

[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 200410029467.5

[45] 授权公告日 2006 年 12 月 20 日

[11] 授权公告号 CN 1291304C

[22] 申请日 2004.3.19

[21] 申请号 200410029467.5

[30] 优先权

[32] 2003. 3. 20 [33] JP [31] 2003 – 076865

[73] 专利权人 株式会社日立制作所

地址 日本东京都

[72] 发明人 海谷佳一 坪木雅直 水主和人

审查员 白雪涛

[74] 专利代理机构 北京银龙知识产权代理有限公司
代理人 郝庆芬

代理人 郝庆芬

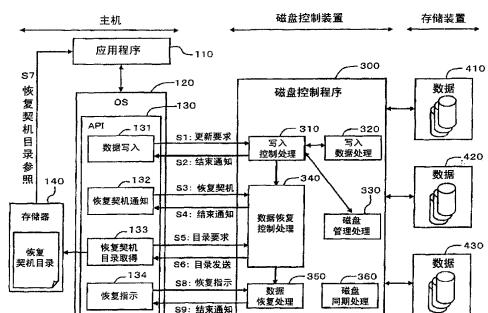
权利要求书 2 页 说明书 19 页 附图 10 页

[54] 发明名称

外部存储装置及外部存储装置的数据恢复方法

[57] 摘要

主机上的应用程序(110)向磁盘控制装置的数据恢复控制处理(340)指示恢复契机的设定(S3)。通过设置包含在运行记录数据中的恢复标志，能够将任意的多个时间点作为恢复可能的时间点使其登记。当发生故障等后要使数据恢复时，应用程序(110)要求示出已设定的恢复契机的一览表的目录(S5)。应用程序(110)根据恢复契机目录，指定使数据恢复的时间点(S8)。磁盘控制装置根据备份磁盘(420)及运行记录磁盘(430)使数据恢复到被指定的时间点。



1. 一种与主机连接的外部存储装置，其特征为，包括以下内容：
 包括存储被所述主机利用的数据的存储装置和控制所述存储装置的
 控制装置，

 所述控制装置包括：关于被所述存储装置存储的数据，登记由所述
 主机设定的恢复可能时间点的登记装置；根据来自所述主机的要求，将
 所述登记的恢复可能时间点的选择用信息发送给所述主机的选择用信息
 发送装置；根据所述恢复可能时间点的选择用信息，将由所述主机指定
 的数据恢复到指定的恢复可能时间点的恢复装置。

2. 如权利要求 1 所述的外部存储装置，其特征为，包括以下内容：
 所述登记装置可将由所述主机设定的任意的多个时间点作为所述恢
 复可能时间点登记。

3. 如权利要求 1 所述的外部存储装置，其特征为，包括以下内容：
 所述存储装置包括将来自所述主机的写入数据作为运行记录数据存
 储的运行记录数据存储装置，

 所述登记装置是根据来自所述主机的指示，通过在所述运行记录数
 据的指定位置对应附加标识信息，来登记所述恢复可能时间点的装置。

4. 如权利要求 3 所述的外部存储装置，其特征为，包括以下内容：
 所述运行记录数据至少包含写入数据、写入位置和作为所述标识信
 息的恢复标志信息而构成，

 所述登记装置是通过设置所述运行记录数据中的指定的恢复标志信
 息，来登记所述恢复可能时间点的装置。

5. 如权利要求 3 所述的外部存储装置，其特征为，包括以下内容：
 所述存储装置包括存储备份数据的备份数据存储装置，
 所述控制装置包括运行记录数据管理装置，
 所述运行记录数据管理装置，是当所述运行记录数据存储装置的未
 用容量不足时，将存储在所述运行记录数据存储装置的最旧的运行记录
 数据移换到备份数据存储装置，增加所述运行记录数据存储装置的未用

容量，并且将在所述被登记的恢复可能时间点中最旧的恢复可能时间点已被变更的意旨通知给所述主机的装置。

6. 如权利要求3所述的外部存储装置，其特征为，包括以下内容：

所述控制装置包括运行记录数据管理装置，

所述运行记录数据管理装置，是在所述运行记录数据存储装置的未用容量不足的时候，利用所述存储装置内的未使用的存储区域，自动扩展运行记录数据存储装置的逻辑容量的装置。

7. 将与主机连接的外部存储装置的数据在该外部存储装置内使其恢复的数据恢复方法，其特征为，包括以下步骤：

对被存储的数据，登记由所述主机在任意的多个时间点上能够设定的恢复可能时间点的登记步骤；按照来自所述主机的要求，将所述登记的恢复可能时间点的选择用信息发送给所述主机的一览发送步骤；根据所述恢复可能时间点的选择用信息，将由所述主机指定的数据恢复到指定的恢复可能时间点的恢复步骤。

外部存储装置及外部存储装置的数据恢复方法

技术领域

本发明涉及例如磁盘装置等的外部存储装置及外部存储装置的数据恢复方法与程序。

背景技术

在处理比较多量数据的业务用应用程序（数据库系统）中，将数据保存在和主机分开形成的磁盘阵列装置上。然后，从主机的数据库系统访问磁盘阵列装置上的数据，进行各种数据操作。所谓磁盘阵列装置，是将多个磁盘装置配设成阵列状而成，根据来自主机的存储命令和读取命令等进行工作。

在此，数据库系统在运转过程中，例如，当因非预期的断电、操作员的误操作、硬件电路及其他程序失常等而发生故障时，有必要使数据库的内容恢复到发生故障前的状态。另外，有时也想将数据操作恢复到故障以外的所希望的时间的状态。

作为技术 1，在通常的数据库系统中，主机上的数据库系统本身将与现实数据分开的运行记录数据（记录数据）写入磁盘阵列装置的指定的磁盘装置。因此，在通常的数据库系统中，根据事前取得的备份数据，数据库系统本身从磁盘装置读取运行记录数据，使其顺序反映到备份数据上。由此，主机上的数据库系统能够在运行记录数据残存的范围内使数据库恢复到所希望的时间点。

在技术 2 中，将第 1 磁盘装置的内容在指定的周期内保存到备份用磁盘装置的同时，将运行记录数据保存到运行记录用磁盘装置。当在第 1 磁盘装置发生故障时，根据备份数据及运行记录数据，在第 2 磁盘装置内生成假想的第 1 磁盘装置，并将向第 1 磁盘装置的数据访问内部切换到假想的第 1 磁盘装置。然后，第 1 磁盘装置的修复结束后，即将假想的第 1 磁盘装置的内容转换到第 1 磁盘装置。

在上述技术 1 中，主机上的数据库系统本身管理运行记录数据，能

在任意的时间点恢复数据。但是，因为数据库系统本身进行数据恢复操作，所以主机的计算机资源（计算存储器）被用在数据恢复处理，在恢复操作中，使应该进行的业务处理及其他业务处理的处理效率降低。另外，数据库系统进行运行记录数据的管理，但当运行记录数据的存储磁盘存满后，只要取不到备份数据，就无法使数据恢复。因此，数据库系统必须进行运行记录数据用磁盘的容量管理等，处理负担变大。进一步，在进行数据的分代管理时，因为制作多代的备份数据，所以处理负担更加增大。

在技术 2 中，通过将访问切换到假想的磁盘装置，能够不中断实际执行中的处理，进行数据恢复操作。但是，只能使数据恢复到稍前的状态，不能在操作员所希望的任意时间点使数据恢复。

发明内容

本发明是鉴于上述问题而提出的，其目的在于提供不使主机侧的处理负担增大，而能够向任意的时间点恢复数据的外部存储装置及外部存储装置的数据恢复方法与程序。本发明的进一步的目的从后述的实施方式的记载中可明确化。

为解决上述问题，基于本发明的第 1 观点的外部存储装置，是与主机连接的装置，其包括存储被主机利用的数据的存储装置和控制存储装置的控制装置。控制装置包括：对被存储装置存储的数据，登记由主机设定的恢复可能时间点的登记装置；根据来自主机的要求，将被登记的恢复可能时间点的选择用信息发送给主机的选择用信息发送装置；根据恢复可能时间点的选择用信息，使被主机指定的数据恢复到被指定的恢复可能时间点的恢复装置。

作为存储装置，例如可采用将多个磁盘装置配置成阵列状的存储装置。主机能够对被存储装置存储的数据设定恢复可能时间点。所谓恢复可能时间点，是指示使该数据恢复的可能的时间点的信息，也可称为复原点。由主机定期、不定期设定的恢复可能时间点，被登记装置登记。

当发生故障等必须恢复数据时，主机向控制装置要求恢复可能时间点的选择用信息。按此要求，选择用信息发送装置将选择用信息发送给

主机。所谓选择用信息，是用于选择恢复可能时间点的信息，例如，可用一览表形式等表示。

主机根据接收到的选择用信息选择想使数据恢复的时间点。由主机选择的恢复可能时间点被通知给恢复装置。之后，恢复装置使被主机指定的数据恢复到被指定的时间点。恢复装置能够通过例如使被指定的到达恢复时间点的运行记录数据按顺序反映到备份数据上恢复数据。由此，事实上几乎不使用主机的计算机资源，就能在外部存储装置内使数据恢复到任意的时间点。

登记装置，能够将由主机设定的任意的多个时间点作为恢复可能时间点进行登记。即，不仅能登记最接近的稍前的最新状态，而且能登记多个任意的时间点。例如，主机每次要求更新处理（提交，日文原文コミット）或每次区分数据操作的分区时，能够通过自动或操作员手动设定恢复可能时间点。

