

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 3/06 (2006.01)



[12] 发明专利说明书

专利号 ZL 200710126604.0

[45] 授权公告日 2009 年 5 月 13 日

[11] 授权公告号 CN 100487639C

[22] 申请日 2007.4.4

[21] 申请号 200710126604.0

[30] 优先权

[32] 2006.4.4 [33] JP [31] 103256/2006

[73] 专利权人 株式会社东芝

地址 日本东京都

共同专利权人 东芝解决方案株式会社

[72] 发明人 小原诚

[56] 参考文献

CN 1606290 A 2005.4.13

JP9-62452A 1997.3.7

CN1540524A 2004.10.27

WO00/60481A1 2000.10.12

审查员 董洪梅

[74] 专利代理机构 北京市中咨律师事务所
代理人 杨晓光 于 静

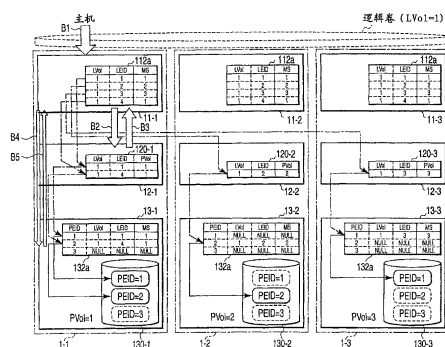
权利要求书 7 页 说明书 30 页 附图 17 页

[54] 发明名称

用于处理来自主机的访问请求的存储系统及方法

[57] 摘要

节点(1-1至1-3)包括控制器(11-1至11-3)、管理服务器(12-1至12-3)以及存储器(13-1至13-3)。每个控制器(11-1至11-3)包括第一映射表(112a)，其指示与构成被提供给主机的逻辑卷的逻辑盘区对应的管理服务器。管理服务器(12-1至12-3)包括第二映射表(120-1至120-3)，其指示哪个节点包括存储了与所述管理服务器对应的逻辑盘区的存储器。存储器(13-1至13-3)包括第三映射表(132a)，其指示在存储器(13-1至13-3)中包括的物理卷(130-1至130-3)的哪个物理盘区存储了所述逻辑盘区。



1. 一种存储系统，其特征在于，包括：

多个节点，该节点的至少一个包括具有物理卷的存储器，该存储器在所述物理卷的物理盘区中映射构成逻辑卷的逻辑盘区并在该物理卷中存储该逻辑盘区；以及

连接所述节点和主机的接口，

其中：

所述节点的至少两个包括为所述主机提供所述逻辑卷并接收来自该主机的访问的控制器；

所述节点的所述至少一个或所述节点的至少另一个包括：管理服务器，其进行位置管理服务，以答复来自任何所述控制器的关于存储所述逻辑卷的逻辑盘区的存储器的位置的询问；

每个所述控制器包括第一映射表，其指示与所述逻辑卷的逻辑盘区对应的管理服务器；

所述管理服务器包括第二映射表，其指示哪个节点包括了存储与所述管理服务器对应的逻辑盘区的存储器；

所述存储器包括第三映射表，其指示在该存储器中包括的物理卷的哪个物理盘区存储了所述逻辑盘区；

当所述控制器的每一个从所述主机接收了指定所述逻辑卷中的逻辑盘区作为访问目的地的访问请求时，所述控制器的所述每一个参考所述第一映射表以指定与所述访问目的地的逻辑盘区对应的管理服务器，并询问所述指定的管理服务器哪个存储器存储了所述访问目的地的逻辑盘区；

当所述管理服务器接收到来自所述控制器中的一个的询问时，该管理服务器参考所述第二映射表以指定被询问的存储器，并作为对所述询问的答复，将指示该指定的存储器的信息通知给作为询问源的所述控制器中的所述一个；

每个所述控制器请求由作为对所述询问的答复的所述信息指示的存储

器，以得到对由所述主机指定的逻辑盘区的访问；以及

当所述控制器中的一个请求所述存储器以得到对逻辑盘区的访问时，所述存储器参考所述第三映射表以在所述物理卷中指定与所述指定的逻辑盘区对应的物理盘区，得到对该指定的物理盘区的访问，并答复作为访问请求源的所述控制器中的所述一个。

2. 根据权利要求1所述的存储系统，其特征在于：

不同于所述节点的所述至少一个的至少一个节点，包括与所述存储器不同的另一个存储器，所述另一个存储器包括与所述第三映射表不同的另一个第三映射表；以及

当发生盘区迁移以使得将形成逻辑盘区的物理盘区中的数据从所述存储器迁移到所述另一个存储器的另一个物理盘区时，所述第三映射表和所述另一个第三映射表被更新以反映该盘区迁移，并且在所述管理服务器中包括的所述第二映射表被更新以显示作为存储所述逻辑盘区的存储器的迁移目的地的所述另一个存储器。

3. 根据权利要求1所述的存储系统，其特征在于：

不同于所述节点的所述至少一个的至少一个节点，包括与所述存储器不同的另一个存储器，所述另一个存储器包括与所述第三映射表不同的另一个第三映射表；以及

当发生盘区复制以将形成逻辑盘区的物理盘区中的数据从所述存储器复制到所述另一个存储器的另一个物理盘区时，所述第三映射表被更新以反映该盘区复制，并且在所述管理服务器中包括的所述第二映射表被更新以显示复制目的地的所述另一个存储器以及作为存储所述逻辑盘区的存储器的复制源的所述存储器。

4. 根据权利要求1所述的存储系统，其特征在于：

所述节点的所述至少一个或不同于所述节点的所述至少另一个的至少另一个节点，包括与所述管理服务器不同的另一个管理服务器，所述另一个管理服务器进行位置管理服务以答复来自所述控制器的任何一个的关于存储了构成所述逻辑卷的逻辑盘区的存储器的位置的询问，所述另一个管

理服务器包括与所述第二映射表不同的另一个第二映射表；以及

当发生位置管理服务迁移以使得将信息迁移到所述另一个第二映射表时，所述控制器的每一个的所述第一映射表、所述第二映射表以及所述另一个第二映射表被更新以反映该位置管理服务迁移，所述信息被包括在所述第二映射表中并指示哪个节点包括了存储逻辑盘区的存储器。

5. 根据权利要求1所述的存储系统，其特征在于：

所述节点的所述至少一个或不同于所述节点的所述至少另一个的至少另一个节点，包括与所述管理服务器不同的另一个管理服务器，所述另一个管理服务器进行位置管理服务以答复来自所述控制器的任何一个的关于存储了构成所述逻辑卷的逻辑盘区的存储器的位置的询问，所述另一个管理服务器包括与所述第二映射表不同的另一个第二映射表；以及

当发生位置管理服务复制以将信息复制到所述另一个第二映射表时，所述控制器的每一个的所述第一映射表和所述另一个第二映射表被更新以反映该位置管理服务复制，所述信息被包括在所述第二映射表中并指示哪个节点包括了存储逻辑盘区的存储器。

6. 根据权利要求1所述的存储系统，其特征在于：

所述第三映射表仅仅指示在创建所述逻辑卷的初始状态存储了构成该逻辑卷的逻辑盘区的物理卷；以及

所述存储器将该存储器中包括的物理卷中的空的物理盘区分配给所述指定的逻辑盘区，并且当作为参考所述第三映射表的结果确定所述物理卷中的物理盘区不与所述指定的逻辑盘区相关联时，更新该第三映射表。

7. 根据权利要求1所述的存储系统，其特征在于：

所述第二映射表指示，在所述逻辑卷被创建的初始状态，对于与包括所述第二映射表的所述管理服务器相对应的逻辑盘区，未分配存储器；

所述第三映射表指示，在所述逻辑卷被创建的初始状态，未分配物理卷和物理盘区给构成所述逻辑卷的逻辑盘区；

所述管理服务器分配存储器给所述访问目的地的逻辑盘区并且当作为参考所述第二映射表的结果确定该存储器未被分配给该逻辑盘区时更新该

第二映射表；以及

所述存储器将该存储器中包括的物理卷中的空的物理盘区分配给所述指定的逻辑盘区并且当作为参考所述第三映射表的结果确定该物理卷中的物理盘区不与所述指定的逻辑盘区相关联时，更新该第三映射表。

8. 根据权利要求1所述的存储系统，其特征在于：

所述控制器的每一个包括高速缓冲存储器，其存储有关所述询问以及对来自所述管理服务器的询问进行答复的信息；以及

在每个所述控制器询问所述管理服务器哪个存储器存储了由所述主机指定的逻辑盘区之前，所述每个控制器检查所述每个控制器的所述高速缓冲存储器是否存储了关于对该询问的答复的信息，并且当该信息被存储在该高速缓冲存储器中时请求由该信息指示的存储器以得到对指定的逻辑盘区的访问。

9. 一种存储系统，其特征在于，包括：

多个节点，该节点包括存储器、控制器和管理服务器；以及
连接所述节点和主机的接口，

其中：

每个所述存储器包括物理卷，所述每个存储器在所述物理卷的物理盘区中映射包括在逻辑卷中的逻辑盘区，并在该物理卷中存储所述逻辑盘区；

每个所述控制器为所述主机提供所述逻辑卷并接收来自该主机的访问，每个所述控制器包括指示与构成所述逻辑卷的逻辑盘区相对应的管理服务器的第一映射表；

每个所述管理服务器进行位置管理服务以答复来自所述控制器的任何一个的询问，所述询问涉及存储了在该逻辑卷中包括的逻辑盘区的存储器的位置，每个所述管理服务器包括指示哪个节点包括了存储与所述管理服务器对应的逻辑盘区的存储器的第二映射表；

每个所述存储器包括指示该存储器中包括的物理卷的哪个物理盘区存储了逻辑盘区的第三映射表；

当所述控制器的每一个从所述主机接收了指定所述逻辑卷中的逻辑盘

区作为访问目的地的访问请求时，所述控制器的所述每一个参考所述第一映射表以指定与所述访问目的地的逻辑盘区对应的管理服务器，并询问所述指定的管理服务器哪个存储器存储了该访问目的地的逻辑盘区；

当所述管理服务器的每一个接收到来自所述控制器中的任何一个的所述询问时，该管理服务器的所述每一个参考所述第二映射表以指定被询问的存储器，并且作为对所述询问的答复将指示该指定的存储器的信息通知给作为询问源的控制器；

每个所述控制器请求由作为对所述询问的答复的所述信息指示的存储器，以得到对由所述主机指定的逻辑盘区的访问；以及

当所述控制器中的任何一个请求每个所述存储器以得到对所述指定的逻辑盘区的访问时，所述每个存储器参考所述第三映射表以在所述物理卷中指定与所述指定的逻辑盘区对应的物理盘区，得到对该指定的物理盘区的访问，并答复作为访问请求源的所述控制器。

10. 根据权利要求 9 所述的存储系统，其特征在于：

每个所述控制器包括高速缓冲存储器，其存储了有关所述询问以及该询问的答复的信息；以及

在每个所述控制器询问所述管理服务器哪个存储器存储了由所述主机指定的逻辑盘区之前，所述每个控制器检查所述每个控制器的所述高速缓冲存储器是否存储了关于对该询问的答复的信息，并且当该信息被存储在该高速缓冲存储器中时请求由该信息指定的存储器以得到对所述指定的逻辑盘区的访问。

11. 根据权利要求 10 所述的存储系统，其特征在于：每个所述节点包括高速缓冲存储器无效化单元，其检查存储了由所述每个节点的管理服务器中包括的所述第二映射表指示的逻辑盘区的存储器与存储了由在所述每个节点的管理服务器中包括的所述高速缓冲存储器中存储的有关答复的信息所指示的逻辑盘区的存储器是否彼此一致，并且当两个存储器彼此不一致时使得存储在该高速缓冲存储器中的所述信息无效。

12. 一种在存储系统中处理来自主机的访问请求的方法，所述存储系

统包括：多个节点，该节点包括至少一个具有物理卷的节点，所述节点的至少两个包括利用在所述至少一个节点中包括的物理卷来为所述主机提供包括多个逻辑盘区的逻辑卷并接收来自该主机的访问的控制器，所述至少一个节点包括在所述物理卷的物理盘区中映射所述逻辑卷的逻辑盘区并在该物理卷中存储该逻辑盘区的存储器，所述至少一个节点或所述节点中的至少另一个包括进行位置管理服务的管理服务器以答复来自所述控制器的任何一个的关于存储了所述逻辑卷的逻辑盘区的存储器的位置的询问，所述方法的特征在于，包括以下步骤：

当所述控制器中的一个从所述主机接收到指定所述逻辑卷中的逻辑盘区的访问请求时，指定与访问目的地的逻辑盘区对应的管理服务器，所述指定管理服务器包括参考在所述控制器的所述一个中包括的第一映射表，该第一映射表指示与所述逻辑卷的逻辑盘区对应的管理服务器；

通过所述控制器中的所述一个询问所述指定的管理服务器哪个存储器存储了所述访问目的地的逻辑盘区；

指定所述被询问的存储器，该指定被询问的存储器包括：参考在所述被询问的管理服务器中包括的第二映射表，并且作为对所述询问的答复、通过所述被询问的管理服务器将指示所述指定的存储器的信息通知给作为询问源的控制器，所述第二映射表指示哪个节点包括了存储与所述管理服务器对应的逻辑盘区的存储器；

通过所述询问源的所述控制器，请求由作为对所述询问的答复的所述信息所指示的存储器以得到对由所述主机指定的逻辑盘区的访问；

在所述被请求的存储器中包括的物理卷中指定与所述指定的逻辑盘区关联的物理盘区，该指定物理盘区包括：参考在所述被请求的存储器中包括的第三映射表，该第三映射表指示在所述存储器中包括的物理卷的哪个物理盘区存储了所述逻辑盘区；

得到对所述指定的物理盘区的访问，该得到访问包括：从所述被请求的存储器答复所述访问请求源的所述控制器；以及

将所述存储器的答复从所述访问请求源的所述控制器返回给所述主

机。

用于处理来自主机的访问请求的存储系统及方法

技术领域

本发明涉及一种由包括具有物理卷的节点的多个节点构成的存储系统。更具体地，本发明涉及一种用于处理来自于主机的访问请求的存储系统及方法。

背景技术

最近已经开发了包括相互连接的多个存储装置（例如，盘存储装置）的存储系统。存储集群（存储集群系统）已经被公知为这种类型的存储系统。

存储集群的特征如下。存在于多个存储装置（在下文中，其每个被称为节点）中的存储区域（在下文中，其每个被称为物理卷）在节点上被绑定在一起以在节点上为主机提供至少一个横向的、虚拟存储区域（在下文中被称为逻辑卷）。这里假定卷被称为对于每个块将被访问的块卷（block volume）。

存储集群的节点是一直在进行工作的存储装置。换言之，一般地，工作中的节点有意地参与存储集群/从存储集群分离，或者由于它们的故障，它们从那里突然分离。在存储集群中，不常发生由于例如日常的通电/断电而导致的节点的参与/分离。

在存储集群中，数据（数据段）被频繁迁移到/适当地复制到存储数据的节点的物理卷或另一个节点的物理卷（自动地或手动地）。数据的迁移/复制在确保从主机对逻辑卷的访问的透明度时，力求使得数据冗余并提高访问性能和可靠性。在形成存储集群的节点中的任何一个突然停止、盘故障等的情况下，需要数据冗余。

简单地通过经由路径在存储集群中访问节点，主机能获得在所有的多个节点上对整个逻辑卷的访问。这里假定故障发生在主机将路径延伸到的节点上且该节点（故障的节点）停止。在这种情况下，还有，如果使得在故障节点中存储的数据对在存储集群中的另一个节点（活动节点，live node）冗余，则主机能将路径再次延伸到该活动节点并继续对逻辑卷的访问。

为了配置上述的存储集群，需要系统来管理在逻辑卷上的数据位于哪个节点的哪个物理卷上的哪个位置。通常，存储集群的节点功能性地包括控制器和存储器（storage）。控制器管理到主机的路径并接收访问。即使主机延伸路径至任一控制器，其也能获得对相同逻辑卷的访问。每个存储器都具有用于存储数据的物理卷。

卷（块卷）包括多个盘区（extent）。每个盘区由一组固定长度的块形成。每个节点管理一个物理卷。另一方面，每个节点的控制器管理通过映射表相互关联的物理卷的物理盘区和逻辑卷的逻辑盘区。因此配置逻辑卷（逻辑卷空间）。在例如日本专利申请公开 9-62452（在下文中称为现有技术文件）中公开了利用映射表的物理盘区和逻辑盘区之间的关联。

以上的映射表被叫做逻辑盘区到物理盘区映射表（LPMT）。LPMT 被用于表示逻辑卷的逻辑盘区由哪个节点（存储器）的物理卷的哪个物理盘区形成。在存储集群中的所有节点的每个控制器都保存相同内容的 LPMT。每个节点利用 LPMT 管理整个逻辑卷。上述意图在于，在将主机和控制器之间设置的路径设置为该主机和另一个控制器之间的路径之后能够立即对相同的逻辑卷进行访问。

如上所述，在现有技术的存储集群（存储系统）中，每个节点都必须通过 LPMT 来管理整个逻辑卷。然而，取决于物理卷的数目和形成物理卷的盘区的数目，LPMT 的大小变得非常大。因此，消耗了控制器上的大量内存。

例如在存储集群中的节点（以下称为第一节点）的第一物理盘区中存储的第一逻辑盘区被迁移到例如另一个节点（以下称为第二节点）的第二

物理盘区。为了减少例如第一节点的存储负担来进行这种迁移。如在上述现有技术文件中所述的，为了提高对节点的存储器的访问性能，在节点的物理盘区中存储的第一逻辑盘区有时候被迁移到该节点的另一个物理盘区中。这种迁移能提高对存储集群的访问性能及其可靠性。

根据上述的存储集群，为了提高对存储集群的访问性能及其可靠性，在物理盘区中存储的逻辑盘区被迁移/复制到在相同存储器的另一个物理盘区或另一个存储器的物理盘区。在这种情况下，存储集群的控制器必须彼此同步以立刻（同时地）更新它们自己的LPMT。同时的更新使LPMT的管理变得复杂并且不利地影响了对存储集群的访问性能。

