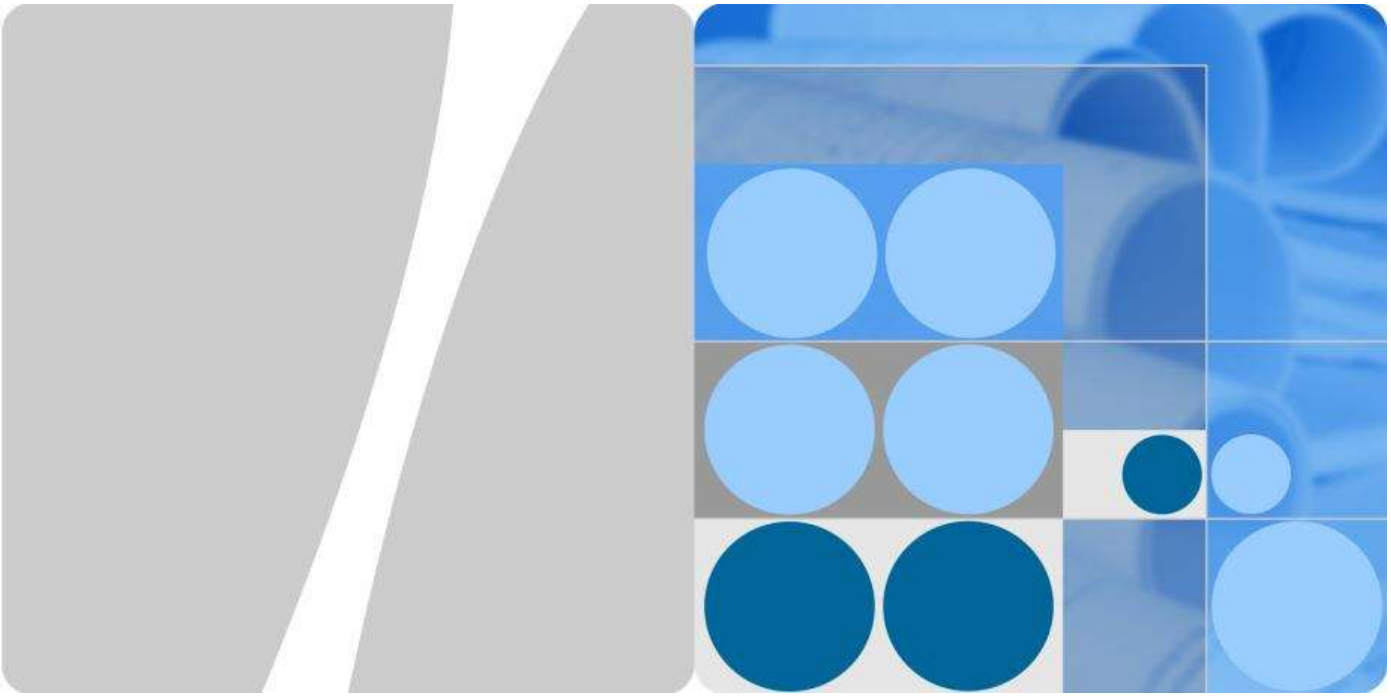


OceanStor 高端存储系统

SmartPartition 技术白皮书

文档版本 01

发布日期 2013-10-12

华为技术有限公司

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://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

