(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 106649401 A
(43)申请公布日 2017.05.10

(21)申请号 201510740419.5

(22)申请日 2015.11.03

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 董乘宇 朱家稷 张海勇 曹锋 王勇 姚文辉 吴均平 吴洋 董元元 吴冬政 陆靖

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
代理人 苏培华

(51)Int.Cl.
G06F 17/30(2006.01)

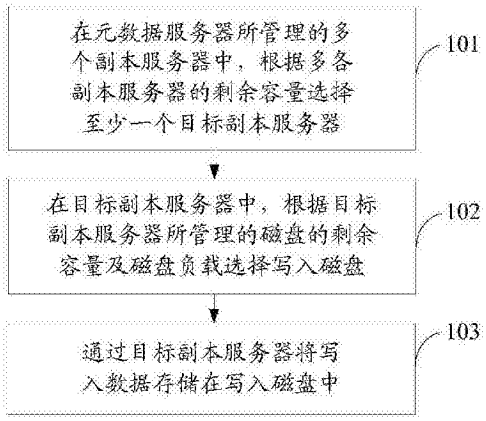
权利要求书2页 说明书9页 附图3页

(54)发明名称

分布式文件系统中的数据写入方法和装置

(57)摘要

本申请实施例提供了一种分布式文件系统中的数据写入方法和装置。该分布式文件系统中的数据写入方法包括:在元数据服务器所管理的多个副本服务器中,根据所述多个副本服务器的剩余容量选择目标副本服务器;在所述目标副本服务器中,根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘;通过所述目标副本服务器将写入数据存储在所述写入磁盘中。本申请实施例通过元数据服务器和副本服务器两级选盘方式,综合考虑磁盘剩余容量以及磁盘负载来确定写入数据的存储磁盘,既避免了磁盘写满,也不会造成磁盘的热点访问。



CN 106649401 A

1.一种分布式文件系统中的数据写入方法，其特征在于，包括：

在元数据服务器所管理的多个副本服务器中，根据所述多个副本服务器的剩余容量选择至少一个目标副本服务器；

在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘；

通过所述目标副本服务器将写入数据存储在所述写入磁盘中。

2.根据权利要求1所述的方法，其特征在于，在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘的步骤包括：

根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率；

根据所述各磁盘被选中的概率及预设条件确定所述写入磁盘。

3.根据权利要求2所述的方法，其特征在于，所述根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率，包括：

根据所述磁盘的剩余容量和总容量之间的比例及其权重，以及所述磁盘的负载及其权重，计算所述磁盘的得分；

对各所述磁盘的得分进行归一化，获得各所述磁盘被选中的概率。

4.根据权利要求3所述的方法，其特征在于，所述根据所述磁盘的剩余容量和总容量之间的比例及其权重，以及所述磁盘的负载及其权重，计算所述磁盘的得分，包括：

分别计算所述磁盘的剩余容量和总容量之间的比例的得分，以及所述磁盘的负载的得分，其中，所述比例的得分与所述比例之间以及所述负载的得分与所述负载之间均满足单调递增函数关系；

