阿里云

Apsara 存储平台

SA2600&SA3700&SA5700

技术白皮书

文 档 版 本 5.1

发 布 日 期 2021-09-30

尊敬的用户：

衷心感谢您选用阿里云计算有限公司（以下简称阿里云）存储！本存储致力于为您提供符合新数据时代需求的存储产品和解决方案。

本手册用于帮助您更详细地了解和便捷地使用本存储，涉及的截图仅为示例，最终界面请以实际设备显示的界面为准。

由于产品版本升级或其他原因，本手册内容会不定期进行更新，如有变动恕不另行通知。除非另有约定，本手册仅作为使用指导，本手册中的所有陈述、信息和建议不构成任何明示或暗示的担保。

阿里云拥有本手册的版权，保留随时修改本手册的权利。未经阿里云许可，任何单位和个人不得以任何形式复制本手册的内容。

如果您对本手册有任何疑问或建议，请向阿里云计算有限公司垂询。

技术服务电话： 4008013260

地　　址： 浙江省杭州市西湖区转塘科技经济区块 12 号

邮　　编： 310000

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

安全声明

我们非常重视数据安全和隐私，且一如既往地严密关注产品和解决方案的安全性，为您提供更满意的服务。在您正式使用本存储之前，请先阅读以下安全声明。

1. 为了保护您的数据隐私，在调整存储产品用途或淘汰存储设备时，请您将存储软件恢复固件出厂设置、删除信息、清除日志。 同时，建议采用第三方安全擦除工具对存储系统软件所在的系统盘进行全面安全擦除。
2. 您购买的存储产品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）。因此，您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。
3. 如需获取存储系统开源软件声明，请直接联系我们的客户服务人员。
4. 存储系统的某些安全特性需要您自行配置，如认证、传输加密、存储数据加密等，这些配置操作可能会对存储系统的性能和使用方便性造成一定影响。您可以根据应用环境，权衡是否进行安全特性配置。
5. 存储系统自带了部分用于生产、装备、返厂检测维修的接口、命令及定位故障的高级命令，如使用不当，可能会导致设备异常或者业务中断，不建议您自行使用。如需使用，请联系我们的客户服务人员。
6. 我们已全面建立产品安全漏洞应急和处理机制，确保第一时间处理产品安全问题。若您在存储产品使用过程中发现任何安全问题，或者寻求有关产品安全漏洞的必要支持，请直接联系我们的客户服务人员。

以上声明中，“我们”指代阿里云计算有限公司；阿里云计算有限公司拥有对以上声明的最终解释权。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

目录

安全声明 ii
1 存储系统架构 1
1.1 统一存储平台 1
1.2 软件组成 1
1.3 集群存储 6
1.4 智能 NAS 7
1.5 NPIV 技术 11
1.6 端口使用说明 12
2 业务定制、弹性存储 14
2.1 存储池 14
2.2 智能异构虚拟化 19
2.3 智能数据迁移 21
3 按需分配、优化工作负载 24
3.1 智能分层 24
3.2 智能精简 27
3.3 智能在线压缩 30
3.4 智能缓存加速 35
3.5 智能缓存分区 38
3.6 智能服务保障 39
4 数据保护、业务连续 41
4.1 智能卷镜像 41
4.2 智能本地复制 43
4.3 智能远程复制 49
4.4 智能双活 54
4.5 智能云分层 59
4.6 智能数据销毁 63
4.7 NAS 快照 65
5 灵活高效的主机插件 67
5.1 SCOM 插件 67
5.2 VSS 插件 68
5.3 InPath-多路径驱动 69
5.4 SRA 73

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

5.5 VCP 75
5.6 Cinder 驱动 76
5.7 Cinder-backup 驱动 77
5.8 K8sCSIPlugin 78
5.9 Manila 驱动 79
6 人性化管理、维护 81
6.1 统一智能存储管理平台 81
6.2 事件监控和维护向导 82
6.3 STAT 工具 82
6.4 Comprestimator 工具 84
6.5 遵循 SMI-S 规范的 CIM 接口 84
6.6 SNMP 接口 85
6.7 RESTful 接口 86
7 缩略语 87

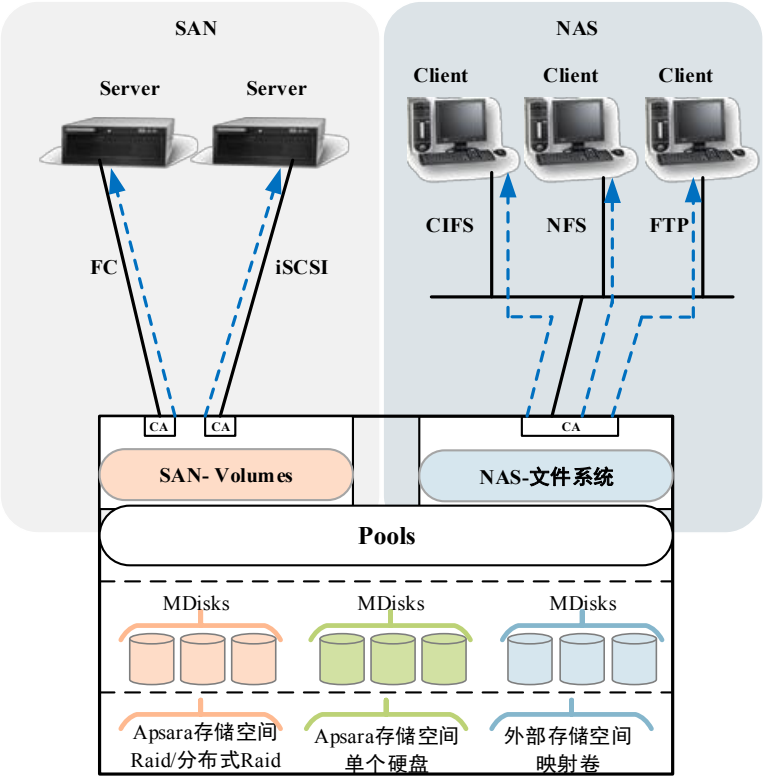
阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

1 存储系统架构

1.1 统一存储平台

本存储系统提供统一的 SAN 、NAS 特性，满足用户数据集中存储以及文件共享的需求，在统一平台上实现多种协议的接入，为企业提供多种存储解决方案，帮助客户灵活部署。存储平台的统一融合架构如图 1-1 所示，存储池是由一个或多个 MDisk 组成的，一个或多个 MDisk 被划分为相同大小的 extent 数据块存放在存储池中的，文件系统和逻辑卷是由存储池中划分出来的 extent 数据块组成的。在这个架构中，文件系统和逻辑卷都是直接与底层存储池交互。 文件系统和逻辑卷各自保持独立，互不影响。

图 1-1 统一融合架构



1.2 软件组成

本存储系统软件能够管理存储设备和存储的数据，协助应用服务器进行数据操作。存储平台提供的软件包括主机插件、MCS 核心软件和系统管理控制平台。通过相互配合，智能、高效、经济地实现各种存储、备份和

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

容灾功能。

存储平台核心软件说明，如表 1-1 所示。

表 1-1 存储平台核心软件

功能模块	功能描述
分布式 RAID（DRAID），也称智能 RAID（InRaid）	分布式 RAID（DRAID）将传统 RAID 的热备磁盘空间分散到 DRAID 的所有成员盘中，是没有独立的热备磁盘的，其中热备磁盘空间称为热备空间。若磁盘出现故障，故障盘里的数据会重新被放置在此 DRAID 的热备空间上，同时，DRAID 中的所有成员盘都会参与到重建工作中，消除了单片磁盘写入的瓶颈，大大提高了重建速度。
存储池	存储池是 MDisks 的一个集合，卷容量的分组。池中所有的 MDisks 被划分为相同大小的数据块。将之后从存储池中划分出的卷映射到主机系统。
集群管理	一个集群系统包含 2 到 8 个控制器。在集群系统中，所有的配置、监听和服务都由系统层承担。集群系统提供了一种将系统配置数据备份到磁盘的方法，以便在灾难事件中恢复集群系统。
NAS	NAS 功能实现文件级共享，不再需要网关设备，使得应用服务器可以简单方便地访问共享的文件。文件系统具备如下特点： 支持多协议访问 支持配额设置 支持文件系统扩容 支持权限控制
智能异构虚拟化	智能异构虚拟化是指一种在逻辑上对存储资源进行抽象，不会因为物理设备的局限性而受到限制，能达到消除不同类型资源之间的差异化的效果，让管理者认为所管理的资源都是来源于同一个资源池设备，并且提供更加丰富和先进的高级功能特性。
智能分层	智能分层功能，能够将不同热度的数据集自动地、智能地分布在不同的存储层，这样就可以优化系统的工作负载。该功能支持多种类型硬盘之间的数据分层。
智能精简	智能精简配置是一项新的容量分配技术，不会一次性分配足够大的存储空间给应用程序，而是根据该项应用实际所需要的容量，分多次、每次少量的分配给应用程序，当该项应用所产生的数据增长，分配的容量空间已不够时，存储系统会再次从后端存储池中补充分配一部分存储空间。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

智能在线压缩	智能在线压缩是基于压缩引擎采用无损数据压缩算法实现的，压缩引擎能够动态地对数据进行压缩，在数据写到存储端时，整个压缩过程是透明进行的，即在主机端是感受不到这个压缩过程的。智能在线压缩也是一种在线实时压缩，在数据写入磁盘之前就已经完成了压缩，同样的数据信息，实际使用的存储容量大大减少。能够很好地满足企业对性能、可靠性以及可扩展性的要求。
智能缓存加速	利用 SSD 的快速读能力，将系统中常用的数据置于 SSD 内，从而满足客户端的快速读取需求。通过 SSD 缓存池对 SSD 资源进行虚拟化管理，已添加 SSD 资源的缓存池，可以创建 SSD 缓存分区，实现 SSD 缓存资源隔离，SSD 缓存分区有两种工作模式， Single 模式和 Double 模式。可以根据业务需求选择不同的工作模式实现单个缓存分区加速单个卷或加速多个卷。
智能缓存分区	通过为关键业务分配私有缓存资源的手段，保障关键业务的性能表现。同时提供了非关键业务的缓存资源共享机制，以提升缓存资源利用率。
智能服务保障	通过重新分配存储系统资源用来满足某些应用程序的特定性能要求，从而保障关键业务的服务质量。
智能卷镜像	镜像卷是从存储池中划分出来并供给主机端使用的逻辑卷，是基于常见卷镜像技术实现的，能让一份镜像卷源卷（主机端正在使用的卷）保持有两份复制物理卷，并且这两份复制卷可以属于不同的存储池或者不同的物理存储系统（如 Apsara 存储自身的磁盘或者后端异构存储阵列），保证方案高可用。
智能本地复制	智能本地复制功能是同一 I/O Group 内的复制功能，是基于时间点的复制，采用写时复制（Copy On Write）原理，可以保证数据在线可用的状态下，创建一份即时可用的一致性拷贝。
NAS 快照	NAS 快照功能是基于智能本地复制功能的，针对 NAS 文件系统，以整个 NAS 文件系统为单位，创建和回滚快照。当共享数据丢失、损坏时，可通过回滚该文件系统的快照立即切换回快照点时的状态，找回丢失数据。
智能远程复制	数据的远程复制是发生在两台 Apsara 存储中容量大小相同的逻辑卷之间，生产中心的主设备作为源卷端，备份设备作为目标卷端。源卷包含了主机端写入的生产数据，并且承担正常的业务应用数据；目标卷包含对源卷的备份数据，并保持与源卷数据的同步更新，用于灾难后的数据恢复。智能远程复制包括同步远程复制、异步远程复制和有变更卷的异步远程复制。
智能双活	智能双活的功能是基于集群存储平台实现的，由两台存储同时管理每个逻辑卷。功能的实现使用了同步远程复制、加入变更卷的异步远程复制和智

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

	能数据迁移的技术。允许业务连续性的硬件故障、设备断电、控制器间通信连接失败、自然灾害（如火灾、水灾）等场景，在灾难发生时保证了业务的连续性。
智能云分层	此功能是指将本地存储平台的卷数据备份到云存储中。本地存储平台支持创建到云服务提供商的连接，以将卷数据副本存储在私有云存储或公有云存储中。
智能数据销毁	当不再需要某个卷时，可以对该卷进行智能数据销毁。销毁单位是卷，智能数据销毁使用的是数据覆写的方式。
事件监控	在存储管理界面中，提供了对存储状态的在线监控，并能将存储平台的重要健康信息及时的反馈给管理员，能够随时反馈各项指标，直观提示，实时监控当前存储平台的运行情况。
DMP 维护向导	包含故障诊断和故障修复的维护步骤，针对存储设备发生的软硬件故障，用户可以根据此步骤和存储智能化的故障恢复机制，进行故障诊断和修复。降低了维护成本，节约了故障修复时间，快速高效的解决故障并使业务恢复正常状态。
统一智能存储管理平台	提供对存储产品的集中管理功能，方便用户同时管理监控多套分散的存储设备。统一管理和监控多个集群的状态变化，容量使用情况，历史性能统计信息，3DC 以及告警信息等。还可以通过内嵌的 iframe 框架登录到单个集群的管理界面，从而实现在一套软件中对多套存储集群进行监控和管理

应用主机插件说明，如表 1-2 所示。

表 1-2 应用主机插件

软件集合	软件名称	功能描述
Microsoft	VSS 插件	在 Windows Server 操作系统上运行，作为微软 Volume Shadow Copy Service（VSS）的被调用者，使 Windows 的卷影拷贝（VSS）功能，能借助 Apsara 存储的本地复制功能。
	SCOM	作为微软 System Center Operation Manager（SCOM）的插件，可以实现在 SCOM 界面上可以对 Apsara 存储进行监控。
VMware	SRA	SRA（Storage Replication Adapter）是 VMware SRM (Site Recovery Manager)的一个适配器，使得 SRM 能够调用 Apsara 存储的智能远程复制功能。
	VCP	包含 3 个部分：配置界面、vCenter Web Client Extension 、后

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

		台。配置界面用于配置 VMware vCenter 和 Apsara 存储的连接信息；vWC Extension 用在 VMware vCenter 的管理页面上，扩展出一些配置用来操作 Apsara 存储的选项和界面。
多路径驱动	InPath for AIX	是 IBM AIX 操作系统平台的多路径驱动。
	InPath for Windows	是 Windows Server 操作系统平台的多路径驱动。
OpenStack	Cinder 驱动	是 OpenStack 中 Cinder 模块的驱动，可以提高 Cinder 的性能。Apsara 的 Cinder 驱动跟随 OpenStack 发布，并且存在于 Cinder 模块的代码中。
	Cinder-backup 驱动	Cinder-backup 主机插件是 OpenStack 系统中 cinder 服务中负责备份的驱动程序，实现了对 cinder 卷进行备份以及备份管理的功能。功能包括对 cinder 卷创建备份、删除备份、利用备份恢复 cinder 卷数据。
	Manila 驱动	Manila 主机插件是 OpenStack 系统中 Manila 服务的驱动程序，实现了通过 Manila 服务对存储进行管理。
容器	K8sCSIPlugin	K8sCSIPlugin 使得本存储可以为 Kubernetes 集群中的应用自动的提供持久化存储。K8sCSIPlugin 的功能主要包含卷的创建、删除、挂载、卸载、克隆，快照的创建、删除。

说明：其他系统的多路径软件请使用系统自带的多路径，也能够很好的兼容 Apsara 存储。

系统管理控制平台说明，如表 1-3 所示。

表 1-3 系统管理控制平台

软件名称	功能描述
GUI	存储平台的图形用户界面，可以通过技术员端口或者管理网口（使用浏览器）登录和使用存储平台的 GUI 界面，实现对存储平台的有效管理和配置。
CLI	存储平台的命令行界面，可以使用第三方终端软件通过串口或者管理网口（使用 SSH 协议）登录和使用存储平台的 CLI 界面，实现对存储平台的有效管理和配置。
服务助手	服务助手是为专业的维护人员提供管理配置的接口，专业人员可以通过访问 GUI 界面，对存储平台进行维护。不建议客户在没有专业指导的情况下使用。
SSH	存储平台支持 SSH 协议，为远程登录会话和其他网络服务提供安全性的协议，可以有效防止远程管理过程中的信息泄露问题。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

SMI-S Server	存储平台支持 SMI-S 协议。用户可以通过支持 SMI-S 协议的第三方网管软件对存储平台进行管理和配置。支持 SMI-S 协议的网管软件种类很多，用户可自行选用。
WebServer	Web 服务器可以解析 HTTP 协议，存储平台支持 HTTP 协议。用户可以通过支持 HTTP 协议的网管软件对存储平台进行管理和配置。
CIM 接口	CIM 客户端通过调用存储提供的 CIM 接口，实现对存储的监控和管理，包含集群管理、主机管理、存储资源管理（池、卷、硬盘）、性能监控、数据保护等功能。
SNMP 接口	本系列存储提供了遵循 SNMP 规范的自定义接口供第三方系统或软件调用，用于管理或监控存储设备。
RESTful 接口	本系列存储提供了符合 REST 约束的接口供第三方系统或软件调用，用于管理或监控存储设备。

1.3 集群存储

Apsara 存储通常是以单个双控存储连接多个扩展柜的模式给前端主机业务提供存储服务，两个控制器都位于同一个机箱内。同时，还可以支持横向扩展的集群存储，集群内的控制器在物理上彼此分离，一个集群系统可以提供多个控制器，详情请参考《产品白皮书》。

为了方便管理多个控制器，系统给集群设置了管理 IP 地址，使用这个 IP 地址可以管理集群内的所有控制器。集群的管理 IP 地址，只能指定到唯一的控制器，如果此控制器发生故障，系统会自动选择一个新的控制器接管集群的管理 IP 地址。

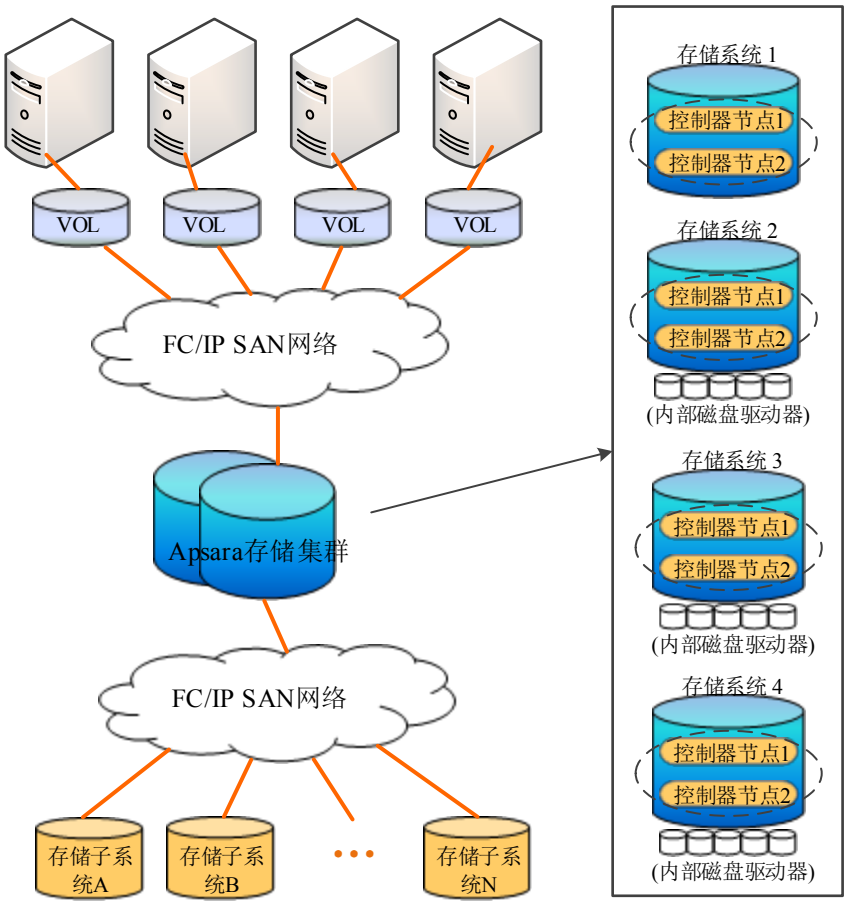
在集群系统中，所有的配置、监听和服务都由系统层承担，配置数据被复制到所有的控制器。集群系统提供了一种将系统配置数据备份到磁盘的方法，以便在灾难事件中恢复故障控制器的集群配置信息。此方法不支持应用数据，只支持存储平台的配置数据。

集群系统给主机应用配置提供了一个连续的可用性平台，即在存储平台任何单一的故障出现时主机都可以保持不间断的访问。在故障期间，无故障控制器会接管与相关应用数据有关的逻辑卷，并保持联机状态，给主机应用提供服务。详情见存储双活功能。

集群存储应用的拓扑图，如图 1-2 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 1-2 集群存储应用拓扑图



说明：右侧的存储平台 1~4，可以带有内部磁盘驱动器，也可以不带，此处仅为举例。

1.4 智能 NAS

Apsara 存储提供的 NAS 功能，实现文件级共享，不再需要网关设备，使得应用服务器可以简单方便地访问共享的文件。

文件系统具备如下特点：

- 支持多协议访问

  存储系统支持应用服务器通过 NFS、CIFS、FTP 协议访问共享的文件。
- 支持配额设置

  存储系统可以设置用户和用户组的可用容量限制，便于对存储空间进行合理的分配与利用。
- 支持文件系统扩容

  文件系统支持在线扩容，不影响系统正常使用，便于用户管理。
- 支持权限控制

  文件系统可以控制共享用户和用户组的访问权限和读写权限，便于用户管理与维护，保证数据的安全。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- 支持文件系统快照

  文件系统支持快照功能，并提供 CIFS 共享和 Windows 系统的卷影复制（Volume Shadow Copy）功能。
- 支持文件系统容灾

  支持创建文件系统的容灾关系，用于主站点故障时备站点可以继续对外提供 NAS 服务（目前仅支持 MiniOSS 服务配置），确保业务的连续性及数据的安全性。
- 支持高性能文件系统创建

  最大支持创建 1PB 的高性能文件系统，可以显著提升大文件的顺序读写性能。

Apsara 存储通过文件网关连接本地存储与云端对象存储，并通过标准的 NFS/CIFS 文件协议实现两个存储系统间数据的无缝交互。用户使用 NFS/CIFS 协议存放和浏览云端对象存储系统中的对象，并将 OSS 中的对象作为 NFS/CIFS 挂载点中的文件进行访问。

约束与限制

- 目前仅支持双控。
- SA2600 仅每控 32GB 和 64GB 内存规格支持 NAS 功能。
- 必须创建 NAS 业务 IP 才能使用 NAS 服务，包括 NFS、CIFS、域服务。
- 配置 NAS 业务 IP 和云服务 IP 必须双控对称，即如果 Node1 的端口 1 配置了 IP，那么 Node2 的端口 1 也必须配置相同网段的 IP。
- 如果配置了白名单，需要将 NAS 的业务、云服务、域、DNS 等 IP 加入白名单。
- 性能监控支持 5 分钟数据显示。
- 文件系统扩容时有容量限制，即文件系统扩容后的容量应小于等于原始容量（第一次创建文件系统的容量）的 10 倍。
- 仅 SA3700&SA5700 支持 NAS 快照功能。
- 文件系统快照业务容量规划建议，因快照的使用容量随着源文件系统写入数据量的增长而增长，请务必在存储池预留足够的容量。若存储池剩余容量小于文件系统容量，快照可能会自动失效。

文件网关

- 建议在上下行带宽 50Mb 以上的情况下使用文件网关功能。
- 元数据存储的容量大于数据存储容量的 1%。
- 数据存储最小 32GB。
- 缓存盘最小 320MB。
- 为了保证 MiniOSS 有更好的性能，建议在 SA3700（单控 128G、256G 内存）平台上调核后使用该功能。
- MiniOSS 支持在系统升级时的 HA 操作。
- Minioss 与 NFS，CIFS 不能同时提供服务，否则可能会由于 CPU 负载过高导致 HA 切换失败。
- 在系统电源及 BBU 同时失效的情况下，MiniOSS 和云文件系统不支持 T3 恢复。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

特性间相互关系

针对 SA3700，NAS 功能和压缩功能不能同时使用。

技术规格

表 1-4 智能 NAS 使用规格

规格	参数
集群内支持 NAS 功能的控制器数	2（来源于 1 个控制柜的 2 个控制器）
最大文件系统数量	SA2600：4 个 SA3700&SA5700：8 个
单个文件系统最小容量	100MB
单个普通文件系统最大容量	SA3700：512TB SA2600&SA5700：256TB
单个高性能文件系统最大容量	1PB
单个文件系统最大文件数	20 亿
单文件最大容量	255TB
每目录最大子目录数	3000 万
最大 SMB 共享数	SA2600：8 个 SA3700&SA5700：16 个
最大 CIFS 共享数	
最大 NFS 共享数	
每控制器 CIFS 最大连接数	512
每控制器 NFS 最大连接数	1024
CIFS 共享的权限控制	只读、读写
NFS 共享的权限控制	只读、读写、限制性读写
最大本地用户和用户组总数	500 个（SA2600） 1000 个（SA3700&SA5700）
单个文件系统最大快照数量	8

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

最大快照数量	16 说明：NAS 快照数量统计在智能本地复制数量内。
最小定时快照间隔周期	5 分钟

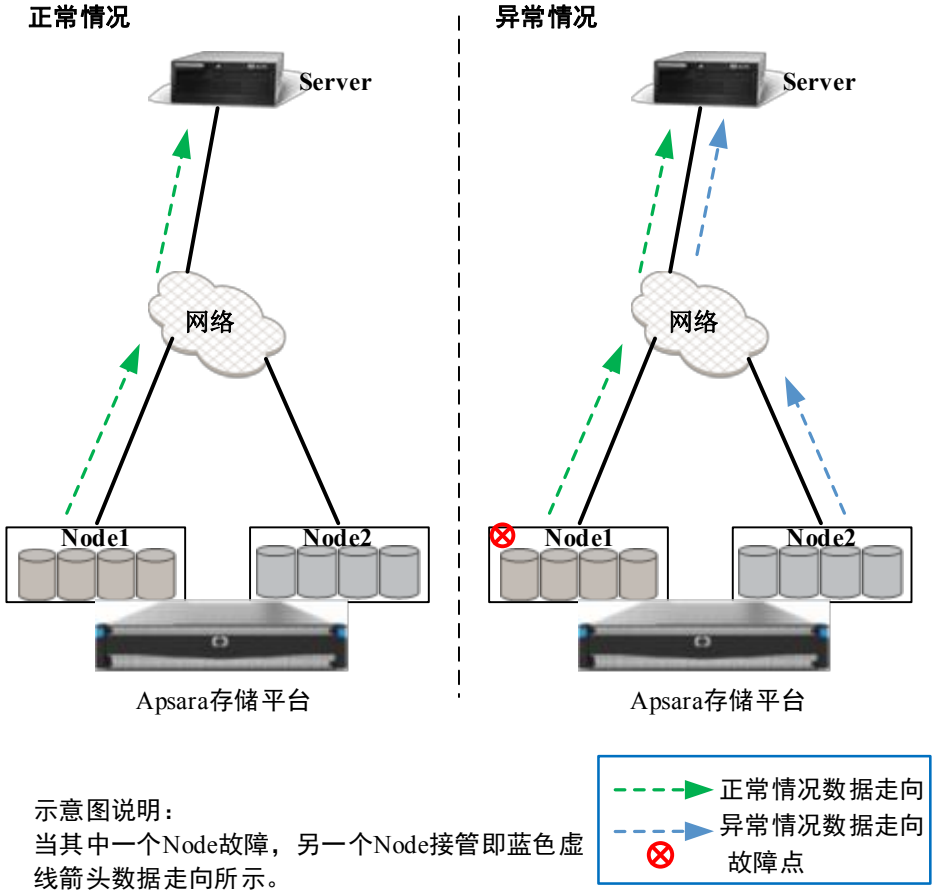
license 许可约束

此功能不受 license 控制，在默认状态下即可正常使用。

技术价值

NAS 存储具有高可用性，保证单点故障不影响业务运行。NAS 集群中一个 Node 故障时，其业务被另一个 Node 接管，共享访问的 IP 地址无需变化，无需重新挂载共享目录。若 Node 发生故障时，存在文件正在读写，则读写过程会中断，只需重新发起读写请求即可；NFSv3 挂载时读写数据不中断。

