(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
2017 年 6 月 22 日 (22.06.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/101642 A1

(51) 国际专利分类号:
H04L 29/08 (2006.01) *G06F 17/30* (2006.01)
H04L 12/24 (2006.01)

(21) 国际申请号: PCT/CN2016/106804

(22) 国际申请日: 2016 年 11 月 22 日 (22.11.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510954639.8 2015 年 12 月 17 日 (17.12.2015) 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: **许瑞亮 (XU, Ruiliang)**; 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)。

(74) 代理人: **北京市浩天知识产权代理事务所(普通合伙) (HYLANDS LAW FIRM)**; 中国北京市朝阳区朝阳门外大街 18 号丰联广场 A 座 15 层 1511, Beijing 100020 (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, DJ, 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, 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)。

[见续页]

(54) **Title**: METHOD AND APPARATUS FOR UPGRADING DATA NODE OF DISTRIBUTED SYSTEM

(54) **发明名称** : 分布式系统的数据节点升级方法及装置

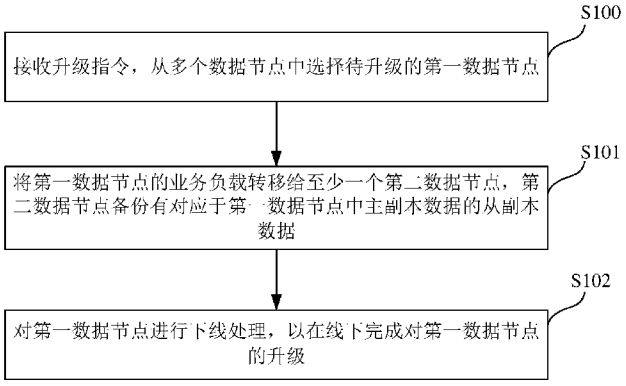


图 1

S100 RECEIVING AN UPGRADE INSTRUCTION, AND SELECTING, FROM A PLURALITY OF DATA NODES, A FIRST DATA NODE TO BE UPGRADED

S101 TRANSFERRING A SERVICE LOAD OF THE FIRST DATA NODE TO AT LEAST ONE SECOND DATA NODE, AND THE SECOND DATA NODE BACKING UP SECONDARY COPY DATA CORRESPONDING TO PRIMARY COPY DATA IN THE FIRST DATA NODE

S102 PERFORMING OFFLINE PROCESSING ON THE FIRST DATA NODE, SO AS TO COMPLETE THE UPGRADE OF THE FIRST DATA NODE OFFLINE

(57) **Abstract**: Disclosed are a method and apparatus for upgrading a data node of a distributed system. In a distributed system, with regard to one piece of data, a plurality of data nodes are respectively used for backing up primary copy data and multiple secondary copy data of this piece of data. The method comprises: receiving an upgrade instruction, and selecting, from a plurality of data nodes, a first data node to be upgraded; transferring a service load of the first data node to at least one second data node, and the second data node backing up secondary copy data corresponding to primary copy data in the first data node; and performing offline processing on the first data node, so as to complete the upgrade of the first data node offline.

(57) **摘要**: 本发明公开了一种分布式系统的数据节点升级方法及装置。在分布式系统中，针对一份数据，由多个数据节点分别用于备份该份数据的主副本数据和多个从副本数据，该方法包括：接收升级指令，从多个数据节点中选择待升级的第一数据节点；将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据；对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级。图 1

WO 2017/101642 A1

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

分布式系统的数据节点升级方法及装置

相关申请的交叉参考

本申请主张在 2015 年 12 月 17 日在中国提交的中国专利申请号 No. 201510954639.8 的优先权，其全部内容通过引用包含于此。

技术领域

本发明涉及计算机网络技术领域，具体涉及一种分布式系统的数据节点升级方法及装置。

背景技术

分布式存储就是将数据分散存储在多台独立的设备上。分布式系统为具有很多数据节点的集群系统,每个数据节点可以理解为一个数据存储服务器。在做数据存储时，将若干条数据存储到对应的数据节点中，这是分布式系统核心的概念。

在现有技术中，当需要对分布式系统的数据节点进行升级时，一般采用的方法是将数据节点关停，停止该数据节点对外的业务服务，也就是说将数据节点下线，在线下对数据节点进行升级，在数据节点升级完成后，再说将数据节点启动，也就是将数据节点上线，使升级后的数据节点对外提供业务服务，然而，在分布式系统的数据节点上一般备份有不同的主副本数据，数据节点在升级后需要重新启动，重新决策数据节点的主副本数据，该决策过程需要耗费很长时间，而且数据节点的启停，导致数据节点的业务服务延时或错误等，从而对分布式系统的业务服务造成影响。

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分布式系统的数据节点升级方法和相应的分布式系统的数据节点升级装置。

根据本发明的一个方面，提供了一种分布式系统的数据节点升级方法，在分布式系统中，针对一份数据，由多个数据节点分别用于备份该份数据的主副本数据和多个从副本数据，方法包括：

接收升级指令，从多个数据节点中选择待升级的第一数据节点；

将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据；

对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级。

根据本发明的另一方面，提供了一种分布式系统的数据节点升级装置，在分布式系统中，针对一份数据，由多个数据节点分别用于备份该份数据的主副本数据和多个从副本数据，装置包括：

选择模块，适于接收升级指令，从多个数据节点中选择待升级的第一数据节点；

转移模块，适于将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据；

升级模块，适于对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级。

根据本发明的又一个方面，提供一种计算机程序，其包括计算机可读代码，当计算机可读代码在计算设备上运行时，导致计算设备执行上述的分布式系统的数据节点升级方法。

根据本发明的再一个方面，提供一种计算机可读介质，其中存储了上述计算机程序。

根据本发明提供的方案，接收升级指令，从多个数据节点中选择待升级的第一数据节点；将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据；对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级，从而实现了分布式系统的平滑升级，避免了数据节点关停和启动对分布式系统的业务服务造成的影响。

