(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 109976663 B
(45)授权公告日 2021.12.28

(21)申请号 201711452479.2

(22)申请日 2017.12.27

(65)同一申请的已公布的文献号
申请公布号 CN 109976663 A

(43)申请公布日 2019.07.05

(73)专利权人 浙江宇视科技有限公司
地址 310000 浙江省杭州市滨江区西兴街道江陵路88号10幢南座1-11层、2幢A区1-3楼、2幢B区2楼

(72)发明人 李华英 王丽红 郭永强

(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371
代理人 金相允

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

(56)对比文件
CN 103761195 A,2014.04.30
CN 103414921 A,2013.11.27
CN 101689131 A,2010.03.31
CN 106662983 A,2017.05.10
US 2017262191 A1,2017.09.14
US 2008109616 A1,2008.05.08
US 2016291901 A1,2016.10.06
CN 105893169 A,2016.08.24
US 2014359218 A1,2014.12.04
US 2017017413 A1,2017.01.19

审查员 刘褚燚

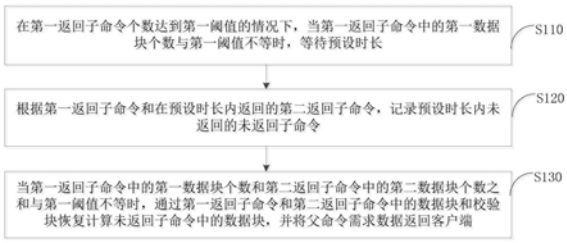
权利要求书2页 说明书8页 附图3页

(54)发明名称

分布式存储响应方法和系统

(57)摘要

本发明提供了分布式存储响应方法和系统，涉及存储应用技术领域，包括在第一返回子命令个数达到第一阈值的情况下，当第一返回子命令中的第一数据块个数与第一阈值不等时，等待预设时长；根据第一返回子命令和在预设时长内返回的第二返回子命令，记录在预设时长内未返回的未返回子命令；当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值不等时，通过第一返回子命令和第二返回子命令中的数据块和校验块恢复计算未返回子命令中的数据块，并将父命令需求数据返回客户端，其中，父命令需求数据为各个子命令中的数据块内容，第一阈值为父命令中数据块的个数，本申请极大程度地提升响应速度。



CN 109976663 B

1.一种分布式存储响应方法,其特征在于,应用于分布式存储系统,所述方法包括:

在第一返回子命令个数达到第一阈值的情况下,当所述第一返回子命令中的第一数据块个数与所述第一阈值不等时,等待预设时长;

根据所述第一返回子命令和在所述预设时长内返回的第二返回子命令,记录所述预设时长内未返回的未返回子命令;

当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值不等时,通过所述第一返回子命令和所述第二返回子命令中的数据块和校验块恢复计算所述未返回子命令中的数据块,并将父命令需求数据返回客户端,其中,所述父命令需求数据为各个子命令中的数据块内容,所述第一阈值为所述父命令中数据块的个数。

2.根据权利要求1所述的分布式存储响应方法,其特征在于,所述方法还包括:当所述第一返回子命令中的第一数据块个数与所述第一阈值相等时,将所述父命令需求数据返回客户端。

3.根据权利要求1或2所述的分布式存储响应方法,其特征在于,在所述当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值不等时,等待预设时长之前,还包括:

判断所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值是否相等;

当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的所述第二数据块个数之和与所述第一阈值相等时,将所述父命令需求数据返回所述客户端。

4.根据权利要求1所述的分布式存储响应方法,其特征在于,所述记录在所述预设时长内未返回的未返回子命令包括:

将超过预设时长仍未返回的未返回子命令进行标记记录,并记录所述未返回子命令所在的各个磁盘;

分别统计所述磁盘中所述未返回子命令的个数,当所述个数超过第二阈值时,进行告警或踢出所述磁盘。

5.根据权利要求1所述的分布式存储响应方法,其特征在于,所述方法还包括:当所述未返回子命令返回时,将所述未返回子命令回收。

6.一种分布式存储响应系统,其特征在于,应用于分布式存储系统,所述系统包括:

等待模块,在第一返回子命令个数达到第一阈值的情况下,当所述第一返回子命令中的第一数据块个数与所述第一阈值不等时,等待预设时长;