根据所述比例的得分及所述比例的权重，以及所述负载的得分及所述负载的权重，计算所述磁盘的得分。

5.根据权利要求3所述的方法，其特征在于，所述磁盘的负载为所述磁盘IO队列的长度。

6.根据权利要求2至5中任意一项所述的方法，其特征在于，所述根据所述各磁盘被选中的概率及预设条件确定所述写入磁盘，包括：

根据所述各磁盘被选中的概率计算所述各磁盘的累加概率值；

对所述各磁盘的累加概率值进行二分查找，将查找到的满足预设条件的磁盘的累加概率值对应的磁盘作为所述写入磁盘。

7.一种分布式文件系统中的数据写入装置，其特征在于，包括：

第一选择单元，被配置为在元数据服务器所管理的多个副本服务器中，根据所述多个副本服务器的剩余容量选择至少一个目标副本服务器；

第二选择单元，被配置为在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘；

存储单元，被配置为通过所述目标副本服务器将写入数据存储在所述写入磁盘中。

8.根据权利要求7所述的装置，其特征在于，所述第二选择单元包括：

概率计算子单元，被配置为根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率；

确定子单元，被配置为根据所述各磁盘被选中的概率及预设条件确定所述写入磁盘。

9.根据权利要求8所述的装置，其特征在于，所述概率计算子单元包括：

得分计算子单元，被配置为根据所述磁盘的剩余容量和总容量之间的比例及其权重，以及所述磁盘的负载及其权重，计算所述磁盘的得分；

概率获得子单元，被配置为对各所述磁盘的得分进行归一化，获得各所述磁盘被选中的概率。

10.根据权利要求9所述的装置，其特征在于，所述得分计算子单元包括：

第一计算子单元，被配置为分别计算所述磁盘的剩余容量和总容量之间的比例的得分，以及所述磁盘的负载的得分，其中，所述比例的得分与所述比例之间以及所述负载的得分与所述负载之间均满足单调递增函数关系；

第二计算子单元，被配置为根据所述比例的得分及所述比例的权重，以及所述负载的得分及所述负载的权重，计算所述磁盘的得分。

11.根据权利要求8所述的装置，其特征在于，所述磁盘的负载为所述磁盘IO队列的长度。

12.根据权利要求8至11中任意一项所述的装置，其特征在于，所述确定子单元包括：

累加子单元，被配置为根据所述各磁盘被选中的概率计算所述各磁盘的累加概率值；

查找子单元，被配置为对所述各磁盘的累加概率值进行二分查找，将查找到的满足预设条件的磁盘的累加概率值对应的磁盘作为所述写入磁盘。

分布式文件系统中的数据写入方法和装置

技术领域

[0001] 本申请涉及数据处理技术领域，特别是涉及一种分布式文件系统中的数据写入方法和一种分布式文件系统中的数据写入装置。

背景技术

[0002] 大型在线分布式存储系统中，单台服务器内有很多块磁盘，不同磁盘可能不同，且在一台服务器内部也可能出现异构的情况，即两种不同容量的磁盘混合。分布式存储系统解决数据存储的问题，当有数据到来时如何选择一块磁盘写入数据是一个需要考虑的问题。

[0003] 传统的分布式文件系统在选择磁盘时采用一致性哈希算法。该方法是事先根据数据分区做哈希来选定磁盘，是一种固定的选盘策略。哈希算法本身保证了平衡性，通常情况下不会造成热点，但是在异常情况，当同一数据分区有集中的数据，被选中的磁盘往往不能躲开大量的写入，可能会写满磁盘，造成写入失败，同时密集的流量会产生热点访问，造成写入时间变长。

[0004] 因此，目前需要本领域技术人员迫切解决的一个技术问题就是：如何可以避免磁盘写满，且不会造成热点访问。

发明内容

[0005] 本申请实施例的发明目的在于提供一种分布式文件系统中的数据写入方法，可以避免磁盘写满，且不会造成热点访问。

[0006] 相应的，本申请实施例还提供了一种分布式文件系统中的数据写入装置，用以保证上述方法的实现及应用。

[0007] 为了解决上述问题，本申请公开了一种分布式文件系统中的数据写入方法，包括：

[0008] 在元数据服务器所管理的多个副本服务器中，根据所述多个副本服务器的剩余容量选择至少一个目标副本服务器；

[0009] 在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘；

[0010] 通过所述目标副本服务器将写入数据存储在所述写入磁盘中。

[0011] 优选地，在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘的步骤包括：

[0012] 根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率；

[0013] 根据所述各磁盘被选中的概率及预设条件确定所述写入磁盘。

[0014] 优选地，所述根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率，包括：

[0015] 根据所述磁盘的剩余容量和总容量之间的比例及其权重，以及所述磁盘的负载及

其权重，计算所述磁盘的得分；

[0016] 对各所述磁盘的得分进行归一化，获得各所述磁盘被选中的概率。

[0017] 优选地，所述根据所述磁盘的剩余容量和总容量之间的比例及其权重，以及所述磁盘的负载及其权重，计算所述磁盘的得分，包括：

[0018] 分别计算所述磁盘的剩余容量和总容量之间的比例的得分，以及所述磁盘的负载的得分，其中，所述比例的得分与所述比例之间以及所述负载的得分与所述负载之间均满足单调递增函数关系；

