FastCube 2910 计算型存储系统

技术白皮书

文档版本 01

发布日期 2023-02-02

北京元亿科技服务有限公司

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

目 录

1 摘要

数据爆炸式增长和丰富的应用使得 IT 系统越发复杂，信息系统严重依赖厂商和集成商，当前无专属方案，中小企业数字化转型挑战愈发明显，需要简单一体化的数据中心。FastCube 2910 计算型存储作为面向中小型企业推出的网存算一体化微型数据中心，系统集成了 FusionCompute 虚拟化平台和 FastCube 混闪存储平台。FusionCompute 是中国云计算软件最早商用的厂家之一，其虚拟化技术久经考验，目前已累计服务于 150+国家和地区。混合闪存架构，为 FastCube 2910 提供了坚实可靠的数据底座，其 A-A 架构、SAN&NAS、智能加速、智能压缩、无损快照等关键技术让系统更高效、更可靠。FastCube 2910 在硬件设计上也极具创新精神，在 4U 空间集成了计算、存储和网络能力，实现了极简组网、计算即插即用能力。FastCube 2910 相对传统超融合升级了算力和存储能力；通过 SAN&NAS、SSD 和 HDD 深度融合技术，实现一机多用，减少客户投资；极简易用和易运维的使用体验，有效减少客户运维成本。FastCube 2910 是中小客户部署 IT 应用平台的最佳性价比之选。该产品具有如下特点：

极简组网

- 计算节点和存储节点通过内部交换机组网，对外通过板载交换口接入客户组网，交付上避免复杂的连线。

极简管理

- 计算、存储、虚拟机、网络的统一管理，统一配置。
- 计算即插即用，免升级工具、免扩容工具、免巡检工具、免安装系统。

极简交付

- 设备出厂预装存储和计算系统软件，从设备开箱到交付可在 1.5 小时完成。
- 支持使用手机 APP 扫码开局。

多设备集中管理

- 支持接入 DME，实现分支机构的集中管理，支持在中心对分支设备集中管理和发放资源。

一体化应用平台

- 内置虚拟化平台(FusionCompute)，可支持虚拟机的快照、克隆、迁移、H-A 等功能。

坚实可靠的数据底座

- 使用 FastCube 混合闪存系统为底座，支持丰富的专业存储特性，如同时支持 SAN 和 NAS 协议实现一机多用；支持 RAID2.0+、快照、CDP、克隆对数据提供安全保障；支持智能加速、数据缩减技术让设备更高效。

灵活扩展

- 支持算力扩展，计算板即插即用，免系统安装，快速完成算力扩展。
- 支持扩容硬盘框，硬盘即插即用，支持管理界面一键扩容。

本文从产品定位、硬件架构、软件架构、特性方面详细介绍了 FastCube 2910 计算型存储的关键技术，以及为客户带来的独特价值。

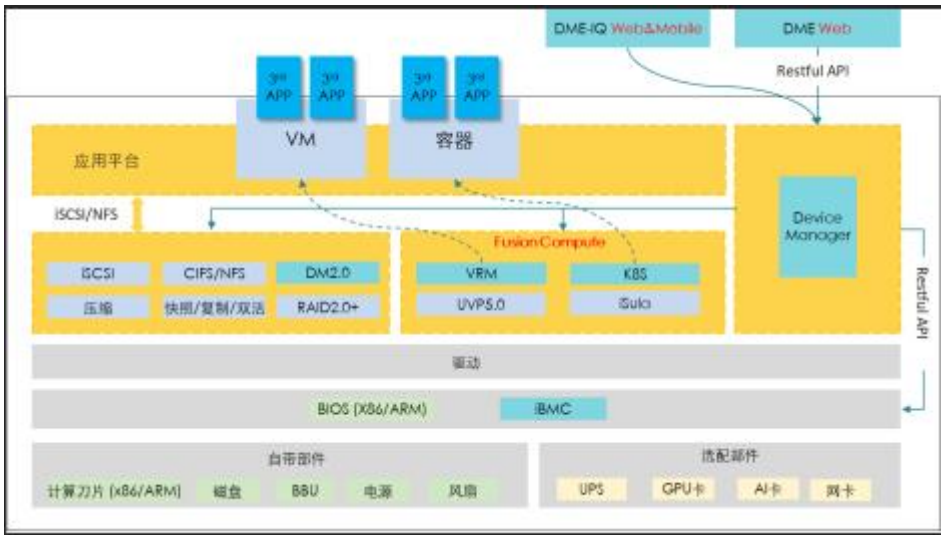
2 简介

本章描述 FastCube 2910 计算型存储及独特的客户价值。

2.1 产品介绍

FastCube 2910 计算型存储产品集成了计算资源虚拟化软件、存储系统、交换网络于一体，针对中小型企业客户打造极简易用、最佳性价比 IT 平台。

图 2-1 整体逻辑架构图



其中存储控制器上部署 FastCube 混合闪存系统，通过对称 Active-Active 架构实现系统的负载均衡，以及配合动态自适应数据布局（DADL）技术，针对 SSD 和 HDD 相结合的混闪特点，进行深度优化，将数据以最优方式放置在不同介质上，充分发挥混闪存储系统性能。

存储控制器上部署了弹性虚拟交换机（EVS），实现计算节点与计算节点、计算节点与存储系统之间的高性能互联。

计算节点上安装新版本 FusionCompute 虚拟化软件，支持虚拟机资源的全生命周期管理。管理角色的计算节点上部署了统一管理平台 DeviceManager，实现系统的一站式极简管理。

说明

- 当前版本不支持部署容器。

2.2 客户价值

过去的几年，数据量的快速增加和愈发丰富的应用使得 IT 系统越发复杂，信息系统严重依赖厂商和集成商，中小客户数字化转型需求基本相同，需要简单一体化的数据中心方案：

- 一体化的建设交付，应用快速上线；
- 简单易用的管理运维方式，免专业 IT 运维人员；
- 数据存储安全可靠，防止数据丢失；
- 设备经济高效，节省投资；

FastCube 2910 计算型存储产品通过软硬件的核心科技能力，实现计算、网络、存储融合，针对商业市场：打造极简易用、最佳性价比 IT 应用平台一体机，设备即插即用，提供专业的产品（企业级特性，极简运维），满足中小客户数字化转型对 IT 系统的述求。

整机高度集成，实现小型化一体机

- 整合 IT 基础设施，包括计算、存储、交换网络系统，整机空间为 4U，减少空间消耗；
- 免外置交换机，降低组网复杂度，2 根管理网线+2 根业务网线即可满足管理和业务网络部署；

预配置，易集成，系统一体化管理，极简运维

- 预集成 FusionCompute 虚拟化组件，存储资源出厂预配置；
- 初始化配置极简，ISV 易集成，设备进场后即插即用；
- 存储、计算、网络、虚拟化资源统一界面管理，故障主动排查，简化日常运维；
- 支持手机 App 扫码开局，支持 DME 云化运维，客户免专业运维；
- 计算节点无系统盘设计，实现计算节点即插即用，在减少客户投资成本的同时，提升计算节点系统盘可靠性；

企业数据存储底座，提供高可靠、经济高效的存储服务能力

- **动态自适应布局技术：**实现 SSD 和 HDD 深度配合协同，利用 ROW 大块顺序写、Cache/Tier 弹性融合性能层、多维智能加速算法，充分激发盘的高性能和长生命周期运行的业务能力。

3 硬件介绍

FastCube 2910 计算型存储采用 4U 高密设计，支持 3 计算节点+2 存储控制器，其中：

- 控制器模块、电源模块、BBU 模块、风扇模块、硬盘单元等核心部件不存在任何单点故障。各种 FRU 均支持在线热插拔，支持在线可更换。同时，存储控制器间采用 RDMA 高速网络互联，通过该网络实现全局缓存的低时延共享访问；
- 计算节点当前版本为 X86 平台，采用开放架构设计，可支持第三方计算主板；支持扩展标准 PCIe 接口卡、FLEX IO 卡（OCP）、AI 卡，具体型号可参考产品规格清单定义。计算节点间、计算到存储系统通过内部 10GE 网络互联。

3.1 硬件能力介绍

计算节点	节点数量	2-3 节点
	CPU	支持第三代英特尔至强可扩展处理器 4310、4314、4316、5318Y、6330、6346、6348 以及澜起科技津逮 C4310、C4314、C4316、C5318Y、C6330、C6346、C6348
	内存	每节点最大支持 32 个内存槽位
	FLEX IO 卡	支持 GE、10GE
	PCIe 标卡	支持 GE、10GE、25GE
	AI 卡	T4、Atlas300I PRO

	硬盘	计算节点免系统盘，操作系统卸载到存储节点
	板载接口	4*10GE
存储节点	存储控制器	2 节点
	CPU	Kunpeng920 28 核 2.6GHz*1/控
	内存	40GB/控
	板载接口	2*10GE（作为存储复制口使用）、4*10GE（与计算节点共用）
	内部接口	2*25GE+16*10GE
	硬盘	系统最大支持 110 盘
	控制框内硬盘	4SSD+6HDD
	扩展硬盘框	4U 3.5 英寸 SAS 硬盘框 *4 或 2U 2.5 英寸 SAS 硬盘框 *4

3.2 控制框介绍

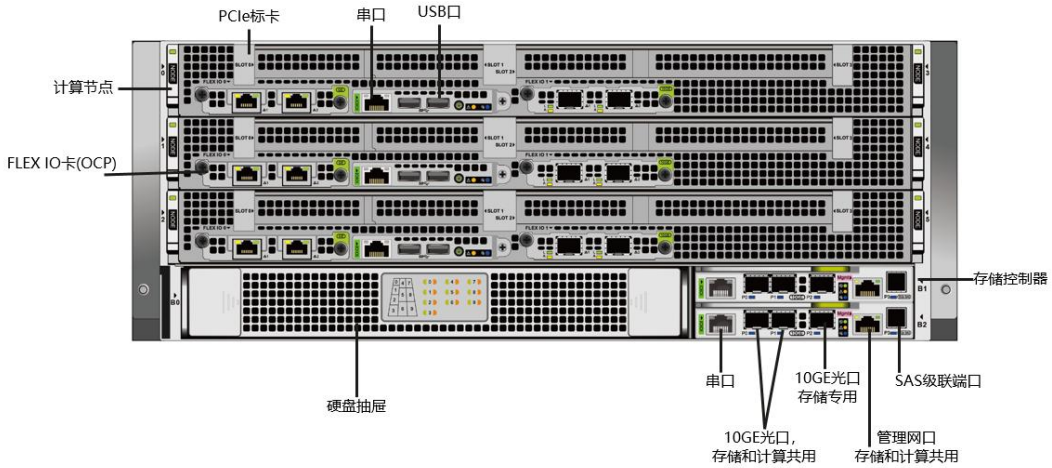
FastCube 2910 采用 4U 的控制框，网、存、算一体形态；包含 3 计算节点+2 存储控制器；支持 4 盘位 SSD+6 盘位 HDD；关键模块 FRU 设计，支持冗余备份可在线更换。

其中存储系统为客户提供 SAN 和 NAS 的统一存储服务，端到端的 SAN（iSCSI）和 NAS（NFS、CIFS)等前端协议服务能力，支持的板载接口包括：2*10GE（作为存储复制口使用）、4*10GE（与计算节点共用）。

存储系统为对称 Active-Active 架构，2 个控制器支持故障的业务接管和正常时负载均衡，两个控制器间通过 RDMA 实现镜像能力；整框前视图和后视图如下：

图 3-1 FastCube 2910 前视图

图 3-2 FastCube 2910 后视图



说明

- BBU 模块只对存储提供备电，每个存储控制器一个 BBU 模块。
- 计算节点面板包含串口、USB 口和独立的电源按钮。
- 存储控制器面板上：
- P0、P1 口默认为主备模式绑定，计算节点和存储系统共用，支持在存储系统中修改端口的绑定模式，支持主备模式、负载均衡模式。
- P2 端口作为存储系统复制口使用。
- 管理网口为存储控制器和计算节点共用。
- 串口为存储系统使用。
- SAS 端口用作存储系统级联硬盘框。

FastCube 2910 计算型存储系统内部通过 10GE 链路实现计算节点与计算节点、计算节点与存储控制器的全互联。

3.3 硬盘框

FastCube 2910 计算型存储支持两种 SAS 硬盘框：

表 3-1 硬盘框类型

硬盘框类型	硬盘类型	端口	盘位数
25 盘位 SAS 硬盘框	双端口 SAS 盘	4x12Gbps SAS 宽端口	25
24 盘位 SAS 硬盘框	双端口 SAS 盘	4x12Gbps SAS 宽端口	24

3.3.1 25 盘位 SAS 硬盘框

25 盘位 SAS 硬盘框，采用 SAS3.0 协议，每个框支持 25 块 2.5 寸 SAS 硬盘。控制框通过板载 SAS 接口或者 SAS 接口模块与 SAS 硬盘框连接。

图 3-3 2U 25 盘位 SAS 硬盘框前视图

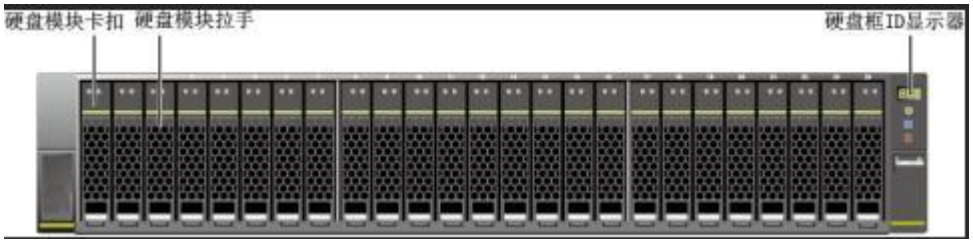
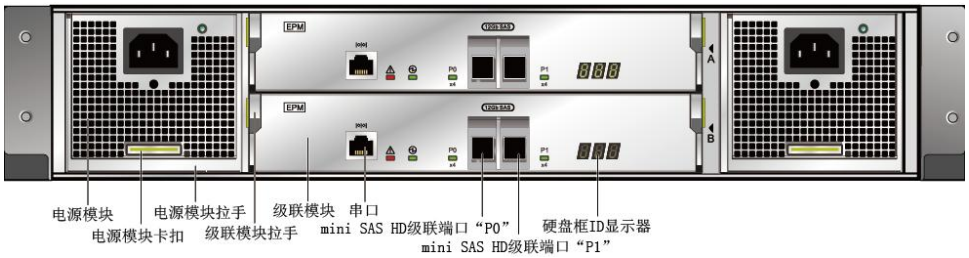


图 3-4 2U 25 盘位 SAS 硬盘框后视图

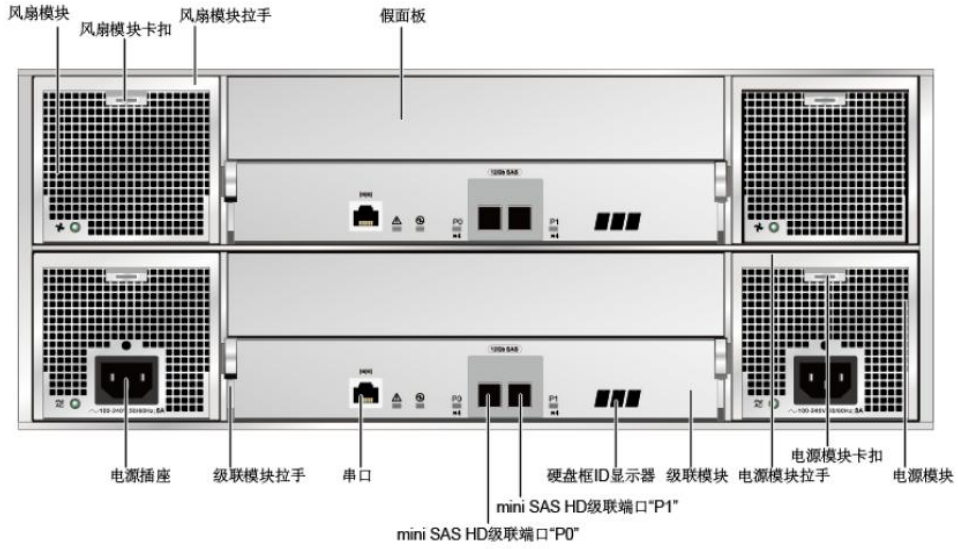


3.3.2 24 盘位 SAS 硬盘框

24 盘位 SAS 硬盘框，采用 SAS3.0 协议，每个框支持 24 块 3.5 寸 SAS 硬盘。控制框通过板载 SAS 接口或者 SAS 接口模块与 SAS 硬盘框连接。

图 3-5 4U 24 盘位 SAS 硬盘框前视图

图 3-6 4U 24 盘位 SAS 硬盘框后视图



3.4 功耗与散热

FastCube 2910 计算型存储通过如下节能设计来满足节能环保要求：

- 采用业界最高转换效率的供电模块
- 通过比例微积分风扇调速算法，提高系统散热效率
- 巧妙的错峰上电设计，降低大功率供电需求

通过高能效设计，节省供电和散热开销。

高效电源

FastCube 2910 计算型存储采用 80 PLUS 白金、钛金电源，在 50%带载时的供电转换效率可达 94%以上，功率因数可达 98%以上，降低供电损耗。钛金电源的转换效率比白金提高 2 个点，可以达到 96%以上，而且轻载时的转换效率可以达到 90%以上，对比普通 PSU 电源的转换效率提升将近 10 个点，最大程度降低产品全负载范围内的供电损耗。PSU 电源满足 80PLUS 认证规范，可提供认证证书。

80 PLUS 效率要求：

80 PLUS 电源认证类型	**电源转换效率（230V 输入 ）**			
负载比（%）	**10%**	**20%**	**50%**	**100%**
80 PLUS 铜	---	81%	85%	81%
80 PLUS 银	---	85%	89%	85%
80 PLUS 黄金	---	88%	92%	88%
80 PLUS 白金	---	90%	94%	91%
80 PLUS 钛金	90%	94%	96%	91%

PID 节能调速技术

FastCube 2910 计算型存储通过支持 PID（Proportional Integral Derivative，比例积微分算法）节能调速算法解决了传统阶梯调速的诸多问题：调速周期长、风扇功耗高、震荡幅度大、噪声高等问题，实现风扇快速响应、快速降温、超低功耗、超低静音的效果。

- 经验证，使用 PID 节能调速算法可以提升产品能效 4%～9%，避免出现风扇反复震荡。
- 通过 PID 调速，风扇响应速度可以提升 22%～53%，明显降低产品噪声。

错峰上电

通过错峰上电，消除产品上电瞬间的供电脉冲，避免多台设备同时上电对机房供电造成风险。

节能认证

产品满足中国环保部的节能认证标准、CQC 的节能认证标准、RoHS 环保认证标准。

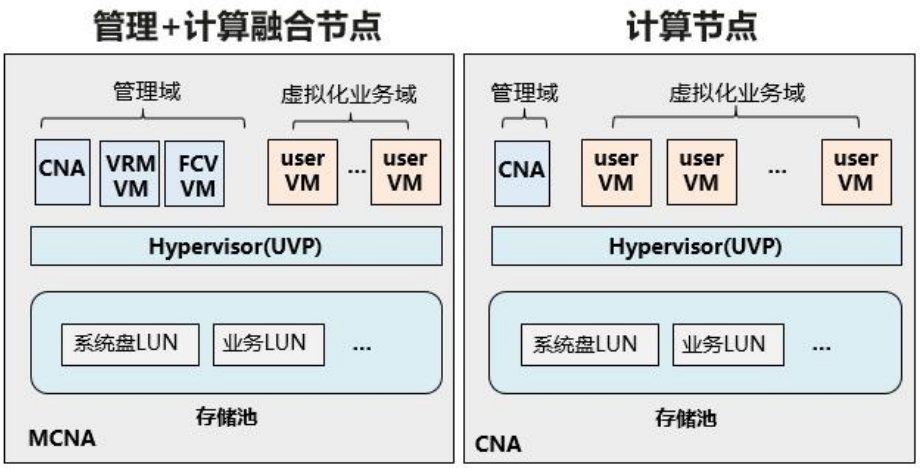
4 虚拟化应用平台

4.1 系统架构

4.1.1 软件架构

系统中 3 个计算节点的出厂预装 FusionCompute 虚拟化管理软件和统一管理软件 DeviceManager。

图 4-1 计算节点软件部署方案

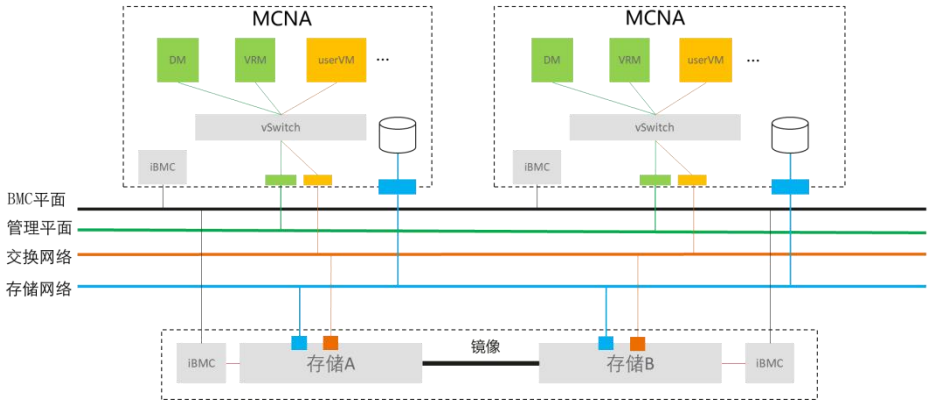


名称	说明	部署原则
MCNA（管理节点）	具有管理功能的节点，其上部署了VRM、2910 DeviceManager 管理虚拟机，同时也可提供计算功能	必须部署 2 个
CNA（计算节点）	具有计算功能的节点，只能提供虚拟化计算资源	根据需要部署 0 个～1 个

4.1.2 网络架构

计算型存储 2910 使用创新网络架构，网元之间的交换网络完全隐藏在系统内部如图所示，这样可以免除互联交换机。而且在大部分情况下，系统自带的交换上行口完全可以满足用户网络诉求，可以免除用户上行交换机。

图 4-2 FusionCompute 网络架构



管理平面：系统管理网络平面，用于系统的业务操作和运维管理；

存储网络：使用 10GE 直连网络，用于 SANBOOT 和块服务；

交换网络：客户业务通信网络平面，支持 TCP/IP 协议，使用 10GE 组网，与管理平面共网络，通过 VLAN 隔离；

BMC 平面：服务器设备管理 IP 平面，使用 GE 网络，用于服务器硬件设备的运维管理；

4.2 虚拟化原子能力

让我们再次回想下云的定义：通过虚拟化技术，将不同的基础设施标准化为相同的业务部件，然后利用这些业务部件，依据用户需求自动化组合来满足各种个性化的诉求。

积木是最巧妙的玩具，就因为它具备的原子特性：构件形态稳定，易于替换，随意组合，可回收重用。云是为敏捷 IT 而生，FusionCompute 平台提供一系列标准化的原子能力，协助用户像搭积木一样的简便快捷的构建自己的系统。

这一章，我们开始讲解标准化后的原子能力。通过理解这些标准化构件，一方面您可以很好的理解云平台为您提供怎样的基础服务，一方面您可以更好的理解这些构件，以便于您利用这些构件来搭建您的系统。

本章将依据 FusionCompute 的产品组合展开，逐一向您讲解 FusionCompute 能提供的原子能力。

4.2.1 虚拟机

x86 虚拟化技术就是将通用的 x86 服务器经过虚拟化软件，对最终用户呈现标准的虚拟机。这些虚拟机就像同一个厂家生产的系列化的产品一样，具备系列化的硬件配置，使用相同的驱动程序。

FusionCompute 就是这样一个虚拟化系统，支持将 x86 服务器虚拟化为多台虚拟机。最终用户可以在这些虚拟机上安装各种软件，挂载磁盘，调整配置，调整网络，就像普通的 x86 服务器一样使用它。

对于最终用户，虚拟机比物理机的优势在于它可以很快速的发放，很方便的调整配置和组网。对于维护人员来讲，虚拟机复用了硬件，这样硬件更少，加上云平台的自动维护能力，维护成本显著降低。对于系统管理员，可以很直观的知道资源使用的总量及变化趋势，以便决策是否扩容。

4.2.2 虚拟存储

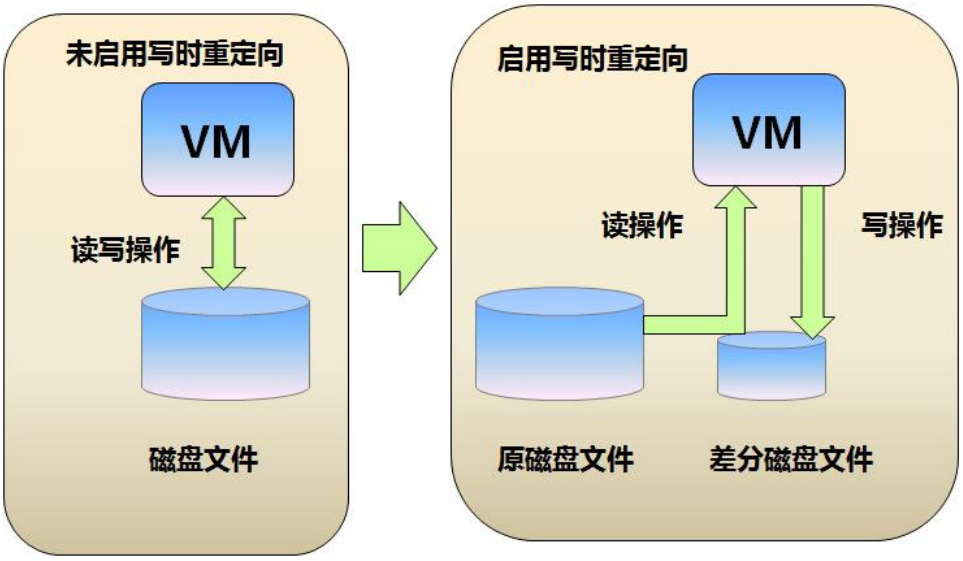
FusionCompute 支持将 SAN 设备，计算节点本地存储，以及分布式存储提供的虚拟存储空间统一管理，以虚拟卷的形式分配给虚拟机使用。

对于最终用户，就像 x86 服务器使用本地硬盘一样的方式使用，可以格式化，安装文件系统，安装操作系统，读写。同时，虚拟化的存储还具备快照能力，可以调整大小，这是物理硬盘不能实现的。

对于管理员来说，虚拟存储卷并不一对一映射到某块具体的磁盘，而是收敛到几台 SAN 设备。由于 SAN 设备已经有了可靠性保障，所以更换硬盘的工作量大规模下降。同时，虚拟存储具备瘦分配，灵活调整，QoS 可限制，可迁移等等比物理盘强的特性，在整体成本方面优势很明显。

磁盘文件的写时重定向技术

FusionCompute 的存储虚拟化具备写时重定向技术（Redirect on Write），能够在虚拟机磁盘文件被修改时，可以不修改原磁盘文件，而是将修改区域记录在另一个差分磁盘中，将差分磁盘的父磁盘指向原磁盘文件，使得虚拟机在从差分磁盘文件中读取数据时，能够自动从原磁盘文件中获得需要的数据。



写时重定向技术可以应用于快照、链接克隆、非持久化磁盘等特性。

快照特性：虚拟机可以将当前状态保存在快照文件中，包括磁盘内容、内存和寄存器数据。用户可以通过恢复快照多次回到这一状态，虚拟机用户在执行一些重大、高危操作前，例如系统补丁，升级，破坏性测试前执行快照，可以用于故障时的快速还原。

链接克隆：链接克隆虚拟机可以基于同一个虚拟机模板，快速发放多个类似的虚拟机。通过对虚拟机模板的系统卷创建多个差分磁盘，将每个差分磁盘挂载给独立的虚拟机。应用于需要大量发放拥有相同或类似数据的虚拟机，且对性能要求不高。

链接克隆加速：在链接克隆场景下，将若干链接克隆虚拟机的共同模板中的热点数据放在主机内存中，达到快速读取的目的，能够极大提升虚拟机的启动和运行速度。

非持久化磁盘：处于保护磁盘数据的目的，在启动虚拟机时，对这种非持久化磁盘先创建差分磁盘，在虚拟机运行过程中，将有更改的数据全部写入差分磁盘，在虚拟机关机后，将差分磁盘删除，达到还原磁盘的目的。应用于公共计算机、计算机数据自动还原的场景。

磁盘文件的存储热迁移

FusionCompute 提供了虚拟机磁盘的冷迁移和热迁移，冷迁移是在虚拟机关机时候，将其磁盘文件从一个存储移动到另一个存储，热迁移可以在不中断业务的前提下，将虚拟机磁盘从一个存储迁移至另一个存储。热迁移的技术原理如下：

1. 热迁移首先使用写时重定向，将虚拟机数据写入目的存储的一个差异磁盘，这样，原磁盘文件就变成只读的。
2. 将源卷的所有的数据块依次读取出来并合并到目标端的差异磁盘中，等数据合并完成后，目的端的差分磁盘就拥有虚拟磁盘的所有最新数据。
3. 去除目的端快照对源卷的依赖，将差分磁盘修改为动态磁盘，这样，目的端磁盘文件可以独立运行。

磁盘文件高级业务

FusionCompute 提供了虚拟机磁盘的高级业务，可以在提供磁盘文件的扩容功能，技术原理如下：

磁盘扩容：在离线或在线状态下支持对磁盘的容量扩充，对于厚置备磁盘，会将数据区域进行扩充，并进行写零。对于厚置备延时置零磁盘，会将数据区域进行扩容，并进行空间预占。对于精简磁盘，仅对数据区域进行扩容。

4.2.3 虚拟网络

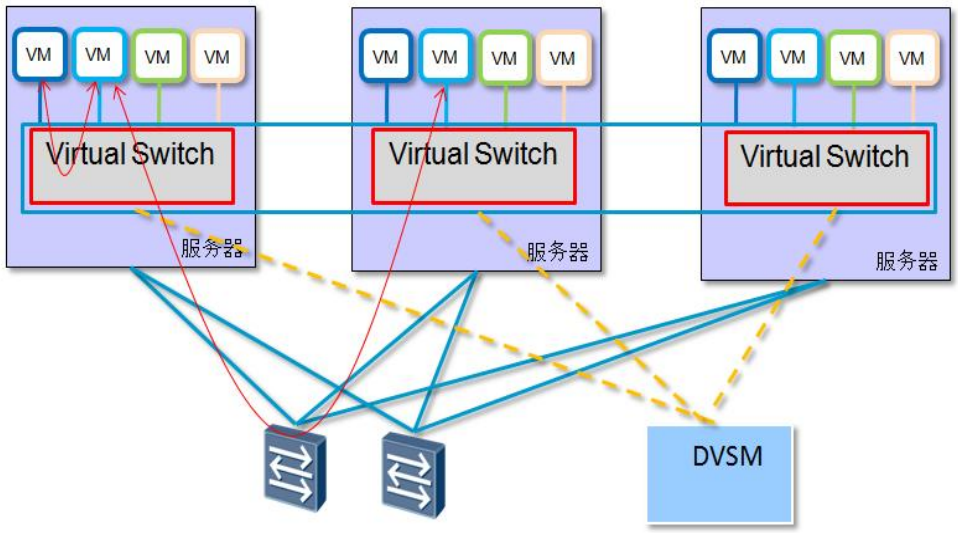
FusionCompute 具备支持分布式虚拟交换，可以向虚拟机提供独立的网络平面。像物理交换机一样，不同的网络平面间通过 VLAN 进行隔离。这种技术具备如下的特点：

- 同一宿主机上的不同虚拟机，如位于不同 VLAN，则不能直接互通；
- 同一宿主机上的不同虚拟机，如位于相同 VLAN，则可以直接二层互通。此时网络流量通过内存交换，不受任何网络带宽限制。
- 不同宿主机上的不同虚拟机，如位于相同 VLAN，则可以通过外部交换机进行互通，就像没有虚拟化一样。
- 管理网络平面及业务网络平面支持 IPv4 和 IPv6 两种网络协议。

通过这种能力，您可以将 VLAN 视为独立的网络平面，通过向不同的虚拟机分配不同的 VLAN，来实现各种业务间的隔离。

虚拟交换提供集中的虚拟交换的管理功能。集中的管理提供统一的 Portal，进行配置管理，简化用户的管理。

图 4-3 虚拟交换场景



通过分布在各物理服务器的虚拟交换机，提供虚拟机的二层通信、隔离、QoS 的能力。

分布式交换机模型基本特征：

1) 虚拟化管理员可以配置多个分布式交换机，每个分布式交换机可以覆盖集群中的多个 CNA 节点；

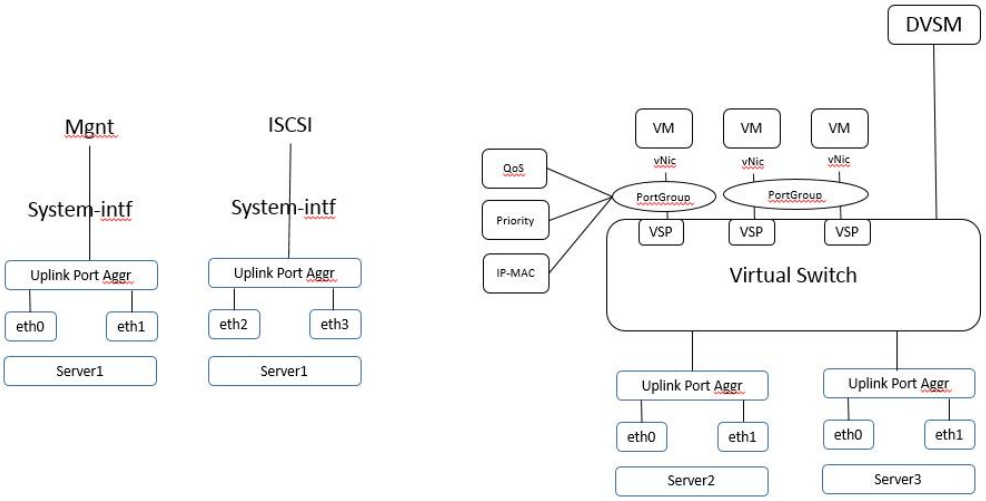
2) 每个分布式交换机具有多个分布式的虚拟端口 VSP，每个 VSP 具有各自的属性(速率)，为了管理方便采用 Port Group 组管理相同属性的一组端口，相同端口组的 VLAN 相同；

3) 虚拟化管理员或业务系统（例如 VDI/IDC）,可选择管理/存储/业务使用的不同物理接口；每个分布式交换机可以配置一个 UpLink 端口或者一个 Uplink 端口聚合组，用于 VM 对外的通信。Uplink 端口聚合组可以包含多个物理端口，端口聚合组可以配置负载均衡策略；

4) 每个 VM 可以具有多个 vNIC 接口，vNIC 可以和交换机的 VSP 一一对接；

5) 虚拟化管理员或业务系统可根据业务需求，选择在一个集群中允许进行 2 层迁移的服务器 创建虚拟二层网络，设置该网络使用的 VLAN 信息；

图 4-4 虚拟交换模型



虚拟化管理员可通过定义端口组 属性（安全/QoS）简化对虚拟机端口属性的设置；设置端口组属性，不影响虚拟机正常工作；

端口组：端口组是网络属性相同的一组端口的属性集合。管理员可以通过配置端口组属性（带宽 QOS、2 层安全属性、VLAN 等）简化对虚拟机端口属性的设置。设置端口组属性，不影响虚拟机正常工作；