记录模块,根据所述第一返回子命令和在所述预设时长内返回的第二返回子命令,记录所述预设时长内未返回的未返回子命令;

恢复计算模块,当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值不等时,通过所述第一返回子命令和所述第二返回子命令中的数据块和校验块恢复计算所述未返回子命令中的数据块,并将父命令需求数据返回客户端,其中,所述父命令需求数据为各个子命令中的数据块内容,所述第一阈值为所述父命令中数据块的个数。

7.根据权利要求6所述的分布式存储响应系统,其特征在于,所述系统还包括返回模

块，当所述第一返回子命令中的第一数据块个数与所述第一阈值相等时，将所述父命令需求数据返回客户端。

8.根据权利要求6或7所述的分布式存储响应系统，其特征在于，还包括判断模块，判断所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值是否相等；当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的所述第二数据块个数之和与所述第一阈值相等时，将所述父命令需求数据返回所述客户端。

9.根据权利要求6所述的分布式存储响应系统，其特征在于，所述记录模块还用于将超过预设时长仍未返回的未返回子命令进行标记记录，并记录所述未返回子命令所在的各个磁盘，分别统计所述磁盘中所述未返回子命令的个数，当所述个数超过第二阈值时，进行告警或踢出所述磁盘。

10.根据权利要求6所述的分布式存储响应系统，其特征在于，所述系统还包括回收模块，当所述未返回子命令返回时，将所述未返回子命令回收。

分布式存储响应方法和系统

技术领域

[0001] 本发明涉及存储应用技术领域，尤其是涉及分布式存储响应方法和系统。

背景技术

[0002] 分布式存储系统，是由分布于不同地理位置的多台存储设备共同组成的集群，通过网络，对外提供存储服务，根据不同的容灾策略，可以容忍多个级别的存储设备故障。在监控视频领域中，分布式存储系统是未来的发展方向。

[0003] 在监控视频应用中，经常需要实时多路视频流长时间写入，这就要求存储系统需具有高带宽且恒定，且IO最大延迟不能超出前端监控视频设备的容忍上限，否则会导致部分视频数据丢失。另外，多路视频回放时，多路并发读取对存储设备的性能要求非常高，视频数据是均衡分布到集群的各个存储节点上，如果某个存储节点响应比较慢，就会出现大量回放卡顿现象，影响被放大。

发明内容

[0004] 有鉴于此，本发明的目的在于提供分布式存储响应方法和系统，极大程度地提升响应速度。

[0005] 第一方面，本发明实施例提供了分布式存储响应方法，应用于分布式存储系统，所述方法包括：

[0006] 在第一返回子命令个数达到第一阈值的情况下，当所述第一返回子命令中的第一数据块个数与所述第一阈值不等时，等待预设时长；

[0007] 根据所述第一返回子命令和在所述预设时长内返回的第二返回子命令，记录所述预设时长内未返回的未返回子命令；

[0008] 当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值不等时，通过所述第一返回子命令和所述第二返回子命令中的数据块和校验块恢复计算所述未返回子命令中的数据块，并将父命令需求数据返回客户端，其中，所述父命令需求数据为各个子命令中的数据块内容，所述第一阈值为所述父命令中数据块的个数。

[0009] 结合第一方面，本发明实施例提供了第一方面的第一种可能的实施方式，其中，所述方法还包括：当所述第一返回子命令中的第一数据块个数与所述第一阈值相等时，将所述父命令需求数据返回客户端。

[0010] 结合第一方面，本发明实施例提供了第一方面的第二种可能的实施方式，其中，在所述当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值不等时，等待预设时长之前，还包括：

