



(12)发明专利申请

(10)申请公布号 CN 108737136 A

(43)申请公布日 2018.11.02

(21)申请号 201710253744.8

(22)申请日 2017.04.18

(71)申请人 微软技术许可有限责任公司

地址 美国华盛顿州

(72)发明人 Z·拉法洛维奇 A·K·齐哈布拉

T·莫希布罗达 E·E·格瑞弗

A·辛格 H·库普塔

B·R·米什拉

(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 胡利鸣

(51)Int.Cl.

H04L 12/24(2006.01)

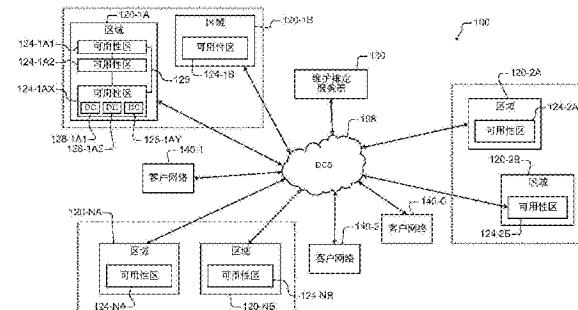
权利要求书3页 说明书11页 附图13页

(54)发明名称

将新虚拟机和容器分配给云网络中的服务器的系统和方法

(57)摘要

一种用于将资源分配给数据中心中的服务器的系统包括被存储在存储器中并由处理器执行的分配应用，并且所述分配应用被配置成将资源分配给数据中心中的服务器。所述资源包括虚拟机和容器实例中的至少一个。分配应用接收维护状态信息，所述维护状态信息标识多个维护波中的哪个维护波正在执行更新。分配应用基于维护状态信息调整经更新的那些服务器和未经更新的那些服务器之间的资源的分配。



1. 一种用于将资源分配给云网络中的服务器的系统,包括:
处理器;
存储器;
分配应用,所述分配应用被存储在所述存储器中并由所述处理器执行并被配置为:
将资源分配给所述数据中心中的所述服务器,其中所述资源包括虚拟机和容器实例中的至少一个;
接收维护状态信息,所述维护状态信息标识多个维护波中的哪个维护波正在执行更新;以及
基于所述维护状态信息调整经更新的那些服务器和未经更新的那些服务器之间的所述资源的分配。
2. 如权利要求1所述的系统,其特征在于,所述多个维护波中的每一个包括第一周期和第二周期,在所述第一周期期间对所述资源的维护被选择性地远程排定,在所述第二周期期间如果维护没有在所述第一周期期间被尝试,则由所述数据中心排定维护。
3. 如权利要求1所述的系统,其特征在于,所述分配应用被进一步配置成在所述多个维护波中的第一个维护波期间将所述资源分配给所述未经更新的那些服务器。
4. 如权利要求3所述的系统,其特征在于,所述分配应用被进一步配置成在所述多个维护波中的第一个维护波之后的所述多个维护波中的后续那些维护波期间将所述资源分配给所述经更新的那些服务器。
5. 如权利要求1所述的系统,其特征在于,进一步包括维护服务器,所述维护服务器包括被配置为生成所述维护状态信息的维护应用。
6. 如权利要求5所述的系统,其特征在于,所述维护应用被进一步配置为在所述多个维护波期间更新主存所述资源的所述服务器。
7. 一种用于将资源分配给云网络中的服务器的方法,包括:
将资源分配给所述数据中心中的所述服务器,所述资源包括虚拟机和容器实例中的至少一个;
接收维护状态信息,所述维护状态信息标识多个维护波中的哪个维护波正在执行更新;以及
基于所述维护状态信息调整经更新的那些服务器和未经更新的那些服务器之间的所述资源的分配。
8. 如权利要求7所述的方法,其特征在于,所述多个维护波中的每一个包括第一周期和第二周期,在所述第一周期期间对所述资源的维护能被远程排定,在所述第二周期期间如果维护没有在所述第一周期期间被尝试,则由所述数据中心排定维护。
9. 如权利要求7所述的方法,其特征在于,进一步包括在所述多个维护波中的第一个维护波期间将所述资源分配给所述未经更新的那些服务器。
10. 如权利要求9所述的方法,其特征在于,进一步包括在所述多个维护波中的第一个维护波之后的所述多个维护波中的后续那些维护波期间将所述资源分配给所述经更新的那些服务器。
11. 如权利要求7所述的方法,其特征在于,进一步包括生成所述维护状态信息。
12. 如权利要求11所述的方法,其特征在于,进一步包括在所述多个维护波期间更新主

存所述资源的所述服务器。

13. 一种用于对云网络中的资源执行维护的系统,包括:

处理器;

存储器;

维护应用,所述维护应用被存储在所述存储器中并由所述处理器执行并被配置为:

协调位于包括第一和第二可用性区的区域中的服务器的更新,

其中所述第一和第二可用性区中的每一个包括至少一个数据中心,并且所述资源包括虚拟机和容器实例中的至少一个;

在多个维护波期间执行对所述服务器的所述更新,

其中多个维护波中的第一个维护波包括第一周期和第二周期,在所述第一周期期间对所述资源的维护能被远程排定,在所述第二周期期间如果维护没有在所述第一周期期间被尝试,则由所述维护应用排定维护;以及

在所述第一周期期间响应于对分别位于所述第一和第二可用性区并与单个实体相关联的第一和第二资源的维护请求:

以下中的至少一者:在所述第一周期期间将所述第一可用性区中的所述第一资源重新部署到经更新的服务器或者在服务器上原地更新所述第一资源;以及

以下中的至少一者:在所述第一周期期间将所述第二可用性区中的所述第二资源重新部署到经更新的服务器或者在服务器上原地更新所述第二资源。

14. 如权利要求13所述的系统,其特征在于,所述维护应用被进一步配置成在所述多个维护波中的第一个维护波的第二周期期间对在所述第一可用性区内但不在所述第二可用性区内的那些资源发起对维护的排定。

15. 一种用于对云网络中的资源执行维护的方法,包括:

协调对位于包括第一和第二可用性区的区域中的服务器的维护,

其中所述第一和第二可用性区中的每一个包括至少一个数据中心,并且所述资源包括虚拟机和容器实例中的至少一个;

在多个维护波期间执行对所述服务器的所述更新,

其中多个维护波中的第一个维护波包括第一周期和第二周期,在所述第一周期期间对所述资源的维护能被远程排定,在所述第二周期期间如果维护没有在所述第一周期期间被尝试,则由所述数据中心排定维护;以及

在所述第一周期期间响应于对分别位于所述第一和第二可用性区并与单个实体相关联的第一和第二资源的维护请求:

以下中的至少一者:在所述第一周期期间将所述第一可用性区中的所述第一资源重新部署到经更新的服务器或者在服务器上原地更新所述第一资源;以及

以下中的至少一者:在所述第一周期期间将所述第二可用性区中的所述第二资源重新部署到经更新的服务器或者在服务器上原地更新所述第二资源。

16. 如权利要求15所述的方法,其特征在于,进一步包括在所述多个维护波中的第一个维护波的第二周期期间在第一可用性区中而在第二可用性区中排定更新。

17. 一种用于在云网络中的服务器上主存资源的系统,包括:

处理器;

存储器；

维护应用，所述维护应用被存储在所述存储器中并由所述处理器执行并被配置为：

在多个维护波期间更新主存资源的所述服务器，其中每个所述资源包括虚拟机和容器实例中的至少一个；

其中所述多个维护波中的每一个维护波包括第一周期和第二周期，在所述第一周期期间对对应的资源中的一个的维护能被远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述维护应用排定维护；

在所述第一周期和所述第二周期中的至少一个期间从所述服务器中的一个接收健康状况度量；以及

在所述第一周期和所述第二周期中的至少一个期间，基于所述健康状况度量来选择性地阻止对所述服务器中的一个的更新。

18. 如权利要求17所述的系统，其特征在于，所述维护应用被配置成：

接收对所述资源中的一个的远程维护请求；以及

基于所述健康状况度量，执行以下之一：

尝试以下中的至少一者：在所述第一周期期间将所述资源中的一个重新部署到经更新的服务器或在服务器上原地更新所述资源中的一个；或者

标记所述资源中的一个，并在所述第二周期期间阻止对包括所述资源中的一个的服务器排定更新。

19. 一种用于在云网络中的服务器上主存资源的方法，包括：

在多个维护波期间更新主存资源的所述服务器，其中每个资源包括虚拟机和容器实例中的至少一个，

其中所述多个维护波中的每一个维护波包括第一周期和第二周期，在所述第一周期期间对对应资源的维护能被远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述数据中心排定维护；

在所述第一周期和所述第二周期中的至少一个期间从所述服务器中的一个接收健康状况度量；以及

在所述第一周期和所述第二周期中的至少一个期间，基于所述健康状况度量来选择性地阻止对所述服务器中的一个的更新。

20. 如权利要求19所述的方法，其特征在于，进一步包括：

接收对所述资源中的一个的远程维护请求；以及

基于所述健康状况度量，执行以下之一：

尝试在所述第一周期期间将所述资源中的一个重新部署到经更新的服务器或在服务器上原地更新所述资源中的一个；或者

标记所述资源中的一个，并在所述第二周期期间阻止对包括所述资源中的一个的服务器排定更新。

将新虚拟机和容器分配给云网络中的服务器的系统和方法

技术领域

[0001] 本公开涉及数据中心，并更具体地涉及用于在数据中心中执行主机维护的 系统和方法。

[0002] 背景

[0003] 在此提供的背景描述是出于概要呈现本公开的上下文的目的。当前署名的 发明人针对在本背景章节中描述的工作的范围的工作，以及以其它方式可能不 符合在申请时作为现有技术的资格的描述的各方面既不是清楚地也非隐含地 要被认可为针对本公开的现有技术。

[0004] 云服务提供者使用虚拟机(VM)或容器来实现基础架构即服务(IaaS)和平台 即服务(PaaS)。数据中心包括主存VM或容器的服务器。每个服务器可主存许多VM和/或容器。VM运行在客操作系统上并与管理程序对接，所述管理程序 共享并管理服务器硬件并隔离各 VM。

[0005] 不同于VM，容器不需要完整的OS被安装或主服务器硬件的虚拟副本。容器可包括一个或多个软件模块和库并要求使用操作系统的某些部分。作为减 少的占用空间的结果，相比较于虚拟机，更多的容器可被部署在服务器上。

[0006] 云服务提供者周期性地在服务器上执行维护。例如，云服务提供者通常需 要执行操作系统更新。虽然可以执行一些不影响VM和/或容器的操作的维护任 务，但是其他维护任务会更多地被涉及并可能要求客户停机时间。客户不喜欢 由维护而造成的不便。

[0007] 虽然服务器上的所有VM和/或容器可能被单个客户所拥有，但服务器更有 可能将包括被一个或多个不同客户所拥有的VM和/或容器。换言之，云服务提 供者通常在多承租环境中提供主存服务。其他云服务(诸如数据库、消息收发 服务、web主存等)也遵循相同的方式。

[0008] 有些客户会喜欢安排维护的定时的能力来限制对其业务进行维护的不利 影响。然而，在多承租环境中，允许单个客户控制服务器上的维护的定时是困 难的，因为这将很可能会影响到其他客户。

