

[19] 中华人民共和国国家知识产权局



[12] 发明专利说明书

专利号 ZL 200480023056.3

[51] Int. Cl.

G06F 9/46 (2006.01)

G06F 9/50 (2006.01)

[45] 授权公告日 2009 年 10 月 14 日

[11] 授权公告号 CN 100549960C

[22] 申请日 2004.8.13

EP0750256A2 1996.12.27

[21] 申请号 200480023056.3

US5933604A 1999.8.3

[30] 优先权

US5721825A 1998.2.24

[32] 2003.8.14 [33] US [31] 60/495,368

审查员 于春晖

[32] 2003.9.3 [33] US [31] 60/500,096

[74] 专利代理机构 北京康信知识产权代理有限公司

[32] 2004.8.12 [33] US [31] 10/917,660

代理人 余刚 尚志峰

[86] 国际申请 PCT/US2004/026506 2004.8.13

[87] 国际公布 WO2005/017746 英 2005.2.24

[85] 进入国家阶段日期 2006.2.13

[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 卡罗尔·科尔雷恩

权利要求书 5 页 说明书 20 页 附图 3 页

[56] 参考文献

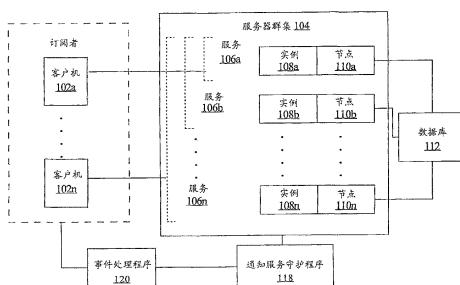
EP1260902A2 2002.11.27  
US2003037146A1 2003.2.20

[54] 发明名称

群集计算系统中改变的快速应用程序通知的方法和系统

[57] 摘要

借助于群集计算系统的改变的快速通知，应用程序可以快速地恢复并且会话可以被快速地再平衡，其中通过群集计算系统为系统状态改变公布多个事件。当与服务相关的资源经历诸如终止或开始/重新开始的状态改变时，立即公布通知事件。通知事件包含信息以使得订阅者能够基于匹配会话签名来识别受状态改变影响的特定会话，并且相应地响应。这允许当资源失败时，会话被快速地中断并且正在进行的处理被快速地终止，并且允许当资源重新开始时，快速地再平衡工作。



1. 一种用于传送关于群集计算环境的改变的方法，所述群集计算环境包括用于容放服务器实例的多个互连节点，所述方法包括以下计算机可执行的步骤：

接收与在群集中执行工作的特定服务相关的资源的状态改变的指示；

响应于所述资源的所述状态改变，立即生成识别所述资源的第一数据和指示所述资源的状态的第二数据；

将所述第一数据和第二数据公布给一组一个或多个订阅者；并且

其中，所述第一数据被订阅者用来基于响应于建立与所述群集的会话而生成并且用于识别与所述会话相关的服务的识别信息，来识别受所述资源的所述状态改变影响的与所述群集的一个或多个会话。

2. 根据权利要求 1 所述的方法，其中，在所述第二数据中被识别的所述资源的所述状态来自以下组中的一项，所述组由 (a) 所述资源的终止，(b) 所述资源的开始，和 (c) 不能够重新开始所述资源组成。
3. 根据权利要求 1 所述的方法，其中，所述群集是数据库群集，并且其中，所述资源在所述第一数据中通过识别受所述状态改变影响的所述数据库群集而被识别。
4. 根据权利要求 3 所述的方法，其中，所述工作与服务相关，并且其中，所述资源在所述第一数据中通过识别受所述状态改变影响的所述服务而被进一步识别。

5. 根据权利要求 4 所述的方法，其中，所述资源在所述第一数据中通过识别受所述状态改变影响的实例和节点而被进一步识别。
6. 根据权利要求 3 所述的方法，进一步包括以下计算机可执行的步骤：

基于使与所述会话相关的所述识别信息和识别所述资源的所述第一数据相匹配，识别与受所述状态改变影响的所述数据库群集的一个或多个会话。
7. 根据权利要求 6 所述的方法，进一步包括以下计算机可执行的步骤：在识别与由所述状态改变影响的所述数据库群集的一个或多个会话之后，中断所述一个或多个会话。
8. 根据权利要求 1 所述的方法，其中，所述资源在所述第一数据中通过识别受所述状态改变影响的节点而被识别。
9. 根据权利要求 1 所述的方法，其中，所述工作与服务相关，并且其中，所述资源在所述第一数据中通过识别服务已经在特定实例终止，并且通过识别所述服务已经终止的所述特定实例而被识别。
10. 根据权利要求 1 所述的方法，其中，所述工作与服务相关，并且其中，所述资源在所述第一数据中通过识别整个服务已经终止，并且通过识别已经终止的所述服务而被识别。
11. 根据权利要求 1 所述的方法，其中，所述资源在所述第一数据中通过识别特定实例已经终止，并且通过识别已经终止的所述特定实例而被识别。

- 
12. 根据权利要求 1 所述的方法，其中，所述资源在所述第一数据中通过识别所有的所述实例已经终止，并且通过识别所述实例与其相关的所述群集而被识别。
  13. 根据权利要求 1 所述的方法，其中，所述工作与服务相关，并且其中，所述资源在所述第一数据中通过识别服务已经在特定实例开始，并且通过识别所述服务已经在其上开始的所述特定实例而被识别。
  14. 根据权利要求 13 所述的方法，其中，所述资源在所述第一数据中通过识别支持已经开始的所述服务的实例的数量而被识别。
  15. 根据权利要求 1 所述的方法，其中，所述工作与服务相关，并且其中，所述资源在所述第一数据中通过识别服务已经在任意实例上开始，并且通过识别已经开始的所述服务而被识别。
  16. 根据权利要求 15 所述的方法，其中，所述资源在所述第一数据中通过识别支持已经开始的所述服务的实例的数量而被识别。
  17. 根据权利要求 1 所述的方法，其中，所述资源在所述第一数据中通过识别特定实例已经开始，并且通过识别已经开始的所述实例而被识别。
  18. 根据权利要求 1 所述的方法，其中，所述资源在所述第一数据中通过识别实例已经开始，并且通过识别所述实例与其相关的所述群集而被识别。

