

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

G06F 11/14 (2006.01)

G06F 17/30 (2006.01)



[12] 发明专利说明书

专利号 ZL 200680005358.7

[45] 授权公告日 2009 年 5 月 20 日

[11] 授权公告号 CN 100489800C

[22] 申请日 2006.2.17

[21] 申请号 200680005358.7

[30] 优先权

[32] 2005.2.18 [33] US [31] 11/061,152

[86] 国际申请 PCT/US2006/005909 2006.2.17

[87] 国际公布 WO2006/089263 英 2006.8.24

[85] 进入国家阶段日期 2007.8.17

[73] 专利权人 甲骨文国际公司

地址 美国加利福尼亚州

[72] 发明人 萨什坎斯·钱德拉塞克拉恩

安杰洛·普鲁希诺

[56] 参考文献

US5951695A 1999.9.14

CN1481534A 2004.3.10

Oracle8 Replication Manual. Bobrowski ET AL. Oracle8 Replication, Release 8.0 Part No. A58245.01. 1997

审查员 毛习文

[74] 专利代理机构 北京康信知识产权代理有限公司

代理人 余刚 尚志峰

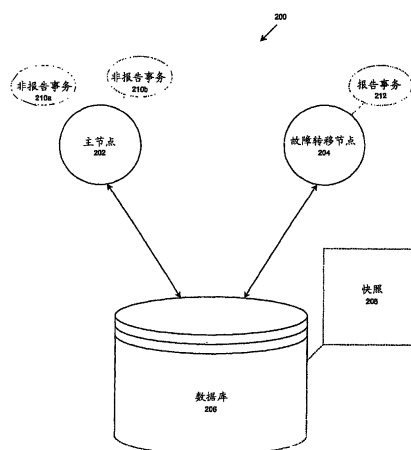
权利要求书 3 页 说明书 11 页 附图 11 页

[54] 发明名称

在数据库系统中处理报告事务的方法和机构

[57] 摘要

本发明披露了用于处理数据库系统中的报告事务的改进方法、系统、和介质。在一些实施例中，在主节点上执行非报告事务的同时，使用数据库快照在故障转移节点上执行报告事务。



1. 一种用于在数据库系统中处理报告事务的方法，所述方法包括：

给数据库拍快照，其中，所述数据库链接至主节点和故障转移节点；

在所述主节点上执行一个或多个非报告事务；以及

在所述主节点上执行所述一个或多个非报告事务的同时，利用所述快照在所述故障转移节点上执行报告事务，其中，执行所述报告事务，而不依赖同步复制的数据。

2. 根据权利要求1所述的方法，还包括：

在所述故障转移节点上创建一个或多个临时表，其中，当在所述故障转移节点上执行所述报告事务时，使用所述一个或多个临时表。

3. 根据权利要求2所述的方法，其中，通过所述报告事务中的查询脚本创建所述一个或多个临时表。

4. 根据权利要求2所述的方法，其中，所述报告事务中的一个以上的查询可访问所述一个或多个临时表中的至少一个。

5. 根据权利要求1所述的方法，还包括：

修改所述数据库中的一个或多个模式，其中，当在所述故障转移节点上执行所述报告事务时，使用所述一个或多个模式。

6. 根据权利要求5所述的方法,其中,所述一个或多个模式不可访问在所述主节点上执行的所述一个或多个非报告事务。
7. 根据权利要求5所述的方法,其中,所述一个或多个模式中的至少一个包括一个或多个表。
8. 根据权利要求1所述的方法,还包括:

访问所述主节点上的一个或多个用户定义的程序,其中,当在所述故障转移节点上执行所述报告事务时,使用所述一个或多个用户定义的程序。
9. 根据权利要求1所述的方法,还包括:

在所述数据库中保留临时空间,其中,当在所述故障转移节点上执行所述报告事务时,使用所述临时空间。
10. 根据权利要求1所述的方法,其中,所述主节点和所述故障转移节点是群集的一部分。
11. 根据权利要求10所述的方法,其中,所述群集包括一个或多个额外的故障转移节点。
12. 根据权利要求1所述的方法,其中,所述一个或多个非报告事务中的至少一个是读写事务。
13. 根据权利要求1所述的方法,其中,所述报告事务和所述一个或多个非报告事务是工作量的一部分。
14. 根据权利要求1所述的方法,其中,所述报告事务通过使用最新的更新提供基本实时报告。

15. 根据权利要求1所述的方法,其中,仅所述主节点可以修改所述数据库。
16. 根据权利要求1所述的方法,其中,响应于用户命令拍所述快照。
17. 根据权利要求1所述的方法,其中,所述快照是只读的。
18. 根据权利要求1所述的方法,其中,所述快照不能被所述主节点修改。
19. 根据权利要求1所述的方法,其中,所述快照和所述数据库共用盘空间。
20. 根据权利要求1所述的方法,其中,所述快照是最新的。
21. 根据权利要求1所述的方法,其中,直接使用所述快照在所述故障转移节点上执行所述报告事务。
22. 一种用于在数据库系统中处理报告事务的系统,所述系统包括:

用于给数据库拍快照的装置,其中,所述数据库链接至主节点和故障转移节点;

用于在所述主节点上执行一个或多个非报告事务的装置;以及

用于在所述主节点上执行所述一个或多个非报告事务的同时,利用所述快照在所述故障转移节点上执行报告事务的装置,其中,执行所述报告事务,而不依赖同步复制的数据。

在数据库系统中处理报告事务的方法和机构

技术领域

本发明涉及数据库系统。更具体地，本发明涉及一种在数据库系统中处理报告事务的方法和机构。

背景技术

许多数据库系统采用故障转移群集（failover cluster）来保证高可用性，这在当今快速发展的市场中是极为重要的。在故障转移群集中，数据库链接至主节点和至少一个故障转移节点（也称为备用节点）。在主节点上运行程序（例如，数据库和网络服务器），直到主节点发生故障。当这种情况发生时，在故障转移节点上重新启动程序。由于故障转移节点和主节点属于单个群集，所以可以使用标准心跳机构（heartbeat mechanism）来检测主节点的故障。

故障转移群集的一个问题在于，故障转移节点不能和主节点同时使用。因而，可能难以估计购买仅在主要硬件故障时使用的其他硬件的成本。某些并行数据库系统通过采用活动/活动群集（active/active cluster）来解决这个问题，其中，在这种群集中，两个或两个以上节点可以同时访问数据库。然而，活动/活动群集需要复杂的同时控制机构来确保，在对群集中的所有节点进行同时读取和修改的情况下，数据库是一致的。

用户面对的另一个问题是需要运行混合的工作量，其中，与其他事务同时执行报告事务。理论上通过每个报告事务提供实时的

报告，即，通过事务中的查询程序来使用最新更新的结果。另外，用户更愿意单独运行报告事务，从而避免在非报告和报告事务之间的硬件资源（例如，对 CPU 或存储器）的竞争。

对于不支持活动/活动群集的数据库系统来说，可以创建复制型数据库并使用其来进行报告。然而，由于复制型数据库是主数据库的完全复制，所以这个解决方案会使存储成本加倍。另外，由于复制主数据库中可能无法即时复制主节点中的变化，所以复制型数据库通常滞后于主数据库。即使能够即时复制，主数据库上的吞吐量也会由于需要将主数据库上的每个提交命令（commit）同步复制到报告数据库中而受到极大影响。

