(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103377281 A
(43)申请公布日 2013.10.30

(21)申请号 201310139393.X

(22)申请日 2013.04.22

(30)优先权数据
13/453,019 2012.04.23 US

(71)申请人 国际商业机器公司
地址 美国纽约

(72)发明人 S·A·巴赛特 R·A·胡森
R·马亨德鲁 H·V·拉马萨米
S·萨卡尔 唐春强 N·G·沃格尔
王龙

(74)专利代理机构 北京市中咨律师事务所 11247
代理人 于静 张亚非

(51)Int.Cl.
G06F 17/30(2006.01)

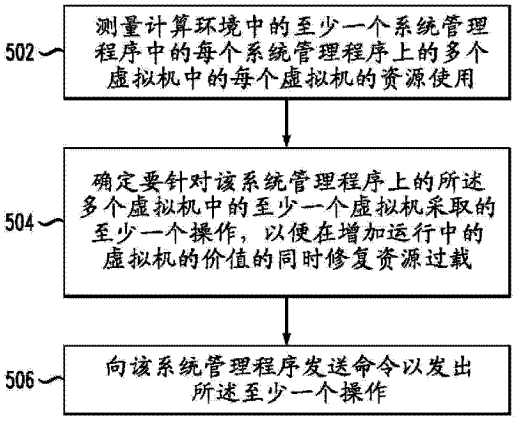
权利要求书3页 说明书11页 附图4页

(54)发明名称

用于修复过度承诺的计算环境中的过载的方法和系统

(57)摘要

本发明涉及一种用于修复过度承诺的计算环境中的过载的方法和系统。所述方法包括:测量计算环境中的至少一个系统管理程序中的每个系统管理程序上的多个虚拟机中的每个虚拟机的资源使用;在检测到所述至少一个系统管理程序中的一个系统管理程序上的资源过载时:确定要针对该系统管理程序上的所述多个虚拟机中的至少一个虚拟机采取的至少一个操作,以便在增加运行中的虚拟机的价值的同时修复资源过载;以及向该系统管理程序发送命令以发出所述至少一个操作。



CN 103377281 A

1. 一种用于修复过度承诺的计算环境中的过载的方法，所述方法包括：

测量计算环境中的至少一个系统管理程序中的每个系统管理程序上的多个虚拟机中的每个虚拟机的资源使用；

在检测到所述至少一个系统管理程序中的一个系统管理程序上的资源过载时：

确定要针对该系统管理程序上的所述多个虚拟机中的至少一个虚拟机采取的至少一个操作，以便在增加运行中的虚拟机的价值的同时修复资源过载；以及

向该系统管理程序发送命令以发出所述至少一个操作；

其中由计算机设备执行上述步骤中的至少一个步骤。

2. 根据权利要求 1 的方法，其中所述至少一个操作包括迁移、静默或恢复中的至少一个。

3. 根据权利要求 1 的方法，其中确定要采取的至少一个操作包括计算该系统管理程序上的每个虚拟机的工作价值，其中工作价值包括以下项中的至少一个：由该虚拟机被静默导致的该虚拟机上的潜在服务损失的测量，以及该虚拟机被静默对至少一个其它虚拟机的执行的影响的测量。

4. 根据权利要求 3 的方法，其中确定要采取的至少一个操作包括标识要静默的具有最低工作价值的至少一个虚拟机和 / 或要恢复以便优化的具有较高工作价值的至少一个虚拟机。

5. 根据权利要求 3 的方法，其中每个虚拟机的工作价值包括至少一个因素的组合，所述至少一个因素包括虚拟机生命周期、价格、收益、服务水平协议、中央处理单元利用率、虚拟机相关性以及工作负载预测。

6. 根据权利要求 3 的方法，其中确定要采取的至少一个操作包括持续测量和利用每个虚拟机的工作价值和资源消耗。

7. 根据权利要求 6 的方法，还包括：

使用工作价值与资源消耗的比率判定是否要针对虚拟机采取迁移、静默或恢复操作中的任何一个。

8. 根据权利要求 1 的方法，还包括：

定期判定是否有足够的资源变得可用以便恢复被静默的虚拟机。

9. 一种用于修复过度承诺的计算环境中的过载的系统，所述系统包括：

至少一个不同软件模块，每个不同软件模块包含在有形的计算机可读介质中；

存储器；以及

至少一个处理器，其耦合到所述存储器并可操作以：

测量计算环境中的至少一个系统管理程序中的每个系统管理程序上的多个虚拟机中的每个虚拟机的资源使用；

在检测到所述至少一个系统管理程序中的一个系统管理程序上的资源过载时：

确定要针对该系统管理程序上的所述多个虚拟机中的至少一个虚拟机采取的至少一个操作，以便在增加运行中的虚拟机的价值的同时修复资源过载；以及

向该系统管理程序发送命令以发出所述至少一个操作。

10. 根据权利要求 9 的系统，其中所述至少一个操作包括迁移、静默或恢复中的至少一个。

11. 根据权利要求 9 的系统，其中所述至少一个处理器还可操作以计算该系统管理程序上的每个虚拟机的工作价值，其中工作价值包括以下项中的至少一个：由该虚拟机被静默导致的该虚拟机上的潜在服务损失的测量，以及该虚拟机被静默对至少一个其它虚拟机的执行的影响的测量。

12. 根据权利要求 11 的系统，其中所述至少一个处理器还可操作以标识要静默的具有最低工作价值的至少一个虚拟机和 / 或要恢复以便优化的具有较高工作价值的至少一个虚拟机。

13. 根据权利要求 11 的系统，其中每个虚拟机的工作价值包括至少一个因素的组合，所述至少一个因素包括虚拟机生命周期、价格、收益、服务水平协议、中央处理单元利用率、虚拟机相关性以及工作负载预测。

14. 根据权利要求 11 的系统，其中所述至少一个处理器还可操作以确定要采取的至少一个操作包括：持续测量和利用每个虚拟机的工作价值和资源消耗。

15. 根据权利要求 11 的系统，其中所述至少一个处理器还可操作以使用工作价值与资源消耗的比率判定是否要针对虚拟机采取迁移、静默或恢复操作中的任何一个。

16. 根据权利要求 11 的系统，其中所述至少一个处理器还可操作以定期判定是否有足够的资源变得可用以便恢复被静默的虚拟机。

17. 一种用于虚拟化环境的自动过载修复方法，所述方法包括：

通过从系统内的多个虚拟机选择所述虚拟机的要迁移的子集以及所述虚拟机的要静默的子集而从系统资源过载情况中恢复；以及

定期选择所述虚拟机的要迁移的子集以及所述虚拟机的要恢复的子集，以便增加运行中的虚拟机数量并增加运行中的虚拟机实现的工作价值。

18. 根据权利要求 17 的方法，其中迁移、静默和恢复选择均考虑每个虚拟机所贡献的工作价值度量。

19. 根据权利要求 18 的方法，其中所述工作价值度量组合一个或多个因素，所述一个或多个因素包括虚拟机生命周期、价格、收益、服务水平协议、中央处理单元利用率、虚拟机相关性以及工作负载预测。