[0019] 根据所述比例的得分及所述比例的权重，以及所述负载的得分及所述负载的权重，计算所述磁盘的得分。

[0020] 优选地，所述磁盘的负载为所述磁盘IO队列的长度。

[0021] 优选地，所述根据所述各磁盘被选中的概率及预设条件确定所述写入磁盘，包括：

[0022] 根据所述各磁盘被选中的概率计算所述各磁盘的累加概率值；

[0023] 对所述各磁盘的累加概率值进行二分查找，将查找到的满足预设条件的磁盘的累加概率值对应的磁盘作为所述写入磁盘。

[0024] 本申请还公开了一种分布式文件系统中的数据写入装置，包括：

[0025] 第一选择单元，被配置为在元数据服务器所管理的多个副本服务器中，根据所述多个副本服务器的剩余容量选择至少一个目标副本服务器；

[0026] 第二选择单元，被配置为在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘；

[0027] 存储单元，被配置为通过所述目标副本服务器将写入数据存储在所述写入磁盘中。

[0028] 优选地，所述第二选择单元包括：

[0029] 概率计算子单元，被配置为根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率；

[0030] 确定子单元，被配置为根据所述各磁盘被选中的概率及预设条件确定所述写入磁盘。

[0031] 优选地，所述概率计算子单元包括：

[0032] 得分计算子单元，被配置为根据所述磁盘的剩余容量和总容量之间的比例及其权重，以及所述磁盘的负载及其权重，计算所述磁盘的得分；

[0033] 概率获得子单元，被配置为对各所述磁盘的得分进行归一化，获得各所述磁盘被选中的概率。

[0034] 优选地，所述得分计算子单元包括：

[0035] 第一计算子单元，被配置为分别计算所述磁盘的剩余容量和总容量之间的比例的得分，以及所述磁盘的负载的得分，其中，所述比例的得分与所述比例之间以及所述负载的得分与所述负载之间均满足单调递增函数关系；

[0036] 第二计算子单元，被配置为根据所述比例的得分及所述比例的权重，以及所述负载的得分及所述负载的权重，计算所述磁盘的得分。

[0037] 优选地，所述磁盘的负载为所述磁盘IO队列的长度。

[0038] 优选地，所述确定子单元包括：

[0039] 累加子单元，被配置为根据所述各磁盘被选中的概率计算所述各磁盘的累加概率值；

[0040] 查找子单元，被配置为对所述各磁盘的累加概率值进行二分查找，将查找到的满足预设条件的磁盘的累加概率值对应的磁盘作为所述写入磁盘。

[0041] 与现有技术相比，本申请实施例包括以下优点：

[0042] 本申请实施例通过元数据服务器和副本服务器两级选盘方式，综合考虑磁盘剩余容量以及磁盘负载来确定写入数据的存储磁盘，既避免了磁盘写满，也不会造成磁盘的热点访问。当各磁盘负载基本相同时，可以优先选择剩余容量较高的磁盘作为写入磁盘，从而可以避免磁盘写满；当剩余容量基本相同时，可以优先选择磁盘负载较低的磁盘作为写入磁盘，从而可以避免造成磁盘的热点访问。

附图说明

[0043] 图1是本申请的一种分布式文件系统中的数据写入方法实施例的步骤流程图；

[0044] 图2是本申请的一种根据目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘方法实施例的步骤流程图；

[0045] 图3是本申请的一种计算各磁盘被选中的概率的方法实施例的步骤流程图；

[0046] 图4是本申请的一种计算磁盘得分的方法实施例的步骤流程图；

[0047] 图5是本申请的一种根据各磁盘被选中的概率及预设条件确定写入磁盘的方法实施例的步骤流程图；

[0048] 图6是本申请的一种分布式文件系统中的数据写入装置实施例的结构框图；

[0049] 图7是本申请的一种第二选择单元实施例的结构框图；

[0050] 图8是本申请的一种概率计算子单元实施例的结构框图；

[0051] 图9是本申请的一种得分计算子单元实施例的结构框图；

[0052] 图10是本申请的一种确定子单元实施例的结构框图。

具体实施方式

[0053] 为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

[0054] 参照图1，示出了本申请的一种分布式文件系统中的数据写入方法实施例的步骤流程图，具体可以包括如下步骤：

[0055] 步骤101，在元数据服务器所管理的多个副本服务器中，根据所述多个副本服务器的剩余容量选择至少一个目标副本服务器。

[0056] 本申请实施例中，首先在分布式文件系统中设置元数据服务器和副本服务器，其中，元数据服务器用于管理名字空间、数据块的副本信息和副本服务器地址信息；副本服务器用于管理数据块的本地副本，提供对所管理数据块副本的读写操作。

