



(12) 发明专利申请

(10) 申请公布号 CN 104205780 A

(43) 申请公布日 2014. 12. 10

(21) 申请号 201480000338. 5

(51) Int. Cl.

(22) 申请日 2014. 01. 23

H04L 29/08 (2006. 01)

(85) PCT国际申请进入国家阶段日

2014. 06. 16

(86) PCT国际申请的申请数据

PCT/CN2014/071224 2014. 01. 23

(71) 申请人 华为技术有限公司

地址 518129 广东省深圳市龙岗区坂田华为
总部办公楼

(72) 发明人 瑞列丹

(74) 专利代理机构 北京中博世达专利商标代理
有限公司 11274

代理人 申健

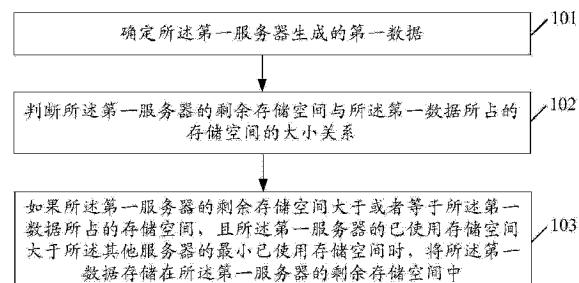
权利要求书4页 说明书20页 附图6页

(54) 发明名称

一种存储数据的方法和装置

(57) 摘要

本发明实施例公开了一种存储数据的方法和装置，涉及存储技术领域，用以减少读取数据时的时延，从而提高系统性能。本发明实施例提供的方法，应用于分布式存储系统，所述分布式存储系统包括第一服务器和其他服务器，所述方法包括：确定所述第一服务器生成的第一数据；判断所述第一服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系；如果所述第一服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间，且所述第一服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时，将所述第一数据存储在所述第一服务器的剩余存储空间中。



1. 一种存储数据的方法,其特征在于,应用于分布式存储系统,所述分布式存储系统包括第一服务器和其他服务器,所述方法包括:

确定所述第一服务器生成的第一数据;

判断所述第一服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系;

如果所述第一服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述第一服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述第一服务器的剩余存储空间中。

2. 根据权利要求 1 所述的方法,其特征在于,所述第一服务器包含一虚拟机,所述第一数据是所述虚拟机生成的数据;所述其他服务器包括第二服务器;所述方法还包括:

如果所述第一服务器的剩余存储空间小于所述第一数据所占的存储空间,则向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时,将所述虚拟机和所述其他数据迁移到所述第二服务器上。

3. 根据权利要求 1 所述的方法,其特征在于,所述其他服务器包括第二服务器;所述方法还包括:

如果所述第一服务器的剩余存储空间小于所述第一数据所占的存储空间,则向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

更新所述第一服务器的剩余存储空间;

当所述第一服务器更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时,向所述第二服务器发送第二指示消息;所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述第一服务器;

接收所述第二服务器发送的所述第一数据;

将所述第一数据存储在所述第一服务器更新后的剩余存储空间中。

4. 根据权利要求 1-3 任一项所述的方法,其特征在于,所述其他服务器还包括第三服务器,所述方法还包括:

向所述第三服务器发送包含所述第一数据的副本的第三指示消息,所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

5. 一种存储数据的方法,其特征在于,应用于分布式存储系统,所述分布式存储系统包括第一服务器和其他服务器,所述方法包括:

确定所述第一服务器生成的第一数据;所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值;

判断所述第一服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系;

如果所述第一服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述第一服务器的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述第一服务器的剩余存储空间中。

6. 一种服务器，其特征在于，应用于分布式存储系统，所述分布式存储系统还包括其他服务器，所述服务器包括：

确定模块，用于确定所述服务器生成的第一数据；

判断模块，用于判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系；

存储模块，用于在所述判断模块确定所述服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间，且所述服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时，将所述第一数据存储在所述服务器的剩余存储空间中。

7. 根据权利要求 6 所述的服务器，其特征在于，所述服务器包含一虚拟机，所述第一数据是所述虚拟机生成的数据；所述其他服务器包括第二服务器，所述服务器还包括：

发送模块，用于在所述判断模块确定所述服务器的剩余存储空间小于所述第一数据所占的存储空间时，向所述第二服务器发送包含所述第一数据的第一指示消息；所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中；

迁移模块，用于在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时，将所述虚拟机和所述其他数据迁移到所述第二服务器上。

8. 根据权利要求 6 所述的服务器，其特征在于，所述其他服务器包括第二服务器，所述服务器还包括：

发送模块，用于在所述判断模块确定所述服务器的剩余存储空间小于所述第一数据所占的存储空间时，向所述第二服务器发送包含所述第一数据的第一指示消息；所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中；

更新模块，用于更新所述服务器的剩余存储空间；

所述发送模块还用于，当所述服务器更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时，向所述第二服务器发送第二指示消息；所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述服务器；

接收模块，用于接收所述第二服务器发送的所述第一数据；

所述存储模块还用于，将所述第一数据存储在所述服务器更新后的剩余存储空间中。

9. 根据权利要求 6-8 任一项所述的服务器，其特征在于，所述其他服务器还包括第三服务器；

所述发送模块还用于，向所述第三服务器发送包含所述第一数据的副本的第三指示消息，所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

10. 一种服务器，其特征在于，应用于分布式存储系统，所述分布式存储系统还包括其他服务器，所述服务器包括：

处理器，用于确定所述服务器生成的第一数据；以及判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系；

存储器，用于在所述处理器确定所述服务器的剩余存储空间大于或者等于所述第一数

据所占的存储空间,且所述服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时,在所述处理器的控制下将所述第一数据存储在所述服务器的剩余存储空间中。

11. 根据权利要求 10 所述的服务器,其特征在于,所述服务器包含一虚拟机,所述第一数据是所述虚拟机生成的数据;所述其他服务器包括第二服务器,所述服务器还包括:

发送器,用于在所述处理器确定所述服务器的剩余存储空间小于所述第一数据所占的存储空间时,向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

所述处理器还用于,在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时,将所述虚拟机和所述其他数据迁移至所述第二服务器上。

12. 根据权利要求 10 所述的服务器,其特征在于,所述其他服务器包括第二服务器;

所述处理器还用于,更新所述服务器的剩余存储空间;

所述发送器还用于,当所述服务器更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时,向所述第二服务器发送第二指示消息;所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述服务器;

所述服务器还包括:接收器,用于接收所述第二服务器发送的所述第一数据;

所述存储器还用于,在所述处理器的控制下将所述第一数据存储在所述服务器更新后的剩余存储空间中。

13. 根据权利要求 10-12 任一项所述的服务器,其特征在于,所述其他服务器还包括第三服务器;

所述发送器还用于,向所述第三服务器发送包含所述第一数据的副本的第三指示消息,所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

14. 一种服务器,其特征在于,应用于分布式存储系统,所述分布式存储系统还包括其他服务器,所述服务器包括:

确定模块,用于确定所述服务器生成的第一数据;所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值;

判断模块,用于判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系;

存储模块,用于在所述判断模块确定所述服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述服务器的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述服务器的剩余存储空间中。

15. 一种服务器,其特征在于,应用于分布式存储系统,所述分布式存储系统还包括其他服务器,所述服务器包括:

处理器,用于确定所述服务器生成的第一数据;以及判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系;所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值;

存储器,用于在所述处理器确定所述服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述服务器的已使用存储空间小于或者等于所述其他服务器的最小

已使用存储空间时，在所述处理器的控制下将所述第一数据存储在所述服务器的剩余存储空间中。

一种存储数据的方法和装置

技术领域

[0001] 本发明涉及存储技术领域，尤其涉及一种存储数据的方法和装置。

背景技术

[0002] 分布式存储系统以可靠性高、可用性强、易于扩展等特点成为研究的热点。分布式存储系统由多个服务器构成，每个服务器上可以安装一个或者多个虚拟机，每个服务器对应 RAID (Redundant Arrays of Independent Disks, 磁盘阵列) 上的一部分存储空间。

[0003] 目前，在分布式存储系统中，一般按照“负载均衡”的原则为服务器上安装的应用 / 虚拟机生成的数据分配物理存储空间，即，按照“HDFS 中的各服务器的已使用存储空间均衡”的原则为服务器上安装的应用 / 虚拟机生成的数据分配物理存储空间；其中，物理存储空间为某一服务器对应的磁盘阵列的存储空间和 / 或该服务器本身的存储空间。具体的，以分布式存储系统为 HDFS (Hadoop Distributed File System, 分布式文件系统) 为例进行说明：

[0004] 在 HDFS 中，最小存储单元为块 (block)，每个块的最大值为 64M。当服务器上安装的应用 / 虚拟机生成的数据所占的存储空间大于 64M 时，该服务器首先对该数据进行分块，一般地，至多有 1 个块所占的存储空间小于 64M，其余块所占的存储空间均等于 64M；其次按照“HDFS 中的各服务器的已使用存储空间均衡”的原则，将各块映射到 HDFS 中的一个或者多个服务器的存储空间上。当服务器上安装的应用 / 虚拟机生成的数据所占的存储空间等于或者小于 64M 时，按照“HDFS 中的各服务器的已使用存储空间均衡”的原则，将该数据映射到 HDFS 中、已使用存储空间最小的服务器的存储空间上。

