 的核心；Thread 即每个 core 的硬件线程数。在操作系统中，可以通过 lscpu 等命令查询该拓扑。

虚拟机的 vCPU 能够模拟真实的物理 CPU 拓扑，并向虚拟机的操作系统呈现；虚拟机 vCPU 拓扑与 vNUMA 特性搭配使用，能够帮助改善特定的应用程序性能。kunpeng boostkit 的 V-Turbo 特性本质采用了 vCPU 拓扑的优化，使虚拟机操作系统内呈现的 2core 实际上映射了 4 个 pCPU thread，测试效果在鲲鹏 920 服务器较使用该特性前性能提升 58%，相比 X86 服务器提升超 85%。



Kunpeng Validated 认证-V Turbo 效果

vCPU 架构

物理 CPU 通用寄存器的位宽有 32/64bit 的区别。32 位处理器可以一次性处理 4 个字节的数据量，64 位处理器可以一次性处理 8 个字节的数据量，比 32 位处理器的处理速率加快一倍。这种差异对应用程序也有比较大的影响，两者的平台、运行要求、内存要求等均有较大差异。虚拟化也能模拟 32bit/64bit 的 CPU 架构，满足不同应用程序的兼容性要求。

在 ARM 平台中，默认架构为 aarch64.

vCPU 工作模式

vCPU 的工作模式包括兼容（custom）、主机匹配（host-model）、直通(host-passthrough)三种模式。

- 兼容模式 custom：虚拟机 CPU 指令集最少，性能最差，但跨不同型号的处理器热迁移时兼容能力最强；
- 主机匹配模式 host-model：系统根据当前宿主机的 CPU 指令集从配置文件 /usr/share/libvirt/cpu_map.xml 中选择一种最相配的 CPU 型号。在这种模式下，虚拟机的指令集往往比宿主机少，性能一般，但热迁移时允许源目 CPU 指令集存在

部分差异；

- 直通模式 host-passthrough：系统直接将物理 CPU 特性透传暴露给虚拟机使用，虚拟机能够最大限度使用指令集故而性能最好，但热迁移时要求源目 CPU 指令集一致；CNware 默认采用这种模式，在实际建设和性能上达到最优。

由此结论

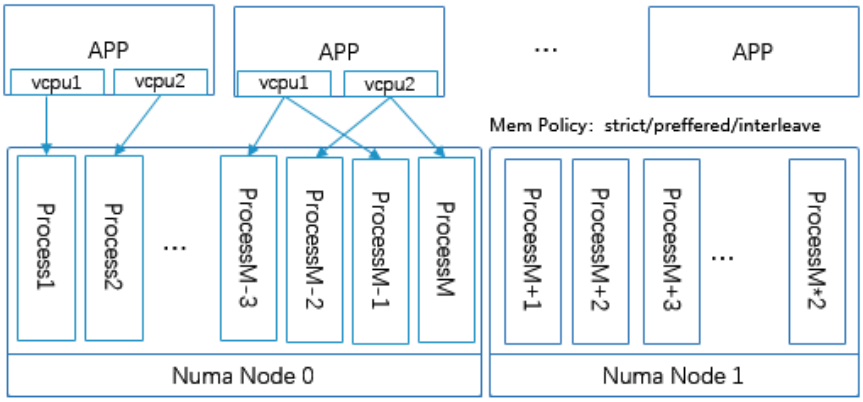
- 从性能排序的角度：host-passthrough>host-model>custom
- 从热迁移通用性的角度：custom>host-model>host-passthrough

不同的模式适用场景不一样，例如兼容模式需要兼顾应用编译，难以适配丰富的硬件；直通模式适合一些需要完全的物理 CPU 特性来支持的应用，但这种模式的兼容性较差，难以迁移到不同型号的 CPU，应当依据实际合理配置。

vCPU 绑定

vCPU 绑定也称为 vCPU 亲和性、vCPU 关联性，其目的是使进程要在指定的处理器上尽量长时间地运行而不被迁移到其他处理器。

不同的 vCPU 实际上在 Hypervisor 看来是不同的线程，而不同的线程是跑在不同的 pCPU 上的；在多核的机器中，每个 CPU 本身自己会有缓存，缓存着进程使用的信息，频繁的切换调度会降低 CPU cache 命中率。因此，将 vcpu 亲和绑定在特定物理 CPU 上，VCPU 只在绑定的物理 CPU 上调度，达到提高虚拟机中的 CPU 执行效率、实现 CPU 资源独享的隔离性、减少虚拟化层调度开销、提升整体性能的目的。



vcpu 绑定

vCPU Qos

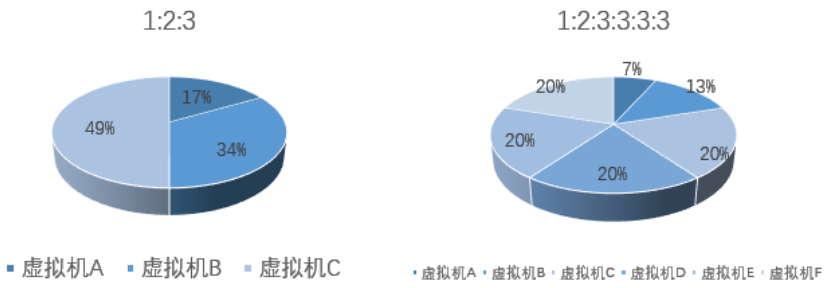
vCPU Qos 的实现原理是，根据分时复用的调度原理给 vCPU 分配运行时间片，消耗完时间片的 CPU 将被限制且直到重新获得时间片。其价值在于为应用提供计算服务质量的保障，确保针对资源的分配是确定的、可衡量的。

Qos 模式包括：

- 限额：定义了分配 CPU 资源的上限，通过计算资源竞争限制保护其他业务；
- 预留：定义了分配 CPU 资源的下限，即最低资源保障，保证服务质量体验；
- 份额：定义了多虚拟机竞抢 CPU 资源时按比例分配，划分不同优先级。

vCPU 优先级

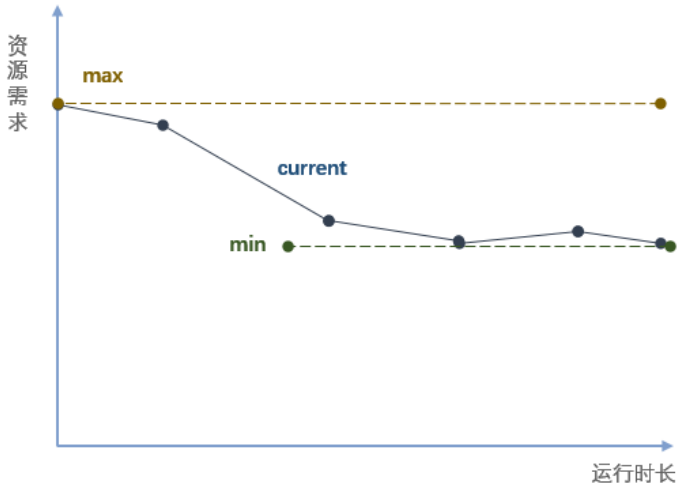
即 vCPU Qos 份额，为了保证企业关键应用的正常运行，适当划分资源份额（即优先等级）实现合理竞争。CNware 提供高、中、低三个挡位，映射为调整虚拟机进程获取 CPU 时间片资源的不同权重，实现同一宿主机上的虚拟机计算性能 Qos;



不同优先级比例的时间片分布

vCPU 预留

即 vCPU Qos 预留，某些重要的应用需要足够的资源才能稳定启动和运行，因此在多虚拟机争抢资源的场景下，可以设置 CPU 预留最低的资源满足应用要求，避免争抢不到足够的资源导致应用程序无法启动或性能瓶颈；



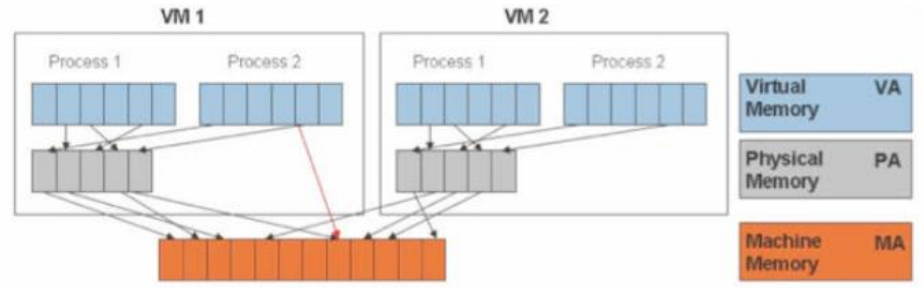
预留最低保障

vCPU 限制

即 vCPU Qos 限额，与预留相对的，某些虚拟机可能会不断占用物理 CPU 资源并难以释放调度，例如感染了病毒或其他持续的进程占用，造成其他的应用瘫痪。为使虚拟机充分隔离避免过度竞争，保证业务体验，可以设置其 CPU 占用的平均上限值。

3.2.3.4. 内存管理

内存虚拟化共享物理内存空间并动态分配给虚拟机，同时管理虚拟页到物理页的映射。除了 CPU 等硬件辅助优化技术外，系统提供气球技术、大页技术等优化内存的共享、调度、性能。



内存虚拟化技术

内存扩缩容

类似于 vCPU，内存应当根据工作负载作出规划。当内存分配不能贴合工作负载的增减趋势，应当扩缩容内存分配的容量。

扩缩容调整选择在较为稳妥的时间窗口进行，并明确可能造成的业务影响，一般建议虚拟机处于离线状态。如果要求在线调整，则需要满足一定的前提条件：包括物理 CPU 的芯片是否支持、虚拟机操作系统内核是否支持热添加特性等。CNware 支持在 X86、ARM 架构下的在线调整特性。

动态内存调整

动态内存调整技术采用 virt balloon 驱动实现，持续优化内存的分配与回收，保持计算节点的内存高效利用。

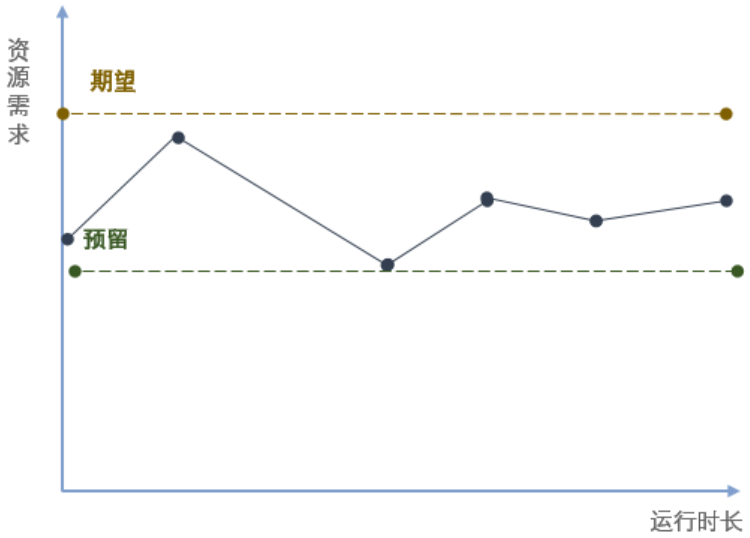
当计算节点的虚拟机几乎分配完主机的物理内存时，系统为了向其他迫切需要内存的虚

拟机提供资源，可以通过该机制动态从已分配的空闲内存中“挤出来”；该机制执行过程中，系统也会控制逼近可回收的极限，保证虚拟机的运行稳定。需要注意的是，内存的动态增加或减少，势必可能使内存被过度碎片化，从而降低内存使用时的性能，因此生产环境中的关键业务尽量使其内存得到充分的保障。

内存预留

内存预留为虚拟机在启动时立即分配足量的真实物理内存，优先保证业务性能。

一般情况下，虚拟机启动时获取的内存不等于申请分配的内存，而是由虚拟化管理调度层动态提供。如果其他业务也在不断争抢内存资源甚至发生内存超分，虚拟化管理调度层会出现难以提供真实物理内存的尴尬局面，导致关键业务无法分配内存空间而降低性能和稳定性，内存预留提供 Qos 的应对方案。



预留最低保障

大页内存

虚拟机可以使用大页内存加速应用；大页内存的涵义描述见 3. 3. 3. 9 章节。

内存分配策略

类似于 vCPU 的亲和性调度，虚拟机的内存也可以指定亲和分配规则，用于指定虚拟机在物理 Numa node 的分配模式，分配策略包括：

- Strict 类型：从指定的 node 节点分配内存，一旦内存耗尽，则无法再分配内存；这种模式优势是性能好，缺点是分配不到内存则容易触发 OOM kill VM 风险；
- Preferred 类型将从最优的 node 节点分配内存，一旦内存耗尽，则从其他节点中分配；
- Interleave 类型将通过轮询算法机制，从多个 node 分布内存占用。

依据内存分配策略的执行，虚拟机从指定的 node 节点分配内存；结合 vCPU 绑定功能，大幅度提高 CPU 访问内存的效率，从而使应用性能得到大幅度的优化提升。

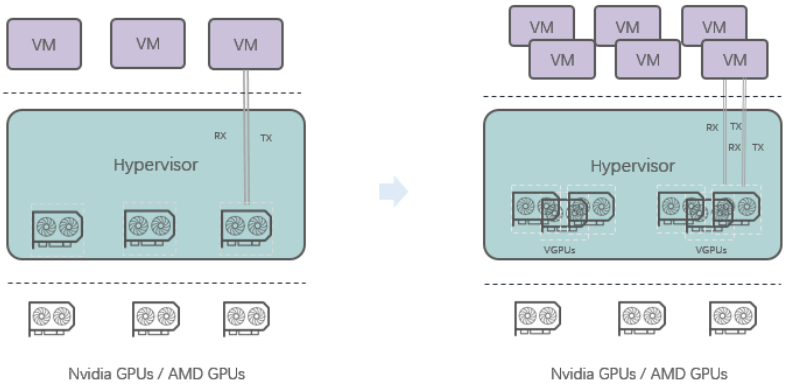
3.2.3.5. GPU 管理

GPU（图形处理器单元）主要进行浮点运算和并行运算，其浮点运算和并行运算速度可以比 CPU 强上百倍之多，适应适用计算密集、图形处理、虚拟现实、深度学习等场景；使用 GPU 虚拟化技术之后，可以让运行在数据中心服务器上的虚拟机实例共享使用同一块或多块 GPU 处理器进行计算。

GPU 在虚拟化领域的应用主要包括 PCI Passthrough 和 SR-IOV 两种方式。PCI Passthrough 即将 GPU 设备整体直通给指定的虚拟机使用，极大提升虚拟机的计算能力，缺点是利用率较差无法在多个虚拟机之间共享。因此，为了提升资源利用率，GPU 也采用了虚拟化技术，称为 vGPU.

vGPU 直通是一项 SR-IOV 的技术实现，CNware 主要支持 Nvdia/AMD vGPU 方案。vGPU

和物理 GPU 类似，都是为了实现桌面（图形处理）、计算（深度学习）等场景的加速，不同的是 vGPU 面向虚拟化和云的场景组成了资源池的方式；相较于物理 GPU 的 PCI Passthrough 方式，vGPU 能够对资源做更细粒度的按比例切割。



GPU 的 PCI Passthrough 方式&SR-IOV 方式

3.2.3.6. 虚拟磁盘管理

虚拟磁盘位于计算节点挂载的存储池，提供虚拟机读写和数据管理能力。

虚拟磁盘管理

支持为虚拟机挂载两种磁盘：

- 空白虚拟磁盘：全新创建的磁盘空间，不包含任何数据；
- 含数据的已有磁盘：即从虚拟机卸载移除、但保留磁盘上的数据的磁盘空间，支持再次挂载使用。

虚拟磁盘的删除包含两种方式：

- 流转至回收站，彻底销毁后删除磁盘逻辑信息；
- 数据安全擦除，磁盘被删除逻辑信息的同时所在的数据块位置被置零擦除，无法再从硬件恢复数据。

磁盘扩容

存储资源不足时，自然考虑为虚拟机扩容更多的数据写入空间。扩容主要包含两种方式，第一种方式是为虚拟机添加空白虚拟磁盘；第二种方式是扩容当前挂载的虚拟磁盘。第二种方式中采用冷扩容则兼容性和安全性较高，采用热扩容则依赖 NFS、ext4、共享文件系统等存储池类型的支持。

总线类型

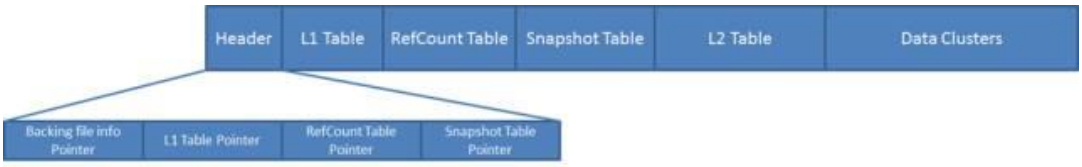
系统支持多种磁盘驱动类型，包括高速、IDE、SCSI、SATA、USB 等类型。

