(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 106055412 B
(45)授权公告日 2019.05.31

(21)申请号 201610670439.4

(22)申请日 2011.07.25

(65)同一申请的已公布的文献号
申请公布号 CN 106055412 A

(43)申请公布日 2016.10.26

(30)优先权数据
12/887,241 2010.09.21 US

(62)分案原申请数据
201180043637.3 2011.07.25

(73)专利权人 亚马逊技术有限公司
地址 美国华盛顿

(72)发明人 李·A·阿奇森 布莱恩·A·怀特
皮特·D·科恩
皮特·N·德桑蒂斯
麦克海尔·盖博

(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 穆童

(51)Int.Cl.
G06F 9/50(2006.01)

(56)对比文件
CN 1302014 A,2001.07.04,
CN 101006423 A,2007.07.25,
CN 1894666 A,2007.01.10,
US 2010228915 A1,2010.09.09,
WO 2007035544 A2,2007.03.29,
WO 2007120663 A3,2008.12.04,

审查员 严颖

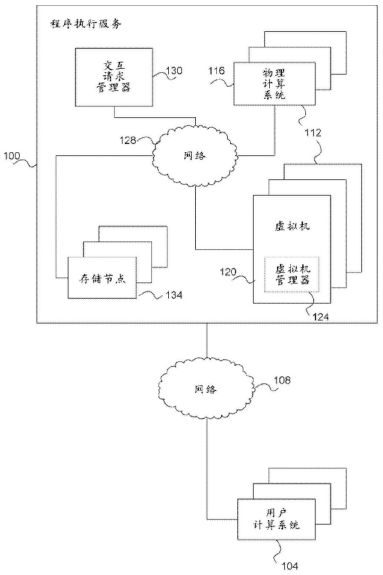
权利要求书2页 说明书16页 附图7页

(54)发明名称

用于动态管理计算容量请求的方法和系统

(57)摘要

描述了用于动态管理对来自计算资源提供者的计算容量的请求的系统和方法的实施方案。举例来说,所述计算资源可以包括程序执行能力、数据存储或管理能力、网络带宽等。所述系统或方法自动分配用于执行与用户相关的一个或多个程序的计算资源。所述系统和方法可以使所述用户能够在所述一个或多个程序的执行已开始之后改变所分配的资源。



CN 106055412 B

1.一种用于动态管理对由程序执行服务PES提供的计算容量的请求的计算机实施方法,所述方法包括:

在提供多个计算节点的PES的控制下,其中每个所述计算节点可被配置来执行所述PES的多个用户的一个或多个程序,所述PES还提供多个可用户自定义的程序服务,所述程序服务包括:(1)负载平衡器,其被配置来允许用户自定义跨所述多个计算节点中的多个计算节点的工作量分布,以及(2)负载调节器,其被配置来响应于工作量的变化而调节计算资源:

从第一用户接收与PES的用户共享针对第一用户生成的第一虚拟化环境的请求,所述第一虚拟化环境包括操作系统、用于在所述多个计算节点的子集上执行的至少一个用户程序、负载平衡器和负载调节器;

更新可用于由PES的用户选择的虚拟环境的列表,以提供包括与第一虚拟化环境相关的信息在内的虚拟化环境的更新列表;

向PES的用户提供虚拟化环境的更新列表;

从PES的第二用户接收对包括在更新列表中的第一虚拟化环境的选择;以及

针对第二用户生成第二虚拟化环境,所述第二虚拟化环境至少部分基于第一虚拟化环境。

2.根据权利要求1所述的方法,还包括:从第二用户接收要对第一虚拟化环境所做的至少一个更改,并至少部分基于所述至少一个更改来生成第二虚拟化环境。

3.根据权利要求2所述的方法,还包括:从第二用户接收与PES的用户共享针对第二用户生成的第二虚拟化环境的请求,并且进一步更新所述更新列表以包括与第二虚拟化环境相关的信息。

4.根据权利要求3所述的方法,还包括:从第三用户接收对第一虚拟化环境或第二虚拟化环境的选择。

5.根据权利要求1所述的方法,还包括:使用第二虚拟化环境来执行由第二用户提供的应用。

6.根据权利要求1所述的方法,还包括:向第二用户推荐包括在虚拟化环境的更新列表中的至少一个虚拟化环境。

7.根据权利要求6所述的方法,其中,所推荐的至少一个虚拟化环境至少部分基于第二用户提供的偏好。

8.根据权利要求1所述的方法,还包括:向第一用户提供费用,所述费用至少部分基于第二用户对第一虚拟化环境的选择。

9.根据权利要求1所述的方法,其中,所述PES还包括以下一个或多个:

监控接口,其被配置来允许用户监控所述至少一个用户程序的执行,或

多个数据库管理服务。

10.一种用于动态管理对由程序执行服务PES提供的计算容量的请求的设备,所述设备包括:

在提供多个计算节点的PES的控制下,其中每个所述计算节点可被配置来执行所述PES的多个用户的一个或多个程序,所述PES还提供多个可用户自定义的程序服务,所述程序服务包括:(1)负载平衡器,其被配置来允许用户自定义跨所述多个计算节点中的多个计算节点的工作量分布,以及(2)负载调节器,其被配置来响应于工作量的变化而调节计算资源:

用于从第一用户接收与PES的用户共享针对第一用户生成的第一虚拟化环境的请求的装置，所述第一虚拟化环境包括操作系统、用于在所述多个计算节点的子集上执行的至少一个用户程序、负载平衡器和负载调节器；

用于更新可用于由PES的用户选择的虚拟环境的列表，以提供包括与第一虚拟化环境相关的信息在内的虚拟化环境的更新列表的装置；

用于向PES的用户提供虚拟化环境的更新列表的装置；

用于从PES的第二用户接收对包括在更新列表中的第一虚拟化环境的选择的装置；以及

用于针对第二用户生成第二虚拟化环境的装置，所述第二虚拟化环境至少部分基于第一虚拟化环境。

11.根据权利要求10所述的设备，还包括：用于从第二用户接收要对第一虚拟化环境所做的至少一个更改的装置，以及用于至少部分基于所述至少一个更改来生成第二虚拟化环境的装置。

12.根据权利要求11所述的设备，还包括：用于从第二用户接收与PES的用户共享针对第二用户生成的第二虚拟化环境的请求的装置，以及用于进一步更新所述更新列表以包括与第二虚拟化环境相关的信息的装置。

13.根据权利要求12所述的设备，还包括：用于从第三用户接收对第一虚拟化环境或第二虚拟化环境的选择的装置。

14.根据权利要求10所述的设备，还包括：用于使用第二虚拟化环境来执行由第二用户提供的应用的装置。

15.根据权利要求10所述的设备，还包括：用于向第二用户推荐包括在虚拟化环境的更新列表中的至少一个虚拟化环境的装置。

16.根据权利要求15所述的设备，其中，所推荐的至少一个虚拟化环境至少部分基于第二用户提供的偏好。

17.根据权利要求10所述的设备，还包括：用于向第一用户提供费用的装置，所述费用至少部分基于第二用户对第一虚拟化环境的选择。

18.根据权利要求10所述的设备，其中，所述PES还包括以下一个或多个：

监控接口，其被配置来允许用户监控所述至少一个用户程序的执行，或

多个数据库管理服务。

用于动态管理计算容量请求的方法和系统

[0001] 分案说明

[0002] 本申请是申请日为2011年7月25日，申请号为201180043637.3，题为“用于动态管理计算容量请求的方法和系统”的中国专利申请的分案申请。

[0003] 相关申请案

[0004] 本申请要求2010年9月21日提交的美国非临时申请第12/887,241号的权益，该申请的公开内容的全文据此以引用的方式并入本文。

背景技术

[0005] 公司和机构运行使众多计算系统互连以支持其运营的计算机网络。计算系统可位于单个地理位置中(例如，作为局域网的一部分)或位于多个不同地理位置中(例如，经由一个或多个专用或公用的中间网络)。数据中心可以置放大量互连的计算系统，例如专用数据中心是由单个机构进行操作并且公用数据中心是由第三方进行操作以把计算资源提供给客户。专用和公用数据中心可以对数据中心、机构或其它客户拥有的硬件提供网络访问、电力、硬件资源(例如，计算和存储)和安全安装设施。

[0006] 为了帮助提高数据中心资源的利用率，虚拟化技术可以允许单个物理计算机主控作为到连接的计算机用户的独立计算机出现而操作作的虚拟机的一个或多个实例。运用虚拟化，单个物理计算装置可以动态方式创建、维持或删除虚拟机。用户又可基于“按照需要”或至少基于“按照请求”请求来自数据中心的计算机资源并且具备不同量的虚拟机资源。

[0007] 随着数据中心的规模和范围增大，提供、支配和管理数据中心的物理和虚拟计算资源已变得越来越复杂。

附图说明

[0008] 在附图各处，参考数字可以重用于指示参考元件之间的对应性。提供附图以示出本文描述的示例性实施方案且并非旨在限制本公开内容的范畴。

[0009] 图1是示意地示出了可经由通信网络把计算资源提供给多个用户计算系统的程序执行服务的示例的网络图；

[0010] 图2A是被配置来管理程序执行服务的用户要用的计算资源的请求的交互请求管理器的阐释性组件的方框图；

[0011] 图2B示意地示出了程序执行服务的用户计算系统与交互请求管理器之间的示例性交互的网络图；和

[0012] 图3A和图3B是示出了由交互请求管理器组件实施的交互请求管理器例程的流程图。

[0013] 图3C是示意地示出了交互请求管理器的实施方案可通过其与用户计算系统进行通信以用于修改分配的计算资源的设置的例程的示例的流程图。

[0014] 图4是示意地示出了交互请求管理器的实施方案可通过其与用户计算系统进行通信以提供供用户选择的多个虚拟化环境的例程的示例的流程图。

具体实施方式

[0015] 描述用于动态管理对来自计算资源提供者(程序执行服务)的计算容量的请求的系统和方法的实施方案。举例来说,计算资源可以包括程序执行能力、数据存储或管理能力、数据库管理能力、网络带宽、应用监控或日志记录、用于采取纠正措施以解决问题的能力等。在某些实施方式中,用户可请求生成可在当前或未来使用时段期间管理用户的计算机资源的虚拟化环境。例如,用户可请求生成可在使用时段期间运行用户的自定义软件应用和为用户管理或预约合适的程序执行容量、数据存储容量、数据库管理选项和/或网络带宽的虚拟化环境。计算资源提供者可确定提供者的计算机资源中的哪些可用于满足用户的请求并且可在请求的使用时段期间把这些计算机资源分配给用户。

[0016] 可以高度灵活地选择用户请求的使用时段和/或其它参数以满足用户的计算机资源需要。用户请求可以包括用于指定用户的偏好、限制和/或需求的一个或多个用户可选择参数。例如,用户请求可指定在使用时段期间执行某个特定程序(或多个特定程序)、在使用时段期间使用特定类型或地理分布的计算机资源、使用时段具有希望的开始日期、结束日期和/或持续时间等等。在某些实施方式中,计算资源提供者对可由用户提交的请求参数的范围施加很少的限制或没有限制。

