(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101243446 B
(45)授权公告日 2012.08.29

(21)申请号 200680029596.1
(22)申请日 2006.06.20
(30)优先权数据
11/204,593 2005.08.15 US
(85)PCT申请进入国家阶段日
2008.02.13
(86)PCT申请的申请数据
PCT/US2006/028344 2006.06.20
(87)PCT申请的公布数据
WO2007/021443 EN 2007.02.22
(73)专利权人 微软公司
地址 美国华盛顿州
(72)发明人 J·库勒扎 R·B·拉詹 S·R·舒米特
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 陈斌
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
US 20050071391 A1,2005.03.31, 全文.
US 20040098425 A1,2004.05.20, 说明书第【0014】,【0018】-【0019】,【0033】-【0037】,【0052】,【0054】-【0056】,【0062】,【0064】-【0065】,【0081】,【0095】段、图3.
US 6205449 B1,2001.03.20, 全文.
审查员 田志刚

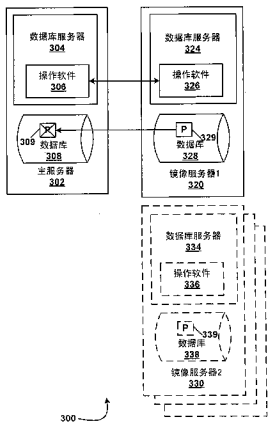
权利要求书 3 页 说明书 8 页 附图 5 页

(54)发明名称

从数据库镜像进行在线页还原

(57)摘要

一组服务器利用镜像映射的数据库的现有数据冗余度来还原页损坏。页还原可在没有从备份介质进行还原的时间和/或管理成本且没有与修复相关联的数据丢失的情况下进行。从数据库镜像进行在线页还原可由计算机系统在损坏检测后自动启动和执行。可还原整个文件或数据库而非个别页或一组页。该机制可用于从镜像还原主服务器上的损坏页,或从主服务器还原镜像上的损坏页。从数据库镜像进行在线页还原允许在无需寻找/加载/扫描并应用数据和日志备份的情况下进行页数据恢复,允许高效且可能自动的数据恢复。



CN 101243446 B

1. 一种用于还原信息的方法，包括：

从存储在第二数据库中相应的未被损坏的信息还原存储在第一数据库中的被损坏的信息，所述还原包括：

在第一数据库中锁定包含所述被损坏的信息的损坏页，其中存储在所述第一数据库中除所述被损坏的信息以外的所有信息保持可访问；

向所述第二数据库发送对于对应于所述第一数据库中的所述被损坏的信息的所述未被损坏的信息的请求，其中所述请求包括所述被损坏的信息的标识符以及与在所述第一数据库中检测到所述被损坏的信息的时刻相关联的日志序列号；

保持等待直到与所述第二数据库相关联的事务日志至少已经被应用到第二数据库中所接收的日志序列号所指示的点上；

从所述第二数据库接收对应于所述第一数据库中的所述被损坏的信息的所述未被损坏的信息；

用来自所述第二数据库的所述未被损坏的信息替换所述第一数据库中的所述被锁定的被损坏的信息以及

对所述第一数据库中被锁定的损坏页进行解锁。

2. 如权利要求 1 所述的方法，其特征在于，所述第一数据库是镜像数据库，所述第二数据库是主数据库。

3. 如权利要求 1 所述的方法，其特征在于，所述第一数据库是主数据库，所述第二数据库是镜像数据库。

4. 如权利要求 3 所述的方法，其特征在于，所述镜像数据库是多个镜像数据库中的第一镜像数据库。

5. 如权利要求 4 所述的方法，其特征在于，所述第一镜像数据库是基于哪个镜像数据库是最新的来选择的。

6. 如权利要求 1 所述的方法，其特征在于，所述方法还包括在检测到所述被损坏的信息之后从所述第二数据库自动还原所述被损坏的信息。

7. 如权利要求 1 所述的方法，其特征在于，所述第二数据库对查询进行服务，该查询是针对所述第一数据库的被损坏的信息的查询。

8. 如权利要求所述的方法，其特征在于，所述被损坏的信息包括所述第一数据库的一页、所述第一数据库的一组页或所述第一数据库。

9. 一种用于还原存储在第一数据库中的被损坏的信息的计算机实现的方法，包括：

响应于检测到存储在所述第一数据库中的被损坏的信息，在第一数据库中锁定包含所述被损坏的信息的损坏页并在无需人工干预的情况下启动对所述被损坏的信息的还原，除存储所述被损坏的信息的所述第一数据库的部分以外的所述第一数据库仍保持可访问；

向第二数据库发送对一个或一组页的请求，所述一个或一组页包括所述第二数据库中对应于所述第一数据库中所述被损坏的信息的未被损坏的信息，所述请求包括所述被损坏的信息的标识符以及与在所述第一数据库中检测到所述被损坏的信息的时刻相关联的日志序列号；

保持等待直到与所述第二数据库相关联的事务日志至少已经被应用到第二数据库中所接收的日志序列号所指示的点上；

从所述第二数据库接收所述相应的未被损坏的信息，并将相应的未被损坏的信息应用于所述第一数据库；以及

对所述第一数据库的被锁定的损坏页进行解锁；

其中，所述第二数据库是多个镜像数据库中的第一镜像数据库；

其中，负载平衡是通过向所述第一镜像数据库发送对要被还原到所述第一数据库的第一页范围的第一请求，并向所述多个镜像数据库中的第二镜像数据库发送对要被还原到所述第一数据库的第二页范围的第二请求而实现的；

其中，所述第一数据库对所述多个镜像数据库中的哪个镜像数据库响应最快保持跟踪，并从响应最快的镜像数据库请求还原页。

10. 如权利要求 9 所述的方法，其特征在于，还包括从所述第二数据库对信息的请求进行服务，该请求是针对所述第一数据库中所述被损坏的信息。

11. 如权利要求 9 所述的方法，其特征在于，还包括基于哪个镜像数据库是最新的或响应时间来选择所述第二数据库。

12. 如权利要求 9 所述的方法，其特征在于，还包括将所述未被损坏的信息存到所述第一数据库中。

13. 一种用于还原信息的方法，包括：

在第一数据库中锁定包含被损坏的信息的数据库的损坏页，所述第一数据库是主数据库；

接收包括以下项的用户输入：所述第一数据库要还原的页的页标识符，所述页包括所述第一数据库上存储所述被损坏的信息的部分，向其请求由所述用户输入的所述页标识符标识的第二数据库的相应页的第二数据库；