在本发明的一种方式中，存储装置包括取得并存储运行记录数据的运行记录数据存储装置。登记装置根据来自主机的指示，通过将标识信息对应附加在运行记录数据的指定位置登记恢复可能时间点。即，外部存储装置内的运行记录数据存储装置独自自动地收集存储运行记录数据。之后，登记装置根据来自主机的设定，通过在运行记录数据的指定位置对应附加标识信息登记恢复可能时间点。标识信息也可以包含在运行记录数据中，也可作为与运行记录数据不同的数据分别管理，通过独特识别代码等将两者联系。

在本发明的一种方式中，运行记录数据至少包含写入数据、写入位置和作为标识信息的恢复标志信息。登记装置通过设置运行记录数据中的指定的恢复标志信息登记恢复可能时间点。

追加恢复标志，并扩展运行记录数据的数据构造。在所有的运行记录数据中预先包含有设置恢复标志的数据区域。对某数据设定恢复可能时间点时，设置对应该数据的恢复标志。如消除恢复标志，能够解除设定的恢复可能时间点。

在本发明的一种方式中，进一步，存储装置包括存储备份数据的备

份数据存储装置。控制装置包括运行记录数据管理装置。另外，运行记录数据管理装置在运行记录数据存储装置的未用容量不足时，将被运行记录数据存储装置存储的最旧的运行记录数据移换到备份数据存储装置，使运行记录数据存储装置的未用容量增加，且将在被登记的恢复可能时间点中的最旧的恢复可能时间点变更的意旨通知给主机。

数据的恢复通过例如将到达作为目标的时间点的运行记录数据顺序反映到某时间点的备份数据上来实现（滚动前进方式）。因此，当运行记录数据不存在时，只能将数据返回到被备份的时间点。另一方面，运行记录数据是数据更新经历的集合体，不断增大。运行记录数据的保存量达到磁盘装置的存储容量后，就不能再存储运行记录数据。因此，当运行记录数据的未用容量不足时，将已经蓄积的运行记录数据中最旧的数据只按必要量移换到的备份数据上，以确保未用容量。移换的必要量，可以是预先设定的固定值，也可以根据运行记录数据的蓄积速度及备份数据存储装置的存储容量等诸因素使其动态变化。在此，所谓将最旧的运行记录数据移换到备份数据上，是指在将最旧的运行记录数据反映到备份数据上之后，删除最旧的运行记录数据的意思。另外，只要存储装置内有未使用的存储区域，就可以自动扩展运行记录数据存储区域，当未使用的存储区域不足时，将最旧的运行记录数据移换到备份数据上。

本发明的其他的外部存储装置的数据恢复方法，特征在于：是将与主机连接的外部存储装置的数据在该外部存储装置内恢复的数据恢复方法，包含：对被存储的数据，登记由主机在任意的多个时间点上能够设定的恢复可能时间点的登记步骤；根据来自主机的要求，将被登记的恢复可能时间点的选择用信息发送给主机的一览发送步骤；根据恢复可能时间点的选择用信息，将被主机指定的数据恢复到被指定的恢复可能时间点的恢复步骤。

登记步骤、一览发送步骤、恢复步骤，可以此顺序执行，也可按不同顺序执行，如并行执行。

基于本发明其他观点的程序，是用于控制与主机连接的外部存储装置的程序。外部存储装置包括：存储被主机利用的数据的存储装置；对

被存储装置存储的数据，登记由主机在任意的多个时间点上能够设定的恢复可能时间点的登记装置；根据主机的要求，将被登记的恢复可能时间点的选择用信息发送给所述主机的选择用信息发送装置；根据恢复可能时间点的选择用信息，将由主机指定的数据恢复到被指定的恢复可能时间点的恢复装置。该程序将这些装置在外部存储装置的计算机上实现。

基于本发明其他观点的程序，是控制利用外部存储装置的主机的程序，将以下装置即：将对被外部存储装置存储的数据在任意的多个时间点上可能设定的恢复可能时间点指示登记在外部存储装置的登记指示装置；要求被外部存储装置登记的恢复可能时间点的选择用信息的选择用信息要求装置；根据从外部存储装置接收的选择用信息，为将所希望的数据恢复到所希望的恢复可能时间点而指示给外部存储装置的恢复指示装置在主机上实现。

该程序，如可以 API (Application Program Interface) 的形式提供，也可以适合被各种业务用应用程序利用。

基于本发明的程序，可以固定放在如磁盘型存媒体、半导体存储器等的各种存储媒体上流通使用，或也可从服务器通过通信网络配信。

附图说明

图 1 是涉及本发明第 1 实施方式的外部存储系统的概略构成图。

图 2 是表示图 1 所示存储装置系统的概略的模块图。

图 3 是表示运行记录数据及写入控制信息的构造的数据构造图。

图 4 是表示主机及磁盘控制装置的程序构造的模块图。

图 5 是表示写入控制处理的流程图。

图 6 是表示运行记录磁盘管理处理的流程图。

图 7 是表示被主机通知恢复契机时的数据恢复控制处理的流程图。

图 8 是表示被主机要求恢复契机目录发送时的数据恢复控制处理的流程图。

图 9 是表示被主机指示恢复时的数据恢复处理的流程图。

图 10 是表示用多代进行数据管理时的模式图。

具体实施方式

以下根据图 1~10 说明本发明的实施方式。

首先根据图 1 说明外部存储系统的整体概要。

先根据图 1 说明系统的整体构成。存储装置系统 60 包括存储设备控制装置 10 和存储设备 30 而构成。存储设备控制装置 10 根据从信息处理装置 20 接收的指令，进行对存储设备 30 的控制。如，存储设备控制装置 10 从信息处理装置 20 接收数据的输出/输入要求后，进行存储在存储设备 30 的数据输出/输入处理。在由存储设备 30 包括的磁盘驱动提供的物理存储区域上，设定有逻辑卷（Logical Unit）（以简称 LU）。LU 是逻辑存储区域，数据被存储在此 LU 上。另外，存储设备控制装置 10 在与信息处理装置 20 之间也进行用于管理存储装置系统 60 的各种指令的接收发送。

信息处理装置 20 是包括 CPU（Central Processing Unit）和存储器等的计算机系统。通过信息处理装置 20 的 CPU 执行各种程序，实现各种功能。信息处理装置 20，可以是如个人计算机和终端站，也可是主机。在图 1 中，为方便说明，用图说明第 1~第 5 的 5 台信息处理装置。为识别各信息处理装置 20，在图 1 中如「信息处理装置 1」、「信息处理装置 2」这样付以连续号码，作为第 1~第 5 的信息处理装置 20。后述的通道控制部 11 及磁盘控制部 14 也同样付以连续号码加以区别。

第 1~第 3 的信息处理装置 20，通过 LAN（Local Area Network）40 与存储设备控制装置 10 连接。LAN40，如可为因特网，也可为专用的网络。第 1~第 3 的信息处理装置 20 与存储设备控制装置 10 间的数据通信，通过 LAN40，按照如 TCP/IP（Transmission Control Protocol/Internet Protocol）协议进行。从第 1~第 3 的信息处理装置 20 向存储装置系统 60 发送由文件名指定发出的数据访问要求（是在文件单位的数据输出/输入要求。以下简称「文件访问要求」）。

备份设备 71 被连接在 LAN40 上。作为备份设备 71 可以采用如 MO（magneto-optic:光磁型存储装置）CD-R（CD-Recordable: 可以读写的微型光盘）、DVD-RAM（Digital Versatile Disk-RAM:可以读写的 DVD）等磁盘系列存储设备，及如 DAT（Digital Audio Tape）磁带、盒式磁带、

开式磁带、卡式磁带等磁带系列存储设备。备份设备 71 通过 LAN40 与存储设备控制装置 10 之间进行通信，由此存储被存储设备 30 存储的数据的备份数据。另外，备份设备 71 可与第 1 信息处理装置 20 连接构成。这种时候，可以通过第 1 信息处理装置 20 取得被存储设备 30 存储的数据的备份数据。

存储设备控制装置 10 由第 1～第 4 通道控制部 11 通过 LAN40 在第 1～第 3 信息处理装置 20 及备份设备 71 间进行通信。第 1～第 4 通道控制部 11 各自受理来自第 1～第 3 信息处理装置 20 的文件访问要求。即 LAN40 上的网络地址（如 IP 地址）分别被分配给第 1～第 4 通道控制部 11。第 1～第 4 各通道控制部 11 分别单独作为 NAS（Network Attached Storage）行动。第 1～第 4 通道控制部 11 就像是各自独立的 NAS，能够向第 1～第 3 信息处理装置 20 提供 NAS 的服务。以下有时将第 1～第 4 通道控制部 11 简称 CHN。这样，通过在 1 台存储装置系统 60 内包括分别提供 NAS 的服务的第 1～第 4 通道控制部 11 的构成，以往在独立的计算机上各自运用的 NAS 服务器被集约成 1 台存储装置系统 60。另外，由此，存储装置系统 60 的集中运用成为可能，能够谋求各种设定・控制及障碍管理、版本管理等维护业务的效率化。

再有，存储设备控制装置 10 的第 1～第 4 通道控制部 11，通过例如在一体化电路板上形成的硬件、由该硬件执行的 OS（Operating System）、在 OS 上运行的应用程序等的软件来实现。在存储装置系统 60 中，作为以往硬件的一部分被安装的功能通过软件来实现。因此，通过采用存储装置系统 60，富于灵活性的系统运用成为可能，能够细致入微地应对多样、变化快的用户需求。