[0005] 由于上述方案中按照“负载均衡”的原则为服务器上安装的应用 / 虚拟机生成的数据分配物理位置，因此，该数据中的部分 / 全部数据可能存储在分布式存储系统中的其他服务器的存储空间上。这样，该服务器往往需要从其他服务器的存储空间上读取该数据，从而造成读取数据时延大的问题，进而导致系统性能较差。

发明内容

[0006] 本发明实施例提供一种存储数据的方法和装置，用以减少读取数据时的时延，从而提高系统性能。

[0007] 为达到上述目的，本发明的实施例采用如下技术方案：

[0008] 第一方面，提供一种存储数据的方法，应用于分布式存储系统，所述分布式存储系统包括第一服务器和其他服务器，所述方法包括：

[0009] 确定所述第一服务器生成的第一数据；

[0010] 判断所述第一服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系；

[0011] 如果所述第一服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间，且所述第一服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时，

将所述第一数据存储在所述第一服务器的剩余存储空间中。

[0012] 结合第一方面,在第一种可能的实现方式中,所述第一服务器包含一虚拟机,所述第一数据是所述虚拟机生成的数据;所述其他服务器包括第二服务器;所述方法还包括:

[0013] 如果所述第一服务器的剩余存储空间小于所述第一数据所占的存储空间,则向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

[0014] 在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时,将所述虚拟机和所述其他数据迁移到所述第二服务器上。

[0015] 结合第一方面,在第二种可能的实现方式中,所述其他服务器包括第二服务器;所述方法还包括:

[0016] 如果所述第一服务器的剩余存储空间小于所述第一数据所占的存储空间,则向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

[0017] 更新所述第一服务器的剩余存储空间;

[0018] 当所述第一服务器更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时,向所述第二服务器发送第二指示消息;所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述第一服务器;

[0019] 接收所述第二服务器发送的所述第一数据;

[0020] 将所述第一数据存储在所述第一服务器更新后的剩余存储空间中。

[0021] 结合第一方面、第一方面的第一种可能的实现方式或者第二种可能的实现方式任一种,在第三种可能的实现方式中,所述其他服务器还包括第三服务器,所述方法还包括:

[0022] 向所述第三服务器发送包含所述第一数据的副本的第三指示消息,所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

[0023] 第二方面,提供一种存储数据的方法,应用于分布式存储系统,所述分布式存储系统包括第一服务器和其他服务器,所述方法包括:

[0024] 确定所述第一服务器生成的第一数据;所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值;

[0025] 判断所述第一服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系;

[0026] 如果所述第一服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述第一服务器的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述第一服务器的剩余存储空间中。

[0027] 第三方面,提供一种服务器,应用于分布式存储系统,所述分布式存储系统还包括其他服务器,所述服务器包括:

[0028] 确定模块,用于确定所述服务器生成的第一数据;

[0029] 判断模块,用于判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系;

[0030] 存储模块,用于在所述判断模块确定所述服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述服务器的剩余存储空间中。

[0031] 结合第三方面,在第一种可能的实现方式中,所述服务器包含一虚拟机,所述第一数据是所述虚拟机生成的数据;所述其他服务器包括第二服务器,所述服务器还包括:

[0032] 发送模块,用于在所述判断模块确定所述服务器的剩余存储空间小于所述第一数据所占的存储空间时,向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

[0033] 迁移模块,用于在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时,将所述虚拟机和所述其他数据迁移至所述第二服务器上。

[0034] 结合第三方面,在第二种可能的实现方式中,所述其他服务器包括第二服务器,所述服务器还包括:

[0035] 发送模块,用于在所述判断模块确定所述服务器的剩余存储空间小于所述第一数据所占的存储空间时,向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

[0036] 更新模块,用于更新所述服务器的剩余存储空间;

[0037] 所述发送模块还用于,当所述服务器更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时,向所述第二服务器发送第二指示消息;所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述服务器;

[0038] 接收模块,用于接收所述第二服务器发送的所述第一数据;

[0039] 所述存储模块还用于,将所述第一数据存储在所述服务器更新后的剩余存储空间中。

[0040] 结合第三方面、第三方面的第一种可能的实现方式或者第三方面的第二种可能的实现方式任一种,在第三种可能的实现方式中,所述其他服务器还包括第三服务器;

[0041] 所述发送模块还用于,向所述第三服务器发送包含所述第一数据的副本的第三指示消息,所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

[0042] 第四方面,提供一种服务器,应用于分布式存储系统,所述分布式存储系统还包括其他服务器,所述服务器包括:

[0043] 处理器,用于确定所述服务器生成的第一数据;以及判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系;

[0044] 存储器,用于在所述处理器确定所述服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时,在所述处理器的控制下将所述第一数据存储在所述服务器的剩余存储空间中。

[0045] 结合第四方面,在第一种可能的实现方式中,所述服务器包含一虚拟机,所述第一

数据是所述虚拟机生成的数据；所述其他服务器包括第二服务器，所述服务器还包括：

[0046] 发送器，用于在所述处理器确定所述服务器的剩余存储空间小于所述第一数据所占的存储空间时，向所述第二服务器发送包含所述第一数据的第一指示消息；所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中；

[0047] 所述处理器还用于，在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时，将所述虚拟机和所述其他数据迁移到所述第二服务器上。

[0048] 结合第四方面，在第二种可能的实现方式中，所述其他服务器包括第二服务器；

[0049] 所述处理器还用于，更新所述服务器的剩余存储空间；

[0050] 所述发送器还用于，当所述服务器更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时，向所述第二服务器发送第二指示消息；所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述服务器；

[0051] 所述服务器还包括：接收器，用于接收所述第二服务器发送的所述第一数据；

[0052] 所述存储器还用于，在所述处理器的控制下将所述第一数据存储在所述服务器更新后的剩余存储空间中。

[0053] 结合第四方面、第四方面的第一种可能的实现方式或者第四方面的第二种可能的实现方式任一种，在第三种可能的实现方式中，所述其他服务器还包括第三服务器；

[0054] 所述发送器还用于，向所述第三服务器发送包含所述第一数据的副本的第三指示消息，所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

[0055] 第五方面，提供一种服务器，应用于分布式存储系统，所述分布式存储系统还包括其他服务器，所述服务器包括：

[0056] 确定模块，用于确定所述服务器生成的第一数据；所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值；

[0057] 判断模块，用于判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系；

[0058] 存储模块，用于在所述判断模块确定所述服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间，且所述服务器的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时，将所述第一数据存储在所述服务器的剩余存储空间中。

[0059] 第六方面，提供一种服务器，应用于分布式存储系统，所述分布式存储系统还包括其他服务器，所述服务器包括：

[0060] 处理器，用于确定所述服务器生成的第一数据；以及判断所述服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系；所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值；

[0061] 存储器，用于在所述处理器确定所述服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间，且所述服务器的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时，在所述处理器的控制下将所述第一数据存储在所述服务器的剩余存储空间中。

[0062] 上述技术方案,应用于分布式存储系统,在一服务器确定其剩余存储空间大于或者等于其生成的一数据所占的存储空间时,优先将该数据存储在该服务器的剩余存储空间中。这样,该服务器可以直接从本地读取该数据,而不需要从网络中的其他服务器上读取该数据,从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中,因通过在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

附图说明

[0063] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例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,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

- [0064] 图 1 为本发明实施例一提供的一种存储数据的方法的流程图;
- [0065] 图 2 为本发明实施例二提供的一种存储数据的方法的流程图;
- [0066] 图 3 为本发明实施例 1 提供的一种存储数据的方法的流程图;
- [0067] 图 4 为本发明实施例 2 提供的一种存储数据的方法的流程图;
- [0068] 图 5 为本发明实施例三提供的一种服务器的结构示意图;
- [0069] 图 6 为本发明实施例三提供的另一种服务器的结构示意图;
- [0070] 图 7 为本发明实施例四提供的一种服务器的结构示意图;
- [0071] 图 8 为本发明实施例四提供的另一种服务器的结构示意图;
- [0072] 图 9 为本发明实施例五提供的一种服务器的结构示意图;
- [0073] 图 10 为本发明实施例六提供的一种服务器的结构示意图。

具体实施方式

[0074] 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。

[0075] 本文中术语“系统”和“网络”在本文中常被可互换使用。本文中术语“和 / 或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A 和 / 或 B,可以表示:单独存在 A,同时存在 A 和 B,单独存在 B 这三种情况。本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0076] 实施例一

[0077] 如图 1 所示,为本实施例提供的一种存储数据的方法,应用于分布式存储系统,所述分布式存储系统包括第一服务器和其他服务器,所述方法包括:

[0078] 101:确定所述第一服务器生成的第一数据。

[0079] 其中,“第一服务器”可以为分布式存储系统中的任一服务器;“其他服务器”是指该分布式存储系统中、除第一服务器之外的所有服务器。“第一数据”可以为第一服务器生成的任一数据。本实施例的执行主体可以为“第一服务器”。

[0080] 分布式存储系统中的服务器可以安装一个或者多个虚拟机,也可以不安装虚拟

机。当服务器上安装有虚拟机时,该服务器与其上安装的虚拟机共享该服务器对应的磁盘阵列。另外,当服务器上安装有一个或者多个虚拟机时,该服务器生成的第一数据可以为:该服务器(物理机)上安装的一应用生成的数据;也可以为该服务器上安装的一虚拟机生成的数据。

