(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
2020 年 7 月 16 日 (16.07.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/143410 A1

(51) 国际专利分类号:
H04L 29/08 (2006.01)

(21) 国际申请号: PCT/CN2019/125871

(22) 国际申请日: 2019 年 12 月 17 日 (17.12.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910023907.2 2019年1月10日 (10.01.2019) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 石建伟 (SHI, Jianwei); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 王辉 (WANG, Hui); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 吴克柱 (WU, Kezhu); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 时晖 (SHI, Hui); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 409, Beijing 100085 (CN)。

(54) **Title:** DATA STORAGE METHOD AND DEVICE, ELECTRONIC DEVICE AND STORAGE MEDIUM

(54) **发明名称**: 数据存储方法及装置、电子设备、存储介质

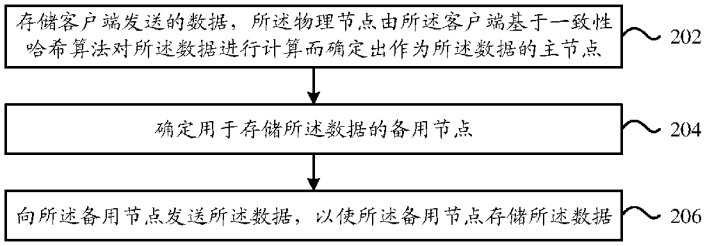


图 2

202 Storing data sent by a client, determining the physical nodes to be a master node of the data by the client performing calculation on the data based on a consistency Hash algorithm
204 Determining a standby node for storing the data
206 Sending the data to the standby node, so as to enable the standby node to store the data

(57) **Abstract:** One or more embodiments of the present description provide a data storage method and device, an electronic device and a storage medium, wherein the method is applied to physical nodes in a distributed key-value pair storage system and can comprises: storing data sent by a client, determining the physical nodes to be a master node of the data by the client performing calculation on the data based on a consistency Hash algorithm; determining a standby node for storing the data; and sending the data to the standby node, so as to enable the standby node to store the data.

(57) **摘要**: 本说明书一个或多个实施例提供一种数据存储方法及装置、电子设备、存储介质，该方法应用于分布式键值对存储系统中的物理节点，可以包括：存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点；确定用于存储所述数据的备用节点；向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

WO 2020/143410 A1

[见续页]

(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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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。

(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))。

数据存储方法及装置、电子设备、存储介质

技术领域

[01] 本说明书一个或多个实施例涉及分布式存储技术领域，尤其涉及一种数据存储方法及装置、电子设备、存储介质。

背景技术

[02] 分布式键值对存储系统大多基于一致性哈希算法设计。通过一致性哈希算法可以保证数据尽可能均匀地分布在系统的物理节点上，并且在物理节点加入或退出系统时，可使得需要迁移的数据尽可能控制在最小必要范围内。

发明内容

[03] 有鉴于此，本说明书一个或多个实施例提供一种数据存储方法及装置、电子设备、存储介质。

[04] 为实现上述目的，本说明书一个或多个实施例提供技术方案如下：

[05] 根据本说明书一个或多个实施例的第一方面，提出了一种数据存储方法，应用于分布式键值对存储系统中的物理节点；所述方法包括：

[06] 存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点；

[07] 确定用于存储所述数据的备用节点；

[08] 向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

[09] 根据本说明书一个或多个实施例的第二方面，提出了一种数据存储方法，应用于分布式键值对存储系统中的物理节点；所述方法包括：

[10] 接收主节点发送的数据；所述数据由客户端发送至主节点，所述主节点由客户端基于一致性哈希算法对所述数据进行计算而确定出；

[11] 存储所述数据。

[12] 根据本说明书一个或多个实施例的第三方面，提出了一种数据存储方法，应用于分布式键值对存储系统中的客户端；所述方法包括：

[13] 基于一致性哈希算法对数据进行计算，以确定出所述分布式键值对存储系统中作为所述数据的主节点；

[14] 向所述主节点发送所述数据，以使所述主节点存储所述数据，以及确定出与自身相关联的备用节点并将所述数据发送至所述备用节点；

[15] 或者，确定出与所述主节点相关联的备用节点；

[16] 分别向所述主节点和所述备用节点发送所述数据，以使得所述主节点和所述备用节点存储所述数据。

[17] 根据本说明书一个或多个实施例的第四方面，提出了一种数据存储装置，应用于分布式键值对存储系统中的物理节点；所述装置包括：

[18] 存储单元，存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点；

[19] 备用节点确定单元，确定用于存储所述数据的备用节点；

[20] 第一发送单元，向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

[21] 根据本说明书一个或多个实施例的第五方面，提出了一种数据存储装置，应用于分布式键值对存储系统中的物理节点；所述装置包括：

[22] 第一接收单元，接收主节点发送的数据；所述数据由客户端发送至主节点，所述主节点由客户端基于一致性哈希算法对所述数据进行计算而确定出；

[23] 存储单元，存储所述数据。

[24] 根据本说明书一个或多个实施例的第六方面，提出了一种数据存储装置，应用于分布式键值对存储系统中的客户端；所述装置包括：

[25] 第一确定单元，基于一致性哈希算法对数据进行计算，以确定出所述分布式键值对存储系统中作为所述数据的主节点；

[26] 第一发送单元，向所述主节点发送所述数据，以使所述主节点存储所述数据，以及确定出与自身相关联的备用节点并将所述数据发送至所述备用节点；

[27] 或者，包括：第二确定单元，确定出与所述主节点相关联的备用节点；

[28] 第二发送单元，分别向所述主节点和所述备用节点发送所述数据，以使得所述主节点和所述备用节点存储所述数据。

[29] 根据本说明书一个或多个实施例的第七方面，提出了一种电子设备，包括:

[30] 处理器;

[31] 用于存储处理器可执行指令的存储器;

[32] 其中，所述处理器通过运行所述可执行指令以实现如上述第一方面所述的数据存储方法。

[33] 根据本说明书一个或多个实施例的第八方面，提出了一种计算机可读存储介质，其上存储有计算机指令，该指令被处理器执行时实现如第一方面所述方法的步骤。

[34] 根据本说明书一个或多个实施例的第九方面，提出了一种电子设备，包括:

[35] 处理器;

[36] 用于存储处理器可执行指令的存储器;

[37] 其中，所述处理器通过运行所述可执行指令以实现如上述第二方面所述的数据存储方法。

[38] 根据本说明书一个或多个实施例的第十方面，提出了一种计算机可读存储介质，其上存储有计算机指令，该指令被处理器执行时实现如第二方面所述方法的步骤。

[39] 根据本说明书一个或多个实施例的第十一方面，提出了一种电子设备，包括:

[40] 处理器;

[41] 用于存储处理器可执行指令的存储器;

[42] 其中，所述处理器通过运行所述可执行指令以实现如上述第三方面所述的数据存储方法。

[43] 根据本说明书一个或多个实施例的第十二方面，提出了一种计算机可读存储介质，其上存储有计算机指令，该指令被处理器执行时实现如第三方面所述方法的步骤。

附图说明

[44] 图 1 是一示例性实施例提供的一种数据存储系统的架构示意图。

[45] 图 2 是一示例性实施例提供的一种数据存储方法的流程图。

[46] 图 3 是一示例性实施例提供的另一种数据存储方法的流程图。

[47] 图 4 是一示例性实施例提供的另一种数据存储方法的流程图。

[48] 图 5 是一示例性实施例提供的一种数据存储方法的交互图。

[49] 图 6A-6C 是一示例性实施例提供的环形哈希空间的示意图。

[50] 图 7 是一示例性实施例提供的在主节点未发生故障时的数据读取方法的交互图。

[51] 图 8 是一示例性实施例提供的在主节点发生故障时的数据读取方法的交互图。

[52] 图 9 是一示例性实施例提供的在主节点发生故障时环形哈希空间的示意图。

[53] 图 10 是一示例性实施例提供的一种设备的结构示意图。

[54] 图 11 是一示例性实施例提供的一种数据存储装置的框图。

[55] 图 12 是一示例性实施例提供的另一种设备的结构示意图。

[56] 图 13 是一示例性实施例提供的另一种数据存储装置的框图。

[57] 图 14 是一示例性实施例提供的另一种设备的结构示意图。

[58] 图 15 是一示例性实施例提供的另一种数据存储装置的框图。

具体实施方式

[59] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本说明书一个或多个实施例相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本说明书一个或多个实施例的一些方面相一致的装置和方法的例子。

[60] 需要说明的是：在其他实施例中并不一定按照本说明书示出和描述的顺序来执行相应方法的步骤。在一些其他实施例中，其方法所包括的步骤可以比本说明书所描述的更多或更少。此外，本说明书中所描述的单个步骤，在其他实施例中可能被分解为多个步骤进行描述；而本说明书中所描述的多个步骤，在其他实施例中也可能被合并为单个步骤进行描述。

[61] 图 1 是一示例性实施例提供的一种数据存储系统的架构示意图。如图 1 所示，该系统可以包括节点统计设备 11、网络 12、分布式键值对存储系统 13 和若干电子设备。其中，分布式键值对存储系统 13 可包括物理节点 131、物理节点 132 和物理节点 133

等；电子设备可包括 PC14、PC15 等。

[62] 节点统计设备 11 可以为包含一独立主机的物理服务器，或者可以为主机集群承载的虚拟服务器。在实施本说明书的数据存储方案的过程中，节点统计设备用于统计分布式键值对存储系统 13 包含的各个物理节点，从而各个物理节点以及与物理节点交互的电子设备可基于统计结果获取物理节点加入和退出分布式键值对存储系统的情况。