20. 根据权利要求 19 的方法，其中问题被表述为可移除在线多背包问题 ROMKP 的变型，包括：

除新物体的到达以外，考虑随时间变化的物体工作价值和资源使用；

允许背包操作，所述背包操作包括从背包移除一个或多个物体、将一个或多个物体从一个背包移动到另一个背包，以及将一个或多个物体放回背包；以及

在决策中考虑所述背包操作的成本。

21. 根据权利要求 20 的方法，其中 ROMKP 问题的所述变型的解包括：

持续测量和利用每个虚拟机的所述工作价值和资源使用；

使用工作价值与资源使用的比率作为决策准则，以便针对一个或多个虚拟机选择迁移、静默或恢复操作；以及

计算每个迁移、静默和 / 或恢复操作的成本，并使用所述成本作为问题公式化的约束。

22. 根据权利要求 18 的方法，其中虚拟机的工作价值被定义为：

1/（资源使用），如果该虚拟机在给定最近时间段 P 内未被静默；以及

无穷大，如果该虚拟机在所述时间段 P 内被静默。

用于修复过度承诺的计算环境中的过载的方法和系统

技术领域

[0001] 本发明的实施例一般地涉及信息技术，更具体地说，涉及虚拟机管理。

背景技术

[0002] 计算云系统中的资源过度订阅正在成为信息技术(IT)服务中的普遍现象。通常供应虚拟机(VM)以托管云系统中的单独服务。传统上，为这些VM供应其以物理资源保证的所有被分配资源(例如，存储器和中央处理单元(CPU))。超过可以由物理资源保证的数量的资源分配(即，资源过度订阅)允许在系统管理程序上托管更多的VM。因为VM通常并不消耗所有被分配资源，所以过度订阅可以增加资源利用率，因此降低服务供应成本。

[0003] 但是，过度订阅带来的风险是在托管的VM用完被分配资源之前，系统管理程序中出现资源过载，并且资源过载(尤其是存储器过载)极大地降低服务性能。

[0004] 试图补救过度订阅的不良后果的现有方法包括仅迁移解决方案。但是，仅迁移解决方案并不足以修复过度承诺(over-committed)的系统中的过载(在这些系统中，存在大量系统管理程序同时接近过载的可能性)，并且在密集型过度承诺的系统管理程序群集中迁移开销可能很大。

[0005] 因此，需要修复过度承诺的计算环境中的过载。

发明内容

[0006] 在本发明的一个方面，提供了用于修复资源过载的技术。一种用于修复过度承诺的计算环境中的过载的示例性计算机实现的方法可以包括以下步骤：测量计算环境中的至少一个系统管理程序中的每个系统管理程序上的多个虚拟机中的每个虚拟机的资源使用；在检测到所述至少一个系统管理程序中的一个系统管理程序上的资源过载时：确定要针对该系统管理程序上的所述多个虚拟机中的至少一个虚拟机采取的至少一个操作，以便在增加运行中的虚拟机的价值的同时修复资源过载；以及向该系统管理程序发送命令以发出所述至少一个操作。

[0007] 在本发明的另一个方面，提供了用于虚拟化环境的自动过载修复方法的技术。一种示例性计算机实现的方法包括以下步骤：通过从系统内的多个虚拟机选择所述虚拟机的要迁移的子集以及所述虚拟机的要静默(quiesce)的子集而从系统资源过载情况中恢复；以及定期选择所述虚拟机的要迁移的子集以及所述虚拟机的要恢复的子集，以便增加运行中的虚拟机数量并增加运行中的虚拟机实现的工作价值。

[0008] 本发明的另一个方面或其元素可以以制品的形式实现，所述制品有形地包含计算机可读指令，当执行所述计算机可读指令时，导致计算机执行在此描述的多个方法步骤。此外，本发明的另一个方面或其元素可以以装置的形式实现，所述装置包括存储器和至少一个处理器，所述至少一个处理器耦合到所述存储器并且可操作以执行说明的方法步骤。此外，本发明的另一个方面或其元素可以以用于执行在此描述的方法步骤的部件或其元素的形式实现；所述部件可以包括(i)硬件模块(多个)，(ii)软件模块(多个)，或(iii)硬件和

软件模块的组合；(i)-(iii)中的任何一个都实现在此说明的特定技术，并且所述软件模块被存储在有形的计算机可读存储介质(或多个此类介质)中。

[0009] 从以下将结合附图阅读的对本发明的示例性实施例的详细描述，本发明的这些和其它目标、特性以及优点将变得显而易见。

附图说明

[0010] 图 1 是示出根据本发明的一个实施例的云中的过载修复系统的实例架构的示意图；

[0011] 图 2 是示出根据本发明的一个实施例的目标云中的过载修复系统的实施架构的示意图；

[0012] 图 3 是示出根据本发明的一个实施例的用于 VM 静默的选择算法的示意图；

[0013] 图 4 是示出根据本发明的一个实施例的用于 VM 恢复的选择算法的示意图；

[0014] 图 5 是示出根据本发明的一个实施例的用于修复过度承诺的计算环境中的过载的技术的流程图；以及

[0015] 图 6 是其中可以实现本发明的至少一个实施例的示例性计算机系统的系统示意图。

具体实施方式

[0016] 如在此所描述的，本发明的一个方面包括修复过度承诺的计算环境中的过载。本发明的至少一个实施例包括提供一种机制，以便修复过载而无需假设始终具有可用于迁移的资源。如在此所详述的，使用工作价值概念比较 VM 的重要性，并将过载修复问题表述为可移除在线多背包问题(ROMKP)的变型。本发明的一个方面包括一种用于求解该优化问题的算法。

[0017] 本发明的一个实施例例如可以在大型商业云环境中实现。当不存在资源可用系统管理程序时，本发明的至少一个实施例包括临时放弃某些计算，即，使一个或多个被确定为不太重要的 VM 静默以便修复持续过载，从而使其余 VM 正常运行。当一组 VM 完成其作业并正常终止或释放资源之后，资源变得可用并且被静默的 VM 恢复以完成其作业。因此，如在此所详述的，本发明的一个方面包括确定应针对哪些 VM 采取哪些操作(迁移、静默或恢复)，以便在可能没有足够资源的过度订阅的计算系统中修复资源过载，同时增加(例如，最大化)运行中的服务的价值。

[0018] 此外，本发明的至少一个实施例包括综合对 VM 的静默、恢复和迁移操作的考虑，以便提供解决该问题变型的算法。相应地，所述算法结合对静默 / 迁移 / 恢复操作的成本的考虑，并且通过试探算法接近全局优化。

[0019] 图 1 是示出根据本发明的一个实施例的云中的过载修复系统的实例架构的示意图。举例来说，图 1 示出了互连的系统管理程序 106 和 110。将 VM (例如，VM102 和 104)的映像放置在跨所有系统管理程序共享的存储装置上。因此，可以在任何系统管理程序上启动或恢复 VM。安装在每个系统管理程序中的监视器(例如由组件 108 和 112 示出)定期测量该系统管理程序上的所有 VM 的资源使用，并向修复中心组件 114 报告资源测量。

[0020] 当系统管理程序中的监视器检测到系统管理程序上的过载时，修复中心知晓云中

的所有系统管理程序的资源使用，并做出要迁移或静默或恢复哪些 VM 的决策。此外，恢复中心向对应的系统管理程序发送命令以发出操作。通过图 1 中的虚线示出监视器和修复中心之间的消息传输。

[0021] 因为 VM 的资源使用随时间变化，并且不时发生 VM 供应和取消供应，所以云中的资源使用也有所变化。因此，修复中心定期检查是否有足够的资源变得可用以便恢复任何被静默的 VM。

[0022] 如在此所详述的，本发明的一个方面包括使用“工作价值”的概念来定义 VM 的重要性。工作价值如在此所使用的，旨在衡量 VM 被静默导致的 VM 上的潜在服务 / 任务损失和 / 或 VM 被静默对其它 VM 的执行的影响。在传统的 VM 放置技术中，并不测量这些价值作为效用，因为此类方法不会使 VM 静默。例如，工作价值包括被静默的 VM 在如果未被静默的情况下完成的预期工作，而这些技术中的效用没有考虑此方面。

[0023] 本发明的至少一个实施例的目标是使执行最少量工作价值的 VM 静默，并恢复执行更多工作价值的 VM。如在此所描述的，使用分析模型评估每个 VM 的工作价值，从而利用 VM 度量的现有分析(例如，季节性时间序列)计算 / 估计工作价值。此类计算可以结合有关被预测工作负载的模式、从过去或工作负载类型进行的生命周期预测、客户风格等知识。

[0024] 此外，本发明的至少一个实施例包括使用相关性模型评估每个 VM 的工作价值。这例如可以包括经由配置分析的静态相关性、动态相关性和 / 或使用其它试探法，例如 VM 价格、当前中央处理单元(CPU)工作等。

