



基于分布式块存储系统的服务质量研究与优化

屠要峰^{1,2}, 杨洪章^{1,2}, 韩银俊^{1,2}, 陈正华^{1,2}

(1.中兴通讯股份有限公司, 广东 深圳 518057;

2.南京中兴新软件有限责任公司, 江苏 南京 210012)

摘要: 在云存储技术中, 资源复用和共享是一把双刃剑, 在为用户降低开销的同时, 也导致了应用之间的相互干扰和竞争, 特别是在资源需求量超出存储系统能力时, 无法提供令人满意的服务质量。提出一种自适应的任务调度技术, 解决了前后端任务的资源平衡问题。还提出一种多级热度的缓存替换技术, 解决了前端任务间的资源平衡问题。在中兴分布式块存储系统中将上述两项技术进行了工程实现和商用测试, 上述优化方案满足了用户对分布式块存储服务质量的期望。

关键词: 分布式存储系统; 缓存分级; 任务调度; 服务质量

中图分类号: TP316.4

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018157

Research and optimization on quality of service in distributed block storage system

TU Yaofeng^{1,2}, YANG Hongzhang^{1,2}, HAN Yinjun^{1,2}, CHEN Zhenghua^{1,2}

1. ZTE Corporation, Shenzhen 518057, China

2. Nanjing Zhongxing New Software Co., Ltd., Nanjing 210012, China

Abstract: In cloud storage technology, resource reuse and sharing is a double-edged sword. While it reducing the overhead for customers, it leads to a situation that the different applications will compete for hardware resources, especially when resource requirement exceeded storage system capability. As a consequence the cloud storage system is unable to provide satisfactory quality of service. An adaptive scheduling technology was proposed, solving the resource balance problem between front tasks and back tasks. A multilevel popularity cache replacement technology was also proposed, solving the resource balance problems between the front tasks themselves. These technology was implemented and tested on the ZTE distributed block storage system. The proposed technology satisfies the expectations of the quality of service from the users.

Key words: distributed storage system, cache tier, task scheduling, quality of service

收稿日期: 2017-12-09; 修回日期: 2018-04-14

通信作者: 杨洪章, yang.hongzhang@zte.com.cn

基金项目: 深圳市科技创新委员会科技应用示范项目资金资助项目 (No.SF20170036)

Foundation Item: Applied Demonstration Programs of Science and Technology Innovation Commission Foundation of Shenzhen City (No.SF20170036)

1 引言

1.1 存储服务质量的迫切需求

自从 2006 年被谷歌首次提出^[1]，“云计算”已悄然发展了 10 年以上，以云桌面、云硬盘、超融合等形态为代表的云计算产品已经步入现代生活的方方面面，按需付费的理念深入人心。软件定义存储^[2-4]作为支撑云计算各项任务的关键技术，成为当仁不让的研究热点。软件定义存储采用虚拟化技术将分布于不同地域的计算机资源进行集中管理和分配，通过资源复用的方式，有效地提高资源利用率。软件定义存储可根据接口分成 3 类^[5]：分布式对象存储系统、分布式文件系统、分布式块存储系统。在云计算的应用环境中，分布式块存储系统是最主要的底层存储系统，它提供弹性、线性的逻辑地址空间。

由于软件定义存储将标准 x86 硬件资源池化、虚拟化、服务化，在同一套云平台中将会支持海量的、并发的、复杂的应用，在公有云抑或私有云中，都十分普遍。通过存储资源复用和共享的方式，用户不需要为每个应用部署专门的存储系统，也降低了软硬件维护代价。然而，资源复用是一把双刃剑^[6]，在为用户降低开销的同时，也导致了应用之间的相互干扰和竞争，难以保障用户的服务质量需求。

从社会大众的认知层面来看，已将存储资源视为与水、电、煤气近似的公共基础设施，用户对该类按需付费的资源有着异曲同工的希望——满足峰值需求和稳定需求。以电力系统为例，用户的基本期望是电压能保证家中所有家用电器同时工作，并且电压保持稳定，不能忽高忽低。再以云存储资源为例，用户的期望是吞吐量能同时保证 VR 视频播放、云桌面办公、大型游戏等应用的需求，且吞吐量保持稳定，不能时断时续。因此，提供高效、稳定的存储服务质量成为用户对云计算产品、对底层分布

式块存储系统的核心期望。

1.2 存储服务质量面临的问题

遗憾的是，现有的分布式块存储系统对服务质量的控制与支持严重不足，特别是在资源需求量超出存储系统能力时，无法提供令人满意的服务质量，迫切需要提出一种满足高效、稳定的存储服务质量要求的解决方案。

存储系统通过块设备接口向应用程序提供服务，如读、写操作等。本文将“通过通用接口显性地向上层应用程序提供的存储服务”称为前端任务，前端任务被视作存储系统的基本服务功能。相较于基本服务功能，存储系统也会提供一些高级服务功能，如负载均衡、缓存脏数据回刷、重复数据删除、数据压缩、定期快照、数据副本同步、副本与纠删码相互转换等。此类高级服务功能对于上层应用程序透明，存储系统不提供显性的调用接口。本文将“不通过通用接口向上层应用程序提供的存储服务”称为后端任务。

对于分布式块存储系统而言，其服务质量应从如下两个维度进行支持。

(1) 前后端任务的资源平衡问题

不论是前端任务还是后端任务，都需要一定的系统资源才能运行，如内存、CPU、I/O、网络带宽等。由于系统资源是有限的，如果不对任务的执行加以控制，会造成前后端任务之间对资源的无序竞争。显然，前端业务的时延对于用户的服务质量体验产生决定性的影响，而后端业务虽然是不可或缺的，但其须为前端业务让路，优先满足前端业务所需的资源。因此，对前端任务和后端任务的执行进行合理的调度，是解决该问题的关键。

