[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
G06F 11/07
G06F 11/16 G06F 11/20
G06F 13/14 G06F 13/36
G06F 12/00

[12]发明专利申请公开说明书

[21]申请号 00804532.1

[43]公开日 2002年5月22日

[11]公开号 CN 1350674A

[22]申请日 2000.2.8 [21]申请号 00804532.1

[30]优先权

[32]1999.3.3 [33]US [31]09/261,906

[86]国际申请 PCT/US00/03275 2000.2.8

[87]国际公布 WO00/52576 英 2000.9.8

[85]进入国家阶段日期 2001.9.3

[71]申请人 约塔约塔股份有限公司

地址 美国华盛顿州

[72]发明人 W·P·布朗恩

M·B·马修斯

[74]专利代理机构 上海专利商标事务所

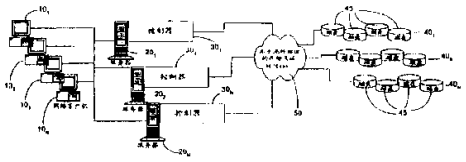
代理人 李 湘

权利要求书5页 说明书18页 附图页数19页

[54]发明名称 用来实现共用磁盘阵列管理功能的方法和系统

[57]摘要

包括在控制器(30)中的阵列管理功有(80)与存储区域欧(SAN),如基于 光纤信道的SAN(50),上的多冗余组有关。多阵列管理功能共用冗余组的管理 职责,每个冗余组典型地包括分布在多磁盘(45)上的多资源。AMF 为相关主机 系统提供并行访问冗余组,当主机请求一 AMF 执行对资源的操作,该 AMF 与共用 对包括将被操作的资源的冗余组的控制的其他 AMF 同步,以取得对该资源的锁定。在执行操作时,该 AMF 发送与该资源相关的复制数据和状态信息,这样,如果该 AMF 失效,任一其他 AMF 能完成该操作以维护数据可靠性及相关性。



ISSN1008-4274

权　利　要　求　书

1、一数据存储系统，它包含；

包括大量资源的一冗余组；

共用对该冗余组的访问的两或更多阵列管理功能（AMF），其中，AMF 为关联主机系统提供对该冗余组的并行访问；以及

用来连接 AMF 与该冗余组的一互连媒体；

其中，当第一 AMF 要求对该冗余组中的第一资源执行行操作，第一 AMF 为对该第一资源的锁定与共同对该冗余组的访问的其他 AMF 仲裁，因此第一 AMF 在第一资源上执行操作并向其他 AMF 并行发送与第一资源相关的复制数据和状态信息使得如果第一 AMF 在执行操作时失效，其他 AMF 中的一个能完成操作。

2、如权利要求 1 所述的数据存储系统，其中，直到第二 AMF 为对第一资源的锁定提出仲裁第一 AMF 才释放对第一资源的锁定。

3．如权利要求 1 所述的数据存储系统，其中，如果第一 AMF 失效，其余的 AMF 为第一资源的锁定提出仲裁，因此第二 AMF 取得锁定并完成该操作。

4、如权利要求 1 所述的数据存储系统，其中，由第一 AMF 在第一资源上执行的操作包括在量步骤，其中，第一 AMF 在该资源上执行该操作的每个步骤，并对每个步骤而言向其余的 AMF 并行发送与第一资源相关的复制数据和状态信息，使得如果第一 AMF 在执行该操作的任一步骤时失效，其余 AMF 中的一个能够完成该操作。

5．如权利要求 1 所述的数据存储系统，其中，一或更多主机系统与通过一互连媒体，一 Internet 连接和一外设部件互连总线这三者中的一个与第一 AMF 通信。

6．如权利要求 1 所述的系统，其中，互连媒体包括一 SCSI 接口，一光纤信道接口，一存储区域网和一光纤信道存储区域网中的至少一个。

7、如权利要求 1 所述的系统，其中，每个 AMF 在一控制器，一主机总线适配器和一主机计算机这三者中的一个中执行。

8、在数据存储系统中，一种在两或更多阵列管理功能（AMF）间动态共用对一冗余组的管理的方法，其中，AMF 能并行访问该冗余组，该冗余组包括大量资源，该方法包含以下步骤：

从主机接收由第一 AMF 执行对第一资源的第一操作的请求。

与其他 AMF 同步以取得对第一资源的访问，以及

与第一 AMF 在第一资源上执行第一读操作并行地，由第三 AMF 在第一资源上执行第二读操作。

9．如权利要求 8 所述的方法，其特征在于第一操作为读取操作，所述方法进一步包含以下步骤：

由 AMF 中的第二个从主机接收请求以完成第一资源上的第二读取操作；以及

在第一 AMF 于第一资源上完成第一读取操作的同时由第二 AMF 在第一资源上完成第二读取操作。

10、如权利要求 8 所述的方法，其中，同步这一步骤包括与一或更多其他 AMF 仲裁取得对第一资源的锁定以便在第一 AMF 维护锁定时其他 AMF 不能访问第一资源这一步骤。

11、如权利要求 10 所述的方法，还包含只有当只一 AMF 就锁定仲裁时才释放对第一资源的锁定这一步骤。

12、如权利要求 8 所述的方法，还包含与执行第一操作这一步骤并行地向其他 AMF 发送与第一资源相关的复制数据和状态信息这一步骤，使得如果第一 AMF 失效，其他 AMF 中的一个能完成第一操作。

13、如权利要求 12 所述的方法，其中，如果第一 AMF 在执行第一操作时失效，该方法还包含用第二 AMF 完成第一操作这一步骤。

14、如权利要求 13 所述的方法，其中，完成第一操作这一步骤包含第二 AMF 与其他 AMF 同步以取得对第一资源的访问这一步骤。

15、如权利要求 8 所述的方法，其中，第一操作包括大量子操作，其中，执行第一操作这一步骤包括在第一资源上执行每个子操作，该方法还包括与每个子操作的执行并行地，向其他 AMF 发送与第一资源相关的复制数据和状态信息，使得如果第一 AMF 失效，其他 AMF 中的一个能完成第一操作。

16、如权利要求 8 所述的方法，还包含以下步骤：

接收一来自于第二主机的，由第二 AMF 在第二资源上执行第二操作的请求。

和第一 AMF 及其他 AMF 同步以取得对第二资源的访问；以及

在第二资源上执行第二操作。

17、如权利要求 16 所述的方法，其中，与第一操作的执行并行地执行至少一部分第二操作。

18、一种数据存储网络系统，它包含：

一或更多冗余组，每个冗余组包括分布在多磁盘上的我资源；

共用一或更多冗余组的冗余组管理的两或更多阵列管理功能（AMF），其中，AMF 能并行访问一或更多冗余组；以及

用来互连 AMF 和一或更多冗余组的一互连媒体。

19、如权利要求 18 所述的系统，其中，第一冗余组包括一替换磁盘，其中，如果第一冗余组中有一个磁盘失效，至少有两个 AMF 都就控制该第一冗余组上的

一或更多资源仲裁，使得这至少两 AMF 都能用替换磁盘并行地重构第一冗余组。

20、如权利要求 18 所述的系统，其中，如果第一冗余组上再加一磁盘，至少有两个 AMF 都就控制该第一冗余组上的一或更多资源仲裁，使得这至少两 AMF 都能用该附加磁盘并行地扩充该冗余组。

21、如权利要求 18 所述的系统，其中，第一 AMF 从主机接收一向两或更多资源写至少两数据集的写命令，其中，第一 AMF 取得对第一数据集将写入的第一资源的锁定，向第一资源写第一数据集并并行地执行一复制操作，其中，与第一资源相关的复制数据和状态信息发送给其他 AMF。使得如果第一 AMF 在执行写操作时失效，其他 AMF 中的一个能够完成该写操作。

22、如权利要求 21 所述的系统，其中，与发送复制数据和状态信息并行，第一 AMF 判断是否向第二资源写第二数据集，如果是则取得对该第二资源的锁定，向第二资源写第二数据集且并行地执行第二复制操作，其中，与第二资源相关的复制数据和状态信息发送给其他 AMF，因此，第一 AMF 等到所有复制操作已完成才向主机发送状态信息。

23、如权利要求 18 所述的系统，其中，每个 AMF 包括一用来流水线复制入局主机数据的方法。

24、如权利要求书 18 所述的系统，还包含两或更多控制器。每个控制器执行两或更多 AMF 中的至少一个，其中，每个 AMF 通过一互连媒体和一或更多外设部件互连总线中的一个彼此通信，其中，该系统还包括一用来集聚消息的方法，其中，收信方是一控制器的许多小消息组合在一起并作为一个消息向该控制器发送。

25、如权利要求 18 所述的系统，其中，互连媒体包括一 SCSI 接口、一光纤信道接口，一存储区域网和一光纤信道存储区域网。

26、一种当一冗余组的一个磁盘在一含有两或更多与该冗余组互连的阵列管理功能（AMF）的数据存储网络系统中失效时用来重构该冗余组的方法，其中，该冗余组包括分布在多磁盘上的多资源，其中，该冗余组还包括一替换磁盘，其中，AMF 全部共用对该冗余组的管理，其中，AMF 能并行访问该冗余组，该方法包含以下步骤

由第一 AMF 就控制第一资源促裁；

由第二 AMF 就控制第二资源仲裁，以及

用该替换磁盘并行地重构第一和第二资源。

27、一种当在一包含两或更多与一冗余组互连的阵列管理功能（AMF）的数据存储网络系统中的该冗余组上加一附加磁盘时用来扩弃该冗余组的方法，其中，该冗余组包括分布在多磁盘上的多资源，其中，AMF 全部共用对该冗余组的管理，

而且其中，AMF 能并行访问该冗余组，该方法包含以下步骤：

第一 AMF 就控制第一资源仲裁；

第二 AMF 就控制第二资源仲裁，以及

用该附加磁盘并行地扩充第一和第二资源。

28、一种在一包含一与两或更多阵列管理功能（AMF）互连的冗余组的数据存储网络系统中流水线复制入局主机数据的方法，其中，该冗余组包括分布在多磁盘上的多资源，其中，该冗余组包括分布在多磁盘上的多资源，其中，AMF 全部共用对该冗余组的管理，并且其中，AMF 能并行地访问该冗余组，该方法包含以下步骤：

第一 AMF 从主机接收从主机接收一向两或更多资源写至少两数据集的写命令；

第一 AMF 取得对第一数据集将写入的第一资源的锁定；

向第一资源写第一数据集；以及并行地

扩行第一复制操作，其中，与第一资源相关的复制数据和状态信息发送给其他 AMF，使得如果第一 AMF 在执行写操作时失效，其他 AMF 中的一个能完成该写操作。

29、如权利要求 28 所述的方法，还包含以下步骤：

和执行第一复制操作并行地，判断是否向第二资源写第二数据集，以及如果是

取得对第二资源的锁定；

向第二资源写第二数据集；以及并行地

执行第二复制操作，其中，与第二资源相关的复制数据和状态信息发送给其他 AMF，以及此后

在所有复制操作完成后向主机发送状态信息。

30、在一数据存储系统中，一种在两或更多阵列管理功能（AMF）间动态共用对一冗余组的管理的方法，其中，AMF 能并行访问该冗余组，该冗余组包括大量资源，该方法包含以下步骤：为第一资源确定一判定器 AMF，其中，判定器 AMF 是两或更多共用对该冗余组的管理的 AMF 之一，而且其中，判定器 AMF 能授予以第一资源的锁定；

把来自第一 AMF 的锁定请求传送给请求对第一资源锁定的判定器 AMF；以及

一旦对第一资源的锁定已由判定器 AMF 授予则第一 AMF 执行在第一资源的操作。

31、如权利要求 30 所述的方法，其中，第二 AMF 有对第一资源的锁定，该方法还包含以下步骤：

判定器 AMF 向第二 AMF 发出释放锁定请求。

把锁定释放消息从第二 AMF 传递给判定器 AMF；以及此后

向第一 AMF 授予对第一资源的锁定以便第一 AMF 能执行该操作。

32、如权利要求 30 所述的方法，其中，当第一 AMF 传送锁定请求时其他 AMF 没有对第一资源的锁定，该方法还包含以下步骤：

立即向第一 AMF 授予对第一资源的锁定以便第一 AMF 能执行该操作。

33、如权利要求 30 所述的方法，其中，第一 AMF 请求的锁定是写锁，而且其中，一旦授予锁定，只到第一 AMF 释放该写锁其他 AMF 才能取得对第一资源的锁定。

34、如权利要求 30 所述的方法，其中，第一 AMF 请求的锁定是读锁，而且其中，其他任一 AMF 能并行地获得对第一资源的读源，使得多 AMF 可并行读第一资源。

35、如权利要求 30 所述的方法，其中，确定判定器 AMF 这步骤包括基于共用第一资源的 AMF 数量和第一资源的位置这两者中的至少一个指配两个或更多 AMF 中的一个作为第一资源的判定器 AMF。

36、如权利要求 35 所述的方法，还包括如果起先的判定器 AMF 失效则重新指配两或更多 AMF 中的不同的一个作为第一资源的判定器 AMF。

37、如权利要求 30 所述的方法，其中，确定判定器 AMF 这一步骤包括指配最后有地第一资源的写锁定的 AMF 作为判定器 AMF。

38、如权利要求 30 所述的方法，其中，判定器能授予对一或更多预取资源的锁定，其中，该方法还包括以下步骤。

第一 AMF 指定第一预取资源，和请求对第一资源的锁定并行地；以及

如果第一预取未尚锁定则向第一 AMF 授予对第一预取资源的锁定，使得第一 AMF 以后请求锁定时已有对第一预取资源的锁定。

01-09-07

说　明　书

用来实现共用磁盘阵列管理功能的方法和系统

对相关申请的相互参考

本申请与1998年9月24日提交的，临时专利申请序列号为60／101742，标题为“用来实现共用磁盘阵列管理功能的方法和系统”的美国临时专利相关，对该临时专利的揭示在此完整引用供参考。

发明领域

本发明通常涉及用来消除数据存储网中及直接服务器附带存储中的瓶颈系统和方法，更具体的说，涉及用来在多磁盘阵列管理功能间完成动态共用冗余组管理的系统和方法。

背景技术

对计算机与数据存储系统间较快通信的需要要求更快，更有效的存储网。近几年，集群技术及存储区域网（SAN）的实现已大大改进存储网性能，例如，在典型存储网中，把N个服务器集群在一起以获得成比例的性能，且在这些服务器与各种RAID（冗余廉价磁盘阵列）存储系统／阵列间加入SAN（例如，基于光纤信道的SAN）。SAN允许任一服务器访问任一存储元件。然而，在该典型存储网中，每个RAID系统有一相关RAID控制器，要访问存储在该RAID系统上的数据必须访问该相关RAID控制器，这会导致系统性能瓶颈，因为只有通过相关RAID控制器才能访问由该RAID控制器管理的存储。而且，如果控制器失效，便不能访问在由该失效控制器管理的RAID系统中维护的信息。

用来提供容错性的一个解决方案包括主从冗余控制器。主控制器进行主控制，只有在主控制器失效的时候从控制器才接替。然而，该解决方案很不有效，因为在主控制器出现失效之前从控制器是不使用的。另一解决方案是使用主从控制器体系结构，但把存储阵列分成冗余组，每个冗余组只由两控制器中的一个控制（每个控制器与其控制的冗余组相较是“主”）。这样，两个控制器同时运行，从而提高系统效率。一旦一个控制器失效，另一控制器承担对效控制器的冗余组的控制，该解决方案还防止“碰撞”，“碰撞”出现在，例如，不止一个控制器想向冗余组写数据。然而，该解决方案也有一些性能缺陷。例如，这样的主从体系结构中的性能受限于主控制器的速度，使得性能不可缩放。

这样，需要提供用来实现系统性能不受限于某控制器速度的对等控制器体系

结构解决方案的技术。而且，这样的系统应提供适当的容错性及性能可缩放性。

发明概要

本发明的提供这样一种用于数据存储管理的对等控制器体系结构解决方案。本发明的系统和方法实现一种新颖的，对建立高扩展的磁盘阵列有用的 RAID 阵列管理功能。特别是，本发明的系统和方法提供共用多（两或更多）阵列管理功能间的冗余组管理。

根据本发明，多阵列管理功能（AMF）与一互连存储媒体上的多冗余组相关。在一实施例中，例如，阵列管理功能与任一存储区域网（SAN），诸如一基于光纤通信的 SAN，上的冗余组有关。多 AMF 共用冗余组的管理职责，每个冗余组

典型地包括分布在多磁盘上的多资源。AMF 为相关主机系统提供对冗余组的并行访问。当主机请求一 AMF 执行对资源的操作，该 AMF 与共用对包括将被操作的资源的冗余的控制的其他 AMF 同步，以获得对该资源的锁定，在执行操作时，该 AMF 发送与该资源相关的复制数据和状态信息，这样，如果该 AMF 失效，任一其他 AMF 能完成该操作以维护数据可靠性及相关性。

如此处所用，术语“阵列管理功能”、“冗余组”和“冗余组管理”的定义同 RAID 咨询委员会（RAB）的系统存储技术手册（第六版）中所给，在此完全引用其内容供各种用途的参考。

“阵列管理功能（AMF）”一般指业或更多磁盘或磁带阵列提供共同的控制及管理的数据本。AMF 把它控制的磁盘或磁带阵列作为一或更多虚拟磁盘或磁带提交给操作环境，AMF 典型地在磁盘控制器，智能主机总线适配器中或在主机中执行。当它在磁盘控制器中执行时，AMF 常称为固件，一或更多 AMF 可如具体应用所需地在每个控制器、适配器或主机中执行。