- 高速（virtio-blk）：半虚拟化类型的驱动，性能比较好，支持热插拔，但虚拟机操作系统需要安装 virtio 驱动；
- IDE：全虚拟化类型的驱动，性能比较差，不支持热插拔，虚拟机操作系统无需任何改动
- 其他模拟类型：模拟物理接口的类型，此处不作展开。

存储格式

CNware 主要支持 qcow2、RAW 类型的磁盘提供块存储服务，其他如 VMDK 等格式也支持转换并导入到虚拟化平台使用。

- 智能 qcow2：一种 qemu 支持稀疏文件格式，具有尺寸小、支持特性丰富的优势
- 高速 raw：KVM 原生支持的裸格式，速度更好，但支持特性简单



qcow2格式图解

缓存方式

缓存模式作用在虚拟化层和宿主机文件系统或块设备之间，主要包括四类模式：

- Writeback 模式：数据更新时只写入缓存 Cache，性能比较好但失去一定的安全性；
- None 模式：IO 操作绕开了 host 的页缓冲，相当于 vm 直接访问 host 的磁盘，性能比 Writethrough 要好。
- Writethrough 模式：数据直接写入磁盘里，不使用缓存；。
- Directsync 模式：与 Writethrough 区别是绕过 host 的页缓存，使用较少。

这几类模式从不同角度有一定的区分：

性能上： writeback > none > writethrough = directsync

安全上 ： writeback < none < writethrough = directsync

限制 IO 速率

由于虚拟磁盘基本都采用共享的存储池承载，很容易导致工作负载独占可用 IO 或其他资源，从而对同一环境中的其他工作负载或租户带来不利影响的这一情形。磁盘 Qos 是一种缓解存储端读写延迟、阻塞的有效手段，通过限制磁盘 IO 带宽速率，对某些 IO 密集的业务有立竿见影的限制效果，防止 IO 抢占过高导致其他业务瘫痪；

限制 IOPS

类似 IO 速率，读、写 IOPS 也可以设置限制值避免不安全抢占；

IO 悬挂时长

通常意义上，如果虚拟磁盘连接的存储链路发生中断，而虚拟磁盘仍收发不断的 I/O 请

求，则虚拟机系统会尝试将虚拟磁盘对应的文件系统只读保护，并返回 IO error 提示；只有存储链路修复后，虚拟机系统重启才恢复正常。

IO 悬挂是一项针对存储网络闪断等场景的智能挂起技术，为防止虚拟化集群发生大规模的虚拟机文件系统只读情况，系统将拒绝所有的 IO 请求；如果在设置的挂起时长内存储链路得到恢复，则虚拟机可以快速恢复访问和提供服务。

置备类型

置备类型包括精简置备、厚置备延迟置零、厚置备置零：

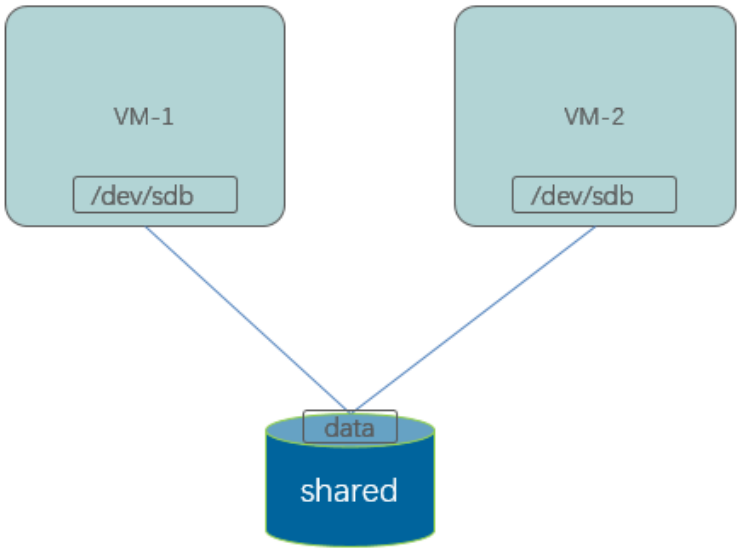
- 精简置备：灵活按需分配实际存储空间，具有节约空间的优势，创建速度快但性能稍差
- 厚置备置零：传统置备模式，预先分配全部空间并立即全部写 0，创建速度慢但性能最好
- 厚置备延迟置零：与置零的核心区别是不会立即写 0 抹掉数据，创建速度和性能均一般

共享/非共享磁盘

通常，虚拟磁盘仅挂载给一台虚拟机，这样能够保证数据安全的读写，这类称为非共享磁盘。在某些特殊的业务场景下，虚拟机之间需要挂载同一个虚拟磁盘，实现文件数据的共享访问，这类我们称为共享磁盘。实现共享磁盘必须作为数据盘而非系统盘。

共享磁盘附加到多个虚拟机，在虚拟机操作系统的层面能够识别到磁盘设备，但数据如何访问则取决于文件系统是否具有群集特性。如果采用非群集的文件系统，例如 ext4，一侧对磁盘进行了写操作，却无法在另一侧立即识别写入内容，或者无法知晓数据发生变更，因此安全可行的办法通常只有一个入口开放读写，其他入口仅能以重新挂载的方式以读模式访

问；由于巨大的不便，采用群集文件系统更能适应共享磁盘的优势，例如 MSCS、Oracle RAC 等，所有的入口均能提供读写模式，但读写的安全性由集群来控制所有的成员实现。



共享卷实现

3.2.3.7. 虚拟网卡管理

虚拟网卡提供虚拟机通讯的能力，链接为虚拟交换机上的端口实现流量传输。

网卡启用/禁用

当虚拟机流量异常，或者存在网络冲突等场景下，允许管理员禁用虚拟机网卡，不接入现有业务网络，提供网络隔离的环境检查系统。

网卡扩缩容

虚拟机网卡数量与业务的网络规划紧密相关。虚拟机需要支持增删虚拟网卡数目，实现特定的网络传输需求；虚拟化支持在线或离线的网卡变更，但在线情况下需要虚拟机操作系统的支持。

网卡类型

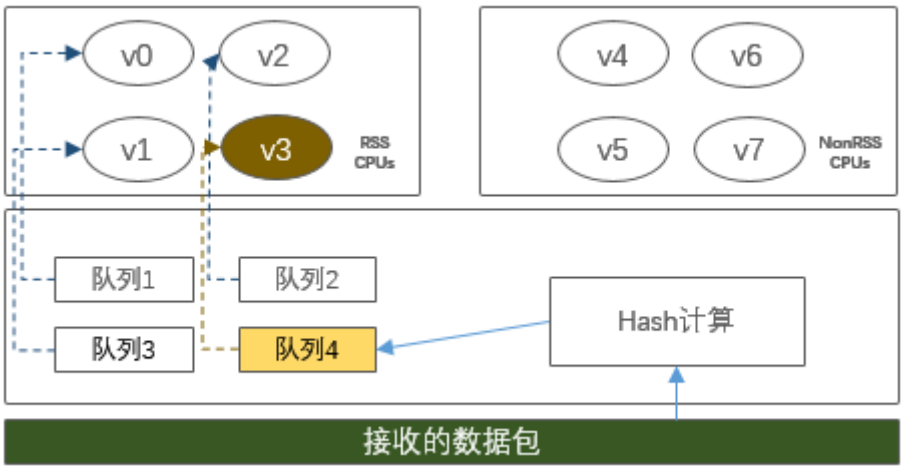
系统支持设置多种网卡类型：

- 高速网卡使用 virtio 驱动，性能比较好，也是默认的模式；
- 普通网卡则使用模拟的 RTL 网卡，较少使用；
- Intel e1000 是特殊的兼容性较好的网卡，较老旧的系统或测试使用；
- SR-IOV 网卡则是将支持 SR-IOV 的网卡进行逻辑切分，并将逻辑网卡映射直通为虚拟网卡，这种模式性能很高，但有迁移方面的限制。

网卡多队列

多队列网卡是一项网络性能优化技术，主要应对日益提升的网络 IO 带宽需求与单核 CPU 处理能力难以满足的挑战。其原理是：每张虚拟网卡通常对应一个队列，所有收到的包从这个队列入，内核从这个队列里取数据处理. 该队列其实是 ring buffer(环形队列)，内核如果取数据不及时，则会存在丢包的情况；一个 CPU 处理一个队列的数据，这个叫中断. 默认是 cpu0(第一个 CPU)处理. 一旦流量特别大，这个 CPU 负载很高，性能存在瓶颈。如果网卡实现多个队列，每个队列对应不同的中断，使用 irqbalance 将不同的中断绑定到不同的核，则能够大幅提升 IO 带宽处理能力。

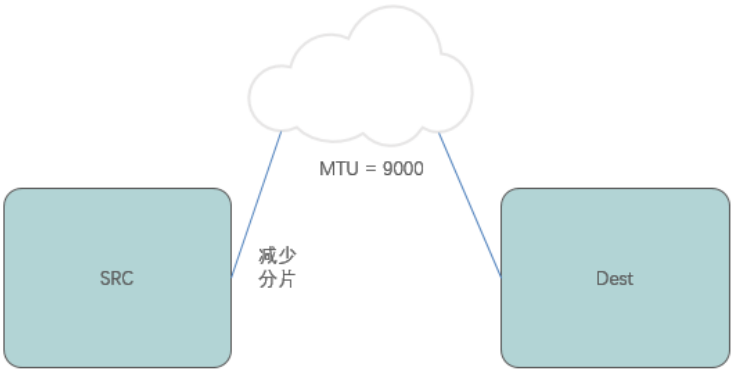
CNware 提供一键式网卡多队列配置，在多核并行、网络容易成为瓶颈的场景下能够明显改善业务处理性能。



网卡多队列原理

MTU 优化

最大传输单元（Maximum Transmission Unit，MTU）是网络调整的重要因素，用来通知对方所能接受数据服务单元的最大尺寸，说明发送方能够接受的有效载荷大小。一般情况下，MTU 默认传统分片大小为 1500bytes；但随着网络吞吐量和效率的提升，巨型帧（通常认为大于 1500，一般被认为最多携带 9000 bytes）的支持对于部分业务场景尤其是网络存储等大量数据传输的领域非常重要，通过减少 CPU 分片和重组报文的开销负担，能够令网络充分发挥性能，数据传输效率能提升 50%～100%。



MTU 传输优化

界面配置 IP

不管是为虚拟机添加或是更改 IP，运维者都希望拥有便捷的、安全的配置方式。借助虚拟化 CNware tools，系统支持在图形界面为虚拟机一键配置 IPv4、IPv6 等类型的 IP 地址，并能够预先检测 IP 是否被占用防止冲突，提升管理运维体验。

3.2.3.8. 远程控制台

虚拟机远程控制台提供安装、访问操作系统的方式。虚拟机控制台的实现支持 VNC 和 SPICE 两种协议：

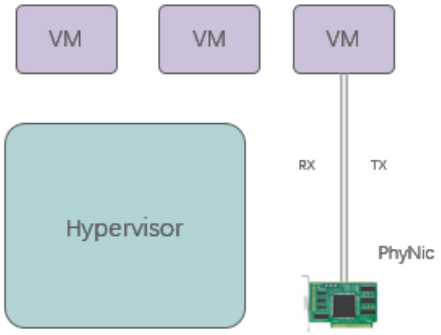
- VNC 控制台：提供基于 HTTPS 协议安全访问的 web 高清访问；
- spice：支持使用兼容 SPICE 协议的如 remote-viewer 等工具进行远程连接。

虚拟机远程控制台需要占用宿主机的一个端口，端口分配规则支持自动或手动的方式；默认没有特殊要求下，建议使用自动分配以避免端口冲突和繁琐的配置。在连接安全性方面，CNware 提供了 HTTPS 加密的通道，以及密码确认的认证机制。

3.2.3.9. PCI 设备直通

虚拟机使用透传设备可以获得设备近乎原生的性能，这里面涉及到设备透传（PCI Passthrough）技术。PCI Passthrough 主要实现虚拟机排他使用主机上的某个 PCI 设备,就像将该设备物理连接到虚拟机上一样，能够大幅提高性能、降低延迟、直接使用裸设备驱动。这项技术拥有广泛的应用场景，例如在银行中需要将 USBKey、GPU 等设备直通给虚拟机使用，或者直通网卡到虚拟机大幅度提升 IO 性能，这依赖于硬件的支持（硬件辅助虚拟化技术），包括主板、处理器需要提供将 PCI 物理地址映射到客户虚拟系统的方法，例如 Intel VT-x，EPT 及 VT-d 技术。

CNware 几乎支持直通所有的 PCI 设备类型，包括如以太网控制器、USB 控制器、HBA 控制器、GPU 显卡等。系统自动识别同步服务器上的设备类型及 PCI Number，持久化到管理平台数据库；在此基础上，系统简化了 PCI 设备加载至虚拟机系统的配置过程，符合产品易用性的目标。值得一提的是，尽管 PCI 设备透传到虚拟机正确识别为硬件文件，也须依赖对应驱动程序才能使用。



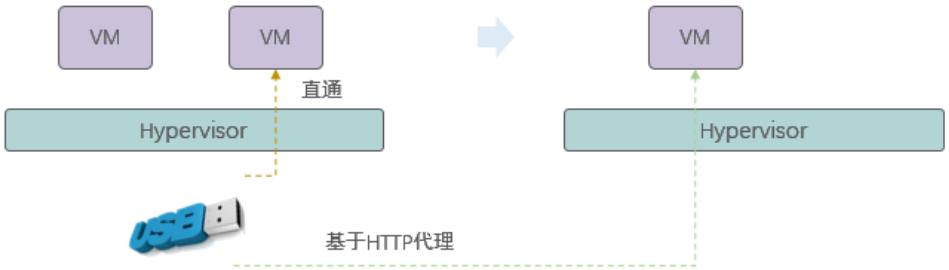
PCI Passthrough 原理

3.2.3.10. USB 直通与重定向

USB，即通用串行总线。作为一种计算机外设，是一种外部总线的标准，用于规范主机与设备之间的通信与连接。USB 系统架构分为三个部分，USB 主机控制器/根集线器（又名，USB 总线接口），USB 集线器，USB 设备。USB 主机控制器接在 PCI/PCIe 总线[1]上，USB 集线器又接在 USB 主机控制器（根集线器）上，USB 设备接在 USB 集线器上。在主机的控制器上，采用分时共享机制控制着所有连接的不同厂商的 USB 设备，但它们都遵循标准的传输规范。

例如 U 盘、加密狗等属于 USB 设备，部分应用程序需要借助 USB Key 的辅助实现安全性、机密性的场景；但由于虚拟机并不像物理机一样能够便捷地连接到硬件设备，因此需要使用 USB 设备直通实现，这项也是 PCI 设备透传技术之一。USB 设备直接分配到虚拟机后，虚拟机中的应用程序可直接访问 USB 设备，而不需要通过虚拟化层进行管理。在这个过程中，USB 设备要从宿主机注销，再向虚拟机虚拟硬件注册该配置。

对于部署在集群内的业务来说，发生位置的迁移改变是相对频繁的事情。USB 设备重定向技术面向的是解决虚拟机发生跨主机的迁移后，仍然能够保持访问关联的 USB 设备这类场景。在多种技术方案中，CNware 实现适应性较强的 HTTP 代理方案，实现 USB 设备跨主机的映射支持、自动加载，无需虚拟机特殊的插件支持。



USB 设备直通与重定向

3.2.3.11. RDM 磁盘直通

我们知道，在虚拟化平台中，虚拟机的磁盘几乎都位于虚拟的存储池。这个存储池构建于一个或若干个物理存储卷，不仅需要经过一层虚拟化的“转换”损耗，同时要共享地分配该卷空间的性能。为了获得更高的空间和性能，虚拟机可能希望直接独占式地访问物理存储卷；RDM 磁盘直通正是解决该问题的技术，通过实现增加硬盘读写速度、减小延迟，从而避免因为虚拟化之后的资源浪费。

CNware 支持直通本地或远程的物理存储设备，包括本地的 SCSI 磁盘、远程的 ISCSI LUN、FC LUN 等。需要明确的是，有优势就会对应有局限性，使用直通模式的磁盘无法有效完成快照、克隆、迁移等特性，在部署决策中必须周全考虑取舍。

3.2.3.12. 虚拟机快照

快照用于保存虚拟机在某个时间点的内存、磁盘和设备的状态数据，且当有需要的时候能够重复的回滚至该时间点。根据被做快照的对象不同，快照可以分为磁盘快照和内存快照，两者也可以叠加构成一个还原点，记录虚拟机在运行时的全部状态；根据做快照时虚拟机是否在运行，快照又可以分为在线快照和离线快照。

磁盘快照根据存储方式的不同，又分为内部快照和外部快照：内部快照只支持 qcow2 格式的虚拟机镜像，把快照及后续变动都保存在原来的 qcow2 文件内；外部快照在创建时，快照被保存在单独一个文件中，创建快照时间点之后的数据被记录到一个新的 qcow2 文件中，原镜像文件成为新的 qcow2 文件的 backing file（只读），在创建多个快照后，这些文件将形成一个链——backing chain。从 V9.3.1 版本开始，CNware 默认提供外部快照的方式。

快照列表

创建多个快照后，将形成类似 VMware vSphere 的一个或多个分支的树状快照列表。快照之间具有父子关系，通过父子关系与分支，可以定位和还原到指定的虚拟机状态。支持通过关键字快速搜索目标快照。

