(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 1645389 B
(45)授权公告日 2010.07.21

(21)申请号 200510005571.5

(22)申请日 2005.01.19

(30)优先权数据
10/761,164 2004.01.20 US

(73)专利权人 国际商业机器公司
地址 美国纽约

(72)发明人 弗兰西斯科·德·拉·克鲁兹
米歇尔·A·波里尼
道格拉斯·斯科特·鲁瑟特
拉德哈克里什南·塞图拉曼

(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 11038
代理人 李颖

(51)Int.Cl.
H04L 29/06(2006.01)
H04L 12/24(2006.01)

(56)对比文件
说明书第[0008]至[0050]段、附图1至5.
EP 1320217 A1,2003.06.18,说明书第[0008]至[0050]段、附图1至5.

审查员 刘长勇

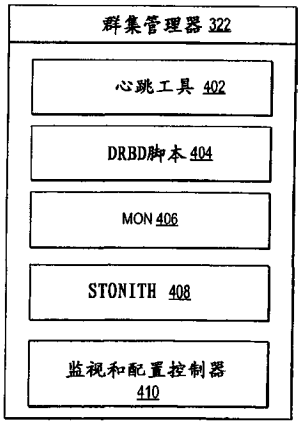
权利要求书 3 页 说明书 12 页 附图 9 页

(54)发明名称

用于实现高可用性系统的远程企业管理的系统和方法

(57)摘要

提供高可用性系统的远程企业管理的方法、系统和程序。多个高可用性系统在企业中连网,并由远程企业服务器整体管理。在每个高可用性系统内,群集管理控制器监视高可用性系统的特定组件的状态,当所述状态指示错误时,作出反应,以调整该高可用性系统。另外,就每个高可用性系统来说,监视控制器检测何时群集管理控制器对特定组件的状态作出反应,并检测该高可用性系统的多个组件的状况。监视控制器随后向远程企业服务器报告所述错误和所述多个组件的状况。使远程企业服务器能够根据从多个高可用性系统中的每一个接收的报告,管理每个高可用性系统。



CN 1645389 B

1. 一种能够实现高可用性系统的远程企业管理的系统，包括：

通过网络与远程企业服务器通信连接的多个高可用性系统的一个特定高可用性系统；

所述特定高可用性系统进一步包括：

主节点，

副节点，

群集管理控制器，和

数据复制器，

其中所述群集管理控制器，用于监视所述主节点的状态，并响应于在所述状态指示错误，作出反应，以调整所述特定高可用性系统，包括启动由所述副节点接管虚拟 IP 地址，关闭所述主节点的电源信号，指令由所述副节点访问数据，

所述群集管理控制器进一步包括：监视控制器，所述监视控制器用于检测何时所述群集管理控制器对所述错误作出反应，并检测所述特定高可用性系统的多个组件的状况，其中所述监视控制器向所述远程企业服务器报告所述错误和所述多个组件的状况，所述远程企业服务器能够根据所述报告管理所述特定高可用性系统，

其中所述数据复制器，用于复制主节点和副节点可访问的数据，而不需要在主节点和副节点之间实际物理共享的存储装置。

2. 按照权利要求 1 所述的实现高可用性系统的远程企业管理的系统，所述特定高可用性系统还包括：

实现由所述群集管理控制器监视的符合 J2EE 的中间件堆栈的多个服务器。

3. 按照权利要求 1 所述的实现高可用性系统的远程企业管理的系统，所述群集管理控制器还包括：

检测所述特定高可用性系统的主节点的状态的心跳监视器。

4. 按照权利要求 1 所述的实现高可用性系统的远程企业管理的系统，所述群集管理控制器还包括：

检测由所述特定高可用性系统的中间件层提供的服务的状态的服务监视后台驻留装置。

5. 按照权利要求 1 所述的实现高可用性系统的远程企业管理的系统，其中所述监视控制器从所述远程企业服务器接收配置请求，并根据所述群集管理控制器将如何作出反应来调整所述特定高可用性系统的配置。

6. 按照权利要求 1 所述的实现高可用性系统的远程企业管理的系统，其中所述监视控制器从所述远程企业服务器接收配置请求，并根据所述请求，调整所述特定高可用性系统的硬件配置。

7. 一种能够实现高可用性系统的远程企业管理的方法，包括：

监视通过网络与远程企业服务器通信连接的多个高可用性系统的一个特定高可用性系统的特定组件的状态，其中所述特定高可用性系统包括主节点与副节点，群集管理控制器和数据复制器，其中所述群集管理控制器用于监视所述主节点的状态并响应于在所述状态指示错误作出反应以调整所述特定高可用性系统，所述数据复制器用于复制主节点和副节点可访问的数据而不需要在主节点和副节点之间实际物理共享的存储装置；

响应于所述状态指示特定组件的错误，作出反应，以调整所述高可用性系统，包括启动由所述副节点接管虚拟 IP 地址，关闭所述主节点的电源信号，指令由所述副节点访问数据；

检测何时所述群集管理控制器对所述特定组件的所述错误作出反应，并检测所述特定高可用性系统的多个组件的状况；和

向所述远程企业服务器报告所述错误和所述多个组件的所述状况，所述远程企业服务器能够根据所述报告管理所述特定高可用性系统。

8. 按照权利要求 7 所述的实现高可用性系统的远程企业管理的方法，还包括：

监视实现符合 J2EE 的中间件堆栈的多个服务器的状态。

9. 按照权利要求 7 所述的实现高可用性系统的远程企业管理的方法，还包括：

由心跳监视器监视所述特定高可用性系统的主节点的状态。

10. 按照权利要求 7 所述的实现高可用性系统的远程企业管理的方法，还包括：

由服务监视后台驻留装置检测由所述特定高可用性系统的中间件层提供的服务的状态。

11. 按照权利要求 7 所述的实现高可用性系统的远程企业管理的方法，还包括：

从所述远程企业服务器接收配置请求；和

根据所述群集管理控制器将如何作出反应来调整所述特定高可用性系统的配置。

12. 按照权利要求 7 所述的实现高可用性系统的远程企业管理的方法，还包括：

从所述远程企业服务器接收配置请求；和

根据所述请求，调整所述特定高可用性系统的硬件配置。

13. 一种远程配置多个高可用性系统的系统，包括：

通过网络与远程企业服务器通信连接的多个高可用性系统的一个特定高可用性系统以及

与所述网络通信连接的远程企业服务器；

其中所述特定高可用性系统进一步包括：

主节点，

副节点，

群集管理控制器，以及

数据复制器，

其中所述群集管理控制器，用于监视所述主节点的状态，并响应于在所述状态指示错误，作出反应，以调整所述特定高可用性系统，包括启动由所述副节点接管虚拟 IP 地址，关闭所述主节点的电源信号，指令由所述副节点访问数据，

所述群集管理控制器进一步包括：监视控制器，用于检测何时所述群集管理控制器对所述错误作出反应，并检测所述特定高可用性系统的多个组件的状况，其中所述监视控制器向所述远程企业服务器报告所述错误和所述多个组件的状况，

其中所述数据复制器，用于复制主节点和副节点可访问的数据，而不需要在主节点和副节点之间实际物理共享的存储装置，和

其中所述远程企业服务器接收关于所述多个高可用性系统中每一个的监视控制器所检测的信息，分析所述被检测的信息，把重构请求发送给提交被检测的信息的所述多个高

可用性系统。

用于实现高可用性系统的远程企业管理的系统和方法

技术领域

[0001] 本发明涉及改进的高可用性群集管理(high availability clustermanagement),具体地说,涉及高可用性系统的远程群集管理。更具体地说,本发明涉及企业网络中的多个高可用性系统的改进远程监控和管理。

背景技术

[0002] 对于工作量和需求不断波动,并且处理客户请求极其重要的零售商、银行及其它在线服务,已开发了高可用性(HA)系统来处理紧要使命操作(mission-critical operation)。通常,HA系统是用于消除由网络系统的组件中的计划的或者非计划的停机引起的服务失败或者使之降至最少的系统。提供HA系统的关键方法是通过分成一群服务器的冗余硬件和软件组件。

[0003] 在HA系统中,冗余至关重要,因为当在群集的一个节点中发生故障时,系统的一个节点执行的处理转移给另一节点。例如,在两节点HA群集中,一个节点通常被指定为主节点,另一节点一般被指定为备用节点。通常,当群集被启动时,主节点一开始运行某一应用程序(application)。另外,备用节点一般被指定成当主节点发生故障时,运行所述应用程序。HA群集系统一般实现群集管理器过程,所述群集管理器过程定期轮询主节点(或者检查主节点的心跳(heartbeat)),以确定主节点是否仍然活动。如果未检测到“心跳”,那么群集管理器把软件进程转移到群集中的另一服务器。

[0004] HA系统的一个重要特征是恢复时间。通常,HA系统中的恢复时间是备用节点从发生故障的主节点接管应用程序所用的时间。在基于销售的HA系统中,恢复时间特别重要,因为如果客户不能很快完成交易,那么零售商会失去有价值的买卖。甚至30秒的恢复时间延迟也会减少零售商的买卖交易。

[0005] HA系统的另一重要特征是在故障转移(failover)期间,几乎不损失数据。特别地,重要的是实现提交数据的几乎不损失。例如,在故障转移期间,失去有价值的客户定单信息或者客户信息是不利的。

[0006] 为了实现短的恢复时间和故障转移期间数据的几乎不丢失,重要的是一开始按照建立HA系统的方式组合硬件和软件。但是,在起动HA系统之后,重要的是监视和调整HA系统的配置,设法提高故障转移和纠正其它错误的效率。

[0007] 当为HA系统配置硬件和软件时,许多开发者已开发了定制HA软件服务,以便控制经常需要新硬件的定制环境中的应用程序。这些解决方案通常费用高昂,并且没有利用允许跨越多个平台的应用程序的可移植性的开放源码技术(open source technology)。此外,通常选择费用高昂的服务器系统,希望服务器系统中的可用能力将自动提高故障转移的效率。

[0008] 作为一种备选方案,开放源码开发者利用当实现HA系统时,可配置的功能扩展开放源码技术。例如,Linux提供一种低廉的与平台无关的操作系统。Linux的开发者继续向该操作系统增加可由其它开发者按照开放源码方式实现的功能。这些功能中的一些,例如