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。

附图概述

通过阅读下文优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。附图仅用于示出优选实施方式的目的，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 示出了根据本发明一个实施例的分布式系统的数据节点升级方法的流程示意图；

图 2 示出了根据本发明另一个实施例的分布式系统的数据节点升级方法的流程示意图；

图 3 示出了根据本发明一个实施例的分布式系统的数据节点升级装置的功能框图；

图 4 示出了根据本发明另一个实施例的分布式系统的数据节点升级装置的功能框图；

图 5 示意性地示出了用于执行根据本发明实施例的分布式系统的数据节点升级方法的计算设备的框图；以及

图 6 示意性地示出了用于保持或者携带实现根据本发明实施例的分布式系统的数据节点升级方法的程序代码的存储单元。

本发明的较佳实施方式

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。

在分布式系统中，针对一份数据，由多个数据节点分别用于备份该份数据的主副本数据和多个从副本数据，也就是说将该份数据的主副本数据和多

个从副本数据分别存储于多个不同的数据节点上。针对不同的数据，在分布式系统的每个数据节点上既可能存储有主副本数据也可能存储有从副本数据，一般情况下，由数据节点的主副本数据提供业务服务。

图 1 示出了根据本发明一个实施例的分布式系统的数据节点升级方法的流程示意图。如图 1 所示，该方法包括以下步骤：

步骤 S100，接收升级指令，从多个数据节点中选择待升级的第一数据节点。

具体地，在接收到升级指令后，开始对分布式系统的数据节点进行升级，为了实现分布式系统的平滑升级，从分布式系统的多个数据节点中选择待升级的数据节点，具体地，可以从多个数据节点中随机选择待升级的数据节点，也可以根据多个数据节点的负载情况选择要进行升级的数据节点，作为待升级的第一数据节点，这里对如何选择待升级的第一数据节点选择不做具体限定。

步骤 S101，将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据。

在本发明实施例中，将第一数据节点的业务负载转移给至少一个第二数据节点，指由至少一个第二数据节点代替第一数据节点提供业务访问服务，将对第一数据节点的主副本数据的业务访问转移给至少一个第二数据节点的从副本数据。

在分布式系统中，第一数据节点可能存储有不同数据的主副本数据，因此，在进行业务负载转移时，需要将第一数据节点存储的所有主副本数据所提供的业务访问都转移给主副本数据的从副本数据。

针对一份数据，会有多个从副本数据，且多个从副本数据分别存储于不同的数据节点，因此，需要查找到备份有对应于第一数据节点的所有主副本数据的从副本数据的数据节点，从中选择承担第一数据节点的业务负载的至少一个第二数据节点。例如，针对一份数据，可以由 3 个数据节点用于备份该份数据的主副本数据和 2 个从副本数据，若第一数据节点备份有 3 份数据的主副本数据，则这 3 份数据的从副本数据至少备份于两个数据节点上，在进行业务负载转移时，一般是转移至一个从副本数据，因此，需要从至少两

个数据节点中选择承担第一数据节点的业务负载的数据节点，作为至少一个第二数据节点。

这里将第一数据节点的业务负载转移给至少一个第二数据节点，能够实现继续提供第一数据节点相应的业务服务，而不会对分布式系统的业务服务造成影响，从而避免了数据节点启停造成业务服务超时。

步骤 S102，对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级。

在利用步骤 S101 将第一数据节点的业务负载转移给至少一个第二数据节点后，第一数据节点仅备份有数据的从副本数据，在上文提到了，业务服务一般是由主副本数据提供，也就说，第一数据节点与分布式系统的业务脱离，可以对第一数据节点进行下线处理，使第一数据节点下线，并在线下对第一数据节点进行升级。

根据本发明上述实施例提供的方法，接收升级指令，从多个数据节点中选择待升级的第一数据节点；将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据；对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级，从而实现了分布式系统的平滑升级，避免了数据节点关停和启动对分布式系统的业务服务造成的影响。

图 2 示出了根据本发明另一个实施例的分布式系统的数据节点升级方法的流程示意图。如图 2 所示，该方法包括以下步骤：

步骤 S200，接收升级指令，对分布式系统的数据节点的业务负载进行分析，将业务负载压力最小的数据节点选择为待升级的第一数据节点。

在本发明实施例中，业务负载指对数据节点的主副本数据的访问量。在分布式系统中，每个数据节点对外提供的业务服务并不相同，有些数据节点的业务负载压力大，有些数据节点的业务负载压力小，在接收到升级指令后，对分布式系统的数据节点的业务负载进行分析，将业务负载压力最小的数据节点选择为待升级的第一数据节点。本发明是对分布式系统的全部数据节点进行升级，因此，待升级的第一数据节点是未升级的数据节点中的业务负载压力最小的数据节点，选择业务负载压力最小的数据节点，可以避免因数据

节点升级，将业务服务转移给其他数据节点而给分布式系统造成影响。

步骤 S201，根据第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中主副本数据的从副本数据的数据节点。

在分布式系统中，一份数据，由多个数据节点分别用于备份该份数据的主副本数据和多个从副本数据，并且为主副本数据和从副本数据设置了相同的标识，以便于根据标识查找主副本数据或从副本数据，针对不同的数据，设置了不同的标识，从而可以准确查找到数据的主副本数据或从副本数据，节省查找时间，提高了升级效率。

在根据步骤 S200 选择出待升级的第一数据节点后，需要将第一数据节点的主副本数据的业务服务转移给其他数据节点的从副本数据，当然，并不是将第一数据节点的主副本数据的业务服务转移给任意数据节点的从副本数据，而是转移给备份有对应于第一数据节点中主副本数据的从副本数据的数据节点，此外，第一数据节点可能备份有不同数据的主副本数据，若要对该第一数据节点进行升级，就需要将第一数据节点中所有主副本数据的业务服务转移给其他数据节点的从副本数据。因此，需要根据第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中所有主副本数据的从副本数据的数据节点。