[0009] 在某些系统中，使用所排定的维护方式。云服务提供者通知客户机需要在 特定时间对服务器执行维护。通常，客户无法控制将被执行的维护的确切时间 或者对将被执行的维护的确切时间的控制量非常有限。维护被应用于同时被主 存在相同的服务器上的所有资源(例如，VM和/或容器)。

[0010] 某些系统可能会将VM或容器重新部署到经更新的服务器。重新部署涉及 在经更新的服务器上创建新的VM或容器，并然后将客户切换到新的VM或容 器上。虽然重新部署VM 和容器使得客户能够控制对其服务有不利影响的时间，但是这要求额外的服务器容量以被用于转换空间。重新部署也会导致VM或容 器丢失在临时硬盘驱动器中的数据。在没有足 够的转换空间的情况下，VM和/ 或容器无法被重新部署，因为经更新的服务器上的空间量 将不足。通常，转换 空间被优化以确保云服务提供者适当的投资回报率。然而，最佳的转换 空间的 量通常留下太少的空间，以至于不能允许VM或容器的重大的重新部署。

[0011] 其他系统可能使用VM或容器实例的实时迁移来减少对客户的影响。实时迁移指的是在不同物理机之间移动正在运行的VM或容器，而不断开客户端或应用。VM或容器的存储器、存储、以及网络连接从原客机器被转移到目的地。然而，实时迁移花费更长时间来完成、要求大量的系统资源、而且还要求一些额外的容量。此外，实时迁移很难普遍适用，在某些别无选择的情况下，只能进行所排定的维护。

[0012] 概述

[0013] 一种用于向数据中心中的服务器分配资源的系统包括被存储在存储器中并被处理器执行的分配应用，并且所述分配应用被配置成向数据中心中的服务器分配资源，其中所述资源包括虚拟机和容器实例中的至少一个。分配应用接收维护状态信息，所述维护状态信息标识多个维护波中的哪个维护波正在执行更新。分配应用基于维护状态信息调整经更新的那些服务器和未经更新的那些服务器之间的资源的分配。

[0014] 在其他特征中，多个维护波中的每一个包括第一周期和第二周期，在所述第一周期期间对资源的维护被选择性地远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述数据中心排定维护。分配应用被进一步配置成在多个维护波中的第一个维护波期间将所述资源分配给未经更新的那些服务器。所述分配应用被进一步配置成在所述多个维护波中的第一个维护波之后的所述多个维护波中的后续那些维护波期间将所述资源分配给所述经更新的那些服务器。

[0015] 在其他特征中，维护服务器包括被配置为生成所述维护状态信息的维护应用。所述维护应用被进一步配置为在多个维护波期间更新主存所述资源的所述服务器。

[0016] 一种用于将资源分配给云网络中的服务器的方法包括将资源分配给所述数据中心中的所述服务器。所述资源包括虚拟机和容器实例中的至少一个。所述方法包括接收标识多个维护波中的哪个维护波正在执行更新的维护状态信息；以及基于所述维护状态信息来调整经更新的那些服务器和未经更新的那些服务器之间的所述资源的分配。

[0017] 在其他特征中，多个维护波中的每一个包括第一周期和第二周期，在所述第一周期期间对资源的维护可被远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述数据中心排定维护。

[0018] 在其他特征中，所述方法包括在多个维护波中的第一个维护波期间将所述资源分配给未经更新的那些服务器。所述方法包括在所述多个维护波中的第一个维护波之后的所述多个维护波中的后续那些维护波期间将所述资源分配给所述经更新的那些服务器。所述方法包括生成所述维护状态信息。所述方法包括在多个维护波期间更新主存所述资源的所述服务器。

[0019] 一种用于对云网络中的资源执行维护的系统包括被存储在存储器中并由处理器执行的维护应用，并且所述维护应用被配置成协调位于包括第一和第二可用性区的区域中的服务器的更新。所述第一和第二可用性区中的每一个包括至少一个数据中心，并且所述资源包括虚拟机和容器实例中的至少一个。所述方法包括在多个维护波期间执行对所述服务器的所述更新。多个维护波中的第一个维护波包括第一周期和第二周期，在所述第一周期期间对资源的维护可被远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述维护应用排定维护。在所述第一周期期间响应于对分别位于所述第一和第二可用性区并与单个实体相关联的第一和第二资源的维护请求，维护应用中

的至少一个在所述第一周期期间将所述第一可用性区中的所述第一资源重新部署到经更新的服务器或者原地更新所述第一资源。所述维护应用至少执行以下一项：在所述第一周期期间将所述第二可用性区中的所述第二资源重新部署到更新的服务器或者在服务器上原地更新所述第二资源。

[0020] 在其他特征中，所述维护应用被进一步配置成在所述多个维护波中的第一个维护波的第二周期期间对在所述第一可用性区内但不在所述第二可用性区内的那些资源发起排定维护。

[0021] 一种用于对云网络中的资源执行维护的方法包括协调对位于包括第一和第二可用性区的区域中的服务器的维护。所述第一和第二可用性区中的每一个包括至少一个数据中心，并且所述资源包括虚拟机和容器实例中的至少一个。所述方法包括在多个维护波期间执行对所述服务器的所述更新。多个维护波中的第一个维护波包括第一周期和第二周期，在所述第一周期期间对资源的维护可被远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述数据中心排定维护。在所述第一周期期间响应于对分别位于所述第一和第二可用性区并与单个实体相关联的第一和第二资源的维护请求，在所述第一周期期间至少执行以下一项：将所述第一可用性区中的所述第一资源重新部署到经更新的服务器或者在服务器上原地更新所述第一资源。所述方法包括在所述第一周期期间至少执行以下一项：将所述第二可用性区中的所述第二资源重新部署到更新的服务器或者在服务器上原地更新所述第二资源。

[0022] 在其他特征中，所述方法包括在所述多个维护波中的第一个维护波的第二周期期间在第一可用性区中而不在第二可用性区中排定更新。

[0023] 一种用于在云网络中的服务器上主存资源的系统包括被存储在存储器中并由处理器执行的维护应用，并且所述维护应用被配置成在多个维护波期间更新主存资源的所述服务器。所述资源中的每一个包括虚拟机和容器实例中的至少一个。多个维护波中的每一个维护波包括第一周期和第二周期，在所述第一周期期间对对应资源的维护可被远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述维护应用排定维护。所述维护应用在所述第一周期和所述第二周期中的至少一个期间从所述服务器中的一个接收健康状况度量。所述维护应用在所述第一周期和所述第二周期中的至少一个期间，基于所述健康状况度量来选择性地阻止对所述服务器中的一个的更新。

[0024] 在其他特征中，所述维护应用被配置成基于所述健康状态度量接收对所述资源中的一个的远程维护请求，并且执行以下之一：尝试以下中的至少一者：在所述第一周期期间将所述资源中的一个重新部署到经更新的服务器或在服务器上原地更新所述资源中的一个，或者标记所述资源中的一个，并在所述第二周期期间阻止对包括所述资源中的一个的服务器排定更新。

[0025] 一种用于在云网络中的服务器上主存资源的方法包括在多个维护波期间更新主存资源的所述服务器。所述资源中的每一个包括虚拟机和容器实例中的至少一个。多个维护波中的每一个维护波包括第一周期和第二周期，在所述第一周期期间对对应资源的维护可被远程排定，在所述第二周期期间如果维护没有在所述第一周期期间被尝试，则由所述数据中心排定维护。所述方法包括在所述第一周期和所述第二周期中的至少一个期间从所述服务器中的一个接收健康状况度量；以及在所述第一周期和所述第二周期中的至

少一个期间，基于 所述健康状况度量来选择性地阻止对所述服务器中的一个的更新。

[0026] 在其他特征中，所述方法包括接收对所述资源中的一个的远程维护请求。基于所述健康状况度量，所述方法包括尝试在所述第一周期期间将所述资源中 的一个重新部署到经更新的服务器或在服务器上原地更新所述资源中的一者；或者标记所述资源中的一个，并在所述第二周期期间阻止对包括所述资源中的 一个的服务器排定更新。

[0027] 本公开的应用性的更多范围将从详细的说明书、权项和附图中变得显而易 见。详细的说明书和具体示例旨在仅仅进行说明的目的，并且不是要限制本公 开的范围。

附图说明

[0028] 图1是根据本公开的云服务提供者的示例的功能框图。

[0029] 图2是根据本公开的数据中心的示例的功能框图。

[0030] 图3A和3B是根据本公开的包括VM和/或容器的服务器的示例的功能框 图。

图3C和3D是根据本公开的包括维护服务器和分配服务器的示例的功能框 图。

[0031] 图4是例示出根据本公开的各波中维护部署的示例的时序图；

[0032] 图5A和5B是例示出根据本公开的用于在主存VM和/或容器的节点上执 行维护的方法的示例的流程图。

[0033] 图6A和6B是例示出根据本公开的用于在主存VM和/或容器的节点上的 维护期间创建转换空间的方法的示例的流程图。

[0034] 图7-8是例示出根据本公开的用于确定是否要重新部署VM或容器或者原 地执行维护的方法的示例的流程图。

[0035] 图9是例示出根据本公开的用于在维护期间向节点供应VM和/或容器的方法的示例的流程图。

[0036] 图10是例示出根据本公开的用于基于节点的健康状况来执行维护的方法 的示例的流程图。

[0037] 图11是例示出根据本公开的在维护波的第一周期期间允许客户跨多个可 用性区对VM和/或容器进行所排定的重新部署的方法的示例的流程图。

[0038] 在附图中，参考标号可被重用以标识相似的和/或相同的元素。

具体实施方式

[0039] 根据本公开的用于操作云服务提供者的系统和方法提供了包括多个维护 波的经计划的维护循环。在第一维护波的第一周期期间，具有支持经计划的维 护特征的服务等级协议 (SLA) 的一些或全部的客户被通知。该通知指定可在 其期间尝试客户发起的对虚拟机 (VM) 和/或容器进行维护重新部署的时间段。在一些示例中，这些客户可通过执行维护重新部署命令来发起对VM和/或其他 容器的维护。如果客户执行维护重新部署命令，则作出尝试来重新部署VM或 容器(或者在某些情况下执行原地维护)。如果客户执行维护重新部署命令并 且维护没有被成功完成，则该VM或容器被标记。

[0040] 在第一波的第二周期期间，云服务提供者控制对服务器或节点的更新的定 时。然而，如果客户的VM或容器在第一阶段被标记，则维护被推迟到后续波，如下文将进一步描述的。对具有经计划的维护或选择退出特征的客户(分别在 第一波的第一周期期间没有发

起维护或选择退出的那些客户机) 和不具有经计划的维护或选择退出特征的那些客户机执行所排定的维护。

[0041] 现在参考图1,云网络100的示例被示出,并且包括多个区域120-1A、120-1B、120-2A、120-2B、...、120-NA以及120-NB(统称为区域120),其中N是大于1的整数。虽然示出了特定的云网络100,但是本公开涉及其他云网络架构。在此示例中,区域120被成对布置(例如,120-1A和120-1B、120-2A 和120-2B等)。区域120中的每一个包括一个或多个可用性区,诸如可用性区124-1A1、124-1A2、...124-1AX(统称为可用性区124),其中X是大于零的整数。区域120通常在用于故障和/或备份域分离的不同维护波期间被配对并被处理。在一些示例中,可用性区124也在用于故障和/或备份域分离的不同维护波期间被处理。区域内的可用性区124通过低等待时间通信链路125连接。每个可用性区包括一个或多个数据中心(DC) 128-1A1、128-1A2、...128-1AY(统称为数据中心128),其中Y是大于零的整数。每个数据中心128包括多个服务器(未在图1中示出)。

