[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 11/10 (2006.01)
G06F 11/16 (2006.01)
G06F 3/06 (2006.01)

[12] 发明专利说明书

专利号 ZL 200510085417.3

[45] 授权公告日 2008 年 6 月 4 日

[11] 授权公告号 CN 100392611C

[22] 申请日 2005.7.18
[21] 申请号 200510085417.3
[30] 优先权
[32] 2005. 3.15 [33] JP [31] 073669/2005
[73] 专利权人 富士通株式会社
地址 日本神奈川县
[72] 发明人 望月信哉 伊藤实希夫
大黑谷秀治郎 池内和彦 高桥秀夫
绀田与志仁 佐藤靖丈 越智弘昭
牧野司 久保田典秀
[56] 参考文献
US5666512A 1997.9.9
JP11 - 85410A 1999.3.30
JP6 - 19632A 1994.1.28
CN1503224A 2004.6.9
审查员 赵 颖
[74] 专利代理机构 北京东方亿思知识产权代理有限责任公司
代理人 赵淑萍

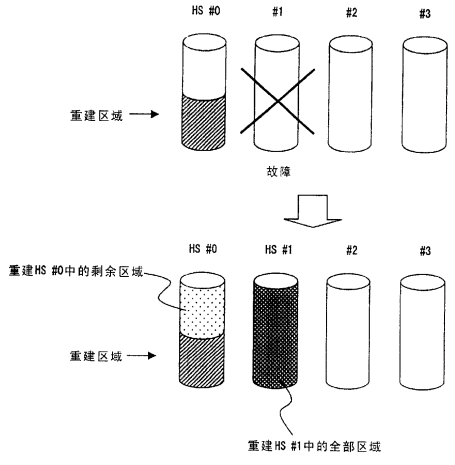
权利要求书 3 页 说明书 26 页 附图 25 页

[54] 发明名称

存储控制装置和方法

[57] 摘要

在将数据和奇偶校验位分散地存储在多个存储设备的系统中，当第一存储设备发生故障时，利用存储在除了第一存储设备外的其他存储设备中的信息来恢复第一存储设备中的信息。 并且当在正恢复第一存储设备中的信息的同时，第二存储设备发生故障时，利用存储在除了第一和第二存储设备外的其他存储设备中的信息来恢复第一存储设备中未恢复区域中的信息和第二存储设备中的信息。



1. 一种用于通过进行控制而实现数据冗余的存储控制装置，其中经由所述控制，数据和奇偶校验位被分散地存储在多个存储设备中，所述存储控制装置包括：

第一重建设备，用于当所述多个存储设备中的第一存储设备发生故障时，利用存储在除了所述第一存储设备外的其他存储设备中的信息来恢复所述第一存储设备中的信息，并将恢复的信息写入第一备用存储设备中；以及

第二重建设备，用于当在正恢复所述第一存储设备中的信息的同时，第二存储设备发生故障时，利用存储在除了所述第一和第二存储设备外的其他存储设备中的信息来恢复所述第一存储设备中未恢复区域中的信息和所述第二存储设备中的信息，并将恢复的信息分别写入所述第一备用存储设备的对应区域中和第二备用存储设备中。

2. 如权利要求 1 所述的存储控制装置，其特征在于：

所述第二重建设备独立且并行地分别执行恢复所述第一存储设备中未恢复区域的信息的处理和恢复所述第二存储设备中的信息的处理。

3. 如权利要求 1 所述的存储控制装置，其特征在于：

所述第二重建设备利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中已恢复区域的区域中的信息，来恢复所述第二存储设备中对应区域中的信息，其后，利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中未恢复区域的区域中的信息，来恢复所述第一存储设备中未恢复区域中的信息和所述第二存储设备中对应区域中的信息。

4. 如权利要求 1 所述的存储控制装置，其特征在于：

所述第二重建设备将所述第一存储设备中恢复进度位置和所述第二存储设备中恢复进度位置之间的差别与阈值相比较，并且当所述恢复进度位置之间的差别等于或大于所述阈值时，独立且并行地分别执行恢复所述第一存储设备中未恢复区域中的信息的处理和恢复所述第二存储设备中的信

息的处理，并且当所述恢复进度位置之间的差别小于所述阈值时，利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中已恢复区域的区域中的信息，来恢复所述第二存储设备中对应区域的信息，其后，利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中未恢复区域的区域中的信息，来恢复所述第一存储设备中未恢复区域中的信息和所述第二存储设备中对应区域中的信息。

5. 如权利要求 1 所述的存储控制装置，其特征在于：

所述第二重建设备利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中未恢复区域的区域中的信息，来恢复所述第一存储设备中未恢复区域中的信息和所述第二存储设备中对应区域中的信息，其后，利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中已恢复区域的区域中的信息，来恢复所述第二存储设备中对应区域中的信息。

6. 如权利要求 1 所述的存储控制装置，其特征在于：

所述第二重建设备并行执行前一个处理和后一个处理，所述前一个处理利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中已恢复区域的区域中的信息，来恢复所述第二存储设备中对应区域中的信息，所述后一个处理利用存储在除了所述第一和第二存储设备外的其他存储设备中、在对应于所述第一存储设备中未恢复区域的区域中的信息，来恢复所述第一存储设备中未恢复区域中的信息和所述第二存储设备中对应区域中的信息。

7. 如权利要求 1 到 6 中任何一个所述的存储控制装置，还包括：

用于保存位图信息的保存设备，所述位图信息指示对于所述第一和第二存储设备中的每个规定区域，是否已恢复了该区域，其中：

所述第二重建设备通过参考所述位图信息，来恢复除了已恢复区域外的其他区域中的信息。

8. 如权利要求 7 所述的存储控制装置，其特征在于：

当发生对所述第一或第二存储设备的访问请求时，所述第二重建设备

恢复作为访问目标的信息，并且在所述位图信息中与作为所述访问目标的信息相对应的位置处，记录指示出所述对应区域已恢复的信息。

9. 一种存储装置，包括：

多个存储设备，用于分散地存储数据和奇偶校验位，以实现数据冗余；

第一重建设备，用于当所述多个存储设备中的第一存储设备发生故障时，利用存储在除了所述第一存储设备外的其他存储设备中的信息来恢复所述第一存储设备中的信息，并将恢复的信息写入第一备用存储设备中；以及

第二重建设备，用于当在正恢复所述第一存储设备中的信息的同时，第二存储设备发生故障时，利用存储在除了所述第一和第二存储设备外的其他存储设备中的信息来恢复所述第一存储设备中未恢复区域中的信息和所述第二存储设备中的信息，并将恢复的信息分别写入所述第一备用存储设备的对应区域中和第二备用存储设备中。

10. 一种用于通过进行控制实现数据冗余的存储控制方法，其中经由所述控制，数据和奇偶校验位被分散地存储在多个存储设备中，所述方法包括以下步骤：

当所述多个存储设备中的第一存储设备发生故障时，利用存储在除了所述第一存储设备外的其他存储设备中的信息来恢复所述第一存储设备中的信息，并将恢复的信息写入第一备用存储设备中；以及

当在正恢复所述第一存储设备中的信息的同时，第二存储设备发生故障时，利用存储在除了所述第一和第二存储设备外的其他存储设备中的信息来恢复所述第一存储设备中未恢复区域中的信息和所述第二存储设备中的信息，并将恢复的信息分别写入所述第一备用存储设备的对应区域中和第二备用存储设备中。

存储控制装置和方法

技术领域

本发明涉及这样的存储控制装置和方法，其中，数据和奇偶校验位（parity）被分散地存储在诸如 RAID（廉价盘冗余阵列）的多个存储设备中，并且在存储设备发生故障时执行数据和奇偶校验位的重建处理。

背景技术

上述 RAID 是这样的技术，在该技术中，多个硬盘被组合并被管理为具有冗余度的一个硬盘。并且 RAID 可以根据将数据分配到盘的方法和实现数据冗余的方法而被分类为从 RAID 0 到 RAID 6 的七个级别。在 RAID 的七个级别中，在 RAID 3 到 RAID 6 中，通过将基于数据生成的奇偶校验位与数据相互分离地存储，从而实现冗余。在盘故障的情形下，执行重建处理以利用奇偶校验位来恢复故障盘中的数据（例如参见日本专利申请公开 No. 03-240123）。

RAID 6 是减轻了两个盘中的故障的 RAID 级别。在 RAID 6 中，两种不同类型的奇偶校验位 P 和 Q 被分别分散地存储在不同的盘中，并且在一个盘故障和两个盘故障的重建处理中，分别采用不同的恢复方法。

例如，当如图 1A 所示，在由五个盘 10 到 14 组成的 RAID 设备中，数据 D0 由于盘 10 的故障而丢失时，利用作为备用盘的热备份 15 来执行一个盘故障的重建处理。这时，基于存储在其他盘 11 到 13 中的数据 D1、D2 和奇偶校验位 P 来恢复数据 D0。

与上述相反，当如图 1B 所示，数据 D0 和 D1 由于盘 10 和 11 中的故障而丢失时，利用热备份 15 和 16 来执行两个盘故障的重建处理。这时，基于存储在其他盘 12 到 14 中的数据 D2 以及奇偶校验位 P 和 Q 来恢复数据 D0 和 D1。

通常在 RAID 6 中，对于每个条带（striping），在故障盘中存储有不

同种类的信息并且为了恢复存储在故障盘中的信息需要这些不同种类的信息，这是因为不同的盘存储每个条带的数据和奇偶校验位。因此，在下面的解释中，存储在每个盘中的信息被称为数据/奇偶校验位。

当一个盘故障的情形变为两个盘故障的情形时，重建处理也从一个盘故障的重建处理切换到两个盘故障的重建处理。例如，当如图 1C 所示，在执行一个盘故障的重建处理的同时（其中第一故障盘#0 被热备份（HS）取代），第二盘#1 发生故障，则通过以上一个盘故障的重建处理不能恢复数据/奇偶校验位。因此，通常的设计是，在停止一个盘故障的重建处理之后，利用热备份 HS #0 和 HS #1 开始两个盘故障的重建处理。

上述两个盘故障的重建处理有下面的问题。

在图 1C 所示两个盘故障的重建处理中，丢弃已存储在热备份 HS #0 的重建区域中的已恢复的数据/奇偶校验位，并且从最开始再次执行重建处理。因此，没有有效地利用已恢复的数据/奇偶校验位。

另外，在热备份 HS #0 和 HS #1 的整个区域上执行两个盘故障的重建处理（该处理比一个盘故障的重建处理开销要大），从而冗余恢复需要更长的时间。

发明内容

本发明的目的是在诸如 RAID 的存储装置（其中利用奇偶校验位实现了数据冗余）中，通过改进在两个用于存储数据/奇偶校验位的存储设备发生故障时的重建处理，来有效地恢复冗余。

根据本发明的存储控制装置包括第一重建设备和第二重建设备，并且通过进行控制而实现了数据冗余，其中经由所述控制，数据和奇偶校验位被分散地存储在多个存储设备中。

第一重建设备当第一存储设备发生故障时，利用存储在除了第一存储设备外的其他存储设备中的信息来恢复第一存储设备中的信息，并将恢复的信息写入第一备用存储设备中。第二重建设备当在正恢复第一存储设备中的信息的同时，第二存储设备发生故障时，利用存储在除了第一和第二存储设备外的其他存储设备中的信息来恢复第一存储设备中未恢复区域的

信息和第二存储设备中的信息，并将恢复的信息分别写入第一备用存储设备的对应区域中和第二备用存储设备中。

附图说明

图 1A 示出了一个盘故障时的数据恢复；

图 1B 示出了两个盘故障时的数据恢复；

图 1C 示出了对两个盘故障的重建处理；

图 2A 示出了根据本发明的存储控制装置的原理；

图 2B 示出了第一存储系统的配置；

图 3 示出了重建处理的计数方案；

图 4 示出了计数方案 1；

图 5 示出了计数方案 2；

图 6 示出了计数方案 4；

图 7 示出了计数方案 5；

图 8 示出了计数方案 6；

图 9 示出了待处理主盘（treated Main）和待处理副盘（treated Sub）中的当前位置；

图 10 示出了每个盘中的恢复状态；

图 11 是重建处理的流程图；

图 12 是计数方案 1 中恢复例程的流程图；

图 13 是计数方案 1 中恢复后处理的流程图；

图 14 是计数方案 2 到 5 中恢复例程的流程图；

图 15 是计数方案 2 中恢复后处理的流程图；

图 16 是计数方案 3 中恢复后处理的流程图；

图 17 是计数方案 4 中恢复后处理的流程图；

图 18 是计数方案 5 中恢复后处理的流程图；

图 19 示出了计数方案 6 中的恢复状态；

图 20 示出了提供程序和数据的方法；

图 21 示出了第二存储系统的配置；以及

图 22 示出了第三存储系统的配置。

具体实施方式

下文中，将参考附图详细解释本发明的优选实施例。

图 2A 示出了根据本发明的存储控制装置的原理。图 2A 中的存储控制装置 101 包括重建设备 111 和 112，并且通过执行以下控制实现了数据冗余，利用该控制，数据和奇偶校验位被分散地存储在多个存储设备 102-1 到 102-N 中。

当存储设备 102-1 发生故障时，重建设备 111 利用存储在除了存储设备 102-1 以外的其他存储设备中的信息，来恢复存储设备 102-1 中的信息，并且将恢复信息写入备用存储设备 103-1 中。当在恢复存储设备 102-1 中的信息的同时，存储设备 102-2 发生故障时，重建设备 112 利用存储在除了存储设备 102-1 和 102-2 以外的其他存储设备中的信息，来恢复存储在存储设备 102-1 的未恢复区域中的信息和存储设备 102-2 中的信息，并且将恢复的信息分别写入备用存储设备 103-1 的对应区域和备用存储设备 103-2 中。

在每个存储设备中，数据或奇偶校验位被存储为信息。利用存储在此刻无故障的正常运行的存储设备中的信息，来恢复故障存储设备中的信息，并且恢复的信息被写入与故障存储设备相对应的备用存储设备中。当在恢复存储设备 102-1 中的信息的同时，存储设备 102-2 发生故障时，存储设备 102-1 的恢复区域中的信息被不加改动地存储到备用存储设备 103-1 中，并且对于存储设备 102-1 的未恢复区域和存储设备 102-2 的整个区域执行恢复处理。

根据上述的重建控制，即使当一个盘故障的情形变为两个盘故障的情形时，也可以有效地利用已恢复的信息，而无需擦除。另外，存储设备 102-1 中的恢复区域不被包括为恢复目标，从而减少了重建处理所耗费的时间。另外，关于存储设备 102-2 中与存储设备 102-1 中的恢复区域相对应的区域，可以施加对一个盘故障的重建处理（其需要的处理开销更小），从而实现了更高的效率。

存储控制装置 101 例如对应于图 2B 中的控制器 211、图 21 中的主机总线适配器 1911 或图 22 中的主机设备 2001，这些将在后面解释。

根据本发明，在将数据和奇偶校验位分散地存储在多个存储设备内的存储装置中，改进了当两个存储设备发生故障时的重建处理，并且有效地恢复了数据冗余。

图 2B 示出了一个实施例中的存储系统的配置示例。图 2B 中的存储系统包括主机设备 201 和 RAID 设备 202。RAID 设备 202 对应于包括控制器 211 和四个盘#0 到盘#3 的存储装置。

每个盘包括一个或多个磁盘，并且主机设备 201 读/写数据，将每个盘识别为一个存储设备。然而注意，连接到控制器 211 的盘的数目并不限于四个，通常有多于四个的盘连接到控制器 211。

控制器 211 包括处理器 221、存储器 222 和缓存 223，并且在盘#0 到盘#3 发生故障时执行重建处理。处理器 221 执行存储在存储器 222 中的程序，从而利用缓存 223 作为数据缓冲器来执行重建处理。

在本实施例中，为了改进两个盘故障时的重建处理，采用了图 3 中所示的计数器方案 1 到 6。首先，参考图 4 到图 8 解释各个计数器方案的概况。

1. 分离执行方案（计数方案 1）

控制器 211 分离地执行对第一盘和第二盘的重建处理。尤其是，在对应于两个盘中故障区域的双重故障区域的重建处理中，只有在一个盘的已恢复数据/奇偶校验位被写入盘中，而另一个盘的已恢复数据/奇偶校验位被丢弃而未写入时，才在数据缓冲器上恢复出两个盘中的数据/奇偶校验位。

例如，解释了如图 4 所示的情形，其中盘#0 首先发生故障，在利用热备份 HS #0 恢复盘#0 中的数据/奇偶校验位的同时，盘#1 发生故障。这种情形下，利用正常运行的盘#1 到#3 中的数据/奇偶校验位，通过对一个盘故障的重建处理，来恢复盘#0 中的数据/奇偶校验位，直到盘#1 发生故障为止。

当盘#1 发生故障时，热备份 HS #0 的重建区域中的数据/奇偶校验位

按原样保存而不被丢弃，并且只有仍然未恢复的数据/奇偶校验位被两个盘故障的重建处理恢复。在这一过程中，利用正常运行的盘#2 到#3 中的数据/奇偶校验位，在热备份 HS #0 中恢复出仍然未恢复的数据/奇偶校验位。这时，同时创建盘#1 中的数据/奇偶校验位，然而，对盘#1 的重建处理是分离并且独立执行的，因此丢弃所创建的数据/奇偶校验位。

至于热备份 HS #1，与热备份 HS #0 的重建处理并行地，利用正常运行的盘#2 和#3 中的数据/奇偶校验位，通过两个盘故障的重建处理来恢复全部数据/奇偶校验位。这时，由两个盘故障的重建处理同时创建的盘#0 的数据/奇偶校验位被丢弃。