19. 根据权利要求 1 所述的方法，其中，所述资源在所述第一数据中通过识别节点已经终止，并且通过识别已经终止的所述节点而被识别。
20. 根据权利要求 1 所述的方法，其中，所述公布的步骤包括通过进程来公布所述第一数据和第二数据，其中所述进程并不是用于管理所述群集的群集件的一部分。
21. 根据权利要求 1 所述的方法，其中，所述第一数据和第二数据的订阅者是连接池管理器，所述连接池管理器通过基于所述第一数据和第二数据将连接重新分配给所述群集，来响应所述状态改变。
22. 根据权利要求 1 所述的方法，其中，所述第一数据和第二数据的订阅者是客户应用程序，所述客户应用程序通过基于所述第一数据和第二数据请求在受所述状态改变影响的所述工作的所述群集内的重新分配，来响应所述状态改变。
23. 根据权利要求 1 所述的方法，其中，所述第一数据和第二数据的订阅者是成批作业，所述成批作业通过基于所述状态改变调用所述群集内的例程的执行，来响应所述第一数据和第二数据。
24. 根据权利要求 1 所述的方法，其中，所述工作与服务相关，并且其中，所述资源在所述第一数据中通过识别所述服务没有重新开始使得订阅应用程序重试以使所述服务被中断而被识别。

25. 一种群集系统，包括：

数据库群集，包括由通信地耦合到数据库的一组互连的节点容放的一组服务器实例；

群集管理软件，用于管理所述群集中的资源以及所述群集中的工作的分配和执行，其中，所述资源与相应的特定服务相关；

通知系统，用于向一个或多个客户端公布关于所述资源的状态改变的信息，以便用于基于当客户端建立与所述群集的会话时被生成并且识别与所述会话相关的资源的识别信息，来识别与受相应状态改变影响的所述群集的一个或多个会话；并且

其中，关于资源的状态改变每个所述信息均包括与所述资源和所述资源的状态相关的一个或多个特定服务的识别。

26. 根据权利要求 25 所述的群集系统，其中，在关于状态改变的所述信息中被识别的所述资源的所述状态构成来自以下组中的一项，所述组由 (a) 所述资源的终止，(b) 所述资源的开始，以及 (c) 不能重新开始所述资源组成。

27. 根据权利要求 25 所述的群集系统，其中，在关于状态改变的所述信息中被识别的所述资源与来自以下组的至少一项相关，所述组由 (a) 服务，(b) 在所述实例的特定实例上执行的服务成员，(c) 所述数据库群集，(d) 所述实例中的一个，以及 (e) 所述节点中的一个组成。

# 群集计算系统中改变的快速应用程序通知 的方法和系统

## 技术领域

本发明一般地涉及群集计算系统，并且更具体地，涉及用于使用事件来实现群集系统中的状态改变的快速通知的技术。

## 背景技术

### 群集计算系统

群集计算系统是用于提供对一组客户应用程序的处理的互连计算元件的集合。每个计算元件被称为节点。节点可以是互连到其它计算机的计算机，或互连到网格中的其它刀片式服务器的刀片式服务器。能够共享地存取存储器（例如，能够共享地磁盘存取一组磁盘驱动器或非易失存储器）并且经由互连件连接的群集计算系统中的一组节点，在此称为工作群集。

群集计算系统被用来容放群集服务器。服务器是集成的软件构件和用于在处理器上执行该集成的软件构件的计算资源（诸如存储器、节点、以及节点上的进程）分配的组合，其中，软件和计算资源的组合用于为该服务器的客户机提供特定类型的功能。服务器的一个例子是数据库服务器。在数据库管理的其他功能之中，数据库服务器管理并且促进对特定数据库的存取，处理客户机对存取数据库的请求。

来自群集计算系统中的多个节点的资源可以被分配，以运行服务器的软件。该服务器的特定节点的资源的每个分配，在此被称为“服务器实例（instance）”或实例。数据库服务器可以是群集的，其中，服务器实例可以总称为群集。数据库服务器的每个实例促进对同一数据库的存取，其中，数据的完整性由全局锁管理器管理。

### 用于根据服务级（service level）来管理应用程序的服务

服务是数据库工作量管理的特征，其划分在数据库中执行的总体工作，以根据服务级来管理工作。资源根据服务级和优先级被分配给服务。服务被衡量和管理以有效地按需要交付资源容量。资源高可用性服务级使用群集的冗余部分的可靠性。

服务是用于管理工作量的逻辑抽象。服务可以用于将在数据库群集中执行的工作划分成相互脱节的类（class）。每个服务可以用共同属性、服务级阈值、和优先级来表示逻辑业务功能（例如，工作量）。服务的分组基于工作的属性，工作可以包括将被调用的应用功能、应用功能执行的优先级、将被管理的作业类、或者在作业类的应用功能中使用的数据范围。例如，电子 - 商务套件可以为每项职责定义服务，诸如总分类帐、应收帐款、定单登记等。服务提供单系统映象以管理竞争应用程序，并且服务允许每个工作量被分离的作为一个单位来管理。服务可以跨越网格中的多个群集或群集中的多个服务器实例，并且单个服务器实例可以支持多个服务。

中间层和客户机/服务器应用程序可以通过，例如，将服务指定为连接的部分来使用服务。例如，应用程序服务器数据源可以被设定为发送到服务。此外，服务器端工作将服务名称设定为工作量定义的一部分。例如，作业类使用的服务在作业类被创建时被定义，并且在执行期间，作业被分配给作业类并且作业类在服务内运行。

## 数据库会话

为了使客户机与数据库群集上的数据库服务器的相互作用，为客户机建立会话。会话（诸如数据库会话）是为客户机到服务器建立的特定连接，诸如数据库实例，客户机通过其发出一系列请求（例如，执行数据库语句的请求）。对于建立在数据库实例上的每个数据库会话，保持反映数据库会话的当前状态的会话状态数据。这样的信息包括，例如，为其建立会话的客户机的身份、以及客户机使用的服务、由执行数据库会话内的软件的进程生成的临时变量值。应用程序可以从连接池“借用”连接，并且在会话结束时将连接放回到连接池中。一般地，会话是用数据库来执行工作的媒介（vehicle）。每个会话都可以具有其自己的数据库进程或者可以共享数据库进程，后者称为多路复用。

## 高可用性

在群集计算系统中出现了某些改变，其降低了高可用性并且导致客户应用程序浪费时间。一般地，这样的改变可以被分类为“停止（down）”改变、“发生（up）”改变、或“非重新开始”改变。当服务、服务器实例、或节点机器（通常为“部件”）终止或“停止”时出现停止改变。当服务、服务器实例、或节点起始或“发生”时出现发生改变。当服务、实例、或节点不再开始时出现“非重新开始”改变。某些改变可以影响现有的会话以及当前没有使用但是已经被创建并与服务、实例、或节点相关的连接。