图 1-3 单点故障下业务持续运行示意图



应用场景

- 适用于企业、政府部门、分支机构等对应用数据整合和数据共享的需求以及跨异构主机平台不同业务应

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

用的需求，以实现对前端所有服务器的一个集中存储和管理。

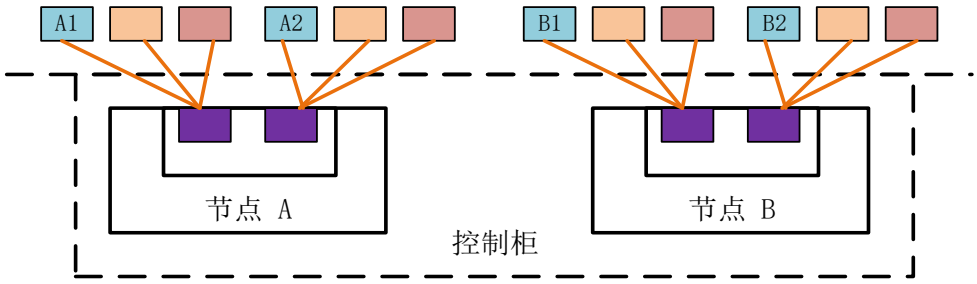
- 适用于办公内网文件共享、流媒体、视频编辑、监控、教育科研等行业应用的文件存储、文件备份。

1.5 NPIV 技术

Apsara 的存储平台支持 NPIV 技术，允许单个 HBA 卡（称为 N_Port）注册多个 WWPNs （全球唯一端口名）和 N_port ID 号码。这使得存储平台在只有一个物理端口的情况下，可以虚拟出多个端口，若要实现与其他设备的互通，需要要求交换机必须同样支持 NPIV 技术。

Apsara 的 NPIV 技术可以将一个 FC 物理端口虚拟出 3 个端口，都拥有各自的 WWPN 用于与其他设备连接。Apsara 存储的 NPIV 技术如图 1-4 所示。

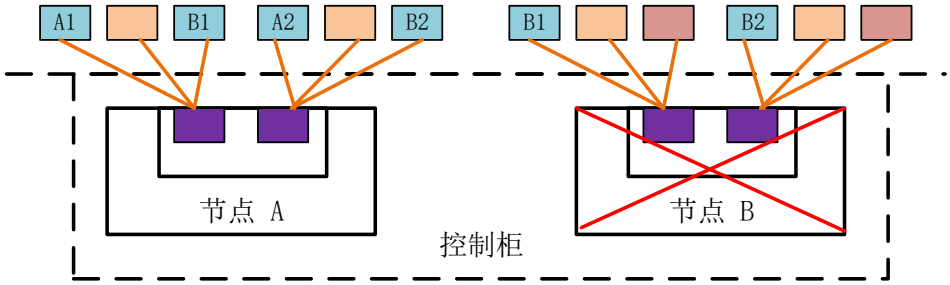
图 1-4 FC 端口的 NPIV 技术



3 个虚拟端口可以应用在以下几个方面。

- 用于连接应用主机，作为 Target 端口。
- 用于连接第三方存储，作为 Initiator 端口。
- 用于集群内部 I/O Group 之间的通信。
- 用于 NVMe over FC 场景，使用不同的虚拟端口承载 NVMe 协议与 SCSI 协议，减少不同业务相互影响。
- 用于 failover 场景，两个控制器相应的 failover 端口具有相同的 WWPN，在发生故障时应用可以无缝切换，原先主机端发往控制器 B 上 B1 和 B2 口的业务将会切换到控制器 A 上，可以替代多路径功能。如图 1-5 所示。

图 1-5 failover 切换



阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

1.6 端口使用说明

本存储系统支持通过 iSCSI 协议进行通信与管理。系统中每个服务使用的网络管理方式、协议类型、端口号、IPv4 和 IPv6 的支持情况信息。

表 1-5 端口使用说明

端口号	协议类型	服务名称	网络管理方式	支持情况		
				仅使用 IPv4	仅使用 IPv6	混合使用 IPv4 和 IPv6
21	TCP	用于访问 FTP 服务	Inbound	Yes	Yes	Yes
22	TCP	用于命令行界面 (CLI) 访问的 SSH	Inbound	Yes	Yes	Yes
25	TCP	电子邮件 (SMTP) 通知和库存报告	Outbound	Yes	Yes	Yes
68	UDP	IPv4 DHCP (节点服务地址)	Outbound	Yes	Yes	Yes
80	TCP	用于 GUI 访问的 HTTP 到 HTTPS 重定向	Inbound	Yes	Yes	Yes
111	TCP	RPCBIND 服务	Inbound	Yes	Yes	Yes
111	UDP	RPCBIND 服务	Inbound	Yes	Yes	Yes
123	UDP	网络时间服务器 (NTP)	Outbound	Yes	Yes	Yes
135	TCP	用于访问 CIFS 服务	Inbound	Yes	Yes	Yes
139	TCP	用于访问 CIFS 服务	Inbound	Yes	Yes	Yes
445	TCP	用于访问 CIFS 服务	Inbound	Yes	Yes	Yes
161	UDP	SNMP 事件通知	Outbound	Yes	Yes	Yes
162	UDP	SNMP 事件通知	Outbound	Yes	Yes	Yes
427	TCP	CIMOM SLPD	Inbound	Yes	Yes	Yes
427	UDP	CIMOM SLPD	Inbound	Yes	Yes	Yes
443	TCP	用于 GUI 访问的 HTTPS 重定向	Inbound	Yes	Yes	Yes

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

514	UDP	Syslog 事件通知	Outbound	Yes	Yes	Yes
547	UDP	IPv6 DHCP (节点服务地址)	Outbound	Yes	Yes	Yes
564	TCP	用于访问 NFS 服务	Inbound	Yes	Yes	Yes
750	UDP	RPCBIND 服务	Inbound	Yes	Yes	Yes
1024	TCP	用于访问 CIFS 服务	Inbound	Yes	Yes	Yes
1025	TCP	用于访问 CIFS 服务	Inbound	Yes	Yes	Yes
1260	TCP	IP 定额功能数据收发端口	Inbound	Yes	Yes	Yes
2049	TCP	RPCBIND 服务	Inbound	Yes	Yes	Yes
3260	TCP	iSCSI	Inbound	Yes	Yes	Yes
3260	TCP	iSCSI iSNS	Outbound	Yes	Yes	Yes
3260	TCP	IP 伙伴关系管理 IP 通信	Inbound/Outbound	Yes	Yes	No
3265	TCP	IP 伙伴关系数据路径连接	Inbound/Outbound	Yes	Yes	No
5989	TCP	CIMOM (HTTPS)	Inbound	Yes	Yes	Yes
8080	TCP	用于 GUI 访问的 HTTP 到 HTTPS 重定向	Inbound	Yes	Yes	Yes
8443	TCP	用于 GUI 访问的 HTTPS	Inbound	Yes	Yes	Yes
8444	TCP	T 口登录 http 服务代理端口	Inbound	Yes	Yes	Yes
8750	TCP	RPCBIND 服务	Inbound	Yes	Yes	Yes
16310	TCP	远程用户认证服务 - HTTP	Outbound	Yes	Yes	Yes
16311	TCP	远程用户认证服务 - HTTPS	Outbound	Yes	Yes	Yes
24007	TCP	RPCBIND 服务	Inbound	Yes	Yes	Yes
33241	TCP	RPCBIND 服务	Inbound	Yes	Yes	Yes
44907	TCP	RPCBIND 服务	Inbound	Yes	Yes	Yes
49152	TCP	RPCBIND 服务	Inbound	Yes	Yes	Yes

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

2 业务定制、弹性存储

2.1 存储池

功能定义

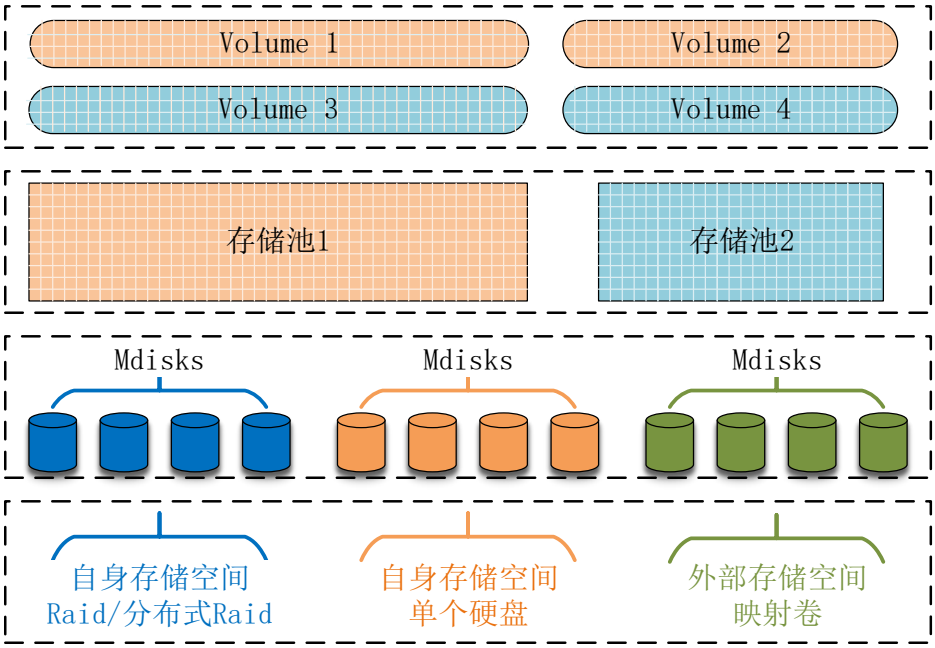
存储池是由一个或多个 MDisk 组成的，一个或多个 MDisk 被划分为相同大小的 extent 数据块存放在存储池中的，卷是由存储池中划分出来的 extent 数据块组成的。

Apsara 存储，由自身存储的硬盘（控制柜和扩展柜的硬盘空间）组合成不同类型的 RAID 或分布式 RAID，每一个 RAID 或者分布式 RAID 就是一个 MDisk；自身存储的一块硬盘也可以组成一个 MDisk。

Apsara 存储，支持外部异构存储虚拟化，详细内容请参考 2.2 章节。从外部存储映射过来的每一个卷，在 Apsara 存储本地均形成一个 MDisk。

存储池的组成示意图如图 2-1 所示。

图 2-1 存储池组成示意图



原理描述

这部分主要讲解分布式 RAID 的实现原理，其他相关内容已是业内非常成熟的技术，在此文中不再讲解。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

传统的 RAID 方式，数据分散在 RAID 组的成员盘中，但热备空间是由独立的磁盘构成的。若 RAID 组中的一块磁盘出现故障，通过 RAID 的校验机制，故障盘里的数据会从其他成员盘读取，并重新被放置在热备磁盘上。

在分布式 RAID 的概念中，热备磁盘是被包含在 RAID 组中的，称为热备空间。热备空间分布在 RAID 中的所有成员盘上的，这意味着包含在 RAID 组中的所有驱动器都有热备空间。若 RAID 组中的一块磁盘出现故障，通过 RAID 的校验机制，故障盘里的数据会从其他成员盘读取，并重新被放置在此 RAID 组的热备空间（在成员盘中）上。

分布式 RAID，不仅热备空间，还有数据和与其相对应的奇偶校验信息均是分布在所有成员盘上。这意味着创建 DRAID 时可以选择更多的硬盘驱动器，如由 5 个驱动器组成的传统 RAID5 分布在 40 个驱动器上（DRAID5 由 40 个驱动器组成）。

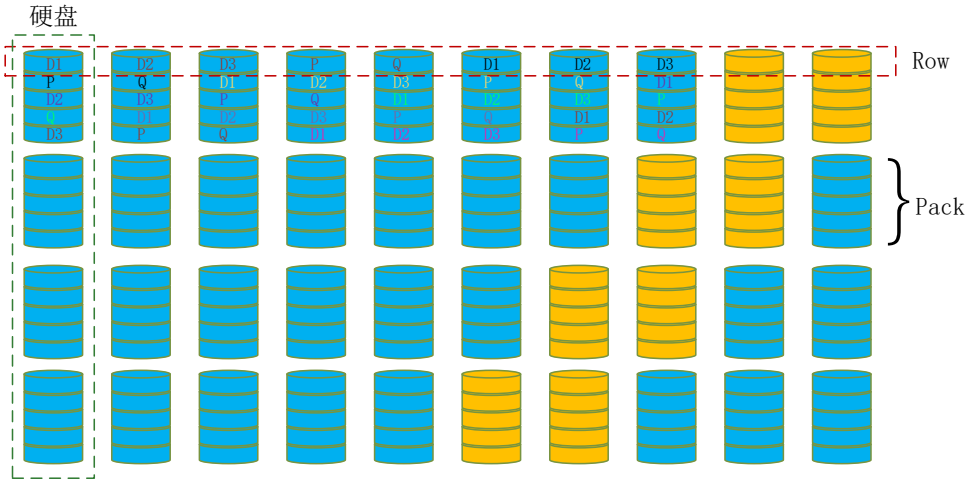
分布式 RAID5 和分布式 RAID6 的热备空间分布取决于 Pack，在数组中的所有驱动器都会被分为 Row 和 Pack。

Row 的宽度是组成这个分布式 RAID 的所有驱动器数量，如图 2-2 中所示 Row 的宽度是 10（DRAID6 由 10 个驱动器组成，由 5 个驱动器组成的传统 RAID6 分布在 10 个驱动器上）。

Pack 是由连续的 Row 组成的，Pack 的数量取决于内部传统 RAID 的条带数量，如图 2-2 所示，内部传统 RAID6 的条带数量是 5（3+P+Q），所以 Pack 值是 5，Pack 值在分布式 RAID 中是恒定不变的。

图 2-2 介绍了分布式 RAID6 的原理。这个分布式 RAID6 的内部是由 5（3+P+Q）个驱动器组成的传统 RAID6 和 2 个热备区（2 个磁盘容量的热备空间）组成的。

图 2-2 分布式的 RAID6



技术规格

存储池使用规格，如表 2-1 所示。

表 2-1 存储池使用规格

属性	限制

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

系统支持 MDisk 数量	4096
单存储池支持的 MDisk 数量	128
系统可管理的容量	32PB
单存储池可管理的容量	32PB
系统支持的存储池数量	1024
系统支持的逻辑卷数量	10000

说明：

- 单个存储池支持的逻辑卷数量不受限制，与系统保持一致。
- 不单独说明 I/O Group 支持数量的，统一与系统（集群）保持一致。
- 以上参数是支持的最大数量，多个参数无法同时满足，需要并行考虑。

传统 RAID 使用的规格，如表 2-2 所示。

表 2-2 传统 RAID 使用规格

属性	数值
系统最大 RAID 数	128
RAID 最大成员盘数	16
RAID-0 成员盘数	1-8
RAID-1 成员盘数	2-2
RAID-5 成员盘数	3-16
RAID-6 成员盘数	5-16
RAID-10 成员盘数	2-16

分布式 RAID 使用的规格，如表 2-3 所示。

表 2-3 分布式 RAID 使用规格

属性	数值
集群最大 DRAID 数	32
I/O Group 最大 DRAID 数	10
DRAID 最大成员盘数	128

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

DRAID-5 成员盘数	4-128
DRAID-6 成员盘数	6-128
单个 DRAID 热备盘数	1-4
DRAID-5 条带宽度	3-16
DRAID-6 条带宽度	5-16
DRAID-5 成员盘最大容量	<8TB

分布式 RAID 推荐使用配置，如表 2-4 所示。

表 2-4 分布式 RAID 推荐使用配置

属性	数值
阵列宽度（DRAID 成员盘数）	48-60
阵列宽度小于 36 的热备盘数	1
阵列宽度 36-72 的热备盘数	2
阵列宽度 73-100 的热备盘数	3
阵列宽度 101-128 的热备盘数	4

存储池中数据块的大小影响普通卷的最大容量，对应关系如表 2-5 所示。

表 2-5 块大小与普通卷容量对应关系

块大小（MB）	普通卷 最大容量（TB）
16	2
32	4
64	8
128	16
256	32
512	64
1024	128
2048	256
4096	512

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

8192	512

存储池中数据块的大小影响 MDisk 的最大容量，对应关系如表 2-6 所示。

表 2-6 块大小与 MDisk 容量对应关系

存储池 块大小（MB）	传统 RAID MDisk 最大容量（TB）	DRAID MDisk 最大容量（TB）
16	2	32
32	4	64
64	8	128
128	16	256
256	32	512
512	64	1024
1024	128	2048
2048	256	4096
4096	512	8192
8192	1024	16384

特性间相互关系

存储池是存储最基本的功能，是使用其他高级特性的基础，因此与其他高级特性之间不存在冲突。

License 许可约束

不受 license 控制的项有：传统 RAID、分布式 RAID、MDisk、存储池、逻辑卷，在默认状态下在可用范围内正常使用。

技术价值

- 形成卷的数据块可以来自于不同的 MDisk，提高读写 I/O 的性能，增加可靠性。
- DRAID 内没有独立空闲的热备磁盘，如有硬盘故障，则所有成员盘都参与到重建工作中，消除单片磁盘写入瓶颈，大大提高了重建速度，降低再次出现磁盘故障时 RAID 失效风险。
- DRAID 所有的驱动器都参与 I/O 操作，没有空闲的驱动器，驱动器利用率和性能都有提升。
- DRAID 以热备块形式分布在所有盘，所有盘都在工作。
- 组成 DRAID 的硬盘数量高达 128 块，能够满足不同的性能需求。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

应用场景

- 大规模硬盘应用场景

  使用 DRAID 功能，单个 I/O Group 支持的硬盘数量高达 10*128=1280 块（不考虑物理限制）；传统 RAID 单个 I/O Group 支持的硬盘数量为 16*48=768 块。因此 DRAID 功能可以使用在硬盘数量较多的场景下。

- 性能、可靠性要求较高的场景

  根据 DRAID 功能的特性价值，存储系统有更高的可靠性和更高的性能。满足客户的不同需求。

2.2 智能异构虚拟化

功能定义

智能异构虚拟化能够在逻辑上对存储资源进行抽象，消除不同类型存储资源之间的差异，实现资源整合和利旧。整合后的存储设备可以拥有与前端存储同样丰富和先进的高级功能，管理者仍然像管理一个存储设备一样同时管理所有存储设备，降低管理成本。

原理描述

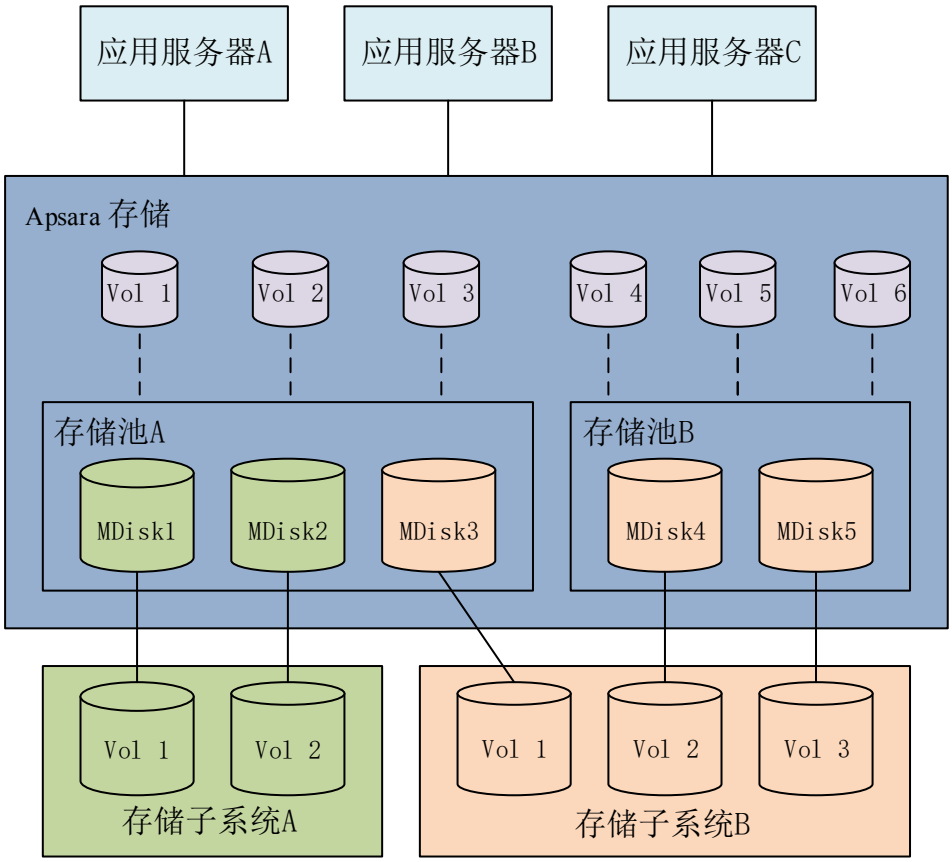
Apsara 存储的异构虚拟化功能是块级别的虚拟化。存储子系统（被虚拟化的外部存储平台）可以将 RAID 阵列分离出存储空间---卷，通过内部的 SAN 网络映射给本存储系统。本存储系统将存储子系统映射过来的卷，形成一个 MDisk，然后，本存储系统会将一个或多个 MDisk 组成虚拟资源池，叫存储池，再将存储池进行划分形成逻辑卷（Volume）映射给前端主机，供其使用。

前端的应用主机只能看到逻辑存在的卷，如果是在 Windows 系统平台下，通过磁盘管理器就能查找到，再进行简单的格式化和新建分区就可以像本机的本地硬盘一样使用了。如果要对卷进行扩容缩减、镜像或者做快照等功能，这些操作都是在 Apsara 存储的虚拟层中进行的，不需要涉及主机资源，也不需要依赖后端虚拟的存储子系统。并且本存储系统可以实现高级功能的传递，即后端存储子系统原本没有的高级功能，存储空间通过 Apsara 存储接管到存储池以后，可以将本存储系统的高级功能传递到该部分存储空间上使用。

智能异构虚拟化的示意图，如图 2-3 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 2-3 块级别异构虚拟化



技术规格

本存储系统与后端存储可以通过 FC 网络进行连接。每个集群系统最大可以支持 4096 个从 FC 协议后端存储映射过来的卷。

特性间相互关系

智能异构虚拟化功能，与其他高级特性没有冲突，在使用此功能的同时，可以结合使用其他所有的存储功能：智能数据迁移、智能分层、智能精简、智能在线压缩、智能卷镜像、本地复制、智能远程复制、集群管理、智能双活等。

License 许可约束

智能异构虚拟化功能受 License 控制，只能使用 License 激活。默认为未激活状态，不能使用智能异构虚拟化功能。使用 license 激活后，支持把第三方存储的卷添加到存储池使用。注销 license 后，已添加的外部存储还可以正常使用，但不能再添加新的外部存储，使用 License 重新激活后才能继续添加外部存储。

技术价值

本存储系统能够虚拟市场上 90%以上的存储阵列，使存储设备易整合、易管理、易使用，满足诸多应用需求。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- 消除平台差异性，增强数据流动性。异构虚拟化能有效解决异构存储环境下的诸多问题，允许企业灵活地按照实际应用需求选择最适合的厂商及存储设备。打破原独立系统的“信息孤岛”，提高了整体存储设备的利用率，提升数据价值。
- 整合设备空间，简化存储管理，提高工作效率。通过重新规划原有旧存储设备，将分散的空闲容量空间集中起来进行管理使用。原有的数据资源也可以通过本存储系统直接映射给前端主机，消除异构阵列之间的差异，无论是块服务还是文件服务都可以在统一的管理界面中进行操作管理。
- 实现数据迁移，有效地保证业务连续性。存储系统虚拟化后的卷可以实现在线迁移，数据迁移在底层的子系统之间完成，对前端主机和应用程序完全透明，不会影响业务的正常运行。在未来扩容升级换代及做数据保护操作时，都不会影响业务。
- 提高数据中心的资源利用率，有效降低存储设备的管理成本。原有存储系统可以使用本存储系统的高级功能，如智能在线压缩、智能精简等，提高存储池中容量空间的利用率。同时，简化存储管理工作，极大降低管理的复杂性，降低管理成本。
- 提升系统性能。虚拟出的卷是在多个阵列及控制器上进行条带化划分的，能够提供额外的缓存容量，能极大地提升系统性能。

应用场景

- 异构设备较多

  数据中心的存储环境复杂，有许多来自不同厂家和不同型号的存储设备，想要实现空间统一、管理统一的标准。就可以通过 Apsara 存储的虚拟化功能来实现空间的虚拟整合和不同设备间的统一管理。
- 实现高级功能

  数据中心中的设备较多，购买时间不一致，对高级功能的需求随着时间的推移和业务的变化有不同的需求，导致一些设备缺少需求的高级特性。就可以通过 Apsara 存储的虚拟化功能来整合异构的存储设备，并且使异构设备的空间拥有与 Apsara 存储同样丰富和先进的高级功能特性。

2.3 智能数据迁移

功能定义

智能数据迁移技术可以实现一个 I/O Group 内的数据迁移，包括同一存储内和不同存储池之间的数据迁移，也可以实现跨 I/O Group 的数据迁移。数据迁移过程中，存储系统将重建数据的存储位置信息，在数据迁移过程完成后，主机端应用直接连接到新的数据卷。

