(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 103905517 A
(43)申请公布日 2014.07.02

(21)申请号 201210587462.9

(22)申请日 2012.12.28

(71)申请人 中国移动通信集团公司
地址 100032 北京市西城区金融大街 29 号

(72)发明人 郭磊涛 钱岭 王娟

(74)专利代理机构 北京鑫媛睿博知识产权代理有限公司 11297
代理人 龚家骅

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

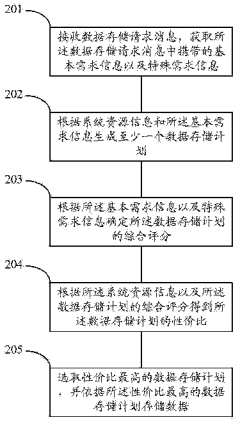
权利要求书2页 说明书6页 附图4页

(54)发明名称

一种数据存储方法及设备

(57)摘要

本发明公开了一种数据存储方法，基于数据存储的基本需求以及特殊需求，确定性价比最高的数据存储计划，并根据性价比最高的数据存储计划进行数据存储，保证了数据存储的 QoS 需求。本发明还同时公开了一种数据存储设备。



CN 103905517 A

1. 一种数据存储的方法，其特征在于，包括：

接收数据存储请求消息，获取所述数据存储请求消息中携带的基本需求信息以及特殊需求信息；

根据系统资源信息和所述基本需求信息生成至少一个数据存储计划；

根据所述基本需求信息以及特殊需求信息确定所述数据存储计划的综合评分；

根据所述系统资源信息以及所述数据存储计划的综合评分确定所述数据存储计划的性价比；

选取性价比最高的数据存储计划，并依据所述性价比最高的数据存储计划存储数据。

2. 如权利要求 1 所述的方法，其特征在于，根据所述基本需求信息以及所述特殊需求信息确定所述数据存储计划的综合评分，具体为：

根据所述基本需求信息生成所述数据存储计划的初始评分，并将所述初始评分依据所述特殊需求信息进行调整。

3. 如权利要求 2 所述的方法，其特征在于，根据所述基本需求信息生成所述数据存储计划的初始评分，具体为：

根据所述基本需求信息中各性能选项所对应的权重因子，以及各性能选项在所述数据存储计划与所述基本需求信息之间的比例，生成所述数据存储计划的初始评分。

4. 如权利要求 1-3 任一项所述的方法，其特征在于，

所述基本需求信息至少包括一种或多种以下性能选项：数据存储空间、数据读写速度、网络带宽；

所述特殊需求信息至少包括一种或多种以下性能选项：数据读写特征、待存储数据与系统数据的操作关系、数据存储时间。

5. 如权利要求 1 所述的方法，其特征在于，在接收数据存储请求消息之前，还包括：

当前节点接收各其他节点发送的节点资源信息，并将所述接收到的节点资源信息按照预设的规则进行分类，生成所述系统资源信息。

6. 一种数据存储设备，其特征在于，包括：

获取模块，用于获取接收到的数据存储请求消息中携带的基本需求信息以及特殊需求信息；

生成模块，用于根据系统资源信息和所述基本需求信息生成至少一个数据存储计划；；

确定模块，用于根据所述基本需求信息以及特殊需求信息确定所述数据存储计划的综合评分，并根据所述系统资源信息以及所述数据存储计划的综合评分得到所述数据存储计划的性价比；