发明内容

本发明的目的是通过减少用于更新由存储系统的每个节点管理的信息所需的处理量来提高存储系统的访问性能。

根据本发明的实施例，提供了一种存储系统。所述存储系统包括：多个节点，该节点的至少一个包括具有物理卷的存储器，该存储器在所述物理卷的物理盘区中映射组成逻辑卷的逻辑盘区并在所述物理卷中存储该逻辑盘区，以及连接所述节点和主机的接口，其中所述节点的至少两个包括为所述主机提供所述逻辑卷并接收来自该主机的访问的控制器，所述节点的所述至少一个或所述节点的至少另一个包括：管理服务器，其执行位置管理服务以答复来自任何控制器的关于存储逻辑卷的逻辑盘区的存储器的位置的询问，每个所述控制器包括指示与逻辑卷的逻辑盘区对应的管理服务的第一映射表，所述管理服务器包括指示哪个节点包括存储与所述管理服务器对应的逻辑盘区的存储器的第二映射表，所述存储器包括指示包括在存储器中的物理卷的哪个物理盘区存储所述逻辑盘区的第三映射表，当所述控制器的每一个从所述主机接收了在逻辑卷中指定逻辑盘区作为访问目的地的访问请求时，所述控制器的所述每一个参考第一映射表以指定与所述访问目的地的逻辑盘区对应的管理服务器，并询问所述指定的管理服务器哪个存储器存储了所述逻辑盘区，当所述管理服务器接收到来自所

述控制器中的一个的询问时, 该管理服务器参考所述第二映射表以指定被询问的存储器, 并通知作为询问源的所述控制器中的所述一个指示该存储器的信息作为对所述询问的答复, 每个所述控制器请求由作为对所述询问的答复的所述信息指示的存储器, 以得到对由所述主机指定的逻辑盘区的访问, 且当所述控制器中的一个请求所述存储器以得到对逻辑盘区的访问时, 所述存储器参考所述第三映射表以在物理卷中指定与所述指定的逻辑盘区对应的物理盘区, 得到对该指定的物理盘区的访问, 并答复作为访问请求源的所述控制器中的所述一个。

附图说明

结合到说明书并构成说明书一部分的附图, 与以上给出的一般描述以及下面给出的实施例的详细描述一起, 阐明了本发明的实施例, 用于解释本发明的原理。

图 1 是根据本发明第一实施例的具有存储集群结构的存储系统的示意性框图;

图 2 是图 1 所示的存储系统中的控制器的框图;

图 3 是图 1 所示的存储系统中的存储器的框图;

图 4 是根据本发明第一实施例的存储系统中当主机发出访问请求时进行的处理的流程的流程图;

图 5 是根据本发明第一实施例的存储系统中当主机发出访问请求时的信息流的示图;

图 6 是图 5 所示的信息流的变型的示图;

图 7 是与图 1 的系统对应的系统的示图, 其被如此构成以使得逻辑卷 ID(LVol)是 1 的逻辑卷包括从逻辑盘区 ID(LEID)1 到逻辑盘区 ID(LEID)4 的四个逻辑盘区;

图 8 是图 7 的系统构成中的逻辑盘区到管理服务器映射表(LMMT)的数据结构的示图;

图 9 是图 7 的系统构成中的逻辑盘区到存储器映射表(LSMT)的数

据结构的示图;

图 10 是图 7 的系统构成中的逻辑盘区到物理盘区映射表 (LPMT) 的数据结构的示图;

图 11 是示图,其说明了在根据本发明第一实施例的存储系统中将形成逻辑盘区的物理盘区的数据迁移到另一个存储器中的物理盘区的操作;

图 12 是示图,其说明了在根据本发明第一实施例的存储系统中将形成逻辑盘区的物理盘区的数据复制到另一个存储器中的物理盘区的操作;

图 13 是示图,其说明了在根据本发明第一实施例的存储系统中将关于逻辑盘区的位置管理服务迁移到另一个节点的操作;

图 14 是示图,其说明了在根据本发明第一实施例的存储系统中将关于逻辑盘区的位置管理服务复制到另一个节点的操作;

图 15 是根据本发明第二实施例的具有存储集群结构的存储系统的框图;

图 16 是根据本发明第三实施例的具有存储集群结构的存储系统的框图;

图 17 是在根据本发明第三实施例的存储系统中当主机发出访问请求时进行的处理的流程的流程图;

图 18 是根据本发明第四实施例的具有存储集群结构的存储系统的框图;

图 19 是在根据本发明第四实施例的存储系统中当主机发出访问请求时进行的处理的流程的流程图;

图 20 是在根据本发明第五实施例的存储系统中当主机发出访问请求时进行的处理的流程的流程图。

具体实施方式

下面将参照附图描述本发明的实施例。

[第一实施例]

图 1 是根据本发明第一实施例的具有存储集群结构的存储系统的示意

性框图。图 1 所示的存储系统（以下称为存储集群）包括多个存储装置（以下称为节点），例如，三个节点 1-1、1-2 和 1-3（#1，#2 以及#3）。节点 1-1、1-2 和 1-3 经由诸如光纤通道和因特网小型计算机系统接口（iSCSI）的接口 2 被连接到主机（主计算机）。主机使用集群存储器。节点 1-1、1-2 和 1-3 分别包括控制器 11-1、11-2 和 11-3，管理服务器 12-1、12-2 和 12-3 以及存储器 13-1、13-2 和 13-3。

控制器 11- i ($i=1, 2, 3$) 为主机提供逻辑卷。这里假定通过绑定存储器 13-1、13-2 和 13-3 的物理卷（后面描述）为主机提供 ID（逻辑卷 ID）是 x （例如“1”）的逻辑卷。以下，逻辑卷 ID 将被表示为 LVol，且逻辑卷 ID 是 LVol 的逻辑卷将被表示为逻辑卷 LVol。进一步，逻辑卷 ID 是 x （LVol= x ）的逻辑卷被表示为逻辑卷 LVol x ，逻辑卷“LVol=LVol x ”或逻辑卷“LVol= x ”。

控制器 11- i 从主机接收访问请求。访问请求包括指示在逻辑卷中将被访问的逻辑块的地址（逻辑块地址）LBA。参照后面描述的 LMMT 112a，控制器 11- i 询问管理服务器 12- j ($j=1, 2, 3$)，哪个节点包括存储了包括被请求的逻辑块的逻辑盘区的存储器。响应于该询问，管理服务器 12- j 通知控制器 11- i 所述节点的存储器，且控制器 11- i 获得对所述存储器的访问。

图 2 是控制器 11- i 的框图。控制器 11- i 包括目标接口 111、服务器指定单元 112 以及输入/输出管理器（IO 管理器）113。

目标接口 111 接收来自主机的访问请求并指定被请求的逻辑卷的逻辑块地址 LBA（由所述逻辑块地址 LBA 指示的逻辑块）位于其中的逻辑盘区。识别逻辑盘区的逻辑盘区 ID（LEID）由如下计算来指定：

$$\text{LEID} = \text{LBA} / \text{盘区大小}$$

通过服务器指定单元 112，目标接口 111 询问管理服务器 12- j 哪个存储器存储了在指定的逻辑卷中的逻辑盘区。此询问包括用于识别逻辑卷/逻辑盘区的逻辑卷 ID（LVol）/逻辑盘区 ID（LEID）。目标接口 111 获得响应于对管理服务器 12- i 的询问的经由服务器指定单元 112 通知的对存储器的访问，并答复主机。

服务器指定单元 112 指定传送来自目标接口 111 的询问的管理服务器 12-j。以下，由包括在来自目标接口 111 的询问中的逻辑卷 ID 和逻辑盘区 ID 所指示的逻辑盘区被称为目标逻辑盘区。管理服务器 12-j 被指定为用于管理存储了目标逻辑盘区的存储器的位置的服务器。

服务器指定单元 112 具有逻辑盘区到管理服务器映射表 (LMMT) 112a。LMMT 112a 是用于指定管理目标逻辑盘区的位置的管理服务器的映射表 (第一映射表)。LMMT 112a 保持有用于提供对存储逻辑盘区的存储器的位置负责 (管理该位置) 的管理服务器的节点的 ID (MS)。

服务器指定单元 112 接收来自控制器 11-1 的询问并将其传输给指定的管理服务器。单元 112 接收来自管理服务器 12-j 的对所述询问的响应并将其发送给目标接口 111。该响应包括在其中存储了目标逻辑盘区的存储器的节点 ID (PVol)。后面将描述存储器的节点 ID (PVol)。单元 112 还具有高速缓冲存储器 112b。后面在第五实施例中还将描述高速缓冲存储器 112b。IO 管理器 113 提供在控制器 11-i 和存储器 13-j 之间的接口。

再次参照图 1，管理服务器 12-i (i=1, 2, 3) 检查存储了由控制器 11-j (j=1, 2, 3) 询问的逻辑盘区的存储器的节点 ID，并通知控制器 11-j 该节点 ID。管理服务器 12-i 具有逻辑盘区到存储器映射表 (LSMT) 120-i。换言之，管理服务器 12-1、12-2 和 12-3 分别具有 LSMT 120-1、120-2 和 120-3。LSMT 120-i (i=1, 2, 3) 是这样的映射表 (第二映射表)，其用于管理哪个节点 ID 被分配给包括存储了管理服务器 12-i 自身负责的逻辑盘区的存储器 (物理卷) 的节点。

存储器 13-1、13-2 和 13-3 分别具有物理卷 130-1、130-2 和 130-3。物理卷 130-1 的 ID (PVol) 是 1，物理卷 130-2 的 ID (PVol) 是 2，且物理卷 130-3 的 ID (PVol) 是 3。在第一实施例中，物理卷 130-i (i=1, 2, 3) 的 ID (PVol) 是 “i”。存储器 13-i 在物理盘区上映射逻辑盘区并将其存储在物理卷 130-i 中。存储器 13-i 接收来自控制器 11-i 的输入/输出请求 (IO 请求)。IO 请求包括将被访问的逻辑卷中的逻辑盘区的 ID (LEID)。存储器 13-i 检查对应于所述逻辑盘区的物理盘区。存储器 13-i 进行对物理盘

区的 IO 操作并答复控制器 11-i。

图 3 示出了存储器 13-i ($i=1, 2, 3$) 的结构。存储器 13-i 包括 IO 驱动器 131、逻辑盘区到物理盘区管理单元 (LP 管理单元) 132 以及盘 133。IO 驱动器 131 提供在存储器 13-i 和控制器 11-i 之间的接口。

LP 管理单元 132 管理在逻辑盘区与物理盘区之间的对应关系。单元 132 保持有逻辑盘区到物理盘区映射表 (LPMT) 132a。LPMT 132a 是这样的映射表 (第三映射表), 其用于指示逻辑盘区对应于并被存储在包括 LPMT 132a 的存储器 13-i 中的物理卷 130-i 的哪个物理盘区。应当注意, 第一实施例的 LPMT 132a 与现有技术的 LPMT 不同并且没有用于整个逻辑卷的映射信息。还应当注意, LPMT 132a 的内容对于各个存储器 13-i 有所不同。

盘 133 具有物理卷 130-i ($PVol=i$)。以物理盘区为单位管理物理卷 130-i。对物理卷 130-i 进行访问的单位是块, 且物理盘区包括多个块。

图 1 所示的存储集群具有控制该集群的结构的管理单元 3。当创建卷时, 构成管理单元 3 确定作为逻辑卷/逻辑盘区的存储目的地的物理卷/物理盘区。单元 3 还确定用于迁移/复制盘区和位置管理服务 (用于冗余的复制) 的迁移源/迁移目的地。进一步, 单元 3 在要求时确定盘区被分配到的目的地。在例如节点 1-1 中设置单元 3。然而, 也可以在节点 1-2 或 1-3 中或独立于节点 1-1 至 1-3 设置单元 3。

在第一实施例中, 构成管理单元 3 通过下面的流程来创建逻辑卷。

(a) 构成管理单元 3 确定哪个节点 ID 被分配给包括了存储逻辑卷的逻辑盘区的存储器的节点。单元 3 把确定的节点 (所述确定的节点 ID 的节点) 的存储器的物理盘区分配给逻辑盘区。单元 3 生成 LPMT 132a 并将它们放置到存储器 13-1 至 13-3 中, LPMT 132a 指示在存储器的物理盘区与该物理盘区被分配到的逻辑盘区之间的对应关系。

(b) 从存储了逻辑盘区被分配到的物理盘区的存储器 (物理卷), 构成管理单元 3 确定节点 1-1 至 1-3 中的哪个具有负责用于管理逻辑盘区的位置的服务 (位置管理服务) 的管理服务器。单元 3 生成 LSMT 120-1 至

120-3 并将它们放置到它们各自的管理服务器 12-1 至 12-3 中, LSMT 120-1 至 120-3 指示在逻辑盘区和负责所述管理服务器的物理盘区之间的对应关系。在后面描述的第二实施例中, 不需要在存储集群中的所有管理服务器 12-1 至 12-3 中设置 LSMT。

(c) 构成管理单元 3 生成 LMMT 112a 并将其分配给所有控制器 11-1 至 11-3。如上所述, LMMT 112a 指示哪个管理服务器被询问了存储了逻辑卷的每个逻辑盘区被分配到的物理盘区的存储器 (物理卷)。

因而创建逻辑卷。因此主机能够得到对所创建的逻辑卷的访问。通过引导从主机到节点 1-1 至 1-3 中的任意一个 1-i 的控制器 11-i 的路径, 进行从主机对逻辑卷的访问。

下面将参照图 4 和 5 描述当路径被从主机延伸到节点 1-1 ($i=1$) 之后主机发出访问请求时进行的处理的流程。图 4 是处理的流程的流程图, 且图 5 是信息流的示图。图 5 中的箭头 A1 至 A8 对应于图 4 的流程图中的它们各自的步骤 S1 至 S8。

首先假定主机发出访问请求 (IO 请求) 到包括在节点 1-1 的控制器 11-1 中的目标接口 111。还假定访问请求指示用于对逻辑卷 “LVol=LVolx” 的逻辑块地址 LBA (=LBAX) 进行访问的请求。在第一实施例中, LVol (=LVolx) 是 1。这里假定逻辑卷 “LVol=LVolx” 的名称 (逻辑卷名称) 与该逻辑卷的 ID (LVol=LVolx) 相符。

目标接口 111 接收来自主机的访问请求 (步骤 S0)。然后, 目标接口 111 通过来自主机的访问请求, 从将被访问的逻辑卷 “LVol=LVolx” 上的逻辑块地址 LBA (=LBAX) 获得 LBAX 位于其中的逻辑盘区的 ID (=LEIDx) (步骤 S1)。在第一实施例中, 目标接口 111 通过下面的等式获得逻辑盘区 ID (=LEIDx):

$$LEIDx = LBAX / \text{盘区大小}$$

这里, 逻辑盘区 ID 在逻辑卷 “LVol=Lvolx (=1)” 上是 LEIDx 的逻辑盘区被表示为逻辑盘区 “LVolx, LEIDx”。逻辑盘区 “LVolx, LEIDx” 有时候被简单地表示为逻辑盘区 LEIDx。

目标接口 111 把逻辑卷名称 (逻辑卷 ID) $LVol (=LVolx=1)$ 和所获得的逻辑盘区 ID ($LEIDx$) 传送给服务器指定单元 112 (步骤 S2)。因而, 目标接口 111 询问哪个节点包括存储了逻辑盘区 “ $LVolx, LEIDx$ ” 的存储器 (即, 逻辑盘区 “ $LVolx, LEIDx$ ” 的位置)。

答复来自目标接口 111 的询问, 服务器指定单元 112 指定提供了管理逻辑盘区 “ $LVolx, LEIDx$ ” 的位置的管理服务器的节点的 ID (步骤 S3)。该 ID 也可以叫做管理服务器的节点 ID。所述管理服务器的节点 ID 被表示为 MS (或 MID), 且管理逻辑盘区 “ $LVolx, LEIDx$ ” 的位置的管理服务器的节点 ID 被表示为 $MIDx$ ($MS=MIDx$)。在这种情况下, $MIDx$ ($MS=MIDx$) 在步骤 S3 被指定为管理服务器的节点 ID。

参考 LMMT 112a 指定节点 ID ($MS=MIDx$)。这是因为 LMMT 112a 保持了与 $LVolx$ 和 $LEIDx$ 关联的节点 ID ($MS=MIDx$), 这从后面描述的数据结构看是显然的 (参见图 8)。节点 1-1 包括指定节点 ID ($MS=MIDx$) 的服务器指定单元 112。假定在第一实施例中指定了节点 1-1 (即, 提供管理服务器 12-1 的节点 1-1) 的 ID ($MS=MIDx=1$)。

在第一实施例中, 节点 1-1、1-2 和 1-3 的存储器 13-1、13-2 和 13-3 分别具有物理卷 130-1、130-2 和 130-3。换言之, 节点 1-1、1-2 和 1-3 的存储器 13-1、13-2 和 13-3 的每个具有一个物理卷。因而, 提供管理服务器 12-1、12-2 和 12-3 的节点 1-1、1-2 和 1-3 的 ID (MS) 是 1、2 和 3, 而包括在节点 1-1、1-2 和 1-3 的存储器 13-1、13-2 和 13-3 中的物理卷 130-1、130-2 和 130-3 的 ID (PVol) 是 1、2 和 3。因此 ID (MS) 和 ID (PVol) 彼此一致。

服务器指定单元 112 将询问从目标接口 111 传送给被指定的节点 1-1 的管理服务器 12-1 (步骤 S4)。换言之, 单元 112 询问管理服务器 12-1 逻辑盘区 (目标逻辑盘区) “ $LVolx, LEIDx$ ” 的位置。逻辑盘区 “ $LVolx, LEIDx$ ” 的位置意味着存储与逻辑盘区 “ $LVolx, LEIDx$ ” 对应的物理盘区的存储器所属的节点, 且该节点的 ID 被询问。存储器所属的节点的 ID 还可以被叫做存储器的节点 ID。存储器的节点 ID 被表示为 PVol (或 SID),

且存储与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器的节点 ID 被表示为 SIDx (PVol=SIDx)。