[0025] 工作价值的一个实例实施例是收益或价格。如果未使用价格作为工作价值，则可以通过服务类型指示工作价值。当提供关键服务(例如电子邮件和网络文件系统(NFS))的 VM 被静默时，可能导致重大损失。可以直接从有关 VM 服务的知识，或者通过利用分析模型(如果没有提供此类直接知识)来估计潜在损失。例如，如果 VM 为多个机器提供备份服务并且每周一次定期执行备份服务，则可以应用季节性时间序列分析以便检测此 VM 的服务的规律性，并预测该 VM 在特定时间的关键性。还可以预测工作负载模式，并使用它确定 VM 的工作在特定时间的重要性。

[0026] 除了使用有关服务的知识、分析模型和工作负载模式来估计工作价值之外，管理员可以指定 VM 在任何时间 t 的工作价值的计算。下面给出了计算工作价值的一个实例(需要正确设置参数 V1、V2、V3 以及函数

[0027]

V_pattern()）：

estimate_work_value(a_vm, t) {

if a_vm is known that it provides critical services all the time

work_value(a_vm, t) = V_1 for any time;

else if a_vm is estimated to be running regularly-scheduled critical services

work_value(a_vm, t)=V_2 for scheduled time;

work_value(a_vm, t)=V_3 for other time;

else if the workload pattern on a_vm is predicted

work_value(a_vm, t)=V_pattern (workload_pattern, t) // with the workload pattern known, we can get its work value at time t

else

work_value(a_vm, t)=average amount of load in a_vm in its past history

}

[0028] 除了被静默的 VM 的服务损失之外，被静默的 VM 还可能影响依赖于该被静默的 VM 的其它 VM。举例来说，WebSphere Application Server（WAS）中的股票交易服务依赖于 DB2 服务器，并且 WAS 和 DB2 服务器存在于两个单独的 VM 中。当 DB2VM 被静默时，WAS VM 不会实现任何工作价值。

[0029] 此外，本发明的至少一个实施例可以包括当选择某些 VM 以便静默、迁移或恢复时，在无相关性情况下优化运行中的 VM 的总工作价值。在无相关性情况下，VM 的工作价值与其它 VM 的工作价值无关。在此类情况下，问题是多背包问题（MKP）的变型，或者更具体地说，是可移除在线 MKP 的变型，因为要供应 VM，从系统管理程序中移除 VM，以及将 VM 从一个系统管理程序重新定位到另一个，并且 VM 具有变化的资源使用和工作价值。除了工作价值之外，在处理该问题时还考虑 VM 静默和迁移的成本。

[0030] 举例来说，考虑具有 m 个系统管理程序 H_1、H_2、…、H_m 的云计算环境。H_i 具有资源量 R_i（为使该实例简单起见，假设具有一种类型的资源；但是将理解，可以扩展本发明的至少一个实施例以支持多种类型的资源）。此外，假设存在 n 个 VM：VM_1、VM_2、…、VM_n。在此使用的变量被定义如下：

[0031] u_j VM_j 要完成的工作价值

[0032] r_j VM_j 使用的资源

[0033] R_i H_i 中的资源

[0034] $x_{i,j}$ 1：VM_j 在 H_i 上运行；0：其他情况

[0035] cq_j 静默 VM_j 的成本

[0036] cr_j 恢复 VM_j 的成本

[0037] cm_j 迁移 VM_j 的成本

[0038] μ 过载检测的阈值百分比

[0039] u_j、r_j 和 $x_{i,j}$ 随时间变化；因此，它们被表示为时间 t 的函数：$u_j(t)$、$r_j(t)$ 和 $x_{i,j}(t)$。形式上，给出以下表示：

[0040]

$$\begin{cases} \sum_{i=1}^{m} x_{i,j}(t) \le 1, \text{对于} 1 \le j \le n \\ x_{i,j}(t) \in \{0,1\}, \text{对于所有} 1 \le j \le n, 1 \le i \le m \end{cases}$$

[0041] 并且系统正在经历资源过载（当修复过载时）或未经历资源过载（当恢复被静默的 VM 时）。即，

[0042]

$$\exists i, \sum_{j=1}^{n} r_j(t) * x_{i,j}(t) \ge R_i * \mu, 1 \le i \le m, \text{或} \sum_{j=1}^{n} r_j(t) * x_{i,j}(t) \le R_i * \mu, \text{对于所有} 1 \le i \le m \text{。}$$

[0043] 本发明的至少一个实施例尝试通过选择受以下约束的 $x_{i,j}(t+1)$ 来获得

[0044]

$$\max\left(\sum_{i=1}^{m} \sum_{j=1}^{n} \left(u_j(t+1) * x_{i,j}(t+1) \right) \right)$$

[0045]

$$\begin{cases} \sum_{j=1}^{n} r_j(t+1) * x_{i,j}(t+1) \le R_i * \mu, \text{对于} 1 \le i \le m \\ \sum_{i=1}^{m} x_{i,j}(t+1) \le 1, \text{对于} 1 \le j \le n \\ x_{i,j}(t+1) \in \{0,1\}, \text{对于所有} 1 \le j \le n, 1 \le i \le m \end{cases}$$

[0046] 以及

[0047]

$$\begin{cases} quiesce\cos t = \sum_{k=1}^{|YQ|} \left(cq_{yq_k} * \sum_{i=1}^{m} x_{i,yq_k}(t+1) \right) \le thrd_q, yq \in YQ \\ resume\cos t = \sum_{k=1}^{|YR|} \left(cr_{yr_k} * \sum_{i=1}^{m} x_{i,yr_k}(t+1) \right) \le thrd_r, yr \in YR \\ migration\cos t = \sum_{k=1}^{|YM|} \left(cm_{ym_k} * \sum_{i=1}^{m} x_{i,ym_k}(t+1) \right) \le thrd_m, ym \in YM \\ YQ = \{ j \mid VM_j \text{ is to be quiesced} \} \\ YR = \{ j \mid VM_j \text{ is to be resumed} \} \\ YM = \{ j \mid VM_j \text{ is to be migrated} \} \end{cases}$$

[0048] $thrd_q$、$thrd_r$ 和 $thrd_m$ 分别是静默、恢复和迁移成本的阈值。这些值可以由系统确

定或者由管理员指定。

[0049] VM_j 要被静默、恢复、迁移还是保持不变将在形式上由 $x_{i,j}(t)$ 和 $x_{i,j}(t+1)$ 之间的关系表示，如下所示：

[0050] $$A=\sum_{i=1}^{m}\left[x_{i,j}(t+1)-x_{i,j}(t)\right];\quad B=\sum_{i=1}^{m}\left[x_{i,j}(t+1)-x_{i,j}(t)\right]^2$$

[0051] (1) VM_j 要被静默⇔ A=-1，B=1

[0052] (2) VM_j 要被恢复⇔ A=1，B=1

[0053] (3) VM_j 要被迁移⇔ A=0，B=2

[0054] (4) VM_j 要被保持不变⇔ A=0，B=0

[0055] ROMKP 问题的该变型是 NP 困难(NP hard)问题。本发明的至少一个实施例包括一种设计的 ROWM 近似算法以便解决该问题。所述算法中存在两个部分：过载修复部分，用于当系统管理程序发生过载时，确定要静默 / 迁移哪些 VM；以及定期恢复部分，用于当资源变得充足时，确定要恢复哪些被静默的 VM。

[0056] 当过载(即，资源使用超过阈值 thrd1)时，标识过载的系统管理程序上的一组 VM，必须迁移或静默这些 VM 以便修复过载。然后，执行测试以便确定是否可以通过针对 MKP 应用试探算法，将 C 中的所有 VM (即，消耗过载的系统管理程序上大部分资源以至其余 VM 消耗的资源少于阈值 thrd2 的 VM) 迁移到其它系统管理程序。当没有足够的资源时，所述测试将失败。因此，计划静默整个云中具有最小的工作价值 - 资源使用比率的 VM，并且再次执行测试。继续该过程直到测试通过。

