(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 105190558 B
(45)授权公告日 2019.01.29

(21)申请号 201480025790.7
(22)申请日 2014.10.10
(65)同一申请的已公布的文献号
申请公布号 CN 105190558 A
(43)申请公布日 2015.12.23
(30)优先权数据
61/891,190 2013.10.15 US
14/261,141 2014.04.24 US
(85)PCT国际申请进入国家阶段日
2015.11.06
(86)PCT国际申请的申请数据
PCT/US2014/060161 2014.10.10
(87)PCT国际申请的公布数据
WO2015/057525 EN 2015.04.23

(73)专利权人 思科技术公司
地址 美国加利福尼亚州
(72)发明人 鲍伯·梅兰德 哈里什·普塔拉希
(74)专利代理机构 北京东方亿思知识产权代理有限责任公司 11258
代理人 李晓冬

(51)Int.Cl.
G06F 9/50(2006.01)
(56)对比文件
US 2012158967 A1,2012.06.21,
CN 102918499 A,2013.02.06,
US 2008201711 A1,2008.08.21,
US 8028071 B1,2011.09.27,
US 2010205601 A1,2010.08.12,
审查员 胡赢

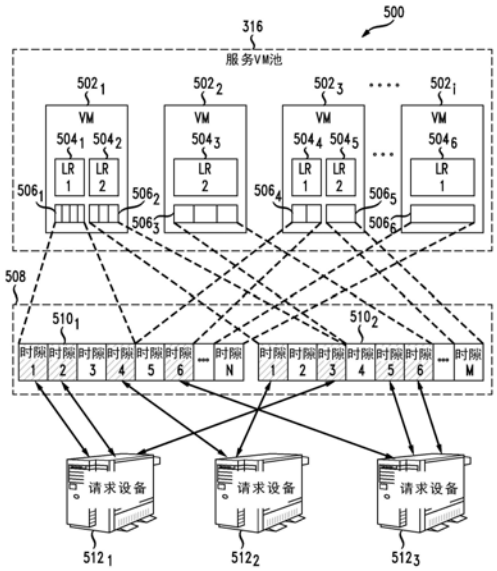
权利要求书3页 说明书15页 附图11页

(54)发明名称

用于创建逻辑资源的方法和系统

(57)摘要

本主题公开涉及用于实例化作为服务虚拟机进行提供的云资源的方法。在一个实施例中，云服务管理系统将多个抽象层时隙中的每一个映射到逻辑资源的虚拟上下文。该虚拟上下文由相应的虚拟机进行托管，该虚拟机是虚拟机池的一部分。该系统标识多个抽象层时隙中的可用抽象层时隙并且保留该时隙，以使得逻辑资源的相应虚拟上下文能够为请求设备服务。该系统然后将可用抽象层时隙标记为不可用。还提供了系统和计算机可读介质。



CN 105190558 B

1.一种方法,包括:

将多个抽象层时隙中的每一个映射到逻辑资源的虚拟上下文,其中,所述虚拟上下文由虚拟机池中相应的虚拟机进行托管;

标识所述多个抽象层时隙中的可用抽象层时隙;

基于所述多个抽象层时隙中可用抽象层时隙的数目来确定可用时隙计数;

保留所述可用抽象层时隙,以使得所述逻辑资源的相应虚拟上下文能够被供应;以及

将所述可用抽象层时隙标记为不可用。

2.根据权利要求1所述的方法,还包括:

从设备接收针对所述逻辑资源的请求;以及

将所述可用抽象层时隙分派给该设备。

3.根据权利要求1所述的方法,其中,所述逻辑资源是虚拟网络资源。

4.根据权利要求3所述的方法,其中,所述逻辑网络资源包括如下各项中的一个:虚拟防火墙、虚拟路由器、虚拟专用网(VPN)、虚拟负载均衡器、虚拟广域网(WAN)优化平台、虚拟深度分组检查器、虚拟流量监控器。

5.根据权利要求1所述的方法,其中,所述多个抽象层时隙中的至少一个抽象层时隙被映射到所述虚拟机池中的某一整个虚拟机。

6.根据权利要求1所述的方法,还包括:

当所述可用时隙计数处于期望范围之外时,执行如下操作之一:

(i)配设至少一个虚拟机并且将该至少一个虚拟机添加到所述虚拟机池,由此由该至少一个虚拟机托管的一个或多个新的虚拟上下文被映射到所述多个抽象层时隙中的一个或多个新的可用抽象层时隙,或者

(ii)从所述虚拟机池中移除至少一个虚拟机,由此被映射到由该至少一个虚拟机托管的虚拟上下文的一个或多个多余的抽象层时隙被从所述多个抽象层时隙中移除;以及

调整所述可用时隙计数。

7.根据权利要求6所述的方法,其中,所述期望范围包括下限和上限,并且其中,所述下限和所述上限之一是基于可用抽象层时隙的目标数目来确定的。

8.根据权利要求1所述的方法,还包括:

当可用抽象层时隙的数目下降到阈值以下时,升起赤字标志;以及

当所述赤字标志被升起时,通过配设对所述逻辑资源的至少一个新的虚拟上下文进行托管的至少一个额外的虚拟机来调整可用抽象层时隙的数目,所述至少一个新的虚拟上下文被映射到所述多个抽象层时隙中的至少一个新的抽象层时隙。

9.根据权利要求1所述的方法,其中,将所述可用抽象层时隙标记为不可用产生了不可用抽象层时隙,该方法还包括:

释放所述不可用抽象层时隙,以使得所述逻辑资源的相应虚拟上下文稍后能够被保留;以及

将所述不可用抽象层时隙标记为可用。

10.根据权利要求9所述的方法,其中,当针对所述逻辑资源的删除触发事件发生时,所述不可用抽象层时隙被释放。

11.一种系统,包括:

处理器；

虚拟机池；以及

存储指令的计算机可读介质，当所述指令被所述处理器运行时使得所述处理器执行操作，所述操作包括：

将多个抽象层时隙中的每一个映射到逻辑资源的虚拟上下文，其中，所述虚拟上下文由所述虚拟机池中相应的虚拟机进行托管；

标识所述多个抽象层时隙中的可用抽象层时隙；

基于所述多个抽象层时隙中可用抽象层时隙的数目来确定可用时隙计数；

保留所述可用抽象层时隙，以使得所述逻辑资源的相应虚拟上下文能够被供应；以及

将所述可用抽象层时隙标记为不可用。

12.根据权利要求11所述的系统，其中，所述计算机可读存储介质存储另外的指令，当这些另外的指令被所述处理器运行时，使得所述处理器执行操作，这些操作还包括：

从设备接收针对所述逻辑资源的请求；以及

将所述可用抽象层时隙分派给该设备。

13.根据权利要求11所述的系统，其中，所述逻辑资源是包括如下项之一的逻辑网络资源：虚拟防火墙、虚拟路由器、虚拟专用网(VPN)、虚拟负载均衡器、虚拟广域网(WAN)优化平台、虚拟深度分组检查器、虚拟流量监控器。

14.根据权利要求11所述的系统，其中，所述多个抽象层时隙中的至少一个抽象层时隙被映射到所述虚拟机池中的某一整个虚拟机。

15.根据权利要求11所述的系统，其中，所述计算机可读存储介质存储另外的指令，当这些另外的指令被所述处理器运行时，使得所述处理器执行操作，这些操作还包括：

当所述可用时隙计数处于期望范围之外时，执行如下操作之一：

(i)配设至少一个虚拟机并且将该至少一个虚拟机添加到所述虚拟机池，由此由该至少一个虚拟机托管的一个或多个新的虚拟上下文被映射到所述多个抽象层时隙中的一个或多个新的可用抽象层时隙，或者

(ii)从所述虚拟机池中移除至少一个虚拟机，由此被映射到由该至少一个虚拟机托管的虚拟上下文的一个或多个多余的抽象层时隙被从所述多个抽象层时隙中移除；以及

调整所述可用时隙计数。

16.根据权利要求15所述的系统，其中，所述期望范围包括下限和上限，并且其中，所述下限和所述上限之一是基于可用抽象层时隙的目标数目来确定的。

17.一种存储指令的非暂态计算机可读存储介质，当所述指令被处理器运行时，使得所述处理器执行操作，所述操作包括：

将多个抽象层时隙中的每一个映射到由虚拟机池中的虚拟机进行托管的逻辑资源；

标识所述多个抽象层时隙中的可用抽象层时隙；

基于所述多个抽象层时隙中可用抽象层时隙的数目来确定可用时隙计数；

保留所述可用抽象层时隙，以使得相应的逻辑资源能够被供应；以及

将所述可用抽象层时隙标记为不可用。

18.根据权利要求17所述的非暂态计算机可读存储介质，其存储另外的指令，当这些另外的指令被所述处理器运行时，使得所述处理器执行操作，这些操作还包括：

当可用抽象层时隙的数目下降到阈值以下时，升起赤字标志；以及

当所述赤字标志被升起时，通过配设对所述逻辑资源进行托管的至少一个额外的虚拟机来调整可用抽象层时隙的数目，所述逻辑资源被映射到所述多个抽象层时隙中的至少一个新的抽象层时隙。

19.根据权利要求17所述的非暂态计算机可读存储介质，其中，将所述可用抽象层时隙标记为不可用产生了不可用抽象层时隙，所述非暂态计算机可读存储介质存储另外的指令，当这些另外的指令被所述处理器运行时，使得所述处理器执行操作，这些操作还包括：

释放所述不可用抽象层时隙，以使得所述相应的逻辑资源稍后能够被保留；以及

将所述不可用抽象层时隙标记为可用。

20.根据权利要求19所述的非暂态计算机可读存储介质，其中，当针对所述逻辑资源的删除触发事件发生时，所述不可用抽象层时隙被释放。

用于创建逻辑资源的方法和系统

[0001] 相关申请

[0002] 本申请要求于2013年10月15日递交的美国临时专利申请No.61/891,190的优先权,该申请的整体通过引用被合并到本文。

技术领域

[0003] 本主题的技术与用于实例化云资源的方法有关,这些云资源作为服务虚拟机进行提供。具体地,本技术的某些方面提供用于对逻辑资源进行近乎实时的创建的系统和方法,这些逻辑资源被托管于云计算环境中的服务虚拟机上。

背景技术

[0004] 通过虚拟机技术,云计算通过允许客户(也被称为"租户")使用服务提供商的虚拟化计算资产(例如,虚拟处理器、虚拟存储装置以及虚拟网络资源)而非必须购买并且完全拥有所有必需的装置来改变基于网络的服务的格局。应当注意,云计算提供商根据若干种基本模型(包括例如,基础设施即服务(IaaS)和平台即服务(PaaS))来提供其服务。传统地,IaaS已经提供了诸如虚拟机(VM)、虚拟网络或虚拟存储装置之类的逻辑基础设施资源,而PaaS已经提供了具有更高抽象等级的资源。然而,多年以来,IaaS与PaaS之间的界线变得日益模糊。

[0005] 用于基础设施即服务(IaaS)和平台即服务(PaaS)环境中的云服务管理(CSM)系统可以向其租户提供诸如虚拟路由器、虚拟防火墙之类的逻辑网络资源。在IaaS和PaaS二者中,逻辑资源通过云API(例如,Amazon® Web服务API和Openstack® API)而变得可用。在封面的背后,可以以各种方法来实现这些资源,例如,使用物理设备或这些设备中的虚拟上下文、以及使用VM或传统的软件。典型地,使用上面提及的方法的组合。

[0006] 与使用专用物理设备的情况相比较,当使用VM来实现云服务中的逻辑资源时,创建必需的逻辑资源需要大量的时间。具体地,物理机器典型地被预配设并且随时可供使用,而逻辑资源通常按需创建。因此,逻辑资源在使得托管资源的服务VM准备就绪并处在服务中的方面遭受时间损失。这额外的准备时间可以包括但不限于:(a)用于选择满足客户的需求的恰当的主机的时间、(b)用于创建VM资产的时间、(c)用于将启动映像复制给主机的时间、以及(d)用于自展启动映像的时间。

[0007] 另一方面,由于对这些资源的需求的高交互性和动态性,因此租户对于这些逻辑资源可能具有不同种预期。例如,当web服务器在网络业务中突然遭受意外的尖波时,租户可能期望在几秒(而不是接下来的半小时)之内实例化并部署附加的资源(例如,虚拟路由器)。这样的延迟是不可取的,因为它们降低了用户体验并且使得使用云服务的应用服务设计更加复杂。

发明内容

[0008] 根据本公开的一个实施例,提供了一种方法,包括:将多个抽象层时隙中的每一个

映射到逻辑资源的虚拟上下文，其中，所述虚拟上下文由虚拟机池中相应的虚拟机进行托管；标识所述多个抽象层时隙中的可用抽象层时隙；基于所述多个抽象层时隙中可用抽象层时隙的数目来确定可用时隙计数；保留所述可用抽象层时隙，以使得所述逻辑资源的相应虚拟上下文能够被供应；以及将所述可用抽象层时隙标记为不可用。