“冗余组”一般指用来提供数据保护的，由 AMF 组织的 P 区域集合。对于一冗余组而言，使用单一类型的数据保护。冗余组中的所有用户数据存储容量受存储在该组中的检验数据保护冗余组外的用户数据容量不受该组中的检验数据保护。冗余组典型地包括由许多诸如条，数据块，高速缓存数据，映射表，配置表，状态表等的资源构成的逻辑实体。

“冗余组管理”一般指与一冗余组相关的一 AMF 的职责，进程及操作。

依照本发明，在各共用冗余组的 AMF 间动态协调并同步冗余组中检验数据的更新，用相干和锁定／解锁技术使这样的更新便利。典型地作为块、一组块，条或一组条的一个功能进行相干和锁定。用任一诸如 MESI 的广为人知的或专有的相干协议动态进行锁定，另外，同步维护与冗余组有关的高速缓存和冗余组中含有的数据间的相干性。

依照本发明的一方面，提供一典型地包含包括许多资源的一冗余组及共用对该冗余组的访问的两或更多阵列管理功能（AMF）的数据存储网。AMF 为关联主机系统提供对该冗余组的并行访问。该网还典型地包括一用来连接 AMF 和该冗余组的存储区域网。在一典型操作中，当第一 AMF 要求对该冗余组中的第一资源执行操作时，第一 AMF 和共用对该冗余组的访问的其他 AMF 为对第一资源的锁定进行仲裁。此后，第一 AMF 执行对第一资源的和并向其他 AMF 关行发送与第一资源相关的复制数据及状态信息。这样，如果第一 AMF 在执行操作时失效，一其他 AMF 能完成操作。

依照本发明的另一方面，提供一种在两或更多阵列管理功能（AMF）间动态共同对一冗余组的管理的方法，其中 AMF 能并行访问该冗余组，该冗余组包括大量资源，该方法典型地包含以下步骤：从一主机接收第一 AMF 要求对第一资源执行第一操作的请求，与其他 AMF 同步以获得对第一资源的访问，及对第一资源执行第一操作。

依照本发明的另一方面，提供一数据存储网系统，该数据存储网系统典型地包含一或更多冗余组，每个冗余组包括分布在多磁盘上的多资源，及两或更多共用对该冗余组或更多冗余组的冗余组管理的阵列管理功能（AMF），其中 AMF 能并行访问该冗余组或更多冗余组。该系统还典型地包含一用来互连 AMF 和冗余组的存储区域网。

依照本发明的另一方面，提供一种在一数据存储网络系统中在一冗余组的磁盘失效时重构该冗余组的方法。该典型的网络系统包含两或更多在一存储区域网上与该冗余组互连的阵列管理有（AMF），其中，所有 AMF 共同对该冗余组的管理且能并行访问该冗余组。该冗余组包括分布在多磁盘上的多资源和一替换磁盘。该方法典型地包含下列步骤：由第一 AMF 仲裁对该冗余组的第一资源的控制，由第二 AMF 仲裁对该冗余组和第二资源的控制。及用该替代磁盘并行地重构第一和第二资源。

依照本发明的另一方面，提供一种在一数据存储网络系统中当一附加磁盘加在一冗余组上时扩充该冗余组的方法。该网络系统典型地包含两或更多在一存储区域网上与该冗余组互连的阵列管理有（AMF）。该冗余组包括分布在多磁盘上的多资源。所有 AMF 共用对该冗余组的管理，并能并行访问该冗余组。该方法典型地包含以下步骤：由第一 AMF 仲裁对第一资源的控制，由第二 AMF 仲裁对第二资源的控制，及用该附加磁盘并行扩充第一和第二资源。

依照本发明的另一方面，提供一种在一数据存储网络系统中流水线复制入局主机数据的方法。该网络系统典型地包含一在一存储区域网上与两或更多阵列管理功能（AMF）互连的冗余组。该冗余组包括分布在多磁盘上的多资源。所有 AMF

共用对该冗余组的管理，并能并行访问该冗余组。该方法典型地包含以下步骤：第一 AMF 从一主机接收向两或更多资源写至少两数据集的写命令，第一 AMF 取得对第一数据集要写入的第一资源的锁定。该方法还典型得对第一数据集要写入的第一资源的锁定。该方法还典型地包括以下步骤：把第一数据集写入第一资源，及并行第一复制操作，其中，与第一资源相关的复制数据和状态信息发送给共他 AMF，这样，如果第一 AMF 在执行写操作时失效，其他 AMF 中的一个能完成写操作。

依照本发明的另一方面，提供一种用来在一数据存储系统中在两或更多阵列管理功能（AMF）间动态共用对一冗余组的管理的方法。这些 AMF 能并行访问该冗余组，该冗余组包括大量资源。该方法典型地包含为第一资源确定一判优程序 AMF 这一步骤，其中，该判优程序 AMF 是两或更多共用对该冗余组的管理的 AMF 中的一个。该判优程序 AMF 能授予对第一资源锁定的资格。该方法还典型地包含以下步骤：把请求锁定第一资源的锁定请从第一 AMF 传递到该判优程序 AMF。及一旦该判优程序 AMF 授权锁定第一资源则由第一 AMF 执行对第一资源的操作。

参考本规范说明书的剩余部分，包括附图和权利要求书，将认识到本发明的共他特点答成点。本发明的另外的特点和优点和本发明的各种实施例的结构有操作一样，在下面进行涉及附图的详细描述。在附图中，相同的参考号表示相同的或功能相近的元件。

附图简述

图 1 到图 7 依照本发明示出对用共用对冗余组的访问和控制的控制器从一或更多冗余组向一或更多主机系统提供数据有用的示例配置；

图 8 依照本发明示出一多控制器配置及这些控制器的内部配置；

图 9 依照本发明的一实施例示出一使用一通用同步序列的操作；

图 10 依照本发明的一实施例示出一使用一通用复序列的操作；

图 11a 依照本发明的一实施例示出当冗余组在正常，不降级方式时读操作的流程；

图 11b 依照本发明的一实施例示出当冗余组在降级方式时读操作的流程；

图 12 依照本发明的一实施例示出用来流水线入局主机数据的复制的流程；

图 13 依照本发明的一实施例示出当冗余组在正常、不降级方式时写操作的流程；

图 13b 依照本发明的一实施例示出当 AMF 更新图 13a 中所示的在完成更新前失效时恢复进程的流程。

图 14a 依照本发明的一实施例示出当冗余组在降级（出现失效驱动）方式时

写操作的流程。

图 14b 依照本发明的一实施例示出当 AMF 更新图 14a 中所示的在完成更新前失效时恢复进程的流程；

图 15 依照本发明的一实施例示出后台重构进程的注程；

图 16 依照本发明的一实施例示出后台扩充进程的通用序列流程；

图 17a 和图 17b 分别说明无本发明的消息聚集技术的 AMF 通信和有本发明的消息聚集技术的 AMF 通信。

图 18a 依照本发明的说明 AMF 请求锁定一资源的基本仲裁进程。

图 19 说明群集配置的两 AMF 间为单一资源的一简化仲裁进程；

图 20 依照本发明说明一包括四 AMF 的群集的示例资源仲裁序列。

详细描述

本发明在多 AMF 间提供共用冗余组管理（SRGM）以便多 AMF 能同时访问同一个冗余组。根据本发明，使用分布同步和复制技术协调所有共用一冗余组的 AMF 的活动并维护数据可靠性。可通过任一包括一共用对一冗余组的 AMF 的活动并维护数据可靠性。可通过任一包括一共用对一冗余组的控制的 AMF 的控制器取得对任一冗余组的访问。共用一资源组的 AMF 因此对等。另外，如果一冗余组为共用访问而配置且一控制器失效，阻塞了通过该失效控制器访问数据，但该冗余组上的数据仍完整，得到保护两不受磁盘失效的影响，且可从包括一正共用该冗余组的 AMF 的任一其他控制器访问。在一指定控制器中，可出现多 AMF，在此情况下冗余组较好的组共用在一组之上。也就是说，一些冗余组可由第一组 AMF 共用，其他冗余组可由第二组 AMF 共用，还有一些冗余组有一对一的关系，即一 AMF 一次只管理一冗余组。或者，一 AMF 同时管理多 AMF。

图 1 依本发明示出一基本的网络配。如图示，大量网络客户机 10，到 10n 可沟通地与大量服务器 20，到 20n 耦合，每个服务器包括一控制器 30。（一般说来，此处“N”用来表示无限多，这样数字“N”用来指一元件时不定等于一不同元件的数字“N”。例如，客户机 10 的数字不，但可能，等于图 1 中服务器 20 的数字。）每个网络客户机 10 耦合到一或多服务器 20 是通过特别应用及相对于服务器 20 的地理位置所需的任一连接方案的，这些大量的连接方案包括，例如，internet 连接，任一局域网（LAN）类型连接，任一广域网（WAN）类型连接，任一专用网连接，等等。每个控制器 30 包括一或更多 AMF，并通过一互连媒体，诸如一存储区域网（SAN）30，与磁盘机 45 的多阵列 40 可沟通地耦合。较好的是，SAN50 是一基于光纤信道的 SAN。然而可不脱离本发明的精神地使用任一 SAN 类型，诸如一基于 SCSI 的 SAN，或诸如一直接 SCSI 或 FC 连接的任一直接服务器互

连。因为每个控制 30 通过 SAN50 有对每个阵列 40 的随机访问，冗余组管理可由所有控制器 30 共用。

基于光纤信道的 SAN 较佳，因为光纤信道标准是一支持包括点对点，交换结构，仲裁环路的几个网络拓朴及这些拓朴的任一组合的开放标准。而且，光纤信道标准支持许多协议，包括 SCSI，异步转网模式（ATM），传输控制协议 / Internet 协议（TCP / IP）、高性能并行接口（HIPPI）、智能外围接口（IPI）及其他。另外，光纤信道在铜电缆和光缆上可提供高达 100Mbps（双向 200Mbps）的数据传送速度，传送距离分别达 30 米和 10 公里。

图 2 依照本发明示出我主机的一个例子，每个多主机有一在基于交换的光纤信道 SAN 中配置的控制器。在该 SAN 中每个控制器 30 通过如图所示的两光纤信道端口耦合到交换器 55。这样，每个控制器 30 和所有其他控制器 30 及磁盘阵列 40 通信。每个控制器 30 通过一外设部件互连总线 35 与其主机系统能信。用如图所示的环路拓朴把交换器 55 耦合到磁盘阵列 40。然而，对一领域中的熟练人员来说显而易见的是，可通过大量交换拓朴中的任何一个支持许多环路。一般说来，环路越多，可支持的数据传送速率越大。图 2 中所示的系统冗余是 N－1，意思是指定 N 个控制器（30，到 30n）中有多达 N－1 个的控制器可失效而当所有控制器 30 配置成共用对阵列 40 的管理时磁盘阵列 40 上的数据不会丢失。一特别主机的控制器失效导致该特别别主机而非整个系统的数据可用性丢失。该系统配置防止诸如电源失效的控制器环境故障，因为依照本发明来自一主机系统的数据与其他主机系统上的数据同步，这一点将在下文中作较详细的描述。

有与控制器失效相关的恢复时间。这是抗毁控制器确定所有关键数据又在群集中得到复制所花的时间。如果另一控制器在此恢复时间期间失效，则数据会丢失。

图 3 依照本发明示出多控制器的一个例子和配置在一基于交换的光纤信道 SAN 中的一单一主机。每个控制器 30 通过如图所示的两光纤信道端口耦合到该 SAN 中的交换器 55，然而，1 到 N 个端口可用于该具体应用所需。这样，每个控制器 30 通过该光纤信道 SAN 和所有其他控制器 30 及磁盘阵列 40 通信。而且，每个控制器 30 通过一或更多外设部件互连总线 35 与主机系统通信。这些控制器 30 还能通过外设部件互连总线 35 彼此通信。用如图所示的环路拓朴把交换器 55 耦合到磁盘阵列 40。然而，对本领域中的熟练人员而言显而易见的是，可通过大量交换拓朴中的任何一个支持许多环路。一般而言，环路越多，可支持的数据传送速率越大。在本配置中，在数据可用性对主机系统丢失前有多达 N－1 个的控制器可失效。提供外部控制器 30N＋1 以防止主机系统失效。如果主机系统失效，当控制器 30N＋1 依照本发明配置成与主机系统控制器 30，至 30N 共用对阵列 40

的管理，则阵列 40 上没有数据丢失。

图 4 依照本发明示出多主机的一个例子，每个多主机有配置在一基于交换的光纤信道 SAN 中的多控制器。每个控制器 30 通过图示的现金光纤信道端口耦合到 SAN 中的交换器 55，然而，1 至 N 个端口可用于该具体应用所需。这样，每个控制器 30 通过光纤信道 SAN 和所有其他控制器 30 及磁盘阵列 40 通信。而且，每个控制器 30 通过一或更多外设部件互连总线 35 与其主机系统通信。这些控制器 30 还能通过外设部件互连总线 35 彼此通信。用图示的环路拓朴把交换器 55 耦合到磁盘阵列 40，然而，对本领域中的熟练人员来说显而易见的是，可通过大量交换拓朴中的任何一种支持许多环路。一般说来，环路越多，可支持的数据传送速率越大。在本配置中，冗余和同步存在于每个主机系统内的两或更多控制器 30 间，如果每个主机系统包括 N 个控制器 30，在对主机系统的数据可用性丢失前有多达 N－1 个的控制器可失效。如果一主机系统失效，当其他主机上的控制器 30 依照本发明配置成与失效主机系统的控制器 30 共用对阵列 40 的管理时，阵列 40 上无数据丢失。

图 5 示出多控制器依照本发明配置在一光纤信道仲裁环路（FC－AL）SAN 中的多主机的一个例子。每个控制器 30 通过图示的两光纤信通端口耦合到该环路。这样，每个控制器 30 通过该 FC－AL 与所有其他控制器 30 及磁盘阵列 40 通信。而且，每个控制器 30 通过一或更多外设部件互连总线 35 与其主机系统通信。在此配置中，冗余和同步存在于每个主机系统内的两或更多控制器间。如果每个主机系统包括 N 个控制器 30，在对主机系统的数据可用性丢失前，多达 N－1 个控制器可失效。而且，如果一主机系统失效，当其他主机上的控制器 30 依照本发明配置成与失效主机系统的控制器 30 共用地阵列 40 的管理时，阵列 40 上无数据丢失。

图 6 示出依照本发明由两独立控制器管理的两独立冗余组。冗余组 A 由主机系统 A 的控制器 30A 管理，而冗余组 B 由主机系统 B 的控制器 30B 管理。还示出外部主机系统 C 和外部主机系统 D。依照一实施例，控制器 30A 和 30B 的 FC 端口既充当设备又充当主机信道。这使得每个控制器 30A 或 30B 能对分别来自于其关联外设部件互连总线 35，诸如外部主机系统 C，外部主机系统 D 的一外部 FC 主机或诸如控制器 30B 或 30A 的另一控制器的服务请作出响应。这样，主机系统 B 可访问冗余组 A 而主机系统 A 可访问冗余组 B。从控制器 30A 的角度看，例如，处理从主机系统 B 接收的一读或写数据的请求就好象该请求是通过关联外设部件互连总线 35 接收的。相似的是，外部主机系统 C 和 D 可通过在光纤信道 SAN 上向合适的控制器 30 发出读或写命令访问冗余组 A 和 B 上的数据。任一数目的主机系统可用这种方式互连。而且，尽管只示出一单控制器配置，可使用其他配置，

诸如图 4 中所示的配置。交换器的使用帮助隔离用来进行性能缩放的磁盘阵列。

图 7 示出依照本发明配置在一基于交换的光纤信道 SAN 中的，包括无源 PCI 底板中的多控制器的外部 RAID 系统的一个例子。控制器 301 至 30N 安装在一或更多无源 PCI 底板中，并配置成从 FC 端口和 / 或关联外设部件互连总线接受主机命令。这样，外部服务器可通过向合适的控制器 30 发出读或写请求访问由控制器 301 至 30N 控制的各种各样的冗余组上的数据。尽管只需要一个控制器 30，当加上较多控制器时性能和冗余按比例缩放。如果交换器 55 的使用不如人意或不可行，则可替换使用与图 5 中所示的配置相似的 FC－AL。

图 8 依照本发明示出一多控制器配置和控制器 30 的内部配置。图 8 所示的控制器 301 至 30N 中的一个或更多可位于一分离主机或在无源 PCI 底板上。例如，每个控制器 30 可位于一分离主机系统中，或每个多主机系统可包括一或更多控制器 30。PCI 主机连接 60 提供一用来从主机系统接收和处理命令及用来向其他控制器提供控制器间链路（ICL）服务的连接路径。光纤信道（FC）主机连接 65 提供 ICL 服务的连接途径，在较佳方面，每个控制器包括两个物理的 FC 端口（图 8 中未示出，但图 2 至图 7 是例子），这两个端口都用作磁盘机接入，接收并处理主机命令和 ICL 服务。对本领域中的熟练人员而言明显的是，每个控制器可包括特别应用所需的 1 至 N 个 FC 端口。

