(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
2015 年 6 月 11 日 (11.06.2015)

WIPO | PCT

(10) 国际公布号
WO 2015/081750 A1

(51) 国际专利分类号:
H04L 29/08 (2006.01)

(21) 国际申请号: PCT/CN2014/087621

(22) 国际申请日: 2014 年 9 月 26 日 (26.09.2014)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201310641751.7 2013 年 12 月 3 日 (03.12.2013) CN

(71) 申请人: **北京京东尚科信息技术有限公司 (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO, LTD.)** [CN/CN]; 中国北京市海淀区杏石口路 65 号西杉创意园四区 11C 楼东段 1-4 层西段 1-4 层, Beijing 100195 (CN)。 **北京京东世纪贸易有限公司 (BEIJING JINGDONG CENTURY TRADING CO., LTD.)** [CN/CN]; 中国北京市经济技术开发区科创十四街 99 号 2 号楼 B168 室, Beijing 100176 (CN)。

(72) 发明人: **罗文彬 (LUO, Wenbin)**; 中国北京市海淀区杏石口路 65 号西杉创意园四区 11C 楼东段 1-4 层西段 1-4 层, Beijing 100195 (CN)。 **张波 (ZHANG, Bo)**; 中国北京市海淀区杏石口路 65 号西杉创意园四区 11C 楼东段 1-4 层西段 1-4 层, Beijing 100195 (CN)。 **杨万元 (YANG, Wanyuan)**; 中国北京市海淀区杏石口路 65 号西杉创意园四区 11C 楼东段 1-4 层西段 1-4 层, Beijing 100195 (CN)。 **杨小龙 (YANG, Xiaolong)**; 中国北京市海淀区杏石口路 65

[见续页]

(54) **Title**: DATA COPY STORAGE METHOD, TERMINAL APPARATUS, AND SERVER APPARATUS

(54) **发明名称**: 一种存储数据副本的方法和终端装置以及服务器装置

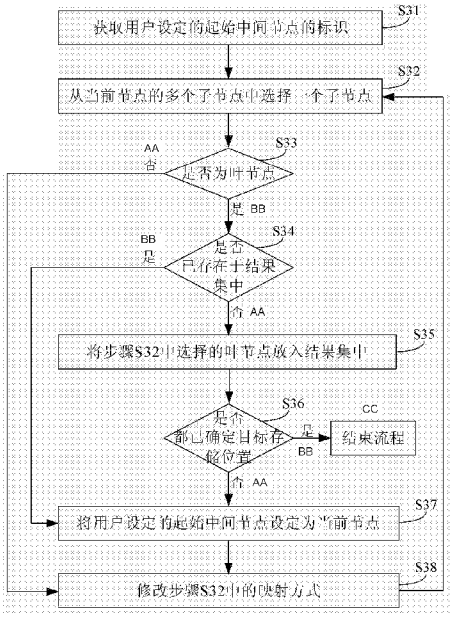


图3 / FIG.3

S31 Obtain an identification of a start intermediate node set by a user
S32 Select a child node from multiple child nodes of a current node
S33 Is a leaf node?
S34 Does the leaf node exist in a result set?
S35 Put the leaf node selected in step S32 into the result set
S36 Are target storage positions determined for all?
S37 Set the start intermediate node set by the user as the current node
S38 Modify a mapping node in step S32
AA No
BB Yes
CC End the process

(57) **Abstract**: Provided are a data copy storage method, a terminal apparatus, and a server apparatus, which help to improve the reliability of a distributed data storage system without a central node and reduce an amount of migrated data when the number of nodes of the system changes. The method comprises: for each data copy to be stored, selecting child nodes level by level from a pre-selected intermediate node until a leaf node according to a tree structure of a distributed data storage system without a central node, and determining the leaf node as a target storage position of the data copy, selection of a child node comprising: combining an identification of data to be stored and an identification of a child node into a primary image, obtaining a mapping value related to the child node by using a preset mapping mode, selecting a sub-node of a central node according to the mapping value, and resetting a mapping relationship before each time when the child node is selected; and storing all data copies to be stored at the target storage positions of the data copies.

(57) **摘要**: 提供一种存储数据副本的方法和终端装置以及服务器装置，有助于提高无中心节点的分布式数据存储系统的可靠性，并降低其在节点数目发生改变时的数据迁移量。该方法包括：根据无中心节点的分布式数据存储系统的树形结构，对于待存数据的每个数据副本，分别从预选中间节点起逐级选择子节点直至叶节点，将该叶节点确定为数据副本的目标存储位置；子节点选择方式为：以待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据映射值选择该中间节点的一个子节点；并且每次选择子节点之前重新设定所述映射关系；将待存数据的所有数据副本保存到各数据副本的目标存储位置。

WO 2015/081750 A1

号西杉创意园四区 11C 楼东段 1-4 层西段 1-4 层, Beijing 100195 (CN)。

(74) **代理人**: **中原信达知识产权代理有限责任公司 (CHINA SINDA INTELLECTUAL PROPERTY LTD.)**; 中国北京市西城区金融街 19 号富凯大厦 B 座 11 层, Beijing 100033 (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。

(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))。

一种存储数据副本的方法和终端装置以及服务器装置

技术领域

本发明涉及计算机技术领域，特别地涉及一种存储数据副本的方法和终端装置以及服务器装置。

背景技术

分布式数据存储系统通常采用多副本的方式存储数据，以提高数据存储的可靠性。存储设备的物理拓扑结构一般是层级式的，参考图 1，图 1 是与本发明有关的分布式数据存储系统中的存储设备的物理拓扑结构的示意图。如图 1 所示，分布式数据存储系统设置在数据中心 10 中，数据中心 10 由 3 个机房组成，每个机房中有多个机架，例如机房 2 中有机架 1、机架 2、……、机架 N1；每个机架上又有多个主机（计算机），例如机架 1 上有主机 1、主机 2、……、主机 N2；每个主机中又有多个存储介质设备（通常是硬盘），例如主机 2 中有硬盘 1、硬盘 2、……、硬盘 N3。为示意清晰起见，图 1 中仅示出了部分设备。可以看出分布式数据存储系统是树形结构，存储介质设备位于叶节点上，而主机、机架、机房则是中间节点。

