



(12) 发明专利

(10) 授权公告号 CN 102652423 B

(45) 授权公告日 2015.04.01

(21) 申请号 201080055666.7

(56) 对比文件

(22) 申请日 2010.11.16

US 2003/0221074 A1, 2003.11.27,

(30) 优先权数据

US 7461130 B1, 2008.12.02,

12/635,702 2009.12.11 US

US 6718361 B1, 2004.04.06,

(85) PCT国际申请进入国家阶段日

CN 1780420 A, 2006.05.31,

2012.06.08

CN 101355476 A, 2009.01.28,

CN 101005372 A, 2007.07.25,

(86) PCT国际申请的申请数据

审查员 马小瑜

PCT/EP2010/067595 2010.11.16

(87) PCT国际申请的公布数据

W02011/069783 EN 2011.06.16

(73) 专利权人 国际商业机器公司

地址 美国纽约

(72) 发明人 J·斯文格勒 T·W·毕施

R-J·Y·特维托 能田毅

(74) 专利代理机构 北京市中咨律师事务所

11247

代理人 于静 杨晓光

(51) Int. Cl.

H04L 29/08(2006.01)

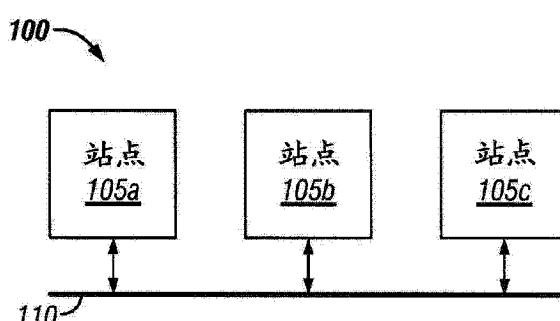
权利要求书3页 说明书16页 附图7页

(54) 发明名称

用于集群选择和协作复制的方法和系统

(57) 摘要

公开了创建用于集群选择和协作复制的集群族的装置、系统和方法。基于它们的关系和角色，将集群分组到集群族的族成员。集群族的成员确定哪个族成员在获得复制信息的最佳位置并且变为与他们的集群族内的累积一致。一旦集群族变为累积一致，在集群族内共享数据使得集群族内所有拷贝是一致的。



1. 一种用于多个集群的协作复制的方法,所述方法包括以下步骤：
将多个集群的至少一个子集安排到集群族的族成员中；
在集群族成员之间进行协商,以确定哪个集群族成员在从族外部的至少一个集群获得至少一个外部数据对象的最佳位置；
选择集群族的一个族成员,以获得外部数据对象；
在集群族内实现与外部数据对象的累积一致性;以及
在集群族成员之间共享外部数据对象,使得集群族内的每个集群与外部数据对象一致。
2. 根据权利要求 1 所述的方法,进一步包括基于集群关系因素和角色因素中的至少一个,在集群族之间创建关系的步骤。
3. 根据权利要求 1 所述的方法,进一步包括以下步骤：
具有一致源的族成员在缓存中保持卷,以使卷可容易地用于对等复制的其它族成员。
4. 根据权利要求 3 所述的方法,进一步包括以下步骤：
所述族成员在缓存中保持所述卷用于其他族成员,使外部集群免除在外部集群的缓存中保持拷贝。
5. 根据权利要求 1 所述的方法,进一步包括以下步骤：
族中 N 个族成员中的每一个复制外部数据对象的 1/N 卷。
6. 根据权利要求 3 至 5 中任意一项所述的方法,进一步包括协作地将所有复制序列化到集群族中的步骤。
7. 根据权利要求 1 所述的方法,进一步包括以下步骤：
复制外部数据对象的 1/N 卷的 N 个族成员中的第一个通知外部集群,第一族成员将保持所述卷用于集群族成员以及免除外部集群在外部集群缓存中保持卷。
8. 根据权利要求 1 至 5 中任意一项所述的方法,进一步包括提供多个集群族的步骤,其中来自每个族的至少一个族成员复制卷。
9. 根据权利要求 1 所述的方法,进一步包括以下步骤 :提供包括多个集群的域,其中集群协作以确保来自每个族的至少一个族成员复制卷并且剩余的集群服从复制要求。
10. 根据权利要求 1 所述的方法,进一步包括以下步骤：
接收拷贝卷到第一集群的拷贝请求；
确定第一集群是否是集群族的族成员；
响应于第一集群是第一族成员,确定指定集群族中的哪一族成员继承拷贝请求；
响应于所述确定,执行拷贝请求并且协作地复制所述卷到集群族中；
在集群族内实现累积一致性;以及
在集群族内共享所述卷,使得所述集群族内的卷的所有拷贝是一致的。
11. 根据权利要求 10 所述的方法,其中,响应于第一集群是第一族成员,确定指定集群族中的哪一族成员继承拷贝请求包括以下步骤：
确定另一族成员是否已经完成拷贝所述卷,
响应于另一族成员还没拷贝所述卷,指定第一族成员继承拷贝请求。
12. 根据权利要求 10 所述的方法,其中,响应于第一集群是第一族成员,确定指定集群族中的哪一族成员继承拷贝请求包括以下步骤：

确定第二族成员是否激活地拷贝所述卷；以及

响应于第二族成员激活地拷贝所述卷，指定第二族成员来继承拷贝请求。

13. 根据权利要求 12 所述的方法，其中，响应于第一集群是第一族成员，确定指定集群族中的哪一族成员继承拷贝请求进一步包括以下步骤：

确定第二族成员是否准备好拷贝，但是此时没有激活地拷贝；

响应于第二族成员准备好拷贝并且此时没有激活地拷贝，降低第二族成员的拷贝优先级且延迟拷贝请求。

14. 根据权利要求 10 所述的方法，进一步包括以下步骤：

响应于确定第二族成员继承拷贝请求，降低第一族成员的拷贝优先级且延迟拷贝请求。

15. 根据权利要求 10 所述的方法，进一步包括以下步骤：

指定第一族成员继承拷贝请求，作为集群族的源集群；

在第一族成员处完成拷贝；

在第一族成员处清除拷贝请求标志；

在第一族成员处为其他族成员设置拷贝请求标志；

其他族成员继承拷贝请求标志；

其他族成员完成拷贝；以及

每个族成员重置拷贝请求标志。

16. 一种用于多个集群的协作复制的系统，所述系统包括：

网络；

通过网络进行通信的多个站点，每个站点包括至少一个主机和存储系统，所述存储系统包括多个集群，每个集群包括被配置为存取在磁带上存储的卷的至少一个磁带驱动器、至少一个磁带卷缓存、以及集群管理器，所述集群管理器包括：

创建模块，用于建立集群族以及将集群分组到集群族的族成员中；以及

协作复制模块，用于选择族成员以协作地复制外部数据对象到集群族中并且实现累积一致性。

17. 根据权利要求 16 所述的系统，其中协作复制模块进一步用于在每个族成员之间共享复制。

18. 根据权利要求 17 所述的系统，其中协作复制模块进一步用于指示族中 N 个族成员中的每一个复制外部数据对象的 1/N 卷。

19. 根据权利要求 16 到 18 中任意一项所述的系统，其中创建模块进一步用于基于集群间关系因素和角色因素中的至少一个，创建集群族。

20. 一种用于多个集群的协作复制的装置，所述装置包括：

创建模块，用于从多个集群创建集群族，其中集群通过网络进行通信并且每个集群包括缓存；以及

协作复制模块，用于协作地复制至少一个外部数据对象到集群族并且实现累积一致性。

21. 根据权利要求 20 所述的装置，其中创建模块进一步用于根据集群关系因素和角色因素中的至少一个建立集群族以及将集群分组到集群族的族成员。

22. 根据权利要求 20 至 21 中任意一项所述的装置，其中协作复制模块进一步用于通过延迟复制来处理持续复制源觉察。
23. 根据权利要求 20 至 21 中任意一项所述的装置，其中协作复制模块进一步用于指示集群族中 N 个族成员中的每一个复制外部数据对象的 1/N 个卷。
24. 根据权利要求 20 至 21 中任意一项所述的装置，其中协作复制模块进一步用于：选择集群族的一个族成员，以获得外部数据对象，并且作为用于集群族内所有族成员的源集群；以及在共享外部数据对象之前，使得源集群负责实现集群族内的累积一致性。
25. 根据权利要求 20 至 21 中任意一项所述的装置，进一步包括：
关系模块，用于保持定义集群族和族成员之间的角色、规则和关系的因素。

用于集群选择和协作复制的方法和系统

技术领域

[0001] 本发明涉及与数据存储系统相关的数据存储，并且更特别地涉及存储系统中的集群。

背景技术

[0002] 存储系统可以包括多个磁带设备，所述磁带设备用于使用库管理器来访问多个磁带。磁带可以被布置在录音带盒中。控制器可以指示传动器将盒式磁带从存储区域转移到磁带驱动器，从而访问在磁带上写入的数据和 / 或将数据写入到磁带中。

[0003] 存储系统可以位于包括多个地理上不同的站点的多个站点。存储系统可以通过一个或多个网络进行通信。每个存储系统可以包括多个集群。每个集群可以包括多个磁带驱动器。将磁带安装到磁带驱动器，从而从磁带读取数据并且将数据写入磁带。

[0004] 可以将每个磁带组织成一个或多个逻辑卷，此处被称为卷。卷可以对主机呈现为不同的存储设备。可以将卷逻辑地“安装”在虚拟磁带驱动器上。如此处所使用的，虚拟磁带驱动器是对主机呈现为磁带驱动器的逻辑构造。

[0005] US 2009/0132657 (Sutani, M R, 等) 公开了分布式结构中跨集群的数据分区，其中缓存节点的动态复制基于伙伴复制的概念。伙伴复制允许由集群内有限数量的节点来复制数据并且提供降低的网络复制业务。

[0006] US 2009/0030986A1 (Bstes, J W) 公开了在复制集群内实现的远程异步数据复制过程，其实现点到点数据复制。通过复制网络上站点之间的双向数据，点到点拓扑允许局部站点的主要存储将数据分布到远程站点。

[0007] 在多集群配置中，每个集群平等地独立于所有其它集群。在没有关系意识的一些分类的情况下，集群不能基于他们角色和 / 或与其它集群的距离按最有效率的方式来操作。在典型的网格配置中，这种对关系的没有意识极大地影响了在安装处理期间选择集群作为卷的源的手段以及集群兑现卷复制的能力。例如，网格可以与城域远程集群相比倾向于选择全球远程源集群用于安装和 / 或拷贝处理。由于集群之间的网络距离，全球远程集群的效率要低得多。尽管可以将实时延迟检查引入以检测所述距离，广域网 (WAN) 的不规则和随机性使得非常难以可靠地测量相对距离。更深一步地，如果组一起工作并且累积地复制数据到组中，并且然后彼此复制，跨组中两个或多个集群距离的自身复制能够更有效率。

[0008] 因此，现有技术中存在解决上述问题的需要。

发明内容

[0009] 提供用于创建集群族、选择集群族成员或多个集群族以及族成员和不同族之间的协作复制的方法、装置和系统。例如，基于集群的关系将集群分组为集群族的族成员。集群族的成员确定哪个族成员处于最好的位置，以获得外部数据对象并且变为与他们的集群族内的外部数据对象累积地一致。一旦集群族变为累积地一致，在集群族内共享数据对象，使

