

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 11/20 (2006.01)



[12] 发明专利说明书

专利号 ZL 200580001870.X

[45] 授权公告日 2008 年 10 月 1 日

[11] 授权公告号 CN 100422949C

[22] 申请日 2005.1.7

[21] 申请号 200580001870.X

[30] 优先权

[32] 2004.1.9 [33] US [31] 10/754,231

[86] 国际申请 PCT/EP2005/050065 2005.1.7

[87] 国际公布 WO2005/069143 英 2005.7.28

[85] 进入国家阶段日期 2006.6.30

[73] 专利权人 国际商业机器公司

地址 美国纽约

[72] 发明人 戴维·M·沙克尔福德

格利高里·E·麦克布里德

[56] 参考文献

US 5555371A 1996.9.10

EP 1349088A3 2003.3.19

审查员 艾攀

[74] 专利代理机构 中国国际贸易促进委员会专利
商标事务所
代理人 鲍进

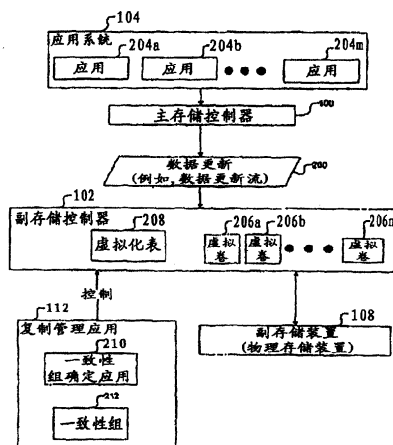
权利要求书 4 页 说明书 15 页 附图 9 页

[54] 发明名称

用于拷贝存储内容的方法和系统

[57] 摘要

本发明提供了用于拷贝存储内容的方法和系统，其中第一单元接收来自第二单元的数据更新。数据更新被存储在与第一单元相关联的多个物理存储位置中。生成指向所述多个物理存储位置中的至少一个物理存储位置的链接，以在第一单元和第二单元之间实现一致数据。



1. 一种用于拷贝存储内容的方法，包括：在第一单元处从第二单元接收数据更新；将所述数据更新存储在与所述第一单元相关联的多个物理存储位置中；并且生成指向所述多个物理存储位置中的至少一个位置的链接，以实现所述第一单元和所述第二单元之间的一致数据。

2. 如权利要求1所述的方法，其中，所述第一单元是耦合到副存储装置的副存储控制器，所述第二单元是耦合到主存储装置的主存储控制器，其中，所述多个物理存储位置与所述副存储装置相关联，并且其中，所述数据更新是在所述第一单元处异步接收的。

3. 如权利要求1所述的方法，其中，所接收的数据只被存储在与所述第一单元相关联的所述多个物理存储位置中一次，所述方法还包括：释放未链接的物理存储位置，以用于存储后续的数据更新。

4. 如权利要求1所述的方法，其中，应用向所述第二单元发送输入/输出请求，其中，所述数据更新对应于来自所述应用的输出请求，并且其中，所述数据更新在所述第一单元处只被存储一次，其中，所述第一单元可以在任意时间点上响应于来自所述应用的输入/输出请求来接替所述第二单元，并且其中，在所述第一和第二单元中的数据在所有时间点上都是一致的。

5. 如权利要求1所述的方法，还包括：在生成所述链接之前判断所存储的数据更新是否形成一致性组；并且，如果所存储的数据更新没有形成一致性组，则等待接收下一数据更新。

6. 如权利要求1所述的方法，还包括：在生成所述链接之前判断所存储的数据更新是否形成一致性组；并且确定提交所述一致性组的所述至少一个物理存储位置，其中，与所述第一单元相关联的虚拟存储装置链接到所确定的至少一个物理存储位置。

7. 如权利要求1所述的方法，还包括：维护将虚拟存储位置映射到所述多个物理存储位置中的所述至少一个物理存储位置的数据结

构, 其中, 所生成的链接与所述数据结构相关联, 并且其中, 多个应用能够执行与所述虚拟存储位置之间的输入/输出操作。

8. 如权利要求 1 所述的方法, 还包括: 维护代表对应于所述数据更新的一致性组的数据结构, 其中, 所维护的数据结构能够指向所述多个物理存储位置; 并且响应于与第一一致性组相关联的第一数据更新的被提交, 删除代表所述第一一致性组的第一数据结构。

9. 如权利要求 1 所述的方法, 还包括: 响应于等待数据更新, 在所述第一单元处接收错误; 并且修改所生成的链接, 以反映所述第一单元和所述第二单元之间的一致数据。

10. 如权利要求 1 所述的方法, 还包括: 响应于生成所述链接, 在所述第一单元处接收错误; 并且修改所述链接, 以反映所述第一单元和所述第二单元之间的一致数据。

11. 一种用于拷贝存储内容的系统, 包括: 处理单元; 多个与所述处理单元相关联的物理存储位置; 包括代码的程序逻辑, 所述代码能够致使所述处理单元执行: (i) 在所述处理单元处接收数据更新; (ii) 将所述数据更新存储在所述多个物理存储位置中; 并且 (iii) 生成指向所述多个物理存储位置中的至少一个物理存储位置的链接, 以实现一致数据。

12. 如权利要求 11 所述的系统, 其中, 所述处理单元是耦合到副存储装置的副存储控制器, 其中, 所述多个物理存储位置与所述副存储装置相关联, 并且其中, 所述数据更新是在所述副存储控制器处被异步接收的。

13. 如权利要求 11 所述的系统, 其中, 所接收的数据只被存储在与所述处理单元相关联的多个物理存储位置中一次, 并且其中, 所述程序逻辑还能够使所述处理单元执行: 释放未链接的物理存储位置, 以用于存储后续的数据更新。

14. 如权利要求 11 所述的系统, 其中, 所述处理单元包括第一处理单元, 其中所述第一处理单元被耦合到第二处理单元, 其中, 应用向所述第二处理单元发送输入/输出请求, 其中, 所述数据更新对

应于来自所述应用的输出请求，并且其中，所述数据更新只被存储在所述第一处理单元处一次，其中，所述第一处理单元可以在任意时间点上响应于来自所述应用的输入/输出请求来接替所述第二处理单元，并且其中，所述第一和第二处理单元中的数据在所有时间点上都是一致的。

15. 如权利要求 11 所述的系统，其中，所述程序逻辑还能够使所述处理单元执行：在生成所述链接之前判断所存储的数据更新是否形成一致性组；并且，如果所存储的数据更新没有形成一致性组，则等待接收下一数据更新。

16. 如权利要求 11 所述的系统，其中，所述程序逻辑还能够使所述处理单元执行：在生成所述链接之前判断所存储的数据更新是否形成一致性组；并且确定提交所述一致性组的所述至少一个物理存储位置，其中，与所述第一处理单元相关联的虚拟存储装置链接到所确定的至少一个物理存储位置。

