(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 107077385 A
(43)申请公布日 2017.08.18

(21)申请号 201580048245.4

(22)申请日 2015.09.10

(30)优先权数据
14/482,841 2014.09.10 US
14/482,812 2014.09.10 US
14/482,789 2014.09.10 US

(85)PCT国际申请进入国家阶段日
2017.03.09

(86)PCT国际申请的申请数据
PCT/US2015/049521 2015.09.10

(87)PCT国际申请的公布数据
WO2016/040699 EN 2016.03.17

(71)申请人 亚马逊技术公司
地址 美国华盛顿州

(72)发明人 A·A·艾彻 M·J·埃迪 R·A·哈曼

(74)专利代理机构 北京纪凯知识产权代理有限公司 11245
代理人 赵蓉民 张全信

(51)Int.Cl.
G06F 9/48(2006.01)
H04L 29/08(2006.01)

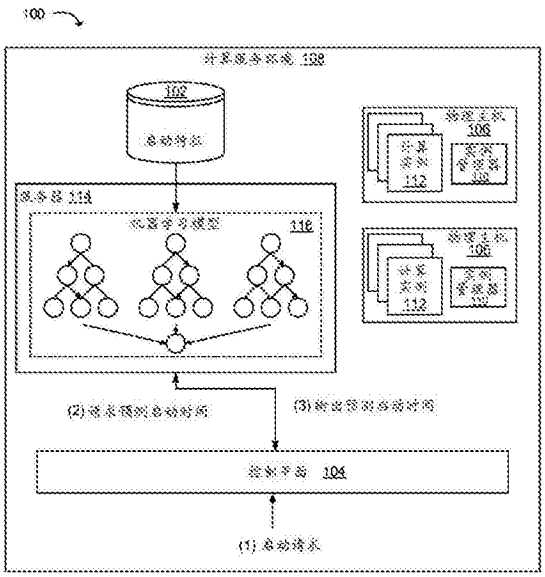
权利要求书2页 说明书35页 附图21页

(54)发明名称

计算实例启动时间

(57)摘要

描述了一种用于预测计算实例的启动时间的技术。示例方法可包括接收对在计算服务环境内的物理主机上启动计算实例的预测启动时间的请求。然后可获得与计算实例的启动特征相关联的数据，其中所述启动特征可被确定为对所述计算实例在计算服务环境内的物理主机上的启动时间具有影响。所述计算实例的所述启动特征然后可被输入至机器学习模型，所述机器学习模型输出用于在所述计算服务环境内启动所述计算实例的所述预测启动时间。



CN 107077385 A

1.一种计算机实现的方法,其包括:

在配置有可执行指令的一个或多个计算机系统的控制下,

接收对在计算服务环境内的物理主机上启动计算实例的预测启动时间的请求;

使用处理器获得与计算实例的启动特征相关联的数据,所述计算实例的所述启动特征被确定为对所述计算实例在计算服务环境内的物理主机上的启动时间具有影响;以及

使用所述处理器,将所述计算实例的所述启动特征输入至机器学习模型,所述机器学习模型输出用于在所述计算服务环境内启动计算实例的预测启动时间。

2.如权利要求1所述的方法,其中获得与启动特征相关联的所述数据还包括获得与机器映像启动特征、物理主机启动特征和客户配置启动特征相关联的数据。

3.如权利要求1所述的方法,其还包括在将所述启动特征输入至所述机器学习模型之前使所述启动特征标准化。

4.如权利要求1所述的方法,其还包括对机器学习参数进行参数值搜索,所述机器学习参数引起所述机器学习模型对所述启动特征的拟合优度。

5.如权利要求4所述的方法,其中进行所述参数值搜索还包括,使用分布式遗传算法进行机器学习参数的参数值搜索。

6.如权利要求1所述的方法,其中获得与启动特征相关联的数据还包括,获得表示多个先前计算实例启动的启动特征的活动训练数据;

从与所述启动特征相关联的所述活动训练数据提取特征;以及

使用来自所述活动训练数据的所述特征训练所述机器学习模型。

7.如权利要求1所述的方法,其中将所述启动特征输入至机器学习模型还包括:将所述启动特征输入至选自以下中的至少一个的机器学习模型:随机森林模型、超随机树模型、AdaBoost模型、随机梯度下降模型或支持向量机器模型。

8.如权利要求1所述的方法,其中将所述启动特征输入至机器学习模型还包括,将所述计算实例的所述启动特征输入至机器学习回归模型。

9.如权利要求1所述的方法,其还包括,从客户接收启动所述计算实例的启动请求;

识别与所述计算实例的计算实例类型相关联的SLA启动时间;以及

通过比较所述计算实例的所述预测启动时间与所述SLA启动时间来确定是否可能满足所述SLA启动时间。

10.如权利要求9所述的方法,其还包括通知计算服务提供商:当确定所述预测启动时间大于所述SLA启动时间时,所述SLA启动时间可能不会实现。

11.如权利要求9所述的方法,其还包括构建用于所述预测启动时间大于所述SLA启动时间的SLA违反特征,并且包括所述SLA违反特征与其它特征输入至机器学习分类模型。

12.如权利要求1所述的方法,其还包括:

识别所述计算实例的SLA启动时间;

通过比较所述计算实例的所述预测启动时间与所述SLA启动时间来确定是否可能满足所述SLA启动时间;以及

分析所述计算实例将被启动的计算服务环境的状态,以确定是否可能进行动作,在已经确定了将可能违反所述SLA启动时间时,所述动作可能防止违反所述SLA启动时间。

13.一种系统,其包括:

处理器；

包括指令的存储器装置，在由所述处理器执行时，所述指令使所述系统：

识别包含在启动配置中的启动特征，所述启动特征已经被确定为对计算实例在计算服务环境内的启动时间具有影响；

从数据源获得用于所述启动特征的数据；

将所述启动特征输入至机器学习模型，所述机器学习模型输出用于在所述计算服务环境内启动计算实例的预测启动时间；以及

通过比较所述预测启动时间与SLA启动时间来确定是否可能满足所述SLA启动时间。

14.如权利要求13所述的系统，其中所述存储器装置包括指令，在由所述处理器执行时，所述指令使所述系统将所述启动特征输入至机器学习回归模型。

15.如权利要求13所述的系统，其中所述存储器装置包括指令，在由所述处理器执行时，所述指令使所述系统获得：用于启动所述计算实例的机器映像的特征、能够托管所述计算实例的物理主机服务器的特征和在启动时附接至所述计算实例的计算实例附接的特征。

计算实例启动时间

[0001] 发明背景

[0002] 用于计算资源的虚拟化技术的出现关于为具有不同需求的许多客户管理大规模计算资源提供了益处并且已经允许由多个客户有效且安全地共享各种计算资源或计算服务。例如，虚拟化技术可通过使用管理程序向每个客户提供由单个物理计算机器托管的一个或多个计算实例来允许在多个客户之间共享单个物理计算机器。每个计算实例可以是充当不同逻辑计算系统的客户机，其向客户提供客户是给定虚拟化硬件计算资源的唯一操作者和管理员的感觉。

[0003] 在单个物理计算机器上启动一个或多个计算实例可需要识别在其上可加载并执行计算实例的可用计算资源(例如，物理主机)。在主机服务器上加载并启动计算实例的时间可由于包含物理主机的计算环境的各个方面和正被启动的计算实例的方面而变化。因此，计算实例的启动时间可从几分钟到几分钟变化。

[0004] 附图简述

[0005] 图1是示出用于预测计算服务环境内计算实例的启动时间的示例系统的框图。

[0006] 图2是示出包括在用于预测计算实例启动时间的系统中的各种示例组件的框图。

[0007] 图3是示出包括预测启动时间服务的示例计算服务环境的框图。

[0008] 图4是示出用于配置和训练用于生成预测启动时间的机器学习模型的示例方法的图。

[0009] 图5是示出用于使用预测启动时间来预测违反SLA(服务水平协议)启动时间的示例方法的流程图。

[0010] 图6是示出用于预测计算实例的启动时间的示例方法的流程图。

[0011] 图7是示出可用于执行用于预测计算实例的启动时间的方法的计算装置的示例的框图。

[0012] 图8是示出根据本技术的示例的包括在用于使用估计启动时间将计算实例放置在计算服务环境中的物理主机上的系统中的各种组件的框图。

[0013] 图9示出根据本技术的示例的用于使用估计启动时间将计算实例放置在计算服务环境中的物理主机上的系统和相关操作。

[0014] 图10示出根据本技术的示例的用于使用估计附接时间来确定将计算实例放置在计算服务环境中的系统和相关操作。

[0015] 图11示出根据本技术的示例的用于使用估计启动时间将计算实例放置在计算服务环境中的系统和相关操作。

[0016] 图12是示出根据本技术的示例的生成用于预测在计算服务环境中启动的计算实例的启动时间的启动时间预测模型的框图。

[0017] 图13是用于使用计算服务环境内的估计启动时间来确定计算实例放置的示例方法的流程图。

[0018] 图14是用于使用计算服务环境内的估计启动时间来确定计算实例放置的另一示例方法的流程图。

[0019] 图15示出根据本技术的示例的用于使用启动时间预测来组织机器映像的缓存以便减少计算服务环境中的计算实例启动时间的系统和相关操作。

[0020] 图16是示出根据本技术的示例的包括在用于使用启动时间预测来组织机器映像的缓存以便减少计算服务环境中的计算实例启动时间的系统中的各种组件的框图。

[0021] 图17示出根据本技术的示例的用于使用启动时间预测来组织机器映像的缓存以便减少计算服务环境中的计算实例启动时间的系统和相关操作。

[0022] 图18示出根据本技术的示例的用于识别计算服务环境中的物理主机来缓存机器映像以便实现用于启动计算实例的期望启动时间的系统和相关操作。

[0023] 图19示出根据本技术的示例的用于在计算服务环境中缓存机器映像以便遵照计算服务环境的服务水平协议(SLA)的系统和相关操作。

[0024] 图20是示出根据本技术的示例的生成用于预测在计算服务环境中启动的计算实例的启动时间的启动时间预测模型的框图。

[0025] 图21是用于减少计算实例启动时间的示例方法的流程图。

[0026] 图22是用于减少计算实例启动时间的另一示例方法的流程图。

具体实施方式

[0027] 描述了一种用于确定计算服务内的计算实例的预测启动时间的技术。在该技术的一个示例中,响应于对预测启动时间(例如,在计算服务内的物理主机上启动计算实例的时间)的请求,与在物理主机上启动计算实例相关联的启动特征可被输入至输出预测启动时间的机器学习模型中。用作至机器学习模型的输入的启动特征可以是已经被确定为对计算实例在物理主机上启动的时间量具有影响的启动特征。如本公开中所提到,计算实例可以是类似物理机器执行应用的虚拟机(例如,计算机的软件实现的实例)。计算服务可以是向客户提供网络可访问计算实例的网络可访问服务。

[0028] 可使用表示来自先前计算实例启动的启动度量的特征来训练用于生成预测启动时间的机器学习模型。用于训练机器学习模型的特征可以是被确定为对启动计算实例的时间量具有影响的特征。在一个示例配置中,可使用从历史启动度量提取的特征(例如,每周使用前一周的数据)离线(例如,在非生产环境中)进行机器学习模型的训练。在另一示例配置中,可使用从最近启动度量提取的特征来在线时(例如,在生产环境中)训练机器学习模型。

[0029] 在一个示例中,计算实例的启动时间可包括执行服务调用以建立计算实例资源(例如,存储和网络接口)、为计算实例选择物理主机并在物理主机上创建计算实例。启动时间可基于计算实例的启动配置而变化。因此,计算服务提供商可能难以提供用于特定计算实例何时可供使用的预期时间范围。作为该技术的结果,计算服务提供商可获得预测启动时间,该预测启动时间然后可用于多个目的。例如,计算服务提供商可向客户提供计算实例何时可用于使用的估计、确定是否可满足SLA(服务水平协议)时间、建议可引起更快启动时间的启动配置以及一些其它目的。

[0030] 图1是示出可用于预测计算服务环境108内计算实例112的启动时间的系统100的高级示例的图。系统100可包括经由实例管理器110(例如,管理程序)执行计算实例112的多个物理主机106,和执行机器学习模型116的服务器114。在一个示例配置中,服务器114可与

多个数据源通信，可从所述数据源获得启动特征102的数据（例如，训练数据和与启动请求相关联的数据）。可使用历史训练数据训练机器学习模型116，之后机器学习模型116可通过使用计算实例启动的启动特征102以确定计算实例启动的预测启动时间来生成计算实例112的预测启动时间。

[0031] 作为说明，执行先前已经训练的机器学习模型116（例如，随机森林回归模型）的服务器114可接收对预测启动时间的请求。预测启动时间可以是接收启动请求（例如，计算实例状态是“待决”）和计算实例引导的开始（例如，计算实例状态是“运行”）之间的时间。对预测启动时间的请求可参考用于识别与启动请求相关联的启动特征102的启动配置。所识别的启动特征102可被机器学习模型116用来确定计算实例112的预测启动时间。作为说明，启动请求(1)可被发送至计算服务的控制平面104，来请求启动计算实例112。在接收到启动请求时，可生成指定用于启动计算实例112的各种参数的启动请求配置。例如，启动请求配置可指定计算实例112的计算实例类型（例如，微型、中小型、大型等以及通用目的、存储器密集型等）、计算实例112的机器映像、与计算实例112相关联的网络类型、附接至计算实例112的存储卷、被选择来托管计算实例112的物理主机106以及其它规范。

[0032] 控制平面104然后可针对预测启动时间对托管机器学习模型116的服务器114做出请求(2)。使用包括在启动请求配置中的信息，可收集被识别为影响启动时间的启动特征，且然后可将启动特征提供至机器学习模型116。作为说明，可参考启动请求配置来获得关于要启动的计算实例112的信息、要附接至计算实例112的附件和关于将托管计算实例112的物理主机106的信息。来自启动请求配置的信息然后可用于识别启动特征，诸如用于创建计算实例112的机器映像和内核映像、操作系统和网络类型、物理主机106将要位于的地理区域、物理主机106能够执行的最大数量的计算实例112等。使用来自启动请求配置的信息识别的启动特征102然后可被提供作为至机器学习模型116的输入，机器学习模型然后可输出(3)计算实例112的预测启动时间。

[0033] 由机器学习模型116生成的预测启动时间可用于任何目的。例如，预测启动时间可以是在分析中用于改进计算实例启动时间的因素，预测启动时间可用于确定计算实例112的物理主机106放置，预测启动时间可用于确定提供至客户的SLA（服务水平协议）启动时间，或预测启动时间可以是建议客户引起更快启动时间的计算实例配置的因素。作为利用预测启动时间的说明，可将SLA启动时间（例如，计算服务提供商和客户之间启动计算实例112的时间的协议）与计算实例112的预测启动时间进行比较以确定是否可能满足SLA启动时间。这样，可通知计算服务提供商和/或客户可能违反计算实例112的SLA启动时间，这可允许计算服务提供商和/或客户响应于通知而采取行动。

[0034] 在将机器学习模型116放置在机器学习模型116接收到对预测启动时间的请求的生产环境中之前，机器学习模型可被训练以便预测各种计算实例启动配置的启动时间。在一个示例配置中，可使用已经被确定为对在计算服务环境108内计算实例112的启动时间具有影响的特征来训练机器学习模型116。在确定哪些特征对计算实例112的启动时间具有影响时，可进行计算实例启动的分析来识别与启动计算实例112相关或与其相关联的特征。作为说明，计算实例112的启动可涉及以下步骤：执行服务调用以为计算实例112设置计算实例资源（例如，存储和网络接口）、为计算实例112选择物理主机106（例如，位置）以及在物理主机上创建计算实例112。