得集群族内所有的集群具有每个外部数据对象的已知拷贝。

[0010] 从一个方面来看，本发明提供一种包括计算机可使用媒介的计算机程序产品，所述计算机可读媒介包括计算机可读程序。当在计算机上执行计算机可读程序时，促使计算机：将多个集群分组到集群族的族成员；确定哪个族成员处于获得来自源的外部数据的最佳位置；选择集群族中的一个或多个族成员以获得数据；将数据复制到集群族；在至少两个外部数据对象的集群族内实现累积一致性；并且共享集群族内的数据，使得集群族内的所有集群具有每个外部数据对象的一致拷贝。

[0011] 从另一方面来看，本发明提供一种用于多个集群的协作复制的方法。所述方法包括将多个集群安排到集群族的族成员中；在族成员之间进行协商，确定哪个族成员位于从源获得数据的最佳位置；选择集群族的一个或多个族成员，以获得数据；协作地将数据复制到集群族中；在集群族内实现累积一致性；并且在集群族内共享数据，使得集群族内的所有数据拷贝是一致的。

[0012] 从另一方面来看，本发明提供一种创建集群族和族成员以执行协作复制的装置。所述装置包括多个模块，被配置为功能性地执行以下步骤：创建集群族和族成员；应用协作复制；以及基于集群关系来选择集群族和族成员。所介绍的实施方式中的这些模块可以包括关系模块、创建模块、协作复制模块、安装处理模块、通信模块以及策略模块或他们的任意组合。

[0013] 关系模块包括在处理器上执行的计算机可读程序并且保持定义角色、规则和集群族和族成员之间的关系的因素。所述集群通过网络进行通信。每个集群包括至少一个缓存，例如，虚拟卷缓存。

[0014] 创建模块包括在处理器上执行的计算机可读程序并且创建用于集群选择和协作复制的集群族。在本发明的优先实施方式中，创建模块通过基于集群的关系和角色将集群分组到族成员中来创建集群族。在可替换的实施方式中，创建模块在配置上将集群分配到族中。在可替换的实施方式中，创建模块创建不同集群之间的关系并且将集群分组到族中。

[0015] 协作复制模块包括在处理器上执行的计算机可读程序，并且跨集群族中集群族成员且跨不同集群族协作地复制数据。

[0016] 安装处理模块在处理器上执行的计算机可读程序，并且支持集群族内的族成员比其它集群族更用于生产目的。

[0017] 从另一方面来看，本发明提供了一种用于多个集群的协作复制的系统。所述系统包括网络；通过网络进行通信的多个站点，每个集群包括至少一个主机和存储系统，所述存储系统包括多个集群，每个集群包括被配置为访问在磁带上存储的卷的至少一个磁带驱动器，至少一个磁带卷缓存，以及集群管理器，被配置为使用处理器和存储器来执行计算机可读程序，其中软件可读程序包括：创建模块，被配置为建立集群组以及将集群组安排到集群族的族成员中；以及协作复制模块，被配置为选择族成员以协作地复制外部数据对象到集群族中并且实现累积族一致性。

附图说明

[0018] 下面将仅通过实例的方式并参考如附图中所示的优选实施方式来介绍本发明，其中：

- [0019] 图 1 是示出了根据本发明的分布式站点的优选实施方式的示意性框图；
- [0020] 图 2A 和 2B 是示出了根据本发明的存储系统的优选实施方式的示意性框图；
- [0021] 图 3 是示出了本发明的集群的优选实施方式的示意性框图；
- [0022] 图 4 是示出了本发明的集群族装置的优选实施方式的示意性框图；
- [0023] 图 5 是示出了本发明的协作复制方法和集群族选择的优选实施方式的示意性流程图；以及
- [0024] 图 6A 和 6B 是示出了本发明的协作复制方法和集群族选择的优选实施方式的示意性流程图。

具体实施方式

[0025] 本说明书中对特征、优点或相似语言的参考并不意味着，可以利用本发明来实现的所有特征和优点应当在或全部在本发明的任意单个实施方式中。而是，将涉及特征和优点的语言理解为表示：结合优选实施方式介绍的特定特征、优点或特性包括在本发明的至少一个实施方式中。因此，整个所述说明书中的特征、优点和相似语言的讨论可以，但不是必须地，指代相同的实施方式。

[0026] 此外，本发明的上述特征、优点和特性可以在一个或多个实施方式中按任意合适方式来结合。所属领域的技术人员将认识到，本发明可以在没有特殊实施方式的一个或多个特定特征或优点的情况下实现。

[0027] 通过参考附图在下面的说明的实施方式中介绍本发明，其中相同的附图标记表示相同或相似的元件。虽然根据用于实现本发明的目标的最佳模式来介绍本发明，所属领域的技术人员应了解的是，在不脱离本发明的范围的情况下根据这些教导可以完成改变。

[0028] 将本说明书中介绍的许多功能单元标记为模块，以更特殊地强调他们的实现方式的独立性。例如，可以将模块实现为硬件电路，包括惯用的超大规模集成电路（VLSI）或门阵列，现成的半导体，例如逻辑芯片、晶体管或其它不相关联组件。还可以在可编程硬件设备中实现模块，可编程硬件设备例如是现场可编程门阵列（FPGA）、可编程阵列逻辑、可编程逻辑设备等。还可以由各种类型的处理器执行的软件中实现模块。可执行模块的识别模块例如包括一个或多个物理或逻辑计算机指令块，所述指令块可以例如被组织成对象、过程或功能。然而，可执行的识别模块不需要在物理位置上处在一起，但是可以包括在不同位置处存储的完全不同的指令，当将上述指令逻辑上连接在一起时，包括模块并且实现用于模块的所表述的目的。

[0029] 实际上，可执行节点的模块可以是单个指令、或多个指令，并且甚至可以在几个不同的代码段上、在不同的程序间、以及跨几个存储器设备上分布。相似地，此处在模块中可以识别和说明可操作的数据。可以将可操作的数据收集作为单个数据集合，或可以在包括不同存储设备的不同位置上分布。

[0030] 本说明书中队“优选实施方式”、“优选实施方式”或相似语言的参考意味着，结合本实施方式介绍的特殊特征、结构或特点包括在本发明的至少一个优选实施方式中。因此，本说明书中出现的短语“在优选实施方式中”、“在优选实施方式中”以及相似语音可以但无须全部指代相同的实施方式。

[0031] 此外，本发明的所述特征、结构或特性可以在一个或多个实施方式中按任意合适

的方式来结合。在下面的说明中，提供大量的具体细节，例如，编程、软件模块、用户选择、网络事务、数据库查询、数据库结构、硬件模块、硬件电路、硬件芯片等的实例，以提供本发明的实施方式的彻底理解。然而，所属领域的技术人员将认识到，可以在不使用一个或多个具体细节、或其它方法、组件、材料等来实现本发明。在其他实例中，没有详细地介绍和示出已知的结构、材料或操作，以避免混淆本发明的方面。

[0032] 图 1 是示出了根据本发明的分布式站点 100 的优选实施方式的示意性框图。分布式站点 100 包括多个站点 105。每个站点 105 可通过网络 110 与其它站点 105 进行通信。网络 110 可以是互联网、局域网 (LAN)、广域网 (WAN)、专用网络、网络的结合等。

[0033] 每个站点 105 可以包括一个或多个存储系统，如此后将介绍的。此外，每个站点 105 可以包括将存储系统连接到网络 110 的网桥、路由器等。

[0034] 图 2A 和 2B 是示出了根据本发明的存储系统 200 的优选实施方式的示意性框图。一个或多个存储系统 200 可以体现在图 1 的每个站点 105 中。

[0035] 存储系统 200 可以将数据存储在不同物理媒介中，包括但不限于存储盒带、磁盘驱动器、固态磁盘 (SSD)、磁盘直接存取存储设备 (DASD)、磁带驱动器、库、以及磁盘驱动器阵列，例如独立磁盘 (RAID) 冗余阵列或磁盘簇。存储盒带的实例是盒式磁带，其包括枢纽卷轴上缠绕的可重写磁带，以及盒带存储器。盒式磁带的一个实例包括基于线性磁带开放 (LTO) 技术的盒带。线性磁带开放 LTO 以及 LTO 标志是 HP、IBM 公司和 Quantum 在美国或其它国家的商标。

[0036] 存储系统 200 可以按不同形式来存储数据，例如逻辑或虚拟数据。此处，可以按各种形式中的任意一种来组织数据，称为“卷”或“对象”，在不参考数据的任意特殊尺寸或安排的情况下选择的数据。

[0037] 如图 2A 和 2B 中所示，存储系统 200 为多个主机系统 210 提供存储器。例如，存储系统 200 包括多个主机 210、多个集群 220、以及网络 215。尽管为了简化的目的，图 2A 中示出了两个 (2) 主机 210a、210b，四个 (4) 集群 220a、220b、220c、220d 以及一个 (1) 网络 215，但是可以使用任意数量的主机 210、集群 220 以及网络 215。因此，存储系统 200 中可以包括任意数量的集群 220。

[0038] 如图 2A 中所示，存储系统 200 可以使用通过网络 215 连接的四个 (4) 集群 220a、220b、220c、220d，每个集群 220 包括用于为主机 210a 仿真磁带驱动器或磁带库的虚拟节点 (“VN”) 260 和存储设备 230。在优选的实施方式中，集群 220a、220b、220c、220d 是虚拟磁带服务器集群。

[0039] 每个集群 220 包括分层存储节点 (“HSN”) 250，用于本地移动和 / 或在存储设备 230 和库 240 之间传递数据。在优选的实施方式中，存储系统 200 包括磁盘存储器 230 和磁盘库 240。在优选的实施方式中，库 240 是自动磁盘库 (“ATL”)。HSN 250 可以用于在本地磁盘存储器 230 和远程磁盘存储器 230 之间远程地传递数据。例如，磁盘存储器 230 可以包括被安排为 RAID、JBOD、SSD 或他们的任意组合的一个或多个磁盘驱动器。

[0040] 每个集群 220 包括如图 3 所示的具有磁带的库管理器 370，将在下面进行介绍。主机 210 可以发起或运行任务或工作，例如磁带工作，其中从集群族 280 和 / 或族成员 220 中的磁带读取数据，并且将数据写入集群族 280 和 / 或族成员 220 中的磁带。主机 210 可以是大型计算机、服务器等。主机 210 可以具有运行或支持多个操作系统的功能。例如，主

机 210 可以运行或可以支持多个操作系统,例如 **Linux®**、**Java®**、**Microsoft®**、**Windows®** 等。Linux 是 Linus Torvalds 在美国和 / 或其它国家的注册商标。Java 和所有基于 Java 的商标和标志是 Oracle 和 / 或其分支机构的商标或注册商标。Microsoft、Windows 以及 Windows 标志是微软公司在美国和 / 或其它国家内的商标。存储系统 200 的主机 210 中的每一个可以用作单个大型计算机,一个或多个服务器、或多个虚拟机。主机 210 可以提供三个级别的虚拟化:通过处理器资源 / 系统管理器 (PR/SM) 工具的逻辑分区 (LPAR);通过 **IBM®z/VM®** 操作系统的虚拟机;以及操作系统,尤其是具有密钥保护的地址空间和面向目标的工作量调度的 **IBM z/OS®**。IBM, z/VM 以及 z/OS 是国际商业机器公司在许多全球管辖区内注册的商标。

[0041] 主机 210 可以通过网络 215 与集群 220 进行通信,以通过下面将介绍的集群族成员 220 访问多个磁带驱动器、磁盘驱动器、或其它存储设备。例如,第一主机 210a 可以在网络 215 上进行通信,以通过第一集群 220a 访问存储设备和磁带。

[0042] 每个集群 220 可以包括分层存储控制器,例如分层存储节点 315,如图 3 所示。集群 220 可以提供用于要被读取和存储的单点管理,聚集了可以容易地将存储器分配给不同主机 210 的存储工具,通过增加存储器或存储器控制节点来扩展存储系统 200,以及用于实现高级功能的平台,例如快写缓存、时间点拷贝、透明数据迁移以及远程拷贝。