接收与在所述第一数据库中检测到所述被损坏的信息的时刻相关联的日志序列号；

标识所述第二数据库并在无需人工干预的情况下启动所述被损坏的信息的还原，除了所述第一数据库上存储所述被损坏的信息的所述页之外的所述第一数据库仍保持可用；

向第二数据库发送对一个或一组页的请求，所述一个或一组页包括所述第二数据库中对应于所述第一数据库中所述被损坏的信息的未被损坏的信息，所述请求包括所述日志序列号；

保持等待直到与所述第二数据库相关联的事务日志至少已经被应用到第二数据库中所接收的日志序列号所指示的点上；

从所述第二数据库接收所述相应的未被损坏的信息；

将所述未被损坏的信息应用于所述第一数据库；以及

对所述第一数据库的被锁定的损坏页进行解锁。

14. 如权利要求 13 所述的方法，其特征在于，还包括：

将所述未被损坏的信息存到所述第一数据库中。

15. 如权利要求 13 所述的方法，其特征在于，还包括：

基于所述第二数据库是最新的来选择所述第二数据库，所述第二数据库包括与所述第一数据库相关联的多个镜像数据库中的第一镜像数据库。

16. 如权利要求 13 所述的方法，其特征在于，还包括：

基于所述第二数据库的响应时间来选择所述第二数据库，所述第二数据库包括与所述

第一数据库相关联的多个镜像数据库中的第一镜像数据库。

17. 如权利要求 15 或 16 所述的方法，其特征在于，还包括：

通过向第一镜像数据库发送页号的第一范围并向第二镜像数据库发送页号的第二范围来对所述被损坏的信息的还原进行负载平衡。

从数据库镜像进行在线页还原

[0001] 背景

[0002] 存储在计算机上的数据每天都在被丢失和损坏 (corruption)。事故、人为错误、病毒攻击、硬件故障和电源问题仅是存储在计算机上的信息丢失和损坏的数千种可能的原因中的某些。为了针对未预料的数据丢失进行保护，聪明的个人（和企业）通常备份其文件。可通过在某种可移动介质上使用备份实用程序简单地制作文件或文件集的副本来进行备份以便在发生故障或原始数据丢失的情况中使用，或者在复制数据时也可对其压缩。当数据丢失或数据损坏发生时，一般从备份中还原受损或丢失的一个或多个文件。就该意义而言，“还原”意味着从可移动介质复制回计算机或如果使用了数据实用程序，则复制数据并对其解压缩。当文件较小且当备份可用时，从备份还原文件是重新获得信息的方便且高效的方式。

[0003] 随着文件随时间改变的大小、重要性和 / 或程度的增加，周期性取得的文件的简单副本不再如此吸引人。例如，假定一企业依赖于频繁改变的一组非常大的文件的可靠的可用性，这种情况例如将在由航空公司维护的数据库文件中发生。数据的周期性快照（在特定时刻取得的一组文件和目录）可能不再是充分的。镜像可能是更好的选择。计算中的镜像是数数据集的直接副本，使得在分开的机器上存在数据的精确重复的副本。这些副本被创建，然后被持续更新，使得副本保持与重要数据库同步。镜像可被维护为硬件级的物理副本或通过数据库机制（有时被称为“复制”）。镜像与快照的不同之处在于，快照表示文件或数据库在特定时刻的状态。相反，镜像是保持随时跟上动态改变的源的活动的、动态的副本。

[0004] 当数据库的小部分被损坏时，从备份还原整个数据库的选择不是最佳的，因为所执行的大多数工作是不必要的（数据库中的大部分是良好的）。还原过程缓慢，需要处理外部介质（备份带或备份盘），且要求人工干预（数据库管理员要选择使用哪些备份等，例如计算机操作员要找到并加载带子，或其它）。而且，在还原过程进行时，数据库一般不可供用户使用。处理页损坏的另一方式是尝试修复该页。修复页较快速，但几乎总是导致页数据的部分或完全丢失，引起数据库内的逻辑不一致性。

[0005] 如果存在快速且不会导致数据丢失或数据不一致性的重新获得损坏的页（页是由 DBMS 识别为一单位的固定数目字节的数据，通常为 8K 字节）上存储的数据的方式，将是有益的。使该过程在检测到数据损坏时自动启动而无需人工干预来进行，不要求对带或其他可移动介质的管理和处理，将是有用的。

[0006] 概述

[0007] 一组服务器利用了镜像映射的数据库的现有数据冗余度来还原页损坏。页还原可在没有从备份介质进行还原的时间和 / 或管理成本且没有与修复相关联的数据丢失的情况下进行。而且，从数据库镜像进行在线页还原可由计算机系统在损坏检测后自动启动和执行。该概念可被扩展来允许还原整个文件或数据库而非个别页或一组页。该机制可用于从镜像还原主服务器上的损坏页，或从主服务器还原镜像上的损坏页。从数据库镜像进行在线页还原允许几乎即时的页损坏修补而没有数据丢失。它也允许在无需寻找 / 加载 / 扫

描并应用数据和日志备份的情况下进行页数据恢复，允许高效且可能自动地数据恢复。

[0008] 因此可在无需提供备份或甚至备份不存在的情况下执行还原。可向一个或多个镜像请求一个或多个页，且可执行验证以确保所返回的页按时赶上主服务器在损坏检测时的页（当页请求由镜像接收时，镜像上的“重做”操作可能未赶上主服务器上的“做”操作）。可在崩溃恢复情形期间或正常操作期间检测到损坏时自动修补页损坏。在崩溃恢复期间，延迟（原文 deterring，错）事务卷回的被损坏的页可在无需人工干预的情况下被自动还原，使得能够在无需个人干预的情况下进行延期事务的卷回。当有多个镜像可用时，被选中来返回所请求页的镜像可基于哪一镜像在历史上具有最快响应时间或基于哪一镜像在沿重新播放来自主服务器的日志中前进最远（即，哪一镜像是最新的）来选择。可对多个镜像上的多页还原进行负载平衡。可在损坏检测之后自动执行一个或多个页还原，或者页还原可以是用户驱动的。可从镜像提供页来用于只读查询直到主服务器上的损坏被修补。可从镜像提供页来用于读 / 写查询，直到主服务器上的损坏被修补。或者，镜像可变为主服务器。主服务器上的损坏可从镜像还原，反之，镜像上的损坏可从主服务器还原。

[0009] 附图简述

[0010] 附图中：

[0011] 图 1 是示出可在其中实现本发明的各方面的示例性计算环境的框图；

[0012] 图 2 是示出如本领域中已知的用于还原数据库中的页的系统的框图；

[0013] 图 3 是示出根据本发明的某些实施例用于从数据库镜像进行在线页还原的系统的框图；

[0014] 图 4 是示出如本领域中已知的用于还原页的方法的流程图；以及

[0015] 图 5 是示出根据本发明的某些实施例用于从数据库镜像进行在线页还原的方法的流程图。

[0016] 详细描述

[0017] 概观

[0018] 图 2 是如本领域中已知的用于还原数据库中的页的系统 200 的框图。诸如单机服务器 202 的计算机上的数据库服务器 204，诸如 Microsoft 的 SQL Server、IBM 的 DB2、Oracle 等可包括修复 / 还原软件 216，该软件允许从图 2 中由备份带 206 等表示的一个或多个备份介质中还原其中一部分被损坏（即，损坏页 208）的数据库 210，这要求用户干预，如用户输入 218（例如，来自计算机操作员和 / 或数据库管理员）所表示的。图 4 是如本领域中已知的用于还原数据库中的页的方法的流程图。在 402，检测到数据库页损坏。此时，一般数据库变为不可用。在 404，数据库管理员或其他人必须决定如何继续。例如，假定 DBA 在 406 启动页修复。在 408，服务器上的软件通常尝试来修补页内容。通常，丢失数据和商业逻辑，且导致了数据库内的不一致性。在 418，一旦修复完成，数据库再次变为可用。或者，在 410，DBA 决定从备份还原该数据库页。在 412，可为所需页寻找、加载和扫描备份介质，或者如果从备份介质重新加载整个数据库，则所有的备份介质必须被按顺序加载以便应用于数据库。在 414，找到该页并将其应用于数据库，或者将整个备份集应用于数据库。在 416，通过应用一个或多个日志来使页面保持最新，且在 418，一旦还原完成，数据库再次变为可用。可以理解，修复选择（步骤 404-408 和 418）可能导致数据丢失和数据库不一致性。还原选择（步骤 404、410-418）可能是漫长的过程，且要求至少一个人的干预。在这两

