

**InCloud dSAN V6**

**技术白皮书**

济南浪潮数据技术有限公司

2024年5月

**声明**

版权声明

版权所有©济南浪潮数据技术有限公司 保留所有权利（包括但不限于增加、删除、改动等），济南浪潮数据技术有限公司拥有最终解释权。

本文件及本文件的相关内容所包含或涉及的（包括但不限于文字、图表、商标、标识或名称、按钮图标、图像、照片、页面设计、数据及软件等）所有知识产权（包括但不限于版权、商标权、专利权、商业秘密等）及相关权利，均归济南浪潮数据技术有限公司或其关联公司所有。未经济南浪潮数据技术有限公司书面许可，任何人不得擅自对本文件及其内容进行使用（包括但不限于转载、修改、复制、发行、出售、发表、或以其他方式展示、传播等）。

安全声明

公司产品不会主动获取或使用用户的个人数据，仅在您同意使用特定功能或服务时，在业务运营或故障定位的过程中可能会获取或使用用户的某些个人数据（如告警邮件接收地址、IP 地址），公司产品在涉及个人数据的收集、存储、使用、传输、删除等全生命周期的处理活动中，已在产品功能上部署了必要的安全保护措施，同时，您也有义务根据所适用国家或地区的法律法规制定必要的用户隐私政策并采取足够的措施以确保用户的个人数据受到充分的保护。

济南浪潮数据技术有限公司高度重视产品数据安全，公司产品在涉及系统运行和安全数据的全生命周期处理活动中，已严格按照相关法律法规及监管要求，在产品功能上部署了必要的安全保护措施。作为系统运行和安全数据处理者，您有义务根据所适用国家或地区的法律法规制定必要的数据安全政策并采取足够的措施以确保系统运行和安全数据受到充分的保护。

特别提示

您购买的产品、服务或特性等应受济南浪潮数据技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，济南浪潮数据技术有限公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新，如有变更，恕不另行通知。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保，济南浪潮数据技术有限公司不对本文档中的遗漏、变更及错误所导致的损失和损害承担任何责任。

警告：在居住环境中，运行此设备可能会造成无线电干扰。

注意：身体部位远离风扇叶片。

注意：为避免电击危险，请在维修前断开所有电源线。

联系我们

售前咨询热线：400-8606708

售后服务热线：400-8600011

浪潮数据官网：<https://www.inspur.com>

存储虚拟化

### 分布式存储

浪潮分布式块存储系统具有高性能，高可靠等特点，专为超融合架构设计，优化的。充分利用了虚拟化平台特点，以及 SSD 磁盘的特性，进行了深入的优化与整合，使得在面对大规模虚拟化的场景中具有高度优势。

#### 分布式存储架构

浪潮分布式存储组成部分如下：

1）集群管理：一个高可用、强一致性的服务发现存储组件，依靠其服务发现、消息发布与订阅、分布式通知与协调能力为分布式存储系统提供集群管理；

2） 资源管理：VDI是对分布式存储中对象资源集合的逻辑抽象，VDI manager作为VDI资源管理的入口，提供南北向接口服务；北向接口为协议层提供块存储接入服务，支持私有协议对接QEMU提供虚拟磁盘、标准的iSCSI协议服务提供LUN、高性能的NVMf协议提供NVMe设备来构建存储资源；南向接口为数据对象提供统一管理，通过存储策略与一致性协议实现对数据对象的管理，同时借助集群管理提供的的数据库服务，实现VDI元数据的管理；

3）存储引擎：Scheduler根据IO请求与元数据管理，将数据对象请求进行拆分、合并与业务均衡访问，由Gateway负责数据对象请求在节点间的分发与数据副本强一致性保障；分布式存储采用基于裸盘管理的存储引擎设计，为物理磁盘设备构建对象存储服务来提供数据对象存储；相较于传统基于文件系统的存储引擎设计可以跨过文件系统，缩短IO路径，提升落盘性能；MD Manager负责磁盘服务的管理，通过一致性哈希算法均衡数据对象请求在磁盘间访问，同时提供磁盘插拔，慢盘、坏盘检测等物理磁盘检测功能；面对大量的数据对象请求，分布式存储设计了智能缓存算法，将磁盘服务划分为缓存层和数据层，通过智能缓存算法区分冷、热数据，提高缓存命中率，提升分布式存储性能。

