

(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
2014 年 4 月 24 日 (24.04.2014)



(10) 国际公布号
WO 2014/059804 A1

- (51) 国际专利分类号:
G06F 12/00 (2006.01)
- (21) 国际申请号: PCT/CN2013/079087
- (22) 国际申请日: 2013 年 7 月 9 日 (09.07.2013)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201210397350.7 2012 年 10 月 18 日 (18.10.2012) CN
- (71) 申请人: 腾讯科技(深圳)有限公司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。
- (72) 发明人: 江兴才 (JIANG, Xingcai); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。 田明 (TIAN, Ming); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。 刘里 (LIU, Li); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。 黄利华 (HUANG, Lihua); 中国广东省深圳市福田区振兴路赛格科技园 2 栋

东 403 室, Guangdong 518000 (CN)。 李钟伟 (LI, Zhongwei); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。

(74) 代理人: 北京市磐华律师事务所 (P.C. & ASSOCIATES); 中国北京市朝阳区建国门外大街 22 号赛特大厦 901-902, Beijing 100004 (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, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, 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,

[见续页]

(54) Title: METHOD AND SYSTEM FOR DATA SYNCHRONIZATION

(54) 发明名称: 数据同步方法及系统

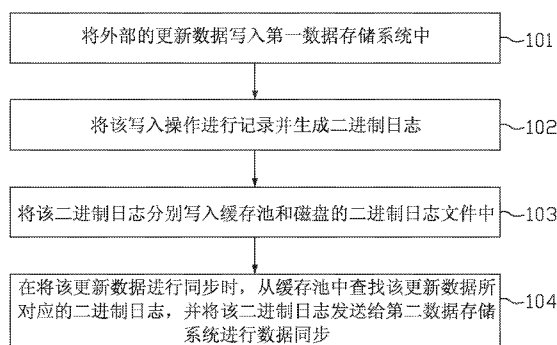


图 1 /FIG. 1

- 101 WRITING EXTERNAL UPDATED DATA INTO A FIRST DATA STORAGE SYSTEM
- 102 RECORDING THE WRITING OPERATION AND GENERATING A BINARY LOG
- 103 WRITING THE BINARY LOG INTO A BUFFER POOL AND THE BINARY LOG FILE IN THE DISK RESPECTIVELY
- 104 WHEN SYNCHRONIZING THE UPDATED DATA, SEARCHING THE BINARY LOG CORRESPONDING TO THE UPDATED DATA IN THE BUFFER POOL, AND SENDING THE BINARY LOG TO A SECOND DATA STORAGE SYSTEM FOR DATA SYNCHRONIZATION

(57) Abstract: Disclosed are a method and system for data synchronization. The data synchronization method comprises the following steps: writing external updated data into a first data storage system; recording the writing operation and generating a binary log; writing the binary log into a buffer pool and the binary log file in the disk respectively; when synchronizing the updated data, searching the binary log corresponding to the updated data in the buffer pool, and sending the binary log to a second data storage system for data synchronization. The embodiment of this invention proposes a synchronous strategy for asynchronous transferring based on the buffer pool and the binary log file. The synchronous strategy separates the data storage system from the data synchronization system. The data synchronization system is responsible for updating every copy of the data to the latest state according to the binary log. In this mode, the overall complexity, coupling, and bandwidth cost of the system significantly decrease without any reduction in system service performance.

(57) 摘要: 本发明实施例公开了一种数据同步方法及系统, 所述数据同步方法包括: 将外部的更新数据写入第一数据存储系统中; 将该写入操作进行记录并生成二进制日志; 将该二进制日志分别写入缓存池和磁盘的二进制日志文件中; 在将

该更新数据进行同步时, 从缓存池中查找该更新数据所对应的二进制日志, 并将该二进制日志发送给第二数据存储系统进行数据同步。本发明实施例提出了一种基于缓存池和二进制日志文件的异步传送到最新的状态, 在该模式下, 系统服务性能没有丝毫降低的同时, 系统的整体复杂度、耦合性、带宽成本都大幅降低。



WO 2014/059804 A1



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))。

数据同步方法及系统

相关申请的交叉引用

- 5 本申请要求于 2012 年 10 月 18 日提交中国专利局、申请号为 201210397350.7、发明名称为“数据同步方法及系统”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

- 10 本发明涉及数据存储与传输技术领域，并且更具体地，涉及一种数据同步方法及系统。

背景技术

- 15 UGC (User Generated Content, 用户原创内容) 是一种用户使用互联网的新方式，即由原来的以下载为主变成下载和上传并重。社区网络、视频分享和博客等都是 UGC 的主要应用形式。随着全球互联网业务的不断发展，UGC 业务正在日渐崛起，引起了业界的广泛关注。

- 20 为了安全运营，系统设计时会引入容灾方案，容灾方案要求系统至少有两份可用的完整数据拷贝，各个数据拷贝在部署上相互独立，并且都能提供全量的实时服务，当其中一个数据拷贝出现异常或者灾难从而导致该数据拷贝无法提供正常的服务时，可以把请求切换到其他可用的数据点，以提供不间断的实时服务，如何使各个数据拷贝之间的数据能够保持一致性，是容灾方案面临的一个难题，如果能有一种简单、高效、低成本的容灾模型，将为该领域带来意义深远的变革。

- 25 现有技术中，数据存储系统负责数据的存储，提供读写服务，同时提供数据的同步服务。用户的一个写操作到达服务进程后，服务进程首先会查询系统一共有多少个可用数据拷贝，假设有可用的数据拷贝数为 N，那么服务进程会把这个写操作复制成 N 份，分别对每一个数据拷贝发送这个写操作，从而使每个数据拷贝都能将数据更新到最新状态。

- 30 (1) 耦合性高。数据存储系统与数据同步系统有依赖关系。首先，数据的存储依赖于数据的同步是否成功，如果某个写操作在主写点成功了，但

是另外一个数据拷贝更新失败了，那么所有数据拷贝的这个写操作都认为是失败的；

5 (2) 设计复杂。两个系统同等重要，为了保证对外的正常服务，某一个系统出现异常时都会影响另一个系统的正常服务，这种设计直接导致运维成本增加；

(3) 搭建新数据拷贝困难。当需要搭建一个新数据拷贝的时候，需要导入原先的历史数据，同时还需要系统停写支持；

(4) 数据拷贝越多性能越差。当可用的数据拷贝越多时，某一个数据拷贝更新失败会导致更多的写操作被判为无效，致使系统性能下降。

10

发明内容

本发明实施例提供了一种数据同步方法及系统，以降低系统的整体复杂度、耦合性，为用户提供高可靠性、可用性的数据同步服务。所述技术方案如下：

15 一种数据同步方法，包括：

将外部的更新数据写入第一数据存储系统中；

将该写入操作进行记录并生成二进制日志；

将该二进制日志分别写入缓存池和磁盘的二进制日志文件中；

20 在将该更新数据进行同步时，从缓存池中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步。

一种数据同步系统，包括：

第一写入模块，用于将外部的更新数据写入第一数据存储系统中；

生成模块，用于将该写入操作进行记录并生成二进制日志；

25 第二写入模块，用于将该二进制日志分别写入缓存池和磁盘的二进制日志文件中；

同步模块，用于在将该更新数据进行同步时，从缓存池中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步。

30 本发明实施例还提供了一种用于数据同步的方法、一种用于数据同步的系统以及一种包括用于数据同步的可执行程序代码的非暂时性计算机程序产品。技术方案如下：

一种用于数据同步的方法，包括：

根据在第一数据存储系统中进行的数据写操作生成二进制日志；

将所述二进制日志写入存储器；以及

独立于所述生成二进制日志和所述将所述二进制日志写入存储器，

- 5 在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给第二数据存储系统，以使所述第二数据存储系统根据所述对应于最新写操作的二进制日志对所述第二数据存储系统中的数据进行同步更新。

