FastCube 1000（虚拟化） 技术白皮书

FastCube 1000（虚拟化） 技术白皮书

文档版本	01
发布日期	2023-03-13

注意

您购买的产品、服务或特性等应受北京元亿科技服务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，北京元亿科技服务有限公司对本文档内容不做任作明示或暗示的声明或保证。

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

北京元亿科技服务有限公司

地址： 北京市朝阳区望京东区保利国际广场 T2-901 邮编：100000

网址： https://www.ed-in.com.cn/

客户服务邮箱： service@ed-in.com.cn

客户服务电话： 400-9652688

前言

概述

本文档介绍了 FastCube 1000 虚拟化场景的产品价值、产品架构、高性能、线性扩展、系统安全以及系统可靠性。

读者对象

本文档主要适用于以下工程师：

- 营销工程师
- 技术支持工程师
- 维护工程师

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
危险	表示如不避免则将会导致死亡或严重伤害的具有高等级风险的危害。
警告	表示如不避免则可能导致死亡或严重伤害的具有中等级风险的危害。
注意	表示如不避免则可能导致轻微或中度伤害的具有低等级风险的危害。
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、设备性能降低或其或其他预知的结果。 “须知”不涉及人身伤害。
说明	对正文中重点信息的补充说明。 “说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。

修改记录

文档版本	发布日期	修改说明
01	2023-03-13	第一次正式发布。

目　录

1 产品概述

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，传统服务器+存储的架构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。
- 系统庞大（不同厂商硬件设备维护、虚拟平台管理），维护难度大。

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。

FastCube 是一个开放的、可扩展的系统，具有计算/存储/网络融合、预集成、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。

2 产品价值

FastCube 遵循开放架构标准，集成服务器、分布式存储及网络交换机为一体，无需外置存储设备，并预集成了分布式存储引擎、虚拟化平台及管理软件，资源可按需调配、线性扩展。主要价值如下：

融合

FastCube 实现了计算、存储和网络资源的融合：

- 硬件融合：计算存储网络高度集成，线性扩容。
- 管理融合：统一运维管理，提高资源利用率，降低 OPEX 费用。
- 应用融合：针对应用业务模型，软硬件深度调优，实现性能提升。

简单

FastCube 实现了预安装、预集成和预验证、上电后的设备自动发现、统一的维护管理，端到端的简化了业务交付：

- 简化安装：硬件预置安装，软件预置集成，设备进场后开箱即用。
- 简捷交付：设备上电自动发现，参数自动配置，实现业务快速上线。
- 简单维护：统一界面管理，故障主动排查，简化日常运维。

优化

FastCube 通过采用业界领先硬件，以及分布式存储软件，为应用提供最优的业务体验：

- 存储优化：通过内置分布式存储，为应用提供了高并发、高吞吐量的存储服务。
- 网络优化：支持 GE、10GE、25GE 网络，提供高带宽的交换网络

开放

FastCube 是基于开放的超融合架构构建的基础设施平台，不绑定特定的上层应用，可以为业界主流虚拟化平台、数据库等提供计算、存储和网络资源：

3 产品架构

FastCube 1000 总体架构主要有：硬件平台、分布式存储软件、安装部署和运维管理平台、虚拟化平台以及相应的备份容灾方案，其中虚拟化平台支持华为自研的 FusionCompute 虚拟化平台。FusionCompute 场景下，FastCube 支持混合部署方案。FastCube 1000 总体架构详细构成如下图所示：

图 3-1 FastCube 1000 总体架构

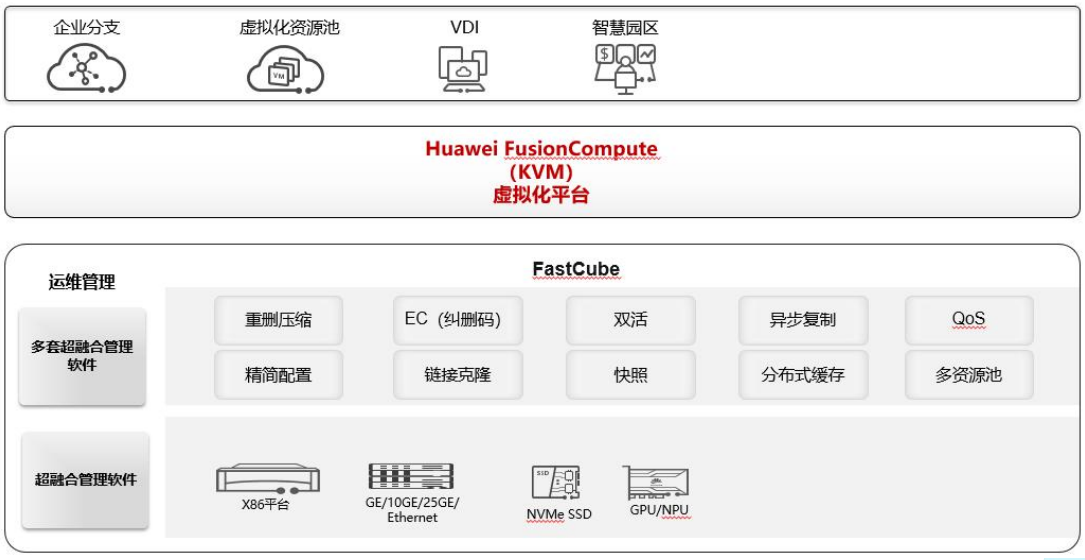


表 3-1 FastCube 1000 总体架构组件说明

名称	说明
FusionCube Vision	FastCube 的管理软件，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。
FusionCube Builder	提供现场快速安装部署 FastCube 系统软件，可用于现场更换虚拟化平台软件或者更新版本。
分布式块存储	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块存储业务。

名称	说明
虚拟化平台	支持华为自研 FusionCompute 虚拟化平台，提供系统虚拟化管理平台。
备份软件	用于备份系统业务虚拟化，主要包括自研的备份软件 eBackup 和第三备份软件 Veeam、CV、爱数等主流备份软件。
容灾软件	提供基于存储双活和存储异步复制的容灾方案，容灾软件主要采用自研的 UltraVR
硬件平台	服务器使用第三方机架服务器，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，支持 GPU，SSD PCIe 等 IO 加速扩展，支持丰富的交换模块 GE、10GE、25GE，根据业务要求灵活配置。

FastCube 遵循开放架构标准，融合服务器、分布式存储及网络交换机为一体，并预集成了分布式存储引擎、虚拟化平台及管理软件，资源可按需调配、线性扩展。

3.1 FusionCompute 场景架构

FusionCompute 虚拟化场景采用 KVM 虚拟化架构，系统主要由兼容适配的第三方服务器、分布式存储系统、FusionCompute 虚拟化平台以及 FusionCube Vision 管理平台构成，其中 FusionCube Builder 提供相应额软件安装操作。配套软件 eBackup、Veeam、UltraVR 等可以给系统提供备份、容灾等高级特性。

3.1.1 架构

在 FusionCompute 虚拟化的部署中，分布式存储软件直接部署在 Hypervisor 内核中，节点的 HDD 和 SSD Cache 存储介质通过分布式存储软件构造成系统共享的存储池资源，同时 FusionCompute 虚拟化平台将节点的计算资源虚拟提供给节点上的业务虚拟机使用。根据节点提供的功能特性差异，又分为管理融合节点、存储融合节点、计算节点，详细的节点架构如下图：

图 3-2 FusionCompute 场景节点架构

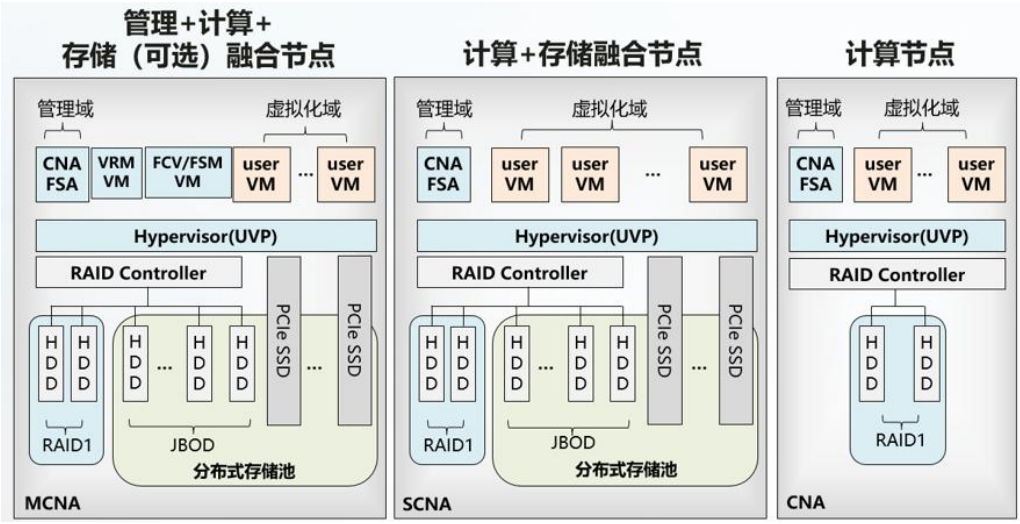


表 3-2 FusionCompute 场景各类节点说明

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、FUSIONCUBE VISION/FSM 等管理虚拟机。同时也可提供存储和计算功能	必须部署 2 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。

3.1.2 典型配置

FusionCompute 虚拟化场景可支持大容量的 HDD+SSD Cache 混合部署场景以及高性能的全 SSD 部署场景。具体的场景配置具体如下：

- **混合部署场景节点典型配置：**

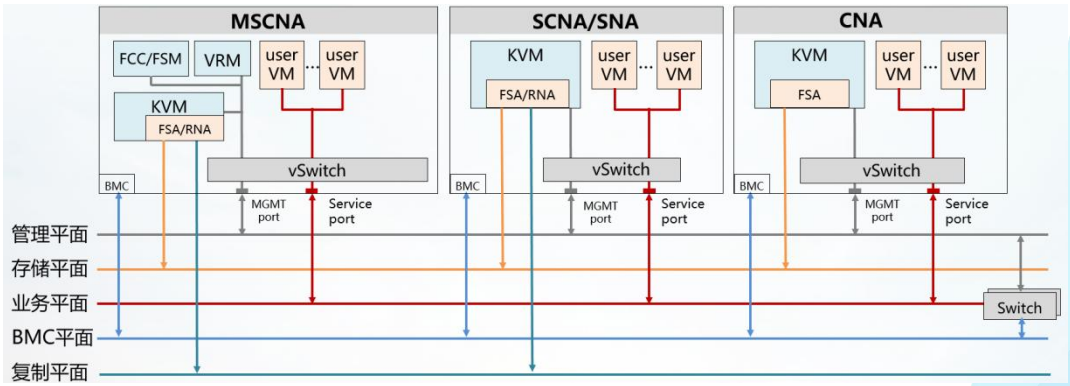
配置项	典型配置	说明
服务器类型	机架服务器	根据客户对机柜空间、磁盘大小、密度、PCIE 网卡数量等选择合适的服务器类型； 机架服务器：优势为灵活，支持各类硬盘类型，预留多个 PCIE 槽位，支持 GPU 卡。不足是空间占用大；
CPU/内存配置	X86 系列 CPU DDR4 最高	CPU/内存配置根据客户的业务规格和配置可以动态调整配置，提

配置项	典型配置	说明
	3200MHz 内存	供更多的计算资源： C4314、C4316、C5318Y、C6330、C6342、C6346、C6348 以及 4314、4316、5318Y、6330、6342、6346、6348
磁盘	4T/6T/8T/10T SATA 盘，1.2T/1.8T/2.4T SAS 盘 操作系统盘为 2*600GB SAS 盘或 2*960GB SAS SSD 盘	分布式存储要求 SATA/SAS 盘必须要采用 3 副本或者 EC 配比为 N+2 及以上的冗余策略。
Cache	3DWPD NVMe SSD	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/6.4TB NVME SSD 盘，缓存主存比建议〉=3%； Cache 类型为华为自研的 NVME SSD
网卡	2*GE+2*10GE+2*10GE 或 2*25GE	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，如果配置了容灾，则再增加 2*10GE 网口用于复制网络平面。

3.1.3 组网

FusionCompute 场景的系统组网包含：管理平面、存储平面、业务平面、BMC 平面以及容灾涉及的复制、仲裁平面。详细的组网情况如下：

图 3-3 FusionCompute 场景系统组网图



通信平面类型说明介绍：

- 管理平面：FastCube 系统的管理网络平面，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；

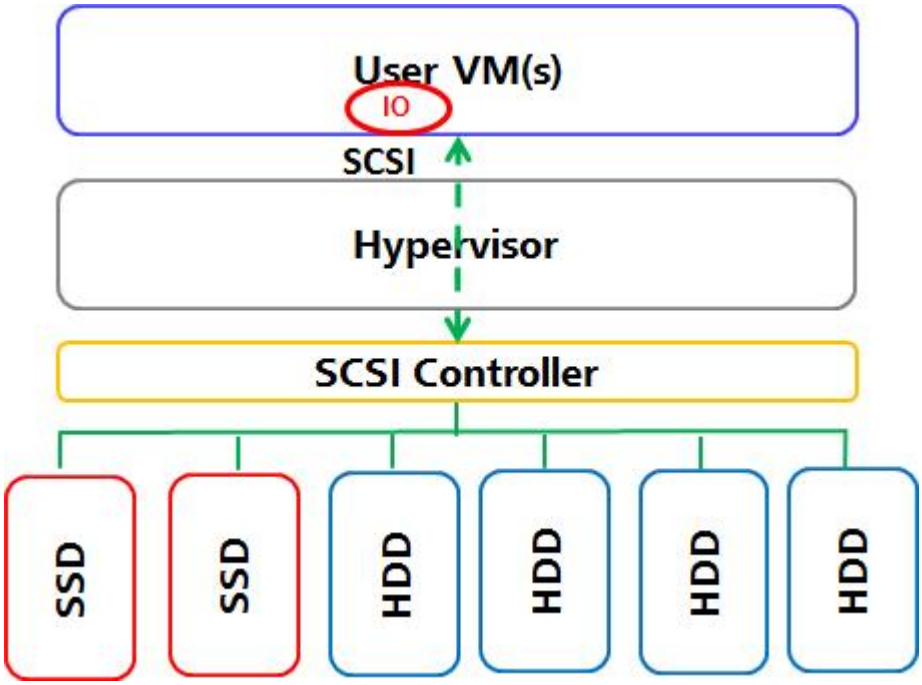
- 存储平面：分布式存储节点间数据读写操作网络平面，支持 TCP/IP 协议，支持 10GE/25GE 组网，建议独占网卡；
- 业务平面：客户业务通信网络平面，支持 TCP/IP 协议，支持 GE/10GE 组网，可以与管理平面共网卡，通过 VLAN 隔离；
- BMC 平面：服务器设备管理平面，访问 FastCube 系统服务器设备的运维管理平台；
- 复制平面：容灾方案中，主备站点间的数据同步网络平面，支持 TCP/IP 协议，支持 10GE 组网，建议系统独占网卡，避免与其他网络平面争抢资源；

3.1.4 工作原理

业务 IO 流程

FusionCompute 场景下，业务虚拟机部署在主机 KVM 虚拟化平台上，存储采用的分布式存储资源，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 流程参考图 3-4。

图 3-4 FusionCompute 场景业务 IO 流程图



FusionCompute 场景下的节点架构 IO 路径短，效率更高。

业务管理与运维

FusionCompute 场景下，支持系统预安装集成能力，系统在出厂时已将系统硬件 BIOS 和磁盘 RAID 配置、虚拟化平台、管理软件、分布式存储软件预安装在节点内，客户在接到产品后，上电后可以快速完成系统的配置、初始化。客户完成系统初始化后，即可在系统进行发放业务虚拟机使用。

在 FusionCube Vision 管理页面上，支持业务虚拟机的创建以及虚拟机生命周期的管理，支持对系统硬件、存储、计算资源、虚拟机等监控管理，将系统各个组件得告警汇总统一上报管理，提供系统的一键式运维能力，包括：一键式扩容、升级、日志收件、健康巡检等。虚拟化平台一些高级特性或配置可通过单点登录的方式跳转至虚拟化管理平台 FusionCompute 页面进行操作，如：计算集群的创建配置、DVS 创建和配置、虚拟机高级特性配置等非常用操作。

4 分布式存储

FastCube 内置分布式存储为业务提供存储服务，分布式存储提供是块存储设备，采用独特的并行架构、创新地缓存算法、自适应的数据分布算法，既消除了热点也提高了性能，并且能够以超快的重建时间实现自动化自修复，提供卓越的可用性和可靠性。

- 线性扩展和弹性

  分布式存储采用全分布式 DHT 架构，将所有元数据按规则分布在各节点，避免了元数据瓶颈，支持线性扩展。分布式存储采用了独特的数据分块切片技术，以及基于 DHT Hash 的数据路由算法，可以将卷的数据均匀地分散到较大的资源池故障区域范围内，使得每个卷可以获得更大的 IOPS 和 MBPS 性能，也使得每个硬件资源的负载相对均衡。

- 高性能

  分布式存储免锁化调度的 IO 软件子系统，彻底解决了分布式锁冲突，使得 IO 路径上无需进行任何锁操作和元数据查询，IO 路径短、时延低；分布式的无状态机头，可以充分发挥各个硬件节点的能力，大大提升了系统的并发 IOPS 和并发 MBPS。同时分布式存储采用分布式的 SSD cache 技术，配合大容量的 SAS/SATA 盘做主存，使得系统的性能可以具备 SSD 的性能和 SAS/SATA 的容量。

- 高可靠性

  分布式存储支持多种数据冗余保护机制，如 2 副本、3 副本、EC 等；在此基础上，分布式存储支持设置灵活的数据可靠性策略，允许将不同的副本放在不同的服务器上，保证在服务器故障的情况下，数据仍然不丢失、仍然可访问。同时采用对有效数据分片进行数据的冗余保护，在硬盘、服务器故障的时候，能够对有效数据进行并行重建，1TB 硬盘的重建时间小于 15 分钟，大大增强系统的可靠性。

- 丰富的存储高级功能

  - 精简配置，当用户对卷进行写操作时才分配实际物理空间，来为用户提供比物理存储资源更多的虚拟存储资源。

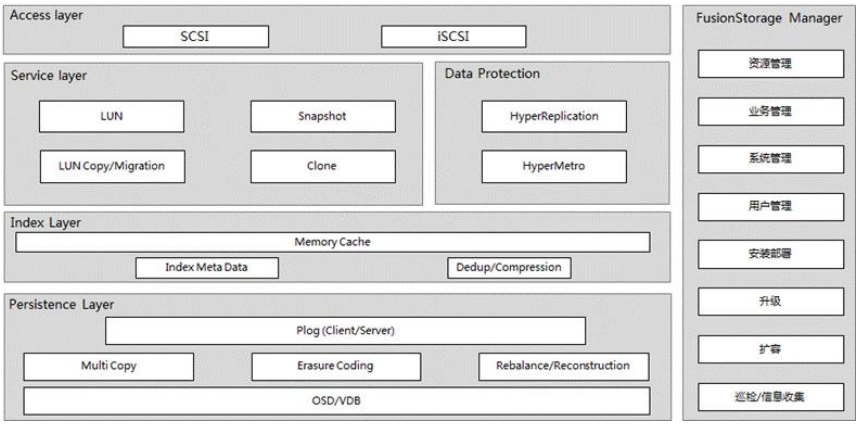
– 卷快照，将用户的逻辑卷数据在某个时间点的状态保存下来，作为快照点；快照不限次数且性能不下降。

– 链接克隆，基于增量快照提供链接克隆，一个快照可以创建出多个克隆卷，各个克隆卷刚创建出来时的数据内容与快照中的数据内容一致，后续对于克隆卷的修改不会影响到原始的快照和其他克隆卷。

4.1 架构概述

分布式存储采用分布式集群控制技术和 DHT 路由技术，提供分布式存储功能特性。分布式存储功能架构如图 4-1 所示。

图 4-1 分布式存储功能框架图



系统描述	类型		描述
业务系统	访问接入		用于应用访问存储系统的标准访问接口，支持 SCSI 标准访问接口协议
	卷特性层		卷提供各种特性，如快照，克隆，迁移，异步复制，双活等企业级特性，均在此层实现
	索引层		用于数据逻辑空间和物理空间的转换，重删压缩等在该层实现
	持久化层		采用 Plog 接口访问（一种 Append Only 的 ROW 写机制）用于数据的存放，包括多副本，EC，数据均衡与重构等，并通过 OSD/VDB 对盘进行管理和数据读写
管理系统	业务管理子系统 FusionStorage Manager	资源管理	存储资源池进行管理和分配，提供数据冗余保护，包括多副本保护和纠错码保护
		业务管理	支持按存储资源池发放块存储服务