快照树视图

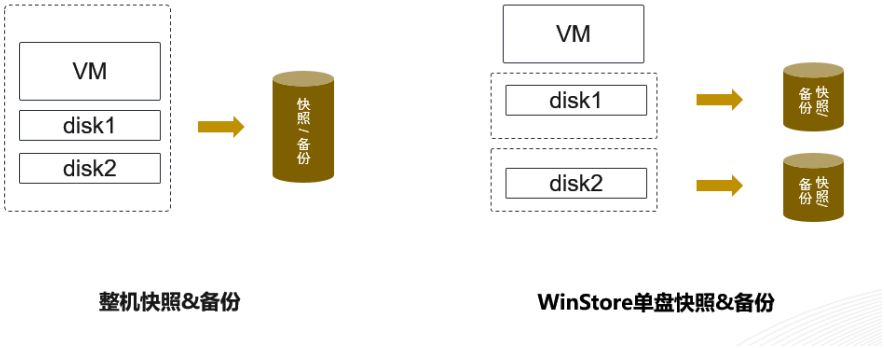
快照支持

支持虚拟机的在线、离线状态创建快照：

- 仅磁盘快照：保留虚拟机系统已往硬盘写入的数据，在线或离线均支持；
- 内存+磁盘快照：除了写入虚拟硬盘的数据，还增加虚拟机操作系统当前在内存

驻留的数据，仅支持在线创建。

＊ 上述快照默认为整机快照，仅 winstore 支持对单块虚拟磁盘创建快照。



还原快照

把虚拟机还原为快照点的状态；

- 如果是仅磁盘快照模式，虚拟机还原为快照点的磁盘数据状态和关机电源状态；
- 如果使用内存+磁盘快照模式，虚拟机还原为快照点的内存数据和磁盘数据，且为开机状态；例如创建内存快照时已打开浏览器程序，还原时仍然是开启浏览器程序的状态。

删除快照

删除任一分支点的快照，且其下含子快照时也允许一并删除。

3.2.3.13. 虚拟机克隆

安装操作系统和应用是一件比较耗时的工作。使用虚拟机克隆技术能够快速从源虚拟机把规格配置、操作系统、应用等数据原封不动地复制出多份“副本”，使用副本的用户能够体验与源虚拟机完全一致的数据，且克隆生成的新虚拟机与原虚拟机磁盘独立，磁盘数据互相不受影响。在分发测试、故障模拟、临时性的备份等场景中，虚拟机克隆是非常快速且有

效的方法。

3.2.3.14. 虚拟机 OVF 模板

OVF 是一种开放的虚拟机模板规范，最早是由 DMTF 组织提出的一项虚拟化管理规范的草案，其目标就是为虚拟化设备制定一个互操作规范，即所谓的“开放虚拟机格式”（Open Virtual Machine Format,OVF），在国家标准 GB/T 35292-2017 也作了细致的规范要求；简单来说，OVF 致力于开发一种“开放的、安全的、可移植的高效以及可扩展的格式”，以便更好地封装与分发虚拟机。OVF 文件可以抽象看做一个由规定的包括描述、封装等几个不同类型的文件所组成的文件包，这个文件包可作为以后不同虚拟机之间一个标准可靠的虚拟文件格式，实现不同虚拟机之间的通用性。

但是，在实际的实施应用中却不能保证 OVF 模板能够完全跨虚拟化兼容，原因可能包括不同的虚拟化厂商对 OVF 规范描述、虚拟设备、虚拟驱动有着不同的解释和要求。考虑到虚拟化市场的衔接和替代，CNware 支持兼容自有平台或 VMware vSphere 等的 OVF 模板格式，实现方便的虚拟机跨平台迁移。

CNware WinSphere 支持将虚拟机导出并压缩处理为 OVF 模板；当部署 OVF 模板时，系统从存储池中解析 CNware/VMware OVF 模板的描述文件，自动生成规格配置、磁盘数据都与原模板完全一致的虚拟机。

3.2.3.15. 虚拟机 HA 高可用

集群 HA 高可用侦测的是虚拟化宿主机的可用性，并作出计划外故障的应对策略，保证业务持续运行；虚拟机 HA 高可用特性也是一项业务连续性保护技术，但不同的是它是针对

虚拟机操作系统本身的可用性侦测和应急保护方案，可以最大程度减少不可预知的系统宕机和服务中断时间，整个过程也无需任何人为干预。

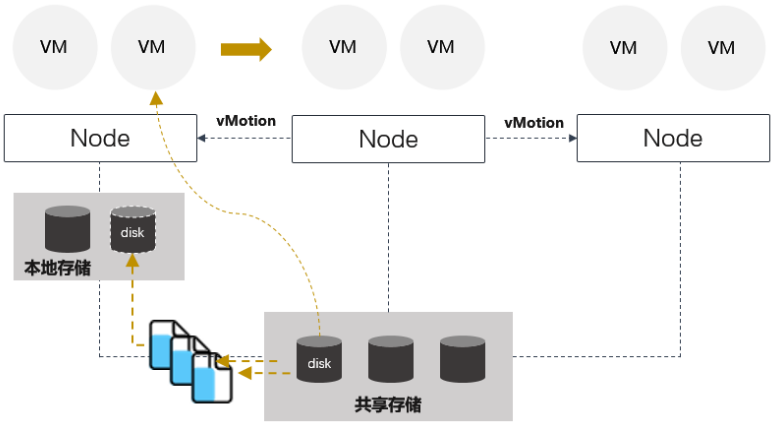
CNware 的虚拟机 HA 实现机制主要是通过判定虚拟机的 CPU、IO 等进程的活动行为状态，系统会持续对加入 HA 故障保护的虚拟机进行侦测，一旦虚拟机系统内发生计划外故障的时候（如蓝屏、关键进程崩溃等），迅速响应并自动触发虚拟机电源硬重启机制，重新拉起该虚拟机的 qemu 进程。

3.2.3.16. 虚拟机迁移

虚拟机迁移是虚拟化的主要目的之一，几乎成为虚拟化集群管理的日常。管理员希望在无需重新配置物理和虚拟网络的情况下改变虚拟机的位置，达到如下的目的：

- 简化系统维护管理
- 优化资源空间利用
- 高系统负载均衡
- 增强系统错误容忍度
- 优化系统电源管理

根据虚拟机运行状态，可以分为在线迁移（热迁移）、离线迁移；根据虚拟机迁移前后相对集群的位置，可以分为集群内、跨集群迁移；根据迁移时是否带内存、磁盘数据，可以分为三种方式：



虚拟机迁移示意

- 更改主机：指仅更改虚拟机归属的宿主机，前提要求虚拟机的所有磁盘均位于共享存储池；源目主机 CPU 型号须尽可能保持一致（与工作模式相关）且计算资源充足。
- 更改数据存储：指仅更改虚拟机的磁盘存储所属的存储池位置；支持共享存储池与非共享存储池互迁；支持将所有磁盘迁移到同一存储池或分别迁移到指定的存储池。
- 更改主机和数据存储:指更改虚拟机及其磁盘的位置，整体迁移到其他的宿主机和存储池，包括本地至共享存储、本地至目标宿主的本地存储。

虚拟机迁移网络

虚拟机迁移过程中，如需迁移存储数据，则会带来大量的数据传输。系统支持指定专门的迁移网络，并且可对网络流量进行限速。进一步提升数据传输安全性的同时，避免了迁移流量对网络带宽带来的压力和冲击。

虚拟机迁移策略

根据虚拟机业务繁忙程度，可以通过搭配迁移策略提高迁移效率。在业务繁忙的情况

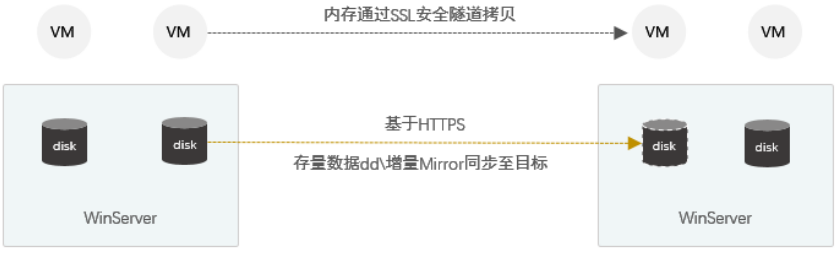
下，选择业务优先并设置自动降频，可以在保障业务连续性的前提下快速完成迁移。

当涉及到迁移虚拟机存储的时候，可以通过对虚拟磁盘的写 IO 速率、写 IOPS 进行限速的方式，选择后台同步或同步写入的方式，以应对虚拟机 IO 压力较大的场景。

后台同步，虚拟机在迁移过程中的增量 IO 数据仅写入源端，后台再将源端数据同步至目标端，当虚拟机的 IO 压力较大时，这种模式对虚拟机 IO 几乎没有影响，但高速增长的增量 IO 数据会导致需要迁移更多的数据，使得迁移时长相较同步写入更长，但对虚拟机的 IO 没有影响；同步写入，虚拟机在迁移过程中的增量 IO 数据同时写入源端和目标端，当虚拟机的 IO 压力较大时，由于增量 IO 数据需要同步写入目标端，因此这种模式对虚拟机的 IO 有一定的影响。

虚拟机迁移安全

虚拟机迁移涉及大量数据拷贝操作，须采用基于 HTTPS 的隧道协议（支持国密算法扩展）保证数据传输过程的加密安全。迁移过程，通过持续的增量拷贝及校验算法，能够动态检查迁移数据的完整性和一致性。



虚拟机迁移安全性设计

虚拟机迁移日志

无论是手工迁移，还是 HA 高可用或 DRS 动态调度触发的自动迁移，运维者可能希望提

供从源端至目标端的完整迁移历史记录，包括触发的对象、时间、是否人为、耗时以及日志详情，在回溯追查时相当有用。

虚拟机批量迁移

虚拟化带来巨量的规模化效应优势，同时引入的管理成本也是线性递增的；当集群内需要迁移数十台虚拟机时，恐怕对管理运维者是个噩梦。CNware 提出了批量迁移的特性，旨在针对性克服虚拟机迁移倍增的工作量，合并多台虚拟机为“一批”和“一键式”来减少大量重复操作的次数，对业务规模庞大、希望提升运维管理效率的用户来说是非常有吸引力的。

虚拟机同构跨芯迁移

同构跨芯指的是同芯片架构（例如同为 ARM）、但芯片厂商不一致（例如鲲鹏-飞腾），虚拟化平台支持（Intel-海光、鲲鹏-飞腾）的迁移能力，虚拟机免应用重构、免修改，即可实现整机迁移至同构的异构芯片。

跨版本迁移

在多套虚拟化环境升级场景中，为了降低手动运维工作量，最大限度减少业务中断时间，系统实现了跨管理平台、跨版本间虚拟机平滑迁移，支持在可视化界面按平台维度批量在线/离线迁移虚拟机。

P2V/V2V 迁移

在业务上云、切换虚拟化云底座等场景中，客户最容易头疼的就是业务的迁移问题：

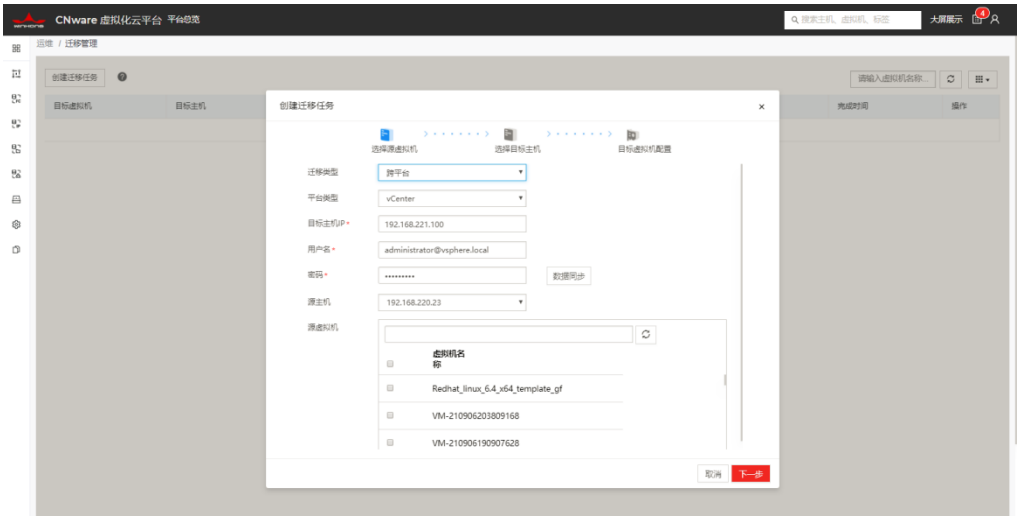
- 老旧系统迁移困难，业务层面无法迁移

- 多厂家难协调，涉及操作系统、数据库、中间件厂商过多，难以配合
- 业务停机窗口长，迁移结果难以预测，担心兼容性

这里包含 P2V、V2V 两类迁移场景，云宏均提供了成熟可靠的方案很好地解决以上的痛点。V2V 迁移指的是，将其他虚拟化类型（例如 VMware、Xen、Hyper-V、开源 KVM、其他商业产品）的虚拟机系统转换到目标虚拟化（CNware）上运行；P2V 迁移指的是，将原本以物理裸机的方式承载的操作系统转换到目标虚拟化（CNware）上运行。

云宏提供的方案如下：

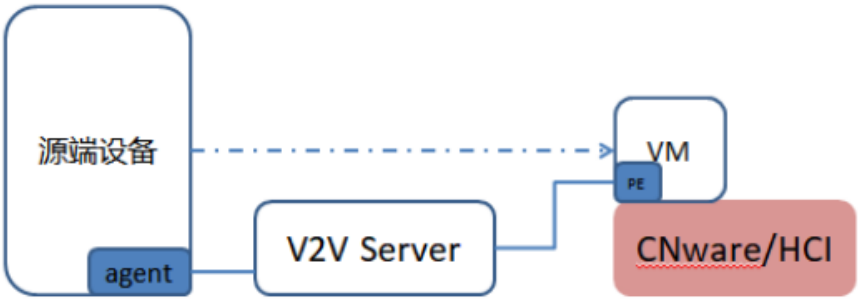
- 针对 VMware vSphere 虚拟化，云宏在平台上提供无成本的两类方式，包括 OVF 模板导入、跨平台直迁，实现 VMware-CNware 快速的、平滑的虚拟机迁移。
- VMware OVF 模板是从 vSphere 平台导出的一项标准描述文件，CNware 通过自动解析配置文件和磁盘文件，实现导入即生成；跨平台直迁提供 VMware vCenter 到 CNware 的连接通道，实现虚拟机的自动发现和批量迁移，在迁移过程也能够方便的指定转换后的目标存储和网络：



- 针对 P2V、其他虚拟化类型，推荐使用云宏的云迁移工具。

云迁移工具的实现原理是，通过源端部署的迁移 agent 代理，与控制台（图中的 V2V

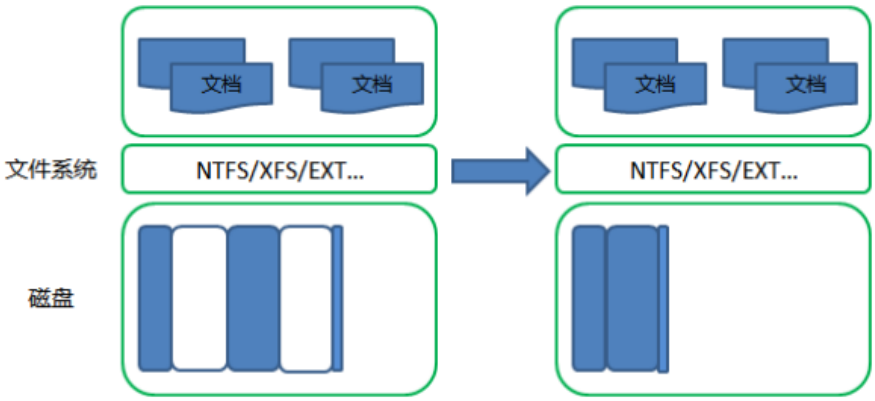
Server）以及目标虚拟机的 PE 系统建立联系，然后采用 P2P 技术实现数据的迁移。



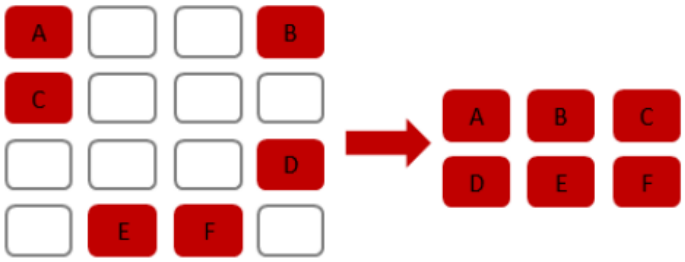
云迁移工具原理

云迁移工具具有如下优势：

- 整机迁移：源端数据通过 P2P 网络实现系统+数据+业务一次性迁入目标端，是目前上云迁移最有效、最高效的方式；
- 在线迁移：用户可以在不影响源端业务的同时进行全量迁移操作，直到数据基本复制完成后协调数分钟的停机窗口进行增量数据的同步，以及源目切换，从而保证数据的一致性；
- 精简迁移：通过智能的磁盘使用空间识别技术，仅拷贝磁盘有效写入的数据空间，极大的减少数据传输量和时间。如图所示，磁盘空白未写入的区域是不作拷贝的，只留下指针记录并在目标端恢复。