当群集系统的状态改变时，应用程序在其经由会话与群集系统的相互作用中浪费相当大量的时间和资源。特别地，当群集系统的状态改变时，群集数据库的客户应用程序浪费时间和资源。例如，当会话正在使用的节点或服务器实例“停止”时，应用程序可以不被长时间中断。特别地，如果节点或网络不能关闭会话套接字

(socket)，则应用程序等待来自局部 TCP/IP 堆栈的 TCP/IP 超时错误。对于另一个例子，当新服务、节点、或实例变得可用时，即“发生”时，工作可以不在支持服务的所有实例中被分配。换句话说，由于当服务、节点、或实例可用时不与其连接，时间被浪费。时间和资源被浪费的另一种方式是，当客户机保持再次尝试与不会变得备用或者还没有重新开始的部件通信时。

一般地，当节点停止时，传统系统的性能非常差。应用程序会话可以等待多达两个小时以被中断。一般地，当失败的系统实体被恢复（即，已经发生的实体）时，传统系统在分配工作给恢复的实体方面性能差。因此，传统系统提供了降低的可用性，并且潜在地减少了服务次数，而可用性和服务次数正是本系统所能提供的。而且，使用传统系统，在失败部件修复或恢复之后，传统的冷故障切换 (cold-failover) 系统通常将整个工作量的低效运行提供给恢复的实体，而不是提供群集中的部件的补充物间的负荷平衡。

在运行时期间，会话通常在涉及相应的数据库服务器实例的四个状态中的一个中。会话可以（1）主动地连接至实例，即，建立与实例的会话；（2）主动地发出命令给实例，诸如发出 SQL 语句；（3）被动地被阻塞，等待对发出的 SQL 语句的响应；以及（4）处理先前的请求，例如，SQL 语句。状态（1）不同于其它状态，因为客户机正在进入 TCP/IP 堆栈。在其它状态中，客户机在 TCP/IP 堆栈内部。

当群集系统的状态改变时，服务器的客户机在其经由会话与群集系统的相互作用中，浪费相当大量的时间和资源。特别地，当数据库群集的状态改变时，群集数据库的客户应用程序浪费时间和资源。例如，当新服务、节点、或实例变得可用时，即“发生”时，工作可以不在支持服务的所有实例间被分配。换句话说，由于当实例或节点上的服务变得可用时而不与其连接，时间被浪费。时间和

资源被浪费的另一种方式是，当客户机保持再次尝试与不会变得备用（即，与死节点通信）或还没有重新开始的部件通信时。例如，当会话正在使用的节点或服务器实例“停止”时，应用程序可以不被长时间（即，通常为两小时）中断。特别地，如果节点或网络不能关闭会话套接字，则应用程序等待来自局部 TCP/IP 堆栈的 TCP/IP 超时错误。

情况（1）可以通过使用通常可用的虚拟 IP 地址来缓和。这是因为客户机在 TCP/IP 堆栈的外部。因此，当节点停止时，IP 地址故障切换到不同节点。然而，当节点重新发生时，不存在寻址系统状态的改变的类似的解决方案。一般地，大多数问题当会话在状态（2）、（3）、或（4）中时出现。当会话在状态（3）中时，当应用程序和/或会话必须等待问题解决时，出现大多数浪费的时间。甚至更糟糕的，大约 90% 的时间，应用程序在状态（3）中。此外，使用连接池客户机，由于通过提供到应用程序的死连接，时间被浪费。

一般地，当失败的或其它终止的系统实体被恢复时，传统系统在分配工作给恢复的实体方面性能差。因此，传统系统提供降低的可用性，并且潜在地减少的服务次数，而可用性和服务次数正是本系统所能提供的。而且，使用传统系统，在失败部件的修复或恢复之后，传统的冷故障切换系统通常将整个工作量的低效运行提供给恢复的实体，而不是提供群集中的部件的补充物间的负荷平衡。此外，当为仅冷故障切换、待机等配置一个节点时，系统中不存在冗余。这样的系统被称为“主动/被动”系统，其中，所有资源对于所有连接的应用程序均可用。

基于前述情况，对群集计算系统中的系统状态改变的反应有改进的空间。

## 发明内容

为了解决上述问题至少之一，本发明的一个实施例提供了一种用于传送关于群集计算环境的改变的方法，所述群集计算环境包括用于容放服务器实例的多个互连节点，所述方法包括以下计算机可执行的步骤：接收与在群集中执行工作的特定服务相关的资源的状态改变的指示；响应于所述资源的所述状态改变，立即生成识别所述特定服务的第一数据和指示所述资源的状态的第二数据；将所述第一数据和第二数据公布给一组一个或多个订阅者；并且其中，所述第一数据被订阅者用来基于响应于建立与所述群集的会话而生成并且用于识别与所述会话相关的服务的识别信息，来识别受所述资源的所述状态改变影响的与所述群集的一个或多个会话。

本发明的另一实施例提供了一种系统，包括：数据库群集，包括由通信地耦合到数据库的一组互连的节点容放的一组服务器实例；群集管理软件，用于管理所述群集中的资源以及所述群集中的工作的分配和执行，其中，所述资源与相应的特定服务相关；通知系统，用于公布关于所述资源的状态改变的信息，以便用于基于当客户端建立与所述群集的会话时被生成并且识别与所述会话相关的资源的识别信息，来识别与受相应状态改变影响的所述群集的一个或多个会话；并且其中，关于资源的状态改变每个所述信息均包括与所述资源和所述资源的状态相关的一个或多个特定服务的识别。

采用根据本发明的技术方案，对群集计算系统中的系统状态改变的反应进行了有效改进。

## 附图说明

通过附图中的实例来描述本发明，但是不局限于此，在附图中相同的参考标号表示类似的元件，其中：

图 1 是示出可以实施实施例的操作环境的框图；

图 2 是概括地示出可以实施本发明的实施例的高可用性 (HA) 系统的框图；以及

图 3 是可以实施本发明的实施例的计算机系统的框图。

## 具体实施方式

### 实施例的功能性综述

描述了用于群集计算系统的改变的快速通知的技术，其中，为系统状态改变公布多个事件，以便允许与群集系统的会话的快速再平衡和快速应用程序恢复。这样的群集计算系统的一个例子是数据库群集，其包括在多节点机器上执行的数据库服务器的多个实例，被设置为响应于来自多客户应用程序的请求而存取和操作来自数据库的共享数据。

