(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102439560 B
(45)授权公告日 2016.02.10

(21)申请号 201080015183.4
(22)申请日 2010.03.29
(30)优先权数据
12/384210 2009.03.31 US
(85)PCT国际申请进入国家阶段日
2011.09.28
(86)PCT国际申请的申请数据
PCT/US2010/000947 2010.03.29
(87)PCT国际申请的公布数据
WO2010/114598 EN 2010.10.07
(73)专利权人 EMC 公司
地址 美国麻萨诸塞州
(72)发明人 H. 钟 D. 莫赫 S.V. 里迪
(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 姜冰 朱海煜
(51)Int.Cl.
G06F 7/00(2006.01)

(56)对比文件
US 6647393 B1,2003.11.11, 第3栏第11行至第12栏第67行及图1-5.
US 6647393 B1,2003.11.11, 第3栏第11行至第12栏第67行及图1-5.
US 7222119 B1,2007.05.22, 第10栏第11行至第44行.
CN 1534518 A,2004.10.06, 全文.
潘群华等. 分布式数据库系统中数据一致性维护方法. 《计算机工程》.2002, 第28卷(第9期), 全文.

审查员 吕鑫

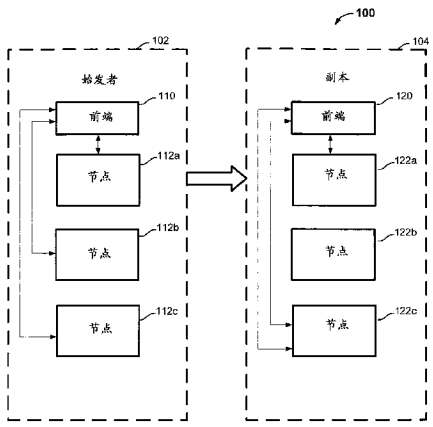
权利要求书3页 说明书4页 附图7页

(54)发明名称

数据复制系统中的数据重新分发

(57)摘要

一种系统包括一个或多个处理器,所述处理器配置成在多个始发者节点之间重新分发一个或多个始发者数据子集并确定与所述多个始发者节点之间所述一个或多个始发者数据子集的重新分发有关的数据重新分发信息。该系统还包括通信接口,该通信接口配置成将数据重新分发信息发送到副本系统。数据重新分发信息由副本系统用于在多个副本节点之间重新分发一个或多个对应的副本数据子集。



CN 102439560 B

1. 一种用于数据复制的系统，包括：

一个或多个处理器，配置成；

在多个始发者节点之间重新分发一个或多个始发者数据子集；以及

确定与所述多个始发者节点之间所述一个或多个始发者数据子集的重新分发有关的数据重新分发信息；以及

通信接口，配置成将数据重新分发信息发送到副本系统；其中

所述数据重新分发信息由所述副本系统用于在多个副本节点之间重新分发一个或多个对应的副本数据子集，

其中所述数据重新分发信息包括所述一个或多个始发者数据子集已从其移动的一个或多个以前始发者节点的标识信息和所述始发者数据子集当前驻留于其上的一个或多个当前始发者节点的标识信息，

其中所述数据重新分发信息包括除备份数据以外的数据。

2. 如权利要求 1 所述的系统，其中所述一个或多个始发者数据子集包括一个或多个数据容器。

3. 如权利要求 1 所述的系统，其中所述重新分发信息包括与所述一个或多个数据容器相关联的元数据。

4. 如权利要求 1 所述的系统，其中所述重新分发信息包括与所述一个或多个数据容器相关联的元数据，并且所述一个或多个数据容器的每个包括一个或多个数据段。

5. 如权利要求 1 所述的系统，其中所述一个或多个始发者数据子集和所述一个或多个副本数据子集包括相同的备份数据。

6. 如权利要求 1 所述的系统，其中所述多个始发者节点包括在文件系统中。

7. 如权利要求 1 所述的系统，其中所述一个或多个始发者数据子集从一个或多个现有节点重新分发到新添加的节点。

8. 如权利要求 1 所述的系统，其中所述一个或多个始发者数据子集被重新分发以重新平衡所述多个始发者节点上的负载。

9. 一种用于数据复制的方法，包括：

在多个始发者节点之间重新分发一个或多个始发者数据子集；以及

确定与所述多个始发者节点之间所述一个或多个始发者数据子集的重新分发有关的数据重新分发信息；以及

将数据重新分发信息发送到副本系统；其中

所述数据重新分发信息由所述副本系统用于在多个副本节点之间重新分发一个或多个对应的副本数据子集，

其中所述数据重新分发信息包括所述一个或多个始发者数据子集已从其移动的一个或多个以前始发者节点的标识信息和所述始发者数据子集当前驻留于其上的一个或多个当前始发者节点的标识信息，

其中所述数据重新分发信息包括除备份数据以外的数据。

10. 如权利要求 9 所述的方法，其中所述一个或多个始发者数据子集包括一个或多个数据容器。

11. 如权利要求 9 所述的方法，其中所述重新分发信息包括与所述一个或多个数据容

器相关联的元数据。

12. 如权利要求 9 所述的方法，其中所述重新分发信息包括与所述一个或多个数据容器相关联的元数据，并且所述一个或多个数据容器的每个包括一个或多个数据段。

13. 如权利要求 9 所述的方法，其中所述一个或多个始发者数据子集和所述一个或多个副本数据子集包括相同的备份数据。

14. 一种用于数据复制的设备，包括 ：

用于在多个始发者节点之间重新分发一个或多个始发者数据子集的部件 ；以及

用于确定与所述多个始发者节点之间所述一个或多个始发者数据子集的重新分发有关的数据重新分发信息的部件 ；以及

用于将数据重新分发信息发送到副本系统的部件 ；其中

所述数据重新分发信息由所述副本系统用于在多个副本节点之间重新分发一个或多个对应的副本数据子集，

其中所述数据重新分发信息包括所述一个或多个始发者数据子集已从其移动的一个或多个以前始发者节点的标识信息和所述始发者数据子集当前驻留于其上的一个或多个当前始发者节点的标识信息，