分布式数据存储系统可以分为带有中心节点和无中心节点两种形式。在带有中心节点的分布式数据存储系统中，一般含有客户端、中心节点及存储节点，同时将数据分块处理，多副本存储，而数据的分布位置由中心节点根据存储节点的负载情况、副本的存储策略来决定数据副本的存储位置。中心节点可以是两台服务器主备配置，也可以是服务器集群。在无中心节点的分布式数据存储系统中，各个节点彼此互相连接，数据随机的分布在各个节点的存储设备上，存储位置可以由节点使用哈希函数计算得到。

对于带有中心节点的分布式数据存储系统来说，从其中读取数据

时，需先访问中心节点以获取数据的存储位置，这使得中心节点负荷过重，处理效率降低，从而形成整个系统的瓶颈，降低了系统性能。并且在中心节点出现故障的情况下，该瓶颈会进一步加剧甚至使整个分布式数据存储系统不可用，可靠性较低。

无中心节点的分布式数据存储系统能够避免上述瓶颈，但其可靠性仍显不足。例如某个机架断电或出现网络故障，并且某项数据的所有副本都在该机架的主机的硬盘上，则该项数据不可获取。另外，在设备数量发生改变，例如增加一块硬盘，或者增加一台主机等情况，无中心节点的分布式数据存储系统必然出现广泛的数据迁移，以下举一个简单的例子对此加以说明。

例如无中心节点的分布式数据存储系统中有 5 个节点（实际上远不止 5 个），由哈希函数计算得到一项待存数据的哈希值为 13，按节点数取模，则为 3，于是该待存数据被保存在节点 3 上；当增加设备，使节点数为 6，则在读取数据时，将哈希值 13 对节点数 6 取模得 2，即从节点 2 上读取数据，此时数据首先会从节点 3 迁移到节点 2。因为节点数改变时取模结果必然不同，所以在节点数变化之后读取节点数变化之前保存的数据时必然发生数据迁移，因此一旦节点数发生变化，数据迁移就成为普遍现象，这将导致系统效率降低以及缩短存储介质寿命。

发明内容

有鉴于此，本发明提供一种存储数据副本的方法和终端装置以及服务器装置，有助于提高无中心节点的分布式数据存储系统的可靠性，并降低其在节点数目发生改变时的数据迁移量。

为实现上述目的，根据本发明的一个方面，提供了一种存储数据副本的方法。

本发明的存储数据副本的方法，应用于具有树形结构的无中心节点的分布式数据存储系统中，所述树形结构包括叶节点和中间节点，存储介质设备位于所述叶节点上，该方法包括：步骤 A：根据所述树形结构，对于待存数据的每个数据副本，分别从所述树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定为该数据副本的目标存储位置；其中所述子节点选择方式为：以所述待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据关联于同一所述中间节点的各个子节点的多个映射值之间的大小关系选择该中间节点的一个子节点；并且对于每个中间节点，在每次选择其子节点之前重新设定所述映射关系；步骤 B：在确定了所述待存数据的所有数据副本的目标存储位置之后，将所述待存数据的所有数据副本保存到各数据副本的目标存储位置。

可选地，对于指定的一个或多个所述中间节点，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若该中间节点被再次选择，则此时从该中间节点的所有兄弟节点中重新进行选择。

可选地，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若对当前数据副本确定出的目标存储位置与另一数据副本的已确定出的目标存储位置相同，则对当前数据副本按所述步骤 A 重新确定目标存储位置，使所述待存数据的所有数据副本的目标存储位置各不相同。

可选地，所述映射方式包括：将所述原像调整之后使用哈希算法计算出唯一的映射值。

可选地，所述映射方式包括：将所述原像调整之后使用哈希算法计算出哈希值；将所述哈希值与预设权重值相乘得到所述映射值；其中所述预设权重值是与所述原像中的子节点连通的所有叶节点的权重

值之和，所述叶节点的权重值与该叶节点上的存储介质设备的存储容量正相关。

可选地，所述重新设定所述映射关系的步骤包括：改变将所述原像调整时采用的参数。

根据本发明的另一方面，提供了一种终端装置。

本发明的终端装置用于将待存数据的多个数据副本保存到具有树形结构的无中心节点的分布式数据存储系统中，所述树形结构包括叶节点和中间节点，存储介质设备位于所述叶节点上，该终端装置包括：选择模块，用于根据所述树形结构，对于待存数据的每个数据副本，分别从所述树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定为该数据副本的目标存储位置；其中所述子节点选择方式为：以所述待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据关联于同一所述中间节点的各个子节点的多个映射值之间的大小关系选择该中间节点的一个子节点；并且对于每个中间节点，在每次选择其子节点之前重新设定所述映射关系；上传模块，用于在所述选择模块确定了所述待存数据的所有数据副本的目标存储位置之后，将所述待存数据的所有数据副本和所述待存数据的所有数据副本的目标存储位置的信息发送给所述分布式数据存储系统中的计算机，以供所述计算机将所述待存数据的所有数据副本保存到各数据副本的目标存储位置。

可选地，所述选择模块还用于：对于指定的一个或多个所述中间节点，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若该中间节点被再次选择，则此时从该中间节点的所有兄弟节点中重新进行选择。

可选地，所述选择模块还用于：在确定所述待存数据的所有数据副本的目标存储位置的过程中，若对当前数据副本确定出的目标存储位置与另一数据副本的已确定出的目标存储位置相同，则对当前数据副本重新确定目标存储位置，使所述待存数据的所有数据副本的目标存储位置各不相同。

可选地，所述映射方式包括：将所述原像调整之后使用哈希算法计算出唯一的映射值。

可选地，所述映射方式包括：将所述原像调整之后使用哈希算法计算出哈希值；将所述哈希值与预设权重值相乘得到所述映射值；其中所述预设权重值是与所述原像中的子节点连通的所有叶节点的权重值之和，所述叶节点的权重值与该叶节点上的存储介质设备的存储容量正相关。