第 3 及第 4 信息处理装置 20，通过 SAN（Storage Area Network）50 与存储设备控制装置 10 连接。SAN50 是将在存储设备 30 提供的存储区域的数据的管理单位信息块作为单位，用于在第 3 及第 4 信息处理装置 20 间进行数据接收/发送的网络。通过 SAN50 进行的第 3 及第 4 信息处理装置 20 与存储设备控制装置 10 间的通信，一般服从光纤通道协议。根据光纤通道协议信息块单位的数据访问要求（以下简称信息块访问要

求)从第3及第4信息处理装置20发送给存储装置系统60。

SAN对应的备份设备70被连接在SAN50上。SAN对应备份设备70通过SAN50与存储设备控制装置10进行通信,由此存储被存储设备30存储的数据的备份数据。

存储设备控制装置10,由第5及第6通道控制部11通过SAN50与第3及第4信息处理装置20及SAN对应备份设备70间进行通信。以下有时将第5及第6通道控制部11简称CHF。

另外,第5信息处理装置20,不通过LAN40及SAN50等与存储设备控制装置10直接连接。作为第5信息处理装置20,例如,可作为主机,但当然不能仅限于此。第5信息处理装置20与存储设备控制装置10间的通信,服从如FICON(Fibre Connection)(注册商标)、ESCON(Enterprise System Connection)(注册商标)、ACONARC(Advanced Connection Architecture)(注册商标)FIBARC(Fibre Connection Architecture)(注册商标)等的通信协议。根据这些通信协议信息块访问要求从第5信息处理装置20发送给存储装置系统60。

存储设备控制装置10,通过第7及第8通道控制部11与第5信息处理装置20间进行通信。以下有时将第7及第8通道控制部11简称CHA。

设置在远离(次场)存储装置系统60的设置场所(主场)的另外的处存储装置系统60被连接到SAN50上。另外的存储装置系统61作为复制或远程复制功能的数据复制处的装置被利用。再有,另外的存储装置系统61有时通过SAN50以外的如ATM等通信线路与存储装置系统60连接。这时,采用包括用于利用上述通信线路的接口的通道控制部11。

下面说明存储设备30的构成。存储设备30包括多个磁盘驱动(物理磁盘),向信息处理装置20提供存储区域。数据存储在逻辑存储区域LU。作为磁盘驱动,可以采用例如硬盘装置、软盘装置、半导体存储装置等各种磁盘。再有,存储设备30也可由多个磁盘驱动构成磁盘阵列。这时,可通过由RAID(Redundant Array Of Independent (Inexpensive)Disks)管理的多个磁盘驱动向信息处理装置20提供存储区域。

存储设备控制装置10与存储设备30,如图1所示,可以直接连接,

也可能通过网络间接地连接。进一步，存储设备 30 可以与存储设备控制装置 10 作为一体构成。

在存储设备 30 内设定的 LU 内，有来自信息处理装置 20 可以访问的用户 LU 及用于控制通道控制部 11 而使用的系统 LU 等。在系统 LU 内也存储有由 CHN11 执行的 OS。另外，各通道控制部 11 预先被对应连接到各 LU 上。由此，可以访问的 LU 按各通道控制部 11 分别被分配。另外，上述对应连接也可以设定为在多个通道控制部 11 共有一个 LU。再有，在以下的说明中，有时将用户 LU 表述为用户磁盘，将系统 LU 表述为系统磁盘。另外，将由多个通道控制部 11 共有的 LU 有时表述为共有 LU 或共有磁盘。

下面说明存储设备控制装置 10 的构成。存储设备控制装置 10 包括通道控制部 11、共有存储器 12、高速缓冲存储器 13、磁盘控制部 14、连接部 15 及管理终端 16。

通道控制部 11，且有用于和信息处理装置 20 间进行通信的通信接口，包括与信息处理装置 20 间发送/接收数据输出/输入指令等的功能。例如，CHN11 接收来自第 1～第 3 信息处理装置 20 的文件访问要求，由此，存储装置系统 60 能够将 NAS 的服务提供给第 1～第 3 信息处理装置 20。另外，CHF11 接收来自第 3 及第 4 信息处理装置 20 的服从光纤通道协议的信息块访问要求。由此，存储装置系统 60 能够将可高速访问的数据存储服务提供给第 3 及第 4 信息处理装置 20。另外，CHA11 接收来自第 5 信息处理装置 20 的服从 FICON、ESCON、ACONARC、FIBARC 等通信协议的信息块访问要求。由此，存储装置系统 60 对像第 5 信息处理装置 20 这样的主机等也能够提供数据存储服务。

各通道控制部 11 与管理终端 16 一道通过内部 LAN17 连接。因此，可以将被通道控制部 11 执行的程序等从管理终端 16 发送给通道控制部 11 使其安装。关于通道控制部 11 的构成进一步后述。

连接部 15，相互连接各通道控制部 11、共有存储器 12、高速缓冲存储器 13、各磁盘控制部 14。在通道控制部 11、共有存储器 12、高速缓冲存储器 13 及磁盘控制部 14 间的数据及指令的发送/接收，通过连接部

15 进行。连接部 15 由通过高速开关进行数据传送的超高速纵横式交换器等的高速总线构成。由于每个通道控制部 11 之间用高速总线连接，比起通过 LAN 将在每个计算机上运行的 NAS 服务器连接起来的情况，通道控制部 11 间的通信性能提高。另外，因此使高速文件共有功能及高速故障越过（日文原文：フェイルオーバ）等成为可能。

共有存储器 12 及高速缓冲存储器 13 是由各通道控制部 11 及各磁盘控制部 14 共有的存储器。共有存储器 12 主要用于存储控制信息及指令等。高速缓冲存储器 13 主要用于存储数据。

例如，某通道控制部 11 从信息处理装置 20 接收的数据输出/输入指令是写入指令时，该通道控制部 11 将写入指令写入共有存储器 12 的同时，将从信息处理装置 20 接收的写入数据写入高速缓冲存储器 13。另一方面，磁盘控制部 14 监视共有存储器 12。磁盘控制部 14 检测出写入指令被写入共有存储器 12 后，根据该指令从高速缓冲存储器 13 读取写入数据，将读取的数据写入存储设备 30。

磁盘控制部 14，进行存储设备 30 的控制。例如，如上所述，磁盘控制部 14 根据通道控制部 11 从信息处理装置 20 接收的写入指令，进行向存储设备 30 的数据的写入。另外，磁盘控制部 14 将向由从通道控制部 11 发送的逻辑寻址产生的 LU 的数据访问要求，变换为向由物理寻址产生的物理磁盘的数据访问要求。磁盘控制部 14 在存储设备 30 的物理磁盘被 RAID 管理的时候，进行根据 RAID 构成的数据的访问。另外，磁盘控制部 14 也进行被存储设备 30 存储的数据的复制管理的控制及备份控制。进一步，磁盘控制部 14 以防止发生灾害时数据丢失等（ディザスタリカバリ）为目的，也进行将主场的存储装置系统 60 的数据的复制存储到设置在次场的另外的存储装置系统 61 上的控制（被称为复制功能或远程复制功能）等。

各磁盘控制部 14 同管理终端 16 一道通过内部 LAN17 连接，能够相互进行通信。由此，能够将使磁盘控制部 14 执行的程序等从管理终端 16 发送给磁盘控制部 14 使其安装。

下面说明管理终端 16。管理终端 16 是用于维护・管理存储装置系统

16 的计算机。通过操作管理终端 16，能够进行例如存储设备 30 内的物理磁盘构成的设定、LU 的设定、用于在通道控制部 11 执行的程序的安装等。在此，作为存储设备 30 内的物理磁盘构成的设定，可以例举如物理磁盘的增设及减设、RAID 构成的变更（从 RAID1 到 RAID5 的变更）等。进一步，从管理终端 16 也可以进行存储装置系统 60 的工作状态的确认及故障部位的确定、在通道控制部 11 执行的 OS 的安装等的操作。另外，管理终端 16 通过 LAN 及电话线等与外部维护中心连接，从外部维护中心利用管理终端 16 进行存储装置系统 60 的故障监视，在故障发生时能迅速应对。故障的发生由如 OS 及应用程序、驱动软件等通知。该通知可以通过如 HTTP (Hyper Text Transfer Protocol) 指令及 SNMP (Simple Network Management Protocol) 指令、电子邮件等进行。这些设定及控制，可以将在管理终端 16 上工作的网络服务器提供的网页作为用户接口通过操作员等的操作进行。操作员等操作管理终端 16，设定故障监视对象及内容，设定故障通知处等。

管理终端 16 可以内置在存储设备控制装置 10 内的构成，也可以外挂在存储设备控制装置 10 上的构成。另外，管理终端 16 也可以由专门进行存储设备控制装置 10 及存储设备 30 的维护管理的计算机构成，或者是由包括维护管理功能的用途广泛的计算机构成。

