(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 107231815 A
(43)申请公布日 2017.10.03

(21)申请号 201480068652.7

(22)申请日 2014.11.11

(30)优先权数据
14/077,146 2013.11.11 US

(85)PCT国际申请进入国家阶段日
2016.06.16

(86)PCT国际申请的申请数据
PCT/US2014/065068 2014.11.11

(87)PCT国际申请的公布数据
WO2015/070241 EN 2015.05.14

(71)申请人 亚马逊技术有限公司
地址 美国华盛顿

(72)发明人 奎斯·塔拉基 麦特·瓦尔辛
维诺德·穆里·马塔尼
詹姆斯·乔纳森·莫里斯
杰拉德·约瑟夫·海因茨II

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 闫晔

(51)Int.Cl.
G06F 15/16(2006.01)

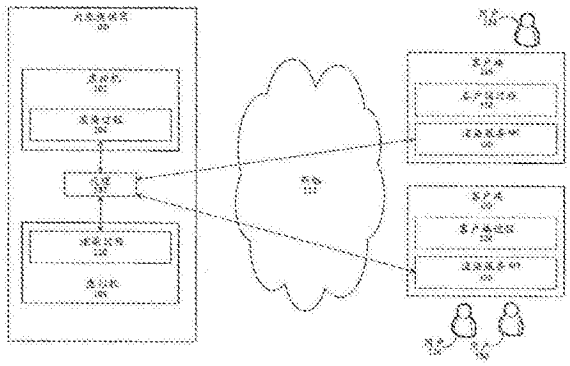
权利要求书2页 说明书18页 附图8页

(54)发明名称

用于流传输服务器的会话空闲优化

(57)摘要

可以通过远程计算设施向客户端设备提供图形渲染服务。可以在于主机计算设备上进行操作的虚拟机上执行一个或多个渲染过程。可以监测客户端状态信息以便检测不活动的时期。可以通过使渲染过程在其上执行的虚拟机暂停来将渲染过程去激活。当活动恢复时，可以通过恢复虚拟机的执行来重新激活渲染过程。



CN 107231815 A

1.一种系统,其包括:

一个或多个计算节点,所述一个或多个计算节点被配置成来操作用于代表一个或多个客户端渲染图形的服务,所述服务包括多个虚拟机;

一个或多个计算节点被配置成至少:

接收指示代表所述一个或多个客户端渲染图形的请求,所述请求包括指示与在所述一个或多个客户端上运行的过程相关联的图形资源集的信息;

确定激活所述多个虚拟机中的虚拟机,所述确定至少部分地基于所述虚拟机被配置成执行对应于所述图形资源集的渲染过程;

响应于尚未接收到将所述渲染过程保持在活动状态的请求的第一确定和自从接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息以来的时间量已超过第一阈值的第二确定,暂停所述虚拟机的操作,其中所述虚拟机在暂停时的第一状态包括用于所述渲染过程的第二状态;以及

响应于接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息,恢复所述虚拟机的操作。

2.如权利要求1所述的系统,所述一个或多个计算节点还被配置成至少:

将所述虚拟机的所述第一状态存储在低延迟高速缓存中;以及

响应于自从接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息以来的所述时间量已超过第二阈值,将所述虚拟机的所述第一状态存储在存储设备上。

3.如权利要求1所述的系统,所述一个或多个计算节点还被配置成至少:

至少部分地基于使所述虚拟机暂停来激活另外的虚拟机。

4.如权利要求1所述的系统,其中对将所述渲染过程保持在活动状态的所述请求是由在所述一个或多个客户端中的客户端上进行操作的过程发送的。

5.一种系统,其包括:

一个或多个处理器;以及

一个或多个计算机可读存储介质,所述一个或多个计算机可读存储介质具有存储在其上的指令,所述指令在由所述一个或多个处理器执行时致使所述一个或多个计算设备至少:

接收指示为一个或多个客户端执行图形渲染服务的请求,所述一个或多个客户端执行与图形资源集相关联的过程;

至少部分地基于选择用于激活的虚拟机被配置成执行能够访问所述图形资源集的渲染过程来激活所述虚拟机;

由所述渲染过程来为所述一个或多个客户端执行所述图形渲染服务;

至少部分地基于确定尚未接收到对保持所述渲染过程为活动的请求并且至少部分地基于确定自从接收到指示由所述一个或多个客户端中至少一个进行的活动的信息以来的时间量已超过第一阈值时间量,暂停所述虚拟机的操作;以及

响应于接收到指示由所述一个或多个客户端中至少一个进行的活动的信息,恢复所述虚拟机的操作。

6.如权利要求5所述的系统,其还包括指令,所述指令在由所述计算设备执行时致使所述计算设备至少:

响应于自从接收到指示由所述一个或多个客户端中至少一个进行的活动的信息以来的所述时间量已超过第二阈值时间量，将所述虚拟机的状态存储在存储设备上。

7.如权利要求5所述的系统，其中暂停所述虚拟机的所述操作包括将所述虚拟机的状态保持在存储器中。

8.如权利要求5所述的系统，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

响应于确定不运行渲染过程的活动虚拟机的数量已低于阈值，激活另外的虚拟机。

9.如权利要求5所述的系统，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

向所述一个或多个客户端中的客户端发送指示所述渲染过程的恢复操作的时间的信息。

10.一种方法，其包括：

接收指示为一个或多个客户端执行图形渲染服务的请求，所述一个或多个客户端执行与图形资源集相关联的过程；

至少部分地基于所选择用于执行渲染过程的虚拟机能够访问所述图形资源集，在所述虚拟机上执行所述渲染过程；

由在所述虚拟机上进行操作的所述渲染过程来为所述一个或多个客户端执行所述图形渲染服务；

至少部分地基于尚未接收到对将所述渲染过程保持在活动状态的请求的第一确定和自从接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息以来的时间量已超过第一阈值的第二确定，暂停所述虚拟机的操作；以及

响应于接收到指示接收对代表所述一个或多个客户端执行图形渲染的请求的信息，恢复所述虚拟机的操作。

11.如权利要求10所述的方法，其中指示接收对代表所述一个或多个客户端执行图形渲染的请求的所述信息对应于由所述一个或多个客户端中至少一个的用户提供的输入。

12.如权利要求10所述的方法，其中对将所述渲染过程保持在活动状态的所述请求至少部分地基于进入没有用户输入预期达一定时间段的状态。

13.如权利要求10所述的方法，其还包括：

将所述虚拟机的状态存储在低延迟高速缓存中达至少等于第二阈值的时间段。

14.如权利要求10所述的方法，其还包括：

将所述虚拟机重新设定到初始状态，所述初始状态对应于在执行所述渲染过程之前的所述虚拟机的状态。

15.如权利要求10所述的方法，其还包括：

向所述一个或多个客户端中的客户端发送指示恢复所述渲染过程的状态的信息。

用于流传输服务器的会话空闲优化

[0001] 相关申请的交叉引用

[0002] 本申请要求2013年11月11日提交的美国专利申请号14/077,146的权益，所述申请的公开内容以引用方式整体并入本文。

[0003] 本申请涉及以下申请：2013年11月11日提交的标题为“VIDEO ENCODING BASED ON AREAS OF INTEREST(基于感兴趣区域的视频编码)”(代理人案号：101058.000083)的美国专利申请号14/076,718；2013年11月11日提交的标题为“ADAPTIVE SCENE COMPLEXITY BASED ON SERVICE QUALITY(基于服务质量的自适应场景复杂性)”(代理人案号：101058.000084)的美国专利申请号14/076,821；2013年11月11日提交的标题为“SERVICE FOR GENERATING GRAPHICS OBJECT DATA(用于生成图形对象数据的服务)”(代理人案号：101058.000086)的美国专利申请号14/077,127；2013年11月11日提交的标题为“IMAGE COMPOSITION BASED ON REMOTE OBJECT DATA(基于远程对象数据的图像合成)”(代理人案号：101058.000087)的美国专利申请号14/077,136；2013年11月11日提交的标题为“MULTIPLE PARALLEL GRAPHICS PROCESSING UNITS(多个并行图像处理单元)”(代理人案号：101058.000110)的美国专利申请号14/077,165；2013年11月11日提交的标题为“ADAPTIVE CONTENT TRANSMISSION(自适应内容传输)”(代理人案号：101058.000114)的美国专利申请号14/077,084；2013年11月11日提交的标题为“VIEW GENERATION BASED ON SHARED STATE(基于共享状态的视图生成)”(代理人案号：101058.000115)的美国专利申请号14/077,180；2013年11月11日提交的标题为“MULTIPLE STREAM CONTENT PRESENTATION(多流内容呈现)”(代理人案号：101058.000116)的美国专利申请号14/077,186；2013年11月11日提交的标题为“DATA COLLECTION FOR MULTIPLE VIEW GENERATION(用于多视图生成的数据收集)”(代理人案号：101058.000124)的美国专利申请号14/077,149；2013年11月11日提交的标题为“STREAMING GAME SERVER VIDEO RECORDER(流式游戏服务器录像机)”(代理人案号：101058.000125)的美国专利申请号14/077,142；2013年11月11日提交的标题为“LOCATION OF ACTOR RESOURCES(演员资源的定位)”(代理人案号：101058.000128)的美国专利申请号14/076,815；2013年11月11日提交的标题为“APPLICATION STREAMING SERVICE(应用流传输服务)”(代理人案号：101058.000139)的美国专利申请号14/077,023；2013年11月11日提交的标题为“EFFICIENT BANDWIDTH ESTIMATION(有效带宽估计)”(代理人案号：101058.000141)的美国专利申请号61/902,740，所述申请中的每一个据此以引用方式整体并入本文。

背景技术

[0004] 计算设备诸如移动电话、平板计算机、游戏控制台等可能未被配备成以对某些应用来说足够的速度和细节来渲染图形。渲染图形，也可以被描述为用于生成在游戏和其他计算机应用中使用的图像的过程，可以利用专用的计算资源，诸如可能在计算设备上不可用的图形处理单元。在一些情况下，资源是可用的但是将消耗过度的功率，或者将以不足的速度运行或提供水平不足的图形质量。

[0005] 可以通过位于远程设施处的计算资源向客户端设备提供图形渲染能力。所述设施例如可以配备有多组图形处理单元(“GPU”)或专用于提供渲染服务的其他硬件。然而,即使是在使用专用硬件的情况下,提供图形渲染服务也可能消耗大量的计算资源。例如,图形渲染可涉及将各种模型、纹理、位图等加载到存储器中。当相关的过程在客户端设备上运行时,这些资源可以保持在存储器中。资源利用率的管理可以改善渲染服务的性能和效率。

[0006] 附图简述

[0007] 结合附图阅读时,可以更好地理解以下详细描述。出于说明的目的,附图中示出本公开的各个方面的示例性实施方案;然而,实施方案并不限于所公开的具体方法和手段。

[0008] 图1是描绘用于向客户端过程提供远程渲染服务的系统的实例的框图。

[0009] 图2是描绘配置成利用远程渲染服务的客户端的框图。

[0010] 图3是描绘在虚拟机实例上执行渲染过程的示例性内容提供商系统的框图。

[0011] 图4是描绘用于激活和去激活在虚拟机实例上执行的渲染过程的示例性过程的流程图。

[0012] 图5是描绘用于将客户端与渲染过程和渲染过程可在其上执行的虚拟机实例相关联的示例性过程的流程图。

[0013] 图6是描绘用于维持被配置成执行渲染过程的虚拟机实例池的示例性过程的流程图。

[0014] 图7是描绘可以在一些实施方案中使用的示例性计算系统的框图。

[0015] 图8是描绘可以在一些实施方案中使用的示例性计算系统的框图。

[0016] 详述

[0017] 根据所公开技术的一些示例性特征,特定内容项(诸如视频游戏)的场景的一个或多个渲染视图可以由内容提供商生成并且从提供商传输到多个不同的客户端,在一些情况下,内容提供商可以生成特定内容项的场景或虚拟环境的多个视图。所述多个视图中的每一个例如可以与一个或多个相应客户端相关联,并且可以从内容提供商传输到相应客户端。例如,每个视图可以呈现从所述视图传输到的相应客户端所控制的特定角色或其他实体的角度来看的场景。视图可基于由内容提供商维持的游戏或其他视频内容的共享状态。在一些情况下,内容提供商可以将特定内容项的场景的相同视图传输到多个客户端。相同视图例如可以传输到从特定角色的角度观看游戏的客户端。