步骤 S202，对备份有对应于第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第二数据节点。

利用步骤 S201 查找到备份有对应于第一数据节点中主副本数据的从副本数据的数据节点，然而，并不是将第一数据节点的业务负载转移给查找到的全部数据节点，例如，针对一份数据，有一个数据节点备份有主副本数据和两个数据节点备份有从副本数据，在将第一数据节点的该份数据的业务负载转移时，仅能转移给备份有从副本数据的两个数据节点中的一个。

在本发明实施例中，通过对备份有对应于第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，分析数据节点的业务负载压力大小，根据负载均衡策略从其中选择出至少一个第二数据节点，其中，负载

均衡策略指分布式系统的数据节点的业务负载压力均衡，而不会出现有些数据节点的业务负载压力很大，而有些数据节点的业务负载压力很小的情况，这里不仅考虑到备份有对应于第一数据节点中主副本数据的从副本数据的数据节点的业务负载，还要考虑到业务负载转移后，数据节点的业务负载情况。本步骤中选择的至少一个第二数据节点是业务负载压力较小的数据节点。这里根据负载均衡策略选择至少一个第二数据节点能够克服随意转移业务负载造成的加重业务负载压力大的数据节点的业务负担的缺陷。

由于第一数据节点可能备份有多个主副本数据，因此根据负载均衡策略选择的数据节点也可能为多个。这里将第一数据节点的业务负载转移给至少一个第二数据节点，能够实现继续提供第一数据节点相应的业务服务，而不会对分布式系统的业务服务造成影响，从而避免了数据节点启停造成业务服务超时。

步骤 S203，将至少一个第二数据节点的从副本数据转换为主副本数据，将第一数据节点的对应主副本数据转换为从副本数据。

在根据步骤 S202 选择出至少一个第二数据节点后，需要将第一数据节点的业务负载转移给至少一个第二数据节点，在分布式系统中，一般由主副本数据提供业务负载，而至少一个第二数据节点是备份有第一数据节点的主副本数据对应的从副本数据的数据节点，因此，需要将至少一个第二数据节点的从副本数据转换为主副本数据，将第一数据节点的对应主副本数据转换为从副本数据，从而将业务负载转移给至少一个第二数据节点。

步骤 S204，对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级。

利用步骤 S203 将第一数据节点的业务负载转移给至少一个第二数据节点后，第一数据节点将不再提供业务服务，此时，即可将第一数据节点下线，在线下对第一数据节点进行升级。

虽然上文提到了一般由数据节点的主副本数据提供业务服务，但也可能存在访问数据节点的从副本数据的情况，因此，方法还包括：

步骤 S205，将针对第一数据节点中从副本数据的访问请求重新定位到其他数据节点中与从副本数据对应的主副本数据/从副本数据。

具体指，在用户对第一数据节点中从副本数据发起访问请求后，将针对第一数据节点中从副本数据的访问请求重新定位到其他数据节点中与从副本数据对应的主副本数据/从副本数据，使定位后的其他数据节点中与从副本数据对应的主副本数据/从副本数据提供业务服务，其中，第一数据节点中从副本数据可以是业务负载转移前，第一数据节点所备份的从副本数据，也可以是业务负载转移后得到的从副本数据。

步骤 S206，在第一数据节点完成升级后，对升级后的第一数据节点进行上线处理。

在线下完成对第一数据节点升级后，需要将第一数据节点上线，以使第一数据节点继续提供业务服务。

在升级后的第一数据节点上线后，需要将至少一个第三数据节点的业务负载转移给升级后的第一数据节点，以使升级后的第一数据节点恢复提供业务服务，其中，升级后的第一数据节点备份有对应于第三数据节点中主副本数据的从副本数据。

在本发明实施例中，将至少一个第三数据节点的业务负载转移给第一数据节点，指由第一数据节点代替至少一个第三数据节点提供业务服务，将至少一个第三数据节点的主副本数据的业务服务转移给第一数据节点的从副本数据。具体可以通过以下方法步骤实现：

步骤 S207，根据升级后的第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有对应于升级后的第一数据节点中从副本数据的主副本数据的数据节点。

在将第一数据节点上线后，需要将其他数据节点的业务负载转移给第一数据节点，第一数据节点才可以提供业务服务。

在将数据备份于数据节点上时，为主副本数据和从副本数据设置了相同的标识，因此，可以根据第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有第一数据节点的从副本数据对应的主副本数据的数据节点。

步骤 S208，对备份有对应于升级后的第一数据节点中从副本数据的主副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至

少一个第三数据节点。

在根据步骤 S207 查找备份有对应于升级后的第一数据节点中从副本数据的主副本数据的数据节点后，对这些数据节点的业务负载进行分析，确定这些数据节点的业务负载情况，根据负载均衡策略从中选择业务负载压力较大的数据节点作为第三数据节点，从而减轻第三数据节点的业务负载，使得分布式系统的业务负载压力均衡。

步骤 S209，将至少一个第三数据节点的主副本数据转换为从副本数据，将升级后的第一数据节点的对应从副本数据转换为主副本数据。

在根据步骤 S208 选择出至少一个第三数据节点后，需要将至少一个第三数据节点的业务负载转移给第一数据节点，在分布式系统中，一般由主副本数据提供业务负载，而至少一个第三数据节点是备份有对应于第一数据节点中从副本数据的主副本数据的数据节点，因此，需要将至少一个第三数据节点的主副本数据转换为从副本数据，将升级后的第一数据节点的对应从副本数据转换为主副本数据，从而将业务负载转移给第一数据节点，以使已升级数据节点恢复提供业务服务。