17. 如权利要求 11 所述的系统，其中，所述程序逻辑还能够使所述处理单元执行：维护将虚拟存储位置映射到所述多个物理存储位置中的所述至少一个物理存储位置的数据结构，其中，所生成的链接与所述数据结构相关联，并且其中，多个应用能够执行与所述虚拟存储位置之间的输入/输出操作。

18. 如权利要求 11 所述的系统，其中，所述程序逻辑还能够使所述处理单元执行：维护代表对应于所述数据更新的一致性组的数据结构，其中，所维护的数据结构能够指向所述多个物理存储位置；并且响应于与第一一致性组相关联的第一数据更新的被提交，删除代表所述第一一致性组的第一数据结构。

19. 如权利要求 11 所述的系统，其中，所述程序逻辑还能够使所述处理单元执行：响应于等待数据更新，在所述第一处理单元处接收错误；并且修改所生成的链接以反映一致数据。

20. 如权利要求 11 所述的系统，其中，所述程序逻辑还能够使所述处理单元执行：响应于生成所述链接，在所述处理单元处接收错

误；并且修改所述链接以反映一致数据。

21. 如权利要求 11 所述的系统，其中，所述处理单元包括第一处理单元，并且还包括：耦合到所述第一处理单元的第二处理单元，其中，在所述第一处理单元处的数据更新是从所述第二处理单元接收的，并且其中，所述一致数据是在所述第一处理单元和所述第二处理单元之间实现的。

22. 一种用于拷贝存储内容的系统，包括：用于接收数据更新的装置；用于将所述数据更新存储在多个物理存储位置中的装置；以及用于生成指向所述多个物理存储位置中的至少一个物理存储位置的链接，以实现一致数据的装置。

23. 如权利要求 22 所述的系统，还包括：在生成所述链接之前判断所存储的数据更新是否形成一致性组的装置；如果所存储的数据更新没有形成一致性组，则等待接收下一数据更新。

24. 如权利要求 22 所述的系统，还包括：在生成所述链接之前判断所存储的数据更新是否形成一致性组的装置；以及确定提交所述一致性组的所述至少一个物理存储位置的装置，其中，虚拟存储装置链接到所确定的至少一个物理存储位置。

25. 如权利要求 22 所述的系统，还包括：维护代表对应于所述数据更新的一致性组的数据结构的装置，其中，所维护的数据结构能够指向所述多个物理存储位置；以及响应于与第一一致性组相关联的第一数据更新的被提交，删除代表所述第一一致性组的第一数据结构的装置。

26. 如权利要求 22 所述的系统，其中，所述一致数据是在第一处理单元和第二处理单元之间实现的。

用于拷贝存储内容的方法和系统

技术领域

本发明涉及以下在同一天递交的共同待审查并且共同转让的专利申请：“Ordering Updates in Remote Copying of Data”，其律所卷号为 no. SJO920030037US1，该申请的全部内容以引用的方式被包含进来。

背景技术

本公开涉及一种使用虚拟技术保持远程异步拷贝的一致性的方法、系统和制品。

发明内容

信息技术系统，包括存储系统在内，可能需要针对现场灾难或断电（outage）的保护。此外，信息技术系统可能需要用于数据迁移、数据备份或数据复制的性能。灾难或断电恢复、数据迁移、数据备份或数据复制的实现方式可以包括在存储系统中建立数据的镜像或拷贝。在某些信息技术系统中，数据从主存储控制器被拷贝到副存储控制器。如果主存储控制器变为不可用，则副存储控制器可被用来接替不可用的主存储控制器。

信息技术系统中的数据拷贝可以是同步的或异步的。同步拷贝涉及从主存储控制器向副存储控制器发送数据，并且在结束对主存储控制器的写操作之前确认这些数据的接收。因此，由于要等待来自副存储控制器的确认，所以同步拷贝使写操作响应时间变慢。然而，同步拷贝在副存储控制器上提供了顺序一致的数据。

异步拷贝可以提供比同步拷贝更好的性能，这是因为对主存储控制器的写操作可以在从副存储控制器确认对所发送数据的接收之

前结束。然而，由于在副存储控制器上接收的数据可能不是按照对主存储控制器进行更新（即，写操作）的顺序，所以不得不要确保数据顺序的一致性。在异步拷贝中，通过在固化位置（例如，日志数据集）中临时存储更新，一直到一组一致的更新可用于施加到与副存储控制器相关联的副存储装置为止，这样就可以实现主副存储控制器之间的跨设备一致性。

提供了一种用于拷贝存储内容的方法和系统，其中第一单元接收来自第二单元的数据更新。数据更新被存储在与第一单元相关联的多个物理存储位置中。生成到所述多个物理存储位置中的至少一个位置的链接，以实现第一单元和第二单元之间的一致性数据。

在另外的实施方案中，第一单元是耦合到副存储装置的副存储控制器，第二单元是耦合到主存储装置的主存储控制器，其中所述多个物理存储位置与副存储装置相关联，并且在第一单元处异步地接收数据更新。

在进一步的实施方案中，接收到的数据只被存储在与第一单元相关联的多个物理存储位置中一次，未链接的物理存储位置被释放以用于存储后续的数据更新。

在另外的实施方案中，应用向第二单元发送输入/输出请求，其中数据更新对应于来自该应用的输出请求，并且数据更新只被存储在第一单元上一次，其中第一单元可以在任意时间点响应于来自应用的输入/输出请求而接替第二单元，并且第一和第二单元中的数据在所有时间点都是一致的。

在进一步的实施方案中，在生成链接之前判断所存储的数据更新是否形成了一致性组。如果所存储的数据更新未形成一致性组，则等待接收下一数据更新。

在另外的实施方案中，在生成链接之前判断所存储的数据更新是否形成了一致性组。确定提交一致性组的所述至少一个物理存储位置，其中与第一单元相关联的虚拟存储装置链接到所确定的至少

一个物理存储位置。

在进一步的实施方案中，维护将虚拟存储位置映射到所述多个物理存储位置中的所述至少一个位置的数据结构，其中所生成的链接与该数据结构相关联，并且多个应用能够与所述虚拟存储位置之间进行输入/输出操作。

在其他实施方案中，维护代表对应于数据更新的一致性组的数据结构，其中所维护的数据结构能够指向所述多个物理存储位置。响应于与第一一致性组相关联的第一数据更新被提交，对代表第一一致性组的第一数据结构进行删除。

在进一步的实施方案中，响应于等待数据更新，在第一单元处接收到错误。对生成的链接进行修改，以反映第一单元和第二单元之间的一致数据。

某些实施方案利用虚拟存储系统实现了异步远程拷贝的一致性。复制管理应用将已被接收但尚未与关联到其他存储控制器的数据一致的数据写入未使用的物理存储装置中。响应于接收到为提供一致性所必需的数据，虚拟化表可以在真正的副位置上被更新，以指向物理存储装置中已被写入数据的位置。在某些实施方案中，在副存储控制器上的接收到的数据只可以被写入与副存储控制器相关联的物理存储装置一次。