[0009] 根据本公开的另一实施例，提供了一种系统，包括：处理器；虚拟机池；以及存储指令的计算机可读介质，当所述指令被所述处理器运行时使得所述处理器执行操作，所述操作包括：将多个抽象层时隙中的每一个映射到逻辑资源的虚拟上下文，其中，所述虚拟上下文由所述虚拟机池中相应的虚拟机进行托管；标识所述多个抽象层时隙中的可用抽象层时隙；基于所述多个抽象层时隙中可用抽象层时隙的数目来确定可用时隙计数；保留所述可用抽象层时隙，以使得所述逻辑资源的相应虚拟上下文能够被供应；以及将所述可用抽象层时隙标记为不可用。

[0010] 根据本公开的又一实施例，提供了一种存储指令的非暂态计算机可读存储介质，当所述指令被处理器运行时，使得所述处理器执行操作，所述操作包括：将多个抽象层时隙中的每一个映射到由虚拟机池中的虚拟机进行托管的逻辑资源；标识所述多个抽象层时隙中的可用抽象层时隙；基于所述多个抽象层时隙中可用抽象层时隙的数目来确定可用时隙计数；保留所述可用抽象层时隙，以使得相应的逻辑资源能够被供应；以及将所述可用抽象层时隙标记为不可用。

附图说明

[0011] 在所附权利要求中提出了本主题技术的某些特征。然而，所包括的用来提供进一步理解的附图阐述了所公开的方面并且与说明书一起用来解释本主题技术的原理。在附图中：

[0012] 图1是包括通过各种通信方法互连的节点/设备的示例计算机网络的示意性框图；

[0013] 图2是示例简化计算设备的示意性框图；

[0014] 图3是示出云服务管理系统的示例的示意性框图；

[0015] 图4是示出以映射到抽象层的虚拟机为特征的示例系统的示意性框图；

[0016] 图5是示出以服务VM池、抽象层以及客户端设备为特征的另一示例系统的示意性框图；

[0017] 图6根据一些实现方式示出了一些可用资源的所期望的范围的示例；

[0018] 图7A-图7D是示出了示例调度功能操作的示意性框图；

[0019] 图8示出了用于创建逻辑资源的示例方法；

[0020] 图9示出了用于执行VM池维护的示例方法；

[0021] 图10示出了用于创建逻辑资源的另一示例方法；以及

[0022] 图11示出了用于删除逻辑资源的示例方法。

具体实施方式

[0023] 1.概述

[0024] 在一个实施例中，系统可以将每个抽象层时隙映射到逻辑资源的虚拟上下文，其中，每个虚拟上下文由来自虚拟机池中的虚拟机进行托管。该系统然后可以识别来自抽象

层时隙中可用的抽象层时隙，并且保留可用的抽象层时隙，以便逻辑资源的相应虚拟上下文能够被供应。接下来，该系统可以将可用的抽象层时隙标记为不可用。

[0025] 2.具体描述

[0026] 下面提出的具体描述旨在作为本主题技术的各种配置的描述，并且不意为仅仅表示本主题技术可以被实施的配置。附图被包括在本文中，并且构成了具体描述的一部分。具体描述包括用于提供对本主题技术更加透彻的理解的具体细节。然而，清楚且显而易见的是，本主题技术不限于本文提出的具体细节，并且可以在无需这些细节的情况下被实施。在一些实例中，结构和组件以框图的形式示出，以免模糊本主题技术的概念。

[0027] 鉴于上面所提出的关于服务VM的实例化的问题，所需要的是当VM用于实现逻辑网络资源时减少资源创建时间的方法。本主题技术通过维护预创建服务VM的备用池解决了上述需求，其中，这些预创建的服务VM在创建之后空闲运行或休眠。换言之，本文提出的各种实施例可以减少或消除涉及如下各项的等待时间：(a)选择主机、(b)创建VM资产、(c)复制启动映像、和/或(d)加载启动映像。服务VM托管各种逻辑网络资源，这些逻辑网络资源然后可以在租户请求逻辑网络资源的任何时候由云系统管理(CSM)系统进行分配和提供。这不仅允许CSM以显著减少的实例化时间来提供逻辑资源，而且其还使得该实例化时间更加可预测且统一。

[0028] 可以通过引入抽象层将该处理进一步流线化，该抽象层以虚拟“时隙”的形式位于逻辑资源与后台资源(即，VM)之间。由于给定的VM可以托管多于一个逻辑资源的虚拟上下文，因此VM上的各个虚拟上下文可以被映射到不同的时隙。替代地，如果VM仅具有一个虚拟上下文，则整个VM可以被映射到单个时隙。由于抽象层降低了关联于与VM进行接合的粒度的等级，因此其有助于简化CSM的任务并且降低管理VM池时引入错误的可能性。

[0029] 此外，CSM可以通过跟踪空闲时隙的数目来将服务VM池维护在其最优大小。例如，如果期望的一组空闲时隙为S，其中，S>0，则空闲时隙的期望范围DR可以被表示为，DR＝INT([f_1(S,...),f_2(S,...)])，其中，f_1和f_2是确定期望范围的下限和上限的函数。当发现空闲时隙的数目处于期望范围之外时，则CSM可以决定启动另外的服务VM或者销毁过量的服务VM以防止池的大小变得太小或太大。CSM可以响应于各种状况而执行这样的维护操作，例如，当租户请求新的资源时、当租户放弃资源时、和/或基于周期而不考虑资源请求。

[0030] 计算机网络是在地理上分布的节点的集合，这些节点通过通信链路和分段进行互连，以在端节点(例如，个人计算机和工作站)之间传输数据。从局域网(LAN)到广域网(WAN)的范围内的许多类型的网络是可用的。LAN通常通过位于大致相同的物理位置(例如，建筑物或校园)的专用私有通信链路来连接节点。另一方面，WAN一般通过长距离通信链路(例如，公共载波电话线、光学光路、同步光网络(SONET)、或同步数字体系(SDH)链路)来连接地理上分散的节点。

[0031] 互连网是WAN的示例，其连接世界范围内分散的网络，从而提供各种网络上的节点之间的全球通信。这些节点典型地根据预定协议(例如，传输控制协议/互连网协议(TCP/IP))通过交换离散帧或数据分组来经由网络进行通信。在该上下文中，协议包括一组定义节点如何相互交互的规则。计算机网络还可以通过中间网络节点(例如，路由器)进行互连，以扩展每个网络的有效“大小”。

[0032] 云计算一般可以被定义为基于互连网的计算，其中，计算资源经由网络(例如，

“云”)按需从可用资源集合中被动态配设并且分配给客户端或用户计算机或者其他设备。云计算资源例如可以包括任意类型的资源,例如,计算、存储、以及网络设备、虚拟机(VM)等。例如,资源可以包括服务设备(防火墙、深度分组检查器、流量监控器等)、计算/处理设备(服务器、CPU、存储器、强力处理能力)、存储设备(例如,网络附连存储装置、存储区域网络设备)等,并且可以用于虚拟机(VM)、数据库、应用(App)等的实例化。

[0033] 云计算资源可以包括“私有云”、“公共云”和/或“混合云”。“混合云”是包括通过技术而结盟或进行互操作的两个或更多个云的云基础设施。本质上,混合云是私有云与公共云之间的交互,其中,私有云加入公共云并且以安全和可扩展的方式来利用公共云资源。

[0034] 图1是示例计算机网络100的示意性框图,该示例计算机网络100说明性地包括通过各种通信方法进行互连的节点/设备。例如,链路可以是有线链路或共享介质(例如,无线链路等),其中,某些节点可以基于物理连接或基于距离、信号强度、当前操作状态、位置等来与其他节点进行通信。本领域技术人员将理解,任意数目的节点、设备、链路等可以用于该计算机网络中,并且本文所示的视图是出于简化的目的。

[0035] 具体地,设备“A”和“B”可以包括具有处理和/或存储能力的任意设备,例如,个人计算机、移动电话(例如,智能电话)、游戏系统、便携式个人计算机(例如,膝上型计算机、平板计算机等)、机顶盒、电视机、交通工具等,并且设备“A”和“B”可以与网络160(互连网或私有网络)进行通信以到达云150。此外,一个或多个服务器(服务器A和B)、网络管理服务器(NMS)、控制中心等也可以与网络160进行互连(或者位于网络160之内)从而到达云150。

[0036] 云150可以是公共云系统、私有云系统、和/或混合云系统。云150包括多个资源(例如,防火墙197、负载均衡器193、(一个或多个)WAN优化平台195、(一个或多个)设备200、(一个或多个)服务器180以及(一个或多个)虚拟机(VM)190)。云资源可以是物理资源和虚拟资源的组合。可以基于来自一个或多个客户端的请求来配设云资源。客户端可以是一个或多个设备(例如,设备A和/或B)或者一个或多个服务器(例如,服务器A和/或B)。

[0037] 可以适当地使用预定的网络通信协议(例如,某些已知的有线协议、无线协议或其他协议)在计算机网络100的节点/设备之间交换数据分组(例如,流量和/或消息)。在该上下文中,协议包括一组定义节点如何相互交互的规则。

[0038] 图2是示例简化计算设备200的示意性框图,该计算设备200可以例如作为服务器180或作为一个或多个设备的表示(如VM 190)与本文所描述的一个或多个实施例一起使用。说明性“设备”200可以包括通过系统总线250进行互连的一个或多个网络接口210、至少一个处理器220和存储器240。(一个或多个)网络接口210包含用于通过耦合到网络100的链路来传输数据的机械电路、电气电路和信令电路。网络接口210可以被配置为使用本领域技术人员所理解的各种不同的通信协议来发送和/或接收数据。存储器240包括用于存储于本文所描述的实施例相关联的软件程序和数据结构且可由处理器220寻址的多个存储位置。处理器220可以包括适于运行软件程序并且操纵数据结构245所必需的元件或逻辑。操作系统242(其部分典型地驻留在存储器240中并且由处理器运行)尤其通过调用支持在设备上运行的软件处理和/或服务的操作来在功能上组织设备。这些软件处理和/或服务可以包括说明性“虚拟资源实例化”处理248,如本文所描述的。

[0039] 本领域技术人员将理解的是,其他处理器和存储器类型(包括各种计算机可读介质)可以被用来存储和运行涉及本文所描述的技术的程序指令。此外,尽管该描述对各种处

理进行了阐述，但应明确认为，各种处理可以被实现为配置为根据本文的技术(例如，根据类似的处理的功能)进行操作的模块。而且，尽管已经独立地示出了处理，但本领域技术人员将理解的是，处理可以是其他处理内的例程或模块。例如，处理器220可以包括一个或多个可编程处理器(例如，微处理器或微控制器、或固定逻辑的处理器)。在可编程处理器的情形中，任何相关联的存储器(例如，存储器240)可以是编码有或存储可以实现程序模块(例如，其上编码有资源分配处理的模块)的指令的任意类型的有形处理器可读存储器(例如，随机存取、只读等)。

[0040] 处理器220还可以包括固定逻辑处理设备(例如，专用集成电路(ASIC)或数字信号处理器)，其被配置有包括能够使得处理器执行本文所描述的功能的逻辑或指令的固件。因此，程序模块可以被编码在一个或多个有形计算机可读存储介质(例如，具有固定逻辑或可编程逻辑(例如，由处理器运行的软件/计算机指令))中用于运行，并且任意处理器可以是可编程处理器、可编程数字逻辑(例如，现场可编程门阵列、或包括固定数字逻辑的ASIC、或其组合)。一般地，任何处理逻辑可以被编码在处理器或计算机可读介质中，该处理器或计算机可读介质编码有由处理器运行的指令，当这些指令由处理器运行时可操作使得处理器执行本文所描述的功能。

[0041] 图3示出了云服务管理(CSM)系统的示例。示例CSM系统302可以管理由VM池316中的VM托管的逻辑资源并且将其提供给任意客户端设备314。在该方面，CSM系统302可以根据客户端设备314的当前需求和未来需求来实例化和销毁各种逻辑资源。

[0042] CSM系统302可以包括若干个子组件，例如，调度功能304、云服务应用编程接口(API)306、池管理(PM)功能308、VM管理(VMM)功能310以及抽象层312。CSM系统302的各种组件可以被实现为硬件和/或软件组件。而且，尽管图3示出了CSM系统302的各种组件的一个示例配置，但本领域技术人员将理解可以以许多不同的方式来配置这些组件。例如，PM 308和VMM 310可以属于一个软件模块而不是两个独立的模块。其他模块可以被组合或者进一步被划分为更多的子组件。

[0043] CSM系统302可以通过其网络接口(未示出)与各个客户端设备314(也被称为租户)进行通信。例如，客户端设备314可以从CSM系统302请求各种服务，包括请求一个或多个逻辑资源。CSM系统302转而可以访问并且操纵VM池316和/或包含在VM池316内的独立VM，以向客户端设备314提供任何所请求的服务。在CSM系统的监管下，客户端设备314也可以直接访问并利用包含在VM池316内的一些VM，以利用由其托管的逻辑资源。客户端设备314可以是通过CSM系统302需要额外云资源的服务器、终端、虚拟机、网络设备等。