- 10 一种用于数据同步的系统，包括第一数据存储系统以及数据同步系统，

所述第一数据存储系统用于根据在所述第一数据存储系统中进行的数据写操作生成二进制日志以及将所述二进制日志写入存储器；以及

- 15 所述数据同步系统用于，独立于所述第一数据存储系统，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给第二数据存储系统，以使所述第二数据存储系统根据所述对应于最新写操作的二进制日志对所述第二数据存储系统中的数据进行同步更新。

一种包括用于数据同步的可执行程序代码的非暂时性计算机程序产品，所述可执行程序代码可操作用于：当执行时，

- 20 根据在第一数据存储系统中进行的数据写操作生成二进制日志；
将所述二进制日志写入存储器；以及

- 25 独立于所述生成二进制日志和所述将所述二进制日志写入存储器，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给第二数据存储系统，以使所述第二数据存储系统根据所述对应于最新写操作的二进制日志对所述第二数据存储系统中的数据进行同步更新。

本发明实施例还提供了一种容灾系统，包括第一数据存储系统、数据同步系统以及第二数据存储系统，

- 30 所述第一数据存储系统用于根据在所述第一数据存储系统中进行的数据写操作生成二进制日志以及将所述二进制日志写入存储器；以及

所述数据同步系统用于，独立于所述第一数据存储系统，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于

最新写操作的二进制日志发送给所述第二数据存储系统；

所述第二数据存储系统用于根据所述对应于最新写操作的二进制日志对所述第二数据存储系统中的数据进行同步更新。

5 本发明实施例还提供了另一种容灾系统，包括多个用于数据同步的系统，其中每个用于数据同步的系统包括数据存储系统和数据同步系统，所述数据存储系统用于：

当在所述数据存储系统中进行数据的写操作时，根据所述写操作生成二进制日志，以及将所述二进制日志写入存储器；

10 当在另一个用于数据同步的系统的数据存储系统中进行数据的写操作时，根据其对应于最新写操作的二进制日志对所述数据存储系统中的数据进行同步更新。

所述数据同步系统用于，当在所述数据存储系统中进行数据的写操作时，独立于所述数据存储系统，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发
15 送给其他用于数据同步的系统的数据存储系统。

本发明实施例提供的技术方案带来的有益效果是：本发明实施例提出了一种基于缓存池和二进制日志文件的异步传送的同步策略，将数据存储系统与数据同步系统分离，数据同步系统负责将各个数据拷贝根据二进制日志更新到最新的状态，在该模式下，系统服务性能没有丝毫降
20 低的同时，系统的整体复杂度、耦合性、带宽成本都大幅降低。

附图说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图
25 仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 是本发明实施例提供的一种数据同步方法的流程图；

图 2 是本发明实施例提供的一种数据同步方法中的系统架构连接关系图；

30 图 3 是本发明实施例提供的一种数据同步系统的组成图；

图 4 是本发明实施例提供的一种用于数据同步方法的流程图；

图 5 是本发明实施例提供的一种用于数据同步的系统的框图；以及图 6 是本发明实施例提供的一种容灾系统的框图。

具体实施方式

5 本发明实施例提供一种数据同步方法及系统。

为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述。

请参考图 1，图 1 为本发明实施例提供的一种数据同步方法的流程图；图 2 是本发明实施例提供的一种数据同步方法中的系统架构连接关系图；请同时参考图 1 与图 2，所述方法包括：

步骤 S101：将外部的更新数据写入第一数据存储系统中；

当用户有写入操作时，服务进程将用户的更新数据写入第一数据存储系统中。其中，服务进程为给用户提供服务写入、读取等服务的模块，服务进程的数量可以为多个，分别对应于不同号段范围内的业务，所谓
15 号段为一个连续的 ID 范围，是部署或迁移的基本单位，例如每十万个连续 ID 成为一个部署号段。

步骤 S102：将该写入操作进行记录并生成二进制日志；

在将用户的更新数据成功写入第一数据存储系统之后，服务进程将本次的写入操作进行记录并生成二进制日志(Binary Log，以下简称
20 “BinLog”)，BinLog 记录了本次写入操作的一些基本信息例如写入时间、写操作序号、写操作内容等；

步骤 S103：将该二进制日志分别写入缓存池和磁盘的二进制日志文件中；

第一数据存储系统中设置了一个缓存池，该缓存池使用共享内存实现，用于存储用户写操作的 BinLog。当外部的更新数据写入第一数据存储系统后，服务进程将记录本次写入操作的 BinLog 写入到缓存池中。缓存池负责存储最近一段时间的 BinLog，当缓存池存储满的时候自动删除最早存入的 BinLog。

在第一数据存储系统的磁盘中还建立了一个二进制日志文件(以下简称“BinLog 文件”)，用于存储用户写入操作的 BinLog。BinLog 写入
30

到缓存池后，服务进程再将该 BinLog 写入到磁盘的 BinLog 文件中，然后对外部返回本次写操作成功的结果。一个 BinLog 文件可以写入的 BinLog 条数可以由系统进行设定，例如一个 BinLog 文件可以写入十万条的 BinLog，当一个 BinLog 文件写满十万条 BinLog，再建立一个新的 BinLog 文件进行写入新的 BinLog 即可。因此，BinLog 除了写入缓存池之外，还一并写入到了磁盘的 BinLog 文件中，由于缓存池对某一 BinLog 的存储有时间限制，当新的 BinLog 写入缓存池时，最早存入缓存池中的 BinLog 将被自动删除掉，而将 BinLog 写入到磁盘的 BinLog 文件中，使得该 BinLog 被保存下来，这样做的目的是即使机器突然掉电重启而导致缓存池的数据丢失，或者机器突然遭遇海量的写操作而导致最早写入缓存池中的 BinLog 还没进行同步即被缓存池自动删除掉的时候，该 BinLog 还能在磁盘的 BinLog 文件中找到，以保证后面同步系统能够读取到需要的同步数据。

步骤 S104：在将该更新数据进行同步时，从缓存池中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步；