(2) 前端任务间的资源平衡问题

前端任务既然被调度执行，那么其相互的资源竞争必须以一种公平的方式对待，特别是对性能影响显著的、有限的缓存资源。在实际的应用



任务中,数据访问热度符合 Pareto (帕拉托) 定律,即少数热数据被频繁访问,因此将热数据存放在缓存中。当产生新的热数据时,需要缓存替换。而现有的缓存替换算法太过陈旧,不符合云环境的新期望。因此,对前端任务之间热数据的精确识别,将真正热数据放在缓存中,及时替换变冷的数据,是解决该问题的关键。

只有解决好这两个维度上的资源竞争问题,分别从预防和妥协两个方面采取措施,才能真正提供高效、稳定的存储服务质量,这已经成为学术界和工业界的共识。

本文设计并实现了中兴分布式块存储系统 (ZTE distributed block storage system, ZTE-DBSS),同时,重点提出了两个方面的优化,从而满足高效、稳定的服务质量的迫切需求。

- 面对前后端任务的资源平衡问题,提出一种自适应的任务调度技术。
- 面对前端任务间的资源平衡问题,提出一种多级热度的缓存替换技术。

最后将上述优化技术应用在 ZTE-DBSS 中,并在真实的云计算应用场景中进行测试。

2 相关研究

2.1 分布式块存储系统

Sheepdog 提供了一个高可用的块级存储卷,提供了对 QEMU 虚拟机的接口、对 iSCSI 协议的支持、基于 FUSE 的文件系统挂载的 SheepFS 以及 SBD (sheepdog block device) 多种块设备接口^[7]。Sheepdog 默认采用多副本模式,另外还提供了日志模式和纠删码存储模式。

Weil 等人^[8]提出了 Ceph RBD 分布式块存储系统,在开源社区中具有广泛的影响力。2010 年 3 月 Linus Torvalds 将 Ceph 合并到了 Linux 内核 2.6.34。Ceph 具有优良的系统架构设计,保证了优异的可靠性和可扩展性。Ceph 主架构由客户端与 RADOS 组成,客户端实现了多种类型的接入,

提供了非常丰富的功能。

Na 等人^[9]提出了 BW-RAID 分布式块存储系统,它是一种前端并行数据传输和后端分布式冗余管理的网络 RAID 存储系统,有着良好的空间利用率和较小的冗余计算开销,同时有效解决了集中式管理方式性能瓶颈的问题。BW-RAID 是国产商业性分布式块存储系统的典范。

Amazon EBS^[10]是 AWS 中著名的商业性弹性块存储系统。对用户来说可以将 EBS 看作一款超大的块设备。因此 EBS 能够提供最底层的服务,可以在这之上建立文件系统、数据库、共享磁盘甚至操作系统等。EBS 的整体架构由分布式系统与集群控制器组成,每个集群控制器可以管理多台物理节点。

以上参考文献提出的分布式块存储系统为开源系统或商业系统,开源系统虽然成本低但服务质量支持不足;商业系统虽然服务质量有一定程度的支持,但效果不佳、成本高昂。

2.2 任务调度技术

张泉^[6]提出了一种支持多维服务质量保证的 I/O 调度算法 Courier,根据应用的带宽需求分配基本的信誉额度,并允许应用透支限定额度的信誉以优先处理紧急请求,同时,奖励负载特征良好的应用以更多的信誉额度,以提高存储系统的整体资源利用效率。

冯盼盼等人^[11]提出一种分布式块存储系统及其存储质量控制方法及装置,该系统利用指标监控服务器,对每个任务使用的各个存储单元及其所在的各个存储节点服务器进行指标监控,对该任务使用的所述各个存储单元进行存储质量评分,根据评分进行存储单元自动切换,从而保证存储服务质量。

Huang^[12]提出了 Stonehenge 算法,根据客户端的需求采用公平的队列算法将请求调度到合适的虚拟磁盘上。它通过存储控制器控制访问共享存储设备上的多个应用数据流,并通过实时调度

和基于反馈控制的设备队列来保证应用所指定的服务质量目标。

开源 Ceph v0.94 代码^[13]采用静态配置文件的方案对前后端任务进行调度控制。可配置如下参数: backfill 最大的操作数、最小对象数、最大对象数、响应请求的最高 OSD Daemon 使用率、recovery 请求活跃个数上限、线程个数上限、优先级等。该种静态配置方式的缺点是: 配置参数复杂、缺乏灵活性、效果较差。系统繁忙时后端任务会阻塞 I/O 请求, 系统空闲时后端任务无法充分利用系统的磁盘、网络带宽等资源。

以上参考文献提出的任务调度技术本质上是资源隔离, 与云存储共享、弹性的理念相违背, 存在资源利用率低下的问题。

2.3 缓存替换技术

Lee 等人^[14]结合最少使用 (least recently used, LRU) 和最不经常使用 (least frequently used, LFU) 页置换算法提出了 LRFU 算法。当替换发生时, LRFU 替换具有最小 CRF 的数据块。CRF 表示数据块会在不久的将来被访问到的概率大小。LRFU 既可作为 LRU, 也可作为 LFU, 因此 LRFU 在时效型和频率型的数据访问模式下都能取得比较好的效果。