“心跳”和分布式复制块装置 (drbd) 由 Linux 操作系统实现，以帮助配置 HA 系统。

[0009] 虽然 Linux 工具提供监视故障，并配置 HA 系统中使用的硬件的构架，仍然需要另外的监视和配置能力。特别地，需要一种监视 HA 系统的硬件和软件方面的故障、错误和其它非理想状况，并且监视何时开放源码 HA 工具检测到故障和错误的方法。此外，需要远程积累所监视的系统状态，随后远程促进 HA 系统的重构。

[0010] 此外，通常在网络中组合多个 HA 系统，形成一个企业系统。每个 HA 系统可服务于对企业内的不同商店的交易请求。需要一种远程积累企业内的多个 HA 系统的所监视的系统状态，比较系统状态与性能要求，跟踪企业内的每个 HA 系统的硬件和软件需要。

[0011] 此外，当利用开放源码操作系统构架实现 HA 系统时，实现符合开放源码的中间件层来处理交易请求应是有利的。特别地，实现 (1) 由与远程企业控制台进行接口的基于开放源码的群集管理控制、并 (2) 能够监视和配置企业网络中的多个 HA 系统的符合 JavaTM2 平台企业版 (J2EE) 的中间件堆栈理应是有利的。

发明内容

[0012] 本发明提供改进的高可用性群集管理，特别提供符合开放源码构架 (framework) 实现的高可用性系统的远程群集管理。更具体地说，本发明涉及企业网络中的多个高可用性系统的改进远程监视和管理。

[0013] 根据本发明的一个方面，多个高可用性系统在企业中连网，并由远程企业服务器整体管理。在每个高可用性系统内，群集管理控制器监视高可用性系统的特定组件的状态，当所述状态指示错误时，作出反应，以调整该高可用性系统。另外，就每个高可用性系统来说，监视控制器检测何时群集管理控制器对特定组件的状态作出反应，并检测该高可用性系统的多个组件的状况。监视控制器随后向远程企业服务器报告所述错误和所述组件的状况。使远程企业服务器能够根据所述报告管理高可用性系统。

[0014] 特别地，高可用性服务器实现由诸如心跳监视器和服务监视后台驻留程序 (daemon) 之类的开放源码功能监视的符合 J2EE 的中间件堆栈 (J2EE compliant middleware stack)。具体地说，心跳监视器检测中间件堆栈驻留其上的特定服务器的状态。服务监视后台驻留程序检测中间件堆栈提供的服务的特殊实例的状态。

[0015] 远程企业服务器可根据报告，确定应进行配置改变，并向高可用性系统发送配置请求。监视控制器随后调整高可用性系统的配置，以调整心跳监视器或服务监视后台驻留程序将如何检测错误并对错误作出反应。此外，高可用性系统内的其它硬件和软件组件可由监视控制器重构。

[0016] 远程企业服务器最好把关于每个高可用性系统的监视信息保存在数据库中。另外，企业服务器最好分析监视信息，确定哪些高可用性系统不满足性能要求。企业服务器可推荐硬件和软件改变，以及配置改变。另外，企业服务器可显示比较性能，并提供高可用性系统以及在每个系统何时检测到错误的实时显示。

附图说明

[0017] 公开了本发明特有的新特征。但是，结合附图，参考例证实施例的下述详细说明，将更好地理解发明本身，以及发明的优选使用模式，其它目的和优点，其中：

[0018] 图 1 是描述其中可实现本发明的方法、系统和程序的服务器系统的方框图；

[0019] 图 2 是描述在故障转移期间，有效转移中间件的高可用性群集的硬件配置的方框图；

[0020] 图 3 是描述根据本发明的方法、系统和程序的群集管理器的方框图；

[0021] 图 4 是描述根据本发明的方法、系统和程序，在故障转移之前，HA 群集的软件结构的一个实施例的方框图；

[0022] 图 5 是描述根据本发明的方法、系统和程序，在故障转移之后，HA 群集的软件结构的一个实施例的方框图；

[0023] 图 6 是描述在 HA 系统中的符合 J2EE 的中间件内，独立软件开发商应用程序的实现的一个实施例的方框图；

[0024] 图 7 是描述把 drbd 分区配置到 HA 群集中的符合 J2EE 的中间件堆栈的进程和程序的高级逻辑流程图；

[0025] 图 8 是描述通过心跳控制器，控制 HA 群集中符合 J2EE 的中间件堆栈的配置和故障转移的进程和程序的高级逻辑流程图；

[0026] 图 9 是描述控制用于监视符合 J2EE 的中间件堆栈提供的服务的 mon 功能的进程和程序的高级逻辑流程图；

[0027] 图 10 是描述根据本发明的方法、系统和程序，包括运行 J2EE 中间件堆栈的多个 HA 系统的企业网络的方框图；

[0028] 图 11 是描述根据本发明的方法、系统和程序，控制 HA 群集管理器内的监视控制器的进程和程序的高级逻辑流程图；

[0029] 图 12 是描述远程控制 HA 系统的群集管理器，从而重构 HA 系统的进程和程序的高级逻辑流程图；和

[0030] 图 13 是描述控制用于管理群集中的多个 HA 系统的远程企业控制台的进程和程序的高级逻辑流程图。

具体实施方式

[0031] 现在参考附图，尤其参考图 1，图 1 中表示了通过其可实现本发明的方法、系统和程序的系统的一个实施例。可在各种系统中执行本发明，包括各种计算系统、服务器系统和企业系统。

[0032] 计算机系统 100 包括总线 122 或者用于在计算机 100 内传递信息的其它通信装置，和与总线 122 耦接，以便处理信息的多个处理器 112a-112n。总线 122 最好包括由桥接器和适配器连接，并在计算机系统 100 内由多个总线控制器控制的低等待时间和高等待时间通路。

[0033] 处理器 112a-112n 可以是在正常操作期间，在可从诸如随机存取存储器 (RAM) 114 之类动态存储装置，和诸如只读存储器 (ROM) 116 之类静态存储装置访问的操作系统和应用软件的控制下，处理数据的通用处理器，例如 IBM 的 PowerPC™ 处理器。在一个优选实施例中，多层软件包含当在处理器 112a-112n 上运行时，执行在图 7、8、9、11、12、13 的流程图中描述的操作，以及这里描述的其它操作的机器可执行指令。另一方面，本发明的步骤可由包含用于实现所述步骤的硬连线逻辑电路的专用硬件组件，或者由编程的计算机组件和定

制硬件组件的任意组合实现。

[0034] 可以包括在机器可读介质上的计算机程序产品的形式提供本发明，所述机器可读介质保存有用于对计算机系统 100 编程，以便执行根据本发明的进程的机器可读指令。这里使用的术语“机器可读介质”包括参与向处理器 112a-112n 或者计算机系统 100 的其它组件提供指令，以便执行的任意介质。这样的介质可采取多种形式，包括（但不限于）非易失性介质，易失性介质和传输介质。非易失性介质的常见形式包括，例如软盘、软磁盘、硬盘、磁带或者任意其它磁性介质，光盘 ROM(CD-ROM) 或者任意其它光学介质，穿孔卡片或者具有小孔图案的任意其它物理介质，可编程 ROM(PROM)，可擦除 PROM(EPROM)，电可擦 PROM(EEPROM)，快速存储器，任意其它存储芯片或盒式存储器，或者计算机系统 100 能够读取，并且适合于存储指令的任意其它介质。在本实施例中，非易失性存储器的一个例子是被描述成计算机系统 100 的内部组件的大容量存储装置 118，但是显然也可由外部装置提供。易失性介质包括诸如 RAM114 之类的动态存储器。传输介质包括同轴电缆、铜导线或者光纤，包括构成总线 122 的导线。传输介质还可采取声波或光波的形式，例如在射频或红外数据通信中产生的那些声波或光波。

[0035] 此外，本发明可被下载为计算机程序产品，其中通过到与总线 122 耦接的通信接口 132 的网络链路 134a-134n 之一，作为包括在载波或者其它传播介质中的数据信号，程序指令可从诸如服务器 140 之类的远程计算机转移到发出请求的计算机系统 100。通信接口 132 提供与多个网络链路 134a-134n 耦接的双向数据通信，网络链路 134a-134n 可与例如局域网 (LAN)、广域网 (WAN) 连接。当实现成服务器系统时，计算机系统 100 一般包括通过与输入 / 输出控制器连接的多个外设部件互连 (PCI) 总线桥接器可访问的多个通信接口。这样，计算机系统 100 允许与多个网络计算机连接。

[0036] 在网络环境中，计算机系统 100 通过网络 102 与其它系统通信。网络 102 可以指的是使用特殊协议，例如传输控制协议 (TCP) 和因特网协议 (IP) 相互通信的网络和网关的全球集合。网络 102 使用电信号、电磁信号或光信号传送数字数据流。携带数字数据往来于计算机系统 100 的各种网络内的信号，以及网络链路 134a-134n 上并通过通信接口 132 的信号是传送信息的载波的例证形式。虽然未示出，不过计算机系统 100 还可包括促进通信的多个外设组件。

[0037] 当计算机系统 100 被实现成 HA 群集中的服务器系统时，可包括另外的网络适配器，以支持与其它服务器系统的本地连接。另外，当被实现成 HA 群集中的服务器系统时，计算机系统 100 可被设计成商用硬件服务器，例如来自 IBM 公司的 xSeries™ 服务器。

[0038] 本领域的普通技术人员会认识到图 1 中描述的硬件可变化。此外，本领域的普通技术人员会认识到描述的例子并不意味着对本发明的结构限制。

[0039] 现在参考图 2，图 2 描述了在故障转移期间，有效地转移中间件的高可用性群集的硬件结构的方框图。如图所示，客户机系统 202 和 204 与网络 102 连接，以便传送服务请求。在本实施例中，客户机系统 202 和 204 向高可用性 (HA) 系统 208 请求服务，HA 系统 208 被配置成具有短的恢复时间，以及故障转移期间最少的提交数据丢失。

[0040] 如图所示，HA 系统 208 包括主节点 210 和副节点 220。如同下面所述，主节点 210 和副节点 220 最好实现当被执行时，提供高可用性系统的冗余硬件和软件。特别地，主节点 210 和副节点 220 实现在优选实施例中，支持 J2EE 应用程序的冗余中间件，后面将说明。中