上行链路：分布式交换机关联的服务器物理网口；管理员可以查询上行链路的名称、速率、模式、状态等信息；

上行链路聚合：分布式交换机关联的服务器绑定网口，绑定网口可以包含多个物理网口，这些物理网口可以配置主备或负载均衡策略。

4.3 GPU 直通技术

4.3.1 方案简介

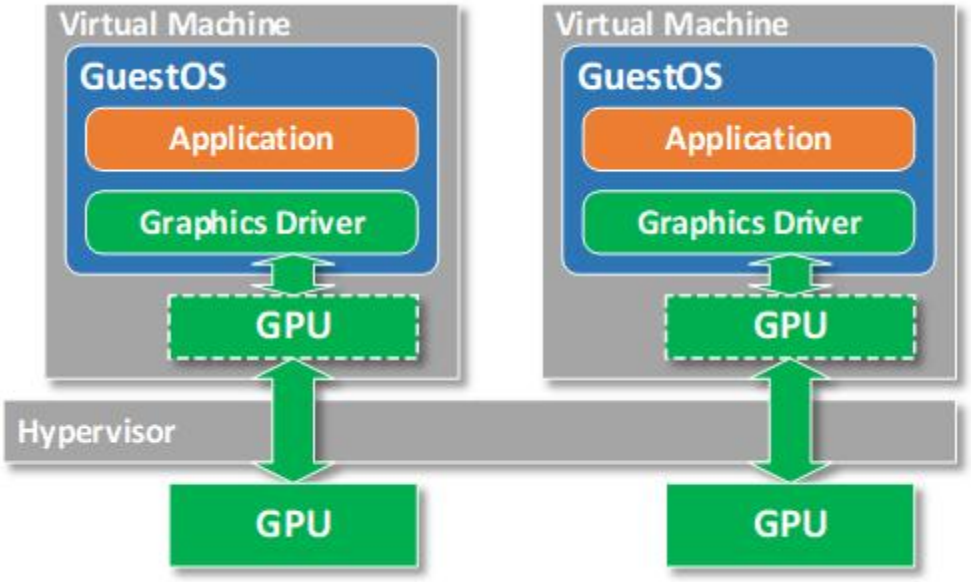
X86 架构的 FusionCompute 针对不同的业务场景及对 GPU 资源的使用情况，提供以下两种有针对性的解决方案：

- 主机 PCI 设备直通

- GPU 资源组

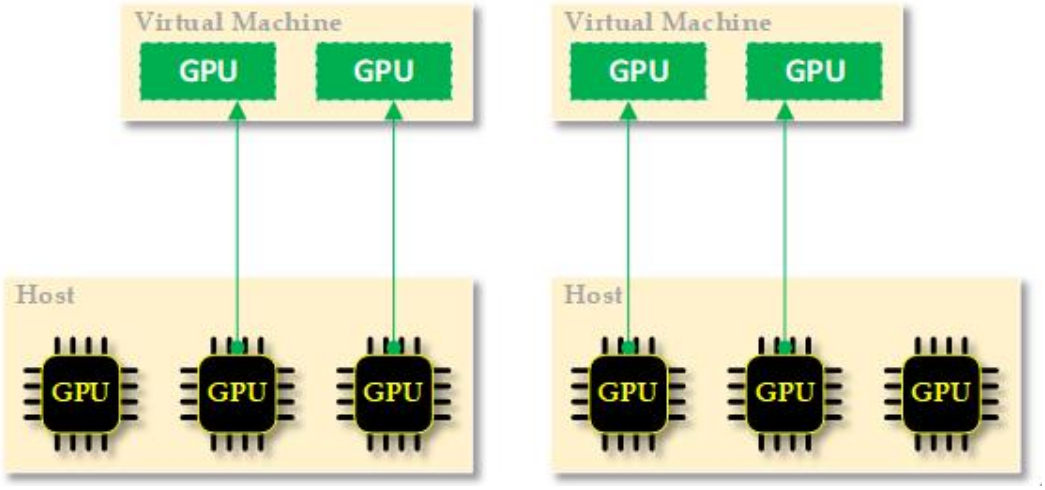
方案类型	解决方案	功能特点	应用场景
PCI 直通	直接将主机上 GPU 对应的 PCI 设备直通给虚拟机使用	1、独占 GPU 设备 2、不允许 GPU 超分配	视频云等
GPU 资源组	将 GPU 设备添加到 GPU 资源组，再从 GPU 资源组中分配 GPU 给虚拟机	1、在虚拟机关闭后，GPU 可被其他虚拟机使用； 2、允许 GPU 超分配	桌面云等

FusionCompute 使用 PCI 直通技术将 GPU 设备直通给虚拟机，即将 GPU 在计算节点中呈现的 PCI 设备直接绑定给虚拟机以呈现成为虚拟机的一个 PCI 设备。



4.3.2 PCI 直通

FusionCompute 在主机（计算节点）启动等时机主动发现其以装配的 GPU 设备，这些 GPU 设备可直接绑定给虚拟机。



使用场景

适用于对 GPU 资源需求明确，需要对 GPU 进行持续独占使用的场景。

例如：地震分析、视频云等领域

使用约束

PCI 直通方案存在以下约束：

1. 虚拟机必须与 GPU 设备所在主机绑定；
2. 已直通 GPU 设备的虚拟机不支持内存快照；
3. 已直通 GPU 设备的虚拟机不支持热迁移、休眠、唤醒操作；
4. 仅支持在 GPU 关闭状态下进行 GPU 设备的绑定与解绑定操作；
5. 一个 GPU 只能绑定给一个虚拟机或一个 GPU 资源组；
6. 需要进行 GPU 直通的虚拟机的内存必须全部预留；
7. 每个虚拟机最多支持直通 8 个 GPU 设备；
8. 需要提前在主机的 BIOS 中开启 VT-d 和 VT-x 支持。不同厂商服务器开启的方式会有区别，请参考具体的服务器帮助文档；

说明

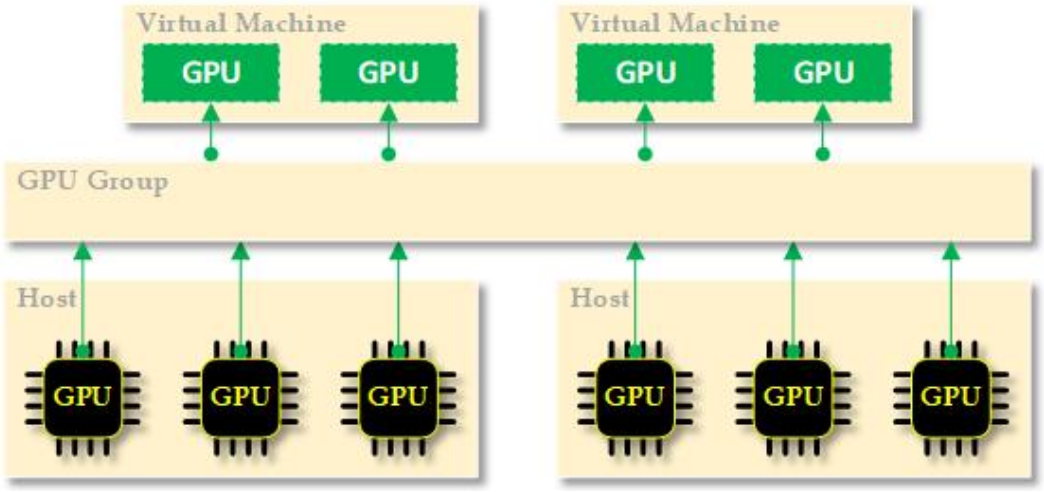
即使绑定 GPU 的虚拟机的电源已关闭，GPU 也不能被其他虚拟机使用

4.3.3 GPU 资源组

GPU 资源组方案需要管理员首先创建以直通方式使用的 GPU 资源组，并将主机上的 GPU 设备添加到资源组中。在需要对虚拟机分配 GPU 资源时，直接将 GPU 资源组绑定给虚拟机，并指定预计在资源组中分配的 GPU 数量。

客户虚拟机启动时，GPU 资源组按照虚拟机所需的 GPU 数量，找到空闲的 GPU 数量满足条件的主机，并从该主机中分配所需数量的 GPU 资源以 PCI 直通的方式直通给虚拟机，以使虚拟机顺利启动。

客户虚拟机在关闭后，GPU 资源会释放给资源组，此时这些 GPU 资源可以被分配给其他虚拟机使用。



相比于 FusionCompute 6.3 之前的版本，该方案支持在一套环境中同时使用多种规格的 GPU 卡，只需将不同的 GPU 卡划分在不同的资源组中即可。该方案可以支持系统中存在需要使用不同 GPU 的应用的情况，同时也可在设备演进中平台地支持 GPU 卡进行更新换代。

多种 GPU 卡是否可以同时直通给一台虚拟机，需要从虚拟机操作系统供应商以及显卡芯片厂商处获取支持。

说明

每个 GPU 资源组可以管理相同产品型号的 GPU 资源，虚拟机可以通过绑定多个不同产品型号的 GPU 资源组以挂载不同型号的 GPU 设备，但是这些设备在虚拟机的操作系统中是否可以正常使用，由 GPU 的驱动程序以及操作系统决定。

使用场景

适用于对 GPU 资源需求存在弹性，允许进行 GPU 分时复用的场景。

场景举例：大数据分析、深度学习、互联网、教育等领域

例如：

分时复用

某公司有 3 种业务 A、B、C 均需要使用 GPU，但各业务运行的高峰时段存在明显差异，如下表所示：

业务	高峰时段	对 GPU 数量的诉求	
		常规时段	高峰时段
A	8:00~10:00 16:00~20:00	2	6
B	11:00~14:00	2	4

C	1:00~4:00	1	3

时段	对 GPU 数量的诉求			
	业务 A	业务 B	业务 C	总数
1:00~4:00	2	2	3	7
8:00~10:00	6	2	1	9
11:00~14:00	2	4	1	7
16:00~20:00	6	2	1	9

在此种场景下，为满足所有业务在高峰时段对于 GPU 的诉求，共需配备 13 块 GPU；而各业务存在明显的高峰时段差异，此时对各高峰时段进行分析得出在每个高峰时段中，所需要的 GPU 数量都不超过 9 个，如下表所示：

在业务系统支持动态扩缩容的情况下，在某个业务处于高峰时段时，业务系统通过扩展节点（虚拟机）的方式进行扩容，在脱离高峰时段时进行缩容，只需配备 9 块 GPU，即可满足所有业务在各自高峰时段对 GPU 的诉求。

此种相同 GPU 在不同时间由不同的系统或应用使用的情况即为分时复用。

使用约束

GPU 资源组方案存在以下约束：

1. 同一台虚拟机所使用的 GPU 必须集中在同一台主机上，当任意主机空闲的 GPU 不足以满足虚拟机运行所需时，即使资源组中的空闲 GPU 总量满足，虚拟机也无法被分配到所需的 GPU 资源；
2. 已直通 GPU 设备的虚拟机不支持内存快照；
3. 已直通 GPU 设备的虚拟机不支持热迁移、休眠、唤醒操作；
4. 仅支持在 GPU 关闭状态下进行 GPU 设备的绑定与解绑定操作；
5. 一个 GPU 只能绑定给一个虚拟机或一个 GPU 资源组；
6. 需要进行 GPU 直通的虚拟机的内存必须全部预留；
7. 每个虚拟机最多支持直通 8 个 GPU 设备；
8. 需要提前在主机的 BIOS 中开启 VT-d 和 VT-x 支持。不同厂商服务器开启的方式会有区别，请参考具体的服务器帮助文档；

4.4 质量保证

FusionCompute 提供的大量的机制来协助管理员实现质量保证，包括主动管理和被动管理方式。

所谓主动管理，就是管理员设定策略，由管理系统依据这些设定自动的进行质量保证。而被动管理则是提供信息给管理员，由管理员依据这些信息进行管理。

如下列出常用的主动管理和被动管理能力

4.4.1 主动管理

FusionCompute 系统中有很多主动管理的功能，大部份都是管理员感知不到的，这里面一些关键功能，比如虚拟机死机检测，休眠检测，虚拟网络的被攻击检测，管理系统本身的故障检测等，这些功能太多就不一一列出。如下仅列出管理员常用的主动管理功能。

4.4.1.1 虚拟机 HA

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

在已经创建的集群中如果高级设置中的 HA 功能已经启用，那么用户在该集群中创建虚拟机时，可以选择是否支持故障重启，即是否支持 HA 功能。

系统周期检测虚拟机状态，当物理服务器宕机等引起虚拟机故障时，系统可以将虚拟机迁移到其他物理服务器重新启动，保证虚拟机能够快速恢复。目前系统能够检测到的引起虚拟机故障的原因包括物理硬件故障、系统软件故障。

重新启动的虚拟机，会像物理机一样重新开始引导，加载操作系统，所以之前发生故障时没有保存到硬盘上的内容将丢失。

对于未启用 HA 功能的虚拟机，当发生故障后，此虚拟机会处于停机状态，用户需要自行操作来启动这台虚拟机。

4.4.1.2 虚拟机 DRS

动态资源调度 (DRS) 动态分配和平衡资源，采用智能调度算法，根据系统的负载情况，对资源进行智能调度，达到系统的负载均衡，保证系统良好的用户体验。

动态资源调度策略针对集群 (Cluster) 设置，可以设置调度阈值、定义策略生效的时间段。在策略生效的时间段内，如果某主机的 CPU、内存负载阈值超过调度阈值，系统就会自动迁移一部分虚拟机到其它 CPU、内存负载低的主机中，保证主机的 CPU、内存负载处于均衡状态。

4.4.1.3 虚拟机 QoS

客户可以自定义必须在同一主机上运行或必须分开主机运行的虚拟机，或者限定某些虚拟机只能在部分主机范围内运行和迁移。

虚拟机 QoS 功能，实现了可衡量的计算能力，用来保证虚拟机的计算能力在一定范围内，隔离了虚拟机间由于业务变化而导致的计算能力的相互影响，满足了不同业务虚拟机的计算性能要求。同时可以更好地控制计算资源，最大程度复用资源，降低成本，提高用户满意度。

虚拟机 QoS 主要体现在 CPU QoS 和内存 QoS。

- CPU QoS

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

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

- 内存 QoS

提供虚拟机内存智能复用功能，依赖内存预留比。通过内存气泡占用等内存复用技术将物理内存虚拟出更多的虚拟内存供虚拟机使用，每个虚拟机都能完全使用分配的虚拟内存。该功能可最大程度的复用内存资源，提高资源利用率，且保证虚拟机运行时至少可以获取到预留大小的内存，保证业务的可靠运行。

系统管理员可根据用户实际需求设置虚拟机内存预留。内存复用的主要原则是：优先使用物理内存。

4.4.1.4 虚拟机自动备份

虚拟机备份是使用 eBackup 备份软件，配合 FusionCompute 快照和 CBT（Changed Block Tracking）功能实现的虚拟机数据备份方案。eBackup 通过与 FusionCompute 配合，实现对指定虚拟机的备份。当虚拟机数据丢失或故障时，可通过备份的数据进行恢复。数据备份的目的端为本地虚拟磁盘或 eBackup 外接的共享网络存储设备（NAS）。

eBackup 支持通过设置备份策略，实现虚拟机的自动定期备份。支持针对不同虚拟机或虚拟机组设置不同备份策略，最多支持 200 个备份策略。

- 支持对全备份与增量备份或差量备份分别设置不同备份周期、备份时间窗口；如支持设置每周进行一次全备、每天进行一次增备，也可只进行一次全备，后续一直进行增备。
- 支持设置备份数据保留时间以自动清除过期备份数据。
- 支持设置备份策略优先级。

4.4.2 被动管理

被动管理向管理员提供常见的展示，报表，统计，告警，事件等能力。管理员可以基于这些能力进行管理。

4.5 自动化能力

4.5.1 概述

云的核心理念是虚拟化，标准化和自动化。前面已经讲解了 FusionCompute 所支持的标准化构件。作为一个云平台，仅仅提供标准化并不能称之为云，必须要有自动化能力才能更加好用。

在自动化方面，FusionCompute 提供了大量的自动化功能，用来简化日常管理动作。

依据 eTOM 模型，在日常管理动作中，主要关心如下的关键点：

- 策略：资源的供给如何满足业务的需求，在不同的业务间怎么调整资源分配策略。
- 基础设施：我们拥有哪些设施，使用情况是怎样的，状态是怎样的。这些基础设施是如何按照既定策略来支持产品的。
- 产品/服务：我们对外提供哪些服务，这些服务的质量是怎样的，基础设施是怎样支撑这些产品和服务的。
- 发放：一个具体的产品/服务实例是怎样发放的。
- 保障：系统怎样保障一个已经发放的产品/服务实例的质量。
- 计量：服务/产品的实例的用量是怎样的，用户偏好是什么，怎样调整资源及策略以便更好的提供产品/服务。

对于 FusionCompute 来说，所提供的产品/服务，就是在上面一章中所描述的标准化部件。FusionCompute 的自动化管理提供了大量的能力，来协助系统管理员和最终用户来更加方便的管理和使用这些服务。

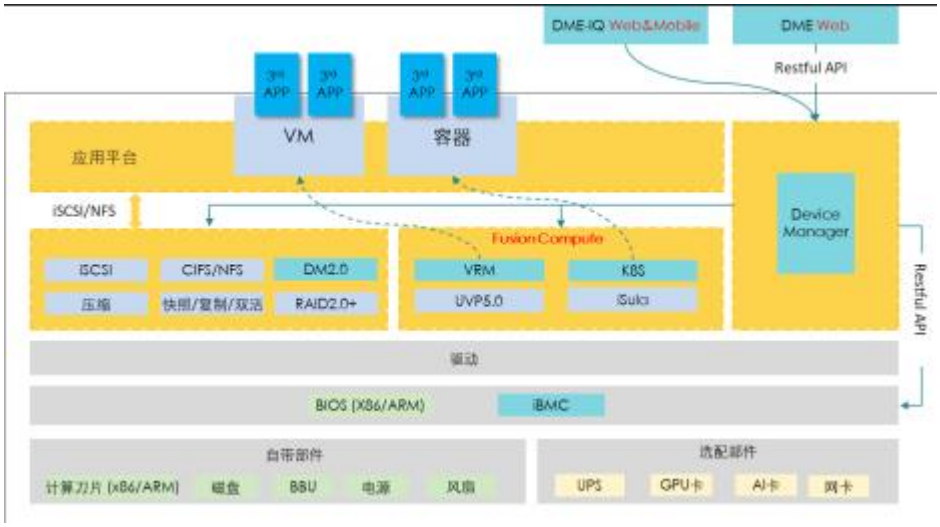
4.5.2 标准化部件发放及使用

所有在上一章描述的标准化部件都提供完备的生面周期管理能力，包括发放，创建，使用，调整，配置，回收。用户可以通过界面进行操作，也可以通过 Web service 接口通过编程访问。

5 企业存储数据底座

FastCube 2910 计算型存储产品集成了计算资源虚拟化软件、存储系统、交换网络于一体，针对中小型企业客户打造极简易用、最佳性价比 IT 平台。

图 5-1 整体逻辑架构图



其中存储控制器上部署 FastCube 混合闪存系统，通过对称 Active-Active 架构实现系统的负载均衡，以及配合动态自适应数据布局（DADL）技术，针对 SSD 和 HDD 相结合的混闪特点，进行深度优化，将数据以最优方式放置在不同介质上，充分发挥混闪存储系统性能。

存储控制器上部署了弹性虚拟交换机（EVS），实现计算节点与计算节点、计算节点与存储系统之间的高性能互联。

计算节点上安装新版本 FusionCompute 虚拟化软件，支持虚拟机资源的全生命周期管理。管理角色的计算节点上部署了统一管理平台 DeviceManager，实现系统的一站式极简管理。

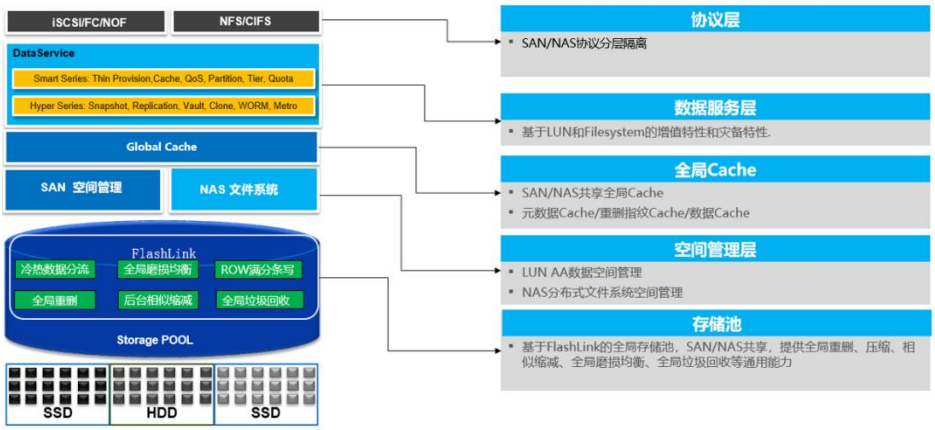
5.1 存储系统软件架构

本章节描述 FastCube 2910 计算型存储中存储控制器上部署的 FastCube 混合闪存系统软件架构，并部署了弹性虚拟交换机（EVS），提供虚拟网络交换功能，使用 DPDK(Data Plane Development Kit)实现计算节点与计算节点、计算节点与存储系统之间的高性能互联。

5.1.1 SAN/NAS 一体化统一存储架构

FastCube 混合闪存系统软件采用 SAN/NAS 一体化设计，不需要单独的 NAS 网关设备，采用 SAN 和 NAS 共 POOL 的方式用一套软硬件同时支持 SAN 和 NAS，支持 NFS、CIFS 文件访问协议和 iSCSI 块访问协议。

图 5-2 软件架构图



FastCube 混合闪存的系统架构自下而上包括以下子系统：

- 存储池(Storage Pool)：提供全局统一的存储池服务，采用 ROW(Redirect on write)技术进行空间分配，为 LUN 和文件系统写入的数据提供存储空间，对元数据和数据进行识别并分流写入 SSD 或 HDD，提供 RAID2.0+全局快速重构功能和全局的后台垃圾回收功能。
- 空间管理层(Space Management)：基于 Thin Provision 按需分配机制，分别为 LUN 和文件系统提供空间分配和回收管理服务。
- 全局 Cache(Global Cache)：提供全局统一的 Cache 服务，为 LUN 和文件系统提供读写缓存和元数据缓存。
- 数据服务层(Data Service Layer)：提供统一的数据复制、复制管理、配置管理和复制网络管理服务，为 LUN 和文件系统提供远程复制、双活等容灾能力。
- 协议层(Protocal Layer)：分别为 LUN 和文件系统提供协议服务，提供协议的解析处理、I/O 收发、错误处理。

FastCube 混合闪存的架构上，文件系统和 LUN 都是直接从存储池直接分配空间并提供服务的，属于 SAN/NAS 平行架构，即文件系统和 LUN 都直接与底层的存储池交互。相比传统在 LUN 基础上创建文件系统和在文件系统上创建 LUN 的架构，这样的软件架构为 LUN 和文件系统都能提供最短的 I/O 路径，让两者的存储效率更高，同时两者之间的空间管理保持独立，互不影响，增强了可靠性。

5.1.1.1 Active-Active 的 SAN 逻辑架构

在 ALUA（Asymmetric Logic Unit Access）架构中，LUN 有归属控制器，客户在创建 LUN 的时候，需要对 LUN 的归属进行规划，以尽可能的实现系统中每个控制器的负载均衡。但是由于不同 LUN 的业务压力各不相同，同一个 LUN 不同时段的压力也不相同，在实际情况下 ALUA 架构难以实现系统的负载均衡。FastCube 混合闪存软件采用 Symmetric Active-Active 架构。通过均衡算法，实现每个控制器接收到的主机读写请求是均衡的；通过全局缓存技术实现 LUN 无归属，每个控制器收到的读写请求，就在本控制器处理（而不像 AULA 存储需要转发到 LUN 归属控制器处理），实现了控制器压力均衡；通过 RAID2.0+技术，数据均匀的分布到存储池内的所有硬盘上，实现盘的压力均衡；

5.1.1.1.1 全局负载均衡

FastCube 混合闪存每个主机读写请求，应该被存储系统的哪个控制器处理，是通过对读写请求的 LBA 进行 HASH 计算来确定的。主机多路径（UltraPath）、前端共享接口模块、以及控制器，协商了相同的 HASH 计算方法和参数，实现读写请求的智慧分发。多路径软件和前端共享接口模块，会尽力把主机的读写请求分发到这个读写请求最终被处理的控制器，避免了在控制器内部的转发。

使用主机多路径软件时，多路径软件把主机 I/O 请求发送到最优的控制器上处理，I/O 请求也不需要在控制间转发。在没有使用多路径软件的场景下，接收到主机 I/O 请求的控制器，会根据对 I/O LBA Hash 的结果，转发到对应的控制器处理，实现各个控制器负载均衡。

5.1.1.2 Active-Active 的 NAS 逻辑架构

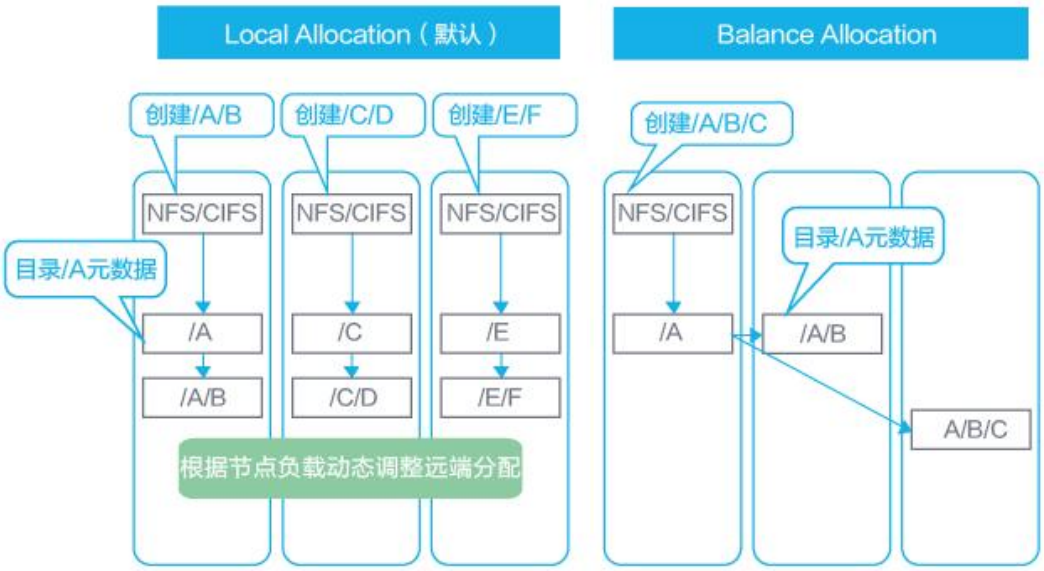
在传统 NAS 文件系统架构中，业界一般采用 Active-Passive 架构，文件系统归属某 1 个控制器，在创建文件系统的时候，需要对文件系统的归属进行规划，采用多个文件系统运行在不同的控制器上，实现系统中每个控制器的负载均衡。这样的 NAS 架构，如果只有 1 个文件系统只能发挥 1 个控制器的硬件性能，无法把多控制器的硬件资源都利用起来获得更高的性能，因此无法支持单一命名空间(SingleNameSpace)。如果采用创建多个文件系统，由于不同文件系统的业务压力各不相同，也很难以实现系统的负载均衡。

FastCube 混合闪存的 NAS 采用分布式文件系统架构，文件系统没有归属控制器，通过均衡算法将文件系统的目录和文件均衡写入每个控制器，实现每个控制器接收到的主机读写请求是均衡的，使得 1 个文件系统也能将整个存储集群的资源充分利用，客户可以根据自己的业务规划灵活的使用单一命名空间的文件系统或者多个文件系统。

5.1.1.2.1 分布式文件系统

FastCube 混合闪存的 NAS 分布式文件系统架构兼具海量小文件和大文件并存的文件共享场景，采用基于目录的分布式打散策略，以目录为粒度将数据均衡打散写入到各个控制器，达到负载均衡的目的。目录与目录下的子文件归属相同的控制器进行 I/O 处理，避免跨控制器转发，以提升目录遍历查询、属性遍历查询、批量属性设置等场景的性能。针对大文件，当大文件写入 FastCube 混合闪存的存储池时，通过 RAID2.0+的全局数据块打散技术，可以将大文件的数据块打散到存储池的所有硬盘上，提升大文件的写入带宽。

图 5-3 NAS 均衡策略

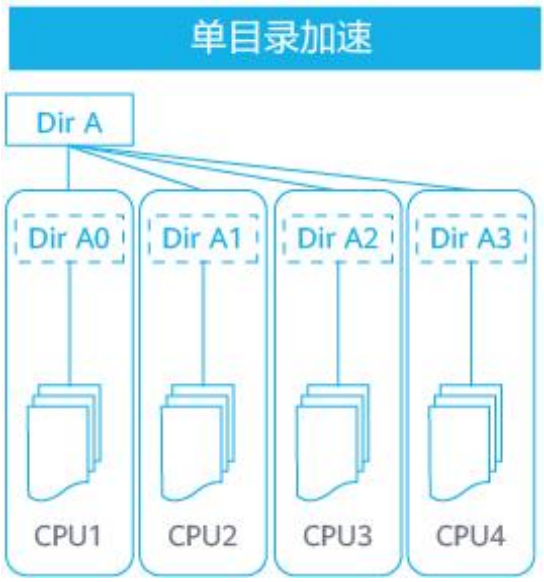


FastCube 混合闪存 NAS 支持两种目录均衡打散策略：

Local Allocation：主机写入的目录优先从所挂载 IP 的控制器直接写入和读取，以获得最佳性能和时延。系统后台根据各个控制器管理的数据容量、文件数量和负载，动态调整远端分配比例，即当某个控制器的负载、管理的数据容量或者管理的文件数量高于其他控制器一定阈值后，系统会自动将远端控制器分配比例调高，以让每个控制器的负载和管理容量达到均衡。此模式为 FastCube 混合闪存的默认分配模式，在存储的每个控制器的每个 IP 被主机均衡挂载时能获得最佳性能，适合 OA 文件共享、票据影像、EDA 后端仿真等对性能和时延要求较高的场景，推荐与 FastCube 混合闪存的内置 DNS 服务功能配合使用，内置 DNS 服务将自动为主机挂载进均衡的分配。

Balance Allocation：主机写入的目录直接进行按控制器负载均衡的打散写入，以在一开始就获得最佳的负载均衡能力，系统后台仍然会根据各控制器的管理容量等信息进行均衡调整。此模式适合主机挂载存储的 IP 地址不均匀、压力不均匀，应用与对时延不高的文件共享场景。

图 5-4 大目录性能均衡

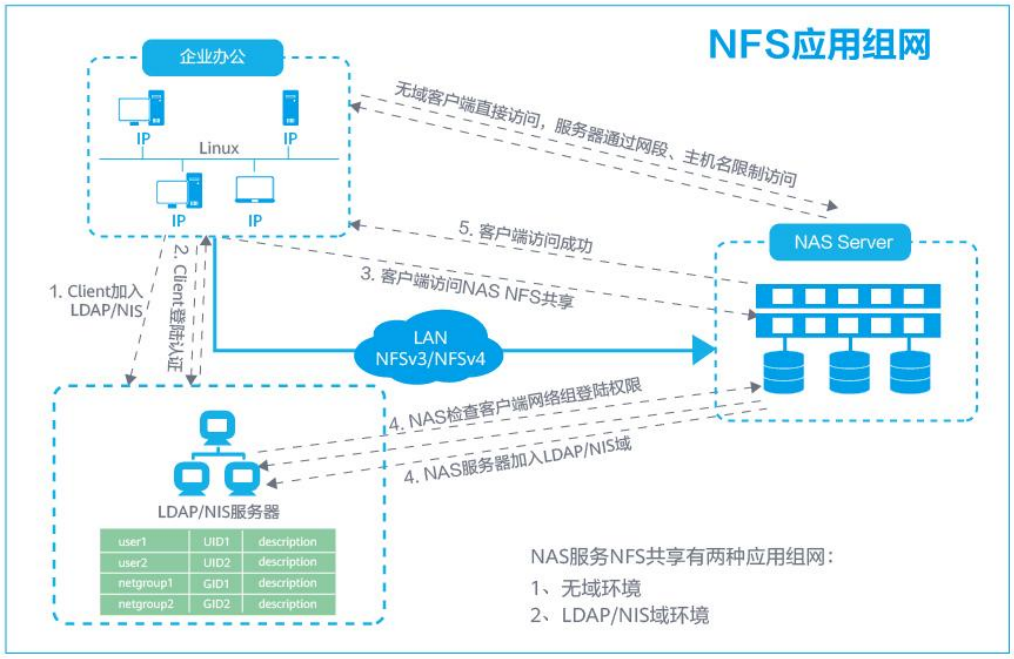


当系统存在热点目录，即某个目录和目录下的子文件访问十分频繁，FastCube 混合闪存支持将热点目录负载到多个 CPU 的多个核心上去并行处理，提高处理效率，避免热点效应。

5.1.1.2.2 NAS 协议

?.1.NFS 协议

图 5-5 NFS 应用组网图



NFS (Network File System)网络文件系统，是在 linux/ unix 环境里面普遍使用的网络文件共享协议。

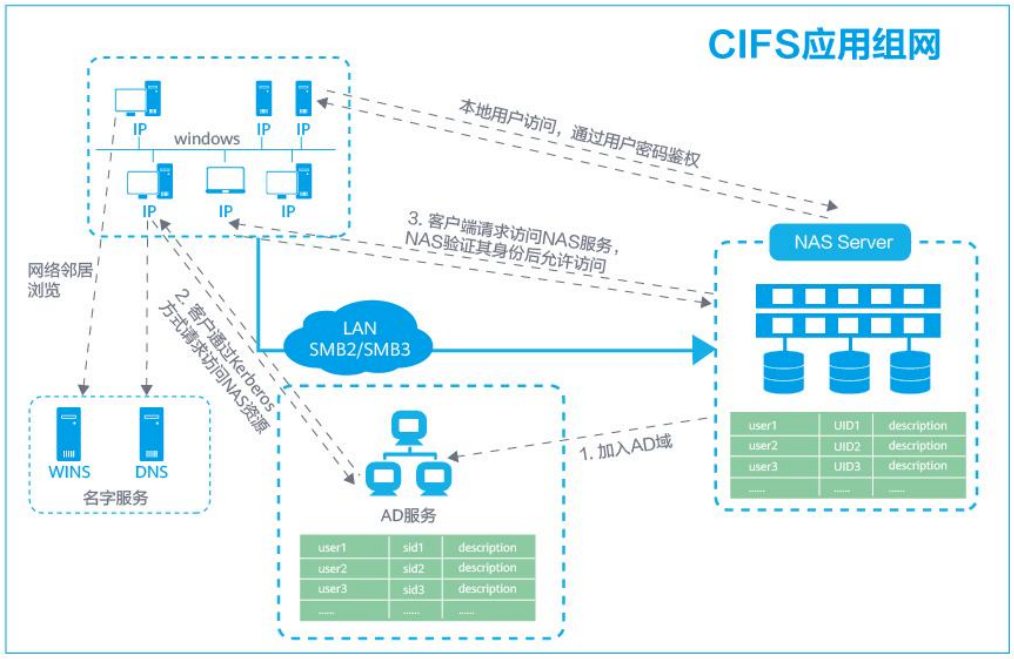
NFS 的应用场景主要有：

- 面向 UNIX/LINUX/AIX/SOLARIS 等类 UNIX 操作系统的网络文件共享。
- Vmware/XEN/虚拟机应用场景
- SAP HANA/Oracle 数据库应用场景

FastCube 混合闪存支持 NFS 协议版本包括 NFS V3、NFS V4.0、NFS V4.1(详细参考产品规格清单)，支持本地用户环境（无域环境）以及 LDAP/NIS 域控管理网络环境，LDAP 支持导入证书提供 LDAPS 的安全域传输。在多租户环境下，支持每个租户单独配置 LDAP/NIS 服务。

?.2.CIFS 协议

图 5-6 CIFS 应用组网图



SMB（Server Message Block）也称为 CIFS（ Common Internet File System ，公共互联网文件系统），是在 Windows 环境里面普遍使用的网络文件共享协议。

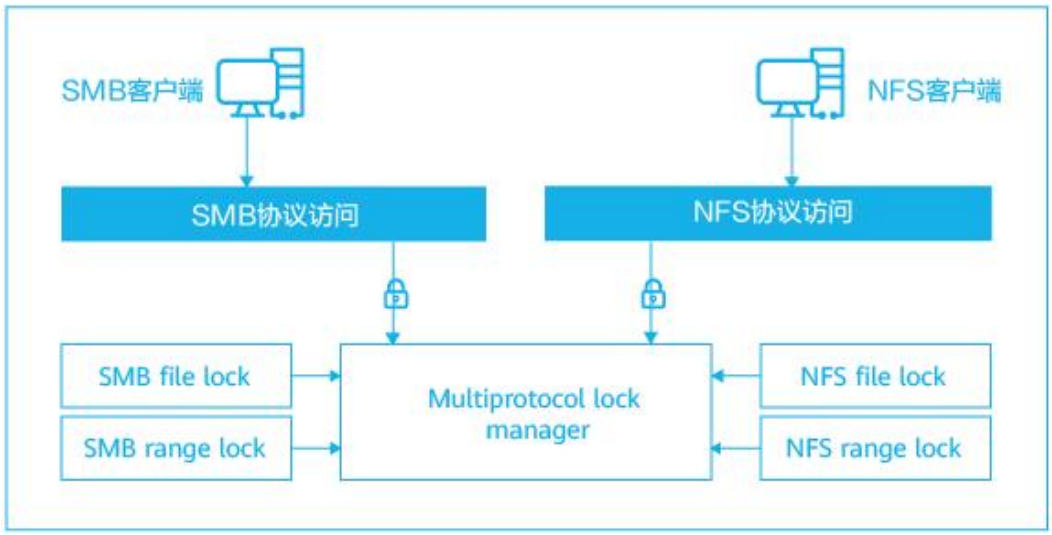
SMB 的应用场景主要有：

- 面向 Windows 操作系统的网络文件共享。
- Hyper-v 虚拟机应用场景

FastCube 混合闪存支持 SMB2.0/SMB3.0 协议，支持本地用户环境（无域环境）以及 AD 域控管理网络环境，能基于 AD 域的 Kerberos 和 NTLM 认证鉴权，支持单一域、父子域、信任域等 AD 域网络环境。在多租户环境下，支持每个租户配置独立的 AD 域。