数据的同步由数据同步系统中的同步进程完成，同步进程与服务进程异步运行，同步进程为负责数据同步的模块，同步进程负责的号段可以与服务进程负责的号段一致。

当有外部的更新数据写入第一数据存储系统中，且同步进程检测到某个数据拷贝（例如第二数据存储系统中的数据拷贝）的数据状态并没有达到最新时，就需要进行数据同步。在将该更新数据进行同步时，同步进程从缓存池中查找需要同步的该更新数据所对应的 BinLog，并将这些 BinLog 按先后顺序发送给该数据拷贝进行数据同步，使得每个数据拷贝（例如第一、第二数据存储系统中的数据拷贝）都达到最新数据状态。在同步数据时，可以一次只发送一个 BinLog，也可以一次发送多个 BinLog。

由于缓存池只保留最近一段时间的 BinLog，当需要同步的更新数据所对应的 BinLog 在缓存池中并没有找到时，上述方法进一步包括：

同步进程再从保存在磁盘中的 BinLog 文件中查找该更新数据所对应的 BinLog，并将该 BinLog 发送给第二数据存储系统，进而完成同步动

作。

同步进程在进行数据同步的时候，服务进程对外进行读写服务，同步进程和服务进程互相独立。

5 如果因为异常原因，磁盘中的二进制日志文件出现异常丢失，例如被误删除或者系统出现故障导致文件丢失时，所述方法进一步包括：

将第一数据存储系统中该二进制日志文件所涵盖的更新数据分别重新生成二进制日志，并将所述重新生成的二进制日志写入一个新的二进制日志文件中。

10 上述实施例中提出了一种基于缓存池和二进制日志文件的异步传送的同步策略，将数据存储系统与数据同步系统分离，第一数据存储系统只负责业务写的基本逻辑，而不用关心其它数据存储系统中数据拷贝的数据状态，数据同步系统负责将各个数据拷贝更新到最新的状态，在该模式下，系统服务性能没有丝毫降低的同时，系统的整体复杂度、耦合性、带宽成本都大幅降低。

15 请参考图 3，图 3 为本发明实施例提供的一种数据同步系统的组成图；所述系统包括：

第一写入模块 301，用于将外部的更新数据写入第一数据存储系统中；

生成模块 302，用于将该写入操作进行记录并生成二进制日志；

20 第二写入模块 303，用于将该二进制日志分别写入缓存池和磁盘的二进制日志文件中；

同步模块 304，用于在将该更新数据进行同步时，从缓存池中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步。

25 缓存池负责存储最近一段时间的二进制日志，第二写入模块 303 还用于当缓存池存储满的时候，自动删除最早存入的二进制日志。

第二写入模块 303 将二进制日志先写入到缓存池，再写入磁盘的二进制日志文件中，之后再对外部返回本次写操作成功的结果。

30 同步模块 304 在进行数据同步时，当需要同步的更新数据所对应的二进制日志在缓存池中没有找到时，同步模块 304 再从保存在磁盘中的

二进制日志文件中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步。

进一步地，所述系统还包括恢复模块 305，用于当磁盘中的二进制日志文件异常丢失时，将第一数据存储系统中该二进制日志文件所涵盖的更新数据分别重新生成二进制日志，并将所述重新生成的二进制日志写入一个新的二进制日志文件中。

关于本实施例的数据同步系统的进一步细节，还可参阅上述实施例的数据同步方法及相关描述。

图 4 示出了根据本发明一个优选实施例的用于数据同步的方法的流程图。如图 4 所示，该用于数据同步方法可以包括步骤 401、步骤 402 和步骤 403。

在步骤 401：根据在第一数据存储系统中进行的数据写操作生成 Binlog；

在步骤 402：将 Binlog 写入存储器；

在步骤 403：独立于生成 Binlog 和将 Binlog 写入存储器，在所写入的 Binlog 中查找对应于最新写操作的 Binlog，并将对应于最新写操作的 Binlog 发送给第二数据存储系统，以使第二数据存储系统根据对应于最新写操作的 Binlog 对第二数据存储系统中的数据进行同步更新。

外部计算设备向第一数据存储系统进行数据写操作。当数据写操作成功后，根据该数据写操作生成 Binlog。将 Binlog 写入存储器。上述过程可以由一个或多个服务进程实现。然后，在所写入的 Binlog 中查找对应于最新写操作的 Binlog，并将对应于最新写操作的 Binlog 发送给第二数据存储系统。该过程可以用一个或多个分别与服务进程相对应的同步进程实现。步骤 401、402 与步骤 403 之间彼此分离、相互独立。换言之，步骤 401、402 无需考虑步骤 403 完成与否，即，无需等待步骤 403 完成，即可再次启动。由此，上述用于数据同步的方法，在保证在多个数据存储系统之间成功同步的同时，方法的复杂度、耦合性显著降低。在多个数据存储系统之间进行同步所要求的带宽成本也大幅降低。

Binlog 可以用于恢复数据写操作。根据本发明一个优选实施例，上述方法还可以包括在第一数据存储系统中的数据丢失的情况下，根据所写入的 Binlog 模拟该写操作来恢复所丢失的数据。具体地，首先根据第

一数据存储系统中现存的数据，确定丢失的数据所对应的写操作。根据丢失的数据所对应的写操作在存储器中查找所写入的 Binlog。根据所写入的 Binlog 模拟写操作来恢复所丢失的数据。通过上述操作，有效保证了第一数据存储系统的安全性。

5 根据本发明一个优选实施例，存储器可以包括缓存池。优选地，缓存池可以使用内存实现。例如，在第一数据存储系统中，可以将内存中的一部分存储空间用作缓存池。当外部的更新数据写入第一数据存储系统后，服务进程将记录本次写操作的 Binlog 写入缓存池。使用内存实现缓存池易于实现，并且能够显著提高存取速率。

10 缓存池容量有限。根据本发明一个优选实施例，将 Binlog 写入存储器包括：当缓存池写满时，在缓存池中用当前待写入的 Binlog 替换最早写入的 Binlog。采用先入先出机制，易于实现。

 根据本发明一个优选实施例，存储器可以包括磁盘。磁盘一般为非易失性存储器。在第一数据存储系统突然掉电时，存储在磁盘中的数据
15 不会丢失，从而保证了 Binlog 的安全性。

 优选地，将 Binlog 写入存储器包括：将 Binlog 写入到磁盘的 Binlog 文件中，其中每个 Binlog 文件能够包括预设数目的 Binlog。每个 Binlog 可以具有唯一序号。这样，系统能够高效地利用磁盘管理 Binlog。

 根据本发明一个优选实施例，将 Binlog 写入存储器包括将 Binlog 写
20 入到缓存池和磁盘。在所写入的 Binlog 中查找对应于最新写操作的 Binlog 包括在缓存池中查找对应于最新写操作的 Binlog。在所写入的 Binlog 中查找对应于最新写操作的 Binlog 还包括对于未能在缓存池找到对应于最新写操作的 Binlog 的情况，在磁盘中查找对应于最新写操作的 Binlog。

