

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2016 年 6 月 30 日 (30.06.2016)



(10) 国际公布号
WO 2016/101759 A1

- (51) 国际专利分类号:
H04L 29/08 (2006.01)
- (21) 国际申请号: PCT/CN2015/095507
- (22) 国际申请日: 2015 年 11 月 25 日 (25.11.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201410832192.2 2014 年 12 月 27 日 (27.12.2014) CN
- (71) 申请人: 北京奇虎科技有限公司 (BEIJING QIHOO TECHNOLOGY COMPANY LIMITED) [CN/CN]; 中国北京市西城区新街口外大街 28 号 D 座 112 室 (德胜园区), Beijing 100088 (CN)。奇智软件 (北京) 有限公司 (QIZHI SOFTWARE (BEIJING) COMPANY LIMITED) [CN/CN]; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼 B 座 2 层、3 层 301-306 室, Beijing 100015 (CN)。
- (72) 发明人: 陈营 (CHEN, Ying); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。李明昊 (LI, Minghao); 中国北京市朝阳区酒仙桥路 6 号院 2

号楼, Beijing 100015 (CN)。宋昭 (SONG, Shao); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。陈宗志 (CHEN, Zongzhi); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。王超 (WANG, Chao); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。

(74) 代理人: 北京润泽恒知识产权代理有限公司 (BEIJING RISEHIGH INTELLECTUAL PROPERTY LAW FIRM); 中国北京市海淀区中关村南大街 31 号神舟大厦 702, Beijing 100081 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

[见续页]

(54) Title: DATA ROUTING METHOD, DATA MANAGEMENT DEVICE AND DISTRIBUTED STORAGE SYSTEM

(54) 发明名称: 一种数据路由方法、数据管理装置和分布式存储系统

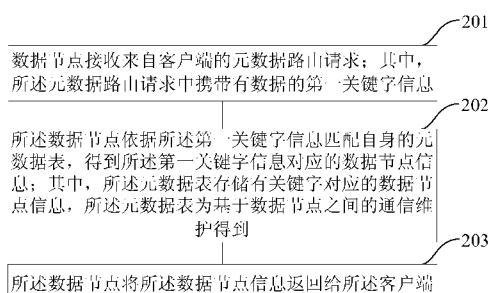


图 2 / Fig.2

- 201 A data node receiving a metadata routing request from a client, wherein the metadata routing request carries first keyword information about data
- 202 The data node obtaining data node information corresponding to the first keyword information according to a metadata table matching the first keyword information itself, wherein the data node information corresponding to the keyword is stored in the metadata table, and the metadata table is obtained based on the maintenance of the communication between data nodes
- 203 The data node returning the data node information to the client

(57) Abstract: Provided are a data routing method, a data management device and a distributed storage system. The method specifically comprises: a data node receiving a metadata routing request from a client, wherein the metadata routing request carries first keyword information about data; the data node obtaining data node information corresponding to the first keyword information according to a metadata table matching the first keyword information itself, wherein the data node information corresponding to the keyword is stored in the metadata table, and the metadata table is obtained based on the maintenance of the communication between data nodes; and the data node returning the data node information to the client. The embodiments of the present invention can guarantee that correct data node information is fed back to the client, thereby increasing the accuracy rate of metadata routing, and furthermore, the operation and maintenance costs of a distributed storage system can be reduced.

(57) 摘要: 本发明实施例提供了一种数据路由方法、数据管理装置和分布式存储系统, 其中的方法具体包括: 数据节点接收来自客户端的元数据路由请求; 其中, 所

述元数据路由请求中携带有数据的第一关键字信息; 所述数据节点依据所述第一关键字信息匹配自身的元数据表, 得到所述第一关键字信息对应的数据节点信息; 其中, 所述元数据表存储有关键字对应的数据节点信息, 所述元数据表为基于数据节点之间的通信维护得到; 以及所述数据节点将所述数据节点信息返回给所述客户端。本发明实施例能够保证向客户端反馈正确的数据节点信息, 提高元数据路由的准确率, 且能够降低分布式存储系统的运维成本。

WO 2016/101759 A1



(84) **指定国** (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ,

CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种数据路由方法、数据管理装置和分布式存储系统

技术领域

本发明涉及分布式存储技术领域，特别是涉及一种数据路由方法、一种数据管理装置和一种分布式存储系统。

5 背景技术

GFS (Google 文件系统, Google File System) 是一个大型的分布式文件系统, 它为 Google 云计算提供海量存储, 并且与 MapReduce (映射化简) 技术结合十分紧密。

参照图 1, 示出了现有技术一种 GFS 的结构示意图, GFS 将整个系统的节点分为三类角色: Client (客户端)、Master (主服务器) 和 Chunk Server (数据块服务器); 其中, Client 是 GFS 提供给应用程序的访问接口; Master 是 GFS 的管理节点, 在逻辑上只有一个, 它保存系统的元数据, 负责整个文件系统的管理; Chunk Server 负责具体的存储工作, 数据以文件的形式存储在 Chunk Server 上, Chunk Server 的个数可以有多个。

15 客户端在访问 GFS 时, 首先向 Master 发送元数据路由请求, 所述元数据路由请求中携带有 key (关键字) 信息, 由 Master 依据存储的元数据表获取 key 对应的 Chunk Server 信息, 然后直接访问这些 Chunk Server 完成数据存取。GFS 的这种设计方法实现了控制流和数据流的分离。Client 与 Master 之间只有控制流, 而无数据流, 这样就极大地降低了 Master 的负载, 使之不
20 成为系统性能的一个瓶颈。Client 与 Chunk Server 之间直接传输数据流, 同时由于文件被分成多个 Chunk 进行分布式存储, Client 可以同时访问多个 Chunk Server, 从而使得整个系统的 I/O 高度并行, 系统整体性能得到提高。

然而, 网络抖动、节点故障等因素容易导致 key 对应的 Chunk Server 发生变化, 而 Master 作为管理节点不能及时获知上述情况下元数据的变化, 因此, 不能保证向客户端反馈正确的 Chunk Server 信息。通常情况下, 客户端
25 在元数据请求失败时, 还需要向 Master 发送更新元数据表的请求, 并持续等待 Master 依据更新后元数据表返回的 Chunk Server 信息。

发明内容

鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的一种数据路由方法、一种数据管理装置和一种分布式存储系统。

5 依据本发明的一个方面，提供了一种数据路由方法，包括：

数据节点接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

10 所述数据节点依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；以及

所述数据节点将所述数据节点信息返回给所述客户端。

15 根据本发明的另一方面，提供了一种计算机程序，包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行如上文所述的数据路由方法。

根据本发明的再一方面，提供了一种计算机可读介质，其中存储了如上文所述的计算机程序。

根据本发明的又一方面，提供了一种数据管理装置，包括：

20 第一接收模块，配置为接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

第一匹配模块，配置为依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；以及

25 第一返回模块，配置为将所述数据节点信息返回给所述客户端。

根据本发明实施例的一种数据路由方法、一种数据管理装置和一种分布式存储系统，数据节点可以利用自身存储的元数据表来处理客户端的元数据

路由请求；由于数据节点可以基于数据节点之间的通信进行元数据表的维护，使得维护的元数据表能够及时反映节点状态的变化，因此，能够保证向客户端反馈正确的数据节点信息，提高元数据路由的准确率；并且，相对于现有技术中采用专门的 Master 来存储并维护元数据表获取，本发明实施例减少了 Master 这一角色，因此能够降低分布式存储系统的运维部署成本。

上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

通过阅读下文可选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出可选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

图 1 示出了现有技术一种 GFS 的结构示意图；

图 2 示出了根据本发明一个实施例的一种数据路由方法的步骤流程示意图；

图 3 示出了根据本发明一个示例的一种分布式存储系统的结构示意图；

图 4 示出了根据本发明一个实施例的一种数据路由方法的步骤流程示意图；

图 5 示出了根据本发明一个实施例的一种数据路由方法的步骤流程示意图；

图 6 示出了根据本发明一个实施例的一种数据节点基于数据节点之间的通信对所述元数据表进行维护的步骤流程示意图；

图 7 示出了根据本发明一个实施例的一种数据路由方法的步骤流程示意图；

图 8 示出了根据本发明一个实施例的一种数据管理装置的结构示意图；

图 9 示出了根据本发明一个实施例的一种数据管理装置的结构示意图；

图 10 示出了根据本发明一个实施例的一种数据管理装置的结构示意图；

图 11 示出了根据本发明一个实施例的一种分布式存储系统的结构示意图；

图 12 示意性地示出了用于执行根据本发明的方法的计算设备的框图；
以及

5 图 13 示意性地示出了用于保持或者携带实现根据本发明的方法的程序代码的存储单元。

具体实施方式

下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不
10 应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

参照图 2，示出了根据本发明一个实施例的一种数据路由方法的步骤流程图示意图，具体可以包括如下步骤：

15 步骤 201、数据节点接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

步骤 202、所述数据节点依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关
20 关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；