根据以上重建处理，当盘#1 发生故障时，盘#0 的重建处理从一个盘故障的重建处理切换到两个盘故障的重建处理，并且恢复区域中的数据/奇偶校验位按原样保存，其中，在对一个盘故障的重建处理中，只有被指定为恢复目标的盘发生故障，而在对两个盘故障的重建处理中，除了被指定为恢复目标的盘以外，还有另一个发生故障的盘。因此，不要求如图 1C 所示的恢复区域中的数据/奇偶校验位的第二次恢复，从而盘#0 的恢复完成的时间要比如图 1C 所示的恢复时间短。

2. 等待进度位置相互对应的方案（计数方案 2）

当第二盘发生故障时，控制器 211 临时暂停对第一盘的重建处理，并且只重建第二盘，直到第二盘中的进度位置对应于第一盘中的进度位置为止。从实现上述对应时开始，同时重建这两个盘。

例如，解释如图 5 所示的情形，其中盘#0 首先发生故障，在利用热备份 HS #0 恢复盘#0 中的数据/奇偶校验位的同时，盘#1 发生故障。直到盘#1 发生故障以前的操作与图 4 相同。

当盘#1 发生故障时，最初利用热备份 HS #0 的重建区域中的数据/奇偶校验位和正常运行的盘#2 和#3 中的数据/奇偶校验位，通过对一个盘故障的重建处理，只恢复与热备份 HS #0 中的重建区域相对应的热备份 HS #1 中的数据/奇偶校验位。当完成以上热备份 HS #1 中的数据/奇偶校验位的恢复时，利用正常运行的盘#2 和#3 中的数据/奇偶校验位，通过对两个盘故障的重建处理，同时恢复分别在热备份 HS #0 和热备份 HS #1 中的剩

余数据/奇偶校验位。

根据以上重建处理，除了获得与计数方案 1 相同的优点之外，还获得了以下优点：当前正恢复的盘#0 的恢复区域中的数据/奇偶校验位被有效地用于盘#1 的恢复，从而可以更有效地执行对盘#1 的重建处理。

3. 组合方案（计数方案 3）

该方案组合了计数方案 1 和 2。控制器 211 在重建处理期间，通过监视两个盘 HS #0 和 HS #1 的重建处理的进度位置，来选择恢复方法。具体地说，控制器 211 检查分别在盘 HS #0 和盘 HS #1 中的进度位置之间的差别，并且当差别等于或大于规定阈值时，应用计数方案 1 从而给予对盘 HS #0 的重建处理以优先级。当差别小于上述阈值时，应用计数方案 2 从而执行对盘 HS #1 的重建处理，直到两个进度位置彼此对应为止。每次恢复规定单元中的数据/奇偶校验位或者每次恢复规定区域中的数据/奇偶校验位时，执行该进度位置的检查。

当应用计数方案 2 时，热备份 HS #0 和 HS #1 中进度位置之间的差别对应于开始恢复冗余丢失的双重故障区域的等待时间，并且进度位置的差别越大，冗余恢复的延迟越大。因而，当上述等待时间长于规定的时间段时，应用计数方案 1 以避免上述延迟太长，从而并行执行对热备份 HS #0 和 HS #1 的重建处理。

然而，在该方案中，与计数方案 1 的区别在于：利用热备份 HS #0 的重建区域中的数据/奇偶校验位和正常运行的盘#2 和#3 中的数据/奇偶校验位，通过对一个盘故障的重建处理，来恢复不在热备份 HS #1 的双重故障区域中的热备份 HS #1 的数据/奇偶校验位。通常，在对两个盘故障的重建处理中，需要比一个盘故障的重建处理更多的计算量。因此，很有可能针对热备份 HS #1 执行的一个盘故障的重建处理要比前述针对热备份 HS #0 执行的两个盘故障的重建处理进行得更快，从而使得进度位置的差别随时间而变小。

根据以上重建处理，除了获得与计数方案 2 相同的优点之外，还获得了以下优点：可以避免由于等待恢复开始而引起的冗余恢复的延迟。另外，通过并行处理可以实现处理器资源等的有效利用。

4. 双重故障区域被优先恢复的方案（计数方案 4）

控制器 211 在第二盘发生故障时将第一盘的当前进度位置保存为恢复完成位置（停止位置，将在后面描述），并且从以上位置同时对两个盘执行重建处理。当以上针对两个盘同时执行的重建处理完成时，从起始端执行重建处理，直到第二盘中的恢复完成位置，并将数据恢复到第二盘的未重建区域中。

例如，解释如图 6 所示的情形，其中盘#0 首先发生故障，在利用热备份 HS #0 恢复盘#0 中的数据/奇偶校验位的同时，盘#1 发生故障。直到盘#1 发生故障以前的操作与图 4 相同。

当盘#1 发生故障时，利用正常运行的盘#2 和#3 中的数据/奇偶校验位，通过对两个盘故障的重建处理，来恢复与热备份 HS #0 中的未重建区域相对应的热备份 HS #1 中的数据/奇偶校验位和热备份 HS #0 中的数据/奇偶校验位。当恢复了以上对应于未重建区域的数据/奇偶校验位时，利用热备份 HS #0 的重建区域中的数据/奇偶校验位和正常运行的盘#2 和#3 中的数据/奇偶校验位，通过对一个盘故障的恢复处理，来恢复热备份 HS #1 中的剩余数据/奇偶校验位。

当在重建处理期间存在对 RAID 6 的正常读/写访问时，检查访问目标数据是否被重建，或者所有的正常读/写访问被简并（degenerate），例如重建数据并且同时针对读请求返回重建数据。

根据以上重建处理，除了获得与计数方案 2 相同的优点之外，还获得了以下优点：通过优先恢复双重故障区域中的数据/奇偶校验位，可以在更短的时间内恢复 RAID 组的冗余。

5. 并行执行对双重故障区域的恢复和使用已恢复的热备份 HS 的恢复的方案（计数方案 5）

控制器 211 无需等待，与对双重故障区域的重建处理并行地，执行从起始端到第二盘中的恢复完成位置的重建处理（在计数方案 4 中稍后执行的处理）。

例如，解释如图 7 所示的情形，其中盘#0 首先发生故障，在利用热备份 HS #0 恢复盘#0 中的数据/奇偶校验位的同时，盘#1 发生故障。直到盘

#1 发生故障以前的操作与图 4 相同。

当盘#1 发生故障时，并行执行恢复热备份#0 和#1 的双重故障区域中的数据/奇偶校验位的处理（称为“前一个处理”）和恢复热备份 HS #1 的剩余区域中的数据/奇偶校验位的处理（称为“后一个处理”），其中，前一个处理是利用正常运行的盘#2 和#3 中的数据/奇偶校验位，通过对两个盘故障的重建处理来执行的，后一个处理是利用热备份#0 的重建区域中的数据/奇偶校验位和正常运行的盘#2 和#3 中的数据/奇偶校验位，通过对一个盘故障的重建处理来执行的。

类似于计数方案 4，当在重建处理期间存在正常的读/写访问时，检查访问目标数据是否被重建，或者访问被简并。

根据以上重建处理，RAID 组的冗余恢复需要比计数方案 4 更长的时间，然而，由于恢复冗余丢失的双重故障区域和冗余保留的其他区域的处理是并行执行的，所以减少了整体恢复的时间。

6. 随机恢复方案（计数方案 6）

控制器 211 利用缓存 223 上的位图对盘上的每个规定区域执行重建处理。采用以上的计数方案 1 到 5 其中之一的顺序作为重建顺序。从数据/奇偶校验位保存的角度出发，基本上希望采用计数方案 4 的顺序，然而，也可以采用其他计数方案的顺序来实现该方案。

另外，当重建操作作为正常读/写访问中的一个处理来执行时，对应区域被重建的事实被记录在位图中。因此，同时采用独立于读/写访问的顺序重建处理以及作为读/写访问中的处理之一的点对点重建处理作为重建处理。在点对点的重建处理中，当响应于读请求恢复数据时，控制器 211 将恢复数据写入盘的对应位置中。

控制器 211 准备例如一个条带对应于一位的位图并管理进度。如果位图丢失，则从头开始重建处理。也可以采用一个逻辑块对应于一位的位图。

另外，当起初无法获得位图的存储区域时，控制器 211 执行下面的操作之一。

- 通过采用不同于本方案的以上计数方案，来执行重建处理。

• 设置位图尺寸的上限，并且当需要超过设置尺寸的位图时，在获得资源后执行重建处理。

控制器 211 可并行执行重建操作，并且具有位图备份/恢复以及接通/断开电源的功能。当控制器 211 冗余（双份）时，在控制器之间也基本上备份有位图，然而，即使在没有备份位图时，数据也不会丢失。如前所述，当位图丢失时，重新开始重建处理。

例如，解释如图 8 所示的情形，其中盘#0 首先发生故障，在利用热备份 HS #0 恢复盘#0 中的数据/奇偶校验位的同时，盘#1 发生故障。

控制器 211 为每个热备份 HS #0 和 HS #1 创建位图，并且将每个条带作为一位数据来管理。对于未重建区域中的条带，记录“1”作为对应位，而对于重建区域中的条带则记录“0”。对于热备份 HS #0 的重建区域中的条带，每次在盘#1 发生故障时记录“0”作为对应位，而对于其他条带，每次在重建对应区域时记录“0”。

根据以上重建处理，除了获得与计数方案 1 到 5 相同的优点之外，还获得了以下优点：通过将由点对点重建处理恢复的数据/奇偶校验位写回到盘中，并将以上数据/奇偶校验位识别为恢复的数据/奇偶校验位，有效地执行了处理。

下面将参考图 9 到 19 详细解释以上的计数方案。

每个计数方案中的重建处理由盘故障触发，或者由其他重建处理触发。控制器 211 将触发盘指定为待处理的主盘，并且在需要的时候将另一个故障盘增加为待处理的副盘。在重建处理中，关于在待处理主盘和待处理副盘之间公共的当前位置执行恢复处理，如图 9 所示。“当前位置”是指当前执行恢复处理的位置。

另外，控制器 211 将与构成 RAID 设备 202 的所有盘有关的信息在缓存 223 中保存为所有计数方案之间共有的控制信息。具体地说，如图 10 所示，关于每个盘保存诸如恢复状态（已恢复、正在恢复和未恢复）、恢复开始位置和停止位置（如果需要的话）之类的信息。至于正常运行的盘，所有区域被设为已恢复。

故障盘的恢复开始位置对应于对被指定为待处理主盘的故障盘执行的

恢复处理的当前位置，并且当不执行恢复处理时，恢复开始位置被设为盘的末端（图 9 中盘的底端）。恢复处理从恢复开始位置沿向上的方向进行。停止位置是指恢复处理不得不停止的位置。

对于每个以诸如条带、逻辑块等为单位的区域来管理恢复状态。利用单位区域的地址或单位区域所属的条带的标识符，来管理诸如当前位置、恢复开始位置、停止位置等位置信息。

图 11 是重建处理的流程图。控制器 211 首先将待处理主盘的末端（图 9 中盘的底端）设为当前位置（步骤 1101），并通过执行恢复例程来执行恢复处理（步骤 1102）。在这一恢复例程中，利用其他盘中的数据/奇偶校验位，来创建被指定为恢复目标的盘中被指定为恢复目标的数据/奇偶校验位，并且将已恢复的数据/奇偶校验位写入对应的热备份中。一旦执行了恢复处理，则恢复出以诸如逻辑块、条带等为规定单位的数据/奇偶校验位。通常，以条带作为规定单位来执行恢复。

随后执行重建控制的恢复后处理（步骤 1103）。在恢复后处理中，进行恢复开始位置的设置、恢复处理是否要结束的判断等等操作。其后，检查在以上恢复后处理中，是否确定恢复处理要结束（步骤 1104）。当确定恢复处理不是要结束时，在当前位置前进刚好一个条带之后（步骤 1105），重复步骤 1102 和后续步骤的处理，并且当确定恢复处理要结束时，重建处理结束。

在步骤 1102 的恢复例程和步骤 1103 的恢复后处理中，在多个计数方案中执行不同的操作，因此，将以从计数方案 1 到计数方案 6 的顺序给出其解释。

1. 计数方案 1

在计数方案 1 中，图 11 中的重建处理由每个盘中的故障触发，并且触发盘被指定为待处理主盘。然而，并不将另一个故障盘加为待处理副盘。因此，在第二盘发生故障之后，并行执行两个重建处理，并且只有待处理主盘是在每个重建处理中被指定为恢复目标的盘。

图 12 是计数方案 1 的恢复例程的流程图。控制器 211 首先将待处理主盘设为恢复目标盘，在盘中设置当前位置（步骤 1201），并检查故障盘的

数目（步骤 1202）。当故障盘的数目为 1 时，确定采用一个盘故障的重建处理作为恢复方法（步骤 1203）。

随后，在属于当前位置处的条带的数据/奇偶校验位当中，从正常运行的盘中读取一个盘故障的重建处理所需的数据/奇偶校验位（步骤 1205），并且检查是否已读取所有上述所需的数据/奇偶校验位（步骤 1206）。当已读取所有上述的数据/奇偶校验位时，利用读取的数据/奇偶校验位，来恢复属于相同条带的恢复目标盘中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入对应的热备份中（步骤 1207）。

当在步骤 1206 中发生读错误时，其被识别为被指定为读目标的盘发生故障。然后，检查故障盘的数目（步骤 1208），并且当上述数目为 2 时，执行步骤 1202 和后续步骤的处理。

然后，恢复方法切换到对两个盘故障的重建处理（步骤 1204），并且从正常运行的盘中读取两个盘故障的重建处理所需的数据/奇偶校验位（步骤 1205）。当已读取所有上述数据/奇偶校验位时，利用读取的数据/奇偶校验位对恢复目标盘中的数据/奇偶校验位进行恢复，并且将恢复的数据/奇偶校验位写入对应的热备份中（步骤 1207）。

当除了两个故障盘外，发生另一个读错误时，被识别为故障盘的盘的数目变为 3（步骤 1208），从而确定恢复已不可能，因此执行错误处理（步骤 1209）。

图 13 是计数方案 1 中恢复后处理的流程图。控制器 211 首先将恢复例程中所用的当前位置设为待处理主盘中的恢复开始位置（步骤 1301），并且检查是否已完成对待处理主盘中整个区域的恢复（步骤 1302）。在该示例中，当恢复例程中所用的当前位置到达待处理主盘的末端（图 9 中盘的顶端）时，则确定已完成对整个区域的恢复。当未完成对整个区域的恢复时，则确定恢复处理必须继续（步骤 1303），并且当已完成对整个区域的恢复时，则确定恢复处理将结束（步骤 1304）。

例如，当盘#0 首先发生故障时，对被指定为待处理主盘的盘#0 触发重建处理，如图 4 所示。此时，故障盘的数目为 1（图 12 中步骤 1202），从而采用对一个盘故障的重建处理（步骤 1203）。因此，在正常

运行的盘#1 到#3 的数据/奇偶校验位中，读取经由对一个盘故障的重建处理来恢复盘#0 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤1205）。然后，利用读取的数据/奇偶校验位来恢复盘#0 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入热备份 HS #0 中（步骤1207）。

盘#0 中的当前位置被设为盘#0 中的恢复开始位置（图 13 中步骤1301），并且确定要继续恢复处理（步骤 1303）。诸如对盘#0 的读/写访问等的其他处理参考所述恢复开始位置。对于每一个条带，反复执行以上的恢复例程和恢复后处理（图 11 中步骤 1105）。

随后，当盘#1 发生故障时，故障盘的数目变为 2（步骤 1202），从而恢复方法切换到对两个盘故障的重建处理（步骤 1204），并且从正常运行的盘#2 和#3 中读取经由对两个盘故障的重建处理来恢复盘#0 和#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1205）。然后，利用读取的数据/奇偶校验位来恢复盘#0 和#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位当中的盘#0 中的数据/奇偶校验位写入热备份#0 中（步骤1207）。

执行与盘#1 中发生故障之前相同的恢复后处理。对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105），并且当已完成对盘#0 中整个区域的恢复时（步骤 1304），确定盘#0 的重建处理结束（步骤1104）。

另外，在盘#1 发生故障时，对被指定为待处理副盘的盘#1 执行另一个重建处理。这时，故障盘的数目为 2（步骤 1202），从而采用对两个盘故障的重建处理（步骤 1204），并且从正常运行的盘#2 和#3 中读取经由对两个盘故障的重建处理来恢复盘#0 和#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1205）。

然后，利用读取的数据/奇偶校验位来恢复盘#0 和#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位当中的盘#1 中的数据/奇偶校验位写入热备份#1 中（步骤 1207）。

执行与盘#0 中发生故障之前相同的恢复后处理。对于每一个条带，反

复执行以上的恢复例程和恢复后处理（步骤 1105），并且当已完成对盘#0中整个区域的恢复时（步骤 1304），确定对盘#1 的重建处理结束（步骤1104）。

2. 计数方案 2

在计数方案 2 中，类似于计数方案 1，图 11 中的重建处理由每个盘的故障触发，并且触发盘被指定为待处理主盘。当第二盘发生故障时，中止对被指定为待处理主盘的第一故障盘的重建处理，并且开始对被指定为待处理主盘的第二故障盘的重建处理。当第二故障盘中的当前位置到达与第一故障盘中相同的进度位置时，第一故障盘被加为待处理副盘。

图 14 是计数方案 2 的恢复例程的流程图。这种情况下，与图 12 中的恢复例程不相似的是，恢复目标盘被表示为待处理主盘和待处理副盘，并且至多两个盘可被设为恢复目标盘。另外，根据恢复目标盘的数目，而不是故障盘的数目来选择恢复方法。

控制器 211 首先设置被指定为恢复目标盘的待处理主盘/待处理副盘，在待处理主盘中设置当前位置（步骤 1401）。当不设置待处理副盘时，只有待处理主盘被设为恢复目标盘。