因此，需要一种用于解决在使用故障转移群集的数据库系统中执行报告事务的这些和其他问题的方法和机构。

发明内容

本发明的实施例提供了用于在数据库系统中处理报告事务的改进方法、系统、和介质。根据实施例，给数据库拍快照。数据库链接至主节点和故障转移节点。然后，在主节点上执行一个或多个非报告事务，并且在主节点上执行一个或多个非报告事务的同时，利用快照在故障转移节点上执行报告事务。

以下，在实施方式、附图、和权利要求书中进一步详细描述了本发明的各个方面、目的和优点。以上的概括描述和以下的详细描述是示例性和说明性的，并不用于限制本发明的范围。

附图说明

本发明中所包括的附图用于提供对本发明的进一步理解，并且连同具体实施方式一起用于说明本发明的原理。

图1是根据本发明实施例的在数据库系统中处理报告事务的方法的流程图。

图2示出了根据本发明的一个实施例的在故障转移群集中执行报告事务。

图3示出了根据本发明的另一个实施例的用于在数据库系统中处理报告事务的方法的流程图。

图4是根据本发明的另一实施例的如何在群集中处理报告事务的实例。

图5示出了在数据库系统中处理报告事务的方法的一个实施例。

图6示出了具有多个故障转移节点的群集。

图7示出了用于在数据库系统中处理报告事务的另一个实施例。

图8示出了示例的数据库系统。

图9是根据本发明的另一个实施例的用于在数据库系统中处理报告事务的方法的流程图。

图 10 示出了根据本发明的另一个实施例的在故障转移群集中执行多个报告和/或非报告事务。

图 11 是可以用于实现本发明实施例的系统架构的示意图。

具体实施方式

本发明公开了如何在数据库系统中处理报告事务。在主节点上执行非报告事务的同时，使用数据库快照在故障转移节点上执行报告事务，而不采用需要复杂的相关性和路由选择机制、或具有单个复制型数据库（需要购买额外硬件）的活动/活动群集（具有可能过期的数据）。这利用了保持空闲的故障转移节点，并提供了在使用最新快照时的近实时报告。

图 1 所示的是在数据库系统中处理报告事务的方法。在 102，给数据库排快照。数据库链接至主节点和故障转移节点。在一些实施例中，只允许主节点修改数据库。客户机连接可以被配置为将所有报告事务导向故障转移节点，而将所有其他事务导向主节点。故障转移节点也可以将可能修改数据库的事务自动路由到主节点。可以通过将事务标记为 READ-WRITE 或 READ-ONLY 来完成该路由选择，这标识该对话是否将要修改数据库。

然后，在主节点上执行一个或多个非报告事务（104），并且在主节点上执行该一个或多个非报告事务的同时，利用快照以在故障转移节点上执行报告事务（106）。报告和/或非报告事务中的每个均包括一个或多个查询。尽管非报告事务可以是读写或只读事务，但是报告事务通常是只读事务。

快照是数据库的点即时复制，并且与除了在拍快照之后修改的数据库块之外的数据库共用相同的盘空间。这可以通过标准写时复

制 (copy-on-write) 机构 (其中, 将改变后的块写入新的位置, 从而使快照仍保持不会被修改) 来实现。由于快照是只读的并不能被主节点修改, 所以在故障转移节点上进行的查询将返回与在不需要与主节点进行协调的情况下所使用的快照相一致的结果。并且, 由于对于整个数据库来说快照是一致的 (即, 快照中的缩略图和查询中所参考的表都是一致的), 所以不需要修改现存的查询执行引擎。可利用各种快照方法并且可以在文件、程序、系统、或数据库层实施这些方法。例如, 可以在 <http://www.netapp.com/techlibrary/3002.html> 找到关于创建文件层快照的描述。

就盘空间和 CPU 的使用而言, 创建快照相对容易, 因为它们都使用相同的盘存储空间用作所有未改变数据的数据库。因此, 数据库系统可以被配置为频繁 (例如, 每隔 10 秒) 拍快照。然而, 例如, 数据库系统也可以基于报告对话所期望的服务质量或其他这样的量度 (metric), 响应于用户命令而生成快照。由于通过在报告事务中进行查询将使用最新的更新, 所以使用最新快照在故障转移节点上执行报告会话将提供近实时的报告。然而, 也可以允许用户指定使用比所获取的最新快照旧的快照。

图 2 示出了具有主节点 202、故障转移节点 204、以及数据库 206 的群集 200。数据库 206 的快照 208 已被拍下来。在主节点 202 上执行多个非报告事务 210a 和 210b 的同时, 使用快照 208 在故障转移节点 204 上执行报告事务 212。在一些实施例中, 非报告事务 210a 和 210b 以及报告事务 212 是工作量的一部分。

图 3 示出了用于在数据库系统中处理报告事务的方法的流程图。根据实施例, 给链接至主节点和故障转移节点的数据库拍快照 (302)。在 304, 在主节点上执行一个或多个非报告事务。在主节点上执行该一个或多个非报告事务的同时, 利用快照在故障转移节点

上执行报告事务 (306)。然后, 当在故障转移节点上执行报告事务时, 创建并使用一个或多个临时表 (308)。

图 4 示出了群集 400。群集 400 包括主节点 402、故障转移节点 404、以及数据库 406。在实例中, 拍快照 408a, 并在主节点 402 上执行非报告事务 410 的同时, 使用该快照 408a 在故障转移节点 404 上执行报告事务 412。在执行报告事务 412 期间, 通过事务 412 中的查询脚本创建临时表 414a 和 414b, 从而存储临时结果。将这些临时表 414a 和 414b 透明转发给主节点 402, 然后, 为临时表 414a 和 414b 分配数据库 406 中的空间。不必将随后在故障转移节点 404 处保存在临时表 414a 和 414b 中的变化转发给主节点 402。

在图 4 中, 给数据库 406 拍新的快照 408b, 从而使报告事务 412 中的后续查询能够访问临时表 414a 和 414b。然而, 在其他实施例中, 后续查询不会连续访问创建的所有临时表。因而, 在完成查询之后, 故障转移节点可以删除临时表并将该删除转发给主节点, 以释放分配给表的数据库空间。

为了确保结果一致, 单个查询通常会使用相同的快照。然而, 如图 4 的实例所示, 相同的对话或事务内的下一次查询可以使用与前一次查询所使用的快照相同的快照、或比前一次查询使用的快照更新的快照。

图 5 所示是在数据库系统中处理报告事务的另一种方法。在 502, 给数据库拍快照。在实施例中, 数据库链接至主节点和故障转移节点。然后, 在主节点上执行一个或多个非报告事务 (504), 并在主节点上执行一个或多个非报告事务的同时: 利用快照在故障转移节点上执行报告事务 (506)。在 508, 当在故障转移节点上执行报告事务时修改并使用数据库中的一个或多个模式。可能已在主节点上创建一个或多个模式, 并且为了能够被故障转移节点上的记