基于 LSMT 120-1, 管理服务器 12-1 指定哪个节点 ID “SIDx” 被分配给包括存储了与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器的节点(步骤 S5)。在此参考与逻辑盘区“LVolx, LEIDx”关联的 LSMT 120-1 中的条目(entry)的 PVol (=SIDx), 来指定以上节点 ID “SIDx”。在步骤 S5, 管理服务器 12-1 通知所述询问的服务器指定单元 112 所述指定的节点 ID “SIDx”。单元 112 将从管理复位服务器 12-1 接收的节点 ID “SIDx” 传送给目标接口 111, 作为对来自目标接口 111 的询问的响应(步骤 S6)。这里假定存储器 13-1 的节点 ID “SIDx” 被传送给目标接口 111。

目标接口 111 通过 IO 管理器 113 请求由传送的节点 ID “SIDx” 指示的节点 1-1 的存储器 13-1 以得到对与逻辑盘区“LVolx, LEIDx”对应的物理盘区的访问(步骤 S7)。如下, 存储器 13-1 的 IO 驱动器 131 在存储器 13-1 (物理卷 130-1) 中指定与所述请求的逻辑盘区“LVolx, LEIDx”相关联的物理盘区(步骤 S8)。首先, IO 驱动器 131 参考 LPMT 132a 并指定与逻辑盘区“LVolx, LEIDx”相关联的物理盘区的 ID (PEID=PEIDx), 这从后面描述的 LPMT 132a 的数据结构来看是显然的(参见图 10)。IO 驱动器 131 通过所述指定的 ID (=PEIDx) 指定目标物理盘区。换言之, IO 驱动器 131 基于 LPMT 132a 指定与逻辑盘区“LVolx, LEIDx”相关联的物理盘区。

物理盘区的 ID (物理盘区 ID) 被表示为 PEID, 且与逻辑盘区“LVolx, LEIDx”相关联的物理盘区的 ID 被表示为 PEIDx。此外, 物理盘区 ID 是 PEIDx 的物理盘区被表示为物理盘区 PEIDx。在这种情况下, 在步骤 S8 IO 驱动器 131 指定与逻辑盘区“LVolx, LEIDx”相关联的物理盘区 PEIDx。

在步骤 S8, IO 驱动器 131 还得到对指定的物理盘区 PEIDx 的访问, 并返回对控制器 11-1 的 IO 管理器 113 的访问的结果作为响应。目标接口 111 接收来自 IO 管理器 113 的响应并将其返回给主机(步骤 S9)。

上面的描述是基于下面的假设作出的。接收来自主机的访问请求的目

标接口、管理哪个存储器存储目标逻辑盘区的管理服务器以及其物理盘区存储了所述目标逻辑盘区的存储器,都属于相同的节点 1-1(第一种情况)。然而,这些目标接口、管理服务器以及存储器可以属于不同的节点(第二种情况)。图 6 示出了在第二种情况下的信息流。在图 6 中,箭头 B1 至 B8 对应于图 4 所示流程图中的步骤 S1 至 S8。

图 7 示出了与图 1 的系统对应的系统,其被如此构成以使得逻辑卷“LVol=1”包括四个逻辑盘区 LEID(逻辑盘区 ID) 1 至 LEID(逻辑盘区 ID) 4。图 8 示出了在图 7 的系统构成中的 LMMT 112a 的数据结构。图 9 示出了在图 7 的系统构成中的 LSMT 120-i(例如节点 1-1 的 LSMT 120-1)的数据结构。图 10 示出了在图 7 的系统构成中的例如节点 1-1 的 LPMT 132a 的数据结构。

LMMT 112a 具有数目等于构成逻辑卷“LVol=1”的逻辑盘区的数目的条目。每个条目相互关联地保持逻辑卷的 ID(LVol)、逻辑盘区的 ID(LEID)以及提供管理服务器 12-j 的节点的 ID(MS)。LEID 指示在由与 LEID 关联的 LVol 表示的逻辑卷中包括的逻辑盘区的 ID。MS 指示提供管理服务器 12-j 的节点的 ID。管理服务器 12-j 管理哪个存储器(物理卷)存储由与 MS 关联的 LVol 和 LEID 表示的逻辑盘区“LVol, LEID”(或者其管理逻辑盘区的位置)。图 8 中示出了逻辑卷“LVol=1”的逻辑盘区是四个逻辑盘区 LEID 1 至 LEID 4。

LSMT 120-i(120-1)具有数目等于由管理服务器 12-i(12-1)管理的逻辑盘区的数目的条目。LSMT 120-i(120-1)的每个条目相互关联地保持逻辑卷的 ID(LVol)、逻辑盘区的 ID(LEID)以及存储器(物理卷)的节点 ID(PVol)。LEID 指示在由与 LEID 关联的 LVol 表示的逻辑卷中包括的逻辑盘区的 ID。PVol 指示存储由与 PVol 关联的 LEID 和 LVol 表示的逻辑盘区的存储器(物理卷)的节点 ID(物理卷号)。

LPMT 132a 具有数目等于构成物理卷 130-i(130-1)的物理盘区的数目的条目。LPMT 132a 的每个条目保持了在物理卷 130-i 中包括的物理盘区的 ID(PEID)、物理卷的 ID(LVol)、逻辑盘区的 ID(LEID)以及

提供位置管理服务（管理服务器 12-j）的节点的 ID（MS）。LVol 指示包括逻辑盘区的逻辑卷的 ID，其中由与 LVol 关联的 PEID 表示的物理盘区被分配给该逻辑盘区。MS 指示提供位置管理服务（管理服务器 12-j）的节点的 ID。位置管理服务（管理服务器 12-j）管理哪个存储器（物理卷）存储逻辑盘区（或者其管理逻辑盘区的位置），其中由与 MS 关联的 PEID 表示的物理盘区被分配给该逻辑盘区。在 LPMT 132a 的每个条目中保持的 MS 被用作位置管理服务的反向索引。图 10 中示出了物理卷 130-i（130-1）的物理盘区是三个物理盘区 PEID 1 至 PEID 3。

在图 7 中，从在控制器 11-1 中包括的 LMMT 112a 的条目向管理服务器 12-1 至 12-3 的 LSMT 120-1 至 120-3 延伸的、长短交替的虚线箭头，表示通过 LMMT 112a 的条目的信息被参考的 LSMT 120-1 至 120-3 的条目。相似地，从 LSMT 120-1 至 120-3 的条目向存储器 13-1 至 13-3 的 LPMT 132a 延伸的、长短交替的虚线箭头，表示通过 LSMT 120-1 至 120-3 的条目的信息被参考的 LPMT 132a 的条目。相似地，从存储器 13-1 至 13-3 的 LPMT 132a 的条目向存储器 13-1 至 13-3 的物理卷 130-1 至 130-3 延伸的、长短交替的虚线箭头，表示将被基于 LPMT 132a 的信息进行访问的物理盘区。

将参照图 7 描述，在路径从主机延伸到节点 1-1 之后，从主机对于将由逻辑卷“LVol=1”/逻辑盘区“LEID=1”来指定的块进行访问的流程。逻辑卷“LVol=1”指示逻辑卷 ID（LVol）是 1 的逻辑卷，且逻辑盘区“LEID=1”指示逻辑盘区 ID（LEID）是 1 的逻辑盘区。逻辑卷“LVol=1”/逻辑盘区“LEID=1”指示在逻辑卷“LVol=1”上的逻辑盘区“LEID=1”。在这种情况下，逻辑盘区“LEID=1”可以被表达为逻辑盘区“LVol=1, LEID=1”。

首先假定主机向节点 1-1 发出对于在逻辑卷“LVol=1”中的块进行访问的请求，如图 7 的箭头 B1 所示。节点 1-1 的控制器 11-1（在控制器 11-1 中包括的目标接口 111）确定将被访问的块被存储在 LEID 是 1 的逻辑盘区（逻辑盘区“LEID=1”）中。

控制器 11-1(目标接口 111)询问控制器 11-1 中的服务器指定单元 112 哪个节点包括了存储逻辑盘区“LEID=1”的存储器(物理卷)(逻辑盘区“LEID=1”的位置)。服务器指定单元 112 参考单元 112 的 LMMT 112a 中保持“LEID=1”的条目(图 8 中的第一行的条目)。因而,单元 112 识别出与逻辑卷“LVol=1”/逻辑盘区“LEID=1”对应的管理服务器存在于节点 1-1 中。控制器 11-1(控制器 11-1 的服务器指定单元 112)询问节点 1-1 中的管理服务器 12-1 的逻辑卷“LVol=1”/逻辑盘区“LEID=1”的位置,如图 7 中的箭头 B2 所示。

管理服务器 12-1 参考管理服务器 12-1 的 LSMT 120-1 中的保持一对“LVol=1”且“LEID=1”的条目(图 9 中的第一行的条目)。因而,管理服务器 12-1 识别出逻辑卷“LVol=1”/物理盘区“LEID=1”的实体被存储在节点 1-1 的存储器(物理卷)中。管理服务器 12-1 向控制器 11-1(控制器 11-1 的服务器指定单元 112)发送节点 11-1 的 ID(SID_x=PVol=1),如图 7 中的箭头 B3 所示(步骤 B3)。然后,控制器 11-1 的服务器指定单元 112 通知控制器 11-1 的目标接口 111 节点 1-1 的 ID(SID_x=PVol=1)。

控制器 11-1(控制器 11-1 的目标接口 111)请求由节点 ID 指定的存储器(或节点 1-1 的存储器 13-1)以得到(经由 IO 管理器 113)对逻辑卷“LVol=1”/逻辑盘区“LEID=1”的访问,如图 7 中的箭头 B4 所示。节点 1-1 的存储器 13-1 参考在存储器 13-1 的 LPMT 132a 中保持“LVol=1”/“LEID=1”的条目(图 10 中所示第一行的条目)。因而,存储器 13-1 识别出与卷“LVol=1”/逻辑盘区“LEID=1”对应的物理盘区的 ID(PEID)为 1。存储器 13-1 对 PEID 是 1 的物理盘区(物理盘区“PEID=1”)进行访问并答复控制器 11-1,如图 7 中的箭头 B5 所示。

在第一实施例中,节点 1-1 至 1-3 的控制器 11-1 至 11-3 的 LMMT 112a 具有相同的内容。换言之,节点 1-1 至 1-3(控制器 11-1 至 11-3)共享 LMMT 112a。因而主机能将路径延伸到构成存储集群的节点 1-1 至 1-3 的任意一个(节点 1-1 至 1-3 的任意一个的控制器)。如果主机延伸路径到的控制器异常停止或如果该控制器的负荷增加,则主机能将路径延伸到另一个控

制器。

节点 1-1 至 1-3 共享的信息（共享信息），诸如 LMMT 112a，被存储在节点 1-1 至 1-3 的控制器 11-1 至 11-3 的非易失性存储器的存储区域、节点 1-1 至 1-3 的存储器 13-1 至 13-3 的存储区域、或者存储器 13-1 至 13-3 的任意一个的存储区域。参考这些存储区域，控制器 11-1 至 11-3 管理所述共享信息。为了所述管理，控制器 11-1 至 11-3 共享所述存储区域的地址。所述共享信息包括逻辑卷构成信息以及 LMMT 112a。逻辑卷构成信息指示了图 1 所示的存储集群为主机提供了什么样的逻辑卷。与现有技术相比，在其中节点 1-1 至 1-3 共享信息的上述构成能减少应由节点 1-1 至 1-3 管理的信息量。

将参考图 11 描述将存储器中的形成逻辑盘区的物理盘区的数据迁移到另一个存储器中的物理盘区的操作。首先，构成管理单元 3 确定数据迁移的源和目的地。这里假定当图 1 所示的存储集群被设定在图 7 所示的状态时，在节点 1-1 的存储器 13-1 中与逻辑盘区“LEID=4”相关联的物理卷“PVol=1”（物理卷 130-1）/物理盘区“PEID=2”被确定为数据的迁移源。此外，假定在节点 1-2 的存储器 13-2 中的物理卷“PVol=2”（物理卷 130-2）/物理盘区“PEID=1”被确定为数据迁移的目的地。物理卷“PVol=p”/物理盘区“PEID=q”的表达行式指示物理卷 ID（PVol）是“p”的物理卷上的物理盘区 ID（PEID）是“q”的物理盘区。

在上面的例子中，在节点 1-1 的存储器 13-1 中的物理卷“PVol=1”/物理盘区“PEID=2”中存储的数据被迁移到节点 1-2 的存储器 13-2 中的物理卷“PVol=2”/物理盘区“PEID=1”。数据迁移源的存储器 13-1 和其迁移目的地的存储器 13-2 的每一个的 LPMT 132a 都被重写。这里，在存储器 13-1 的 LPMT 132a 中保持“PEID=2”的条目中的 LVol（=1），LEID（=4）和 MS（=1）全部被重写为“NULL（空）”。存储器 13-2 的 LPMT 132a 中的保持“PEID=1”的条目中的 LVol（=NULL），LEID（=NULL）和 MS（=NULL）分别被重写为 1、4 和 1，如图 11 中的箭头 C2 所示。

进一步，参考逻辑盘区“LEID=4”（或者负责关于逻辑盘区“LEID=4”

的位置管理服务)的管理服务器 12-1 的 LSMT 120-1 被重写。这里,在 LSMT 120-1 中的保持“LEID=4”的条目中的 PVol (=1) 被重写为 2,如图 11 中的箭头 C3 所示。

如上所述,在第一实施例中,仅须对管理逻辑盘区 (LEID=4) 的管理服务器 (12-1) 作出逻辑盘区 (LEID=4) 的迁移的通知 (更新)。换言之,对构成存储集群的所有节点 1-1 至 1-3 不需要作出该通知。因此,能降低通知的代价。

将参考图 12 描述将存储器中的形成逻辑盘区的物理盘区的数据复制到另一个存储器中的物理盘区的操作。为了通过制造物理盘区冗余来提高容错并通过并行处理增加读访问的速度,而进行该复制操作。

首先,构成管理单元 3 确定数据复制的源和目的地。这里假定当图 1 所示的存储集群被设定在图 11 所示的状态时,与逻辑盘区“LEID=4”关联的、在节点 1-2 的存储器 13-2 中的物理卷“PVol=2” (物理卷 130-2) / 物理盘区“PEID=1”被确定为数据的复制源。此外,假定节点 1-3 的存储器 13-3 中的物理卷“PVol=3” (物理卷 130-3) / 物理盘区“PEID=2”被确定为数据复制的目的地。

在这种情况下,如图 12 中的箭头 D1 所示,在节点 1-2 的存储器 13-2 的物理卷“PVol=2” / 物理盘区“PEID=1”中存储的逻辑卷“LVol=1” / 逻辑盘区“LEID=4”的数据被复制到节点 1-3 的存储器 13-3 中的物理卷“PVol=3” / 物理盘区“PEID=2”。通过,例如,存储器 13-2 和 13-3 的 IO 驱动器 131 进行该复制。

数据的复制目的地的存储器 13-3 的 LPMT 132a 被重写。这里,如图 12 中的箭头 D2 所示,在存储器 13-2 的 LPMT 132a 中保持“PEID=1”的条目中的 LVol (=1)、LEID (=4) 和 MS (=1) 被复制到在存储器 13-3 的 LPMT 132a 中保持“PEID=2”的条目中的 LVol、LEID 和 MS。

此外,参考逻辑盘区“LEID=4”的管理服务器 12-1 的 LSMT 120-1 被通过例如存储器 13-1 的 IO 驱动器 131 重写。这里,如图 12 中的箭头 D3 所示,“3”被加到由节点 1-1 的管理服务器 12-1 管理的 LSMT 120-1

中保持“LEID=4”的条目中的 PVol。如果 LSMT 120-1 中的条目被重写，则除了节点 1-2 的物理卷“PVol=2”之外，逻辑盘区“LEID=4”还与节点 1-3 的物理卷“PVol=3”关联。换言之，形成逻辑盘区“LEID=4”的物理盘区被使得对节点 1-2 的物理卷“PVol=2”和节点 1-3 的物理卷“PVol=3”冗余。

假定在使得形成逻辑盘区的物理盘区冗余之后，主机请求控制器 11-i 以得到对逻辑盘区的写访问。在这种情况下，控制器 11-i 将数据写到与所述逻辑盘区对应的所有物理盘区。因此，如果发出对逻辑盘区“LEID=4”的写访问，则数据被写入节点 1-2 的物理卷“PVol=2”/物理盘区“PEID=1”和节点 1-3 的物理卷“PVol=3”/物理盘区“PEID=2”。数据读取不限于此，换言之，从物理卷“PVol=2”/物理盘区“PEID=1”和物理卷“PVol=3”/物理盘区“PEID=2”中的一个读出数据。

下面将参考图 13 描述将关于逻辑盘区的位置管理服务迁移到另一个节点的操作。为了位置管理服务的负荷分配或为了形成节点的硬件的交换/更新来进行所述迁移。

当关于逻辑盘区的位置管理服务被迁移时，例如，构成管理单元 3 确定服务被迁移的源和目的地。这里假定当图 1 所示的存储集群被设定在图 12 所示的状态时，节点 1-1 的管理服务器 12-1 被确定为位置管理服务的迁移源。管理服务器 12-1 负责关于逻辑卷“LVol=1”/逻辑盘区“LEID=4”的命名服务。并且，假定节点 1-2 的管理服务器 12-2 被确定为命名服务的迁移目的地。

然后，如图 13 中的箭头 E1 所示，涉及迁移源的节点 1-1 的逻辑卷“LVol=1”/逻辑盘区“LEID=4”的 LSMT 120-1 中的条目的信息被迁移到迁移目的地的节点 1-2 的 LSMT 120-2。例如，迁移源的节点 1-1 的管理服务器 12-1，或例如迁移目的地的节点 1-2 的管理服务器 12-2 通知所有控制器 11-1 至 11-3：关于逻辑卷“LVol=1”/逻辑盘区“LEID=4”的位置管理服务被迁移到节点 1-2。

如图 13 中的箭头 E2 所示，控制器 11-1 至 11-3 的每个从节点 1-1（节

点 1-1 的管理服务器 12-1) 的 ID (MS=1) 到节点 1-2 (节点 1-2 的管理服务器 12-2) 的 ID (MS=2) 重写涉及逻辑卷 “LVol=1” / 逻辑盘区 “LEID=4” 的 LMMT 112a 中的条目的 MS。