根据本发明的又一方面，提供了一种服务器装置。

本发明的服务器装置，设置在具有树形结构的无中心节点的分布式数据存储系统中的服务器内，用于将待存数据的多个数据副本保存到所述分布式数据存储系统中，所述树形结构包括叶节点和中间节点，存储介质设备位于所述叶节点上，该服务器装置包括：选择模块，用于根据所述树形结构，对于待存数据的每个数据副本，分别从所述树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定为该数据副本的目标存储位置；其中所述子节点选择方式为：以所述待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据关联于同一所述中间节点的各个子节点的多个映射值之间的大小关系选择该中间节点的一个子节点；并且对于每个中间节点，在每次选择其子节点之前重新设定所述映射关系；保存模块，用于在所述选择模块确定了所述待存数据的所有数据副本的目标存储

位置之后，将所述待存数据的所有数据副本保存到各数据副本的目标存储位置。

可选地，所述选择模块还用于：对于指定的一个或多个所述中间节点，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若该中间节点被再次选择，则此时从该中间节点的所有兄弟节点中重新进行选择。

可选地，所述选择模块还用于：在确定所述待存数据的所有数据副本的目标存储位置的过程中，若对当前数据副本确定出的目标存储位置与另一数据副本的已确定出的目标存储位置相同，则对当前数据副本重新确定目标存储位置，使所述待存数据的所有数据副本的目标存储位置各不相同。

可选地，所述映射方式包括：将所述原像调整之后使用哈希算法计算出唯一的映射值。

可选地，所述映射方式包括：将所述原像调整之后使用哈希算法计算出哈希值；将所述哈希值与预设权重值相乘得到所述映射值；其中所述预设权重值是与所述原像中的子节点连通的所有叶节点的权重值之和，所述叶节点的权重值与该叶节点上的存储介质设备的存储容量正相关。

可选地，还包括确认模块，用于在所述分布式数据存储系统中的节点发生改变的情况下触发当前所述分布式数据存储系统中保存的多项数据的各自的一个预选副本所在的服务器中的所述选择模块；所述选择模块还用于在被所述确认模块触发的情况下将所述预选副本以及与该副本相同的其他副本作为所述待存数据，确定该待存数据中的各个数据副本的新的目标存储位置，然后将该各个数据副本中原存储位置不为该新的目标存储位置的数据副本迁移到该新的目标存储位置。

根据本发明的技术方案，在确定副本的目标存储位置时，对于待存数据的每个数据副本，分别从分布式数据存储系统的树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定为该数据副本的目标存储位置，也就是说，从选定的中间节点起层层选择子节点直至存储介质设备所在的叶节点，并且对于每个中间节点来说，每次选择其子节点时的选择方式有变化从而使选择结果尽可能地改变，在层层选择子节点的情况下，这种改变能够使多个副本的目标存储位置尽可能地分散，有助于提高分布式数据存储系统在设备出现故障时的鲁棒性，从而提高了分布式数据存储系统的可靠性。并且可以指定一个或多个中间节点，使一项待存数据只有一个副本在该中间节点的存储设备上，也就是说将数据副本分布在不同的中间节点例如机架或主机上，这样当一个中间节点失效例如机架或者一台主机不能服务时，仍有其他副本能够服务，从而减少数据丢失以及不可获取性的风险及概率，提高分布式数据存储系统的可靠性。在分布式数据存储系统新增存储设备的情况下，有相当多的数据副本并不需要迁移，数据迁移量相当小，对系统效率的影响也较低，使得分布式数据存储系统比较易于扩容。

附图说明

附图用于更好地理解本发明，不构成对本发明的不当限定。其中：

图 1 是与本发明有关的分布式数据存储系统中的存储设备的物理拓扑结构的示意图；

图 2 是根据本发明实施例的终端设备与分布式数据存储系统连接状态的示意图；

图 3 是根据本发明实施例的存储数据副本的方法的示意图；

图 4 是根据本发明实施例的终端装置的基本结构的示意图；

图 5 是根据本发明实施例的服务器装置的基本结构的示意图。

具体实施方式

以下结合附图对本发明的示范性实施例做出说明，其中包括本发明实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本发明的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

图 2 是根据本发明实施例的终端设备与分布式数据存储系统连接状态的示意图。如图 2 所示，多个终端设备 21、22、……、2M 与分布式数据存储系统 20 连接，本实施例中该分布式数据存储系统 20 的物理拓扑同样采用图 1 所示的树状结构。各个终端设备保存有该物理拓扑的信息，从而终端设备能够计算要存储到系统 20 中的各数据副本的目标存储位置。

终端设备对于待存到系统 20 中的数据的多个数据副本，依次计算每个数据副本在系统 20 中的目标存储位置，也就是说为每个数据副本确定出系统 20 中的一个叶节点，数据副本保存在位于该叶节点的存储介质设备中。在确定叶节点时，从用户在系统 20 中指定的一个中间节点起，逐级选择子节点直至被选子节点为叶节点。确定一个数据副本的目标存储位置的流程如图 3 所示。图 3 是根据本发明实施例的存储数据副本的方法的示意图。

步骤 S31：获取用户设定的起始中间节点的标识。从本步骤起，通过循环的流程，对一项待存数据的各个副本的目标存储位置加以确定，得到一个结果集，该结果集是一项待存数据的所有副本的目标存储位置的集合。在以下的步骤中，参考图 1，以起始中间节点为机房 2 为例进行说明。

步骤 S32：从当前节点的多个子节点中选择一个子节点。在第一次执行本步骤时，该当前节点为用户设定的起始中间节点。在选择子节点时，先得出每个子节点的得分，优选的方式是根据待存数据的标

识（例如数据的文件名）和子节点的标识按照策略中的映射方式得出一个映射值，该映射值即为该子节点对于本次待存数据的得分。在上述的映射方式中，具体可以采用哈希算法进行计算。因为在后面的步骤 S36 中需要改变映射方式，所以在本步骤中，对于待存数据的标识和子节点的标识以及一个可调整的参数这三者的组合，采用哈希函数进行计算得到一个哈希值，将该哈希值作为上述得分。在得出起始中间节点的每个子节点的得分之后，根据各个得分的大小选择一个子节点，例如选择得分最高的子节点。