录事务使用，已对上述一个或多个模式进行了“标记”或“保留”。另外，在与主节点不协调的情况下，可以对一个或多个模式作出改变。

数据库模式是对象的集合。例如，模式对象包括但不限于表、视图、序列、或存储程序。表通常是数据库中的基本结构单元，并且包括存储在各个行和列中的数据。视图是在一个或多个表中的数据的定制显示。视图的数据来自于表（即，基本表）。而基本表可以是表、或可以是视图本身。视图的实例是减去表中的两列数据后得到的表。

序列是用于识别一个或多个数据库表中的数字列的唯一数字的连续清单。通常，其通过自动生成单个表或多个表的行的唯一数值来简化程序的编程。使用了序列，通常，一个以上的用户可以同时将数据键入表中。所存储的程序通常是组合在一起作为可执行单元的用于执行特定任务的一组计算机语句。

图 6 示出了具有主节点 **602**、两个故障转移节点 **604a** 和 **604b**、以及数据库 **606** 的群集 **600**。给数据库 **606** 拍快照 **608**。在实施例中，与通过快照 **608** 仅对故障转移节点 **604a** 和 **604b** 开放的数据库 **606** 的其余部分不同，在读写模式下，数据库 **606** 中的模式 **614a** 和 **614b** 可用于故障转移节点 **604a** 和 **604b**。在这种情况下，可以通过分别在故障转移节点 **604a** 和 **604b** 上执行的报告事务 **612a** 和 **612b** 来修改模式 **614a** 和 **614b**。由于在故障转移节点 **604a-604b** 和主节点 **602** 之间没有共用包含在模式 **614a** 和 **614b** 中的数据，所以在主节点 **602** 上执行的非报告事务 **610** 不能访问数据库 **606** 中的模式 **614a** 和 **614b**。

图 7 示出了用于在数据库系统中处理报告事务的方法的流程图。在 **702**，给链接至主节点和故障转移节点的数据库拍快照，在

704，在主节点上执行一个或多个非报告事务。然后，在主节点上执行一个或多个非报告事务的同时，利用快照在故障转移节点上执行报告事务（**706**）。

在实施例₁中，当在故障转移节点上执行报告事务时，访问并使用主节点上的一个或多个用户定义的程序（**708**）。用户定义的程序一般用于更简单地准备复杂报告，并且通常在主节点上被创建和编译。与任何其他数据库对象一样，可以从故障转移节点访问这些程序。

图 8 示出了数据库系统 **800**。虽然该图仅示出了用户 **802**、客户机 **804**、主节点 **806**、故障转移节点 **808**、以及数据库 **810**，但是系统 **800** 可以包括其他群集、节点、用户、数据库、和客户机。在实例₁中，用户 **802** 通过客户机 **804** 在主节点 **806** 上定义程序 **818a** 和 **818b**。在给数据库 **810** 拍快照 **812** 之后，在主节点 **806** 上执行非报告事务 **814** 的同时，使用快照 **812** 以及用户定义的程序 **818a** 和 **818b**，在故障转移节点 **808** 上执行报告事务 **816**。如图 8 所示，与用户定义的程序 **818a** 和 **818b** 不同，直接使用快照 **812**，即，在不通过主节点 **806** 的情况下使用快照 **812**。

图 9 示出了在数据库系统中处理报告事务的另一种方法。根据该方法，在 **902**，给数据库拍快照。数据库链接至主节点和次节点。然后，在 **904**，在主节点上执行一个或多个非报告事务，以及在 **906**，在主节点上执行该一个或多个非报告事务的同时，利用快照在故障转移节点上执行报告事务。在数据库中保留临时空间，并且当在故障转移节点上执行报告事务时，使用临时空间（**908**）。

为了保留数据库中的临时空间，故障转移节点由于保留通常需由主节点执行的用于避免相干性问题的目录改变而可以将消息发送给主节点。一旦为故障转移节点保留暂存盘（scratch disk）空

间，则可以在没有来自主节点的干预的情况下，执行将其自身写入临时空间的过程。暂存空间允许创建临时文件。有时需要这些临时文件来存储不适合主存储器的临时操作的结果，例如，分类的中间结果、在 JOIN 方法中使用的散列表等。

图 10 示出了具有主节点 1002 和三个故障转移节点 1004a、1004b、以及 1004c 的群集 1000，所有的节点均链接至数据库 1006。在图中，可以在主节点 1002 上找到用户定义的程序 1012 以及读写事务 1010a 和只读事务 1010b。在故障转移节点 1004a 上执行报告事务 1014c，另外，在故障转移节点 1004b 上执行报告事务 1014c，同时在故障转移节点 1004c 上执行报告事务 1014d、1014e、和 1014f。在不同的时间获取数据库 1006 的三个快照 1008a、1008b、和 1008c。可以使用这些快照中的一个来执行每个报告事务。然而，同一故障转移节点上的报告事务不必利用同一个快照。例如，故障转移节点 1004c 上的报告事务 1014d、1014e、和 1014f 中可以各自使用不同的快照 1008。

如图 10 所示，已在数据库 1006 中分别为故障转移节点 1004a、1004b、和 1004c 保留了三个临时空间 1016a、1016b、和 1016c。故障转移节点 1004a、1004b、和 1004c 中的每个发送请求给主节点 1002，以请求保留其各自的暂存空间。在其他实施例中，故障转移节点 1004a、1004b、和 1004c 可以共用一个或多个临时空间。

系统结构概述

图 11 是适用于实施本发明实施例的计算机系统 1100 的框图。计算机系统 1100 包括总线 1102 或用于传送信息的其他通信机构，它将子系统和装置（例如，处理器 1104、系统存储器 1106（例如，RAM）、静态存储装置 1108（例如，ROM）、盘驱动器 1110（例如，磁盘或光盘）、通信接口 1112（例如，调制解调器或以太网卡）、显

示器 1114 (例如, CRT 或 LCD)、输入装置 1116 (例如, 键盘)、以及光标控件 1118 (例如, 鼠标或跟踪球)) 相互连接起来。

根据本发明的一个实施例, 通过处理器 1104 执行包含在系统存储器 1106 中的一个或多个指令的一个或多个序列的处理器, 计算机系统 1100 执行特定操作。可以将这样的指令从另一个计算机可读取介质 (例如, 晶态存储装置 1108 或硬盘驱动器 1110) 读取到系统存储器 1106 中。在可选实施例中, 可以使用硬连线电路来代替软件指令或将其与软件指令相结合来实现本发明。

在本文中所使用的术语“计算机可读取介质”是指参与将用于执行的指令提供给处理器 1104 的任何介质。这种介质可以采取许多形式, 包括但不限于非易失性介质、易失性介质、和传输介质。例如, 非易失性介质包括光盘或磁盘 (例如, 盘驱动器 1110)。易失性介质包括动态存储器 (例如, 系统存储器 1106)。传输介质包括同轴电缆、铜线、以及光纤 (包括了包含总线 1102 的电线)。传输介质还可以采取声波或光波的形式 (例如, 在无线电波和红外线数据通信期间所产生的声波或光波)。