随后，控制器 211 检查恢复目标盘的数目（步骤 1402）。当恢复目标盘的数目为 1 时，确定采用一个盘故障的重建处理作为恢复方法（步骤1403）。从正常运行的盘中读取在属于当前位置的条带的数据/奇偶校验位中为一个盘故障的重建处理所需的数据/奇偶校验位（步骤 1405），并且检查是否已读取所有上述的数据/奇偶校验位（步骤 1406）。当已读取所有上述的数据/奇偶校验位时，利用读取的数据/奇偶校验位来恢复属于相同条带的恢复目标盘中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入对应的热备份中（步骤 1407）。

在步骤 1406 中发生读错误时，其被识别为被指定为读目标的盘发生故障。然后检查故障盘的数目（步骤 1408），当以上数目为 1 时，故障盘被加为恢复目标盘（步骤 1410），并且执行步骤 1402 和后续步骤的处理。

然后，恢复方法切换到对两个盘故障的重建处理（步骤 1404），并且

从正常运行的盘中读取两个盘故障的重建处理所需的数据/奇偶校验位（步骤 1405）。当已读取所有上述的数据/奇偶校验位时，利用读取的数据/奇偶校验位来恢复恢复目标盘中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位分别写入对应的热备份中（步骤 1407）。

当除了两个故障盘之外，发生另一个读错误时，被识别为故障盘的盘的数目变为 3（步骤 1408），从而确定不可能恢复，因此执行错误处理（步骤 1409）。

图 15 是计数方案 2 中恢复后处理的流程图。控制器 211 首先将恢复例程结束时待处理主盘中的当前位置设为待处理主盘/待处理副盘中的恢复开始位置（步骤 1501），并且检查是否满足下面的条件 a（步骤 1502）。

条件 a：除了待处理主盘外，存在另一个故障盘，在待处理主盘和另一个故障盘中都没有设置停止位置，并且另一个故障盘中的恢复开始位置比待处理主盘中的要靠后（低）。

以上另一个故障盘中的恢复开始位置比待处理主盘中的要靠后这一事实意味着在另一个故障盘中的恢复处理的执行与待处理主盘中的相比有一个延迟。当条件 a 满足时，待处理主盘中的恢复开始位置被设为另一个故障盘中的停止位置（步骤 1506），并且确定恢复处理将结束，以中止对待处理主盘的重建处理（步骤 1508）。

另外当条件 a 不满足时，检查是否已完成对待处理主盘中整个区域的恢复（步骤 1503）。当已完成对整个区域的恢复时，确定恢复处理将结束（步骤 1508）。

当对整个区域的恢复未完成时，检查在待处理主盘中是否设置了停止位置，同时待处理主盘中的当前位置是否对应于以上停止位置（步骤 1504）。当当前位置对应于停止位置时，另一个故障盘被加为待处理副盘（步骤 1507），并且确定恢复处理将继续（步骤 1505）。

当当前位置并不对应于停止位置并且未设置停止位置时，确定恢复处理将按原样继续（步骤 1505）。

例如，如图 5 所示，当盘#0 首先发生故障时，对作为待处理主盘的盘#0 触发重建处理。因而，由于恢复目标盘的数目为 1（图 14 中步骤

1401），所以采用对一个盘故障的重建处理（图 14 中步骤 1403），并且从正常运行的盘#1 到#3 中读取经由对一个盘故障的重建处理来恢复盘#0 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。利用读取的数据/奇偶校验位来恢复盘#0 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入热备份 HS #0 中（步骤 1407）。

盘#0 中的当前位置被设为盘#0 中的恢复开始位置（图 15 中步骤 1501），并且由于不存在另一个故障盘（步骤 1502），所以确定继续恢复处理（步骤 1505）。对于每一个条带，反复执行以上的恢复例程和恢复后处理（图 11 中步骤 1105）。

随后，当盘#1 发生故障时，条件 a 满足（步骤 1502），这是因为被指定为待处理主盘的盘#0 中的恢复开始位置对应于盘#0 中的当前位置（步骤 1501），并且作为另一个故障盘的盘#1 中的恢复开始位置对应于盘#1 的底端。然后，盘#0 中的恢复开始位置被设为盘#1 中的停止位置（步骤 1506），并且确定恢复处理将结束（步骤 1508）。从而中止对被指定为待处理主盘的盘#0 的重建处理（步骤 1104）。

此后，触发对被指定为待处理主盘的盘#1 的另一个重建处理。由于恢复目标盘的数目为 1（步骤 1401），所以采用对一个盘故障的重建处理（步骤 1403），并且从正常运行的盘#0、#2 和#3 中读取经由对一个盘故障的重建处理来恢复盘#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。然而，关于盘#0，只读取已被写入热备份 HS #0 中的已恢复数据/奇偶校验位。

然后，利用读取的数据/奇偶校验位来恢复盘#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入热备份 HS #1 中（步骤 1407）。

盘#1 中的当前位置被设为盘#1 中的恢复开始位置（步骤 1501），并且由于在盘#1 中已经设置了停止位置，所以不满足条件 a（步骤 1502）。另外，由于盘#1 中的当前位置还未到达盘#1 中的停止位置（步骤 1504），所以确定继续恢复处理（步骤 1505）。对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105）。

当盘#1 中的当前位置到达盘#1 中的停止位置时（步骤 1504），将恢

复处理被中止的盘#0 加为待处理副盘（步骤 1507），并且确定继续恢复处理（步骤 1505）。从而更新当前位置（步骤 1105）。

从而，恢复目标盘的数目变为 2（步骤 1404），恢复方法切换到对两个盘故障的重建处理（步骤 1404），并且从正常运行的盘#2 和#3 中读取经由对两个盘故障的重建处理来恢复盘#0 和#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。利用读取的数据/奇偶校验位来恢复盘#0 和#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位分别写入热备份 HS #0 和 HS #1 中（步骤 1407）。

盘#1 中的当前位置被设为盘#1 中的恢复开始位置（步骤 1501），并且由于盘#1 中的当前位置已超过盘#1 中的停止位置（步骤 1504），所以确定继续恢复处理（步骤 1505）。

对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105）。当已完成对盘#1 中整个区域的恢复时（步骤 1508），被指定为待处理主盘的盘#1 的重建处理结束（步骤 1104）。另外，由于在以上时刻作为待处理副盘的盘#0 中的当前位置已到达盘#0 的顶端，所以对盘#0 的恢复也一同结束。

3. 计数方案 3

在计数方案 3 中，类似于计数方案 2，图 11 中的重建处理由每个盘的故障触发，并且触发盘被指定为待处理主盘。当第二盘发生故障时，根据两个故障盘中各自的进度位置之间的差别来选择计数方案 1 或计数方案 2。

当进度位置之间的差别等于或大于阈值时，选择计数方案 1 从而并行执行两个重建处理。然而，在该方案中，与计数方案 1 的区别在于，利用第一故障盘的重建区域中的数据/奇偶校验位和正常运行的盘中的数据/奇偶校验位，通过对一个盘故障的重建处理来恢复第二故障盘中的数据/奇偶校验位。

当进度位置之间的差别变得小于上述阈值时，选择计数方案 2 从而中止对被指定为待处理主盘的第一故障盘的重建处理，并且开始被指定为待处理主盘的第二故障盘的重建处理。并且当第二故障盘中的当前位置到达

与第一故障盘中相同的进度位置时，第一故障盘被加为待处理副盘。

计数方案 3 中恢复例程的流程图与计数方案 2 类似，并且图 16 示出了恢复后处理的流程图。图 16 中的恢复后处理采用的配置除了包括图 15 中的恢复后处理外，还包括步骤 1603 的判决步骤。

在步骤 1603 中，控制器 211 将分别在另一个故障盘和待处理主盘中的恢复开始位置之间的差别与阈值相比较。当恢复开始位置之间的差别小于阈值时，待处理主盘中的恢复开始位置被设为另一个故障盘中的停止位置（步骤 1607），并且确定恢复处理将结束（步骤 1609）。另外，当以上恢复开始位置之间的差别等于或大于阈值时，执行步骤 1604 和后续步骤的处理。

因此，当计数方案 2 中所描述的条件满足并且分别在另一个故障盘和待处理主盘中的恢复开始位置之间的差别小于阈值时，在另一个故障盘中设置停止位置（步骤 1607）。在不同于以上情形的其他情形中，不设置停止位置。

当例如图 5 所示盘#0 首先发生故障时，触发对被指定为待处理主盘的盘#0 的重建处理，并且执行与计数方案 2 中相似的处理，直到盘#1 发生故障。

随后，当盘#1 发生故障时，条件 a 满足（图 16 中步骤 1602），从而将分别在盘#0 和盘#1 中的恢复位置之间的差别与阈值相比较（步骤 1603）。这时，如果自盘#0 发生故障开始已经过去足够长的时间，则认为对盘#0 的恢复处理已进行的程度足以使恢复开始位置之间的差别超过阈值。这种情况下，执行步骤 1604 和后续步骤的处理，并且由于在盘#0 中没有设置停止位置（步骤 1605），所以确定继续恢复处理（步骤 1606）。

这时，由于恢复目标盘的数目为 1（图 14 中步骤 1401），所以采用对一个盘故障的重建处理（步骤 1403），并且从正常运行的盘#1 到#3 中读取经由对一个盘故障的重建处理来恢复盘#0 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。

这时，当读取盘#1 中的数据/奇偶校验位时，发生读错误，并且盘#1

被加为恢复目标盘（步骤 1410），从而恢复目标盘的数目变为 2。

因此，恢复方法切换到对两个盘故障的重建处理（步骤 1404），并且从正常运行的盘#2 和#3 中读取经由对两个盘故障的重建处理来恢复盘#0 和#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。利用读取的数据/奇偶校验位来恢复盘#0 和#1 中的数据/奇偶校验位，并且将以上恢复的数据/奇偶校验位当中的盘#0 中的恢复数据/奇偶校验位写入热备份 HS #0 中（步骤 1407）。

盘#0 中的当前位置被设为盘#0 中的恢复开始位置（步骤 1601），并且条件 a 满足（步骤 1602），然而，恢复开始位置之间的差别仍然大于阈值（步骤 1603）。另外，由于在盘#0 中没有设置停止位置（步骤 1605），所以确定继续恢复处理（步骤 1606）。对于每一个条带，反复执行以上的恢复例程和恢复后处理（图 11 中步骤 1105）。

另外，当盘#1 发生故障时，触发对被指定为待处理主盘的盘#1 的另一个重建处理。由于恢复目标盘的数目为 1（步骤 1401），所以采用对一个盘故障的重建处理（步骤 1403），并且从正常运行的盘#0、#2 和#3 中读取经由对一个盘故障的重建处理来恢复盘#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。然而，关于盘#0，只读取被写入热备份 HS #0 中的恢复数据/奇偶校验位。

然后，利用读取的数据/奇偶校验位来恢复盘#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入热备份 HS #1 中（步骤 1407）。

盘#1 中的当前位置被设为盘#1 中的恢复开始位置（步骤 1601），并且由于被指定为另一个故障盘的盘#0 中的恢复开始位置已超过了被指定为待处理主盘的盘#1 中的恢复开始位置，所以条件 a 不满足（步骤 1602）。另外，由于在盘#1 中没有设置停止位置（步骤 1605），所以确定继续恢复处理（步骤 1606）。对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105）。

从而，并行执行对被指定为待处理主盘的盘#0 的两个盘故障的重建处理以及对被指定为待处理主盘的盘#1 的一个盘故障的重建处理，使得盘#1 中的恢复开始位置逐渐接近盘#0 中的恢复开始位置。

当在对被指定为待处理主盘的盘#0 的重建处理中，分别在盘#0 和盘#1 中的恢复开始位置之间的差别变得小于阈值时（步骤 1603），盘#0 中的恢复开始位置被设为盘#1 中的停止位置（步骤 1607），并且确定恢复处理将结束（步骤 1609）。从而，中止对被指定为待处理主盘的盘#0 的重建处理（步骤 1104）。

其后，只继续对被指定为待处理主盘的盘#1 的重建处理，然而，由于在盘#1 中已经设置了停止位置，所以条件 a 不满足（步骤 1602）。另外，由于盘#1 中的当前位置还未到达盘#1 中的停止位置（步骤 1605），所以确定继续恢复处理（步骤 1606）。

当盘#1 中的当前位置到达盘#1 中的停止位置时（步骤 1605），将恢复处理被中止的盘#0 加为待处理副盘（步骤 1608），并且确定继续恢复处理（步骤 1606）。从而，更新当前位置（步骤 1105）。

从而，恢复目标盘的数目变为 2（步骤 1401），因而采用对两个盘故障的重建处理（步骤 1404），并且从正常运行的盘#2 和#3 中读取经由对两个盘故障的重建处理来恢复盘#0 和盘#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。利用读取的数据/奇偶校验位来恢复盘#0 和#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位分别写入热备份 HS #0 和 HS #1 中（步骤 1407）。

盘#1 中的当前位置被设置为盘#0 和#1 中的恢复开始位置（步骤 1601），并且盘#1 中的当前位置已超过盘#1 中的停止位置（步骤 1605），因此，确定继续恢复处理（步骤 1606）。

对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105）。并且当已完成对盘#1 中整个区域的恢复时（步骤 1609），被指定为待处理主盘的盘#1 的重建处理结束（步骤 1104）。另外，由于在以上时刻被指定为待处理副盘的盘#0 中的当前位置已到达盘#0 的顶端，所以对盘#0 的恢复也一同结束。

4. 计数方案 4

在计数方案 4 中，图 11 中的重建处理由盘中的故障触发，或者由其他重建处理触发。并且触发盘被指定为待处理主盘。通过前一个触发，对

于每一个 RAID 组只触发一个重建处理。因此，如果已经触发 RAID 组的重建处理，则即使第二盘发生故障，也不会触发另一个重建处理。

当第二盘发生故障时，第一故障盘中的当前位置被设置为第二故障盘中的停止位置，第二故障盘被加为待处理副盘，并且继续重建处理。并且当已完成对第一故障盘的恢复时，从底端开始，执行对被指定为待处理主盘的第二故障盘的重建处理，直到第二故障盘中的停止位置。计数方案 4 中的恢复例程的流程图与计数方案 2 中的相类似。

图 17 是计数方案 4 中恢复后处理的流程图。控制器 211 首先在恢复例程结束时将待处理主盘中的当前位置设为待处理主盘/待处理副盘中的恢复开始位置（步骤 1701），并且检查是否满足下面的条件 b（步骤 1702）。

条件 b：除了待处理主盘外，还存在另一个故障盘，并且在待处理主盘和另一个故障盘中都没有设置停止位置。

当条件 b 满足时，待处理主盘中的恢复开始位置被设为另一个故障盘中的停止位置，并且另一个故障盘被加为待处理副盘（步骤 1706）。然后，检查是否已完成对待处理主盘中整个区域的恢复（步骤 1703）。当条件 b 不满足时，按原样执行步骤 1703 中的处理。

当已完成对待处理主盘中整个区域的恢复时，检查是否存在另一个故障盘。当存在另一个故障盘时，触发对被指定为待处理主盘的另一个故障盘的另一个重建处理（步骤 1707）。并且确定恢复处理将结束（步骤 1708）。当不存在另一个故障盘时，确定重建处理将结束，而不触发另一个重建处理（步骤 1708）。

当未完成对整个区域的恢复时，检查待处理主盘中的当前位置是否对应于待处理主盘中的停止位置（步骤 1704）。当当前位置对应于停止位置时，则确定恢复处理结束（步骤 1708）。

当上述当前位置不对应于停止位置并且未设置停止位置时，则确定恢复处理将按原样继续（步骤 1705）。

例如，如图 6 所示，当盘#0 首先发生故障时，触发对被指定为待处理主盘的盘#0 的重建处理。这时，由于恢复目标盘的数目为 1（图 14 中步骤 1401），所以采用对一个盘故障的重建处理（步骤 1403），并且从正

常运行的盘#1 到#3 中读取经由对一个盘故障的重建处理来恢复盘#0 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。利用读取的数据/奇偶校验位来恢复盘#0 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入热备份 HS #0 中（步骤 1407）。

盘#0 中的当前位置被设为盘#0 中的恢复开始位置（图 17 中步骤 1701），并且由于不存在 另一个故障盘（步骤 1702），所以确定继续恢复处理（步骤 1705）。对于每一个条带，反复执行以上的恢复例程和恢复后处理（图 11 中步骤 1105）。

随后，当盘#1 发生故障时，条件 b 满足（步骤 1702），盘#0 中的恢复开始位置被设为盘#1 中的停止位置，并且盘#1 被加为待处理副盘（步骤 1706）。然而，由于在盘#0 中没有设置停止位置（步骤 1704），所以确定继续重建处理（步骤 1705）。

这时，恢复目标盘的数目变为 2（步骤 1401），从而恢复方法切换到对两个盘故障的重建处理（步骤 1404），并且从正常运行的盘#2 和#3 中读取经由两个盘故障的重建处理来恢复盘#0 和#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。利用读取的数据/奇偶校验位来恢复盘#0 和#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位分别写入热备份 HS #0 和 HS #1 中（步骤 1407）。

盘#0 中的当前位置被设为盘#0 和盘#1 中的恢复开始位置（步骤 1701），并且由于在盘#1 中已经设置了停止位置，所以条件 b 不满足（步骤 1702）。另外，由于在盘#0 中没有设置停止位置（步骤 1704），所以确定继续重建处理（步骤 1705）。

对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105）。并且当已完成对盘#0 中整个区域的恢复时（步骤 1703），触发对被指定为待处理主盘的盘#1 的另一个重建处理（步骤 1707），并且确定恢复处理将结束（步骤 1708）。从而，对被指定为待处理主盘的盘#0 的重建处理结束（步骤 1104）。在以上时刻，被指定为待处理副盘的盘#1 的当前位置已到达盘#1 的顶端。