[0017] 作为一个可能的示例,用户可能够请求在可以包括在一个或多个地理位置的计算机资源的一组计算机资源上执行特定程序。用户可能够使用应用编程接口(API)或其它类型的计算接口把程序和程序执行参数传达到程序执行服务,用于生成虚拟化环境。例如,用户可以使用Web应用档案文件(如Java WAR文件)上传软件应用。接着,程序执行服务可以自动配置虚拟化环境(例如,“应用容器”),其可是包括用于用户程序的应用软件堆栈以及用于在程序执行服务上执行用户程序的一个或多个基础结构服务的运行时环境。应用容器可包括用户可选择的操作系统(例如,Linux、Windows等)、应用服务器(例如,Apache Tomcat)、系统或应用配置等。虚拟化环境可被配置来寄宿在特定URL处。基础结构服务可包括但不限于:负载平衡器,其用于跨请求的计算资源分布工作量;负载调节器,其响应于负载或需求变化而调节计算资源;监控接口,其允许用户监控程序、数据存储资源(例如,可伸缩容量块存储装置)等的执行。在某些实施方案中,用户可能够选择可包括在容器中的一个或多个程序或服务。例如,用户可能够从多个数据库模型(例如,关系数据库、SQL数据库、Oracle数据库等)中进行选择。在某些实施方案中,基础结构服务可被自定义用于用户而非作为多个用户之间共享的资源。例如,在某些这类实施方案中,负载平衡器可被单独自定义用于用户应用而非在程序执行服务的多个用户之间进行共享或分布。

[0018] 虚拟化环境的特定实施方案的可能优点是计算系统允许用户在希望的情况下具有灵活度和对应用容器的内容的控制。例如,在某些情况下,用户可以仅提供用户程序,并且计算系统可以自动管理由虚拟化环境使用的所有剩余基础结构的部署。在其它情况下,用户可以选择和/或配置包括在虚拟化环境中的一个或多个基础结构服务。在某些情况下,用户可以选择在不同地理区域中的计算资源,以便实现用于用户应用的希望的部署拓扑。可以以任何希望的方式配置或选择部署拓扑。例如,部署拓扑可以基于用户的客户的位置或区域,以便改善应用的性能(例如,减小网络延时)。作为另一示例,部署拓扑可以被配置来使用一个或多个区域或地带中的计算资源以例如通过改善应用对特定区域或地带中的计算资源的故障(例如,由于一个区域或地带中的不利天气状况)的复原能力来改善用户应

用的稳健性。此外，在特定实施方案中，用户在执行程序期间保持对计算资源的访问，并且用户可以在执行期间控制某些或所有计算资源。例如，在某些这类实施方案中，用户在希望的情况下可以选择使用户应用脱离虚拟化环境。

[0019] 在特定收费实施方式中，计算机资源提供者可以向用户收取针对请求的预约费(例如，当准许该请求时)和/或对于在使用时段期间提供可用计算机资源的使用的使用费。各种类型或级别的费用安排是有可能的。例如，可以由用户请求用于直接用户的计算机资源(“按需资源”)。在某些此类情况下，用户可能不会支付预约费但可能支付更高使用费。作为另一示例，用户可能会预约计算机资源以便在未来使用时段期间获得保证的可用性(“预约的资源”)。可以向用户收取进行预约的预约费并且还可基于在使用时段期间实际使用的计算机资源量向用户收取使用费。在某些这类情况下，预约资源的使用费可以从按需资源的使用费打折扣并且/或者可以在更接近使用时段之时而非更接近进行请求之时收取预约费。在另一示例中，计算机资源提供者可以允许用户对未使用的计算机资源(“现货资源”)进行竞价。在某些这类情况下，计算机资源提供者可以设置基于资源供应和需求而变化的现货价格，并且可使资源可用于其竞价达到或超过现货价格的那些用户。

[0020] 现将参考旨在阐释而非限制本公开内容的特定实施例和实施方案描述本公开内容的各个方面。

[0021] 图1是示意地示出了可经由通信网络108把计算资源提供给多个用户计算系统104的程序执行服务100的示例的网络图。例如，程序执行服务100可管理来自用户的请求以代表用户执行程序或程序集。至少一些用户计算系统104可在远离程序执行服务100之处。在此示例中，用户可使用计算系统104以通过通信网络108访问程序执行服务100。网络108可以例如是可能由各个不同方运营的链接网络的可公开访问的网络，如互联网。在其它实施方案中，网络108可以是专用网，例如非特权用户完全或部分无法访问的企业或大学网络。在其它实施方案中，网络108可以包括可访问互联网和/或从互联网访问的一个或多个专用网。

[0022] 程序执行服务100提供用于管理多个用户的程序的执行的多种功能。在图1所示的示例中，程序执行服务100包括可代表用户执行程序的多个计算节点112。计算节点112可以包括一个或多个物理计算系统116和/或寄宿在一个或多个物理计算系统上的一个或多个虚拟机120。例如，主机计算系统可以提供多个虚拟机120并且包括用于管理这些虚拟机的虚拟机(“VM”)管理器124(例如，系统管理程序或其它虚拟机监控器)。

[0023] 在图1所示的示例中，每个计算节点112具有可用于执行一个或多个程序的一定量的计算资源。每个计算节点112可以被配置来提供可以例如按处理容量(例如，处理单元的数量和/或大小)、存储器容量、存储容量、网络带宽容量、非网络通信带宽等中的一个或多个的组合来测量的特定量的程序执行容量。在某些实施方案中，程序执行服务100可以提供预先配置的计算节点112，其中每个预先配置的计算节点具有可用于代表用户执行程序的类似和/或等量的资源。在其它实施方案中，程序执行服务100可以提供对各个不同计算节点112的选择，用户可以从这些节点中选择节点来代表用户执行程序。在其它实施方案中，程序执行服务100可以生成特定于用户和用户程序的执行的各种计算节点。在某些这类实施方案中，计算节点112可以具有不同量和/或类型的计算资源(例如，处理单元的大小、速度和/或类型；处理单元的数量；存储器和/或存储装置的量；平台配置，如32位或64位操作

系统等)。

[0024] 程序执行服务100可以提供可访问提供数据、程序和其他用户信息的大容量存储的存储节点134的用户计算系统104。存储节点134可以包括任何类型的持久性数据存储装置,例如非易失性存储装置,如硬盘驱动器、光盘驱动器等。在图1所示的示例中,计算节点112可经由网络128访问存储节点134。网络128可以包括多个联网装置(未示出),如交换机、边缘路由器、核心路由器等。网络128可以但不必须是与图1所示的网络108不同的网络。

[0025] 程序执行服务100的用户可经由交互请求管理器130与程序执行服务100进行交互以请求程序执行服务的优选和/或所要资源(例如,程序执行容量和/或存储资源)。交互请求管理器130可经由网络128连接到计算节点112和存储节点134。交互请求管理器130可通过网络108从用户计算系统104接收资源请求。用户可以经由交互请求管理器130请求服务100提供用于代表用户(或由该用户授权的其他用户)执行程序的一个或多个计算节点。在某些实施方案中,用户可以经由交互请求管理器130请求服务100生成可管理和预约可能被需要用于代表用户执行程序的计算资源的一个或多个计算节点。在不同实施方案中,可以在请求代表用户执行程序之时和/或在一个或多个其它时间(如当用户进行注册和/或预订程序执行服务100的使用服务时)指定计算资源。在某些实施方案中,交互请求管理器130可以对一个或多个用户提供预订和/或注册服务,使得用户可以指定与代表用户执行的一个或多个程序有关的信息(例如,程序、源代码、一个或多个程序的可寻址位置等)、帐户信息(例如,用户名、记账信息等)、使用条款等。在某些实施方案中,在用户与交互请求管理器130进行交互以为了服务预订和/或注册之后,可以向用户发出与该用户相关并且要结合代表该用户执行程序所使用的一个或多个请求识别符(例如,密钥、令牌、用户名、密码等)。在其它实施方案中,可以提供除交互请求管理器130外的其它模块以执行与程序执行服务100的预订和/或注册服务有关的各种操作。

[0026] 在某些实施方案中,由一或多个物理或虚拟计算系统执行或具体实施交互请求管理器130。例如,在某些实施方案中,具有包括CPU、I/O组件、存储装置和存储器的组件的服务器计算系统可以用来执行交互请求管理器130。I/O组件包括显示器、到网络128的网络连接、计算机可读介质驱动器和其它I/O装置(例如,键盘、鼠标、扬声器等)。交互请求管理器130的实施方案可以作为一个或多个可执行程序模块存储在服务器的存储器中,并且交互请求管理器130可通过网络128与计算节点112(例如,物理计算系统116和/或VM 120)进行交互。交互请求管理器130可经由网络108从用户接收对程序执行服务100的计算资源的请求。

[0027] 图2A是被配置用于管理对代表用户执行程序的请求的交互请求管理器130的实施方案的说明性组件的方框示意图。在这个实施方案中,交互请求管理器包括资源生成模块204、资源调度模块208、监控和报告模块212以及记账模块216。

[0028] 资源生成模块204从用户接收对程序执行服务100的计算资源的请求,例如生成一个或多个计算节点以管理和预约可用于在使用时段期间执行用户程序的计算资源的请求。用户可以请求程序执行计算节点立即可用、可以请求在未来时间生成程序执行计算节点或可以请求基于其它条件生成程序执行计算节点。可以由资源生成模块204以各种方式接收对程序执行计算节点的请求。例如,可直接从用户(例如,经由由程序执行服务提供的交互控制台或其它GUI)、从自动开始执行其它程序或其自身的其它实例的用户的执行程序、从

编程工具(例如,命令行工具、集成开发环境(例如,Eclipse)等)、从经由由程序执行服务提供的应用编程接口(“API”)(例如,使用Web服务的API)与交互请求管理器进行交互的程序等等接收请求。

[0029] 对计算节点的请求可以包括计算节点的数量和/或类型、要使用的计算节点的最小和/或最大数量、在其期间保证计算节点的可用性的使用时段、请求的到期时间等。请求可以指定在使用时段期间仅在生成的计算节点上执行特定程序。对计算节点的请求可以包括其它类型的偏好、需求和/或限制(例如,存储容量或网络带宽的量、节点的地理和/或逻辑位置、终止条件等)。

[0030] 在某些实施方案中,请求包括用户程序,并且资源生成模块204自动提供基础结构资源使得可由程序执行服务100执行用户程序。例如,资源生成模块204可以自动配置虚拟化环境,其可是包括用于用户程序的应用软件堆栈以及用于执行用户程序的一个或多个基础结构服务的运行时环境。虚拟化环境可包括用户可选择的操作系统(例如,Linux、Windows等)、应用服务器(例如,Apache Tomcat)和系统或应用配置等。虚拟化环境可被配置来寄宿在特定URL处。基础结构服务可包括但不限于:负载平衡器,其用于跨请求的计算资源分布工作量(如应用流量);负载调节器,其响应于负载或需求变化而调节计算资源;监控接口,其允许用户监控程序、数据存储资源(例如,可伸缩容量块存储装置)等的执行。在某些实施方案中,用户可能够选择由程序执行服务100提供且可包括在虚拟化环境中的一个或多个程序或服务。例如,用户可能够从多个数据库模型(例如,关系数据库、SQL数据库、Oracle数据库等)中进行选择。在某些实施方案中,基础结构服务可被自定义用于用户而非作为程序执行服务100的多个用户之间共享的资源。例如,在某些这类实施方案中,负载平衡器可被单独自定义用于用户应用而非在程序执行服务100的多个用户之间被共享或分布。在某些实施方案中,虚拟化环境可被配置来允许用户应用产生进程线程或打开用于与其它应用进行通信的任何合适端口(例如,除端口80外的端口)。例如,虚拟化环境可以支持超文本传输协议安全端口(HTTPS)。