[0057] 本实施例采取两级选盘机制，该分布式文件系统中的数据写入装置(以下简称装置)首先在元数据服务器所管理的副本服务器中选择副本服务器来进行写入数据，具体可以选择剩余容量大的副本服务器，记为目标副本服务器。

[0058] 步骤102，在目标副本服务器中，根据目标副本服务器所管理的磁盘的剩余容量及

磁盘负载选择写入磁盘。

[0059] 本步骤中，在该装置确定出目标副本服务器后，进一步在该目标副本服务器中选择进行写入数据存储的磁盘，记为写入磁盘。该装置综合考虑了剩余容量和负载两方面来确定写入磁盘，具体的，该装置可以对各磁盘的剩余容量及磁盘负载进行打分，然后结合剩余容量和负载各自的权重获得该磁盘被选中的概率，进而根据该概率确定出写入磁盘。具体请参见后续实施例的描述。

[0060] 步骤103，通过目标副本服务器将写入数据存储在写入磁盘中。

[0061] 在该装置确定出写入磁盘后，即可通过该目标副本服务器将客户端写入的数据存储到上步骤确定的写入磁盘上了。

[0062] 本申请实施例通过元数据服务器和副本服务器两级选盘方式，综合考虑磁盘剩余容量以及磁盘负载来确定写入数据的存储磁盘，既避免了磁盘写满，也不会造成磁盘的热点访问。当各磁盘负载基本相同时，可以优先选择剩余容量较高的磁盘作为写入磁盘，从而可以避免磁盘写满；当剩余容量基本相同时，可以优先选择磁盘负载较低的磁盘作为写入磁盘，从而可以避免造成磁盘的热点访问。

[0063] 在本申请的另一实施例中，在目标副本服务器中，根据目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘的过程，如图2所示，可以包括：

[0064] 步骤201，根据目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率。

[0065] 本步骤中，首先获得目标副本服务器所管理的各磁盘的剩余容量及各磁盘的负载，然后根据剩余容量和负载计算各磁盘被选中的概率。具体计算方法有多种，以下仅为举例。

[0066] 其中一种计算各磁盘被选中的概率的方法，可以如图3所示，包括：

[0067] 步骤301，根据磁盘的剩余容量和总容量之间的比例及其权重，以及磁盘的负载及其权重，计算磁盘的得分。

[0068] 目标副本服务器中的各磁盘采用相同方法计算获得磁盘的得分。可以先根据比例及其权重计算比例的得分，根据负载及其权重计算负载的得分，然后综合两个得分获得磁盘的得分，或者综合两个得分并引入指定函数关系式获得磁盘的得分；还可以先计算比例的得分以及负载的得分，然后根据各自的权重计算获得磁盘的得分，其中也可以引入指定的函数关系式计算磁盘得分，具体请参见后续实施例的描述。

[0069] 其中，磁盘的负载为可以用磁盘IO队列的长度来衡量，磁盘IO队列长度越大表示该磁盘的负载越高。

[0070] 步骤302，对各磁盘的得分进行归一化，获得各磁盘被选中的概率。

[0071] 在获得各磁盘的得分后，对得分进行归一化，进而可获得各磁盘被选中的概率。

[0072] 步骤202，根据各磁盘被选中的概率及预设条件确定写入磁盘。

[0073] 在获得各磁盘被选中的概率后，即可将概率满足预设条件的磁盘确定为写入磁盘。该预设条件可以是用户自主设定等。

[0074] 在本申请的另一实施例中，一种计算磁盘得分的方法，如图4所示，具体可以包括：

[0075] 步骤401，分别计算磁盘的剩余容量和总容量之间的比例的得分，以及磁盘的负载的得分。

[0076] 其中，比例的得分与比例之间以及负载的得分与负载之间均满足单调递增函数关系，此处的单调递增函数的定义域从负无穷到正无穷，值域范围为从0到1，形状为S型，多存在于分类评定模型、逻辑回归模型，属于多重变数分析范畴，例如arctan函数等。本实施例中，磁盘的负载为用磁盘IO队列的长度来衡量。

[0077] 假设目标副本服务器中包含N块磁盘，各磁盘的剩余容量和N块磁盘的总容量之间的比例依次为R1，R2，……，RN。N块磁盘当前磁盘IO队列的长度(也即磁盘负载)依次为Q1，Q2，……，QN。

