[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 200610151366.4

[45] 授权公告日 2008 年 10 月 8 日

[11] 授权公告号 CN 100424626C

[22] 申请日 2006.9.7
[21] 申请号 200610151366.4
[30] 优先权
[32] 2005. 9. 9 [33] US [31] 11/223,559
[73] 专利权人 国际商业机器公司
地址 美国纽约
[72] 发明人 维卡斯·阿鲁瓦利亚
斯科特·A.·派博 维普尔·保罗
[56] 参考文献
CN86108426A 1987.7.29
CN1363096A 2002.8.7
US5961613A 1999.10.5
CN1564138A 2005.1.12
US5493670A 1996.2.20
审查员 孙薇薇

[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 李镇江

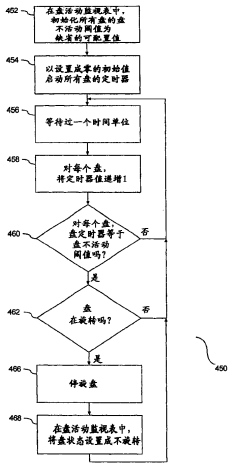
权利要求书 2 页 说明书 10 页 附图 8 页

[54] 发明名称

用于在分布式文件系统中管理功率的方法与系统

[57] 摘要

提供了用于在分布式文件系统中管理单个物理盘旋转状态的方法与系统。 该方法包括：支持多台客户端机器对存储介质的同时访问；及响应数据请求，异步控制所述存储介质中物理盘的旋转状态，其中异步控制所述存储介质中物理盘的旋转状态的步骤包括选自包括停旋不活动物理盘和起旋适于为数据请求提供服务的物理盘的组的命令。 旋转控制消息与 I/O 命令异步地并在由物理盘接收数据请求之前转发到指定的物理盘。 这使得物理盘的旋转状态能够以最小的延迟响应 I/O 命令。



1、一种用于在分布式文件系统中管理功率的方法，包括：

支持多台客户端机器对存储介质的同时访问；及

响应数据请求，异步控制所述存储介质中物理盘的旋转状态，

其中异步控制所述存储介质中物理盘的旋转状态的步骤包括选自包括停旋不活动物理盘和起旋适于为数据请求提供服务的物理盘的组的命令。

2、如权利要求 1 所述的方法，其中所述客户端机器选自包括同构和异构的组。

3、如权利要求 1 所述的方法，还包括在由所述物理盘接收所述数据请求之前起旋所述物理盘。

4、如权利要求 1 所述的方法，还包括响应将数据写到所述存储介质的请求在活动物理盘上分配空间。

5、如权利要求 1 所述的方法，还包括跟踪所述物理盘随时间进行的 I/O 活动。

6、一种计算机系统，包括：

分布式文件系统，具有与至少一台服务器和物理存储介质同时通信的至少两台客户端机器；

管理器，该管理器适于响应与物理盘关联的活动异步控制所述存储介质中所述物理盘的旋转状态，

其中所述管理器适于控制所述物理存储介质的旋转活动，而所述控制选自包括停旋不活动物理盘和起旋适于为数据请求提供服务的物理盘的组。

7、如权利要求 6 所述的计算机系统，其中所述客户端机器选自包括同构和异构的组。

8、如权利要求 6 所述的计算机系统，还包括适于组织所述物理存储介质随时间进行的 I/O 活动的表。

9、如权利要求 6 所述的计算机系统，还包括适于由所述管理器

异步传送到所述物理存储介质的起旋命令。

10、如权利要求9所述的计算机系统，其中所述起旋命令适于在所述数据请求之前由所述物理盘接收。

用于在分布式文件系统中管理功率的方法与系统

技术领域

本发明涉及管理物理存储介质的活动。更具体而言，本发明涉及在支持两个或多个客户端机器对存储介质同时访问的分布式文件系统中控制物理存储介质的运行速度。

背景技术

大部分个人计算机包括形式为至少一个硬盘驱动器的物理存储介质。当个人计算机运行时，一个硬盘消耗个人计算机总功率的20%到30%。在本领域已知管理个人计算机以便当不需要访问硬盘时降低硬盘运行速度到空闲状态及当需要访问硬盘时提高硬盘运行速度的不同技术。硬盘速度的管理使个人计算机能够有更高的运行效率。

图1是分布式文件系统的现有技术方框图（10），该系统包括服务器集群（20）、多个客户端机器（12）、（14）和（16）、存储区域网络（SAN）（30）及独立的元数据存储设备（42）。每个客户端机器在数据网络（40）上与服务器集群（20）中的一个或多个服务器机器（22）、（24）和（26）通信。类似地，每个客户端机器（12）、（14）和（16）与服务器集群（20）中的每个服务器机器与存储区域网络（30）通信。存储区域网络（30）包括仅包含用于所关联文件的数据块的多个共享盘（32）和（34）。类似地，服务器机器（22）、（24）和（26）管理位于元数据存储设备（42）中的关于所关联文件的位置与属性的元数据。每个客户端机器可以访问存储在SAN（30）的文件数据空间（38）上的对象或多个对象，但也可以不访问元数据存储设备（42）。在打开SAN（30）中存储介质上现有文件对象的内容时，客户端机器联系服务器机器中的一个来获得对象元数据与锁。

一般来说，元数据向客户端提供关于文件的信息，例如其属性和在存储设备上的位置。锁向客户端提供其打开文件及读和/或写数据所需的权限。服务器机器在元数据存储设备（42）中执行所请求文件的元数据信息的查找。服务器机器向发出请求的客户端机器传送许可的锁信息和文件元数据，其包括构成文件的所有数据块的地址。一旦客户端机器持有锁并知道了数据块地址，客户端机器就可以直接从连接到SAN（30）的共享存储设备（32）或（34）访问文件数据。系统（10）中包括集群中服务器节点、客户端机器和存储介质的元件的数量仅仅是说明性的数量。该系统可以放大到包括附加元件，而且类似地，该系统可以减小到包括更少的元件。因此，图 1 所示的元件不应当看作是限制因素。

如图 1 所示，所说明的分布式文件系统分别存储元数据与数据。在一种例子中，服务器集群（20）中的一个服务器保留关于共享对象的信息，包括存储设备中客户端可以访问的数据块的地址。为了读共享对象，客户端从服务器获得文件的元数据，包括数据块地址，然后从存储设备给定的块地址读数据。类似地，当向共享对象写时，客户端请求服务器为数据创建存储块地址，然后请求数据将要写入的所分配块地址。元数据可以包括关于对象的大小、创建时间、上次修改时间和安全性属性的信息。