系统描述	类型		描述
		系统管理	支持对系统进行初始化配置和必要的业务功能配置，设备拓扑管理，可提供系统设备拓扑关系图，方便查看和管理设备间的拓扑关系。
		用户管理	支持对用户的增删改查，包括用户的等级，权限等
		安装部署	完成系统的初始安装，部署
		升级	支持对系统的升级，包括软件升级，操作系统升级，固件升级
		扩容	完成系统的在线扩容缩容
		巡检/信息收集	设备详情管理，可提供设备详细配置和运行状态信息收集，方便了解设备配置信息和健康状态。

分布式存储逻辑架构如图 4-2 所示。

图 4-2 分布式存储逻辑架构图

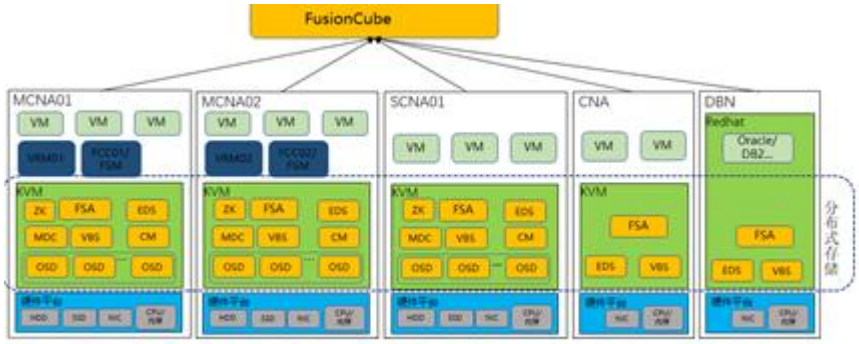


表 4-1 分布式存储逻辑组件

名称	说明
FSM	FusionStorage Manager 缩写，FusionStorage 管理模块，提供告警、监控、日志、配置等操作维护功能。与 FusionCube Vision 共部署在一起，工作在主备模式下
FSA	Fusiostorage Agent 缩写，代理进程，部署在各节点上，实现各节点与 FusionStorage Manager 通信，可收集各节点的监控与告警信息或在升级本节点软件组件时接收升级包与执行升级。
ZK	Zookeeper 缩写。一个系统需部署 3、5、7 等奇数个 Zookeeper 组成。Zookeeper 集群，为 MDC 集群提供选择主

名称	说明
	仲裁，Zookeeper 至少 3 个，必须保证大于总数一半的 Zookeeper 处在活跃可访问状态。
MDC	元数据控制软件，实现对分布式集群的状态控制，以及控制数据分布规则、数据重建规则等。一个系统至少部署 3 个 MDC，形成 MDC 集群，系统启动时由 Zookeeper 集群在多个 MDC 中选举主 MDC，主 MDC 对其他 MDC 进行监控，主 MDC 故障时产生新的主 MDC。每个资源池有一个归属 MDC，当某池的归属 MDC 故障时，主 MDC 指定另外的 MDC 托管这个资源池，一个 MDC 最多管理两个资源池。MDC 作为一个进程可以在每个存储节点启动，增加资源池会自动启动 MDC，一个系统最多启动 96 个 MDC。
VBS	虚拟块存储管理组件，执行卷元数据管理，VBS 通过 SCSI 或 iSCSI 接口提供分布式存储接入点服务，使计算资源能够通过 VBS 访问分布式存储资源。VBS 与其所能访问的资源池的所有 OSD 点对点通信，使 VBS 能并发访问这些资源池的所有硬盘。每个节点上默认部署一个 VBS 进程，多个节点上的 VBS 形成 VBS 集群，VBS 启动时与主 MDC 连接并协调主 VBS。节点上也可以通过部署多个 VBS 来提升 IO 性能。
OSD	KV 设备服务，执行具体的 I/O 操作。在每个节点上部署多个 OSD 进程，一块磁盘默认对应部署一个 OSD 进程。在 SSD 卡作主存时，为了充分发挥 SSD 卡的性能，可以在 1 张 SSD 卡上部署多个 OSD 进程进行管理，例如 2.4TB 的 SSD 卡可以部署 6 个 OSD 进程，每个 OSD 进程负责管理 400GB。
EDS	EEnterprise Data Service 组件，接收到来自 VBS 的 I/O 业务之后，执行具体的 I/O 操作。在 EDS 服务里面，会执行有关快照、克隆等与模块相关的特性，同时还对存储空间的做管理，将块的数据与存储空间建立索引关系，确保每块数据通过索引都能找到对应的存储位置；同时在数据存储到物理空间之前，可以进行重删压缩处理。
CM	Cluster Manager，集群管理软件，用于管理整个存储集群的状态信息，包括各组件的状态信息，实时监控各组件的状态，当组件出现故障时，根据组件状态触发相关措施来恢复错误。
CCDB	Cluster Configuration Database，集群配置数据库，用于保存用户配置信息的数据库，当前在 EDS 组件中会采用 CCDB 存放配置信息。

4.2 关键业务流程

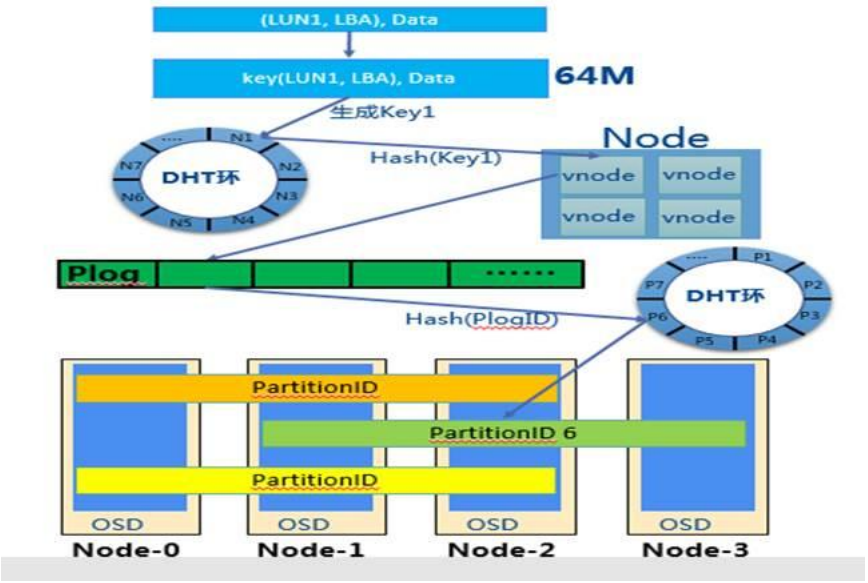
4.2.1 数据路由

分布式存储数据路由采取分层处理方式：

- VBS 通过计算确定数据存放在哪个节点的哪块硬盘上。
- OSD 通过计算确定数据存放在硬盘的具体位置。

具体流程如下图所示：

图 4-3 分布式存储数据路由示意图



- 第一层 DHT hash 环的目的是通过 hash 算法将数据分发到计算出来的存储服务器节点处理该数据，通过该 hash 算法，确保每个数据都有对应的服务器节点来处理，保证了业务处理的均衡。系统根据 LUNID 和 LBA 定位到服务器节点，然后再定位到该服务器上的 vnode 上，由该 vnode 逻辑处理单元来处理该数据；vnode 是一种逻辑处理单元，将物理服务器节点分为 4 个逻辑处理单元，即 4 个 vnode，例如：一个由 6 个物理服务器组成的一个存储集群，当其中 1 个物理服务器故障时，该服务器上的 4 个 vnode 处理的业务，可以分别被该集群中另外的 4 个物理服务器去接管，这样剩下的 5 个物理服务器中，有 4 个物理服务器运行有 5 个 vnode，1 个物理服务器运行 4 个 vnode，通过 vnode 机制，可以确保故障节点的业务可以分散到不同的服务器节点上去接管，就可以防止只用一个物理服务器接管带来的业务处理瓶颈问题。该 DHT hash 环打散粒度是按 64MB 对齐打散。
- 第二层 DHT hash 环的目的是通过 hash 算法将数据转到对应存储空间去保存，完成数据的持久化。通过该 hash 算法，确保数据存储空间的均衡性。系统根据 PlogID 和 Offset 定位到硬盘应该存放的具体位置，避免在海量数据中进行查找和计算，该 DHT 路由技术，采用华为自研算法，不仅能保证数据在各个硬盘的均衡性，而且在硬件增减（故障或扩容）时，自动快速调整，并保证数据迁移的有效性，确保自动快速自愈，自动资源均衡。

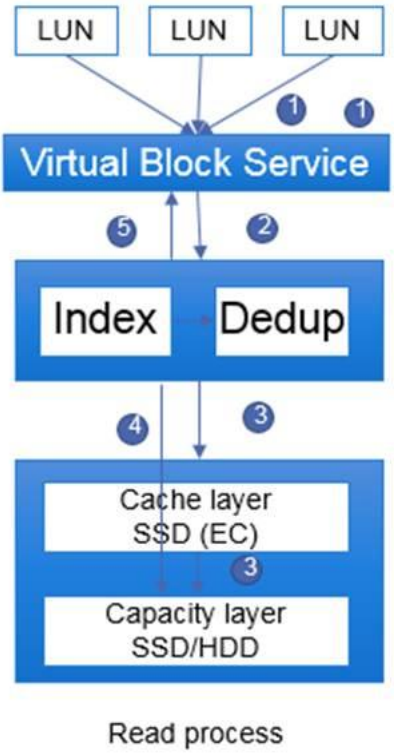
存储空间根据可靠性有机柜级、节点级、硬盘级，默认是跨节点组织副本/EC。

4.2.2 IO 路径

读 IO 流程

分布式存储系统中的读 IO（EC）流程如图 4-4 所示。

图 4-4 分布式存储读 IO 流程



① 上层应用下发读 IO 请求到存储服务，存储服务的 VBS（Virtual Block Service）模块收到该 IO 请求，根据第一层的 DHT hash 算法将数据转到指定服务器；

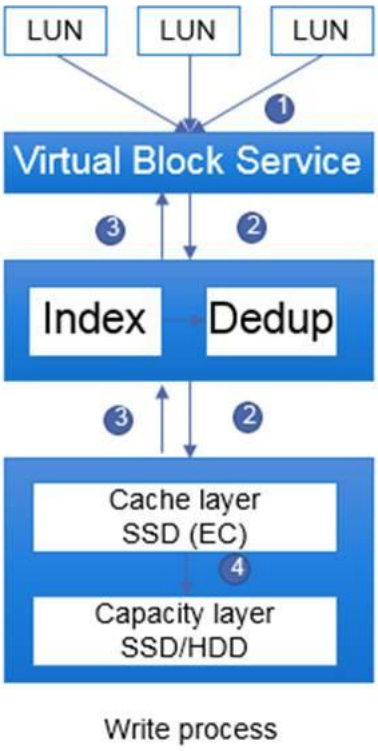
② 服务器上的 EDS（Index+Dedup）模块处理该数据。EDS 接收到读 IO 请求后，优先在内存地写缓存中查找，如果找到就返回给 VBS。

③ 如果内存写缓存中没有命中，则再在内存读缓存中去查找，如果仍然没有找到，则到存储介质中去读，先在 SSD Cache 中去读，如果还不命中，则到存储介质）中去读（详细见 Read Cache 章节说明）。

写 IO 流程

分布式存储系统中地写 IO（EC）流程如图 4-5 所示。

图 4-5 分布式存储写 IO（EC）流程



1. 上层应用下发写 IO 请求到存储服务，存储服务的 VBS（Virtual Block Service）模块收到该 IO 请求（图中①），根据第一层的 DHT hash 算法将数据转到指定服务器；

2. 由这个服务器上的 EDS（Index+Dedup）模块处理该数据（图中②上）；

3. EDS 接收到写 IO 请求后，以小比例 EC 形式写入 Cache Layer 层的 SSD 缓存盘上（图中②下），同时该 EDS 所在服务器的内存中仍然保持一份该数据，EDS 返回写 IO 成功给 VBS（图中③），再由 VBS 返回给上层应用。

4. 待内存中的数据聚合到更大的块，走刷盘流程异步刷入（图中④）到 Capacity Layer 的存储介质中。

4.2.3 Cache 机制

分布式存储采用多级 Cache 机制提升存储 IO 性能，读、写 Cache 机制采用不同流程。

Write Cache

VBS 发送地写 IO 操作（图中 Write IO From Host）时，会将 Write IO 在 Memory Write Cache 内存中保存一份，同时同步以日志的方式（采用固定的 2+2 小分片 EC）记录到 SSD WAL Cache 中并返回成功完成本次写操作，这个流程通常称为 Host Write IO 流程。

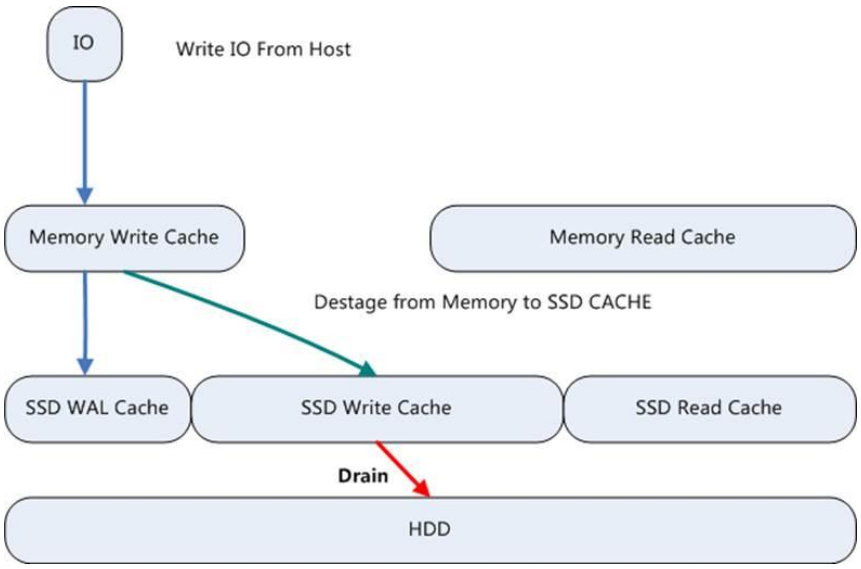
通常 SSD Disk Cache 分为两个部分：SSD Write Cache 和 SSD Read Cache。Memory Write Cache 中的数据会进行 IO 排序重整并等待满分条以副本或 EC 的方式直

接写入到 SSD Write Cache 中并返回；对于大块 IO 则直接由 Memory Write Cache 直通写到 HDD 中，而不驻留在 SSD Write Cache 里；

当 SSD Write Cache 中的保存数据水位达到 40%时，则由 SSD Write Cache 往 HDD 中搬迁。

随着 Memory Write Cache 中的数据逐步刷盘到 SSD Write Cache 时，SSD WAL Cache 中的数据将逐步淘汰掉，我们通常会进行异步的垃圾回收。

图 4-6 分布式存储写 Cache 机制示意图



相比较传统的副本方式写入 SSD Cache，然后异步地再从 SSD Cache 中读出满分条到持久化存储层 HDD，分布式存储的 SSD WAL Cache 方案带来 4 大优势：

- 分布式存储的 SSD WAL Cache 地写放大比较小，2+2 的 EC 的 Overhead 为 2；而副本方式的 SSD Cache，OverHead 最低必须为 2。
- 由于写放大较小，分布式存储对网络的带宽消耗也较低
- 分布式存储的 SSD WAL Cache 可靠性高，是+2 的冗余保护。
- 分布式存储的数据往主存上刷盘通常是由 RAM 中触发完成的，比传统的后台异步先从 SSD Cache 读出再写到主存中的效率高。

Read Cache

分布式存储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，SSD Cache 采用热点解读机制，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。

OSD 在收到 VBS 发送的读 I/O 操作时，会进行如下步骤处理：

步骤 1　从内存“Memory Write Cache”中查找是否存在所需 I/O 数据，如果存在，则直接返回，同时调整该 IO 数据到“读 Cache”LRU 队首，否则执行步骤 2；

步骤 2　从内存“Memory Read Cache”中查找是否存在所需 IO 数据，如果存在，则直接返回，同时增加该 IO 数据的热点访问因子，否则执行步骤 3；

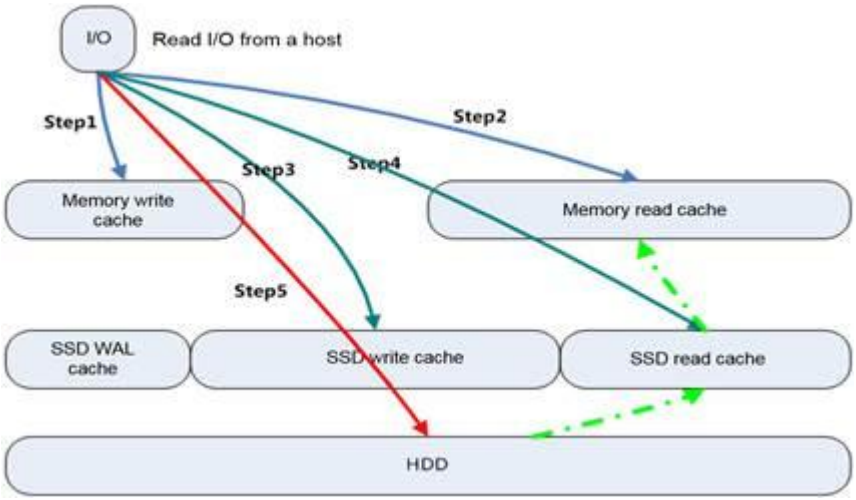
步骤 3　从 SSD 的“SSD Write Cache”中查找是否存在所需 IO 数据，如果存在，则直接返回，如果不存在，执行步骤 4；

步骤 4　从 SSD 的“SSD Read Cache”中查找是否存在所需 IO 数据，如果存在，则直接返回，同时增加该 IO 数据的热点访问因子；如果热点访问因子达到阈值，则会被缓存在 SSD 的“SSD Read Cache”中，如果不存在，执行步骤 5；

步骤 5　从硬盘中查找到所需 IO 数据并返回，同时增加该 IO 数据的热点访问因子，如果热点访问因子达到阈值，则会被缓存在 SSD 的“SSD Read Cache”中。

---一结束

图 4-7 分布式存储读 Cache 机制示意图



4.3 存储管理

4.3.1 存储集群管理

分布式存储通过集群管理软件完成集群的管理工作，功能包括集群基本信息监控、性能监控、告警管理、用户管理、license 管理、硬件管理。

- 集群基本信息监控：查看集群的基本信息，包括集群名称、健康状态、运行状态、节点信息、节点进程信息。
- 性能监控：查看 CPU 利用率、内存利用率、带宽、IOPS、时延、磁盘利用率、存储池利用率统计。
- 告警管理：提供查看告警信息、清除告警、屏蔽告警的功能。
- 用户管理：系统管理员可以创建新的管理员，为该管理员赋予一定的管理权限，以便多个管理员按照所授权权限进行系统或资源管理。对用户的操作包括：查询、删除、创建、解锁、冻结用户等。支持设置密码策略以提升系统安全。
- License 管理：提供查看已激活的 license 和导入新 license 功能。
- 硬件管理

硬件管理包括服务器管理和磁盘管理。服务器管理对系统中的所有服务器集中管理，可查看服务器的软件安装状态、软件版本号、是否加入集群，可查看在服务器上创建的存储池状态以及存储池在该服务器的拓扑信息，支持将服务器设置为维护模式以方便对服务器进行故障恢复处理，支持对服务器的 CPU、内存进行性能监控。磁盘管理将系统中所有的磁盘集中管理，支持查看磁盘的状态、槽位号、序列号、磁盘使用率、类型等，支持磁盘包括 IOPS、时延、带宽、利用率等监控性能统计。

4.3.2 存储服务化

分布式存储的管理平台用户按角色分为“系统管理员”“系统操作员”和“系统查看员”，提供的管理功能可分为资源接入和配置、资源管理和维护、系统管理和维护三类。资源管理维护包括系统概览汇总信息、存储池管理、块客户端管理、卷管理、虚拟文件系统管理、硬件管理等。

- 存储池管理

存储池管理可查看选定存储池的统计信息，查看选定存储池的硬盘拓扑，为选定存储池扩容、减容，以及删除存储池。还提供创建新存储池功能。

- 块客户端管理

块客户端管理提供创建、删除客户端功能。也提供查看块客户端的挂载信息与 CPU 及内存的监控统计信息，为块客户端进行挂载和卸载卷等操作。

- 卷管理

卷管理提供卷的创建和删除功能。创建卷需指定资源池、卷名、卷大小等信息。对于创建后的卷若按 SCSI 协议使用需要挂载卷，若按 iSCSI 协议使用需要做 iSCSI 映射。还提供 iSCSI 卷映射界面完成创建主机/主机组、配置启动器、配置 CHAP 认证、为主机/主机组映射/解映射卷等操作。