[0081] 进一步地,当一数据为服务器上安装的一应用生成的数据时,该服务器会记录该数据与该服务器之间的对应关系;当一数据为服务器上安装的一虚拟机生成的数据时,该服务器会记录该数据与该虚拟机之间的对应关系。具体可以通过生成数据路由表来记录该对应关系。一般地,分布式存储系统中的每个服务器均可共享该数据路由表。在本文中,当生成一数据的应用/虚拟机所在的服务器与存储该数据的存储空间所对应的服务器为同一服务器时,认为实现了对该数据的本地化。

[0082] 102:判断所述第一服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系。

[0083] 其中,“服务器的剩余存储空间”是指该服务器对应的磁盘阵列和/或该服务器本身的存储空间中未存储数据的、可用存储空间。剩余存储空间的大小可以为0,也可以包括一个或者多个存储单元。每个服务器的剩余存储空间可能因删除已存储的数据、增加待存储的数据、改变该服务器对应的磁盘阵列的大小等而更新。服务器可以包括但不限于按照以下几种方式获取其剩余存储空间:周期性获取其剩余存储空间,定期获取其剩余存储空间,生成一数据时获取其剩余存储空间等。

[0084] 本实施例对“第一数据所占的存储空间”的大小不进行限定。

[0085] 在步骤102之前,该方法还可以包括:获取所述第一数据所占的存储空间,以及所述第一服务器的剩余存储空间。

[0086] 103:如果所述第一服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述第一服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述第一服务器的剩余存储空间中。

[0087] 其中,步骤103具体可以包括:所述第一服务器在确定其剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述第一服务器的已使用存储空间大于所述其他服务器的最小已使用存储空间时,为所述第一数据分配存储地址,并将所述第一数据存储在所述存储地址对应的存储单元中;所述存储地址为所述第一服务器的剩余存储空间中的一个或者多个存储单元对应的存储地址。

[0088] 示例性的,通过一个具体示例对“其他服务器的最小已使用存储空间”进行说明:假设其他服务器由服务器1、服务器2、服务器3构成,该3个服务器的已使用存储空间的大小分别为:A、B、C,且A>B>C;那么,其他服务器的最小已使用存储空间为C。

[0089] 进一步地,为了与现有技术方案进行详细对比,下面从第一数据所占的存储空间与分布式存储系统的最小存储单元的最大值的大小关系的角度对步骤103进行说明。具体的,在包含特征“第一服务器的剩余存储空间大于或者等于第一数据所占的存储空间”的情况下,步骤103可以包括以下场景1和场景2:

[0090] 场景1:第一数据所占的存储空间大于分布式存储系统的最小存储单元的最大值,且第一服务器的已使用存储空间大于其他服务器的最小已使用存储空间。

[0091] 场景2:第一数据所占的存储空间小于或者等于分布式存储系统的最小存储单元

的最大值,且第一服务器的已使用存储空间大于其他服务器的最小已使用存储空间。

[0092] 其中,针对场景 1,在现有技术方案中,第一服务器需要对第一数据进行分块,并按照“负载均衡”的原则,分别将每个块存储在分布式存储系统的多个服务器的剩余存储空间中。针对场景 2,在现有技术方案中,第一服务器需要将第一数据存储在其他服务器中的最大剩余存储空间中。

[0093] 需要说明的是,实际实现时,还可能出现以下场景 A 和场景 B:

[0094] 场景 A:第一数据所占的存储空间大于分布式存储系统的最小存储单元的最大值,且第一服务器的已使用存储空间小于或者等于其他服务器的最小已使用存储空间。

[0095] 场景 B:第一数据所占的存储空间小于或者等于分布式存储系统的最小存储单元的最大值,且第一服务器的已使用存储空间小于或者等于其他服务器的最小已使用存储空间。

[0096] 其中,本文中提供了针对场景 A 的实现方法,在下述实施例二中有相应描述。针对场景 B,与现有技术方案的实现方法相同,具体的,第一服务器将第一数据存储在其剩余存储空间中。

[0097] 综述,由上述场景 1、场景 2、场景 A 和场景 B 可知,实际实现时,本发明实施例提供的存储数据的方法不需要限定第一数据所占的存储空间的大小,也不需要限定第一服务器的已使用存储空间与其他服务器的已使用存储空间的大小关系。具体实现方式包括但不限于下述实施例 1 和实施例 2 所示的方法。

[0098] 在本发明的一个实施例中,所述第一服务器包含一虚拟机,所述第一数据是所述虚拟机生成的数据;所述其他服务器包括第二服务器;所述方法还包括下述步骤 A1-A2:

[0099] 步骤 A1:如果所述第一服务器的剩余存储空间小于所述第一数据所占的存储空间,则向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中。

[0100] 步骤 A2:在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时,将所述虚拟机和所述其他数据迁移到所述第二服务器上。

[0101] 其中,“第二服务器”可以为该分布式存储系统中的、除第一服务器之外的、满足“剩余存储空间大于或者等于所述虚拟机对应的除第一数据之外的其他数据所占的存储空间”条件的任一服务器。

[0102] 步骤 A1 可以包括:A11) 如果所述第一服务器在确定其剩余存储空间小于所述第一数据所占的存储空间,则确定与所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间的大小;A12) 将剩余存储空间大于或者等于该存储空间的服务器作为第二服务器;A13) 向所述第二服务器发送包含所述第一数据的第一指示消息。

[0103] 示例性的,“将所述虚拟机和所述虚拟机对应的其他数据迁移到所述第二服务器上”的具体实现方式如现有技术,此处不再描述。

[0104] 需要说明的是,实际实现时,步骤 A2 可以在步骤 A1 之后立即执行;也可以在所述虚拟机需要读取第一数据时执行。其中,前者可以描述为:在将第一数据写入第二服务器时实现第一数据的本地化;后者可以描述为:在所述虚拟机需要读取所述第一数据时实现第一数据的本地化。

[0105] 该实施例中，在“第一服务器将所述虚拟机迁移到第二服务器上”之后，所述虚拟机成为安装在第二服务器上的一虚拟机，第一服务器上不再存在所述虚拟机。在“第一服务器将与所述其他数据迁移到第二服务器上”之后，所述其他数据均存储在第二服务器的存储空间中。另外，可以将该可选的实施例称为通过虚拟机迁移的方式实现第一数据的本地化。

[0106] 在本发明的另一个实施例中，所述其他服务器包括第二服务器；所述方法还包括下述步骤B1-B4：

[0107] 步骤B1：如果所述第一服务器的剩余存储空间小于所述第一数据所占的存储空间，则向所述第二服务器发送包含所述第一数据的第一指示消息；所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中。

[0108] 步骤B2：更新所述第一服务器的剩余存储空间。

[0109] 步骤B3：当所述第一服务器更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时，向所述第二服务器发送第二指示消息；所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述第一服务器。

[0110] 步骤B4：接收所述第二服务器发送的所述第一数据；将所述第一数据存储在所述第一服务器更新后的剩余存储空间中。

[0111] 其中，在该实施例中，第二服务器在向第一服务器发送第一数据之后，第二服务器可以将本地存储的第一数据删除，也可以不删除。在前者的实现方式中，可以将本实施例称为通过第一数据迁移的方式实现第一数据的本地化；在后者的实现方式中，可以认为分布式存储系统中存储的第一数据的副本由n副本变成了n+1副本，其中，存储第一数据的副本的方法可以参见下文相关的实施例。需要说明的是，为了与上述可选的实施例对应，下文中将该可选的实施例称为通过第一数据迁移的方式实现第一数据的本地化。

[0112] 在本发明的又一实施例中：所述第一服务器包含一虚拟机，所述第一数据是所述虚拟机生成的数据，所述其他服务器还包括第三服务器，所述方法还包括以下步骤C：

[0113] 步骤C：所述第一服务器向所述第三服务器发送包含所述第一数据的副本的第三指示消息，所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

[0114] 其中，“第三服务器”可以用于存储第一数据的副本，分布式存储系统中可以包含一个或者多个第三服务器。实际实现时，分布式存储系统中的每个服务器均可获知该分布式存储系统中所有服务器的剩余存储空间。该实施例中，第一服务器可以按照“负载均衡”的原则，选择分布式存储系统中的一个或者多个服务器作为第三服务器。

[0115] 该实施例通过存储第一数据的副本，可以达到增强分布式存储系统的性能的有益效果。具体的：当第一数据损坏或者丢失时，可以通过调用第一数据的副本使分布式存储系统正常运行，从而增强系统的稳定性。

[0116] 本发明实施例提供的存储数据的方法，应用于包含第一服务器和其他服务器的分布式存储系统中，在第一服务器确定其剩余存储空间大于或者等于其生成的一数据所占的存储空间，且第一服务器的已使用存储空间大于其他服务器的最小已使用存储空间时，优先将该数据存储在该服务器的剩余存储空间中。这样，第一服务器可以直接从本地读取该数据，而不需要从网络中的其他服务器上读取该数据，从而达到缩短时延、提高系统性能的

有益效果。解决了现有技术中,因在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

[0117] 实施例二

[0118] 本实施例提供的存储数据的方法,应用于分布式存储系统,所述分布式存储系统包括第一服务器和其他服务器。本实施例描述的是上述实施例一中的场景 A 下存储数据的方法。

[0119] 如图 2 所示,包括:

[0120] 201:确定所述第一服务器生成的第一数据;所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值。

[0121] 202:判断所述第一服务器的剩余存储空间与所述第一数据所占的存储空间的大小关系。

[0122] 203:如果所述第一服务器的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述第一服务器的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述第一服务器的剩余存储空间中。

