(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104182184 B
(45)授权公告日 2017.08.25

(21)申请号 201410426148.1

(22)申请日 2014.08.27

(65)同一申请的已公布的文献号
申请公布号 CN 104182184 A

(43)申请公布日 2014.12.03

(73)专利权人 浪潮电子信息产业股份有限公司
地址 250101 山东省济南市高新区舜雅路1036号

(72)发明人 孟圣智 魏盟

(74)专利代理机构 济南信达专利事务所有限公司 37100
代理人 姜明

(51)Int.Cl.
G06F 3/06(2006.01)
G06F 11/14(2006.01)

(56)对比文件
CN 103870202 A,2014.06.18,
US 2006/0161810 A1,2006.07.20,
US 7631155 B1,2009.12.08,
CN 102541461 A,2012.07.04,
CN 102594849 A,2012.07.18,
US 2013/0254481 A1,2013.09.26,
US 2012/0124307 A1,2012.05.17,
CN 102012853 A,2011.04.13,
CN 102073739 A,2011.05.25,
US 8285758 B1,2012.10.09,

审查员 陈国耀

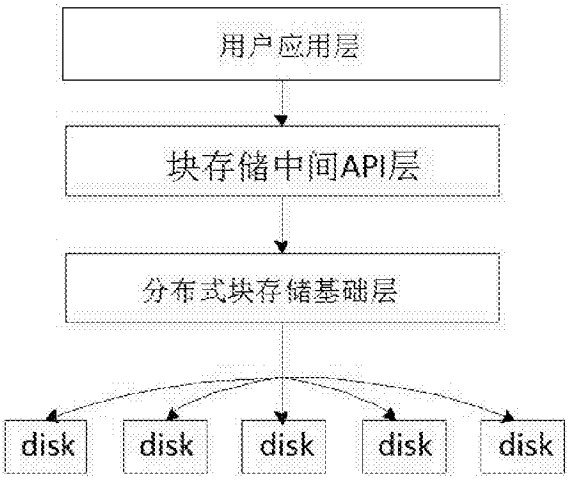
权利要求书2页 说明书3页 附图1页

(54)发明名称

一种分布式块存储克隆方法

(57)摘要

本发明公开了一种分布式块存储克隆方法，提出了一个存储系统，所述存储系统包括分布式块存储基础层、块存储中间API层和用户应用层；通过对存储系统的块设备卷建立快照，克隆时用户指定该卷的某个时期的一个快照ID建立新的块设备卷，用户所指定的卷称为父卷，所创建的新块设备卷称为克隆卷；然后建立克隆卷与父卷之间的映射关系，当向克隆卷发出读写请求时，从父卷的对应快照中获取数据复制到为克隆卷申请的存储空间中，再将新数据予以覆盖写入。通过该分布式块存储克隆方法，能够最大程度的保证克隆的高效性和克隆数据的一致性，使得用户的数据不会发生意外丢失和篡改，显著提高了克隆的稳定性和安全性。



CN 104182184 B

1.一种分布式块存储克隆方法,其特征在于,提出了一个存储系统,该存储系统包括分布式块存储基础层、块存储中间API层和用户应用层;其中,

所述分布式块存储基础层:完成对底层块设备的模拟,块设备为单独的物理磁盘disk或逻辑卷,所有块设备卷被编号并分组;同时负责接收并响应块存储中间API层发来的读写请求、消息调度、自适应调整权重,且负责完成所有对底层块设备的操作以及对块存储中间API层的接口支持;

所述块存储中间API层:进行接口封装,封装出的接口能够进行随机读写、克隆快照管理块存储方面的应用;同时,该层针对多线程进行优化,对数据缓存进行控制,监控分布式块存储基础层卷的状态和读写行为;

所述用户应用层:负责收集用户的传入参数和操作类型,并向块存储中间API层发起请求,并在执行完毕后给用户反馈结果;

所述分布式块存储克隆方法包括:对块设备卷建立快照的步骤,建立克隆卷与父卷之间映射关系的步骤以及对克隆卷发起读写请求的步骤;

对所述存储系统的块设备卷建立快照,克隆时用户指定该卷的某个时期的一个快照ID建立新的块设备卷,用户所指定的卷称为父卷,所创建的新块设备卷称为克隆卷;然后建立克隆卷与父卷之间的映射关系,当向克隆卷发出读写请求时,从父卷的对应快照中获取数据复制到为克隆卷申请的存储空间中,再将新数据予以覆盖写入。

2.根据权利要求1所述的一种分布式块存储克隆方法,其特征在于,所述对块设备卷建立快照的步骤包括:

完整记录存储系统的块设备卷中所有的可用数据,将卷各时期的数据进行分层,同时期的数据拥有相同的一个快照ID号,各快照之间通过ID号彼此区分,通过对卷建立快照将卷的数据存放形式变成层状分布D。

3.根据权利要求2所述的一种分布式块存储克隆方法,其特征在于,所述建立克隆卷与父卷之间映射关系的步骤包括:

用户指定父卷的一个快照ID,分布式块存储基础层接收到创建命令之后,根据用户给出的新卷名称创建一个克隆卷,将其基本信息注册到存储系统中;建立一个对象数据文件来记录克隆卷与父卷的对应关系,把克隆卷名称和父卷快照ID组成key-value对,通过LevelDB的接口将所述key-value对记录到存储系统中;同时,所述对象数据文件还记录着父卷所在的存储池、父卷名称以及所使用的父卷快照ID号。

4.根据权利要求3所述的一种分布式块存储克隆方法,其特征在于,所述对克隆卷发起读写请求的步骤包括:对克隆卷发起读请求和对克隆卷发起写请求两个过程。

5.根据权利要求4所述的一种分布式块存储克隆方法,其特征在于,当用户对克隆卷发起读请求时,所述存储系统先通过记录克隆卷和父卷之间映射关系的对象数据文件,查看克隆卷是否存在父卷并获得父卷的快照ID,将针对克隆卷的读请求转换成针对父卷对应快照的读请求,再配合用户给出的数据偏移量和读取数据长度,能够从父卷中获取数据。

6.根据权利要求5所述的一种分布式块存储克隆方法,其特征在于,当用户对克隆卷发出写请求时,所述存储系统先通过记录克隆卷和父卷之间映射关系的对象数据文件,查看克隆卷是否存在父卷并获得父卷的快照ID,然后再根据数据要写入的偏移地址和长度判断是否越界,在确定此次操作合法之后,先从父卷的对应快照中获取数据复制到为克隆卷申

请的存储空间中，然后再将新数据予以覆盖写入。

一种分布式块存储克隆方法

技术领域

[0001] 本发明涉及块存储领域，具体地说是一种分布式块存储克隆方法。

背景技术

[0002] 克隆技术的应用主要是为了满足人们对重复数据的需求。在实际应用中，往往要面对这样的场景，面对一份存放在块设备上的数据，既想对其进行修改，又不希望完全丢失其原始数据，比较快速的解决方法就是克隆出一个复本，在复本上进行修改和覆盖，而其原本保持不变。传统的克隆解决方案一般会选择完整地备份原始数据，这样的处理方式不仅效率低下，还会在业务高峰期占用较多的带宽，从而极大地影响用户使用，最大的隐患在于备份过程中如果原始数据被修改，则有可能造成克隆前后的数据不一致。

发明内容

[0003] 针对现有技术的不足之处，本发明提出一种分布式块存储克隆方法。

[0004] 本发明所述一种分布式块存储克隆方法，解决上述技术问题采用的技术方案如下：本发明所述分布式块存储克隆方法提出了一个存储系统，所述存储系统包括分布式块存储基础层、块存储中间API层和用户应用层；其中，所述分布式块存储基础层完成对底层物理块设备的模拟，底层设备为物理磁盘disk或逻辑卷，同时负责接收并响应块存储中间API层发来的读写请求；所述块存储中间API层监控分布式块存储基础层的状态和读写行为，所述用户应用层接收用户的管理类操作命令和读写命令，并向块存储中间API层发起请求。

[0005] 本实施例所述分布式块存储克隆方法，通过对存储系统的块设备卷建立快照，克隆时用户指定该卷的某个时期的一个快照ID建立新的块设备卷，用户所指定的卷称为父卷，所创建的新块设备卷称为克隆卷；然后建立克隆卷与父卷之间的映射关系，当向克隆卷发出读写请求时，从父卷的对应快照中获取数据复制到为克隆卷申请的存储空间中，再将新数据予以覆盖写入。克隆卷的原始数据完全来自于父卷的快照数据，没有额为申请空间，只有当覆盖写入时才真正使用物理空间，极大地提高了克隆操作的效率。所述分布式块存储克隆方法包括：对块设备卷建立快照的步骤，建立克隆卷与父卷之间映射关系的步骤以及对克隆卷发起读写请求的步骤。

[0006] 本发明所述一种分布式块存储克隆方法具有的有益效果：通过所述分布式块存储克隆方法，不再需要完整的备份所有数据，引入快照ID使得克隆可以瞬间完成，能够最大程度的保证克隆的高效性，不会在业务高峰期占用较多带宽，能够最大限度的减小对用户使用的影响；同时写请求的过程，使得用户的数据不会发生意外丢失和篡改，提高了克隆数据的一致性，从而显著提高了克隆的稳定性和安全性，该分布式块存储克隆方法更加适用于将来大规模普及的分布式块存储中。