原理描述

智能数据迁移是以逻辑卷为单位，手动进行的数据迁移，数据迁移过程中不会影响前端业务。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

本存储系统支持以下三种级别的数据迁移：

- 可以将卷迁移到同一个存储池内。
- 可以将卷迁移到同一 I/O Group 的另一个存储池内。
- 可以将卷迁移到一个集群内的另一个 I/O Group。

同一个 I/O Group 内手动的数据迁移，是在存储平台内运行的，对前端业务完全透明。当数据迁移过程完成以后，数据块的位置会重新指向新的位置，不涉及主机访问的链路问题。而实现不同 I/O Group 之间的数据迁移，在迁移之前，需要保证主机能够同时访问两个 I/O Group 里的卷，并且此卷需要两个 I/O Group 能同时访问。迁移过程主要分为以下几个步骤。

1. 源端将缓存中的数据刷写到硬盘，此过程中主机端正在进行的所有的 IO 都是 write-through 模式，直接将数据写入硬盘。
2. 增加目标端的 I/O Group 对源卷的访问权限。
3. 建立新的路径使主机端可以访问迁移的目标卷。
4. 执行卷迁移过程。
5. 迁移完成后，删除源端 I/O Group 对卷的访问权限。

特性间相互关系

普通卷、精简卷、压缩卷、镜像卷（某一个拷贝卷或两个拷贝卷）都可以进行手动的数据迁移。

智能本地复制与智能远程复制的源卷（主卷）和目标卷（从卷），与数据迁移的关系，请参考 4.3 章节。

License 许可约束

此功能不受 license 控制，在默认状态下即可正常使用。

技术价值

- 自动触发的数据迁移，有效保证了存储设备的高可靠性。
- 数据迁移可以改变卷的存放位置，能够随时满足前端业务对性能的要求。
- 迁移过程中无需停机，对后端存储子系统的环境完全透明，对主机系统和应用均不造成影响。
- 将多个异构设备的数据，迁移到同一个存储设备上。

应用场景

- 平衡负载、提高 IO 性能

  客户现场有多种应用环境，每个应用对存储平台的性能要求都不完全一样，由于前期因素的影响，导致对性能要求较高的业务部署在性能较低的存储池中，用户可以通过数据迁移功能，将主机业务迁移到性能较高的存储池中。

- 异构存储之间数据迁移

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

当客户现有环境中存在两家不同品牌厂商的存储设备，客户想将 A 存储设备上的数据迁移到 B 中，但是这两家的存储设备不能互相通信，这时候可以通过 Apsara 的异构虚拟化和数据迁移功能，轻松实现数据的迁移。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

3 按需分配、优化工作负载

3.1 智能分层

功能定义

智能分层功能是将不同热度的数据块自动地、智能地分布在不同的存储层，存储池、卷都可以实现智能分层功能，以 extent 数据块为单位，实现数据在不同存储层间的浮动，并且不会损坏数据的完整性。

原理描述

一般情况下，存储平台是在卷的层面对 I/O 操作进行监控，而卷是处于存储层的内部。因此如果需要对 I/O 操作进行统计，如果不使用自动算法，而是手动地将它们移动到合适的存储层，并且还要能够反应工作负载的变化，是很难实现的过程。本存储系统对内部空间的 RAID/DRAID 或者外部存储的 MDisks 中的 SSD 和 HDD 成员盘是区别对待的。按照硬盘类型，存储池支持多类硬盘的任意组合的数据分层，如表 3-1 所示。

表 3-1 智能分层支持组合列表

组合	Tier0_Flash	Tier1_Flash	Tiersata_Flash	Tier_Enterprise	Tier_NearLine
组合 1	1	-	-	-	-
组合 2	1	2	-	-	-
组合 3	1	2	2	-	-
组合 4	1	2	-	3	-
组合 5	1	2	-	-	3
组合 6	1	2	2	3	-
组合 7	1	2	2	-	3
组合 8	1	2	-	2	3
组合 9	1	2	2	2	3
组合 10	1	-	2	-	-

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

组合 11	1	-	2	3	-
组合 12	1	-	2	-	3
组合 13	1	-	2	2	3
组合 14	1	-	-	2	-
组合 15	1	-	-	2	3
组合 16	1	-	-	-	2
组合 17	-	2	-	-	-
组合 18	-	2	2	-	-
组合 19	-	2	2	3	-
组合 20	-	2	2	-	3
组合 21	-	1	1	2	3
组合 22	-	2	-	3	-
组合 23	-	1	-	2	3
组合 24	-	2	-	-	3
组合 25	-	-	2	-	-
组合 26	-	-	2	3	-
组合 27	-	-	1	2	3
组合 28	-	-	2	-	3
组合 29	-	-	-	2	-
组合 30	-	-	-	2	3
组合 31	-	-	-	-	3

说明:

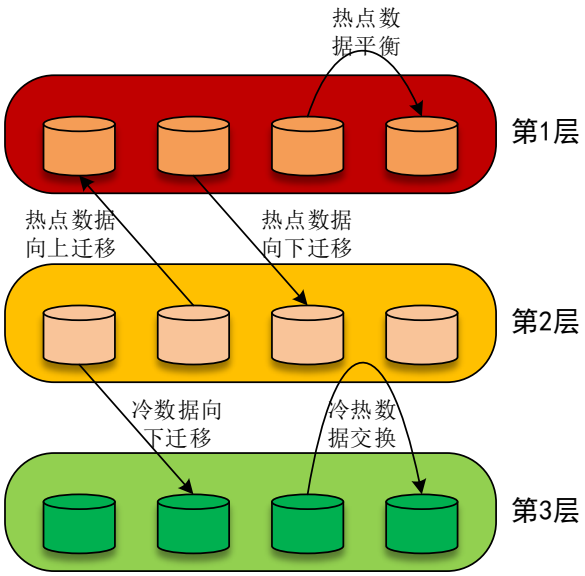
- 表格内容的数字（1,2,3）表示在存储池中位于的存储层。
- Tier0_Flash 代表普通 SAS SSD 盘，来自于自身存储空间或外部存储空间。
- Tier1_Flash 代表 DWPD 值较低的 SAS SSD 盘，主要用于读密集型应用，来自于自身存储空间或外部存储空间。
- Tiersata_Flash 代表普通 SATA SSD 盘，来自于外部存储空间。
- Tier_Enterprise 代表 10K 或 15K rpm 的 SAS 硬盘，来自于自身存储空间或外部存储空间。
- Tier_NearLine 代表 7200 rpm 的 NL-SAS 硬盘，来自于自身存储空间或外部存储空间。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

数据初次写入存储时，落入的存储层优先级排序为 2>3>1。其中，第二层或第三层可能包含多种类型的盘，在包含多种类型盘的组合中盘的优先级遵循以下顺序： Tier0_Flash > Tier1_Flash > Tiersata_Flash > Tier_Enterprise > Tier_NearLine。

智能分层对热点数据 extent 数据块进行移动的示意图，如图 3-1 所示。

图 3-1 智能分层



智能分层可以看作一种对性能优化的功能，这项功能可以自动将卷或池中的 extent 数据块在不同的存储层之间进行迁移。以 24 小时为一个周期（24 小时是默认的时长），智能分层会对开启了这项功能的卷或池进行 I/O 操作的活动性和 extent 的延迟程度进行监控，基于这份性能分析日志，分层算法会生产一个迁移计划，动态地将活跃度高的热点数据块迁移到高性能的存储层。反过来，将冷数据下移也是同样的逆向过程。

智能分层可以开启加速模式，加速模式是指根据性能分析日志得出迁移计划，实现数据块更快速度的迁移。加速模式并不适合每天开启，在需要时、系统活动较少时开启，因为加速模式需要更高的工作负载支持完成，迁移结束后，应及时关闭加速模式。

特性间相互关系

智能分层功能是在存储池这一层实现的，对存储池的 extents 进行分析与迁移，适用于所有类型的卷，并且与其他高级特性之间不存在冲突。

license 许可约束

智能分层功能受 license 控制，支持试用激活和使用 license 激活。默认为可试用状态，不能使用智能分层功能。试用激活和使用 license 激活支持对存储池开启智能分层功能，启动数据分析和热点数据迁移。试用激活时间为 90 天，使用 license 激活是永久激活。试用到期后进入试用失效状态，无法继续使用智能分层功能。对于已开启智能分层功能的存储池，不能进行数据分析和数据迁移，使用 License 重新激活后才能启动数据

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

分析和数据迁移。注销与试用失效的处理方法一致。

技术价值

- 可以实现存储池均衡，在同一个存储池里，可以通过评估 extent 自动地将其均衡到所有的 MDisk 中去，平衡整个池中相同性能层级的数据块。
- 可以优化系统的工作负载，提高 SSD 硬盘的使用率。
- 尽可能地使当前环境下的存储性能达到最优。
- 可以减小热点数据区域的 I/O 延迟。

应用场景

数据分层主要应用在磁盘中某一块区域的数据有规律地读/写活动。例如：

- 音频行业，热门影视剧的观看。
- 铁路售票系统，经常查询余票数量。
- 多类型应用环境，可以自动优化不同应用程序所产生数据的存放位置，使存储达到最优的使用效果。

3.2 智能精简

功能定义

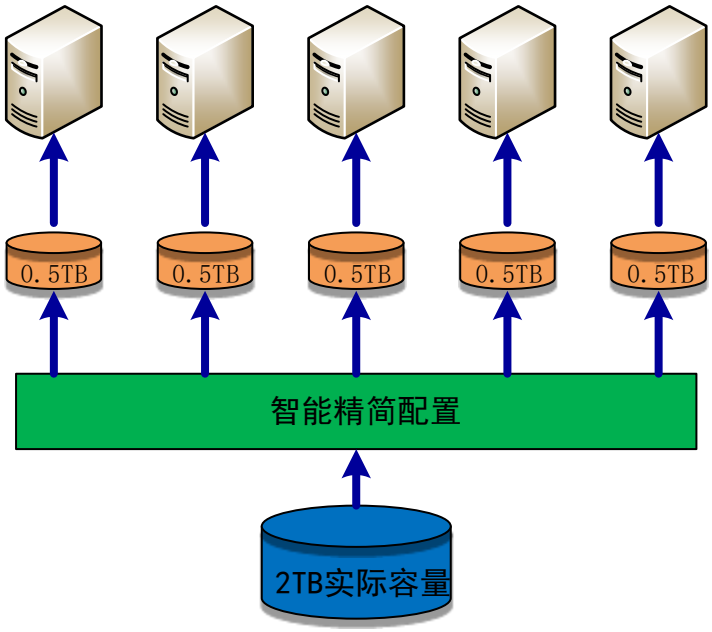
智能精简配置是一项新的容量分配的技术，不会一次性分配足够大的存储空间给应用程序，而是根据该项应用实际所需要的容量，分多次、每次少量的分配给应用程序，当该项应用所产生的数据增长，分配的容量空间已不够的时候，存储平台会再次从后端存储池中补充分配一部分存储空间。

原理描述

使用智能精简功能，存储平台可以提供比可用的实际容量更多的容量给前端的应用主机，Apsara 存储可以通过 FC 或者 iSCSI 提供预分配的卷。如图 3-2 是智能精简的一个实例：一个存储平台可用的实际存储容量为 2TB，但是在 5 个主机端，每个主机端却接收到存储映射来的 0.5TB 的容量。也就是实际只有 2TB 容量空间的存储，在主机端可以看到 2.5TB 的可用容量空间。管理员只需监控系统实际容量的使用情况，在真的需要实际空间时，再去增加空间容量。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 3-2 智能精简配置实例



Apsara 存储的卷能够配置成自动精简卷和足额卷，自动精简卷包含了实际的和虚拟的容量空间。初始创建精简卷时，默认分配 2%的实际物理空间，其他 98%是虚拟给主机端的，没有分配物理空间。实际空间有一小部分容量用来存放初始化源数据，也可以设置从 0%开始。当写操作请求，写入的数据按照固定大小的实际空间（颗粒度）写入精简卷，颗粒度可以在创建卷的时候设定，分别有 32KB、64KB、128KB、256KB 几种颗粒度可根据不同应用要求供用户选择，越小的颗粒度可以节省越多的空间，但是相对的也会产生更大的检索表，不同的应用场景对性能的影响也不一样，所以用户建议根据自身应用需求，合理设置颗粒度。如果需要对精简卷做快照时，要求精简卷和快照拥有同样大小的颗粒度。

自动精简卷扩容有两种操作：自动扩容和非自动扩容，并且这两种模式可以随时更变。如果选择自动扩容，存储平台便会根据实际需求自动地增加固定大小的实际空间给该精简卷，可以支持实际存放容量大于当初配置的精简卷容量，如图 3-2 可以超出映射给主机的 0.5T。自动扩容精简卷会维持一定的实际空间空闲，防止用户存放数据量与精简卷实际容量不匹配时，造成意外错误。对于非自动扩容的精简卷，没有实际空间空闲，当实际需求的数据量大于该卷配置的存储容量时，该卷便会立即掉线。对于这类情况，可以通过设定告警阈值，当需求容量超过此阈值，由存储平台提供的监控机制，如邮件或 SNMP 对用户进行告警。

技术规格

自动精简卷在使用时的规格说明如表 3-2 所示。

表 3-2 自动精简卷使用规格

属性	数值
系统支持的精简卷数量	10000

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

单个精简卷最大容量	SA3700：512TB SA2600&SA5700：256TB
单个精简卷最小容量	512B
是否支持扩容/缩容	是

存储池中 extent 数据块的大小影响精简卷的最大容量，对应关系如表 3-3 所示。

表 3-3 块大小与精简卷容量对应关系

存储池 块大小（MB）	精简卷 最大虚拟容量（GB）
16	2000
32	4000
64	8000
128	16000
256	32000
512	65000
1024	130000
2048	260000
4096	520000
8192	524288

精简卷按照写入操作，以固定的颗粒度增加实际空间，颗粒度的大小影响最大的虚拟容量，对应关系如表 3-4 所示。

表 3-4 颗粒度与最大虚拟容量对应关系

精简卷 颗粒度（KiB）	精简卷 最大虚拟容量（GB）
32	260000
64	520000
128	1040000
256	2080000

特性间相互关系

- 自动精简卷可以应用在智能本地复制、智能远程复制、智能双活场景。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- 自动精简卷和普通卷在使用上是完全一样的，可以创建镜像，可以实现数据迁移。
- 自动精简卷可以转化成普通卷、压缩卷。

license 许可约束

此功能不受 license 控制，在默认状态下即可正常使用。

技术价值

- 智能精简是优化可用存储容量利用率的一种方法，这种方法能够消除几乎所有的空闲空间，有效地避免低利用率的情况出现。
- 自动精简卷还有一个特点，就是能根据不同应用需求，通过智能卷镜像功能实现精简卷与足额卷（精简卷是实际用到多少空间就分配多少空间给用户，足额卷是初始分配时就将容量一次性分配给用户）之间无误差的转换。比如，用户可以对一个足额的生产卷复制一份精简卷的副本，当同步完成以后，便可以将生产卷中的数据删掉。
- 当客户真正需要更多的存储容量时，才去扩容购买更多的硬盘，这样就有效的降低了前期对存储设备的投入成本，控制预算。

应用场景

- 云存储空间分配

  现在流行的云存储空间，已经广泛应用到每一个人。并且云存储空间的容量分配的比较大，对于实际的应用而言，大多数人们使用的仅仅是一小部分的空间。这种情况下，提供的云空间可以是自动精简卷，按照用户的实际空间使用量去分配空间，大大降低了云存储设备对硬盘数量的要求。

- 合理利用存储空间

  对于应用较多的环境下，在最初的空间规划时，往往规划的不是那么精确。那么用户可以使用智能精简配置，给应用分配足够的虚拟空间，根据应用程序对实际容量的需求去分配物理空间，大大提高了空间利用率，并降低了前期规划不合理的风险。

3.3 智能在线压缩

功能定义

智能在线压缩功能只适用于 Apsara SA3700&SA5700 存储平台。考虑成本和散热问题，本存储支持两种基于硬件的压缩方式：内置 QAT 压缩和外置压缩卡压缩。

- SA3700：系统主板不带 QAT，使用外置压缩卡，该卡插在控制器的 PCIe slot2 或 slot3 插槽。压缩卡遵循了半长、半高 PCIe 卡的设计。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

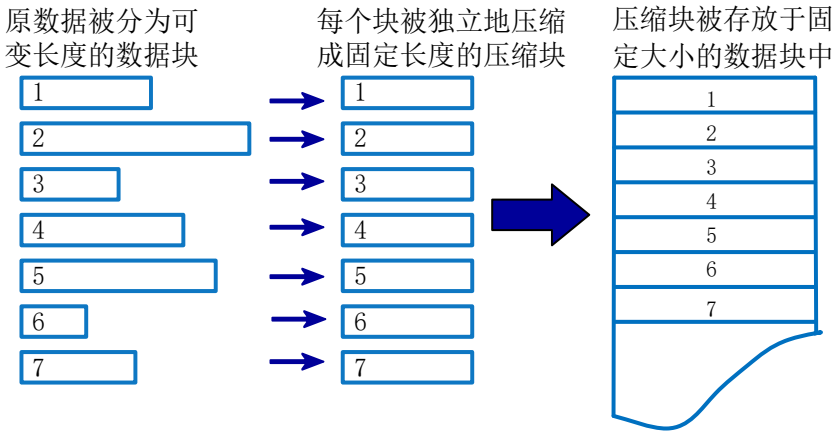
- SA5700：系统主板自带 QAT 压缩引擎，可以为存储提供压缩功能，降低压缩成本的同时，压缩设备的吞吐量、压缩性能提升；支持 3 个 QAT 引擎。

原理描述

实时数据压缩技术是基于随机访问压缩引擎（RACE），这个引擎是集成在存储平台里面的，它不会改变系统数据的属性，并且兼容以前压缩卷的各项特性。存储平台的实时压缩技术基于这种压缩引擎采用无损数据压缩算法，能够动态地对数据进行压缩，压缩是一种在线实时压缩，在数据写入磁盘之前就已经完成了压缩，在数据写到存储平台时，整个压缩过程是透明进行的，主机端是感受不到这个压缩过程的。

随机访问压缩引擎提供一种大小可变的数据块，让原数据写入这些数据块，压缩处理后输出大小固定的压缩块。其中压缩后的数据是以大小固定的数据块存放在存储设备中的，这种方式提供了一种高效统一的方式与途径去检索被压缩的数据。随机访问压缩示意图，如图 3-3 所示。

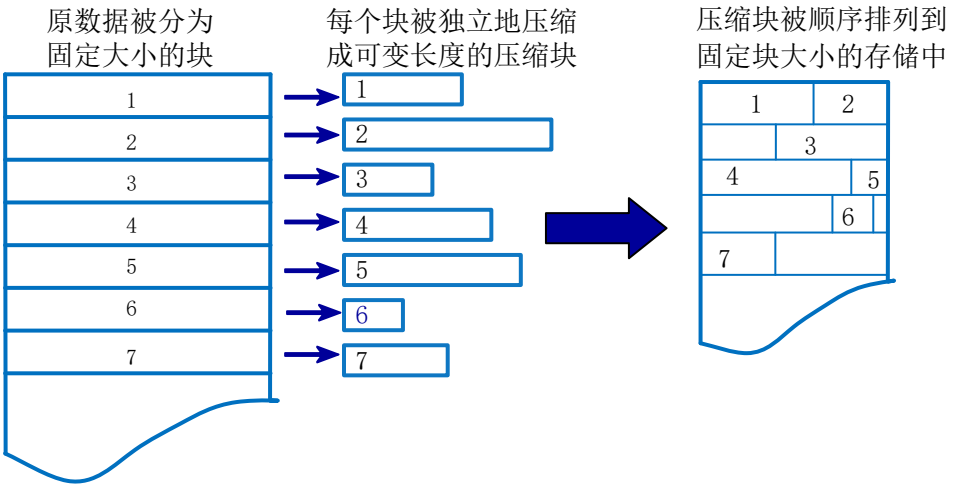
图 3-3 随机访问压缩方式



传统压缩一般有应用比较广泛的压缩工具诸如 zip、gzip 等，写进来的数据会被分离到一些固定的组块中，这些块的压缩和提取都是各个独立的。这就会带来问题，当压缩块中有数据需要更新到相应组块中时，整个块的容量都会被选中，提取和重压缩带来了繁重的 I/O 压力，如果能够实现只选择其中变化数据的一小块压缩块来提取和重压缩，那么需要压缩的比率和负担都会减少。传统压缩方式的使用性能是比较低的，缺陷是不能实现真正的数据随机访问，如图 3-4 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 3-4 传统压缩方式



本存储系统的数据压缩技术具有如下优势和特点：

- 预判断机制

  对于某些数据块，它本身的压缩度已经很高了，如果还是与其他数据块一样进行压缩，不仅同样地消耗了 CPU 和缓存资源，还只能节约出来一小部分的空间。因此，为了避免在处理这些数据块时浪费资源，本存储系统引入了一种更高效的压缩算法，预判断机制。当数据块的可压缩比率低于设定的压缩比率时，压缩引擎便会跳过该数据块，以达到节约 CPU 和缓存资源的目的。压缩算法控制着数据块的压缩是否跳过，当然，一个数据块如果被判定达不到压缩要求，同样可以被压缩，只需要对其进行标记然后再压缩。

- 基于时间压缩

  本存储系统的随机访问压缩引擎提供了一种基于时间压缩的技术。当主机有数据向压缩引擎写数据时，这些数据就被压缩，同时向一个大小固定的组块中填充，这个组块叫做可压缩块。当同一时间同时写入多路数据流时，所有数据可被压缩后集中存储到这个单一的可压缩块中，对应的循环检索表也被存储在这个块中。因为这些写入的数据流大部分都是来源于同一个应用程序，并且数据类型也相同，压缩算法可以方便地检测更多的重复数据。这种压缩方式，是在数据被同一时间写入同一个可压缩块时进行的重复数据检索并压缩，因此叫做基于时间压缩，这种压缩由于让压缩的数据均匀地写到一个可压缩块中，因此可以拥有更高的压缩比率，减少检索表数量。

- 压缩过程透明

  压缩引擎集成在存储系统里面，它不会改变系统数据的属性，只会改变数据存放，并且兼容以前压缩卷的各项特性。基于这种压缩引擎，存储系统的实时压缩技术采用无损数据压缩算法，能够动态地对数据进行压缩，很好地满足企业对性能、可靠性以及可扩展性的要求。在数据写到存储系统时，整个压缩过程是透明的，主机端感受不到这个压缩过程。

- 多线程绑核机制

  RACE 采用多线程绑核机制，来保证能够发挥出压缩设备的最大性能。RACE 分为两个进程 A 和 B，每个

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

进程独占两个 CPU 物理核。每个进程按照压缩引擎的数量创建最少 6 个线程（保证每个压缩引擎至少 3 个线程）。然后 RACE 把线程和压缩设备建立多对一的对应关系，RACE 的每个线程都指向一个压缩设备，同时 RACE 的几个线程绑定一个固定的 CPU 核（避免竞争，造成性能波动）。每当一个 I/O 请求下发到压缩模块时，RACE 根据合理的选择策略选线程来进行数据的压缩和解压。

- 采用更高性能的压缩芯片

  QAT 硬件压缩功能采用更高性能的 LBG 系列芯片。LBG 系列芯片在处理 I/O 请求时的性能（吞吐量、IOPS 以及延时）都有较大的提升。在压缩算法上，通过对压缩算法的优化，能够提供更高的压缩比性能。

技术规格

压缩卷在存储平台使用过程中的技术规格如表 3-5 所示。

表 3-5 压缩卷使用的规格

属性	数值
I/O Group 支持的压缩卷数量	缓存 32G：支持 200 个 缓存 64G/128G/256G：支持 512 个
系统支持的压缩卷数量	2048
单个压缩卷最大容量	96T
单个压缩卷最小容量	512B
是否支持扩容/缩容	是

特性间相互关系

- 针对 SA3700，NAS 功能和压缩功能不能同时使用。
- 压缩卷可以应用在智能本地复制、智能远程复制、智能双活场景。
- 压缩卷可以创建镜像，可以实现数据迁移。
- 压缩卷可以转化成普通卷、自动精简卷。

License 许可约束

智能在线压缩功能受 license 控制，此功能的工作状态包含未激活、试用激活、使用 license 激活、试用失效、注销。默认情况下是未激活状态，这个状态不支持使用压缩卷。试用激活和使用 license 激活支持使用压缩卷，包括对压缩卷的配置与管理。试用激活时间为 90 天，使用 license 激活是永久激活。试用到期后进入试用失效状态，已创建的压缩卷支持删除、读写、迁移到其他池，不支持创建新卷、卷拷贝、卷复制、卷扩容、卷缩容、转换卷为压缩卷等操作。注销与试用失效的处理方法一致。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

技术价值

本存储系统的压缩技术在压缩比率、检索表的数量、读写性能上都有大幅的提升和优化能够节约 2 倍以上，提升各项业务性能，有效降低了 TCO 成本。

- 提高 I/O 性能

  实时压缩技术可以压缩活跃的主要数据。提高了数据被写入磁盘的 I/O 性能。
- 节省磁盘空间

  当主机将数据写入存储时，已经完成了数据压缩。因此，同样的数据，减少了实际占用的存储空间。
- 减少设备数量

  使用在线压缩技术后，同样的存储空间能够存放更多的数据，从而减少所需存储设备的数量，同时降低空间、采购、冷却、管理和维护等 TCO 成本。

应用场景

- 通用压缩卷

  大多数的数据类型都是可以高度压缩的，如目录索引数据，设计数据，石油与天然气地质地震数据。在存放这些类型的数据时，压缩功能可以实现在线压缩，降低整体使用空间。更多的空间可以提供给其他应用。

  压缩卷可以存放多个类型的数据，对于多类型数据的压缩比例，根据实际经验得出的值为 50%~60%。
- 数据库应用

  数据库的数据信息以表空间文件为主，压缩这类文件可以实现很高的压缩比。可以实时压缩，如 DB2、Oracle 和 Microsoft SQLserver 这三种数据库。预期的压缩比为 50%~80%。
- 虚拟化应用

  虚拟化应用已经广泛的应用在现在的业务需求中，并且在不断的扩充市场需求。这就需要有更多的存储空间去支撑虚拟化应用，以保证能够存储更多的虚拟服务器镜像数据和备份数据。使用压缩功能后，减少了对存储空间的需求。虚拟化应用软件支持主流的 VMware、Hyper-V、基于内核的虚拟机。预期的压缩比是 45%~75%。
- 日志服务器

  日志文件对于任何公司和部门都是非常关键的数据。因此，为了方便管理员能够顺利并快速的登录系统日志服务器查询日志也是非常关键的。由于压缩功能的透明性，不会改变系统数据的属性，能够很好地满足企业对性能、可靠性以及可扩展性的要求。其次由于日志文件的特殊性，选择压缩功能的效果是非常明显的，预期的压缩比是高达 90%。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