目 录

1 缓存分区技术概述 1

1.1 IT 架构的融合趋势对存储的要求 1

1.2 SmartPartition 技术简介 2

2 SmartPartition 技术原理 3

2.1 SmartPartition 工作原理简介 3

2.2 SmartPartition 读写缓存 IO 流程 4

2.3 SmartPartition 在系统中的分布 6

2.4 SmartPartition 主机并发控制 7

2.5 SmartPartition 后端并发控制 7

3 SmartPartition 配置 8

3.1 创建 SmartPartiton 分区 8

3.2 修改 LUN 归属的 SmartPartiton 分区 9

3.3 修改 SmartPartiton 分区大小 9

3.4 删除 SmartPartiton 分区 10

3.5 SmartPartiton 配置限制 10

4 SmartPartition 技术特点及应用场景 11

4.1 SmartPartition 技术特点 11

4.2 SmartPartition 技术应用场景 11

5 缩略语表/Acronyms and Abbreviations 12

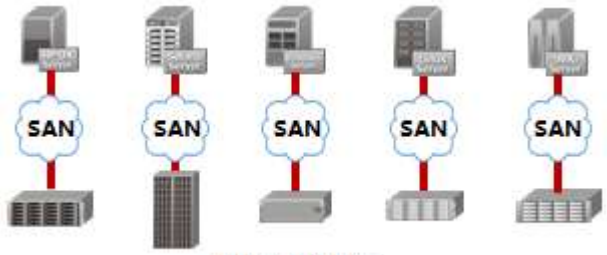
修订记录/Change History

日期	修订版本	描述	作者
20131012	V1.0		王雪松/61123 秦烜/204091
20131014	V1.1	产品更名	秦烜/204091

1 缓存分区技术概述

1.1 IT 架构的融合趋势对存储的要求

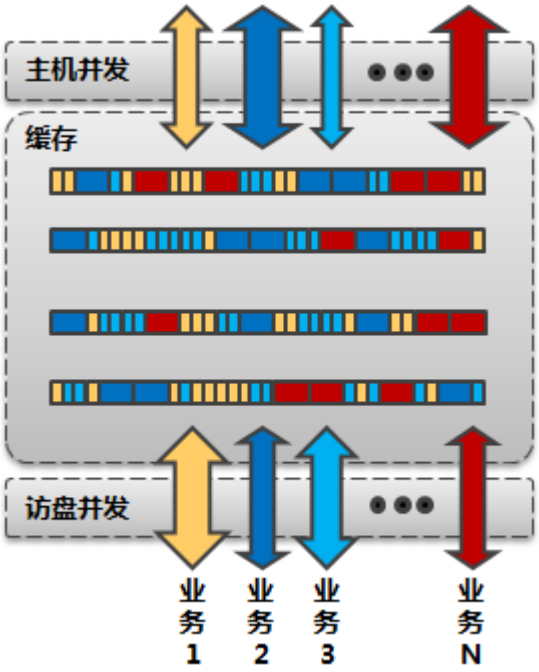
典型的 IT 系统架构由计算、网络、存储三大部分组成，在传统的“烟囱式”架构中，不同的应用系统相对独立，单个存储需要面对的应用数量不多（一般在 5 个以下）。



传统数据中心

然而，在**虚拟化**和 **FCoE** 等技术的推动下，IT 架构中计算的融合和网络的融合已成必然，这种大融合的趋势也导致了存储的融合；其次，为了更加灵活地分配存储资源，简化管理，同时为将来的数据备份、容灾做好准备，越来越多的数据中心开始采用数据集中存储的方式，其结果就是单个存储系统要面对的应用数量剧增，有些甚至是数量级上的变化，比如针对 VDI（Virtual Desktop Infrastructure）环境的集中存储，可能需要同时对上千个应用提供服务，启动风暴带来的海量并发 IO，以及桌面应用负载特征的巨大差异（大 IO/小 IO， 随机/顺序等），都给存储系统的性能和服务质量保障能力（QoS - Quality of Service）提出了更高的要求。为了应对这种趋势的变化，存储业界提出了很多技术手段，比如 IO 优先级、应用限流和分区技术等。

下图展示了多业务混合场景下存储面临的性能问题：



其中，每种颜色代表了一种不同的业务。可以看到，这些业务的 IO 模式是不同的，对主机并发、缓存及访盘并发的需求也是不同的。多种业务混合争抢并发资源和缓存资源，导致服务质量不可保证。比如，业务 N 抢占的访盘并发同其抢占的主机并发并不匹配，其结果就是缓存中业务 N 的数据会越来越多，对其他业务性能造成影响；而业务 3 抢占的主机并发非常小，如果是关键业务，就意味着其业务性能不能得到保证。

1.2 SmartPartition 技术简介

SmartPartition（智能缓存分区）是华为技术有限公司（以下简称华为）OceanStor 高端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 高端存储系统的其他 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。