25 根据本发明的一个优选实施例，Binlog 写入缓存池后，再将该 Binlog 写入保存在磁盘中的 Binlog 文件中。具体地，一方面，Binlog 写入缓存池可以保证同步进程能够迅速地从缓存池中查找到 Binlog。另一方面，可以保证同步进程在没有从缓存池中查找到所写入的 Binlog 的情况下，例如待查找的 Binlog 已经被后写入的 Binlog 所替换的情况，能够在保存在
30 磁盘中的 Binlog 文件中查找到所写入的 Binlog。这样，保证了数据同步系统能够读取到需要的 Binlog。

 根据本发明的一个实施例，本发明的用于数据同步的方法还包括：

比较所写入的 Binlog 的数目以及对第二数据存储系统中的数据进行同步更新的次数。其中，在所写入的 Binlog 的数目大于对第二数据存储系统中的数据进行同步更新的次数的情况下，进行在所写入的 Binlog 中查找对应于最新写操作的 Binlog，并将对应于最新写操作的 Binlog 发送给第二数据存储系统。

例如，在一个实施例中，所写入的 Binlog 数目是 6，对第二数据存储系统中的数据进行同步更新的次数是 4。因为 6 大于 4，所以从存储器中的所写入的 Binlog 中查找对应于最新写操作的 Binlog（即，第 5 个和第 6 个 Binlog）并将其按照先后顺序发送给第二数据存储系统，以使第二数据存储系统根据该 Binlog 对其数据拷贝进行同步更新。

上述检测第二数据存储系统中的数据状态的方法，可操作性强。

根据本发明一个实施例，该方法还可以包括：在将 Binlog 写入存储器之后，向执行写操作的计算设备返回写操作成功的信息。从而，使计算设备可以及时进行新的写操作。如前所述，该方法中的步骤 401、402 和 403 是相互独立的。因此，只要将 Binlog 写入存储器，即可认为写操作已经成功。无论第一数据存储系统以外的其他数据存储系统中的数据是否已经更新，都可以继续对第一数据存储系统进行新的写操作。

根据本发明另一方面，还提供了一种用于数据同步的系统。图 5 是本发明实施例提供的一种用于数据同步的系统的框图。如图 5 所示，该系统包括第一数据存储系统和数据同步系统。

第一数据存储系统用于根据在第一数据存储系统中进行的数据写操作生成 Binlog，以及将 Binlog 写入存储器。在图 5 中存储器示出为缓存池和磁盘中的 Binlog 文件。数据同步系统用于独立于第一数据存储系统，在所写入的 Binlog 中查找对应于最新写操作的 Binlog，并将对应于最新写操作的 Binlog 发送给第二数据存储系统，以使所述第二数据存储系统根据所述对应于最新写操作的 Binlog 对所述第二数据存储系统中的数据进行同步更新。

优选地，第一数据存储系统还用于在将 Binlog 写入存储器之后，向执行写操作的计算设备返回写操作成功的信息。

优选地，数据同步系统还用于在第一数据存储系统中的数据丢失的情况下，根据所写入的 Binlog 模拟写操作来恢复所丢失的数据。

本领域普通技术人员参考上面所详细描述的进行数据同步的方法，可以理解用于数据同步的系统的具体操作。为了简洁，在此不再赘述。

根据本发明一个实施例，还提供了一种包括用于数据同步的可执行程序代码的非暂时性计算机程序产品。可执行程序代码可操作用于：当
5 执行时，根据在第一数据存储系统中进行的数据写操作生成 Binlog；将 Binlog 写入存储器；以及独立于生成 Binlog 和将 Binlog 写入存储器，在所写入的 Binlog 中查找对应于最新写操作的 Binlog，并将对应于最新写操作的 Binlog 发送给第二数据存储系统，以使第二数据存储系统根据对应于最新写操作的 Binlog 对第二数据存储系统中的数据进行同步更新。

10 本领域普通技术人员可以理解，上述可执行程序代码进一步可操作用于当执行时，实施用于数据同步的方法的所有步骤。为了简洁，在此不对可执行程序代码的附加功能进行进一步描述。注意，代码可直接使处理器实施指定操作、经编译以使处理器实施指定操作、和/或其他软件、硬件、和/或固件元件（例如用于实施标准功能的库）结合以使处理器实施
15 指定操作。

根据本发明另一方面，还提供了一种容灾系统，如图 2 所示。所述系统包括第一数据存储系统、数据同步系统以及第二数据存储系统。

第一数据存储系统用于根据在第一数据存储系统中进行的数据写操作生成 Binlog 以及将 Binlog 写入存储器。数据同步系统用于独立于第一
20 数据存储系统，在所写入的 Binlog 中查找对应于最新写操作的 Binlog，并将对应于最新写操作的 Binlog 发送给第二数据存储系统。第二数据存储系统用于根据对应于最新写操作的 Binlog 对第二数据存储系统中的数据进行同步更新。

根据本发明一个优选实施例，第一数据存储系统和数据同步系统用
25 同一计算设备实现。

图 6 是根据本发明另一个实施例的容灾系统的框图。该容灾系统包括多个用于数据同步的系统。其中每个用于数据同步的系统包括数据存储系统和数据同步系统。

数据存储系统用于当在数据存储系统中进行数据的写操作时，根据
30 写操作生成 Binlog，以及将 Binlog 写入存储器。数据存储系统还用于当在另一个用于数据同步的系统的数据存储系统中进行数据的写操作时，根据其对应于最新写操作的 Binlog 对数据存储系统中的数据进行同步更

新。

数据同步系统用于当在数据存储系统中进行数据的写操作时，独立于数据存储系统，在所写入的 Binlog 中查找对应于最新写操作的 Binlog，并将对应于最新写操作的 Binlog 发送给其他用于数据同步的系统的数
5 据存储系统。

本领域普通技术人员参考上面所详细描述的进行数据同步的方法，可以理解上述容灾系统的具体操作。为了简洁，在此不再赘述。

上述容灾系统，每个用于数据同步的系统均包括数据存储系统和数据同步系统。因此，每个用于数据同步的系统均可配置为接收外部更新
10 数据。从而当当前用于接收外部更新数据的数据同步的系统故障时，另一个用于数据同步的系统可配置为替代它接收外部更新数据。由此，保持容灾系统一直正常运行。

上述实施例中提出的数据同步方法及系统，具有以下优点：

同步进程与服务进程异步工作，降低了耦合性，两个系统可以独立
15 设计、开发、上线以及维护，设计简单，运维成本降低，提高了同步的成功率，用户写入操作只要写 BinLog 文件成功就可以对外返回成功的结果；缓存池的引入大大减少了同步进程读磁盘的次数，提高了整个系统的性能；BinLog 文件保证了任何同步数据都可以查找到，新搭建一个数据拷贝的时候，不用停止业务的写服务，可以同步 BinLog 将新的数据拷
20 贝更新到最新状态。

需要说明的是，在本文中，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、物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、物品或者设备所固有的要素。
25 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、物品或者设备中还存在另外的相同要素。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可
30 存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存