[0043] 集群 220 可以遵循“带内”方法。带内方法可以导致通过集群族成员 220 来处理所有的输入 / 输出 (I/O) 请求和所有的管理和配置请求。

[0044] 集群 220 中的每一个可以在他们自己之间连接并且可以通过网络 215 连接到主机 220,以访问在磁带上写入的数据和 / 或将数据写入到磁带中。多个集群 220 可以形成存储系统 200 的域 205。域 205 可以代表多个集群或网格配置。域 205 可以包括两个或多个集群 220。

[0045] 存储系统 200 的网络 215 可以是存储区域网络 (SAN),令牌环网络、局域网 (LAN)、广域网 (WAN)、互联网、专用网络和网络的结合等。SAN 可以包含一种“构造”,主机 210 可以通过所述“构造”在网络 215 上与集群 220 进行通信。构造可以包括光纤通道网络、以太网等。所有的元件不能共享用于通信的相同构造。第一主机 210a 可以通过一种构造与第一集群 220a 进行通信。此外,第一主机 210a 可以通过另一构造与第三集群 220c 进行通信。

[0046] 每个存储系统 200 可以包括集群族 280。集群族 280 可以包括多个集群族成员 220,将所述多个集群族成员 220 安排、配置、组织和 / 或分组到集群族 280 中。例如,如图 2B 中所示,存储系统 200 包括集群族 280(1) 和集群族 280(2)。集群族 280(1) 包括被分组到集群族 280(1) 的族成员中的多个集群 220(a)、220(b)。集群族 280(2) 包括被分组到集群族 280(2) 的族成员中的多个集群族成员 220(b)、220(c)。集群族 280(1) 和集群族 280(2) 经由网络 (例如网络 110、215) 彼此进行通信。可以为每个集群族 280 指定或分配名称。例如,可以将集群族 280(1) 命名为城市 A,并且可以将集群族 280(2) 命名为城市 B。

[0047] 尽管为了简化的目的,图 2B 示出了具有两个集群族 280 的存储系统 200。可以使用任意数量的存储系统 200、集群族 280 以及集群族成员 220。

[0048] 存储系统 200 的实例是 IBM TS7700 虚拟磁带服务器。

[0049] 图 3 是示出了本发明的集群 220 的优选实施方式的示意性框图。例如,集群 220

可以代表图 2A 和 2B 的集群族 280 的集群族成员 220。集群 220 的说明引用图 1 至 2 的元件,相同的数字表示相同的元件。集群 220 可以包括虚拟化节点 310、分层存储节点 315、卷缓存 365 以及库管理器 370。

[0050] 例如,存储设备 230 可以包括被安排为独立磁盘冗余阵列 (RAID) 或磁盘簇 (JBOD)、或固态磁盘 (SSD) 等的一个或多个磁盘驱动器。存储设备 230 可以包括卷存储器 365。卷缓存 365 可以用作虚拟卷缓存和 / 或磁带卷缓存 (TVC)。

[0051] 例如,存储设备 230 包括虚拟卷缓存 365。虚拟卷缓存 365 可以用作 TVC,其中 TVC 包括快速访问存储器设备,例如硬盘驱动器。在优选的实施方式中,集群 220 用于缓存到 TVC 365 的数据。

[0052] TVC 365 可以缓存从逻辑卷读取的数据和 / 或缓存要被写入到逻辑卷的数据。主机 210 可以重复地写入到逻辑卷。TVC 365 可以在硬盘驱动器 230 上存储写入的数据,而不将数据写入到逻辑卷的磁带。在稍后的时间,TVC 365 可以将缓存的数据写入到磁带库 240 内的磁带。因此,可以通过 TVC 365 来路由诸如用于安装逻辑卷的虚拟磁带驱动器的读取操作和写入操作的操作。

[0053] 主机 210 可以发起和运行集群 220 上的任务和 / 或工作。例如,第一主机 210a 访问可能导致库管理器 370 的传动器由物理磁带管理器 335 控制,将盒式磁带从存储区域传递到磁带驱动器,以访问在磁带上写入的数据和 / 或将数据写入磁带和 / 或 TVC 365。

[0054] 虚拟化节点 310 可以是具有到网络 215 的多个连接的独立的基于处理器的服务器。虚拟化节点 310 可以包括电池备份单元 (BBU) 和 / 或可以访问不间断电源 (UPS)。虚拟化节点 310 可以包含看门狗定时器。看门狗定时器可以确保能够重启不能和 / 或花费较长时间来恢复的故障虚拟化节点 310。

[0055] 虚拟化节点 310 可以包括一个或多个磁带后台程序 312。磁带后台程序 312 可以将集群 220 到主机 210 的磁带驱动器仿真为虚拟磁带驱动器。磁带后台程序 312 可以在 TVC 365 上操作文件,和 / 或可以通过远程文件访问在另一集群 220 的远程 TVC 365 中操作文件。

[0056] 分层存储节点 315 可以包括集群管理器 320、远程文件访问 325、数据移动器 330、物理磁带管理器 335、缓存管理器 340、回调管理器 345、数据库 350、管理接口 355 以及媒体管理器 360。集群管理器 320 可以在多个集群或网格拓扑中的多个集群 220 之间协调操作。

[0057] 集群管理器 320 可以使用令牌来确定哪个集群 220 具有数据的当前拷贝。可以将令牌存储在数据库 350 中。集群管理器 320 还可以协调集群 220 之间的拷贝数据。集群管理器 320 可以包括一个或多个处理器,被配置为执行所属领域的技术人员所了解的计算机可读程序。

[0058] 远程文件访问 325 可以是服务器、一个或多个处理器等。远程文件访问 325 可以提供到 TVC 365 的用于由任意远程集群 220 访问的链接。集群管理器 320 可以包括计算机可读程序。

[0059] 数据移动器 330 可以控制用于在集群 220 之间执行的拷贝的实际数据传递操作,并且还可以在物理磁带媒体和 TVC 365 之间传递数据。数据移动器 330 可以包括计算机可读程序。

[0060] 物理磁带管理器 335 可以控制集群 220 中的物理磁带。物理磁带管理器 335 可以

管理多个池中的物理磁带、改造、从共同的暂存池借卷以及将卷返回给暂存池，并且在池之间传递磁带。物理磁带管理器 335 可以包括计算机可读程序。

[0061] 缓存管理器 340 可以控制从 TVC 365 到物理磁带的数据拷贝，以及随后的从 TVC 365 的数据冗余拷贝的移除。缓存管理器 340 可以提供控制信号以不同组件和 TVC 365 之间的数据流。缓存管理器 340 可以计算机可读程序。

[0062] 回调管理器 345 可以对从物理媒体到 TVC 365 的数据的回调进行排队和控制，用于集群管理器 320 所请求的虚拟磁带驱动器或拷贝。回调管理器 345 可以包括计算机可读程序。

[0063] 数据库 350 可以是在硬盘驱动器上存储的记录的结构收集。记录可以包括磁带上的数据的位置。主机 210 可以使用数据库地址将数据写入集群 220 的磁带和 / 或可以从磁带访问数据，以将数据提供给用户。

[0064] 管理接口 355 可以提供与集群 220 相关的信息给用户。同样，管理接口 355 可以允许用户控制和配置集群 220。管理接口 355 可以包括计算机阴极射线管 (CRT)、液晶显示器 (LCD) 屏幕、键盘等，或作为基于网络的接口而存在。

[0065] 媒体管理器 360 可以管理集群 220 的磁带的物理处理。同样，媒体管理器 360 可以管理集群 220 的磁带的错误恢复。媒体管理器 360 可以诊断错误且可以确定错误是否是由物理磁带驱动器或由物理磁带媒体引起的。此外，媒体管理器 360 可以采取用于错误恢复的适当动作。

[0066] 库管理器 370 可以包括多个物理磁带驱动器、机器人存取器以及多个物理磁带媒体。库管理器 370 的机器人存取器可以将磁带传递到被分配给 TVC 365 的磁带驱动器。虚拟磁带驱动器可以是对主机 210 来说是物理磁带驱动器的逻辑结构。如所属领域技术人员公知的，可以通过读取 / 写入通道从磁带驱动器的磁带读取数据，或将数据写入磁带驱动器的磁带。

[0067] 多个集群 220 中的每一个磁带驱动器可以使用一个或多个磁带以存储数据。磁带可以用作存储系统 200 中数据的存储媒体。集群 220 可以使用任意数量的磁带驱动器和磁带。例如，存储系统 200 可以使用两个 (2) 磁带驱动器以及两百五十六 (256) 个虚拟驱动器。

[0068] TVC 365 可以包含来自被操作的磁带卷的数据并且可存储用于快速存取的附加卷数据。可以通过 TVC 365 来路由安装卷的虚拟磁带驱动的诸如读取操作和写入操作的操作。因此，选择集群 220 可以选择集群的 TVC 365。可以将磁带驱动器的所有磁带组织为一个或多个逻辑卷。可以使用先入先出 (FIFO) 和 / 或最近使用 (LRU) 算法来管理 TVC 365 中的卷。

[0069] TVC 365 可以是快速存取存储器设备。例如，TVC 365 可以是具有五千四百吉比特 (5400GB) 存储器容量的硬盘驱动器等。在存储系统 200 中，磁带驱动器可以缓存从逻辑卷读取的去往 TVC 365 的数据和 / 或可以缓存要被写入到逻辑卷的数据。例如，主机 210 可以重复地写入虚拟磁带驱动器。TVC 365 可以在硬盘驱动器上存储写入的数据，而不将数据写入虚拟磁带。在稍后的时间，缓存管理器 340 可以将缓存的数据写入到集群 220 的磁带。

[0070] 可以将存取卷的虚拟化节点 310 称为安装点。选择用于逻辑卷的最近安装点的远程集群 TVC 365 可以改进对卷的存取。TVC 365 的高可使用性、快速写入存储器允许主机

210 将数据写入 TVC 365, 而不必等待要被写入到物理磁盘的数据。

[0071] 在优选的实施方式中, 每个站点 105 包括存储系统 200。每个存储系统 200 包括被分组在一起的两个或多个集群族成员 220, 以创建集群族 280。例如, 集群族 280(1) 包括集群族成员 220(a) 和 220(b) 的组, 并且集群族 280(2) 包括集群族成员 220(c) 和 220(d) 的组。集群族 280(1) 可以用于生产目的, 并且例如, 集群族 280(2) 可以用于灾难 (DR) 或归档的目的。因此, 集群族 280 可以实现与其它集群族 28 相关的不同角色。此外, 集群族 280 的集群族成员 220 可以实现集群族 280 内彼此相关的不同角色。因此, 集群族 280 的集群族成员 220 可以实现与非族成员的不同角色。

[0072] 在优选的实施方式中, 可以在全球距离、都市距离或其组合上配置集群族 280。相似地, 可以在全球距离、都市距离或其组合上配置集群族成员 220。此外, 在集群族 280 中, 集群族成员 220 可以具有彼此不同的远离分级。相似地, 集群族 280 可以具有彼此之间的不同的远离分级。虽然可以将远离分级用作定义角色和集群族 280 和集群族成员 220 之间的关系的因素, 但这仅是带来集群族成员 220 和基站族 280 之间的关系感知的因素。因此, 将集群 220 安排或分组到集群族 280 的集群族成员中并不被限制为距离。

[0073] 此外, 由于每个存储系统 200 包括通过将两个或多个集群 220 分组到族成员中所创建的集群族 280, 每个存储系统 200 或存储系统 200 的结合可以代表多集群配置或网格。