步骤 203、所述数据节点将所述数据节点信息返回给所述客户端。

本发明实施例可以应用于各种业务的分布式存储系统中，用于提高元数据路由的准确率。

参照图 3，示出了根据本发明一个示例的一种分布式存储系统的结构示意图，具体可以包括：客户端 301 和数据节点 302；其中，客户端 301 作为
25 业务请求的发起方，可以通过元数据路由请求从所述数据节点 302 获取所述元数据路由请求对应的数据节点信息，从而可以访问所述数据节点信息对应数据节点 302 以完成数据存取操作；数据节点 302 上可以存储有元数据表，

且数据节点 302 之间可以通信,从而可以基于数据节点 302 之间的通信维护所述元数据表,以保证所述元数据表能够及时反映节点状态的变化。

在本发明的一种可选实施例中,所述数据节点依据所述第一关键字信息匹配自身的元数据表,得到所述第一关键字信息对应的数据节点信息的步骤,具体可以包括:

子步骤 A1、计算所述第一关键字信息的哈希值;

子步骤 A2、依据所述哈希值匹配所述元数据表,得到所述哈希值对应数据节点信息。

在具体实现中,可以采用单向散列算法等哈希算法计算所述第一关键字信息的哈希值,本发明实施例对具体的哈希算法不加以限制。

在实际应用中,所述数据节点信息具体可以包括如下信息中的一项或多项:节点编号信息、节点属性信息和节点通信速率信息。其中,节点属性信息具体可以包括:节点可用或节点不可用等信息,节点通信速率信息可以为节点的通信速率值等等。

在本发明的一种应用示例中,所述哈希值对应数据节点的节点编号信息可以为所述数据节点自身的编号,也可以为其它数据节点的编号,无论何种情况,所述数据节点均可以将所述哈希值对应数据节点的节点编号返回给客户端,以满足客户端进一步的节点访问请求。

综上,数据节点可以利用自身存储的元数据表来处理客户端的元数据路由请求;由于数据节点可以基于数据节点之间的通信进行元数据表的维护,使得维护的元数据表能够及时反映节点状态的变化,因此,能够保证向客户端反馈正确的数据节点信息,提高元数据路由的准确率;并且,相对于现有技术中采用专门的 Master 来存储并维护元数据表获取,本发明实施例减少了 Master 这一角色,因此能够降低分布式存储系统的运维部署成本。

参照图 4,示出了根据本发明一个实施例的一种数据路由方法的步骤流程图示意图,具体可以包括如下步骤:

步骤 401、数据节点接收来自客户端的元数据路由请求;其中,所述元

数据路由请求中携带有数据的第一关键字信息；

步骤 402、所述数据节点依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护
5 得到；

步骤 403、所述数据节点将所述数据节点信息返回给所述客户端；

步骤 404、当所述第一关键字信息对应的数据节点信息非所述数据节点自身时，所述数据节点将所述元数据表返回给所述客户端。

相对于图 2 所示实施例，本实施例在所述第一关键字信息对应的数据节点信息非所述数据节点自身时，所述数据节点还可以将所述元数据表返回给
10 所述客户端，也即，可以向客户端提供当前最新的元数据表，方便客户端通过查询自身的元数据表来进行元数据的路由。

对于客户端而言，其可以通过如下两种方式进行元数据的路由：方式一为通过向数据节点发送元数据路由请求来进行元数据的路由，方式二为通过
15 查询自身的元数据表来进行元数据的路由，其中方式一具有准确率高的优点，方式二具有节省流量的优点。可以理解，本领域技术人员可以根据实际需求确定采用上述两种方式中的任一或全部，例如，在对准确率要求比较严格时，可以采用上述方式一，又如，在对流量要求比较严格时，可以采用上述方式二，或者，为了保证路由的成功率，可以同时采用上述方式一和方式
20 二等等。

参照图 5，示出了根据本发明一个实施例的一种数据路由方法的步骤流程图示意图，具体可以包括如下步骤：

步骤 501、数据节点接收来自客户端的元数据路由请求；其中，所述元
25 数据路由请求中携带有数据的第一关键字信息；

步骤 502、所述数据节点依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护

得到;

步骤 503、所述数据节点将所述数据节点信息返回给所述客户端;

步骤 504、所述数据节点基于数据节点之间的通信对所述元数据表进行维护。

5 参照图 6, 示出了根据本发明一个实施例的一种数据节点基于数据节点之间的通信对所述元数据表进行维护的步骤流程图, 具体可以包括如下步骤:

子步骤 541、协调者节点向所有执行事务 T 的参与者节点发送准备消息;

子步骤 542、各参与者节点确定是否提交事务 T, 若是, 则向所述协调者节点返回准备就绪消息, 否则, 向所述协调者节点返回异常中止消息;

子步骤 543、当从所有参与者节点获得的消息均为准备就绪消息时, 所述协调者节点向所有参与者节点发出正式提交消息;

子步骤 544、在接收到正式提交消息后, 各参与者节点正式完成事务 T, 释放在整个事务 T 期间内占用的, 并向协调者节点发送完成消息;

15 子步骤 545、所述协调者节点在收到所有参与者节点反馈的完成消息后, 完成事务;

子步骤 546、当所述协调者节点从所有参与者节点获得的消息中存在异常中止消息, 或者, 在超时之前无法获取所有参与者节点的响应消息时, 所述协调者节点向所有参与者节点发出回滚消息;

20 这里, 超时可以表示从发出正式提交消息到现在超出了预置周期。

子步骤 547、各参与者节点在接收到所述回滚消息后, 对事务 T 执行回滚操作, 释放在整个 T 事务期间内占用的资源, 并向所述协调者节点发送回滚完成消息;

子步骤 548、协调者节点接收到所有参与者节点反馈的回滚完成消息后, 取消事务 T。

25 相对于图 1 所示实施例, 本实施例的数据节点具体可以包括: 协调者节点和参与者节点, 并且, 增加了所述数据节点基于数据节点之间的通信对所述元数据表进行维护; 其中, 上述子步骤 541-子步骤 548 涉及的元数据维护

过程具体通过两阶段提交协议来保证所有参与者节点所维护元数据表的一致性和完整性。

两阶段提交协议的主要思想可以为：当一个事务 T 要对多个数据库进行操作，必须确保该多个参与者节点的数据库都提交成功，事务 T 才能成功，
5 所以，协调节点可以先对多个参与者节点发出预提交，多个参与者节点返回是否可以提交，如果所有参与者节点都能提交，则协调节点可以正式提交事务 T。

上述子步骤 541-子步骤 548 涉及的元数据维护过程中，子步骤 541-子步骤 542 为第一阶段的步骤，子步骤 543-子步骤 548 为第二阶段的步骤。其中，
10 在第一阶段，协调者节点将通知事务的参与者节点准备提交或取消事务，由参与者节点向协调者节点告知自己的决策：同意（准备就绪消息）或取消（异常中止消息）；在第二阶段，协调者节点将基于所有参与者节点反馈的信息进行决策：提交或取消，当且仅当所有的参与者节点同意提交事务时，协调者节点才通知所有的参与者节点提交事务，否则协调者节点将通知所有的参与者节点取消事务。
15

需要说明的是，上述通过两阶段提交协议来保证所有参与者节点所维护元数据表的一致性和完整性的方案只是作为保证所有参与者节点所维护元数据表的一致性和完整性的可选方案，可以理解，本领域技术人员还可以根据实际需要，采用其它方案来保证所有参与者节点所维护元数据表的一致性和完整性，如三阶段提交协议等等，本发明实施例对保证所有参与者节点所
20 维护元数据表的一致性和完整性的具体方案不加以限制。

参照图 7，示出了根据本发明一个实施例的一种数据路由方法的步骤流程图，具体可以包括如下步骤：

25 步骤 701、数据节点接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

步骤 702、所述数据节点依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有

关键字对应的数据节点信息,所述元数据表为基于数据节点之间的通信维护得到;

步骤 703、所述数据节点将所述数据节点信息返回给所述客户端;

5 步骤 704、所述数据节点接收来自客户端的读请求;其中,所述读请求中携带有数据的第二关键字信息;

步骤 705、所述数据节点依据所述读请求携带的第二关键字信息匹配自身的元数据表,得到所述第二关键字信息对应的数据节点信息;

步骤 706、所述数据节点依据所述第二关键字信息对应的数据节点信息,判定所述读请求对应数据是否在所述数据节点自身;

10 步骤 707、当所述第二关键字信息对应的数据节点信息为所述数据节点自身时,所述数据节点依据所述读请求查询自身的数据引擎,并将查询得到的数据返回给所述客户端;

15 步骤 708、当所述第二关键字信息对应的数据节点信息非所述数据节点自身时,将所述读请求转发给第二关键字信息对应的数据节点信息所对应的第一数据节点;