3.4 智能缓存加速

功能定义

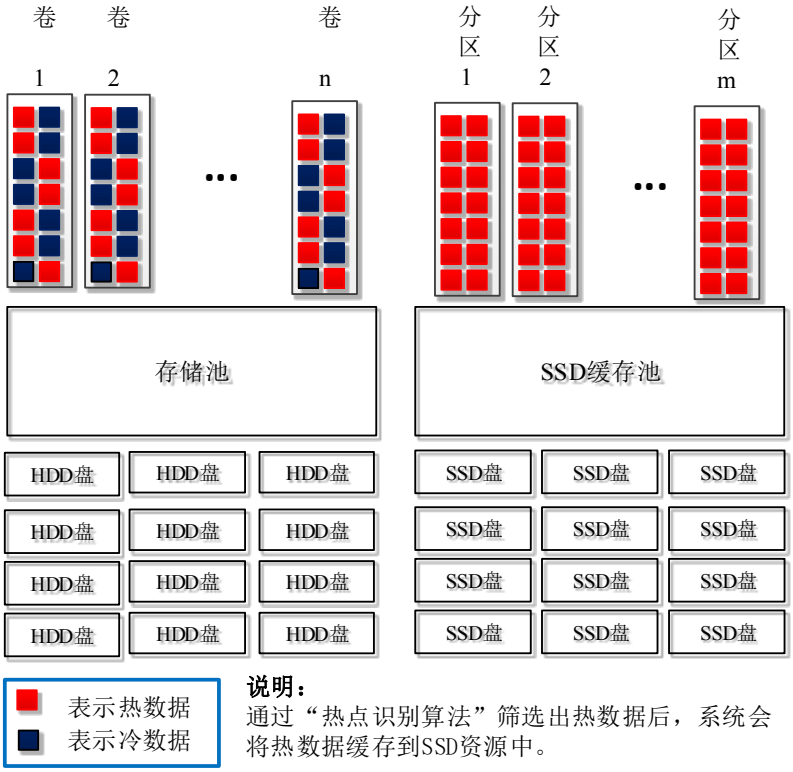
智能缓存加速（InFlashCache）功能是利用 SSD 的快速读能力，将系统常用数据置于 SSD 内，从而满足客户端的快速读取需求。通过 SSD 缓存池对 SSD 资源进行虚拟化管理，已添加 SSD 资源的缓存池，可以创建 SSD 缓存分区，实现 SSD 缓存资源隔离，SSD 缓存分区有两种工作模式， Single 模式和 Double 模式。单个缓存分区可以加速单个卷，也可以加速多个卷。可以根据业务需求选择不同的工作模式来实现。

原理描述

InFlashCache 功能通过 SSD 缓存池对 SSD 资源进行虚拟化管理，只有被添加到 SSD 缓存池中的 SSD 盘，才能用于卷加速。

高效的“热点识别算法”对热点的统计计算，在短时间内（1~2 小时左右）就能体现数据加速效果。并且由于 InFlashCache 的数据缓存开销较小，数据的缓存过程对性能影响较小，可以随时更新 SSD 资源上的热点。

图 3-5 InFlashCache 功能



InFlashCache 功能支持 SSD 缓存分区功能，SSD 资源被添加到 SSD 缓存池中后，就可以根据需要创建 SSD 缓存分区。通过 SSD 缓存分区可以将单个 SSD 缓存池划分为一系列用于不同业务加速需求的分区，实现 SSD 缓存资源的隔离。同时，单个缓存分区可以加速多个卷，将通过“热点识别算法”筛选所有被加速的卷业务中的热点数据，缓存到 SSD 资源上。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

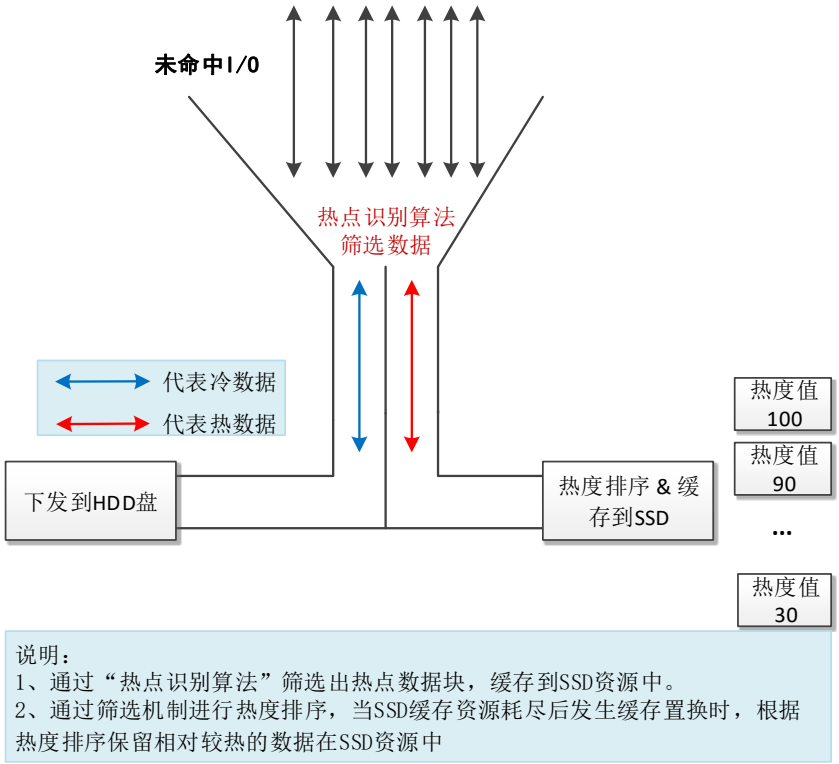
SSD 缓存分区的两种工作模式：

- Single 模式：SSD 缓存分区具备 Node 属性，只能用于加速与分区 Node 属性一致的卷，主要用于 ALUA 模式下的卷加速。GUI 界面会自动过滤非本 node 的缓存分区。
- Double 模式：SSD 缓存分区可以加速 I/O Group 内任一 Node 上的卷。

Single 模式的 SSD 缓存分区扩展子分区后切换为 Double 模式，Double 模式的 SSD 缓存分区删除子分区后切换为 Single 模式。Single 模式下，SSD 缓存分区只有一个子分区，只能加速与 SSD 缓存分区 Node 属性一致的卷；Double 模式下，SSD 缓存分区对应两个子分区，分别加速本 IO_Group 内两个 Node 上的卷，即 Double 模式的 SSD 缓存分区可以加速本 I/O Group 内所有的卷。

InFlashCache 功能通过“热点识别算法”识别出业务中“热点数据块”，然后将其缓存到 SSD 资源中，从而提高整个存储平台的读加速性能。InFlashCache 特性支持高效的“热点识别算法”，提供热点数据筛选机制，通过筛选会识别出热点数据，然后会对这些热点数据进行热度排序，当 SSD 缓存资源耗尽后发生缓存置换时，通过热度排序尽量保留相对较热的数据在 SSD 资源中。值得注意的是，只有筛选出的热点数据才会被缓存到 SSD 资源中，同时通过热度排序机制保证热度值高的数据能够不被替换掉，从而降低 SSD 盘的写次数。如图 3-5 图 3-6 所示。上述机制能够有效避免假热点被写入到 SSD 资源，同时也能够有效降低真正的热点数据由于被置换掉导致重复写入 SSD 的几率，保证性能最优同时，有效提高 SSD 盘的使用寿命。

图 3-6 热点识别算法



技术规格

InFlashCache 在使用时的规格说明如表 3-6 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

表 3-6 InFlashCache 使用规格

规格	参数
每个系统最大支持的 SSD 缓存池数量	8 个
每个系统最大支持 SSD 缓存分区数量	32 个
每个控制器最大支持的 SSD 缓存容量	1TB（内存 16GB）
	2TB（内存 32GB）
	4TB（内存 64GB）
	10TB（内存 128/256GB/512GB）

特性间相互关系

- 目前仅支持 striped 类型的卷（sequential、image 类型的卷暂不支持）
- 被添加到 SSD 缓存池中的 SSD 盘不能用于其他用途
- 智能双活功能和智能缓存加速功能同时使用时，有以下限制：
- 双活拓扑下，仅能配置缓存分区工作模式为 Double 模式。
- 如果存在单缓存分区的情况下，不能从标准拓扑转换到双活拓扑。请删除单缓存分区或者将单缓存分区转换成双缓存分区模式。

license 许可约束

此功能受 license 控制，此功能的工作状态包含未激活、试用激活、使用 license 激活、试用失效、注销。默认情况下是未激活状态，这个状态不支持使用 InFlashCache。试用激活和使用 license 激活支持使用 InFlashCache，包括对 InFlashCache 的配置与管理。试用激活时间为 90 天，使用 license 激活是永久激活。试用到期后进入试用失效状态，所有卷的 InFlashCache 加速功能将被关闭，已创建的 SSD 缓存池和 SSD 缓存分区只支持查询、删除操作，不支持新建 SSD 缓存池、SSD 缓存分区。注销与试用失效的处理方法一致。

技术价值

- 满足客户端的快速读取需求，提高存储平台整体读加速性能。
- 可以优化系统的工作负载，提高 SSD 硬盘的使用率。
- 可以减小热点数据区域的 I/O 延迟。

应用场景

适用于大量读需求的应用环境，比较典型的互联网应用中。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

3.5 智能缓存分区

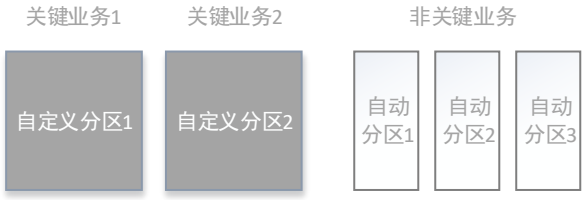
功能定义

智能缓存分区是一种高级性能特性，通过为关键业务分配私有缓存资源的手段，保障关键业务的性能表现。同时提供了非关键业务的缓存资源共享机制，以提升缓存资源利用率。

原理描述

缓存分区是以池为单位的，仅支持父池、不支持子池。对于关键业务，可自定义对应的缓存分区容量大小，该容量会由系统预留，不会被其他缓存分区占用。对于非关键业务，可设置对应的缓存分区为自动模式，则各分区实际容量会随着业务负载的变化而变化。所有自动缓存分区可以使用的缓存资源总量受限。

图 3-7 智能缓存分区视图



技术规格

智能缓存分区在使用时的规格说明如表 3-7 所示。

表 3-7 智能缓存分区使用规格

规格	参数
每个系统最大支持的缓存分区数量	128 个
每个 I/O Group 下，所有自定义缓存分区的容量限制	总和不超过 100%

特性间相互关系

与其他高级特性之间不存在冲突。

License 许可约束

此功能不受 License 控制，在默认状态下即可正常使用。

技术价值

- 隔离缓存资源的使用

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

实现了缓存资源的独立性，避免不同的业务应用之间互相干扰，保障了产品质量。

- 保障关键业务性能

  关键业务的缓存资源不被非关键业务抢占，确保了关键业务的性能及可靠性。

- 提高非关键业务缓存利用率

  缓存资源比较充裕时，允许业务应用尽可能多占用缓存，以提高性能；缓存资源紧张时，缓存资源动态均衡，使得非关键业务的性能表现趋于均衡。

- 限制慢盘资源占用

  通过限制“慢盘”占用缓存资源的上限，避免个别分区过多的占用缓存资源，保障了系统的整体服务质量。

应用场景

可应用在互联网服务的会员与非会员性能分级场景中，可通过将会员与非会员用户的业务规划到不同的缓存分区中，以满足不同用户的需求。

在会员用户所在分区分配充足的私有缓存资源，以保障对会员用户的性能服务质量和可靠性。

而非会员用户的性能要求低，并且业务负载波动大。在非会员用户所在分区，当业务负载小时，可以一定程度的多用缓存资源，以提升性能。当业务负载大时，进行缓存资源均衡，以均衡性能。

3.6 智能服务保障

功能定义

智能服务保障（InQoS）是一种高级性能特性，通过重新分配存储系统资源用来满足某些应用程序的特定性能要求，从而保障关键业务的服务质量。

原理描述

流量控制是基于存储端 IO 队列的管理程序，实现对存储资源的重新分配。系统支持前端到存储端，以及存储端到端的限流功能。可以根据客户需要，在线开启/关闭存储限流功能。系统根据客户设置的限流功能，可以限制前端到存储端 LUN，或存储端 LUN 到 LUN 的 IOPS 或者带宽，降低部分低优先级业务的处理，从而降低低优先级业务对系统性能的占用，保障高优先级业务性能。

技术规格

智能服务保障在使用时的规格说明如表 3-8 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

表 3-8 智能服务保障使用规格

规格	参数
每个系统支持设置 IOPS 范围	0 ~33554432
每个系统支持设置带宽范围	0MB~131072MB

说明：数值为 0 代表不限制 IOPS 或带宽。

特性间相互关系

与其他高级特性之间不存在冲突。

License 许可约束

此功能不受 License 控制，在默认状态下即可正常使用。

技术价值

- 保障关键业务性能，可以通过设置， 对低优先级业务进行限流控制，从而降低系统资源占用率，保障高优先级业务性能。
- 对关键业务进行过载控制，保障存储系统的高稳定性和高可靠性。
- 对存储端到端的业务进行流量限制，保障前端到存储端的性能要求。

应用场景

可应用在需要控制前端到存储端或者存储端到端的 IOPS 或者带宽的场景中。通过限制非关键业务流量，动态分配存储系统资源，以保证关键业务以及高级别用户的业务性能。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

4 数据保护、业务连续

4.1 智能卷镜像

功能定义

Apsara 存储提供一种卷镜像的技术，卷即从存储池中划出并供给主机端使用的逻辑卷。这种技术是基于常见的卷镜像技术，能让一份数据同时保持有两份拷贝卷，并且两份拷贝卷可以是来自于不同的物理存储存储平台（如 Apsara 存储自身的磁盘或者后端虚拟化的阵列），保证数据安全高可用。

原理描述

镜像卷是有两个拷贝卷的，这两份拷贝卷有主次关系，当主机端向镜像卷写入数据时，存储平台会同时将数据写入两份拷贝卷。当主机端要向镜像卷读取数据时，主拷贝卷会向主机端提供所需数据。如果某个拷贝卷意外故障，另一个拷贝卷会立即接管并提供数据，保证主机端应用不中断。当两个拷贝卷重新连接上配对以后，存储平台会重新同步两个拷贝卷上的数据。

通过卷镜像，主机写下来的 I/O 数据能够同时写入两个拷贝卷中并保存。在本存储系统中，这项功能可以通过指定“延迟”和“冗余”的优先级，对 2 个副本的写算法进行优化调节，可以防止两份副本数据同时丢失。

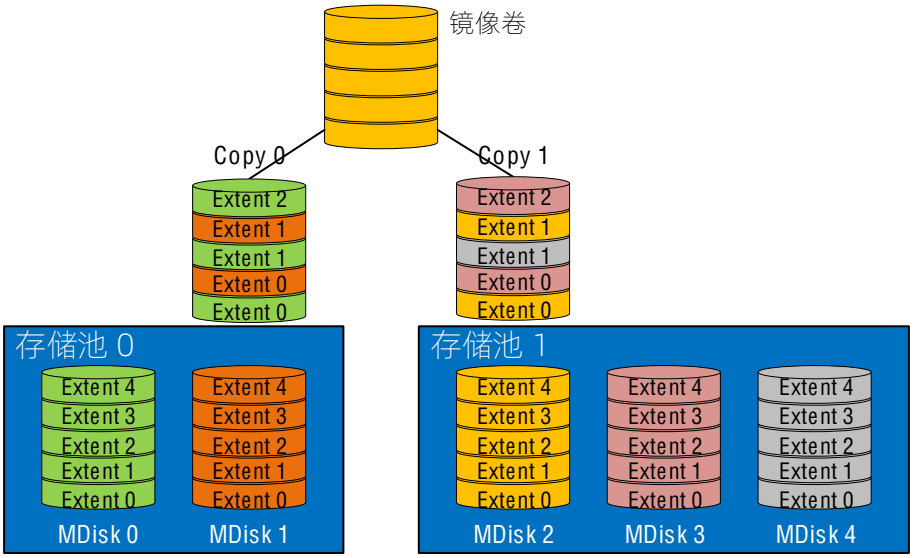
当指定以下优先级时，代表以下含义：

- 指定“延迟”优先，表示当主机端写入 I/O 时，允许响应慢的拷贝卷变为不同步模式。如果另一拷贝卷完成数据写入的操作，即向主机端反馈写 I/O 操作完成。
- 指定“冗余”优先，表示当主机端写入 I/O 时，即使有响应慢的拷贝卷，也要保证两个拷贝卷都完成写入操作，才向主机端反馈写 I/O 操作完成。

智能卷镜像组成示例，如图 4-1 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 4-1 智能卷镜像组成示例



技术规格

镜像卷在使用时的规格说明如表 4-1 所示。

表 4-1 镜像卷使用规格

属性	数值
每个卷的拷贝数量	2
系统支持的拷贝卷数量	10000
每个 I/O Group 的镜像卷总容量	2048 TB

特性间相互关系

- 镜像卷可以应用在智能本地复制、智能远程复制、智能双活场景。
- 镜像卷和普通卷在使用上是完全一样的，可以实现数据迁移。
- 镜像卷可以转化为非镜像卷，只要将其中一个拷贝卷删除或者把拷贝卷分离并创建为一个新卷。
- 可以给普通卷、精简卷、压缩卷创建镜像。

license 许可约束

此功能不受 license 控制，在默认状态下即可正常使用。

技术价值

- 能让一份数据同时保持有两份拷贝卷，并且两份拷贝卷可以是来自于不同的物理存储存储平台（如 Apsara 存储自身的磁盘或者后端虚拟化的阵列），保证数据安全高可用。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- 镜像卷的两个拷贝，数据是保持实时同步的，当两个拷贝卷中的某一个卷发生无法避免的故障时，另外一个拷贝卷可以快速接管故障卷的业务访问，而且无数据丢失。
- 镜像卷的两个拷贝卷可以实时同步或被拆分为独立的卷。当某一个 RAID 保护卷发生无法避免的故障时，始终保持同步的副本可以快速恢复，而且无数据丢失。

应用场景

- 重要数据备份
  镜像卷是由两个拷贝卷组成的，两个拷贝卷在存储系统端是物理存在的，同一份数据，同时保留两份，并且实时同步，保证数据一致性，可以应用在需要对重要数据进行备份的应用。
- 业务不可中断
  对业务连续性较高的应用，可以使用镜像功能，在不同的物理平台上创建两个拷贝卷，彼此互相独立。当一个拷贝卷出现故障，另一个拷贝卷可以快速接管故障卷的业务访问，而且无数据丢失。当故障卷恢复后，两个拷贝卷会重新进行同步，实时保证业务的不中断。

4.2 智能本地复制

功能定义

智能本地复制（LocalCopy）是同一集群内的复制功能，又被成为基于时间点的复制，采用写时复制（Copy On Write）原理，可以保证数据在线可用的状态下，创建一份即时可用的一致性拷贝。Apsara 存储是基于数据块的拷贝，不仅支持同一源卷创建多个目标卷，还支持级联，即可以创建拷贝的拷贝，允许一个卷在不同的关系中，既是源卷又是目标卷。在不影响原本地复制关系情况下，可以继续创建目标卷的拷贝。一个一致性组可以包含多个卷，实现多个卷同时创建拷贝、同时删除拷贝等操作。并且可以通过一致性组保证多个卷同一时间点数据的一致性，确保拷贝数据的完整性。

Apsara 存储针对不同的应用，提供了三种类型的智能本地复制，如下：

- 快照型本地复制
- 克隆型本地复制
- 增量备份型本地复制

原理描述

快照型本地复制在创建时是没有分配实际容量的，会自动地创建一个自动精简的目标卷，快照可以转化为克隆。克隆和增量备份可以调节拷贝进程的优先级，实现拷贝速率可变。

在创建本地复制关系之前，控制器缓存默认是处于写通模式（write-through）。因为本地复制模块是在位于上层缓存之下，确认控制器缓存模式是为了保证创建本地复制时，主机端的数据视图和本地复制模块的数据视图保持一致。智能本地复制在各种状态下，源卷和目标卷对应状态，如表 4-2 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

表 4-2 智能本地复制源卷目标卷对应状态

智能本地复制状态	源卷		目标卷	
	在线/离线	缓存状态	在线/离线	缓存状态
空闲/已拷贝	在线	写回	在线	写回
拷贝中	在线	写回	在线	写回
已停止	在线	写回	离线	N/A
停止中	在线	写回	在线：如果后台拷贝完成，则目标卷是在线状态 离线：如果后台拷贝没有完成，则目标卷是离线状态	N/A
暂停中	离线	写回	离线	N/A
准备中	在线	写通	在线，但不可访问	N/A
准备完毕	在线	写通	在线，但不可访问	N/A

对指定卷开启本地复制的拷贝后，bitmap 就会被建立，源卷和目标卷之间便会存在一个可用的拷贝关系。对于每个目标卷，都存在一个 bitmap 记录，bitmap 记录了源卷和目标卷之间的数据拷贝进展和差异。也就是说用来记录哪些数据已经被拷贝到目标端，而哪些数据还没有被拷贝到目标端。当活动的数据块从源卷拷贝到目标卷，相应的 bitmap 也会对应更新。

图 4-2 智能本地复制拷贝原理-智能本地复制建立

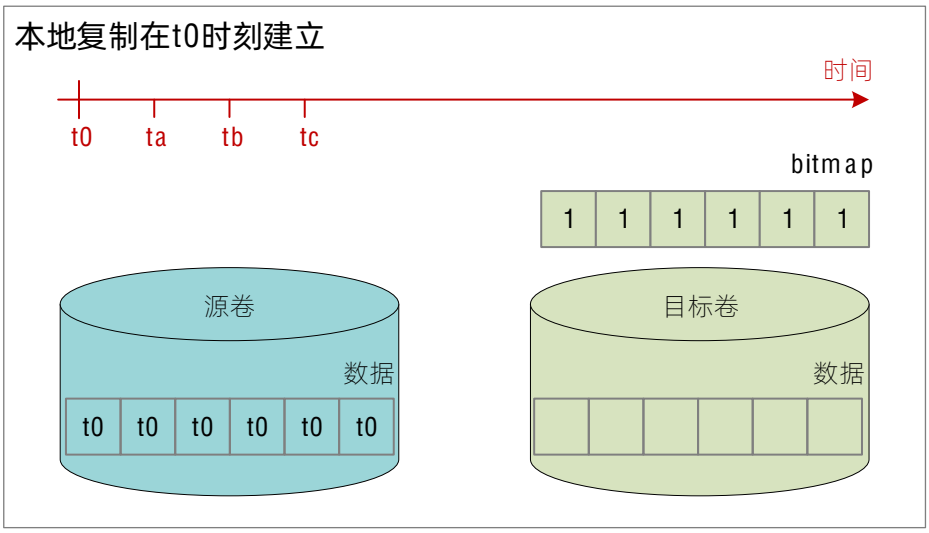
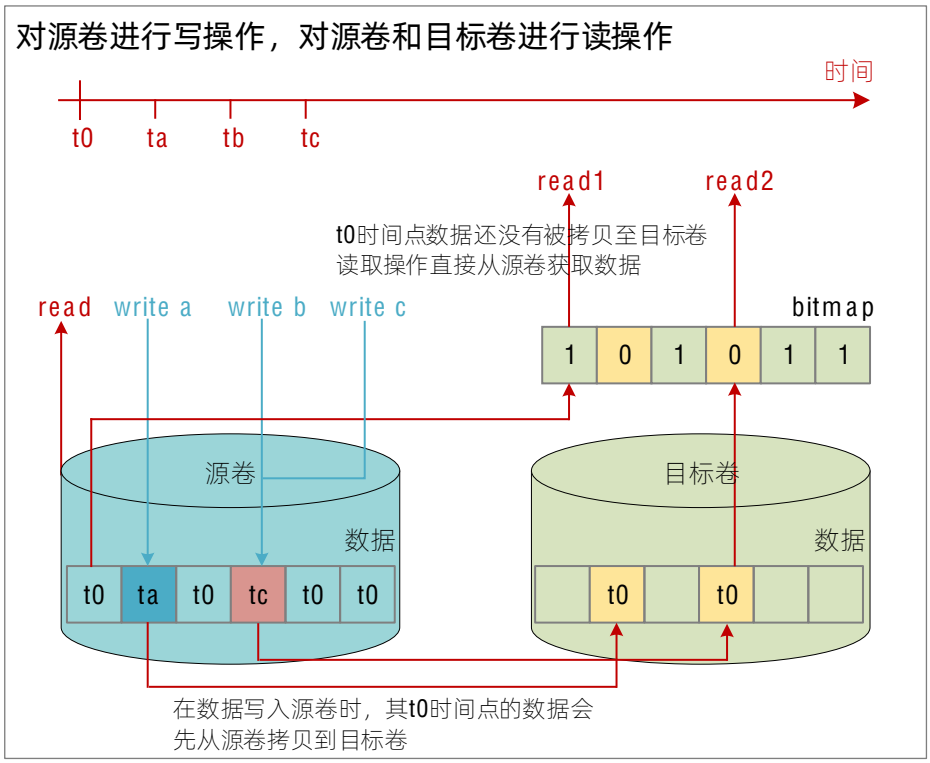


图 4-2 中，在快照建立开始的时间是 t0，此时生成的 bitmap 中的值全部为 1。用户在使用目标卷数据的时候，事实上数据并没有被完全拷贝到目标卷，我们分几种情况来进行阐述：

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- 用户从源卷读取数据。此时采用的是最简单的方式，数据直接从源卷被获取。
- 用户对源卷进行写操作。当快照关系存在时，在开始对所有源卷进行写操作之前，存储都会保证写之前的数据被拷贝到目标卷。为了达成此目标，在对源卷进行写操作时，会先分析之前建好的 bitmap。如果从 bitmap 上获知数据已经被拷贝过，那么写操作直接进行；如果分析 bitmap 发现数据还没有被拷贝，则先将该数据拷贝到目标卷，然后再进行对该数据的写操作。如图 4-3 所示，对源数据的写操作 write a，wirte b 以及 write c 在落盘之前，其 t0 时间点的数据首先会被保证拷贝到目标卷。