[0031] 交互请求管理器130的特定实施方案的可能优点是用户可以仅把自定义应用上传到程序执行服务100,且接着交互请求管理器130可生成包括用来在程序执行服务100上执行用户自定义应用的资源和应用堆栈的虚拟化环境。使得能够执行用户程序所需的与交互请求管理器130的用户交互量范围可以从相对少的交互(例如,与使用Java WAR文件上传程序差不多)到相对高度的交互(例如,用户可以自定义大致上虚拟化环境的所有方面)。因此,交互请求管理器130的实施方案可以提供虚拟化环境的相对高度的灵活性和可自定义能力。在某些实施方式中,程序执行服务100(或其它提供者)可以使一个或多个标准或默认虚拟化环境可供服务100的用户使用。

[0032] 对程序执行容量的请求可以指定使用时段,在该使用时段期间可使计算资源可用于用户。在某些实施方式中,程序执行服务100可以对用户提供在使用时段期间请求的计算资源将可用的保证。在不同实施方案中,可以以不同方式指定使用时段。例如,使用时段可以指示在开始时间开始并且在到期时间结束的指定持续时间(例如,小时、天、周、月、年等数目)。开始时间和/或到期时间可以包括一天中的时间(例如,早上7:00)和日期(例如,2010年1月23日)。开始时间可在某个未来时间,例如未来的一小时或数小时、一天或数天、一周或数周或者一年或数年。在某些情况下,未来使用时段的开始时间可以比请求(或请求

的确认)时间至少晚特定时段,例如未来的至少一小时、一天、一周、一月或更长时间。

[0033] 在交互请求管理器130的某些实施方式中,延迟时段可以发生在由交互请求管理器130接收到对计算节点的请求的时间与准许该请求的时间或将确认提供给用户的时间之间。例如,延迟时段可能是因由交互请求管理器130或程序执行服务100执行的各种处理操作、管理操作、会计操作等而发生。在某些这类实施方式中,请求的使用时段指的是在这些延迟时段被考虑之后(或大致在其之后)的时段。例如,在特定实施方式中,延迟时段可以是数秒、数分钟或几小时。在特定的此类实施方式中,请求的未来使用时段的开始时间可以是超过这样的延迟时段的未来时间。在交互请求管理器130的特定其它实施方式中,开始时间可以是由程序执行服务100提交、接收或准许用户请求的时间。

[0034] 在某些情况下,请求可以指示使用时段直到被用户明确终止才到期(例如,可以不设置到期时间)。持续时间可以在从一小时到一周、一周到一月、一月或数月、一年或数年或某个其它持续时间的范围中。在某些实施方案中,使用时段可以包括上述(或其它)因素的组合以便对用户提供调度计算机资源中的高度灵活性。

[0035] 在某些情况下,在程序服务100生成满足用户请求的计算节点之后,用户可对程序或一个或多个计算节点的设置进行一个或多个更改。例如,用户可以改变与一个或多个计算节点相关的存储装置或网络带宽的量或类型、可以改变与一个或多个计算节点相关的使用时段或终止条件、可以终止程序的执行等。用户可以各种方式请求一个或多个更改,如本文所述。例如,用户可经由GUI、命令行工具、集成开发环境(例如,Eclipse)、API调用等请求一个或多个更改。

[0036] 在由资源生成模块204接收到对计算节点的请求之后,资源调度模块208可调度和分配计算节点以满足请求。例如,在接收到对程序执行容量的请求之后,资源调度模块208可以确定用于程序执行的一个或多个计算节点112。在某些实施方案中,即使请求是针对未来的可用性,也在请求之时执行对要使用的计算节点112的确定。在其它实施方案中,计算节点的确定推迟到后来某个时间(例如,在使用时段开始之前),使得确定可基于那时可用的信息。

[0037] 资源调度模块208可以分配来自计算节点112的一个或多个计算节点以供用户在请求的使用时段期间利用。在某些实施方案中,分配一个或多个特定计算节点112(例如,一个或多个特定物理计算节点116和/或虚拟计算节点120)以供用户(或授权用户)在整个使用时段期间优先使用。

[0038] 在其它实施方案中,资源调度模块208可以分配来自计算节点池的计算节点,而非把特定计算节点分配给特定用户用于使用时段。计算节点池可以包括有足够资源用于满足用户或授权用户的程序执行请求的适量计算节点。在某些这类实施方案中,在使用时段期间接收到执行一个或多个程序的请求之后,可以从计算节点池中选择足以执行该一个或多个程序的适量计算节点,并且在选定节点上开始程序执行。在选定量的计算节点不再用于执行请求之后(例如,在请求的执行终止和/或完成之后),可将这些计算节点返回到计算节点池以供用户或其它授权用户在使用时段期间使用。在某些实施方式中,分配计算节点池的节点以供用户(或授权用户)专用、排他使用或优先使用。在某些这类实施方式中,用户(或授权用户)未在使用的计算节点池的节点可以分配给其他用户用于程序执行,并且如果用户(或授权用户)需要这些节点来满足请求的容量,那么可以终止其他用户的程序。

[0039] 在使用时段期间，用户(或授权用户)可以把对在分配的计算节点上执行一个或多个程序的请求提交到交互请求管理器130。程序执行请求可以包括用于开始一个或多个程序的执行的各种信息，如要执行的程序的可执行或其它副本、先前为执行而注册或以其它方式供应的程序的指示以及要同时执行的程序实例的数量(例如，表示成实例的单个希望数量、希望实例的最小和最大数量等)。请求可以指定用于执行程序的计算节点的数量和/或类型、要使用的计算节点的最小和/或最大数量、请求的到期时间、执行的优选执行时间和/或时段等。请求可以包括对于执行一个或多个程序的其它类型的偏好和/或需求(例如，资源分配、执行的地理和/或逻辑位置、执行与其它程序和/或计算节点的接近度、时间相关的条件、终止条件等)。

[0040] 资源调度模块208可以以不同方式(包括基于请求中指定或以其它方式指定的针对程序和/或相关用户的任何偏好、限制和/或需求)确定哪些分配的计算节点要用于执行每个程序实例。例如，如果确定了用于执行程序实例的优选和/或所要资源(例如，存储器和/或存储装置；CPU类型、周期或其它性能度量；网络容量；平台类型等)的条件，那么用于执行程序实例的适当计算节点的确定可以至少部分基于计算节点是否具有可用于满足这些资源条件的足够资源。

[0041] 在使用时段期间，由资源生成模块204接收到的代表用户或授权用户在分配的计算节点上执行程序的请求可以导致在分配的计算节点中的一个或多个上开始程序执行。在某些情况下，可以在使用时段期间接收足够的对程序执行的请求使得所有分配的计算节点是在使用中(例如，执行程序)。在使用时段期间接收到的对程序执行的另外的请求可以被拒绝，或者可以被资源调度模块208保留或排入队列直到一个或多个节点变得可用为止。

[0042] 在某些实施方案中，资源调度模块208可以执行关于满足请求的一个或多个管理操作，例如强制执行与请求相关的使用时段或其它限制、释放计算资源来满足请求、授权和/或认证请求和/或请求用户等。例如，在某些情况下，来自用户的请求可以指定在使用时段期间仅特定用户被授权可以访问分配的计算节点。在某些情况下，来自用户的请求可以指定在使用时段期间仅在分配的节点上执行一个或多个指定程序。其它限制可包括对程序执行持续时间的限制、对在程序执行期间发生的费用的限制等。一个或多个上述限制(或其它限制)的组合可以由用户指定并由交互请求管理器130在允许访问分配的计算节点之前检查。

[0043] 在某些实施方式中，在使用时段到期之后，资源调度模块208释放分配的计算节点(例如，专用计算节点或计算节点池中的节点)以供其他用户使用。在某些这类实施方式中，终止在使用时段到期时执行的程序。在其它实施方式中，不终止此类执行的程序并且允许其继续执行直到更高优先级的用户请求访问计算节点为止。

[0044] 在图2A所示的实施方案中，监控和报告模块212在使用时段期间监控和追踪分配的计算节点的使用并且向用户报告有关使用的信息和统计数据。例如，监控和报告模块212可以追踪在分配的计算节点上执行程序的用户的使用模式。使用模式可包括访问节点的用户的数量或身份、程序执行的开始/结束时间和持续时间和/或其他用户指定模式或诊断。在某些此类实施方案中，监控和报告模块212可以把交互反馈(包括例如何时程序可能在计算节点上执行和/或要执行多久、实际或预测的节点需求等的指示)提供给用户。在某些实施方案中，监控和报告模块212可生成详述或汇总使用统计数据的报表并经由电子邮件把

该报表传达给用户或提供经由Web服务的对报表、使用统计数据或交互反馈的访问。

[0045] 某些程序执行服务100可收费，使得服务代表用户执行程序或分配计算资源来获取由用户支付的一笔或多笔费用。在某些收费服务中，交互请求管理器130可以可选地包括图2A中示意地示出的记账模块216。例如，在某些实施方案中，可以基于被分配用于代表用户执行一个或多个程序的程序执行容量的量和/或类型(如基于被分配用于执行用户的程序的处理单元的数量、存储器的量、存储装置的量、网络资源的量等中的一项或多项)对用户收取费用。在某些实施方案中，费用可以基于其它因素，如用来执行程序的计算资源的各种特征，例如基于CPU能力或性能、平台类型(例如，32位、64位等)等。在某些实施方案中，可以基于多种使用因素(如每次使用服务的价格、使用计算服务的单位时间的价格、每个使用的存储装置的价格、每个传入和/或传出的数据的价格等)收取费用。

[0046] 费用可以基于诸如与程序执行容量请求有关的各种其它因素和/或与执行程序有关的各种特性(例如，执行连续性、容错等)。在至少某些实施方案中，程序执行服务可以提供用于代表多个用户执行程序的各种服务或功能级别、类型和/或等级中的一个或多个，且在某些这类实施方案中，各种费用可以与各种服务级别、类型和/或等级相关。记账模块216可监控和追踪计算机资源的使用并计算应收取的使用费。

[0047] 可以对用户收取用于预约计算容量的固定费用支付(例如，预付或周期性地开账单)，且在某些情况下对用户收取其它使用费(例如，与各种资源(如电、物理机柜空间、网络利用等)的使用相关的可变费用)。例如，当提出在使用时段期间利用计算资源的请求时或当程序执行服务100准许该请求时，可以对提出请求的用户收取预约费。预约费可以基于例如请求的资源量、使用时段的开始时间和/或持续时间、服务是否将被需要来购买额外计算硬件来满足请求等。例如，如果开始时间是在不久的将来，那么预约费可以比开始时间是在更远的将来的情况下更高。此外，可以对用户(或授权用户)收取针对在使用时段期间利用资源的使用费。例如，可以基于例如程序执行的持续时间、用来执行程序的资源的类型等对请求在使用时段期间在分配的计算节点上执行程序的授权用户收取使用费。如上文论述，各种类型或级别的费用安排是有可能的。例如，可以不对请求用于立即使用的按需资源的用户收取预约费，但是可以对其收取比对支付预约费以预约用于未来使用时段的资源的用户收取的使用费更高的使用费。