随后，在对被指定为待处理主盘的盘#1 的重建处理中，盘#1 的底端

被设为当前位置（步骤 1101）。这时，由于恢复目标盘的数目为 1（步骤 1401），所以采用对一个盘故障的重建处理（步骤 1403），并且从正常运行的盘#0、#2 和#3 中读取经由一个盘故障的重建处理来恢复盘#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。然而，关于盘#0，只读取已被写入热备份 HS #0 中的恢复数据/奇偶校验位。

然后，利用读取的数据/奇偶校验位来恢复盘#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入热备份 HS #1 中（步骤 1407）。

盘#1 中的当前位置被设置为盘#1 中的恢复开始位置（步骤 1701），并且在盘#1 中已经设置了停止位置，因此条件 b 不满足（步骤 1702）。另外，由于盘#1 中的当前位置还未到达盘#1 中的停止位置（步骤 1704），所以确定继续恢复处理（步骤 1705）。

对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105），并且盘#1 中的当前位置到达停止位置。此时，盘#1 中的当前位置还未到达盘#1 的顶端，从而确定未完成对整个区域的恢复（步骤 1703）。然而，当前位置对应于停止位置（步骤 1704），确定恢复处理将结束（步骤 1708）。从而，对被指定为待处理主盘的盘#1 的重建处理结束（步骤 1104），并且对盘#1 的恢复已完成。

5. 计数方案 5

在计数方案 5 中，类似于计数方案 1，图 11 中的重建处理由每个盘的故障触发，并且触发盘被指定为待处理主盘。

类似于计数方案 4，当第二盘发生故障时，第一故障盘中的当前位置被设置为第二故障盘中的停止位置，第二故障盘被加为待处理副盘，并且继续重建处理。同时，触发对被指定为待处理主盘的第二故障盘的重建处理，以与被指定为待处理主盘的第一故障盘的重建处理并行执行。

计数方案 5 中恢复例程的流程图与计数方案 2 中的相类似，并且图 18 示出了恢复后处理的流程图。图 18 中的恢复后处理所采用的配置是从图 17 中的恢复后处理中除去了步骤 1707 的处理。

当如图 7 所示，盘#0 首先发生故障时，触发对被指定为待处理主盘的盘#0 的重建处理，并且执行与计数方案 4 中相类似的处理，直到盘#1 发生

故障。

随后，当盘#1 发生故障时，条件 b 满足（步骤 1802），盘#0 中的恢复开始位置被设为盘#1 中的停止位置，并且盘#1 被加为待处理副盘（步骤 1806）。然而，由于在盘#0 中没有设置停止位置（步骤 1804），所以确定继续恢复处理（步骤 1805）。

这时，恢复目标盘的数目变为 2（步骤 1401），从而恢复方法切换到对两个盘故障的重建处理（步骤 1404），并且从正常运行的盘#2 和#3 中读取经由两个盘故障的重建处理来恢复盘#0 和#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。利用读取的数据/奇偶校验位来恢复盘#0 和#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位分别写入热备份 HS #0 和 HS #1 中（步骤 1407）。

盘#0 中的当前位置被设置为盘#0 和盘#1 中的恢复开始位置（步骤 1801），并且由于在盘#1 中已经设置了停止位置，因此条件 b 不满足（步骤 1802）。另外，由于在盘#0 中没有设置停止位置（步骤 1804），所以确定继续恢复处理（步骤 1805）。

对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105），并且当已完成对盘#0 中整个区域的恢复时（步骤 1803），确定恢复处理结束（步骤 1807）。从而，对被指定为待处理主盘的盘#0 的重建处理结束（步骤 1104）。在以上时刻，被指定为待处理副盘的盘#1 中的当前位置已到达盘#1 的顶端。

另外，当盘#1 发生故障时，触发对被指定为待处理主盘的盘#1 的另一个重建处理，并且盘#1 的底端被设为当前位置（步骤 1101）。由于恢复目标盘的数目为 1（步骤 1401），所以采用对一个盘故障的重建处理（步骤 1403），并且从正常运行的盘#0、#2 和#3 中读取经由一个盘故障的重建处理来恢复盘#1 中的数据/奇偶校验位时所需的数据/奇偶校验位（步骤 1405）。然而，关于盘#0，只读取已被写入热备份 HS #0 中的恢复数据/奇偶校验位。

然后，利用读取的数据/奇偶校验位来恢复盘#1 中的数据/奇偶校验位，并且将恢复的数据/奇偶校验位写入热备份 HS #1 中（步骤 1407）。

盘#1 中的当前位置被设置为盘#1 中的恢复开始位置（步骤 1801），并且在盘#1 中已经设置了停止位置，因此条件 b 不满足（步骤 1802）。另外，由于盘#1 中的当前位置还未到达盘#1 中的停止位置（步骤 1804），所以确定继续恢复处理（步骤 1805）。

对于每一个条带，反复执行以上的恢复例程和恢复后处理（步骤 1105），并且盘#1 中的当前位置到达停止位置。此时，盘#1 中的当前位置还未到达盘#1 的顶端，从而确定还未完成对整个区域的恢复（步骤 1803）。然而，当前位置对应于停止位置（步骤 1804），确定恢复处理将结束（步骤 1807）。从而，对被指定为待处理主盘的盘#1 的重建处理结束（步骤 1104）。

并行执行对被指定为待处理主盘的盘#0 的两个盘故障的重建处理以及对被指定为待处理主盘的盘#1 的一个盘故障的重建处理，并且当以上两个处理都结束时，对盘#1 的恢复完成。

6. 计数方案 6

在计数方案 1 到 5 的重建处理中，用于指示每个盘中诸如条带、逻辑块等之类的每个规定区域的恢复状态的位图被加为控制信息。由计数方案 1 到 5 中的一个来进行对所有盘的进度控制。

在执行重建处理或恢复例程时，控制器 211 参考在位图中对应于恢复位置的位信息。然后，如图 19 所示，当作为读/写访问等的处理之一已恢复了某一区域时，则跳过对已恢复区域的恢复。从而，减少了用于恢复处理的不必要开销。

图 20 示出了提供控制器 211 中的处理器 221 处理时所用的程序和数据的方法。存储在外部设备 1801 或诸如信息处理设备等之类的移动式存储介质 1802 中的程序和数据被加载到 RAID 设备 202 的存储器 222。

外部设备 1801 生成用于携带程序和数据的载波信号，并将程序和数据经由通信网络上的任意传输介质传输到 RAID 设备 202。移动式存储介质 1802 是诸如存储器卡、软盘、光盘、磁光盘等的任意计算机可读存储介质。处理器 221 利用存储介质中的数据执行程序，并执行所需的处理。

图 21 和图 22 分别示出了存储系统的其他配置示例。在图 21 示出的示

例中，位于主机设备中的主机总线适配器执行重建处理。在图 22 示出的示例中，位于主机设备中的软件执行重建处理。在两种配置中，以与 RAID 设备 202 的情形中相同的方式提供了必要的程序和数据。

图 21 中的存储系统包括主机设备 1901 和盘#0 到#3。主机设备 1901 包括主机总线适配器 1911。主机总线适配器 1911 包括处理器 1921、存储器 1922 和缓存 1923，并且在盘#0 到#3 发生故障时执行重建处理。这时，处理器 1921 执行存储在存储器 1922 中的程序，从而执行上述的重建处理。

图 22 中的存储系统包括主机设备 2001 和盘#0 到#3。主机设备 2001 包括处理器 2011、存储器 2012 和 2013，并且在盘#0 到#3 发生故障时执行重建处理。这时，处理器 2011 执行存储在存储器 2012 中的程序，从而在存储器 2013 上执行上述的重建处理。

另外，在以上实施例中，采用磁盘设备作为盘设备，然而，本发明也可应用于使用其他盘设备的存储系统或诸如磁带设备的其他存储设备，其他盘设备例如光盘设备、磁光盘设备等。

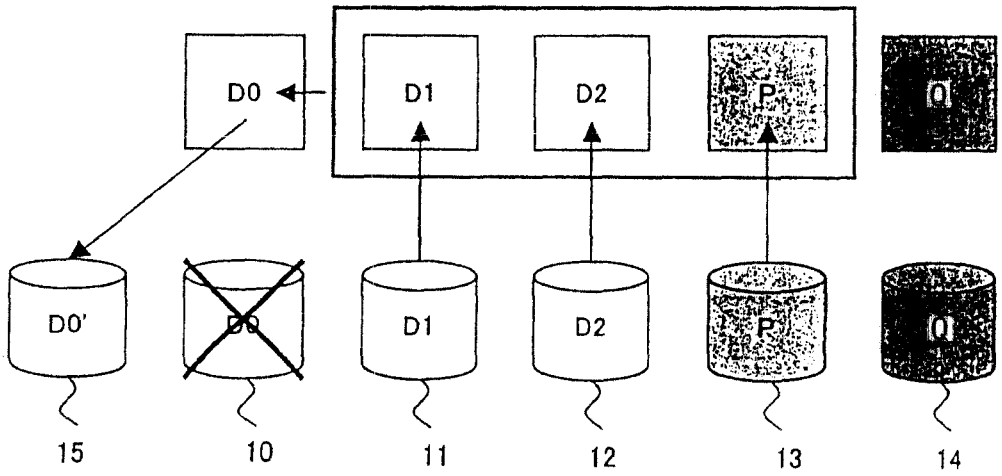


图1A

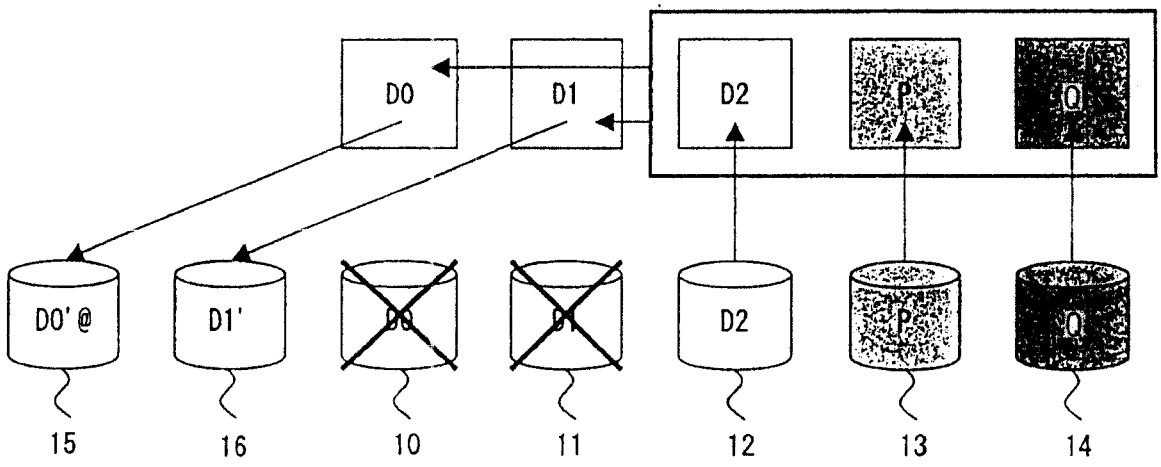


图1B

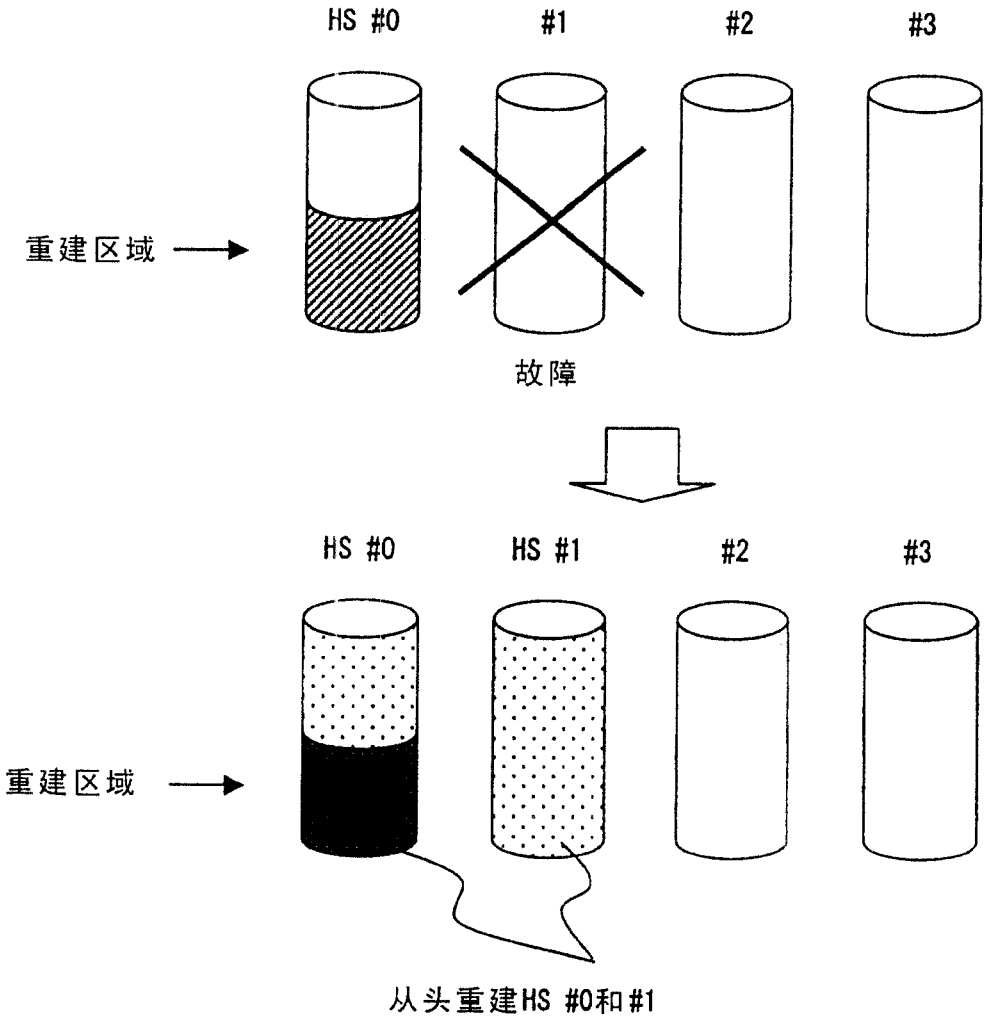


图1C

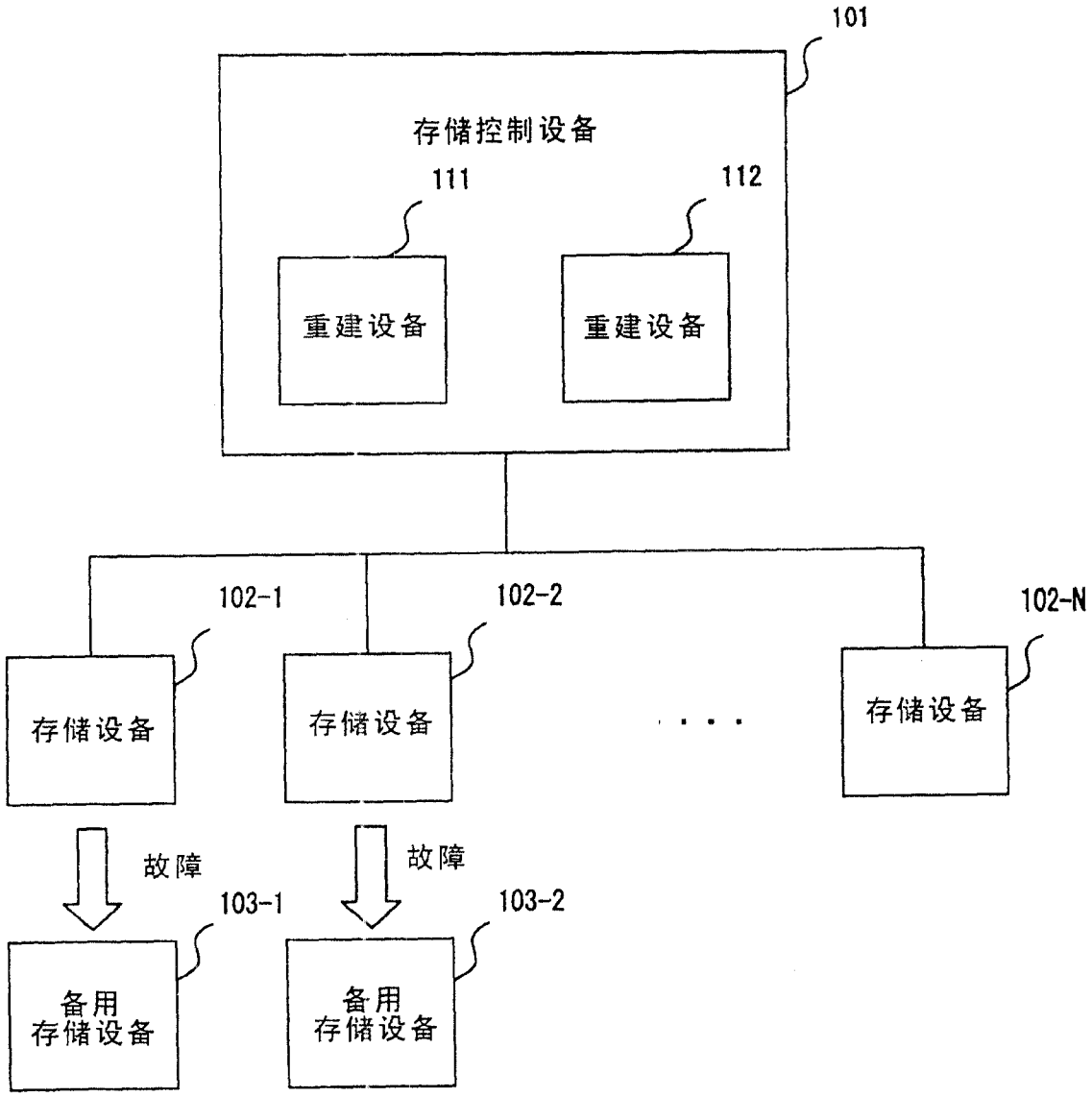

101
存储控制设备
111
112
重建设备
重建设备
102-1
102-2
102-N
存储设备
存储设备
存储设备
故障
故障
103-1
103-2
备用
存储设备
备用
存储设备