[0123] 进一步地,针对本实施例的场景(即场景 A),在现有技术方案中,第一服务器需要对第一数据进行分块,并按照“负载均衡”的原则,分别将每个块存储在分布式存储系统的多个服务器的剩余存储空间中。

[0124] 可选的,在本发明的一个实施例中,所述方法还可以包括:上述实施例一中的步骤 A1-A2,或者上述实施例一中的步骤 B1-B4;另外,还可以包括上述实施例一中的步骤 C。

[0125] 需要说明的是,本实施例中的相关解释可以参考上述实施例一的相关部分。

[0126] 本发明实施例提供的存储数据的方法,应用于包含第一服务器和其他服务器的分布式存储系统中,在第一服务器确定其剩余存储空间大于或者等于其生成的一数据所占的存储空间,且第一服务器的已使用存储空间小于或者等于其他服务器的最小已使用存储空间时,优先将该数据存储在该服务器的剩余存储空间中;其中,该数据所占的存储空间大于分布式存储系统的最小存储单元的最大值。这样,第一服务器可以直接从本地读取该数据,而不需要从网络中的其他服务器上读取该数据,从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中,因在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

[0127] 下面通过几个具体的实施例对上述实施例一和实施例二提供的存储数据的方法进行示例性说明:

[0128] 实施例 1

[0129] 本实施例中第一服务器上不包含虚拟机。本实施例中的“本地”是指生成数据的应用所在的服务器。

[0130] 如图 3 所示,为本实施例提供的一种存储数据的方法,包括:

[0131] 301:第一服务器(物理机)上安装的应用生成一数据。

[0132] 示例性的,这里的“应用”可以为现有技术中的任何一种应用,例如,文字处理应用、音频处理应用、视频图片处理应用、计算机控制应用、计算机辅助设计应用、科学仿真应用等。

[0133] 假设 4 次执行该步骤 301,第一服务器(物理机)上安装的应用共生成 4 个数

据,分别为 D1、D2、D3、D4。

[0134] 302 :第一服务器获取第一服务器的剩余存储空间和该数据所占存储空间。

[0135] 具体地,步骤 302 可以实现为:第一服务器上的分布式存储程序获取第一服务器的剩余存储空间和该数据所占存储空间

[0136] 示例性的,按照步骤 301 中的示例,4 次执行步骤 302,获取的第一服务器的剩余存储空间分别为:X1、X2、X3、X4 ;数据 D1、D2、D3、D4 所占的空间的大小分别为:M1、M2、M3、M4。

[0137] 303 :第一服务器判断其剩余存储空间是否大于或者等于该数据所占存储空间。

[0138] 若是,则执行步骤 304 ;若否,则执行步骤 305。

[0139] 具体地,步骤 303 可以实现为:第一服务器上的分布式存储程序判断第一服务器的剩余存储空间是否大于或者等于该数据所占存储空间。

[0140] 示例性的,按照步骤 302 中的示例,假设:X1 > M1、X2 > M2、X3 = M3、X4 < M4。

[0141] 304 :第一服务器将该数据存储在第一服务器的剩余存储空间中。具体包括:第一服务器上的分布式存储程序为该数据分配存储地址,并将该数据存储在该存储地址对应的存储模块中;该存储地址为第一服务器的剩余存储空间中的一个或者多个存储模块对应的存储地址。

[0142] 在执行步骤 304 之后,则结束。

[0143] 示例性的,按照步骤 303 中的示例,D1、D2、D3 分别存储第一服务器的(剩余)存储空间中。假设第一服务器为 D1、D2、D3 分配的存储地址分别为:S1/D1/offset1、S1/D2/offset2、S1/D3/offset3。其中,S1 表示第一服务器,S1/D1/offset1 表示:数据 D1 的存储地址位于第一服务器 S1 的存储空间的 offset1 处,其他的存储地址不再一一解释。

[0144] 第 3 次执行步骤 301-306 之后,分布式存储系统中记录的数据路由表如表 1 所示:

[0145] 表 1

[0146]

数据	本地	剩余存储空间	数据存储地址
D1	S1	X1	S1/D1/offset1

[0147]

D2	S1	X2	S1/D2/offset2
D3	S1	X3	S1/D3/offset3

[0148] 305 :第一服务器向第二服务器发送包含该数据的第一指示消息;第一指示消息用于指示第二服务器将该数据存储在第二服务器的剩余存储空间中。

[0149] 在执行步骤 305 之后,执行步骤 306。

[0150] 306 :第二服务器根据第一指示消息将该数据存储在第二服务器的剩余存储空间中。具体包括:第二服务器上的分布式存储程序为该数据分配存储地址,并将该数据存储在该存储地址对应的存储模块中;该存储地址为第二服务器的剩余存储空间中的一个或者多个存储模块对应的存储地址。

[0151] 示例性的,按照步骤 303 中的示例,D4 存储在第二服务器的(剩余)存储空间中。假设第二服务器为 D4 分配的存储地址分别为:S2/D4/offset4。其中,S2 表示第二服务器,

S2/D4/offset4 表示数据 D4 的存储地址位于第二服务器 S2 的存储空间的 offset4 处。

[0152] 第 4 次执行步骤 301-306 之后,分布式存储系统中记录的数据路由表如表 2 所示 :

[0153] 表 2

[0154]

数据	本地	剩余存储空间	数据存储地址
D1	S1	X1	S1/D1/offset1
D2	S1	X2	S1/D2/offset2
D3	S1	X3	S1/D3/offset3
D4	S1	X4	S2/D4/offset4

[0155] 由表 2 可知,获取数据 D4 的服务器与存储数据 D4 的服务器为同一服务器。

[0156] 307 :第一服务器更新第一服务器的剩余存储空间,并周期性检测第一服务器更新后的剩余存储空间。

[0157] 具体地,步骤 307 可以实现为:第一服务器上的分布式存储程序更新第一服务器的剩余存储空间,并周期性检测第一服务器更新后的剩余存储空间。

[0158] 示例性的,按照步骤 306 中的示例,将第 4 次执行步骤 301-306 之后,执行步骤 307 时,获得的第一服务器更新后的剩余存储空间表示为 $X4'$ 。假设 $X4' > M4$ 。

[0159] 308 :第一服务器判断其更新后的剩余存储空间是否大于或者等于该数据所占的存储空间。

[0160] 若否,则返回步骤 307 ;若是,则执行步骤 309。

[0161] 具体地,步骤 308 可以实现为:第一服务器上的分布式存储程序判断第一服务器更新后的剩余存储空间是否大于或者等于该数据所占的存储空间。

[0162] 示例性的,按照步骤 307 中的示例,步骤 308 具体为:第一服务器判断 $X4'$ 是否大于或者等于 $M4$ 。

[0163] 309 :第一服务器向第二服务器发送第二指示消息;第二指示消息用于指示第二服务器将该数据发送到第一服务器。

[0164] 310 :第二服务器根据第二指示消息向第一服务器发送该数据。

[0165] 311 :第一服务器将该数据存储在第一服务器更新后的剩余存储空间中。

[0166] 执行步骤 311 之后,则结束。

[0167] 具体地,步骤 311 可以实现为:第一服务器上的分布式存储程序将该数据存储在第一服务器更新后的剩余存储空间中。

[0168] 示例性的,按照步骤 307 中的示例,执行步骤 311 之后,分布式存储系统中记录的数据路由表如表 3 所示 :

[0169] 表 3

[0170]

数据	本地	剩余存储空间	数据存储地址

D1	S1	X1	S1/D1/offset1
D2	S1	X2	S1/D2/offset2
D3	S1	X3	S1/D3/offset3
D4	S1	X4'	S1/D4/offset4

[0171] 由表 3 可知, 获取数据 D4 的服务器与存储数据 D4 的服务器为同一服务器。其中, S1/D4/offset4 表示数据 D4 的存储地址位于第一服务器 S1 的 offset4 处。

[0172] 可选的, 在步骤 301 之后, 该方法还可以包括以下步骤 A 和步骤 B :

[0173] 步骤 A : 第一服务器向第三服务器发送包含该数据的副本的第三指示消息, 第三指示消息用于指示第三服务器将该数据的副本存储在第三服务器的剩余存储空间中。

[0174] 示例性的, 按照实施例 1 中的示例, 将 D1、D2、D3、D4 的副本分别表示为 :D1'、D2'、D3'、D4' 。

[0175] 步骤 B : 第三服务器根据第三指示消息将该数据的副本存储在第三服务器的剩余存储空间中。

[0176] 具体地, 步骤 B 可以实现为 : 第三服务器上的分布式存储程序根据第三指示消息将该数据的副本存储在第三服务器的剩余存储空间中。

[0177] 示例性的, 按照实施例 1 中的示例, 假设第三服务器为 D1、D2、D3、D4 分配的存储地址分别为 :S3/D1' /offset1、S3/D2' /offset2、S3/D3' /offset3、S3/D4' /offset4, 其中, S3 表示第三服务器, S3/D1' /offset1 表示。数据 D1 的副本 D1' 存储地址位于第三服务器 S3 的存储空间的 offset1 处, 其他的存储地址不再一一解释。

[0178] 在第 3 次执行步骤 301-306 以及步骤 A、步骤 B 之后, 分布式存储系统中记录的数据路由表如表 4 所示 :

[0179] 表 4

[0180]

数据	本地	剩余存储空间	数据存储地址	副本	副本存储地址
D1	S1	X1	S1/D1/offset1	D1'	S3/D1' /offset1
D2	S1	X2	S1/D2/offset2	D2'	S3/D2' /offset2
D3	S1	X3	S1/D3/offset3	D3'	S3/D3' /offset3

