

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 12/00 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200580032766.7

[43] 公开日 2007 年 9 月 5 日

[11] 公开号 CN 101031891A

[22] 申请日 2005.9.30

[21] 申请号 200580032766.7

[30] 优先权

[32] 2004.10.1 [33] US [31] 10/956,493

[86] 国际申请 PCT/US2005/035201 2005.9.30

[87] 国际公布 WO2006/039493 英 2006.4.13

[85] 进入国家阶段日期 2007.3.28

[71] 申请人 伊姆西公司

地址 美国马萨诸塞州

[72] 发明人 瓦季姆·隆吉诺夫

本杰明·W.·约德

马克·J.·霍尔斯特德

戴维·梅雷

[74] 专利代理机构 北京金信立方知识产权代理有限公司

代理人 黄 威

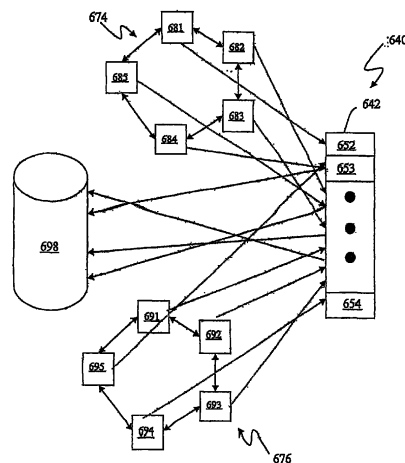
权利要求书 2 页 说明书 30 页 附图 14 页

[54] 发明名称

虚拟排序的写

[57] 摘要

处理对数据存储设备(62)进行排序的写包括：将第一数据块(56)接收到数据存储设备的高速缓存槽(642)的槽(652)中并使用散列表(702)在包含来自第一数据块的数据的高速缓存槽和数据存储设备(62)的逻辑存储设备(698)的位置之间提供关联。处理排序的写还可包括保存指向包含第一数据块的高速缓存槽的指针的循环链表(674)。处理排序的写还可包括将第二数据块(58)接收到数据存储设备(62)的高速缓存槽中并接收指示所有对应于第一数据块的数据已被传给数据存储设备(698)的消息。处理排序的写还可包括，响应于接收消息，将来自第一数据块的数据保存到逻辑存储设备(698)上。处理排序的写还可包括使用散列表(702)在包含来自第二数据块的数据的高速缓存槽(652)和数据存储设备(62)的逻辑存储设备(698)的位置之间提供关联。



- 1、处理数据存储设备的排序的写的方法，包括：
将第一数据块接收到数据存储设备的高速缓存槽中；及
使用散列表在包含来自第一数据块的数据的高速缓存槽和数据
5 存储设备的逻辑存储设备的位置之间提供关联。
- 2、根据权利要求1的方法，还包括：
将第二数据块接收到数据存储设备的高速缓存槽中；及
接收指示所有对应于第一数据块的数据均已被传给数据存储设
备的消息。
- 10 3、根据权利要求2的方法，还包括：
响应于接收所述消息，将来自第一数据块的数据保存到逻辑存储
设备上。
- 4、根据权利要求2的方法，还包括：
使用散列表在包含来自第二数据块的数据的高速缓存槽和数据
15 存储设备的逻辑存储设备的位置之间提供关联。
- 5、根据权利要求1的方法，其中散列表是指向高速缓存槽的指
针表。
- 6、根据权利要求5的方法，还包括：
响应于碰撞，散列表的项包含指向第一槽的指针，及第一槽包含
20 指向第二槽的指针。
- 7、数据存储设备，包括：
至少一包含数据的磁盘驱动器；
连到至少一磁盘驱动器的至少一主机适配器，其接收将被保存在
至少一磁盘驱动器上的数据；及
25 连到至少一磁盘驱动器及至少一主机适配器的至少一远程适配
器，其接收将被保存在数据存储设备上的数据，其中数据存储设备将
第一数据块接收到数据存储设备的高速缓存槽中，并使用散列表在包
含来自第一数据块的数据的高速缓存槽和数据存储设备的逻辑存储
设备的位置之间提供关联。

8、根据权利要求 7 的数据存储设备，其中数据存储设备保存指向包含第一数据块的高速缓存槽的指针的循环链表。

9、根据权利要求 7 的数据存储设备，其中数据存储设备将第二数据块接收到数据存储设备的高速缓存槽中，并接收指示所有对应于
5 第一数据块的数据均已传给数据存储设备的消息。

10、根据权利要求 9 的数据存储设备，其中响应于接收所述消息，数据存储设备将来自第一数据块的数据保存到逻辑存储设备上。

虚拟排序的写

发明背景

5 1. 技术领域

本发明涉及计算机存储设备，特别是涉及在存储设备之间传输数据的领域。

2. 背景技术

10 主处理机系统可使用存储设备保存和检索数据，存储设备包含多个主机接口装置（主机适配器）、磁盘驱动器、及磁盘接口装置（磁盘适配器）。这样的存储设备可由麻萨诸塞州 Hopkinton 的伊姆西公司提供，且其在授权给 Yanai 等的美国专利 5,206,939、授权给 Galtzur 等的美国专利 5,778,394、授权给 Vishlitzky 等的美国专利
15 5,845,147、及授权给 Ofek 等的美国专利 5,857,208 中公开。主机系统通过多个随其提供的通道访问存储设备。主机系统通过通道提供数据和访问控制信息给存储设备，存储设备也通过通道提供数据给主机系统。主机系统不直接寻址存储设备的磁盘驱动器，而是访问对主机系统表现为多个逻辑磁盘机的装置。逻辑磁盘机可以也可不对应于实际的磁盘驱动器。允许多个主机系统访问单个存储设备将允许主机系
20 统共享保存于其中是数据。

在某些情形下，可能希望将数据从一个存储设备复制到另一存储设备。例如，如果主机写数据到第一存储设备，则可能希望将该数据复制到位于不同位置的第二存储设备，使得当致使第一存储设备不能
25 工作的灾难出现时主机（或另一主机）可使用第二存储设备的数据恢复运行。这样的性能可由麻萨诸塞州 Hopkinton 的伊姆西公司提供的远程数据镜像（RDF）产品提供。具有 RDF 功能的第一存储设备（标注为“主存储设备”或“R1”）连接到主机。一个或多个其它存储设备（称为“辅助存储设备”或“R2”）接收由主机写到主存储设备的

数据的拷贝。主机与主存储设备直接相互作用，但主存储设备的任何数据变化均使用 RDF 自动提供给一个或多个辅助存储设备。主和辅助存储设备可由数据链路连接，数据链路如 ESCON 链路、光纤通道链路、和/或吉比特以太网链路。RDF 功能可使用在每一存储设备处提供的

5 RDF 适配器 (RA) 而得以促进。

RDF 允许同步数据传输，其中在数据从主机写到主存储设备之后，该数据通过使用 RDF 从主存储设备传输到辅助存储设备，由辅助存储设备向主存储设备确认数据的接收，主存储设备接下来将写确认提供回主机。因此，在同步模式中，主机不从主存储设备接收写确认，直

10 到对辅助存储设备的 RDF 传输已被完成并由辅助存储设备确认为止。

同步 RDF 系统的缺点在于每一写操作的等待时间将由于等待 RDF 传输的确认而增加。在主存储设备和辅助存储设备之间距离较长时，该问题更加突出。由于传输延迟，进行 RDF 传输及随后在传输完成后等待确认回馈所需要的时延可能是不可接受的。

也可能以半同步方式使用 RDF，在这种情况下，数据从主机写到主存储设备，主存储设备立即确认写，接着，同时开始将数据传输到辅助存储设备的步骤。因此，对于单一数据传输，该方案克服了以同步方式使用 RDF 的一些缺点。然而，对于数据完整性目的，半同步传输不允许主存储设备传输数据给辅助存储设备，直到先前的传输由辅

15 助存储设备确认为止。因此，与以同步方式使用 RDF 相关联的瓶颈在于由一次重复造成的简单的延迟，因为第二数据量的传输不会发生，直到先前数据的传输为辅助存储设备确认为止。

20

另一种可能是使主机以异步方式写数据到主存储设备并使主存储设备在背后将数据复制到辅助存储设备。后台复制包括顺序轮转通过主存储设备的每一磁轨，当确定特定的数据块自上次复制该数据块后已被修改时，数据块被从主存储设备传输到辅助存储设备。尽管该机制可减轻与同步及半同步数据传输方式相关的等待时间问题，但困难依然存在，因为不能保证主存储设备和辅助存储设备之间的数据一致性。如果发生问题，如主系统不能工作，则辅助系统可能以使得数

25

据不可用的无序变化结束工作。

所提出的解决该问题的方案为对称自动复制（SAR）方法，其于未决美国专利申请 10/224,918 和 10/225,021 中描述，两申请均于 2002 年 8 月 21 日申请。SAR 使用可镜像标准逻辑设备的设备（BCV）。BCV 设备还可在镜像后从其标准逻辑设备分离并可在分离后重新同步（即重新建立为镜像）到标准逻辑设备。然而，使用 SAR 方法要求大量连续分离和重新同步 BCV 设备的辅助操作。SAR 方法还使用主机控制和管理，其依赖于控制主机运行。此外，SAR 方法的实际实施的轮转时间为 20-30 分钟，因此，当 RDF 链路和/或主设备出故障时，可能丢失的数据量为 20-30 分钟的数据价值。

因此，希望具有这样一种 RDF 系统，其具有上述不同技术中的每一种的部分有益品质，同时还减少了缺点。该系统对每一主机写可呈现低等待时间性，不管主设备和辅助设备之间的距离的长短如何，并可在出现故障时提供辅助设备的一致性（可复原性）。

发明内容

根据本发明，处理对数据存储设备进行排序的写包括：将第一数据块接收到数据存储设备的高速缓存槽中并使用散列表在包含来自第一数据块的数据的高速缓存槽和数据存储设备的逻辑存储设备的位置之间提供关联。处理排序的写还可包括保存指向包含第一数据块的高速缓存槽的指针的循环链表。处理排序的写还可包括将第二数据块接收到数据存储设备的高速缓存槽中并接收指示所有对应于第一数据块的数据已被传给数据存储设备的消息。处理排序的写还可包括，响应于接收消息，将来自第一数据块的数据保存到逻辑存储设备。处理排序的写还可包括使用散列表在包含来自第二数据块的数据的高速缓存槽和数据存储设备的逻辑存储设备的位置之间提供关联。处理排序的写还包括保存指向包含第二数据块的高速缓存槽的指针的循环链表。散列表可以是指向高速缓存槽的指针表。处理排序的写还

可包括，响应于碰撞，散列表的项包含指向第一槽的指针，及第一槽包含指向第二槽的指针。

进一步根据本发明，处理对数据存储设备进行排序的写的计算机软件包括：将第一数据块接收到数据存储设备的高速缓存槽中的可执行代码，及使用散列表在包含来自第一数据块的数据的高速缓存槽和数据存储设备的逻辑存储设备的位置之间提供关联的可执行代码。计算机软件还可包括保存指向包含第一数据块的高速缓存槽的指针的循环链表的可执行代码。计算机软件还可包括将第二数据块接收到数据存储设备的高速缓存槽中的可执行代码，及接收指示所有对应于第一数据块的数据已被传给数据存储设备的消息的可执行代码。计算机软件还可包括，响应于接收消息而将来自第一数据块的数据保存到逻辑存储设备的可执行代码。计算机软件还可包括使用散列表在包含来自第二数据块的数据的高速缓存槽和数据存储设备的逻辑存储设备的位置之间提供关联的可执行代码。计算机软件还包括保存指向包含第二数据块的高速缓存槽的指针的循环链表的可执行代码。散列表可以是指向高速缓存槽的指针表。计算机软件还可包括，响应于碰撞，将指向第一槽的指针提供为散列表的项及在第一槽中提供指向第二槽的指针的可执行代码。

进一步根据本发明，数据存储设备包括至少一包含数据的磁盘驱动器、连到至少一磁盘驱动器从而接收将被保存在至少一磁盘驱动器上的数据的至少一主机适配器、及连到至少一磁盘驱动器及至少一主机适配器的至少一远程适配器，该远程适配器接收将被保存在数据存储设备上的数据，其中数据存储设备将第一数据块接收到数据存储设备的高速缓存槽中，并使用散列表在包含来自第一数据块的数据的高速缓存槽和数据存储设备的逻辑存储设备的位置之间提供关联。数据存储设备可保存指向包含第一数据块的高速缓存槽的指针的循环链表。数据存储设备可将第二数据块接收到数据存储设备的高速缓存槽中，并接收指示所有对应于第一数据块的数据均已传给数据存储设备

的消息。响应于接收所述消息，数据存储设备可将来自第一数据块的数据保存到逻辑存储设备。

附图简要说明

5 图 1 为连同在此描述的系统使用的主机、本地存储设备及远程数据存储设备的示意图。

图 2 为连同在此描述的系统使用的主机、本地存储设备及远程数据存储设备之间的数据流的示意图。

10 图 3 为根据在此描述的系统，示出用于构建和利用本地存储设备上的数据块的项目的示意图。

图 4 示出了连同在此描述的系统使用的槽的数据结构。

图 5 为根据在此描述的系统，响应于主机的写，主机适配器（HA）的操作流程图。

15 图 6 为根据在此描述的系统，将数据从本地存储设备传输到数据存储设备的流程图。

图 7 为根据在此描述的系统，示出用于构建和利用远程存储设备上的数据块的项目的示意图。

图 8 为根据在此描述的系统，由远程存储设备执行的、与从本地存储设备接收提交指示符有关的步骤的流程图。

20 图 9 为根据在此描述的系统，将所传输的数据保存在远程存储设备中的流程图。

图 10 为根据在此描述的系统，所执行的与本地存储设备增加序号有关的步骤的流程图。

25 图 11 为根据在此描述的系统的另一实施例，用于构建和利用本地存储设备上的数据块的项目的示意图。

图 12 为根据在此描述的系统的另一实施例，响应于主机的写，主机适配器（HA）的操作流程图。

图 13 为根据在此描述的系统的另一实施例，将数据从本地存储设备传输到远程存储设备的流程图。

图 14 为根据在此描述的系统，用于构造和操纵远程存储设备上的数据块的项目的示意图。

图 15 为根据在此描述的系统，用于将逻辑设备位置映射到包含远程存储设备接收的数据的槽的表的简图。

5 图 16 为根据在此描述的系统，示出了用于将逻辑设备位置映射到包含远程存储设备接收的数据的槽的表的另一实施例。

具体实施方式

参考图 1，图表 20 示出了主机 22、本地存储设备 24 和远程存储设备 26 之间的关系。主机 22 经主机适配器 (HA) 28 从本地存储设备读数据及写数据到本地存储设备，主机适配器便于主机 22 和本地存储设备 24 之间的连接。尽管图表 20 仅示出了一个主机 22 和一个 HA28，本领域一般技术人员可以意识到的是，可使用多个 HA，且一个或多个 HA 可有一个或多个主机与其连接。