[0074] 此外, 存储系统 200 的集群 220 可以形成分布式存储配置。例如, 第二集群 220(b) 可以创建卷的第二实例。第二实例可以与第一集群 220(a) 上的第一拷贝同步, 其中在更新第一拷贝的任意时间更新第二拷贝。可以在位于远程站点 105 的另一集群族 280 中存储第二实例, 以确保在第一实例变为不可使用的情况下数据可用性。未来的安装点存取可以选择第二拷贝作为第一拷贝。当增加、移除和 / 或重新平衡数据到磁带时, 可以使用透明数据迁移。

[0075] 尽管通过参考图 1 至 2 来讨论本发明的优选实施方式, 但是这仅用于说明的目的。所属领域的技术人员将明白的是, 本发明并不限于任意特定的网格配置并且可以在任意多集群或网格配置中实现。例如, 可以将来自站点 105(a) 的一个或多个集群 220 与来自不同站点 105(站点 105(b)) 的一个或多个集群 220 分为一组, 以创建第一集群族 280。同样地, 可以将来自站点 105(c) 和站点 105(a) 的一个或多个集群 220 分组到族成员中, 以创建第二集群族 280。因此, 可以将集群 220 的任意组合分组到组成员中, 以创建集群族 280。

[0076] 图 4 是示出了本发明的集群族装置 400 的优选实施方式的示意性框图。装置 400 可以体现在主机 210 和 / 或集群 220 中。在优选的实施方式中, 装置 400 体现在集群管理器 320 中。装置 400 的说明指图 1 至 3 的元件, 相同的数字指代相同的元件。装置 400 可以包括关系模块 405、创建模块 410、协作复制模块 415、安装处理模块 420、通信模块 425 和策略模块 430 或其任意组合。

[0077] 关系模块 405 包括在处理器上执行的计算机可读程序, 处理器例如是集群管理器 320 的处理器。此外, 集群关系模块 405 包括定义集群族 280 和族成员 220 之间的关系和角色的因素。例如, 与哪些族成员属于哪些族相关的因素, 相邻族和 / 或组成员之间的距离分级, 以及哪些族成员用于生产目的且哪些族成员用于 DR(灾难恢复) 和 / 或实现目的。

[0078] 集群族成员 220 通过诸如网络 110 和 / 或网络 215 的网络进行通信。每个集群族成员 220 可包括具有至少一个磁带驱动器的库管理器 370, 所述至少一个磁带驱动器被配

置为存取在磁带和至少一个 TVC 365 上的卷。

[0079] 创建模块 410 包括在处理器上执行计算机可读程序,所述处理器例如是集群管理器 320 的处理器。通过将集群 22 分组到一起以通过准则、规则和 / 或目的的共同集合来进行操作,创建模块 410 选择集群 220 并且将集群 220 安排到集群族 280 的族成员中。

[0080] 创建模块 410 将集群 220 分组到集群族 280 中,以允许族成员 220 遵从规则或准则的共同集合。这样允许集群组,例如族 280(1)、280(2) 一起工作以更有效率地完成特殊任务,或允许不同组的集群 220 和 / 或族 280 具有网格内的不同目的。

[0081] 创建模块 410 可以被用于通过配置属性允许族 280 内族成员 220 的可定制行为。例如,参考图 2B,可以允许集群族成员 220(a)、220(b) 的组充当遵从对生产工作量有益的规则集合的生产族 280(1)。可以允许域 205 中的另一组集群族成员 220(c)、220(d) 充当归档或灾难恢复族 280(2),具有使族成员 220(c)、220(d) 在从生产族 280(1) 复制数据时更有效地进行操作的规则。

[0082] 此外,创建模块 410 管理族 280 的族成员 220 的关系以及不同集群族 280 之间的关系。例如,创建模块 410 可以基于集群族成员的关系和角色来管理集群族成员 220。在优选的实施方式中,关系模块 405 可提供这种信息给创建模块 410。基于族成员和相邻族的关系和 / 或角色,集群族成员 220 将在彼此间进行协调以确定,哪个族成员 220 在从族 280 外部的多个集群获得外部数据的最佳位置。创建模块 410 还可使用这个信息来支持族 280 的成员 220 作为 TVC 集群,或允许相对于仅集群或全网格的族 280 上的存取限制或其它特殊情况行为。

[0083] 创建模块 410 可使用管理接口 355 来显示页面,其中用户(例如,客户)可以创建具有字符名称的集群族,例如 8 个字符名称。然后,用户可以使用创建模块 410 增加一个或多个集群到族。创建模块 410 可以在集群持续的重要产品数据内存储所述信息,使得多个集群或网格配置中的所有集群知道他们集群的角色以及其所驻留的族。创建模块 410 可以确定正在选择用于族的集群已经被选择用于另一族。为了避免使任意一个集群在相同时间出现在两个族中,创建模块 410 可以通知用户被选择的集群已经存在于另一族成员中。此外,创建模块 410 可以使用规则集合来阻止同一时间将一个集群选择到两个族中。

[0084] 策略模块 430 包括在处理器(诸如集群管理器 320 的处理器)上执行的计算机可读程序。在优选的实施方式中,策略模块 430 可包括与应当将哪个集群族成员 220 用于生产以及应当将哪个族成员用于 DR/ 归档目的有关的特定策略。这些策略可以包括管理数据复制的数据集合。用户可以经由管理接口 355 输入用于管理多个集群族 280 和族成员 220 的策略。

[0085] 参考图 2A 和 2B,集群族创建模块 410 可以用于创建名称为“城市 A”的集群族 280(1) 以及创建名称为“城市 B”的集群族 280(2)。集群族 280(1) 可以包括集群族成员 220(a)、220(b) 的组,并且集群族 280(2) 可以包括集群族成员 220(c)、220(d) 的组。此外,创建模块 410 可以用于增加族成员 220 到集群族 280 或从集群族 280 移除族成员 220,以及将集群族成员重新分组到不同的集群族 280 中。

[0086] 由于创建模块 410 基于族的创建期间集群彼此间的关系和 / 或角色建立集群并将集群安排到族组中,网格或多集群配置中的所有集群直到彼此的角色以及他们所驻留的族。因此,创建模块 410 可通过管理接口 355 警告或通知用户,被增加到族 280(1) 的集群

220(d) 当前例如是另一族 280(2) 的族成员。然后, 用户可以从族 280(2) 取消选择 220(d), 并且为族 280(1) 增加或重新选择 220(d)。因此, 创建模块 410 允许域 205(例如, 网格) 中的所有集群 220 直到他们自己的角色以及与他们所驻留的族的关系、与其它族成员的关系以及与驻留在其它族中的非族成员的关系。

[0087] 在优选的实施方式中, 创建模块 410 可为集群族指派名称。例如, 在配置期间, 用户可使用管理接口 355 为集群族指派名称。

[0088] 协作复制模块 415 包括在处理器上执行的计算机可读程序, 所述处理器例如是集群管理器 320 的处理器。此外, 协作复制模块 415 加强现有拷贝的管理以使属于集群族 280 的集群 220 的组能够更有效率地一起工作, 以实现族 280 以及族 280 内(例如, 族成员 220) 各个集群 220 之间的一致性。

[0089] 协作复制模块 415 允许族 280(例如, DR 或归档族) 内两个或多个集群族成员 220 共享入境复制工作量。因此, 当为复制选择了源集群时, 使用协作复制模块 415 的 DR/ 归档集群族成员 220 的族 280 能够从改进的 TVC 选择获益。

[0090] 协作复制模块 415 允许集群族成员在属于相同族的其它集群族成员之间共享拷贝工作量。例如, 在优选的实施方式中, 域 205 包括 Y 个集群 220, 其中 Y 代表域 205 中包括的集群 220 的数量。将集群分组到具有 N(两个或多个) 集群族成员 220 的集群族 280 中。因此, 域 205 由 Y 个集群 220 组成, 其中将集群 220 中的一些分组到集群族 280 的 N 个集群族成员中。

[0091] 例如, 参照图 2B, 域 205 中存在四个集群 220(a)、220(b)、220(c)、和 220(d), 因此 Y 代表 4 个集群 ($Y = 4$)。将两个集群 220(a) 和 220(b) 分组到第一集群族 280(1) 的 $N = 2$ 的集群族成员中, 并且将两个集群 220(c) 和 220(d) 分组到第一集群族 280(2) 的 $N = 2$ 的集群族成员中。在这种网格配置中, 域 205 由 Y(4) 个集群组成, 其中将 $N = 2$ 的集群的子集分组到集群族 280 的族成员中。因此, $N = 2$, 作为族中集群族成员的数量。

[0092] 通过当第一次将集群带入到族中时序列化任意一个卷的复制, 协作复制模块 45 协作地复制基站的族组。例如, 协作复制模块 415 指示族 280(2) 中的每个集群成员复制外部卷中的 $1/N$, 其中 N 是需要拷贝的族中的集群数量。一旦将所有的外部卷复制到族 280(2) 中且族 280(2) 是累积一致的, 于是相同族 280(2) 内不一致的集群在每一个间共享外部数据。

[0093] 例如, 在没有本发明的情况下, 如果可能, 从微代码的级别, 由于集群 220 不知道彼此之间的关系和角色, 每个集群 220 彼此独立的工作。例如, 如果我们假设集群 220(a) 包括需要被复制到集群 220(c) 和 220(d) 的 20 个卷。由于集群 220(c) 和 220(d) 彼此独立地工作, 集群 220(c)、220(d) 中的每一个可以在网络 215 上拉取 20 卷的原始数据。

[0094] 现在参照图 2A 和 2B, 在优选的实施方式中, 例如, 存在四个集群, 其中通过创建模块 410 将两个集群 220(a)、220(b) 分组到族 280(1) 中并且将两个集群 220(c)、220(d) 分组到族 280(2) 中。所有的族成员 220 知道彼此并且知道他们所属的族 280, 并且知道需要被复制到族中相邻集群的所有卷。

[0095] 例如, 集群族成员 220(c) 和 220(d) 知道彼此并且存在来自不同集群族 280(1) 内非族成员 220(a) 的需要被复制到它们的族 280(2) 的 20 个卷。使用协作复制模块 415, 族成员 220(c) 拉取 10 个唯一的卷并且族成员 220(d) 拉取其它 10 个唯一的卷。即, 每个集

群族成员 220(c)、220(d) 拉取卷的 $1/N$, 其中 $N = \text{族中集群族成员的数量}$ 。由于在这个实例中存在属于族集群 280(2) 的两个集群族成员 220(c)、220(d), 每个族成员拉取 $1/2$ 的卷 (例如, 每个拉取 10 个唯一的卷) 以获得整体 20 个卷。然后, 集群族成员 220(c)、220(d) 彼此共享 10 个唯一的卷。

[0096] 通过经由协作复制模块 415 协作地复制, 由于任意一个卷仅通过远程链路 110/235 拉取一次而不是 N 次, 集群族 280(2) 或 DR 位置可以变得累积地一致更快 N 倍。于是, 由于它们之间的相对距离, 集群族成员 220 可更快地变得彼此间可用性一致。因此, 通过每个集群 220 独立地源自从相同远程生产集群 220, 变得 DR 一致和高度可用 (HA) 一致的整体时间可以极大地增强。

[0097] 因此, 可以优化拷贝吞吐量以及提升整体时间, 以实现集群族 280 内的卷一致性。例如, 在受限带宽系统或具有多个存档站点的网格中, 协作复制模块 415 允许族 280 中的每个族成员 220 参与到用于所有入境拷贝的复制过程, 而不复制任何努力。一旦族 280 内集群 220 的组 (族成员) 达到聚集一致状态时, 在相同族内的对等集群间共享族 280 内各个集群 220 中的一致性拷贝。

