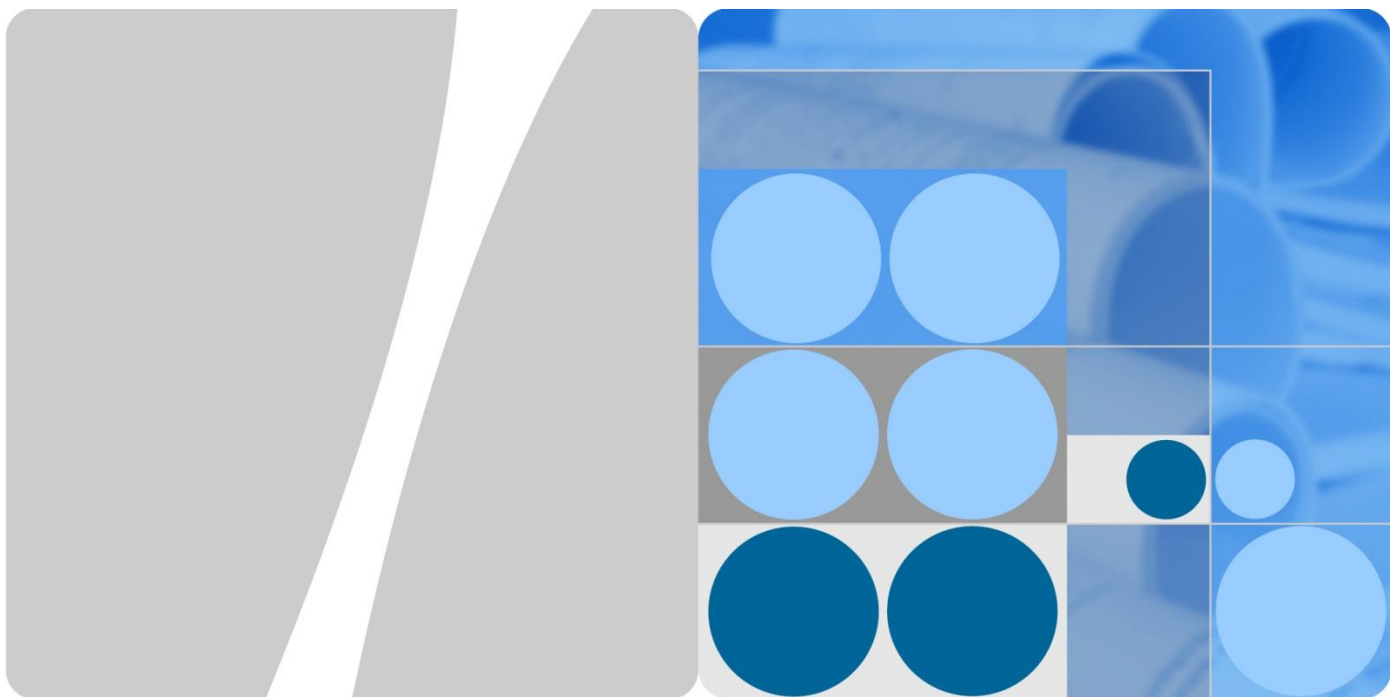




华为存储整合解决方案技术白皮书

文档版本 1.0
发布日期 2015-03-31

华为技术有限公司



版权所有 © 华为技术有限公司 2015。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为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，华为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。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。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： <http://enterprise.huawei.com>

目 录

目 录.....	ii
插图目录.....	iv
1 概述.....	1
1.1 运营商 BSS 业务存储面临的挑战	1
1.1.1 资源利用率低，管理复杂	1
1.1.2 高可用性需求	2
1.1.3 持续扩容难	2
1.2 存储整合场景分析	3
1.3 存储整合方案概述	3
1.4 方案亮点	4
1.5 主要内容	4
2 方案架构.....	5
2.1 旧设备替换方案架构	6
2.2 存储集中化方案架构	7
2.3 存储虚拟化与利旧方案架构	8
2.4 方案组件	9
2.4.1 关键组件介绍	10
3 关键技术.....	12
3.1 异构虚拟化技术	12
3.1.1 SmartVirtualization 工作原理	13
3.1.2 SmartVirtualization 技术特点	1
3.1.3 SmartVirtualization 应用场景	2
3.2 数据迁移技术	5
3.2.1 基于异构虚拟化的数据迁移方案	5
3.2.2 基于存储功能（LUN 拷贝）的数据迁移方案	9
3.2.3 基于存储功能（远程复制）的数据迁移方案	13
3.3 服务质量保障技术	16
3.3.1 SmartQoS 工作原理	1
3.3.2 SmartQoS 技术特点	7

3.3.3 SmartQoS 技术应用场景 7

3.4 卷镜像技术..... 8

3.4.1 HyperMirror 工作原理..... 8

3.4.2 HyperMirror 技术特点10

3.4.3 HyperMirror 应用场景.....10

3.5 缓存分区技术.....12

3.5.1 SmartPartition 工作原理13

3.5.2 SmartPartition 技术特点18

3.5.3 SmartPartition 技术应用场景.....18

3.6 多租户技术.....18

3.6.1 多租户工作原理18

3.6.2 多租户技术特点19

4 缩略语21

插图目录

图 1-1 存储整合场景分析	3
图 2-1 旧设备替换方案架构图.....	6
图 2-2 旧设备替换方案架构图.....	7
图 2-3 旧设备替换方案架构图.....	8
图 2-4 华为 OceanStor 5800/6800 V3 系列.....	10
图 2-5 华为 OceanStor 18000 系列.....	11
图 3-1 异构虚拟化工作原理	13
图 3-2 异构阵列接管	2
图 3-3 异构数据搬迁	3
图 3-4 异构容灾	4
图 3-5 异构数据保护	5
图 3-6 异构虚拟化数据迁移方案架构图	5
图 3-7 异构虚拟化迁移步骤示意图	6
图 3-8 LUN 拷贝数据迁移方案架构图	9
图 3-9 LUN 拷贝迁移步骤示意图.....	10
图 3-10 远程复制数据迁移方案架构图	13
图 3-11 远程复制迁移步骤示意图	14
图 3-12 IO 优先级调度原理图	2
图 3-13 并发资源优先级分配流程图	3
图 3-14 CPU 资源优先级调度流程图.....	4
图 3-15 页面分配优先级调度流程图	5
图 3-16 页面分配优先级调度流程图	6
图 3-17 镜像 LUN 的创建过程.....	8
图 3-18 同步过程的原理	9
图 3-19 分裂过程的原理	10

图 3-20 本地 LUN 数据镜像 11

图 3-21 异构 LUN 可靠性 11

图 3-22 异构阵列间镜像 12

图 3-23 SmartPartition 工作原理 14

图 3-24 写缓存后对 I/O 过程的处理 15

图 3-25 配置读缓存后对 I/O 过程 16

图 3-26 SmartPartition 在系统中的分布 17

图 3-27 多租户管理面 19

1 概述

关于本章

- 1.1 运营商 BOSS 存储系统面临的挑战
- 1.2 存储整合场景分析
- 1.3 存储整合方案概述
- 1.4 方案亮点
- 1.5 主要内容

1.1 运营商 BSS 业务存储面临的挑战

运营商业务云化发展趋势可分为三个步骤：集中化，虚拟化，云化。目前国内及海外运营商 IT 支撑系统的部署呈现出集中化的趋势。进一步的集中化，能够集中企业的支撑能力、服务能力和创新能力，实现企业 IT 能力的重点突破；同时，更高层次的集中易于实现低成本高效运行，易于打造企业内部专业化 IT 服务。

随着 IT 支撑系统集中化建设规模的不断扩大，系统所需硬件资源越来越多，机器数量规模越来越庞大，占用面积越来越大，产生很多技术问题。其中存储系统也面临几大挑战：资源利用率低，管理复杂；可用性要求高，需保障业务连续性和数据完整性；数据量每年翻倍激增，持续扩容难。

1.1.1 资源利用率低，管理复杂

传统存储系统基于烟囱式建设模式，CBS、CRM、BI 等功能部件采用各自独立的存储部署方式，互相之间互为孤岛。各系统独自根据业务峰值规划系统资源需求，资源配比不灵活，无法共享，资源利用率低。

- 容量和性能需求不平衡：如事务型 CBS、CRM 业务要求高性能，配置大量硬盘后存储容量利用率不足 35%。分析型 BI 业务要求大容量，硬盘的性能资源不能充分利用。容量和性能需求无法平衡，导致大量硬盘资源浪费。
- 冷数据长期存在：如 CRM、BI 系统，基于长期数据查询和统计分析的需要，占比 60% 以上的话单等历史数据长期占用大量高性能硬盘空间。

- 存储系统分散部署：多套存储系统分散部署，资源无法有效均衡利用，扩容频繁，组网复杂，且各厂商的存储管理软件互不兼容，增加了管理和维护成本。

1.1.2 高可用性需求

存储承载着运营商关键的计费、营业系统，必须确保业务连续性和数据完整性。

- 业务中断代价昂贵：运营商核心业务系统，业务连续性至关重要，存储需提供从架构、硬件、软件方案级的整体可靠性设计。
- 数据不容丢失：业务系统包含重要的用户资料、账单及运营数据，而传统 RAID 技术存在双盘失效、数据重构时间长等风险。
- 关键业务存储资源无法保障：存储系统缺乏资源 Qos 管控，关键业务高峰期资源被抢占，造成用户体验差甚至业务中断。

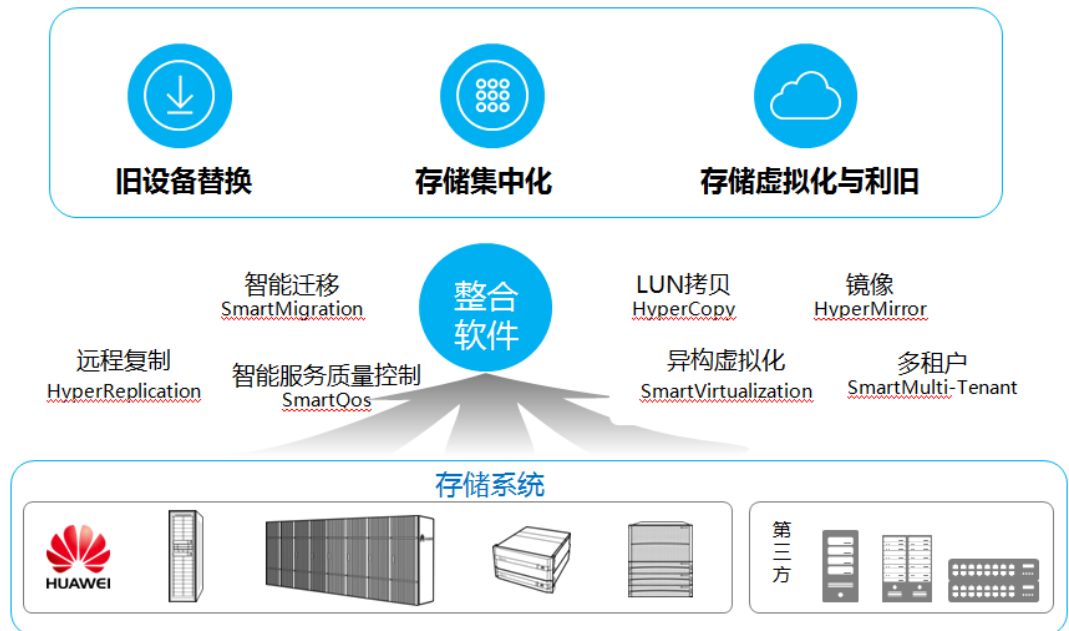
1.1.3 持续扩容难

随着用户规模及数据类新业务的爆发式增长，BSS 系统数据量每年翻倍激增，扩容和老旧替换机会不断。

- 已有存储系统无法满足扩容需求：已有存储受制于控制器规格及架构限制，扩容能力有限，无法满足数据快速增长带来的长期、持续性扩容需求。
- 存储架构不能支撑容量和性能的线性增长：传统存储系统扩容后性能和容量不能线性同步增长。
- 前期投资无法得到保护：扩容存储无法兼容和接管已有存储，造成前期投资浪费。

1.2 存储整合场景分析

图1-1 存储整合场景分析



华为存储整合解决方案主要针对运营商 BSS 业务，可分为三大场景：

旧设备替换场景：采用一套中高端存储系统替换一套过保的老旧存储系统。提供更高性能、更好扩展性的存储。

存储集中化场景：采用一套高端存储，一次性替换现网多套老旧存储系统，实现存储集中化。简化管理，消除多厂商异构问题，提供专业的数据迁移能力，提高资源利用率。

存储虚拟化与利旧场景：采用一套高端存储，虚拟现网各厂商异构存储，形成统一存储资源池。广泛兼容异构厂商，实现现网存储利旧，简化管理。

1.3 存储整合方案概述

华为存储整合解决方案根据运营商 BSS 业务三大场景，分别采用华为 OceanStor V3 和 18000 系列产品整合客户现网设备。以高性能高规格的 OceanStor V3 或 18000 设备应对一对一旧设备替换场景；以高端 OceanStor18000 的高性能及 Smart 特性应对存储集中化场景；以高端 OceanStor18000 设备及其异构兼容能力应对存储虚拟化利旧场景。

存储整合解决方案提供专业的数据迁移能力，如下有三种迁移方法：

- 基于异构虚拟化和智能迁移功能的数据迁移方法可以将满足兼容性的第三方厂商的存储阵列产品或华为存储阵列产品上的业务数据迁移到华为新存储阵列产品上。

- 基于存储的 LUN 拷贝迁移方法可以在满足兼容性列表的存储阵列间进行数据迁移。
- 基于存储远程复制功能的迁移方法现阶段仅适用于华为存储之间的长距离数据迁移。

存储整合解决方案提供高效的异构兼容能力，屏蔽第三方存储系统硬件差异，实现异构存储资源接管和统一运维管理，资源利旧，保护客户既有投资。

存储整合解决方案提供多业务 Qos 保障，为不同的业务指定不同性能目标，以满足关键业务的运行，实现存储资源和性能的灵活分配。

存储整合解决方案使用卷镜像功能，在主存储和异构存储中写双份，达到老旧存储的再利用，也达到了数据高可靠效果。

通过实际验证，该方案满足 BSS 业务系统高可用、数据高可靠、业务高连续性等需求。

1.4 方案亮点

方案特点

- 高效的数据迁移，满足业务停歇时间窗要求
- 可靠的存储平台保障 CRM/Billing/BI 等多业务承载
- 简单的管理可快速完成存储资源的扩展与卸载

客户价值

- 系统 7×24 小时持续运行，保障核心业务不中断。
- 异构存储，旧设备利用，保护已有投资。
- 多品牌设备统一可视化管理，节省维护成本。
- 设备平滑更替，提高资源利用率
- 数据迁移有保障，业务无需中断
- 多设备空间整合，简化空间分配

1.5 主要内容

白皮书涵盖了以下主题：

- 介绍华为存储整合解决方案的设计架构，描述存储虚拟化、智能迁移、LUN 拷贝、远程复制、QoS 保障、卷镜像、多租户等技术，说明存储整合三大场景的具体组网、基本原理和配置。
- 介绍方案可靠性验证及性能验证结果。