[0018] 内容提供商可以通过渲染过程提供渲染服务。在一些情况下并且在一些实施方案中,渲染过程可以对应于可执行应用的实例。然而,渲染过程可替代地包括组件启用、程序库调用、对象、可执行应用的多个实例等。

[0019] 渲染过程可以与一组一个或多个客户端相关联。每个客户端继而可以与一个或多个用户相关联。渲染过程可以与特定内容项(诸如游戏、动画、电视节目等)相关联。渲染过程还可以与内容项的实例相关联。例如,多玩家游戏可以提供一组客户端之间的交互。在此类情况下,实施方案可以将渲染过程与所述一组客户端相关联。在一些情况下并且在一些实施方案中,多个渲染过程可以与一个游戏实例相关联,并且每个渲染过程可以与一个或多个客户端相关联。例如,在大规模的多玩家游戏中或群组观看电影或电视节目时可以采用这种方法。

[0020] 为了使得能够生成显示内容的单个共享状态并且从那个状态内选择场景的一个

或多个视图，不同参与客户端中的每一个可以收集相应客户端状态信息。客户端状态信息可包括例如关于在相应客户端处执行的操作(诸如由相应客户端控制的相应角色或其他实体执行的移动或其他动作)的信息。客户端可以周期性地将其相应客户端状态信息的更新传输到内容提供商。内容提供商随后可以使用从每个客户端接收到的客户端状态信息更新来更新由内容提供商维持的共享内容项状态信息。内容提供商随后可以使用所述共享内容项状态信息来生成传输到不同参与客户端的一个或多个视图。

[0021] 客户端状态信息还可包括与客户端的用户提供的输入相关的或反映所述输入的信息。例如，客户端的用户可以按压按钮、移动操纵杆、对麦克风说话等。来自客户端(诸如所呈现的实例)的输入可导致由相应客户端控制的相应角色或其他实体执行的移动或其他动作。在一些情况下并且在一些实施方案中，输入可以不与游戏角色的控制相关。输入还可以对应于用户的存在。例如，运动敏感的摄像机可以指示客户端的用户的存在。

[0022] 在一些实施方案中，客户端状态信息可包括关于旁观者状态、记录状态等的信息。例如，用户可以是多玩家游戏中的旁观者，在这种情况下客户端状态信息可包含指示旁观者的存在的信息，即使旁观者可能不向客户端提供输入，或间歇地提供输入。在另一个实施方案中，客户端设备可能正在记录信息，在这种情况下客户端状态信息可以包括正在记录所传输内容的指示。一些实施方案可以利用渲染服务以使得能够观看电影、电视、体育事件等。在一个这种实施方案中，客户端状态信息可涉及从对所显示内容做出评论的用户接收文本输入、音频输入或其他输入。

[0023] 接收客户端状态信息可以指示启动、恢复或继续与一个或多个客户端相关联的渲染过程的活动状态。内容提供商或渲染过程可以接收包含对应于或指示客户端状态信息的数据的各种传输或消息。指示渲染过程的活动状态的客户端状态信息包括但不限于由用户提供的输入、用户的存在、旁观者状态、记录状态等。

[0024] 渲染过程可以在虚拟机的实例上执行，或在除虚拟机实例之外的其他类型的计算节点上执行。虚拟机实例还可以被称为虚拟机。在一些实施方案中，可能存在渲染过程与计算节点的一对一的关联。在其他实施方案中，多个渲染过程可以在同一个计算节点上执行。实施方案可以将对应于同一个内容项的渲染过程与同一个计算节点相关联。这种方法可以使得渲染过程能够共享内容相关的图形资源。

[0025] 内容提供商可以通过使可用于活动的渲染过程的计算资源最大化来提高其操作效率。在其上执行渲染过程的虚拟机可以消耗否则可用于执行其他渲染过程的其他虚拟机的资源。为了使可用于运行活动的渲染过程的虚拟机的计算资源最大化，可以使正在运行不活动的渲染过程的那些虚拟机暂停。不活动的渲染过程可以是其执行状态保持在暂停的虚拟机的状态中的执行过程。

[0026] 不活动的渲染过程可包括其客户端当前不需要渲染服务的渲染过程。在各种情况下并且在各种实施方案中，客户端状态信息可以由内容提供商接收。客户端状态信息可以指示当前对渲染服务的需要，其例如可对应于以下因素：诸如内容的元素的移动、由客户端的用户提供的输入、旁观者模式、记录模式等。在各种情况下并且在各种实施方案中，中止接收客户端状态信息可以指示当前没有对渲染服务的需要。这可以例如在客户端设备已关闭时发生。另一种可能性是客户端状态可能已改变以使得不再需要渲染服务。例如，在客户端上运行的游戏可能已经结束并且新游戏尚未开始。

[0027] 图1描绘用于管理虚拟机上的渲染过程的系统的与本发明实施方案的各方面一致的实例。用户114可以与客户端116进行交互。客户端可包括硬件设备，诸如移动电话、平板电脑、游戏控制台、个人计算机等。客户端还可包括呈各种组合的形式的应用程序118、操作系统或其他软件模块。客户端还可包括渲染服务应用编程接口(“API”)120，其可提供能够访问渲染过程(诸如渲染过程104)的客户端过程118。客户端过程还可以直接与渲染过程通信。例如，客户端过程124可以与渲染过程110通信，或者调用渲染服务API 126的功能以便与渲染过程110通信。

[0028] 对渲染过程104的访问可以是直接的或间接的。例如，直接访问可包括客户端116通过网络112与渲染过程104通信。间接访问可包括客户端122通过内容提供商100来访问渲染过程110。在一些实施方案中，代理132可以在内容提供商100内进行操作以便代理对渲染过程(诸如渲染过程110)的访问。代理还可以在虚拟机102或虚拟机108内进行操作。

[0029] 多个用户诸如用户128和用户130可以与客户端诸如客户端122进行交互。与客户端诸如客户端122的交互可包括各种动作和事件。交互的非限制性实例包括游戏控制器移动、麦克风输入、摄像机输入等。交互可以并入客户端状态信息中。客户端状态信息可包括指示客户端的用户的数量的信息。所述信息可包括允许扣除活动用户的数量的信息。例如，客户端状态信息可包括由多个用户完成的动作。多个客户端可基于对应于由客户端的用户(诸如客户端122的用户128或用户130)做出的输入动作的最后时间度量推断出。与客户端相关联的活动用户的计数可用于确定是否维持渲染过程(诸如虚拟机108上的渲染过程110)的活动状态。例如，如果用户128和用户130停止与客户端122进行交互，那么渲染过程110可以进入不活动状态以使资源消耗最小化。这可以例如通过使虚拟机108暂停来完成。另一种可能性是用户128和用户130两者激活用于客户端过程124的中止功能。这可以例如由用户128和用户130两者切换到替代的应用而发生。

[0030] 图2描绘了包括至少一个客户端过程202和至少一个渲染服务API 204的客户端200的实施方案。客户端过程202可包括实现应用的功能的各个方面的一个或多个模块。应用包括但不限于视频游戏、视频回放服务、视频编辑、动画、模拟等。模块可包括计算机可读指令、电路等的任何组合。模块例如可包括静态和动态链接的程序库、商业对象、组件对象、图形处理单元、物理处理单元等。

[0031] 渲染服务API可以充当客户端过程202的组件。它可包括实现功能的各个部分的一个或多个模块。尽管渲染服务API 204在图2中被描绘成包含实现功能的特定部分(诸如应用关闭206)的各个模块，但是本领域的普通技术人员将了解，包括添加、减少、替换和重组的各种组合是可能的。

[0032] 应用关闭206描绘渲染服务API 204的接收与客户端过程202的运行状态相关的指示的模块。它例如可以接收客户端过程202将被关闭、暂停或中止的通知。然后，它可以向内容提供商发送通知，所述内容提供商包括在内容提供商内进行操作的任何代理或渲染过程。通知指示可以暂停与那个应用相关联的服务。

[0033] 用户输入监测208可涉及跟踪与客户端过程202相关的用户活动。它还涉及将用户活动表示为客户端状态信息，所述客户端状态信息可以从客户端传输到内容提供商。

[0034] 服务利用控制210可涉及监测影响客户端过程202有可能请求的渲染过程的水平的事件或对所述事件做出响应。这些事件包括游戏状态的转换，诸如从涉及激活游戏的状

态到涉及显示预先渲染的场景的状态，后者目前不需要渲染服务。客户端过程202可以通过在渲染服务API 204的模块上调用的功能调用、方法调用或类似技术来控制事件。服务利用事件可以例如作为客户端状态信息发送到内容提供商。

[0035] 硬件事件212可以检测各种硬件事件并对其做出响应，所述硬件事件包括但不限于客户端设备关闭、控制器关闭、显示器关闭、系统暂停、系统恢复等。这些事件还可以指示客户端过程202有可能请求的渲染过程的水平。指示硬件事件的信息可以例如作为客户端状态信息传输到内容提供商。

[0036] 服务利用统计214可以与涉及游戏、游戏发布商、硬件提供商或其他实体使用渲染服务的信息集合相关。在一些实施方案中，这些服务的某些方面可以由渲染服务API 204直接执行。这例如可以包括导致将要生成和/或传输的报告。报告可涉及各种使用统计，诸如代表特定客户端执行的使用的水平。在一些实施方案中，服务利用统计214可间接参与相关统计的生成，例如通过将识别信息传输到内容提供商。识别信息可包括用户信息、客户端设备信息、游戏信息、发布商信息等。可以采用加密技术来防止所提供的身份被篡改。

[0037] 服务重新定位216可涉及将内容提供商、代理、虚拟机或渲染过程与不同的地址(诸如不同的互联网协议(“IP”)地址)重新相关联。重新相关联可基于各种事件诸如利用平衡、故障转移或其他情境而发生。例如，对应于暂停的虚拟机的状态可以从一个主机移动到另一个，并且可以在移动之后指派不同的IP地址。实施方案还可以将渲染过程从一个虚拟机重新定位到另一个。

[0038] 图3描绘了在虚拟机302上托管渲染服务的内容提供商300的实施方案。尽管图3描绘了一个虚拟机，但内容提供商可以托管多个虚拟机。内容提供商可包括一个或多个计算设施，诸如数据中心、服务器集群、单独的服务器等。各个虚拟机可以在内容提供商内进行操作。例如，计算设施可以操作管理程序和一个或多个虚拟机在其上进行操作的各个计算主机。各种控制设施可以用来创建、移除、关闭、重新启动、暂停和再激活虚拟机。

[0039] 内容提供商可以管理虚拟机的状态信息。虚拟机可以具有可存储在存储设备上的相关联的状态。暂停或关闭的虚拟机的状态可存储在低延迟高速缓存中。低延迟高速缓存包括但不限于随机存取存储器、存储器内数据库、固态存储设备等。暂停的虚拟机的状态还可以存储在与低延迟选项相比可以被描述为高延迟的存储设备上。低延迟存储选项与高延迟存储选项之间的区别可基于除底层存储设备的质量之外的因素。替代地，在一些实施方案中，低延迟存储装置和高延迟存储装置可以通过总体系统延迟来区别，所述总体系统延迟可能受诸如网络速度、通信量拥塞等的因素影响。

[0040] 可以指派渲染过程304在虚拟机302上进行操作。渲染过程可以提供与图形处理相关的各种服务，诸如管理图形资源和渲染图形场景。渲染过程可以执行渲染流水线的一个或多个步骤。渲染流水线可包括各种数据结构、资源和阶段(诸如形状缓冲、顶点缓冲、纹理缓冲、输入装配、纹理映射、阴影、渲染目标等)。实施方案可包括对各种类型的流水线(包括但不限于2维和3维渲染、物理处理流水线等)的支持。

