文章编号: 1009 - 2552(2015)08 - 0100 - 04 DOI: 10.13274/j.cnki.hdzj.2015.08.026

基于 Ceph 的云存储容错机制研究与实现

漆晓芳，倪 明

(中国电子科技集团第三十二研究所，上海 200030)

摘 要: Ceph 是当前热门的开源云存储项目，具有良好的可扩展性，可轻松扩展至 PB 级，提供对象存储，块存储，文件存储三种存储服务，研究了 Ceph 对象存储中数据容错机制，对副本容错和纠删码容错机制进行对比分析，并提出一种基于冷热数据分层的云存储容错机制。实验表明该机制可以提高存储空间利用率和存储可用性。

关键词: 云存储; Ceph; 容错; 纠删码; 副本复制

中图分类号: TP302.8 **文献标识码:** A

Research and implementation of fault-tolerant mechanism in ceph storage

QI Xiao-fang, NI Ming

(No. 32 Research Institute of China Electronic Technology Group Corporation, Shanghai 200030, China)

Abstract: Ceph is a popular open source storage project, it offers services including object storage, block storage and file storage. This paper describes the fault-tolerant mechanism in Ceph storage, analyzes and compares replication mechanism and erasure-code mechanism, then it present a mechanism based on data stratification. The test results indicate that this mechanism can not only improve the space availability but also improve the reliability of cloud storage.

Key words: cloud storage; Ceph; fault-tolerant; erasure-code; replication

0 引言

随着信息科技的快速发展，数据量呈现爆炸式增长趋势，数据存储需求不断扩大，存储即服务也成为云计算中的一种服务需求，而且数据的价值往往高于硬件设备价值，因此为保证数据可用性，可靠性，分布式云存储需要具备高性能，高可靠等特性。

云存储系统庞大，因断电、人为误操作等因素，节点失效频繁发生，Facebook 的 Hadoop 集群中的 3000 个节点，涉及 45PB 数据。这些数据平均每天有 22 个节点失效，有时候一天的失效节点超过 100 个。为保证数据可用性，数据容错机制对于保证云存储可靠性十分关键。目前最广泛使用的是副本冗余策略，如 GPFS，HDFS，GlusterFS，Ceph [1] 等都提供这种副本冗余容错策略。

另一种容错机制是纠删码 [2] 策略，纠删码在通信中应用广泛，近几年被用到云存储中。当前 Google、Facebook、Microsoft、Amazon 等互联网巨头已开始研究纠删码存储技术，并应用在各自的存储系统中，GlusterFS 将在新版本中添加纠删码存储功能，Ceph 最新的 0.8 版本系列添加了纠删码存储功能。可见纠删码在存储中的应用会越来越广泛。众多学者也对云存储中纠删码的应用进行了研究，文献 [3] 研究了云文件系统中纠删码技术的应用。

在实际应用中，云存储系统中的文件的访问频率大不相同，有些文件经常被读取或修改，有些文件则访问频率偏低，但间歇性被读取，另外一些文件则是几乎不被读取，文献 [4] 和 [5] 都提出一种基于访问统计的自适应容错机制，针对冷热两种数据，采用不同的容错机制。与文献 [4] 和 [5] 不同的是，本文

收稿日期: 2014 - 08 - 04

作者简介: 漆晓芳(1991 -)，女，硕士，研究方向为云计算。

主要以 Ceph 为存储平台，针对冷热温三种数据，研究其容错机制，提出一种基于冷热数据分层容错机制，并通过对象存储实验，分析和验证了这种容错机制的可行性和可靠性。

1 Ceph 简述

Ceph 是一种性能优越，可靠性和可扩展性良好的统一的分布式云存储系统，提供对象存储，块存储，文件存储三种存储服务。Ceph 不是刚刚出现的开源项目，它从最初发布到逐渐流行，经历了 7 年。当前在 Openstack 社区，Ceph 是最受欢迎的开源存储项目，且国内外知名厂商如 Intel、Cisco、Dell、华为等，均采用 Ceph 作为存储后备。

Ceph 的生态系统分为四部分，如图 1 所示。包括监视器集群，对象存储设备集群，元数据服务器集群，客户端。监视器集群由多个监控器组成，负责监控整个存储集群的状态，对象存储集群由多个对象存储设备组成，主要负责存储对象，元数据服务器集群由多个元数据服务器组成，元数据服务器集群只有在文件存储服务中才用得到，对象存储和块存储都无需使用元数据服务器。