其中所述数据重新分发信息包括除备份数据以外的数据。

15. 一种用于数据复制的系统，包括 ：

接口，配置成接收来自始发者系统的数据重新分发信息，所述数据重新分发信息与多个始发者节点之间一个或多个始发者数据子集的重新分发有关 ；以及

一个或多个处理器，配置成根据所述数据重新分发信息，在多个副本节点之间重新分发一个或多个对应的副本数据子集，

其中所述数据重新分发信息包括所述一个或多个始发者数据子集已从其移动的一个或多个以前始发者节点的标识信息和所述始发者数据子集当前驻留于其上的一个或多个当前始发者节点的标识信息，

其中所述数据重新分发信息包括除备份数据以外的数据。

16. 一种用于数据复制的方法，包括 ：

接收来自始发者系统的数据重新分发信息，所述数据重新分发信息与多个始发者节点之间一个或多个始发者数据子集的重新分发有关 ；以及

根据所述数据重新分发信息，在多个副本节点之间重新分发一个或多个对应的副本数据子集，

其中所述数据重新分发信息包括所述一个或多个始发者数据子集已从其移动的一个或多个以前始发者节点的标识信息和所述始发者数据子集当前驻留于其上的一个或多个当前始发者节点的标识信息，

其中所述数据重新分发信息包括除备份数据以外的数据。

17. 一种用于数据复制的设备，包括 ：

用于接收来自始发者系统的数据重新分发信息的部件，所述数据重新分发信息与多个始发者节点之间一个或多个始发者数据子集的重新分发有关 ；以及

用于根据所述数据重新分发信息，在多个副本节点之间重新分发一个或多个对应的副本数据子集的部件，

其中所述数据重新分发信息包括所述一个或多个始发者数据子集已从其移动的一个或多个以前始发者节点的标识信息和所述始发者数据子集当前驻留于其上的一个或多个当前始发者节点的标识信息，

其中所述数据重新分发信息包括除备份数据以外的数据。

数据复制系统中的数据重新分发

背景技术

[0001] 在许多现有数据复制系统中，在始发者与副本之间同步数据。有关始发者的任何更改被发送到副本并被镜像。频繁的数据更新消耗了许多带宽，并导致效率低下。在始发者和副本分隔有广域网 (WAN) 及带宽有限的环境中，该问题特别明显。

附图说明

[0002] 在下面的详细描述和附图中，公开了本发明的不同实施例。

[0003] 图 1 是示出数据复制环境的实施例的框图。

[0004] 图 2 是示出用于数据复制的过程的一实施例的流程图。

[0005] 图 3 是示出用于数据复制的过程的另一实施例的流程图。

[0006] 图 4 是示出容器的实施例的数据结构图。

[0007] 图 5A-5C 是示出重新分发数据的示例情形的一系列图形。

具体实施方式

[0008] 本发明能够以多种方式实现，包括作为过程、设备、系统、组合物、计算机可读存储媒体上包含的计算机程序产品和 / 或处理器，例如配置成执行耦合到处理器的存储器上所存储的和 / 或由该存储器所提供的指令的处理器。在此说明书中，这些实现或本发明可采用的任何其它形式可称为技术。通常，在本发明范围内可改变公开的过程的步骤顺序。除非另有说明，否则，被描述为配置成执行某个任务的诸如处理器和存储器等组件可实现为暂时配置成在给定时间执行该任务的通用组件或制造为执行该任务的特定组件。在本文中使用时，术语“处理器”指配置成处理诸如计算机程序指令等数据的一个或多个装置、电路和 / 或处理核。

[0009] 本发明一个或多个实施例的详细描述在下面与显示本发明原理的附图一起提供。本发明结合此类实施例进行描述，但本发明并不限于任何实施例。本发明的范围只受权利要求的限制，并且本发明包括许多备选、修改和等同。许多特定的细节在下面的描述中陈述以便提供本发明的详细理解。这些细节被提供以用于示例的目的，并且本发明可在一些或所有这些特定细节不存在的情况下根据权利要求来实践。为了清晰的目的，与本发明相关的技术领域已知的技术材料未详细描述以免不必要地混淆本发明。

[0010] 图 1 是示出数据复制环境的实施例的框图。在此示例中，数据复制系统 100 包括始发者系统 102（也称为源系统）和副本系统 104（也称为目的地系统）。这些系统由诸如局域网或广域网等一个或多个网络来分隔。

[0011] 始发者系统包括始发者前端装置 110 和多个始发者节点 112a、112b 和 112c（也称为始发者后端装置）。副本系统包括副本前端装置 120 和多个副本节点 122a、122b 和 122c（也称为副本后端装置）。不同数量的节点和前端装置与节点的不同布置是可能的。例如，前端装置和节点的功能能够集成到单个物理装置中。