例如, 计算机可读介质的一般形式包括软盘、软磁盘、硬盘、磁带、任何其他磁介质、CD-ROM、任何其他光介质、穿孔卡、纸带、具有孔的图案的任何其他物理介质、RAM、PROM、EPROM、FLASH-EPROM、任何其他存储芯片或盒式磁带、载波、或者计算机可以读取的任何其他介质。

在本发明的实施例中, 通过单个计算机系统 1100 执行用于实现本发明的指令序列。根据本发明的其他实施例, 通过通信链路 1120 (例如, LAN、PTSN、或无线网络) 连接的两个或多个计算机系统 1100 可以彼此配合地执行用于实现本发明所需的指令序列。

计算机系统 1100 可以通过通信链路 1120 和通信接口 1112 发送和接收消息、数据、和指令（包括程序，即，程序代码）。当接收到程序代码时，可以通过处理器 1104 执行所接收到的程序代码，然后将其存储在盘驱动器 1110 或其他非易失性存储器中，以便稍后运行该程序。

在以上说明中，参考具体实施例描述了本发明。然而，显而易见的是，在不脱离本发明的精神和范围的条件下，可以对本发明进行各种修改和变化。例如，参考特定顺序的处理动作描述了上述处理流程。然而，在不影响本发明的范围或操作的情况下，可以改变所述描述的处理操作的顺序。因此，说明书和附图仅用于说明而非限制。