[0041] 各个模块可执行与渲染过程304的操作相关的功能。尽管图3将这些模块描绘成不同的实体，但应当理解，所描绘的模块可以各种组合的方式进行重新排序、重新布置、组合、改变或省略。此外，尽管在图3中被描绘为与渲染过程304和虚拟机302不同的元素，但由所描绘模块执行的功能中的一些或全部可以由渲染过程304和/或虚拟机302执行。

[0042] 用于客户端状态监测306的模块可以接收和/或处理涉及渲染过程的客户端的客户端状态信息。实施方案可以监测客户端状态信息以便确定用户活动的存在与否，并且可以进一步将所述活动分类为指示维持与客户端相关联的渲染过程的活动状态。渲染过程的活动状态可以由以下因素指示：诸如用户输入、客户端对维持活动状态的请求(不管不活动的时期)、客户端上新应用的初始化等。可基于接收到指示客户端上的活动的信息而将渲染过程维持在活动状态，所述信息可作为客户端状态信息传输到内容提供商。客户端上的活动例如可以是指接收到用户输入、对渲染服务的请求、渲染过程的活动执行等。

[0043] 客户端状态信息还可以指示到不活动状态的转换。缺乏活动可以对应于可暗示到渲染过程的不活动状态的转换的各种状况。例如，中断的网络连接可以中断渲染过程与客户端之间的通信，在这种情况下，实施方案可以致使渲染过程转换到不活动状态。客户端状态监测306可以跟踪或接收涉及接收到最后一次用户输入、新应用的初始化或涉及客户端活动的其他信息的信息。如果这个时间超过阈值，那么客户端状态监测306可以确定渲染过程304应当进入不活动状态。在接收对应于渲染过程的活动状态的信息时，实施方案可以致使不活动的渲染过程转换到活动状态。

[0044] 渲染过程可通过使渲染过程在其上运行的虚拟机暂停而转换到活动状态。在一些实施方案中，暂停的虚拟机可以保持在其主机的主存储器中，但是不消耗中央处理单元周期。在另一个实施方案中，暂停的虚拟机的状态可以存储在低延迟高速缓存中，或存储在其他形式的存储设备上。在另一个实施方案中，基于对应的渲染过程已不活动的时间段，暂停的虚拟机的状态可以从存储器内转换到低延迟高速缓存再到较高延迟存储装置。延迟最小化模块308可以执行各种动作以使访问渲染过程时的延迟最小化。这些动作可包括但不限于：基于与在虚拟机上执行的渲染过程相关联的客户端的状态，将暂停的虚拟机的状态移动到主存储器、高速缓存或相对较高延迟的存储装置或从它们移动暂停的虚拟机的状态。

[0045] 容量和利用管理模块310可以执行各种动作以便提高资源利用。在一些情况下并且在一些实施方案中，虚拟机302可以托管多个渲染过程。实施方案可基于以下因素将渲染过程指派给虚拟机：包括但不限于公共客户端设备、公共用户、公共游戏内容集等。实施方案还可以尝试通过根据改善诸如资源利用、延迟等的因素将渲染过程指派给虚拟机来改善这些因素。在一些实施方案中，容量和利用管理310可涉及基于计算资源的处理能力将渲染过程和/或虚拟机指派给这些资源。例如，在一些情况下并且在一些实施方案中，客户端状态信息可包含对渲染服务的请求水平的指示。基于所述指示，可选择虚拟机主机以部分地基于可用于所选择主机的处理能力来操作渲染过程。

[0046] 另一个模块可以执行与容量和利用统计312相关的各种动作。这些动作可包括维持关于提供给用户、客户端、应用、应用发布商等的渲染服务的记录。在一些实施方案中，内容提供商可以监测使用统计，诸如吞吐量、渲染的页面、消耗的CPU周期等。内容提供商可以采用诸如这些的各种使用统计来对各实体诸如游戏发布商征收服务费。

[0047] 图4描绘了用于管理渲染过程的方法的实施方案。尽管被描绘为操作序列，但本领域的普通技术人员应当理解，所描绘的次序不应被解释为限制本公开的范围，并且所描绘的操作中的至少一些可以改变、省略、重新排序或并行地执行。

[0048] 操作400可涉及接收指示向客户端提供渲染服务的信息。实施方案可以接收对应于将使用渲染服务的应用的初始化的客户端状态信息。初始化的应用实例和/或初始化的

应用实例在其上运行的客户端可以与渲染过程相关联。实施方案可以执行映射或查找操作以便确定应用实例和/或客户端是否已经与渲染过程相关联。如果没有，那么可以创建新的渲染过程。

[0049] 操作402描绘分配渲染过程可在其上执行的虚拟机。分配虚拟机可涉及重新使用虚拟机的现有实例、创建新实例、复制预先存在的实例等。实施方案例如可以形成虚拟机状态的预定义图像的副本，其中预定义的图像对应于被配置成执行渲染过程的虚拟机的初始状态。在一些实施方案中，虚拟机的图像可以在渲染过程已经开始执行但尚未服务于任何客户端的情况下使用。在一些情况下并且在一些实施方案中，虚拟机图像可包含已开始执行并且已加载对应于特定应用(诸如特定游戏程序)的资源的渲染过程。此时，可以在资源加载之后但在已将服务提供给客户端之前记录虚拟图像。当虚拟机准备好向客户端提供渲染服务时，可以存储并且随后加载对应于这些和其他初始状态的虚拟机的图像。

[0050] 可基于虚拟机能够访问与在渲染服务的客户端上运行的程序(诸如游戏)相关联的图形资源来分配用于执行渲染过程的虚拟机。在客户端上运行的游戏可依赖于在虚拟机上执行的渲染过程可访问的图形资源。能够访问图形资源可包括与其上存储有图形资源的存储设备的连通性。许多变化中的另一种可能的变化包括虚拟机图像，其中渲染过程已预先加载与游戏或其他应用相关的图形资源。

[0051] 操作404描绘将在虚拟机上执行的渲染过程与客户端相关联。各实施方案可以维持客户端与渲染过程和/或渲染过程在其上操作的虚拟机之间的关联的记录。在一些实施方案中，这些关联可以维持在独立于虚拟机进行操作的数据库或其他数据存储区中。所接收的客户端状态信息可以基于所记录的关联与渲染过程关联。例如，客户端状态信息可以由内容提供商接收，并且对应的渲染过程可基于所记录的关联来确定。

[0052] 渲染过程可以被描述为在虚拟机上执行或操作。术语“在...上执行”和“在...上操作”旨在包括渲染过程开始或继续在虚拟机上执行的多种状态。

[0053] 操作406描绘接收指示将渲染过程转换到不活动状态的信息。实施方案可基于超过了预测的最大不活动时期来确定渲染过程应转换到不活动状态。在各种实施方案中，可以记录时间以指示何时最后一次接收客户端状态信息，其中客户端状态信息指示对应的渲染过程应当保持为活动的。然后，实施方案可以将自从最后一次接收这种信息以来的时间量与阈值时间量进行比较，如果时间量超过阈值水平，那么实施方案可以将渲染过程转换到不活动状态。预测的最大不活动时期可以用作阈值。预测的最大值可以是部分地基于先前的活动或不活动时期确定的固定值或动态值。

[0054] 操作408描绘使虚拟机暂停以便去激活在虚拟机上进行操作的渲染过程。在一些实施方案中，暂停的虚拟机可保持在其主机的主存储器中，但其虚拟处理器可以停止执行。在另一个实施方案中，虚拟机可以从存储器中移除并且存储在存储设备上。实施方案例如可以将对应于虚拟机的状态存储在低延迟高速缓存中，或存储在其他存储设备上。实施方案可以执行这些动作的各种组合。例如，实施方案可基于由客户端给出的预期或预测的不活动时期来确定执行这些动作中的一个。例如，如果客户端状态信息指示游戏已进入已知的不活动时期(例如，当正在显示预先渲染的视频时)，虚拟机可保持不活动但处于存储器中。另一方面，如果客户端已关闭，那么虚拟机可以移动到相对较高延迟的存储设备或删除。

[0055] 当确定虚拟机上活动的渲染过程的数量已低于阈值水平时，可以使虚拟机暂停。在一些实施方案中，诸如将渲染过程一对一地映射到虚拟机的那些，阈值水平可以是一。其他实施方案可以在虚拟机上托管多个渲染过程。因此，当由虚拟机托管的所有渲染过程都已变成不活动时，实施方案可以使虚拟机暂停。当活动的渲染过程的数量低于阈值数量时，其他实施方案可以将活动的渲染过程转移到另一个虚拟机。

[0056] 操作410描绘接收指示将渲染过程从不活动状态转换到活动状态的信息。例如，渲染过程可能已基于中止事件而去激活。在接收指示客户端已被解除中止的客户端状态信息时，可以重新激活与客户端相关联的渲染过程。

[0057] 操作412描绘通过重新激活渲染过程正在其上执行的虚拟机来重新激活渲染过程。重新激活虚拟机可包括诸如从存储装置或从低延迟高速缓存检索虚拟机的状态的步骤。重新激活还可包括恢复与虚拟机相关联的虚拟处理器的执行。内容提供商可以发送涉及恢复渲染服务的操作的状态信息。例如，内容提供商可以发送指示恢复渲染服务的操作预期将花费的时间量的客户端信息。这可包括例如估计的完成时间、完成百分比等。

[0058] 图5描绘了用于管理渲染过程的方法的另一个实施方案。尽管被描绘为操作序列，但本领域的普通技术人员应当理解，所描绘的次序不应被解释为限制本公开的范围，并且所描绘的操作中的至少一些可以改变、省略、重新排序或并行地执行。

[0059] 客户端可以与可向客户端提供渲染服务的渲染过程相关联。渲染过程可以在基于多种因素所选择的虚拟机上执行。操作500至506描绘可用于将客户端与渲染过程相关联和/或选择用于执行渲染过程的虚拟机的因素的非限制性实例。

[0060] 操作500描绘将客户端与渲染过程一对一地相关联，以使得每个客户端被映射到仅向那个客户端提供渲染服务的渲染过程。实施方案还可以将渲染过程一对一地映射到应用，以使得给定的渲染过程仅向一个应用提供渲染服务。在一些实施方案中，可以将多个渲染过程指派给一个客户端，或者将多个客户端指派到一个渲染服务。

[0061] 操作502描绘将客户端与预先配置成渲染所请求内容的渲染过程相关联。例如，实施方案可以保持处于以下状态的虚拟机的图像：其中渲染过程已开始执行以准备提供与特定内容诸如特定游戏应用相关的渲染服务。这可以允许在渲染服务与客户端之间相关联之前预先加载各种线框、纹理等。

[0062] 操作504描绘确定在虚拟机上执行渲染过程，所述虚拟机基于正由在同一个虚拟机上执行的其他过程渲染的内容来选择。这种方法可允许在为同一个内容提供渲染服务的多个渲染过程之间共享包含图形资源的存储器块。

[0063] 操作506描绘在基于在一个虚拟机上为特定客户端分组渲染过程所选择的虚拟机上、或在最小数量的虚拟机上执行渲染过程。例如，客户端可以执行两个应用，每个应用都需要渲染服务。那么，应用可以与其自身的专用渲染过程相关联。可以将这些渲染过程分组以便在同一个虚拟机上执行。在一些实施方案中，与同一个客户端相关联的渲染过程在虚拟机上分组，以便将与其他客户端相关联的渲染过程从那个虚拟机上排除。

[0064] 操作508描绘管理在虚拟机上执行的渲染过程的活动状态和不活动状态。这可包括使虚拟机暂停以便去激活在虚拟机上进行操作的一个或多个渲染过程。在一些实施方案中，在采用多个渲染过程的情况下，可以在所有渲染过程都应转换到不活动状态时使虚拟机暂停。实施方案可以使在虚拟机上进行操作的所有渲染过程都处于活动状态，直到每个

渲染过程都能够转换到不活动状态，并且然后使虚拟机暂停。实施方案可以确定当发生各种事件时将渲染过程转换到不活动状态，所述事件包括但不限于在接收指示需要渲染服务的客户端状态信息时的延时。

[0065] 实施方案可以采用各种方法来准备用于执行渲染过程的虚拟机。图6描绘了用于维持虚拟机池(pool)的过程的实施方案。尽管被描绘为涉及操作序列，但本领域的普通技术人员应当理解，所描绘的次序不应被解释为限制本公开的范围，并且所描绘的操作中的至少一些可以改变、省略、重新排序或并行地执行。