[0078] 则磁盘的剩余容量和总容量之间的比例的得分$S_i(R_i)$及磁盘的负载的得分$S_i(Q_i)$，i＝1，……，N，可以通过以下公式获得：

[0079] $$S_i(x_i)=\frac{1}{\pi}\left[\arctan(ax_i+b)+c\right]+\frac{1}{2}$$

[0080] $$a=\frac{6}{high-low}$$

[0081] $$b=\frac{-3(high-low)}{high+low}$$

[0082] c＝-arctan(a×bottom+b)

[0083] x可以是R或Q。也即比例的得分与比例之间以及负载的得分与负载之间均满足单调递增函数关系。该单调递增函数可以为arctan函数。

[0084] 其中，high，low和bottom是实验确定的值。x_i取值需要保证在(low，high)之间，因此ax_i+b保证了取值范围保证是在(-3，3)之间。bottom保证了当一个很低的x_i时，$P_i=0$。

[0085] 由于arctan函数是一个单调递增函数，其特点是，较高的得分表示磁盘剩余容量比例高或者负载较低，而反之则表示磁盘剩余容量比例低或者负载较高。

[0086] 步骤402，根据比例的得分及比例的权重，以及负载的得分及负载的权重，计算磁盘的得分。

[0087] 在获得各磁盘的剩余容量和总容量之间的比例的得分，以及各磁盘的负载的得分后，在本步骤中，即可结合磁盘剩余容量的权重及磁盘负载的权重计算各磁盘的得分了。

[0088] $S_i=S_i(R_i)^w\times S_i(Q_i)^{1-w}$

[0089] 其中，w是磁盘剩余容量对应的权重，1-w即为磁盘负载的权重，该权重值可以根据经验值等自主设定。

[0090] 在按照上述步骤401～402获得各磁盘的得分后，即可执行前述步骤302，对各磁盘的得分进行归一化，获得各磁盘被选中的概率。

[0091] 对N块磁盘的得分做归一化，用Si来表示第i块磁盘的得分，则：

[0092] $$\sum_{i=1}^{N}S_i=1$$

[0093] 因此，第i块磁盘被选中的概率Pi是，如下所示：

[0094] $$P_i=\frac{S_i}{\frac{1}{N}\sum_{i=1}^{N}S_i}$$

[0095] 在本申请的另一实施例中，根据各磁盘被选中的概率及预设条件确定写入磁盘的方法，如图5所示，可以包括：

[0096] 步骤501，根据各磁盘被选中的概率计算各磁盘的累加概率值。

[0097] 以第i块磁盘被选中的概率为Pi为例：

[0098] $$P_i = \frac{S_i}{\frac{1}{N}\sum_{i=1}^{N} S_i}$$

[0099] 所有磁盘被选中的概率累加等于1。

[0100] 定义各磁盘的累加概率值依次为A1，A2，……，AN，

[0101] $$A_i = \sum_{j=1}^{i} P_j \qquad i=1, \ \ldots\ldots N。$$

[0102] 因此A_i是从小到大排序的，取值范围是[0,1]，定义$A_0=0$。

[0103] 步骤502，对各磁盘的累加概率值进行二分查找，将查找到的满足预设条件的磁盘的累加概率值对应的磁盘作为写入磁盘。

[0104] 本实施例中，可以生成一个[0,1]随机数发生器，产生一个随机数r，该预设条件即为磁盘i需满足$A_{i-1}<r\leq A_i$。

[0105] 该装置对排序后的各磁盘的累加概率值进行二分查找，在查找到磁盘i满足$A_{i-1}<r\leq A_i$时，该磁盘i即确定为写入磁盘。其中，二分查找又称折半查找，优点是比较次数少，查找速度快，平均性能好，通过二分查找方式可以在排序后的累加概率值中很快找到满足预设条件的磁盘。

[0106] 在确定出写入磁盘后，即可执行前述步骤103进行写入数据的存储。

[0107] 需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请实施例并不受所描述的动作顺序的限制，因为依据本申请实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本申请实施例所必须的。