间件是开发、集成和管理 web 应用程序和系统的软件。如后所述，中间件能够实现通信、进程、数据的集成，和交易容量与系统管理的自动化。

[0041] 特别地，Java™2 平台企业版 (J2EE) 提供用于建立 web 应用程序的可重复使用的组件模型。J2EE 定义 J2EE 平台的标准应用程序模型，托管应用程序的标准平台，兼容性要求和操作定义。这种开放源码模型的一个优点在于多个开发者能够用另外的组件和配置实现 J2EE 模型，而所有 J2EE 应用程序将在基于 J2EE 的系统上运行。

[0042] 国际商用机器公司 (IBM™) 的开发人员已开发出实现 J2EE 模型的软件。该软件通常填充 J2EE 构架中未规定的空白。例如，IBM™ 已开发了当在一群服务器上被实现时，支持 J2EE 软件产品的符合 J2EE 的软件产品的中间件堆栈。一般来说，中间件堆栈包括 web 服务器、数据库服务器和通用因特网应用程序服务器。具体地，该堆栈可包括诸如 IBM DB™，UDB 企业版，IBM HTTP 服务器和 IBMWebSphere™ 应用程序服务器之类的产品。

[0043] 另外，主节点 210 和副节点 220 实现监视 HA 群集中符合 J2EE 的中间件堆栈和硬件的故障及错误的监视和配置控制器。作为监视和配置控制器的一个例子，可实现 Tivoli™ 监视控制器，它填补了监视在 J2EE 构架中运行的软件的空白，并且简化了运行 J2EE 构架的系统的配置。

[0044] 按照使每个节点能够快速检查另一节点的心跳的简单、可靠方式，连接主节点 210 和副节点 220。在该实施例中，这种连接由连接在每个节点的网络适配器之间的跨接电缆 (cross-over cable)218 实现。特别地，跨接电缆 218 最好能够实现以太网连接，以便传送心跳数据。另一方面，在跨接电缆 218 发生故障的情况下，也可通过网络 102，跨越公共 IP 连接传送心跳数据。显然也可实现其它硬件来提供主节点 210 和副节点 220 之间的心跳通信通道，并且除了基于网络的连接之外，还可实现串行连接。

[0045] 特别地，当通过跨接电缆 218，在主节点 210 和副节点 220 之间发送心跳信号时，如果心跳失败，那么在该故障之前，副节点 220 将接管主节点 210 提供的服务。但是，如后所述，根据本发明的一个优点，在副节点 220 接管主节点 210 提供的服务之前，中间件组件会进一步分析心跳故障，并提供关于该故障的额外信息。此外，如后所述，通过跨接电缆 218，既能监视基于 Linux 的心跳，又能监视基于非 Linux 的心跳。

[0046] 主节点 210 和副节点 220 访问数据存储系统 214 和 224。有利的是，这里描述成 drbd 分区 230 的数据复制器包括每个数据存储系统 214 和 224 的一部分，用于复制主节点 210 和副节点 220 可访问的数据，而不需要在主节点 210 和副节点 220 之间实际物理共享的存储装置。根据本发明的一个优点，drbd 被配置成在该分区上运行，以便简化故障转移期间，从主节点 210 到副节点 220 的数据的转移。显然虽然关于 drbd 脚本管理的 drbd 分区描述了本发明，不过也可实现其它分布式数据复制系统。

[0047] 不间断电源 (UPS)212 和 UPS222 分别向主节点 210 和副节点 220 提供独立电源。最好，在 UPS212 与副节点 220，以及 UPS222 与主节点 210 之间也建立连接。在一个实施例中，从主节点 210 到 UPS222 设置串行电缆 216，从副节点 220 到 UPS212 设置串行电缆 226。但是，显然也可实现其它类型的连接硬件。

[0048] 根据本发明的一个优点，当在主节点 210 中发现故障时，在故障转移之后，副节点 220 开始接收先前被引向主节点 210 的请求。由于在主节点 210 上运行的硬件、软件或网络可能只有一部分发生故障，确保在故障转移之后，主节点 210 不试图更新数据的唯一方式

是关闭 UPS212。有利的是，如后所述，当检测到对待机节点 220 的故障转移时，群集管理器实现这里更详细说明的 STONITH，以便把命令从待机节点 220 引向 UPS212，从而关闭电源。

[0049] 现在参考图 3，图 3 描述了根据本发明的方法、系统和程序的群集管理器的方框图。如图所示，群集管理器包括用于实现有效故障转移的多个组件，包括心跳工具 402、drbd 脚本 404、mon406 和 stonith 函数 408。显然其它组件也可包括在群集管理器中，管理群集的其它方面。此外，显然另外的组件可包括在群集管理器 322 中，以管理故障转移。

[0050] 心跳工具 402 最好包括 Linux 的心跳软件包，配置成利用符合 J2EE 的中间件堆栈，管理 HA 群集内的故障转移。特别地，心跳工具 402 一般通过在群集中的两个节点之间发送“心跳”请求来起作用。如图 2 中所示，可通过每个节点的网络适配器之间的跨接电缆，发送心跳请求。当被应用于在服务器系统群集上运行的符合 J2EE 的中间件堆栈时，心跳工具 402 发送的心跳请求被分布在堆栈的不同层周围。

[0051] 如果心跳请求未能被返回，那么副节点会假定主节点发生故障，并接管 IP、数据和在主节点上运行的服务。当副节点接管 IP、数据和在主节点上运行的服务时，心跳工具 402 启动副节点的在待机模式下等待的组件，把 IP 地址分配给副节点的组件，并执行其它故障转移任务。

[0052] drbd404 是具有管理 HA 群集中的数据，以便提高故障转移期间数据的转换的相关脚本的内核模块。这通过对 drbd404 管理的块装置进行镜像来实现。drbd 是装入 drbd 模块，并配置以 HA 群集中的相关系统和共享的存储装置的 IP 地址的脚本。

[0053] 当被应用于符合 J2EE 的中间件堆栈时，drbd 管理的块装置提供中间件堆栈可在其上运行的存储空间。首先，配置群集，并安装 drbd 分区，以致只有主节点能够读写 drbd 管理的块装置。当发生故障转移时，drbd404 的数据盘脚本由心跳工具 402 运行，以便安装 drbd 分区，从而只有副节点能够读 / 写 drbd 管理的块装置。

[0054] mon406 是定期运行监视符合 J2EE 的中间件堆栈内的紧要系统服务的监视脚本的服务监视后台驻留程序。如果发现某一服务已发生故障或者被异常终止，那么 mon406 重新启动该服务，以确保中间件堆栈的所有组件仍然在主服务内运行。可因例如编程错误，或者灾难性操作系统事件，例如关于 RAM 的临时紧要资源约束，发生异常终止。特别地，当 mon 重启某一服务时，它用不同于停用服务的进程标识符 (PID)，但是相同的虚拟 IP 地址，重启该服务的新的实例。

[0055] stonith406 是心跳工具 402 调用的函数，以确保故障转移期间的数据完整性。特别地，stonith406 包括如图 2 中所示的到 UPS212 和 222 的串行电缆的配置。当心跳工具 402 调用 stonith406 时，该调用指定要关闭的节点。stonith 发送关闭被请求 UPS 的电源的信号。

[0056] 监视和配置控制器 410 包括被指定用于监视 HA 群集内的硬件和软件的状态的多个监视控制器。根据本发明的一个优点，HA 群集的多个硬件和软件组件的状态信息被转发给远程集中式企业控制台。最好，监视和配置控制器 410 补充 Java™ 管理扩展 (JMX) 以监视 HA 群集的硬件和软件组件，检测瓶颈和潜在的问题，自动从危险情形恢复群集。在一个实施例中，监视控制器由把监视的信息转发给 Tivoli™ 企业控制台 (TEC) 的 Tivoli™ 监视实现。

[0057] 特别地，在心跳工具 402 和 mon406 监视节点内的特定组件和特定服务实例的状态

的时候，监视和配置控制器 410 检测由这些工具监视的状况，并检测当心跳工具 402 被触发以启动故障转移，或者 mon406 被触发以重启服务器时，系统的整体状态。从而，监视和配置控制器 410 通过编辑当发生故障、错误和非理想情况时，节点的多个组件的状态，补充开放源码工具。

[0058] 根据本发明的一个优点，远程集中式监视控制台能够使用收集的信息确定配置变化。具体地说，根据本发明的一个优点，监视和配置控制器 410 的监视控制器都被配置成监视 HA 群集中的每个硬件组件，以及符合 J2EE 的中间件堆栈的每一层。从而，根据与所述硬件和各层中间件相关的监视信息，控制台能够确定哪些中间件层需要更多的存储空间来高速缓存请求，需要更多的线程来处理请求，或者需要按照某一其它方式被重新配置。控制台能够把配置变化发送给监视和配置控制器 410 的配置控制器，所述监视和配置控制器随后调整 HA 群集的配置。在一个实施例中，配置控制器是管理 HA 群集的配置特征的 Tivoli™ 配置管理器。

[0059] 根据本发明的一个优点，在企业系统中，控制台使用收集的信息确定哪些 HA 群集需要硬件和软件升级。例如，对于监视信息来说，控制台能够确定哪些商店具有似乎正在发生故障，需要被更换的硬件，哪些商店具有已达到容量，需要升级的硬件，和哪些商店具有正在发生故障或者未可靠运行的软件。

[0060] 根据本发明的另一优点，监视和配置控制器 410 与群集管理器 322 内的其它监视组件交互作用，收集发送给控制台的状态信息。例如，当 mon406 检测到任意被监视服务的故障时，监视和配置控制器 410 向远程集中式监视控制台发送通知，以致能够编译系统中故障的更大图象。此外，当心跳工具 402 启动系统的一个节点到另一节点的故障转移时，监视和配置控制器 410 向远程集中式监视控制台发送通知，以致能够收集节点故障统计数字。

[0061] 现在参见图 4，图 4 是根据本发明的方法、系统和程序，在故障转移之前的 HA 群集的软件结构的一个实施例的方框图。如图所示，主节点 210 和副节点 220 代表服务器系统的群集，每个被分配一个 IP 地址。

[0062] 根据本发明的一个优点，群集管理器 322 在主节点 210 和副节点 220 上运行，以便监视故障，重启服务，和当检测到故障时控制故障转移。如图所示，群集管理器 322 建立位于在主节点 210 和副节点 220 之间共享的存储器上的 drbd 分区 230。