存储模块，用于选取性价比最高的数据存储计划，并依据所述性价比最高的数据存储计划存储数据。

7. 如权利要求 6 所述的设备，其特征在于，所述确定模块，具体用于：

根据所述基本需求信息生成所述数据存储计划的初始评分，并将所述初始评分依据所述特殊需求信息进行调整。

8. 如权利要求 7 所述的设备，其特征在于，所述确定模块，具体用于：

根据所述基本需求信息中各性能选项所对应的权重因子，以及各性能选项在所述数据

存储计划与所述基本需求信息之间的比例，生成所述数据存储计划的初始评分。

9. 如权利要求 6-8 任一项所述的设备，其特征在于，

所述基本需求信息至少包括一种或多种以下性能选项：数据存储空间、数据读写速度、网络带宽；

所述特殊需求信息至少包括一种或多种以下性能选项：数据读写特征、待存储数据与系统数据的操作关系、数据存储时间。

10. 如权利要求 6 所述的设备，其特征在于，还包括：处理模块，

所述处理模块，具体用于接收各其他节点发送的节点资源信息，并将所述接收到的节点资源信息按照预设的规则进行分类，生成所述系统资源信息。

一种数据存储方法及设备

技术领域

[0001] 本发明涉及通信技术领域，特别涉及一种数据存储方法，本发明还涉及一种数据存储设备。

背景技术

[0002] 目前，分布式存储系统已经成为大数据应用解决方案的基础组件，有大量的软件系统和应用基于分布式存储开发和运行，而每一种软件和应用对分布式存储系统具有不同的服务质量 QoS 的需求。例如，对于分布式数据挖掘应用来说，其需要存储大量的数据并以此进行分布式计算分析，典型应用如“大云”BC-PDM（parallel data mining）其运行在 HDFS 分布式文件系统上，并利用 MapReduce 并行计算框架进行分析。BC-PDM 对分布式存储系统的主要需求是大容量以及节点间数据交换能力。而对于同样构建在 HDFS 上的“键值对”系统 HBase 来说，其对 HDFS 除了容量需求外，还需要大的内存空间需求以及快速的磁盘读写 IO。甚至，在 HBase 系统内部的各个模块对存储系统的需求也是不一样的，例如 HBase 系统的元数据信息和数据信息均保存在分布式文件系统中，由于元数据信息经常更新和读取，其需要存储在性能更高的磁盘上。

[0003] 如图 1 所示，为分布式存储系统的结构示意图，系统一般由多个从节点、一个或多个主节点和客户端组成。数据被分散存储到各个从节点上，主节点负责管理系统中的元数据信息，从节点负责数据的存储，客户端通过与主节点和从节点直接交互实现数据的读写。

[0004] 为了满足数据存储的需求，现有技术提出了如下一些解决方案：

[0005] （1）调整副本数满足读写性能的需求：如监控数据的热度，根据热度调整数据副本的数量；

[0006] （2）调整数据放置的位置满足网络的需求：如通过将数据副本放置在边缘服务器，从而满足网络带宽的要求。

[0007] 以上方案仅通过简单的增加副本或改变数据放置位置来满足读写性能及网络带宽的要求，无法提供有 QoS 保证的数据存储服务。

发明内容

[0008] 本发明提供一种数据存储方法，根据数据存储的基本需求以及特殊需求确定性价比最高的数据存储计划，并以此存储数据，从而保证数据存储的 QoS 需求。

[0009] 为达到上述目的，本发明一方面提供了一种数据存储的方法，包括：

[0010] 接收数据存储请求消息，获取所述数据存储请求消息中携带的基本需求信息以及特殊需求信息；

[0011] 根据系统资源信息和所述基本需求信息生成至少一个数据存储计划；

[0012] 根据所述基本需求信息以及特殊需求信息确定所述数据存储计划的综合评分；

[0013] 根据所述系统资源信息以及所述数据存储计划的综合评分确定所述数据存储计划的性价比；

[0014] 选取性价比最高的数据存储计划，并依据所述性价比最高的数据存储计划存储数据。

[0015] 另一方面，本发明还提供了一种数据存储设备，包括：

[0016] 获取模块，用于获取接收到的数据存储请求消息中携带的基本需求信息以及特殊需求信息；

[0017] 生成模块，用于根据系统资源信息和所述基本需求信息生成至少一个数据存储计划；；

[0018] 确定模块，用于根据所述基本需求信息以及特殊需求信息确定所述数据存储计划的综合评分，并根据所述系统资源信息以及所述数据存储计划的综合评分得到所述数据存储计划的性价比；

[0019] 存储模块，用于选取性价比最高的数据存储计划，并依据所述性价比最高的数据存储计划存储数据。

[0020] 与现有技术相比，本发明具有以下优点：

[0021] 本发明基于数据存储的基本需求以及特殊需求，确定性价比最高的数据存储计划，并根据性价比最高的数据存储计划进行数据存储，保证了数据存储的 QoS 需求。