[0042] 一个或多个客户网络140-1、140-2、...140-C中的计算机可被用于经由诸如因特网之类的分布式通信系统108来访问云网络100。维护服务器130可被用于排定对云网络100的维护。通常在第一维护波或第一波期间在某一成对的区域中的第一个的服务器上执行维护。如上所述,第一波包括第一周期,在该第一周期期间至少一些客户尝试发起对VM或容器的维护排定,接着是第二周期,在该第二周期期间云服务提供者排定并尝试发起对至少一些VM或容器的维护。

[0043] 然后,以与第一波相似的方式,在第二波期间对该成对的区域中的第二个执行维护。如果在第一波期间没有完成对该成对的区域中的第一个中的服务器的维护,则在第二波之后的第三维护波中尝试维护。如果在第二波期间没有完成对该成对的区域中的第二个中的服务器的维护,则在第三波之后的第四维护波中尝试维护。可按需要执行附加波以完成对所有节点的维护。

[0044] 现在参考图2,示出了数据中心128的示例。数据中心包括防火墙/路由器 150、安全服务器154、分配服务器162、以及维护服务器166。虽然防火墙/路由器150、安全服务器154、分配服务器162、以及维护服务器166被示为单独的服务器,但是这些服务器中的一个或多个的功能可被结合或进一步被分布。安全服务器154执行用户认证。维护服务器166与维护服务器130进行通信,并对数据中心128中的服务器执行维护,如下文将进一步描述的。分配服务器 162将新的VM或容器分配给数据中心128中的服务器,如下文将进一步描述的。

[0045] 数据中心128包括多个机架170-1、170-2、...、170R-1、以及170-R(统称为机架170),其中R是大于1的整数。每个机架170包括一个或多个路由器和一个或多个服务器。例如,机架170-1包括路由器174和一个或多个服务器180-1、180-2、...、以及180-S(统称为服务器180)。每个服务器180支持一个或多个虚拟机(VM) 和/或容器。

[0046] 现在参考图3A和3B,示出了用于主存虚拟机的服务器180的各示例。在图3A中,使用本机管理程序的服务器被示出。服务器180包括硬件188,诸如有线或无线接口190、一个或多个处理器192、易失和非易失存储器194以及大容量存储196,诸如硬盘驱动器或闪存驱动器。管理程序198直接在硬件188 上运行以控制硬件188并管理虚拟机204-1、204-2、...、204-V(统称为虚拟机 204) 和对应的客操作系统208-1、208-2、...、208-V(统称为客操

作系统208)，其中V是大于1的整数。在此示例中，管理程序198在传统的操作系统上运行。客操作系统208作为进程在主操作系统上运行。管理程序的各示例包括微软的 Hyper-V、Xen、Oracle的VM Server for SPARC(针对SPARC的VM服务器)、Oracle的VM Server for x86(针对x86的VM服务器)、Citrix的XenServer 以及VMware的ESX/ESXi，但是其他管理程序也可被使用。

[0047] 现在参考图3B，第二类型的管理程序可以被使用。服务器180包括硬件 188，诸如 有线或无线接口190、一个或多个处理器192、易失和非易失存储器 194以及大容量存储 196，诸如硬盘驱动器或闪存驱动器。管理程序224在主 操作系统220上运行。虚拟机204-1、204-2、…、204-V(统称为虚拟机204) 和对应的客操作系统208-1、208-2、…、208-V(统称为客操作系统208)。客 操作系统208是从主操作系统220中被抽象的。这种第二类型的示例包括 VMware Workstation(工作站)、VMware Player(播放器)、VirtualBox、Parallels Desktop for Mac and QEMU(针对Mac和QEMU的Parallels桌面)。尽管示 出了管理程序的两个示例，但是其它类型的管理程序也可被使用。

[0048] 服务器180可包括健康状况监控应用223，该健康状况监控应用监控各种 软件层 和/或硬件的健康状态，并且在周期性或事件的基础上将健康状况报告给 维护服务器，如 下文将进一步描述的。在一些示例中，健康状况监控应用223 在服务器180的VM或容器中运 行。

[0049] 现在参考图3C和3D，分别示出了维护服务器166和分配服务器162的示 例。在图3C 中，维护服务器166包括硬件250，诸如 有线或无线接口254、一 个或多个处理器258、易失和 非易失存储器62以及大容量存储264，诸如硬盘 驱动器或闪存驱动器。维护服务器166进一 步包括运行维护应用276的操作系 统272，如下文将进一步描述的。

[0050] 在图3D中，分配服务器162包括硬件280，诸如 有线或无线接口282、一 个或多个处理器286、易失和非易失存储器290以及大容量存储292，诸如硬 盘驱动器或闪存驱动器。维 护服务器166进一步包括运行分配应用296的操作 系统294，如下文将进一步描述的。

[0051] 现在参考图4，在一些示例中，对服务器的维护以波的方式被执行。区域 被成对组 织。在第一波300-1期间，对区域对的第一区域执行维护。在第二波300-2期间，对区域对的 第二区域执行维护。在第三波300-3期间，对区域对的 第一区域执行维护。在第四波300-4 期间，对区域对的第二区域执行维护。持 续该过程直到对第一和第二区域中的所有服务器 完成维护。可将服务器分割成 区域对，以确保故障和/或备份域被实施。

[0052] 在第一波300-1的第一周期300-1A期间，云服务提供者联系需要维护的服 务器或 节点上的一些或全部的客户。被联系的客户可能具有服务等级协议 (SLA)，该服务等级协 议具有经计划的维护特征(允许由客户排定的客户发 起维护)和/或选择退出特征。云服务 提供者指示在第一波300-1的第一周期 300-1A期间，客户可发起维护(如果适用的话，和/ 或选择退出)。如果客户 发起维护重新部署(例如使用维护重新部署命令)，则对VM或容器 的维护重 新部署被尝试。

[0053] 如果对客户的维护重新部署成功，则可从需要维护的VM或容器列表中移 除该客 户。如果维护重新部署不成功，则该客户的VM或容器被标记并且在第一波300-1的第二周 期300-1B期间维护不被尝试。本质上，如果客户在波的第一周期期间尝试发起维护，则客 户不会在该波的第二周期期间经受强制维护排 定的惩罚。同样，选择退出第一波的客户在

第一波300-1的第二周期300-1B期间将不具有被执行的维护。

[0054] 如果客户在第一波300-1的第一周期300-1A期间没有尝试维护重新部署并且没有选择退出，则在第一波300-1的第二周期300-1B期间（在云服务提供者所排定的时刻）云服务提供者将在该节点上尝试执行经排定的维护。

[0055] 在第二波300-2期间，以类似的方式对区域对的第二区域执行维护。在第三波300-3期间，云服务提供者返回到区域对的第一区域，以尝试对在第一波300-1中未经更新的节点执行维护。对于这些剩余的节点，过程类似于在第一波300-1期间所使用的过程。在第三波300-3期间，客户可能能够或可能不能够选择退出。在附加的波中继续该过程直到所有的维护被执行。

[0056] 现在参考图5A和5B，示出了用于操作维护应用276的方法400。在图5A中，如果可能的话维护应用在410处创建转换空间，如下文将进一步描述的。当第一波的第一周期在416开始时，维护应用向要求维护的第一区域中的具有VM或容器的客户发送消息。在一些示例中，该消息仅被发送给具有经计划的维护和/或选择退出特征的客户。该消息允许客户在第一波的第一周期期间排定维护（或在某些情况下选择退出）。如果客户在第一波的第一周期期间没有尝试发起维护重新部署或选择退出，则云服务提供者将在第一波的第二周期期间排定维护。

[0057] 在420，维护应用确定在第一周期期间是否接收到针对客户的维护重新部署（MR）命令。如果420为真，则维护应用在426尝试将VM或容器重新部署到另一个节点，或尝试执行原地维护。如果在428所确定的维护成功，则客户VM或容器从维护列表中被移除。如果在440所确定的第一波的第一周期没有结束，则维护应用返回到420。如果420为假，则维护应用在434确定客户是否已选择退出维护。在一些示例中，选择客户可被提供选择退出特征。如果在428所确定的执行客户经排定的维护的尝试失败，或者客户在434已选择退出，则客户在438被标记，并且所述方法在440处继续。如果维护重新部署成功，则在441处VM或容器从维护列表中被移除（如果该节点上的维护已原地执行，则通知也从维护列表中被移除）。如上所述，经标记的客户在第一波的第二周期期间未被更新，而在后续波期间被处理。

[0058] 现在参考图5B，当在450所确定的第二周期开始的情况下，维护方法在463对维护列表中剩余的节点执行所排定的维护。在一些示例中，此时不对具有经标记的VM或容器的未经更新的节点进行更新。如果在464所确定的维护成功，则所述方法在470处继续并从维护列表中移除该节点。如果没有，则该节点不会从维护列表中被移除。在472，维护应用确定第二周期是否已结束。如果472为真，则所述方法返回。如果472为假，则维护应用在476确定对其他节点的维护是否应该被尝试。作出是否应对其他节点执行维护的确定可能取决于是否有任何节点剩余在维护列表中、是否存在足够的转换空间以允许维护被执行、是否可对任何节点执行原地维护以及其他类似的准则。如果476为假，则所述方法返回到462。如果476为真，则所述方法返回。

[0059] 在一些示例中，在该波的第二周期期间，具有经计划的维护特征的客户可被允许在该波的第二周期期间执行由云服务提供者所排定的维护的时刻之前的任何时刻发起维护重新部署。在一些示例中，在该波的第二周期期间，具有选择退出特征的客户也可被允许在该波的第二周期期间执行由云服务提供者所排定的维护的时刻之前的任何时刻选择

退出。

[0060] 现在参考图6A和6B,示出了用于在发起维护之前(例如在第一周期期间(诸如图5A中的410期间))创建转换空间的示例方法。在510,维护应用确定是否存在不具有任何VM或容器的节点。如果510为真,则维护应用对该节点执行维护。在维护被执行之后,该节点被更新并且可作为用于后续重新部署VM或容器的转换空间的源。在一些示例中,节点可能具有可以以可接受的成本/效益使用实时迁移来移动的一个或几个VM或容器。不具有需要维护通知的SLA的一些VM和/或容器可被重新部署或被执行原地维护,以便创建转换空间。

[0061] 在518,维护应用可确定节点是否包括与PaaS相关联的VM或容器。如果518为真,则维护应用在不移动PaaS VM或容器的情况下对节点执行维护。如果如在526所确定的存在足够的转换空间,则所述方法返回。如果如在526所确定的转换空间不足,则维护应用可使用实时迁移或本文所描述的其他技术来创建转换空间。实时迁移的示例可包括将VM或容器移动到未经更新的节点以释放一个或多个节点,使得节点可被更新并被用于转换空间,如528所示。其他示例包括移动和/或协同定位经标记的VM或容器以释放节点,如图6B中的529所示。