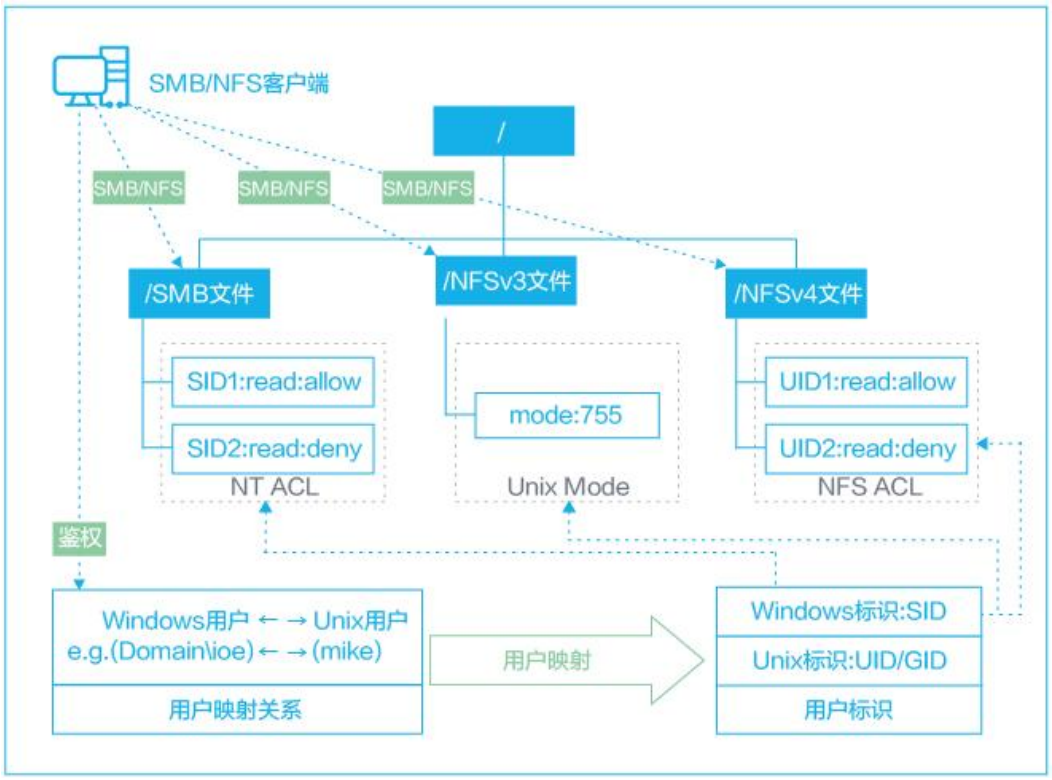
?.3.跨协议访问

FastCube 混合闪存支持 NFS/SMB 跨协议访问，能为一个文件系统既配置 NFS 共享服务又配置 SMB 共享服务，系统通过 Multiprotocal Lock manager 进行分布式锁管理，保证 NFS/SMB 能互斥的访问相同的文件，不会造成文件损坏或数据不一致。

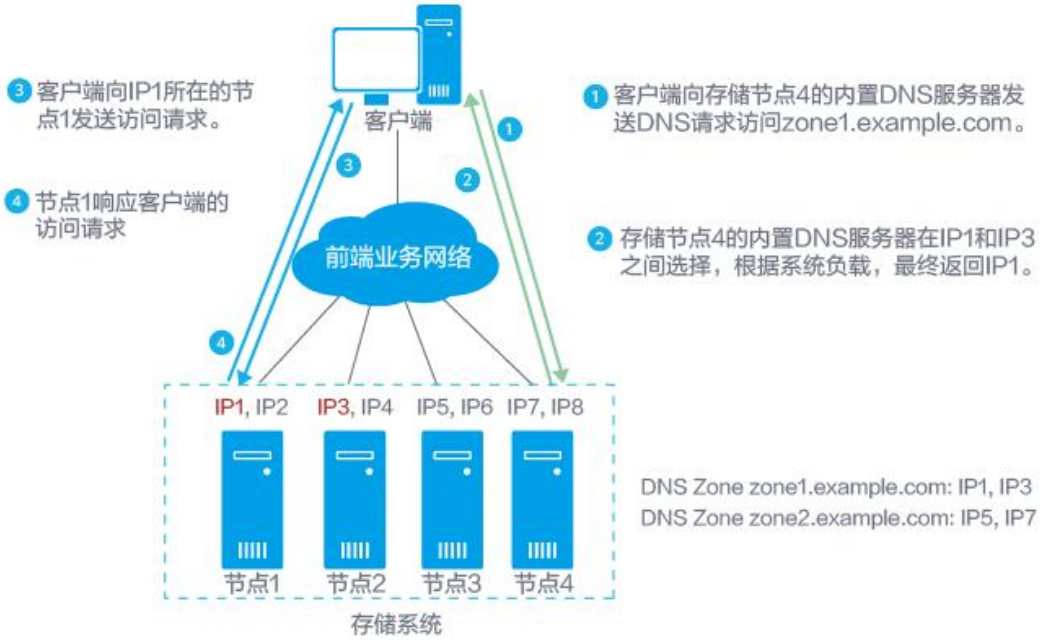


跨协议访问场景下的安全模式，FastCube 混合闪存支持 4 种安全模式：

- NT 模式：仅允许在 windows 端通过 SMB 设置文件的属性和 ACL 权限，系统会自动根据配置好的用户映射关系，将 SMB 用户的文件权限映射到 NFS 端，从而 NFS 端使用对应的 Linux 用户来访问该文件时可以成功鉴权。在 NFS 端该模式会禁止用户设置文件的 Mode 和 ACL 权限。
- UNIX 模式：仅允许在 Linux/UNIX 端通过 NFS 设置文件的权限，系统会自动根据配置好的用户映射关系，将 Linux 用户的文件权限映射到 SMB 端，从而 SMB 端使用对应的用户来访问该文件时可以成功鉴权。在 SMB 端该模式会禁止用户设置文件的 ACL 权限。
- Mixed 模式：UNIX 客户端和 windows 客户端均可进行权限设置，文件权限会相互覆盖，以最后设置的权限为准。
- Native 模式：UNIX 客户端和 windows 客户端均可进行权限设置，NFS/SMB 端权限独立存储，各自鉴权。



5.1.1.2.3 内置 DNS 负载均衡



在 NAS 环境中，主机可通过域名访问存储阵列上的服务。当同一域名下有多个 IP 地址时，可通过 DNS 实现不同 IP 间的负载均衡。外置的 DNS 服务器无法感知阵列上各 IP 所在节点的 CPU 利用率、所在端口的带宽利用率等负载情况，通常只能提供简单的 round-robin 策略，无法真实的反应各 IP 的负载情况，均衡效果不理想。

使用传统外置 DNS 服务器主要有以下问题：

- 只能提供 round-robin 均衡策略，无法感知阵列上各 IP 所在节点的 CPU 利用率、所在端口的带宽利用率等负载情况，均衡效果不理想。
- 依赖外置 DNS 服务器稳定性，阵列不能提供高可靠的域名解析服务。

FastCube 混合闪存提供一种新的内置 DNS 负载均衡技术，可以感知阵列上各 IP 的负载情况，获得更好的负载均衡效果，从而提升存储业务性能和可靠性。

FastCube 混合闪存 DNS 负载均衡特性支持的策略有轮循方式、按节点 CPU 利用率、按节点连接数、按节点带宽利用率、按节点综合负载，其选择参考如表所示。

名称	说明	优点	缺点
加权轮询	当客户端通过域名访问时，根据性能数据计算权重，同一域名下需负载的各 IP 以相等的概率选中处理客户端业务。	NAS 各业务本身差异不大时，均衡效果最佳。	• 仅在各 IP 间均衡，无法感知所在节点的实时负载情况。 • 任何 DNS 域名请求（如客户端的 ping、nslookup 请求或 showmount 命令等），超时或认证失败的业务连接都会影响负载均衡。
CPU 利用率	当客户端通过域名访问时，根据各节点 CPU 使用率性能数据计算权重，以权值作为概率选择节点处理客户端业务。	依赖的数据为 CPU 使用率，能够以权值概率来选择性能数据中 CPU 使用率最低的节点处理客户端业务，同时也能应对并发场景。	CPU 使用率需要各节点承担业务才会显著变化，先挂载的客户端必须跑业务，后续才能有效的按 CPU 利用率进行负载均衡，具有一定的时延性。
带宽利用率	当客户端通过域名访问时，根据各节点总带宽利用率性能数据计算权重，以权值作为概率选择节点处理客户端业务。	依赖的数据为节点总带宽使用率，能够以权值概率来选择性能数据中端口总带宽使用率最低的节点处理客户端业务，同时也能应对并发场景。	• 均衡粒度较粗，只能在各节点间负载均衡，不能细分到各物理端口上。 • 带宽利用率需要各节点承担业务才会显著变化，先挂载的客户端必须跑业务，后续才能有效的按端口带来利用率进行负载均衡，具有一定的时延性。

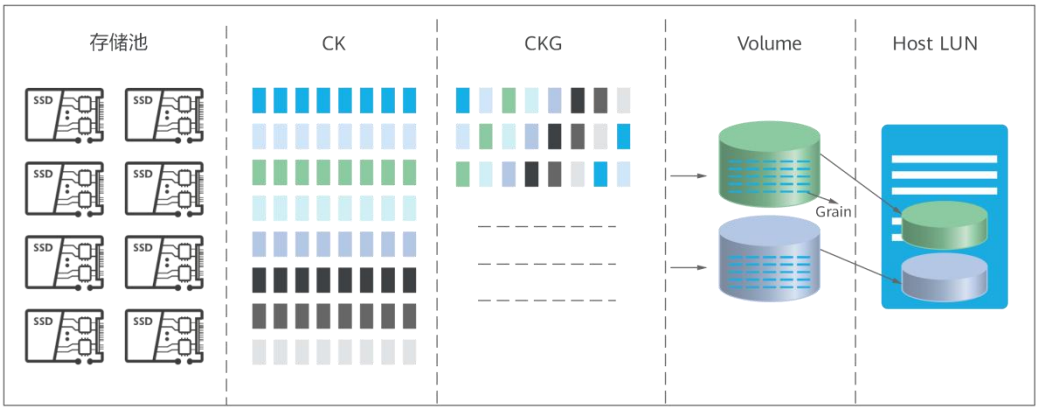
名称	说明	优点	缺点
NAS 连接数	当客户端通过域名访问时，根据各节点 NAS 连接数性能数据计算权重，以权值作为概率选择节点处理客户端业务。	依赖的数据为节点 NAS 连接数，能够以权值概率来选择性能数据中连接数最少的节点处理客户端业务，同时也能应对并发场景。	• 性能数据中的连接数是各节点的非实时性能数据，会影响客户端的负载均衡。 • 如果已挂载的 NFS 业务没有报文交互时，节点会清除连接信息，但节点的挂载点还存在，新的客户端根据连接数可能还会挂载到该节点，影响负载均衡。
综合负载	当客户端通过域名访问时，根据性能数据的节点综合负载选择节点处理客户端业务。根据 CPU 利用率、带宽利用率和 NAS 连接数计算节点负载，负载越低则被选中的概率越高。	• 综合负载能考虑到节点的 CPU 使用率和吞吐量，选择负载最低的节点来承担客户端业务。 • 综合负载不会一直选择负载最低的节点承担业务，而是让负载低的节点尽可能多的承担业务，逐步达到系统的负载均衡。	综合负载统计需要节点承担业务才会显著变化，先挂载的客户端必须进行业务，后续才能有效的按综合负载进行负载均衡，具有一定的时延性。

5.1.1.3 RAID 2.0+

存储的数据，最终都会存储到硬盘上，如果有些盘片上存放的数据不均匀，就可能导致某些压力大的硬盘成为系统的瓶颈。FastCube 混合闪存通过 RAID 2.0+技术，通过细粒度的划分，实现所有 LUN 的数据均衡的分布在每个硬盘上，实现盘的负载均衡。RAID 2.0+ 的机制：

- 多个硬盘组成一个存储池（Storage Pool）；
- 每个硬盘被切分成固定大小的 Chunk（简称 CK，通常大小为 4MB）进行逻辑空间管理；
- 来自不同硬盘的 CK 按照客户配置 RAID 冗余级别组成 Chunk 组（CKG）。
- CKG 再被划分为更细粒度的 Grain，通常为 8K，为 Voulume 使用的最小粒度

图 5-7 RAID2.0+映射图



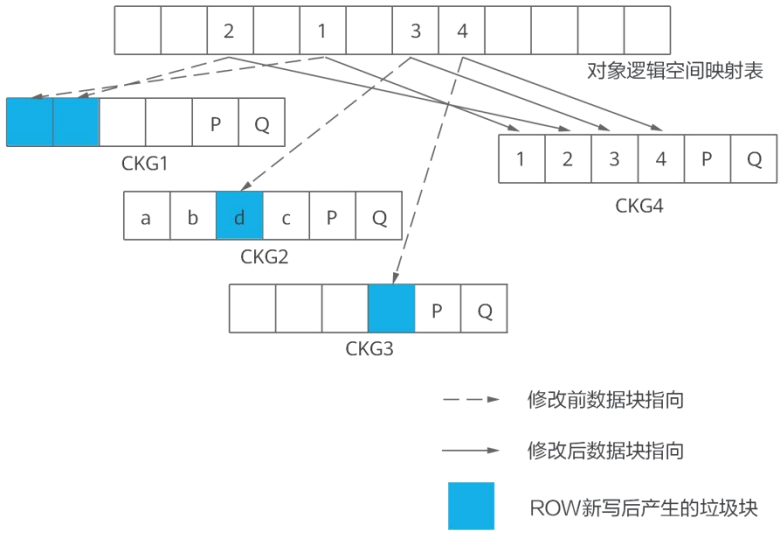
5.1.2 动态自适应数据布局（DADL）

动态自适应数据布局（DADL）技术的核心是以 ROW 大块顺序写架构为底座，打破传统机械盘的随机 IOPS 性能瓶颈，通过 Cache/Tier 弹性融合的性能层提供全场景的热数据加速，以多维智能的全局感知和冷热协同算法实现数据的最高效流动和放置，最大限度的发挥混闪系统的性能。其主要包含如下关键技术：ROW 大块顺序写、Cache/Tier 弹性融合性能层、多维智能的全局冷热感知和数据协同算法等。

5.1.2.1 ROW 大块顺序写

不管是 SSD 还是 HDD，大块顺序写都是比随机小 IO 写更加友好的写入模型。对于 HDD，由于机械特性的限制，磁道摆臂和寻道速度限制了其随机 IO 的吞吐能力，IOPS 非常有限。尽管 NL_SAS 盘目前可提供的单盘容量越来越大，但受限于其较低的 IOPS 能力，单位 TB 提供的 IOPS 性能密度往往无法满足高性能应用的要求，因此性能稍好一些但容量小很多的 10K SAS 盘仍以补充性能密度的方式长期存在。而 FastCube 全混合闪存采用全新的 ROW 大块顺序写机制，将随机小 IO 访问转换为顺序大块写，避免了传统 RAID 写流程所需的数据读和校验修改而产生 RAID 写惩罚，直接打破了机械盘传统访问模式的性能瓶颈，可大幅减少为满足性能密度而大量堆砌的硬盘数量，同时有效降低了写入过程阵列控制器的 CPU 开销。相比传统的 RAID 覆盖写的方式，ROW 大块顺序写方式使得各种 RAID 级别都能实现高性能。

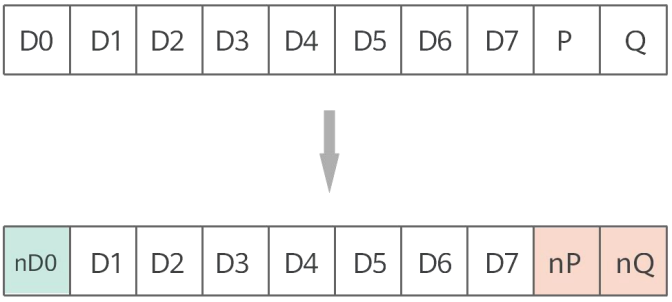
图 5-8 ROW 大块顺序写



上图以 RAID6（4+2）为例，对已有数据进行改写，改写写入的数据为 1、2、3、4。采用传统的覆盖写方式，对每个数据所在的 CKG 均需要进行修改写。以 CKG2 为例，写入新数据 3 时，需要读取校验列 P、Q 和原始数据 d，通过冗余算法计算出新的校验位 P’、Q’，再把 P’、Q’和数据 3 写入 CKG2 中。而采用 ROW 大块顺序写设计，写入数据 1、2、3、4 时，直接使用数据 1、2、3、4 计算出 P、Q 作为一个新的 RAID 分条写入硬盘，再修改 LBA 的指针指向新的 CKG，整个过程无需额外的预读。

对于传统 RAID，以 RAID 6 为例，D0 数据发生变化，需要先读 D0、P 和 Q；再写新的 nD0、nP 和 nQ，因此其读放大是 3，写放大也是 3。通常对于传统 RAID（xD+yP）的随机小 I/O 写其读写放大为 y+1。

图 5-9 传统 RAID6 的写放大



下表为各种传统 RAID 级别的写放大数据。

表 5-1 传统 RAID 的写放大

传统 RAID 类型	随机小 I/O 写产生的写放大	随机小 I/O 写产生的读放大	顺序写 I/O 写放大

传统 RAID 类型	随机小 I/O 写产生的写放大	随机小 I/O 写产生的读放大	顺序写 I/O 写放大
RAID5(7D+1P)	2	2	1.14（8/7）
RAID6(14D+2P)	3	3	1.14（16/14）
RAID-TP（传统 RAID 不支持）	-	-	-

FastCube 全混合闪存，以 RAID 6 采用 22D+2P（P、Q 校验列）、RAID-TP 为 21D+3P（P、Q、R 校验列）为例，下图展示了其采用 ROW 大块顺序写在典型场景下的写放大比较。

表 5-2 ROW 大块顺序写放大率

	随机小 I/O 写产生的写放大	随机小 I/O 写产生的读放大	顺序写 I/O 写放大
RAID 6（22D+2P）	1.09（24/22）	0	1.09
RAID-TP（21D+3P）	1.14（24/21）	0	1.14

5.1.2.2 Cache/Tier 弹性融合性能层

传统 Cache 技术和 Tier 技术是两种独立的加速技术，Cache 往往用于加速一部分读的热数据，Tier 尽管可以加速读和写，但冷热识别的精度和数据迁移的频度往往差强人意，并且 Tier 的各种策略和迁移周期很多时候需要基于对应用的了解，通过人工进行配置才能取得较好效果。同时这两种技术相互割裂，往往都需要自己独立配置高性能介质，各自的算法和策略也自成一套，既不灵活，也很难通过协同获得最佳的效率。FastCube 全混合闪存提供了一个 Cache/Tier 弹性融合的性能层，打破 Cache 和 Tier 的边界，通过整体协同实现最优的数据布局。

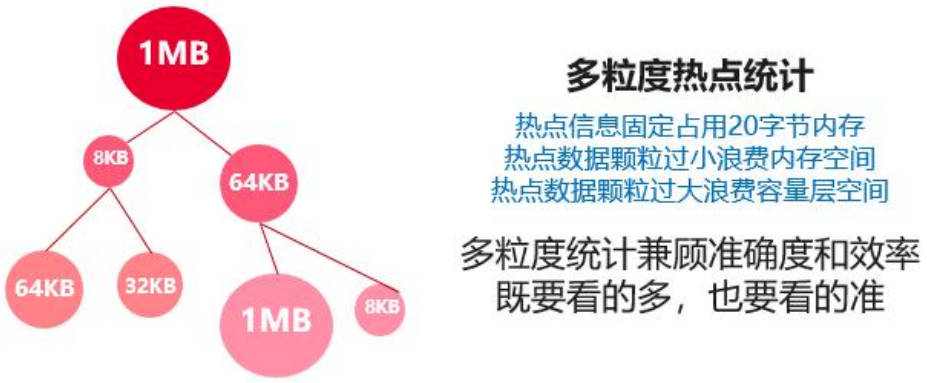
FastCube 全混合闪存的性能层不再需要 Cache 和 Tier 各自独立的物理盘配置，由统一的可弹性伸缩的加速层配额进行自动配置，Cache 和 Tier 可基于场景相互转化。同时性能层采用全局统一的冷热感知和数据协同算法，避免了传统 Cache 和 Tier 技术由于各自算法割裂带来的热点采集点不一致、统计粒度不一致、CPU 和内存重复浪费等低效行为。

图 5-10 Cache 和 Tier 深度融合

5.1.2.3 多维智能加速算法

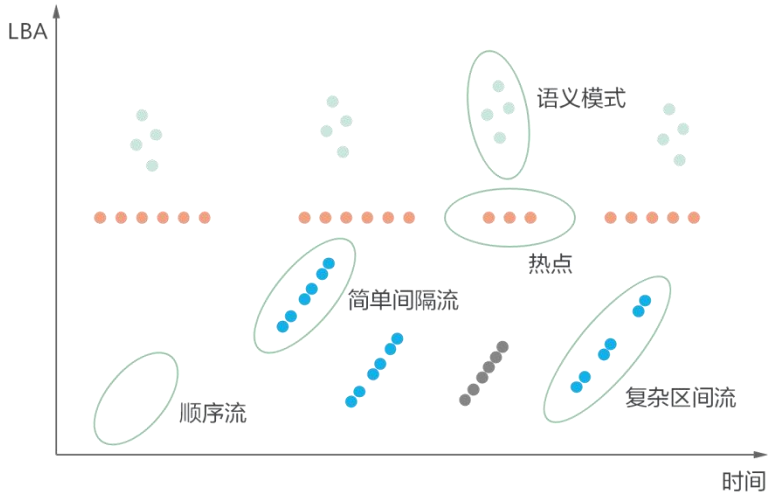
FastCube 全混合闪存采用了多维智能的全局冷热感知和数据协同算法，配合统一的性能层实现“全场景、全范围、全模型”的冷热数据感知，将混闪系统的性能发挥到最佳。

- 传统的热点统计算法基本都采用固定的数据结构，混合闪存热点统计算法采用了多粒度学习型的统计结构，可随模型的变化自适应调整数据结构，能更好的适应大小 IO 模型的转换和顺序随机模型的变化，在各种业务下都能实现效率最大化。



- 传统的冷热感知算法通常只考虑历史统计，全混合闪存采用多个维度的特征分析来识别全量热点，除传统的历史统计之外，还纳入了机器学习的时序预测算法。对于冷热变化的场景，基于预测结合历史统计来综合判断数据应该放置到性能层还是容量层，可实现多维特融合的冷热感知，确保所有的热数据都能被有效识别和加速。
- 传统的数据预取通常只考虑顺序流预取，FastCube 全混合闪存除了传统的顺序预取外，也兼顾了间隔顺序流、IO 关联流等多种流式模型的预取。

图 5-11 业务流模式



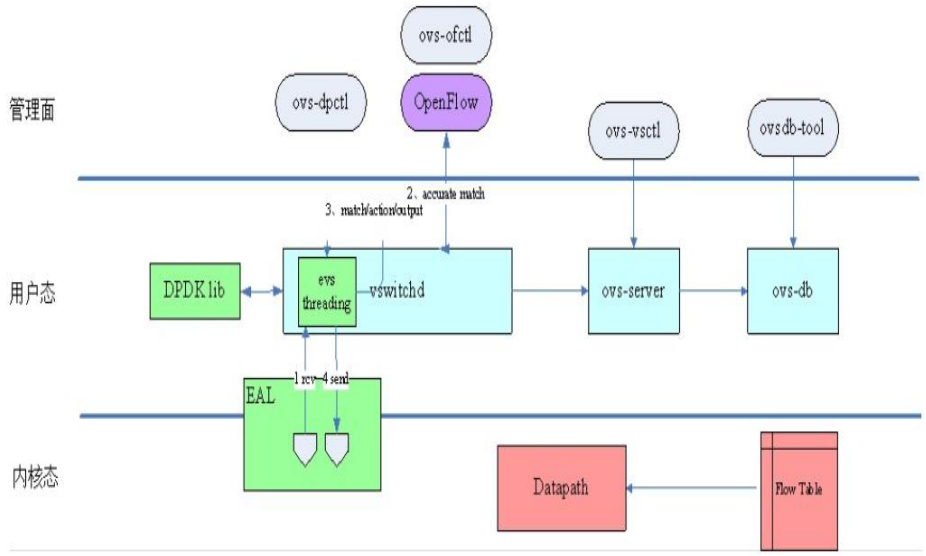
- 传统的 Cache 和 Tier 技术通常只基于单一的业务场景进行冷热感知，FastCube 全混合闪存结合全局 Workload 在更多的维度上进行系统资源的动态调配，例如 SAN 和 NAS 共存的场景，可基于各自不同的负载动态调节算法资源，以更好的匹配混合型场景加速。

由于采用多个维度的智能化全新冷热感知和数据协同技术，FastCube 全混合闪存的性能层能够适应更加灵活的业务模型，适应更加多变的冷热变化，适应更加多样的数据关联，最终实现最佳的数据流动和分层放置。

5.1.3 弹性虚拟交换机（EVS）

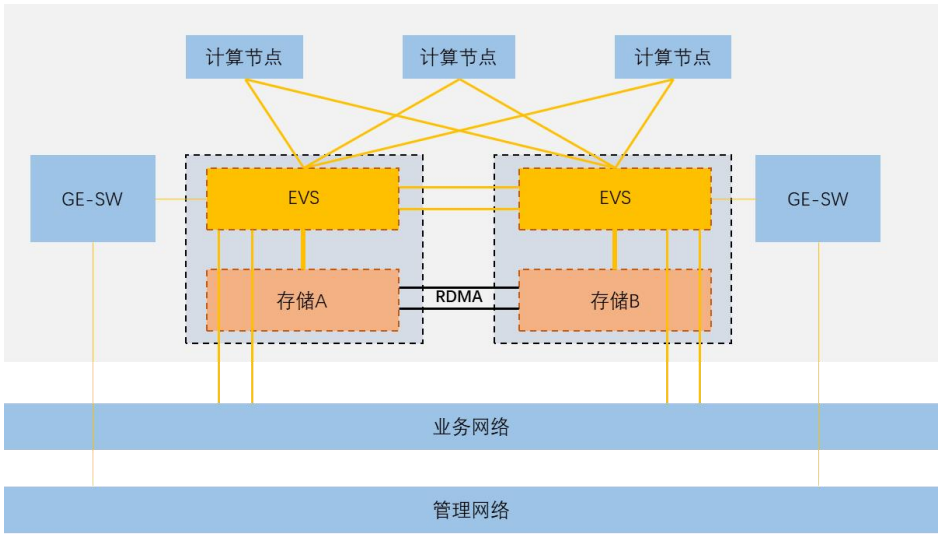
EVS(Elastic Virtual Switch)是华为研发的，基于 OVS 转发技术，通过使用 DPDK 数据收发技术提升其 IO 性能的弹性化虚拟交换机。通过用户态进程接管网卡数据收发，采用“IO 独占核”和大页内存技术，即每个端口分配一个核专门用于数据收发，这种轮询式的处理方式显然比中断式的处理更高效，结合预留的大页内存减少内存数据拷贝次数，收发包和数据转发都在用户态完成，减少了内核态与用户态间切换带来的开销，因而 IO 性能方面有显著提升。

图 5-12 EVS 高性能数据转发架构



FastCube 2910 计算型存储通过在存储控制器部署 EVS，实现系统内部计算节点与计算节点、计算节点与存储系统、系统外部管理网络和业务网络的互联互通。

图 5-13 内部交换网络架构



5.1.4 丰富增值特性

FastCube 2910 计算型存储系统提供了用于系统效率提升的 Smart 软件系列、用于数据保护的 Hyper 系列软件，实现数据的全生命周期管理。

- 效率提升系列（Smart 系列）：智能精简配置（SmartThin）、压缩（SmartCompression）、服务质量控制（SmartQoS）、智能数据迁移

（SmartMigration）、数据销毁（SmartErease）、智能配额（SmartQuota）、智能加速（SmartAcceleration）、多租户（SmartMulti-tenant），主要为用户提供存储效率提升方面的功能，降低用户的 TCO。

- 数据保护系列（Hyper 系列）：安全快照（HyperSnap）、WORM（HyperLock）、持续数据保护（HyperCDP）、克隆（HyperClone）、一体化备份（HyperVault）、远程复制（HyperReplication），主要为用户提供数据容灾备份、安全相关的功能。

5.2 增值特性：Smart 系列

FastCube 2910 计算型存储的存储系统提供了丰富的 Smart 系列软件。包括提升提升空间使用效率的软件特性，智能精简配置（SmartThin）、数据缩减（SmartCompression）；提升性能提供不同业务的服务质量的软件特性，服务质量控制软件（SmartQoS）、智能加速（SmartAcceleration）、多租户（SmartMulti-tenant）、智能配额（SmartQuota，仅限 NAS）；对系统进行生命周期管理和数据安全的软件特性，智能数据迁移（SmartMigration，仅限 SAN）、数据销毁（SmartErase）。

5.2.1 数据缩减（SmartCompression）

FastCube 2910 计算型存储支持 SmartCompression 特性系统根据用户数据特征自适应进行数据压缩，能够有效节省存储空间。本章主要介绍数据压缩的实现原理。

5.2.1.1 压缩

数据压缩主要包括两个过程，首先是采用压缩算法对输入的数据块进行压缩，形成一个个的小数据块，然后再把这些压缩后的数据块拼接在一起（数据压紧，Data Compaction）下盘。

5.2.1.1.1 压缩处理

本章主要介绍压缩处理的实现原理。

压缩预处理

FastCube 2910 计算型存储在数据压缩前，采用压缩预处理算法，根据数据格式等方法识别出待压缩数据中较难压缩的部分（使用通用压缩算法压缩效果差的部分），并对压缩数据进行重排，从而获得达到更高的压缩率。

通过预处理，待压缩的数据会分成两部分：

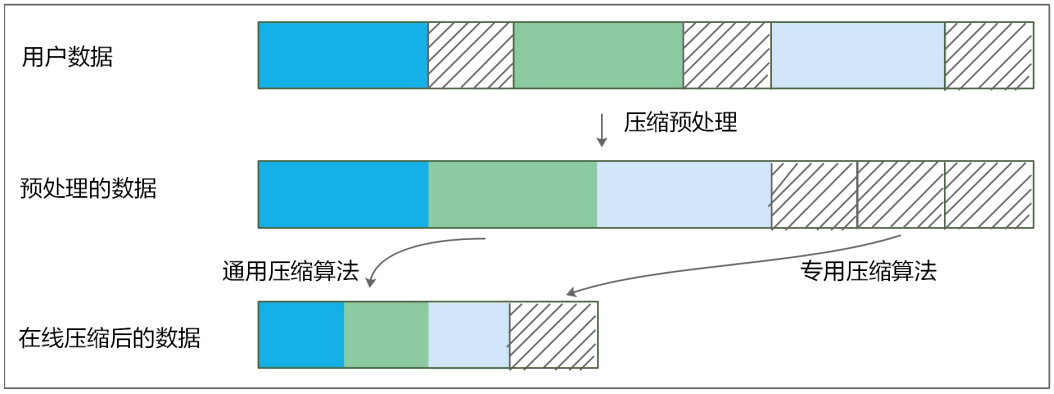
- 较难压缩的部分，将使用专用压缩算法进行压缩。
- 待压缩数据的其他部分，使用通用压缩算法进行压缩。

专用压缩

根据压缩预处理的结果，对于较难压缩的部分，FastCube 2910 计算型存储采用专用压缩算法进行压缩。

专用压缩算法采用特殊的编码规则，在不增加元数据的同时，对通用压缩算法难以压缩的数据进行压缩。专有压缩算法拥有高性能的特性，不影响在线读写。压缩预处理及专用压缩的效果原理如图 5-14 所示。

图 5-14 压缩预处理及专用压缩



通用压缩

FastCube 2910 计算型存储提供的在线压缩技术，采用压缩算法。该算法采用了 LZ 匹配+Huffman 熵编码算法。保证算法实现在相同性能情况下，提供了比 LZ4 算法的高 20%左右压缩率。

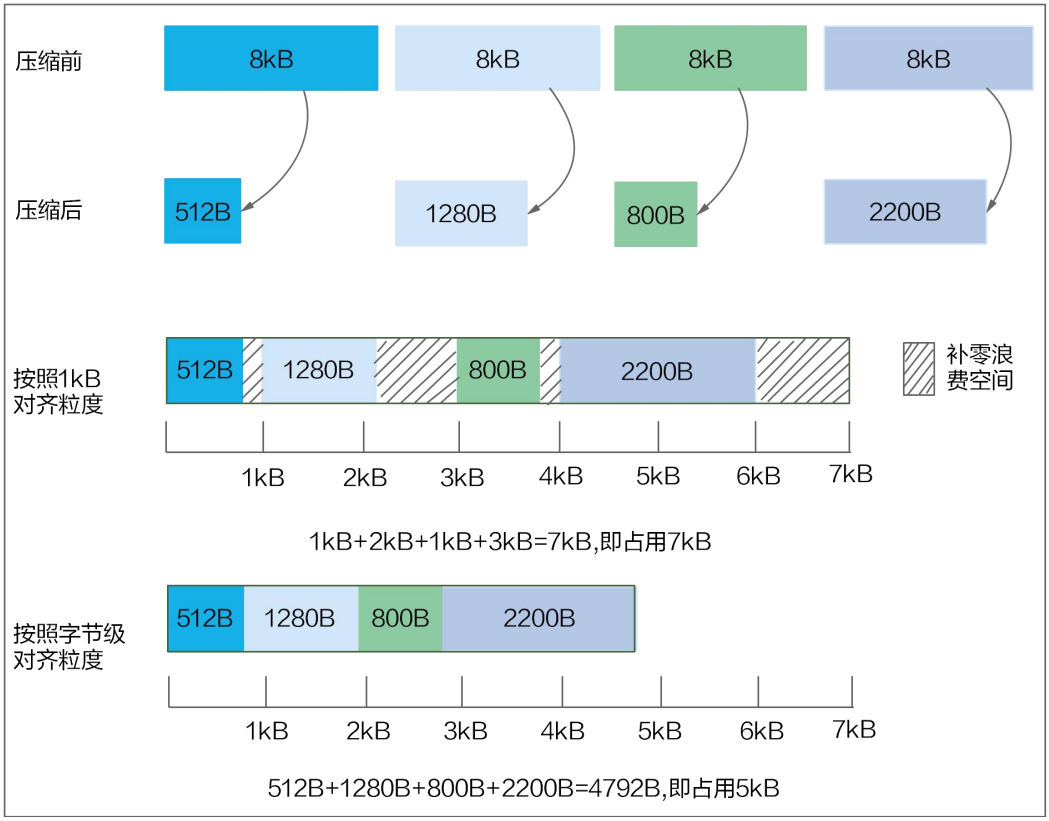
5.2.1.1.2 数据压紧（Data Compaction）

本章主要介绍数据压紧的实现原理。

FastCube 2910 计算型存储压缩后的用户数据按照字节级对齐，进行数据压紧，减少浪费物理空间，相对于 1kB（相对业界主流对齐粒度）粒度对齐，提供更高缩减率。

如图 5-15 所示，通过字节级对齐，减少物理空间的浪费，相对于 1kB 粒度对齐节省了 2kB 物理空间，从而提升压缩率。

图 5-15 压缩后数据字节级对齐原理



5.2.2 服务质量控制（SmartQoS）

SmartQoS 可以通过动态地分配存储系统的资源来满足某些应用程序的特定性能目标。SmartQoS 特性允许用户根据应用程序数据的一系列特征（IOPS、占用带宽、响应时延）对特定应用程序设置特定的上限/下限目标。存储系统根据设定的目标，准确限制应用程序的性能，避免非关键应用程序抢占过多存储系统资源，影响关键应用程序的性能。