[0098] 此外, 协作复制模块 415 通过推迟复制来处理持续复制源觉察。例如, 可以通知具有一致源的集群成员 220 以保持缓存中所述源卷, 从而使其容易地可用于对等复制的其它族成员 220。具有一致性源的集群继承原始安装源集群或包含主机创建 / 修改的原始拷贝的集群的角色。一旦族中的一个集群复制其 $1/N$ 卷中的一个, 协作复制模块 415 首先通知原始安装源集群, 说明其族中所有其它集群 (包括其自己) 的理由并且生产集群可以解除代表目标族中集群的其自己的角色。这样免除生产集群以组织后端磁带外的卷 (假设没有其它族或需要拷贝的生产集群), 因此提供更多的缓存可用性。第二, 发起用于卷的复制的 DR 族集群继承角色并且想起其族内的哪些集群仍需要拷贝。通过这种继承, 可以在缓存中支持卷, 直到所有的其对等族集群完成了拷贝。

[0099] 在优选的实施方式中, 协作复制模块 415 可以使用级联拷贝需求标志。例如, 随着集群族变为一致的, 协作复制模块 415 将拷贝需求标志的所有权从一个集群族移动到另一个。通过级联拷贝需求标志, 协作复制模块 415 可以允许标志的益处从一个族移动到另一个族, 因此释放其参与的原始 TVC。通过从 TVC 继承拷贝需求标志, 例如, 一旦族成员获得拷贝, 其可以允许 TVC 集群迁移卷并且在缓存中为其他新工作量分配空间。

[0100] 一个实例可以是包括产品或与 DR/ 归档族相连的默认族的域。TVC 集群可以是生产或默认族的成员, 并且可以开始管理拷贝需求标志。一旦 DR/ 归档族的成员从 TVC 集群获得了拷贝, DR/ 归档族可以通知 TVC 集群清除与 DR/ 归档族的成员相关的所有拷贝需求标志。与此相结合, DR/ 归档族可以继承管理用于其族成员的这些拷贝需求标志的责任。

[0101] 例如, 在存储系统 200 (未示出) 的另一实施方式中, 域 205 可以包括: 第一族集群 280(1), 其包括集群族成员 220(a)、220(b)、220(c); 第二族集群 280(2), 其包括集群族成员 220(d)、220(e)、220(f); 以及第三族集群 280(3), 其包括集群族成员 220(g)、220(h)、220(i)。每个族 280 包括三个集群族成员 220 并且每个族成员代表一个比特。由于存在三个族且每个族具有三个族成员 (3 比特), 比特集合中总共有 9 个比特。例如, 族集群 280(1) 包括需要被复制到族集群 280(2)、280(3) 中的原始数据对象。

[0102] 可能的是, 集群 220(a) 可以持有缓存中的卷, 直到所有九个集群 220 在网络 110

或 215 之间拉取了拷贝为止。例如，集群 220(a) 可以包括 9 比特集合并且，当每个集群 220 拉取拷贝时，集群 220(a) 可清除其掩码中的比特。由于集群 220(a) 在其缓存中持有用于所有九个集群 220 的拷贝，集群 220(a) 不能为附加工作量分配空间。

[0103] 通过允许每个集群族 280 继承管理用于其族成员的这些拷贝需求标志的责任，集群 220(a) 可以清除用于这些集群族 280(2)、280(3) 的剩余 6 个比特，并且在其缓存中仅保持用于其自己的驻留在族 280(1) 中的两个族成员 220(b)、220(c) 的拷贝。一旦其自己的族成员 220(b) 和 220(c) 具有拷贝，集群 220(a) 于是可清除其掩码以为更多的工作量分配空间。

[0104] 在本实例中，族 280(2) 的集群 220(d) 在网络 215 上拉取拷贝并且通知族集群 280(1) 的集群 220(a) 由于 220(d) 将在其缓存中保持拷贝直到其族成员 220(e)、220(f) 接收了拷贝位置，则其不再需要在缓存中持有用于族 280(2) 的拷贝。这样使集群 220(a) 无需在缓存中持有用于集群族 280(2) 的所有集群成员的拷贝。相似地，属于族 280(3) 的族成员 220(g) 指示集群 220(a)，将在其缓存中保持其族成员 220(h)、220(i) 的拷贝。因此，族 220(a) 免除在其缓存中持有用于属于 280(3) 的所有族成员的拷贝。

[0105] 此外，协作复制模块 415 可以通过在域中使用更多链接来执行拷贝而不是主要依赖于来自 TVC 的拷贝，在低带宽环境中增加性能，并且改进了用于族内集群变为一致的整体时间。例如，使用协作复制模块 415 的族 280 合作，从而实现跨族的一致性。在接收到全族的一致性时，族成员 220 于是可以一起工作以在族成员之间共享数据，以将每个各自成员带至族的一致性级别。

[0106] 安装处理模块 420 包括在处理器上执行的计算机可读程序，处理器例如是集群管理器 320 的处理器。当利用集群的逻辑卷发生安装时，安装处理模块 420 支持和选择在其族外部的集群上的其自己的族内的集群族成员。例如，对产品集群的安装可通过主要用于 DR 或电子跳马远程集群来支持相同族 280(1) 中的另一生产集群。当生产数据需要保持局部且快速复制的高度可用性，安装处理模块 420 可被用于支持通过灾难恢复的可用性，并且因此通过 DR 族在生产族内选择族成员。

[0107] 安装处理模块 420 可以通过在需要远程安装时支持集群族成员来改进控制和性能。可以配置族和 / 或族成员（例如，使用创建模块 410）以在选择远程 TVC 时相比于其它基站更喜欢特定集群。这样在从非生产集群区别产品集群集合时是有益处的。优选相同产品族内的族成员可以在生产集群内保持 TVC 选择，而不是潜在地选择旨在 DR 或归档目的的距离上远程的集群。

[0108] 此外，由于集群族 280 内的集群族成员 220 是安装的目标并且在相同集群族 280 中得到支持，可以改进 TVC 选择处理。

[0109] 在优选的实施方式，存储系统 200 可以包括多个集群，其中将两个或多个集群 220 的子集分组到第一集群族 280 中，并且将两个或多个集群 220 的子集分组到第二集群族 280 中。族组的分组可以基于族成员的角色、关系和 / 或彼此间和 / 或与其它非族成员集群之间的距离。集群族 280 的每个集群族成员知道他们彼此之间的关系。族成员之间的这种关系意识允许组一起有效率地工作，以累积地复制数据到组中并且然后在彼此间进行复制。

[0110] 站点 105 可以包括集群族 280 或集群族 280 的组合。例如，站点 105(a) 可以包括第一集群族 280 和第二集群族 280。第一集群族 280 可以包括生产集群 220(a)、220(b)，并

且第二集群族 280 可以包括生产集群 220(c)、220(d)。此外,可以从站点 105 的结合中选择集群 220 以创建集群族 280。例如,可以通过选择不同站点 105(例如,105(a) 和 105(b) 处的集群 220 来创建集群族 280,其中站点 105(a) 处集群 220(a)、220(b) 用于生产目的并且站点 105(b) 处集群 220(c)、220(d) 用于 DR 和 / 或归档目的。

[0111] 在优选的实施方式中,集群 220(c)、220(d) 用于归档数据。在优选的实施方式中,集群 220(c)、220(d) 用于 DR。在另一实施方式中,诸如集群 220(c) 的一个集群用于 DR,并且诸如集群 220(d) 的另一集群用于归档。

[0112] 一般地,下面的示例性流程图详述了逻辑流程图。这样,所描述的顺序和标记的步骤指示本方法的优选实施方式。可以将功能上、逻辑上或效果上等价的其它步骤和方法设想为所述方法的一个或多个步骤,或其部分。此外,提供所使用的格式和符号以解释所述方法的逻辑步骤,并且将所使用的格式和符号理解为不限制本方法的范围。尽管可以在流程图使用各种箭头类型和线类型,不将他们理解为限制相应方法的范围。实际上,一些箭头和其它连接器可以用于仅指示方法的逻辑流。例如,箭头可以指示所描述的方法的列举的步骤之间非特定持续的等待或监控时间段。此外,特定方法发生的顺序可以限制为附着的相应示出步骤的顺序,或不限制为附着的相应示出步骤的顺序。

[0113] 图 5 是示出了本发明的集群族选择和协作复制方法的优选实施方式的示意性流程图。方法 500 实质上包括执行上面与图 1 至 4 的介绍的装置和方法的操作相关呈现的共更能的步骤。在优选的实施方式中,利用计算机程序产品来实现所述方法,所述计算机程序产品包括具有计算机可读程序的计算机可读媒介。可以将计算机可读程序集成到计算机系统中,所述计算机系统例如集群管理器 320 和 / 或主机 210,其中与计算系统相结合的程序能够执行所述方法 500。

[0114] 方法 500 开始并且在步骤 510,将一组集群安排到集群族的族成员中。例如,基于集群间彼此的关系以及与域中其它集群的关系来对集群进行分组。基于各种因素和 / 或功能来创建集群族,所述功能包括角色(例如,生产源、DR、归档等)、范围、距离(例如,族之间的距离比率)等。此外,用户可以将字符名称指派给集群族。例如,如图 2B 中所示,可以使用字符名称(“城市 A”)来创建集群族并且可以使用字符名称“城市 B”来创建另一族。

[0115] 在优选的实施方式中,创建模块 410 用于创建集群族,其中用户可以使用管理接口 355 在配置期间创建集群族,以创建集群族名称、增加一个或多个集群到族、在相邻族间分配角色和 / 或距离比率、以及使用配置属性来教导集群。例如,创建模块 410 可以使用这些持续设置以为一个或多个集群或族成员带来关系意识以及带来族之间的相关属性,例如距离。

[0116] 在优选的实施方式中,关系模块 405 保持用于集群族和族成员的这些持续设置。

[0117] 此外,可以使用自主功能来检测角色和集群间的关系。例如,在创建模块 410 中执行自主功能簇。

[0118] 在步骤 515 中,族成员在彼此间进行协商,以确定族的哪个族成员位于获得外部数据对象的最佳位置。例如,如图 2B 所示,集群族 280(1) 包括两个或多个族成员 220(a)、220(b),可以在城市距离上配置集群族并且集群族可以用于生产目的。集群族 280(2) 包括两个或多个族成员 220(c)、220(d),可以在全球距离上参照族 280(1) 来配置集群族并且集群族可以用于 DR 目的。集群族 280(1) 可以经由准备好拷贝数据对象的网络 110、215 与集

群族 280(2) 进行通信。由于每个族的集群成员以及集群自身知道其它每个集群的角色和彼此间的关系, 族成员 220(c)、220(d) 可以在彼此间进行协商以确定族 280(2) 的哪一个族成员位于获得数据对象外部的拷贝的最佳位置。

[0119] 在优选的实施方式中, 例如, 属于集群族 280 的集群族成员 220 使用共同拷贝工作队列按 FIFO 顺序工作。在拷贝的工作之前, 每个集群族成员 220 首先确保集群族 280 中没有其它集群族成员已经在拷贝或已经拷贝了。如果不是, 一个或多个集群族成员 220 可执行拷贝。如果另一族成员已经正在发生拷贝或另一族成员已经发生拷贝, 一个或多个集群族成员可以将拷贝移动到延迟队列。一些时间之后, 毕竟已经将激活的生产内容拷贝到集群族 280 中, 族成员开始在他们应当彼此共享内容的延迟队列上进行工作。如果原始获得拷贝的对等族成员不在, 那么其仍然可从外部集群或另一族成员获得拷贝。