[0011] 判断所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值是否相等；

[0012] 当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的所述第

二数据块个数之和与所述第一阈值相等时，将所述父命令需求数据返回所述客户端。

[0013] 结合第一方面，本发明实施例提供了第一方面的第三种可能的实施方式，其中，所述记录在所述预设时长内未返回的未返回子命令包括：

[0014] 将超过预设时长仍未返回的未返回子命令进行标记记录，并记录所述未返回子命令所在的各个磁盘；

[0015] 分别统计所述磁盘中所述未返回子命令的个数，当所述个数超过第二阈值时，进行告警或踢出所述磁盘。

[0016] 结合第一方面，本发明实施例提供了第一方面的第四种可能的实施方式，其中，所述方法还包括：当所述未返回子命令返回时，将所述未返回子命令回收。

[0017] 第二方面，本发明实施例还提供分布式存储响应系统，应用于分布式存储系统，所述系统包括：

[0018] 等待模块，在第一返回子命令个数达到第一阈值的情况下，当所述第一返回子命令中的第一数据块个数与所述第一阈值不等时，等待预设时长；

[0019] 记录模块，根据所述第一返回子命令和在所述预设时长内返回的第二返回子命令，记录所述预设时长内未返回的未返回子命令；

[0020] 恢复计算模块，当所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值不等时，通过所述第一返回子命令和所述第二返回子命令中的数据块和校验块恢复计算所述未返回子命令中的数据块，并将父命令需求数据返回客户端，其中，所述父命令需求数据为各个子命令中的数据块内容，所述第一阈值为所述父命令中数据块的个数。

[0021] 结合第二方面，本发明实施例提供了第二方面的第一种可能的实施方式，其中，所述系统还包括返回模块，当所述第一返回子命令中的第一数据块个数与所述第一阈值相等时，将所述父命令需求数据返回客户端。

[0022] 结合第二方面，本发明实施例提供了第二方面的第二种可能的实施方式，其中，还包括判断模块，判断所述第一返回子命令中的第一数据块个数和所述第二返回子命令中的第二数据块个数之和与所述第一阈值是否相等；当所述第一返回子命令和所述第二返回子命令中的数据块个数与所述第一阈值相等时，将所述父命令需求数据返回所述客户端。

[0023] 结合第二方面，本发明实施例提供了第二方面的第三种可能的实施方式，其中，所述记录模块还用于将超过预设时长仍未返回的未返回子命令进行标记记录，并记录所述未返回子命令所在的各个磁盘，分别统计所述磁盘中所述未返回子命令的个数，当所述个数超过第二阈值时，进行告警或踢出所述磁盘。

[0024] 结合第二方面，本发明实施例提供了第二方面的第四种可能的实施方式，其中，所述系统还包括回收模块，当所述未返回子命令返回时，将所述未返回子命令回收。

[0025] 本发明实施例提供了分布式存储响应方法和系统，应用于分布式存储系统，包括在第一返回子命令个数达到第一阈值的情况下，当第一返回子命令中的第一数据块个数与第一阈值不等时，等待预设时长；根据第一返回子命令和在预设时长内返回的第二返回子命令，记录在预设时长内未返回的未返回子命令；当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值不等时，通过第一返回子命令和第二返回子命令中的数据块和校验块恢复计算未返回子命令中的数据块，并将父命令需求数

据返回客户端，其中，父命令需求数据为各个子命令中的数据块内容，第一阈值为父命令中数据块的个数，极大程度地提升响应速度。

[0026] 本发明的其他特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点在说明书、权利要求书以及附图中所特别指出的结构来实现和获得。

[0027] 为使本发明的上述目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合所附附图，作详细说明如下。

附图说明

[0028] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案，下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施方式，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0029] 图1为本发明实施例提供的分布式存储响应方法流程图；

[0030] 图2为本发明实施例提供的又一分布式存储响应方法流程图；

[0031] 图3为本发明实施例提供的分布式存储响应系统功能框图。

具体实施方式

[0032] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合附图对本发明的技术方案进行清楚、完整地描述，显然，所描述的实施例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。

[0033] 目前，在监控视频应用中，经常需要实时多路视频流长时间写入，这就要求存储系统需具有高带宽且恒定，且IO最大延迟不能超出前端监控视频设备的容忍上限，否则会导致部分视频数据丢失。另外，多路视频回放时，多路并发读取对存储设备的性能要求非常高，视频数据是均衡分布到集群的各个存储节点上，如果某个存储节点响应比较慢，就会出现大量回放卡顿现象，影响被放大。