[0035] 分析计算实例启动的步骤可识别与启动计算实例112相关联的特征。例如，与设置计算实例资源相关联的特征、与选择物理主机106相关联的特征以及与计算实例112的配置相关联的特征(例如，用于创建计算实例112的机器映像)。然后可根据特征对启动时间的影响对所识别的那些特征进行排序或排名。例如，可根据对启动时间具有最大影响的特征进行排名，且与对启动时间几乎没有影响的那些特征相比，对启动时间具有最大影响的那些特征可接收更高排名。当确定预测启动时间时，可选择并使用具有较高排名的那些特征。

[0036] 在另一示例中，被确定为对计算实例112的启动时间具有影响的特征可选自多个特征类别。说明性地，特征类别可包括机器映像特征、物理主机特征和客户配置特征(例如，在客户控制下进行修改的启动配置的特征)。可选择并使用来自这些类别的特征来确定预测启动时间。

[0037] 被选择为对计算实例112的启动时间具有影响的那些特征的特征数据可从各自数据源(例如，活动训练数据或历史训练数据)检索并且用于训练机器学习模型116。特征数据可以是例如来自计算服务108内的先前计算实例启动的启动度量。在一些示例中，当特征数据冗余或较大时，特征数据可被转换为特征的缩减表示集合(例如，特征向量)。此外，特征数据可在训练机器学习模型116之前被标准化。

[0038] 在一个示例配置中，可使用历史训练数据(例如，与启动计算服务108内的计算实例112相关联的存档数据)离线训练(例如，在机器学习模型116放置于生产之前)机器学习模型116。在使用历史训练数据训练机器学习模型116之后，可在线(例如，在生产环境中)放置机器学习模型116，其中机器学习模型116可处理对预测启动时间的请求。在一些示例中，周期性地，可使用从上次训练机器学习模型116以来累积的历史训练数据离线提取并再次训练机器学习模型116。

[0039] 在另一示例配置中，机器学习模型116可最初使用历史训练数据来训练并且置于生产中，其中机器学习模型116可处理对预测启动时间的请求。随后，可在使用活动训练特征(例如，与在计算服务环境108内启动计算实例112相关联的最近特征数据)的生产时重新训练机器学习模型116。例如，在过去数分钟、小时或天内累积的特征数据可用于重新训练机器学习模型116或进一步改进机器学习模型116的训练。在相对较短时间间隔内累积的特征数据集可足够小，使得机器学习模型116可在短时间段(例如，分钟)内重新训练，而不必使机器学习模型116停产。

[0040] 图2示出可执行本技术的示例系统200的组件。系统200可包括多个客户装置228可经由网络226访问的计算服务环境202。计算服务202可向客户提供网络可访问服务，诸如在物理主机236上执行的计算实例。包括在计算服务环境202中的可以是托管启动时间预测模块218的服务器204，启动时间预测模块可用于生成在物理主机236上启动的计算实例的预测启动时间。除了启动时间预测模块218之外，服务器204还可包含训练模块222、启动特征模块220和具有包含在服务器204上的模块可访问的数据的一个或多个数据存储区206。

[0041] 在一个示例配置中，启动时间预测模块218可被配置为使用机器学习模型生成预测启动时间。启动时间预测模块218可为放置在位于计算服务环境202的特定部分内的物理主机236上的计算实例提供预测启动时间。例如，如图2所示，启动时间预测模块218可在计算服务环境202内执行，并且可为在计算服务环境202中启动的计算实例提供预测启动时间。在另一示例配置中，启动时间预测模块218可处于任何计算服务外，并且可通过网络的

方式从任何数量的计算服务接收对预测启动时间的请求。

[0042] 可由启动时间预测模块218用来预测启动时间的机器学习模型的示例可包括回归模型诸如随机森林模型、超随机树模型、AdaBoost模型、随机梯度下降模型、支持向量机模型以及这里未特别提及的其它类型的机器学习模型。

[0043] 训练模块222可被配置为从各种数据源获得特征(其然后用于训练由启动时间预测模块218使用的机器学习模型)。在一个示例中,可从数据仓库224检索特征和训练数据。特征数据可以是来自已经存储至数据仓库224的计算服务202内的先前计算实例启动的启动度量。说明性地,信息管理服务238可将启动相关的数据推送(例如,上传)至数据仓库224,从而使得训练模块222可访问数据。从数据仓库224检索的数据可以是与计算实例启动相关联的最近数据(例如,过去的秒、分钟或小时)或历史数据(例如,过去的天、周或月)。

[0044] 从数据仓库224检索的特征数据可与被确定为对计算实例的启动时间具有影响的启动特征208匹配。说明性地,可进行分析以确定哪些启动特征208影响启动时间且然后可构建从数据仓库224选择启动特征208的特征数据的查询。在一些示例中,当特征数据可能较大或冗余时,启动特征208的特征数据可被处理并汇总。例如,特征数据可被处理为启动特征的缩减表示集(例如,特征向量)。在获得启动特征208之后,然后可使用启动特征208训练机器学习模型。

[0045] 如前所述,机器学习模型可最初使用历史数据来训练,且然后置于生产中,其中机器学习模型可根据按需提供预测启动时间。训练模块222可被配置为从数据仓库224获得启动特征208的历史数据,并将历史数据提供至机器学习模型。历史数据可用于初始训练机器学习模型。可通过使机器学习模型停产(例如离线)并使用历史数据(例如,来自先前天、周、月等的数据)训练机器学习模型来进行机器学习模型的后续训练。或者,可在机器学习模型使用最近数据(例如,来自先前分钟、小时、天等的数据)投入生产(例如,在线)的同时进行后续训练。

[0046] 启动特征模块220可被配置为获得与预测启动时间的请求相关联的启动特征208。然后可将所获得的启动特征208提供为至机器学习模型的输入。作为说明,用于(例如,经由客户装置228)启动计算实例的请求可由用于计算服务202的控制平面240接收。启动请求可以是针对单个计算实例或任何数量的计算实例(例如,数十、数百或数千计算实例)。在接收到启动请求时,可为计算实例确定启动配置,该启动配置尤其指定机器映像特征、物理主机特征和客户配置特征(例如,存储装置、网络类型、地理区域等)。启动配置(或对启动配置的引用)然后可包括在预测启动时间的请求中。

[0047] 在服务器204接收到预测启动时间的请求时,启动配置可被提供至启动特征模块220,在该模块上,可评估启动配置,并且可收集对应于启动配置的启动特征208的数据。然后基于启动配置的规范,可获得启动特征208的数据。

[0048] 针对启动特征208收集的数据可被提供至启动时间预测模块218并且输入至机器学习模型。启动时间预测模块218然后可经由通过评估提供至启动时间预测模块218的启动特征208来确定预测启动时间的算法来生成预测启动时间。作为一个示例,由启动时间预测模块218使用的机器学习模型可包括多个决策树,其中启动特征208被输入至决策树中,并且使用回归,从决策树的输出计算预测启动时间。然后由机器学习模型生成的预测启动时间可用于与如前所述的计算服务202相关联的各种目的。

[0049] 包括在系统200中的物理主机236可以是被配置为执行实例管理器(即,管理程序、虚拟机监视器(VMM)或另一类型的程序)的服务器计算机,其管理单个物理主机236上的多个计算实例。物理主机236可位于各个地理区域210内的数据中心中。因此,计算实例的启动时间可基于被选择为托管计算实例的物理主机236的地理区域210而受影响。此外,启动时间可受物理主机236的其它属性(诸如架构、品牌等)影响。

[0050] 机器映像216可以是可由实例管理器执行的预配置的虚拟机映像(例如,虚拟装置)。机器映像216可包括用于计算实例的机器可执行包,其可包括操作系统、应用服务器和各种应用,其中任何一个都可影响计算实例的启动时间。此外,机器映像216可包括在启动计算实例时附接至对应计算实例的存储卷的映射。

[0051] 说明性地,机器映像216可存储在块级存储卷或网络文件存储服务中。机器映像216的存储位置可影响计算实例的启动时间。例如,当将机器映像216存储在网络文件存储服务中时,机器映像216可被压缩以便于通过网络传送机器映像216。因此,在将机器映像216传送至被选择来托管计算实例的物理主机236之后,解压缩机器映像216的进一步操作可增加计算实例的启动时间。

[0052] 包含在系统200内的各种过程和/或其它功能可在与一个或多个存储器模块232通信的一个或多个处理器230上执行。系统200可包括多个计算装置(例如,物理主机236和服务器204),其布置在例如一个或多个服务器组或计算机组或其它装置中。

[0053] 术语“数据存储区”可指能够存储、访问、组织和/或检索数据的任何装置或装置的组合,其可包括任何集中式、分布式或集群环境中的任何组合和数量的数据服务器、关系数据库、面向对象的数据库、集群存储系统、数据存储装置、数据仓库、平面文件和数据存储配置。数据存储区的存储系统组件可包括存储系统诸如SAN(存储区域网络)、云存储网络、易失性或非易失性RAM、光学介质或硬盘驱动型介质。如可理解,数据存储区可表示多个数据存储区。

[0054] 在一些示例中,客户可利用客户装置228来请求启动计算实例且然后访问计算实例。客户装置228可包括能够通过网络226发送并接收数据的任何装置。客户装置228可包括例如基于处理器的系统,诸如计算装置。

[0055] 网络226可包括任何有用的计算网络,包括内联网、因特网、局域网、广域网、无线数据网络或任何其它这样的网络或其组合。用于这样系统的组件可至少部分地取决于所选的网络和/或环境的类型。通过网络进行的通信可通过有线或无线连接及其组合来实现。

[0056] 图2示出可结合该技术讨论的某些处理模块且这些处理模块可被实现为计算服务。在一个示例配置中,模块可被认为是具有在服务器或其它计算机硬件上执行的一个或多个过程的服务。这样的服务可以是集中托管的功能或可接收请求并向其它服务或消费者装置提供输出的服务应用。例如,提供服务的模块可被认为是托管在服务器、虚拟化服务环境、网格或集群计算系统中的按需计算。可为每个模块提供API,以使得第二模块能够向第一模块发送请求并从第一模块接收输出。这样的API还可允许第三方与模块对接并做出请求并从模块接收输出。虽然图2示出可实现上述技术的系统的示例,但是许多其它类似或不同环境也是可能的。上面讨论和示出的示例环境仅仅是代表性的而不是限制性的。

[0057] 图3是示出可用于执行并管理多个计算实例304a-d的示例计算服务环境300的框图。具体地,所描述的计算服务环境300示出可使用本文所描述的技术的一个环境。计算服

务环境300可以是包括可用于例如托管计算实例304a-d的各种虚拟化服务资源的一种类型的环境。

[0058] 计算服务环境300可能够将计算、存储和网络能力作为软件服务传递至最终接收者的社区。在一个示例中，可由组织或代表组织为组织建立计算服务环境300。即，计算服务环境300可提供“私有云环境”。在另一示例中，计算服务环境300可支持多租户环境，其中多个客户可独立地操作(即，公共云环境)。一般来说，计算服务环境300可提供以下模型：基础设施即服务(“IaaS”)、平台即服务(“PaaS”)和/或软件即服务(“SaaS”)。可提供其它模型。对于IaaS模型，计算服务环境300可提供计算机作为物理机或虚拟机和其它资源。虚拟机可由管理程序作为客户机运行，如下面进一步描述。PaaS模型提供可包括操作系统、编程语言执行环境、数据库和web服务器的计算平台。

[0059] 应用开发商可在计算服务平台上开发并运行它们的软件解决方案，而不会引起购买和管理基础硬件和软件的成本。SaaS模型允许在计算服务环境300中安装并操作应用软件。例如，最终客户可使用联网的客户端装置(诸如运行web浏览器或其它轻量级客户端应用的台式计算机、膝上型计算机、平板计算机、智能手机等)访问计算服务环境300。熟悉本领域的人员应认识到，计算服务环境300可被描述为“云”环境。

[0060] 具体示出的计算服务环境300可包括多个物理主机302a-d。虽然示出了四个物理主机，但是可使用任何数量，且大数据中心可包括数千个物理主机302a-d。计算服务环境300可提供用于执行计算实例304a-d的计算资源。计算实例304a-d可以例如是虚拟机。虚拟机可以是类似物理机器执行应用的机器(即，计算机)的软件实现的实例。在虚拟机的示例中，物理主机302a-d中的每个都可被配置为执行能够执行实例的实例管理器308a-d。实例管理器308a-d可以是管理程序、虚拟机监视器(VMM)或被配置为使得能够在单个物理主机上执行多个计算实例304a-d的另一类型的程序。此外，计算实例304a-d中的每个都可被配置为执行一个或多个应用。

[0061] 一个或多个服务器计算机314和316可被保留以执行用于管理计算服务环境300和计算实例304a-d的操作的软件组件。例如，服务器计算机314可执行可响应对在物理主机302a-d上启动的计算实例的预测启动时间的请求的预测启动时间服务。

[0062] 服务器计算机316可执行管理组件318。客户可访问管理组件318以配置由客户购买的计算实例304a-d的操作的各个方面。例如，客户可建立计算实例304a-d并对计算实例304a-d的配置作出改变。

[0063] 部署组件322可用于辅助客户进行计算实例304a-d的部署。部署组件322可对与计算实例304a-d相关联的帐户信息(诸如帐户所有者的名称、信用卡信息、所有者的国家等)具有访问权。部署组件322可从客户接收包括描述可如何配置计算实例304a-d的数据的配置。例如，配置可包括操作系统、提供要安装在计算实例304a-d中的一个或多个应用、提供要执行以用于配置计算实例304a-d的脚本和/或其它类型的代码、提供指定应如何准备应用缓存的缓存逻辑以及其它类型的信息。部署组件322可利用客户提供的配置和缓存逻辑来配置、初始化并启动计算实例304a-d。配置、缓存逻辑和其它信息可由访问管理组件318的客户指定或者通过将该信息直接提供至部署组件322来指定。

[0064] 客户账户信息324可包括与多租户环境的客户相关联的任何期望信息。例如，客户账户信息可包括客户的唯一标识符、客户地址、计费信息、许可信息、用于启动实例的定制

参数、调度信息等。如上所述，客户账户信息324还可包括用于对API请求的异步响应的加密的安全信息。“异步”意味着API响应可在初始请求之后的任何时间并利用不同网络连接作出。

[0065] 网络310可用于互连计算服务环境300、物理主机302a-d和服务器计算机316。网络310可以是局域网(LAN)，并且可连接至广域网(WAN)312或因特网，使得终端客户可访问计算服务环境300。虽然图3中所示的网络拓扑结构已经被简化，但是可利用更多的网络和网络装置来互连本文公开的各种计算系统。

[0066] 现在转到图4，图示出用于配置并训练用于生成预测启动时间的机器学习模型416的示例方法400。如在框406中，可通过分析各种计算实例启动来进行启动特征选择，以确定对计算实例启动时间具有影响的启动特征。例如，可识别在其中能够观察到特征的计算服务环境内的物理主机上启动计算实例的各种特征。

[0067] 启动特征的示例可包括但不限于：物理主机上的多个竞争计算实例、物理主机上的多个运行计算实例、包含用于创建计算实例的机器映像的数据存储类型、由计算实例使用的内核映像、物理主机的架构、计算实例的虚拟化类型、物理主机能够托管的最大数量的计算实例、在计算实例启动的开始时计算实例对物理主机的占用的百分比、物理主机所位于的地理区域、物理主机的硬件类型、物理主机的硬件供应商以及计算实例的操作系统、网络类型、数据存储和大小。