另一种映射方式即计算上述的得分的方式是考虑到尽量把数据保存到容量较大的存储介质设备中，因此对每个叶节点赋予权重，该权重与该叶节点上的存储介质设备的存储容量正相关，例如 1TB 的硬盘的权重为 1，3TB 的硬盘的权重为 3。中间节点的权重是其所有子节点权重之和。节点的得分即在上述的哈希值的基础上再乘以节点的权重。这样，在按照得分最高来选择一个中间节点的子节点时，存储容量更大的节点就有更大的可能性被选取。

步骤 S33：判断步骤 S32 中选择的节点是否为叶节点。若是，则进入步骤 S34，否则进入步骤 S38。在步骤 S38 之后返回步骤 S32，此时步骤 S32 中的当前节点为步骤 S33 中被判断的节点。

步骤 S34：判断步骤 S32 中选择的叶节点是否已存在于结果集中。为了提高可靠性，要使一项待存数据的多个副本分布在不同的存储介质设备中。所以本步骤的判断结果若为是，则舍弃本次对叶节点的选择，并且进入步骤 S37，并在再次执行的步骤 S32 中对当前副本的目标存储位置重新加以确定；否则保留本次对叶节点的选择，进入步骤 S35。

步骤 S35：将步骤 S32 中选择的叶节点放入结果集中。本步骤之后进入步骤 S36。

步骤 S36：判断当前待存数据的所有副本是否都已确定目标存储位置，若是，结束流程，否则进入步骤 S37。

步骤 S37：将用户设定的起始中间节点设定为当前节点。然后进入步骤 S38。

步骤 S38：修改步骤 S32 中的映射方式。即改变步骤 S32 中的可调整的参数。这样做的目的是改变节点的得分从而改变步骤 S32 中对节点的选择结果。对该可调整的参数进行调整的方式可以有很多，只要达到了改变最终的得分即可，可以是将生成的随机数作为该参数，也可以是每次调整时对该参数累加一个大于零的固定或随机值。当然，也可以采用改变哈希函数的方法来修改步骤 S32 中的映射方式。本步骤之后返回步骤 S32。

从以上的流程可以看出，在确定副本的目标存储位置时，是从选定的中间节点起层层选择子节点直至存储介质设备所在的叶节点，并且对于每个中间节点来说，每次选择其子节点时的选择方式有变化从而使选择结果尽可能地改变，在层层选择子节点的情况下，这种改变能够使多个副本的目标存储位置尽可能地分散，有助于提高分布式数据存储系统在设备出现故障时的鲁棒性，从而提高了分布式数据存储系统的可靠性。

当然，如上所述，在以上的流程中每次选择其子节点时的选择方式有变化从而使选择结果尽可能地改变，但这种改变并不是绝对的，在实际中仍以一个较小的概率会出现两次选择结果相同的情况。例如副本 1 被确定在机架 1 下的某个主机的硬盘上，副本 2 也被确定在机架 1 下的某个主机的硬盘上，这样这两个副本就会因为机架 1 的整体故障（例如断电、网络故障等）变得不可用。另外在实际的分布式数据存储系统中，更为理想的方式是将每个副本都分布在某一层级的不同存储位置上，例如有 8 个副本，系统中有 10 个机架，则最好是这 8

个副本分布在不同的机架上。或者因某些实际情况，有些机架的可靠性较低，则最好不要使多个副本都处于这样的机架上的存储设备中。所以在本实施例中，一个优选的方式是指定一个或多个中间节点，当执行图 3 所示的流程时，向步骤 S32 和步骤 S33 之间增加一个步骤，在该步骤中判断步骤 S32 中选择的子节点是否曾经被选择过。这里分两种情况，一种情况是对于一个数据副本，在第一次确定其目标存储位置之后发现该目标存储位置上已有其他数据副本（即步骤 S34 中判断结果为是的情形），在该重新确定该数据副本的目标存储位置的过程中，选择的子节点与第一次确定该数据副本的目标存储位置时选择的子节点相同。这种情况的概率较低，因为中间要经过步骤 S38。另一种情况是对于不同的数据副本，在确定它们的目标存储位置的过程中选择了同一子节点，例如确定数据副本 1 的目标存储位置是主机 2 中的硬盘 2，确定数据副本 2 的目标存储位置的过程中，在一次执行步骤 S32 时也选中了主机 2。此时在步骤 S32 和步骤 S33 之间增加的步骤中判断当前选择的中间节点是否曾经被选择过，并且判断结果为“是”，则直接进入步骤 S38，并且在接下来重新执行的步骤 S32 中，将被重复选择的中间节点的父节点作为当前节点，也就是说在该被重复选择的中间节点的兄弟节点中重新选择。如上例，确认主机 2 曾经被选择过，则在主机 2 所在机架中重新选择另一台主机。通过以上的处理，能够使一项待存数据在指定的中间节点例如机架、主机等上只有一个副本，也就是说将数据副本分布在不同的机架或主机上，这样当一个机架或者一台主机不能服务时，仍有其他副本能够服务，从而减少数据丢失以及不可获取性的风险及概率，提高分布式数据存储系统的可靠性。

由终端设备执行图 3 所示的流程，能够对一项待存数据的各个数据副本在分布式数据存储系统中的目标存储位置得以确定。对于该分布式数据存储系统来说，由于数据量的不断增加，系统中常常需要新增存储设备，例如增加一个硬盘、增加一台主机、或者增加一个机架等，也就是说节点会发生改变；并且需要将已有设备中的一些数据迁移到该新增存储设备中。对于这种情况，在本实施例中，分布式数据

存储系统中的主机也具备上述的终端设备所具有的确定数据副本的目标存储位置的功能，从而能够在分布式数据存储系统新增存储设备的情况下，对该系统中的已存数据的新的存储位置加以确定，需要说明的是，只要系统中出现新增存储设备的情况，各个主机都会进行计算以做出该确定。当然这种计算仅是在系统中出现新增存储设备时进行，而且从图 3 的流程可以看出这种计算比较简单，占用的 CPU 资源较少，所以主机重新确定已存数据的新的存储位置的工作并不会对系统性能带来太大的影响。