15 来自本地存储设备 24 的数据经 RDF 链路 29 复制到远程存储设备 26 以使远程存储设备 26 上的数据与本地存储设备 24 上的数据一样。尽管只示出了一个链路 29，在存储设备 24、26 之间及在存储设备 24、26 之一或二者与其它存储设备 (未示出) 之间具有另外的链路是可能的。注意，在从本地存储设备 24 到远程存储设备 26 的数据传输之间可能有时延，从而远程存储设备 26 可能，在某些时间点，某些数据与本地存储设备上的数据不一样。使用 RDF 的通信在美国专利 20 5,742,792 中描述，其通过引用组合于此。

本地存储设备 24 包括第一多个 RDF 适配器单元 (RA) 30a、30b、30c，远程存储设备 26 包括第二多个 RA32a-32c。RA30a-30c、32a-32c 均连接到 RDF 链路 29 且均类似于主机适配器 28，但用于在存储设备 24、26 之间传输数据。连同 RA30a-30c、32a-32c 使用的软件将在下文 25 文中详细描述。

存储设备 24、26 可包括一个或多个磁盘，每一磁盘包含保存在每一存储设备 24、26 上的数据的不同部分。图 1 示出了存储设备 24

包括多个磁盘 33a、33b、33c，及存储设备 26 包括多个磁盘 34a、34b、34c。在此描述的 RDF 功能可被应用，使得本地存储设备 24 的磁盘 33a-33c 的至少一部分的数据可使用 RDF 复制到远程存储设备 26 的磁盘 34a-34c 的至少一部分中。存储设备 24、26 的其它数据不在存储设备 24、26 之间进行复制是可能的，因而不完全相同。

磁盘 33a-33c 中的每一个被连接到对应的磁盘适配器单元 (DA) 35a、35b、35c，其提供数据给磁盘 33a-33c 中的对应磁盘并从磁盘 33a-33c 中的对应磁盘接收数据。类似地，远程存储设备 26 的多个 DA36a、36b、36c 均用于提供数据给磁盘 34a-34c 中的对应磁盘并从磁盘 34a-34c 中的对应磁盘接收数据。在本地存储设备 24 的 DA35a-35c、HA28 和 RA30a-30c 之间存在内部数据通路。类似地，在远程存储设备的 DA36a-36c 和 RA32a-32c 之间也存在内部数据通路。注意，在其它实施例中，DA 服务于一个以上的磁盘是可能的，且一个以上的 DA 服务于一个磁盘也是可能的。

本地存储设备 24 还包括全局存储器 37，其可用于帮助在 DA35a-35c、HA28 和 RA30a-30c 之间的数据传输。存储器 37 可包含将由 DA35a-35c、HA28 和 RA30a-30c 中的一个或多个执行的任务，及包含用于从磁盘 33a-33c 中的一个或多个取回的数据的高速缓存器。类似地，远程存储设备 26 包括全局存储器 38，其可包含将由 DA36a-36c 和 RA32a-32c 中的一个或多个执行的任务，及包含用于从磁盘 34a-34c 中的一个或多个取回的数据的高速缓存器。存储器 37、38 的使用将在下文中详细描述。

本地存储设备 24 中对应于磁盘 33a-33c 的存储空间可被再分为多个卷或逻辑设备。逻辑设备可以也可不对应于磁盘 33a-33c 的物理存储空间。因而，例如，磁盘 33a 可包含多个逻辑设备，或者单个逻辑设备可横跨两个磁盘 33a、33b。类似地，远程存储设备 26 的包括磁盘 34a-34c 的存储空间可被再分为多个卷或逻辑设备，其中每一逻辑设备可以也可不对应于磁盘 34a-34c 中的一个或多个。

假设在本地存储设备 24 和远程存储设备 26 的部分之间进行映射

的 RDF 包括, 在远程存储设备 26 上建立逻辑设备, 其是对本地存储设备 24 上的逻辑设备的远程镜像。主机 22 从本地存储设备 24 上的逻辑设备读数据并写数据到该逻辑设备, 且 RDF 映射使得已修改的数据将通过使用 RA30a-30c、32a-32c 和 RDF 链路而从本地存储设备 24 传输到远程存储设备 26。在稳态工作情况下, 远程存储设备 26 上的逻辑设备包含的数据与本地存储设备 24 上的逻辑设备的数据完全一样。本地存储设备 24 上为主机 22 所访问的逻辑设备被称为“R1 卷” (或“R1”), 而远程存储设备 26 上包含 R1 卷上的数据的拷贝的逻辑设备被称为“R2 卷” (或“R2”)。因此, 主机从 R1 卷读数据并将数据写到 R1 卷, 及 RDF 处理从 R1 卷到 R2 卷的数据的自动复制和更新。

参考图 2, 其示出了从主机 22 到本地存储设备 24 和远程存储设备 26 的数据通路。如本地存储设备 24 的数据元 51 所示, 从主机 22 写到本地存储设备 24 的数据被本地保存。由主机 22 写到本地存储设备 24 的数据还被本地存储设备 24 保持并由本地存储设备 24 经链路 29 发送给远程存储设备 26。

在在此描述的系统中, 主机 22 的每一数据写 (如记录、多个记录、磁轨等的写) 均被分配以序号。序号可在与写相关联的适当数据字段中提供。在图 2 中, 主机 22 的写被示为被分配以序号 N。由主机执行的、被分配以序号 N 的所有写均被收集在一个数据块 52 中。数据块 52 代表由主机 22 执行的、几乎在同时发生的多个单独的写。

通常, 本地存储设备 24 积聚一个序号的数据块, 同时将先前积聚的数据块 (具有在前的序号) 传输给远程存储设备 26。因而, 当本地存储设备 24 正积聚来自主机的、被分配以序号 N 的写时, 已发生的、在前序号 (N-1) 的写由本地存储设备 24 经链路 29 传输给远程存储设备 26。数据块 54 代表来自主机 22 的、被分配以序号 N-1 的、尚未传输给远程存储设备 26 的写。

远程存储设备 26 从对应于被分配以序号 N-1 的写的数据块接收数据并为具有序号 N-1 的主机写建立新的数据块 56。数据可使用适当的 RDF 协议进行传输, 其确认跨链路 29 发送的数据。当远程存储

设备 26 已从数据块 54 接收所有的数据时，本地存储设备 24 发送提交信息给远程存储设备 26 以提交被分配以 $N-1$ 序号的所有数据，其对应于数据块 56。通常，一旦对应于特定序号的数据块被提交，则数据块可被写到逻辑存储设备。这在图 2 中用数据块 58 图示，其对应于被分配以序号 $N-2$ （即，主机 22 到本地存储设备 24 的写正使用的当前序号之前的第二个序号）的写。在图 2 中，数据块 58 被示出为正被写到数据元 62，其代表远程存储设备 26 的磁盘存储。因而，远程存储设备 26 正接收和积聚对应于序号 $N-1$ 的数据块 56，同时对应于在前序号（ $N-2$ ）的数据块 58 正被写到远程存储设备 26 的磁盘存储，如数据元 62 所示。在一实施例中，数据块 58 的数据因为写（但不是必需立即写）而被标记，而数据块 56 的数据则未被标记。

因此，在运行中，主机写数据到本地存储设备 24，该数据被本地保存在数据元 51 中并被积聚在数据块 52 中。一旦特定序号的所有数据均已被积聚（在本文的其它地方描述），本地存储设备 24 递增序号。对应于小于当前序号的序号的数据块 54 的数据经链路 29 从本地存储设备 24 传输到远程存储设备 26。数据块 58 对应于已由本地存储设备 24 通过发送消息给远程存储设备 26 进行提交的序号的数据。数据块 58 的数据被写到远程存储设备 26 的磁盘存储。

应注意的是，在数据块 52、54、56、58 特定之一内的写不是必须被排序。然而，如本文中其它地方所详细描述，在开始对应于序号 $N-1$ 的数据块 54、56 的任何写之前，对应于序号 $N-2$ 的数据块 58 的每一写均已开始。此外，在开始对应于序号 N 的数据块 52 的任何写之前，对应于序号 $N-1$ 的数据块 54、56 的每一写均已开始。因此，如果本地存储设备 24 和远程存储设备 26 之间通信失败，远程存储设备 26 可简单地结束最后提交的数据块（在图 2 的例子中为数据块 58）的写，并可被保证：远程存储设备 26 处的数据的状态被排序，其中数据元 62 包含在某一时间点之前开始的所有写且不包含该时间点之后开始的写。因而， $R2$ 总是包含 $R1$ 的即时拷贝且从 $R2$ 设备重新建立一致的映射是可能的。

参考图 3, 图表 70 示出了用于构建和保持数据块 52、54 的项目。标准逻辑设备 72 包含由主机 22 写入的数据并对应于图 2 的数据元 51 和图 1 的磁盘 33a-33c。标准逻辑设备 72 包含由主机 22 写到本地存储设备 24 的数据。

5 两个指针的链表 74、76 与标准逻辑设备 72 一起使用。链表 74、76 对应于可被保存在本地存储设备 24 的存储器 37 中的数据。链表 74 包含多个指针 81-85, 每一指针指向与本地存储设备 24 一起使用的高速缓存器 88 的槽。类似地, 链表 76 包含多个指针 91-95, 每一指针指向高速缓存器 88 的槽。在一些实施例中, 高速缓存器 88 可被
10 提供在本地存储设备 24 的存储器 37 中。高速缓存器 88 包含多个缓存槽 102-104, 其可连同向标准逻辑设备 72 写一起使用, 同时, 可连同链表 74、76 一起使用。

链表 74、76 中的每一个均可用于数据块 52、54 之一, 使得链表 74 可对应于序号 N 的数据块 52, 而链表 76 可对应于序号 N-1 的数据块 54。因而, 当数据由主机 22 写到本地存储设备 24 时, 数据被
15 提供给高速缓存器 88, 在某些情况下 (在本文其它地方描述), 链表 74 的适当指针被产生。应注意, 数据将不会从高速缓存器 88 去除, 直到数据被降级 (destage) 到标准逻辑设备 72 且数据不再被链表 74 的任一指针 81-85 指向为止, 如本文其它地方所述。

20 在在此的实施例中, 链表 74、76 之一被视为“活动态”, 而另一个被视为“非活动态”。因此, 当序号 N 为偶数时, 链表 74 可以是活动态, 而链表 76 是非活动态。链表 74、76 中的活动态链表处理来自主机 22 的写, 而非活动态链表对应于正从本地存储设备 24 传输到远程存储设备 26 的数据。

25 在由主机 22 写的数据通过使用链表 74、76 中的活动态链表进行积聚时 (对于序号 N), 对应于链表 74、76 中的非活动态链表的数据 (对于在前的序号 N-1) 被从本地存储设备 24 传输到远程存储设备 26。RA30a-30c 使用链表 74、76 确定数据从本地存储设备 24 传输到远程存储设备 26。

一旦对应于链表 74、76 之一中的指针的特定指针的数据已被传输给远程存储设备 26，指针中的特定指针可从链表 74、76 中的适当链表去除。此外，假如槽中的数据不再因其它目的需要（如降级到标准逻辑设备 72），数据还可被标记以从高速缓存器 88 去除（即槽可被返回到槽池以用于随后的、不相干的使用）。一机制可被用于确保数据不被从高速缓存器 88 去除，直到所有设备不再使用该数据为止。这样的机制在 1996 年 7 月 16 日授予的美国专利 5,537,568、及 2001 年 7 月 7 日申请的美国专利申请 09/850,551 中有所描述，二者均通过引用组合于此。

10 参考图 4，槽 120，与高速缓存器 88 的槽 102-104 之一类似，包括标题 122 和数据 124。标题 122 对应于系统用以管理槽 120 的内务信息。数据 124 为来自磁盘的、正被（暂时）保存在槽 120 中的相应数据。标题 122 中的信息包括指回到磁盘的指针、时间标签等。

标题 122 还包括与在此描述的系统一起使用的缓存标签 126。在此的实施例中，缓存标签 126 为 8 个字节。其中两个字节为“密码”，其表明槽 120 是否正由在此描述的系统使用。在其它实施例中，密码可以是一个字节，而其后的字节用于填充。如本文其它地方所描述的，等于特定值的两字节密码（或一字节）表明槽 120 被链表 74、76 的至少一项指向。不等于特定值的密码表明槽 120 未被链表 74、76 的
20 项指向。密码的使用将在本文其它地方描述。

缓存标签 126 还包括一两字节字段，其指明槽 120 的数据 124 的序号（如 N、N-1、N-2 等）。如本文其它地方所描述的，缓存标签 126 的序号字段可用于帮助在此描述的处理。缓存标签 126 的其余四个字节可用于指针，如本文其它地方所描述的那样。当然，序号的两个字节和指针的四个字节仅在密码等于特定值时是有效的，其中特定
25 值表明槽 120 被列表 74、76 之一中的至少一项指向。

参考图 5，流程图 140 示出了连同主机 22 执行写操作，由 HA28 执行的步骤。当然，当主机 22 执行写时，处理出现以用一般方式处理写，而不管数据是否是 R1/R2 RDF 组的一部分。例如，当主机 22

对磁盘的一部分写数据时,对缓存槽的写发生,其最后被降级到磁盘。缓存槽或可以是新的缓存槽,或可以是已经存在的、连同先前对同一磁轨进行的读和/或写操作产生的缓存槽。

处理开始于第一步骤 142,其中对应于写的槽被锁定。在在此的
5 实施例中,高速缓存器 88 的每一槽 102-104 对应于标准逻辑设备上的数据磁道。在步骤 142 锁定槽阻止另外的过程在 HA28 执行的、对应于流程图 140 的步骤的处理期间在相应的槽上运行。

在步骤 142 之后是步骤 144,其中序号 N 的值被设定。如本文别处所描述的,在步骤 144 获得的序号的值在 HA28 执行的整个写操作
10 期间均被保持,同时槽被锁定。如本文别处所描述的,序号被分配给每一写以设定写属于数据块 52、54 中的哪一个。主机 22 执行的写被分配以当前序号。单一写操作始终保持同一序号是有用的。

在步骤 144 之后是测定步骤 146,其确定缓存槽的密码字段是否有效。如上所述,在此描述的系统将密码字段设定为预先确定的值以
15 表明缓存槽已经在指针的链表 74、76 之一中。如果在测定步骤 146 确定密码字段无效(表明槽是新的且没有列表 74、76 的指针指向该槽),则控制从步骤 146 转到步骤 148,新槽的缓存标签被设定,其中将密码设定为预先确定的值、将序号字段设定为 N、将指针字段设定为空。在其它实施例中,指针字段可被设定为指向槽本身。

20 在步骤 148 之后是步骤 152,指向新槽的指针被添加到指针列表 74、76 中的活动态列表。在在此的一实施例中,列表 74、76 为循环双向链表,且新指针被以传统方式添加到循环双向链表中。当然,其它适当的数据结构也可用于管理列表 74、76。在步骤 152 之后是步骤 154,其中标志被设定。在步骤 154, RDF_WP 标志(RDF 写悬挂标志)
25 被设定以表明该槽需要通过使用 RDF 传输给远程存储设备 26。此外,在步骤 154, IN_CACHE 标志被设定以表明该槽需要被降级到标准逻辑设备 72。在步骤 154 之后是步骤 156,由主机 22 和 HA28 写的的数据被写到该槽中。在步骤 156 之后是步骤 158,该槽被解锁。在步骤 158 之后,处理结束。