下面参照图 2 说明基于本发明的数据恢复方法的一种实例。图 2 是抽出图 1 所述存储装置系统主要部分的概略构成图。在图 2 示出的外部存储系统，分别如后面所述，大致区分为主机 10 和外部存储装置。外部存储装置大致区分为磁盘控制装置 200 和大容量存储装置 400。在此，简单说明图 1 和图 2 的对应关系：图 1 中的存储装置系统 60 与图 2 中的磁盘控制装置 200 对应，图 1 中的通道控制部 11 与图 2 中的通道口 210 及微处理器 220 对应，图 1 中的共有存储器 12 及高速缓冲存储器 13 与图 2 中的缓冲存储器 230 对应，图 1 中的连接部 15 与总线及开关类等对应(图未示)，图 1 中的磁盘控制部 14 与图 2 中的微处理器 220 对应，图 1 中的存储设备 30 与图 2 中的存储装置 400 对应，图 1 中的信息处理装置 20 与图 2 中的主机 100 对应。微处理器 220 可以存在于通道控制部 11 或磁

盘控制部 14 的任何一侧。

主机 100 是由如个人计算机及工作台等构成的，拥有处理数据库的应用程序 110（以下简称应用程序）。另外，虽省略了图示，但是主机 100 还包括通过如指向设备、键盘开关、监视器等用于和操作员进行信息交换的用户接口。应用程序 110 通过磁盘控制装置 200 访问存储装置 400 内的数据，由此处理指定的业务。

磁盘控制装置 200 是控制存储装置 400 的装置，包括通道口 210、微处理器 220 及缓冲存储器 230。微处理器 220 通过通道口 210 与主机 100 进行双方向的数据通信。微处理器 220 执行磁盘控制程序 300。在磁盘控制程序 300 中包含有写入控制处理 310、写入数据处理 320、磁盘管理处理 330、数据恢复控制处理 340、数据恢复处理 350、数据同期处理 360。

关于主要的处理，后面进一步详述。写入控制处理 310 主要是管理数据写入时的写入控制信息（运行记录控制信息）的程序。写入数据处理 320 是进行向指定的磁盘装置的数据写入的程序。磁盘管理处理 330 主要是进行运行记录数据存储磁盘 430 的管理的程序。数据恢复控制处理 340 是将由主机 100 设定的恢复契机的登记和被登记的恢复契机的目录数据发送给主机 100 的程序。数据恢复处理 350 是使指定的磁盘装置的数据恢复到指定的时间点的程序。数据同期处理 360 是按照主机的指示进行数据的备份处理的程序。

在缓冲存储器 230 上，存储如恢复数据信息 D10、运行记录数据 D20、写入控制信息 D30、更新数据 D40。恢复数据信息 D10 是数据恢复处理的履历信息，存储如数据恢复处及恢复时间点等。运行记录数据 D20 是数据操作的更新履历，从缓冲存储器 230 被顺序地转移给运行记录存储磁盘 430。写入控制信息 D30 包含有为在任意时间点使数据恢复的必要的信息。更新数据 D40 是由应用程序 110 指示更新的数据，从缓冲存储器 230 被转移到数据存储磁盘 410。再有，以上数据不必同时存在缓冲存储器 230 上。另外，为便于说明，将缓冲存储器 230 作为单一的存储器来表示，但是，其可作为例如多个种类的存储器的集合体而构成。

大容量存储装置 400 包括数据存储磁盘 410、备份数据存储磁盘 420

及运行记录数据存储磁盘 430。在数据存储磁盘 410 上存储有当前使用的最新数据（现实数据）。在备份数据存储磁盘 420 上存储有某时间点的备份数据。在运行记录数据存储磁盘 430 上存储运行记录数据。再有，各磁盘 410～430 正确地说是磁盘装置，分别包括多个磁盘。以下将数据存储磁盘称为数据磁盘，将备份数据存储磁盘称为备份磁盘，将运行记录数据存储磁盘称为运行记录磁盘。

图 3 是表示运行记录数据 D20 及写入控制信息 D30 的概略构造图。

由本实施方式产生的运行记录数据 D20 包含有写入控制信息 D30 及更新数据（写入数据）D40。写入控制信息 D30 是发挥作为运行记录控制信息的功能的信息，包含有如数据写入位置 D31、数据大小 D32、时间标记 D33、恢复标志 D34、其他控制信息 D35 等的信息。数据写入位置 D31 是指示数据被写入哪个磁盘的哪个地方的位置信息。数据大小 D32 是指示被写入数据大小的信息。时间标记 D33 是指示数据写入时刻的信息。恢复标志 D34 是指示恢复可能时间点（恢复点）的标识信息，设置恢复标志 D34 后，作为可以恢复的数据被设定，消除恢复标志 D34，恢复点的设定被解除。在其他的控制信息 D35 中包含有例如专门用于特定写入控制信息 D30 的控制号码及数据类别等的其他必要信息。

在本实施方式中，如图 3 所示，独自扩展运行记录数据 D20 的构造，在运行记录数据 D20 内设有恢复标志 D34。由此，通过只追加少量的数据就能将任意的时间点作为恢复可能时间点自由地设定，能在任意的时间点恢复数据。但不仅限于此，也可以是分离运行记录数据 D20 和恢复标志 D34，用独特的 ID（识别码）等将两者对应连接的构成。

下面，图 4 是表示主机 100 及磁盘控制装置 200 的程序构造的概略的模块图。

应用程序 110 通过主机 100 的 OS120 与磁盘控制程序 300 进行双向数据通信。OS120 包括 API（Application Program Interface）群 130。在 API130 群中包含有数据写入用 API131、恢复契机通知用 API132、恢复契机目录取得要求用 API133、恢复指示用 API134。应用程序 110 通过适时调用这些 API131～134，能够将所希望的时间点作为恢复契机设定，

读出设定完的恢复契机目录，选择所希望的时间点，指示数据的恢复。

参照图4简单说明整体操作。应用程序110通过数据写入用API131向磁盘控制装置200指示数据更新要求后(S1)，磁盘控制程序300的写入控制处理310通过写入数据处理320将数据写入指定的磁盘，并将处理更新要求的意旨通知应用程序110(S2)。

应用程序110在业务处理过程中，能够例如定期或不定期地将所希望的时间点作为可能恢复的时间点的恢复契机(恢复点)设定。应用程序110通过调用恢复契机通知用API132，将设定恢复契机的数据指示给磁盘控制装置200(S3)。恢复契机被通知后，磁盘控制程序300的数据恢复控制处理340设置被指定的数据的恢复标志，并将恢复契机被设定的意旨通知给应用程序110(S4)。

根据故障发生等的主要原因恢复数据时，应用程序110调用恢复契机目录取得要求用API133，向磁盘控制装置200要求可能恢复的时间点的目录信息(S5)。要求目录后，数据恢复控制处理340检查运行记录磁盘430，取得设置恢复标志的数据信息，制成恢复契机目录。数据恢复控制处理340将恢复契机目录返送给应用程序110(S6)。

应用程序110参照存储在存储器140中的恢复契机目录，至少选择1个希望恢复的时间点。应用程序110通过调用恢复指示用API134，向磁盘控制装置200发出使指定磁盘的数据恢复到所希望的时间点的指示(S8)。数据恢复处理350接收到从应用程序110发来的恢复指示后，使用备份磁盘420及运行记录磁盘430，使被指定的数据恢复到被指定的时间点。恢复处理350将恢复处理结束的意旨通知给应用程序110(S9)。

接下来参照图5～图9说明各部的详细控制。首先，图5是写入控制处理的流程图。而且，以下的说明也同样，附图所示的流程图是为理解发明而表示操作主要部分的图示，可能与实际的程序不同。图中将「步骤」简略为「S」。

应用程序110提出写入要求后，在缓冲存储器230上的数据D40被更新(S21)的同时，缓冲存储器230上的写入控制信息D30被更新(S22)。接着，判断运行记录磁盘430是否有充足的未用容量(S23)。例如，能

判断运行记录磁盘 430 当前未用容量是否超过之后将要写入的数据的数据大小。当运行记录磁盘 430 的未用容量不足时 (S23: NO) , 与图 6 同样, 执行后述的运行记录磁盘管理处理以确保未用容量 (S24), 必要时更新缓冲存储器 230 上的写入控制信息 (S25)。所谓的必要时, 是指例如后述的通过运行记录自动扩展, 运行记录数据的写入位置变动等的时候。

当运行记录磁盘 430 存在充足的未用容量时 (S23: YES) 及运行记录磁盘 430 被确保包括充足未用容量时, 将写入数据 D40 及写入控制信息 D30 (即运行记录数据 D20) 追加写入运行记录磁盘 430 (S26)。并且将缓冲存储器 230 上的写入数据 D40 写入数据磁盘 410 的指定位置 (S27), 并将数据写入结束的意旨通知给主机 100 (确切地是主机 100 上的应用程序 110。下同) (S28)。

另外, S26 及 S27 也可以在与本写入控制处理不同的另外契机 (非同期) 进行。那时, 可通过例如在缓冲存储器上的该数据上设置是否向磁盘反映的标志来进行管理。

之后, 判断备份更新标志是否接通 (S29)。所谓备份更新标志, 是表示为确保运行记录磁盘 430 的未用容量而将最旧的运行记录数据移换到备份磁盘 420 的标识信息。通过运行记录数据的移换, 从备份数据变更恢复可能的最旧的时间点, 所以当备份更新标志被设置为接通状态时 (S29: YES), 将备份数据被更新的意旨通知给主机 100 (S30)。将备份更新通知给主机 100 后, 使备份更新标志复位到断开状态 (S31)。

下面, 图 6 是表示图 5 中的运行记录磁盘管理处理 S24 的详细内容的流程图。首先, 判定运行记录磁盘 430 的自动扩展方式是否被设定 (S41), 所谓自动扩展方式, 是指搜索未使用的磁盘、未使用的存储区域, 自动扩展运行记录磁盘 430 的逻辑容量的方式。