附图说明

现在参考附图，其中相近的标号通篇代表对应的部分：

图 1 根据本发明的某些描述方面图示了一个计算环境的框图；

图 2 根据本发明某些描述的实施方式图示了与所述计算环境有关的数据结构和设备的框图；

图 3 根据本发明某些描述的实施方式图示了一致性组的框图；

图 4 根据本发明某些描述的实施方式图示了使用虚拟技术保持一致性的逻辑；

图 5 根据本发明某些描述的实施方式图示了在一个典型实施方

案中的数据结构的第一状态的框图;

图 6 根据本发明某些描述的实施方式图示了一个典型实施方案中的数据结构的第二状态的框图;

图 7 根据本发明某些描述的实施方式图示了一个典型实施方案中的数据结构的第三状态的框图;

图 8 根据本发明某些描述的实施方式图示了用于保持一致性和灾难恢复的逻辑; 并且

图 9 根据本发明某些描述的实施方式图示了一个计算机体系结构的框图。

具体实施方式

在以下描述中参考构成本说明书的一部分并且图示说明几种实施方式的附图。可以理解, 可以采用其他实施方式并且可以作出结构和操作上的改变, 而不会偏离当前实施方式的范围。

图 1 根据本发明的某些方面图示了一个计算环境的框图。主存储控制器 100 耦合到副存储控制器 102。包括一个或多个应用的应用系统 104 可以对主存储控制器 100 执行 I/O 操作, 包括写操作。在某些实施方案中, 应用系统 104 可以驻留在经由主机总线适配器耦合到主存储控制器 100 的主机计算设备中。

主存储控制器 100 和副存储控制器 102 可以分别对主存储装置 106 和副存储装置 108 存取数据, 其中主存储装置 106 被耦合到主存储控制器 100, 副存储装置 108 被耦合到副存储控制器 102。另外, 主存储控制器 100 和副存储控制器 102 可以分别控制主存储装置 106 和副存储装置 108 的操作。主存储装置 106 和副存储装置 108 可以包括非易失性存储装置, 例如硬盘驱动器、RAID、直接访问型存储设备或其他类型的物理存储装置。

在某些实施方案中, 主存储控制器 100 可能不在运行中, 与副存储控制器 102 相关联的数据可被恢复系统 110 用于处理。在本发明的实施方案中, 关联于副存储控制器 102 的数据与关联于主存储

控制器 100 的数据保持一致。一致性通过复制管理应用 112 来保持。

复制管理应用 112 被耦合到主存储控制器 100 和副存储控制器 102，在某些实施方案中，复制管理应用 112 可以将数据从主存储控制器 100 镜像映射到副存储控制器 102。在一些实施方案中，可以通过将数据从主存储控制器 100 异步地拷贝到副存储控制器 102 来完成镜像映射。

在某些实施方案中，复制管理应用 112 可以分布在主存储控制器 100 和副存储控制器 102 上。在其他实施方案中，复制管理应用 112 可以驻留在不同于主存储控制器 100 和副存储控制器 102 的单独系统上。在其他的实施方案中，复制管理应用 112 可以只驻留在主存储控制器 100 和副存储控制器 102 之一中。

在某些实施方案中，复制管理应用 112 保持从应用系统 104 接收的数据更新的一致性，其中，所述数据更新是从主存储控制器 100 异步拷贝到副存储控制器 102 的。复制管理应用 112 可以进行对被耦合到副存储控制器 102 的副存储装置 108 的虚拟，以保持主存储控制器 100 和副存储控制器 102 间的数据一致性。在某些实施方案中，虚拟包括物理副存储装置 108 到虚拟卷的映射。

这样，图 1 图示了复制管理应用 112 虚拟副存储控制器 102 并且保持主存储控制器 100 和副存储控制器 102 间的数据一致性的实施方案，其中数据从主存储控制器 100 异步拷贝到副存储控制器 102。

图 2 根据本发明某些描述的实施方式图示了与复制管理应用 112 和副存储控制器 102 有关的数据结构和设备的框图。

副存储控制器 102 可以接收数据更新 200，该数据更新 200 是作为从一个或多个应用 204a...204m 到主存储控制器 100 的写操作的结果而生成的，其中一个或多个应用 204a...204m 可以构成应用系统 104。在某些实施方案中，来自应用系统 104 的数据更新 200 经由主存储控制器 100 异步到达副存储控制器 102，并且数据更新 200 可被称为数据更新流。

复制管理应用 112 创建一个或多个虚拟设备，例如与副存储控制器 102 相关联的虚拟卷 206a...206n。主存储控制器 100 也可以具有与虚拟卷 206a...206n 相对应的虚拟卷。在某些实施方案中，对应于虚拟卷 206a...206n 的数据被存储在物理副存储装置 108 中的位置上。虚拟卷 206a...206n 到物理副存储装置 108 中的位置的映射关系可以被存储在与副存储控制器 102 相关联的虚拟化表 208 中。在某些实施方案中，虚拟化表 208 可被耦合到虚拟卷 206a...206n。

应用系统 104 对于与主存储控制器 100 相关联的虚拟卷执行 I/O 操作，并且对应的虚拟卷 206a...206n 也被关联到副存储控制器 102。

在某些实施方案中，复制管理应用 112 可以包括一致性组确定应用 210 以及对应于一致性组 212 的关联数据结构。

这样，图 2 描述了复制管理应用 112 虚拟副存储控制器 102 并且在主存储控制器 100 和副存储控制器 102 间保持数据一致性的实施方案。

图 3 根据本发明某些描述的实施方式图示了由复制管理应用 112 创建的典型的一致性组的框图。

一致性组是一组更新，在该组中的更新可以跨越多个存储卷，并且其中为了保持包含在所述多个存储卷的每个存储卷中的数据之间的相互数据一致性，所述更新必须被一起写。提供一个非限制性的例子：第一命令将与主存储控制器 100 相关联的卷 A1 拷贝到与副存储控制器 102 相关联的卷 B1，第二命令将与主存储控制器相关联的卷 A2 拷贝到与副存储控制器相关联的卷 B2。要求卷 B1 和 B2 应当代表在特定时间点卷 A1 和 A2 中的数据集的一致状态。在对这些卷的特定操作序列中，可以发生由应用系统 104 完成的下面一组相关写操作（其中第二操作发生在第一操作之后）：

1. 写到卷 A1 上的数据集（数据被更新）
2. 写到卷 A2 上的数据集（数据被更新）

当卷 A1 和 A2 分别经由数据更新 200 被异步拷贝到卷 B1 和 B2 时，则下面的作为非限制性的例子的操作序列可以在卷 B1、B2 中产

生相对于卷 A1、A2 的不一致状态：

将卷 A1 拷贝到卷 B1

2. 写到卷 A1 上的数据集（数据被更新）

3. 写到卷 A2 上的数据集（数据被更新）

4. 将卷 A2 拷贝到卷 B2

在所有拷贝操作都结束后，即，在第四操作结束后，卷 B2 包含卷 A2 的数据更新，而卷 B1 不包含卷 A1 的数据更新。卷 B1、B2 的集合相对于卷 A1、A2 的集合处于不一致的状态。使用与副存储控制器 102 相关联的卷 B1、B2 的应用 204a...204m 不能从存储在卷 B1、B2 中的备份拷贝进行恢复。

