(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106230998 A
(43)申请公布日 2016.12.14

(21)申请号 201610874206.6

(22)申请日 2016.10.08

(71)申请人 深圳市云舒网络技术有限公司
地址 518000 广东省深圳市南山区南山街道高新南一道006号TCL工业研究院大厦A座九楼A902室

(72)发明人 陈仲涛

(51)Int.Cl.
H04L 29/08(2006.01)

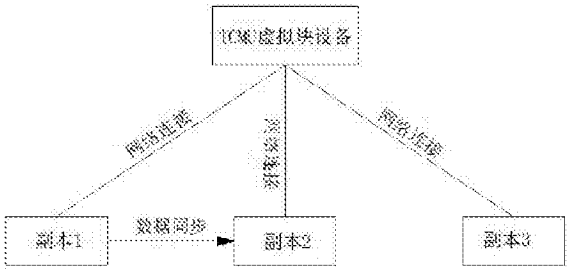
权利要求书1页 说明书3页 附图2页

(54)发明名称

一种基于TCMU虚拟块设备的多后端数据同步方案

(57)摘要

本发明公开了一种基于TCMU虚拟块设备的多后端数据同步方案，包括前端TCMU虚拟块设备和后端存储设备，后端存储设备包括三个数据副本，前端的虚拟块设备对应后端的三个数据副本，本发明的前端应用为TCMU虚拟块设备，其最小写IO为4K，并且数据4K对齐，使用bitmap来记录数据的同步情况，内存消耗比较小，但是，如果只使用一级bitmap，bitmap中1位代表4K数据时，bitmap还是很消耗内存，并且扫描一遍的时间比较久，如果bitmap中1位代表较大数据时，如4M，内存占用会小很多，但是前端如果只修改了4K数据，又必须重新同步4M数据，这样会降低速度，影响网络和磁盘带宽。本发明综合两者的优点，使用二级bitmap，即占用小的内存，又不影响同步速度，网络和磁盘带宽。



CN 106230998 A

1.一种基于TCMU虚拟块设备的多后端数据同步方案，包括前端TCMU虚拟块设备和后端存储设备，其特征在于：所述的后端存储设备包括三个数据副本，前端的虚拟块设备对应后端的三个数据副本。

2.根据权利要求1所述的一种基于TCMU虚拟块设备的多后端数据同步方案，其特征在于：假设数据大小为40G，副本2数据丢失，副本1向副本2进行数据同步，数据同步的具体的流程步骤如下：

副本1初始化一级bitmap，每一位代表4M数据，共10240位，初始化值全为1；

副本1按照一级bitmap的顺序，发包请求副本2对位置的数据哈希值。

3.根据权利要求2所述的一种基于TCMU虚拟块设备的多后端数据同步方案，其特征在于：副本2根据请求包携带的偏移量和数据长度，计算数据的哈希值，然后回包。

4.根据权利要求3所述的一种基于TCMU虚拟块设备的多后端数据同步方案，其特征在于：副本1收到副本2的回包后，计算对应数据的哈希值，然后跟副本2回包携带的哈希值比较，如果相等，数据不需要同步，把一级bitmap对应位置0，如果不相等， 发送同步数据包，把一级bitmap对应位置0。

5.根据权利要求4所述的一种基于TCMU虚拟块设备的多后端数据同步方案，其特征在于：副本2收到同步数据包，把数据写入，回包返回写入成功或者失败；

副本1如果收到副本2返回失败，把一级bitmap对应位置重新设置为1，表示需要重新同步，如果有对应的二级bitmap，删除该二级bitmap。

6.根据权利要求5所述的一种基于TCMU虚拟块设备的多后端数据同步方案，其特征在于：由于后端数据同步的过程中，前端还有数据写入，使用二级bitmap来记录被修改的数据，前16M数据已经同步完成，如果前端的写入的数据偏移量大于16M，对应的一级bitmap为1，表示还没进行同步，所以之间写入数据，不用生成对应的二级bitmap，如果前端的写入的数据偏移量小于16M，对应的一级bitmap为0，表示已经同步过，此时需要生成对应的二级bitmap，记录已经被修改的数据，二级bitmap每一位代表4K数据，共1024位，二级bitmap保存在hash map中，键值为一级bitmap的索引号。