如果在测定步骤 146 确定槽的密码字段有效(表明该槽已经被列表 74、76 的至少一指针指向),则控制从步骤 146 转到测定步骤 162,其确定该槽的序号字段是否等于当前序号 N。应注意,对于具有有效密码的槽的序号字段有两种有效的可能性。序号字段等于 N(当前序号)是可能的。这在槽对应于具有序号 N 的在前写的情况下发生。另一种可能性是序号字段等于 N-1。这在槽对应于具有序号 N-1 的在前写的情况下发生。序号字段的任何其它值均是无效的。因此,对于某些实施例,在步骤 162 中包括错误/有效性检查或在单独的步骤中进行错误/有效性检查是可能的。这样的错误可以任何适当的方式处理,其可包括提供消息给用户。

如果在测定步骤 162 确定槽的序号字段中的值等于当前序号 N,则不要求特殊的处理且控制从步骤 162 转到步骤 156,如上所述,数据被写到槽中。否则,如果序号字段的值为 N-1(唯一另外的有效值),则控制从步骤 162 转到步骤 164,新槽被获得。在步骤 164 获得的新槽可用于保存正被写的的数据。

在步骤 164 之后是步骤 166,来自旧槽的数据被复制到小步骤 164 获得的新槽。应注意,所复制的数据包括 RDF_WP 标志,其应在该槽被第一次创建时已被设定(步骤 154)。在步骤 166 之后是步骤 168,新槽的缓存标签被设定,其将密码字段设定为适当的值、将序号字段设定为当前序号 N、及将指针字段设定为指向旧槽。在步骤 168 之后是步骤 172,指向新槽的指针被添加到链表 74、76 中的活动态链表。在步骤 172 之后是步骤 156,如上所述,数据被写到槽中,在该例子中,其是新槽。

参考图 6,流程图 200 示出了连同 RA30a-30c 扫描链表 72、74 中的非活动态链表所执行的步骤以将 RDF 数据从本地存储设备 24 传输到远程存储设备 26。如上所述,当 N 循环正由主机使用链表 72、74 中的活动态链表写到 R1 设备时,链表 72、74 中的非活动态链表指向对应于 R1 设备的 N-1 循环的槽。

处理开始于第一步骤 202,其确定在列表 72、74 的非活动态表

中是否有任何项。当数据被传输时，相应的项被从列表 72、74 的非活动态列表中去掉。此外，新的写被提供给列表 72、74 中的活动态列表且通常不提供给列表 72、74 中的非活动态列表。因此，列表 72、74 中的非活动态列表在某些时间不包含数据是可能的（也是所希望的，如本文别处所述的）。如果在步骤 202 确定没有数据将被传输，则列表 72、74 中的非活动态列表继续被轮询直到数据可用为止。用于发送的数据变得可用同时循环切换（在本文别处描述），其中列表 72、74 中的非活动态列表变成列表 72、74 中的活动态列表，反之亦然。

10 如果在步骤 202 确定有数据可用于发送，则控制从步骤 202 转到步骤 204，其中槽被验证为是正确的。在步骤 204 执行的处理是随意的“精神健全检查”，其可包括验证密码字段是正确的及验证序号字段是正确的。如果在槽中有不正确的（意外的）数据，则可执行错误处理，其可包括通知用户该错误及可能的错误校正处理。

15 在步骤 204 之后是步骤 212，数据以传统方式经 RDF 进行发送。在在此的一实施例中，整个槽均未被传输。而是，只有槽内具有适当的镜像位设置（表明记录已改变）的记录被传输给远程存储设备 26。然而，在其它实施例中，则可能传输整个槽，假如远程存储设备 26 只写对应于具有适当镜像位设置的数据并忽略该磁轨的其它数据，其可能是也可能不是有效的。在步骤 212 之后是测定步骤 214，其确定被传输的数据是否已被 R2 设备确认。如果没有，则数据被按照从步骤 214 回到步骤 212 的流程的指示重新发送。在其它实施例中，不同的及更多有关的处理可用于发送数据及确认其接收。这样的处理可包括错误报告及在一定数量的发送数据的努力已失败之后执行的另外的处理。

20

25

一旦在测定步骤 214 确定数据已被成功发送，控制从步骤 214 转到步骤 216 以清除 RDF_WP 标志（因为数据已经 RDF 成功发送）。在步骤 216 之后是测定步骤 218，其确定该槽是否是连同向已经具有列表 72、74 的非活动态列表中的现存项的槽进行写而创建的复制槽。这

种可能性已在上面连同步骤 162、164、166、168、172 进行描述。如果在测定步骤 218 确定该槽是复制槽，则控制从步骤 218 转到步骤 222，该槽被返回到可用的槽池中（将被重新使用）。此外，槽还可被老化（或具有某些应用于其的其它适当机制）以在其它槽之前提供用于立即的重新使用，因为该槽中提供的数据对于任何其它目的均不是有效的。在步骤 222 或步骤 218（如果槽不是复制槽）之后是步骤 224，槽的标题的密码字段被清零，使得当槽被重新使用时，在图 5 的步骤 146 的测定可正确地将该槽分类为新槽。

在步骤 224 之后是步骤 226，列表 72、74 的非活动态列表中的项被移除。在步骤 226 之后，控制转回到步骤 202，如上所述，其确定在列表 72、74 的非活动态列表上是否有另外的项对应于需要被传输的数据。

参考图 7，图表 240 示出了远程存储设备 26 所使用的数据块 56、58 的产生和处理。由远程存储设备 26 接收的数据经链路 29 提供给远程存储设备 26 的高速缓存器 242。高速缓存器 242 可被提供在远程存储设备 26 的存储器 38 中。高速缓存器 242 包括多个缓存槽 244-246，每一缓存槽可被映射到标准逻辑存储设备 252 的磁轨。高速缓存器 242 类似于图 3 的高速缓存器 88 并可包含可被降级到远程存储设备 26 的标准逻辑存储设备 252 的数据。标准逻辑存储设备 252 对应于图 2 中所示的数据元 62 及图 1 中所示的磁盘 34a-34c。

远程存储设备 26 还包含一对只缓存虚拟设备 254、256。只缓存虚拟设备 254、256 对应可能保存在远程存储设备 26 的存储器 38 中的设备表。每一只缓存虚拟设备 254、256 的表的每一磁轨项（track entry）或指向标准逻辑设备 252 的磁轨或指向高速缓存器 242 的槽。只缓存虚拟设备在 2003 年 3 月 25 日申请的、题为“CACHE-ONLY 虚拟设备”的美国未决专利申请 10/396,800 中描述，其通过引用组合于此。

多个缓存槽 244-246 可与向标准逻辑设备 252 进行写一起使用，同时，也可与只缓存虚拟设备 254、256 一起使用。在在此的实施例

中，只缓存虚拟设备 254、256 的每一磁轨表项包含一空值以表明该磁轨的数据被保存在标准逻辑设备 252 的对应磁轨上。否则，每一只缓存虚拟设备的磁轨表中的项包含指向高速缓存器 242 中的槽 244-246 之一的指针。

5 每一只缓存虚拟设备 254、256 对应于数据块 56、58 之一。因而，例如，只缓存虚拟设备 254 可对应于数据块 56，而只缓存虚拟设备 256 可对应于数据块 58。在在此的实施例 中，只缓存虚拟设备 254、256 之一可被视为“活动态”，而只缓存虚拟设备 254、256 中的另一个可被视为“非活动态”。只缓存虚拟设备 254、256 中的非活动态虚
10 拟设备可对应于正从本地存储设备 24 接收的数据（即数据块 56），而只缓存虚拟设备 254、256 中的活动态虚拟设备对应于正被恢复（写）到标准逻辑设备 252 的数据。

来自本地存储设备 24 的、经链路 29 接收的数据可被放置在高速缓存器 242 的槽 244-246 之一中。只缓存虚拟设备 254、256 中的非
15 活动态虚拟设备的相应指针可被设置为指向所接收的数据。随后的具有相同序号的数据可以类似的方式进行处理。在某一时间，本地存储设备 24 提供消息，其提交所有使用同一序号发送的数据。一旦特定序号的数据已被提交，只缓存虚拟设备 254、256 中的非活动态设备变成活动态，反之亦然。在这时，来自只缓存虚拟设备 254、256 的
20 当前活动态设备的数据被复制到标准逻辑设备 252，而只缓存虚拟设备 254、256 中的非活动态设备被用于接收从本地存储设备 24 传输给远程存储设备 26 的新数据（具有新的序号）。

当从只缓存虚拟设备 254、256 中的活动态设备移除数据时（在本文别处描述），只缓存虚拟设备 254、256 中的活动态设备中的对应
25 项可被设为空值。此外，假如槽中的数据不再为其它目的所需要（如将被降级到标准逻辑设备 252），数据还可从高速缓存器 244 中移除（即该槽被返回到空闲槽池中以供后用）。一机制可用于确保数据不从高速缓存器 242 中移除，直到所有镜像（包括只缓存虚拟设备 254、256）均不再使用该数据为止。这样的机制在 1996 年 7 月 16 日授予

的美国专利 5,537,568 及 2001 年 7 月 7 日申请的美国专利申请 09/850,551 中描述,二者均通过引用组合于此。

在本文别处讨论的一些实施例中,远程存储设备 26 可包含链表 258、262,类似于本地存储设备 24 使用的链表 74、76。链表 258、
5 262 可包含这样的信息,其确定相应的只缓存虚拟设备 254、256 中已被修改的槽,其中列表 258、262 之一对应于只缓存虚拟设备 254、256 之一,及列表 258、262 中的另一列表对应于只缓存虚拟设备 254、256 中的另一个。如本文别处所描述的,列表 258、262 可用于帮助从只缓存虚拟设备 254、256 恢复数据到标准逻辑设备 252 中。

10 参考图 8,流程图 270 示出了与处理用于序号提交的数据有关的、由远程存储设备 26 执行的步骤,其中所述数据由本地存储设备 24 传输给远程存储设备 26。如本文别处所描述的,本地存储设备 24 定期递增序号。当其发生时,本地存储设备 24 结束在前序号的所有数据的传输并继而发送一用于在前序号的提交信息。

15 处理开始于第一步骤 272,其接收提交。在步骤 272 之后是测定步骤 274,其确定远程存储设备 26 的只缓存虚拟设备 254、256 中的活动态虚拟设备是否是空的。如本文别处所描述的,远程存储设备 26 的只缓存虚拟设备 254、256 中的非活动态虚拟设备用于积聚来自本地存储设备 24 的使用 RDF 发送的数据,而只缓存虚拟设备 254、
20 256 中的活动态虚拟设备被恢复到标准逻辑设备 252。

如果在测定步骤 274 确定只缓存虚拟设备 254、256 中的活动态虚拟设备不是空的,则控制从测定步骤 274 转到步骤 276,其中,在执行进一步的处理之前只缓存虚拟设备 254、256 中的活动态虚拟设备的恢复被完成。从只缓存虚拟设备 254、256 中的活动态虚拟设备
25 恢复数据将在本文别处详细描述。在处理提交并开始恢复下一序号的数据之前只缓存虚拟设备 254、256 中的活动态虚拟设备为空是有用的。

在步骤 276 之后或在步骤 274 之后(如果只缓存虚拟设备 254、256 中的活动态虚拟设备被确定为空)是步骤 278,只缓存虚拟设备

254、256 中的活动态虚拟设备被使得为非活动态。在步骤 278 之后是步骤 282，只缓存虚拟设备 254、256 中先前非活动态的虚拟设备（即在执行步骤 278 之前为非活动态的虚拟设备）被变成活动态。在步骤 278、282 交换活动态及非活动态只缓存虚拟设备 254、256 准备了只缓存虚拟设备 254、256 中的当前非活动态（及空）的虚拟设备以开始从本地存储设备 24 接收下一序号的数据。

在步骤 282 之后是步骤 284，只缓存虚拟设备 254、256 中的活动态虚拟设备被恢复到远程存储设备 26 的标准逻辑设备 252。将只缓存虚拟设备 254、256 中的活动态虚拟设备恢复到标准逻辑设备 252 将在下文中详细描述。然而，应注意，在一些实施例中，恢复过程在步骤 284 开始，但不必在步骤 284 就结束。在步骤 284 之后是步骤 286，已从本地存储设备 24 发送到远程存储设备 26 的提交被确认回到本地存储设备 24，使得本地存储设备 24 被通知：提交成功。在步骤 286 之后，处理结束。

参考图 9，流程图 300 更详细地示出了图 8 的步骤 276、284，其中远程存储设备 26 恢复只缓存虚拟设备 254、256 中的活动态虚拟设备。处理开始于第一步骤 302，指针被设定为指向只缓存虚拟设备 254、256 中的活动态虚拟设备的第一槽。指针用于重复通过只缓存虚拟设备 254、256 中的活动态虚拟设备的每一磁轨表项，每一磁轨表项均被单独处理。在步骤 302 之后是测定步骤 304，其确定只缓存虚拟设备 254、256 中的活动态虚拟设备的正被处理的磁轨是否指向标准逻辑设备 252。如果是，则没有东西需要恢复。否则，控制从步骤 304 转到步骤 306，只缓存虚拟设备 254、256 中的活动态虚拟设备的相应槽被锁定。

在步骤 306 之后是测定步骤 308，其确定标准逻辑设备 252 的相应槽是否已经在远程存储设备 26 的高速缓存器中。如果是，则控制从测定步骤 308 转到步骤 312，标准逻辑设备的槽被锁定。在步骤 312 之后是步骤 314，来自只缓存虚拟设备 254、256 中的活动态虚拟设备的数据与用于标准逻辑设备 252 的高速缓存器中的数据合并。在步

骤 314 合并数据包括用只缓存虚拟设备 254、256 中的活动态虚拟设备的新数据盖写标准逻辑设备的数据。应注意，在对记录级标志提供的实施例中，则可能简单地将新记录从只缓存虚拟设备 254、256 中的活动态虚拟设备或 (OR) 到高速缓存器中的标准逻辑设备 252 的记录。即，如果记录被交叉存取，则只需使用来自只缓存虚拟设备 254、256 中的活动态虚拟设备的已改变的记录，并将记录提供给标准逻辑设备 252 的缓存槽。在步骤 314 之后是步骤 316，标准逻辑设备 252 的槽被解锁。在步骤 316 之后是步骤 318，只缓存虚拟设备 254、256 中的活动态虚拟设备的正被处理的槽也被解锁。

如果在测定步骤 308 确定标准逻辑设备 252 的相应槽不在高速缓存器中，则控制从测定步骤 308 转到步骤 322，改变标准逻辑设备 252 的槽的磁轨项以表明标准逻辑设备 252 的槽在高速缓存器中（如可设置 IN_CACHE 标志）并需要被降级。如本文别处所述，在一些实施例中，只有具有适当镜像位设置的磁轨的记录才需要被降级。在步骤 322 之后是步骤 324，可设置磁轨的标志以表明该磁轨的数据在高速缓存器中。