Jiang 等人^[15]提出 LIRS 算法, 通过使用两次访问同一数据块之间的距离动态地对访问数据块排序, 从而进行数据块替换算法。LIRS 将每个存储层划分为两部分: LIR 空间和 HIR 空间。该算法能够保证在 LIR 空间里的所有数据块都比 HIR 空间里的数据块具有更小的访问间隔。因此, HIR 空间中的数据块将会被优先考虑替换。

Megiddo 等人^[16]提出了 ARC 算法, 将每个存储层划分为两部分, 分别存储只访问过一次的数据块和访问次数超过一次的数据块。ARC 可以根据具体的数据访问工作流自动调整这两部分空间的大小来提升命中率。

顾荣^[17]提出一种通用化分层存储系统缓存调度框架, 将参考文献[14-16]中的不同的数据访问模式和缓存策略以及上层的大数据应用融合起来, 并为用户提供一组覆盖多种不同数据访问模式的高效缓存调度策略, 从而加速上层大数据应用的数据读写访问性能。

以上参考文献提出的缓存替换技术均是对 LRU、LFU 等经典算法的局部改进或重新组合。

2.4 分布式块存储系统的优化设计方案

本文设计并实现了中兴分布式块存储系统 ZTE-DBSS, 其采用无中心的分布式架构^[18], 如图 1 所示。其中, MDS 主要用于集群配置和元数据管理; ZK 用于实现分布式数据的强一致更新; CSD (Chunk 存储服务) 主要负责实际的数据存储; CLI 实现内核态客户端, 提供块存储的访问接口; OMM 为操作维护平台, 用于集群管理。

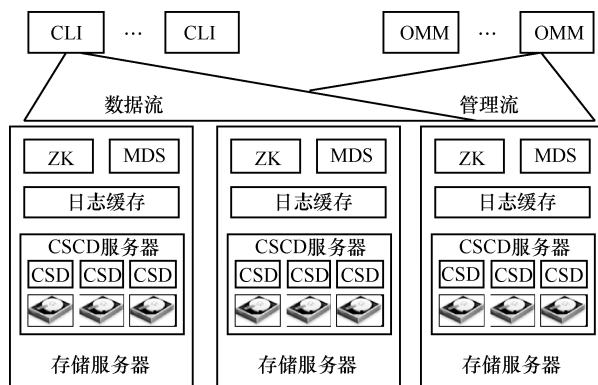


图1 ZTE-DBSS 系统整体架构

ZTE-DBSS 在线性地址空间内可以任意创建虚拟卷, 虚拟卷切分成固定大小的数据片分布在不同的物理存储节点上。每个数据片均有 3 份副本分别保存在不同的存储设备中, 提供了高可用性。通过虚拟块设备驱动, 将用户数据切分为小粒度数据分片, 并通过路由算法确定当前主节点, 发送到对应的服务端进行处理。根据集群中主机状态、容量、负载等情况, 进行路由算法的动态调整, 从而满足可用性、扩展性、一致性要求。本系统支持异构的存储设备, 能够自动根据不同设



备的负载能力,按照一定的比例关系使用每台存储设备的存储资源,实现负载均衡分布、提升设备的利用率。如图2所示,ZTE-DBSS分为两级寻址。

- 首次寻址:根据一致性散列计算本次读写请求归属的CG(Chunk group),根据CG的路由定位请求应该发到那个节点的哪个CSD。
- 二次寻址:请求磁盘后,调用裸盘K/V引擎对物理数据块进行读写,利用裸盘K/V的内部B+元数据管理完成二次寻址。

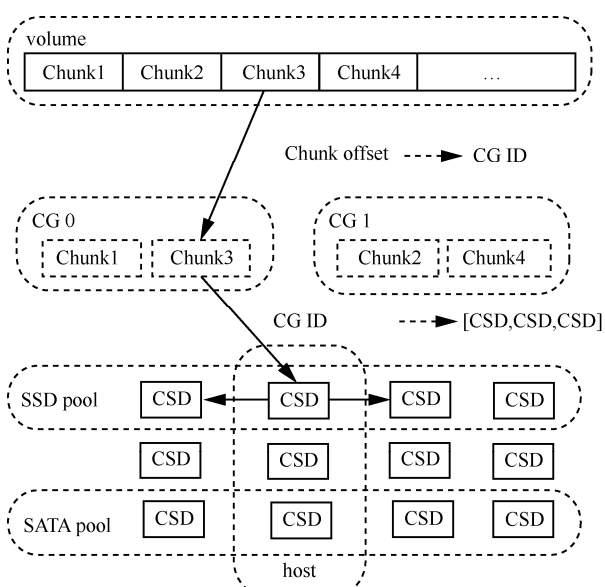


图2 ZTE-DBSS 两级寻址

为提升 I/O 利用率和系统吞吐量,如图3所示,ZTE-DBSS 采用定制化的裸盘文件系统替代常规文件系统。直接管理物理磁盘的空间分配,采用段/区/页三级组织形式,避免了文件系统常见的碎片化现象。不同 CG 间共享存储空间,动态分配,实现块设备精简配置特性。同一块硬盘支持不同 Chunk 大小卷共享存储,每个 CG 创建的时候可以指定其尺寸,适应不同任务类型,在同一存储池中满足不同场景的性能要求。

