(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 106933868 B
(45)授权公告日 2020.04.24

(21)申请号 201511024615.9

(22)申请日 2015.12.30

(65)同一申请的已公布的文献号
申请公布号 CN 106933868 A

(43)申请公布日 2017.07.07

(73)专利权人 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 张海勇 陆靖 姚文辉 董乘宇 朱家稷

(74)专利代理机构 北京睿博行远知识产权代理有限公司 11297
代理人 龚家骅

(51)Int.Cl.
G06F 16/13(2019.01)
G06F 3/06(2006.01)
H04L 29/08(2006.01)

(56)对比文件
CN 102414673 A,2012.04.11,
CN 104932956 A,2015.09.23,
CN 102150150 A,2011.08.10,
CN 103384272 A,2013.11.06,
CN 102609508 A,2012.07.25,
CN 103701916 A,2014.04.02,
CN 103984737 A,2014.08.13,
US 9020984 B1,2015.04.28,
US 8789050 B2,2014.07.22,
US 7383381 B1,2008.06.03,
US 2008155537 A1,2008.06.26,
Srikanth Kandula et al.The Nature of Datacenter Traffic:Measurement & Analysis.《Proceedings of the 9th IMC》.2009,第202-208页.
Jakub Konecny.Federated Optimization: Distributed Optimization Beyond the Datacenter.《Mathematics》.2015,第1-5页.
张鹏.数据中心网络的流量管理和优化问题研究.《中国博士学位论文全文数据库 信息科技辑》.2013,第2013年卷(第12期),第I137-1页.

审查员 张罗

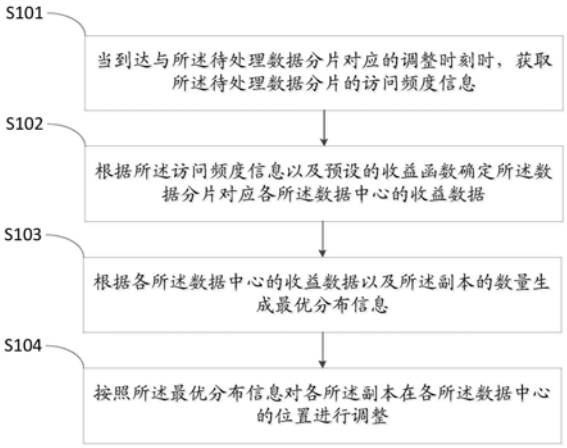
权利要求书3页 说明书10页 附图3页

(54)发明名称

一种调整数据分片分布的方法及数据服务器

(57)摘要

本发明公开了一种调整数据分片分布的方法。当到达与待处理数据分片对应的调整时刻时，获取待处理数据分片的访问频度信息，随后根据访问频度信息以及预设的收益函数确定数据分片对应各数据中心的收益数据，最后根据各数据中心的收益数据以及副本的数量生成最优分布信息，并按照最优分布信息对各副本在各数据中心的位置进行调整。从而在无需额外设置用于存储的内存或者硬盘的情况下，根据数据分片的访问频度和特性动态优化数据分片的分布情况，从而降低了数据中心之间的传输带宽需求。



CN 106933868 B

1.一种调整数据分片分布的方法，其特征在于，所述方法应用于包括多个数据中心的分布式文件存储系统中，待处理数据分片的多个副本存储于所述分布式文件存储系统中的一个或多个数据中心，该方法包括：

当到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息；

根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据；

根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息；

按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

2.如权利要求1所述的方法，其特征在于，当到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息，具体为：

在根据预设的时间周期确定当前时刻为调整时刻时，获取所述时间周期内各所述数据中心上报的子访问频度信息；

或，在接收到调整触发消息时，获取在预设的时间周期内各所述数据中心上报的子访问频度信息。

3.如权利要求1所述的方法，其特征在于，所述访问频度信息由所述待处理数据分片在各所述数据中心的子访问频度信息组成，所述子访问频度信息至少包括数据分片大小、从与所述子访问频度信息对应的数据中心访问所述数据分片的访问次数、所述数据分片从所述数据中心所产生的数据流量，以及平均跨机房带宽。

4.如权利要求3所述的方法，其特征在于，所述收益数据与所述访问次数、所述数据流量以及所述平均跨机房带宽成正比，以及与所述数据分片大小成反比。

5.如权利要求1所述的方法，其特征在于，根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息，具体为：

按照收益数据从大到小的顺序依次排列所述数据中心；

获取与所述副本的数量相同的排名之内的数据中心的标识，并将已获取的标识作为所述最优分布信息。

6.如权利要求5所述的方法，其特征在于，按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整，具体为：

获取所述数据分片的实时分布信息，所述实时分布信息由各所述副本当前所在的数据中心的标识组成；

判断所述实时分布信息是否与所述最优分布信息一致；

若所述实时分布信息与所述最优分布信息不一致，根据所述实时分布信息与所述最优分布信息中不相同的标识生成数据复制任务，以将各所述副本存储至与所述最优分布信息中的标识对应的数据中心。

7.如权利要求6所述的方法，其特征在于，还包括：

所述访问次数在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前加一；

所述数据流量在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前增加所述副本的数据量。

8.如权利要求7所述的方法,其特征在于,还包括:

当接收到用户通过所述客户端发送的数据写入请求时,获取所述数据写入请求中携带的待写入数据分片,并判断所述数据写入请求中是否还携带写入选项信息;