根据本发明上述实施例提供的方法，在对分布式系统的数据节点的业务负载进行分析后，将业务负载压力最小的数据节点选择为待升级的第一数据节点，可以避免对分布式系统的业务服务造成影响，根据第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中主副本数据的从副本数据的数据节点，可以准确查找到数据的主副本数据或从副本数据，节省查找时间，提高了升级效率，对备份有对应于第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第二数据节点，能够克服随意转移业务负载造成的加重业务负载压力大的数据节点的业务负担的缺陷，能够实现继续提供第一数据节点相应的业务服务，而不会对分布式系统的业务服务造成影响，从而避免了数据节点启停造成业务服务超时，在第一数据节点完成升级后，对第一数据节点进行上线处理，根据第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中从副本数据的主副本数据的数据节点，对备份有对应于第一数据节点中从副本数据的主副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一

个第三数据节点，从而减轻第三数据节点的业务负载，实现了分布式系统的平滑升级。

图 3 示出了根据本发明一个实施例的分布式系统的数据节点升级装置的功能框图。如图 3 所示，该装置包括：选择模块 300、转移模块 310、升级模块 320。

选择模块 300，适于接收升级指令，从多个数据节点中选择待升级的第一数据节点。

转移模块 310，适于将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据。

升级模块 320，适于对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级。

根据本发明上述实施例提供的装置，接收升级指令，从多个数据节点中选择待升级的第一数据节点；将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据；对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级，从而实现了分布式系统的平滑升级，避免了数据节点关停和启动对分布式系统的业务服务造成的影响。

图 4 示出了根据本发明另一个实施例的分布式系统的数据节点升级装置的功能框图。如图 4 所示，该装置包括：选择模块 400、转移模块 410、升级模块 420。

选择模块 400，适于接收升级指令，从多个数据节点中选择待升级的第一数据节点。

转移模块 410，适于将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据。

升级模块 420，适于对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级。

可选地，装置还包括：处理模块 430，适于在第一数据节点完成升级后，

对升级后的第一数据节点进行上线处理；

转移模块 410 进一步适于：将至少一个第三数据节点的业务负载转移给升级后的第一数据节点，以使升级后的第一数据节点恢复提供业务服务，其中，升级后的第一数据节点备份有对应于第三数据节点中主副本数据的从副本数据。

可选地，转移模块 410 进一步适于：将至少一个第二数据节点的从副本数据转换为主副本数据，将第一数据节点的对应主副本数据转换为从副本数据。

可选地，装置还包括：查找模块 440 和分析模块 450。

查找模块 440，适于根据第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中主副本数据对应的从副本数据的数据节点。

分析模块 450，适于对备份有对应于第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第二数据节点。

可选地，转移模块 410 进一步适于：将至少一个第三数据节点的主副本数据转换为从副本数据，将升级后的第一数据节点的对应从副本数据转换为主副本数据。

可选地，查找模块 440 进一步适于：根据升级后的第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有对应于升级后的第一数据节点中从副本数据的主副本数据的数据节点；

分析模块 450 进一步适于：对备份有对应于升级后的第一数据节点中从副本数据的主副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第三数据节点。

可选地，选择模块 400 进一步适于：对分布式系统的数据节点的业务负载进行分析，将业务负载量最小的数据节点选择为待升级的第一数据节点。

可选地，装置还包括：定位模块 460，适于将针对第一数据节点中从副本数据的访问请求重新定位到其他数据节点中与从副本数据对应的主副本数据/从副本数据。

根据本发明上述实施例提供的装置，在对分布式系统的数据节点的业务负载进行分析后，将业务负载压力最小的数据节点选择为待升级的第一数据节点，可以避免对分布式系统的业务服务造成影响，根据第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中主副本数据的从副本数据的数据节点，可以准确查找到数据的主副本数据或从副本数据，节省查找时间，提高了升级效率，对备份有对应于第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第二数据节点，能够克服随意转移业务负载造成的加重业务负载压力大的数据节点的业务负担的缺陷，能够实现继续提供第一数据节点相应的业务服务,而不会对分布式系统的业务服务造成影响，从而避免了数据节点启停造成业务服务超时，在第一数据节点完成升级后，对第一数据节点进行上线处理，根据第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中从副本数据的主副本数据的数据节点，对备份有对应于第一数据节点中从副本数据的主副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第三数据节点，从而减轻第三数据节点的业务负载，实现了分布式系统的平滑升级。

在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述，构造这类系统所要求的结构是显而易见的。此外，本发明也不针对任何特定编程语言。应当明白,可以利用各种编程语言实现在此描述的本发明的内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

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。

类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映

的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书（包括伴随的权利要求、摘要和附图）中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书（包括伴随的权利要求、摘要和附图）中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

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分布式系统的数据节点升级设备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网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

例如，图 5 示出了可以实现根据本发明的获取应用信息的方法的计算设备。该计算设备传统上包括处理器 510 和以存储设备 520 形式的计算机程序产品或者计算机可读介质。存储设备 520 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

储设备 520 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 531 的存储空间 530。例如，存储程序代码的存储空间 530 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 531。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 6 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 5 的计算设备中的存储设备 520 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 531'，即可以由诸如 510 之类的处理器读取的代码，当这些代码由计算设备运行时，导致该计算设备执行上面所描述的方法中的各个步骤。

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是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。

权 利 要 求 书

1、一种分布式系统的数据节点升级方法，在所述分布式系统中，针对一份数据，由多个数据节点分别用于备份该份数据的主副本数据和多个从副本数据，所述方法包括：

接收升级指令，从多个数据节点中选择待升级的第一数据节点；

将所述第一数据节点的业务负载转移给至少一个第二数据节点，所述第二数据节点备份有对应于所述第一数据节点中主副本数据的从副本数据；

对所述第一数据节点进行下线处理，以在线下完成对所述第一数据节点的升级。

2、根据权利要求 1 所述的方法，其中，所述方法还包括：

在所述第一数据节点完成升级后，对升级后的所述第一数据节点进行上线处理，将至少一个第三数据节点的业务负载转移给升级后的所述第一数据节点，以使升级后的所述第一数据节点恢复提供业务服务，其中，升级后的所述第一数据节点备份有对应于所述第三数据节点中主副本数据的从副本数据。