[0066] 虚拟机初始化600涉及创建、启动和配置虚拟机以用于执行渲染过程的各个方面。可以执行各种操作，诸如通过操作602、604和606所描绘的那些，以便将虚拟机实例化以用于执行渲染过程。通过602、604和606所描绘的操作可以各种方式组合以便形成本公开的其他实施方案。

[0067] 操作602描绘检索被配置成执行渲染过程的虚拟机的状态信息。虚拟机状态可以存储为文件，并且有时可以被称为虚拟机图像。在各种实施方案中，虚拟机状态可对应于已被预先配置成执行渲染过程的虚拟机。例如，虚拟机可具有安装的操作系统。用于执行渲染过程的任何文件都可能已经被复制到虚拟机，并且任何必要的配置步骤都可能已经执行。

[0068] 操作604描绘检索已经在执行渲染过程的虚拟机的状态信息。实施方案可以利用虚拟机图像，其中虚拟机的状态反映当前正在执行但尚未与客户端相关联或正在向客户端提供服务的渲染过程。

[0069] 操作606描绘检索被配置成执行(或已经在执行)与特定内容相关的渲染过程的虚拟机的状态信息。例如，可以在渲染过程已开始执行并且已加载了用于特定应用(诸如游戏)的资源时保存虚拟机图像。所述资源可包括位图、纹理、线框模型等。

[0070] 实施方案可以从所述池中选择虚拟机，基于所选择的虚拟机被配置成执行能够访问与应用相关的图形资源的渲染过程，正在提供的图形渲染服务用于所述应用。在一些情况下，渲染过程可以被特定地配置成能够访问图形资源。在其他情况下，所选择的虚拟机可以被配置成能够访问图形资源。

[0071] 池组织608是指可以执行来形成用于执行渲染过程的虚拟机池的各种操作。如本文所用，术语池可以是指对象的各种集合，诸如虚拟机的集合。所述集合可以通过数据结构诸如列表或阵列、或通过各种分类方法进行组织。在各种实施方案中，池可以是根据需要可从所述池中抽取的对象或资源的集合或集。例如，池可以包含可从所述池中抽取并且用于执行渲染过程的未使用的虚拟机集。

[0072] 操作610描绘维持虚拟机池。维持池可涉及创建虚拟机、将虚拟机放置在池中、从池中抽取虚拟机以及将虚拟机返回到池。实施方案可在池中维持最小和/或最大数量的自由虚拟机。

[0073] 操作612描绘基于内容维持虚拟机池。这可涉及将池中的虚拟机分类为归属于特定的内容集。当从池中抽取虚拟机时，可以抽取适于特定内容集的虚拟机。例如，可以将被配置成预先加载用于特定游戏的资源的虚拟机分组到池中。当客户端请求用于那个特定游戏的渲染服务时，可以使用来自那个池的虚拟机。与不同应用相关联的客户端可以获得从不同池中提取的虚拟机。

[0074] 操作614描绘基于延迟最小化维持虚拟机池。实施方案可以采用多种因素来组织

此类池。在一个实施方案中，根据每个虚拟机的主机的地理位置来将虚拟机分组。可基于主机的位置和请求渲染服务的客户端的位置来从池抽取虚拟机。还可根据速度、容量等来组织池。

[0075] 实施方案可以将池组合成池和子池的各种组合。例如，可以将虚拟机分组到通过内容组织的池中，并且分组到通过地理区域组织的子池中。

[0076] 池维持616涉及执行来抽取、利用虚拟机并且将虚拟机返回到虚拟机池的各种操作。如通过操作618所描绘的，可以从池中抽取虚拟机以便向新客户端提供渲染服务。操作620描绘在不将立即虚拟机返回到池的情况下使虚拟机暂停以便暂时性地去激活渲染过程。在稍后的某个时间，可以将虚拟机返回到池。操作622描绘一个实例，其中在与客户端断开连接时将虚拟机返回到池。一些实施方案可以在不再需要虚拟机时将它们删除。操作624描绘重新补足虚拟机池，例如通过执行与虚拟机初始化600相关联的操作中的一个或多个。可以重新补足虚拟机以便保持池中可用的最小数量的虚拟机，或以便替换已从池中移除和随后删除的虚拟机。

[0077] 在一些情况下，内容提供商可以渲染内容项视图并通过电子网络诸如互联网将其传输到客户端。在一些情况下，可以根据客户端的请求使用例如流传输内容递送技术来提供内容。现在将详细描述能够渲染内容并将其传输到客户端的示例性计算环境。具体地，图7示出可在实现本文所述的实施方案的示例性计算环境。图7是示意性地示出数据中心710的实例的图，所述数据中心710可经由通信网络730、通过用户计算机702a和702b(其在本文中以单数形式可被称为一个计算机702或以复数形式可被称为这些计算机702)向用户700a和700b(其在本文中以单数形式可被称为一个用户700或以复数形式可被称为这些用户700)提供计算资源。数据中心710可以被配置成提供用于永久性地或根据需要执行应用的计算资源。由数据中心710提供的计算资源可包括各种类型的资源，诸如网关资源、负载平衡资源、路由资源、联网资源、计算资源、易失性和非易失性存储器资源、内容递送资源、数据处理资源、数据存储资源、数据通信资源等。计算资源可以是通用的或者可用于多种特定配置。例如，数据处理资源可以用作可被配置成提供各种web服务的虚拟机实例。此外，资源的组合可通过网络而变得可用并且可以被配置为一个或多个web服务。所述实例可以被配置成执行包括web服务的应用，所述web服务诸如应用服务、媒体服务、数据库服务、处理服务、网关服务、存储服务、路由服务、安全服务、加密服务、负载平衡服务、应用服务等。这些服务可以被配置成具有设定或定制的应用，并且可以在大小、执行、成本、延迟、类型、持续时间、可访问性以及任何其他维度方面进行配置。这些web服务可以被配置为可用于一个或多个客户端的基础结构，并且可包括被配置为用于一个或多个客户端的平台或软件的一个或多个应用。这些web服务可通过一个或多个通信协议而变得可用。数据存储资源可包括文件存储设备、区块存储设备等。

[0078] 每种类型或配置的计算资源可以具有不同大小，诸如由许多处理器、大量存储器和/或大的存储容量组成的大型资源和由较少的处理器、较少量存储器和/或较小的存储容量组成的小型资源。例如，客户可选择分配如web服务器的许多小型处理资源和/或如数据库服务器的一个大型处理资源。

[0079] 数据中心710可包括提供计算资源的服务器716a～b(其在本文中以单数形式可被称为一个服务器716或以复数形式可被称为这些服务器716)。这些资源可用作裸露金属资

源或用作虚拟机实例718a-d以及(其在本文中以单数形式可被称为一个虚拟机实例718或以复数形式可被称为这些虚拟机实例718)。虚拟机实例718c和718d是共享状态虚拟机("SSVM")实例。SSVM虚拟机实例718c和718d可以被配置成执行共享内容项状态技术和根据本公开并且下文详细描述的任何其他所公开技术的全部或任何部分。如应当理解的,虽然图7所示的特定实例在每个服务器中包括一个SSVM虚拟机,但这仅仅是一个实例。服务器可包括多于一个SSVM虚拟机或者可能不包括任何SSVM虚拟机。

[0080] 用于计算硬件的虚拟化技术的可用性已提供多种益处,用于向客户提供大型计算资源并且允许在多个客户之间有效地且安全地共享计算资源。例如,通过向每个用户提供物理计算设备托管的一个或多个虚拟机实例,虚拟化技术可允许在多个用户之间共享物理计算设备。虚拟机实例可以是充当不同逻辑计算系统的特定物理计算系统的软件仿真。这种虚拟机实例在共享给定的物理计算资源的多个操作系统之间提供隔离。此外,一些虚拟化技术可以提供跨越一个或多个物理资源的虚拟资源,诸如具有跨越多个不同物理计算系统的多个虚拟处理器的虚拟机实例。

[0081] 参考图7,通信网络730例如可以是链接网络中的公共可访问网络并且可能由各个不同方进行操作,诸如互联网。在其他实施方案中,通信网络730可以是专用网络,诸如完全或部分地不能被非特权用户访问的公司或大学网络。在另外其他的实施方案中,通信网络730可包括具有对互联网的访问和/或从互联网进行访问的一个或多个专用网络。

[0082] 通信网络730可以提供对计算机702的访问。用户计算机702可以是由用户700或数据中心710的其他客户使用的计算机。例如,用户计算机702a或702b可以是服务器、台式或便携式个人计算机、平板计算机、无线电话、个人数字助理(PDA)、电子书阅读器、游戏控制台、机顶盒或能够访问数据中心710的任何其他计算设备。用户计算机702a或702b可以直接连接到互联网(例如,通过电缆调制解调器或数字用户线路(DSL))。尽管仅描绘了两个用户计算机702a和702b,但是应当理解,可以存在多个用户计算机。

[0083] 用户计算机702也可用来配置由数据中心710提供的计算资源的各方面。就这一点而言,数据中心710可提供网关或web接口,通过所述网关或web接口,可通过使用在用户计算机702上执行的web浏览器应用程序来配置数据中心的操作的各方面。或者,在用户计算机702上执行的独立应用程序可访问由数据中心710公开的应用编程接口(API)以便执行配置操作。还可以利用用于配置在数据中心710可用的各种web服务的操作的其他机制。

[0084] 图7所示的服务器716可以是被适当地配置用于提供上述计算资源的标准服务器,并且可以提供用于执行一个或多个web服务和/或应用的计算资源。在一个实施方案中,计算资源可以是虚拟机实例718。虚拟机实例还可以被称为虚拟机。如上所述,虚拟机实例718中的每一个可被配置成执行应用的全部或一部分。在虚拟机实例的实例中,数据中心710可以被配置成执行能够执行虚拟机实例718的实例管理程序720a或720b(其在本文中以单数形式可被称为一个实例管理程序720或以复数形式可被称为这些实例管理程序720)。例如,实例管理程序720可以是虚拟机监视程序(VMM)或被配置成允许在服务器716上执行虚拟机实例718的另一类型的程序。应当理解,实例管理程序720的配置(如图7所描绘的)可以改变,并且实例管理程序720例如可以被配置成作为路由器714的前端进行操作。在一些实施方案中,实例管理程序720可以托管在服务器716上或其他计算节点上。

[0085] 应当理解,尽管上文所公开的实施方案讨论了虚拟机实例的情形,但是其他类型

的实施方式可以与本文所公开的概念和技术一起使用。例如，本文所公开的实施方案也可以与并不使用虚拟机实例的计算系统一起使用。

[0086] 在图1所示的示例性数据中心710中，路由器714可用来互连服务器716a和716b。路由器714也可连接到网关740，网关740连接到通信网络730。路由器714可以连接到一个或多个负载平衡器，并且可以单独地或以组合方式在数据中心710中管理网络内的通信，例如通过基于此类通信的特性(例如，包括源地址和/或目的地地址、协议标识符、大小、处理要求等的标头信息)和/或专用网络的特性(例如，基于网络拓扑等的路由)来适当地转发数据包或其他数据通信。应当理解，为简单起见，示出了这个实例的计算系统和其他设备的各方面，而并未展示某些常规的细节。另外计算系统和其他设备在其他实施方案中可以互连并且可以以不同的方式互连。

[0087] 应当理解，图7所示的网络拓扑已大大简化，并且可以利用更多的网络和联网设备来互连本文所公开的各种计算系统。这些网络拓扑和设备对本领域的技术人员来说应是明显的。

[0088] 也应当理解，图7中所描述的数据中心710仅仅是说明性的，并且可利用其他实现方式。另外，应当理解，本文所公开的功能可以以软件、硬件或软件与硬件的组合来实现。其他实现方式对本领域的技术人员来说是应明显的。也应理解，服务器、网关或其他计算设备可包括可交互并且执行所述类型的功能的硬件或软件的任何组合，其包括但不限于台式计算机或其他计算机、数据库服务器、网络存储设备和其他网络设备、PDA、平板计算机、移动电话、无线电话、寻呼机、电子管理器、互联网电器、(例如，使用机顶盒和/或个人/数字视频记录器的)基于电视机的系统和包括适当通信能力的各种其他消费品。此外，由所示模块提供的功能在一些实施方案中可以以较少的组件来组合或分布于另外的模块中。类似地，在一些实施方案中，可以不提供所示模块中的一些的功能和/或可以使用其他另外的功能。