若所述数据写入请求携带跨数据中心分布的写入选项信息,按照所述用户指定的分布信息确定用于分配所述待写入数据分片的数据中心,并将确定结果返回至所述客户端,以使所述客户端按照所述确定结果写入所述待写入数据分片;

若所述数据写入请求携带默认写入选项信息或未携带任何写入选项信息,根据所述数据写入请求中携带的所述客户端所在的数据中心的标识确定用于分配所述待写入数据分片的数据中心,并将确定结果返回至所述客户端,以使所述客户端按照所述确定结果写入所述待写入数据分片。

9.如权利要求8所述的方法,其特征在于,还包括:

当接收到所述用户通过所述客户端发送的数据读取请求时,将所述数据读取请求对应的数据分片的分布信息返回至所述客户端,以使所述客户端根据所述分布信息选择与自身所在的数据中心对应的数据分片进行读取。

10.一种调整数据分片分布的设备,其特征在于,所述设备应用于包括多个数据中心的分布式文件存储系统中,待处理数据分片的多个副本存储于所述分布式文件存储系统中的一个或多个数据中心,该设备包括:

获取模块,在到达与所述待处理数据分片对应的调整时刻时,获取所述待处理数据分片的访问频度信息;

确定模块,根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据;

生成模块,根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息;

调整模块,按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

11.如权利要求10所述的设备,其特征在于,所述获取模块具体用于:

在根据预设的时间周期确定当前时刻为调整时刻时,获取所述时间周期内各所述数据中心上报的子访问频度信息;

或,在接收到调整触发消息时,获取在预设的时间周期内各所述数据中心上报的子访问频度信息。

12.如权利要求10所述的设备,其特征在于,所述访问频度信息由所述待处理数据分片在各所述数据中心的子访问频度信息组成,所述子访问频度信息至少包括数据分片大小、从与所述子访问频度信息对应的数据中心访问所述数据分片的访问次数、所述数据分片从所述数据中心所产生的数据流量,以及平均跨机房带宽。

13.如权利要求12所述的设备,其特征在于,所述收益数据与所述访问次数、所述数据流量以及所述平均跨机房带宽成正比,以及与所述数据分片大小成反比。

14.如权利要求10所述的设备,其特征在于,所述生成模块还包括:

排列子模块,按照收益数据从大到小的顺序依次排列所述数据中心;

处理子模块,获取与所述副本的数量相同的排名之内的数据中心的标识,并将已获取的标识作为所述最优分布信息。

15.如权利要求10所述的设备,其特征在于,所述生成模块还包括:

获取子模块，获取所述数据分片的实时分布信息，所述实时分布信息由各所述副本当前所在的数据中心的标识组成；

判断子模块，判断所述实时分布信息是否与所述最优分布信息一致，当所述实时分布信息与所述最优分布信息不一致，根据所述实时分布信息与所述最优分布信息中不相同的标识生成数据复制任务，以将各所述副本存储至与所述最优分布信息中的标识对应的数据中心。

16.如权利要求10所述的设备，其特征在于，还包括：

计数模块，在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前，所述访问次数加一；

计量模块，在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前，所述数据流量增加所述副本的数据量。

17.如权利要求16所述的设备，其特征在于，还包括：

写入模块，在接收到用户通过所述客户端发送的数据写入请求时，获取所述数据写入请求中携带的待写入数据分片，并判断所述数据写入请求中是否还携带写入选项信息；

在所述数据写入请求携带跨数据中心分布的写入选项信息，所述写入模块按照所述用户指定的分布信息确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片；

在所述数据写入请求携带默认写入选项信息或未携带任何写入选项信息，所述写入模块根据所述数据写入请求中携带的所述客户端所在的数据中心的标识确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片。

18.如权利要求17所述的设备，其特征在于，还包括：

读取模块，在接收到所述用户通过所述客户端发送的数据读取请求时，将所述数据读取请求对应的数据分片的分布信息返回至所述客户端，以使所述客户端根据所述分布信息选择与自身所在的数据中心对应的数据分片进行读取。

19.一种分布式文件存储系统，其特征在于，包括至少一个客户端，所述文件存储系统还包括：一个或多个数据中心，所述数据中心用于存储待处理数据分片的多个副本；

调整数据分片分布的设备，所述设备用于在到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息；根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据；根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息；按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

一种调整数据分片分布的方法及数据服务器

技术领域

[0001] 本发明涉及通信技术领域，特别涉及一种调整数据分片分布的方法。本发明同时还涉及一种数据服务器。

背景技术

[0002] 在云计算和大数据处理环境下，分布式文件系统作为底层的存储层，向上层的业务提供接近无限扩展的存储服务能力。然而，随着数据中心规模的增大以及全球部署的流行，数据中心因为各种物理问题(例如挖断光纤、机房起火等等)而发生整体下线的事故越来越多，一旦数据中心下线，将会导致严重的服务可用性问题。

[0003] 为了提高数据服务的可用性和延续性，业界一种常见的做法是将数据的多份拷贝分布到一定区域内的多个数据中心中，通过数据中心之间的数据冗余来提高服务的可用性，数据运营商通过采用跨数据中心数据分布的技术方案，从而使自己的部分服务能承受任意一个数据中心离线。