因此，并非所有的序列数据更新可以形成一组一致的数据更新。在图 3 中，表 300 的行代表不同的设备，而列代表不同的时间。时间是相对时间而非绝对时间。例如，t3（标号 306）是 t2（标号 304）后的时间，t2（标号 304）是 t1（标号 302）后的时间。在表 300 主体内的字母-数字组合标识了在某一时间上对某一设备的更新，用字母标识一个应用，用数字标识一个用于该应用的更新序列。例如，B1（标号 308）是来自名为 B 的应用的第一数据更新，这里，该更新在相对时间 t1（标号 302）到达并用于设备 D3（标号 310）。在表 300 的条目中的不同阴影标识了一组具有数据一致性的更新，并且不一定正好是表中条目的垂直片段。例如，表 300 具有三个一致性组 312、314、316。在一个一致性组中的更新数据可能需要被一起施加到副控制器 102，用于与副控制器 102 相关联的数据，以便保持与关联到主控制器 100 的数据之间的一致。在表 300 中的一致性组 312、314、316 的确定可以以本领域公知的任何方式来完成。

这样，图 3 图示了可以由复制管理应用 112 通过处理从主存储控制器 100 异步到达副存储控制器 102 的数据更新流 200 而生成的典型一致性组 312、314、316 的实施方案。为了与主存储控制器 102 之间保持数据一致性，首先确定并提交一致性组，然后通过副存储控制器 108 的虚拟卷 206a...206n 中的指针或链接来反映该一致性组

的数据更新。

图 4 根据本发明某些描述的实施方式图示了在副存储控制器 102 中实现的通过虚拟来保持一致性的逻辑。

控制开始于框 400，其中复制管理应用 112 创建虚拟化表 208，该表对应于与副存储控制器 102 相关联的虚拟卷 206a...206n，其中虚拟化表 208 指向物理存储装置 108 中的位置，并且可以定义虚拟卷 206a...206n 的一致数据内容。虚拟卷 206a...206n 的数据内容的一致性是相对于与主存储控制器 100 相关联的数据内容而言的。

复制管理应用 112（在框 402）接收对与副存储控制器 102 相关联的虚拟卷 206a...206n 的数据更新 200。例如，在某些实施方案中，数据更新 200 可以是数据更新 B1（标号 308）。

复制管理应用 112（在框 404）将数据更新 200 写入未使用的物理存储装置。例如，复制管理应用 112 可以将数据更新 200 写入副存储装置 108 的未使用位置。

复制管理应用 112（在框 406）判断是否已接收到用于一致性组 212（例如一致性组 312、314、316）的所有数据更新。如果已收到，则复制管理应用 112（在框 408）更新虚拟化表 208，以指向物理存储装置 108 中构成对包括在一致性组 212 中的数据更新的提交的位置。更新后的虚拟化表 208 定义了与虚拟卷 206a...206n 相关联的新数据。因此，与副存储控制器 102 的虚拟卷 206a...206n 相关联的数据与关联到主存储控制器 100 的数据是一致的。

作为虚拟化表 208 的更新的结果，复制管理应用 112（在框 410）释放物理存储装置 108 中的空间。例如，某些在框 404 完成的写入未使用物理存储装置的数据更新 200 可能是不需要的，可被释放。复制管理应用 112（在框 402）接收下一个数据更新 200。

如果复制管理应用 112（在框 406）确定用于一致性组 212（例如一致性组 312、314、316）的数据更新还没有全部收到，则复制管理应用 112（在框 402）接收下一个数据更新 200。

这样，图 4 描述了某些实施方案，其中，复制管理应用 112 将

所有数据更新 200 写入未使用物理存储装置 108 中的位置，并且在确定一致性组 212 之后，可以通过将虚拟化表 208 更新为指向一定数量的写入数据更新从而仅仅固化所述一定数量的写入数据更新。写入的数据更新不需要拷贝。

图 5 根据本发明某些描述的实施方式图示了在一个典型实施方案中的数据结构的第一状态的框图，其中原始数据与副存储控制器 102 相关联，对应于第一和第二一致性组的数据更新已到达副存储控制器 102 但尚未提交。

在图 5 中，示出了一个典型的盘 500（例如物理存储装置 108）的线性表示。典型的盘 500 具有十个物理块，第一物理块包含名为 VB1 的第一虚拟块（VB），它是盘 500 中的原始数据的一部分。当数据更新 200 到达并且与不同的一致性组（CG）关联时，数据更新被写入盘 500。例如，第四虚拟块（VB4）是第二一致性组（CG2）的一部分，用于该虚拟块的数据更新 200 被写入物理块 5。所述虚拟块对应于虚拟卷 206a...206n。

图 5 还图示了指向盘 500 中的原始数据的物理块的当前数据指针 502、指向与构成第一一致性组的数据更新相对应的物理块的第一一致性组指针 504 和指向与构成第二一致性组的数据更新相对应的物理块的第二一致性组指针 506。

还维护代表指针 502、504、506 的虚拟块到物理块映射关系的表。例如，当前数据指针表 508 示出在盘 500 中虚拟块到物理块的当前映射关系。在当前数据指针表 508 中的映射关系示出了被固化或提交的数据，即在主存储控制器 100 和副存储控制器 102 之间一致的数据。访问与副存储控制器 102 相关联的数据的应用 204a...204n 作用于当前数据指针 502 所指向的数据。

第一一致性组指针表 510 针对形成第一一致性组的一部分的数据更新示出了在盘 500 中虚拟块到物理块的映射关系。在第一一致性组指针表 510 中表示的数据不被固化，因为第一一致性组尚未被提交。

类似地，第二一致性组指针表 512 针对形成第二一致性组的一部分的数据更新示出了在盘 500 中虚拟块到物理块的映射关系。在第二一致性组指针表 512 中表示的数据不被固化，因为第二一致性组尚未被提交。

这样，图 5 就示出了一种实施方案中的数据结构的第二状态，其中，用于第一和第二一致性组的数据更新已经到达副存储控制器 102 但尚未被提交。

图 6 根据本发明某些描述的实施方式图示了在一个典型实施方案中的数据结构的第二状态的框图，其中对应于第一一致性组指针表 510 的数据更新 200 已被固化，即，第一一致性组已被提交，并且对应于第三一致性组的未提交数据更新 200 已到达副存储控制器 102。

图 6 示出了指向盘 500 中的原始数据的物理块的已更新当前数据指针 502、指向与构成未提交的第二一致性组的数据更新相对应的物理块的第二一致性组指针 506 和指向与构成第三一致性组的未提交数据更新相对应的物理块的新的第三一致性组指针 600。