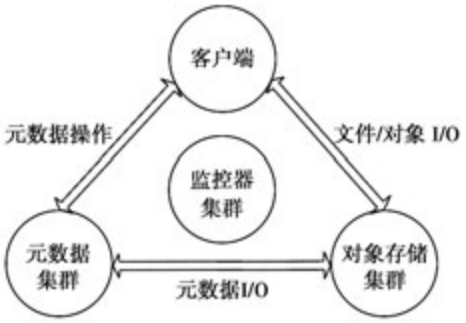

图 1 Ceph 集群组件

2 云存储容错机制简述

2.1 基于副本冗余的容错机制

基于副本冗余的容错机制是将原始数据复制成多份，每一份称为一个副本。将这些副本分别存放在集群中的不同节点上，当集群中有些节点出现故障时，只要其余健康节点中任一个节点拥有副本，用户就可以获取该数据。当前众多存储系统采用副本数为 3 的副本冗余容错机制，这种机制能很好地保证数据可靠性，但也会极大降低存储空间利用率。

2.2 基于纠删码的容错机制

纠删码的基本思想是，通过编解码实现数据冗余。首先将原数据文件分成 k 个大小相等的数据块，然后按照纠删码编码算法进行编码，会得到 n 个带有一定冗余的编码数据，将这些码块存储在不同数据节点上，读取数据时，只需要取其中任意 k 个数据块就能恢复原始数据。纠删码算法众多，经典算法有 RS(Reed – Solomn) 编码[6]、EVENODD 码[7] 和 LDPC(low – density parity – check code) 编码[8] 等，其中 RS 编码是在 Galois 域 GF(2^w) 上进行多项式域运算的编码方式，它是唯一可满足任意数据磁盘数(n) 和冗余磁盘数(m) 的 MDS 编码方式。以 RS(4,2) 为例，是将文件分为 4 个大小相等的数据块，编码之后生成另外 2 个校验码块，从所有的 6 个码块中只要任意取 4 个就可以恢复原始数据。这种方式可容错为 2，即最多允许 2 个码块丢失。通常根据生成矩阵的不同，RS 纠删码分为范德蒙 RS 纠删码和柯西 RS 纠删码。

2.2.1 范德蒙 RS 纠删码

RS 编码是用生成矩阵 G 与数据列向量 x 的乘积来计算信息列向量 y，即 $y = G * x$，其解码过程为生成矩阵的逆矩阵与信息列向量的乘积，即 $x = G^{-1} * y$，为保证解码运算的成功进行，生成矩阵 G 一定要是可逆的。范德蒙矩阵有着良好的特性，在 $GF(2^w)$ 域上，对范德蒙矩阵进行初等变换，将其前 n 行变为单位矩阵，即可保证生成矩阵可逆，基于这个生成矩阵的 RS 纠删码为范德蒙 RS 纠删码。

2.2.2 柯西 RS 纠删码

柯西 RS 纠删码用柯西矩阵代替范德蒙矩阵，得到更为简单的生成矩阵。研究者通过简单改造，使解码过程更为简单。柯西码解码不用求大矩阵的逆，而是把乘法除法运算分别转化为有限域上的加法和减法运算，可用异或实现。因此，柯西 RS 纠删码运算复杂度低于范德蒙 RS 纠删码。

3 Ceph 容错机制

3.1 Ceph 容错机制现状

Ceph 0.8 之前的版本均采用副本冗余策略，用户可以根据需要创建存储池，并设置存储池中数据的副本数目，每个数据副本被分到不同的对象存储设备(OSD) 上，当存储设备中有故障，可以从其他健康的设备上获取数据。

Ceph 0.8 之后的版本添加了纠删码冗余策略，采用开源纠删码库 Jerasure[9]，提供不同的纠删码算法，用户可根据需要选择纠删码算法类型，并创建相应的纠删码池，由于 Ceph 新版本发行时间不久，功能有待完善。

3.2 一种基于冷热数据分层的数据容错机制

存储中数据可根据访问热度分为三种: 热数据，温数据和冷数据。热数据通常需要在高性能、高度可用、高要求的环境下即时存取。温数据处于近线

或在线备份环境中，用户需要快速访问这些数据，但访问的次数较少。冷数据通常访问次数极少，通常用于归档备份。针对云存储中数据访问热度不同，提出一种基于数据热度分层的容错机制。

所有数据先按照副本策略存储，本机制对存入系统的数据，实时统计该数据的被访问频率，设定热数据阈值，温数据阈值，冷数据阈值，高于热数据阈值则判断为热数据，低于冷数据阈值则判定为冷数据，在冷热数据阈值之间的判定为温数据。每 3 个月进行一次数据热度划分，数据被访问频率高于热数据阈值时，判断为热数据，存放在副本池里，该存储池采用副本容错机制。数据访问频率低于冷数据阈值时，判定为冷数据，存放在范德蒙 RS 纠删码池里，该存储池采用范德蒙 RS 纠删码容错机制。对于温数据，存放在柯西纠删码池里，该存储池采用改进的柯西 RS 纠删码容错机制。流程如图 2 所示。