当与服务相关的资源经历状态改变时，通知事件立即被公布以便该事件的不同订阅者使用。例如，只要服务在一个实例上变得可用并且只要服务在一个实例上变得不可用，就发出通知事件。当服务和支持该服务的资源（诸如特定实例、实例、节点、或数据库群集）的状态改变时，发生通知事件。当由一个或多个实例提供的服务开始时，通知事件 (UP) 被发出，其可以被用于启动取决于该服务的应用程序。当由一个或多个实例提供的服务终止时，以及当实例或节点终止时，通知事件 (DOWN) 被发出以暂停从属应用程序。

当因为服务已经超过其失败阈值，管理群集件（clusterware）不再管理该服务时，通知事件（NOT\_RESTARTING）被发出以中断应用程序重试该服务。在一个实施例中，NOT\_RESTARTING 事件启动到灾难服务的切换。

在连接到群集时，唯一签名（unique signature）（即，定位器）被生成用于相关的会话并且被记录在句柄（handle）上作为连接的一部分。在实施例中，该签名包括服务标识符、节点标识符、数据库唯一名称、以及实例标识符，其每个都与会话相关。在数据库群集环境中，通知事件包含信息以使得订阅者能够识别特定会话，该特定会话受状态（即，受影响的会话的签名）改变的影响。对于某些类型的事件而言，被用于识别受影响的会话的信息包括与状态改变相关的数据库和服务的识别。对于其它类型的事件而言，用于识别受影响的会话的信息还包括与状态改变相关的节点和实例的识别。受影响的会话是与包括在事件有效负载中的签名相匹配的签名的会话。

因此，当相关的资源改变状态时，应用程序和会话被快速地通知。通过使用这些技术，可以使用通知事件来克服如前面所述的传统系统的问题。

### 操作环境

图 1 是示出可以实施实施例的操作环境的框图。使用从 a 到 n 的元件标识符，例如，客户机 **102a-102n**、服务 **106a - 106n**、实例 **108a-108n**、和节点 **110a-110n**，并不意味着要求这样的元件的相同标号。换句话说，n 对于各个元件而言并不必然相同。相反，这样的标识符在一般意义上使用，以便标识多个类似的元件。

## 群集计算环境

一个和多个客户机 **102a-102n** 可通信地耦合至服务器群集 **104** (“服务器”), 服务器群集连接到共享数据库 **112**。服务器 **104** 泛指服务器实例 **108a-108n** 和实例可在其上执行的节点 **110a-110n** 的群集。其它元件也可当作服务器 **104** 的一部分, 诸如通知服务守护程序 (daemon) **118** 和事件处理程序 **120**。然而, 其中已经配置了前述元件的实际结构可以随着执行的不同而改变。客户机 **102a-102n** 可以是由例如经由网络互连到应用程序服务器或互连到在客户机和服务器 **104** 之间的某些其它中间件元件的计算机执行的应用程序。此外, 一个服务器实例可以是另一个服务器实例的客户机。正如在此所描述的, 任何或所有客户机 **102a-102n** 可以作为公布的事 件的订阅者来操作。

在数据库群集环境中, 数据库 **112** 包括存储在持久存储机制上的数据和元数据, 持久存储机制例如可通信地耦合到节点 **110a-110n** 的一组硬盘, 其每个均能够容放一个或多个实例 **108a-108n**, 每个实例都容放一个或多个服务的至少一部分。这样的数据和元数据可以逻辑地存储在数据库 **112** 中, 例如, 根据相关的数据库构造、多维数据库构造、或相关的数据库构造和多维数据库构造的组合。节点 **110a-110n** 可以被实施作为传统计算机系统, 诸如图 3 中所示的计算机系统 **300**。

如上所述, 数据库服务器是集成的软件构件和用于在处理器上执行该集成的软件构件的计算资源 (诸如存储器和进程) 分配的组合, 其中, 软件和计算资源的组合用于管理诸如数据库 **112** 的特定数据库。在数据库管理的其它功能中, 数据库服务器通常通过处理来自客户机的对存取数据库 **112** 的请求来促进对数据库 **112** 的存取。

---

实例 **108a-108n** 与相应的节点 **110a-110n** 结合，用于容放服务 **106a-106n**。

## 服务 **106**

如上所述，服务通常是用于管理工作量的逻辑抽象。特别地对于本发明的实施例的环境，诸如服务 **106a-106n** 的服务具有名称和域名，并且可以具有相关的目标、服务级、优先级、和高可用性属性。被执行作为服务的一部分的工作包括计算机资源的任何使用或开销，包括，例如，CPU 处理时间、在易失性存储器中存储和存取数据、从持久存储器（即，磁盘空间）读取和向持久存储器写入、以及使用网络或总线带宽。

在一个实施例中，服务是由数据库服务器在会话期间执行的工作，并且通常包括被执行以处理和/或计算请求存取特定数据库的查询的工作。这里所使用的术语查询指的是遵循诸如 SQL 的数据库语言的语句，并且包括规定添加、删除、或修改数据以及创建和修改数据库对象（例如表、对象视图（view）、以及可执行的例程）的操作的语句。包括群集计算系统的系统可以支持多种服务。

服务可以由一个或多个数据库服务器实例提供。因此，多个服务器实例可以共同工作以向客户机提供服务。在图 1 中，用虚线括号将服务 **106a**（例如，FIN）描述为由实例 **108a** 提供，将服务 **106b**（例如，PAY）描述为由实例 **108a** 和 **108b** 提供，并且将服务 **106n** 描述为由实例 **108a-108n** 提供。

通常，在此所描述的技术是以服务为中心的，其中，服务器 **104** 内发生的事件可以基于受事件影响的服务被识别和/或表征。通知事件的有效负载在下文中描述。

## 通知系统

一般地，守护程序是在后台运行并且在预定的时间或响应于特定事件而执行规定的操作的进程。一般地，事件是其公布被进程检测的动作或发生的事。通知服务守护程序 **118** 是从服务器 **104**（诸如从被设置为管理实例 **106a-106n** 的群集的群集件）接收系统状态改变信息的进程。这样的状态改变信息可以包括，例如，服务、实例、和节点发生或停止事件信息。正如在此所描述的，当条件在群集中改变时，服务器 **104** 公布事件。

通知服务守护程序 **118** 与事件处理程序 **120** 具有发布者 - 订阅者关系，通过该发布者 - 订阅者关系，由守护程序 **118** 从服务器 **104** 接收的系统状态改变信息被作为通知事件传输到事件处理程序 **120**。一般地，事件处理程序是包含响应于事件被执行的程序语句的函数或方法。响应于接收到来自守护程序 **118** 的事件信息，事件处理程序 **120** 至少传递在此所描述的事件类型和属性。单个事件处理程序 **120** 在图 1 中被描述为服务于所有订阅者。然而，不同的事件处理程序可能与不同的订阅者相关。由不同的订阅者执行的通知事件的处理的方式对于这样的事件而言并不重要，并且其可以随着执行的不同而改变。