[0012] 节点用于存储数据。在各种实施例中，节点使用任何适当类型的装置来实现，例如

存储装置或包括存储组件的文件服务器。前端装置也能够使用多种装置来实现，例如运行数据复制管理软件的通用服务器。每个前端装置与其相应节点通信，协调节点上的数据存储以实现虚拟化的文件系统。换而言之，对于通过前端装置访问数据的外部装置，前端装置好像是管理单个文件系统的文件系统服务器。在一些实施例中，前端和后端节点在带有分开的存储分区的一个物理装置上共存。

[0013] 如下面将更详细描述的，始发者和副本系统相互通信。更具体地说，始发者系统能够将备份信息发送到副本前端装置，包括有关新数据的信息和有关现有数据的分发的信息。通信可在前端装置之间进行，或者直接在节点之间进行。

[0014] 在一些实施例中，备份数据流由前端装置接收和处理，并且分发到始发者节点以进行存储。在图 1 所示的示例中，副本上的数据保持作为始发者上数据的镜像映像。在新数据变得可用时，它存储在始发者上，并复制在副本上。在例如 100 等其中始发者和副本具有相同节点配置的系统中，特定始发者节点上的数据复制在对应副本节点（有时称为“好友”）上。例如，节点 112b 上存储的新数据复制在好友节点 122b 上。在一些实施例中，在前端装置上保持有关节点及其好友的知识。各个节点可相互直接通信，并且始发者节点将要复制的数据直接发送到其好友。备选的是，始发者节点与始发者前端装置通信，该前端装置又与副本前端装置通信以便将复制的数据传送到适当的副本节点。

[0015] 在一些情况下，始发者上的现有数据能够从一个始发者节点移到另一始发者节点。例如，如果数据分发变得不均匀，换而言之，太多数据存储在某些节点上，而太少数据存储在其它节点上，则系统将在节点之间重新平衡数据分发。导致数据重新分发的另一种情况是在新节点添加到系统时 - 数据从现有节点重新分发到新节点。在数据重新分发出现时，与重新分发的数据有关的信息从始发者发送到副本，以便数据能够在副本上以相同的方式重新分发。然而，数据本身不重新发送。由于不再要求将已复制数据复制到新副本节点，然后删除旧副本节点上存储的相同数据，因此，整个系统有效地处理数据重新分发。

[0016] 图 2 是示出用于数据复制的过程的一实施例的流程图。在一些实施例中，过程 200 在诸如 102 等始发者系统上执行。在一些实施例中，该过程由前端装置 110 实现。在 202，在多个始发者节点之间重新分发一个或多个始发者数据子集。换而言之，始发者数据从某些始发者节点移到其它始发者节点。重新分发在系统执行负载平衡、新节点变得添加到网络、现有节点变得从网络删除时、或由于任何其它适当的原因而可能发生。在一些实施例中，数据子集是在下面更详细描述的数据容器。在 204，确定与如何重新分发数据子集有关的数据重新分发信息。在一些实施例中，数据重新分发信息包括与已从其移动了数据子集的源始发者节点和始发者数据子集移到的目的地始发者节点有关的信息。在 206，数据分发信息经通信接口发送到副本系统，副本系统使用数据重新分发信息在副本节点之间重新分发对应的副本数据子集。

[0017] 图 3 是示出用于数据复制的过程的另一实施例的流程图。在一些实施例中，过程 300 在诸如 104 等副本系统上执行。在一些实施例中，该过程由前端装置 120 实现。在 302，从始发者接收数据重新分发信息。数据重新分发信息可从实现过程 200 的始发者发送。在 304，根据数据重新分发信息，在副本系统上重新分发一个或多个对应的副本数据子集。如前面所述，数据重新分发信息包括与重新分发的数据子集相关联的源节点和目的地节点有关的信息。假设每个始发者节点具有对应的好友副本节点，并且最初相同的始发者数据子

集和副本数据子集分别存储在始发者节点和对应的副本节点上，以及数据子集在始发者节点之间的最初分发与在副本节点之间的分发相同。因此，给定数据重新分发信息时，副本系统能够以与始发者系统相同的方式重新分发其现有数据子集，而不带来重复的数据传送开销。

[0018] 在一些实施例中，上述过程中使用的数据子集是容器。在各种实施例中，容器大小可以是几兆字节。例如，在一些实施例中使用 4.5 MB 的容器。节点可存储多个容器。图 4 是示出容器的实施例的数据结构图。在此示例中，容器 400 包括备份数据部分 404 和元数据部分 402。备份数据部分包括要求备份的实际数据，并且元数据部分包括与备份数据部分有关的信息，其用于促进数据备份。备份数据部分包括多个数据段，这些数据段是数据存储子单元，并且可以有不同大小。在始发者上接收数据时，例如，在由前端装置读取数据流时，将数据分成数据段，并且生成适当的段标识符 (ID)。前端装置还执行诸如检查数据段以验证未收到重复段等功能。数据段如何在数据流中布置以便在以后可重构数据流的记录保持在前端装置上或者存储在一个或多个节点中。

[0019] 数据段打包到适当的容器中，并且其对应偏移和段 ID 记录在元数据部分。元数据部分包括多个偏移 / 段标识符 (ID) 对。偏移指示数据段的开始的偏移。段 ID 用于标识数据段。在一些实施例中，使用了唯一地标识数据段的指纹或修改的指纹。元数据部分中还包括有用于标识此容器的容器 ID、用于标识容器当前驻留的节点(即，容器移到的目的地节点)的当前节点 ID 及用于标识容器以前驻留的节点(即，从其移动了容器的源节点)的以前节点 ID。在一些实施例中，容器 ID、当前节点 ID 和以前节点 ID 用于在复制期间促进容器重新分发过程。