[0004] 当数据在多个数据中心(以下简称DataCenter或者DC)分布时，将会存在跨DC之间的数据读写过程，比如用户作业跨机房读写数据或者因为软硬件故障导致文件系统需要从内向外进行数据复制。这对跨DC的网络连接和带宽提出了较高的要求。现有的技术方案往往是数据运营商自建数据中心和网络，从而能够保证数据中心之间有充足的网络带宽。但是数据中心之间的线路需要租用，成本高昂，并不能保障有充足的带宽，

[0005] 针对以上问题，目前存在一种方案，就是在每个数据中心内部加上一层缓存来尽量避免跨数据中心的数据读取。这种方法虽然能够有效规避跨数据中心读数据所带来的网络流量。但由于缓存的数据放在内存中，相比磁盘内存的容量小上几个数量级(TB vs GB)，因此缓存的效果会随着数据量的增大而降低，而且缓存系统在将缓存放到硬盘上会挤占用户数据的可用IO能力，此外，缓存和底层的文件系统配合比较困难。如某个文件数据改写会导致整个文件的缓存数据失效，影响缓存的使用效率。

[0006] 由此可见，如何优化数据分布来节省网络访问的带宽，成为本领域技术人员亟待解决的技术问题。

发明内容

[0007] 本发明提供了一种调整数据分片分布的方法，用以在降低带宽需求的同时能最大化数据访问性能，该方法应用于包括多个数据中心的分布式文件存储系统中，待处理数据分片的多个副本存储于所述分布式文件存储系统中的一个或多个数据中心，该方法包括：

[0008] 当到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息；

[0009] 根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据；

[0010] 根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息；

[0011] 按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。优选地，当到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息，具体为：

[0012] 在根据预设的时间周期确定当前时刻为调整时刻时，获取所述时间周期内各所述数据中心上报的子访问频度信息；

[0013] 或，在接收到调整触发消息时，获取在预设的时间周期内各所述数据中心上报的子访问频度信息。

[0014] 优选地，所述访问频度信息由所述待处理数据分片在各所述数据中心的子访问频度信息组成，所述子访问频度信息至少包括数据分片大小、从与所述子访问频度信息对应的数据中心访问所述数据分片的访问次数、所述数据分片从所述数据中心所产生的数据流量，以及平均跨机房带宽。

[0015] 优选地，所述收益数据与所述访问次数、所述数据流量以及所述平均跨机房带宽成正比，以及与所述数据分片大小成反比。

[0016] 优选地，根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息，具体为：

[0017] 按照收益数据从大到小的顺序依次排列所述数据中心；

[0018] 获取与所述数量相同的排名之内的数据中心的标识，并将已获取的标识作为所述最优分布信息。

[0019] 优选地，按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整，具体为：

[0020] 获取所述数据分片的实时分布信息，所述原始分布信息由各所述副本当前所在的数据中心的标识组成；

[0021] 判断所述实时分布信息是否与所述最优分布信息一致；

[0022] 若所述实时分布信息与所述最优分布信息不一致，根据所述实时分布信息与所述最优分布信息中不相同的标识生成数据复制任务，以将各所述副本存储至与所述最优分布信息中的标识对应的数据中心。

[0023] 优选地，还包括：

[0024] 所述访问次数在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前加一；

[0025] 所述数据流量在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前增加所述副本的数据量。

[0026] 优选地，还包括：

[0027] 当接收到用户通过所述客户端发送的数据写入请求时，获取所述数据写入请求中携带的待写入数据分片，并判断所述数据写入请求中是否还携带写入选项信息；

[0028] 若所述数据写入请求携带跨数据中心分布的写入选项信息，按照所述用户指定的分布信息确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片；

[0029] 若所述数据写入请求携带默认写入选项信息或未携带任何写入选项信息，根据所述数据写入请求中携带的所述客户端所在的数据中心的标识确定用于分配所述待写入数

据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片。

[0030] 优选地，还包括：

[0031] 当接收到所述用户通过所述客户端发送的数据读取请求时，将所述数据读取请求对应的数据分片的分布信息返回至所述客户端，以使所述客户端根据所述分布信息选择与自身所在的数据中心对应的数据分片进行读取。

[0032] 相应地，本申请还提出一种调整数据分片分布的设备，其特征在于，所述设备应用于包括多个数据中心的分布式文件存储系统中，待处理数据分片的多个副本存储于所述分布式文件存储系统中的一个或多个数据中心，该设备包括：

[0033] 获取模块，在到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息；

[0034] 确定模块，根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据；

[0035] 生成模块，根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息；

[0036] 调整模块，按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

[0037] 优选地，所述获取模块具体用于：

[0038] 在根据预设的时间周期确定当前时刻为调整时刻时，获取所述时间周期内各所述数据中心上报的子访问频度信息；

[0039] 或，在接收到调整触发消息时，获取在预设的时间周期内各所述数据中心上报的子访问频度信息。

[0040] 优选地，所述访问频度信息由所述待处理数据分片在各所述数据中心的子访问频度信息组成，所述子访问频度信息至少包括数据分片大小、从与所述子访问频度信息对应的数据中心访问所述数据分片的访问次数、所述数据分片从所述数据中心所产生的数据流量，以及平均跨机房带宽。