[0034] 基于此，本发明实施例提供的分布式存储响应方法和系统，极大程度地提升响应速度。

[0035] 为便于对本实施例进行理解，首先对本发明实施例所公开的分布式存储响应方法进行详细介绍，

[0036] 随着高清摄像头越发普及，平安城市等大规模视频监控部署，同步需要配套大量后端存储设备。在监控视频应用中，引入大量分布式存储系统开源项目，比如HDFS、Swift、CEPH等，而视频监控应用中，由于视频码流需要多路视频长时间同时写入同一个存储设备，且要求此存储系统能长期稳定工作的写入特点，分布式存储与传统存储相比可以提供更好的容灾策略选择，副本或纠删码策略部署在整个存储集群上，可以设置其故障保护域为主机或机架(不同机架独立电源)等不同级别，这样副本或纠删码的分片数据就会存储到不同的主机或不同机架的主机设备上，当一台主机故障或一个机架电源断电，其它正常提供服务的存储节点只要达到副本或纠删码算法最小可用数目要求(比如3副本，当有2台存储节

点故障，另外1台存储节点仍可对外提供正常读写服务；纠删码8+2策略，当有两台存储节点故障，另外八台存储节点仍可对外提供正常读写服务)，整个集群仍可对外提供正常存储服务，不会导致视频监控数据丢失；

[0037] 这里，目前存在的各种分布式存储系统，管理方式在宏观上基本一致，具体软件架构和技术手段各不相同。存储集群在收到读写命令后，一般处理流程：1)对接收到的读写命令按接收顺序加入一个队列等待处理线程处理；2)处理线程(单线程或多线程)从队列头依次取出命令进行处理；3)对于一个读写命令，先从数据库或元数据服务器读取相关元数据管理信息，确定写入/读取位置及是首次写入还是修改写等信息，然后执行写入/读取动作(比如根据副本或纠删码策略，对父命令进行拆分校验后，分发子命令到其它节点执行具体分片子命令的读写处理)；

[0038] 这里，本发明实施例适用于应用纠删码保护策略或多副本策略的分布式存储集群系统中；

[0039] 图1为本发明实施例提供的分布式存储响应方法流程图。

[0040] 参照图1，分布式存储响应方法，应用于分布式存储系统，包括以下步骤：

[0041] 步骤S110，在第一返回子命令个数达到第一阈值的情况下，当第一返回子命令中的第一数据块个数与第一阈值不等时，等待预设时长；

[0042] 这里，当返回的第一返回子命令个数达到第一阈值时，此时第一返回子命令中的第一数据块个数与第一阈值不等，即此时第一返回子命令中的第一数据块个数小于第一阈值，第一返回子命令中的第一数据块和校验块的个数之和等于第一阈值；

[0043] 为了获取第一阈值个数的数据块，继续判断等待预设时长后的数据块个数；

[0044] 步骤S120，根据第一返回子命令和在预设时长内返回的第二返回子命令，记录预设时长内未返回的未返回子命令；

[0045] 具体地，因为刚开始下发的子命令是已知的，且在达到预设时长后，根据等待之前已经返回的第一返回子命令和等待之后返回的第二返回子命令，能够获知在等待之后仍然未返回的未返回子命令都是哪些；

[0046] 步骤S130，当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数与第一阈值不等时，通过第一返回子命令和第二返回子命令中的数据块和校验块恢复计算未返回子命令中的数据块，并将父命令需求数据返回客户端，其中，父命令需求数据为各个子命令中的数据块内容，第一阈值为父命令中数据块的个数。

[0047] 这里，等待之前已经返回的第一返回子命令中的数据块个数(第一数据块)和等待之后返回的第二返回子命令中的数据块个数(第二数据块)之和都没达到第一阈值时，可知晓，此时在未返回子命令中包含数据块，通过数据块和校验块的共同校验恢复计算，还原出未返回子命令中的数据块，此时，则获得了各个子命令的数据块内容(即父命令需求数据)，并将父命令需求数据返回客户端，完成分布式存储系统的响应，与传统的分布式存储系统需要等待所有子命令均返回(即返回全部数据块和校验块)，再将父命令需求数据返回客户端的响应方法相比，响应速度更快；

[0048] 其中，子命令为读命令；

[0049] 具体地，在分布式存储系统中，由于多副本或纠删码应用在整个存储集群上，每一个读命令都会拆分为多个子命令(根据副本或纠删码策略)，下发到不同存储节点处理，待