附图说明

[0022] 图 1 为现有技术中分布式存储系统的结构示意图；

[0023] 图 2 为本发明实施例提供的一种数据存储方法流程示意图；

[0024] 图 3 为本发明具体实施例提供的一种数据存储方法流程示意图；

[0025] 图 4 为本发明实施例提供的一种数据存储设备结构示意图。

具体实施方式

[0026] 基于现有技术方案的不足，本发明实施例提供了一种数据存储方法，根据数据存储的基本需求以及特殊需求确定性价比最高的数据存储计划，并以此存储数据，从而保证数据存储的 QoS 需求。

[0027] 下面结合附图对本发明实施例进行详细描述。

[0028] 参见图 2，为本发明提出的一种数据存储方法，该方法具体包括以下步骤：

[0029] 步骤 201，接收数据存储请求消息，获取所述数据存储请求消息中携带的基本需求信息以及特殊需求信息。

[0030] 由于各种应用对数据的操作特征不同，因此在接收到数据存储请求消息后，首先获取其中所携带的基本需求信息以及特殊信息，这些信息均在应用或客户端发送数据存储请求消息前预先设置在请求消息之中。

[0031] 步骤 202，根据系统资源信息和所述基本需求信息生成至少一个数据存储计划。

[0032] 根据上一步骤中所获取的基本需求以及已有资源信息表，生成可满足存储需求的数据存储计划。该步骤中可预先生成所有的数据存储计划，再依据基本需求信息从中选择，或者是按照基本需求信息逐项地生成数据存储计划，以上方式可根据实际情况灵活选择。

[0033] 步骤 203，根据所述基本需求信息以及特殊需求信息确定所述数据存储计划的综合评分。

[0034] 该步骤中，将进一步地根据数据的基本需求以及特殊需求对上一步骤生成的数据存储计划进行评估和调整；其中的特殊需求至少包括读写特征（少写多读、多写少读、读写均衡等）、与已有数据的操作关系（如，先读已有的数据，再读目前写入的数据；或者并发读已有的数据和目前写入的数据），以及数据存储时长等。

[0035] 步骤 204，根据所述系统资源信息以及所述数据存储计划的综合评分得到所述数据存储计划的性价比。

[0036] 步骤 205，选取性价比最高的数据存储计划，并依据所述性价比最高的数据存储计划存储数据。

[0037] 为了进一步阐述本发明的技术思想，现结合具体的应用场景，对本发明的技术方案进行说明。

[0038] 如图 3 所示，为本发明具体实施例所提出的一种数据存储方法流程，该流程包括以下步骤：

[0039] 步骤 301，节点接收各其他节点发送的自身资源信息，根据规则将所有节点资源信息分类汇总，生成系统资源信息。

[0040] 在分布式存储系统中，由于各节点硬件资源异构性无法避免，因此系统中的节点预先接收各其他节点发送的自身资源信息，并根据设定的规则对各类资源进行分类。例如，存储资源分为高、中、低性能等几类，分别提供不同级别的 IO 读写能力。对于具有高速网络设备的存储资源，对该资源标记网络属性为高，对于具有大容量内存的存储资源，对该资源标记内存属性为高等。由于每个节点的每个磁盘的性能数据并不完全相同，需要对每个磁盘性能进行归一化，具体地，当前节点在接收到其他节点发送的资源信息后，其构建的系统资源信息表如下所示：

[0041]

节点 ID	存储目录	存储容量	存储性能特征	其他资源属性
Node1	/data1	128GB	高速 SSD	网络：千兆以太网 MEM：32GB
	/data2	1TB	中速 SAS 盘	
	/data3	2TB	低速 SATA 盘	
Node2	/data1	256GB	高速 SSD	网络：万兆以太网 MEM：64GB
	/data2	1TB	中速 SAS 盘	
	/data3	2TB	低速 SATA 盘	
	/data4	1TB	低速 SATA 盘	