在分布式文件系统中，如图 1 所示的系统，SAN 可以包括形式为盘的多个存储介质。桌面计算机系统中硬盘的功率消耗大约是总系统功率的 20-30%。给定 SAN 中硬盘的数量，很显然有许多系统功率要控制。用于控制与 SAN 中存储介质关联的功率的现有技术方法包括如果盘在设定数量的时间内不使用就将其停旋。当需要访问盘时，盘就起旋，当盘达到合适的速度后，它就准备好接收数据。但是，这种方法涉及当盘从不活动状态变成活动状态时的延迟。存储介质可用性的延迟影响了响应时间和系统性能。在具有多个客户端机器的分布式文件系统和具有多个硬盘的 SAN 中，单个客户端机器不能有效地管理 SAN 中可以由其它客户端机器共享的每个硬盘的功率运行。因此，需

要一种能够有效管理 SAN 中每个硬盘速度和运行而不会严重损害响应时间和系统性能的方法和/或管理器。

发明内容

本发明包括用于解决存储区域网络中可以由多个客户端机器同时访问的物理存储介质的旋转状态的控制的方法与系统。

在本发明的一方面，提供了一种用于在分布式文件系统中管理功率的方法，包括：支持多台客户端机器对存储介质的同时访问；及响应数据请求，异步控制所述存储介质中物理盘的旋转状态，其中异步控制所述存储介质中物理盘的旋转状态的步骤包括选自包括停旋不活动物理盘和起旋适于为数据请求提供服务的物理盘的组的命令。

在本发明的另一方面，提供了一种计算机系统，包括：分布式文件系统，具有与至少一台服务器和物理存储介质同时通信的至少两台客户端机器；管理器，该管理器适于响应与物理盘关联的活动异步控制所述存储介质中所述物理盘的旋转状态，其中所述管理器适于控制所述物理存储介质的旋转活动，而所述控制选自包括停旋不活动物理盘和起旋适于为数据请求提供服务的物理盘的组。

而在本发明的另一方面，提供了具有计算机可用介质的制造物，该计算机可用介质包含了用于管理分布式文件系统中功率的计算机可用程序代码。该程序代码包括支持多个客户端机器对存储介质的同时访问的指令。此外，该程序代码还包括用于响应数据访问请求异步控制存储介质中物理盘的旋转状态的指令。

本发明的其它特征与优点将从以下本发明目前优选的实施方式的结合附图的具体描述中变得显而易见。

附图说明

图 1 是分布式文件系统的现有技术方框图。

图 2 是分布式文件系统中服务器机器与客户端机器的方框图。

图 3 是演示具有存储介质功率管理的读命令处理的流程图。

图 4 是演示具有存储介质功率管理的写命令处理的流程图。

图 5 是演示针对高速缓存数据的并具有存储介质功率管理的写命令处理的流程图。

图 6 是演示用于将逻辑盘区转换成物理盘区的处理的流程图。

图 7 是说明监视表中组件的方框图。

图 8 是说明根据本发明优选实施方式并建议印在所公布专利的首页上的用于监视 SAN 中物理盘的盘活动的处理的流程图。

具体实施方式

概述

如存储区域网络的共享存储介质通常包括多个物理盘。控制共享存储设备中每个物理盘的旋转状态管理功率消耗并使得能够有效管理存储介质。起旋命令可以与读和/或写命令异步地传送到处于空闲状态的单个物理盘，以避免与激活空闲盘关联的延迟。因此，与异步发消息结合的功率管理扩展到了单个物理盘，更具体是扩展到了共享存储系统的单个存储盘的旋转状态。

技术细节

图 2 是跨图 1 的分布式文件系统通信的服务器机器（110）与客户端机器（120）的例子的方框图（100）。服务器机器（110）包括存储器（112）和存储器（112）中的元数据管理器（114）。在一种实施方式中，元数据管理器（114）是管理与文件对象关联的元数据的软件。客户端机器（120）包括存储器（122）和存储器中的文件系统驱动器（124）。在一种实施方式中，文件系统驱动器（124）是用于方便 I/O 请求的软件。存储器（122）向操作系统提供向存储介质读和写数据的接口。在一种实施方式中，如将对对象的访问限定到每次一个客户端的文件系统，元数据管理器可以是文件系统驱动器的一部分。

对文件对象的读或写访问请求称为 I/O 请求。当生成 I/O 请求时，

客户端机器的操作系统负责处理这种请求并负责将该请求重定向到文件系统驱动器（124）。I/O 请求包括以下参数：对象名、读/写的对象偏移及读/写的对象大小。因为对象偏移和对象大小是参照逻辑卷或盘分区上文件对象空间的逻辑相邻映射的，所以它们称为逻辑盘区。通常，逻辑盘区与汇集的物理盘区级连在一起，汇集的物理盘区即计算机文件系统中为文件保留的存储设备的相邻区域。一旦由操作系统接收到 I/O 请求，I/O 请求就转发到管理所关联文件对象的逻辑卷的文件系统驱动器（124）。在一种实施方式中，可以有多个客户端机器，而 I/O 请求指向管理文件对象所驻留在的逻辑卷的文件系统驱动器。将请求从文件系统驱动器（124）传送到元数据管理器（114），元数据管理器（114）将 I/O 文件系统参数转换成以下：盘号、盘偏移读/写和读/写的对象大小。盘号、盘偏移读/写和读/写的对象大小称为物理盘区。因此，文件系统驱动器用于将 I/O 请求的逻辑盘区转换成一个或多个物理盘区。

图 3 是说明用于结合物理存储介质的管理处理分布式文件系统中读请求的处理的流程图（200）。一开始，由客户端机器接收读命令（202）。在接收到读命令后，就进行测试，以确定从读命令请求的数据是否可以从高速缓存的数据提供（204）。如果对步骤（204）的测试的响应是肯定的，则高速缓存的数据拷贝到读命令的缓冲区（206），读命令完成（208）。但是，如果对步骤（204）的测试的响应是否定的，则信息转发到驻留在一台服务器上的元数据管理器，以便将读命令的逻辑 I/O 范围转换成物理存储介质中对应的物理盘盘区（210）。在一种实施方式中，信息从文件系统驱动器传送到元数据管理器。逻辑盘区转换的细节在图 6 中示出。在步骤（210）的转换之后，读命令发布到对应于当前命令逻辑范围的每个物理盘盘区的所有物理盘（212）。在一种实施方式中，为 I/O 命令提供服务的物理盘从元数据管理器接收异步信息，以确保在接收 I/O 命令之前盘处于合适的旋转状态。客户端等待，直到所有发布的盘盘区的读都完成（214）。在步骤（214）所有发布的读都完成后或者在步骤（206）将高速缓存的数