[0181] 在第 4 次执行步骤 301-306 以及步骤 A、步骤 B 之后, 分布式存储系统中记录的数据路由表如表 5 所示 :

[0182] 表 5

[0183]

数据	本地	剩余存储空间	数据存储地址	副本	副本存储地址

D1	S1	X1	S1/D1/offset1	D1'	S3/D1' /offset1
D2	S1	X2	S1/D2/offset2	D2'	S3/D2' /offset2

[0184]

D3	S1	X3	S1/D3/offset3	D3'	S3/D3' /offset3
D4	S1	X4	S2/D4/offset4	D4'	S3/D4' /offset3

[0185] 进一步地,该步骤 A 和步骤 B 用于增强系统的稳定性,具体的,若第二服务器的存储空间中存储的该数据丢失或者损坏时,第一服务器可以向第三服务器发送将该数据的副本发送至第一服务器的指示消息;使得第三服务器根据该指示消息向第一服务器发送该数据的副本;第一服务器将该数据的副本存储在第一服务器更新后的剩余存储空间中。

[0186] 本发明实施例提供的存储数据的方法,在第一服务器确定其剩余存储空间大于或者等于其安装的一应用生成的一数据所占的存储空间时,优先将该数据存储在该服务器的剩余存储空间中;在第一服务器确定其剩余存储空间小于该数据所占的存储空间时,将该数据存储在第二服务器的剩余存储空间中,并在确定第一服务器更新后的剩余存储空间大于或者等于该数据所占的存储空间时,将该数据存储在第一服务器更新后的剩余存储空间中。这样,第一服务器上安装的应用可以直接从本地(第一服务器的存储空间)读取该数据,不需要从网络中的其他服务器上读取该数据,从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中,因在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

[0187] 实施例 2

[0188] 本实施例中第一服务器上包含虚拟机。本实施例中的“本地”是指生成数据的虚拟机所在的服务器。

[0189] 如图 4 所示,为本实施例提供的一种存储数据的方法,包括:

[0190] 401:第一服务器上安装的虚拟机生成一数据。

[0191] 示例性的,假设第一服务器上安装有 4 个虚拟机(VM1、VM2、VM3、VM4),4 执行该步骤 401,具体为:VM1、VM2、VM3、VM4 分别获取数据 D1、D2、D3、D4。

[0192] 402:第一服务器获取第一服务器的剩余存储空间和该数据所占存储空间。

[0193] 403:第一服务器判断其剩余存储空间是否大于或者等于该数据所占存储空间。

[0194] 若是,则执行步骤 404;若否,则执行步骤 405。

[0195] 404:第一服务器将该数据存储在第一服务器的剩余存储空间中。

[0196] 在执行步骤 404 之后,则结束。

[0197] 需要说明的是,步骤 402-步骤 404 的示例可以参考上述实施例 1 中的步骤 302-304,此处不再赘述。

[0198] 405:第一服务器确定与步骤 401 中的虚拟机对应的除所述第一数据之外的其他数据所占的存储空间的大小,将其标记为 W。

[0199] 406:第一服务器确定分布式存储系统中、剩余存储空间大于或者等于 W 的服务器,作为第二服务器。

[0200] 407 :第一服务器向第二服务器发送包含该数据的第一指示消息 ;第一指示消息用于指示第二服务器将该数据存储在第二服务器的剩余存储空间中。

[0201] 408 :第二服务器根据第一指示消息将该数据存储在第二服务器的剩余存储空间中。

[0202] 需要说明的是,步骤 407-408 中的具体示例可以参考上述实施例 1 中的步骤 305-306 相同,此处不再赘述。另外,实施例 1 中的表 1 和表 2 也可以适用于本实施的对应步骤中。

[0203] 进一步地,为了清楚对比实施例 1 和实施例 2,在本实施例中,上述表 1 和表 2 可以分别表示为以下表 1' 和表 2' :

[0204] 表 1'

[0205]

虚拟机	数据	本地	剩余存储空间	数据存储地址
VM1	D1	S1	X1	S1/D1/offset1
VM2	D2	S1	X2	S1/D2/offset2
VM3	D3	S1	X3	S1/D3/offset3

[0206] 表 2'

[0207]

虚拟机	数据	本地	剩余存储空间	数据存储地址
VM1	D1	S1	X1	S1/D1/offset1
VM2	D2	S1	X2	S1/D2/offset2
VM3	D3	S1	X3	S1/D3/offset3
VM4	D4	S1	X4	S2/D4/offset4

[0208] 由表 2' 可知,获取 D4 的虚拟机 VM4 所在的服务器 S1 与存储 D4 的服务器 S2 不是同一服务器。

[0209] 409 :第一服务器将该虚拟机和该虚拟机生成的其他数据迁移到第二服务器上。

[0210] 执行步骤 409 之后,则结束。

[0211] 示例性的,假设数据路由表中记录的、与该虚拟机存在对应关系的所有数据为 :D1、D2、D3、D4,则步骤 409 具体为 :第一服务器将该虚拟机和 D1、D2、D3 迁移到第二服务器上。

[0212] 示例性的,按照步骤 408 中的示例,执行步骤 409 之后,分布式存储系统中记录的数据路由表如表 3' 所示 :

[0213] 表 3'

[0214]

虚拟机	数据	本地	剩余存储空间	数据存储地址
VM1	D1	S1	X1	S1/D1/offset1
VM2	D2	S1	X2	S1/D2/offset2
VM3	D3	S1	X3	S1/D3/offset3
VM4	D4	S2	X4	S2/D4/offset4

[0215] 由表 3' 可知, 获取 D4 的虚拟机 VM4 所在的服务器 S2 与存储 D4 的服务器 S2 为同一服务器。

[0216] 可选的, 该方法还可以包括上述实施例 1 中的步骤 A 和步骤 B。

[0217] 为了清楚对比实施例 1 和实施例 2, 在本实施例中, 上述表 4 和表 5 可以分别表示为以下表 4' 和表 5' :

[0218] 表 4'

[0219]

虚拟机	数据	本地	剩余存储空间	数据存储地址	副本	副本存储地址
VM1	D1	S1	X1	S1/D1/offset1	D1'	S3/D1' /offset1
VM2	D2	S1	X2	S1/D2/offset2	D2'	S3/D2' /offset2
VM3	D3	S1	X3	S1/D3/offset3	D3'	S3/D3' /offset3

[0220] 表 5'

[0221]

虚拟机	数据	本地	剩余存储空间	数据存储地址	副本	副本存储地址
VM1	D1	S1	X1	S1/D1/offset1	D1'	S3/D1' /offset1
VM2	D2	S1	X2	S1/D2/offset2	D2'	S3/D2' /offset2
VM3	D3	S1	X3	S1/D3/offset3	D3'	S3/D3' /offset3
VM4	D4	S1	X4	S2/D4/offset4	D4'	S3/D4' /offset3

[0222]

[0223] 本发明实施例提供的存储数据的方法, 在第一服务器确定其剩余存储空间大于或者等于其上的一虚拟机生成的一数据所占的存储空间时, 优先将该数据存储在该服务器的剩余存储空间中; 在第一服务器确定其剩余存储空间小于该数据所占的存储空间时, 将该数据存储在第二服务器的存储空间中, 并将该虚拟机迁移到第二服务器上。这样, 虚拟机可以直接从本地(第二服务器的存储空间)读取数据, 不需要从网络中的其他服务器上读取数据, 从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中, 因在网络中的其

他服务器上读取数据导致的时延长、系统性能较差的问题。

[0224] 实施例三

[0225] 如图 5 所示,为本实施例提供的一种服务器 1,应用于分布式存储系统,所述分布式存储系统还包括其他服务器,所述服务器 1 用以执行图 1 所示的存储数据的方法,所述服务器 1 包括:

[0226] 确定模块 51,用于确定所述服务器 1 生成的第一数据;

[0227] 判断模块 52,用于判断所述服务器 1 的剩余存储空间与所述第一数据所占的存储空间的大小关系;

[0228] 存储模块 53,用于在所述判断模块 52 确定所述服务器 1 的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述服务器 1 的已使用存储空间大于所述其他服务器的最小已使用存储空间时,将所述第一数据存储在所述服务器 1 的剩余存储空间中。

[0229] 可选的,所述服务器 1 包含一虚拟机,所述第一数据是所述虚拟机生成的数据;所述其他服务器包括第二服务器;如图 6 所示,所述服务器 1 还包括:

[0230] 发送模块 54,用于在所述判断模块 52 确定所述服务器 1 的剩余存储空间小于所述第一数据所占的存储空间时,向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

[0231] 迁移模块 55,用于在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时,将所述虚拟机和所述其他数据迁移到所述第二服务器上。

[0232] 可选的,所述其他服务器包括第二服务器,如图 6 所示,所述服务器 1 还包括:

[0233] 发送模块 54,用于在所述判断模块 52 确定所述服务器 1 的剩余存储空间小于所述第一数据所占的存储空间时,向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

[0234] 更新模块 56,用于更新所述服务器 1 的剩余存储空间;

[0235] 所述发送模块 54 还用于,当所述服务器 1 更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时,向所述第二服务器发送第二指示消息;所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述服务器 1;

[0236] 接收模块 57,用于接收所述第二服务器发送的所述第一数据;

[0237] 所述存储模块 53 还用于,将所述第一数据存储在所述服务器 1 更新后的剩余存储空间中。

[0238] 可选的,所述其他服务器还包括第三服务器;