[0020] 图 5A-5C 是示出由于新节点添加到系统而重新分发数据的示例情形的一系列图形。在图 5A 中，数据复制系统 100 配置成包括始发者系统 102 和副本系统 104。在始发者系统上，数据容器 115、117 和 119 分别分发在始发者节点 112a、112b 和 112c 上。每个节点还包括图中未示出的另外容器。在对始发者系统镜像的副本系统上，对应的复制数据容器 125、127 和 129 分发在副本节点 122a、122b 和 122c 上。这些副本容器以前从始发者复制。虽然在此示例中，诸如始发者系统中的前端装置、节点和数据容器等始发者组件示为具有与副本系统中的组件不同的标签 /ID，但在一些实施例中，始发者组件和副本上的其对应相对物共享相同的标识符。只要副本系统能够将始发者组件和副本上的其相对物相关联，各种标识方案便能够被使用。

[0021] 在图 5B 中，新节点 112d 添加到始发者系统，并且对应的新节点 122d 也添加到副本系统。因此，始发者和副本系统上存储的数据应重新平衡。在此示例中，例如 200 的过程在始发者系统 102 上进行。具体而言，在始发者系统上，重新分发容器 115、117 和 119。不重新发送这些容器到副本，而是确定数据分发信息。在此情况下，容器 115、117 和 119 已移到新节点 112d。因此，数据重新分发信息发送到副本系统。在此情况下，数据重新分发信息包括与重新分发的容器有关的元数据信息的紧密集合，其包括容器的 ID、容器以前驻留的相应节点的 ID 及容器重新分发到的和容器当前驻留的当前节点的 ID。诸如容器中数据段等实际备份数据在此示例中未发送，并且带宽被保留。

[0022] 在图 5C 中，例如 300 的过程在此示例中在副本系统 104 上进行。在从始发者系统接收数据重新分发信息时，根据数据重新分发信息而重新分发副本系统上的数据容器。在

此示例中，前端装置 120 接收和解析重新分发信息，并且与副本节点协调以便以与始发者上重新分发对应容器相同的方式来重新分发数据容器。基于所给的数据重新分发，数据容器 125、127 和 129（分别对应于容器 115、117 和 119）移到新节点 122d。

[0023] 上述过程也可被执行以响应负载平衡。在一个示例中，节点 112a-c 和 122a-c 是现有节点，并且节点 112d 和 122d 也是现有节点，而不是新添加的节点。确定的是，节点 112a、112b 和 112c 上存储了太多数据，并且节点 112d 和 122d 上未存储足够数据。因此，执行类似于图 5A-5C 中所述的过程，以重新分发数据和平衡各种节点上存储的数据量。通过使用数据重新分发信息，不必跨网络发送数据容器，并且能够快速、有效地实现负载平衡。