中兴分布式块存储系统 ZTE-DBSS 目前已在多个公有云、私有云场景下支撑了多样的任务应用,如服务器虚拟化、数据库、气象预报大数据、

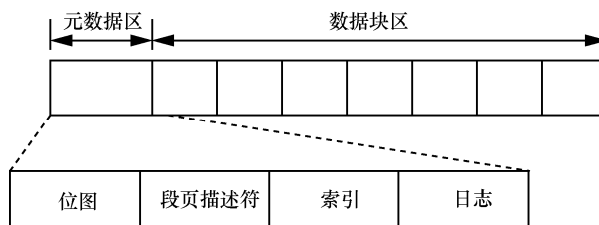


图3 ZTE-DBSS 裸盘文件系统

地质勘探大数据、云盘、远程办公、云桌面、VR 视频等。在系统上线之初存储服务质量支持较弱,在前后端任务的资源平衡问题上采取的技术方案是类似于考参考文献[10]的静态限速技术,在前端任务间的资源平衡问题上采取的技术方案是类似于参考文献[13]的二级链表技术,存储服务质量难以符合用户的预期,有较大的提升空间。

4 自适应的任务调度技术

如前文所述,现有的任务调度技术^[3,8-10]往往采用静态限速的方法,对后端任务设置资源上限,如 IOPS (input/output operations per second, 每秒进行读写操作的次数) 限定在某阈值以下。该方法虽较好地保证了前端任务执行时所需要的资源,但是在系统空闲时却无法充分利用资源,造成了资源的浪费,以及后端任务执行效率的低下。

一种合理的解决办法是,分布式块存储系统需要优先满足前端任务的运行,尽量在系统的空闲时间段执行后端任务,并根据前端应用负载自适应地调整后端任务的速率,由此保证存储服务质量。

针对现有静态限速技术中存在的灵活性低下、系统空闲时资源无法有效利用的问题,本文提供一种通过跟踪系统资源负载的轻重程度,自适应地执行后端任务并限速的技术,避免对前端任务性能的影响,从而保证存储服务质量。

本技术包括两个方面:首先通过基于逻辑阀的任务开关机制来决定后端任务是否可以执行、哪些后端任务可以执行;然后通过基于令牌桶的速率控制机制对准许执行的后端任务进

行速率控制。

(1) 基于逻辑阀的任务开关机制

在后端任务执行控制方面, 采用一个逻辑阀来调控。是否在下一时间周期 P_{i+1} 内执行后端任务, 依赖于当前时间周期 P_i 内负载 WL_i 的轻重。周期 P 的值不能设置得太长。

对于某一具体的后端任务而言: 当前时间周期 P_i 内负载 WL_i 较轻时, 下一时间周期 P_{i+1} 执行后端任务; 反之, 当前周期 P_i 内负载 WL_i 较重时, 下一周期 P_{i+1} 不执行后端任务。

对于全体后端任务而言: 将后端任务的优先级进行预设, 负载阈值见表 1。根据负载的轻重程度, 可以调控不同优先级的后端任务的执行情况。在一些简化场景中, 后台业务无需区分优先级, 可仅设置 2 个繁忙等级, 即空闲和繁忙。

表 1 负载阈值

繁忙等级	负载范围下界	负载范围上界	后台业务可否执行
空闲	0	第 1 负载阈值	全部后台业务
1 级繁忙	第 1 负载阈值	第 2 负载阈值	第 1 至第 $N-1$ 优先级后台业务
2 级繁忙	第 2 负载阈值	第 3 负载阈值	第 1 至第 $N-2$ 优先级后台业务
...	...	...	...
$N-1$ 级繁忙	第 $N-1$ 负载阈值	第 N 负载阈值	第 1 优先级后台业务
N 级繁忙	第 N 负载阈值	1	否

本文主要用“时间周期 P_i 内, 破坏时延上限的用户 I/O 所占比率 VR_i ”作为负载 WL_i 的量化方式。令 N_i 为时间周期 P_i 内完成的用户 I/O 请求个数, $L_i(k)$ 是在周期 P_i 内的第 k 个用户 I/O 的服务时间, 周期 P_i 的时延上限为 LC_i 。若 $L_i(k) > LC_i$, 则认为该用户 I/O 破坏了时延上限。

$$VR_i = \frac{|\{k : L_i(k) > LC_i, k = 1, \dots, N_i\}|}{N_i} \quad (1)$$

如果在周期 P_i 内, $VR_i > \delta$ 则判断负载 WL_i 为高负载, 反之, 则系统负载较轻。这里, δ 称为松弛因子。

(2) 基于令牌桶的速率控制机制

基于逻辑阀的任务开关机制虽然对后台任务的执行进行了初步的把控, 但是如果不对这些准许执行的后台任务进行限速的话, 则仍旧会对前台任务造成严重的资源竞争。因此, 提出一种基于令牌桶机制的速率控制机制。

通过跟踪系统中的实时负载, 进而动态调整后端任务的资源, 控制其执行速率。采用令牌(token)量化存储硬件的能力, 并通过控制令牌消耗的方式给不同的前后端任务分配不同的令牌, 即对后端任务限速。在前台任务繁忙、对资源需求量大时, 减少后端任务令牌数量, 减轻其对前端任务性能的影响; 在前台任务空闲、对资源需求量小时, 增加后端任务令牌数量, 使得后端任务能够尽快完成。

如图 4 所示, 在 ZTE-DBSS 中进行实现时, 在 CSD 层面进行重分布速率控制, 并根据硬件设备单盘 IOPS 上限设置令牌的生成速率, 前端任务请求和后端任务请求共同消耗令牌, 这样就能根据前端任务的 I/O 速率, 自适应地调整后端任务的 I/O 速率。为了让后端任务不与前端任务争夺令牌, 只有在生成的令牌数量超过 $\min_tokens$ 阈值时, 才能执行后端任务, 否则只能由前端任务来消耗令牌。