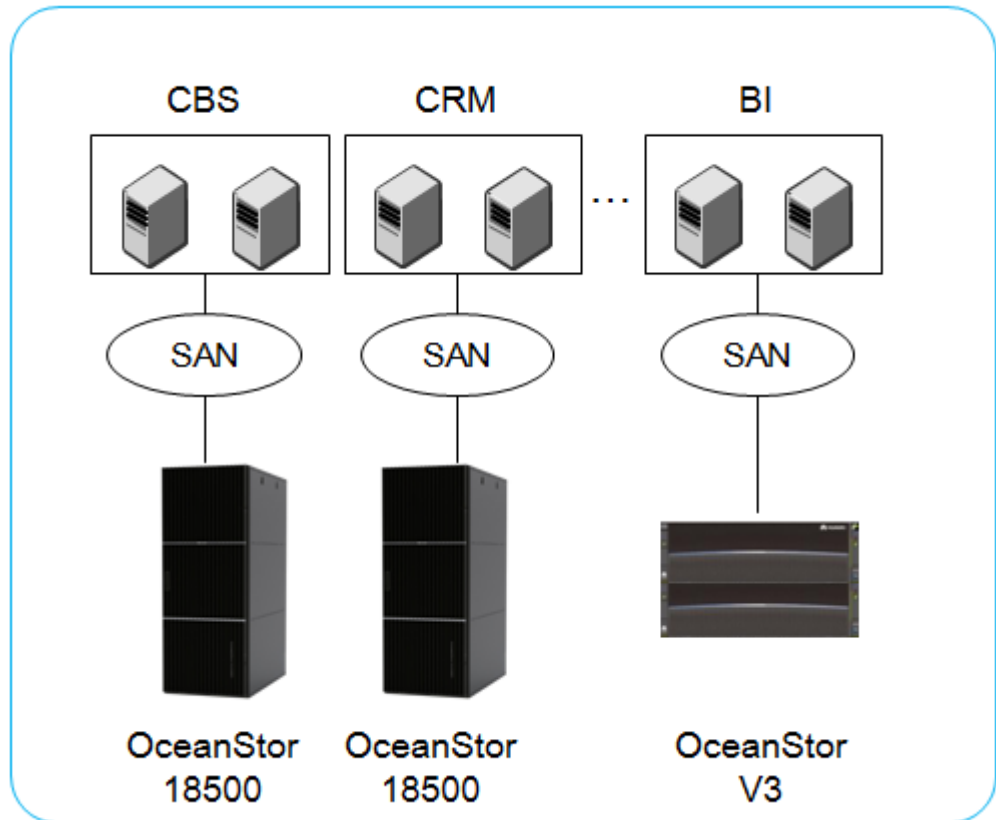
2 方案架构

关于本章

- 2.1 旧设备替换方案架构
- 2.2 存储集中化方案架构
- 2.3 存储虚拟化与利旧方案架构
- 2.4 方案组件

2.1 旧设备替换方案架构

图2-1 旧设备替换方案架构图



以高性能高规格的 OceanStor V3 或 18000 设备应对一对一旧设备替换场景。华为中高端存储系统能提供更高性能，具有更好的架构和扩展性，有如下优势：

- 性能和规格全面领先：

V3 融合存储首家采用 Intel 全新多核 Ivy bridge 处理器，支持 16Gbps FC/10Gbps FCoE/56Gbps InfiniBand 等主机接口，内部采用新一代 PCI-E 3.0 总线，并采用 12Gbps SAS 高速硬盘接口，可提供高达 40GB/s 的系统带宽，满足视频、大文件等高带宽应用场景所需。V3 融合存储性能可达到百万级 IOPS，最大可横向扩展至 8 个控制器，1TB 缓存容量及 8PB 存储空间，规格全面领先。

- 智能 IO 卡：

单张接口卡可根据用户需求设定不同的接口协议，支持 8Gbps FC、16Gbps FC、10Gbps iSCSI 和 10Gbps FCoE 等协议类型。

- 智能数据协处理卡：

可支持性能无损的数据重删/压缩功能，有效降低数据存储成本。也可支持数据加密功能，确保用户数据安全无忧。

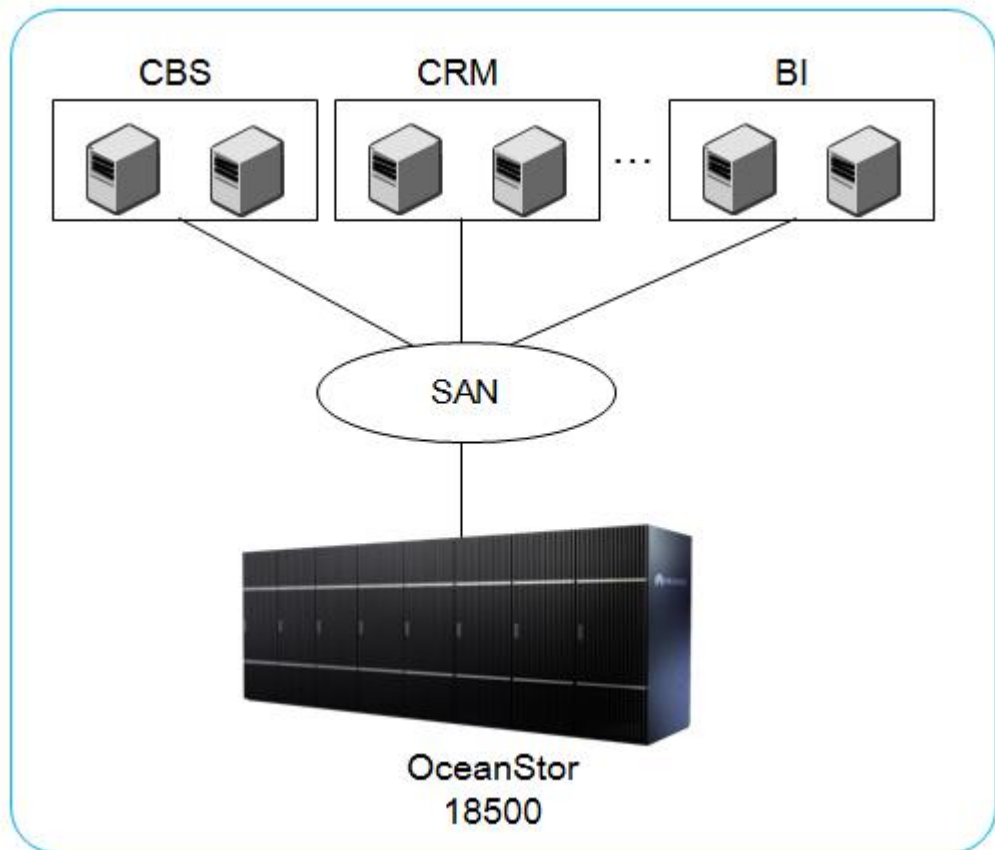
方案优势如下：

- 使用更高规格更高性能存储，保障性能需求；

- 使用存储异构虚拟化和智能迁移技术，使新旧设备平滑替换，业务不中断。

2.2 存储集中化方案架构

图2-2 旧设备替换方案架构图



以高端 OceanStor18000 的高性能及 Smart 特性应对存储集中化场景；全面基于最先进的 SAN 技术，应用虚拟化技术和智能迁移技术，将客户不同业务数据整合到单套华为系统中，整个系统达到 99.999%可靠性。

针对运营商关键业务（如 OCS 计费系统等），利用存储整合一系列软件特性保障服务质量：

- SmartQos 实现 CPU、内存、IO 资源按照优先级调用，以保障高优先级业务的服务质量；
- SmartPartition 设置 Cache 限额，防止非关键业务（如 Report）抢占 Cache 资源，以保障关键业务的服务质量；
- SmartMulti-Tenant 有效隔离租户间的资源数据，由租户的管理员对租户内的资源进行自我管理，提高存储资源分配效率。

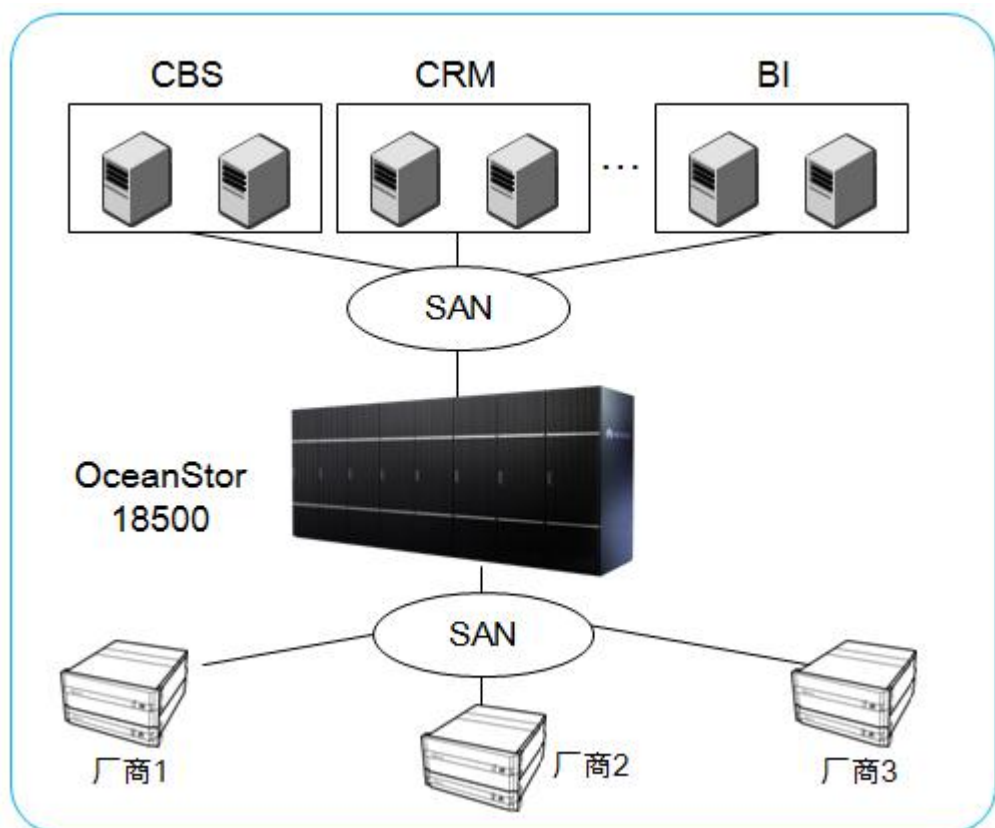
针对运营商一些业务的热点数据（如 CRM 的 Base DB 热点数据表、用户索引表空间），采用 SmartTier 分级存储软件，自动识别数据冷热程度，将热点数据搬迁至 SSD 以提高访问性能。

方案优势如下：

- 存储集中化方案使用存储异构虚拟化和智能迁移技术，使新旧设备平滑替换，业务不中断。
- 从烟囱式架构过度到资源共享，提高了资源利用率，且简化管理，消除多厂商问题。
- 通过业务分析，实现可衡量/预测的服务质量，多业务 QoS 保障，满足关键业务稳定运行。

2.3 存储虚拟化与利旧方案架构

图2-3 旧设备替换方案架构图



以高端 OceanStor18000 设备及其异构兼容能力应对存储虚拟化利旧场景。存储虚拟化与利旧方案，将各类存储资源都接管到统一的 SAN 存储资源池中，充分利用存储资源池强大的软件特性，满足运营商混合业务场景按需使用存储的要求。

方案优势如下：

- 提升存储资源使用效率：异构存储，利旧存储资源；冷热数据分层存储，最大化存储使用效率。
- 简化管理，降低 Opex：资源池屏蔽各厂商存储差异，提供统一访问界面，简化管理；
- 提高系统可靠性：RAID2.0+技术，快速重构；使用华为存储设备与第三方设备构建高可靠方案，实现数据镜像保护策略。
- 构建数据保护方案，保障业务连续：构建双活、两地三中心容灾方案，保障业务连续。

2.4 方案组件

本次方案涉及 OceanStor 18500/18800 和 OceanStor 5800/6800V3 控制框、硬盘框、异构虚拟化、LUN 迁移、数据分级、服务质量控制、多租户、LUN 拷贝、远程复制等配置。下表描述了解决方案所需要的必备配置。

组件名称	数量	描述信息
OceanStor 18500/18800 控制框	1	双控或多控（4/6/8），XXX GB 内存
OceanStor 18500/18800 硬盘框	xx	
OceanStor 5800/6800 V3 控制框	1	双控或多控（4/6/8），XXX GB 内存
OceanStor 5800/6800 V3 硬盘框	xx	
数据分级软件		
异构虚拟化		
LUN 迁移		
服务质量控制		
多租户		
LUN 拷贝		
远程复制		

2.4.1 关键组件介绍

2.4.1.1 华为 OceanStor V3 系列产品

OceanStor 5800 V3/6800 V3 存储系统是华为根据存储产品应用现状和存储技术未来发展趋势，推出的新一代中端存储产品，能够满足大中型企业用户对海量数据存储、高速数据存取、高可用性、高利用率、绿色环保和易于使用的要求。

OceanStor 5800 V3/6800 V3 存储系统采用灵活、可扩展的设计理念，具有业界领先的硬件规格、RAID2.0+底层虚拟化架构和多种数据保护技术，能够满足大型数据库 OLTP/OLAP（Online Transaction Processing/Online Analytical Processing）、高性能计算、数字媒体、因特网运营、集中存储、备份、容灾和数据迁移等不同业务应用的需求。

OceanStor 5800 V3/6800 V3 存储系统不但能够为应用服务器提供性能出色的存储服务，而且支持各种先进的数据备份和容灾技术，保证数据业务顺利、安全的运行。除此之外，OceanStor 5800 V3/6800 V3 存储系统还提供易于使用的管理方式和方便快捷的本地/远程维护方式，大大降低了设备管理和维护的成本。

图2-4 华为 OceanStor 5800/6800 V3 系列



获取 OceanStor V3 系列更多信息，参考如下链接：

<http://e.huawei.com/cn/products/cloud-computing-dc/storage/unified-storage/mid-range>

<http://e.huawei.com/cn/products/cloud-computing-dc/storage/unified-storage/high-end>

2.4.1.2 华为 OceanStor 18000 系列产品

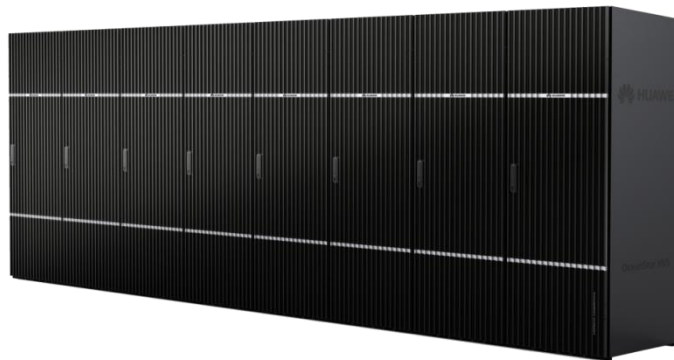
OceanStor 18000 系列高端存储系统是华为自主研发的存储高端旗舰产品系列，包括 OceanStor 18500 和 OceanStor 18800 两种型号，致力于为企业级数据中心提供安全可信、弹性高效的核心存储解决方案。

OceanStor 18000 系列基于 Smart Matrix 智能矩阵架构和全虚拟化专用操作系统，最大扩展至 16 个控制器、3,216 块硬盘、7,168TB 容量，192GB 系统带宽、3TB Cache，192 个主机接口（FC/FCoE/iSCSI），最大支持 65,536 台服务器共同使用。

软件方面，OceanStor 18000 系列配备 SMART 系列资源管理软件和 Hyper 系列数据保护软件，最大提升资源利用效率、提供关键业务智能优先保障和 7*24 业务可用性。

OceanStor 18000 系列采用整柜交付的模式，支持机柜分散布放。

图2-5 华为 OceanStor 18000 系列



获取 OceanStor 18000 系列更多产品信息，参考如下链接：

<http://e.huawei.com/cn/products/cloud-computing-dc/storage/unified-storage/18000>

3 关键技术

关于本章

- 3.1 异构虚拟化技术
- 3.2 数据迁移技术
- 3.3 服务质量保障技术
- 3.4 卷镜像技术
- 3.5 缓存分区优化技术
- 3.6 多租户技术

3.1 异构虚拟化技术

异构虚拟化技术就是着眼于客户的存量设备，为异构设备提供丰富的异构虚拟化功能。通过异构接管功能降低用户管理不同异构阵列的复杂度并提升异构 LUN 性能；通过提供异构 LUN 在线迁移实现在不断业务的情况下实现异构 LUN 之间数据平滑的迁移；通过提供异构远程复制技术实现异构 LUN 间的容灾；通过提供异构快照技术实现异构 LUN 的快速备份。

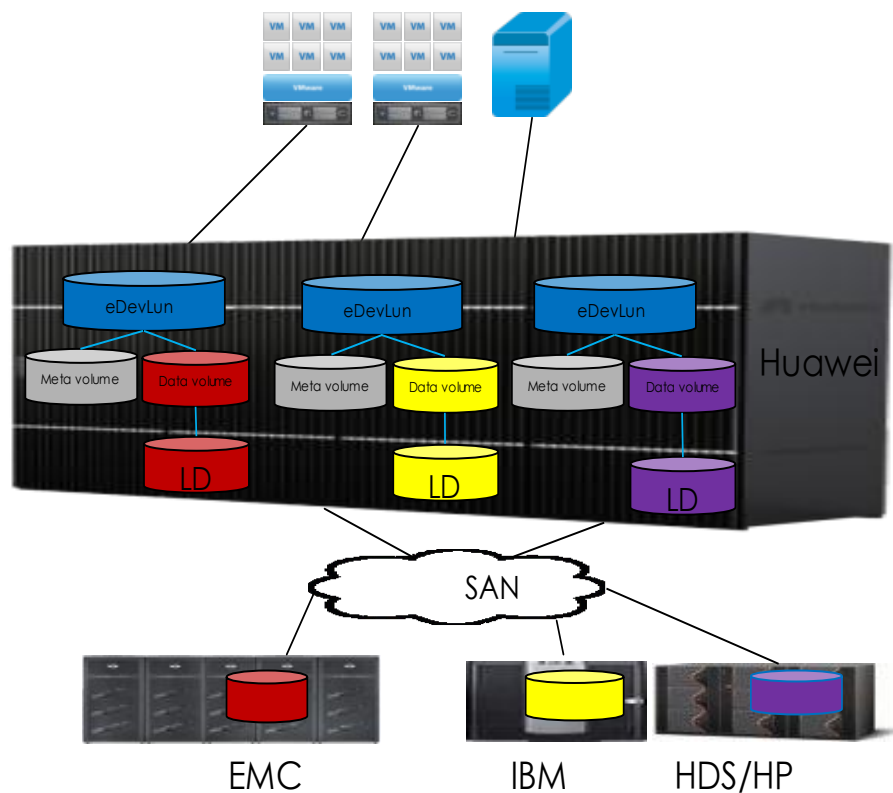
华为存储设备通过上述异构虚拟化技术，充分利用客户存量存储设备价值，保护好了客户现有投资，同时也降低了客户投资成本。