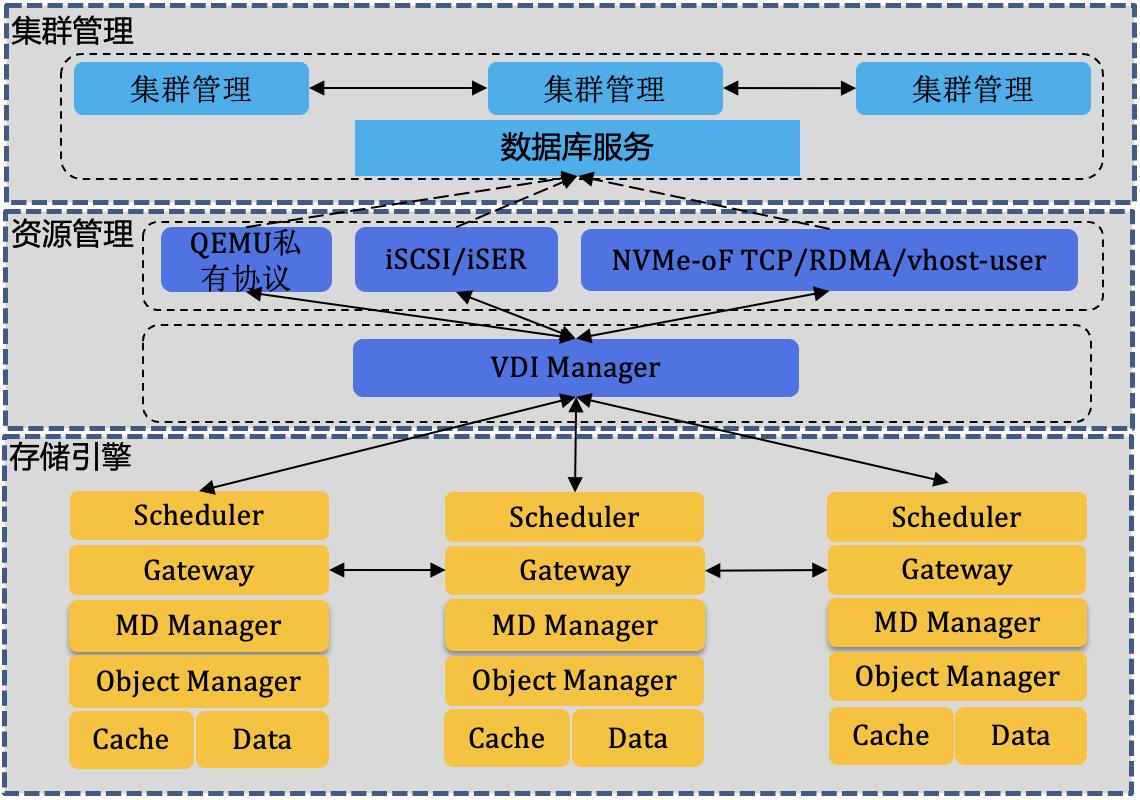


图4.3.1-1分布式存储架构

#### 分布式存储池管理

存储池是对存储介质进行组织的基本单元。存储服务集群中，不同的存储介质可以组成不同架构的存储池，使得存储池具备不同的存储特性，主要分为SSD/HDD 混合存储池和全闪 SSD 存储池，服务器是存储池成员管理的最小单位。一份数据的所有副本，都会保存在同一个存储池中。数据在存储池之间并不共享。如果数据需要在存储池之间进行移动，需要触发数据全量拷贝。

1. SSD/HDD 混合存储池：以SSD作为缓存盘、HDD作为数据盘构建存储池，数据盘承载核心数据存储，基于两层RLU缓存设计将IO命中数据迁移至缓存盘中，加速IO读写性能；
2. 全闪SSD存储池：以SSD作为数据盘，无缓存盘设置，所有数据操作均发生于SSD中，充分发挥全闪极致性能。