[0048] 记账模块216可以追踪使用、计算适当费用和开账单给用户和/或授权用户(或把记账信息提供给会计模块或服务)。在某些情况下，用户请求可以指示把授权用户产生的某些或所有使用费开账单给用户而非授权用户。在某些此类情况下，记账模块216可以适当地在用户和授权用户之间进行费用分配。

[0049] 可与图2A所示不同地配置交互请求管理器130。例如，可组合、重新安排、添加或删除由所示模块提供的各种功能。在某些实施方案中，额外或不同的处理器或模块可以执行参考图2A中所示的示例性实施方案描述的某些或所有功能。许多实施方式变化是有可能的。

[0050] 虽然按照程序执行容量的管理进行了大致描述，但是在其它实施方案中，交互请求管理器130可被配置来管理供多个用户使用的额外或替代类型的计算相关资源和对这些计算相关资源的可用性提供灵活的保证。这些资源可以包括一个或多个以下项：持久性数据存储能力(例如，在非易失性存储器装置如硬盘驱动器上)；临时数据存储能力(例如，在

易失性存储器如RAM上)；消息排队和/或传递能力；其它类型的通信能力(例如，网络套接字、虚拟通信电路等)；数据库管理能力；专用带宽或其它网络相关资源；非网络带宽；输入装置能力；输出装置能力；CPU周期或其它指令执行能力；等。

[0051] 图2B是示意地示出了用户计算系统104a与程序执行服务100的交互请求管理器130之间的示例性交互的网络图。程序执行服务100可把计算资源提供给多个用户计算系统104a、104b、...、104n。在这个说明性示例中，程序执行服务100对用户计算系统104a、104b、...、104n提供API以用编程方式与交互请求管理器130进行交互。图2B说明性地示出了用户计算系统104a使用请求API传达对在程序执行服务100的计算资源上执行程序的请求。请求API(1)是经由网络108传达并且(2)是由程序执行服务100的交互请求管理器130接收。请求API可包括关于用户的请求的信息，例如用户的程序(例如，要执行的程序的可执行或其它副本或先前为执行而注册或以其它方式供应的程序的指示等)、计算节点的数量和/或类型、要使用的计算节点的最小和/或最大数量、请求在其期间计算节点的可用性(或计算节点被保证可用)的使用时段、请求的到期时间等。请求API可包括关于请求的其它信息，例如与用户程序或用户对计算资源的需求有关的偏好、需求和/或限制。例如，请求API可包括关于在使用时段期间准许哪些用户访问计算资源、可在使用时段期间执行哪个程序(或哪些程序)、存储容量或网络带宽的量、节点的地理和/或逻辑位置、终止条件等的信息。

[0052] 在图2B所示的示例中，交互请求管理器130经由网络108传达作为由用户计算系统104a接收的(4)的确认API(3)。确认API可包括与在请求的使用时段期间(或在不同的使用时段期间)程序执行服务100是否可准许请求(全部或部分)有关的信息。确认API还可以包括与用户请求相关且要在使用时段期间结合访问分配的计算资源而使用的一个或多个请求识别符(例如，密钥、令牌、用户名、密码等)。确认API可包括其它信息，例如确认可满足用户的偏好、需求和/或限制的信息。

[0053] 图2B说明性地示出了用户计算系统104a经由API与程序执行服务100的交互请求管理器130以编程方式进行交互。程序执行服务100可经由API从其它用户计算系统(例如，用户计算系统104b、...、104n)接收对服务的计算资源的可用性的请求并且可经由API把确认传达给其它用户计算系统(在图2B的说明性示例中未示出这样的请求和确认)。交互请求管理器130(或其它合适组件)可调度来自多个用户计算系统的请求并且可在各个请求的使用时段期间分配计算资源。程序执行服务100与用户计算系统之间的其它类型的编程交互(此外或替代地)是有可能的。例如，可直接从用户(例如，经由由程序执行服务提供的交互控制台或其它GUI)、从自动开始执行其它程序或其自身的其它实例的用户的执行程序等接收请求。

[0054] 作为用户计算系统104a与程序执行服务100的交互请求管理器130之间的交互的额外说明性示例，用户可以请求交互请求管理器130生成用于用户程序的应用容器。使用请求API(1)，用户可诸如经由Java WAR文件上传程序。用户可使用Web浏览器、命令行工具或集成开发环境上传程序。程序可以是用户希望的任何应用。例如，程序可以是Web应用。接着，交互请求管理器130可接收程序(例如，请求API(2))。接着，交互请求管理器30可以通过自动生成用于用户程序的应用容器来处理用户请求。应用容器可包括使用户程序能够可伸缩和容错的用于用户的其它服务。在某些实施方案中，交互请求管理器30在没有来自用户的进一步输入的情况下自动生成应用容器(包括基础结构服务)。在其它实施方案中，用户

可选择(例如,通过选择和/或配置包括在容器中的基础结构服务)对应用容器的生成施加控制度。可创建代表用户生成应用堆栈、自动伸缩、负载平衡、版本控制、存储和/或其它服务的应用容器。例如,交互请求管理器130可以生成Linux操作系统、Apache Tomcat服务器、Oracle关系数据库的实例以及一个或多个服务,例如负载平衡、监控和自动伸缩,用于执行用户程序。接着,交互请求管理器130可以经由确认API(3)和(4)向用户确认应用容器的创建。

[0055] 在接收到确认API(4)之后,用户可以在本地改变程序。例如,用户可以使用集成开发环境在本地更新和/或测试应用并且可以(例如,经由后续请求API(1)、(2))再次把更新的程序发送到交互请求管理器130。如上文论述,交互请求管理器130可以(在具有或没有进一步的用户输入的情况下)自动生成或更新用于修订的程序的应用容器。在生成应用容器之后,用户可改变由交互请求管理器130生成的服务。用户可以在用户程序的执行开始之前或之后使用网页、交互控制台等进行这些更改。例如,用户可以优选Windows操作系统而非Linux操作系统或优选Microsoft SQL数据库而非Oracle数据库并且可例如使用API或交互控制台请求交互请求管理器130进行这些更改。作为另一示例,用户可能希望改变Apache Tomcat服务器实例的设置(如使用的端口)并且使用例如API或交互控制台请求更改。作为另一示例,用户可以优选关掉负载平衡并且可请求交互请求管理器130停用负载平衡。因此,用户可以请求任何类型的希望的变化,并且交互请求管理器130可尝试实施用户希望的变化。用户还可与任何其他用户共享应用容器。用户还可以随时请求删除用户程序和取消分配应用容器的计算资源。

[0056] 如本文论述,应用容器还可以包括用于用户程序的监控服务。监控服务可监控应用容器内的用户程序和/或服务的性能。例如,监控服务可以监控负载平衡和/或自动伸缩服务的性能或监控服务器的性能。监控服务可以把性能的报告提供给用户。如果在执行用户程序期间发生任何问题,那么监控服务还可采取纠正措施。例如,如果监控服务检测到用户程序没有响应(例如,一个或多个计算节点已出现故障),那么监控服务可重启该一个或多个计算节点和/或请求额外计算节点来接替有故障的节点。作为另一示例,如果对用户应用的需求增加使得存在影响用户程序的性能的风险,那么监控服务可代表用户程序分配额外计算资源和/或向用户通知该分配。用户可向交互请求管理器130指定如何或何时被通知。例如,用户可以使用API或交互控制台指定当出现任何错误时交互请求管理器130把电子邮件消息传达给用户。作为另一示例,用户可以指定交互请求管理器130把关于任何错误的电子邮件消息传达给用户并且在采取任何纠正措施之前从用户接收授权。用户可以任何希望的方式配置通知设置。前述旨在作为用户与交互请求管理器130的各个实施方案之间的交互的类型的说明性示例且并非旨在限制。

[0057] 图3A和图3B是示意地示出了交互请求管理器例程300的示例性实施方案的流程图。在某些实施方式中,例程300可通过参考图1和图2描述的程序执行服务100的交互请求管理器130的实施方案提供。示例性例程300是按照如下情形描述的:第一用户请求在使用时段期间的程序执行容量(例如,在程序执行服务的一个或多个计算节点上执行程序)(参见例如图3A),并且第二用户请求对在使用时段期间的程序执行容量进行更改(参见例如图3B)。如下文将论述的,第一用户和第二用户不需要是不同用户并且可指同一用户。示例性例程300旨在说明而非限制交互请求管理器130的各个方面。

[0058] 参考图3A，在方框304处，由交互请求管理器130从第一用户接收对在使用时段期间由程序执行服务100执行程序的请求。如上文论述，请求可以包括程序、计算节点的数量和/或类型、要使用的计算节点的最小和/或最大数量、在其期间要保证计算节点的可用性的未来使用时段、请求的到期时间等。请求可以指定在使用时段期间仅准许特定用户访问计算节点或在使用时段期间仅在计算节点上执行特定程序。对计算资源的请求可以包括其它类型的偏好、需求和/或限制(例如，存储器或存储容量的量、网络带宽、节点的地理和/或逻辑位置、终止条件等)。

[0059] 在方框308处，交互请求管理器130确定是否可满足请求。例如，在某些情况下，程序执行服务100可以具有用于满足请求的足够容量，或者使用时段是在足够久远的未来，使得可获取额外计算资源(如果需要)。如果可满足请求，那么在方框320处把可满足请求的确认提供给第一用户。例如，可以经由电子邮件给第一用户传达消息，或者程序执行服务可以经由Web服务或经由由程序执行服务提供的交互控制台或其它GUI提供确认。可以经由如参考图2B论述的确认API提供确认。

[0060] 如果无法全部或部分地满足请求，那么例程300继续进行到方框312，在方框312，交互请求管理器尝试确定是否可与所请求不同地全部或部分地满足请求。例如，例程300可以确定可在不同的使用时段期间满足请求或可在请求的使用时段期间部分地(例如，使用比所请求少的节点)满足请求。在某些情况下，在方框312处，例程300可以确定可取决于一个或多个额外事件在请求的使用时段期间满足请求。例如，例程300可以确定可取决于由程序执行服务获取足够的额外计算资源并取决于在请求的使用时段开始之前已交付和安装这些额外资源来满足请求。在方框316处，例程300把关于与请求有关的一个或多个可能的修改或偶发事件的信息提供给第一用户且接着例程300结束。例如，消息可以经由电子邮件传达给第一用户，或者程序执行服务可以经由Web服务或经由由程序执行服务提供的交互控制台或其它GUI提供信息。可以经由API提供信息(参见例如图2B)。第一用户可使用关于与请求有关的可能的修改或偶发事件的信息并且接着如果需要的话重新提交新请求。

[0061] 在所示实施方案中，如果可满足请求，那么例程300继续进行到方框324，在方框324，交互请求管理器生成包括用户程序和用于在程序执行服务上执行该程序的合适基础结构的虚拟化环境(例如，应用容器)。交互请求管理器确定可执行虚拟化环境的一组计算节点。在某些实施方式中，用户可以请求要用于执行程序的计算节点的特定数量、地理分布等。计算节点组中计算节点的数量(和/或地理分布)可以但无需与用户请求的计算节点的数量不同。例如，组中计算节点的数量可以小于请求的数量，因为如果在使用时段期间实际请求的计算节点的数量大于组中计算节点的数量，程序执行服务也具有足够的额外计算容量。在其它情况下，组中计算节点的数量可以大于请求的数量，以尝试确保将存在足够计算节点来可靠地满足在使用时段期间的预期需求(例如，提供储备节点，以防组中一个或多个计算节点发生故障)。在不同实施方案中，用户可对与虚拟化环境或节点组相关的设置(例如，存储容量或网络带宽的量或类型、节点的地理和/或逻辑位置、终止条件等)进行一个或多个更改。例程300可验证可履行或满足这一个或多个更改。在方框328处，分配可供第一用户在使用时段期间利用的计算节点组。如上文参考资源调度模块208所论述的，分配的计算节点组可以包括特定计算节点或选自计算节点池的节点。