[0057] 随后，检查 Q 中被计划静默的 VM 以便选择不必要地被计划静默的那些 VM。在这些步骤之后，获得最终迁移和静默计划。定期恢复部分的步骤类似于上面对过载修复部分的描述，只是执行测试以便确定是否可以恢复 C 中所有被静默的 VM。

[0058] 该算法具有时间复杂度 O(nlogn+nkm)，其中 n 是云中的 VM 数，k 是要被迁移 / 恢复的 VM 数，m 是系统管理程序数。它是多项式时间开销，并且例如可以针对包含数百个系统管理程序和数千个 VM 的云实现。

[0059] 举例来说，本发明的至少一个实施例可以在大型云平台上被实现为 MAPE (监视 - 分析 - 计划 - 执行) 循环。图 2 是示出根据本发明的一个实施例的目标云中的过载修复系统的实施架构的示意图。图 2 中的带阴影组件是那些要被添加到目标云以支持过载修复的组件。图 2 中的不带阴影组件是云本身的组件。

[0060] 云管理器 214 是负责 VM 202 的 VM 供应、取消供应、迁移和其它管理操作的软件包。数据仓库 212 存储由系统管理程序 204 中的监视器 206 定期报告的测量数据。客户可以通过门户 216 向云管理器发出命令以便执行所需的操作。在门户和云管理器之间利用 REST (表现状态传输) 应用编程接口 (API) 以便请求云管理器中的特定操作。此外，可以在云管理器和门户之间实施消息队列组件 218。

[0061] 与图 1 中示出的基本过载修复相比，图 2 中示出的系统包括修复数据库 (DB) 208 以支持持久性。所述数据库存储由系统管理程序监视器生成的事件 (例如，“资源过载”事件、“已修复过载”事件等)、云管理器执行的操作的状态 (例如，“迁移成功”、“恢复 VM 失败”等)，以及所有被管理 VM 的当前状态 (例如，“正常运行”、“过载”、“静默”等)。修复中心 210 是在单独计算机上运行的过程，其定期查询修复 DB 以便处理所述 DB 接收的任何新事件。当

处理这些事件时，修复中心从数据仓库212取回执行数据（当修复算法中需要这些数据时）。当进行有关要迁移、静默和/或恢复哪些VM的决策时，向云管理器发送命令以执行决策。

[0062] 图3是示出根据本发明的一个实施例的用于VM静默的选择算法的流程图。在步骤302，Q={计划要通过该算法静默的VM；初始为空}。在步骤304，C={消耗过载的系统管理程序上大部分资源以至其余VM消耗少于阈值（例如，85%）的物理资源的VM}。步骤306包括针对所有系统管理程序上的所有运行中的VM，根据其U/R比率按升序排序。步骤308包括进行测试以便确定是否可以在不会导致资源使用超过设置阈值（例如，85%）并具有可接受的迁移成本的情况下，将C中的VM迁移到其它系统管理程序（Q中的VM被认为已静默）。系统管理员可以根据其经验和专业知识指定可接受的迁移成本的准则；例如，准则为在某一时间段内少于80%的网络带宽。此外，在测试期间应用最佳拟合MKP试探算法。如果测试成功（如在步骤310中所确定），则执行工作并且所述算法继续到步骤316；否则，所述算法继续到步骤312。

[0063] 在步骤312，Q=Q+{具有次最小的U/R值并具有可接受的静默成本的VM}。步骤314包括确定是否修复了过载。如果是，则所述算法继续到步骤316；否则，所述算法返回到步骤308。因此，将迁移计划成功地创建为M，并且步骤316包括从Q中移除不必要地通过步骤312添加到Q中的VM。<M,Q>是最终结果。

[0064] 图4是示出根据本发明的一个实施例的用于VM恢复的选择算法的示意图。在步骤402，Q={计划要通过该算法静默的VM；初始为空}。在步骤404，C={被静默的VM}。步骤406包括针对所有系统管理程序上的所有运行中的VM和被静默的VM，根据其U/R比率按升序排序。步骤408包括进行测试以便确定是否可以在不会导致资源使用超过阈值（例如，85%）并具有可接受的恢复成本的情况下，在系统管理程序上恢复C中的VM（Q中的VM被认为已静默）。此外，在所述测试期间应用最佳拟合MKP试探算法。

[0065] 步骤410确定所述测试是否成功。如果成功，则执行工作并且所述算法继续到步骤414。否则，所述算法继续到步骤412。在步骤412，Q=Q+{具有次最小的U/R值并具有可接受的静默成本的VM}。随后，所述算法返回到步骤408。

[0066] 将恢复计划成功地创建为M，并且步骤414包括从Q中移除不必要地通过步骤412添加到Q中的VM。<M,Q>是最终结果。

[0067] 图5是示出根据本发明的一个实施例的用于修复过度承诺的计算环境中的过载的技术的流程图。步骤502包括测量计算环境中的至少一个系统管理程序中的每个系统管理程序上的多个虚拟机中的每个虚拟机的资源使用。

[0068] 在检测到所述至少一个系统管理程序中的一个系统管理程序上的资源过载时，执行步骤504和506。步骤504包括确定要针对该系统管理程序上的所述多个虚拟机中的至少一个虚拟机采取的至少一个操作，以便在增加或最大化运行中的虚拟机的价值的同时修复资源过载。如在此所详述的，所述操作可以包括迁移、静默或恢复中的至少一个。

[0069] 确定要采取的至少一个操作可以包括计算该系统管理程序上的每个虚拟机的工作价值，其中工作价值可以包括以下项中的至少一个：由该虚拟机被静默导致的该虚拟机上的潜在服务损失的测量，以及该虚拟机被静默对至少一个其它虚拟机的执行的影响的测量。每个虚拟机的工作价值可以包括至少一个因素的组合，所述至少一个因素包括虚拟机生命周期、价格、收益、服务水平协议、中央处理单元利用率、虚拟机相关性以及工作负载预

测。

[0070] 此外，确定要采取的至少一个操作可以包括标识要静默的具有最低工作价值的至少一个虚拟机和 / 或要恢复以便优化的具有较高工作价值的至少一个虚拟机。此外，确定要采取的至少一个操作可以包括持续测量和利用每个虚拟机的工作价值和资源消耗。本发明的至少一个实施例还可以包括使用工作价值与资源消耗的比率判定是否要针对虚拟机采取迁移、静默或恢复操作中的任何一个。

[0071] 步骤 506 包括向该系统管理程序发送命令以发出所述至少一个操作。此外，图 5 中示出的技术可以包括定期确定是否有足够的资源变得可用以便恢复被静默的虚拟机。

[0072] 还如在此所详述的，本发明的至少一个实施例包括一种用于虚拟化环境的自动过载修复方法。此类方法包括通过从系统内的多个虚拟机选择所述虚拟机的要迁移的子集以及所述虚拟机的要静默的子集而从系统资源过载情况中恢复。此外，本发明的此类实施例可以包括定期选择所述虚拟机的要迁移的子集以及所述虚拟机的要恢复的子集，以便增加运行中的虚拟机数量并增加运行中的虚拟机实现的工作价值。如在此所详述的，迁移、静默和恢复选择均考虑每个虚拟机所贡献的工作价值度量。所述工作价值度量组合一个或多个因素，所述一个或多个因素包括虚拟机生命周期、价格、收益、服务水平协议、中央处理单元利用率、虚拟机相关性以及工作负载预测。

[0073] 此外，在本发明的至少一个实施例中，问题被表述为可移除在线多背包问题 (ROMKP) 的变型。此类表述的问题包括 ：除新物体的到达以外，考虑随时间变化的物体工作价值和资源使用 ；允许背包操作，所述背包操作包括从背包移除一个或多个物体、将一个或多个物体从一个背包移动到另一个背包，以及将一个或多个物体放回背包。此类问题还包括在决策中考虑所述背包操作的成本。