另外，对于一项已存数据，只需由一台主机的一个线程来执行上述计算，该是保存有该项已存数据的一个副本的主机，可以在终端设备确定该项已存数据的各个数据副本的目标存储位置时就设定该主机。例如，终端设备依次确定了待存数据 A 的第 1 个至第 8 个副本的目标存储位置，其中第 1 个副本将保存在第 2 机架的第 4 主机的一个硬盘上，然后将数据发送给各个目标存储位置所在的主机，当分布式数据存储系统新增存储设备时，由保存数据 A 的第 1 个副本（即最先被终端设备确定目标存储位置的副本）的主机即第 2 机架的第 4 主机计算数据 A 的 8 个副本的新的目标存储位置，此时该第 4 主机可以按图 3 的流程来执行。从图 3 的流程可以看出，对于这 8 个副本来说，新的目标存储位置有可能与原有存储位置相同，特别是在物理拓扑上存储位置与新增存储设备距离较远的数据副本，存储位置发生改变的可能性相当低。也就是说分布式数据存储系统新增存储设备时，有相当多的数据副本并不需要迁移，数据迁移量相当小，对系统效率的影响也较低，使得分布式数据存储系统比较易于扩容。

以下对于本发明实施例中的终端装置的基本结构做出说明。该终端装置设置在终端设备例如个人计算机中，通过互联网与具有树形结构的分布式数据存储系统连接。图 4 是根据本发明实施例的终端装置的基本结构的示意图。如图 4 所示，终端装置 40 主要包括选择模块 41 和上传模块 42。

选择模块 41 用于根据分布式数据存储系统的树形结构，对于待存数据的每个数据副本，分别从该树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定为该数据副本的目标存储位置；其中子节点选择方式为：以待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据关联于同一中间节点的各个子节点的多个映射值之间的大小关系选择该中间节点的一个子节点；并且对于每个中间节点，在每次选择其子节点之前重新设定映射关系。

上传模块 42 用于在选择模块 41 确定了待存数据的所有数据副本的目标存储位置之后，将待存数据的所有数据副本和待存数据的所有数据副本的目标存储位置的信息发送给分布式数据存储系统中的计算机，以供该计算机将待存数据的所有数据副本保存到各数据副本的目标存储位置。

选择模块 41 还可用于：对于指定的一个或多个中间节点，在确定待存数据的所有数据副本的目标存储位置的过程中，若该中间节点被再次选择，则此时从该中间节点的所有兄弟节点中重新进行选择。

选择模块 41 还可用于：在确定待存数据的所有数据副本的目标存储位置的过程中，若对当前数据副本确定出的目标存储位置与另一数据副本的已确定出的目标存储位置相同，则对当前数据副本重新确定目标存储位置，使待存数据的所有数据副本的目标存储位置各不相同。

图 5 是根据本发明实施例的服务器装置的基本结构的示意图。该服务器装置设置在具有树形结构的分布式数据存储系统中的服务器内，用于将待存数据的多个数据副本保存到该分布式数据存储系统中。如图 5 所示，服务器装置 50 主要包括选择模块 51 和保存模块 52。其

中选择模块 51 可以采用与选择模块 51 类似的结构，而保存模块 52 用于在选择模块 51 确定了待存数据的所有数据副本的目标存储位置之后，将待存数据的所有数据副本保存到各数据副本的目标存储位置。

服务器装置 50 还可以包括确认模块（图中未示出），该确认模块用于在分布式数据存储系统中的节点发生改变的情况下触发当前分布式数据存储系统中保存的多项数据的各自的一个预选副本所在的服务器中的选择模块 51；这样，选择模块 51 还用于在被确认模块触发的情况下将该预选副本以及与该副本相同的其他副本作为待存数据，确定该待存数据中的各个数据副本的新的目标存储位置，然后将该各个数据副本中原存储位置不为该新的目标存储位置的数据副本迁移到该新的目标存储位置。

以上结合具体实施例描述了本发明的基本原理，但是，需要指出的是，对本领域的普通技术人员而言，能够理解本发明的方法和设备的全部或者任何步骤或者部件，可以在任何计算装置（包括处理器、存储介质等）或者计算装置的网络中，以硬件、固件、软件或者它们的组合加以实现，这是本领域普通技术人员在阅读了本发明的说明的情况下运用他们的基本编程技能就能实现的。

因此，本发明的目的还可以通过在任何计算装置上运行一个程序或者一组程序来实现。所述计算装置可以是公知的通用装置。因此，本发明的目的也可以仅仅通过提供包含实现所述方法或者装置的程序代码的程序产品来实现。也就是说，这样的程序产品也构成本发明，并且存储有这样的程序产品的存储介质也构成本发明。显然，所述存储介质可以是任何公知的存储介质或者将来开发出的任何存储介质。

还需要指出的是，在本发明的装置和方法中，显然，各部件或各步骤是可以分解和/或重新组合的。这些分解和/或重新组合应视为本发明的等效方案。并且，执行上述系列处理的步骤可以自然地按照说明

的顺序按时间顺序执行，但是并不需要一定按照时间顺序执行。某些步骤可以并行或彼此独立地执行。

上述具体实施方式，并不构成对本发明保护范围的限制。本领域技术人员应该明白的是，取决于设计要求和其他因素，可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明保护范围之内。

权 利 要 求 书

1. 一种存储数据副本的方法，应用于具有树形结构的无中心节点的分布式数据存储系统中，所述树形结构包括叶节点和中间节点，存储介质设备位于所述叶节点上，其特征在于，该方法包括：

步骤 A：根据所述树形结构，对于待存数据的每个数据副本，分别从所述树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定为该数据副本的目标存储位置；

其中所述子节点选择方式为：以所述待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据关联于同一所述中间节点的各个子节点的多个映射值之间的大小关系选择该中间节点的一个子节点；并且对于每个中间节点，在每次选择其子节点之前重新设定所述映射关系；

步骤 B：在确定了所述待存数据的所有数据副本的目标存储位置之后，将所述待存数据的所有数据副本保存到各数据副本的目标存储位置。

2. 根据权利要求 1 所述的方法，其特征在于，对于指定的一个或多个所述中间节点，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若该中间节点被再次选择，则此时从该中间节点的所有兄弟节点中重新进行选择。