[0024] 虽然为了理解的清楚性目的而以一定的细节描述了上述实施例，但本发明并不限于提供的细节。实施本发明有许多备选方式。公开的实施例是说明性而非限制性的。

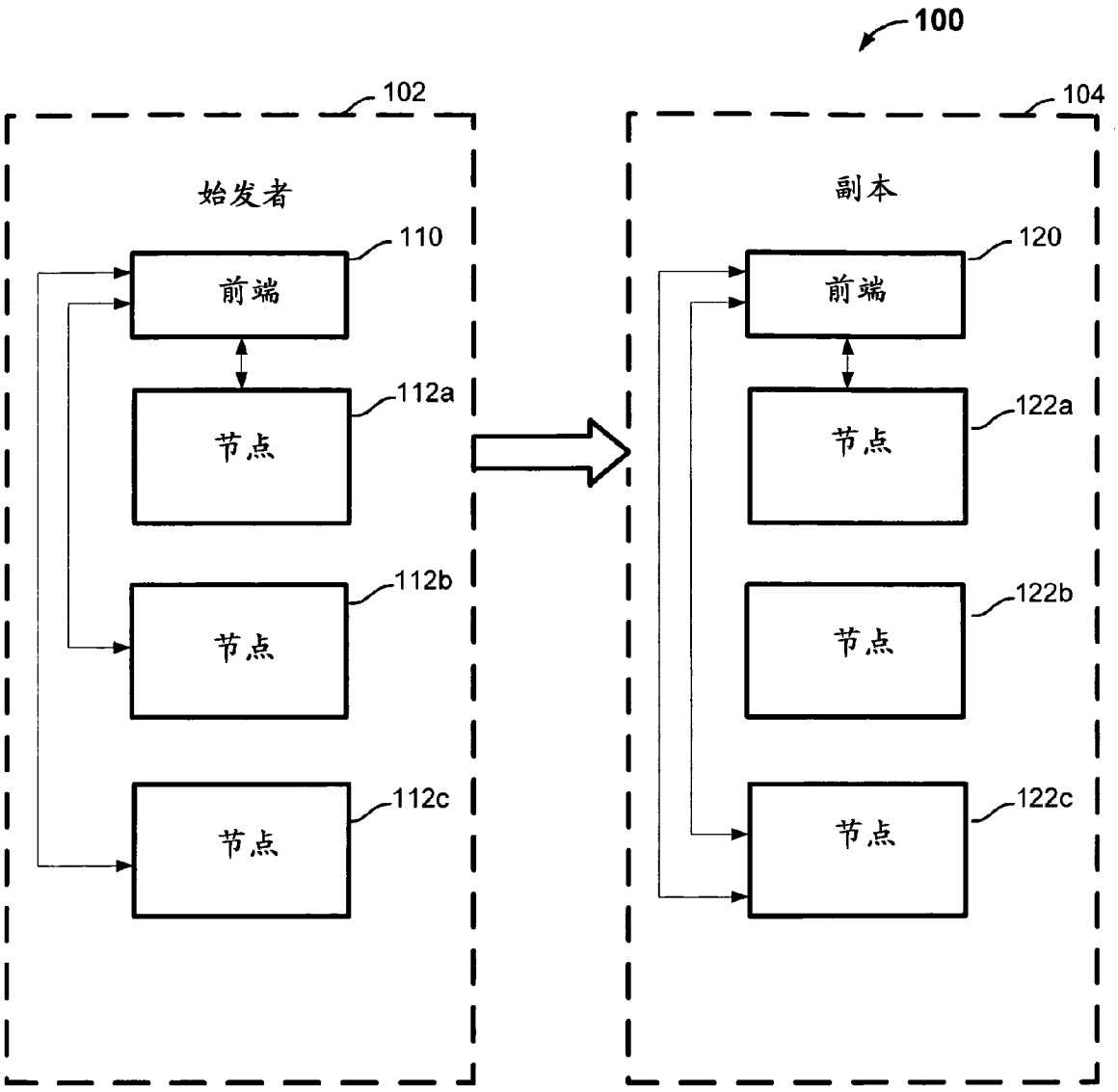

100
102
始发者
110
前端
112a
节点
112b
节点
112c
节点
104
副本
120
前端
122a
节点
122b
节点
122c
节点

图 1

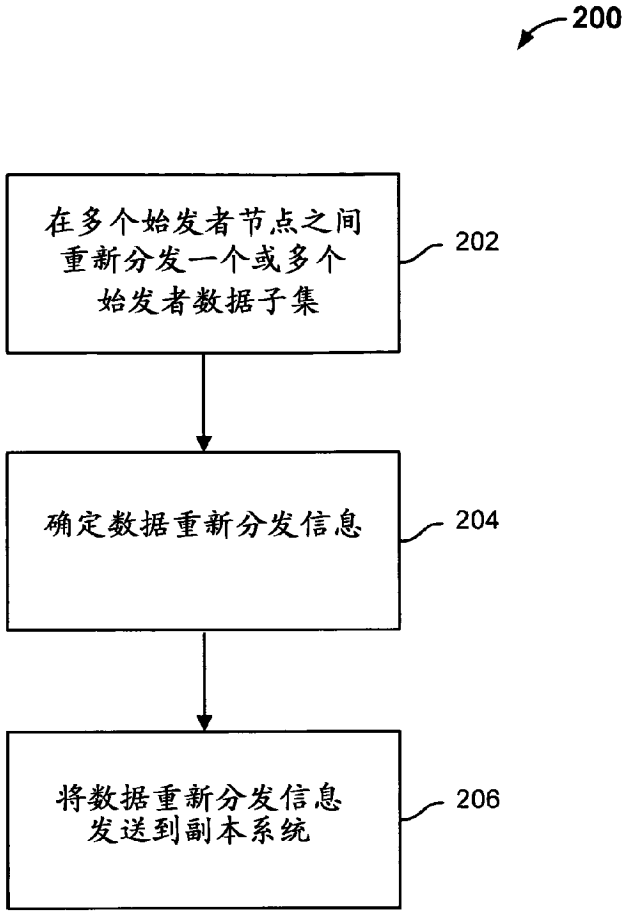

200
在多个始发者节点之间重新分发一个或多个始发者数据子集
202