所有子命令返回后，父命令才对客户端返回结果。而分布式存储系统由分布于不同地理位置的多台存储设备组成，可以允许很多异构存储设备加入集群，即集群中的存储设备硬件之间可能存在很多差异，比如CPU/内存以及磁盘的类型(SAS/SATA)/容量等。这样，在使用过程中，集群中个别存储节点的个别磁盘响应时间可能比其它一些磁盘要长。因此，对于一个父命令，可能因为一个子命令返回较慢，而导致整个父命令整体返回也慢。在分布式存储系统中，这种慢速磁盘的现象要比在传统存储中发生的概率大很多。在监控视频应用中，这会产生非常大的影响，对于监控多路视频回放的场景(对实时响应要求非常高)，可能导致大量视频回放卡顿；

[0050] 此外，一个慢磁盘由于上层被虚拟化到很多不同的数据下发的单元中，这个慢磁盘的IO延时会很快成为集群性能的瓶颈，导致大量的应用IO需要等待最慢的IO返回才能下发完毕；针对响应慢的磁盘会严重影响整个集群的性能，本发明实施例对分布式存储系统的命令处理流程进行优化，提升读命令响应速度，降低IO延迟时间；

[0051] 进一步的，方法还包括：步骤S108，当第一返回子命令中的第一数据块个数与第一阈值相等时，将父命令需求数据返回客户端。

[0052] 这里，第一返回子命令和第一数据块个数相等，都等于第一阈值，说明，子命令中全部的数据块都已被获得，此时，将父命令需求数据返回客户端；

[0053] 进一步的，在步骤S130之前，还包括：

[0054] 步骤S126，判断第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值是否相等；

[0055] 这里，判断此时经等待预设时长后，两次获取的子命令中数据块数目之和是否等于第一阈值；

[0056] 步骤S128，当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值相等时，将父命令需求数据返回客户端。

[0057] 这里，当前后两次获取的子命令中数据块数目之和等于第一阈值，即此时获取了全部的数据块的数据，即得到了父命令需要的数据，能够获取全部数据块中的内容，再将父命令需要的数据返回客户端；

[0058] 进一步的，在上述实施例中，步骤S120中记录在预设时长内未返回的未返回子命令，还可采用以下步骤实现，包括：

[0059] 步骤S122，将超过预设时长仍未返回的未返回子命令进行标记记录，并记录未返回子命令所在的各个磁盘；

[0060] 步骤S124，分别统计磁盘中未返回子命令的个数，当个数超过第二阈值时，进行告警或踢出磁盘。

[0061] 这里，本发明实施例主要考虑到CPU性能远远高于慢速磁盘的响应速度。对于存在怠工磁盘或其它原因导致响应速度较慢的磁盘(不同型号/不同接口/不同容量/不同位置/寻道时间等原因)的分布式存储系统，可以大大缩短最大IO的延迟时间，提升读命令响应时间。且可以找到预设时长的配置大小，并通过预先配置的预设时长t，逐步找到性能较差的磁盘，通过逐步踢出这些磁盘提高整体性能；