[0042] 表 1. 系统资源信息表

[0043] 步骤 302，接收数据存储请求信息，其中包括基本需求信息以及特殊需求信息，其中的基本需求信息包括但不限于以下内容：存储空间、读写 IO 性能需求；特殊需求信息包

括但不限于以下内容：数据读写特征、与已有数据的操作关系等。

[0044] 步骤 303，根据系统资源信息，生成满足基本性能需求的所有数据存储计划，由于可能会得出关于基本性能需求的多个分配计划，例如，将该数据以 2 个副本分别存储在 2 个节点的高性能磁盘上，或将该数据以 6 个副本分别存储在 6 个节点的低性能磁盘上等，因此每个数据分配计划可以采取如下格式进行标识：

[0045] <数据标识，副本数，<存储节点，存储磁盘号>列表，存储时长>

[0046] 除了基本信息外，上述格式内容还包括了将数据存储为几个副本，每个副本应该存储到哪个存储节点的哪个磁盘上，以及存放的时间等信息。

[0047] 步骤 304，确定每个数据存储计划的初始评分；

[0048] 该步骤中，每个数据存储计划的初始评分可表示如下：Performance=A*（当前计划中存储磁盘的总 IO 吞吐量 / 实际需求 IO 吞吐量）+B*（当前计划中存储所在节点网络带宽量 / 实际需求网络带宽量)+C*（当前计划中存储所在节点计算能力 / 实际需求计算能力)，其中 A、B 和 C 为这几个参数的比例，A+B+C=1。

[0049] 需要说明的是，以上针对数据存储计划所采用的评分计算方案仅仅只是本发明根据实际情况所提出的一个示例，其中的因子“吞吐量”、“网络带宽量”、“计算能力”及其所对应的参数比例在不同的应用场景下可以进行适当的调整或更换其他的因子，这样的改变并不影响本发明的保护范围。

[0050] 步骤 305，根据特殊需求信息，如数据读写特征、该数据与系统中已有数据的操作关系、数据存储时长等，对各个数据存储计划进行调整。

[0051] 在本具体实施例中，分别以数据读写特征、该数据与系统中已有数据的操作关系、数据存储时长三者为例，对该步骤进行详细说明：

[0052] (1)数据读写特征包括：写少读多、写多读少、读写均衡等。如果数据为写少读多，则将采用副本数较多的分配计划的 Performance 数值提高，尽量将数据读的压力分摊给多个节点。如果数据为写多读少，则将采用副本数较多的分配计划的 Performance 数值降低，因为当副本数较多时，保持副本一致性的额外开销会增加。

[0053] (2) 当前写入数据与存储系统中已有数据的操作关系特征包括：当前数据与已有数据是顺序读、当前数据与已有数据是并行读等。对于分布式文件系统一个大文件的多个数据块在进行 MapReduce 计算时，是并行读。对于如 HBase 存放到 HDFS 上的数据，新加入的数据和旧的数据一般呈现顺序读的特征。当当前数据与已有数据是并行读时，为了更好的使多个磁盘并发读取，则将分配计划中数据放置位置与已有数据不在一个磁盘的计划 Performance 提高，从而提高并发读的效率；当当前数据与已有数据是顺序读时，由于当前数据与已有数据不存在 IO 竞争，则将分配计划中数据放置位置与已有数据位于同一个磁盘的计划 Performance 提高。

[0054] (3) 数据存放时长特征：当数据存储时间短时，则对数据放置在频繁擦写对磁盘寿命影响较小的数据存储方案，提高其 Performance。如果存放时间较长，为了避免其过长时间占用高效存储资源，则对存储在低速磁盘且多副本的数据分配计划的 Performance 提高。

[0055] 以上针对各特殊需求信息的处理方法仅为本发明实施例所提出的若干优选实施方式，本领域技术人员可以在此基础上进行其他适应性的特征增加或是调整，这些改进均

属于本发明的保护范围。

[0056] 步骤 306，确定调整后的数据存储计划的性价比，选择性价比最高的数据存储计划。

[0057] 该步骤中，首先采用如下公式计算各数据存储计划的费用：

[0058] Cost= 提供当前 IO 能力所需的存储资源量 * 单位价格 + 提供当前网络带宽所需的网络资源量 * 单位价格 + 提供当前计算能力所需的计算资源量 * 单位价格。