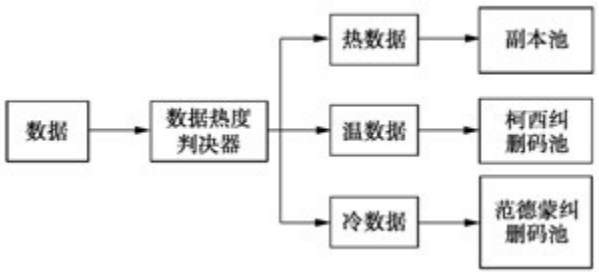

图 2　基于冷热数据分层的数据容错流程

4　实验及分析

4.1　实验环境

首先搭建 Ceph 存储集群，本实验集群由 1 个 Monitor，6 个 OSD 组成，本实验采用 Ceph 的对象存储服务，因此未使用元数据服务器。集群中每个节点基本配置如下：

CPU：Intel © Atom © S1260@2.0GHz

内存：4G

OS：ubuntu 12.04

每个 OSD 容量为 250GB

网络：千兆网

4.2　实验内容

创建三个存储池，分别为副本数为 2 的副本池 Repool，范德蒙 RS(4,2) 纠删码池 Rspool 和柯西 RS(4,2) 纠删码池 Capool。通过数据热度判决选择相应的容错机制，将数据以对象存储的方式放入相应的存储池，进行读写实验。为了更好地管理存储资源，自主开发了一套基于 Ceph 的 Web 管理平台，如图 3 所示，本文的热度判决是依据此 Web 管理平台的资源访问统计日志实现的。

随着各行各业数据量剧增，同样的数据随着时间的推移，访问频率会发生变化，很多数据都会经历从热数据变为温数据，最后成为冷数据的过程，下面实验就是对相同数据不同时间段被划分为不同热度数据时的存储分析。热数据以副本策略进行存储，存放在副本池里，温数据以柯西纠删码策略进行存储，存放在柯西纠删码池，冷数据直接以范德蒙 RS 纠删码策略进行存储，存放在范德蒙纠删码池。

图 3　基于 Ceph 的 Web 管理平台登陆界面

按照 Ceph 对象读写操作指令，编写 shell 脚本，并统计出对象存储读写的操作时间。

4.3　实验分析

4.3.1　时间性能分析

图 4 是执行对象写入操作时，副本池，范德蒙 RS 纠删码池，柯西 RS 纠删码池的时间对比，图 5 是集群所有 OSD 健康时，从三种池对象读取的时间对比，图 6 是集群中有一个 OSD 坏了的情况下，对三种池对象读取的时间对比。

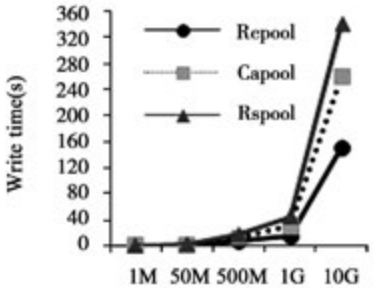

图 4　对象写入时间对比

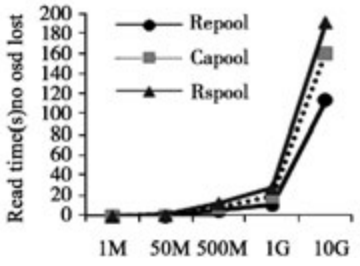

图 5　对象读取时间对比(没有 OSD 损坏)

由图 4－5 可知，副本池的对象读写时间都小于纠删码池对象读写时间，而对比两个纠删码池，采用柯西纠删码策略的存储池读写性能要优于采用范德蒙纠删码策略的存储池。

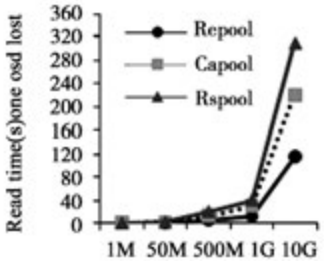

图 6　对象读取时间对比(1 个 OSD 损坏)

由于纠删码涉及到编码解码过程,相比于副本池,有额外的计算成本,对于小文件对象存储,时间差别不是太大,但对于大文件,时间差会很明显,对于数据量庞大可达 PB 级,ZB 级的场景,纠删码池用于编解码的时间代价很大,从而导致性能降低。而对比两个纠删码池,由于改进的柯西纠删码使用异或操作大大节省了矩阵运算乘法的时间,因此柯西纠删码池的读写性能优于范德蒙纠删码池,综上可知柯西纠删码池对象读写性能在副本池和范德蒙纠删码池之间。