[0063] 主节点 210 包括中间件堆栈的所有活动组件：负载均衡器 312，HTTP 服务器 314，web 应用程序服务器 (WAS)316，消息接发控制器 318，和数据库服务器 320。副节点 220 包括活动 HTTP 服务器 334 和 WAS336，但是，负载均衡器 332，消息接发控制器 338 和数据库 340 处于待机模式。

[0064] 负载均衡器 312 和 332 最好在也可被群集的 HTTP 和 WAS 服务器之间均衡请求的负载。最好，负载均衡器 312 和 314 通过利用服务器可用性，能力，工作量和其它标准，实现智能负载均衡。根据一个实施例，可通过 IBM WebSphere™ Edge 服务器实现负载均衡器 312 和 332。

[0065] 如图所示，负载均衡器 312 和 332 可实现与基于 Linux 的心跳无关的心跳。另一方面，基于 Linux 的心跳监视 332 和 342 可监视负载均衡器 312 和 332 的状态。

[0066] HTTP 服务器 314 和 334 可包括用于接收请求，和在 WAS316 及 336 间分布 HTTP 请

求的服务器群集。另外，当收到其它请求，例如对小服务程序（servlet）和 EJB 的请求时，使 HTTP 服务器 314 和 334 能够呼叫 enabler，例如小服务程序容器和 Enterprise Java™ Bean(EJB) 容器。根据一个实施例，可通过与 IBM 的 WebSphere™，尤其是 WebSphere™v5.0 捆绑的 HTTP 服务器来实现 HTTP 服务器 314 和 334。WebSphere™5.0 有利，因为能够从一个位置控制 WebSphere™ 组件的多个副本。从而，可在实现位于多个服务器系统上的软件组件的多个实例的一个位置完成配置变化。

[0067] 根据本发明的一个优点，HTTP 服务器 314 和 334 在活动 / 活动配置下运行，在活动 / 活动配置下，在主节点启动并运行之后，群集管理器 322 的心跳工具激活 HTTP 服务器。通过在活动 / 活动配置下运行 HTTP 服务器 314 和 334，可在两个（或更多的）服务器间分割请求负载，从而提高处理客户机请求的速度。另外，通过在活动 / 活动配置下运行 HTTP 服务器 314 和 334，故障转移时的启动时间被降低。

[0068] WAS316 和 336 最好包括使得能够支持向客户提供紧要使命服务的 web 应用的服务器群集，特别是使这些服务器支持 J2EE 应用。根据一个实施例，WAS316 和 336 是托管小服务程序，EJB 及支持 J2EE 应用和服务所需的其它 J2EE 组件的 IBM 的 WebSphere™5.0 支持的 WebSphere™ 应用服务器。

[0069] WAS316 与消息接发控制器 318 及数据库服务器 320 交互作用，提供与消息接发控制和数据库集成的应用服务器功能。根据本发明的一个优点，WAS316 和 WAS336 在活动 / 活动配置下运行。特别地，当初始化系统时，一旦消息接发控制器 318 和数据库服务器 320 可用，那么群集管理器 322 的心跳工具启动 WAS336 以产生活动 / 活动配置。通过运行活动 - 活动配置，可在多个系统群集间分割请求负载，以提高处理客户机请求的速度。另外，通过运行活动 / 活动配置，那么减少故障转移时的启动时间。

[0070] 消息接发控制器 318 和 338 包括监听异步请求，并把这些请求保存在本地队列中，提供队列以便与基于 J2EE 的系统通信的控制器。消息接发控制器 318 和 338 可实现 IBM MQSeries™，IBMWebSphere™ MQ，或者补充 Java™ 消息接发服务（JMS）的其它消息控制器。

[0071] 根据本发明的一个优点，消息接发控制器 318 和 338 在活动 / 待机配置下运行，在活动 / 待机配置下，群集管理器 322 的 drbd 管理 drbd 分区 230 中消息接发队列中的持久资源，群集管理器 322 的心跳工具控制故障转移中，消息接发控制器 338 的启动。

[0072] 数据库服务器 320 和 340 提供对永久存储器的控制。数据库 320 和 340 可通过诸如 IBM DB2UDB 企业版之类的数据库控制系统或者其它关系数据库管理系统来实现。

[0073] 根据本发明的一个优点，数据库服务器 320 和 340 在活动 / 待机配置下运行，在活动 / 待机配置下，群集管理器 322 的 drbd 管理 drbd 分区 230 中的数据库中的持久资源，群集管理器 322 的心跳工具控制故障转移中数据库服务器 340 的启动。

[0074] 为了消息接发控制器 318 和 338 及数据库服务器 320 和 340 在活动 / 待机配置下运行，并且以最小的数据损失快速实现故障转移，消息接发控制器 318 和数据库服务器 320 被配置成指向 drbd 分区 320 被安装成队列和数据库的存储空间的根的位置。另外，群集管理器 322 用消息接发控制器 318 和数据库服务器 320 的虚拟 IP 地址配置 drbd 和心跳工具。

[0075] 此外，根据本发明的一个优点，群集管理器 322 的 mon 功能定期运行监视紧要系统服务，例如消息接发控制器 318 和数据库服务器 320 提供的服务的监视脚本。如果发现某一服务已失败或者异常终止，mon 重启服务，以确保中间件堆栈的所有组件仍然在主服务内

运行。

[0076] 重要的是注意配置每层中间件以实现有效故障转移，并通过群集管理器 322 控制每层中间件的方法可被应用于其它类型的中间件。从而，随着可从与 J2EE 兼容的中间件软件堆栈获得的功能继续扩展，每个中间件组件可用活动 / 活动配置或活动 / 待机配置来构成，由群集管理器 322 监视，并在故障转移期间受到控制。

[0077] 现在参见图 5，图 5 描述根据本发明的方法、系统和程序，在故障转移之后，HA 群集的软件结构的一个实施例的方框图。如图所示，故障转移之后，主节点 210 被标记成失效节点。副节点 220 接管成为活动节点。

[0078] 当检测到故障，并且副节点 220 把主节点 210 标明为“停用”时，存在硬件和软件问题。具体地说，主节点 210 可能在要求的时间内不应答心跳请求，但是，会在之后不久即可工作。为了避免主节点 210 和副节点 220 都工作的情况，如前所述，群集管理器 322 的心跳工具调用 STONITH 关闭主节点 210 的 UPS。通过实现廉价的可由 STONITH 控制的 UPS，能够实现数据完整性，避免当主节点未真正被停用时，会发生的 HA 的“裂脑”问题。

[0079] 随后，在故障转移期间，负载均衡器心跳管理负载均衡器 332 的启动。当被激活时，群集管理器 322 的心跳工具把主节点 210 的虚拟 IP1 地址分配给负载均衡器 33。因此，对虚拟 IP 地址的请求被重定向到负载均衡器 332，从而不发生负载均衡群集的 IP 地址的任何变化。

[0080] 在故障转移期间，由于 HTTP 服务器 334 和 WAS336 已活动，群集管理器 322 的心跳工具不必启动这些组件。但是，由于消息接发控制器 338 和数据库服务器 340 处于待机状态，群集管理器 322 的心跳工具需要管理这些层的故障转移。首先，心跳工具将接收虚拟 IP2 地址。随后，心跳工具将启动 drbd 的数据盘服务，配置和安装 drbd 镜像分区。最后，心跳工具将启动配置到虚拟 IP2 地址的消息接发控制器 338 和数据库服务器 340，同时在镜像 drbd 分区 230 上启动消息队列和数据库实例。另一方面，虽然未示出，数据库服务器 340 可以处于活动模式，而不是待机模式，因为每次虚拟 IP2 地址只适用于一个节点。由于在请求到达之前，数据库服务器 340 并不试图接触 drbd 分区 230 上的数据，因此在故障转移时，数据库服务器 340 被配置到虚拟 IP2 地址，在请求到达之前，镜像 drbd 分区 230 是可访问的。相反，一些层，例如消息接发控制器 338 在启动时直接装入数据，从而如果在故障转移之前在副节点 220 上被启动，那么将会崩溃，因为在故障转移之前，drbd 分区 230 上的数据对副节点 220 不可用。

[0081] 现在参见图 6，图 6 描述 HA 系统中，符合 J2EE 的中间件内的独立软件开发商 (ISV) 应用程序的一个实现例子的方框图。如图所示，活动 WAS602，活动 IBM MQSeries™ 服务器 610 和活动 IBM DB2 服务器 614 举例说明与 drbd 分区 630 进行接口的符合 J2EE 的中间件堆栈的主节点的一部分。如附图标记 620 所示，在活动的 WebSphere™ 应用服务器 602 接收物品销售或交易完成。ISV 可对小服务程序或 EJB 编程，以处理特殊类型的到来请求。例如，如附图标记 620 所示，查寻小服务程序 604 是当在收银机扫描物品的价格时，处理价格查寻 (PLU)，检查物品的价格的 ISV web 应用程序。查寻小服务程序 602 随后张贴该零售交易将由另一组件，例如交易小服务程序 608 或者另一小服务程序或者 EJB 异步完成的请求。但是，首先，如附图标记 622 所示，信息被传送给 MQ 监听器 612，并放置在 MQ 队列 632 上，以释放查寻小服务程序 604 接收下一到来请求，并确保通过 MQ 队列 632，按照顺序一次

准确记录交易。随后，如附图标记 624 所示，随后调用 MDB606，从 MQ 队列 632 取出该交易，并如附图标记 626 所示，把该交易送给交易小服务程序 626。交易小服务程序 626 最终处理该 PLU，并如附图标记 628 所示，把结果提交给 IBM DB2 控制器 616，以便保存在 DB2634 中。

[0082] 特别地，图 6 图解说明了在故障转移期间，HA 系统中的符合 J2EE 的中间件堆栈的优点，因为即使在请求已开始在堆栈的各层之间转移之后，发生故障转移，该堆栈也能确保每个交易将被立即准确记录。另外，图 6 图解说明了在故障转移期间，HA 系统中的符合 J2EE 的中间件堆栈的优点，因为活动各层 MQSeries™ 服务器 610 和 DB2 服务器 614 与只有主节点能够访问、但是被快速重新安装以便在故障转移期间由副节点访问的 drbd 分区 630 进行接口。