SmartQoS 采用基于被控对象的 I/O 优先级调度技术和 I/O 流量控制技术（上限流控、下限保障）两种方式来保证数据业务的服务质量，其可支持的被控对象及配置项如下表：

表 5-3 SmartQoS 策略及被控对象

特性	被控对象	配置项
上限流控（含突发流控）	SAN：LUN、快照、LUN 组、HOST、 NAS：FS	IOPS、带宽
下限保障	SAN：LUN、快照、LUN 组 NAS：FS	IOPS、带宽、最大时延

5.2.2.1 功能特性

质量控制服务，包括上限流控和下限保障。上限控制主要用于多业务部署的情况下防止部分业务的流量过大照成其它业务无法正常运行的场景，即防止扰邻；下限保障则更多用于多业务部署情况下对关键业务的保障作用个，尤其是延时敏感的业务进行资源保障，确保该业务的访问响应能力。

5.2.2.1.1 上限流控

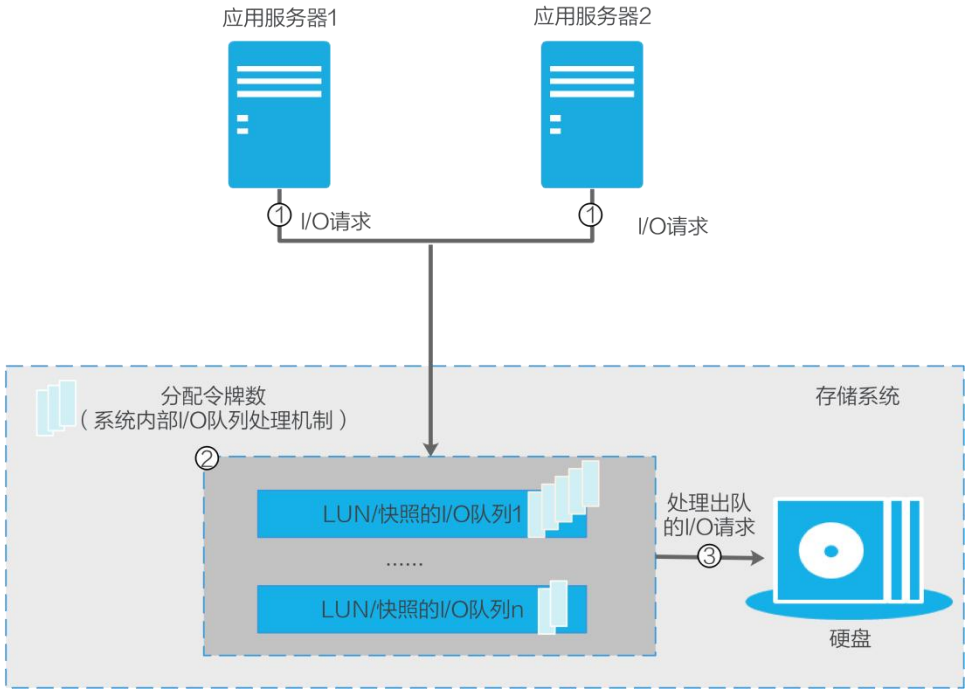
SmartQoS 的 I/O 流量控制技术需要用户配置 SmartQoS 策略，并往策略中添加被控对象，被控对象包括 SAN 资源对象（LUN/快照、LUN 组、主机/主机组）和 NAS 资源对象（FS）。通过限制策略中对象的的总体 IOPS、带宽，来达到限制系统中某些应用的性能，避免这些应用由于突发流量过大，影响系统中其它业务的正常性能。

SmartQoS 流控管理通过对控制对象中 I/O 队列管理、令牌分发和出队控制三部分实现。

当用户为某个 SmartQoS 策略设置性能上限目标，系统就会根据性能上限目标分配一定数量的令牌，通过控制令牌的发放来实现流控。在存储系统中，如果用户要限制的流量类型是 IOPS，那么一个 I/O 会根据转换关系转换成标准 I/O（大小为 8KB 的 I/O），并消耗对应数目的令牌；如果设定的性能目标是带宽，那么一个字节对应一个令牌。

基于 I/O 队列管理通过令牌机制实现存储资源的分配，某个对象的 I/O 队列所拥有的令牌数越多，系统分配给这个 LUN/文件系统或快照的 I/O 资源也越多。实现原理如图 5-16 所示。

图 5-16 上限流控



1. 应用服务器 I/O 下发到相应的被控对象的 I/O 队列中。

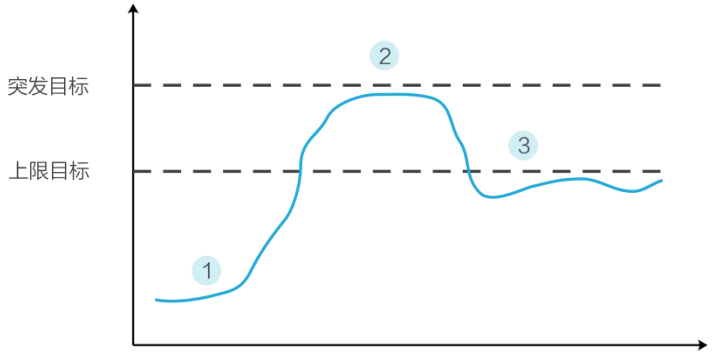
2. 存储用户可根据被控对象的业务流量配置不同的流量控制策略，通过将某些对象加入配置上限目标的流量控制策略，保证其不占有系统过多资源，影响其它业务的服务。
3. 处理出队的请求时，仅当请求拿到令牌时才能出队进入系统处理。

突发流控管理

作为上限流控的一种特别增强，针对部分时延非常敏感的业务，允许其短时间突破上限流控目标。SmartQoS 提供了突发流控管理能力，支持对被控对象配置突发能力，指定其突发期的 IOPS、带宽、突发时长。

其工作原理为积攒下前期未消耗的性能，当业务压力突然增大的情况下，可消耗此前积攒下的性能，短时间获得突破上限目标的性能。长时间来看，其平均流量应低于上限目标的。

图 5-17 上限突发



1. 当对象在过去时间段内流量未达到上限值，则可在未来时间段内系统整体没有过载的情况下该对象的业务流量短暂的超过 QoS 配置上限，满足业务峰值诉求达到突发值。突发的最大时长和比例可配置。
2. 突发流量控制是通过积攒突发时长来实现的，当对象在某一秒内性能低于上限性能目标一定阈值时，则积攒一秒的突发时长，当业务压力突然增长时，性能可突破上限值达到突发流量，突发时长为此前积攒的突发时长，最长不超过配置值。
3. 消耗完积攒时长或达到配置的最长突发时长，则流量被控制到上限值。

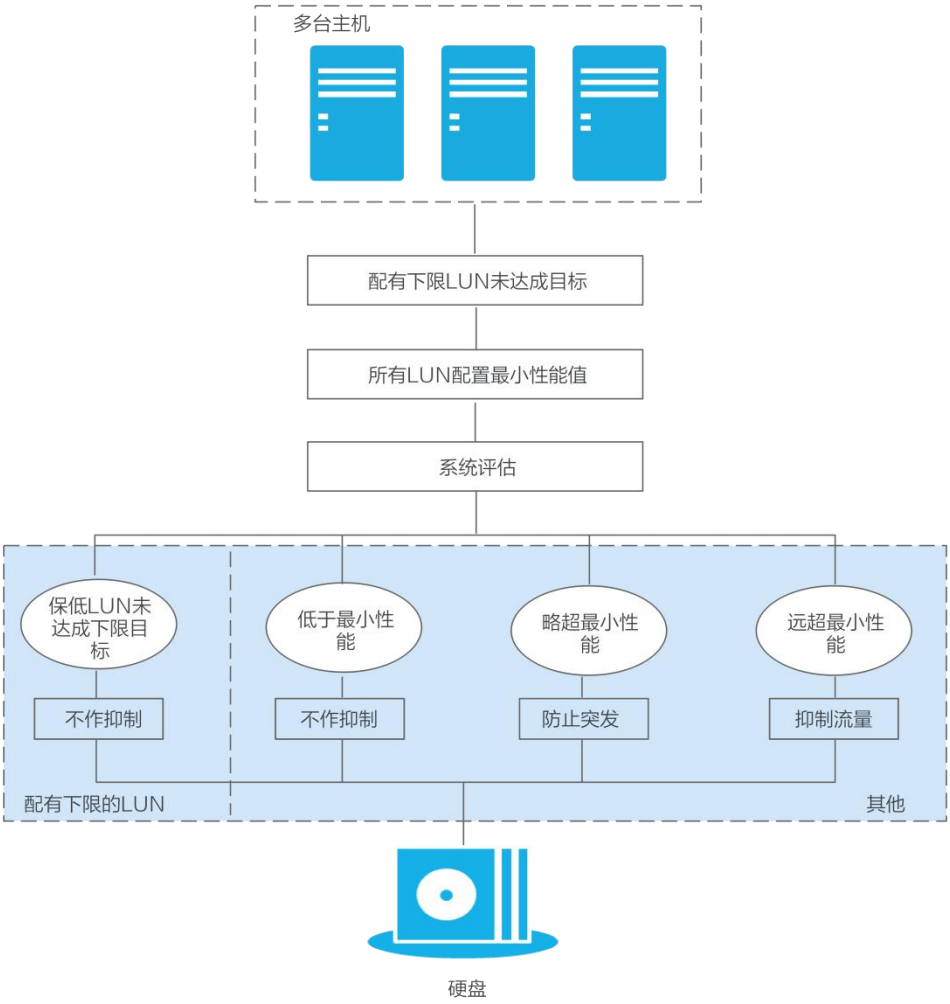
5.2.2.1.2 下限保障

下限保障功能专用于保障系统关键业务。下限保障工作原理是针对系统内所有业务对象（支持 LUN、快照、LUN 组、文件系统）进行流量控制，以释放足够资源给未达成下限目标的对象，尽可能达成对象的下障保障目标。由于其工作原理是尽力保障的过程，需要达到较好效果，仅适宜应用于少量关键业务。针对所有业务配置下限保障最终会导致所有业务均无法保证。

当系统中存在被控对象未达成下限目标时，针对系统所有对象进行负载评级，针对中低负载对象依据负载情况给定较宽松的流量条件，高负载对象则依据其当前流量按下限保障缺口进行流量抑制，直至其释放出足够资源使系统中所有对象达成下限目标。

下限达成但性能未远超下限的对象，仅需要防止其突发流量即可。下限未达成对象不作限制，充许其抢占抑制高负载对象释放资源。

图 5-18 LUN 下限流控



时延保障则是针对配有时延目标的对象的请求提升优先级，在系统内部作缓存。此外，时延保障未达成情况下，会将时延目标转化为流量控制，依照流量下限保障的方法进行保障。

5.2.2.2 策略管理

5.2.2.2.1 分层管理

SmartQoS 流量控制策略可以分为两种类型：普通策略和分层策略。分别如下：

- 普通策略：策略中的对象仅能为对象的策略叫普通策略，主要用于管理单一应用的流量控制策略。如：VDI 应用中存在短暂的启动风暴，在启动风暴的时间段配置普通策略，避免影响其他在线业务。
- 分层策略：可以添加普通策略作为其对象的策略叫分层策略，主要用于管理多种应用混合的流量控制策略。如：VMware 虚拟机场景，客户配置了虚拟机业务之

后，然后又对虚拟机的某一个 vmdk 做上限控制。这需要用户使用分层策略对虚拟机设置分层策略，对虚拟机内的 vmdk 设置普通策略。

说明

- 流控控制的对象，包括：LUN 或快照，LUN 组，主机，文件系统；一个普通策略中只能包含这几类对象中的一种。
- 分层策略可以包含多个普通策略，每个普通策略可以包含不同类的对象；
- 每一个 LUN 或快照只能加入一条流量调控策略。
- 每一个 LUN 组只能加入一条流量调控策略。
- 每一个主机只能加入一条流量调控策略。
- 每一个文件系统只能加入一条流量调控策略。

两者关系如下面两图以 LUN 关联资源讲解所示：

图 5-19 每个策略可添加 1-512 个 LUN 或快照

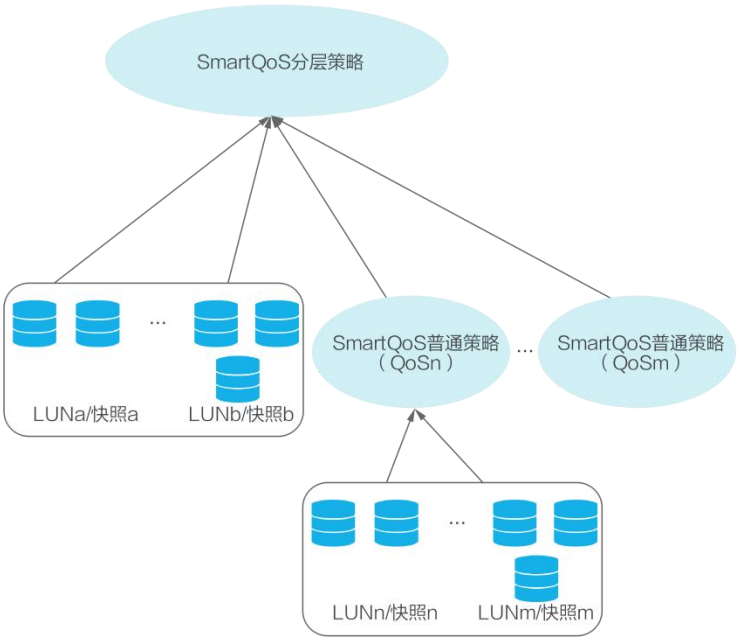
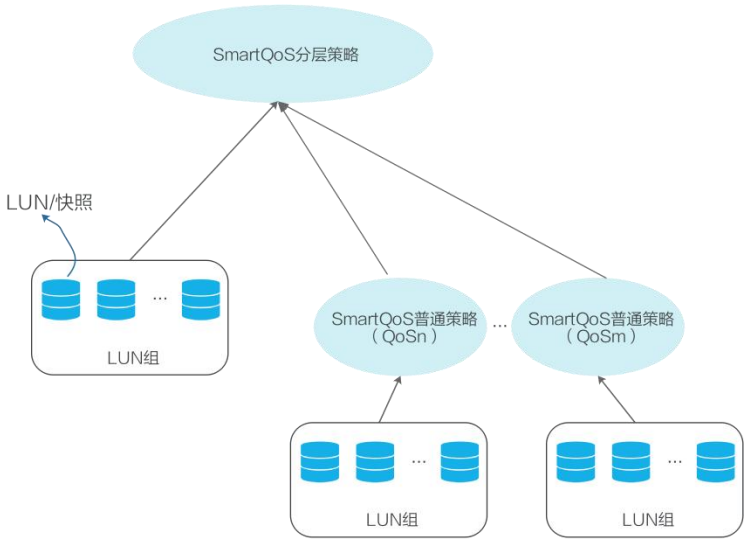


图 5-20 每个策略仅可添加单个 LUN 组



策略具体规格参考产品的规格清单。

5.2.2.2.2 策略分配

SmartQoS 策略用户配置的上限目标值为共享型，即添加到策略中的所有对象共享此策略的流控目标。所以 SmartQoS 模块会定期采集流量控制策略中所有对象的性能数据及需求数据，然后通过分配算法，将流量控制策略总体的控制目标分配给策略中具体的对象。

目前采用的分配算法为经过优化后的 max-min 权重分配算法，主要思路是通过识别策略的上限目标值、下限目标值、对象的需求数据，决策对象的实际目标值。该算法优先满足对象的需求量，所谓需求量是指配置对象接收到的请求数，包含成功数和拒绝数。剩余的上限目标值将会均分至各对象，并对连续多个周期的分配结果进行滤波平滑处理，使得最终结果相对稳定。

配置对象重叠处理:

一个 LUN/文件系统或快照可以单独加入一条流量控制策略，也可以通过一个 LUN 组被加入一条流量控制策略，当一个 LUN/文件系统或快照被加入多条流量控制策略中时，该 LUN/文件系统或快照的上限分配值将会取多个上限分配值中的较小值。

5.2.2.2.3 推荐配置

上限配置

- 针对多租户场景，可针对不同租户配置不同的分层策略，确保单租户占用的资源不超过限定值。针对单租户下各不同业务，可配置普通策略，限定其不同业务的性能上限及下限。
- 针对混合负载业务场景，可以针对非关键业务尤其是业务压力波动大的业务设置上限能力，针对关键且时延敏感的业务设置下限。
- 针对对于业务不均匀又对延迟敏感的工作负载，建议配置突发策略。

下限配置

针对系统中的少量关键业务，可配置下限策略，设置业务的最小 IOPS、最小带宽、最大响应时延。由于下限策略的工作原理是在系统资源紧张时抑制大部分非关键业务来保障少数关键业务，当由于系统资源紧张导致下限策略无法保障时，系统会发送告警，用户可依据告警提示信息调整下限策略。所以应尽量确保配置下限策略的业务总性能不要占用超系统性能的 50%，所有业务均配置下限策略与不配置没有区别。

5.2.3 智能数据迁移（SmartMigration）

FastCube 2910 计算型存储通过 LUN 迁移（SmartMigration）提供了智能化的数据迁移手段。可以在不中断原有业务的情况下实现将源 LUN 上的数据完整地迁移到目标 LUN 上。LUN 迁移不仅支持存储系统内部的数据迁移，还支持存储系统和与其兼容的异构存储系统之间的数据迁移。

SmartMigration 特性通过把源 LUN 的数据完整的复制到目标 LUN，在复制过程中采用源 LUN 和目标 LUN 双写、差异日志记录等技术，复制完成后采用 LUN 信息交换由目标 LUN 接管源 LUN 业务，实现数据的在线迁移。

SmartMigration 的实现过程分为数据同步和 LUN 信息交换两个阶段。

数据同步

1. 迁移前，客户需要配置迁移的源 LUN 和目标 LUN。
2. 迁移开始时，数据由源 LUN 后台复制到目标 LUN。
3. 主机此时可以继续访问源 LUN。主机写入源 LUN 数据时，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。
4. 写入的数据同时向源 LUN 和目标 LUN 双写。
   - 等待源 LUN 和目标 LUN 的写处理结果都返回。如果都写成功，清除日志；否则保留日志，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。
   - 返回主机写请求处理结果，以写源 LUN 的处理结果为准。
5. 在数据完全复制到目标 LUN 之前，保持上述双写和日志机制，直到数据复制完成。

LUN 信息交换

数据复制完成后，源 LUN 和目标 LUN 进行信息交换。交换过程中，LUN ID、WWN 等信息保持不变，交换源 LUN 和目标 LUN 的数据卷 ID，使得源 LUN ID 和目标卷 ID 形成新的映射关系。在交换完成后，主机仍然通过源 LUN ID 识别到源 LUN，但实际访问到的物理空间已修改为目标 LUN 对应的数据卷。

SmartMigration 可以满足以下场景的需求：

- 结合 SmartVirtualization 特性实现存储系统升级换代。现有老旧设备上数据迁移到华为新阵列上，提升业务的性能和数据的可靠性。
- 由于容量、性能、可靠性调整等原因所需要的数据迁移。比如，把一个 LUN 从一个存储池迁移到另一个存储池。

迁移目标 LUN 支持增值配置

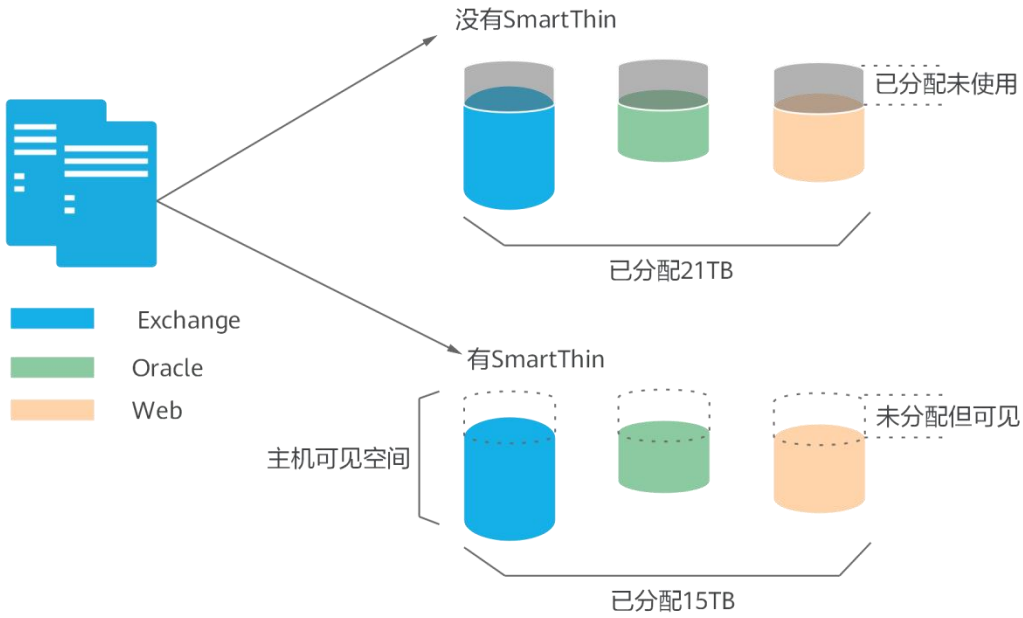
跨设备的数据迁移，数据迁移持续时间和具体的数据量强相关，从几小时到几周甚至几个月不等。支持目标 LUN 配置跨阵列的远程复制和双活特性，可以实现迁移期间灾备数据能够跨数据中心的迁移，确保迁移完成后能够尽快完成数据保护，提升客户的解决方案可靠性。

备注：由于数据迁移期间，目标 LUN 的数据为非完整数据，因此迁移完成前和灾备数据同步完成前，都不支持灾备侧接管业务。

5.2.4 智能精简配置（SmartThin）

智能精简配置以一种按需分配的方式来管理存储设备。智能精简配置不会预先分配所有的空间，而是将大于物理存储空间的容量形态呈现给用户，使用户看到的存储空间远远大于系统实际分配的空间。用户对这部分空间的使用实行按需分配的原则。如果用户的存储空间不足，可通过扩充存储资源池的方式来进行系统扩容，整个扩容过程无需业务系统停机，对用户完全透明。FastCube 混合闪存只支持 Thin LUN 模式，下图以 LUN 为例展示了 SmartThin 提供的空间价值。

图 5-21 thin LUN 与传统 LUN 对比



5.2.5 数据销毁（SmartErase）

硬盘不再按原场景使用时，盘上数据就不再需要了；如果不进行任何处理，可能出现非授权用户利用残留数据恢复原始数据，存在信息泄露风险。为避免出现信息泄露问题，需要对盘数据进行有效擦除，确保数据安全。FastCube 2910 计算型存储数据销毁功能基于盘级实现，不针对具体的服务业务类型，SAN 和 NAS 的应用都适用。

加密硬盘数据销毁

- 支持在删除加密硬盘域时对整个硬盘域的 SED 进行数据销毁，也支持针对单个 SED 进行数据销毁。
- 技术原理：通过更改 SED 的 DEK，使盘上的旧数据无法解密，从而达到秒级数据销毁的目的，无需通过反复耗时的擦写硬盘来销毁数据。

全盘数据销毁

- 擦除的数据无法恢复，有效保障信息安全。
- 支持三种数据擦除机制：
  - block_erase：块级别数据擦除，同时擦除用户数据及映射关系。
  - cryptographic_erase：对于加密硬盘，通过擦除安全密钥方式擦除用户数据及映射关系。
  - overwrite：通过特定的十六进制数重复写入覆盖用户数据，达到擦除数据的目的。当前支持的 overwrite 标准包括“DoD 5220.22-M(E)”、“DoD 5220.22-M(ECE)”、“VSITR”和“Custom”。

    DoD 5220.22-M(E)：DoD 5220.22-M 标准。依次写入 0x55、0xAA、伪随机数。

    DoD 5220.22-M(ECE)：DoD 5220.22-M (ECE)标准。依次写入 0x55、0xAA、伪随机数、伪随机数、0x55、0xAA、伪随机数。

    VSITR：VSITR 标准。依次写入 0x00、0xFF、0x00、0xFF、0x00、0xFF、伪随机数。

    Custom：用户自定义标准，可自定义设置写入的十六进制数以及写入次数。
- 支持校验数据擦除。
- 支持导出硬盘数据销毁报告。
- 单盘数据擦除。进行单盘数据擦除时，支持以下两种擦除方式：
  - 保留硬盘认证信息擦除
  - 全盘擦除（不保留硬盘认证信息）

5.2.6 配额(SmartQuota)

SmartQuota，即文件系统配额技术，下文称配额。配额的主要作用是方便系统管理员管控资源使用者（包括目录、用户、用户组）的存储资源，以限制指定使用者可使用的磁盘空间，从而避免出现某些用户过度占用资源的问题。

在 SmartQuota 特性中，DTree（目录配额树）是一个重要的概念。DTree 是文件系统内的特殊目录，可以被创建在文件系统的任意一级，DTree 管理其目录下所有的子目录和文件的配额（包括递归的普通子目录和文件，DTree 目录下不能再创建 DTree）。DTree 只能通过管理终端（命令行或 GUI 管理界面）来创建、删除和修改，而不能通过客户端主机来修改。另外，作为配置配额时的载体，目录配额、用户配额、以及组配额都只能在 DTree 上配置。文件系统的根目录也是 DTree，因此也能设置相关配额。

总结一下 DTree 与普通目录存在的差别：

（1） DTree 只能由管理员通过命令行或 GUI 管理界面创建、删除、重命名等操作，DTree 可以创建在文件系统的任意一级。

（2） DTree 可以通过协议进行共享，共享时不允许被改名和删除。

（3） 不允许跨 DTree 进行 NFS 协议下的移动文件或 SMB 协议下的剪切文件操作，即不允许在两个不同 DTree 目录之间进行文件的 MV 操作（NFS 协议）或剪切操作（SMB 协议）。因为一个文件或目录同时只能属于同一个 DTree。

（4） 不允许跨 DTree 的硬链接，即不允许在两个不同的 DTree 之间进行硬链接操作。

SmartQuota 支持如下配额类型：

- 容量软配额（space soft quota）：用于空间容量告警的配置值，当配额对象已用空间超过所设置的容量软配额时，向系统告警提示空间资源紧张，提醒用户删除不用的文件或扩大配额，此时用户仍然可以继续写入数据。
- 容量硬配额（space hard quota）：配额对象上用于限制最大可用容量的配置值。当配额对象已用空间到达所设置的硬配额时，向用户返回空间不足的错误。
- 文件软配额（file soft quota）：配额对象上用于文件数告警的配置值，当配额对象已用文件数超过所设置的文件软配额时，向系统告警提示文件资源紧张，提醒用户删除不用的文件或扩大配额，此时用户可以继续创建文件或目录。
- 文件硬配额（file hard quota）：配额对象上用于限制最大可用文件数的配置值。与容量硬配额一样，当配额对象的已用文件数到达所设置的硬配额时，向用户返回空间不足的错误。

SmartQuota 支持如下设置配额的对象：

- DTree 配额：限制 DTree 下的所有的文件和目录的总体配额，包括容量和文件数量。
- 用户配额：限制某个用户创建的文件和目录的总体配额，包括容量和文件数量。
- 用户组配额：限制某个用户组创建的文件和目录的总体配额，包括容量和文件数量。

5.2.7 智能加速（SmartAcceleration）

在机械盘为主的时代，为解决机械盘性能不足的问题存储业界诞生了 Cache 和 Tier 技术，可以实现一定程度的性能加速。但传统的 Cache 和 Tier 受限于其技术存在一定的使用限制，例如单独的 Cache 很难解决数据生命周期冷热的问题，而 Tier 又往往需要基于人工经验进行策略配置和周期配置，易用性和性能加速效果都面临不小的挑战。

到了全闪存时代，面向机械盘设计的传统存储软件栈过于厚重，需要通过架构创新打破系统瓶颈，充分发挥闪存介质的效能。为此，业界诞生了原生全闪存技术，通过顺序追加写机制（ROW）结合更加灵活的元数据索引设计和盘控配合技术，实现了全闪存系统在数据缩减、极致性能、稳定时延多个维度的全面突破。

而面向混闪场景，如何优化整体性能，由 SSD 提供性能加速，由 HDD 提供容量，使得混闪存储系统也能获得和全闪接近的性能体验是其最重要的设计目标。因此，FastCube 全混合闪存结合了机械盘时代和全闪时代的优势技术，通过创新的 SmartAcceleration 特性，进一步提升存储系统的性能和效率。SmartAcceleration 其核心是动态自适应数据布局架构（DADL），以全闪时代的 ROW 大块顺序写机制为底座，以全新的全局冷热感知算法为引擎，采用 Cache 和 Tier 弹性融合的统一性能层加速技术，实现全场景最优的数据流动和布局，从而打破传统机械盘在随机小 IO 场景下的性能瓶颈，最大限度的发挥混闪系统的效能。

5.2.7.1 SmartAcceleration 基本原理

SmartAcceleration 基于统一的性能层配合全局冷热感知算法实现全场景性能加速，性能层融合了 Cache 和 Tier 两种加速技术，并通过配额分配的方式将性能层资源灵活地分配到各个存储池。

SmartAcceleration 默认将 SSD 硬盘资源创建为一个统一的性能层，在性能层基础之上可以创建混闪存储池或者全闪存储池。创建混闪存储池时先选择所需要的 HDD 硬盘，根据选出的 HDD 硬盘数量和容量，系统会自动推荐分配到每个混闪存储池的性能层容量配额（实际就是 SSD 空间）。创建全闪存储池时则无需选择 HDD 盘，直接输入全闪存储池容量需求即可。

相对传统的 Tier 或 Cache 配置，统一性能层具备如下特点：

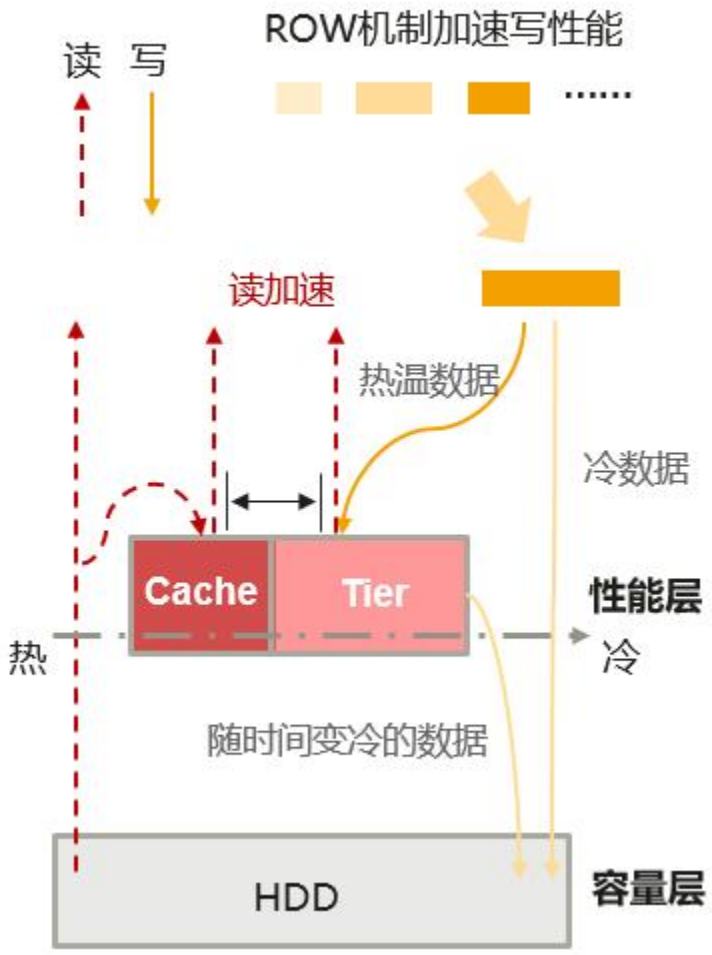
- 性能层支持多 POOL 共享，不再要求物理独占，利用率和灵活性更高
- 性能层由系统默认按推荐配额分配到各个 POOL 做智能加速，也可灵活伸缩（增/减）
- Tier/Cache 在性能层实现统一融合，系统内部自动配置，以达到整体效率最佳

当新的数据写入时，根据智能算法进行分流，热数据或温数据写入性能层，冷数据写入容量层。性能层基于当前的水位情况和数据的冷热变化，选择适当的时机将相对变冷的数据淘汰或迁移到容量层。

当数据读取时，先在内存读缓存中查找，不命中则到性能层查找，如果不命中则到容量层进行查找。容量层的数据根据智能算法推荐，当数据再次变热时会重新流动到性能层，以提升整体访问的命中率。

相对传统算法，全新算法采用了多粒度结构树的热点统计，对于混合 IO 模型有更精准更高效的冷热感知。同时除了采用传统的历史统计方法外，还引入了机器学习的时间序列预测算法进行综合判断，对冷热变化的场景适应度更强，命中率更高。

图 5-22 SmartAcceleration 基本原理



由于 SmartAcceleration 内置了强大的全局冷热算法和数据流动算法，并且将 Cache 和 Tier 的技术联合统一，可实现全场景、全周期、全范围的高效数据加速，提升混闪系统的性能体验。

5.2.7.2 SmartAcceleration 应用场景

SmartAcceleration 针对有冷热访问特征的 IOPS 或 OPS 密集型的 NAS 和 SAN 应用都有较好的加速效果，具体加速后的性能水平取决于真实场景的热点区域的大小和访问强度。例如，SAN 应用的中低性能要求的生产测试场景，NAS 应用的 OA 办公共享、票据影像、医疗 PACS 等混合型加速场景。

如果应用有极致的性能要求，全时段都期望极高的 IOPS 和极低的访问时延，则推荐使用全闪层配置。尽管 SmartAcceleration 可以提供一定的命中率获得近似闪存的效果，但受限于应用自身的访问特征，命中率通常无法做到 100%，无法替代全闪存的全时段极致性能体验。

此外，SmartAcceleration 针对纯大 IO 访问的带宽场景或无冷热特征的场景，加速效果并不明显，因此需针对具体场景进行综合推荐。

5.2.8 多租户（SmartMulti-tenant）

随着单阵列的业务能力提升，单设备上会承载越来越多的客户业务系统，多业务混合应用上客户就期望能够实现一定程度的隔离。FastCube 混合闪存的多租户特性就承载着业务隔离的目标，提供 NAS 业务和 SAN 业务的配置隔离。

多租户特性主要解决租户之间逻辑资源的隔离问题，包括业务隔离、网络隔离。用户不能跨越租户进行数据访问，以此来达到安全隔离的效果。

- **业务隔离**：每个租户都有自己的存储服务、用户访问鉴权，用户能通过租户的 LIF 或 FC 口来访问所属业务服务。
- **网络隔离**：对于 NAS 业务租户的网络由 VLAN 和 LIF 隔离，以防止非法主机访问租户的存储资源。对于 SAN 业务，由于 FC 协议是点对点通信，用户可以通过指定租户使用的 FC 端口实现网络隔离。

业务隔离

多租户特性以租户为单位进行业务隔离，支持 NAS 业务、SAN 业务的租户隔离。对于 NAS 业务，租户之间的 FS 和 NAS 用户管理相互隔离，独立配置和管理。NAS 租户的资源对象包括：FS/Dtree、NAS 共享协议配置、NAS 用户鉴权（本地用户和域用户配置）、租户级特性配置（比如：审计日志、Quota）。对于 SAN 业务，不同租户间无法访问对方的 LUN，从而确保数据均是按租户隔离的。

说明

租户隔离只涉及 SAN 和 NAS 的服务资源的隔离，不涉及存储池的隔离。存储池隔离采用多 Pool 技术实现，即客户可以在设备上规划配置多个 pool 的方式来实现空间隔离。

网络隔离

对于 NAS 业务，租户的网络资源采用 LIF（Logical Interface）逻辑端口进行管理，实现端口虚拟化的管理和隔离，实现资源的灵活安全应用。

对于 SAN 业务，用户可以通过指定租户使用的 FC 端口实现网络隔离。。

5.3 增值特性：Hyper 系列

为满足客户本地保护以及远程容灾的需求，FastCube 2910 计算型存储提供丰富的 Hyper 系列软件。通过 HyperSnap 及 HyperCDP，实现本地逻辑错误的恢复；通过 HyperClone，实现了本地完整数据副本，父对象的数据完整性不影响克隆对象的数据完整性能力，保证了故障域的隔离。通过 HyperReplication 特性，实现了远程容灾保护；而 HyperMetro for SAN 既保障了业务连续性，又提供了容灾能力。