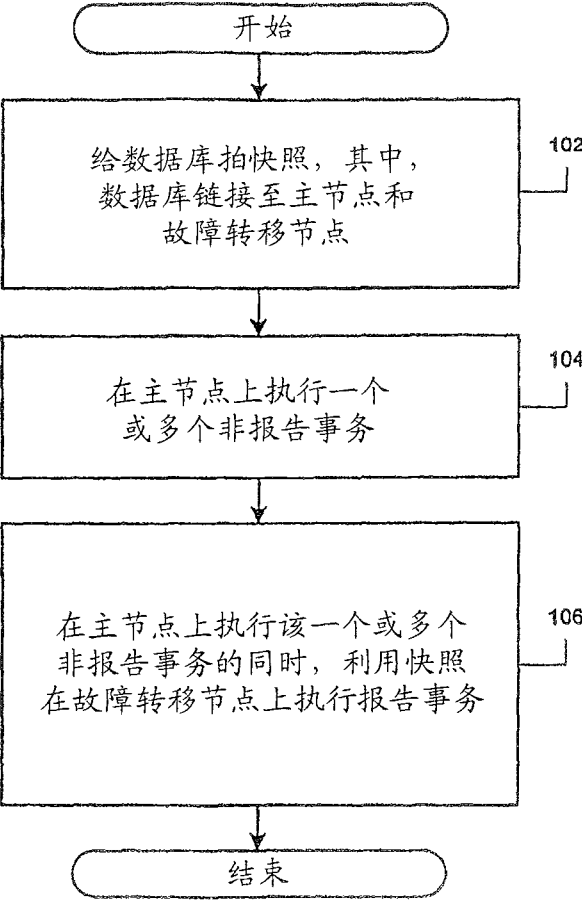


图 1

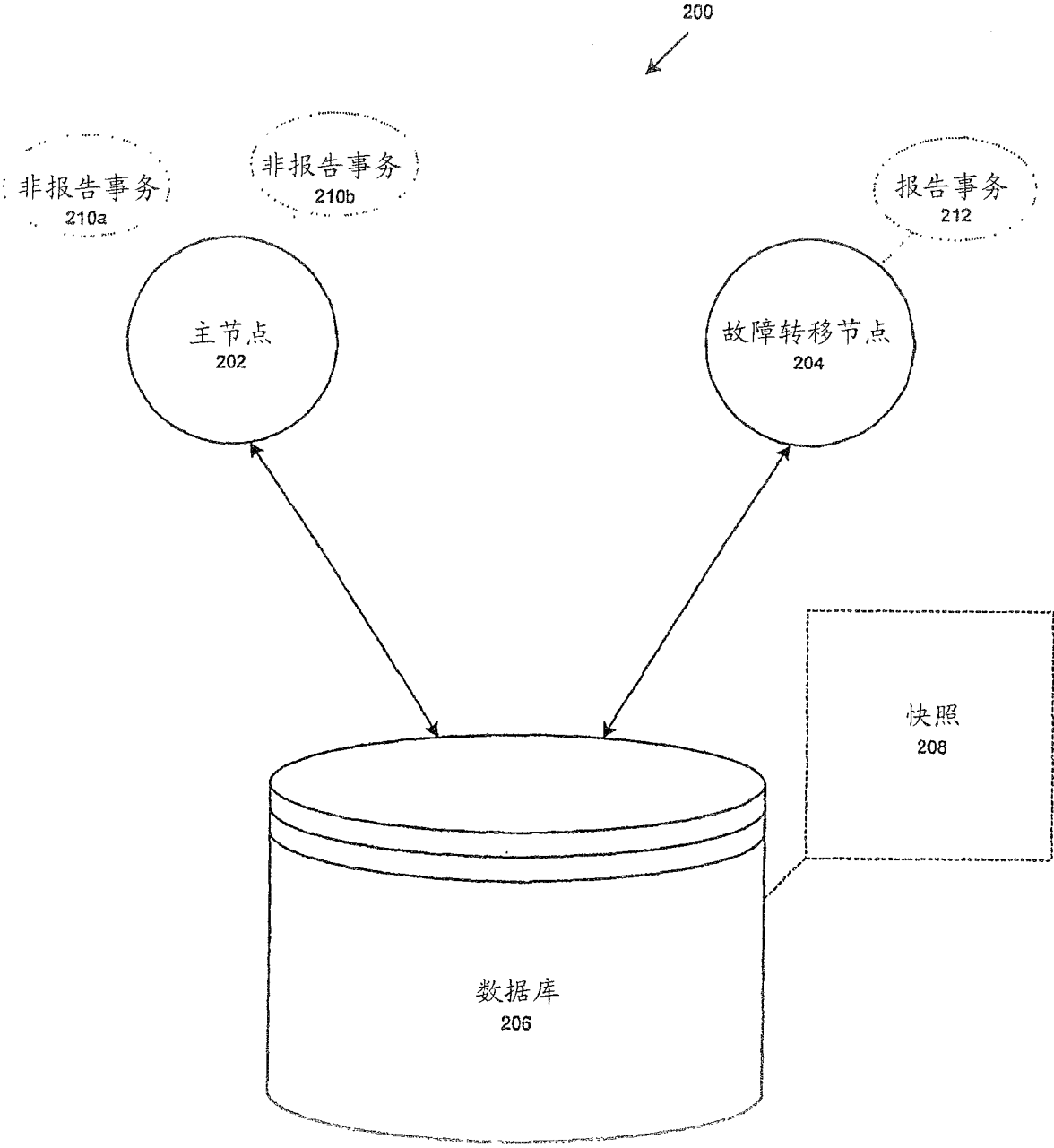


图 2

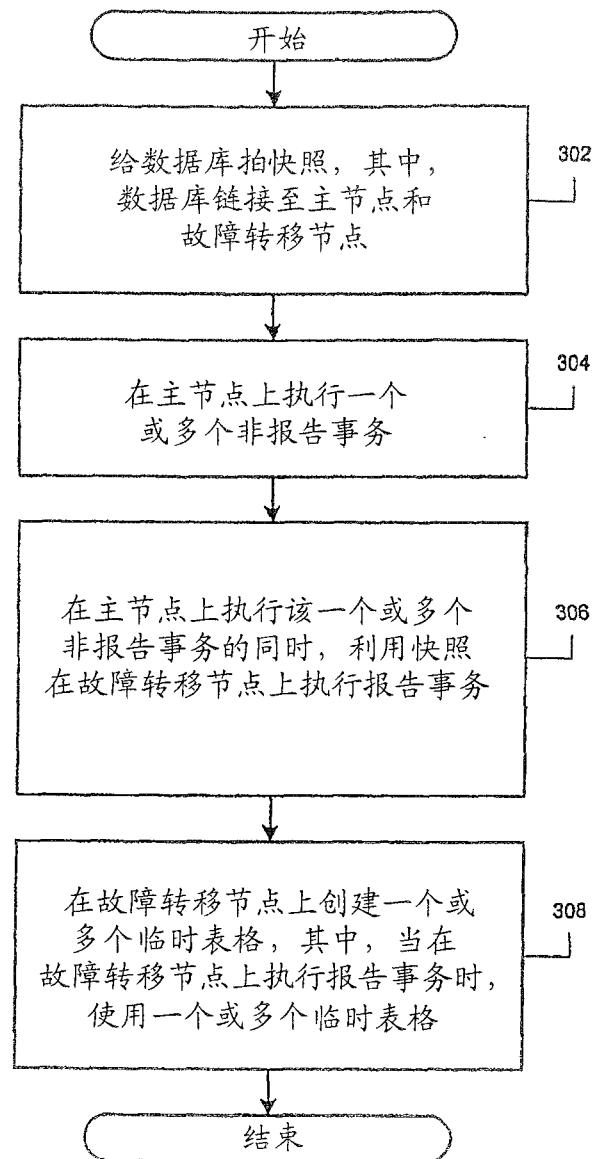