[63] PC14-15 只是用户可以使用的一种类型的电子设备。实际上，用户显然还可以使用诸如下述类型的电子设备: 手机、平板设备、掌上电脑(PDAs, Personal Digital Assistants)、可穿戴设备（如智能眼镜、智能手表等）等，本说明书一个或多个实施例并不对此进行限制。在实施本说明书的数据存储方案的过程中，该电子设备可作为客户端与分布式键值对存储系统 13 中的物理节点进行交互，以将数据存储至物理节点中，以及向物理节点读取数据。

[64] 分布式键值对存储系统 13 通过多个物理节点提供“按照键写入值”和“按照键读取值”两种服务。其中，不同的键值对互相独立，不存在关联关系。在实施本说明书的数据存储方案的过程中，各个物理节点可存储客户端发送的数据，以及响应于客户端的读取请求，向客户端返回相应的数据。

[65] 而对于节点统计设备 11、分布式键值对存储系统 13 和电子设备之间进行交互的网络 12，可以包括多种类型的有线或无线网络。例如，该网络 12 可以包括因特网。当然，本说明书一个或多个实施例并不对此进行限制。同时，分布式键值对存储系统 13 中的各个物理节点之间也可以通过该网络 12 进行通讯交互。

[66] 请参见图 2，图 2 是一示例性实施例提供的一种数据存储方法的流程图。如图 2 所示，该方法应用于分布式键值对存储系统中的物理节点（作为待存储数据的主节点），可以包括以下步骤:

[67] 步骤 202，存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点。

[68] 在一实施例中，分布式键值对存储系统可基于一致性哈希算法设计，以使得所需存储的数据在物理节点间的分配尽可能均匀，以及在存在物理节点加入或退出系统的情况下尽可能少地迁移数据。进一步的，客户端在确定当前待存储的数据时，可利用一致性哈希算法计算对应于该数据的散列值，再按照预设方向在哈希取值空间中查找(例如，固定按照顺时针方向在环形哈希空间中查找）与该散列值距离最近的虚拟节点，从而将

该虚拟节点对应的物理节点作为存储该数据的主节点。其中，哈希取值空间可由节点统计设备来维护。例如，节点统计设备可统计当前加入分布式键值对存储系统的所有物理节点，并基于一致性哈希算法将各个物理节点的虚拟节点尽可能均匀地映射到环形哈希空间（又称哈希环或哈希桶）中。那么，客户端和各个物理节点均可向节点统计设备获取该环形哈希空间，或者由节点统计设备主动发送至客户端和各个物理节点；换言之，客户端和物理节点侧均记录有相同的环形哈希空间。在确定出主节点后，客户端可向主节点发送待存储的数据，以使主节点存储该数据。

[69] 步骤 204，确定用于存储所述数据的备用节点。

[70] 在一实施例中，除主节点存储客户端发送的数据之外，还可进一步从哈希取值空间中选取出预设数量的备用节点来存储该数据，以防止因主节点退出系统而导致该数据丢失。因此，主节点在接收到客户端发送的数据后，可同样基于该一致性哈希算法确定对应于该数据的散列值，并按照预设方向在哈希取值空间中查找（承接于上述举例，可固定按照顺时针方向在环形哈希空间中查找）距离该散列值最近的第一虚拟节点（承接于上述举例，实际上对应于第一虚拟节点的物理节点便是主节点），并在该哈希取值空间中查找出距离第一虚拟节点最近的预设数量的第一备用虚拟节点，从而可将对应于第一备用虚拟节点的物理节点作为备用节点。其中，第一备用虚拟节点均属于区别于自身（即主节点）的各个不同物理节点；换言之，各个备用节点为相互不同的物理节点。

[71] 步骤 206，向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

[72] 在一实施例中，主节点在接收到客户端发送的数据后，可先存储该数据并在完成存储后向客户端返回回执以告知客户端存储成功，而针对向备用节点发送该数据的操作，可采用异步复制该数据的方式来实现将该数据备份至备用节点中，从而降低对响应客户端请求存储数据的速度的影响，即符合弱一致性的要求即可。

[73] 在一实施例中，基于将客户端请求存储的数据分别存储在主节点和备用节点中，当后续客户端请求读取该数据时，若主节点处于正常工作状态，则由主节点来响应读取数据的请求（备用节点无需响应）即可。其中，客户端在发送针对该数据的读取请求时，若主节点处于正常工作状态（即未退出分布式键值对存储系统），则按照上述计算散列值的方式仍然可以定位到该主节点，即该主节点仍然可以接收到客户端发送的针对该数据的读取请求。同时，主节点在接收到读取请求后，可同样校验自身是否为该数据的主节点，并在基于一致性哈希算法确定出自身作为该数据的主节点时，向客户端返回该数据。当然，物理节点（无论是否为该数据的主节点）在接收到读取请求后，可无需执行

上述校验的操作，直接向客户端返回相应的数据。

[74] 而当主节点退出分布式键值对存储系统时（例如主节点发生故障），可平滑切换至备节点，由备用节点(同样存储有客户端请求读取的数据)来响应该读取数据的请求，从而可以避免在主节点处于非正常工作时客户端无法读取所需数据（即主节点上存储的数据）的情况。同时，在宏观上，不同的数据所对应的主节点和备用节点是不同的，因此各个物理节点的负载仍然可以达到均衡，不存在明显的主备差异。例如，客户端基于本地维护的系统中各个物理节点的状态（来自于节点统计设备），可确定所请求读取的数据的主节点是否发生故障，以及在主节点发生故障时进一步确定出与主节点相关联的备用节点（即采用上述步骤204中确定备用节点的方式），从而向备用节点发送读取请求以获取相应的数据。

[75] 在一实施例中，基于上述利用主节点和备用节点分别存储数据的机制，当分布式键值对存储系统中新加入其他物理节点时，针对任一数据，与之对应的主节点和备用节点也将随之发生变化。换言之，该新加入的其他物理节点，将作为部分数据的主节点，也将作为部分其他数据的备用节点。因此，需要将这些数据备份至新加入的其他物理节点中，以使得该其他物理节点在加入系统后可正常向客户端提供读取数据的服务。

[76] 以上述物理节点作为主节点所存储的数据为例，当其他物理节点加入分布式键值对存储系统，且该其他物理节点作为该数据的主节点或备用节点时，向该其他物理节点发送该数据以使该其他物理节点存储该数据。其中，可通过以下方式确定新加入的其他物理节点是否为该数据的主节点或者备用节点：基于一致性哈希算法（在实施本说明书的数据存储方案时，客户端和物理节点均采用同一一致性哈希算法）确定对应于该数据的散列值，当哈希取值空间中在预设方向上距离该散列值最近的第二虚拟节点属于该其他物理节点时，可判定该其他物理节点为该数据的主节点；当该其他物理节点的虚拟节点属于第二备用虚拟节点时，可判定该其他物理节点为该数据的备用节点。其中，第二备用虚拟节点为在哈希取值空间中距离第二虚拟节点最近的预设数量（即备用节点的数量）的虚拟节点，且第二备用虚拟节点均属于区别于该数据的主节点的其他不同物理节点。

[77] 进一步的，针对任一数据，与之对应的备用节点可能在分布式键值对存储系统中新加入其他物理节点后，不再作为该任一数据的备用节点；换言之，在此情况下该任一数据对于该备用节点来说为无需存储的冗余数据。因此，该备用节点可在该其他物理节点加入并备份完数据之后，将该任一数据删除以节省自身的存储空间。例如，各个物理

节点可按照预设频率来删除冗余数据，或者在通过节点统计设备得知存在新的物理节点加入系统后来执行删除冗余数据的操作。其中，各个物理节点可利用上述计算散列值的方式来确定出自身是否仍然为所存储数据的备用节点，并将自身并非与之对应的备用节点的数据从本地删除。

[78] 在一实施例中，可将备用节点的数量设定为与分布式键值对存储系统的安全等级、存储空间、处理资源呈正相关。换言之，当分布式键值对存储系统的安全等级要求越高时，可配置数量相对较多的备用节点，从而防止数据丢失。

[79] 相应的，请参见图3，图3是一示例性实施例提供的另一种数据存储方法的流程图。如图3所示，该方法应用于分布式键值对存储系统中的物理节点（作为待存储数据的备用节点），可以包括以下步骤：

[80] 步骤302，接收主节点发送的数据；所述数据由客户端发送至主节点，所述主节点由客户端基于一致性哈希算法对所述数据进行计算而确定出。

[81] 在一实施例中，除主节点存储客户端发送的数据之外，还可进一步选取出预设数量的备用节点来存储该数据，以防止因主节点退出系统而导致该数据丢失。因此，主节点在接收到客户端发送的数据后，可从分布式键值对存储系统中选取部分物理节点作为备用节点，并向备用节点发送自身存储的数据以使得备用节点存储该数据，从而实现对该数据的备份。其中，客户端确定主节点以及主节点确定出与自身相关联的备用节点的具体过程，可参考上述图2所示实施例的相应部分，在此不再赘述。