据拷贝到读命令的缓冲区后，读命令完成。因此，如果数据不在高速缓冲存储器中，则文件系统模块中的读与元数据管理器通信，以便获得满足读命令的物理盘盘区。

图4是说明用于结合物理存储介质的管理处理分布式文件系统中写请求的处理的流程图（250）。一开始，由客户端机器接收写命令（252）。在接收到写命令后，就进行测试，以便确定从写命令请求的数据是否可以高速缓存（254）。如果对步骤（254）的测试的响应是肯定的，则数据从写缓冲区拷贝到高速缓冲存储器，对高速缓存数据的指定范围设置脏位（256），不发生盘I/O。在设置脏位的步骤之后，写命令完成（258）。但是，如果对步骤（254）的测试的响应是否定的，则信息转发到驻留在一台服务器上的元数据管理器，以便将写命令的逻辑I/O范围转换成对应的物理盘盘区（260）。逻辑盘区转换的细节在图6中示出。在步骤（260）的转换之后，写命令发布到对应于当前命令逻辑范围的每个物理盘盘区的所有物理盘（262）。在一种实施方式中，为I/O命令提供服务的物理盘从元数据管理器接收异步信息，以确保在接收I/O命令之前盘处于合适的旋转状态。其后，客户端等待，直到所有发布的盘的盘区的写都完成（264）。在步骤（264）所有发布的盘的盘区的写都完成后或者在步骤（256）设置脏位后，写命令完成。因此，如果数据不是要写到高速缓冲存储器而是要直接写到盘，则文件系统模块中的写与元数据管理器通信，以便获得满足写命令的物理盘盘区。

除了图4所示的写处理，还有关于高速缓存数据管理的可选写处理。这种处理由文件系统驱动器以规律的时间间隔调度。图5是说明这种可选写处理的流程图（300）。一开始，进行测试，以便确定是否有任何高速缓存的数据具有脏位设置（302）。对步骤（302）的测试的肯定响应之后是到元数据管理器的信息，以便将脏的高速缓存数据的逻辑I/O范围转换成对应的物理盘盘区（304）。逻辑盘区转换的细节在图6中示出。其后，写命令发布到对应于脏的高速缓存数据当前命令逻辑范围的每个物理盘盘区的所有物理盘（306）。在一种实施方

式中，为 I/O 命令提供服务的物理盘从元数据管理器接收异步信息，以确保在接收 I/O 命令之前盘处于合适的旋转状态。其后，客户端等待，直到所有发布的盘的盘区的写都完成（308），写命令完成。在步骤（308）之后，用于已经冲刷到一个或多个物理盘的高速缓存数据的脏位被清除（310）。如果对步骤（302）的测试的响应是否定的，或者在步骤（310）高速缓存数据的脏位清除之后，在返回步骤（302）以确定脏的高速缓存数据的存在之前，处理等待预定义的可配置时间间隔（312）。因此，图 5 概述的处理是关于高速缓存数据的，更具体而言是关于为脏的高速缓存数据传送从逻辑 I/O 范围到一个或多个物理盘盘区的转换的。

逻辑盘区到物理盘区的转换是由元数据管理器模块处理的。在一种实施方式中，元数据管理器模块是驻留在一台服务器的存储器中的软件组件，如图 2 所示。图 6 是说明根据本发明优选实施方式用于将逻辑盘区转换成物理盘区的处理的流程图（350）。一旦从文件系统模块接收到将逻辑盘区转换成对应物理盘盘区的请求（352），如步骤（210）、（260）和（304）所示，就检查盘区转换表（354）并建立用于该逻辑 I/O 范围的对应物理盘盘区的列表（356）。该盘区转换表是元数据存储设备的一部分。元数据管理器从 SAN 上的元数据存储设备读盘区转换表。其后，从在步骤（356）建立的盘区列表检索物理成员（358），然后向元数据管理器发送带关于正在访问的物理盘的信息的消息（360）。这种信息可以包括 I/O 需要发生时物理盘的地址。然后，进行测试，以便确定来自步骤（360）的物理盘是否在旋转（362）。在一种实施方式中，在集群中一台服务器的存储器中维护盘活动表。盘活动表存储盘的旋转状态及以所设定时间周期监视活动或不活动的定时器。对步骤（362）的测试的否定响应将导致元数据管理器向物理盘发送提高其速度的命令，即起旋（364）。一旦盘在旋转，发出请求的客户端就可以有效地使用该物理盘。在步骤（364）之后或者对步骤（362）的测试的肯定响应之后，进行后续测试，以便确定在盘区列表中是否有更多的项（366）。对步骤（366）的测试的肯定响应将返回

步骤（358），以检索盘区列表中的下一成员，而对步骤（366）的测试的否定响应将导致盘区转换请求的完成（368）。因此，元数据管理器负责起旋与所返回盘区列表中成员关联的物理盘。

如上所示，物理盘可以响应接收到读或写命令而接收提高其速度即起旋的命令。在一种实施方式中，提供盘活动监视表来跟踪文件系统中物理盘的速度。图 7 是说明监视表(405)组件例子的方框图(400)。在一种实施方式中，该表存储在服务器之一的存储器中。如图所示，表（405）包括以下四列：盘号（410）、盘旋转状态（412）、不活动阈值时间（414）和盘定时器（416）。盘号列（410）存储在共享存储设备中指定给每个盘的编号。盘旋转状态列(412)存储相应盘的状态。不活动阈值时间列(414)存储相应盘从活动状态被放置到空闲状态保持不活动的最小时间间隔。盘定时器列（416）存储从相应盘上次被访问开始过去的时间间隔（416）。当盘定时器值超过不活动阈值时间值时，相应的盘被放置到空闲状态。相反，如果不活动阈值时间大于盘定时器，则相应盘保持活动旋转状态。例如，如第一行所示，盘定时器具有 500 的值而不活动阈值设置成 200。因此，由于盘定时器值超过阈值时间值，所以关联的盘被放置到空闲状态，而且旋转状态在表中反映出来。因此，盘活动表监视共享存储设备中每个盘的状态。

图 8 是说明用于监视 SAN 中物理盘的盘活动的处理例子的流程图（450）。一开始，对每个盘的不活动设置阈值（452）。在一种实施方式中，在客户端机器启动时，客户端机器向元数据管理器传送其对物理盘期望的空闲时间。同构客户端，即具有相同操作系统的客户端，可以配置成具有不同的空闲时间。阈值设置其后不活动盘将被放置到空闲状态的时间周期。当元数据管理器看到盘不活动时间大于其阈值时间时，元数据管理器停旋该不活动盘。处于空闲状态的盘比处于活动状态的盘消耗较少的功率。例如，如果物理盘有 2 分钟保持不活动，而其空闲时间设置成 1 分钟，则其旋转状态可以减慢到空闲状态，直到 I/O 请求需要物理盘起旋到为数据请求提供服务的时候。在步骤（452）的阈值建立之后，为每个物理盘设置定时器，该定时器的