步骤 709、接收所述第一数据节点返回的所述读请求对应的数据,并返回给所述客户端。

20 相对于图 1 所示实施例,本实施例除了可以通过执行步骤 701-步骤 703 来处理来自客户端的元数据路由请求外,还可以通过步骤 704-步骤 709 处理来自客户端的读请求,特别地,在处理所述读请求的过程中,当所述读请求中所携带第二关键字信息对应的数据节点信息非所述数据节点自身时,所述数据节点还可以担当网络代理的角色,也即,将所述读请求转发给第二关键字信息对应的数据节点信息所对应的第一数据节点,所述第一数据节点返回的所述读请求对应的数据,并返回给所述客户端;所述代理客户端转发读请求并向客户端返回读出数据的过程,能够避免客户端尝试向多个不同的数据节点发送读请求,从而节省了客户端的流量。

25 需要说明的是,数据节点除了可以代理客户端转发读请求并向客户端返回读出数据外,还可以代理客户端转发写请求并向客户端返回响应结果,由

于代理客户端转发写请求的过程与代理客户端转发读请求的过程类似，故在此不作赘述，相互参照即可。

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，本发明实施例并不受所描述的动作顺序的限制，因为依据本发明实施例，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可选实施例，所涉及的动作并不一定是本发明实施例所必须的。

参照图 8，示出了根据本发明一个实施例的一种数据管理装置的结构示意图，具体可以包括如下模块：

第一接收模块 801，配置为接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

第一匹配模块 802，配置为依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；以及

第一返回模块 803，配置为将所述数据节点信息返回给所述客户端。

在实际应用中，上述数据管理装置可以为数据节点本身，也可以为管理各数据节点的装置，本发明实施例对上述数据管理装置的具体位置不加以限制。

在本发明的一种可选实施例中，所述第一匹配模块 802，具体可以包括：哈希计算子模块，配置为计算所述第一关键字信息的哈希值；及哈希匹配子模块，配置为依据所述哈希值匹配所述元数据表，得到所述哈希值对应数据节点信息。

在本发明的另一种可选实施例中，所述数据节点还可以包括：

第二返回模块，配置为当所述第一关键字信息对应的数据节点信息非所述数据节点自身时，所述数据节点将所述元数据表返回给所述客户端。

参照图 9，示出了根据本发明一个实施例的一种数据管理装置的结构示意图，具体可以包括：协调者节点 901 和参与者节点 902；

其中，所述协调者节点 901 具体可以包括：配置为基于数据节点之间的通信对所述元数据表进行维护的第一维护模块 911；

5 所述参与者节点 902 具体可以包括如下模块：

 第一接收模块 921，配置为接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

 第一匹配模块 922，配置为依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存
10 储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；

 第一返回模块 923，配置为将所述数据节点信息返回给所述客户端；及

 第二维护模块 924，配置为基于数据节点之间的通信对所述元数据表进行维护的；

15 其中，所述第一维护模块 911，具体可以包括：

 准备发送子模块 9111，配置为向所有执行事务 T 的参与者节点发送准备消息；

 正式提交发送子模块 9112，配置为当从所有参与者节点获得的消息均为准备就绪消息时，向所有参与者节点发出正式提交消息；

20 事务完成子模块 9113，配置为在收到所有参与者节点反馈的完成消息后，完成事务；

 回滚发送子模块 9114，配置为当从所有参与者节点获得的消息中存在异常中止消息，或者，在超时之前无法获取所有参与者节点的响应消息时，所述协调者节点向所有参与者节点发出回滚消息；及

25 事务取消子模块 9115，配置为在接收到所有参与者节点反馈的回滚完成消息后，取消事务 T；

 其中，所述第二维护模块 924，具体可以包括：

 准备响应子模块 9241，配置为确定是否提交事务 T，若是，则向所述协

调者节点返回准备就绪消息，否则，向所述协调者节点返回异常中止消息；

事务执行子模块 9242，配置为在接收到正式提交消息后，正式完成事务 T，释放在整个事务 T 期间内占用的，并向协调者节点发送完成消息；及

回滚子模块 9243，配置为各参与者节点在接收到所述回滚消息后，对事务 T 执行回滚操作，释放在整个 T 事务期间内占用的资源，并向所述协调者节点发送回滚完成消息。

参照图 10，示出了根据本发明一个实施例的一种数据管理装置的结构示意图，具体可以包括如下模块：

10 第一接收模块 1001，配置为接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

第一匹配模块 1002，配置为依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；

15 第一返回模块 1003，配置为将所述数据节点信息返回给所述客户端；

第二接收模块 1004，配置为接收来自客户端的读请求；其中，所述读请求中携带有数据的第二关键字信息；

20 第二匹配模块 1005，配置为依据所述读请求携带的第二关键字信息匹配自身的元数据表，得到所述第二关键字信息对应的数据节点信息；

判定模块 1006，配置为依据所述第二关键字信息对应的数据节点信息，判定所述读请求对应数据是否在所述数据节点自身；

25 查询模块 1007，配置为当所述第二关键字信息对应的数据节点信息为所述数据节点自身时，所述数据节点依据所述读请求查询自身的数据引擎，并将查询得到的数据返回给所述客户端；

转发模块 1008，配置为当所述第二关键字信息对应的数据节点信息非所述数据节点自身时，将所述读请求转发给第二关键字信息对应的数据节点信息所对应的第一数据节点；以及

第三返回模块 1009,配置为接收所述第一数据节点返回的所述读请求对应的数据,并返回给所述客户端。

5 本发明还提供了一种分布式存储系统,其具体可以包括:客户端和前述的数据管理装置。

参照图 11,示出了根据本发明一个实施例的一种分布式存储系统的结构示意图,具体可以包括:客户端 1101 和多个数据节点 1102;

其中,所述数据节点 1102 具体可以包括如下模块:

10 第一接收模块 1121,配置为接收来自客户端的元数据路由请求;其中,所述元数据路由请求中携带有数据的第一关键字信息;

第一匹配模块 1122,配置为依据所述第一关键字信息匹配自身的元数据表,得到所述第一关键字信息对应的数据节点信息;其中,所述元数据表存储有关键字对应的数据节点信息,所述元数据表为基于数据节点之间的通信维护得到;以及

15 第一返回模块 1123,配置为将所述数据节点信息返回给所述客户端。

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。

20 本发明的各个部件实施例可以以硬件实现,或者以在一个或者多个处理器上运行的软件模块实现,或者以它们的组合实现。本领域的技术人员应当理解,可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明实施例的数据路由方法、数据管理装置和分布式存储系统中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如,计算机程序和计算机程序产品)。这样的实现本发明的程序可以存储在计算机可读介质上,或者

25 可以具有一个或者多个信号的形式。这样的信号可以从因特网平台上下下载得到,或者在载体信号上提供,或者以任何其他形式提供。

例如,图 12 示出了可以实现根据本发明上述方法的计算设备,例如搜

索引服务器。该计算设备传统上包括处理器 1210 和以存储器 1230 形式的计算机程序产品或者计算机可读介质。存储器 1230 可以是诸如闪存、EEPROM（电可擦除可编程只读存储器）、EPROM、硬盘或者 ROM 之类的电子存储器。存储器 1230 具有存储用于执行上述方法中的任何方法步骤的程序代码 1251 的存储空间 1250。例如，存储程序代码的存储空间 1250 可以包括分别用于实现上面的方法中的各种步骤的各个程序代码 1251。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。这些计算机程序产品包括诸如硬盘，紧致盘（CD）、存储卡或者软盘之类的程序代码载体。这样的计算机程序产品通常为例如图 13 所示的便携式或者固定存储单元。该存储单元可以具有与图 12 的计算设备中的存储器 1230 类似布置的存储段、存储空间等。程序代码可以例如以适当形式进行压缩。通常，存储单元包括用于执行根据本发明的方法步骤的计算机可读代码 1251’，即可以由诸如 1210 之类的处理器读取的代码，当这些代码由服务器运行时，导致该服务器执行上面所描述的方法中的各个步骤。

本文中所称的“一个实施例”、“实施例”或者“一个或者多个实施例”意味着，结合实施例描述的特定特征、结构或者特性包括在本发明的至少一个实施例中。此外，请注意，这里“在一个实施例中”的词语例子不一定全指同一个实施例。

在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下被实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。

在列举了若干装置的单元权利要求中，这些装置中的若干个可以通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

此外，还应当注意，本说明书中使用的语言主要是为了可读性和教导的目的而选择的，而不是为了解释或者限定本发明的主题而选择的。因此，在不偏离所附权利要求书的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。对于本发明的范围，对本发明所做的公开是说明性的，而非限制性的，本发明的范围由所附权利要求书限定。