每个控制器 30 包括一或更多虚拟磁盘端口 70，每个磁盘端口提供对一虚拟磁盘 75 的访问，虚拟磁盘 75 基本上是一阵列的段。（“冗余组”通常与“阵列”同义）。每个阵列可按需分段成许多虚拟磁盘 75。每个虚拟磁盘是相关的并由一或更多关联 AMF80 控制。许多虚拟磁盘端口 70 可为同一虚拟磁盘 75 存在，但每个端口必须位于一分离控制器上。例如，如图 8 所示，与虚拟磁盘 YR 相关的虚拟磁盘端口 70YR，和 70YRN 分别提供给控制器 30，和控制器 30N 上的主机，每个虚拟磁盘端口 YR 提供对同一虚拟磁盘 YR 的访问，虚拟磁盘 YR 是阵列 Y 的一段，对虚拟磁盘 YR 的控制和管理由 AMF80Y 和 80YN 分担。

通过建立虚拟磁盘并向该虚拟磁盘端口指配一 IO 端口地址则可把虚拟磁盘端口加在一控制器上。通常，虚拟磁盘必须存在于创建虚拟磁盘端口并指配给虚拟磁盘之前，但虚拟磁盘端口紧接着冗余组的创建而创。虚拟磁盘端口然后可在此后的任一时间建立，但虚拟磁盘的建立只进行一次。也可在任一时间删除虚拟磁盘端口。允许所有在端口上进行的主机操作完成。当在完成这些操作时，柜绝新的主机操作，例如，通过向主机回一未就绪状态信号。

假定多 AMF80 出现在每个控制器 30 上，则按组地组共用冗余组。如图 8 所示，有三组 SMF，在三组 AMF，Y，G 和 T，各自共用一阵列。不同控制器上的关联 AMF 彼此同步和复制，这一点将在下面作较详细的描述。例如，如图 8 所示，AMF80Y1

和 AMF80YN 同步并复制（并与任一其他与阵列 Y 相关的 AMF，例如，AMF80Y2（未示出））。类似的是，AMF80G，与 AMF80GN 同步并复制，而 AMF80T1 与 AMF80TN 同步并复制。另外，一控制器上的虚拟磁盘端口与其他控制器上的相关虚拟磁盘端口同步并复制。

分布同步和复制

同步和复制保证由共用一冗余组（阵列）的不同 AMF 执行的操作不破坏性地互相干扰（例如，不出现“碰撞”）。同步要求任一需访问一与一共用冗余组相关的资源的 AMF 在使用该资源前就访问权（锁定）与其他 AMF 仲裁。仲裁通过在 PCI 和 / 或 FC　ICL 链路上在 AMF 间发送仲裁消息完成。

图 9 依照要发明的一实施例示出一操作的通用同步序列。在步骤 110 中，开始操作。例如，主机可发送一要求将数据写至一资源的请求。在步骤 120，AMF 判断是否已锁定所需资源。如果没有，AMF 在步骤 130 中就对所需资源的锁定与其他 AMF 仲裁。一旦取得锁定（或确定 AMF 已有锁定），在步骤 140 中由 AMF 对该源执行所需的和。一旦 AMF 取得锁定，较好直到另一 AMF 需要该锁定（即，另一 AMF 就该锁定仲裁)时才释放，以帮助削减许多应用中的共用冗余组客量(SRGM)开锁，。根据一施例，使用先来先得到服务仲裁方案，但基于优先权，或可以使用任一基他仲裁方案。如下文参考图 18、19 和 20 所作的较详细的描述，仲裁典型地包括向资源控制器（典型地为软件，但有时基于硬件）提出使用资源的请求。资源管理器基于所用的仲裁算法授予对该资源的访问权利。如果提出请求时资源在用，请求方要等到该资源空闲且共他所有在它之前的请求方已得到服务时。

每个 AMF 能够在一冗余组上搪行许多类型的操作，包括，例如，主机读，主机写，后台写，重写，重构，在线扩充，奇偶性消等等。这种操作的扩展序列取名为“进程”。进程的例子包括重构，在线扩充和奇偶性取消。所有 AMF 操作类型需同步仲裁以前进。一旦指定 AMF 完成对一资源的操作，其他 AMF 能够使用该资源。

较好在操作级而非进程级执行同步。也就是说，对每个将要执行的操作执行图 9 中所示的基本的同步序列。对于某一功能需在整个冗余组上执行的一进程(诸如重构）处理分解为一序列操作。如果每个操作在不同的资源上操作，则一操作所需资源的同步的仲裁与该进程中其他操作所需资源的同步的仲裁独立进行。在操作级而非进程级使用同步使得 AMF 在一处理在进行中时能共用资源。假如在进程级而非操作级执行同步，一些 AMF 将不得不等到整个处理完成才能使用这些资源，从而引起主机超时。

复制调节 AMF 失效。复制资源及其状态信息以便如果一 AMF 失效则通过另一

有该资源及其状态信息的拷贝的 AMF 提供该资源及其状态信息。在某资源的更新（写）期间，修改资源和／或该资源的操作状态的考贝发送给其他共用该资源的 AMF。这些其他的 AMF 称为复制伙伴。例如，参考图 8，AMF80Y1 和 AMF80YN 是都共用阵列 Y 的控制的复制伙伴。万一 AMF 更新资源在操作期间失效，由复制伙伴用该复制信息完成操作。

图 10 依照本发明的一实施例示出一操作的通用复制序列。该序列的开始是图 9 的基本同步序列。在步骤 210 中，操作开始。例如，主机可发送一要求向一资源写数据的请求。在步骤 220 中，AMF 判断是否已有对所需资源的锁定。如果没有，在步骤 320 中 AMF，就对所需资源的锁定与其他 AMF 仲裁。一旦取得锁定则可执行该操作。如图示，由 AMF 执行的操作分成一数目个，i 个，步骤。在步骤 2401 中，与该资源及第一操作步骤相关的复制数据和状态信息发送给每个复制伙伴。在步骤 2501 中，执行该操作的第一步骤。如图 10 所示，和复制步骤 2402 至 240i 一样，按顺序执行后面的操作步骤 2502 至 250i，在每个复制步骤 2401 至 240i 中，复制信息发送给与指定资源相关的复制伙伴。通常，如果 AMF 群集（即，共用一资源的那些 AMF）内有该资源及其状态信息的 N 个拷贝则调节 N－1 个并行 AMF 失效，其中 N 定义为复制度。这样，复制信息发送给与指定资源相关的 N－1 年复制伙伴。

可把复制指向具体复制组或具体 AMF。在一实施例中，无规定复制组地执行 N 向复制。在本实施例中，复制发生于群集中共用正在复制的资源的任何 N－1 个 AMF。或者，复制执行于 N－1 个其他复制组。复制组是一组以复制组而非本身的组局 AMF 复制关键数据的 AMF。这一点的例子是一组在一特理位置的控制器和在另一物理位置的另外一组控制器，每个控制器包括一或更多 AMF。另一个例子是主机系统中的一组控制器与主机外部的另一组控制器。使用复制组有助于确保如果一组控制器全部失铲，其他组控制器有维护数据可靠性所必需的信息。

复制可对准具体复制组或具体 AMF。一指定复制组较好地用正在复制的 AMF 复制组之外的任一 AMF 复制。然而，可用一操作者具体指明一指定复制组复制到的复制组群。而且，同步复制是较佳的复制方式，在同步复制方式中，在所有操作对象已收到复制数据后操作的完成状态信息反回主机。侯选复制方式包括异步复制和伪同步复制。在异步复制方式中，传送复制数据前操作的完成状态信息返回到主机。在伪同步复制方式中，传送复制数据之后但在所有复制对象角认接收数据之前，操作的完成状态信息返回到主机。

依照本发明，多 AMF 能够并行读一资源。也就是说，多读锁可在住一时间未认可。然而，只调节一个一次写入该资源的 AMF。允许多 AMF 同时读一资源大大改进读性能。如果一 AMF 正向该资源写，不允许其他 AMF 读该资源。同步协议允

许并加强这一方面。

图 11a 依照一发明的一实施例示出当冗余组（RG）在正常，非降级方式时读操作的通用序列流程。“非降级”通常指冗余组中的所有驱动器操作的情况，而“降级”通常指冗余组中的一或更多驱动器已失效的情况。在步骤 310 中，读操作开始。例如，AMF 接收来自主机的读一特别资源的请求。在步骤 320 中，需要锁定该特别资源。这基本上与图 9 的步骤 120 和 130 相同。在较佳方面，多锁可未认可。这使多 AMF 能并行读一资源。

图 11b 依照本发明的一实施例示出当冗余组（RG）在降级方式时读操作的通用序列流程。在步骤 10 中，读操作开始。例如，AMF 接收来自主机的，读一特别资源的请求。在步骤 320 中，需要锁定该特别资源。一旦取得锁定，在降级读方式，AMF 在步骤 340 读来自该特别资源的数据和奇偶性，并在步骤 350 再生任何丢失的数据。在步骤 360 中，该数据（再生的）传送到发出读请求的主机。

图 12 依照本发明的一实施例示出写操作的用来以流水线方式复制入局主机数据的通用序列流程。流水线化复制数据有助于最小化复制等待时间，在步骤 410 中，操作开始，例如，主机发出向一或更多资源写一或更多数据块的写命令。在步骤 420 中，从主机接收该主机命令并进行语法分析。该主机命令作为一序列数据集进行处理。在步骤 430 中，取得对开始块集的适当的锁定。在步骤 440 中，开始块集从主机传送给 AMF。在步骤 450 中，开始开始集的块集复制。AMF 不等块集复制完成；AMF 立即判断是否还有块集需要在步骤 460 处理。如果有，AMF 立即开始取得适当的锁定以得到步骤 430 中的下一块集，并为下一块集重复步骤 440、450 和 460，如果已接收并处理所有块集，AMF 等待所有复制操作在步骤 470 中完成。当每个操作完成时 AMF 在步骤 480 中向主机发送状态。

图 13a 依照本发明的一实施例示出当冗余组（RG）在正常，非降级方式时写操作的通用序列流程。在步骤 510 中，开始操作。例如，主机发出向一资源写数据的写命令。步骤 520 是如图 9 所示的取得对所需资源同步锁定的过程。在本例中该资源是一条状写锁，但亦可是其他任一由具体操作所需的锁定。取得条状写锁后，在步骤 530 中 AMF 读来自 RG 的磁盘的老数据和奇偶性。

和步骤 530 的磁盘读操作并行，在步骤 540 中，AMF 向其复制伙伴发送该资源的状态通知信号。复制伙伴包括该 AMF 复制该特别资源的状态信息和数据朝向的全部其他 AMF。如果 N 是复制度则复制伙伴的数目等于 N－1。较好的是，复度 N 是 1 至 8，但 N 可是所需的任一数目。状态通知信号是一“开始更新”类型信号，告诉复制伙伴条状更新已始。复制伙伴需道该信息，因为万一写 AMF 在完成操作前失效，复制伙伴将负责清理。

一旦“开始更新”通知步骤 530 和读老数据及奇偶性步骤 540 完成，AMF 在

步骤 550 中向 RG 存储磁盘写新数据。和新数据写步骤 550 并行的是步骤 560 中新奇偶性的生成。此后，在步骤 570 中 AMF 向 RG 存储磁写新奇偶性。一旦奇偶性写操作完成，AMF 在步骤 580 中向其复制伙伴发送一“停止更新”通知，一收到该通知，复制伙伴释放其与条状更新相关的状态数据。

图 13b 依照本发明的一实施例示出当图 13a 所示的 AMF 更新条在完成更新前失效时恢复过程的通用序列流程。当 AMF 更新条在完成更新前失效时，复制伙伴承担从失效更新操作中恢复的责任。在步骤 610 中，当一或更多复制伙伴 AMF 检测到失效或得知失效时恢复操作开始，复制伙伴在步骤 620 中为条锁的所有权仲裁，赢得该仲裁的 AMF（恢复 AMF）负责执行失效更新操作的恢复。

失效通知典型地来自于控制器间链路（ICL）部件。如果一控制器失效，该控制器上的 AMF 和与其共用该冗余组的其他 AMF 失去通信。ICL 周期性地向与其共用该冗余组的所有其他 AMF 发送一“砰”消息。如果没有 AMF 对该强制回应消息作出响应，则发送该强制回应消息的 AMF 认为 AMF 已失效开始恢复行动。如果 ICL 在向目的地 AMF 发送同步或复制消息时遭遇传送失效也引发恢复。

恢复过程包括两基本步骤：重新计算条奇偶性及重写数据。在步骤 630 中，恢复 AMF 读受失效更新操作影响的条段的所有数据。和数据读步骤 630 并行，恢复 AMF 在步骤 640 中向它的所有复制伙伴指配一或更多新复制伙伴并发送一“开始更新”通知。当数据读操作完成，恢复 AMF 在步骤 650 中生成新奇偶性，该新奇偶性计算不包括新数据，只是 RG 存储磁盘上数据的奇偶性再生。

一旦奇偶性生成步骤 650 和“开始更新”通知步骤 640 完成，恢复 AMF 在步骤 660 中向 RG 存储磁盘写新奇偶性。在奇偶性写步骤 660 完成后，恢复 AMF 在步骤 670 中向复制伙伴发送一“结束更新”通知。在一些时间间隔后，高速缓存写回调度算法在步骤 680 中使一复制伙伴向 RG 存储磁盘写新数据，这是图 13a 所示的一正常（非恢复方式）条更新操作。高速缓存功能是 AMF 的一部分。

图 14a 依照本发明的一实施例，示出当冗余组（RG）在降级（有一失效驱动）方式时写操作的流程。该序列类似于图 13a 所示的非降级情况的序列，包括下面将描述的再生步骤 744 和复制步骤 746。

在步骤 710 中，操作开始，例如，主机发出向一资源写数据的写命令。步骤 720 是图 9 所示的取得对所需资源的同步锁定的过程。在此例中该资源是一条写锁，但也可是特别操作所需的任一其他锁，在取得条写锁后，AMF 在步骤 730 中读来自 RG 的磁盘的老数据和奇偶性。

与步骤 730 的磁盘读操作并行，在步骤 740 中，AMF 向其复制伙伴发送一该资源的状态通知信号。复制伙伴包括该 AMF 复制该特别资源的状态信息和数据朝向的所有其他 AMF。状态通知信号是一“开始更新”类型信号，告诉复制伙伴条

更已开始。复制伙伴需要知道该认息，因为万一写 AMF 在完成操作前失效复制伙伴将负责清理。一旦读老数据和奇偶性步骤 540 完成，AMF 在步骤 744 中再生失效磁盘上的数据。在步骤 746 中，老数据，包括再生数据，复制到复制伙伴。万一更新 AMF 在完成操作前失效，把这些数据复制复制伙伴对恢复而言是必要的。

一旦复制步骤 746 完成，步骤 750 中新数据写给 RG 存储磁盘。和新数据写步骤 750 并行的是步骤 760 中的新奇偶性的生成。此后，在步骤 770 中 AMF 向 RG 存储磁盘写新奇偶性。一旦奇偶性写操作完成，AMF 在步骤 780 中向其复制伙伴发送一“结束更新”通知。一收到此通知，复制伙伴释放与条更新相关的状态数据。

图 14b 依照本发明的一实施例示出当图 14a 所示的 AMF 更新条在完成更新前失效时恢复过程的通用序列流程。该情况与图 13b 所示的非降级恢复情况不同的是：恢复 AMF 使用复制的老数据得检受该更新 AMF 失效影响的 RG 磁盘段。当 AMF 更条在完成更新前失效时，复制伙伴承担从失效更新操作恢复的责任。在步骤 810 中，当一或更多复制伙伴 AMF 检测到失效或收到例如由主机所发的失效通知时，恢复操作开始。在步骤 820 中复制伙伴就条锁的所有权仲裁。赢得该仲裁的 AMF（恢复 AMF）负责执行失效更新操作的恢复。

在步骤 830 中，从由图 14a 的复制步骤 746 所供的老数据中生成新奇偶性，与步骤 830 并行，在步骤 840 中恢复 AMF 向它的所有复制伙伴指配一或更多新复制并发送一“开始更新”通知，在步骤 850 中，向磁盘写老数据。在步骤 860 中，通知复制伙伴已把老数据写回至磁盘。复制伙伴现在可废弃它们对老数据的拷贝。在此之后，恢复序列同非降级恢复序列。特别的是，在步骤 870 中要向 RG 存储磁盘写的新奇偶性。在奇偶性写步骤 870 完成后，恢复 AMF 在步骤 880 中向复制伙伴发送一“结束更新”通知。在一些时间间隔后，在步骤 890 中高速缓存写回调度算法使一复制伙伴向 RG 成员磁盘写新数据，这是图 13a 所示的一正常（非恢复方式）条更新操作。

分布重构

图 15 依照本发明的一实施例，示出后台重构过程的通用序列流程。每个操作开始于步骤 910，在步骤 920 中取得适当的锁定，本例中是条锁。在步骤 930 中，AMF 读条的数据和奇偶性。在步骤 940 中，AMF 再生丢失的数据，而在步骤 950 中向替换磁写数据。写数据之后，在步骤 960 中 AMF 更新映射表以反映，原本映射到失效磁盘的块现在映射到替换磁盘这样一个事实。映射表向 RG 成员磁盘和这些磁盘上的块映射主机逻辑块。在步骤 970 中，判断是否还有条需重构。只要还有需要构到替换磁盘上的条，该序列例重复。