[0108] 参照图6，示出了本申请一种分布式文件系统中的数据写入装置实施例的结构框图，具体可以包括如下单元：

[0109] 第一选择单元601，被配置为在元数据服务器所管理的多个副本服务器中，根据多个副本服务器的剩余容量选择至少一个目标副本服务器。

[0110] 第二选择单元602，被配置为在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘。

[0111] 存储单元603，被配置为通过所述目标副本服务器将写入数据存储在所述写入磁盘中。

[0112] 本申请实施例通过上述单元综合考虑磁盘剩余容量以及磁盘负载来确定写入数据的存储磁盘，既避免了磁盘写满，也不会造成磁盘的热点访问。当各磁盘负载基本相同时，可以优先选择剩余容量较高的磁盘作为写入磁盘，从而可以避免磁盘写满；当剩余容量基本相同时，可以优先选择磁盘负载较低的磁盘作为写入磁盘，从而可以避免造成磁盘的热点访问。

[0113] 在另一实施例中，如图7所示，第二选择单元602可以进一步包括：

[0114] 概率计算子单元701，被配置为根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率。

[0115] 确定子单元702，被配置为根据所述各磁盘被选中的概率及预设条件确定所述写入磁盘。

[0116] 其中，如图8所示，概率计算子单元701又可以进一步包括：

[0117] 得分计算子单元801，被配置为根据所述磁盘的剩余容量和总容量之间的比例及其权重，以及所述磁盘的负载及其权重，计算所述磁盘的得分。

[0118] 概率获得子单元802，被配置为对各所述磁盘的得分进行归一化，获得各所述磁盘被选中的概率。

[0119] 如图9所示，得分计算子单元801又可以进一步包括：

[0120] 第一计算子单元901，被配置为分别计算所述磁盘的剩余容量和总容量之间的比例的得分，以及所述磁盘的负载的得分，其中，所述比例的得分与所述比例之间以及所述负载的得分与所述负载之间均满足单调递增函数关系。

[0121] 第二计算子单元902，被配置为根据所述比例的得分及所述比例的权重，以及所述负载的得分及所述负载的权重，计算所述磁盘的得分。

[0122] 其中，磁盘的负载为所述磁盘IO队列的长度。

[0123] 在另一实施例中，如图10所示，确定子单元702可以包括：

[0124] 累加子单元1001，被配置为根据所述各磁盘被选中的概率计算所述各磁盘的累加概率值。

[0125] 查找子单元1002，被配置为对所述各磁盘的累加概率值进行二分查找，将查找到的满足预设条件的磁盘的累加概率值对应的磁盘作为所述写入磁盘。

[0126] 本申请实施例还提供了一种电子设备，包括存储器和处理器。

[0127] 处理器与存储器通过总线相互连接；总线可以是ISA总线、PCI总线或EISA总线等。所述总线可以分为地址总线、数据总线、控制总线等。

[0128] 其中，存储器用于存储一段程序，具体地，程序可以包括程序代码，所述程序代码包括计算机操作指令。存储器可能包含高速RAM存储器，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。

[0129] 处理器用于读取存储器中的程序代码，执行以下步骤：

[0130] 在元数据服务器所管理的多个副本服务器中，根据多个副本服务器的剩余容量选择至少一个目标副本服务器；

[0131] 在所述目标副本服务器中，根据所述目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘；

[0132] 通过所述目标副本服务器将写入数据存储在所述存储器的所述写入磁盘中。

[0133] 对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0134] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0135] 本领域内的技术人员应明白，本申请实施例的实施例可提供为方法、装置、或计算

机程序产品。因此，本申请实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0136] 在一个典型的配置中，所述计算机设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非持续性的电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0137] 本申请实施例是参照根据本申请实施例的方法、终端设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0138] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0139] 这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上，使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0140] 尽管已描述了本申请实施例的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请实施例范围的所有变更和修改。

[0141] 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品

或者终端设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

[0142] 以上对本申请所提供的一种分布式文件系统中的数据写入方法和一种分布式文件系统中的数据写入装置，进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