还维护代表指针 502、506、510 的虚拟块到物理块映射关系的表。例如，当前数据指针表 508 示出在盘 500 中虚拟块到物理块的当前映射关系。在当前数据指针表 508 中的映射关系示出了在第一一致性组的数据更新 200 已被提交后被固化或提交的数据。

第三一致性组指针表 602 针对形成第三一致性组的一部分的数据更新示出了在盘 500 中虚拟块到物理块的映射关系。在第二一致性组指针表 512 或第三一致性组指针表 602 中表示的数据不被固化，因为第二和第三一致性组尚未被提交。在图 6 中，第一一致性组指针表 510 已被删除，因为用于第一一致性组的数据更新已被提交。

这样，图 6 示出了一种实施方案，其中，用于第一、第二和第三一致性组的数据更新 200 已到达副存储控制器 102，但只有第一一致性组的数据更新已被提交。

图 7 根据本发明某些描述的实施方式示出了在一个典型实施方

案中的数据结构的第三状态。在图 7 中，对应于第二和第三一致性组的数据更新 200 已被提交。

由于第一、第二和第三一致性组的数据更新都已被提交，所以第一一致性组指针表 510、第二一致性组指针表 512 和第三一致性组指针表 602 都被示为已删除。当前数据指针指向盘 500 的物理块 5、6、8、9，并且分别对应于虚拟块 VB4、VB2、VB1 和 VB3。

这样，图 7 示出了一种实施方案，其中，用于第一、第二和第三一致性组的数据更新 200 已经到达副存储控制器 102 并且所有的数据更新都已提交。

图 8 根据本发明某些描述的实施方式图示了在复制管理应用 112 中实现的用于保持一致性和灾难恢复的逻辑。

控制开始于框 800，其中当前一致性组被初始化为 1。复制管理应用 112（在框 802）确定是否所有的存储控制器，即主存储控制器 100 和副存储控制器 102 都已接收到当前一致性组的数据更新。如果是的，则复制管理应用 112 提交当前一致性组的数据更新，并且对于所有的存储控制器（在框 804）将当前指针设置为指向当前一致性组数据更新。在某些实施方案中，当前指针可以通过与虚拟化表 208 和/或指针表 508、510、512、602 相关联的数据结构来实施。

复制管理应用 112（在框 806）确定是否所有的指针都已更新。如果是的，则复制管理应用（在框 808）删除当前一致性组指针。

复制管理应用 112（在框 810）将当前一致性组增一。例如，如果在框 802 到 808 的第一次迭代中，处理的是第一一致性组的数据更新，那么在框 802 到 808 的第二次迭代中，将要处理的则是第二一致性组的数据更新。复制管理应用 112（在框 812）等待当前一致性组的数据更新到达。如果在等待的过程中没有错误的话，则控制前进到框 802，在这里复制管理应用 112 确定是否所有的存储控制器都已接收到当前一致性组的数据更新。

如果复制管理应用 112（在框 802）确定并非所有的存储控制器都已收到当前一致性组的数据更新，则复制管理应用 112（在框 812）

等待当前一致性组的所有数据更新到达。如果（在框 812）等待时发生了错误或灾难，那么复制管理应用 112（在框 814）确定在所有存储控制器上的当前一致性组，然后（在框 816）确定在每个存储控制器中最后的可用一致性组。复制管理应用 112（在框 818）确定在所有存储控制器上可用的最大一致性组，然后（在框 820）将所有存储控制器上的指针更新为对应于在所有存储控制器上最后的可用一致性组。

如果复制管理应用 112（在框 806）确定并非所有的指针都已更新，则复制管理应用 112（在框 822）等待所有的指针被更新。如果在等待时没有错误，则控制返回框 806，在这里复制管理应用 112 确定是否已更新了所有的指针。

如果（在框 822）等待时发生了错误，则复制管理应用 112（在框 824）开始对所有的存储控制器执行 FOR 循环。在每次迭代中为一个存储控制器（在框 824）执行对 FOR 循环的控制。如果 FOR 循环没有结束，即，不是所有的存储控制器都已被处理，则复制管理应用 112（在框 826）确定指针是否对应于正在处理的存储控制器的当前一致性组。如果是的，则复制管理应用 112（在框 828）使指针倒退，以对应于前面的一致性组，并且控制（在框 824）前进到 FOR 循环针对下一存储控制器的下一次迭代。

如果指针（在框 826）不对应于当前一致性组，则控制（在框 824）前进到 FOR 循环针对下一存储控制器的下一次迭代。在 FOR 循环（在框 830）结束时，存储控制器中的数据相对于前面的一致性组是一致的。

这样，图 8 的逻辑就使所有存储控制器上的数据都保持相互一致，此外，如果在更新指针或接收数据更新时有错误或灾难降临于存储控制器，则图 8 的逻辑可以使存储控制器中的指针倒退，以反映与先前处理的一致性组相关联的数据更新，使得存储控制器中的数据相互一致。

这些实施方案通过虚拟存储系统实现了异步远程拷贝的一致

性。复制管理应用将已被接收但尚未与其他存储子系统中的数据一致的数据写入未使用的物理存储装置。响应于为提供一致性所必需的数据的被接收，可以在真正的目标位置上更新指针和表，以指向在数据已被写入的物理存储装置中的位置。

因此，所述实施方案不需要在副存储控制器上进行两阶段提交以确保数据一致性。在两阶段提交中，数据更新可在第一阶段中被写入与副存储控制器相关联的日志数据集，并且在一致性组被提交的第二阶段中可以拷贝适当的数据更新。所述实施方案在不使用日志数据集的情况下来实现。响应于一致性组的被提交，对指针进行调整，以指向物理存储装置中的适当位置，使得调整后的指针代表在所有存储控制器上的一致数据集。

此外，在更新指针或等待数据更新时发生错误或灾难的情况下，所述实施方案可以调整存储控制器中的指针，以反映与先前处理的一致性组相关联的数据更新，使得存储控制器中的数据相互一致。

此外，由于数据更新可被保存在盘中，因此与数据更新被保存在高速缓存中相比，在空间利用率或易失性方面的限制要小一些。另外，要包含对同时进行中的大量一致性组的更新，只需要充足的额外盘空间就可以了。

此外，由于没有任何数据更新被物理覆写，所以在某些实施方案中，有可能在存储控制器中保留或者倒退回多个先前版本的一致数据集。

使用标准的程序设计和/或工程技术来制作软件、固件、硬件或它们的组合，可以将所描述的技术实现为一种方法、设备或制品。术语“制品”用在这里是指用硬件逻辑（例如，集成电路芯片、可编程门阵列(PGA)、专用集成电路(ASIC)等）或计算机可读介质（例如，磁存储介质（例如硬盘驱动器、软盘、磁带）、光存储（例如 CD-ROM、光盘等）、易失性或非易失性存储器件（例如 EEPROM、ROM、PROM、RAM、DRAM、SRAM、固件、可编程逻辑等））实现的代码或逻辑。计算机可读介质中的代码由处理器来访问并执行。完