初始值为零（454）。允许过去一个时间单位（456），其后该定时器值对每个盘以值一进行递增（458）。在步骤（458）的递增之后，进行测试，以便对所监视的每个盘确定盘定时器是否大于在步骤（452）所设置的盘不活动阈值（460）。对步骤（460）的测试的否定响应之后将返回步骤（456）。这指示被监视的物理盘处于空闲的时间周期都没有超过在步骤（452）中所设置的阈值。但是，对步骤（460）的测试的肯定响应之后将是后续测试，以便确定在大于所设置阈值的时间内处于空闲的每个盘是否在旋转（462）。旋转不活动盘浪费能量。如果盘不旋转，则处理返回步骤（456），继续监视每个被监视盘的旋转状态。但是，如果在步骤（462）确定不活动盘在旋转，则转发停旋不活动盘的命令（464）。停旋盘的动作之后是在表中将该盘的盘状态设置成不旋转状态，即空闲状态（466）。在盘被放置到空闲状态且这种改变已经记录到盘活动表中之后，处理返回步骤（456），以继续监视处理。因此，旋转状态控制处理需要跟踪物理盘的活动并且如果它们在超过所设置阈值时间间隔的时间保持不活动状态则停旋盘。

被指定为该命令提供服务的物理盘在接收I/O命令之前的异步发消息技术使得可以管理物理盘而没有为 I/O 命令提供服务的延迟。使用异步发消息技术的一个例子是当新客户端启动时。在客户端机器启动时，客户端机器向元数据管理器传送其对物理盘期望的空闲时间。这种信息记录到由元数据管理器管理的盘活动表中。在一种实施方式中，客户端到元数据管理器的通信可以异步发生，以便按客户指定的偏好更新所有盘的盘不活动阈值。使用异步发消息技术的另一例子是当元数据管理器接收盘需要被访问的通知时。这种通知可以异步传送到元数据管理器。这种通知优选地包括将用于正被访问的物理盘的时间计数复位成零和将物理盘设置成旋转状态的指令。通过向元数据管理器异步转发这些列出的消息，由于其在为命令提供服务之前提供了另外的起旋的空闲盘时间，所以可以无延迟地为所接收的 I/O 命令提供服务。因此，异步发消息技术的实现使得能够以为 I/O 命令提供服务时最小的延迟或无延迟地对各物理存储盘的旋转状态进行控制。

优于现有技术的优点

元数据管理器将与读和写命令关联的 I/O 指向物理存储介质。元数据管理器维护盘活动表并在发布 I/O 命令之前询问该表以确定物理存储介质的旋转状态。类似地，如果盘处于空闲状态且没有处于活动旋转状态的可用的可选物理盘，则元数据管理器可以在发布 I/O 命令之前向指定盘发布开始起旋处理的异步消息。异步消息的发布避免了与物理盘起旋关联的延迟。因此，共享存储设备中盘的物理旋转状态通过元数据管理器被监视并控制，以便有效地管理与之相关的功率消耗。

在一种实施方式中，由于元数据管理器（114）和文件系统驱动器（116）包含机器可读格式的数据，因此其可以是存储在计算机可读介质上的软件组件。为了这种描述，计算机可用、计算机可读及机器可读介质或格式可以是能够包含、存储、传送、传播或传输由指令执行系统、装置或设备所使用或与其结合使用的程序的任何装置。因此，功率管理工具及关联的组件可以全部是计算机系统中的硬件元件或计算机可读格式的软件元件或软件与硬件结合的形式。

可选实施方式

应当理解，尽管本发明的特定实施方式在此已经为了说明而进行了描述，但在不背离本发明主旨与范围的情况下可以进行各种修改。特别地，当为第一次写分配盘空间时，元数据管理器将尝试将来自客户端的请求映射到具有匹配不活动阈值时间的物理盘。但是，如果没有匹配的物理盘可用，则元数据管理器可以将写请求指向不处于空闲状态的物理盘。此外，响应不能从高速缓存数据提供服务的读或写命令，元数据管理器可以在接收到实际的 I/O 命令之前开始盘的起旋。这种起旋盘的抢先处理避免了与完成 I/O 命令关联的延迟。优选地，盘起旋命令从元数据管理器异步发送到物理盘。因此，本发明的保护范围只能由以下权利要求及其等价物限定。