[0062] 现在参考图7,示出了用于在第一周期期间(诸如图5A中的426期间)响应于维护重新部署命令而重新部署或更新节点的示例。在550,维护应用确定特定节点上的(一个或多个)VM或(一个或多个)容器是否(都)与单个客户相关联。如果550为真,则维护应用对该节点执行原地维护。如果该节点上有多个VM或容器,则维护重新部署命令可能需要附加字段,其中当客户对一个VM排定维护并且所述客户在相同节点上具有其他VM时,客户可指定对节点的维护可原地继续。

[0063] 如果550为假,则维护应用确定节点上的其他VM或容器是否是PaaS VM或容器。如果一个或多个VM或容器与单个客户相关联,并且该节点上所有剩余的VM和容器都是PaaS VM或容器,则此步骤也可为真。如果558为真,则维护应用对该节点执行原地维护。如果558为假,则如果可能的话,维护应用将VM或容器重新部署到经更新的节点。此决定可包括将VM的需求与经更新的节点上的可用空间、打包考量和/或其他准则进行比较。

[0064] 现在参考图8,示出了用于在第二周期期间(诸如图5B中的426期间)响应来重新部署或更新节点的示例。在570,维护应用确定是否存在没有任何经标记的VM或容器的节点。如果570为真,则维护应用在574尝试对该节点执行原地维护。在原地更新节点之后,维护应用确定是否存在有足够的可用转换空间来从包括一个或多个经标记的VM和/或容器的节点移动未经标记的VM或容器。如果578为真,则维护应用尝试将这些节点上的未经标记的VM和容器重新部署到经更新的节点。

[0065] 现在参考图9,示出了由分配应用执行的方法600。在610,分配服务器确定在一个或多个维护波期间是否已接收到供应VM或容器的请求。如果610为假,分配应用使用正常的分配过程来分配VM或容器。如果610为真,则分配应用确定维护波是否处于该区域的第一波中。如果618为真,则分配应用在624将VM或容器分配给尚未更新的节点。如果618为假(并且该维护处于第二波或后续波中),则分配应用在622将VM或容器分配给已被更新的节点。

[0066] 现在参考图10,示出了由维护应用执行的方法650。机架170中的服务器180在操

作期间监控一个或多个软件层和/或硬件组件的各种健康状况度量。对 服务器的健康状况度量的适当监控的示例在共同转让的美国专利申请 9,274,842号“Flexible and Safe Monitoring of Computers(计算机的柔性和安全监 控)”,和美国专利申请8,365,009号“Controlled Automatic Healing of Data-Center Services(受控的数据中心服务的自动复原)中被示出并描述,其整体通过引用 结合于此。”

[0067] 服务器180周期性地或在事件的基础上将健康状况度量发送给维护服务器 166。服务器180可发送一个或多个健康状况度量。在一些示例中,可为不同的软件层和/或不同的硬件系统生成多个健康状况度量。维护服务器166可基于 一个或多个健康状况度量与一个或多个预定范围的比较来确定服务器的健康 状况是好的/坏的。替换地,服务器180可向维护服务器发送针对一个或多个软 件层和/或硬件组件的好/坏的健康状况指示符(例如二进制指示符)。

[0068] 在652,维护服务器166从节点接收一个或多个健康状况度量。维护服务 器166基于节点度量来确定该节点是否具有坏的健康状况。替换地,如上所述,服务器180可发送好的/坏的健康状况指示符。在节点具有坏的健康状况的情况 下(如在656所确定的),维护服务器166在660禁用对该节点的维护。如果 节点包括如在664所确定的第一周期期间由客户 为VM或容器排定维护重新部 署,则VM或容器在668被标记并且在下一波期间维护被执行。如果656为假 (并且节点的健康状况是好的),则维护被启用。

[0069] 现在参考图11,示出了用于在第一波的第一周期期间跨两个或多个可用性 区执行对两个或更多个VM或容器的经排定的维护的方法800。在一些示例中, 每个维护波可被限制于一个可用性区而不是一个区域。一些客户可能具有位于 两个或更多个可用性区中的多个VM或容器。客户可能希望同时在多个VM或 容器上执行维护重新部署以减少停机时间。在一些示例中,某时刻在一个可用 性区上执行维护,因此这是不可能的。因此,具有位于两个或更多个可用性区 中的多个VM或容器的客户不能在大致同一时间对这些VM或容器 排定维护。本文描述的系统和方法允许客户在第一波的第一周期期间在两个或更多个VM 或容器上排定维护重新部署。

[0070] 在810,维护应用确定是否存在将被执行的维护。在810为真的情况下, 如果客户 具有在不同可用性区中的VM或容器,则维护应用向客户发送消息, 从而允许在第一波的第一周期期间为不同可用性区中的VM或容器排定维护。在一些示例中,客户被告知如果客户 不为VM排定维护,则维护服务器将在与 不同波(和不同周期) 相关联的两个或多个不同的 第二周期期间排定维护。

[0071] 在818当第一波的第一周期开始时,维护应用在822在第一波的第一周期 期间使用维护重新部署命令,跨两个或更多个可用性区实现客户所排定的对 VM或容器的维护重新部署。在824,维护应用确定是否接收到对两个或更多 个可用性区中的两个或更多个VM 或容器的维护重新部署命令。如果824为真, 则维护应用在826尝试对跨两个或更多个可用性区的VM或容器执行维护重新 部署。

[0072] 在830,维护应用确定针对每个VM或容器的维护重新部署是否成功。如 果830为 真,则维护应用从列表中移除已成功更新的VM或容器。来自830(如 果为假)或834的所述方法在838处继续,其中维护应用确定第一周期是否结 束。如果838为假,则所述方法在824处继续。在838为真的情况下,所述方 法在842处继续并确定第二周期是否开始。在842为真的

情况下,所述方法在 846处继续,并每次将维护波的第二周期期间的维护限制到一个可用性区。在 850,所述方法确定第二周期是否结束。在850为假的情况下,所述方法在846 继续,否则所述方法返回。

[0073] 前面的描述本质上仅仅是说明性的,并且并非旨在限制本公开、其应用或使 用。本公开的广泛示教可以用各种方式来实现。因此,尽管本公开包括特定示 例,但本公开的真实范围不受这样的限制,因为依据附图、说明书和下述权项 的研究,其它修改将变得显而易见。应该理解,在方法中的一个或多个步骤可 以以不同的顺序(或同时)被执行,而无需改变本公开的原理。而且,尽管每 个实施例如上所述为具有某些特征,参考本公开的任何实施例描述的其它特征 的任意一个或多个可以被实现在任意其它实施例中和/或结合任意其它实施例 来实现,即使该组合并未明确描述。换句话说,所述的实施例并不是互斥的, 并且一个或多个实施例彼此的置换仍然在本公开的范围内。

[0074] 在元件之间(例如在模块、电路元件、半导体层等之间)的空间和功能性 关系使用各种术语被描述,包括“相连的”、“啮合的”、“耦合的”、“相邻的”、“在… 旁”、“在… 顶上”、“之上”、“之下”和“布置”。除非明确被描述为是“直接”,当 第一和第二元件之间的关系在上述公开中被描述时,该关系可以是直接关系, 其中在第一和第二元件之间不存在其它中介元件,但是也可以是间接关系,其 中在第一和第二元件之间(空间上的或功能性地)存在一个或多个中介元件。如在此使用的,在至少一个A、B和C处的短语应该被解释为意指一个逻辑(A OR B OR C), 使用非排他性的逻辑OR,并且不应该被解释为意指“至少一个A、至 少一个B和至少一个C”。

[0075] 在附图中,如由箭头所指示的箭形的方向通常表明图解感兴趣的信息流 (诸如数据或指令)。例如,当元件A和元件B交换各种信息,但从元件A 传送到元件B的信息与图解有关,箭形可以从元件A指向元件B。这种未定向 箭形不是隐含了没有其它信息被从元件B传送到元件A。而且,对于从元件A 发送到元件B的信息,元件B可以将对信息的请求发送给元件A,或将信息的 接收确认发送给元件A。

[0076] 在该应用中,包括下述定义,术语“模块”或术语“控制器”可以用术语“电路” 替代。术语“模块”可以是指下述项的部分或包括下述项:专用集成电路(ASIC); 数字、模拟或 混合模拟/数字分立电路;数字、模拟或混合模拟/数字集成电路; 组合的逻辑电路;现场可 编程门阵列(FPGA); 执行代码的处理器电路(共享的、专用的或分组); 存储由处理器电路 执行的代码的存储器电路(共享的、专用 的或分组); 提供期望功能性的其它合适的硬件组 件; 或上述项的一些或全部 的组合,诸如在片上系统中。

[0077] 所述模块可以包括一个或多个接口电路。在一些示例中,接口电路可以包 括被连接到局域网(LAN)、因特网、广域网(WAN)或其组合的有线或无线接口。本开的任意给定模 块的功能性可以被分布在通过接口电路相连的多个模块之 中。例如,多个模块可以允许负 载平衡。在另外的示例中,服务器(也称为远 程或云)模块可以代表客户端模块来完成某些 功能性。

[0078] 术语代码,如上所述,可以包括软件、固件和/或微代码,并且可以指代程 序、例 程、功能、类、数据结构和/或对象。术语共享处理器电路包括单个处理 器电路,其执行来自 多个模块的一些或所有的代码。术语分组处理器电路包括 处理器电路,该处理器电路与附 加的处理器电路相组合以执行来自一个或多个 模块的一些或所有的代码。对多个处理器

电路的引用包括在分立管芯上的多个 处理器电路、在单个管芯上的多个处理器电路、在单个处理器电路上的多个核、单个处理器电路的多个线程或上述的组合。术语共享存储器电路包括单个存储 器电路，其存储来自多个模块的一些或所有的代码。术语分组存储器电路包括 存储器电路，该存储器电路与附加的存储器相组合以存储来自一个或多个模块 的一些或所有的代码。

[0079] 术语存储器电路是术语计算机可读介质的子集。如在此使用，术语计算机 可读介质不包括传播通过介质(诸如在载波上的)瞬态电子或电磁信号；因此，术语计算机可读介质可以被认为是有形和非瞬态的。非瞬态、有形的计算机可 读介质的非限制性示例是非易失存储器电路(诸如闪存存储器电路、可擦除可 编程只读存储器电路，或掩模只读存储器)、易失存储器电路(诸如静态随机 存取存储器电路或动态随机存取存储器电路)、磁存储器介质(诸如模拟或数 字磁带或硬盘驱动器)，以及光学存储介质(诸如CD、DVD或蓝光盘)。

[0080] 在本申请中，被描述为具有特定属性或执行特定操作的装置元件被具体配 置为具有那些特定属性并执行那些特定操作。具体而言，执行一个动作的元件 的描述意指该元件被配置成执行所述动作。元件的配置可以包括对元件的编 程，诸如通过对与所述元件相 关联的非瞬态、有形计算机可读介质上的指令进 行编码。