[0239] 所述发送模块 54 还用于,向所述第三服务器发送包含所述第一数据的副本的第三指示消息,所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

[0240] 示例性的,本实施例中的服务器 1 具体可以为上述实施例一中描述的“第一服务器”,本实施例中的“第二服务器”可以为上述实施例一中描述的“第二服务器”。

[0241] 本发明实施例提供的服务器 1,应用于还包含其他服务器的分布式存储系统中,在

服务器 1 确定其剩余存储空间大于或者等于其生成的一数据所占的存储空间,且服务器 1 的已使用存储空间大于其他服务器的最小已使用存储空间时,优先将该数据存储在该服务器的剩余存储空间中。这样,服务器 1 可以直接从本地读取该数据,而不需要从网络中的其他服务器上读取该数据,从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中,因在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

[0242] 实施例四

[0243] 针对实施例三,在硬件实现上,其中的发送模块可以为发送器,接收模块可以为接收器,且该发送器和接收器可以集成在一起构成收发器;存储模块可以为存储器,确定模块、判断模块、迁移模块等可以以硬件形式内嵌于或独立于服务器 1 的处理器中,也可以以软件形式存储于服务器 1 的存储器中,以便于处理器调用执行以上各个模块对应的操作,该处理器可以为中央处理单元(CPU)、微处理器、单片机等。

[0244] 如图 7 所示,为本发明实施例提供的一种服务器 1,应用于分布式存储系统,所述分布式存储系统还包括其他服务器,所述服务器 1 用以执行图 1 所示的存储数据的方法,所述服务器 1 包括:总线系统 71、存储器 72、处理器 73。

[0245] 其中,存储器 72 和处理器 73 之间是通过总线系统 71 耦合在一起的,其中总线系统 71 除包括数据总线之外,还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见,在图中将各种总线都标为总线系统 71。

[0246] 存储器 72,用于存储一组代码;

[0247] 存储器 72 中存储的代码用于控制处理器 73 确定所述服务器 1 生成的第一数据;以及判断所述服务器 1 的剩余存储空间与所述第一数据所占的存储空间的大小关系;

[0248] 存储器 72,还用于在所述处理器 73 确定所述服务器 1 的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述服务器 1 的已使用存储空间大于所述其他服务器的最小已使用存储空间时,在所述处理器的控制下将所述第一数据存储在所述服务器 1 的剩余存储空间中。

[0249] 可选的,所述服务器 1 包含一虚拟机,所述第一数据是所述虚拟机生成的数据;所述其他服务器包括第二服务器;如图 8 所示,所述服务器 1 还包括:

[0250] 发送器 74,用于在所述处理器 73 确定所述服务器 1 的剩余存储空间小于所述第一数据所占的存储空间时,向所述第二服务器发送包含所述第一数据的第一指示消息;所述第一指示消息用于指示所述第二服务器将所述第一数据存储在所述第二服务器的剩余存储空间中;

[0251] 所述处理器 73 还用于,在所述虚拟机对应的除所述第一数据之外的其他数据所占的存储空间小于或者等于所述第二服务器的剩余存储空间时,将所述虚拟机和所述其他数据迁移到所述第二服务器上。

[0252] 可选的,所述其他服务器包括第二服务器;

[0253] 所述处理器 73 还用于,更新所述服务器 1 的剩余存储空间;

[0254] 所述发送器 74 还用于,当所述服务器 1 更新后的剩余存储空间大于或者等于所述第一数据所占的存储空间时,向所述第二服务器发送第二指示消息;所述第二指示消息用于指示所述第二服务器将所述第一数据迁移到所述服务器 1;

[0255] 如图 8 所示,所述服务器 1 还包括:接收器 75,用于接收所述第二服务器发送的所

述第一数据；

[0256] 所述存储器 72 还用于，在所述处理器 73 的控制下将所述第一数据存储在所述服务器 1 更新后的剩余存储空间中。

[0257] 可选的，所述其他服务器还包括第三服务器；

[0258] 所述发送器 74 还用于，向所述第三服务器发送包含所述第一数据的副本的第三指示消息，所述第三指示消息用于指示所述第三服务器将所述第一数据的副本存储在所述第三服务器的剩余存储空间中。

[0259] 示例性的，本实施例中的服务器 1 具体可以为上述实施例一中描述的“第一服务器”，本实施例中的“第二服务器”可以为上述实施例一中描述的“第二服务器”。

[0260] 本发明实施例提供的服务器 1，应用于还包含其他服务器的分布式存储系统中，在服务器 1 确定其剩余存储空间大于或者等于其生成的一数据所占的存储空间，且服务器 1 的已使用存储空间大于其他服务器的最小已使用存储空间时，优先将该数据存储在该服务器的剩余存储空间中。这样，服务器 1 可以直接从本地读取该数据，而不需要从网络中的其他服务器上读取该数据，从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中，因在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

[0261] 实施例五

[0262] 如图 9 所示，为本实施例提供的一种服务器 2，应用于分布式存储系统，所述分布式存储系统还包括其他服务器，所述服务器 2 用以执行图 2 所示的存储数据的方法，所述服务器 2 包括：

[0263] 确定模块 91，用于确定所述服务器 2 生成的第一数据；所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值；

[0264] 判断模块 92，用于判断所述服务器 2 的剩余存储空间与所述第一数据所占的存储空间的大小关系；

[0265] 存储模块 93，用于在所述判断模块 92 确定所述服务器 2 的剩余存储空间大于或者等于所述第一数据所占的存储空间，且所述服务器 2 的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时，将所述第一数据存储在所述服务器 2 的剩余存储空间中。

[0266] 示例性的，本实施例中的服务器 2 具体可以为上述实施例一中描述的“第一服务器”，

[0267] 本发明实施例提供的服务器 2，应用于还包含其他服务器的分布式存储系统中，在服务器 2 确定其剩余存储空间大于或者等于其生成的一数据所占的存储空间，且服务器 2 的已使用存储空间小于或者等于其他服务器的最小已使用存储空间时，优先将该数据存储在该服务器的剩余存储空间中；其中，该数据所占的存储空间大于分布式存储系统的最小存储单元的最大值。这样，服务器 2 可以直接从本地读取该数据，而不需要从网络中的其他服务器上读取该数据，从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中，因在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

[0268] 实施例六

[0269] 针对实施例五，在硬件实现上，其中的存储模块可以为存储器，确定模块、判断模块可以以硬件形式内嵌于或独立于服务器 2 的处理器中，也可以以软件形式存储于服务器

2 的存储器中,以便于处理器调用执行以上各个模块对应的操作,该处理器可以为中央处理单元(CPU)、微处理器、单片机等。

[0270] 如图 10 所示,为本发明实施例提供的一种服务器 2,应用于分布式存储系统,所述分布式存储系统还包括其他服务器,所述服务器 2 用以执行图 2 所示的存储数据的方法,所述服务器 1 包括:总线系统 10A、存储器 10B、处理器 10C。

[0271] 其中,存储器 10B 和处理器 10C 之间是通过总线系统 10A 耦合在一起的,其中总线系统 10A 除包括数据总线之外,还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见,在图中将各种总线都标为总线系统 10A。

[0272] 存储器 10B,用于存储一组代码;

[0273] 存储器 10B 中存储的代码用于控制处理器 10C 确定所述服务器 2 生成的第一数据;以及判断所述服务器 2 的剩余存储空间与所述第一数据所占的存储空间的大小关系;所述第一数据所占的存储空间大于所述分布式存储系统的最小存储单元的最大值;

[0274] 存储器 10B,还用于在所述处理器 10C 确定所述服务器 2 的剩余存储空间大于或者等于所述第一数据所占的存储空间,且所述服务器 2 的已使用存储空间小于或者等于所述其他服务器的最小已使用存储空间时,在所述处理器 10C 的控制下将所述第一数据存储在所述服务器 2 的剩余存储空间中。

[0275] 示例性的,本实施例中的服务器 2 具体可以为上述实施例一中描述的“第一服务器”,

[0276] 本发明实施例提供的服务器 2,应用于还包含其他服务器的分布式存储系统中,在服务器 2 确定其剩余存储空间大于或者等于其生成的一数据所占的存储空间,且服务器 2 的已使用存储空间小于或者等于其他服务器的最小已使用存储空间时,优先将该数据存储在该服务器的剩余存储空间中;其中,该数据所占的存储空间大于分布式存储系统的最小存储单元的最大值。这样,服务器 2 可以直接从本地读取该数据,而不需要从网络中的其他服务器上读取该数据,从而达到缩短时延、提高系统性能的有益效果。解决了现有技术中,因在网络中的其他服务器上读取数据导致的时延长、系统性能较差的问题。

[0277] 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,装置和模块的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0278] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统,装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述模块的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个模块或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或模块的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0279] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的,作为模块显示的部件可以是或者也可以不是物理模块,即可以位于一个地方,或者也可以分布到多个网络模块上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0280] 另外,在本发明各个实施例中的各功能模块可以集成在一个处理模块中,也可以是各个模块单独物理包括,也可以两个或两个以上模块集成在一个模块中。上述集成的模

块既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

[0281] 上述以软件功能模块的形式实现的集成的模块，可以存储在一个计算机可读取存储介质中。上述软件功能模块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部分步骤。而前述的存储介质包括：U 盘、移动硬盘、ROM (Read-Only Memory, 只读存储器)、RAM (Random Access Memory, 随机存取存储器)、磁碟或者光盘等各种可以存储程序代码的介质。