管理逻辑卷 “LVol=1” / 逻辑盘区 “LEID=4” 的节点 1-2 和 1-3 的存储器 13-2 和 13-3 改变存储器 13-2 和 13-3 的 LPMT 132a 中的关于逻辑卷 “LVol=1” / 逻辑盘区 “LEID=4” 的位置管理服务的反向索引。如图 13 中的箭头 E3 所示, 节点 1-1 (节点 1-1 的管理服务器 12-1) 的 ID (MS=1) 被改变为节点 1-2 (节点 1-2 的管理服务器 12-2) 的 ID (MS=2)。存储器 13-2 和 13-3 返回与对位置管理服务的迁移的通知源节点的改变有关的信息。

当控制器 11-1 至 11-3 的每个都重写它们自己的 LMMT 112a 时, 仅须应用下面三种方法中的一种。第一种方法采用通过管理服务器 (管理服务器 12-1 或 12-2) 的广播。第二种方法采用通过控制器 (控制器 11-1 或 11-2) 的服务器指定单元的广播。第三种方法采用从迁移源对控制器 11-1 至 11-3 的每一个的更新通知。

例如, 节点 1-1 的管理服务器 12-1 确定来自节点 1-1 至 1-3 的每一个的响应并在 LSMT 120-1 中删除关于管理服务器 12-1 的逻辑卷 “LVol=1” / 逻辑盘区 “LEID=4” 的条目。于是, LSMT 120-1 中关于管理服务器 12-1 的逻辑卷 “LVol=1” / 逻辑盘区 “LEID=4” 的条目的数据被迁移到迁移目的地的节点 1-2 的 LSMT 120-2。换言之, 涉及逻辑卷 “LVol=1” / 逻辑盘区 “LEID=4” 的迁移源的节点 1-1 的 LSMT 120-1 中的条目和迁移目的地的节点 1-2 的 LSMT 120-2 中的条目被重写。

下面将参考图 14 描述在节点上复制关于逻辑盘区的位置管理服务的操作。当位置管理服务没有被多路复用, 如果负责位置管理服务的节点故障, 则不能进行从主机的访问请求。因而, 需要在节点上复制位置管理服务并对其多路复用 (使得其冗余), 从而提高容错和负荷分配。

当在节点上复制 (使得冗余) 位置管理服务时, 例如, 构成管理单元 3 确定复制的源和目的地。这里假定当图 1 所示的存储集群被设定在图 13

所示的状态时，节点 1-2 的管理服务器 12-2 被确定为位置管理服务的复制源。管理服务器 12-2 负责关于逻辑卷“LVol=1”/逻辑盘区“LEID=4”的命名服务。并且，假定节点 1-3 的管理服务器 12-3 被确定为所述命名服务的复制目的地。

然后，如图 14 中的箭头 F1 所示，LSMT 120-2 中涉及复制源的节点 1-2 的逻辑卷“LVol=1”/逻辑盘区“LEID=4”的条目的信息被复制到复制目的地的节点 1-3 的 LSMT 120-3。例如，复制源的节点 1-2 的管理服务器 12-2 或者，例如，复制目的地的节点 1-3 的管理服务器 12-3 通知所有控制器 11-1 至 11-3，节点 1-3 及节点 1-2 共享关于逻辑卷“LVol=1”/逻辑盘区“LEID=4”的位置管理服务的功能。

如图 14 中的箭头 F2 所示，控制器 11-1 至 11-3 的每一个将节点 1-3（节点 1-3 的管理服务器 12-3）的 ID（MS=3）添加到 LMMT 112a 中关于逻辑卷“LVol=1”/逻辑盘区“LEID=4”的条目的 MS。在添加节点 1-3 的 ID（MS=3）之前，节点 1-2（节点 1-2 的管理服务器 12-2）的 ID（MS=2）被设定到条目的 MS（列项 MS）。因此，在 LMMT 112a 中关于逻辑卷“LVol=1”/逻辑盘区“LEID=4”的条目与节点 1-3 以及节点 1-2 关联。

有了节点 1-3 的 ID 的添加，节点 1-3 新近管理逻辑卷“LVol=1”/逻辑盘区“LEID=4”。如图 14 中的箭头 F3 所示，节点 1-3 的存储器 13-3 将在存储器 13-3 的 LPMT 132a 中关于逻辑卷“LVol=1”/逻辑盘区“LEID=4”的位置管理服务的反向索引（MS）与节点 1-3 及节点 1-2 关联。存储器 13-3 返回与对位置管理服务的复制的通知源节点的改变有关的信息。当控制器 11-1 至 11-3 的每一个重写它自己的 LMMT 112a 时，在上述位置管理服务的迁移的情况中仅须应用上述第一至第三种方法中的一种。

以上描述了在多个节点（存储器）上的盘区的迁移/复制以及位置管理服务的迁移/复制。然而，盘区能够在在一个节点（存储器）的物理卷内被迁移/复制，并且，在这种情况下，仅须更新节点（存储器）的 LPMT 132a。此外，例如，为了冗余，能够复制管理服务器 12-i 的 LSMT 120-i 和存储

器 13-i 的 LPMT 132a。

下面将描述新将节点添加（加入）到图 1 所示的存储集群的操作。在这个操作中，至少 LMMT 112a 被从构成存储集群的节点 1-1 至 1-3 中的任意一个复制到新近添加到该存储集群的节点（新节点）。因而，所述新节点能进行 IO 操作。

为了与所述新节点共享节点 1-1 至 1-3 中的任意一个管理服务器（位置管理服务）的处理，例如，与构成管理单元 3 关联，仅须迁移位置管理服务。与构成管理单元 3 关联，能分别地使用通过加入新节点而在存储器中新近增加的物理卷/物理盘区。

根据第一实施例，如上所述，当由多个节点（在本实施例中为三个节点）构成存储集群时，能降低应该由每个节点管理的信息量。当在节点（节点的存储器）的物理卷内或在节点（节点的存储器）的物理卷间迁移/复制与逻辑盘区对应的物理盘区时，能够获得如下的优势。在存储集群中包括的所有控制器仅须重写管理服务器的信息且不需要更新该信息以使得彼此同步地通过控制器管理的信息项是相同的。在第一个实施例中，能简化在迁移/复制盘区时进行的处理并通常能改善访问性能。

[第二实施例]

在上述的第一实施例中，存储集群的节点 1-1，1-2 和 1-3 分别包括控制器 11-1，11-2 和 11-3，管理服务器 12-1，12-2 和 12-3，以及存储器 13-1，13-2 和 13-3。然而，每个节点不需要总是包括所有这些控制器、管理服务器和存储器。下面是本发明的第二实施例的描述，在其中存储集群的某些节点的每个包括控制器、管理服务器和存储器中的一个或两个。

图 15 是根据本发明第二实施例的具有存储集群结构的存储系统的框图。在图 15 中，为了方便，与图 7 中所示相同的组件用相同的参考数字来表示。与第一实施例类似，图 15 所示的存储系统（存储集群）包括三个节点 1-1 至 1-3。第二实施例的存储集群与第一实施例的存储集群的不同之处在于，三个节点的节点 1-2 和 1-3 的每个包括控制器、管理服务器和存储器中的一个和两个。在图 15 中，节点 1-1 包括与第一实施例相同的控制器

11-1、管理服务器 12-1 和存储器 13-1，而节点 1-2 只包括控制器 11-2 并且节点 1-3 只包括管理服务器 12-3 和存储器 13-3。图 15 所示的存储集群的操作与第一实施例的存储集群的操作相同。

假定在第二实施例中，虽然存储集群包括三个节点，但是控制器的数目、管理服务器的数目以及存储器的数目的每个都是两个。管理服务器的数目和存储器的数目的每个能被在整个存储集群中设定为一个。换言之，在包括多个节点的存储集群中，控制器的数目仅须为至少两个，且管理服务器的数目和存储器的数目的每个仅须为至少一个。在该限制条件下，可以适当地增加或减小控制器的数目、管理服务器的数目以及存储器的数目。利用第二实施例，当配置存储集群时，仅需根据情形扩展必要的功能。

[第三实施例]

下面将描述本发明的第三实施例。第三实施例的特征在于，由于直到主机实际请求对逻辑盘区的首次访问，在存储器侧不将物理盘区分配给该逻辑盘区，有效地使用了盘。

图 16 是根据本发明第三实施例的具有存储集群结构的存储系统的框图。在图 16 中，为了方便，与图 7 所示相同的组件由相同的参考数字来表示。与第一实施例类似，图 16 所示的存储系统（存储集群）包括三个节点 1-1 至 1-3。第三实施例的存储集群与第一实施例的存储集群的不同之处在于，直到主机请求对逻辑盘区的首次访问（例如，写访问），在存储器（13-i）侧不将物理盘区分配给逻辑盘区。

在图 16 的例子中，存储器 13-1 的 LPMT 132a 指示节点 1-1 中的物理卷 130-1 的物理盘区未被分配给逻辑卷“LVol=1”/逻辑盘区“LEID=1”或者逻辑卷“LVol=1”/逻辑盘区“LEID=4”。类似地，存储器 13-2 的 LPMT 132a 指示节点 1-2 中的物理卷 130-2 的物理盘区未被分配给逻辑卷“LVol=1”/逻辑盘区“LEID=2”。此外，存储器 13-3 的 LPMT 132a 指示节点 1-3 中的物理卷 130-3 的物理盘区未被分配给逻辑卷“LVol=1”/逻辑盘区“LEID=3”。

通过条目来指示不分配物理盘区，在该条目中信息 NA（标志 NA）被

设定给 LPMT 132a 中的 PEID (列项 PEID)。然而, 逻辑卷的信息、逻辑盘区的信息以及节点的信息在条目中分别被设定给其它的列项 LVol、LEID 和 MS。这指示存储器 (物理卷) 被分配给逻辑卷/逻辑盘区。换言之, 这指示在由 LPMT 132a 中的条目的 MS 指定的节点上预留了由该条目 (LVol, LEID) 指定的逻辑卷/逻辑盘区所必要的区域。应当注意, 为了方便, 在图 16 中改变了 LPMT 132a 的所述条目中的列项的顺序。

当物理盘区被分配给由 LPMT 132a 中的条目 (LVol, LEID) 指定的逻辑卷/逻辑盘区时, 该条目中的 PEID 被从 NA 重写为所分配的物理盘区的 ID。图 16 所示的存储集群的基本操作与根据第一实施例的存储集群的基本操作相同。然而, 在当创建逻辑卷时和当得到对逻辑盘区的首次访问 (写访问) 时作出的行为中, 前面的存储集群不同于在后的存储集群。

下面将描述在第三实施例的存储集群和第一实施例的存储集群之间的行为中的不同。根据第一实施例, 由上述的程序 (a)、(b) 和 (c) 来创建逻辑卷。根据第三实施例, 由于当得到对逻辑盘区的首次写访问时物理盘区被指定给该逻辑盘区, 因此用以下程序 (a1) 来替换上述程序 (a):

(a1) 构成管理单元 3 确定哪个节点 ID (PVol) 被分配给包括存储了形成要被创建的逻辑卷的逻辑盘区的存储器 (物理卷) 的节点, 以确保逻辑卷的容量。与第一实施例的程序 (a) 不同, 没有物理盘区被分配给各个逻辑盘区。构成管理单元 3 为存储器 13-1 至 13-3 的每一个生成指示没有物理盘区被分配给各逻辑盘区的 LPMT 132a, 并在存储器中保持 LPMT 132a。

下面将参考图 17 所示的流程图并利用图 1 至 3 来描述当主机得到对逻辑盘区的写访问时进行的图 16 所示的存储集群的操作。在图 17 中, 用相同的参考数字来表示与图 4 所示的流程图相同的组成部分。图 17 的流程图与图 4 的不同之处在于, 用步骤 S18 替换步骤 S8。步骤 S18 的特征在于: 步骤 S18 包括确定逻辑盘区被分配给的物理盘区的处理, 这是因为当对逻辑盘区的首次访问是写访问时没有物理盘区被分配给逻辑盘区。

与第一实施例中一样, 假定主机请求对节点 1-1 的控制器 11-1 中包括

的目标接口 111 的访问（写访问）。与在图 4 中所示的流程图中一样，首先执行步骤 S0 至 S7。在图 17 中，省略了步骤 S0 至 S7 中的步骤 S1 至 S6。在步骤 S7，主机请求节点 ID 是 SID_x 的存储器，例如节点 1-1 的存储器 13-1，以得到对与逻辑盘区“LVol $_x$, LEID $_x$ ”对应的物理盘区的访问。然后，在存储器 13-1 中，如下进行步骤 S18 的处理。

首先，存储器 13-1 的 IO 驱动器 131 参考在 LPMT 132a 中与逻辑盘区“LVol $_x$, LEID $_x$ ”关联的条目的 PEID（步骤 S18a）。根据 PEID 是否是非 NA（non-NA），存储器 13-1 确定物理盘区是否被分配给逻辑盘区“LVol $_x$, LEID $_x$ ”（步骤 S18b）。

如果 PEID 是 NA，并且因而没有物理盘区被分配给逻辑盘区“LVol $_x$, LEID $_x$ ”，则 IO 驱动器 131 与构成管理单元 3 关联地执行步骤 S18c。在步骤 S18c 中，IO 驱动器 131 将存储器 13-1（物理卷 130-1）中的适当（空）物理盘区 PEID $_x$ 分配给逻辑盘区“LVol $_x$, LEID $_x$ ”。在步骤 S18c，IO 驱动器 131 还将 LPMT 132a 中与逻辑盘区“LVol $_x$, LEID $_x$ ”关联的条目的 PEID 从 NA 更新为物理盘区 PEID $_x$ 的 ID（=PEID $_x$ ）。然后，IO 驱动器 131 得到对物理盘区 PEID $_x$ 的访问并答复控制器 11-1 的 IO 管理器 113（步骤 S18d）。另一方面，如果 PEID 是非 NA，并且因而物理盘区（例如，物理盘区 PEID $_x$ ）已经被分配给逻辑盘区“LVol $_x$, LEID $_x$ ”（步骤 S18b），则 IO 驱动器 131 跳过步骤 S18c 并执行步骤 S18d。

根据第三实施例，当对逻辑盘区的首次访问是写访问时，物理盘区被分配给逻辑盘区。另一方面，当对逻辑盘区的首次访问是读访问时，或当对没有数据被写入的逻辑盘区发出读访问时，物理盘区能被或不需要被分配给逻辑盘区。当它被分配时，IO 驱动器 131 仅须执行步骤 S18。在任一种情况下，仅须返回适当的数据（例如零数据）作为对读请求的响应。根据第三实施例的分配物理盘区的方法可以被应用于第二实施例。根据第三实施例，在每个存储器中，当主机请求访问时，能根据情形分配最佳的物理盘区。

[第四实施例]

下面将描述本发明的第四实施例。第四实施例的特征在于，直到主机实际请求对逻辑盘区的首次访问，不预留逻辑盘区的区域。

图 18 是根据本发明第四实施例的具有存储集群结构的存储系统的框图。在图 18 中，为了方便，与图 7 所示相同的组成部分由相同的参考数字来表示。与第一实施例中类似，图 18 所示的存储系统（存储集群）包括三个节点 1-1 至 1-3。第四实施例的存储集群与第一实施例的存储集群的不同之处在于，直到主机实际请求对逻辑盘区的首次访问，不但物理盘区不被分配给逻辑盘区，而且存储器（物理卷）也不被分配给逻辑盘区（不预留）。

在图 18 所示的存储集群中，直到请求对逻辑盘区的首次访问，逻辑盘区的区域不被预留。因此，当创建逻辑卷时，不需要预留存储器 13-1 至 13-3 的 LPMT 132a 的条目。响应于来自控制器 11-i（ $i=1, 2, 3$ ）的询问，由管理服务器 12-j（ $j=1, 2, 3$ ）进行关于逻辑盘区的对存储器（物理卷）和物理盘区的分配。例如，通过构成管理单元 3 计划该分配。

图 18 示出了没有存储器（物理卷）已经被分配给逻辑卷的各个逻辑盘区。更具体地，管理服务器 12-1 的 LSMT 120-1 指示没有存储器（物理卷）被分配给逻辑卷“LVol=1”/逻辑盘区“LEID=1”或逻辑卷“LVol=1”/逻辑盘区“LEID=4”。类似地，管理服务器 12-2 的 LSMT 120-2 指示没有存储器（物理卷）被分配给逻辑卷“LVol=1”/逻辑盘区“LEID=2”。此外，管理服务器 12-3 的 LSMT 120-3 指示没有存储器（物理卷）被分配给逻辑卷“LVol=1”/逻辑盘区“LEID=3”。通过在其中将 NA（标志 NA）设定给在 LSMT 120-1 至 120-3 中的 PVol（列项 PVol）的条目来指示存储器（物理卷）的未分配。应当注意，在图 18 中，从 LSMT 120-1 至 120-3 的条目到存储器 13-1 至 13-3 的 LPMT 132a 的条目没有绘制箭头。

图 18 所示的存储集群的基本操作与根据第一实施例的存储集群的基本操作相同。然而，当在创建逻辑卷时和当得到对逻辑盘区的首次访问时作出的行为中，前面的存储集群不同于在后的存储集群。

下面将描述在第四实施例的存储集群和第一实施例的存储集群之间的行为中的差别。根据第一实施例，由上述的程序（a）、（b）和（c）来创

建逻辑卷。根据第四实施例，当得到对逻辑盘区的首次访问时确定存储该逻辑盘区的存储器（物理卷），且在确定的存储器（物理卷）中对该逻辑盘区分配物理盘区。用下面的程序（a2）和（b2）来替换上述的程序（a）和（b）。

（a2）构成管理单元 3 不确定哪个节点 ID（PVol）被分配给包括存储了形成将被创建的逻辑卷的逻辑盘区的存储器（物理卷）的节点。如图 18 所示，构成管理单元 3 为存储器 13-1 至 13-3 的每一个生成空的 LPMT 132a，并在存储器中保持该 LPMT 132a。