OceanStor 存储系统的异构虚拟化技术软件名称为 SmartVirtualization。

3.1.1 SmartVirtualization 工作原理

异构虚拟化的工作原理就是把异构阵列映射到本端阵列的 LUN，作为可为本端阵列提供存储空间的逻辑盘 LD，再在该逻辑盘 LD 上创建为可对主机映射的异构设备 LUN eDevLun，逻辑盘 LD 为 eDevLun 的数据卷 Data Volume 提供了全部的数据存储空间，eDevLun 的元数据卷 Meta Volume 的存储空间由本地存储提供。异构虚拟化可保证外部 LUN 数据完整性不被破坏。eDevLun 与原异构 LUN 具有不同的全球唯一名称 WWN。

图3-1 异构虚拟化工作原理



由于 eDevLun 与本地 LUN 基本上具有相同的 LUN 属性，所以，通过 SmartMigration 技术为异构 LUN 提供在线 LUN 迁移功能，通过 HyperReplication/S 技术为异构 LUN 提供同步远程复制功能，通过 HyperReplication/A 技术为异构 LUN 提供异步远程复制功能，通过 HyperSnap 为异构 LUN 提供异构快照功能。同时通过 SmartQos 和 SmartPartition 技术，以及 CACHE 可回写策略提升异构 LUN 性能。

3.1.2 SmartVirtualization 技术特点

可靠的兼容性

由于不同厂商的异构阵列在实现的过程中，对 SCSI 协议的理解和遵从性都不一样，导致接管异构阵列带来很大的兼容性挑战。异构虚拟化技术通过针对不同的阵列做兼容性的差异化处理，能很好的识别和处理与异构设备间的各种兼容性问题，如及时识别和处理异构设备的 LUN 路径故障，以此达到与异构阵列间高可靠的兼容。对支持的异构设备，都通过了华为兼容性实验室认证。

后端多路径

异构虚拟化技术通过后端多路径软件支持以多路径冗余方式连接异构阵列，防止与异构阵列间的物理链路连接出现因为单个链路故障而导致业务中断。由于针对单个异构阵列进行了专门的兼容性处理，后端多路径软件能很好的识别异构 LUN 路径，如识别出异构 LUN 的优选路径，非优选路径，或备用路径等，并在这些路径中选择最合适的路径进行 IO 下发，并提供多种后端多路径负载均衡选路算法。

异构 LUN 迁移

异构虚拟化技术通过与 SmartMigration 技术结合，为异构阵列间的 LUN 提供高可靠的在线 LUN 间数据迁移，异构 LUN 数据迁移过程中不需要主机中断在线业务。迁移速度可配置。迁移过程中，如出现目标 LUN 故障，会记录 LUN 间 IO 差异日志，故障恢复后，只进行差异同步。

异构远程复制

异构虚拟化技术通过与 HyperReplication 技术结合，为异构 LUN 提供同步和异步远程复制功能。其主要特点如下：

- 异构同步复制 RPO 为 0；异构异步复制通过多时间片的缓存技术使 RPO 可达到秒级。
- 支持分裂模式，在分裂模式下，生产主机的写请求只会写到主 LUN，并通过差异日志来记录主、从 LUN 数据之间的差异。当用户希望重新保持主、从 LUN 数据一致时，用户可手动重新开启同步操作。分裂模式可以灵活的满足用户在某些特定场景下的需求：如暂时性的链路维修，网络带宽扩容，需要从 LUN 保持某一时间点的数据等等。
- 快速响应故障和故障恢复，系统在检测到系统故障情况下，复制进入断开状态，在断开状态下，生产主 LUN 记录 IO 差异，故障恢复后，根据策略自动完成主从 LUN 数据差异同步。
- 支持异构远程复制从 LUN 可写。从而实现用户在不影响主 LUN 业务的情况下，通过从 LUN 进行数据分析和挖掘等用户使用场景。
- 支持异构远程复制的 LUN 主从切换。
- 支持异构远程复制 LUN 一致性组功能。

异构快照

异构虚拟化技术通过和 HyperSnap 技术结合，为异构 LUN 提供快照功能。快照采用 COW（Copy-On-Write）技术，保证快照不影响原 LUN 读 IO 性能。具有快照瞬间完成，占用磁盘空间小等特点。另外还支持对异构快照创建多个副本，用于进行数据分析和数据挖掘等应用。

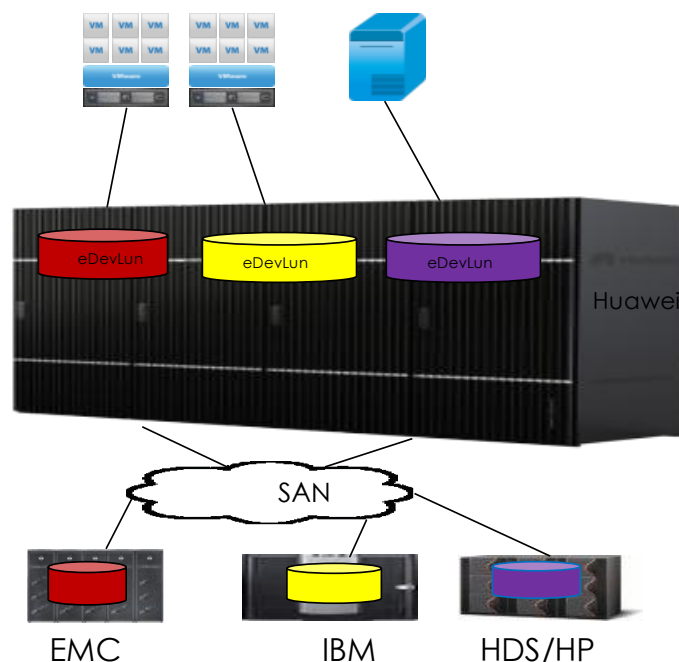
3.1.3 SmartVirtualization 应用场景

异构阵列接管

用户的数据中心通过长期的建设，数据中心可能存在来自不同异构厂商的存储阵列。如何很好的管理和应用好来自不同厂商的异构阵列给存储管理员提出了很高的技术要求。存储管理员可通过异构虚拟化接管功能，可大大降低对管理异构阵列的技术难度和复杂度。存储管理员只需要管理好华为阵列就能达到管理好所有异构阵列的目的，从而减轻了存储管理员的工作负担。其场景特点就是简化用户管理。

异构接管的最大容量可以根据产品型号支持的最大异构 LUN 个数乘以单 LUN 最大容量得出，例如：OceanStor 18800 支持 2048 个异构 LUN，最大容量就是 $2048 * 64 = 131072$ TB。

图3-2 异构阵列接管

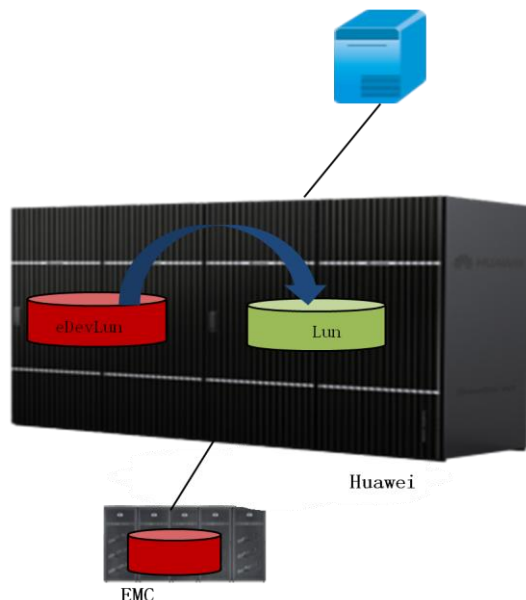


异构数据搬迁

用户数据中心可能存在大量的存量异构设备，有些设备可能即将过保或性能不能再满足业务要求，在购买华为 OceanStor 存储后，客户可能希望把存量 LUN 上的业务迁移到新购阵列中。客户可通过异构虚拟化 LUN 在线迁移功能，在线迁移异构 LUN 数据到新购阵列中，数据迁移过程中，主机业务正常运行，但在进行数据迁移前，需

要对异构 LUN 进行接管。其场景特点就是异构 LUN 数据搬迁过程中，主机业务不中断。

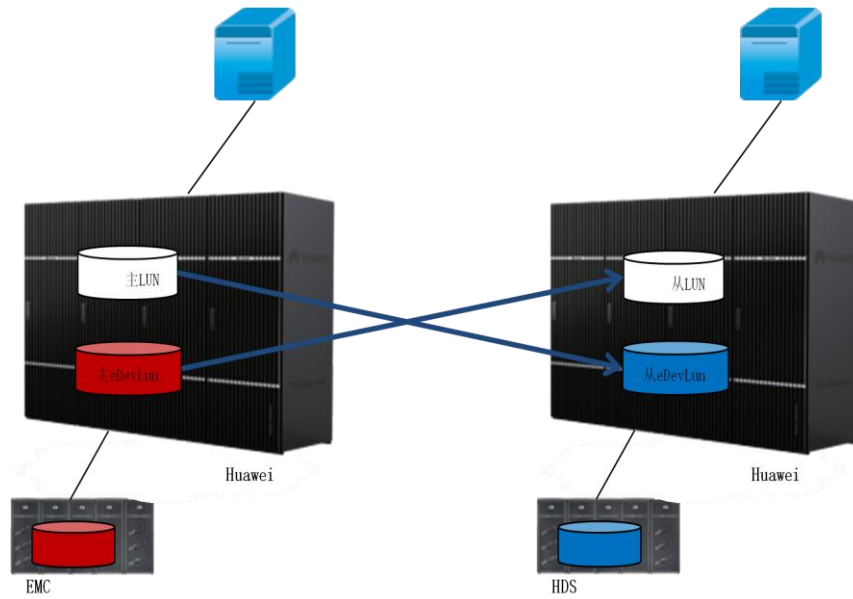
图3-3 异构数据搬迁



异构容灾

如果客户业务数据分散在不同的站点，且对业务的持续性要求较高时，需要各业务站点的数据互为备份和业务切换。当灾难发生时，能够通过互为备份的站点进行业务数据的接管和数据恢复。但是，可能由于数据站点的阵列来自于不同的异构厂商，导致异构阵列间的数据无法做到互为备份。异构虚拟化提供的异构同步和异步复制功能，可以使异构阵列间的 LUN 数据做到互相备份，做到站点间的数据容灾。

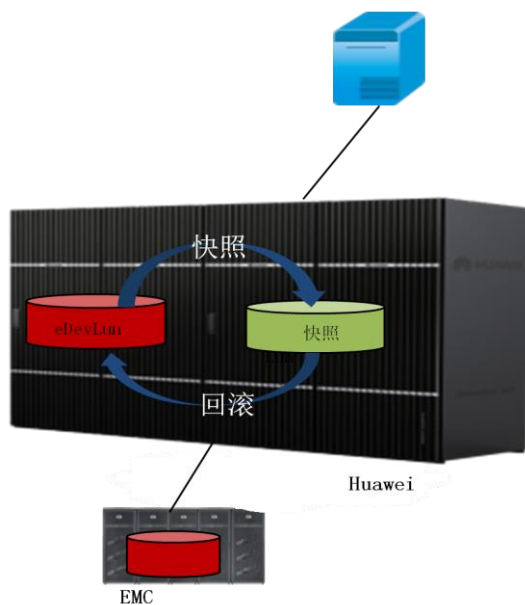
图3-4 异构容灾



异构数据保护

客户异构阵列上的 LUN 数据可能受到病毒或其他原因导致 LUN 数据受到破坏。异构虚拟化提供异构快照技术可为异构 LUN 提供快照备份，快照瞬间完成，当数据被破坏后，可通过快照迅速的回滚到指定的快照时间点的数据，迅速恢复数据。

图3-5 异构数据保护

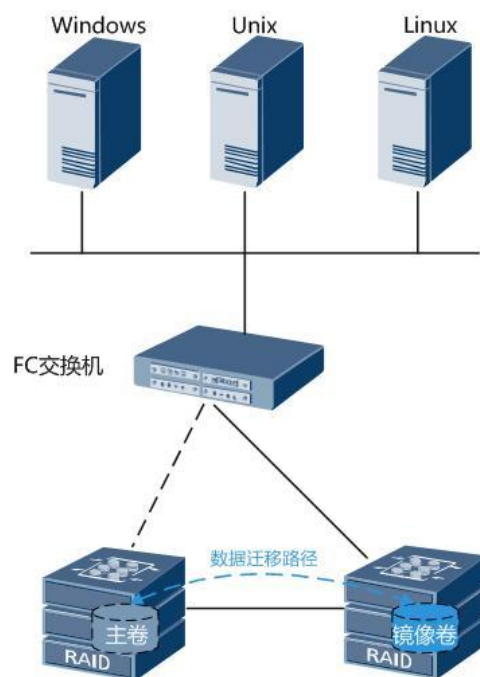


3.2 数据迁移技术

3.2.1 基于异构虚拟化的数据迁移方案

数据迁移方案架构图

图3-6 异构虚拟化数据迁移方案架构图

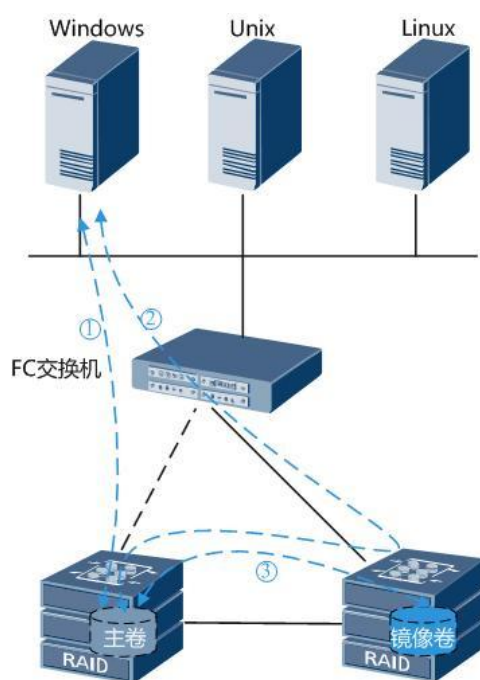


数据迁移工作原理

基于异构虚拟化的数据迁移方案可以将满足兼容性的第三方厂商的存储阵列产品或华为存储阵列产品上的业务数据迁移到华为新存储阵列产品上。异构虚拟化方案在迁移的初始阶段接管源存储映射的 LUN，将远端存储阵列的 LUN 在确保数据不丢失的情况下初始化为本端存储阵列可以管理的 eDevLUN。利用异构虚拟化提供的增值特性，如 LUN 迁移，使得数据在源存储阵列和本端存储阵列之间进行快速的迁移，异构接管过程中需短暂中断业务，数据复制过程中可保持业务在线运行。

迁移步骤

图3-7 异构虚拟化迁移步骤示意图



迁移前，I/O 路径为“服务器主机—源存储”。

停止业务，将华为新存储产品（带有异构虚拟化增值特性及 License）加入客户的应用系统中。连接方式由“服务器主机—源存储”方式改为“服务器主机—华为新存储产品—源存储”。在原始数据阵列上将所属服务器主机的 LUN 映射给华为新存储产品，经由华为新存储产品接管后再映射给服务器主机。此操作需要中断客户业务系统，此步骤的中断时间长度可以按照每个标准系统（双机集群单阵列）2 小时+5 分钟/LUN 的时间估算，具体包括：