[0059] 根据以上费用 Cost，再以 Performance/Cost 计算各个数据存储计划对应的性价比，选择性价比最高的的数据存储计划。

[0060] 由于每个数据存储计划所对应的性价比最终为一个数值，因此只要针对当前所有数据存储计划所采用的性价比计算方案保持一致的前提下，本领域技术人员也可以采取其他的方式计算各数据存储计划的性价比，这样的改变均属于本发明的保护范围。

[0061] 步骤 307，按照性价比最高的数据存储计划将数据进行存储。

[0062] 同时，本发明还提出了一种数据存储设备，如图 4 所述，该数据存储设备包括：

[0063] 获取模块 410，用于获取接收到的数据存储请求消息中携带的基本需求信息以及特殊需求信息；

[0064] 生成模块 420，用于根据系统资源信息和所述基本需求信息生成至少一个数据存储计划；；

[0065] 确定模块 430，用于根据所述基本需求信息以及特殊需求信息确定所述数据存储计划的综合评分，并根据所述系统资源信息以及所述数据存储计划的综合评分得到所述数据存储计划的性价比；

[0066] 存储模块 440，用于选取性价比最高的数据存储计划，并依据所述性价比最高的数据存储计划存储数据。

[0067] 进一步地，在具体的应用场景中，所述确定模块 430，具体用于：

[0068] 根据所述基本需求信息生成所述数据存储计划的初始评分，并将所述初始评分依据所述特殊需求信息进行调整。

[0069] 进一步地，在具体的应用场景中，所述确定模块 430，具体用于：

[0070] 根据所述基本需求信息中各性能选项所对应的权重因子，以及各性能选项在所述数据存储计划与所述基本需求信息之间的比例，生成所述数据存储计划的初始评分。

[0071] 进一步地，在具体的应用场景中，

[0072] 所述基本需求信息至少包括一种或多种以下性能选项：数据存储空间、数据读写速度、网络带宽；

[0073] 所述特殊需求信息至少包括一种或多种以下性能选项：数据读写特征、待存储数据与系统数据的操作关系、数据存储时间。

[0074] 进一步地，在具体的应用场景中，还包括：处理模块 450，

[0075] 所述处理模块 450，具体用于接收各其他节点发送的节点资源信息，并将所述接收到的节点资源信息按照预设的规则进行分类，生成所述系统资源信息。

[0076] 由此可见，本发明基于数据存储的基本需求以及特殊需求，确定性价比最高的数据存储计划，并根据性价比最高的数据存储计划进行数据存储，保证了数据存储的 QoS 需求。