图2A

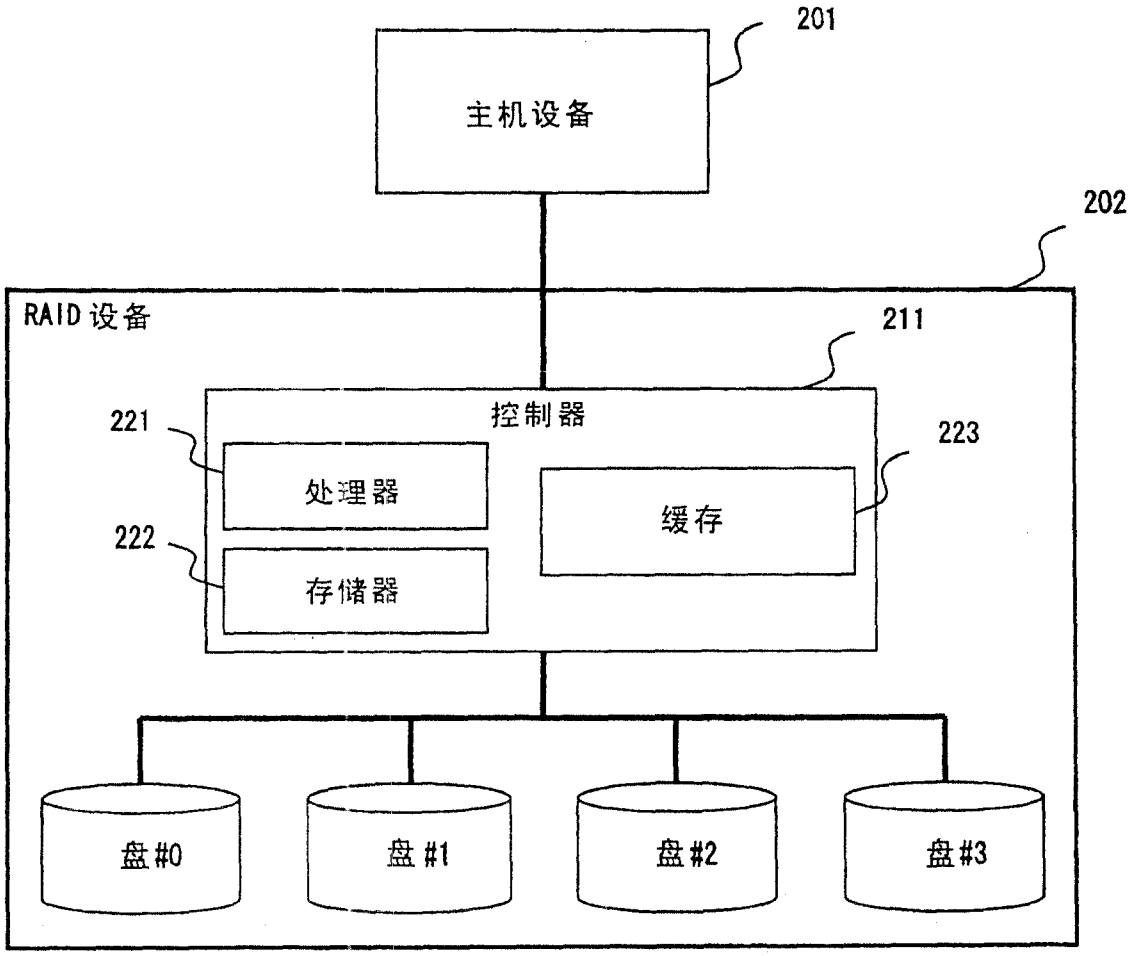


图2B

	方案	管理进度的方法	人工小时数/难度
1	分离执行的方案	与传统技术相同	容易
2	等待进度位置相互对应的方案	与传统技术相同	容易
3	组合方案	与传统技术相同	中等
4	双重故障区域被优先恢复的方案	除了传统技术外， 还管理恢复完成位置	对正常操作有影响
5	并行执行对双重故障区域的恢复 和使用已恢复热备份HS的恢复的方案	除了传统技术外， 还管理恢复完成位置	
6	随机恢复方案	位图管理	人工小时数更多， 且对正常操作有影响