3. 根据权利要求 1 所述的方法，其特征在于，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若对当前数据副本确定出的目标存储位置与另一数据副本的已确定出的目标存储位置相同，则对当前数据副本按所述步骤 A 重新确定目标存储位置，使所述待存数据的所有数据副本的目标存储位置各不相同。

4. 根据权利要求 1 所述的方法，其特征在于，所述映射方式包括：

将所述原像调整之后使用哈希算法计算出唯一的映射值。

5. 根据权利要求 1 所述的方法，其特征在于，所述映射方式包括：

将所述原像调整之后使用哈希算法计算出哈希值；

将所述哈希值与预设权重值相乘得到所述映射值；

其中所述预设权重值是与所述原像中的子节点连通的所有叶节点的权重值之和，所述叶节点的权重值与该叶节点上的存储介质设备的存储容量正相关。

6. 根据权利要求 4 或 5 所述的方法，其特征在于，所述重新设定所述映射关系的步骤包括：改变将所述原像调整时采用的参数。

7. 一种终端装置，用于将待存数据的多个数据副本保存到具有树形结构的无中心节点的分布式数据存储系统中，所述树形结构包括叶节点和中间节点，存储介质设备位于所述叶节点上，其特征在于，该终端装置包括：

选择模块，用于根据所述树形结构，对于待存数据的每个数据副本，分别从所述树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定为该数据副本的目标存储位置；其中所述子节点选择方式为：以所述待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据关联于同一所述中间节点的各个子节点的多个映射值之间的大小关系选择该中间节点的一个子节点；并且对于每个中间节点，在每次选择其子节点之前重新设定所述映射关系；

上传模块，用于在所述选择模块确定了所述待存数据的所有数据副本的目标存储位置之后，将所述待存数据的所有数据副本和所述待存数据的所有数据副本的目标存储位置的信息发送给所述分布式数据存储系统中的计算机，以供所述计算机将所述待存数据的所有数据副本保存到各数据副本的目标存储位置。

8. 根据权利要求 7 所述的终端装置，其特征在于，所述选择模块还用于：对于指定的一个或多个所述中间节点，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若该中间节点被再次选择，则此时从该中间节点的所有兄弟节点中重新进行选择。

9. 根据权利要求 7 所述的终端装置，其特征在于，所述选择模块还用于：在确定所述待存数据的所有数据副本的目标存储位置的过程中，若对当前数据副本确定出的目标存储位置与另一数据副本的已确定出的目标存储位置相同，则对当前数据副本重新确定目标存储位置，使所述待存数据的所有数据副本的目标存储位置各不相同。

10. 根据权利要求 7 所述的终端装置，其特征在于，所述映射方式包括：将所述原像调整之后使用哈希算法计算出唯一的映射值。

11. 根据权利要求 7 所述的终端装置，其特征在于，所述映射方式包括：

将所述原像调整之后使用哈希算法计算出哈希值；

将所述哈希值与预设权重值相乘得到所述映射值；

其中所述预设权重值是与所述原像中的子节点连通的所有叶节点的权重值之和，所述叶节点的权重值与该叶节点上的存储介质设备的存储容量正相关。

12. 一种服务器装置，设置在具有树形结构的无中心节点的分布式数据存储系统中的服务器内，用于将待存数据的多个数据副本保存到所述分布式数据存储系统中，所述树形结构包括叶节点和中间节点，存储介质设备位于所述叶节点上，其特征在于，该服务器装置包括：

选择模块，用于根据所述树形结构，对于待存数据的每个数据副本，分别从所述树形结构中的预选中间节点起，按预设的子节点选择方式逐级选择子节点直至被选子节点为叶节点，然后将该叶节点确定

为该数据副本的目标存储位置；其中所述子节点选择方式为：以所述待存数据的标识和子节点的标识的组合为原像，按预设的映射方式得出关联于该子节点的映射值，再根据关联于同一所述中间节点的各个子节点的多个映射值之间的大小关系选择该中间节点的一个子节点；并且对于每个中间节点，在每次选择其子节点之前重新设定所述映射关系；

保存模块，用于在所述选择模块确定了所述待存数据的所有数据副本的目标存储位置之后，将所述待存数据的所有数据副本保存到各数据副本的目标存储位置。

13. 根据权利要求 12 所述的服务器装置，其特征在于，所述选择模块还用于：对于指定的一个或多个所述中间节点，在确定所述待存数据的所有数据副本的目标存储位置的过程中，若该中间节点被再次选择，则此时从该中间节点的所有兄弟节点中重新进行选择。

14. 根据权利要求 12 所述的服务器装置，其特征在于，所述选择模块还用于：在确定所述待存数据的所有数据副本的目标存储位置的过程中，若对当前数据副本确定出的目标存储位置与另一数据副本的已确定出的目标存储位置相同，则对当前数据副本重新确定目标存储位置，使所述待存数据的所有数据副本的目标存储位置各不相同。

15. 根据权利要求 12 所述的服务器装置，其特征在于，所述映射方式包括：将所述原像调整之后使用哈希算法计算出唯一的映射值。

16. 根据权利要求 12 所述的服务器装置，其特征在于，所述映射方式包括：

将所述原像调整之后使用哈希算法计算出哈希值；

将所述哈希值与预设权重值相乘得到所述映射值；

其中所述预设权重值是与所述原像中的子节点连通的所有叶节点的权重值之和，所述叶节点的权重值与该叶节点上的存储介质设备的

存储容量正相关。

17. 根据权利要求 12 至 16 中任一项所述的服务器装置，其特征在于，

还包括确认模块，用于在所述分布式数据存储系统中的节点发生改变的情况下触发当前所述分布式数据存储系统中保存的多项数据的各自的一个预选副本所在的服务器中的所述选择模块；

所述选择模块还用于在被所述确认模块触发的情况下将所述预选副本以及与该副本相同的其他副本作为所述待存数据，确定该待存数据中的各个数据副本的新的目标存储位置，然后将该各个数据副本中原存储位置不为该新的目标存储位置的数据副本迁移到该新的目标存储位置。