储记忆体 (Read-Only Memory, ROM) 或随机存储记忆体 (Random Access Memory, RAM) 等。

以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

5

权利要求书

1、一种数据同步方法，其特征在于，包括：

将外部的更新数据写入第一数据存储系统中；

5 将该写入操作进行记录并生成二进制日志；

将该二进制日志分别写入缓存池和磁盘的二进制日志文件中；

在将该更新数据进行同步时，从缓存池中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步。

2、根据权利要求1所述的方法，其特征在于，当所述缓存池存满二进制日志时，自动删除最早存入的二进制日志。

3、根据权利要求1所述的方法，其特征在于，所述二进制日志先写入到缓存池，再写入磁盘的二进制日志文件中。

4、根据权利要求1所述的方法，其特征在于，当需要同步的该更新数据所对应的二进制日志在缓存池中没有找到时，所述方法进一步包括：

15 再从保存在磁盘中的二进制日志文件中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步。

5、根据权利要求4所述的方法，其特征在于，当磁盘中的二进制日志文件异常丢失时，所述方法进一步包括：

20 将第一数据存储系统中该二进制日志文件所涵盖的更新数据分别重新生成二进制日志，并将所述重新生成的二进制日志写入一个新的二进制日志文件中。

6、一种数据同步系统，其特征在于，包括：

第一写入模块，用于将外部的更新数据写入第一数据存储系统中；

生成模块，用于将该写入操作进行记录并生成二进制日志；

25 第二写入模块，用于将该二进制日志分别写入缓存池和磁盘的二进制日志文件中；

同步模块，用于在将该更新数据进行同步时，从缓存池中查找该更新数据所对应的二进制日志，并将该二进制日志发送给第二数据存储系统进行数据同步。

30 7、根据权利要求6所述的系统，其特征在于，所述第二写入模块还

用于当所述缓存池存满二进制日志时,自动删除最早存入的二进制日志。

8、根据权利要求6所述的系统,其特征在于,所述第二写入模块将所述二进制日志先写入到缓存池,再写入磁盘的二进制日志文件中。

9、根据权利要求6所述的系统,其特征在于,所述同步模块在进行数据同步时,当需要同步的更新数据所对应的二进制日志在缓存池中没
5 有找到时,所述同步模块再从保存在磁盘中的二进制日志文件中查找该更新数据所对应的二进制日志,并将该二进制日志发送给第二数据存储系统进行数据同步。

10、根据权利要求9所述的系统,其特征在于,所述系统还包括恢复模块,用于当磁盘中的二进制日志文件异常丢失时,将第一数据存储
10 系统中该二进制日志文件所涵盖的更新数据分别重新生成二进制日志,并将所述重新生成的二进制日志写入一个新的二进制日志文件中。

11、一种用于数据同步的方法,包括:

根据在第一数据存储系统中进行的数据写操作生成二进制日志;

15 将所述二进制日志写入存储器;以及

独立于所述生成二进制日志和所述将所述二进制日志写入存储器,在所写入的二进制日志中查找对应于最新写操作的二进制日志,并将所述对应于最新写操作的二进制日志发送给第二数据存储系统,以使所述第二数据存储系统根据所述对应于最新写操作的二进制日志对所述第二
20 数据存储系统中的数据进行同步更新。

12、根据权利要求11所述的方法,其特征在于,所述方法进一步包括:在所述第一数据存储系统中的所述数据丢失的情况下,根据所述所写入的二进制日志模拟所述写操作来恢复所丢失的数据。

13、根据权利要求11所述的方法,其特征在于,所述存储器包括缓存池。
25

14、根据权利要求13所述的方法,其特征在于,所述缓存池用内存实现。

15、根据权利要求14所述的方法,其特征在于,所述将所述二进制日志写入存储器包括:当所述缓存池写满时,在所述缓存池中用当前待
30 写入的二进制日志替换最早写入的二进制日志。

16、根据权利要求11所述的方法,其特征在于,所述存储器包括磁

盘。

17、根据权利要求 16 所述的方法，其特征在于，所述将所述二进制日志写入存储器包括：

5 将所述二进制日志写入到所述磁盘的二进制日志文件中，其中每个二进制日志文件能够包括预设数目的二进制日志。

18、根据权利要求 11 所述的方法，其特征在于，

所述将所述二进制日志写入存储器包括将所述二进制日志写入到缓存池和磁盘；

10 所述在所写入的二进制日志中查找对应于最新写操作的二进制日志包括：

在所述缓存池中查找所述对应于最新写操作的二进制日志；以及

对于未能在所述缓存池找到所述对应于最新写操作的二进制日志的情况，在所述磁盘中查找所述对应于最新写操作的二进制日志。

15 19、根据权利要求 11 所述的方法，其特征在于，所述方法进一步包括：

比较所述所写入的二进制日志的数目以及对所述第二数据存储系统中的数据进行同步更新的次数；

20 其中，在所述所写入的二进制日志的数目大于对所述第二数据存储系统中的数据进行同步更新的次数的情况下，进行所述在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给第二数据存储系统。

20、根据权利要求 11 所述的方法，其特征在于，所述方法进一步包括：在将所述二进制日志写入存储器之后，向执行所述写操作的计算设备返回所述写操作成功的信息。

25 21、一种用于数据同步的系统，包括第一数据存储系统以及数据同步系统，

所述第一数据存储系统用于根据在所述第一数据存储系统中进行的数据写操作生成二进制日志以及将所述二进制日志写入存储器；以及

30 所述数据同步系统用于，独立于所述第一数据存储系统，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给第二数据存储系统，以使所述第二数据存储系统根据所述对应于最新写操作的二进制日志对所述第二数据存储

系统中的数据进行同步更新。

22、根据权利要求 21 所述的用于数据同步的系统，其特征在于，所述数据同步系统进一步用于在所述第一数据存储系统中的所述数据丢失的情况下，根据所述所写入的二进制日志模拟所述写操作来恢复所丢失的数据。

23、根据权利要求 21 所述的用于数据同步的系统，其特征在于，所述第一数据存储系统进一步用于在将所述二进制日志写入存储器之后，向执行所述写操作的计算设备返回所述写操作成功的信息。

24、一种包括用于数据同步的可执行程序代码的非暂时性计算机程序产品，所述可执行程序代码可操作用于：当执行时，

根据在第一数据存储系统中进行的数据写操作生成二进制日志；

将所述二进制日志写入存储器；以及

独立于所述生成二进制日志和所述将所述二进制日志写入存储器，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给第二数据存储系统，以使所述第二数据存储系统根据所述对应于最新写操作的二进制日志对所述第二数据存储系统中的数据进行同步更新。

25、一种容灾系统，包括第一数据存储系统、数据同步系统以及第二数据存储系统，

所述第一数据存储系统用于根据在所述第一数据存储系统中进行的数据写操作生成二进制日志以及将所述二进制日志写入存储器；以及

所述数据同步系统用于，独立于所述第一数据存储系统，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给所述第二数据存储系统；