[0068] 可对被确定为对计算实例的启动时间具有影响的启动特征进行分类。例如，启动特征的类别可基于计算实例启动的各个方面。作为说明，启动特征可被分类为机器映像启动特征、物理主机启动特征和客户配置启动特征。

[0069] 在一个示例中，可根据启动特征对计算实例启动时间的影响来对所识别的启动特征进行排序或排名，且可选择对启动时间具有最大影响的那些启动特征作为要用于预测启动时间的特征。例如，可分析启动特征以确定各个启动特征对启动时间具有的贡献的百分比。被识别为对启动时间具有最大贡献的启动特征可被选择为至机器学习模型的输入。应注意，可选择任何数量的启动特征，并且启动特征的选择可不限于仅对启动时间具有最大影响的那些启动特征。

[0070] 在识别了启动特征之后，然后可从包含与启动特征相关联的数据的数据源获得启动特征的启动特征数据402。如图所示，可从包含例如计算服务管理数据、库存数据(例如，物理主机信息)以及与计算服务相关联的其它数据的数据存储区获得启动特征数据402。启动特征数据402可被标准化为使得从不同数据源获得的启动特征数据402被输入至机器学习模型416中。启动特征数据402可被划分为训练数据410、交叉验证数据412和测试数据414。例如，启动特征数据402的百分比可被随机选择为测试数据414和交叉验证数据412，且剩余的启动特征数据402可用作训练数据410来训练机器学习模型416。

[0071] 可从任何可用的机器学习算法中选择机器学习模型416。在一个示例中，可测试多个回归机器学习模型来确定提供启动时间的可接受近似的机器学习模型。如在框408中，生成机器学习模型的一个方面可以是进行机器学习参数的参数值搜索，其引起机器学习模型416对启动特征的拟合优度(goodness-of-fit)。机器学习参数(即，用于配置机器学习模型416的参数，诸如设置决策树的深度)可影响机器学习模型416如何拟合训练数据410。在一个示例中，网格搜索或梯度下降算法可用于进行参数值搜索。在另一示例中，当机器学习模

型416的参数空间可能太大而不能进行彻底的参数值搜索时，可使用进化算法(例如，分布式遗传算法)、群算法(例如，粒子群优化)、模拟退火等算法。

[0072] 在选择机器学习模型416之后，可使用训练数据410训练机器学习模型416。然后，交叉验证数据412和测试数据414可通过机器学习模型416运行，以测试机器学习模型的输出是否表示额外的历史情况。此后，如在框418中，可进行数据分析以确定机器学习模型416能够有多好地预测启动时间(与实际启动时间相比)。在测试两个或更多个机器学习模型416之后，如在框420中，可比较机器学习模型416的结果以识别更好进行的机器学习模型416，然后可选择该模型并且将其置于生产环境中。

[0073] 图5是示出可使用预测启动时间的方法500的一个示例的流程图。所示出的示例方法500用于使用预测启动时间来预测可能违反SLA启动时间。在一个示例中，SLA启动时间可以是计算服务提供商已同意作为服务合同的一部分提供的计算实例的启动时间。这样，计算服务提供商可能希望被通知在实际违反SLA启动时间之前可能将违反SLA，从而允许计算服务提供商相应地采取行动。

[0074] 开始于框502，可接收请求在计算服务内启动一个或多个计算实例的启动请求。例如，可由希望在计算服务环境内启动一个计算实例或一组计算实例的客户作出请求。在接收到启动请求时，启动服务可识别要启动的一个或多个计算实例的启动配置。

[0075] 如在框504中，可识别与作出启动请求的客户相关联的SLA。其中，SLA可指定计算实例的SLA启动时间。说明性地，SLA启动时间可以是从计算服务接收启动请求到计算实例正在运行的时间(例如，开始引导过程)之间的时间。因此，作出启动请求的客户可预期计算实例将在SLA启动时间内准备就绪。

[0076] 在接收到启动请求并识别启动配置和SLA启动时间之后，如在框506中，可获得计算实例的预测启动时间。例如，可对预测启动时间服务作出请求，该预测启动时间服务生成如前所述的预测启动时间。作为说明，对预测启动时间的请求可包括用于一个或多个计算实例的启动配置或对启动配置的引用。预测启动时间服务然后可通过至少部分地基于启动配置来识别启动特征并将启动特征输入至输出预测启动时间的机器学习模型中来生成一个或多个计算实例的预测启动时间。

[0077] 如在框508中，然后可比较预测启动时间与SLA时间，以确定(如在框510中)预测启动时间是否大于SLA启动时间。预测启动时间和SLA启动时间的比较可提供可能实现还是违反SLA启动时间的指示。

[0078] 在预测启动时间不大于SLA启动时间的情况下，如在框514中，可启动一个或多个计算实例。在预测启动时间大于SLA时间的情况下，如在框512中，可响应于潜在SLA启动时间违反进行预定动作。预定动作的一个示例可包括通知计算服务操作者和/或客户SLA启动时间可能不会实现。这样，计算服务操作者和/或客户可通过进行可增加启动时间的动作来尝试减少或防止可能的SLA启动时间违反。例如，计算服务提供商可从提供计算能力的一组物理主机中移除引起增加启动时间的物理主机。可替代地或附加地，可建议客户修改可在客户控制之内的计算实例的启动配置的那些方面，以及计算服务操作者和/或客户可进行的其它动作(这里没有具体描述)。

[0079] 在一个示例配置中，在确定将可能会违反SLA启动时间时，计算过程可分析计算实例要在其中启动的计算服务环境的状态，以确定是否可进行防止违反SLA启动时间的动作。

作为可进行的动作的一个示例，可分析可用计算能力以确定添加增加计算能力的附加物理主机是否可增加启动时间。例如，一组物理主机可提供可用计算能力来托管多个计算实例。可分析该组物理主机以确定该组物理主机能够托管多少计算实例并确定该组物理主机当前托管多少计算实例(例如，运行计算实例)。基于分析的结果，可向该组物理主机添加附加物理主机以增加可用计算能力。

[0080] 作为可响应于可能的SLA启动时间违反而进行的动作的另一示例，可分析包括在提供计算能力的一组物理主机中的各个物理主机，以确定物理主机是否可负面地影响启动时间。作为具体示例，包括在一组物理主机中的过载物理主机可由于在过载物理主机上同时启动多个计算实例而影响启动时间。例如，过载物理主机可看起来具有用于托管计算实例的可用计算能力，但是由于过载物理主机正在处理的计算实例启动的数量，计算实例在过载物理主机上的启动时间可超过SLA启动时间。这样，可从被认为可用于托管计算实例的该组物理主机中移除过载物理主机。具体地，在生成计算实例的第二预测启动时间(例如，因为第一预测启动时间包括过载物理主机)之前，可从可用计算能力中移除过载物理主机。然后可生成第二预测启动时间，这与基于包括过载物理主机的可用计算能力的第一预测启动时间相比可引起更快启动时间的预测。

[0081] 作为可响应于可能SLA启动时间违反而进行的动作的又另一示例，可分析计算实例的启动配置，以确定对启动配置的改变是否可引起增加启动时间。作为说明，启动配置可指定用于启动计算实例的参数和计算资源。这些参数和计算资源可影响计算实例的预测启动时间。这样，可分析启动配置以确定对启动配置的改变是否可引起不违反SLA启动时间的预测启动时间。作为具体示例，启动配置可指定在其中启动计算实例的地理区域。可进行分析以确定在不同地理区域中启动计算实例是否将引起更好预测启动时间。在分析确定不同地理区域可引起更好预测启动时间的情况下，启动配置可被修改为包括不同地理区域。

[0082] 作为上述操作的替代或附加，可提供表示SLA启动时间违反的特征(例如，SLA违反特征)作为至机器学习分类模型的输入，该机器学习分类模型输出指示计算实例启动是否可违反SLA启动时间的分类。例如，SLA违反特征可与其它特征一起被考虑提供至机器学习分类模型。使用算法(例如，分类器)，提供至机器学习模型的输入特征数据可被映射到类别。因此，在预测启动时间特征可大于SLA启动时间特征的示例中，机器学习分类模型可输出指示计算实例的启动时间将可能违反SLA启动时间的分类。

[0083] 图6是示出用于预测计算实例的启动时间的示例方法600的流程图。开始于框610，可接收与在计算服务环境内的物理主机上启动计算实例相关联的预测启动时间的请求。预测启动时间可以是从计算实例处于待决状态(即，执行服务调用以建立计算实例资源、识别托管计算实例的物理主机并在物理主机上创建计算实例)的时间到计算实例处于执行状态(即，引导计算实例的开始)的时间。在一些示例中，客户接收可用计算实例(例如，引导的计算实例)的时间可通过包括计算实例的引导时间而被包括在预测启动时间中，这可受到计算实例的内部配置的影响。

[0084] 如在框620中，可获得与被确定为对计算实例在计算服务环境内的物理主机上的启动时间具有影响的计算实例的启动特征相关联的数据。例如，可被确定为对启动时间具有影响的启动特征可包括但不限于：机器映像启动特征(例如，用于创建计算实例的机器映像的特征)、物理主机启动特征(例如，被选择为托管计算实例的物理主机的特征)和可由客

户控制的启动配置特征(例如,机器映像配置、地理区域、同时启动的多个计算实例等)。在一个示例中,在获得与启动特征相关联的数据之后,然后可对数据进行标准化。

[0085] 如在框630中,启动特征(即,启动特征的数据)可被输入至机器学习模型,该机器学习模型输出用于在计算服务环境内的所选物理主机上启动计算实例的预测启动时间。在一个示例中,机器学习模型可是回归模型(例如,随机森林回归模型)。

[0086] 可使用历史数据来训练机器学习模型,且然后将其置于生产环境中,其中机器学习模型接收对预测启动时间的活动请求。在一个示例配置中,可使用历史数据(例如,先前的天、周或月启动特征)来周期性地训练机器学习模型。在另一示例配置中,机器学习模型可通过在机器学习模型处于生产环境中时从活动数据(例如,先前的秒、分钟、小时)中提取启动特征并重新训练机器学习模型来训练,从而使机器学习模型适应于在计算服务内发生的变化。

[0087] 然后可响应于该请求提供由机器学习模型生成的预测启动时间。作为一个示例,预测启动时间可被提供至计算服务内的各种服务,诸如为计算实例选择物理主机的计算实例放置服务。作为另一示例,预测启动时间可被提供至客户,从而通知客户预测启动时间,或者通知客户是否可能实现SLA启动时间。作为又另一示例,预测启动时间可被提供至计算服务操作者,从而允许计算服务操作者根据预测启动时间来分析并修改计算服务环境。如应理解,预测启动时间可用于任何目的并且因此不限于本文公开的示例。描述了用于使用估计启动时间来确定计算实例在计算服务环境中的物理主机内的放置的技术。在计算实例的放置期间,可使用估计启动时间和其它放置准则来在计算服务环境中识别具有可用计算槽(例如,用于启动计算实例的计算资源)的物理主机或提供减少启动时间的一组物理主机。计算实例可放置在物理主机(也称为服务器计算机)上,并且计算实例可在计算服务环境内的物理主机上启动或执行。

[0088] 在一个示例中,可接收在计算服务环境中启动计算实例的请求。可从期望来自计算服务环境的计算服务的客户接收请求。可在从客户接收到启动计算实例的请求时进行关于计算服务环境中的哪个物理主机要为计算实例提供放置的确定。例如,与计算服务环境中的其它物理主机相比,可识别为计算实例提供减少启动时间的物理主机,并且可在该物理主机上启动计算实例。因此,计算实例可被放置在物理主机上,以为计算实例提供减少启动时间。术语“启动时间”通常可指1)接收启动计算实例的请求和2)在被选择来启动计算实例的物理主机上引导计算实例之间的时间段。

[0089] 在一个配置中,当确定计算实例或可从其生成计算实例的机器映像的放置时,可识别与请求中涉及的计算实例相关联的实例特征。实例特征可描述或表征计算实例。例如,实例特征可包括但不限于计算实例的大小、计算实例使用的计算实例映像类型(例如,机器映像或内核映像)、计算实例的架构类型(例如,32位架构或64位架构)、计算实例的虚拟化类型(例如,半虚拟化或硬件虚拟机)以及计算实例使用的数据存储区的类型。实例特征可包括用户控制的特征,诸如用于启动计算实例的操作系统(OS)的类型和网络类型(例如,虚拟专用网络)。

[0090] 在一个配置中,当确定计算实例的放置时,可识别与计算服务环境中的物理主机相关联的物理主机特征。物理主机特征可在给定时间(例如,当计算实例将被启动时)描述或表征计算服务环境中的物理主机的各方面。或者,物理主机特征可描述计算服务环境中

的所定义的一组物理主机。物理主机特征可包括但不限于可托管在物理主机处的最大数量的计算实例、与物理主机相关联的硬件类型、与物理主机相关联的硬件供应商、当计算实例被启动时计算实例在物理主机处的占用百分比以及物理主机所位于的区域。此外，物理主机特征可包括当前在物理主机上待决和/或运行的多个计算实例。

[0091] 与计算实例相关联的实例特征和与计算环境中的物理主机相关联的物理主机特征可被提供至启动时间预测模型。启动时间预测模型可使用实例特征和物理主机特征来预测在计算服务环境中的物理主机上启动计算实例的估计启动时间。更具体地，启动时间预测模型可预测用于在物理主机A上启动计算实例的估计启动时间、用于在物理主机B上启动计算实例的估计启动时间、用于在物理主机C上启动计算实例的估计启动时间等等。启动时间预测模型可以是已经使用历史启动时间信息和用于多个先前启动的计算实例的特征来训练的机器学习模型(例如，回归模型)，以便预测要在计算服务环境中启动的计算实例的估计启动时间。

[0092] 作为实例特征的非限制性示例，要在计算服务环境中启动的计算实例可以是：大小相对较小、包括32位架构、使用硬件虚拟机(HVM)和/或使用限定类型的数据存储区。计算实例可在物理主机A、物理主机B或物理主机C上启动。物理主机A可被占用80％(即，当前正在使用物理主机A的计算资源的80％)，并且当前正在启动十个其它计算实例。物理主机B可被占用50％，并且当前正在启动六个其它计算实例。物理主机C可被占用20％，并且当前正在启动两个其它计算实例。启动时间预测模型可接收所识别特征，并且确定在物理主机A上启动计算实例的估计启动时间是70秒，在物理主机B上启动计算实例的估计启动时间是40秒，且在物理主机C上启动计算实例的估计启动时间是15秒。因此，当确定哪个物理主机要为计算实例提供放置时，可考虑估计启动时间。

[0093] 在上面的示例中，与其它物理主机相比，可提供减少启动时间的物理主机可被选择用于计算实例的放置。在上述示例中，因为与物理主机A和物理主机B相比，物理主机C可提供减少启动时间，所以可选择物理主机C来启动计算实例。

[0094] 在替代配置中，因为在物理主机上同时启动的数量可增加用于启动计算实例的启动时间，所以为计算实例的放置而选择的物理主机可具有同时被启动的最少数量的计算实例(与计算服务环境中的其它物理主机相比)。作为非限制性示例，在正在进行放置决策时，物理主机A可正启动十个计算实例，物理主机B可正启动两个计算实例，且物理主机C可正启动一百个计算实例。因此，因为可推断物理主机B提供最少启动时间(与物理主机A和物理主机C相比)，所以可选择物理主机B来启动计算实例。