[0062] 参考图3B，在使用时段期间，第二用户可以进行与在已分配给第一用户的计算节

点上第一用户的程序的执行相关的一个或多个更改。如上述，第二用户可以但无需与第一用户不同。在一个示例性情形中，第一用户可能已请求执行程序。在使用时段期间，第一用户可提交对在计算节点上执行的程序的设定的一个或多个更改。在这个示例性情形中，第二用户将与第一用户是同一用户。在某些这类情形中，第一用户的请求可指示仅第一用户(且没有其他用户)可对程序执行或分配的计算节点进行更改。

[0063] 在其它示例性情形中，第二用户可以是与第一用户不同的用户。例如，第一用户的请求可以指示特定的第二用户被授权进行相对于在使用时段期间在计算节点上执行的程序的更改。在这个示例性情形中，第二用户可是与第一用户不同的用户。例如，第一用户可以是最初上传并请求程序执行容量的应用开发者，而第二用户可以是监控正在进行的程序执行的网络管理员或应用开发管理员。

[0064] 在另一示例性情形中，第一用户的请求可以指示程序执行服务100的任何用户皆可对在使用时段期间的程序执行进行更改，前提是这个(第二)用户要提交正确识别符信息。在此类情形中，第一用户可以把识别符(例如，密钥、令牌、密码等)传达给各第二用户。接着，这些第二用户中的任何一个将在请求在使用时段期间进行更改时使用程序识别符。在某些这类情形中，如果用户请求了对在分配的计算节点上执行的程序的更改但却不具有(或未随请求一起提交)程序识别符，那么交互请求管理器将拒绝请求。

[0065] 在方框336处，交互请求管理器可对第二用户提供接口以使第二用户能够请求对被提供用于在计算节点上执行程序的虚拟化环境进行一个或多个更改。如上文论述，可以多种方式提供该接口。例如，可通过经由交互控制台或其它GUI、命令行工具、网页和集成开发环境等提供该接口。在方框340处，由交互请求管理器接收来自第二用户的更改请求。例如，该请求可以是终止程序的执行实例、启动一个或多个新实例来执行程序、修改一个或多个实例的运行时设置等。如上文论述，请求可以是更改分配的计算节点，例如资源分配的类型或量、执行的地理和/或逻辑位置、时间相关条件、终止条件等。

[0066] 在方框344处，交互请求管理器确定是否可准许或满足来自第二用户的更改请求。例如，第一用户的请求可能已指定对可能进行的对计算节点的更改施加一个或多个要求或限制，并且如果第二用户的请求不满足某些或所有要求或限制，那么可以拒绝第二用户的请求。在其它情况下，对计算节点的更改可能需要额外资源或计算节点，使得无法在第二用户请求之时满足第二用户的请求。在这种情况下，在各个实施方案中，交互请求管理器可以拒绝第二用户的请求或可以保留第二用户的请求或把其排入队列直到可产生供第二用户使用的足够计算节点为止。在某些实施方式中，交互请求管理器可以把有关何时可满足请求的预计时间、可如何修改请求使得请求可被立即满足等的信息提供给第二用户。

[0067] 在方框348处，如果可满足来自第二用户的程序执行请求，那么程序执行服务在可能正在执行第一用户的程序的计算节点上实施一个或多个更改。

[0068] 参考图3A，例程300在方框352处继续进行，在方框352处，交互请求管理器监控和追踪在被分配用于程序执行的节点组上的程序执行的使用。如参考监控和报告模块212论述的，交互请求管理器监控在分配的计算节点上执行程序的用户(例如，第一用户)的使用模式。使用模式可包括对分配的节点进行更改的用户的数量或身份、程序执行的开始/结束时间和持续时间和/或其它用户指定的模式或诊断。在某些实施方案中，在方框352处，可以把交互反馈(包括例如可能是在何时在分配的计算节点上执行程序和/或要执行多久、对节

点的实际或预测需求等的指示)提供给第一用户或第二用户。在某些实施方案中,可生成详述或汇总使用统计数据的报表并经由电子邮件或经由由程序执行服务提供的交互控制台或其它GUI把报表提供给第一用户。

[0069] 在程序执行服务要收费的实施方案中,在方框356处,交互请求管理器(或其它会计或记账管理器)可计算一笔或多笔费用。例如,可以向第一用户收取对于请求计算容量的预约费,并且可以向第一用户或第二用户收取针对在使用时段期间在分配的节点上执行程序的使用费。

[0070] 在方框360处,可选地可由例程300的实施方案执行其它服务。例如,可执行包括在使用时段到期之后释放计算节点以供其他用户使用的各种内务操作。接着,例程300继续进行到方框364处并结束。

[0071] 图3C是示意地示出了通过其交互请求管理器的实施方案可与用户计算系统进行通信以用于确认对计算资源的请求的例程370的示例的流程图。在某些实施方式中,可通过参考图1和图2描述的程序执行服务100的交互请求管理器130的实施方案实施例程370。如参考图3A的方框304和308论述的,交互请求管理器可从第一用户接收对在使用时段期间的程序执行容量的请求并且可确定是否可满足该对程序执行容量的请求。

[0072] 在图3C的方框374处继续进行,如果可满足请求,那么交互请求管理器把确认提供给第一用户。例如,如参考图2B和图3A的方框320论述的,确认可包括与在请求的使用时段期间(或在不同的使用时段期间)程序执行服务是否可准许请求(全部或部分地)有关的信息。确认还可以包括与第一用户的请求相关并且要结合在使用时段期间对计算资源进行更改而使用的一个或多个请求识别符(例如,密钥、令牌、用户名、密码等)。确认可包括其它信息,例如确认可满足用户的偏好、需求和/或限制的信息。在某些实施方式中,经由确认API传达确认(参见例如图2B)。

[0073] 在某些情况下,在确认请求之时(在方框374处)与使用时段开始之时之间第一用户的程序执行需求可以改变。在某些这类情况下,第一用户可以把请求的修改提交给交互请求管理器。例如,修改的请求可以包括关于要执行的修改的程序、计算节点的修改数量或与计算节点相关的设置、使用时段的修改的开始时间、终止时间和/或持续时间或第一用户的其它偏好或需求的变化的信息。修改的请求可以是全部或部分地取消初始的请求。相应地,在此类情况下,在方框378处,交互请求管理器可以从第一用户接收修改的请求并且确定是否可(全部或部分地)满足修改的请求。

[0074] 在方框382处,交互请求管理器把更新的确认提供给第一用户,更新的确认可包含与在请求的使用时段期间(其可能已在修改的请求中被修改)程序执行服务是否可准许修改的请求(全部或部分地)或在不同的使用时段期间程序执行服务是否可准许修改的请求(全部或部分地)有关的信息。更新的确认还可以包括与第一用户的修改的请求相关并且要结合在(可能更新的)未来使用时段期间对计算资源进行一个或多个更改而使用的一个或多个更新的请求识别符(例如,密钥、令牌、用户名、密码等)。更新的确认可包括其它信息,例如确认可满足用户的(可能更新的)偏好、需求和/或限制的信息。在某些实施方式中,更新的确认经由确认API传达到第一用户(参见例如图2B)。

[0075] 在方框386处,交互请求管理器可在使用时段期间从第二用户接收对对分配给第一用户的计算容量进行一个或多个更改的请求。如参考图3B大致描述,交互请求管理器可

以处理来自第二用户的请求。例如，在某些实施方式中，从第二用户接收的请求可以包括用于在方框374处(和/或如果从第一用户接收到修改的请求那么在方框382处)传达给第一用户的分配的计算容量的请求识别符。

[0076] 图4是示意地示出了通过其交互请求管理器的实施方案可与用户计算系统进行通信以用于提供供用户选择的多个虚拟化环境(例如，应用容器)的例程400的示例的流程图。在某些实施方式中，可通过参考图1和图2描述的程序执行服务100的交互请求管理器130的实施方案实施例程400。如参考图3B的方框340论述的，交互请求管理器可被配置来接收与为用户生成的虚拟化环境有关的至少一个更改。

[0077] 在图4的方框404处继续进行，交互请求管理器从第一用户接收共享虚拟化环境的请求。例如，在交互请求管理器已生成用于第一用户的虚拟化环境之后，第一用户可以决定与其他用户共享该虚拟化环境。为了启用虚拟化环境共享，第一用户可以把共享虚拟化环境的请求传达给交互请求管理器。该请求可以指定要共享哪些虚拟化环境。在某些实施方式中，经由请求API传达请求(参见例如图2B)。

[0078] 在示例性例程400的方框408处，交互请求管理器把虚拟化环境添加到虚拟化环境列表以创建虚拟化环境的更新的列表。更新的列表可以包括可以提供给程序执行服务的其他用户的虚拟化环境的列表。在某些情况下，更新的列表可以包括不仅由第一用户而且由程序执行服务和/或由程序执行服务的其他用户提供的虚拟化环境。某些虚拟化环境可以免费提供给用户，而其它环境可收费才可用。更新的列表可以包括例如虚拟化环境的描述、生成虚拟化环境的用户的名字、使用虚拟化环境的费用、对适合于与虚拟化环境一起使用的应用的类型的推荐、到类似或有关虚拟化环境的链接、与虚拟化环境的性能有关的统计数据、虚拟化环境的属性的设置等。在某些实施方式中，在交互请求管理器创建更新的列表之后，将确认经由确认API传达给第一用户(参见例如图2B)。

[0079] 在方框412处，交互请求管理器可以把应用容器的更新的列表提供给程序执行服务的第二用户。第二用户可与第一用户相同或不同。在某些实施方式中，交互请求管理器经由Web接口、GUI、API调用等把更新的列表提供给第二用户。例如，第二用户可以请求交互请求管理器在使用时段期间执行程序(参见例如图3A)。在请求过程期间，交互请求管理器可以提供虚拟化环境的更新的列表，使得第二用户可请求应将哪个虚拟化环境用于执行第二用户的应用。在某些实施方式中，第二用户可以提供用于执行第二用户的应用的一个或多个偏好或需求，并且交互请求管理器可以推荐满足某些或所有第二用户的偏好或需求的一个或多个虚拟化环境。例如，第二用户可以指示对第二用户的应用的预期需求、用户的客户的地理位置、希望的程序执行容量(例如，CPU、存储器、存储装置、带宽等)等等。交互请求管理器可以识别匹配某些或所有第二用户的偏好的一个或多个虚拟化环境并且把这个推荐提供给第二用户。在某些情况下，交互请求管理器可以根据满足第二用户的偏好或需求的可能性对推荐的虚拟化环境进行排名或排序。