种选择中，数据库在修复或还原过程期间一般不可用。

[0019] 根据本发明的实施例，以下将更全面地描述完全基于软件的灾难恢复解决方案。简而言之，一个简单情形可能是：

[0020] 1. 检测到损坏页

[0021] 2. 数据库管理员执行还原功能（例如，“从镜像中对页 x 还原数据库”）或还原功能由计算机自动启动（无需人工干预）

[0022] 3. 服务器在数据库中锁定损坏页

[0023] 4. 从主服务器向至少一个镜像发送要求该数据库页的请求。该请求包括一个或多个被损坏页的页标识符和主服务器上当前时刻的日志序列号 (LSN)。提供当前时刻的日志序列号是因为不能够信任被损坏页上的 LSN。

[0024] LSN 是重要的，因为 SQL Server 将对数据库进行的改变写入事务日志，使得如果事务开始但无法完成，则来自该日志的改变可被检索并重新应用（“卷回”）或被取消完成。当事务提交时，SQL Server 将关于该事务的所有日志记录写入磁盘上的持久存储。因此，即使系统在 SQL Server 将改变后的数据页写入磁盘之前发生故障，日志记录也位于磁盘上。当 SQL Server 再次启动时，日志提供恢复或前进 (roll forward) 已完成但其相应的数据页未写入磁盘的任何事务的足够信息。写入事务日志的每一记录被分配（一般递增的）顺序日志序列号，提供了容易的方式以跟踪任务被应用的次序。

[0025] 5. 镜像等待其“重做”操作以通过该请求中提供的 LSN，以确保对所请求的页的所有改变都从日志中被重新播放了，并将其应用于该页。

[0026] 6. 镜像从其缓冲池或其磁盘上取回该页。可保证现在由镜像持有的该页与主服务器上的数据库一致，因为在这期间不能对该页进行更新（因为在步骤 3 中将其锁定），且日志被重新播放以通过该锁定点（步骤 5）。

[0027] 7. 使用数据库镜像映射通信基础架构，使用新消息类型将该页从镜像发送到主服务器。

[0028] 8. 在接收该页之后，主服务器将该页写入磁盘以保存还原并释放锁定，使得该被修补的页再次可用于查询。

[0029] 可提供与在镜像不可用时会发生什么、在镜像挂起时会发生什么、如果在还原操作期间触发了数据库镜像映射故障转移则会发生什么等相关联的错误处理。

[0030] 在自动模式中，在崩溃恢复期间或常规操作期间检测到损坏页之后，可在无需人工干预的情况下由计算机自动启动该过程。当检测到损坏页时，自动对于对该页进行更新的事务保持锁定。延期事务是（异常中止或提交）直到某些外部事件发生之前不能被解决的事务。在本上下文中，所指示的事件是恢复一致性页，这可被自动生成。（传统上，所引起的“延期操作”需要管理员干预来解决底层问题。）当调用从数据库镜像中进行在线页还原的机制的自动模式特征时，损坏页（由页 id 标识）被锁定，从镜像完成页还原，然后可调用用于卷回延期事务的代码，得到对数据库页损坏的无缝修复。

[0031] 所述技术可被扩展来从镜像中还原整个文件（例如，在磁盘崩溃的情况中）。或者，数据库可能故障而转移到所述镜像，所述镜像成为活动（主）数据库，而发生故障的副本成为镜像。在此情况中，通过从新主服务器运送数据，镜像可变为自动修复的目标。如果调用手动模式（需要人工干预），则对原始位置不可用的情况下，可任选地指定文件位置。

对于自动模式，可尝试默认位置，否则服务器可等待手动操作被执行。

[0032] 如果自动模式被关闭或未被实现，则可使用底层页还原机制来从镜像提供用于只读操作的页，直到损坏页被修复。在此选项中，上述方法仍被遵循，但从镜像接收到的页不被写回到主服务器上的磁盘。这允许更大的数据可用性同时仍允许管理员维持对还原的手动控制。

[0033] 优化包括：

[0034] 1) 采用多个镜像，主服务器跟踪最快响应的镜像（较快响应时间可能由于多种因素，包括网络差异、物理位置等），并从最快响应的镜像请求还原页。

[0035] 2) 如果多个镜像处于追逐从主服务器接收的日志的重做操作的不同阶段，则主服务器可要求用于还原页的最新的（当前）镜像。

[0036] 3) 当还原多个页时，可通过向不同镜像要求页的块来对镜像进行负载平衡。

[0037] 示例性计算环境

[0038] 图 1 和以下讨论旨在提供对可在其中实现本发明的合适的计算环境的简要一般描述。然而，应理解，构想了供结合本发明使用的手持、便携式和所有种类的其他计算设备。尽管以下描述通用计算机，但这仅是一个示例，本发明仅要求具有网络服务器互操作性和交互能力的瘦客户机 (thin client)。因此，本发明可在隐含很少或最少客户机资源的网络托管服务的环境中实现本发明，例如其中客户机设备仅用作万维网的浏览器或接口的联网环境。

[0039] 尽管不是必需的，但本发明可由应用程序编程接口 (API) 实现，供用户使用和 / 或包括在网络浏览软件中，它将在诸如程序模块等由诸如客户机工作站、服务器或其他设备的一台或多台计算机执行的计算机可执行指令的一般上下文中描述。一般而言，程序模块包括例程、程序、对象、组件、数据结构等，它们执行特定任务或实现特定抽象数据类型。一般，程序模块的功能可在各个实施例中按需组合或分布。而且，本领域的技术人员可以理解，本发明可以使用其它计算机系统配置来实现。适合于本发明使用的其他公知计算系统、环境和 / 或配置包括但不限于：个人计算机 (PC)、自动柜员机、服务器计算机、手持或膝上型设备、多处理器系统、基于微处理器的系统、可编程消费者电子产品、网络 PC、小型机、大型计算机等。本发明也可以在分布式计算环境中实现，其中任务由通过通信网络或其他数据传输介质连接的远程处理设备来执行。在分布式计算环境中，程序模块可以位于包括存储器存储设备的本地和远程计算机存储介质中。

[0040] 图 1 示出了可在其中实现本发明的合适的计算系统环境 100 的示例，尽管如上已很清楚，计算系统环境 100 只是合适的计算环境的一个示例，并不旨在对本发明的使用范围或功能提出任何限制。也不应该把计算环境 100 解释为对示例性操作环境 100 中示出的任一组件或其组合有任何依赖性或要求。