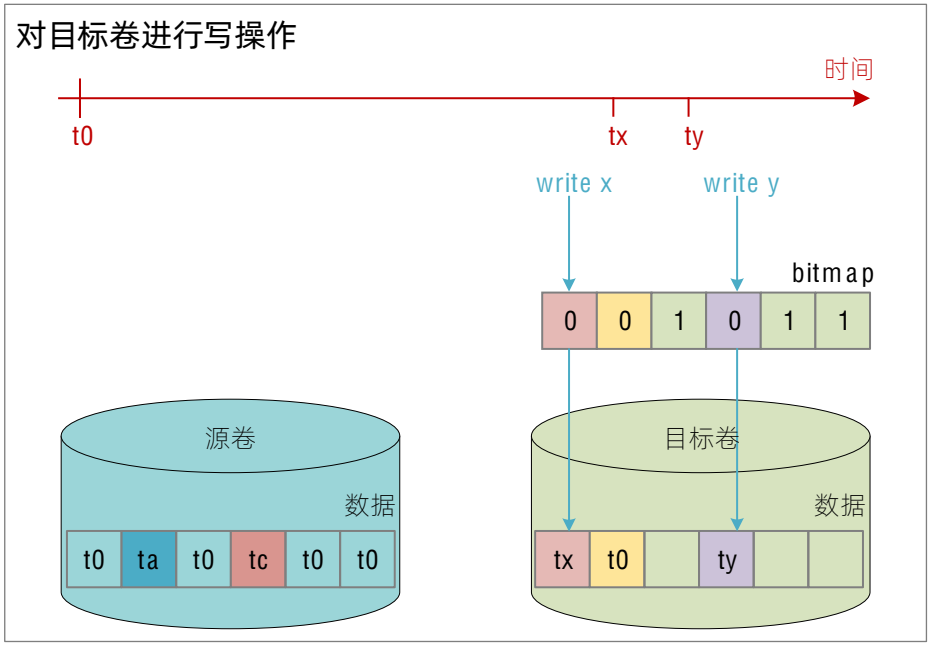
图 4-3 智能本地复制拷贝原理



- 用户从目标卷读取数据。从目标卷读数据时，也会先对 bitmap 进行检查。如果 bitmap 显示数据已经被拷贝到目标卷，则直接从目标卷读取数据；如果 bitmap 显示数据还没有被拷贝至目标卷，则读取操作直接从源卷获取数据。如图 4-3 所示，读操作 read 1 开始时，存储发现 bitmap 中对应的值为 1，也就是其初始值，说明数据还没有被拷贝到目标卷，因此该读操作实际上是对快照源卷的相应数据进行的。而读操作 read 2 进行时，bitmap 中对应的值已经变成 0，说明数据已经被拷贝到目标卷，因此读操作直接在目标卷相应位置进行。
- 用户对目标卷进行写操作。在这种情况下，存储会保证 bitmap 被正确的更新。如果某个数据块还没有被拷贝，那么 bitmap 会被更新为已经拷贝。这样做是为了保护对目标的写不会被本地复制覆盖掉。如图 4-4 所示，写操作 write x 在开始之前，检查到 bitmap 中的值为 0，表示数据已经被拷贝至目标卷，则写操作直接进行。写操作 write y 在开始之前，检查到 bitmap 中的值为 1，表示数据还没有被拷贝至目标卷，则在写操作之前，bitmap 会被更新为 0，然后再进行写操作落盘。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 4-4 智能本地复制拷贝原理



Apsara 存储针对不同的应用，提供了三种类型的智能本地复制功能：

- 快照（后台拷贝速率等于 0，完成后映射关系不删除）
  快照是对源卷数据创建了一个瞬时点的映射视图，它并不是数据的一个独立的拷贝，而是依赖于源卷。当这种类型的快照被创建，快照和源卷数据之间就会保持一个映射关系，并且目标卷与源卷始终保持映射关系。这种快照在创建之初是没有分配容量的，会自动地创建一个自动精简的目标卷，即只有数据写入了源卷，快照目标卷才会有容量分配，在没有数据变化的时候数据空间占用几乎为 0。这种方式常用于数据检索。
- 克隆（后台拷贝速率大于 0，完成后映射关系删除）
  这种本地复制是源卷的一个完全的精准拷贝，当拷贝完成以后，它与源卷是相互独立的，对其进行操作或者修改，不会影响源卷，并且与源卷之间的映射关系也会断裂。当预设为克隆型的本地复制后，后台数据会以指定的速度进行拷贝，创建一个与源卷属性完全一样的目标卷，当拷贝完成到 100%时，克隆与源卷的映射关系会自动删除。这种方式常用于数据的测试或修改。
- 增量备份（后台拷贝速率大于 0，完成后映射关系不删除）
  这种本地复制方式是按照首次全拷贝，再根据数据变化对源卷进行增量拷贝，当拷贝完成以后，增量备份与源卷的映射视图会被更新但不会断裂，还是保持关系，这种本地复制是以最小的数据拷贝量从源卷拷贝给目标卷。增量备份创建的目标卷属性与源卷也是一致的，并以一定的速度进行的增量拷贝。这种方式常用于数据的备份与恢复。

技术规格

后台拷贝速率是一个本地复制关系的属性，目的为取消目标卷对源卷的依赖，从而使拷贝完成后，目标卷成为一份独立于源卷的克隆卷，可设置 0~150。每个本地复制关系的后台拷贝速率都可以可动态调整，设置值

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

与拷贝速率的关系如表 4-3 所示。

表 4-3 后台拷贝速率与设置值关系

设置值	后台拷贝速率（每秒钟）
0	0（关闭）
1-10	128 KB
11-20	256 KB
21-30	512 KB
31-40	1 MB
41-50	2 MB
51-60	4 MB
61-70	8 MB
71-80	16 MB
81-90	32 MB
91-100	64 MB
101-110	128 MB
111-120	256 MB
121-130	512 MB
131-140	1GB
141-150	2GB

本地复制在使用时的规格限制，如表 4-4 所示。

表 4-4 本地复制使用规格

属性	数值
系统支持的最大本地复制数量	5000
单个源卷支持的最大的目标卷数量	256
单个一致性组支持的最大本地复制数量	512
系统支持的最大一致性组数量	499

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

单个 I/O Group 支持的最大本地复制目标卷容量	4096TB

特性间相互关系

智能本地复制不限制源卷和目标卷的类型，可以支持普通卷、精简卷、压缩卷、镜像卷任意组合，但源卷和目标卷的容量必须保持一致。

智能本地复制与远程复制的源卷（主卷）和目标卷（从卷）之间的关系，请参考 4.3 章节。

本地复制源卷、目标卷的其他操作支持状态，如表 4-5 所示。

表 4-5 卷操作支持状态

属性	本地复制源/目标卷
Imaged 模式的卷	支持
存储池内的块迁移	支持，不影响主机 IO
存储池间的卷迁移	支持，不影响主机 IO
I/O Group 间的卷迁移	受限支持，本地复制关系应处于非 prepared 状态，且 bitmap 保留在原 I/O Group 中
扩容	仅克隆关系的目标卷支持，其他类型不支持
缩容	不支持
I/O Group 内更改首选控制器	支持

克隆卷扩容条件及限制如下。

条件：

- 仅能对 copying 状态的克隆关系的目标卷执行扩容操作。
- 该卷只能是一个本地复制关系的目标卷（不能是多个本地复制关系的目标卷，不能在级联本地复制关系中既是源卷又是目标卷）。
- 该卷所在的克隆关系的源卷容量必须是 64KB 或 256KB 的整数倍。

限制：

- 克隆卷扩容之后，不允许回滚到源卷。
- 克隆卷扩容之后，不允许取消克隆关系的 autodelete 属性。
- 克隆卷扩容之后，若克隆关系意外停止，不允许再次启动该克隆关系。
- T3、T4 无法恢复扩容后的克隆关系。
- 克隆卷扩容后，如果克隆关系还存在且处于 copying 状态，则不允许以克隆卷为源卷创建本地复制关系。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

license 许可约束

此功能不受 license 控制，在默认状态下即可正常使用。

技术价值

Apsara 存储一个源卷可以被创建多个目标卷，多个目标卷可以在不需要损坏源数据的同时实现对数据进行分析、测试等功能。

Apsara 存储还支持卷的级联本地复制，也就是说可以给卷创建拷贝的拷贝，允许一个卷在不同关系中既是源卷又是目标卷。并且创建级联本地复制不影响原本地复制的映射关系，能够同时保证多个卷的数据一致。

应用场景

用户可以通过智能本地复制功能解决条件苛刻和要求较高的业务需求，如用户需要对数据组做拷贝，并需要保持一致性时，数据卷可能处于在线状态，正在被使用。之所以能够达到这样的要求，是因为智能本地复制功能是在块级别实现的，这些操作是在缓存的底层进行的，而主机无法感知底层的变化。主要应用在以下场景：

- 快速创建动态变化的数据一致性备份。
- 快速创建生产数据一致性的备份，进行主机之间的数据传输或迁移。
- 快速创建应用程序开发和测试生产数据集的备份。
- 快速创建生产数据的备份，进行审核和数据挖掘。
- 快速创建生产数据集的备份，用于保障质量。

4.3 智能远程复制

功能定义

智能远程复制是为了达到数据保护或灾难恢复的目的，实现一份数据复制成两份拷贝的目的。两份卷可以位于同一个系统中（单个 I/O Group 或者集群内部），也可以位于不同系统中（不同集群的 I/O Group 之间）。

智能远程复制有三种复制模式：同步复制、异步复制、有变更卷的异步复制。同步复制在同一时间向主站点和从站点编写数据，从而保证主从数据的实时一致。而异步复制，数据写到从站点是有延迟的。由于异步复制可以实现更远的距离工作，且需要更少的带宽，对于灾难恢复来说是非常不错的选择。但是，异步复制有丢失数据的危险，因为从卷的数据和主卷数据不是实时同步的。

对不同的容灾环境，Apsara 存储提供了三种不同的数据远程复制模式来应对，用于数据保护，实现灾难恢复。在远程复制中，需要先给在距离上分离的两台 Apsara 存储建立伙伴关系，两套系统完成伙伴关系的配置后，就可以通过 FC 链路或者 IP 链路进行通信。数据的远程复制是发生在两个容量大小相同的逻辑卷之间，生产

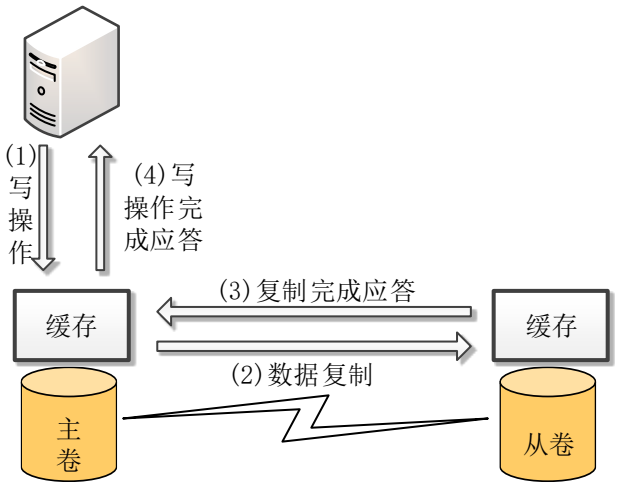
阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

中心作为主端，备份中心作为从端。主卷包含了主机端写入的生产数据，并且承担正常的业务应用数据；从卷包含对主卷的备份数据，并保持与主卷数据的同步更新，用于灾难后的数据恢复。Apsara 存储的这种卷复制关系可以在一台设备中建立，也可以在两台设备之间建立。

原理描述

通过同步远程复制，主机应用数据写入主卷后，只有在相应数据被同步到从卷之后，主机才会收到写操作完成的确认信号，并进行下一次的写入操作。这种模式确保了主从卷之间的数据一致性，即在首次复制完成以后，同步模式下会时刻同步更新主卷和从卷的数据。一次写操作的同步远程复制示意图，如图 4-5 所示。

图 4-5 同步远程复制



同步远程复制步骤如下：

(1) 主机端进行一次写操作，到达主卷的缓存中

(2) 主卷端缓存数据同步复制到从卷端的缓存中

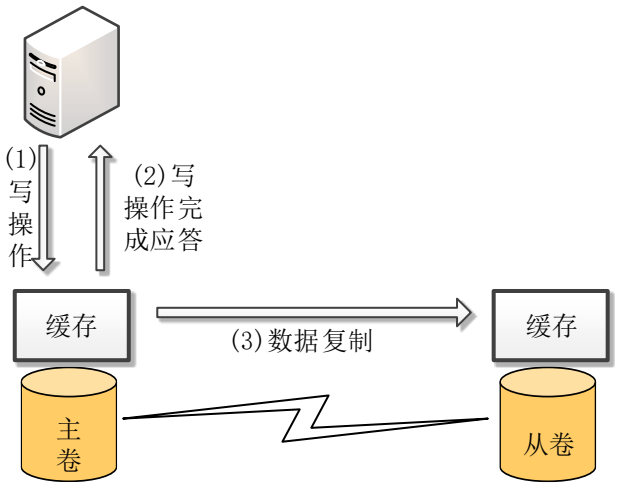
(3) 从卷端给主卷端反馈数据复制完成应答

(4) 主卷端给主机应用反馈写操作完成应答

异步远程复制，在每个时间点上从卷的数据并没有及时地与主卷数据保持一致。这种异步模式相比同步，有可能会造成数据的不一致性，但是异步模式下的算法可以在从端保持一个映射的一致性，通过识别主卷当前活动的 I/O，来获得一致性的映射，当有许多 I/O 同时发生时，从端将按照顺序进行多个 I/O 的操作，因此不会受到远距离链路延迟的限制。一次写操作的异步远程复制示意图，如图 4-6 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 4-6 异步远程模式



异步远程复制步骤如下：

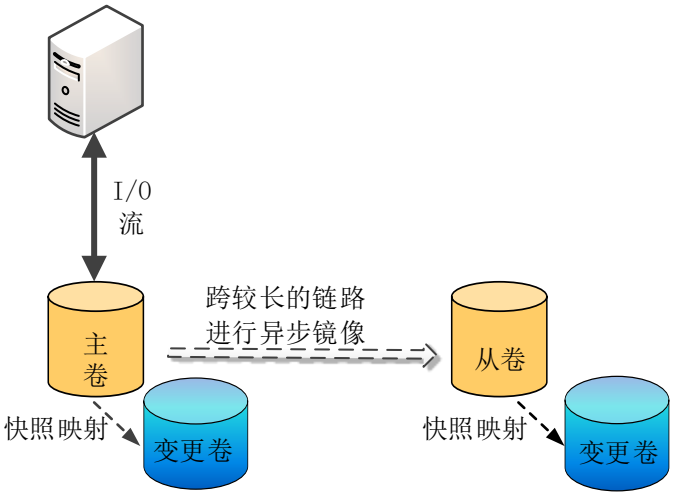
(1) 主机端进行一次写操作，到达主卷的缓存中

(2) 主卷端给主机应用反馈写操作完成应答

(3) 主卷端所有 IO 数据按顺序传输到从端

有变更卷的异步复制，是一种为低带宽网络环境下提供的远程复制模式，该模式是在创建复制关系的时候会自动为主卷和从卷各自创建一个变更卷。变更卷是使用了智能本地复制快照的功能，但是又不同于快照，不像快照一样能够被操作。有变更卷的异步复制，是在一个固定周期内（默认 300 秒，可在 60 秒到 24 小时之间进行调整）从主端到从端传输一系列的快照拷贝。有变更卷的异步复制模式，如图 4-7 所示。

图 4-7 有变更卷的异步远程复制



有变更卷的异步远程复制流程如下：

1. 建立有变更卷的远程复制关系，初始化同步，并建立主卷和从卷的变更卷。
2. 每隔固定循环周期，主端自动启动一次异步远程复制到从端，每次循环的流程如下：

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

a. 启动主卷变更卷快照关系
b. 启动从卷变更卷快照关系
c. 参考主卷变更卷将数据传输到从卷，等待完成
d. 停止主卷变更卷快照关系
e. 停止从卷变更卷快照关系

远程复制链路支持 FC 链路和 IP 链路，使用 IP 链路实现不同站点的设备之间的远程复制是应用最广泛的。在 IP 远程复制过程中，采用了以下技术。

- 采用了压缩技术

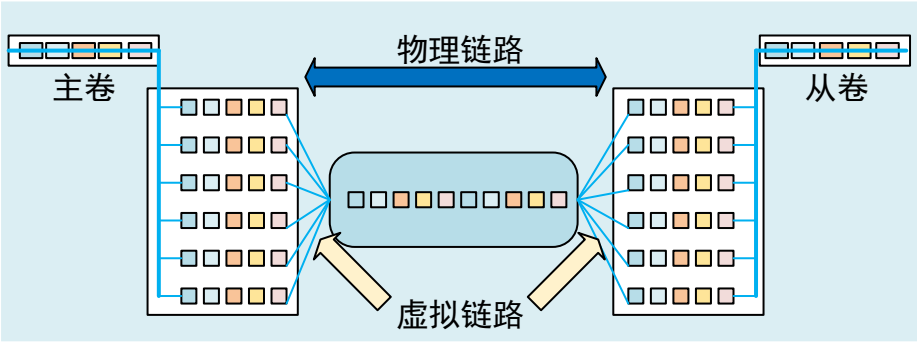
  压缩技术可以支持软件压缩和硬件压缩，软件压缩会消耗系统的 CPU 资源，硬件压缩使用的是智能在线压缩技术，并且可以通过压缩技术，减少复制链路上实际需要传输的数据量。

- 采用了网络传输优化技术

  没有采用网络优化之前，主端发出复制数据到从端需要等待从端反馈复制完成的应答，主端才能开始下一次数据的发送。在长距离的 IP 网络中，网络延迟是制约传输速率的很关键因素，延迟越高传输速率就会越低。

  采用链路虚拟化技术，解决了延迟带来的影响。链路虚拟化是指将一条物理的链路虚拟成多条链路，由多条链路共同去传输数据。如果数据包从任何一条虚拟链路中丢失，数据将被重新发送。虚拟链路的数量由 AI 引擎控制，AI 引擎还监视数据传输期间的链路性能，它可以适当地调整虚拟链路的数量，此信息保留在控制器中，以便如果链接停止并再次启动，则将使用先前设置的虚拟链路重新启动。虚拟链路技术如图 4-8 所示。

图 4-8 虚拟链路技术



技术规格

智能远程复制功能在使用中的规格限制，如表 4-6 所示。

表 4-6 智能远程复制功能使用规格

属性	数值
每个集群支持的智能远程复制个数（3 个模式共用）	10000

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

单个集群内支持的有变更卷的异步远程复制数量	2500
每个一致性组内的远程复制关系	10000
每个集群的远程复制一致性组数量	256
每个 I/O Group 内远程复制卷的总容量	2048 TB
每个集群支持的伙伴关系（FC+IP）总数	3
每个集群支持的 IP 伙伴关系数量	1

特性间相互关系

远程复制不限制主卷和从卷的类型，可以支持普通卷、精简卷、压缩卷、镜像卷任意组合，但主卷和从卷的容量必须保持一致。

同一个卷，在远程复制和本地复制共同使用时的交互性，如表 4-7 所示。

表 4-7 一个卷在同时使用复制功能的交互性

本地复制	远程复制主卷	远程复制从卷
本地复制源卷	支持，两特性分别独立完成复制功能	支持
本地复制目标卷	受限支持，需要两个复制关系都在同一个 I/O Group	受限支持，本地复制关系必须处于 idle_copied 状态

远程复制主卷、从卷的其他操作支持状态，如表 4-8 所示。

表 4-8 卷操作支持状态

属性	远程复制源/目标卷
Imaged 模式的卷	支持
存储池内的块迁移	支持，不影响主机 IO
存储池间的卷迁移	支持，不影响主机 IO
I/O Group 间的卷迁移	不支持
扩容	不支持
缩容	不支持
I/O Group 内更改首选控制器	支持

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

license 许可约束

此功能受 license 控制，此功能的工作状态包含未激活、试用激活、使用 license 激活、试用失效、注销。默认情况下是未激活状态，这个状态可以创建远程复制关系但不能进行数据同步。试用激活和使用 license 激活支持创建远程复制关系，并且保持数据同步。试用激活时间为 90 天，使用 license 激活是永久激活。试用到期后进入试用失效状态，保留已创建的远程复制关系，但停止它们之间的数据同步，以便使用 license 激活后恢复远程复制的数据同步。注销与试用失效的处理方法一致。

技术价值

- 同步复制可以允许两台存储平台的部署距离达到 300KM，实现了远距离的数据实时容灾。对灾难恢复，提供了一种简单的保持主卷与从卷之间数据一致的方法。
- 异步复制模式，数据传输是按照 I/O 的写入顺序，顺序的复制到从端，不会影响前端主机业务的延迟，主机应用等待的时间也会减少，并可以增加部署两台设备之间的距离。同时也确保了整个系统的吞吐量随着系统容量的增加而增加，保持了不断增长的数据组的一致性。
- 加入变更卷的异步复制，可以保证在从端始终保有一份一致性数据，并且在不断的同步中，保护用户数据的一致性。该模式专为低带宽异步复制模式设计的，使得带宽能够根据对不同数据延迟程度的需求而变化。
- IP 远程复制使用压缩技术和虚拟链路技术结合，不仅减少了传输的数据量，降低了对带宽的要求，还增加了虚拟链路共同去传输数据，使数据传输速率接近满带宽。

应用场景

- 容灾场景

  Apsara 存储提供三种模式的远程复制，用户可根据站点距离和对应用的要求，自主选择合适的复制模式。支持同城容灾环境：不同机房、不同楼宇间、不同地点间的容灾环境；还支持远距离的容灾场景，距离可达 8000 公里。
- 网络延迟较高、带宽较低场景

  如果是在一个网络带宽比较低、拥堵或者超载的主机应用中，会产生较坏 的影响，影响业务运行、影响主从端数据一致性，甚至会是出现许多错误。在这种环境下，可以选择优化了的 IP 异步远程复制，结合网络优化技术，降低网络问题造成的风险。

4.4 智能双活

功能定义

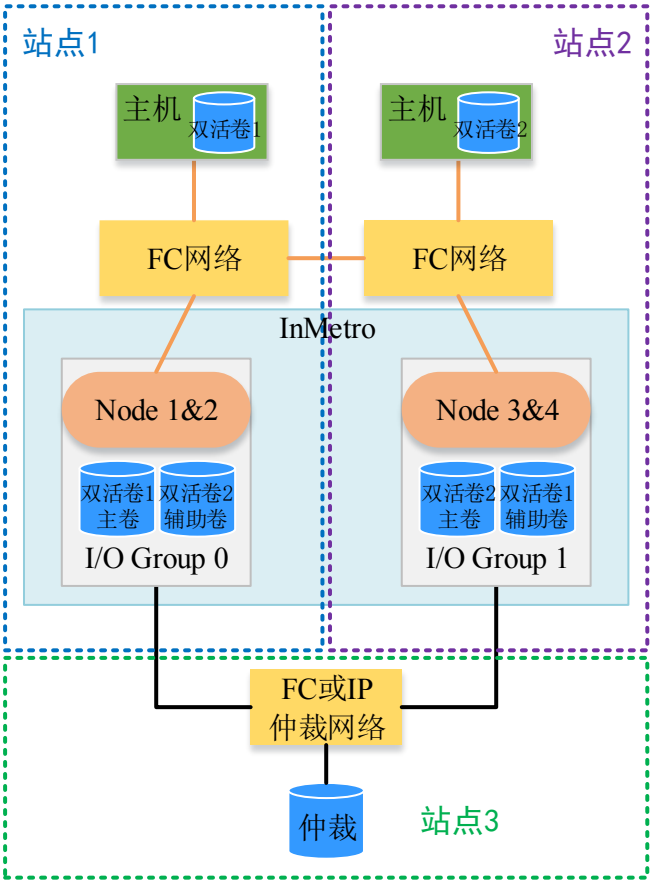
智能双活（InMetro）功能是基于集群存储平台的，由两个 I/O Group 同时管理每个逻辑卷。InMetro 功能使用了同步远程复制、变更卷技术和智能数据迁移的技术。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

原理描述

智能双活存储的站点部署如图 4-9 所示。主机和存储分别分布在两个不同的站点。每个 I/O Group 的控制器处在一个站点，每个卷作为一个单一的对象对于两个站点（I/O 组）可见、可访问。InMetro 允许用户指定主机的站点，使其可以访问最近的卷副本，主机会优先使用最近的副本，但是也能在发生故障时转移到另外一个副本。

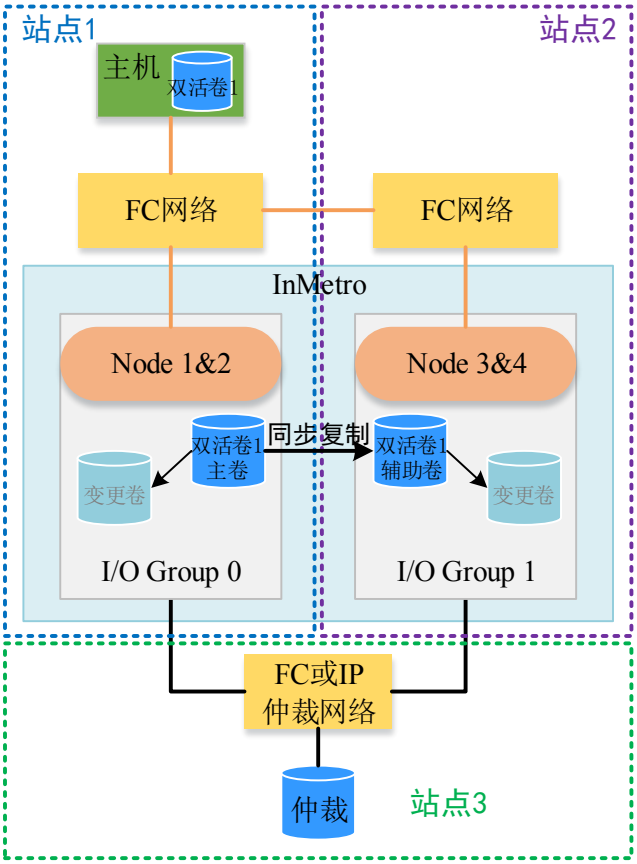
图 4-9 站点分布



智能双活通过同步复制来实现两个 I/O Group 之间的同步。正常情况下，由主卷提供数据访问服务，辅助卷不映射给主机且为 Offline 状态。集群内部通过同步远程复制实时同步数据。同步复制关系由存储系统自动控制，无法手动进行启停操作。I/O Group 同步方式，如图 4-10 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 4-10 I/O Group 同步方式