[82] 步骤304，存储所述数据。

[83] 在一实施例中，基于将客户端请求存储的数据分别存储在主节点和备用节点中，当后续客户端请求读取该数据时，若主节点处于正常工作状态，则由主节点来响应读取数据的请求（备用节点无需响应）即可；否则，可由备用节点（同样存储有客户端请求读取的数据）来响应该读取数据的请求，从而可以避免在主节点处于非正常工作时客户端无法读取所需数据（即主节点上存储的数据）的情况。同时，在宏观上，不同的数据所对应的主节点和备用节点是不同的，因此各个物理节点的负载仍然可以达到均衡，不存在明显的主备差异。

[84] 例如，客户端在确定出所请求读取的数据的主节点发生故障后，可进一步确定出与该主节点相关联的备用节点（即采用上述步骤204中确定备用节点的方式），从而向备用节点发送读取请求以获取相应的数据。那么，备用节点可接收客户端发送的针对该

数据的读取请求（即该读取请求由客户端在主节点退出分布式键值对存储系统时发送），并向客户端返回该数据。

[85] 由于数据的主节点退出分布式键值对存储系统，备用节点可切换为该数据的主节点来响应客户端发送的读取请求。而在该备用节点切换为主节点后，自身作为主节点仍然存在退出分布式键值对存储系统的可能。因此，为了防止数据丢失，可进一步确定与自身相关联的备用节点，并向确定出的备用节点发送该数据，以使备用节点存储该数据。其中，该确定备用节点的过程，与上述图 2 所示主节点确定与自身相关联的备用节点的过程类似，在此不再赘述。

[86] 相应的，请参见图 4，图 4 是一示例性实施例提供的另一种数据存储方法的流程图。如图 4 所示，该方法应用于分布式键值对存储系统中的客户端，可以包括以下步骤：

[87] 步骤 402，基于一致性哈希算法对数据进行计算，以确定出所述分布式键值对存储系统中作为所述数据的主节点。

[88] 步骤 404，向所述主节点发送所述数据，以使所述主节点存储所述数据，以及确定出与自身相关联的备用节点并将所述数据发送至所述备用节点。

[89] 步骤 406，确定出与所述主节点相关联的备用节点。

[90] 步骤 408，分别向所述主节点和所述备用节点发送所述数据，以使得所述主节点和所述备用节点存储所述数据。

[91] 在一种情况下，客户端可仅确定出待存储的数据的主节点，而与主节点相关联的备用节点由主节点自身来确定即可，同时该数据也由主节点发送至备用节点，从而可降低客户端的处理压力。在另一种情况下，客户端在确定出待存储的数据的主节点后，可进一步确定与主节点相关联的备用节点，并分别向主节点和备用节点发送该数据，从而降低主节点的压力（主节点无需执行确定备用节点和向备用节点发送数据的操作）。

[92] 为了便于理解，下面结合场景和附图对本说明书的数据存储方案进行详细说明。

[93] 请参见图 5，图 5 是一示例性实施例提供的一种数据存储方法的交互图。如图 5 所示，该交互过程可以包括以下步骤：

[94] 步骤 502，客户端计算对应于待存储的数据的散列值。

[95] 步骤 504，确定存储该数据的主节点。

[96] 在一实施例中，可将哈希算法的取值空间首尾相接来形成环形哈希空间。进一步

的，针对每个物理节点，计算相应的多个虚拟节点的散列值并映射到环形哈希空间。那么当需要读写数据时，可根据数据的键值计算散列值映射到环形哈希空间，再按照固定方向（比如顺时针方向）查找与该散列值距离最近的虚拟节点对应的物理节点，而该物理节点便是存储该数据的主节点。

[97] 举例而言，如图 6A 所示，环形哈希空间 P 的取值空间为 $0\sim2^{32}$，虚拟节点包括 B4、C3、A1、C1、A2、D2 等。其中，虚拟节点 A1、A2 属于物理节点 A，虚拟节点 B4 属于物理节点 B，虚拟节点 C1、C3 属于物理节点 C，虚拟节点 D2 属于物理节点 D。假定对应于待存储的数据的散列值为 K，散列值 K 在环形哈希空间 P 中的位置如图中箭头所示。以按照顺时针查找主节点为例，由图 6A 可知，距离散列值 K 最近的第一个虚拟节点为 B4（即虚拟节点 B4 为第一虚拟节点），那么与虚拟节点 B4 对应的物理节点 B 则可作为待存储的数据的主节点。

[98] 步骤 506，客户端向确定出的主节点发送存储请求。

[99] 步骤 508，主节点存储该待存储的数据。

[100] 步骤 510，主节点向客户单返回存储成功的回执。

[101] 在一实施例中，主节点在接收到客户端发送的数据（包含于存储请求中）后，可先存储该数据并在完成存储后向客户端返回回执以告知客户端存储成功，而针对向备用节点发送该数据的操作，可采用异步复制该数据的方式来实现将该数据备份至备用节点中，从而降低对响应客户端请求存储数据的速度的影响，即符合弱一致性的要求即可。

[102] 步骤 512，主节点确定与自身相关联的用于存储该数据的备用节点。

[103] 在一实施例中，可预先设定备用节点的数量。其中，备用节点的数量与分布式键值对存储系统的安全等级、存储空间、处理资源呈正相关。例如，分布式键值对存储系统的安全等级要求越高，则在存储数据时可配置越多的备用节点；分布式键值对存储系统中各物理节点的存储空间越大，则在存储数据时可配置越多的备用节点；分布式键值对存储系统中各物理节点的处理资源越多，则在存储数据时可配置越多的备用节点。

[104] 承接于上述举例，如图 6A 所示，假定在存储数据时设定 2 个备用节点，那么在环形哈希空间 P 中距离第一虚拟节点 B4 最近，且不属于物理节点 B 的 2 个互不相同的第一备用虚拟节点分别为 C3 和 A1，从而可将物理节点 C 和物理节点 A 作为备用节点。

[105] 步骤 514，主节点向确定出的备用节点发送数据。

[106] 在一实施例中，在一种情况下，客户端可仅确定出待存储的数据的主节点，而与主节点相关联的备用节点由主节点自身来确定即可，同时该数据也由主节点发送至备用节点（即上述步骤 512-514），从而可降低客户端的处理压力。在另一种情况下，客户端在确定出待存储的数据的主节点后，可进一步确定与主节点相关联的备用节点，并分别向主节点和备用节点发送该数据，从而降低主节点的压力（主节点无需执行确定备用节点和向备用节点发送数据的操作）。其中，客户端确定主节点和备用节点的过程，与上述由主节点来确定主节点和备用节点的过程相同，在此不再赘述。

[107] 步骤 516，备用节点存储数据。

[108] 在一实施例中，基于上述利用主节点和备用节点分别存储数据的机制，当分布式键值对存储系统中新加入其他物理节点时，针对任一数据，与之对应的主节点和备用节点也将随之发生变化。换言之，该新加入的其他物理节点，将作为部分数据的主节点，也将作为部分其他数据的备用节点。因此，需要将这些数据备份至新加入的其他物理节点中，以使得该其他物理节点在加入系统后可正常向客户端提供读取数据的服务。

[109] 以图 5 所示实施例中物理节点 B 作为主节点所存储的数据（以下简称为目标数据）为例，如图 6B 所示，假定当物理节点 E 加入分布式键值对存储系统时，与物理节点 E 对应的虚拟节点 E1 在环形哈希空间 P 中的位置在散列值 K 与虚拟节点 B4 之间（即虚拟节点 E1 为环形哈希空间 P 中在顺时针方向上距离散列值 K 最近的第二虚拟节点），那么此时针对目标数据的主节点则为新加入的物理节点 E。进一步的，在物理节点 E 加入分布式键值对存储系统之前作为目标数据的主节点的物理节点 B，需要将目标数据备份至物理节点 E 中。在备份完成后，后续客户端在需要读取目标数据时，便可根据散列值 K 定位出作为目标数据的主节点的物理节点 E，从而向物理节点 E 发送读取请求以获取目标数据。例如，在备份完成后可告知节点统计设备，使得客户端可通过节点统计设备获取到物理节点 E 已备份完成的信息，那么后续客户端在需要读取目标数据时，便可基于加入物理节点 E 的虚拟节点后的环形哈希空间 P 来确定主节点。

[110] 如图 6C 所示，假定当物理节点 E 加入分布式键值对存储系统时，与物理节点 E 对应的虚拟节点 E1 在环形哈希空间 P 中的位置在虚拟节点 C3 与 A1 之间（即虚拟节点 E1 为在环形哈希空间中距离虚拟节点 B4 最近的 2 个第二备用虚拟节点中的其中一个，另外一个为虚拟节点 C3），那么此时针对目标数据的备用节点则为新加入的物理节点 E 和之前的物理节点 C。进一步的，作为目标数据的主节点物理节点 B 需要将目标数据备份至物理节点 E 中，从而保证主节点始终有 2 个相关联的备用节点。

[111] 进一步的，针对任一数据，与之对应的备用节点可能在分布式键值对存储系统中新加入其他物理节点后，不再作为该任一数据的备用节点；换言之，在此情况下该任一数据对于该备用节点来说为无需存储的冗余数据。因此，该备用节点可在该其他物理节点加入并备份完数据之后，将该任一数据删除以节省自身的存储空间。例如，图 6C 所示的虚拟节点 A1，在物理节点 E 加入后不再作为目标数据的备用节点，那么对应于虚拟节点 A1 的物理节点 A 可将目标数据删除。作为一示例性实施例，各个物理节点可按照预设频率来删除冗余数据，或者在通过节点统计设备得知存在新的物理节点加入系统后来执行删除冗余数据的操作。