从服务器 **104** 到通知服务守护程序 **118** 的事件信息的传输，以及从通知服务守护程序 **118** 到事件处理程序 **120** 的事件信息的传输是“带外的 (out of band)”。在本说明书中，带外的意味着这样的传输不通过会话通信路径，因为该路径可能由于该事件被阻塞。例如，带外通知可以通过网关程序被异步地发出，该网关程序不是管理该群集的群集件的一部分。对于非限制性的例子，通知服务守护程序 **118** 可以使用 Oracle 通知系统 (Oracle Notification System)

(ONS)API, 其是允许应用程序元件基于 Java2 平台、企业版(J2EE)来创建、发送、接收、和读取消息的通信机制。

“订阅者”表示可以为了多种相应的目的而订阅并响应通知事件的多种实体。订阅者的非限制性的例子包括客户机 **102a-102n**、连接池管理器、中间层应用程序、成批作业、呼出(callout)、呼叫和警报机制、高可用性日志等。

### 通知事件

当给定的服务的状态改变时, 即, 当参予执行给定服务的工作的群集资源的状态改变时, 新状态通过通知事件被通知给感兴趣的订阅者。例如, 应用程序可以使用该通知, 以实现失败的快速检测以便停止处理在先结果, 从而整理连接池会话, 并用于在失败之后以及失败的元件被修复时平衡连接池。例如, 当服务在实例开始时, 事件可以被用于立即触发工作执行该实例。当服务在实例终止时, 事件可以被用于中断使用在该实例的服务的应用程序。

多个客户机订阅者对通知事件的使用排除了, 例如, 在失败之后并且在被中断之前, 客户机等待 TCP 超时或在客户机的对最后结果的无用的处理。在没有这样的通知的情况下, 如果节点在没有关闭套接字的情况下失败, 则阻塞 I/O 等待(读取或写入)的任何会话将在数分钟到数小时内等待超时, 并且处理最后结果的会话将不接收中断, 直到下一个数据被请求。

### 通知事件有效负载

在实施例中, 通知事件包括事件类型和事件属性的识别。根据本发明的实施例, 表 1 描述了与通知事件相关的参数。

参数	描述
事件类型	用于群集元件的事件类型：服务、服务_成员、数据库、实例、节点。
服务名称	服务名称。
数据库名称	支持服务的唯一数据库。
实例	支持服务的实例的名称。
节点名称	支持服务或已经停止的节点的节点名称。
状态	新状态：UP、DOWN、NOT_RESTARTING。
具像(incarnation)	日期和时间标记；可以被用于命令通知事件。
基数	支持服务的实例的数目。
原因	系统改变的原因：计划的或未计划的。

表 1

当整个服务停止时，即当服务在支持该服务的每个实例上停止时，“服务”事件类型被触发。当服务在特定实例上停止时，“服务\_成员”事件类型被触发。当整个数据库停止时，即管理该数据库的每个实例停止时，“数据库”事件类型被触发。当节点机器停止时，“节点”事件类型被触发，因此，在该节点上运行的实例不能够用来支持任何服务。“节点名称”是管理群集件已知的节点的名称。DOWN事件通常在UP事件之前，例如当服务成员在支持实例失败时故障切换至另一个实例时，以及当失败的元件被修复时。

每个事件类型的事件有效负载包含事件特性中的每个是必要的。根据本发明的实施例，表2表示了包括在每个事件类型的事件有效负载中的特性。

事件类型	服务名称	数据库名称	实例名称	节点名称	状态	时间标记	基数
服务	x	x			x	x	
服务成员	x	x	x	x	x	x	x
数据库	x	x			x	x	x
实例	x	x	x	x	x	x	
节点				x	x	x	

表 2

---

根据实施例，事件可以被发布到在此所描述的通知系统（其为事件提供程序化的接口）、服务器端呼出（callout）、以及调用接口回叫（callback）。

下面是系统状态改变的情况的一些例子。

当实例失败时，公布几个通知事件：（1）事件类型 = 实例的通知事件，用于通知实例停止；（2）在终止的实例上运行的每个服务的事件类型 = 服务\_成员的通知事件，用于通知服务在该特定实例上停止。此外，如果服务在另一个支持实例（即，可以用作备用实例以支持服务的实例）上重新开始，事件以事件类型 = 服务\_成员被公布，以通知该服务在特定实例上新近可用。如果终止的实例被修复并且重新开始，则可以为没有可用的支持实例（即，没有备用实例来支持该服务）的每个服务公布通知事件：（1）事件类型 = 实例的通知事件，用于通知先前终止的实例发生；以及（2）事件类型 = 服务\_成员的通知事件，用于通知该服务在先前终止的实例上发生。

因为用于前述情况的每个通知事件均是“实例”或“服务\_成员”类型，因此表 2 示出所有特性均包括在事件有效负载中。如果数据库配置有数据库域名（诸如 us.acme.com），那么将希望域名资格表示为事件特性中的数据库名称和服务名称。例如，数据库 = databaseX.us.acme.com 并且服务 = serviceY.us.acme.com。此外，通过响应于 UP 事件来使用基数，工作可以被再分配从而以平衡的方式使用可用资源。

当支持服务的所有实例停止时，即整个服务停止时，公布几个通知事件：（1）用事件类型 = 服务\_成员触发事件用于支持该服务的每个实例，从而通知服务在每个相应的实例上停止；（2）用事件类型 = 服务触发事件，从而通知整个服务停止；以及（3）用事件

---

类型 = 实例触发事件用于支持该服务的每个实例，从而通知每个相应的实例停止。同样地，当终止的实例重新开始时，类似的事件用 UP 状态触发，从而通知服务在每个相应的实例上发生、整个服务发生、以及实例发生。

对于“服务”类型事件，表 2 示出实例名称、节点名称、和基数特性可以从该类型的事件的有效负载中去除。这是因为，通过定义，“服务”类型事件意味着整个服务发生或停止，即，服务在所有实例和支持该服务的相关节点上发生或停止，其中，用于给定的服务的支持实例和节点被映射并且在其它地方可用。因此，实例、节点、和基数特性对于响应这样的事件是不必要的。类似地，对于“数据库”类型事件，表 2 示出实例名称和节点名称特性可以从该类型的事件的有效负载中去除。这是因为，通过定义，数据库类型事件意味着整个数据库发生或停止，即，该数据库中的所有实例和节点发生或停止，其中，数据库群集的配置在其它地方可用。