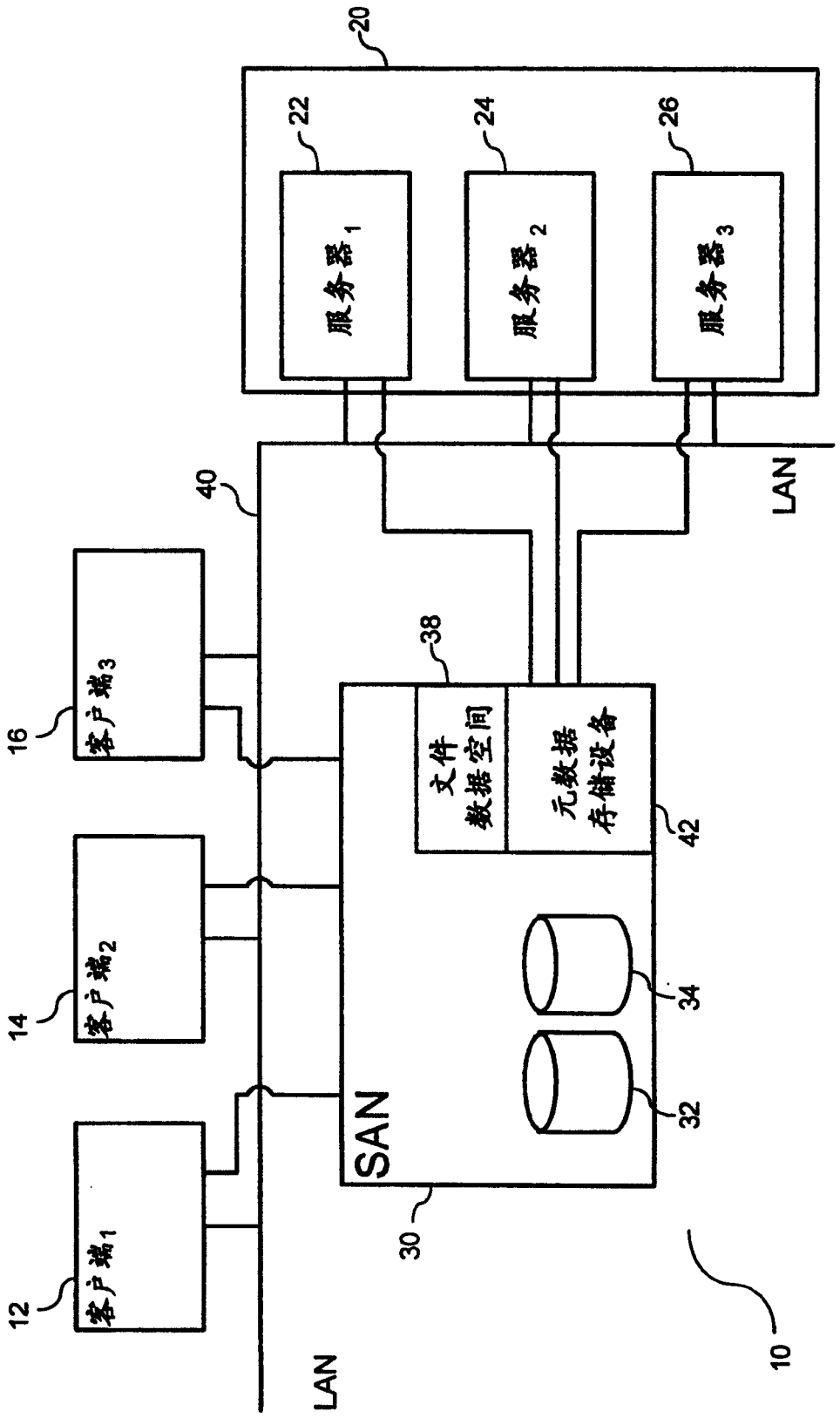

20
22
24
26
服务器1
服务器2
服务器3
LAN
40
38
文件数据空间
元数据存储设备
42
34
32
SAN
30
客户端3
客户端2
客户端1
16
14
12
LAN
10

图1(现有技术)

