



## [12] 发明专利说明书

专利号 ZL 200410090977.3

[45] 授权公告日 2006 年 12 月 27 日

[11] 授权公告号 CN 1292352C

[22] 申请日 2004.11.11

[21] 申请号 200410090977.3

[30] 优先权

[32] 2003.12.31 [33] US [31] 10/749,879

[73] 专利权人 国际商业机器公司

地址 美国纽约

[72] 发明人 M·E·布朗 G·R·怀特威克

审查员 李科

[74] 专利代理机构 北京市中咨律师事务所

代理人 于静 李峰

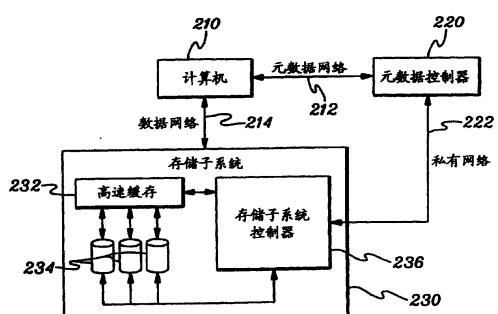
权利要求书 2 页 说明书 10 页 附图 6 页

## [54] 发明名称

利用存储元数据处理管理数据访问请求的方法和系统

## [57] 摘要

提供了一种用于在通信环境中管理数据访问请求的方法和系统。在本发明的一个方面，一请求管理器接收到一个与元数据相关的请求，该元数据对应于与该元数据分开存储的数据。在本发明的另一方面，该请求管理器通知一数据对象管理器一预期的、该数据对象管理器将接收到的请求，以使之为该预期请求做准备。被通知该预期请求后，该数据对象管理器开始为该预期请求做准备，以利于减少数据访问时间。在利用本发明的一个或更多方面的计算环境的一例子中，一存储子系统包括一个数据对象管理器从存储介质预取数据块到高速缓存来为该预期请求做准备。



1. 一种在通信环境中管理请求的方法，该方法包括：

由一管理器接收一个与元数据相关的请求，该元数据对应于与该元数据分开保存的数据；以及

由该管理器通知另一个管理器—预期的、该另一个管理器将接收的请求，以使该另一个管理器为该预期请求做准备。

2. 如权利要求1所述的方法，进一步包括由该另一个管理器为该预期请求做准备，该准备响应于所述通知。

3. 如权利要求2所述的方法，其中所述准备包括管理数据存储子系统中的高速缓存的内容。

4. 如权利要求2所述的方法，其中所述准备包括管理用户或客户计算机对该数据的访问。

5. 如权利要求2所述的方法，进一步包括：

由该管理器在所述通知的几乎同时，向通信单元发送一个响应所述请求的应答；以及

由所述另一个管理器接收该预期请求，其中所述准备在由所述另一个管理器进行的该接收前开始。

6. 如权利要求3所述的方法，其中所述管理数据存储子系统中的高速缓存的内容包括从数据存储子系统的一个或更多存储介质预取一个或更多数据块，从而将所述一个或更多数据块存储在高速缓存里，所述一个或更多数据块包含至少一部分该数据。

7. 如权利要求3所述的方法，其中所述管理数据存储子系统中的高速缓存的内容包括释放高速缓存的存储单元，以使所述存储单元可用于存储其他数据，所述存储单元存储包含至少一部分该数据的数据块。

8. 一种用于通信环境的请求管理系统，该系统包括：

用于由一管理器接收一个与元数据相关的请求的装置，该元数据对应于与该元数据分开保存的数据；以及

---

用于由该管理器通知另一个管理器—预期的、该另一个管理器将接收的请求，以使该另一个管理器为该预期请求做准备的装置。

9. 如权利要求 8 所述的系统，进一步包括用于由该另一个管理器为该预期请求做准备的装置，该用于做准备的装置响应于所述用于通知的装置。

10. 如权利要求 9 所述的系统，其中所述用于做准备的装置包括用于管理数据存储子系统中高速缓存的内容的装置。

11. 如权利要求 9 所述的系统，其中所述用于做准备的装置包括用于管理用户或客户计算机对该数据的访问的装置。

12. 如权利要求 9 所述的系统，进一步包括：

用于由该管理器在通知所述另一个管理器将接收的预期请求的几乎同时，向通信单元发送一个响应所述请求的应答的装置；以及

用于由所述另一个管理器接收该预期请求的装置，其中所述用于做准备的装置在所述用于接收的装置接收该预期请求之前开始为该预期请求做准备。

13. 如权利要求 10 所述的系统，其中所述用于管理内容的装置包括用于从数据存储子系统的一个或更多存储介质预取一个或更多数据块的装置，从而所述数据块存储在高速缓存里，所述数据块包含至少一部分该数据。

14. 如权利要求 10 所述的系统，其中所述用于管理内容的装置包括用于释放高速缓存的存储单元以使存储单元可用于存储其他数据的装置，所述存储单元存储包含至少一部分该数据的数据块。

## 利用存储元数据处理管理数据访问请求的方法和系统

### 技术领域

本发明通常涉及通信环境中的数据访问请求管理，并且更具体地，涉及通知数据对象管理器一个预期的对存储介质中的数据的访问请求，此预期访问请求是基于接收到的与相应于该数据的元数据相关的请求。

