基于 Ceph 存储系统的冗余存储技术研究

刘媛媛
(通信信息集团有限公司,北京 100191)

摘要:存储系统的可靠性和安全性至关重要,保证存储系统可靠性的方式则是对数据进行冗余存储。广泛应用的数据冗余存储技术主要有两种:副本冗余和纠删码冗余。Ceph 作为一种分布式云存储解决方案越来越流行,文章以 Ceph 为平台,研究与分析了 Ceph 的副本冗余和纠删码冗余机制,分析对比了这两种方式各自的优缺点,并以此为基础提出一种综合二者优点的冗余方法。该方法在节省存储空间和数据访问性能方面达成一个折中,同时也保证了系统的可靠性。

关键词:存储系统;Ceph;副本冗余;纠删码
中图分类号:TP333 文献标识码:A 文章编号:1673-1131(2018)09-0091-02

近年来,企业存储需求爆发性增长,研究表明,许多公司的数据占用空间每年翻一番。对存储系统来说,统一、分布式、可靠、高性能、能够大规模扩展至关重要。Ceph 作为一个开源的、可大规模扩展、高性能且无单点故障的分布式存储系统,它可以运行在通用的商用硬件上,因而正变成一个流行的云存储解决方案。为了保证数据存储的可靠性,Ceph 采用了副本冗余存储及纠删码存储两种方式,也是目前被广泛应用的两种冗余方式。本文基于 Ceph 存储系统,研究了 Ceph 存储中使用的两种冗余存储方式,并在此基础上提出基于文件分类的冗余存储方式。不同类型的文件数据采用不同的冗余方式,既保证了系统的可靠性,又能节省存储空间。

1 Ceph 存储系统

Ceph 的生态系统包括五部分:监视器(monitor)、管理器(manager)、对象存储设备(OSD)、元数据服务器(MDS)、客户端。监控器负责监控整个集群的状态,包括监控器、管理器、OSD 等的状态,同时也负责集群节点与客户端之间的认证相关工作。管理器负责追踪集群的运行时参数及当前状态,包括存储利用率、当前性能指标、系统负载等。对象存储设备负责存储数据,处理数据的复制、恢复、再平衡,并向监视器提供一些监视信息。元数据服务器存储元数据信息,Ceph 存储集群支持对象存储、块存储和文件系统存储,只有在文件系统存储中才需要元数据服务器,而对于块存储和对象存储,则不需要元数据服务器。

2 冗余机制

存储系统的冗余机制一般采用两种:副本冗余存储和纠删码存储。副本冗余存储是把原始数据复制多份进行存储备份,每份为一个副本,把每个副本分别存储在不同的节点,当其中某个节点失效时,其他节点可继续提供服务,保证系统的可靠性。纠删码存储主要是通过某种纠删码算法把要存储的数据进行编码得到冗余数据,然后把原始要存储的数据及冗余数据进行存储,当其中某些数据失效时,可以通过算法恢复或重新生成原始数据。

Ceph 通过池(Pool)来提供简单的存储管理,Ceph 的池是一个用来存储对象的逻辑分区,有复制池和纠删码池,分别对应冗余存储机制的副本存储和纠删码存储。

2.1 副本方式

在 Ceph 中引入了 PG(数据归置组)概念,PG 是一组对象的逻辑集合,主要用途是为了在成千上万的 OSD 上管理和跟踪数以百万计的对象的复制和传播,以及管理这些对象所消耗的计算资源。每个池中有大量的 PG,在 Ceph 系统中只需要管理包含大量对象的 PG。使用 Ceph 存储数据时,系统会把客户端要存储的数据分散到各个数据归置组中。当 Ceph 集群接收到数据存储请求时,会首先把数据分解成一组对象,利用 CRUSH 算法并根据对象 ID、池名称以及池中的 PG 数量执行散列操作,将结果生成 PG ID。每个 PG 中的对象会被复制分发到 Ceph 集群的多个 OSD 上。这些 OSD 称为主 OSD、第二 OSD 和第三 OSD 等,这个包含了一个特定 PG 的副本的 OSD 集合称为 acting 集合。