[0081] 在本申请中所述的装置和方法可以是部分或全部由专用计算机来实现的，所述 专用计算机通过将通用计算机配置为执行在计算机程序中实现的一个或 多个特定功能来 被创建。如上所述的功能框、流程图组件和其它元件用作软件 规范，其可以通过本领域的 技术人员或程序员的例行工作被转换成计算机程 序。

[0082] 计算机程序包括被存储在至少一个非瞬态、有形计算机可读介质上的处理 器可 执行指令。计算机程序还可以包括所存储的数据或依赖于所存储的数据。计算机程序可以 包括与专用计算机的硬件交互的基本输入/输出系统(BIOS)、与 专用计算机的特定设备交 互的设备驱动器、一个或多个操作系统、用户应用、背景服务、背景应用等。

[0083] 计算机程序可以包括：(i)要被解析的描述性文本，诸如JavaScript Object Notation (JSON)、超文本标记语言 (HTML) 或可扩展标记语言 (XML)，(ii)汇编代码，(iii)由编译器从源代码生成的目标代码，(iv)用于由解释器执行的源代码，(v)用于 由即时编 译器编译并执行的源代码等。仅作为示例，源代码可以使用来自各语 言的语法来编写，所 述语言包括C、C++、C#、Objective C、Haskell、Go、SQL、R、Lisp、Java®、Fortran、Perl、Pascal、Curl、OCaml、Javascript®、HTML5、Ada、ASP(活动服务器页面)、PHP、Scala、Eiffel、Smalltalk、Erlang、Ruby、Flash®、VisualBasic®、Lua和Python®。

[0084] 在权利要求中引用的元件都不是旨在要成为35U.S.C. §112(f) 中的装置+ 功能元 件，除非使用术语“用于…的装置”明确地引用该元件，或者在方法权利 要求使用短语“用 于…的操作”或“用于…的步骤”的情况下。