### 背景技术

存储子系统通常包括若干盘驱动器，这些驱动器可以聚合起来而对一个或多个客户计算机表现为虚拟盘驱动器。为了改进性能，存储子系统通常利用高速缓存来保持频繁访问的盘块。需要高速缓存的盘块的选择对整体系统性能可能有重要影响。一些存储子系统试图通过检查盘块访问的历史样式来预测哪些盘块可能被客户计算机请求。这种高速缓存管理算法的性质是预测性的。

尽管目前有技术用于管理通信环境中对数据的访问请求，但由于其预测本质这些技术可能使存储子系统把预期时间内不会被访问的数据装载入高速缓存。因此，依需要进一步的、促进计算机环境中对数据的访问请求管理的技术。

### 发明内容

根据本发明的一个方面，一种在通信环境中管理请求的方法，该方法包括：

由一管理器接收一个与元数据相关的请求，该元数据对应于与该元数据分开保存的数据；以及

由该管理器通知另一个管理器一预期的、该另一个管理器将接收的请求，

---

以使该另一个管理器为该预期请求做准备。

根据本发明的另一个方面，一种用于通信环境的请求管理系统，该系统包括：

用于由一管理器接收一个与元数据相关的请求的装置，该元数据对应于与该元数据分开保存的数据；以及

用于由该管理器通知另一个管理器一预期的、该另一个管理器将接收的请求，以使该另一个管理器为该预期请求做准备的装置。

通过提供一种管理请求的方法克服了现有技术的缺点并提供了附加的优点。在本发明的一方面，一个管理器接收到一个与相应于数据（此数据与元数据分开存储）的元数据相关的请求。在本发明的另一方面，此管理器通知另一个管理器一预期的、该另一个管理器将接收到的请求，以使之为预期请求做准备。

---

与上述方法对应的系统和计算机程序产品在此也做了描述并提出了权利要求。

通过本发明的技术实现了附加的特征和优点。在此详细描述了本发明的其他实施例和方面，实施例并认为它们是本发明的一部分。

#### 附图说明

在说明书结尾部分的权利要求中特别指出了并清楚地要求保护了被认为是本发明的主题。通过结合附图阅读下面的详细描述，本发明的前述的以及其他的对象、特征和优点是显然的，在附图中：

图 1 表示了依照本发明的一个方面，一种用于在计算机环境中管理数据请求的技术的一个实施例的流程图；

图 2 表示了依照本发明的一个方面，一种用于在计算机环境中管理数据请求的技术的另一个实施例的流程图；

图 3 表示了依照本发明的一个方面，一种用于在数据与关于该数据的元数据分开存储的环境中管理数据请求的技术的一个实施例；

图 4 表示了依照本发明的一个方面，利用了数据访问请求管理技术的环境的一个例子；

图 5 表示了依照本发明的一个方面，利用了数据请求管理技术的环境的另一个例子；

图 6 表示了依照本发明的一个方面，利用了数据请求管理技术的环境的第三个例子。

#### 具体实施方式

在本发明的一个方面，一个管理器收到一个与元数据相关的请求。此管理器通知另一个管理器—预期的、该另一个管理器将接收到的请求，使之为该预期请求做准备。

下面结合图 1 中所示的请求管理流程图 60 描述一种依照本方面的一个方面，在计算机环境中管理与数据相关的请求的技术。首先，步骤 61 包括

一个请求管理器接收到一个与元数据相关的请求。然后在步骤 62，请求管理器通知数据对象管理器关于数据对象的元数据的改变。在步骤 63 数据对象管理器作出一个数据对象管理决策，并且如果必要，在步骤 64 执行这个数据对象管理决策。

下面结合图 2 中流程图 50 描述依照本发明的一个方面，一种用于在计算机环境中管理与数据对象有关的请求的技术的进一步的方面。首先，步骤 51 包括一个请求管理器接收到使用请求。接着，如果在步骤 52 中确定通信单元被授权允许使用数据对象，则分别在步骤 53 和 55 中请求管理器发送一个请求管理消息给一个数据对象管理器，并且几乎同时地回复一个使用请求响应给一个通信单元。分别在步骤 53 和 55 之后，在步骤 54 中数据对象管理器为预期请求做准备，而在步骤 56 中在通信单元和数据对象管理器之间传输数据使用通信。在一个例子中，传输数据使用通信的步骤包括通信单元传输数据块请求，及数据对象管理器传输包含数据对象的被请求数据块的数据。反之，如果通信单元未被授权允许使用一个数据对象，则步骤 57 包括请求管理器向通信单元回复一个使用请求响应。

下面结合图 3 描述依照本发明的一个方面，利用了数据对象请求管理技术的通信环境的一个例子。在一个数据对象及其元数据可分开存储的环境中，请求管理器 10 通过请求管理网络 12 从通信单元 20 接收使用请求。请求管理器 10 通过私有网路 14 传送一个请求管理消息给数据对象管理器 30，并通过请求管理网络 12 向通信单元 20 回复一个使用请求响应。如果通信单元 20 被授权允许使用数据对象 40，则通过数据网络 16 在通信单元 20 和数据对象管理器 30 之间传输支持数据对象 40 使用的通信。