5.3.1 快照（HyperSnap）

FastCube 2910 计算型存储的快照特性叫 HyperSnap。由于 SAN 和 NAS 的业务场景有些区别，快照功能特性上也有些区分；SAN 的快照是可读可写，在主机视图是一个独立的 LUN 对象需要单独添加映射。NAS 快照则是只读快照，部署在 FS 的快照目录下，通过文件系统的共享挂载直接可以访问。

5.3.1.1 SAN 快照（HyperSnap for SAN）

本章主要介绍 SAN 的 HyperSnap 技术原理以及关键功能。

5.3.1.1.1 快照基本原理

FastCube 2910 计算型存储的 SAN 的快照是可读可写的快照，主机访问需要单独将快照添加到主机的映射中。本章主要介绍 HyperSnap 的关键技术—TP（Time Point）。

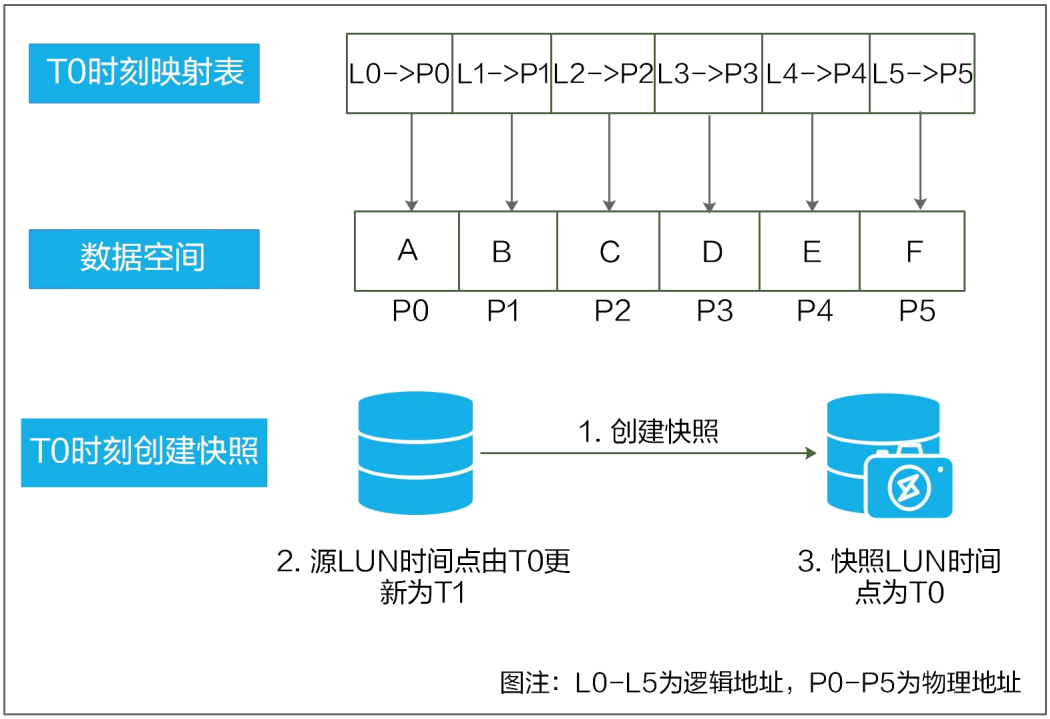
TP 技术

多时间点 TP（Time Point）技术是 FastCube 2910 计算型存储实现数据保护类特性的基础技术，所有的本地数据和远程数据保护都使用到该技术能力获取数据副本和实现一致性保护。

LUN TP 即为 LUN 的数据增加时间点属性。表现为：当为 LUN 创建快照时，源 LUN 的时间点属性标识字段数字递增，快照 LUN 的时间点属性仍为创建该快照时源 LUN 的时间点属性。

下面以图 5-23 为例进行说明。假设源 LUN 的当前时间点为 T0，此时用户对源 LUN 创建快照。

图 5-23 快照原理

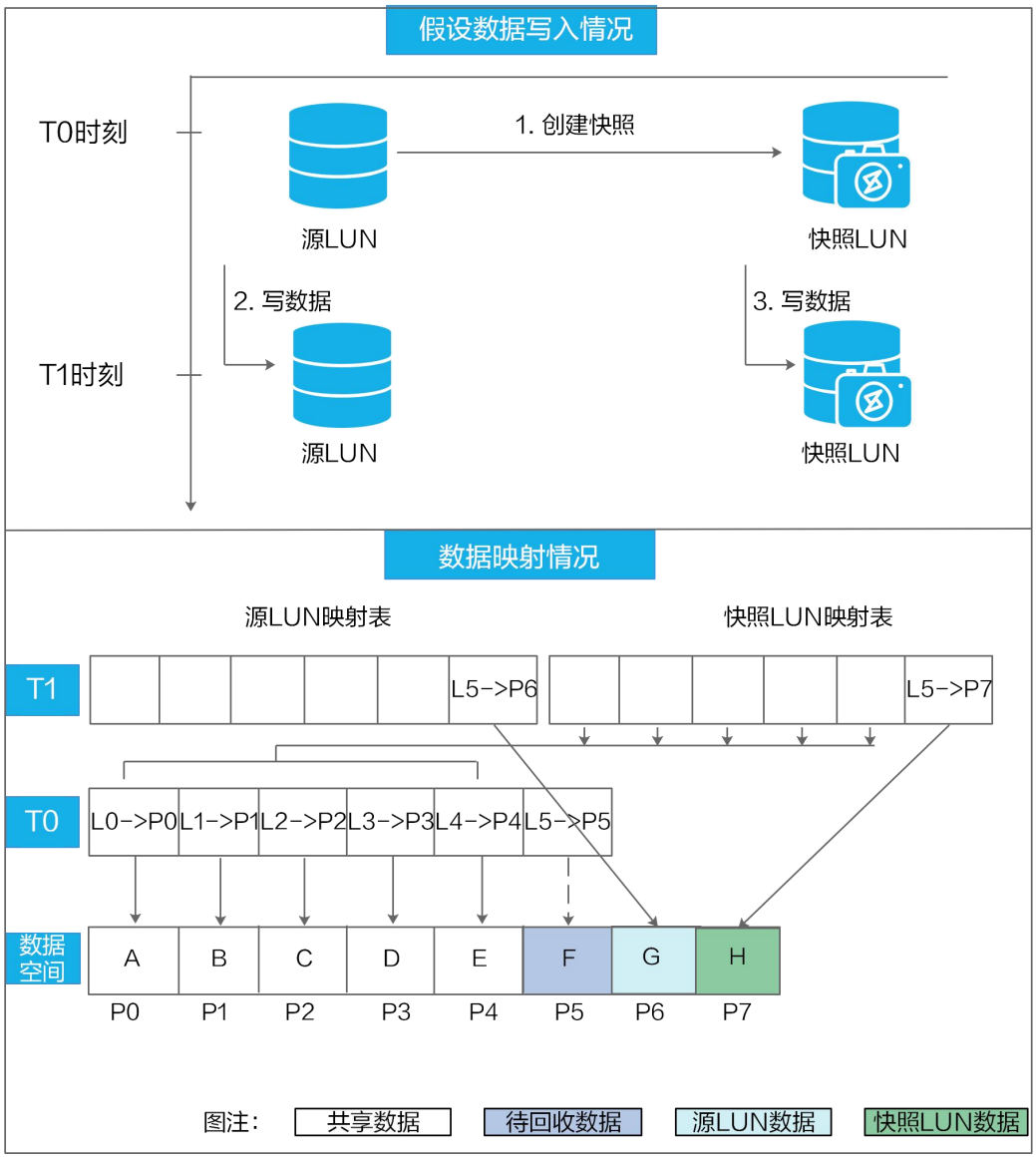


如上图：由于时间点 T0 创建了快照，导致源 LUN 的时间点属性由 T0 变更为 TP1。快照的时间点属性为创建的时间点 T0。因此对源 LUN 进行读取时，将读到下一个时间点 T1 的数据，而对快照进行读取时 ，读到的是 T0 时刻的数据 ABCDEF。

快照读写

当主机进行读写时，读写 I/O 使用最新时间点读写源 LUN 数据，使用更新前的时间点读写快照与源 LUN 的共享数据，如图 5-24 所示。

图 5-24 快照读写



- 读源 LUN：

  当源 LUN 新建快照后，源 LUN 的最新时间点由 T0 变更为 T1。读取源 LUN 时，将读取源 LUN [T0, T1]时间点范围内的数据，此时根据映射表中的映射表项按时间点从新到老读取数据。整个过程无新增的性能开销。

- 读快照：

快照的最新时间点为 T0。读取快照的最新时间点数据时，如果快照的映射表项非空，直接返回该时间点数据。否则，触发时间点重定向，将读取源 LUN 时间点 T0 的数据。

- 写源 LUN：

  源 LUN 新写数据，数据携带源 LUN 的最新时间点 T1 写入系统。将新写数据的逻辑地址以及时间点 T1 做为 Key，Key 对应的值为新数据存放在 SSD 存储池中的地址。

- 写快照：

  快照新写数据，携带快照的最新时间点 T0 写入系统。将新写数据的逻辑地址以及时间点 T0 做为 Key，Key 对应的值为新数据存放在 SSD 存储池中的地址。

由于的读写源 LUN 或者读写快照的过程中，I/O 均携带了对应的时间点信息，可以快速的定位到元数据信息，因此对性能开销较小。

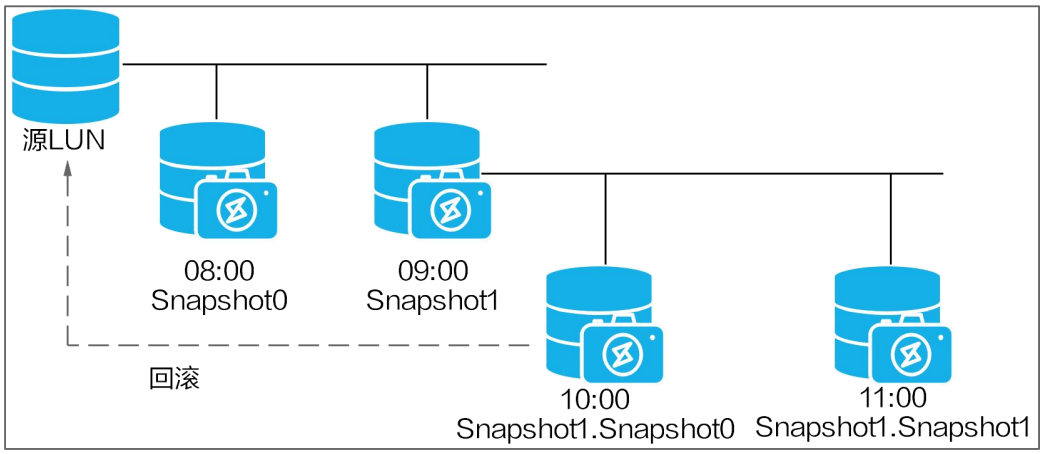
5.3.1.1.2 级联快照

如果需要对可写快照再进行保护，可以使用级联快照。本章主要介绍级联快照。

级联快照是指对快照再创建快照。FastCube 2910 计算型存储的 HyperSnap 支持最多 8 层级联快照。

级联快照可以跨级回滚，跨级回滚是指相同源 LUN 的快照之间可以进行回滚，且没有层级约束。如图 5-25 所示, Snapshot1 为源 LUN 在 9 点时刻的快照，Snapshot1.Snapshot0 为 Snapshot1 在 10:00 创建的快照。系统支持将源 LUN 直接回滚至 Snapshot1.Snapshot0，也支持源 LUN 回滚至 Snapshot1。同时，还支持 Snapshot1 回滚至 Snapshot1.Snapshot0。

图 5-25 级联快照和跨级回滚原理图



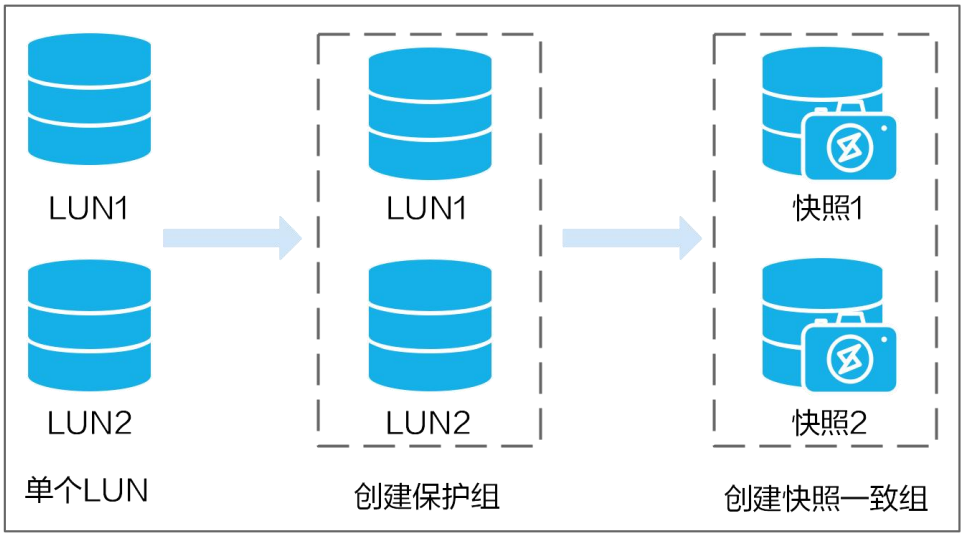
5.3.1.1.3 快照一致性组

HyperSnap 支持快照一致性组功能。本章主要介绍级联快照一致性组。

对有数据依赖关系的多个 LUN，通过创建快照一致性组的方式对快照中的多个 LUN 同时创建快照能保证多个 LUN 之间数据的一致性。比如 Oracle 数据库应用中，数据文

件、配置文件、日志文件通常会分布在不同的 LUN 中，在创建快照时，必然要对这些文件所在的 LUN 在同一时间创建快照，才能实现在数据恢复时应用数据的一致性。

图 5-26 快照一致性组原理



1. 创建 LUN 保护组，向 LUN 保护组添加 LUN。

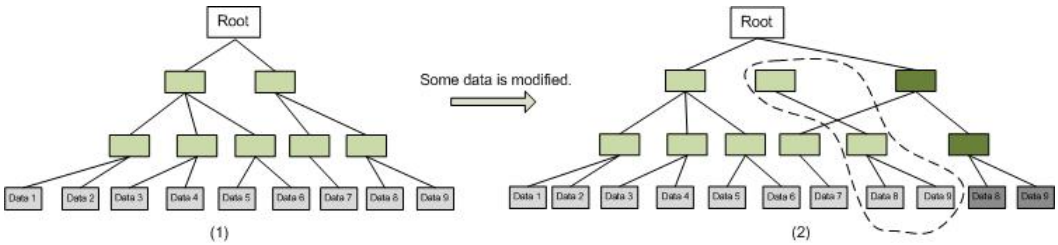
说明

用户可以向 LUN 保护组中最多添加 4096 个 LUN。

2. 对保护组创建快照一致性组（快照一致性组里快照的时间点是一致的）。

5.3.1.2 NAS 快照 (HyperSnap for NAS)

NAS 的快照基础技术基本原理也是基于 TP 技术，配合 NAS 业务对象的管理模型实现 Thin 空间和性能无损快照能力。NAS 快照为只读快照，可写快照能力由克隆特性提供。



FastCube 2910 计算型存储的文件系统快照是基于本产品的 ROW 型（Redirect On Write，写时重定向）文件系统来实现的。所谓 ROW 型文件系统，是指向文件系统新写入或者修改写入数据时，新数据不会覆盖掉原来的旧数据，而是在存储介质上新分配空间来写入数据，此种方式保证了数据的高可靠性和文件系统的高扩展性。基于 ROW 技术的文件系统快照，可实现快速创建（秒级），并且除非原始文件被删除或者更改，快照数据并不占用额外的磁盘空间。

为文件系统创建快照时，仅只需要拷贝文件系统的根节点并保存，就形成了一个文件系统的快照。整个过程中不需要拷贝任何用户数据，因此整个过程耗时极少，通常在

一两秒内完成。并且在数据被修改之前，快照的文件集与源文件系统共同使用文件系统空间，无需单独为快照分配特定的空间。

刚创建的快照不包含任何实际数据，只包含了指向源文件系统数据的入口指针，当用户访问快照数据时，实际上访问的是源文件系统中的数据。只有当源文件系统的数据发生变化后，快照才会引入其独有的数据，这部分数据由于受快照保护，故而不能直接删除，只有等快照被删除后，这部分空间才能够被释放。

随着源文件系统不断被更新，原有的数据块会逐步的变成快照的占用空间，但是新写入的数据，不会计入快照占用空间中，因为快照所映像的只是生成快照时刻的源文件系统映像。当用户需要恢复出快照点时刻的数据时，可通过快照数据的回滚快速实现，通过回滚，文件系统可将数据恢复到快照点时刻，从而避免了快照点后因为人为的错误或者病毒的入侵等引起的源文件系统损坏造成的数据丢失。

需要说明的是，快照的回滚是不可逆的，回滚只能将数据恢复到某一特定的时间点，但该时间点之后的数据包括快照将会丢失。如果仅仅是特定的几个文件被损坏、误修改、误删除，则无需进行整个文件系统的回滚，可以直接从特定时间的快照中将这些文件手动拷贝到源文件系统中即可。另外，由于快照回滚会导致文件系统数据或快照丢失，如果用户正在访问这部分丢失的数据，有业务中断的风险，用户需要谨慎使用。

NAS 安全快照

为了防止文件系统只读快照被意外或者故意删除，支持设置其安全属性把快照变成安全快照。

文件系统只读快照支持设置其安全属性。快照的安全属性如下：

- 是否为安全快照。
- 安全快照过期时间，支持过期时间为 1 天~20 年。
- 安全快照过期后是否自动删除。

用户可以通过以下方式来设置快照的安全属性：

- 创建快照时，设置该快照的安全属性。
- 快照创建完成后，修改该快照的安全属性。

当文件系统快照被设置了安全属性后，该快照在安全期内不能被删除。

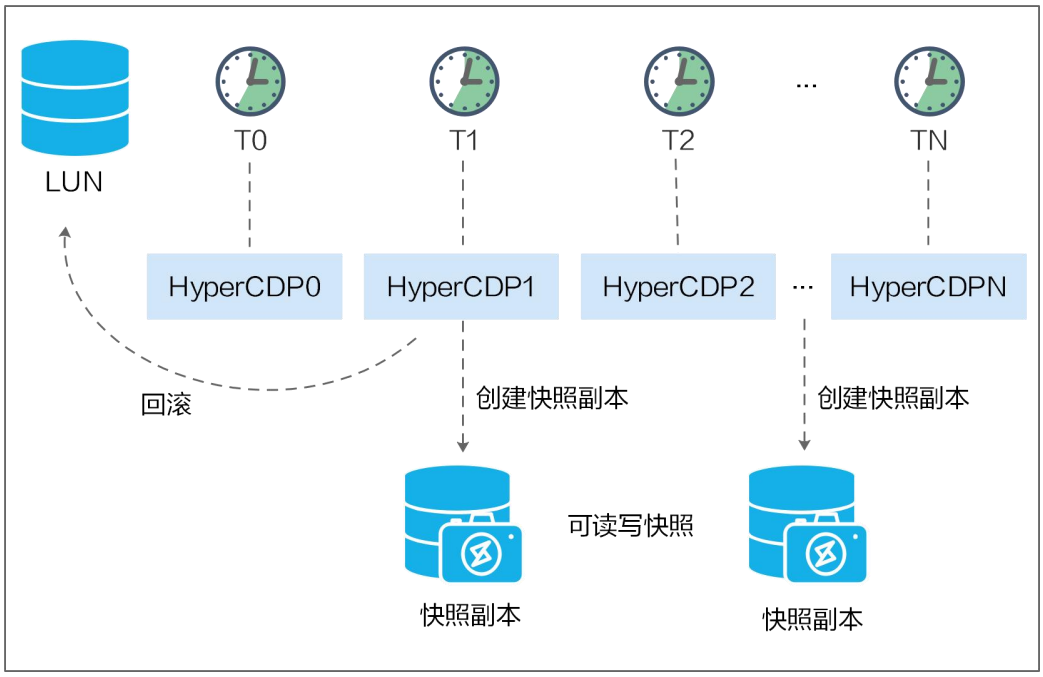
5.3.2 持续数据保护（HyperCDP）

HyperCDP 特性是一种高密快照特性，本章主要介绍 HyperCDP 的原理。

误操作、病毒入侵等都可能导致数据破坏，持续数据保护就是持续对数据创建间隔周期非常短的快照，帮助客户能够把数据恢复到期望的时刻。

FastCube 2910 计算型存储的 HyperCDP 特性可实现 LUN 和文件系统的持续保护。HyperCDP 基于无损快照技术（TP、ROW 技术），每个 HyperCDP 对应源 LUN 的一个时间点，其原理如图 5-27 所示。

图 5-27 HyperCDP 快照原理（以 LUN 为例）



定时策略

HyperCDP 特性内置定时计划功能，支持间隔固定周期、每天、每周、每月 4 种定时策略，通过组合多种定时策略满足用户远疏近密的备份诉求。

表 5-4 HyperCDP 定时策略

策略类型	策略描述
固定周期	• 按秒设置，默认 10 秒执行一次定时 HyperCDP 计划（NAS 最小 15 秒）。 • 按分钟设置，默认 1 分钟执行一次计划。 • 按小时设置，默认每个 1 个小时执行一次计划。 **说明** SAN：单 LUN 支持保留 60000 个 HyperCDP 的快照，系统最大支持 2000K 个 HyperCDP 的快照。 NAS：单文件系统支持保留 4096 个 HyperCDP 的快照，系统最大支持 128K 个 HyperCDP 的快照。
每天执行	设置每天定时执行 HyperCDP，以存储设备所在时区显示，取值范围为 00:00 到 23:59。 **说明** 保留个数为 1 到 256。

策略类型	策略描述
每周执行	设置每周定时执行 HyperCDP，以存储设备所在时区显示，取值范围为周一到周日的 00:00 到 23:59。 **说明** 保留个数为 1 到 256。
每月执行	设置每月定时执行 HyperCDP，以存储设备所在时区显示，取值范围为 1 到 31 日的 00:00 到 23:59。 **说明** 保留个数为 1 到 256。

数据保护密集、持久

单 LUN 可支持创建 6 万个 HyperCDP，最短定时间隔支持 3 秒。单 FS 的可支持创建 4096 个 HyperCDP 对象，最短定时时间间隔为 15s。CDP 的保存策略客户可以通过定时策略文件来配置。

一致性组（仅 SAN）

数据库应用中，数据文件、配置文件、日志文件通常会分布在不同的 LUN 中，通过 HyperCDP 一致性组功能，可保证一致性组中的数据时间点一致，即业务恢复时应用数据的一致性。

安全快照

在金融、证券或者银行的应用中，客户对重要的数据配置了 HyperCDP 进行数据备份，这些 HyperCDP 有长期保存的需求。为了防止 HyperCDP 被意外或者故意删除，支持设置保留期。在保留期段内不能被删除；到期后，可以选择手动或者自动删除。

- 支持对单个 LUN 或单个文件系统创建安全快照；支持对 LUN 一致性组创建安全快照一致性组。保留期可设置为 1 天-20 年，可配置到期是否自动删除。
- 支持将普通类型 HyperCDP 修改为安全快照；支持将普通类型的 HyperCDP 一致性组修改为安全快照一致性组。保留期可设置为 1 天-20 年，可配置到期是否自动删除。
- 支持修改安全快照的保留期和自动删除开关；支持修改安全快照一致性组的保留期和自动删除开关。保留期可以延长，但是不能缩短。
- 支持配置创建安全快照的定时策略。

说明

- 安全快照在保留期内，任何人都不能删除，保留期不能缩短，所以需要谨慎设置保留期。
- 当 HyperCDP 自动删除开关打开，存储池已用容量达到即将耗尽容量告警阈值，或者保护容量达到高阈值时，安全快照不会被删除。
- 安全快照时钟是系统内部维护的一个时钟，不受系统时间修改的影响。该时钟每隔 1 分钟推进一次，系统关机后该时钟就不会往前推进。

5.3.3 克隆（HyperClone）

FastCube 2910 计算型存储支持 HyperClone 功能，通过创建源 LUN/FS 和目标 LUN/FS 的 HyperClone 关系，可以为目标 LUN/FS 同步源 LUN/FS 完整的数据拷贝。目标 LUN/FS 可以是已经存的 LUN/FS，也可以在创建 HyperClone 关系时自动创建。创建 HyperClone 关系时，需要源 LUN/FS 和目标 LUN/FS 的容量相等。目标 LUN/FS 可以是空的，也可以是已有数据的 LUN/FS。如果目标 LUN/FS 已有数据，则旧数据将被删除。克隆 LUN/FS 和源 LUN/FS 的数据访问视图保持独立，即对任何一方的修改不影响另外一方的数据。

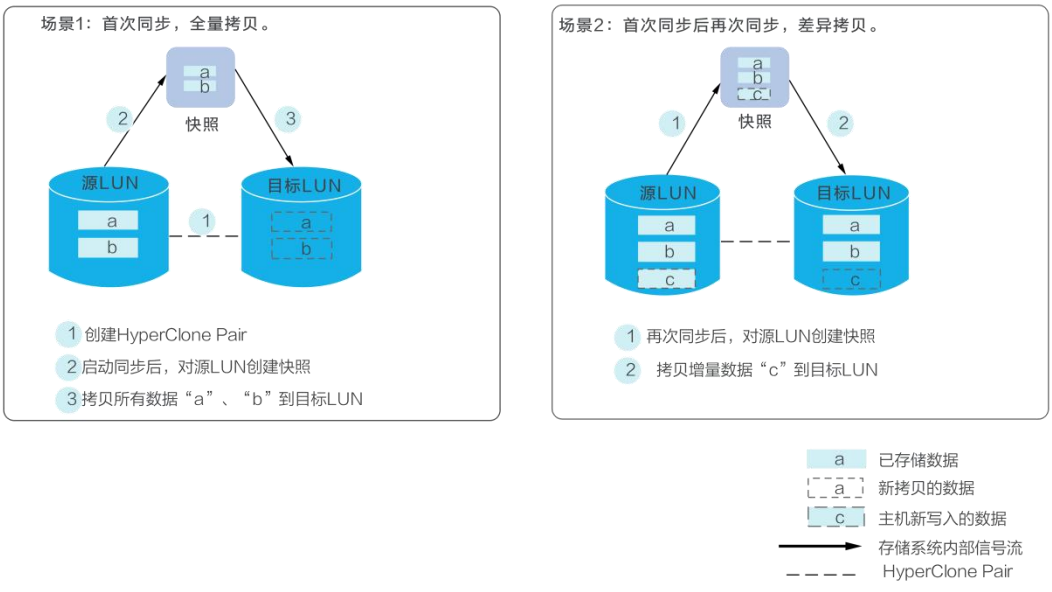
5.3.3.1 SAN 克隆（HyperClone for SAN）

LUN 克隆创建完成后，克隆 LUN 和源 LUN 复用同一份数据。数据的读写模型同快照 LUN 和源 LUN 的。对克隆 LUN 用户可以通过分裂操作启动数据后台拷贝。数据同步状态不影响目标 LUN 读写状态，无需等待后台拷贝完成，目标 LUN 可以立即读写。LUN 克隆数据同步支持增量同步和反向增量同步。可以通过 LUN 的保护组创建 HyperClone 一致性组，为一组源 LUN 的数据提供一致性保护和完整备份。

5.3.3.1.1 正向数据同步

数据同步开始时，系统将对源 LUN 生成一个即时的快照，将源 LUN 该时刻的快照数据全量同步到目标 LUN，并对后续的写操作都记录到差异日志中。后续用户再执行数据同步时，通过对比目标 LUN 和源 LUN 的差异数据增量同步到目标 LUN，目标 LUN 两次同步间修改的数据将被覆盖。用户可以在数据同步操作前通过对 HyperClone 关系中的目标 LUN 创建快照的方式，保留对目标 LUN 数据的修改。

图 5-28 正向同步

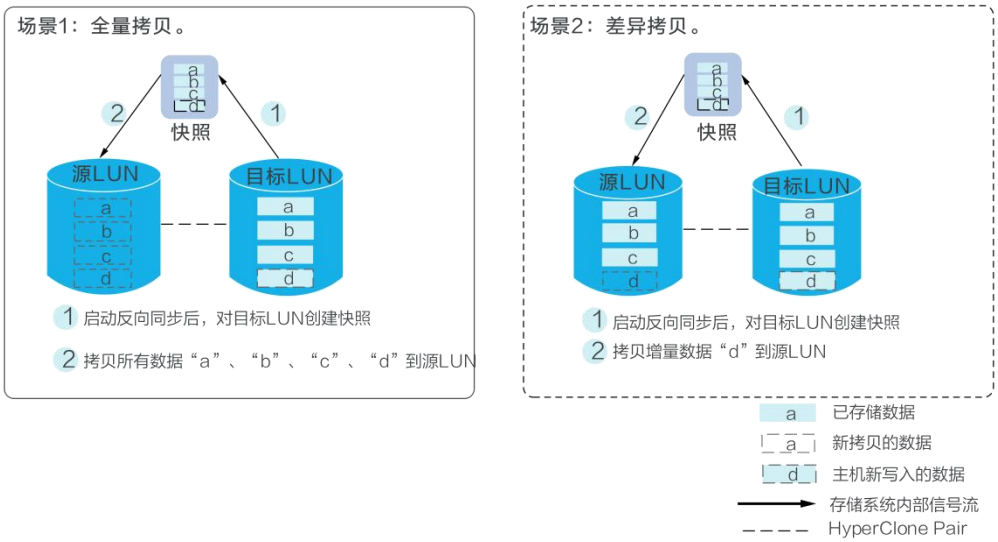


5.3.3.1.2 反向数据同步

当源 LUN 损坏时，可以通过把目标 LUN 数据反向同步到源 LUN 实现对源 LUN 的保护。反向同步支持全量同步和增量同步两种。反向数据同步启动时系统对目标 LUN 生

成快照，将目标 LUN 该时刻的快照数据全量同步到源 LUN；对于增量同步，通过对比目标 LUN 和源 LUN 的差异数据，进行增量数据同步。

图 5-29 反向同步



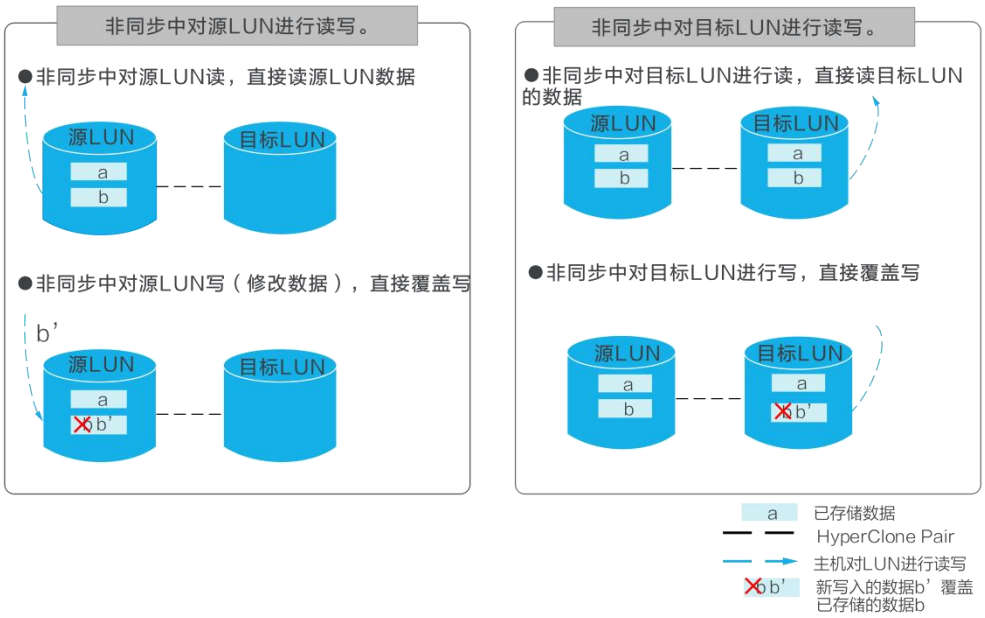
5.3.3.1.3 Clone LUN 即时可用

HyperClone 数据同步状态分为同步中、同步暂停、未同步和正常四个状态。不同的状态对源 LUN 和目标 LUN 的读写 I/O 处理不同，每一种状态都支持克隆 LUN 立即可用。

1. 未同步及正常状态时的读写原理：

   对源 LUN 或目标 LUN 的读写，直接读写源 LUN 或目标 LUN。

图 5-30 未同步及正常状态时的读写原理



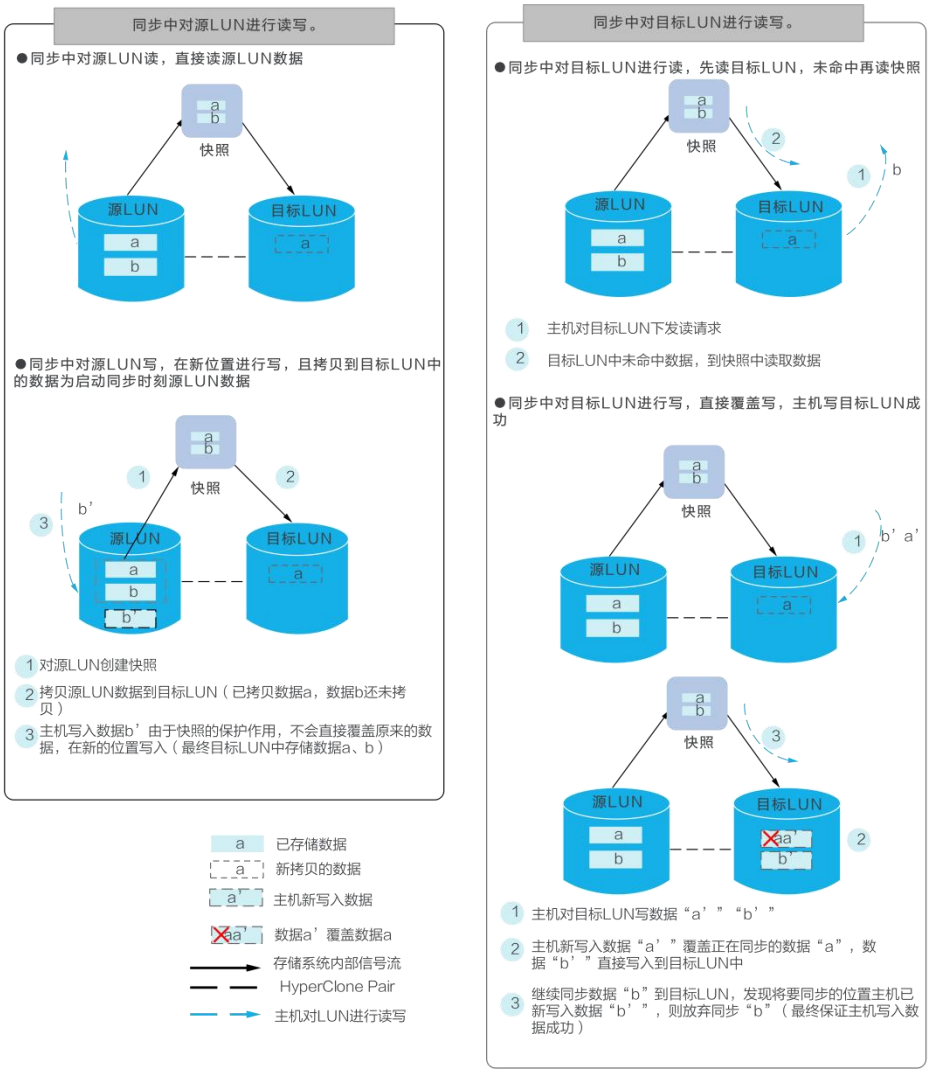
2. 同步中或者同步暂停状态的读写原理如下：

   对源 LUN 的读写，直接读写源 LUN。

   对目标 LUN 的读操作，如果读数据在目标 LUN 命中（数据已同步）则直接读取；如果读数据在目标 LUN 未命中（数据尚未同步），则到源 LUN 的快照中读取。

   对目标 LUN 的写操作，如果数据已经同步，则进行覆盖写；未同步数据进行新写，待数据同步到目标 LUN 时，如果发现该地址主机已经写入数据则放弃同步。这样保证了目标 LUN 在同步未完成时也可以读写。