（b2）构成管理单元 3 确定哪个节点包括了负责逻辑盘区的位置管理服务的管理服务器；然而，它不确定哪个存储器（物理卷）存储该逻辑盘区。因此，构成管理单元 3 生成包括条目信息项的 LSMT 120-1 至 120-3，并在管理服务器 12-1 至 12-3 中设定它们，其中所述条目信息项的 PVol 是 NA 且其涉及管理服务器 12-1 至 12-3 负责的逻辑盘区。与第二实施例中一样，不需要为存储集群中的节点 1-1 至 1-3 的每一个设置管理服务器（LSMT）。

下面将参考图 19 所示的流程图并利用图 1 至 3 来描述当主机得到对逻辑盘区的写访问时进行的图 18 所示的存储集群的操作。在图 19 中，与图 4 或图 17 所示的流程图相同的组成部分用相同的参考数字来表示。

与第一实施例中相同，假定主机请求对节点 1-1 的控制器 11-1 中包括的目标接口 111 的访问。如图 4 所示的流程图，首先执行步骤 S0 至 S4。在图 19 中，省略了步骤 S0 至 S4 中的步骤 S1 至 S3。在步骤 S4 中，例如，控制器 11-1 的服务器指定单元 112 询问例如管理服务器 12-1 的逻辑盘区（目标逻辑盘区）“LVolx, LEIDx”的位置。换言之，单元 112 询问存储了与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器的节点 ID “SIDx”。然后，在管理服务器 12-1 中，如下进行步骤 S25 的处理。

首先，管理服务器 12-1 参考 LSMT 120-1 以检查哪个节点 ID “SIDx”被分配给包括存储了与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器（物理卷）的节点（步骤 S25a）。这里，管理服务器 12-1 参考与逻

辑盘区“LVolx, LEIDx”关联的在 LSMT 120-1 中的条目的 PVol(=SIDx)。

根据 PVol 是否是非 NA, 管理服务器 12-1 确定存储器 (物理卷) 是否被分配给逻辑盘区“LVolx, LEIDx” (步骤 S25b)。如果没有存储器 (物理卷) 被分配给逻辑盘区“LVolx, LEIDx”, 则管理服务器 12-1 检查在存储器 13-1 至 13-3 (存储器 13-1 至 13-3 的物理卷 130-1 至 130-3) 中是否有空的物理盘区 (步骤 S25c)。通过例如管理服务器 12-1 与存储器 13-1 至 13-3 的 IO 驱动器 131 的通信来实现所述检查。如果在存储器 13-1 至 13-3 中没有空的物理盘区 (或如果没有得到对来自主机的访问请求的响应), 则 IO 错误发生 (步骤 S25d)。在这种情况下, 返回所述错误作为对来自主机的访问请求的响应 (步骤 S9)。

如果在存储器 13-1 至 13-3 的任一个中有空的物理盘区, 则管理服务器 12-1 选择具有空的物理盘区的存储器 (例如, 存储器 13-1) 作为适当的一个 (步骤 S25e)。在步骤 S25e, 管理服务器 12-1 将选择的存储器分配给逻辑盘区“LVolx, LEIDx”。在步骤 S25e, 管理服务器 12-1 还更新 LSMT 120-1 中的条目以使得在该条目中反映对逻辑盘区“LVolx, LEIDx”的存储器的分配。这里, 与逻辑盘区“LVolx, LEIDx”关联的 LSMT 120-1 中的条目的 PVol 被从 NA 更新到被分配给逻辑盘区“LVolx, LEIDx”的存储器 13-1 的节点 ID (物理卷 130-1 的节点 ID), 即, 1 (PVol=SIDx=1)。

当在步骤 S25e 中更新 LSMT 120-1 时, 类似第一实施例中图 4 所示的流程图的步骤 S4 执行步骤 S5 和 S6。因而, 管理服务器 12-1 经由询问源的服务器指定单元 112 通知控制器 11-1 被分配给逻辑盘区“LVolx, LEIDx”的存储器 13-1 的节点 ID (物理卷 130-1 的节点 ID), 即, 1 (PVol=SIDx=1)。当存储器 (物理卷) 被分配给逻辑盘区“LVolx, LEIDx” (步骤 S25b) 时, 立刻执行步骤 S5 和 S6。

在通过服务器指定单元 112 接收到节点 ID (=SIDx=1) 后, 目标接口 111 经由 IO 管理器 113 请求由节点 ID (=SIDx=1) 指示的节点 1-1 的存储器 13-1 以得到对与逻辑盘区“LVolx, LEIDx”对应的物理盘区的访问, 与第一和第三实施例一样 (步骤 S7)。在步骤 S7 之后的操作与在第三实

施例中的操作相同,且在存储器 13-1 中执行包括步骤 S18a 至 S18d 的步骤 S18。因此,当没有物理盘区被分配给逻辑盘区“LVolx, LEIDx”时,在物理卷 130-1 中的空的物理盘区被分配给逻辑盘区“LVolx, LEIDx”。这里假定物理盘区 PEIDx 被分配到那里。在这种情况下,将与逻辑盘区“LVolx, LEIDx”关联的 LPMT 132a 中的条目的 PEID 从 NA 更新到物理盘区 PEIDx 的 ID (=PEIDx)。

根据第四实施例,不仅当主机发出访问请求时能根据每个存储器的条件分配最佳的物理盘区,而且能创建其容量超出总物理卷容量的上限的逻辑卷。

[第五实施例]

下面将描述本发明的第五实施例。第五实施例的特征在于,与来自控制器 11-i (i=1, 2, 3) 的服务器指定单元 112 的管理服务器 12-j (j=1, 2, 3) 的询问以及对该询问的响应(询问的结果)有关的信息被存储在高速缓冲存储器 112b 中。该信息包括“LVolx, LEIDx”和“SIDx”。信息“LVolx, LEIDx”表示逻辑卷 LVolx/逻辑盘区 LEIDx,或逻辑盘区“LVolx, LEIDx”。信息“SIDx”表示存储了与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器的节点 ID。换言之,信息“SIDx”表示逻辑盘区“LVolx, LEIDx”的位置。当需要相同的询问时,如果使用高速缓冲存储器 112b,则能从高速缓冲存储器 112b 快速地获得与对询问的响应相对应的信息。

下面将参考图 20 所示的流程图并利用图 1 至 3 和 7,描述在第五实施例中当主机获得对逻辑盘区的访问时进行的操作。在图 20 中,用相同的参考数字来表示与图 4 所示的流程图相同的组成部分。

现在假定主机发出对节点 1-1 的控制器 11-1 中包括的目标接口 111 的访问请求。在这种情况下,与图 4 所示的流程图相同,首先执行步骤 S0 至 S2。在图 20 中,省略步骤 S0 至 S2 中的步骤 S0 和 S1。在步骤 S2 中,控制器 11-1 的目标接口 111 询问控制器 11-1 的服务器指定单元 112 哪个节点包括了存储逻辑盘区“LVolx, LEIDx”的存储器(逻辑盘区“LVolx, LEIDx”的位置)。

然后，不同于第一实施例，服务器指定单元 112 执行步骤 S31。在步骤 S31，单元 112 参考由其自身管理的高速缓冲存储器 112b 并检查存储了与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器的节点 ID “SIDx”。换言之，单元 112 检查表示逻辑盘区“LVolx, LEIDx”的位置的节点 ID “SIDx”。

基于步骤 S31 的执行结果，服务器指定单元 112 确定目标节点 ID “SIDx” 是否与逻辑盘区“LVolx, LEIDx”相关联地被存储在高速缓冲存储器 112b 中（即打击（hit）或误击（mishit））（步骤 S32）。如果发生误击，则如第一实施例中来自目标接口 111 的询问（步骤 S2）的情况，单元 112 执行步骤 S3 和 S4。

在第五实施例中，假定由节点 1-1 的管理服务器 12-1 管理存储了与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器的位置，且该存储器的节点 ID（=PVol）是 SIDx（=1），如第一实施例。基于 LSMT 120-1 获得节点 ID “SIDx”。在这种情况下，管理服务器 12-1 通知询问源的服务器指定单元 112，存储了与逻辑盘区“LVolx, LEIDx”对应的物理盘区的存储器的节点 ID “SIDx” 以及 LSMT 120-1 的序列号 SN（LSMT 序列号）（步骤 S5a）。第五实施例的步骤 S5a 与第一实施例的不同之处在于，不仅将节点 ID 而且将用于获得该节点 ID 的 LSMT 120-1 的序列号 SN 通知给单元 112。例如，每次更新 LSMT 120-1，就由管理服务器 12-1 增加序列号 SN。LSMT 120-2 和 LSMT 120-3 也是一样。

在接收到来自管理服务器 12-1 的通知后，服务器指定单元 112 产生包括了将被存储在高速缓冲存储器 112b 中的{(LVolx, LEIDx), SIDx, SN}的信息（即，关于询问及其响应的信息）（步骤 S5b）。当再次询问哪个节点包括存储了逻辑盘区“LVolx, LEIDx”的存储器时，单元 112 能在步骤 S31 的处理中快速地从高速缓冲存储器 112b 获得与所述询问的响应相对应的信息“SIDx”。

当发生打击时，在其中如上所述能从高速缓冲存储器 112b 获得与逻辑盘区“LVolx, LEIDx”关联的节点 ID “SIDx”（步骤 S32），服务器指

定单元 112 能立刻将节点 ID “SID_x” 传送给目标接口 111（步骤 S6）。因此，能减少来自服务器指定单元 112 的对于管理服务器的询问的数目，并且能提高访问速度。

然而，当更新管理服务器 12-1 至 12-3 上的 LSMT 120-1 至 120-3 时，需要在高速缓冲存储器 112b 中反映所述更新。原因如下。与第一实施例中一样，假定在存储器之间管理物理盘区的复制（冗余结构）（参见图 12）。如果，在这种情况下，在高速缓冲存储器 112b 中未反映 LSMT 120-1 至 120-3 的更新，则通过写访问重写物理盘区中的仅仅一个，且将会破坏冗余结构。为了避免这种情况，在第五实施例中对高速缓冲存储器的更新使用下面的机制。

首先，序列号 SN 被分配给 LSMT 120-1 至 120-3 中的每一个。每次通过 IO 驱动器 131 更新 LSMT 120-1 至 120-3 时，就通过例如存储器 13-1 至 13-3 的 IO 驱动器 131 来增加（更新）序列号 SN。用于在步骤 S5a 获得节点 ID “SID_x” 的 LSMT 120-j（j=1, 2, 3）的序列号 SN 通过服务器指定单元 112 与 “（LVol_x, LEID_x），SID_x” 结合被存储在高速缓冲存储器 112b 中（步骤 S5b）。当单元 112 确定在步骤 S32 发生误击时，执行步骤 S3, S4, S5a 和 S5b，然后单元 112 将节点 DI “SID_x” 传送给目标接口 111（步骤 S6）。

当得到对存储器 13-j 的访问时，控制器 11-i（i=1, 2, 3）将序列号 SN 传送给存储器 13-j。存储器 13-j 的 IO 驱动器 131 将该序列号 SN 与由 IO 驱动器 131 自身识别的 LSMT 120-j 的最新序列号 SN 进行比较。如果两个序列号彼此一致，则 IO 驱动器 131 确定在由控制器 11-i 的服务器指定单元 112 管理的高速缓冲存储器 112b 中未反映 LSMT 120-j 的最新更新。在这种情况下，存储器 13-j 的 IO 驱动器 131 通知控制器 11-i 其在高速缓冲存储器 112b 中的对应信息被无效。换言之，IO 驱动器 131 作为高速缓冲存储器无效化单元。在对所述通知的答复中，如单元 112 在步骤 S32 确定发生误击的情况，控制器 11-i 的服务器指定单元 112 再次询问管理服务器 12-j（步骤 S3 和 S4）。因此，单元 112 能从管理服务器 12-j 获取 LSMT

120-j 的最新信息并将其反映在高速缓冲存储器 112b 中。

如下所述，管理服务器 12-j 可作为高速缓冲存储器无效化单元。换言之，管理服务器 12-j 能监控控制器 11-j 的服务器指定单元 112 上的高速缓冲存储器 112b，并确定在高速缓冲存储器 112b 和由服务器 12-j 自身管理的 LSMT 120-j 之间是否发生误击。如果确定发生误击，则管理服务器 12-j 仅须通知服务器指定单元 112 其在高速缓冲存储器 112b 中的对应信息应该被放弃或更新。

对本领域技术人员来说将容易产生另外的优点或变型。因此，本发明在其更宽的方面不限于在此所示出和描述的具体细节和代表性实施例。于是，不脱离由所附权利要求及其等同内容所定义的一般发明构思的精神或范围，可以作出各种修改。