[0083] 现在参考图 7，图 7 描述了把 drbd 分区配置到 HA 群集中的符合 J2EE 的中间件堆栈的进程和程序的高级逻辑流程图。如图所示，进程开始于方框 700，之后进行到方框 702。方框 702 描述配置和安装 drbd 分区。接下来，方框 704 表示激活 drbd 分区上的消息队列和数据库。之后，方框 706 图解说明记录访问 drbd 分区的消息接发服务器和数据库服务器的虚拟 IP 地址，以便在故障转移期间，有效转移对 drbd 分区的访问，之后该进程结束。

[0084] 现在参见图 8，图 8 描述通过心跳控制器，控制 HA 群集中符合 J2EE 的中间件堆栈的配置和故障转移的进程和程序的高级逻辑流程图。如图所示，该进程开始于方框 800，之后进行到方框 802。方框 802 图解说明激活主节点的中间件层。之后，方框 804 说明激活副节点的 HTTP 服务器和 WAS 中间件层。另外，指定成在活动 - 活动配置下运行的其它中间件层被激活。之后，方框 806 描述定期启动从副节点到主节点的心跳请求。方框 808 描述副节点是否检测到心跳返回的确定。如果检测到心跳返回，那么进程返回方框 806。如果没有检测到心跳返回，那么进程进行到方框 810。

[0085] 方框 810 描述调用 STONITH 关闭主节点的电源。随后，方框 812 描述从主节点接管虚拟 IP 地址，以便分配给副节点中的冗余组件。之后，方框 814 描述调用数据盘脚本，重新安装 drbd 分区，以便由副节点访问，之后进程结束。随后，方框 816 描述激活副节点上的待机中间件层，并启动 drbd 分区上的数据。显然心跳工具和故障转移期间的其它群集管理服务可执行另外的步骤。

[0086] 现在参见图 9，图 9 是描述控制用于监视符合 J2EE 的中间件堆栈提供的服务的 mon 功能的进程和程序的高级逻辑流程图。如图所示，该进程开始于方框 900，之后进行到方框 902。方框 902 描述配置用于监视由中间件提供的服务的时间表。随后，方框 904 描述预定的监视时间是否被触发的确定。如果预定的监视时间未被触发，那么该进程在方框 904 重复。如果预定的监视时间被触发，那么进程进行到方框 906。方框 906 描述监视预定服务的状态。之后，方框 908 确定服务是否以某种方法被检测为停用或故障。如果服务未被检测为停用，那么进程结束。如果服务被检测为停用，那么进程进行到方框 910。方框 910 描述用新的 PID 重启相同服务，之后进程结束。

[0087] 现在参见图 10，图 10 描述根据本发明的方法、系统和程序，包括运行 J2EE 中间件堆栈，并由远程企业控制器管理的多个 HA 系统的企业网络的方框图。如图所示，HA 系统 1202 和 HA 系统 1204 与远程企业控制台 1210 通信连接，远程企业控制台 1210 通过网络 102 监视并遥控 HA 系统 1202 和 1204。显然单个或多个远程中央控制台可监视和控制多个 HA 系统。

[0088] 根据本发明的一个优点，每个 HA 系统 1202 和 1204 能够处理零售交易和其它紧要使命操作。根据一个实施例，每个 HA 系统 1202 和 1204 可通过能够实现 J2EE 应用程序的冗余的符合 J2EE 的中间件堆栈，例如图 4 和 5 中图解说明的中间件堆栈实现高可用性。特别地，每个 HA 系统 1202 和 1204 包括运行监视和配置控制器 410 的群集管理器，如图 3 中所示。

[0089] 有利的是，当在任意 HA 系统 1202 和 1204 发出错误、故障或非理想情况时，监视和配置控制器 410 检查错误、故障或其它非理想情况时系统的状况，随后编译该信息，向远程企业控制台 1210 报告。根据本发明的一个优点，如果心跳监视器或 mon 功能检测到故障或错误，那么监视和配置控制器 410 被触发，以便检测所述故障或错误，并确定故障或错误时的系统状况。

[0090] 远程企业控制台 1210 最好把监视信息保存在数据库中。随后，远程企业控制台 1210 最好包括用于分析从 HA 系统 1202 和 1204 接收的错误或故障信息，并且可能把配置变化返回给 HA 系统，以便试图防止和改善故障转移的效率的第一控制器。另外，远程企业控制台 1210 可包括比较从多个 HA 系统接收的故障、错误和其它信息，确定哪些系统需要修理和升级，哪些系统不满足性能要求的第二控制器。远程企业控制台 1210 可收集并控制 HA 系统 1202 和 1204 的性能统计数字的显示。

[0091] 现在参见图 11，图 11 描述根据本发明的方法、系统和程序，控制 HA 群集管理器内的监视控制器的进程和程序的高级逻辑流程图。如图所示，该进程开始于方框 1000，之后进行到方框 1002。方框 1002 描述确定心跳监视器，mon，或者监视 HA 系统中的中间件堆栈的其它监视控制器是否检测到故障或错误。如果没有检测到故障或错误，那么该进程在方框 1002 重复。如果检测到故障或错误，那么进程进行到方框 1004。方框 1004 描述收集并分析故障或错误时的可用系统信息。随后，方框 1006 描述把故障或错误和可用系统信息发送给监视 HA 系统的远程中央控制台，之后进程结束。

[0092] 现在参见图 12，图 12 是描述远程控制 HA 系统的群集管理器，从而重构 HA 系统的进程和程序的高级逻辑流程图。如图所示，该进程开始于方框 1100，之后，进行到方框 1102。方框 1102 描述确定是否从远程企业控制台收到重构运行中间件堆栈的 HA 系统的配置请求。如果没有收到请求，那么该进程在方框 1102 重复。如果收到请求，那么进程进行到方框 1104。方框 1104 描述调用心跳监视器重构 HA 系统故障转移设置，之后进程结束。另外，HA 系统的群集管理器内的其它控制器可被调用，以调整 HA 系统的其它软件和硬件配置。

[0093] 现在参见图 13，图 13 描述控制用于管理群集中的多个 HA 系统的远程企业控制台的进程和程序的高级逻辑流程图。如图所示，进程开始于方框 1300，之后进行到方框 1302。方框 1302 描述确定是否从 HA 系统收到监视信息。如果没有收到监视信息，那么进程在方框 1302 重复。如果收到监视信息，那么进程进行到方框 1304。特别地，远程企业控制台定期向每个 HA 系统发送关于监视信息的请求，每个 HA 系统也可自动发送监视信息。

[0094] 方框 1304 描述把监视信息加入企业数据库中，所述企业数据库中保存来自多个 HA 系统的监视信息。随后，方框 1306 描述如果监视信息触发重构，那么请求 HA 系统的重构。特别地，远程企业控制台可包括当在监视信息中检测到一些特定类型的错误时，将被请求的预定配置。另一方面，系统管理员可为特定类型的错误，推荐配置的类型。之后，方框

1308 描述根据监视信息，重新计算 HA 系统的性能统计数字。特别地，可以只对某些类型的监视错误或波动，触发性能统计数字的计算。随后，方框 1312 描述比较该 HA 系统的性能与企业网络中的其它 HA 系统的性能和企业网络的性能要求集合。随后，方框 1314 描述以图表的形式显示比较的性能结果。例如，图表可描述 HA 系统的位置的图形表示，提供哪些系统已发生故障的图形指示符，并提供表示每个 HA 系统相对于其它 HA 系统的性能的图形指示符。此外，可显示每个系统的实时性能和报告的任意错误。随后，方框 1316 描述推荐针对 HA 系统弱点的纠正动作，之后进程结束。例如，推荐可指示哪些 HA 系统需要被更换，哪些 HA 系统需要升级，哪些 HA 系统需要软件升级或微调。显然图 13 中描述的进程是可对从多个高可用性服务器接收的监视信息执行的各种进程的例子，并且在不脱离本发明的范围的情况下，可执行其它类似的分析和输出。

[0095] 虽然参考优选实施例具体说明了本发明，但是对本领域的技术人员来说，在不脱离本发明的精神和范围的情况下，显然可做出形式和细节方面的各种变化。

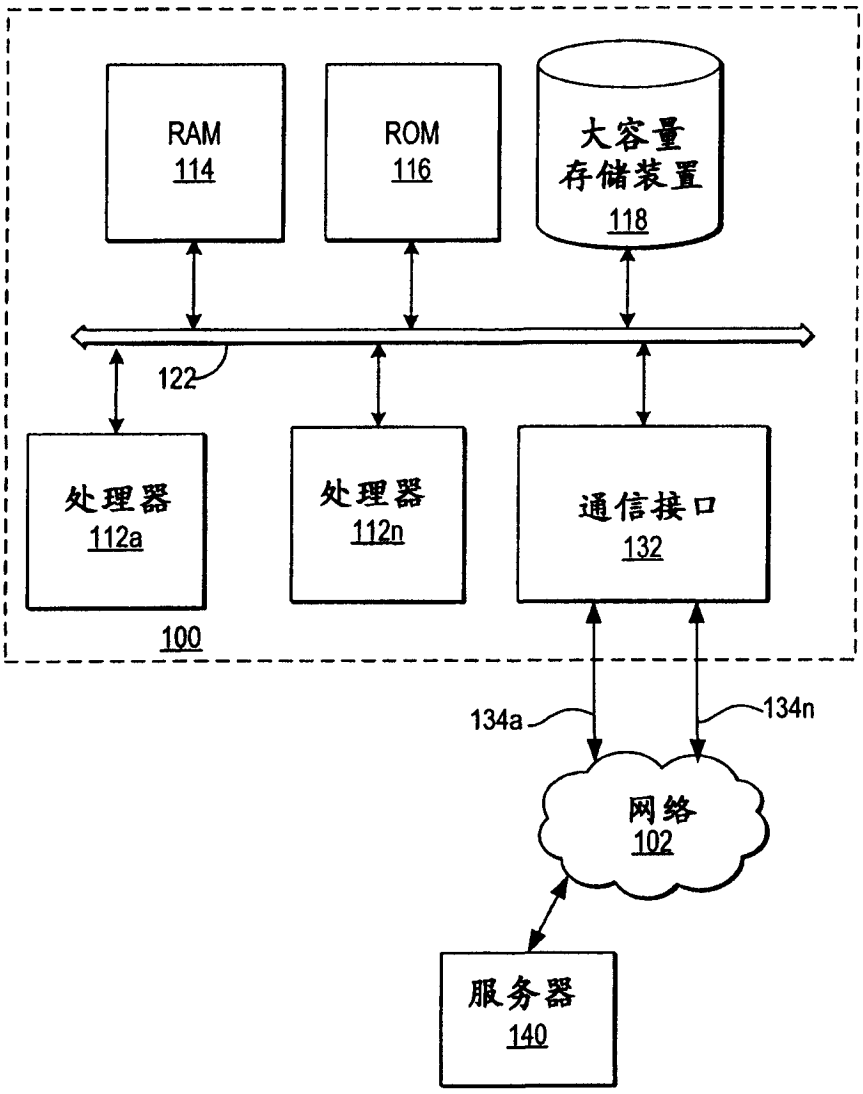

RAM
114
ROM
116
大容量
存储装置
118
122
处理器
112a
处理器
112n
通信接口
132
100
134a
134n
网络
102
服务器
140