确定数据重新分发信息
204
将数据重新分发信息发送到副本系统
206

图 2

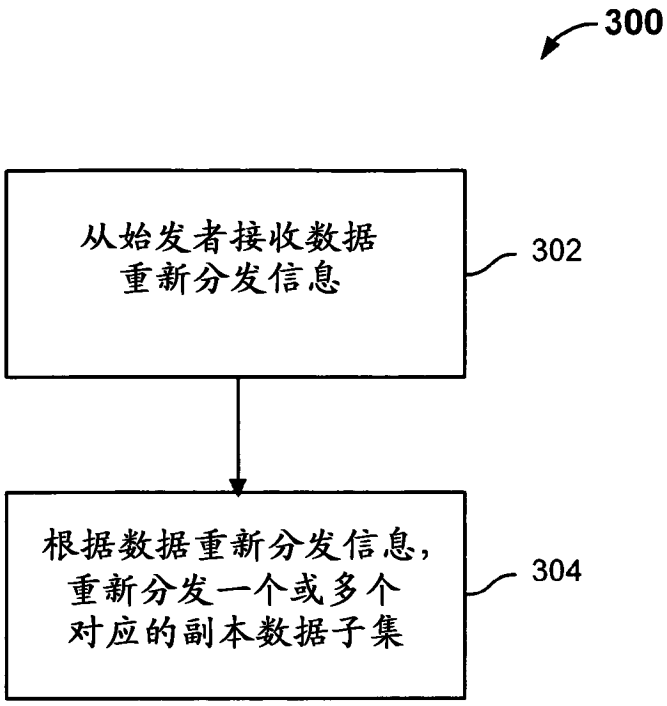

300
从始发者接收数据
重新分发信息
302
根据数据重新分发信息，
重新分发一个或多个
对应的副本数据子集
304

图 3

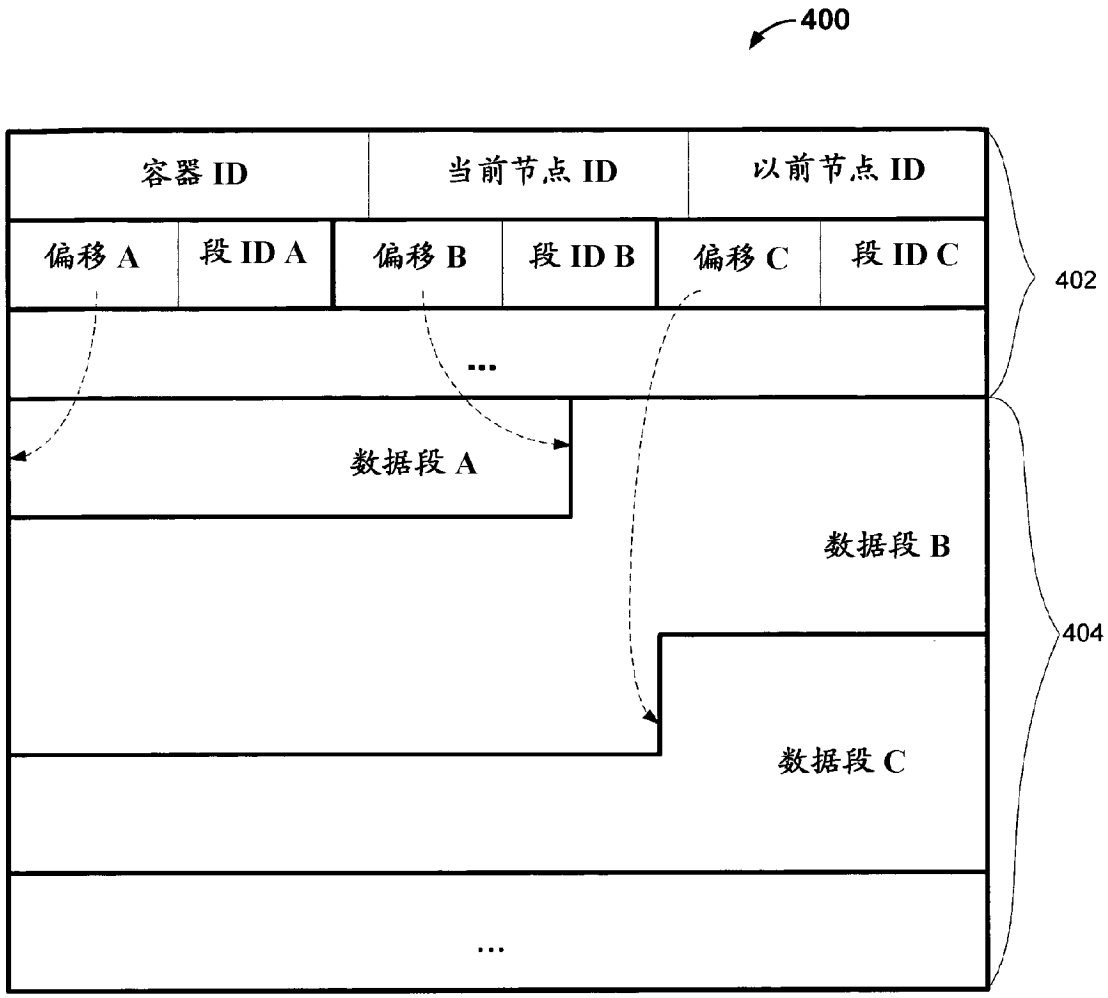

400
容器 ID
当前节点 ID
以前节点 ID
偏移 A
段 ID A
偏移 B
段 ID B
偏移 C
段 ID C
402
...
数据段 A
数据段 B
404
数据段 C
...