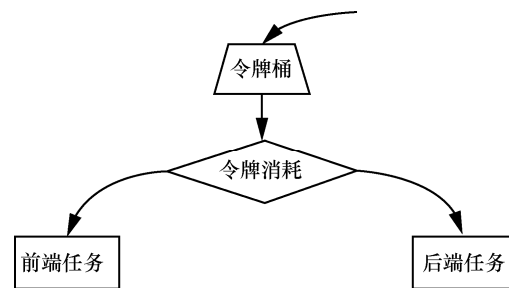


图 4 令牌桶限速

与现有技术相比, 本文提出的自适应的任务调度技术在分布式块存储系统执行后端任务时, 会根据系统的前端任务负载轻重来动态调整后端任务开关和速率, 能够实现在保证前端任务执行



的前提下, 后端任务亦能够快速完成。

5 多级热度的缓存替换技术

如前文所述, 现有的缓存替换技术^[11-14]将产生于 20 世纪的经典缓存替换算法 LRU、LFU 进行改进和重组, 主要问题包括以下方面。

(1) 热度判断的主体局限于进入缓存的数据

现有缓存替换算法不仅缺乏对被换出数据的长期跟踪, 而且缺乏对全体数据的热度判断。随着新型存储器件层出不穷, 存储金字塔结构更加复杂, 涉及多个存储层级之间的冷热数据的判断与迁移, 需要对全体数据进行热度统计与计算。

(2) 热度判断准确率低下

现有缓存替换算法没有同时考虑访问时间和历史访问频率, 导致一些只是近期没有被访问到, 但实际上访问频率很高的数据被替换出去, 热度判断的合理性存在较大提升空间。同时, 被换出数据再次被换入时其热度等同于首次进入缓存的数据, 但其热度是不同的。

(3) 未能充分发挥硬件性能

受制于经典缓存替换算法的惯性思维以及 20 世纪 80 年代的 CPU 计算能力和内存容量限制, 现有的缓存替换算法仍然十分节约 CPU 和内存, 以链表作为最主要的缓存管理方法, 处理效率低下。

一种合理的解决方案是, 应该综合考虑数据近期的访问时间和历史的访问频率, 避免将只是近期未被访问, 但实际上历史访问频率很高的数据换出。因此, 不能将最近的数据元素和已过去一定时间的数据元素截然分开, 不能只对最近的访问情况进行考察, 应对全体数据进行考察, 采用可量化的计算方式对数据的热度进行评价。

在本技术中, 对每一个数据元素的访问视作一个数据流, 将访问记录按时间来划分为若干段, 数据流中的每个数据即各个时间段内, 该

对象被访问的次数, 将其表示为 $\text{Visit}[a][t]$, 其中 $\text{Visit}[a][0]$ 是数据 a 距离当前最远的一条访问记录, $\text{Visit}[a][t-1]$ 为当前访问记录。令 k 为热度衰减常数, 取值范围为 $k \in [0, 1]$, 典型值如 10^{-3} 、 10^{-2} 。

数据 a 在时间片 t 之后的热度评分计算式为:

$$\text{Visit}[a][t] = \text{Score}[a][t-1] * k + \text{Visit}[a][t] \quad (2)$$

其中, k 值越大, 历史访问对当前热度评分的影响越显著, 反之则越不显著。数据热度的评分越高, 则表示该对象被访问得越频繁。数据流中的元素权重取决于其离当前时间的远近, 流中越早的元素权重越小。与此相比, 在 LRU 算法中, 对一个大小为 n 的固定大小的窗口内的元素进行加权求和时, 会对最近 n 个元素都赋予权重 1, 而对所有更早的元素赋予权重 0。两者区别如图 5 所示。

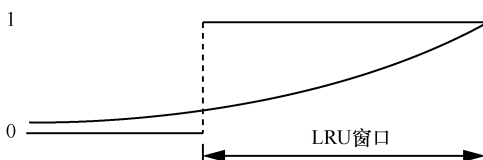


图 5 多级热度的缓存替换技术与 LRU 对比

在多级缓存的分布式块存储环境中, 热度阈值见表 2, 热度分级与存储系统的缓存层数有直接关系。热度阈值与存储系统中各缓存层的容量有直接关系, 可按照经验值设置。

表 2 热度阈值

热度级别	热度范围下界	热度范围上界	对应存储器件
第 1 热度	第 1 热度阈值	$+\infty$	第 1 存储器件
第 2 热度	第 2 热度阈值	第 1 热度阈值	第 2 存储器件
...	...	...	...
第 N 热度	0	第 $N-1$ 热度阈值	第 N 存储器件

这样一来, 就可以将数据元素的当前热度与热度阈值进行比对, 确定数据元素所属的热度级别, 若新旧级别不相同则将数据元素迁移到对应的存储器件中。

若出现某热度的存储器件容量在警戒阈值之

上,则调整热度阈值。警戒阈值为可设经验值,典型值为 95%,目的在于在容量满之前及时调整阈值,提前迁移数据,避免出现因容量已满而导致新换入数据失败的情况。缓存容量达到警戒值说明热度阈值表设置不当,需对其进行调整。调整的方法包括:将该热度阈值的范围下界调低或上界调高。

与现有技术相比,本项技术在分布式块存储系统执行前端任务时,热度判断的主体全面、热度判断准确率高、充分发挥了硬件性能,从而以一种公平的方式对待前端任务对缓存资源的竞争。