[0062] 其中，上述磁盘包括虚拟磁盘和硬件磁盘；

[0063] 进一步的，方法还包括：步骤S140，当未返回子命令返回时，将未返回子命令回收。

[0064] 这里，当超出预设时长的未返回子命令返回时，将这些未返回子命令直接进行回收，即使这些未返回子命令中包括数据块也不返回至客户端；

[0065] 图2为本发明实施例提供的又一分布式存储响应方法流程图。

[0066] 参照图2，这里，N为数据块个数，M为校验块个数，在纠删码(N+M)保护策略应用中，针对慢速磁盘进行性能优化步骤如下：

[0067] 步骤S201，根据N+M测试构造N+M个读命令下发；

[0068] 步骤S202，判断第一返回子命令返回数目是否为N；

[0069] 步骤S203，若否，继续等待第一返回子命令；

[0070] 步骤S204，若是，判断第一返回子命令是否包括N个数据块；

[0071] 若包括，则跳转到步骤S209；

[0072] 步骤S205，若不包括，继续等待预设时长后，判断第一返回子命令和第二返回子命令是否等于N+M；

[0073] 若等于，则跳转到步骤S209；

[0074] 步骤S206，若不等，标记未返回子命令；

[0075] 步骤S207，判断第一返回子命令和第二返回子命令中是是否包括N个数据块；

[0076] 若是，则跳转到步骤S209；

[0077] 步骤S208，若不是，根据已返回子命令中的数据块和校验块计算出未返回子命令中数据块；

[0078] 步骤S209，将父命令需求数据返回客户端；

[0079] 这里，客户端就能获得读命令中N个数据块中的内容；

[0080] 其中，此处N为第一阈值；

[0081] 具体地，A.对于读命令，父命令会下发N+M个子命令，对于下发的各个读子命令，不必等待所有子命令均返回(正常需要返回N个数据块和M个校验块)，才对客户端返回父命令处理结果。而是在接收到最快返回的最小可能数目N个子命令后，判断获取的数据块数是否等于N(不包括校验块)，如果等于N则返回父命令需求数据给客户端；如果小于N，等待预设时长后(用户可以根据具体的应用配置预设时长的大小，如果CPU的性能很强，不考虑计算校验值的性能损耗，可以将预设时长配置为0)，再次判断获取的数据块数；

[0082] B.在上述步骤等待的预设时长内，如果第一返回子命令和第二返回子命令为N+M个，则返回N个数据块的内容给客户端；

[0083] C.如果经等待时长后，返回的子命令数目大于等于N且小于N+M个：

[0084] 1)如果返回子命令中的数据块数目等于N，则直接返回客户端N个数据块的内容(父命令需求数据)；

[0085] 2)如返回子命令中的数据块数目小于N，则通过已返回子命令中的数据块和校验块恢复出未返回子命令中数据块的内容，即可获取完整的N个数据块的内容，并对客户端返回N个完整的数据块内容(父命令需求数据)；

[0086] D.对于那些超过预设时长仍然没有返回的子命令，记录这些未返回子命令，记录下这些命令所在的磁盘，统计这些慢速盘的未返回子命令个数，如果慢速盘的未返回子命令个数超过第二阈值(用户可配置)，则需要对用户告警(或者配置直接踢盘)；

[0087] 这里，用户根据实际情况，来确定是否踢掉集群中这些慢速盘，以提高集群的整体

性能；

[0088] E.等这些慢速盘的未返回子命令返回时，直接回收；

[0089] 这里，对于多副本策略，通过上面通过设置预设时长来统计集群中较慢磁盘的策略同样适用，在经过预设时长后，对没收到的副本子命令进行记录，并在超过预设次数的情况下，发出告警或踢出相应磁盘；

[0090] 图3为本发明实施例提供的分布式存储响应系统功能框图。

[0091] 参照图3，分布式存储响应系统，应用于分布式存储系统，包括：

[0092] 等待模块，在第一返回子命令个数达到第一阈值的情况下，当第一返回子命令中的第一数据块个数与第一阈值不等时，等待预设时长；

[0093] 记录模块，根据第一返回子命令和在预设时长内返回的第二返回子命令，记录预设时长内未返回的未返回子命令；

[0094] 恢复计算模块，当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值不等时，通过第一返回子命令和第二返回子命令中的数据块和校验块恢复计算未返回子命令中的数据块，并将父命令需求数据返回客户端，其中，父命令为各个子命令中的数据块内容，第一阈值为父命令中数据块的个数。

[0095] 进一步的，系统还包括返回模块，当第一返回子命令中的第一数据块个数与第一阈值相等时，将父命令需求数据返回客户端。

[0096] 进一步的，还包括判断模块，判断第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值是否相等；当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值相等时，将父命令需求数据返回客户端。

[0097] 进一步的，记录模块还用于将超过预设时长仍未返回的未返回子命令进行标记记录，并记录未返回子命令所在的各个磁盘，分别统计磁盘中未返回子命令的个数，当个数超过第二阈值时，进行告警或踢出磁盘。

[0098] 进一步的，系统还包括回收模块，当未返回子命令返回时，将未返回子命令回收。

[0099] 本发明实施例提供了分布式存储响应方法和系统，应用于分布式存储系统，包括在第一返回子命令个数达到第一阈值的情况下，当第一返回子命令中的第一数据块个数与第一阈值不等时，等待预设时长；根据第一返回子命令和预设时长内返回的第二返回子命令，记录在预设时长内未返回的未返回子命令；当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值不等时，通过第一返回子命令和第二返回子命令中的数据块和校验块恢复计算未返回子命令中的数据块，并将父命令需求数据返回客户端，其中，父命令需求数据为各个子命令中的数据块内容，第一阈值为父命令中数据块的个数，极大程度地提升响应速度。