[0041] 参考图 1，用于实现本发明的一个示例性系统包括计算机 110 形式的通用计算设备。计算机 110 的组件可以包括，但不限于，处理单元 120、系统存储器 130 和将包括系统存储器在内的各种系统组件耦合至处理单元 120 的系统总线 121。系统总线 121 可以是若干类型的总线结构中的任一种，包括存储器总线或存储器控制器、外围总线和使用各种总线体系结构中的任一种的局部总线。作为示例，而非限制，这样的体系结构包括工业标准体系结构 (ISA) 总线、微通道体系结构 (MCA) 总线、扩展的 ISA(EISA) 总线、视频电子技术标准

协会 (VESA) 局部总线和外围部件互连 (PCI) 总线 (也被称为 Mezzanine 总线)。

[0042] 计算机 110 通常包括各种计算机可读介质。计算机可读介质可以是能够被计算机 110 访问的任何可用介质，且包括易失性和非易失性介质、可移动和不可移动介质。作为示例，而非限制，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以任何方法或技术实现的用于存储诸如计算机可读指令、数据结构、程序模块或其它数据等信息的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括，但不限于，RAM、ROM、EEPROM、闪存或其它存储器技术 ;CD-ROM、数字多功能盘 (DVD) 或其它光盘存储 ; 磁带盒、磁带、磁盘存储或其它磁性存储设备 ;或能用于存储所需信息且可以由计算机 110 访问的任何其它介质。通信介质通常具体化为诸如载波或其它传输机制等已调制数据信号中的计算机可读指令、数据结构、程序模块或其它数据，且包含任何信息传递介质。术语“已调制数据信号”指的是这样一种信号，其一个或多个特征以在信号中编码信息的方式被设定或更改。作为示例，而非限制，通信介质包括诸如有线网络或直接线连接的有线介质，以及诸如声学、RF、红外线和其它无线介质的无线介质。上述中任一个的组合也应包括在计算机可读介质的范围之内。

[0043] 系统存储器 130 包括易失性和 / 或非易失性存储器形式的计算机存储介质，诸如只读存储器 (ROM)131 和随机存取存储器 (RAM)132。基本输入 / 输出系统 133(BIOS) 包含有助于诸如启动时在计算机 110 中元件之间传递信息的基本例程，它通常被存储在 ROM131 中。RAM132 通常包含处理单元 120 可以立即访问和 / 或目前正在操作的数据和 / 或程序模块。作为示例，而非限制，图 1 示出了操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137。

[0044] 计算机 110 也可以包括其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质。仅作为示例，图 1 示出了从不可移动、非易失性磁介质中读取或向其写入的硬盘驱动器 141，从可移动、非易失性磁盘 152 中读取或向其写入的磁盘驱动器 151，以及从诸如 CD ROM 或其它光学介质等可移动、非易失性光盘 156 中读取或向其写入的光盘驱动器 155。可以在示例性操作环境下使用的其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质包括，但不限于，盒式磁带、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等。硬盘驱动器 141 通常由诸如接口 140 的不可移动存储器接口连接至系统总线 121，磁盘驱动器 151 和光盘驱动器 155 通常由诸如接口 150 的可移动存储器接口连接至系统总线 121。

[0045] 以上描述和在图 1 中示出的驱动器及其相关联的计算机存储介质为计算机 110 提供了对计算机可读指令、数据结构、程序模块和其它数据的存储。例如，在图 1 中，硬盘驱动器 141 被示为存储操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147。注意，这些组件可以与操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137 相同或不同。操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147 在这里被标注了不同的标号是为了说明至少它们是不同的副本。用户可以通过输入设备，诸如键盘 162 和定点设备 161(通常指鼠标器、跟踪球或触摸垫) 向计算机 110 输入命令和信息。其它输入设备 (未示出) 可以包括话筒、操纵杆、游戏手柄、圆盘式卫星天线、扫描仪等。这些和其它输入设备通常由耦合至系统总线 121 的用户输入接口 160 连接至处理单元 120，但也可以由其它接口或总线结构，诸如并行端口、游戏端口或通用串行总线 (USB) 连接。

[0046] 监视器 191 或其它类型的显示设备也经由接口，诸如视频接口 190 连接至系统总

线 121。诸如北桥的图形接口 182 也可连接至系统总线 121。北桥是与 CPU 或主机处理单元 120 通信的芯片组，并承担加速图形端口 (AGP) 通信的责任。一个或多个图形处理单元 (GPU)184 可与图形接口 182 通信。就此而言，GPU184 一般包括片上存储器存储，诸如寄存器存储，且 GPU184 与视频存储器 186 通信。然而，GPU184 仅是协处理器的一个示例，因此可在计算机 110 中包括各种协处理设备。监视器 191 或其它类型的显示设备也经由接口，诸如视频接口 190 连接至系统总线 121，视频接口 190 又与视频存储器 186 通信。除监视器 181 以外，计算机也可以包括其它外围输出设备，诸如扬声器 197 和打印机 196，它们可以通过输出外围接口 195 连接。

[0047] 计算机 110 可使用至一个或多个远程计算机，诸如远程计算机 180 的逻辑连接在网络化环境下操作。远程计算机 180 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其它常见网络节点，且通常包括上文相对于计算机 110 描述的许多或所有元件，尽管在图 1 中只示出存储器存储设备 181。图 1 中所示逻辑连接包括局域网 (LAN)171 和广域网 (WAN)173，但也可以包括其它网络。这样的联网环境在办公室、企业范围计算机网络、内联网和因特网中是常见的。

[0048] 当在 LAN 联网环境中使用时，计算机 110 通过网络接口或适配器 170 连接至 LAN171。当在 WAN 联网环境中使用时，计算机 110 通常包括调制解调器 172 或用于与诸如因特网等 WAN173 上建立通信的其它装置。调制解调器 172 可以是内置或外置的，它可以通过用户输入接口 160 或其它合适的机制连接至系统总线 121。在网络化环境中，相对于计算机 110 描述的程序模块或其部分可以存储在远程存储器存储设备中。作为示例，而非限制，图 1 示出了远程应用程序 185 驻留在存储器设备 181 上。可以理解，所示的网络连接是示例性的，且可以使用在计算机之间建立通信链路的其它手段。

[0049] 本领域的普通技术人员可以理解，计算机 110 或其他设备可被部署为计算机网络的一部分。就此而言，本发明涉及具有任何数目的存储器或存储单元、以及跨任何数目的存储单元或卷进行的任何数目的应用程序和进程的任何计算机系统。本发明可适用于网络环境中部署了服务器计算机和客户机计算机并具有远程或本地存储的环境。本发明也可适用于具有编程语言功能、解释和执行能力的单机计算设备。