- 增量迁移：用户关心的由于业务的原因暂时无法切换至目标端产生新的数据如何重新同步的问题，通过磁盘变化记录能够解决，仅需同步增量变化的部分，有效缩短迁移和停机窗口。



依据记录磁盘块变化

- 兼容全面：覆盖主流的、多个版本的 Linux、Windows 操作系统。

3.2.3.17. 批量部署

批量部署的目的是提升业务交付效率，同时保持环境的一致性。

部署模式包括：

- 普通部署：生成的虚拟机与镜像数据完全相同，且所有新虚拟机的磁盘均为独立完整的存储卷拷贝，互不影响。
- 快速部署：生成的虚拟机与镜像数据完全相同，但同一宿主机的所有新虚拟机的磁盘使用共享的基础卷，因此系统会大幅提高创建虚拟机的速度且节省存储空间；但由于使用共享的基础卷，因此会损失安全性，一般只适用于测试场景。

策略部署

基于策略的部署是面向集群批量分发生成虚拟机的一项特殊特性，通过结合宿主机监控

和调度策略控制虚拟机的期望落点，实现更优的负载分布。

策略包括：

- 集中式策略：按 CPU、内存利用率排序选择宿主机，根据设置的宿主机利用率阈值，镜像将向策略推举的主机集中分发虚拟机；
- 分散式策略：默认在集群中轮询部署，镜像将向策略推举的主机轮询式分发虚拟机。

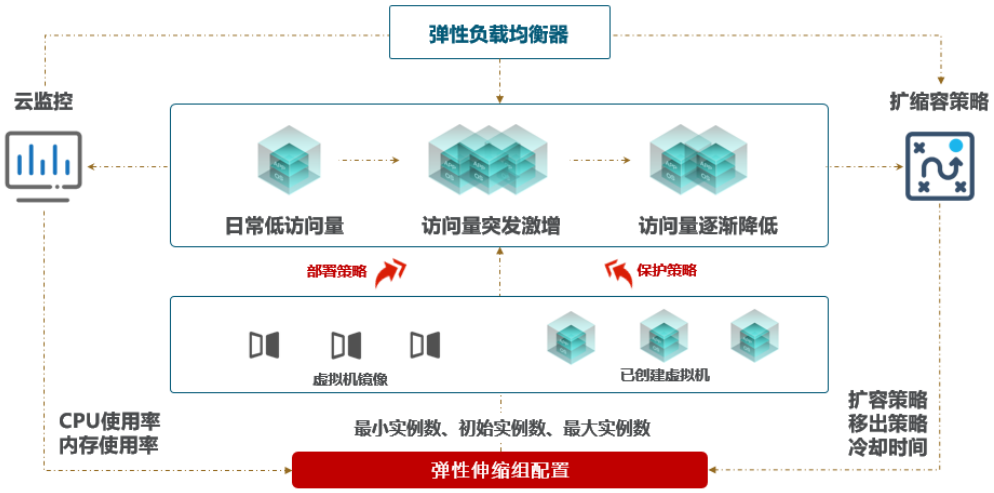
3.2.4. 资源动态调整

弹性快速扩缩容是云计算的目标之一，尤其适合脉冲型、突发式的、难以预判资源消耗例如应用网站、数据处理等业务场景，为了在业务访问增长时能够自动扩容以保证充沛的计算及服务能力，或者业务访问下降时自动缩容以释放算力占用来节约成本，降低人为反复调整，仓促应对业务高低峰的工作量，具有自动调整、提高可用性和容错能力的优势。

在国产化的环境下，国产芯片硬件性能、稳定性参差不齐，承载的业务出现波峰波谷的几率迥异，动态调整计算资源的灵活性能够一定程度上减少这类影响因素。

3.2.4.1. 弹性伸缩组

弹性伸缩组通过设置监控规则和伸缩规则，根椐虚拟机的历史性能表现，增加或减少虚拟机实例，来实现自动化横向增加或缩减计算资源，以确保业务持续高效地运行。



伸缩组定义

弹性伸缩组实例模板包含最小虚拟机数量、起始虚拟机数量、最大虚拟机数量，平台从镜像依据集中/分散部署策略初始化生成期望数量的虚拟机，之后在最小与最大数量之间按触发规则执行扩缩容。

伸缩规则

扩容、缩容规则采用虚拟机 CPU、内存的平均使用率作为依据。根据或与条件判断，以及监控持续时间和间隔时间检测触发条件。

此外，缩容时还定义了规则：

- 移除最新加入的实例
- 移除最旧加入的实例

关联负载均衡

负载均衡器是将访问流量根据转发策略分发到后端多台虚拟机的流量分发控制服务，弹性伸缩组与负载均衡器实例联动，使扩缩容更新的实例快速同步注册到负载均衡监听。

关联虚拟机

除了从镜像生成虚拟机加入伸缩组外，系统支持将已创建的虚拟机实例加入伸缩组，并支持指定策略：

- 保护策略：即缩容时实例处于保护状态，系统拒绝将指定实例移出弹性伸缩组；
- 备用策略：即实例处于备用状态，系统将实例暂停服务但不会移出弹性伸缩组，当需要扩容时实例可快速重新上线提供服务；
- 无策略：即实例与其他新建的实例无差别对待，缩容时系统按最新或最旧策略进行删除。
- 备用策略、无策略可能影响现有业务的运行，须完全知晓其工作原理并合理配置。

3.2.4.2. 虚拟机热添加

虚拟机热添加通过热添加策略，依据虚拟机的历史性能表现，自动化增加虚拟机实例的计算资源（CPU、内存）的规格，来实现纵向扩展虚拟机的计算能力，以确保业务持续高效地运行。

热添加策略

热添加策略包含触发热添加的性能阈值、热添加的规格设置以及关联的虚拟机。

依据热添加策略，系统自动分配额外的资源给虚拟机，而无需停止虚拟机或对其进行重新启动，满足应用程序对计算能力的需求，同时确保持续的运行和可伸缩性。

功能优点

- 实时性：无需停机或重启虚拟机即可增加计算资源，快速满足应用程序的需求。
- 灵活性：管理员可以根据实际需求动态地调整虚拟机的规格，避免资源浪费。
- 可伸缩性：随着工作负载的增长，通过热添加策略可以轻松扩展虚拟机的计算能力，提高系统性能和响应能力。
- 降低维护成本：避免了停机时间和重新配置的繁琐过程，减少了对管理员的工作负担。

3.3.存储虚拟化

存储虚拟化的目的是对存储硬件资源进行抽象化表现，将数据也从磁盘中抽象出来。在传统的存储与计算互联架构中，受限或依赖于阵列存储的协作，形成难以互通的数据孤岛，这就需要依赖存储虚拟化技术的突破。

第一个层次，虚拟化存储首先打破的是存储的供给形式，迁移到虚拟化环境下提供共享访问的，允许按需扩展的存储池空间，承载抽象化的虚拟机磁盘；第二个层次，虚拟化推动了基于虚拟存储的卷复制、快照、迁移、自动精简等高级特性，数据几乎能在任意的存储系统之间动态共享，完全不受制于连接的系统；第三个层次是基于通用服务器及硬盘并通过软件定义的方式实现“SAN 存储”，实现分布式的、去中心化的虚拟化存储池，能够线性任意规模的扩展。

从虚拟化软件的角度，不论存储虚拟化技术如何发展，最终都是为了服务数据的承载；而虚拟化软件如何稳定兼容传统的、新兴的各种存储类型，才是长期面对的最大挑战。

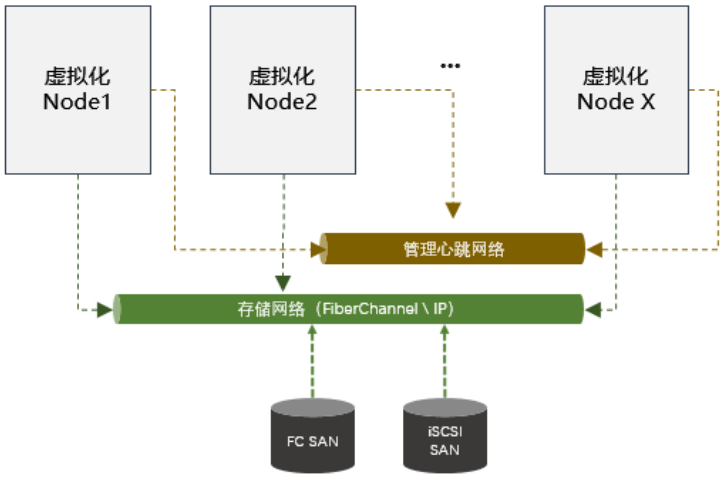
3.3.1. 共享文件系统

共享文件系统是云宏深度优化的一种专门服务于虚拟化的集群文件系统，通过分布式锁机制实现多个节点协同，确保被访问的资源不会因为并发而出现不一致的情况，因此可以封装为存储池并挂载给虚拟化集群，提供共享的数据存储服务。

存储后端类型

FC、iSCSI 等存储设备提供 LUN 块存储作为共享文件系统的后端（backend）。此外，CNware 也支持华为 NVMeoF 等全新的存储技术。

原理是，系统提供了自动扫描存储设备和格式化的机制，仅需存储侧协同映射好存储。例如添加 FC LUN，系统会自动过滤计算节点识别的 LUN 设备和设备 ID；例如添加 iSCSi LUN，系统会根据指定的目标 IQN 过滤扫描得到的 LUN 设备和设备 ID。存储设备须被格式化后才能挂载给集群访问。



共享文件系统与存储后端的关系

3.3.2. 存储池管理

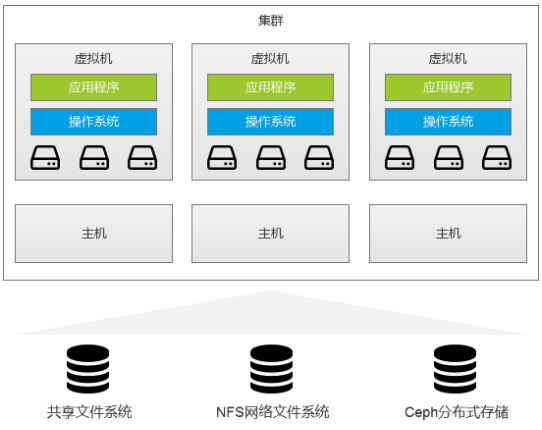
如前所述，虚拟化存储池提供对后端真实存储提供者、前端的虚拟机消费者的数据抽象化表现。即存储提供者只需要负责存储空间的映射供给，虚拟磁盘作为使用者只需要关注有无空间可以存取，两者并不关心数据如何读写以及如何呈现给虚拟机，这部分的工作由虚拟化存储池解决。

共享存储池

字面理解可知该类型的存储池具有共享、多点连接的属性。共享存储池提供给集群内所有节点可连接访问的数据存储能力，且支持虚拟机迁移、HA 等高级特性基础。

共享存储池由以下存储后端类型完成封装：

- 共享文件系统：云宏优化的一种集群文件系统；
- NFS：网络文件系统，一般称之为 NAS 存储；
- 分布式存储：基于 RBD 协议访问的分布式存储，包括云宏自有的 WinStore 或第三方基于 ceph 的存储。

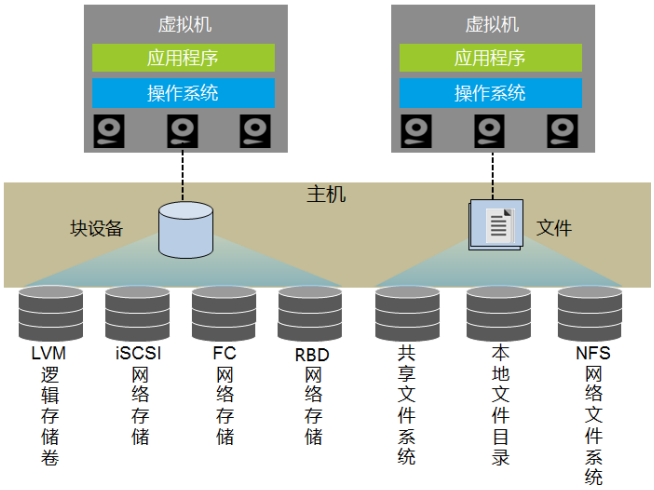


共享存储池供给类型

本地存储池

相比于集群共享，本地存储池是指非共享的、仅限挂载到指定的宿主机可使用的数据存储。

- 本地目录：宿主机本地格式化的 ext4、xfs 等类型的文件系统；
- iSCSI 网络存储：映射给宿主机的 iSCSI LUN，激活为 VG 卷组
- FC 网络存储：映射给宿主机的 FC LUN，激活为 VG 卷组
- 本地 LVM 存储：宿主机的本地磁盘，激活为 VG 卷组



虚拟机使用存储资源的方式

本地存储池供给类型

存储池管理

存储池的连接挂载支持如下管理操作：

- 启动：激活指定的存储池，存储池下的虚拟磁盘卷允许挂载使用。
- 暂停：抑制指定的存储池，虚拟磁盘不允许使用，但虚拟磁盘的内容将会保留，同时虚拟机用来访问虚拟磁盘的元数据信息也将保留。
- 删除：断开当前主机与指定存储池的连接关系，同时虚拟机连接到存储池上的虚拟磁盘信息将永久销毁。

3.3.3. 存储卷管理

存储卷包括放置在存储池内的虚拟磁盘、ISO 文件（须文件系统类型的存储池），用于挂载给虚拟机使用。

存储卷类型

存储卷类型主要包括磁盘类型、置备类型两个维度。

磁盘类型包括智能和高速：

- 智能 qcow2：一种 qemu 支持稀疏文件格式，具有尺寸小、支持特性丰富的优势
- 高速 raw：KVM 原生支持的裸格式，速度更好，但支持特性简单

置备类型包括精简置备、厚置备延迟置零、厚置备置零：

- 精简置备：灵活按需分配实际存储空间，具有节约空间的优势，创建速度快但性能稍差
- 厚置备置零：传统置备模式，预先分配全部空间并立即全部写 0，创建速度慢但性能最好
- 厚置备延迟置零：与置零的核心区别是不会立即写 0 抹掉数据，创建速度和性能均一般

存储卷管理

管理员新建虚拟磁盘，可以指定名称、大小、磁盘类型、置备类型；删除存储卷时包括软删除和硬删除方式。

- 软删除：存储卷将从存储池中移入回收站，仅彻底销毁后才重新回收空间；
- 硬删除：勾选擦除磁盘，则存储卷所在数据块数据将被彻底删除无法恢复，从而

达到安全删除的要求

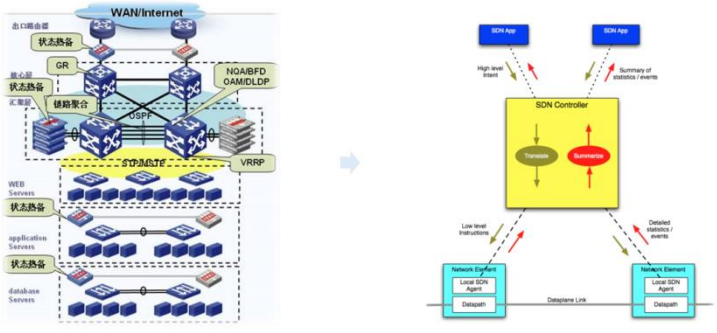
存储卷导入导出

导入导出提供了一种存储卷跨平台、跨环境转移的方式。存储卷能够通过浏览器导出到本地，通过导入功能可再次将存储卷存放到指定的存储池。

存储池与存储卷的关系

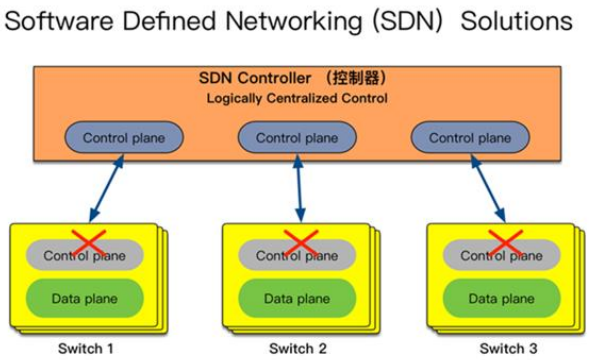
3.4.网络虚拟化

网络虚拟化是实现软件定义数据中心美好愿景的最漫长的旅途，其目的是在一个物理网络上模拟出多个逻辑网络，以支持更丰富的、比之传统网络架构更高效的特性。随着云网络技术的发展，网络虚拟化衍生了更为广泛的概念，一般大致可以分为：网络虚拟化抽象（NV）、软件定义网络（SDN）、网络功能虚拟化（NFV）。