[0041] 优选地，所述收益数据与所述访问次数、所述数据流量以及所述平均跨机房带宽成正比，以及与所述数据分片大小成反比。

[0042] 优选地，所述生成模块还包括：

[0043] 排列子模块，按照收益数据从大到小的顺序依次排列所述数据中心；

[0044] 处理子模块，获取与所述数量相同的排名之内的数据中心的标识，并将已获取的标识作为所述最优分布信息。

[0045] 优选地，所述生成模块还包括：

[0046] 获取子模块，获取所述数据分片的实时分布信息，所述原始分布信息由各所述副本当前所在的数据中心的标识组成；

[0047] 判断子模块，判断所述实时分布信息是否与所述最优分布信息一致，当所述实时分布信息与所述最优分布信息不一致，根据所述实时分布信息与所述最优分布信息中不相同的标识生成数据复制任务，以将各所述副本存储至与所述最优分布信息中的标识对应的数据中心。

[0048] 优选地，还包括：

[0049] 计数模块，在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前，所述访问次数加一；

[0050] 计量模块，在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前，所述数据流量增加所述副本的数据量。

[0051] 优选地，还包括：

[0052] 写入模块，在接收到用户通过所述客户端发送的数据写入请求时，获取所述数据写入请求中携带的待写入数据分片，并判断所述数据写入请求中是否还携带写入选项信息；

[0053] 在所述数据写入请求携带跨数据中心分布的写入选项信息，所述写入模块按照所述用户指定的分布信息确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片；

[0054] 在所述数据写入请求携带默认写入选项信息或未携带任何写入选项信息，所述写入模块根据所述数据写入请求中携带的所述客户端所在的数据中心的标识确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片。

[0055] 优选地，还包括：

[0056] 读取模块，在接收到所述用户通过所述客户端发送的数据读取请求时，将所述数据读取请求对应的数据分片的分布信息返回至所述客户端，以使所述客户端根据所述分布信息选择与自身所在的数据中心对应的数据分片进行读取。

[0057] 相应地，本申请另一方面还提出了一种分布式文件存储系统，其特征在于，包括至少一个客户端，所述文件存储系统还包括：

[0058] 一个或多个数据中心，所述数据中心用于存储待处理数据分片的多个副本；

[0059] 调整数据分片分布的设备，所述设备用于在到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息；根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据；根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息；按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

[0060] 由此可见，通过应用本发明的技术方案，当到达与待处理数据分片对应的调整时刻时，获取待处理数据分片的访问频度信息，随后根据访问频度信息以及预设的收益函数确定数据分片对应各数据中心的收益数据，最后根据各数据中心的收益数据以及副本的数量生成最优分布信息，并按照最优分布信息对各副本在各数据中心的位置进行调整。从而在无需额外设置用于存储的内存或者硬盘的情况下，根据数据分片的访问频度和特性动态优化数据分片的分布情况，从而降低了数据中心之间的传输带宽需求。

附图说明

[0061] 图1为本申请提出的一种调整数据分片分布的方法的流程示意图；

[0062] 图2为本申请实施例提出的一种分布式存储系统的结构示意图。

[0063] 图3为本申请提出的一种调整数据分片分布的设备的结构示意图。

具体实施方式

[0064] 如背景技术所述，若通过加大机房间带宽来规避跨机房流量的限制，其成本将会十分的高昂，而通过在每个机房内部设置缓存来规避跨机房流量的话，又会受到内存限制并且总体存储效率会降低。因此，本申请将机房中的数据划分为各个不同的数据分片，该数据分片也可以称为数据块，其作为一种数据的物理记录方式，是一组逻辑上连续排列在一起的记录，每个记录由多个副本构成，数据分片的副本是数据中心与输入、输出设备或其他数据中心之间进行传输的一个数据单位。基于对各个不同机房中的数据分片，通过动态调整数据分布来优化跨数据中心的网络流量，从而在降低带宽需求的同时能最大化数据访问性能。

[0065] 如图1所示，为本申请提出的一种调整数据分片分布的方法的流程示意图，由于本申请旨在根据用户的访问情况优化现有数据分片在多个数据中心中的分布，因此该方法应用于包括多个数据中心的分布式文件系统中，待处理数据分片的多个副本存储于所述分布式文件系统中的一个或多个数据中心(即多个副本可全部存储于多个数据中心中的一个数据中心，或者是分散存储于当前的数据中心)，具体地，包括以下步骤：

[0066] S101，当到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息。

[0067] 其中，所述访问频度信息由所述待处理数据分片在各所述数据中心的子访问频度信息组成，所述子访问频度信息至少包括数据分片大小、从与所述子访问频度信息对应的数据中心访问所述数据分片的访问次数、所述数据分片从所述数据中心所产生的数据流量，以及平均跨机房带宽。

[0068] 在本申请的优选实施例中，对于数据分片的调整可以设置系统自发进行，也可以由人工手动触发，对于系统自动触发的情况，在根据预设的时间周期确定当前时刻为调整时刻时，获取所述时间周期内各所述数据中心上报的子访问频度信息；而在由人工触发时，则是在接收到调整触发消息的情况下获取在预设的时间周期内各所述数据中心上报的子访问频度信息。