[0050] 从数据库镜像进行在线页还原

[0051] 图 3 和 5 描述了本发明的示例性实施例。系统 300 可驻留在诸如以上关于图 1 所述的一个或多个计算机上。系统 300 可包括以下组件中的一个或多个 :主数据库 (在图 3 中，驻留在主服务器 302 上的数据库 308) 以一个或多个镜像数据库 (由驻留在一个或多个镜像服务器 320、330 等上的数据库 328、338 等表示)。因此，主服务器 302 可包括以下中的一个或多个 :诸如 Microsoft 的 SQL Server、IBM 的 DB2、Oracle 等的主数据库服务器的实例 304 以及主数据库 (在图 3 中由数据库 308 表示)。主数据库服务器 304 可包括执行如此处所述的从数据库进行在线页还原机制的功能的软件模块 306。类似地，一个或多个镜像数据库服务器 320、330 等可包括以下中的一个或多个 :诸如 Microsoft 的 SQL Server、IBM 的 DB2、Oracle 等的镜像数据库服务器的实例 324、334 等以及镜像数据库 (在图 3 中由数据库 328、338 等表示)。镜像数据库服务器 324、334 等可包括执行如此处所述的从数据库进行在线页还原机制的功能的软件模块 326、336。

[0052] 在本发明的某些实施例中，从数据库进行在线页还原模块 306、326、336 等包括完

全是基于软件的灾难恢复解决方案，将如下更详细描述。在本发明的某些实施例中，在线页还原模块检测主服务器上的一个或多个损坏页。在手动模式中，模块 306 可接收执行还原功能的指令。例如，示例性、非限定性的指令可以是“从镜像中对页 x 还原数据库”。或者，在自动模式中，当检测到损坏页后，还原软件 306 可在没有人工干预的情况下由计算机自动调用。主服务器 302 然后可锁定主服务器上的数据库中的一个或多个损坏页（在图 3 中由页 309 表示），并向至少一个镜像发送要求镜像上对应于该被损坏页的页（如果所选镜像为镜像 1320 则为页 329，如果所选镜像为镜像 2330 则为页 339，依此类推）的请求。在某些实施例中，页由页标识符表示。可发送主服务器上当前时刻的日志序列号 (LSN)，因为被损坏的页上的 LSN 不可信。镜像 (320、330 等）等待其重做操作以通过请求中提供的 LSN，来确保从日志中重新播放了对所请求的页的所有改变并将其应用于镜像上的该页（页 329、339 等），使得最新（当前）页被发送到主服务器 302。镜像从其缓冲池或其磁盘上取回页（页 329、339 等）。可保证现在由镜像持有的页与主服务器上的数据库一致，因为不能对该页进行更新，因为它被锁定，且日志被重新播放以通过该锁定时刻。使用数据库镜像映射通信基础架构，使用特殊消息类型将该页从镜像发送到主服务器，以将该页标识为要用于还原损坏页的页。在接收该页之后，主服务器将被还原的页写入磁盘以保存还原。锁定可被释放，使得该被修补的页再次可用于查询。

[0053] 如果一个或多个镜像不可用、镜像被挂起、或如果在还原操作期间触发了数据库镜像映射故障转移，则执行错误处理。

[0054] 在自动模式中，在崩溃恢复期间或常规操作期间检测到损坏页之后，可在无需人工干预的情况下由计算机自动启动该过程。当检测到损坏页时，对该页进行更新的事务的锁定被自动地保持（传统上，所引起的“延期操作”需要管理员干预来解决底层问题）。当调用从数据库镜像中进行在线页还原的机制的自动模式特征时，损坏页（由页 id 标识）被锁定，从镜像完成页还原，然后可调用用于卷回延期事务的代码，得到对数据库页损坏的无缝修复。

[0055] 所述技术可被扩展来从镜像中还原整个文件（例如，在磁盘崩溃的情况中）。如果调用手动模式（需要人工干预），则对原始位置不可用的情况下，可任选地指定文件位置。对于自动模式，可尝试默认位置，否则服务器可等待手动操作被执行。

[0056] 如果自动模式被关闭或未被实现，则可使用底层页还原机制来从镜像提供用于只读操作的页，直到损坏页被修补。在此选项中，上述方法仍被遵循，但从镜像接收到的页不被写回到主服务器上的磁盘。这允许更大的数据可用性，同时仍允许管理员维持对还原的手动控制。

[0057] 优化包括：

[0058] 1) 采用多个镜像，主服务器跟踪最快响应的镜像（较快响应时间可能由于多种因素，包括网络差异、物理位置等），并从最快响应的镜像请求还原页。

[0059] 2) 如果多个镜像处于追逐来自从主服务器接收的日志的重做操作的不同阶段，则主服务器可要求用于还原页的最新的（当前）镜像。

[0060] 3) 当还原多个页时，可通过向不同镜像要求页的块来对镜像进行负载平衡。

[0061] 图 5 是根据本发明的某些实施例，并如以上参考图 3 所述，示出用于从数据库进行在线页还原的示例性方法的流程图。在 502，检测到损坏。损坏可在崩溃恢复或正常操作期

间检测。损坏可被限于单个页或一组页，或可涉及整个文件或数据库。如果在线页还原在手动模式中操作，则需要人工干预(506)。诸如数据库管理员的某个人可执行指定要还原的指定页、要还原的一组页、整体要还原的文件（诸如文件系统文件或数据库文件）的命令。另外，可指定这一个或多个所属的数据库和从中接收相应的未被损坏的一个或多个页的一个或多个镜像。要被还原的一个或一组页可由页 id 或页 id 范围来标识。也可指定与检测到损坏的时间相关联的 LSN。可以理解，与被损坏的页相关联的 LSN 是不可靠的，因为 LSN 可能被损坏。

[0062] 如果在线页还原在自动模式中操作，则在 504 处检测到损坏后，由运行在计算机上的软件启动还原，且无需人工干预。在自动模式中，在前一段中描述的参数由运行在计算机上的在线页还原软件设置。在 508，不论是在手动还是自动模式中，锁定主数据库中被损坏的一个或多个页。在锁定的时刻，被损坏的一个或多个页变为不可用，但数据库的其余部分仍可访问（例如，可用于用户查询和更新等）。向至少一个镜像请求被损坏的一个或一组页。可基于历史上具有最快响应时间的镜像、基于最新（当前）镜像或基于任何其他合适的准则来选择接收该页请求的一个或多个镜像。如果要还原大量页或整个文件，则可通过向多个镜像发送对所需页的子集的（可任选为非重叠）请求来执行负载平衡。为易于理解，假定单个页被损坏，且选择了适当的镜像，例如镜像 1 来接收页请求。然而，可以理解，所构想的本发明并不如此限制。在 508，主服务器可向镜像 1 发送对页 idX 的请求，以及如上所述的 LSN。在 510，该镜像可接收该请求，且可等待直到其日志更新被应用于所接收的 LSN 以确保该该页是最新的（在损坏检测之前对该页进行的所有改变已被应用）。一旦，所述日志被应用到至少所接收的 LSN 所指定的时刻之后，对应于被损坏页的一个或多个镜像页就可从镜像上的缓冲池或从镜像盘中取回。可生成将消息标识为在线页还原消息的指定类型的消息，并可将其发送到主服务器。在 512，可在主服务器接收该页，并将其应用于数据库。该页可被写入磁盘以保存还原后的页。锁定可被释放 (514)，如由数据库的特性所确定地，使得还原后的页可用于查询和更新。