传统数据中心网络　　云数据中心网络

- NV（Network Virtualization）：这是云数据中心发展的必然趋势，云要实现更灵活快速的网络交付、租户网络要隔离、安全策略要随业务迁移等，必须在物理网络中构建抽象的、叠加（Overlay）的虚拟网络(隧道封装技术可以是 VXLAN、NVGRE、Geneve 等)，最终还能与云平台融合紧密协同。
- SDN(Soft Define Network)：SDN 应用前景光明，不仅在数据中心内，甚至在园区网、城域网、骨干网均有广泛的应用场景，只要有网络灵活多变，业务部署复杂，就是 SDN 的目标所在。SDN 拥有三个典型特征：控制平面与转发平面分离、控制平面集中化、网络可编程。其中，业界最为关注的是转发平面的通用化，并推动了 openflow 等一系列令人瞩目的技术诞生。



- SDN 集中式控制平面
- NFV（Network Function Virtualization）：主要由 ETSI(即欧洲电信标准化协会)等运营商群体提出，其目标是采用通用的硬件及虚拟化技术取代昂贵的专用硬件，降低通讯网元的建设成本。但目前发展的局限性在于通用的 CPU 性能难以企及专用的 NP、ASIC 芯片。

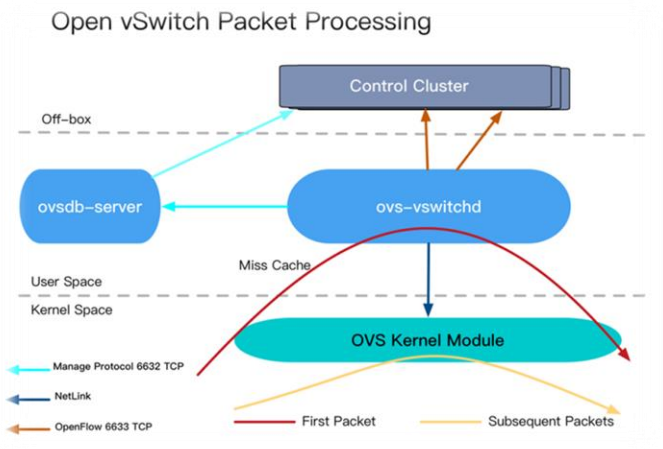
三者提出的背景有别，解决的问题也有别。总结而言，SDN、NV、NFV 分别关注网络的

面、角、点，三者在技术或思想上的借鉴与融合能够带来下一代的网络巨变。

3.4.1. 虚拟交换机

虚拟交换机即通过软件实现的交换机，具备硬件交换机的功能，但与硬件交换机相比减少了采购成本，且软件层面更具灵活性。虚拟交换机主要使用在云环境中，在虚拟机之间、虚拟机和外部网络之间实现网络的连通。

CNware WinSphere 采用主流的 openvSwitch 技术实现虚拟交换机。Open VSwitch（OVS）是一种具有生产级质量的多层虚拟交换机，OVS 主要包括用户态的转发逻辑处理线程 ovs-vswitchd、数据库配置线程 ovsdb-server 和内核态的 datapath 模块。Ovs-vswitchd 线程是 switch 功能的实现核心，同时支持 openflow，包含多个配置工具如 ovs-dpctl 等；Ovsdb-server 线程用于处理 switch 和 flow table 的配置；Datapath 模块根据 flowtable 规则，处理数据包转发的任务。



OVS数据包处理过程

与硬件交换机相比，最大的优势在于：

- 配置灵活：一台普通的服务器可以配置出多台虚拟交换机且端口数目可以灵活选

择，也可以通过可编程扩展来实现大规模网络自动化配置、管理和维护。同时，OVS 支持现有标准管理接口和协议，如 netFlow，sFlow，SPAN，RSPAN，LACP 等；

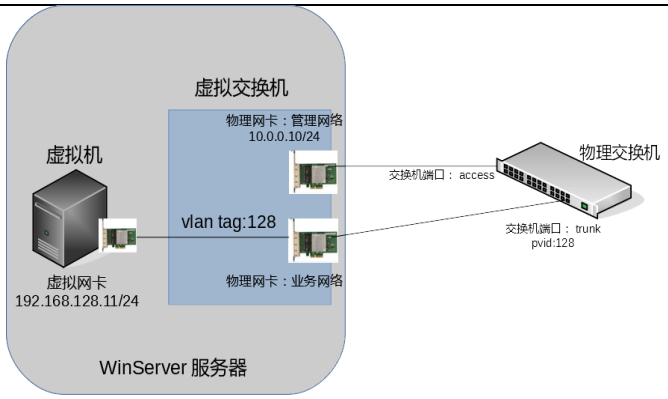
- 成本更低：硬件交换机以昂贵成本才能达到的性能用 OVS 即可达到

与上一代的 Linux Bridge 虚拟交换技术相比，其优势在于：

- 隧道隔离模型：支持更前延和扩展性更好的 VXLAN 等模型；
- SDN 的结合：支持 openflow，支持插件或 SDN 控制器（OpenDayLight 等）集成
- Qos 配置：按需配置虚拟机端口所需网络速率和带宽；
- 流量监控：支持配置 Netflow、sFlow 的功能配置，实现网络流量监控和分析；
- 端口镜像：OVS 可以配置各种 span（SPAN, RSPAN, ERSPAN），把端口的数据包镜像复制到指定端口，再通过 tcpdump 抓包分析；
- 数据处理优化：支持 dpdk，ebpf 等高级功能

CNware 的实现中包括普通、特殊的两种虚拟交换机，区别如下：

- 普通的虚拟交换机作为 VLAN 网桥转发虚拟机流量，包括系统生成的 vSwitch0 和手动生成的虚拟交换机；
- 特殊的名为 br-int 的虚拟交换机由系统启用 SDN 模块时自动生成，且只作为隧道网络解封包处理的网桥，仅允许 VPC 网络内的虚拟机使用。



CNware虚拟交换机连接示意

网卡绑定

网卡绑定是通过多张物理网卡绑定为一张逻辑网卡，从而提高网络可靠性、增加链路带宽、实现负载均衡的一项配置能力。

网络绑定可带来以下价值：

- 提升容量：虚拟化交换机通过链接多个物理网卡，增加业务带宽容量支撑更高的访问能力；
- 网络冗余：当其中的某个网卡发生故障，系统自动将流量切换至于可用的网卡，避免业务网络中断；
- 负载均衡：多个网卡能够均衡分担流量传输，减少冲突拥堵。

CNware WinSphere 支持动态或静态的链路聚合模式，应用过程中可选择主备、平衡负载、高级平衡负载等方式。

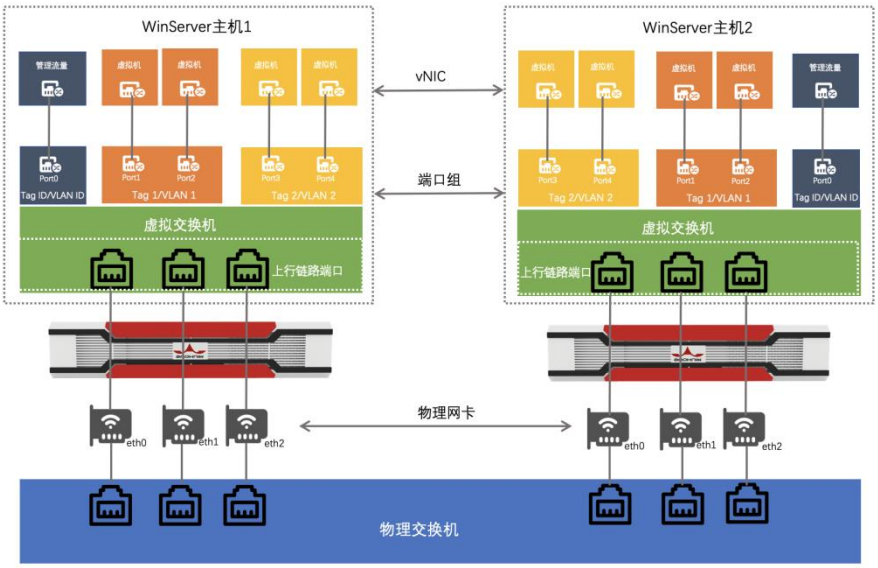
虚拟端口面板

在 OVS 中，端口的概念与物理交换机的端口概念类似，每个端口都归属于某个虚拟交换机。CNware 提供了图形化的虚拟端口面板，实时监控虚拟交换机的虚拟端口、流量统计

等信息，能够查询虚拟端口对应的虚拟机名称、MAC 地址、接收字节数、接收错包数、发送字节数、发送错包数、接收报文数、发送报文数等信息，方便管理员观察业务的流量分布情况，辅助验证虚拟交换机的转发性能。

3.4.2. 标准虚拟交换机

标准虚拟交换机是宿主机上运行的一种实现方式，作用范围仅在宿主机本地。上行链路（物理网卡）连接到标准交换机，提供外部网络连接条件；虚拟机以虚拟端口的方式逻辑连接到标准虚拟交换机，由标准虚拟机完成流量的转发处理。

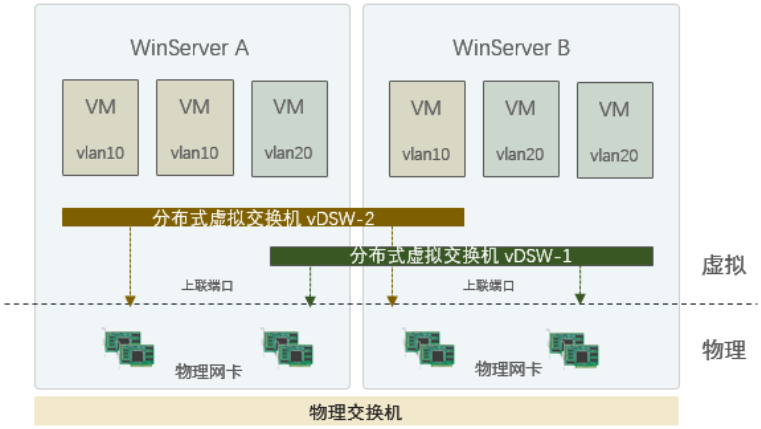


3.4.3. 分布式虚拟交换机

相对标准虚拟交换机而言，分布式虚拟交换机是一种更加便捷的虚拟网络管理方式，作用范围能够支持跨集群甚至整个虚拟数据中心。在分布式虚拟交换机中，虚拟机的内部流量转发仍然依赖物理上行链路的连通，但通过逻辑集中对多台宿主机的虚拟交换机的物理端口和虚拟端口的管理（端口应用是分布式的），实现灵活的网络连接和简便、统一的网络配

置，并保证虚拟机在宿主机之间迁移时网络连接的一致性。

需要指出的是，本小节的分布式虚拟交换机概念范围仅作用于基于 VLAN 的经典网络；在 VPC 网络中，也有分布式虚拟交换机的实现方式，CNware 将其封装为虚拟私有网络，见下文。



CNware分布式虚拟交换机连接示意

3.4.4. 端口组

顾名思义，端口组是作用于虚拟端口（即虚拟机网卡）的一组策略，关联于标准虚拟交换机、分布式虚拟交换机，实现 VLAN、优先级、Qos、安全拦截等网络调度能力，能够区分通过虚拟交换机的流量类型，以及充当通信与安全策略的边界。

网络 Qos

Qos 的保证对容量有限且积极争抢的网络资源是十分重要的，管控包括出入方向的平均带宽、突发缓冲、IP 广播包、ARP 广播包等，能够很好解决网络延迟和阻塞问题，保障核心应用的网络资源质量。

DHCP 报文拦截

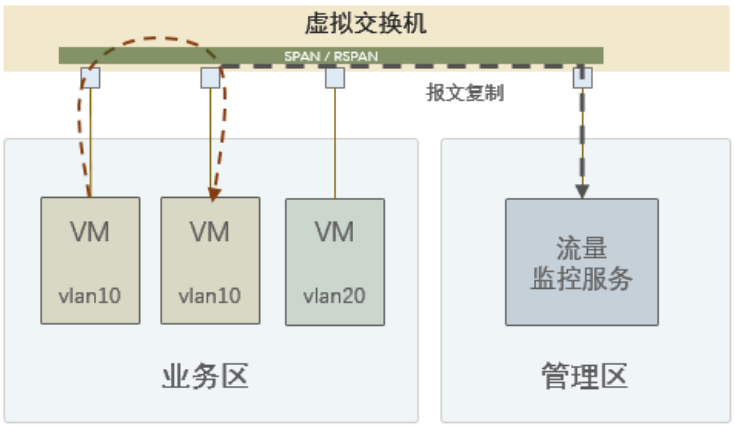
即对虚拟机网卡所在的端口进行安全报文拦截，抑制无意或恶意的 DHCP 报文对外广播，影响业务正常使用。

3.4.5. 端口镜像

端口镜像功能即将一个或多个源端口的数据流量转发到某一个指定端口，实现对网络的监听：例如观察网络流量是否有访问安全异常、运维故障排查等场景。

镜像流量抓取

镜像流量抓取可指定数据包长度、源端口、目标端口等，其中源端口可指定流量方向，包括输入、输出、输入/输出。



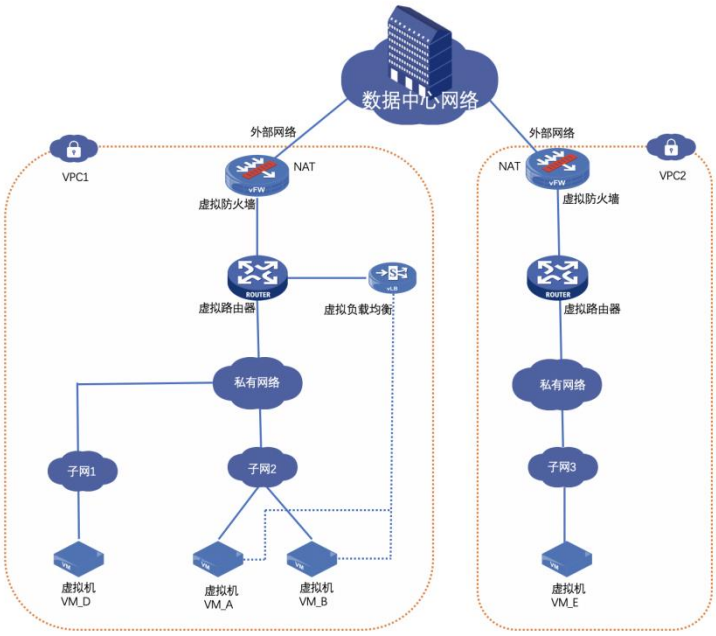
端口镜像原理

3.4.6. VPC 网络

VPC（(Virtual Private Cloud）是一个逻辑上隔离的专有云，是实现租户网络隔离的基础，由各类虚拟网络元素和虚拟机实例构成。与基于传统 VLAN 技术、二层转发的虚拟网络实现不同，VPC 的实现包含基于隧道封装技术（geneve、vxlan 等）的数据转发平面和集中控制平面，依赖于强大的 SDN（Soft Define Network，软件定义网络）控制器。

基于 openvSwitch 项目、控制与转发分离的理念，CNware 打造了轻量级 SDN 网络操作系统——WinFabric。数据转发平面由计算节点的 openvSwitch 虚拟交换机完成，但转发策略及网络元素的抽象编排由 SDN 控制器下发基于 openflow 的流表规则实现；SDN 控制平面、开放插件 provider、控制器高可用全部深度自研，以松耦合的方式和极低的资源消耗集成到虚拟化管理平台，实现云网联动。

VPC 实例即由 WinFabric 系统实现，核心是构建了完全隔离、基于隧道封装构建的网络地址空间。VPC 的私有网络承载在经典网络之上，如果将经典 VLAN 网络按第 2 层来理解，VPC 私有网络可以看作是 2.5 层网络。此外，VPC 内还包括虚拟路由器、虚拟负载均衡器、安全组等能够在物理数据中心网络拓扑对应的元素。



VPC 网络元素与关系

3.4.7. IPv4/IPv6 双栈

IPv6 是英文“Internet Protocol Version 6”（互联网协议第 6 版）的缩写，是互联网工程任务组（IETF）设计的用于替代 IPv4 的下一代 IP 协议，其地址数量号称可以为全世界的每

一粒沙子编上一个地址。由于 IPv4 最大的问题在于网络地址资源不足，严重制约了互联网的应用和发展。IPv6 的使用，不仅能解决网络地址资源数量的问题，而且也解决了多种接入设备连入互联网的障碍。

IPv4、IPv6 双栈支持是现今平台的基础能力要求。IPv6 的改造和支持行业呼声愈高，互联网头部厂商几乎已完成数据中心全网 IPv6 升级，推动了 IPv6 下一代互联网全面部署和大规模商用。CNware WinSphere 覆盖多层级的 IPv6 网络类型支持，包括平台管理网络、虚拟机业务网络、虚拟交换机、共享文件系统等，帮助企业无缝部署到升级的网络基础架构，支持更广、更深的全网应用改造。

3.5.异构管理

3. 5. 1. VMware

由于 IT 发展的历史原因，企业架构中仍然存在大量的 VMware 虚拟化环境；从信息化自主的角度，VMware 不但绑定了 X86 设备，还绑架了用户的信息安全。尽管 VMware 的国产化替代是必然趋势，但推进的过程有来自应用、基础架构、企业自身的发展等多方面的压力，因此短期甚至中长期内仍然需要提供 VMware 虚拟化环境的统一管理能力，驱动后续的逐步分阶段替换。