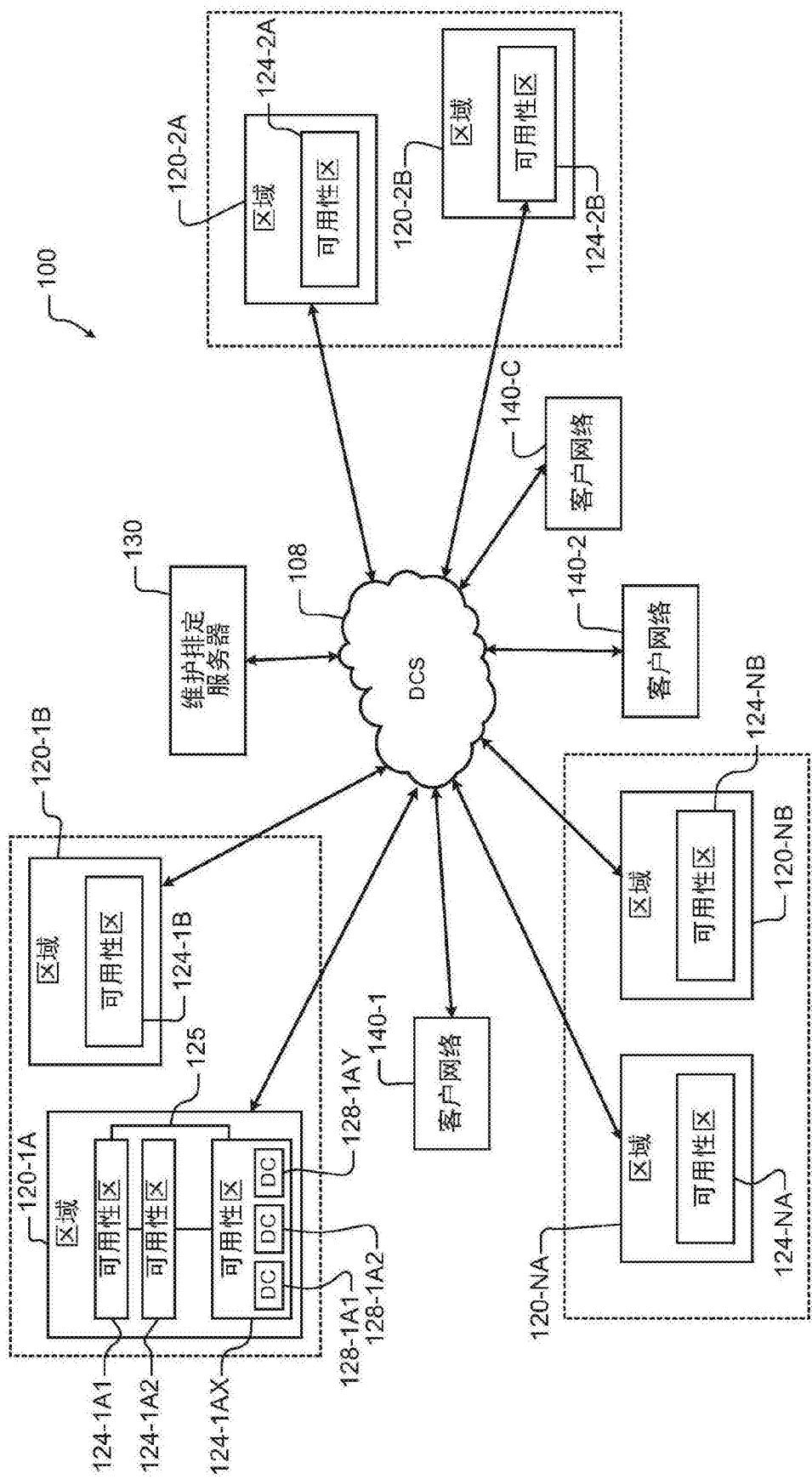


图1

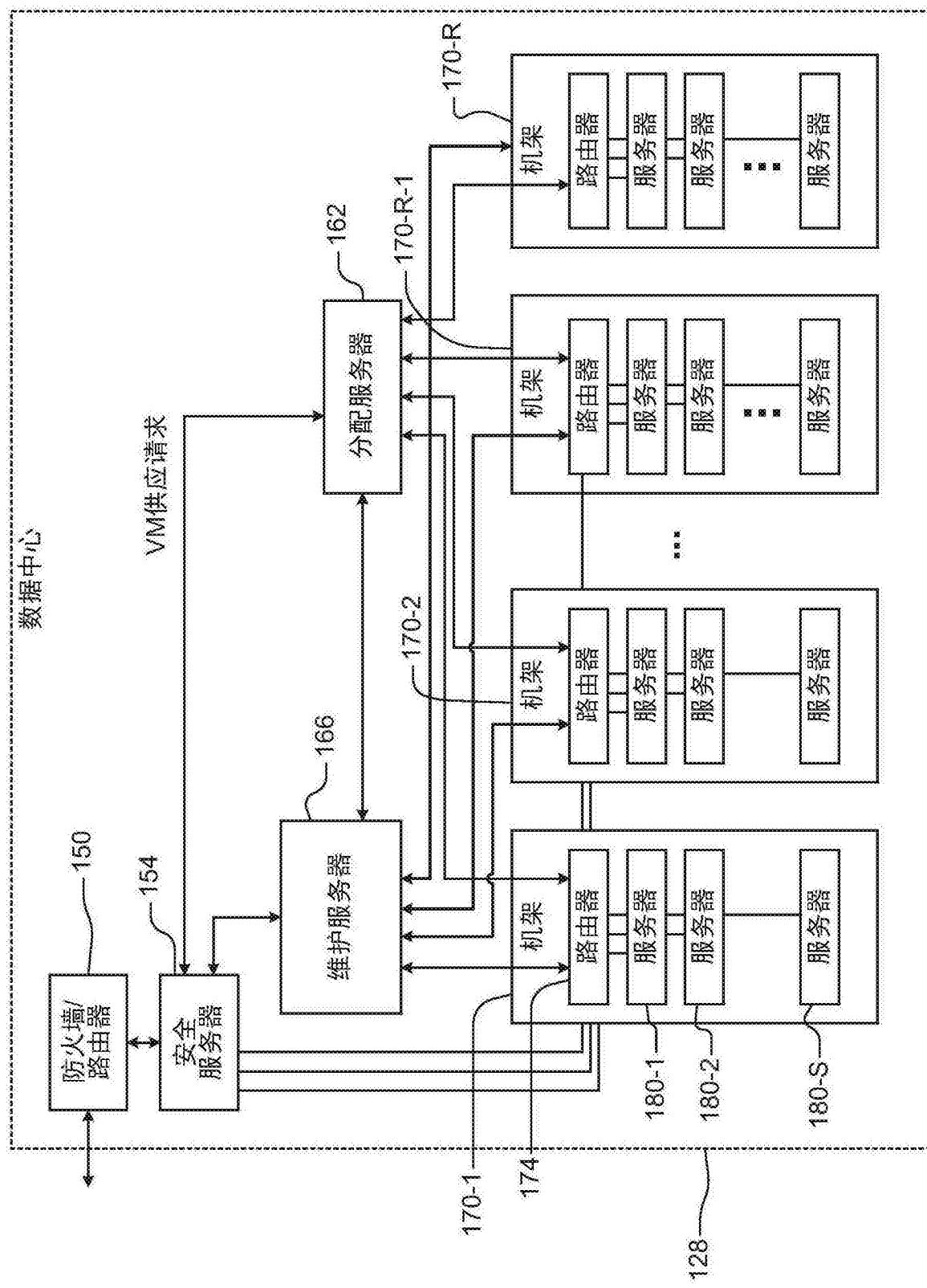


图2

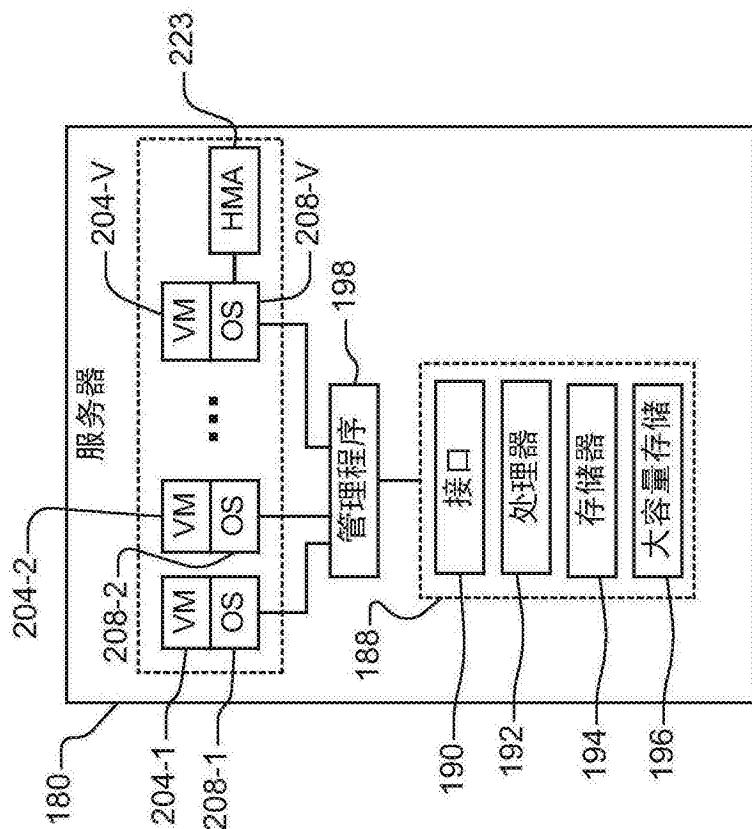


图3A

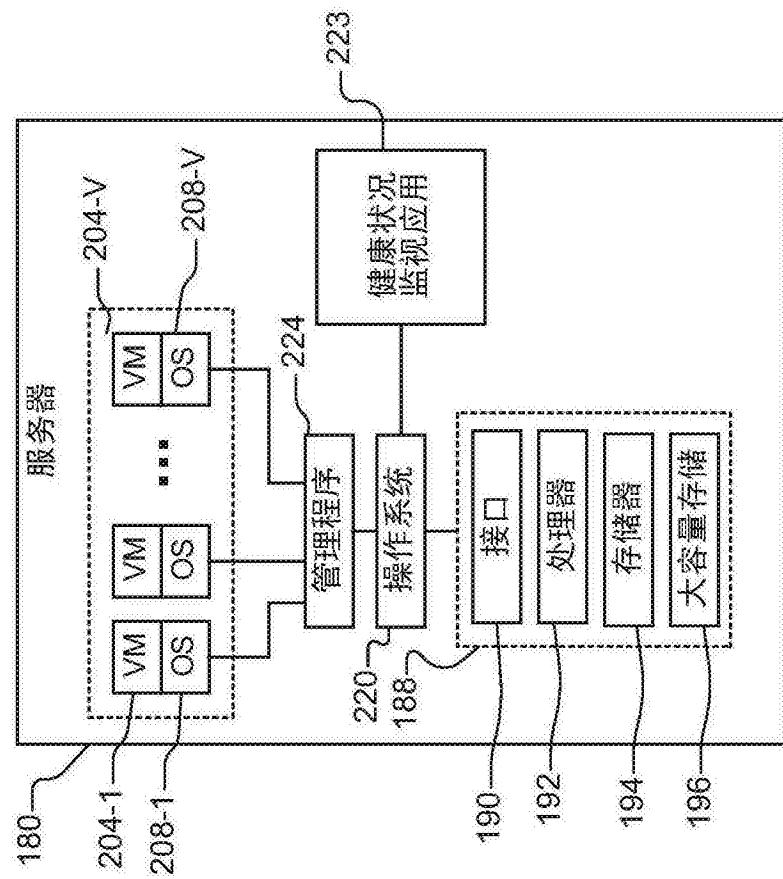


图3B

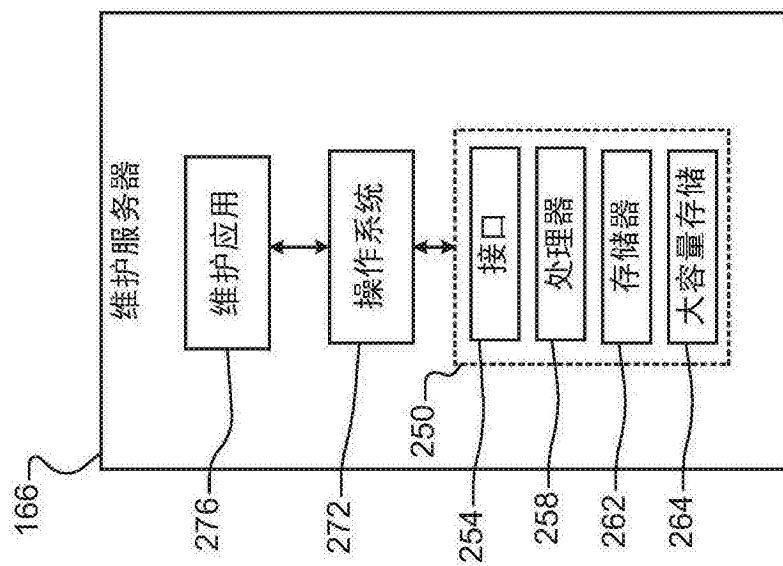


图3C