所述第二数据存储系统用于根据所述对应于最新写操作的二进制日志对所述第二数据存储系统中的数据进行同步更新。

26、根据权利要求 25 所述的容灾系统，其特征在于，所述第一数据存储系统和所述数据同步系统用同一计算设备实现。

27、一种容灾系统，包括多个用于数据同步的系统，其中每个用于数据同步的系统包括数据存储系统和数据同步系统，

所述数据存储系统用于：

当在所述数据存储系统中进行数据的写操作时，根据所述写操作生

成二进制日志，以及将所述二进制日志写入存储器；

当在另一个用于数据同步的系统的数据存储系统中进行数据的写操作时，根据其对应于最新写操作的二进制日志对所述数据存储系统中的数据进行同步更新。

- 5 所述数据同步系统用于，当在所述数据存储系统中进行数据的写操作时，独立于所述数据存储系统，在所写入的二进制日志中查找对应于最新写操作的二进制日志，并将所述对应于最新写操作的二进制日志发送给其他用于数据同步的系统的数据存储系统。

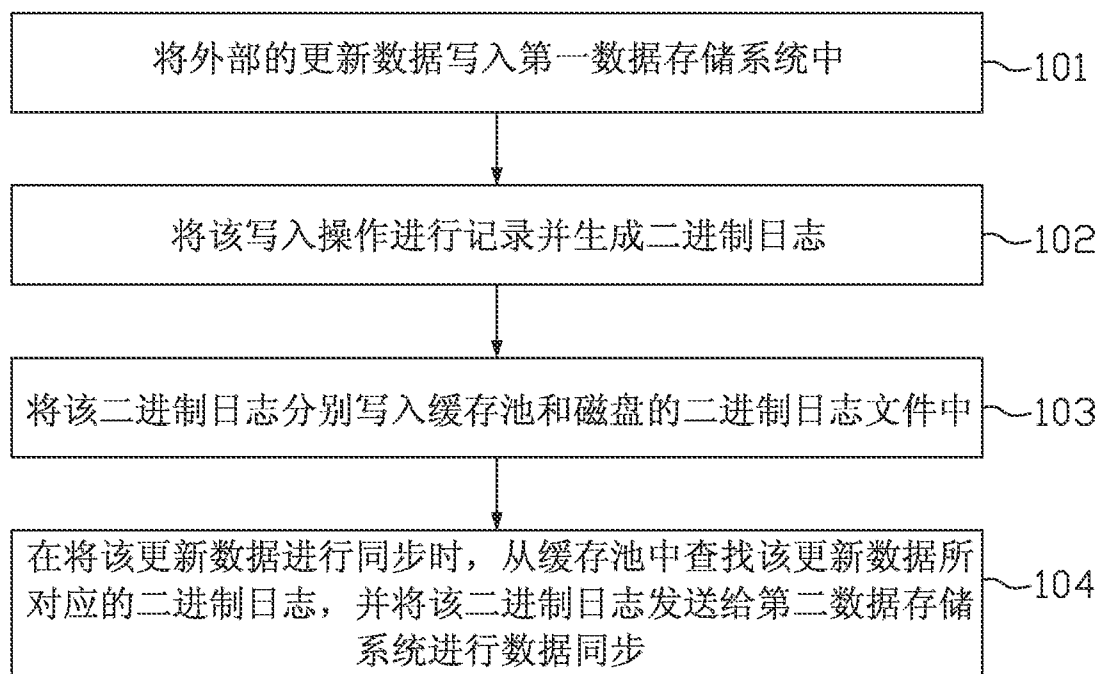


图 1

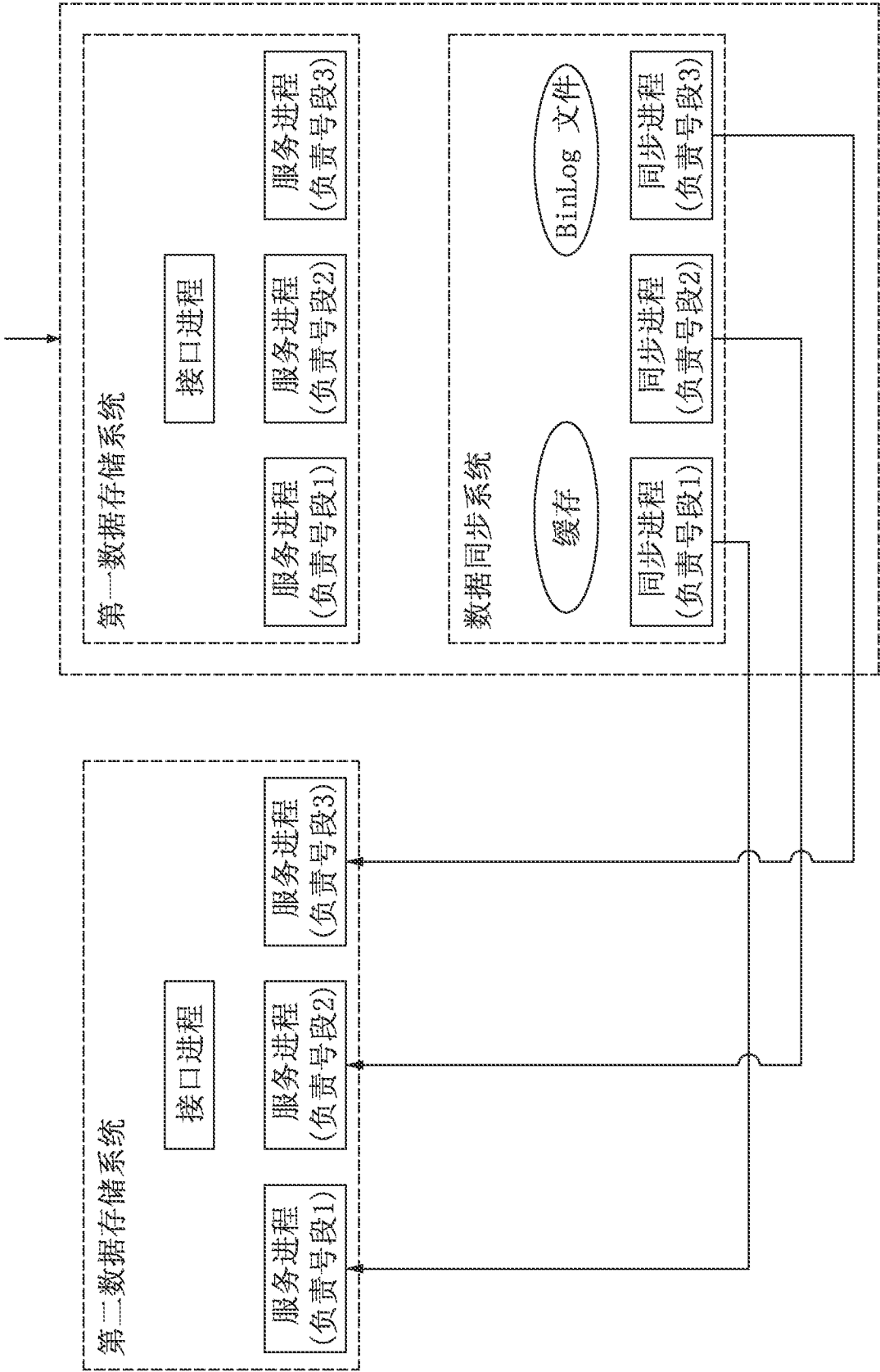


图 2

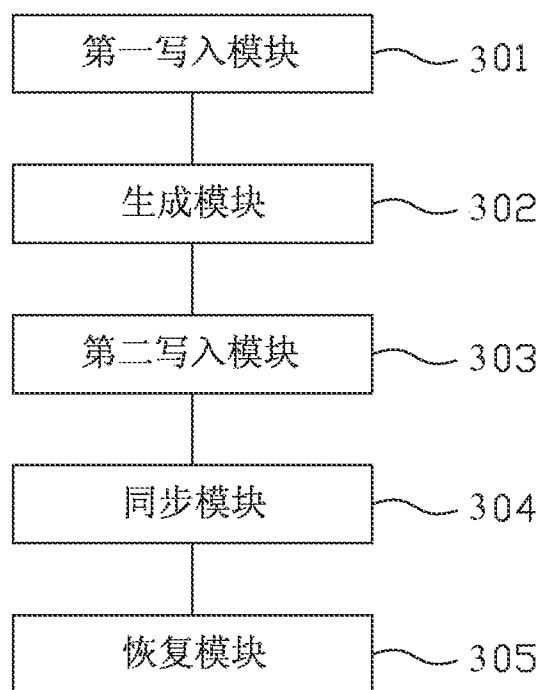


图 3

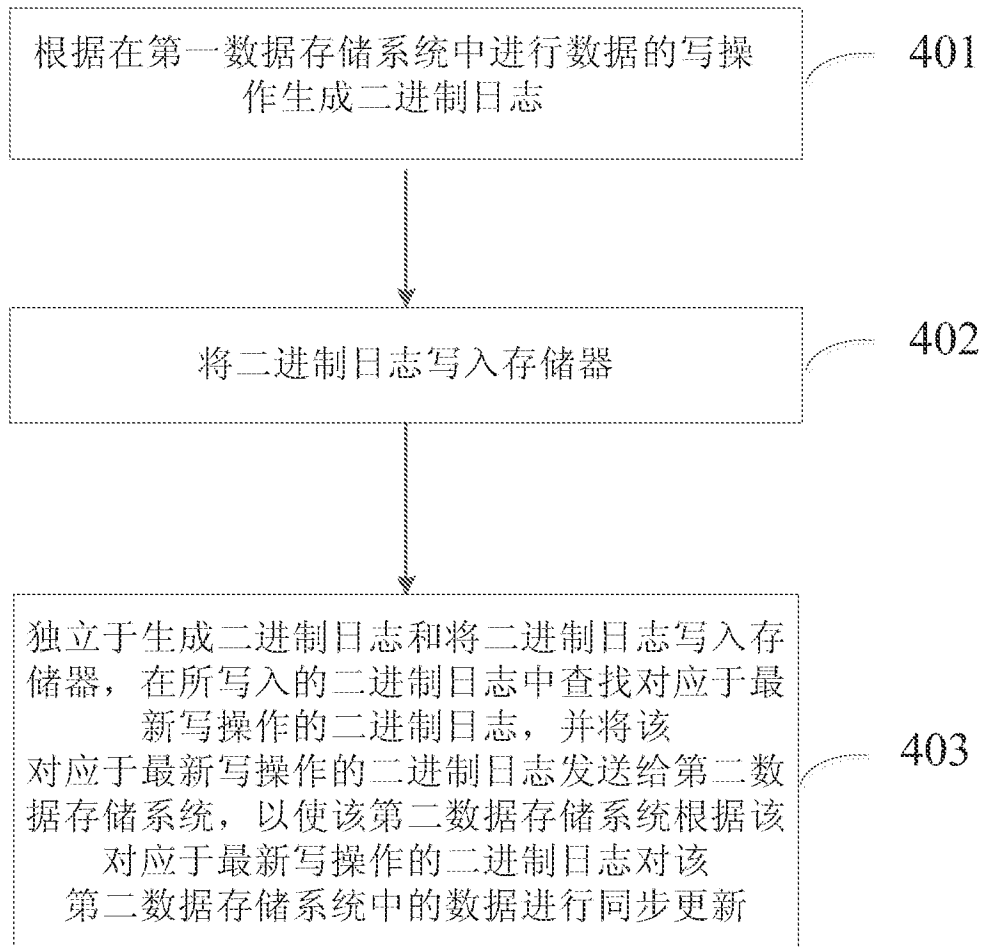


图4

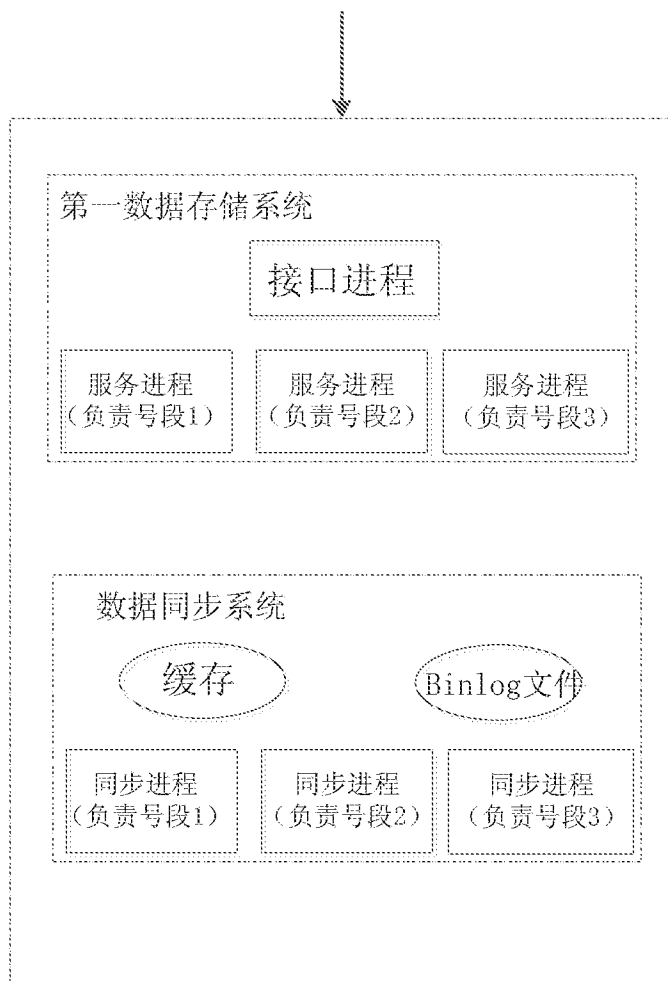


图5

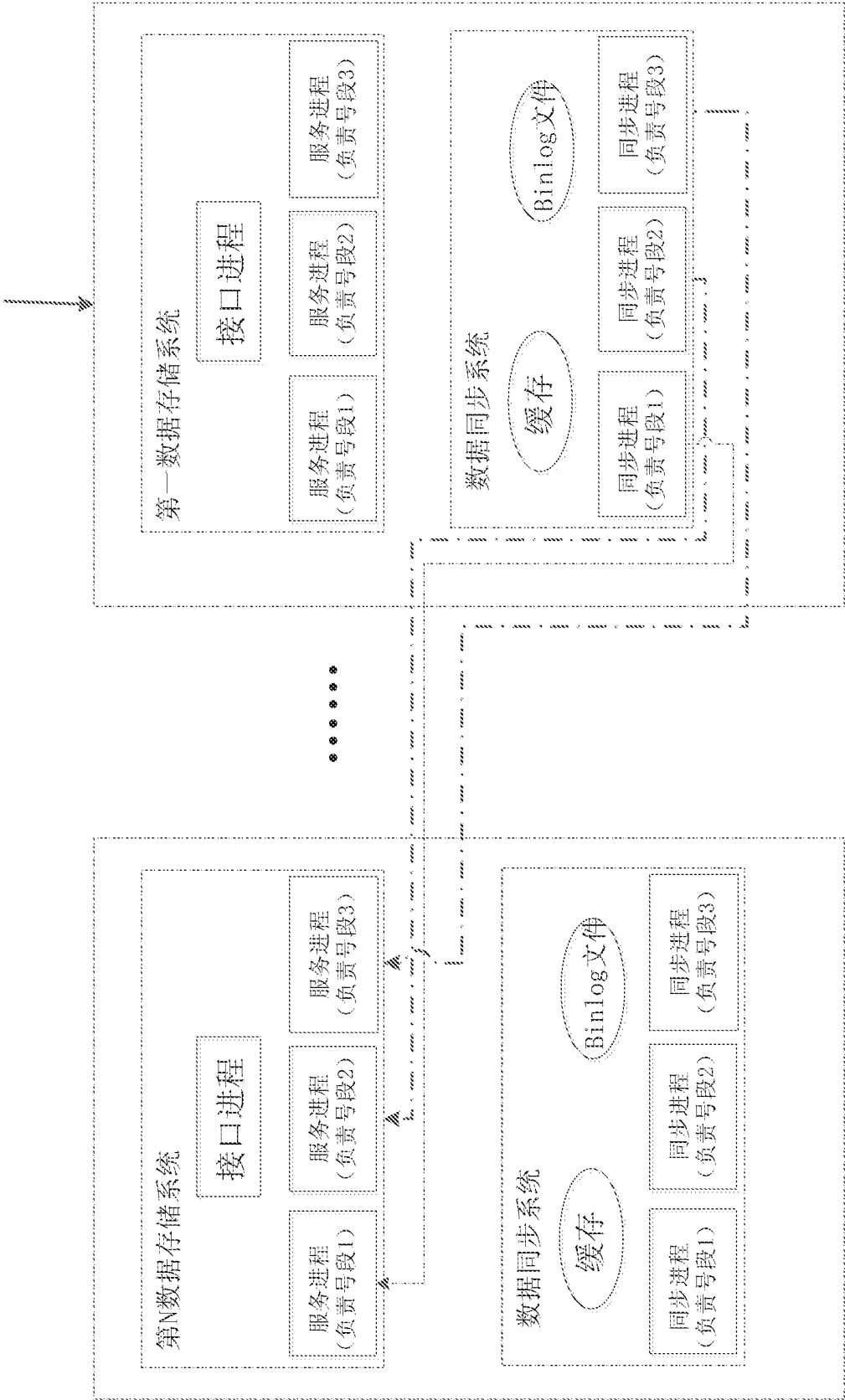


图6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/079087

A. CLASSIFICATION OF SUBJECT MATTER

G06F 12/00 (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)

IPC: G06F; 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, GOOGLE, CNPAT, CNKI, IEEE: update, backup, data, store, memory, synchronous, asynchronous, log, cache, binary, file

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 1753373 A (HUAWEI TECHNOLOGIES CO., LTD.), 29 March 2006 (29.03.2006), description, pages 7-8, and figures 4-5	1-27