图 5-31 同步中或者同步暂停状态的读写原理



5.3.3.1.4 HyperClone 一致性组

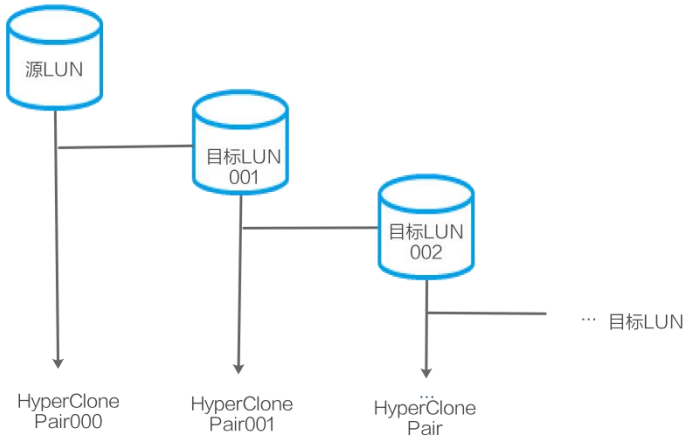
HyperClone 支持对 LUN 保护组创建 HyperClone 一致性组。用户创建 LUN 保护组后，可以通过自动模式或者手动模式为 LUN 保护组中的成员选择目标 LUN，并最终将多个 HyperClone 关系对添加到 HyperClone 一致性组中。HyperClone 一致性组可以进行数据同步、反向同步等操作，在进行这些操作时，一致性组的成员 LUN 数据始终保持在一个一致性点上，从而保证数据的完整性和可用性。

HyperClone 一致性组最多支持 4096 个成员。

5.3.3.1.5 级联 HyperClone

当 HyperClone 的目标 LUN 完成数据同步后，系统支持对 HyperClone 的目标 LUN 再次创建 HyperClone。如图 5-32 所示：

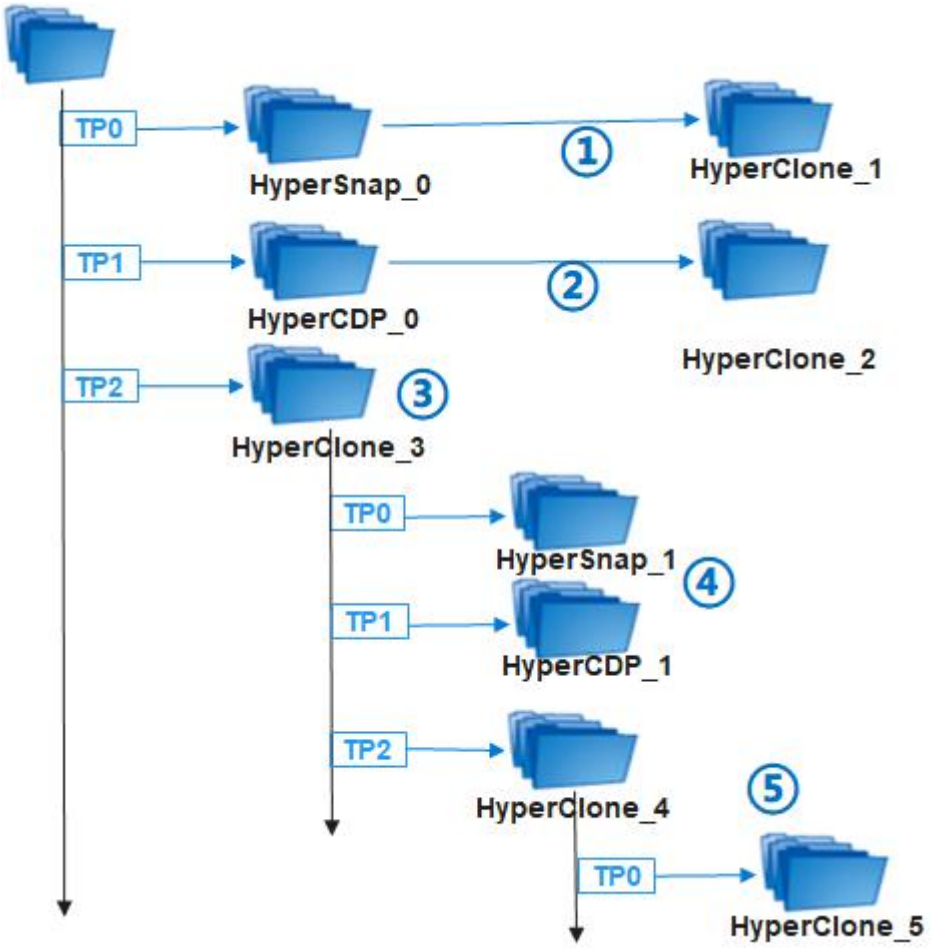
图 5-32 级联 Clone



HyperClone 特性无级联深度限制。

5.3.3.2 NAS 克隆（HyperClone for NAS）

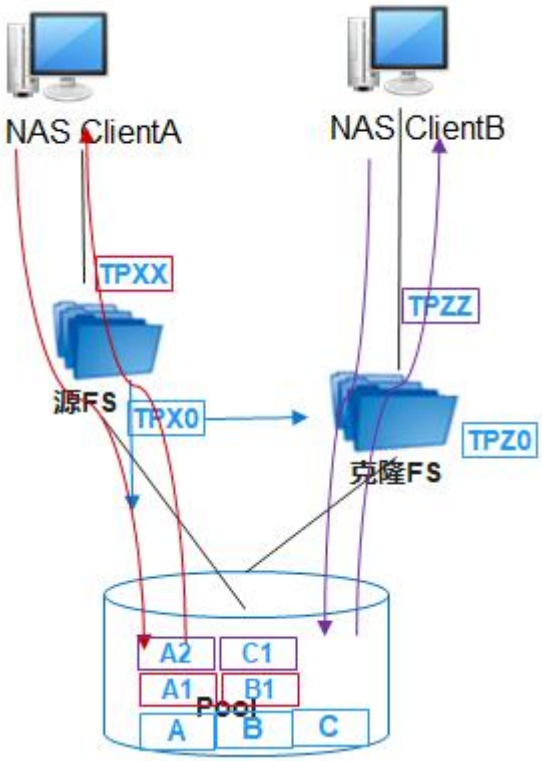
NAS 克隆功能，包括文件系统级克隆和文件级克隆，FastCube 混合闪存当前仅支持文件系统级克隆，暂不支持文件级克隆。文件系统级克隆指用户可以通过制定的源 FS 或 FS 快照来创建克隆文件系统，创建出的克隆文件系统的数据和创建时刻父文件系统的数据保持一致，且支持立即可读写。克隆功能和快照、CDP 功能的组合使用，灵活实现本地数据保护和数据管理。



1. 基于用户快照创建克隆
2. 基于 CDP 的快照创建克隆
3. 基于源 FS 直接创建克隆
4. 克隆 FS 创建快照/CDP
5. 级联克隆（最大 8 级，该功能依赖克隆分裂，分裂后才能支持。）

FS 克隆的读写原理

克隆 FS 是一个独立的 FS，和父 FS 共享已有数据和 pool 池。克隆时从父 FS 继承其子对象及其子对象的属性配置；不继承父对象的用户管理配置：父文件系统的增值配置、协议共享配置。因此克隆 FS 读写之前需要配置协议共享并在客户端指定克隆 FS 挂载后访问。



克隆 FS 的对象依赖:

源 FS 在 TPX0 时刻创建克隆 FS，该时刻的文件系统的数据在克隆 FS 上以时间点 TPZ0 对应；同时源 FS 时间点增加到 TPXX，克隆 FS 的时间点增加到 TPZZ。所有的数据都存放在 pool 中(ABC)。

源 FS 的访问:

- 写：新写和修改数据以 TPXX 时间点写入 pool 保存(A1/B1)；
- 读：请求以 TPXX 时间点到 pool 命中（A1/B1），没有则取和 TPXX 最近时间点的数据，即读的数据视图为(A1/B1/C)。

克隆 FS 的访问：

- 写：新写和修改数据以 TPZZ 时间点写入到 pool 保存(A2/C1)；
- 读：请求以 TPZZ 时间点到 pool 命中，如果不命中则取 TPZZ 到 TPZ0 之间靠 TPZZ 最近时间点的数据(A2/C1)，如果没有则表示克隆 FS 没有修改过这部分数据，系统会取 TPZ0 对应的源 FS 在 TPX0 及其之前时间点数据(B)。即都的数据视图为(A2/B/C)

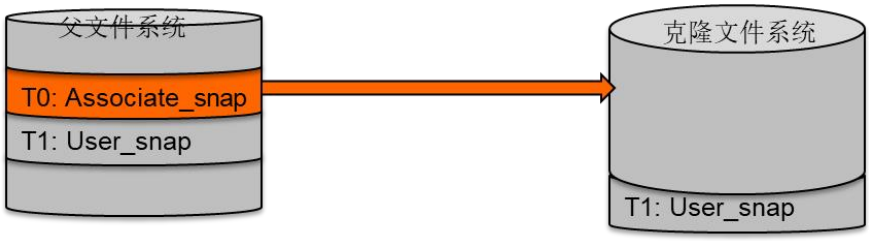
克隆文件系统分裂

克隆文件系统基于父文件系统的快照创建完成后，就继承了对应快照（该快照称为克隆文件系统的关联快照）的全量数据，克隆文件系统可以访问继承于父文件系统的数据而无需占用额外空间保存对应数据。当克隆文件系统需要独立于父文件系统而存在时，可以选择分裂克隆文件系统，克隆文件系统分裂完成后，则退化为普通文件系统，不再引用原父文件系统的任何数据。

克隆文件系统分裂原理：

- 克隆分裂时，需要将关联快照时间点的数据全量拷贝写入克隆文件系统对应时间点上。但其中已被克隆文件系统修改过的数据，则无需重新拷贝。

图 5-33 克隆文件系统分裂



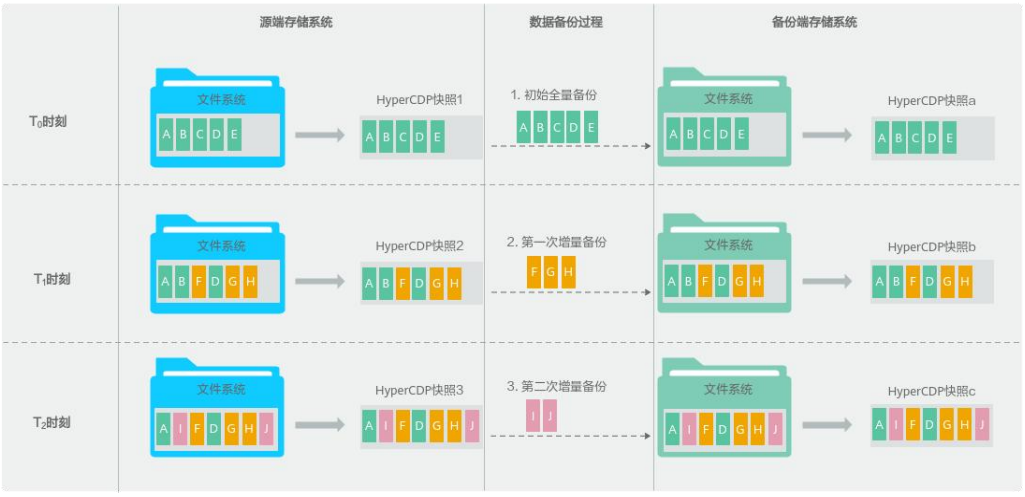
5.3.4 一体化备份（HyperVault）

采用文件系统 CDP 和远程复制快照同步功能，支持与 FastCube 混合闪存、Dorado 之间实现文件系统数据备份和恢复；用户无需单独购买额外的商业备份软件，给用户提供了一个低成本的备份方案。用户无需额外部署备份服务器和介质服务器就可以达到基本的本地备份和远程备份的功能，节省客户投资和管理的复杂性。

基本原理：

一体化备份功能以文件系统为单位提供存储系统内的备份和存储系统间的备份，备份副本基于 HyperSnap 实现，主机和应用不需要感知备份过程和副本的生成过程。在备份存储端，每一份备份副本都是一个源端文件系统备份时间点的全量业务数据，且所有备份副本相互独立，删除任何一份备份副本不会影响其它副本数据的有效性。主端 CDP 策略提供了丰富的备份策略，用户可以在本地配置较频繁的备份策略，在异地配置较稀疏的备份策略，实现将较旧的数据逐渐迁移到远端的目的，降低主端的空间和性能压力。

主端和从端的备份副本是通过 HyperSnap 实现，在主存储端文件系统生成快照，备份时将本次快照和上次同步的快照之间差异数据传输到备份存储端，备份完成后备份存储端生成快照，称之为备份副本。备份存储端的快照为主存储文件系统备份点的全量数据，因此备份存储端的快照为主存储端文件系统的远程快照。



每次备份传输的数据为主端文件系统本次时刻点数据和上次备份数据之间变化的数据块。例如初始备份时将主端文件系统的快照 1 全量备份到备份端存储中，备份完成后，在备份端存储文件系统上生成快照 a（备份副本）。

达到下一次备份时间后，主端文件系统生成快照 2，快照 2 相对于快照 1，只改变了（F、G、H 块）。备份快照 2 时，此次备份传输的数据块为（F、G、H）。备份完成后，在备份端存储文件系统生成快照 b（备份副本）。主端删除快照 1，用户主机/应用在从端访问快照 a 和快照 b 时都可以访问到主端文件系统备份点数据的全集，即备份存储文件系统的快照 a 和快照 b 都为业务数据的全量数据。

达到下一次备份时间后，主端文件系统生成快照 3，备份快照 3 相对快照 2 改变了（I、J）块。备份快照 3 时，此次备份传输的数据块为（I、J）。备份完成后，在备份端存储生产快照 c（备份副本）。主端删除快照 2，用户主机/应用在从端访问快照 a、快照 b 和快照 c 时都可以访问到主端文件系统备份点数据的全集，即备份存储文件系统的快照 a、快照 b 和快照 c 都为业务数据的全量数据。

该方案无需移动或备份源端文件系统未变化的数据，在链路上更少的传输数据和并执行更为频繁地备份，可以做到一次全量数据远端备份永久增量备份，节省物理带宽，节约从端的存储空间，提高备份效率。

一体化备份的主要功能点有：

1. 本地备份：在主存储系统内部备份文件系统的数据。
2. 异地备份：将主存储系统中的数据，备份到备份存储端中。
3. 本地恢复：将主存储系统的副本，恢复到文件系统中。
4. 异地恢复：将备份存储端系统的副本，恢复到主存存储文件系统中。

6 系统可靠性设计

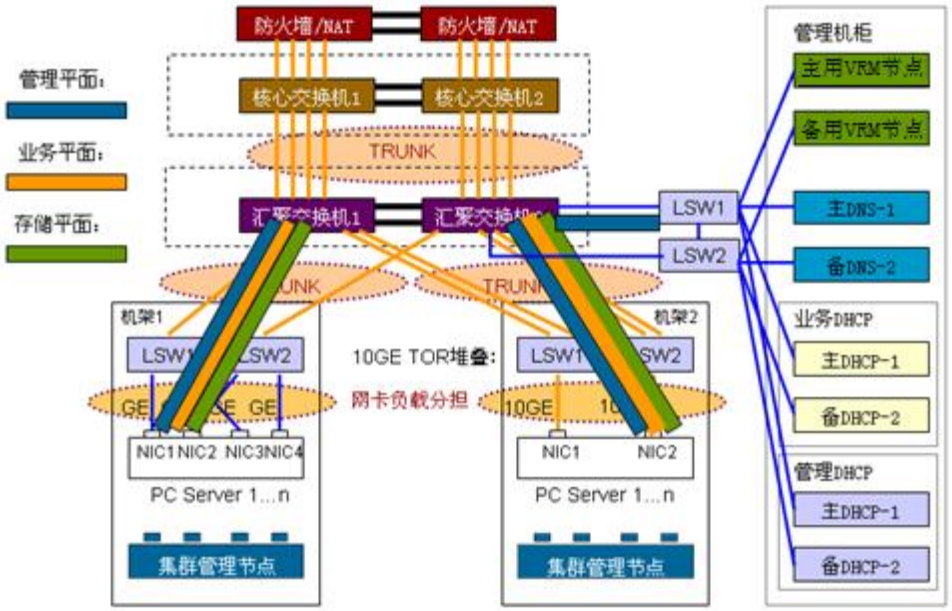
6.1 系统可靠性

6.1.1 网络分平面通信

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

下图给出了从服务器－接入层交换设备－汇聚层交换设备间的网络连接图：

图 6-1 网络分平面通信隔离示意图



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

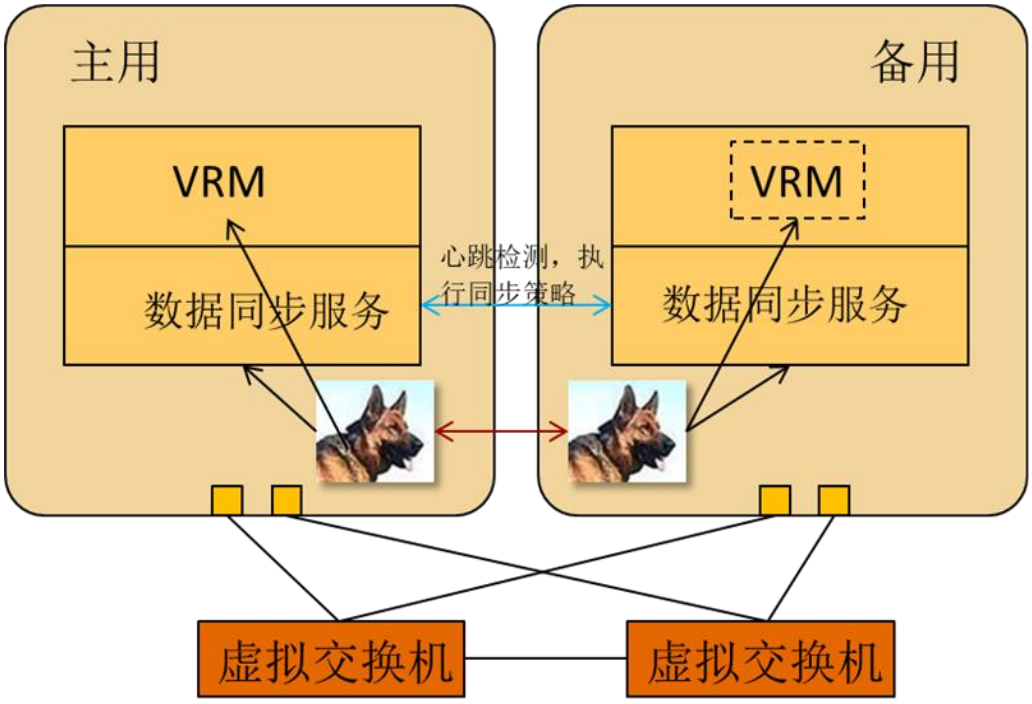
6.1.2 管理节点 HA

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

6.1.3 进程僵死保护

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

图 6-2 管理节点 HA 示意图



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

6.1.4 流量控制

为向用户提供稳定的高可用的并发业务和避免大流量冲击导致系统崩溃，管理节点针对系统关键流程设计了完善的流量控制机制。首先在 VRM 接入点采用操作流控措施，从前端抑制系统过载，保证系统的稳定性。其次是针对系统内部的瓶颈环节，增加了镜像文件下载流控，鉴权、虚拟机相关业务流控（包括虚拟机迁移，虚拟机 HA，虚拟机的创建，虚拟机的休眠和唤醒，启动和停止），O&M 流控，确保各个环节不因为流量过载导致业务失效。

6.1.5 故障检测

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

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

另外 FusionCompute 解决方案提供了健康检查工具，为技术支持工程师和维护工程师提供的一套日常检查工具，并能输出各部件健康检查报告，方便技术支持工程师和维护工程师快速了解系统的健康状况。通过检查系统当前信息和运行状态，反映系统健康或亚健康状态，在开局、巡检、升级等维护场景中使用。

6.1.6 数据一致性审计

FusionCompute 提供了数据一致性审计功能，除了系统本身针对关键资源提供的自审计和恢复能力之外，还支持定时审计 VM，卷，网络等关键资源的数据和状态的一致性，发现有异常，会自动记录或出告警，并针对记录情况提供操作指导，以便维护人员做相应的判断和恢复措施，从而保证系统内部各种相互关联数据的一致性，防止残留资源数据对系统的影响。

6.1.7 管理数据备份与恢复

系统提供管理节点配置数据和业务数据定期本地和异地备份能力，支持与第三方 FTP /FTPS Server 对接配置的能力。当管理节点服务异常无法自动修复时，通过本地备份的数据立即恢复；当由于灾难性的故障导致管理节点双点同时故障且不能通过重启等操作进行恢复，可使用异地备份数据立即恢复（1 个小时之内完成），减少故障恢复时间。

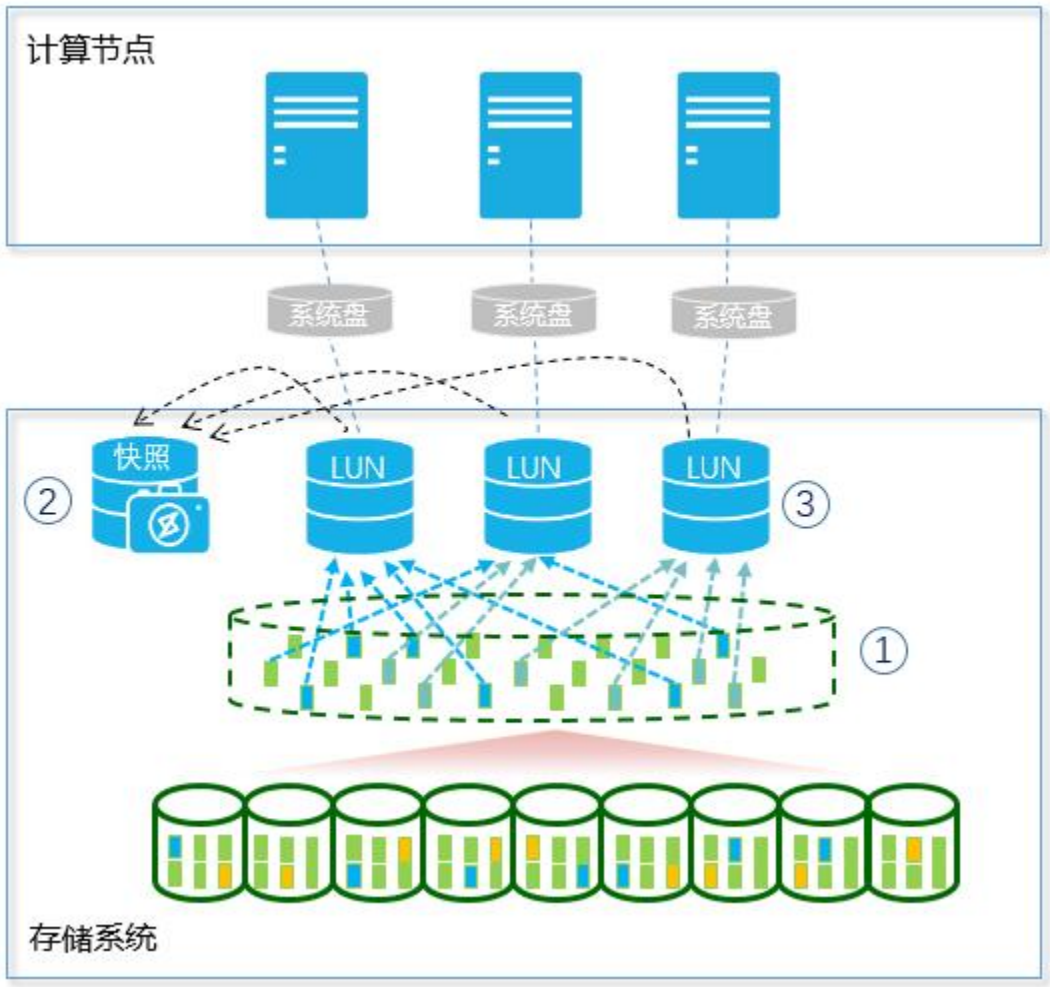
6.1.8 全局时间同步

FusionCompute 解决方案系统内部提供了时钟同步功能，可以保证所有管理节点、计算节点、虚拟机等时间一致，还支持外接 NTP 时钟源设备，可以保证全局时间统一且精准，方便系统维护以及各个网元的正常消息交互。

6.2 系统盘可靠性

FastCube 2910 计算型存储使用从存储系统划分的 LUN 作为计算节点的系统盘，通过专业存储系统的数据可靠性能力，消除了传统计算服务器上系统盘可靠性低的缺陷，并通过企业存储丰富的数据缩减和可靠性特性（如（1） RAID2.0+、（2） Hyper CDP、（3）SmartThin 和 SmartCompression 等），减少计算节点系统盘空间浪费的同时，在硬盘故障或环境异常导致计算节点操作系统异常时可快速修复系统。

图 6-3 系统盘可靠性保护

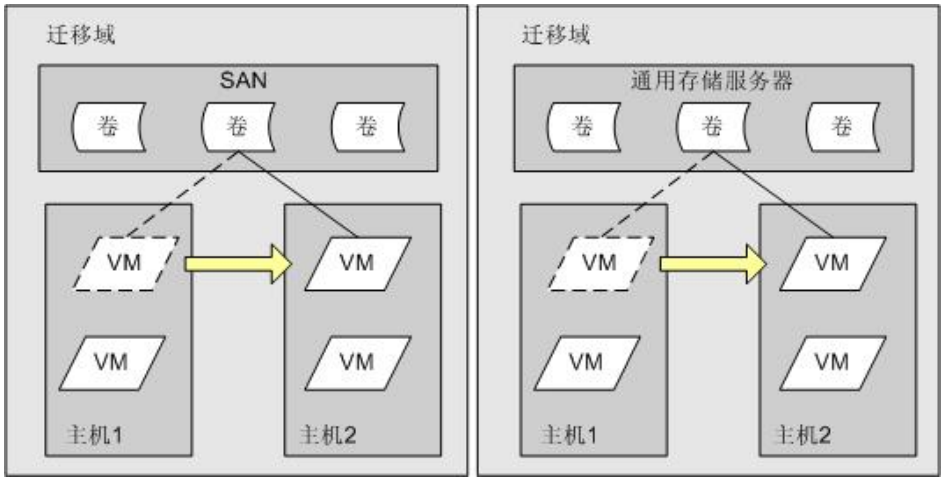


6.3 虚拟机可靠性

6.3.1 虚拟机热迁移

虚拟机是弹性计算服务的资源实体，为保证虚拟机的可用性，规避业务中断的风险，系统提供虚拟机热迁移能力，即虚拟机在不中断业务的情况下实现迁移。虚拟机迁移时，管理系统会在迁移的目的端创建该虚拟机的完整镜像，并在源端和目的端进行同步。同步的内容包括内存，寄存器状态，堆栈状态，虚拟 CPU 状态，存储以及所有虚拟硬件的动态信息。在迁移过程中，为保证内存的同步，虚拟机管理器（Hypervisor）提供了内存数据的快速复制技术，从而保证了在不中断业务的情况下将虚拟机迁移到目标主机（图示如下）。同时，通过共享存储保证了虚拟机迁移前后持久化数据不变。

图 6-4 虚拟机热迁移特性示意图



降低客户的业务运行成本：根据时间段的不同，客户的服务器会在一定时间内处于相对空闲状态，此时若将多台物理机上的业务迁移到少量或者一台物理机上运行，而将没有运行业务的物理机关闭，就可以降低客户的业务运行成本，同时达到了节能减排的作用。

保证客户系统的高可靠性：如果某台物理机运行状态出现异常，在进一步恶化之前将该物理机上运行的业务迁移到正常运行的物理机上，就可以为客户提供高可用性的系统。

硬件在线升级：当客户需要对物理机硬件进行升级时，可先将该物理机上的所有虚拟机迁移出去，之后对物理机进行升级，升级完成再将所有虚拟机迁移回来，从而实现在不中断业务运行的情况下对硬件进行升级，保证服务的持续可用性。

虚拟机热迁移典型应用场景：

- 根据需要按照迁移目的手动把虚拟机迁移到空闲的物理服务器
- 根据资源利用情况将虚拟机批量迁移到空闲的物理服务器

6.3.2 存储冷热迁移

FusionCompute 提供了虚拟机磁盘的冷迁移和热迁移，冷迁移是在虚拟机关机时候，将其磁盘文件从一个存储移动到另一个存储，热迁移可以在不中断业务的前提下，将虚拟机磁盘从一个存储迁移至另一个存储。

图 6-5 存储冷迁移原理架构

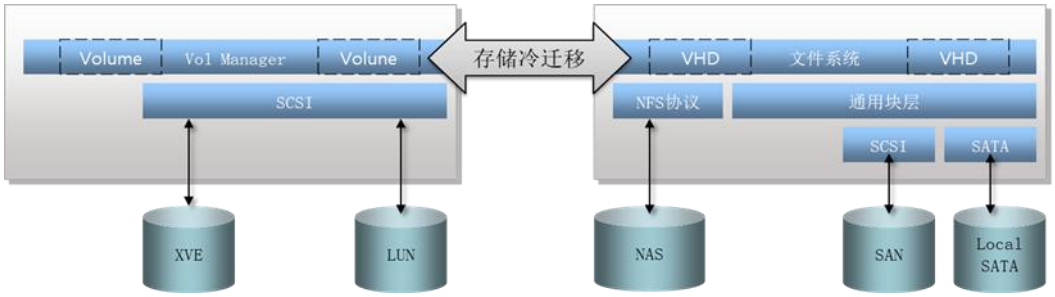
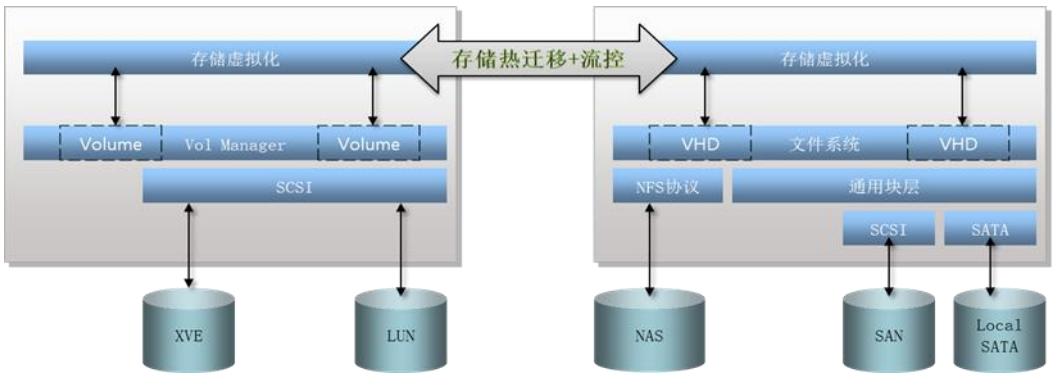


图 6-6 存储热迁移原理架构

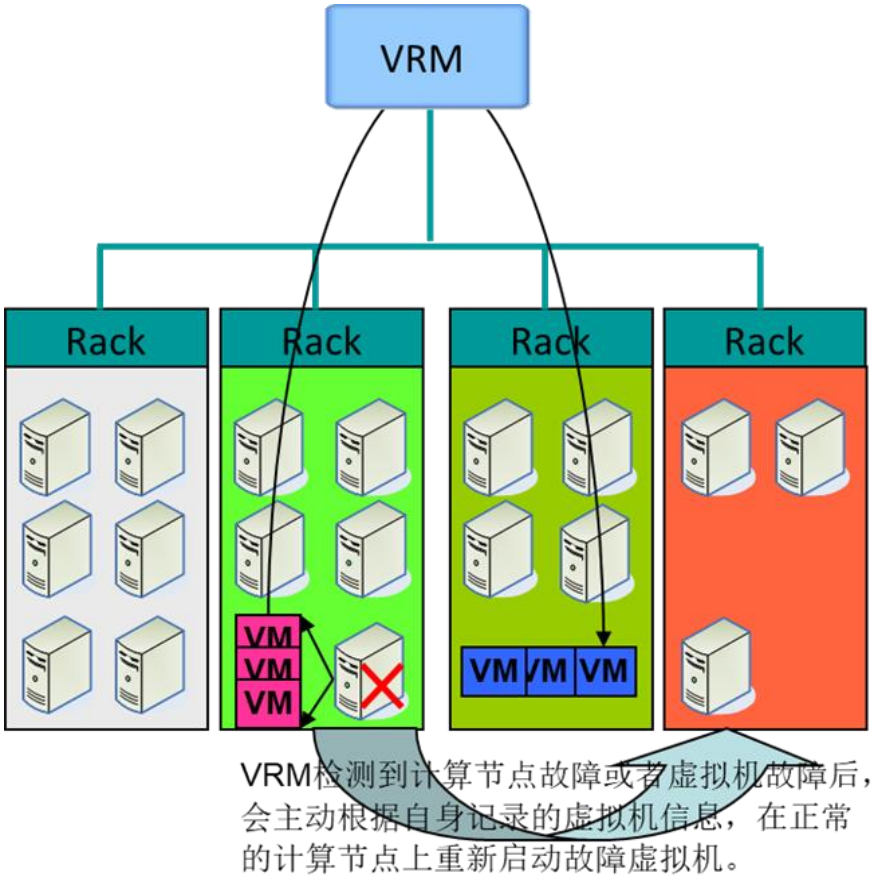


6.3.3 虚拟机 HA

当 CNA 物理服务器宕机或者重启，系统可以将具有 HA 属性的虚拟机故障迁移到其他计算服务器，保证虚拟机能够快速恢复。

当计算服务器宕机后，由于单个集群内可以运行上千个虚拟机，为避免大量虚拟机迁移造成网络拥塞和目的服务器过载，系统会根据网络流量、目的服务器负荷选择将虚拟机迁移到不同的目的服务器。

图 6-7 虚拟机 HA 特性示意图



当 VRM 与 CNA 的心跳中断超过 30 秒则会触发虚拟机 HA，当一个虚拟机有运行状态突然异常消失也会触发 HA 在其他正常的计算节点上快速恢复业务。

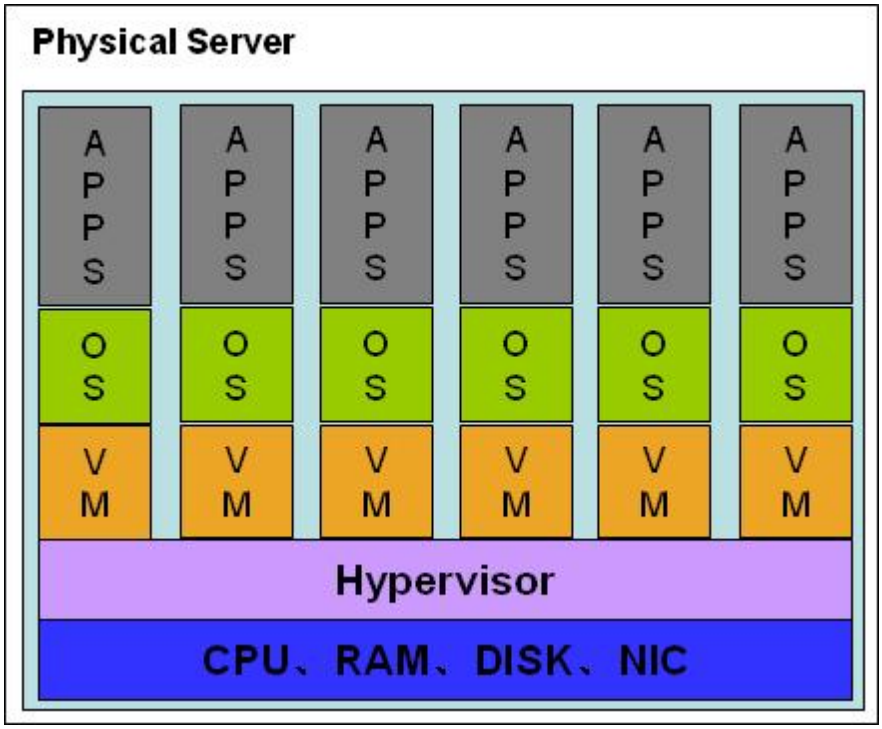
防止脑裂：通过存储层面的锁机制防止同一个虚拟机实例在多个 CNA 上同时启动。

CNA 节点的掉电恢复：CNA 节点掉电恢复后，业务进程开机自启动恢复，其上之前运行的虚拟机全部故障迁移至其他计算节点。

6.3.4 虚拟机故障隔离

虚拟机的本质就是通过虚拟化技术，将一台物理服务器虚拟成多个计算机。虚拟机之间彼此相互独立，一个虚拟机故障不会影响其他虚拟机。用户对虚拟机的使用体验和对传统物理机的体验相同。