[112] 由上述实施例可见，基于将客户端请求存储的数据分别存储在主节点和备用节点中，当后续客户端需要请求读取该数据时，若主节点处于正常工作状态，则由主节点来响应读取数据的请求（备用节点无需响应）即可；否则，可平滑切换至备节点以由备用节点（同样存储有客户端请求读取的数据）来响应该读取数据的请求，从而可以避免在主节点处于非正常工作时客户端无法读取所需数据（即主节点上存储的数据）的情况。同时，在宏观上，不同的数据所对应的主节点和备用节点是不同的，因此各个物理节点的负载仍然可以达到均衡，不存在明显的主备差异。下面结合图 7-9 对客户端请求读取数据的过程进行详细说明。

[113] 请参见图 7，图 7 是一示例性实施例提供的在主节点未发生故障时的数据读取方法的交互图。如图 7 所示，该交互过程可以包括以下步骤：

[114] 步骤 702，客户端计算对应于目标数据的散列值。

[115] 在一实施例中，以客户端向主节点请求读取上述图 5 中所存储的目标数据为例，对本说明书读取数据的过程进行详细说明。

[116] 步骤 704，确定存储目标数据的主节点。

[117] 在一实施例中，当主节点未发生故障（即主节点未退出分布式键值对存储系统）时，客户端直接向主节点发送针对目标数据的读取请求即可。例如，在图 6A 中，若物理节点 B 未发生故障，则客户端按照上述步骤 504 的方式仍然可以确定出目标数据的主节点为物理节点 B。

[118] 步骤 706，向主节点发送针对目标数据的读取请求。

[119] 步骤 708，主节点校验自身是否为目标数据的主节点或备用节点。

[120] 在一实施例中，物理节点在接收到读取请求后，可校验自身是否为相应数据的主

节点或备用节点，并且只在自身为相应数据的主节点（校验过程与上述确定主节点和备用节点的过程类似）时，向客户单返回该相应数据。当然，物理节点在接收到读取请求后，也可以无需执行上述校验的操作，默认返回相应数据即可。而在本实施例中，物理节点 B 作为主节点并未发生故障，因此可直接返回目标数据。

[121] 步骤 710，主节点读取目标数据。

[122] 步骤 712，主节点向客户端返回目标数据。

[123] 请参见图 8，图 8 是一示例性实施例提供的在主节点发生故障时的数据读取方法的交互图。如图 8 所示，该交互过程可以包括以下步骤:

[124] 步骤 802，客户端计算对应于目标数据的散列值。

[125] 在一实施例中，以客户端向主节点请求读取上述图 5 中所存储的目标数据为例，对本说明书读取数据的过程进行详细说明。

[126] 步骤 804，确定存储目标数据的备用节点。

[127] 步骤 806，向备用节点 1 发送针对目标数据的读取请求。

[128] 在一实施例中，当主节点发生故障（即主节点退出分布式键值对存储系统）时，客户端可切换为向备用节点发送针对目标数据的读取请求。

[129] 举例而言，如图 9 所示，当物理节点 B 发生故障时，与散列值 K 距离最近的虚拟节点为 C3 和 A1，从而可确定出目标数据的备用节点包括物理节点 C（备用节点 1）和物理节点 A（备用节点 2）。例如，客户端可选取在物理节点 B 退出系统后距离散列值 K 最近的虚拟节点 C3，从而向与虚拟节点为 C3 对应的物理节点 C 发送针对目标数据的读取请求。

[130] 步骤 808，备用节点 1 校验自身是否为目标数据的主节点或备用节点。

[131] 步骤 810，备用节点 1 切换为目标数据的主节点。

[132] 步骤 812，备用节点 1 读取目标数据。

[133] 步骤 814，备用节点 1 向客户端返回目标数据。

[134] 在一实施例中，物理节点在接收到读取请求后，可校验自身是否为相应数据的主节点或备用节点，并且只在自身为相应数据的主节点（校验过程与上述确定主节点和备用节点的过程类似）时，向客户单返回该相应数据。当然，物理节点在接收到读取请求

后，也可以无需执行上述校验的操作，默认返回相应数据即可。

[135] 举例而言，如图 9 所示，物理节点 C（备用节点 1）在接收到针对目标数据的读取请求后，可确定出距离散列值 K 最近的是虚拟节点 B4，而通过节点统计神可知与虚拟节点 B4 对应的物理节点 B 已经发生故障。同时，下一个距离散列值 K 最近的是虚拟节点 C3，从而确定出自身为目标数据的备用节点，那么将自身切换为目标数据的主节点以响应客户端发送的读取请求。

[136] 步骤 816，备用节点 1 向备用节点 3 发送目标数据。

[137] 步骤 818，备用节点 3 存储目标数据。

[138] 在一实施例中，在备用节点 1 切换为主节点后，自身作为主节点仍然存在退出分布式键值对存储系统的可能。因此，为了防止数据丢失，可进一步确定与自身相关联的备用节点（与上述主节点接收到存储请求时确定备用节点的方式相同），并向确定出的备用节点发送目标数据，以使该备用节点存储目标数据。换言之，作为任一数据的主节点，应始终保证存在预设数量的与自身相关联的备用节点。

[139] 举例而言，如图 6A 所示，在物理节点 B 发生故障之前，距离散列值 K 最近的 3 个分别属于不同物理节点的虚拟节点为 B4、C3 和 A1；如图 9 所示，在物理节点 B 发生故障之后，距离散列值 K 最近的 3 个分别属于不同物理节点的虚拟节点为 C3、A1 和 D2。可见，物理节点 D 在物理节点 B 发生故障之前，并非目标数据的备用节点，而在物理节点 B 发生故障之后作为目标数据的备用节点。因此，需将目标数据备份至物理节点 D（即备用节点 3）中。其中，由于读取请求仅由作为主节点的物理节点 C 来响应，物理节点 C 同样可以先响应读取请求，以降低对响应读取请求的速度的影响，然后再异步将目标数据备份至物理节点 D 中。

[140] 图 10 是一示例性实施例提供的一种设备的示意结构图。请参考图 10，在硬件层面，该设备包括处理器 1002、内部总线 1004、网络接口 1006、内存 1008 以及非易失性存储器 1010，当然还可能包括其他业务所需要的硬件。处理器 1002 从非易失性存储器 1010 中读取对应的计算机程序到内存 1008 中然后运行，在逻辑层面上形成数据存储装置。当然，除了软件实现方式之外，本说明书一个或多个实施例并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[141] 请参考图 11，在软件实施方式中，该数据存储装置应用于分布式键值对存储系统

中的物理节点，可以包括：

[142] 存储单元 1101，存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点；

[143] 备用节点确定单元 1102，确定用于存储所述数据的备用节点；

[144] 第一发送单元 1103，向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

[145] 可选的，所述备用节点确定单元 1102 具体用于：

[146] 基于所述一致性哈希算法确定对应于所述数据的散列值；

[147] 按照预设方向在哈希取值空间中查找距离所述散列值最近的第一虚拟节点；

[148] 在所述哈希取值空间中查找出距离所述第一虚拟节点最近的预设数量的第一备用虚拟节点，所述第一备用虚拟节点均属于区别于自身的各个不同物理节点；

[149] 将对应于所述第一备用虚拟节点的物理节点作为所述备用节点。

[150] 可选的，还包括：

[151] 接收单元 1104，接收所述客户端发送的针对所述数据的读取请求；

[152] 数据返回单元 1105，当基于所述一致性哈希算法确定出自身作为所述数据的主节点时，向所述客户端返回所述数据。

[153] 可选的，还包括：

[154] 第二发送单元 1106，当其他物理节点加入所述分布式键值对存储系统，且所述其他物理节点作为所述数据的主节点或备用节点时，向所述其他物理节点发送所述数据，以使所述其他物理节点存储所述数据。

[155] 可选的，还包括：

[156] 散列值确定单元 1107，基于所述一致性哈希算法确定对应于所述数据的散列值；

[157] 第一判定单元 1108，当哈希取值空间中在预设方向上距离所述散列值最近的第二虚拟节点属于所述其他物理节点时，判定所述其他物理节点为所述数据的主节点；

[158] 第二判定单元 1109，当所述其他物理节点的虚拟节点属于第二备用虚拟节点时，判定所述其他物理节点为所述数据的备用节点，所述第二备用虚拟节点为在哈希取值空间中距离所述第二虚拟节点最近的预设数量的虚拟节点，且所述第二备用虚拟节点均属

于区别于所述数据的主节点的其他不同物理节点。

[159] 可选的，所述备用节点的数量与所述分布式键值对存储系统的安全等级、存储空间、处理资源呈正相关。

[160] 图 12 是一示例性实施例提供的另一种设备的示意结构图。请参考图 12，在硬件层面，该设备包括处理器 1202、内部总线 1204、网络接口 1206、内存 1208 以及非易失性存储器 1212，当然还可能包括其他业务所需要的硬件。处理器 1202 从非易失性存储器 1212 中读取对应的计算机程序到内存 1208 中然后运行，在逻辑层面上形成数据存储装置。当然，除了软件实现方式之外，本说明书一个或多个实施例并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[161] 请参考图 13，在软件实施方式中，该数据存储装置应用于分布式键值对存储系统中的物理节点，可以包括:

[162] 第一接收单元 1301，接收主节点发送的数据；所述数据由客户端发送至主节点，所述主节点由客户端基于一致性哈希算法对所述数据进行计算而确定出；

[163] 存储单元 1302，存储所述数据。

[164] 可选的，还包括:

[165] 第二接收单元 1303，接收所述客户端发送的针对所述数据的读取请求，所述读取请求由所述客户端在所述主节点退出所述分布式键值对存储系统时发送；

[166] 返回单元 1304，向所述客户端返回所述数据。

[167] 可选的，还包括:

[168] 切换单元 1305，切换为所述数据的主节点；

[169] 确定单元 1306，确定与自身相关联的备用节点；

[170] 发送单元 1307，向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

[171] 图 14 是一示例性实施例提供的另一种设备的示意结构图。请参考图 14，在硬件层面，该设备包括处理器 1402、内部总线 1404、网络接口 1406、内存 1408 以及非易失性存储器 1414，当然还可能包括其他业务所需要的硬件。处理器 1402 从非易失性存储器 1414 中读取对应的计算机程序到内存 1408 中然后运行，在逻辑层面上形成数据存储装置。当然，除了软件实现方式之外，本说明书一个或多个实施例并不排除其他实现方式，

比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[172] 请参考图 15，在软件实施方式中，该数据存储装置应用于分布式键值对存储系统中的客户端，可以包括:

[173] 第一确定单元 1501，基于一致性哈希算法对数据进行计算，以确定出所述分布式键值对存储系统中作为所述数据的主节点;

[174] 第一发送单元 1502，向所述主节点发送所述数据，以使所述主节点存储所述数据，以及确定出与自身相关联的备用节点并将所述数据发送至所述备用节点;

[175] 或者，包括: 第二确定单元 1503，确定出与所述主节点相关联的备用节点;

[176] 第二发送单元 1504，分别向所述主节点和所述备用节点发送所述数据，以使得所述主节点和所述备用节点存储所述数据。

[177] 上述实施例阐明的系统、装置、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为计算机，计算机的具体形式可以是个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件收发设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任意几种设备的组合。

[178] 在一个典型的配置中，计算机包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

[179] 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（flashRAM）。内存是计算机可读介质的示例。

[180] 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、磁盘存储、量子存储器、基于石墨烯的存储介质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暂存电脑可读媒体（transitory media），如调制的数据信号和载波。

[181]还需要说明的是，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、商品或者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、商品或者设备所固有的要素。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、商品或者设备中还存在另外的相同要素。

[182]上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

[183]在本说明书一个或多个实施例使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[184]应当理解，尽管在本说明书一个或多个实施例可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[185]以上所述仅为本说明书一个或多个实施例的较佳实施例而已，并不用以限制本说明书一个或多个实施例，凡在本说明书一个或多个实施例的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本说明书一个或多个实施例保护的范围之内。

权利要求书

1. 一种数据存储方法，应用于分布式键值对存储系统中的物理节点；所述方法包括:

存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点;

确定用于存储所述数据的备用节点;

向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

2. 根据权利要求 1 所述的方法，所述确定用于存储所述数据的备用节点，包括:

基于所述一致性哈希算法确定对应于所述数据的散列值;

按照预设方向在哈希取值空间中查找距离所述散列值最近的第一虚拟节点;

在所述哈希取值空间中查找出距离所述第一虚拟节点最近的预设数量的第一备用虚拟节点，所述第一备用虚拟节点均属于区别于自身的各个不同物理节点;

将对应于所述第一备用虚拟节点的物理节点作为所述备用节点。

3. 根据权利要求 1 所述的方法，还包括:

接收所述客户端发送的针对所述数据的读取请求;

当基于所述一致性哈希算法确定出自身作为所述数据的主节点时，向所述客户端返回所述数据。

4. 根据权利要求 1 所述的方法，还包括:

当其他物理节点加入所述分布式键值对存储系统，且所述其他物理节点作为所述数据的主节点或备用节点时，向所述其他物理节点发送所述数据，以使所述其他物理节点存储所述数据。

5. 根据权利要求 4 所述的方法，还包括:

基于所述一致性哈希算法确定对应于所述数据的散列值;

当哈希取值空间中在预设方向上距离所述散列值最近的第二虚拟节点属于所述其他物理节点时，判定所述其他物理节点为所述数据的主节点;

当所述其他物理节点的虚拟节点属于第二备用虚拟节点时，判定所述其他物理节点为所述数据的备用节点，所述第二备用虚拟节点为在哈希取值空间中距离所述第二虚拟节点最近的预设数量的虚拟节点，且所述第二备用虚拟节点均属于区别于所述数据的主节点的其他不同物理节点。

6. 根据权利要求 1 所述的方法，所述备用节点的数量与所述分布式键值对存储系统的安全等级、存储空间、处理资源呈正相关。

7. 一种数据存储方法，应用于分布式键值对存储系统中的物理节点；所述方法包括:

接收主节点发送的数据；所述数据由客户端发送至主节点，所述主节点由客户端基于一致性哈希算法对所述数据进行计算而确定出；

存储所述数据。

8. 根据权利要求 7 所述的方法，还包括:

接收所述客户端发送的针对所述数据的读取请求，所述读取请求由所述客户端在所述主节点退出所述分布式键值对存储系统时发送；

向所述客户端返回所述数据。

9. 根据权利要求 8 所述的方法，还包括:

切换为所述数据的主节点；

确定与自身相关联的备用节点；

向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

10. 一种数据存储方法，应用于分布式键值对存储系统中的客户端；所述方法包括:

基于一致性哈希算法对数据进行计算，以确定出所述分布式键值对存储系统中作为所述数据的主节点；

向所述主节点发送所述数据，以使所述主节点存储所述数据，以及确定出与自身相关联的备用节点并将所述数据发送至所述备用节点；

或者，确定出与所述主节点相关联的备用节点；

分别向所述主节点和所述备用节点发送所述数据，以使得所述主节点和所述备用节点存储所述数据。

11. 一种数据存储装置，应用于分布式键值对存储系统中的物理节点；所述装置包括:

存储单元，存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点；

备用节点确定单元，确定用于存储所述数据的备用节点；

第一发送单元，向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

12. 根据权利要求 11 所述的装置，所述备用节点确定单元具体用于:

基于所述一致性哈希算法确定对应于所述数据的散列值；

按照预设方向在哈希取值空间中查找距离所述散列值最近的第一虚拟节点；

在所述哈希取值空间中查找出距离所述第一虚拟节点最近的预设数量的第一备用

虚拟节点，所述第一备用虚拟节点均属于区别于自身的各个不同物理节点；

将对应于所述第一备用虚拟节点的物理节点作为所述备用节点。

13. 根据权利要求 11 所述的装置，还包括：

接收单元，接收所述客户端发送的针对所述数据的读取请求；

数据返回单元，当基于所述一致性哈希算法确定出自身作为所述数据的主节点时，向所述客户端返回所述数据。

14. 根据权利要求 11 所述的装置，还包括：

第二发送单元，当其他物理节点加入所述分布式键值对存储系统，且所述其他物理节点作为所述数据的主节点或备用节点时，向所述其他物理节点发送所述数据，以使所述其他物理节点存储所述数据。

15. 根据权利要求 14 所述的装置，还包括：

散列值确定单元，基于所述一致性哈希算法确定对应于所述数据的散列值；

第一判定单元，当哈希取值空间中在预设方向上距离所述散列值最近的第二虚拟节点属于所述其他物理节点时，判定所述其他物理节点为所述数据的主节点；

第二判定单元，当所述其他物理节点的虚拟节点属于第二备用虚拟节点时，判定所述其他物理节点为所述数据的备用节点，所述第二备用虚拟节点为在哈希取值空间中距离所述第二虚拟节点最近的预设数量的虚拟节点，且所述第二备用虚拟节点均属于区别于所述数据的主节点的其他不同物理节点。

16. 根据权利要求 11 所述的装置，所述备用节点的数量与所述分布式键值对存储系统的安全等级、存储空间、处理资源呈正相关。

17. 一种数据存储装置，应用于分布式键值对存储系统中的物理节点；所述装置包括：

第一接收单元，接收主节点发送的数据；所述数据由客户端发送至主节点，所述主节点由客户端基于一致性哈希算法对所述数据进行计算而确定出；

存储单元，存储所述数据。

18. 根据权利要求 17 所述的装置，还包括：

第二接收单元，接收所述客户端发送的针对所述数据的读取请求，所述读取请求由所述客户端在所述主节点退出所述分布式键值对存储系统时发送；

返回单元，向所述客户端返回所述数据。

19. 根据权利要求 18 所述的装置，还包括：

切换单元，切换为所述数据的主节点；

确定单元，确定与自身相关联的备用节点；

发送单元，向所述备用节点发送所述数据，以使所述备用节点存储所述数据。

20. 一种数据存储装置，应用于分布式键值对存储系统中的客户端；所述装置包括：

第一确定单元，基于一致性哈希算法对数据进行计算，以确定出所述分布式键值对存储系统中作为所述数据的主节点；

第一发送单元，向所述主节点发送所述数据，以使所述主节点存储所述数据，以及确定出与自身相关联的备用节点并将所述数据发送至所述备用节点；

或者，包括：第二确定单元，确定出与所述主节点相关联的备用节点；

第二发送单元，分别向所述主节点和所述备用节点发送所述数据，以使得所述主节点和所述备用节点存储所述数据。

21. 一种电子设备，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器通过运行所述可执行指令以实现如权利要求 1-6 中任一项所述的方法。

22. 一种计算机可读存储介质，其上存储有计算机指令，该指令被处理器执行时实现如权利要求 1-6 中任一项所述方法的步骤。

23. 一种电子设备，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器通过运行所述可执行指令以实现如权利要求 7-9 中任一项所述的方法。