[0095] 计算实例的估计启动时间可以是在确定计算实例的放置时使用的多个放置因素中的一个。例如，与计算实例的放置相关的其它因素可包括物理主机利用率、许可成本、灾难影响等。可向每个放置因素(包括估计启动时间)分配与放置因素的重要性级别相关的加权值。例如，估计启动时间可占据放置决策的50％，物理主机利用率可占据放置决策的30％，许可成本可占据放置决策的20％，且灾难影响可占据放置决策的10％。

[0096] 图7示出可执行本技术的模块的计算装置710。示出了可执行本技术的高级示例的计算装置710。计算装置710可包括与多个存储器装置720通信的一个或多个处理器712。计算装置710可包括用于计算装置中的组件的本地通信接口718。例如，本地通信接口718可以是本地数据总线和/或可能需要的任何相关地址或控制总线。

[0097] 存储器装置720可包含可由处理器(多个)712执行的模块724和用于模块724的数据。例如,存储器装置720可包含训练模块和启动特征模块。模块724可执行先前描述的功能。数据存储区722还可位于存储器装置720中,以用于存储与模块724和其它应用相关的数据以及可由处理器712执行的操作系统。

[0098] 其它应用也可存储在存储器装置720中,并且可由处理器(多个)712执行。在本具体实施方式中讨论的组件或模块可使用高编程级语言以软件的形式实现,使用这些方法的混合来编译、解释或执行。

[0099] 计算装置还可访问可由计算装置使用的I/O(输入/输出)装置714。网络装置716和类似通信装置可包括在计算装置中。网络装置716可以是连接至因特网、LAN、WAN或其它计算网络的有线或无线网络装置。

[0100] 被示为存储在存储器装置720中的组件或模块可由处理器(多个)712执行。术语“可执行的”可意味着程序文件,该程序文件是可由处理器712执行的形式。例如,较高级语言的程序可被编译为可被加载至存储器装置720的随机存取部分中并由处理器712执行的格式的机器代码,或者可由另一可执行程序加载并且被解释为在由处理器执行的存储器的随机存取部分中生成指令的源代码。可执行程序可存储在存储器装置720的任何部分或组件中。例如,存储器装置720可以是随机存取存储器(RAM)、只读存储器(ROM)、闪存、固态驱动器、存储卡、硬盘驱动器、光盘、软盘、磁带或任何其它存储器组件。

[0101] 处理器712可表示多个处理器,并且存储器720可表示与处理电路并行操作的多个存储器单元。这可为系统中的过程和数据提供并行处理通道。本地接口718可用作网络以便于多个处理器和多个存储器中的任何之间的通信。本地接口718可使用被设计为用于协调通信的附加系统,诸如负载平衡、批量数据传输和类似系统。

[0102] 图8示出根据本技术的一个示例的示例计算服务环境800的组件。计算服务环境800可包括经由网络850与多个客户端装置860通信的服务器计算机810,并且服务器计算机可以是用于服务提供商环境800的控制平面的一部分。服务器计算机210可包含数据存储区830和用于确定计算实例的放置的多个模块。此外,计算服务环境800可包括执行多个计算实例的多个服务器计算机840a-c。

[0103] 服务器计算机840a-c可具有可用于执行计算实例的可用计算槽842a-c(例如,空闲计算资源)。可用计算槽842a-c可被分配至其可随后利用可用计算槽842a-c来执行计算实例的客户。计算实例的示例可包括按需计算实例、保留计算实例和可中断计算实例。按需计算实例可以是客户可根据请求购买并执行的计算实例。保留计算实例可以是以下计算实例的保留:客户可在限定的时间段内购买,使得当客户请求计算实例时计算实例可用,且可中断计算实例可以是以下计算实例:其可在计算槽842a-c中执行且不被另一计算实例类型使用,除非对于可中断计算实例支付的价格低于当前投标价格。

[0104] 存储在数据存储区830中的数据可包括实例特征832。实例特征832可与要在计算服务环境800中启动的计算实例相关联。此外,实例特征832可与从其的计算实例在计算服务环境800中启动的计算实例映像相关联。实例特征832可描述或表征要在计算服务环境800中启动的计算实例。例如,实例特征可具有数值或其它标量值。

[0105] 存储在数据存储区830中的数据可包括物理主机特征834。物理主机特征834可与计算服务环境800中的多个物理主机相关联。物理主机特征834可描述或表征潜在地可启动

计算实例的计算服务环境800中的物理主机。例如，物理主机特征834可具有数值或其它标量值。

[0106] 存储在数据存储区830中的数据可包括估计启动时间836。估计启动时间836可用于要在计算服务环境800中启动的多个计算实例。估计启动时间836可为给定计算实例指示估计启动时间，以在计算服务环境800中的多个物理主机中的每个物理主机(或服务器计算机)上启动计算实例。估计启动时间836可存储在数据存储区830中以用于质量控制目的、记录保存或其它用途。可使用启动时间预测模型来确定估计启动时间836。在一个示例中，当确定估计启动时间836时，启动时间预测模型可使用与计算实例相关联的实例特征832和与多个物理主机相关联的物理主机特征834。作为非限制性示例，用于在三个不同物理主机上启动计算实例的估计启动时间836可分别是10秒、50秒或两分钟。

[0107] 服务器计算机810可包括计算实例请求模块822、估计启动时间预测模块824、物理主机选择模块826以及本文未详细讨论的其它应用、服务、过程、系统、引擎或功能。计算实例请求模块822可被配置为接收在计算服务环境800中启动一个或多个计算实例的请求。可从期望来自计算服务环境800的计算服务的客户接收请求。请求可包括要启动的多个计算实例以及要启动的计算实例的类型或大小。在一个示例中，请求可指定特定地理区域或区以启动计算实例。

[0108] 估计启动时间预测模块824可被配置为接收或识别与请求中的计算实例相关联的实例特征和与计算服务环境800中的多个物理主机相关联的物理主机特征。实例特征可包括计算实例的大小、用于启动计算实例的机器映像、计算实例的架构类型、计算实例的虚拟化类型、由计算实例使用的数据存储区的类型等。此外，对于计算服务环境800中的每个物理主机或对于计算服务环境800中的限定组的物理主机，物理主机特征可包括可托管在物理主机上的最大数量的计算实例、硬件类型、硬件供应商、占用百分比、物理主机所位于的地理区域、当前在物理主机上待决或运行的多个实例等。

[0109] 估计启动时间预测模块824可识别用于在计算服务环境800中启动计算实例的估计启动时间。估计启动时间预测模块824可使用机器学习模型来识别估计启动时间。给定实例特征和物理主机特征，机器学习模型可预测估计启动时间。实例特征和物理主机特征可影响计算实例的估计启动时间。例如，某些实例特征和/或物理主机特征(例如，物理主机上的多个同时计算实例启动、计算实例的大小)可增加计算实例的估计启动时间，而其它实例特征和/或物理主机特征可减少计算实例的估计启动时间。在一个示例中，机器学习模型可以是回归模型，其使用多个先前启动的计算实例的历史启动时间信息来预测要在计算服务环境800中启动的计算实例的估计启动时间。

[0110] 物理主机选择模块826可被配置为从可提供计算实例的放置的计算服务环境800中的一组物理主机中选择物理主机。物理主机选择模块826可基于计算实例的估计启动时间来选择物理主机。在一个示例中，物理主机选择模块826可选择可以减少估计启动时间或最低估计启动时间(与该组物理主机中的其它物理主机相比)提供计算实例的放置的物理主机。此外，当确定计算实例的放置时，物理主机选择模块826可使用附加放置因素。附加因素可包括但不限于物理主机利用率、许可成本和灾害影响。当确定计算实例的放置时，可向估计启动时间和附加放置因素每个分配与放置因素的重要性级别相关的加权值。计算实例可在放置时加载在物理主机上并在物理主机上执行，以便向客户提供计算服务。

[0111] 图9示出用于将计算实例放置在计算服务环境900中的物理主机上的示例性系统和相关操作。可启动计算实例以便在放置到物理主机上时提供计算服务。可在计算服务环境900处接收用于启动计算实例的计算实例请求910。例如，客户可进行计算实例请求910，以便从计算服务环境900获得计算服务。可根据预定目标来选择将计算实例放置在其上的物理主机。预定目标可由客户和/或计算服务环境900来限定。在一个示例中，预定目标可包括将计算实例放置在物理主机上，该物理主机可提供最快启动时间(与该计算服务环境900中的其它物理主机相比)。

[0112] 可识别与包括在计算实例请求910中的即将被启动的计算实例相关联的实例特征915和计算服务环境900中的多个物理主机的物理主机特征920。例如，物理主机950-960可被直接查询对应于物理主机特征920的数据。实例特征915和物理主机特征920可分别描述计算服务环境900中的计算实例和物理主机950-960。实例特征915和物理主机特征920的识别可使得计算实例能够放置到物理主机上。更具体地，被选择用于放置的物理主机可取决于实例特征915和物理主机特征920。如前所述，实例特征915可包括计算实例的大小、由计算实例使用的机器映像、计算实例的架构类型、计算实例的虚拟化类型、由计算实例使用的数据存储区的类型等。此外，对于计算服务环境900中的每个物理主机或对于计算服务环境900中的所限定组的物理主机，物理主机特征920可包括可托管在物理主机上的最大数量的计算实例、硬件类型、硬件供应商、占用百分比、物理主机所位于的地理区域、当前在物理主机上待决或运行的多个实例等。

[0113] 实例特征915和物理主机特征920可被提供至机器学习模型930。机器学习模型930可以是回归模型，其基于实例特征915和物理主机特征920预测用于在给定物理主机上启动计算实例的估计启动时间。机器学习模型930可使用先前启动的计算实例的历史信息(例如，先前启动的计算实例的类型、计算实例的启动时间、在启动了计算实例的物理主机上同时启动的数量等)来训练，以便预测计算实例的估计启动时间。

[0114] 在一个示例中，机器学习模型930可预测用于在计算服务环境900中的每个可用物理主机上启动计算实例的估计启动时间。或者，机器学习模型930可预测用于在所限定组的物理主机中的各个物理主机上启动计算实例的估计启动时间。例如，机器学习模型930可预测用于在物理主机950上启动计算实例的估计启动时间，以及用于在物理主机960上启动计算实例的估计启动时间。

[0115] 机器学习模型930可向放置模块940提供估计启动时间。当对计算实例进行放置决策(即，哪个物理主机要托管或启动计算实例)时，放置模块940可使用估计启动时间。此外，当进行放置决策时，放置模块940可使用附加放置因素935。附加放置因素935可包括但不限于物理主机利用率放置因素、许可成本放置因素和灾害影响放置因素。物理主机利用率放置因素可表示使包括在计算服务环境900中的物理主机之中的物理主机利用率最大化的预定目标。许可放置因素可表示使与在包括在计算服务环境900中的物理主机上放置计算实例相关联的软件许可成本最小化的预定目标。灾难影响放置因素可表示使计算服务故障(例如，物理主机故障、机架故障、可用区域故障或硬件故障)对客户的执行计算实例的影响最小化的预定目标。

[0116] 当确定计算实例的放置时，可向由机器学习模型930确定的估计启动时间以及附加放置因素935每个分配指示放置因素的各自重要性的加权值。换言之，可根据计算服务环

境900可如何受计算实例的放置的影响来将加权值分配至每个放置因素。例如，在期望维持高物理主机利用率的情况下，物理主机利用率因素可接收相对较高加权值。在优化软件许可成本对计算服务环境具有较小值的情况下，许可成本放置因素可接收较低加权值(与分配至利用率放置因素的加权值相比)。在计算实例放置在整体计算服务环境900中当前可能对由于系统故障而受影响的多个客户具有负面影响的情况下，分配至灾难影响放置因素的加权值可以是相对较高值。作为非限制性示例，当放置模块840进行放置决策时，计算实例的估计启动时间可被加权50％，物理主机利用率放置因素可被加权20％，许可成本放置因素可被加权15％，且灾害影响放置因素可被加权15％。

[0117] 放置模块940可从机器学习模型930接收用于启动计算实例的估计启动时间以及附加放置因素935。放置模块940可确定哪个物理主机要接收计算实例以便遵照计算服务环境900的预定目标。在一个示例中，放置模块940可选择为计算实例提供减少启动时间或最低启动时间的用于放置的物理主机。

[0118] 作为非限制性示例，放置模块940可确定物理主机950可在28秒内启动计算实例。此外，放置模块940可确定物理主机960可在30秒内启动计算实例。因为物理主机950提供较低启动时间，所以放置模块940可选择用于计算实例的放置的物理主机950。

[0119] 在一个配置中，可使用包括在计算实例请求910中的放置约束来选择用于提供计算实例的放置的物理主机。在一个示例中，请求要启动计算实例的客户可提供放置约束。放置约束可指示计算实例请求910是否是用于启动计算实例的集群的计划。放置约束可指示在启动计算实例时要使用的特定类型的硬件、操作系统或网络类型。此外，放置约束可指示计算实例是否将在相对彼此靠近的一组物理主机(与相对分散开的一组物理主机对照)中启动。

[0120] 图10示出当确定在计算服务环境1000中放置计算实例时使用估计附接时间的示例性系统和相关操作。可(例如，从客户)接收启动计算实例的请求。在一个示例中，请求可包括附接请求1010。附接请求1010可是在启动计算实例时将网络接口和/或网络存储装置附接至计算实例。由于在启动计算实例时使用的附接的数量和/或附接的大小可影响计算实例的启动时间，因此可在对计算实例进行放置决策时考虑附接时间。

[0121] 在一个示例中，机器学习模型1030可预测估计附接时间，即，进行附接的时间量。当预测估计附接时间时，机器学习模型1030可使用附接请求1010以及与附接请求1010相关联的附接特征1020。附接特征可包括但不限于包括在附接请求1010中的附接的数量、附接的大小、附接是否与数据存储装置或网络接口相关等。在一个示例中，机器学习模型1030可以是使用与过去附接请求相关的历史信息来预测估计附接时间的回归模型。

[0122] 机器学习模型1030可向放置模块1040提供估计附接时间。放置模块1040可基于估计附接时间来选择用于计算实例的放置的物理主机。例如，放置模块1040可根据估计附接时间将计算实例放置在物理主机1050、物理主机1060或物理主机1070中的一个上。在一个示例中，放置模块1040可选择可向计算实例的放置提供减少估计附接时间的物理主机。

[0123] 在另一配置中，客户可请求临时附接(即，在已经启动计算实例之后对附加存储的请求)。机器学习模型1030可基于临时附接请求的特性(诸如请求中的附加存储的大小等)来预测用于获得附加存储的估计时间量。换言之，机器学习模型1030可基于过去附加存储请求来确定用于获得附加存储的估计时间量。在一个示例中，可经由用户接口向客户提供