A	CN 1670709 A (HITACHI LTD.), 21 September 2005 (21.09.2005), the whole document	1-27
A	CN 102567338 A (SYSCOM COMPUTER ENGINEERING CO.), 11 July 2012 (11.07.2012), the whole document	1-27
A	US 2009/0259698 A1 (BROTHER KOGYO KABUSHIKI KAISHA), 15 October 2009 (15.10.2009), the whole document	1-27
A	US 2005/0027748 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION), 03 February 2005 (03.02.2005), the whole document	1-27

☐ 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 26 September 2013 (26.09.2013)	Date of mailing of the international search report 17 October 2013 (17.10.2013)
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 WU, Shaohong Telephone No.: (86-10) 62413694

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2013/079087

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 1753373 A	29.03.2006	None	
CN 1670709 A	21.09.2005	EP 1577775 A1	21.09.2005
		US 2005210073 A1	22.09.2005
		CN 101231606 A	30.07.2008
		JP 2005267301 A	29.09.2005
CN 102567338 A	11.07.2012	None	
US 2009/0259698 A1	15.10.2009	JP 2008165328 A	17.07.2008
		EP 2098960 A1	09.09.2009
		WO 2008078450 A1	03.07.2008
US 2005/0027748 A1	03.02.2005	None	

A. 主题的分类

G06F 12/00 (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)