[0044] VM池316也被称为服务VM池，其为可以托管各种逻辑资源的一个或多个虚拟机的集合。换言之，VM池316可以是准备好的(即，被创建并正在运行的)、闲置或休眠的服务VM的“备用”池。如其名称所隐含的，虚拟机是虚拟化的或者仿真的计算环境，其主要通过软件来实现，但其通常包括软件和硬件组件二者。通过虚拟化技术，一个物理计算设备(例如，服务器)可以(并行地)运行多个虚拟机。每个虚拟机可以在与彼此和/或主机设备不同的操作系统(OS)上运行。每个VM可以具有其自己的上下文、存储装置、通信接口等。服务VM是可以用于在后台实现网络服务的虚拟机。根据其上所加载的网络操作系统的类型，服务VM可以提供多种不同类型的网络服务。在该上下文中，服务VM对于客户端/租户而言可以是不可见的，并且对于客户的明确请求可能不是不可用的。此外，尽管服务VM可以被装备有其他VM可

以附连的虚拟端口，但服务VM在由客户端创建的VM之间可能不是可见的。VM池316中的活动VM的数目可以根据客户端设备314的需求等级而被动态调整，以使得仅仅最小数目或者最优数目的VM在任意给定的时刻是运作的。这有助于缩减能量开销以及维护基于云的基础设施所需的资源量。

[0045] VM池316中的VM可以在其使用之前被创建和启动，以使得其可以在需求产生时能更快地被部署。例如，当客户端设备314之一从CSM系统300请求逻辑资源的实例(例如，虚拟路由器)时，CSM系统300不是从头开始配设新的VM，而是可以简单地选择并分派由VM池316中的VM之一托管的逻辑资源的实例以进行更快速的部署。

[0046] 属于VM池316的单个VM和/或集体VM可以形成用于托管和提供包括逻辑资源的各种云服务的后台基础设施。换言之，逻辑资源可以通过服务VM的方式在云提供商后台处实现。逻辑资源是运作起来很像其硬件配对物的基于软件的资源。逻辑资源可以是虚拟的网络资源。例如，由服务VM托管的虚拟路由器将具有与物理路由器类似的接口及其相关联的行为。从与资源进行交互的客户端设备的角度看，在使用逻辑资源与使用物理资源之间可能仅存在可忽略的差别。逻辑资源的类型可以包括但不限于防火墙、路由器、虚拟专用网(VPN)、负载均衡器、WAN优化器、深度分组检查器、流量监控器等。

[0047] 单个服务VM可以托管多于一个的逻辑资源实例。也就是说，VM对于给定的逻辑资源可以具有相互独立运行的一个或多个虚拟上下文。虚拟上下文可以独立于VM的全局上下文。例如，VM路由器可以具有八个不同的虚拟上下文，每个具有其自己的一组环境变量、状态、配置、用户偏好等。虚拟上下文的另一示例是虚拟路由和转发(VRF)。每个虚拟上下文可以被分派给不同的客户端设备。在一些实例中，多于一个的虚拟上下文可以被分派给同一客户端设备。尽管驻留在同一VM中的虚拟上下文可以共享该VM的相同硬件资源(例如，处理器、存储器、总线、存储装置等)，但从单个客户端设备314的角度来看，每个虚拟上下文实质上类似于不同的物理资源来运行。因此，例如，具有128个虚拟上下文的VM防火墙可以在逻辑上等同于具有128个物理防火墙设备。

[0048] 而且，一个服务VM可以托管多于一种类型的逻辑资源，每种逻辑资源可能具有多于一个虚拟上下文。例如，对于单个虚拟机而言，托管虚拟路由器的四个虚拟上下文和虚拟负载均衡器的六个虚拟上下文是可能的。因此，逻辑资源不一定基于一对一来映射到VM。而且，托管一种类型的逻辑资源的VM可以被重新配设为托管不同类型的逻辑资源。例如，如果CSM系统302确定客户端设备314的需求是需要更多的虚拟路由器且更少的虚拟防火墙，则CSM系统302可以停止使用VM池316中正提供防火墙服务的一些VM并且将这些VM重新用作托管虚拟路由器的示例。

[0049] 客户端设备314可以通过云服务API 306来与CSM系统302进行通信。面向租户的云服务API 306可以包括可用于每个客户端设备314请求服务、发送/接收数据、操作资源等的各种功能、例程、方法等。例如，客户端设备可以使用云服务API 306来从CSM系统302请求逻辑资源、取消请求、放弃资源等。因此，云服务API 306在涉及VM池316的维护和VM的分配的工作流程中十分重要。

[0050] 抽象层312可以位于逻辑资源与后台资源(即，实现逻辑资源的VM)之间。抽象层312可以利用软件、硬件、或二者的组合来实现。尽管图3将抽象层312示出为CSM 302的一部分，但抽象层312可以位于CSM系统302之外。例如，抽象层312可以是VM池316或者VM池316中

的单个VM的一部分。抽象层可以具有其自己的一组API命令，CSM 302可以使用该组API命令来与VM池316中的服务VM进行接合。抽象层312允许CSM 302更加有效地利用VM提供的资源，因为典型的VM提供的粒度等级可能相当高，而不具有这样的额外的抽象层。换言之，通过隐藏VM池316中的VM的一些技术细节，抽象层312允许CSM 302更加有效地管理VM池316。

[0051] 抽象层312对于CSM系统302隐藏这些细节的方式可以通过使用虚拟“时隙”。类似于数据网络设备中存在的物理时隙的时隙为符号和逻辑的隐喻，其可以用于管理由VM托管的逻辑资源的各个方面。每个时隙可以被映射到逻辑资源。替代地，当适用时，时隙可以被映射到VM中的虚拟上下文。尤其在VM只具有一个虚拟上下文的情况下，时隙还可以被映射到整个VM本身。CSM系统302可以使用该虚拟时隙隐喻来向客户端设备分派可以被映射到逻辑资源的时隙，从而客户端设备可以对所映射的资源进行排他访问。

[0052] 当时隙被映射到逻辑资源或逻辑资源的虚拟上下文时该时隙为空闲的或可用的，但其不被分派给客户端设备。换言之，一旦CSM 302将时隙分派给客户端设备，则该时隙变得不可用，并且直到该时隙再次变得可用之前，其他设备不能使用该特定逻辑资源或其虚拟上下文。例如，当特定的服务VM出现并运行时，其可以提供X个空闲时隙，其中，X是VM可以托管的最多的虚拟上下文的数目。如果VM可以托管32个虚拟上下文，则X＝32。在另一方面，如果整个VM被映射到单个时隙，则X＝1。然后，当被映射到时隙之一的逻辑资源被分派给客户端设备时，VM剩下X-1个空闲时隙。随后，当时隙变得再次可用时(例如，由于客户端设备不再需要该时隙)，VM将再次具有X个可用时隙。出于标识的目的，可以向各个时隙设定序列号或者名称。

[0053] 而且，CSM 304可以具有多于一组时隙，或者替代地，多于一组抽象层，以分别跟踪不同类型的逻辑资源。例如，CSM系统304可以具有一个抽象层，该抽象层具有一组用于管理VM池316中所有虚拟路由器的时隙，并且CSM系统304可以具有不同的抽象层，该不同的抽象层具有其自己一组用于管理虚拟防火墙的时隙。多个抽象层或时隙组可以被分等级地排列。例如，VM池316中的虚拟路由器VM可以具有其自己的时隙组，并且CSM 302可以维护较高等级的抽象层，该较高等级的抽象层巩固各组时隙，如图5所示。

[0054] 调度功能(SCH)304可以主要负责管理抽象层312中的虚拟时隙。具体地，SCH 304可以将各个服务VM、逻辑资源以及虚拟上下文映射到时隙，并且当客户端设备314请求服务时将这些时隙中的一些时隙经由云服务API 306进行分派。当CSM系统302接收到来自客户端设备的新的服务请求时，SCH 304选择空闲时隙(以及为该时隙负责的VM)以提供所请求的逻辑资源。SCH 304可以尝试维护抽象层312中所期望的一组空闲时隙S，其转换为VM池316中所期望的一些可用资源，其中，大小S>0。

[0055] SCH 304可以尝试使得空闲时隙的实际数目S_A保持在期望范围DR之内。例如，期望范围DR可以由如下公式来表示DR＝INT({f_1(S,...),f_2(S,...)])，其中，f_1和f_2是S和确定期望范围的下限和上限的任意其他相关参数的函数，以使得0＜f_1(S,...)≤f_2(S,...)。其他参数例如可以是目前由CSM 302服务的客户端设备314的数目、来自客户端设备314的预计服务需求、服务请求的数目、资源请求速率、时间、VM池316目前的大小、VM池316的最大容量、VM的平均配设时间(例如，启动时间)、所请求的逻辑资源类型之间的比例等。

[0056] 这些各种参数可以作为确定可用资源的理想数目以及其他边限的因素。在一方面，上限和下限可以由函数f_1＝S-M和f_2＝S+M来定义，其中，M是可配置的边限。其他更精确

的公式可以用来确定更令人满意的边限。在一个实施例中，当CSM 302正被初始化时，VM池316可以被填充以其期望的大小S，然而，(例如，在从客户端设备314接收各种请求以及向客户端设备314提供服务的过程中)一旦实际空闲时隙的数目S_A落入期望范围DR之外，则CSM 302可以通过配设更多VM来添加更多的空闲时隙或者通过从VM池316中移除VM来移除过量的空闲时隙。

[0057] 可选地，SCH 304可以具有赤字标志(未示出)，其可以被“升起”以表示可用时隙的数目已经下降到期望范围以下并且需要对时隙进行相应调整。在一个实施例中，赤字标志被连接到跟踪可用时隙的数目的物理传感器或者输入设备。在另一实施例中，利用软件来实现赤字标志。在又一实施例中，赤字标志包括硬件组件和软件组件二者。标志可以是布尔(Boolean)变量。SCH 304可以具有多于一个的赤字标志来跟踪不同组的虚拟时隙。SCH 304还可以依赖于其他类型的逻辑标志来向CSM系统302的其他组件(例如，池管理功能(PM)308和VMM 310)发送关于调度功能304和/或抽象层312的各种状态的信号。例如，SCH 304可以使用标志来指示VM池316具有太多的运行VM。一旦解决了与升起的标志有关的问题，该标志可以被SCH 304或CSM系统302的其他组件“降落”。

[0058] 一旦空闲时隙的数目落入期望范围之外，池管理功能(PM)308可以向备用服务VM池316添加或移除实例，该备用服务VM池316尝试维护准备好进行部署的大约S个空闲时隙。添加或移除空闲时隙的指令可以由SCH 304发布。在另一实施例中，PM 308可以检测到赤字标志或任何其他标志被升起，然后确定其自己的空闲时隙的数目可能需要调整。PM功能308可以静态运行(即，只运行固定次数或以预定计划来运行)或者其可以动态运行(即，连续地运行或每当产生需要时运行)。对于该目的，PM功能308可以接收诸如资源请求速率之类的输入。

[0059] 优选地，每当有来自客户端设备304的请求时，PM功能308可以运行。例如，在向客户端设备314分派时隙或者释放时隙之后，PM 308可以运行其自己的维护例程以确保VM池的大小保持在期望边界之内。可以当创建或删除逻辑资源时进行维护。也可以周期性地进行维护。因此，池管理和逻辑资源的调度不需要紧密耦合。而且，PM 308可以考虑输入、参数和测量(例如，资源请求速率)并且在台后增加或减少VM池316的大小，目的在于对于可能请求逻辑资源的任何租户设备保持足够的可用逻辑资源。

[0060] 虚拟机管理功能(VMM)310可以由PM 308或CSM系统302的其他组件要求来创建和删除服务VM。VMM 310能够直接与VM池316中的单个VM进行接合，以创建、配置、配设、操纵和删除VM。VMM 310可以对VM进行启动、设置和安装应用以及将其关闭。在该方面，VMM 310的操作与抽象层312紧密相关。替代地，VMM 310可以是隐藏关于VM的操作的粒度细节的抽象层312的一部分。

[0061] 图4是示出以映射到抽象层408的虚拟机402为特征的示例系统400的框图。VM 402可以是图3中示出的VM池316的一部分。在一个实施例中，抽象层408是CSM系统302的一部分。在另一实施例中，抽象层408由虚拟机402自己进行管理。抽象层408可以纯粹基于软件。虚拟机402可以被配置为托管一个或多个逻辑资源404(只示出了一个逻辑资源)。逻辑资源404可以是虚拟网络资源，例如，防火墙、路由器、虚拟专用网(VPN)、负载均衡器、广域网(WAN)优化器、深度分组检查器、流量监控器等。

[0062] 每个逻辑资源404中可以具有一个或多个虚拟上下文406_1、406_2、406_3，…，406_N(统

称为“406”)，该一个或多个虚拟上下文406可以作为独立的逻辑资源彼此独立地进行操作。虚拟上下文406可以被映射到时隙410_1、410_2、410_3,…,410_N(统称为“410”)。当另外的虚拟上下文或另外的虚拟机在线(即，完成启动)时，它们还可以作为额外的时隙被添加到抽象层408。尽管图4示出了抽象层408具有与虚拟上下文406的数目相同的时隙410数目，但本领域技术人员将理解的是，虚拟时隙410的数目可能比虚拟上下文406的数目更多或更少，在该情形中，将存在过量的虚拟上下文或时隙。