在步骤 324 之后是步骤 326，标准逻辑设备 252 的槽指针被改变为指向高速缓存器中的槽。在步骤 326 之后是测定步骤 328，其确定在步骤 322、324、326 执行的操作是否已成功。在某些情况下，一称为“比较和交换”的操作可用于执行步骤 322、324、326。如果这些操作因任何原因而没有成功，则控制从步骤 328 转回到步骤 308 以重新检查标准逻辑设备 252 的相应磁轨是否在高速缓存器中。否则，如果在测定步骤 328 确定先前的操作均已成功，则控制从测定步骤 328 转到步骤 318，如上所述。

在步骤 318 之后是测定步骤 332，其确定只缓存虚拟设备 254、256 中的活动态虚拟设备的缓存槽（其正被恢复）依然在使用。在某些情况下，只缓存虚拟设备 254、256 中的活动态虚拟设备的槽可能依然正由另一镜像使用。如果在测定步骤 332 确定只缓存虚拟设备的槽正由另一镜像使用，则控制从测定步骤 332 转到步骤 334，释放槽

以为其它过程使用（如恢复到可用槽池中，如本文别处所述）。在步骤 334 之后是步骤 336，指向下一槽以处理只缓存虚拟设备 254、256 中的活动态虚拟设备的下一槽。应注意，步骤 336 还可从测定步骤 332 到达，如果在步骤 332 确定只缓存虚拟设备 254、256 中的活动态虚拟设备仍然由另一镜像使用的话。应注意，如果在步骤 304 确定对于正处理的槽，只缓存虚拟设备 254、256 中的活动态虚拟设备指向标准逻辑设备 252，步骤 336 也可从测定步骤 304 到达。在步骤 336 之后是测定步骤 338，其确定只缓存虚拟设备 254、256 中的活动态虚拟设备是否有更多的槽要被处理。如果不是，则处理结束。否则，控制从测定步骤 338 转回到步骤 304。

在另一实施例中，可能对所接收的数据块 56 建立已修改的槽的列表，数据块 56 对应于远程存储设备 26 上的 N-1 循环，所述列表如图 7 中所示的列表 258、262。当数据被接收时，远程存储设备 26 建立已修改的槽的链表。所建立的链表可以是循环的、线性的（具有 NULL 终止）、或任何其它适当的设计。链表接着可用于恢复只缓存虚拟设备 254、256 中的活动态虚拟设备。

图 9 的流程图 300 示出了两个二选一的通路 342、344，其图示了使用已修改的槽的列表的实施例的实施。在步骤 302，指针（用于反复通过已修改的槽的列表）被使得指向列表的第一单元。在步骤 302 之后是步骤 306，其由另一通路 342 到达。在使用已修改的槽的列表的实施例中，不需要测定步骤 304，因为列表上没有槽应指向标准逻辑设备 252。

在步骤 306 之后，如上以先前实施例所述，处理继续，除了步骤 336 为遍历已修改的槽的列表而不是指向 COVD 中的下一槽以外。类似地，在步骤 338 的测试确定指针是否在列表的末尾（或在循环链表的情况下回到开始）。同样，如果在步骤 338 确定有更多的槽要处理，则控制从步骤 338 转到步骤 306，如另一通路 344 所示。如上所述，对于使用已修改的槽的列表的实施例，步骤 304 可被排除。

参考图 10，流程图 350 示出了连同本地存储设备 24 增加序号执

行的步骤。处理开始于第一步骤 352，在增加序号之前本地存储设备 24 等待至少 M 秒。在在此的实施例中，M 为 30，当然，M 可以为任何数字。如果存储设备 24、26 之间的通信中断，较大的 M 值增加了可能丢失的数据的量。然而，较小的 M 值可能因更频繁地递增序号而增加内务操作的总量。

在步骤 352 之后是测定步骤 354，其确定是否本地存储设备 24 的所有 HA 已设置位，其表明 HA 已完成在前序号的所有 I/O。当序号改变时，每一 HA 注意到该改变并设置位，其表明在前序号的所有 I/O 完成。例如，如果序号从 N-1 变为 N，HA 将在 HA 已完成序号 N-1 的所有 I/O 时设置位。注意，在某些情况下，HA 的单一 I/O 可能需要很长时间并可能在序号已改变之后仍然在进行之中。还应注意，对于某些系统，可使用不同的机制来确定所有 HA 已完成它们的 N-1 I/O。不同的机制可包括检查存储器 37 中的设备表。

如果在测定步骤 354 确定来自在前序号的 I/O 已被完成，则控制从步骤 354 转到测定步骤 356，其确定列表 74、76 中的不活动列表是否是空的。应注意，可不进行序号转换，除非及直到对应于列表 74、76 的不活动列表的所有数据已使用 RDF 协议全部从本地存储设备 24 传输到远程存储设备 26 为止。一旦列表 74、76 中的不活动列表被确定为空，则控制从步骤 356 转到步骤 358，在前序号的提交从本地存储设备 24 发送到远程存储设备 26。如上所述，远程存储设备 26 接收特定序号的提交信息将使得远程存储设备 26 开始恢复对应于序号的数据。

在步骤 358 之后是步骤 362，对列表 74、76 中的不活动列表的数据复制被暂停。如本文别处所述，不活动列表被扫描以将相应的数据从本地存储设备 24 发送到远程存储设备 26。暂停复制数据直到序号转换完成为止是有用的。在在此的实施例中，暂停通过发送消息给 RA 30a-30c 实现。然而，本领域一般技术人员应该意识到的是，对于使用其它组件来帮助使用在此描述的系统发送数据的实施例，暂停复制可通过发送适当的消息/指令给其它组件实现。

在步骤 362 之后是步骤 364，其增加序号。在步骤 364 之后是步骤 366，在测定步骤 354 中使用的 HA 的位均被清零，使得这些位可连同序号的增加而被再次设置。在步骤 366 之后是测定步骤 372，其确定远程存储设备 26 是否已确认在步骤 358 发送的提交信息。确认提交信息已在上面结合图 8 进行了描述。一旦确定远程存储设备 26 已确认在步骤 358 发送的提交信息，控制从步骤 372 转到步骤 374，在步骤 362 提供的暂停复制设定被解除，使得复制可重新开始。在步骤 374 之后，处理结束。注意，从步骤 374 回到步骤 352 以开始新的循环从而连续增加序号是可能的。

10 使用 R1 设备上的 COVD 来收集与活动态数据和非活动态数据块相关的槽也是可能的。在这种情况下，正象 R2 设备处那样，一 COVD 可与非活动态序号相关联，而另一 COVD 可与活动态序号相关联。这将在下面讨论。

参考图 11，图表 400 示出了用于构建和保持数据块 52、54 的项目。标准逻辑设备 402 包含由主机 22 写的数据并对应于图 2 的数据单元 51 和图 1 的磁盘 33a-33c。标准逻辑设备 402 包含由主机 22 写到本地存储设备 24 的数据。

两个只缓存虚拟设备 404、406 可与标准逻辑设备 402 一起使用。只缓存虚拟设备 404、406 对应可能保存在本地存储设备 24 的存储器 20 37 中的设备表。每一只缓存虚拟设备 404、406 的表的每一磁轨项或指向标准逻辑设备 402 的磁轨或指向连同本地存储设备 24 使用的高速缓存器 408 的槽。在某些实施例中，高速缓存器 408 可被提供在本地存储设备 24 的存储器 37 中。

高速缓存器 408 包含多个缓存槽 412-414，其可连同对标准逻辑设备 402 进行写一起使用，同时，还可连同只缓存虚拟设备 404、406 一起使用。在在此的实施例中，只缓存虚拟设备 404、406 的每一磁轨表项包含一空值以指向标准逻辑设备 402 的相应磁轨。否则，每一只缓存虚拟设备 404、406 的磁轨表中的项包含指向高速缓存器 408 中的槽 412-414 之一的指针。

每一只缓存虚拟设备 404、406 可用于数据块 52、54 之一，从而，例如，只缓存虚拟设备 404 可对应于数据块 52（序号 N），而只缓存虚拟设备 406 可对应于数据块 54（序号 N-1）。因此，当数据由主机 22 写到本地存储设备 24 时，数据被提供给高速缓存器 408 且只缓存虚拟设备 404 的适当指针被调整。应注意，数据将不从高速缓存器 408 中移除，直到数据被降级到标准逻辑设备 402 为止，且数据还由只缓存虚拟设备 404 释放，如本文别处所述。

在在此的实施例中，只缓存虚拟设备 404、406 之一被视为“活动态”，而另一只缓存虚拟设备被视为“非活动态”。因而，例如，当序号 N 为偶数时，只缓存虚拟设备 404 可以为活动态，而只缓存虚拟设备 406 为非活动态。只缓存虚拟设备 404、406 中的活动态虚拟设备处理来自主机 22 的写，而只缓存虚拟设备 404、406 中的非活动态虚拟设备对应于正从本地存储设备 24 传输到远程存储设备 26 的数据。

当主机 22 所写的数据使用只缓存虚拟设备 404、406 中的活动态虚拟设备进行积聚时（对于序号 N），对应于只缓存虚拟设备 404、406 中的非活动态虚拟设备的数据（对于在前序号 N-1）从本地存储设备 24 传输到远程存储设备 26。对于本实施例及有关的实施例，本地存储设备的 DA 35a-35c 处理对只缓存虚拟设备 404、406 的不活动态虚拟设备的扫描以将复制请求发送给一个或多个 RA 30a-30c 从而将数据从本地存储设备 24 传输到远程存储设备 26。因而，上述与暂停和重新开始复制有关的步骤 362、374 可包括提供消息/指令给 DA 35a-35c。

一旦数据已被传输给远程存储设备 26，只缓存虚拟设备 404、406 的非活动态虚拟设备中的对应项可被设定为空值。此外，如果槽中的数据不再因另外的目的所需要（如要被降级到标准逻辑设备 402），数据还可从高速缓存器 408 中移除（即槽被返回到槽池中以供后用）。一机制可用于确保数据不被从高速缓存器 408 移除，直到所有镜像（包括只缓存虚拟设备 404、406）不再使用该数据为止。这样的机

制在 1996 年 7 月 16 日授予的美国专利 5,537,568 及 2001 年 7 月 7 日申请的美国专利申请 09/850,551 中描述,二者均通过引用组合于此。

参考图 12,流程图 400 示出了连同主机 22 执行写操作由 HA28 执行的步骤,其是针对两个 COVD 均由 R1 设备使用以提供在此描述的系统的实施例。处理开始于第一步骤 442,对应于写的槽被锁定。在此的实施例中,高速缓存器 408 的每一槽 412-414 对应于标准逻辑设备 402 上的数据磁轨。在步骤 442 锁定槽可在 HA28 执行的处理期间阻止另外的过程在相应的槽上运行,所述 HA28 执行的处理对应于流程图 440 的步骤。

在步骤 442 之后为步骤 444,序号 N 的值被设定。正象使用列表而不是在 R1 侧上使用 COVD 的实施例那样,在步骤 444 获得的序号的值在 HA28 执行的整个写操作期间均被保持,且槽被锁定。如本文别处所述,序号被分配给每一写以设定数据块 52、54 之一属于哪一写。主机 22 执行的写被分配以当前序号。单一写操作始终保持相同的序号是有用的。

在步骤 444 之后是测定步骤 446,其确定只缓存虚拟设备 404、406 中的非活动态虚拟设备是否已经指向在步骤 442 锁定的槽。这可能在序号小于当前序号时如果对同一槽的写已被提供的情况下发生。对应于在前序号的写的的数据可能尚未被传输给远程存储设备 26。

如果在测定步骤 446 确定只缓存虚拟设备 404、406 中的非活动态虚拟设备没有指向该槽,则控制从测定步骤 446 转到另一测定步骤 448,其确定只缓存虚拟设备 404、406 中的活动态虚拟设备是否指向该槽。如果在序号与当前序号相同时先前已有对该槽的写,则只缓存虚拟设备 404、406 中的活动态虚拟设备指向该槽是可能的。如果在测定步骤 448 确定只缓存虚拟设备 404、406 中的活动态虚拟设备没有指向该槽,则控制从测定步骤 448 转到步骤 452,为数据获得新槽。在步骤 452 之后是步骤 454,只缓存虚拟设备 404、406 中的活动态虚拟设备被使得指向该槽。

在步骤 454 之后或在步骤 448 之后（如果只缓存虚拟设备 404、406 的活动态虚拟设备指向该槽）是步骤 456，设置标志。在步骤 456，RDF_WP 标志（RDF 写未决标志）被设定以表明该槽需要使用 RDF 传输给远程存储设备 26。此外，在步骤 456，IN_CACHE 标志被设定以表明该槽需要被降级到标准逻辑设备 402。应注意，在某些情况下，如果只缓存虚拟设备 404、406 的活动态虚拟设备已经指向该槽（如在步骤 448 确定），则 RDF_WP 和 IN_CACHE 标志在执行步骤 456 前已被设定是可能的。然而，在步骤 456 设置标志确保标志被正确地设置，而不管先前的状态如何。

在步骤 456 之后是步骤 458，指向该槽的磁轨表中的间接标志被清零，以表明有关数据被提供在该槽中而不是在间接指向其的另外的槽中。在步骤 458 之后是步骤 462，正由主机 22 和 HA28 写的的数据被写到该槽中。在步骤 462 之后是步骤 464，其解锁槽。在步骤 464 之后，处理结束。

如果在测定步骤 446 确定只缓存虚拟设备 404、406 中的非活动态虚拟设备指向该槽，则控制从步骤 446 转到步骤 472，其获得新槽。在步骤 472 获得的新槽可用于只缓存虚拟设备 404、406 的非活动态虚拟设备以实现 RDF 传输，同时旧槽可与只缓存虚拟设备 404、406 的活动态虚拟设备相关联，如下所述。

在步骤 472 之后是步骤 474，旧槽的数据被复制到在步骤 472 获得的新槽。在步骤 474 之后是步骤 476，（上述的）间接标志被设定以表明只缓存虚拟设备 404、406 的非活动态虚拟设备的磁轨表项指向旧槽及表明数据在旧槽所指向的新槽中。因此，在步骤 476 设置间接标志影响只缓存虚拟设备 404、406 的非活动态虚拟设备的磁轨表从而导致磁轨表项表明数据在新槽中。

在步骤 476 之后是步骤 478，新槽中的记录的镜像位被调整。当数据在步骤 474 从旧槽复制到新槽时，被复制的任何本地镜像位均被清零，因为新槽的目的只是简单地实现对只缓存虚拟设备的非活动态虚拟设备的 RDF 传输。旧槽将用于处理任何本地镜像。在步骤 478 之