图 4

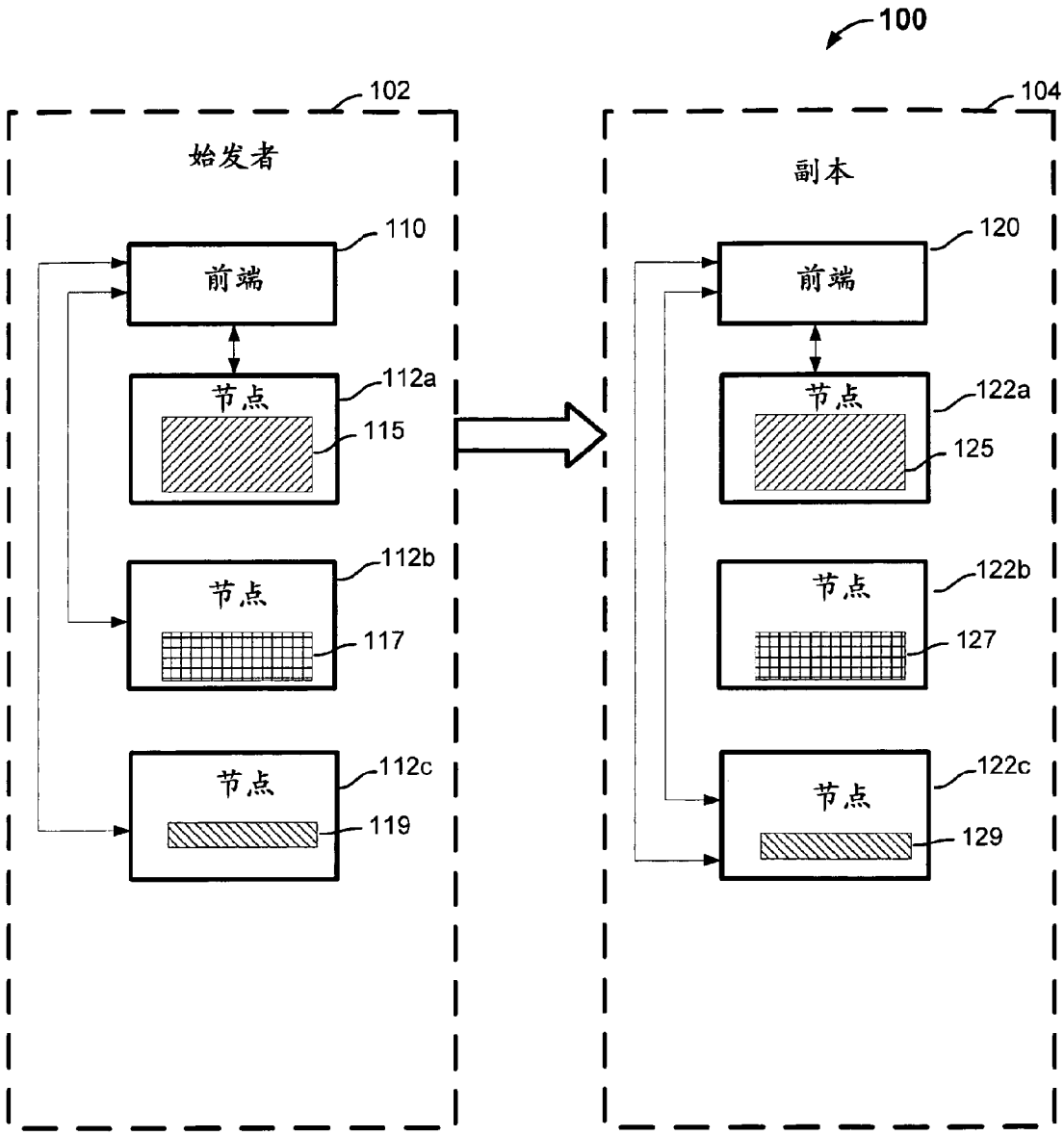

100
102
始发者
前端
110
节点
112a
115
节点
112b
117
节点
112c
119
104
副本
前端
120
节点
122a
125
节点
122b
127
节点
122c
129