图 6-8 虚拟化环境下的软件协议栈示意图



因此在一个虚拟机内的任何操作，不对同一台物理服务器上的其它虚拟机和虚拟化平台自身的可用性产生危害。即使虚拟机的运行出现故障，比如操作系统崩溃、应用程序错误导致死机等情况，同一物理服务器上的虚拟化平台以及其它虚拟机仍然可以正常运行，继续为用户提供服务。

6.3.5 虚拟机 OS 故障检测

当虚拟机本身发生故障时，系统能够根据用户预先设置的故障处理策略，通过主机定期检测虚拟机是否蓝屏，并决定在本地或异地重新启动虚拟机，尽快恢复业务的运行。用户也可以设置为虚拟机发生故障后不作处理，在这种故障处理策略下，系统即使检测到虚拟机发生故障，只上报告警。对于虚拟机 OS 内部故障，如 Windows 虚拟机的蓝屏故障，这类故障系统能检测到并处理。

- 增强系统的自动化维护手段，减少了维护人力投入。
- 最大限度的减少了虚拟机业务中断时间，缩短了平均故障恢复时间，提升系统可靠性。

6.3.6 黑匣子

虚拟化软件和虚拟化管理软件支持黑匣子功能，在管理节点或者计算节点出现系统崩溃、进程死锁或异常复位故障时，会将“临死信息”备份到本地目录，用于后续故障定位。

黑匣子主要用于管理节点和计算节点上收集并存储操作系统异常退出前的内核日志、诊断工具的诊断信息等数据，以便操作系统出现死机后，系统维护人员能将黑匣子功能保存的数据导出分析。为了让这些系统定位数据不丢失，黑匣子支持把操作系统死

机前收集的数据通过 netpoll 方式实时发送至远端服务器进行备份，如果网络异常则会保存在本地。

6.3.7 管理节点虚拟化部署

FusionCompute 解决方案管理软件可以选择部署到虚拟机中，即管理节点支持虚拟化部署模式。管理节点部署到虚拟机上，其本身支持主备冗余，热迁移，HA，另外：

- 主备管理节点虚拟机支持使用本地存储，除了主备管理虚拟机本身高可靠性之外，主备管理虚拟机存储还支持使用 RAID 组，进一步提升了系统的可靠性。
- FusionCompute 支持管理节点虚拟机开机自启动（即管理节点虚拟机所在主机上电，VRM 管理节点虚拟机支持自启动）。

6.3.8 主机故障恢复

CNA 节点故障更换支持如下场景：整机，硬盘，主板，网卡，RAID 卡。

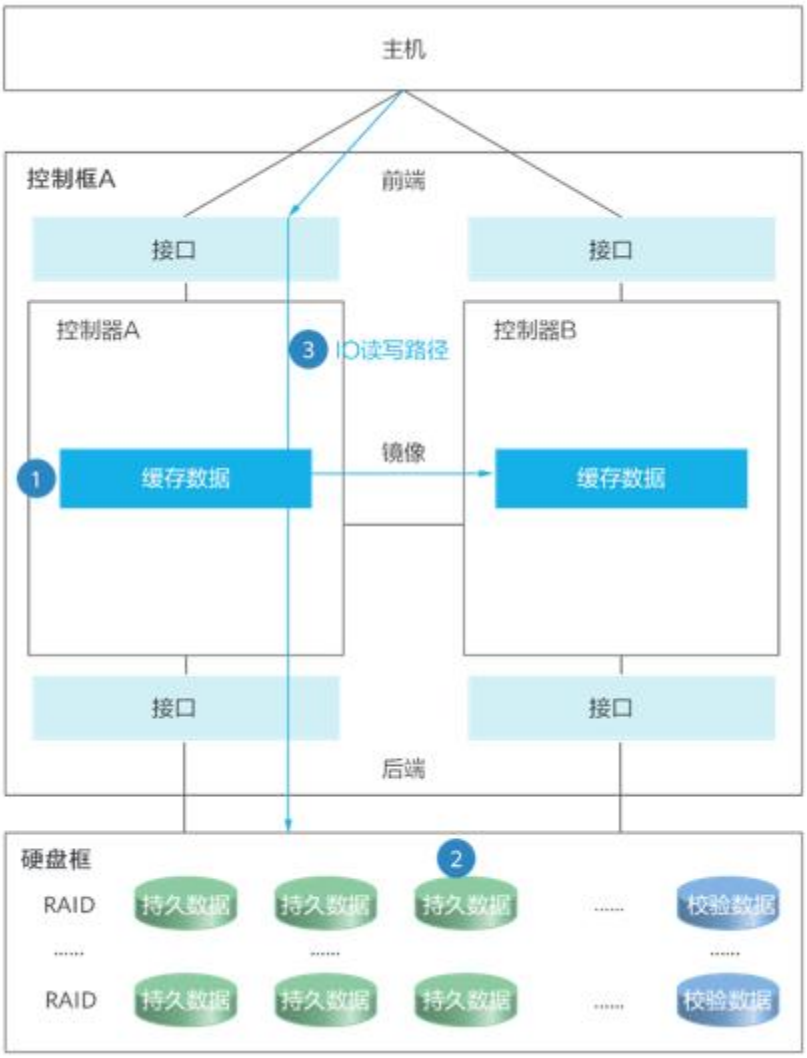
当整机或者 CNA 节点操作系统故障，通过重启或者对应的告警处理无法恢复时，FusionCompute 支持对该节点进行更换，并支持一键式或命令行方式恢复其上原有的业务和配置。主机恢复后其上绑定主机的虚拟机能够自动被拉起，并且之前在添加主机时进行的网络，存储，计算，ntp 等公共配置能够自动恢复。

6.4 存储可靠性设计

6.4.1 数据可靠性设计

针对主机写入到存储的数据，存储系统会经历（1）数据缓存，（2）盘上持久化，（3）数据路径传输三个过程。下面我们分别对这三个过程的数据可靠性措施进行说明。

图 6-9 数据可靠性全景



6.4.1.1 缓存数据可靠性保证

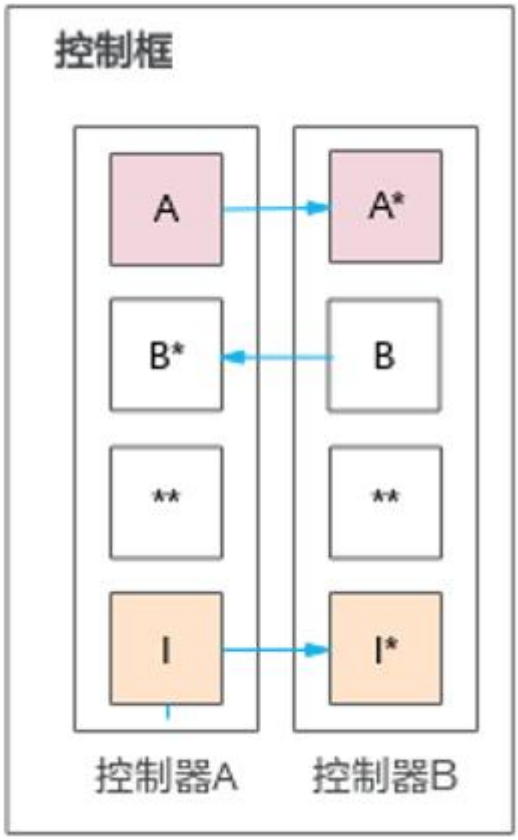
为了提升写入数据的速度，FastCube 2910 计算型存储提供了写缓存机制。即数据写入到控制器的内存缓存及其跨控的副本就向主机返回成功，然后后台将缓存数据下盘。

储存于控制器内存中的用户数据，一旦遇到系统掉电或是控制器故障，将存在内存数据丢失进而导致内存数据丢失的风险。为了应对上述故障模式，系统提供了缓存跨控多副本和掉电保护功能保证数据可靠性。

6.4.1.1.1 缓存多副本

FastCube 2910 计算型存储支持两副本，实现了控制器故障，写缓存数据不丢失，业务不中断。如下图所示，写数据在缓存到控制器 A 的时候同时镜像到控制器 B，保证 AB 控任何一个控制器故障、数据不丢失。

图 6-10 缓存多副本



6.4.1.1.2 掉电保护

FastCube 2910 计算型存储配置了 BBU 模块（备电），当系统在发生供电故障时，能利用 BBU 模块将各控制器内存中的缓存数据刷入到保险箱中。待系统电力恢复后，存储系统在启动时，再将保险箱中的缓存数据恢复到内存，保证数据不丢失。系统的掉电刷保险箱流程，由底层系统完成，不依赖上层软件单元，确保了刷盘过程不会受业务影响，进一步提升了用户数据的可靠性。

6.4.1.2 持久数据可靠性保证

FastCube 2910 计算型存储通过盘内 RAID 技术保证单盘数据可靠性，保证盘内数据不丢失；通过系统 RAID2.0+技术、HyperZoom 保证系统级的数据可靠性，即冗余范围内的单盘故障，多盘故障也不会导致数据丢失或是冗余度下降。

6.4.1.2.1 盘内 RAID

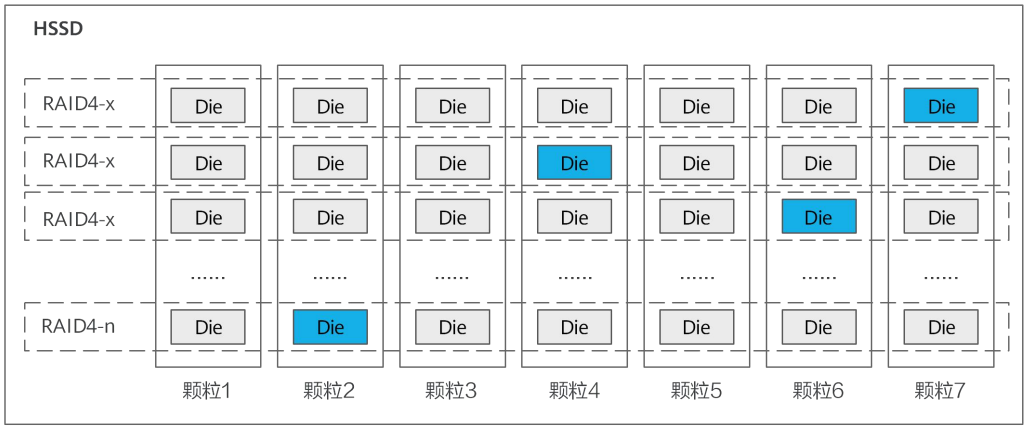
硬盘除了会出现整体故障外，用于存储数据的颗粒也会出现区域性损坏，我们称之为静默失效（坏块）。这些坏块不会体现为单盘故障，但会导致盘上相应的数据无法访问。

通过常规的坏块扫描可以提前发现静默失效数据并进行修复，但因为扫描对盘访问会占用较多资源，为了避免影响前台业务，扫描速率必须有所控制，在盘容量且数量较

大时，扫描一次花费时间少则数周，多则数月才能完成所有盘的扫描工作。而在扫描间隔期间，出现坏块并叠加盘故障，则可能导致数据损坏无法恢复。

在坏块扫描基础之上，为了应对扫描隔离内的静默失效问题，FastCube 2910 计算型存储结合 HSSD(Huawei SSD)，提供了盘内 RAID 特性。即针对盘内的数据，以 Die 为单位建立 RAID，通过 RAID4 形成冗余，容忍单个 Die 失效，盘内数据不丢失。

图 6-11 SSD 盘内 RAID



6.4.1.2.2 RAID2.0+

传统 RAID 的存储系统中 RAID 组的成员盘是固定的几个物理盘，用户使用的 LUN/FS 从某一个 RAID 组中划分。由于系统中每个 LUN/FS 的访问频度不同，这就导致系统中一部分 RAID 组中的硬盘异常繁忙，形成“热点”，而其它 RAID 组的硬盘即使空闲，也无能为力。另外，硬盘如果长期工作，它的故障率就会明显升高，结果快速故障。因此，传统 RAID 存储系统中的“热点”硬盘，时刻面临负载不均和部分盘着“过劳死”的风险。

FastCube 2910 计算型存储将每个 SSD 盘或 HDD 盘切分成固定大小的 Chunk（简称 CK，大小为 4MB），将所有盘上的 Chunk 按 RAID 冗余组成 Chunk 组，形成 RAID2.0+，相对于传统 RAID 机制，RAID2.0+具备如下优势：

- 业务负载均衡，避免热点。数据打散到资源池内所有硬盘上，没有热点，硬盘负荷平均，避免个别盘因为承担更多的写操作而提前达到寿命的上限。
- 快速重构，缩小风险窗口。当硬盘故障时，故障盘上的有效数据会被重构到资源池内除故障盘外的所有盘上，实现了多对多的重构，速度快，大幅缩短数据处于非冗余状态的时间。
- 全盘参与重构。资源池内所有硬盘都会参与重构，每个盘的重构负载很低，重构过程对上层应用无影响。

?.1.盘级冗余 RAID

盘级冗余以盘为单位进行 RAID 成员选盘，满足 RAID2.0+，均衡随机选盘，每个 Chunk 组，在每个盘上最多选 1 个 Chunk，能保证冗余内盘数故障后数据不丢失，比如 4 个硬盘框，每硬盘框 25 硬盘，共 100 块硬盘，RAID6 能保证任意坏 2 块盘业务不中断，FastCube 2910 计算型存储的冗余级别为 RAID6。

6.4.1.2.3 缩列重构（HyperZoom）

当硬盘域因为连续盘故障，盘更换等原因导致可用成员盘数小于 RAID 成员盘数时，将导致原有的重构无法进行，用户数据冗余无法保证。为了应对上述问题，FastCube 2910 计算型存储重构采用动态 RAID 重构（缩列重构）。即存储池总的可用硬盘小于 RAID 成员盘数，缩列重构时保持 M（校验列）不变，减少 N（数据列）的方式进行重构，重构前后 RAID 校验列数不变，数据列数变少。缩列重构完成后，RAID 组成员盘数减少，但是 RAID 冗余级别不变。

故障盘更换完成后，系统会根据存储池内的可用硬盘数，增加 N（数据列），新写数据就会采用新的 RAID 方式，故障期间写的数据也会逐渐转换为新的 RAID 方式。

说明

缩列重构会导致系统的总可用容量降低，当出现多盘故障时请务必及时处理盘故障并关注存储池使用利用率情况。

6.4.1.3 I/O 路径数据可靠性保证

数据在存储系统内部传输时，会经过了多个部件、多种传输通道和复杂的软件处理，其中任意一个错误都可能会导致数据错误。如果这种错误无法被立即检测出来，就可能将这些错误的数据写入到持久化的盘、内部计算或是将错误的数据返回给主机，造成业务异常。

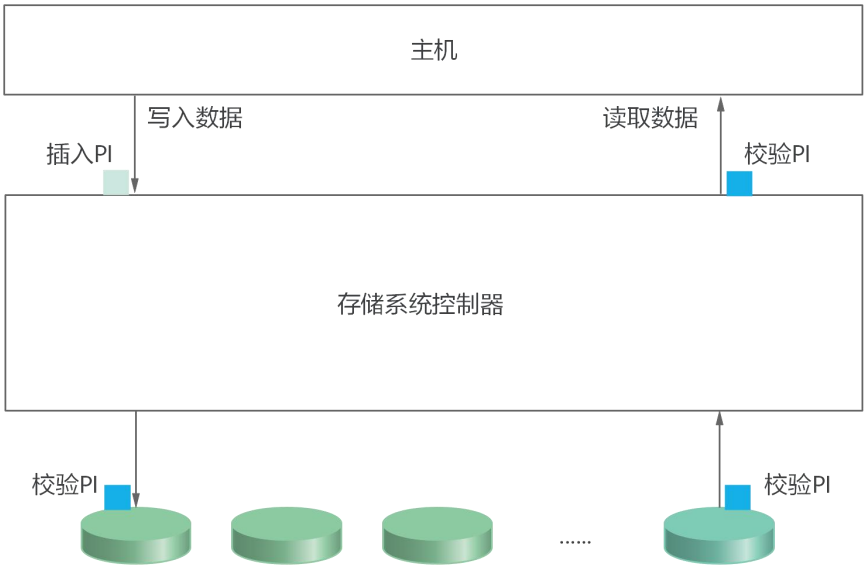
为了解决上述问题，FastCube 2910 计算型存储通过端到端 PI 功能保证传输路径数据错误（数据块内部跳变）能被检测出来并纠错。通过矩阵校验功能保证数据整体跳变（整块数据被老数据或其它数据覆盖）也能被检测出来。通过上述措施，保证 I/O 路径数据可靠性。

6.4.1.3.1 端到端 PI

FastCube 2910 计算型存储通过支持 ANSI T10 PI（Protection Information）来保证存储系统内部的数据完整性保护。当数据从主机下发到达阵列后，阵列会在收到数据后，首先对每 512 个字节插入 8 字节的 PI，然后再做内部处理。

如下所示：绿色的点表示对数据插入 PI，蓝色的点表示对 512 字节的数据计算 PI 并与存储的 8 字节的 PI 比对，识别数据是否出现错误。

图 6-12 端到端 PI

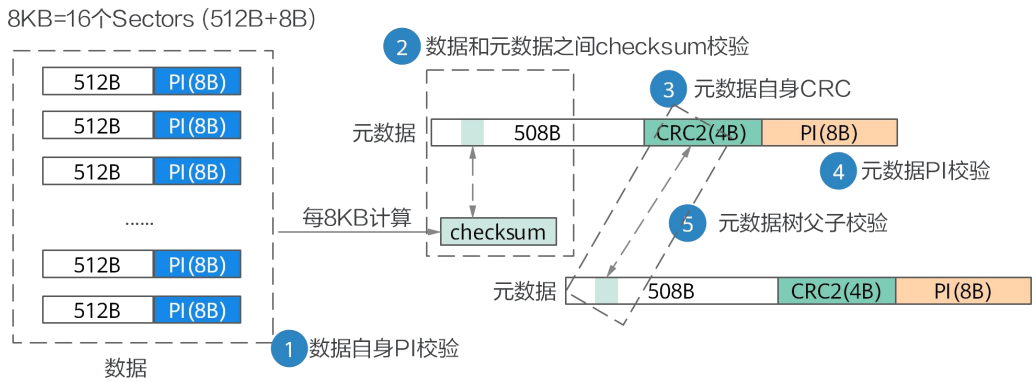


当主机读取数据时，为了防止盘上数据跳变，盘在读取数据后会对数据进行校验，一旦出错会通知上层控制器软件通过 RAID 冗余恢复跳变前的数据。为了防止数据在盘-阵列前端返回之间路径出现问题，会在向主机返回前再次校验数据是否跳变，一旦出现问题会降级读（通过 RAID 的其它成员盘计算出错误盘上的数据）恢复故障数据，从而端到端保证了阵列前端到后端数据可靠性。

6.4.1.3.2 矩阵校验

盘本身内部结构复杂或是读盘路径较长（经历多个硬件部件），极有可能出现硬件内部软件原因导致漏写(新数据未写到盘却返回成功，盘上为老数据)，整体读偏（读取 A 数据却返回了 B 数据）或写偏（本该写入 A 地址的数据写到了 B 地址）的情况发生。一旦出现此类错误，通过数据本身的 PI 校验必然通过，如果继续使用，则可能导致向主机返回错误的数据(如老数据)。

图 6-13 父子校验

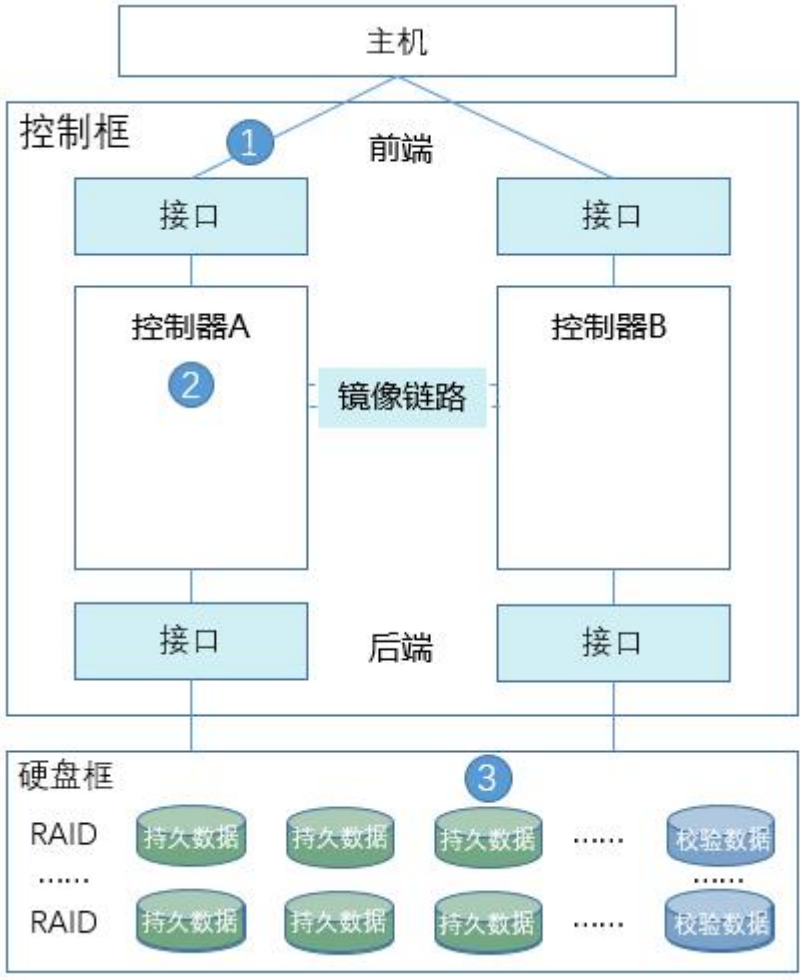


为了应对盘可能出现的漏写，读偏，写偏，FastCube 2910 计算型存储提供了矩阵校验。如上图所示：每个数据是由 512 字节的数据和 8 字节 PI 构成，其中 PI 内的 2 个字节作为 CRC 检验，横向保证 512 数据的可靠性（保护点 1）。我们将 16 个用户数据 PI 区中的 CRC 提取出来，计算 CRC 形成 checksum，然后存储在其归属的元数据节点中。一旦单个或多个数据(512+8)出现整体偏移，因为其 PI 区也产生变化，会导致 16 个数据计算的 checksum 也发生变化，这样就和其元数据中存储的 checksum 不一致。纵向保证了数据的可靠性。系统检测到数据损坏后，通过 RAID 冗余恢复故障数据，形成矩阵校验。

6.4.2 业务可用性设计

在主机访问阵列的整个路径，存储阵列提供了多重冗余保护能力。即在 I/O 经过的接口模块或链路（1）；控制板（2）；存储介质（3）出现单点故障，都能通过冗余部件和和容错措施保证业务不中断。

图 6-14 系统多冗余设计



6.4.2.1 接口模块/链路冗余保护

FastCube 2910 计算型存储是一个全冗余的存储系统。对接主机的前端，连接硬盘的后端，以及控制器之间的通讯，均有链路/接口冗余保护。当控制器故障或是更换时，其余控制器接管业务后，I/O 下发到接管控制器，保证业务连续性。控制器上板载的 P0、P1 端口在双控间为主备模式，端口或控制器故障时网络自动切换到剩余控制器。

6.4.2.2 控制器冗余保护

FastCube 2910 计算型存储为存储的关键部件控制器提供了高冗余的可靠性保证。典型场景下，缓存数据除在当前控制器存在外，还会选择另外一个控制器作为其副本，确保单控制器故障时，业务能切换到冗余的缓存副本归属控制器，保证业务连续性。

6.4.2.3 存储介质冗余保护

FastCube 2910 计算型存储除了保证单盘本身的高可靠性之外，还利用多盘冗余能力保证单盘故障损坏后的业务可用能力。即通过算法及时发现单盘故障或是亚健康，及时隔离，避免长期影响业务，然后再利用冗余技术恢复故障盘数据，持续对外提供业务能力。

6.4.2.3.1 盘故障快速隔离

FastCube 2910 计算型存储在盘正常运行过程中，会监测在位/复位信号，当盘因为更换拔出/故障时，软件感知后能快速响应此事件，将不在位/故障的盘进行隔离，新的 I/O 在其它盘触发写入，读 I/O 则通过降级读后向主机返回。

另外，经过长时间的不间断工作，硬盘会出现部分器件老化、颗粒失效等，此时硬盘响应 I/O 的速度比较慢，并可能影响到业务。这时响应慢的硬盘会被快速检测并隔离，避免对业务造成进一步的影响。

FastCube 2910 计算型存储的慢盘快速检测与隔离基于聚类算法，对硬盘类型、接口类型、所属硬盘域进行类的划分，按类统计硬盘 I/O 的平均服务时间，并建立模型进行横向比较，在较短周期内检测出慢盘并从业务中隔离，减少慢盘影响主机业务时间。

6.4.2.3.2 多盘冗余

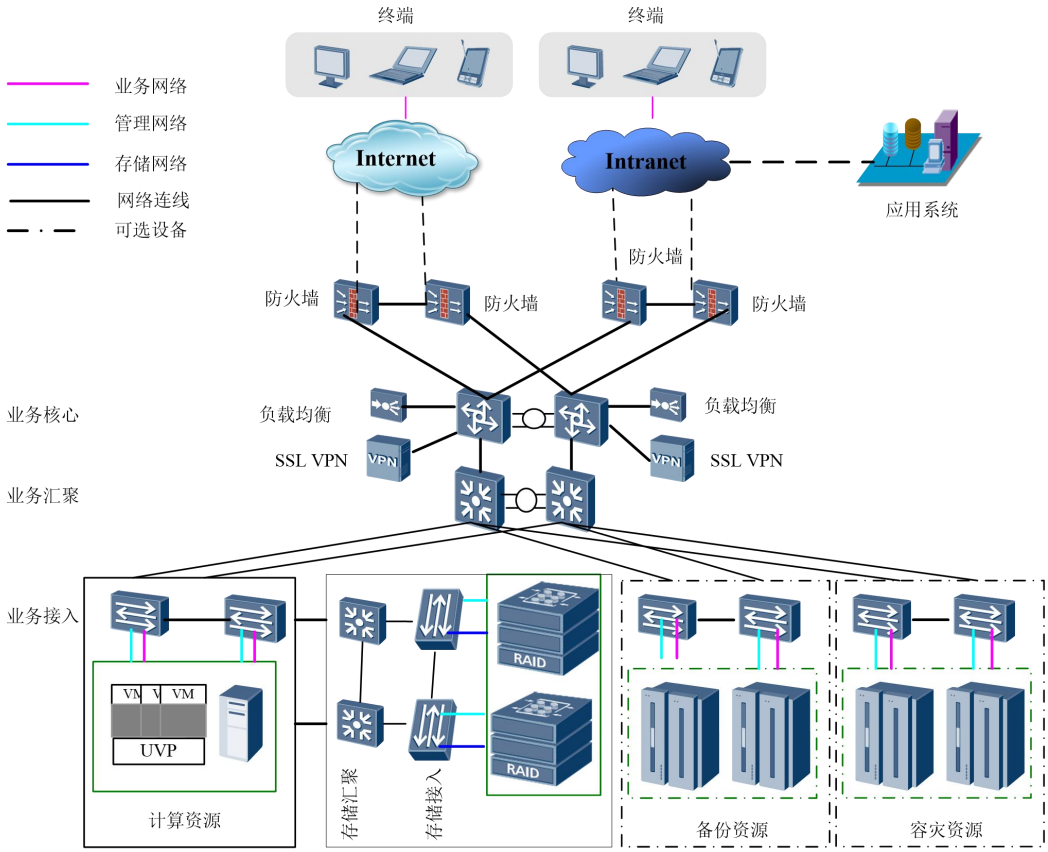
FastCube 2910 计算型存储出厂预制 RAID 6 级别存储池，采用 EC-2 算法，每个校验条带生成 2 个校验数据，最大可支持存储池内任意 2 块硬盘同时故障，数据不丢失，业务不中断能力。

6.5 网络可靠性

网络子系统主要采取以下四个措施来增强系统的可靠性。分别是：通过网卡绑定技术提高服务器端口的可用性；可以通过交换机堆叠技术将两台交换机虚拟成一台使用在提高链路的利用效率的同时大大提高了接入交换机的可靠性；同时通过 Trunk 后的 SmartLink 技术接入汇聚交换机。最后在核心路由器侧，采用 VRRP 技术部署主备两台路由器以便提高网络核心部分的可用性。

数据中心网络总体方案如下：

图 6-15 数据中心网络总体方案示意图



整体网络划分为三层，分别为：

1）接入层

服务器和存储设备上行接入到接入层交换机。

服务器侧建议采用 6 网卡（业务+管理+存储）方式进行组网，业务、管理平面分别通过两网卡聚合确保链路冗余，存储平面通过多路径确保链路冗余。

在接入交换机划分 VLAN，将管理、业务、存储三个平面逻辑隔离。为简化组网提高组网可靠性，建议接入交换机采用堆叠方式：

业务平面网络：用于承载虚拟机业务数据。

管理平面网络：用于承载管理服务器以及资源服务器之间的内部管理消息流量。

存储平面网络：用于承载服务器和磁盘阵列之间的专用数据访问。

2）汇聚层

接入交换机上行到汇聚层交换机。汇聚交换机建议采用交换机集群的方式，接入交换机采用 ETH-TRUNK 上行至汇聚交换机，汇聚交换机堆叠之后，无需启用 VRRP 功能，如果需要汇聚交换机提供网关功能，则直接将 VLAN IF 接口作为用户网关地址。

3）核心层

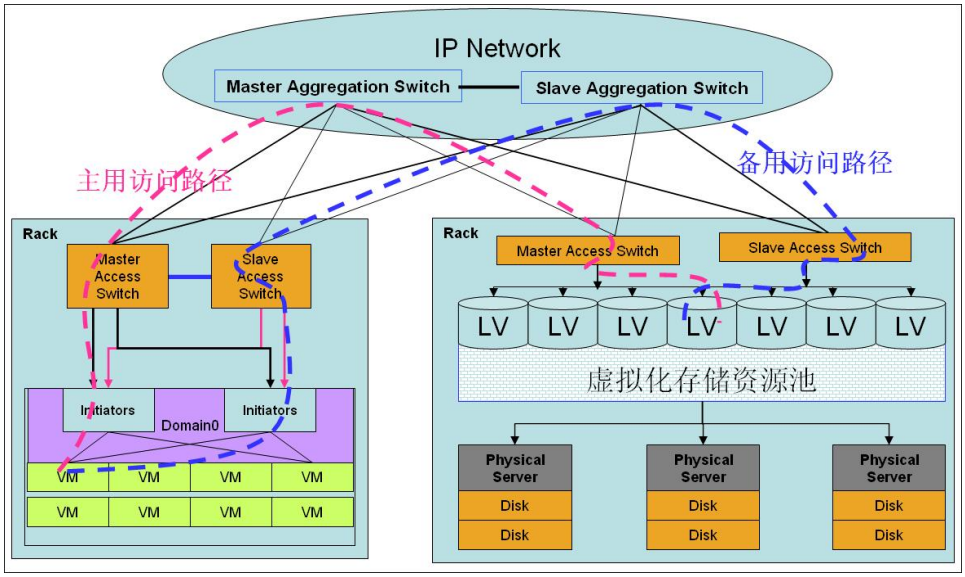
汇聚交换机上行接入核心层交换机。核心交换机也建议采用集群的方式。核心交换机采用 OSPF 或者静态路由的方式同上层设备进行对接：

当采用 OSPF 对接时，OSPF 发布地址包括核心交换机互联地址，直连路由地址以及 loopback 地址。当采用静态路由方式时,建议核心交换机同上级设备采用 VRRP 地址为网关地址。

6.5.1 存储多路径访问

计算节点支持存储 Initiators 模块的冗余部署，其上虚拟机通过标准协议（iSCSI 等）访问存储系统，并通过多块网卡的负荷分担技术、交换机的堆叠和集群技术提供存储路径的物理冗余。

图 6-16 数据存储多路径访问示意图



上图给出了计算节点和存储节点使用协议通信时的多路径访问流程，任意一个虚拟机对所挂载的任意一个虚拟卷，都将至少有两个完全冗余的路径来实现卷的多路径访问，并通过多路径软件来实现访问多路径的控制和故障切换，从而避免单点故障带来的系统可靠性问题。

6.5.2 虚拟化网络流量控制

虚拟化网络流量控制提供基于端口组的带宽配置控制能力。

支持基于端口组的保留带宽，上限带宽，带宽优先级控制能力，保证虚拟机的网络通信质量，同时避免不同端口组之间的拥塞互相影响。当某一类虚拟机由于业务需要，要求对其某个虚拟网卡使用的带宽提供保证，以保证虚拟机在拥塞的情况下仍然保持高质量的网络通信，可通过设置虚拟机网卡端口组的保留带宽来实现。当管理员需要限制某一虚拟机可占用的带宽的上限时，可通过设置虚拟机网卡的上限带宽来实现。当管理员需要拥塞情况下，对于不同的虚拟机有不同的带宽抢占能力时，可通过配置端口组带宽优先级来实现，使优先级高的虚拟机抢到更多的带宽。

6.5.3 网卡负荷分担

对于物理服务器提供的多块网卡，出于可靠性以及流量负载均衡的考虑，系统采用了Bonding 模式（支持主备和负荷分担绑定模式）。使用绑定模式之后，网卡被绑定成逻辑上的“一块网卡”后，同步一起工作，对服务器的访问流量被均衡分担到多块网卡上，这样每块网卡的负载压力就很多，抗并发访问的能力提高，保证了服务器访问的稳定和畅快，而且当其中一块发生故障的时候，另外的网卡立刻接管全部负载，过程是无缝的，服务不会中断。避免单个网卡或者链路故障引发的业务中断。

服务器绑定多网卡的实际意义在于当系统采用绑定多网卡形成阵列之后，不仅可以扩大服务器网络进出口带宽，而且可以实现有效负载均衡和提高容错能力，避免服务器出现传输瓶颈或者因某块网卡故障而停止服务。

6.6 硬件可靠性

6.6.1 内存可靠性

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

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

6.6.2 支持磁盘在线定时故障检测和预警

FastCube 2910 采用了业界先进的 S.M.A.R.T.技术标准来实现对基于 ATA 和 iSCSI 接口的硬盘进行检测和可靠性管理，检查其可靠性并预测磁盘错误。他的技术原理是主要时通过侦测硬盘各属性，如数据吞吐性能、马达起动时间、寻道错误率等属性值和标准值进行比较分析，推断硬盘的故障情况并给出提示信息，帮助用户避免数据损失。

S.M.A.R.T.是 Self-Monitoring Analysis and Reporting 系统的英文系统缩写，中文就是自监测、分析和报告技术。这个技术是现在普遍应用于硬盘的数据可靠性技术，主要是在硬盘工作的时候，对硬盘中的电机、电路、磁盘、磁头的状态进行分析，当有异常的时候就会上报告警，甚至在某些情况下，会自动降低磁盘转速并备份数据。

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

一般情况下 S.M.A.R.T.的几个主要的关键检测属性包含如下：

- Read Error Rate 错误读取率
- Start/Stop Count 启动/停止次数(又称加电次数)
- Relocated Sector Count 重新分配扇区数
- Spin up Retry Count 旋转重试次数(即硬盘启动重试次数)

- Drive Calibration Retry Count 磁盘校准重试次数
- ULTRA DMA CRC Error Rate (ULTRA DMA 奇偶校验错误率)
- Multi-zone Error Rate 多区域错误率

6.6.3 电源可靠性

FastCube 2910 提供电源故障告警，支持电源 2+2 冗余和热插拔，可以在一组电源故障后，系统持续运行而不影响业务；并且可以在线更换故障电源。

6.6.4 系统检测

FastCube 2910 支持对 CPU，内存等热关键器件的温度实时检测，配合智能的风扇调速和检测，确保系统运行的可靠性。支持对风扇，电源，硬盘等关键器件的运行状态检测，设备故障时会产生告警，可以灵活对支持热插拔设备进行在线更换，不支持热插拔设备提前安排好业务后进行下电更换。

6.6.5 板载软件可靠性

BMC 软件支持双 Image，当 Flash 中的某个 Image 遭到破坏时，支持从另一个 Image 启动 BMC 系统，而不会造成系统无法启动的情况。

BMC 软件支持进程检测，某个进程死掉后，支持重启恢复功能。

7 系统性能设计

FastCube 2910 计算型存储通过全新的硬件设计，以及从前端网络、CPU、后端网络等全 I/O 路径的优化，为客户提供极致的高 IOPS 和低时延。按照从主机到 SSD/HDD 盘的 I/O 下发的全流程，针对当前的问题和痛点，来阐述整个产品的关键性能的设计。