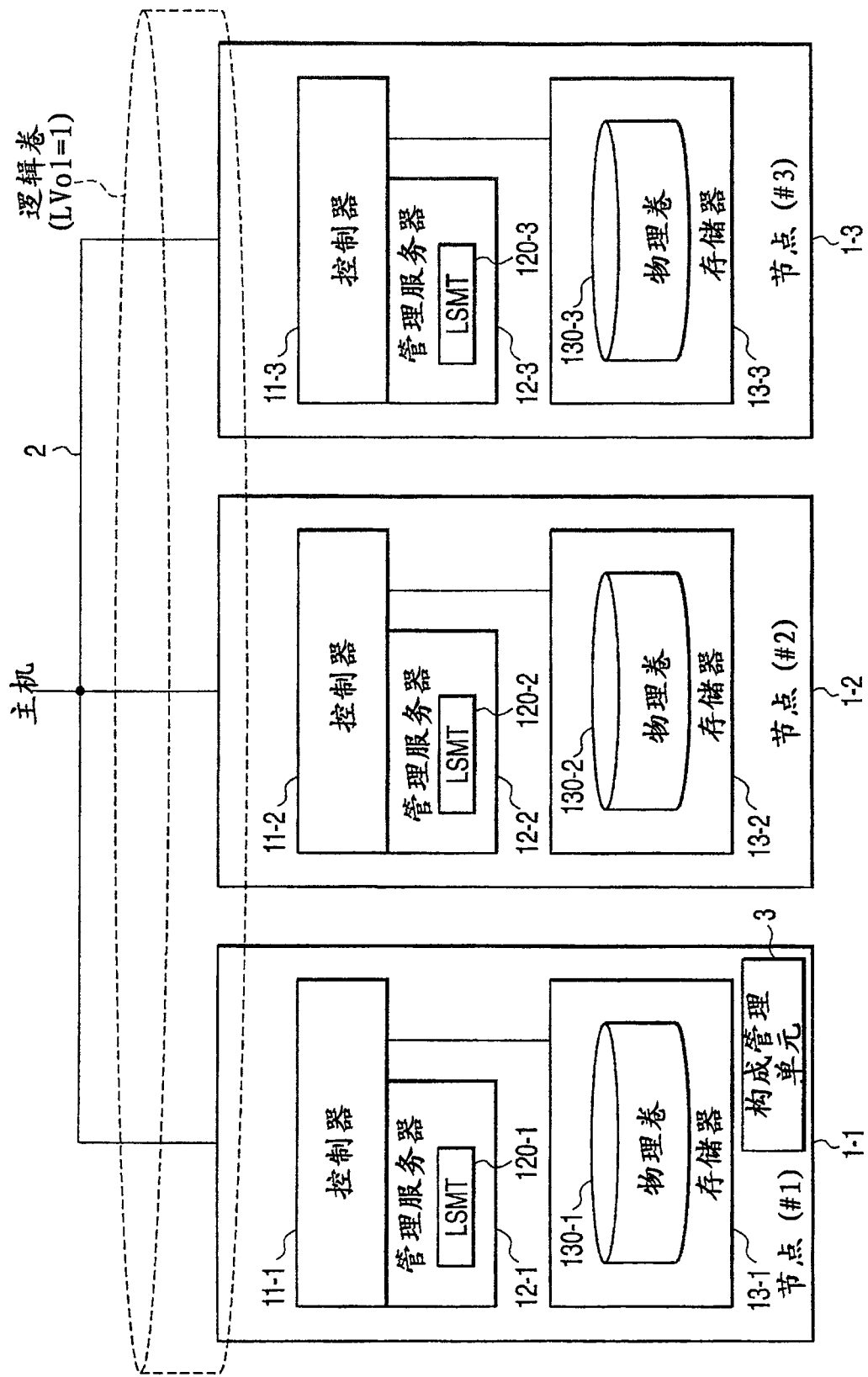


图 1

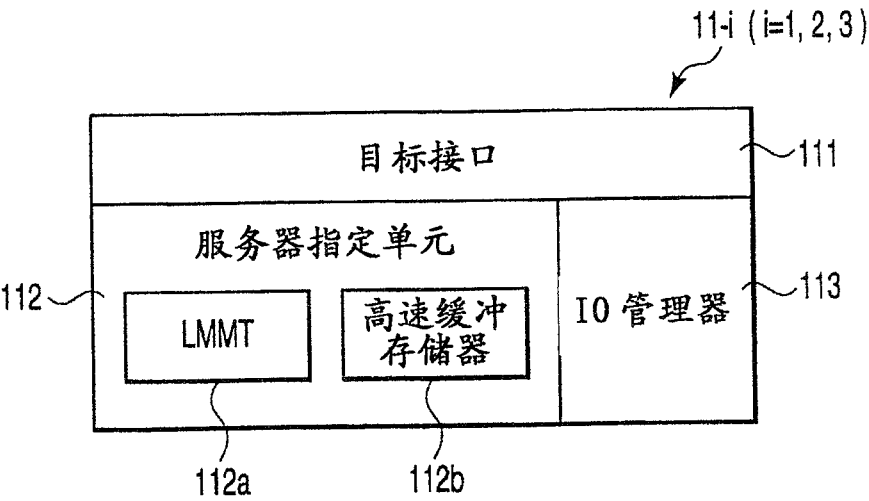


图 2

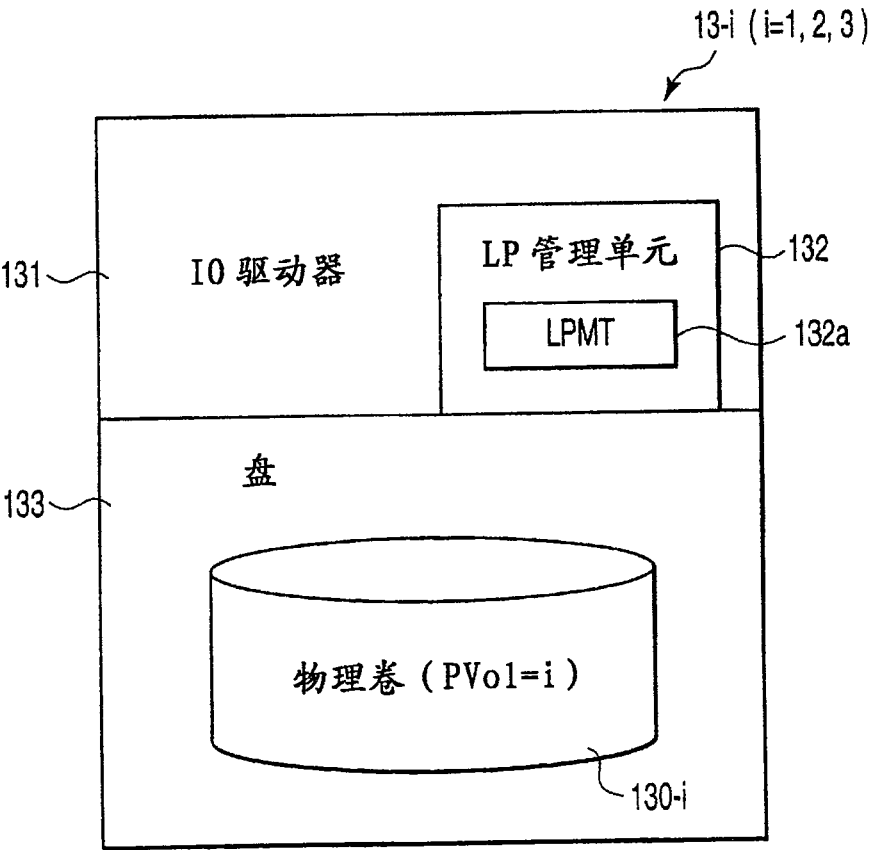


图 3

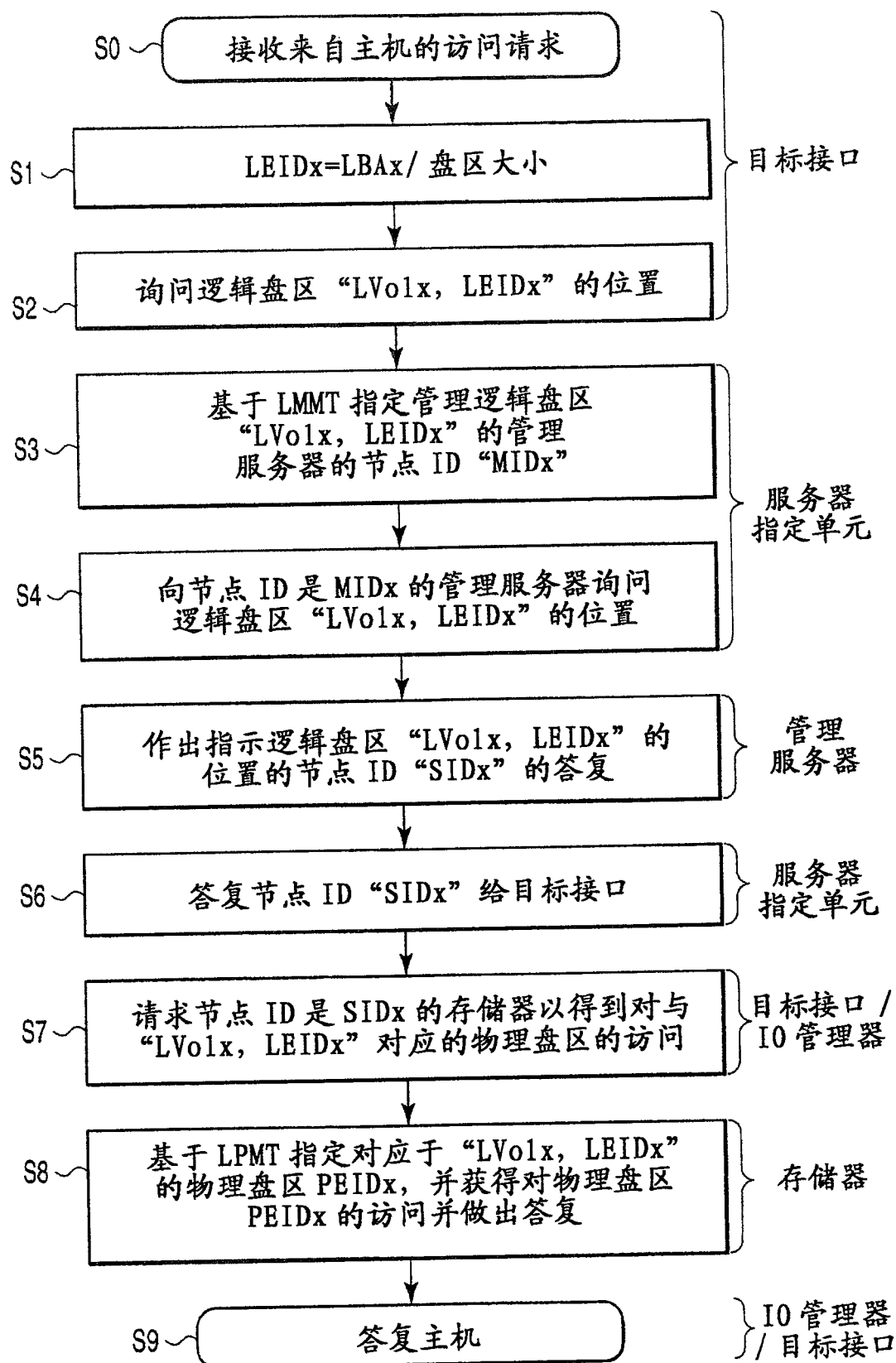


图 4

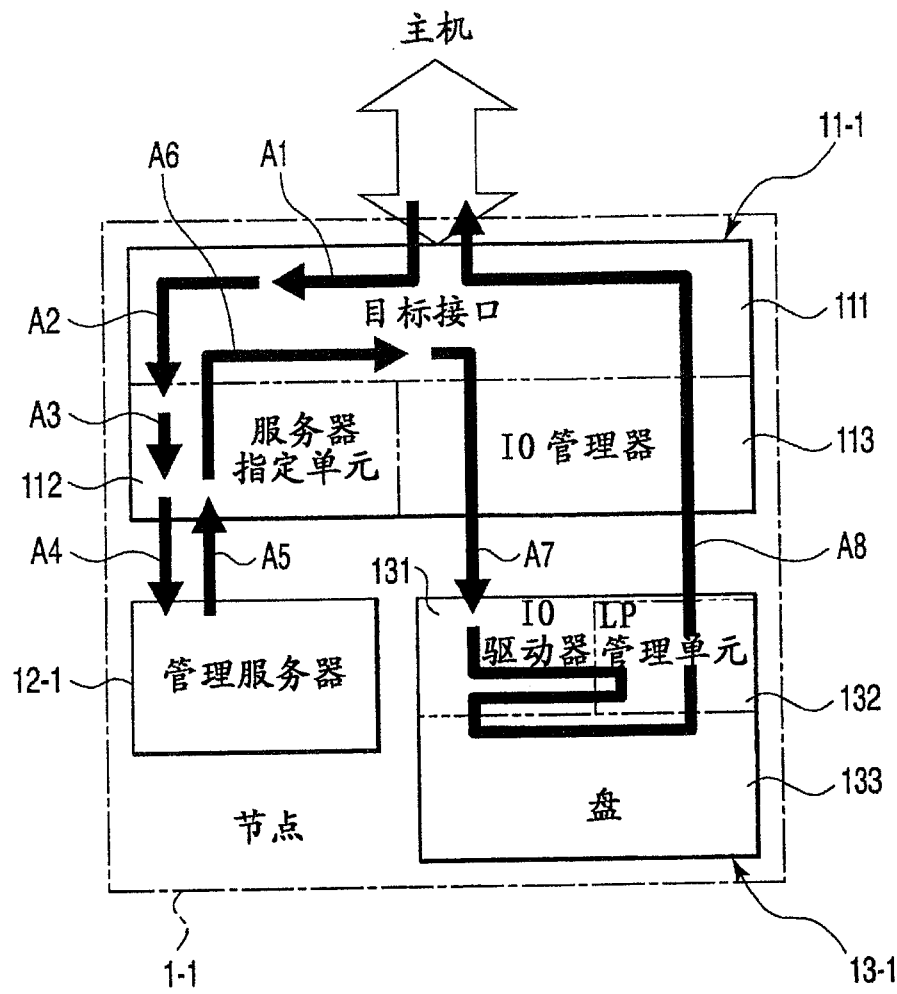


图 5

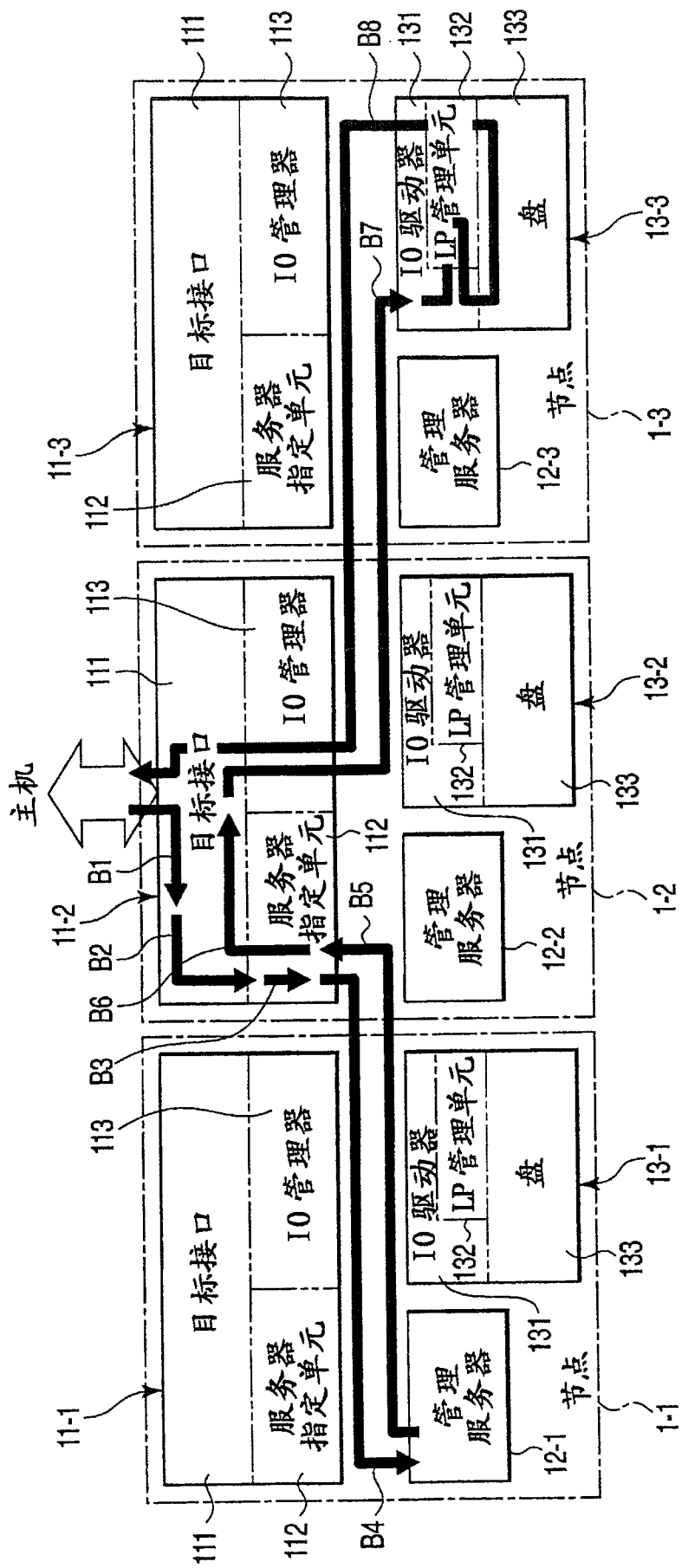


图 6

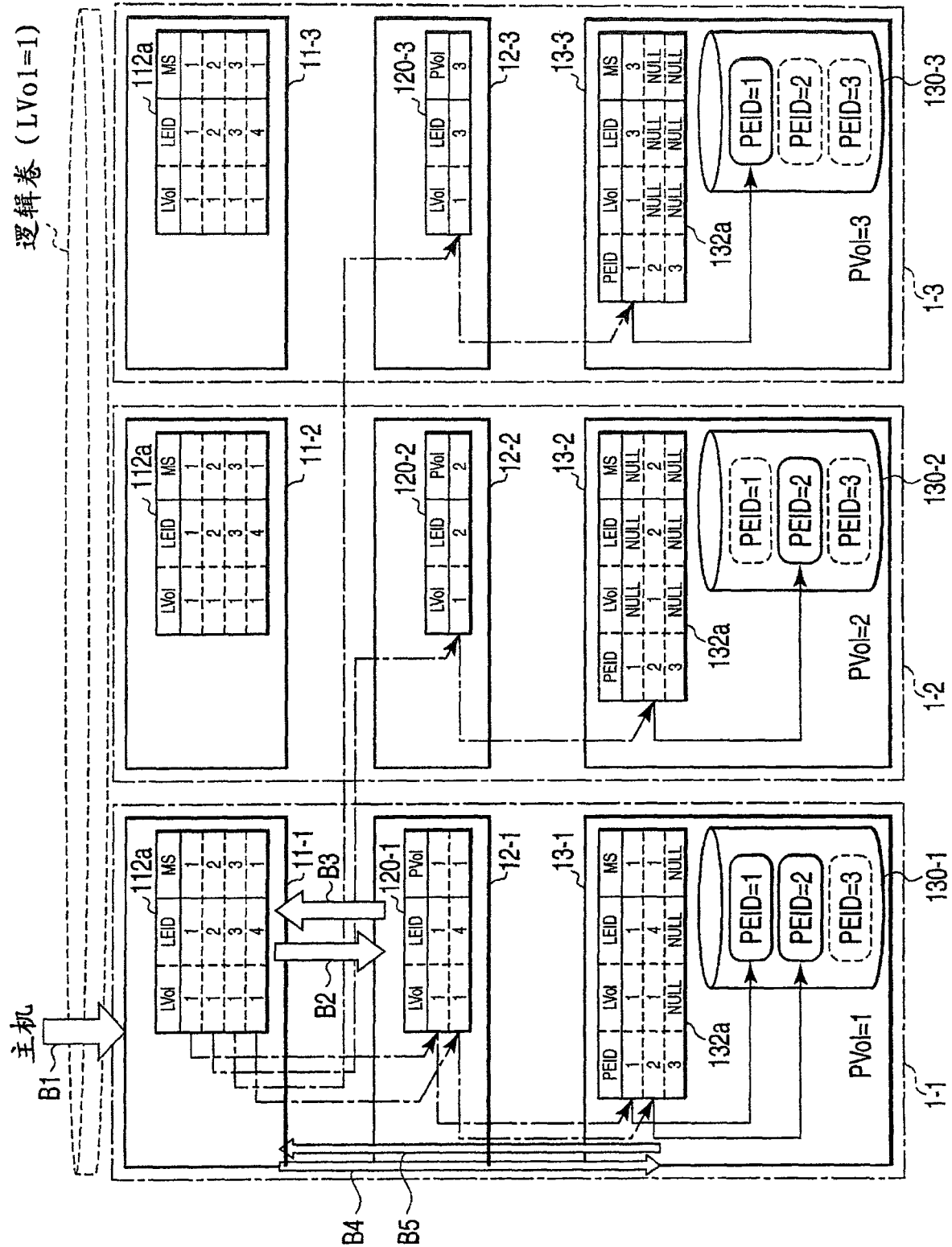


图 7

LMMT

112a

LVol	LEID	MS
1	1	1
1	2	2
1	3	3
1	4	1

图 8

LSMT

120-i (i=1)