[0080] 在示例性例程400的方框416处，交互请求管理器可以从第二用户接收对更新的列表中包括的至少一个虚拟化环境的选择。在某些实施方式中，第二用户可以在选择之前或之后对一个或多个虚拟化环境进行至少一个更改。例如，第二用户可以在把关于第二用户的选择的信息提供给交互请求管理器之前改变特定虚拟化环境的一个或多个设置。接着，交互请求管理器可以使用选定的虚拟化环境执行第二用户的应用的实例(参见例如图3A)。

[0081] 在某些实施方案中，在第二用户已选择至少一个虚拟化环境之后，可将费用提供给提供了选定的虚拟化环境的第一用户。例如，如上文论述，可以向第二用户收取用于使用至少一个选定的虚拟化环境执行用户的应用的一笔或多笔费用(例如，预约费、使用费等)。一笔或多笔费用的一部分可以提供给第一用户。例如，提供的费用可以是固定费用或向第二用户收取的某些或所有费用的一定比例。提供的费用还可以是基于第二用户已对至少一个选定的虚拟化环境进行的更改的数目、已选择特定虚拟化环境的用户的数量、特定虚拟化环境已存在或被使用的持续时间等的分级费用。

[0082] 在示例性例程400的某些实施方式中，第二用户可以选择改变选定的虚拟化环境。例如，第一用户可能已提供基于例如Java的虚拟化环境，而第二用户可以修改该环境使得其是基于另一编程语言(例如，Ruby)。第二用户可以选择与程序执行服务的其他用户共享包括由第二用户进行的更改的经过修改的虚拟化环境。第二用户可以请求交互请求管理器与其他用户共享这个虚拟化环境(参见例如方框404)并且(可选地)被提供使用费。例如，第三用户可以选择第一用户的虚拟化环境，第一用户的虚拟化环境的第二用户的修改版本和/或由程序执行服务和/或由其他用户提供的虚拟化环境。相应地，可以开发包括很多种虚拟化环境(免费和/或收费)的市场并且使其对程序执行服务的用户开放。

[0083] 在某些实施方案中，可提供被配置来管理用户程序的执行的计算系统。该系统可以包括被配置来管理程序执行服务的用户的程序的执行的交互请求管理器组件。交互请求管理器组件可以被配置来从程序执行服务的用户接收生成用于在使用时段期间执行用户应用的虚拟化环境的请求，该请求包括允许程序执行服务至少部分基于用户应用来执行程序的与用户应用相关的信息，程序执行服务提供可被配置来执行程序执行服务的多个用户的程序并生成虚拟化环境的多个计算节点，虚拟化环境包括一个或多个程序服务，程序服务包括：(1)负载平衡器，其被配置来跨虚拟化环境的计算资源分布工作量，(2)监控接口，其被配置来允许用户监控程序的执行，(3)负载调节器，其被配置来响应于对程序执行的需求的变化而调节计算资源，和(4)多个数据库管理服务。交互请求管理器组件还可以被配置来在一个或多个计算节点的组上执行虚拟化环境的一个或多个实例、确定对虚拟化环境或对在执行虚拟化环境的一个或多个实例期间在一个或多个计算节点的组上执行虚拟化环境的一个或多个实例的至少一个更改。交互请求管理器组件还可以被配置来在执行虚拟化环境的一个或多个实例期间实施至少一个更改。

[0084] 在某些实施方式中，可以提供被配置来管理用户程序的执行的计算系统。该系统可以包括被配置来管理程序执行服务的用户的程序的执行的交互请求管理器组件。交互请求管理器组件可以被配置来从第一用户接收与程序执行服务的用户共享为第一用户生成的第一虚拟化环境的请求、更新可供程序执行服务的用户选择的虚拟化环境的列表以提供包括与第一虚拟化环境有关的信息的虚拟化环境的更新的列表，以及把虚拟化环境的更新的列表提供给程序执行服务的用户。交互请求管理器组件还可以被配置来从程序执行服务的第二用户接收对更新的列表中包括的第一虚拟化环境的选择并生成用于第二用户的第二虚拟化环境，第二虚拟化环境至少部分基于第一虚拟化环境。

[0085] 前面段落中描述的每个过程、方法和算法可以以由一个或多个计算机或计算机处理器执行的代码模块的形式具体实施并且通过这样的代码模块完全或部分地自动化。代码模块可以存储在任何类型的非瞬时性计算机可读介质或计算机存储装置(如硬盘驱动器、

固态存储器、光盘等等)上。系统和模块还可以作为生成的数据信号(例如,作为载波或者其它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(包括基于无线和基于有线/电缆的介质)上传输,并且可以采取多种形式(例如,作为单用或复用模拟信号的一部分、或作为多个离散数字数据包或帧)。可以部分或全部以专用电路的形式实施过程和算法。公开的过程和过程步骤的结果可以永久或以其它方式存储在任何类型的非瞬时性计算机存储装置(例如易失性或非易失性存储装置)中。

[0086] 上文描述的各个特征和过程可以独立于彼此使用,或者可以以各种方式进行组合。所有可能的组合和子组合旨在落于本公开内容的范围内。此外,在某些实施方式中可以省略特定方法或过程块(方框)。本文描述的方法和过程也不限于任何特定顺序,并且与其相关的块或状态可按其它适当顺序执行。例如,描述的块或状态可以按具体公开的次序外的次序执行,或者多个块或状态可以组合成单个块或状态。示例性块或状态可以串行、并行或以某种其它方式执行。块或状态可以添加到公开的示例性实施方案中或从其中移除。可以与所描述不同地配置本文描述的示例性系统和组件。例如,元件可以添加到公开的示例性实施方案中、从其中移除或与其相比而进行重新安排。

[0087] 除非另有特别说明或在如使用的上下文内另外所理解,否则本文使用的条件语言(如尤其是“可”、“可能”、“可以”、“例如”等等)通常旨在表达特定实施方案包括而其它实施方案不包括特定特征、元件和/或步骤。因此,这样的条件语言通常并非旨在暗示特征、元件和/或步骤是以任何方式对一个或多个实施方案所必需的或一个或多个实施方案必定包括用于在具有或没有程序设计者输入或提示的情况下决定是否包括这些特征、元件和/或步骤或其是否要在任何特定实施方案中被执行的逻辑。术语“包括”(“comprising”、“including”)、“具有”(“having”)等等是同义的且以开放方式按包括之意使用,并且不排除额外元件、特征、动作、操作等等。此外,术语“或”是以其包括之意(而非以其排除之意)使用,使得当例如用来连接一系列元件时,术语“或”意指其中的一个、某些或所有元件。

[0088] 虽然已描述特定示例性实施方案,但是这些实施方案仅按举例方式来呈现,且并非旨在限制本文公开的发明的范围。因此,前文描述中的任何部分并非旨在暗示任何特定特征、特性、步骤、模块或块是必要的或不可缺少的。实际上,本文描述的新颖方法和系统可以以多种其它形式具体实施;此外,可以在不脱离本文公开的本发明的精神的情况下对本文描述的方法和系统的形式进行各种省略、置换和改变。权利要求书和其等效物旨在涵盖诸如将落于本文公开的特定发明的范围和精神内的形式或修改。

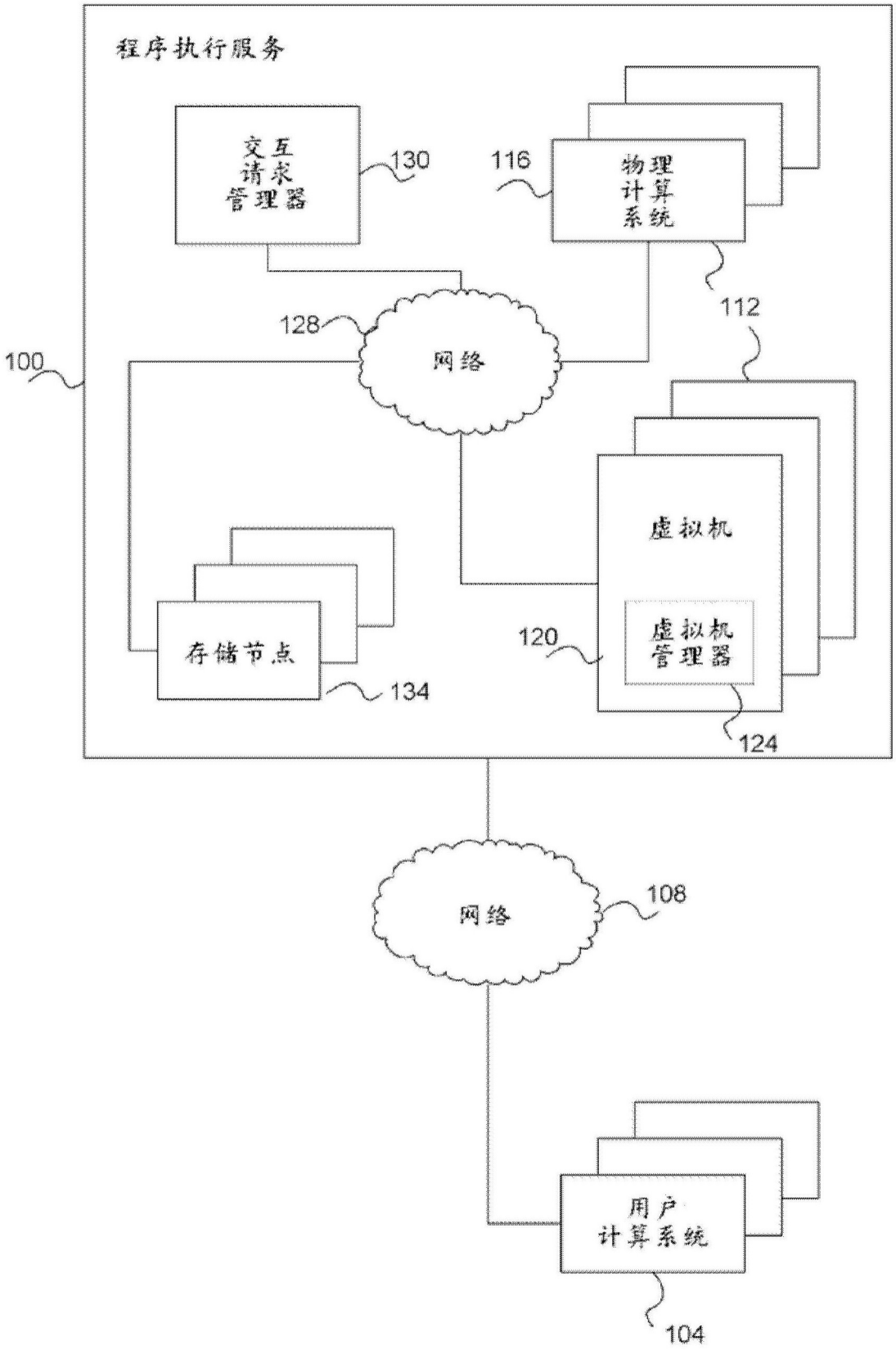

程序执行服务
交互请求管理器
130
116
物理计算系统
112
128
网络
100
虚拟机
存储节点
134
120
虚拟机管理器
124
网络
108
用户计算系统
104