2 SmartPartition 技术原理

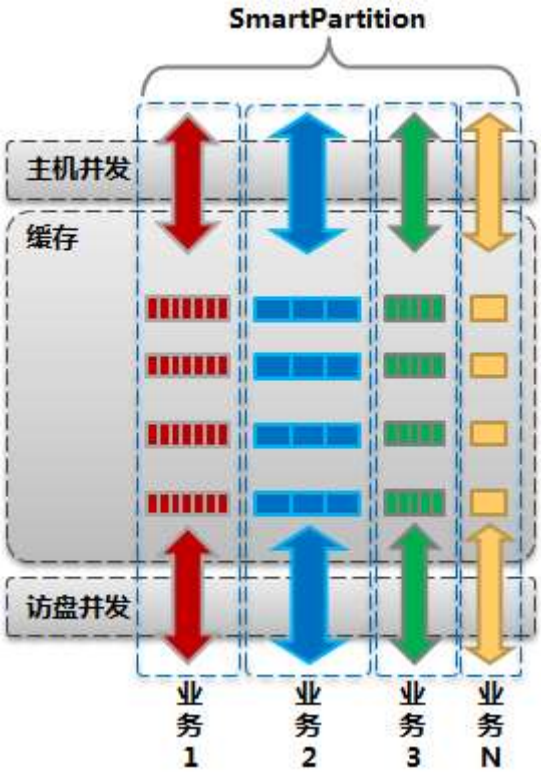
2.1 SmartPartition 工作原理简介

分区技术被设计来通过隔离不同的业务所需要的资源，保证某些业务的服务质量。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 IO 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 IO 延时的增加；反之，则意味着存储资源的浪费。
- **缓存**：该业务当前可以占有存储的缓存大小。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好的写合并率、写命中率和更好的访盘顺序度；对读业务来说，更多的缓存意味着更好的读命中率。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，缓存数量不需要很大，只需要满足 IO 合并要求即可；而对随机类业务来说，更大的缓存数量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 IO 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。

合理的分配存储系统中这几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 高端存储系统会自动根据分区大小与实际的 IO 流合理分配主机并发与访盘并发，保证关键业务的服务质量。

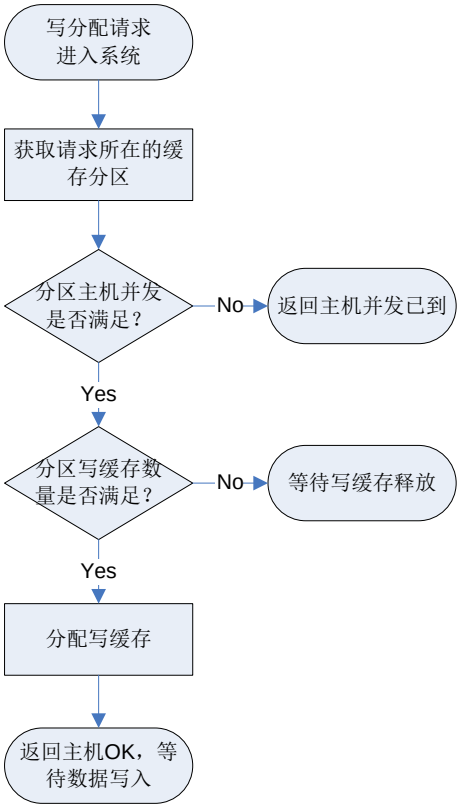


2.2 SmartPartition 读写缓存 IO 流程

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 IO 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。

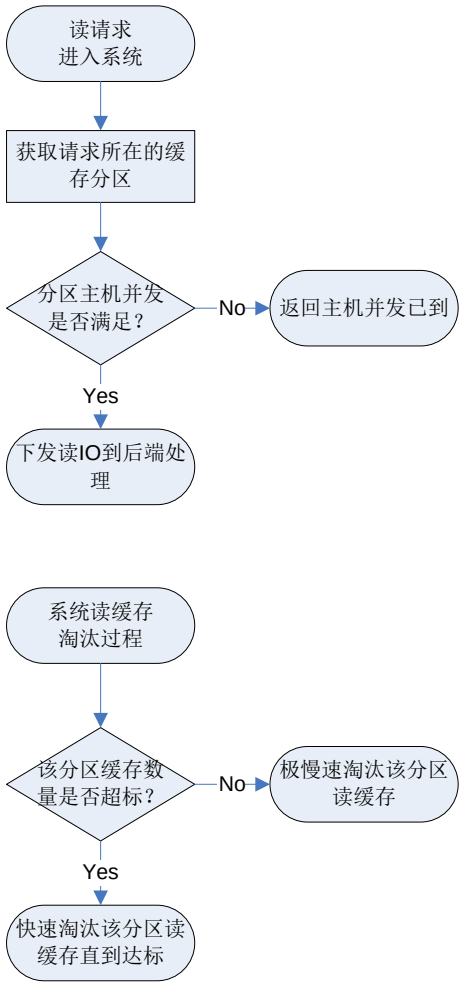
由于目的不同，配置读写缓存对 IO 的处理流程也是不同的。

配置写缓存后对 IO 过程的处理如下：



对写 IO 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 IO 过程的影响：



对读 IO 的影响分为两个部分：第一个部分与写 IO 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 IO 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。