[0089] 在至少一些实施方案中，实现本文中所述的技术中的一种或多种的部分或全部的服务器可包括通用计算机系统，所述通用计算机系统包括一种或多种计算机可访问介质或被配置成访问一种或多种计算机可访问介质。图8描绘了包括一种或多种计算机可访问介质或被配置成访问一种或多种计算机可访问介质的通用计算机系统。在所示的实施方案中，计算设备800包括通过输入/输出(I/O)接口830耦接到系统存储器820的一个或多个处理器810a、810b和/或810n(其在本文中以单数形式可被称为“一个处理器810”或以复数形式可被称为“这些处理器810”)。计算设备800还包括耦接到I/O接口830的网络接口840。

[0090] 在各种实施方案中，计算设备800可为包括一个处理器810的单处理器系统，或包括若干处理器10(例如两个、四个、八个或另一合适数量)的多处理器系统。处理器810可为能够执行指令的任何合适的处理器。例如，在各种实施方案中，处理器810可为实现多种指令集架构(ISA)中任何一种架构的通用或嵌入式处理器，所述架构诸如x86、PowerPC、SPARC、或MIPS ISA或任何其他合适的ISA。在多处理器系统中，每一个处理器810可通常但不一定实现相同的ISA。

[0091] 在一些实施方案中，图形处理单元(“GPU”)812可以参与提供图形渲染和/或物理处理能力。GPU例如可以包括专门用于图形计算的高度并行处理器架构。在一些实施方案中，处理器810和GPU 812可以实现为相同类型的设备中的一个或多个。

[0092] 系统存储器820可以被配置成存储可由处理器810访问的指令和数据。在各种实施

方案中，系统存储器820可使用任何合适的存储器技术来实现，所述存储器技术诸如静态随机存取存储器(“SRAM”)、同步动态RAM(“SDRAM”)、非易失性/快闪型存储器或任何其他类型的存储器。在所示的实施方案中，实现一种或多种所需功能的程序指令和数据(诸如以上所述的那些方法、技术和数据)被示出作为代码825和数据826存储在系统存储器820内。

[0093] 在一个实施方案中，I/O接口830可以被配置成协调处理器810、系统存储器820和设备中的任何外围装置之间的I/O通信量，所述外围装置包括网络接口840或其他外围接口。在一些实施方案中，I/O接口830可以执行任何必需的协议、时序或其他数据转换以便将来自一个组件(例如，系统存储器820)的数据信号转换成适于由另一个组件(例如，处理器810)使用的格式。在一些实施方案中，I/O接口830可包括对于通过各种类型的外围总线附接的设备的支持，所述外围总线例如像外围组件互连(PCI)总线标准或通用串行总线(USB)标准的变化形式。在一些实施方案中，I/O接口830的功能可分离到两个或更多个单独的组件中，例如像北桥和南桥。另外，在一些实施方案中，I/O接口830的一些或所有功能，诸如到系统存储器820的接口，可直接并入处理器810中。

[0094] 网络接口840可以被配置成允许数据在计算设备800与附接到一个或多个网络850的其他一个或多个设备860(例如像其他计算机系统或设备)之间进行交换。在各种实施方案中，网络接口840可以支持通过任何合适的有线或无线通用数据网络(例如像以太网网络类型)进行通信。另外，网络接口840可以支持通过电信/电话网络诸如模拟语音网络或数字光纤通信网络、通过存储区域网络诸如光纤信道SAN(存储区域网络)或通过任何其他合适类型的网络和/或协议进行通信。

[0095] 在一些实施方案中，系统存储器820可以是被配置成存储如上所述的程序指令和数据以用于实现对应方法和装置的实施方案的计算机可访问介质的一个实施方案。然而，在其他实施方案中，可以在不同类型的计算机可访问介质上接收、发送或存储程序指令和/或数据。一般来说，计算机可访问介质可以包括非暂时性存储介质或存储器介质，诸如磁性介质或光学介质，例如通过I/O接口830耦接到计算设备800的磁盘或DVD/CD。非暂时性计算机可访问存储介质还可以包括可作为系统存储器820或另一类型的存储器被包括在计算设备800的一些实施方案中的任何易失性或非易失性介质，诸如RAM(例如，SDRAM、DDR SDRAM、RDRAM、SRAM等)、ROM等。另外，计算机可访问介质可以包括传输介质或信号，诸如经由通信介质(诸如网络和/或无线链路)传送的电信号、电磁信号或数字信号，诸如可通过网络接口840来实现的那些。多个计算设备的部分或全部(诸如图8所示的那些)可用于实现各种实施方案中所述的功能；例如，在多个不同设备和服务器上运行的软件组件可以协作来提供功能。在一些实施方案中，除了或代替使用通用计算机系统来实现，所描述功能的部分可以使用存储设备、网络设备或专用计算机系统来实现。如本文所用，术语“计算设备”是指至少所有这些类型的设备并且不限于这些类型的设备。

[0096] 计算节点(compute node)，也可以被称为计算性节点(computing node)，可以在广泛多种计算环境上实现，诸如平板计算机、个人计算机、智能电话、游戏控制台、商用硬件计算机、虚拟机、web服务、计算集群和计算器具。为方便起见，这些计算设备或环境中的任一个可以被描述为计算节点或计算性节点。

[0097] 由实体(诸如公司或公共部门组织)建立的用于向一组分布式客户端提供通过互联网和/或其他网络可访问的一个或多个web服务(诸如各种类型的基于云的计算或存储)

的网络可以被称为提供商网络。这种提供商网络可包括托管各种资源池的多种数据中心，诸如物理和虚拟化计算机服务器、存储设备、网络设备等的集合，所述资源池是实现并分布由提供商网络提供的基础设施和web服务所需要的。在一些实施方案中，所述资源可以以与web服务相关的各种单元提供给客户端，诸如用于存储的存储容量的量、用于处理的处理能力、作为实例、作为相关服务集等。虚拟计算实例例如可包括具有指定的计算能力(其可以通过指示CPU的类型和数量、主存储器大小等来指定)和指定的软件堆叠(例如，特定版本的操作系统，其继而可在管理程序之上运行)的一个或多个服务器。

[0098] 多种不同类型的计算设备可以单独地或以组合方式用来在不同实施方案中实现提供商网络的资源，所述实施方案包括通用或专用计算机服务器、存储设备、网络设备等。在一些实施方案中，可以向客户端或用户提供对资源实例的直接访问，例如通过给予用户管理员登录名和密码。在其他实施方案中，提供商网络运营商可允许客户端代表适用于应用的执行平台(诸如应用服务器实例、Java™虚拟机(JVM)、通用或专用操作系统、支持各种解释或编译编程语言诸如Ruby、Perl、Python、C、C++等的平台或高性能计算平台)上的客户端，指定对所指定客户端应用和应用的调度执行的执行要求，而不例如要求客户端直接访问实例或执行平台。在一些实现方式中，给定的执行平台可以利用一个或多个资源实例；在其他实现方式中，多个指定平台可以映射到单个资源实例。

[0099] 在许多环境中，实现不同类型的虚拟化计算、存储和/或其他网络可访问功能的提供商网络的运营商可允许客户以各种资源获取模式预订或购买对资源的访问。计算资源提供商可以向客户提供设施，以便选择和启动所期望计算资源、将应用组件部署到计算资源并且维持应用在环境中执行。此外，计算资源提供商可以向客户提供其他设施，以便根据应用改变的需要或容量要求、手动地或通过自动缩放快速且容易地放大或缩小分配到应用的资源的数量和类型。计算资源提供商提供的计算资源可以以离散单元变得可用，所述离散单元可以被称为实例。实例可表示物理服务器硬件平台、在服务器上执行的虚拟机实例、或两者的某种组合。各种类型和配置的实例都可以变得可用，包括执行不同操作系统(OS)和/或管理程序并且具有各种安装的软件应用、运行时间等的不同大小的资源。实例还可以在特定可用性区中可用，所述特定可用性区表示例如逻辑区域、容错区域、数据中心或底层计算硬件的其他地理位置。实例可以在可用性区内或跨可用性区复制，以便改善实例的冗余，并且实例可以在特定可用性区内或跨可用性区迁移。作为一个实例，可用性区中客户端与特定服务器通信的延迟可以小于客户端与不同服务器通信的延迟。因此，实例可以从较高延迟服务器迁移到较低延迟服务器以便改善总体客户端体验。

[0100] 在一些实施方案中，提供商网络可被组织成多个地理区域，并且每个区域可包括一个或多个可用性区。可用性区(其还可以被称为可用性容器)继而可包括一个或多个不同的位置或数据中心，其被配置成使得给定可用性区中的资源可以与其他可用性区中的故障隔离或绝缘。也就是说，预期一个可用性区中的故障可能不会导致任何其他可用性区中的故障。因此，一个资源实例的可用性简档旨在独立于不同可用性区中的资源实例的可用性简档。客户端可能够通过启动相应可用性区中的多个应用实例来保护其应用免于单个位置处的故障。同时，在一些实现方式中，可以在驻留于同一地理区域内的资源实例之间提供廉价且低延迟的网络连通性(并且同一可用性区的资源之间的网络传输甚至可以更快)。

[0101] 在前述部分中所描述的过程、方法和算法中的每一个可体现在由一个或多个计算

机或计算机处理器执行的代码模块中，并且完全或部分地由所述代码模块自动进行。所述代码模块可存储在任何类型的非暂时性计算机可读介质或计算机存储设备(诸如硬盘、固态存储器、光盘和/或类似设备)上。所述过程和算法可部分地或全部地在专用电路中实现。所公开的过程和处理步骤的结果可永久地或以其他方式存储在任何类型的非暂时性计算机存储装置(例如像易失性或非易失性存储装置)中。

[0102] 鉴于以下条款，上述内容可更好地理解：

[0103] 1.一种系统，其包括：

[0104] 一个或多个计算节点，所述一个或多个计算节点被配置成来操作用于代表一个或多个客户端渲染图形的服务，所述服务包括多个虚拟机；

[0105] 一个或多个计算节点被配置成至少：

[0106] 接收指示代表所述一个或多个客户端渲染图形的请求，所述请求包括指示与在所述一个或多个客户端上运行的过程相关联的图形资源集的信息；

[0107] 确定激活所述多个虚拟机中的虚拟机，所述确定至少部分地基于所述虚拟机被配置成执行对应于所述图形资源集的渲染过程；

[0108] 响应于尚未接收到将所述渲染过程保持在活动状态的请求的第一确定和自从接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息以来的时间量已超过第一阈值的第二确定，暂停所述虚拟机的操作，其中所述虚拟机在暂停时的第一状态包括用于所述渲染过程的第二状态；以及

[0109] 响应于接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息，恢复所述虚拟机的操作。

[0110] 2.如条款1所述的系统，所述一个或多个计算节点还被配置成至少：

[0111] 将所述虚拟机的所述第一状态存储在低延迟高速缓存中；以及

[0112] 响应于自从接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息以来的时间量已超过第二阈值，将所述虚拟机的所述第一状态存储在存储设备上。

[0113] 3.如条款1所述的系统，所述一个或多个计算节点还被配置成至少：

[0114] 至少部分地基于使所述虚拟机暂停来激活另外的虚拟机。

[0115] 4.如条款1所述的系统，其中对将所述渲染过程保持在活动状态的所述请求是由在所述一个或多个客户端中的客户端上进行操作的过程发送的。

[0116] 5.一种非暂时性计算机可读存储介质，其具有存储在其上的指令，所述指令在由一个或多个计算设备执行时致使所述一个或多个计算设备至少：

[0117] 接收指示为一个或多个客户端执行图形渲染服务的请求，所述一个或多个客户端执行与图形资源集相关联的过程；

[0118] 至少部分地基于选择用于激活的虚拟机被配置成执行能够访问所述图形资源集的渲染过程来激活所述虚拟机；

[0119] 由所述渲染过程来为所述一个或多个客户端执行所述图形渲染服务；

[0120] 至少部分地基于确定尚未接收到对保持所述渲染过程为活动的请求并且至少部分地基于确定自从接收到指示由所述一个或多个客户端中至少一个进行的活动的信息以来的时间量已超过第一阈值时间量，暂停所述虚拟机的操作；以及