本发明的共用冗余组管理（SRGM）技术的一个优点是：SRGM 允许对所有共用一 RG 的 AMF 的分布重构处理。这样的结果是：较快的重构时间，及主机在重构期间通常遇到的响应时间增加的量的普遍减少。分布重构通过使共用一RG 的 AMF 的一子集（或全部）就哪些条将各自负责重构进行协调而完成。如果有一些 AMF 在重构期间失效或停工，剩下的 AMF 重新仲裁重构责任。例如，假设有 N 个共用一冗余级组的 AMF 需要重构。这些 AMF 互相交谈（通过发送消息）并确定哪些 AMF 将参加重构，例如，N 的一个子集，用 M 表示。这些 M 个 AFM 通过确定哪些 AMF 将重构哪一条或哪些条确定重构责任。这可利用任一算法确定。在本发明的一实施例中，例如，每个 AMFi 重构条 Mo(s/i)，其中 S 是条号。

分布 RG 扩充

正如重构较好地分布在共用一 RG 的 AMF 的一子集间，RG 扩充也是如此，RG 扩充是对一现有 RG 增加驱动成员。SRGM 的一独特优点在于：允许扩充处理分布给所有共用一 RG 的 AMF。这样的结果是：较快的扩充时间，及主机在扩充期间通常遇到的响应时间增加的量的普遍减少。

分布扩充通过使共用一RG 的 AMF 的一子集（或全部）就它们将负责扩充哪些条进行仲裁而完成。如果有一些 AMF 在扩充期间失效或停工，其余的 AMF 重新仲裁充责任。例如，假设有 N 个共用一冗余组的 AMF 需扩充。这些 AMF 互相交谈（通过发送消息）并确定哪些 AMF 将参加扩充，例如，N 的一个子集，表示为 M。这些 M 个 AMF 通过确定哪些 AMF 将扩充哪一条或哪些条确定扩充责任。这可用任一算法确定。在本发明的一实施例中，例如，每个 AMFi 扩充 Mod(s/i)，共中 S 是条号。

图 16 依照本发明的一实施例示出后台扩充过程的通用序列流程。该过程始于步骤 1010，适当的条锁定在步骤 1020 中获得。扩充情况与前面的例子不同的是：必须取得多锁定，一扩充操作将涉及 2 或更多的条。一个条是条宽将从 W 增加到 W+N 的扩充条。涉及的其他条是含有将从这些条迁移到该扩充条的主机数据条。

在步骤 1030，读要扩充的条上的数据。在步骤 1040 中，复制这些数据以便如果该操作在完成前失效，复制伙伴能够在失效后进行清理且继续该扩充过程，与步骤 1030 及 1040 并行，在步骤 1015 中，读含有将迁移到该扩充的条的数据的源数据条。步骤 1040 和 1045 完成后，在步骤 1050 中 AMF 通知它的复制伙伴：它正开始扩充条更新，并行地，在步骤 1055 中，AMF 生成该扩充条的奇偶性信息。

通知开始更新完成后，在步骤 1060 中将该扩充条的数据写到磁盘。一旦奇偶性生成步骤 1055 和通知开始更新步骤 1050 完成，在步骤 1080 中 AMF 通知其复制伙伴：更新完成。复制伙伴然后更新其映射表以及映增加条宽和迁移的主机数

据。复制伙伴还丢弃在步骤 1040 中复制的数据。映射表向 RG 成员磁盘及这些磁盘上的块映射主机逻辑块。

在步骤 1090 中，确定是否还有由该 AMF 扩充的条。如果有，则重复该序列。只要还有条需扩充以利用新 RG 成员磁盘的容量，该序列重复。注意这是过程——成其为过程的是，步骤 1090 中产生的环路。步骤 1020 到 1090 包含一操作。

消息集聚

人们对 SRGM 关心的一个问题是与用来支持 SRGM 的同步和复制相关的处理开销及 IO 信道负载。为使同步及复制技术简便，较好使用 ICL（控制器间链路）消息集聚。消息集聚是，大体上，一种把发往一群集节点（即，一控制器，其中可驻留许多 AMF）的许多小消息组合成一个大消息包并把它作为一个消息发送给该节点的算法。这极大地减少处理开锁及 IO 信道负载，并与局一群集节点发送单独消息的方法形成对比。

图 17a 说明无本发明的消息聚技术的 AMF 通信。如图所示，一 AMF 集 1100 和一控制器间链路（ICL）实体 1105 构成一 SRGM 节点 1110。节点典型地是诸如控制器的一硬件实体。ICL1105 是负责一 AMF 到另一 AMF 的路由同步和复制消息 1120 的一软件实体。如图 17a 所示，许多相似的节点中只有一个占 SAN1130 连接。

节点 1110 中的 AMF1100 正在共用同一冗余组的其他节点上的其他 AMF 收发同步和复制消息。节点 1100 内的每个 AMF 生成独立的同步和复制消息流，，全部发往 SAN 1130 上的一或更多其他节点，正由一 AMF 收发的消息独立于正由同一节点上其他 AMF 收发的消息。如图 17a 所示，三个 AMF1100 正向其他节点上的 AMF 发送总共九个消息。而且，没有消息集聚，由一 SAN 节点中的所有 AMF 生成的所有同步和复制消息都通过该 SAN 单独地处理和发送。每个消息带来一固定数量的处理开涣，而不问大小。

图 17b 说明有本发明的消息集聚技术的 AMF 通信。消息集聚是把发往一节点的许多较小的消息在一起以形成一较大的消息。该较大消息可通过 SAN1130 作为一个消息发送并在接收节点上解压为单独的消息。例如如图所示，九个消息 1120 要发往三个不同的节点。在此例中，如果使用消息集聚，ICL1105 只需发送三个消息 1150——每个节点一个（不计自身）。ICL1105 承担压缩和解压单独 AMF 消息的责任。

仲裁和分布式资源中裁

图 18a 依照本发明说明当 AMF 请求锁定一资源时的一基本仲裁过程。AMF1200 和 AMF1210 各自请求锁定一资源，锁定请求在一仲裁队列 1205 中排队。当一请

求置于仲裁队列 1205 中时一 AMF 的仲裁过程开始。按某一顺序处理该请求，按优先级顺序满足所有请求。通过任一广为人知的算法（例如，FIFO，LIFO）建立请求队列优先级。每个请求 AMF 必须等到处理其请求以取得锁定。每个 AMF 在成功完成仲裁过程时获得对资源的锁定，如果仲裁失效则 AMF 不能锁定该资源。

图 18b 依照本发明说明广义仲裁过程的通用过程流程。仲裁包括协调两 AMF 间的资源锁定传送：请求 AMF1225 和有资源锁定的 AMF1230。AMF1225 向判优器 1220（执行仲裁过程的实体）发送一请求锁定消息，判优器 1220 让该消息排队到由判优器 1220 的优先级算法确定的一时间。适合时，判优器 1220 通过向目前有资源锁定的 AMF1230 发出一释放锁定消息处理该请求。一旦到 AMF1230 能安全地不继续使用该资源时，AMF1230 便释放锁定并通知判优器 1220 锁定释放。判优器 1220 接着发信号给请求 MF1225 告诉已授予资源锁定。AMF1225 继续保持该锁定直到判器 1220 要求它释放该资源。

当一或更多 AMF 请求对一资源的读锁时优化仲裁队列是有可能的。在较好方面，在这些情形中仲裁过程同时授予读锁，只要保持命令顺序。

一 AMF（或控制器）管理冗余组群集中一资源的仲裁过程，该 AMF 称为资源仲裁器。可用多种方法（例如，所有资源的单一仲裁器，负载平衡分配，等等）中的任何一种完成一资源的仲裁器的指配。依照本发明，较佳的仲裁指配方法基于控制器数目和资源范围。对有一个或两个 AMF 的群集配置，指配的仲裁器是有写锁的最后的 AMF。对有多于两个 AMF 的群集配置，仲裁器指配根据基于 AMF 数量及资源位置（例如，条 ID，块 ID，等等）的固定映射算法。

图 19 说明群集配置的两 AMF 间就单一资源的一简化仲裁过程。第一 AMF1300（AMF＃1）向第二 AMF1310（AMF＃2）发出对一资源的读锁请求 1320，AMF1310 目前有对该资源的写锁。AMF#2 向 AMF#1 发出一授锁（读）消息 1330，表示已授予资源锁定。AMF#1 现在有对该资源的读访问。当 AMF#1 向 AMF#2 发出一写锁请求 1340 时该序列继续。AMF#2 响应以一授锁（写）消息 1350。继续地，AMF#1 发出一读锁请求 1360，而由于 AMF#1 已有写锁。它便处理自身的仲裁并使写锁降级为读锁，接下来，AMF#2 此时没有对该资源的锁定，所以它无需通知。AMF#2 向 AMF#1 发出一读锁请求 1370，AMF#1 立即用一授锁（读）消息 1380 响应，因为该资源支持多读锁。对于两 AMF 的群集配置，当有一写锁的 AMF 在请求一读时不需要与其他 AMF 仲裁。在所有其他情况中，第二 AMF 必须仲裁锁定。

图 20 根据本发明说明的一包括四个 AMF 的群集的示例资源仲裁序列。含有三个或更多 AMF 的群集的较佳仲裁器指配方法是用一固定的晨射算法选择仲裁器。这样有把一仲裁器与单一 AMF 永久粗连的效果。万一 AMF 资源仲裁失效，则根据该射算法重新指配资源仲裁器。

参考图 20，第一 AMF1400（AMF#1）向第三 AMF1410（AMF#3）上的资源 X 仲裁器发出一写锁请求 1420。AMF#3 上的仲裁器向第二 AMF1405（AMF#2）发出一释放锁定请求 1422 以释放它对资源 X 的锁定。当资源 X 的锁请求 1420 在进行时，AMF#1 发出资源 Y 的读锁请求 1424。第四 AMF1415（AMF#4）是资源 Y 的指配仲裁器。AMF#4 立即授予一读锁 1426，因为其他 AMF 目前无锁定。接下来，AMF#4 发出资源 X 的一写锁请求 1428，AMF#3 上的仲裁器把它排在队列中，因为它目前正在处理写锁请求 1420。

AMF#2 向 AMF#3 发送一锁释放消息 1430，AMF#3 向 AMF#1 发送一授锁（写）消息 1432。一表示 AMF#1 在结束时应释放该锁定的标识嵌入在授锁消息 1432 中。这种优化不需要 AMF#3 向 AMF#1 发送一释放锁定消息。当对资源 X 完成时，AMF#1 向 AMF#3 发送一锁释放消息 1434，AMF#3 向 AMF#4（接下来排队等对资源 X 的写锁）发送一授锁消息 1436。

以请求锁定（读）消息 1440 开始的序列显示出一多读锁情况。除授锁（读）消息 1442 外，AMF#2 和 AMF#1 有对资源 Y 的同时读锁。由 AMF#3 发送的写锁请求 1444 使 AMF#4 分别向 AMF#2 和 AMF#1 发出释锁消息 1446 和 1448，结果是，锁释放消息 1450 和 1852 都发送给 AMF#4。在 AMF#4 授锁给 AMF#3 之前，AMF#2. 1 发送一请求读锁消息 1454，AMF#4 把它排入队列中。AMF#3 接收资源 Y 的授写锁消息 1456，内有一表示它在结束时应该释放该锁定的标志。当对资源 Y 完成时 AMF#3 发出一锁释放消息 1458。AMF#4 然后发出一授锁（读）消息 1460，通知 AMF#1 它已取得对资源 Y 的读锁。

这种锁定管理方法的一个好处在于：AMF 间发送的消息数只是锁定未认可数的一个函数。在群集配置中它不依赖于 AMF 数。扩展两控制器群集配置的第一种方法，例如如图 19 所示，以支持多于三个控制器将使仲裁消息数随群集配置中的 AMF 数增长。有可能有其他仲裁优化方法，但这两种方法说明了本发明的较佳实施例。

资源预取

在本发明的较佳实施例中，资源仲裁通过锁定预取的使用也得到优化。一 AMF 在就一锁定仲裁时可确定另外的预取资源。如果全部或一些预取资源未锁定，判定器也将为 AMF 锁定它们。这样，当 AMF 请对这些预取资源锁定时，（在某后来的时间）它能快速获得锁定（因为它已有该锁定）。

尽管通过例子并以具体实施例的形式描述了本发明，应该理解的是，本发明不限于所揭示的实施例。相反，本发明覆盖对本领域中的熟练人员而言显而易见的各种修改和类似的安排。因此，所附权利要求的范围应给予最大范围的理解经

包含所有这种修改和类似的安排。

说 明 书 附 图

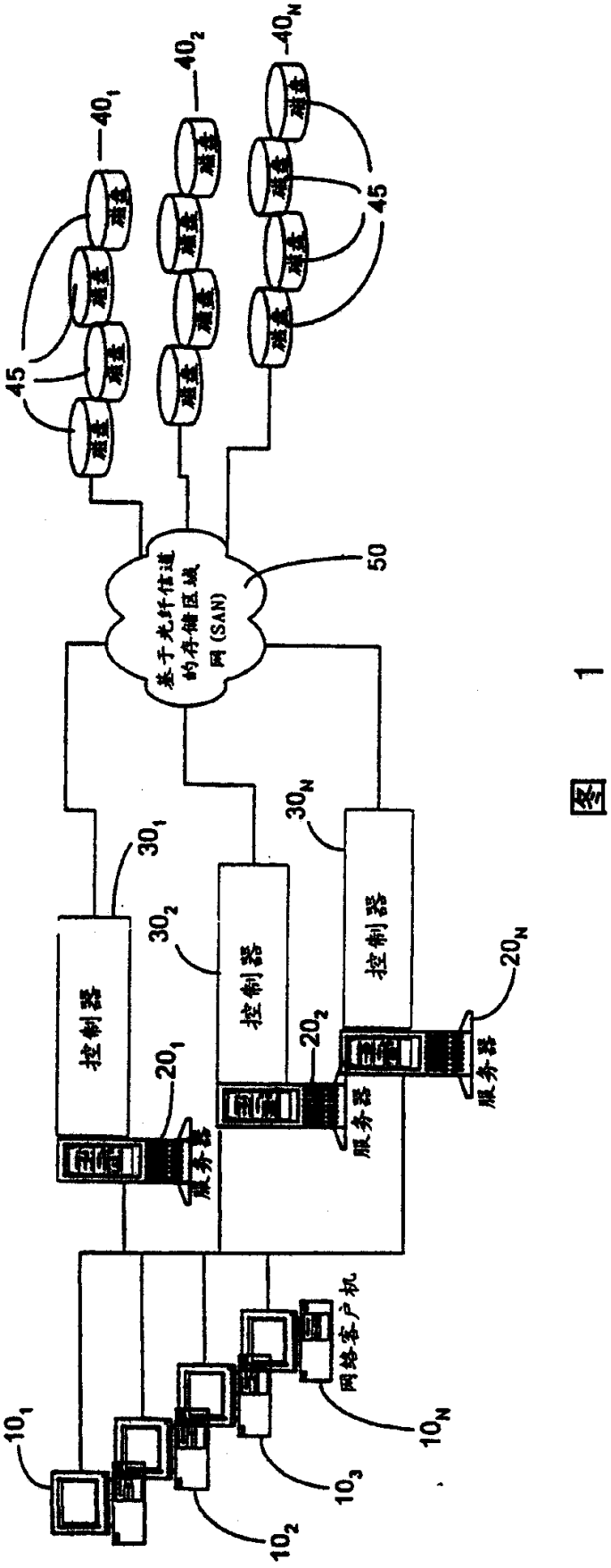

40_1
40_2
40_N
45
45
通道
基于光纤信道的存储区域网 (SAN)
50
30_1
30_2
30_N
控制器
控制器
控制器
20_1
20_2
20_N
服务器
服务器
服务器
网络客户机
10_1
10_2
10_3
10_N

图 1

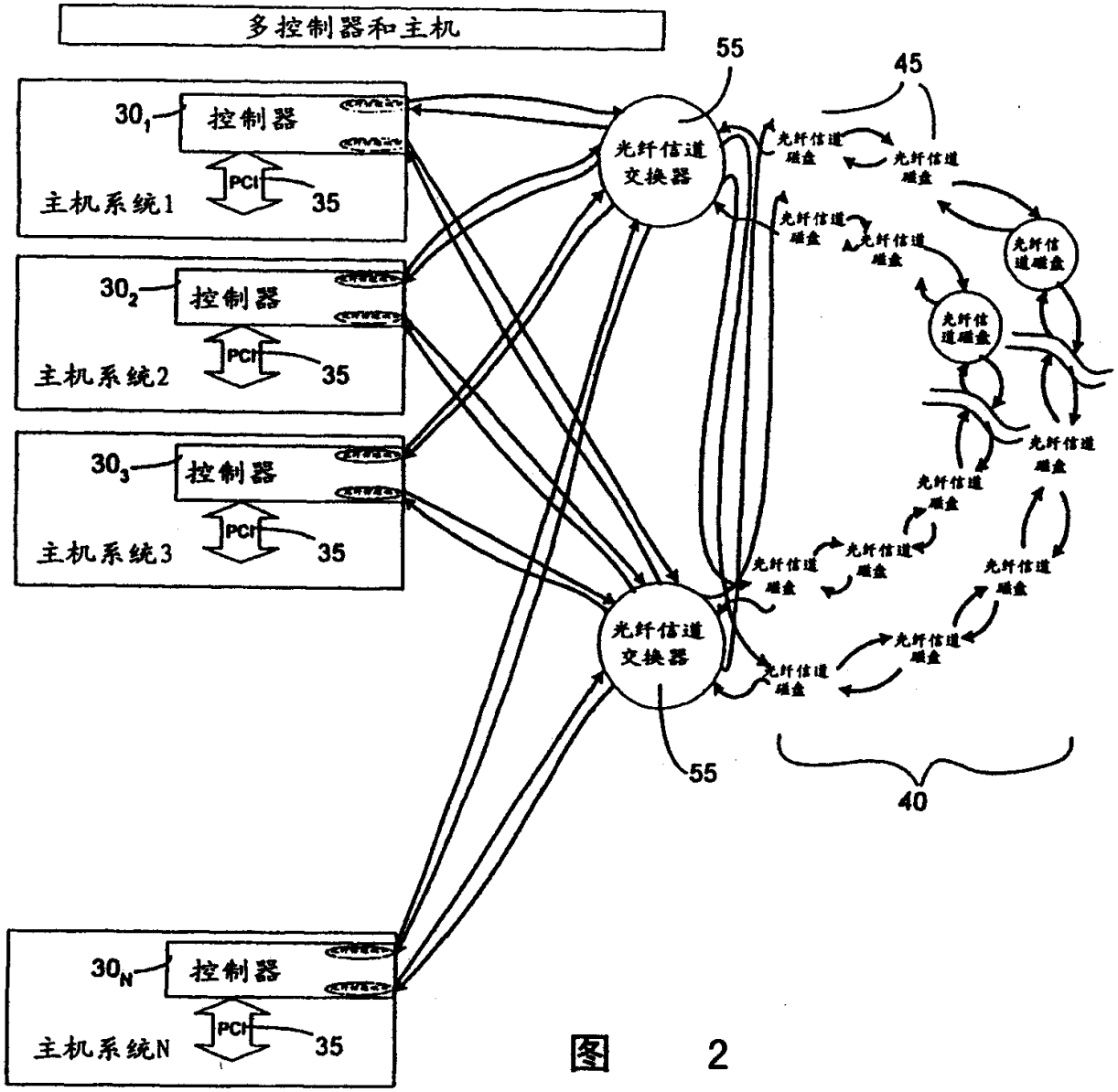

多控制器和主机
30_1
控制器
PCI
35
主机系统1
30_2
控制器
PCI
35
主机系统2
30_3
控制器
PCI
35
主机系统3
30_N
控制器
PCI
35
主机系统N
光纤信道交换器
光纤信道交换器
55
55
45
光纤信道磁盘
40