3、根据权利要求 2 所述的方法，其中，所述将第一数据节点的业务负载转移给至少一个第二数据节点进一步包括：

将所述至少一个第二数据节点的从副本数据转换为主副本数据，将所述第一数据节点的对应主副本数据转换为从副本数据。

4、根据权利要求 2 或 3 所述的方法，其中，在将所述第一数据节点的业务负载转移给至少一个第二数据节点之前，所述方法还包括：

根据所述第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于所述第一数据节点中主副本数据的从副本数据的数据节点；

对所述备份有对应于所述第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第二数据节点。

5、根据权利要求 2-4 任一项所述的方法，其中，所述将至少一个第三数

据节点的业务负载转移给升级后的所述第一数据节点进一步包括：

将所述至少一个第三数据节点的主副本数据转换为从副本数据，将升级后的所述第一数据节点的对应从副本数据转换为主副本数据。

6、根据权利要求 4 所述的方法，其中，在将至少一个第三数据节点的业务负载转移给升级后的所述第一数据节点之前，所述方法还包括：

根据升级后的所述第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有对应于升级后的所述第一数据节点中从副本数据的主副本数据的数据节点；

对所述备份有对应于升级后的所述第一数据节点中从副本数据的主副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第三数据节点。

7、根据权利要求 1-6 任一项所述的方法，其中，所述从多个数据节点中选择待升级的第一数据节点进一步包括：

对所述分布式系统的数据节点的业务负载进行分析，将业务负载量最小的数据节点选择为待升级的第一数据节点。

8、根据权利要求 1-7 任一项所述的方法，其中，在升级后的所述第一数据节点上线之前，所述方法还包括：

将针对所述第一数据节点中从副本数据的访问请求重新定位到其他数据节点中与所述从副本数据对应的主副本数据/从副本数据。

9、一种分布式系统的数据节点升级装置，在所述分布式系统中，针对一份数据，由多个数据节点分别用于备份该份数据的主副本数据和多个从副本数据，所述装置包括：

选择模块，适于接收升级指令，从多个数据节点中选择待升级的第一数据节点；

转移模块，适于将所述第一数据节点的业务负载转移给至少一个第二数据节点，所述第二数据节点备份有对应于所述第一数据节点中主副本数据的从副本数据；

升级模块，适于对所述第一数据节点进行下线处理，以在线下完成对所

述第一数据节点的升级。

10、根据权利要求 9 所述的装置，其中，所述装置还包括：

处理模块，适于在所述第一数据节点完成升级后，对升级后的所述第一数据节点进行上线处理；

所述转移模块进一步适于：将至少一个第三数据节点的业务负载转移给升级后的所述第一数据节点，以使升级后的所述第一数据节点恢复提供业务服务，其中，升级后的所述第一数据节点备份有对应于所述第三数据节点中主副本数据的从副本数据。

11、根据权利要求 10 所述的装置，其中，所述转移模块进一步适于：将所述至少一个第二数据节点的从副本数据转换为主副本数据，将所述第一数据节点的对应主副本数据转换为从副本数据。

12、根据权利要求 10 或 11 所述的装置，其中，所述装置还包括：

查找模块，适于根据所述第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于所述第一数据节点中主副本数据对应的从副本数据的数据节点；

分析模块，适于对所述备份有对应于所述第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第二数据节点。

13、根据权利要求 10-12 任一项所述的装置，其中，所述转移模块进一步适于：将所述至少一个第三数据节点的主副本数据转换为从副本数据，将升级后的所述第一数据节点的对应从副本数据转换为主副本数据。

14、根据权利要求 12 所述的装置，其中，所述查找模块进一步适于：根据升级后的所述第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有对应于升级后的所述第一数据节点中从副本数据的主副本数据的数据节点；

所述分析模块进一步适于：对所述备份有对应于升级后的所述第一数据节点中从副本数据的主副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第三数据节点。

15、根据权利要求 9-14 任一项所述的装置，其中，所述选择模块进一步适于：

对所述分布式系统的数据节点的业务负载进行分析，将业务负载量最小的数据节点选择为待升级的第一数据节点。

16、根据权利要求 9-15 任一项所述的装置，其中，所述装置还包括：定位模块，适于将针对所述第一数据节点中从副本数据的访问请求重新定位到其他数据节点中与所述从副本数据对应的主副本数据/从副本数据。

17、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 1-8 中的任一项所述的分布式系统的数据节点升级方法。