IPC: G06F; 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, GOOGLE, CNPAT, CNKI, IEEE: 数据, 存储, 同步, 异步, 更新, 日志, 缓存, 内存, 二进制, 文件, 备份, data, store, memory, synchronous, asynchronous, log, cache, binary, file

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 1753373 A (华为技术有限公司) 29.3 月 2006 (29.03.2006) 说明书第 7-8 页、附图 4-5	1-27
A	CN 1670709 A (株式会社日立制作所) 21.9 月 2005 (21.09.2005) 全文	1-27
A	CN 102567338 A (凌群电脑股份有限公司) 11.7 月 2012 (11.07.2012) 全文	1-27
A	US 2009/0259698 A1 (BROTHER KOGYO KABUSHIKI KAISHA) 15.10 月 2009 (15.10.2009) 全文	1-27
A	US 2005/0027748 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 03.2 月 2005 (03.02.2005) 全文	1-27



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

26.9 月 2013 (26.09.2013)

国际检索报告邮寄日期

17.10 月 2013 (17.10.2013)

ISA/CN 的名称和邮寄地址:

中华人民共和国国家知识产权局

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

吴少鸿

电话号码: (86-10) 62413694

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2013/079087

检索报告中引用的 专利文件	公布日期	同族专利	公布日期
CN 1753373 A	29.03.2006	无	
CN 1670709 A	21.09.2005	EP 1577775 A1	21.09.2005
		US 2005210073 A1	22.09.2005
		CN 101231606 A	30.07.2008
		JP 2005267301 A	29.09.2005
CN 102567338 A	11.07.2012	无	
US 2009/0259698 A1	15.10.2009	JP 2008165328 A	17.07.2008
		EP 2098960 A1	09.09.2009
		WO 2008078450 A1	03.07.2008
US 2005/0027748 A1	03.02.2005	无	