[0074] 对 ROMKP 问题的所述变型的解可以包括 ：持续测量和利用每个虚拟机的所述工作价值和资源使用 ；使用工作价值与资源使用的比率作为决策准则，以便针对一个或多个虚拟机选择迁移、静默或恢复操作 ；以及计算每个迁移、静默和 / 或恢复操作的成本，并使用所述成本作为问题公式化的约束。此外，在本发明的至少一个实施例中，虚拟机的所述工作价值被定义为 1/ (资源使用) (如果该虚拟机在给定最近时间段 P 内未被静默)以及无穷大 (如果该虚拟机在所述时间段 P 内被静默)。

[0075] 如在此所描述的，图 5 中示出的技术还可以包括提供一种系统，其中所述系统包括不同软件模块，所述不同软件模块的每一个都包含在有形的计算机可读可记录存储介质中。例如，所有模块 (或其任何子集)可以在相同介质中，或者每一个可以在不同介质中。所述模块可以包括附图中示出的任何或全部组件。在本发明的一个方面，所述模块例如可以在硬件处理器上运行。然后可以使用所述系统的所述不同软件模块 (如上所述，在硬件处理器上执行)执行所述方法步骤。此外，一种计算机程序产品可以包括有形的计算机可读可记录存储介质，其具有适合于被执行的代码以便执行在此描述的至少一个方法步骤，包括为所述系统供应所述不同软件模块。

[0076] 此外，图 5 中示出的技术能够通过可以包括计算机可用程序代码的计算机程序产品来实现，所述计算机可用程序代码被存储在数据处理系统内的计算机可读存储介质中，并且其中所述计算机可用程序代码通过网络从远程数据处理系统下载。此外，在本发明的一个方面，所述计算机程序产品可以包括被存储在服务器数据处理系统内的计算机可读存

储介质中的计算机可用程序代码，并且其中所述计算机可用程序代码通过网络下载到远程数据处理系统，以便在计算机可读存储介质中与所述远程系统一起使用。

[0077] 如本领域的技术人员将理解的，本发明的各方面可以体现为系统、方法或计算机程序产品。因此，本发明的各方面可以采取完全硬件实施例、完全软件实施例（包括固件、驻留软件、微代码等）或组合了在此通常可以被称为“电路”、“模块”或“系统”的软件和硬件方面的实施例的形式。此外，本发明的各方面可以采取体现在计算机可读介质（在介质中包含计算机可读程序代码）中的计算机程序产品的形式。

[0078] 本发明的一个方面或其元素可以以装置的形式实现，所述装置包括存储器和至少一个处理器，所述至少一个处理器耦合到所述存储器并且可操作以执行示例性方法步骤。

[0079] 此外，本发明的一个方面可以使用在通用计算机或工作站上运行的软件。参考图6，此类实施方式例如可以采用处理器602、存储器604和输入／输出接口（例如，由显示器606和键盘608形成）。术语“处理器”如在此所使用的，旨在包括任何处理设备，例如包括CPU（中央处理单元）和／或其它形式处理电路的处理设备。此外，术语“处理器”可以指多个单独的处理器。术语“存储器”旨在包括与处理器或CPU关联的存储器，例如RAM（随机存取存储器）、ROM（只读存储器）、固定存储器设备（例如，硬盘驱动器）、可移动存储器设备（例如，软盘）、闪存等。此外，词组“输入／输出接口”如在此所使用的，旨在包括例如用于将数据输入到所述处理单元中的机构（例如，鼠标），以及用于提供与所述处理单元关联的结果的机构（例如，打印机）。作为数据处理单元612的一部分，处理器602、存储器604和输入／输出接口（例如显示器606和键盘608）例如可以通过总线610互连。还可以例如通过总线610为网络接口614（例如网卡，可以提供它以便与计算机网络连接）以及介质接口616（例如软盘或CD-ROM驱动器，可以提供它以便与介质618连接）提供适合的互连。

[0080] 因此，如在此所描述的，包括用于执行本发明的所述方法的指令或代码的计算机软件可以被存储在关联的存储设备（例如，ROM、固定或可移动存储器）中，并且当准备使用时，被部分或全部加载（例如，加载到RAM中）并由CPU实现。此类软件可以包括但不限于固件、驻留软件、微代码等。

[0081] 适合于存储和／或执行程序代码的数据处理系统将包括至少一个通过系统总线610直接或间接连接到存储器元件604的处理器602。所述存储器元件可以包括在程序代码的实际执行期间采用的本地存储器、大容量存储装置以及提供至少某些程序代码的临时存储以减少必须在执行期间从大容量存储装置检索代码的次数的高速缓冲存储器。

[0082] 输入／输出或I/O设备（包括但不限于键盘608、显示器606、指点设备等）可以直接（例如通过总线610）或通过中间I/O控制器（为清楚起见而被省略）与所述系统相连。

[0083] 网络适配器（例如网络接口614）也可以被连接到所述系统以使所述数据处理系统能够通过中间专用或公共网络变得与其它数据处理系统或远程打印机或存储设备相连。调制解调器、电缆调制解调器和以太网卡只是当前可用的网络适配器类型中的少数几种。

[0084] 如在此（包括权利要求）所使用的，“服务器”包括运行服务器程序的物理数据处理系统（例如，如图6中示出的系统612）。将理解，此类物理服务器可以包括也可以不包括显示器和键盘。

[0085] 如所说明的，本发明的各方面可以采取包含在计算机可读介质中的计算机程序产品的形式，所述计算机可读介质具有包含在其中的计算机可读程序代码。此外，可以使用一

个或多个计算机可读介质的任意组合。所述计算机可读介质可以是计算机可读信号介质或计算机可读存储介质。计算机可读存储介质例如可以是(但不限于)电、磁、光、电磁、红外线或半导体系统、装置或设备或上述任意适合的组合。所述计算机可读存储介质的更具体的实例(非穷举列表)将包括以下项:具有一条或多条线的电连接、便携式计算机软盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦写可编程只读存储器(EPROM 或闪存)、光纤、便携式光盘只读存储器(CD-ROM)、光存储设备、磁存储设备或上述任意适合的组合。在本文档的上下文中,计算机可读存储介质可以是任何能够包含或存储由指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合的程序的有形介质。

[0086] 计算机可读信号介质可以包括其中包含计算机可读程序代码(例如,在基带中或作为载波的一部分)的传播数据信号。此类传播信号可以采取各种形式中的任一种,包括但不限于电磁、光或其中任意适合的组合。计算机可读信号介质可以是任何不属于计算机可读存储介质并且能够传送、传播或传输由指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合的程序的计算机可读介质。

[0087] 可以使用适当的介质(包括但不限于无线、线缆、光缆、RF 等或上述任意适合的组合)来传输包含在计算机可读介质中的程序代码。

[0088] 用于执行本发明的各方面的操作的计算机程序代码可以使用包含至少一种编程语言的任意组合来编写,所述编程语言包括诸如 Java、Smalltalk、C++ 之类的面向对象的编程语言以及诸如“C”编程语言或类似的编程语言之类的常规过程编程语言。所述程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为独立的软件包、部分地在用户计算机上并部分地在远程计算机上执行,或者完全地在远程计算机或服务器上执行。在后者的情况中,所述远程计算机可以通过包括局域网(LAN)或广域网(WAN)的任何类型网络与用户的计算机相连,或者可以与外部计算机进行连接(例如,使用因特网服务提供商通过因特网连接)。

[0089] 在此参考根据本发明的实施例的方法、装置(系统)和计算机程序产品的流程图和/或方块图对本发明的各方面进行描述。将理解,所述流程图和/或方块图的每个方块以及所述流程图和/或方块图中的方块的组合可以由计算机程序指令来实现。这些计算机程序指令可以被提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器以产生机器,以便通过所述计算机或其它可编程数据处理装置的处理器执行的所述指令产生用于实现在一个或多个流程图和/或方块图方块中指定的功能/操作的装置。