[0121] 响应于接收到指示由所述一个或多个客户端中至少一个进行的活动的信息，恢复

所述虚拟机的操作。

[0122] 6.如条款5所述的非暂时性计算机可读介质，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

[0123] 将所述虚拟机的状态存储在低延迟高速缓存中。

[0124] 7.如条款5所述的非暂时性计算机可读介质，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

[0125] 响应于自从接收到指示由所述一个或多个客户端中至少一个进行的活动的信息以来的时间量已超过第二阈值时间量，将所述虚拟机的状态存储在存储设备上。

[0126] 8.如条款5所述的非暂时性计算机可读介质，其中暂停所述虚拟机的所述操作包括将所述虚拟机的状态保持在存储器中。

[0127] 9.如条款5所述的非暂时性计算机可读介质，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

[0128] 响应于确定不运行渲染过程的活动虚拟机的数量已低于阈值，激活另外的虚拟机。

[0129] 10.如条款5所述的非暂时性计算机可读介质，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

[0130] 至少部分地基于确定所述虚拟机上活动的渲染过程的数量已低于阈值水平，暂停所述虚拟机的操作。

[0131] 11.如条款5所述的非暂时性计算机可读介质，其中对保持所述渲染过程为活动的所述请求是由在所述一个或多个客户端中的一个上运行的过程发起的。

[0132] 12.如条款11所述的非暂时性计算机可读介质，其中响应于进入没有用户输入预期达一定时间段的状态，所述过程发起所述请求。

[0133] 13.如条款5所述的非暂时性计算机可读介质，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

[0134] 向所述一个或多个客户端中的客户端发送指示所述渲染过程的恢复操作的时间的信息。

[0135] 14.如条款5所述的非暂时性计算机可读介质，其还包括指令，所述指令在由所述计算设备执行时致使所述计算设备至少：

[0136] 将渲染过程的状态从所述虚拟机传递到另一个虚拟机。

[0137] 15.一种方法，其包括：

[0138] 接收指示为一个或多个客户端执行图形渲染服务的请求，所述一个或多个客户端执行与图形资源集相关联的过程；

[0139] 至少部分地基于所选择用于执行渲染过程的虚拟机能够访问所述图形资源集，在所述虚拟机上执行所述渲染过程；

[0140] 由在所述虚拟机上进行操作的所述渲染过程来为所述一个或多个客户端执行所述图形渲染服务；

[0141] 至少部分地基于尚未接收到对将所述渲染过程保持在活动状态的请求的第一确定和自从接收到指示由所述一个或多个客户端中至少一个的用户提供的输入的信息以来的时间量已超过第一阈值的第二确定，暂停所述虚拟机的操作；以及

[0142] 响应于接收到指示接收对代表所述一个或多个客户端执行图形渲染的请求的信息，恢复所述虚拟机的操作。

[0143] 16.如条款15所述的方法，其中指示接收对代表所述一个或多个客户端执行图形渲染的请求的所述信息对应于由所述一个或多个客户端中至少一个的用户提供的输入。

[0144] 17.如条款15所述的方法，其中对将所述渲染过程保持在活动状态的所述请求至少部分地基于进入没有用户输入预期达一定时间段的状态。

[0145] 18.如条款15所述的方法，其还包括：

[0146] 将所述虚拟机的状态存储在低延迟高速缓存中达至少等于第二阈值的时间段。

[0147] 19.如条款15所述的方法，其还包括：

[0148] 将所述虚拟机重新设定到初始状态，所述初始状态对应于在执行所述渲染过程之前的所述虚拟机的状态。

[0149] 20.如条款15所述的方法，其还包括：

[0150] 向所述一个或多个客户端中的客户端发送指示恢复所述渲染过程的状态的信息。

[0151] 上文所述的各种特征和过程可以彼此独立地使用，或者可以以各种方式进行组合。所有可能的组合和子组合意图落入本公开的范围内。此外，在一些实现方式中，某些方法和过程块可省略。本文所述的方法和过程也不限于任何特定的顺序，并且与之相关的块或状态可以按其他适当的顺序执行。例如，所述的块或状态可以按不同于已特别公开的次序的次序执行，或多个块或状态可组合在单个块或状态中。示例性块或状态可串行地、并行地或以某种其他方式执行。块或状态可被添加到所公开的示例性实施方案或可从这些实施方案中移除。本文所述的示例性系统和组件可以以不同于所述方式的方式进行配置。例如，与所公开的示例性实施方案相比，元件可被添加、移除或重新布置。

[0152] 还应了解，各种项目被示出为在使用时存储在存储器中或存储设备上，且为了存储器管理和数据完整性，可以在存储器与其他存储设备之间传递这些项目或它们的部分。或者，在其他实施方案中，一些或所有的软件模块和/或系统可以在另一设备上的存储器中执行，并且通过计算机间通信来与所示的计算系统通信。另外，在一些实施方案中，可以其他方式、诸如至少部分地在固件和/或硬件中来实现或提供所述系统和/或模块中的一些或全部，所述硬件包括但不限于一个或多个专用集成电路(ASIC)、标准集成电路、控制器(例如，通过执行适当的指令并且包括微控制器和/或嵌入式控制器)、现场可编程门阵列(FPGA)、复杂可编程逻辑设备(CPLD)等。所述模块、系统和数据结构中的一些或全部也可(例如作为软件指令或结构化数据)存储在计算机可读介质(诸如硬盘驱动、存储器、网络或通过适当驱动器或经由适当连接来读取的便携式媒体物品)上。所述系统、模块和数据结构也可作为所生成的数据信号(例如，作为载波或其他模拟或数字传播信号的一部分)在多种计算机可读传输介质(包括基于无线的介质和基于有线/电缆的介质)上传输，并且可采取多种形式(例如，作为单个模拟信号或多路复用的模拟信号的一部分，或作为多个离散的数字数据包或帧)。在其他实施方案中，此类计算机程序产品也可采取其他形式。因此，本发明可用其他计算机系统配置来实践。

[0153] 本文所用的条件语言，诸如尤其是“可(can)”、“可以(could)”、“可能(might和may)”、“例如(e.g.)”等，除非另外明确说明或在使用的背景下以其他方式进行理解，否则通常旨在传达：某些实施方案包括、而其他实施方案不包括某些特征、元件和/或步骤。因

此，这种条件语言通常并非意图暗示所述特征、元件和/或步骤无论如何都是一个或多个实施方案所必需的，或者并非暗示一个或多个实施方案必须包括用于在借助或不借助输入或者提示的情况下决定是否包括这些特征、元件和/或步骤或是否在任何特定实施方案中实施这些特征、元件和/或步骤的逻辑。术语“包括”、“包含”、“具有”等是同义的，并以开放的方式包含性地使用，而且不排除另外的元件、特征、动作、操作等。另外，术语“或者”以其包含性意义(并且不以其排除性意义)使用，从而使得当(例如)用来连接一列表元件时，术语“或者”意味着所述列表中元件的一个、一些或全部。

[0154] 尽管已经描述某些示例性实施方案，但是这些实施方案仅通过实例呈现，且并非意图限制本文所公开的发明的范围。因此，在前文描述中没有内容意在暗示任何特定特征、特性、步骤、模块或方框是必须的或不可缺少的。实际上，本文所述的新颖方法和系统可通过各种其他形式来体现；另外，在不脱离本文所公开的发明的精神的情况下，可对本文所述的方法和系统的形式做出各种省略、替代以及改变。所附权利要求书和其等效物意图涵盖将会落在本文所公开的发明的范围和精神内的此类形式或修改。

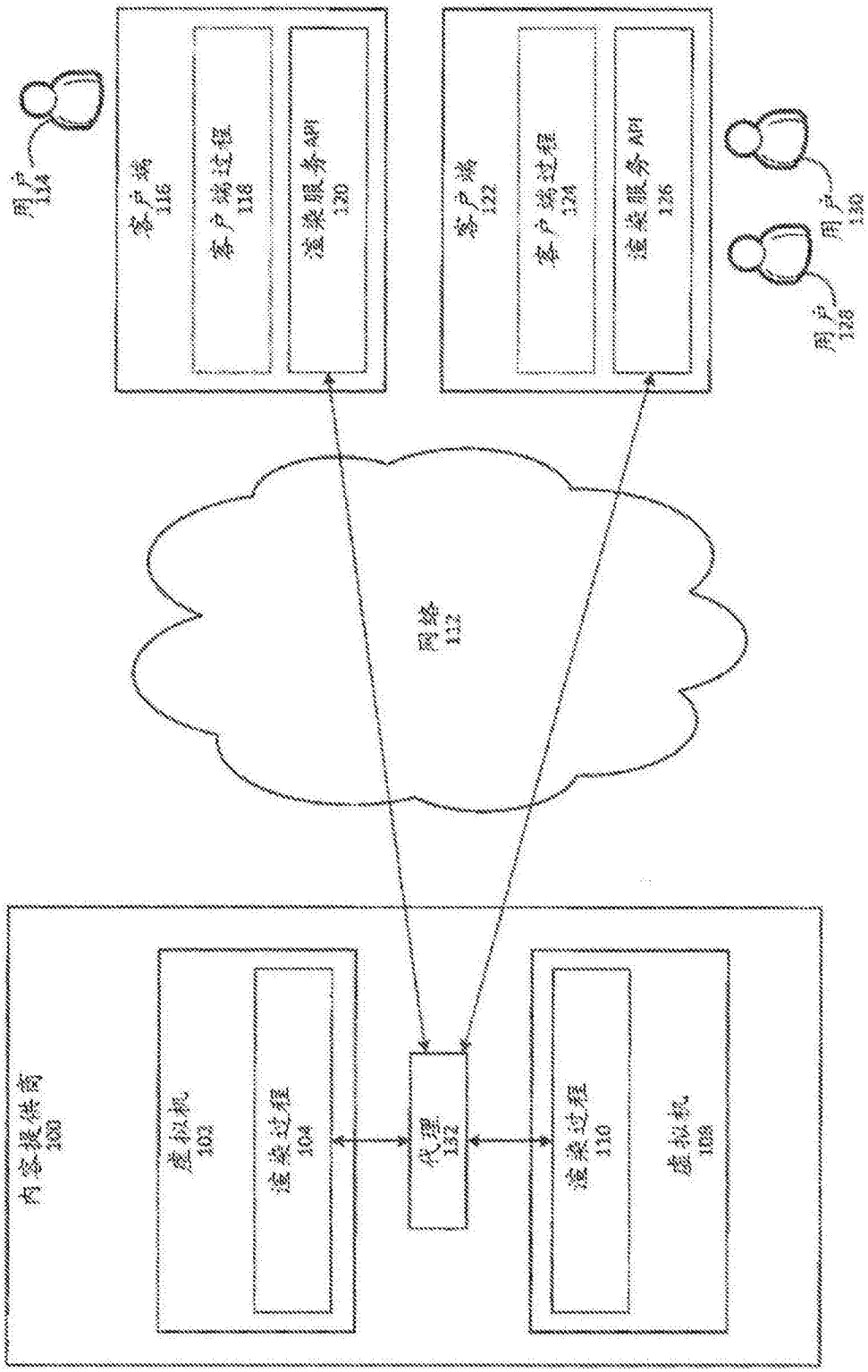

用户
114
客户端
116
客户端过程
118
渲染服务API
120
客户端
122
客户端过程
124
渲染服务API
126
用户
130
用户
128
网络
112
内容提供商
100
虚拟机
102
渲染过程
104
代理
132
渲染过程
110
虚拟机
108