后是步骤 462，数据被写到槽中。在步骤 462 之后是步骤 464，解锁槽。在步骤 464 之后，处理结束。

参考图 13，流程图 500 示出了连同本地存储设备 24 将数据块 54 传输给远程存储设备 26 所执行的步骤。传输实质上包括：对只缓存
5 虚拟设备 404、406 的非活动态虚拟设备先前在活动态时的反复操作期间已写到那里的磁轨，扫描只缓存虚拟设备 404、406 的非活动态虚拟设备。在该实施例中，本地存储设备 24 的 DA35a-35c 扫描只缓存虚拟设备 404、406 的非活动态虚拟设备以通过一个或多个 RA30a-30c、使用 RDF 协议将需要传输的数据复制到远程存储设备 26。

10 处理开始于第一步骤 502，指向只缓存虚拟设备 404、406 的非活动态虚拟设备的第一磁轨以开始通过所有磁轨的反复过程。在第一步骤 502 之后是测定步骤 504，其确定 RDF_WP 标志是否已被设定。如本文别处所述，RDF_WP 标志被用于表明槽（磁轨）包含需要经 RDF 链路传输的数据。被设置的 RDF_WP 标志表明槽（磁轨）中的至少部
15 分数据将使用 RDF 传输。在在此的实施例中，整个槽均不被传输。而是，只有槽内具有适当的镜像位设置（表明记录已改变）的记录才被传输给远程存储设备 26。然而，在其它实施例中，假如远程存储设备 26 只写对应于具有适当镜像位设置的记录的数据并忽略磁轨的其它数据，则可能传输整个槽，其可以是也可不是有效的。

20 如果在测定步骤 504 确定被处理的缓存器具有 RDF_WP 标志设置，则控制从步骤 504 转到测定步骤 505，其确定槽是否包含数据或确定槽是否是指向包含相关数据的另一槽的间接槽。在某些情况下，槽可能不包含对应于槽的磁盘部分的数据。即槽可能是指向包含数据的另一槽的间接槽。如果在步骤 505 确定槽是间接槽，则控制从步骤 505
25 转到步骤 506，获取数据（从间接槽指向的槽获取）。因而，如果槽是直接槽，正由 RDF 发送的数据被保存在该槽中，而如果槽是间接槽，则正由 RDF 发送的数据被保存在由间接槽所指向的另一槽中。

在步骤 506 之后或步骤 505（如果槽是直接槽）是步骤 507，正被发送的数据（直接或间接从槽）由 DA 35a-35c 之一复制以使用 RDF

协议从本地存储设备 24 发送到远程存储设备 26。在步骤 507 之后是测定步骤 508，其确定远程存储设备 26 是否已确认数据的接收。如果否，则控制从步骤 508 转回到步骤 507 以重新发送数据。在其它实施例中，不同的及更多有关的处理可用于发送数据并确认其接收。这样的处理可包括错误报告及其它处理，其是在一定量的发送数据的尝试失败之后执行的处理。

一旦在测定步骤 508 确定数据已被成功发送，控制从步骤 508 转到步骤 512 以清零 RDF_WP 标志（因为数据已经 RDF 成功发送）。在步骤 512 之后是步骤 514，适当的镜像标志被清零以表明至少 RDF 镜像（R2）不再需要数据。在在此的实施例中，作为槽（磁轨）的一部分的每一记录具有各自的镜像标志，其表明哪些镜像使用特定的记录。R2 设备是每一记录的镜像之一且是对应于 R2 设备的标志在步骤 514 被清零。

在步骤 514 之后是测定步骤 516，其确定正被处理的磁轨的任意记录是否具有任意其它镜像标志设置（对于其它镜像设备）。如果否，则控制从步骤 516 转到步骤 518，其释放槽（即不再被使用）。在一些实施例中，未使用的槽可被保持在可用槽池中。应注意，对于槽的一些记录，如果另外的标志仍被设置，则可能意味着记录需要被降级到标准逻辑设备 402 或正由一些其它镜像（包括另一 R2 设备）使用。在步骤 518 之后，或在步骤 516 之后（如果更多的镜像标志存在）是步骤 522，用于反复通过只缓存虚拟设备 404、406 的非活动态虚拟设备的每一磁轨项的指针被使得指向下一磁轨。在步骤 522 之后是测定步骤 524，其确定只缓存虚拟设备 404、406 中的非活动态虚拟设备是否有更多的磁轨要被处理。如果否，则处理结束。否则，控制转回到测定步骤 504，如上所述。注意，如果确定对于正被处理的磁轨 RDF_WP 标志未被设置，步骤 522 还可从测定步骤 504 到达。

在于此描述的系统的另一实施例中，可能不像图 7 的图 240 中所示的那样将 COVD 用于 R2 设备。也就是说，可能不在 R2 使用 COVD 的情况下实现异步数据的 R2 接收。

参考图 14, 图 640 示出了高速缓存 642, 其被提供在接收数据的远程存储设备 26 中。高速缓存 642 包括多个槽 652-654, 从本地存储设备 24 接收的异步数据放于其中。还示出了第一循环链表 674 和第二循环链表 676, 其包含指向高速缓存 642 的槽 652-654 的指针。

- 5 因此, 例如, 循环链表 674 包括多个指针 681-685, 每一指针指向高速缓存 642 中的槽 652-654 之一。类似地, 循环链表 676 包括多个指针 691-695, 每一指针指向高速缓存 642 中的槽 652-654 之一。标准的逻辑设备 698 也被映射到高速缓存 642 的部分。

- 10 在于此的实施例中, 链表 674、676 之一对应于非当前数据块(如图 2 中所示的数据块 56), 而链表 674、676 中的另一个对应于当前数据块(如图 2 中所示的数据块 58)。接收的数据使用数据块中的非当前数据块累积, 而数据块中的当前数据块用于将数据保存在标准逻辑设备 698, 如本说明书别处结合图 7 的图 240 及相应的文本所述。因而, 当新的数据到达时, 其被放在高速缓存 642 中, 新指针被添加
- 15 到循环链表 674、676 中在数据被接收时对应于非当前数据块的那一循环链表。

- 在一些情形下, 能够确定标准逻辑设备 698 (或任何其它逻辑设备)的一部分是否在对应于所接收数据的高速缓存 642 中具有与其相关联的槽是有用的。当然, 总是可能遍历链表 674、676 以确定在高速缓存 642 中是否有相应的槽。然而, 如果有一种使用特定设备、柱面、及磁头值确定在高速缓存 642 的槽 652-654 中有相应的槽等待被降级到所述设备的方法, 则这将更有用。
- 20

- 参考图 15, 图 700 示出了包含多个项 704-706 的散列表 702。在于此的实施例中, 项 704-706 中的每一项或包含空值指针或指向对应于已被接收但尚未保存在标准逻辑设备 698 (或另一标准逻辑设备)
- 25 上的数据的高速缓存槽 652-654 中之一。表 702 使用执行数学运算的散列函数进行索引, 其使用设备、柱面及磁头的特定值在表 702 内产生索引从而找到相应的项。因而, 当数据由 R2 设备接收时, 散列函数被应用于设备、柱面及磁头以找到其在表 702 内的索引值, 之后,

指针被写到项 704-706 中指向高速缓存 642 中的特定槽 652-654 的那一项。一旦所接收的数据被适当地降级到标准逻辑设备 698（或另一设备），项 704-706 中的相应项被设为空值，如此，散列表 702 使能够快速确定标准逻辑设备的特定部分是否对应于已接收但尚未被降级的数据。对于在此描述的系统，任何适当的散列函数可用于产生表 702 中的索引。

在一些情况下，使用特定设备、柱面及磁头值在表 702 内产生的索引与用不同的设备、柱面及磁头值产生的索引一样是可能的。这被称为“碰撞”。在出现碰撞的情况下，表 702 中的第二项对应于所提供的相同索引，第二项被链接到第一项，从而特定索引将对应于一个以上的项。这由链接到表 702 的元素 705 的元素 708 说明。因而，第一设备、柱面及磁头被散列以产生并索引到项 705，同时不同的设备、柱面及磁头被输入到散列函数以产生相同的索引值。在于此的实施例中，项 705 用于指向高速缓存 642 中对应于第一设备、柱面及磁头的数据，而项 708 用于指向高速缓存 642 中对应于第二设备、柱面及磁头的数据。当然，当数据被降级到适当的设备时，项 705、708 中的相应项可从表 700 中删除。

应注意，任何数量的项可对应于单一索引，从而，例如，如果导致设备、柱面及磁头的三组分开的值产生相同索引的碰撞出现时，则在表 702 内的特定索引处将有三个（或更多个）项链接在一起。还应注意，其它适当的技术也可用于处理碰撞，包括提供另外的表（如第二表、第三表、第四表等）。

参考图 16，图 720 示出了散列表 722 的另一实施例，其包含多个项 724-726。图 16 的实施例与图 15 的实施例类似，仅有如在此所述的很少区别。项 724-726 中的每一项或包含空值指针或指向图 720 中所示的高速缓存槽 728、732、734 之一，所述缓存槽对应于已被接收但尚未保存在标准逻辑设备 698（或另一标准逻辑设备）上的数据。表 722 使用执行数学运算的散列函数进行索引，其使用设备、柱面及磁头的特定值在表 722 内产生索引从而找到相应的项。因而，当数据

由 R2 设备接收时，散列函数被应用于设备、柱面及磁头以找到其在表 722 内的索引值，之后，指针被写到项 724-726 中指向特定槽 728、732、734 的那一项。一旦所接收的数据被适当地降级到标准逻辑设备 698（或另一设备），项 724-726 中的相应项被适当调整。如此，

5 散列表 722 使能快速确定标准逻辑设备的特定部分是否对应于已接收但尚未被降级的数据。对于在此描述的系统，任何适当的散列函数可用于产生表 722 中的索引。

对于图 16 中所示的实施例，在出现碰撞的情况下，由表项指向的第一槽指向导致碰撞的第二槽。因而，例如如果槽 732 和槽 736 导致在表项 725 的碰撞，表项 725 指向槽 732，同时槽 732 指向槽 736。

10 因而，当随后添加槽时，碰撞不会导致表 722 的任何变化，因为添加随后的槽简单地包括改变先前槽的指针值。当然，任何数量的槽可对应于单一的表项。

应注意，任何数量的项可对应于单一索引，从而，例如，如果导致设备、柱面及磁头的三组分开的值产生相同索引的碰撞出现时，则在表 722 内的特定索引处将有三个（或更多个）项链接在一起。还应注意，其它适当的技术也可用于处理碰撞，包括提供另外的表（如第二表、第三表、第四表等）。