用于提供附加存储的估计时间量。

[0124] 图11示出用于将计算实例放置在从计算服务环境1100中的多个拓扑层1150中的至少一个中选择的物理或地理区域中的物理主机(例如,服务器)上的示例性系统和相关操作。可从客户接收启动计算实例的计算实例请求1110。与计算服务环境1100中的计算实例和物理主机相关联的特征1120可被提供至机器学习模型1130。机器学习模型1130可确定用于在变化拓扑层1150内的各个区域中的物理主机(诸如特定地理区域、区、数据中心、数据架、物理主机、计算槽等中的物理主机)上启动计算实例的估计启动时间。在一个示例中,地理区域可包括多个区,每个区都可包括多个数据中心,每个数据中心都可包括多个数据架,每个数据架都可包括多个物理主机,且每个物理主机都可包括多个计算槽。机器学习模型1130可确定将计算实例放置在特定拓扑层1150中的物理主机上是否可引起提高启动时间。例如,机器学习模型1130可指示将计算实例放置在特定区中的第一数据中心中的物理主机上可引起更快启动时间(与将计算实例放置在特定区中的第二数据中心中相比)。机器学习模型1130可将拓扑层1130的估计启动时间传送至放置模块1140。放置模块1140可在确定计算实例的放置时使用估计启动时间(即,选择哪个拓扑层1150来托管计算实例)。

[0125] 图12是示出生成机器学习模型1250以预测在计算服务环境中启动的计算实例的启动时间的示例性框图1200。可使用实际启动时间预测数据1210来创建机器学习模型1250。实际启动时间输入数据1210可包括先前已经在计算服务环境中启动的多个计算实例的信息(例如,启动度量)。因此,实际启动时间输入数据1210可包括与在计算服务环境中先前启动的计算实例相关的历史信息。此外,实际启动时间输入数据1210可包括计算服务环境中的多个物理主机的历史信息。实际启动时间输入数据1210可被变换以用于训练机器学习模型1250,如稍后所讨论。

[0126] 作为非限制性示例,实际启动时间输入数据1210可指示计算实例A花费60秒来启动,而计算实例A的大小相对较大、使用第一类型的数据存储区、使用32位架构,并在同时启动其它五个计算实例的物理主机上启动。作为另一非限制性示例,实际启动时间输入数据1210可指示计算实例B花费15秒来启动,而计算实例B的大小相对较小、使用第二类型的数据存储区、使用64位架构,并在未同时启动其它计算实例的物理主机上启动。

[0127] 实际启动时间输入数据1210可被提供至特征选择和标准化模块1220。特征选择和标准化模块1220可将实际启动时间输入数据1210转换为模型特征。换言之,模型特征可涉及先前启动的计算实例的特性以及先前在其上启动计算实例的物理主机的特性。模型特征可被分类为实例特征和物理主机特征。

[0128] 实例特征可包括但不限于计算实例的大小、计算实例使用的机器映像(例如,机器映像或内核映像)、计算实例的架构类型(例如,32位架构或64位架构)、计算实例的虚拟化类型(例如,半虚拟化或硬件虚拟机)以及由计算实例使用的数据存储区的类型。实例特征可包括用户控制的特征,诸如用于启动计算实例的操作系统(OS)的类型和网络类型(例如,虚拟私有云)。

[0129] 物理主机特征可包括但不限于物理主机可托管的最大数量的计算实例、与物理主机相关联的硬件类型、与物理主机相关联的硬件供应商、在计算实例要被启动时物理主机的占用百分比以及物理主机所位于的区。物理主机特征可包括在启动计算实例的物理主机(即,目标物理主机)上待决计算实例和/或运行计算实例的平均、最小和最大数量。此外,物

理主机特征可包括当前在启动计算实例的物理主机(即,目标物理主机)上处于待决状态和/或运行状态的多个计算实例。

[0130] 特征选择和标准化模块1220可对模型特征进行标准化(即,将在不同尺度上测量的值调整至标称的共同尺度),以便创建启动时间预测训练数据1230。启动时间预测训练数据1230可表示在启动时间预测输入数据1210中识别的多个计算实例的聚合特征。启动时间预测训练数据1230可被提供至机器学习选择模块1240。机器学习选择模块1240可使用启动时间预测训练数据1230来训练各种机器学习模型1242。例如,可训练回归模型。回归模型1242可包括但不限于支持向量机、随机梯度下降、自适应引导、附加树和随机森林。各种回归模型1242可对应于具有各种级别成功的启动时间预测训练数据1230。在一个示例中,随机森林回归量可相对于启动时间预测训练数据1230提供相对较高精确度,且因此,当估计计算实例的启动时间时,机器学习选择模块1240可使用随机森林回归量。

[0131] 机器学习模型1250可接收启动计算实例的请求,并且基于与计算实例相关联的实例特征和物理主机特征,机器学习模型1250可预测计算实例的启动时间。在一个示例中,机器学习模型1250可确定在同一物理主机上进行同时计算实例启动的数量、由计算实例使用的存储数据的类型、由计算实例使用的架构类型以及与计算实例相关联的计算实例映像可对计算实例的启动时间具有更大影响(与其它模型特征相比)。

[0132] 在一些情况下,来自计算实例的预测启动时间可偏离计算实例的实际启动时间。与计算实例相关联的实例特征和物理主机特征以及启动计算实例的实际启动时间可用于进一步训练机器学习模型1250,以便改进未来启动时间预测。

[0133] 图13是示出用于在确定计算服务环境内的计算实例放置的示例方法的流程图。可接收在计算服务环境中启动计算实例的请求,如在框1310中。可从期望来自计算服务环境的计算服务的客户接收启动计算实例的请求。

[0134] 可将与计算实例相关联的实例特征和与计算服务环境中的一组物理主机相关联的物理主机特征提供至机器学习模型,如在框1320中。实例特征可根据请求描述或表征要启动的计算实例。物理主机特征可在给定时间(即,当根据请求启动计算实例时)描述或表征计算服务环境中的每个物理主机。

[0135] 可使用机器学习模型来确定用于在计算服务环境中的各个物理主机上启动计算实例的估计启动时间,如在框1330中。给定实例特征和物理主机特征,机器学习模型可预测估计启动时间。在一个示例中,机器学习模型可以是回归模型,其使用多个先前启动的计算实例的历史启动时间信息来预测要在计算服务环境中启动的计算实例的估计启动时间。

[0136] 根据较低估计启动时间(与该组物理主机中的其它物理主机相比),来自该组物理主机的物理主机可被选择提供计算实例的放置,如在框1340中。此外,可使用包括在启动计算实例的请求中的放置约束来选择用于提供计算实例的放置的物理主机。估计启动时间可以是在选择用于放置计算实例的物理主机时使用的多个放置因素中的一个。在一个示例中,当使用多个放置因素确定放置时,可将加权值分配至计算实例的估计启动时间,并且可部分地基于分配至估计启动时间的加权值来选择提供计算实例的放置的物理主机。

[0137] 图14是示出用于确定在计算服务环境内的计算实例放置的另一示例方法的流程图。可接收在计算服务环境中启动计算实例的请求,如在框1410中。可从请求来自计算服务环境的计算服务的客户接收启动计算实例的请求。

[0138] 可识别在一组物理主机中的物理主机上启动计算实例的估计启动时间，如在框1420中。估计启动时间可包括在从客户接收计算实例启动请求和在物理主机上引导计算实例之间的时间段。可使用预测计算实例的启动时间的回归模型来识别用于启动计算实例的估计启动时间。可基于与计算实例相关联的实例特征和与该组物理主机中的物理主机相关联的物理主机特征来识别用于启动计算实例的估计启动时间，其中与计算实例相关联的实例特征包括用户选择的特征。

[0139] 部分地基于计算实例的估计启动时间，该组物理主机中的物理主机可被选择以提供计算实例的放置，并且可选地包括与计算实例的放置相关的附加因素，如在框1430中。与计算实例的放置相关的附加因素可包括物理主机利用率放置因素、许可成本放置因素和灾害影响放置因素。计算实例可被加载在物理主机上以便提供计算服务。

[0140] 在一个配置中，可将每个物理主机上的计算实例的估计启动时间与该组物理主机中的其它物理主机的估计启动时间进行比较。可选择该组物理主机中的物理主机，其可向计算实例的放置提供较低估计启动时间(与该组物理主机中的其它物理主机相比)。或者，可选择包括同时启动的较少量计算实例(与该组物理主机中的其它物理主机相比)的物理主机。在一个示例中，被选择来执行计算实例的物理主机可被验证为不同时执行超过预定阈值的多个计算实例。

[0141] 在另一配置中，可部分地基于区域或区中的计算实例的估计启动时间来选择区域或区以用于计算实例的放置。在一个示例中，当使用多个放置因素确定计算实例的放置时，可将加权值分配至计算实例的估计启动时间，并且可部分地基于分配至估计启动时间的加权值选择用于计算实例的放置的物理主机。此外，可识别启动与计算实例相关联的附接的估计时间量，并且可选择可向计算实例的放置提供较低估计附接时间(与该组物理主机中的其它物理主机相比)的物理主机。

[0142] 描述了用于使用启动时间预测来组织计算服务环境中的机器映像的缓存的技术。机器映像可提供用于在计算服务环境中启动计算实例的信息(即，可从机器映像启动计算实例)。例如，机器映像可指示用于启动计算实例的数据存储区的类型、启动许可等。机器映像可被缓存或存储在计算服务环境中的物理主机(也被称为服务器计算机)上以便减少启动计算实例的启动时间。换言之，与通过网络从数据存储区检索机器映像相比，将机器映像缓存至启动相关联计算实例的物理主机本地可为计算实例提供相对较快启动时间。术语“启动时间”通常是指在接收到启动计算实例的请求与将与计算实例相关联的机器映像引导至被选择来启动计算实例的物理主机上之间的时间段。

[0143] 在一个配置中，可识别计算服务环境的预期流量模式。预期流量模式可指示在限定时间段(以及可能在限定地理位置)期间可能在计算服务环境中启动的特定计算实例。例如，预期流量模式可指示计算实例A可能在星期二上午8:30启动。在一个示例中，可使用与计算服务环境中的过去流量模式相关的启发式规则来识别预期流量模式。在另一示例中，可使用机器学习模型来识别预期流量模式，该机器学习模型使用计算服务环境的历史流量信息以便预测计算服务环境的预期流量模式。

[0144] 计算服务环境的预期流量模式(例如，预期在限定时间段期间启动的计算实例的特征)可被提供至启动时间预测模型。启动时间预测模型可确定在计算服务环境中的预定义位置处抢先缓存与计算实例相关联的机器映像是否可引起减少用于启动计算实例的估

计启动时间。换言之，启动时间预测模型可确定在特定位置处缓存机器映像是否可减少估计启动时间(与没有缓存机器映像或在不改进估计启动时间的其它位置缓存机器映像相比)。预定义位置可包括计算服务环境中的特定物理主机、多组物理主机或者本地存储位置(例如，本地网络附接存储装置)。作为示例，启动时间预测模型可以是回归模型，其使用多个先前启动的计算实例的历史启动时间信息(包括历史计算实例缓存信息)用于确定预期在计算服务环境中启动的计算实例的估计启动时间。

[0145] 作为非限制性示例，计算实例A可能根据预期流量模式而在计算服务环境中启动。物理主机X和物理主机Y可被识别为可用于缓存与计算实例A相关联的机器映像。启动时间预测模型可确定在物理主机X上缓存机器映像可使计算实例A的预测启动时间是60秒。此外，启动时间预测模型可确定在物理主机Y上缓存机器映像可使计算实例A的预测启动时间是30秒。机器映像可被缓存在物理主机Y中，其中预期计算实例A可由客户在未来请求且当计算映像被缓存在物理主机Y上时机器映像可能最快启动。

[0146] 因此，当与计算实例相关联的机器映像被缓存在物理主机上时(如使用启动时间预测模型确定)，被预测为计算实例提供减少启动时间的物理主机可被选择来缓存机器映像。被选择来缓存机器映像的物理主机可包括在缓存布局中。包括在缓存布局中的物理主机可以是可用的和/或能够缓存机器映像。在一个示例中，缓存布局可识别可用于缓存机器映像的单个物理主机。或者，缓存布局可识别可用于缓存机器映像的一组物理主机。包括在缓存布局中的物理主机可具有用于执行计算实例的可用计算槽(例如，计算资源)。此外，可用计算槽可支持机器映像的类型或大小。

[0147] 图15是示出缓存机器映像减少在计算服务环境1500中的计算实例启动时间的图。可针对计算服务环境1500识别用于启动计算实例的预期流量模式1510。预期流量模式1510可指示可能在限定时间段期间(以及可能在限定地理位置)在计算服务环境1500中启动的计算实例。计算实例可与机器映像1512相关联。在一个示例中，计算实例和/或机器映像1512的特征可被提供至启动时间预测模型1530。这些特征可包括机器映像1512是否被缓存、计算实例的大小、机器映像可被缓存的位置等。启动时间预测模型130可基于计算实例和/或机器映像1512的特征来确定将机器映像1512缓存在计算服务环境1500中的预定义位置可减少计算实例的启动时间。预定义位置可包括经由高速网络连接而连接的某些物理主机或本地存储位置，诸如服务器架上或具有服务器的建筑物中的网络附接的存储装置(NAS)。

[0148] 缓存布局模块1540可使用启动时间预测模型1530来确定用于在计算服务环境1500中缓存机器映像1512的缓存布局。缓存布局可包括计算服务环境1500中的物理主机(或物理主机)，其具有可用的和/或能够缓存机器映像1512的缓存槽。此外，选择以在缓存布局中使用的物理主机可在与计算实例相关联的机器映像1512被缓存在物理主机上时提供启动计算实例的减少启动时间。

[0149] 作为非限制性示例，计算服务环境1500可包括多个物理主机1550、1560和1570。缓存布局模块1540可使用启动时间预测模型1530来确定将机器映像1512缓存在物理主机1550可使计算实例的估计启动时间是1580秒。此外，缓存布局模块1540可确定将机器映像1512缓存在物理主机1560上或物理主机1570上可分别使计算实例的估计启动时间是165秒和190秒。因此，缓存布局模块1540可选择用于缓存机器映像1512的物理主机1560(即，缓存

布局包括物理主机1560)，以便实现启动计算实例的减少启动时间(与在物理主机1550上或在物理主机1570上缓存机器映像1512相比)。

[0150] 图16示出根据本技术的一个示例的示例计算服务环境1600的组件。计算服务环境1600可包括经由网络1650与多个客户端装置1660通信的服务器计算机1610，并且服务器计算机可以是用于服务提供商环境1600的控制平面的一部分。服务器计算机1610可包含数据存储区1630和用于确定机器映像的缓存放置的多个模块。此外，计算服务环境1600可包括执行多个计算实例的多个服务器计算机240a-b。

[0151] 服务器计算机1640a-b可具有可用于执行计算实例的可用计算槽1642a-b(例如，空闲计算资源)。可用计算槽1642a-b可被分配至客户，其可随后利用可用计算槽1642a-b来执行计算实例。此外，服务器计算机1640a-b可具有可用于缓存与要执行的计算实例相关联的机器映像的可用缓存槽1644a-b。计算实例的示例可包括按需计算实例、保留计算实例和可中断计算实例。按需计算实例可以是客户可根据请求购买并执行的计算实例。保留计算实例可以是客户可在限定时间段内购买的计算实例的保留，使得当客户请求计算实例时，计算实例可用，且可中断计算实例可以是可在计算槽1642a-b中执行而不被另一计算实例类型使用的计算实例，除非对于可中断计算实例支付的价格低于当前投标价格。

[0152] 存储在数据存储区1630中的数据可包括基于历史数据的预期流量模式1632。预期流量模式1632可识别预期在限定时间段期间在计算服务环境1600中启动的计算实例。例如，预期流量模式1632可指示计算实例Z可能在星期六晚上7点启动。在一个示例中，可基于计算服务环境1600的历史流量信息来确定预期流量模式1632。例如，预期流量模式1632可指示计算实例Z可能在星期六晚上7点启动，因为计算实例Z已在过去两个月在类似时间启动。在一个配置中，可使用与计算服务环境1600中的过去流量模式相关的启发式规则来识别预期流量模式1632。在又另一配置中，可训练机器学习模型以使用计算服务环境1600的历史流量模式来确定预期流量模式1632。