物理链路连接

光纤交换机划 zone

源存储上删除 LUN 到服务器的映射

源存储上添加 LUN 到华为新存储产品的映射

华为新存储产品上扫描发现源存储映射的 LUN

华为新存储产品上建立 eDevLUN 和 LUN 组

华为新存储产品上建立主机组和主机，添加启动器

华为新存储产品建立映射视图

配置主机多路径

服务器上发现华为新存储产品映射的 LUN

重新识别新存储数据，让业务应用从华为新存储产品启动运行

说明

*此步骤的中断时间长度可以按照每个标准系统（双机集群单阵列）2 小时+5 分钟/LUN 的时间估算。例如一个双机业务系统有 5 个 LUN 则此步骤中断时间为 2 小时+5*5 分钟=2.41 小时。此时间为售前阶段需申明的时间，实际操作可能短于此估算时间。

将目标存储阵列上计划分配给服务器主机的 LUN 作为迁移的目标 LUN，在华为新存储产品中对源数据阵列和数据迁移目标阵列之间进行在线数据复制操作。

确认阵列间数据复制完成后，断开源数据阵列与华为新存储产品的逻辑连接，由数据迁移目标阵列对业务系统提供存储空间。

方案特点

- 基于异构虚拟化的数据迁移方案可以针对主流的操作系统（Windows、AIX、HP-UX、Solaris、SuSE Linux、Redhat Linux 和其它 Linux 平台）在不同厂商的 SAN 存储阵列产品之间进行数据迁移。
- 数据迁移过程中数据复制通过华为新存储产品完成，对业务系统没有业务压力。不会占用服务器主机的 CPU、内存等计算机资源。
- 数据迁移过程中，业务不会中断，可以持续进行，但存储访问性能会有一定影响；
- 无须在现有环境中安装任何软件或代理，方便快捷的进行数据迁移；
- 华为存储产品提供人性化的图形化管理界面，操作简单，图形直观，大大提升了数据迁移效率。

停机时间和数据复制时间估算

异构虚拟化方案为在线数据迁移方案，此方案数据迁移时间分为停机接管以及数据复制时间两部分，停机接管是将华为新存储产品接入到业务系统中的时间，此部分必须通过停止业务来保证数据一致性，因此必须短暂停机。业务重新启动后可以在线数据复制，此时间段内业务可以继续运行，为在线数据迁移。

停机接管、迁出时间估算方法：

- **接管（离线）：**此步骤的中断时间长度可以按照每个标准系统（双机集群单阵列）2 小时+5 分钟/LUN 的时间估算。例如一个双机业务系统有 5 个 LUN 则此步骤中断时间为 2 小时+5*5 分钟=2.41 小时。此时间为售前阶段需申明的时间，实际操作可能短于此估算时间。

- **数据复制（在线）：**此步骤时间可以由总容量/迁移速率来评估，异构虚拟化迁移方案一般可以按 250MB/s 来估算*，例如 1TB 数据，同步时间为 $1\text{TB} \times 1024 \times 1024 / 250\text{MB/s} / 3600$ 秒每小时=1.165 小时。

说明

*此速率按常用速率估算，某些特殊情况，比如客户使用的是 2Gbps 接口，则链路最大带宽只能到 200MB/s 左右，具体迁移速率依现网实际情况而定。

因此，以一个双机，单存储使用异构虚拟化方案迁移的例子对迁移和停机时间进行说明：

例 1：客户一套系统需搬迁的 LUN 原始裸容量（LUN 的总容量）为 20TB，一共有 5 个 LUN，则阶段时间如下：

异构虚拟化接管（离线）：2 小时+5*5 分钟=2.41 小时

数据复制（在线）： $20\text{TB} \times 1024 \times 1024 / 250\text{MB/s} / 3600$ 秒每小时=23.3 小时

总共客户需停业务：145 分钟

总共迁移时间： $2.41 + 23.3 = 25.71$ 小时

使用建议

- 客户对业务连续性要求较高的大容量通用迁移场景。
- 源存储系统侧创建华为新存储产品主机时，主机类型请参考最新发布的兼容性列表，否则可能导致本端存储系统不能接管异构存储系统的存储资源。
- 使用异构虚拟化特性对异构存储设备上的 Thin LUN 进行接管后，该 LUN 上的业务可以正常运行，接管后对应的 eDevLUN 识别为 Thick LUN。

技术限制

源存储的限制

- 源存储必须满足华为新存储产品异构虚拟化的兼容性要求（异构虚拟化支持的阵列类型以兼容性列表为准，有项目需求请找数据迁移小组确认）。
- 异构阵列只支持 FC 组网，我司阵列支持 FC 组网，部分支持 iSCSI 组网，请以最新发布的兼容性列表为准。
- 原阵列 HBA adapter 速率 $\geq 2\text{Gbps}$

新存储产品的限制

- 迁移的 LUN 上要求无增值业务。
- LUN 迁移源/目标 LUN 不支持带增值迁移。
- 源 LUN、目标 LUN 的归属控制器必须相同。
- 引擎故障后，引擎上 LUN 的业务会中断；
- 最大 LUN 容量 64TB
- 当映射给本端存储系统的 LUN 数量达到或超过规格限制时，添加新的 LUN 映射后，无法被本端存储系统识别。请参考本端存储系统的最新产品规格。
- 其他限制请参考对应的新存储产品限制表

光纤交换机的限制

- 光纤交换机有足够的空余端口供华为新存储产品使用，具体需要端口数与组网方式强相关，每台光纤交换机至少需要 4 个空余口。
- 光纤交换机端口速率 $\geq 2\text{Gbps}$

服务器的限制

- 操作系统与替代阵列多路径软件兼容

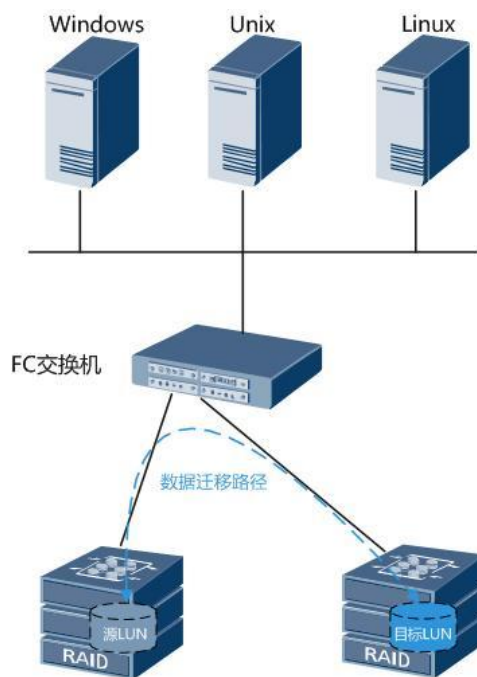
选择此方案的优选场景

- 满足异构兼容性的存储搬迁场景优选此方案
- 客户可以允许短暂的业务中断的场景

3.2.2 基于存储功能（LUN 拷贝）的数据迁移方案

数据迁移方案架构图

图3-8 LUN 拷贝数据迁移方案架构图

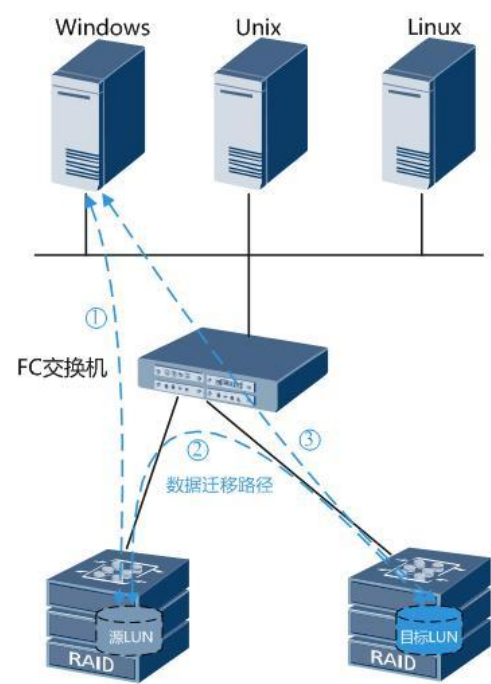


数据迁移工作原理

基于存储的 LUN 拷贝迁移方案可以在满足兼容性列表的存储阵列间进行数据迁移。在将目标存储系统接入存储网络后，关闭业务，解除源存储对业务主机的映射，并建立源存储到目标存储的映射，然后在目标存储系统上进行源 LUN 与目标 LUN 的 LUN 拷贝操作，当所有源 LUN 都成功进行了 LUN 拷贝后，从目标存储系统上拉起业务。

迁移步骤

图3-9 LUN 拷贝迁移步骤示意图



迁移前准备，包括目标存储配置与 LUN 的创建（依照源存储上 LUN 的大小与个数进行创建）等操作，此时业务的 I/O 路径为“务器主机—源存储”

将目标存储接入业务网络中。在业务离线后，解除源存储对业务主机的映射，并添加到目标存储的映射，待完成 LUN 拷贝后，建立目标存储系统对主机的映射，然后重新拉起业务，具体步骤包括：

- 物理链路连接，完成目标存储接入
- 业务离线，并解除源存储对业务主机的映射
- 建立源存储到目标存储的映射
- 根据实际需求，更改业务主机的多路径软件
- 目标存储上发现源存储映射的 LUN，并创建 LUN 拷贝
- 待所有 LUN 完成 LUN 拷贝后，删除 LUN 拷贝关系
- 建立目标存储对业务主机的映射

说明

*LUN 拷贝时的迁移速率可按常用的 500MB/s 进行估算，现网实际操作中，由于客户网络环境不同，该速率可能会有较大波动，比如客户使用的是 2Gbps 接口，则链路最大带宽只能到 200MB/s 左右，具体迁移速率依现网实际情况而定。

启动客户业务系统，验证业务运行是否正常，此时的 I/O 路径为“服务器主机—目标存储”

方案特点

- 基于存储系统的 LUN 拷贝迁移方案，适用于华为存储产品间的数据迁移操作。

- 存储系统之间的数据拷贝操作，在迁移速度上比网络层或者主机层进行数据迁移效率更高，速度更快。
- 数据迁移过程中，为了保证数据的一致性，需要在暂停业务的情况下进行数据拷贝操作。
- 无须在现有环境中安装任何软件或代理，方便快捷的进行数据迁移；
- 数据迁移过程中，数据复制在阵列间完成，对同一网络的其它业务系统影响很小，同时其也不会占用主机的相关资源。
- 提供人性化的图形化管理界面，操作简单，图形直观，大大提升了数据迁移效率。



说明

*基于存储系统的 LUN 拷贝迁移方案，也可用于满足华为存储兼容性的友商存储系统数据搬迁。

迁移时间和停机时间估算

LUN 拷贝为离线数据迁移方案，此方案在数据迁移过程中需要停机进行拷贝操作，整个操作过程主要分为二个阶段。

迁移时间与停机时间估算方法：

- **迁移前准备（在线）：**此步骤包括目标存储的组装与配置，并依据源存储系统上 LUN 的大小与个数，完成目标存储系统上 LUN 的创建等操作，待以上操作完成后，将目标存储系统接入业务网络。
- **LUN 拷贝实施（离线）：**

停业务，解除源存储对业务主机的映射，并进行交换机配置，完成源存储对目标存储的映射，依据实际情况更新主机多路径软件等操作，操作时间按每个标准系统（双机集群单阵列）2 小时+5 分钟每个 LUN 进行估算，例如一个双机业务系统有 5 个 LUN 共 1TB 数据量，则此步骤耗时：2 小时+5*5/60=2.42 小时。

在目标存储系统上进行 LUN 拷贝操作，LUN 拷贝耗时估算方法：容量/500MB/s+5 分钟每个 LUN。例如一个双机业务系统有 5 个 LUN 共 1TB 数据量，则此步骤耗时： $1\text{TB} \times 1024 \times 1024 / 500\text{MB/s} / 3600 + 5 \times 5 / 60 = 0.96$ 小时。在实际的存储系统中，多个 LUN 拷贝可以同时进行，业务中断时间会缩减。此时间为售前阶段需申明的时间，实际操作可能短于此估算时间。

待 LUN 拷贝完成后，添加目标存储对业务主机的映射，并拉起业务，耗时估算方法：业务拉起操作时间按每个标准系统（双机集群单阵列）1 小时+5 分钟每个 LUN 进行估算，例如一个双机业务系统有 5 个 LUN 共 1TB 数据量，此步骤需要 1 小时+5*5/60=1.42 小时。

则一个双机业务系统有 5 个 LUN 共 1TB 数据量，迁移时需要停业务的时长为： $2.42 + 1.42 + 0.96 = 4.8$ 小时。



说明

*现网实际操作过程中，可以根据存储的规格限制，采用多个 LUN 拷贝同时进行的方式来缩短迁移数据的时间。

例如，以一个双机，单存储使用 LUN 拷贝方案迁移的例子对迁移和停机时间进行说明：

例 1：客户一套系统需搬迁的 LUN 原始裸容量（LUN 的总容量）为 20TB，一共有 5 个 LUN，时间计算如下：

LUN 拷贝实施（离线）：

停业务：2 小时+5*5/60=2.42 小时

LUN 拷贝：20TB*1024*1024/500MB/s/3600+5*5/60=11.54 小时

拉起业务：1 小时+5*5/60=1.42 小时

总共客户需停业务：2.42+1.42+11.54=15.38 小时

总共迁移时间：15.38 小时

使用建议

- 适用于客户对业务离线时间要求不高的场景。
- 离线迁移成本较低，适合对价格敏感的客户。

技术限制

源存储的限制

- 适合于华为存储更替时的数据迁移
- 可用于满足目标存储兼容性的友商存储数据搬迁（兼容性不在公布的兼容性列表中的设备问题请联系数据迁移小组确定评估）

目标存储的限制

- 目标存储必须具有 LUN 拷贝的 License（可申请临时 License）
- 进行 LUN 拷贝的源 LUN 与目标 LUN 不能都是外部 LUN
- LUN 拷贝过程中最大 LUN 拷贝数量限制与最大目标 LUN 数量限制依具体产品的具体版本而定，请参考其《产品描述说明》
- LUN 拷贝过程中支持的最大远端存储设备数量依不同的存储产品与版本而定，具体限制请参考对应产品的《产品描述说明》，如 T 系列 V200R002C00 版本，最大支持可连接的远端存储设备数量为 32
- LUN 拷贝的源 LUN 与目标 LUN 的增值互斥关系请参考《某型号存储某版本的 LUN 拷贝特性指南》
- 替代阵列的容量≥原阵列的容量

光纤交换机的限制

- 光纤交换机有足够的空余端口供目标存储系统使用，具体需要端口数与组网方式强相关，每台光纤交换机至少需要 2 个端口。

服务器的限制

- 操作系统与替代阵列多路径软件兼容。

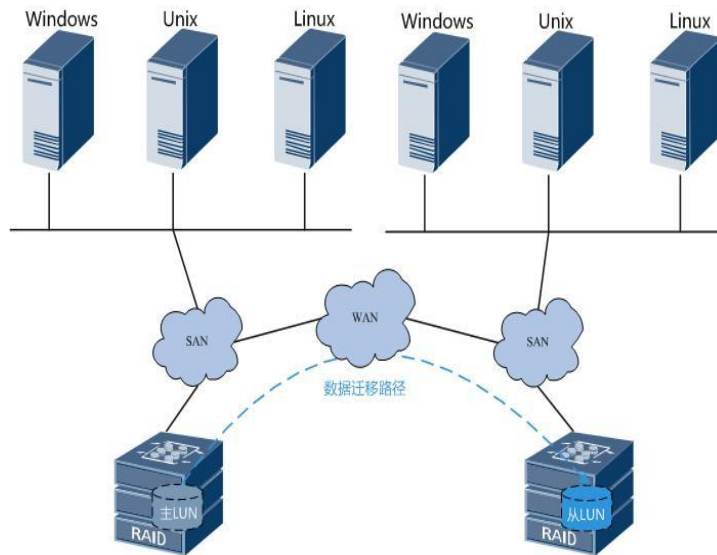
选择此方案的优选场景

- 适用于客户对业务离线时间要求不高的场景。
- 迁移报价较低的场景。

3.2.3 基于存储功能（远程复制）的数据迁移方案

数据迁移方案架构图

图3-10 远程复制数据迁移方案架构图



数据迁移工作原理

基于存储远程复制功能的迁移方案现阶段仅适用于华为存储之间的长距离数据迁移或数据中心整合。其原理是在将目标存储系统接入存储网络后，建立源存储与目标存储的逻辑链路，然后在阵列之间配置远程复制进行数据的迁移，待阵列之间初始同步完成后，关闭业务，进行差异数据同步，在目标存储与源存储数据一致时，从目标存储系统上拉起业务，完成迁移操作。