图1

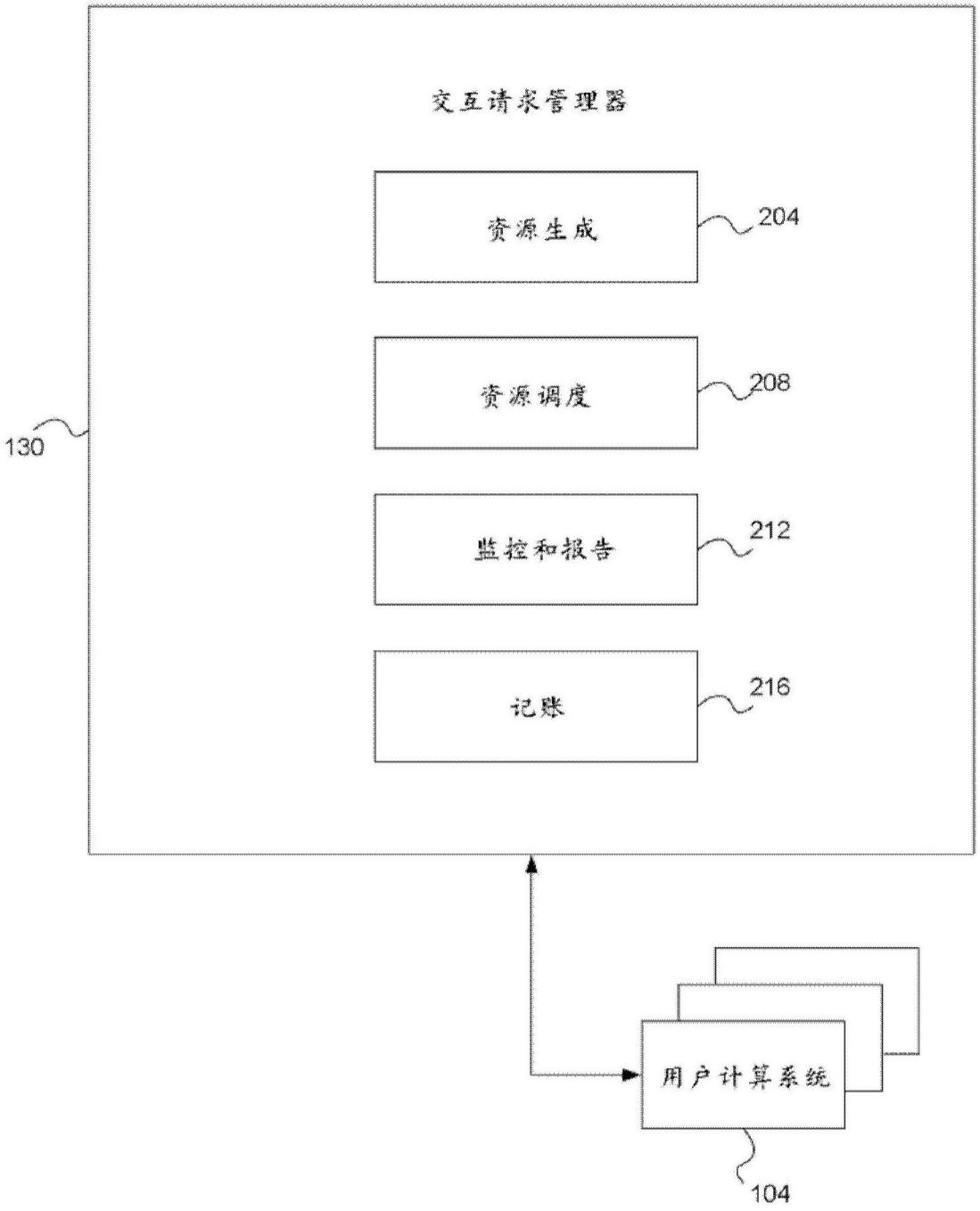


图2A

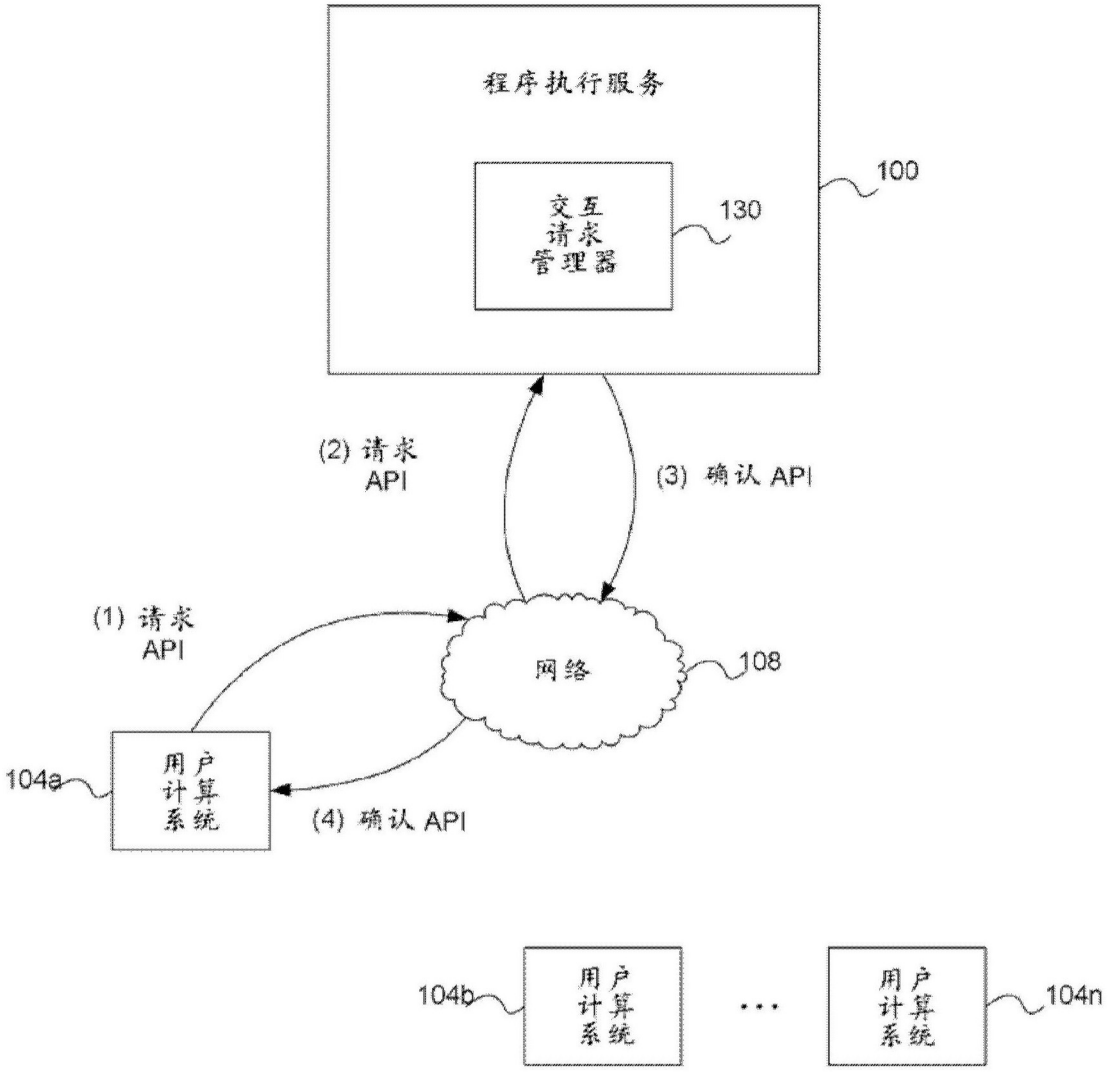

程序执行服务
交互请求管理器
130
100
(2) 请求API
(3) 确认 API
(1) 请求API
网络
108
104a
用户计算系统
(4) 确认 API
104b
用户计算系统
...
用户计算系统
104n

图2B

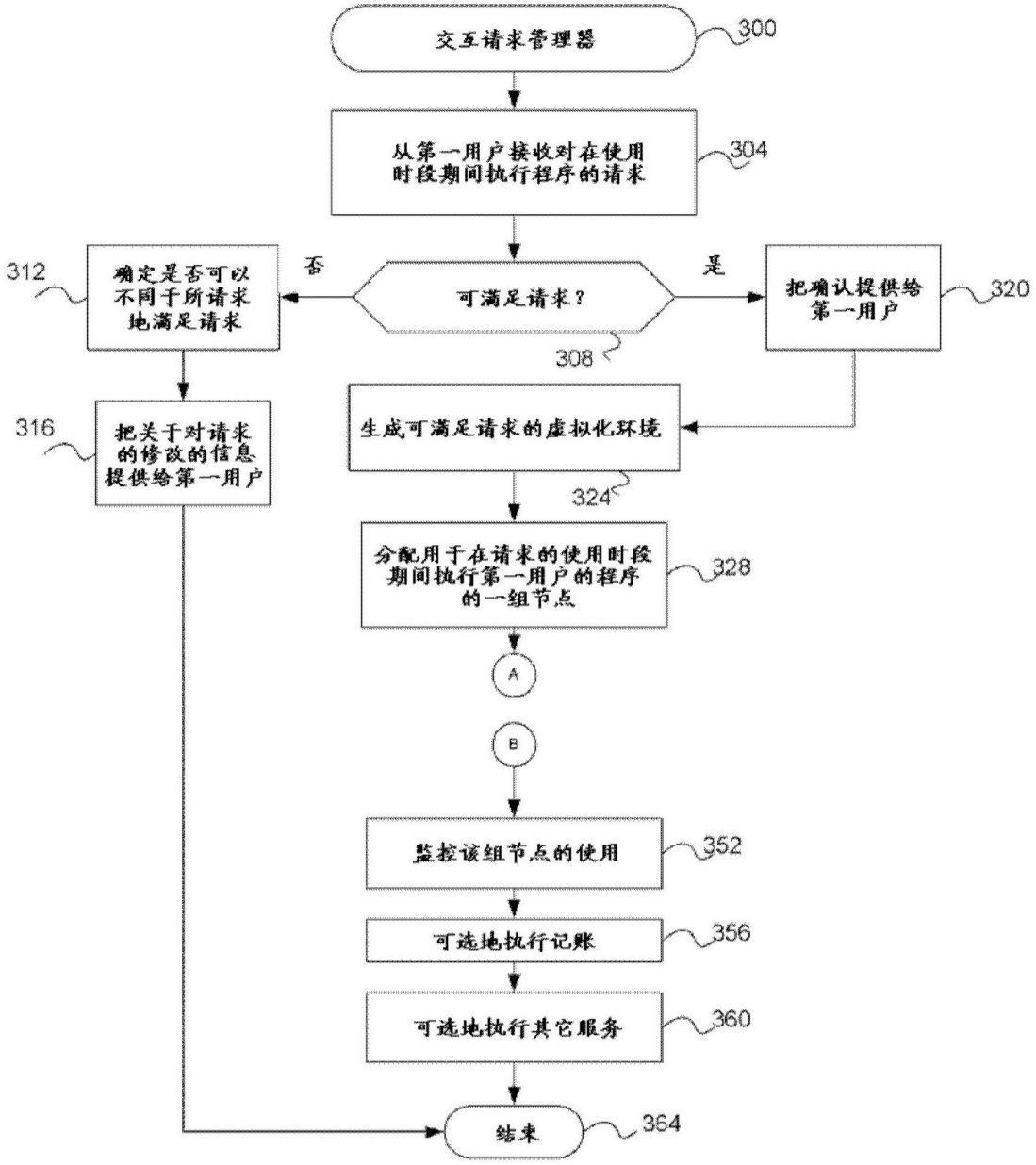

交互请求管理器
300
从第一用户接收对在使用时段期间执行程序的请求
304
可满足请求？
否
是
确定是否可以不同于所请求地满足请求
312
把确认提供给第一用户
320
308
把关于对请求的修改的信息提供给第一用户
316
生成可满足请求的虚拟化环境
324
分配用于在请求的使用时段期间执行第一用户的程序的一组节点
328
A
B
监控该组节点的使用
352
可选地执行记账
356
可选地执行其它服务
360
结束
364