[0282] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；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：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

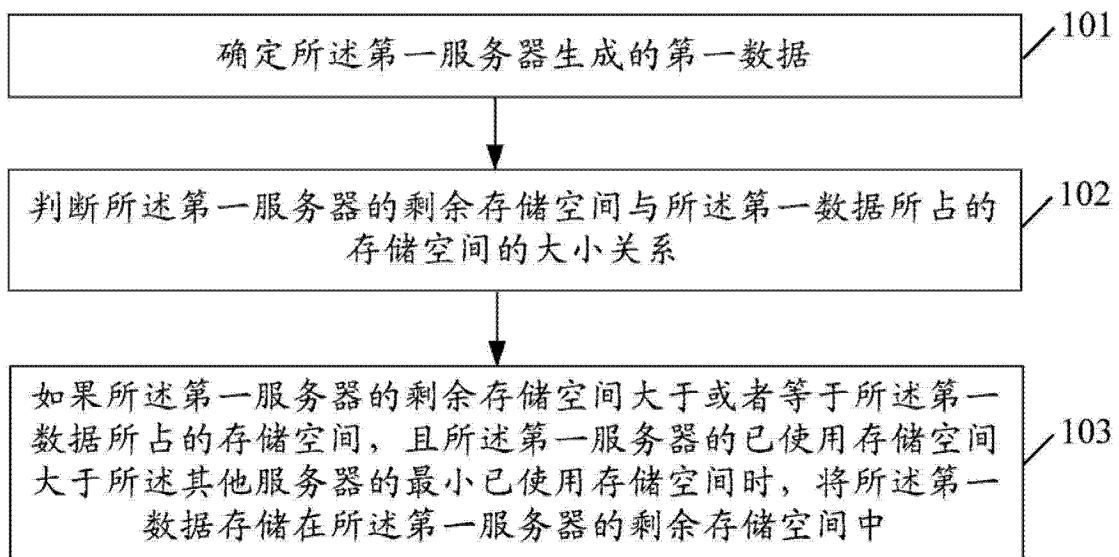


图 1

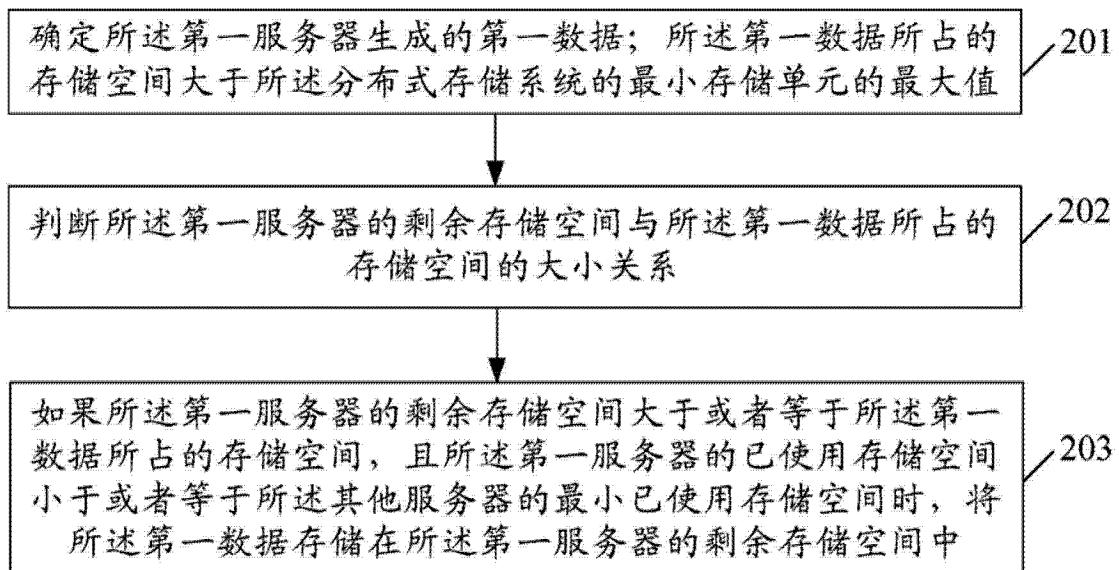


图 2

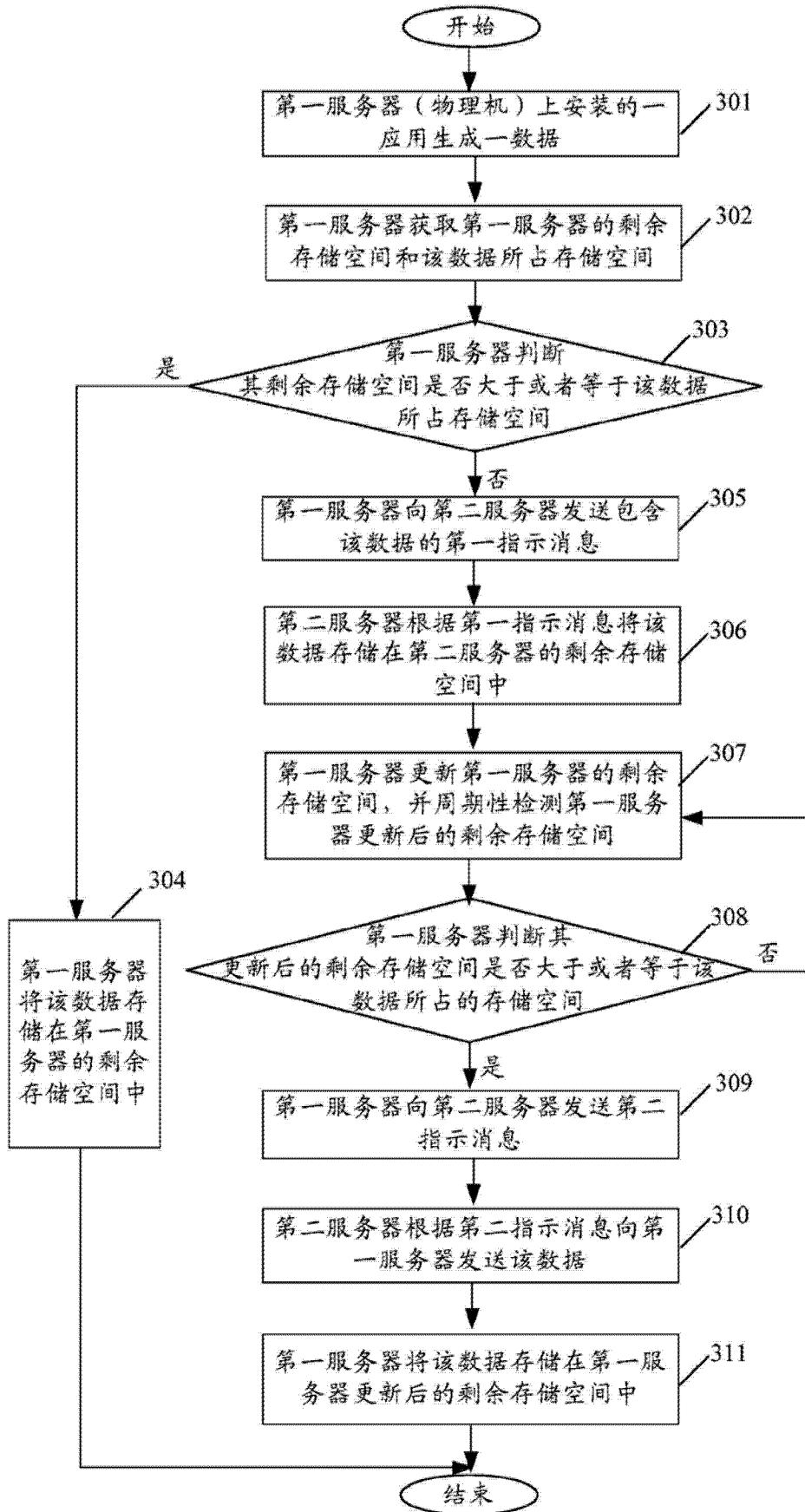


图 3