[0069] 由于本申请的技术方案需要针对各个数据中心判断数据分片的副本是否适合存储于该数据中心，因此访问频度信息由待处理数据分片在各数据中心的子访问频度信息组成。对于每个数据中心来说，其不仅要记录数据分片大小以及平均跨机房带宽，而且还要在采集每个周期中访问数据分片的访问次数以及数据分片从数据中心所产生的数据流量，从而为后续确定收益提供依据。

[0070] 需要说明的是，作为子访问频度信息中的变量，访问次数需要在数据中心将客户端所请求的数据分片对应的副本返回给客户端之前加一；而数据流量则需要在数据中心将客户端所请求的数据分片对应的副本返回给客户端之前增加副本的数据量。

[0071] 本申请在实现动态调整数据分布时，不可避免地会出现数据分片的写入和读取过程。因此，如果针对数据分片的读写过程进行优化，可以有效降低数据分片在读写时所产生的跨数据中心的流量。

[0072] 在数据分片写入过程中，默认处理方式是将所有数据分片放置到客户端所在的数据中心，可以规避数据分片写入时所产生的跨数据中心的流量。但是如果用户指定的写入选项是跨数据中心写入，则按照用户指定的写入选项来写入数据分片，此时可能会产生跨

数据中心的流量。故本申请实施方式中提出了数据分片写入过程的对应方法步骤，具体如下：

[0073] a)当接收到用户通过所述客户端发送的数据写入请求时，获取所述数据写入请求中携带的待写入数据分片，并判断所述数据写入请求中是否还携带写入选项信息；

[0074] b)若所述数据写入请求携带跨数据中心分布的写入选项信息，按照所述用户指定的分布信息确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片；

[0075] c)若所述数据写入请求携带默认写入选项信息或未携带任何写入选项信息，根据所述数据写入请求中携带的所述客户端所在的数据中心的标识确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片。

[0076] 在数据分片读取过程中，则只通过读取放置在客户端所在数据中心的数据分片，来规避数据分片读取时所产生的跨数据中心的流量。故本申请实施方式中提出了数据分片读取过程的对应方法步骤，具体如下：

[0077] a)当接收到所述用户通过所述客户端发送的数据读取请求时，将所述数据读取请求对应的数据分片的分布信息返回至所述客户端，以使所述客户端根据所述分布信息选择与自身所在的数据中心对应的数据分片进行读取。

[0078] 本申请实施方式中通过上述方式对数据分片的读写过程进行优化，使得数据分片读写时所产生的跨数据中心的流量大幅降低，进而降低了带宽需求。

[0079] 为便于清楚阐述本申请的技术方案，如图2所示，本申请的具体实施例提供了一种包括主服务器以及客户端的分布式存储系统，其中客户端由一个或多个客户端组件组成，而主服务器则包括数据服务器、元数据服务器以及数据分布管理组等多个模块，具体地，各个模块的介绍如下：

[0080] 客户端组件：客户端组件在打开文件时从主服务器获取到数据分片的位置信息，并选择距离最近的DC(同DC优先于外部DC)去访问；在写数据时，默认总是将所有数据分片放置到写者所在的数据中心，以规避写数据所产生的跨数据中心流量。

[0081] 数据服务器：管理文件的副本的数据和访问频度信息，提供对所管理副本的读写操作。通过周期性的汇报机制数据服务器将频度信息报告给元数据服务器。因数据是多副本，因此会有多个数据服务器汇报同一个数据分片的访问数据。

[0082] 元数据服务器：记录文件的数据分片位置信息，并汇总某个数据分片从任何一个DC访问的频次信息(在一定时间内累计，如1天内)。

[0083] 数据分布管理组件：在元数据服务器上的一个组件。该组件周期性的(如一天)计算所有数据分片的位置信息和访问频度信息，并计算数据重新分布的收益。收益超过一定权值则异步触发数据分片位置调整请求完成数据分布调整。

[0084] 基于上述各个模块，在数据分片的调整之前的数据写入流程以及数据读取流程如下：

[0085] (1)数据写入流程

[0086] 步骤a)客户端程序C在dc1中，收到写数据的请求。C向主服务器请求数据分片的位置，在请求中C会带上自身所处的数据中心名称dc1。默认情况下主服务器会将所有数据分

片分配到到dc1中。这样后继的写不会产生跨DC的流量。而用户若指定了写入选项是跨DC分布，则maste会尽量按照用户指定的分布来分配数据分片。这时后继的写有可能产生跨DC的流量。

[0087] 步骤b)客户端程序根据主服务器所分配的数据分片完成数据写入。

[0088] (2)数据读取流程

[0089] 步骤a)客户端程序C向主服务器申请打开文件f进行数据读取

[0090] 步骤b)主服务器将数据分片的位置返回给C

[0091] 步骤c)C优先选择同一数据中心的数据分片，直接连接对应的的数据服务器进行数据读取操作，在请求中C会将自身所在的DC名称dc1携带给数据服务器，并指定要访问和所读取的数据分片d。

[0092] 步骤d)数据服务器在返回数据给C之前记录对应数据分片d的访问频次+1，并且对应的访问数据量加上这次请求的数据

[0093] 步骤e)C得到数据服务器回吐的数据并返回给用户。

[0094] S102，根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据。

[0095] 其中，所述收益数据与所述访问次数、所述数据流量以及所述平均跨机房带宽成成正比，以及与所述数据分片大小成反比。

[0096] 在本申请的具体实施例中，收益函数的一个参考公式如下：

[0097] $$f(d,dc)=\frac{aA(d,dc)+bB(d,dc)}{2}*\frac{C}{S(d)};$$

[0098] 其中，d对应某个特定的数据分片，S(d)是该分片的大小(MB)。A(d,dc)是数据分片d从机房dc访问的次数，B(d,dc)是数据分片d从机房dc所产生的数据流量。C是跨机房带宽(MB)除上整个集群server的数量得到的一个平均每台CS所能拿到的带宽。