3.5.1.1. VMware vSphere

CNware 通过 VMware vSphere 的 API 接口，实现 vSphere 6.0/6.5//6.7/7.0 等主流版本的统一资源监控与管理，为资源融合、服务融合、跨云编排和盘活利旧资产提供基础。

CNware 与 VMware 在客户环境内共存，提供一致的管理运维体验，覆盖以下特性：

- DataCenter 下的集群自动发现、资源总量及利用率实时侦测
- 宿主机自动发现、状态同步、资源总量及使用率侦测、性能实时监控告警；
- 虚拟机自动发现、操作系统版本同步、资源配置实时更新、性能实时监控告警等；支持虚拟机创建、IP 设置、控制台、电源管理、一键迁移；
- 存储池（datastore）及类型自动发现、资源总量及利用率侦测、虚拟磁盘更新及统计；
- 虚拟网络类型（包括标准交换机、分布式交换机）自动发现、支持端口组及 VLAN ID 同步；

3.5.2. 物理裸机管理

物理裸机作为专享型的计算资源适用于特殊要求的场景，与虚拟化主机是相辅相成的云计算基础设施。从传统服务器管理的角度，应当包括服务器从到货验收、上架、上线服务、下架等生命周期管理过程，因此不可避免引入资产的属性；从管理的便捷性角度，应当提供远程的 BMC 带外、电源控制、硬件级的监控；从云化的角度，应当引入一定程度的自动化和弹性，实现虚拟化或操作系统的自动化部署。

CNware 提供三类接入模型：

- 本地管理平台：实现资产化视图管理、电源管理、硬件基础监控等基础能力。
- Cobbler 平台：一项能够满足 Linux 服务器快速网络安装的开源服务，整体非常小巧轻便，支持通过网络启动(PXE)的方式来快速安装、重装服务器操作系统；
- 硬件部署平台：低成本实现的裸机 OS 自动化部署，以及更为丰富的硬件发现与监控能力，同时也支持对接开源的 cobbler 平台。

3.5.2.1. 设备监控

CNware 实现服务器“纳管即监控”。针对绝大部分的服务器设备，基于标准接口，系统能够实时同步各类硬件状态信息，包括硬件温度、硬件功耗、风扇转速、电源功率等，辅助管理员掌握硬件健康状态。

特别的，系统针对鲲鹏服务器（Huawei TaiShan 系列）提供更丰富的特性支持。

- 远程虚拟控制台：得益于鲲鹏服务器的固件支持，系统能够一键打开鲲鹏服务器的远程虚拟控制台，无需再繁琐地登录至服务器 BMC 带外管理界面。

相比其他服务器，系统能够获取更为详尽的硬件信息提升管理体验：

鲲鹏标准化硬件信息

3.5.2.2. 自动化部署

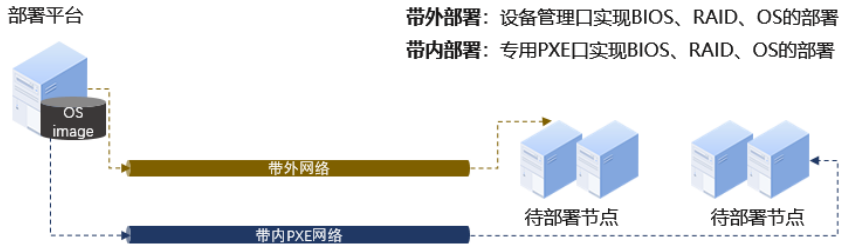
操作系统的部署是比较重复性的繁琐工作，如果通过尽可能标准化、自动化、快速的方式去完成，从而释放运维人员去从事更多有价值的事情，不能不说是绝佳的方案。尽管有部分平台能够实现这一工作，但实际落地中又会遇到很多问题：

- 网络合规要求：传统的平台依赖 PXE+DHCP 技术实现装机，但 DHCP 很容易引起广播风暴，在生产环境中也不太允许开启 DHCP 报文的转发；同时，PXE 的方式

是容易造出网络带宽的占用的，需要提前规划；如果网络跨区域部署，又要考虑到网络的打通和协同；

- 硬件的配置：这里指的是 BIOS、RAID 等硬件的配置，如果这里是个管辖的盲点，那么有可能是不符合安全合规要求的；另外，对资源的申请者来说不一定擅长或熟悉硬件方面的配置；
- 操作系统的兼容：按用户实际使用的操作系统清单，一般来说都会数十个类型和版本的操作系统，形成一大挑战；
- 并发装机：当规模变得庞大，装机有可能变成日常高频的操作，批量并发部署的效率影响业务交付的进度。

CNware 内置的硬件部署平台提供用户一套适用于带内（网管与业务数据同一生产网）或带外（网管与业务数据分离），适应多种网络区域、多种服务器硬件（含 X86、鲲鹏/飞腾等），多种操作系统类型的自动化、批量化、标准化的装机解决方案，同时支持 RAID/BIOS 的批量配置和升级，单个管理平台并发装机数不低于 200 台。



设备自动化部署方式

3.6.运维管理

3. 6. 1.　监控告警

告警管理功能用于统计和查看管理员需要关注的告警信息。目前，CNware 统计的告警类型包括：集群资源告警、宿主机资源告警、虚拟机资源告警、存储池资源告警、故障告警和运维告警。

3.6.1.1.　实时告警

当系统运行出现异常或与期望的状态不符时，系统会立即产生告警，令运维人员及时关注平台的稳定性、性能等方面的风险，修正系统的异常错误，避免造成更大范围的影响。

告警处理

系统产生的告警事件会在主机池、集群、宿主机、虚拟机等层级汇报，最终汇入实时告警列表。例如，非管理平台操作的虚拟机创建、启停、重启等事件，虚拟化底层能够通过管道上报告警模块，管理员在实时列表及时查询及处理此类告警事件。

告警事件包含告警级别、确认状态、来源、类型、信息详情、时间等信息。管理员可以执行确认、删除、导出、一键清理、刷新时间间隔等操作。

告警数据管理

实现告警数据收敛，突出告警重点，避免重复消息对运维人员产生干扰。

同时支持告警数据保存期（按天）设置，防止告警数据过多降低系统性能；系统提供告警转储 NFS 路径设置，将告警数据备份到远程介质存储归档。

SNMP 配置

简单网络管理协议（SNMP）是一项设计于网络发现、管理和解决问题的标准协议，第三方监控与告警平台通过该协议能够实现虚拟化平台的监控发现与告警上报。虚拟化平台提供详细的 SNMP MIB 库，随附于产品配套发布文档。

3.6.1.2. 告警策略

3.6.1.3. 阈值配置

告警阈值指的是触发告警的最低值、告警级别或触发告警的时间间隔，包含主机、虚拟机、集群、故障等维度。指标覆盖：

- 计算：如 CPU 利用率、内存利用率、
- IO：磁盘利用率、磁盘 IOPS、磁盘吞吐量、网络吞吐量
- 故障：主机故障、网卡故障、FC 链路故障、虚拟机故障等；

告警阈值及相应包含紧急、重要、次要、提示等 4 类级别。

实现告警策略配置，支持配置告警生效范围，可针对集群、宿主机、虚拟机、存储池自定义配置告警策略。

3.6.1.4. 告警通知

系统支持多种方式界面、邮件、短信（阿里云短信网关）、钉钉和企业微信等方式发送管理员订阅的告警通知，及时告知管理员并辅助解决告警故障，保障业务尽快恢复安全运行的状态。告警通知内容包括平台告警、硬件告警、巡检报告等内容，管理员可以配置仅关注的告警级别，避免收取信息泛滥轰炸的过度告警现象。

3.6.2. 日志管理

3.6.2.1. 任务日志

展示主机下的所有调度任务，包含任务名称、目标名称、任务描述、任务进度、任务状态、创建人、创建人 IP、创建时间、结束时间、任务日志详情。通过调度任务日志，管理员可关注任务的执行状态和结果，便于跟踪、排错等。

3.6.2.2. 日志收集

平台提供方便的一键式日志收集和下载服务，帮助用户集中、快速地获取管理节点和计算节点的日志，避免多点登录拷贝日志的复杂性。

3.6.2.3. Syslog 同步

Syslog 协议广泛用于系统日志的记录和集中采集，实现日志的审计和解析。企业架构内通常架设 syslog 服务器，虚拟化平台支持配置 syslog 服务器地址及端口等，实现日志的传递和同步。

3.6.3. 一键巡检

“巡检”提供对平台核心组件、资源告警等关键指标作全面的筛查分析，一键即可执行“对象-计划-报告-分析”的完整任务，确保虚拟化环境持续、健康、稳定地运行，从而提供高水平的业务保护服务。系统支持巡检对象可勾选、自定义异常检测阈值及周期计划配置，并在汇总的巡检结果中提供修复建议。

3.6.4. 回收站

回收站提供了一种温和的资源处理方式。即虚拟机（及磁盘）被删除时，被放入回收站缓冲保护而不是立即从系统中移除，这为用户数据的安全性提供多一道的保障。

回收策略

资源流转至回收站包含多种处理策略：

- 冻结并移入回收站
- 关闭并移入回收站。
- 设置虚拟机回收保存期，超出保存期限会立即触发数据销毁动作。默认策略为永久保留，请谨慎配置，云宏不承担任何由策略引发的数据损失责任。

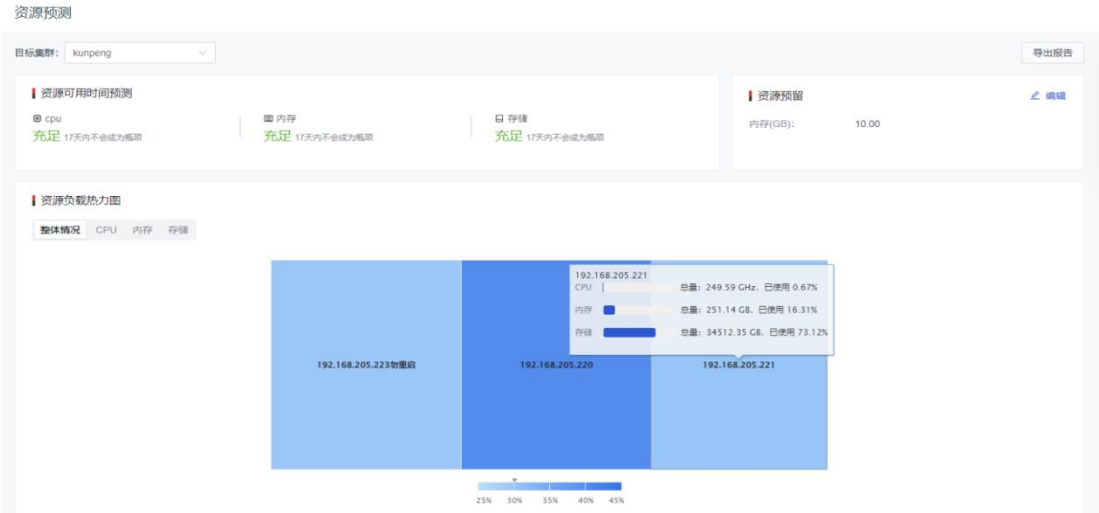
资源回收

资源被删除后包含的状态和处理方式：

- 回收站状态：资源流转到回收站，但系统会保持其数据状态和资源占用并不会有实质的操作。
- 回退还原：从回收站可以立即批量还原资源与其状态，业务重新使用；
- 彻底销毁：确定资源及其数据失去价值，管理员可以利用系统自动回收策略或手动执行批量销毁。

3.6.5. 资源预测

资源预测从计算资源及存储资源层面对集群进行监控，主要根据集群历史的 CPU、内存与数据空间使用情况，用户可以获得未来这些资源使用量的预测。以便用户有效地避免或减少资源不足所造成的影响，能对资源情况快速且准确地预测从而协助用户提前进行硬件规划、资源腾挪，保障流量洪峰下的扩容诉求，确保业务稳态运行。



3.6.6. 用户管理

用户对应现实的身份代表，任何身份代表须通过指定的角色访问虚拟化平台。用户管理涉及用户信息的管理与行为管理。

基本管理

用户的增删改查，冻结/解冻、角色关联、密码重置、限制登录源等管理能力。

AD/LADP 支持

系统支持基于 LDAP/LDAPS 实现域架构与用户的同步，在企业架构在实现一致的访问凭证。

登录限制

登录限制指设置限制用户登录的时间段、IP 地址，达到安全访问平台的目的；

权限与行为

系统记录用户的权限与行为日志，包括上次登录成功时间、上次成功登录 IP、上次登录失败时间、上次登录失败 IP、是否冻结、关联角色、操作（增加/修改/删除/重置密码）等。

3.6.7. 角色管理

系统默认预置 3 类角色类型：系统管理员、资源管理员、运维管理员。

支持通过角色克隆快速生成新的角色、批量为用户授予角色。角色权限支持灵活的细粒度控制，包括：

- 菜单权限：访问计算、安全、存储资源、存储池、运维、网络、系统等模块
- 操作权限：访问主机池、集群、虚拟机、异构、系统、模板、业务组、存储、安全、备份、运维等操作集
- 资源权限：访问的集群、宿主机、虚拟机范围

3.6.8. 补丁管理

系统为虚拟化平台专门制作和发布的补丁包，考虑到业务安全，补丁定义了多种影响级别：如 W0 无影响、W3 需重启服务。

管理员可上传到补丁库列表和执行补丁任务，集中维护升级底层虚拟化版本。

3.6.9. NTP 配置

Network Time Protocol（NTP，网络时间协议）是用于使计算机之间时间同步化的一种协议。NTP 的目的是在无序的 Internet 环境中提供精确和健壮的时间服务，依赖于准确而可靠的时钟源提供高精度的时间校正；计算机主机一般同多个时间服务器保持连接和容错设计，利用统计学的算法过滤来自不同服务器的时间，以选择最佳的路径和来源来校正主机时间。

整个网络保持时间准确是十分重要的，即使小小的时间误差也可能引起巨大的问题和后果，因此时间同步是管理的基本要求。管理平台与计算节点时间不一致会导致监控数据、日志服务出现问题；虚拟化系统支持最多指定 5 个 NTP 服务器。

3.7.可靠性

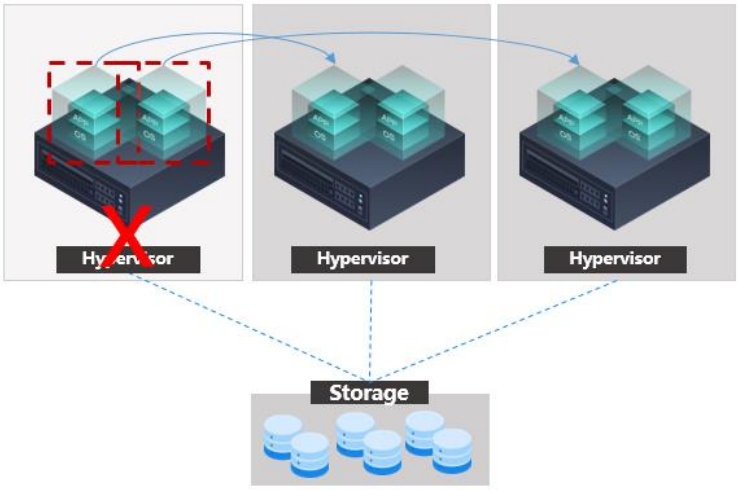
3.7.1. 计算高可靠

虚拟化是一个复杂的 IT 支撑环境，没有人能保证任何环节不会发生意外故障（计划外的故障）。虚拟化环境内部故障可能包括物理机本身的硬件损毁、虚拟化系统关键进程的崩溃，外部故障可能包括网络链路故障、存储链路故障等。这就要求，虚拟化系统必须提供成熟可靠的 HA 机制尽可能保障业务不中断或极短暂的中断。

集群 HA 高可用是虚拟化集群的高级特性，当主机发生计划外故障的时候，WinServer 支持在若干秒内（通常不会超过 1 分钟）自动的把故障主机上的虚拟机（位于共享存储）迁移至其它可用的主机上启动，由此实现保障业务连续性。

CNware 的集群 HA 机制主要结合 Pacemaker、Corosync、Stonith、IPMI 等多种技术实现。Pacemaker 是一个集群资源管理器，利用集群基础构件（OpenAIS 、heartbeat 或 corosync）提供的消息和成员管理能力来探测并从节点或资源级别的故障中恢复，以实现集群服务；Corosync 是集群管理套件中负责运行心跳检测的一部分，通常会组合使用并在传递信息的时候可以通过一个简单的配置文件来定义信息传递的方式和协议等，例如 Redhat 的 RHCS 集群套件就是基于 corosync 实现。Stonith 是 NodeFencing 的一种技术，能够运行在节点的守护进程响应远程或 HA 判断节点死亡后自动触发的命令，实现资源级别的隔离机制来防止节点抢占资源。IPMI 是一项独立于 BIOS 的智能管理接口，能够结合实现 Fence 机制确保节点下线，尽最大程度避免集群中“脑裂”（split-brain）现象的出现。