[0063] 在本发明的某些实施例中，在损坏页正被还原时，可从镜像来对针对一个或多个损坏页的查询进行服务，允许数据更大的可用性。

[0064] 此处所述的各种技术可结合硬件或软件或其适当组合来实现。因此，本发明的方法和装置或其某些方面或部分，可采取体现为在现实介质中的程序代码（即，指令）形式，这些现实介质诸如有软盘、CD-ROM、硬盘驱动器或任何其他机器可读存储介质，其中当程序代码被加载到诸如计算机的机器中并由其执行时，该机器成为用于实现本发明的装置。在程序代码在可编程计算机上执行的情况中，计算设备一般包括处理器、该处理器可读的存储介质（包括易失性和非易失性的存储器和 / 或存储元件）、至少一个输入设备以及至少一个输出设备。利用本发明的领域 - 专用编程模型特征而创建和 / 或实现的一个或多个程序（例如通过使用数据处理 API 等），优选地以高级过程或面向对象编程语言实现，以与计算机系统通信。然而，如果需要，也可以使用汇编或机器语言来实现程序。在任何情况中，该语言可以是编译或解释语言，且与硬件实现相结合。

[0065] 尽管结合各个附图的优选实施例描述了本发明，但可以理解，可使用其他类似的实施例，或者可对所述实施例进行修改和添加以便于执行本发明的相同功能而不与之背离。从而，本发明不应限于任何单个实施例，而应根据所附权利要求书的广度和范围来解释。

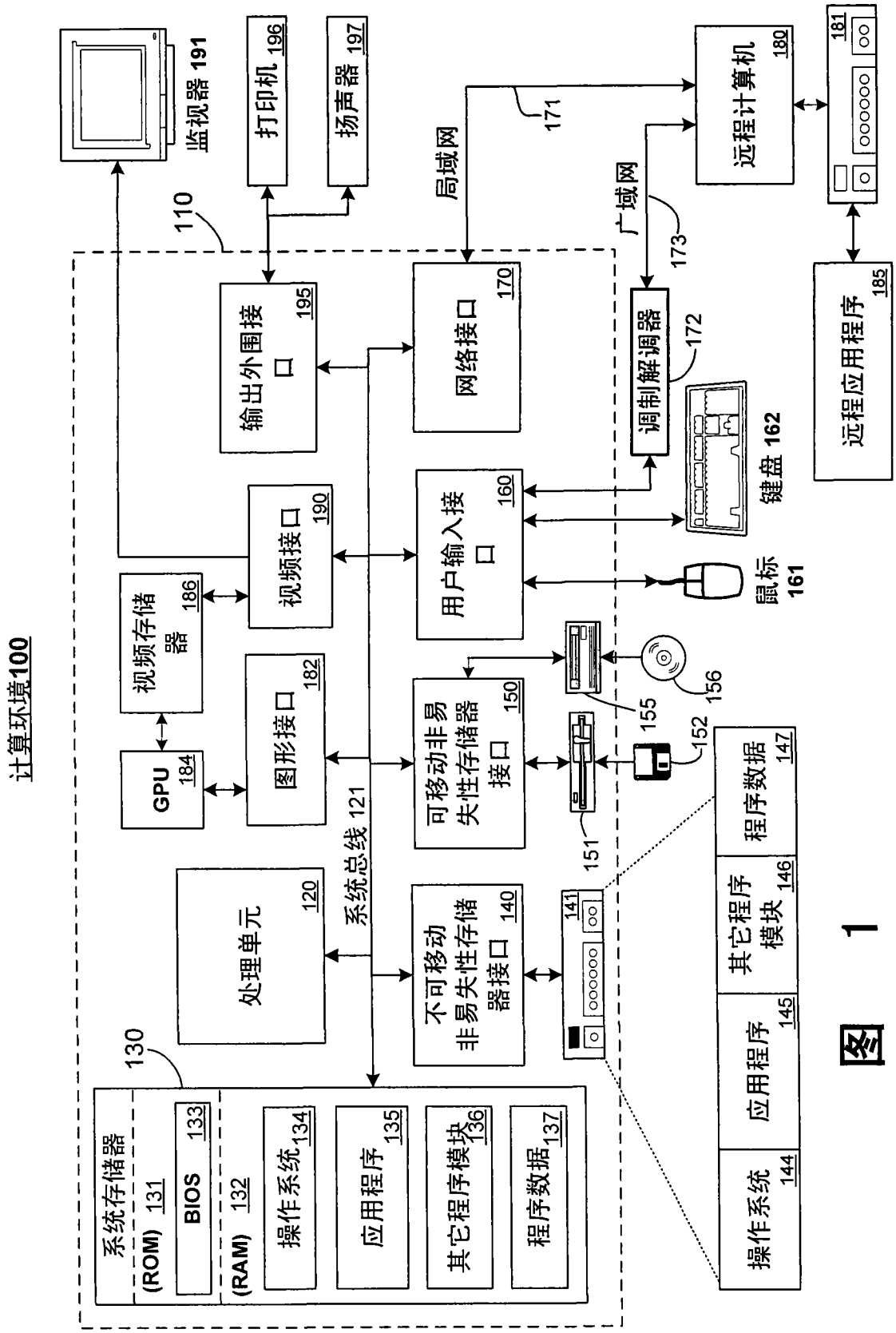

计算环境100
监视器 191
打印机 196
扬声器 197
局域网
171
远程计算机 180
181
110
广域网
173
输出外围接口 195
网络接口 170
调制解调器 172
远程应用程序 185
视频存储器 186
视频接口 190
用户输入接口 160
键盘 162
鼠标 161
GPU 184
图形接口 182
可移动非易失性存储器接口 150
155
156
152
151
系统总线 121
处理单元 120
不可移动非易失性存储器接口 140
141
130
系统存储器
(ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
其它程序模块 136
程序数据 137
操作系统 144
应用程序 145
其它程序模块 146
程序数据 147