当自动扩展方式未被设定时 (S41: NO), 选择存储在运行记录磁盘 430 的运行记录数据中最旧的数据, 使其反映在备份磁盘 420 上 (S42)。被移换到备份磁盘 420 上的最旧的运行记录数据被从运行记录磁盘 430 上删去 (S43)。由此运行记录磁盘 430 的未用容量增加。到运行记录磁

盘 430 的未用容量达到规定值，从最旧的运行记录数据开始顺序地移换到备份磁盘 420 (S44)。当运行记录磁盘 430 的未用容量达到规定值时 (S44: YES)，将备份更新标志设置为接通状态 (S45)。由此，如图 5 中 S30 所示，备份数据被更新，可能恢复的最旧的时间点从备份数据被变更的意旨被通知给主机 100。且 S44 中的规定值，可以是预先设定的固定值，也可以是例如根据备份磁盘的未用容量及被写入数据磁盘 410 的数据大小等动态变化的值。

另一方面，当运行记录磁盘 430 的自动扩展方式设定时 (S41)，从连接的磁盘装置中检索未使用的存储区域（称为未使用区域），判断保存运行记录数据可能的未使用区域是否存在。(S46, S47)。未使用区域未被发现时 (S47: NO)，移给 S42，如上所述，通过将最旧的运行记录数据移换给备份磁盘 420，以确保运行记录磁盘 430 上的未用容量。当未使用区域被发现时 (S47: YES)，在将被发现的未使用区域作为运行记录磁盘利用，扩展运行记录磁盘 430 的逻辑容量的同时，更新磁盘管理映像 (S48)。然后，判断由运行记录磁盘 430 的逻辑容量扩展产生的未用容量是否达到规定值 (S49)，到运行记录磁盘 430 的未用容量达到规定值为止，一边反复进行 S46~S49 的处理，一边将未使用区域作为运行记录数据的存储区域自动扩展。

接下来，图 7 表示来由主机 100 指示的恢复契机的登记处理。如上所述，在本实施方式，主机 100 能够将任意的时间点作为可能恢复的契机（恢复点）进行多个设定。

可以登记的恢复契机从主机 100 被通知给磁盘控制装置 200 后，数据恢复控制处理 340 就检索存储在运行记录磁盘 430 的最新数据的位置 (S51)，将对应最新写入数据的写入控制信息中的恢复标志设置为接通状态进行更新 (S52)。然后，向主机 100 报告恢复契机设定结束的意旨的同时，通知用于特定写入控制信息的控制号码 (S53)。这样，主机 100 的应用程序 110 在数据写入时，对任意的时间点的数据能够设定指示恢复契机。

接下来，图 8 表示根据来自主机 100 的要求，送回恢复契机目录信

息的恢复契机目录的发送处理。首先，在运行记录磁盘 430 内选择对应被主机 100 指定恢复的数据的磁盘，将指示指向在被选择的磁盘中的最旧的运行记录数据上（S61）。

之后，从最旧的运行记录数据读入（S62），检查涉及读入的运行记录数据的写入控制信息中的恢复标志是否被设置为接通状态（S63），当恢复标志被设置时，将读入的运行记录数据追加记录在恢复契机的目录信息上（S64）。直到读出存储在由 S61 选择的磁盘上的最终数据，上述 S62～S64 反复进行。（S65）。这样，将对应被指定数据的运行记录数据从最旧的数据到最新的数据顺序检查，抽出被设置恢复标志的运行记录数据，生成恢复契机目录。生成的恢复契机目录，与结束报告一同或不同期发送给主机 100（S66）。

接下来，图 9 表示数据恢复处理。主机 100 上的应用程序 110，根据由图 8 所示的处理所取得的恢复契机的目录信息，能够指示到所希望的时间点的数据恢复。

从主机 100 通知恢复指示后，数据恢复处理 350 在备份磁盘 420 及运行记录磁盘 430 中分别选择对应恢复被指定的数据的磁盘（S71）

接着，判断从主机 100 作为数据恢复处被指定的磁盘是否是备份磁盘 420（S72）。总之，在本实施方式中，在备份磁盘 420 以外的其他磁盘装置能够恢复到被指定的时间点的数据。作为恢复处被指定的磁盘装置是备份磁盘 420 以外的其他磁盘装置时，将存储在备份磁盘 420 的备份数据复制到指定的磁盘装置，完成成为数据恢复基础的备份数据的准备。（S73）。

接着，从运行记录磁盘 430 检索最旧的运行记录数据（S74），从最旧的运行记录数据顺序地读出数据，将其反映到指定恢复处的磁盘的存储内容上（S75）。到数据恢复到主机 100 指定的时间点为止，读出运行记录数据，更新恢复处磁盘的存储内容（S76）。

当数据恢复到被指定的时间点时，将数据恢复结束的意旨通知给主机 100（S77）。另外，将恢复时间点及恢复处信息记录在恢复数据信息 D10 上（S78）。

通过本实施方式，因在外部存储装置内自动进行数据恢复，所以不必为数据恢复处理消费主机 100 的计算机资源，不会降低主机 100 上的其他的业务处理的效率。特别是，在采用大容量的外部存储装置的应用程序 110 中，因为处理大规模的数据，所以数据恢复处理的负担变大，大量地消费计算机的资源。因此，在主机 100 上进行的其他业务的处理速度降低，而且到数据恢复结束的处理时间也变长。但是，在本实施方式中，因是采用在主机 100 上只执行恢复契机的设定指示、恢复契机目录的取得要求及恢复指示这些仅有的处理，将实际的数据恢复处理委托给外部存储装置这样的构成，能够减轻主机 100 的负担。在外部存储装置进行数据恢复期间，主要 100 可以有效地处理其他业务。

另外，可以将任意的多个时间点作为恢复契机设定，可以将数据恢复到所希望的时间点，所以与单纯地只将稍前的数据恢复的以往技术不同，很便利。

进一步，在本实施方式中，准备了用于从主机 100 侧进行恢复契机的设定指示及恢复契机目录的取得要求等的 API131～134，正是因为主机包括这些独自的 API，才使利用基于本发明的外部存储装置成为可能。

另外，在本实施方式中，因在外部存储装置内自动收集运行记录数据的同时，进行运行记录磁盘 430 的管理，所以能够预先防止运行记录磁盘 430 存储满而无法进行数据恢复的情况发生。

另外，在本实施方式中，因是扩展运行记录数据 D20 的数据构造，在运行记录数据 D20 内（在作为运行记录控制信息的写入控制信息 D30 内）设定恢复标志的构成，虽是比较简易的构成，但能够实现向任意的多个时间点的数据恢复。

图 10 表示本发明的第 2 种实施方式。在本实施方式中进行多代的数据管理。即，加上保存最新数据的数据磁盘 410，像存储 1 代前数据的 1 代前数据磁盘 410(1GA)、存储 2 代前的数据 2 代前数据磁盘 410(2GA) 等这样，可以用多代管理数据。

例如，在 1 代前数据磁盘 410(1GA) 上恢复备份磁盘 420 记录内容后，读出存储在运行记录磁盘 430 的数据 dB 的运行记录数据，将其反映

到 1 代前数据磁盘 410 (1GA) 上，这样能够返回到 1 代前的数据。同样，在 2 代前数据磁盘 410 (2GA) 上复制备份数据之后，通过反映数据 dB 及数据 dC 的运行记录数据，能够返回到 2 代前的数据。这样，当用多代管理数据时，基于本发明，也能不增加主机 100 的负担在外部存储装置内构筑管理多代的数据。