15

在本发明已连同多个实施例进行公开的同时，其修改对本领域技术人员而言应是显而易见的。因此，本发明的实质和范围在下述权利要求中提出。

20

20 →

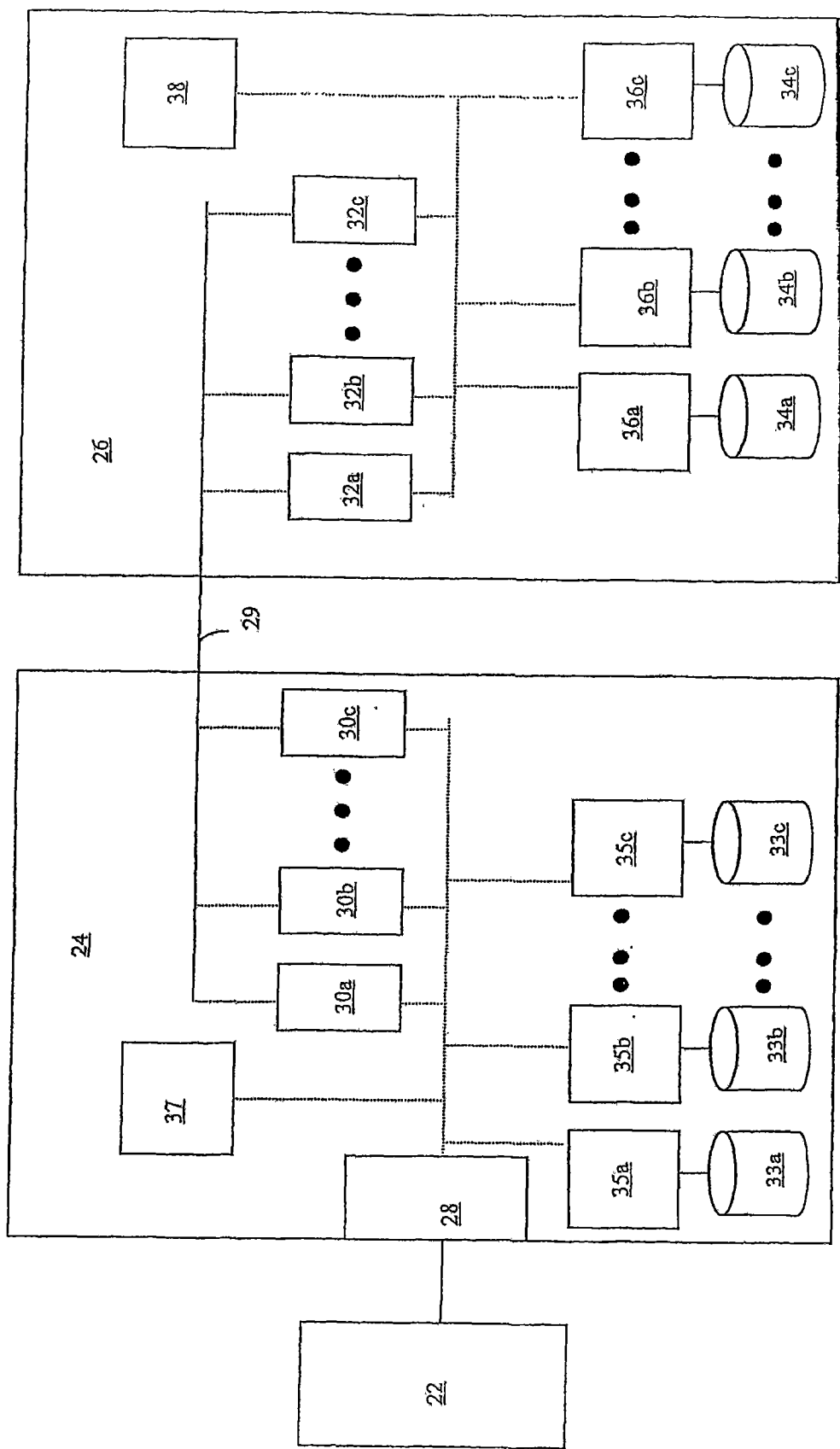


图 1

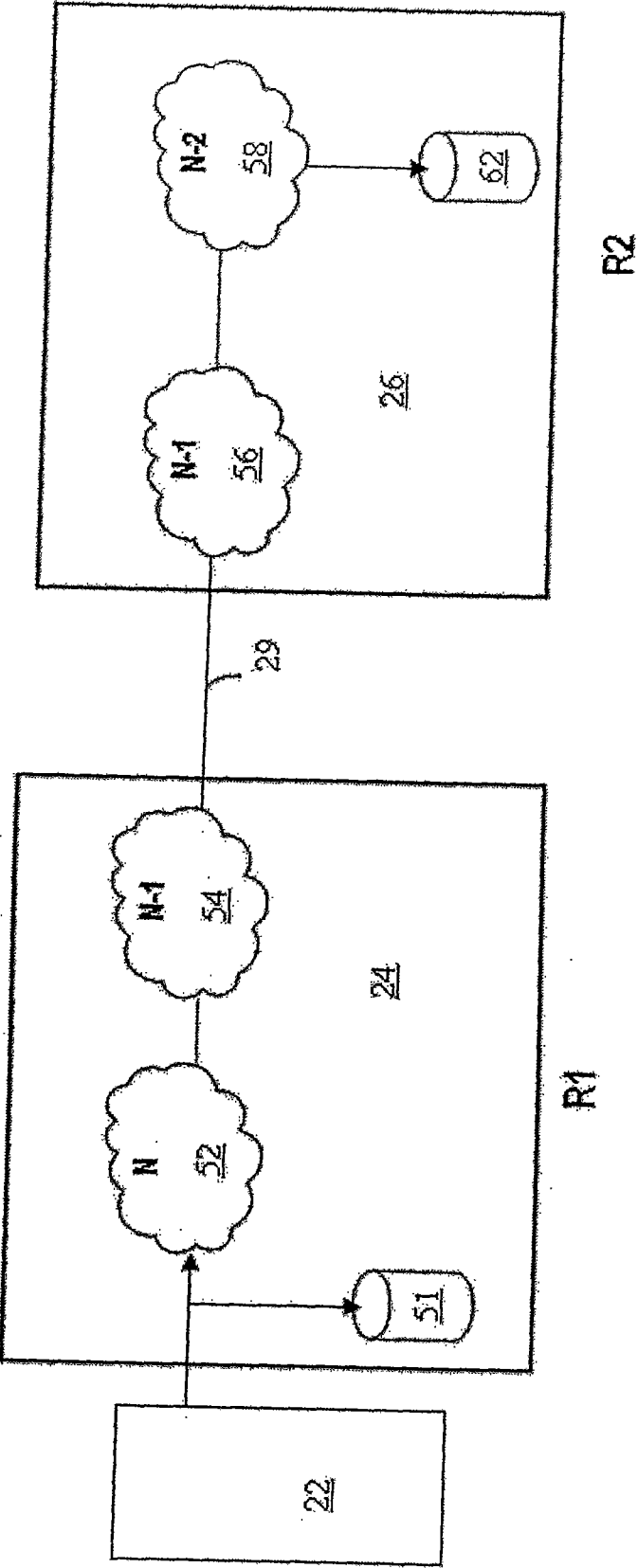


图2

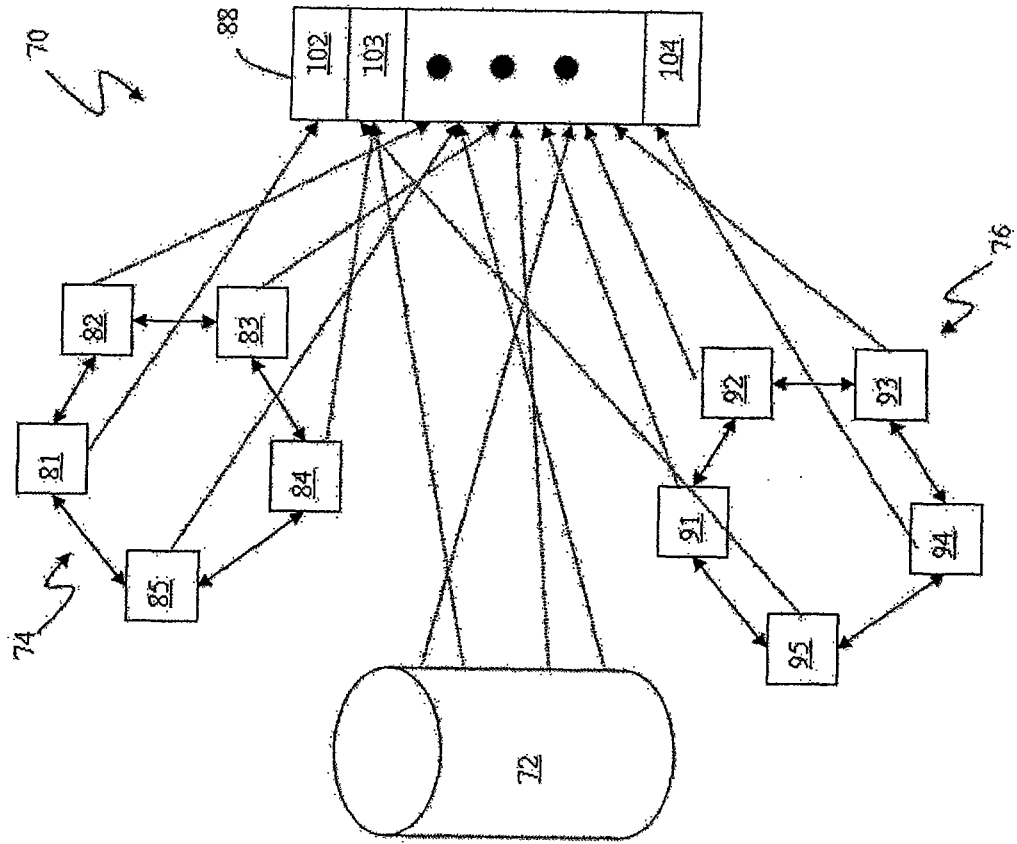


图3

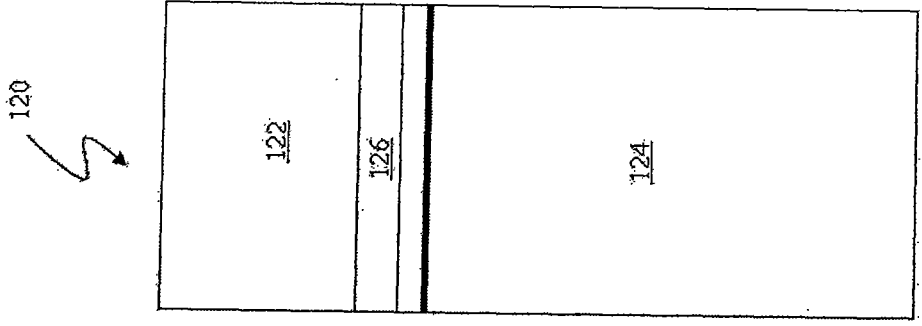


图4