当节点失败时，事件类型 = 节点的事件被触发，通知该节点停止。在一个实施例中，没有其它事件被触发，即，由于节点到实例和服务的可用的映射，没有服务、服务\_成员、或实例类型的事件是必要的。节点事件包括群集具像，因而有助于消除复制(duplicate)事件的复制处理。表 2 示出服务名称、数据库名称、实例名称、和基数特性可以从该类型的事件的有效负载中去除。如果由于任何原因特定节点不能够被重新开始，那么事件类型 = 节点并且状态为 NOT\_RESTARTING 的事件被触发，指示要求干涉。

## 事件处理

一般地，通过识别受系统状态改变的影响并且与给定的订阅者相关的一个或多个会话，通过将事件有效负载中的信息与当建立会话连接时被记录的会话位置信息相匹配，给定的订阅者响应通知事件。会话位置信息（即，会话签名）识别会话位置，诸如什么服务、数据库、实例、节点和数据库与会话相关。类似的“事件”签名被提供作为事件有效负载的一部分，其可以与在会话建立时被记录的一个或多个会话的签名精确地匹配，以确定受系统状态改变影响的会话。

会话位置信息被记录的方式可以随着执行的不同而改变。对于非限制性的例子而言，会话位置信息可以被公布在不同的订阅者可以访问的“公告牌”机构上，或者会话位置信息可以被存储在索引表或散列表中。不同的订阅者可以响应于通知事件而执行的详细的动作超出了本说明书的范围。

例如，连接池管理器被描述作为通知事件一个潜在订阅者或客户机，其中，通知事件响应于群集中资源的状态改变被发布。连接池管理器是软件构件，该软件构件管理连接池并要求会话与服务器**104** 的连接（图 1）。在标题为“Fast Reorganization of Connections In Response To An Event In a Clustered Computing System”的美国专利申请第 10/XXX,XXX（案号 No.50277-2335）中，详细描述了连接池管理器可以响应不同类型的通知事件的方式，诸如响应在服务器**104** 的实例**108a-108n** 中重新分配来自连接池的连接。

例如，连接池管理器为每个服务保持到物理位置（即实例、节点、数据库）的连接的映射。因此，包括客户机-服务器会话和批处理会话的每个会话的位置被唯一地识别。无论何时连接建立，连

接池管理器记录该连接的位置。无论何时接收到系统状态改变通知事件，这些数据都被用于快速重新分配连接池。通过确保连接池具有当会话请求被接收时即可以用的连接，响应于事件的连接的快速再分配有助于群集中的工作的运行时分配。

对于其它非限制性的例子，(1) 响应于服务器端的呼出，工作和批处理程序可以在服务因为任何原因停止时被停止，并且在服务开始时立即开始/重新开始；(2)呼出可以被用于呼叫并发送电子邮件给报警机构；以及(3)呼出可以被用于高可用性正常运行时间记录，其中，由于系统改变的原因，诸如“计划的”（例如，由用户发起）或“未计划的”（例如，由失败引起），可以区分正常运行时间（uptime）和故障时间（downtime）。

### 执行机制

正如在此所描述的，用于群集计算系统中的快速应用程序通知的技术，可以以多种方式来执行，并且本发明并不局限于任何特定的执行。本方法可以被集成到系统或装置中，或者可以被执行作为独立机制。此外，本方法可以在计算机软件、硬件、或软件和硬件的组合中执行。

图 2 是概括地示出可以实施本发明的实施例的高可用性 (HA) 系统的框图。所述的 HA 系统 200 的“层”包括：群集服务节点从属关系模块；内部事件系统；HA 架构，其包括服务、数据库、和实例；其通信地耦合到外部事件系统。

HA 系统 200 执行用于资源（即，服务、数据库、和实例）的开始和停止动作。开始动作公布 UP 事件，并且停止动作公布 DOWN 事件。当资源不再执行时，HA 系统 200 公布 NOT\_RESTARTING 事件。

## 硬件综述

图 3 是示出可以执行本发明的实施例的计算机系统 300 的框图。计算机系统 300 包括用于传递信息的总线 302 或其它通信装置以及用于处理信息的与总线 302 连接的处理器 304。计算机系统 300 还包括诸如随机访问存储器 (RAM) 或者其它动态存储装置的主存储器 306，其连接至总线 302 用于储存信息和将由处理器 304 执行的指令。在执行将由处理器 304 执行的指令期间，主存储器 306 还可用于储存临时变量或其他中间信息。计算机系统 300 进一步包括只读存储器 (ROM) 308 或连接至总线 302 的其他静态存储装置，用于存储静态信息和处理器 304 的指令。提供诸如磁盘或光盘的存储设备 310，并连接至总线 302 用于存储信息和指令。

计算机系统 300 可以经由总线 302 连接至诸如阴极射线管 (CRT) 的显示器 312，用于向计算机用户显示信息。包括字母数字键和其他键的输入装置 314 连接至总线 302，用于将信息和指令选择传递到处理器 304。另一种类型的用户输入装置是光标控制 316，诸如鼠标、跟踪球、或光标方向键，用于将方向信息和命令选择传递到处理器 304 并用于控制显示器 312 上的光标移动。输入装置通常在两个轴上（第一个轴（例如 X 轴）和第二个轴（例如 Y 轴））具有两个自由度，使装置能指定平面上的位置。

本发明涉及计算机系统 300 的使用，用于执行在此描述的技术。根据本发明的一个实施例，通过计算机系统 300 响应于执行包括在主存储器 306 中的一个或多个指令的一个或多个序列的处理器 304，来实现这些技术。这样的指令可以从诸如存储装置 310 的其它计算机可读介质读入主存储器 306。包括在主存储器 306 中的指令序列的执行，使得处理器 304 执行此处所述的处理步骤。在可选实施例中，可以使用硬连线电路（hard-wired circuitry）来取代软件指令或

者与软件指令结合来实施该发明。因此，本发明的实施例将不限于硬件电路和软件的任何特定组合。

这里使用的术语“计算机可读介质”是指参与向处理器 304 提供指令用于执行的任何介质。这种介质可以采取多种形式，包括但不限于非易失性介质、易失性介质、和传递介质。非易失性介质举例来说包括光盘或磁盘，诸如存储装置 310。易失性介质包括动态存储器，诸如主存储器 306。传输介质包括同轴电缆、铜线、和光纤，包括组成总线 302 的导线。传输介质还可采取声波或光波形式，例如那些在无线电波和红外线数据通信过程中产生的声波和光波。

通常形式的计算机可读介质包括如软盘、软性盘、硬盘、磁带，或者任何其它磁性介质、CD-ROM、任何其它光介质、打孔纸、纸带、或者任何带孔图样的物理介质、RAM、PROM、EPROM、FLASH-EPROM、或者其他任何存储芯片或者盒式磁带，或者以下提到的载波、或者计算机可读的任何其他介质。