图 3

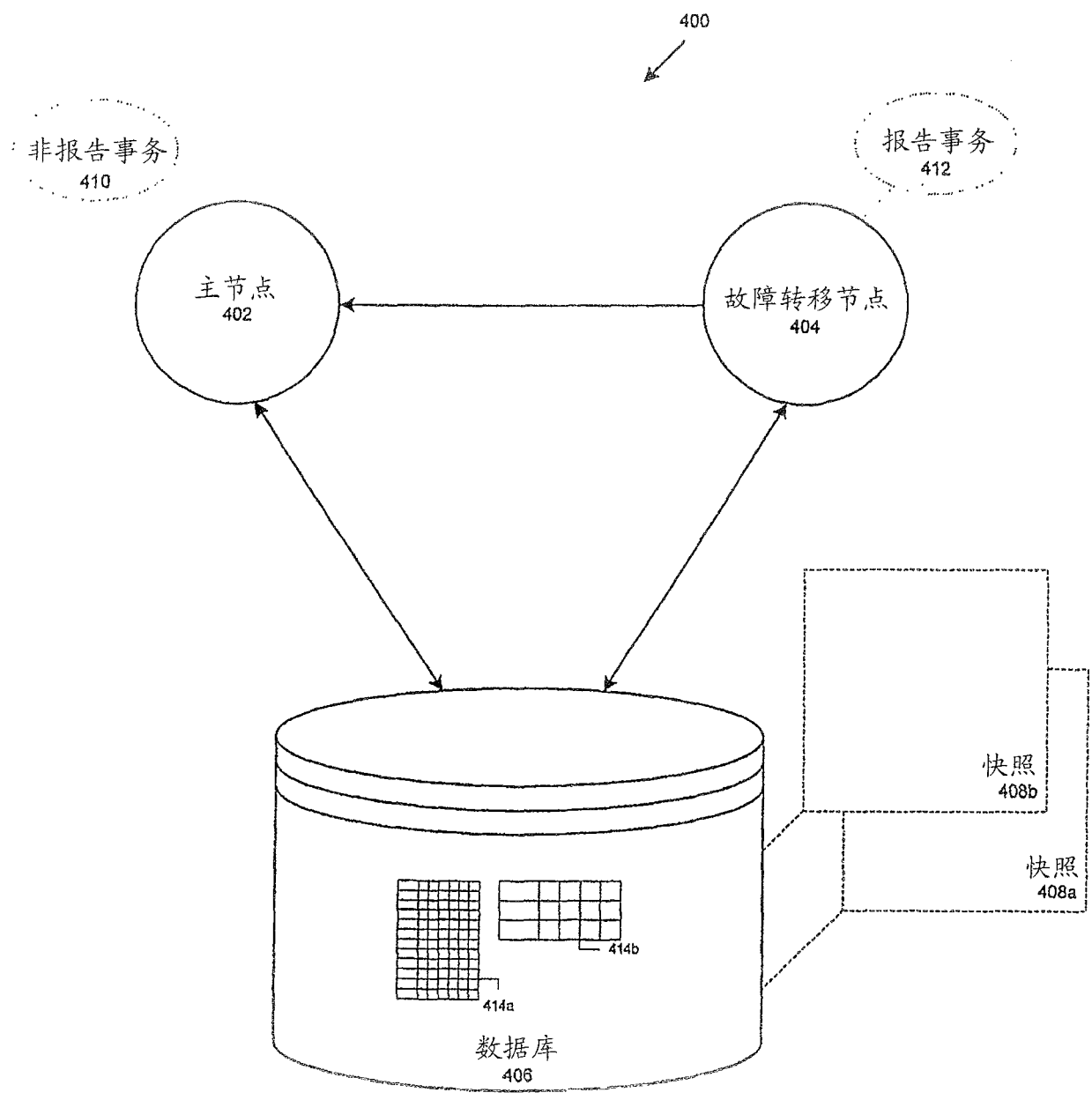


图 4

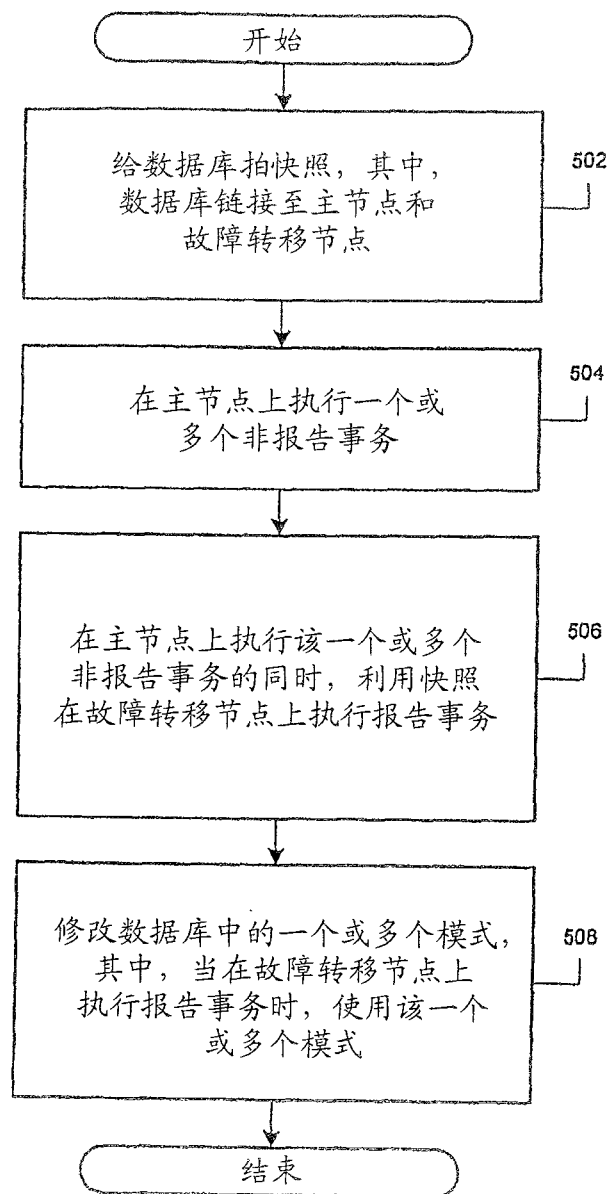


图 5

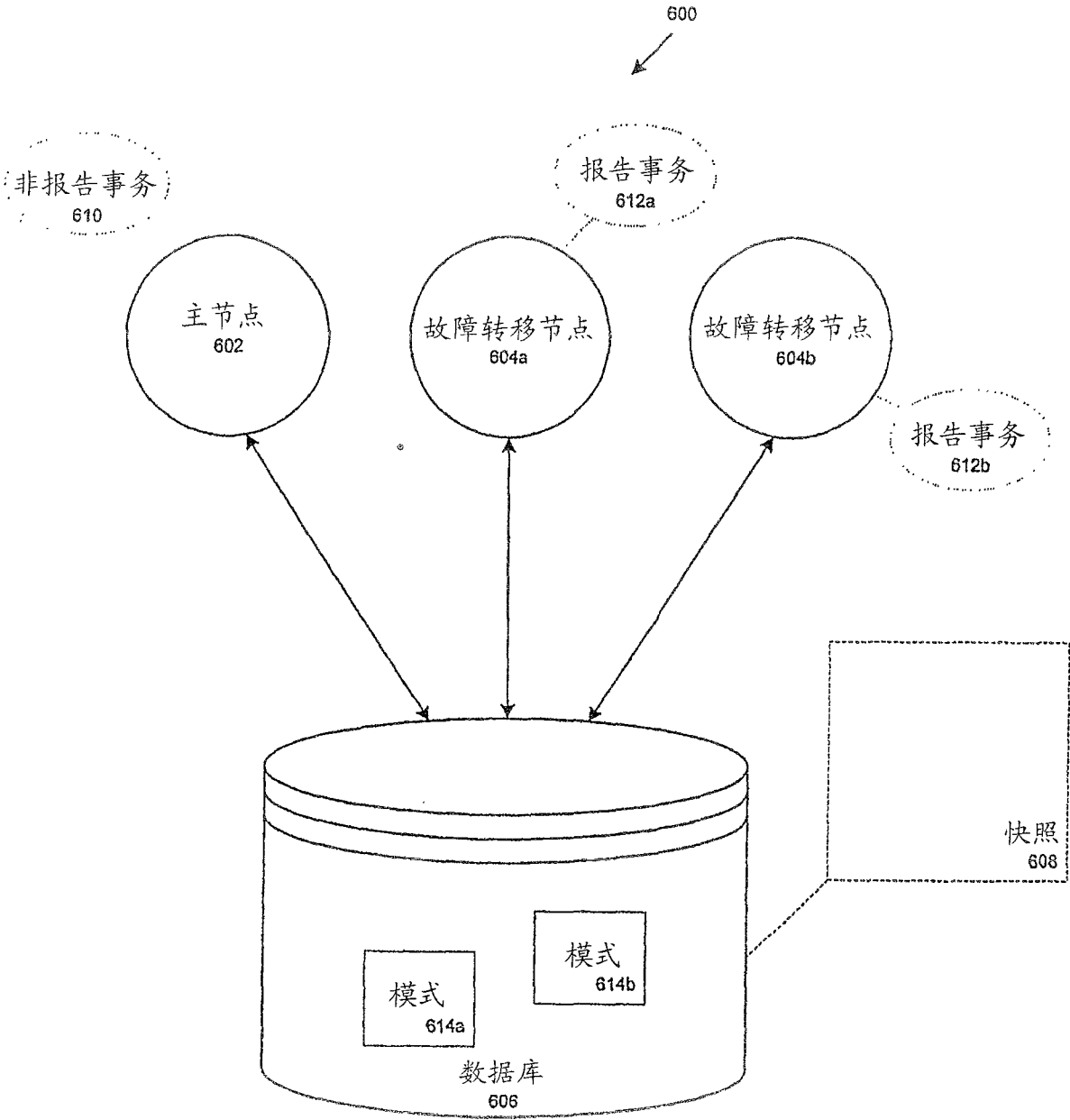