[0153] 存储在数据存储区1630中的数据可包括用于要在计算服务环境1600中启动的计算实例的缓存布局1634。缓存布局1634可识别被选择用于缓存机器映像的计算服务环境1600中的物理主机，这是因为物理主机可用和/或能够缓存与计算实例相关联的机器映像。缓存布局1634可识别可用于缓存机器映像的单个物理主机或可用于缓存机器映像的一组物理主机。包括在缓存布局1634中的物理主机可具有支持机器映像的限定类型或大小的可用计算槽(例如，用于执行计算实例的计算资源)。

[0154] 在一个示例中，缓存布局1634可识别用于缓存机器映像的特定区域或区中的物理主机。此外，可响应于计算服务环境1600中的改变来修改或更新缓存布局1634。例如，当用于缓存机器映像的先前在缓存布局1634中识别的物理主机变得过载或满时(例如，引起减少启动时间)，可更新缓存布局1634以包括用于缓存机器映像的其它物理主机，这会引起减少计算实例的启动时间。因此，可周期性地更新缓存布局1634。

[0155] 服务器计算机1610可包括预期流量模式识别模块1622、估计启动时间预测模块1624、缓存布局模块1626、缓存建立模块1628以及本文未详细讨论的其它应用、服务、过程、系统、引擎或功能。预期流量模式识别模块1622可被配置为识别计算服务环境1600中的预期流量模式。预期流量模式可指示与计算实例相关联的机器映像，该计算实例可能在限定时间段期间在计算服务环境中启动。预期流量模式识别模块1622可使用机器学习模型来识

别预期流量模式，该机器学习模型使用计算服务环境1600的历史流量信息来预测预期流量模式。在一个示例中，预期流量模式识别模块1622可使用启发式规则来识别计算服务环境1600中的预期流量模式。

[0156] 估计启动时间预测模块1624可确定在计算服务环境1600中的预定义位置中缓存机器映像可减少计算实例的启动时间(与不缓存机器映像相比)。预定义位置可包括计算服务环境1600中的特定物理主机、一组物理主机或相对于物理主机的本地存储组件。在一个示例中，估计启动时间预测模块1624可使用启动时间预测模型来确定是否在预定义位置处缓存机器映像。换言之，当与计算实例相关联的机器映像被缓存在特定物理主机上时，启动时间预测模型可提供用于启动计算实例的估计启动时间。

[0157] 缓存布局模块1626可被配置为确定缓存布局以使得能够在计算服务环境1600中缓存机器映像。缓存布局可识别可用于缓存机器映像的处于预定义位置处的物理主机以便减少与机器映像相关联的计算实例的启动时间。在缓存布局中指示的物理主机可具有足够资源和能力来缓存机器映像。在一个示例中，缓存布局模块1626可使用启动时间预测模型来选择包括在缓存布局中的物理主机。换言之，缓存布局中的物理主机可能已经通过启动时间预测被识别为在机器映像被缓存在物理主机上时可能为计算实例提供减少启动时间。此外，当选择要包括在缓存布局中的物理主机时，缓存布局模块1626可识别物理主机的各种特性(例如，硬件类型、寻址)。作为另一示例，当选择要包括在缓存布局中的物理主机时，缓存布局模块1626可使用遗传技术或粒子群优化。因此，缓存布局模块1626可使用遗传技术或粒子群优化来选择在机器映像被缓存在物理主机上时提供减少启动时间的物理主机。

[0158] 缓存建立模块1628可被配置为根据缓存布局将机器映像存储在计算环境1600中的至少一个物理主机上。在一个示例中，缓存建立模块1628可将机器映像存储在与某些拓扑层(例如，特定区域、区、服务器架和物理主机)相关联的物理主机上。因此，可能存在与特定拓扑层相关联的本地存储装置。例如，可以区域、区或服务器架级提供缓存装置(例如，NAS)。通过抢先缓存机器映像，可减少用于启动计算实例的启动时间。在一个示例中，在预期启动计算实例并且计算实例可驻留在缓存位置中达限定时间段的限定时间段之前，缓存建立模块1628可将机器映像发送至缓存位置。当限定时间段已经结束时，缓存建立模块1628可从缓存中清除机器映像。作为非限制性示例，当预期在星期六上午8点至9点启动特定计算实例时，与计算实例相关联的机器映像可在星期六上午6点至10点之间被缓存在物理主机上，并随后从物理主机移除。

[0159] 图17示出用于在计算服务环境1700中缓存机器映像以便减少计算实例启动时间的系统和相关操作。机器映像可提供用于在计算服务环境中启动计算实例的信息(即，可从机器映像启动计算实例)。例如，机器映像可指示用于启动计算实例的数据存储区的类型、启动许可等。机器映像可存储在计算服务环境1700中的至少一个物理主机上，以便减少计算映像启动时间。换言之，与通过网络从单独数据存储区获取机器映像相比，本地缓存机器映像可提供相对更快计算实例启动时间。

[0160] 可识别计算服务环境1700的预期流量模式1710。预期流量模式1710可指示预期在某个时间段和/或在某个地理位置启动计算实例。作为非限制性示例，预期流量模式1710可指示计算实例可能在星期一上午8点启动。在一个配置中，可使用启发式规则1712来识别预期流量模式1710。启发式规则1712可涉及计算服务环境1700的过去流量模式。作为示例，使

用启发式规则1712可推断出如果计算实例在过去在星期一上午8点被启动则计算实例可能在相似时间启动。在另一配置中，可使用机器学习模型1714来识别预期流量模式1710。机器学习模型1714可使用来自计算服务环境400的历史流量信息，以便预测计算服务环境1700中的预期流量模式1710。

[0161] 在一个示例中，可将与预期在某一时间段期间启动的计算实例相关的实例特征提供至启动时间预测模型1730。实例特征可与由计算实例使用的机器映像1722和/或要在计算服务环境1700中启动的计算实例相关。例如，实例特征可包括计算实例的大小、计算实例的架构类型(例如，32位架构或64位架构)、计算实例的虚拟化类型(例如，半虚拟化或硬件虚拟机)和/或由计算实例使用的数据存储区的类型。

[0162] 启动时间预测模型1730可基于实例特征确定在计算服务环境1700中的预定义位置处抢先缓存机器映像1722是否可减少计算实例的启动时间。此外，与将机器映像1722缓存在计算服务环境1700中的另一位置处相比，启动时间预测模型1730可确定将机器映像1722缓存在特定位置(例如，特定物理主机)可减少启动时间。预定义位置可包括计算服务环境1700中的特定物理主机或一组物理主机。

[0163] 启动时间预测模型1730可以是由多个先前启动的计算实例的历史启动时间信息训练的机器学习模型，且启动时间预测模型1730可用于确定预期在计算服务环境1700中启动的计算实例的估计启动时间。因此，启动时间预测模型1730可在训练中使用与机器映像被缓存时的先前启动时间、计算实例未被缓存时的先前启动时间、机器映像被缓存在特定位置处时的先前启动时间等相关的历史启动时间信息，以预测预期要启动的计算实例的估计启动时间。在一个示例中，启动时间预测模型1730可以是用于预测估计启动时间的回归模型。

[0164] 作为非限制性示例，计算服务环境可包括物理主机1750、1760。启动时间预测模型1730可提供信息以帮助确定是否将机器映像1722缓存在物理主机1750、1760中的一个上以便减少计算实例的启动时间。启动时间预测模型1730可确定将机器映像1722缓存在物理主机1750上可使计算实例的估计启动时间是30秒。此外，启动时间预测模型1730可确定将机器映像1722缓存在物理主机1760上可使估计启动时间是25秒。当预测估计启动时间时，除了机器映像1722是否被缓存之外，启动时间预测模型1730还可使用各种类型的信息(诸如计算实例的大小、在物理主机处同时启动的数量、对物理主机占用的百分比等)。

[0165] 缓存布局模块1740可使用来自启动时间预测模型1730的预测信息来确定用于在计算服务环境1700中缓存机器映像1722的缓存布局。缓存布局可包括至少一个物理主机，该物理主机被选择来缓存机器映像1722以便减少计算实例的启动时间。为了确定缓存布局，缓存布局模块1740可识别可用于和/或能够缓存机器映像1722的处于计算服务环境1700中的预定义位置处的至少一个物理主机。缓存布局模块1740可经由启动时间预测模型1730来比较用于在每个可用物理主机上启动机器映像1722的估计启动时间，以确定机器映像1722是否将被缓存在每个各自物理主机上。缓存布局模块1740可比较用于在物理主机上启动计算实例的估计启动时间，并选择可为计算实例提供减少启动时间的物理主机。换言之，可提供减少启动时间的物理主机可包括在缓存布局中。在一个示例中，相同机器映像1722的数十或数百个副本可被缓存在一组物理主机中，以便提供减少启动时间。

[0166] 在一个配置中，缓存布局模块1740可使用遗传技术或粒子群优化来识别计算服务

环境1700中的物理主机或该组物理主机，其可缓存机器映像1722，以便提供计算实例的减少启动时间。机器映像1722可存储在所选物理主机上并且在计算服务环境1700中启动计算实例时从物理主机加载。

[0167] 作为非限制性示例，缓存布局模块1740可确定物理主机1750、1760可用于缓存机器映像1722。缓存布局模块1740可经由启动时间预测模型1730确定在物理主机1750、1760中的一个上缓存机器映像422可分别使估计启动时间是30秒、25秒或40秒。因此，缓存布局模块1740可选择要包括在缓存布局中的物理主机1760，因为与将机器映像1722缓存在物理主机1750上相比，将机器映像1722缓存在物理主机1760上可引起较少启动时间。或者，缓存布局模块1740可确定将机器映像1722缓存在包括在计算服务环境400中的网络附接存储(NAS)装置1770上，以便减少启动时间。

[0168] 在一个示例中，缓存布局模块1740可将机器映像1722存储在能够支持机器映像1722的大小的物理主机的可用缓存槽中。例如，物理主机1750可包括具有第一类型的可用缓存槽。物理主机1760可包括具有第二类型和第三类型的可用缓存槽。机器映像1722可以多种大小和类型来配置。缓存布局模块1740可验证物理主机中可用的缓存槽的类型是否能够存储该类型的机器映像1722。

[0169] 在另一示例中，机器映像1722可存储在与用于变化拓扑层的各个区域相关联的物理主机或物理装置上。拓扑层可包括特定区域、区、数据中心、服务器架、物理主机、缓存槽等。作为示例，缓存布局可包括特定区或该区中的特定数据中心(其将用来存储机器映像1722，诸如NAS 1770)。拓扑层可提供缓存(其又可为计算实例提供减少启动时间)。

[0170] 在一个配置中，可在计算服务环境1700中接收启动计算实例的请求。例如，可从期望来自计算服务环境1700的计算服务的客户接收该请求。与计算实例相关联的机器映像1722可被识别为处于暂停状态。换言之，用于机器映像1722的域创建过程可以是完整的，但是尚未开始域来启动计算实例。计算实例可通过加载机器映像1722且然后将机器映像1722从暂停状态切换至运行状态来启动，从而最小化计算实例的启动时间。在一个示例中，最流行的机器映像或最近使用的机器映像可以暂停状态存储在计算服务环境1700中。因此，启动这些机器映像可以最小启动时间来进行。

[0171] 图18示出用于将机器映像1814缓存在计算服务环境1800中以便实现用于启动与机器映像1814相关联的计算实例的期望启动时间1812的示例性系统和相关操作。在一个示例中，客户可提供指定用于启动计算实例的期望启动时间1812的启动请求1810。例如，客户可请求在少于45秒内启动计算实例。

[0172] 启动请求1810可被提供至缓存布局模块1840。缓存布局模块1840可确定用于缓存机器映像1814的缓存布局，使得计算实例的启动时间基本上满足客户的要求(例如，45秒)。换言之，缓存布局模块1840可识别可缓存机器映像1814的计算服务环境1800中的物理主机，使得可实现期望启动时间1812。

[0173] 在一个示例中，缓存布局模块1840可使用启动时间预测模型530来选择哪个物理主机缓存机器映像1814。缓存布局模块1840可确定是否将机器映像1814缓存在物理主机1850、1860或1870上。缓存布局模块1840可预测在机器映像1814被缓存在物理主机1850上时用于启动计算实例的估计启动时间。缓存布局模块1840可类似地预测机器映像1814被缓存在物理主机1860上或缓存在物理主机1870上时的估计启动时间。缓存布局模块1840可基

于来自启动时间预测模型1830的信息确定将机器映像1814缓存在物理主机上，该物理主机可提供对应于由客户指定的期望启动时间1812的启动时间。作为非限制性示例，缓存布局模块1840可确定将机器映像1814缓存在物理主机1850上可对应于用于启动计算实例的期望启动时间1812(例如，45秒)。

[0174] 在另一示例中，可从客户接收用于启动计算实例的期望启动时间1812，并且作为响应，缓存布局模块1840可识别已经缓存了与将要启动的计算实例相关联的机器映像的物理主机。缓存布局模块1840可使用来自启动时间预测模型1830的信息来确定在已经缓存了机器映像的物理主机上启动计算实例是否符合期望启动时间1812，且如果预测启动时间符合期望启动时间1812，则可在物理主机上启动计算实例。

[0175] 图19示出根据服务水平协议(SLA)在计算服务环境1900中缓存机器映像的示例性系统和相关操作。可为在计算服务环境1900中启动的计算实例确定计算实例实际启动时间1910。可在计算实例已经被成功启动之后确定计算实例的实际启动时间。可将计算实例的实际启动时间提供至SLA比较模块1920。SLA比较模块1920可比较计算实例的实际启动时间与计算服务环境1900的SLA。在一个示例中，SLA比较模块1920可确定计算实例的实际启动时间与计算服务环境1900的SLA一致。

[0176] 或者，SLA比较模块1920可比较计算实例的实际启动时间与计算服务环境1900的SLA，并确定计算实例的实际启动时间与计算服务环境1900的SLA不一致。例如，SLA可指定计算实例的启动时间小于10分钟。然而，SLA比较模块可确定实际启动时间大于10分钟。当实际启动时间与SLA不一致时，SLA比较模块1920可通知缓存布局模块1940。缓存布局模块1940可确定用于缓存与计算实例相关联的机器映像1912的缓存布局，使得计算实例的启动时间与SLA相符(例如，减少启动时间)。在一个示例中，缓存布局模块1940可通过将机器映像1912存储在计算服务环境1900中的附加物理主机上来修改现有缓存布局。在如图19所示的示例中，缓存布局模块1940可将机器映像1912存储在物理主机1950和物理主机1970上，但不存储在物理主机1950上，以减少启动时间并符合计算服务环境1900的SLA。

[0177] 图20是示出生成启动时间预测模型2050以预测在计算服务环境中启动的计算实例的启动时间的示例性框图2000。启动时间预测模型2050可以是使用实际启动时间预测数据2010创建的机器学习模型。实际启动时间输入数据2010可包括用于先前已经在计算服务环境中启动的多个计算实例的信息(例如，启动度量)。因此，实际启动时间输入数据2010可包括与计算服务环境中的先前启动的计算实例有关的历史信息。此外，实际启动时间输入数据2010可包括计算服务环境中的多个物理主机的历史信息。如下所述，实际启动时间输入数据2010可被变换以用于训练启动时间预测模型2050。