[0063] 一旦被映射到时隙，虚拟上下文406或逻辑资源404可以被分派给租户314。通过检查被占用或分派的时隙410的状态，CSM系统302可以确定哪些逻辑资源或虚拟上下文是可用的以及有多少逻辑资源或虚拟上下文是可用的。例如，在图4中，如果时隙410_1和时隙410_3(以及引申的虚拟上下文410_1和虚拟上下文410_3)被分派给一些客户端设备314，则CSM系统302可以确定空闲时隙的数目(从而确定可用资源的数目)为N-2。

[0064] 图5是示出以服务VM池316、抽象层508以及客户端设备512_1、512_2、512_3(统称为“512”)为特征的另一示例系统500的框图。CSM系统(未示出)也可以涉及将逻辑资源504_1、504_2,…,504_6(统称为“504”)映射到抽象层508，并且随后将时隙510_1、510_2分派给请求设备512。服务VM池316可以是一个或多个服务VM 502_1、502_2,…,502_i(统称为“502”)的集合。VM 502可以在其上托管各种类型的逻辑资源504。客户端设备512可以通过CSM系统302请求访问逻辑资源504中的一个或多个。CSM系统302然后可以向每个请求客户端设备512分派空闲时隙。

[0065] VM 502可以托管一种或多种类型的逻辑资源504。例如，逻辑资源504_1、504_4、504_6可以是类型1，并且逻辑资源504_2、504_3、504_5可以是类型2。作为进一步说明，类型1逻辑资源可以是虚拟防火墙，并且类型2逻辑资源可以是VPN。如图5所示，虚拟机502_2可以仅托管一种类型的逻辑资源504_3，并且虚拟机502_1可以托管两种或更多种类型的逻辑资源504_1、504_2。每个VM 502还可以托管给定逻辑资源的多个实例。例如，VM 502_1可以运行逻辑资源1(504_1)的四个虚拟上下文和逻辑资源2(504_2)的三个虚拟上下文，而VM 502_2可以具有逻辑资源2(504_3)的三个虚拟上下文但不具有逻辑资源1的虚拟上下文。

[0066] 抽象层506_1、506_2,…,506_6(统称为“506”)可以以映射到VM 502中的虚拟上下文的虚拟时隙为特征。尽管抽象层506在图5中被描绘为VM 502的一部分，但抽象层506不一定必须驻留在任何VM中。抽象层506的软件实现方式和/或逻辑数据结构可以被存储于VM 502、CSM系统302、或任意其他计算设备中。每个VM 502对于其逻辑资源504可以具有其自己的一组时隙506。例如，VM 502_1可以在抽象层506_1中具有映射到逻辑资源1(504_1)的四个虚拟上下文的四个时隙以及在抽象层506_2中具有映射到逻辑资源2(504_2)的三个虚拟上下文的三个时隙。在另一示例中，VM 502_i可以在抽象层506_6中只具有映射到其唯一的逻辑资源504_6的一个时隙。

[0067] 可选地，CSM系统302可以聚合多个VM 502的虚拟时隙506并且将其安排到抽象层508的另一层中。抽象层508可以是以分层的方式排列的与抽象层506不同的层。替代地，抽象层508可以简单地为涉及抽象层506的信息的集合和/或重新排列。例如，对于逻辑资源1，抽象层506_1中的四个时隙、抽象层506_4中的两个时隙、以及抽象层506_6中的一个虚拟时隙可以在抽象层510_1中被重新排列并且被重新编号为时隙1-7。由此，CSM系统302可以利用一组虚拟时隙510_1来管理相同资源类型(即，逻辑资源1)的每个实例。类似地，逻辑资源2的虚拟

上下文(其遍及多个VM 502)可以被映射到时隙510_2的一个主要集合。

[0068] 在一个实施例中,CSM系统302可以针对不同的逻辑资源类型维护不同的抽象层(即,虚拟时隙的不同集合)。例如,CSM系统302可以将虚拟路由器的所有虚拟上下文映射到编号为0-1023的一组时隙,并且将虚拟防火墙的所有虚拟上下文映射到编号为0-511的另一组时隙,类似于图5中所示。在另一实施例中,CSM系统302可以具有一大组虚拟时隙,其将两种或更多种类型的逻辑资源进行组合。例如,CSM系统302可以将虚拟路由器或虚拟防火墙的每一实例映射到编号为0-1535的一组时隙。

[0069] 当租户设备512请求访问一个或多个逻辑资源时,CSM 302可以查看抽象层508的当前状态,并且确定所请求的资源类型的实例是否可用于分派。具体地,通过检查抽象层508中的给定时隙是否已经被占用(如图5中阴影部分所示),CSM 302可以确定该时隙是否可用于分派。例如,逻辑资源类型1的时隙1和2目前被分派给请求设备512_1,而时隙4和6分别被分派给请求设备512_2和请求设备512_3。同样,逻辑资源类型2的时隙1被分派给请求设备512_2,时隙3被分派给请求设备512_1,并且时隙5和6被分派给请求设备512_3。

[0070] 图6示出了可用资源的数目的期望范围的示例。为了实现资源请求与VM池316中资源可用性之间的最优性能和最小的等待时间,PM 308对于抽象层508中可用时隙的期望数目可以具有预定的值S 602,其还可以与VM池316中可用资源或未使用资源的数目相对应。换言之,值S 602可以是CSM 302所估计的PM 308在抽象层508中尽力维护的空闲时隙的理想数目或目标数目。在VM池316中运行一些空闲的VM(从而有一些额外的逻辑资源)使得CSM系统302能够一经通知便向租户提供服务。同时,VM池316中具有太多未使用的VM可能会使成本较高并且造成浪费。

[0071] 因此,值S 602可以利用数学公式基于许多不同变量(包括客户端设备314的数目、预计的服务需求、未决的服务请求的数目、资源请求速率、日历时间(例如,一天中的某一时间、一周中的某一天、节日等)、VM池大小、VM池容量、VM配设时间(即,启动时间)、VM故障率等)来计算。值S 602可以在随着这些依赖性变量中的一些随时间改变而进行动态改变。例如,当来自客户端设备314的服务请求速率增加时,空闲时隙的期望数目S 602也可以增加以补偿增加的需求。在另一示例中,在停机期间,例如,在午夜时,值S 602可以被调整以减少空闲时隙的数目。当VM池316中可用资源的数目下降到值S 602以下时,CSM 302可以启动一个或多个额外的VM以满足目标资源数目。另一方面,当空闲资源的数目超过目标值S 602,则一些过量的资源可以被销毁。

[0072] 替代地,CSM 302对于可用逻辑资源的数目可以具有期望范围DS 606。换言之,CSM系统302或其PM子组件308将尝试将空闲时隙的数目保持在期望范围DS 606之内,并且当空闲时隙的数目处于范围DS 606的下限和上限之外时,可以对服务VM或逻辑资源的实例的数目进行相应调整。DS 606可以基于针对空闲时隙的期望数目的值S 602来确定。例如,DR 606可以被表示为$INT[f_1(S),f_2(S)]$,其中,INT([])表示包括上限和下限的区间,并且其中,$f_1(S)$和$f_2(S)$分别是表示下限和上限的S的函数。然而,本领域技术人员将理解,期望范围DR 606可以通过不同的公式来确定。

[0073] 在一些实现方式中,函数$f_1(S)$和$f_2(S)$也可以基于其他变量,例如,客户端设备314的数目、预计的服务需求、未决的服务请求的数目、资源请求速率、资源请求速率的一阶导数、资源请求速率的二阶导数、平均资源使用时间、预测的资源释放时间、日历时间、VM池大

小、VM池容量、VM配设时间、VM故障率等。

[0074] 作为示例，期望范围DR 606的下限和上限可以通过函数$f_1(S)$和$f_2(S)$来表示，以使得$f_1(S)=S-M_1$和$f_2(S)=S+M_2$，其中，M_1和M_2是表示下边限和上边限的非负整数。在该示例中，$S=6$，$M_1=2$，并且$M_2=1$(602)，这使得期望范围DR 606等于INT([4,7])。换言之，CSM 302将尝试使得空闲时隙的数目(以及可用资源的数目)保持在4和7之间，并且为了满足VM池大小要求而按需创建或销毁VM。

[0075] 图7A-图7D是示出了对VM池的示例调度功能操作的框图。抽象层700以一组虚拟时隙(统称为“702”)为特征，这组虚拟时隙可以被映射到由服务VM池316中的服务VM 502托管的逻辑资源。图中阴影部分示出被分派给客户端设备314的时隙。相反地，无阴影的时隙表示可以被分派给新的客户端的空闲时隙。标志704当被升起为704_1时，其可以表示空闲时隙的数目已经落入期望范围DR 606以外，并且需要通过创建另外的VM或销毁过量的VM来重新调整可用时隙的数目。升起或降落标志704例如可以通过将二进制标志比特在0(即，“降落”位置704_2)与1(即，“升起”位置704_1)之间进行切换来完成。在一个实施例中，可能存在多于一个的标志。例如，赤字标志可以被排他地用来通过信号通知空闲时隙的数目已经下降到DR 606以下，并且另一标志可以被排他地用来通过信号通知空闲时隙的数目已经超过期望范围DR。抽象层700与标志二者可以全部利用软件来实现或者被实现为硬件和软件二者的组合。

[0076] 抽象层700可以包含涉及对VM池316进行管理的其他信息。例如，每个时隙可以包含关于如下内容的信息：该时隙被映射到的VM的身份、所映射的虚拟上下文的身份、映射的时间、分派状态(例如，租户标识符、分派时间、调度释放时间等)、该时隙是否可以被多于一个的设备共享、预留队列等。对虚拟时隙向客户端314的调度和分派可以由SCH 304进行处理，而PM 308和VMM 310可以分别调整池大小和创建/销毁VM。

[0077] 在图7A中，抽象层700目前具有七个时隙702_1、702_2，…，702_7，每个时隙被映射到逻辑资源或逻辑资源的虚拟上下文。换言之，七个时隙702表示逻辑资源的七个独立的实例，这七个独立的实例转而可以是七个物理资源的逻辑等同。映射到时隙702的逻辑资源可以由一个服务虚拟机托管或者遍及VM池316中的多个服务虚拟机。然而，从SCH 304的视角，这些细节中的一些可以被隐藏。目前，七个虚拟时隙中的四个(即，702_1、702_2、702_4、702_7)被分派给一个或多个客户端设备314。因此，抽象层700当前具有三个空闲时隙702_3、702_5、702_6。在其周期性维护例程之一的过程中，PM 308可以发现空闲时隙的数目(即，$S_A=3$)已经下降到期望范围DR＝INT([4,7])606的下限以下。PM 308可以通过将标志704升起到其升起的位置704_1来警告CSM系统302的其他组件。升起的标志704_1可以指示请求速率正在上升。

[0078] 在图7B中，VMM 310可以检测到标志704已被设置到其升起的位置704_1，并且确定要么VM池316需要额外的VM要么现有VM需要运行逻辑资源的更多实例(即，虚拟上下文)。VMM 310例如通过启动一个或多个额外服务VM来继续实例化逻辑资源的多三个实例。尽管数目3在该示例中已被选定为要产生的额外资源的数目，以使得可用时隙的总数目符合可用时隙的期望数目值$S=6$(602)，但本领域技术人员将理解的是，只要可用时隙的结果数目会落入期望范围DR＝INT([4,7])之内，那么可以创建更多时隙或更少时隙。例如，VMM 310仅可以产生新资源的绝对最小数目，以使得空闲时隙的数目符合期望范围DR。在VMM 310完成其工作之后，标志704可以被设置到其降落的位置704_2以防止未来任何重复资源创建操

作。当新创建的资源在线并且可访问时，PM 308可以创建新的虚拟时隙702_8、702_9、702_{10}，并且将其映射到逻辑资源的三个新的可用实例。因此，现在可以将空闲时隙计数S_A从3调整到6。

[0079] 在图7C中，一些客户端设备314已经终止与CSM系统302的服务。因此，先前被分派到一个或多个客户端设备314的时隙702_1、702_7通过调度功能304而被释放，并且在未来分派中变得可用。可用时隙计数S_A因而还增加了2个，变为8个。PM 308在其路由维护会话之一的过程中，其可以检测到空闲时隙计数太高，这可能导致对VM池316中资源的低效利用以及浪费。PM 308可以升起标志704_1以警告VMM 310。

[0080] 在图7D中，VMM 310检测到标志704_1已被升起，并且继而关闭一些VM以减少空闲资源的数目。在该示例中，VMM 310终止映射到时隙702_9、702_{10}的逻辑资源或虚拟上下文。这两个时隙702_9、702_{10}也被从抽象层700移除，以使得它们不再能够被分派给客户端。CSM 302也可以将可用时隙计数减少2以使得$S_A=6$，并且将标志704设置到其降落的位置704_2。尽管在该示例中选定数目2使得所产生的空闲时隙计数能够等于空闲时隙的期望数目值(即，$S_A=S=6$)，但只要所产生的空闲时隙计数落入期望范围DR之内，那么任何数目的时隙可以被删除。一旦完成了所有的维护操作，则标志704可以被设置到其降落位置704_2以通过信号通知此时不需要做出任何进一步的时隙计数调整。