说明

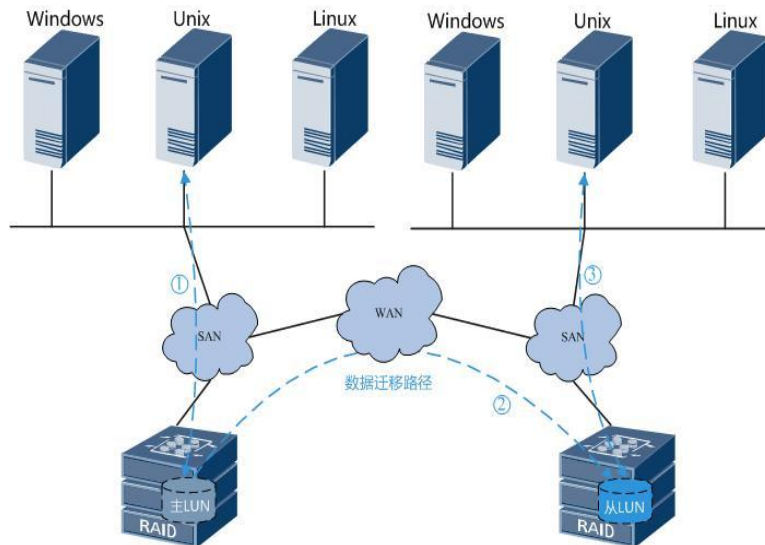
*存储的远程复制功能，其工作模式可分为同步远程复制与异步远程复制，前者为保证业务的正常运行，要求传输时延小，适用于近距离或者同城的数据迁移；而后者适用于较长距离的数据迁移或者数据中心整合，具体情况依赖于差异数据产生量、网络传输速率和传输时延等。

*现阶段基于远程复制的数据迁移方案仅适用于带有远程复制功能的华为存储之间的数据搬迁，但是在结合异构虚拟化功能后，可以实现对满足兼容性要求的第三方存储实施基于远程复制功能的数据迁移。

*远程复制功能是存储系统的一项增值特性，需要 License 进行激活。

迁移步骤

图3-11 远程复制迁移步骤示意图



迁移前准备，包括 IP 地址规划，目标存储配置与 LUN 的创建（依照源存储上 LUN 的大小与个数进行创建）等操作，此时业务的 I/O 路径为“服务器主机—源存储”

将目标存储接入客户网络中，建立源存储与目标存储的逻辑链路后，在源存储上创建远程复制，并根据业务的实际情况添加一致性组，待初始同步完成后，暂停业务，进行主 LUN 与从 LUN 的差异数据同步，在主从 LUN 的数据一致后，从目标存储端拉起业务，具体步骤包括：

目标存储接入，远程复制 License 检查

建立源存储与目标存储的逻辑链路

在源存储上创建远程复制，并根据业务实际情况添加远程复制一致性组

进行主从 LUN 的初始同步

初始同步完成后，暂停业务，进行差异数据的同步

差异数据同步完成后，建立目标存储对业务主机的映射，并在目标存储上拉起业务

 说明

远程复制的迁移速率可按常用的 40MB/s 进行估算，现网实际操作中，远程复制的迁移速率受逻辑链路带宽与远程复制迁移速率设置有关，具体迁移速率依现网实际情况而定。

启动客户业务系统，验证业务运行是否正常，此时的 I/O 路径为“服务器主机—目标存储”

方案特点

- 基于存储系统的远程复制的迁移方案，适用于华为存储产品间的数据迁移操作。
- 该方案适用于长距离的数据迁移操作或数据中心整合。
- 数据迁移过程中，为了保证数据的一致性，需要在短暂关闭业务的情况下进行差异数据的同步操作。
- 无须在现有环境中安装任何软件或代理，方便快捷的进行数据迁移；

- 数据迁移过程中，数据复制在阵列间完成，对同一网络的其它业务系统影响很小，同时其也不会占用主机的相关资源。
- 提供人性化的图形化管理界面，操作简单，图形直观，大大提升了数据迁移效率。

迁移时间和停机时间估算

远程复制为需短暂停机的数据迁移方案，初始同步过程可以在线进行，但是为了保证数据的一致性，在差异数据同步阶段需要离线进行操作，整个操作过程主要分为三个阶段。

迁移时间与停机时间估算方法：

- **迁移前准备（在线）：**

此步骤包括 IP 地址规划，目标存储的组装与配置，并依据源存储系统上 LUN 的大小与个数，完成目标存储系统上 LUN 的创建等操作，待以上操作完成后，将目标存储系统接入业务网络。

- **远程复制初始同步（在线）：**

目标存储接入后，建立与源存储的逻辑链路，并在源存储上配置远程复制，操作时间按逻辑链路配置 1 小时+10 分钟每个 LUN 进行估算，例如采用远程复制迁移一个双机业务系统，其有 5 个 LUN 共 1TB 数据量，则此步骤耗时：1 小时+5*10/60=约 1.84 小时。

远程复制初始同步耗时计算方法：容量/50MB/s。例如一个双机业务系统有 5 个 LUN 共 1TB 数据量，则此步骤耗时：1TB/50MB/s/3600=5.56 小时。在实际的存储系统中，多个远程复制可以同时进行，业务中断时间会缩减。此时间为售前阶段需申明的时间，实际操作可能短于此估算时间。

- **远程复制差异数据同步（离线）：**

待初始同步完成后，暂停业务，进行差异数据同步，耗时估算方法：每个标准系统（双机集群单阵列）1 小时+5 分钟每个 LUN+差异数据量/50MB/s 进行估算，例如一个双机业务系统有 5 个 LUN 共 1TB 数据量，则耗时：1 小时+5*5/60+50GB/50MB/s/3600=1.70 小时。

迁移完成后，从目标存储拉起业务，每个标准系统（双机集群单阵列）2 小时。

则一个双机业务系统有 5 个 LUN 共 1TB 数据量，迁移时需要停业务的时长为：1.84+5.56+1.70+2=11.10 小时。

例如，以一个双机，单存储使用远程复制方案迁移的例子对迁移和停机时间进行说明：

例：客户一套系统需搬迁的 LUN 原始裸容量（LUN 的总容量）为 20TB，一共有 5 个 LUN，时间计算如下：

远程复制初始同步（在线）：

链路配置与远程复制创建：1 小时+5*10/60=1.84 小时

初始同步：20TB*1024*1024/50MB/s/3600=116.51 小时

远程复制差异数据同步（离线）：

差异数据同步：1 小时+5*5/60+1TB/50MB/s/3600=6.98 小时

业务拉起：2 小时

总共客户需停业务：6.98+2=8.98 小时

总共迁移时间：8.98+1.84+116.51=127.33 小时

使用建议

- 适用于华为存储之间的长距离的数据迁移场景。
- 适用于采用华为存储的数据中心整合搬迁。

技术限制

源存储的限制

- 源存储具有远程复制功能，且具有远程复制的 License（可申请临时 License）
- 远程复制过程中最大远程复制组数量限制、可支持连接的远端存储设备数量限制，请参考对应存储、对应版本的《产品描述说明》
- 远程复制过程中一致性组的数量最大值限制与一致性组的最大成员数限制，请参考对应存储、对应版本的《产品描述说明》
- 远程复制的源 LUN 与目标 LUN 的增值互斥关系请参考《某型号存储某版本的远程复制特性指南》
- 异步远程复制迁移时要求最小链接带宽不小于双向 10Mbps，传输延迟小于 50ms（单向）
- 同步远程复制迁移时要求距离小于 200KM，最小链接带宽不小于 64Mbps，传输延迟小于 1ms（单向）
- 异步远程复制时资源池有足够的容量来保存差异数据

目标存储的限制

- 目标存储具有远程复制功能，且具有远程复制的 License（可申请临时 License）
- 替代阵列的容量≥原阵列的容量

组网的限制

- 不支持局域网中存在 NAT 设备的组网情况

服务器的限制

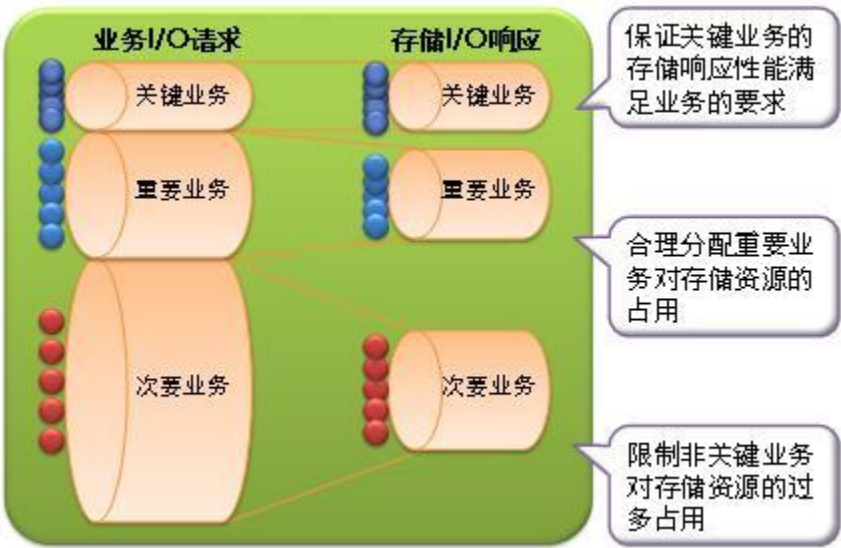
- 操作系统与替代阵列多路径软件版本兼容

选择此方案的优选场景

- 华为存储间长距离的数据迁移场景。

3.3 服务质量保障技术

SmartQoS 特性是华为 OceanStor V3 融合存储系统上提供的一项存储 QoS 功能。如下图所示，它能够对存储系统中的计算资源，缓存资源，并发资源以及硬盘资源进行智能分配和调节，在整个存储路径上进行端到端的细粒度控制，从而满足多种不同重要性业务在同一台存储设备上的不同 QoS 要求。



3.3.1 SmartQoS 工作原理

SmartQoS 工作原理简介

企业业务对数据响应的需求纷繁复杂，SmartQoS 特性可以选择从以下三个方面来保证数据业务的服务质量：

- **IO 优先级调度技术：**通过区分不同业务的重要性来划分业务响应的优先级。在存储系统为不同业务分配系统资源的时候，优先保证高优先级业务的资源分配请求。在资源紧张的情况下，为高优先级的资源分配较多的资源，以此尽可能保证高优先级业务的服务质量。当前用户可以配置的优先级分为高、中、低三个等级；
- **IO 流量控制技术：**基于传统的令牌桶机制，针对用户设置的性能控制目标（IOPS 或者带宽）进行流量限制，通过 IO 流控机制，限制某些业务由于流量过大而影响其它业务；
- **IO 性能保证技术：**基于流量扼制的方式，允许用户为高优先级业务指定最低性能目标（最小 IOPS/带宽或最大时延），当该业务的最低性能无法保障时，系统内部通过对其他低优先级业务的 IO 逐级增加时延的方式来限制其流量，从而尽力保障该业务的最低性能目标。

下面将分别针对 OceanStor V3 融合存储系统 SmartQoS 特性的主要功能和关键技术原理进行阐述。

IO 优先级调度技术

OceanStor V3 融合存储系统 SmartQoS 特性的 IO 优先级调度技术是以存储资源的分配和调度为设计出发点来实现的。存储系统在不同应用场景中性能取决于不同场景下对存储资源的消耗水平，因此只要实现了资源尤其是瓶颈资源的合理调度和分配，就能有效地对系统的性能产生影响。该技术通过监控对性能影响最大的并发，计算，缓存和硬盘四个资源的使用情况，并在出现资源瓶颈时进行资源调度的方式，尽可能满足高优先级的资源需求，较好地解决了关键业务在不同场景下的服务水平保证问题。

IO 优先级调度对象和策略

SmartQoS 特性的 IO 优先级调度技术主要针对存储系统 IO 路径上关键瓶颈资源进行调度，主要调度的资源包括并发资源，计算资源，缓存资源以及硬盘资源。调度策略基于用户配置的 LUN 的优先级来进行，不同的优先级对应不同的调度策略。LUN 的优先级由用户根据部署在该 LUN 上的业务重要性来指定，目前用户可以配置高，中，低三个优先级。

优先级调度通过控制前端主机并发，系统 CPU 资源，阵列内缓存资源，后端硬盘资源四种瓶颈资源的分配来达到控制每个调度对象在存储系统内部的响应时间。

- **前端并发资源**的优先级调度在存储系统前端进行，即针对主机的并发访问来进行控制。由于存储系统对主机的最大并发访问的承载能力是有限的，因此当系统达到最大并发数时，SmartQoS 将结合业务优先级进行控制，限制的原则是保证优先级高的业务能获得更多并发数，保证业务量更大的业务获得更多并发数；

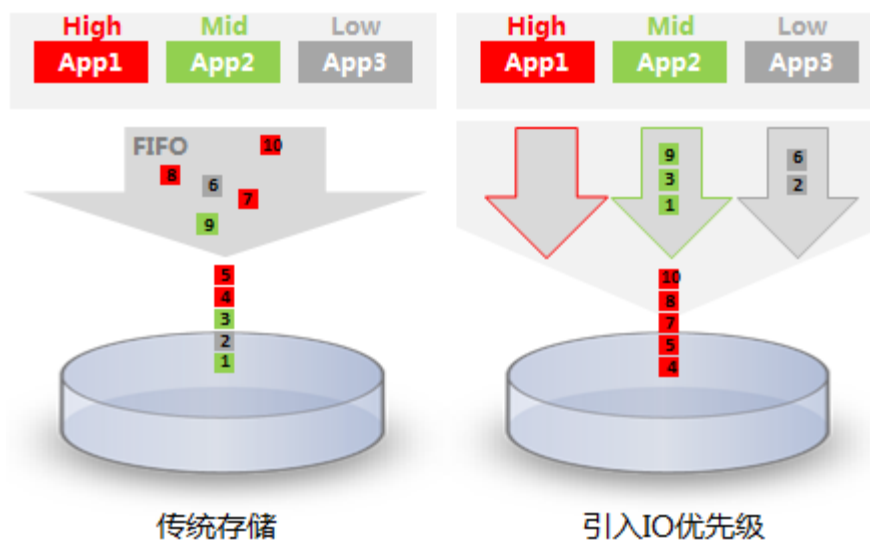
- **计算资源**的优先级调度主要通过控制 CPU 运行时间的分配来实现。SmartQoS 会根据高、中、低三个优先级各自的权重来分配每个优先级业务占用的 CPU 计算时间，当 CPU 成为系统性能瓶颈时，会通过优先级调度保证高优先级的业务获得更多的 CPU 计算时间；
- **缓存资源**的优先级调度主要通过控制缓存页面资源的分配来实现。SmartQoS 会根据每个优先级的权重来对不同优先级的页面分配请求进行调度，优先满足较高优先级业务的页面分配请求；
- **硬盘资源**的优先级调度主要通过控制 IO 的下盘顺序来实现。SmartQoS 会根据 IO 的优先级，在访问硬盘时让高优先级的 IO 优先访盘。当出现硬盘繁忙，大部分 IO 在硬盘侧出现排队时，通过引入基于硬盘资源的优先级调度机制，可以减小高优先级 IO 的排队时间，总体上减小高优先级 IO 的时延。

优先级调度实现原理

SmartQoS 特性的优先级调度技术基于 LUN 的优先级实现。因此每个 LUN 都有一个优先级属性，这个属性由用户配置并保存在数据库中，当一个 IO 从主机（SCSI 目标器）发送到阵列，这个 IO 将会根据其归属的 LUN 来获得这个优先级属性，并且在整个 IO 路径上携带这个优先级信息。

首先，用户在创建 LUN 的时候需要指定所创建的 LUN 的优先级。如果用户不进行指定，则所创建的 LUN 的优先级属性默认为低优先级。LUN 创建后，其优先级属性可以根据用户的需要手动修改。