附图说明

[0007] 附图1为本发明所述存储系统的示意图。

具体实施方式

[0008] 为使本发明的目的、技术方案和优点更加清楚明白，下文中将结合附图和实施例，对本发明的一种分布式块存储克隆方法进行详细说明。

[0009] 如附图1所示，本发明所述分布式块存储克隆方法提出了一个存储系统，所述存储系统包括分布式块存储基础层、块存储中间API层和用户应用层；其中，所述分布式块存储基础层完成对底层物理块设备的模拟，底层设备为物理磁盘disk或逻辑卷，同时负责接收并响应块存储中间API层发来的读写请求；所述块存储中间API层监控分布式块存储基础层的状态和读写行为，所述用户应用层接收用户的管理类操作命令和读写命令，并向块存储中间API层发起请求。

[0010] 下面对本发明所述存储系统的三个组成部分进行详细说明：

[0011] 所述分布式块存储基础层：该层完成对底层物理块设备的模拟，块设备为单独的物理磁盘disk或逻辑卷，它们是该层识别的最基本的块存储单元；所有块设备(卷)被编号并分组，以便通过一致性哈希将待存放的数据映射到对应的存储介质上；同时，所述分布式块存储基础层还负责接收并响应上层发来的读写请求、消息调度、自适应调整权重，因此，所有对底层块设备的操作以及对上层的接口支持均在该层完成；

[0012] 所述块存储中间API层：该层进行接口封装，封装出的接口能够进行随机读写、克隆快照管理块存储方面的应用；同时，该层也会针对多线程进行优化，对数据缓存进行控制，监控分布式块存储基础层卷的状态和读写行为；

[0013] 所述用户应用层：该层用于接收用户在终端或服务端所发出的管理类操作命令和读写命令，进行解析和筛选，并分别调用对应的中间层API(应用程序编程接口)完成功能；即用户应用层负责收集用户的传入参数和操作类型，并向块存储中间API层发起请求，执行完毕后给用户反馈结果。

[0014] 上述分布式块存储基础层、块存储中间API层和用户应用层三个部分是本存储系统的核心模块，三者共同协调配合完成包括克隆在内的所有块设备卷的高级功能。本存储系统的块存储单元卷都会拥有一个默认的快照ID，当对该卷的当前数据添加快照之后，此快照ID会自动增1，成为最新的快照的ID号；同时对卷进行读写时需要指定快照的ID号，这样可以确定用户当前需要对卷的哪一部分数据进行读写操作。

[0015] 实施例：

[0016] 下面通过一个实施例，对本发明所述分布式块存储克隆方法的优点和设计内容，进行详细说明。

[0017] 本实施例所述分布式块存储克隆方法，提出了一个存储系统，通过对存储系统的块设备卷建立快照，克隆时用户指定该卷的某个时期的一个快照ID建立新的块设备卷，用户所指定的卷称为父卷，所创建的新块设备卷称为克隆卷；然后建立克隆卷与父卷之间的映射关系，当向克隆卷发出读写请求时，从父卷的对应快照中获取数据复制到为克隆卷申请的存储空间中，再将新数据予以覆盖写入。克隆卷的原始数据完全来自于父卷的快照数据，没有额为申请空间，只有当覆盖写入时才真正使用物理空间，极大地提高了克隆操作的效率。所述分布式块存储克隆方法包括：对块设备卷建立快照的步骤，建立克隆卷与父卷之

间映射关系的步骤以及对克隆卷发起读写请求的步骤。

[0018] 所述对块设备卷建立快照的步骤包括：完整记录存储系统的块设备卷中所有的可用数据，将卷各时期的数据进行分层，同时期的数据拥有相同的一个快照ID号，各快照之间通过ID号彼此区分，通过对卷建立快照将卷的数据存放形式变成层状分布；需要访问块设备卷某时期的数据时，只需指定其对应的快照ID。

[0019] 所述建立克隆卷与父卷之间映射关系的步骤包括：用户指定父卷的一个快照ID，分布式块存储基础层接收到创建命令之后，根据用户给出的新卷名称创建一个克隆卷，将其基本信息注册到存储系统中；

[0020] 建立一个对象数据文件来记录克隆卷与父卷的对应关系，使用LevelDB以key-value的形式标识出克隆卷与其父卷的对应关系，即把克隆卷名称和派生它的父卷快照ID组成key-value对，通过LevelDB的接口将所述key-value对记录到存储系统中；所述LevelDB是Google开源非常高效的kv数据库，提供单进程的服务，具有非常高的随机读写性能。同时，所述对象数据文件还记录着父卷所在的存储池、父卷名称以及所使用的父卷快照ID号等信息，以此来锁定克隆卷和父卷的对应关系。