图1

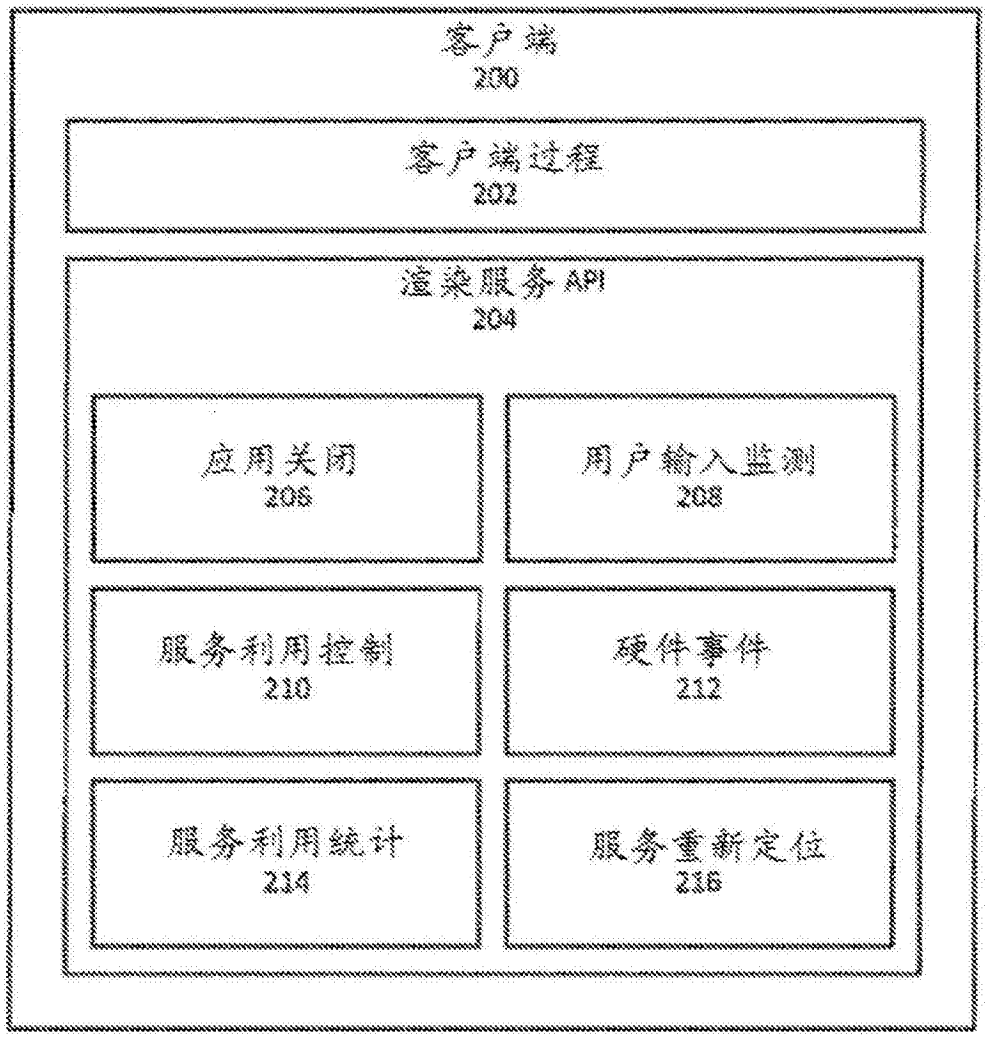

客户端
200
客户端过程
202
渲染服务API
204
应用关闭
206
用户输入监测
208
服务利用控制
210
硬件事件
212
服务利用统计
214
服务重新定位
216

图2

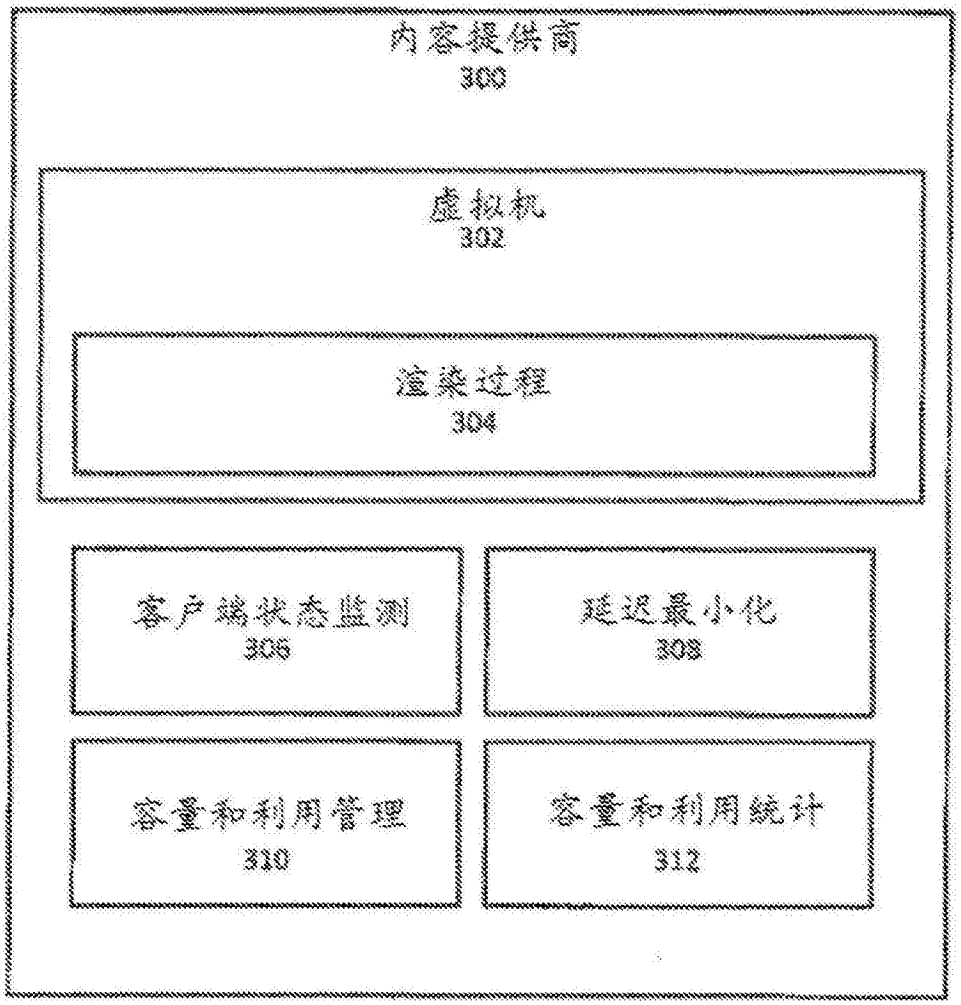

内容提供商
300
虚拟机
302
渲染过程
304
客户端状态监测
306
延迟最小化
308
容量和利用管理
310
容量和利用统计
312

图3

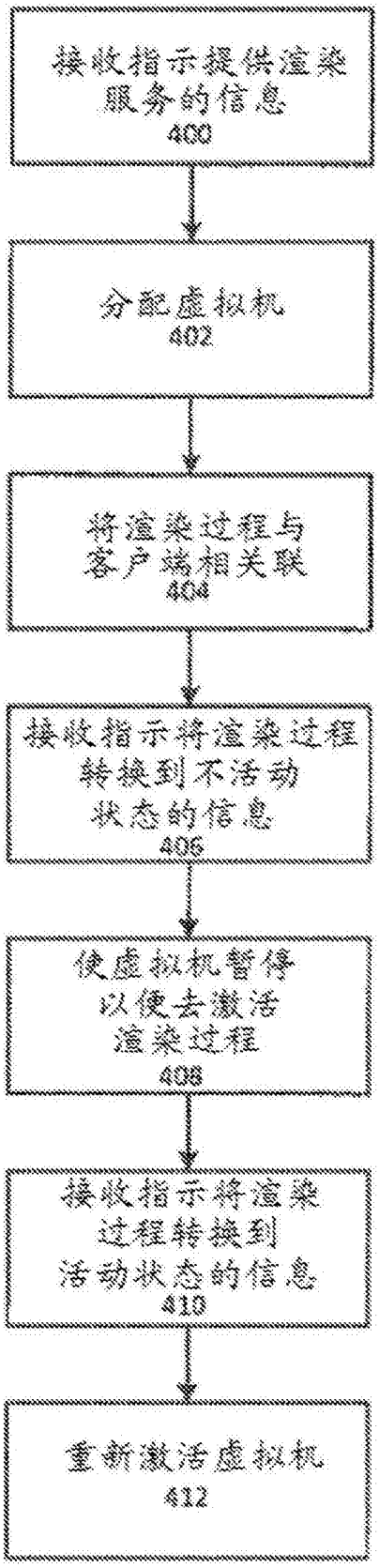

接收指示提供渲染服务的信息
400
分配虚拟机
402
将渲染过程与客户端相关联
404
接收指示将渲染过程转换到不活动状态的信息
406
使虚拟机暂停以便去激活渲染过程
408
接收指示将渲染过程转换到活动状态的信息
410
重新激活虚拟机
412

图4

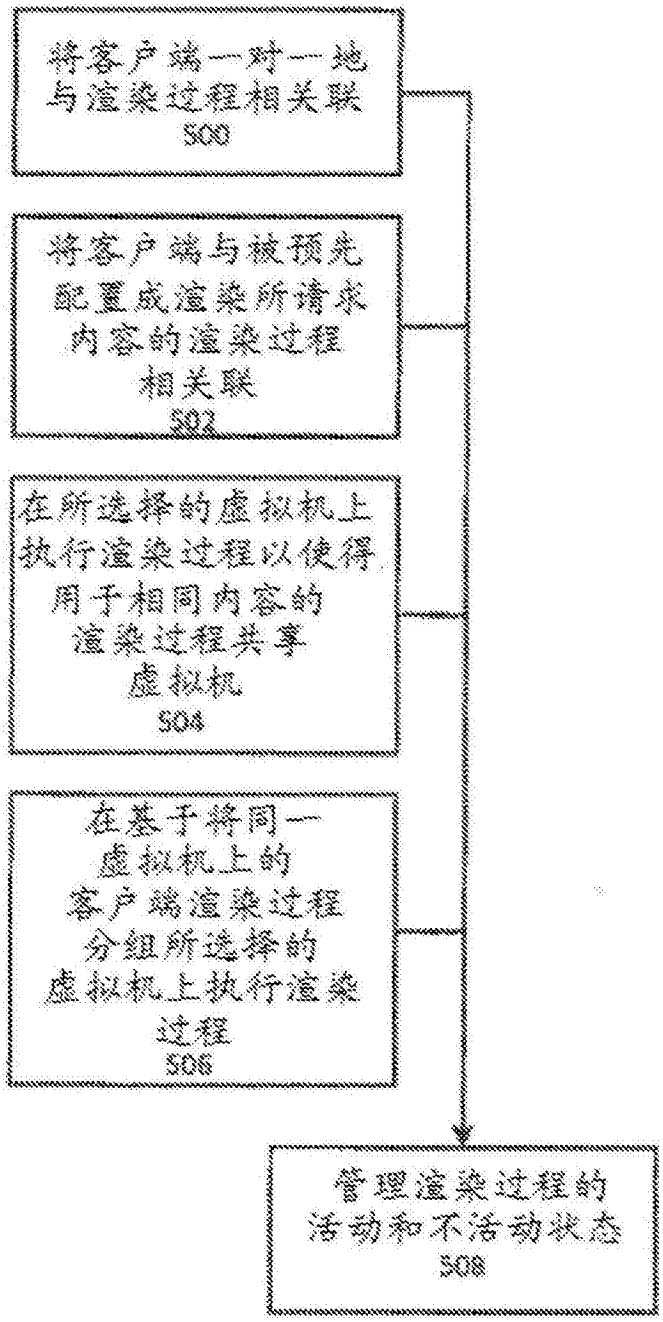

将客户端一对一地与渲染过程相关联
502
将客户端与被预先配置成渲染所请求内容的渲染过程相关联
504
在所选择的虚拟机上执行渲染过程以使得用于相同内容的渲染过程共享虚拟机
506
在基于将同一虚拟机上的客户端渲染过程分组所选择的虚拟机上执行渲染过程
508
管理渲染过程的活动和不活动状态