图3

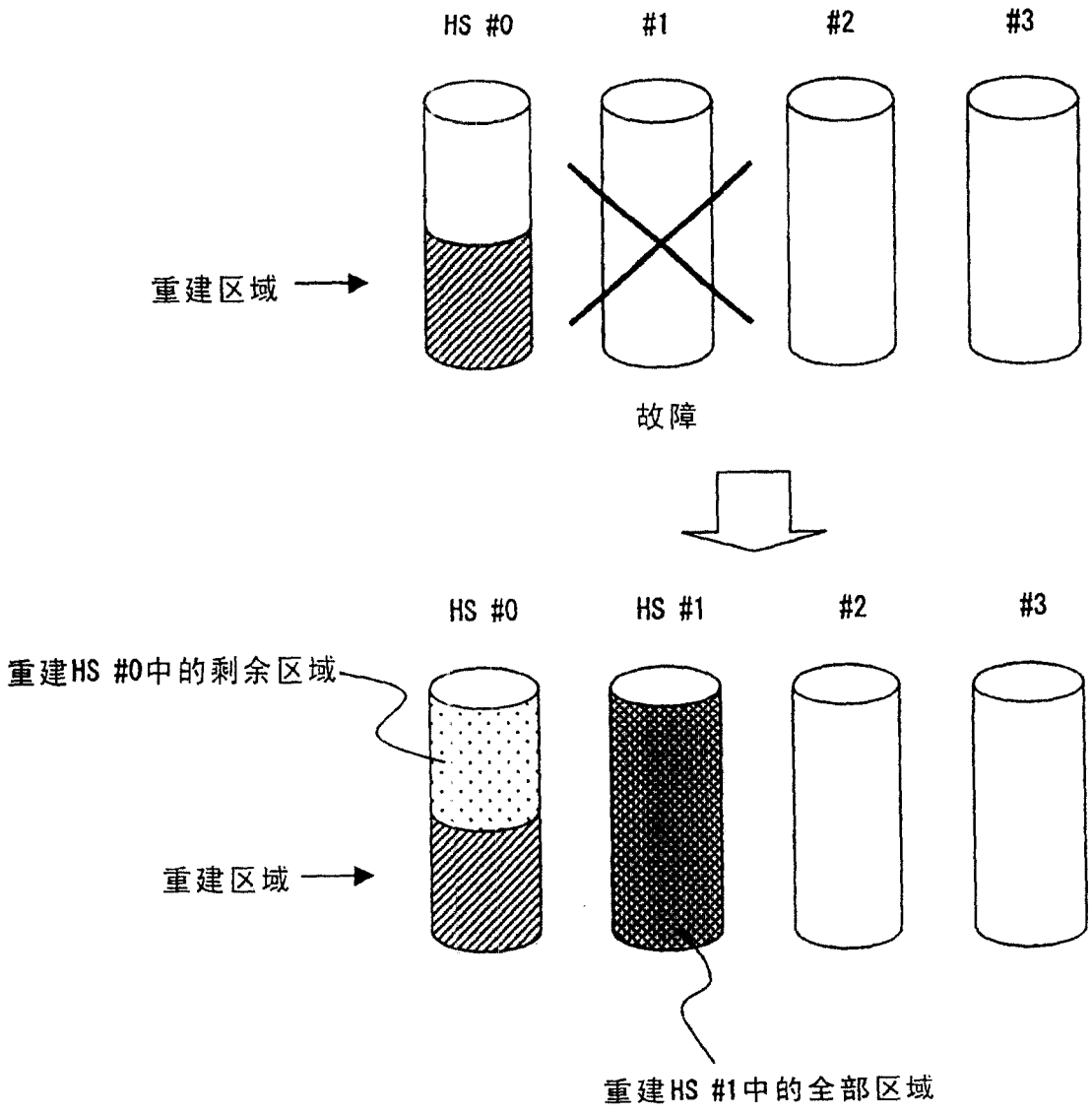


图4

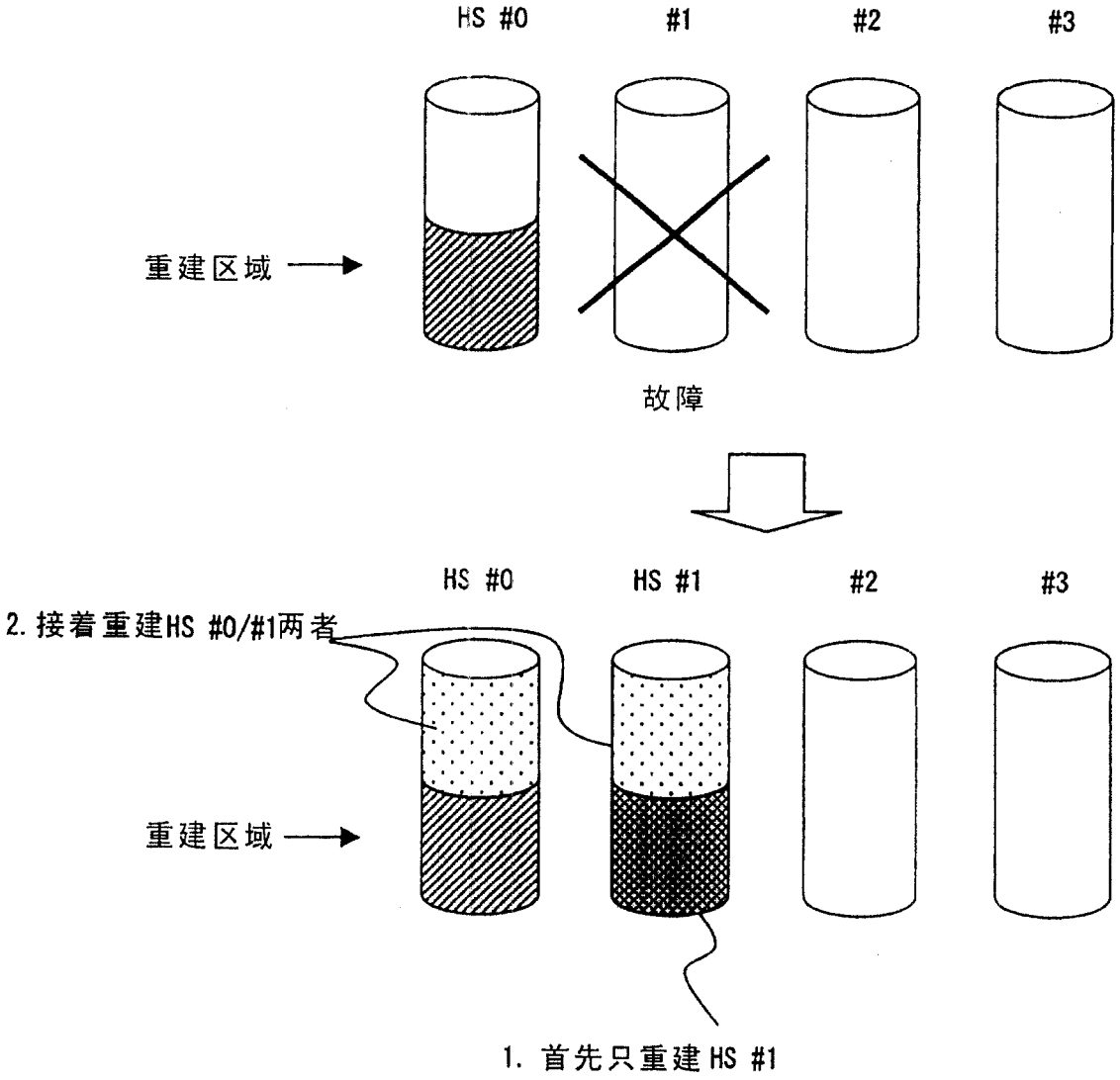


图5

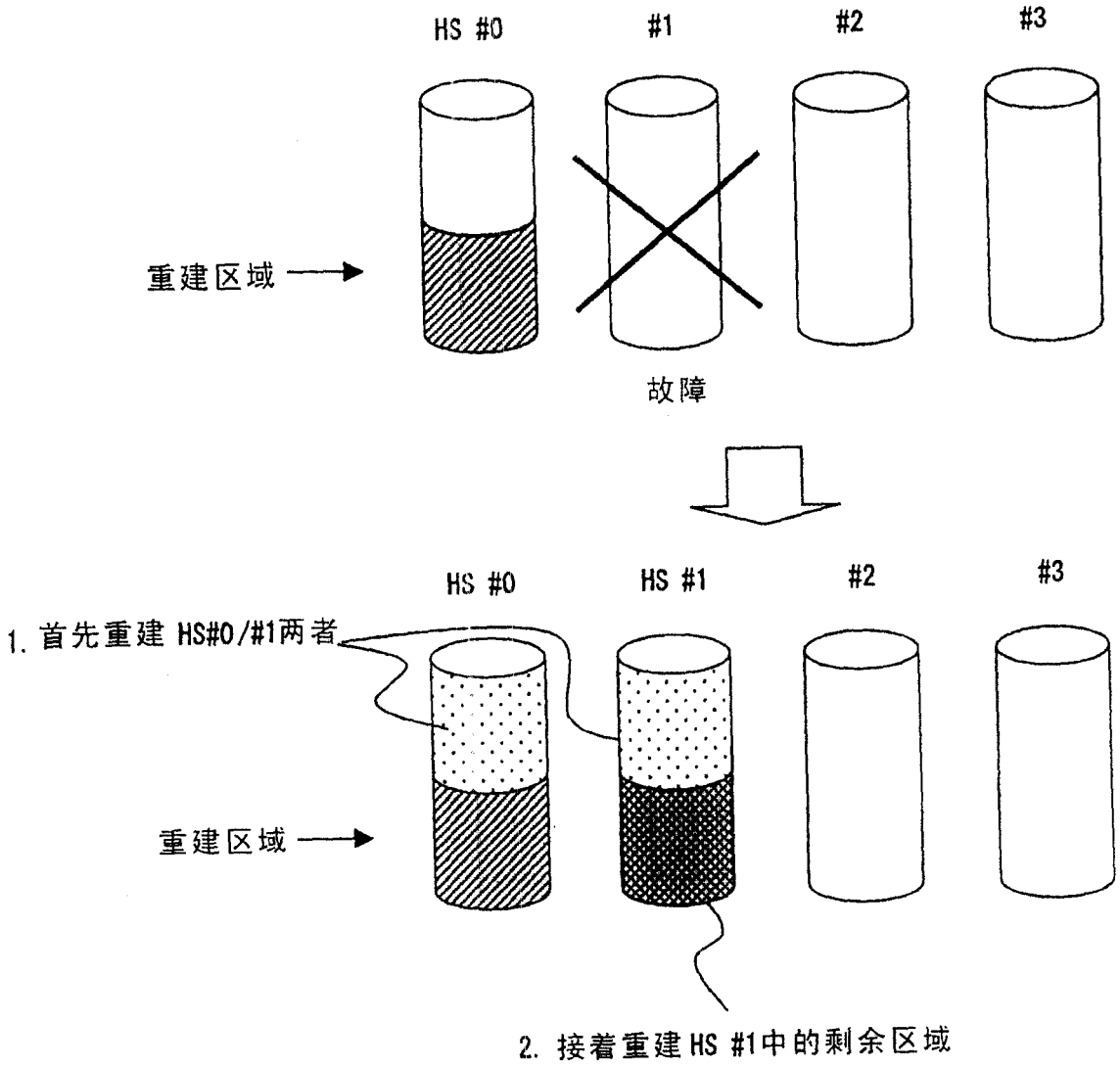


图6

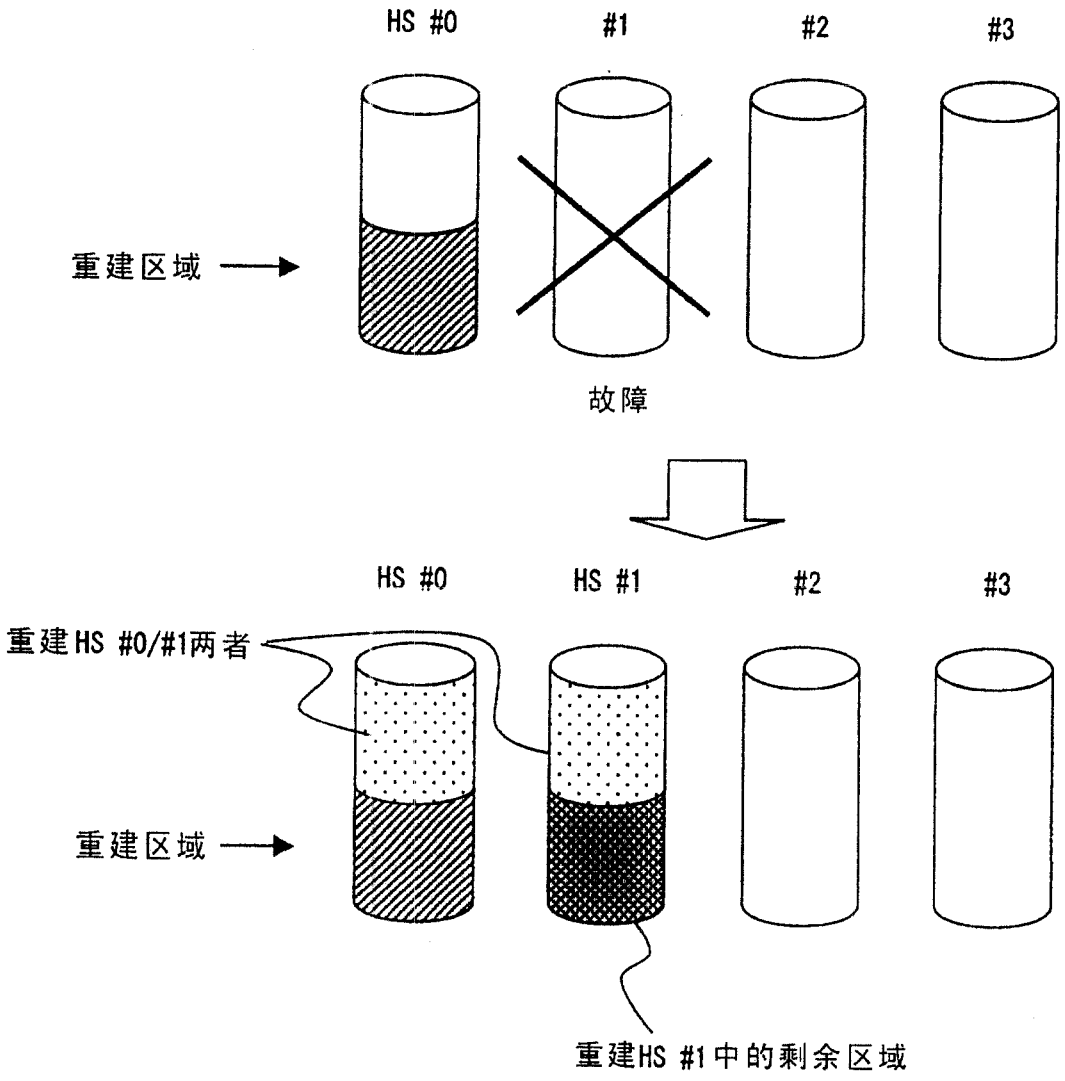


图7

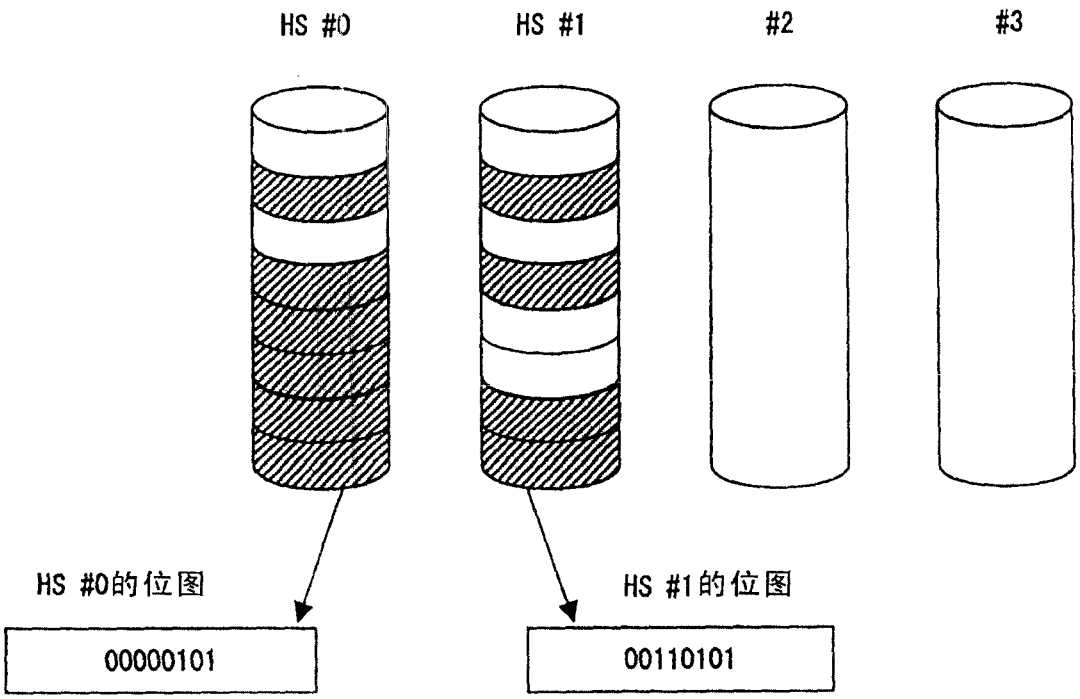


图8

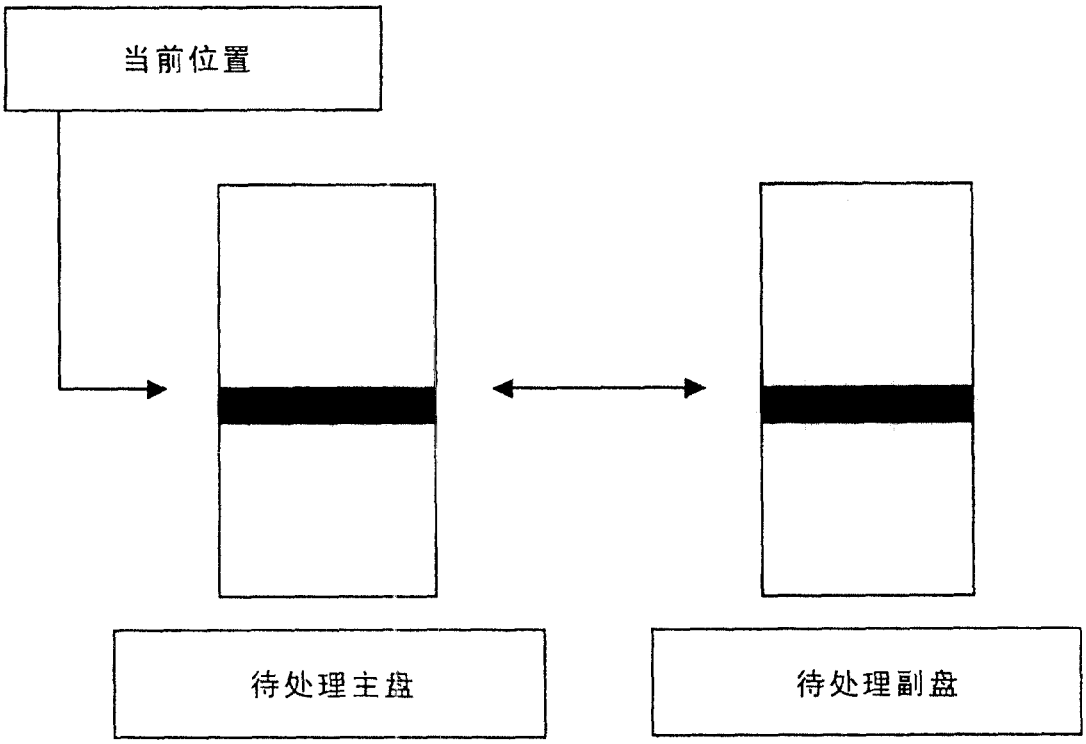


图9

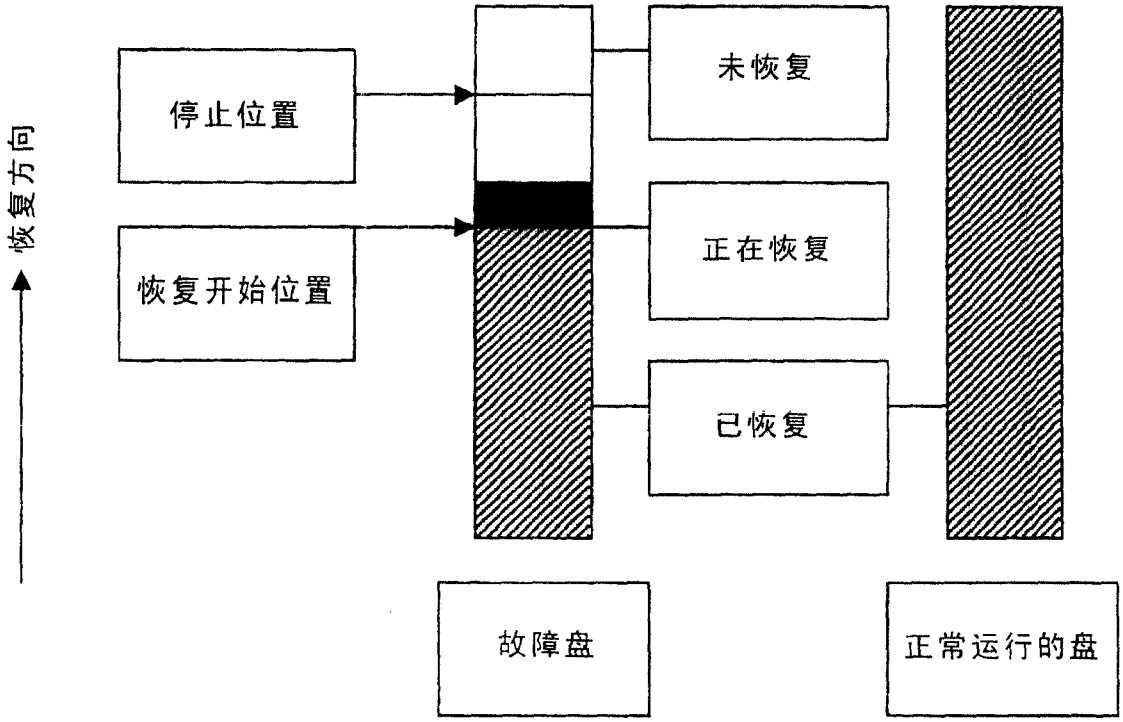


图10

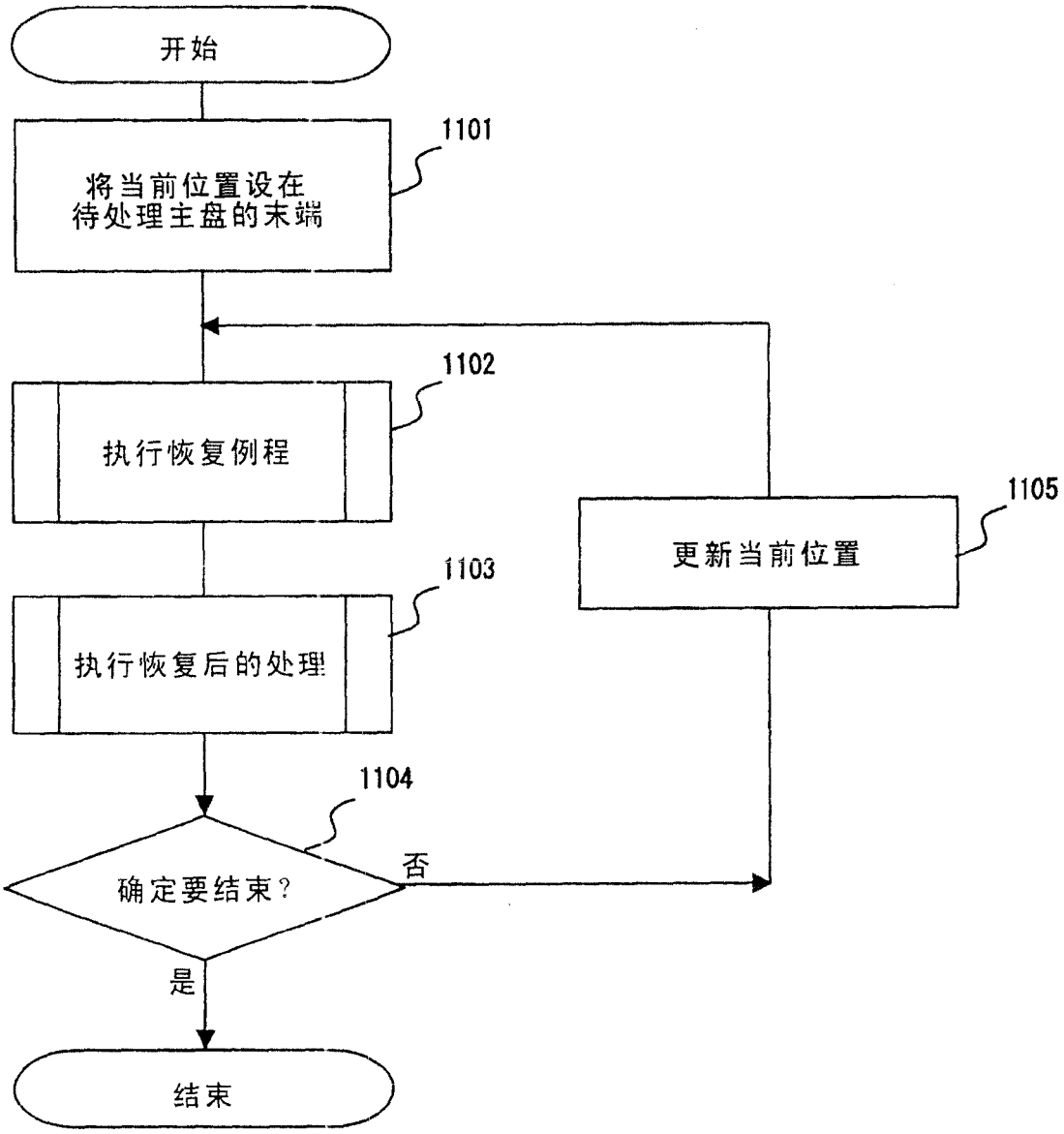

开始
1101
将当前位置设在
待处理主盘的末端
1102
执行恢复例程
1103
执行恢复后的处理
1104
确定要结束？
否
是
结束
1105
更新当前位置

图11

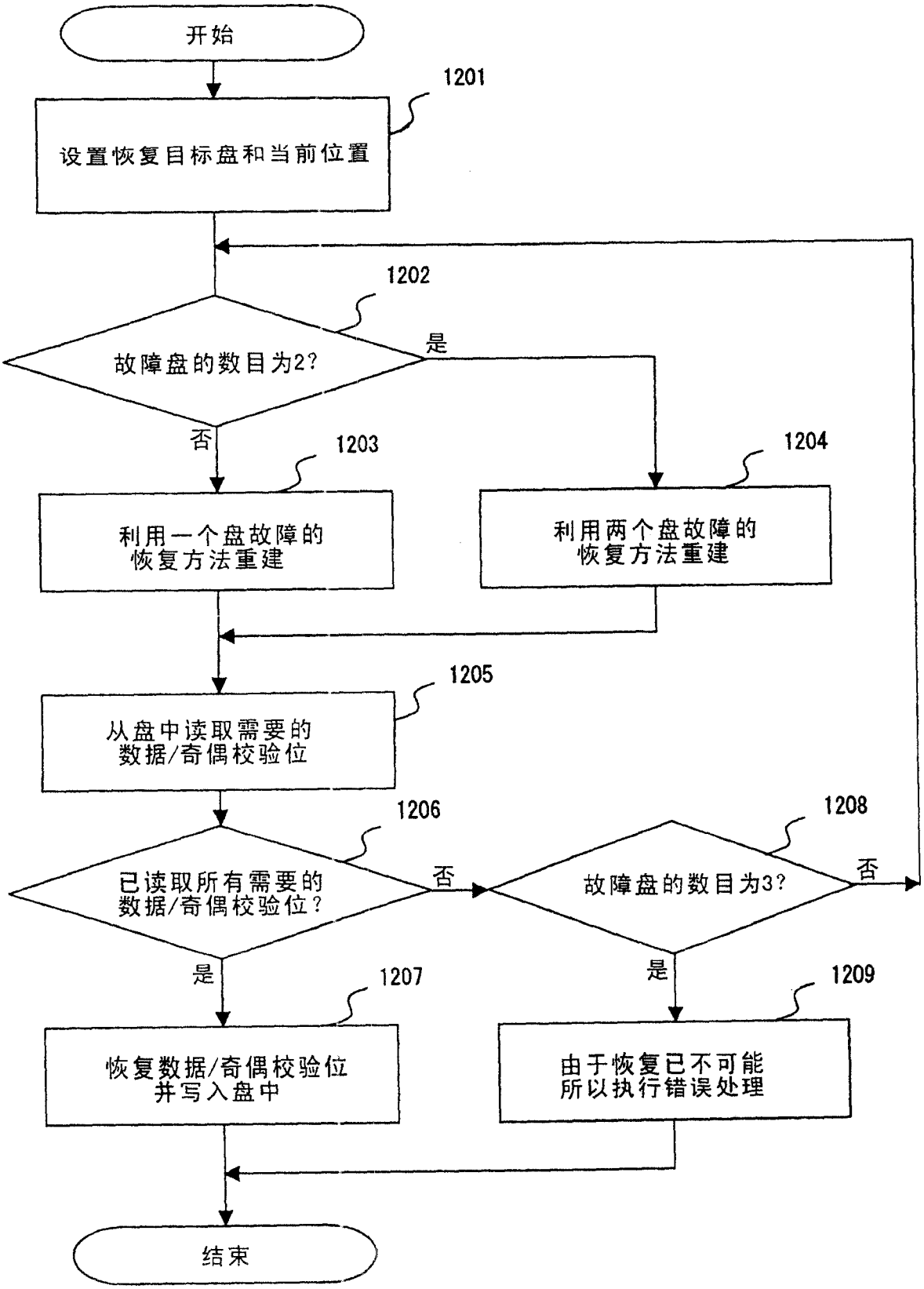

开始
1201
设置恢复目标盘和当前位置
1202
故障盘的数目为2?
是
否
1203
利用一个盘故障的
恢复方法重建
1204
利用两个盘故障的
恢复方法重建
1205
从盘中读取需要的
数据/奇偶校验位
1206
已读取所有需要的
数据/奇偶校验位?
否
1208
故障盘的数目为3?
否
是
1207
恢复数据/奇偶校验位
并写入盘中
是
1209
由于恢复已不可能
所以执行错误处理
结束