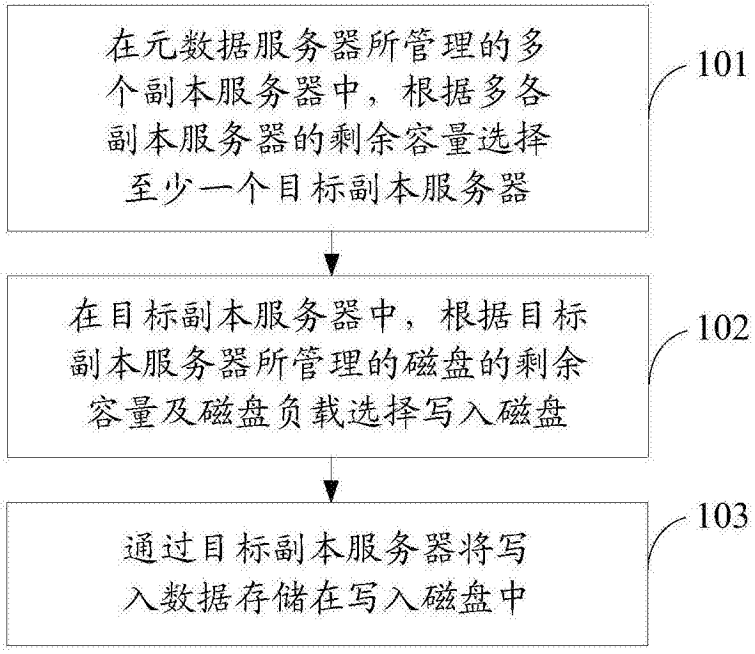

在元数据服务器所管理的多个副本服务器中，根据多各副本服务器的剩余容量选择至少一个目标副本服务器
101
在目标副本服务器中，根据目标副本服务器所管理的磁盘的剩余容量及磁盘负载选择写入磁盘
102
通过目标副本服务器将写入数据存储在写入磁盘中
103