图 1

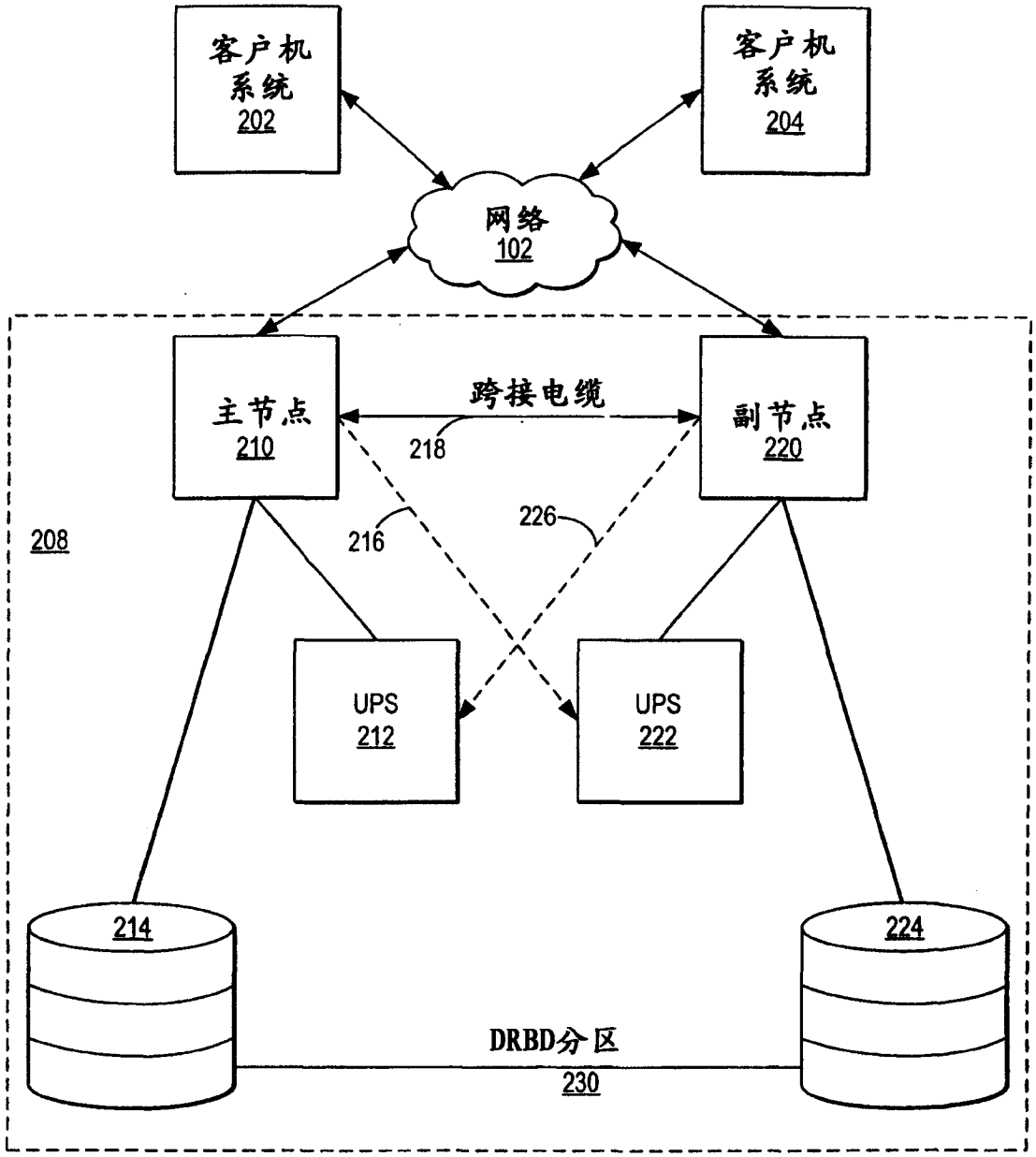

客户机
系统
202
客户机
系统
204
网络
102
主节点
210
跨接电缆
218
副节点
220
208
216
226
UPS
212
UPS
222
214
224
DRBD分区
230

图 2

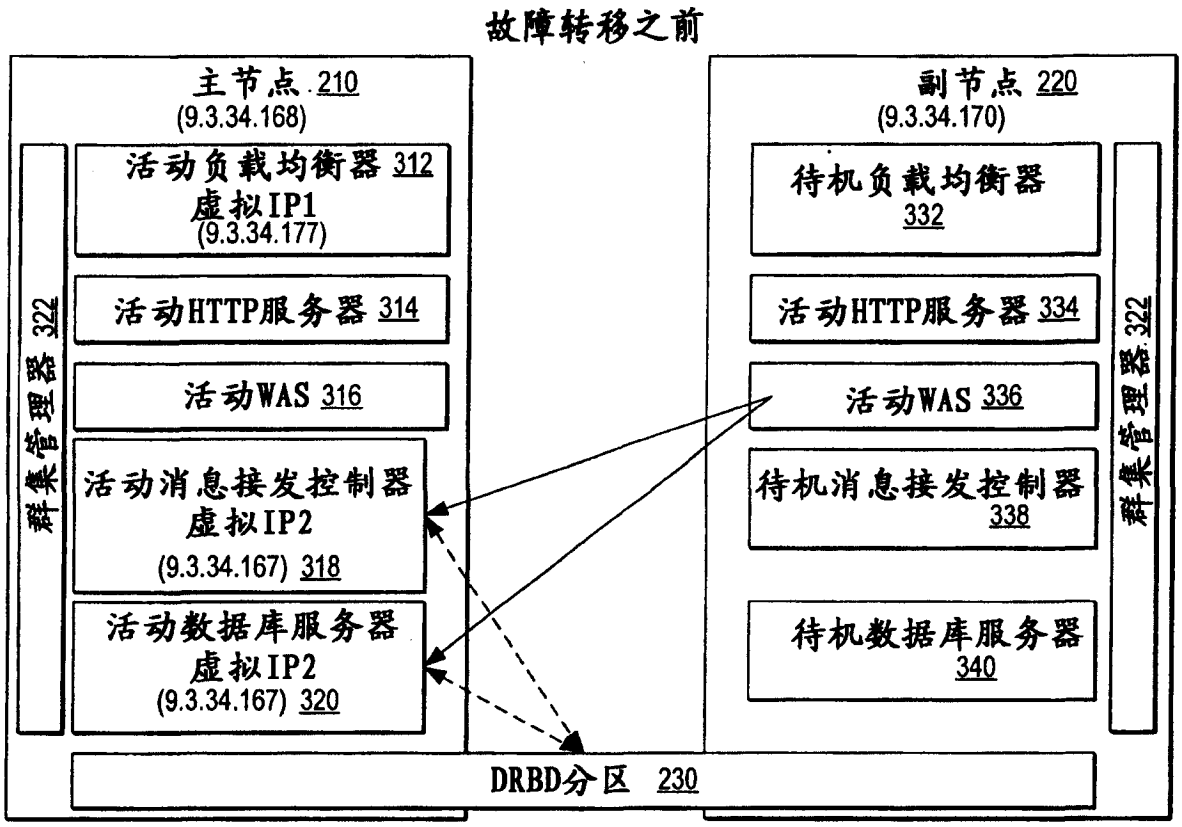

故障转移之前
主节点 210
(9.3.34.168)
活动负载均衡器 312
虚拟IP1
(9.3.34.177)
活动HTTP服务器 314
活动WAS 316
活动消息接发控制器
虚拟IP2
(9.3.34.167) 318
活动数据库服务器
虚拟IP2
(9.3.34.167) 320
群集管理器 322
副节点 220
(9.3.34.170)
待机负载均衡器
332
活动HTTP服务器 334
活动WAS 336
待机消息接发控制器
338
待机数据库服务器
340
群集管理器 322
DRBD分区 230