集群中单点故障的概率很高,考虑集群中有一个 OSD 损坏时,副本策略和纠删码策略的数据恢复性能,当集群中节点损坏时,副本策略的数据恢复是最快的,只需要取另外的可用副本即可,对于纠删码,则需要获取指定数目的码块进行解码操作才能恢复原始数据,从图 5 -6 也可知,副本池的数据恢复性能优于纠删码池,柯西纠删码池性能优于范德蒙纠删码池。对比也可知,当集群中有一个 OSD 损坏时,副本池的读性能和没有 OSD 损坏时相差不大,但纠删码池的性能则会优于 OSD 损坏,性能极大下降。

热数据通常需要在高性能、高度可用、高要求的环境下即时存取,采用副本冗余策略的副本池有着良好的读写性能及可靠性,因此适合热数据存储。对于温数据,用户需要快速访问这些数据,但访问的次数较少,不需要用太高的性能提供服务,因此可以考虑使用柯西纠删码池进行存储,冷数据通常访问次数极少,因此可以考虑用范德蒙纠删码池进行存储。

4.3.2　空间利用率分析

虽然副本策略比纠删码冗余策略的读写性能更好,但在存储空间利用率上,副本策略并不占优势,实验中采用 RS(4,2) 的纠删码策略,所需存储开销为原文件大小的 1.5 倍,而对于 RS(8,2) 的纠删码池,所需存储开销为原文件大小的 1.25 倍,对于副本数为 2 的副本策略,其存储开销为文件大小的 2 倍,当前众多存储场合中,使用副本数为 3 的副本冗余策略,其存储开销为文件大小的 3 倍,对比可知,副本策略的空间利用率比纠删码冗余策略的低。

随着云计算的发展,存储即服务成为一种趋势,存储成本和可靠性成为众厂商关心的问题,综合分析可知,对于温数据和冷数据,可以采用纠删码容错冗余机制,这样可以极大地节省存储空间,提高空间利用率,由于冷数据访问频率过低,且冷数据的场景并不要求高的读写性能所以可以用范德蒙纠删码容错机制,对于温数据,访问频率虽然也不高,但一旦访问时需要保证读写效率,因此可以考虑柯西纠删码容错机制,在保证空间利用率的同时,兼顾读写性能。对于热数据,可以采用副本容错机制,满足热数据场景的高性能,高可靠性要求。

5　结束语

云存储容错机制对数据可用性和数据可靠性尤为重要,通过合理的容错机制可以保证云存储中数据的高可靠性和可用性,而且可以极大地减少存储空间,提升存储空间利用率,有利于节约存储成本。因此研究云存储纠删码技术有重要的理论意义和实验价值。本文提出的基于 Ceph 的云存储冗余容错机制,根据数据访问热度采取不同的数据容错机制,对于热度数据采用副本策略,对于温数据采用柯西 RS 纠删码策略,对于冷数据,使用范德蒙 RS 纠删码策略。实验表明此策略可以很好地兼顾数据访问和存储空间利用率的提高。

由于 Ceph 中的纠删码功能还不够完善,目前不支持纠删码池中的数据更新操作,在将来的工作中,将进一步研究新型纠删码在 Ceph 中的应用,从而更好地保障云存储的高性能,高可用性及高存储利用率。

参 考 文 献:

[1] Sage A Weil,Scott A Brandt,Ethan L Miller,et al. Ceph: A Scalable, High-Performance Distributed File System [C]. Proceedings of the 7th symposium on Operating systems design and implementation,2006.

[2] 罗象宏,舒继武. 存储系统中的纠删码研究综述 [J]. 计算机研究与发展,2012,49(1): 1 -11.

[3] 程振东,栾钟治,孟由,等. 云文件系统中纠删码技术的研究与实现 [J]. 计算机科学与探索,2013,7(4): 315 -325.

[4] 杨东日,王颖,刘鹏. 一种副本复制和纠错码融合的云存储文件系统容错机制 [J]. 清华大学学报: 自然科学版,2014(1): 137 -144.

[5] 聂瑞华,张科伦,梁军. 一种改进的云存储系统容错机制 [J]. 计算机应用研究,2013,30(12): 3724 -3728.

[6] 万武南. 分布式安全存储系统纠删码技术的研究 [D]. 成都: 中国科学院研究生院(成都计算机应用研究所),2006.

[7] 常乾,许胤龙,项利萍,等. 基于 EVENODD 码的单盘故障快速恢复算法 [J]. 计算机应用与软件,2011,28(6): 15 -18.

[8] 王鹏,王新梅. LDPC 码的快速编码研究 [J]. 西安电子科技大学学报,2005,31(6): 934 -938.

[9] Plank J S, Simmerman S, Schuman C D. Jerasure: A library in C/C ++ facilitating erasure coding for storage applications-Version 1.2 [Z]. University of Tennessee, Tech. Rep. CS-08-627, 2008, 23.

责任编辑: 么丽苹