再有，本发明不限定上述各实施方式，如果是本领域人员，可在本发明的范围内进行各种追加及变更等。

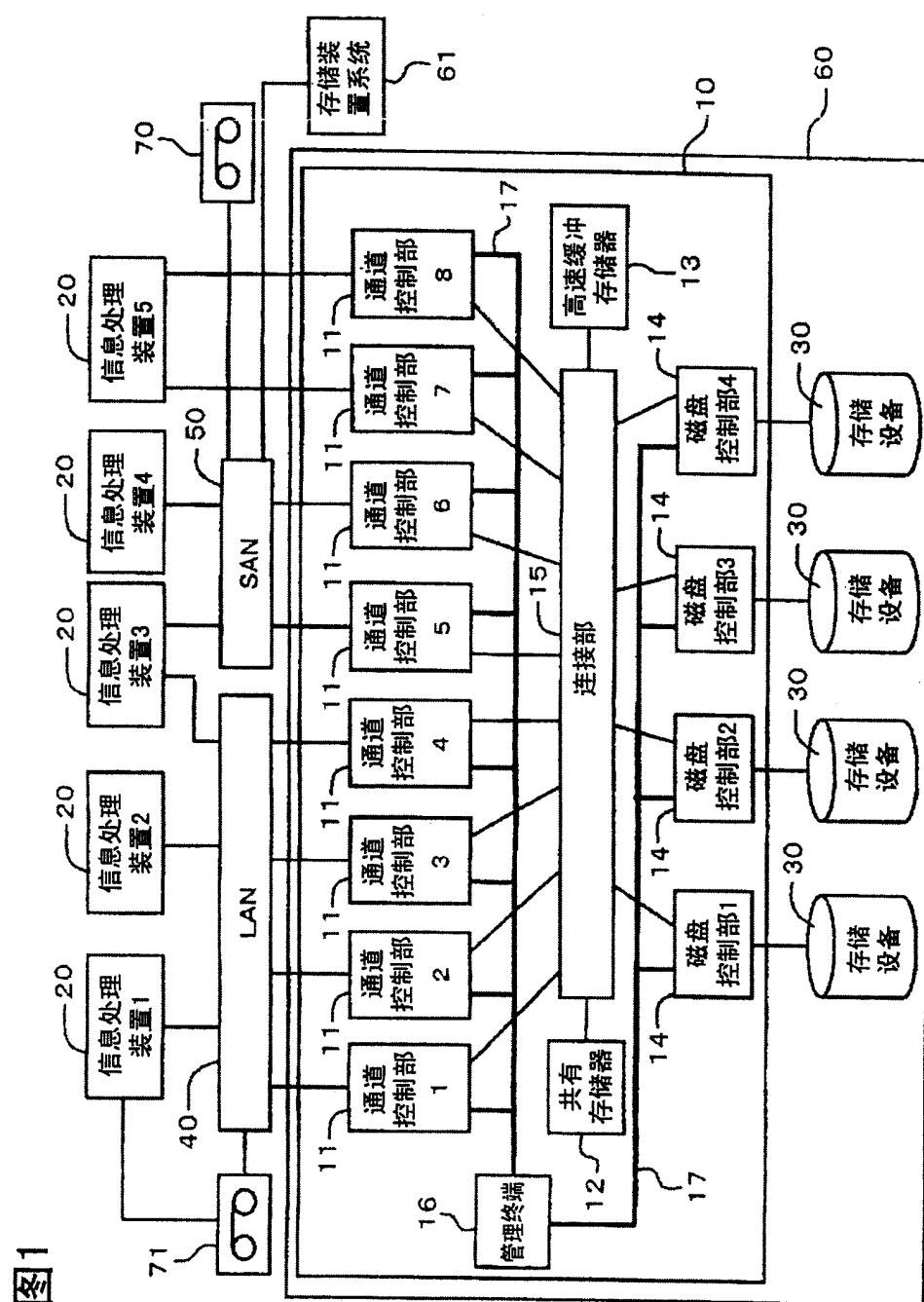


图2

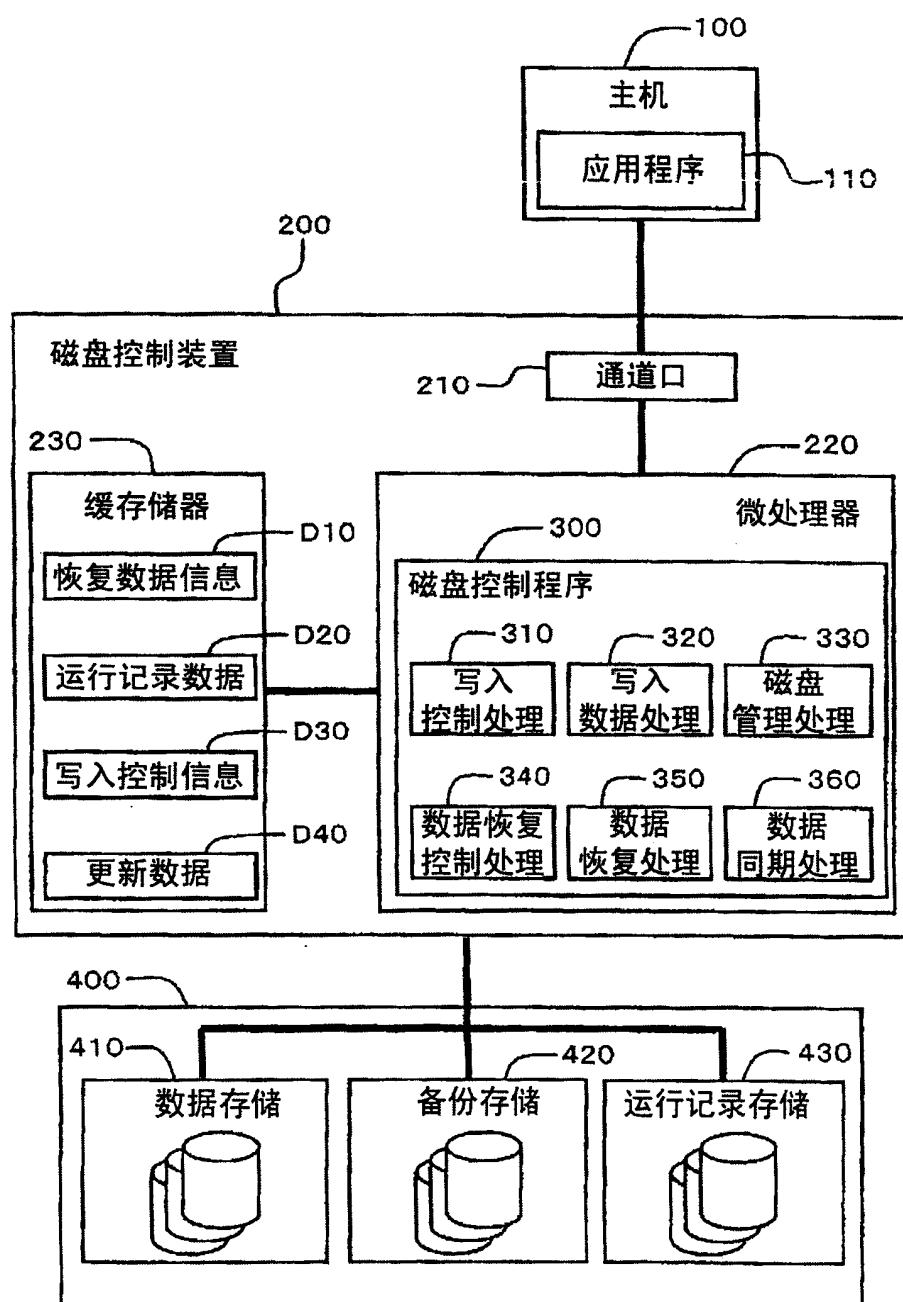