通常，在计算机通信环境中，元数据和用户数据都与数据对象相关联。用户数据是对某个用户或处理此数据的程序有意义的信息。用户数据的例子如一个 Freelance Graphics<sup>®</sup> 演示的内容，或者存储在关系型数据库中的雇员信息。元数据是关于用户数据的数据。与数据对象相关的元数据的例子包括有访问权限的客户计算机的标识、数据对象类型、与一组盘块相关的文件名、文件长度、组成文件的块列表、用户访问权限信息、文件创建

或更新的日期和时间。数据对象包含数据。数据对象类型包括数据文件、检验点文件、文件系统、逻辑卷、以及日志文件系统(JFS)逻辑卷日志。

在图 3 所示的计算机环境中使用此技术带来的优点包括：改进的数据对象访问安全性，数据对象访问速度调节，数据对象访问优先权仲裁，以及数据对象访问速度的提高。

新出现的一类存储环境把用户数据和元数据分开存储并提供分别的网络来传输用户数据和元数据。这种存储环境的一个例子是 IBM 公司的 Storage Tank™文件系统，其中 Storage Tank™客户（计算机）从一存储子系统（使用块传输协议通过存储区域网（SAN））访问用户数据，从一中心 Storage Tank™元数据控制器（使用 TCP/IP 协议通过以太网）访问元数据。用户数据与元数据的分离可以是逻辑的或者物理的。存储子系统通常包括若干可以聚合起来并向一个或多个客户计算机显示为虚拟盘驱动器的盘驱动器，并且经常使用高速缓存来保持频繁访问的盘块以改进输入输出性能。

本发明的一个或多个方面利用了这样一个事实，即用户数据与元数据分离时，与文件访问同时的元数据处理提供了额外的信息，其可以用来通知存储子系统未来的输入/输出（I/O）访问请求。存储子系统可以利用这些信息来促进内部高速缓存的管理。

下面结合图 4 描述依照本发明的一个方面，利用处理文件元数据获得的信息管理数据存储子系统中高速缓存内容的一个例子。当客户计算机 210 需要读取并更新与存储子系统 230 上的一文件相关联的一些盘块时，客户计算机 210 必须被授权在相关盘块上的排他锁。客户计算机 210 向元数据控制器 220 发起请求锁的事务。此锁请求表示客户计算机 210 希望未来在某一定范围的盘块上执行输入/输出操作。如果元数据控制器 220 可以授权此请求锁，元数据控制器 220 会传输一个“提示”给存储这些块的存储子系统 230，表明存储子系统可以预期接收来自客户计算机的对某一特定范围的盘块的输入/输出请求。元数据控制器（MDC）220 通过私有网络 222 传送这个“提示”到存储子系统 230。基本上同时地，元数据控制器

---

220 通过元数据网络 212 给客户计算机 210 发信号来授权该锁。在这个示例性的实施例中，MDC 220 是一个请求管理器的例子，存储子系统 230 的存储子系统控制器 236 是一个数据对象管理器的例子。

如果存储子系统 230 确定请求的盘块不在高速缓存 232 中，它把请求的块从存储盘 234 中预取到高速缓存 232 中。接收到请求的锁后，客户计算机 210 通过数据网络 214 向存储子系统 230 发起一个输入/输出操作，以存取接收到其上的锁的盘块中的至少一些盘块。当存储子系统 230 接收到客户发起的输入/输出请求时，由于预取，存储子系统可能在其高速缓存中已经有了请求的盘块。即使没有，由于先前已经从元数据控制器 220 接收到提示，存储子系统 230 已经开始了必要的物理输入/输出操作来把请求的块装载到高速缓存 232 中。当请求的盘块在高速缓存 232 中时，通过数据网络 214 从高速缓存 232 将它们传送到客户计算机 210。存储子系统 230 在接收到客户计算机 210 的请求之前发起盘输入/输出操作，以把客户计算机将要请求访问的盘块存储在高速缓存 232 中，其结果是数据访问的等待时间减少了。

本发明的方法也用在图 4 所示的示例环境的操作中当客户计算机把盘块写到存储子系统时。如果客户计算机 210 发送一个事务到元数据控制器 220，表明客户计算机已经关闭了一个文件，并完成了把块写到存储子系统，元数据控制器 220 通过私有网络 222 向存储子系统 230 传送这个信息。存储子系统 230 根据这个文件关闭消息以及可能接收到的有关其他未来数据访问请求的“提示”决定是否释放存储了构成该关闭的文件的盘块的、高速缓存 232 中的存储单元。释放高速缓存 232 中的存储单元允许为快速访问而在高速缓存中存储其他盘块。

依照本发明的一个方面、利用处理文件元数据获得的信息管理数据存储子系统的高速缓存内容的另一个例子涉及计算机写一个不太可能被读的大文件。这种文件的一个例子是长时间运行的计算作业产生的一检验点/重启文件或一数据库日志文件。这些文件一般在计算机崩溃后用来恢复一个计算工作负荷的状态。因为计算机崩溃很少发生，检验点/重启一般

---

定期写，但很少读。对这个信息的认识可以用来通知存储子系统在检验点/重启文件写入盘后不要对其进行高速缓存。

结合图 4 的示例性环境对这个例子作了进一步描述。当计算机 210 希望写一个检验点/重启文件时，它通过元数据网络 212 向元数据控制器 220 请求写权限，并通知元数据控制器 220 这个文件不应该被活动地高速缓存。另一个选择是，元数据控制器 220 能自动识别出此文件是一特殊类型（例如检验点/重启文件）。元数据控制器 220 授权允许计算机 210 写该文件，提供文件应写入的块列表，并同时通过私有网络 222，通知存储子系统 230 的存储控制器 236 计算机 210 准备写一个不应在存储子系统的高速缓存 232 中被活动地高速缓存的大文件。