注：默认情况下 iSCSI 功能是关闭的，若要使用 iSCSI 功能需要先开启 iSCSI 功能并添加 iSCSI 监听的 IP 地址和端口。

- QoS 策略管理

QoS 策略管理支持创建、删除 QoS 策略，及分页查看 QoS 策略信息。

- 快照管理

快照管理支持分页出查看快照列表，列表信息包括快照名称、容量、所属存储池和创建时间；支持创建链接克隆卷、设置 QoS 策略和删除快照。

4.4 数据冗余

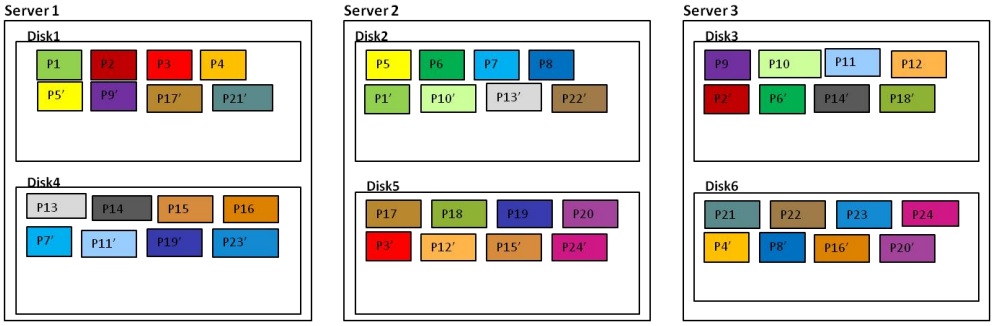
分布式存储支持两种数据冗余保护机制，一种是多副本方式，一种是 Erasure Code（EC，纠错码）方式。

4.4.1 多副本

分布式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，默认按照 1MB 进行分片，分片后的数据按照 DHT 算法保存集群节点上。

如图 4-8 所示，对于节点 Server1 的磁盘 Disk1 上的数据块 P1，它的数据备份为节点 Server2 的磁盘 Disk2 上 P1，P1 和 P1 构成了同一个数据块的两个副本。例如，当 P1 所在的硬盘故障时，P1 可以继续提供存储服务。

图 4-8 分布式存储多副本示意图

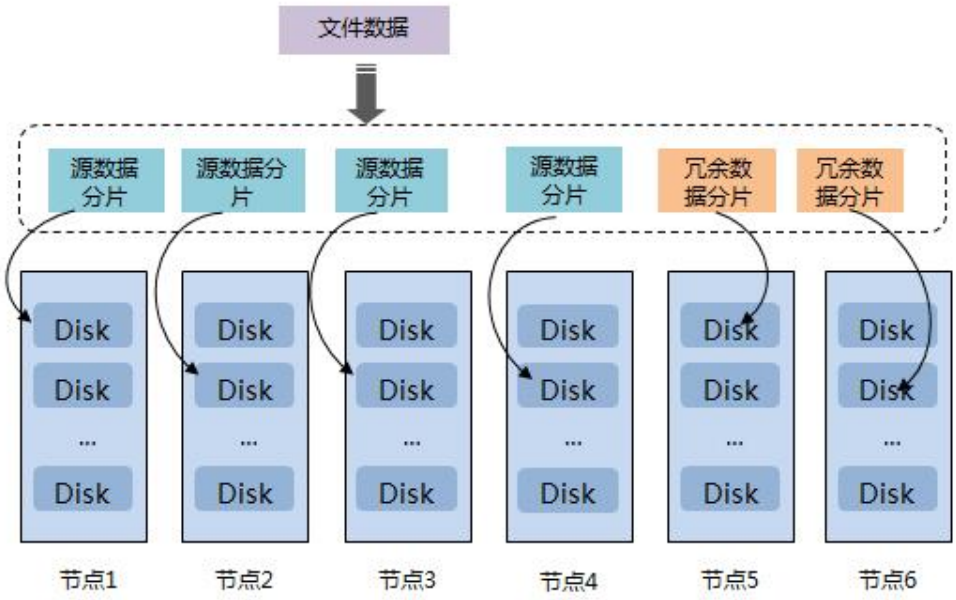


4.4.2 Erasure Code

分布式存储也可以采用 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。

基于 EC 的分布式存储的数据保护技术，是建立在分布式、节点间冗余的基础上的。分布式存储采用自研 LDEC（Low Density Erasure Code）算法，基于 XOR 和伽罗华域乘法相结合的一种 MDS 阵列码，编码最小粒度 512B，支持 Intel 指令加速，支持各种主流配比。数据进入系统之后，首先被切分为 N 个数据条带，然后计算出 M 个冗余数据条带，并最终保存在 N+M 个不同的节点中。

图 4-9 分布式存储 EC 示意图



图示以4份数据切片2份冗余切片存储在6个节点上举例

弹性 EC 是一种增强型数据冗余保护机制，广泛应用于分布式存储领域。EC 在分布式存储系统中使用 N 个数据块和 M 个校验块保证数据的可靠性，这 N+M 个数据块中有任意 M 个块数据损坏，都可以通过其他 N 个块上的数据恢复 M 个块的数据。

相比于副本存储方式，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硬盘利用率，从而降低成本。比如一个 4M 的 IO，在三副本存储方式下，共占有 12M 的硬盘空间，而在 4+2 配比的 EC 存储方式下，4 个数据节点每个占用 1M 空间，2 个校验节点各占用 1M 空间，共 6M 空间，在提供相同可靠性的前提下，EC 比三副本节省了 6M 硬盘空间。

EC 在节点扩容时支持扩列功能，对于 N+M 配比扩至规则为 2*N+M，如 4+2 的 EC 扩列时直接扩到 8+2，然后到 16+2

EC 在节点故障时，如果节点数不满足 EC 最小节点数时，就会采用缩列方式，确保可靠性不下降；对于 N+M 的缩列机制，通常采用 N/2+M 的方式缩列，如 4+2 的 EC 缩列时直接缩到 2+2，8+2 的 EC 缩列则缩到 4+2，10+2 的 EC 缩列则缩到 4+2（不能采用奇数数据列数，如果为奇数则向下偶数取整）

EC 的性能通常比副本的性能高 15%左右，在高比例 EC 配比中，最大能支持到 22+2、20+3 和 20+4 三种最大配比。

4.5 特性介绍

4.5.1 SCSI 块接口

分布式存储通过 VBS 以 SCSI 或 iSCSI 方式提供块接口。SCSI 方式可为安装 VBS 的本机提供存储访问，物理部署、FusionCompute 或 KVM 等采用 SCSI 方式。

SCSI 协议支持 SCSI-3 持久预留锁和非持久预留锁：

- 持久预留锁可用于 HANA 集群。
- 非持久预留锁可用于 MSCS 集群。

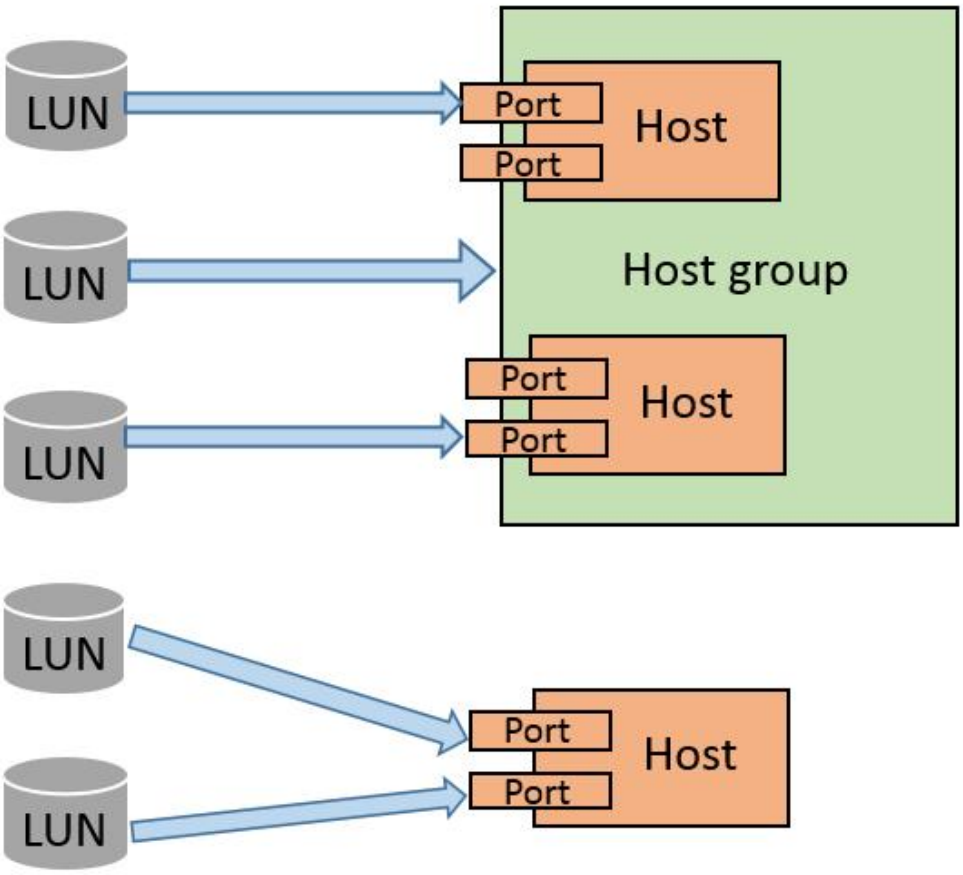
对于 iSCSI 协议的支持是通过 VBS 提供 iSCSI Target，块存储使用方通过本机的 Initiator 与 iSCSI Target 连接来访问存储。

对于 iSCSI 协议需要保证安全访问，分布式存储支持以下安全访问的标准：

- 支持 **CHAP** 身份验证以保证客户端的访问是可信与安全的。CHAP 全称是 PPP 询问握手认证协议（Challenge Handshake Authentication Protocol）。该协议可通过三次握手周期性的校验对端的身份，可在初始链路建立时以及链路建立之后重复进行。通过递增改变的标识符和可变的询问值，可防止来自端点的重放攻击，限制暴露于单个攻击的时间。
- 支持 **LUN MASKING** 给 Host 对 Lun 的访问进行授权。对于 SAN 存储主机将 Lun 当作本地设备，在主机端进行数据的维护，需要对各主机对 Lun 的访问进行隔离，避免各主机互相破坏对方的数据。LUN Masking 将 Lun 与主机的 HBA WWN 地址绑定，通过 LunMasking 功能保证 Lun 只能被指定的 Host 或 Host 集群访问， 未授权的 Host 将无法访问。 主机与 Lun 间既可有多对一的关系也可有一对多的关系，一对多能满足虚拟化场景小 LUN 方式使用存储的需求，多对一能满足 Oracle RAC 等集群系统使用共享卷的需求。

LunMasking 核心功能由 Port、Host、HostGroup、LUN 几大组件相互建立映射关系实现。

图 4-10 LunMasking 组件关系图



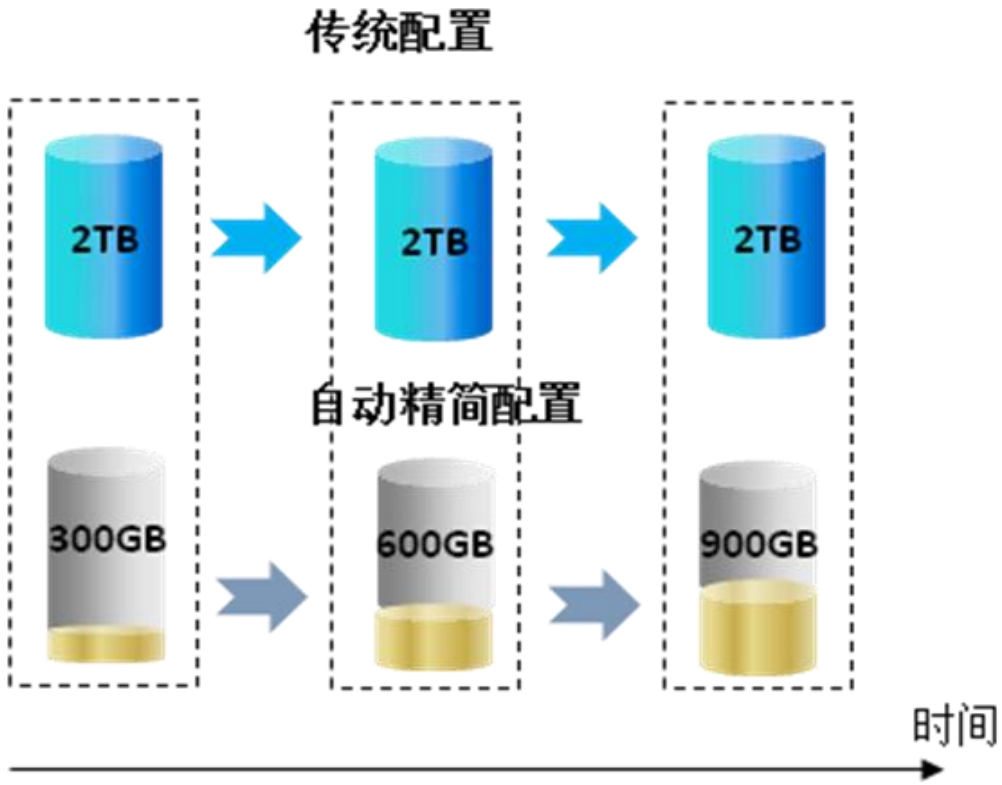
支持 **LUN MAPPING** 将 Lun 与存储端的端口绑定，主机端连接不同的端口使用不同的 Lun。当一个存储系统同时为多个应用系统提供数据存储服务，且不同的应用系统的主机分别处于不同的地理地址时，可用到 LUN Mapping。

4.5.2 精简配置

分布式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。

采用 DHT 路由技术，系统无需使用专门的集中元数据来记录卷的精简分配情况，和传统 SAN 相比，不会带来性能下降。

图 4-11 分布式存储自动精简配置



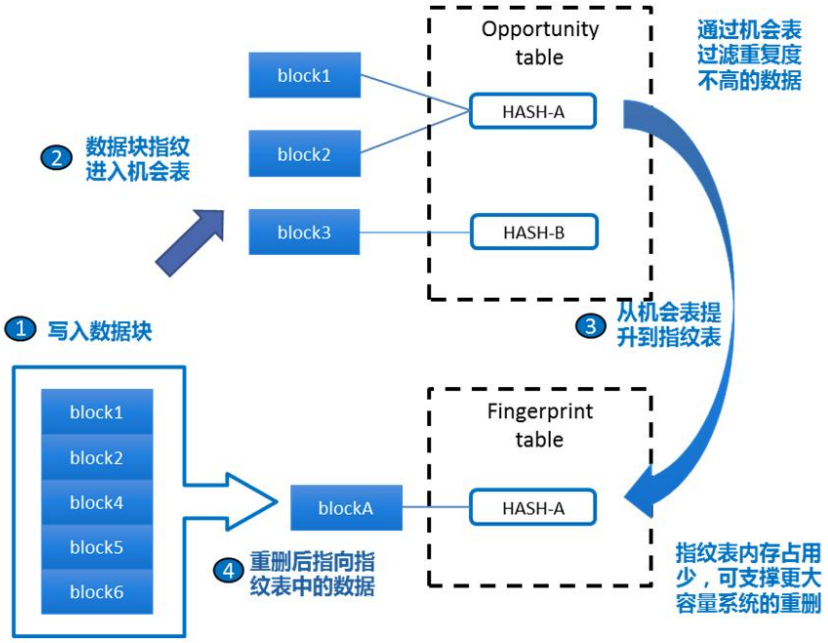
4.5.3 重删压缩

为了应对管理数据增加带来的运营成本的增长，各存储设备的厂商都在存储设备中增加了对应的数据缩减手段，如重复数据删除（Deduplication，重删）和数据压缩（Compression，压缩）技术以减少实际需要保存的数据量，从而降低企业的运营成本。SmartDedupe 和 SmartCompression 是华为公司自主研发的、基于 FastCube 分布式块服务地重删压缩特性。

FastCube 1000 采用了智能的自适应重删技术，以用户需求为导向，在用户数据处理负载较高的情况下，前重删会自动关闭，优先确保性能，由后处理完成数据缩减。在负载较低的情况下自动开启前重删，避免了后处理的读写放大。自适应技术以用户为导向，在用户不感知的情况下根据负载自动切换重删方式，避免了在线重删和后重删两种方案的缺点。

为了获取较好地重删压缩效果，FastCube 采用了全局重删。同时分布式存储空间巨大，为了减少指纹表的内存空间消耗，引入指纹机会表的机制（如下图）。数据块的指纹先进入机会表计数，只有相同数据块被多次写入达到阈值（默认 3，可配置），方可进入指纹表执行重删。

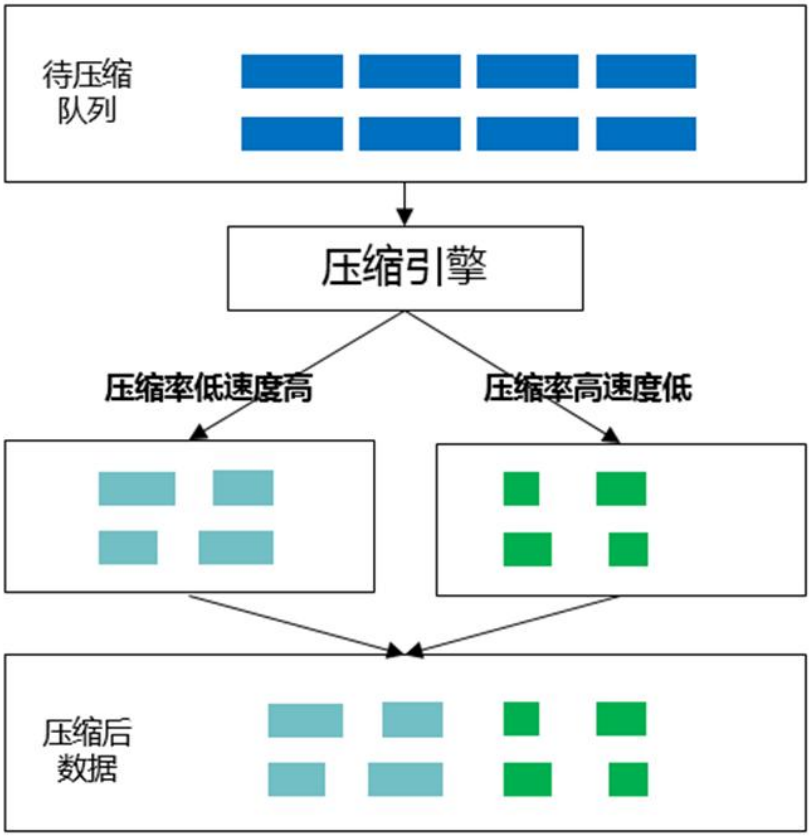
图 4-12 分布式存储重删流程



如果压缩未开启，则直接申请该数据块长度存储空间保存数据。开启了压缩的情况下，则会进行压缩。数据块将被压缩引擎压缩，然后以最小 512 字节的粒度进行保存。

分布式存储压缩引擎采用不同两种不同压缩算法组合运行，一种是压缩率低速度高的算法，一种是压缩率高但速度低的算法。通过配置两种压缩算法不同的执行比例，可以得到不同的性能和数据缩减率。在同一个存储池只能选择一个压缩算法。存储池压缩算法的修改不会影响已经压缩的数据，已经压缩的数据在读取时会根据压缩时的算法进行解压。