18、一种计算机可读介质，其中存储了如权利要求 17 所述的计算机程序。

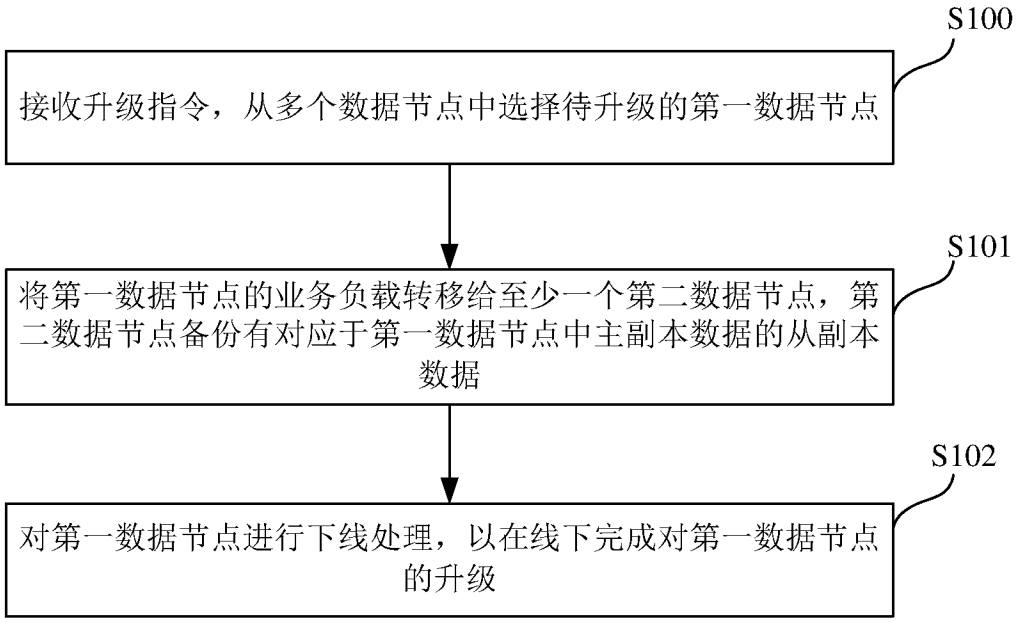

S100
接收升级指令，从多个数据节点中选择待升级的第一数据节点
S101
将第一数据节点的业务负载转移给至少一个第二数据节点，第二数据节点备份有对应于第一数据节点中主副本数据的从副本数据
S102
对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级