图3-12 IO 优先级调度原理图



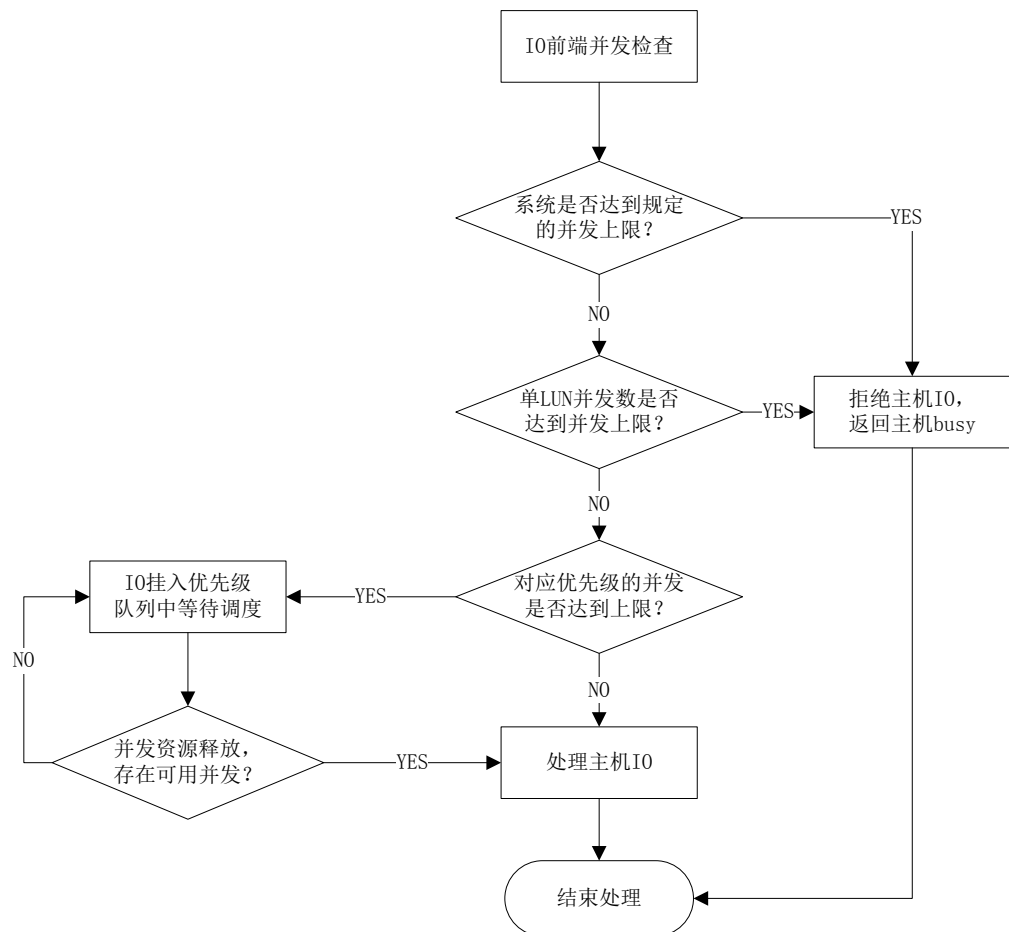
● 前端并发资源的优先级控制

前端并发资源的优先级控制在读写请求分配系统资源之前进行。在系统中，SmartQoS 针对每一个优先级的并发数维护一个计数值，当一个 IO 请求进入存储系统后，首先要从 SmartQoS 申请并发数，SmartQoS 会根据其占用的并发资源折算出一个并发数，并将此并发数累加到这个 IO 对应的优先级并发计数值上。当此 IO 被系统处理完成，在返回主机之前，其占用的并发数会从相应的并发计数中减除。

通过上述机制，每个优先级对应的并发计数值反映了这个优先级的业务在存储系统上占用的并发资源的数量。当每个优先级的并发计数值达到预设的最大值时，SmartQoS 会将这个 IO 请求挂入对应的优先级队列中排队等待，直到有 IO 返回，有空闲的并发资源后，再从队列中依次取出 IO 请求进行处理。同时，系统还维护了一个系统总并发计数和单个 LUN 的总并发计数，当其中一个达到最大限制时，SmartQoS 会拒绝这个 IO 请求，系统返回给主机繁忙信息，告知主机当前存储系统已经达到最大处理能力，此 IO 需要等待重试。

此外，SmartQoS 会根据一定规则对优先级的并发进行动态调整，以最大化并发资源利用率。初始时，系统为每个优先级分配较少数量的并发，剩余的并发置于公共并发池中，当某个优先级并发不够时，可以从并发池中获取，直到池中没有并发为止；同时，各优先级多余的并发也会及时归还到并发池中。当高优先级出现并发不足，且未达到其对应的最大并发数时，会强制其它优先级归还并发资源。

图3-13 并发资源优先级分配流程图



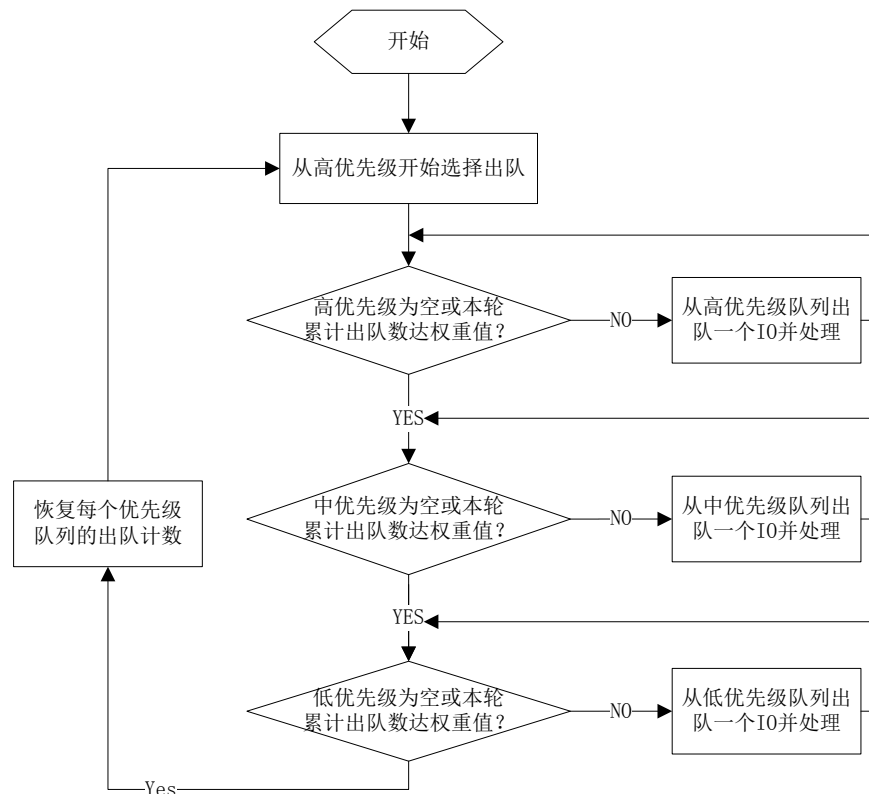
基于以上的设计原则，在并发资源充裕的情况下，可以尽量满足每个优先级的并发资源申请。而在并发资源紧张时，以确定的比例来给每个优先级分配并发资源，并确保较高的优先级能够获得相对更多的并发资源，业务需求更大的业务可以获得更多的并发资源。

● 计算资源的优先级控制

计算资源主要指系统中 CPU 的运行时间。当一个 IO 请求在需要获得 CPU 处理的时候，会根据其所携带的优先级信息被投入到相应的优先级等待队列中去，SmartQoS 根据 IO 调度算法每次从一个优先级等待队列中取一个 IO 请求进行处理。

SmartQoS 对 CPU 计算资源的控制是基于权重的优先级调度策略来实现的，调度策略会优先从高优先级的非空等待队列选择出队处理的 IO，即如果高优先级队列不为空，则从高优先级队列选取，如果高优先级队列的出队个数达到其权重值，则从下一个非空的优先级队列选择出队的 IO。按照这个规则直到低优先级，如果低优先级队列为空或低优先级队列的出队个数达到其权重值，则回到高优先级选择。

图3-14 CPU 资源优先级调度流程图



● 缓存资源的优先级控制

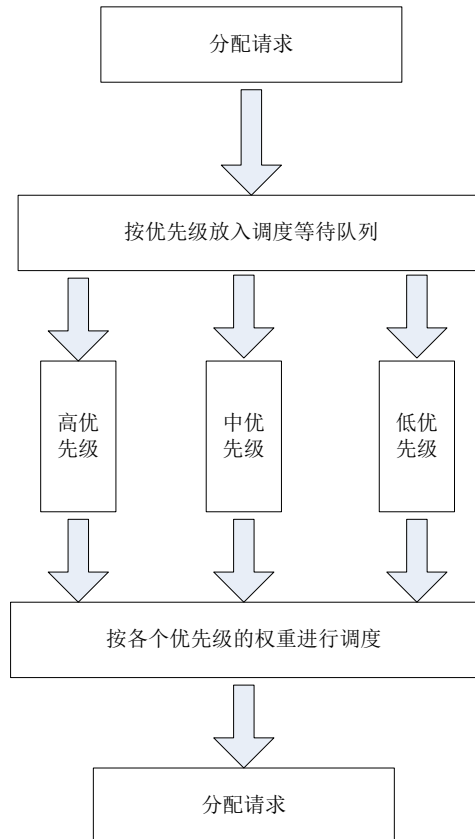
缓存资源主要指存储系统内部的 Cache 页面资源，由于写入存储系统的数据在写入硬盘之前都是存放在页面资源中的，因此缓存资源的优先级控制实际上就在 IO 申请页面的时候对页面的分配进行优先级控制。

缓存资源优先级控制也采用基于权重的优先级策略，每一种优先级对应一个权重值，每为一种优先级分配一定数量的页面，此优先级就消耗一定数量的权重，当一种优先级的权重消耗尽之后，此优先级将无法再得到调度，直到它的权重值得以恢复。

当 IO 申请页面的时候，SmartQoS 会根据这个 IO 携带的优先级信息将此 IO 投递到对应的优先级队列中等待分配。此外，SmartQoS 还会提供一个优先级调度策略，根据每个优先级的权重来选择当前应该从哪个优先级队列中调度一个 IO 来处理。

每种优先级的权重的恢复条件有两个：一是当所有优先级队列都为空时，将所有优先级的权重值恢复为初始值；二是当最低优先级的权重都耗尽时，将所有优先级的权重值恢复为初始值。

图3-15 页面分配优先级调度流程图



● 硬盘资源的优先级控制

SmartQoS 特性对于硬盘资源的控制，主要是通过 BDM（块设备管理）侧引入优先级调度机制，通过 BDM 的 PFE（优先级前向电梯）算法，来保证高优先级的 IO 能够被优先调度访盘。

PFE 算法即优先级前向电梯算法，此算法是在 Deadline 算法基础上发展出来的一种 IO 调度算法，总体性能不低于 Deadline，且在继承了 Deadline 算法优点（防止 IO 饿死、前向电梯保证顺序性）的同时，PFE 算法还能够支持 IO 优先级以及支持读 IO 优先的工作模式。

SmartQoS 特性 IO 流量控制技术

OceanStor V3 融合存储系统 SmartQoS 特性的 IO 流量控制技术通过限制存储系统中的一个 LUN 或者多个 LUN 的总体 IOPS 或者带宽，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。

IO 流量控制技术针对特定 LUN 上的数据业务限制可使用的数据处理资源，流控对象主要有以下两类：首先是需要限制的 IO 类型（限制读、限制写，或者同时限制读和

写)，其次是需要限制的流量类型（IOPS 或带宽），最终针对特定 LUN 得到一个流控限制的二元组（IO 类型，流量类型）。

SmartQoS 特性的 IO 流量控制技术根据确定好的二元组，通过 IO 分类来实现流量的限制。每个 IO 分类即对应一个流控组，即包含一定数量 LUN，并被设置了最大流量限制的 LUN 组。IO 分类流控的功能通过 IO 分类队列管理，令牌分发和出队控制几部分共同实现。

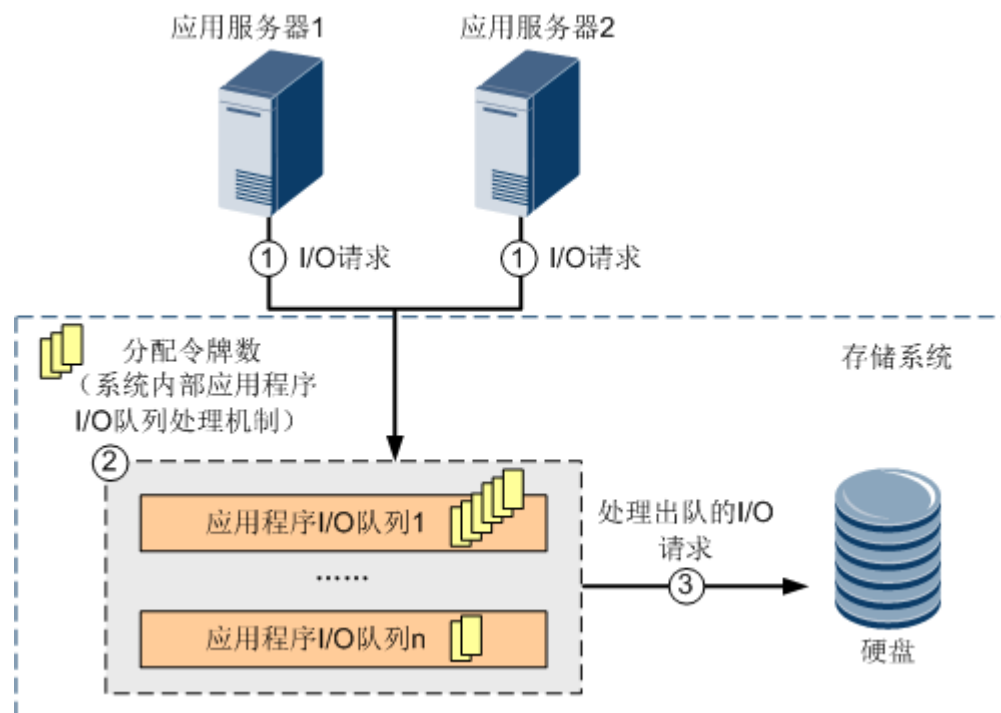
● 流控队列

每一个流控组对应于系统中的一个流控队列。所有发送到存储系统的 IO 请求首先会按照其所操作的 LUN 和读写类型进行分类，判断它是否属于系统中某个已经配置的并处于工作状态的流控组，如果是，这个 IO 将被投递到其对应的流控队列中等待处理。

SmartQoS 会实时地从每个流控队列中取出等待处理的 IO 进行下一步处理，一个流控队列中的 IO 能否被处理取决于这个队列对应的令牌桶中是否有令牌。

● 基于令牌桶的流控算法

图3-16 页面分配优先级调度流程图



SmartQoS 特性的 IO 流量控制技术的性能目标是基于令牌的分发和控制实现的。当用户为某个流控组设置了性能上限，那么这个性能上限会被转化成对应的令牌。在存储系统中，如果用户设定的性能目标是 IOPS，那么一个 IO 即对应一个令牌；如果设定的性能目标是带宽，那么一个扇区对应一个令牌。

每个流控队列都有一个令牌桶，SmartQoS 会定期向每个流控队列的令牌桶中放入一定数量的令牌，令牌的数量取决于用户设定的这个流控组的性能上限。即如果用户设定性能上限为 $IOPS = 10000$ ，则令牌分发算法就会在每秒将此流控组的令牌桶中的令牌数设置为 10000。

当流控队列进行出队处理的时候，会查看此队列的令牌桶中是否有足够的令牌，如果有，则取出一个 IO 进行处理，并消耗这个 IO 对应的令牌；如果没有令牌，则必须等待令牌桶中有足够的令牌可以用为止。

SmartQoS 特性 IO 时延控制技术

OceanStor V3 融合存储系统 SmartQoS 特性的 IO 时延控制技术通过保高限低的方式来保证某些关键业务的最低性能要求。用户可以为高优先级的业务设置最低性能指标，当此业务的最低性能指标无法达成之后，系统会通过依次限制低优先级和中优先级业务的性能来保障设置了最低性能目标的高优先级业务的性能。

SmartQoS 通过给中、低优先级的业务逐步增加时延的方式来做到对于其性能的限制，为了避免对系统性能产生较大抖动，当逐步增加的时延达到最大时延时，将不再进行增加；同时，当需要保证最低性能的业务性能达到最低性能指标的 1.2 倍时，系统将逐步消除中、低优先级所增加的时延。

3.3.2 SmartQoS 技术特点

SmartQoS 的技术优势主要体现在以下几个方面：

- 全 IO 路径的优先级调度：

SmartQoS 通过控制并发、计算、缓存、硬盘等关键系统资源的调度和分配，在整个 IO 路径进行优先级控制。相比友商的优先级控制策略，控制粒度更细，控制点更丰富彻底。

- 针对不同应用场景提供不同的 QoS 策略：

SmartQoS 细分用户场景，针对不同的应用场景提供不同的 QoS 策略。在用户有明确的性能指标需求时可以通过流量限制或者最小性能保障策略来满足用户需求；当用户无法精确指定性能指标时，可以基于应用的重要性采用优先级机制来保证高优先级业务的性能。