[0021] 所述对克隆卷发起读写请求的步骤包括：对克隆卷发起读请求和对克隆卷发起写请求两个过程；

[0022] 当用户对克隆卷发起读请求时，所述存储系统先通过记录克隆卷和父卷之间映射关系的对象数据文件，查看该卷是否存在父卷并获得父卷的快照ID，将针对克隆卷的读请求转换成针对父卷对应快照的读请求，再配合用户给出的数据偏移量和读取数据长度，即可从父卷中获取数据。而在上层用户看来，像是从一个真正的新卷里获取数据；

[0023] 当用户对克隆卷发出写请求时，所述存储系统先通过记录克隆卷和父卷之间映射关系的对象数据文件，查看该卷是否存在父卷并获得父卷的快照ID，然后再根据数据要写入的偏移地址和长度判断是否越界，在确定此次操作合法之后，存储系统为保证父卷和克隆卷数据的一致性，先从父卷的对应快照中获取数据复制到为克隆卷申请的存储空间中，然后再将新数据予以覆盖写入。从父卷将相应偏移处的数据进行拷贝，是为克隆卷申请存储空间的过程，这样做能在最大限度上保证父卷和克隆卷数据的一致性，确保用户不想修改部分的数据的安全。

[0024] 上述具体实施方式仅是本发明的具体个案，本发明的专利保护范围包括但不限于上述具体实施方式，任何符合本发明的权利要求书的且任何所属技术领域的普通技术人员对其所做的适当变化或替换，皆应落入本发明的专利保护范围。

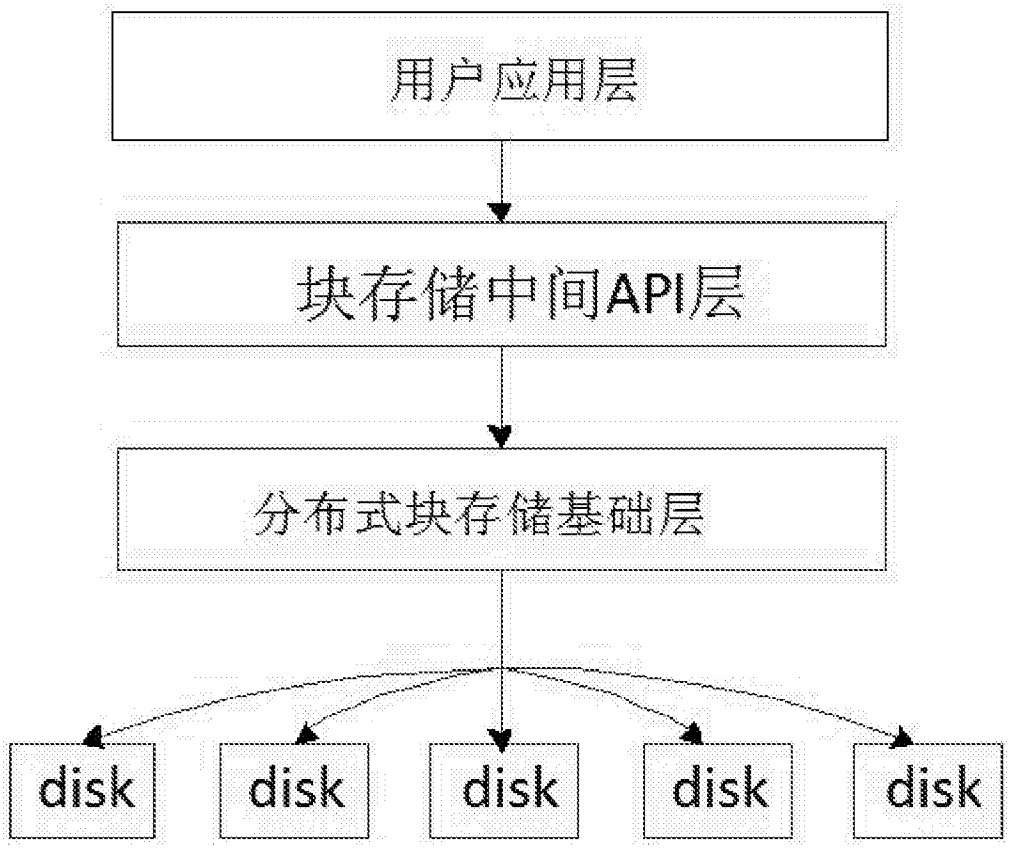


图1