图 1

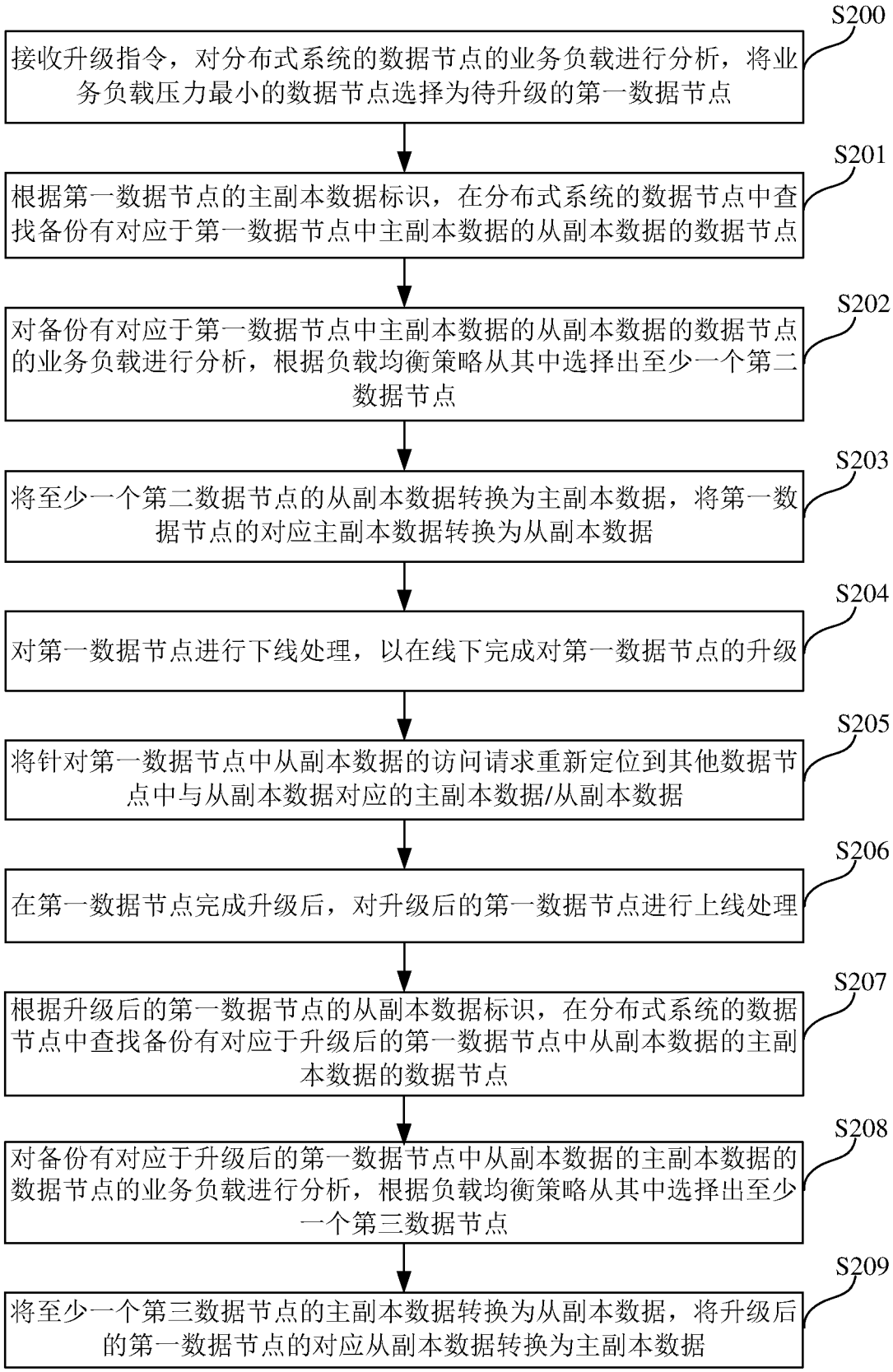

S200
接收升级指令，对分布式系统的数据节点的业务负载进行分析，将业务负载压力最小的数据节点选择为待升级的第一数据节点
S201
根据第一数据节点的主副本数据标识，在分布式系统的数据节点中查找备份有对应于第一数据节点中主副本数据的从副本数据的数据节点
S202
对备份有对应于第一数据节点中主副本数据的从副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第二数据节点
S203
将至少一个第二数据节点的从副本数据转换为主副本数据，将第一数据节点的对应主副本数据转换为从副本数据
S204
对第一数据节点进行下线处理，以在线下完成对第一数据节点的升级
S205
将针对第一数据节点中从副本数据的访问请求重新定位到其他数据节点中与从副本数据对应的主副本数据/从副本数据
S206
在第一数据节点完成升级后，对升级后的第一数据节点进行上线处理
S207
根据升级后的第一数据节点的从副本数据标识，在分布式系统的数据节点中查找备份有对应于升级后的第一数据节点中从副本数据的主副本数据的数据节点
S208
对备份有对应于升级后的第一数据节点中从副本数据的主副本数据的数据节点的业务负载进行分析，根据负载均衡策略从其中选择出至少一个第三数据节点
S209
将至少一个第三数据节点的主副本数据转换为从副本数据，将升级后的第一数据节点的对应从副本数据转换为主副本数据