6 实验与测试

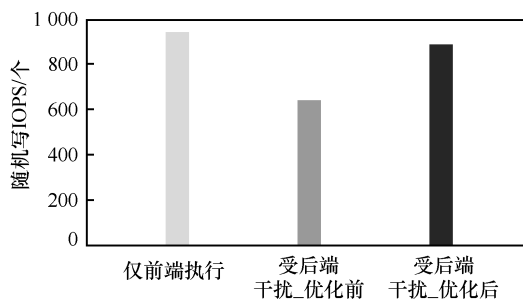
6.1 测试方案

在本节中,不但针对优化前后的任务调度技术和缓存替换技术分别进行了对比测试,而且在实际的分布式业务场景中对整体服务质量进行了实际测试和对比,包括存储服务效率和稳定性。为保证实验结果可靠,每组实验重复 10 次,去掉最大和最小的 2 个值后取平均值。

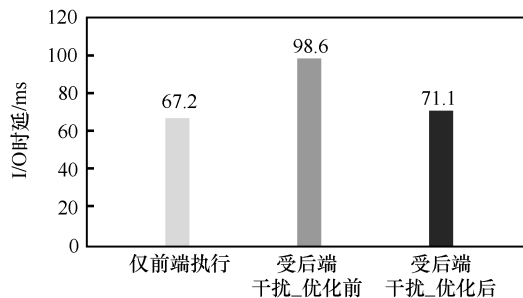
6.2 自适应的任务调度测试结果

在本测试中:前端业务请求为 4 KB 随机写,I/O 深度 64;后端业务请求为数据重分布,迁移量约 22 GB。优化前采用静态限速的技术,优化后采用本文自适应的任务调度技术。

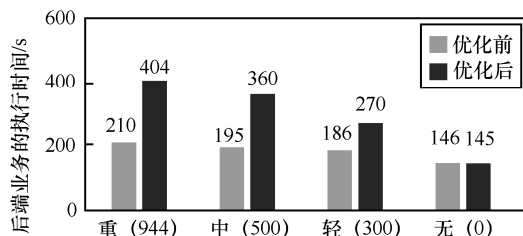
如图 6(a)及图 6(b)所示,在优化前,当后端任务来临时,前端任务的 IOPS 和时延受到显著的干扰;在优化后,后端业务尽量少地参与资源竞争,前端任务的 IOPS 和时延与无后端任务的执行结果十分接近。与优化前相比,本文的调度技术在 IOPS 上提升 38%,在时延上降低 39%,达到了优先满足前端任务的目标。如图 6(c)所示,与优化前相比,优化后的技术在前端业务繁忙时,后端业务的执行时间优化后显著拉长,避免参与资源竞争。在前端业务空闲时,后端业务能做到全速完成。达到了优先满足前端任务的目标。



(a) 对前端业务 IOPS 的影响



(b) 对前端业务 I/O 时延的影响



(c) 对后端业务执行时间的影响

图 6 任务调度测试结果

6.3 多级热度的缓存替换测试结果

为了体现缓存的效果,采用随机读写来测试,主要由两个指标来衡量:4 KB 粒度的随机读写下的 IOPS 以及 1 MB 粒度的随机读写下的吞吐量。另测试了真实工作负载产生的 I/O trace 情况下的系统 I/O 性能。优化前采用 ARC 算法,优化后采用本文提出的缓存替换算法,SSD 缓存大小为 40 GB。

随机负载时的测试结果如图 7 所示,4 KB 随机读写的 IOPS 分别提升 68%、69%,播放真实 trace 时测试结果见表 8,4 种 trace 提升 11%~59%,平均提升 35%。

6.4 存储服务质量的整体测试结果

采用商用云计算的真实业务场景——国内某通信运营商西南区数据中心进行测试。该测试案例包括 3 个前端应用,分别为视频播放、云桌面

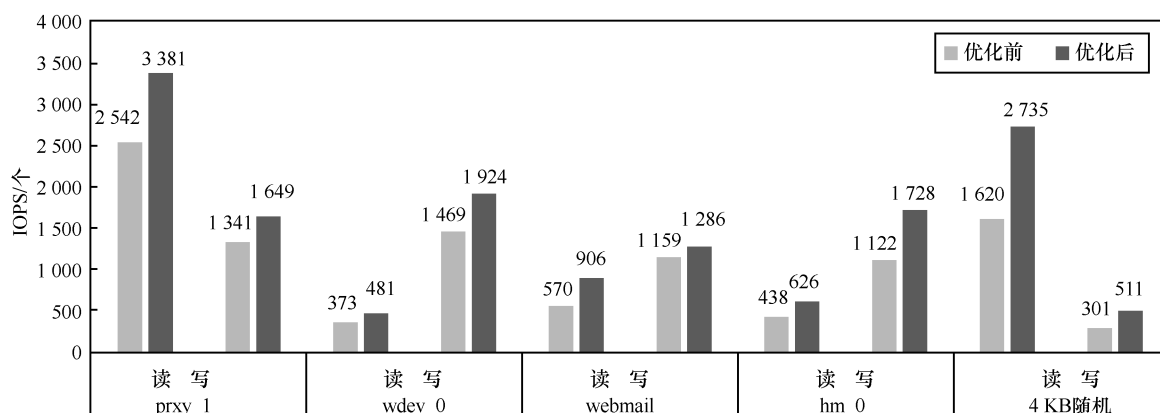


图7 缓存替换测试结果

办公、大型游戏，2 个后端任务，分别为热点数据迁移、数据库负载均衡。在测试案例进行到 22 s 时触发后端业务，从而体现突发性的资源竞争对于前端业务的稳定性影响。在运行相同测试程序时，对比优化前和优化后的 ZTE-DBSS 在存储服务质量上的表现。存储服务效率的评价指标为：在前后端业务同时执行期间，前端业务的持续稳定的 IOPS 平均值。存储服务稳定性的评价指标为：突发性的资源竞争对于前端业务 IOPS 的影响程度。