当计算机 210 通过数据网络 214 把文件写到存储子系统 230 时，存储子系统控制器 236 决定分配高速缓存 232 的多大空间来存储全部或部分大文件，并尽快地把大文件的内容写入存储子系统的存储盘 234。一旦文件的内容（或部分内容）写入存储子系统的存储盘 234，高速缓存 232 中的相关文件数据可以立刻被丢弃，因为这个文件需要再被读出的可能性很小。因此，基于处理文件的元数据获得的被存储数据的类型的知识来使用高速缓存，有利于优化高速缓存资源的使用。

与上述例子有关的一例子涉及一计算机在崩溃后读取一前面描述的检验点/重启文件来恢复一计算工作负荷的状态。可以利用这个文件很少被读的认识，来通知存储子系统在这个文件从盘中读出后不要高速缓存它。需要指出的是，下面描述的关于这个例子的数据存储子系统中的高速缓存内容的管理适用于可能不经常访问的任何大文件。

参考图 4，计算机 210 试图读取一检验点/重启文件，因此计算机 210 通过元数据网络 212 从元数据控制器 220 请求允许以便读取文件，并通知元数据控制器 220 此文件不应该被主动高速缓存。或者，元数据控制器 220 自动地识别出此文件是特殊类型（例如一检验点/重启文件），而不是需要计算机 210 通知。元数据控制器 220 授权允许计算机 210 读取文件，提供文件所在的块列表，并且同时，通过私有网络 222 通知存储子系统 230

的存储子系统控制器 236，计算机 210 将要读取一个不应该在存储子系统的高速缓存 232 中被活动地高速缓存的大文件。当计算机 210 通过数据网络 214 从存储子系统 230 中读文件时，存储子系统控制器 236 决定分配高速缓存 232 的多大空间用来从存储盘 234 读文件。一旦文件的内容（或部分内容）传输到计算机 210，高速缓存 232 中的相关文件数据就可以立刻丢弃，因为此文件需要被再次读取的可能性很小。因此，基于处理文件的元数据确定的、被访问数据的类型来使用高速缓存，释放了高速缓存资源，有利于提高对其他文件的访问速度。

下面结合图 5 描述依照本发明的一个方面，使用一种管理数据对象访问请求的技术，以便利用处理文件元数据获得的信息，来实现数据存储子系统的高速缓存内容管理的一种环境的另一个例子。在此例子中，若干计算机（310、311 和 312）是一数据库群集的成员。计算机 310、311 和 312 中的每一个都运行群集数据库的一个实例。当计算机 310 希望读取并更新与位于存储子系统 330 的一数据库表相关联的一些盘块时，计算机 310 必须被授权一个在相关盘块上的排他锁。计算机 310 通过在外部网络 314 上发送一个请求来从数据库锁管理器 320 请求一个在这些盘块上的锁。如果数据库锁管理器 320 可以授权请求的锁，则数据库锁管理器 320 通过在外部网络 314 上向计算机 310 发送一消息来授权锁，并且几乎同时地发送一消息给存储子系统 330，通知存储子系统 330 可以预期从客户计算机接受一个访问特定范围的盘块的输入/输出请求。数据库锁管理器 320 通过私有网络 322 向存储子系统 330 传送这个消息。与前面的例子类似，数据库锁管理器 320 是一请求管理器的例子，而存储子系统 330 构成了一数据对象管理器。

如果存储子系统 330 确定被授权锁的盘块不在高速缓存 332 中，存储子系统 330 向存储盘 334 发起一输入/输出操作。计算机 310 因为已经获得请求的盘块上的锁授权，所以向存储子系统 330 发起一个输入/输出操作。当存储子系统 330 通过数据网络 316 接收到计算机 310 发起的输入/输出请求时，存储子系统可能已经预取了请求的盘块并将它们存储在高速缓存

332 中。即使请求的盘块还没有全部装入高速缓存，存储子系统已经在收到计算机 310 的请求之前发起了从存储盘 334 的物理输入/输出操作。因此，从高速缓存 332 提供请求盘块的等待时间小于没有根据数据库锁管理器的“提示”进行预取时的等待时间。

在结合图 6 描述的实现了本发明的计算机环境的另一个例子中，一个集中的存储元数据控制器 434 与一个存储子系统 420 共存在一起。在此环境中，一个或多个计算机通过一数据网络连接到数据存储子系统。作为例子，如图 6 所示，一个计算机 410 通过数据网络 412 与存储子系统 420 交换数据。存储子系统 420 包括一个连接到高速缓存 322 及存储盘 424 的服务器 430。服务器 430 包括逻辑分区 431 和 432。

在该例子中，元数据控制器 434 和存储子系统控制器 433 的功能分别由运行在服务器 430 的逻辑分区 (LPAR) 432 和 431 上的软件执行。使用在元数据控制器 LPAR 432 与存储子系统控制器 LPAR 431 之间的虚拟输入/输出总线 436，关于预期的对存储子系统 420 的未来输入/输出请求的“提示”以极高的速度和低的等待时间从元数据控制器 434 传送到存储子系统控制器 433。预取盘块到存储子系统高速缓存的优点通过在元数据控制器与存储子系统控制器之间使用高速低等待时间的通信得到了提高。在本例中，元数据控制器 434 与存储子系统控制器 433 分别是请求管理器及数据对象管理器的例子。