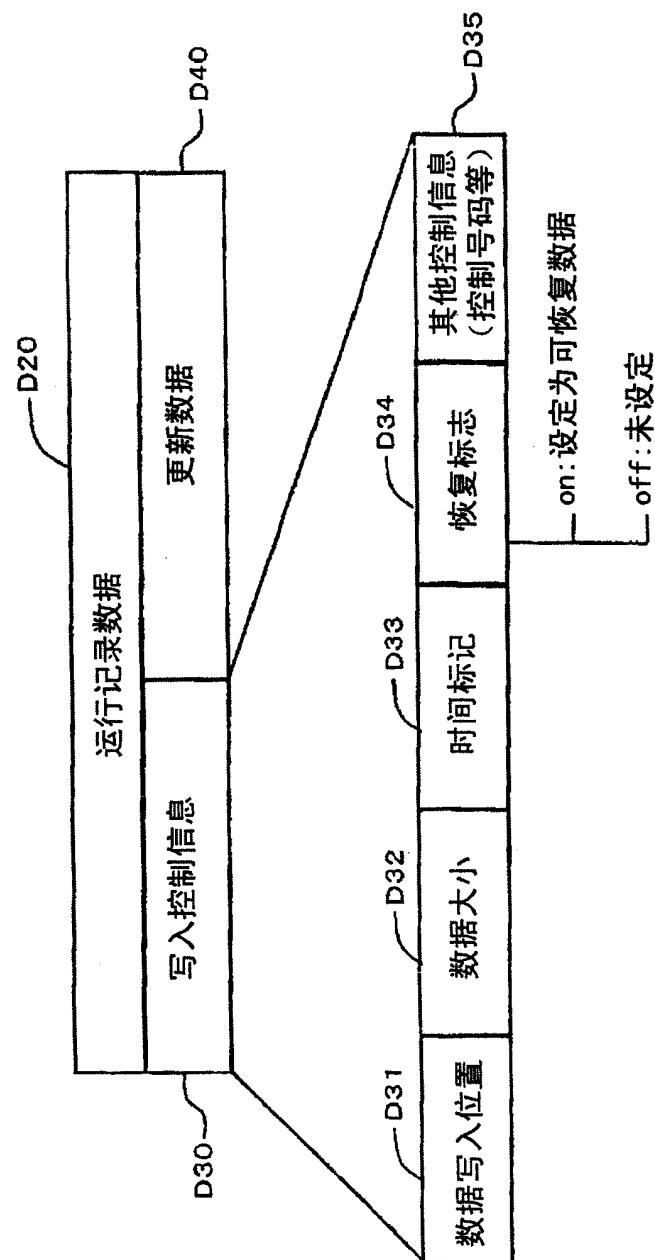
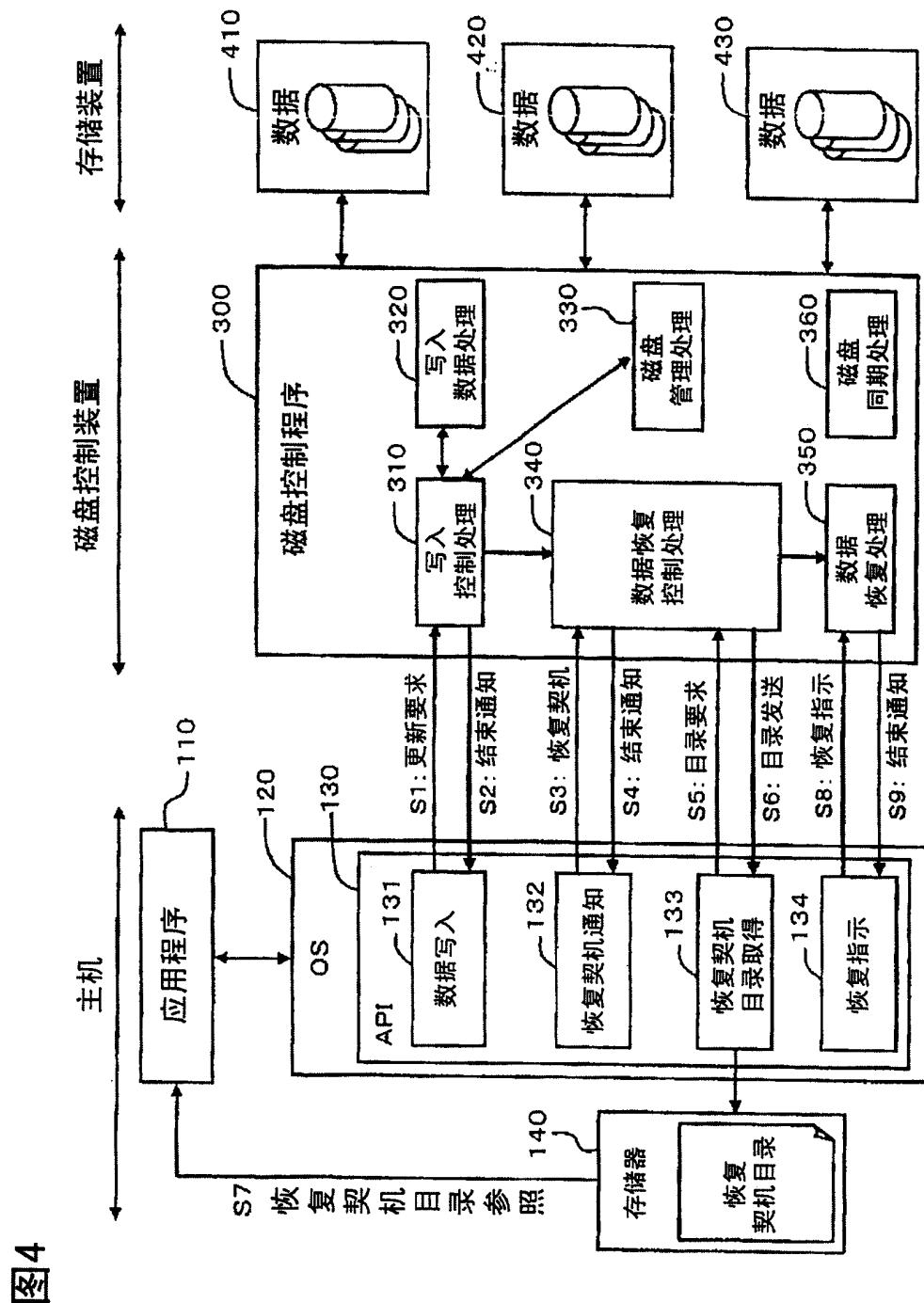
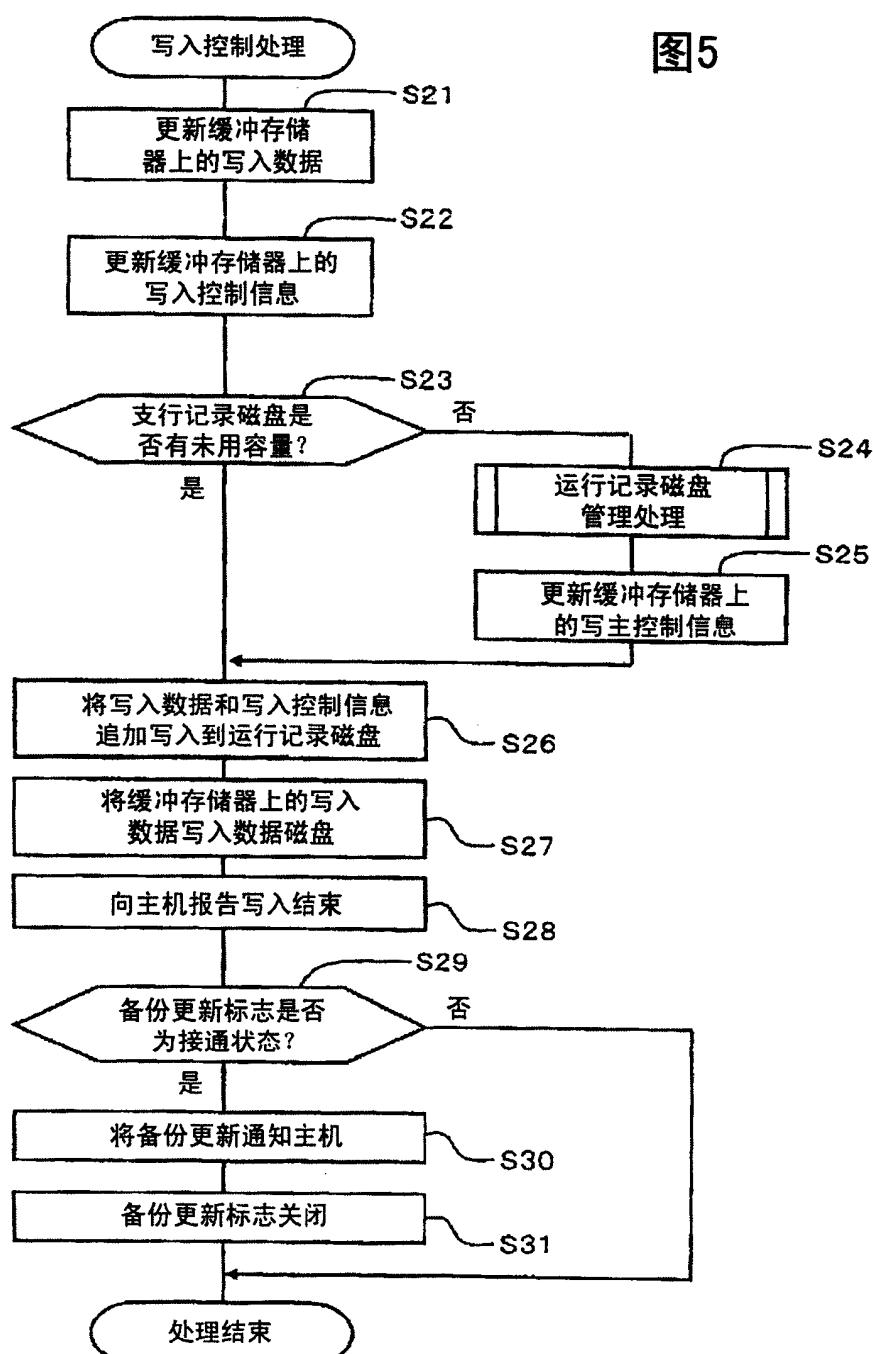
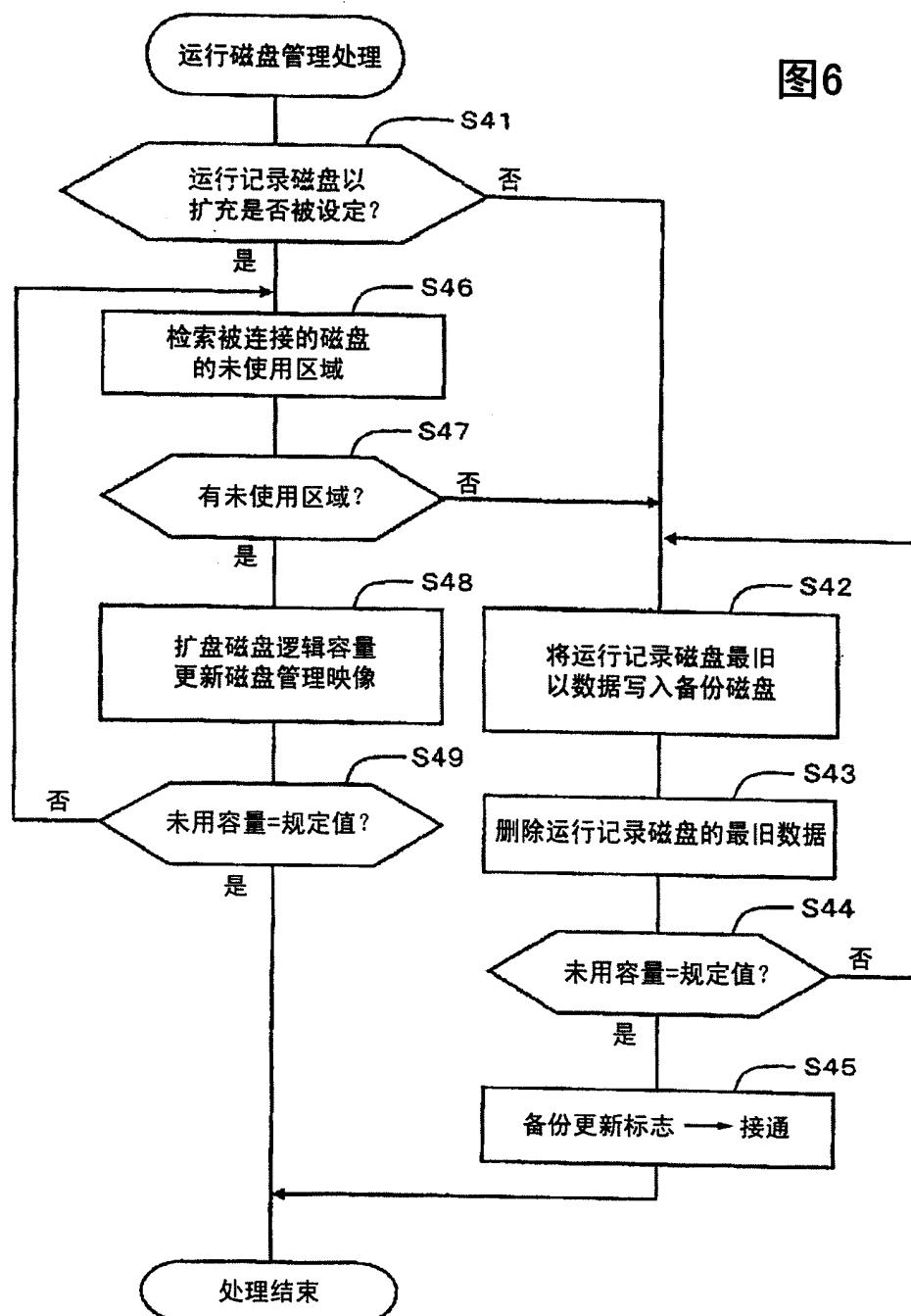