#### 集群管理

InCloud dSAN基于自研实现分布式存储中集群管理：包括节点离开、加入，网络异常，集群重启，集群内消息同步与配置信息共享等；

集群消息同步：采用集群管理提供的的KV数据库以及监听订阅机制实现集群内节点间消息的订阅与通知

节点管理：定义节点角色、状态，通过集群管理服务提供的租约Key机制，实现节点上线、离线动态通知，服务重启管理

集群管理：定义集群状态，通过节点状态变化实现集群状态的管理

分布式任务流处理框架：采用集群管理服务的消息订阅机制实现任务流在分布式系统中的状态同步与分布式集群锁语义，KV数据库能力实现任务的保存与查询

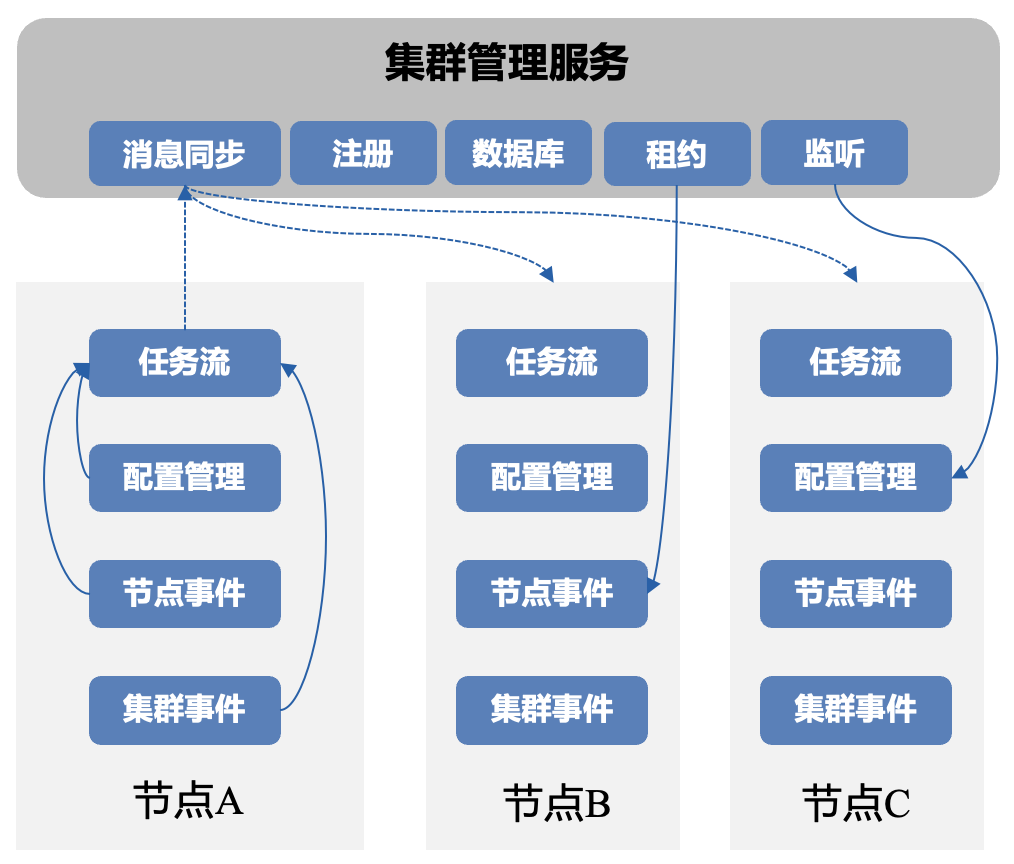


图4.3.1-2基于集群管理服务组件的集群管理技术

InCloud dSAN基于自研集群管理服务设计和实现了原型系统中的集群管理功能与分布式任务流处理框架，该框架适用于配置管理和数据恢复，实现了统一流程管理；

#### 存储资源

InCloud dSAN将不同存储协议下的资源抽象管理，摒弃协议层之间的差异，统一通过协议层对外提供服务的方式，包括：

* **资源定义**：针对不同存储协议类型，存储后端的提供的逻辑设备不同，如iSCSI协议的Tragte和LUN，NVMe-oF协议的Subsystem和Bdev，将不同逻辑资源抽象为设备类型进行管理
* **统一抽象**：统一上层存储协议后，针对不同存储协议层的逻辑设备，设计统一的VDI卷机制，通过抽象元数据、数据、配置、IO等入口适配上层存储业务逻辑
* **数据块管理：**将VDI卷拆分为元数据和数据，元数据主要记录资源大小、名称、精简、对象等信息，数据则按照既定规划拆分成相同大小的数据块
* **协议扩展**：从性能角度分析超融合场景下以bypass内核方式访问VDI的路径，扩展设计了私有协议
* **网络协议扩展**：面向RDMA编程，扩展协议层，构建前端协议RDMA技术支持