图 2

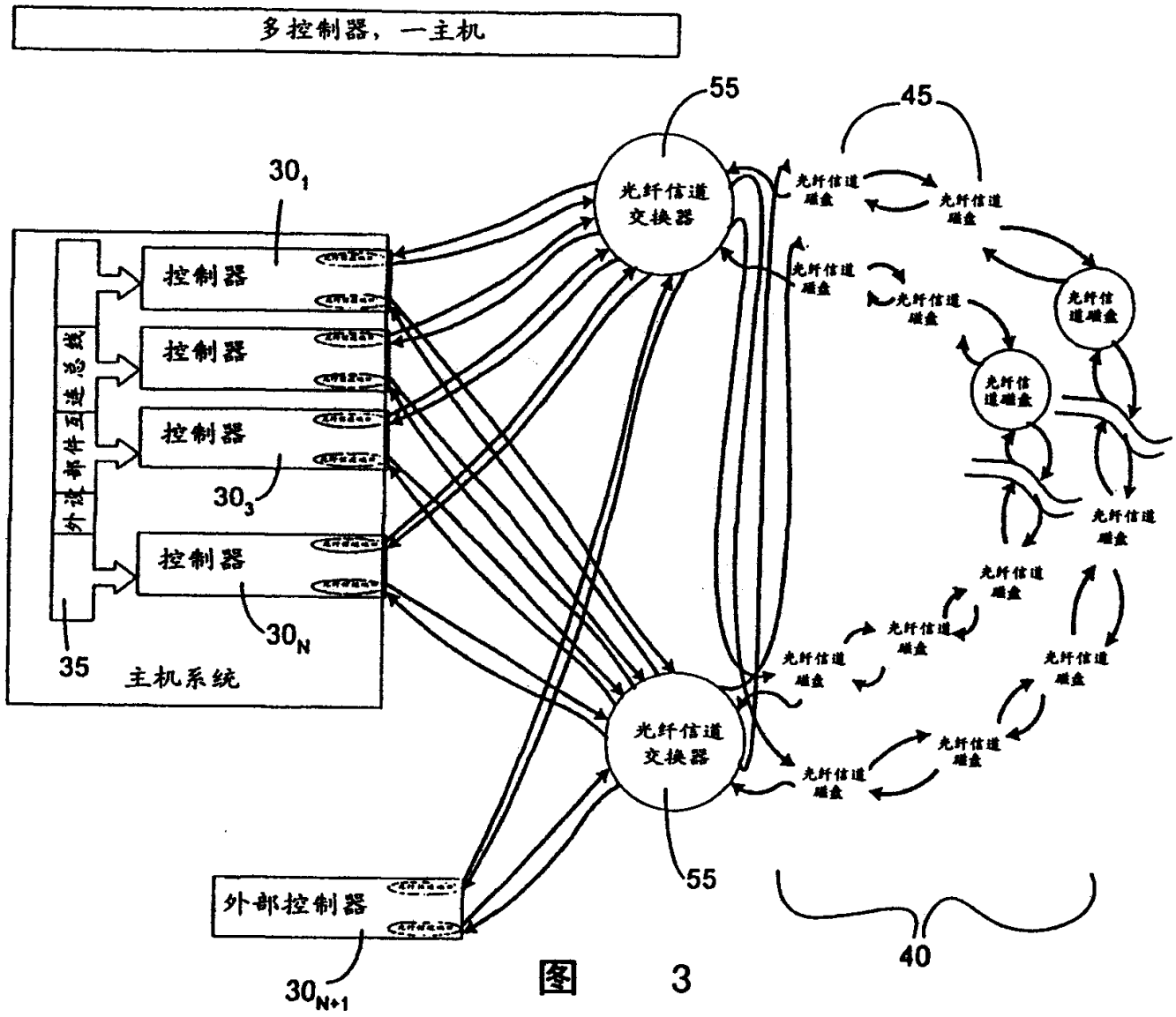

多控制器，一主机
55
45
30_1
光纤信道
交换器
控制器
控制器
控制器
30_3
控制器
30_N
外设部件互连总线
35
主机系统
光纤信道
交换器
55
外部控制器
30_{N+1}
光纤信道
硬盘
40

图 3

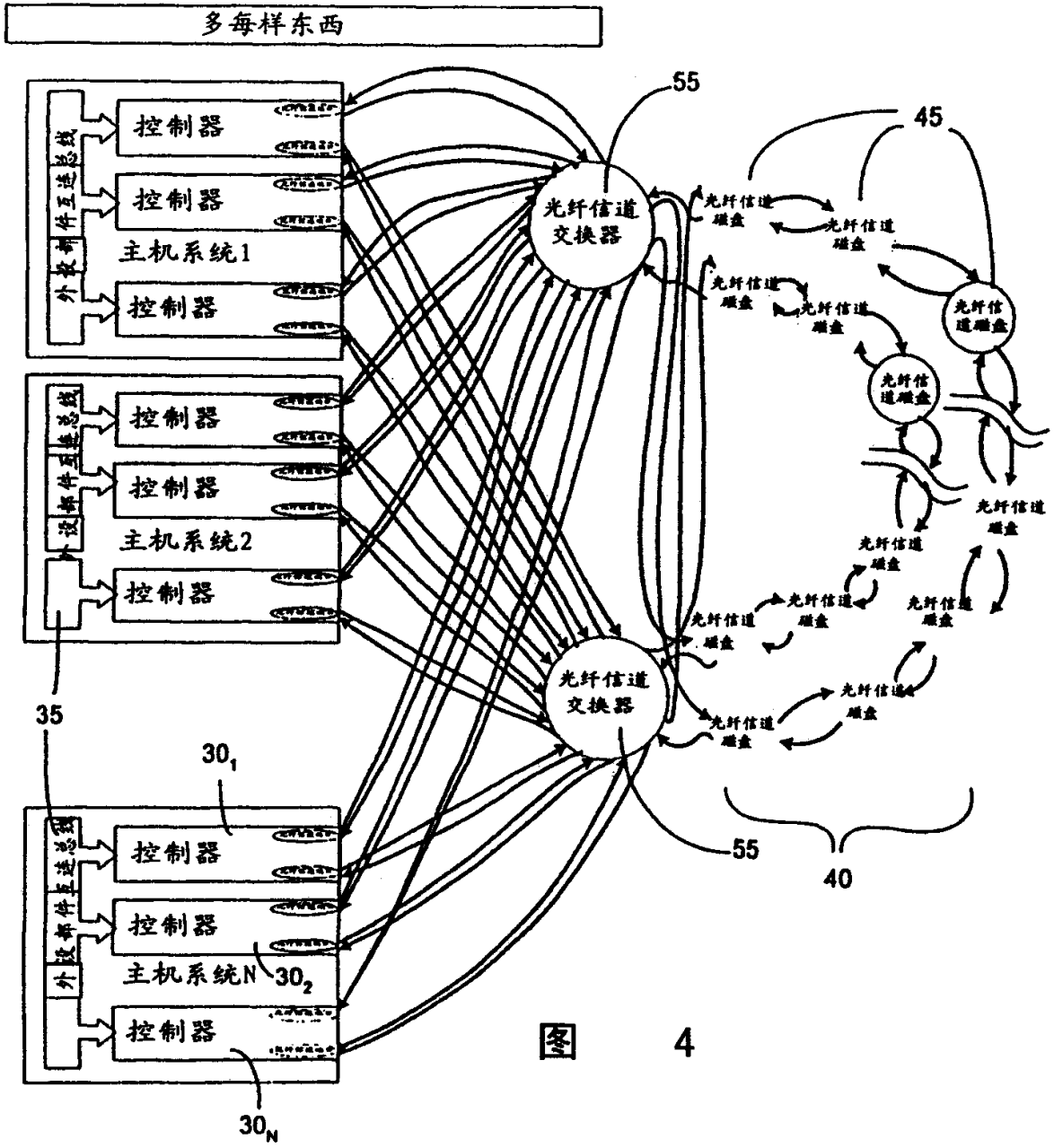

多每样东西
控制器
控制器
主机系统1
控制器
控制器
控制器
主机系统2
控制器
35
控制器
控制器
主机系统N
控制器
30_1
30_2
30_N
光纤信道交换器
光纤信道交换器
55
45
55
40
光纤信道磁盘

图 4

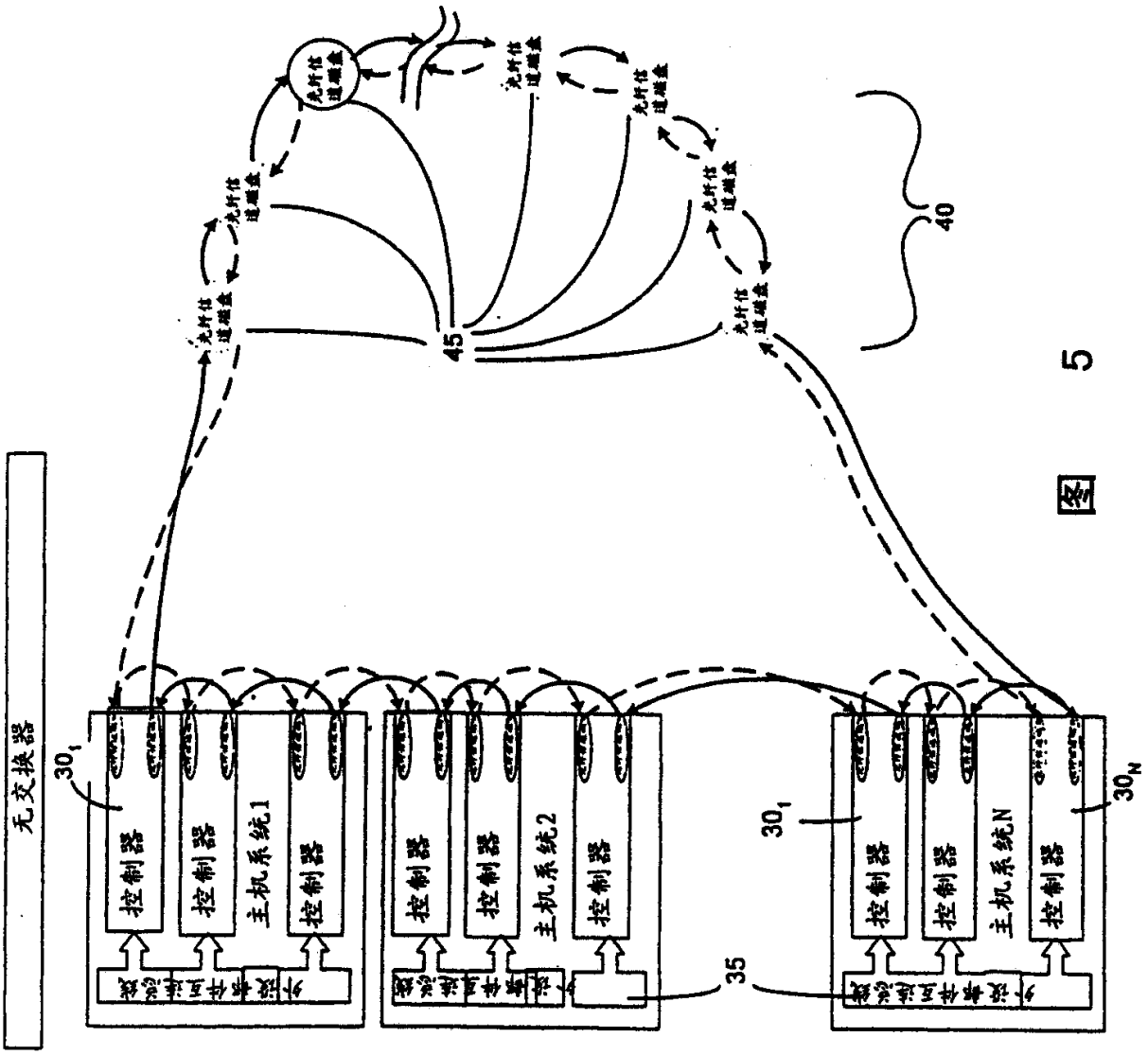

无交换器
30_1
控制器
控制器
主机系统1
控制器
控制器
控制器
主机系统2
控制器
30_1
控制器
控制器
主机系统N
控制器
30_N
35
45
40
光纤信道端盒