假设创建池时,设置副本数量 size=3,则池中的每个 PG 中的对象会以 PG 为单位存储到三个 OSD 上。当客户端向集群写入数据时,它会访问主 OSD,主 OSD 将对象写入存储器,并负责写副本到其他副本 OSD 上。在主 OSD 将对象写到存储器,并且收到来自副本 OSD 的确认,即副本 OSD 写入对象完成之后,主 OSD 才会向客户端返回存储成功确认信息,从

4 结语

元件串并联的几种识别方法与二端口的连接方式,为分析与设计电路提供了一定理论支持。认清电路的结构,无论是对电路的故障排查还是对电路的改造升级,都至关重要,此外,显明的电路结构可以在设计复杂电路时起到化繁为简的作用,提高开发者的效率。在实际问题中应用以上方法时,应首先明确电路的特点,根据电路特征选取相适的方法,即使电路千变万化,也必能使其“现出原形”。

参考文献:

[1] 欧忠祥.识别串并联电路四法[J].物理教学参考,1996(10):15.
[2] 童诗白,华成英.模拟电子技术基础[M].北京:高等教育出版社,2015.
[3] 游祥岭.串并联电路识别的方法与运用[J].课程教育研究:新教师教学,2013(28)
[4] 邱关源,罗先觉.电路[M].北京:高等教育出版社,2006.
[5] 史源平,于京生,蔡文霞,袁莉.二端口网络的难点分析及应用[J].石家庄学院学报,2013,15(3):21-23

基金项目:教育部 2017 年产学合作协同育人项目“高水平应用型在线课程《工程电路分析》”(201701079019);湖北省教研项目“基于多维学习空间的在线开放课程建设与教学创新”(2017234)。
作者简介:王晓晨(1998-)男,本科生,研究方向:电子电路的分析、设计;张加齐(1998-)男,本科生,研究方向:电子电路的分析、设计;通讯作者:黄敬华(1968-),湖北武汉人,硕士,副教授,研究方向:电路与系统。

而保证集群中所有副本数据一致性。如 1 所示:acting 集合包含了三个 OSD,分别是 osd.2、osd.20、osd.40,其中 osd.2 是主 OSD,osd.20 和 osd.40 分别是第二 OSD 和第三 OSD。 osd.2 负责其他两个 osd 的数据写入。

在 Ceph 中,使用副本冗余存储方式时可以随时改变副本数。

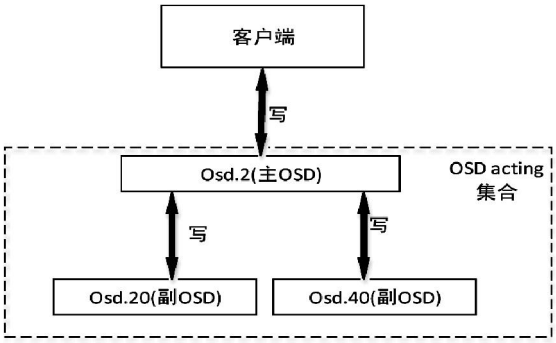

图 1 副本方式存储图

2.2 纠删码方式

Ceph 创建一个池的时候默认为副本方式存储,在 Ceph 0.8 版本中引入了纠删码方式。纠删码存储的基本原理是:首先把原数据分割成 K 个原始数据块,然后把 K 个原始数据块用纠删码编码算法进行编码,生成 M 个带有冗余信息的编码数据块,把这些 K+M 个数据块分别存储到不同的 OSD 中。当其中任意的 M 块数据(包括原始数据和冗余数据)出错,都可以通过 CURSH 算法恢复出原来的 K 块数据。利用这种方式存储数据可以允许 M 个 OSD 同时失效而不丢失数据。

使用纠删码方式存储数据,在数据丢失的情况下需要通过解码算法进行数据解码恢复,所以这种方式的数据访问效率比较低。如图 2 所示,在 Ceph 集群中使用纠删码向集群存储数据,首先创建一个纠删码池,设置 K=3,M=2。数据对象 obj 首先被分解成 3 个原始数据块 obj 1,obj 2 和 obj 3,通过编码算法生成两个冗余编码块 obj 4 与 obj 5。分别把这五个数据块存储到集群不同的 OSD 中,当其中任意 2 个 OSD 失效时,可以通过纠删码算法恢复数据。