图12

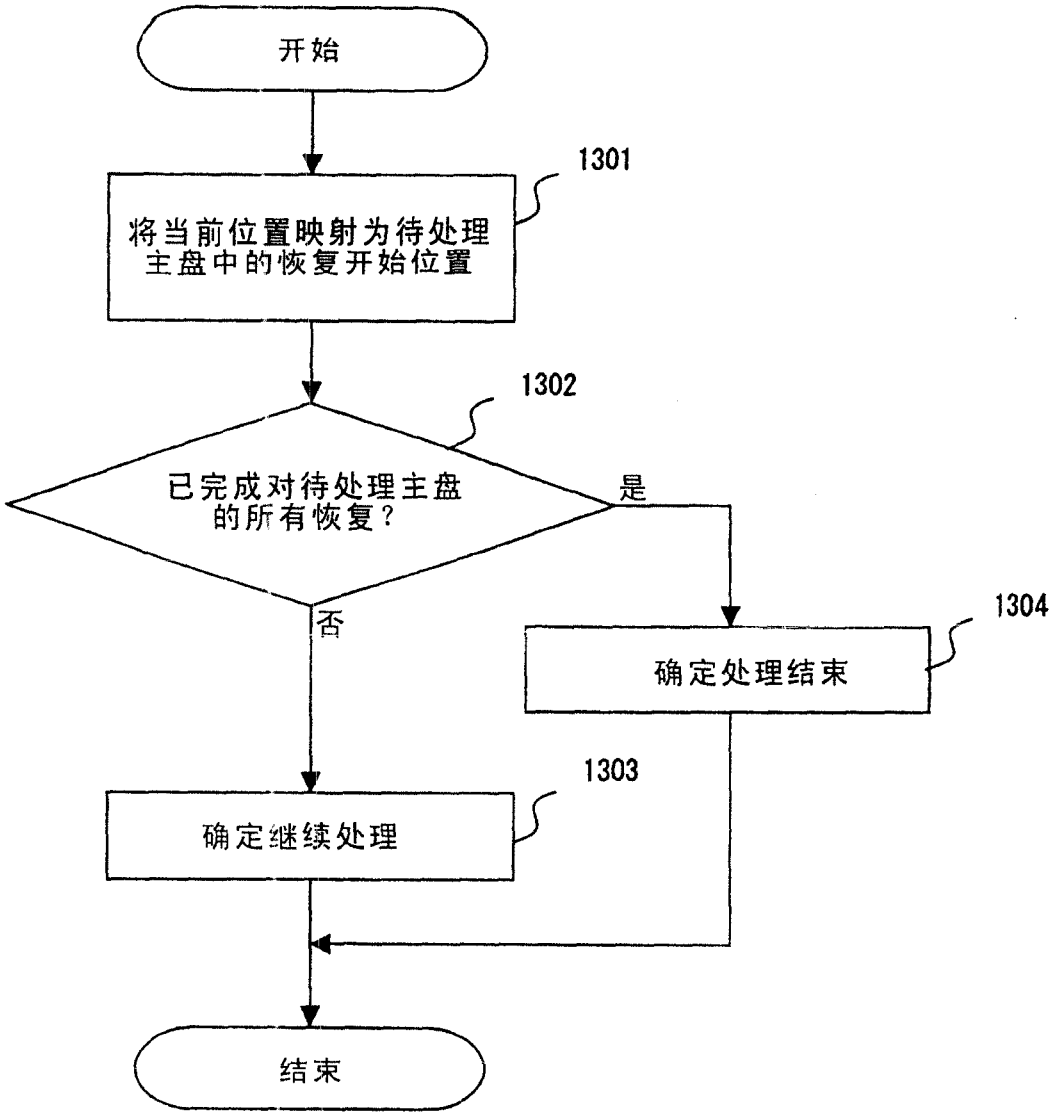

开始
1301
将当前位置映射为待处理主盘中的恢复开始位置
1302
已完成对待处理主盘的所有恢复？
是
否
1304
确定处理结束
1303
确定继续处理
结束

图13

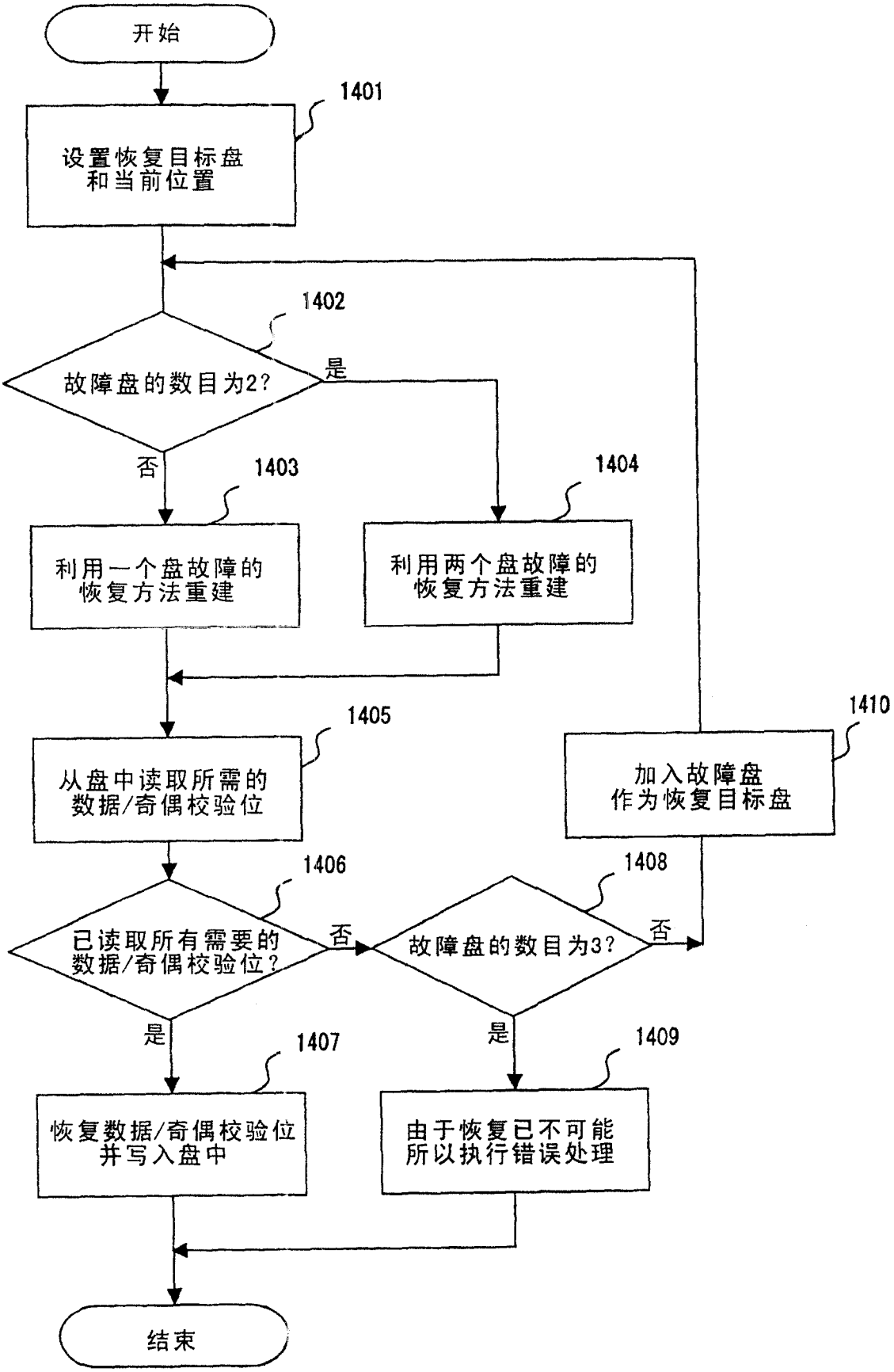

开始
1401
设置恢复目标盘
和当前位置
1402
故障盘的数目为2?
是
否
1403
利用一个盘故障的
恢复方法重建
1404
利用两个盘故障的
恢复方法重建
1405
从盘中读取所需的
数据/奇偶校验位
1406
已读取所有需要的
数据/奇偶校验位?
否
是
1407
恢复数据/奇偶校验位
并写入盘中
1408
故障盘的数目为3?
否
是
1409
由于恢复已不可能
所以执行错误处理
1410
加入故障盘
作为恢复目标盘
结束

图14

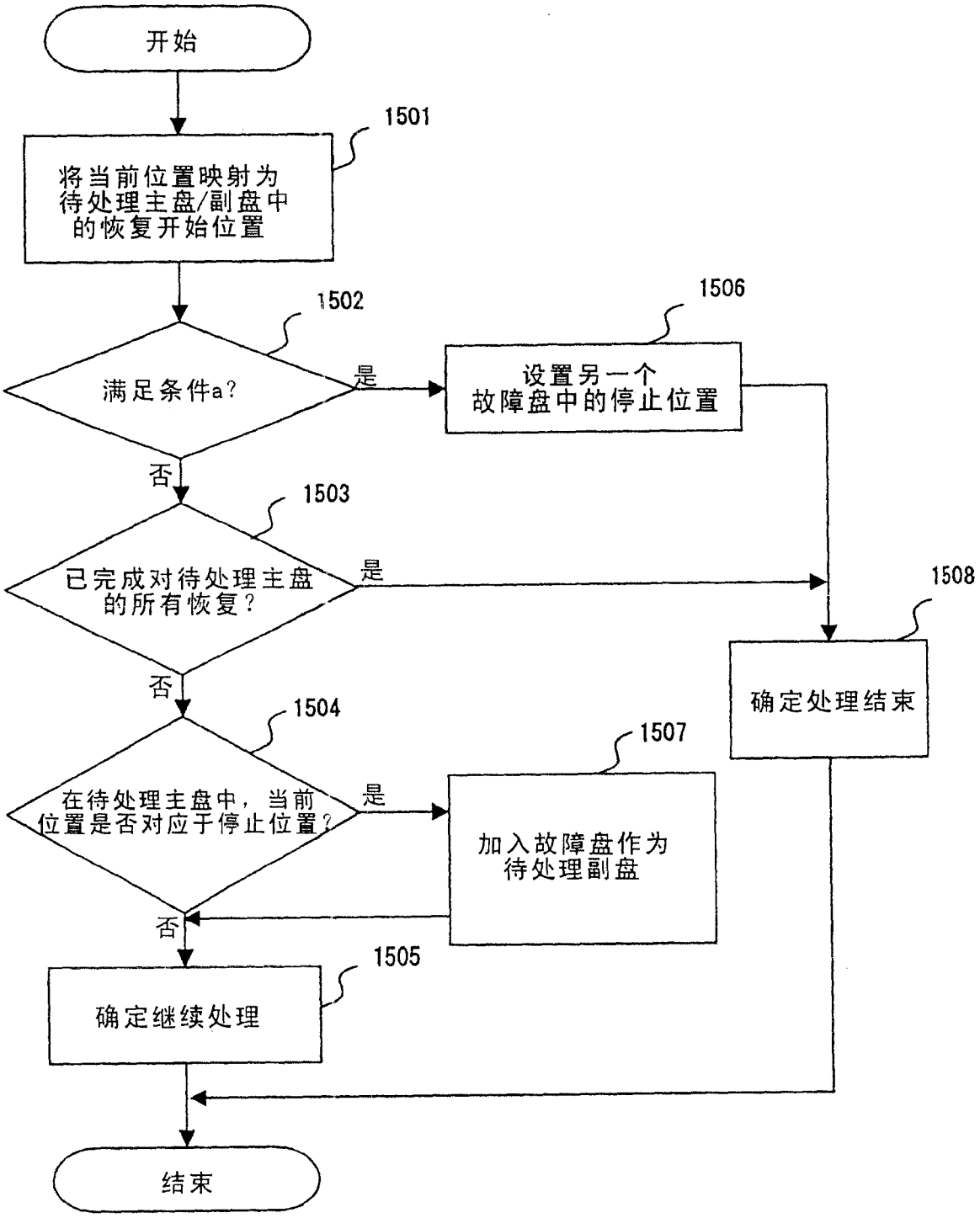

开始
1501
将当前位置映射为待处理主盘/副盘中的恢复开始位置
1502
满足条件a?
是
1506
设置另一个故障盘中的停止位置
否
1503
已完成对待处理主盘的所有恢复?
是
1508
确定处理结束
否
1504
在待处理主盘中，当前位置是否对应于停止位置?
是
1507
加入故障盘作为待处理副盘
否
1505
确定继续处理
结束

图15

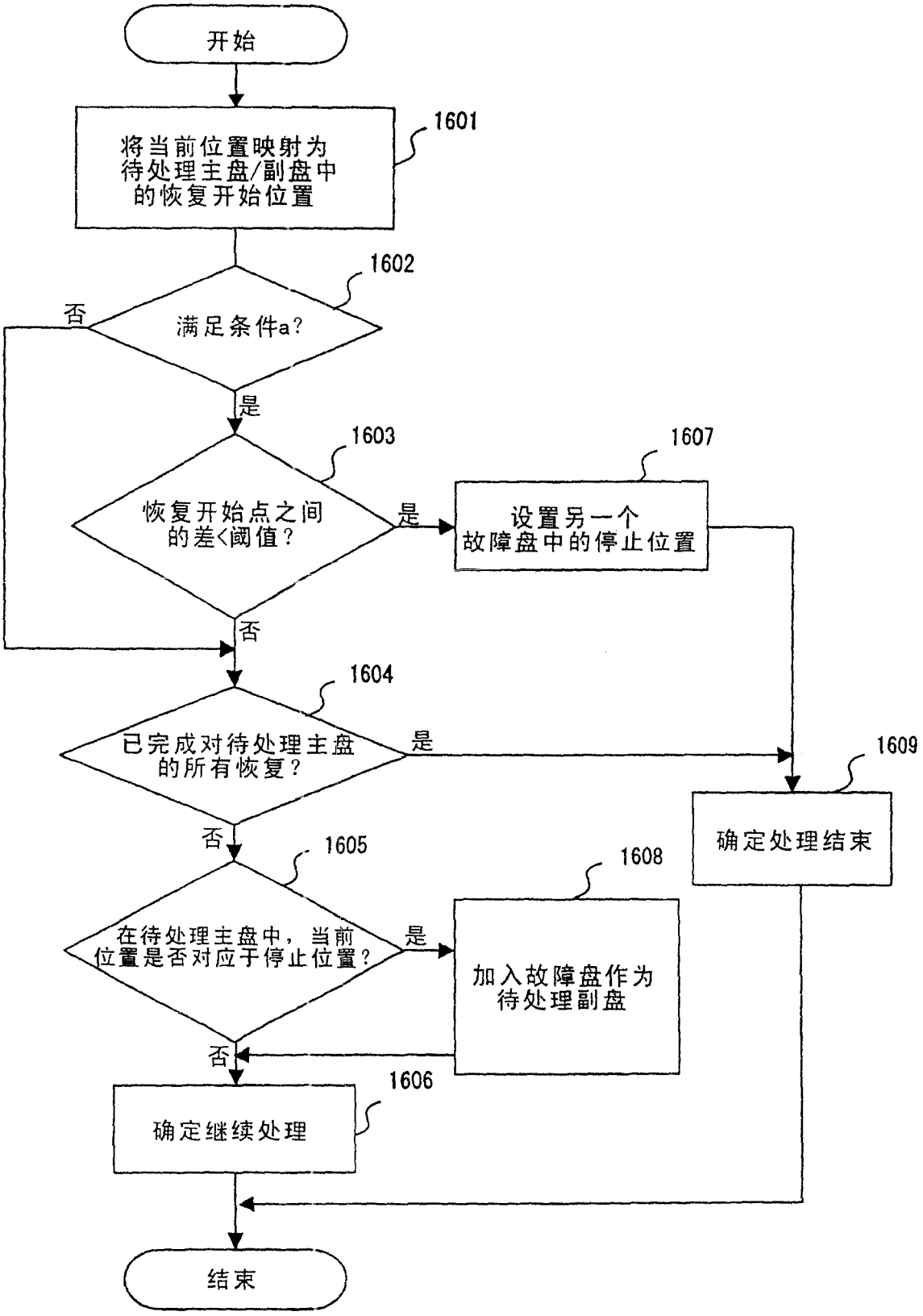

开始
将当前位置映射为待处理主盘/副盘中的恢复开始位置
1601
满足条件a?
1602
否
是
恢复开始点之间的差<阈值?
1603
是
设置另一个故障盘中的停止位置
1607
否
已完成对待处理主盘的所有恢复?
1604
是
确定处理结束
1609
否
在待处理主盘中，当前位置是否对应于停止位置?
1605
是
加入故障盘作为待处理副盘
1608
否
确定继续处理
1606
结束