成所述实施方式的代码还可以通过传输介质来访问或者通过网络从文件服务器获得。在这些情况下，实现代码的制品可以包括传输介质，例如网络传输线、无线传输介质、通过空间传播的信号、无线电波、红外信号等。当然，本领域的技术人员会认识到：可以对这种配置进行很多修改，而不会偏离所述实施方式的范围，并且制品可以包括本领域公知的任何信息承载介质。

图 9 图示了可以实现本发明的某些方面的计算机体系结构的框图。图 9 图示了存储控制器 100、102、包括应用系统 104 的主机以及包括复制管理应用 112 的任意计算设备的一种实施方式。存储控制器 100、102、包括应用系统 104 的主机、以及包括复制管理应用 112 的任意计算设备可以实现一个计算机体系结构 900，它具有处理器 902、存储器 904（例如，易失性存储器件）和存储装置 906（例如，非易失性存储装置、磁盘驱动器、光盘驱动器、磁带驱动器等）。存储装置 906 可以包括内部存储装置、外接存储装置或网络可访问的存储装置。存储装置 906 中的程序可以利用本领域公知的方式被加载到存储器 904 中并由处理器 902 来执行。该体系结构还可以包括网卡 908，以使得能够与网络通信。该体系结构还可以包括至少一个输入设备 910 和至少一个输出设备 912，输入设备例如是键盘、触摸屏、笔、声动输入等，输出设备例如是显示设备、扬声器、打印机等。

图 4-8 描述了按特定顺序发生的具体操作。此外，所述操作不但可以顺序执行，还可以并行执行。在可选的实施方式中，某些逻辑操作可以按不同的顺序来执行，可以被修改或删除，但仍能实现本发明的实施方式。此外，可以向上述逻辑中添加步骤而仍遵循所述实施方式。另外还有一些步骤可以利用单个过程或分布式过程来完成。

虽然所述实施方案是关于异步对等远程拷贝而言的，但是也可以使用需要两阶段提交来进行拷贝的另外的实施方案。例如，如果使用同一虚拟化引擎使得同时包含数据库日志（log）和表空间，那

么事务可被写入所述日志，然后可以使用在所述实施方案中描述的虚拟技术来完成提交。因为不再需要向存储装置中转移数据以供提交，所以数据库处理的性能会有所提高。

为了图示说明方便，用单独的模块描述了很多软件和硬件组件。这样的组件可被集成到更少的组件中或者被划分到更多的组件中。另外，被描述为由某一特定组件完成的某些操作可以由其他组件来完成。

这样，为图示说明和描述的目的在上面对具体实施方式进行了说明。但这不是穷尽性的，也不是要将本发明限制为所公开的精确形式。遵照以上教导可以作出很多修改和变体。本发明的范围不应由上述详细描述来限定，而是要用所附的权利要求来限定。以上说明、例子和数据提供了对本发明的构成部分的制造和使用的完整描述。由于在不偏离本发明的精神和范围的条件下可以完成本发明的很多实施方式，所以本发明表现为这里所附的权利要求。