[0120] 在步骤 520, 一个或多个集群族成员获得和复制信息或源卷。例如, 选择属于集群族 280 的一个或多个集群族成员 220 以在远程网络 110、215 上拉取数据或源卷, 拷贝 / 复制数据或源卷、并且将其带入到集群族 280 中。例如, 族 280(2) 的族成员 220(c) 将外部数据对象通过网络 110、215 拉到族 280(2)。族成员 220(c) 现在具有一致的源并且可以要求族成员 220(c) 在缓存 (例如, TVC 365) 中保持源卷, 以使其容易地可用于对等复制。

[0121] 在步骤 525 中, 在族的族成员之间协作地复制源卷。例如, 当第一次将集群带入到族中, 集群的族组通过序列化任意一个卷的复制来进行协作。族中的每个集群均可用作复制 1/N 卷的角色, 其中 N 是族内需要拷贝的集群的数量。

[0122] 通过协作地复制, 由于仅跨远距离链路拉取任意一个卷一次而不是多次, 集群族或 DR 位置可以 N 倍更快地变为累积一致。然后, 由于它们之间的相对距离, 集群可以更快地变为用于可用性的彼此间一致。变为 DR 一致和 HA (高可用性) 一致的整体时间可在每个集群上极大地增加, 而与相同远程生产集群的源相独立。

[0123] 在步骤 530, 集群族实现累积一致性。即, 完成需要被复制到集群族的集群族外部的所有卷。集群族作为整体与所有外部数据对象一致。现在, 集群族成员可以在彼此间共享, 使得集群族内每个单独族成员具有其自己的拷贝。

[0124] 在步骤 535, 在将所有的卷复制到族中并且族是累积一致的之后, 相同族内不一致的集群于是彼此间共享卷 (即, 数据对象)。

[0125] 因此, 本发明的实现方法 500 协作地执行复制, 其中由于仅跨远距离链路拉取任意一个卷一次而不是多次, 集群族或 DR 位置可以 N 倍更快地变为累积一致。然后, 由于它们之间的相对距离, 集群可以更快地变为用于可用性的彼此间一致。变为 DR 一致和 HA (高可用性) 一致的整体时间可在每个集群上极大地增加, 而与相同远程生产集群的源相独立。

[0126] 此外, 当客户仅需要 N 个拷贝以及这些 N 个拷贝彼此远离时, 方法 500 使用复制到 X 个集群 (其中, X 代表集群的数量) 的更有效方法的族。这样允许客户在距离 / 族上传播拷贝, 而不明确集群接收拷贝。例如, 用户可能不关心哪个集群包含拷贝, 只要存在 N 个拷贝 (其中 N 小于 X); 并且客户要求 N 个拷贝全部存在于独立的族中。因此, 域中的所有基站可以协作, 以确保来自每个族的至少一个成员复制卷, 并且于是剩余的集群可满足其复制要求。如是, 可以在 N 个族中以 N 个拷贝作为结束, 而不会在任意一个区域中具有 N 个拷贝中的过多拷贝。

[0127] 可以在安装处理中按任意结合来使用方法 500 的步骤。例如, 通过在步骤 510 中

配置的集群族，使用步骤 515 到 535，方法 500 可以支持在其族外部的集群上其自己族的集群。例如，对生产集群的安装可以支持远程集群上的另一生产集群（在系统族中），所述远程集群最初用于灾难恢复（电子跳马）。由于用户可能趋于想要产品数据，以保持本地的和快速复制的高可用性（支持通过灾难恢复的可用性），根据短期目标且仍然不影响长期目标，获得产品集群更为有效。

[0128] 参考图 6A 和 6B，是示出了本发明的集群族选择和协作复制方法的优选实施方式的示意性流程图。方法 600 实质上包括执行上面关于图 1 至 4 介绍的装置和系统的操作呈现的功能的步骤。在优选的实施方式中，利用计算机程序产品来实现所述方法，所述计算机程序产品包括具有计算机可读程序的计算机可读媒介。可以将计算机可读程序集成到计算机系统，例如集成管理器 320 和 / 或主机 210，其中与计算系统结合的程序能够执行方法 600。

[0129] 方法 600 开始且在步骤 605，拷贝过程开始。例如，城市 A 中的外部数据对象需要在城市 B 中复制（例如，图 2B）。

[0130] 在步骤 610，控制确定接收拷贝请求的集群是否是集群族成员。如果不是，在步骤 615，拷贝卷，且不执行协作复制。例如，协作复制模块 415 可以管理拷贝请求，而没有延迟或优先级改变。

[0131] 此外，协作复制模块 415 可以在族中选择至少一个族成员，以跨遥远的链路或网络拉取数据。一旦确定了集群是族成员且没有其它族成员跨网路拉取数据，可以执行选择。

[0132] 如果这是集群族成员，在步骤 620，控制确定其它族成员中的一个是否已经完成了拷贝所述卷。如果是，在步骤 625，由于其他族成员中的一个已经拷贝了卷，为拷贝卷提供较低优先级并且将卷拷贝回队列。

[0133] 如果族成员中的一个还没有完成对所述卷的拷贝，在步骤 630，控制确定另一族成员是否激活地包括所述卷。如果是，在步骤 635，降低用于拷贝这个卷的优先级并且在回到队列前存在延迟。在将拷贝请求发送回队列之前的延迟确保，例如，激活地拷贝卷的另一族成员没有遭遇拷贝卷的任何问题。

[0134] 在步骤 630，如果没有激活地拷贝所述卷的其它族成员，那么在步骤 640，控制确定另一族成员是否也已经准备拷贝所述卷，但是此时没有激活地拷贝。如果不是，在步骤 645，控制确定此时没有激活地拷贝的这种其它族成员是否应当继承拷贝需求标志。如果是，在步骤 645，这个集群降低拷贝优先级且延迟回到队列。

[0135] 如果在步骤 645 中的否，方法 600 移动到步骤 655 并且所述族成员赢得两个集群成员之间的连接中断器并且继承拷贝标志。因此，在步骤 645，控制确定将指定哪个族成员来继承拷贝标志。非指定族成员降低拷贝优先级并且延迟回到队列（例如，步骤 650）。

[0136] 返回步骤 640，如果另一族成员没有准备好拷贝所述卷，那么在步骤 655，控制确定在所述集群族中仅有一个族成员准备好拷贝所述卷且指定族成员作为集群来继承拷贝标志且完成复制。

[0137] 应当注意的是，在步骤 640，控制可以确定存在准备拷贝并且此时没有激活地拷贝的另一族，但如步骤 645 所示的，确定其它集群将不会继承拷贝标志。因此，步骤 640 中，集群将继承拷贝标志，如步骤 655 所示。

[0138] 在步骤 660，在步骤 655 中继承拷贝标志的指定集群完成拷贝。

[0139] 在步骤 670, 控制清除源集群处的拷贝需求标志并且通过设置用于集群族的族成员的拷贝需求标志, 协作以累积地将族变为一致的。

[0140] 在步骤 675, 集群成员的其它族成员完成他们的拷贝, 并且重置在步骤 655 中设置的在被指定集成拷贝标志的集群内的他们的拷贝需求标志。

[0141] 图 1 至 3 可以是多集群配置的指示。在多集群配置或 (网格配置) 中, 从微代码的角度, 每个集群可能不知道其与自己以及其它集群的关系和角色, 因此与所有其它期间平等独立地工作。例如, 当从一个或两个生产集群全球远程地配置两个或多个集群时, 他们可以通过在远程网络上“拉取”数据而单独地复制。由于集群没有关系意识, 他们不能够基于它们的角色和 / 或与其它集群的距离来按最有效的方式进行操作。

[0142] 此外, 在多集群配置中, 由于这种对关系的无意识, 很大程度上影响选择集群以在安装期间获得卷的装置以及集群获得卷复制的能力。例如, 生产集群可以在城市远程集群上选择全球远程源集群, 以用于安装和 / 或拷贝处理。由于集群间的网络距离, 全球远程集群的效率要低得多。

[0143] 本发明的实现方式能够通过为多集群或网格配置中族成员和族之间引入关系意识来解决这些问题。此外, 实现本发明可以提高数据拷贝和 / 或复制的性能、效率和优化。例如, 协作地复制到族, 从而实现累积族一致性更快 N 倍以及通过相比于让每个集群独立地从相同的远程生产集群获取降低变为 DR 一致和 HA 一致的整体时间, 仅使用累积网络吞吐量的 1/N 可以提升效率和性能。

[0144] 参照图 1 至 6, 本发明的实现方式可以涉及软件、固件、微代码、硬件和 / 或任意组合。实现方式可以采用在媒介中实现的代码或逻辑的形式, 媒介例如是分层存储节点 315 的存储器、存储和 / 或电路, 其中媒介可以包括硬件逻辑 (例如, 集成电路芯片、可编程门阵列 [PGA]、应用专用集成电路 [ASIC]、或其它电路、逻辑或设备)、或计算机可读存储媒介, 例如磁存储媒介 (例如, 电、磁、光、电磁、红外、或半导体系统、半导体或固态存储器、磁带、可移动计算机磁带、以及随机存取存储器 [RAM]、只读存储器 [ROM]、硬磁盘和光盘、致密盘 - 只读存储器 [CD-ROM]、致密盘 - 读 / 写 [CD-R/W] 以及数字视频磁碟 (DVD))。

[0145] 所属领域的技术人员容易了解的是, 可以对上述讨论的方式进行改变, 包括对步骤顺序的改变。此外, 所属领域的技术人员将了解的是, 可以与此处示出的那些方式使用不同的特定组件安排。

[0146] 虽然此处已经详细地说明了本发明的优选实施方式, 应当了解的是, 在不脱离下列权利要求所述的本发明的范围的情况下, 所述领域的技术人员可以对这些实施方式进行各种修改和改变。