[0077] 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。

[0078] 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。

[0079] 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。

[0080] 上述本发明序号仅仅为了描述，不代表实施场景的优劣。

[0081] 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。

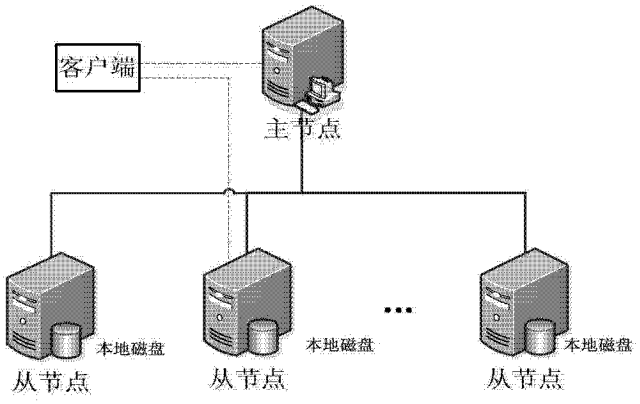


图 1

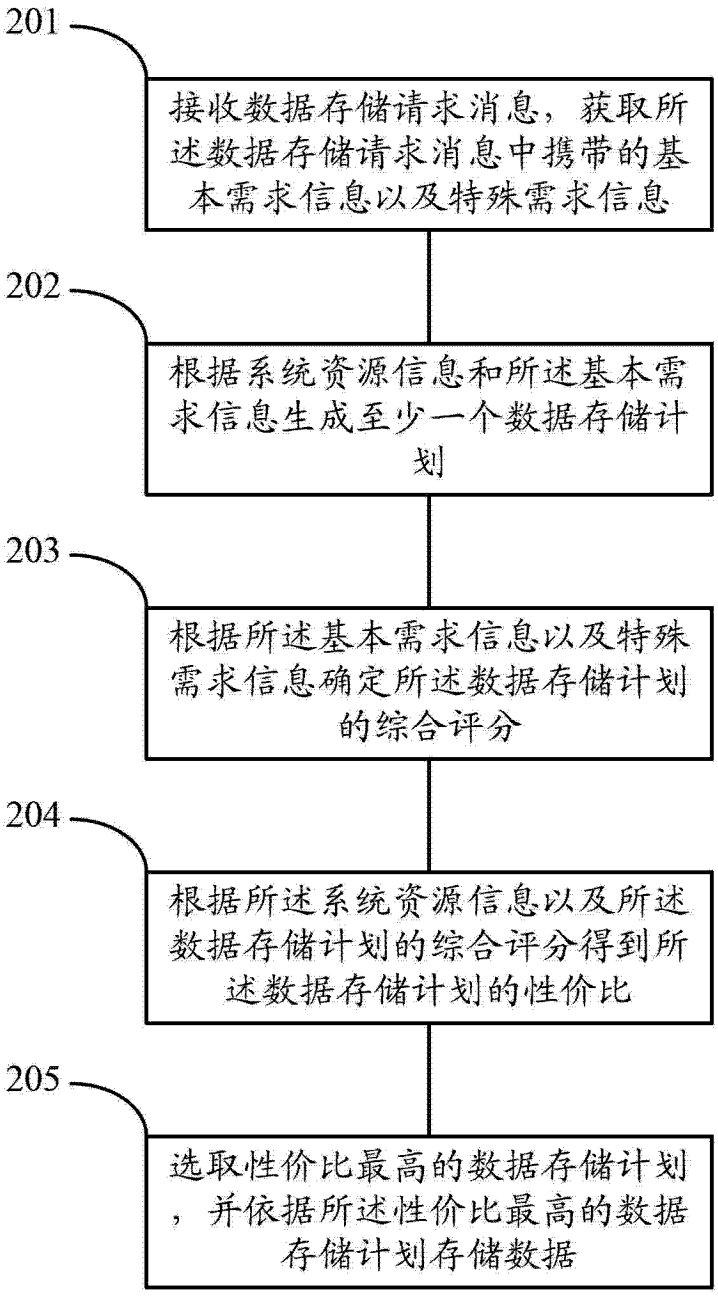

201
接收数据存储请求消息，获取所述数据存储请求消息中携带的基本需求信息以及特殊需求信息
202
根据系统资源信息和所述基本需求信息生成至少一个数据存储计划
203
根据所述基本需求信息以及特殊需求信息确定所述数据存储计划的综合评分
204
根据所述系统资源信息以及所述数据存储计划的综合评分得到所述数据存储计划的性价比
205
选取性价比最高的数据存储计划，并依据所述性价比最高的数据存储计划存储数据