3.3.3 SmartQoS 技术应用场景

SmartQoS 特性适用于多种业务混合的应用场景，如：

- 多业务系统中确保关键业务的性能

基于保高限低的原则，将系统中的关键业务设置为高优先级，当系统内部出现资源抢占或者系统出现故障时，通过内部的优先级调度机制，保证高优先级的业务在整个系统内部能够被优先调度，从而保证关键业务的性能。

适合于一些没有明确性能目标，但是整个系统中的各个业务存在重要性之分的场景，比如企业内部的混合存储应用环境。

- 避免多业务之间的扰邻现象

基于限流机制，对于系统中的一些突发业务进行最大性能限制，避免突发业务流量对于系统中其它业务的冲击。

适合于一些存在明确性能限制目标的场景，比如在线业务和备份业务混合的场景，通过限制备份业务的最大突发流量，从而在确保备份时间窗的同时，减小对于在线业务的性能影响。

- 确保系统中关键业务的最低性能保证

基于保高限低的原则，当系统中的关键业务的最低性能指标无法保证时，通过对中、低优先级业务增加时延的方式，达到保证关键业务的具体性能指标。

适合于有明确性能需求的业务场景。

3.4 卷镜像技术

OceanStor 存储系统的卷镜像软件名称为 HyperMirror。

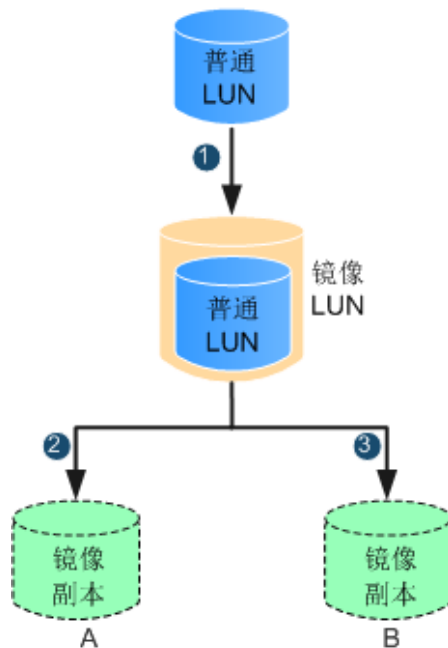
3.4.1 HyperMirror 工作原理

卷镜像的主要用途是为本地 LUN 或外部 LUN 提供多个可用的镜像副本。如果其中一个镜像副本故障不可用，主机仍然可以正常访问 LUN，主机侧业务无任何影响；同时，待故障镜像副本从故障中恢复后，镜像副本会自动同步镜像 LUN 的数据，最终达到镜像副本与镜像 LUN 的数据完全一致。

卷镜像的实现过程分为三个阶段：创建镜像 LUN、同步和分裂。

创建镜像 LUN

图3-17 镜像 LUN 的创建过程



1. 对一个普通 LUN（本地 LUN 或外部 LUN）执行创建镜像 LUN 操作，此时镜像 LUN 完全继承普通 LUN 的存储空间；同时继承普通 LUN 的基本属性和业务，主机侧不中断业务。
2. 创建镜像 LUN 过程中会在本地自动生成一个镜像副本 A，普通 LUN 变为镜像 LUN，并将数据存储空间交换到镜像副本 A，镜像 LUN 从镜像副本 A 中同步数据。

3. 此后需再给镜像 LUN 添加一个镜像副本 B，创建之初从镜像副本 A 同步数据。此时普通 LUN 具有空间镜像功能，同时拥有镜像副本 A 和镜像副本 B 两份镜像数据。

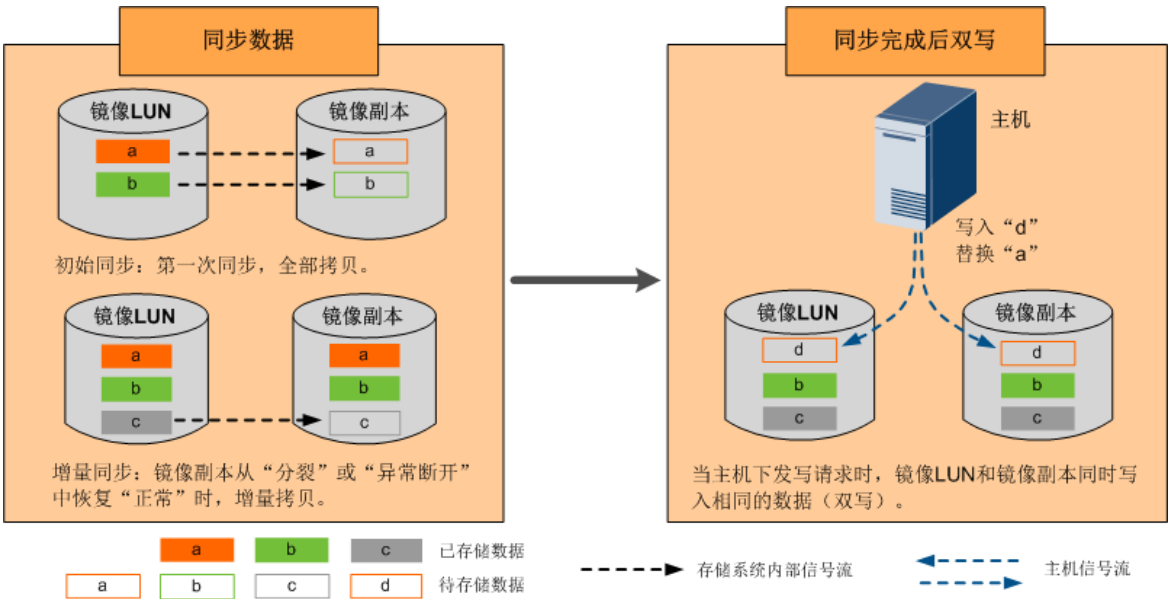
镜像 LUN 创建完成后，主机下发 I/O 的情况：

1. 当主机对镜像 LUN 下发读请求时，存储系统会以轮询方式在镜像 LUN 和镜像副本之间进行读操作。当镜像 LUN 或者某个镜像副本故障时，主机侧业务不受影响。
2. 当主机对镜像 LUN 下发写请求时，存储系统会以双写方式对镜像 LUN 和镜像副本进行写操作。

同步

同步过程的原理如下图所示。当一个副本从故障恢复或从数据不完整恢复到数据完整的过程中，增量从完整的镜像副本同步数据，最终达到镜像副本间的数据完全一致。

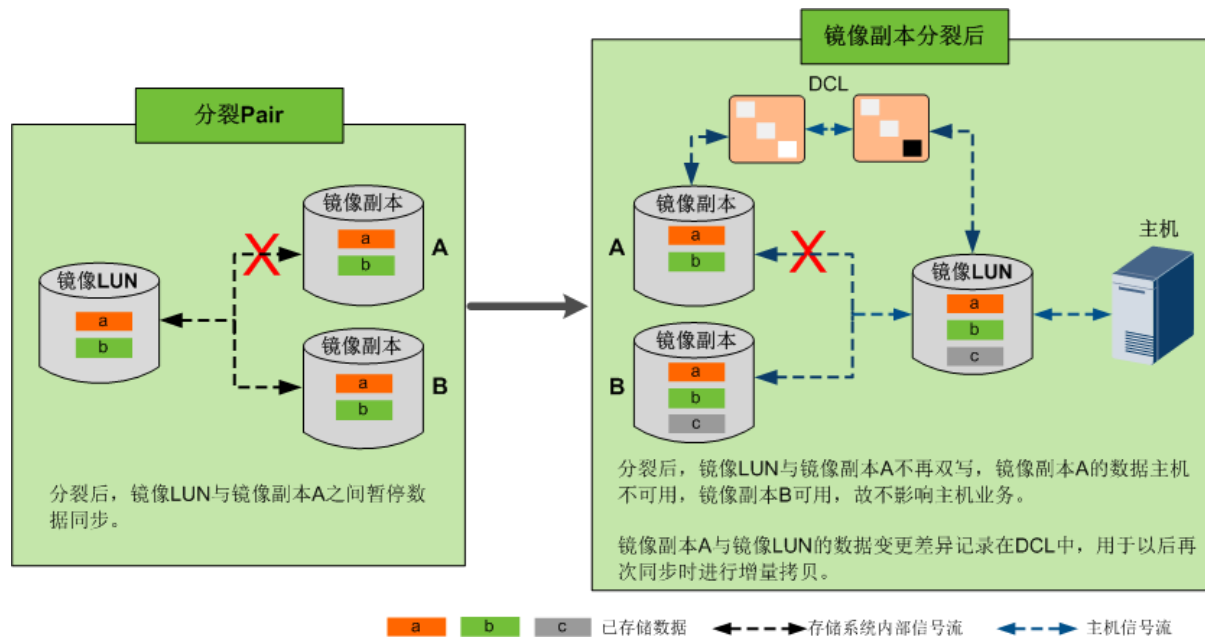
图3-18 同步过程的原理



分裂

在业务需要隔离一个副本时，可执行分裂操作，将副本从镜像中隔离，此时镜像不具备数据镜像功能，同时镜像副本记录此后的数据差异，当需要重新建立镜像关系时，根据差异增量同步差异数据。

图3-19 分裂过程的原理



3.4.2 HyperMirror 技术特点

- 存储系统内数据可靠性

通过对业务 LUN 创建卷镜像，使其同时拥有两份完全独立的数据副本，其中一份数据副本故障，不影响主机业务的运行，大大提高了数据的可靠性。

- 异构阵列的可靠性

通过 SmartVirtualization 特性，将异构阵列的 LUN 接管，对其创建镜像 LUN，创建本一个本地镜像副本，防止由于异构阵列不稳定或链路异常导致的业务中断。

- 主机性能影响小

镜像数据副本位于业务 LUN 的 CACHE 之下实现，且镜像空间之间实现并发写和轮询读技术，不影响主机业务性能。

- 主机业务连续性

通过在线将正在运行业务的 LUN 创建镜像副本，主机业务不感知 LUN 数据空间的变化。

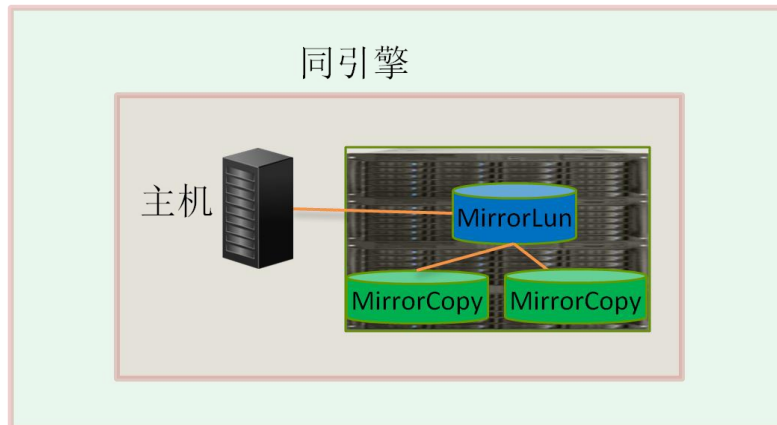
3.4.3 HyperMirror 应用场景

本地 LUN 数据镜像

为了解决单 LUN 的双盘失效导致的业务中断和数据丢失，采用为本阵列的 LUN 寻找一个伙伴，即在本阵列再建立一个和源 LUN 容量大小一样的 LUN，两者建立镜像关系，通过冗余提升 LUN 的可靠性。在这种组网方式下，建议两个镜像副本数据存放

在不同的故障域中（即两个 LUN 存在于不同的 diskDomain），并且使用性能参数相同的环境（如磁盘类型、Raid 级别、磁盘个数）。

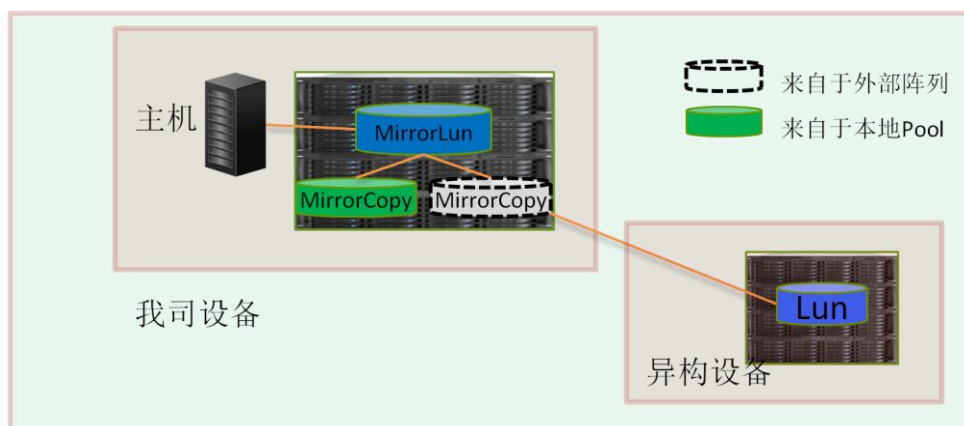
图3-20 本地 LUN 数据镜像



异构 LUN 可靠性

随着当前数据量越来越大，网络环境越来越复杂，客户环境中存在各厂家的存储设备，为了便于集中管理，统一控制，会采用异构接管了外部阵列的 LUN 后，在一个阵列上统一管控，如下图。由于外部阵列的多样性，很可能出现意料外的兼容性问题，导致客户业务中断甚至丢失。那么可以使用本阵列上的存储阵列本地 LUN 和异构接管 LUN 组成镜像 LUN，将镜像 LUN 映射给主机读写，通过数据冗余提高了异构接管 LUN 的可靠性，同时保证数据的高可用。由于是在双写都成功后才返回主机 IO 响应，所以通常使用 FC 链路，且在同城范围以内。

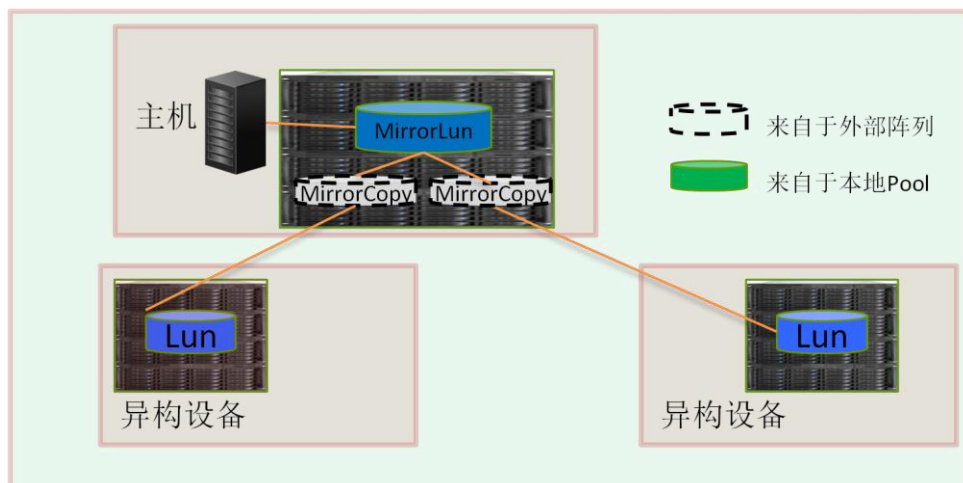
图3-21 异构 LUN 可靠性



异构阵列间镜像

单个阵列可靠性较低，为了充分利用就设备，将多个异构阵列组成冗余备份关系，提高单个阵列的可靠性。则如下配置：使用两个接管的异构阵列 LUN 构建镜像关系，将镜像 LUN 映射给主机读写，通过冗余提升异构 LUN 的可靠性，解决连接异构设备的链路故障或异构设备自身原因导致的故障。

图3-22 异构阵列间镜像



3.5 缓存分区技术

SmartPartition（智能缓存分区）是华为技术有限公司（以下简称华为）OceanStor V3 融合存储系统为应对存储融合趋势下 QoS 的挑战而设计的智能缓存分区技术，其核心思想是通过对系统核心资源的分区，保证关键应用的性能。管理员可以针对不同的应用配置不同大小的缓存分区，系统将保证该分区中的缓存资源被该应用独占，并根据业务实际情况实时动态调配不同分区中的前后端并发，从而保证位于该分区的应用性能。

SmartPartition 还可以与 OceanStor V3 融合存储系统的其他 QoS 技术（如智能服务质量控制 SmartQoS）相配合，从而达到更好的服务质量保证效果。