图3









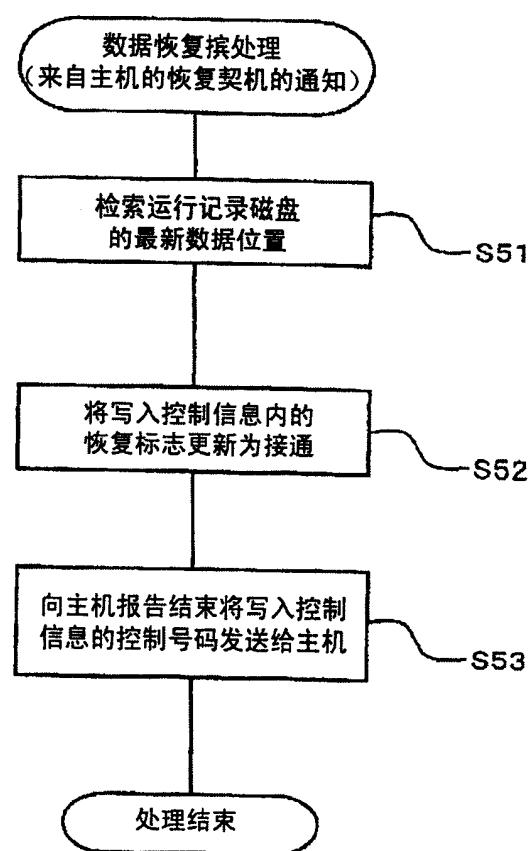


图7

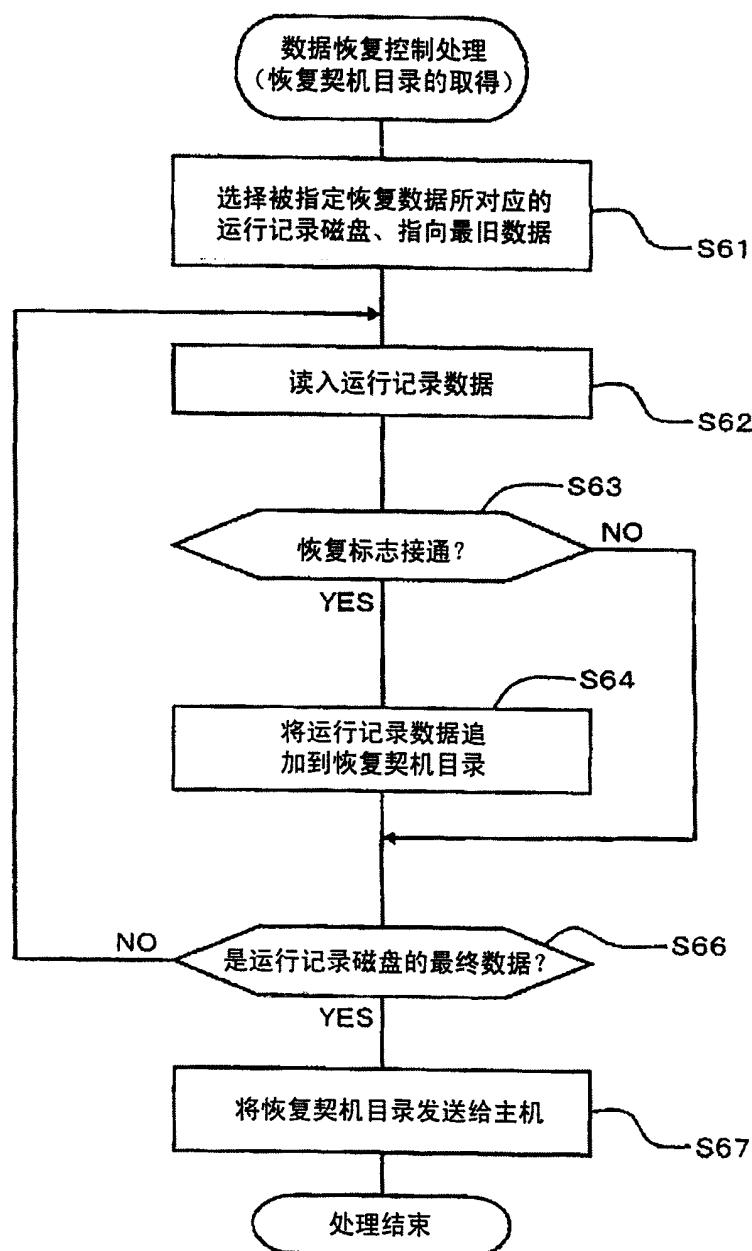


图8

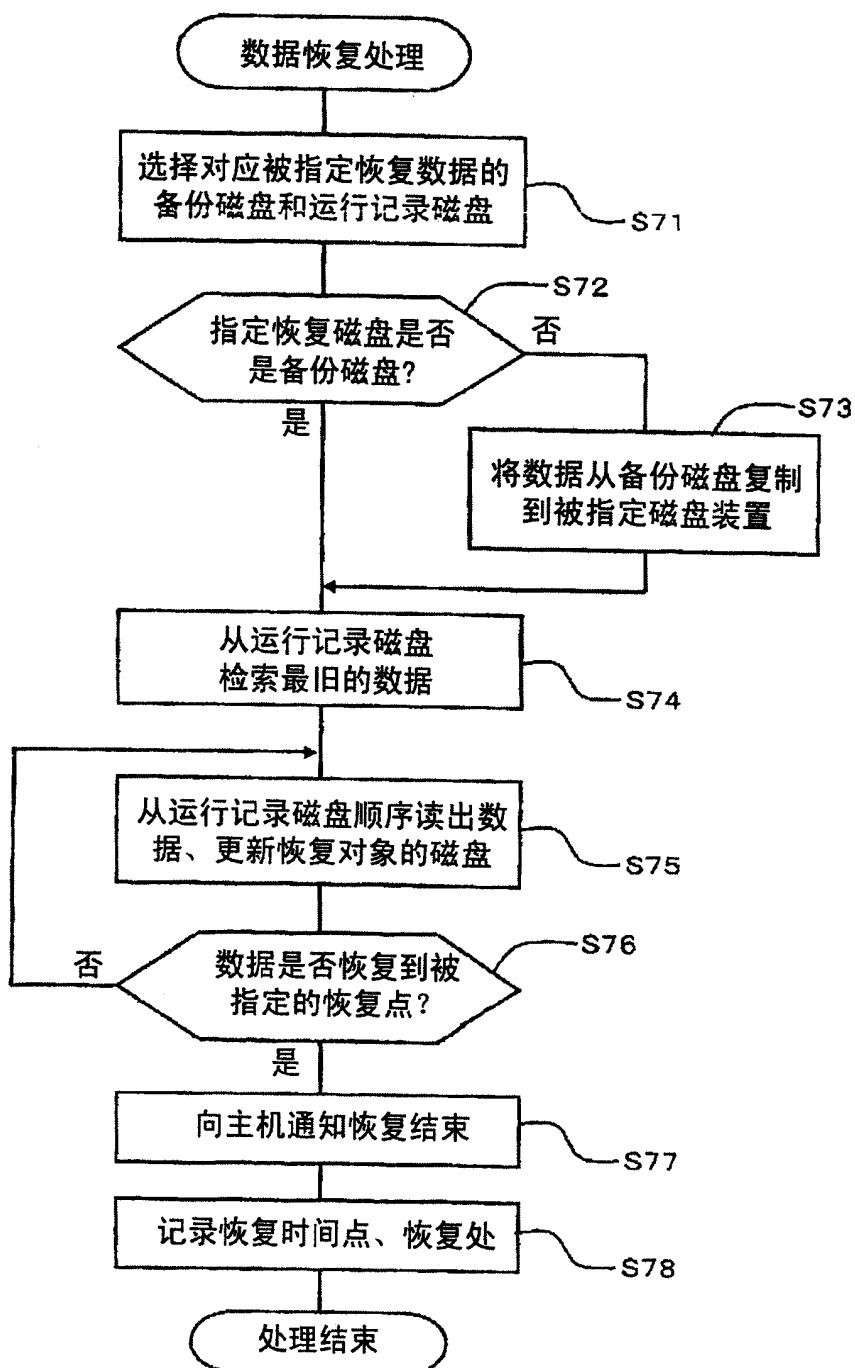


图9

图10