集群的 HA 策略需要建立管理网络心跳，这依赖于交换机需要支持组播报文的转发。同时，仲裁方式可选 IPMI 或 SBD 方式。选择 IPMI 方式，要求虚拟化管理节点与所有计算节点的 IPMI 带外地址连通，这种方式在计算节点故障时能够很好的完成断电隔离；如选择 SBD 方式，则需指定一个共享存储池作为存储心跳，计算节点需要同时检测网络心跳与存储心跳来判断自我故障并实现自我隔离。



集群 HA 原理

3.7.2. 存储高可靠

3.7.2.1. 多路径

存储多路径是指存储设备通过一条或多条链路与主机连接，通过存储设备的控制器控制数据流的路径，实现数据流的负荷分担，保证存储设备与主机连接的可靠性。WinServer 支持存储多路径配置，通过容错、I/O 流量负载均衡甚至更细粒度的 I/O 调度策略调校，实现更高的可用性和性能。

虚拟化宿主机提供存储多路径服务的配置管理，内置来源于 HUAWEI、MacroSAN、TOYOU 等主流国产存储厂商的参数标准模板，模板内囊括如 path_checker、path_grouping_policy、path_selector、no_path_retry、failback 等策略项，默认情况下，multipath 已经支持大部分常见的存储型号，但是针对特定存储或者多个存储设备，默认规则并不能完全通用。因此额外提供在线编辑能力，允许按需定制修改多路径规则。

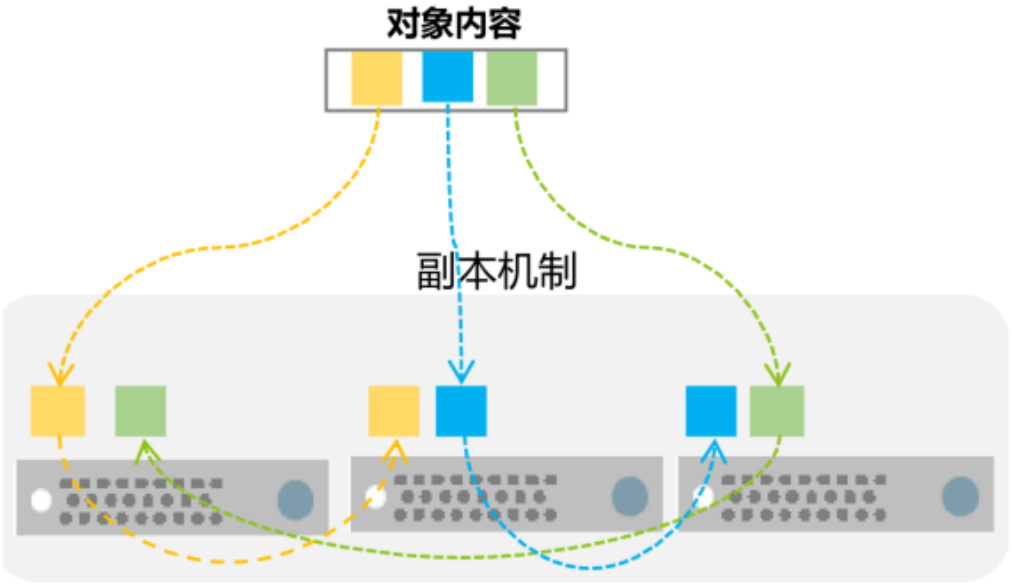
3.7.2.2. 数据冗余

存储系统中，数据先写入硬盘中。随着硬盘数量增多，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。为了保障数据的安全性，系统支持多副本和纠删码两种数据冗余技术。

多副本

在副本冗余策略下，各副本数据最终写入不同的硬盘。采用 DHT 分布式哈希算法，保证各个硬盘的容量使用均衡。副本策略支持服务器、机架级别故障域，应用层不感知副本存在什么硬盘上。

例如客户端写入一个对象，此对象按 2 副本写入。默认服务器级别时，两个硬盘属于两个不同的服务器节点。 此时即使一台服务器硬盘损坏，数据也不会丢失。



当某个节点故障超过一定时间(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，即利用现有的副本恢复出原有数据，保证数据的可靠性。

纠删码

副本冗余技术下，硬盘使用率不高，除了简单的副本外，企业开始尝试将纠删码技术应用于数据存储中。纠删码(Erasure Coding， EC)最早是通信过程中的编码校验容错技术，其基本原理就是把传输的信号分段加入一定的校验再让各段间发生相互关联，即使在传输过程中丢失部分信号，接收端仍然能通过算法将完整的信息计算出来。人们自然也想到使用纠删码来增加存储系统中数据的可靠性。

3.7.3. 备份容灾

3.7.3.1. 虚拟机备份

针对企业关键的业务和信息数据，尤其在国产化架构的背景下，必须建立起有效的数据保护体系。例如，面对人为误操作、病毒、网络攻击、软硬件故障、自然灾害等意外事件时，极易造成数据丢失，甚至将严重影响到业务的正常开展，带来巨大的经济损失以及负面影响。另外，重大故障发生时，需要重新安装操作系统、所有应用程序，然后才能恢复数据，而重新恢复应用耗时长，因此操作系统也需要进行保护。

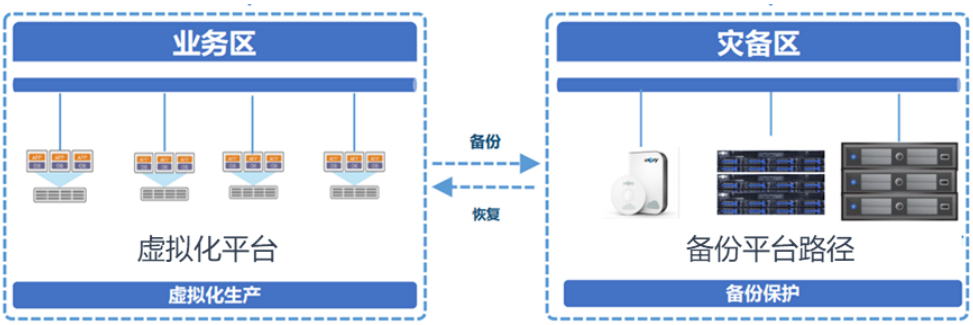
“容灾备份”可以帮助您保护数据免受意外的损失；随着企业逐步引入云计算、虚拟化技术，同样需要提高对灾难事件的抵御能力。CNware 自带虚拟机备份方案，在无须第三方工具的支持下，配合快照技术实现指定虚拟机基于策略的全量备份、增量备份；当虚拟机数据丢失或故障时，可通过备份的数据尽快恢复业务。数据备份是容灾的基础，如果需要更加专业可靠的数据灾备保护方案，例如 CDP 持续保护、异地灾备保护、应急恢复等，CNware 也支持与第三方专业灾备平台无缝整合。

备份虚拟机

虚拟机支持 FTP、SCP 等传输方式传输到指定的本地或远端目录，备份方式包括：

- 即时备份：立即创建并执行备份任务
- 计划备份：定义备份任务及备份策略，包括备份周期（按天/周/月）、执行时间、重复次数等。

  * 上述默认为整机备份，仅基于winstore支持备份单块虚拟磁盘。



虚拟机备份与恢复

备份类型

备份类型包括全量备份、增量备份：

- 全量备份：完全备份所有系统数据，数据完整直观，缺点是速度冗长、占据大量存储空间；
- 增量备份：仅备份上一次备份后的增加、改动部分的数据，优势是备份窗口短、大幅节约存储空间，缺点是数据恢复相对较慢。

备份保留规则

每次备份均产生一次备份记录和文件，持续备份会增加存储成本。因此系统允许设置备份保留规则（保留天数或个数），平衡备份有效性和存储成本。

备份文件管理

每次执行备份任务会生成记录到备份文件列表，按执行的时间顺序归档，管理员可以查阅备份记录，在必要时恢复至到虚拟机磁盘，或者新建一个虚拟机。

备份池管理

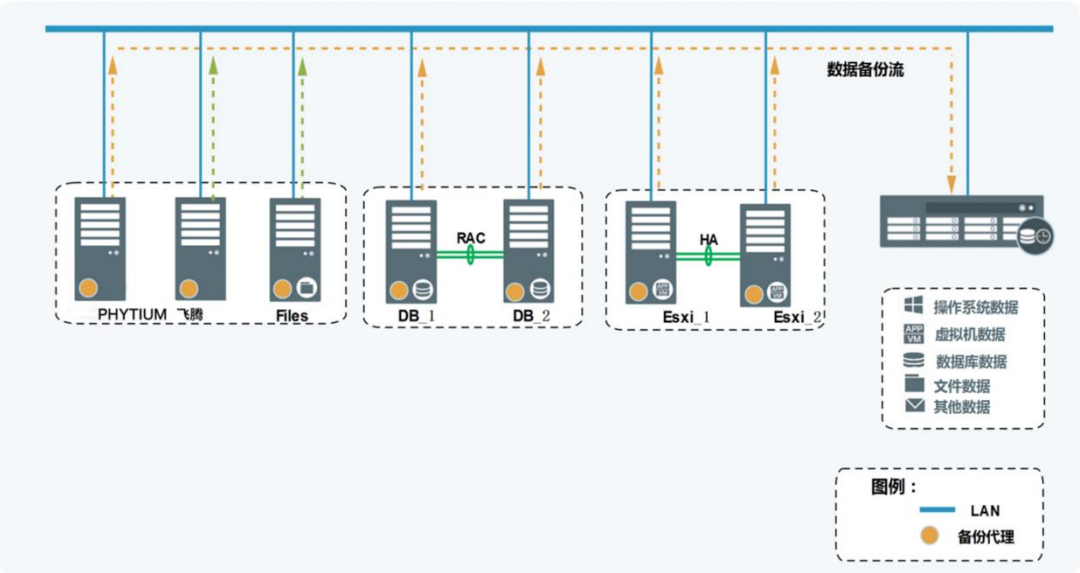
立足虚拟化数据中心全局，支持多集群、多宿主机、多虚拟机、多磁盘等维度，实现批量备份任务管理，包括批量任务策略、备份状态、下一次备份时间、备份数量、任务日志详情等，提供备份任务执行的月度统计盘点。

第三方备份整合

第三方备份平台提供无代理、有代理两种集成方式。无代理方式通过对接虚拟化接口，无需在虚拟机系统部署客户端代理，结合虚拟化的快照和卷拷贝技术实现虚拟机相对粗粒度的定期备份；有代理的方式通过在虚拟机操作系统部署保护代理，基于系统的数据块级别的变化拷贝实现 CDP 实时备份保护，同时支持更多的快速恢复、压缩传输、断点续传、数据重删、备份文件自动校验等高级特性。

方案具备如下价值：

- 全方位实时保护，兼容性强大，包括各版本的操作系统及数据库；
- 安全可信，支持备份副本使用军密、商密、国密等高位加密技术，确保网络传输和存储安全；
- 一致性保护，通过自动校验机制确保备份副本数据的完整；
- 支持多种备份方式，包括 LanBase、LanFree 等。



备份保护方案

3.8. 安全性

安全是生产的第一红线。在云计算技术的助力下，企业的业务市场蓬勃发展，但规模化发展的光环下也随之带来更多的信息安全隐忧，既有敌对势力对我国实施“陆、海、空、天、网”全方位、立体式、多维度、不断翻新的信息监控和情报窃取手段的外因，也有企业内部系统网络肌体薄弱、安全保障水平低、安全管理制度不足的内因。国家尤为重视网络安全，陆续出台《中华人民共和国网络安全法》、《网络安全等级保护管理条例》等法律法规，指导实际的信息化系统安全建设。CNware 作为企业信息化系统的关键设施，积极强化自身系统的技术防护并灵活融入等级保护安全体系，坚持统一性、多重保护、适应性的原则，遵循 PPDR 模型的思想综合引入多种策略和方法化被动为主动，与安全技术、安全运营、安全制度协同形成牢不可破的高安全云基础设施体系。

3.8.1. 身份鉴别和管理

3.8.1.1. 用户密码策略

密码复杂性要求提供若干组合选择，校验用户的密码复杂度是否满足指定要求：

- 必须混合使用字母和数字
- 必须混合使用字母、数字、特殊字符
- 必须混合使用大写字母、小写字母、数字、特殊字符。

密码最小长度要求用户提供的密码必须满足最小位数，否则认定为系统无法接受、无效的弱密码；

密码有效期设置了定期失效的属性，密码超过设定的有效时限（天）后，密码自动失效并强制要求修改。

3.8.1.2. 双因子身份认证

双因子身份认证的保护机制有效性逐渐受企业重视和接纳，单一的密码形式似乎容易被击破或仿冒。双因子认证提出在密码之外，使用另一种非密码形式的验证因素来确认使用者的身份。CNware 采取的是基于数字签名技术的 x509 证书+密码的组合认证方式。

数字签名技术大多基于哈希摘要和非对称密钥加密体制来实现。如果签名者想要对某个文件进行数字签名，他必须首先从可信的第三方机构(数字证书认证中心 CA)取得私钥和公钥。实现数字签名的流程是，使用私钥对数据进行签名，输出一段特定长度的数字签名（指纹）。通过使用对应的公钥、原始数据、数字签名进行运算，可以校验数据是否被篡改或者发行者身份的合法性。

X.509 是密码学里公钥证书的格式标准。X.509 证书已应用在包括 TLS/SSL 在内的众多 Intenet 协议里(如：当前 HTTPS 使用 X.509 V3 版)。同时它也用在很多非在线应用场景里，比如电子签名服务。X.509 证书里含有公钥、身份信息(比如网络主机名，组织的名称或个体名称等)和签名信息（可以是证书签发机构 CA 的签名，也可以是自签名）。对于一份经由可信的证书签发机构签名或者可以通过其它方式验证的证书，证书的拥有者就可以用证书及相应的私钥来创建安全的通信，对文档进行数字签名。

3.8.2. 访问控制和权限

3.8.2.1. 登录策略

- 在线超时时长：如用户的会话一直无操作，超过设定时长后，系统自动登出；防止忘记退出账号时，系统长时间保持在登录状态，被进行恶意操作；
- 连续登录失败次数：即同一会话最多能登录失败的次数，超过次数后，会话将被冻结；
- 系统登录失败：增加动态图形码验证，防止恶意破解；
- 连续登录失败锁定时长：同一会话超过限制的登录失败次数后，用户会被冻结若干分钟，解冻前无法登录系统；
- 单一用户单一登录：不允许同一个账号多客户端登陆，如若超出限制则已登录账号被强制下线。
- 单一用户最大连接数：配置单个用户的最大会话连接数（上限 100）后，若登录会话数超过设定数，最先登录的将会被踢掉。
- 登录限制：限制用户登录 IP 与登录时间，非指定时间范围内或使用非指定 IP 地址的用户终端将无法登录系统。

3.8.2.2. 远程维护

当业务系统发生故障时，如果能够立即与技术维护人员取得一对一的连接，由技术专家对系统进行远程诊断和维护，这对于及时排除系统故障和隐患，节省系统维护费用，保证系统长期稳定、可靠地运行具有重要意义。

远程维护高效、便捷的同时，伴随着的安全风险也日益突出。为了保障远程维护的安全性，系统支持严格的访问控制，主要功能如下：

- 登录密钥：虚拟机操作系统支持 SSH 协议登录前提下，支持虚拟机挂载密钥对，实现通过密钥方式免密登录虚拟机操作系统。
- 网络地址：主机高级设置中支持配置防火墙，限制远程连接至主机的网络地址。非网络地址白名单中的客户端，禁址访问主机。

3.8.2.3. 三员管理

三员管理是一种权限管理策略，系统的管理权限一分为三：系统管理员、安全管理员、安全审计员，增加系统安全性。默认情况下，该功能不启用，系统超级管理员具有全部权限。开启后，权限分化如下：

- 系统管理员拥有除“日志管理”、“权限管理”外的全部权限
- 安全管理员具有“安全权限管理”的权限
- 安全审计员具有“日志管理”的权限

3.8.3. 数据传输保护

3.8.3.1. DDoS 防护

分布式拒绝服务(DDoS)攻击是通过大规模互联网流量淹没目标服务器或其周边基础设施,以破坏目标服务器、服务或网络正常流量的恶意行为。总体而言，DDoS 攻击好比高速公路发生交通堵塞，妨碍常规车辆抵达预定目的地。

在虚拟化环境中往往通过子网隔离、安全组策略甚至部分 SDN 的流表策略以降低整体受影响面规模，但仍然无法很好针对局部（例如同一宿主机及虚拟交换机）的风险作反应。CNware 采取的策略是在虚拟化内核集成自研 DDoS 识别及防御模块，作用于 OVS 虚拟交换机的 port 及 datapath，能够针对 SYN Flood、HTTP Flood、Smurf 等多类主流供给类型作出应对，快速拦截过滤不合法流量，达到保护宿主机、业务虚拟机的重要目标。该项直接与虚拟化底层深度融合的 DDoS 防护技术为云宏首创，填补了国际空白。

3.8.3.2. IP/MAC 防欺诈

MAC 欺骗、IP 欺骗、ARP 欺骗是非常典型的攻击行为，其手法一般是伪造 IP 或 MAC 的合法身份来获取访问特权。在网络交换机等设备中，通常配置 IP Source Guard 结合 DHCP snooping 等安全策略实现有效的防护。