图1

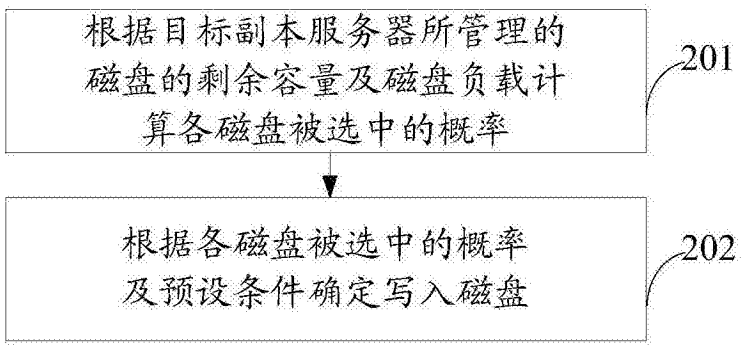

根据目标副本服务器所管理的磁盘的剩余容量及磁盘负载计算各磁盘被选中的概率
201
根据各磁盘被选中的概率及预设条件确定写入磁盘
202

图2

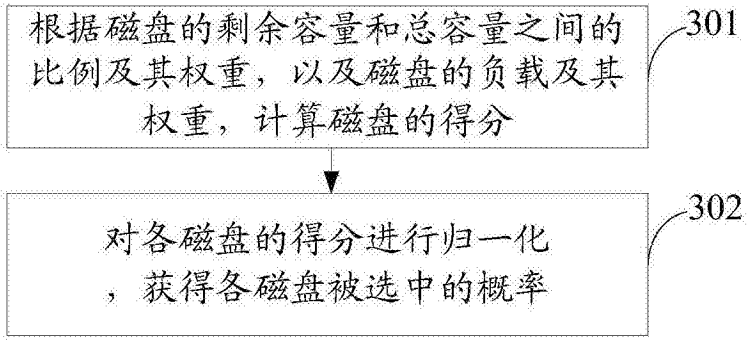

根据磁盘的剩余容量和总容量之间的比例及其权重，以及磁盘的负载及其权重，计算磁盘的得分
301
对各磁盘的得分进行归一化，获得各磁盘被选中的概率
302

图3

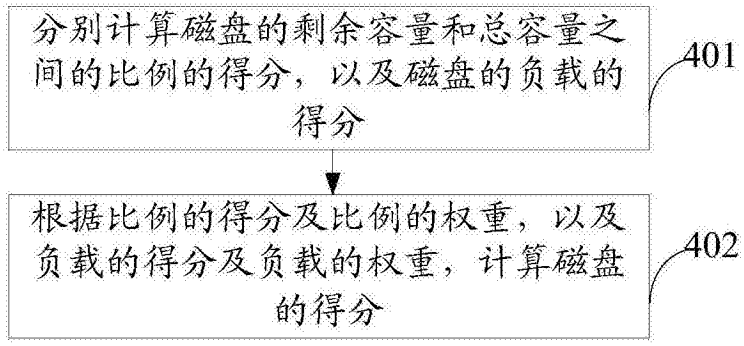

分别计算磁盘的剩余容量和总容量之间的比例的得分，以及磁盘的负载的得分
401
根据比例的得分及比例的权重，以及负载的得分及负载的权重，计算磁盘的得分
402

图4

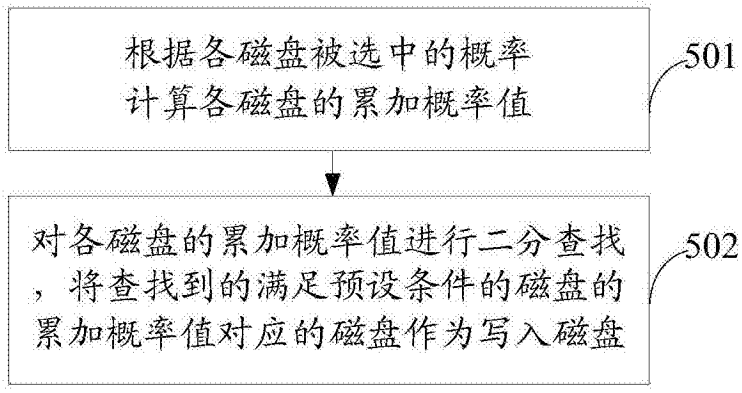

根据各磁盘被选中的概率计算各磁盘的累加概率值
501
对各磁盘的累加概率值进行二分查找，将查找到的满足预设条件的磁盘的累加概率值对应的磁盘作为写入磁盘
502

图5

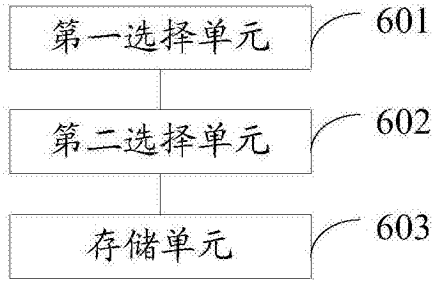

第一选择单元
601
第二选择单元
602
存储单元
603

图6

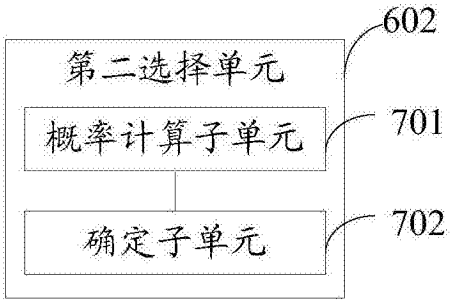

602
第二选择单元
概率计算子单元
701
确定子单元
702

图7

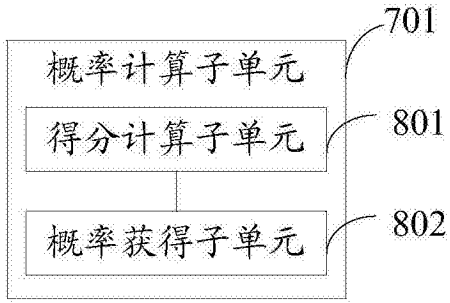

701
概率计算子单元
得分计算子单元
801
概率获得子单元
802

图8

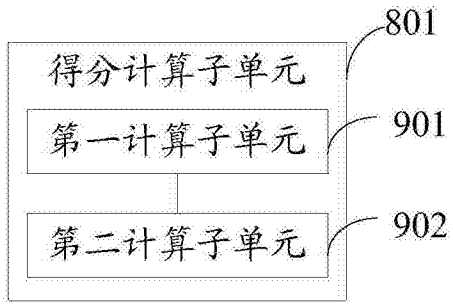

801
得分计算子单元
901
第一计算子单元
902
第二计算子单元

图9

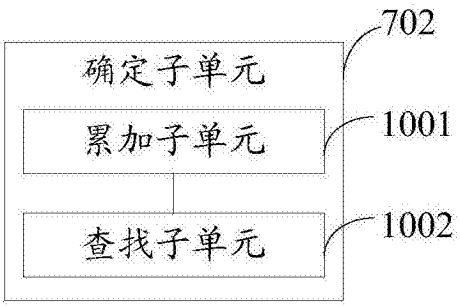

702
确定子单元
1001
累加子单元
1002
查找子单元

图10