权 利 要 求 书

1、一种数据路由方法，包括：

数据节点接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

所述数据节点依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；以及

所述数据节点将所述数据节点信息返回给所述客户端。

2、如权利要求 1 所述的方法，其中，所述方法还包括：

当所述第一关键字信息对应的数据节点信息非所述数据节点自身时，所述数据节点将所述元数据表返回给所述客户端。

3、如权利要求 1 或 2 所述的方法，其中，所述数据节点包括：协调者节点和参与者节点；

则所述方法还包括：所述数据节点基于数据节点之间的通信对所述元数据表进行维护；

其中，所述数据节点基于数据节点之间的通信对所述元数据表进行维护的步骤，包括：

所述协调者节点向所有执行事务 T 的参与者节点发送准备消息；

各参与者节点确定是否提交事务 T，若是，则向所述协调者节点返回准备就绪消息，否则，向所述协调者节点返回异常中止消息；

当从所有参与者节点获得的消息均为准备就绪消息时，所述协调者节点向所有参与者节点发出正式提交消息；

在接收到正式提交消息后，各参与者节点正式完成事务 T，释放在整个事务 T 期间内占用的，并向协调者节点发送完成消息；

所述协调者节点在收到所有参与者节点反馈的完成消息后，完成事务；

当从所有参与者节点获得的消息中存在异常中止消息，或者，在超时之前无法获取所有参与者节点的响应消息时，所述协调者节点向所有参与者节点发出回滚消息；

各参与者节点在接收到所述回滚消息后,对事务 T 执行回滚操作,释放在整个 T 事务期间内占用的资源,并向所述协调者节点发送回滚完成消息;

协调者节点接收到所有参与者节点反馈的回滚完成消息后,取消事务 T。

4、如权利要求 1 或 2 或 3 所述的方法,其中,所述方法还包括:

所述数据节点接收来自客户端的读请求;其中,所述读请求中携带有数据的第二关键字信息;

所述数据节点依据所述读请求携带的第二关键字信息匹配自身的元数据表,得到所述第二关键字信息对应的数据节点信息;

所述数据节点依据所述第二关键字信息对应的数据节点信息,判定所述读请求对应数据是否在所述数据节点自身;

当所述第二关键字信息对应的数据节点信息为所述数据节点自身时,所述数据节点依据所述读请求查询自身的数据引擎,并将查询得到的数据返回给所述客户端;

当所述第二关键字信息对应的数据节点信息非所述数据节点自身时,将所述读请求转发给第二关键字信息对应的数据节点信息所对应的第一数据节点;

接收所述第一数据节点返回的所述读请求对应的数据,并返回给所述客户端。

5、如权利要求 1 或 2 或 3 所述的方法,其中,所述数据节点依据所述第一关键字信息匹配自身的元数据表,得到所述第一关键字信息对应的数据节点信息的步骤,包括:

计算所述第一关键字信息的哈希值;

依据所述哈希值匹配所述元数据表,得到与所述哈希值对应的数据节点的数据节点信息。

6、如权利要求 1 或 2 或 3 所述的方法,其中,所述数据节点信息包括如下信息中的一项或多项:节点编号信息、节点属性信息和节点通信速率信息。

7、一种计算机程序,包括计算机可读代码,当所述计算机可读代码在

计算设备上运行时，导致所述计算设备执行根据权利要求 1 至 6 中的任一项所述的数据路由方法。

8、一种计算机可读介质，其中存储了如权利要求 7 所述的计算机程序。

9、一种数据管理装置，包括：

第一接收模块，配置为接收来自客户端的元数据路由请求；其中，所述元数据路由请求中携带有数据的第一关键字信息；

第一匹配模块，配置为依据所述第一关键字信息匹配自身的元数据表，得到所述第一关键字信息对应的数据节点信息；其中，所述元数据表存储有关键字对应的数据节点信息，所述元数据表为基于数据节点之间的通信维护得到；以及

第一返回模块，配置为将所述数据节点信息返回给所述客户端。

10、如权利要求 9 所述的数据管理装置，其中，所述数据管理装置还包括：

第二返回模块，配置为当所述第一关键字信息对应的数据节点信息非所述数据节点自身时，所述数据节点将所述元数据表返回给所述客户端。

11、如权利要求 9 或 10 所述的数据管理装置，其中，所述数据管理装置包括：协调者节点和参与者节点；

则所述协调者节点包括：配置为基于数据节点之间的通信对所述元数据表进行维护的第一维护模块；

所述参与者节点包括：前述第一接收模块、第一匹配模块、第一返回模块、以及配置为基于数据节点之间的通信对所述元数据表进行维护的第二维护模块；

其中，所述第一维护模块，包括：

准备发送子模块，配置为向所有执行事务 T 的参与者节点发送准备消息；

正式提交发送子模块，配置为当从所有参与者节点获得的消息均为准备就绪消息时，向所有参与者节点发出正式提交消息；

事务完成子模块，配置为在收到所有参与者节点反馈的完成消息后，完成事务；

回滚发送子模块，配置为当从所有参与者节点获得的消息中存在异常中止消息，或者，在超时之前无法获取所有参与者节点的响应消息时，所述协调者节点向所有参与者节点发出回滚消息；及