InCloud dSAN通过一套抽象的分布式存储资源，规范存储协议、存储卷、数据块等不同维度的接口规范，便于存储资源的管理和扩展，具体如下图所示：

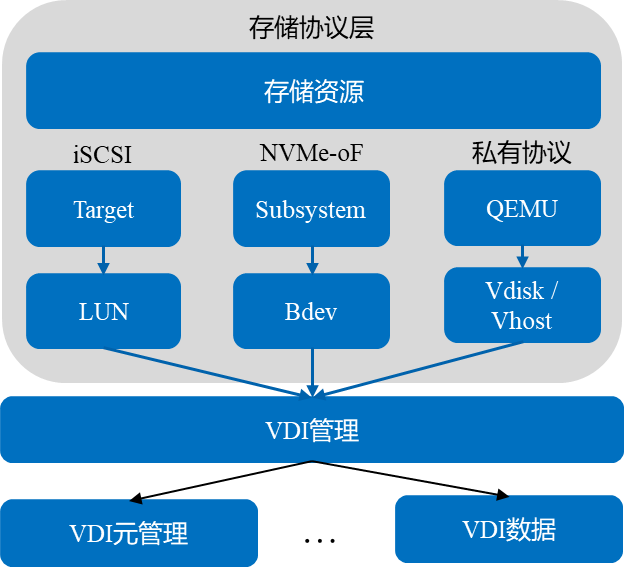


图4.3.1-5统一存储资源管理

#### 存储协议

InCloud dSAN在VDI（Visual Disk Image）基础上添加iSCSI服务，支持iSER访问，添加NVMe-of服务，支持TCP和RDMA，同时探索如和缩短虚拟机到后端存储的路径以降低损耗的技术，包括：

* **协议高可用**：通过集群管理服务实现存储协议接入服务的高可用VIP方案，采用iscsi redirect机制配合VIP高可用方案，达到client端对存储后端故障无感知的目标
* **协议扩展**：在iSCSI基础上基于RDMA编程实现iSER功能，扩展iSCSI协议；参考SPDK 实现NVMe-of的协议扩展，同时支持TCP和RDMA；同时支持iSCSI的PR锁，ATS锁与其他VAAI语义
* **高性能接入协议**：使用 SPDK vhost-user 技术，加速 KVM 虚拟机中 virtio-blk/virtio-scsi 存储设备的 I/O 性能；采用libiscsi和libiser接入方式，缩短QEMU到存储后端的IO路径
* **存储特性扩展**：支持SCSI语义扩展，如PR锁，UNMAP等以支持VAAI存储特性