图 2

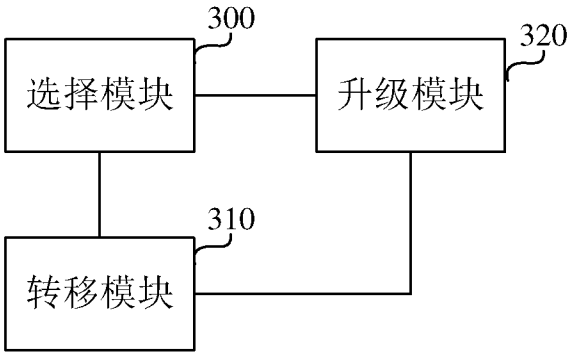

300
320
选择模块
升级模块
310
转移模块

图 3

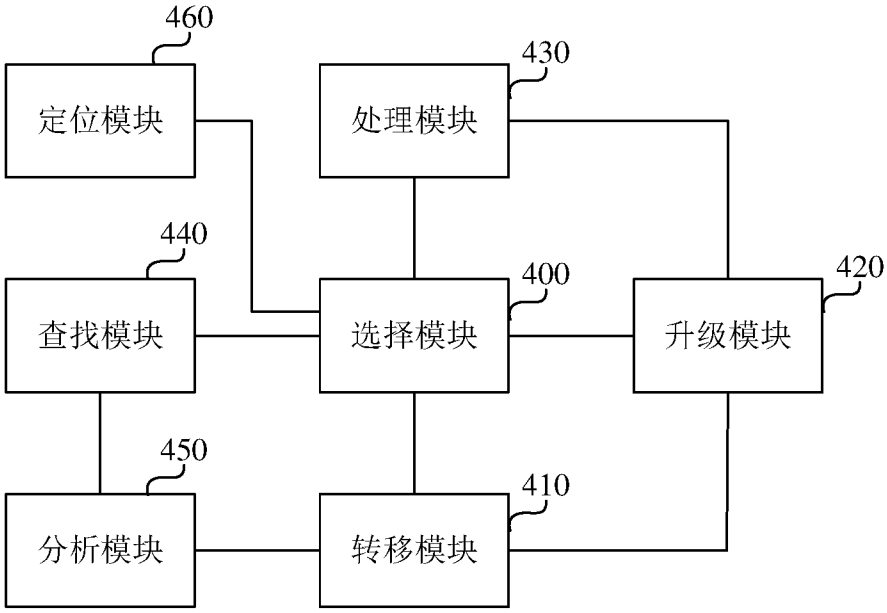

460
430
定位模块
处理模块
440
420
400
查找模块
选择模块
升级模块
450
410
分析模块
转移模块

图 4

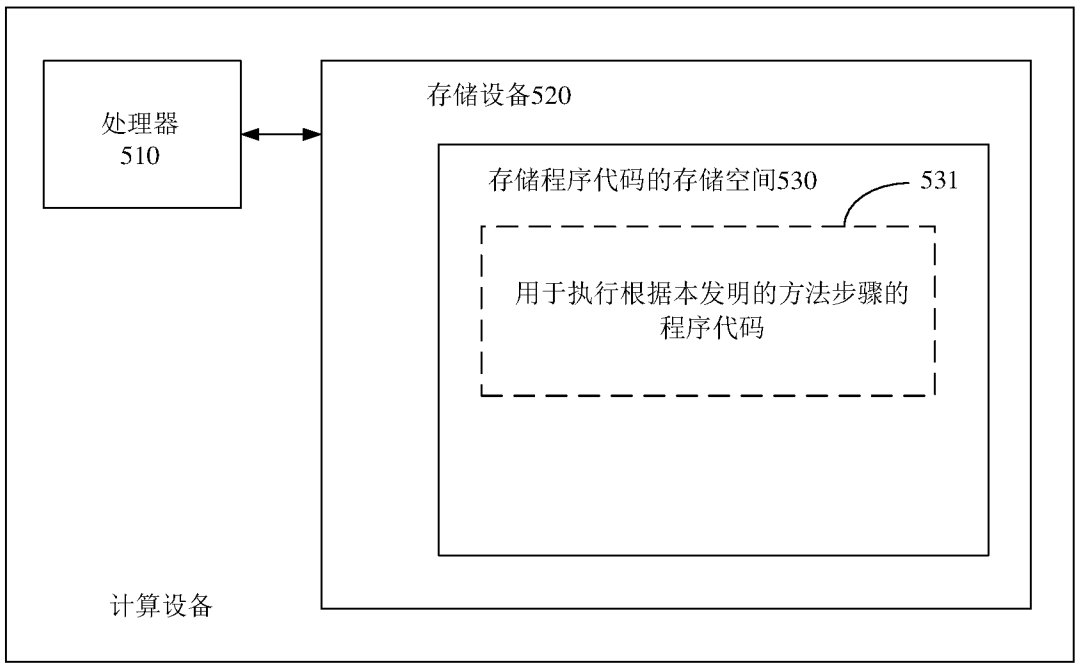

处理器
510
存储设备520
存储程序代码的存储空间530
531
用于执行根据本发明的方法步骤的
程序代码
计算设备

图 5

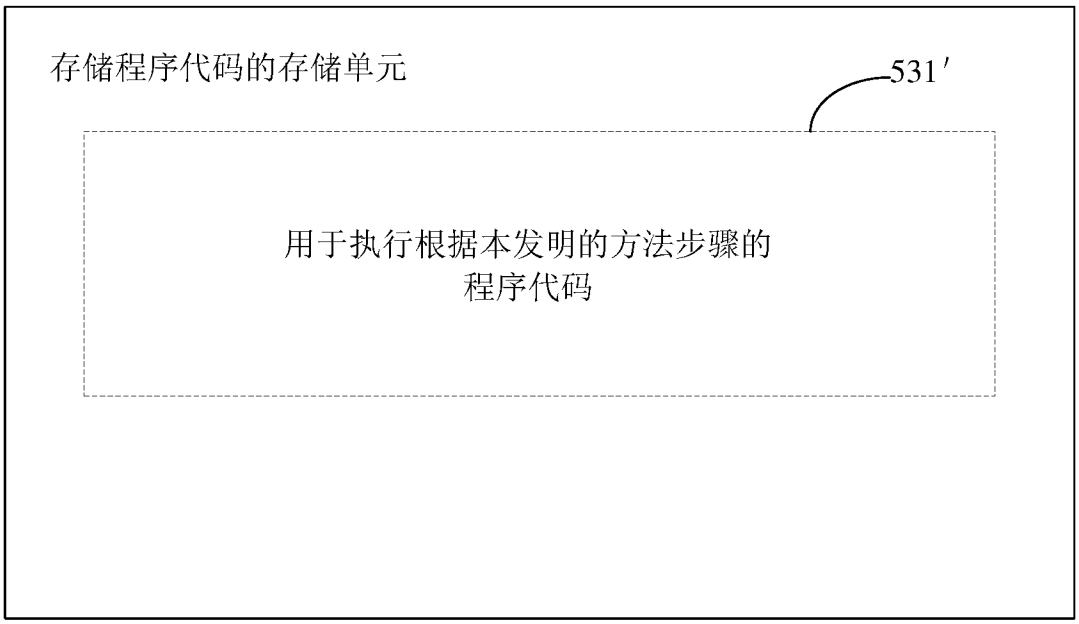

存储程序代码的存储单元
531′
用于执行根据本发明的方法步骤的
程序代码

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/106804

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08 (2006.01) i; H04L 12/24 (2006.01) n; G06F 17/30 (2006 01) n
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; G06F

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)

CNTXT; CNABS; CNKI; VEN: distribute, cache, storage, buffer, server, node, online, update, upgrade, backup

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 105610903 A (BEIJING QIHOO SCIENCE & TECHNOLOGY CO., LTD. et al.) 25 May 2016 (25.05.2016) claims 1-10, and description, paragraphs [0062]-[0088], and figures 1-4	1-18
X	CN 103970567 A (INSPUR ELECTRONIC INFORMATION INDUSTRY CO., LTD.) 06 August 2014 (06.08.2014) description, paragraphs [0015]-[0020], and figure 1	1, 7-9, 15-18
Y	CN 103970567 A INSPUR ELECTRONIC INFORMATION INDUSTRY CO., LTD.) 06 August 2014 (06.08.2014) description, paragraphs [0015]-[0020], and figure 1	2-6, 10-14
Y	CN 104954444 A (HUAWEI TECHNOLOGIES CO., LTD.) 30 September 2015 (30.09.2015) description, paragraph [0004]	2-6, 10-14
A	US 2013246498 A1 (ZUCKNOVICH STEPHEN et al.) 19 September 2013 (19.09.2013) the whole document	1-18

☐ 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
06 January 2017	08 February 2017
Name and mailing address of the ISA 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 LV, He Telephone No. (86-10) 62411400

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/106804

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105610903 A	25 May 2016	None	
CN 103970567 A	06 August 2014	None	
CN 104954444 A	30 September 2015	None	
US 2013246498 A1	19 September 2013	US 9444904 B2	13 September 2016

国际检索报告

国际申请号
PCT/CN2016/106804

A. 主题的分类

H04L 29/08(2006.01)i; H04L 12/24(2006.01)n; G06F 17/30(2006.01)n

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; G06F

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:分布，存储，缓存，服务器，节点，在线，更新，升级，备份 VEN:distribute, cache, storage, buffer, server, node, online, update, upgrade, backup

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 105610903 A (北京奇虎科技有限公司 等) 2016年 5月 25日 (2016-05-25) 权利要求1-10，说明书第【0062】-【0088】段，附图1-4	1-18
X	CN 103970567 A (浪潮电子信息产业股份有限公司) 2014年 8月 6日 (2014-08-06) 说明书第【0015】-【0020】段，附图1	1，7-9，15-18
Y	CN 103970567 A (浪潮电子信息产业股份有限公司) 2014年 8月 6日 (2014-08-06) 同上	2-6，10-14
Y	CN 104954444 A (华为技术有限公司) 2015年 9月 30日 (2015-09-30) 说明书第【0004】段	2-6，10-14
A	US 2013246498 A1 (ZUCKNOVICH STEPHEN等) 2013年 9月 19日 (2013-09-19) 全文	1-18

☐ 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
2017年 1月 6日	2017年 2月 8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	受权官员 吕鹤 电话号码 (86-10)62411400

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/106804

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 105610903 A	2016年 5月 25日	无	
CN 103970567 A	2014年 8月 6日	无	
CN 104954444 A	2015年 9月 30日	无	
US 2013246498 A1	2013年 9月 19日	US 9444904 B2	2016年 9月 13日