[0100] 本发明实施例提供的分布式存储响应系统，与上述实施例提供的分布式存储响应方法具有相同的技术特征，所以也能解决相同的技术问题，达到相同的技术效果。

[0101] 本发明实施例所提供的分布式存储响应方法和系统的计算机程序产品，包括存储了程序代码的计算机可读存储介质，所述程序代码包括的指令可用于执行前面方法实施例中所述的方法，具体实现可参见方法实施例，在此不再赘述。

[0102] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统

和装置的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0103] 另外，在本发明实施例的描述中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。

[0104] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0105] 在本发明的描述中，需要说明的是，术语“中心”、“上”、“下”、“左”、“右”、“竖直”、“水平”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”、“第二”、“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性。

[0106] 本发明实施例还提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，处理器执行计算机程序时实现上述实施例提供的分布式存储响应方法的步骤。

[0107] 本发明实施例还提供一种计算机可读存储介质，计算机可读存储介质上存储有计算机程序，计算机程序被处理器运行时执行上述实施例的分布式存储响应方法的步骤。

[0108] 最后应说明的是：以上所述实施例，仅为本发明的具体实施方式，用以说明本发明的技术方案，而非对其限制，本发明的保护范围并不局限于此，尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化，或者对其中部分技术特征进行等同替换；而这些修改、变化或者替换，并不使相应技术方案的本质脱离本发明实施例技术方案的精神和范围，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应所述以权利要求的保护范围为准。

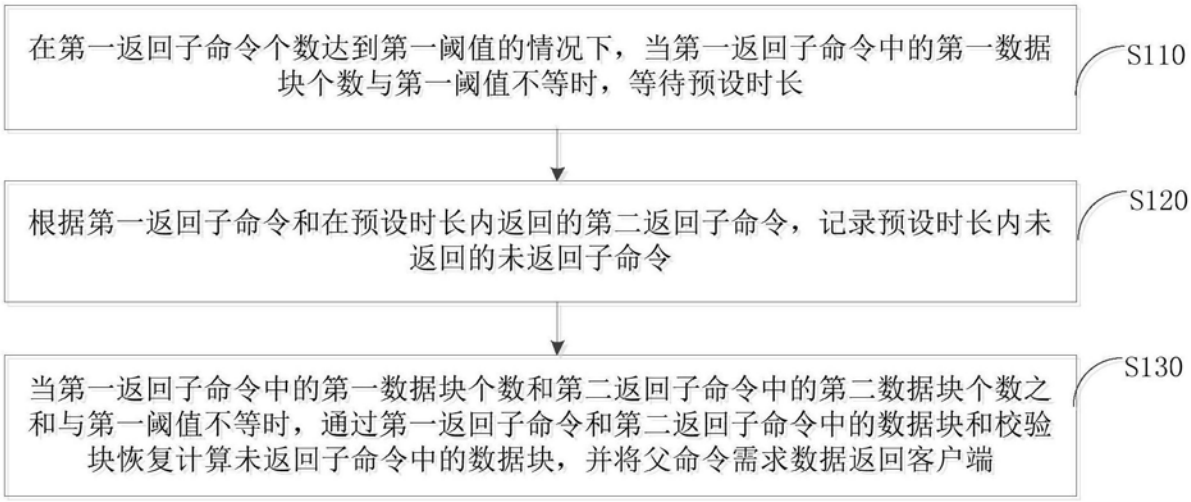

在第一返回子命令个数达到第一阈值的情况下，当第一返回子命令中的第一数据块个数与第一阈值不等时，等待预设时长
S110
根据第一返回子命令和在预设时长内返回的第二返回子命令，记录预设时长内未返回的未返回子命令
S120
当第一返回子命令中的第一数据块个数和第二返回子命令中的第二数据块个数之和与第一阈值不等时，通过第一返回子命令和第二返回子命令中的数据块和校验块恢复计算未返回子命令中的数据块，并将父命令需求数据返回客户端
S130