图 5A

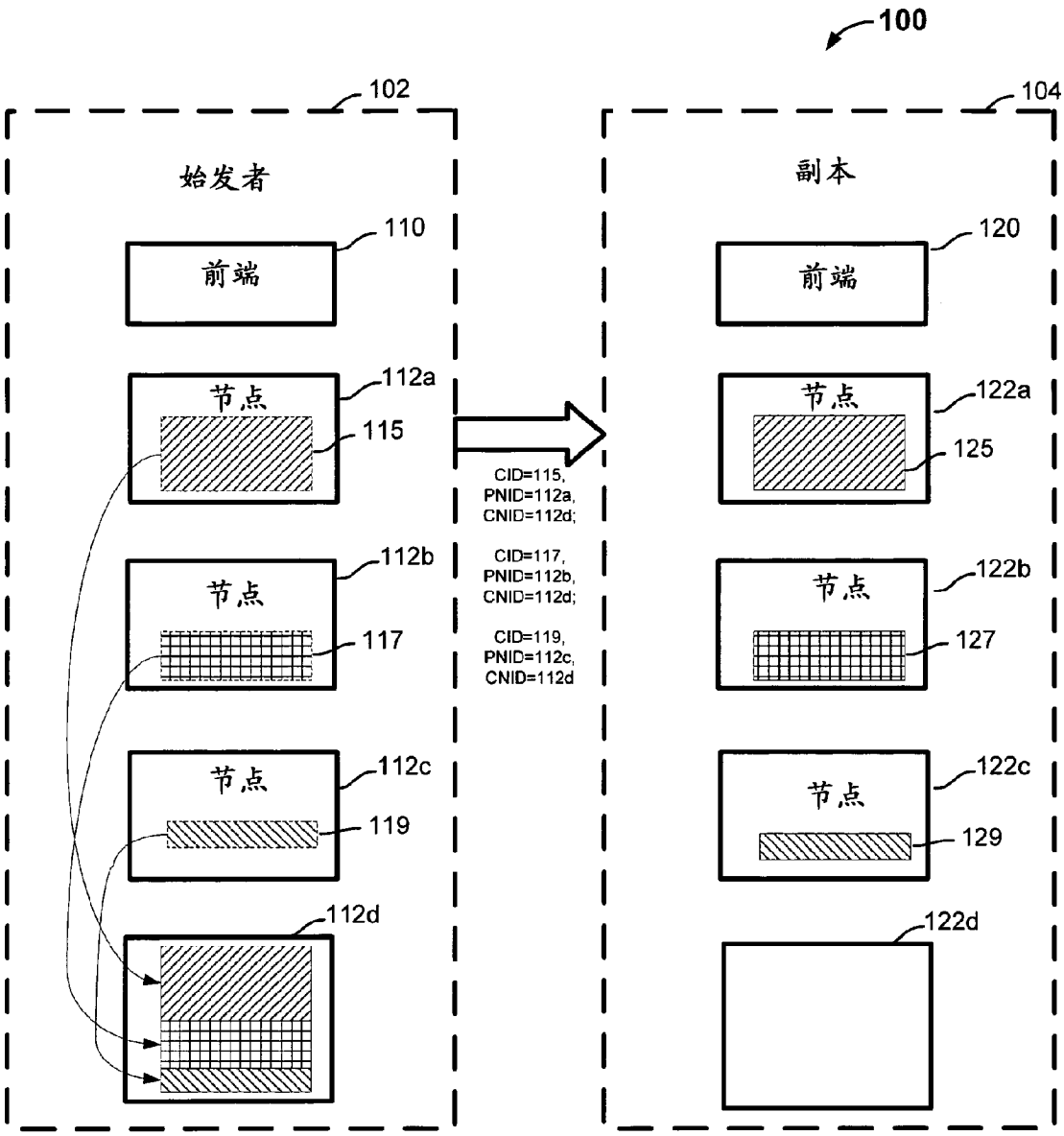

100
102
始发者
前端
110
节点
112a
115
节点
112b
117
节点
112c
119
112d
CID=115,
PNID=112a,
CNID=112d;
CID=117,
PNID=112b,
CNID=112d;
CID=119,
PNID=112c,
CNID=112d
104
副本
前端
120
节点
122a
125
节点
122b
127
节点
122c
129
122d

图 5B

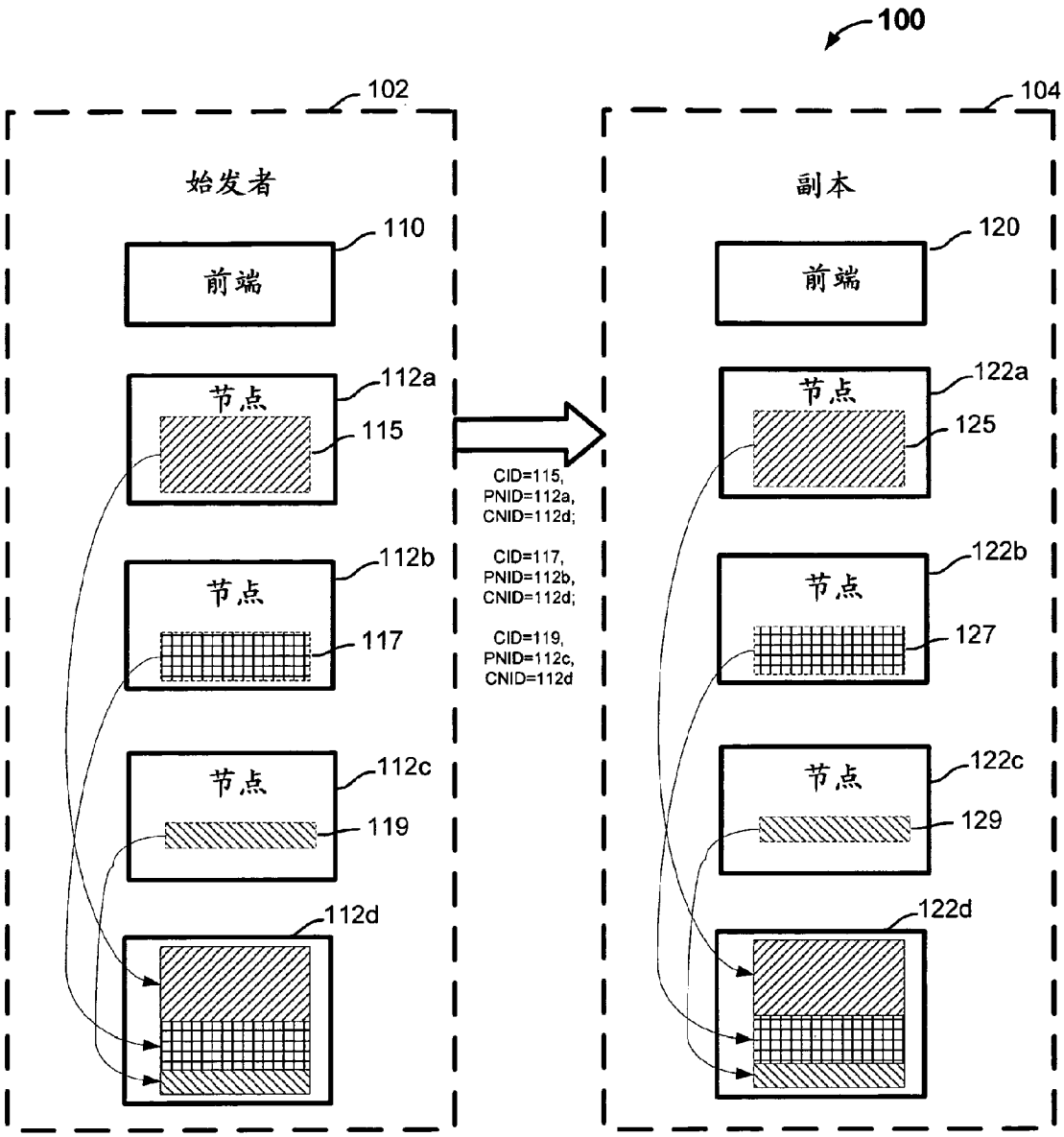

100
102
始发者
110
前端
节点
112a
115
节点
112b
117
节点
112c
119
112d
CID=115,
PNID=112a,
CNID=112d;
CID=117,
PNID=112b,
CNID=112d;
CID=119,
PNID=112c,
CNID=112d
104
副本
120
前端
节点
122a
125
节点
122b
127
节点
122c
129
122d

图 5C