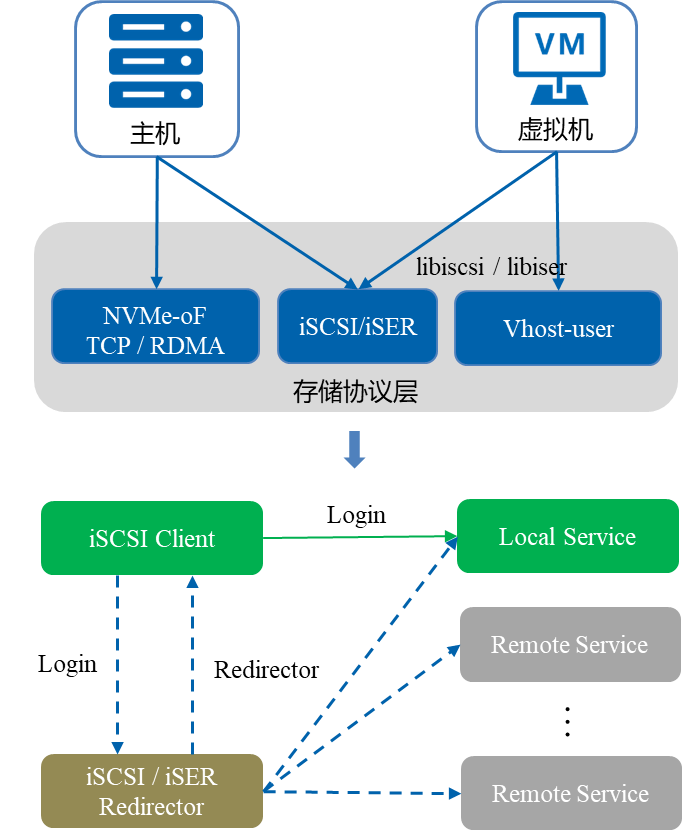


图4.3.1-6高可用、高性能、可扩展存储协议层设计

InCloud dSAN实现了iSCSI高可用方案，达到iSCSI客户端对存储后端存储故障无感知的效果；同时，实现了NVMe-oF协议，并支持TCP和RDMA模式，扩展iSCSI支持VAAI存储特性（Full Copy, Block Zero, ATS, Thin Provisoning Unmap），支持Windows故障转移；通过上述设计，InCloud dSAN达到了高可用、高性能、可扩展的存储协议支持。

#### 元数据

InCloud dSAN采用自研元数据管理方式，将数据分配位置存放在元数据中，IO访问对象时，首先查询元数据中心，之后将IO发往目标位置，由目标节点完成对IO对象的IO操作，提高对象读写IO性能。

元数据管理对象保存在内存中，提高元数据访问效率，降低对IO的性能影响，同时元数据永久保存在SSD磁盘中，加速元数据变更时IO速度，降低元数据管理对性能的影响。

元数据通过多副本方式保留在集群中多个节点上的缓存盘中，从而避免单点故障导致集群不可用。

#### IO路径

InCloud dSAN在基于自研元数据管理机制下，通过对象聚合、IO链路检测与动态切换优化数据对象的随机读性能，包括：

* **IO链路监测**：在随机读场景，如果在本地节点不存在该对象，则会转发副本所在节点，通过监控读转发的请求的节点，通过统计确定LUN的热点访问的节点
* **链路动态切换**：根据链路监测的热点访问节点信息，使用iSCSI协议的redirector机制动态切换session，上层应用无感知

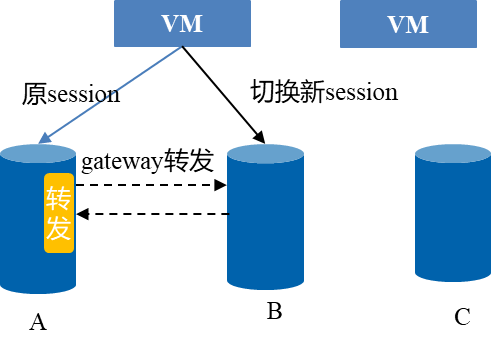


图4.3.1-7 IO链路检测与动态切换

InCloud dSAN通过单Target单LUN为虚拟机提供存储资源的机制，确定虚拟机与iSCSI Target的单session连接，通过副本本地聚合方式，提供IO本地化，同时通过建立智能IO链路监测，动态切换session实现无法聚合场景下随机读性能提升。

#### 存储引擎

InCloud dSAN通过基于裸磁盘管理的单机存储引擎设计方案，避免基于文件系统设计带来的性能开销，同时设计实现类似文件系统操作日志的功能实现掉电场景下的数据可靠性，包括：

* **磁盘布局优化**：将裸磁盘划分为Meta、Journal和Data

1. Meta主要记录data区域的分配，同时记录data读写的log，Meta主要负责数据分配管理，读写类型，ino号，extent，alloc，offset和length等
2. Journal主要用于存储覆盖写场景下上一次数据内容，便于覆盖写不成功后，回滚到上一时刻数据
3. Data区域主要存储数据区域。Data区域以4M为单位切分，以object为维度管理