图 5

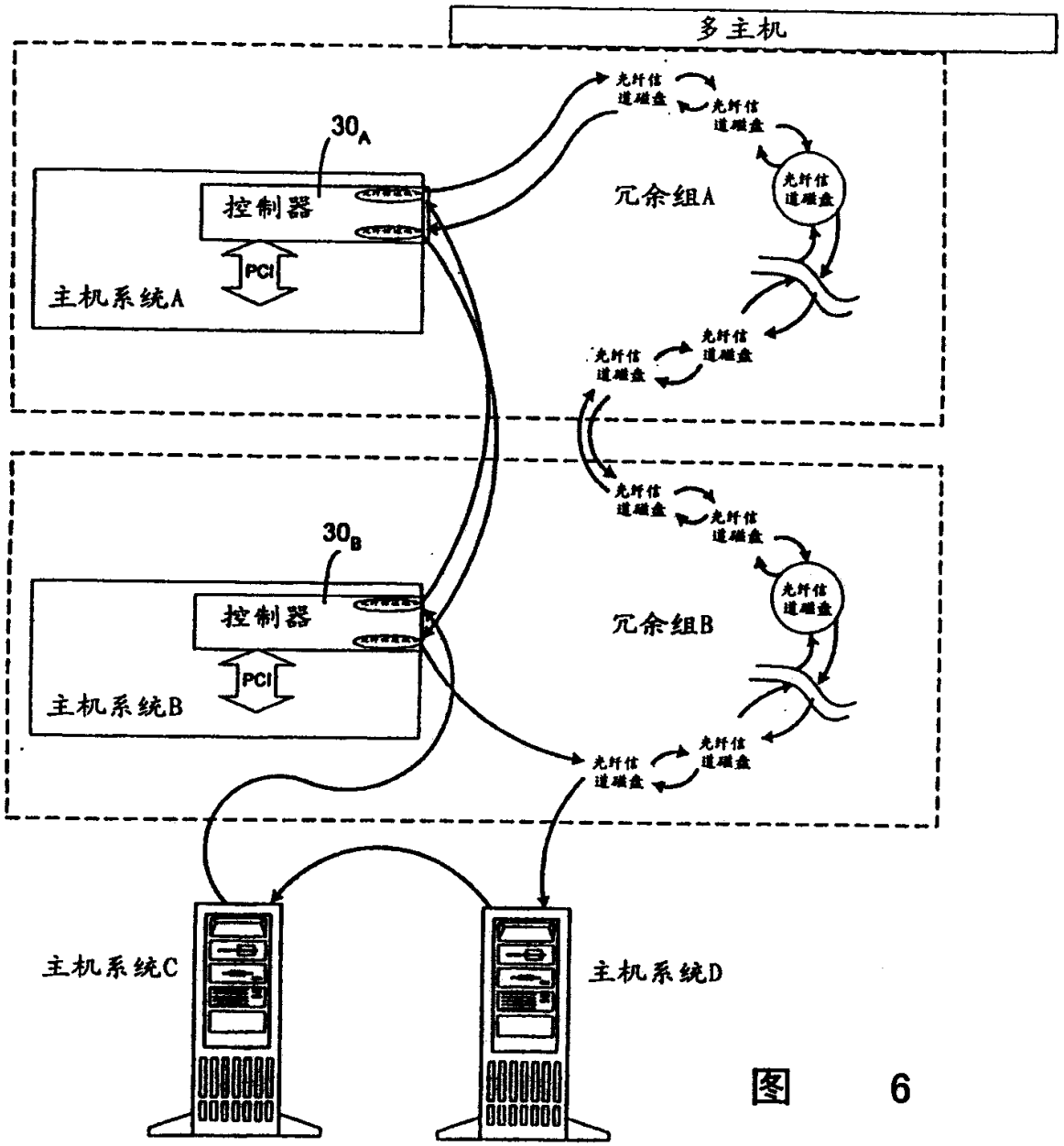

多主机
30A
控制器
PCI
主机系统A
光纤信道磁盘
冗余组A
30B
控制器
PCI
主机系统B
冗余组B
主机系统C
主机系统D

图 6

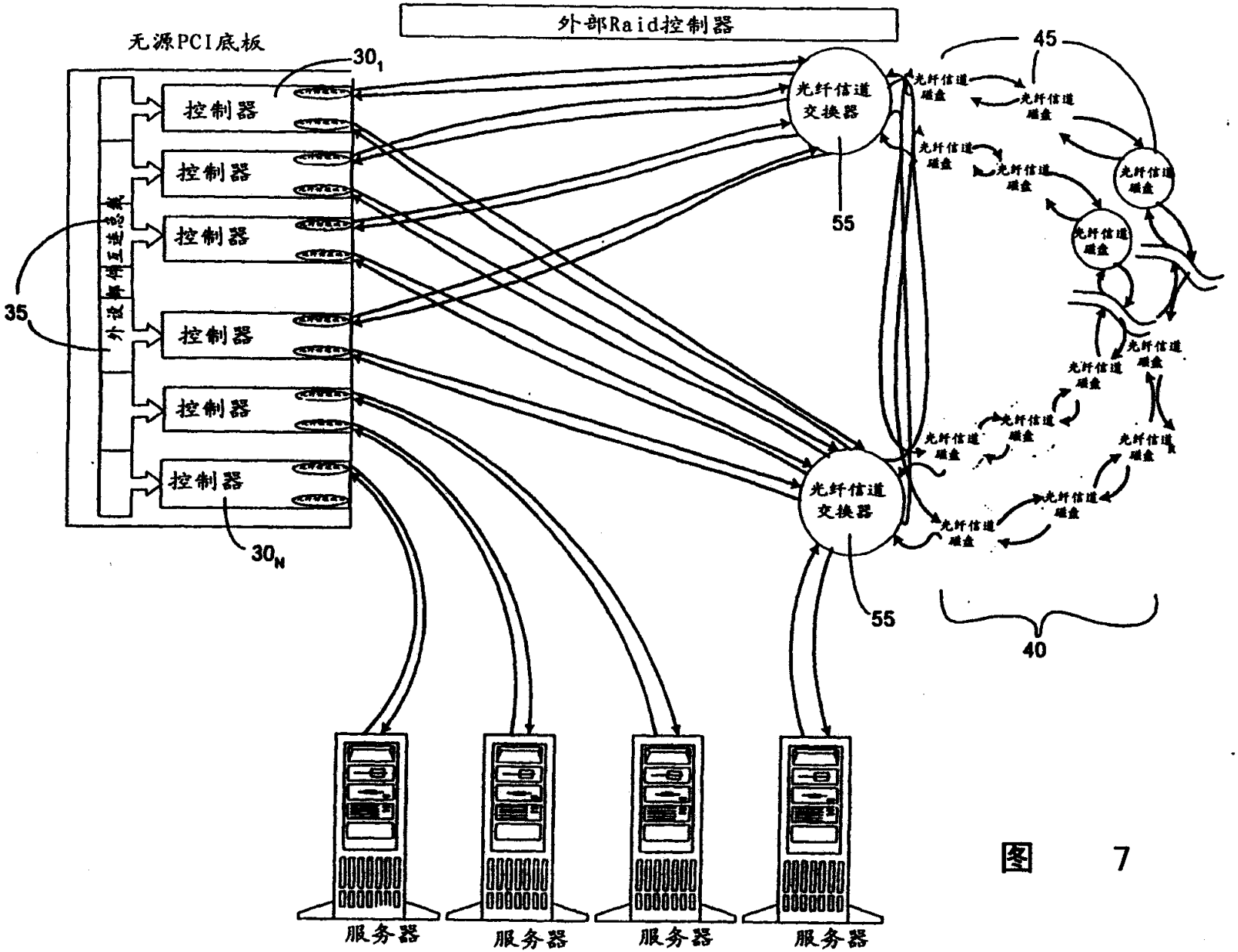

外部Raid控制器
无源PCI底板
30_1
控制器
控制器
控制器
控制器
控制器
控制器
外设部件互连总线
35
30_N
光纤信道交换器
光纤信道交换器
55
55
45
光纤信道磁盘
40
服务器
服务器
服务器
服务器

图 7

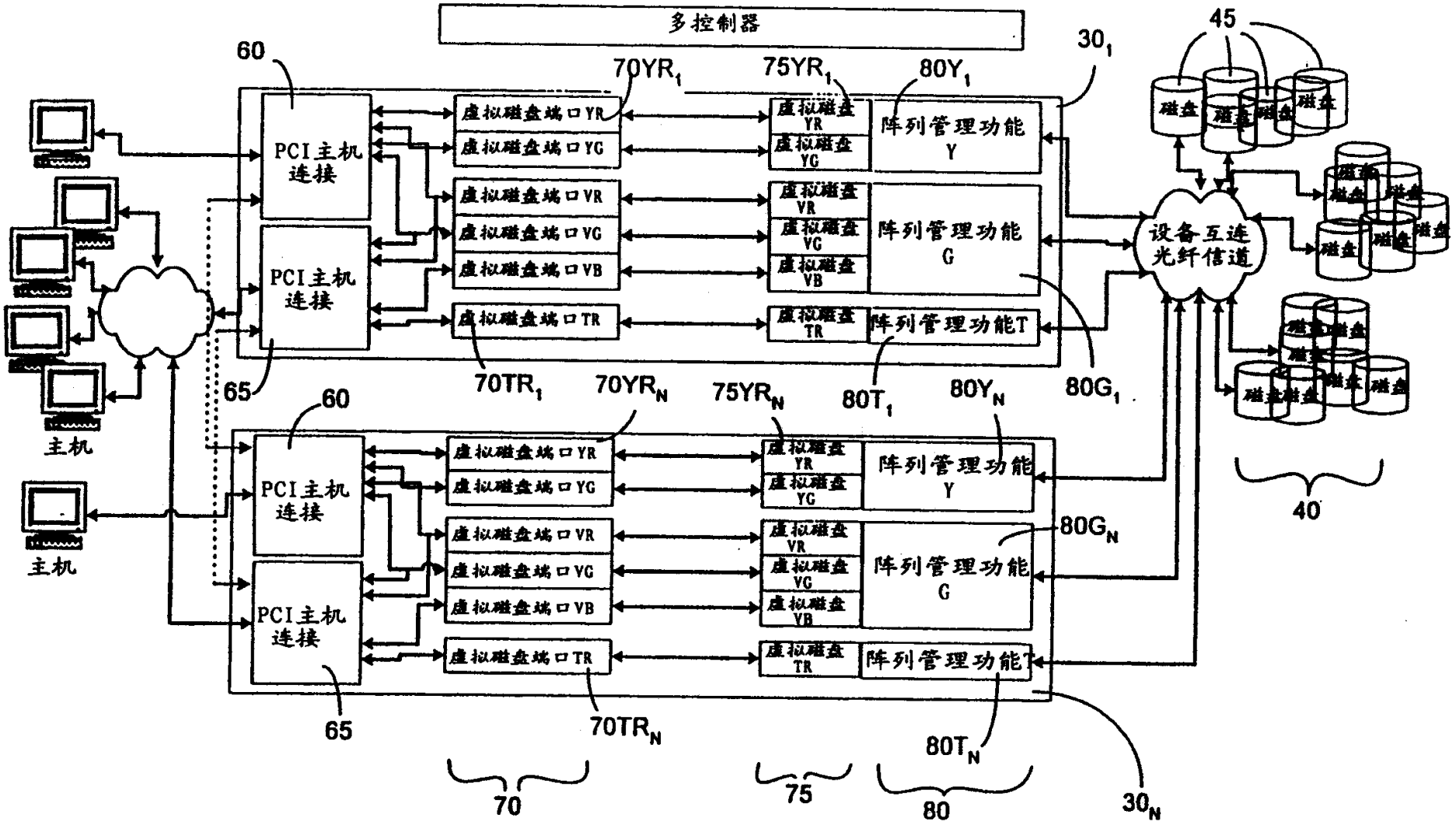

多控制器
60
PCI主机连接
PCI主机连接
65
虚拟磁盘端口YR
虚拟磁盘端口YG
虚拟磁盘端口VR
虚拟磁盘端口VG
虚拟磁盘端口VB
虚拟磁盘端口TR
70YR1
75YR1
80Y1
30 1
虚拟磁盘YR
虚拟磁盘YG
阵列管理功能Y
虚拟磁盘VR
虚拟磁盘VG
虚拟磁盘VB
阵列管理功能G
虚拟磁盘TR
阵列管理功能T
70TR1
80T1
80G1
70YRN
75YRN
80YN
80GN
80TN
30N
70
75
80
45
磁盘
设备互连光纤信道
40
主机
主机

图 8

基本同步序列

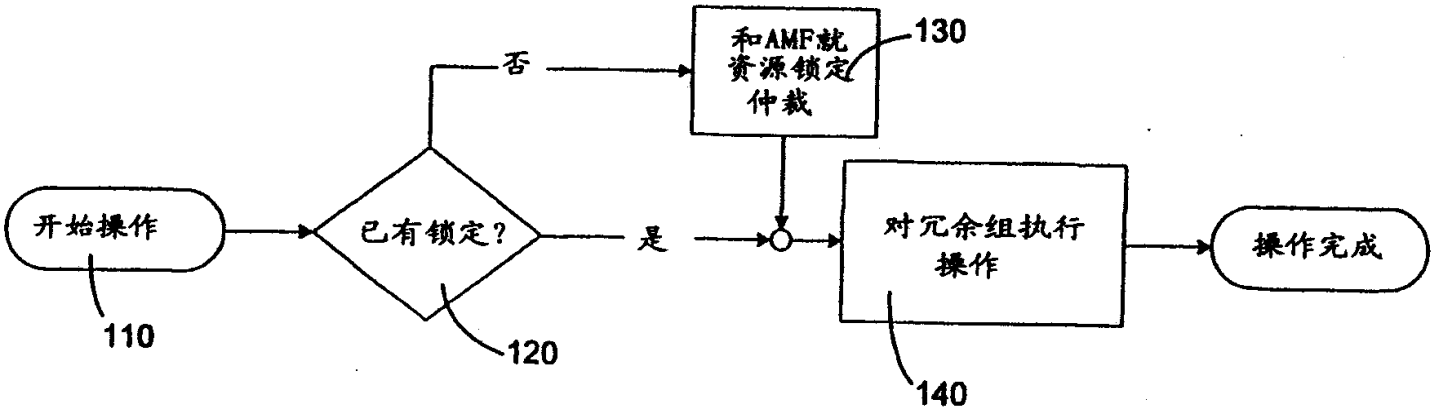


图 9

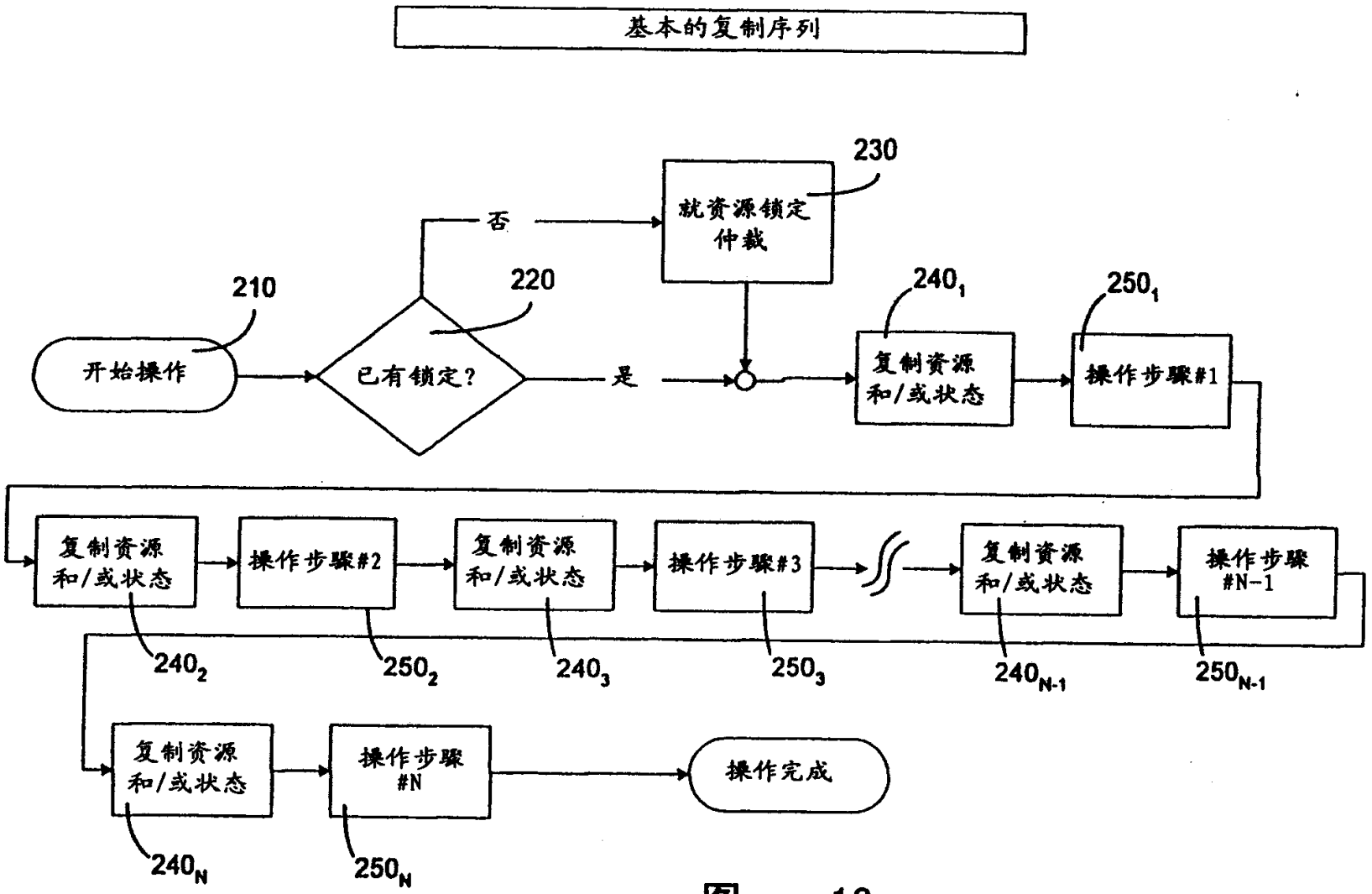

基本的复制序列
开始操作
210
已有锁定?
220
否
是
就资源锁定仲裁
230
复制资源和/或状态
240₁
操作步骤#1
250₁
复制资源和/或状态
240₂
操作步骤#2
250₂
复制资源和/或状态
240₃
操作步骤#3
250₃
复制资源和/或状态
240N-1
操作步骤#N-1
250N-1
复制资源和/或状态
240N
操作步骤#N
250N
操作完成