图16

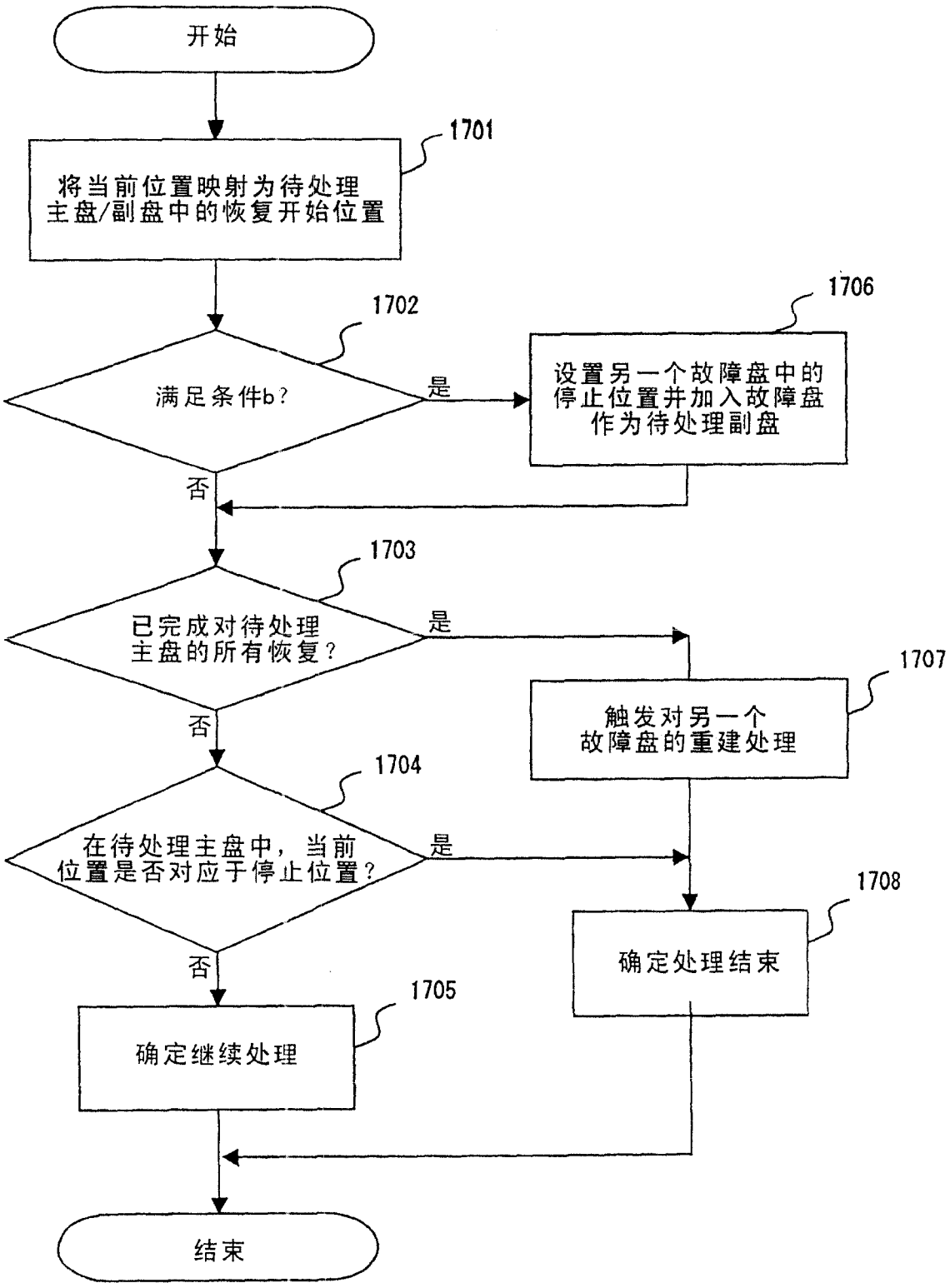

开始
1701
将当前位置映射为待处理主盘/副盘中的恢复开始位置
1702
满足条件b?
是
1706
设置另一个故障盘中的停止位置并加入故障盘作为待处理副盘
否
1703
已完成对待处理主盘的所有恢复?
是
1707
触发对另一个故障盘的重建处理
否
1704
在待处理主盘中，当前位置是否对应于停止位置?
是
1708
确定处理结束
否
1705
确定继续处理
结束

图17

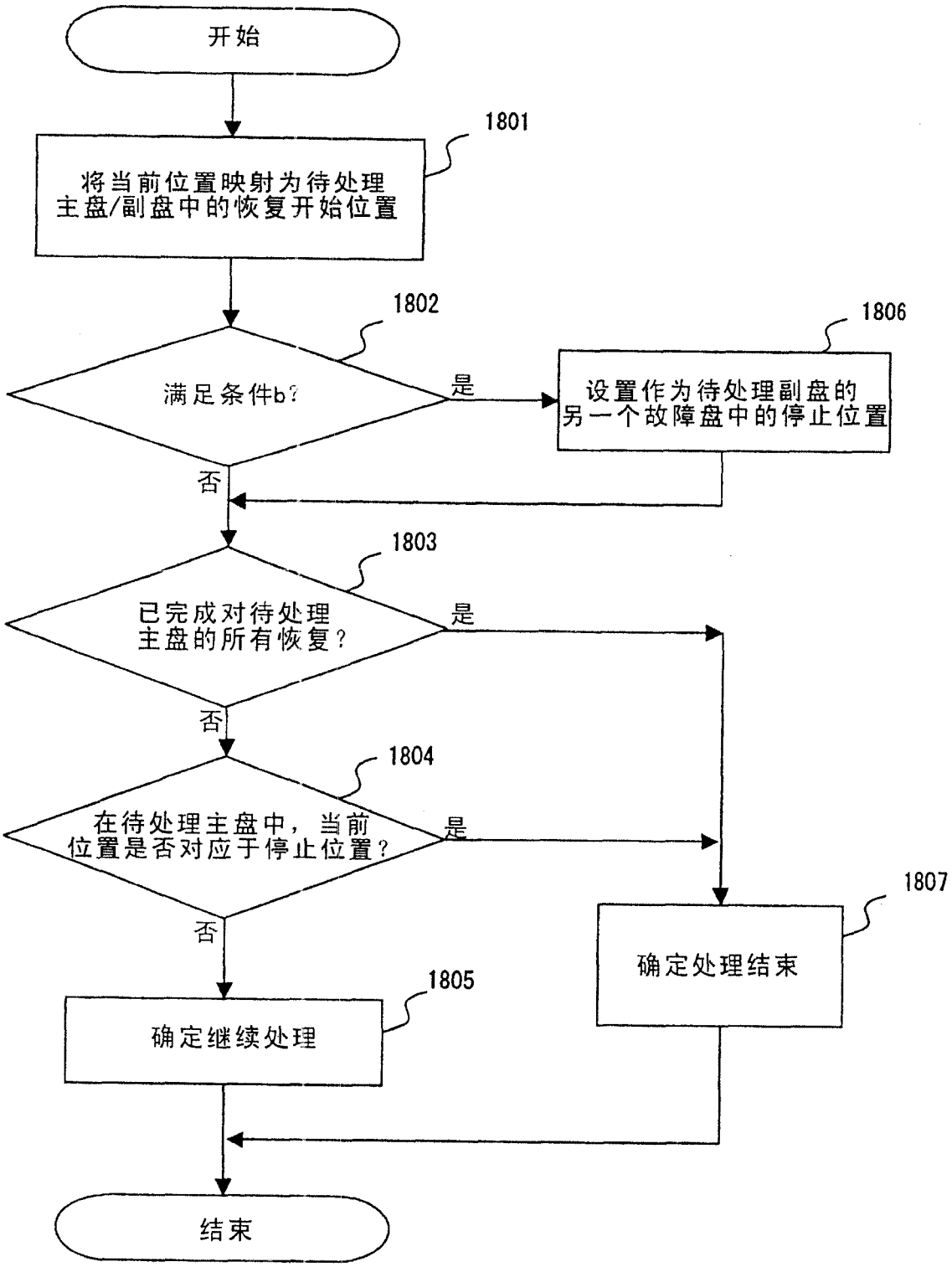

开始
1801
将当前位置映射为待处理主盘/副盘中的恢复开始位置
1802
满足条件b?
是
1806
设置作为待处理副盘的另一个故障盘中的停止位置
否
1803
已完成对待处理主盘的所有恢复?
是
否
1804
在待处理主盘中，当前位置是否对应于停止位置?
是
否
1807
确定处理结束
1805
确定继续处理
结束

图18

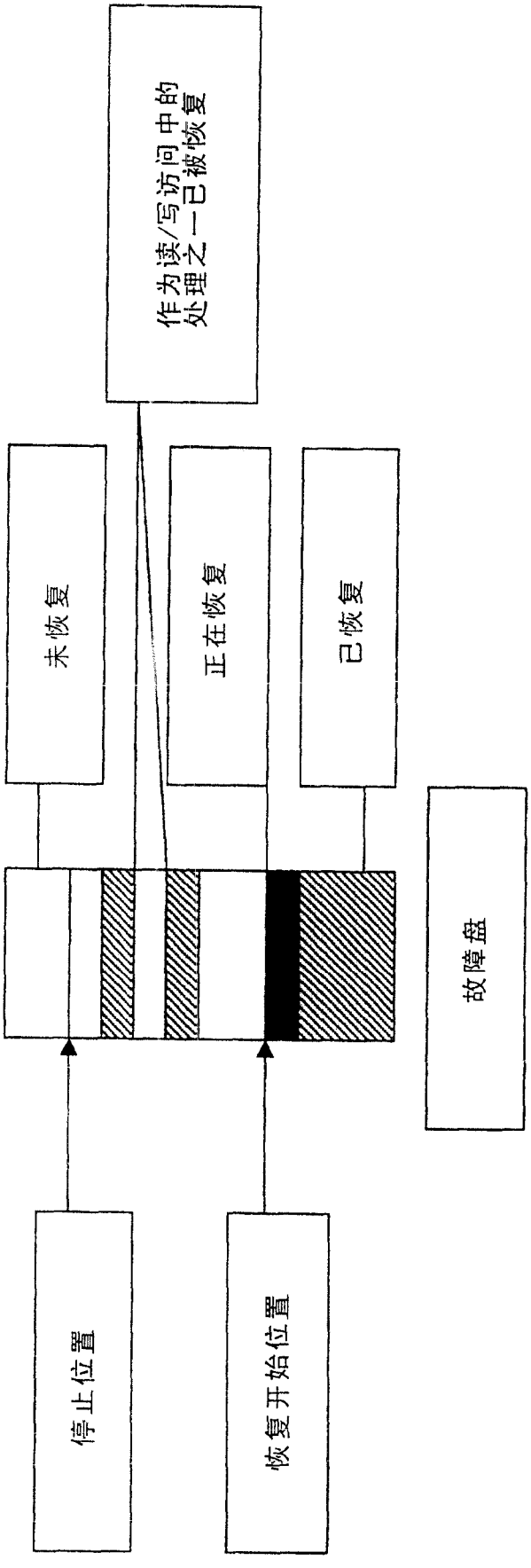


图19

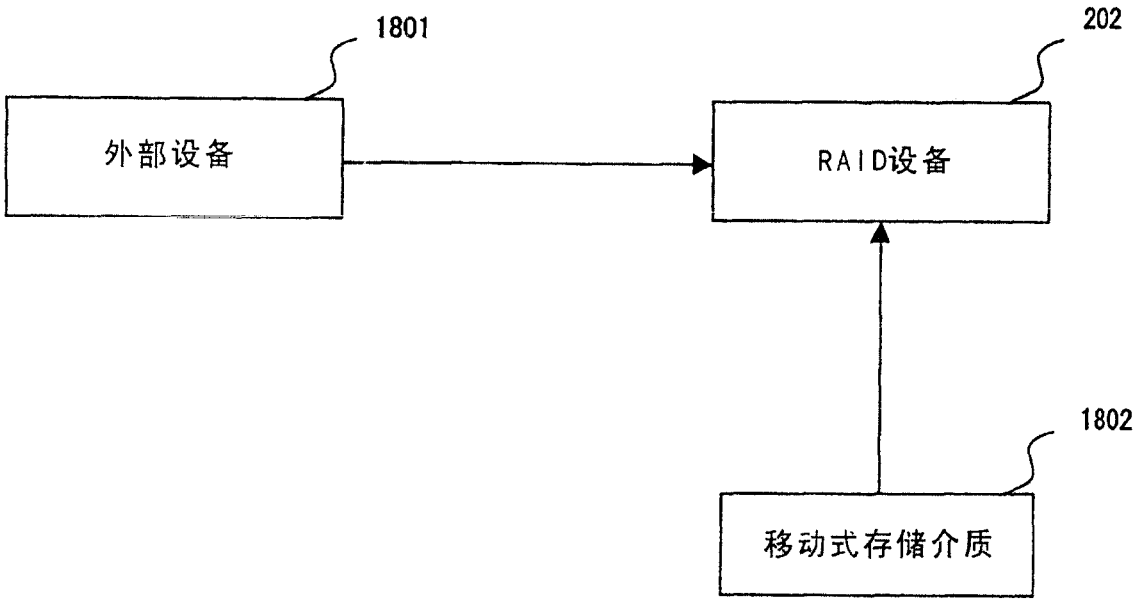


图20

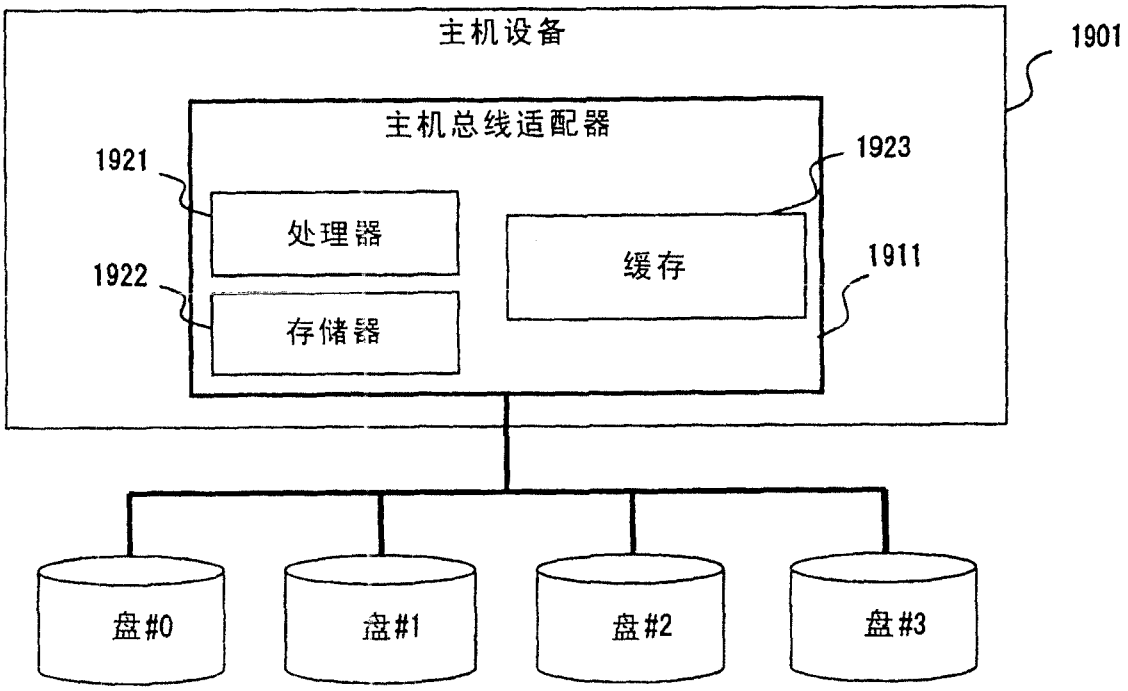


图21

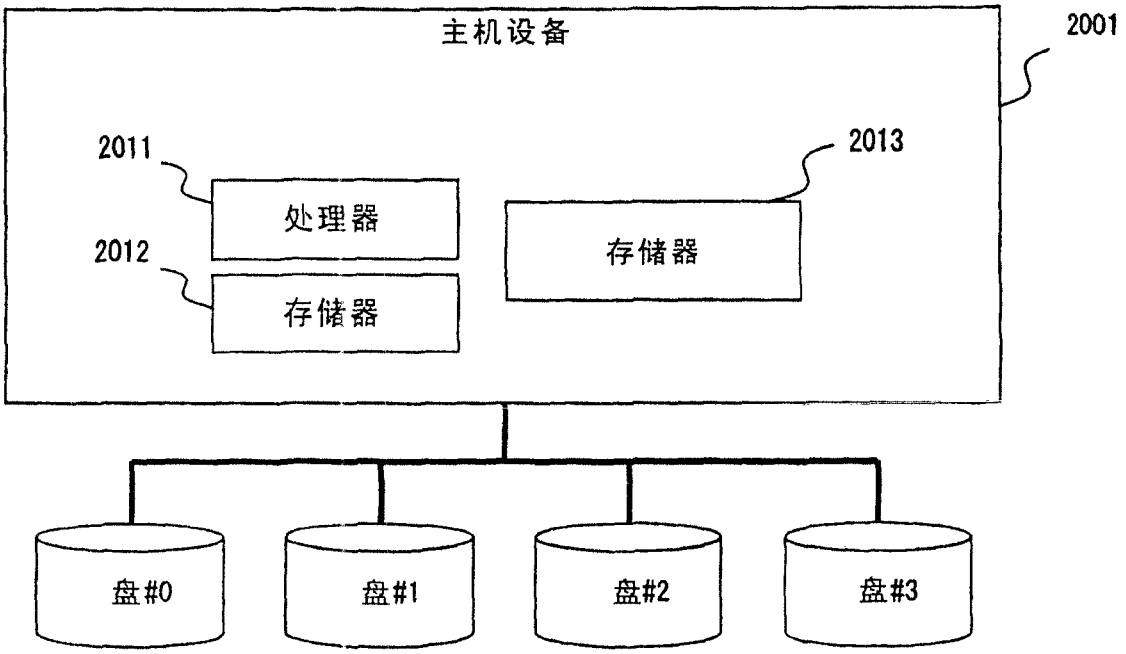


图22