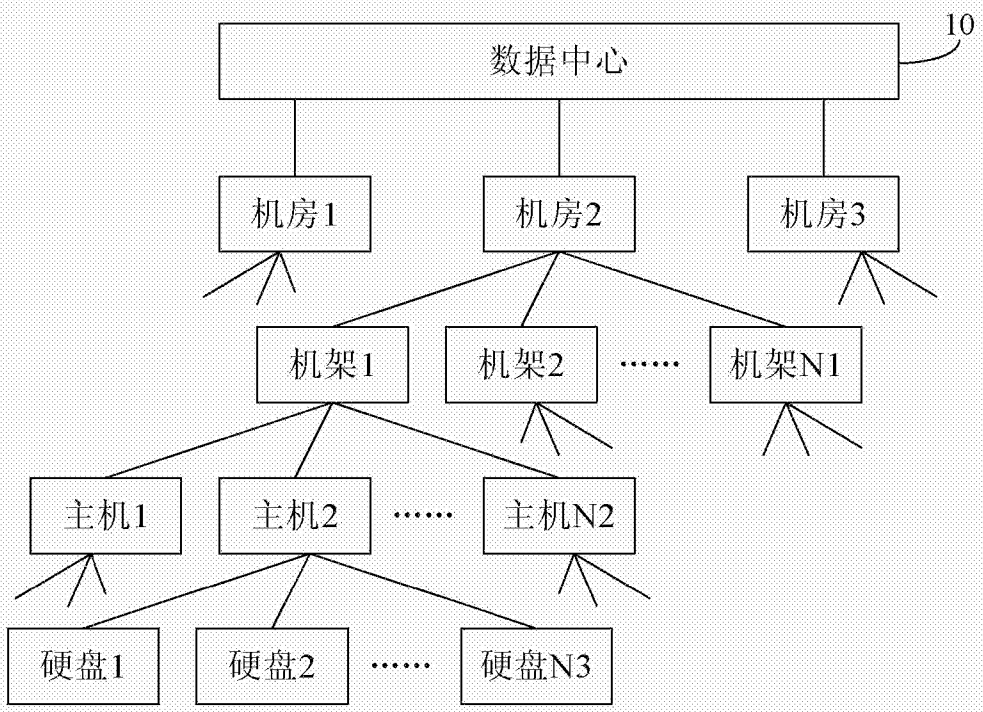

数据中心
10
机房1
机房2
机房3
机架1
机架2
……
机架N1
主机1
主机2
……
主机N2
硬盘1
硬盘2
……
硬盘N3

图1

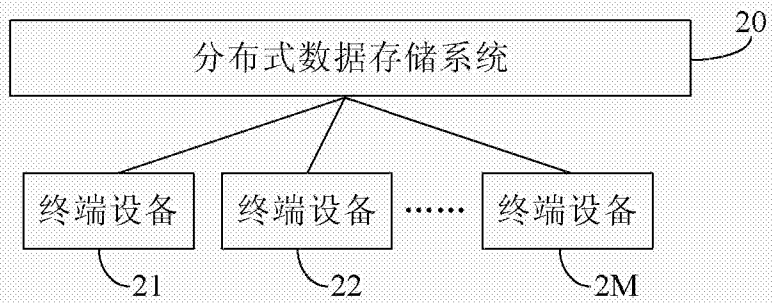

分布式数据存储系统
20
终端设备
终端设备
……
终端设备
21
22
2M

图2

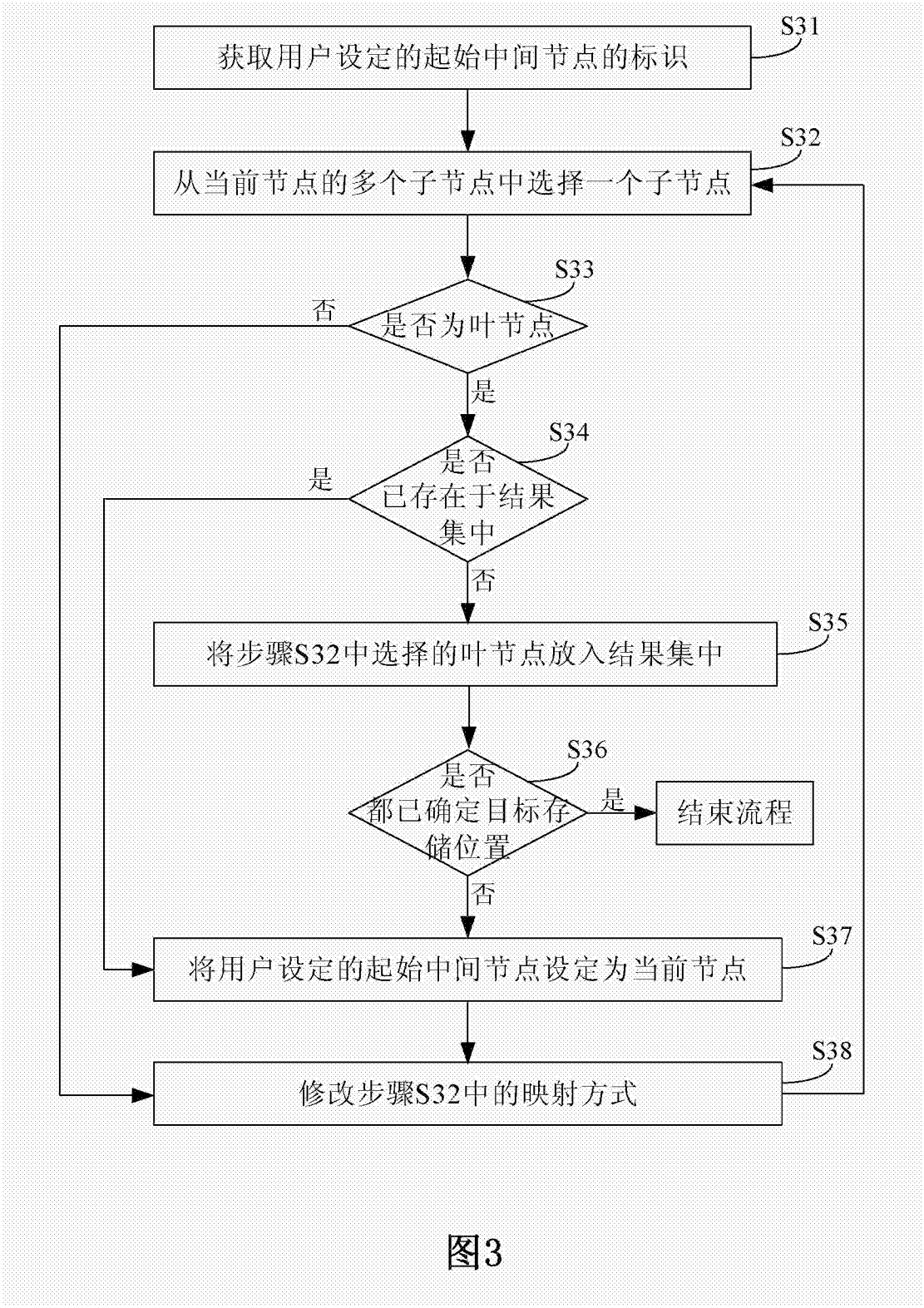

获取用户设定的起始中间节点的标识
S31
从当前节点的多个子节点中选择一个子节点
S32
是否为叶节点
S33
否
是
是否已存在于结果集中
S34
是
否
将步骤S32中选择的叶节点放入结果集中
S35
是否都已确定目标存储位置
S36
是
结束流程
否
将用户设定的起始中间节点设定为当前节点
S37
修改步骤S32中的映射方式
S38

图3

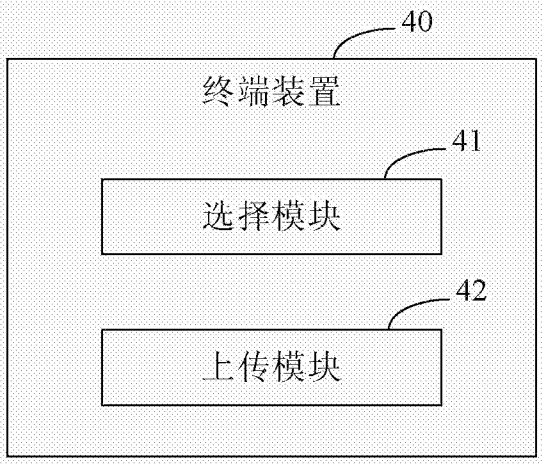


图4

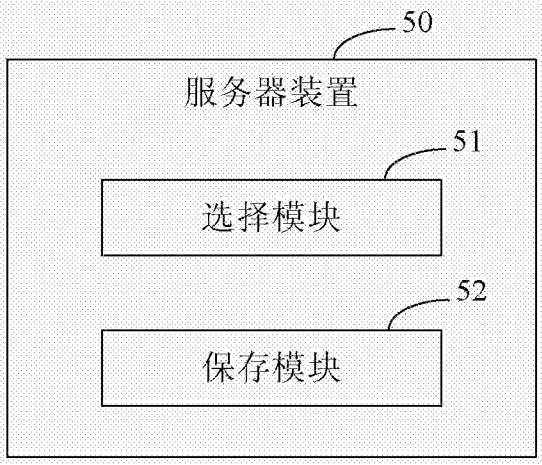


图5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2014/087621

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, H04W

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)

CNPAT, CNKI, WPI, EPOOC: distributed, storage, distributed storage, node, map+, hash, weight

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 103634401 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.), 12 March 2014 (12.03.2014), claims 1-17	1-17
A	CN 102571991 A (UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA), 11 July 2012 (11.07.2012), description, paragraphs [0001] and [0049]	1-17
A	CN 102916811 A (INSTITUTE OF INFORMATION ENGINEERING，CHINESE ACADEMY OF SCIENCES), 06 February 2013 (06.02.2013), the whole document	1-17
A	US 2010146128 A1 (NATIONAL ICT AUSTRALIA LIMITED), 10 June 2010 (10.06.2010), the whole document	1-17

☐ 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
08 December 2014 (08.12.2014)	**30 December 2014 (30.12.2014)**
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 **LIU, Man** Telephone No.: (86-10) **62414025**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2014/087621