24. 一种计算机可读存储介质，其上存储有计算机指令，该指令被处理器执行时实现如权利要求 7-9 中任一项所述方法的步骤。

25. 一种电子设备，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器通过运行所述可执行指令以实现如权利要求 10 所述的方法。

26. 一种计算机可读存储介质，其上存储有计算机指令，该指令被处理器执行时实现如权利要求 10 所述方法的步骤。

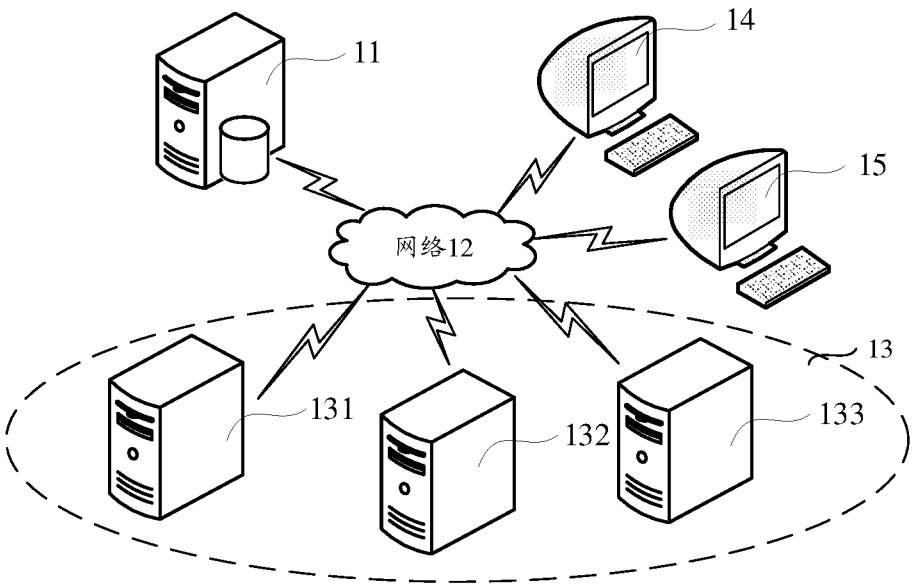

11
14
15
网络12
13
131
132
133

图 1

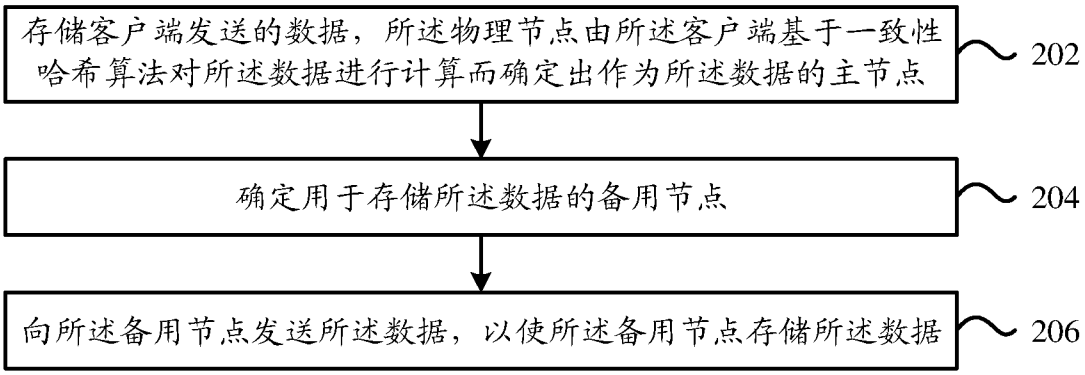

存储客户端发送的数据，所述物理节点由所述客户端基于一致性哈希算法对所述数据进行计算而确定出作为所述数据的主节点
202
确定用于存储所述数据的备用节点
204
向所述备用节点发送所述数据，以使所述备用节点存储所述数据
206

图 2

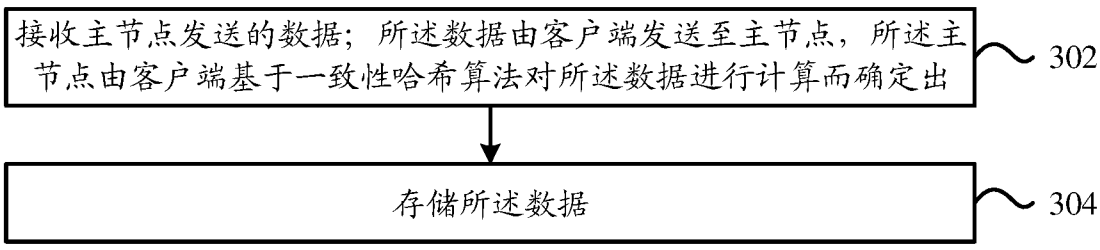

接收主节点发送的数据；所述数据由客户端发送至主节点，所述主节点由客户端基于一致性哈希算法对所述数据进行计算而确定出
302
存储所述数据
304

图 3

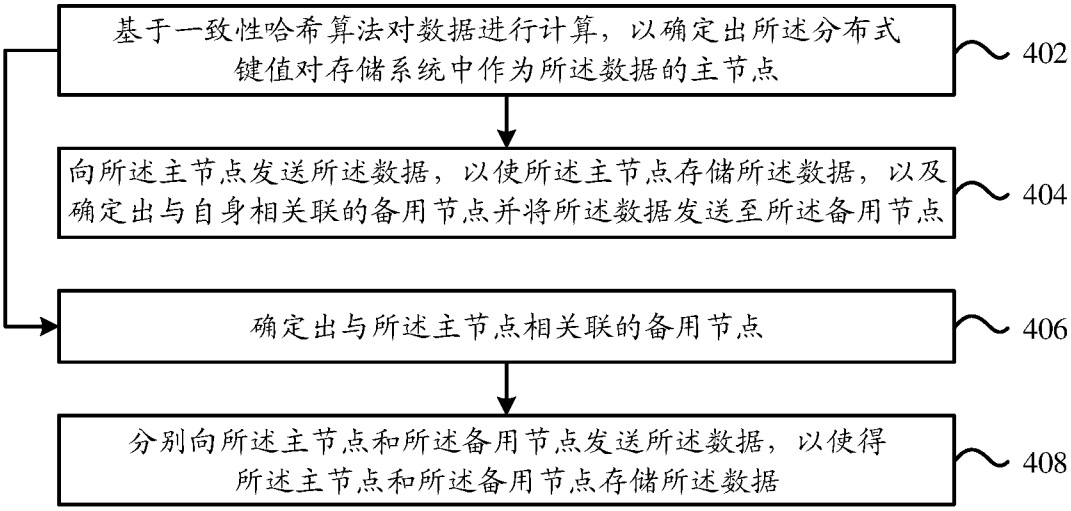

基于一致性哈希算法对数据进行计算，以确定出所述分布式键值对存储系统中作为所述数据的主节点
402
向所述主节点发送所述数据，以使所述主节点存储所述数据，以及确定出与自身相关联的备用节点并将所述数据发送至所述备用节点
404
确定出与所述主节点相关联的备用节点
406
分别向所述主节点和所述备用节点发送所述数据，以使得所述主节点和所述备用节点存储所述数据
408

图 4

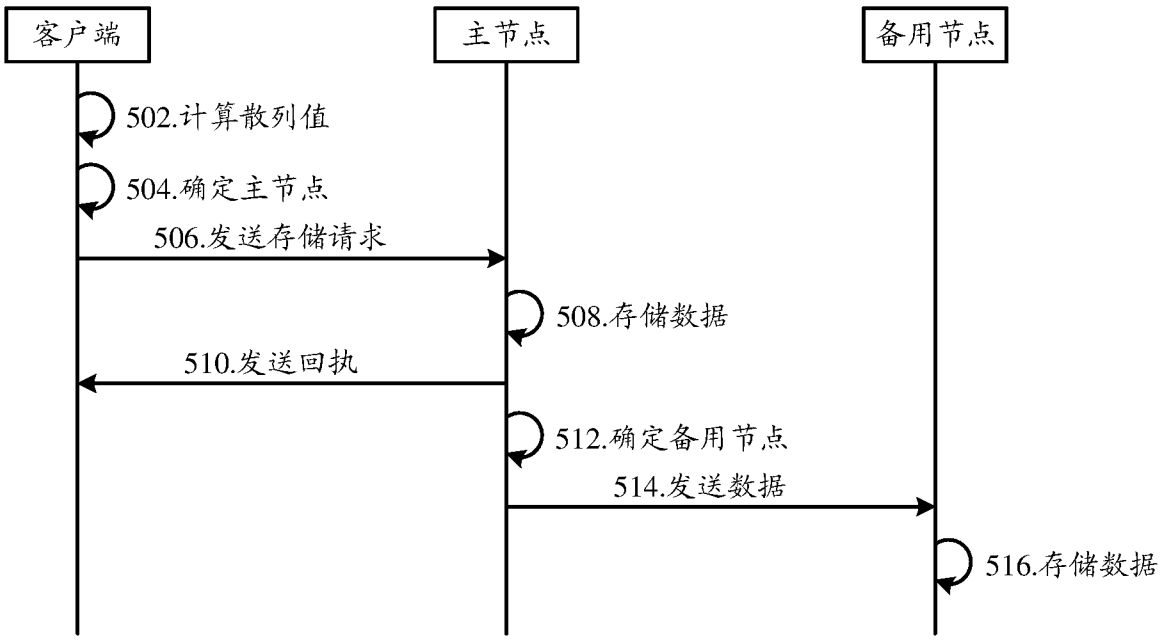

客户端
主节点
备用节点
502.计算散列值
504.确定主节点
506.发送存储请求
508.存储数据
510.发送回执
512.确定备用节点
514.发送数据
516.存储数据