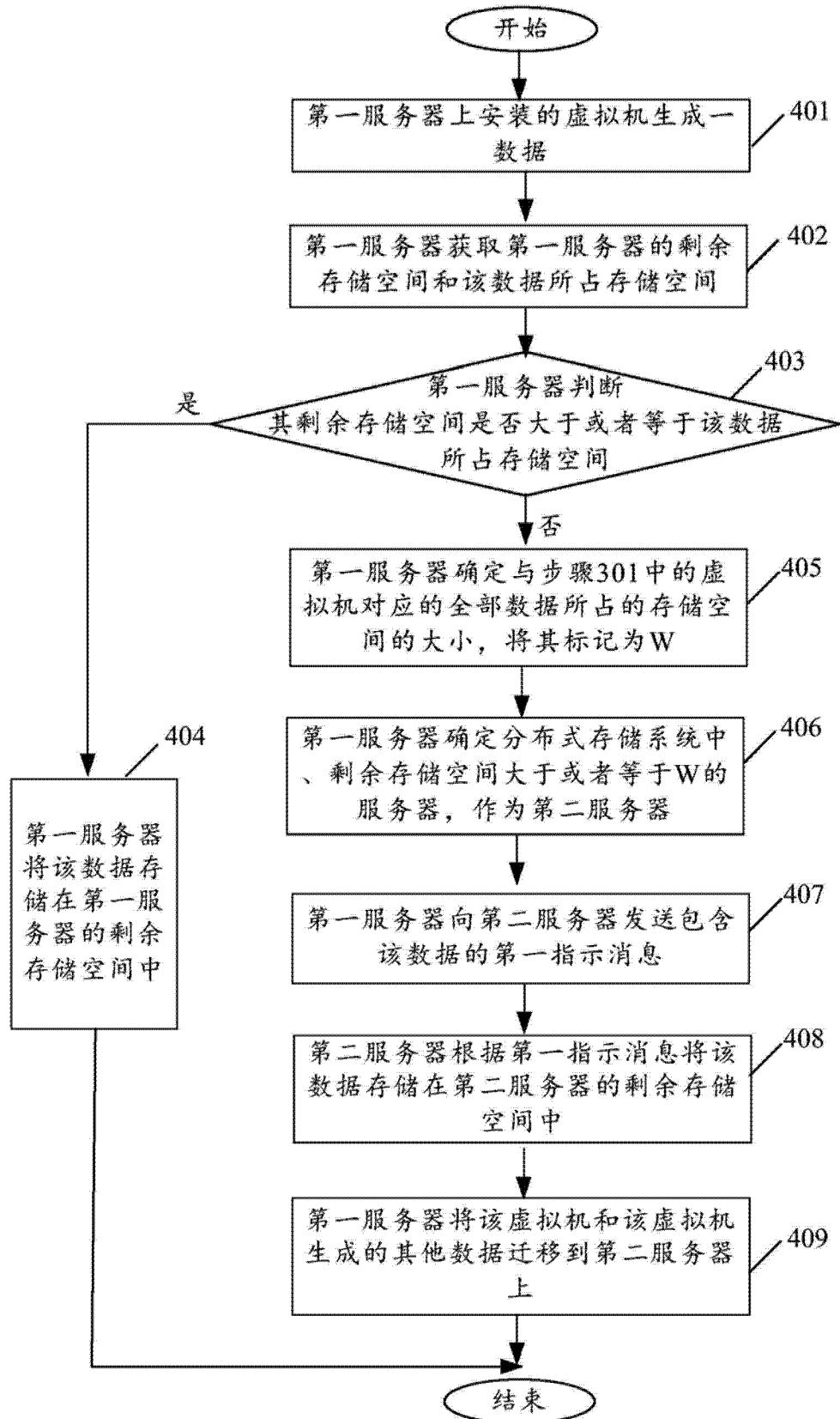


图 4

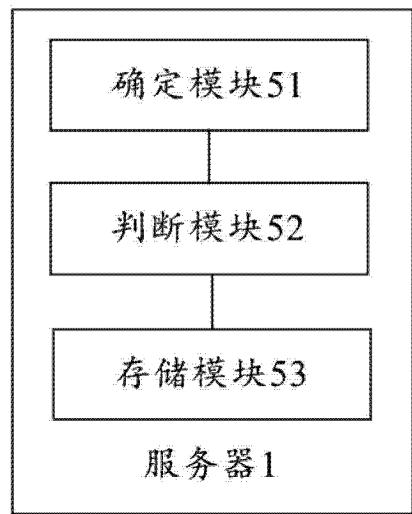


图 5

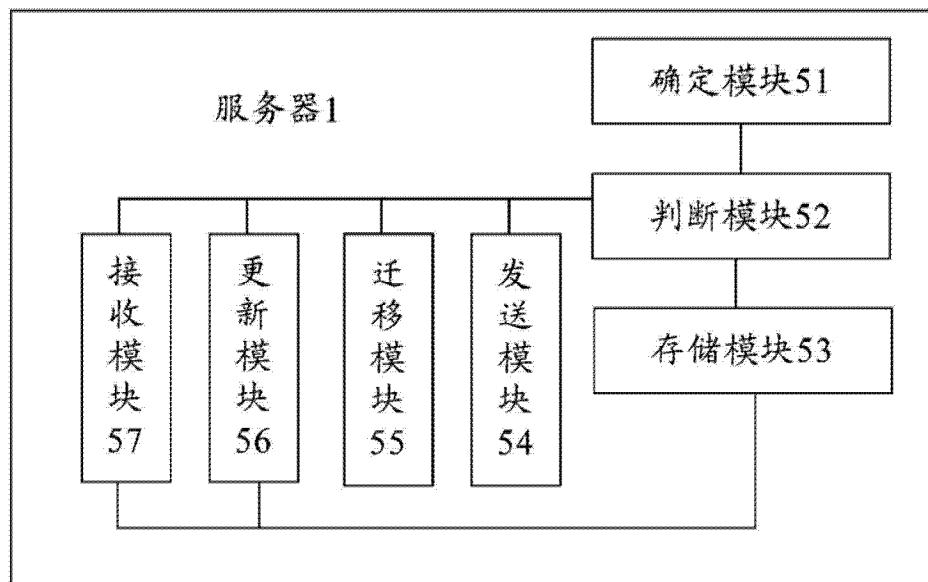


图 6

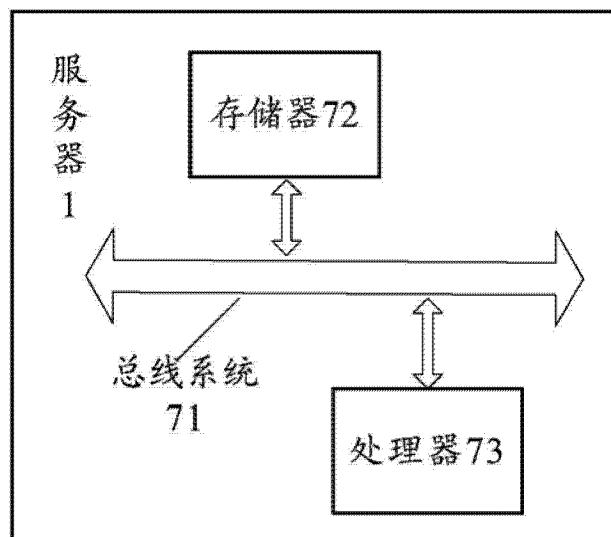


图 7

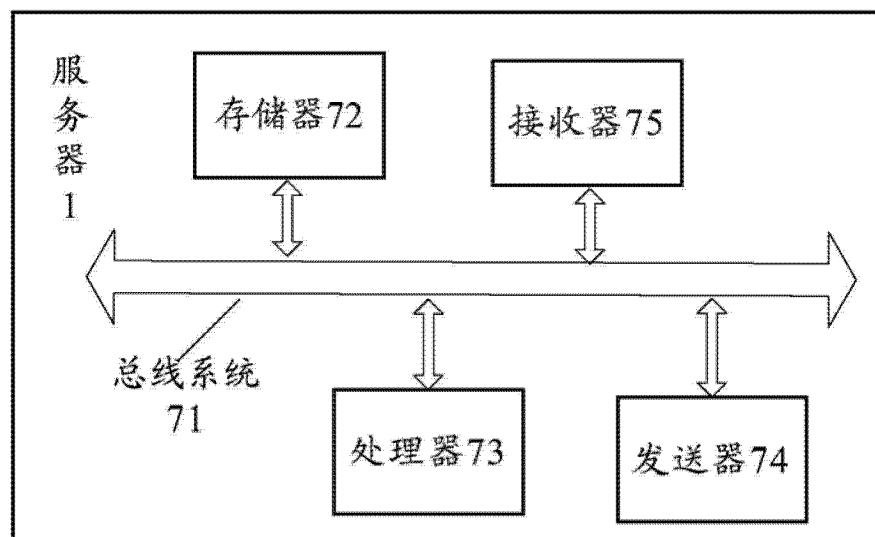


图 8

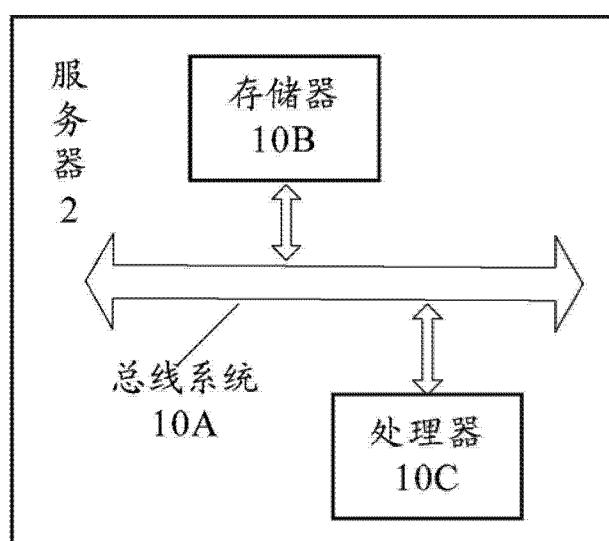
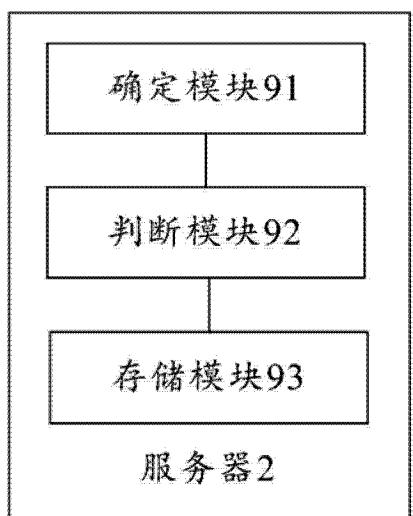


图 9

图 10