图 4

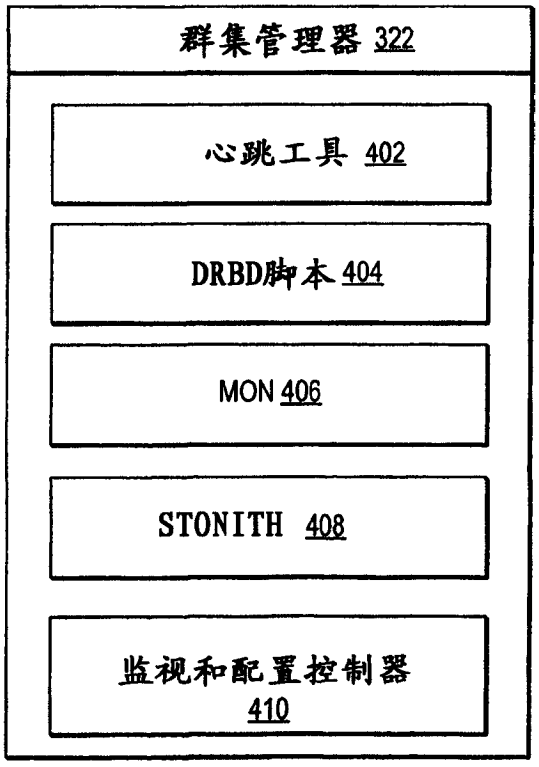

群集管理器 322
心跳工具 402
DRBD脚本 404
MON 406
STONITH 408
监视和配置控制器
410

图 3

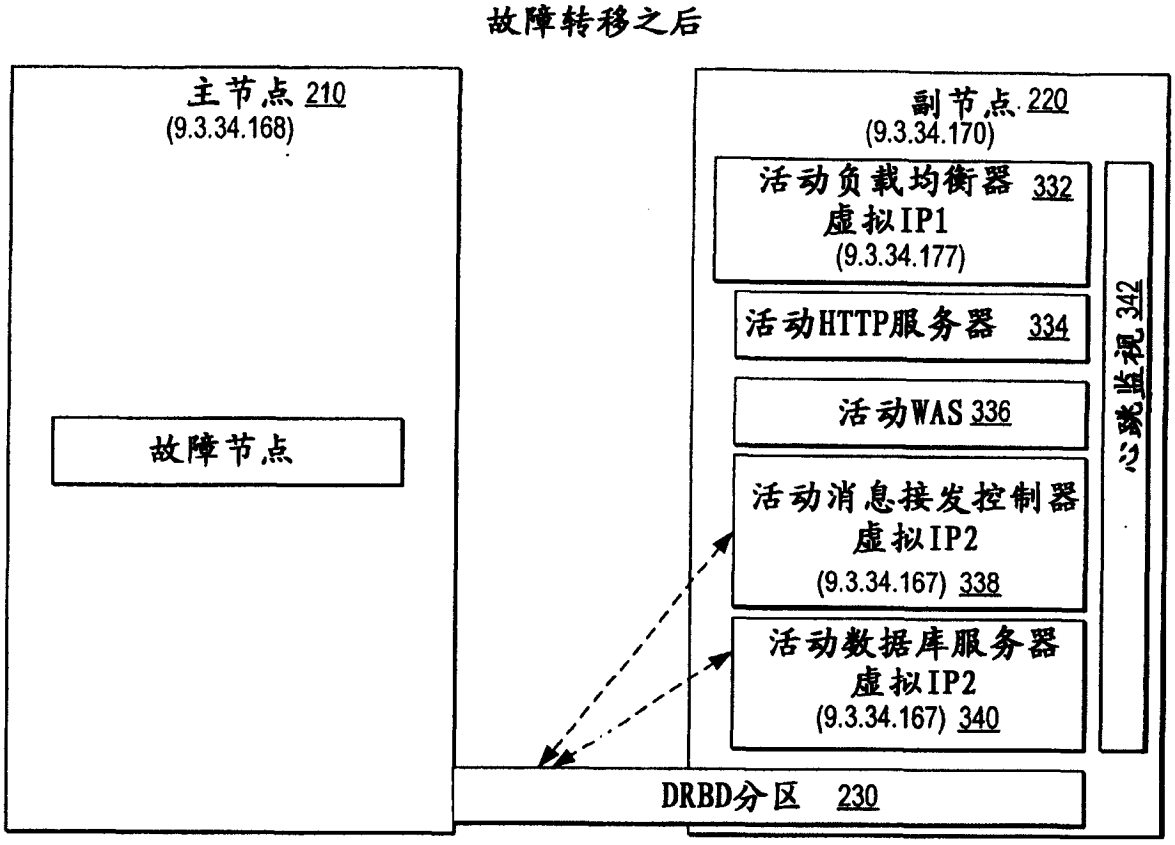

故障转移之后
主节点 210
(9.3.34.168)
故障节点
副节点 220
(9.3.34.170)
活动负载均衡器 332
虚拟IP1
(9.3.34.177)
活动HTTP服务器 334
活动WAS 336
活动消息接发控制器
虚拟IP2
(9.3.34.167) 338
活动数据库服务器
虚拟IP2
(9.3.34.167) 340
DRBD分区 230
监视 342

图 5

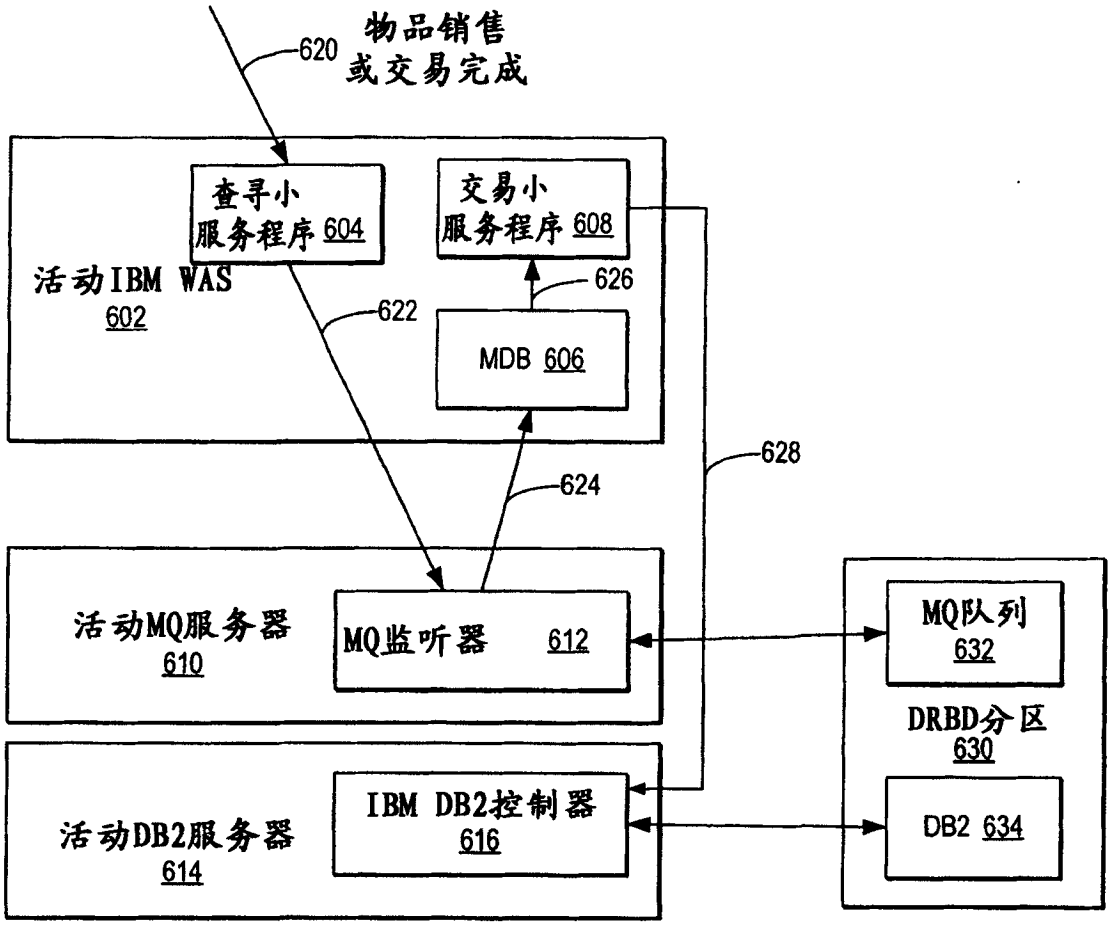

物品销售
或交易完成
620
查寻小
服务程序 604
交易小
服务程序 608
活动IBM WAS
602
626
622
MDB 606
628
624
活动MQ服务器
610
MQ监听器 612
MQ队列
632
DRBD分区
630
活动DB2服务器
614
IBM DB2控制器
616
DB2 634

图 6

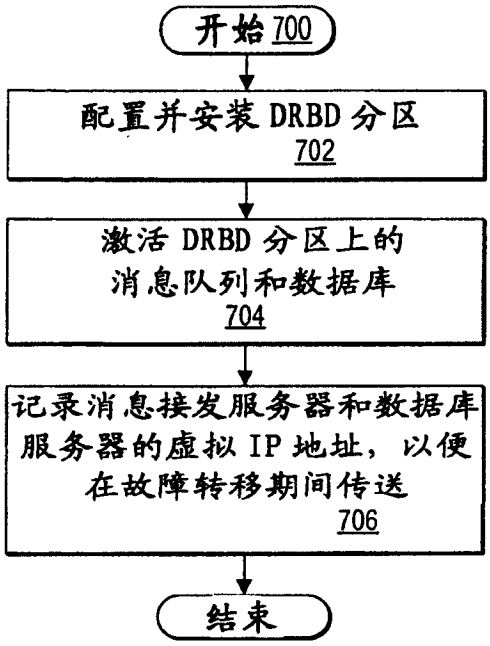

开始 700
配置并安装 DRBD 分区
702
激活 DRBD 分区上的
消息队列和数据库
704
记录消息接发服务器和数据库
服务器的虚拟 IP 地址，以便
在故障转移期间传送
706
结束

图 7

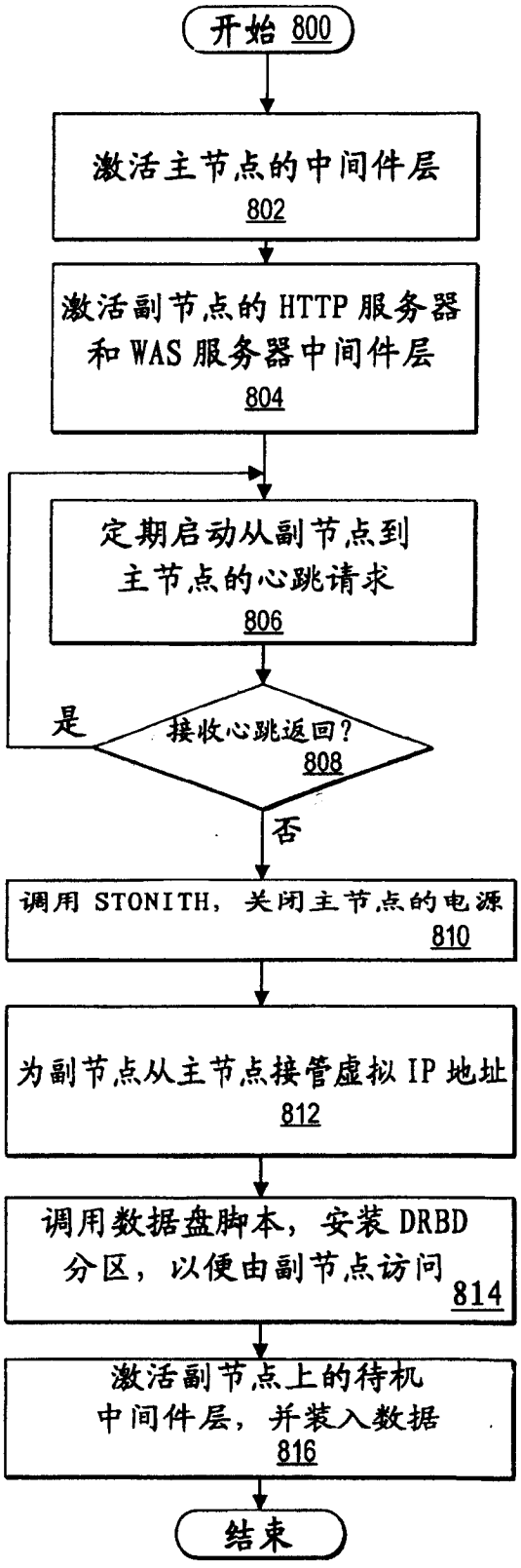

开始 800
激活主节点的中间件层
802
激活副节点的 HTTP 服务器
和 WAS 服务器中间件层
804
定期启动从副节点到
主节点的心跳请求
806
是
接收心跳返回?
808
否
调用 STONITH，关闭主节点的电源
810
为副节点从主节点接管虚拟 IP 地址
812
调用数据盘脚本，安装 DRBD
分区，以便由副节点访问
814
激活副节点上的待机
中间件层，并装入数据
816
结束