图 1

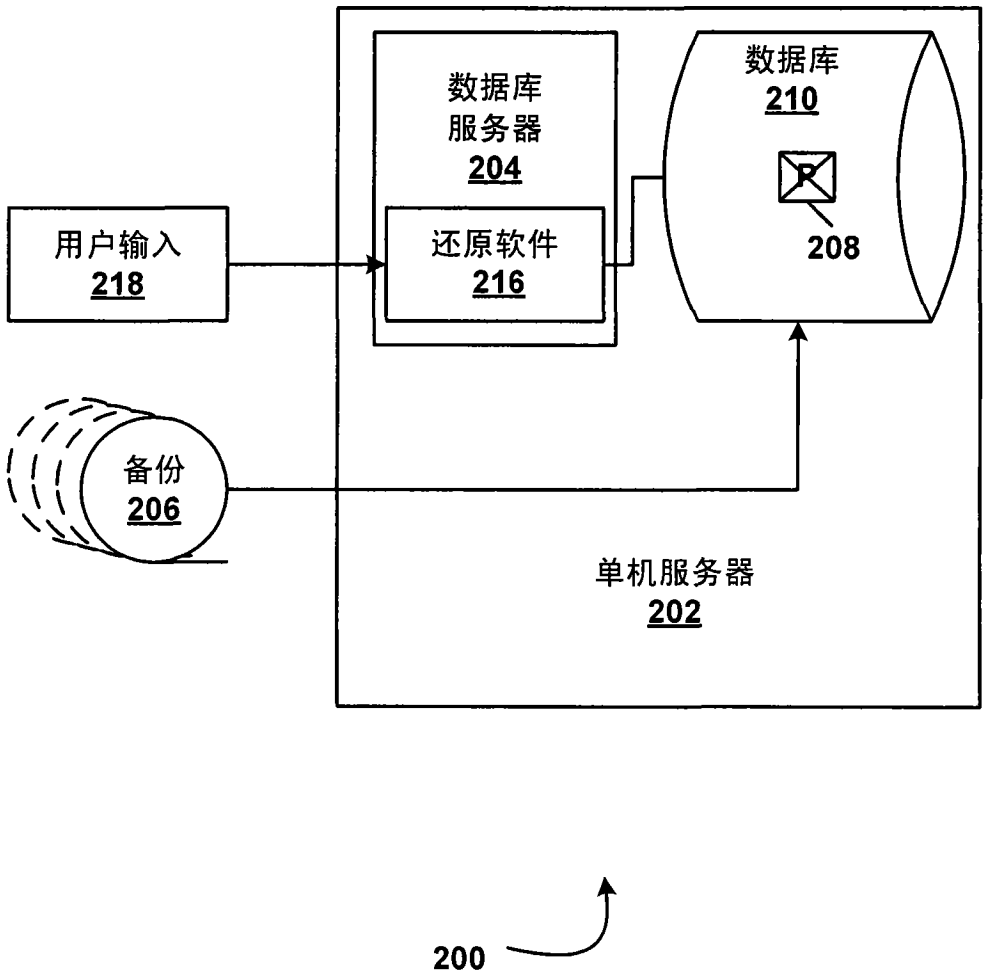


图 2
现有技术

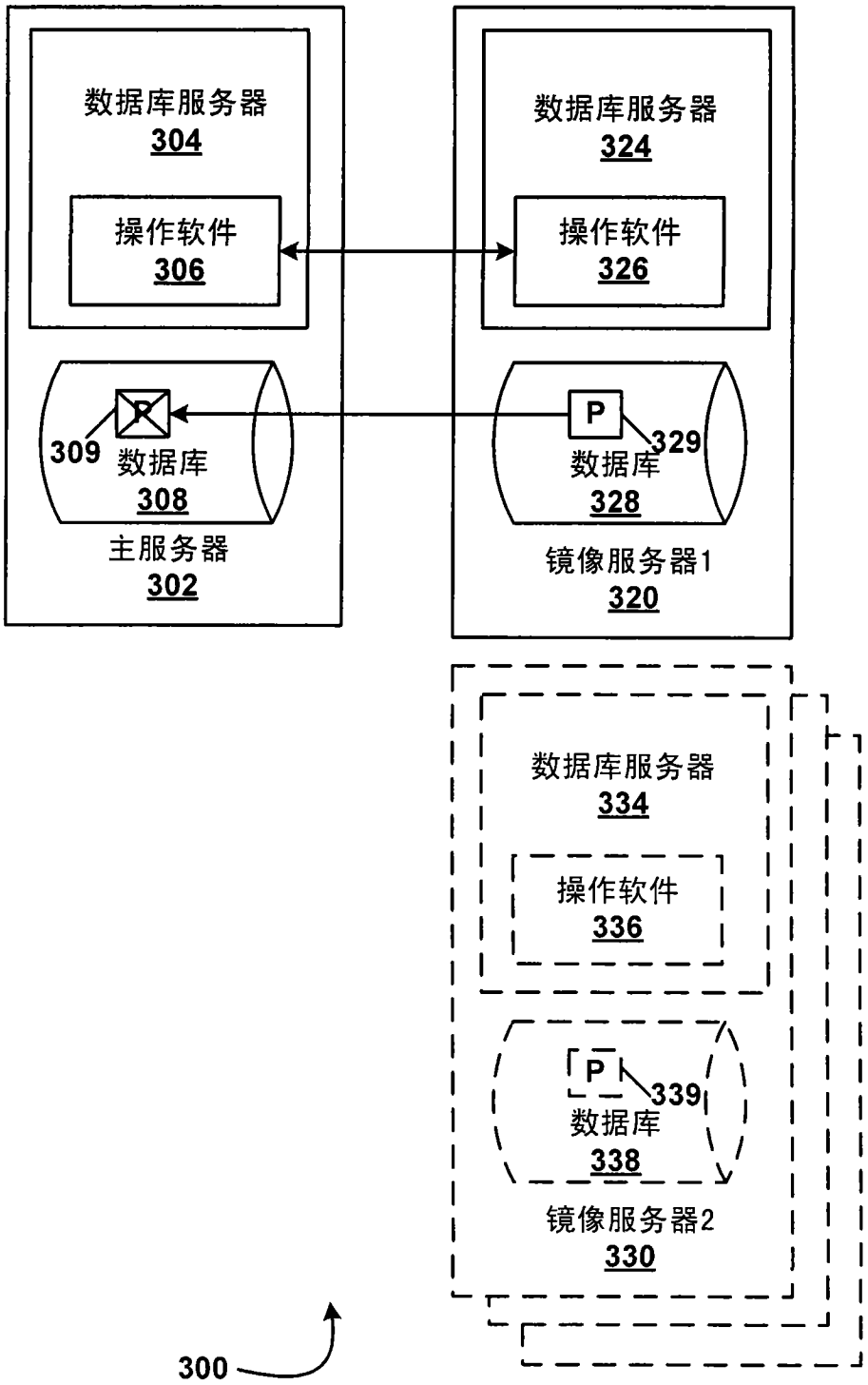

数据库服务器
304
操作软件
306
数据库服务器
324
操作软件
326
P
309
数据库
308
主服务器
302
P
329
数据库
328
镜像服务器1
320
数据库服务器
334
操作软件
336
P
339
数据库
338
镜像服务器2
330
300