各种形式的计算机可读介质可参与将一个或者多个指令的一个或多个序列承载到处理器 304 用于执行。例如，指令开始可承载在远程计算机的磁盘中。远程计算机可以将指令加载到其动态存储器中，然后使用调制解调器通过电话线发送指令。计算机系统 300 本地的调制解调器可接收电话线上的数据，并使用红外发射器将数据转换成红外信号。红外探测器可以接收红外信号携带的数据，并且合适的电路可以将数据放到总线 302 上。总线 302 将数据承载到主存储器 306，处理器 304 从主存储器取回并执行这些指令。在由处理器 304 执行这些指令之前或之后，由主存储器 306 接收的指令可随意地储存在存储装置 310 上。

计算机系统 300 还包括连接至总线 302 的通信接口 318。提供双向数据通信的通信接口 318，连接到与局域网 322 连接的网络链

路 320。例如，通信接口 318 可以是综合业务数字网（ISDN）卡或者调制解调器，用于提供到相应类型的电话线的数据通信连接。又如，通信接口 318 可以是局域网（LAN）卡，用于提供至兼容局域网（LAN）的数据通信连接。也可以使用无线链路。在任何这样的实施中，通信接口 318 发送和接收承载表示各种类型的信息的数字数据流的电信号、电磁信号、和光学信号。

网络链路 320 通常可通过一个或者多个网络向其它数据装置提供数据通信。例如，网络链路 320 可通过局域网 322 与主机 324 连接，或者与互联网服务提供商（ISP）326 操作的数据设备连接。ISP 326 又通过目前通称为“互联网” 328 的全球分组数据通信网络提供数据通信服务。局域网 322 和互联网 328 都使用承载数字数据流的电信号、电磁信号、或光学信号。通过各种网络的信号和网络链路 320 上的信号以及通过通信接口 318 的信号，都传送数字数据给计算机系统 300 或者传送来自计算机系统的数字数据，是传输信息的载波的示例性形式。

计算机系统 300 能通过网络、网络链路 320、和通信接口 318 发送消息和接收数据（包括程序代码）。在互联网的实例中，服务器 330 可通过互联网 328、ISP 326、局域网 322、和通信接口 318，传送用于应用程序的所请求的程序代码。

所接收的代码可以在其被接收时由处理器 304 执行，和/或储存 在存储装置 310 或者其它非易失性介质中用于随后执行。按照这种方式，计算机系统 300 可以以载波的形式获得应用代码。