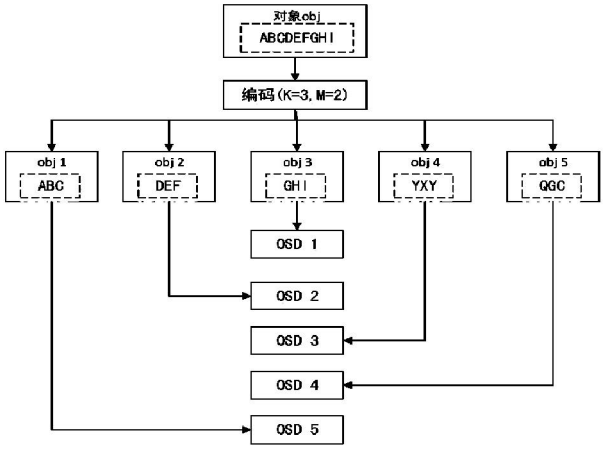

图 2 纠删码存储方式

2.3 副本冗余与纠删码冗余对比

在对数据进行冗余存储时,数据修复能力是衡量存储系统性能的关键因素,数据修复能力就是在节点失效后对数据的恢复能力。副本方式通过指定对象的副本数,设定了在不丢数据的情况下可以允许失效的 OSD 数量;而纠删码方式,则是指定数据块的数量。

副本方式与纠删码方式各有特点。在提供同级别可靠性的情况下,即数据修复能力同级时,纠删码方式比副本方式更节省存储空间。例如 对一个副本数为 2 的复制池来说,要存储 1TB 的数据需要 2TB 存储容量,但是纠删码池只需要 1.5TB 的存储容量。在数据丢失时,纠删码方式因为需要通过纠删码算法计算恢复原始数据,访问性能不如副本存储方式。如表 1 所示,列出了两种方式在提供同级别可靠性的情况下,各自的优缺点。

表 1 副本方式与纠删码方式对比

冗余存储机制	存储开销	数据访问性能	实现复杂度
副本方式	较高	良好	简单
纠删码方式	较低	一般	复杂

总之,副本与纠删码的选择是存储开销、数据可用性和数据访问性能等多种因素的综合权衡与折衷。在同一个 Ceph 集群的不同存储池中可以分别使用这两种数据冗余存储技术。

3 基于文件分类标记的冗余存储

根据副本与纠删码两种冗余存储机制的优缺点,提出一种结合二者优点的折中方案。此方案是在客户端对文件数据进行分类,可根据文件大小,以及访问频率对文件进行分类标记。对于比较大的文件可采用纠删码方式进行存储,比较小的文件采用副本方式存储;而对于客户端访问频繁的数据则采用副本方式存储,访问周期较长的数据采用纠删码方式存储。如图 3 所示:

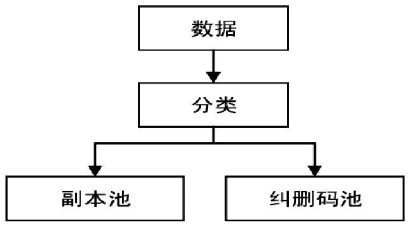

图 3 基于文件分类标记的冗余存储

4 结语

存储系统中冗余存储机制对数据可用性和数据可靠性尤为重要,通过合理的冗余机制可以保证 Ceph 存储集群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。本文通过对 Ceph 分布式存储中的两种冗余存储机制研究分析,结合他们的优点,根据数据的大小与访问频率对文件进行分类,不同类型的数据采用不同的冗余机制,对占用空间大的数据,采用纠删码方式,对占用空间小的数据,采用副本方式,充分节省存储空间;对访问频率高的数据,采用副本方式,对访问频率低的数据,采用纠删码方式,提高数据访问性能。实验表明,该种方案在保证系统可靠性和可用性的基础上,有效地节省了存储空间,同时也有较好的数据访问性能。

参考文献:

[1] 程振东,栾钟治,孟由,李亮淑,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(4).

[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).

[3] Ceph. http://docs.ceph.com/docs/master/rados/operations/, 2016.