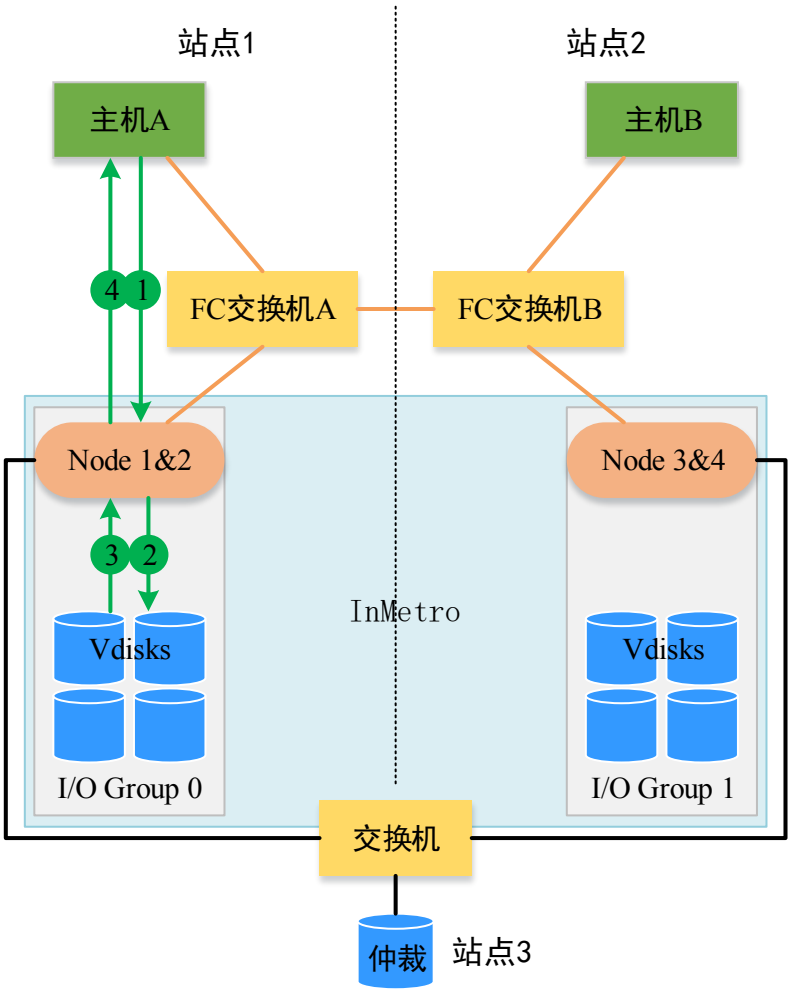
InMetro 的读过程，如图 4-11 所示，每个步骤的说明如下：

1. 主机端发送读请求
2. 存储端发送读请求
3. 存储端反馈数据
4. 数据发送给主机端

如果需要读取的数据位于存储的缓存中，则不经过步骤 2 和 3。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 4-11 InMetro 读过程



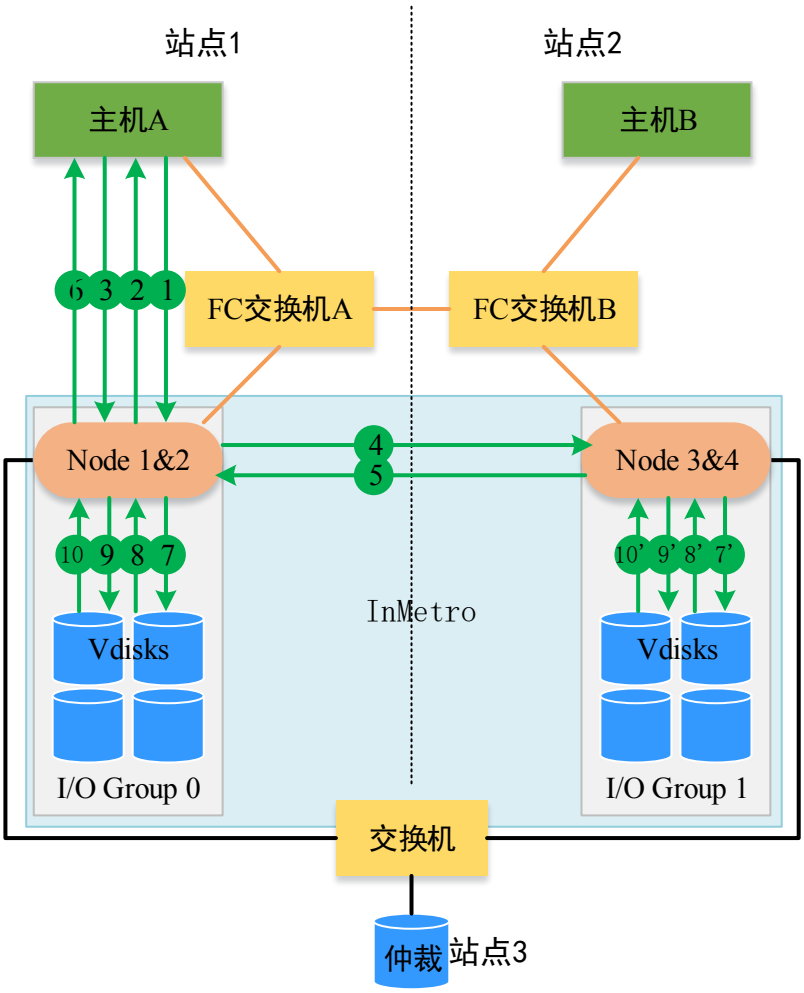
InMetro 的写过程，如图 4-12 所示，每个步骤的说明如下：

1. 主机端发送写请求
2. 反馈数据可以写入
3. 主机端写入数据到缓存
4. 写缓存数据复制到远端节点
5. 远端反馈已写完
6. 反馈写操作完成
7. 存储端发送写请求
8. 反馈数据可以写入
9. 存储端写入数据到磁盘
10. 反馈写操作完成

步骤 1-6 是同步操作，会影响主机应用性能，增加延时。步骤 4-5 在双活的应用中写 I/O 仅传送一次数据到远端。步骤 7-10 是异步操作，不影响主机应用性能。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 4-12 InMetro 写过程



技术规格

智能双活功能在使用时的规格限制如表 4-9 所示。

表 4-9 智能双活功能使用规格

属性	数值
每个集群支持的双活关系个数	1250
集群支持的双活最大容量	4PB
单个 IO Group 支持的双活最大容量	2PB

特性间相互关系

智能双活功能和智能缓存加速功能同时使用时，有以下限制：

- 双活拓扑下，仅能配置缓存分区工作模式为 Double 模式。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- 如果存在单缓存分区的情况下，不能从标准拓扑转换到双活拓扑。请删除单缓存分区或者将单缓存分区转换成双缓存分区模式。

其他特性与智能远程复制功能相同，请参考 4.3 章节。

license 许可约束

此功能受 License 控制，此功能的工作状态包含未激活、试用激活、使用 License 激活、试用失效、注销。试用激活和使用 License 激活支持创建智能双活关系，并且保持数据同步。试用激活时间为 90 天，使用 License 激活是永久激活。试用到期后进入试用失效状态，保留已创建的双活拓扑关系，但停止它们之间的数据同步，以便使用 License 激活后恢复双活拓扑的数据同步。注销与试用失效的处理方法一致。

如果需要永久激活智能双活功能，需要注册远程复制 License 许可。详细内容请参考 4.3 章节。

技术价值

- 在智能双活存储的场景下，由于一个 I/O Group 的所有控制器在同一机柜，两个 I/O Group 里又有对方的副本数据，即使只有一个 I/O Group 也能保证主机端完整的读和写的性能。
- InMetro 支持双活卷一致性组，应用中每个卷的数据都可以被一致地维护，极大地提高了使用 InMetro 灾备的应用场景，因为是在单一集群内，是集群内的数据复制，不涉及跨集群的远程数据复制。
- 允许业务连续性的硬件故障、设备断电、控制器间通信连接失败、自然灾害（如火灾、水灾）等场景。
- 差异数据增量同步时，提供一致性时间点保护数据。

应用场景

- 业务不可中断

  对业务连续性较高的应用，可以使用智能双活功能，同一份数据同时保存在两个 I/O Group 上，彼此互相独立。当其中一个站点出现任何故障，另一个站点即可快速接管故障站点的业务访问。此过程不仅有效保证了业务的连续性，而且还保证无数据丢失。
- 可靠性要求较高场景

  支持多种场景故障，整个站点、I/O Group 双控制器、存储子系统故障场景，I/O 可以无缝切换到备用资源，对上层应用是透明的。整个站点故障场景下，远端的 I/O Group 仍有写缓存保护机制，避免直接进入直写模式带来的性能下降。

4.5 智能云分层

功能定义

智能云分层（Transparent cloud tiering）功能是指将本地存储平台的卷数据备份到云存储中。本地存储平台支持创建到云服务提供商的连接，以将卷数据副本存储在私有云存储或公有云存储中。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

云卷（cloud volume）是指启用了智能云分层功能的任意卷。在创建云卷之前，必须配置支持的云服务提供商的有效连接。

云快照（cloud snapshot）是对云卷或云卷组创建的基于时间点的副本，并传输到由云服务提供商管理的云存储中。这些快照也可以复原到本地存储平台以用于灾难恢复目的。

云帐户（cloud account）是指系统连接到云服务提供商所使用的一组特定凭证。

原理描述

如果需要将数据复制到云存储，必须创建从本地存储平台到云服务提供商的连接。这就需要本地存储平台使用云帐户凭证向云服务提供商进行认证，从而建立起与云存储服务的通信。如果系统通过认证，那么本地存储可以访问云存储以将数据复制到云存储或者将复制到云存储的数据恢复到本地存储平台。

云账户凭证根据指定的云服务提供商类型的不同而有所不同。大多数云服务提供商都需要云服务提供商的主机名和关联密码，而某些云服务提供商还需要用于认证云存储用户的证书。公有云使用认证中心签署的证书。私有云服务提供商可以使用自签名证书或者由可信认证中心签署的证书。这些凭证在云服务提供商处定义，并通过云服务提供商的管理员传递到系统。系统支持一个云帐户对应于一个云服务提供商。不支持在提供商之间进行数据迁移。

每个云服务提供商都将云存储划分为多个分段，以用于使用云存储的每个本地存储平台，每个分段仅存储指定存储平台的数据。在云存储端每个本地存储平台的名称以在为系统创建帐户时指定的前缀开头，以便能够支持多个独立的存储平台将数据存储到单个云帐户存储中。这些独立存储平台的名称的命名规则，根据指定的云服务提供商类型的不同而有所不同。

智能云分层支持万兆云备份功能，即针对一些特殊云环境，可以通过存储的万兆网口和第二个集群 IP 进行数据传输与备份。

智能云分层功能使用的是与智能本地复制的快照相同的技术，使用此功能的对象为单个卷或卷组，对卷启用此功能后，可以在云存储上创建和存储这些卷的云快照。多个快照可以在不同的时间创建，每个版本都带有时间戳并存储在云存储中。云存储上的每个快照版本表示特定时间点的数据，因此必要时可以恢复不同的快照版本。

系统支持的云快照分为两种类型：

- 完整快照（Full snapshots）

  当首次为卷创建云快照时，所有数据都会复制到云存储。该卷后续的任意云快照仅包含自初始快照以来发生更改的数据。卷的每个快照版本都带有时间戳记并存储到云存储上。因此，完整快照和增量快照的工作方式能确保所有数据都复制到云存储，并在需要执行复原操作时可供使用。如果卷包含大量数据，完整快照可能需要较长时间。

- 增量快照（Incremental snapshots）

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

该版本包含自上次为所选卷创建云快照以来发生更改的数据的快照副本。与完整快照相比，增量快照完成时间较短，需要的云存储空间更小。

除了创建卷数据快照以传输到云存储之外，智能云分层还支持将卷数据从云存储复原到系统。像快照一样，每个复制操作的版本都带有时间戳记并存储在云上，这允许向原始系统或另一系统执行指定的时间点复原操作。只能对单个卷执行复原快照操作，也可以同时复原多个卷，但需要对每个卷分别启动复原快照操作。较旧版本的快照可以同时从系统和云存储中删除以实现空间管理。

快照版本与正在将快照版本复原到的目标卷，必须具有相同大小。不能删除正在用于复原的版本。您可以通过以下方法复原快照版本：

- 复原到生产卷
  如果将快照版本复原到生产卷（创建快照的原始卷），快照版本使用云存储上存储的数据替换当前卷上存在的数据。生产卷在复原操作期间为脱机状态。
- 复原到新卷
  当快照版本复原到新卷时，复原操作会创建一个新卷，并将此快照版本的数据写入新卷。新卷在复原操作期间为脱机状态。

注意

当复原较旧版本的云快照到生产卷时，云存储上所有较新版本的快照会被删除，包括正在进行的快照操作。

在通过公共网络将系统连接到云服务提供商时，由于系统需要访问外部网络，可能会存在数据泄露的风险。但是，Apsara 存储可以实现基于加密的安全性，从而在云中传输和存储数据时提供保护。标准协议 TLS 可以对系统和云服务提供商之间传输的数据进行加密，以起到保护作用。安全通信对于这些连接而言是必需的，并要求在云服务提供商和系统之间交换公用证书。通过安全通信，数据在传输到云时进行加密，但在云中存储时可能处于解密状态。当数据存储到云存储后，由每个云服务提供商提供安全措施来保护数据。使用云服务提供商的客户可以添加额外的加密方法，以在数据存储到云之后加以保护。

技术规格

表 4-10 智能云分层技术规格

属性	数值
每个集群支持的最大云账户数量	1
每个集群支持的最大云卷数量	1024
每个卷支持的最大云快照数量	256
每个集群支持的最大卷组数量	256

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

特性间相互关系

如果在以下任何情况中使用卷，那么不能对该卷启用云快照：

- 卷是智能远程复制关系的一部分。
- 卷为 VVol 卷。
- 卷为智能本地复制映射的一部分。
- 卷在其他存储池中具有副本。
- 正在不同的存储池之间进行数据迁移的卷。
- 如果将云存储设置为导入方式。
- 如果系统中存在的云卷数目已达到最大值 1024。

说明

如果系统超出该限制，那么可以禁用现有云卷上的云快照并从云存储中删除其关联的快照，以容纳新云卷上的快照。

License 许可约束

此功能不受 license 控制，在默认状态下即可正常使用。

技术价值

- 通过智能云分层，可以将较旧数据移到云存储来释放系统上的容量。
- 通过智能云分层，可以在系统上创建基于时间点的数据快照，然后将其复制并存储到云存储中。
- 外部云服务提供商负责管理云存储，因此降低系统的存储成本和维护成本。

应用场景

当需要保护卷数据，并需要传输到云存储时，可以使用此功能。Apsara 存储除了支持以下这四家主流服务商之外，用户还能通过 Apsara 存储拥有的独立云网关，接入特定的云服务商，获得定制的云备份服务。

- SoftLayer
  系统可以连接到 SoftLayer，这是基于 OpenStack 对象存储标准针对云存储解决方案构建的云服务提供商。如果您连接到该云服务提供商，可以在私有和公有云存储中创建卷数据快照。
- OpenStack Swift
  OpenStack Swift 是标准的云计算体系结构，管理员可以借助它来管理单个私有云环境中的存储和联网资源。标准应用程序可编程接口 (API) 可以用于为私有云解决方案构建可定制解决方案。
- Amazon S3
  Amazon Simple Storage Service (Amazon S3) 为程序员和存储管理员提供了灵活安全的公有云存储。Amazon S3 也基于对象存储标准，提供了基于 Web 的界面来通过 Web 管理、备份和恢复数据。
- Aliyun OSS

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

系统可以连接到阿里云，OSS 为海量、安全和高可靠的云存储服务，使用云存储的资源实现对系统数据的备份。如果您连接到该云服务提供商，可以在私有和公有云存储中创建卷数据快照。

4.6 智能数据销毁

功能定义

当不再需要某个 vdisk 时，可以对该 vdisk 进行智能数据销毁。销毁单位是 vdisk，智能数据销毁使用的是数据覆写的方式。

原理描述

Apsara 存储系统提供两种智能数据销毁方式：

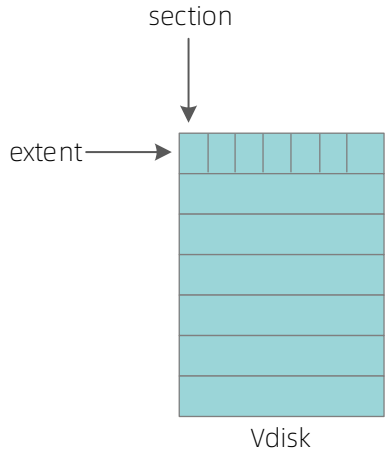
- DoD：系统根据美国国防部 DoD 标准对所选的 vdisk 进行数据覆写，覆写 3 次。
- 自定义：系统根据设定的覆写次数对所选的 vdisk 进行数据覆写。覆写次数取值范围为 3~100，默认为 7 次。

两种智能数据销毁方式覆写数据规则如下：

- 第 1 次，全 0 覆盖写。
- 第 2 次，全 1 覆盖写。
- 第 3~100 次，每次均为随机数。

vdisk 是由存储池中一个或若干个 extent 组成，数据销毁功能将每个 extent 划分为若干个以 256KB 为单位的 section。如图 4-13 所示。

图 4-13 Extent & Section

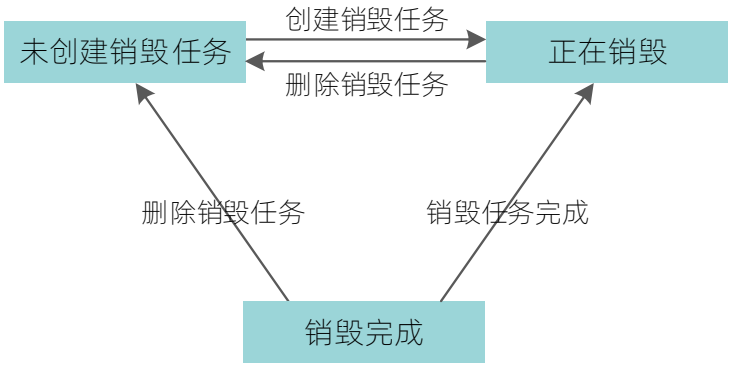


数据覆写的最小单位是 section，执行数据销毁任务时，会从 vdisk 的第一个 extent 的第一个 section 进行覆写，直至整个 vdisk 全部覆写完成，此数据销毁任务完成。vdisk 执行数据销毁任务的状态转换图，如图 4-14

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

所示。

图 4-14 销毁状态转换图



特性间相互关系

- 销毁功能只针对非精简卷进行销毁，精简卷由于写时分配以及空间释放回收的特性，无法完成所有空间的销毁。
- 已配置增值业务（如本地复制、远程复制、双活、云卷、NAS 卷）的卷，不允许销毁。
- 正在格式化、迁移、扩容/缩容的卷以及缓存分区的卷，不允许销毁。
- 在对卷进行覆写之前，需确保卷的状态是正常的，并且要取消卷和主机之间的映射。

License 许可约束

此功能受 License 控制，此功能的工作状态包含未激活、试用激活、使用 License 激活、试用失效、注销。默认情况下是未激活状态，这个状态不能创建销毁任务。试用激活和使用 License 激活支持创建销毁任务。试用激活时间为 90 天，使用 License 激活是永久激活。试用到期后进入试用失效状态，此时正在执行的销毁任务仍然可以执行，直到任务完成。但不允许创建新的销毁任务。注销与试用失效的处理方法一致。

技术价值

通过销毁功能将已作废的数据一键智能销毁，不可恢复，保障机密数据的安全性。

应用场景

- 容灾场景

  本存储提供三种模式的智能远程复制，用户可根据站点距离和对应用的要求，自主选择合适的复制模式。支持同城容灾环境：不同机房、不同楼宇间、不同地点间的容灾环境；还支持远距离的容灾场景，距离可达 8000 公里。

- 网络延迟较高、带宽较低场景

  如果是在一个网络带宽比较低、拥堵或者超载的主机应用中，会产生较坏 的影响，影响业务运行、影响主从端数据一致性，甚至会是出现许多错误。在这种环境下，可以选择优化了的 IP 异步远程复制，结合

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

网络优化技术，降低网络问题造成的风险。

4.7 NAS 快照

功能定义

NAS 快照功能是针对 NAS 文件系统提供快照功能，可以以整个 NAS 文件系统为单位创建和回滚快照。本功能基于智能本地复制功能，采用写时复制（Copy On Write）原理，可以保证 NAS 业务共享持续可用下，创建一份只读的一致性拷贝。当共享数据丢失、损坏时，可通过回滚该文件系统的快照立即切换回快照点时的状态，找回丢失数据。

NAS 快照功能同时提供 CIFS 共享和 Windows 系统的卷影复制（Volume Shadow Copy）功能，用户挂载 CIFS 共享时可从挂载点中浏览和恢复历史数据。

原理描述

本功能实现采用 NAS 业务端使用 NAS 资源管理快照，底层快照实现调用智能本地复制功能接口。

NAS 快照的创建、删除、查询和回滚操作为智能本地复制功能对应接口的封装，即由 NAS 模块调用对应命令、管理由智能本地复制功能创建的快照卷和映射关系：

- 创建快照时，NAS 模块暂停源文件系统对卷的读写并创建快照卷和映射关系，然后恢复卷读写、将快照卷挂载至 NAS 虚机进行管理。期间源文件系统对卷的读写会短暂停止，对外提供的共享则持续可用。
- 删除快照时，NAS 虚机卸载快照卷，并停止映射关系、删除快照卷。删除快照不影响源文件系统和共享。
- 查询快照时，NAS 模块同时输出快照卷、映射关系和快照卷挂载的状态。
- 回滚快照时，NAS 模块停止文件系统上的共享、卸载文件系统，然后执行从快照卷到源文件系统卷的回写映射过程。开始执行回写后，源文件系统卷的视图实际已变为快照创建时间点的一致性视图，此时挂载文件系统，共享立即恢复可用，无需等待回滚完成。

NAS 快照提供 CIFS 共享和 Windows 系统的卷影复制（Volume Shadow Copy）功能。快照 CIFS 共享是在导出的共享目录中自动创建链接，链接到快照卷中对应的共享目录，用户可通过该链接浏览和恢复快照点时刻的数据。卷影复制为微软设计的备份功能接口，NAS 模块导出的 CIFS 共享均实现对该接口的支持，即共享目录在 windows 端挂载后，可在目录或文件的历史版本中检出快照版本数据。

技术规格

NAS 快照在使用时的规格说明如表 4-11 所示。

表 4-11 NAS 快照使用规格

规格	参数

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

最大快照数量	16 个 **说明：** NAS 快照数量统计在智能本地复制数量内。

特性间相互关系

与其他高级特性之间不存在冲突。

License 许可约束

此功能不受 license 控制，在默认状态下即可正常使用。

技术价值

实现对 NAS 文件系统的快照备份功能，使 NAS 文件系统具备与 SAN 卷同等的备份和恢复能力。

应用场景

- 快速创建动态变化的数据一致性备份，用于文件系统数据恢复。
- 快速创建生产数据集的备份，用于保障质量。
- 快速创建指定时间点的一致性视图，用于纵向对比。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

5 灵活高效的主机插件

Apsara 针对主机应用提供了更丰富的接口软件，在减少用户的管理成本、增强存储的易操作性、可用性、易用性等方面增加了更广泛的功能。Apsara 提供了以下这些主机插件，应用在不同的数据中心环境中，这些软件是不受 license 控制的，可以免费获取并使用。

5.1 SCOM 插件

Microsoft System Center Operations Manager (SCOM)是微软推出的专业系统监控软件，可以监控部署在网络中的服务器、应用系统和客户端，提供图形化视图，使管理员可以监控目标计算机存在的故障和产生的告警。微软所提供的 SCOM 软件支持的主机系统范围有限，基本上仅包含微软自己的系统与软件，同时包含主流的操作系统，如 Linux、Unix。

微软 SCOM 支持用户通过自定义管理工具来扩展 SCOM 功能，Apsara 提供的 SCOM 插件管理包通过对管理工具的定制，可以使得微软 SCOM 能够对 Apsara 的存储平台进行监控，并能够获得运行状态和警告等相关信息。

SCOM 插件是安装在微软 SCOM 管理服务器或运行 SCOM 代理的服务器上，在微软 SCOM 软件环境安装完成后，导入 Apsara 存储相应的 SCOM 插件管理包，通过 Apsara SCOM 插件管理包自带的命令行工具对 SCOM 管理服务器进行配置，添加需要被监控的 Apsara 存储后，可以实现在微软 SCOM 的界面中对本存储系统的活动告警和事件进行收集，并能够对存储的逻辑组件和物理组件健康状况进行监视。SCOM 插件的主要功能如下：

- 报警收集：列出所有存储平台的报警信息。
- 事件收集：列出捕获的存储平台事件信息。
- 状态信息显示：
  - 系统信息：显示存储平台和它的所有组成部分的整体健康状况信息。
  - 逻辑组件信息：显示定义的逻辑组件的健康状况信息，逻辑组件包括存储池、卷、主机映射关系、本地复制映射关系、本地复制一致性组、远程复制关系、远程复制一致性组
  - 物理组件信息：显示物理组件的监控状况信息，物理组件包括 Arrays，Drives，Enclosures，Nodes，MDisks，iSCSI Ports，Fibre Channel Ports，Hosts。

Apsara 的 SCOM 主机插件，可以支持微软 SCOM 软件的版本信息如下：

- Microsoft System Center Operations Manager 2012

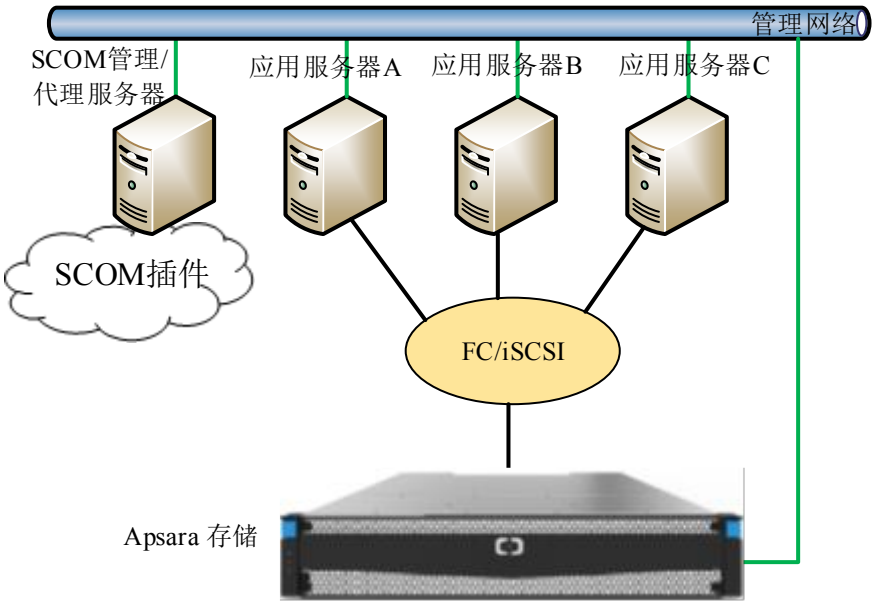
阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- Microsoft System Center Operations Manager 2012 SP1
- Microsoft System Center Operations Manager 2012 R2

可以把 Apsara 的 SCOM 插件管理包安装在已使用微软 SCOM 软件的任意版本 Microsoft Windows 或 Windows Server 操作系统主机上。