图 5

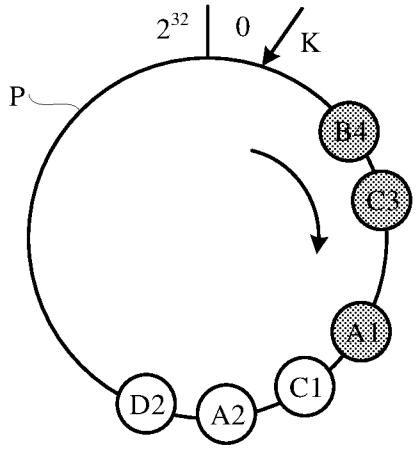

2^{32}
0
K
P
B4
C3
A1
C1
A2
D2

图 6A

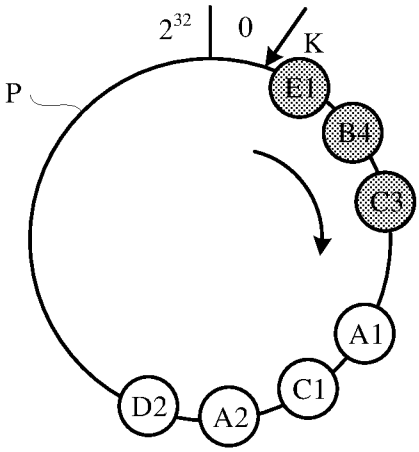

2^{32}
0
K
P
E1
B4
C3
A1
C1
A2
D2

图 6B

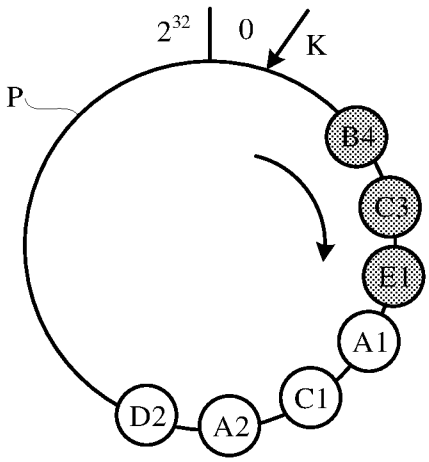

2^{32}
0
K
P
B4
C3
E1
A1
C1
A2
D2

图 6C

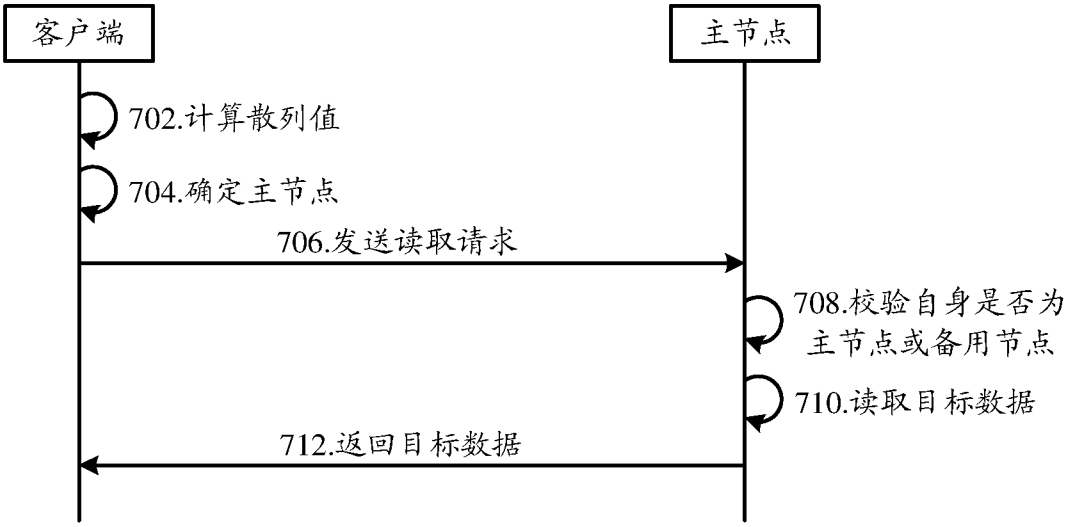

客户端
主节点
702.计算散列值
704.确定主节点
706.发送读取请求
708.校验自身是否为
主节点或备用节点
710.读取目标数据
712.返回目标数据

图 7

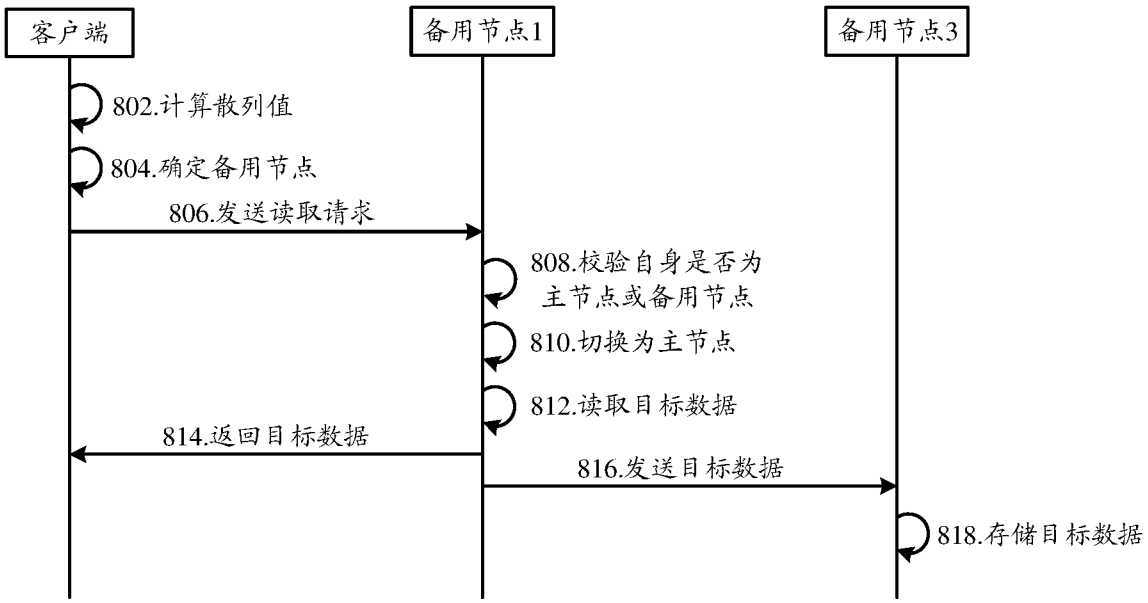

客户端
备用节点1
备用节点3
802.计算散列值
804.确定备用节点
806.发送读取请求
808.校验自身是否为
主节点或备用节点
810.切换为主节点
812.读取目标数据
814.返回目标数据
816.发送目标数据
818.存储目标数据

图 8

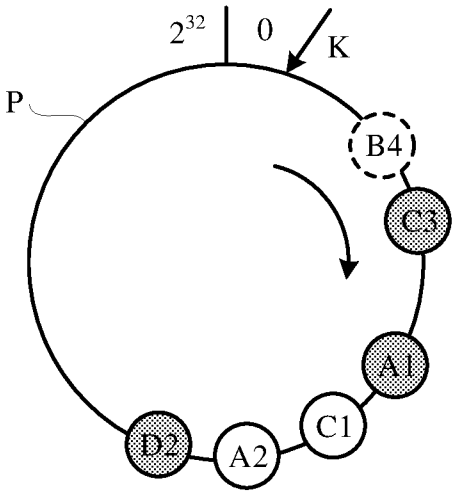

2^{32}
0
K
P
B4
C3
A1
C1
A2
D2

图 9

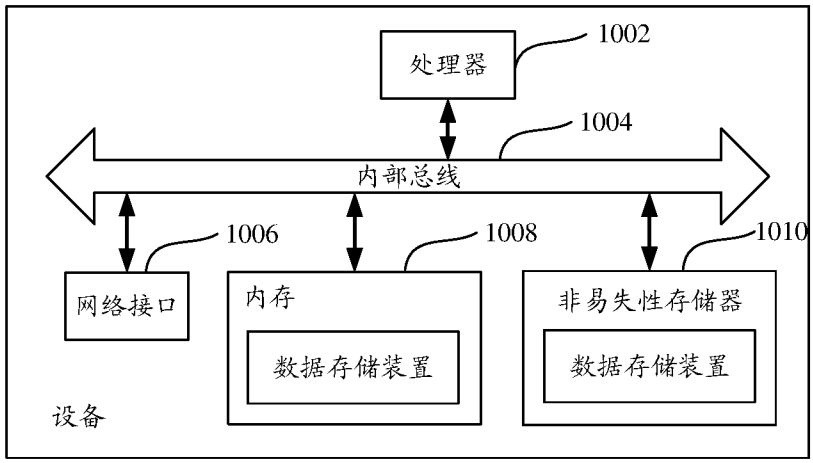

处理器
1002
1004
内部总线
1006
1008
1010
网络接口
内存
数据存储装置
非易失性存储器
数据存储装置
设备

图 10

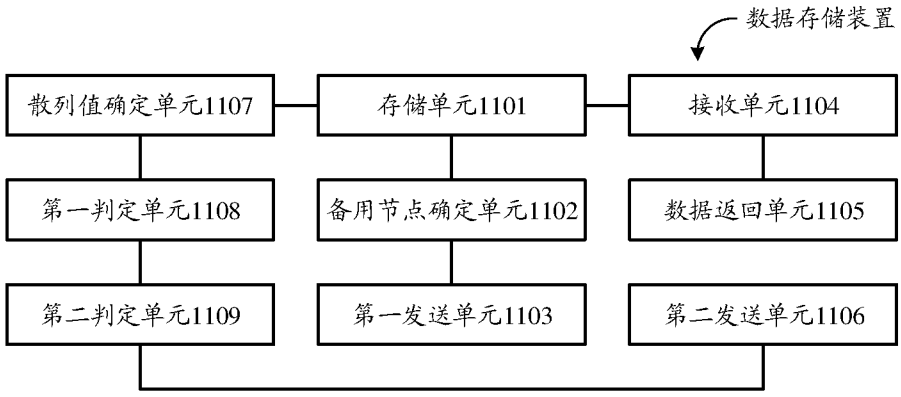

数据存储装置
散列值确定单元1107
存储单元1101
接收单元1104
第一判定单元1108
备用节点确定单元1102
数据返回单元1105
第二判定单元1109
第一发送单元1103
第二发送单元1106