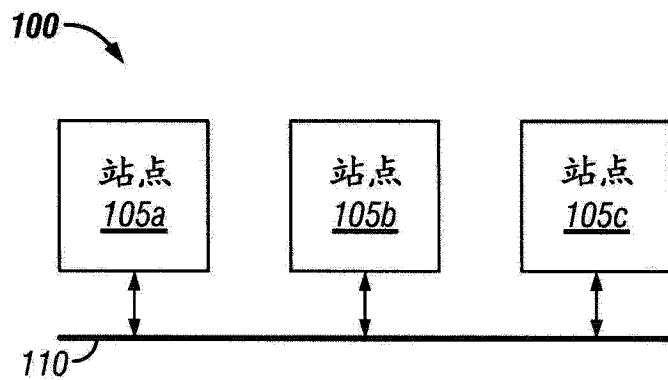


图 1

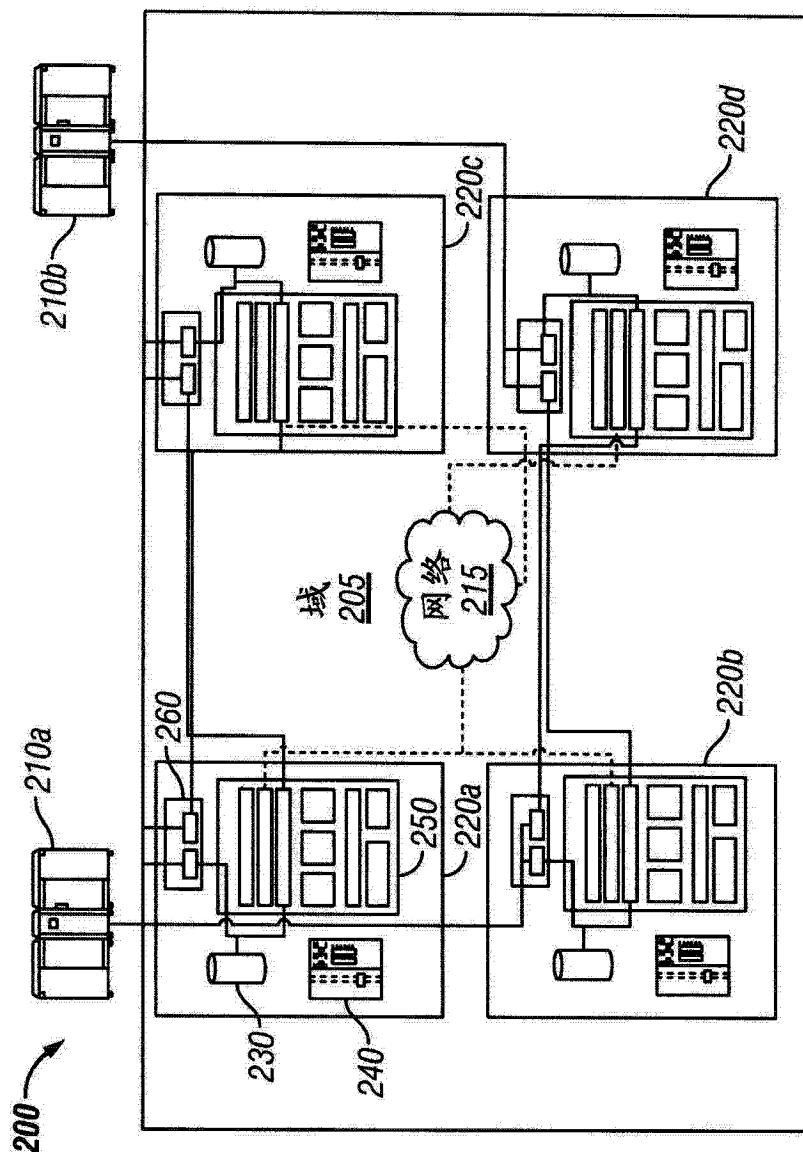


图 2A

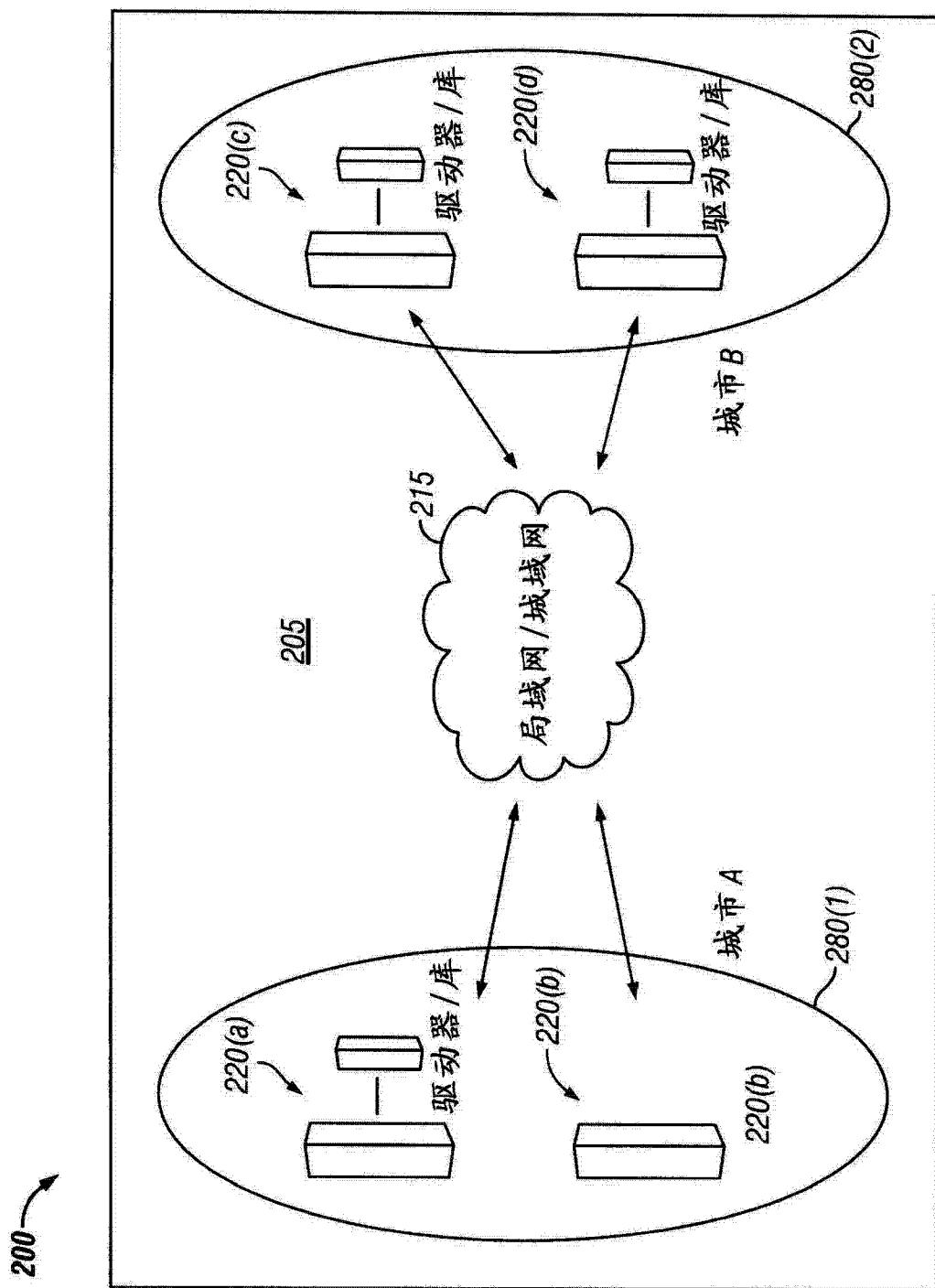


图 2B

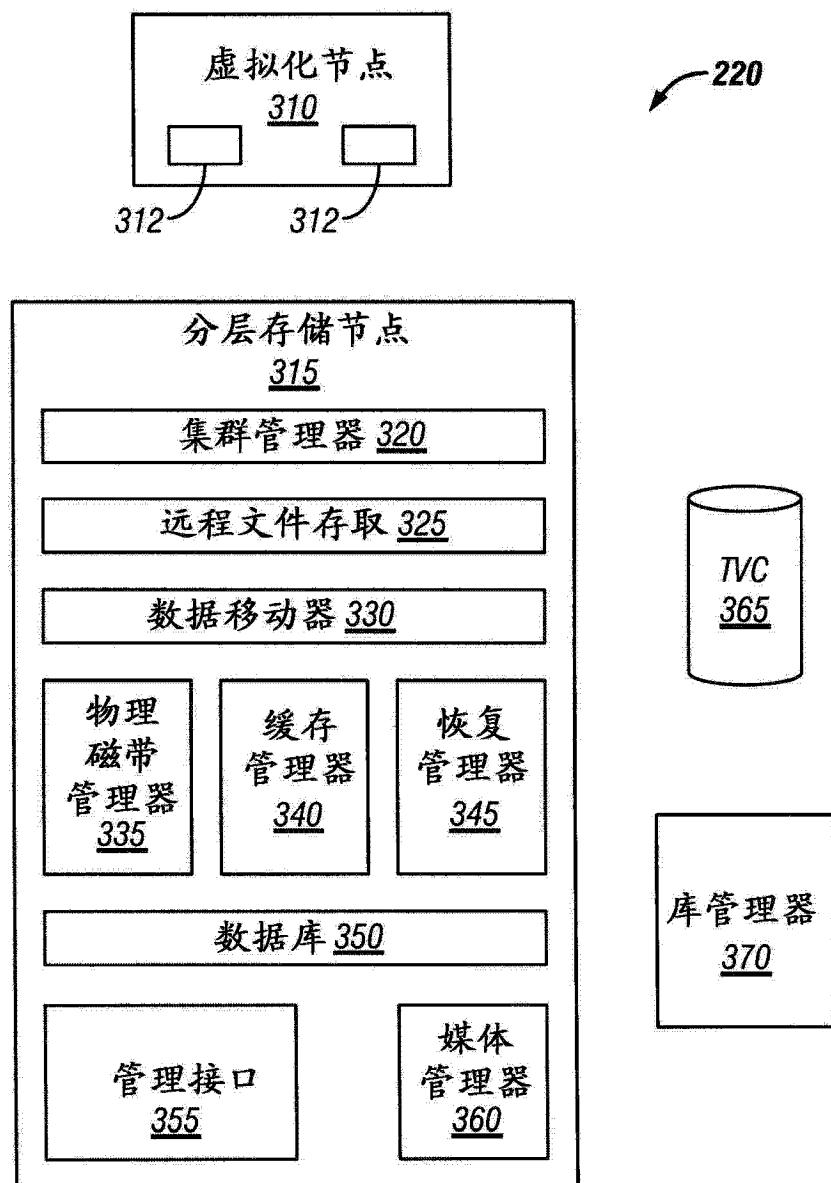


图 3