图 4-13 分布式存储多策略压缩

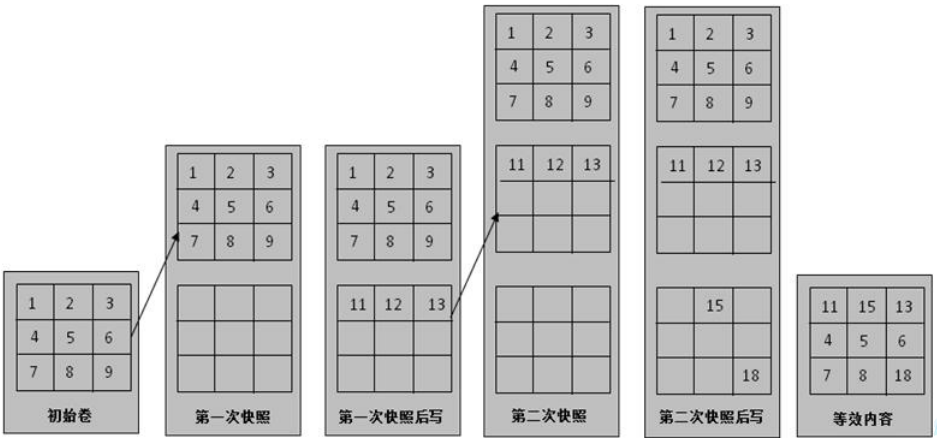


4.5.4 快照

分布式存储提供了快照机制，将用户的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。

分布式存储快照数据在存储时采用 ROW（Redirect-On-Write）机制，快照不会引起原卷性能下降。

图 4-14 分布式存储快照



分布式存储也支持共享卷的快照，提供共享卷备份能力。

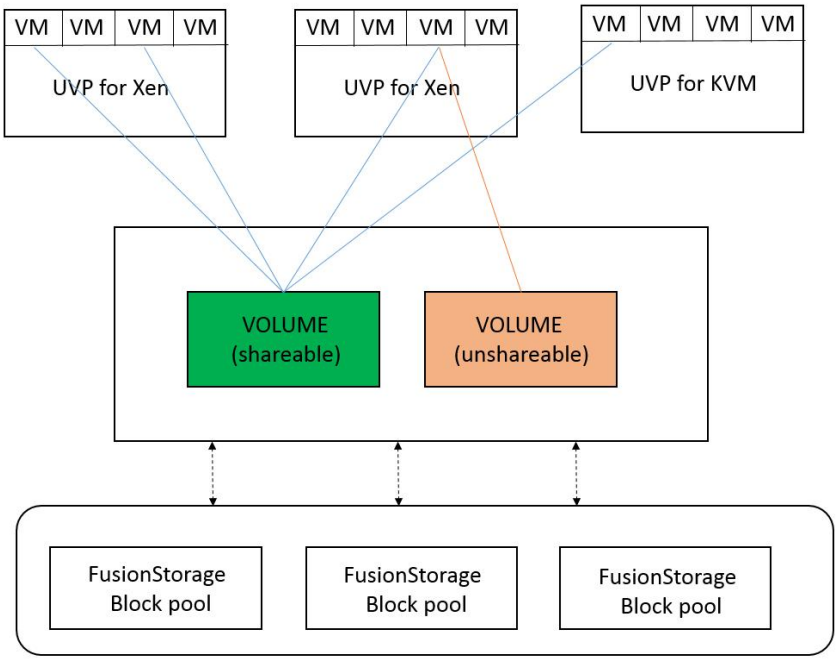
存在多个挂载点的 SCSI 卷称为共享卷，而对 iSCSI 卷而言，所有的 iSCSI 卷都是共享卷。

和普通卷快照流程不同，由于共享卷存在多个挂载点，多个挂载点的 VBS 都可能下 IO，在对共享卷打快照时需要多个节点协同配合来完成悬挂 IO 的操作，分布式存储采用两阶段来实现该功能：

1. 由主 VBS 向所有的挂载点 VBS 发出 prepare 消息，挂载点 VBS 在收到 prepare 消息后进行 IO 悬挂的操作，并回复 OK 给主 VBS。

2. 主 VBS 向所有的挂载点 VBS 发送 commit 消息，并带上需要更新的卷元数据信息，参与者在收到消息后更新本地元数据信息并解开 IO 悬挂，然后回复 OK 给主 VBS 后，完成本次事务。

图 4-15 分布式存储共享卷快照



一致性快照用于整机备份，一个虚拟机通常挂载了多个卷，对虚拟机做整机备份时所有卷快照的应处于同一时间点，才能保证数据恢复的可靠性。

分布式存储支持一致性快照的能力， 对上层发起的一致性快照请求分布式存储会保证多个卷的快照属于同一个时间点。

分布式存储对多个卷执行悬挂 IO 后，再更新快照信息操作，保证了多个卷快照时间点的一致性。

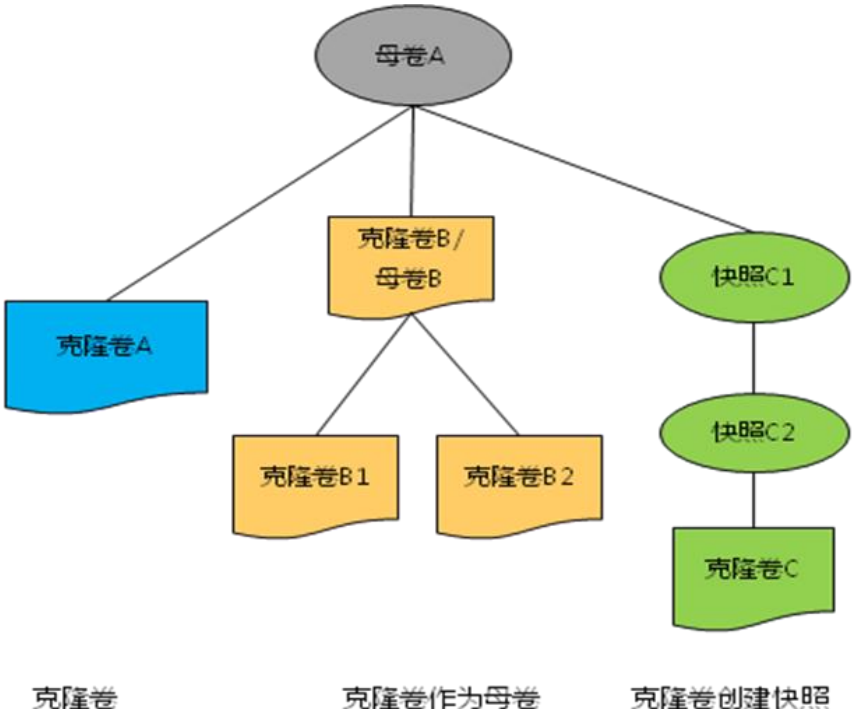
4.5.5 链接克隆

分布式存储提供链接克隆机制，支持基于一个卷快照创建出多个克隆卷，各个克隆卷刚创建出来时的数据内容与卷快照中的数据内容一致，后续对于克隆卷的修改不会影响到原始的快照和其他克隆卷。

支持 1:256 的链接克隆比，提升存储空间利用率。

克隆卷继承普通卷所有功能：克隆卷可支持创建快照、从快照恢复以及再次作为母卷进行克隆操作。

图 4-16 分布式存储 链接克隆



4.5.6 多资源池

为了满足使用不同性能存储介质以及故障隔离，分布式存储支持多资源池特性。一套分布式存储 Manager 管理多个资源池。多个资源池共用同一个分布式存储集群，包括 Zookeeper 和主 MDC 都是共用的。每个资源池有归属 MDC，创建资源池时会自动启动 MDC 为其归属资源池，资源池最大规模为 128，MDC 控制在 96 个，超过 96 资源池后会指定已有 MDC 为其归属 MDC，每个 MDC 最多管理两个资源池。资源池的归属 MDC 负责该资源池的初始化，初始化对存储资源进行 Partition 划分，并将 Partition 与 OSD 的视图存储到 ZK 盘。当资源池的归属 MDC 故障后，主 MDC 会为该资源池指定一个托管 MDC。

支持在多个资源池间做离线卷迁移。

多资源池规划遵循以下原则：

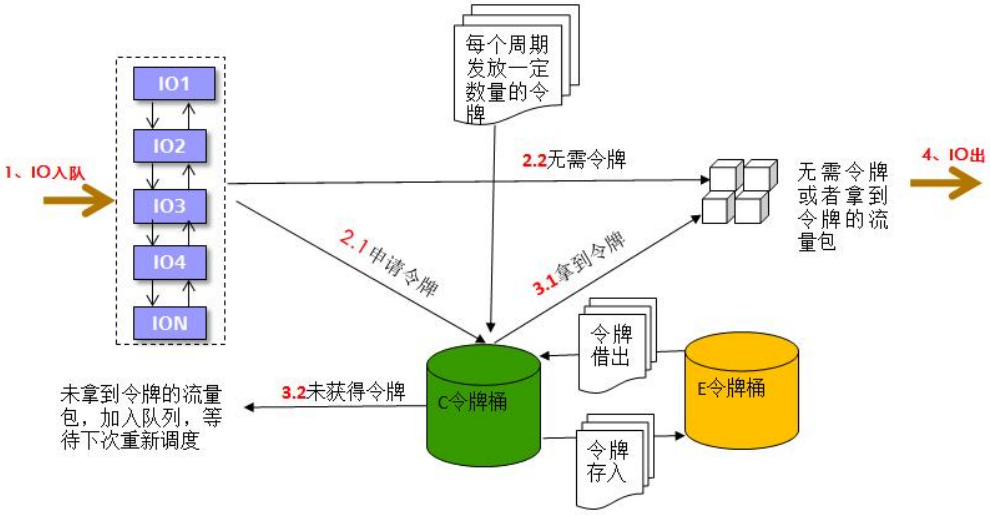
- 资源池可以为副本（2 副本或者 3 副本）或者 EC（支持 2+2、4+2、6+2、8+2 等冗余比）。
- 同一资源池的硬盘类型一样，为不同类型的硬盘分别规划资源池；同一资源池的硬盘容量规格要求一样，否则都按小容量规格的硬盘使用。
- 同一资源池的缓存介质一样，为不同类型的缓存介质分别规划资源池。
- 资源池创建时，各存储节点的硬盘数最好相同。允许相差不超过两个，同时硬盘差不能超过硬盘最大数的 33%。
- 同一服务器可有不同类型的硬盘，创建资源池时支持将同一服务器不同类型的硬盘纳入不同的资源池。

4.5.7 QoS

分布式存储 QoS 提供了对卷的 I/O 精细化控制并提供 burst 功能（所谓 burst 功能，是指当卷的需求超出了基准 IOPS（带宽）时，允许在一定时间内使用超出基准性能的配额）。

分布式存储 QoS 采用双令牌桶的算法实现对卷的 I/O 控制、带宽控制以及 burst 功能。

图 4-17 分布式存储 QoS 机制示意图



其中，C 令牌桶用来控制 IO，E 令牌桶用来存放令牌余额，两个桶配合来实现 burst 功能。

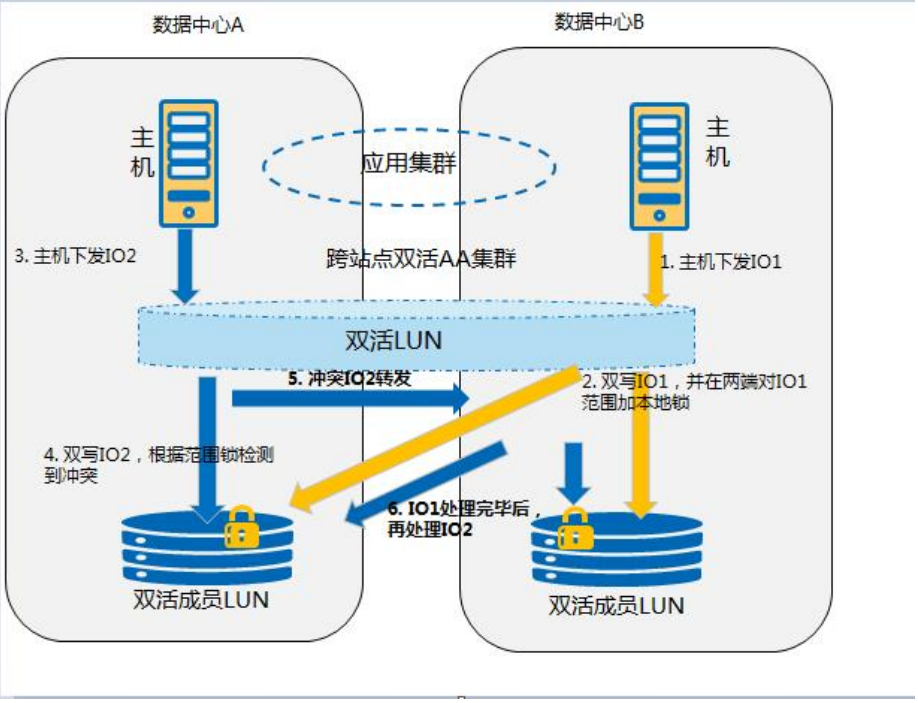
4.5.8 存储双活

基于 AB 两个站点的两套分布式存储集群构建双活容灾关系，基于两套分布式存储的卷虚拟出一个双活卷，两站点业务的主机能同时进行读写服务。任意站点故障，数据零丢失，业务能迅速切换到另外一个站点运行，保证业务连续性。

在原有基础服务基础上，引入复制集群，按服务化的架构提供双活业务。

支持优先站点仲裁和第三方仲裁的双仲裁模式，故障自动倒换，无需人工介入。

图 4-18 分布式存储双活原理

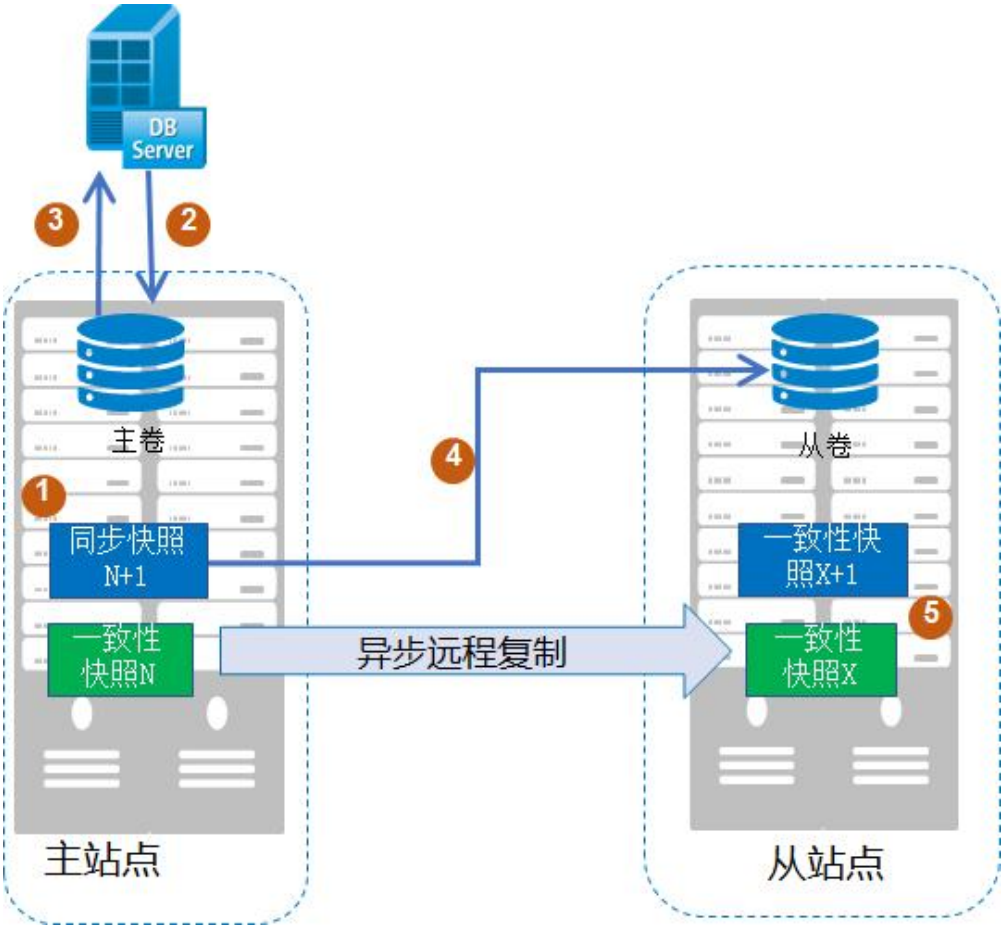


4.5.9 存储异步复制

FastCube 系统的存储异步远程复制采用了对比快照获取差异的方式，其实现原理如下：

- 当主站点的主卷和远端从站点地从卷建立异步远程复制关系以后，首次启动为全量同步，主站点对主卷创建一个同步快照，通过全量拷贝快照将主卷数据全量拷贝到从卷；
- 初始同步完成后，主站点的同步快照角色变换为一致性快照，从卷数据状态变为“一致”，并且创建一个一致性快照（即从卷的数据为主卷启动全量同步时刻的一致性拷贝），然后开始按照下面的流程进行 I/O 处理：

图 4-19 分布式存储异步复制流程示意图



图中序号说明：

a. 每个复制周期启动时在主卷创建一个新的同步快照（主端新建的同步快照为 N+1，主端上个同步周期结束后的一致性快照为 N，从端上个同步周期结束后产生的一致性快照为 X）；

b. 主机新写入的数据写入主卷；

c. 响应主机写完成；

d. 主站点通过对比一致性快照 N 和同步快照 N+1 获取差异，读取快照 N+1 的数据将差异数据写入从卷中；

e. 同步完成后，主端删除老的一致性快照 N 并将新的同步快照 N+1 角色转换为一致性快照，从端创建新的一致性快照 X+1，并删除老的一致性快照 X。

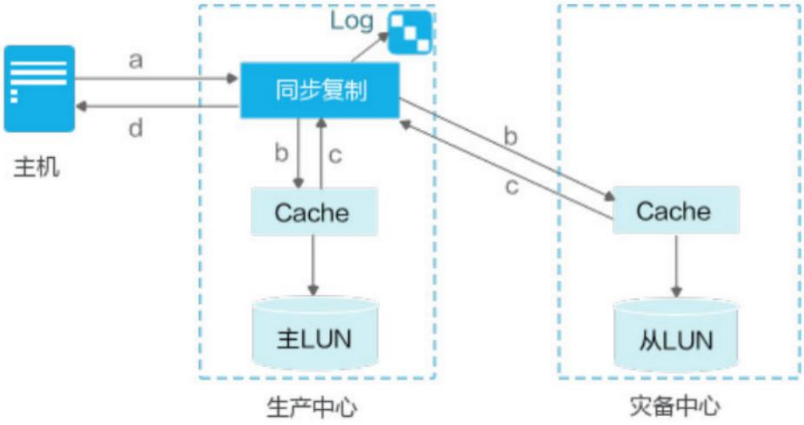
- 每间隔一个同步周期（由用户设定，范围为 150s～1440min），系统会自动启动一个将主站点数据增量同步到从站点的同步过程（如果同步类型为手动，则需要用户来触发同步）。

4.5.10 存储同步复制

结构化同步远程复制对于每个主机地写 I/O，都会同时写到主 LUN 和从 LUN，直到主 LUN 和从 LUN 都返回处理结果后，才会返回主机处理结果。因此，同步远程复制可以 实现 RPO 为 0，其实现原理如下： 1. 初始同步：生产中心的主 LUN 和灾备

中心地从 LUN 建立同步远程复制关系，启动 初始同步。 - 将主 LUN 数据全量拷贝到从 LUN。 - 初始同步中主 LUN 收到主机写请求也会同样写到从 LUN。 2. 双写状态：初始同步完成以后，进入正常状态，此时主、从 LUN 数据相同。正常 状态下的 I/O 处理流程如下：

图 2-3 同步远程复制示意图



a. 生产存储收到主机写请求。同步远程复制将该请求记录日志。日志中只记录地址信息，不记录数据内容。

b. 将该请求写入主 LUN 和从 LUN。通常情况下数据会写入 Cache。

c. 同步远程复制等待主 LUN 和从 LUN 地写处理结果都返回，如果写从 LUN 超时 或失败时则同步远程复制关系断开。如果都写成功，清除日志；否则保留日志，写入 DCL 中（Data Change Log 数据变更日志）元数据所在持久化空间 中，进入异常断开状态，后续启动同步时重新复制该日志地址对应的数据 块。

d. 返回主机写请求处理结果，以写主 LUN 的处理结果为准，如果主 LUN 写失败，即使从 LUN 写成功，仍然返回主机为失败。

3. 单写状态：用户下发分裂命令、复制链路断开、I/O 双写失败等均会使远程复制进 入单写状态。

- 同步远程复制单写状态下主机 I/O 写到主端 Cache 即返回主机。
- Cache 数据刷盘时需要记录 DCL 差异。

5 硬件设备平台

5.1.1 机架服务器

形态	2U 机架服务器
处理器	2 个 x86 系列 CPU：
内存插槽	最大 32 个 DDR4 DIMM 插槽，最高 3200MT/s
硬盘数量	12 个 3.5 英寸 SATA 硬盘 25 个 2.5 英寸 SAS 硬盘
RAID 支持	支持 RAID1
PCIe 扩展	支持 1 个 RAID 控制扣卡专用的 PCIe 扩展槽位， 8 个标准的 PCIe 扩展槽位

6 安装部署和运维管理

FastCube 系统在安装运维方面致力给用户提供部署简单、运维便捷的体现，实现一站式交付、系统部署当天业务上线，傻瓜式运维，降低对 IT 管理人员的技能要求。其中安装部署方面，提供生产预安装以及自动化安装部署工具 FusionCube Builder；运维管理方面提供了统一的运维管理平台 FusionCube Vision 管理工具。