图3A

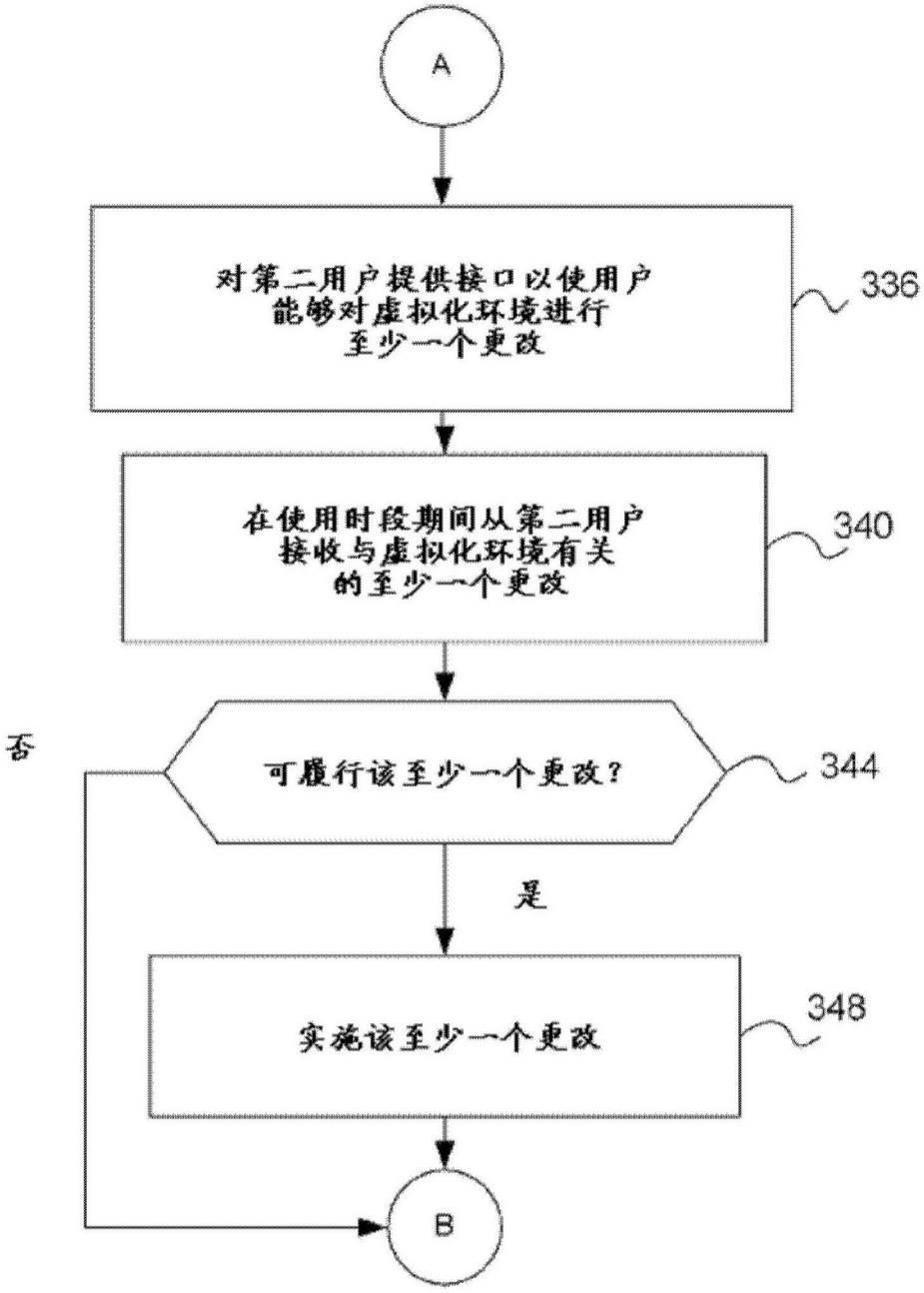

A
对第二用户提供接口以使用户能够对虚拟化环境进行至少一个更改
336
在使用时段期间从第二用户接收与虚拟化环境有关的至少一个更改
340
可履行该至少一个更改？
344
否
是
实施该至少一个更改
348
B

图3B

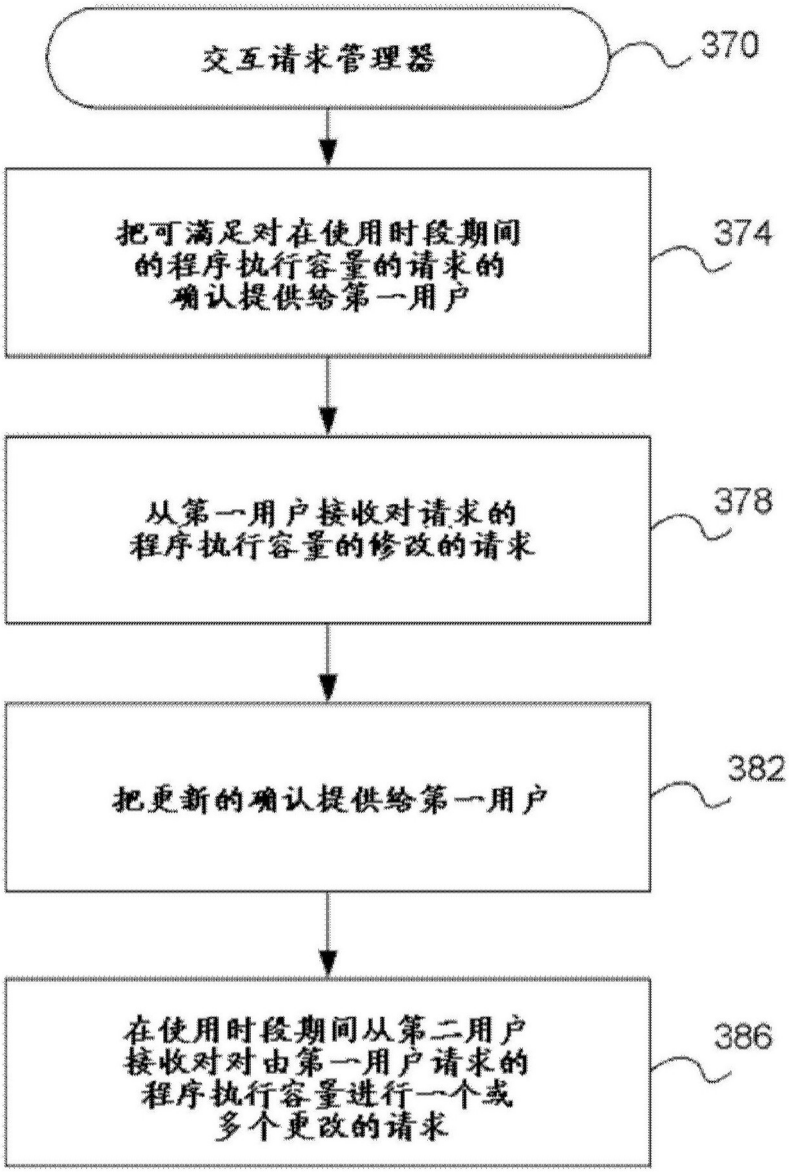

交互请求管理器
370
把可满足对在使用时段期间的程序执行容量的请求的确认提供给第一用户
374
从第一用户接收对请求的程序执行容量的修改的请求
378
把更新的确认提供给第一用户
382
在使用时段期间从第二用户接收对对由第一用户请求的程序执行容量进行一个或多个更改的请求
386

图3C

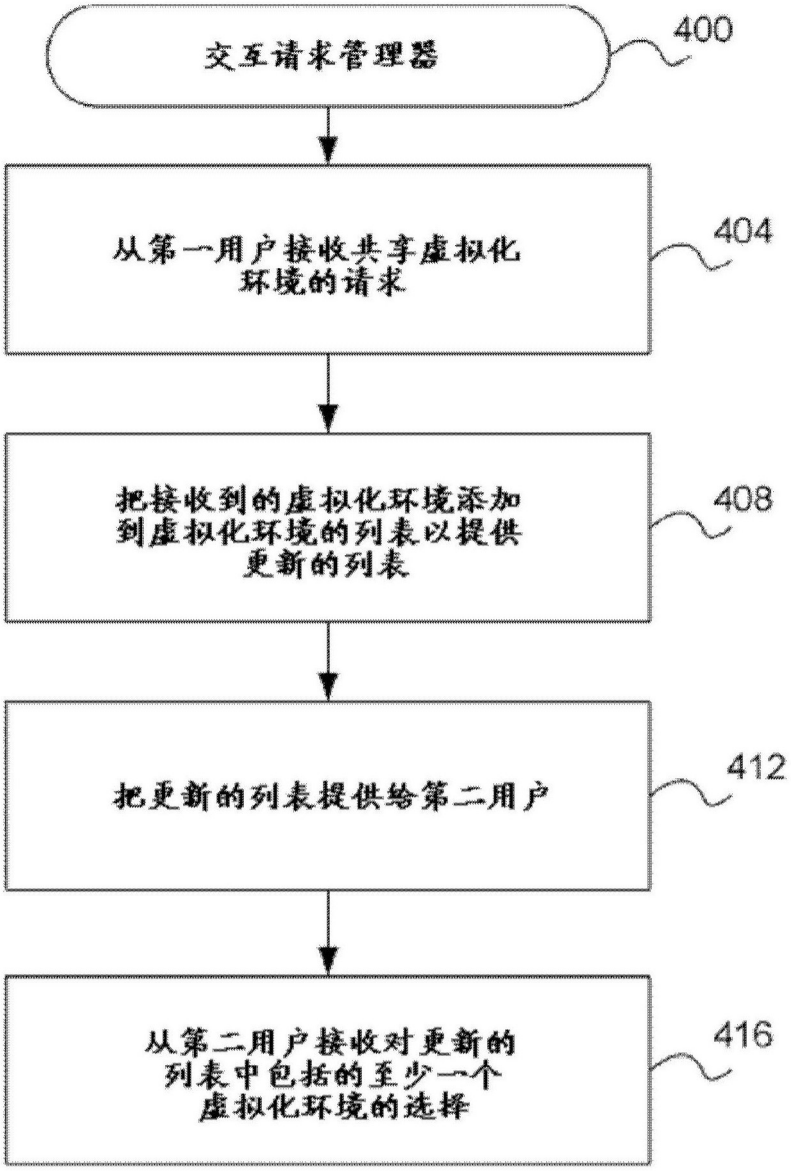

交互请求管理器
400
从第一用户接收共享虚拟化环境的请求
404
把接收到的虚拟化环境添加到虚拟化环境的列表以提供更新的列表
408
把更新的列表提供给第二用户
412
从第二用户接收对更新的列表中包括的至少一个虚拟化环境的选择
416

图4