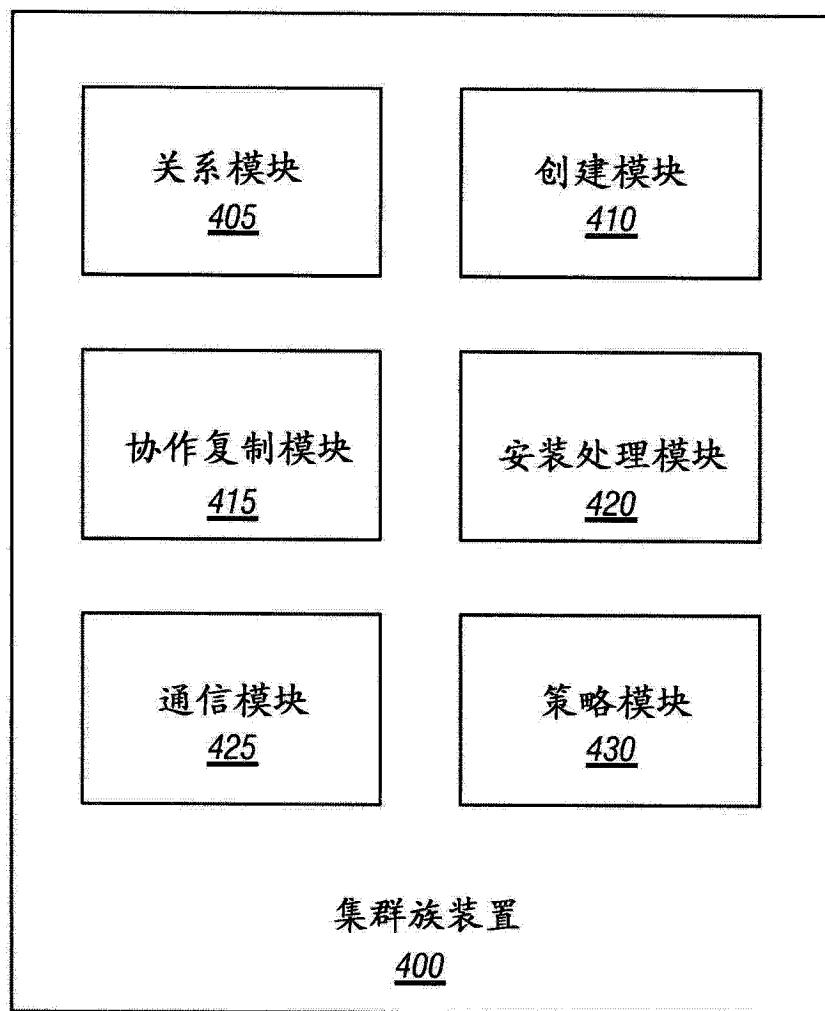


图 4

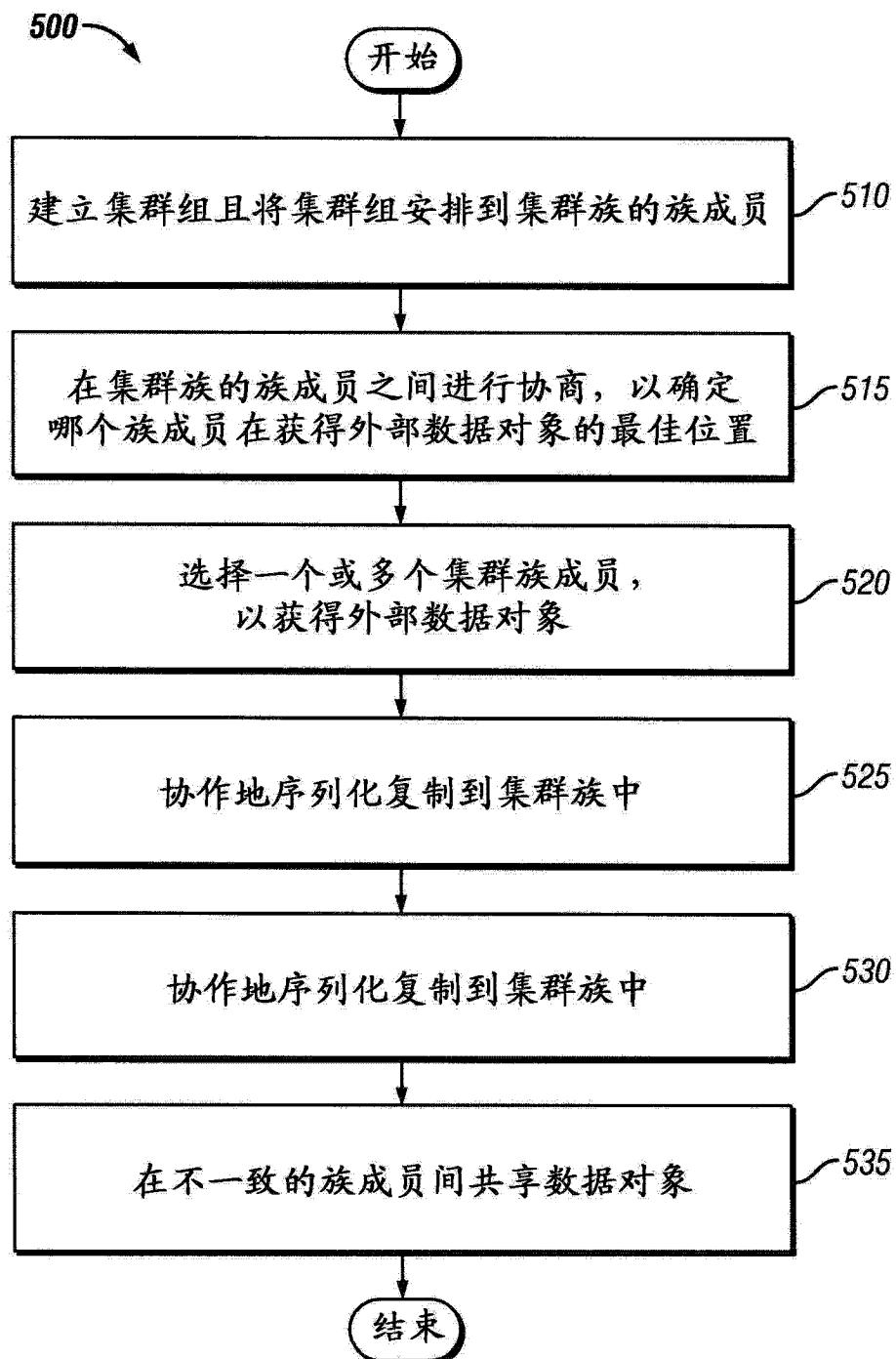


图 5

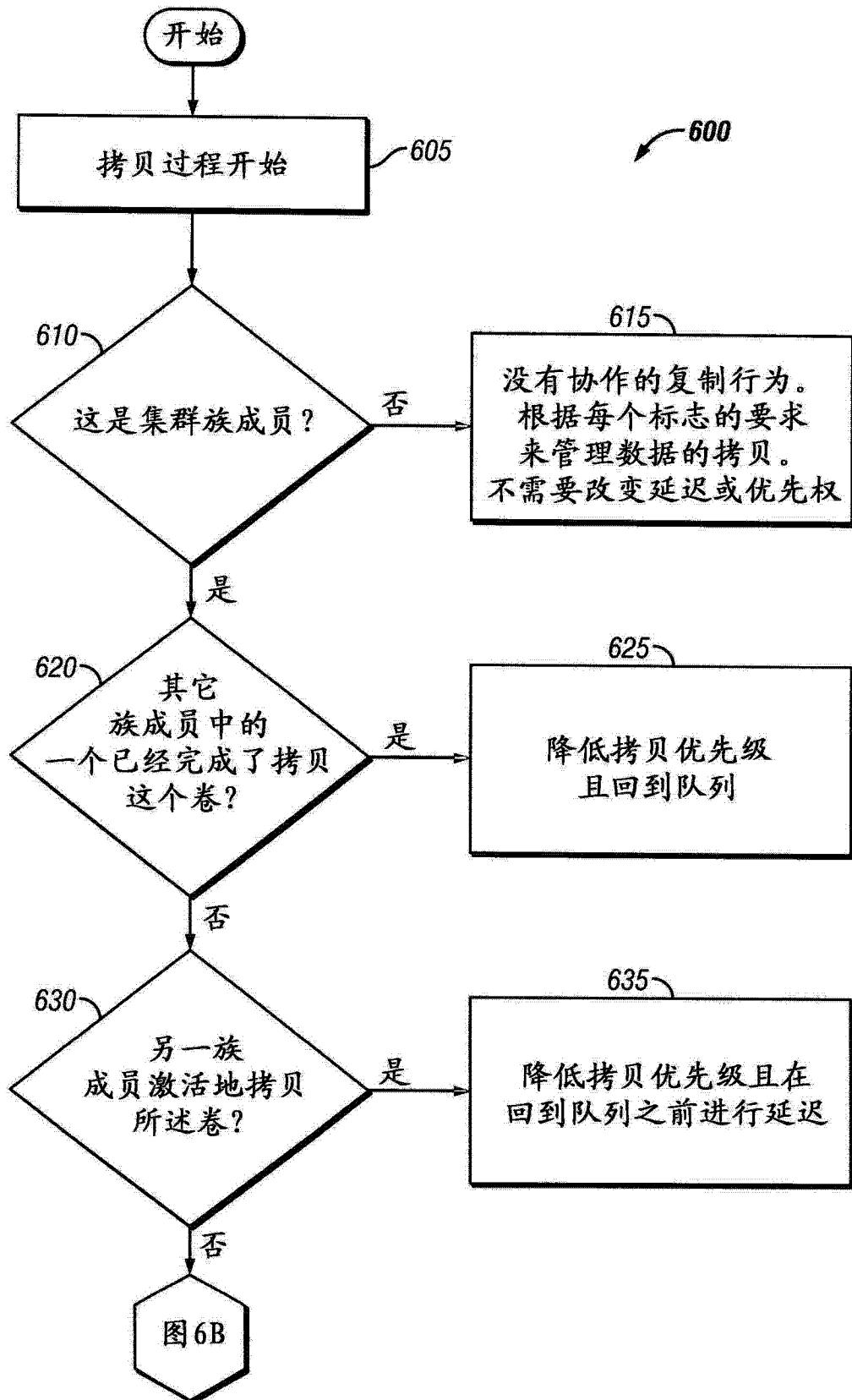


图 6A

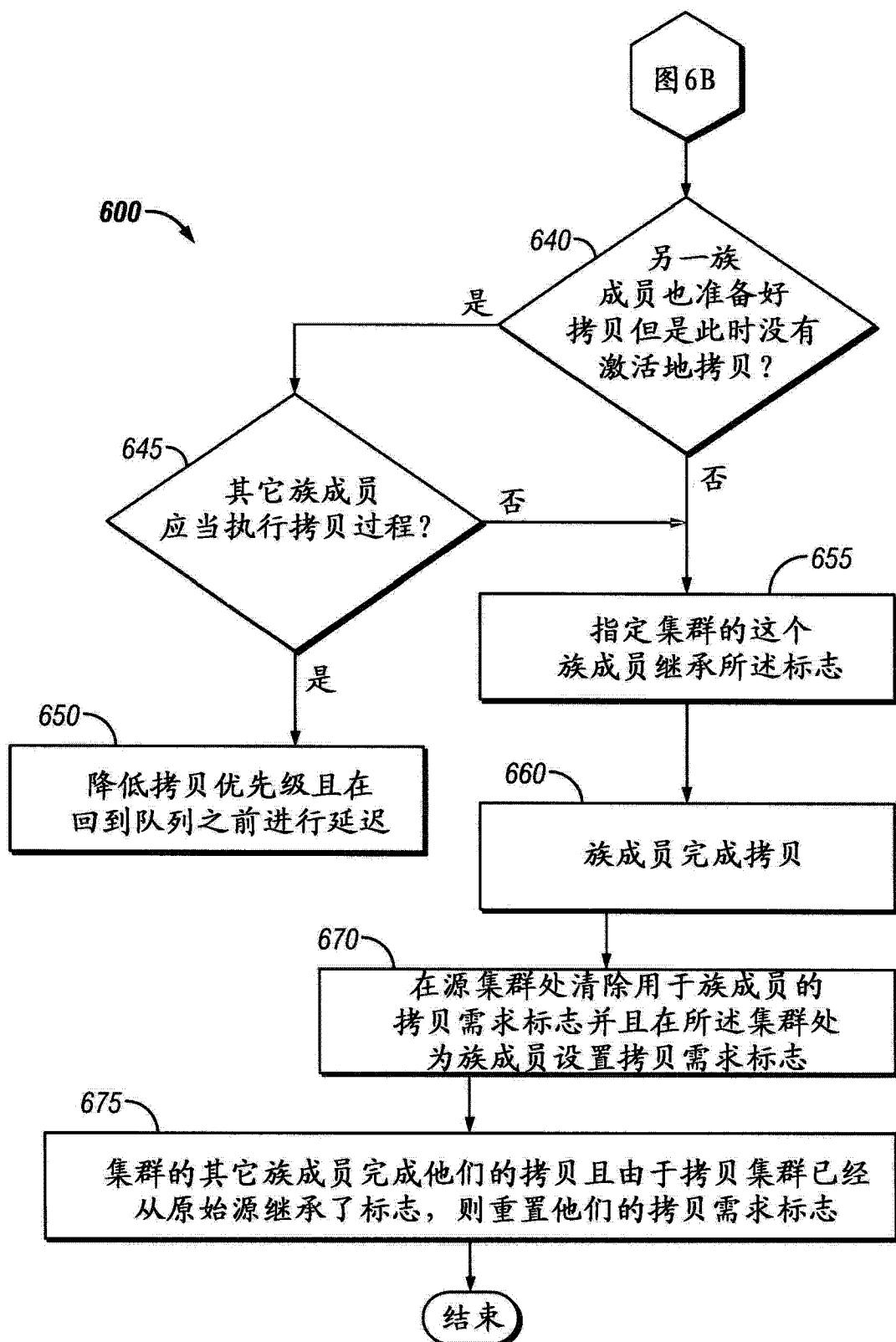


图 6B