[0081] 本公开已经公开了一些基本的系统组件和概念，现在转向图8-图11中示出的一些示例性方法实施例。出于清楚的目的，针对图1中所示的示例系统100讨论了方法，该示例系统100被配置为实施这些方法。应当理解，本文所概述的步骤是出于对主题技术的某些实施例进行说明的目的来提供的，但是也可以使用其他组合，包括排除、添加、修改某些步骤的组合。

[0082] 图8示出了用于创建或实例化逻辑资源的示例方法。实际上，系统100可以将多个抽象层时隙中的每一个映射到逻辑资源的虚拟上下文，其中，每个虚拟上下文由来自虚拟机池中的相应的虚拟机进行托管(802)。多个抽象层时隙可以是基于软件的数据结构，其被存储于云服务管理系统或虚拟机中。在一个实施例中，抽象层时隙可以被映射到多于一种类型的逻辑资源的虚拟上下文。逻辑资源可以是诸如防火墙、路由器、虚拟专用网(VPN)、负载均衡器或WAN优化器之类的虚拟网络资源。虚拟机可以托管多于一个的逻辑资源以及资源的多于一个的实例或虚拟上下文。

[0083] 系统100然后可以从设备接收针对逻辑资源的请求(804)。请求设备可以是经由API进行请求的客户端设备或者租户。该请求可以规定诸如所需的资源类型、优先级、使用的持续时间、最低性能要求等项目。资源创建可以发生在其他逻辑资源“创建”触发事件发生时。系统100在多个抽象层时隙中标识可用抽象层时隙(806)。对可用抽象层时隙的标识可以通过调度功能来完成。一旦被分派给客户端设备，则抽象层时隙以及其相关联的逻辑资源对于其他客户端设备而言可变得不可用。因此，当系统100标识可用的抽象层时隙时，映射到该时隙的逻辑资源、逻辑资源的虚拟上下文、或托管该逻辑资源的服务VM也可以被标识。

[0084] 系统100保留可用的抽象层时隙，以使得逻辑资源的相应虚拟上下文可以被供应(808)。对可用抽象层时隙的保留可以意味着请求设备具有对该时隙和映射到该时隙的逻辑资源(或其虚拟上下文之一)的排他使用。换言之，该时隙对于其他设备而言不再可用于

访问。系统100然后将该可用抽象层时隙标记为不可用(810)。因此，系统100的空闲时隙计数减少1。将时隙标记为不可用可以有助于避免向多个请求设备分派任何特定的抽象层时隙。然而，在一些实施例中，一个抽象层时隙可以被分派给两个或更多个请求设备，并且相关联的逻辑资源可以在多个请求设备之间进行共享。

[0085] 系统100向设备分派可用的抽象层时隙(812)。作为分派的结果，该设备可以对映射到该抽象层时隙(其现在被标记为不可用)的逻辑资源进行排他访问。将时隙标记为不可用以及将时隙分派给设备的时序可以互换。换言之，可以在将时隙分派给请求设备之后将该时隙标记为不可用。可选地，系统100可以执行VM池维护(814)以使得VM池的大小保持在值的预期范围之内。

[0086] 图9示出了用于执行VM池维护的示例方法。VM池维护可以确保空闲时隙的数目S_A保持在期望范围DR的边界之内。可以在检测到触发事件(例如，对逻辑资源或服务VM进行创建、实例化、生成、移除、或删除)时进行VM池维护。替代地，触发也可以作为系统100内部的一些逻辑的结果而发生。也可以周期性地或者根据预定计划来执行VM池维护。VM池维护可以通过调度功能、池管理器、或云服务管理系统的VM管理器来执行。

[0087] 作为VM池维护例程的一部分，系统100可以标识可用时隙计数(902)。可用时隙计数通常与可用的或空闲的逻辑资源的数目相对应。系统100然后确定可用时隙数量是否处于期望范围之外。具体地，系统100可以确定可用时隙计数是否处于期望范围以下(904)。期望范围是系统100认为可接受的、理想的或最优的空闲时隙的数目的值的范围。该范围可以基于空闲时隙的期望数目来确定。如果空闲时隙计数确实低于期望范围，则系统100可以创建或配设至少一个虚拟机并且向虚拟机池添加该新的虚拟机(906)。可选地，赤字标志(例如，布尔值)可以被设置为“真(TRUE)”，这可以表明VM池中资源消耗的速率高于时隙返回的速率。换言之，升起的标识可以通过信号通知:VM池不足了。

[0088] 在一些实施例中，对服务VM的创建可以通过由外部实体或用户向系统100进行API调用来触发。在其他实施例中，虚拟机可以由于其他触发事件来准备。例如，系统100可以检测到季节性峰值时间要来临并且需要更多的虚拟机。新创建的虚拟机可以托管可被分派给客户端设备来使用的一个或多个实例或逻辑资源。一旦新的虚拟机(从而新的逻辑资源)被创建，则系统100能够通过将时隙计数增加该资源的新的实例数目来调整可用时隙计数(908)。在VM池维护期间，期望的VM池大小S或下限和上限函数f_1和f_2也可以基于上面提及的各种因素(包括预计服务需求、未决服务请求的数目、资源请求速率等)进行动态调整。

[0089] 系统100还可以确定可用时隙计数是否处于期望范围以上(910)。如果是，则系统100可以从虚拟机池中移除至少一个虚拟机(912)。因而，曾由所移除的虚拟机托管的任何逻辑资源或逻辑资源的实例也可以被删除。替代地，一个或多个虚拟上下文可以被去激活。系统然后可以通过从该计数中减去所移除的资源数目来调整可用时隙计数(914)。可选地，可以以递归的方式来配设或移除更多的VM，直到可用时隙计数处于期望范围之内。

[0090] 图10示出了用于创建逻辑资源的另一示例方法。系统100检测到逻辑资源“创建”触发事件(1002)。在一些实施例中，“创建”触发事件可以是来自请求逻辑资源的客户端设备的API调用。在其他实施例中，触发事件可以是对需求激增的预测。系统100然后可以确定可用时隙的数目是否小于阈值(1004)。该条件可以在创建过程中被早期评估，以使得系统100可以尽快开始准备任何所需的新的VM。阈值可以是系统100所估计的抽象层中空闲时隙

的最优数目。替代地，阈值可以是系统100所估计的空闲时隙的期望范围的下限。如果已经有足够的空闲时隙(因而有足够的资源)，则该过程可以跳过，进入到选择步骤1010。

[0091] 然而，如果空闲时隙的数目低于阈值，则系统100可选地将赤字标志的值设置为“真”(1006)。该标志可以是可具有两个状态(“真”和“假(FALSE)”)之一的布尔变量，其可以由二进制比特1和0来表示。系统100的组件(例如，VM管理器)可以检测到标志的“真”状态，并且创建可以托管另外的逻辑资源的新的VM。系统100还可以明确地请求创建新的VM(1008)。一旦新的VM被创建，则其可以加入服务VM池中的其他服务VM的行列。系统100可以从VM池中选择VM(1010)。这样的选择可以通过使用抽象层来完成，该抽象层在逻辑上将由VM托管的资源或VM本身映射到抽象层中的虚拟时隙。在这一情形中，该系统可以分派可用时隙和/或将时隙标记为已使用，以使得与该时隙相关联的资源不会被重复分派给其他设备(1012)。

[0092] 图11示出了用于删除逻辑资源和/或释放虚拟时隙的示例方法。系统100检测到逻辑资源“删除”触发事件(1102)。删除触发事件可以是API调用、周期性VM池维护、服务到期等。例如，租户设备可以明确地请求对正在使用的逻辑资源进行释放，或者租户与系统100之间对于资源的服务协定可能自然到期。系统100可以释放与要被删除的逻辑资源相对应的不可用或被占用的抽象层时隙(1104)。因此，新释放的时隙可以变得可用于重新分派。系统100可能必须强制使得资源与客户端断开连接。在替代实施例中，相应的VM可以被关闭并且时隙可相应地被移除。

[0093] 接下来，系统100可以执行清除操作(1106)。该步骤可以由调度功能(SCH)或池管理(PM)功能来执行。作为清除操作的一部分，任何旧的配置可以被清除，此时不可用的抽象层时隙可以被再次标记为可用。随后，可以相应调整可用时隙计数。可选地，系统100可以执行VM池维护(1208)。资源删除之后的VM池维护可以基本类似于图9中所示的过程。

[0094] 应当理解，上面所示的步骤仅仅是用于说明的示例，并且可以按需包括或排除某些步骤。而且，尽管示出了步骤的特定顺序，但该顺序仅仅是说明性的，并且在不背离本文的实施例的范围的情况下，可以使用对步骤的任何适当的排列。

[0095] 因此，本文所描述的技术提供了改善用户体验、简化使用云服务的应用服务设计、并且更加可预测地建立虚拟资源实例化时间。

[0096] 尽管已经示出并描述了提供作为服务VM来提供的对云资源的加速实例化的说明性实施例，但应当理解，可以在本文的实施例的范围和精神之内做出各种其他适应和修改。例如，本文已经针对云网络示出和描述了实施例。然而，这些实施例在其广义上是不被限制的，事实上，这些实施例可以用于其他类型的共享网络。而且，尽管一些实施例针对虚拟网络资源进行描述和示出，但诸如服务设备、计算/处理设备、存储设备之类的其他类型的资源也可以作为逻辑资源进行托管。

[0097] 前面的描述针对具体实施例。然而，显而易见的是，可以对所描述的实施例做出其他变更和修改，而实现其一些优势或全部优势。例如，应明确认为，本文所描述的组件和/或元件可以被实现为存储于有形(非暂态)计算机可读介质(例如，盘/CD/RAM/EEPROM/等)上的软件，其具有在计算机、硬件、固件或其组合上运行的程序指令。相应地，该说明书仅以示例的方式进行，并且不以其他方式限制本文实施例的范围。因此，所附权利要求的目的在于覆盖落入本文实施例的真实精神和范围之内的所有这样的变更和修改。

[0098] 应当理解，在所公开的处理中的步骤的任何特定顺序或架构是示例性方法的说明。基于设计偏好，应当理解，这些处理中的步骤的特定顺序或架构可以被重新排列，或者仅执行所示出的步骤的一部分。一些步骤可以被同时执行。例如，在某些情景中，多任务和并行处理可能是有益的。而且，在上面所描述的实施例中对各种系统组件的分离不应当被理解为所有实施例中需要这样的分离，而是应当理解，所描述的程序组件和系统一般可以被一起集成在单个软件产品中或被封装在多个软件产品中。

[0099] 先前的描述被提供来使得本领域技术人员能够实施本文所描述的各个方面。对这些方面的各种修改对于本领域技术人员而言将是显而易见的，并且本文定义的一般原理可以适用于其他方面。因此，权利要求不意为限制本文所示的方面，而是根据与描述的权利要求一致的全部范围，其中，除非另有说明，否则以单数形式对元件的引用不意为“一个且只有一个”，而是“一个或多个”。

[0100] 诸如“方面”之类的短语不暗示该方面对于主题技术而言是必需的，或者该方面适应于主题技术的所有配置。关于某一方面的公开可以适用于所有配置、或者一个或多个配置。诸如方面的短语可以指代一个或多个方面，反之亦然。诸如“配置”之类的短语不暗示这样的配置对于主题技术而言是必需的，或者这样的配置适应于主题技术的所有配置。关于某一配置的公开可以适用于所有配置、或者一个或多个配置。诸如配置之类的短语可以指代一个或多个配置，反之亦然。