[0099] 通过上述收益函数，计算结果f(d,dc)就是数据分片d分部到dc中时所能带来的收益。在后续过程中即可利用收益函数对所有数据分片d进行计算，并且对所有dc(包括数据所在的dc)进行计算。对计算结果进行排序，确保收益最高的机房中存在数据分片，并且分片的数目同数据访问频度成正比。

[0100] 需要说明的是，以上公式仅为本申请具体实施例提出的一种优选方案，然而，在保证收益数据与所述访问次数、所述数据流量以及所述平均跨机房带宽成正比，以及与所述数据分片大小成反比的前提下，本领域技术人员也可以对该收益函数进行修改或者变形，这些都属于本申请的保护范围。

[0101] S103，根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息。

[0102] 为了能够使数据分片的副本能够均匀分布到各个数据中，同时兼顾各个数据中心的收益数据，在本申请的优选实施例中，首先按照收益数据从大到小的顺序依次排列所述数据中心，获取与所述数量相同的排名之内的数据中心的标识，并将已获取的标识作为所述最优分布信息。

[0103] S104，按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

[0104] 具体地，在本申请的优选实施例中，首先获取所述数据分片的实时分布信息，需要

说明的是，该原始分布信息由各所述副本当前所在的数据中心的标识组成，因此后面可判断所述实时分布信息是否与所述最优分布信息一致，若所述实时分布信息与所述最优分布信息不一致，根据所述实时分布信息与所述最优分布信息中不相同的标识生成数据复制任务，以将各所述副本存储至与所述最优分布信息中的标识对应的数据中心。

[0105] 以S101中的分布式存储系统为例，该具体实施例中的数据分片调整过程如下：

[0106] 步骤a)数据服务器定期将数据分片的访问频度信息汇报给主服务器

[0107] 步骤b)主服务器根据数据分片综合汇聚该信息

[0108] 步骤c)主服务器的数据分布管理组件定期(每天)或者在系统管理员的手动触发下重新计算所有数据分片的分布收益函数f(d,dc1)，计算过程参考上文中的公式。根据计算结果将数据分片的分布数据更新。如对于数据分片d，在计算前的分布时(dc1,dc1,dc1)，计算后调整为(dc1,dc2,dc2)

[0109] 步骤d)数据分布管理模块在后台扫描数据分布情况，如果发现当前数据的分布和理想分布(在步骤C中调整后的)不一致，则发起低优先级的数据复制任务，重新组织数据的布局。

[0110] 步骤e)Client(客户端)在后继读取中优先访问本机房的数据。

[0111] 为达到以上技术目的，本申请还提出一种调整数据分片分布的设备，所述设备应用于包括多个数据中心的分布式文件系统中，待处理数据分片的多个副本存储于所述分布式文件系统中的一个或多个数据中心，该设备包括：

[0112] 获取模块310，在到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息；

[0113] 确定模块320，根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据；

[0114] 生成模块330，根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息；

[0115] 调整模块340，按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

[0116] 在具体应用场景中，所述获取模块具体用于：

[0117] 在根据预设的时间周期确定当前时刻为调整时刻时，获取所述时间周期内各所述数据中心上报的子访问频度信息；

[0118] 或，在接收到调整触发消息时，获取在预设的时间周期内各所述数据中心上报的子访问频度信息。

[0119] 在具体应用场景中，所述访问频度信息由所述待处理数据分片在各所述数据中心的子访问频度信息组成，所述子访问频度信息至少包括数据分片大小、从与所述子访问频度信息对应的数据中心访问所述数据分片的访问次数、所述数据分片从所述数据中心所产生的数据流量，以及平均跨机房带宽。

[0120] 在具体应用场景中，所述收益数据与所述访问次数、所述数据流量以及所述平均跨机房带宽成正比，以及与所述数据分片大小成反比。

[0121] 在具体应用场景中，所述生成模块还包括：

[0122] 排列子模块，按照收益数据从大到小的顺序依次排列所述数据中心；

[0123] 处理子模块，获取与所述数量相同的排名之内的数据中心的标识，并将已获取的标识作为所述最优分布信息。

[0124] 在具体应用场景中，所述生成模块还包括：

[0125] 获取子模块，获取所述数据分片的实时分布信息，所述原始分布信息由各所述副本当前所在的数据中心的标识组成；

[0126] 判断子模块，判断所述实时分布信息是否与所述最优分布信息一致，当所述实时分布信息与所述最优分布信息不一致，根据所述实时分布信息与所述最优分布信息中不相同的标识生成数据复制任务，以将各所述副本存储至与所述最优分布信息中的标识对应的数据中心。

[0127] 在具体应用场景中，还包括：

[0128] 计数模块，在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前，所述访问次数加一；

[0129] 计量模块，在所述数据中心将客户端所请求的数据分片对应的副本返回给所述客户端之前，所述数据流量增加所述副本的数据量。

[0130] 在具体应用场景中，还包括：

[0131] 写入模块，在接收到用户通过所述客户端发送的数据写入请求时，获取所述数据写入请求中携带的待写入数据分片，并判断所述数据写入请求中是否还携带写入选项信息；