7.根据权利要求1-6所述的一种基于TCMU虚拟块设备的多后端数据同步方案，其特征在于：重复以上步骤，直到一级bitmap全为0，再遍历二级bitmap，根据二级bitmap组装同步数据包，二级bitmap同步中不用先验证数据的哈希值，副本2返回写入成功时，对应二级bitmap位置0，如果一个二级bitmap所有位全为0，删除该二级bitmap。

8.根据权利要求7所述的一种基于TCMU虚拟块设备的多后端数据同步方案，其特征在于：重复上述步骤，直到只存在少数的二级bitmap，此时暂停前端处理写IO，再次重复上述步骤，直到二级bitmap为空，数据同步完成，再次启动处理前端写IO和启动副本2的前端写IO。

一种基于TCMU虚拟块设备的多后端数据同步方案

技术领域

[0001] 本发明涉及多后端数据同步技术领域，具体为一种基于TCMU虚拟块设备的多后端数据同步方案。

背景技术

[0002] 随着大数据时代的来临，基于互联网或者基于通信网络的服务提供者需要存储海量的数据，以支持其运营，同时，服务提供者通过对海量数据的分析，可向用户提供更便利且更具有个性化的服务，从而达到提高其服务水平的目的。

[0003] 目前，通常由分布式存储系统来存储海量的数据，例如，技术人员将服务器集群分散布置，将海量的数据按照某些规则分散存储在不同的服务器中；这些服务器通常可以根据用户的请求，为用户提供数据查询、数据更新等服务，为了提高数据的安全性和可靠性，分布式存储系统通常还设置有一定数量的备份服务器，以用于备份数据。

[0004] 分布式存储系统，是将数据分散存储在多台独立的设备上，传统的网络存储系统采用集中的存储服务器存放所有数据，存储服务器成为系统性能的瓶颈，也是可靠性和安全性的焦点，不能满足大规模存储应用的需要，分布式网络存储系统采用可扩展的系统结构，利用多台存储服务器分担存储负荷，利用位置服务器定位存储信息，它不但提高了系统的可靠性、可用性和存取效率，还易于扩展。

[0005] 现有的分布式存储系统都是通过网络通信，网络的不稳定性容易造成后端数据不一致，如果不能快速进行数据同步，分布式存储系统的数据完整性和高可用性就大大降低，现有技术数据同步过慢，占用过高的网络带宽，并且在有写IO的情况下很难达到数据一致性。

发明内容

[0006] 本发明要解决的技术问题是克服现有的缺陷，提供一种基于TCMU虚拟块设备的多后端数据同步方案，结合TCMU虚拟块设备以及二级bitmap占用小的内存的优点，使得该基于TCMU虚拟块设备的多后端数据同步方案具有高速度，低内存消耗，低网络带宽消耗，低磁盘带宽消耗的特点，可以有效解决背景技术中的问题。

[0007] 为实现上述目的，本发明提供如下技术方案：一种基于TCMU虚拟块设备的多后端数据同步方案，包括前端TCMU虚拟块设备和后端存储设备，所述的后端存储设备包括三个数据副本，前端的虚拟块设备对应后端的三个数据副本。

[0008] 作为本发明的一种优选技术方案，假设数据大小为40G，副本2数据丢失，副本1向副本2进行数据同步，数据同步的具体的流程步骤如下：

副本1初始化一级bitmap，每一位代表4M数据，共10240位，初始化值全为1。

[0009] 副本1按照一级bitmap的顺序，发包请求副本2对位置的数据哈希值。

[0010] 作为本发明的一种优选技术方案，副本2根据请求包携带的偏移量和数据长度，计算数据的哈希值，然后回包。

[0011] 作为本发明的一种优选技术方案，副本1收到副本2的回包后，计算对应数据的哈希值，然后跟副本2回包携带的哈希值比较，如果相等，数据不需要同步，把一级bitmap对应位置0，如果不相等，发送同步数据包，把一级bitmap对应位置0。