事务取消子模块，配置为在接收到所有参与者节点反馈的回滚完成消息后，取消事务 T；

其中，所述第二维护模块，包括：

准备响应子模块，配置为确定是否提交事务 T，若是，则向所述协调者节点返回准备就绪消息，否则，向所述协调者节点返回异常中止消息；

事务执行子模块，配置为在接收到正式提交消息后，正式完成事务 T，释放在整个事务 T 期间内占用的，并向协调者节点发送完成消息；及

回滚子模块，配置为各参与者节点在接收到所述回滚消息后，对事务 T 执行回滚操作，释放在整个 T 事务期间内占用的资源，并向所述协调者节点发送回滚完成消息。

12、如权利要求 9 或 10 或 11 所述的数据管理装置，其中，所述数据管理装置还包括：

第二接收模块，配置为接收来自客户端的读请求；其中，所述读请求中携带有数据的第二关键字信息；

第二匹配模块，配置为依据所述读请求携带的第二关键字信息匹配自身的元数据表，得到所述第二关键字信息对应的数据节点信息；

判定模块，配置为依据所述第二关键字信息对应的数据节点信息，判定所述读请求对应数据是否在所述数据节点自身；

查询模块，配置为当所述第二关键字信息对应的数据节点信息为所述数据节点自身时，所述数据节点依据所述读请求查询自身的数据引擎，并将查询得到的数据返回给所述客户端；

转发模块，配置为当所述第二关键字信息对应的数据节点信息非所述数据节点自身时，将所述读请求转发给第二关键字信息对应的数据节点信息所

对应的第一数据节点；

第三返回模块，配置为接收所述第一数据节点返回的所述读请求对应的数据，并返回给所述客户端。

13、如权利要求 9 或 10 或 11 所述的数据管理装置，其中，所述第一匹配模块，包括：

哈希计算子模块，配置为计算所述第一关键字信息的哈希值；及

哈希匹配子模块，配置为依据所述哈希值匹配所述元数据表，得到与所述哈希值对应的数据节点的数据节点信息。

14、如权利要求 9 或 10 或 11 所述的数据管理装置，其中，所述数据节点信息包括如下信息中的一项或多项：节点编号信息、节点属性信息和节点通信速率信息。

15、一种分布式存储系统，包括：客户端和前述权利要求 9 至 14 中任一所述的数据管理装置。