类似 VMware vSphere 产品也提供混杂模式下的“MAC 地址更改”、“伪传输”等特性，在云宏 CNware 虚拟化平台的地址管理策略中，支持基于 VLAN 的传统网络、VPC 网络双栈的“MAC 地址防火墙”，简单配置开启 IP/MAC 的静态绑定策略，拒绝非法修改的源 MAC，即达到安全保护的目的。

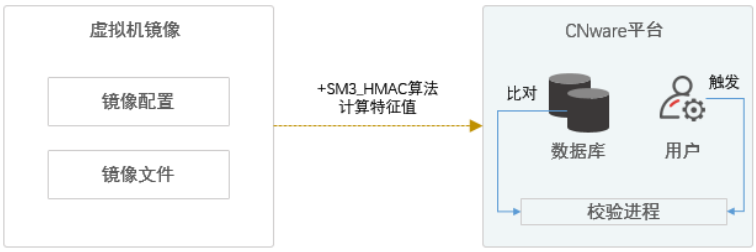
3.8.4. 数据保护

3.8.4.1. 虚拟机防护

镜像完整性校验

虚拟机镜像存在被有意或无意破坏的风险，如果被恶意篡改替换，可能会给用户带来数据泄露的巨大担忧。除了在镜像过程中提供基于国密算法的加密方法外，系统增加对镜像的完整性校验能力，镜像初次生成时系统会自动采用 SM3_HMAC 算法计算并记录完整性校验码 MAC 值，作为初始值在数据库中加密存储；用户通过镜像部署前，可以手动发起校验，系统再次计算镜像的完整校验码并与初始值比对，一旦比对不匹配则认定镜像异常。

除了镜像外，日志等重要数据也逐渐加入完整性校验框架，实现增强安全等级。



完整性校验原理

虚拟机杀毒

虚拟机杀毒是云宏与奇安信网神漏洞扫描系统合作的内置增值功能，实现产品级融合。网神漏洞扫描系统是一款综合性的漏洞扫描产品，其产品的设计以“黑客”攻击前期的漏洞扫描器为开发视角进行产品研发的，能够从操作系统、数据库、网络设备、防火墙、Web 系统、弱口令等多方位多视角对目标进行安全漏洞扫描检查；通过轻代理的方式向 CNware 平

台的虚拟机提供实时保护能力，兼容数十项主流操作系统版，支持快速扫描、全盘扫描、定时扫描、指定目录扫描等多项安全策略。该项能力能够应用于实战攻防演练、满足等保要求、分保要求、合规监管、与态势感知和 SOC 联动等平台数据联动等场景。

虚拟机安全分析

安全分析是杀毒漏扫基础之上提供的增值能力。奇安信网神安全引擎结合了 CVE、CNVE、Bugtraq、CNVD、CNNVD、CVSS 等多个权威漏洞检测标准，通过系统扫描+WEB 扫描+数据库检测+弱口令检测+基线核查多检测引擎合一体方式全方位检测后，自动将结果上报给安全引擎在线分析，形成各类如网页病毒、间谍软件、广告软件等恶意程序及受感染业务分布的分析报告；同时在病毒威胁视图特别生成病毒感染的趋势图、操作系统分布，以及专门罗列已隔离的病毒、木马列表及受威胁文件名称，体系化帮助用户深度洞察业务虚拟机的脆弱性、差异化对比。

虚拟机安全保护机制

虚拟机主要是由虚拟磁盘、配置描述文件组成，在某些高安全要求的场景下需要对虚拟机作出防止非法篡改窃取数据的保护机制。开启该功能后，对虚拟机的配置文件、磁盘文件一旦被以非法手段进行写入活动，虚拟机立即关闭且无法再运行。

虚拟机设备管控

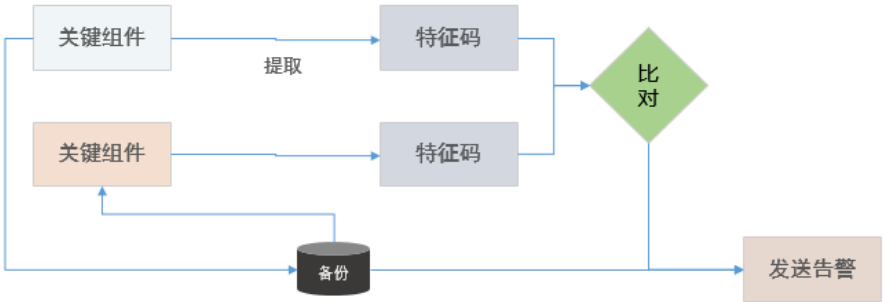
尽管数据窃取手段多样化，技术手段越来越高，但根据调研实际上很大比例的信息安全事件发生在初级的管理纰漏上。例如，采用可存储介质接触式窃取或破坏就是最为简单且常见的方式。

虚拟机的常见可移动存储介质包括 USB 设备、ISO、共享磁盘卷等设备类型，在虚拟化平台

能够将常见的设备类型设置管控方法，实现挂载权限的限制，避免系统设备发起的数据拷贝。

3.8.4.2. 主机防护

宿主机的配置文件关系虚拟化系统的稳定性，任何不正确的修改更新都可能导致极其严重的后果，甚至彻底宕机。系统提供宿主机关键配置文件防篡改的能力，持续监测文件的写入活动，一旦非法修改则立即告警。



防篡改安全设计

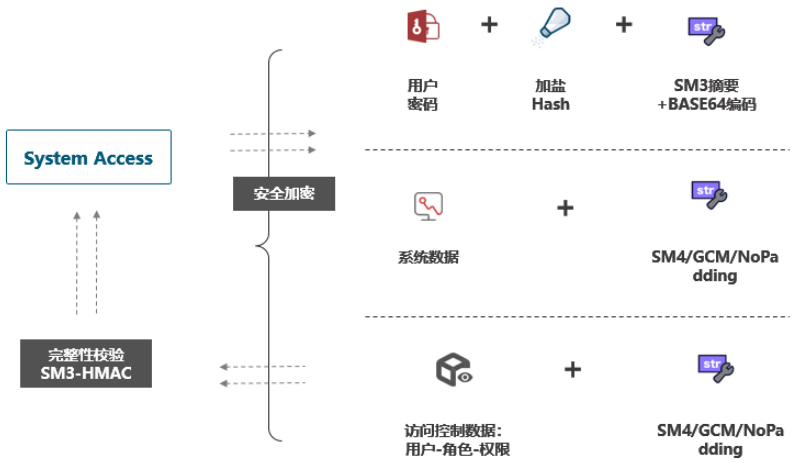
3.8.4.3. 安全加密

数据安全加密第一项体现在增强的方法上，用户的密码等敏感数据在系统中默认以非对称算法直接进行加密存储，理论上达到较高水平的保护。但是，如果无法完全避免数据脱库的情况，则仅直接对敏感数据加密仍然存在较大被反推的可能性。因而，数据“加盐（hash）”的方法能够极大地提高反推原始数据的困难，从而实现更高级别的加密保护。

数据安全加密第二项体现在增强的算法上，CNware 非常注重产品的安全性，取得了国家保密局颁发的高水平证书。在产品的安全设计中，尤其在日志、镜像等重要数据保护、传输安全、完整性校验等方面引入了国家商用密码算法替代 MD5 等通用的公开算法。

- SM2 为基于椭圆曲线密码的公钥密码算法标准，包含数字签名、密钥交换和公钥加密，用于替换 RSA/Diffie-Hellman/ECDSA/ECDH 等国际算法。安全性方面，
- 160 位的 SM2 密钥与 1024 位的 RSA 密钥安全性相同；
- 128 位安全加密需要 3,072 位 RSA 密钥，却只需要一个 256 位 SM2 密钥；
- 256 位安全加密需要 15,360 位 RSA 密钥，却只需要一个 512 位 SM2 密钥
- SM3 为密码哈希算法，用于替代 MD5/SHA-1/SHA-256 等国际算法。SM3 设计安全性为 128 比特，安全性与 256 比特椭圆曲线/SM2、SM4、AES-128 等同。
- SM4 为分组密码，用于替代 DES/AES 等国际算法。设计安全性等同于 AES-128（也有研究表明，稍弱于 AES-128）

以下为系统数据+国密算法的安全设计：



数据安全加密及算法

3.8.4.4. 密钥管理

密钥管理提供了对企业敏感资料实施有效监管和加密、解密的管理方法。密钥具有高度的保密性、完整性、可用性能力，能够满足企业多应用多业务的管理需求，符合监管与合规要求。

密钥管理拥有完整的生命周期管理能力，包括创建、修改、归档、计划删除等，轻松完成企业密钥管理计划的保护和执行。密钥生成方式默认支持系统生成（内置基于国密算法封装的 SDK）、自定义上传（企业自有密钥资料），同时支持与专业商密认证安全产品的 HSM 硬件模块对接生成。密钥管理能够与虚拟机、虚拟磁盘等无缝集成，实现例如磁盘加密、证书签名等核心数据的高性能海量加密场景。作为密钥管理的服务基础平台同样需要规划高可用性高容错的数据中心环境，企业与服务商也需共同承担起安全管理责任。

3.8.4.5. 商密环境安全

密码技术作为网络与信息安全保障的核心技术和基础支撑，在解决网络身份的真实性、数据机密性、数据完整性保护和行为抗抵赖等方面发挥着不可替代的作用。在信息化建设过程中，国家相关主管部门相继颁布了《网络安全等级保护基本要求》(简称“等保 2.0 基本要求”)、《国家政务信息化项目建设管理办法》、《中华人民共和国密码法》等政策法规，要求落实使用商用密码进行关键信息基础设施的保护，并开展商用密码应用安全性评估。

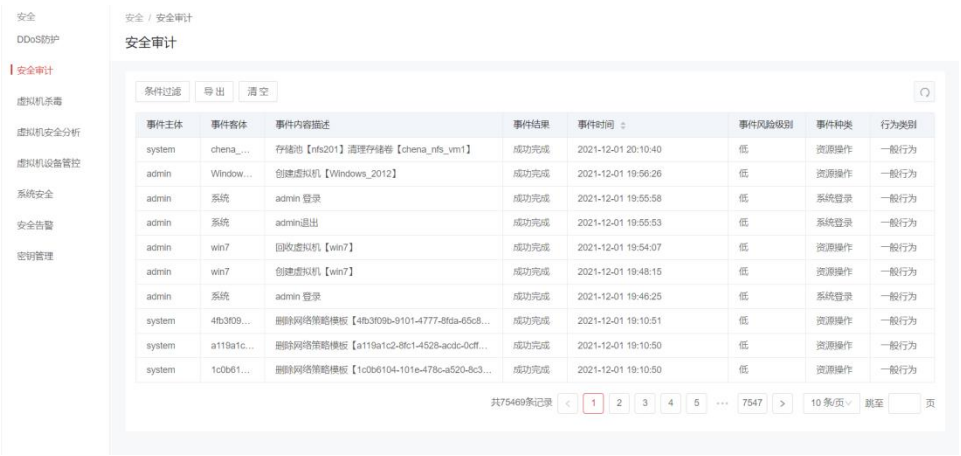
云宏作为虚拟化厂商，业内最早与得安、渔翁信息、卫士通等专业密码安全方案提供商取得联合验证，双方对《GMT 0054-2018 信息系统密码应用基本要求》进行了细致研读并合作通过项目密码类安全测评。CNware 本身支持基于 SM 算法对关键数据完成软加密，防止用户窃取分析数据；同时，支持对接密码服务商的硬件加密机、数字证书认证系统、密钥管

理系统、签名验签、时间戳、SSL VPN 等外部设备，完全符合商密测评规范要求。

密评环境对接

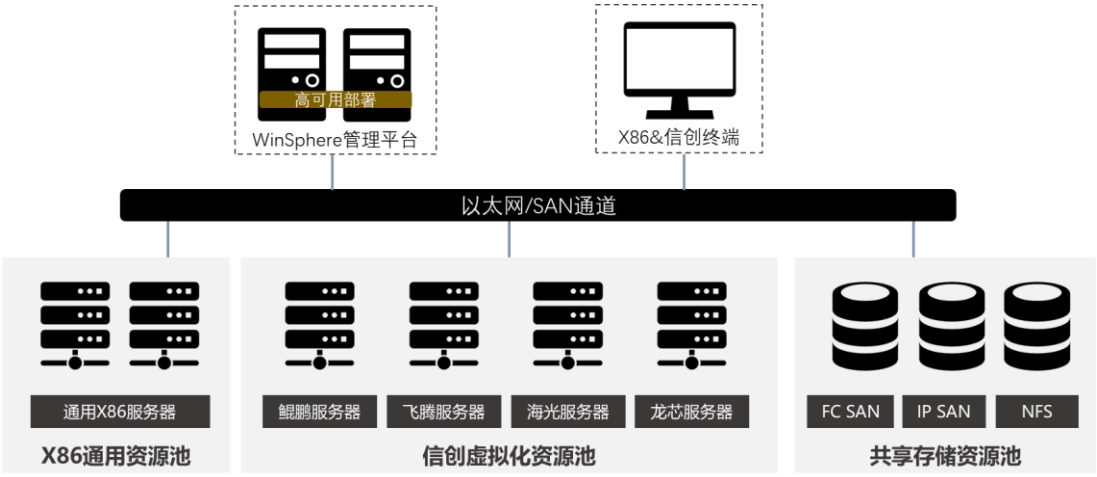
3.8.5. 日志审计与合规性

操作日志用于详细记录所有操作员在 CNware WinSphere 平台登录与操作事件，包括事件主体、事件客体、事件内容、事件结果、事件时间、风险级别、事件种类、行为类别等信息，为审计人员提供直接的告警、审计、追溯手段。



第4章 部署要求

4.1.部署架构



4.2.管理平台配置要求

项目	要求
处理器	1.处理器支持 64 位寻址技术 2.支持 X86 芯片、六大国产芯片 3.处理器至少具有两个物理核心
内存	1.最低 32GB，建议 64GB 或更高

存储	1.磁盘空间：本地存储（SATA、SCSI），RAID1 等级，建议使用 240GB 或更高容量的磁盘空间； 2.磁盘控制器：完全支持兼容性列表内的 SATA、SAS、RAID 及 HBA 卡
网络	1.千兆/万兆网卡环境，推荐两个专用的管理接口网卡
备注	支持裸金属或虚拟机方式部署，高可用要求至少两台

4.3. 节点配置要求

项目	要求
节点数	单个虚拟化集群不少于 3 个节点
处理器	1.支持 X86 芯片，例如 Intel Xeon 系列等，建议 Purley 4110 或以上； 2.支持六大国产芯片全系列型号，见兼容性列表
内存	1.至少 64GB，建议更高配置
系统盘	1.每节点两块系统盘，SAS 盘容量≥600GB 或 M.2 容量≥480GB，可做 RAID1/镜像，RAID1 配置为 Write-Through 模式
缓存盘	1.每节点至少一块（建议两块）读写混合型 SSD 用于缓存盘 2.SSD 缓存盘必须为 DC SSD（数据中心级），DWPD 必须≥1 3.单块 SSD 缓存盘容量不低于 240GB，建议 480GB 或以上 4.节点缓存盘与数据盘容量比建议是 1:10，最低不能低于 1:15

数据盘	1.每节点建议至少四块数据盘 2.数据盘容量建议不低于 1.2TB
阵列卡	1.阵列卡支持 JBOD/NON_RAID 模式 2.除系统盘外，缓存盘和数据盘必须运行在硬盘直通模式（NON_RAID/JBOD）
网络	1.至少两个专用的管理网口，支持千兆/万兆 2.至少两个业务接口，推荐万兆，非强物理隔离要求可复用管理接口 3.至少两个存储万兆网口，不可复用管理和业务网口
备注	裸金属架构部署，支持组建集群

4.4.存储资源要求

项目	要求
本地存储	1.支持 IDE、SATA、SCSI 和 SAS 控制器 2.符合 WinServer 硬件兼容列表要求 3.支持本地化挂载，不支持共享
NAS/NFS 共享存储	1.为计算节点规划专用的两个存储网络接口 2.为存储接口配置创建绑定网络、独立的子网 IP 3.为 NFS 存储网络配置冗余的交换机连接 4.存储接口分别连接到冗余交换机 5.提供路径完全挂载、读写权限

IP-SAN 存储	1.为计算节点规划至少 2 个专用的存储网络接口 2.存储接口分别连接到冗余交换机 3.为每个存储接口规划一个独立的存储子网 4.存储映射 iSCSI LUN 的块大小必须为 512 字节 5.采用多路径管理软件配置可用性
FC-SAN 存储	1.至少配置一台附合光纤通道协议的磁盘阵列柜产品，具体型号规格建议符合 WinServer 的兼容性认定； 2.存储设备要求配置至少两个控制器，每个控制器配置至少两个光纤端口； 3.存储网络要求配置两个型号及规格均相同的光纤通道交换机，交换机端口数量不小于 HBA 卡的端口数量与存储设备的端口数量之和； 4.计算节点的 HBA 卡分别连接到两台存储交换机，每个存储控制器的光纤端口分别与两台存储交换机连接。

第5章 感谢使用

尊敬的用户与合作伙伴，云宏感谢您选用我们的产品。您在使用过程中遇到的任何产品或文档的问题和改进建议都可以通过 http://support.winhong.com/向我们反馈，我们将不断改进并向您提供更为优质的产品与服务产品。