[0132] 在所述数据写入请求携带跨数据中心分布的写入选项信息，所述写入模块按照所述用户指定的分布信息确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片；

[0133] 在所述数据写入请求携带默认写入选项信息或未携带任何写入选项信息，所述写入模块根据所述数据写入请求中携带的所述客户端所在的数据中心的标识确定用于分配所述待写入数据分片的数据中心，并将确定结果返回至所述客户端，以使所述客户端按照所述确定结果写入所述待写入数据分片。

[0134] 在具体应用场景中，还包括：

[0135] 读取模块，在接收到所述用户通过所述客户端发送的数据读取请求时，将所述数据读取请求对应的数据分片的分布信息返回至所述客户端，以使所述客户端根据所述分布信息选择与自身所在的数据中心对应的数据分片进行读取。

[0136] 另一方面，本申请另一方面还提出了一种分布式文件存储系统，其特征在于，包括至少一个客户端，所述文件存储系统还包括：

[0137] 一个或多个数据中心，所述数据中心用于存储待处理数据分片的多个副本；

[0138] 调整数据分片分布的设备，所述设备用于在到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息；根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据；根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息；按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整。

[0139] 在应用以上方案进行数据分片存储之后，数据服务器中所保存的频度信息只需要在内存中保存，如果数据服务器因为任何原因crash则对应的数据清零。异常宕机是小概率

事件，在整个集群环境下所带来的访问频度不准确问题影响不大。并且能随着下一个周期的数据分片分布调整自动恢复合理布局。

[0140] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到本发明可以通过硬件实现，也可以借助软件加必要的通用硬件平台的方式来实现。基于这样的理解，本发明的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)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场景所述的方法。

[0141] 本领域技术人员可以理解附图只是一个优选实施场景的示意图，附图中的模块或流程并不一定是实施本发明所必须的。

[0142] 本领域技术人员可以理解实施场景中的装置中的模块可以按照实施场景描述进行分布于实施场景的装置中，也可以进行相应变化位于不同于本实施场景的一个或多个装置中。上述实施场景的模块可以合并为一个模块，也可以进一步拆分成多个子模块。