图1

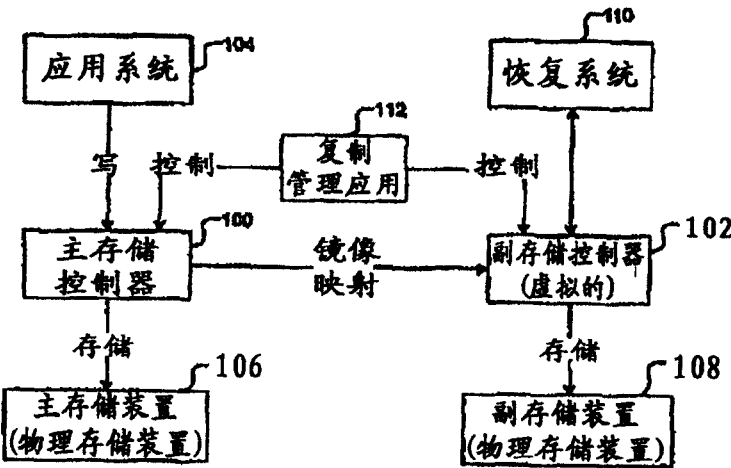


图2

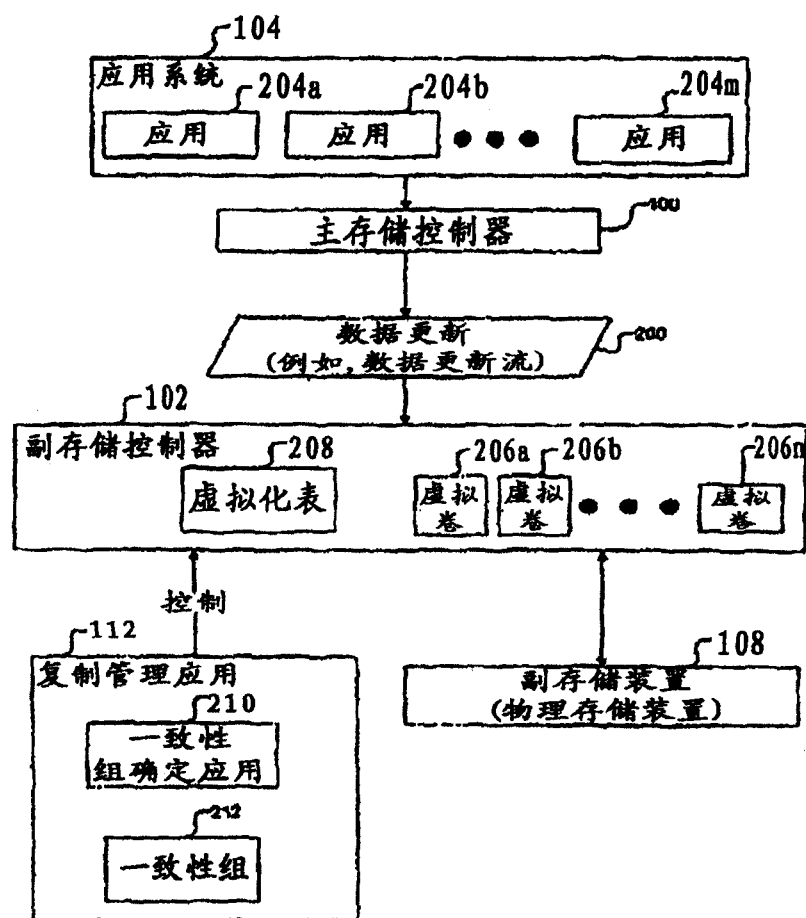


图 3

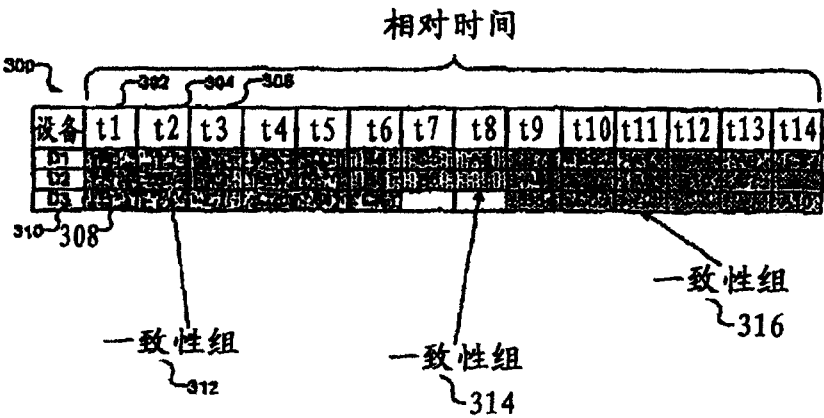


图 4

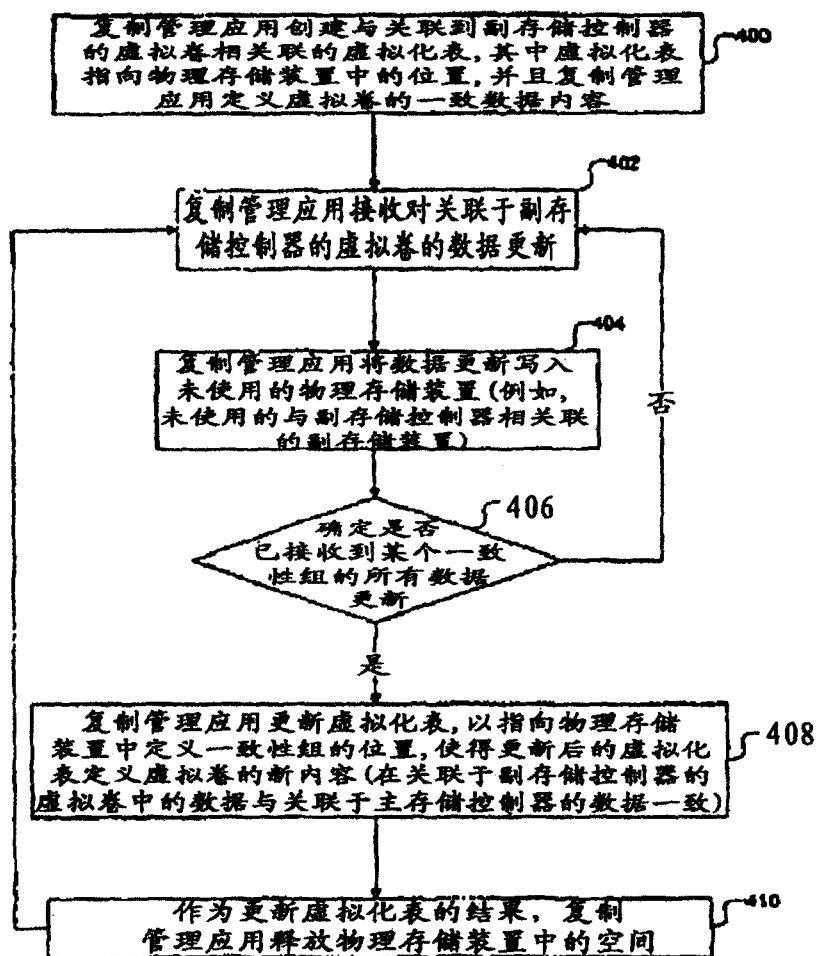