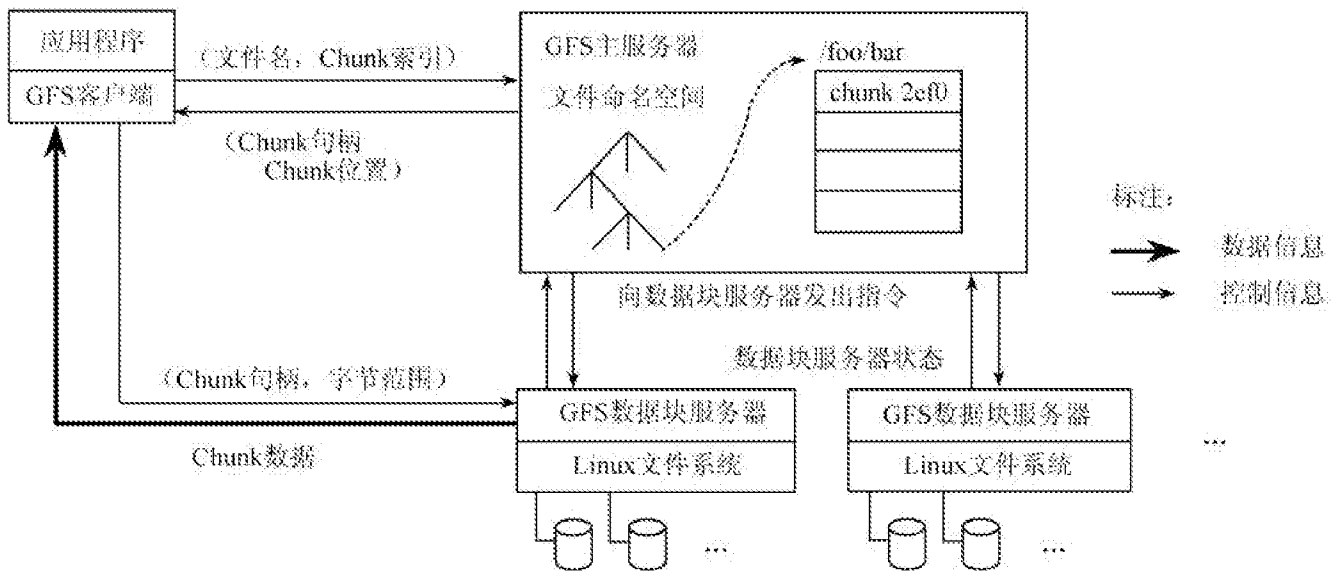


图 1

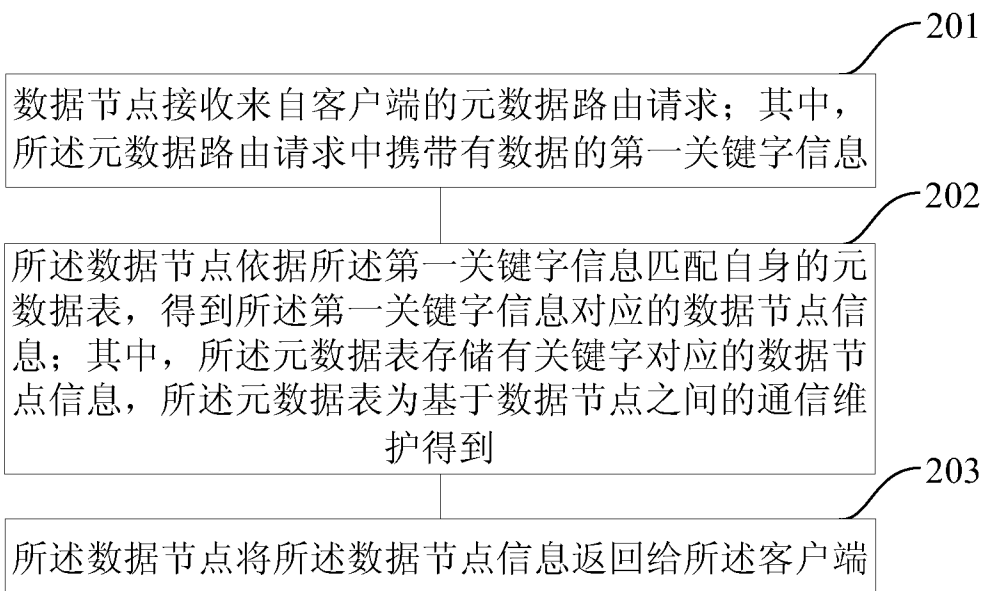


图 2

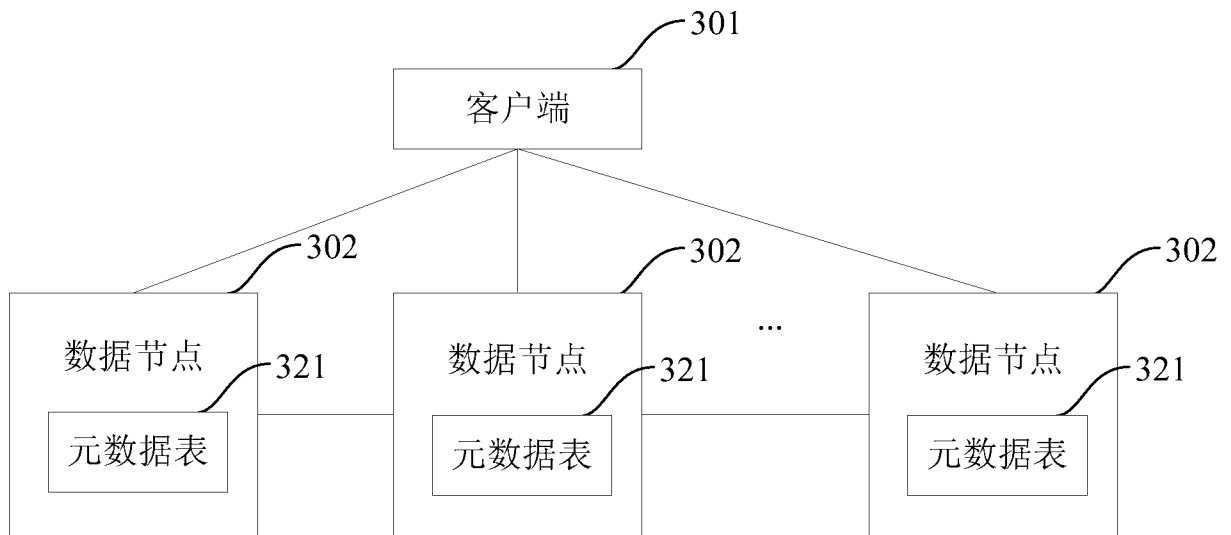


图 3

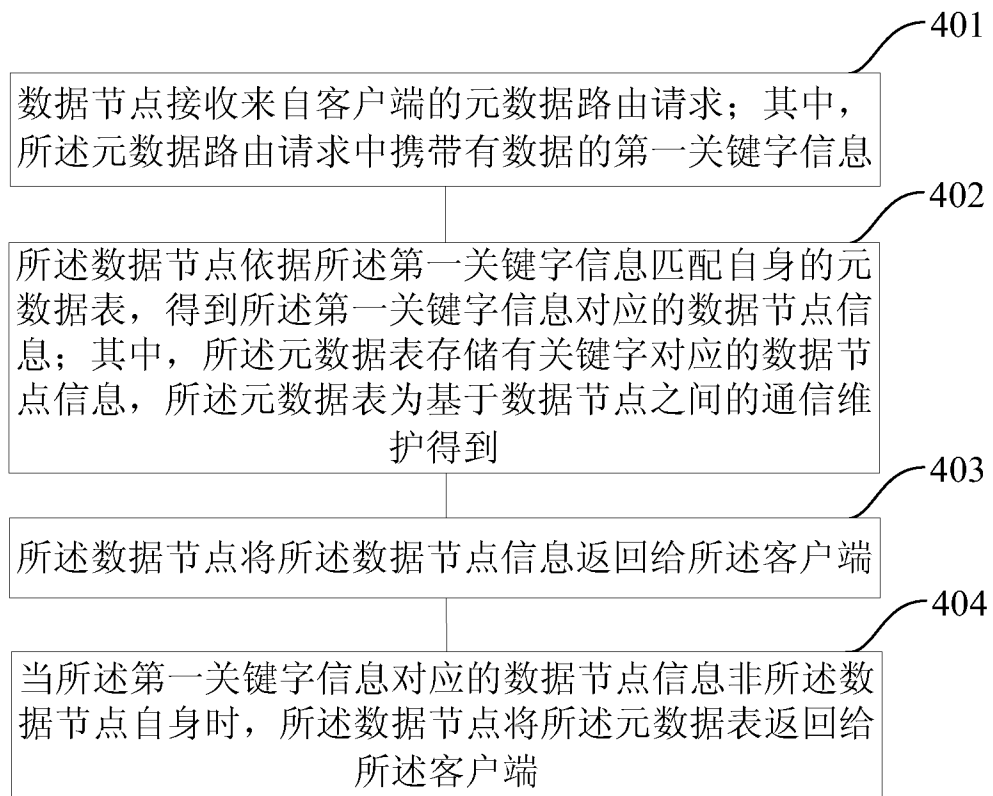


图 4

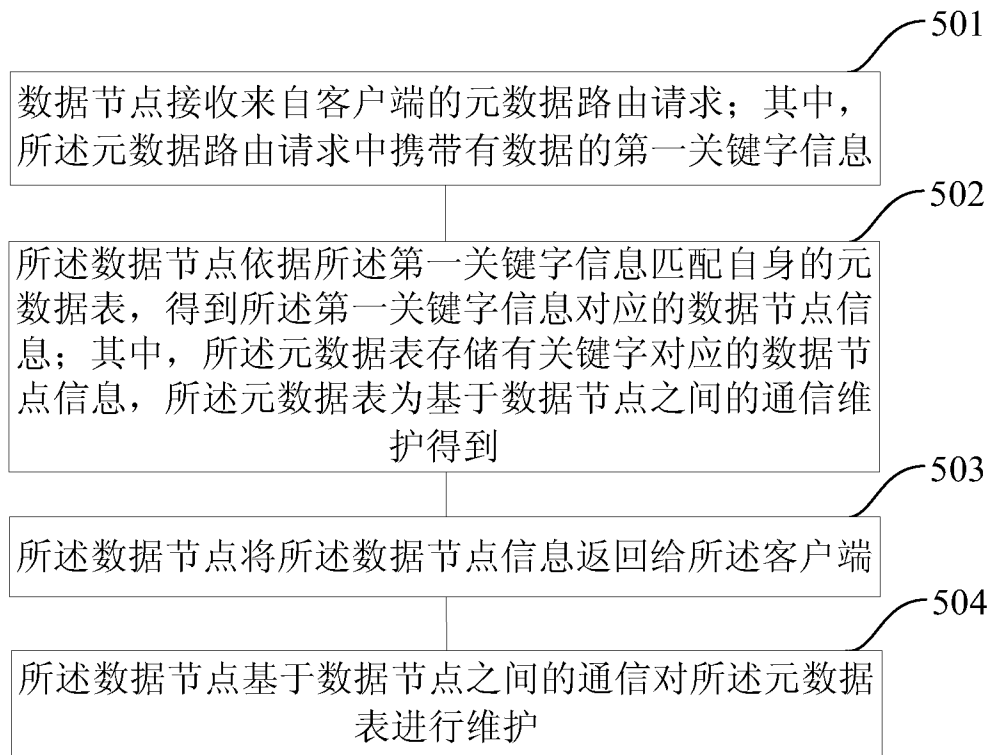


图 5

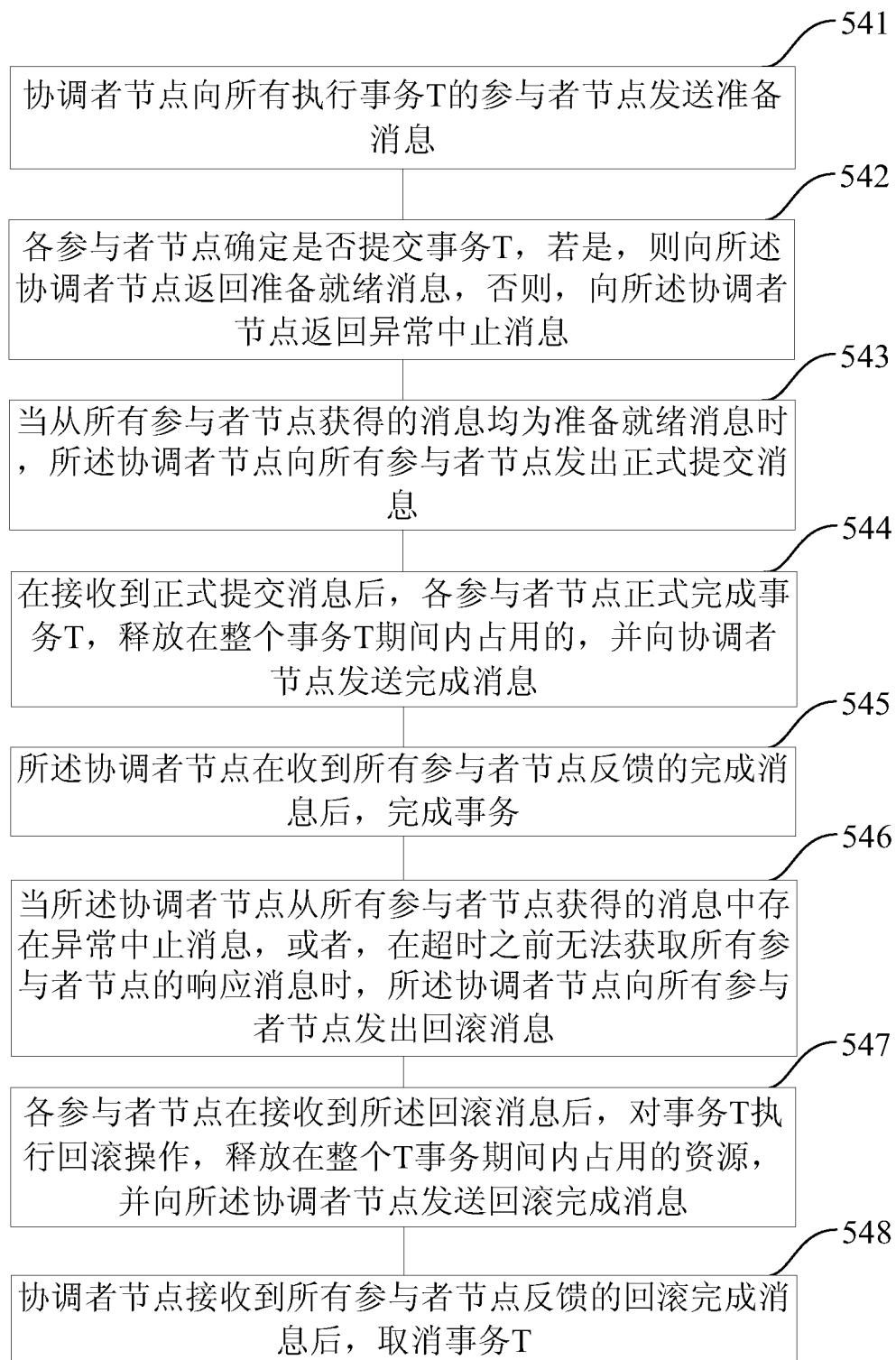


图 6

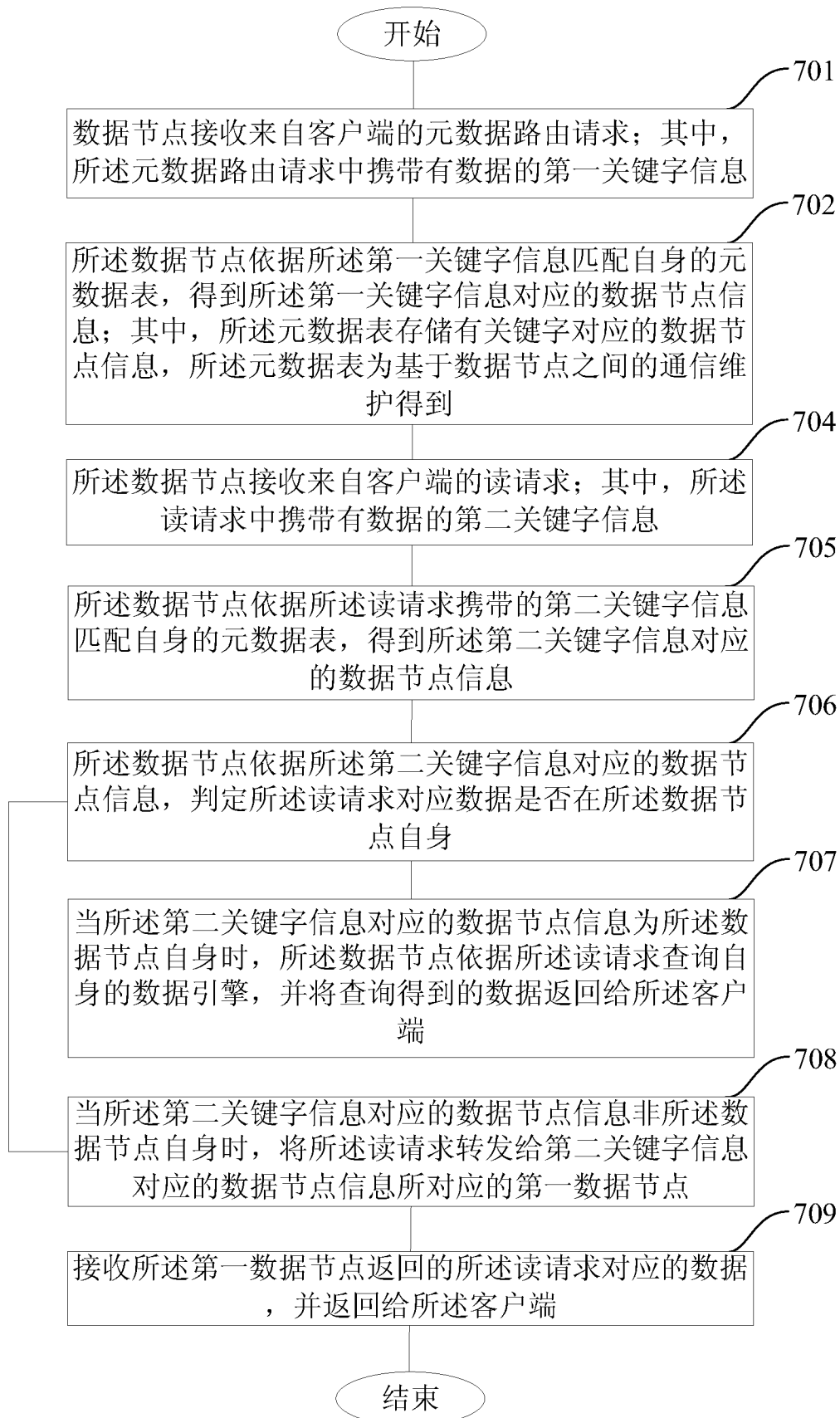


图 7

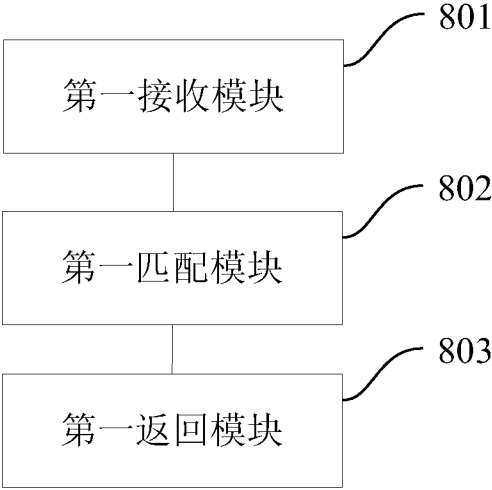


图 8

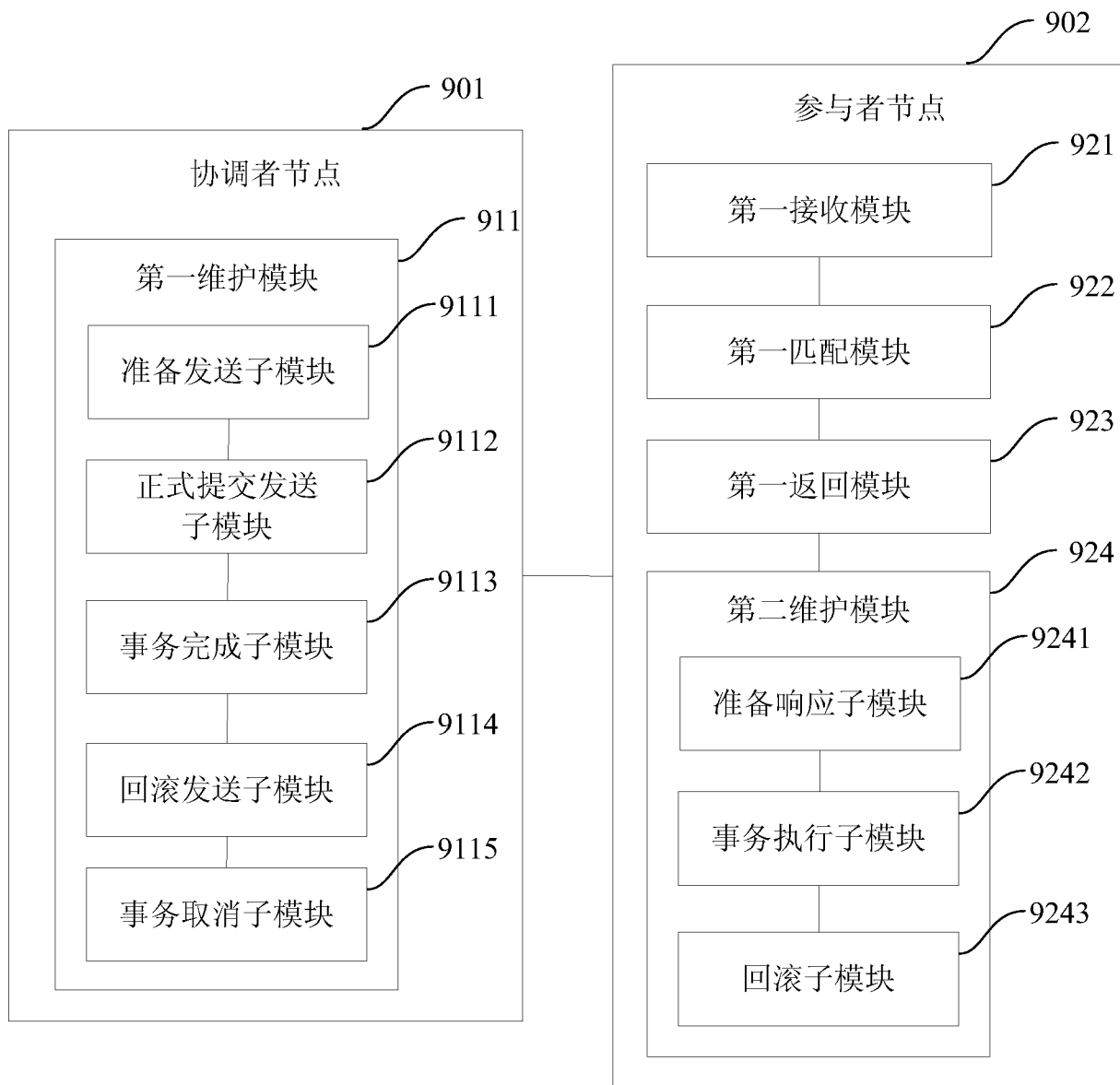


图 9

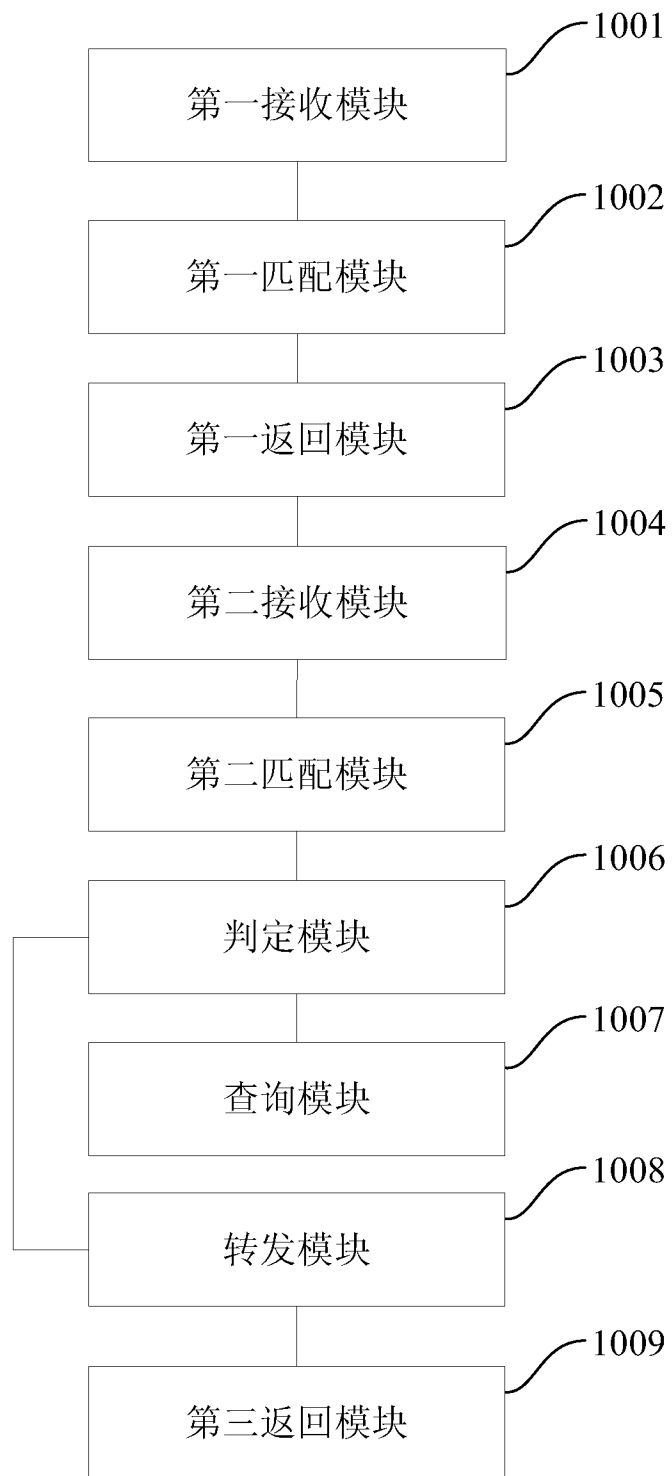


图 10

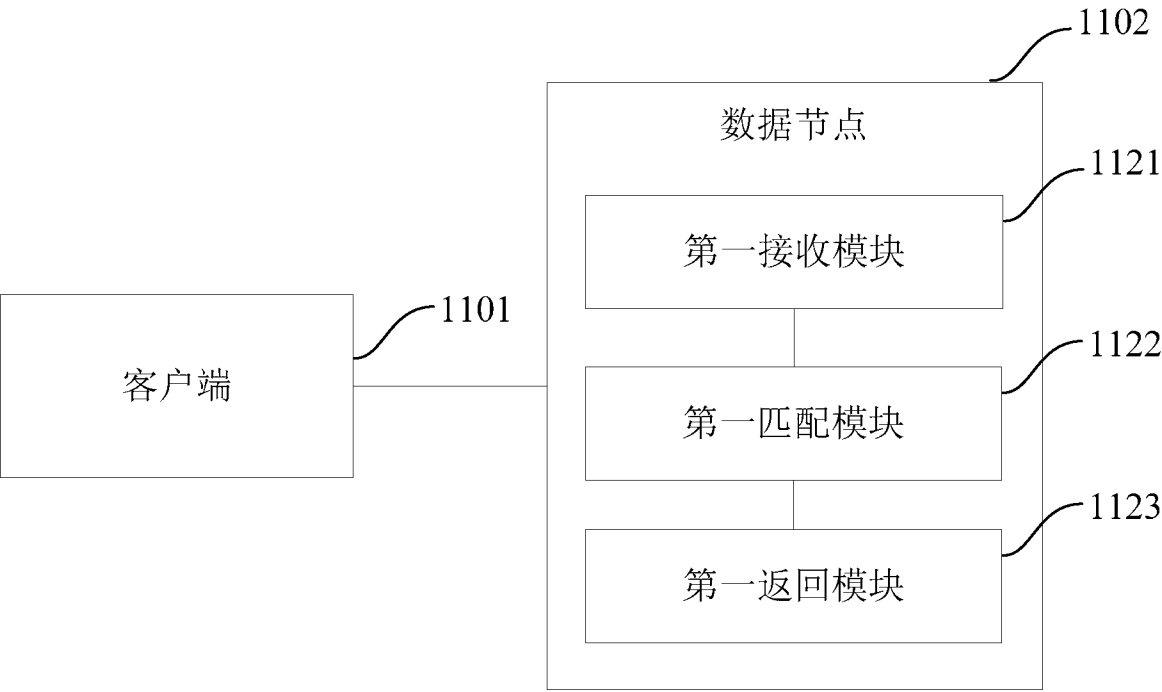


图 11

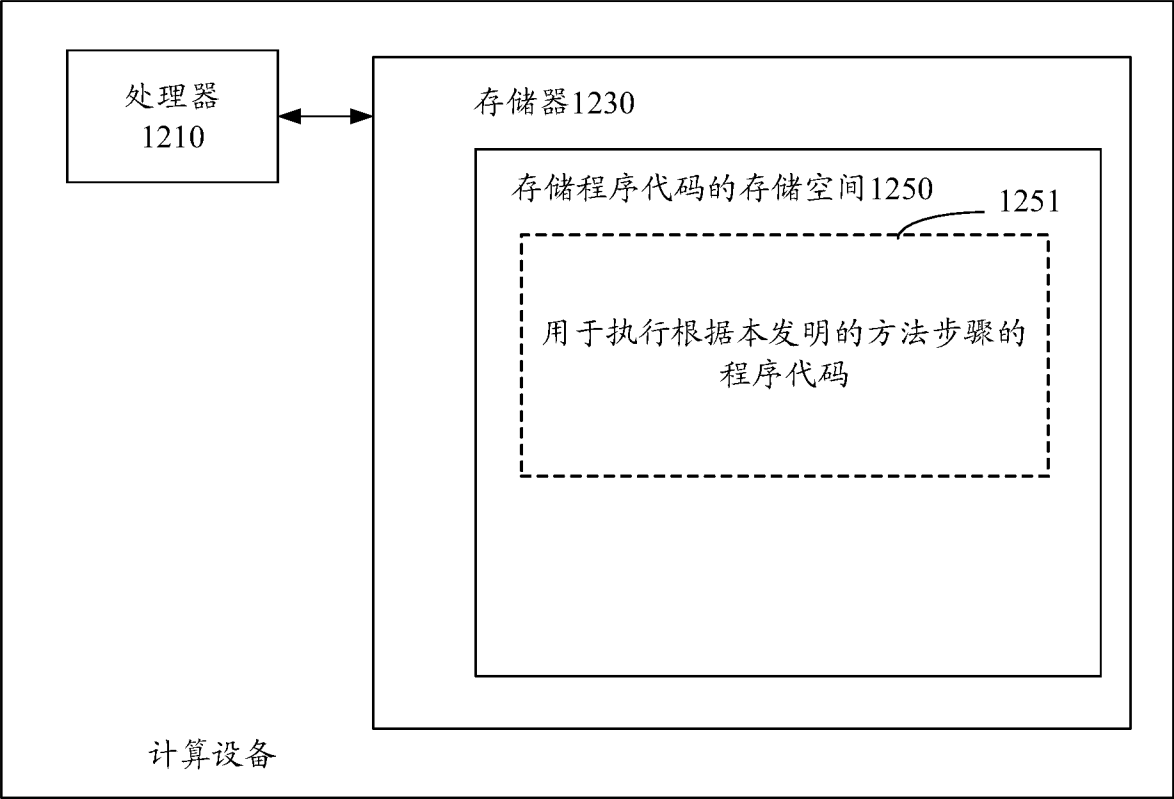


图 12

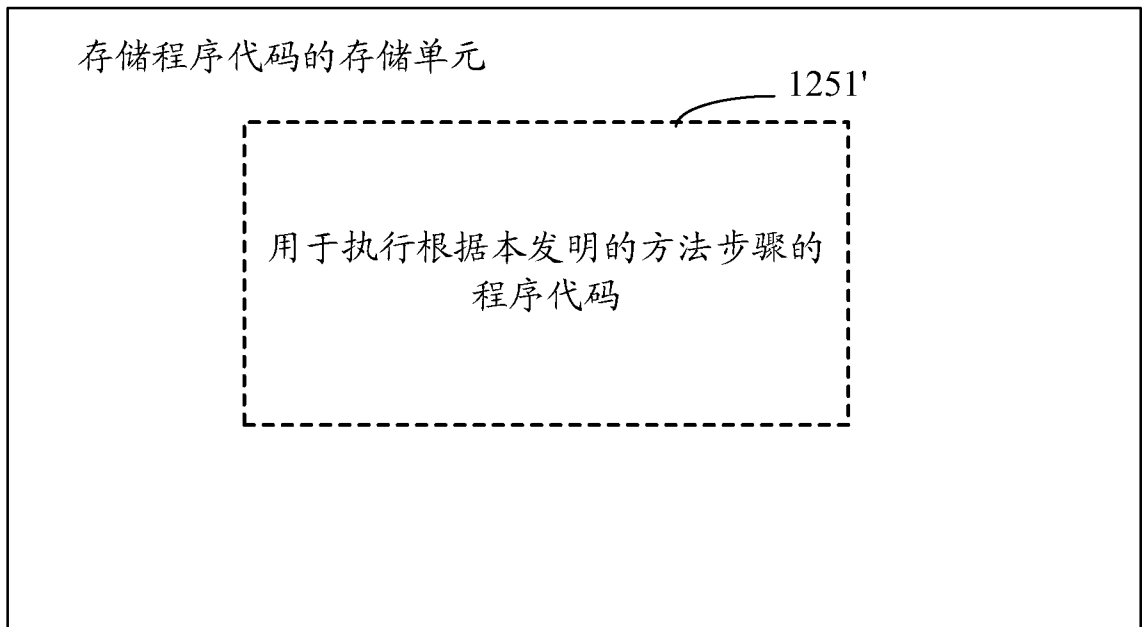


图 13

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/095507

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNXTX; CNABS; CNKI; VEN: metadata, keyword, master server, chunk server, meta, key, distribute, file, master, chunk, server, client