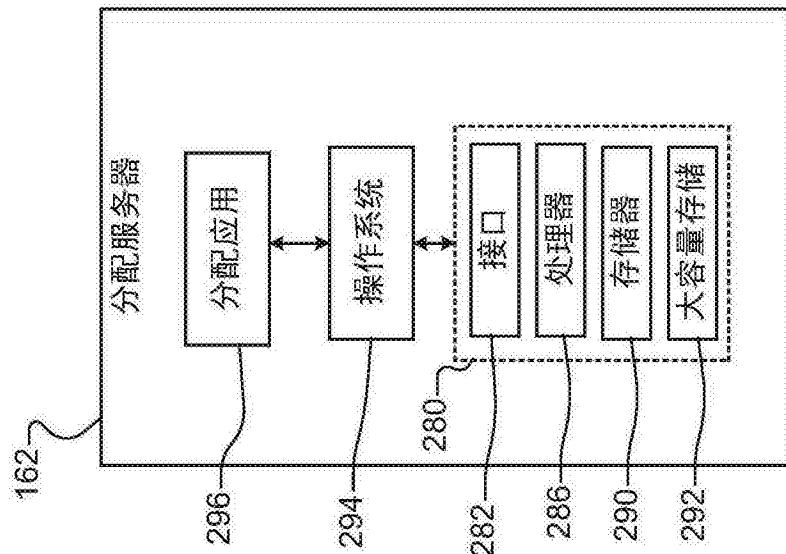


图3D

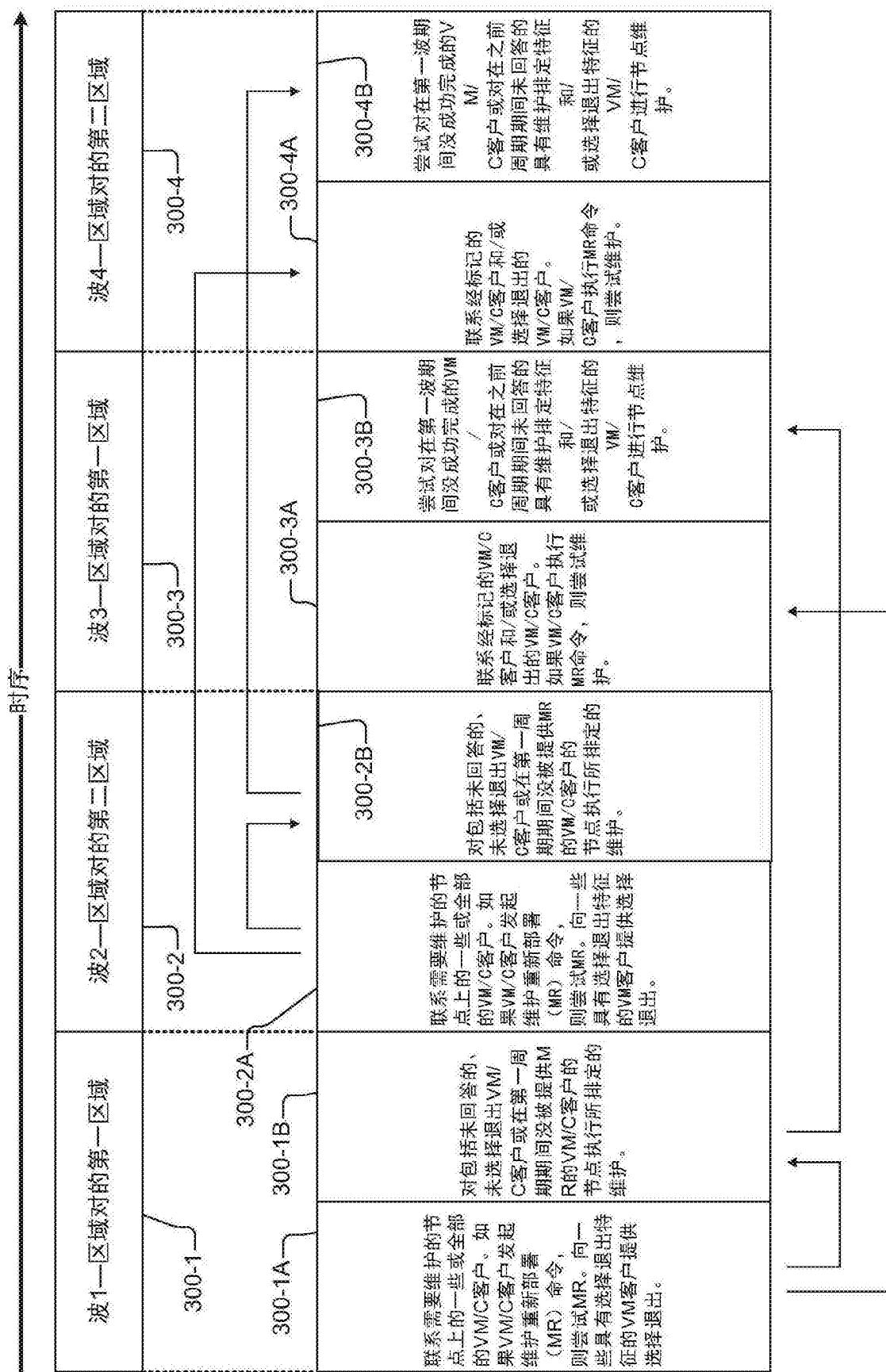


图4

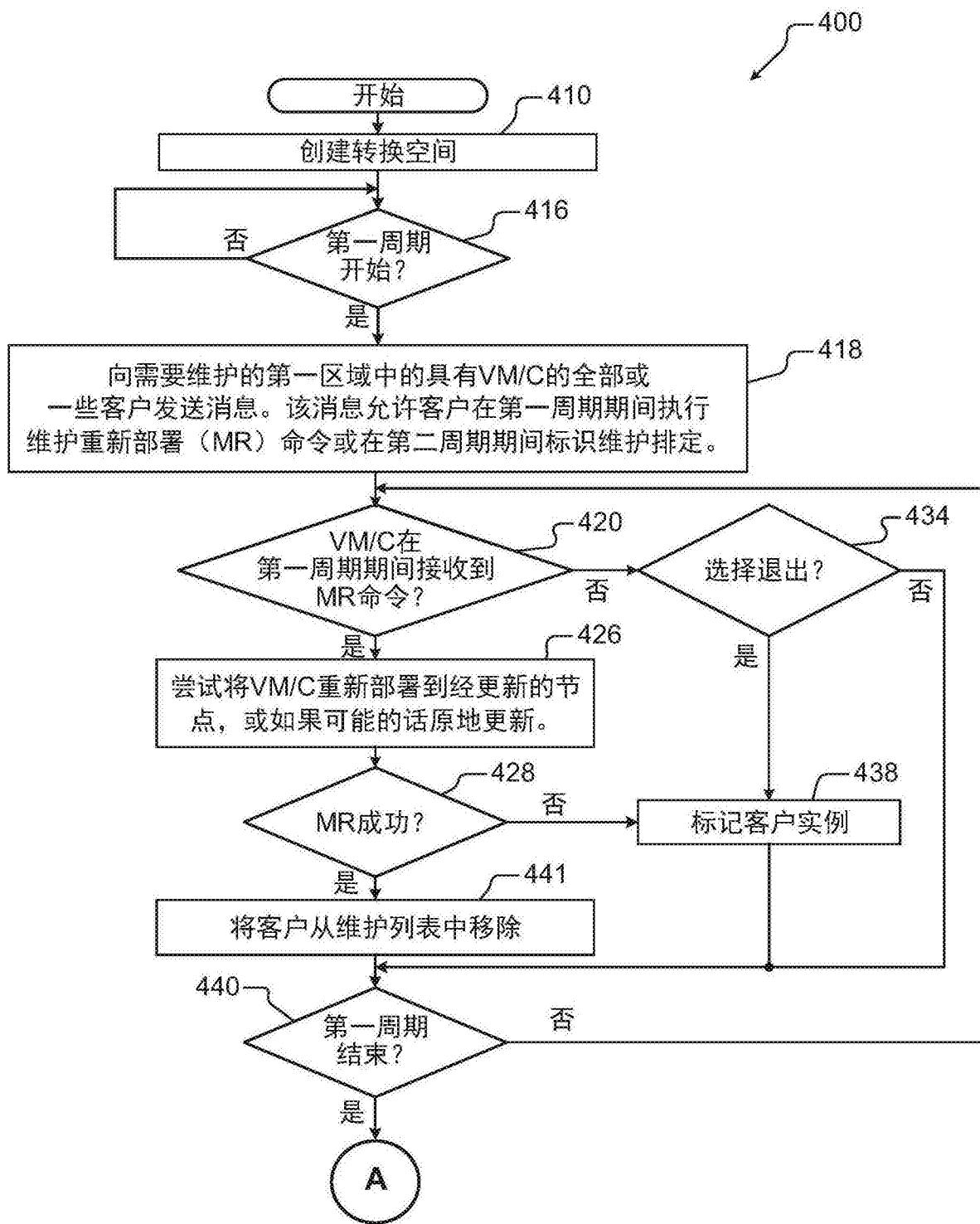


图5A

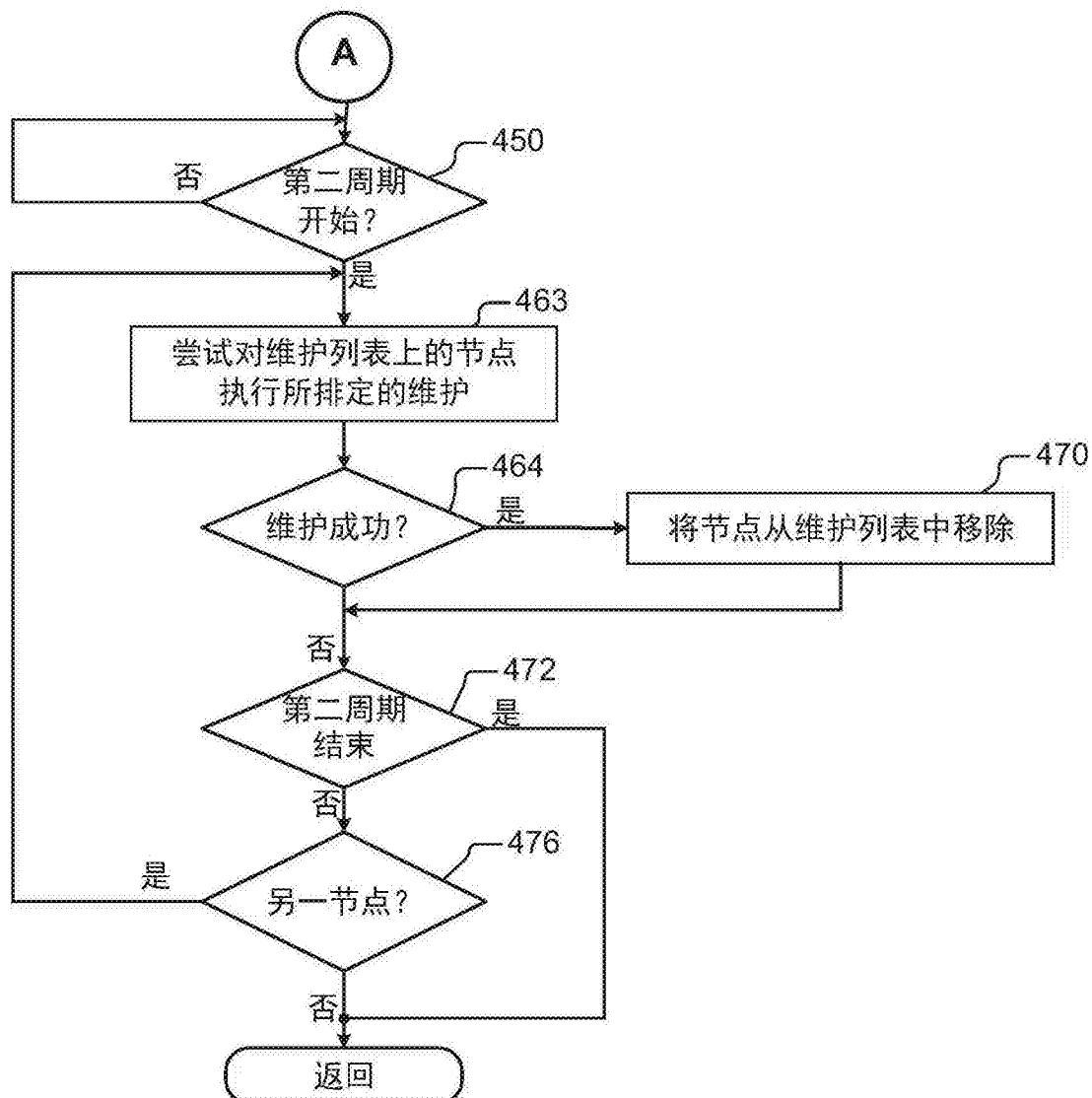


图5B

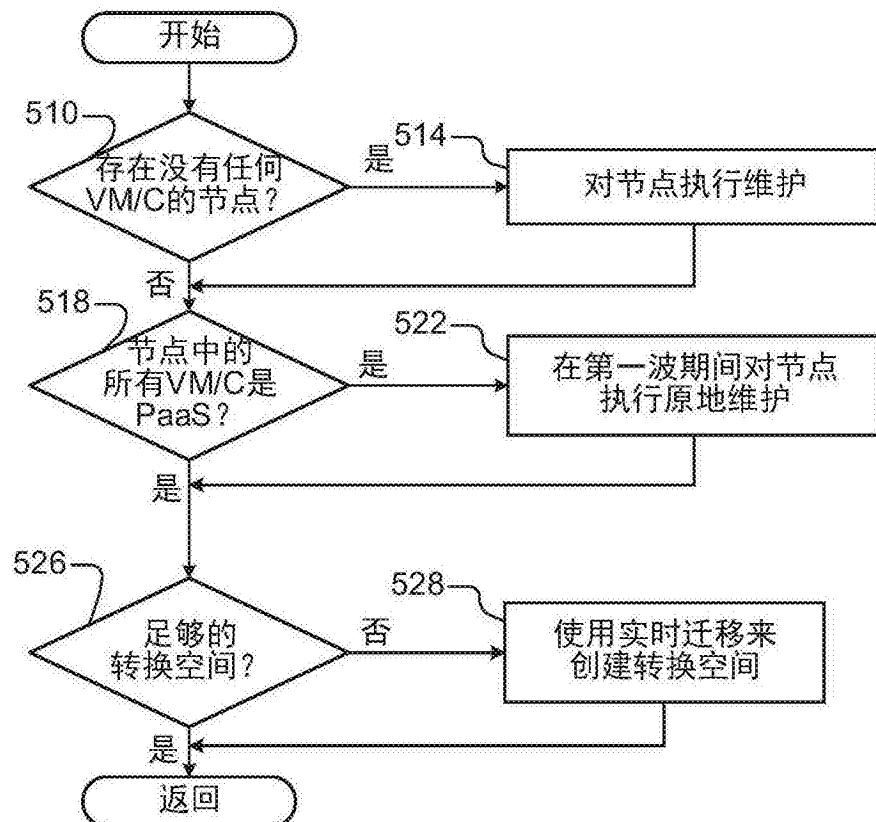


图6A

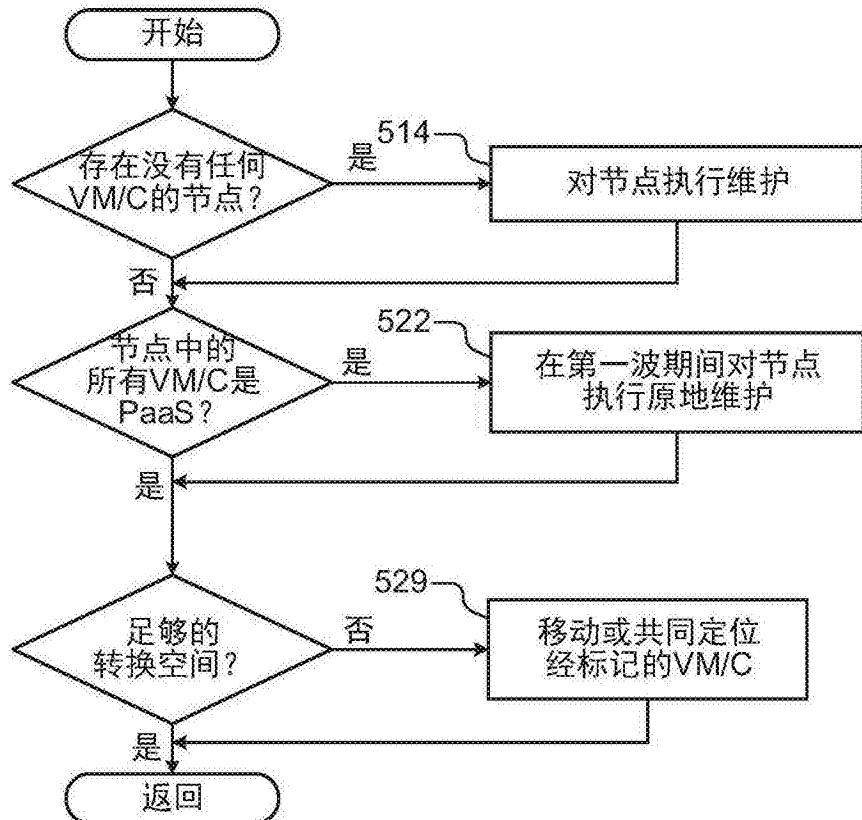