图 10

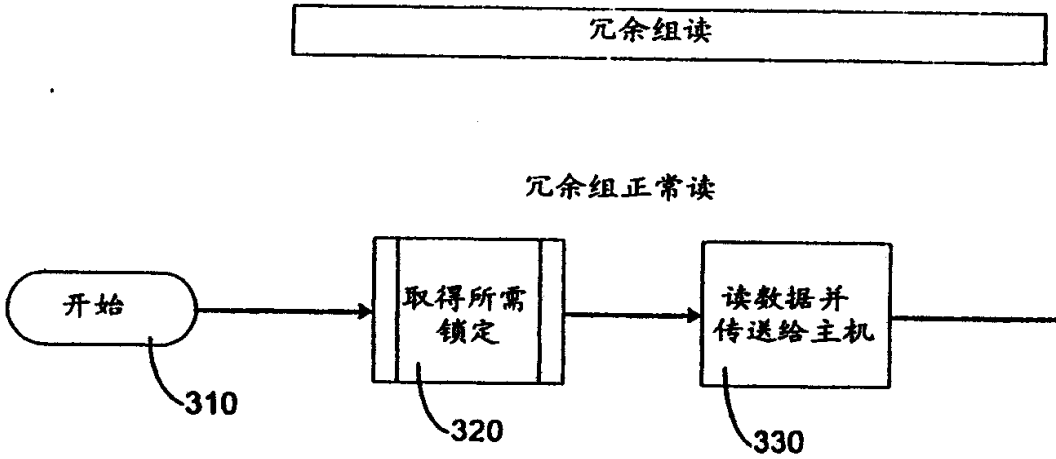

冗余组读
冗余组正常读
开始
取得所需锁定
读数据并传送给主机
完成
310
320
330

图 11a

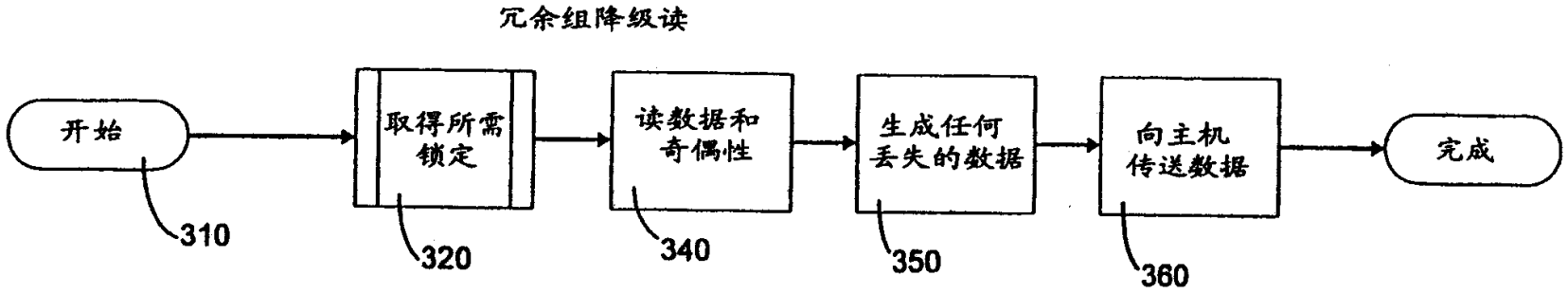

冗余组降级读
开始
取得所需锁定
读数据和奇偶性
生成任何丢失的数据
向主机传送数据
完成
310
320
340
350
360

图 11b

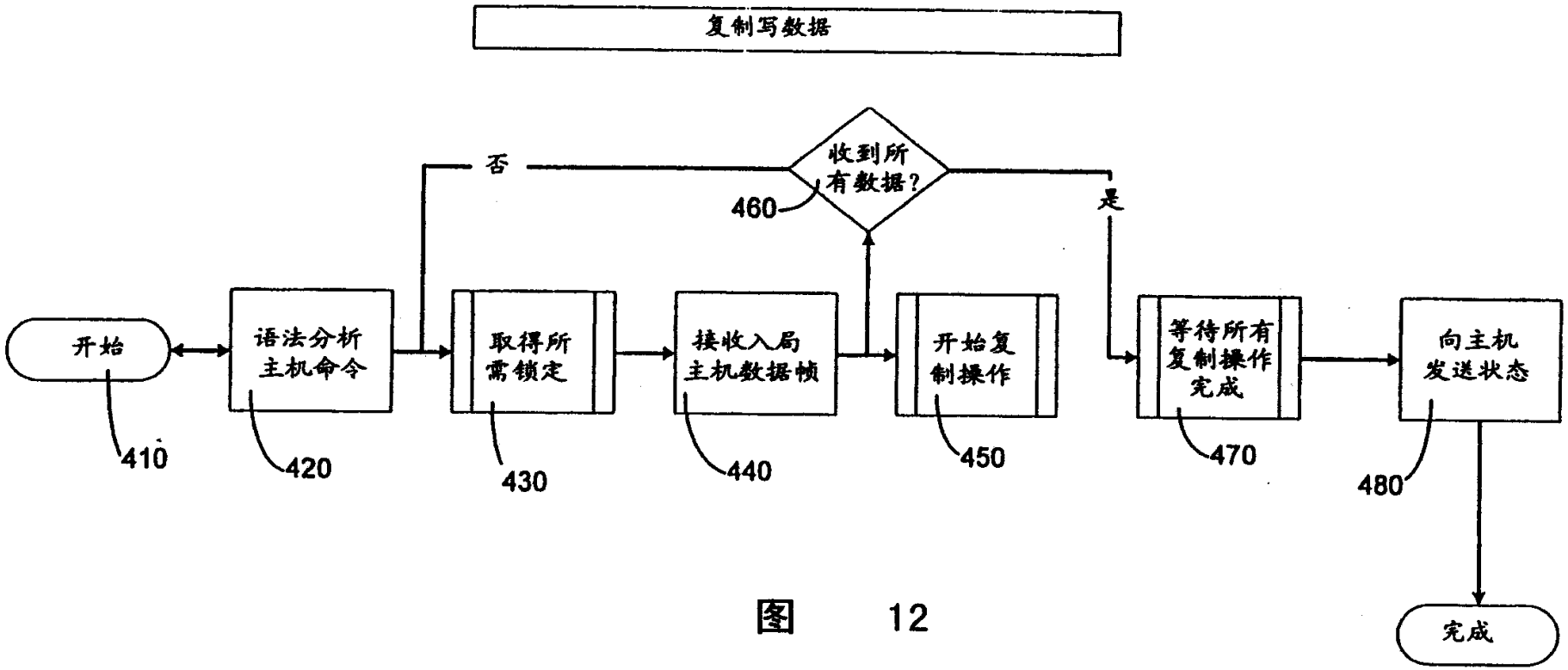

复制写数据
开始
410
语法分析
主机命令
420
取得所
需锁定
430
接收入局
主机数据帧
440
开始复
制操作
450
收到所
有数据?
460
否
是
等待所有
复制操作
完成
470
向主机
发送状态
480
完成

图 12

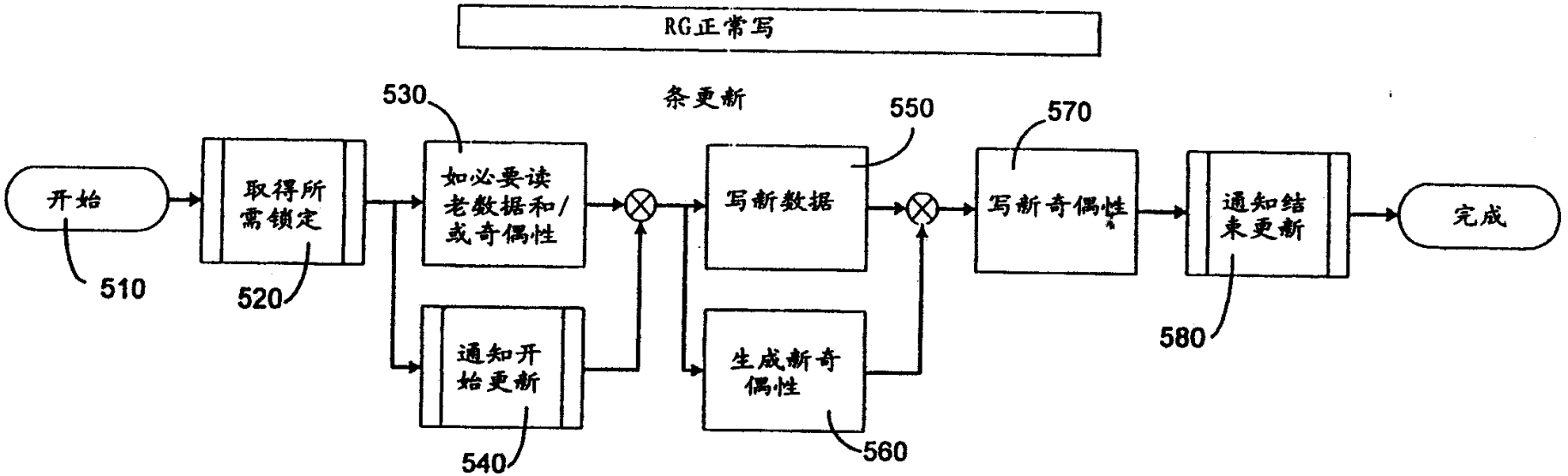

RG正常写
条更新
开始
取得所需锁定
如必要读老数据和/或奇偶性
写新数据
写新奇偶性
通知结束更新
完成
通知开始更新
生成新奇偶性
510
520
530
540
550
560
570
580

图 13a

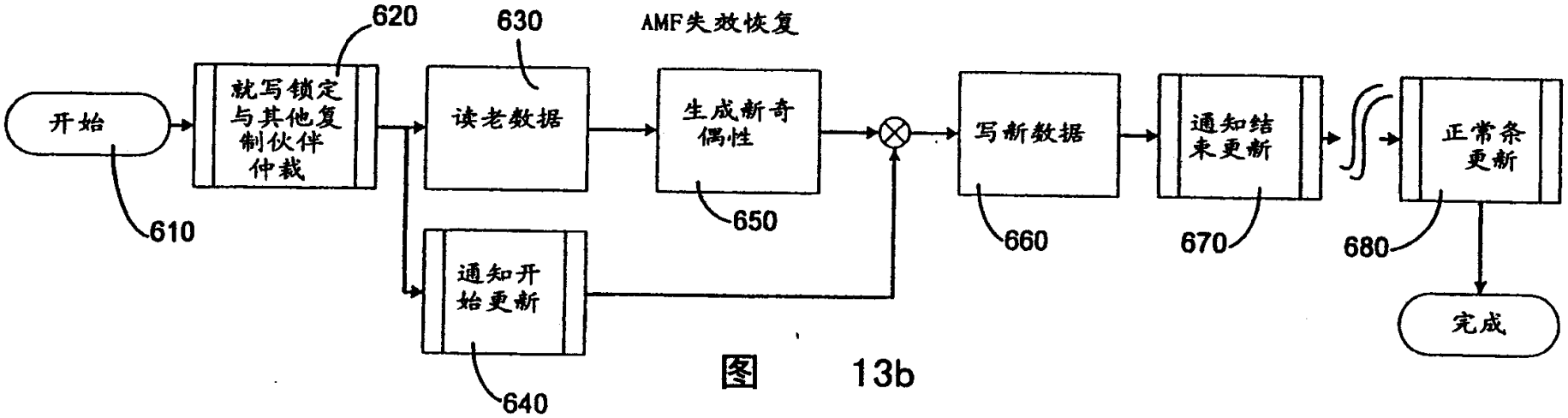

AMF失效恢复
开始
就写锁定与其他复制伙伴仲裁
读老数据
生成新奇偶性
写新数据
通知结束更新
正常条更新
完成
通知开始更新
610
620
630
640
650
660
670
680

图 13b

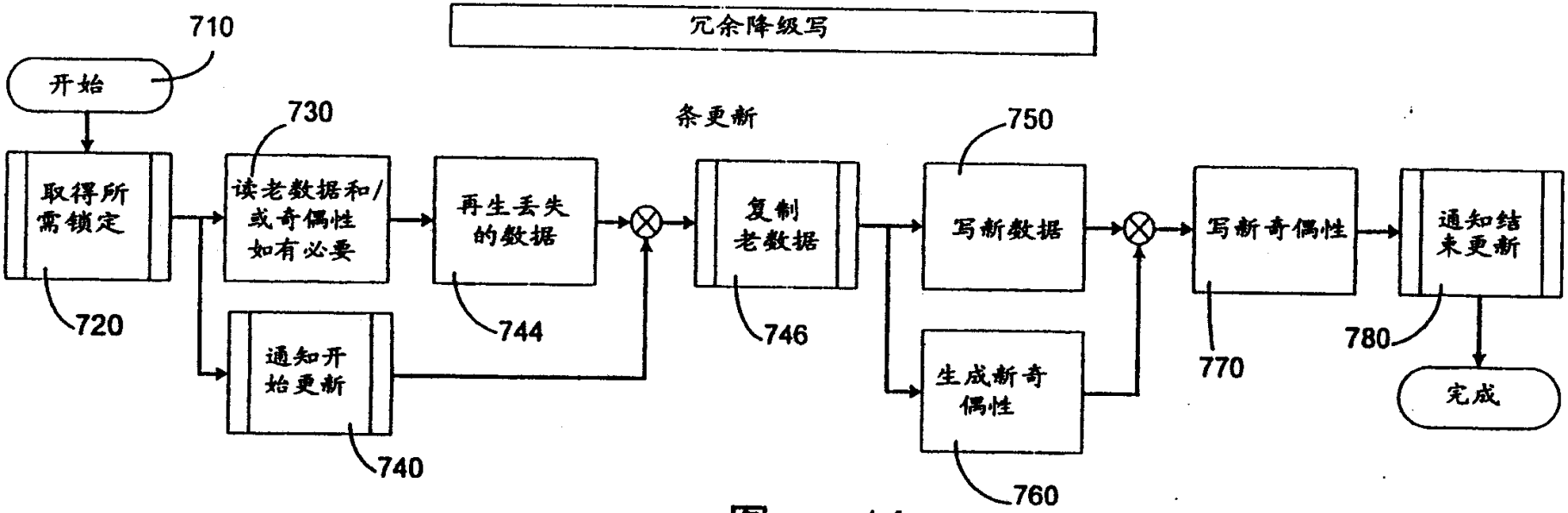

冗余降级写
710
开始
730
条更新
750
取得所需锁定
读老数据和/或奇偶性如有必要
再生丢失的数据
复制老数据
写新数据
写新奇偶性
通知结束更新
720
744
746
770
780
通知开始更新
生成新奇偶性
完成
740
760

图 14a

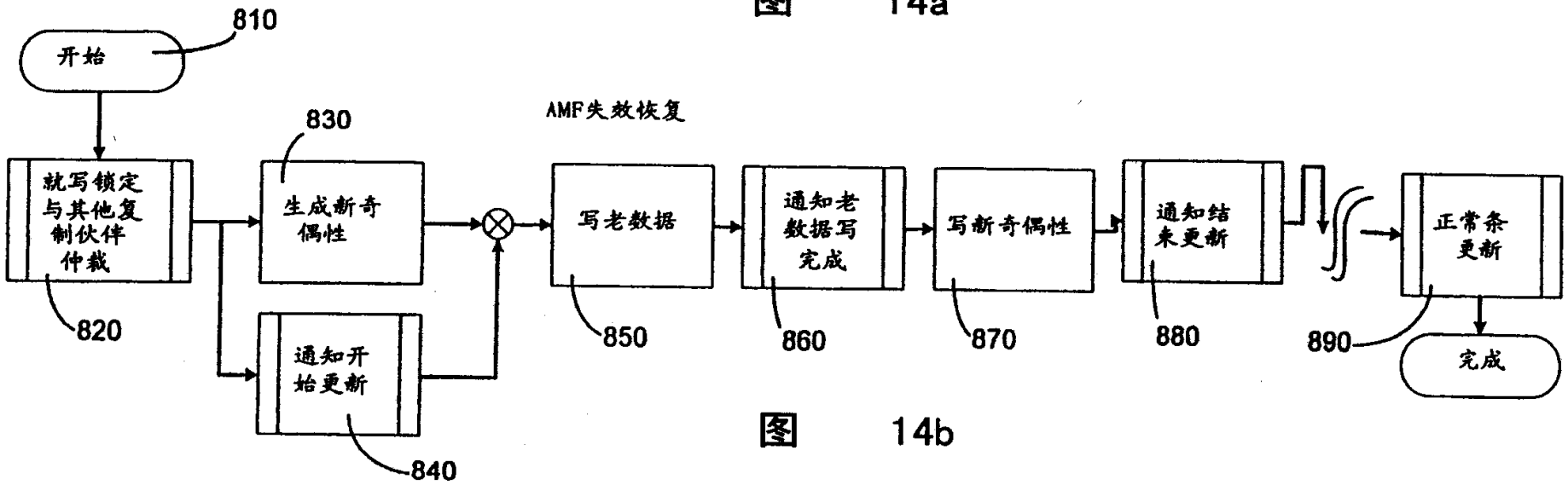

810
开始
AMF失效恢复
830
就写锁定与其他复制伙伴仲裁
生成新奇偶性
写老数据
通知老数据写完成
写新奇偶性
通知结束更新
正常条更新
820
850
860
870
880
890
通知开始更新
完成
840

图 14b

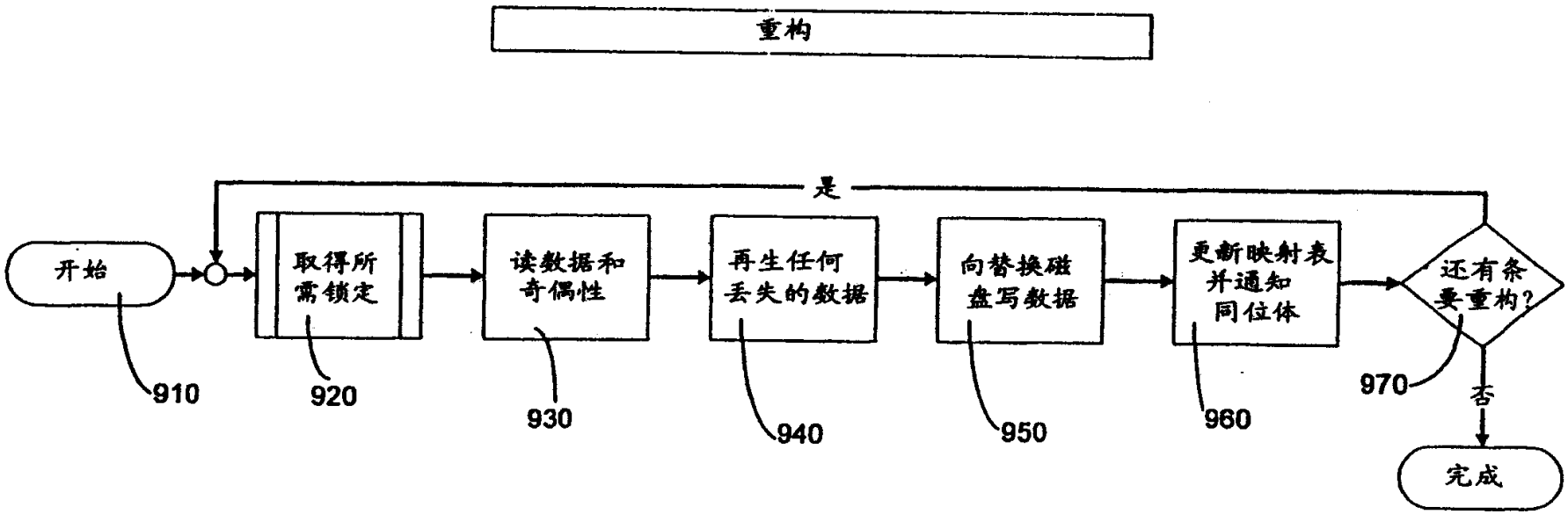

重构
是
开始
取得所需锁定
读数据和奇偶性
再生任何丢失的数据
向替换磁盘写数据
更新映射表并通知同位体
还有条要重构?
否
完成
910
920
930
940
950
960
970