图 11

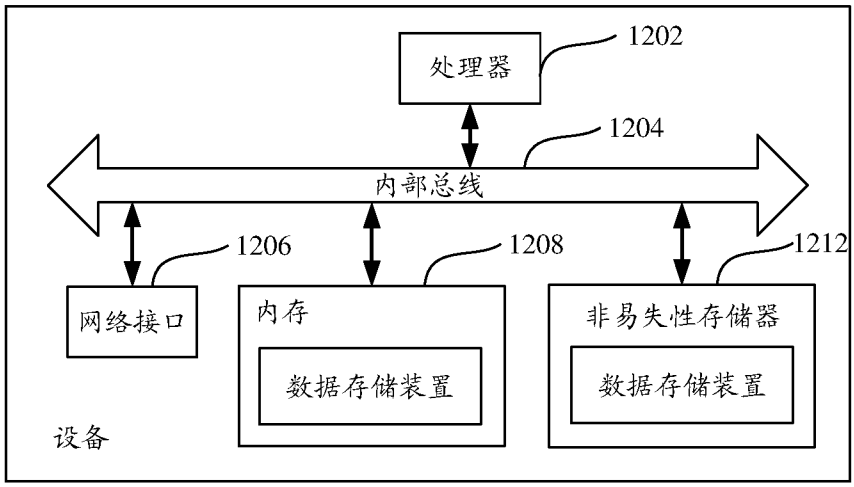

处理器
1202
1204
内部总线
1206
1208
1212
网络接口
内存
数据存储装置
非易失性存储器
数据存储装置
设备

图 12

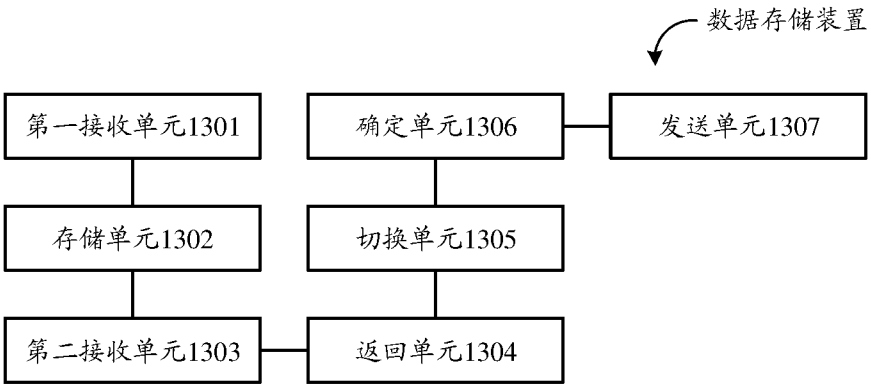

数据存储装置
第一接收单元1301
确定单元1306
发送单元1307
存储单元1302
切换单元1305
第二接收单元1303
返回单元1304

图 13

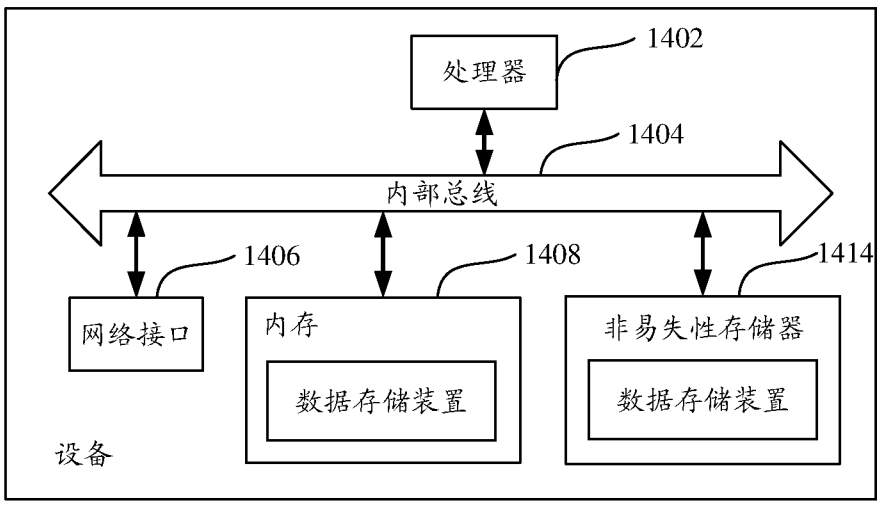

处理器
1402
1404
内部总线
1406
1408
1414
网络接口
内存
数据存储装置
非易失性存储器
数据存储装置
设备

图 14

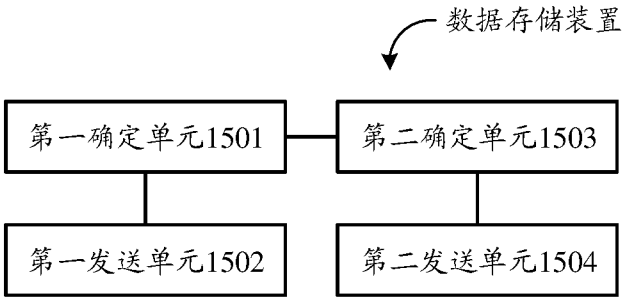


图 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/125871

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)

WPI, EPODOC, CNKI, CNPAT, IEEE: 数据, 存储, 分布式, 键值对, 哈希, 环, 主节点, 备用节点; data, storage, distributed, key-value pair, hash, ring, primary node, standby node

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110049091 A (ALIBABA GROUP HOLDING LIMITED) 23 July 2019 (2019-07-23) description, paragraphs 3-180	1-26
X	CN 106572153 A (LETV HOLDING (BEIJING) CO., LTD. et al.) 19 April 2017 (2017-04-19) description, paragraphs 34-119	1-26
X	WO 2011157144 A2 (HUAWEI TECHNOLOGIES CO., LTD.) 22 December 2011 (2011-12-22) description, page 5, paragraph 5 to page 16, paragraph 5	1-26
A	CN 106254240 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 21 December 2016 (2016-12-21) entire document	1-26

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"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)
"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
"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
"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
"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
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
02 March 2020	**18 March 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/125871

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110049091	A	23 July 2019	None			
CN	106572153	A	19 April 2017	None			
WO	2011157144	A2	22 December 2011	US	8938604	B2	20 January 2015
				US	2013111187	A1	02 May 2013
				CN	102918509	A	06 February 2013
				CN	102918509	B	04 June 2014
				EP	2698718	A2	19 February 2014
CN	106254240	A	21 December 2016	CN	106254240	B	05 July 2019

国际检索报告

国际申请号
PCT/CN2019/125871

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））

WPI,EPODOC,CNKI,CNPAT,IEEE:数据，存储，分布式，键值对，哈希，环，主节点，备用节点；data, storage, distributed, key-value pair, hash, ring, primary node, standby node

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
PX	CN 110049091 A (阿里巴巴集团控股有限公司) 2019年 7月 23日 (2019-07-23) 说明书第3-180段	1-26
X	CN 106572153 A (乐视控股北京有限公司等) 2017年 4月 19日 (2017-04-19) 说明书第34-119段	1-26
X	WO 2011157144 A2 (华为技术有限公司) 2011年 12月 22日 (2011-12-22) 说明书第5页第5段至第16页第5段	1-26
A	CN 106254240 A (腾讯科技深圳有限公司) 2016年 12月 21日 (2016-12-21) 全文	1-26

☐ 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	国际检索报告邮寄日期
2020年 3月 2日	2020年 3月 18日
ISA/CN的名称和邮寄地址 中国国家知识产权局(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	受权官员 汤广强 电话号码 86-（10）-53961734

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/125871

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	110049091	A	2019年 7月 23日	无			
CN	106572153	A	2017年 4月 19日	无			
WO	2011157144	A2	2011年 12月 22日	US	8938604	B2	2015年 1月 20日
				US	2013111187	A1	2013年 5月 2日
				CN	102918509	A	2013年 2月 6日
				CN	102918509	B	2014年 6月 4日
				EP	2698718	A2	2014年 2月 19日
CN	106254240	A	2016年 12月 21日	CN	106254240	B	2019年 7月 5日