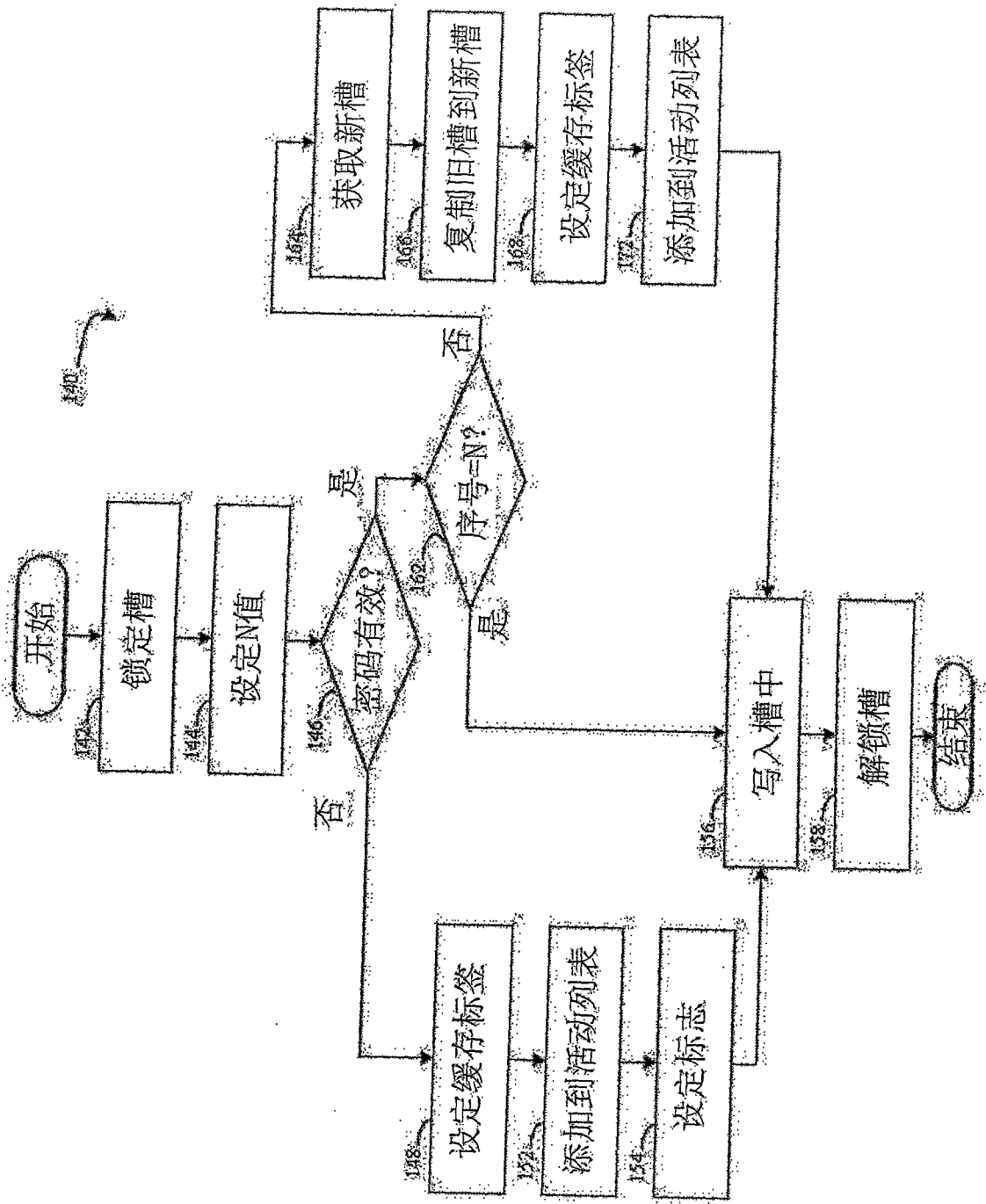


图5

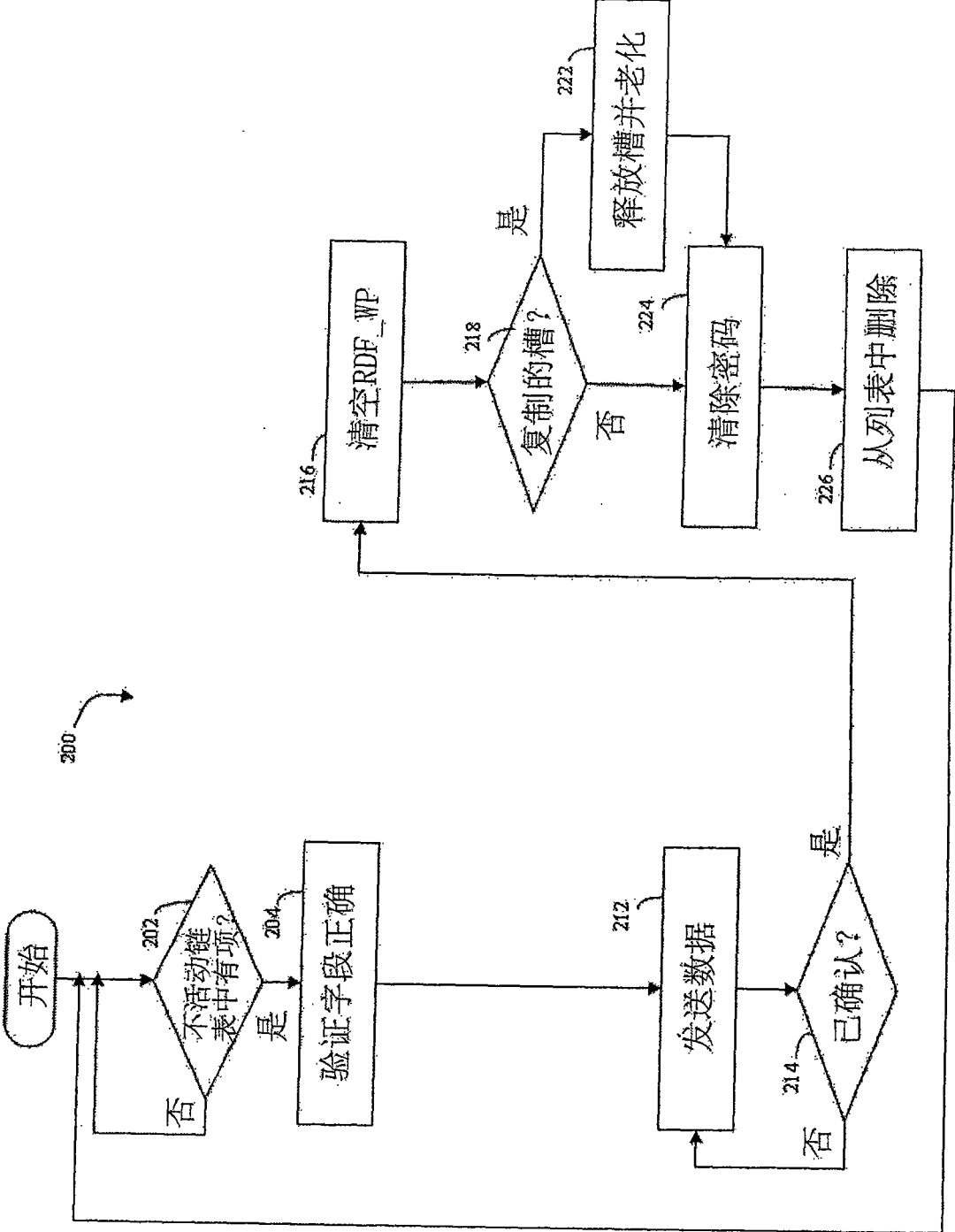


图6

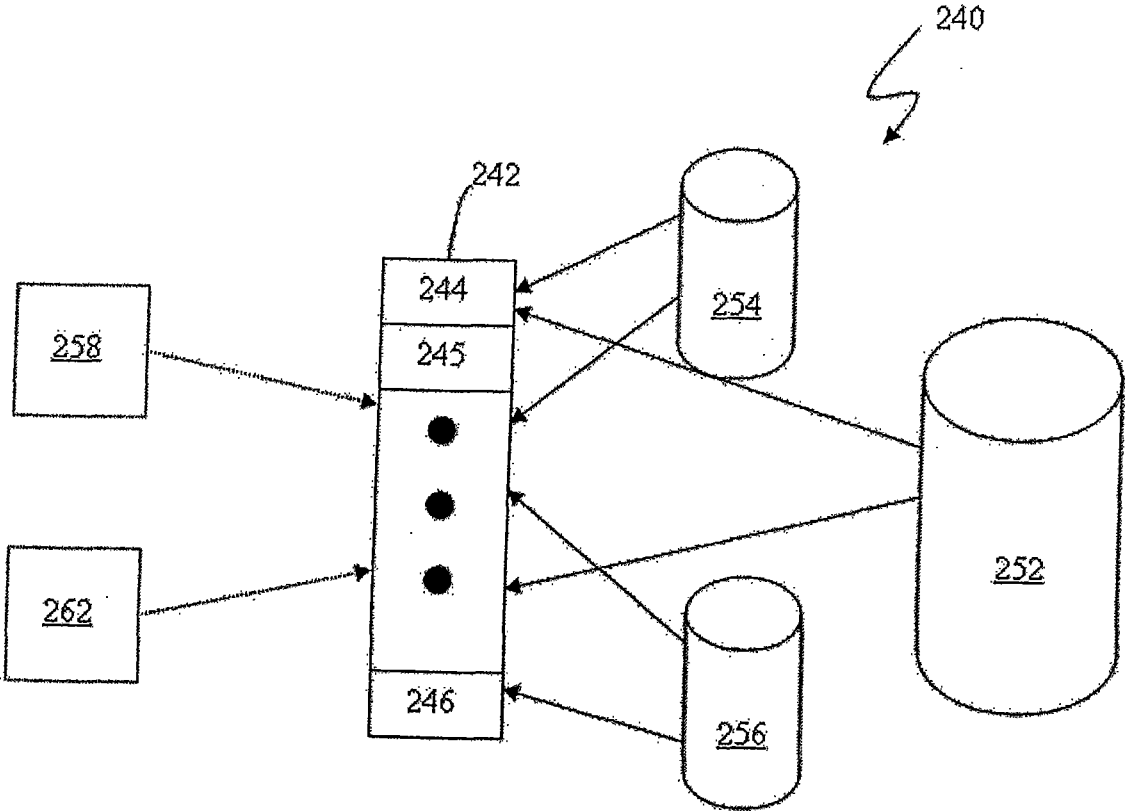


图7

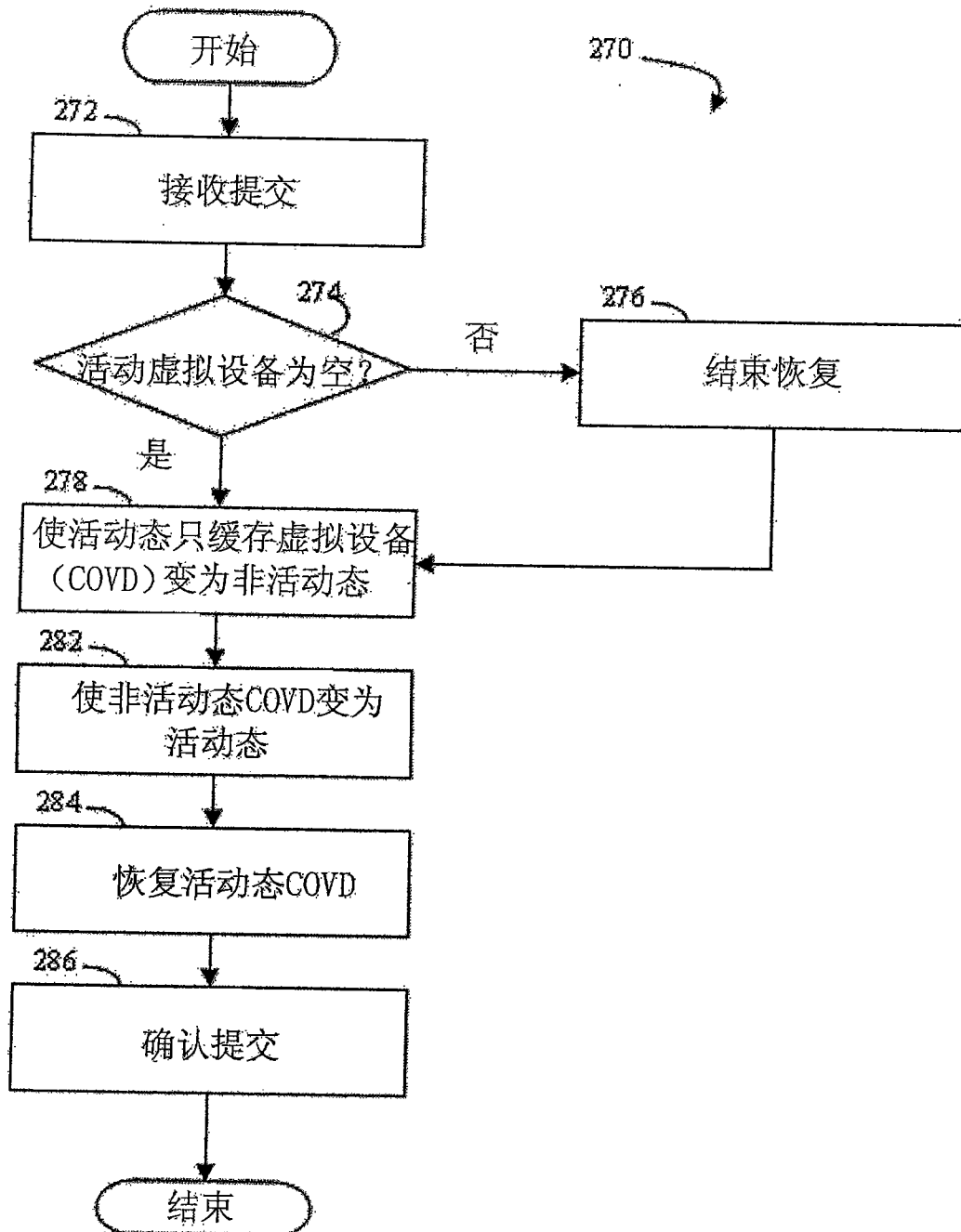


图8

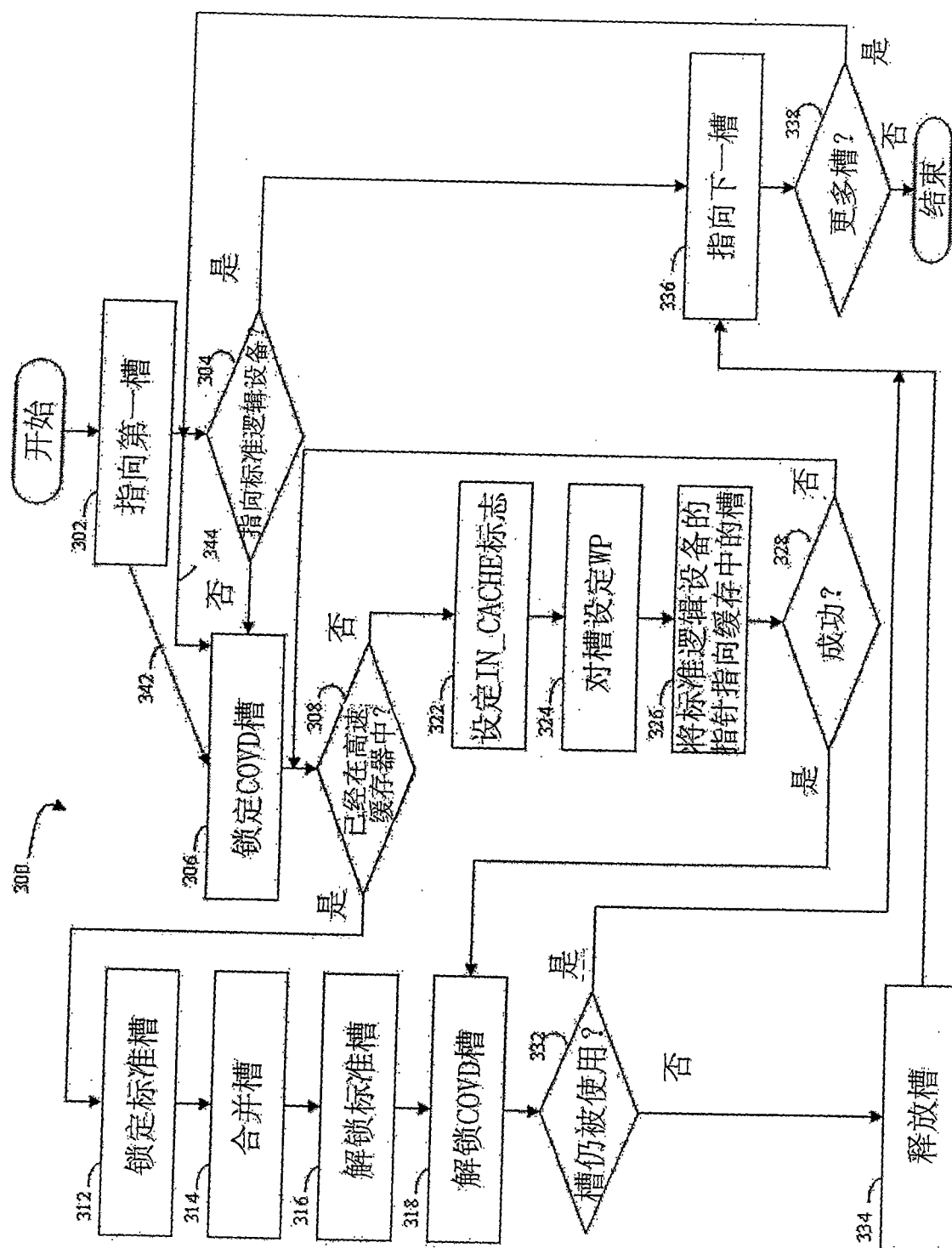


图9

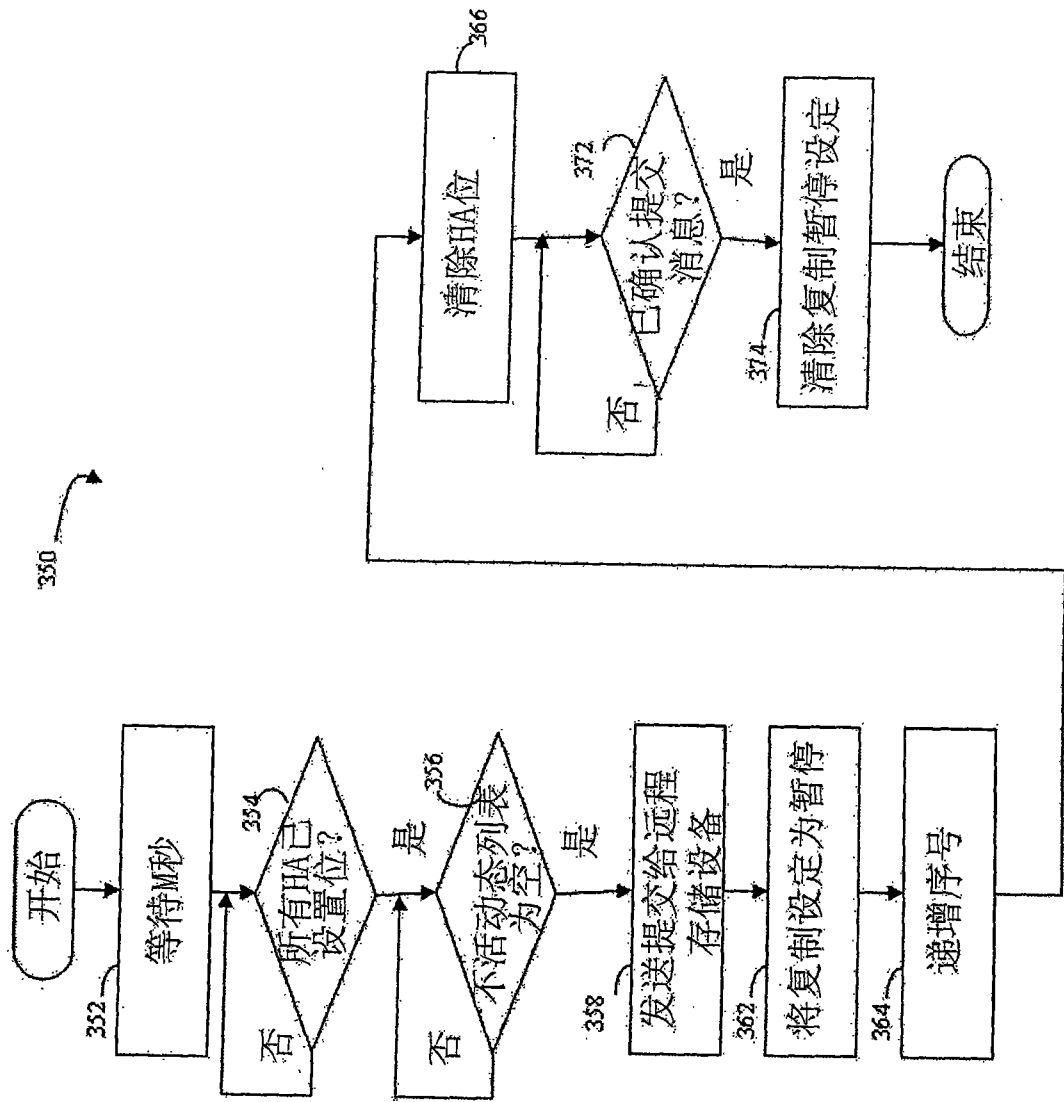


图10