本发明可以被包含在一种具有例如计算机可用介质的产品中（例如一个或多个计算机程序产品）。此介质在其中具有例如计算机可读的程序代码装置或逻辑（例如指令、代码、命令等），来提供并促进本发明的能力。该产品可以作为计算机的系统的一部分或者单独销售。

另外，可以提供至少一个机器可读的程序存储装置，它体现了至少一个机器可执行的指令程序来实现本发明的能力。

这里描述的流程图只是示例。在不脱离本发明精神的情况下，这里描述的图或步骤（或操作）可以有许多变种。例如，可以按不同的顺序执行各步骤，或者增加、删除、或修改一些步骤。所有这些变种均视为本发明

的一部分。

尽管这里详细刻画并描述了优选的实施例，对熟悉相关技术的人来讲，很显然可以在不偏离本发明精神的情况下做出各种修改、增加、替换等等，因而，这些均被视为属于如权利要求书中所定义的本发明范围以内。

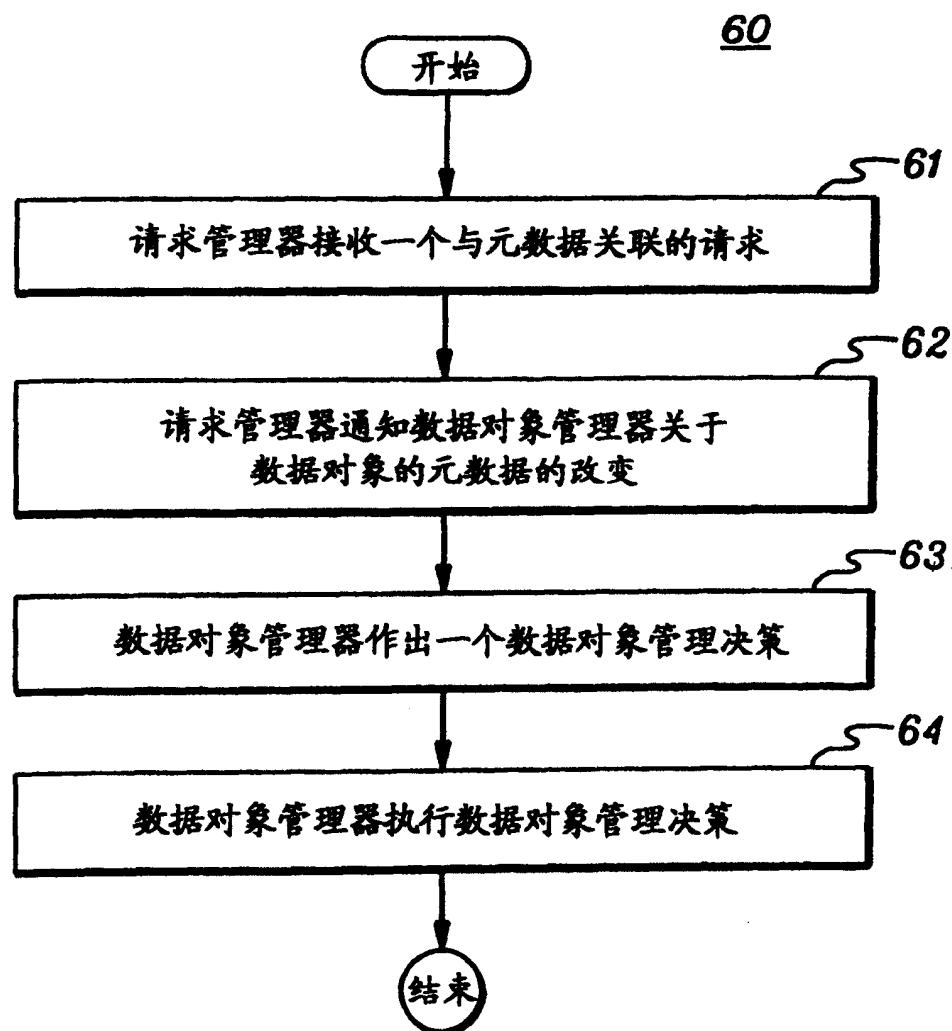
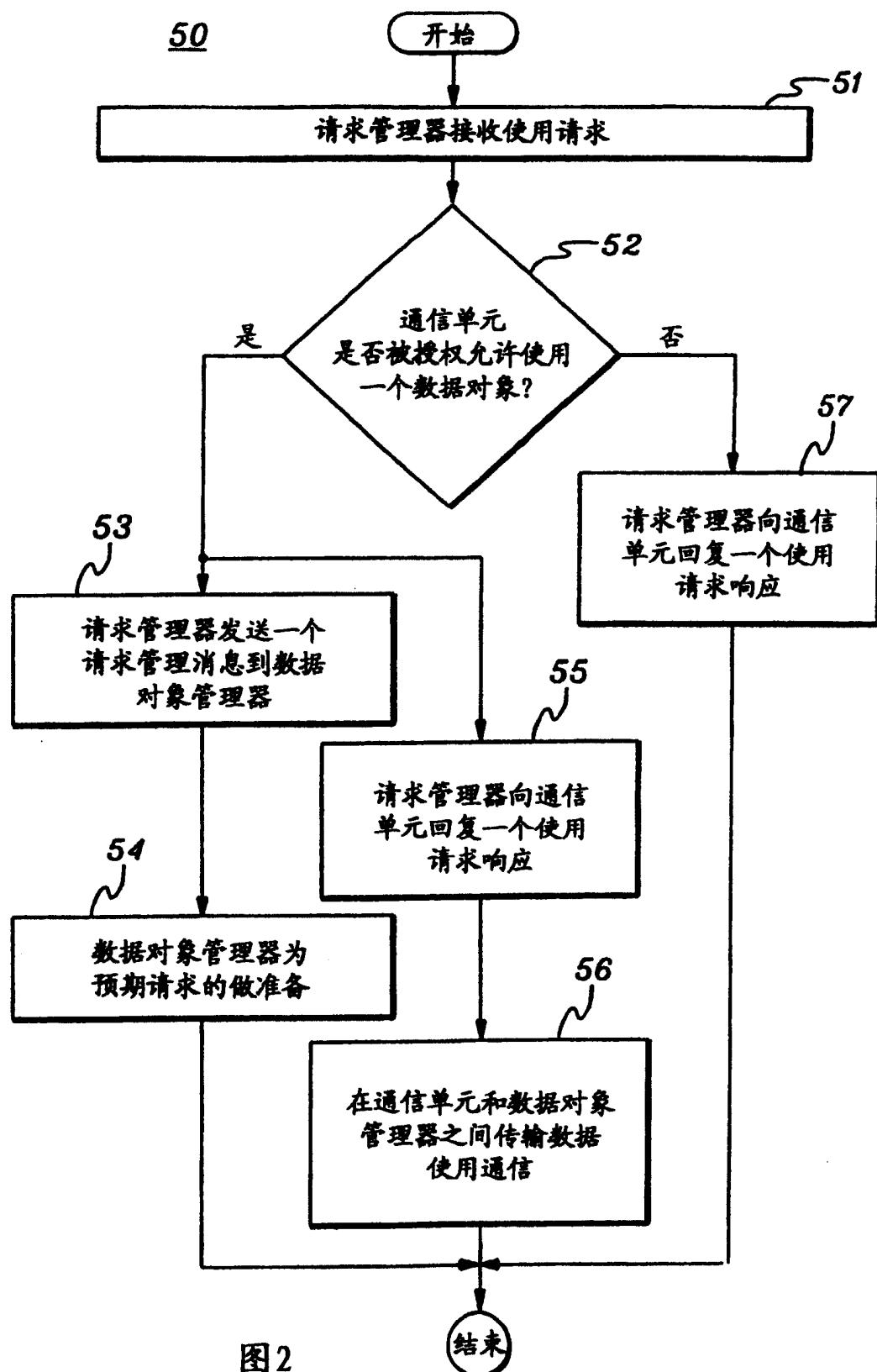