如图 8 所示，优化前，尽管各前端业务对性能的需求不同，但前端业务受到相同的限速，对性能需求较小的 3 号业务与性能需求较大的 1 号业务的 IOPS 接近，这是十分不合理的；在后端业

务对资源竞争时，前端业务的 I/O 波动较大，且前后端受到相同的限速，本应为前端业务让路的后端业务与各前端任务无差别地共享资源，这同样是不合理的。优化后，虽然前后端业务性能总需求超过了存储资源的总体 I/O 服务能力，但所有的应用都按照一定比例进行分享，该种结果对于各业务而言是公平的；在后端业务对资源竞争时，前端业务的 IOPS 波动较小，后端业务尽量少干扰。优化前后的对比结论：从服务效率来看，在后端业务竞争资源时，优化前的前端业务平均 IOPS 仅为 400，而优化后达到平均 710，因此前端业务服务效率平均提升 77.81%；从服务稳定性来看，在后端业务竞争资源时，优化前的前端业务 IOPS 下降率平均高达 63.31%，而优化后仅平

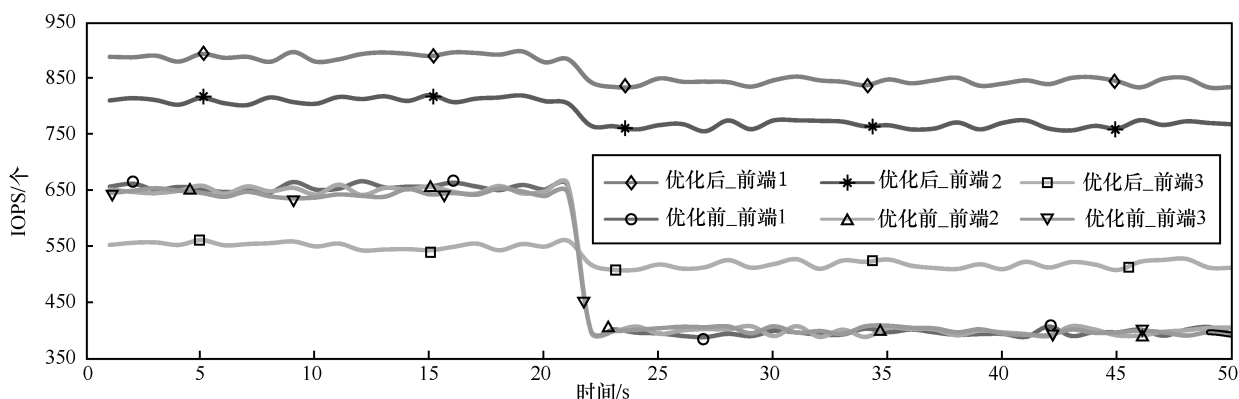


图8 存储服务质量的整体测试结果——优化前后各业务 IOPS

均下降 5.98%，因此前端业务稳定性平均提升 57.33%。

如图 9 所示，在相同的测试环境、测试用例下，本文测试了 cephv10.2.z 在服务质量上的表现。可以看到，Ceph 与优化前的 ZTE-DBSS 表现相近，前端业务性能不佳，且在受到后端业务干扰时前端业务性能平均下降达到 56%。优化后的 ZTE-DBSS 的性能明显高于 Ceph，受到前端业务干扰明显小于 Ceph。

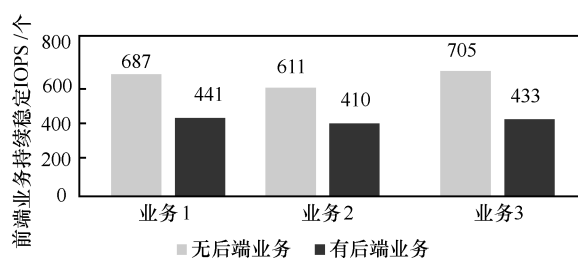


图9 与 Ceph 对比存储服务质量

7 结束语

本文针对现有的分布式块存储系统对服务质量的控制与支持不足的缺陷，提出一种自适应的任务调度技术，解决了前后端任务的资源平衡问题，与静态限速技术相比性能提升 38%。提出一种多级热度的缓存替换技术，解决了前端任务间的资源平衡问题，与 ARC 算法相比性能提升 35%。并在中兴分布式块存储系统中进行了工程实现和商用测试。测试结果表明，在典型的云计算应用场景中，相较于优化前，存储服务效率平均提升 77.81%，存储服务稳定性平均提升 57.33%。ZTE-DBSS 满足了用户对分布式块存储服务质量的期望，满足了对云计算产品的核心期望。

本文提出的任务调度技术和缓存替换技术均是在系统正常情况下的服务质量控制手段，然而，在异常情况下的服务质量是下一阶段重点关注的内容，如坏盘预测、新型纠删码等内容。坏盘预测是一种运用人工智能的技术手段，对存储器件

的健康状况进行预测，从而能够在系统空闲时提前迁移数据，避免突发性的异常坏盘导致的服务质量降级。纠删码是一种数据可靠性保障技术，在数据故障发生后，通过一定的算法能够将丢失的数据找回，但是会引起大量的计算开销，严重影响存储服务质量，迫切地需要提出新型纠删码算法，在保障可靠性的前提下降低计算量。

参考文献:

- [1] 孙健, 贾晓菁. Google 云计算平台的技术架构及对其成本的影响研究[J]. 电信科学, 2010, 26(1): 38-44.
SUN J, JIA X J. A study of the influence of technical architecture on the total cost of google cloud computing platform[J]. Telecommunications Science, 2010, 26(1): 38-44.
- [2] 李丹, 刘方明, 郭得科, 等. 软件定义的云数据中心网络基础理论与关键技术[J]. 电信科学, 2014, 30(6): 48-59.
LI D, LIU F M, GUO D K, et al. Fundamental theory and key technology of software defined cloud data center network [J]. Telecommunications Science, 2014, 30(6): 48-59.
- [3] 樊自甫, 伍春玲, 王金红. 基于 SDN 架构的数据中心网络路由算法需求分析[J]. 电信科学, 2015, 31(2): 42-51.
FAN Z F, WU C L, WANG J H. Requirements analysis of data center network routing algorithm based on SDN architecture [J]. Telecommunications Science, 2015, 31(2): 42-51.
- [4] 孙振正, 龚靖, 段勇, 等. 面向下一代数据中心的软件定义存储技术研究[J]. 电信科学, 2014, 30(1): 39-43.
SUN Z Z, GONG J, DUAN Y, et al. Research on software defined storage technology for the next-generation data center[J]. Telecommunications Science, 2014, 30(1): 39-43.
- [5] 杨洪章, 罗圣美, 施景超, 等. 面向移动通信大数据的云存储系统优化[J]. 计算机应用, 2017, 37(a01): 27-33.
YANG H Z, LUO S M, SHI J C, et al. Cloud storage system optimizations for telecom big data[J]. Journal of Computer Applications, 2017, 37(a01): 27-33.
- [6] 张泉. 面向云数据中心的存储服务质量技术研究[D]. 武汉: 华中科技大学, 2014.
ZHANG Q. Research on storage quality of service for cloud data centers[D]. Wuhan: Huazhong University of Science and Technology, 2014.
- [7] 贾博文, 张文军, 李小勇. 面向虚拟机的分布式块存储系统设计及实现[J]. 微型电脑应用, 2015(3): 32-37.
JIA B W, ZHANG W J, LI X Y. Design and implementation of a distributed block storage system for virtual machines in the IaaS cloud[J]. Microcomputer Applications, 2015(3): 32-37.
- [8] WEIL S A, BRANDT S A, MILLER E L, et al. Ceph: a scalable, high-performance distributed file system[C]//The 7th symposium on Operating systems design and implementation, No-



- vember 6-8, 2006, Seattle, Washington. [S.l.]: USENIX Association, 2006: 307-320.
- [9] NA W W, MENG X X, SI C, et al. A novel network RAID architecture with out-of-band virtualization and redundant management[C]//The 14th IEEE International Conference on Parallel and Distributed Systems, December 8-10, 2008, Melbourne, Victoria, Australia. [S.l.:s.n.], 2008: 105-112.
- [10] DIKAIKAKOS M D, KATSAROS D, MEHRA P, et al. Cloud Computing: distributed internet computing for IT and scientific research[J]. IEEE Internet Computing, 2009, 13(5):10-13.
- [11] 冯盼盼, 卢明燮. 一种分布式块存储系统及其存储质量控制方法及装置: CN201510544239.X [P]. 2015-08-28.
FENG P P, LU M F. A distributed block storage system and its storage quality control method and device: CN201510544239.X[P]. 2015-08-28.
- [12] HUANG L. Stonehenge: a high-performance virtualized IP storage cluster with QoS guarantees[M]. New York: ACM Press, 2003.
- [13] 刘飞, 蒋德钧, 张欢, 等. 异构存储感知的 Ceph 存储系统数据放置方法[J]. 计算机科学, 2017, 44(6): 17-22.
LIU F, JIANG D J, ZHANG H, et al. Heterogeneous storage aware data placement of ceph storage system computer science[J]. Computer Science, 2017, 44(6):17-22.
- [14] LEE D, CHOI J, KIM J H, et al. LRFU: a spectrum of policies that subsumes the least recently used and least frequently used policies[J]. IEEE Transactions on Computers, 2001, 50(12): 1352-1361.
- [15] JIANG S, ZHANG X. LIRS: an efficient low inter-reference recency set replacement policy to improve buffer cache performance[J]. ACM SIGMETRICS Performance Evaluation Review, 2002, 30(1): 31-42.
- [16] MEGIDDO N, MODHA D S. ARC: a self-tuning, low overhead replacement cache[C]// The 2003 USENIX Conference on File and Storage Technologies, March 31 - April 2, 2003, San Francisco, California, USA. [S.l.]: USENIX Association, 2003: 9.
- [17] 顾荣. 大数据分布式存储技术与系统研究[D]. 南京: 南京大学, 2016.
GU R. Research on techniques and systems for big data processing[D]. Nanjing: Nanjing University, 2016.
- [18] 屠要峰. 高性能、高可靠中兴超融合架构关键技术解析[J]. 通信世界, 2017(9): 55-56
TU Y F. Key technical analysis of high performance and high reliability ZTE hyper-convergence architecture[J]. Communications World, 2017(9): 55-56

[作者简介]



屠要峰 (1972-), 男, 中兴通讯股份有限公司研究员级高级工程师, 中国计算机学会会员, 主要研究方向为云存储。



杨洪章 (1989-), 男, 中兴通讯股份有限公司高级工程师, 中国计算机学会信息存储专委会委员, 主要研究方向为文件系统。



韩银俊 (1977-), 男, 中兴通讯股份有限公司工程师, 主要研究方向为云计算平台。



陈正华 (1983-), 男, 中兴通讯股份有限公司工程师, 主要研究方向为超融合存储系统。