[0090] 这些计算机程序指令也可以被存储在能够引导计算机、其它可编程数据处理装置或其它设备以特定方式执行功能的计算机可读介质中,以便存储在所述计算机可读介质中的所述指令产生一件包括实现在一个或多个流程图和/或方块图方块中指定的功能/操作的指令的制品。因此,本发明的一个方面包括有形地包含计算机可读指令的制品,当执行所述计算机可读指令时,导致计算机执行在此描述的多个方法步骤。

[0091] 所述计算机程序指令还可被加载到计算机、其它可编程数据处理装置或其它设备,以导致在所述计算机、其它可编程装置或其它设备上执行一系列操作步骤以产生计算机实现的过程,从而在所述计算机或其它可编程装置上执行的所述指令提供用于实现在一个或多个流程图和/或方块图方块中指定的功能/操作的过程。

[0092] 附图中的流程图和方块图示出了根据本发明的各种实施例的系统、方法和计算机

程序产品的可能实施方式的架构、功能和操作。在此方面，所述流程图或方块图中的每个方块都可以表示模块、组件、段或代码部分，它们包括用于实现指定的逻辑功能（多个）的至少一个可执行指令。还应指出，在某些备选实施方式中，在方块中说明的功能可以不按图中说明的顺序发生。例如，示为连续的两个方块可以实际上被基本同时地执行，或者某些时候，取决于所涉及的功能，可以以相反的顺序执行所述方块。还将指出，所述方块图和／或流程图的每个方块以及所述方块图和／或流程图中的方块的组合可以由执行指定功能或操作的基于专用硬件的系统或专用硬件和计算机指令的组合来实现。

[0093] 应指出，在此描述的任何方法都可以包括提供一种系统的附加步骤，所述系统包括包含在计算机可读存储介质中的不同软件模块；所述模块例如可以包括在此详述的任何或全部组件。然后可以使用所述系统的所述不同软件模块和／或子模块（如上所述，在硬件处理器 602 上执行）执行所述方法步骤。此外，计算机程序产品可以包括计算机可读存储介质，其具有适合于被执行的代码以便执行在此描述的至少一个方法步骤，包括为所述系统供应所述不同软件模块。

[0094] 在任何情况下，应理解，在此示出的组件可以以各种形式的硬件、软件或它们的组合来实现；例如，专用集成电路（多个）（ASIC）、功能电路、具有关联存储器的经过适当编程的通用数字计算机等。给予了在此提供的本发明的教导后，本领域的技术人员将能够构想本发明的所述组件的其它实施方式。

[0095] 在此使用的术语只是为了描述特定的实施例并且并非旨在作为本发明的限制。如在此所使用的，单数形式“一”、“一个”和“该”旨在同样包括复数形式，除非上下文明确地另有所指。还将理解，当在此说明书中使用时，术语“包括”和／或“包含”指定了声明的特性、整数、步骤、操作、元素和／或组件的存在，但是并不排除其它特性、整数、步骤、操作、元素、组件和／或其组的存在或增加。

[0096] 以下的权利要求中的对应结构、材料、操作以及所有功能性限定的装置或步骤的等同替换，旨在包括任何用于与在权利要求中具体指出的其它单元相组合地执行该功能的结构、材料或操作。所给出的对本发明的描述其目的在于示意和描述，并非是穷尽性的，也并非是要把本发明限定到所表述的形式。对于所属技术领域的普通技术人员来说，在不偏离本发明范围和精神的情况下，显然可以作出许多修改和变型。对实施例的选择和说明，是为了最好地解释本发明的原理和实际应用，使所属技术领域的普通技术人员能够明了，本发明可以有适合所要的特定用途的具有各种改变的各种实施方式。

[0097] 本发明的至少一个方面可以提供有益的效果，例如修复系统管理程序过载而无需假设始终具有可用于迁移的资源。

[0098] 出于示例目的给出了对本发明的各种实施例的描述，但所述描述并非旨在是穷举的或是限于所公开的实施例。在不偏离所述实施例的范围和精神的情况下，对于本领域的技术人员来说许多修改和变化都将是显而易见的。在此使用的术语的选择是为了最佳地解释实施例的原理、实际应用或对市场中的技术的技术改进，或者使本领域的其它技术人员能够理解在此公开的实施例。