[0101] “示例性”一词在本文中用来指“作为示例或说明”。本文作为“示例性”进行描述的任何方面或设计不一定被理解为相对于其他方面或设计而言是优选的或有利的。

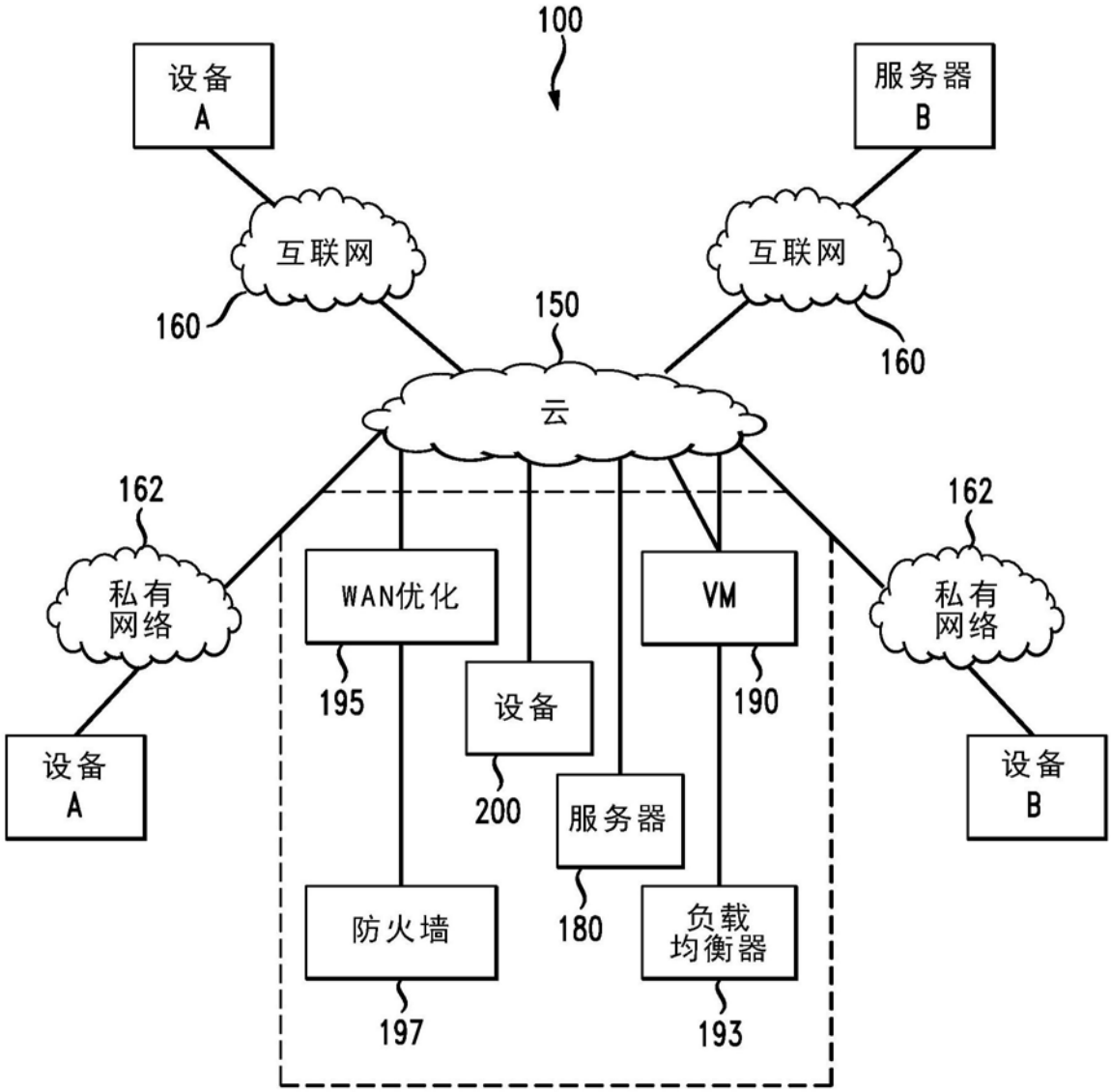

100
设备
A
服务器
B
互联网
160
互联网
160
150
云
162
162
私有
网络
私有
网络
WAN优化
VM
195
设备
190
设备
A
200
服务器
设备
B
180
防火墙
负载
均衡器
197
193

图1

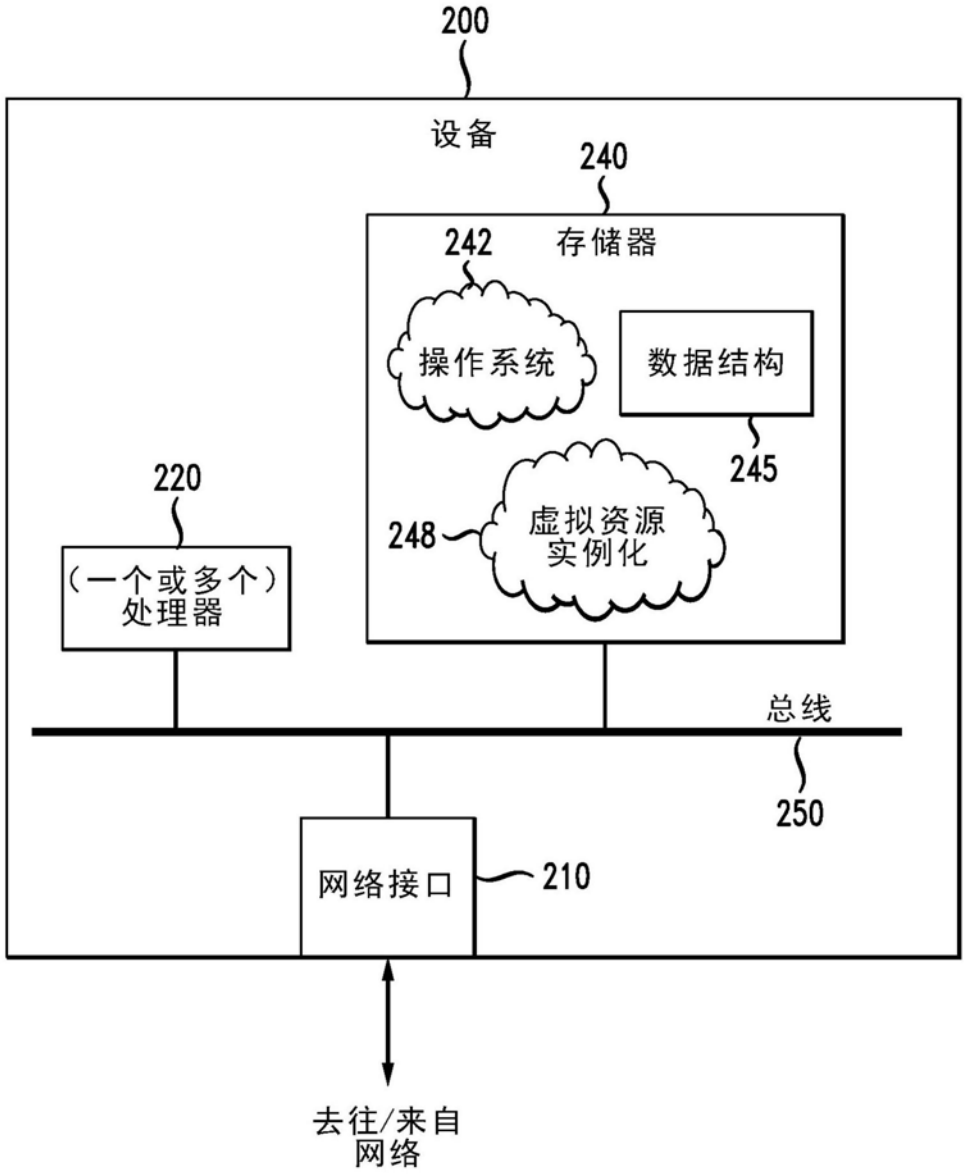

200
设备
240
存储器
242
操作系统
数据结构
245
220
(一个或多个)
处理器
248
虚拟资源
实例化
总线
250
网络接口
210
去往/来自
网络

图2

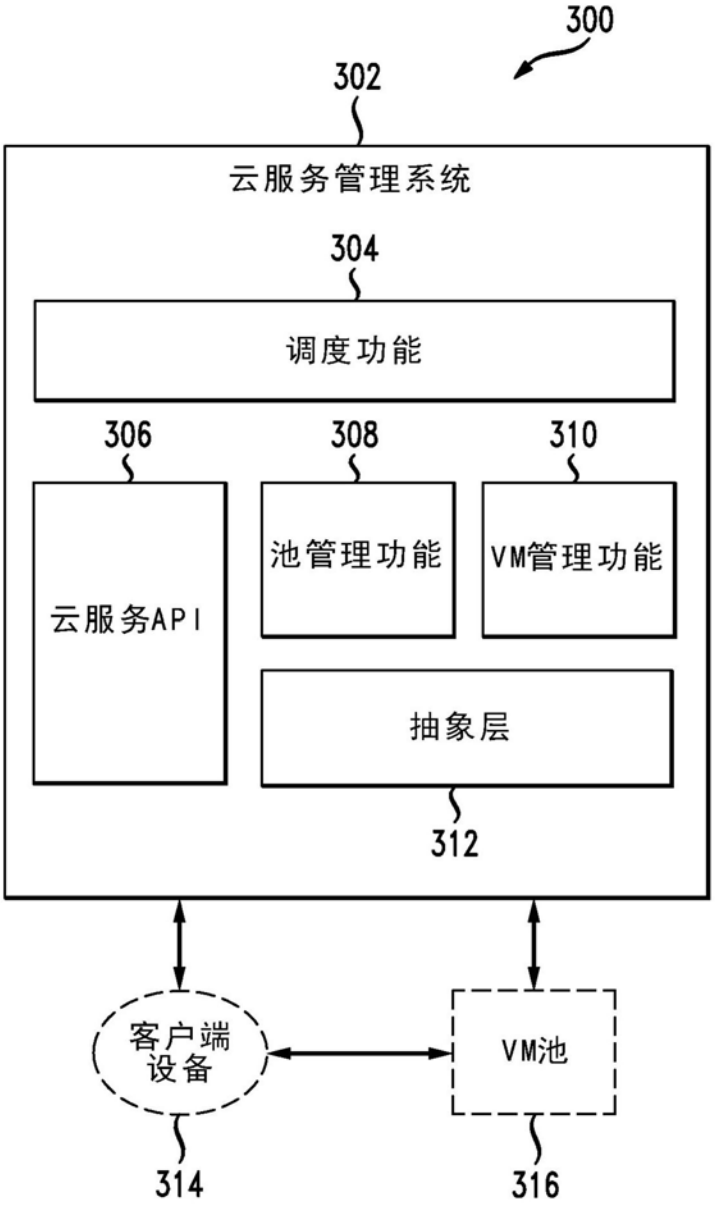

300
302
云服务管理系统
304
调度功能
306
308
310
云服务API
池管理功能
VM管理功能
抽象层
312
客户端设备
VM池
314
316

图3

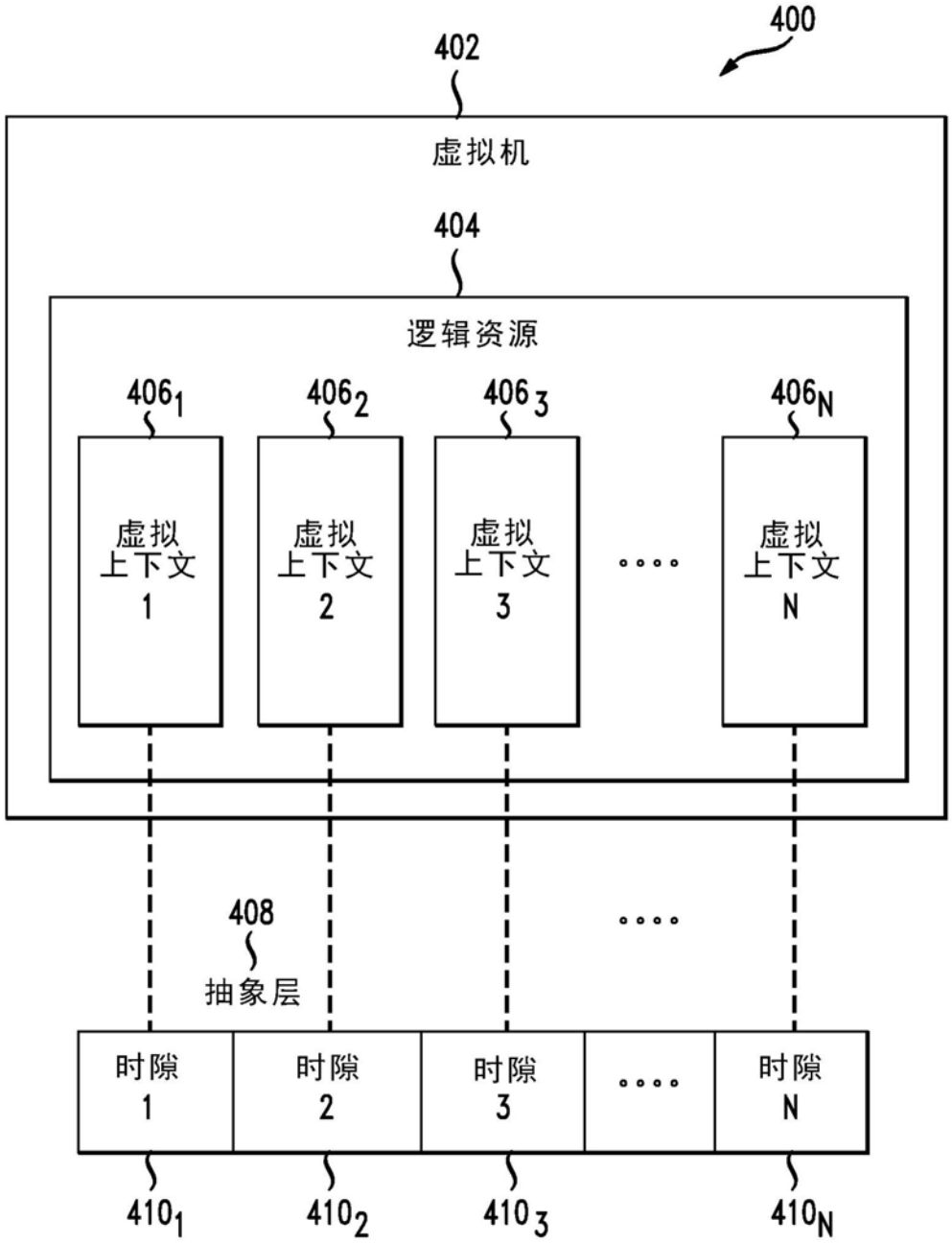

400
402
虚拟机
404
逻辑资源
406 1
406 2
406 3
406 N
虚拟 上下文 1
虚拟 上下文 2
虚拟 上下文 3
虚拟 上下文 N
408
抽象层
时隙 1
时隙 2
时隙 3
时隙 N
410 1
410 2
410 3
410 N