图 8

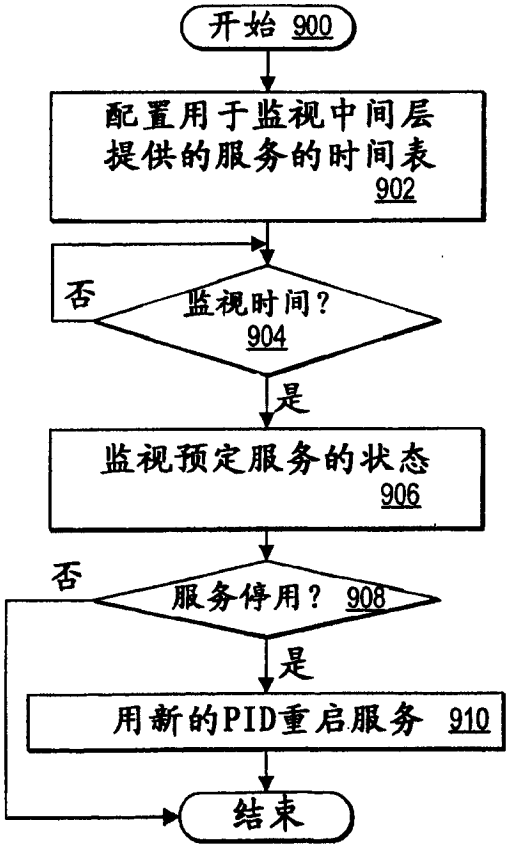

开始 900
配置用于监视中间层提供的服务的时间表 902
否
监视时间？ 904
是
监视预定服务的状态 906
否
服务停用？ 908
是
用新的PID重启服务 910
结束

图 9

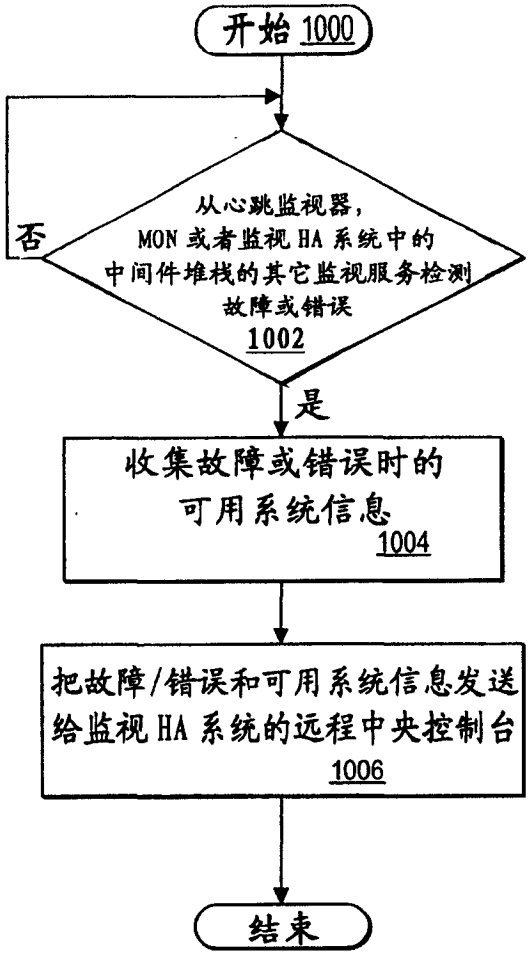

开始 1000
否
从心跳监视器，MON 或者监视 HA 系统中的中间件堆栈的其它监视服务检测故障或错误 1002
是
收集故障或错误时的可用系统信息 1004
把故障/错误和可用系统信息发送给监视 HA 系统的远程中央控制台 1006
结束

图 11

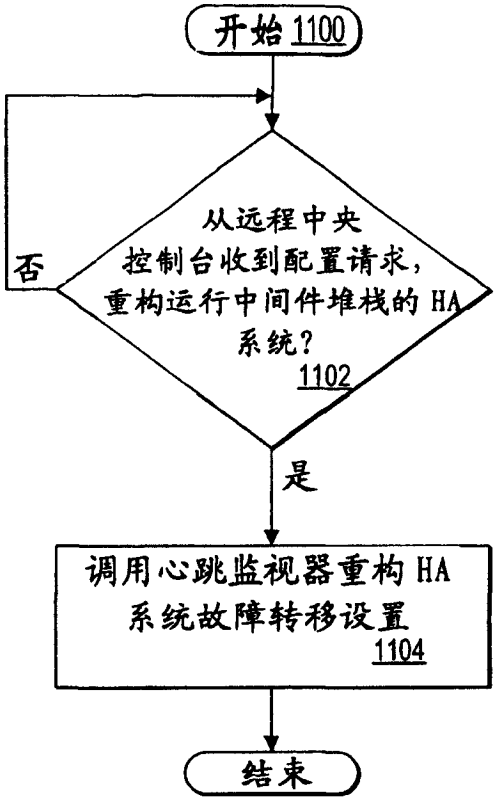

开始 1100
从远程中央控制台收到配置请求，重构运行中间件堆栈的 HA 系统？
1102
否
是
调用心跳监视器重构 HA 系统故障转移设置
1104
结束

图 12

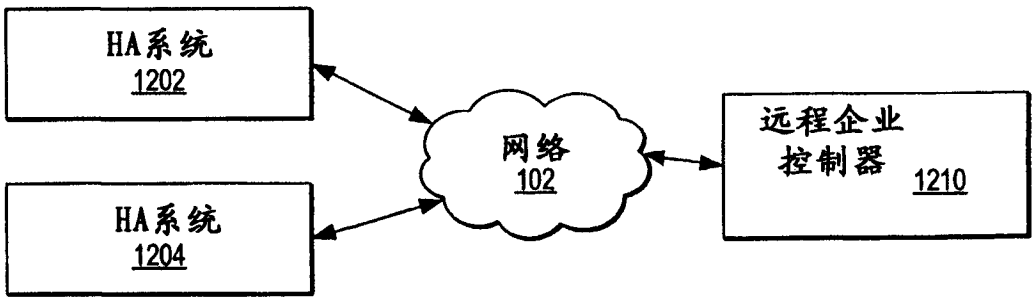

HA系统
1202
HA系统
1204
网络
102
远程企业控制器
1210

图 10

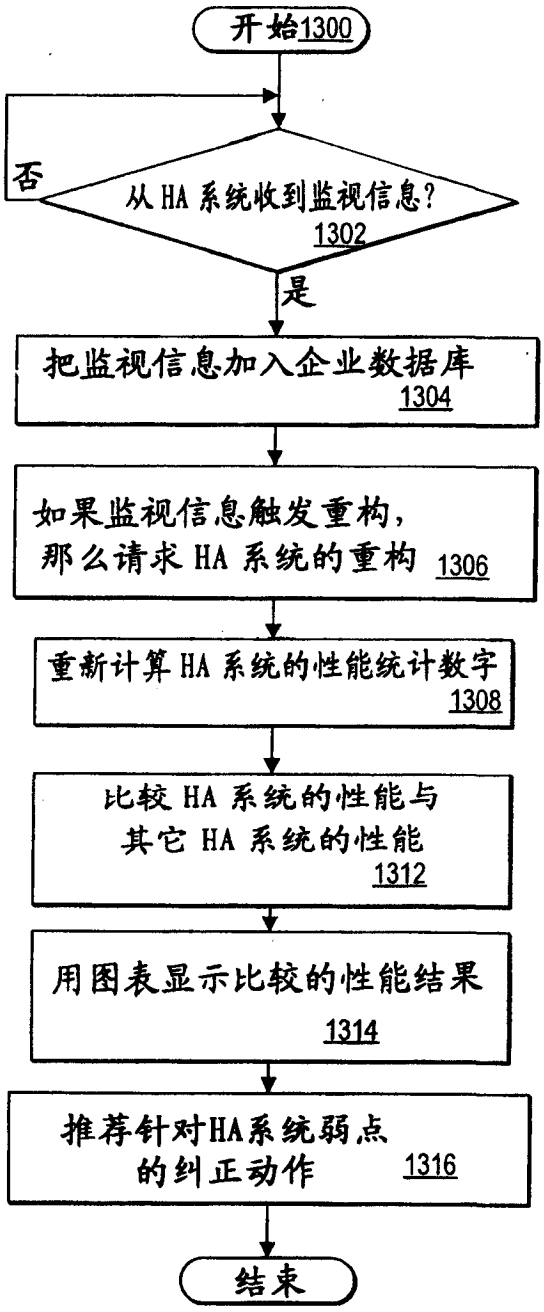

开始1300
否
从 HA 系统收到监视信息?
1302
是
把监视信息加入企业数据库
1304
如果监视信息触发重构，
那么请求 HA 系统的重构 1306
重新计算 HA 系统的性能统计数字
1308
比较 HA 系统的性能与
其它 HA 系统的性能
1312
用图表显示比较的性能结果
1314
推荐针对HA系统弱点
的纠正动作 1316
结束

图 13