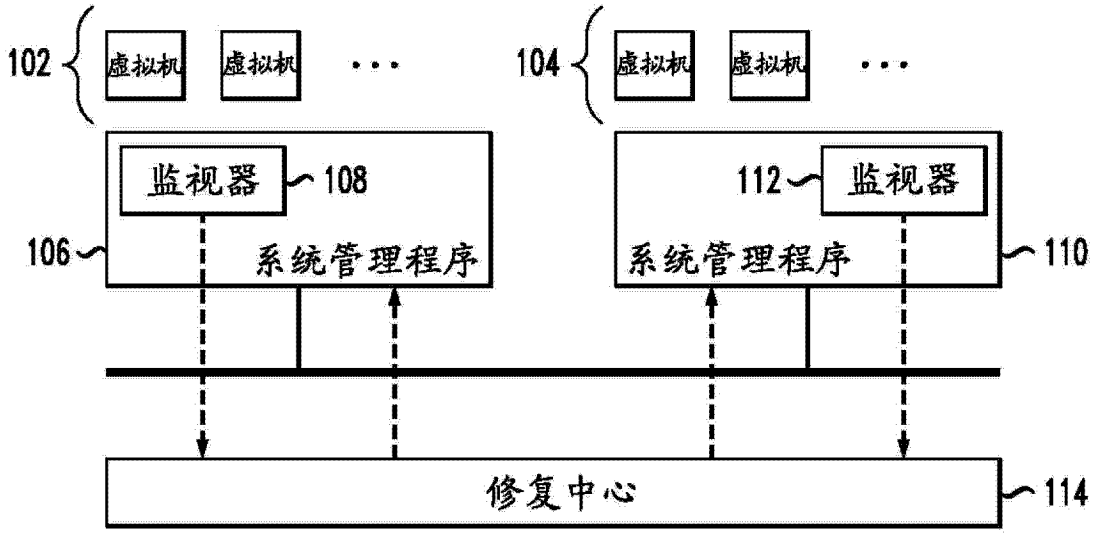

102
虚拟机
虚拟机
104
虚拟机
虚拟机
监视器
108
106
系统管理程序
112
监视器
系统管理程序
110
修复中心
114

图 1

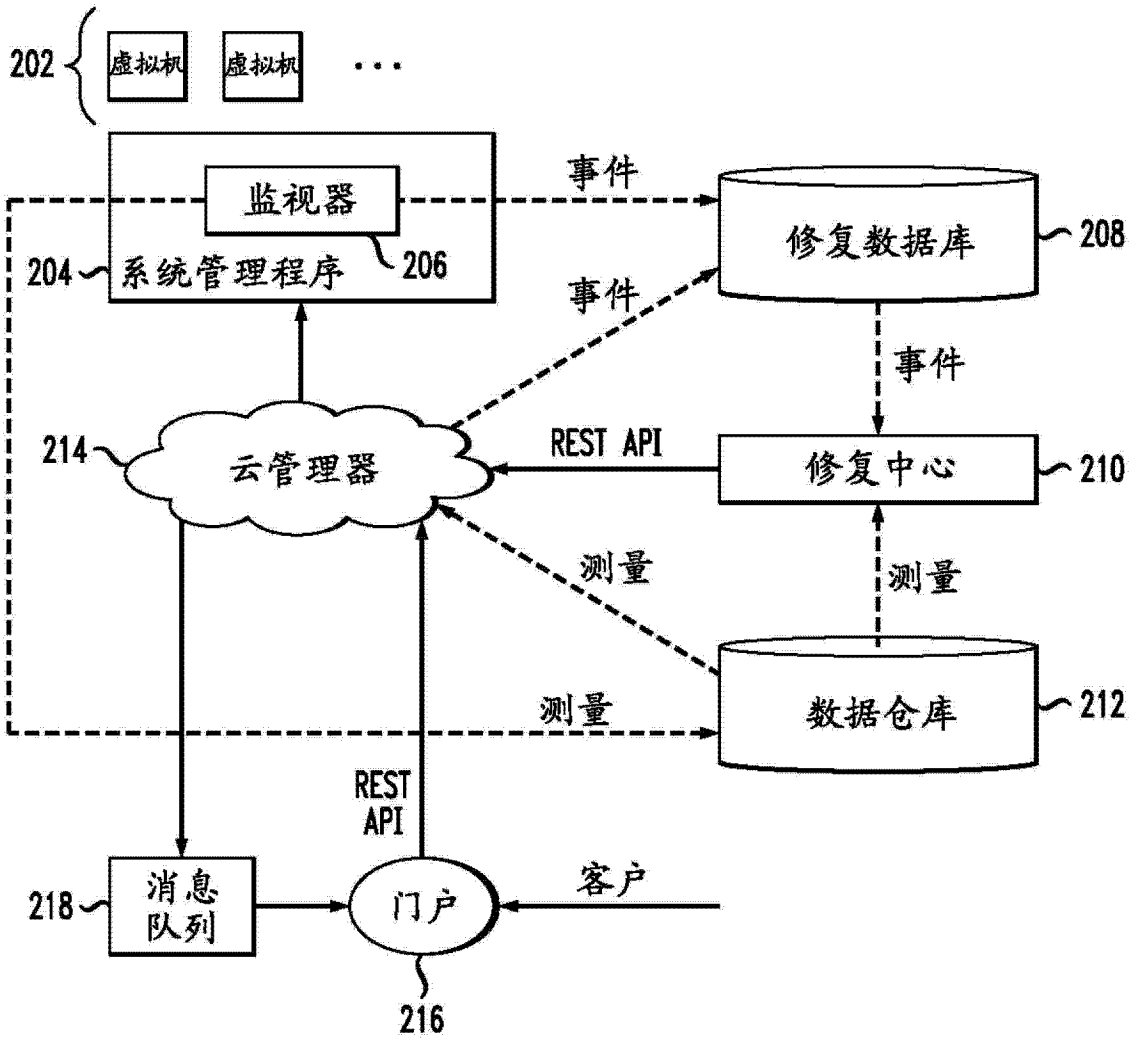

202
虚拟机
虚拟机
监视器
206
204
系统管理程序
事件
修复数据库
208
事件
事件
214
云管理器
REST API
修复中心
210
测量
测量
测量
数据仓库
212
REST
API
218
消息
队列
门户
客户
216

图 2

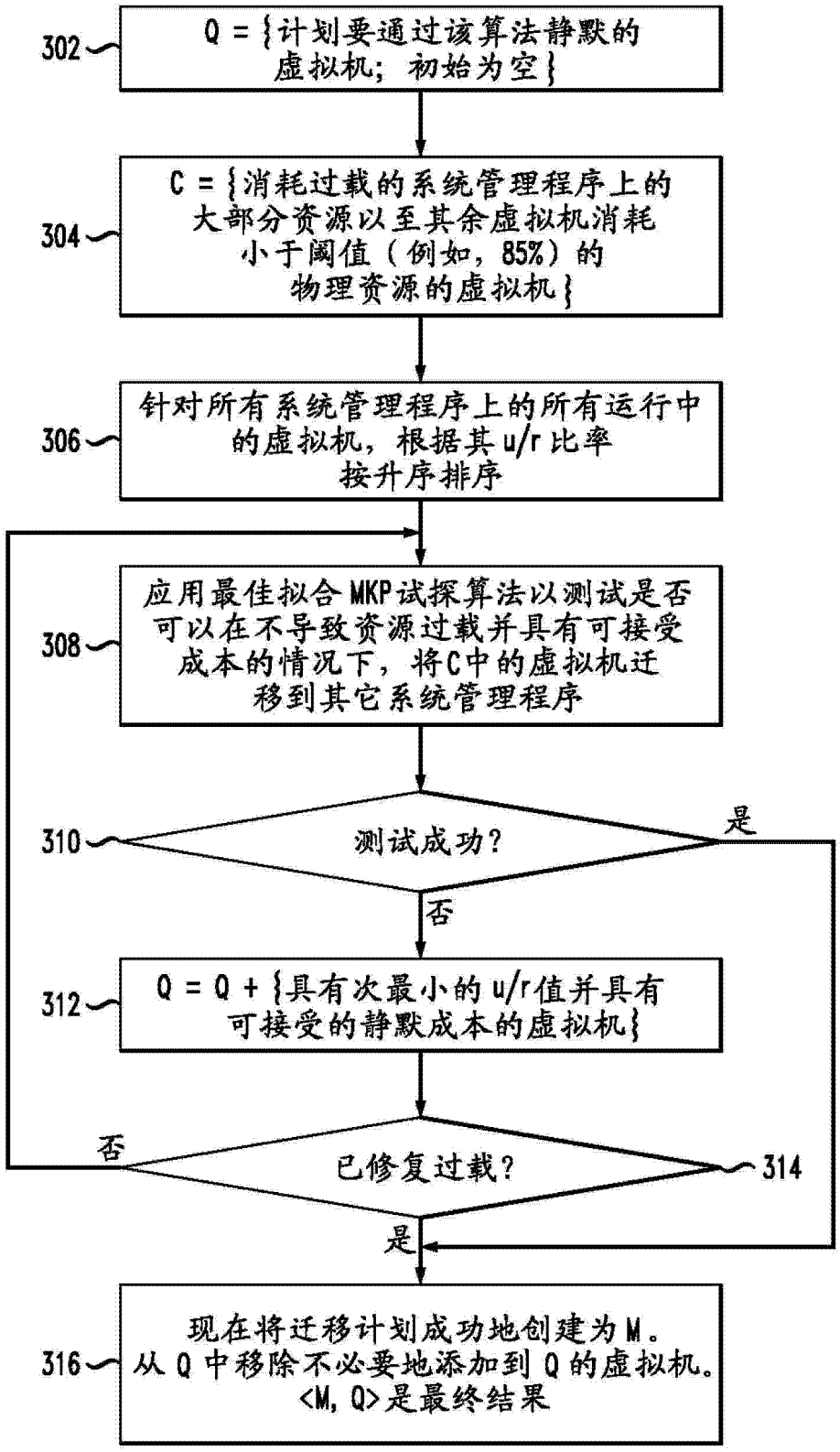

302
Q = {计划要通过该算法静默的虚拟机；初始为空}
304
C = {消耗过载的系统管理程序上的大部分资源以至其余虚拟机消耗小于阈值（例如，85%）的物理资源的虚拟机}
306
针对所有系统管理程序上的所有运行中的虚拟机，根据其 u/r 比率按升序排序
308
应用最佳拟合 MKP 试探算法以测试是否可以在不导致资源过载并具有可接受成本的情况下，将 C 中的虚拟机迁移到其它系统管理程序
310
测试成功？
是
否
312
Q = Q + {具有次最小的 u/r 值并具有可接受的静默成本的虚拟机}
否
已修复过载？
314
是
316
现在将迁移计划成功地创建为 M。从 Q 中移除不必要地添加到 Q 的虚拟机。<M, Q> 是最终结果