[0178] 作为非限制性示例，实际启动时间输入数据2010可指示计算实例A花费60秒来启动，而计算实例A的大小相对较大、使用第一类型的数据存储区、使用32位架构并在同时启动其它五个计算实例的物理主机上启动。作为另一非限制性示例，实际启动时间输入数据2010可指示计算实例B花费15秒来启动，而计算实例B的大小相对较小、使用第二类型的数据存储区、使用64位架构并在未同时启动其它计算实例的物理主机上启动。

[0179] 实际启动时间输入数据2010可被提供至特征选择和标准化模块2020。特征选择和标准化模块2020可将实际启动时间输入数据2010转换为模型特征。换言之，模型特征可涉及先前启动的计算实例的特性以及先前在其上启动计算实例的物理主机的特性。模型特征

可被分类为实例特征和物理主机特征。

[0180] 实例特征可包括但不限于计算实例的大小、计算实例使用的机器映像(例如,机器映像或内核映像)、在启动计算实例时机器映像是否被缓存在物理主机上、计算实例的架构类型(例如,32位架构或64位架构)、计算实例的虚拟化类型(例如,半虚拟化或硬件虚拟机)和由计算实例使用的数据存储区的类型。实例特征可包括用户控制的特征,诸如用于启动计算实例的操作系统(OS)的类型和网络类型(例如,虚拟私有云)。

[0181] 物理主机特征可包括但不限于物理主机可托管的最大数量的计算实例、与物理主机相关联的硬件类型、与物理主机相关联的硬件供应商、在计算实例要被启动时物理主机的占用百分比以及物理主机所位于的区。物理主机特征可包括在启动计算实例的物理主机(即,目标物理主机)上待决计算实例和/或运行计算实例的平均、最小和最大数量。此外,物理主机特征可包括当前在启动计算实例的物理主机(即,目标物理主机)上处于待决状态和/或运行状态的多个计算实例。

[0182] 特征选择和标准化模块2020可对模型特征进行标准化(即,将在不同尺度上测量的值调整至标称的共同尺度),以便创建启动时间预测训练数据2030。启动时间预测训练数据2030可表示在启动时间预测输入数据2010中识别的多个计算实例的聚合特征。启动时间预测训练数据2030可被提供至机器学习选择模块2040。机器学习选择模块2040可使用启动时间预测训练数据2030以训练各种机器学习模型2042。例如,可训练回归模型。回归模型2042可包括但不限于支持向量机、随机梯度下降、自适应引导、附加树和随机森林。各种回归模型2042可对应于具有各种级别成功的启动时间预测训练数据2030。在一个示例中,随机森林回归量可相对于启动时间预测训练数据2030提供相对较高精确度,且因此,当估计计算实例的启动时间时,机器学习选择模块2040可使用随机森林回归量。

[0183] 启动时间预测模型2050可接收对启动计算实例的请求,并且基于与计算实例相关联的实例特征和物理主机特征,启动时间预测模型2050可预测计算实例的启动时间。在一个示例中,启动时间预测模型2050可确定在相同物理主机上同时计算实例启动的数量、由计算实例使用的存储的数据的类型、由计算实例使用的架构类型以及与计算实例相关联的机器映像可对计算实例的启动时间具有更大影响(与其它模型特征相比)。

[0184] 在一些情况下,来自计算实例的预测启动时间可偏离计算实例的实际启动时间。与计算实例相关联的实例特征和物理主机特征以及启动计算实例的实际启动时间可用于进一步训练启动时间预测模型2050,以便改进未来启动时间预测。

[0185] 图21是示出用于减少计算实例启动时间的示例方法的流程图。可识别计算服务环境中的预期流量模式,如在框2110中。预期流量模式可指示与预期在预定时间段期间在计算服务环境中启动的计算实例相关联的机器映像。在一个示例中,启发式规则可用于识别计算服务环境的预期流量模式。启发式规则可涉及计算服务环境的历史流量模式。

[0186] 机器映像可被确定为被缓存在计算服务环境中的预定义位置,以便减少计算实例的启动时间(与不缓存机器映像相比),如框2120中。换言之,将机器映像缓存至启动计算实例的物理主机本地可提供相对更快启动时间(与通过网络从数据存储区获取机器映像相比)。在一个示例中,启动时间预测模型可用于确定将机器映像缓存在预定义位置。

[0187] 可确定能够在计算服务环境中缓存机器映像的缓存布局,如在框2130中。缓存布局可识别可用于缓存机器映像的处于计算服务环境中的预定义位置处的物理主机。缓存布

局可识别可用于缓存机器映像的物理主机或每个可用于缓存机器映像的副本的一组物理主机。

[0188] 机器映像可根据缓存布局存储在计算环境中的至少一个物理主机上，如在框2140中。将机器映像存储在物理主机上可减少用于启动计算实例的启动时间。在一个示例中，机器映像可存储在多个物理主机上，以便基于相同机器映像减少多个计算实例的启动时间。

[0189] 在一个示例中，可接收用于启动计算实例的期望启动时间，并且可确定缓存布局以使得能够缓存机器映像，使得用于启动计算实例的实际启动时间基本上类似于期望启动时间。在又另一示例中，可识别要在计算服务环境中缓存的各种类型的机器映像，并且可确定缓存布局包括在能够缓存各种类型的机器映像的物理主机上的一个或多个缓存槽上。

[0190] 图22是示出用于减少计算实例启动时间的另一示例方法的流程图。可识别预期在限定时间段期间在计算服务环境中启动的计算实例，如在框2210中。在一个示例中，启发式规则可用于识别预期在计算服务环境中启动的计算实例。启发式规则可涉及计算服务环境的历史流量模式。在一个示例中，可使用预测计算服务环境的预期流量模式的机器学习模型来识别预期在限定时间段期间在计算服务环境中启动的计算实例。

[0191] 可作出在计算服务环境中缓存机器映像的确定，以便减少用于启动计算实例的启动时间(与不缓存机器映像相比)，如在框2220中。在一个示例中，启动时间预测模型可用于确定缓存机器映像。启动时间预测模型可以是回归模型，其部分地基于与计算实例相关联的实例特征和与计算服务环境中的一组物理主机相关联的物理主机特征来预测计算实例的启动时间。

[0192] 可使用启动时间预测模型来选择可用于缓存计算实例的机器映像的计算服务环境中的至少一个物理主机，如在框2230中。此外，物理主机可能够缓存机器映像(例如，物理主机包括具有足够容量来缓存机器映像的可用存储槽)。

[0193] 机器映像可存储在物理主机中，如在框2240中。机器映像可存储在物理主机上，以便减少用于启动计算实例的启动时间。在一个示例中，机器映像可根据所选拓扑层存储在计算服务环境中的物理主机或物理存储装置上。在另一示例中，可根据缓存布局存储机器映像。缓存布局可指示可用于缓存机器映像的计算服务环境中的多个物理主机，以便减少计算实例(多个)的启动时间。在又另一示例中，机器映像可存储在物理主机上，使得计算实例的启动时间与计算服务环境的服务水平协议(SLA)一致。

[0194] 在一个配置中，可接收启动计算实例的请求。可识别与在计算服务环境中的物理主机上缓存的计算实例相关联的机器映像。机器映像可被加载在物理主机上以便提供计算服务。在另一配置中，可识别要在计算服务环境中缓存的机器映像的大小。可选择具有能够缓存该大小的机器映像的槽的计算服务环境中的物理主机。机器映像可被缓存在计算服务环境中的预定义位置，以便减少与机器映像相关联的启动时间。

[0195] 在一个示例中，可识别用于启动计算实例的期望启动时间。可从访问计算服务环境的客户接收期望启动时间。可选择可用于存储与计算实例相关联的机器映像的计算服务环境中的物理主机。物理主机可被验证为能够使用启动时间预测模型根据期望启动时间来启动计算实例。机器映像可存储在计算环境中的物理主机中，使得用于启动计算实例的实际启动时间基本上类似于期望启动时间。

[0196] 在另一示例中，计算实例的启动时间可被确定为与计算服务环境的服务水平协议

(SLA)不一致。与计算实例相关联的机器映像可存储在计算服务环境中的附加物理主机上，以进一步减少计算实例的启动时间，以便符合SLA。当同时启动多个计算实例时，将机器映像存储在附加物理主机上可减少启动时间。此外，将机器映像存储在附加物理主机上可允许选择提供相对最低启动时间的物理主机来启动计算实例。在又另一示例中，可接收在计算服务环境中启动计算实例的请求。与计算实例相关联的机器映像可被识别为处于暂停状态。计算实例可通过将机器映像从暂停状态切换至运行状态而在计算服务环境中启动，从而最小化在计算服务环境中启动计算实例的启动时间。可根据以下条款来描述本公开的实施方案：

[0197] 1.一种上面包含指令的非暂时性机器可读存储介质，在由处理器执行时，所述指令：

[0198] 获得表示多个先前计算实例启动的启动特征的训练数据；

[0199] 使用所述训练数据训练随机森林回归模型；

[0200] 接收对在计算服务环境内的物理主机上启动计算实例的预测启动时间的请求；

[0201] 识别已经被确定为对所述计算实例的启动时间具有影响的与在所述计算服务环境内启动所述计算实例相关联的启动特征；以及

[0202] 将与启动所述计算实例相关联的所述启动特征输入至机器学习回归模型中，所述机器学习回归模型输出用于在所述计算服务环境内启动所述计算实例的预测启动时间。

[0203] 2.如条款1所述的非暂时性机器可读存储介质，其中当由所述处理器执行时，指令进一步从包括机器映像启动特征、物理主机启动特征和客户配置启动特征的启动特征选择被确定为对计算实例的启动时间具有影响的启动特征。

[0204] 3.如条款1所述的非暂时性机器可读存储介质，其中当由所述处理器执行时，指令通过比较所述预测启动时间与SLA启动时间来进一步确定是否可能满足所述SLA(服务水平协议)启动时间。

[0205] 4.一种计算机实现的方法，其包括：

[0206] 在配置有可执行指令的一个或多个计算机系统的控制下，

[0207] 接收对在计算服务环境内的物理主机上启动计算实例的预测启动时间的请求；

[0208] 使用处理器获得与计算实例的启动特征相关联的数据，所述计算实例的所述启动特征被确定为对所述计算实例在计算服务环境内的物理主机上的启动时间具有影响；以及

[0209] 使用所述处理器，将所述计算实例的所述启动特征输入至机器学习模型，所述机器学习模型输出用于在所述计算服务环境内启动计算实例的预测启动时间。

[0210] 5.如条款4所述的方法，其中获得与启动特征相关联的所述数据还包括获得与机器映像启动特征、物理主机启动特征和客户配置启动特征相关联的数据。

[0211] 6.如条款4所述的方法，其还包括在将所述启动特征输入至所述机器学习模型之前使所述启动特征标准化。

[0212] 7.如条款4所述的方法，其还包括对机器学习参数进行参数值搜索，所述机器学习参数引起所述机器学习模型对所述启动特征的拟合优度。

[0213] 8.如条款7所述的方法，其中进行所述参数值搜索还包括，使用分布式遗传算法进行机器学习参数的参数值搜索。

[0214] 9.如条款4所述的方法，其中获得与启动特征相关联的数据还包括，获得表示多个

先前计算实例启动的启动特征的活动训练数据；

[0215] 从与所述启动特征相关联的所述活动训练数据提取特征；以及

[0216] 使用来自所述活动训练数据的所述特征训练所述机器学习模型。

[0217] 10.如条款4所述的方法，其中将所述启动特征输入至机器学习模型还包括将所述启动特征输入至选自以下中的至少一个的机器学习模型：随机森林模型、超随机树模型、AdaBoost模型、随机梯度下降模型或支持向量机器模型。

[0218] 11.如条款4所述的方法，其中将所述启动特征输入至机器学习模型还包括，将所述计算实例的所述启动特征输入至机器学习回归模型。

[0219] 12.如条款4所述的方法，其还包括，从客户接收启动所述计算实例的启动请求；

[0220] 识别与所述计算实例的计算实例类型相关联的SLA启动时间；以及

[0221] 通过比较所述计算实例的所述预测启动时间与所述SLA启动时间来确定是否可能满足所述SLA启动时间。

[0222] 13.如条款12所述的方法，其还包括通知计算服务提供商：当确定所述预测启动时间大于所述SLA启动时间时，所述SLA启动时间可能不会实现。

[0223] 14.如条款12所述的方法，其还包括构建所述预测启动时间大于所述SLA启动时间的SLA违反特征，并且包括所述SLA违反特征与其它特征输入至机器学习分类模型。

[0224] 15.如条款4所述的方法，其还包括：

[0225] 识别所述计算实例的SLA启动时间；

[0226] 通过比较所述计算实例的所述预测启动时间与所述SLA启动时间来确定是否可能满足所述SLA启动时间；以及

[0227] 分析所述计算实例将被启动的计算服务环境的状态，以确定是否可能进行动作，在已经确定了将可能违反所述SLA启动时间时，所述动作可能防止违反所述SLA启动时间。

[0228] 16.如条款15所述的方法，其中进行的动作还包括，从可用于托管所述计算实例的一组物理主机中移除物理主机。

[0229] 17.如条款15所述的方法，其中进行的动作还包括，将至少一个物理主机添加至可用于托管所述计算实例的一组物理主机。

[0230] 18.一种系统，其包括：

[0231] 处理器；

[0232] 包括指令的存储器装置，在由所述处理器执行时，所述指令使所述系统：

[0233] 识别包含在启动配置中的启动特征，所述启动特征已经被确定为对计算实例在计算服务环境内的启动时间具有影响；

[0234] 从数据源获得用于所述启动特征的数据；

[0235] 将所述启动特征输入至机器学习模型，所述机器学习模型输出用于在所述计算服务环境内启动计算实例的预测启动时间；以及

[0236] 通过比较所述预测启动时间与SLA启动时间来确定是否可能满足所述SLA启动时间。

[0237] 19.如条款18所述的系统，其中所述存储器装置包括指令，在由所述处理器执行时，所述指令使所述系统将所述启动特征输入至机器学习回归模型。

[0238] 20.如条款18所述的系统，其中所述存储器装置包括指令，在由所述处理器执行

时，所述指令使所述系统获得：用于启动所述计算实例的机器映像的特征、能够托管所述计算实例的物理主机服务器的特征和在启动时附接至所述计算实例的计算实例附接的特征。

[0239] 21.一种包括上面包含的指令的非暂时性机器可读存储介质，在由处理器执行时所述指令使系统：

[0240] 接收在计算服务环境中启动计算实例的请求；

[0241] 将与所述计算实例相关联的实例特征和与所述计算服务环境中的一组物理主机相关联的物理主机特征提供至机器学习模型；

[0242] 使用所述机器学习模型来确定用于在所述计算服务环境中的各个物理主机上启动所述计算实例的估计启动时间；以及

[0243] 根据较低估计启动时间(与该组物理主机中的其它物理主机相比)，从该组物理主机选择物理主机来提供所述计算实例的放置。

[0244] 22.如条款21所述的非暂时性机器可读存储介质，其还包括在执行时进一步使所述系统进行以下操作的指令：

[0245] 向所述计算实例的所述估计启动时间分配加权值；以及

[0246] 部分地基于分配至所述估计启动时间的所述加权值来选择用于所述计算实例的放置的所述物理主机。

[0247] 23.如条款21所述的非暂时性机器可读存储介质，其还包括在执行时进一步使所述系统进行以下操作的指令：使用包括在启动所述计算实例的所述请求中的放置约束，从该组物理主机中选择所述物理主机，以提供所述计算实例的放置。