* **磁盘驱动管理**：HDD/SSD采用kernel\_device管理，NVMe采用SPDK管理，Optane采用pmdk管理

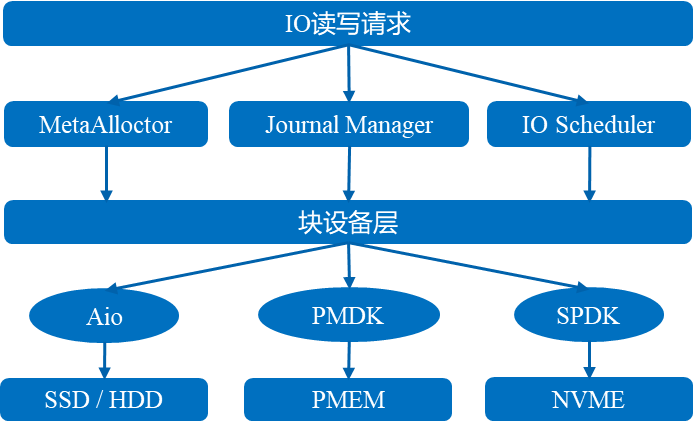


图4.3.1-8基于裸磁盘管理的高性能单机存储引擎设计

#### 物理磁盘管理

InCloud dSAN研究分布式存储系统中物理磁盘管理技术，包括磁盘的挂载，卸载，手动拔出磁盘的处理以及对系统状态的影响，同时研究坏盘、慢盘、坏块与坏道的定义、识别、处理方案，设计了物理磁盘管理设计如下：

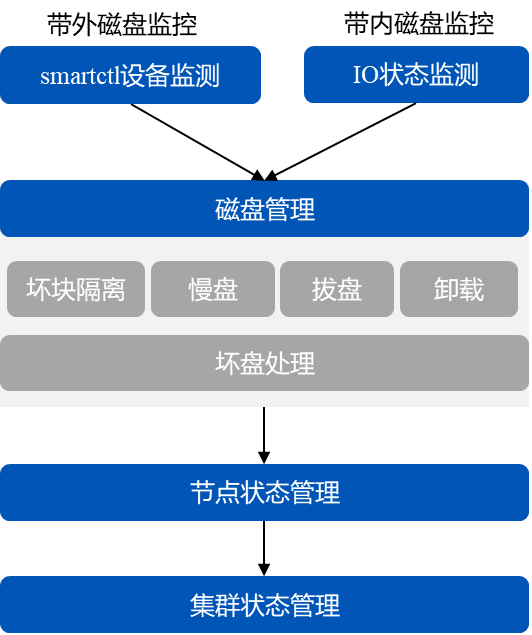


图4.3.1-9物理磁盘管理技术

支持计划内的磁盘运维操作和计划外的异常磁盘处理，包括：

* 磁盘运维
  1. 拔盘：
     1. udev事件：通过捕获系统的udev事件通知来处理拔盘事件，进行盘时间处理
     2. IO驱动：统计一定时间内设备IO错误次数判断磁盘被拔出，进行盘时间处理
  2. 盘漫游：在线磁盘更换插槽，在该过程中不出现数据迁移和恢复
* 磁盘盘异常
  1. 坏盘：统计盘处理IO的中的错误次数，达到定义标准后进行主动踢盘
  2. 慢盘：统计磁盘处理IO的超时次数，达到定义标准后进行主动踢盘
  3. 坏道隔离：通过带内IO探测和带外坏块监控上报坏块，对坏块位置进行隔离和标记，为其分配正常可用内存空间，坏块达到标准后，进入坏盘处理

InCloud dSAN支持自定义坏盘、慢盘标准与处理策略，提前识别和预测到异常磁盘，提升系统的稳定性；同时支持通过磁盘运维指令实现磁盘间的漫游处理以及节点（缓存、容量）盘全部被拔出场景下节点仍具备计算能力。

#### 缓存设计

InCloud dSAN缓存是系统中最主要的性能层，虚拟机的读写操作可以直接访问数据写入SSD硬盘中，完成后立即向上层返回成功，充分发挥缓存层的性能优势；缓存层上已有的数据会逐步下刷至容量层的HDD中，从而保障缓存空间能够加速更多的客户业务。