图 3

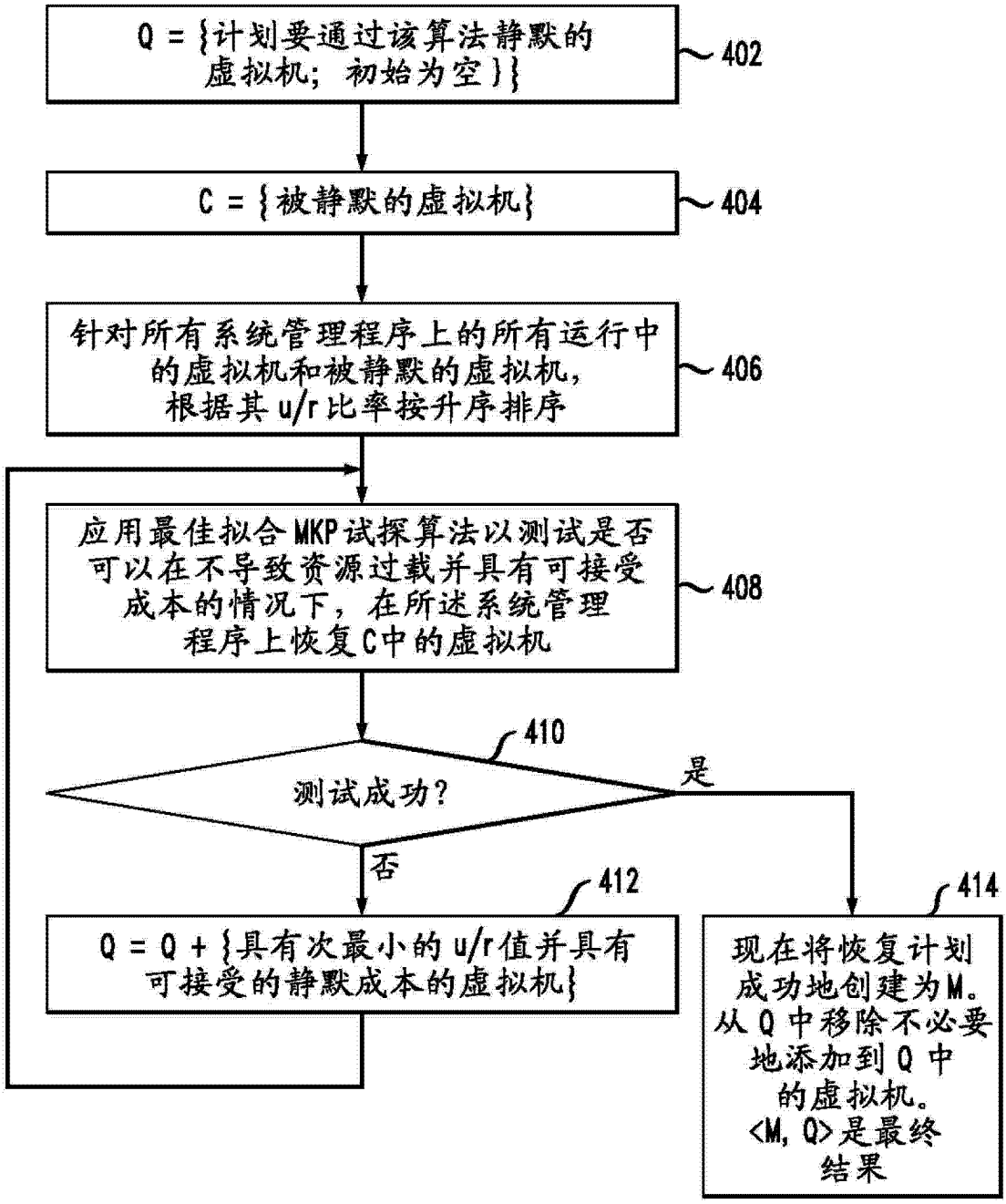

Q = {计划要通过该算法静默的虚拟机；初始为空}
402
C = {被静默的虚拟机}
404
针对所有系统管理程序上的所有运行中的虚拟机和被静默的虚拟机，根据其 u/r 比率按升序排序
406
应用最佳拟合 MKP 试探算法以测试是否可以在不导致资源过载并具有可接受成本的情况下，在所述系统管理程序上恢复 C 中的虚拟机
408
410
测试成功？
是
否
412
Q = Q + {具有次最小的 u/r 值并具有可接受的静默成本的虚拟机}
414
现在将恢复计划成功地创建为 M。从 Q 中移除不必要地添加到 Q 中的虚拟机。<M, Q> 是最终结果

图 4

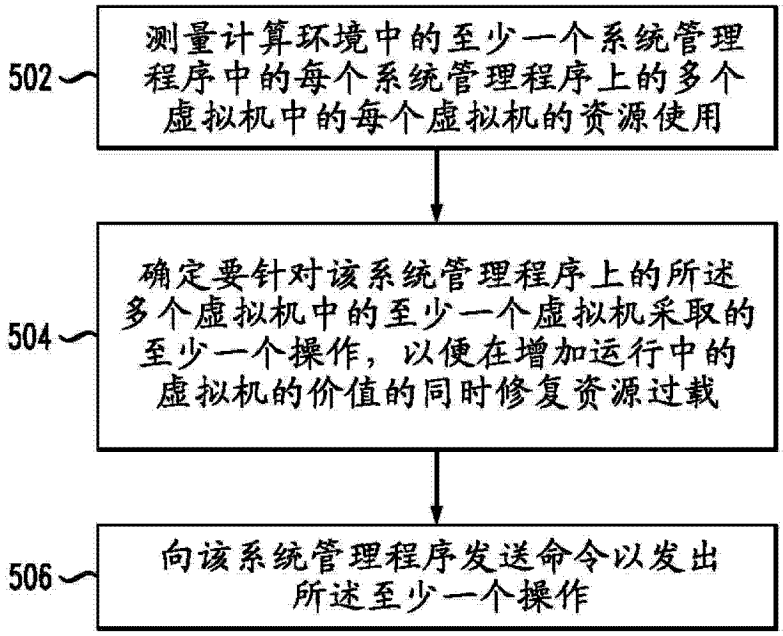

502
测量计算环境中的至少一个系统管理程序中的每个系统管理程序上的多个虚拟机中的每个虚拟机的资源使用
504
确定要针对该系统管理程序上的所述多个虚拟机中的至少一个虚拟机采取的至少一个操作，以便在增加运行中的虚拟机的价值的同时修复资源过载
506
向该系统管理程序发送命令以发出所述至少一个操作

图 5

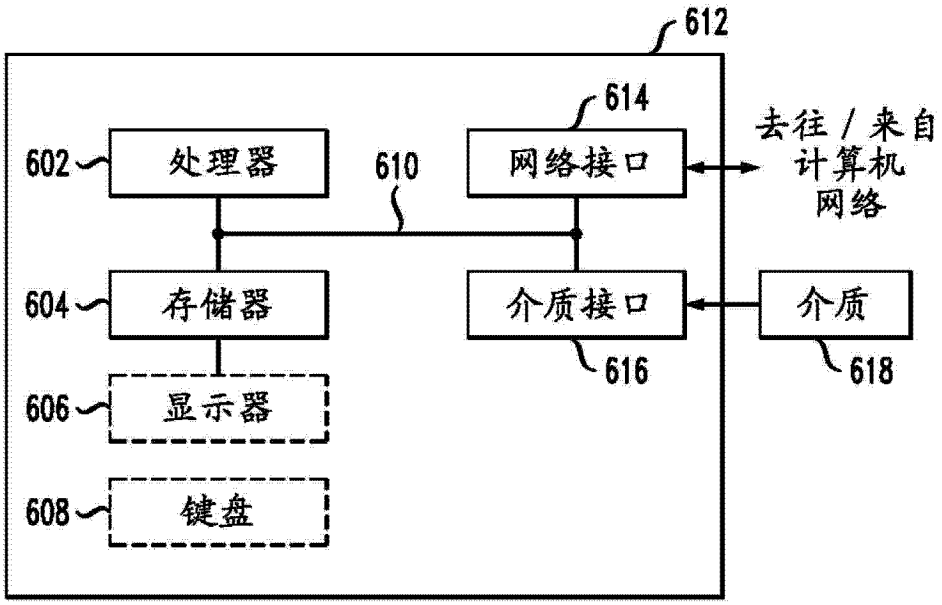

612
614
602
处理器
610
网络接口
去往 / 来自计算机网络
604
存储器
介质接口
介质
616
618
606
显示器
608
键盘

图 6