[0012] 作为本发明的一种优选技术方案，副本2收到同步数据包，把数据写入，回包返回写入成功或者失败。

[0013] 副本1如果收到副本2返回失败，把一级bitmap对应位置重新设置为1，表示需要重新同步，如果有对应的二级bitmap，删除该二级bitmap。

[0014] 作为本发明的一种优选技术方案，由于后端数据同步的过程中，前端还有数据写入，使用二级bitmap来记录被修改的数据，前16M数据已经同步完成，如果前端的写入的数据偏移量大于16M，对应的一级bitmap为1，表示还没进行同步，所以之间写入数据，不用生成对应的二级bitmap，如果前端的写入的数据偏移量小于16M，对应的一级bitmap为0，表示已经同步过，此时需要生成对应的二级bitmap，记录已经被修改的数据。二级bitmap每一位代表4K数据，共1024位，二级bitmap保存在hash map中，键值为一级bitmap的索引号。

[0015] 作为本发明的一种优选技术方案，重复以上步骤，直到一级bitmap全为0，再遍历二级bitmap，根据二级bitmap组装同步数据包，二级bitmap同步中不用先验证数据的哈希值，副本2返回写入成功时，对应二级bitmap位置0，如果一个二级bitmap所有位全为0，删除该二级bitmap。

[0016] 作为本发明的一种优选技术方案，重复上述步骤，直到只存在少数的二级bitmap，此时暂停前端处理写IO，再次重复上述步骤，直到二级bitmap为空，数据同步完成，再次启动处理前端写IO和启动副本2的前端写IO。

[0017] 与现有技术相比，本发明的有益效果是：本基于TCMU虚拟块设备的多后端数据同步方案的前端应用为TCMU虚拟块设备，其最小写IO为4K，并且数据4K对齐，使用bitmap来记录数据的同步情况，内存消耗比较小，但是，如果只使用一级bitmap，bitmap中1位代表4K数据时，bitmap还是很消耗内存，并且扫描一遍的时间比较久，如果bitmap中1位代表较大数据时，如4M，内存占用会小很多，但是前端如果只修改了4K数据，又必须重新同步4M数据，结合TCMU虚拟块设备以及二级bitmap占用小的内存的优点，使得该基于TCMU虚拟块设备的多后端数据同步方案具有高速度，低内存消耗，低网络带宽消耗，低磁盘带宽消耗的特点。

附图说明

[0018] 图1为本发明的分布式存储系统结构示意图；

图2为本发明的一级bitmap结构示意图；

图3为本发明的二级bitmap结构示意图。

具体实施方式

[0019] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0020] 请参阅图1-3，本发明提供一种技术方案：一种基于TCMU虚拟块设备的多后端数据

同步方案，包括前端TCMU虚拟块设备和后端存储设备，所述的后端存储设备包括三个数据副本，前端的虚拟块设备对应后端的三个数据副本。

[0021] 假设数据大小为40G，副本2数据丢失，副本1向副本2进行数据同步，数据同步的具体的流程步骤如下：

步骤1：如图2所示，副本1初始化一级bitmap，每一位代表4M数据，共10240位，初始化值全为1。

[0022] 步骤2：副本1按照一级bitmap的顺序，发包请求副本2对位置的数据哈希值。

[0023] 步骤3：副本2根据请求包携带的偏移量和数据长度，计算数据的哈希值，然后回包。

[0024] 步骤4：副本1收到副本2的回包后，计算对应数据的哈希值，然后跟副本2回包携带的哈希值比较，如果相等，数据不需要同步，把一级bitmap对应位置0，如果不相等， 发送同步数据包，把一级bitmap对应位置0。

[0025] 步骤5：副本2收到同步数据包，把数据写入，回包返回写入成功或者失败。

[0026] 步骤6：副本1如果收到副本2返回失败，把一级bitmap对应位置重新设置为1，表示需要重新同步，如果有对应的二级bitmap，删除该二级bitmap。