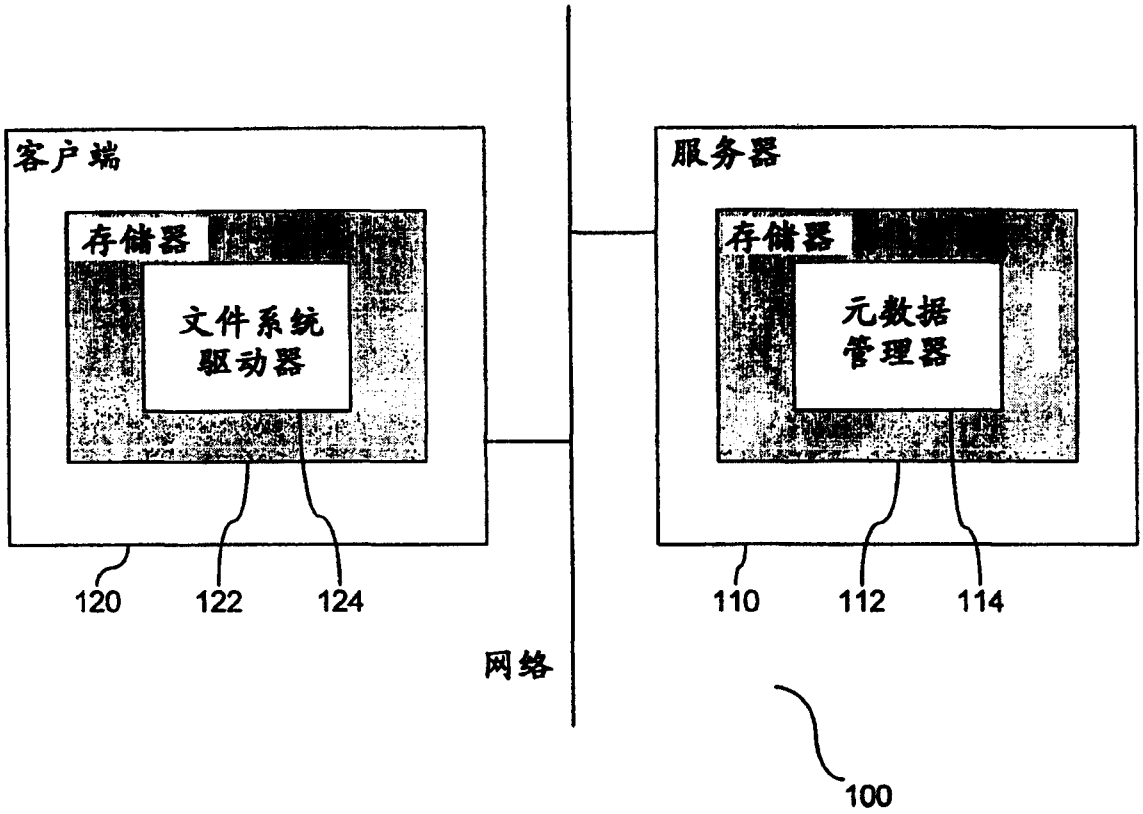


图2

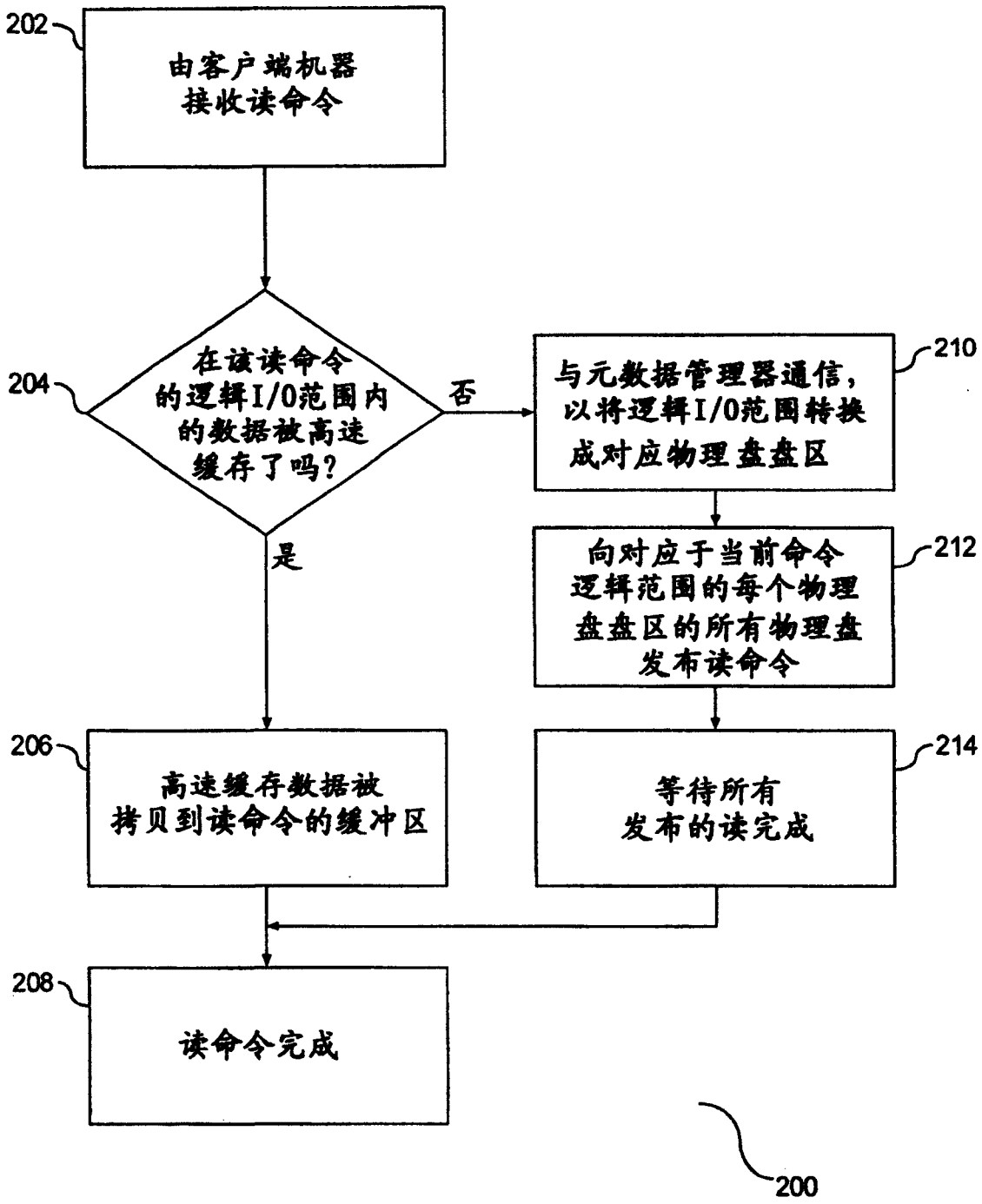

202
由客户端机器
接收读命令
在该读命令
的逻辑I/O范围内
的数据被高速
缓存了吗?
204
否
是
与元数据管理器通信,
以将逻辑I/O范围转换
成对应物理盘盘区
210
向对应于当前命令
逻辑范围的每个物理
盘盘区的所有物理盘
发布读命令
212
高速缓存数据被
拷贝到读命令的缓冲区
206
等待所有
发布的读完成
214
读命令完成
208
200

图 3

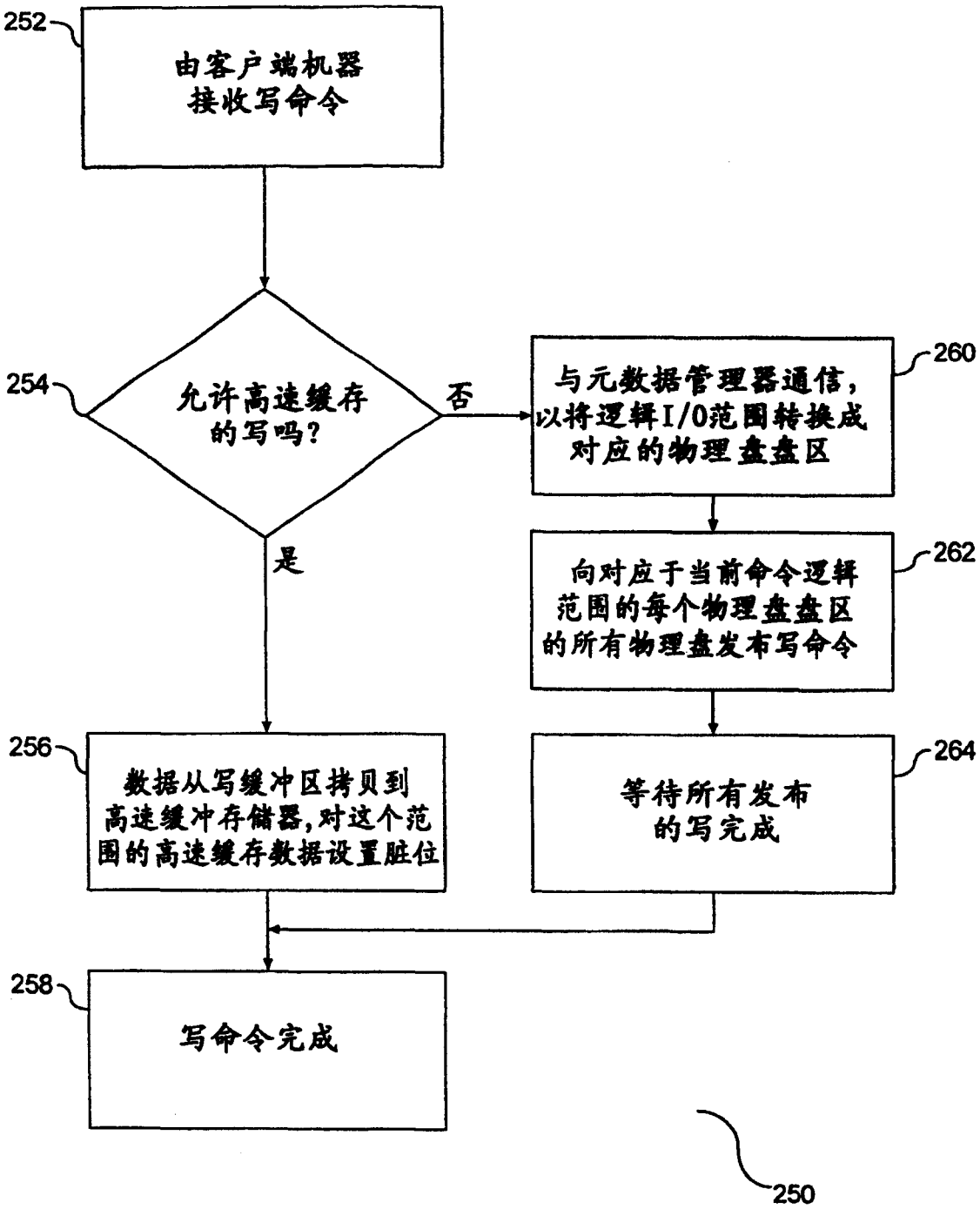

252
由客户端机器
接收写命令
254
允许高速缓存
的写吗?
否
是
260
与元数据管理器通信,
以将逻辑I/O范围转换成
对应的物理盘盘区
262
向对应于当前命令逻辑
范围的每个物理盘盘区
的所有物理盘发布写命令
264
等待所有发布
的写完成
256
数据从写缓冲区拷贝到
高速缓冲存储器,对这个范
围的高速缓存数据设置脏位
258
写命令完成
250