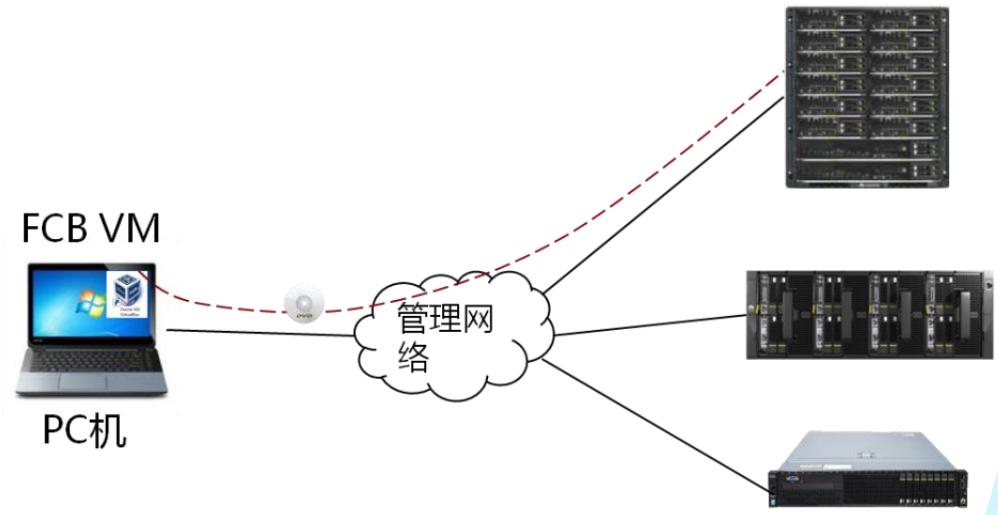
6.1 自动化部署

FastCube 提供 FusionCube Builder（简称 FCB）快速安装部署工具完成系统软件的安装。同时 FastCube 支持一键式系统初始化功能，只需要完成系统基本参数配置后，则自动完成各节点网络配置，管理集群和存储集群的创建。

6.1.1 FusionCube Builder

FusionCube Builder（简称 FCB）是为满足现场快速安装部署 FastCube 系统需求而开发的安装工具。

图 6-1 FCB 安装系统软件示意图

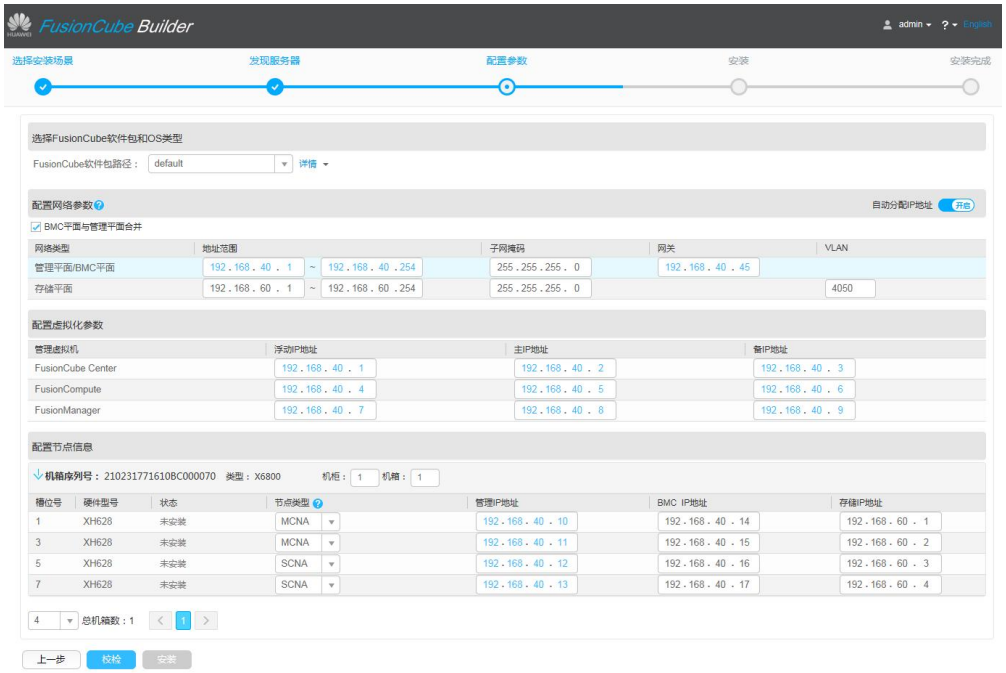


- FCB 安装工具可部署在 PC 机或者虚拟机中。
- FCB 通过 Simple Service Discovery Protocol（SSDP）简单服务发现协议或者扫描 IP 方式发现服务器，读取服务器信息。
- FCB 连接服务器 BMC，使用 KVM 挂载光盘功能进行引导，启动安装。最大支持 8 个节点并行安装。

- 安装过程中使用 NFS 共享传送安装软件包和配置。

FCB 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，FCB 根据相关的安装配置快速自动完成系统软件的安装。

图 6-2 FCB 安装向导和配置界面



6.1.2 系统初始化

首次登录 FusionCube Vision 管理系统时，支持管理系统管理 IP 地址修改和系统初始化功能。

系统初始化流程如下：

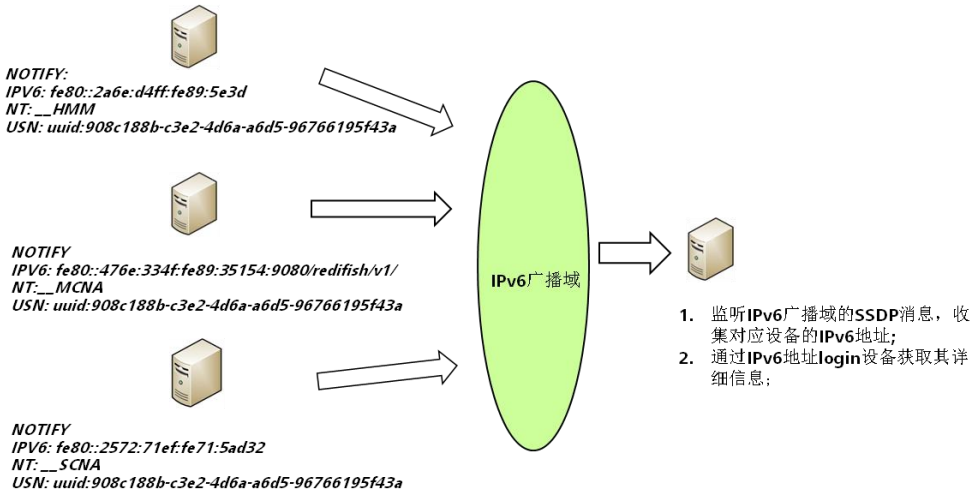
1. 进入系统初始化界面，系统自动发现设备（包括物理节点和管理虚拟机），并在初始化参数配置页面显示节点相关信息。
2. 完成初始化参数（包括网络参数和存储池参数）配置，参数校验通过后则可进行系统初始化构建。
3. 系统自动完成初始化，包括各节点的网络配置，管理集群和存储集群、存储池的创建。FusionCompute 场景还支持计算集群创建，并将主机加入计算集群

系统完成初始化后即具备 VM 发放功能，只需要完成设备上行网络、NTP 等配置即可交付使用。

6.1.3 设备自动发现

FastCube 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 6-3 SSDP 设备自动发现示意图



FCB 系统安装过程中的 SSDP 设备自动发现流程如下：

1. FastCube 使用到的服务器 BMC 中内嵌了 SSDP 协议，设备上电后会自动通过 IPv6 地址广播 SSDP 消息。

2. FCB 中部署了 SSDP 服务端，负责监听 IPv6 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，获取相应设备的详细信息。

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 VM、CVM 以及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. FusionCube Vision 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 地址登录设备，获取相应设备的详细信息。

6.2 统一运维管理

FastCube 通过 FusionCube Vision 管理系统实现整个系统的统一管理，功能包括资源管理、性能监控、告警管理、操作日志管理、权限管理、硬件管理、健康检查和日志收集。

FusionCube Vision 包括如下主要页面：

- 首面

站点的系统监控仪表盘，包括了告警、容量、性能、健康度、任务等详细信息，以及部分常用操作的快捷键；

- 资源
  - 包括虚拟化、存储、网络等资源监控管理；

- FusionCompute 场景下，“虚拟化”页签下支持创建管理虚拟机，“存储”页签支持创建虚拟机卷设备以及挂载操作，“网络”页签支持查看虚拟化网络配置，“数据库”主要支持混合场景下数据节点的相应操作；

- 硬件

监控管理站点的硬件设备，包括：机箱、服务器、交换机。可查看硬件设备的节点类型、型号、IP 地址以、硬件信息以及资源的监控；

- 监控

站点的告警以及性能监控，其中告警包括系统的各个部件的所有告警清单、告警的设置以及统计信息；性能包括系统的历史性能数据和性能 TOP 统计；

- 系统
  - 站点的系统配置、权限管理、任务与日志、系统维护；
  - 系统配置提供时间管理、管理数据备份、邮件服务器配置、eService 配置、SNMP 配置、桌面云接入和系统超时时间；
  - 权限管理主要是用户和角色串讲管理，密码策略和域认证信息；
  - 系统维护主要提供一键式运维功能：扩容、健康检查、日志收集。

6.2.1 业务发放管理

当前只支持在 FusionCompute 场景下支持虚拟机相关业务的发放管理特性，包括：虚拟机发放管理、磁盘创建管理以及网络端口组管理。

虚拟机发放管理

FusionCompute 场景下，FusionCube Vision 管理平台提供了虚拟机发放管理特性，提供了虚拟机的创建以及常用的日常操作特性，包括：虚拟机上下电、重启关闭，虚拟机迁移，虚拟机导出导入，虚拟机规格调整，性能监控，快照等管理以及虚拟机模板管理等特性；

磁盘管理

虚拟机磁盘管理，提供了虚拟机卷设备创建、绑定虚拟机等操作，系统可提供：普通、共享卷设备，支持 IDE、VIRTIO、SCSI 的接口类型，系统默认提供的卷设备为瘦分配卷，可以有效地提升系统的磁盘利用率。

网络管理

网络管理主要为提供虚拟机发放中需要的网络资源，主要为 vlan、端口组以及 MAC 地址。FusionCube Vision 提供了 VLAN 池、端口组、MAC 池的创建配置等功能，分布式交换机（DVS）暂只支持查看系统中已有的 DVS，创建管理需要跳转至 FusionCompute 虚拟化平台上进行操作。

6.2.2 一键式运维

一键式运维功能作为 FastCube 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，主要提供了扩容、健康检查、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，导致在健康巡检、日志收集以及升级等一键式运维操作中，无法对服务器硬件设备进行运维操作。

一键式扩容

FastCube 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。FusionCompute 场景下，节点出厂即已安装好节点上系统。节点扩容前准备好后，在 FusionCube Vision 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，即可点击进行**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，即点击**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。

一键式日志收集

在 FusionCube Vision 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；

- 可支持收集日志项包括：硬件（BMC、SSD、NIC 等组件）、分布式存储、FusionCompute、FusionCube Vision 等系统组件的相关日志项。
- 日志收集一次收集的时间短暂只支持 2 天时间的日志文件，支持并发收集节点日志。

在 FusionCube Vision 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。

一键式健康巡检

在 FusionCube Vision 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。整个的监控检查包括系统的检查和硬件兼容性检查两部分，系统检查主要排查分布式存储、虚拟化系统、硬件的健康状态；硬件兼容性检查主要排查系统的硬件的版本、驱动和固件版本是否符合当前版本的要求。

在 FusionCube Vision 管理平台上的健康检查页面，选择需要巡检的部件以及节点，然后点击检查，对系统进行健康排查。

6.2.3 Call Home

FusionCube Vision 管理平台可以配置将系统的告警通过 eServer 对接将告警发送至元亿，主要通过将告警信息发送至华为邮箱，然后元亿运维中心接收到局点环境的告警信息后，判断用户的系统是否需要立马进行问题故障处理，然后通知客户安排相应的运维人员进行故障排查处理。

7 性能和可扩展性

7.1 系统高性能

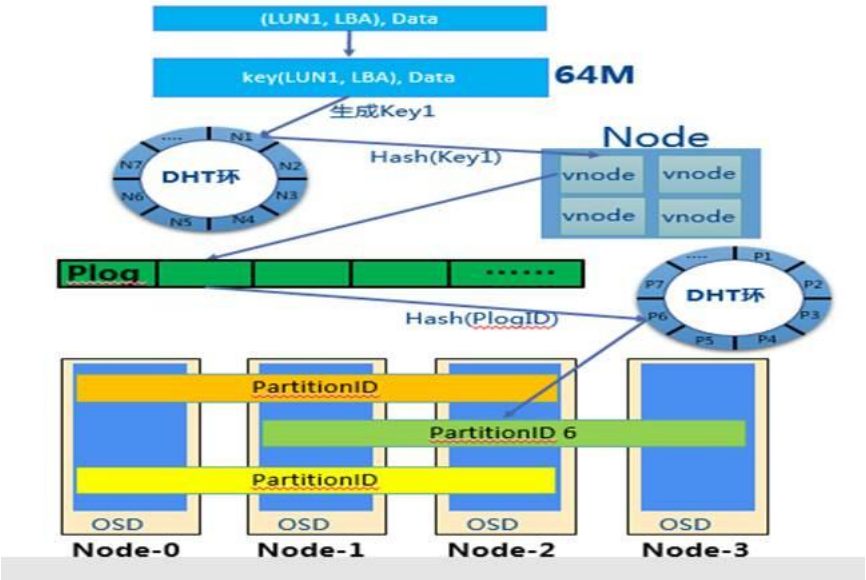
分布式存储采用全分布式大资源池架构，无集中接入组件和模块，消除单个组件/模块瓶颈导致的性能不足；采用分布式哈希数据路由算法，将业务数据在资源池中所有盘中进行均衡打散存放，不会出现单个组件/模块成为热点， 同时资源池所有的都可用作资源池的热备盘，在组件/硬件故障时能够快速恢复，同时能保障业务性能的一致性。在分布式存储中，将各个存储节点上的 SSD 组建成为一个共享的分布式 SSD Cache 资源池，提供给系统所有业务共同所用，有效地提升系统的 IO 性能。

同时，FastCube 在系统中引入了性能更高的 NVME SSD 固态硬盘。

7.1.1 分布式 I/O 环

分布式存储采用 DHT 路由技术，实现从业务 I/O 快速定位到硬盘应该存放的具体位置，避免在海量数据中进行查找和计算，该 DHT 路由技术，采用华为自研算法，不仅能保证数据在各个硬盘的均衡性，而且在硬件增减（故障或扩容）时，自动快速调整，并保证数据迁移的有效性，确保自动快速自愈，自动资源均衡的目的：

图 7-1 分布式存储数据路由示意图



- 第一层 DHT hash 环的目的是通过 hash 算法将数据分发到计算出来的存储服务器节点处理该数据，通过该 hash 算法，确保每个数据都有对应的服务器节点来处理，保证了业务处理的均衡。系统根据 LUNID 和 LBA 定位到服务器节点，然后再定位到该服务器上的 vnode 上，由该 vnode 逻辑处理单元来处理该数据；vnode 是一种逻辑处理单元，将物理服务器节点分为 4 个逻辑处理单元，即 4 个 vnode，例如：一个由 6 个物理服务器组成的一个存储集群，当其中 1 个物理服务器故障时，该服务器上的 4 个 vnode 处理的业务，可以分别被该集群中另外的 4 个物理服务器去接管，这样剩下的 5 个物理服务器中，有 4 个物理服务器运行有 5 个 vnode，1 个物理服务器运行 4 个 vnode，通过 vnode 机制，可以确保故障节点的业务可以分散到不同的服务器节点上去接管，就可以防止只用一个物理服务器接管带来的业务处理瓶颈问题。该 DHT hash 环打散粒度是按 64MB 对齐打散。

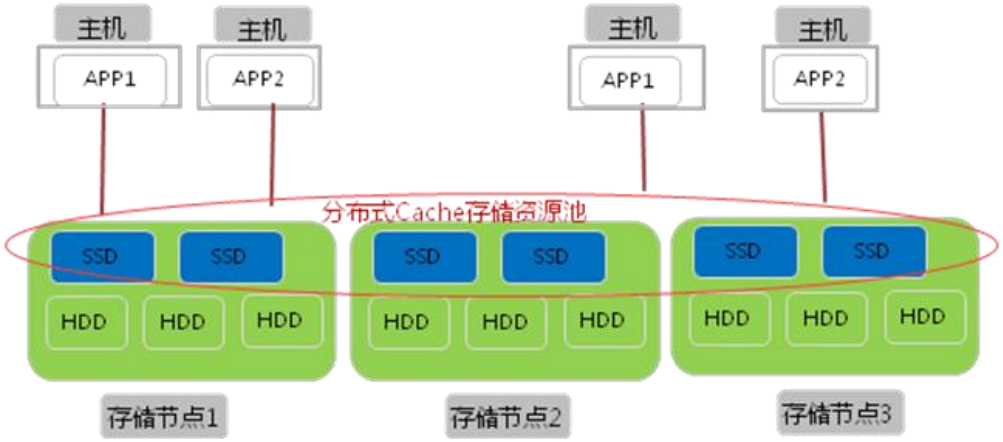
- 第二层 DHT hash 环的目的是通过 hash 算法将数据转到对应存储空间去保存，完成数据的持久化。通过该 hash 算法，确保数据存储空间的均衡性。系统根据 PlogID 和 Offset 定位到硬盘应该存放的具体位置，避免在海量数据中进行查找和计算，该 DHT 路由技术，采用华为自研算法，不仅能保证数据在各个硬盘的均衡性，而且在硬件增减（故障或扩容）时，自动快速调整，并保证数据迁移的有效性，确保自动快速自愈，自动资源均衡。

7.1.2 分布式 SSD Cache 加速

传统的 HDD 受机械原理的影响，虽然在容量上有比较大的增长，但在性能方面，几十年来基本上没有任何变化，随机 I/O 时延从几毫秒到十几毫秒，严重影响用户体验和性能的发挥。而 SSD 虽然相对于 HDD 有很大的提升，但是价格比较贵。目前业界使用 SSD 作为系统 Cache 或 Tier 层，实现了性能和成本之间的平衡。

华为将分布到各个存储节点上的 SSD 组建成为一个共享的分布式 Cache 资源池，供所有的业务共同所用。这样充分利用所有 SSD 的资源。

图 7-2 分布式存储分布式 Cache 逻辑架构图



7.1.2.1 Read/Write Cache

Read Cache

FusionStorage 块存储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，SSD Cache 采用热点解读机制，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。

OSD 在收到 VBS 发送的读 I/O 操作时，会进行如下步骤处理：

步骤 1 从内存“Memory Write Cache”中查找是否存在所需 I/O 数据，如果存在，则直接返回，同时调整该 IO 数据到“读 Cache”LRU 队首，否则执行步骤 2；

步骤 2 从内存“Memory Read Cache”中查找是否存在所需 IO 数据，如果存在，则直接返回，同时增加该 IO 数据的热点访问因子，否则执行步骤 3；

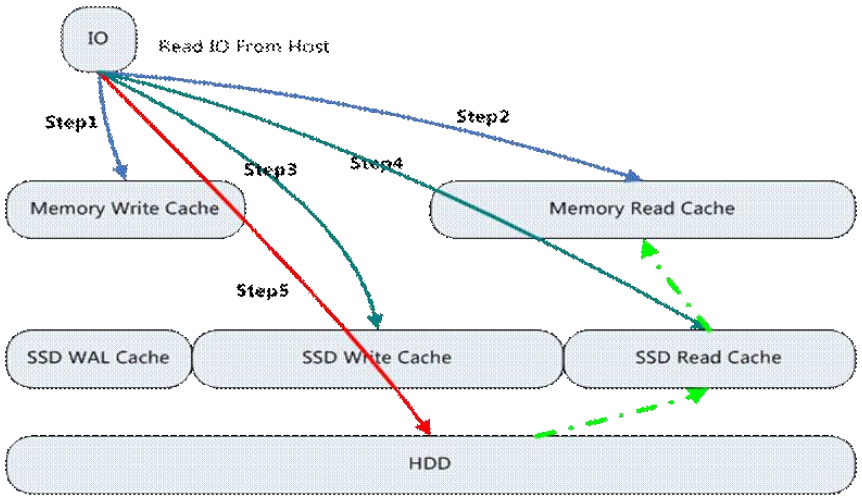
步骤 3 从 SSD 的“SSD Write Cache”中查找是否存在所需 IO 数据，如果存在，则直接返回，如果不存在，执行步骤 4；

步骤 4 从 SSD 的“SSD Read Cache”中查找是否存在所需 IO 数据，如果存在，则直接返回，同时增加该 IO 数据的热点访问因子；如果热点访问因子达到阈值，则会被缓存在 SSD 的“SSD Read Cache”中，如果不存在，执行步骤 5；