图1



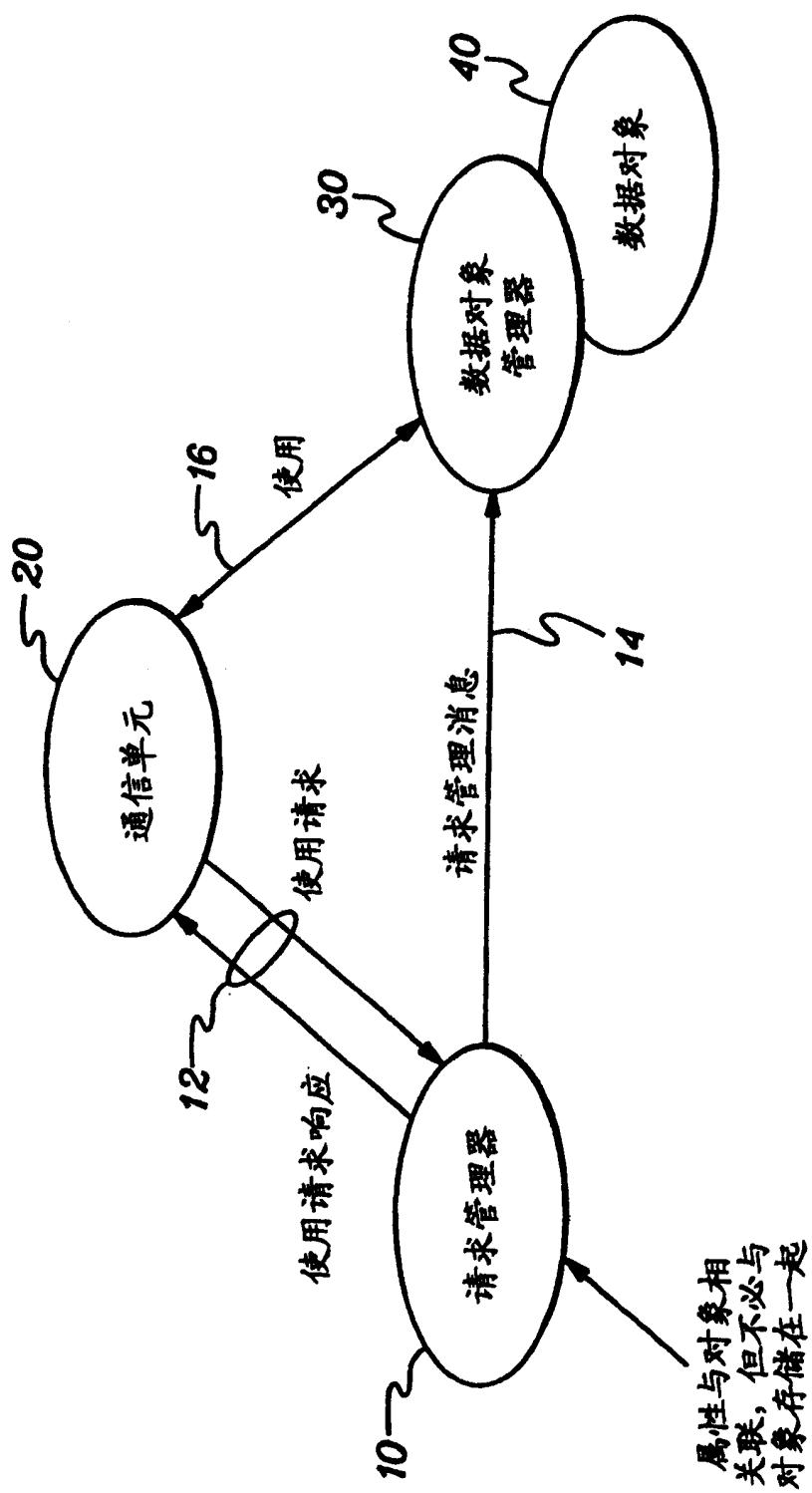


图 3

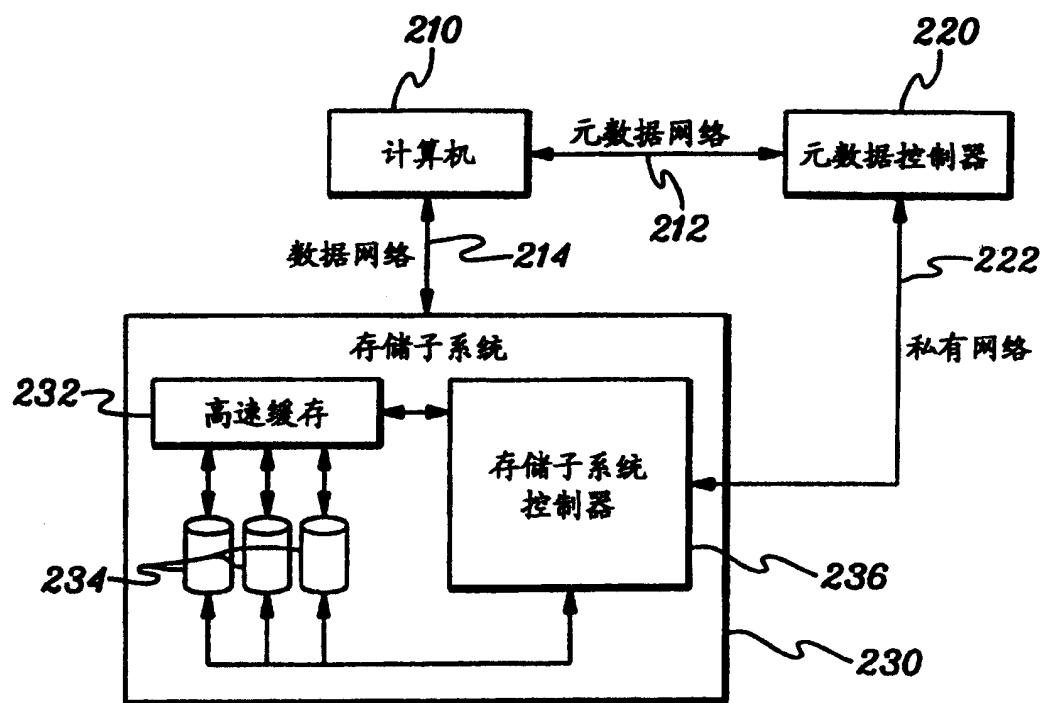