LVol	LEID	PVol
1	1	1
1	4	1

图 9

LPMT

132a

PEID	VLol	LEID	MS
1	1	1	1
2	1	4	1
3	NULL	NULL	NULL

图 10

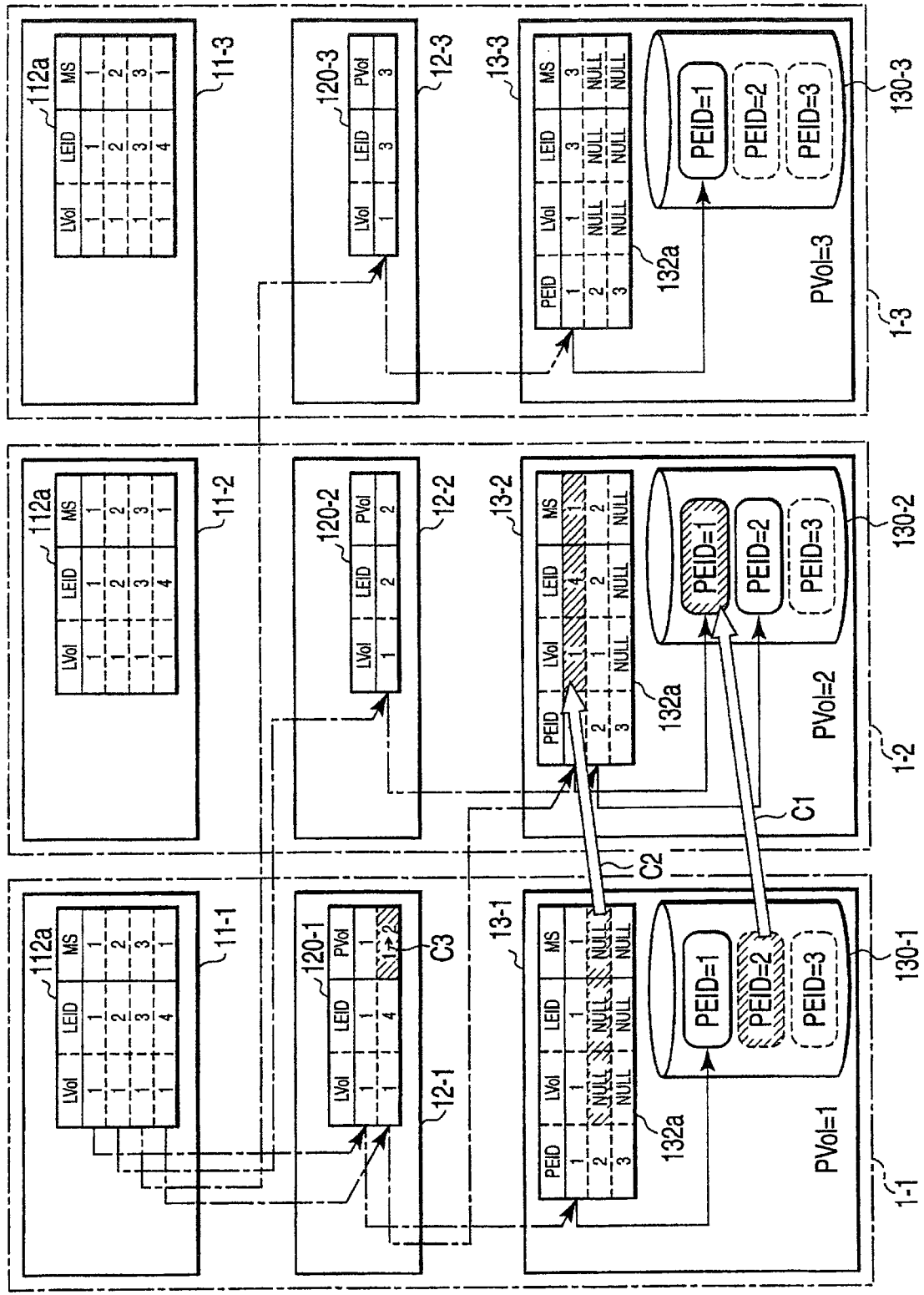


图 11

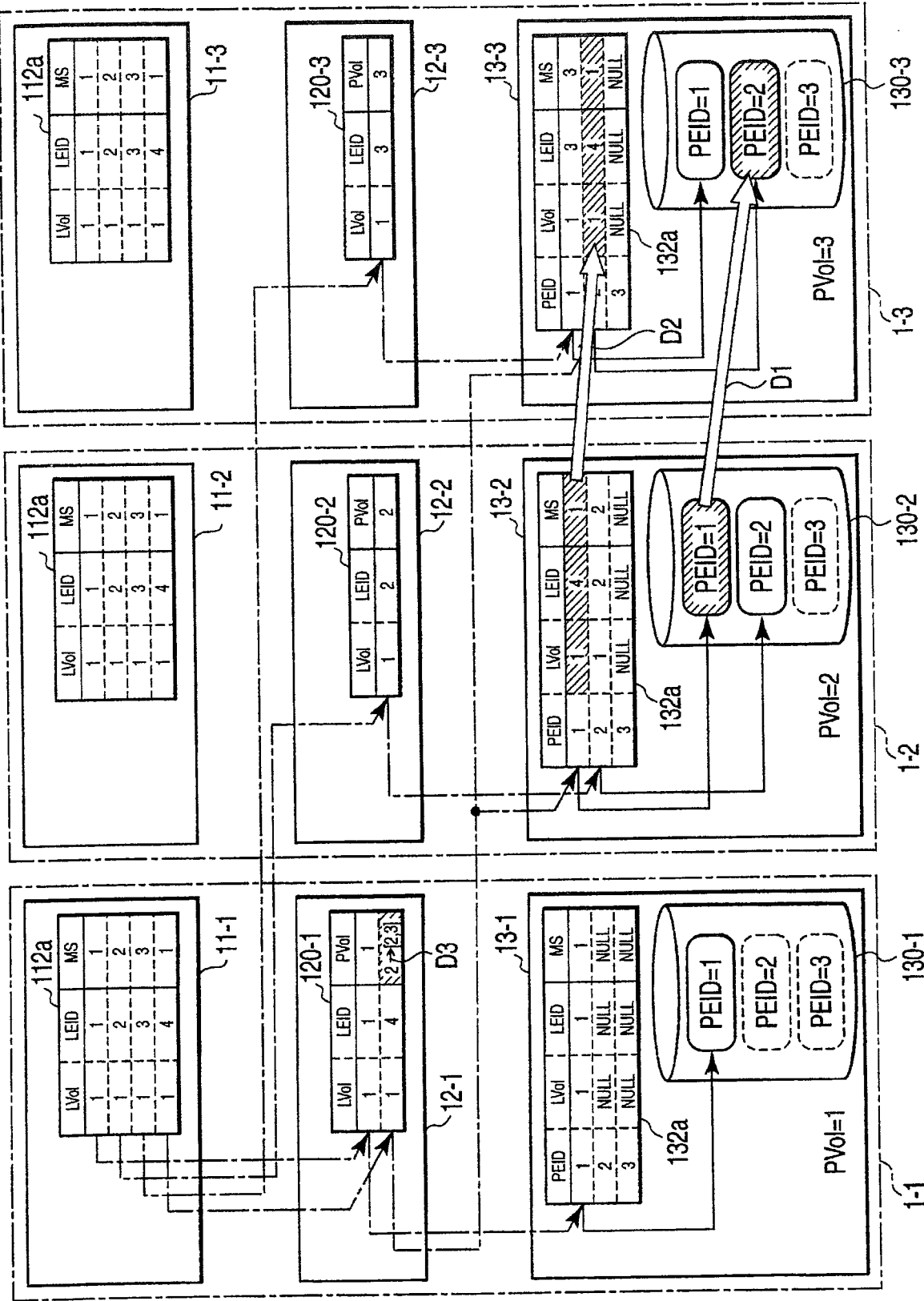


图 12

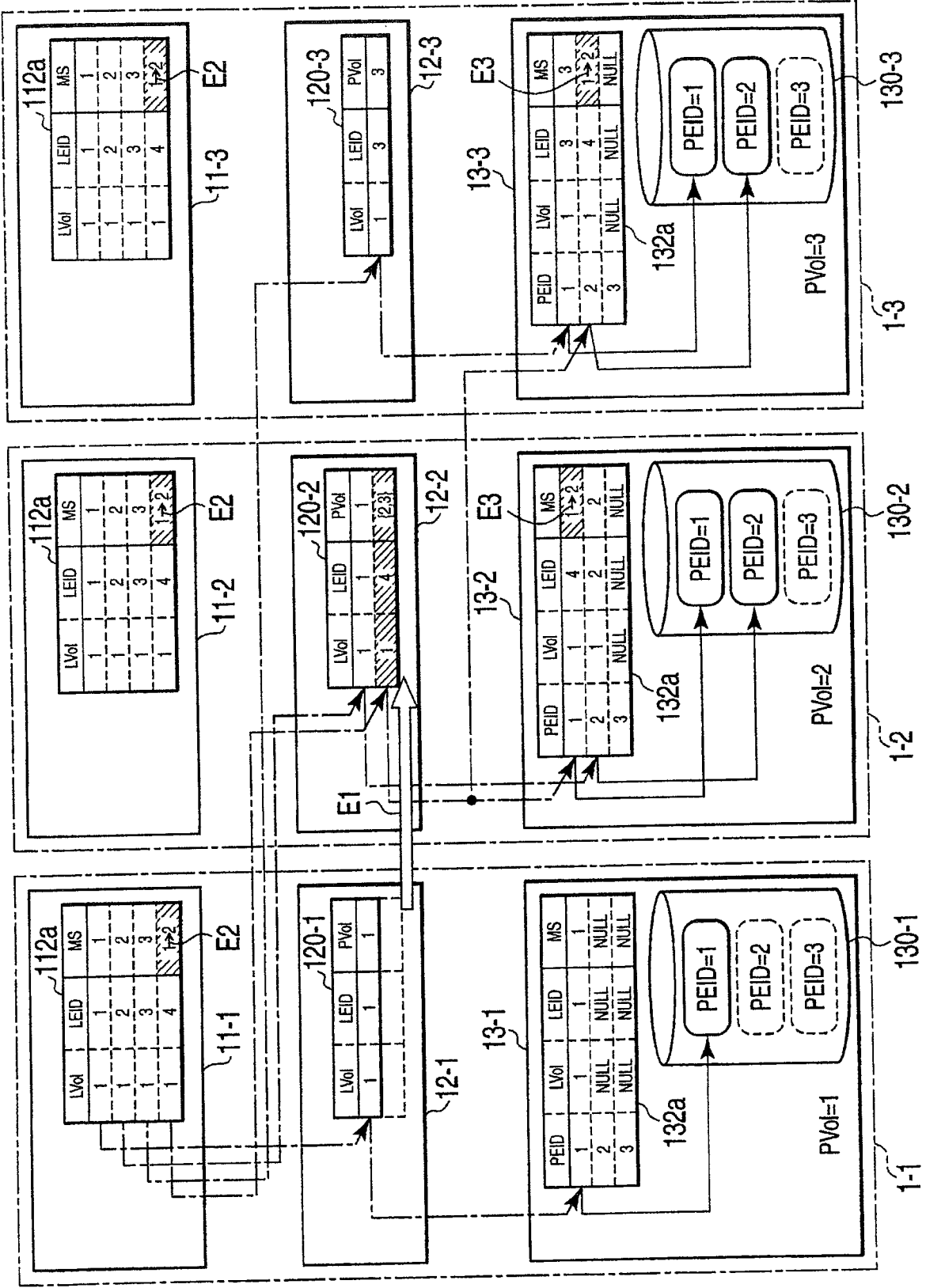


图 13

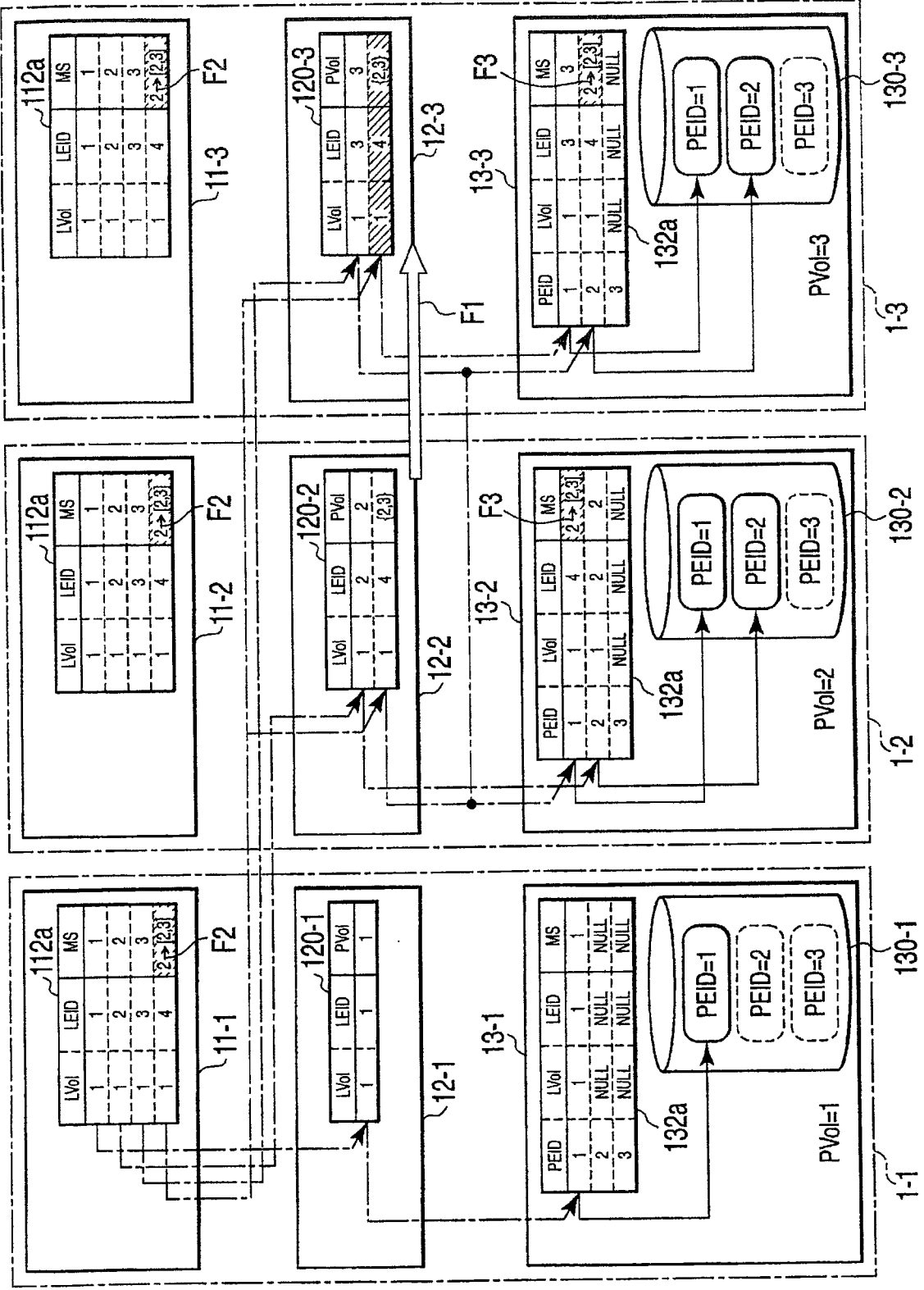