表 7-1 关键性能设计

I/O 全流程	面临的挑战	关键设计	性能设计原理简述
主机选路	SAN：读写请求下发到控制器后，再转发一次会增加 CPU 开销和增大时延	全局负载均衡	多路径软件（UltraPath）的选路从单纯的负载均衡模式，优化为多路径软件与控制器协商，把 I/O 下发到最终处理该 I/O 的控制器，从而做到全局负载均衡
	NAS：客户端选择服务 IP 建立访问路径，容易选择同一个 IP 使得 IP 所在网卡的性能成为瓶颈	DNS 服务 IP 动态均衡	设备内嵌 DNS 服务能力，支持不同的 IP 均衡策略，满足不同业务模型下的服务 IP 的分配策略达成业务均衡
前端	减少系统调度时延	iSCSI、NFS、SMB、交换网络驱动轮询调度	• 通过工作线程对周期性的轮询接口模块的接收队列，减少了有请求时候再唤醒工作线程带来的时延； • 支持前端、镜像和后端网络负载均衡，充分利用 CPU 能力

I/O 全流程	面临的挑战	关键设计	性能设计原理简述
控制器	如何发挥多 CPU 多核的计算能力	智能众核技术	• CPU 间按 CPU 分组分发，减少跨 CPU 调度的时延 • CPU 内按服务区分进行分区设计，减少服务相互干扰 • 服务分区内无锁设计，减少锁冲突
	不同场景情况下 CPU 分组间的负载不均	全局负载均衡	高密计算开销的任务在 CPU 内各业务分组间实现负载均衡的任务调度
后端	由于写放大导致 SSD 寿命短、性能低	多流技术	冷热数据分流减少盘内写惩罚
		ROW 满分条写	ROW 满分条写入的设计降低随机写的写放大
	SSD 后台擦除、写操作影响前台读时延	SSD 读优先	SSD 软硬件结合存储设计，提升盘内读 I/O 优先级，降低读盘时延

7.1 前端网络优化

前端网络优化主要是优化应用程序与存储器之间的时延，包括：在服务器侧的多路径选路算法优化、针对通用场景的调度优化。

7.2 CPU 计算优化

动态负载均衡

传统的 CPU 分组技术能解决各个业务的冲突域问题，但是也带来了不同的场景下各个 CPU 分组间资源利用不均衡的问题。FastCube 2910 计算型存储的动态负载均衡技术根据任务的计算开销不同差异化定义调度策略，使得任务在 CPU 内各核分组间实现负载均衡，将高密计算的任务与普通密度计算任务区分开来，作为各分组间负载均衡的调度单元，防止任务间调度相互干扰有效提升任务执行流水效率。

7.3 后端网络优化

多流技术

FastCube 2910 计算型存储的性能层采用多流技术，通过 SSD 驱动和控制器软件配合，通过系统的数据集视图有效区分数据的变更热度并将不同热度数据区分 block 进行存储布局。比如：系统产生的元数据（热数据）以及用户数据（温数据）存放在不同的 Block 中，从而增加 Block 中数据同时无效的概率，达到减少 GC 过程中搬移有效数据的数据量，提升 SSD 盘访问性能，延长寿命。

ROW 大块顺序写

FastCube 2910 计算型存储采用 ROW 大块顺序写入的设计，ROW 大块顺序写对所有写入数据（含新写和修改写）都采用满分条新写模式。这样不需要因为传统 RAID 写流程所需的数据读和校验修改写而产生 RAID 写惩罚的写数据量放大和多块同时修改一致性风险，有效降低了写入过程阵列控制器的 CPU 开销与对 SSD 盘的读写压力，并简化写操作过程中的处理逻辑提升系统容错能力。相比传统的 RAID 覆盖写（Write In Place）的方式，ROW 大块顺序写方式在随机写场景会带来更高的性能和容错处理的效率。

SSD 读优先

FastCube 2910 计算型存储采用最 SSD 盘，盘上读时延相比上一代平均下降>50us。通常盘内部的闪存介质（颗粒）有三种操作：读、写、擦除，一般擦除时延在 5~15ms，写时延在 2~4ms，而读操作通常在几十到一百 us。当闪存颗粒正在进行写或擦除操作时，读就必须等待相关操作完成后才能进行，这会导致读时延抖动很大。通过读优先技术，正在擦除或写时，如果识别到有高优先级的读请求，则取消当前操作优先处理读。通过此关键技术手段消除了盘上读的大时延。

8 系统可服务性设计

FastCube 2910 计算型存储致力给用于提供部署简单、运维便捷的体现，实现一站式交付、业务快速上线，免专业 IT 人员的傻瓜式运维，降低对 IT 管理人员的技能要求。其中安装部署方面，FastCube 2910 计算型存储在生产预安装 FusionCompute 主机系统软件并预配置存储系统，并支持使用手机 APP 扫码初始化设备，实现设备极简初始化；运维管理方面 FastCube 2910 计算型存储提供了统一的运维管理平台 DeviceManager，支持计算、存储、网络、虚拟化资源的一站式管理。

8.1 自动化部署

FastCube 2910 计算型存储在生产完成主机系统软件的安装，并支持手机 APP 扫码开局和一键式系统初始化功能，只需要完成系统基本参数配置后，则自动完成系统网络配置、计算节点集群组建和存储资源创建等系统初始化配置。

8.1.1 手机 APP 扫码开局

FastCube 2910 计算型存储支持使用手机 APP 进行设备初始化，实现设备的免专业技能极简开局：

1）根据客户网络规划，在设备进场前提前在云端完成设备的管理 IP 规划、申请 License 及其他准备活动。

2）服务人员使用 APP 扫描设备二维码识别设备后，从云端自动下载配置。

3）按照 APP 提示连接设备，自动进行设备初始化并导入 License 等配置。

4）自动上传设备信息完成 eService 建档。

图 8-1 手机扫描开局原理

使用手机 APP 完成设备初始化后，重新将设备接入客户管理网络即可交付使用。

8.1.2 系统初始化

FastCube 2910 计算型存储首次登录 DeviceManager 管理系统时，自动引导用户进行系统初始化，包括各节点的网络配置，计算集群、存储资源的创建、管理帐户密码初始化等。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。

8.2 统一运维管理

FastCube 2910 计算型存储通过 DeviceManager 管理界面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、设备管理、系统升级、健康检查和日志收集等功能。

8.2.1 业务发放管理

FastCube 2910 计算型存储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

FastCube 2910 计算型存储的 DeviceManager 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；

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。FastCube 2910 计算型存储的 DeviceManager 管理平台提供了 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。

8.2.2 一键式运维

FastCube 2910 计算型存储系统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、升级、健康检查、日志收集等功能。

一键式扩容

FastCube 2910 计算型存储系统扩容计算节点，将待扩容节点插入控制框，无需额外联线，启动待扩容节点。计算节点出厂时已安装好操作系统；节点扩容前准备好后，在 DeviceManager 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、管理账户密码等参数，即可进行系统扩容操作，自动将待扩容节点加入系统集群中并添加存储资源，扩容完成后即可部署主机业务。

一键式日志收集

FastCube 2910 计算型存储在 DeviceManager 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、BIOS 等组件）、存储系统、FusionCompute、OS、DeviceManager 等系统组件的相关日志项。

FastCube 2910 计算型存储在 DeviceManager 计算节点日志收集页面，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。

FastCube 2910 计算型存储在 DeviceManager 存储系统日志收集页面，可根据需要选择收集近期日志、全部日志和关键日志。

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

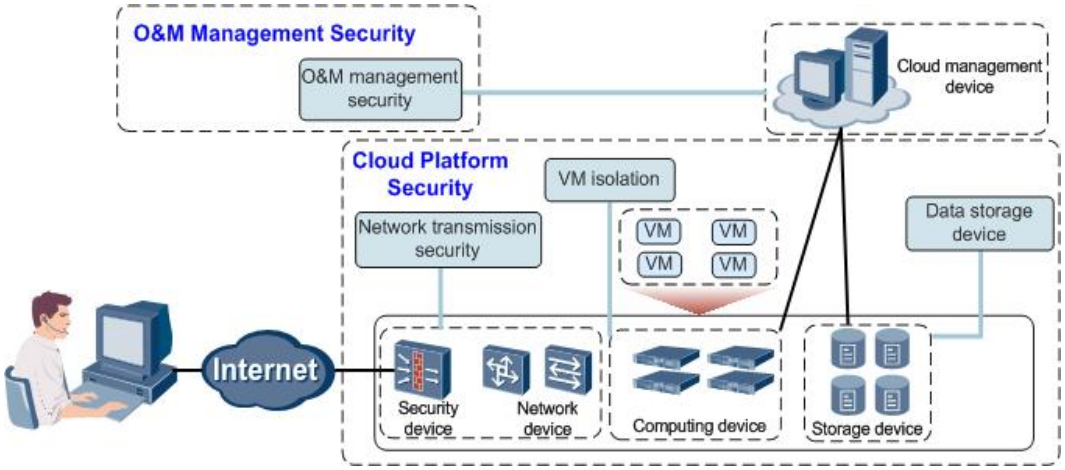
FastCube 2910 计算型存储 DeviceManager 管理界面上集成系统健康检查功能，能一键式对于系统的各个部件以及节点进行相应的系统排查，检查系统的健康状态，是否存在健康风险或故障，提供一定的排查建议，最终输出相应的巡检报告。整个的监控检查包括系统的检查和硬件兼容性检查两部分，系统检查主要排查存储系统、虚拟化系统、硬件的健康状态；硬件兼容性检查主要排查系统的硬件的版本、驱动和固件版本是否符合当前版本的要求。

9 系统安全设计

9.1 总体安全框架

根据 IT 系统面临的威胁与挑战，FastCube 2910 提供虚拟化平台安全解决方案，如图所示。

图 9-1 安全解决方案框架



分层简要介绍如下：

- 虚拟化平台安全
  - 数据存储安全

    从隔离用户数据、控制数据访问、备份数据等方面保证用户数据的安全和完整性。
  - 虚拟机隔离

    实现同一物理机上不同虚拟机之间的资源隔离，避免虚拟机之间的数据窃取或恶意攻击，保证虚拟机的资源使用不受周边虚拟机的影响。终端用户使用虚拟机时，仅能访问属于自己的虚拟机的资源（如硬件、软件和数据），不能访问其他虚拟机的资源，保证虚拟机隔离安全。
  - 网络传输安全

    通过网络平面隔离、引入防火墙、传输加密等手段，保证业务运行和维护安全。
- 运维管理安全

  从帐号密码、用户权限、日志、传输安全等方面增强日常运维管理方面的安全措施。

除上述安全方案外，还通过修复 Web 应用漏洞、对操作系统和数据库进行加固、安装安全补丁和防病毒软件等手段保证各物理主机的安全。

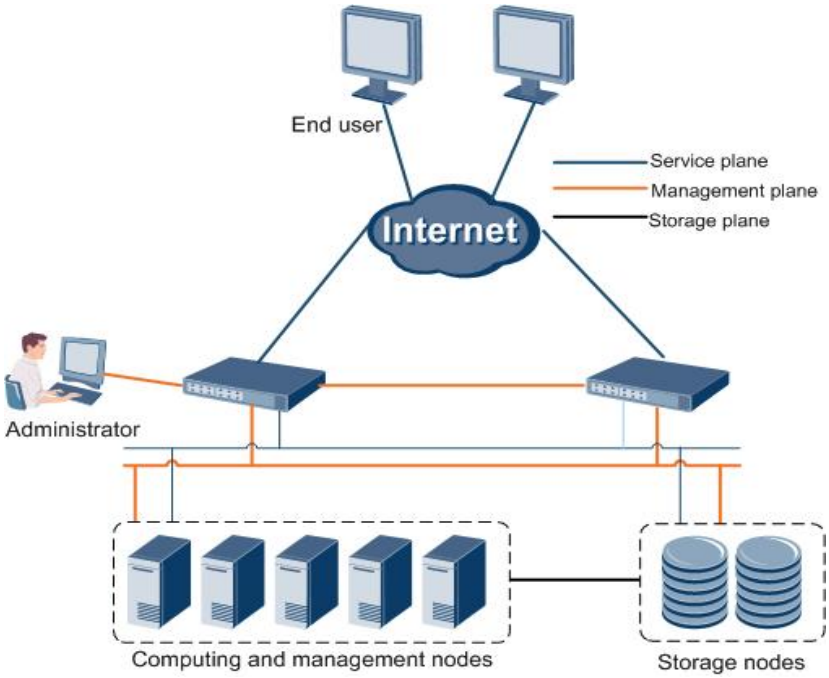
9.2 网络安全

9.2.1 网络平面隔离

将 FusionCompute 的网络通信平面划分为业务平面、存储平面和管理平面，且三个平面之间是隔离的。存储平面与业务平面、管理平面间物理隔离；管理平面与业务平面间是逻辑隔离。通过网络平面隔离保证管理平台操作不影响业务运行，最终用户不能破坏基础平台管理。

FusionCompute 的网络平面隔离如图所示。

图 9-2 网络平面隔离示意图



- 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 iSCSI 存储设备提供通信平面，并为虚拟机提供存储资源，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、业务部署、系统加载等流量的通信。

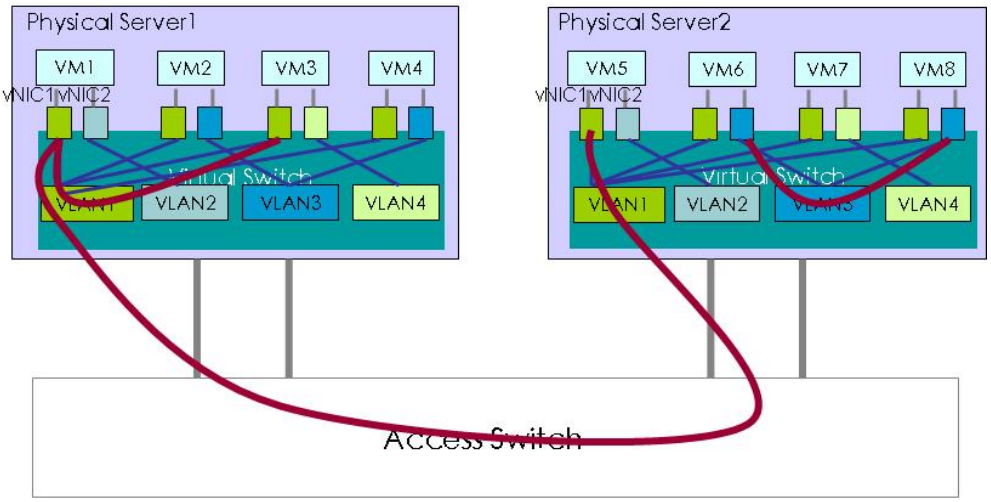
9.2.2 VLAN 隔离

通过虚拟网桥实现虚拟交换功能，虚拟网桥支持 VLAN tagging 功能，实现 VLAN 隔离，确保虚拟机之间的安全隔离。

虚拟网桥的作用是桥接一个物理机上的虚拟机实例。虚拟机的网卡 eth0，eth1，…，称为前端接口（front-end）。后端（back-end）接口为 vif，连接到 Bridge。这样，虚拟机的上下行流量将直接经过 Bridge 转发。Bridge 根据 mac 地址与 vif 接口的映射关系作数据包转发。

Bridge 支持 VLAN tagging 功能，这样，分布在多个物理机上的同一个虚拟机安全组的虚拟机实例，可以通过 VLAN tagging 对数据帧进行标识，网络中的交换机和路由器可以根据 VLAN 标识决定对数据帧路由和转发，提供虚拟网络的隔离功能。

图 9-3 VLAN 组网图



如图所示，处于不同物理服务器上的虚拟机通过 VLAN 技术可以划分在同一个局域网内，同一个服务器上的同一个 VLAN 内的虚拟机之间通过虚拟交换机进行通信，而不同服务器上的同一 VLAN 内的虚拟机之间通过交换机进行通信，确保不同局域网的虚拟机之间的网络是隔离的，不能进行数据交换。

9.2.3 防 IP 及 MAC 仿冒

通过 IP 和 MAC 绑定方式实现：防止虚拟机用户通过修改虚拟网卡的 IP、MAC 地址发起 IP、MAC 仿冒攻击，增强用户虚拟机的网络安全。具体通过生成 IP-MAC 的绑定关系，使用 IP 源侧防护(IP Source Guard)与动态 ARP 检测（DAI）对非绑定关系的报文进行过滤。

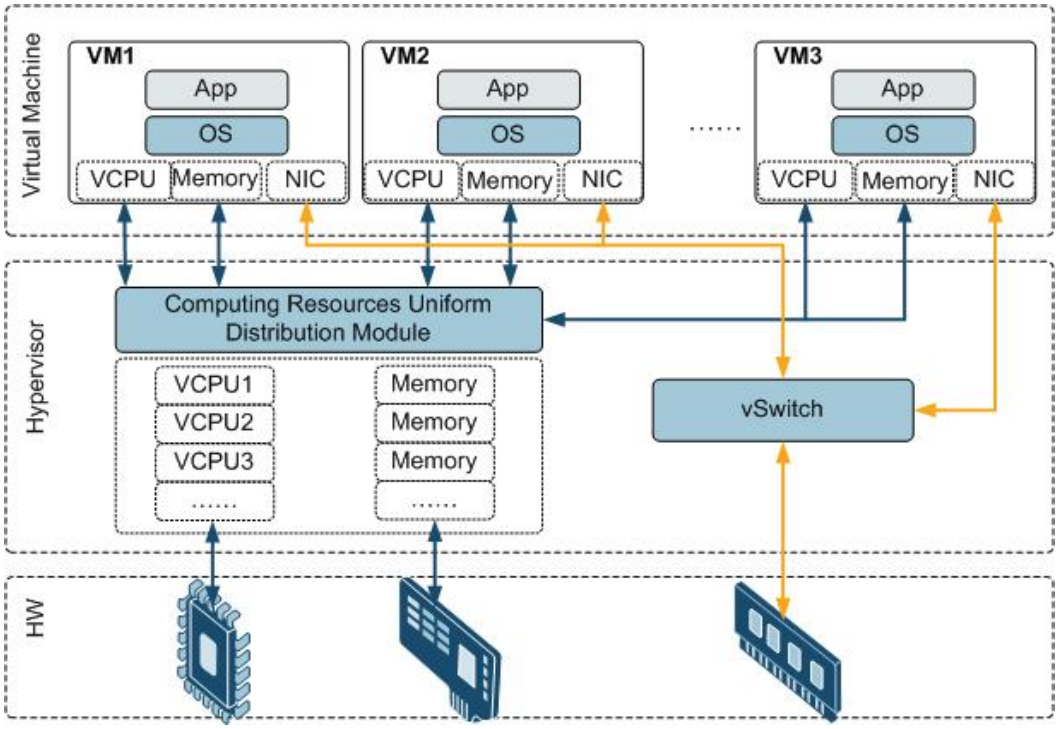
9.2.4 端口访问限制

为了减少外部服务攻击面，服务端口使用系统 iptables 机制限制端口只在固定平面监听业务消息， 并且服务进程建立 socket 时进行端口绑定固定 IP（不绑定到 0.0.0.0），通过双重手段保证端口访问安全。

9.3 虚拟化安全

Hypervisor 能实现同一物理机上不同虚拟机之间的资源隔离，避免虚拟机之间的数据窃取或恶意攻击，保证虚拟机的资源使用不受周边虚拟机的影响。终端用户使用虚拟机时，仅能访问属于自己的虚拟机的资源（如硬件、软件和数据），不能访问其他虚拟机的资源，保证虚拟机隔离安全。虚拟机隔离如图所示。

图 9-4 虚拟机相关资源隔离



9.3.1 vCPU 调度隔离安全

虚拟化平台支持裸金属架构虚拟化，借助物理 CPU 提供的 VT-X 技术，在单一物理平台上支持多个虚拟机同时运行。虚拟化层借用 Linux 内核公平调度算法，以分时复用的方式调度虚拟机的虚拟 CPU（vcpu）到物理 CPU（pcpu）上运行，每个 vcpu 寄宿在 Qemu 线程进行调度。

每个 vcpu 对应一个 VMCS（Virtual-Machine Control Structure）结构，当 vcpu 被从 pcpu 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 pcpu 上。通过这种方式，实现各 vcpu 之间的隔离。

9.3.2 内存隔离

虚拟机通过内存虚拟化来实现不同虚拟机之间的内存隔离。内存虚拟化技术在客户机已有地址映射（虚拟地址和机器地址）的基础上，引入一层新的地址——“物理地址”。在虚拟化场景下，客户机 OS 将“虚拟地址”映射为“物理地址”；Hypervisor 负责将客户机的“物理地址”映射成“机器地址”，实际物理地址后，再交由物理处理器来执行。

9.3.3 内部网络隔离

Hypervisor 提供内部网络隔离机制，每个客户虚拟机都有一个或者多个在逻辑上的网络接口 VIF（Virtual Interface）。从一个虚拟机上发出的数据包，先到达 Hypervisor，由 Hypervisor 通过 VLAN、数据包完整性校验、安全组等规则来实现数据过滤和隔离；经过 Hypervisor 处理后，合规的数据包会转发给目的虚拟机。

9.3.4 磁盘 I/O 隔离

Hypervisor 采用分离设备驱动模型实现 I/O 的虚拟化。该模型将设备驱动划分为前端驱动程序、后端驱动程序和原生驱动三个部分，其中前端驱动在虚拟机中运行，而后端驱动和原生驱动则在 Hypervisor 中运行。前端驱动负责将虚拟机的 I/O 请求传递到 Hypervisor 中的后端驱动，后端驱动解析 I/O 请求并映射到物理设备，提交给相应的设备驱动程序控制硬件完成 I/O 操作。换言之，虚拟机所有的 I/O 操作都会由 Hypervisor 截获处理；Hypervisor 保证虚拟机只能访问分配给它的物理磁盘空间，从而实现不同虚拟机存储空间的安全隔离。

9.4 数据安全

9.4.1 数据加密

对于存储在本地的敏感数据使用安全加密算法（PBKDF2， AES_128_CBC）加密保存。

对于传输的敏感数据使用 TLS 传输通道，保证数据的机密性、完整性。

9.4.2 用户数据隔离

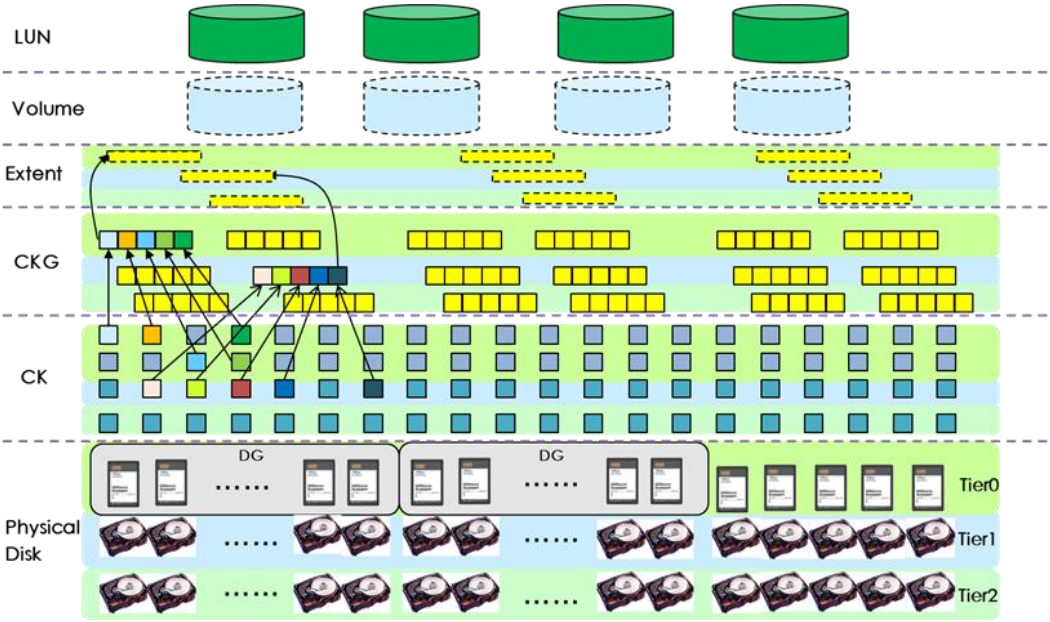
Hypervisor 采用分离设备驱动模型实现 I/O 的虚拟化。该模型将设备驱动划分为前端驱动程序、后端驱动程序和原生驱动三个部分，其中前端驱动在虚拟机中运行，而后端驱动和原生驱动则在主机中运行。前端驱动负责将虚拟机的 I/O 请求传递到主机中的后端驱动，后端驱动解析 I/O 请求并映射到物理设备，提交给相应的设备驱动程序控制硬件完成 I/O 操作。换言之，虚拟机所有的 I/O 操作都会由 VMM 截获处理；VMM 保证虚拟机只能访问分配给它的物理磁盘空间，从而实现不同虚拟机硬盘空间的安全隔离。

9.4.3 数据访问控制

系统对每个卷定义不同的访问策略，没有访问该卷权限的用户不能访问该卷，只有卷的真正使用者（或者有该卷访问权限的用户）才可以访问该卷，每个卷之间是互相隔离的。

9.4.4 剩余信息保护

存储采用 RAID 创新技术，系统会将存储池空间划分成多个小粒度的数据块，基于数据块来构建 RAID 组，使得数据均匀地分布到存储池的所有硬盘上，然后以数据块为单元来进行资源管理，大小范围是 256KB~64MB（可调），默认 4MB。



虚拟机删除或数据卷删时，系统进行卷（Volume）资源回收时，小数据块链表将被释放，进入资源池。存储资源重新利用时，再重新组织小数据块，这样从新分配的虚拟磁盘恢复原来数据的可能性很小。

数据中心的物理硬盘更换后，需要数据中心的系统管理员采用消磁或物理粉碎等措施保证数据彻底清除。

9.4.5 数据备份

FusionCompute 的数据存储采用多重备份机制，每一份数据都可以有一个或者多个备份，即使存储载体（如硬盘）出现了故障，也不会引起数据的丢失，同时也不会影响系统的正常使用。

系统对存储数据按位或字节的方式进行数据校验，并把数据校验信息均匀的分散到阵列的各个磁盘上；阵列的磁盘上既有数据，也有数据校验信息，但数据块和对应的校验信息存储于不同的磁盘上，当某个数据盘被损坏后，系统可以根据同一带区的其他数据块和对应的校验信息来重构损坏的数据。

9.4.6 软件包完整性保护

软件包完整性保护方面包括：

系统安装包在官网发布时携带软件包的数字签名文件，用户在安装前可根据产品资料验证软件包的完整性。

系统的升级包也有数字签名完整性保护机制，升级包上传过程会自动校验完整性。

9.5 运维管理安全

在运维管理方面，主要的安全威胁包括：

- 管理员权限不支持精细化控制。

- 采用弱密码，且长期不进行修改，导致密码泄露。
- 管理员恶意行为无法监控、回溯。

9.5.1 管理员分权管理

管理员通过 Portal 登录管理云系统，包括查看资源、发放虚拟机等。

系统支持对 Portal 用户进行访问控制，支持分权管理，便于维护团队内分职责共同有序地维护系统。

9.5.2 账号密码管理

管理员支持设置密码策略，确保密码的保密性。例如：可以设置密码最小长度、密码是否含特殊字符、密码有效时长等。

密码在系统中不会明文存储。

所有账户的密码均支持修改。

系统同时支持密码和公私钥对身份进行认证，公私钥支持替换。

支持弱口令校验。系统默认启用弱口令校验，包括 Web 账户、操作系统账户，并支持自定义弱口令字典，提升系统运维安全性。

9.5.3 日志管理

FusionCompute 支持以下三类日志：

- 操作日志

  操作日志记录操作维护人员的管理维护操作，日志内容详实，包括用户、操作类型、客户端 IP、操作时间、操作结果等内容，以支撑审计管理员的行为，能及时发现不当或恶意的操作。操作日志也可作为抗抵赖的证据。

- 运行日志

  运行日志记录各节点的运行情况，可由日志级别来控制日志的输出。

  各节点的运行日志包括级别、线程名称、运行信息等内容，维护人员可通过查看运行日志，了解和分析系统的运行状况，及时发现和处理异常情况。

- 黑匣子日志

  黑匣子日志记录系统严重故障时的定位信息，主要用于故障定位和故障处理，便于快速恢复业务，计算和管理节点异常时产生的黑匣子日志本地存放。

9.5.4 传输加密

管理员访问管理系统，均采用 HTTPS 方式，传输通道采用 TLS 加密。

安全证书默认使用华为提供的，支持替换为客户自己的或第三方机构颁发的。

客户端支持认证服务端证书，如果发现不合法，通过告警方式通知运维人员，防止恶意用户伪造服务端。

9.5.5 数据库备份

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。FusionCompute采用华为研发的高斯数据库，支持本地在线备份方式和异地备份方式：

- 本地备份：数据库每天定时执行备份脚本完成备份。
- 本地主机备份：VRM 虚拟化部署时，系统默认将管理数据备份到 VRM 虚拟机所在主机上。
- 异地备份：数据异地备份到第三方备份服务器。

9.6 基础设施安全

基础设施安全是指 FusionCompute 中各设备、节点以及组件的操作系统、数据库等安全性。例如，系统中大量使用的 OS、DB 等通用软件，其软件自身的漏洞、不安全的账号和口令、不当的配置和操作、开启不安全的服务等等为病毒、黑客、蠕虫、木马等的入侵提供了方便之门，使得系统容易遭受病毒入侵、漏洞攻击、拒绝服务等等安全威胁，从而影响系统的运营。保障基础设施的安全性，是维持系统正常运行、构建网络安全和应用安全的基础。

9.6.1 操作系统加固

FusionCompute 中计算节点、管理节点均使用华为欧拉 Linux 操作系统，为保证此类设备的安全，必须对欧拉 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：

- 最小化服务： 禁用多余或危险的系统后台进程和服务，如邮件代理、图形桌面、telnet、编译工具、调试工具等。
- 服务加固：对 SSH 等常用服务进行安全加固。
- 内核参数调整：修改内核参数，增强操作系统安全性，如禁用 IP 转发、禁止响应广播请求、禁止接受/转发 ICMP 重定向消息。
- 文件目录权限设置：结合业界加固规范及应用要求，保证文件权限最小化。
- 帐号口令安全 ：启动口令复杂度检查、密码有效期、登录失败重试次数等。
- 系统认证和授权：禁止 root 远程登录、尽量不用 root 账号安装运行进程。
- 日志和审计：记录服务、内核进程运行日志，可以与日志服务器对接。
- 系统完整性检查：默认安装入侵检测程序 AIDE，支持系统文件完整性异常告警。
- SUDO 配置加固：检查 Sudoers 配置、检查定时任务配置、自定义脚本权限及参数合法性，确保不会出现 ROOT 提权。

9.6.2 Web 安全

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。通过在 Web 服务端设置 Strict-Transport-Security 强制要求客户端使用 HTTPS 连接。

- 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

  跨站请求伪造是指用户登录 A 网站且在 Session 未超时情况下，同时登录 B 网站（含攻击程序），攻击者可在这种情况下获取 A 网站的 Session ID，登录 A 网站窃取用户的关键信息。

- 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。

- 帐号密码安全

  Web 帐号和密码满足系统账号密码安全原则。

9.6.3 数据库加固

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

- 设置高复杂度的帐户密码。
- 记录数据库的操作日志。
- 支持防暴力破解。
- 主备数据同步使用 TLS 加密通道。

9.6.4 Web 容器加固

虚拟化平台对外 web 服务使用 Tomcat 作为 Web 容器，主要的安全加固包括：

- 使用非 root 用户运行
- 禁止自动部署
- 定制错误页面信息

9.6.5 安全补丁

软件因自身设计缺陷而存在很多漏洞，需要定期为系统安装安全补丁以修补这些漏洞，以防止病毒、蠕虫和黑客利用操作系统漏洞对系统进行攻击。虚拟化提供安全补丁方案如下：

- 虚拟化平台安全补丁

  用户可以通过升级工具，将系统安全补丁安装到虚拟化平台上。

- 用户虚拟机安全补丁

  FusionCompute 没有针对用户虚拟机提供额外的安全补丁机制。用户根据操作系统安全补丁官方的发布情况，结合实际的使用需求，为用户虚拟机定期安装安全补丁。

9.6.6 安全编译

自研二进制和开源通过源码编译的二进制使用安全编译选项，通过缓冲区保护、安全异常终止、地址空间随机化等多个安全编译技术，全面提高产品安全性。

9.6.7 防病毒

在 X86 场景下各管理节点或虚拟机上部署防病毒软件，防止 FusionCompute 遭受病毒入侵。

- 管理节点防病毒

  管理节点提供外部操作维护 Portal，与外界存在交互操作，存在病毒感染风险。但管理节点采用加固的 Linux 操作系统，病毒感染风险低。

  用户可以自行选择为管理节点部署兼容欧拉 linux 的防病毒软件。

说明

1、用户为管理节点部署防病毒软件时，需对该防病毒软件做兼容性测试。

2、管理节点指 VRM 节点。

3、对计算节点（CNA），由于采用裁剪及安全加固的 Linux 操作系统，病毒感染风险极低，不建议安装防病毒软件。

9.6.8 深度报文检测（DPI）

在 X86 场景下虚拟化平台提供深度报文检测接口，配合深度报文检测软件，为 FusionCompute 虚拟化平台提供报文检测能力。其部署方案如下：

- GVM：安全用户虚拟机，使用虚拟机 DPI 功能的最终用户虚拟机。
- SVM：为安全用户虚拟机提供网络入侵检测、网络漏洞扫描、防火墙服务。

说明

1. 为 SVM 部署 DPI 检测软件时，需对该检测软件做兼容性测试。
2. DPI 检测软件由第三方安全厂商提供。

10 缩略语

英文缩写	英文全称	中文全称
FRU	Field Replaceable Unit	现场可更换单元
FlashLink®	FlashLink®	盘控配合技术
CK	Chunk	数据块
CKG	Chunk Group	数据块组
DIF	Data Integrity Field	数据完整性字段
RDMA	Remote Direct Memory Access	远程直接数据存取
FC	Fibre Channel	光纤通道
FTL	Flash Translation Layer	FLASH 转换层
GC	Garbage Collection	垃圾回收
LUN	Logical Unit Number	逻辑单元号
OLAP	On-Line Analytical Processing	联机分析处理系统
OLTP	On-Line Transaction Processing	联机事务处理系统
OP	Over-Provisioning	预留空间
RAID	Redundant Array of Independent Disks	独立磁盘冗余阵列
RAID-TP	Redundant Array of Independent Disks-Triple Parity	独立磁盘冗余阵列-3 盘冗余
SAS	Serial Attached SCSI	串行 SCSI
SCSI	Small Computer System Interface	小型计算机系统接口

英文缩写	英文全称	中文全称
SSD	Solid State Disk	固态硬盘
T10 PI	T10 Protection Information	T10 数据保护信息
VDI	Virtual Desktop Infrastructure	虚拟桌面架构
VSI	Virtual Server Infrastructure	服务器虚拟化架构
WA	Write Amplification	写入放大
Wear Leveling	Wear Leveling	磨损均衡
TCO	Total Cost of Ownership	总体拥有成本
DC	Data Center	数据中心
DCL	Data Change Log	数据变更日志
TP	Time Point	时间点
GUI	Graphical User Interface	图形用户界面
CLI	Command Line Interface	命令行界面
eDevLun	External Device LUN	由第三方阵列创建的阵列逻辑空间
FIM	Front-end Interconnect I/O Module	前端共享卡
SCM	Storage Class Memory	存储级内存
FRU	Field Replaceable Unit	现场可更换单元
PI	Protection Information	保护信息
SFP	Similar Fingerprint	相似指纹
DTOE	Direct TCP/IP Offloading Engine	TOE 直通技术
NUMA	Non-uniform Memory Access	非一致性内存访问
ROW	Redirect-on-Write	写时重定向
PID	Proportional Integral Derivative	比例微积分算法
SMP	Symmetrical Multiprocessor System	对称多处理器系统
NAS	Network Attached Storage	网络存储设备
CSI	Container Storage Interface	容器存储接口
SMB	Server Message Block	服务器信息块协议，用于在计算机间共享文件、打印机、串口等。

英文缩写	英文全称	中文全称
NFS	Network File System	网络文件系统，主要应用在 UNIX 环境下。
CIFS	Common Internet File System	通用互联网文件系统，主要应用在 NT/Windows 环境下。