图 6

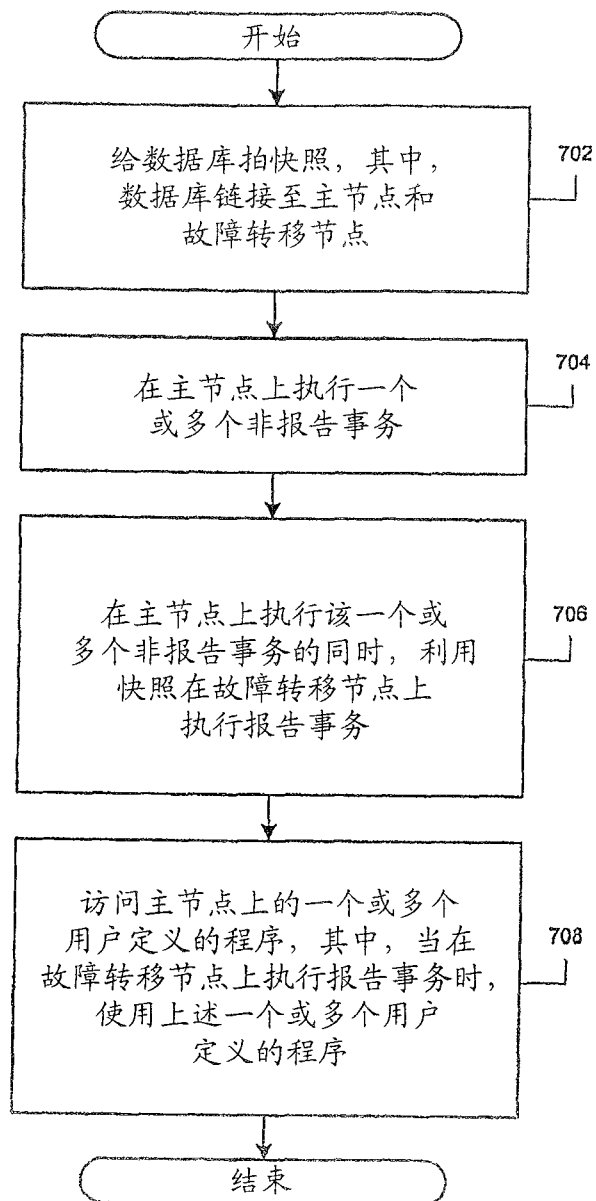


图 7

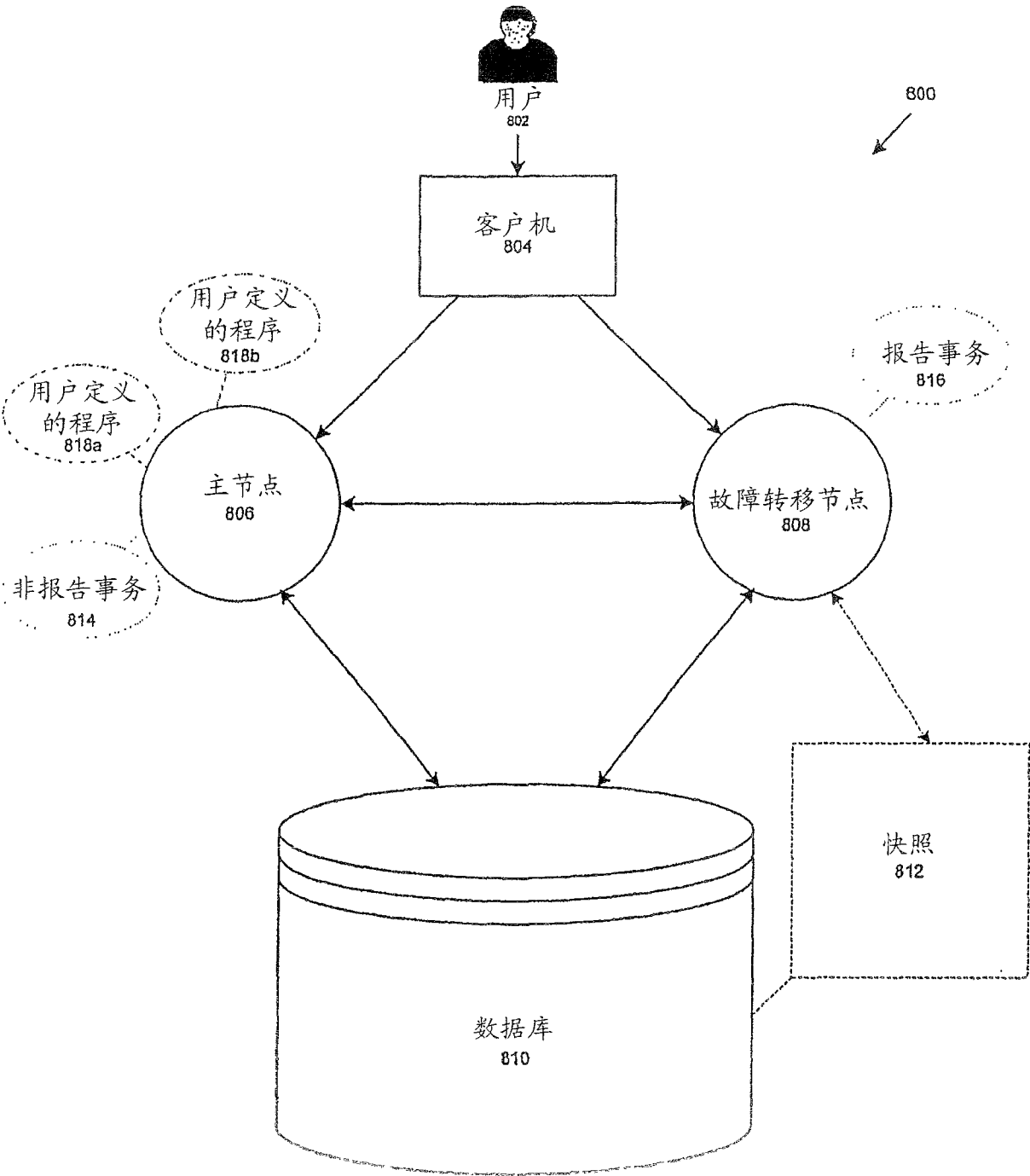


图 8