图 14

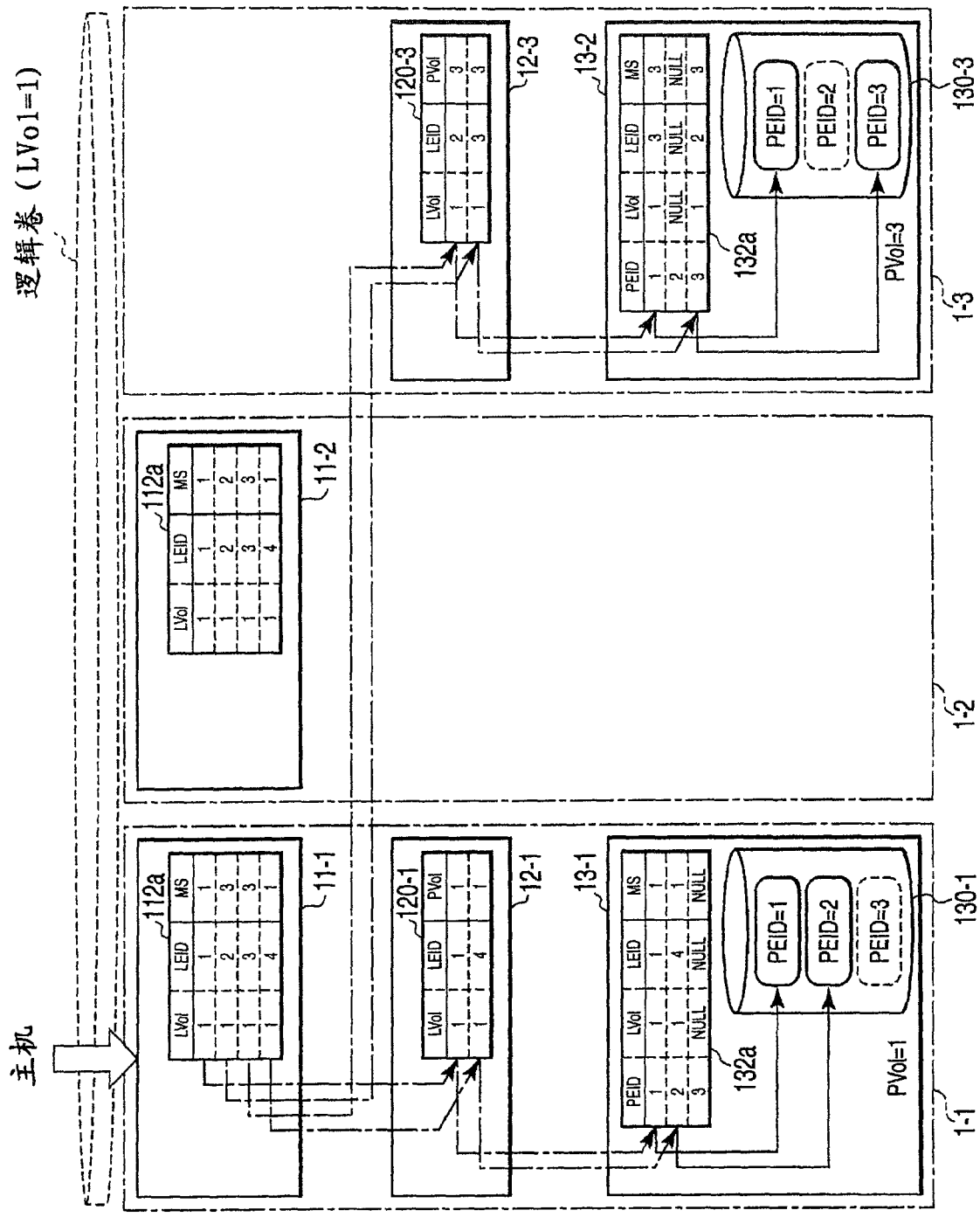


图 15

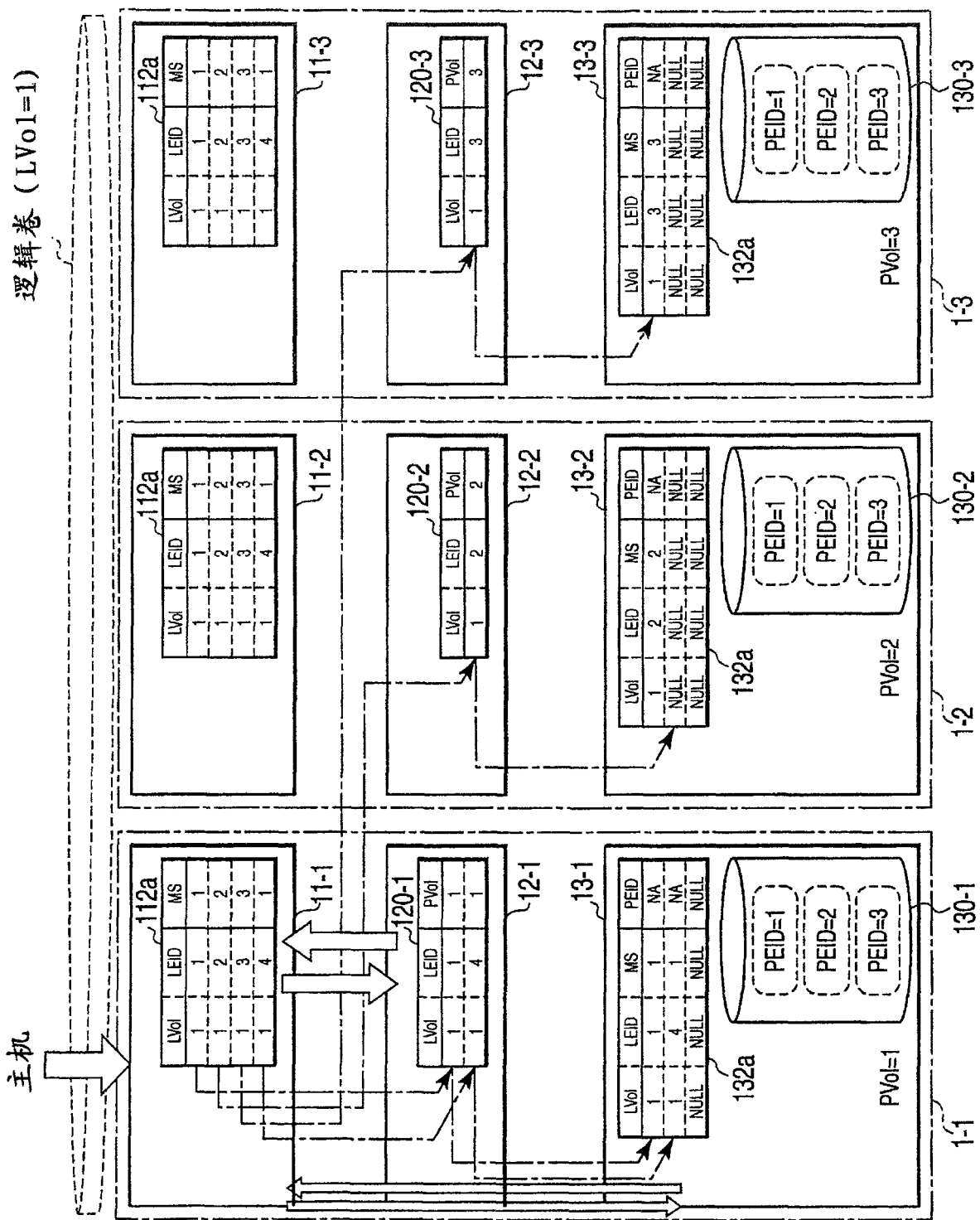


图 16

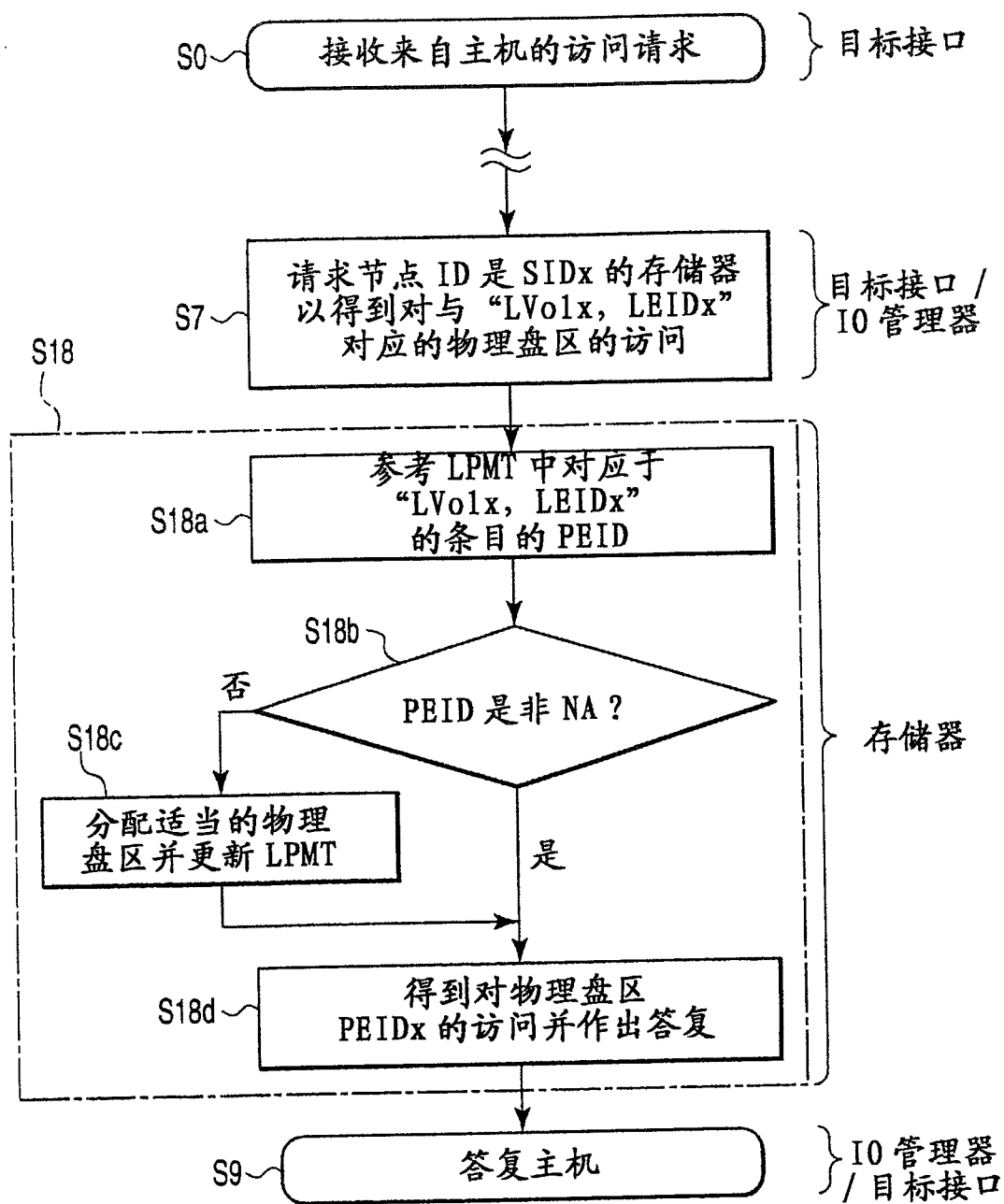


图 17

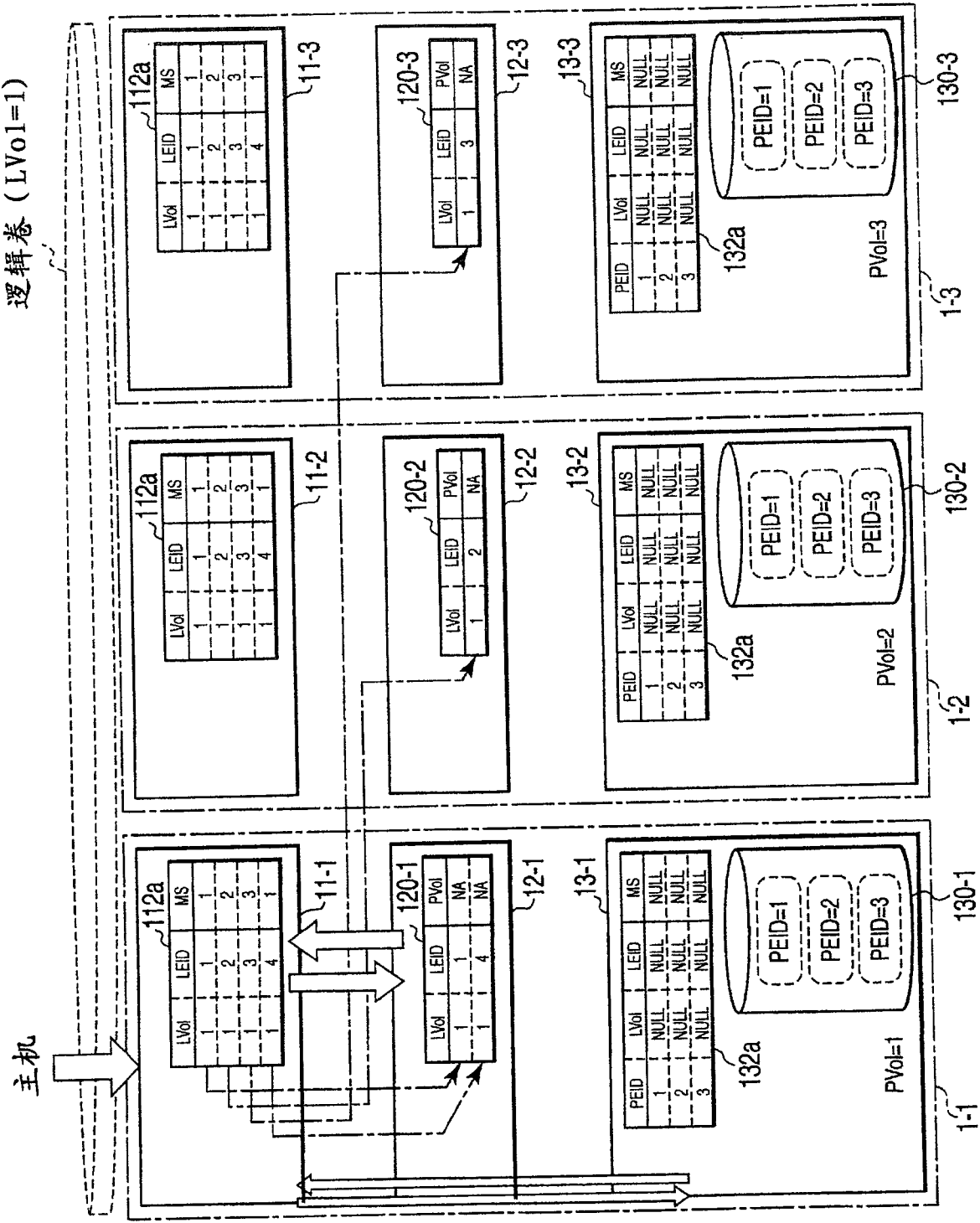


图 18

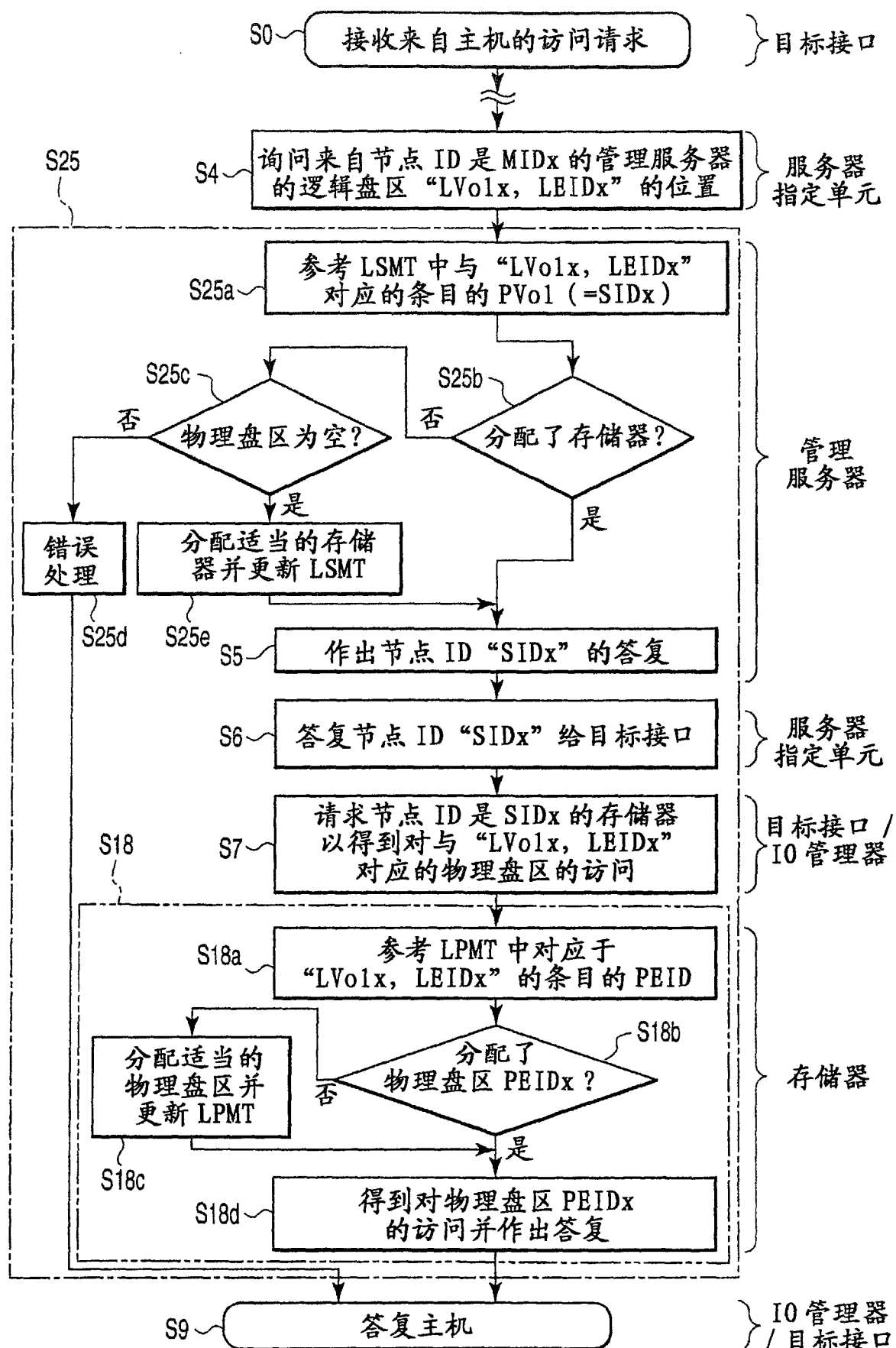


图 19

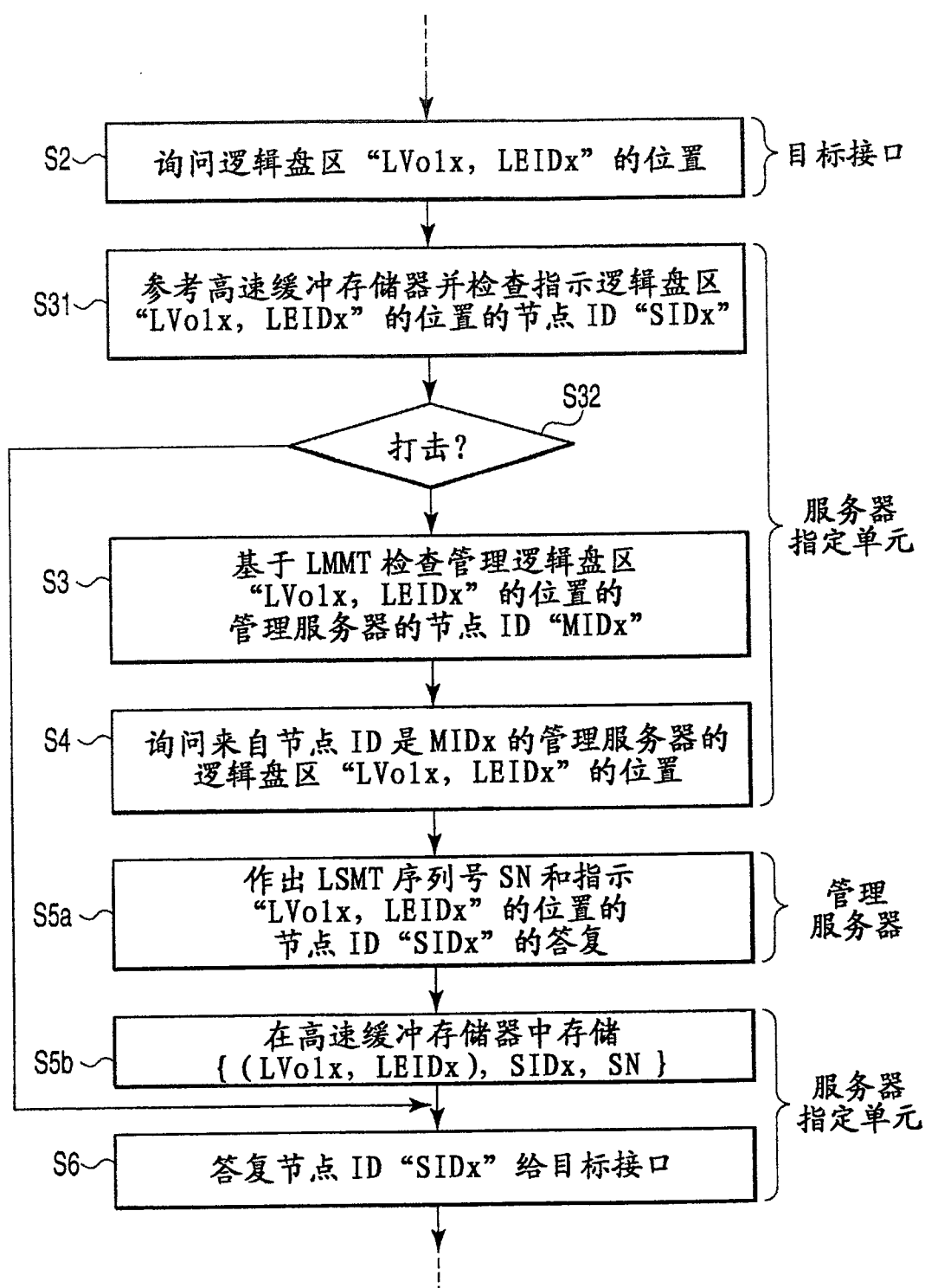


图 20