图4

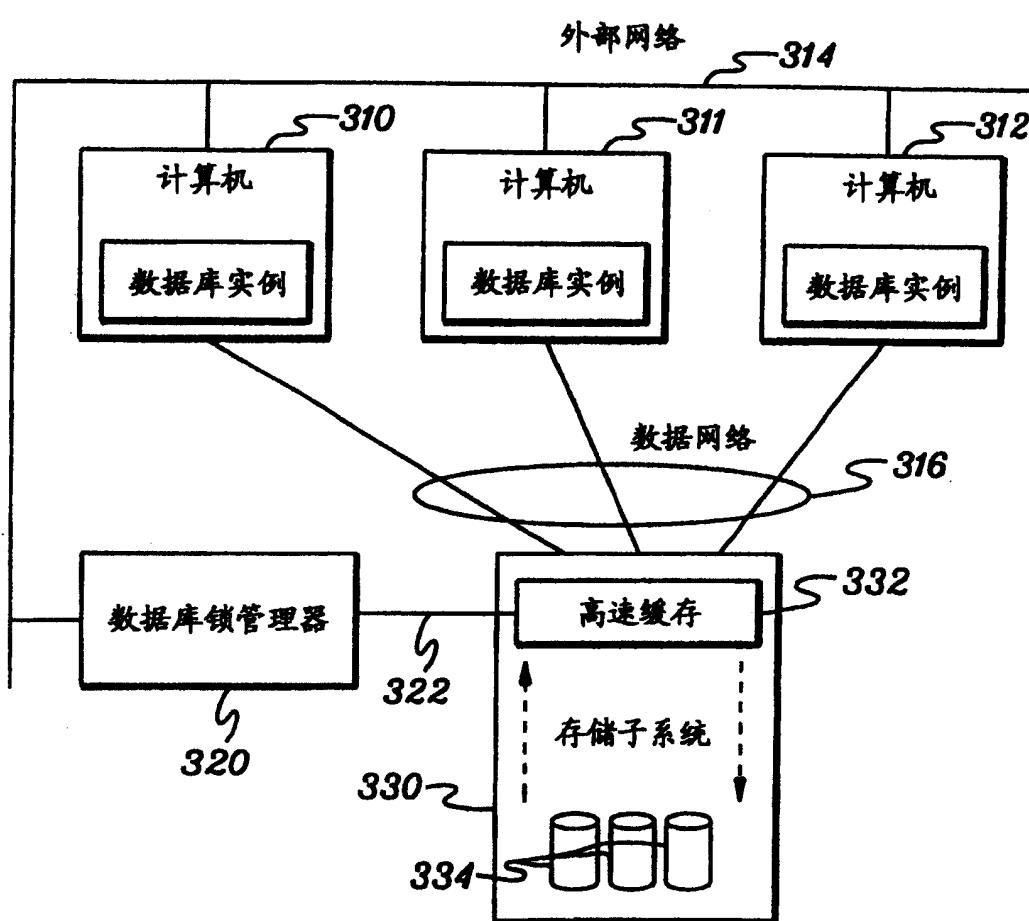


图5

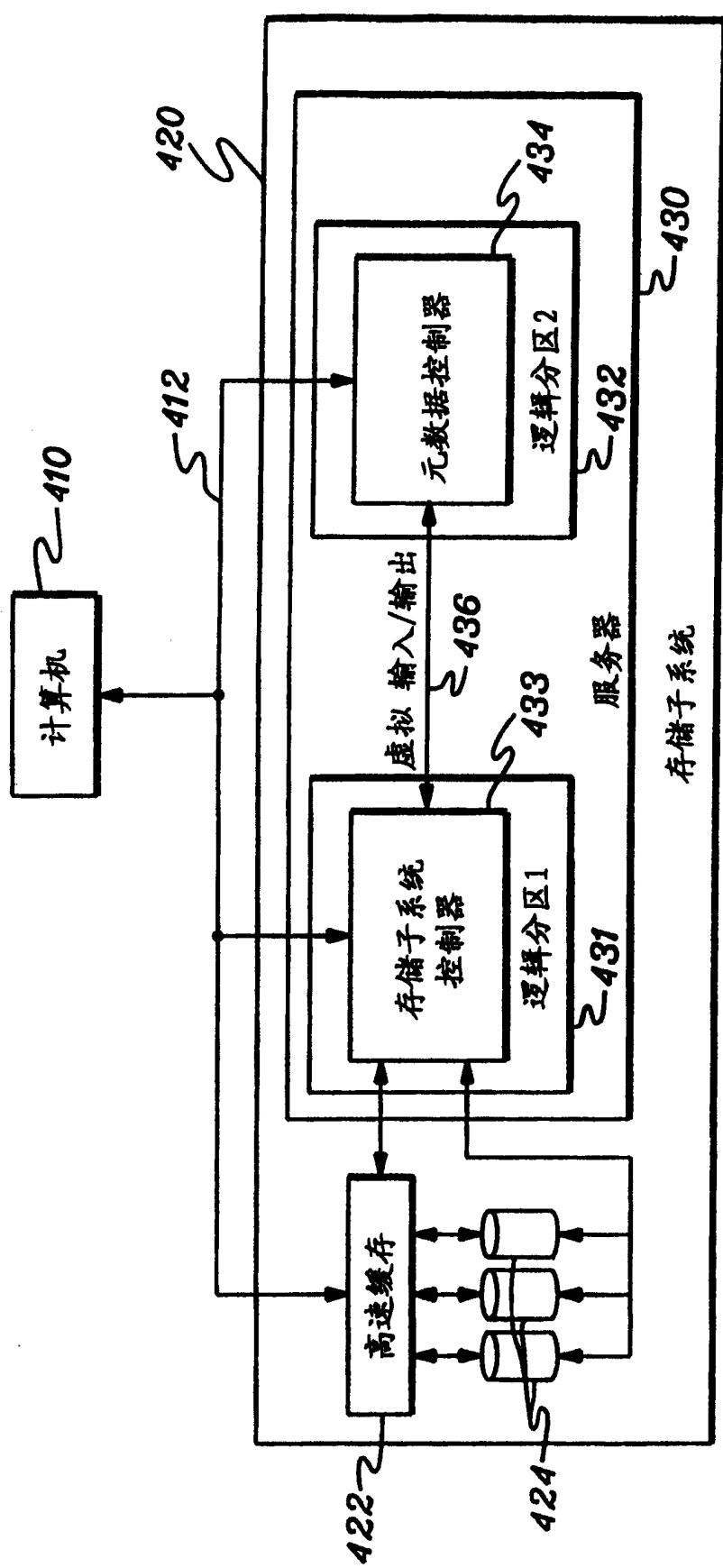


图6