SmartPartition 本质上就是一种 Cache 分区技术，Cache 分区在业界是比较成熟的一项技术，推出的时间也比较长。当前主流的存储厂商，比如 EMC 和 HDS，均推出了各自的 Cache 分区特性。从技术层面来看，各大厂商实现的方式基本一致，都是将有限的 Cache 资源划分为多个逻辑区域进行管理，区别主要在于 Cache 分区内部的具体的一些算法和调整策略。

3.5.1 SmartPartition 工作原理

Cache 分区技术通过隔离不同的业务所需要的资源，保证某些关键业务的服务质量。

在存储系统中，影响某个业务服务质量的主要因素是该业务对以下几个资源的占有量：

主机并发：该业务当前有多少主机 I/O 可以在存储内并行执行，更大的并发通常意味着更好的服务质量。如果存储分配给一个业务的主机并发太小，会导致主机侧 I/O 延时的增加；反之，则意味着存储资源的浪费。

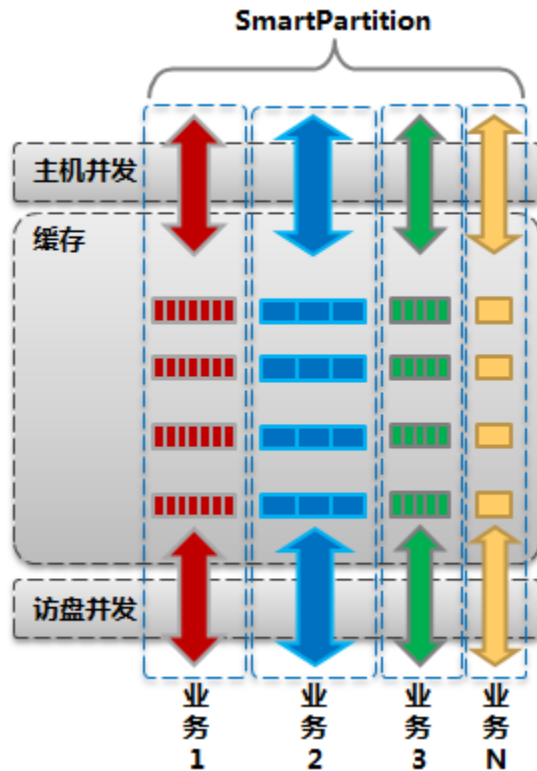
缓存：该业务在该存储系统上可以占有的缓存大小。缓存是影响存储系统性能的最主要因素：对写业务来说，更多的缓存意味着更高的写合并率、写命中率（同一块数据在缓存中被再次写中的比率）和更好的访盘顺序度；对读业务来说，更多的缓存通常意味着更高的读命中率。同时，不同类型的业务对缓存的需求也有很大不同：对顺序类业务来说，缓存不需要很大，只需要满足 I/O 合并要求即可；而对随机类业务来说，更大的缓存通常意味着更好的访盘顺序度，从而带来性能的提升。

访盘并发：该业务当前可以有多少 I/O 可以并发访问后端硬盘，体现了该业务对后端性能的占有能力。这个并发需要与主机并发匹配，即保证后端能力能满足前端要求，且不会有浪费。

合理分配这三个关键资源是提高存储系统服务质量的主要手段。

SmartPartition 可以针对不同的业务（实际控制对象为 LUN）分配不同大小的缓存分区资源，OceanStor V3 融合存储系统会根据分区大小与实际的 I/O 流自动分配合理的主机并发与访盘并发，从而保证关键业务的服务质量。

图3-23 SmartPartition 工作原理



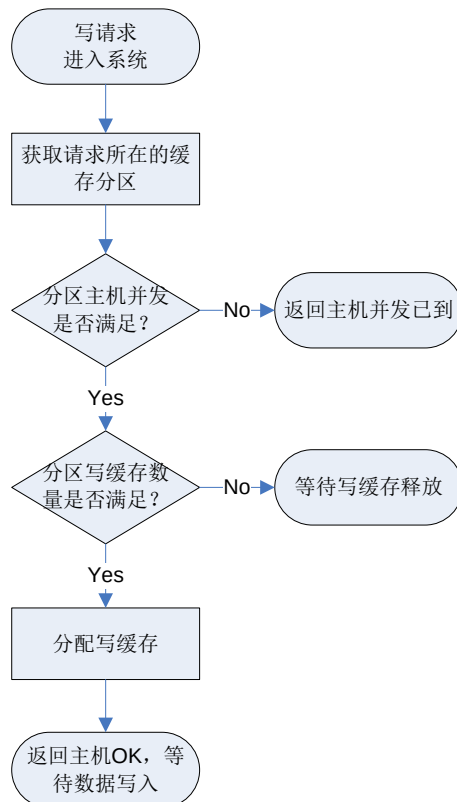
SmartPartition 读写缓存 I/O 流程

缓存从类型上分为读缓存和写缓存。读缓存的主要作用是通过读预取及数据保有提高主机读 I/O 的命中率；写缓存的主要作用是通过合并、命中、排序等手段提高访盘性能。不同的业务对读写缓存大小有不同的要求，SmartPartition 支持用户单独为分区配置读、写缓存的大小，从而满足不同类型业务的要求。

由于目的不同，配置读写缓存对 I/O 的处理流程也是不同的。

配置写缓存后对 I/O 过程的处理如下：

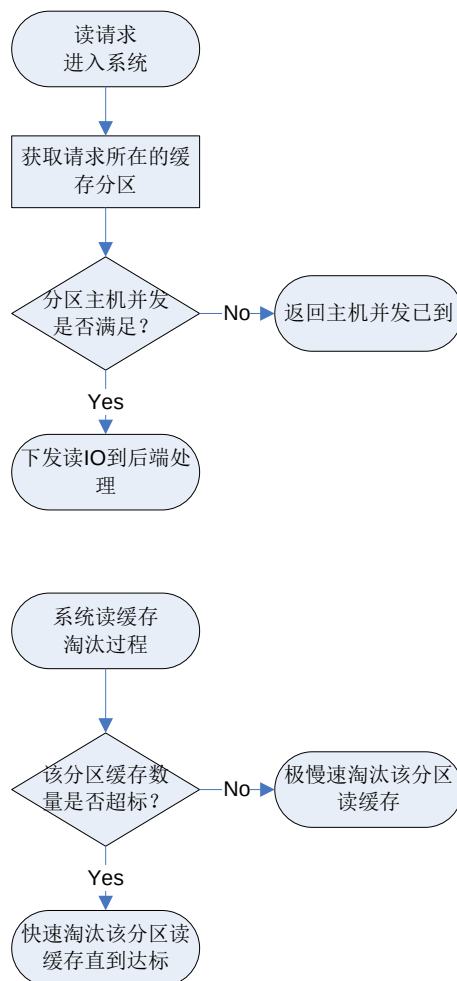
图3-24 写缓存后对 I/O 过程的处理



对写 I/O 的影响主要在缓存资源分配阶段，需要做分区主机并发的判断和写缓存数量的判断。选择在这个阶段处理的原因在于这是写过程的最初阶段，且该阶段中并没有实际占用存储缓存。如果放到后面的阶段，由于回写的原因，是无法控制业务占用的写缓存数量的。

配置读缓存后对 I/O 过程的影响：

图3-25 配置读缓存后对 I/O 过程



对读 I/O 的影响分为两个部分：第一个部分与写 I/O 类似，需要判断分区主机并发是否满足，如果不满足，则需要将 I/O 返回。由于读缓存的目的是控制读数据占用的缓存大小，而读缓存的大小是由读缓存淘汰过程控制的，所以第二个部分在淘汰过程中控制。如果分区内的读缓存没有达到设定值，此时需要以极慢的速度淘汰；反之，则需要快速淘汰，保证读缓存数量在设定值以内。

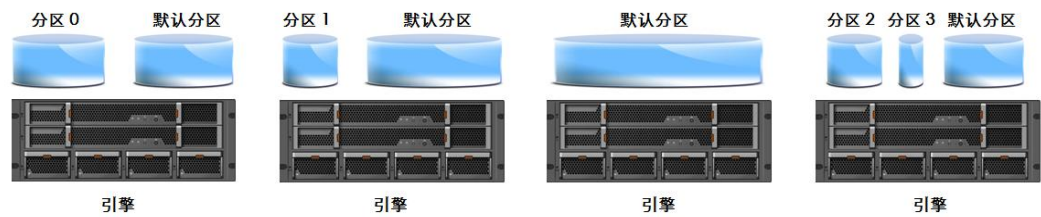
说明

系统不允许将某分区的读或写缓存配置为 0，最小的规格为 256MB。

SmartPartition 在系统中的分布

OceanStor V3 融合存储系统采用多控制器 Scale-out 架构，在整个系统中，缓存在物理上是分离到各个引擎中的，因此，SmartPartition 所设置的分区也是分布在各个引擎中的，其分布方式如下图所示：

图3-26 SmartPartition 在系统中的分布



OceanStor V3 融合存储系统最多支持 64 个分区，每个引擎最多可创建 8 个用户分区，但是一个分区只能分布在一个引擎上。

OceanStor V3 融合存储系统会为每个引擎预留一定量的分区资源配额以保证最基本的 I/O 处理能力，剩余的分区资源都保留为一个默认分区，以满足不在用户分区中的 LUN 的 I/O 需求。用户可以从默认分区中配置用户分区（读分区/写分区）的大小，但必须保证默认分区的大小要大于或者等于可分配分区资源的 50%。

SmartPartition 主机并发控制

相对于主机应用，存储系统的处理资源是有限的。所以，存储需要限制系统内总的主机并发。针对每个分区，当然也需要控制其并发以保证服务质量。

SmartPartition 中分区的主机并发并不是固定的，而是基于多种因素，采用优先级加权算法计算出来的，这些因素包括：

- 上个统计周期内分区中活动的 LUN 的个数
- 上个统计周期内分区中活动的 LUN 的优先级
- 上个统计周期内各 LUN 的完成 I/O 个数
- 上个统计周期内各 LUN 由于分区并发已到而返回主机的 I/O 个数

通过这些因素的加权，可以实现系统内主机并发最大化利用的同时兼顾分区的服务质量保证。

当一个统计周期结束后，分区可能需要根据新的统计结果调整其并发。这个过程由 SmartPartition 逻辑控制，按一定并发步长逐步调整，以保证其调整尽量平滑，避免对主机造成强烈的性能波动。

SmartPartition 后端并发控制

与主机并发控制类似，后端并发控制的目的是最大化利用系统资源的同时兼顾分区的服务质量保证。这个并发也是多种因素按优先级加权计算出来的，这些因素包括：

上个统计周期内分区中各种优先级 LUN 的脏数据数量

上个统计周期内分区中 LUN 的刷盘时延

上个统计周期内分区中 LUN 的实际刷盘并发

调整周期和调整方式也与主机并发控制类似，不再赘述。

3.5.2 SmartPartition 技术特点

SmartPartition 的技术优势主要体现在以下几个方面：

- 智能的分区控制：

SmartPartition 不仅仅是缓存的分区，而是系统核心资源的分区。SmartPartition 根据用户配置的缓存大小以及其他 QoS 策略自动调配系统并发资源，达到系统服务质量的最大化和分区质量保证达标。

- 简便易用：

SmartPartition 配置简单，用户界面友好，所有操作立即生效，系统自动将分区的创建、归还、切换以及并发调整等复杂的流程和操作在后台实现，不需要用户参与，从而大大提升了分区功能的易用性。

3.5.3 SmartPartition 技术应用场景

SmartPartition 功能适用多种应用混合的应用场景，如：

- 多业务系统中确保核心业务性能。
- VDI 场景中针对重要客户的性能保证。
- 云计算系统中的多租户场景。

3.6 多租户技术

SmartMulti-Tenant（多租户）是华为公司在 T 系列阵列上实现基于管理面隔离的多租户功能，它将一部分资源的管理和监控功能下放给各个租户，从而降低人员的资源管理成本。

3.6.1 多租户工作原理

多租户的设计层面分为数据面、程序面、系统面、管理面，而华为公司在当前版本仅实现了基于管理面的功能。

在设计多租户管理功能时，需要考虑以下几个用户面临的普遍调整：

租户间数据隔离：租户间的资源数据分域管理

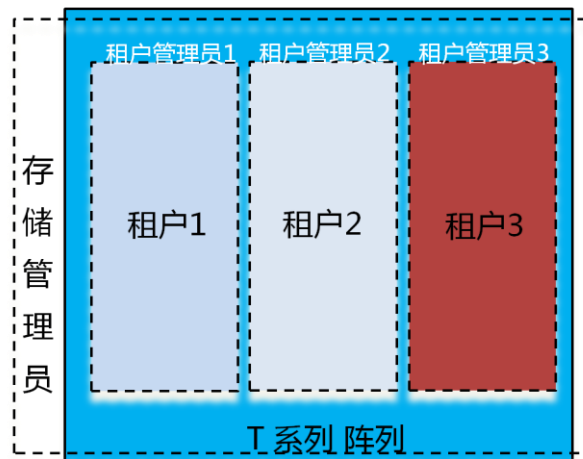
租户间权限隔离：租户管理员只能管理和查看租户内的数据

业务功能权限范围：对开放给租户的业务管理功能有效限制

简单来说，分权分域有效隔离租户间的资源数据，将一部分资源的管理功能下放给各自的租户，由租户的管理员对租户内的资源进行自我管理，不仅能提高资源的管理效率，还能将存储管理员从繁琐的资源管理工作中解脱出来。

从管理面区分租户和存储管理员的职能划分，如下图所示：

图3-27 多租户管理面



其中存储管理员能对阵列内所有资源进行管理，包括对所有租户的管理。而租户内只能由租户管理员对资源进行管理。

3.6.2 多租户技术特点

存储管理员在对大量存储资源进行管理的时候，疲于应付租户各式各样的管理需求，加大了存储管理员的工作负荷，同时也容易因为误操作产生诸多问题。如果把一部分管理功能下放给租户自我管理，不仅减轻了存储管理员的负担，同时也降低了管理成本。多租户的分权分域和资源管理两大特性能有效提升多租户管理能力。

分权分域

分权分域功能由网管系统的 vStore 模块完成。每一个 vStore 就对应一个租户。网管系统允许存储管理员创建 vStore 实例，并且给每个实例分配一个租户管理员，租户管理员通过统一登录入口登录管理系统，可以对 vStore 实例内的所有存储资源进行管理。

提供给租户的管理视图，包括以下功能：

- 磁盘（LUN）管理：可以查看磁盘的状态和容量，并且可以查看磁盘映射给主机的信息列表。
- 性能监控：可以查看磁盘的读带宽和写带宽性能数据。

提供给存储管理员的管理 vStore 的视图，包括以下功能：

- 支持给 vStore 分配一位租户管理员，可以对在线的租户管理员进行强制下线操作。
- 支持给 vStore 分配磁盘（LUN）。已分配给租户的磁盘，可以在租户视图下进行管理。

资源管理

资源管理功能，涉及对资源的分配、调优、映射等管理功能，在一定程度上如果全部放开给租户，可能会出现由于资源滥用，随意分配等情况导致的资源闲置等一系

列的问题。所以华为公司在对资源管理功能，没有完全对租户开放。仅给租户提供资源查询能力。

在当前版本提供的资源管理能力，包括如下：

- 支持查询磁盘（LUN）
- 支持查询磁盘（LUN）和主机映射关系
- 支持实时查看磁盘的读带宽和写带宽能力

4 缩略语

缩略语	英文解释	中文解释
eDevLun	External Device Lun	外部设备逻辑单位号
IOPS	Input and Output Per Second	每秒输入输出数
LUN	Logical Unit Number	逻辑单元号
OLAP	On Line Analysis Process	在线分析处理
OLTP	On Line Transaction Process	在线交易处理
QoS	Quality of Service	服务质量保障
RAID	Redundant Array of Independent Disks	独立磁盘冗余阵列
SAS	Serial Attached SCSI	串行 SCSI
SCSI	Small Computer System Interface	小型计算机系统接口
SSD	Solid State Disk	固态硬盘
TCO	Total Cost Ownership	整体拥有成本
WWN	World Wide Number	全球唯一标示号
RTO	Recovery Time Object	目标恢复时间
RPO	Recovery Point Object	目标恢复点
SAN	Storage Area Network	存储区域网络
FCoE	Fibre Channel over Ethernet	以太网光纤通道
VDI	Virtual Desktop Infrastructure	虚拟桌面设备
QoS	Quality of Service	服务质量

缩略语	英文解释	中文解释
BDM	Block Device Management	块设备管理
PFE	Priority Forward Elevator	优先级前向电梯