整个缓存设计基于调优的LRU策略，动态识别IO过程中的热点数据，将最优价值的数据驻留在缓存层中，降低无效扫描导致的缓存淘汰客户有效数据，提高客户业务访问速度。根据时间的迁移与数据的访问频度智能调整缓存对象在冷热数据间的迁移，保证高命中率的同时提高缓存空间利用率

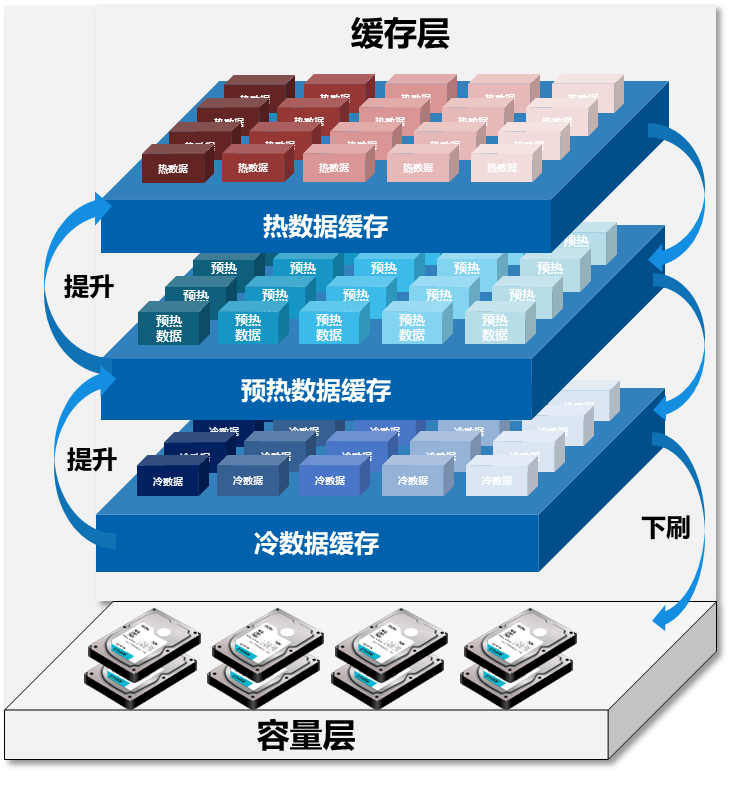


图4.3.1-10智能缓存管理技术

#### 副本策略

副本策略管理包含如下几个部分：

**1）副本分配策略：**为了保证数据的安全，浪潮分布式存储系统支持数据对象的多副本存储，同时采用机架感知和在集群中节点划分不同故障域的方式，将多副本的数据分别存储在集群不同故障域的节点中，避免了因磁盘或单一节点故障等因素造成数据丢失的风险。在副本分配时采用元数据管理，通过副本聚合和容量均衡等策略进行副本分布，根据集群中不同节点的容量和机架感知将不同副本分布在不同节点上，在保证集群中各节点数据分布的均衡性，提高IO的访问效率，降低IO访问时延。

**2）副本迁移策略：**生产集群中的环境会经常发生变化，例如有服务器磁盘故障、服务器故障、集群扩容、虚拟机迁移等等。在出现相关情况时，浪潮分布式存储系统会自动触发检测到相关事件并通知到集群中的各个节点，节点在收到通知后会根据事件类型对副本进行节点内或者跨节点数据迁移，保证了集群中副本数据的分布均衡场景下，发挥扩容节点的能，提高分布式存储的性能。

**3）数据访问策略**：浪潮分布式存储系统采用强一致性的副本策略，保证集群中不同副本内容始终处于相同的状态。在处理写数据请求时，集群必须等待所有副本所在节点返回成功，写的请求才算正常完成。这样在处理读数据请求时，由于写过程中的可靠性和强一致性，所以读取任一副本所在节点的数据即可返回成功。

#### 数据保护

产品支持如下几种数据保护技术：

**1）副本**

不同于硬件RAID技术，通过软件的方式实现多副本的方式来提供安全的保证。当集群发生故障时，可以对数据进行自我修复。

每一个虚拟磁盘都会被划分成以 4MB 为长度的对象（obj）。采用副本技术以后，同一个对象在集群中会有多个副本（拷贝）。同一个对象的多个副本会分布到不同的物理节点上，从而保证即使发生服务器节点宕机，只要还有可以访问到的副本，就不会影响数据的可用性和可靠性。