图4

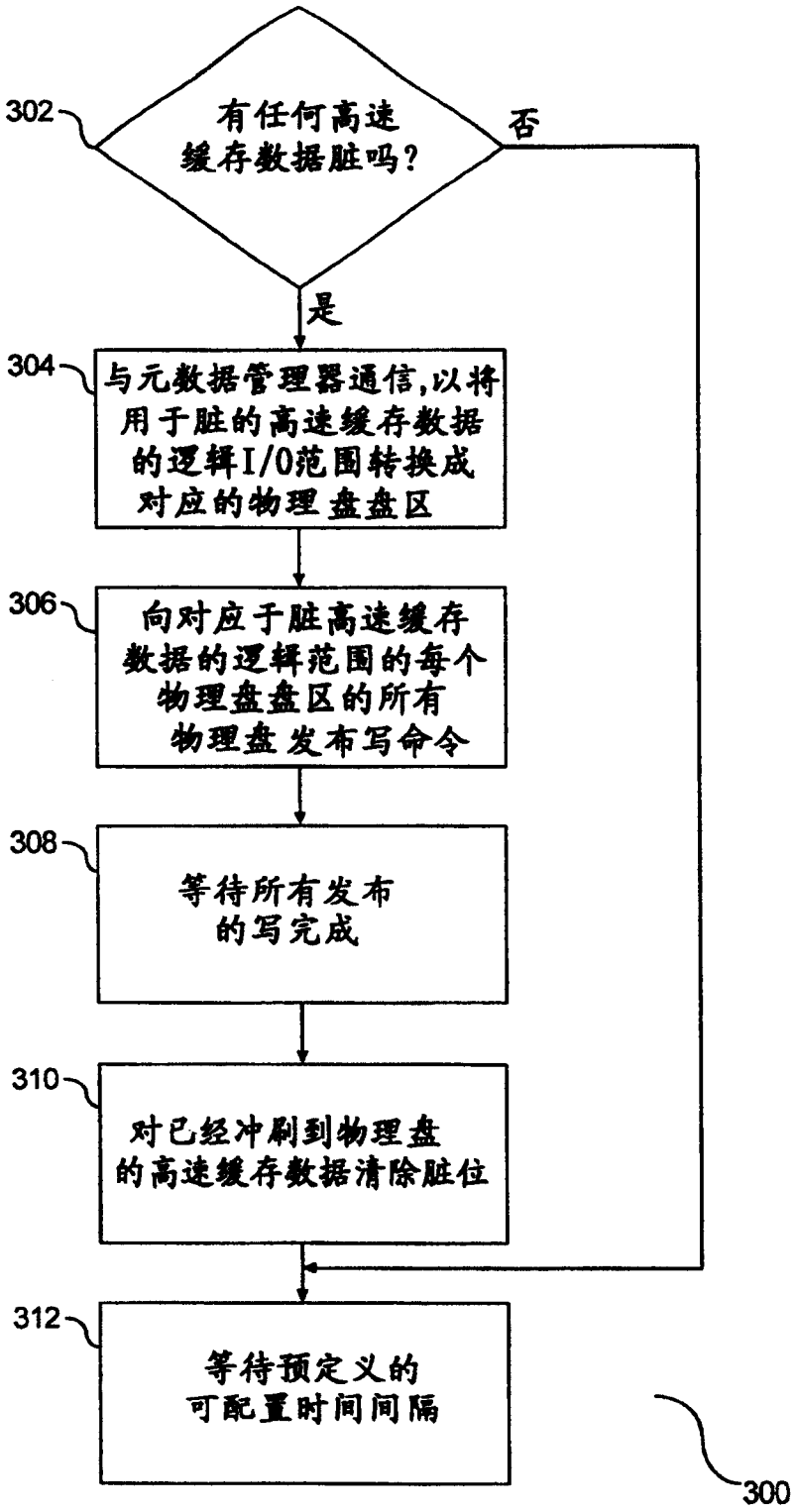
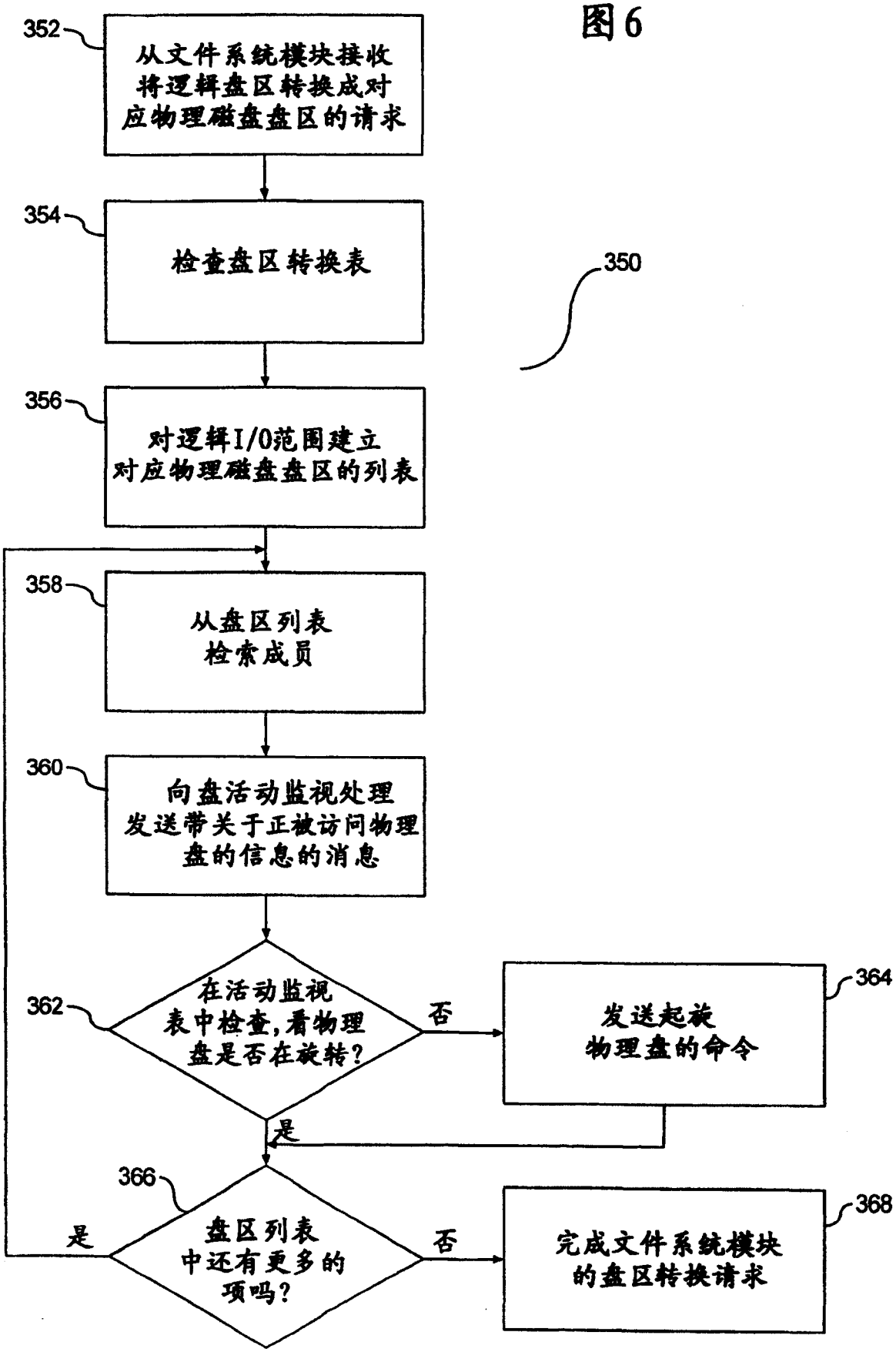