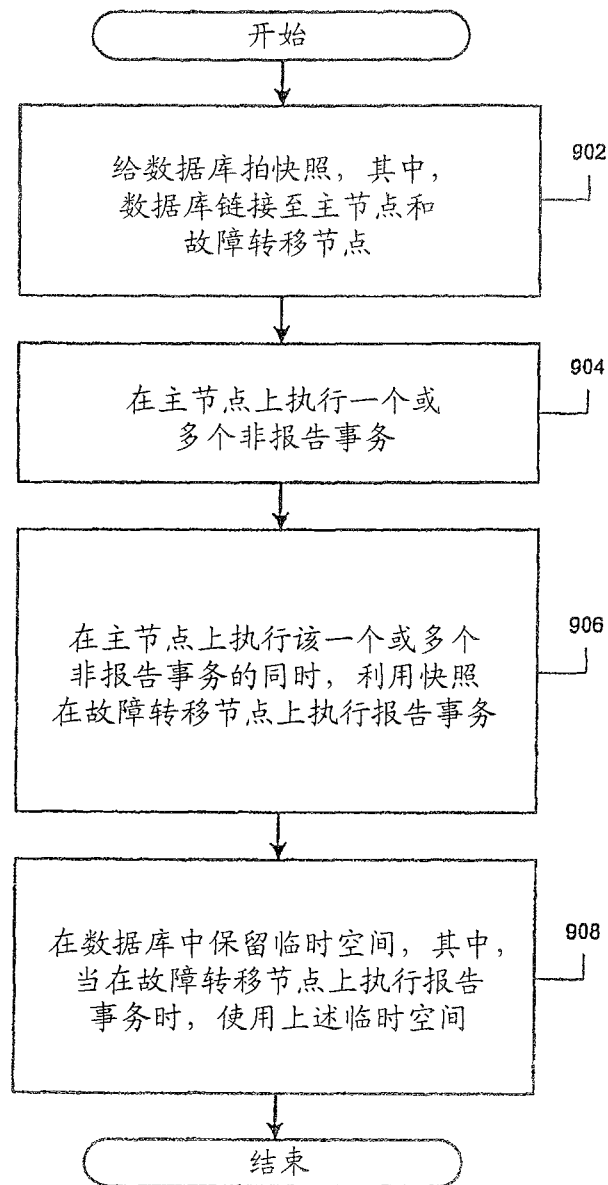


图 9

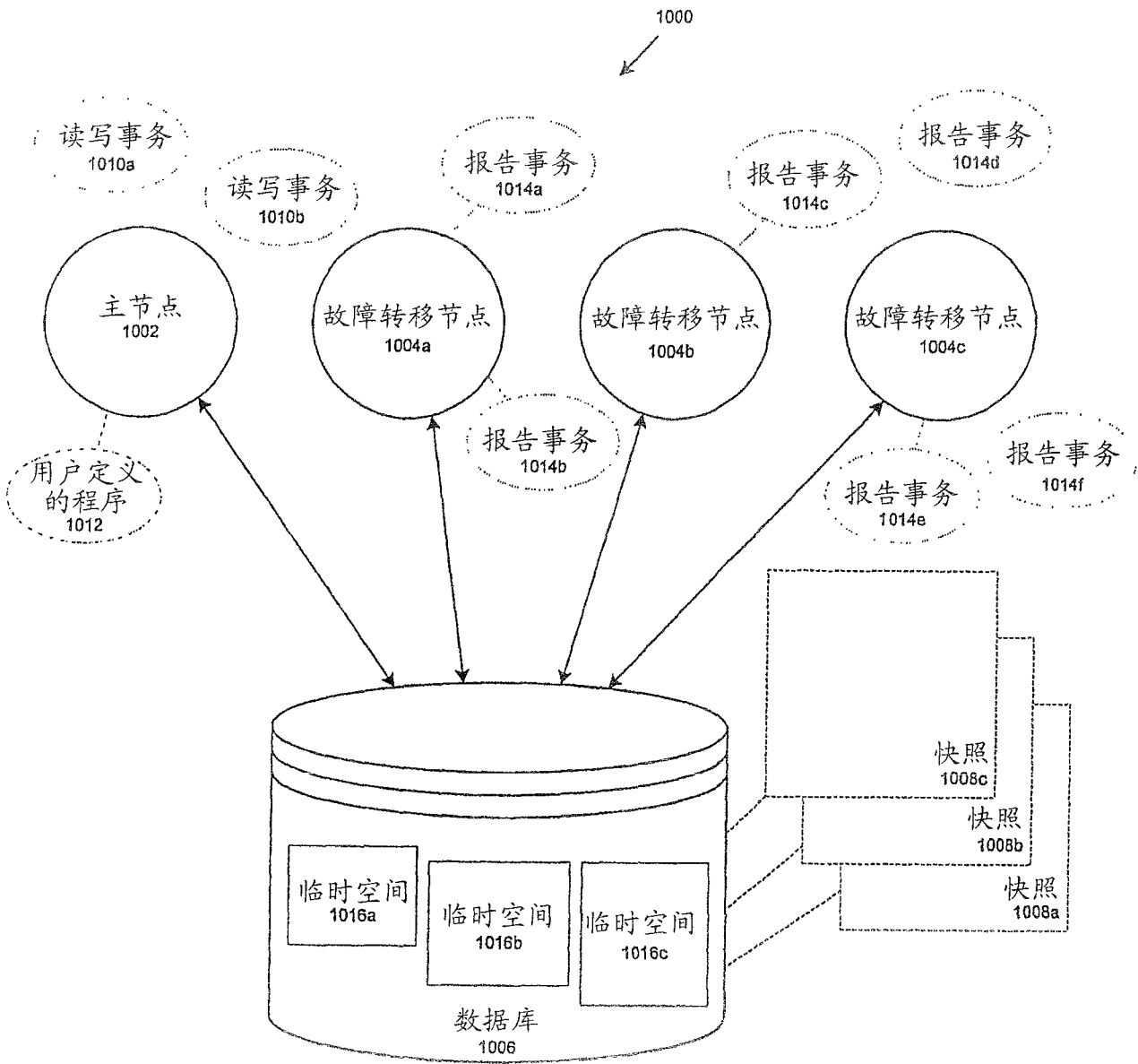


图 10

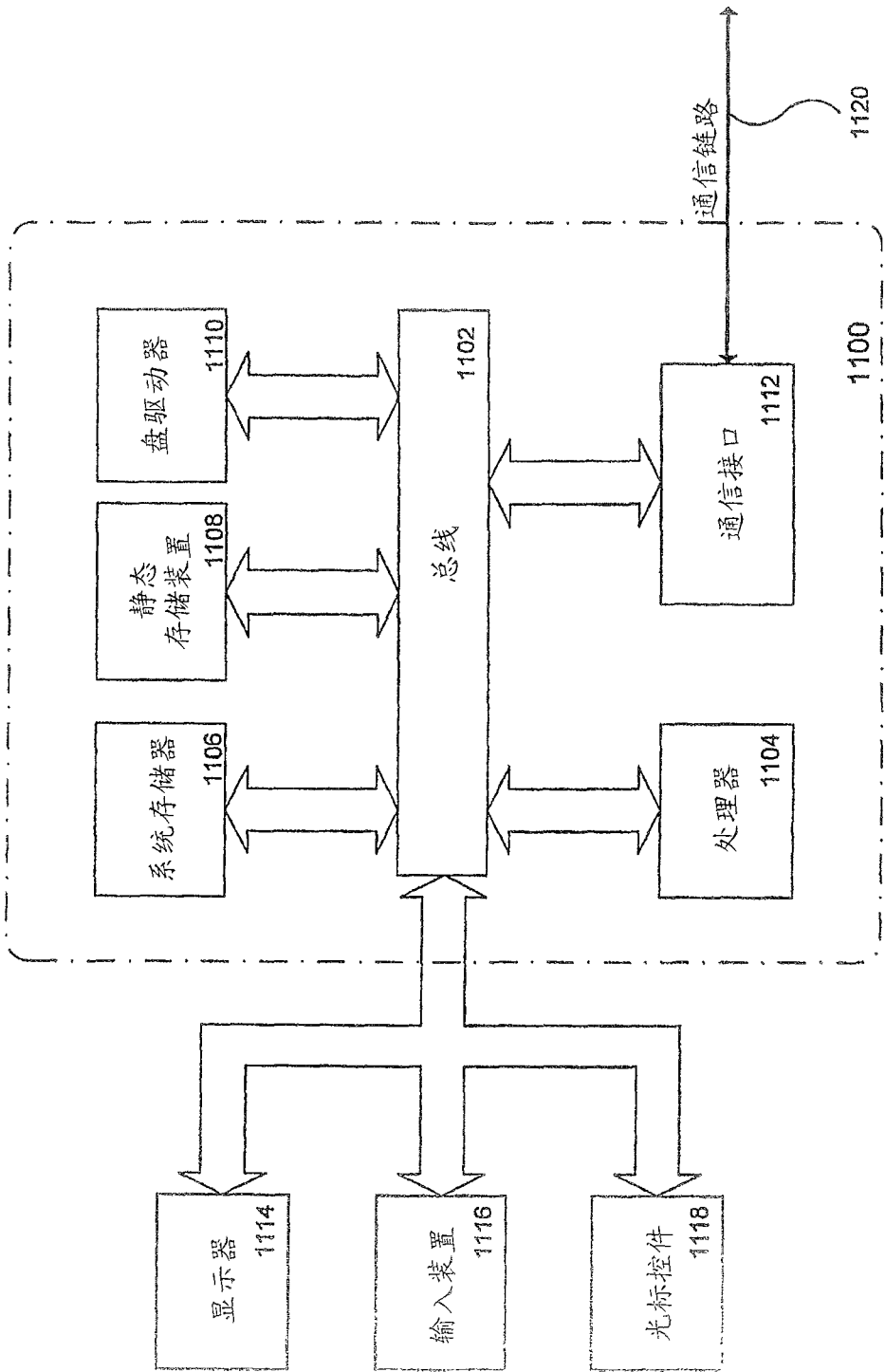


图 11