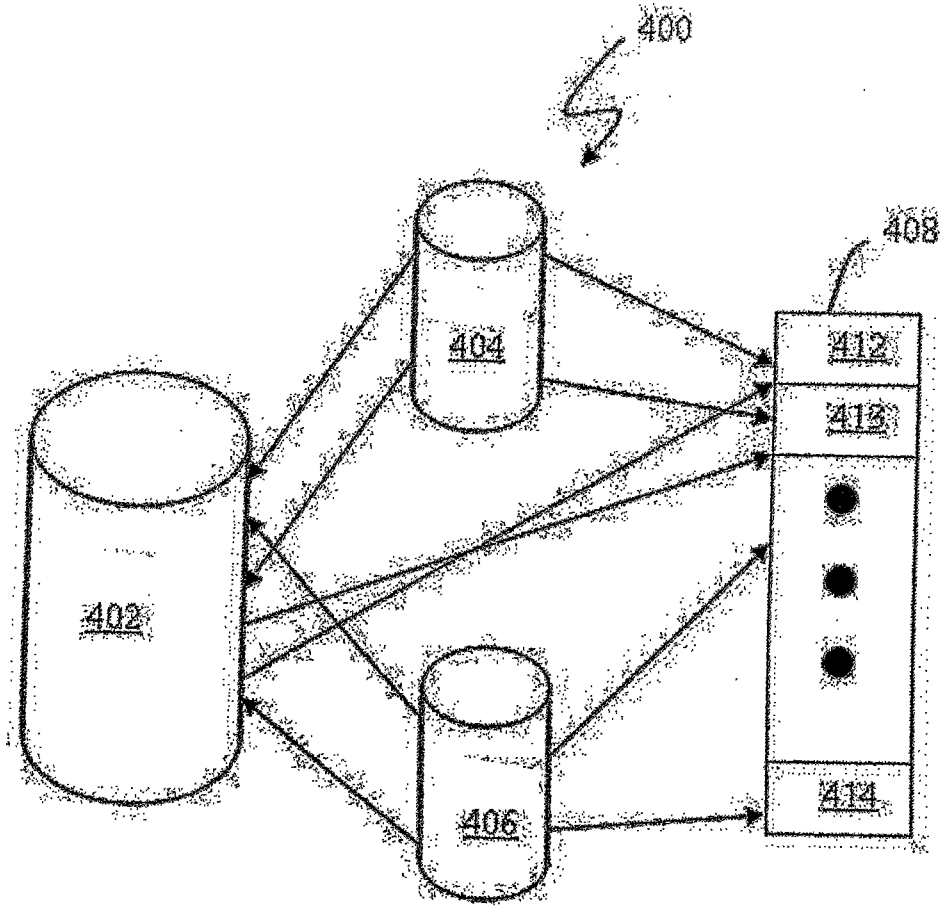


图11

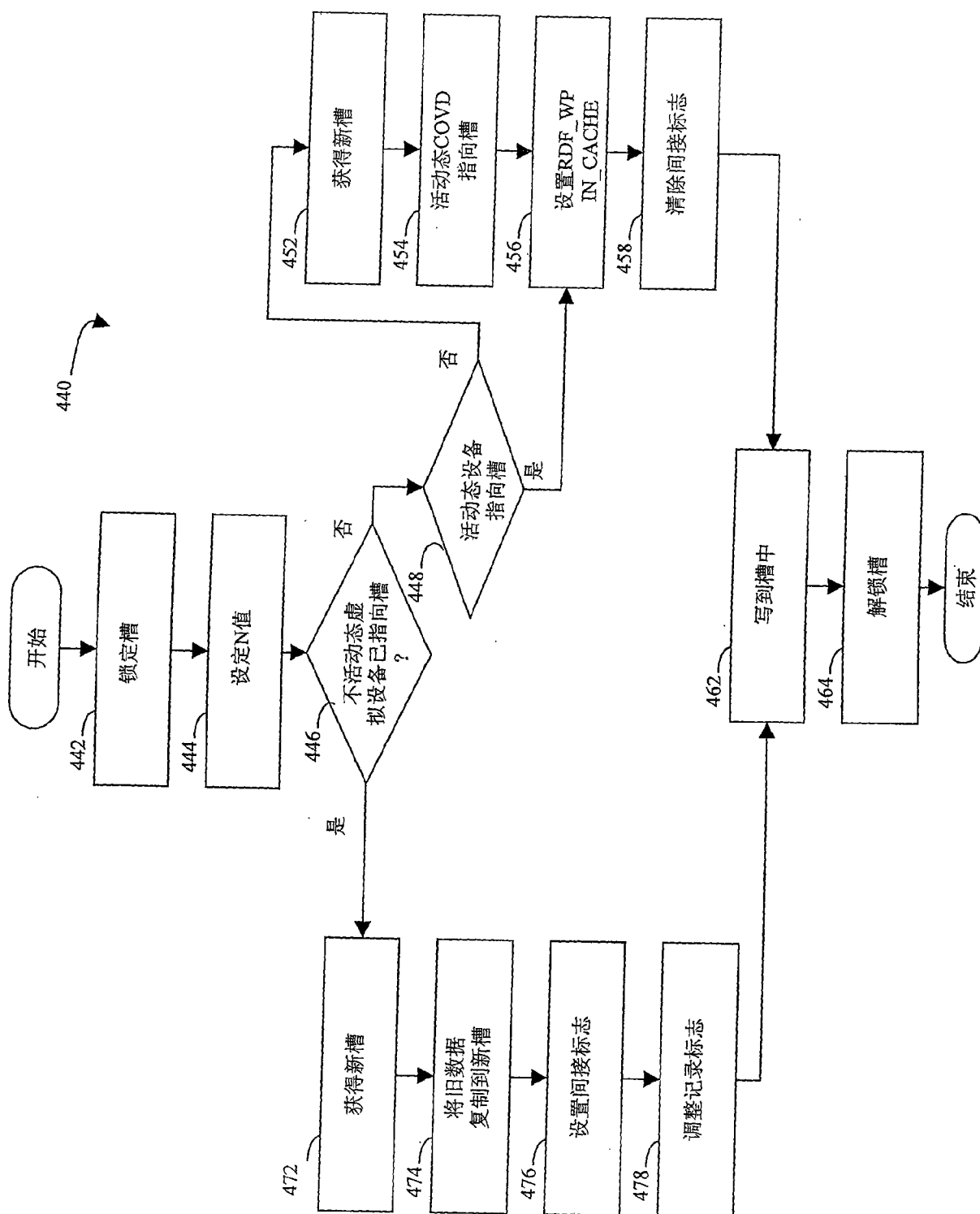


图12

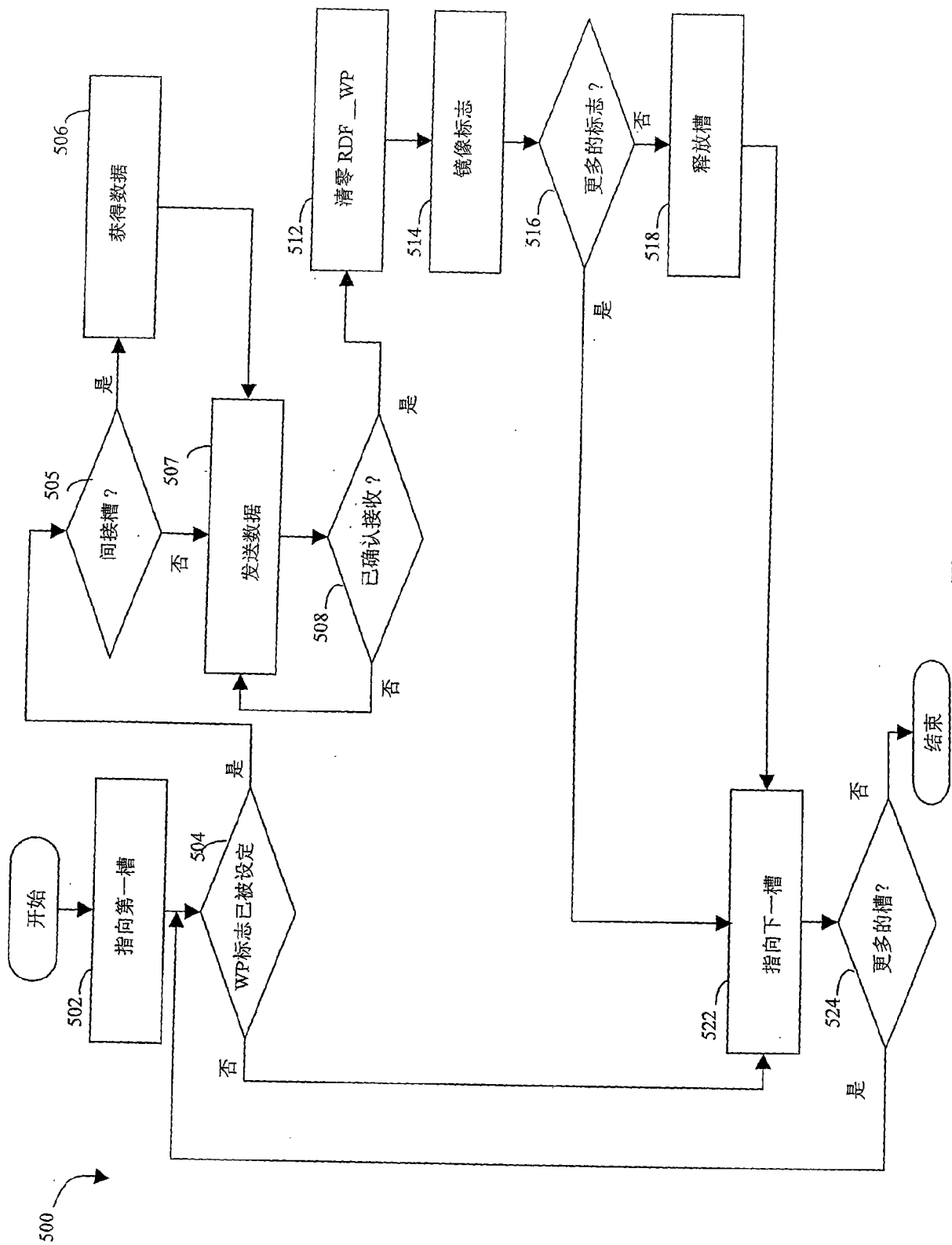


图13

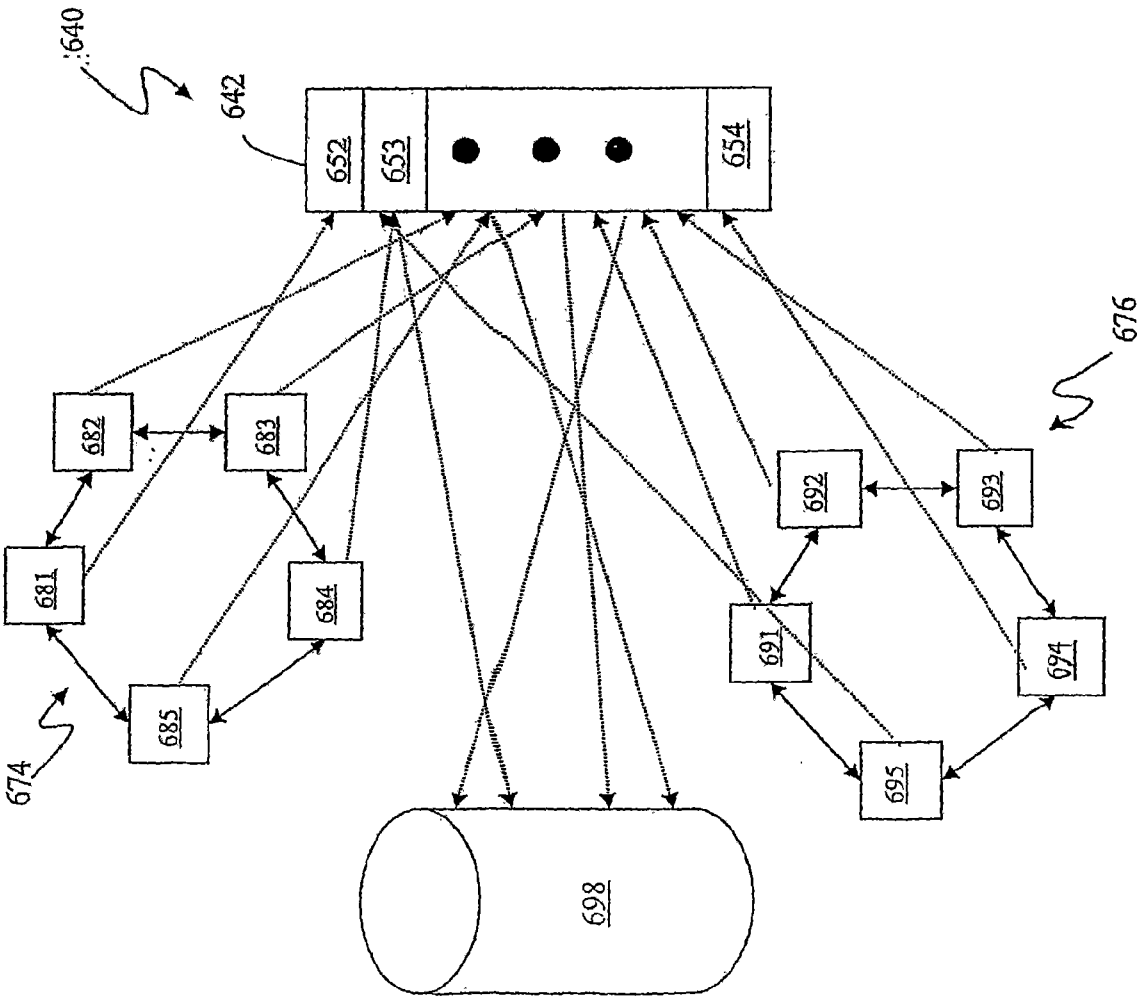


图 14

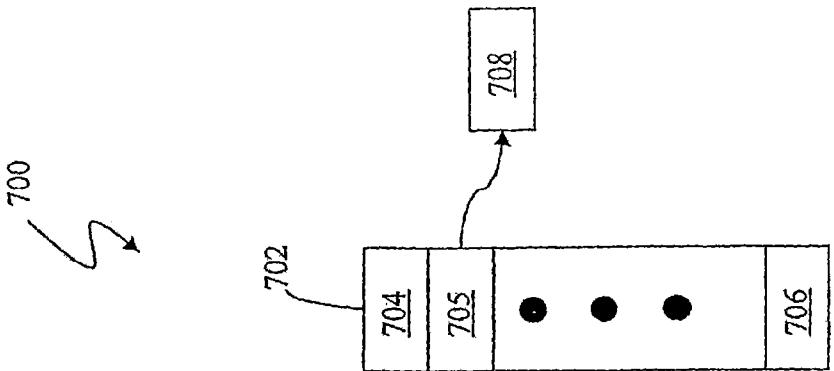


图 15

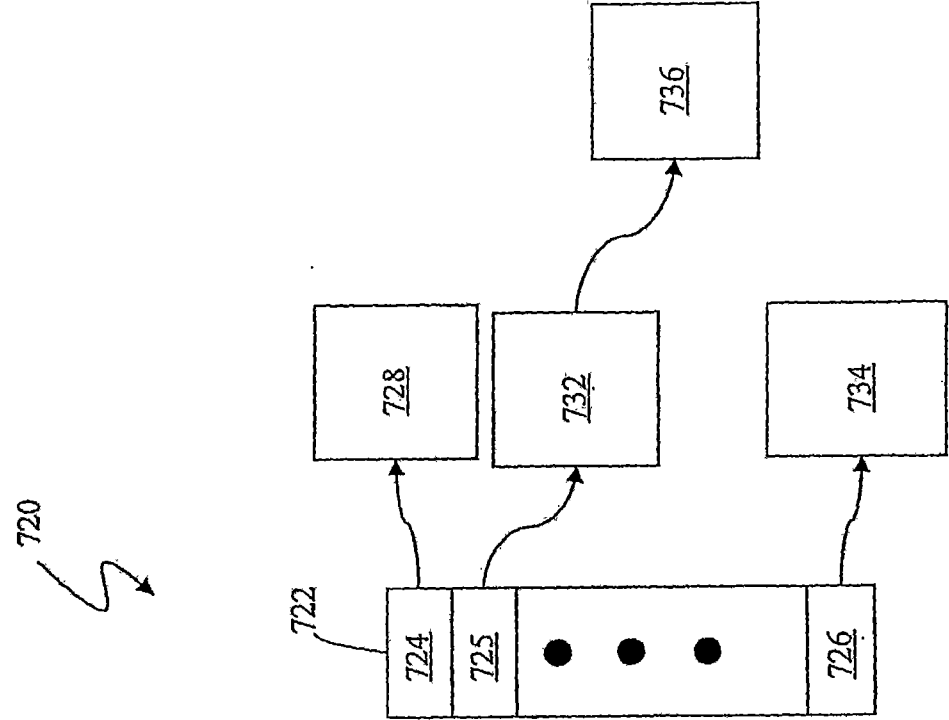


图 16