图 2

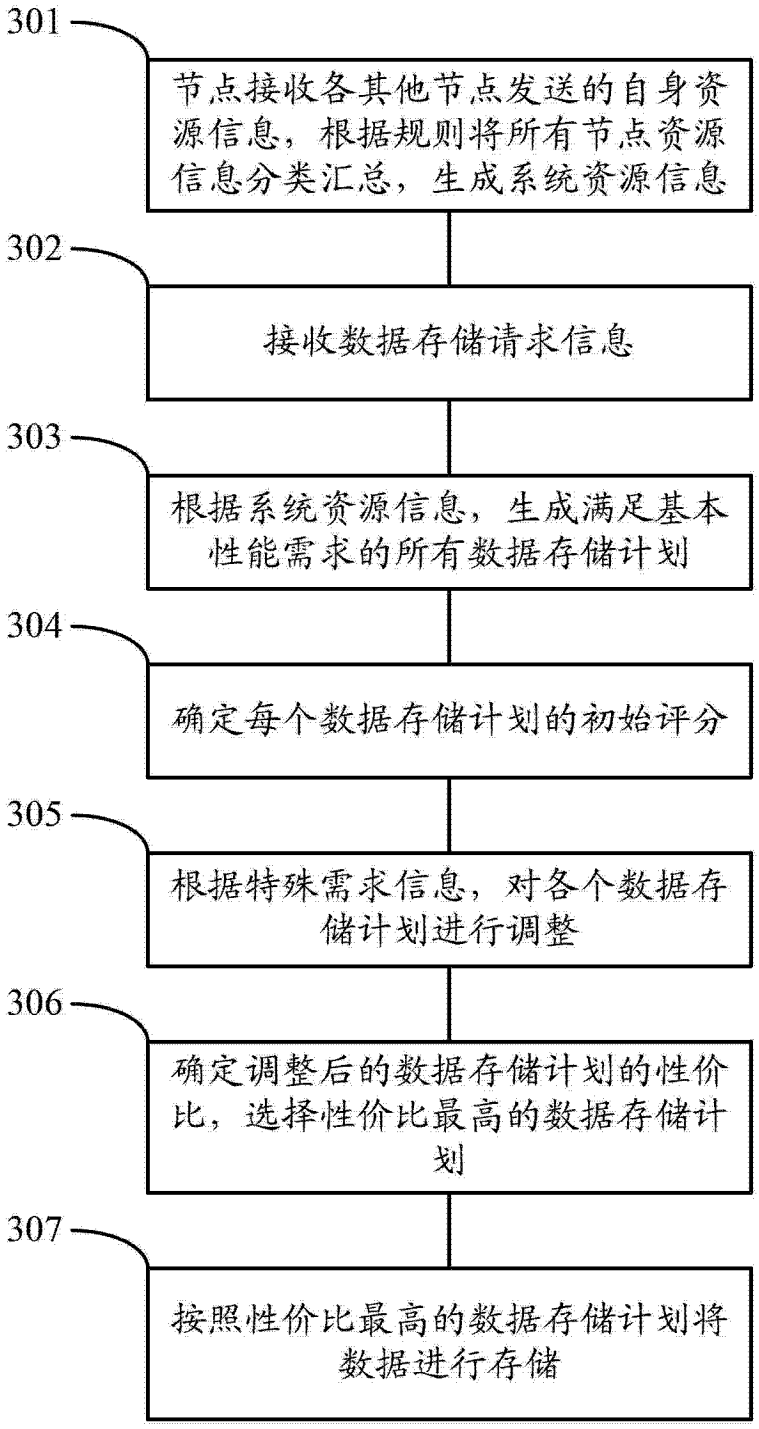

301
节点接收各其他节点发送的自身资源信息，根据规则将所有节点资源信息分类汇总，生成系统资源信息
302
接收数据存储请求信息
303
根据系统资源信息，生成满足基本性能需求的所有数据存储计划
304
确定每个数据存储计划的初始评分
305
根据特殊需求信息，对各个数据存储计划进行调整
306
确定调整后的数据存储计划的性价比，选择性价比最高的数据存储计划
307
按照性价比最高的数据存储计划将数据进行存储

图 3

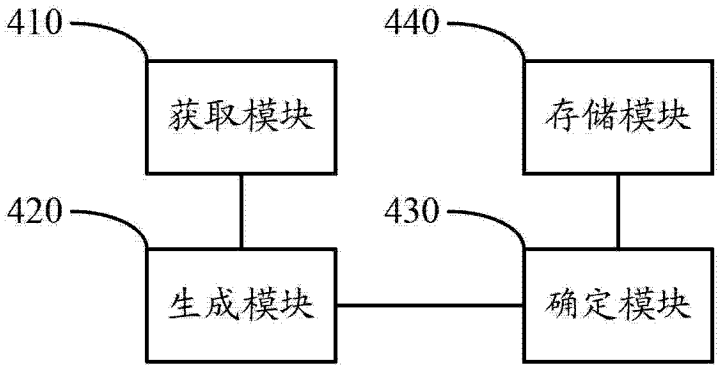


图 4