以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

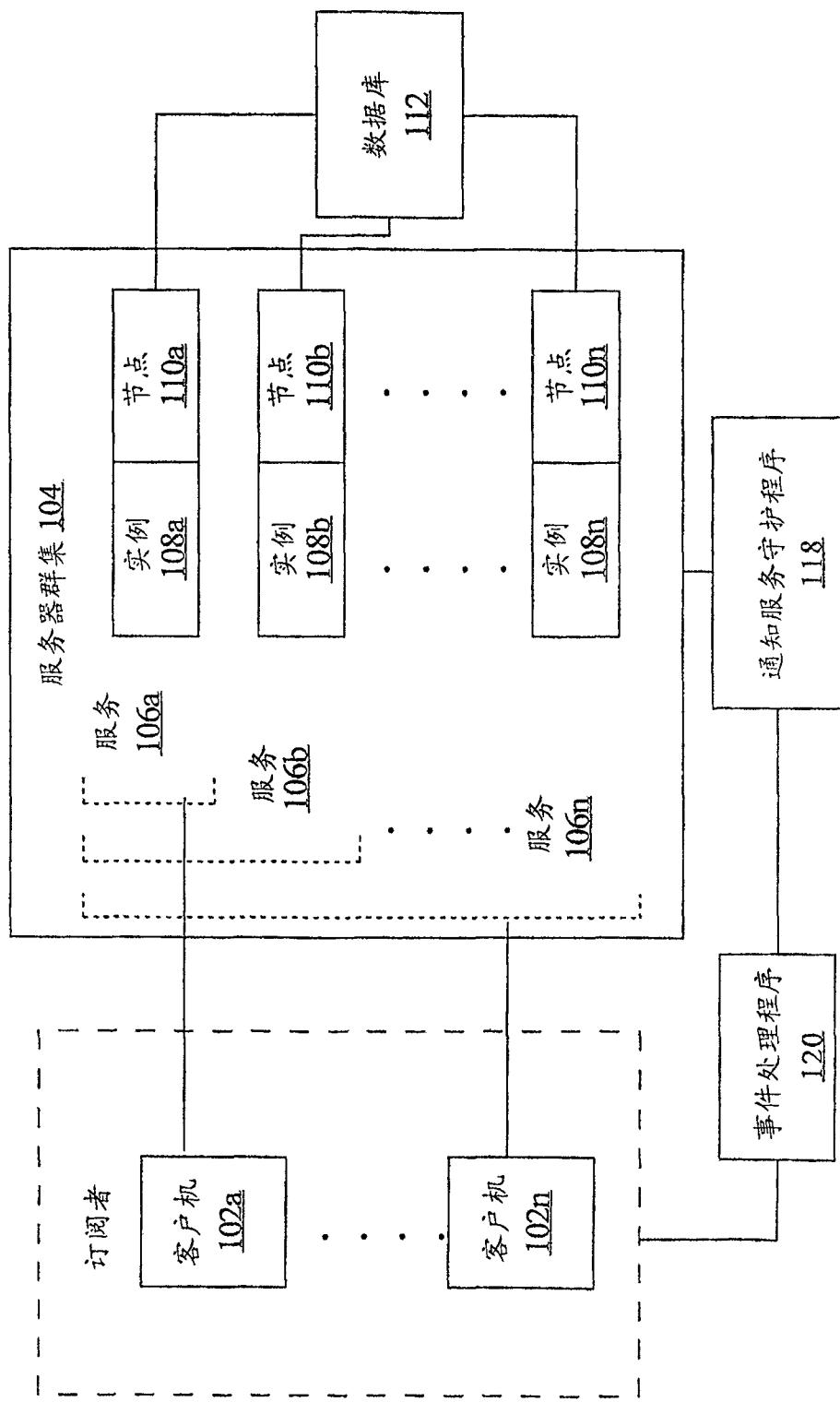


图 1

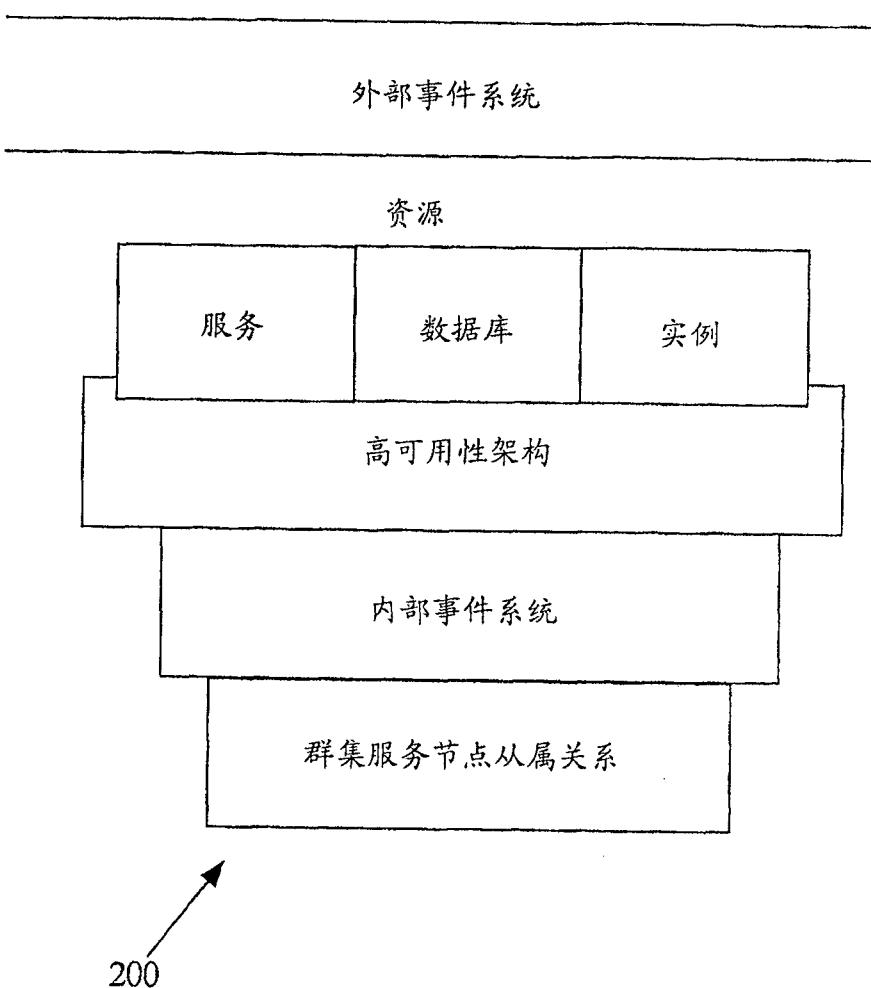


图 2

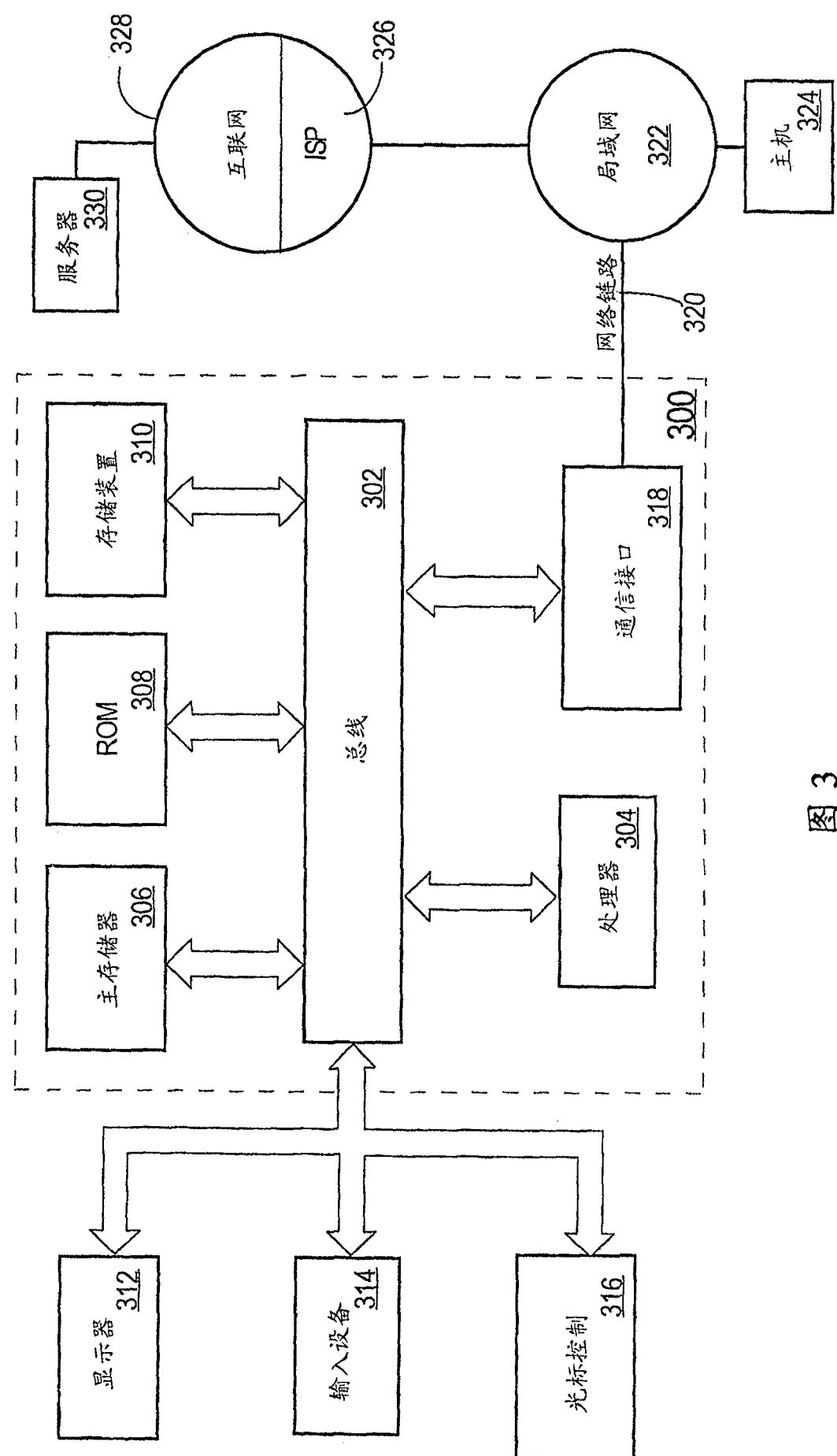


图 3