[0027] 步骤7：由于后端数据同步的过程中，前端还有数据写入，使用二级bitmap来记录被修改的数据，如图3所示，前16M数据已经同步完成，如果前端的写入的数据偏移量大于16M，对应的一级bitmap为1，表示还没进行同步，所以之间写入数据，不用生成对应的二级bitmap，如果前端的写入的数据偏移量小于16M，对应的一级bitmap为0，表示已经同步过，此时需要生成对应的二级bitmap，记录已经被修改的数据，二级bitmap每一位代表4K数据，共1024位，二级bitmap保存在hash map中，键值为一级bitmap的索引号。

[0028] 步骤8：重复以上步骤，直到一级bitmap全为0，再遍历二级bitmap，根据二级bitmap组装同步数据包，二级bitmap同步中不用先验证数据的哈希值，副本2返回写入成功时，对应二级bitmap位置0，如果一个二级bitmap所有位全为0，删除该二级bitmap。

[0029] 步骤9：重复步骤8，直到只存在少数的二级bitmap，此时暂停前端处理写IO，再次重复步骤8，直到二级bitmap为空，数据同步完成，再次启动处理前端写IO和启动副本2的前端写IO。

[0030] 本发明提出一种快速、内存开销小、节省网络带宽的块设备多后端数据同步方案，前端应用为TCMU虚拟块设备，其最小写IO为4K，并且数据4K对齐，使用bitmap来记录数据的同步情况，内存消耗比较小，但是，如果只使用一级bitmap，bitmap中1位代表4K数据时，bitmap还是很消耗内存，并且扫描一遍的时间比较久，如果bitmap中1位代表较大数据时，如4M，内存占用会小很多，但是前端如果只修改了4K数据，又必须重新同步4M数据，这样会降低速度，影响网络和磁盘带宽，本发明综合两者的优点，使用二级bitmap，即占用小的内存，又不影响同步速度，网络和磁盘带宽。

[0031] 尽管已经示出和描述了本发明的实施例，对于本领域的普通技术人员而言，可以理解在不脱离本发明的原理和精神的情况下可以对这些实施例进行多种变化、修改、替换和变型，本发明的范围由所附权利要求及其等同物限定。