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103634401 A	12 March 2014	None	
CN 102571991 A	11 July 2012	None	
CN 102916811 A	06 February 2013	None	
US 2010146128 A1	10 June 2010	KR 20090075847 A	09 July 2009
		AU 2007304834 A1	10 April 2008
		EP 2080110 A1	22 July 2009
		WO 2008040092 A1	10 April 2008
		CN 101523372 A	02 September 2009

国际检索报告

国际申请号
PCT/CN2014/087621

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, H04W

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））

CNPAT,CNKI,WPI,EPOOC:分布式，存储，节点，映射，散列，哈希，权值，权重，加权，distributed storage, node, map+, hash, weight

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 103634401 A (北京京东尚科信息技术有限公司 等) 2014年 3月 12日 (2014-03-12) 权利要求1-17	1-17
A	CN 102571991 A (电子科技大学) 2012年 7月 11日 (2012-07-11) 说明书第[0001], [0049]段	1-17
A	CN 102916811 A (中国科学院信息工程研究所) 2013年 2月 06日 (2013-02-06) 全文	1-17
A	US 2010146128 A1 (NATIONAL ICT AUSTRALIA LIMITED) 2010年 6月 10日 (2010-06-10) 全文	1-17

☐ 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
2014年 12月 08日	2014年 12月 30日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国	刘曼
传真号 (86-10)62019451	电话号码 (86-10)62414025

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2014/087621

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	103634401	A	2014年 3月 12日	无			
CN	102571991	A	2012年 7月 11日	无			
CN	102916811	A	2013年 2月 06日	无			
US	2010146128	A1	2010年 6月 10日	KR	20090075847	A	2009年 7月 09日
				AU	2007304834	A1	2008年 4月 10日
				EP	2080110	A1	2009年 7月 22日
				WO	2008040092	A1	2008年 4月 10日
				CN	101523372	A	2009年 9月 02日