图 15

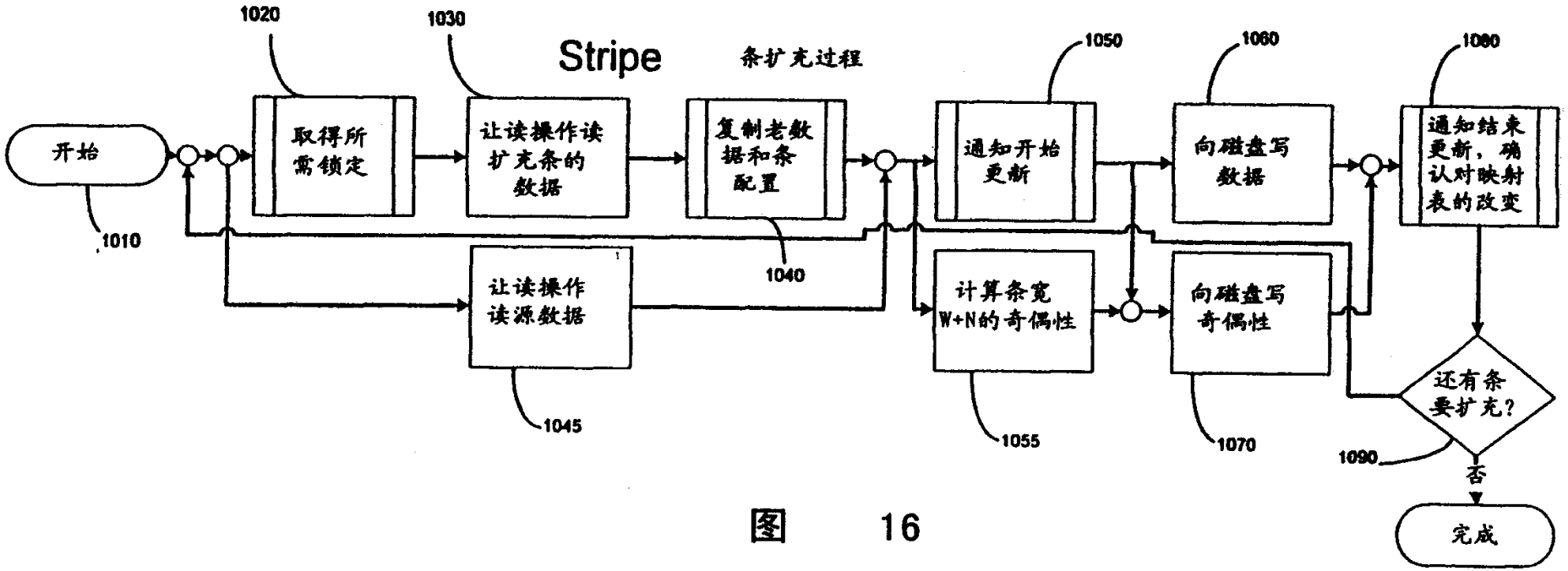

Stripe
条扩充过程
开始
1010
取得所需锁定
1020
让读操作读扩充条的数据
1030
复制老数据和条配置
1040
让读操作读源数据
1045
通知开始更新
1050
计算条宽W+N的奇偶性
1055
向磁盘写数据
1060
向磁盘写奇偶性
1070
通知结束更新，确认对映射表的改变
1080
还有条要扩充?
1090
否
完成

图 16

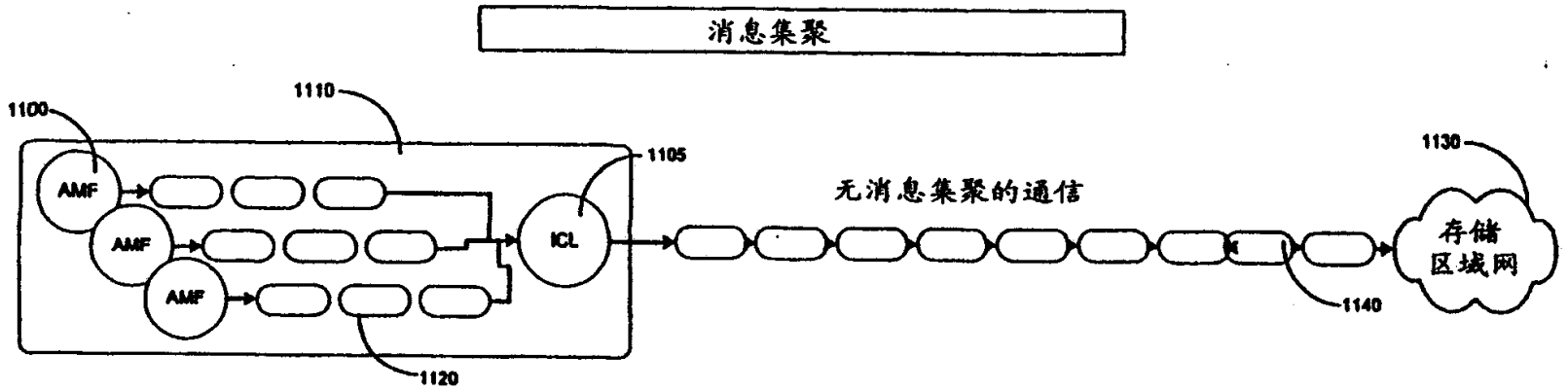

消息集聚
1100
1110
AMF
AMF
AMF
1105
ICL
无消息集聚的通信
1130
存储
区域网
1140
1120

图 17a

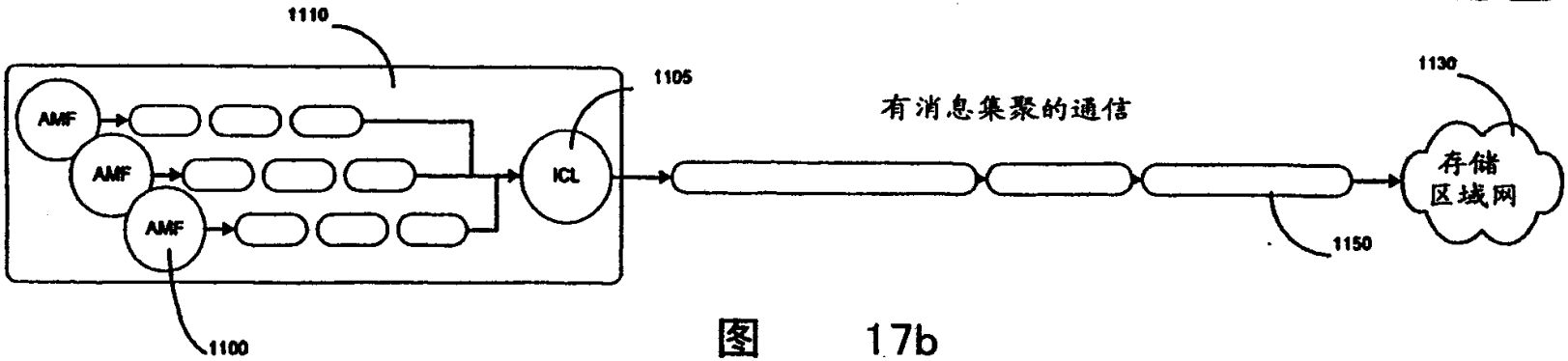

1110
AMF
AMF
AMF
1105
ICL
有消息集聚的通信
1130
存储
区域网
1150
1100

图 17b

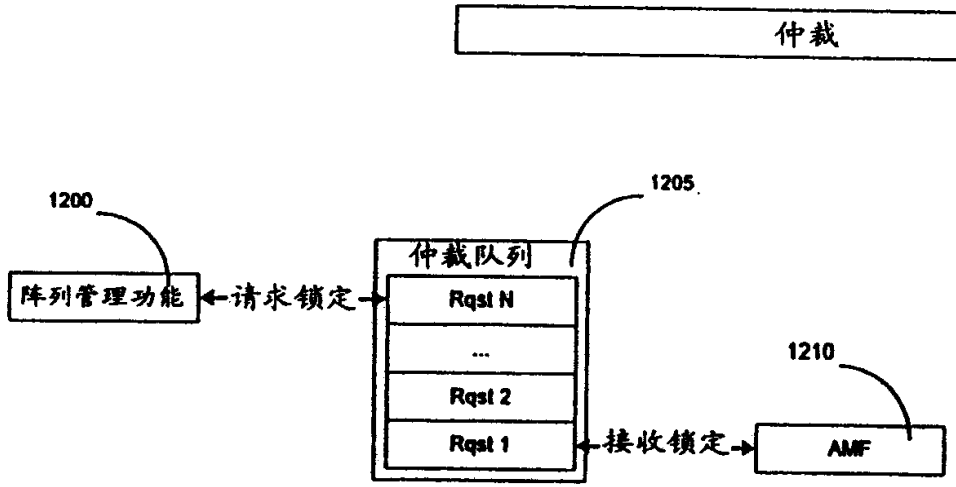


图 18a

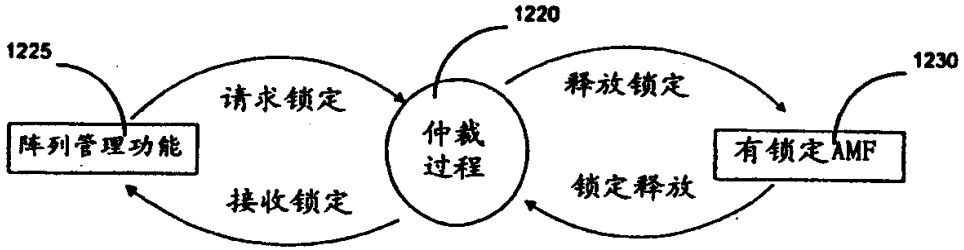


图 18b

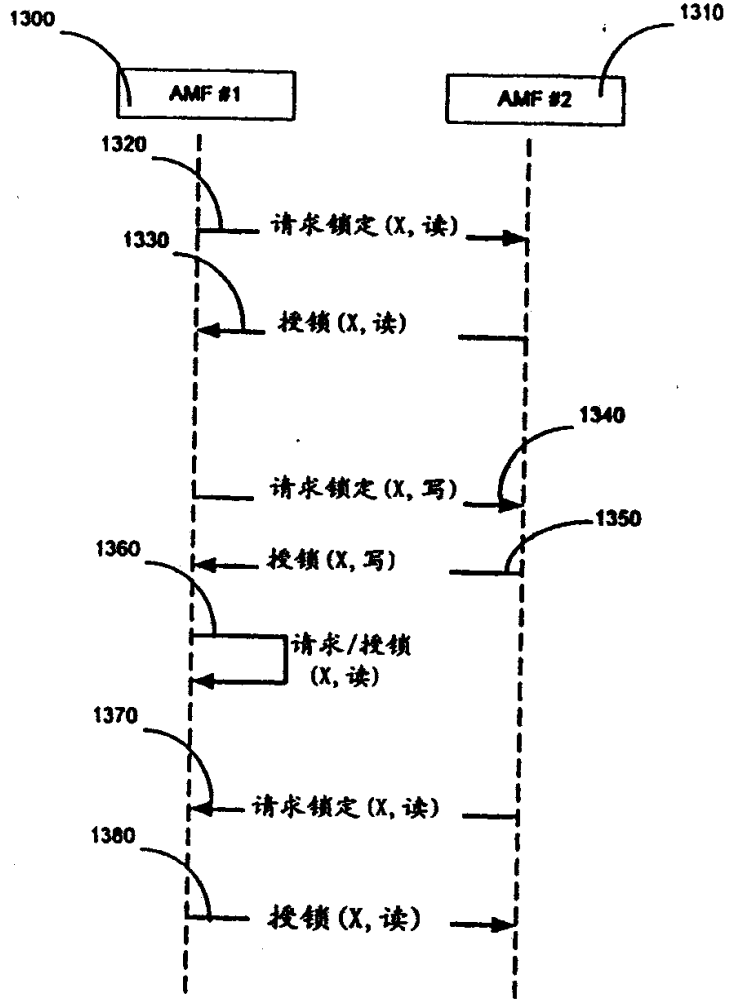


图 19

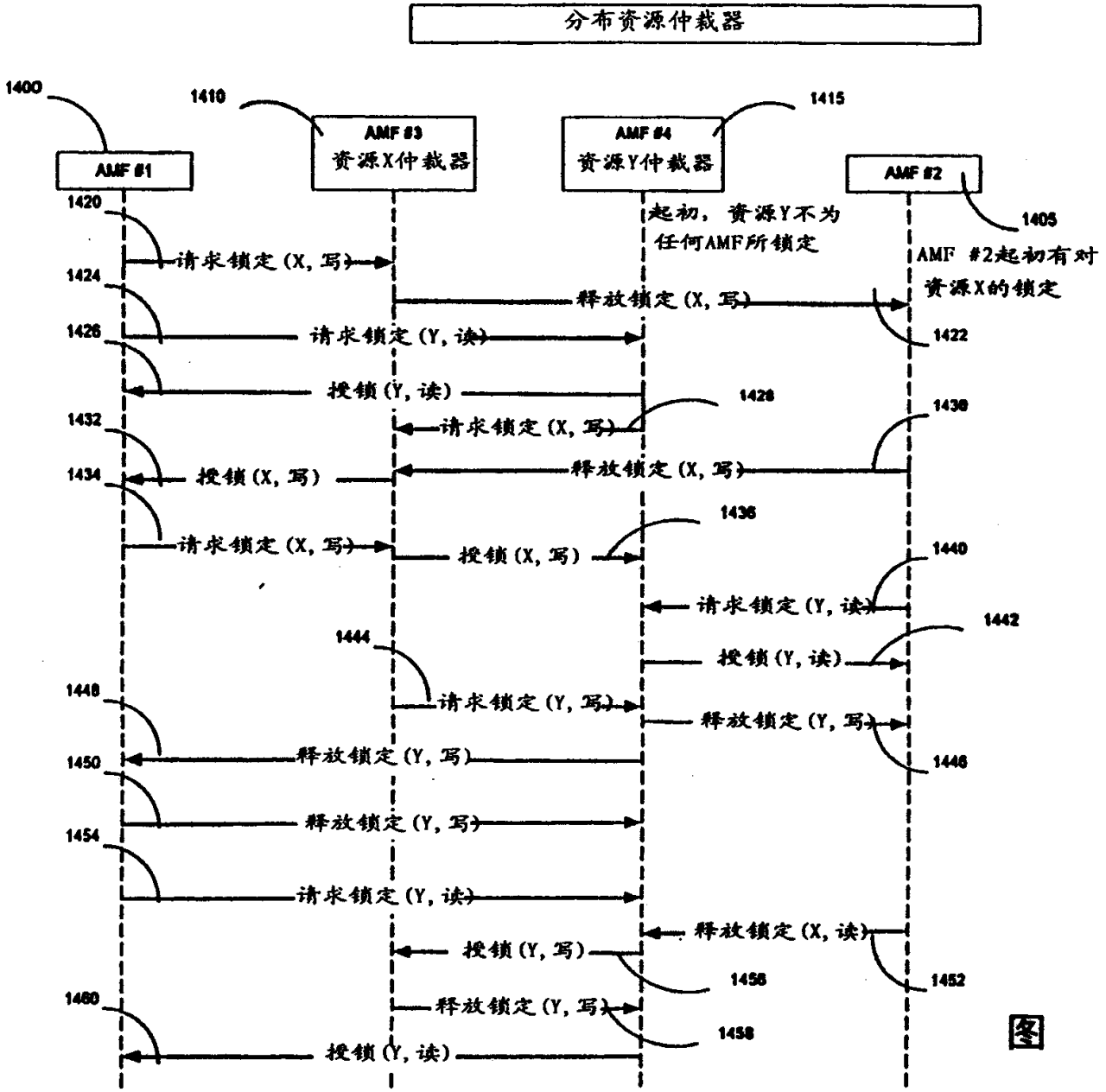

分布资源仲裁器
1400
AMF #1
1410
AMF #3
资源X仲裁器
1415
AMF #4
资源Y仲裁器
AMF #2
1405
起初，资源Y不为
任何AMF所锁定
AMF #2起初有对
资源X的锁定
1420
请求锁定(X, 写)
1424
释放锁定(X, 写)
1422
1426
请求锁定(Y, 读)
授锁(Y, 读)
1428
请求锁定(X, 写)
1430
1432
释放锁定(X, 写)
1434
授锁(X, 写)
请求锁定(X, 写)
1436
授锁(X, 写)
1440
请求锁定(Y, 读)
1442
1444
授锁(Y, 读)
请求锁定(Y, 写)
1448
释放锁定(Y, 写)
1446
1450
释放锁定(Y, 写)
1454
请求锁定(Y, 读)
释放锁定(X, 读)
授锁(Y, 写)
1456
1452
1460
释放锁定(Y, 写)
1458
授锁(Y, 读)

图 20