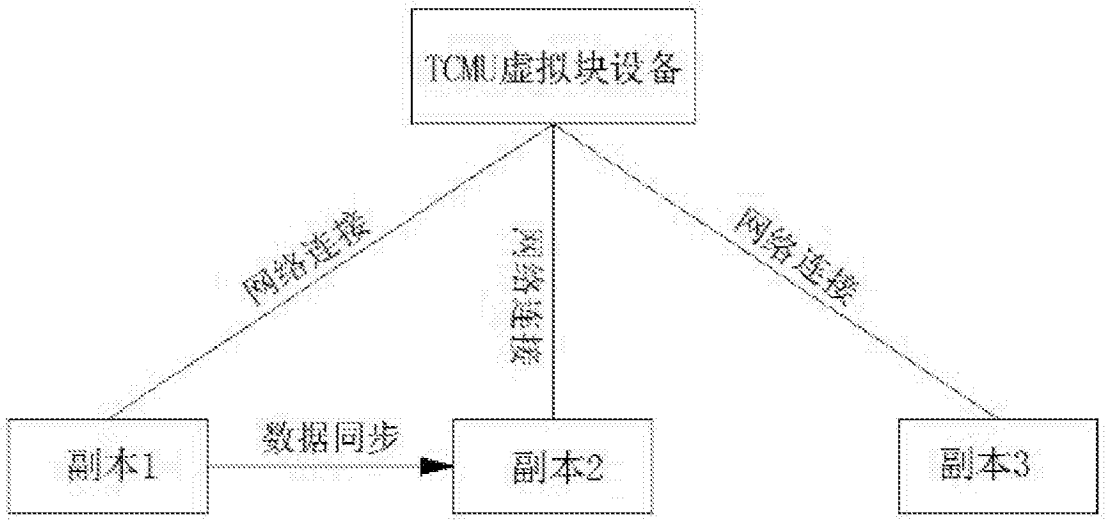

TCM虚拟块设备
网络连接
网络连接
网络连接
副本1
数据同步
副本2
副本3

图1

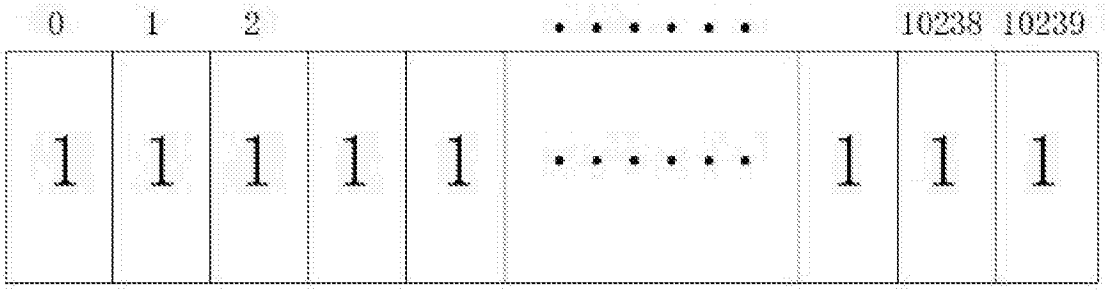

0
1
2
10238
10239
1
1
1
1
1
1
1
1

图2

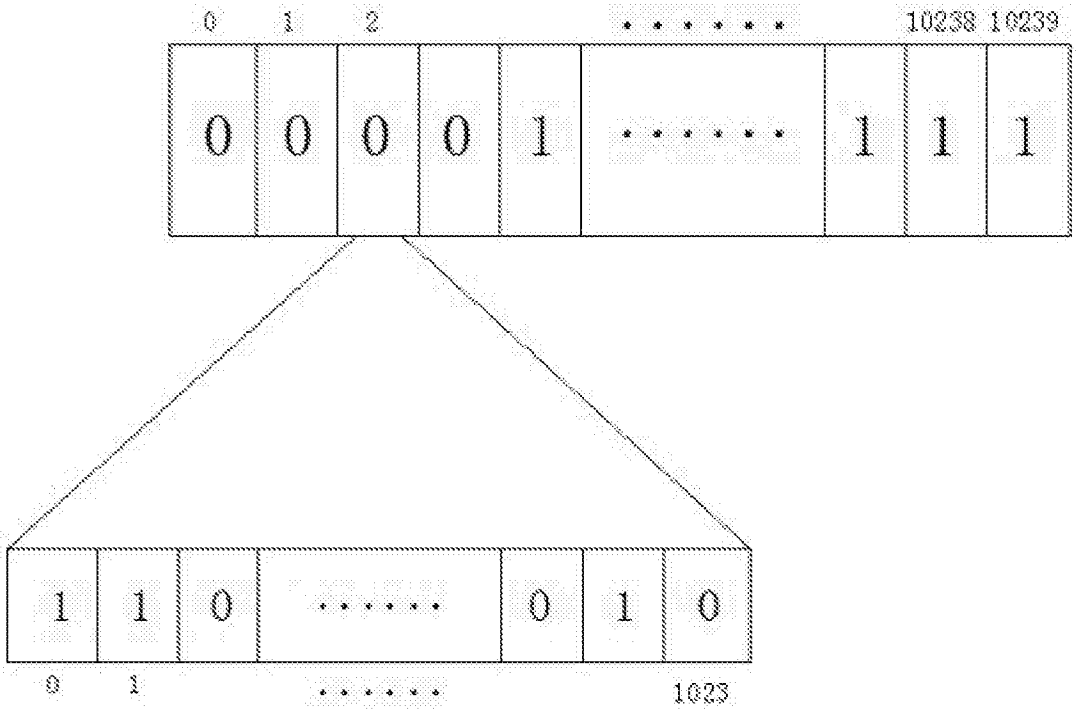


图3