步骤 5 从硬盘中查找到所需 IO 数据并返回，同时增加该 IO 数据的热点访问因子，如果热点访问因子达到阈值，则会被缓存在 SSD 的“SSD Read Cache”中。

---一结束

图 7-3 Read Cache 示意图



Write Cache

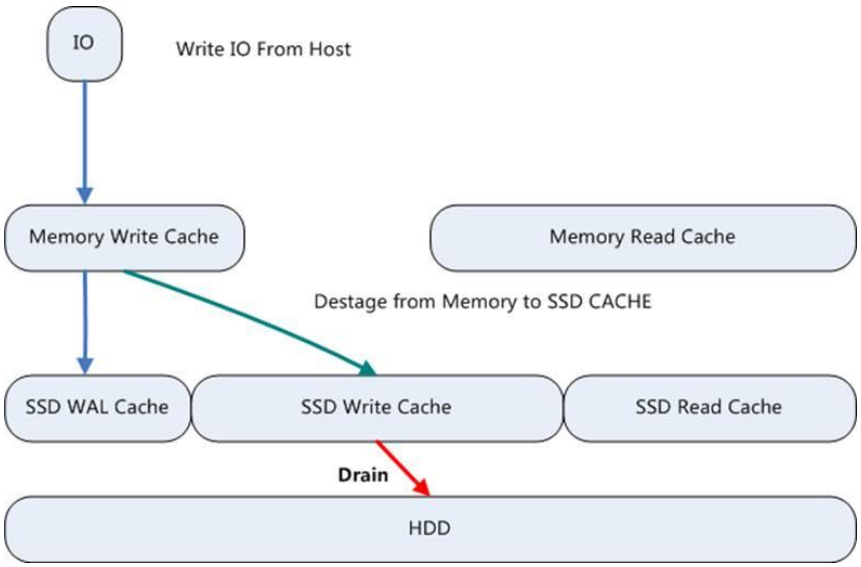
VBS 发送地写 IO 操作（图中 Write IO From Host）时，会将 Write IO 在 Memory Write Cache 内存中保存一份，同时同步以日志的方式（采用固定的 2+2 小分片 EC）记录到 SSD WAL Cache 中并返回成功完成本次写操作，这个流程通常称为 Host Write IO 流程。

通常 SSD Disk Cache 分为两个部分：SSD Write Cache 和 SSD Read Cache。Memory Write Cache 中的数据会进行 IO 排序重整并等待满分条以副本或 EC 的方式直接写入到 SSD Write Cache 中并返回；对于大块 IO 则直接由 Memory Write Cache 直通写到 HDD 中，而不驻留在 SSD Write Cache 里；

当 SSD Write Cache 中的保存数据水位达到 40%时，则由 SSD Write Cache 往 HDD 中搬迁。

随着 Memory Write Cache 中的数据逐步刷盘到 SSD Write Cache 时，SSD WAL Cache 中的数据将逐步淘汰掉，我们通常会进行异步的垃圾回收。

图 7-4 Write Cache 示意图



相比较传统的副本方式写入 SSD Cache，然后异步地再从 SSD Cache 中读出满分条到持久化存储层 HDD，FusionStorage 的 SSD WAL Cache 方案带来 4 大优势：

- FusionStorage 的 SSD WAL Cache 地写放大比较小，2+2 的 EC 的 Overhead 为 2；而副本方式的 SSD Cache，OverHead 最低必须为 2。
- 由于写放大较小，FusionStorage 对网络的带宽消耗也较低
- FusionStorage 的 SSD WAL Cache 可靠性高，是+2 的冗余保护。
- FusionStorage 的数据往主存上刷盘通常是由 RAM 中触发完成的，比传统的后台异步先从 SSD Cache 读出再写到主存中的效率高。

7.1.2.2 大块 Pass Throught

下面是不同介质的性能对比数据，对于随机小 I/O，SSD 比 HDD 存在几十到上百倍的性能优势，但是顺序 I/O 来说，优势其实并不明显。

表 7-1 性能对比数据

介质类型	4k 随机写 IOPS	4k 随机读 IOPS	1M 写带宽	1M 读带宽	平均时延
SAS	180	200	150MB	150MB	3~5ms
NL-SAS	100	100	100MB	100MB	7~8ms
SATA	100	100	80MB	80MB	8~10ms
SSD 盘	70K	40K	500MB	500MB	<1ms
SSD 卡	600K	800K	2GB	3GB	<1ms

- HDD 盘的性能数据是关闭 HDD 写 Cache 的数据，对于组建存储系统的 HDD 来说，必须关闭 HDD 地写 Cache，确保可靠性。
- HDD 盘原理相差并不大，不同厂商的性能有变化，但是变化并不大，一般在 10%之内。

- SSD 盘和 SSD 卡相差比较大，上面只是以一款作为示例。SSD 盘带宽性能无法提高的主要原因是受制于 SAS/SATA 接口带宽（测试使用目前最常用的 6Gb/s SATA 接口）。
- 从上面的性能数据看，对于小块随机 IOPS， SSD 对 HDD 存在明显的性能优势，但是在大块顺序 IO 的场景下，虽然 SSD 卡存在较大带宽优势，但 SSD 盘受 SAS/SATA 接口带宽的影响，和 HDD 相比，优势并不明显。考虑到一个 SSD 盘会同时会给多个硬盘作为 Cache 使用，当一个 SSD 盘同时给超过 5 个 HDD 作为 Cache 时，直接操作 HDD 反而性能会更高。
- 分布式存储支持大块 I/O 直接 bypass SSD Cache，直接操作 HDD，由此带来如下好处：

  大块 IO 性能会更佳。释放原来大块 I/O 占用的 Cache 空间，可以缓存更多的随小块 I/O，变相提高了随机小块 I/O 的 Cache 命中率，提升系统整体性能，提高写 I/O 操作次数，提升 SSD 卡使用寿命。

7.1.3 硬件加速

固态 SSD 磁盘/卡

分布式存储支持为高性能应用提供全闪存 SSD 存储池方案，支持华为 ES3000 SSD 设备，包括 NVME SSD 盘，SAS SSD 磁盘，提供比传统的机械硬盘（SATA/SAS）更高的读写性能。

ES3000 V6 NVME SSD 性能指标如下：

- 大容量：1.6TB/3.2TB
- 高 IOPS：读 IOPS 达 900K@4KB；写 IOPS 达 200K@4KB
- 高带宽：PCIe4.0 x4 接口，顺序读带宽高达 7000MB/s
- 采用华为自研控制芯片 Hi1812，多个逻辑硬件协同工具，降低读写时延

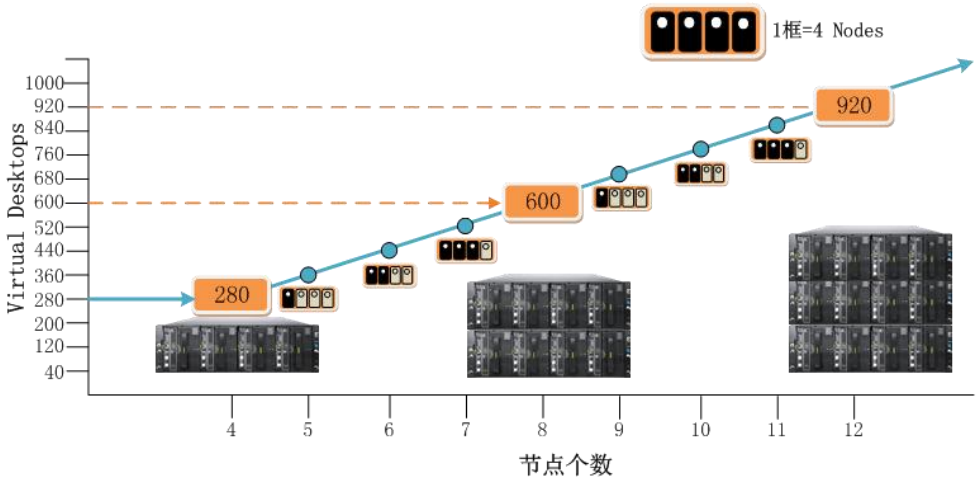
ES3000 V6 SAS SSD 性能指标如下：

- 大容量：960GB/1.92TB/3.84TB/7.68TB
- 高 IOPS：读 IOPS 达 200K@4KB；写 IOPS 达 150K@4KB
- 高带宽：读宽高达 1000MB/s；写带宽高达 1000MB/s
- 采用华为自研控制芯片 Hi1812，多个逻辑硬件协同工具，降低读写时延

7.2 线性扩展

FastCube 系统具有良好的扩展性，至少 2 个节点起配，可通过增加服务器/节点的方式实现系统的扩展，单集群最大支持 1024 个节点。

图 7-5 FastCube 系统扩展示意图

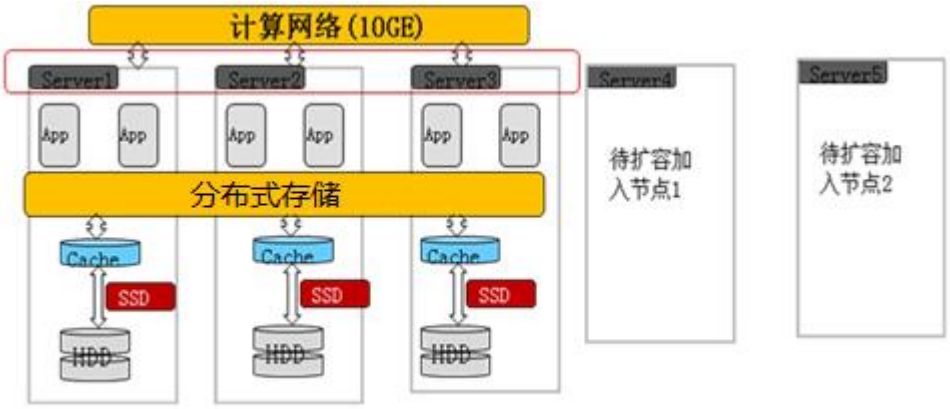


7.2.1 存储平滑扩容

分布式存储的分布式架构具有良好的可扩展性，支持超大容量的存储：

- 扩容存储节点后不需要做大量的数据搬迁，系统可以快速达到负载均衡状态。
- 支持灵活的扩容方式，可以独立扩容计算节点、硬盘、存储节点，或者同时进行扩容。在扩容计算节点时同步扩容存储空间，扩容后的系统仍旧可以使计算和存储融合。
- 软件机头、存储带宽和 Cache 都均匀分布到各个节点上，系统 IOPS、吞吐量和 Cache 随着节点的扩容而线性增加。

图 7-6 存储平滑扩容示意图



7.2.2 性能线性扩展

分布式存储通过创新的架构把分散的 SATA 机械硬盘组织成一个高效的类 SAN 存储池设备，提供比 SAN 设备更高的 I/O，把性能发挥到了极致。

分布式机头

分布式存储采用无状态的分布式软件机头，机头部署在各个节点上，无集中式机头的性能瓶颈。单个节点上软件机头只占用较少的 CPU 资源，提供比集中式机头更高的 IOPS 和吞吐量。例如：假设系统中有 20 台节点需要访问 FusionStorage 提供的存储资源，每台节点提供给存储平面的带宽为 2*10Gb，我们在每台节点中部署 1 个 VBS 模块（相当于在每台节点中部署 1 个存储机头），20 台节点意味着可部署 20 个存储机头，所能获取到的总吞吐量最高可达 20*2*10Gb=400Gb，随着集群规模的不断扩大，可以线性增加的存储机头，突破了传统的双控或者多控存储系统集中式机头的性能瓶颈。

分布式缓存

- 分布式存储缓存和带宽都均匀分布到各个节点上。
- 分布式存储集群内各节点的硬盘使用独立的 IO 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。
- 分布式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，分布式存储仍然可以发挥很高的 IO 性能，整体性能提升 1~3 倍。
- 分布式存储支持 SSD 用作数据缓存，除具备通常地写缓存外，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。

全局负载均衡

分布式存储的 DHT 机制可以保证上层应用对数据的 I/O 操作会均匀分布在不同节点的不同硬盘上，不会出现局部的热点，实现全局复负载均衡。

- 系统自动将每个卷的数据块打散存储在不同节点的不同硬盘上，冷热不均的数据会均匀分布在不同的节点上，不会出现集中的热点。
- 数据分片分配算法保证了主用副本和备用副本在不同节点和不同硬盘上的均匀分布，换句话说，每块硬盘上的主用副本和设备副本数量是均匀的。
- 扩容节点或者故障减容节点时，数据恢复重建算法保证了重建后系统中各节点负载的均衡性。

7.2.3 一键式扩容

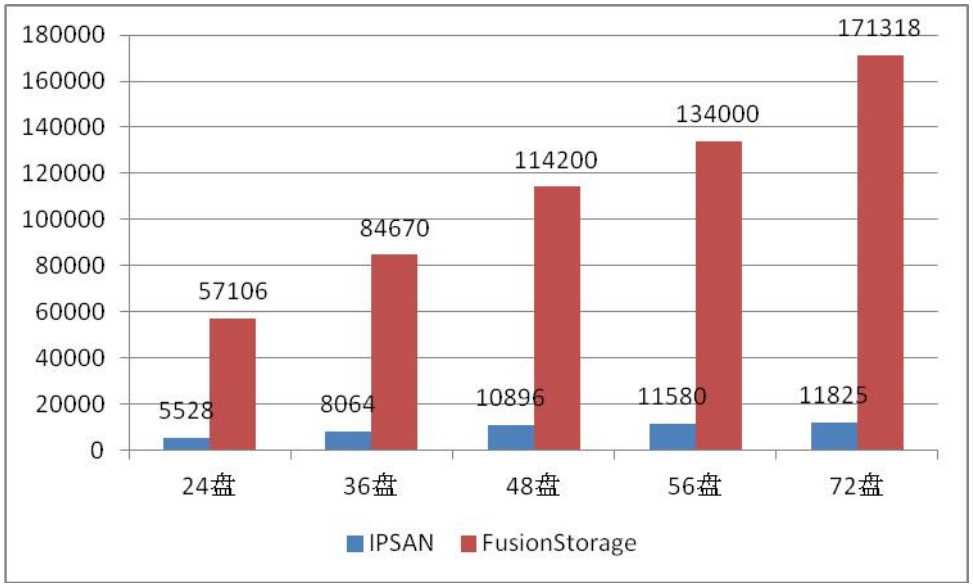
FastCube 提供一键式扩容功能，只需简单的操作即可完成系统扩展，大大简化系统扩容操作：

- 在 FusionCube Vision 的系统扩容界面点击扩容按钮，进入扩容操作。
- 系统自动发现设备，并在界面统一显示。
- 完成网络和存储相关参数并检验成功后，点击扩容按钮进行系统扩容。
- 系统根据配置自动完成所有的扩容配置，包括节点的网络配置以及将节点加入存储集群，扩展存储池或者新建存储池。

7.3 分布式存储相对于传统 SAN 的性能优势

7.3.1 更高的性能

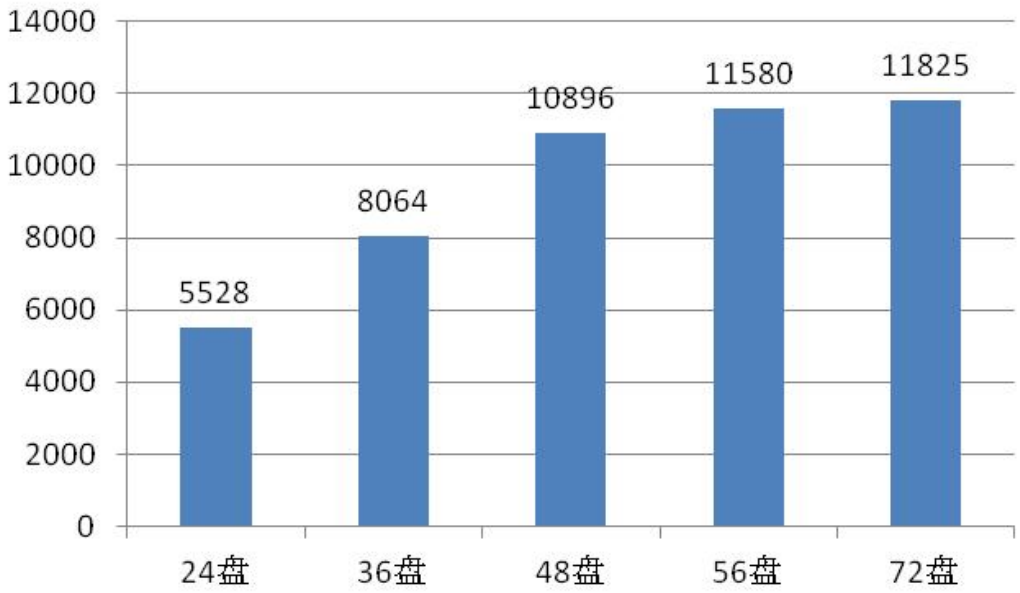
图 7-7 FastCube 与 IP SAN 对比测试



在和 IP SAN 对比测试中，在同样的测试条件下，FastCube 相比 IP SAN 存在十倍以上的性能。随着支持的盘数越多，性能优势越明显。

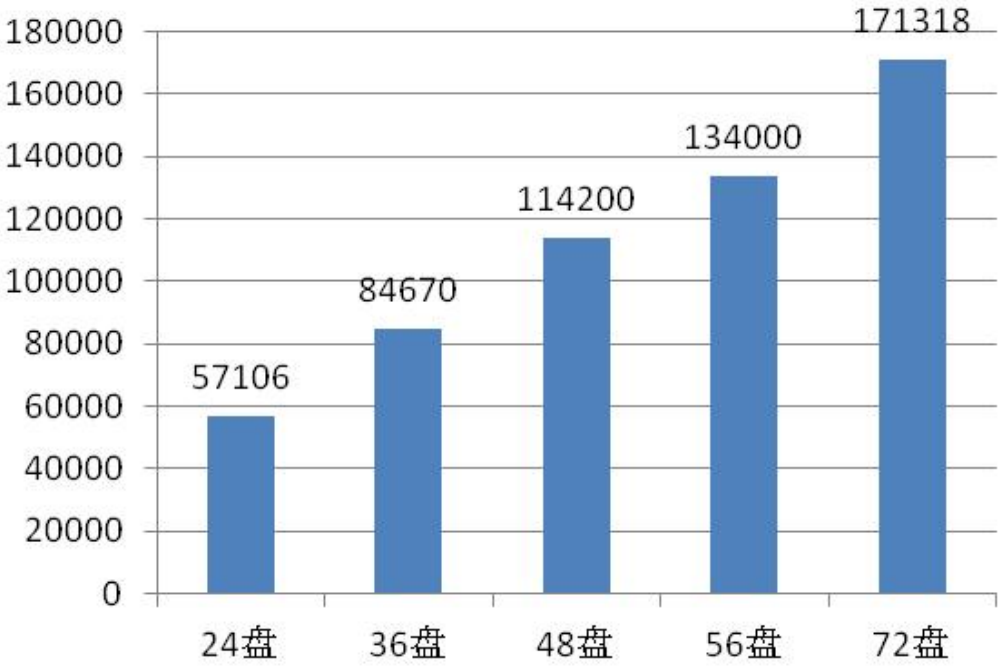
7.3.2 线性 Scale-up/Scale-out

图 7-8 IP SAN 扩容性能情况（iops）



大多数 IP SAN 只能 Scale-up，不支持 Scale-out，即使在 Scale-up 的情况下，也很难保证线性。从上图可以看出，IP SAN 在 48 盘以下基本上可以保持线性，但是随着盘数的增加，受制机头处理能力的影响，很难保证能够线性扩展。因此 IP SAN 在高性能配置时，一般不会按照系统最大能够携带的磁盘框进行配置。