当一个对象存在多个副本时，为保证数据的一致性，对数据的每一次写都是同时写多份副本，并且当所有副本都写成功后再返回给客户端。

对象的副本数取决于它所从属的虚拟磁盘的存储策略。为保证数据可靠性，InCloud dSAN会主动感知物理环境故障，以确定当前系统中的可以访问的对象。当发现物理环境故障后会主动触发自动修复功能，以确保对象的副本个数满足配置。如用户设置某个虚拟磁盘的副本个数为三份，同一份数据会复制到三个不同的节点上，当其中一个节点故障导致可访问的对象副本个数减少，集群会主动以对象为单位进行数据恢复，从健康的副本所在的节点上读取正确的对象文件并恢复到另外一个节点上，以确保对象始终保持在三副本状态。

对比传统RAID重建时经常导致的存储系统不可用，InCloud Rail在恢复副本的时候正常I/O访问仍然可以被正常执行，并不会出现中断，并且通过控制恢复流量所占的 I/O 带宽保证了正常业务性能被限制在可以接受的范围。

**2）副本策略**

可以通过存储策略，为每一个虚拟磁盘设置不同的副本数。目前支持 2，3 个副本。不同的副本数定义了不同的数据安全级别。当虚拟磁盘的副本数为3时，允许同时有2台服务器发生宕机事件。当虚拟磁盘的副本数为 2 时，允许同时有 1 台服务器发生宕机事件。

### 虚拟化存储池

#### 存储池

由于存储设备的能力、接口协议等差异性，利用InCloud Rail的存储虚拟化技术，可以将存储设备转换为统一管理的存储池，存储池可理解为一种逻辑容器，它以文件或块的形式来存取虚拟机的相关资源，如磁盘镜像、快照和备份等数据。

InCloud Rail支持基于分布式存储的CFS存储池，CFS存储池采用的是集群文件系统，其可以被集群中的多个主机访问使用，为HA和DRS等高级功能的实现提供了支持。

#### 存储池扩容

当InCloud Rail数据存储的使用量达到一定的阈值时，为了不影响系统的性能，需要向系统动态的增加存储的容量，以保证系统的高可用性，即存储扩容。存储的扩容不影响原数据存储的使用，新添加的数据磁盘可立即作为数据存储直接使用。

为了保证存储扩容的安全性，提供了维护模式方案，用户将存储池进入维护模式，中断业务，扩容完成后，退出维护模式自动恢复业务。

#### 存储池柔性卸载

基于存储池柔性卸载特性，当存储设备无法下发IO时，不会引发主机重启，从而保证客户业务在极端情况下不会因为主机重启而中断。当主机连接不到存储设备心跳盘，则该主机会卸载相应的存储池，虚拟机通过HA机制在其他可正常访问存储设备的主机上重启。

#### 存储域通信模式

共享CFS存储池的节点位于同一个存储域中，通过节点间通信完成对共享资源的并发访问控制。存储域支持三种通信模式：网络通信模式、磁盘通信模式和混合通信模式。网络通信模式适用于网络稳定的环境，性能好但对网络波动容忍度差。磁盘通信模式不依赖于网络，但性能要略差于网络通信模式，且对集群规模有一定限制。混合通信模式也称网络/磁盘主备通信模式，优先使用网络通信模式，若遇到网络故障则自动切换到磁盘通信模式。混合通信模式适用于普通网络环境，兼顾网络通信模式性能的同时又能够容忍网络波动，大幅提升CFS存储池的稳定性。

#### 裸设备存储池

InCloud Rail提供了基于小LUN的高性能虚拟磁盘技术方案，通过将存储设备上创建的pool以iSCSI或FC的形式挂载给主机，InCloud Rail能够创建以pool为单位的裸设备存储池。用户可以在裸设备存储池上创建裸设备磁盘，每个裸设备磁盘对应了pool里面的一个LUN。这种类型的虚拟磁盘不需要在存储设备上创建集群文件系统，从而有效降低I/O路径性能损耗。