[0143] 上述本发明序号仅仅为了描述，不代表实施场景的优劣。

[0144] 以上公开的仅为本发明的几个具体实施场景，但是，本发明并非局限于此，任何本领域的技术人员能思之的变化都应落入本发明的保护范围。

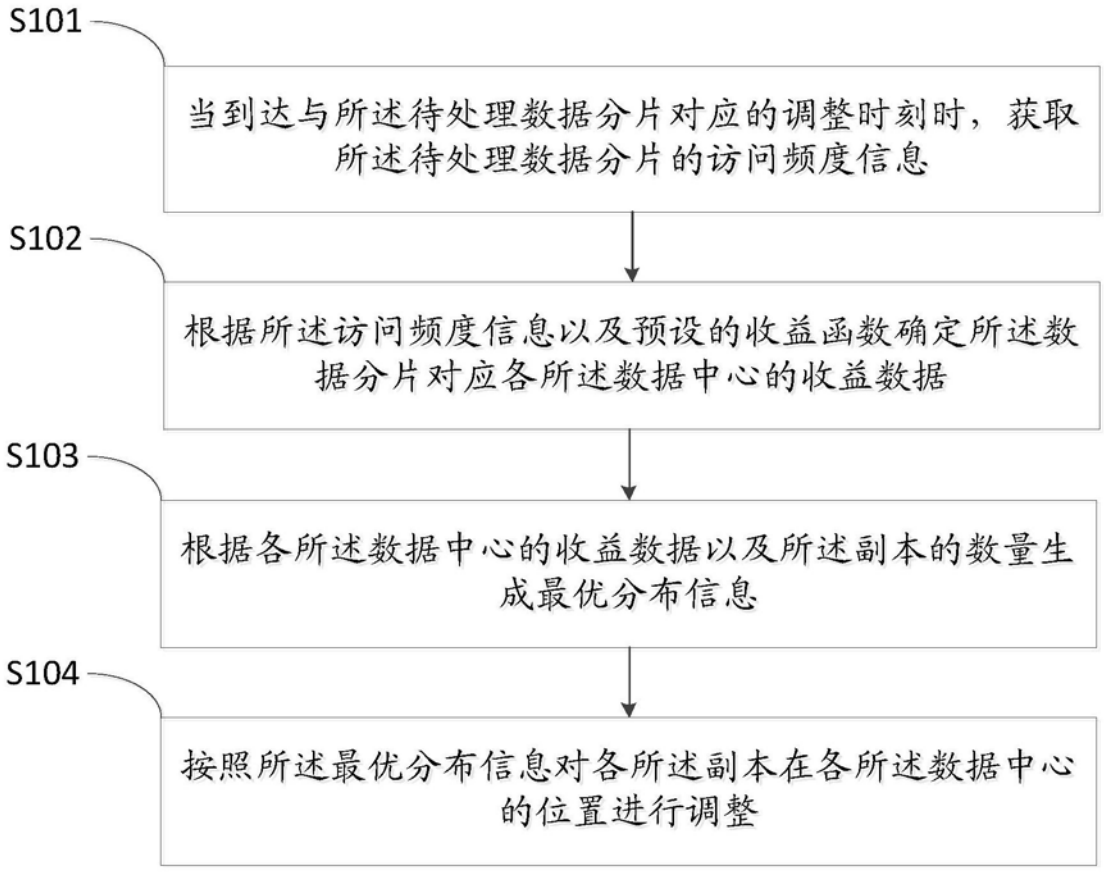

S101
当到达与所述待处理数据分片对应的调整时刻时，获取所述待处理数据分片的访问频度信息
S102
根据所述访问频度信息以及预设的收益函数确定所述数据分片对应各所述数据中心的收益数据
S103
根据各所述数据中心的收益数据以及所述副本的数量生成最优分布信息
S104
按照所述最优分布信息对各所述副本在各所述数据中心的位置进行调整

图1

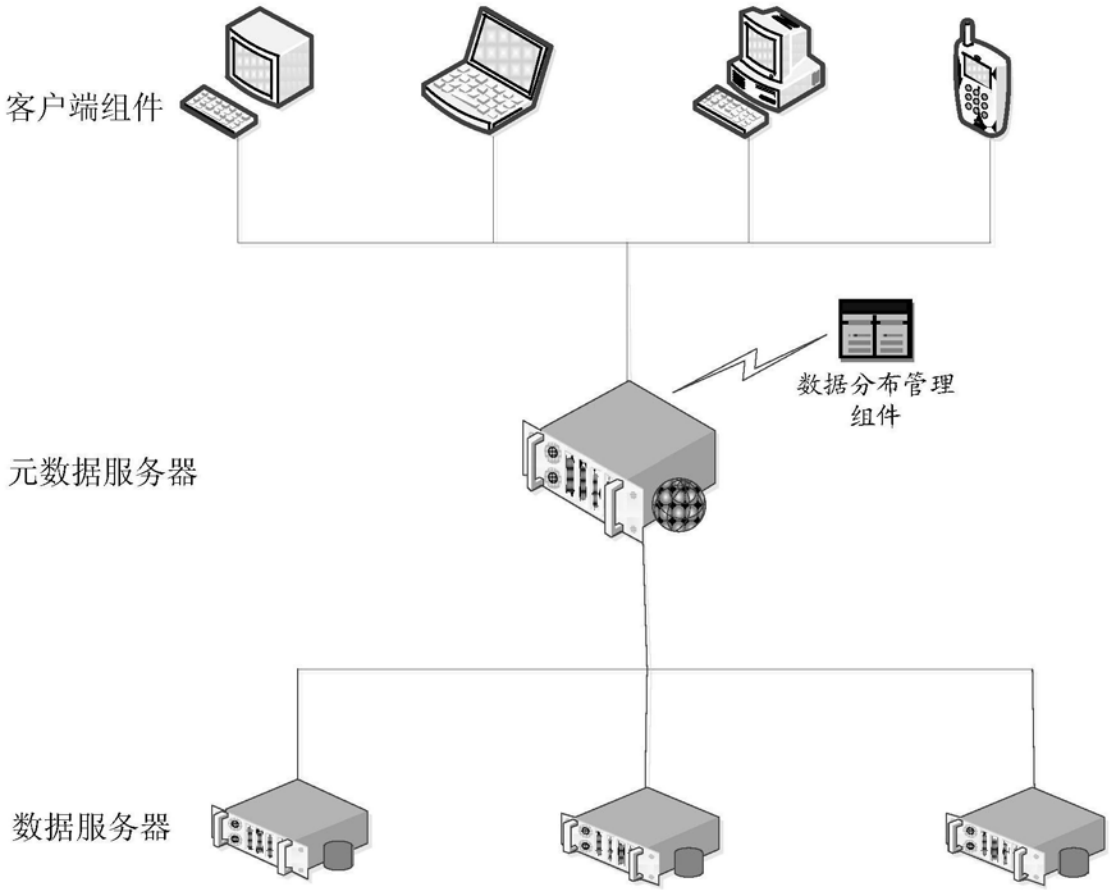


图2

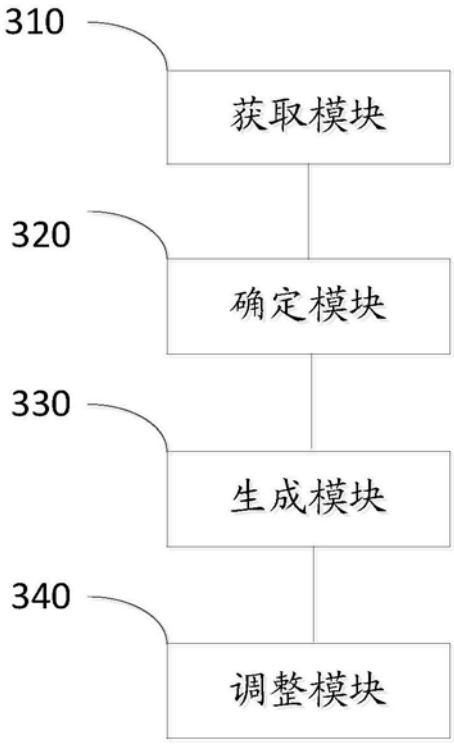


图3