Apsara 的 SCOM 插件主要应用在，帮助用户使用统一的 SCOM 界面实现查看受监控计算机系统和受监控存储平台的状态、健康和性能等信息。在应用环境中的部署网络图，如图 5-1 所示。

图 5-1 SCOM 主机插件安装部署



5.2 VSS 插件

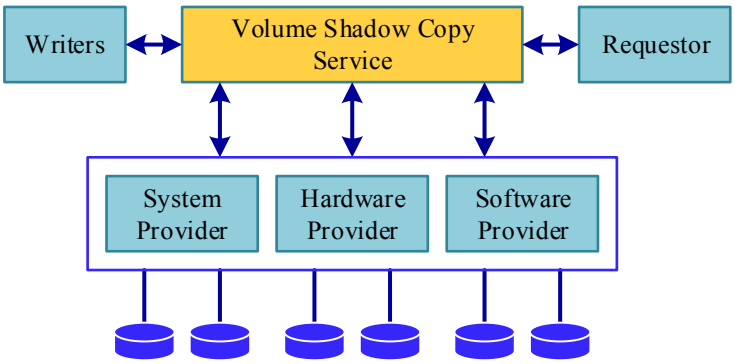
微软 VSS (Volume Shadow copy Service，卷映射拷贝服务)，最初是在 Windows Server 2003 被引入的存储技术，它是一个软件模块，运行在微软 Windows 服务器上，用来促进 VSS 各个组件之间能够更好的协同工作，组件包含：VSS 本身服务、数据保护备份软件（requestor）、应用程序（writer）、插件（provider）。

通过微软 VSS，可以创建基于时间点(Point-In-Time)的映像(Image)，并可以恢复某一时间点的数据状态，从而实现数据的快速备份和恢复，不再因为无意删除数据而造成困扰。如果所有组件都支持 VSS 服务的话，还可以通过 VSS 实现在线备份应用程序数据的功能，并且不再需要中断应用程序。

一个完整的 VSS 解决方案包括以下四部分，如图 5-2 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 5-2 完整的 VSS 解决方案



- VSS service：在 Windows 服务器上运行，是 Windows 操作系统的一部分，可以确保各个组件/模块之间能够互相通信、协同一致工作。
- VSS writer：该模块通常是指应用程序的一部分，保证备份的数据具有一致性，比如 SQL Server 或者 Exchange Server。非 Windows 系列应用程序，需要保证备份期间的数据一致性。
- VSS requestor：通常指备份软件，可以是 Windows 服务器上的备份软件（system center data protection manager），也可以是其他运行在 Windows 服务器上的非微软的备份软件。用来发起管理（创建、引入、删除）VSS 卷映射的请求。
- VSS provider：包含 system provider、software provider、hardware provider。Windows 操作系统自带的 system provider 采用 copy-on-write（写时复制）技术。hardware provider 可以是 SAN 环境中提供存储空间的存储设备。Apsara 的存储平台可以给 VSS 服务提供本地复制功能，可以把主机操作系统端管理 VSS 卷映射的工作任务和资源占用转移到存储端，那么安装基于 SAN 的 VSS hardware provider 是非常有必要的。

Apsara 的 VSS 插件（hardware provider）安装在主机端（Windows 操作系统主机），安装后以服务的形式运行在操作系统中，用来调用本存储系统端的智能本地复制功能，采用微软 VSS 框架作为它的管理界面，用来创建和维护本地复制，将本地复制数据保存在本存储系统中。

Apsara 的 VSS 主机插件，可以支持微软操作系统有版本信息如下：

- Microsoft Windows Server 2008 SP2 x86 & x64
- Microsoft Windows Server 2008 R2 SP1 x64
- Microsoft Windows Server 2012 x64
- Microsoft Windows Server 2012 R2 x64

5.3 InPath-多路径驱动

存储多路径（multipath I/O）是指在主机（Host）与存储设备（Storage Array）之间存在多条物理连接，通过主机操作系统提供的多路径访问机制，在多条物理路径上访问存储设备。目的是实现主机与存储设备之间的高可用性，高吞吐率。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

InPath for Windows

微软存储多路径解决方案允许存储设备制造商开发与自己产品相关的插件，实现对存储设备的优化访问，以及特色功能。此插件即 InPath for windows，InPath 与 Windows 操作系统的多路径驱动程序（mpio.sys）协同工作，实现存储多路径的各种操作。

Apsara 基于 Windows 平台开发的存储多路径模块 InPath，实现了 Windows 平台下访问 Apsara 存储设备的高可用性，高安全性，高吞吐率。提供了故障切换，故障恢复，路径实时选择，读写性能监控，日志跟踪等特色功能。InPath 支持多种接口协议，包括 FC，SAS，iSCSI 。

阿里云“InPath for Windows”与 Windows 平台多路径驱动程序协同工作，实现以下功能与特色：

- 存储多路径故障切换
  当一条路径发生 I/O 故障时，将 I/O 请求提交到另一条路径执行。
- 存储多路径故障恢复
  当一条路径从故障状态恢复时，实时将路径加入可用路径集。
- 主机集群功能支持
  主机集群功能是指多台主机以共享方式访问存储设备中的逻辑单元。
- 存储多路径负载均衡
  支持负载均衡算法，从可用路径集中选择最优路径，提高存储设备访问效率。目前支持以下负载均衡算法：
  - shortest queue service time
  - least IO weigth
  - least queue depth
  - least current outstanding bytes
  - round robin
  - round robin with subset
  - single path(first good path found)
- 负载均衡算法实时设置
  支持实时的多路径负载均衡算法设置，实时设置，实时生效。
- 路径实时开启/关闭
  支持实时的路径开关设置，实时设置，实时生效。
- 路径性能监控
  通过 Windows 系统性能监视器，实时监控路径 I/O 性能。或者通过 InPath 安装包中的工具，读取实时的路径 I/O 性能。
- Trace log 实时输出
  InPath trace log 包含 InPath 运行时的日志信息，方便查找导致存储系统故障或瓶颈的原因。
- Windows Event log
  InPath 输出标准的 Windows Event 日志，方便查找存储系统状态变化。

图 5-3 为 Windows 平台存储 IO 栈示意图，多路径模块位于存储 IO 栈的 Class 层。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 5-3 Windows 平台存储 IO 栈

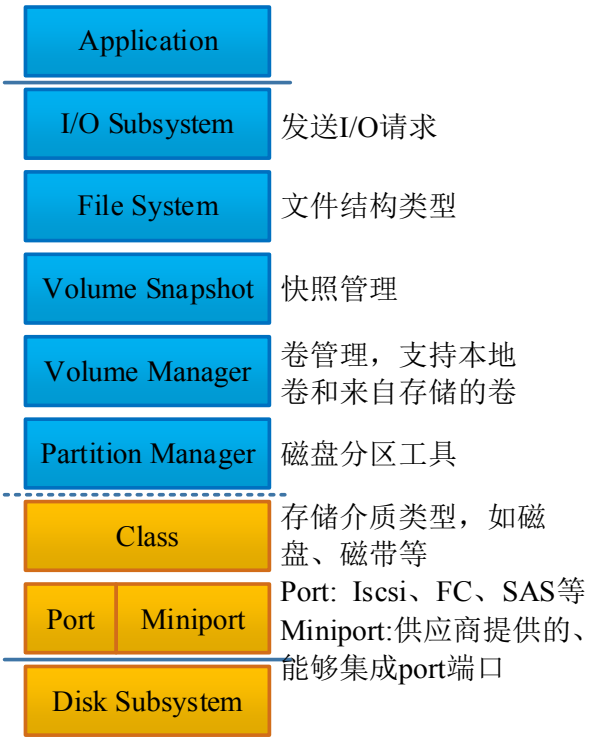
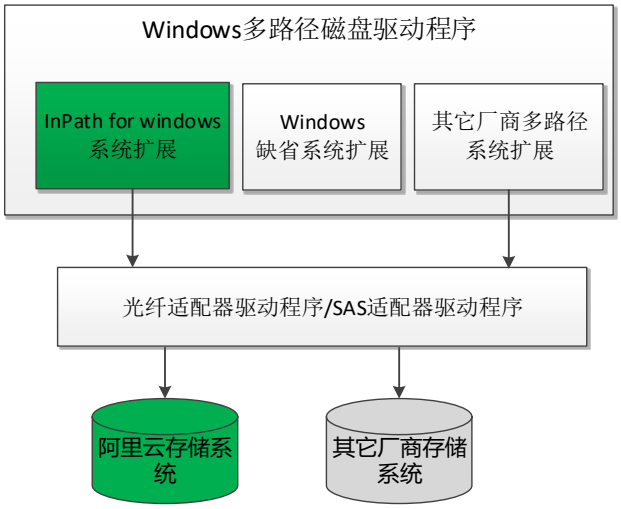


图 5-4 为 Windows 平台多路径模块示意图。多路径模块由多路径驱动程序（mpo.sys）以及 InPath 组成。多路径驱动程序可以同时与多个 InPath 协同工作，包括 Windows 平台提供的 MSDSM，Apsara 研发的 InPath，以及其它存储厂商的多路径程序。系统发现多路径存储设备后，会先询问系统中已经存在的非 MSDSM 是否支持发现的存储设备，如果支持，则关联存储设备与此 InPath 模块，如果所有非 MSDSM 都不支持此存储设备，则默认关联 MSDSM。

图 5-4 Windows 平台多路径模块



InPath for AIX

Apsara 基于 AIX 平台开发的存储多路径模块 InPath for AIX，实现了 AIX 平台下访问 Apsara 存储设备的高可

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

用性，高安全性，高吞吐率。提供了故障切换，故障恢复，路径实时选择，读写性能监控，日志跟踪等特色功能。

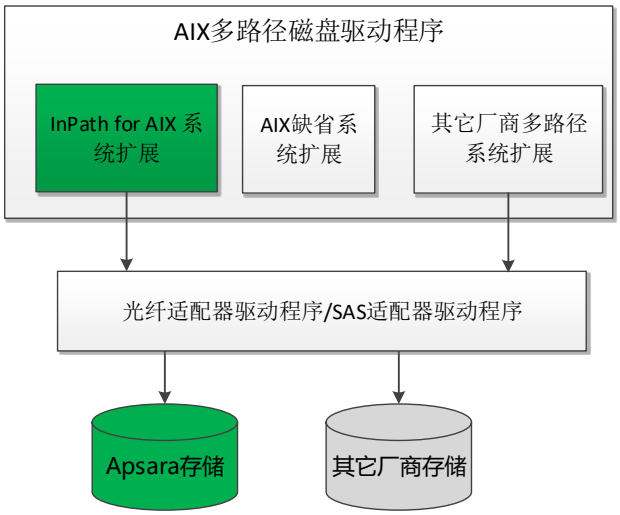
Apsara InPath 是一个系统内核扩展模块，通过配置文件配置相关支持参数，在运行时加载入内核，为 AIX 系统提供多路径相关的功能调用，实现对 Apsara 存储设备的多路径访问。

InPath 支持以下多路径功能及特色：

- 存储多路径故障切换
  当一条路径发生 I/O 故障时，将 I/O 请求提交到另一条路径执行。
- 存储多路径故障恢复
  当一条路径从故障状态恢复时，将路径加入可用路径集。
- 存储多路径负载均衡
  支持负载均衡算法，从可用路径集中选择最优路径，提高存储设备访问效率。目前支持以下负载均衡算法：
  - least IO weigth.
  - round robin.
  - single path(first good path found).
- 负载均衡算法实时设置
  支持实时的多路径负载均衡算法设置，实时设置，实时生效。
- 路径实时开启/关闭
  支持实时的路径开关设置，实时设置，实时生效。
- Trace log 实时输出
  InPath trace log 包含 SDDPCMcia 运行时的日志信息，方便查找导致存储系统故障或瓶颈的原因。

PCM 作为 AIX 系统内核扩展，其模块架构遵守标准的 AIX 内核扩展结构。InPath 提供了 AIX 系统所要求的所有接口调用。InPath 装载入内核后与相关模块的框架关系请见图 5-5 所示。

图 5-5 AIX 平台多路径模块



阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

Linux 多路径

Linux 系统平台通常会包含多路径模块：multipath-tools，根据多路径模块的版本号不同，支持阿里云存储的多路径配置也不同：

- 从版本 0.7.5 开始（包含 0.7.5），multipath-tools 的默认支持设备包含了阿里云存储设备，无需修改配置文件，即可支持本存储设备的多路径功能。
- 当 multipath-tools 模块的版本号低于 0.7.5 时，需要修改多路径模块的配置文件，实现对本存储设备的多路径功能支持。

在 Linux 系统下配置多路径工具 multipath-tools，用以支持阿里云存储设备多路径功能。multipath-tools 是 Linux 平台用于配置存储设备多路径支持的工具，其作用是配置 Linux 内核模块 device-mapper，实现对存储设备的多路径支持。此处以 Linux 发行版 RHEL 6.6（Red Hat Enterprise Level 6.6）为例，介绍在 Linux 平台配置阿里云系列设备多路径支持。

- RHEL 6.6 发行版本基于 Linux 2.6.32 内核版本，其多路径工具版本为 multipath-tools 0.4.9。
- 其它 RHEL 发行版本的多路径工具，与此例类似，请参考其对应的模块说明。
- 系统 CentOS、Oracle Linux、Kylin 3.2-4F、Ubuntu 与 RHEL 系统类似，请参考其对应的模块说明。
- 系统 SuSE 与 RHEL 存在差别，请参考文中备注对应的模块说明。

5.4 SRA

Storage Replication Adapter，作为 VMware SRM（Site Replication Manager）的适配器，使得 SRM 操作能够调用 Apsara 存储的智能远程复制功能。

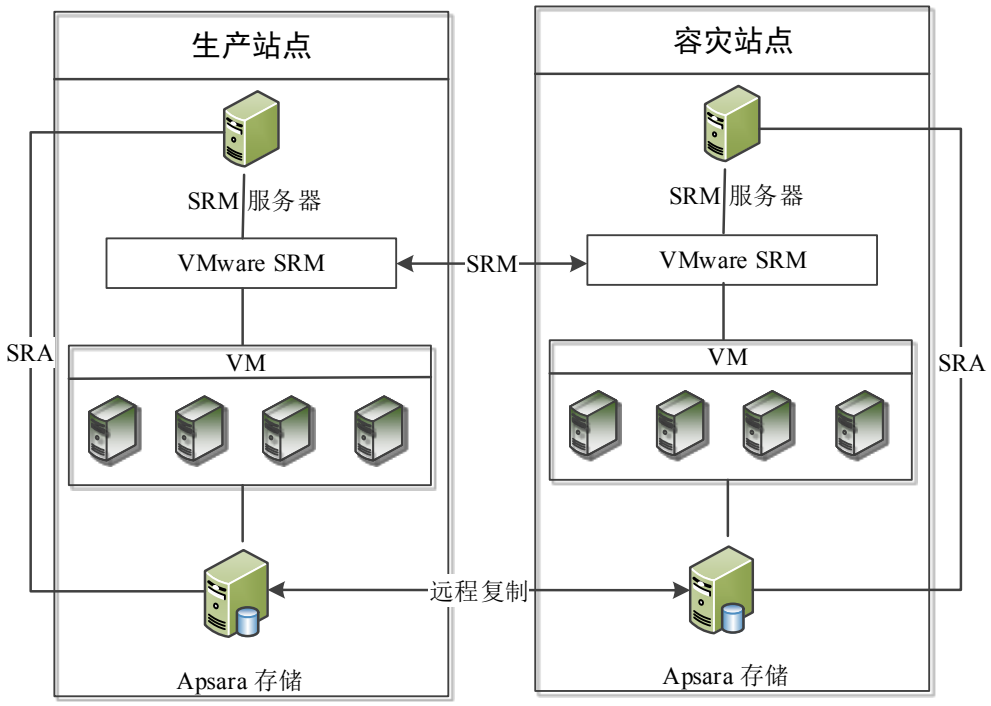
随着大数据的兴起，虚拟化应用越来越广泛，同时，虚拟化应用的数据安全性及业务的连续性、高可用性的要求，越来越高。

而 Apsara 系列存储，已通过 VMware 认证，与 VMware 进行了深度合作，在客户部署虚拟化解决方案时可与 Apsara 存储设备实现无缝连接。通过 VMware 及 Apsara 存储实现灾备管理，在灾难发生时，可以快速、安全的恢复，实现业务不中断，同时确保数据的安全。

例如，针对生产中心建立对应的容灾中心，分别安装 VMware SRM 及 Apsara 存储 SRA，通过 Apsara 存储之间的数据拷贝，来确保数据的一致性，如图 5-6 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 5-6 SRA 解决方案



在拥有成熟 VMware SRM 虚拟化容灾环境的前提下，通过 SRA 集成 Apsara 存储，确保数据的安全性、完整性及业务的连续性。其中，SRA 安装在 SRM 服务器上 ，通过 SRM 管理功能，发现存储阵列并配置主站点及备份站点之间的数据备份关系。针对不可避免的灾难(台风、地震、暴雨、停电等)，通过 SRM 提供的灾难恢复功能，切换到容灾站点服务器，以确保业务的连续性和数据的安全性。

Apsara SRA 目前支持两种配置模式：

- pre-configured 预先配置

  如果通过 SRA Utility 选中此配置方式，就需要预先配置好生产站点与容灾站点之前的映射关系。
- non pre-configured 非预先配置

  如通过 SRA Utility 选中此配置方式（默认为 non pre-configured），则不需要预先配置生产站点与容灾站点之间的映射关系，可以通过 SRA 建设这些关系（需要相应的管理权限）。

Apsara SRA 结合 VMware SRM 实现了以下功能：

- Test Failover

  测试灾备功能，执行此功能，可以测试灾备功能是否有效（pre-configured 及 non pre-configured 两种模式）。（SRM 调用 Apsara SRA 实现）
- Failover

  执行灾备功能，包括计划内及计划外两种：
  - 计划内，指在原有的计划内执行的灾备功能，在一些预知的风险或者故障发生时，通过 SRM 的 Failover 功能，切换到灾备站点，恢复业务的使用（SRM 调用 Apsara SRA 实现）。
  - 计划外，指一些非计划发生的自然灾害或者突然故障，造成整个保护站点/生产站点故障，不可访问，这样通过 SRM 的 Failover 功能，启用灾备站点，恢复业务的使用（SRM 调用 Apsara SRA 实现）。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

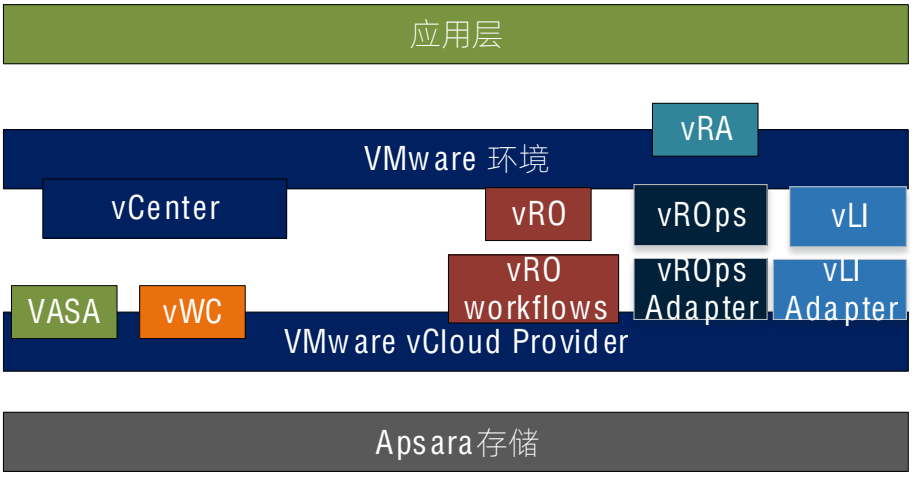
- 重新保护

  执行站点的重新保护。在生产站点发生故障，并且故障修复之后，重新保护生产站点，建立生产站点与灾备站点之间的关系（SRM 调用 Apsara SRA 实现）。

5.5 VCP

VCP（VMware vCloud Provider）是一个统一、集中式的服务器系统，用来整合一系列存储设备的存储资源调配（包括卷创建，调整大小，重命名，映射，取消映射，多路径策略实施和删除）、自动化及监控的解决方案。VCP 提供了一个单独的服务器后台，可以针对不同型号存储系统的存储资源，进行集中式的统一管理。VCP 主要针对 VMware 虚拟化平台，提供存储方面的对接解决方案，如图 5-7 所示。

图 5-7 VCP 解决方案



解决方案组件（VASA、vWC、vRO workflows、vROps Adapter、vLI Adapter）简称接口接口，均被包含在 VCP 软件包中，是 VCP 系统功能的一部分：

- Storage Enhancements for VMware vSphere Web Client

  集成在 VMware vSphere Web Client 平台上，可以让 VMware 管理员在该平台上独立、集中的管理其在 vSphere 环境中使用的本存储资源。

  vWC 插件有基于 Flex 技术和 Html 技术两种，两者在功能上相同，分别对应 VMware vSphere Web Client 上的 Flex 和 Html 入口，提供不同技术架构下的解决方案支持。
- Storage Provider for VMware VASA

  提升 VMware 平台针对存储的监控、自动化等操控能力，包含了对 VASA2.0 和 VVOL 的支持。另外，它使用了 SPBM(Storage Policy Based Management)技术用来最佳化虚拟机的配置流程。
- Storage Plug-in for VMware vRealize Orchestrator

  允许 VMware 管理员在 vRealize Orchestrator 中，通过插件，方便管理和使用本存储资源。
- Storage Management Pack for Vmware vRealize Operations Manager

  允许 VMware 管理员在 vRealize Operations Manager 中监视本存储各项性能指标。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- Storage Plug-in for VMware vRealize LogInsight

  允许 VMware 管理员在 vRealize LogInsight 中监视、接收本存储事件信息，并通过 dashboards 进行趋势分析。

5.6 Cinder 驱动

Cinder 是 OpenStack 三大核心组件（Nova、Neutron、和 Cinder）之一，此模块主要提供块存储服务，包含对卷、快照、备份、一致性组的操作及管理。

OpenStack 是一个开源的云计算管理平台项目，由几个主要的组件组合起来完成具体的工作。OpenStack 从 Folsom 开始使用 Cinder 替换原来的 Nova-Volume 服务，为 OpenStack 云平台提供块存储服务。

Cinder 的功能实现依赖于存储设备，需要通过相应的驱动与存储设备进行连接，存储设备才能为 Cinder 提供服务。Cinder 的主要功能包括 卷、快照、拷贝、一致性组等一系列的操作，如增删改查等。Cinder 进行相应的操作时，会调用相应的驱动接口，驱动通过 ssh 连接存储设备执行相应的命令，进行相应的操作。Apsara 存储可以通过 FC 或者 iSCSI 进行与 OpenStack 虚拟机的进行连接。

Cinder 通过 CLI、Dashboard 发起请求命令，通过 scheduler 的调度最终会落到 Volume 的控制器上，并通过 Volume 调用厂商的驱动进行卷、快照、备份等的操作。Cinder 是 OpenStack 的其中一个组件主要负责存储的功能。它是用来给用户提供在 nova 计算中的存储功能。Cinder 主要分为三大模块，API、Scheduler、Volume。

API service：Cinder 构架图的核心部分，负责接受和处理 Rest 请求，并将请求放入 RabbitMQ 队列。

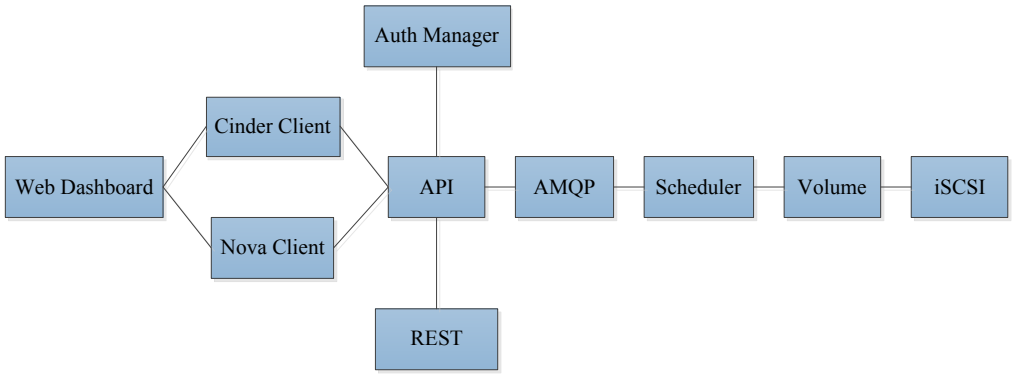
Scheduler Service：处理任务队列的任务，并根据预定策略选择合适的 Volume Service 控制器来执行任务。所以构架图中，Scheduler 一端连接 MQ 队列，处理 MQ 队列中的任务；一端连接 Volume 设备，任务在 Volume 控制器上执行。

Volume Service：该服务运行在存储控制器上，管理存储空间。每个存储控制器都有一个 Volume Service，若干个这样的存储控制器联合起来可以构成一个存储资源池。

Cinder 的架构如图 5-8 所示。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

图 5-8 Cinder 架构图



OpenStack 云存储架构相比传统存储架构的优势主要体现在以下几个方面，如表 5-1 所示。

表 5-1 OpenStack 云存储架构和传统存储架构对比

方向	传统存储架构	openstack 云存储架构
海量数据承载能力	扩展方式是通过增加硬件配置实现，属于 scaleup 方式	存储平台可以达到 PB 级别的扩展空间更合适海量数据的存储、scaleout
高可用	昂贵的硬件保证系统的高可用	通过系统自身的机制，即软件完成的自动化、智能机制来保证系统可用性
存储资源动态调配的能力	存储资源分配给应用后，难以回收再分割	计算和存储资源虚拟化，可以按照需求分配，动态调整
	低资源利用率，高能耗	35%-75%的 TCO 节省，30%以上的软件成本节省，CPU 利用率提升到 60%-80%，70%-80%运营成本节约
运维效率和成本	运维效率低，维护成本高，硬件准备周期长	部署时间缩短到分钟级，减少硬件准备周期
管理复杂度	高	低

5.7 Cinder-backup 驱动

Apsara 存储 Cinder-backup 主机插件是 OpenStack 系统中 Cinder 服务中负责备份的驱动程序，实现了对 Cinder 卷进行备份以及备份管理的功能。功能包括对 Cinder 卷创建备份、删除备份、利用备份恢复 Cinder 卷数据。Cinder-backup 驱动向上实现了 OpenStack 的相关接口，完成与 OpenStack 平台的对接；向下连接本存储，通过调用相关 CLI 命令利用存储快照完成备份操作。Cinder-backup 驱动支持在不同的存储设备中创建备份，即支持源卷与备份卷在不同的存储系统中。