图4

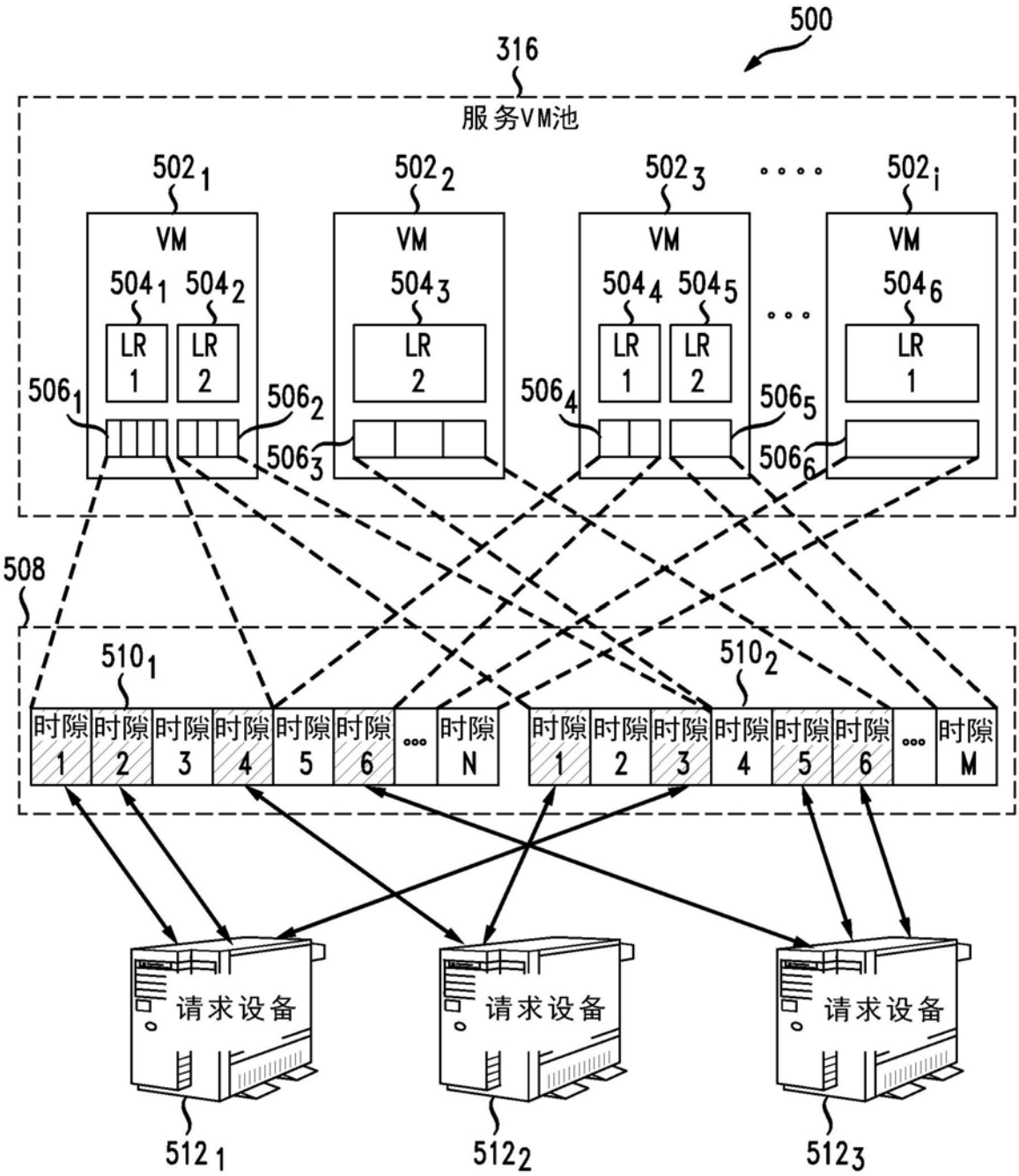

500
316
服务VM池
502 1
502 2
502 3
502 i
VM
504 1
504 2
504 3
504 4
504 5
504 6
LR 1
LR 2
506 1
506 2
506 3
506 4
506 5
506 6
508
510 1
510 2
时隙 1
时隙 2
时隙 3
时隙 4
时隙 5
时隙 6
时隙 N
时隙 M
请求设备
512 1
512 2
512 3

图5

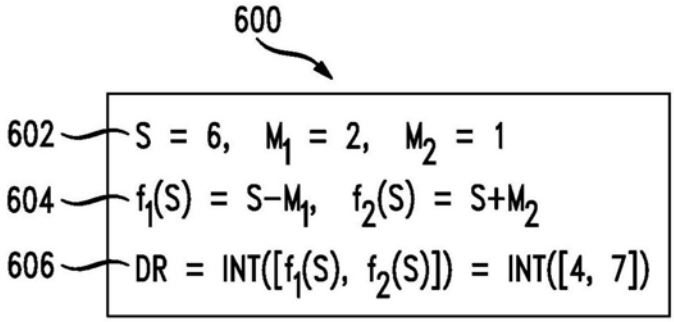

600
602
S = 6, M1 = 2, M2 = 1
604
f1(S) = S−M1, f2(S) = S+M2
606
DR = INT([f1(S), f2(S)]) = INT([4, 7])

图6

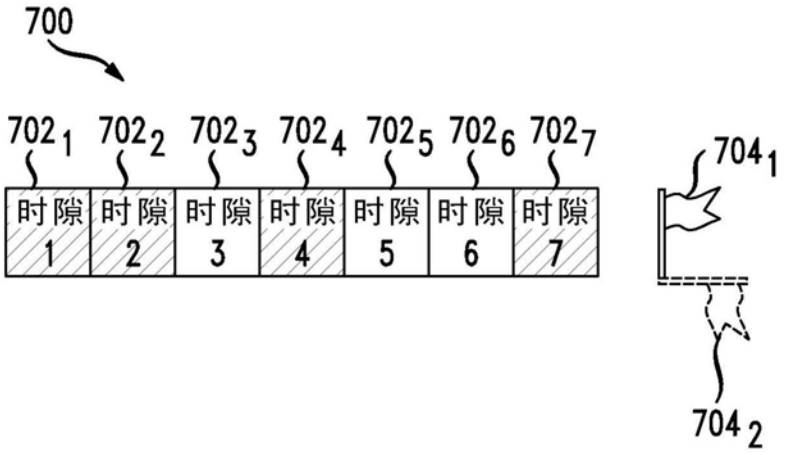

700
702 1
702 2
702 3
702 4
702 5
702 6
702 7
时隙 1
时隙 2
时隙 3
时隙 4
时隙 5
时隙 6
时隙 7
704 1
704 2

图7A

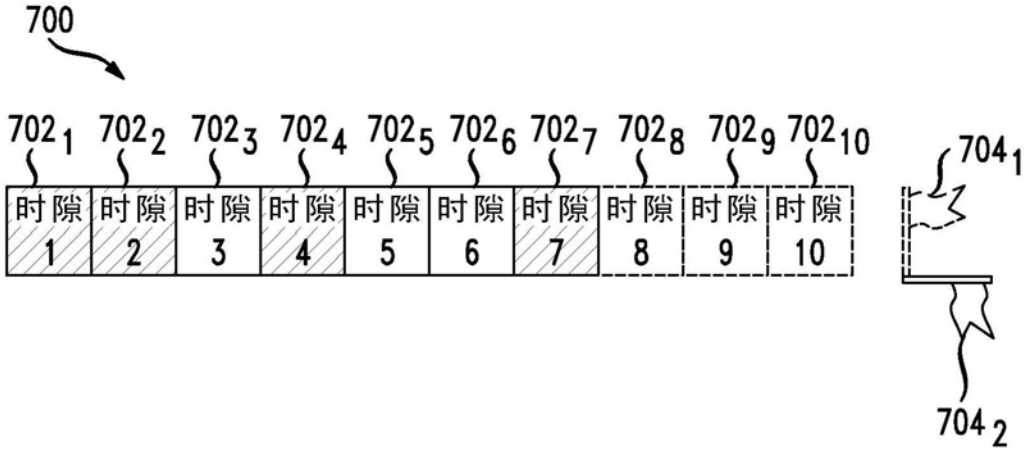

700
702 1
702 2
702 3
702 4
702 5
702 6
702 7
702 8
702 9
702 10
时隙 1
时隙 2
时隙 3
时隙 4
时隙 5
时隙 6
时隙 7
时隙 8
时隙 9
时隙 10
704 1
704 2

图7B

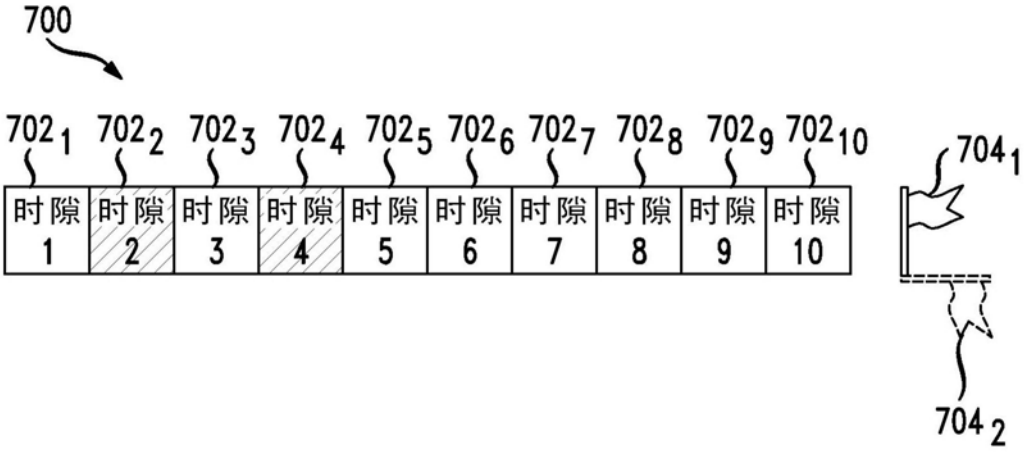

700
702_1 702_2 702_3 702_4 702_5 702_6 702_7 702_8 702_9 702_{10}
时隙 1
时隙 2
时隙 3
时隙 4
时隙 5
时隙 6
时隙 7
时隙 8
时隙 9
时隙 10
704_1
704_2

图7C

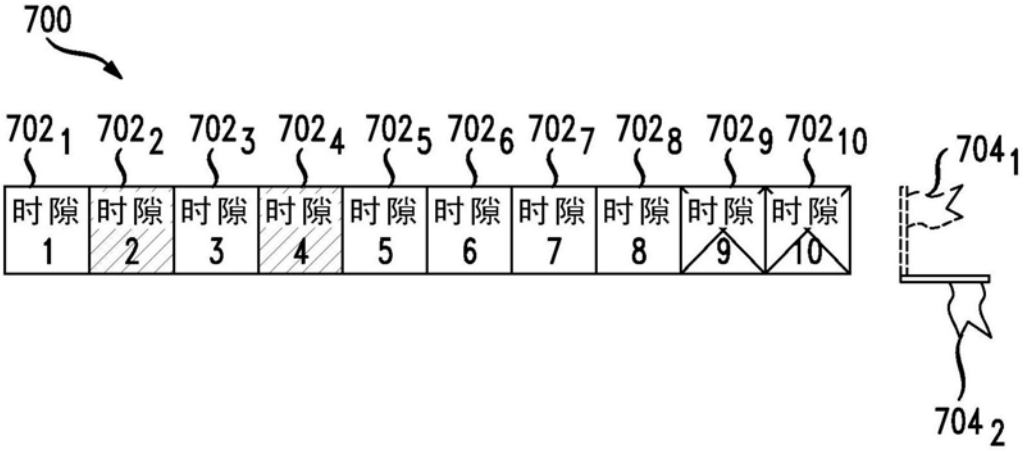

700
702_1 702_2 702_3 702_4 702_5 702_6 702_7 702_8 702_9 702_{10}
时隙 1
时隙 2
时隙 3
时隙 4
时隙 5
时隙 6
时隙 7
时隙 8
时隙 9
时隙 10
704_1
704_2