图1

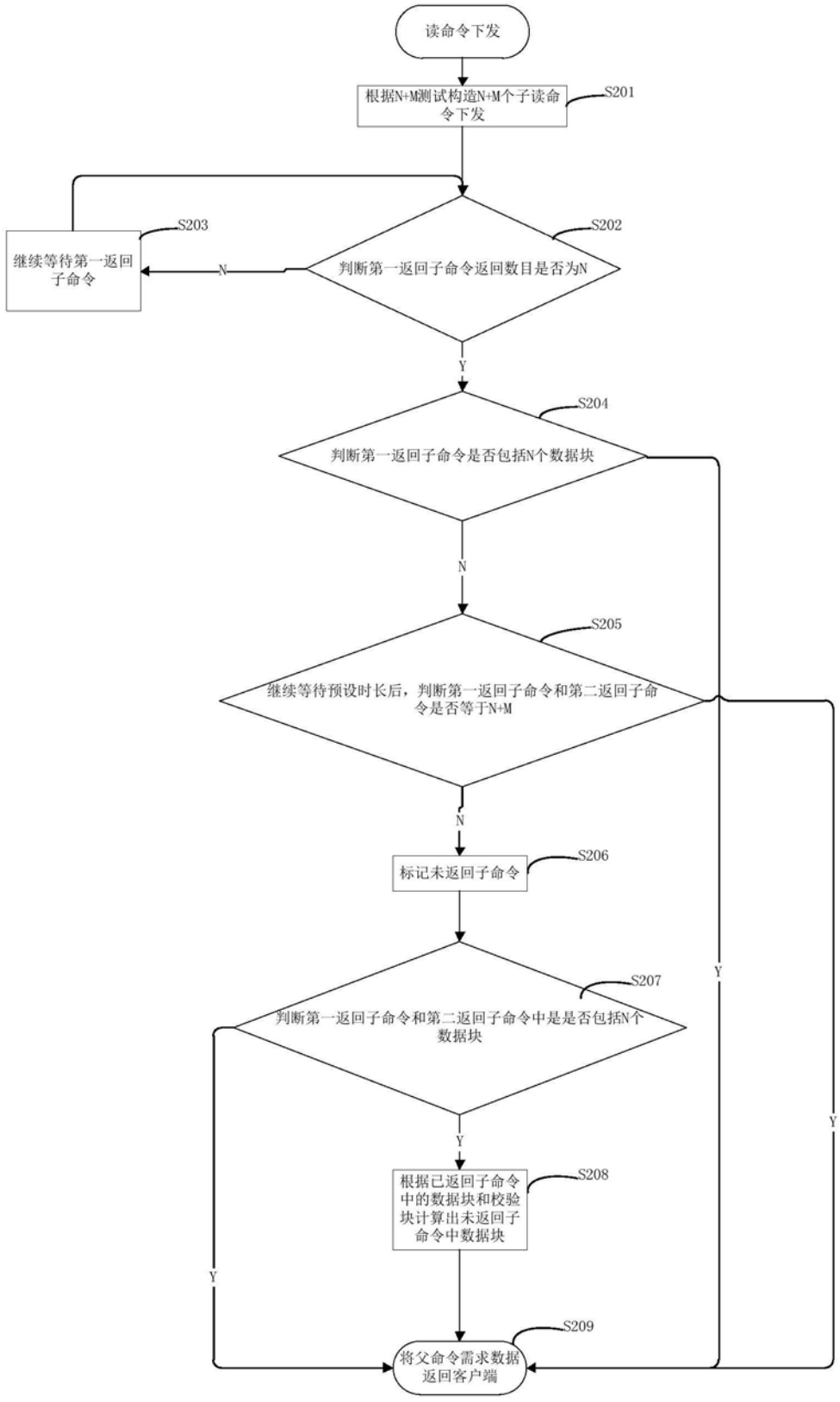

读命令下发
根据N+M测试构造N+M个子读命令下发
S201
判断第一返回子命令返回数目是否为N
S202
N
继续等待第一返回子命令
S203
Y
判断第一返回子命令是否包括N个数据块
S204
N
继续等待预设时长后，判断第一返回子命令和第二返回子命令是否等于N+M
S205
N
标记未返回子命令
S206
判断第一返回子命令和第二返回子命令中是是否包括N个数据块
S207
Y
Y
Y
根据已返回子命令中的数据块和校验块计算出未返回子命令中数据块
S208
Y
将父命令需求数据返回客户端
S209

图2

图3