图 3

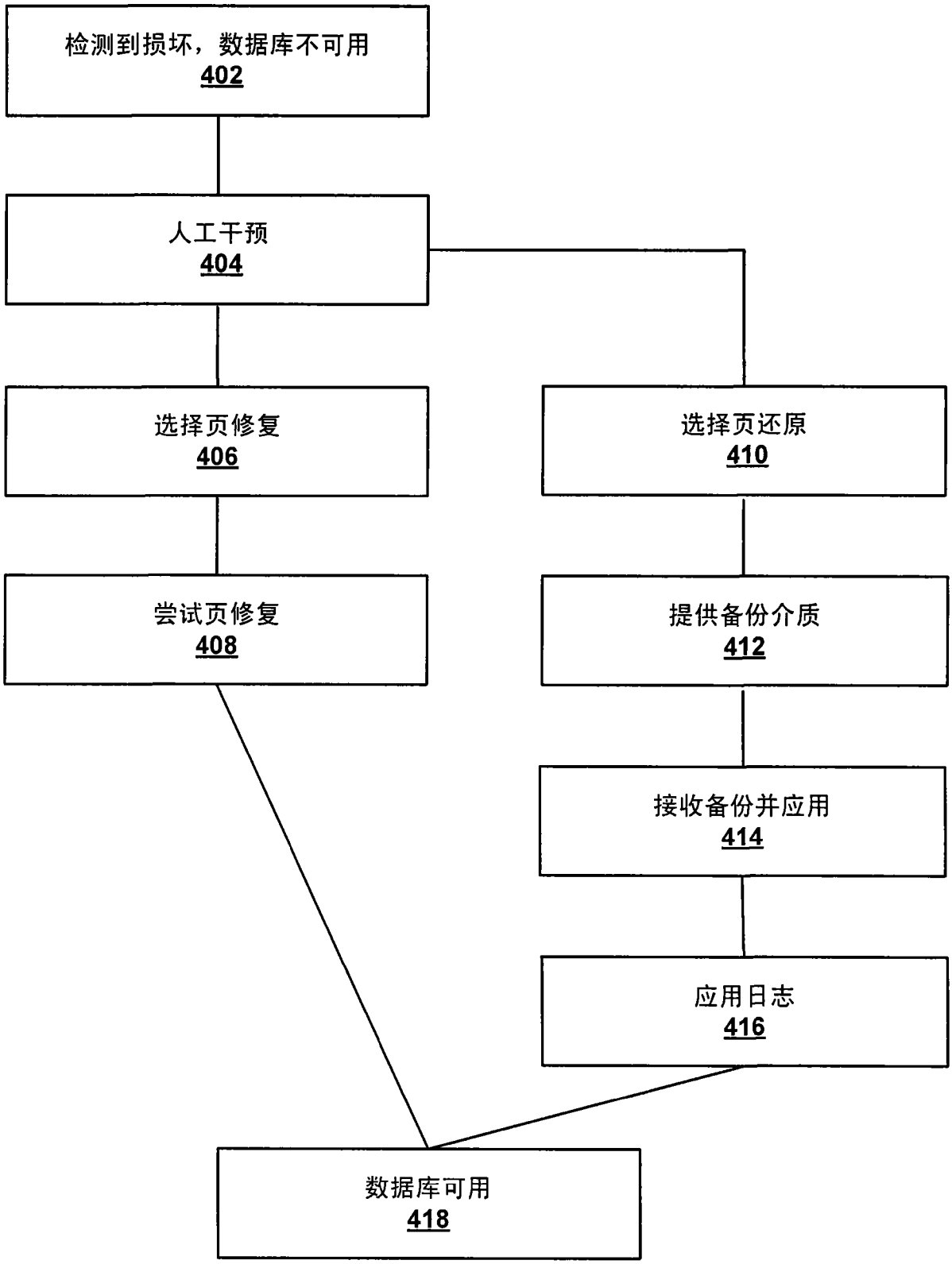

检测到损坏，数据库不可用
402
人工干预
404
选择页修复
406
尝试页修复
408
选择页还原
410
提供备份介质
412
接收备份并应用
414
应用日志
416
数据库可用
418

图 4

现有技术

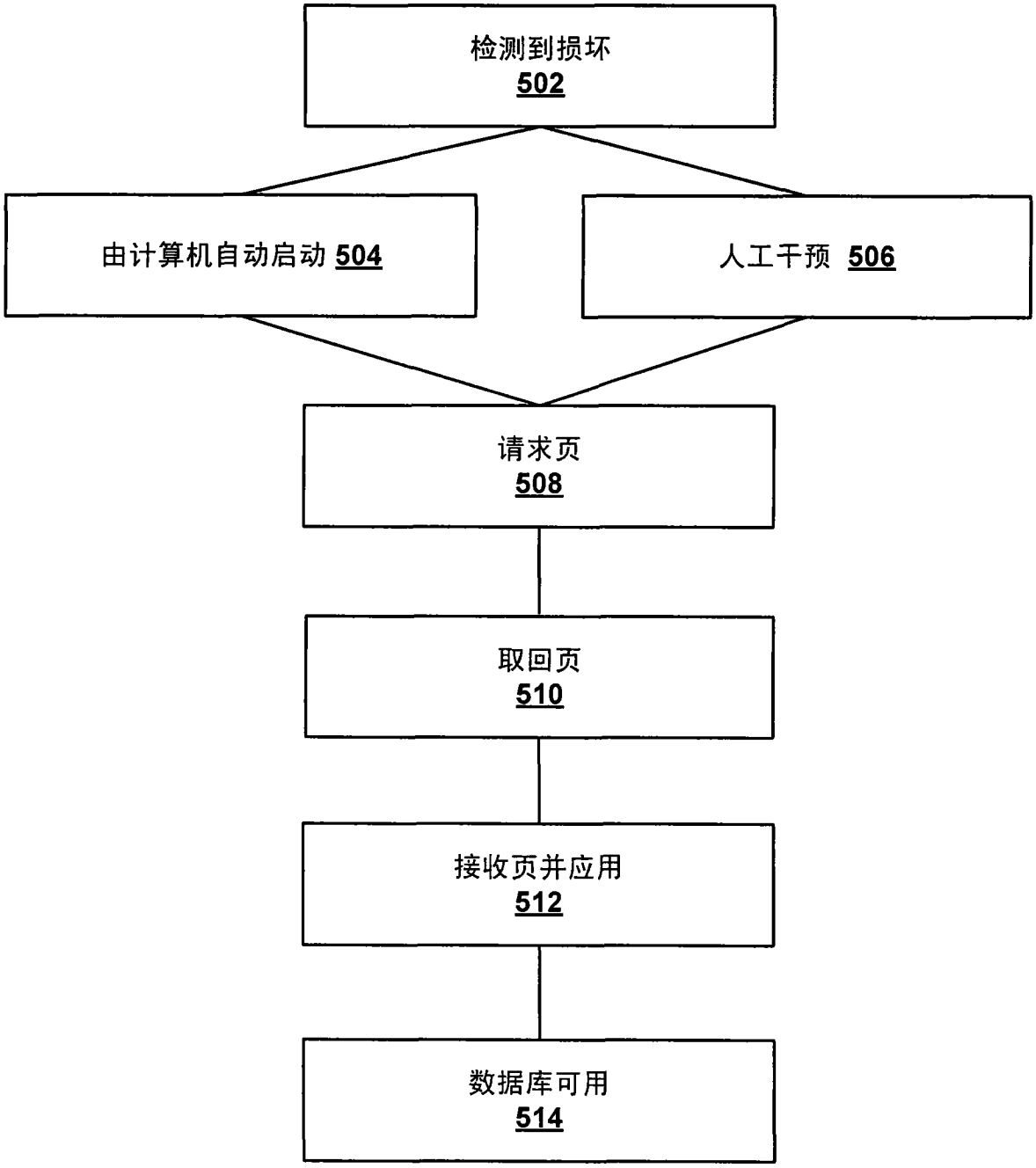

检测到损坏
502
由计算机自动启动 504
人工干预 506
请求页
508
取回页
510
接收页并应用
512
数据库可用
514

图 5