2.3 SmartPartition 在系统中的分布

OceanStor 高端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：

OceanStor 高端存储系统最多支持 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 高端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 IO 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 IO 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不小于可分配分区资源的 50%。

2.4 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 IO 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 IO 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。

2.5 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也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：

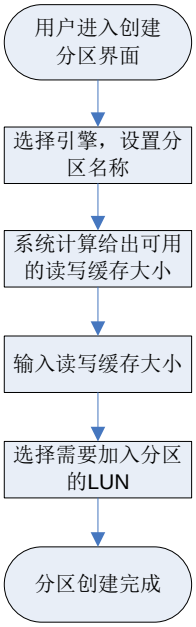
- 上个统计周期内分区中各种优先级 LUN 的脏数据数量
- 上个统计周期内分区中 LUN 的刷盘时延
- 上个统计周期内分区中 LUN 的实际刷盘并发

调整周期和调整方式也与主机并发控制类似，不再赘述。

3 SmartPartition 配置

本章描述了 SmartPartition 的最主要配置流程。

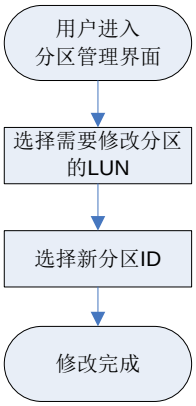
3.1 创建 SmartPartiton 分区



分区创建后立即可用。其缓存由系统在后台协调，不需要全部分配完成才能使用。

LUN 可以在创建分区的时候加入分区，也可以分区创建好之后再加入分区。

3.2 修改 LUN 归属的 SmartPartiton 分区

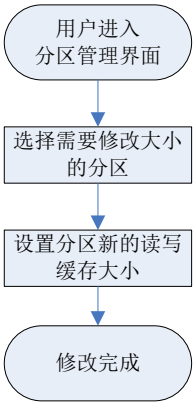


LUN 移入、移出分区过程不需要停止 IO 或刷盘，数据配额转换后系统后台自动完成。

移入新分区时，LUN 上当前的读写配额会占用新分区的配额。如果新分区的缓存出现了超出配置配额的情况，系统会启动后台归还机制，以一定比例平缓的将配额归还给系统，最终将配额控制到配置配额内。

OceanStor 高端存储系统支持在一个引擎内修改 LUN 的归属分区，支持默认分区修改到用户分区，或从用户分区修改到默认分区。

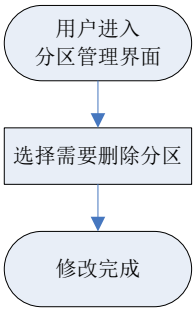
3.3 修改 SmartPartiton 分区大小



分区修改后的大小也遵循创建分区时的原则，修改后立即生效。

分区修改后也可能导致实际占用的缓存超出配置配额，这种情况下同样会启动后台归还机制。

3.4 删除 SmartPartiton 分区



分区删除立即生效，删除后其配额将归还给系统默认分区。

说明

分区删除前需确保没有 LUN 归属在该分区中，否则删除会失败。

3.5 SmartPartiton 配置限制

SmartPartition 配置限制如下：

- 读分区、写分区大小最小为 256MB。
- 一个 LUN 只能归属到与它的工作引擎 ID 一致的分区上。
- 一个 LUN 只能在用户分区和默认分区间切换。
- 删除一个分区前，必须把该分区上的所有 LUN 均移出该分区。

4 SmartPartition 技术特点及应用场景

4.1 SmartPartition 技术特点

SmartPartition 的技术优势主要体现在以下几个方面：

(一) 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。

(二) 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。

4.2 SmartPartition 技术应用场景

SmartPartition 功能适用多种应用混合的应用场景，如。

（1）多业务系统中确保核心业务性能。

（2）VDI 场景中针对重要客户的性能保证。

（3）云计算系统中的多租户场景。

5 缩略语表/Acronyms and Abbreviations

表5-1 缩略语清单

英文缩写	英文全称	中文全称
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