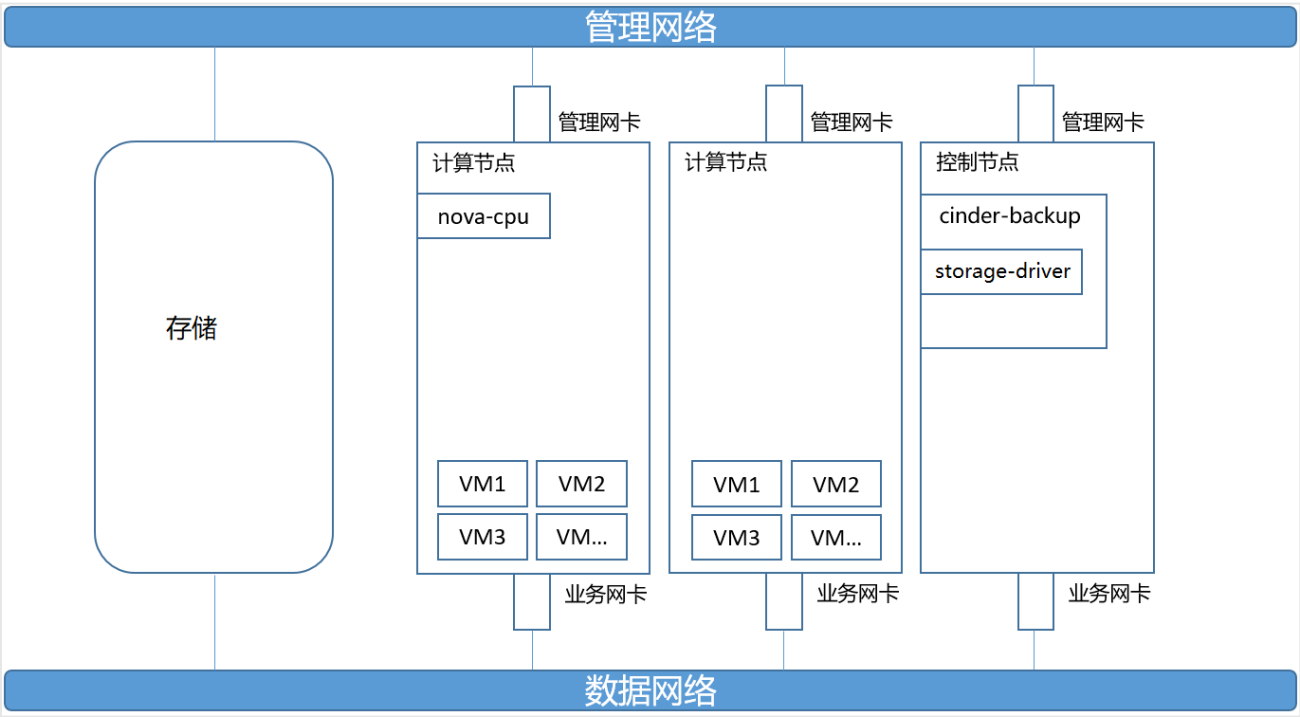
阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

通过配置该驱动，Cinder-backup 可以使用本存储的快照功能完成备份的创建、删除、回滚操作，详细功能如表 5-2 所示。

表 5-2Cinder-backup Driver 实现的功能

序号	功能模块	描述
1	创建备份	对 Cinder 卷创建一个备份
2	删除备份	删除 Cinder 卷的一个备份
3	利用备份恢复数据	利用备份，将 Cinder 卷恢复至过去某一时刻的状态

图 5-9 Cinder-backup 应用拓扑图



5.8 K8sCSIPlugin

K8sCSIPlugin 插件使得阿里云存储可以为 Kubernetes 集群中的应用自动的提供持久化存储。K8sCSIPlugin 插件功能主要包含卷的创建、删除、挂载、卸载、克隆，快照的创建、删除。

当 Kubernetes 集群中的一个服务需要使用阿里云存储提供的卷时，用户会首先在 Kubernetes 上创建一个 PVC(Persistent Volume Claim)资源信息，Kubernetes 根据 PVC 中的信息，通过接口，将消息传递给 instorage-csi，instorage-csi 收到消息后会自动的在存储上创建满足要求的卷，并在 Kubernetes 上生成一个 PV（Persistent Volume）信息，之后服务就可以使用该卷。当 Kubernetes 删除 PV 信息时，消息也将传递给 instorage-csi，instorage-csi 根据里面的具体信息，自动的将存储上与之对应的卷删除。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

当 Kubernetes 集群中的一个使用持久化存储的服务启动时，首先需要将服务所使用的持久化存储映射到某个主机的指定目录上，然后服务才可以使用这些持久化存储。在这个过程中，Kubernetes 会调用与相应 PV 信息对应的卷挂载/卸载插件，K8sCSIPlugin 插件中也实现了该部分功能。当 Kubernetes 需要挂载一个卷时，instorage-csi 会自动的在存储上完成卷与主机信息的绑定，并在主机端将对应的卷扫描出来，根据需要，会自动对卷进行格式化，然后将卷上的文件系统挂载到 Kubernetes 指定的目录上，随后，Kubernetes 便可以将该目录映射到服务的容器中，供服务使用。

当前插件实现的详细功能，如表 5-3 所示。

表 5-3 K8sCSIPlugin 插件实现的功能

功能模块	操作
卷-主机操作	挂载卷到主机
	mount 卷上的文件系统到主机
	从主机上 unmount 卷上的文件系统
	从主机上卸载卷
存储接口协议	iSCSI
	FC
多路径	支持多路径设备
卷操作	创建卷（支持普通卷、镜像卷、双活卷）
	删除卷
	克隆卷（支持基于卷克隆和基于快照克隆）
	卷在线扩容
快照操作	创建快照
	删除快照

5.9 Manila 驱动

Apsara 存储的 Manila 驱动是 OpenStack 系统中 Manila 服务的驱动程序，实现了通过 Manila 服务对存储进行管理。Manila 项目全称是 File Share Service，文件共享即服务，是 OpenStack 系统中的一个组件，用来提供云上的文件共享服务，支持 CIFS 协议和 NFS 协议及其他协议。它的作用是向 OpenStack 环境中的虚拟机提供文件共享服务，通过 Manila 调度底层的 Driver，配合使用本存储，使本存储的文件共享服务接入到

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

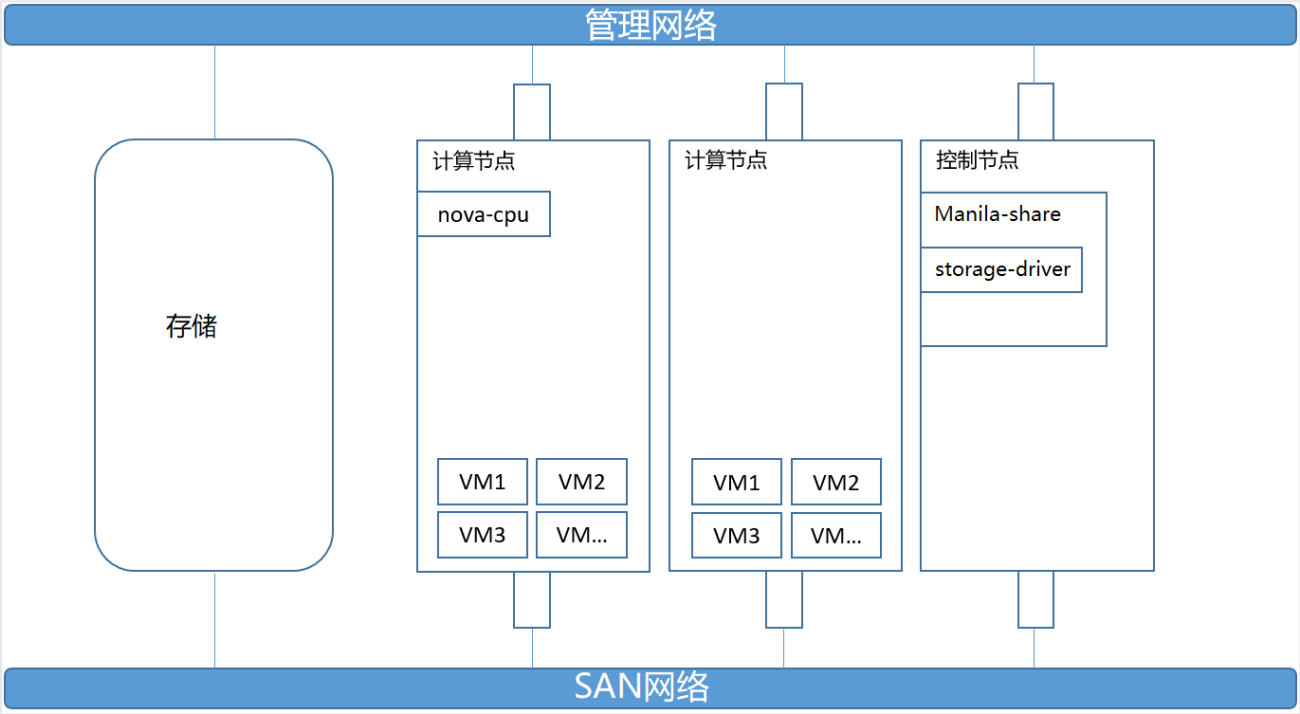
OpenStack 中，Manila 配合 Neutron 服务，配置网络使虚拟机可以访问到存储的文件共享服务。阿里开发了上述中的插件形式的驱动，通过配置驱动插件并管理存储，Manila 服务可以使用本存储设备为用户提供所需的文件共享服务，使 OpenStack 环境中的虚拟机可以使用本存储的 NFS 和 CIFS 共享服务。

通过配置该驱动，Manila 可以使用本存储为 OpenStack 提供文件共享服务，并且可将 Manila 的共享、共享权限的操作在存储上实现，详细功能如表 5-4 所示。

表 5-4 Manila 驱动实现的功能

序号	功能模块	操作
1	共享操作	创建共享
2		删除共享
3	共享--容量	共享扩容
4	共享--权限操作	更新共享权限

图 5-10 Manila 驱动应用拓扑图



阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

6 人性化管理、维护

6.1 统一智能存储管理平台

InView 是面向 Apsara 存储平台的多设备智能管理软件，采用大数据和 AI 技术，构建智能化运维管理平台，实现对存储平台自动化部署、状态监测、容量预测、性能优化、远程巡检、故障诊断、资产管理、“两地三中心”容灾等智能化运营，解决 IT 系统运维难题，提高运维效率，保证数据的可靠性和可用性，降低运维成本，优化用户体验，保障业务连续性，为客户提供一体化的智能运维综合解决方案。

Inview 主要特征与优势如下：

- 多套存储平台统一管理
  - 统一入口，存储平台精细管理
  - 嵌入式设计，对外始终只呈现一个系统页面
  - 标签栏切换式操作，多设备间自如切换
- 容量智能预测
  - 对存储的集群和池的已用容量在未来的变化趋势进行预测
  - 容量 1-90 天预测
  - 提供容量过载告警，避免紧急扩容
  - 指导用户根据当前业务合理扩容
- 性能智能预测
  - 集群存储的 CPU 利用率，读写缓存利用率，读写 IOPS，读写带宽，读写时延的性能趋势预测
  - 性能 1-90 天预测
  - 设备的负载状态，发现瓶颈，合理规划设备配置
- 智能化存储平台管理
  - 中英文图形化界面展示，全局信息集中展示
  - 告警信息汇总，故障分析定位
  - 容量智能预测，性能瓶颈分析
  - 快照增强，分钟级数据自动备份
  - 多副本复制，轻松构建跨设备容灾

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

6.2 事件监控和维护向导

本存储系统提供的统一图像化管理界面，可以通过本机浏览器输入管理地址，实现对存储的管理、查看和配置。这个功能提供了在线的存储状态监控，并能够及时反馈多项重要的健康信息，主要包括完整的失败信息反馈、用户操作信息或者进程状态改变等信息，随时反馈各项指标，直观提示，有效保障当前存储平台运行情况。

DMP 维护向导是针对事件信息中的故障或错误事件提供的修复向导，涉及存储平台的软件问题和硬件问题，实现智能化故障恢复功能。向导的步骤详细、准确、通俗易懂，允许用户自己根据向导修复并解决问题。

事件监控界面提供了三种信息筛选选项：

- 推荐操作

  只显示要求关注的警告信息，警告是按紧急程度的优先级呈现的，然后顺序地用可用的修复程序进行修复。选中每一项错误，用户可以进行如下两种操作：使用修复程序或者查看相关属性。
- 未修复信息和警告

  显示了未被修复的警告和相关信息，同样用户可以进行如下操作：运行修复程序；将事件标记为已修复；过滤信息以分钟、小时或者天来显示；重置日期过滤器；查看属性。
- 显示所有

  无论是否已修复或者未修复，全部事件显示。同样可以进行如下操作：运行修复程序；将事件标记为已修复；过滤信息以分钟、小时或者天来显示；重置日期过滤器；查看属性。

Apsara 存储能够迅速地向用户反馈问题报告，并进行相关修复。除了图形化管理界面可以提供事件信息反馈，用户还可以选择其他通知方式，有如下几种通知机制可供使用：

- 邮件

  一个事件的通知信息能够被发送给一个或多个邮箱地址，这种机制提供灵活的方式去通知管理员，包括移动终端，只要能够查看邮件，就能够接收到存储平台有关事件的通知
- SNMP

  一个简单的网络管理协议报告会被发送给数据中心管理系统，让管理员进行处理，管理员就能整合多个系统的 SNMP 的报告。通过这种机制，用户能在单一的工作站点对数据中心进行监控。
- 系统日志

  系统的日志报告会被送往数据中心管理系统，用于整理来自多个系统的日志报告，这样用户就可以在一个地点监控整个数据中心的情况。

6.3 STAT 工具

智能分层（InTier）功能是 Apsara 存储的软件特性之一，可以自动的将经常被访问的数据（热点数据）从低转速硬盘 HDD 迁移到高性能高转速硬盘 SDD 中。

为了配合使用和分析智能分层的功能，Apsara 同时提供了一个工具 STAT，集成在 GUI 管理界面。主要是分

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

析存储平台生成的热点数据文件，以一种直观的方式(HTML)将使用智能分层功能后的效果（将热点数据从 HDD 迁移到 SDD）展示出来，给用户提供有效的数据支持。

STAT 工具可以针对使用了分层功能的存储池、卷，生成数据分析报告，内容有：工作负载分类、工作负载偏差曲线、每日数据移动报告。

工作负载分类

工作负载分类提供了一种方法，可以对 InTier 管理的数据活动进行分类，此分类允许以更通俗易懂的方式来重新定义和比较各个存储层中的热数据。

工作负载分类有以下作用：

- 随着不断变化的热数据的使用情况，热数据衡量的标准变得越来越稳定，能够长期以一致的标准判断并提供数据活动变化信息。
- 显示并比较存储池中的不同存储层中数据活动的信息。
- 帮助避免发生对数据工作负载评估出现高偏差的情况。
- 使数据更易于解释，例如显示 InTier 可以识别并提升到 SSD 的活动数据。
- 依据偏差曲线中的比较结果，可以为工作负载分配提供协作视图。

工作负载偏差曲线

InTier 支持工作负载统计收集数据信息，工作负载偏差曲线演示了工作负载 I/O 活动的偏差。

工作负载分类以每 5 分钟为一个周期来跟踪 InTier 活动（升级、降级、交换、自动调整以及 InTier 程序本身）。工作负载偏差曲线随后汇总每存储层上的工作负载分布，以帮助判断是否需要进行跨层的数据迁移。

工作负载曲线上的信息有以下作用：

- 提供存储层的配置信息。根据工作负载偏差，可以使用曲线信息来识别哪些 SSD 配置应该处理预期的工作负载。此信息还可以提供指导信息来确认需要多少 SSD 容量才能处理一定比例的工作负载。InTier 按照热点以固定颗粒度 对降序排列的数据进行迭代，并将累计的工作负载汇总到热点文件报告中。
- 验证 InTier 的行为。例如，根据当前的 InTier 算法，当包含的 I/O 活动不是很大的时候，SSD 层很可能包含较小的 I/O 活动，HDD 层包含较大的 I/O 活动。

每日移动数据报告

每日数据移动报告提供有关 InTIer 管理的工作负载如何随时间变化的信息。

每日数据移动报告有以下作用：

- 验证 InTier 如何报告配置变化：
  - 当添加新层时（例如 SSD），期望看到升级到 SSD 存储层的热数据。
  - 当使用 SAS 层添加新的 NL-SAS 层时，期望看到数据降级到 NL-SAS 层。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

- 添加容量扩展现有层时，期望看到数据在现有层的自动调整变化。
- 验证 InTier 如何报告工作负载变化：
  - 当数据工作负载需要迁移到较高存储层数据块中时，如果较高存储层超负荷，应期望看到更多数据或“稍冷”的热数据需要升级。
  - 当数据工作负载需要移出较高存储层数据块中时，应期望看到更多的数据需要降级。

6.4 Comprestimator 工具

Comprestimator 是一种卷压缩率评估工具，集成在 GUI 管理界面。能够评估出某个卷在开启实时压缩功能后可节省的空间比例。可用于评估特定环境的预期压缩的优势以及评估某一存储介质的预期压缩率。由于该工具采用先进的数学和统计算法进行取样和分析过程，根据卷中的数据类型分析得出这个卷的压缩率，在分析数据存储量大、利用率很高的卷空间时，其预估的压缩率结果会更加准确。

在分析传统卷（足额卷）时，重要的是需要了解存储块设备的行为。传统卷在创建时，没有自动归零旧数据块设备的情况下，数据在文件系统级别将不能被访问和可视化。当使用 Comprestimator 分析这样的卷，预期的压缩结果可以反映这个数据块真实的压缩率，包括旧数据的痕迹。这个模拟卷镜像过程的分析设备为一个压缩卷。之后，当卷镜像用于压缩存储平台上的数据时，它将处理设备上的所有数据(包括活动数据和旧的痕迹数据)并压缩。当在压缩卷上存储更多的活动数据时，旧数据的痕迹将开始被新的数据写入到该卷中。由于设备上积累的活动数据更多，压缩率将进行调整，来反应活动数据的准确结果。

无论对哪一类型的块设备扫描，同样重要的是需要了解一些常见的文件系统空间管理的特征。当文件从文件系统中删除，删除前他们所占据的空间将被释放,即使数据在磁盘上的文件系统并不是真正删除，而是文件系统索引和指针被更新来反映这种变化。当使用 Comprestimator 来分析被文件系统所使用的块设备时，所有在设备上的相关数据都将会被分析，无论该数据所属的文件是否已经从文件系统中删除。例如：你可以填充一个 100GB 的文件系统，使它的使用率是 100%，然后再删除文件系统中所有的文件，使它的使用率为 0。当扫描存储文件系统的块设备时，Comprestimator 也将访问那些已经被我们删除的文件数据。

为了减少块设备和文件系统行为的影响，使用 Comprestimator 分析包含尽可能多的活动数据的数据卷，而不是空数据卷。这将提高精度水平，并且减少对已删除但仍然可能在设备留下了痕迹的旧数据的分析风险。

6.5 遵循 SMI-S 规范的 CIM 接口

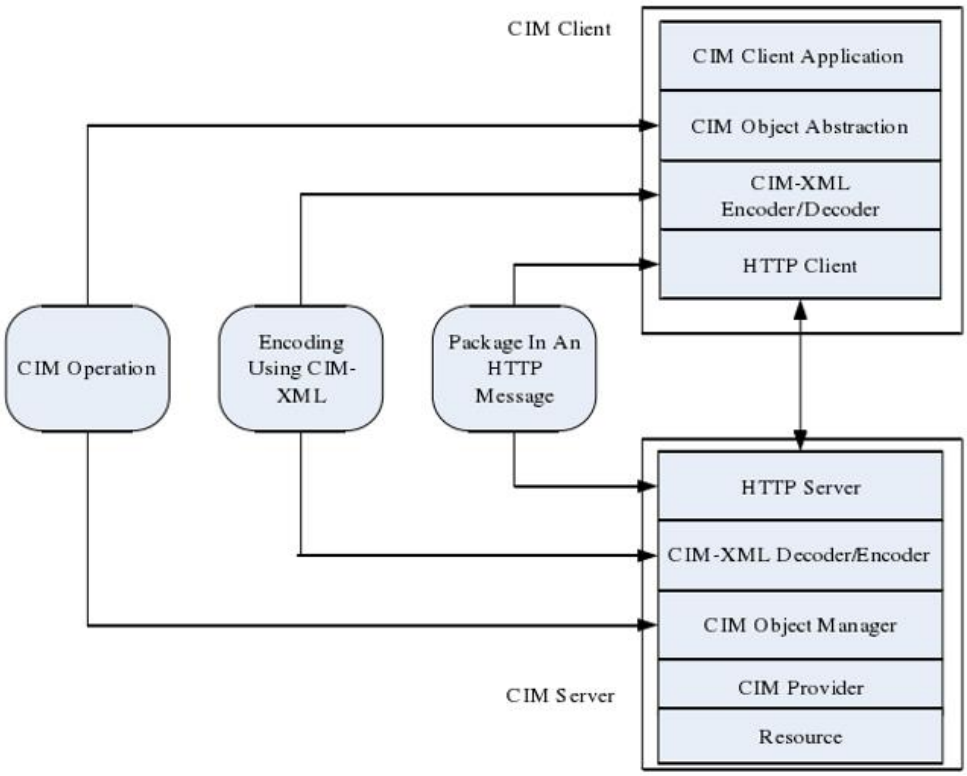
CIM（Common Information Model，公共信息模型）是由 DMTF 组织提出的一个开放（开源）标准。它定义了 IT 环境中如何对象化地表现受管对象以及受管对象之间的关系。其目的是保持系统管理的一致性，不依赖于制造商和提供商。SMI-S （Storage Management Initiative Specification，存储管理倡议标准）是 SNIA 发起和主导众多存储厂商共同参与开发的一种 CIM 标准管理接口。

Apsara 系列存储提供了遵循 SMI-S 规范的 CIM 接口，并已通过 SNIA 官方认证。Apsara 存储提供的 CIM 接

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

口，功能涵盖集群管理、主机管理、存储资源管理（池、卷、硬盘）、性能监控、数据保护等功能。CIM 客户端通过调用 Apsara 存储提供的 CIM 接口，实现对 Apsara 存储的监控和管理。

图 6-1 CIM 架构



- CIM Client 来发起 CIM 操作(CIM Operation)，请求或接收并处理 CIM 操作的返回。
- CIM Server 接收并且处理 CIM 操作的请求，并将 CIM 操作的结果返回给 CIM Client。
- 在 CIM Client 与 CIM Server 之间有一个 CIM 对象抽象层（CIM Object Abstraction Layer），该抽象层负责实现无具体协议依赖性的 CIM 对象和 CIM 操作。在 Client 端该层表现为 CIM 对象抽象（CIM Object Abstraction）；在 Server 端表现为对象管理器（CIM Object Manager， CIMOM）。
- CIMOM 是所有 CIM 类实例的一个数据库，所有实现的类在这里都有体现，是访问管理资源的重点。

6.6 SNMP 接口

SNMP（ Simple Network Management Protocol）是基于 TCP/IP 协议族的网络管理标准，是一种在 IP 网络中管理网络节点（如服务器、工作站、路由器、交换机等）的标准协议。

本存储提供了遵循 SNMP 规范的自定义接口供第三方系统或软件调用，用于管理或监控存储设备。SNMP GET/SET 支持 v1、v2c、v3 三个版本，Trap 支持 v2c、v3 两个版本。

- 通过 SNMP 接口，可以获取如下信息：
  - 集群状态：包括 CPU 使用率、集群状态。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

  - 容量信息：包括硬盘、MDisk、存储池容量信息。
  - 组件信息：包括 BBU、风扇、PSU、控制器信息。
  - 端口状态：包括 iSCSI、SAS、FC 端口状态信息。
  - 硬盘状态：包括硬盘信息、利用率统计信息。
  - IP 配置：包括集群 IP、网关、掩码信息。
- 通过 SNMP 接口，可以设置如下信息：
  - 集群 IP 配置。
  - 集群名称。
  - 系统时间。
- 当存储系统产生告警事件时，主动上报包含此事件的 SNMP TRAP 报文，包含告警级别及事件内容。

6.7 RESTful 接口

REST（Representational State Transfer，表述性状态转移） 指的是一组架构约束条件和原则，满足这些约束条件和原则的应用程序或设计就是 RESTful。

RESTful 具有如下特点：

- 每一个 URI 代表 1 种资源。
- 客户端使用 GET、POST、PUT、DELETE 4 个表示操作方式的对服务端的资源进行操作。
- 通过操作资源的表现形式来操作资源。
- 资源的表现形式是 XML 或者 HTML。
- 客户端与服务端之间的交互在请求之间是无状态的，从客户端到服务端的每个请求都必须包含理解请求所必需的信息。

Apsara 系列存储提供了符合 REST 约束的接口供第三方系统或软件调用，用于管理或监控存储设备。

通过 RESTful 接口可以操作存储系统如下信息：

- 安全类：查询、修改、创建、删除用户。
- 设备类：查询机柜、控制器、硬盘、电池、电源、风扇、端口信息。
- 集群类：节点、集群信息。
- 事件告警类：获取系统事件信息。
- 块存储类：获取 MDisk、卷、存储池、主机信息。
- 性能类：获取集群实时性能数据。

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

7 缩略语

B		
BBU	Battery Backup Unit	电池备份单元
C		
CIFS	Common Internet File Systems	通用互联网文件系统
CLI	Command-Line Interface	命令行界面
CPU	Central Processing Unit	处理器
D		
DMP	Directed maintenance procedures	维护向导
DRAID	Distribute RAID	分布式 RAID
F		
FC	Fibre Channel	光纤通道
FTP	File Transfer Protocol	文件传输协议
G		
GUI	Graphical User Interface	图形用户界面
H		
HDD	Hard Disk Drive	硬盘驱动器
I		
I/O	Input/Output	输入/输出
iSCSI	Internet Small Computer System Interface	网络小型计算机系统接口
L		
LUN	Logical Unit Number	逻辑单元号

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司

M		
MPIO	Multi-Path I/O	多路径 I/O 管理
N		
NAS	Network Attached Storage	网络附加存储
NFS	Network File System	网络文件系统
NL-SAS	Near Line SAS	近线 SAS
P		
PSU	Power Supply Unit	电源装置组件
R		
RACE	Random Access Compression Engine	随机访问压缩引擎
RAID	Redundant Array of Disks	独立磁盘冗余阵列
S		
SAN	Storage Area Network	存储区域网络
SAS	Serial Attached SCSI	串行 SCSI
SCSI	Small Computer System Interface	小型计算机系统接口
SMI-S	Storage Management Initiative specification	存储管理接口标准
SNMP	Simple Network Management Protocol	简单网络管理协议
SSD	Solid State Disk	固态硬盘
SSH	Secure Shell	安全外壳协议
T		
TCO	Total Cost Ownership	总体拥有成本

阿里云专有和保密信息
版权所有 © 阿里云计算有限公司