图6B

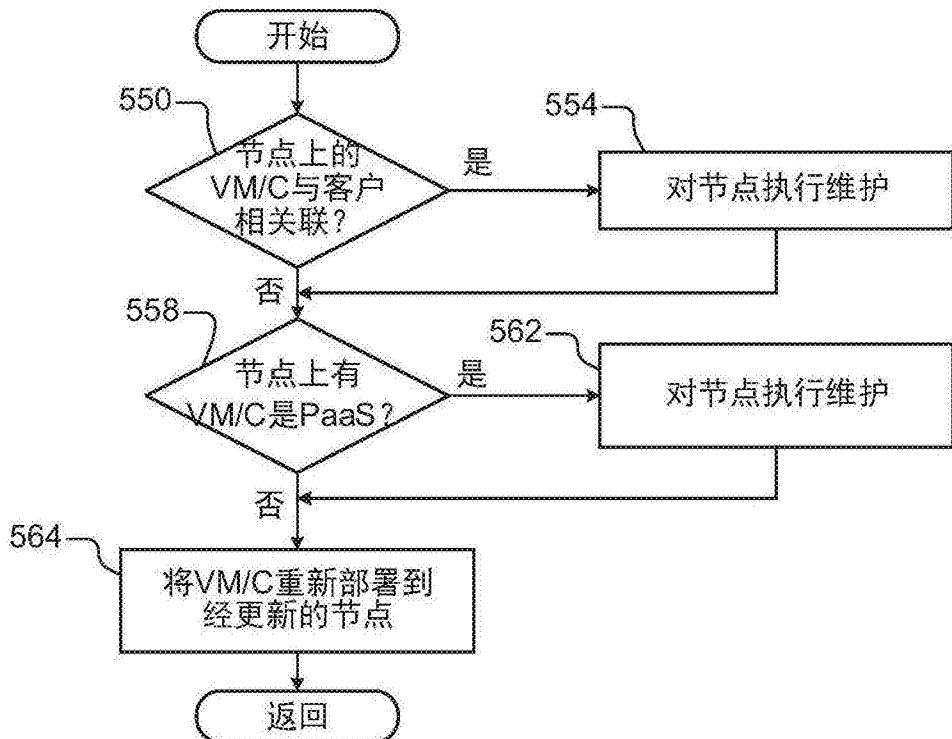


图7

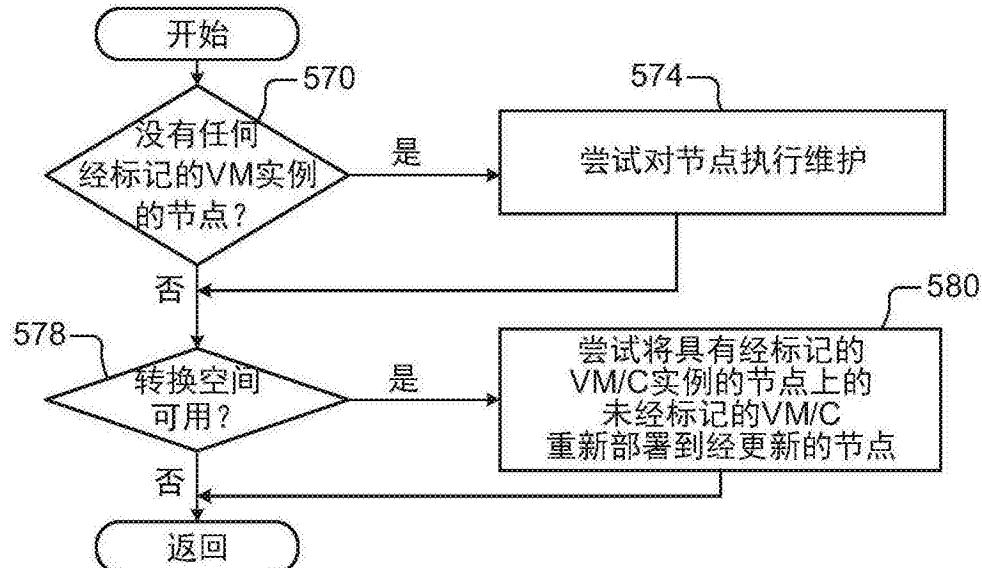


图8

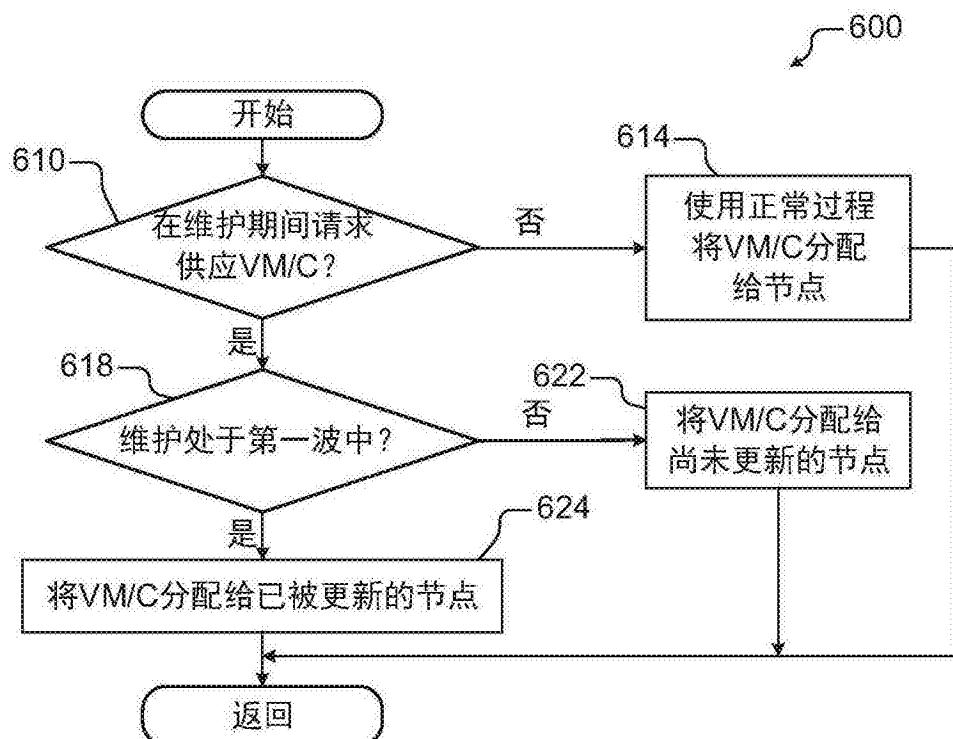


图9

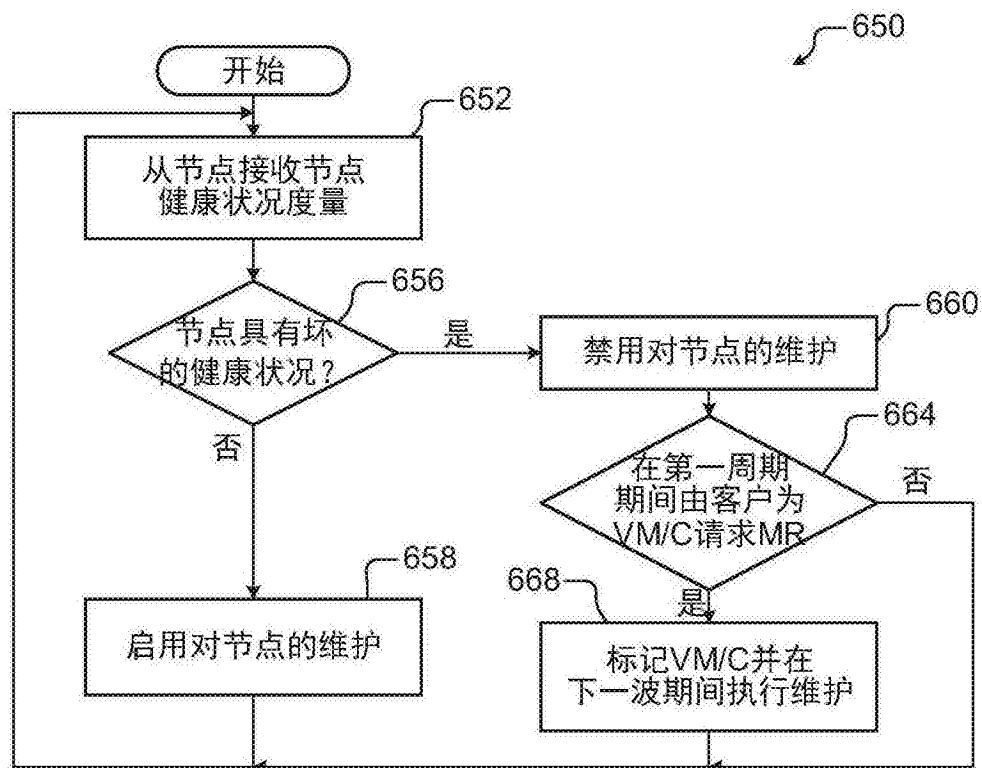


图10

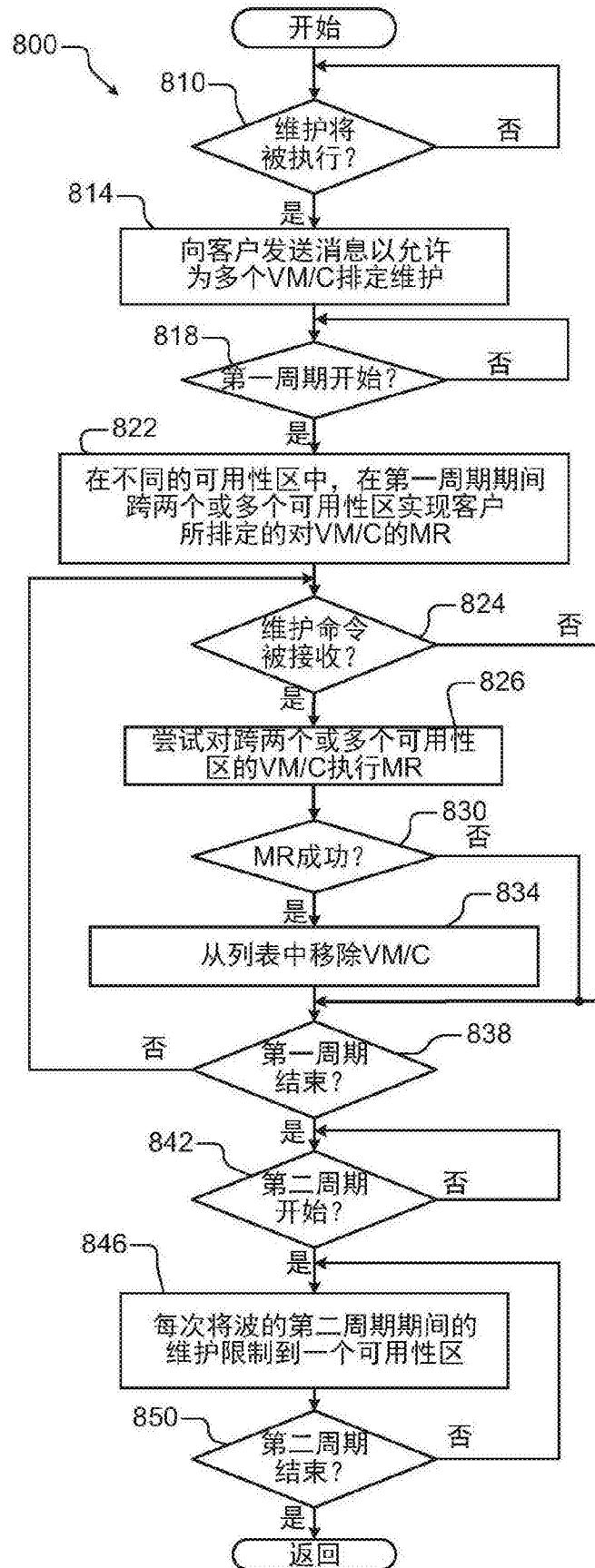


图11