图7D

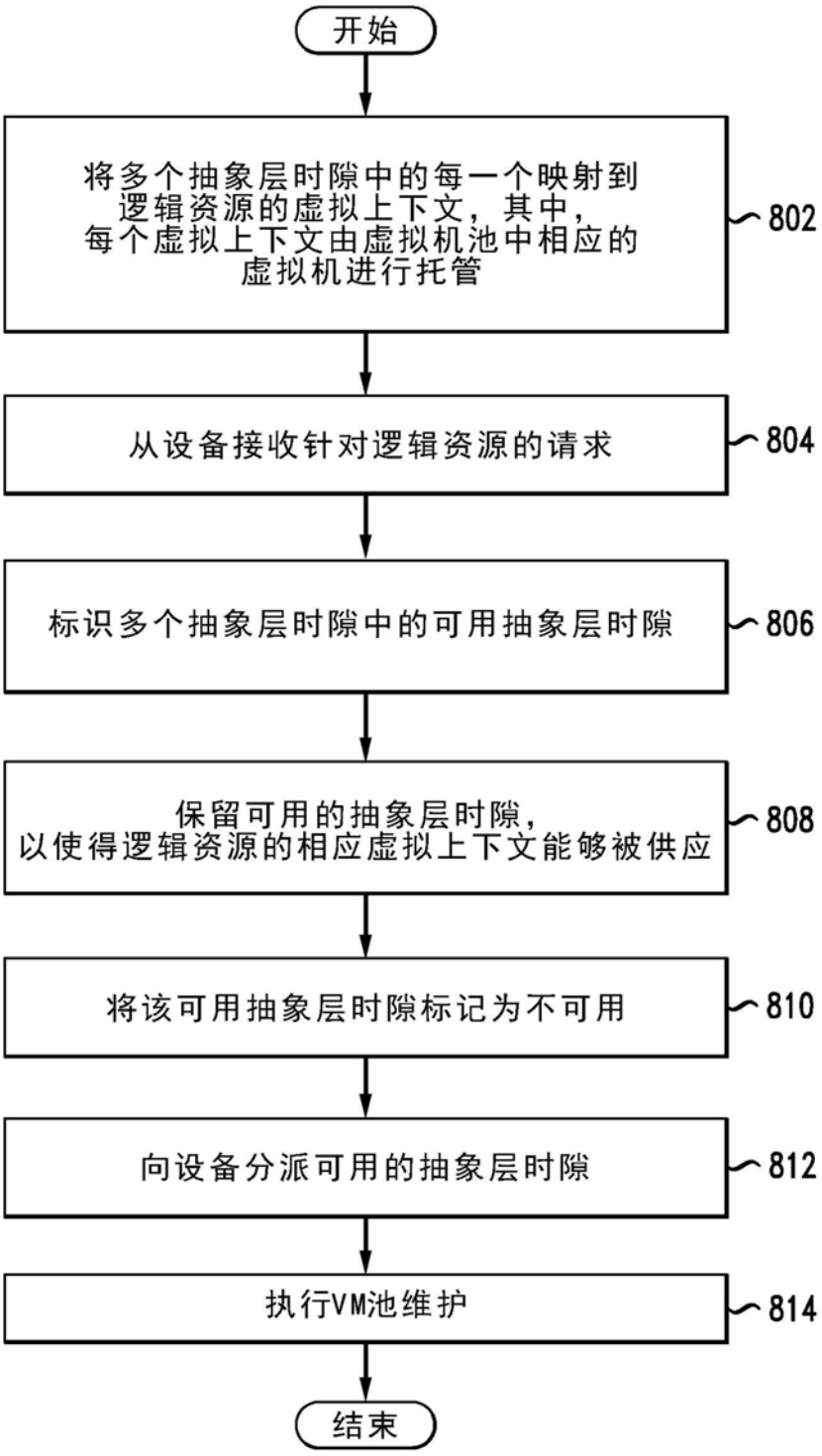

开始
将多个抽象层时隙中的每一个映射到逻辑资源的虚拟上下文，其中，每个虚拟上下文由虚拟机池中相应的虚拟机进行托管
802
从设备接收针对逻辑资源的请求
804
标识多个抽象层时隙中的可用抽象层时隙
806
保留可用的抽象层时隙，以使得逻辑资源的相应虚拟上下文能够被供应
808
将该可用抽象层时隙标记为不可用
810
向设备分派可用的抽象层时隙
812
执行VM池维护
814
结束

图8

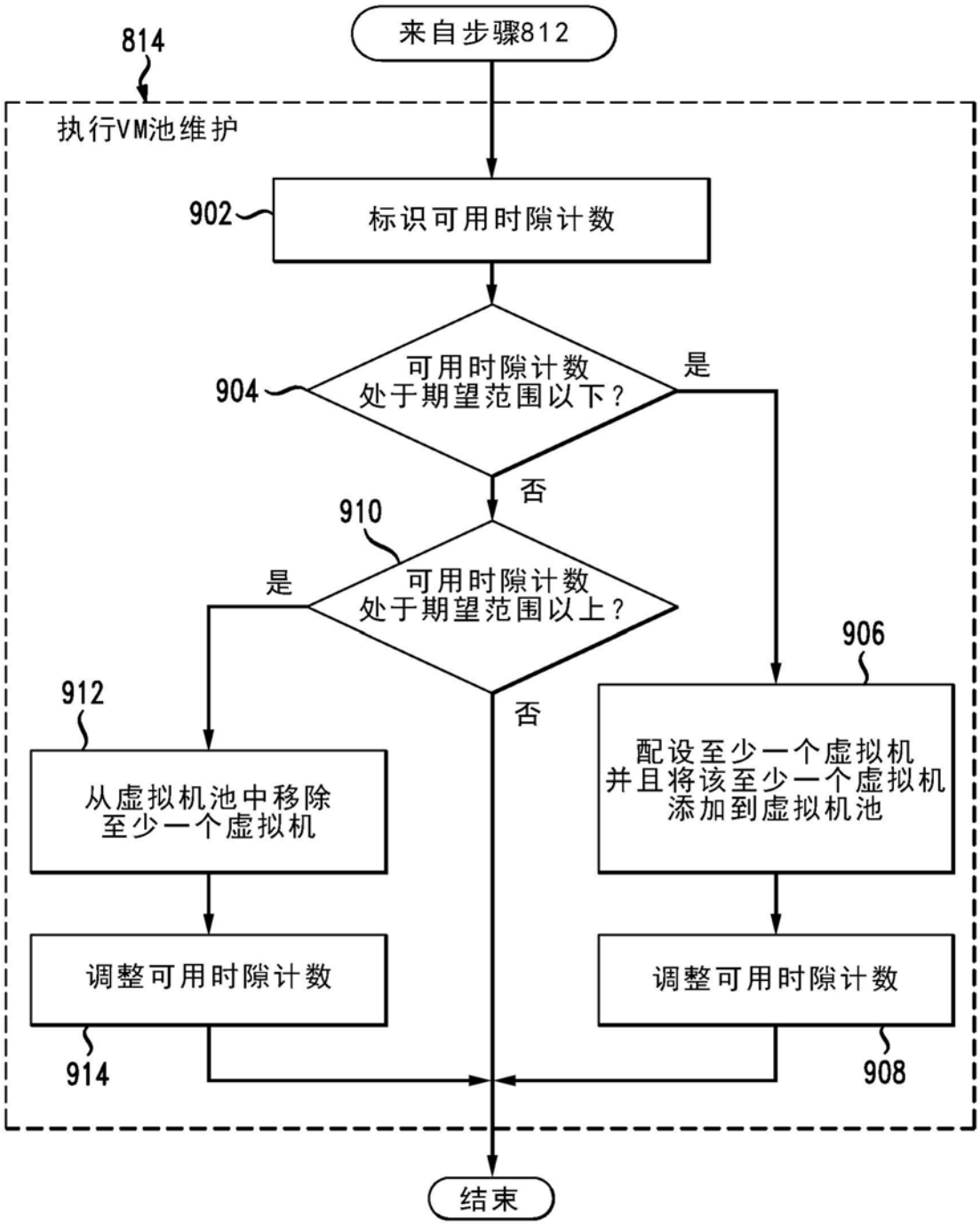

814
来自步骤812
执行VM池维护
902
标识可用时隙计数
904
可用时隙计数
处于期望范围以下？
是
否
910
可用时隙计数
处于期望范围以上？
是
否
906
配设至少一个虚拟机
并且将该至少一个虚拟机
添加到虚拟机池
912
从虚拟机池中移除
至少一个虚拟机
调整可用时隙计数
调整可用时隙计数
914
908
结束

图9

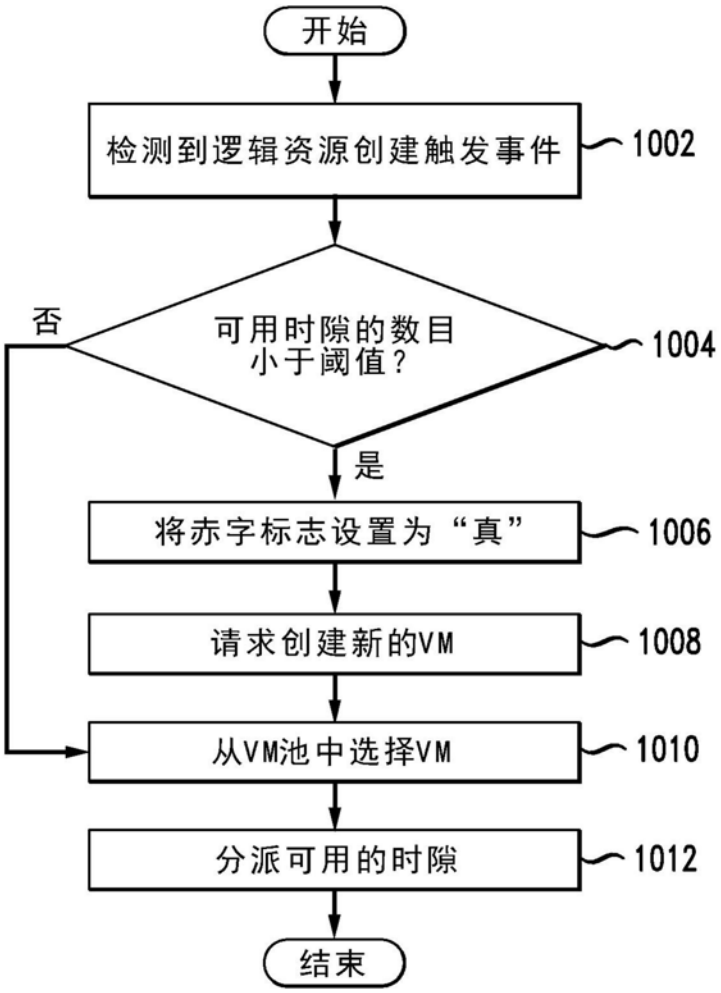

开始
检测到逻辑资源创建触发事件
1002
可用时隙的数目
小于阈值？
1004
否
是
将赤字标志设置为“真”
1006
请求创建新的VM
1008
从VM池中选择VM
1010
分派可用的时隙
1012
结束

图10

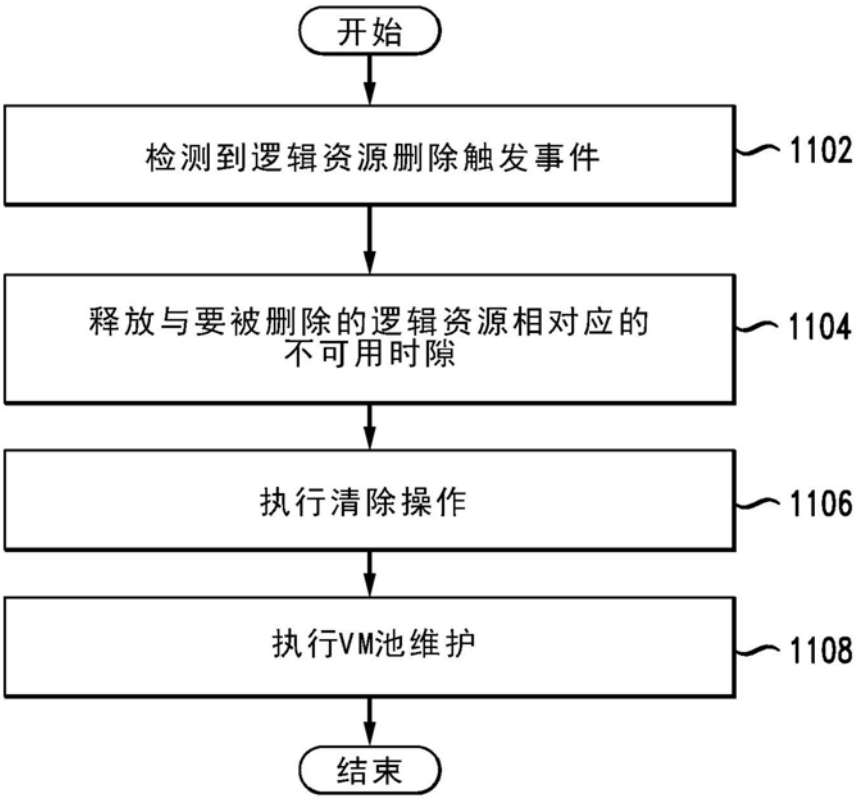

开始
检测到逻辑资源删除触发事件
1102
释放与要被删除的逻辑资源相对应的不可用时隙
1104
执行清除操作
1106
执行VM池维护
1108
结束

图11