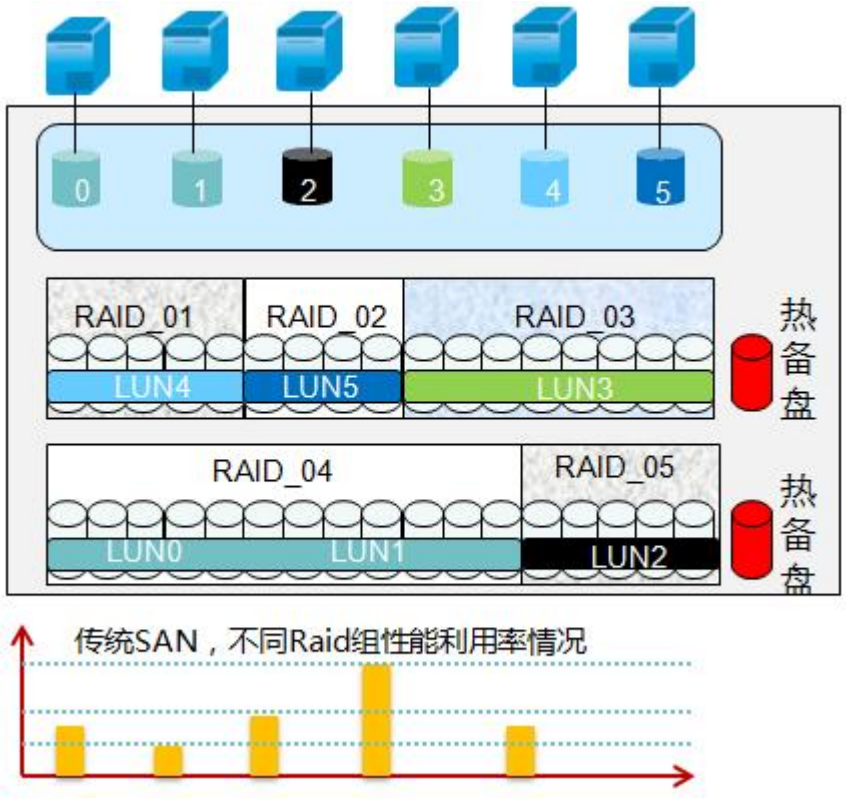
图 7-9 FastCube 扩容性能情况（iops）



分布式存储系统随着盘数的增加，系统性能线性提升。

7.3.3 大池 POOL

图 7-10 传统 IP SAN RAID 示意图



传统 IP SAN 对外提供业务前，需要组建不同的 RAID 组，然后在 RAID 上划分不同的 LUN 对外提供业务。受可靠性影响，一般 RAID 组能够容纳的盘比较少，因此单个 RAID 组对外提供的业务性能有限，这样就导致如下问题：

业务规划复杂，调整困难：当业务性能在原来 RAID 组上无法满足时，需要将该 LUN 迁移到另外一个 RAID 组上。特别当出现系统总体性能还比较富裕，但是任何一个 RAID 组都无法满足该 LUN 性能要求时，会需对多个 RAID 组，多个 LUN 进行调整。这种调整不仅对业务正常运行带来风险，而且很多情况下，即使大范围的调整也无法满足性能要求，系统性能浪费严重，不同 RAID 组性能无法共享使用，如上图，有些 RAID 组性能要求低，有些 RAID 组性能要求高，会造成系统整体系统资源虽然有很大冗余，但是业务无法使用的情况，导致浪费存在如下几种情况：

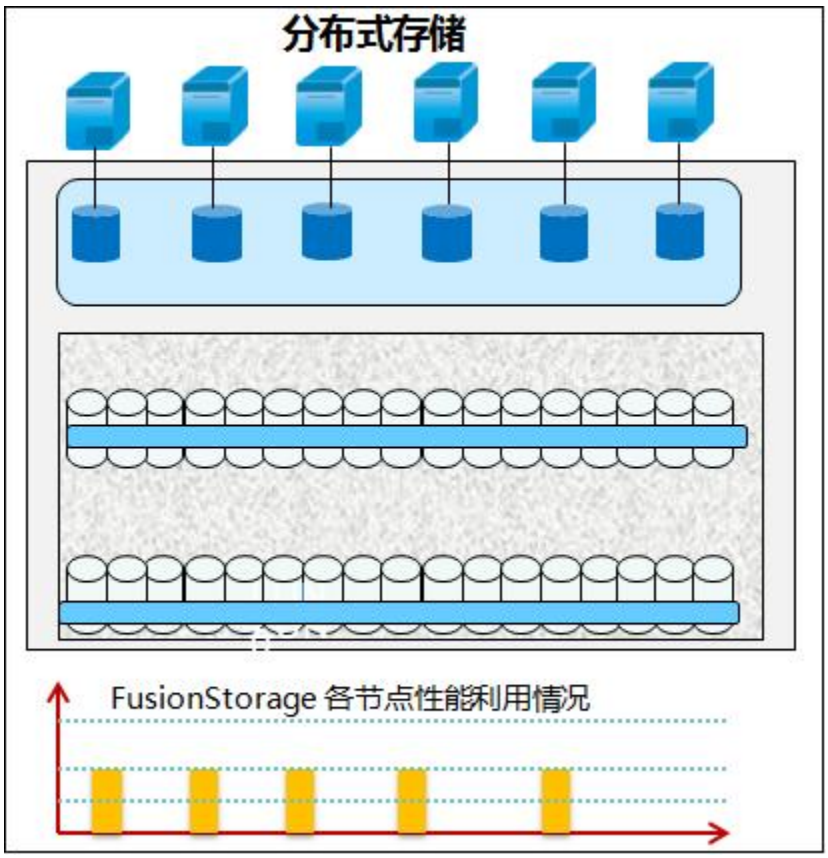
- 不同业务性能要求不同。

用户需要按照业务进行规划不同的 RAID 组，不同的 RAID 组不能性能共享，导致性能浪费。

- 同一类业务在不同时刻，对性能要求不同。

系统需要按照该 RAID 组最大性能进行规划，而这个最大性能要求，可能只占一天/周/月很短的时间，但还是必须按照最大性能进行规划。大多数情况下闲置资源无法得到使用。

图 7-11 分布式存储大资源池架构示意图



分布式存储由于采用大资源池，所有的硬件资源都参与到任意一个业务，用户只要保证系统整体性能和容量满足要求，就可以直接增加业务（新业务或原有业务提升性能），不需要额外进行性能规划和调整。面对 IP SAN 遇到的问题，分布式存储能轻松应对，不仅使系统性能利用最大化，而且最大程度地降低用户维护投入，降低业务中断的风险。

7.3.4 SSD Cache vs SSD Tier

传统 SAN 使用机头内的内存作为存储 I/O 的 Cache，但是内存大小有限，一般配置 8G/16G/32G，这和所处理的业务需求相差非常大，特别是对于复杂混合多业务以及性能要求高的场景，机头内存 Cache 能够缓存的内容非常少，很难真正发挥明显的作用。因此传统存储厂商也相继启用 SSD 盘对存储业务进行加速，但是大多数存储厂商将它作为 SSD Tier 而不是 SSD Cache。SSD Tier 对于稳定单一，热点固定且热点持续时间长的业务比较有效，已经无法满足当前多业务，热点变化快的场景。

分布式存储利用 SSD 作为 Cache 层，能及时感知业务热点变化，快速反应，能持续保障高性能。

表 7-2 SSD Cache 与 SSD Tier 对比

描述项	SSD Cache	SSD Tier
价值	将热点数据放到高速介质（SSD 卡，SSD 盘），提	将热点数据放到高速介质（SSD 卡，SSD 盘），提

描述项	SSD Cache	SSD Tier
	升存储系统处理性能	升存储系统处理性能
数据变化	只是将热点数据从 HDD 移到 SSD 上，并不在 HDD 上删除数据。当热点数据变冷时，只是释放 SSD 空间	在将热点数据搬迁到 SSD 时，删除 HDD 中的内容，当热点数据变冷，除释放 SSD 空间外，还会将该数据写回到 HDD 中
容量	SSD 只是作为 Cache，并不会增加系统总容量	SSD 作为 Tier，会增加系统对外提供的总容量空间
SSD 空间利用率	因热点数据在 HDD 中有备份，在 SSD 中不需要使用 RAID 等可靠性技术保障可靠性。利用率高	因热点数据从 HDD 中删除，数据搬迁到 SSD 时，需要进行使用 Raid 技术保障可靠性
SSD 性能利用率	利用率高。搬迁到 SSD 中的数据直接写入，无写惩罚。热点变冷后，可以被新热点直接覆盖	要使用 Raid 可靠性技术，搬迁到 SSD 卡时会遇到 Raid 组地写惩罚。 热点数据变冷后，需要重新读出写入到 HDD 后，才能被其他新数据覆盖
热点统计周期	非常短，几分钟。能及时捕捉到热点变化，快速反应。	非常长， 一般几个小时。反应较慢
数据管理块大小	小块，一般为 8K、16k、32K。Cache 利用的效率非常高，浪费少	大块，一般为 1M、2M、4M。Cache 利用空间浪费比较大。
适用场景	热点变化比较大，无相对固定热点场景 混合业务，尤其是业务热点变化比较大	热点相对稳定，热点持续时间长的场景 单一业务，热点稳定性比较高的场景

8 系统可靠性

分布式存储系统提供了数据跨节点的保护能力在多个硬盘或者节点故障时也能够继续提供服务，将数据放置到同一个节点池内不同节点的不同硬盘上，数据获得了跨节点的可靠性和故障快速恢复的能力。同时通过硬件的冗余配置提供系统的可用性。

8.1 数据可靠性

8.1.1 块存储集群可靠性

分布式存储采用集群管理方式，从架构上保证了系统不会出现单点故障，一个节点或者一块硬盘故障自动从集群内隔离出来，不影响整个系统业务的使用。具体为：

- Zookeeper

Zookeeper 也称作 ZK。为 MDC 提供选主仲裁。ZK 还存储系统初始化时产生的元数据，包括“分区－硬盘”映射关系等数据路由信息。一个系统部署奇数个 Zookeeper 组成集群，最少部署 3 个，必须保证大于总数一半的 Zookeeper 处于活跃可访问状态。一旦系统部署不能再扩容 ZK 数量。

- MDC

MDC 为元数据控制软件，实现对分布式集群的状态控制。系统初始至少部署 3 个 MDC 模块，增加资源池自动为该资源池启动或指定一个 MDC。多个 MDC 利用 Zookeeper 选举一个主 MDC，主 MDC 监控其他 MDC，发现 MDC 故障则重启 MDC 或为资源池指定托管 MDC。当主 MDC 故障时，通过选举产生新的主 MDC。

- OSD

OSD 为对象存储设备服务，执行具体的 IO 操作。采用主备模式，MDC 实时监控 OSD 的状态，当指定 Partition 所在的主 OSD 故障时，存储服务会实时自动切换到备 OSD，保证了业务的连续性。

每个节点上有多个 OSD，分别管理节点上的磁盘或者 SSD 虚拟磁盘。OSD 跟磁盘或者 SSD 虚拟磁盘一一对应，但不跟磁盘或者 SSD 绑定，支持节点内任意一个存储磁盘/SSD 交换位置，可防止维护时的误操作，提升系统可靠性。

8.1.2 数据一致性

FastCube 底层分布式存储采用华为 Pacific 分布式存储软件。Pacific 采用多种技术手段，包括强一致性复制协议、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。

强一致性复制协议

Pacific 采用强一致性复制协议来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 Pacific 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，Pacific 会暂时不写这个副本，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，Pacific 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，再通过重建机制使得数据在各个磁盘上的分布均匀。

读修复技术

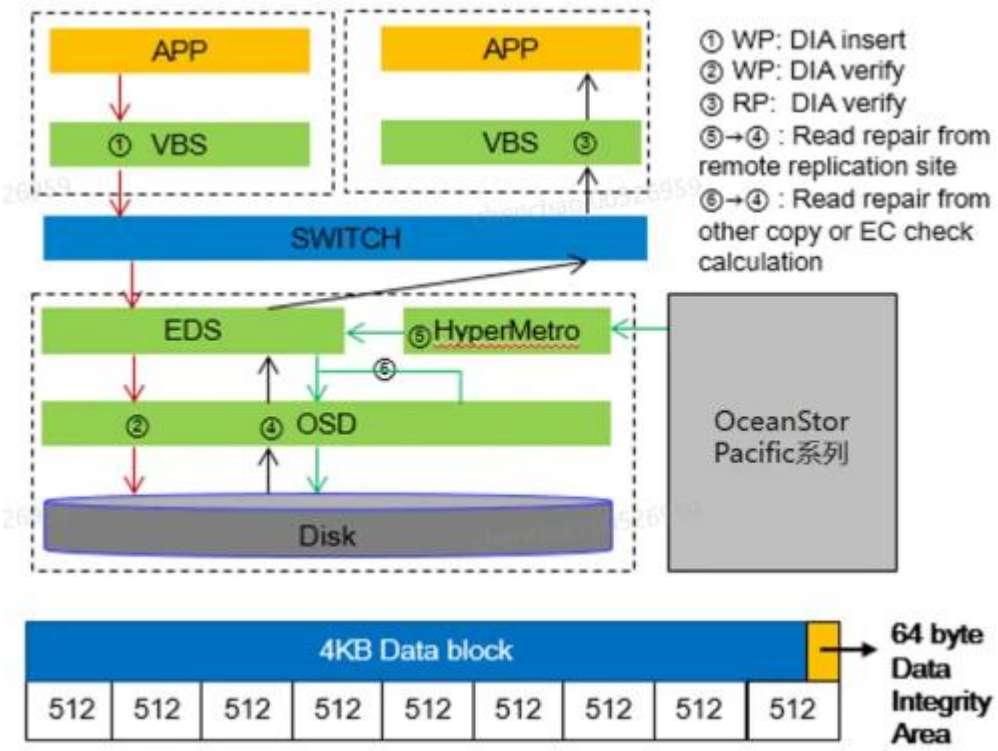
Pacific 还实现读修复（Read Repair）技术，读修复机制是指在读数据失败时，系统会判断错误类型，如果是磁盘扇区读取错误，系统会自动从其他节点保存的副本读取数据，然后重新写入该副本数据到硬盘扇区错误的节点，从而保证数据副本总数不减少和副本间的数据一致性。

数据完整性保护

Pacific 块存储系统通过 IO 实时端到端数据完整性校验、后台周期性数据校验以及损坏数据实时自愈纠错机制来解决静默数据破损场景。

Pacific 提供 IO 极端到端的数据完整性保护方案，能够有效检测跳变、读写偏等各种静默数据破坏场景，当检测到数据静默破坏后会实时对数据进行纠错自愈，避免数据损坏扩散。如下图展示了 IO 路径关键静默数据错误检测位置，使用 CRC32 保护用户 4KB 数据，除此外支持主机 LBA 校验，盘 LBA 校验等。

图 8-1 FusionStorage 数据校验



对于存储介质由于器件老化、电磁/信号干扰、工艺缺陷等原因导致静默数据破坏。通过周期性数据校验可以提前识别风险并进行处理，能有效防止静默数据破坏累积导致数据丢失。FusionStorage 块存储系统采用了后台自适应周期性校验方式来防止数据出现错误。无论在主机 IO 还是后台周期性 IO 识别到静默数据破坏时，均会触发自动的用户无感知的损坏数据纠错自愈机制。

另外系统还提供后台周期性数据校验，系统会周期性从存储介质上读取数据，并校验是否存在静默错误数据，如果存在，则会立即触发后台修复流程进行纠错自愈；此后台任务会自动适应业务压力，当主机业务压力较大时会以较慢速率运行，当主机业务压力较小时会以较快速率运行。

8.1.3 数据冗余保护

分布式存储支持两种数据冗余保护机制，一种是多副本方式，一种是 Erasure Code（EC，纠错码）方式。

分布式存储多副本冗余保护机制当前只支持 2、3 副本两种机制，其中两副本冗余策略支持系统随意故障一块数据盘或故障一个节点，系统存储运行正常。当前两副本只支持主存为 SAS 盘或 SSD 盘场景，且单个存储池最大支持 96 块盘，确保系统单个存储池的高可靠性；三副本冗余策略支持系统随意故障两块数据或故障两个节点（配置 5 个 ZK 元数据管理节点），系统存储运行正常。当前三副本可支持版本兼容的所有硬盘类型，单个存储池可支持 2048 块盘。在分布式存储副本机制中，版本推荐使用 3 副本，提供更高的可靠性。

分布式存储的 EC 冗余策略，在提供高可靠性的同时也能够提供相对副本更高的利用率，且系统的性能会相对副本不会有较明显的下降，甚至在大块场景下有一定提升。

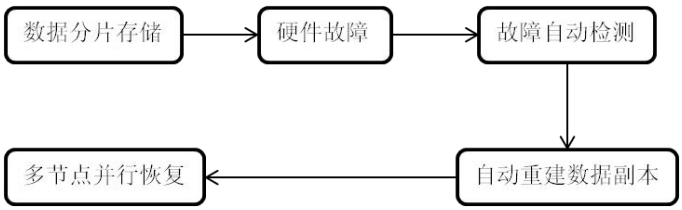
FastCube 当前支持的 EC 配比类型包括：2+2，4+2（：1），6+2（：1），8+2（：1），其中版本默认推荐配比为 4+2。分布式存储在 EC 特性实现中采用了数据拐弯的算法，可支持在 3 节点上部署 4+2（：1）配置的 EC 冗余策略，采用数据拐弯算法后，系统可靠性还是能保障同时故障两个磁盘，但不能同时故障两个节点，可靠性相对正常的 4+2 配置会有一定下降。

8.1.4 快速数据重建

分布式存储的每个硬盘都保存了多个数据块（Partition），这些数据块的副本按照策略分散在系统中的其他节点。当分布式存储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。

数据故障自动重建流程如图 8-2 所示。

图 8-2 分布式存储数据重建流程图



分布式存储支持并行、快速故障处理和重建：

- 数据块（Partition）及其副本分散在整个资源池内，硬盘故障后，可在资源池范围内自动并行重建。
- 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。
- HDD 场景下，1TB 数据重构最少需要 30min；SSD 场景下，1TB 数据重构最短仅需要 15min 即可。

8.2 硬件可靠性

FastCube 选用高可靠的硬件服务器，通过系统冗余设计保证系统可靠性，具有如下特点：

- 具有掉电可保数据的缓存备电设计，保证掉电数据安全。
- 采用可热拔插 SAS 专用系统盘，支持 RAID1 保护。
- 整机冗余电源、风扇设计，保障系统可用性。
- 网络双平面设计。
- ES3000 NVME SSD 盘/卡支持 DIF，提供数据完整性校验。

8.3 系统亚健康增强

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。

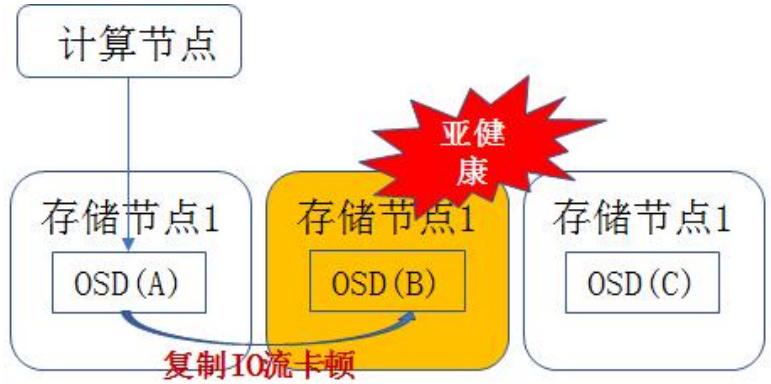
- 当单个 IO 时延过大时，为保证业务 IO 快速恢复，判断此 OSD 处于亚健康状态。

切换机制：

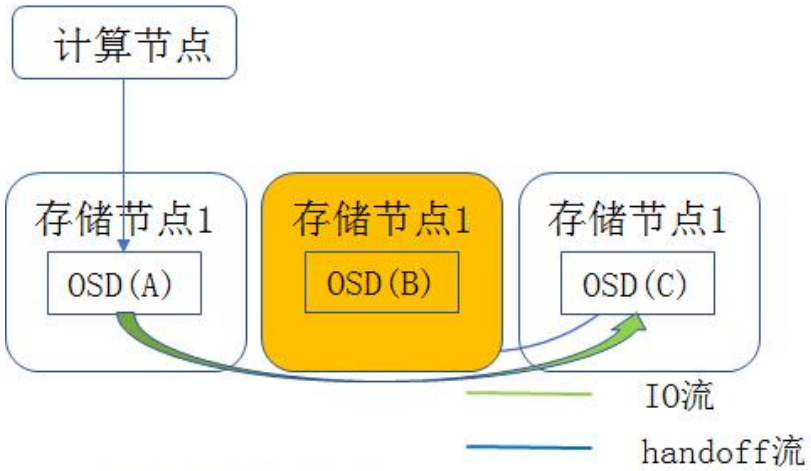
OSD 集群控制节点对判断出的亚健康 OSD，生成临时视图将其服务切换到其他 OSD（Hint 节点），主节点的复制 IO 写到 Hint 节点后即可返回。