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103019960 A (HUAWEI TECHNOLOGIES CO., LTD.), 03 April 2013 (03.04.2013), description, paragraphs [0029]-[0092], and figures 1-3	1-2, 5-10, 13-15
A	CN 103019960 A (HUAWEI TECHNOLOGIES CO., LTD.), 03 April 2013 (03.04.2013), the whole document	3-4, 11-12
A	CN 102984280 A (BEIJING UNIVERSITY OF TECHNOLOGY), 20 March 2013 (20.03.2013), the whole document	1-15

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 25 January 2016 (25.01.2016)	Date of mailing of the international search report 05 February 2016 (05.02.2016)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer BAI, Jingxin Telephone No.: (86-10) 62089447

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2015/095507

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103019960 A	03 April 2013	None	
CN 102984280 A	20 March 2013	CN 102984280 B	20 May 2015

A. 主题的分类 H04L 29/08 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类		
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) H04L 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNTXT; CNABS; CNKI; VEN: 元数据, 关键字, 分布, 文件, 主服务器, 数据块服务器, 客户端, meta, key, distribute, file, master, chunk, server, client		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 103019960 A (华为技术有限公司) 2013年 4月 3日 (2013 - 04 - 03) 说明书第[0029]-[0092]段以及附图1-3	1-2、5-10、13-15
A	CN 103019960 A (华为技术有限公司) 2013年 4月 3日 (2013 - 04 - 03) 全文	3-4、11-12
A	CN 102984280 A (北京工业大学) 2013年 3月 20日 (2013 - 03 - 20) 全文	1-15
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期 2016年 1月 25日		国际检索报告邮寄日期 2016年 2月 5日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10) 62019451		授权官员 白晶心 电话号码 (86-10) 62089447

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/095507

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	103019960	A	2013年 4月 3日	无	
CN	102984280	A	2013年 3月 20日	CN	102984280 B 2015年 5月 20日