图 5

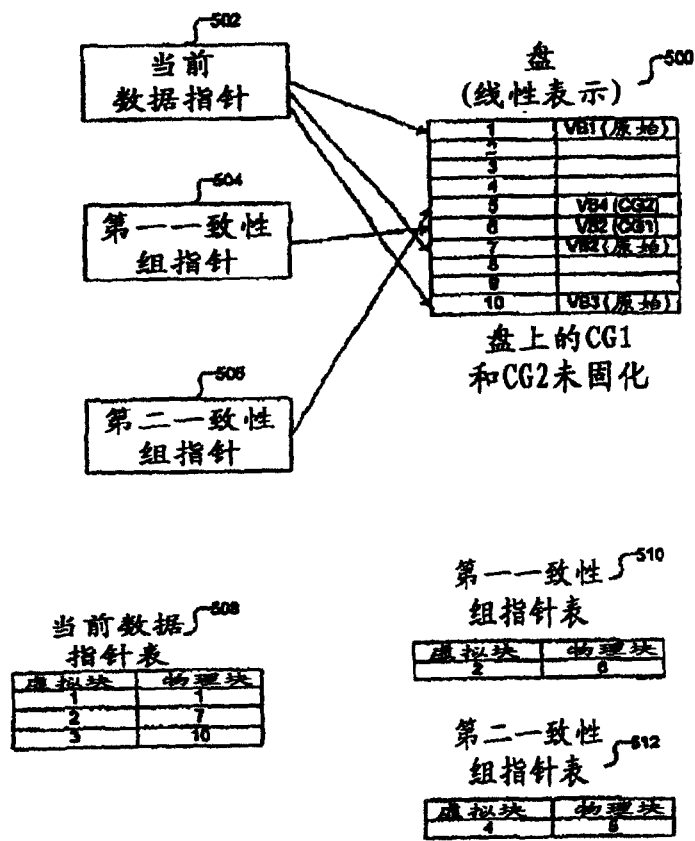


图6

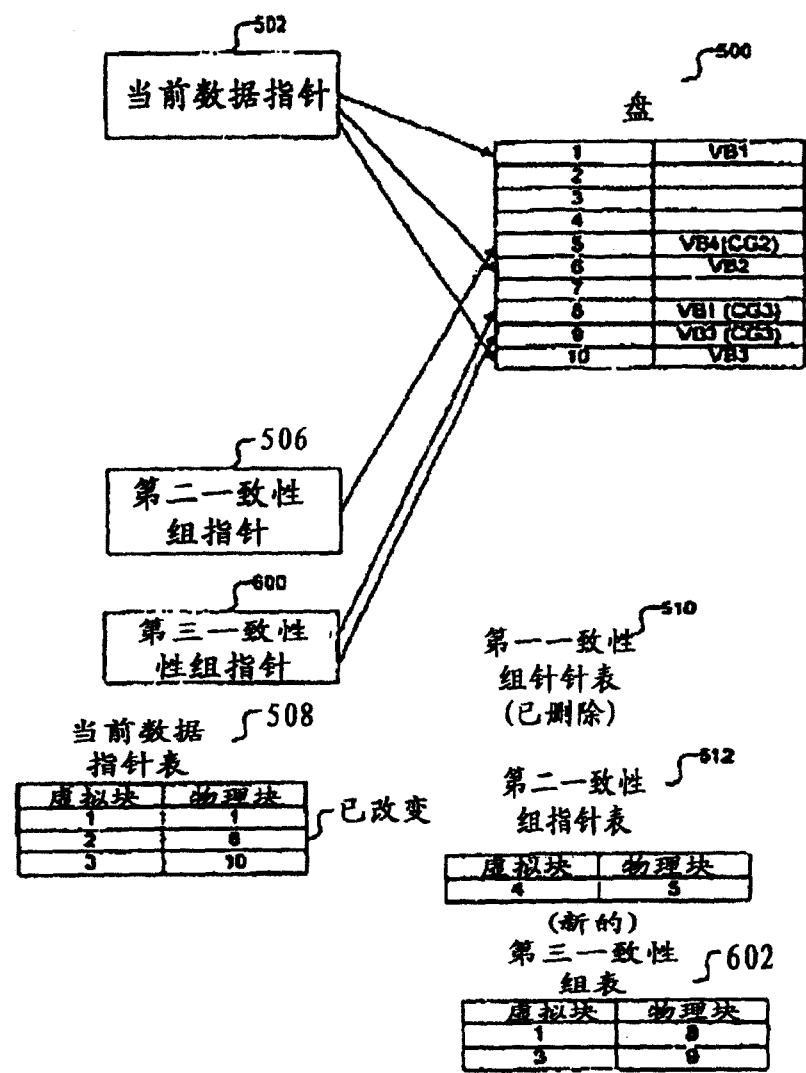


图 7

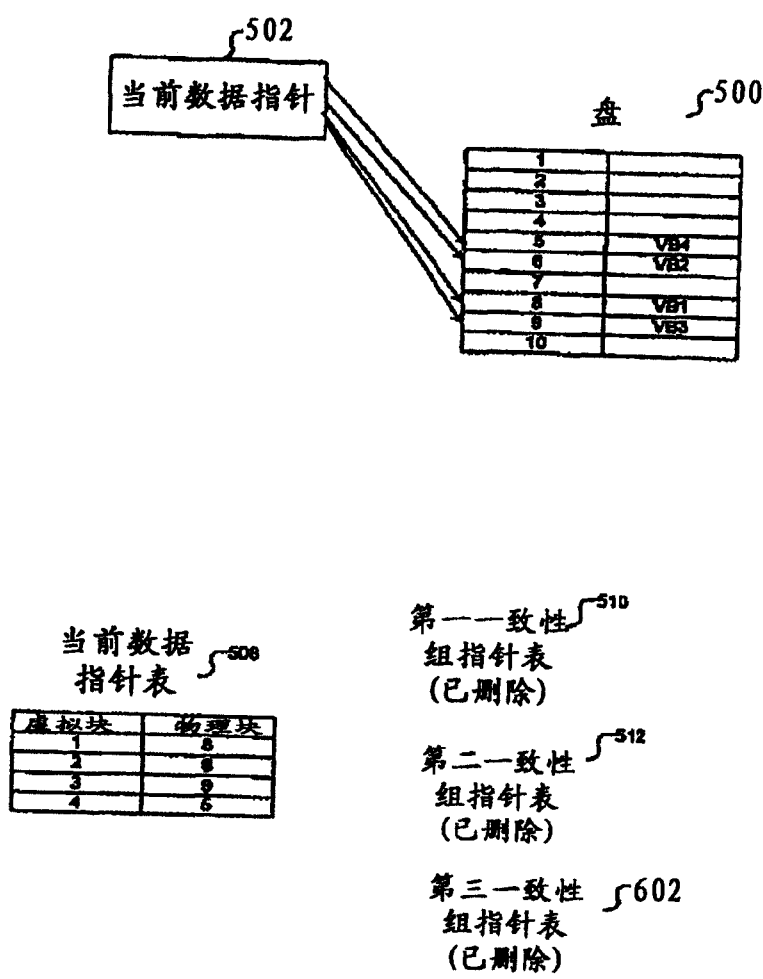


图 8

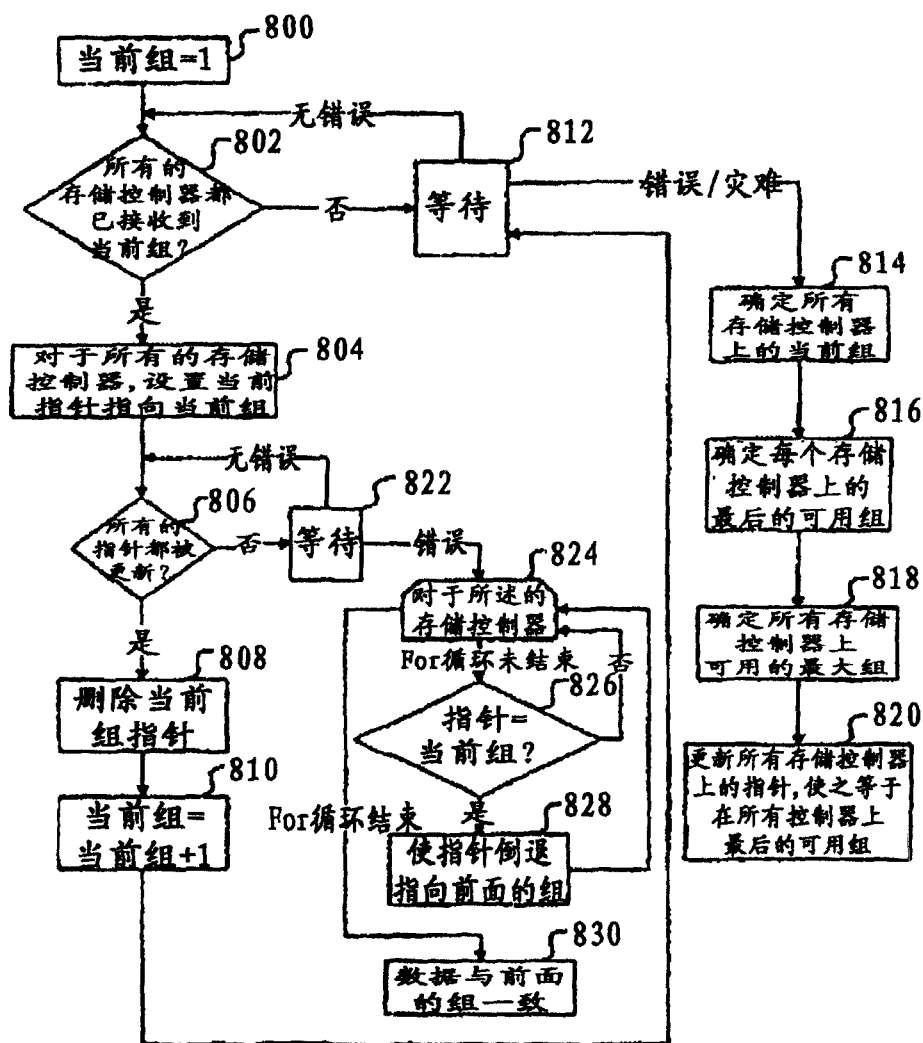


图9