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）

- **复制集群节点亚健康检测与处理**

检测机制：

系统在访问复制节点时有检测机制，检测访问复制节点的 IO 路径时延，上报时延相关信息到复制主节点。

复制主节点收集各节点上报的访问复制节点时延信息，当超时 IO 数量占 IO 总数的比例较高时认为复制节点处于亚健康状态。

隔离机制：

复制主节点对诊断出的亚健康复制节点，在隔离节点数量不超过限制的情况下进行隔离处理。

恢复机制：

复制主节点在等待一段时间以后，会重新加入已被隔离的复制节点。

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：

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。

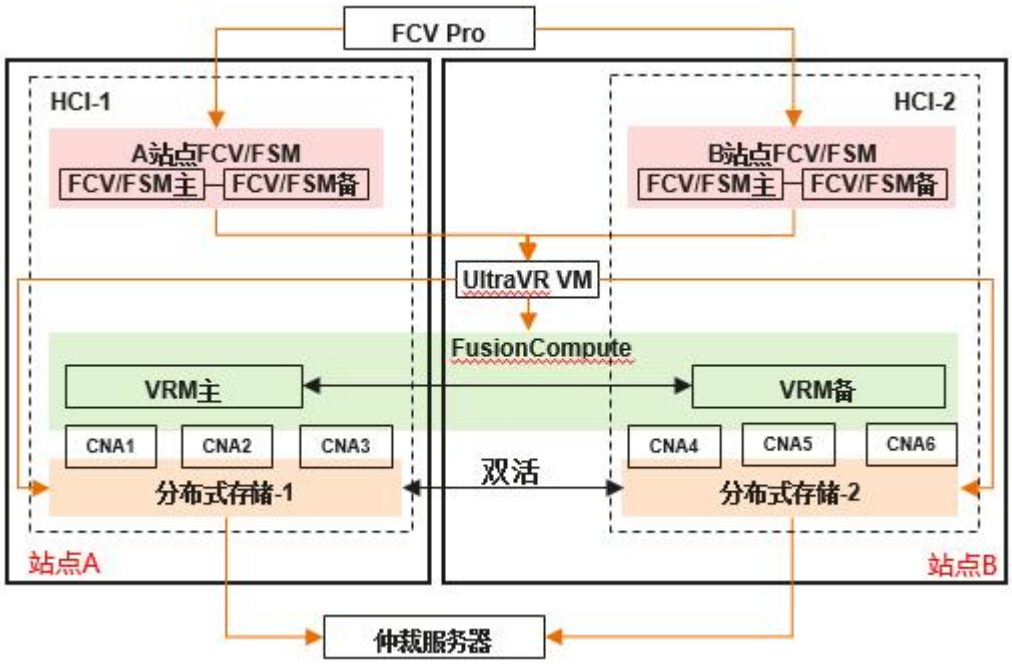
8.4 容灾恢复

针对业务的连续性，企业需要考虑如何在发生自然或人为灾难、操作员出错或是应用出现故障的情况下，保护数据并快速进行业务恢复。为了应对这些挑战，除了本地备份，还需要一个有效的方式将数据发布到远程位置。如果没有有效的数据保护和远程发布措施，可能会导致大量的收入损失。容灾系统正是解决以上挑战的解决方案，据客户对于业务宕机时间以及灾备中心的距离等差异，FastCube 提供了基于存储双活技术的端到端双活解决方案和基于存储异步复制技术的主备站点冗余解决方案。

端到端双活解决方案可以在两个站点间实现负载均衡和灾难自动切换，提供更高的资源利用率。双活方案中，数据是在同步双写到两个站点成功后再返回成功的，所以系统可提供近乎等于 0 的 RTO 和 RPO 时间。

基于存储异步复制的容灾方案，由于采用的数据异步同步刷新至灾备站点存储池中，所以可以支持在异地部署或者站点传输时延大的场景适用。

8.4.1 虚拟化高可用解决方案（FusionCompute）



FastCube 1000 虚拟化高可用方案是基于 FusionCompute 虚拟化的 HA 特性和分布式存储的双活特性构建的端到端方案：

FastCube 1000 虚拟化高可用方案由以下组件构成：

- FusionCube Vision：负责端到端的虚拟化高可用管理，包括接入、创建、删除、查看等

- FusionCompute：负责计算资源的提供以及计算资源的高可靠保证，如主机部件故障、主机故障、存储故障等
- 分布式存储：负责已裸盘数据资源的高可靠保证，如单节点故障、集群故障、网络故障等。
- 仲裁服务器：提供分布式存储双活的仲裁。
- UltraVR：负责容灾配置，被 FusionCube Vision 调用实现容灾配置
- FusionCube Vision Pro：负责跨站点的资源管理。

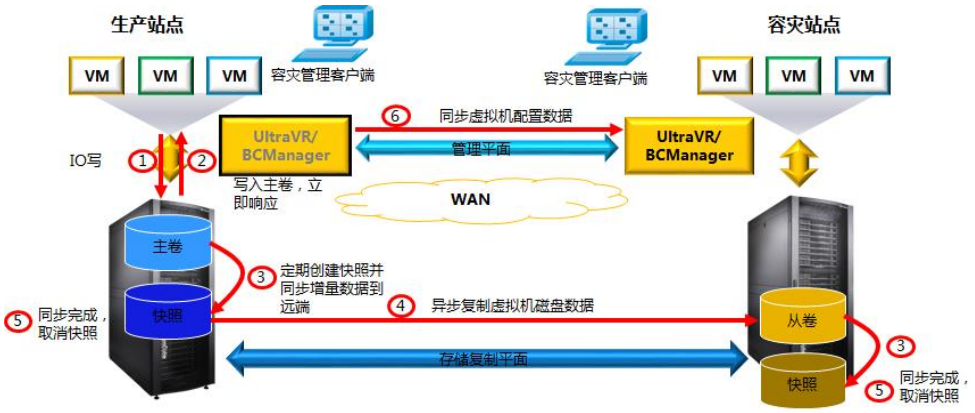
分布在不同地域或者机房的两个站点，通过部署两套 FastCube 并配置虚拟化高可用功能实现虚拟机级别的跨站点保护，其中站点 A 通过 FusionCube Builder 标准部署模式部署，站点 B 通过 FusionCube Builder 的 B 站点部署模式部署，部署过程中，站点 B 自动将站点 A 的 FusionCompute 集群延伸至站点 B；允许站点 A 运行后，再部署站点 B。

当两个站点部署完成后每个站点可见一套完整的 FastCube，用户单虚拟机可部署在站点 A 或站点 B 中任意一个，当发生故障时，虚拟化高可用会自动判断本站点内有无满足条件节点恢复业务以保证最佳体验，当本站点无满足条件时则在对端站点恢复业务，保证业务连续性，同时对于多个虚拟机，可均匀分布在站点 A、站点 B 上合理利用资源、就近部署，使站点 A 和站点 B 达到负载均衡的工作。

FastCube 1000 虚拟化高可用常见故障场景处理如下：

- 分布式存储实时检测分布式工作状态，基于分布式存储的双活机制将 A、B 站点数据做双向同步保证数据一致性，当 A、B 站点中发生网卡、线路、服务器部件、服务器整机、存储多节点、存储集群故障时通过冗余与仲裁机制快速切换到备用卡件、线路、服务器、站点中，实现存储数据的实时可用
- FusionCompute 实时检测计算侧集群工作故障，基于 FusionCompute 的 DRS 和 HA 机制，当 A、B 站点中发生网卡、线路、服务器部件、服务器整机故障时，FusionCompute 自动调用策略进行处理，优先选择本主机的备用网卡、线路、部件；如本主机不可用，则自动选择本站点可用服务器资源；如本站点资源不可用则选择跨站点资源进行切换。

8.4.2 异步复制解决方案



FastCube 1000 异步复制架构是基于两套 FastCube 的分布式存储构建异步复制关系，叠加上层 UltraVR 或 BCManager 容灾管理软件构成整套的异步复制解决方案。存储异步复制通过快照对比方案，周期性的同步主、从卷的数据，上一次同步周期数据同步完成后，到当前同步周期开始，主卷上产生的所有数据会在本次同步时写到从卷上。

用户能按需部署存储容灾集群，该集群是提供复制服务的逻辑对象，用于管理容灾集群内集群节点、集群元数据、复制 pair、一致性组、数据搬移操作。容灾集群和系统业务存储共同部署在存储节点上。容灾集群具有良好弹性扩展能力，单容灾集群 3 节点起步、最大 64 个节点，整系统最大支持 8 个集群；单容灾集群支持 64000 个卷和 16000 个一致性组，满足后续快速增长的容灾业务需求。

容灾管理软件 UltraVR 和 BCManager 以应用视角管理容灾业务，对 FastCube 系统以业务虚拟机维度进行保护；流程化引导容灾业务配置，包括一键式容灾测试、容灾保护策略、主站点故障恢复操作等。其中 UltraVR 配套 FastCube 的 FusionCompute 场景提供异步复制能力。

8.4.3 同步复制解决方案

FastCube 1000 同步复制相对于异步复制，对于每个主机地写 I/O，都会同时写到主 LUN 和从 LUN，直到主 LUN 和从 LUN 都返回处理结果后，才会返回主机处理结果。因此，同步远程复制可以 实现 RPO 为 0

基于两套 FastCube 的分布式存储构建同步复制关系，叠加上层 UltraVR 或 BCManager 容灾管理软件构成整套的同步复制解决方案。存储同步复制通过快照对比方案，周期性的同步主、从卷的数据，上一次同步周期数据同步完成后，到当前同步周期开始，主卷上产生的所有数据会在本次同步时写到从卷上。

用户能按需部署存储容灾集群，该集群是提供复制服务的逻辑对象，用于管理容灾集群内集群节点、集群元数据、复制 pair、一致性组、数据搬移操作。容灾集群和系统业务存储共同部署在存储节点上。容灾集群具有良好弹性扩展能力，单容灾集群 3 节点起步、最大 64 个节点，整系统最大支持 8 个集群；单容灾集群支持 64000 个卷和 16000 个一致性组，满足后续快速增长的容灾业务需求。

容灾管理软件 UltraVR 和 BCManager 以应用视角管理容灾业务，对 FastCube 系统以业务虚拟机维度进行保护；流程化引导容灾业务配置，包括一键式容灾测试、容灾保护策略、主站点故障恢复操作等。其中 UltraVR 配套 FastCube 的 FusionCompute 场景提供同步复制能力。

9 系统安全

9.1 系统安全威胁

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、泪滴攻击等。

- 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 或者即时通信进行。

- 零日攻击

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

来自内部网络的安全威胁

- 攻击方法日新月异，内部安全难以防范

内网 ARP 欺骗与恶意插件滥用问题等将产生新的安全威胁。被攻破的内网主机，容易被攻击者作为“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，也容易受到攻击。

- 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

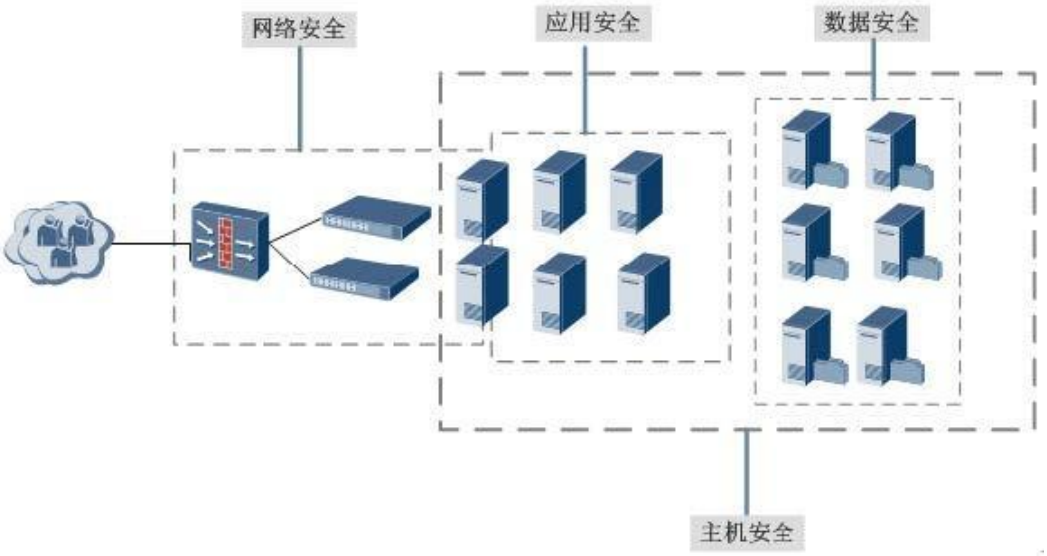
U 盘、光驱、打印、红外、串口、并口等外设，由于使用方便，已成为数据泄密、病毒感染的出入口。通过封贴端口、制度要求等方式无法灵活对外设进行管理，特别是对 USB 接口的管理，因此，需通过其他技术手段解决存在的问题。

- 管理制度缺乏技术依据，安全策略无法有效落实

9.2 总体安全框架

依据系统面临的安全威胁和风险，FastCube 产品提供安全解决方案，如图 9-1 所示。FastCube 安全框架通过网络、主机、应用以及数据四个维度上来保证系统的安全性。

图 9-1 FastCube 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 FastCube 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

从集群容灾、备份、数据完整性、数据包密性等方面保证用户数据的安全。

9.2.1 网络安全

FastCube 的网络通信平面划分为业务平面、存储平面和管理平面：

- 业务平面

为用户提供业务通道，为虚拟机虚拟网卡的通信平面，对外提供业务应用。

- 存储平面

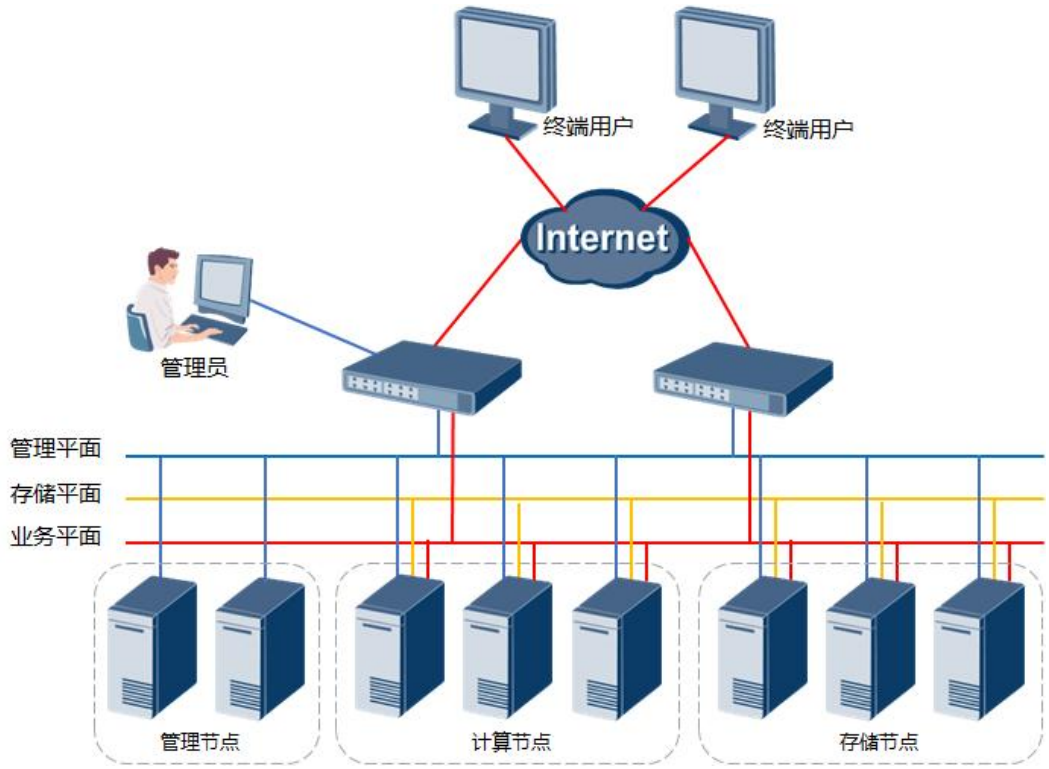
为虚拟机提供访问存储资源的通信平面，但不直接与虚拟机通信，而通过虚拟化平台转化。

- 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 9-2 所示。

图 9-2 平面隔离示意图



9.2.2 应用安全

9.2.2.1 权限管理

FastCube 支持分权管理。通过用户分权，使得不同用户具有不同的权限，从而保证系统的安全。

分权即区分“操作权限”，它由“角色”进行控制。一个“角色”可拥有一个或多个不同的“操作权限”，一个“用户”可拥有一个或多个不同的“角色”。通过绑定“用户”和“角色”，实现“用户”和“操作权限”的绑定。如果一个“用户”拥有多个“角色”，其拥有的“操作权限”是多个“角色”拥有的“操作权限”的并集。

9.2.2.2 Web 安全

FastCube 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。

9.2.2.3 数据库加固

FastCube 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：

- 访问控制

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

- 最小授权原则

数据库超级管理员之外的其他用户，均按照最小权限的需求设定角色。

- 目录保护

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

- 敏感文件保护

对于数据库的核心配置文件，属主为安装用户，权限控制为读写。

- 连接数限制

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

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

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

9.2.2.4 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。

9.2.3 主机安全

9.2.3.1 操作系统加固

FastCube 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。

9.2.4 数据安全

FastCube 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

FastCube 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。

9.2.4.1 数据加密

FastCube 分布式块存储系统支持数据加密特性，通过配置加密盘和内置加密管（FastCube 1000 分布式块存储系统自带密钥管理系统），和存储系统配合完成静态数据加密，从而保证数据的安全性。

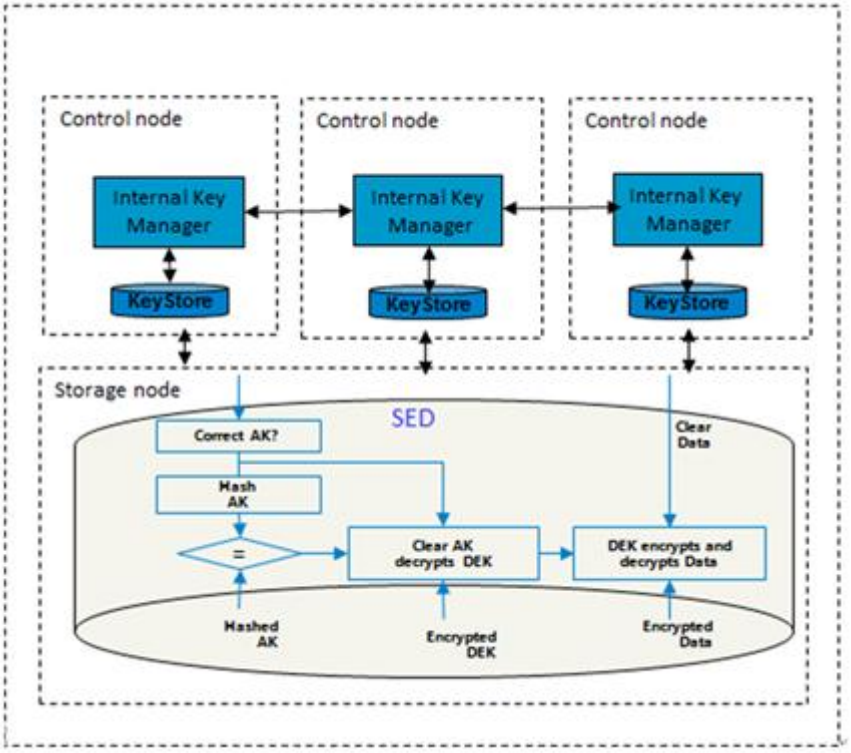
使用内置加密管+加密盘来保证数据的安全性，具有如下特点和优势：

- 采用 SED（Self-Encrypting Drive），数据在盘内进行加解密计算，对业务应用处理流程无影响；
- 无上层增值服务影响，由于数据加密在数据下盘后完成，因此不影响上层的数据重删压缩服务；
- 快速数据销毁，通过密钥销毁，达到快速数据销毁；

- 易部署、易配置和易管理的内嵌于存储系统的密钥管理应用（Internal Key Manager）。

分布式存储系统采用分布式的架构来管理内置密管，多节点的内置密管协同为加密盘提供安全的密钥服务。参见下图：

图 9-3 分布式存储数据加密系统图



- AK（Authentication Key）认证原理：当在分布式存储 8.0.0 分布式块存储系统上启用数据加密特性时，存储会打开加密硬盘的 AutoLock 功能，使用由 Internal Key Manager 分配的 AK 对加密盘的接入进行认证，控制对加密硬盘的访问。此时访问已由 SED 的 AutoLock 功能进行保护，只能由存储系统本身访问。硬盘每次接入时，需要存储系统从密管服务器获取硬盘的 AK，如果与硬盘上的 AK 匹配，硬盘就将加密后的 DEK 解密，用于数据加解密。如果 AK 与硬盘上的 AK 不匹配，则任何读写操作都将失败。

- DEK（Data Encryption Key）加密技术原理：当硬盘成功通过 Autolock 认证后，对硬盘进行读写时，硬盘通过自身的加密芯片和内部的数据密钥（Data Entrypt Key）完成写入数据加密和读取数据解密的功能。用户下发写操作时，明文数据通过 AES 加密引擎的 DEK 加密变成加密数据，然后被写入介质。用户下发读操作时，在介质中的加密数据通过 AES 加密引擎的 DEK 解密，被还原成明文数据取出。DEK 本身无法获取，意味着硬盘被拆除后，通过直接读取的方式无法还原原始信息。