302
有任何高速缓存数据脏吗?
否
是
304
与元数据管理器通信,以将用于脏的高速缓存数据的逻辑I/O范围转换成对应的物理盘盘区
306
向对应于脏高速缓存数据的逻辑范围的每个物理盘盘区的所有物理盘发布写命令
308
等待所有发布的写完成
310
对已经冲刷到物理盘的高速缓存数据清除脏位
312
等待预定义的可配置时间间隔
300

图5

图6


352
从文件系统模块接收将逻辑盘区转换成对应物理磁盘盘区的请求
354
检查盘区转换表
350
356
对逻辑I/O范围建立对应物理磁盘盘区的列表
358
从盘区列表检索成员
360
向盘活动监视处理发送带关于正被访问物理盘的信息的消息
362
在活动监视表中检查,看物理盘是否在旋转?
否
364
发送起旋物理盘的命令
是
366
盘区列表中还有更多的项吗?
是
否
368
完成文件系统模块的盘区转换请求

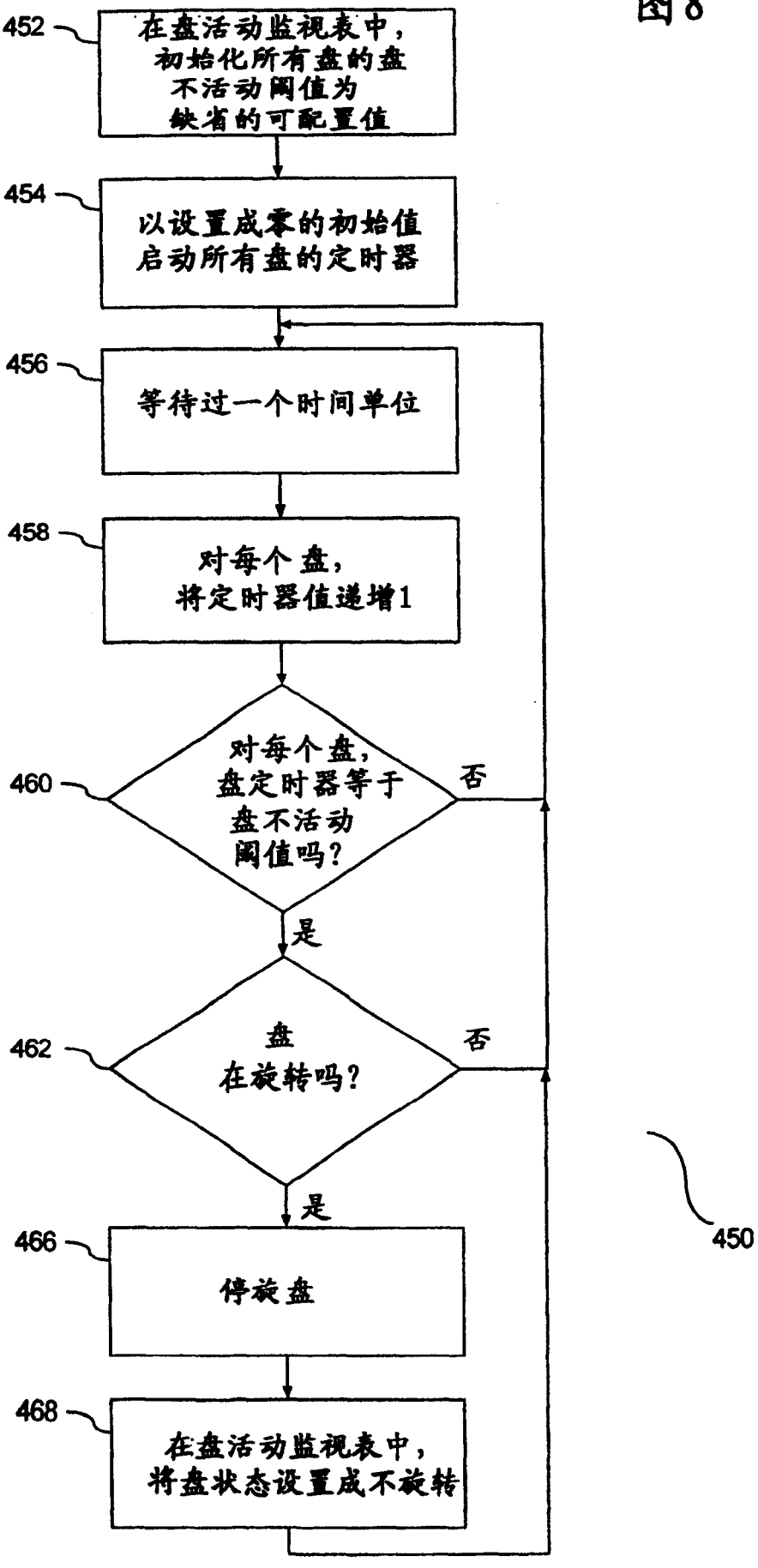
410 412 414 416

盘号	盘 旋转状态	不活动 阈值时间	盘 定时器
0	空闲	200	500
1	旋转/活动	200	100
2	旋转/活动	150	20
3	空闲	200	210

405

图7

图8


452
在盘活动监视表中，初始化所有盘的盘不活动阈值为缺省的可配置值
454
以设置成零的初始值启动所有盘的定时器
456
等待过一个时间单位
458
对每个盘，将定时器值递增1
460
对每个盘，盘定时器等于盘不活动阈值吗？
否
是
462
盘在旋转吗？
否
是
466
停旋盘
468
在盘活动监视表中，将盘状态设置成不旋转
450