图5

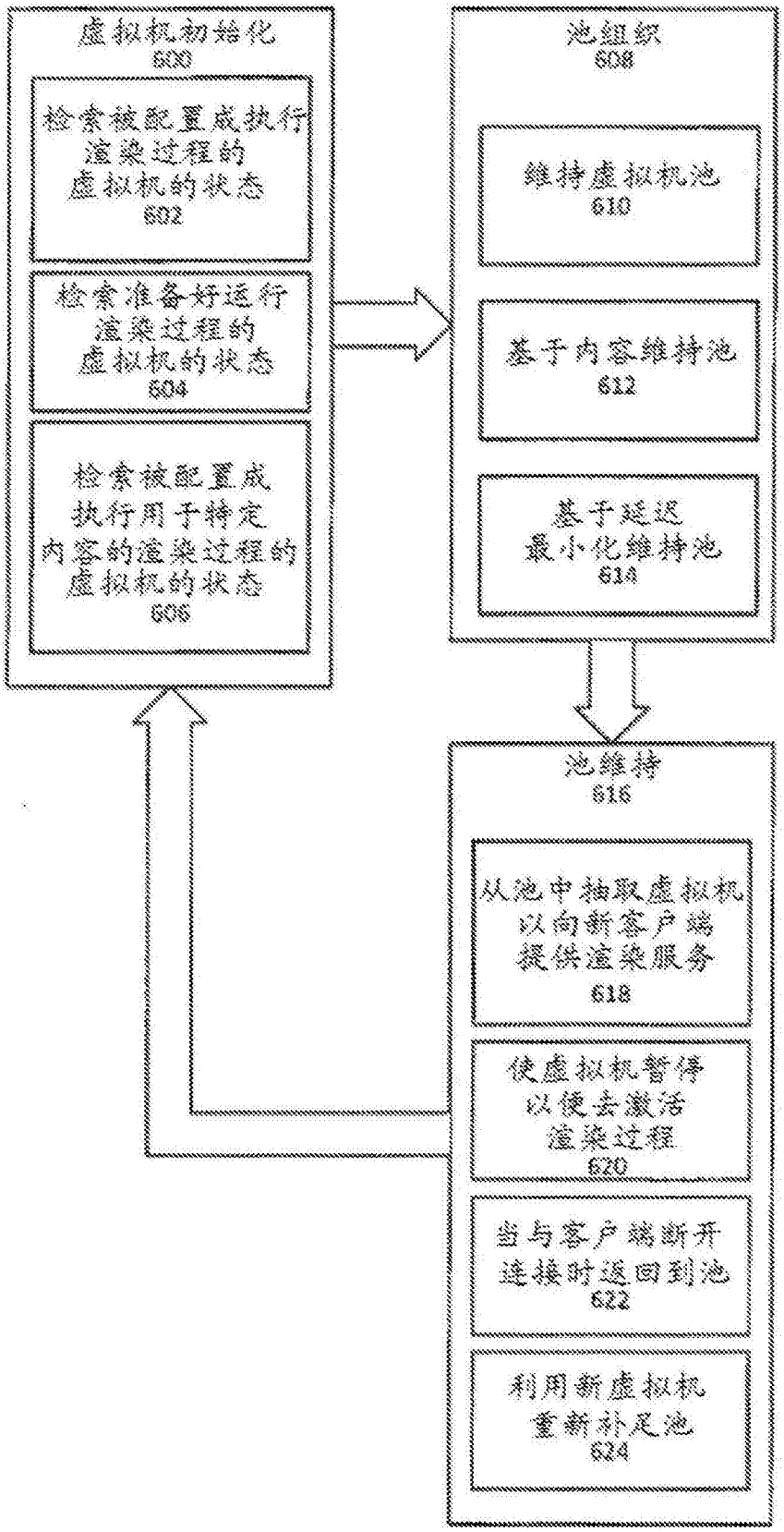

虚拟机初始化
600
检索被配置成执行渲染过程的虚拟机的状态
602
检索准备好运行渲染过程的虚拟机的状态
604
检索被配置成执行用于特定内容的渲染过程的虚拟机的状态
606
池组织
608
维持虚拟机池
610
基于内容维持池
612
基于延迟最小化维持池
614
池维持
616
从池中抽取虚拟机以向新客户端提供渲染服务
618
使虚拟机暂停以使去激活渲染过程
620
当与客户端断开连接时返回到池
622
利用新虚拟机重新补足池
624

图6

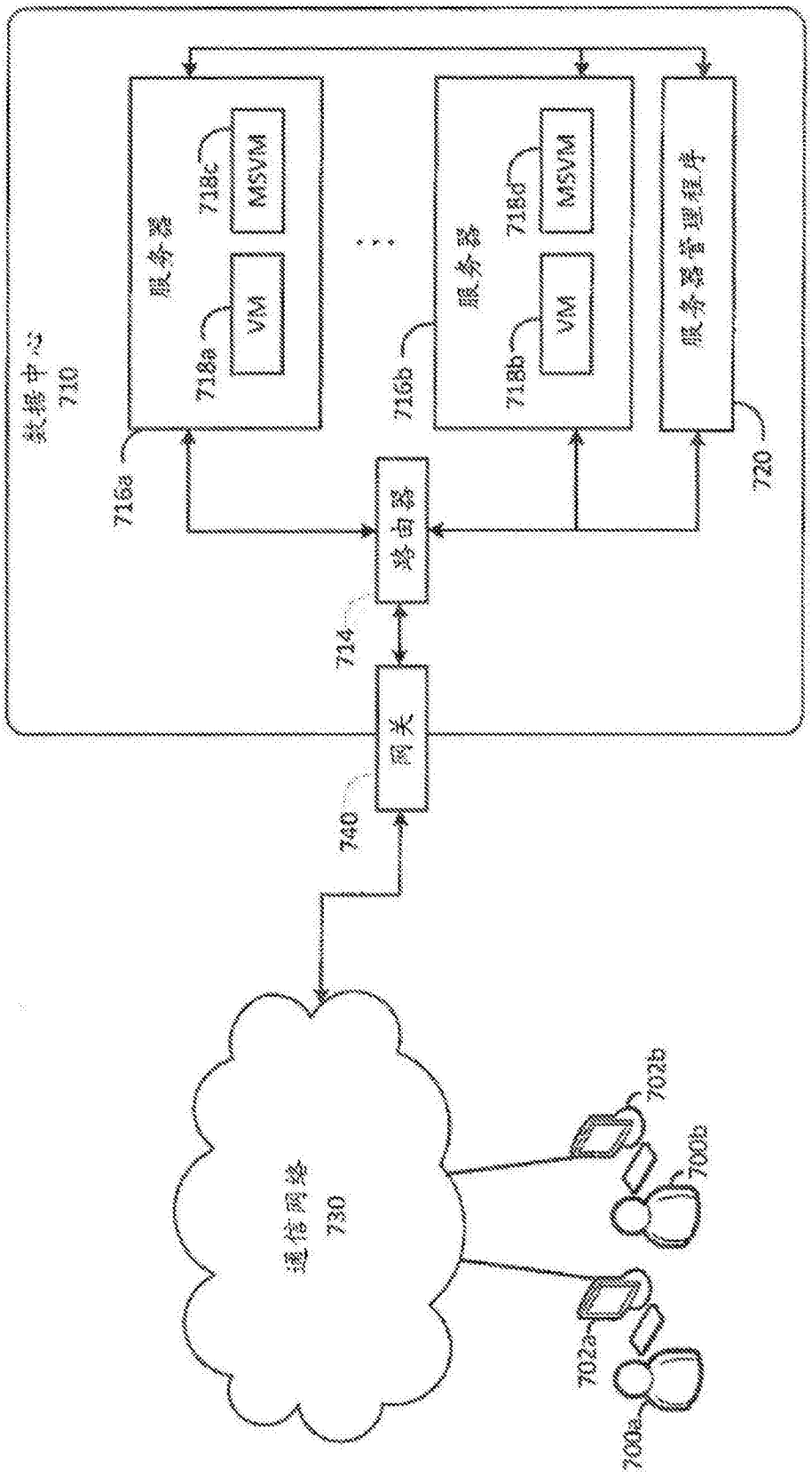

数据中心
710
服务器
716a
VM
718a
MSVM
718c
服务器
716b
VM
718b
MSVM
718d
服务器管理程序
720
路由器
714
网关
740
通信网络
730
702a
702b
702c

图7

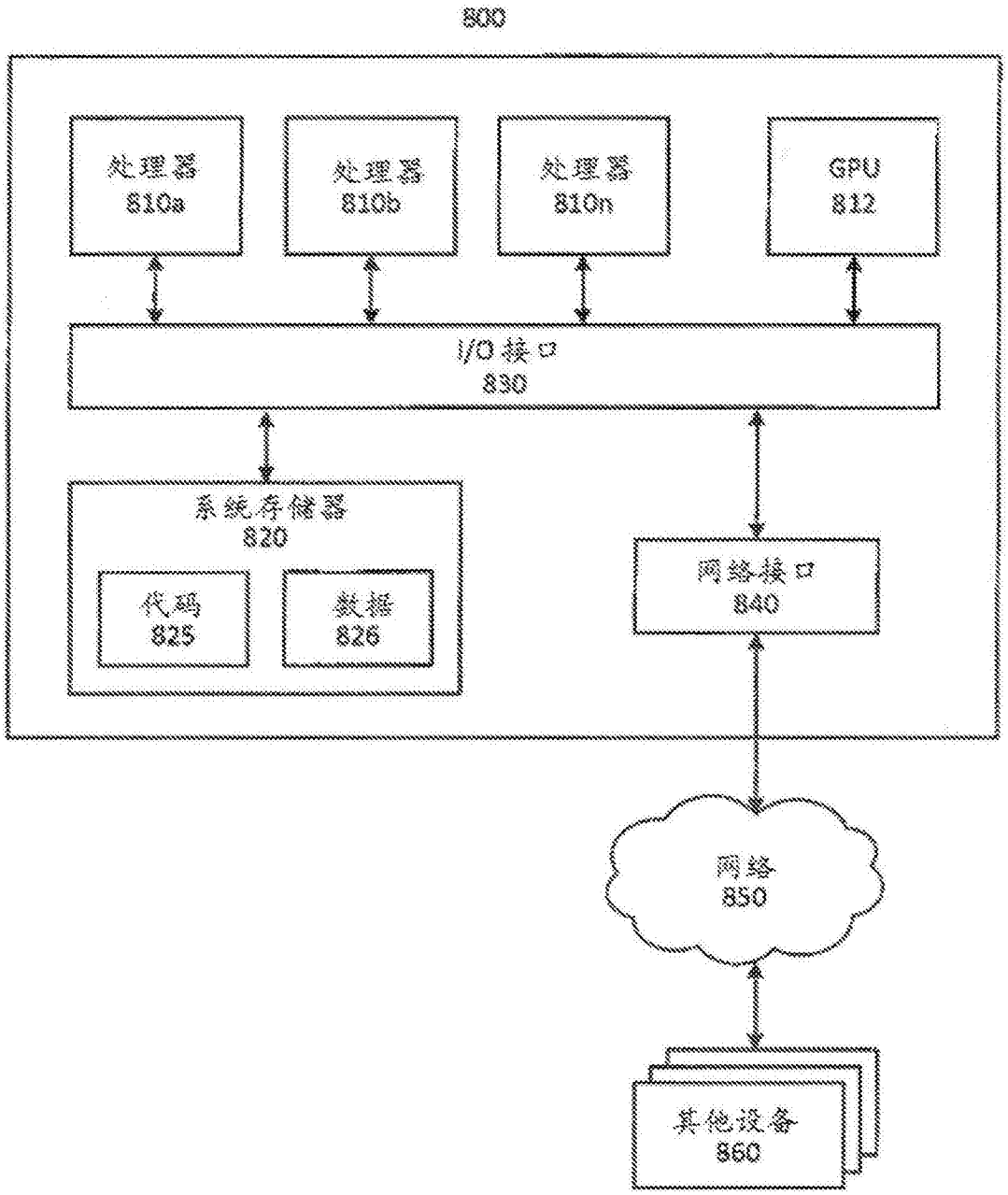

800
处理器
810a
处理器
810b
处理器
810n
GPU
812
I/O接口
830
系统存储器
820
代码
825
数据
826
网络接口
840
网络
850
其他设备
860

图8