[0248] 24.如条款21所述的非暂时性机器可读存储介质，其中所述估计启动时间是在选择用于放置所述计算实例的所述物理主机时使用的多个因素中的一个，所述多个因素包括以下中的至少一个：物理主机利用率放置因素、许可成本放置因素和灾害影响因素。

[0249] 25.一种计算机实现的方法，其包括：

[0250] 在配置有可执行指令的一个或多个计算机系统的控制下：

[0251] 使用所述计算机系统的一个或多个处理器接收在计算服务环境中启动计算实例的请求；

[0252] 使用所述计算机系统的所述一个或多个处理器来识别在物理主机组中的物理主机上启动所述计算实例的估计启动时间；以及

[0253] 使用所述计算机系统的所述一个或多个处理器部分地基于所述计算实例的所述估计启动时间来选择该组物理主机中的物理主机以放置所述计算实例。

[0254] 26.如条款25所述的方法，其还包括使所述计算实例在所述物理主机上被启动。

[0255] 27.如条款25所述的方法，其还包括：

[0256] 比较所述计算实例在每个物理主机上的所述估计启动时间与在该组物理主机中的其它物理主机；以及

[0257] 选择具有较低估计启动时间(与该组物理主机中的其它物理主机相比)的该组物理主机中的所述物理主机。

[0258] 28.如条款25所述的方法，其还包括部分地基于与所述计算实例相关联的实例特征和与该组物理主机中的所述物理主机相关联的物理主机特征来识别用于启动所述计算实例的所述估计启动时间，其中与所述计算实例相关联的所述实例特征包括用户选择的特

征。

[0259] 29.如条款25所述的方法，其还包括使用回归模型来识别用于启动所述计算实例的所述估计启动时间。

[0260] 30.如条款25所述的方法，其还包括：

[0261] 向所述计算实例的所述估计启动时间分配加权值；以及

[0262] 部分地基于分配至所述估计启动时间的所述加权值来选择用于所述计算实例的放置的所述物理主机。

[0263] 31.如条款25所述的方法，其还包括使用附加放置因素选择该组物理主机中的所述物理主机以放置所述计算实例，所述附加放置因素包括以下中的至少一个：物理主机利用率放置因素、许可成本放置因素或灾害影响放置因素。

[0264] 32.如条款25所述的方法，其还包括选择包括同时启动的较低数量(与该组物理主机中的其它物理主机相比)的计算实例的所述物理主机。

[0265] 33.如条款25所述的方法，其还包括部分地基于用于区域或区中的计算实例的所述估计启动时间来选择用于放置所述计算实例的区域或区。

[0266] 34.如条款25所述的方法，其还包括验证被选择来执行所述计算实例的所述物理主机不同时执行超过预定阈值的多个计算实例。

[0267] 35.如条款25所述的方法，其中所述估计启动时间包括在从客户接收计算实例启动请求与在所述物理主机上引导所述计算实例之间的时间段。

[0268] 36.如条款25所述的方法，其还包括：

[0269] 识别启动与所述计算实例相关联的附接的估计时间量；以及

[0270] 选择可向所述计算实例的放置提供较低估计附接时间(与该组物理主机中的其它物理主机相比)的所述物理主机。

[0271] 37.一种用于确定计算实例放置的系统，其包括：

[0272] 处理器；

[0273] 包括指令的存储器装置，在由所述处理器执行时，所述指令使所述系统：

[0274] 接收在计算服务提供商上启动计算实例的请求；

[0275] 部分地基于与所述计算实例相关联的实例特征和与被配置为托管计算实例的物理主机相关联的物理主机特征来识别在物理主机上启动所述计算实例的估计启动时间；以及

[0276] 部分地基于所述计算实例的所述估计启动时间选择物理主机以放置所述计算实例。

[0277] 38.如条款37所述的系统，其中所述存储器装置包括指令，当由所述处理器执行时，所述指令使所述系统选择具有同时被启动的较低数量(与其它物理主机相比)的计算实例的所述物理主机。

[0278] 39.如条款37所述的系统，其中所述存储器装置包括指令，所述指令在由所述处理器执行时使所述系统验证被选择来执行所述计算实例的所述物理主机不同时执行超过预定阈值的多个计算实例。

[0279] 40.如条款37所述的系统，其中所述存储器装置包括指令，所述指令在由所述处理器执行时使所述系统使用回归模型来识别所述计算实例的所述估计启动时间，所述回归模

型部分地基于所述实例特征和所述物理主机特征预测计算实例的启动时间。

[0280] 41.一种上面包含指令的非暂时性机器可读存储介质，所述指令由处理器执行以改进计算实例启动时间，所述方法包括：

[0281] 识别指示与预期在限定时间段期间在计算服务环境中启动的计算实例相关联的机器映像的预期流量模式；

[0282] 使用启动时间预测模型来确定在所述计算服务环境中的预定义位置处缓存所述机器映像将减少所述计算实例的启动时间(与不缓存所述机器映像相比)；

[0283] 确定缓存布局以使所述机器映像能够缓存在所述计算服务环境中，所述缓存布局识别可用于缓存所述机器映像的处于所述计算服务环境中的预定义位置处的物理主机；以及

[0284] 根据所述缓存布局将所述机器映像存储在所述计算环境中的至少一个物理主机上。

[0285] 42.如权利要求41所述的非暂时性机器可读存储介质，其还包括使用启发式规则来识别所述计算服务环境的所述预期流量模式，所述启发式规则与所述计算服务环境的历史流量模式相关。

[0286] 43.如权利要求41所述的非暂时性机器可读存储介质，其还包括：

[0287] 接收用于启动计算实例的期望启动时间；以及

[0288] 确定所述缓存布局以能够缓存与所述计算实例相关联的机器映像，使得用于启动所述计算实例的实际启动时间基本上类似于所述期望启动时间。

[0289] 44.如权利要求41所述的非暂时性机器可读存储介质，其还包括：

[0290] 识别要在所述计算服务环境中缓存的各种类型的机器映像；以及

[0291] 确定所述缓存布局以包括所述物理主机上能够缓存所述各种类型的机器映像的一个或多个槽。

[0292] 45.一种计算机实现的方法，其包括：

[0293] 在配置有可执行指令的一个或多个计算机系统的控制下：

[0294] 使用所述计算机系统的一个或多个处理器来识别预期在限定时间段期间在计算服务环境中启动的计算实例；

[0295] 使用所述计算机系统的所述一个或多个处理器经由启动时间预测模型确定在所述计算服务环境中缓存所述计算实例的机器映像将减少启动所述计算实例的启动时间(与不缓存所述机器映像相比)；

[0296] 使用所述计算机系统的所述一个或多个处理器在所述计算服务环境中选择可用于缓存所述机器映像的至少一个物理主机，以减少由所述启动时间预测模型预测的所述计算实例的所述启动时间；以及

[0297] 使用所述计算机系统的所述一个或多个处理器将所述机器映像存储在所述物理主机中。

[0298] 46.如权利要求45所述的方法，其还包括：

[0299] 接收启动所述计算实例的请求；

[0300] 识别与缓存在所述计算服务环境中的所述物理主机上的所述计算实例相关联的所述机器映像；以及

[0301] 将所述机器映像加载在所述物理主机上以便提供计算服务。

[0302] 47.如权利要求45所述的方法，其还包括使用启发式规则来识别所述计算服务环境的预期流量模式，所述启发式规则与所述计算服务环境的历史流量模式相关。

[0303] 48.如权利要求45所述的方法，其还包括使用机器学习模型来识别预期在所述限定时间段期间和限定地理位置在所述计算服务环境中启动的所述计算实例，所述机器学习模型预测所述计算服务环境的预期流量模式。

[0304] 49.如权利要求45所述的方法，其还包括：

[0305] 识别要在所述计算服务环境中缓存的所述机器映像的大小；

[0306] 在所述计算服务环境中选择具有能够缓存该大小的机器映像的槽的所述物理主机，所述机器映像被缓存在所述计算服务环境中的预定义位置处，以便减少与所述机器映像相关联的所述启动时间。

[0307] 50.如权利要求45所述的方法，其还包括根据所选拓扑层将所述机器映像存储在所述计算服务环境上。

[0308] 51.如权利要求45所述的方法，其还包括根据缓存布局存储所述机器映像，所述缓存布局指示可用于缓存所述机器映像以便减少所述启动时间的所述计算服务环境中的多个物理主机。

[0309] 52.如权利要求45所述的方法，其还包括：

[0310] 识别用于启动所述计算实例的期望启动时间；

[0311] 在所述计算服务环境中选择可用于存储与所述计算实例相关联的所述机器映像的所述物理主机；以及

[0312] 将所述机器映像存储在所述计算环境中的所述物理主机中，使得用于启动所述计算实例的实际启动时间基本上类似于所述期望启动时间。

[0313] 53.如权利要求45所述的方法，其还包括将所述机器映像存储在所述物理主机上，使得所述启动时间与所述计算服务环境的服务水平协议(SLA)一致。

[0314] 54.如权利要求45所述的方法，其还包括：

[0315] 确定所述计算实例的所述启动时间与所述计算服务环境的服务水平协议(SLA)不一致；以及

[0316] 将与所述计算实例相关联的所述机器映像存储在所述计算服务环境中的附加物理主机上，以进一步减少与所述机器映像相关联的所述计算实例的所述启动时间，以便符合所述SLA。

[0317] 55.如权利要求45所述的方法，其还包括：

[0318] 接收在所述计算服务环境中启动所述计算实例的请求；

[0319] 识别与所述计算实例相关联的所述机器映像处于暂停状态；以及

[0320] 通过将所述机器映像从所述暂停状态切换至运行状态而在所述计算服务环境中启动所述计算实例，从而最小化在所述计算服务环境中启动所述计算实例的所述启动时间。

[0321] 56.如权利要求45所述的方法，其中所述启动时间预测模型是回归模型，其部分地基于与所述计算实例相关联的实例特征和与所述计算服务环境中的一组物理主机相关联的物理主机特征来预测所述计算实例的所述启动时间。

[0322] 57.一种用于减少计算实例启动时间的系统，其包括：

[0323] 处理器；

[0324] 包括指令的存储器装置，在由所述处理器执行时，所述指令使所述系统：

[0325] 识别计算服务环境中的预期流量模式，所述预期流量模式指示与预期在限定时间段期间在所述计算服务环境中启动的计算实例相关联的机器映像；

[0326] 使用启动时间预测模型来确定在所述计算服务环境中的预定义位置中缓存所述机器映像将减少所述计算实例的启动时间(与不缓存所述机器映像相比)；

[0327] 确定缓存布局以使所述机器映像能够缓存在所述计算服务环境中，所述缓存布局识别可用于缓存所述机器映像的处于所述计算服务环境中的所述预定义位置处的物理主机；以及

[0328] 根据所述缓存布局将所述机器映像存储在所述计算环境中的至少一个物理主机上。

[0329] 58.如权利要求57所述的系统，其中所述存储器装置包括指令，在由所述处理器执行时，所述指令使所述系统将所述机器映像存储在所述物理主机上，使得所述启动时间与所述计算服务环境的服务水平协议(SLA)一致。

[0330] 59.如权利要求57所述的系统，其中所述存储器装置包括指令，在由所述处理器执行时，所述指令使所述系统根据所选拓扑层将所述机器映像存储在所述计算服务环境上。

[0331] 60.如权利要求57所述的系统，其中所述存储器装置包括指令，在由所述处理器执行时，所述指令使所述系统：

[0332] 接收启动所述计算实例的请求；

[0333] 识别与在所述计算服务环境中的所述物理主机上缓存的所述计算实例相关联的所述机器映像；以及

[0334] 加载所述机器映像以便提供计算服务。

[0335] 虽然针对该技术呈现的流程图可暗示特定的执行顺序，但是执行顺序可不同于所示顺序。例如，可相对于所示顺序重新布置两个以上框的顺序。此外，可并行或部分并行地执行连续示出的两个框或更多个框。在一些配置中，可省略或跳过流程图中所示的一个或多个框。为了增强效用、计费、性能、测量、故障排除或类似原因的目的，可将任何数量的计数器、状态变量、警告信号或消息添加至逻辑流中。

[0336] 在本说明书中描述的一些功能单元已经被标记为模块，以便更具体地强调它们的实现独立性。例如，模块可实现为包括定制VLSI电路或门阵列、现成半导体(诸如逻辑芯片、晶体管或其它分立组件)的硬件电路。模块还可在可编程硬件装置(诸如现场可编程门阵列、可编程阵列逻辑、可编程逻辑装置等)中实现。

[0337] 模块还可在用于由各种类型的处理器执行的软件中实现。可执行代码的标识模块可例如包括一个或多个计算机指令块，其可被组织为对象、过程或函数。然而，所识别的模块的可执行体不需要物理地位于一起，但是可包括存储在包括该模块的不同位置的不同指令，并且在逻辑上一起连接时实现该模块的所述目的。

[0338] 实际上，可执行代码的模块可以是单个指令或许多指令，并且甚至可分布在几个不同的代码段上、在不同程序之间并跨几个存储器装置。类似地，在这里，操作数据可在模块内被识别并示出，并且可以任何合适形式体现并且在任何合适类型的数据结构内组织。

操作数据可被收集为单个数据集，或者可分布在不同位置(包括在不同存储装置)上。模块可以是被动的或主动的，包括可操作以进行期望功能的代理。

[0339] 这里描述的技术还可存储在计算机可读存储介质上，该计算机可读存储介质包括使用用于存储信息(诸如计算机可读指令、数据结构、程序模块或其它数据)的任何技术实现的易失性和非易失性、可移动和不可移动介质。计算机可读存储介质包括但不限于非暂时性介质，诸如RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字通用盘(DVD)或其它光学存储装置、磁带盒、磁性磁带、磁盘存储装置或其它磁性存储装置或可用于存储期望信息和所描述的技术的任何其它计算机存储介质。

[0340] 本文描述的装置还可包含允许装置与其它装置通信的通信连接或网络装置和网络连接。通信连接是通信介质的示例。通信介质通常在诸如载波或其它传输机制的调制数据信号中包含计算机可读指令、数据结构、程序模块和其它数据，并且包括任何信息传递介质。“调制数据信号”是指具有以在信号中编码信息的方式设置或改变其一个或多个特性的信号。通过示例而非限制的方式，通信介质包括诸如有线网络或直接有线连接的有线介质，和诸如声学、射频、红外和其它无线介质的无线介质。这里使用的术语计算机可读介质包括通信介质。

[0341] 参考附图中所示的示例，并且本文使用特定语言来描述相同的示例。然而，应理解，不意图因此限制本技术的范围。本文所示的特征的改变和进一步修改以及本文所示的示例的附加应用被认为都在本说明书的范围内。

[0342] 此外，所描述的特征、结构或特性可以任何合适的方式在一个或多个示例中组合。在前面的描述中，提供了许多具体细节，诸如各种配置的示例，提供了对所描述的技术的示例的透彻理解。然而，应认识到，可在没有一个或多个具体细节的情况下，或者通过其它方法、组件、装置等来实现本技术。在其它情况下，没有详细示出或描述公知的结构或操作以避免模糊本技术的各方面。

[0343] 虽然已经以结构特征和/或操作特定的语言描述了主题，但是应理解，所附权利要求中限定的主题不一定限于上述具体特征和操作。相反，上述具体特征和动作被公开作为实现权利要求的示例形式。在不脱离所描述的技术的精神和范围的情况下，可设计出许多修改和替代布置。

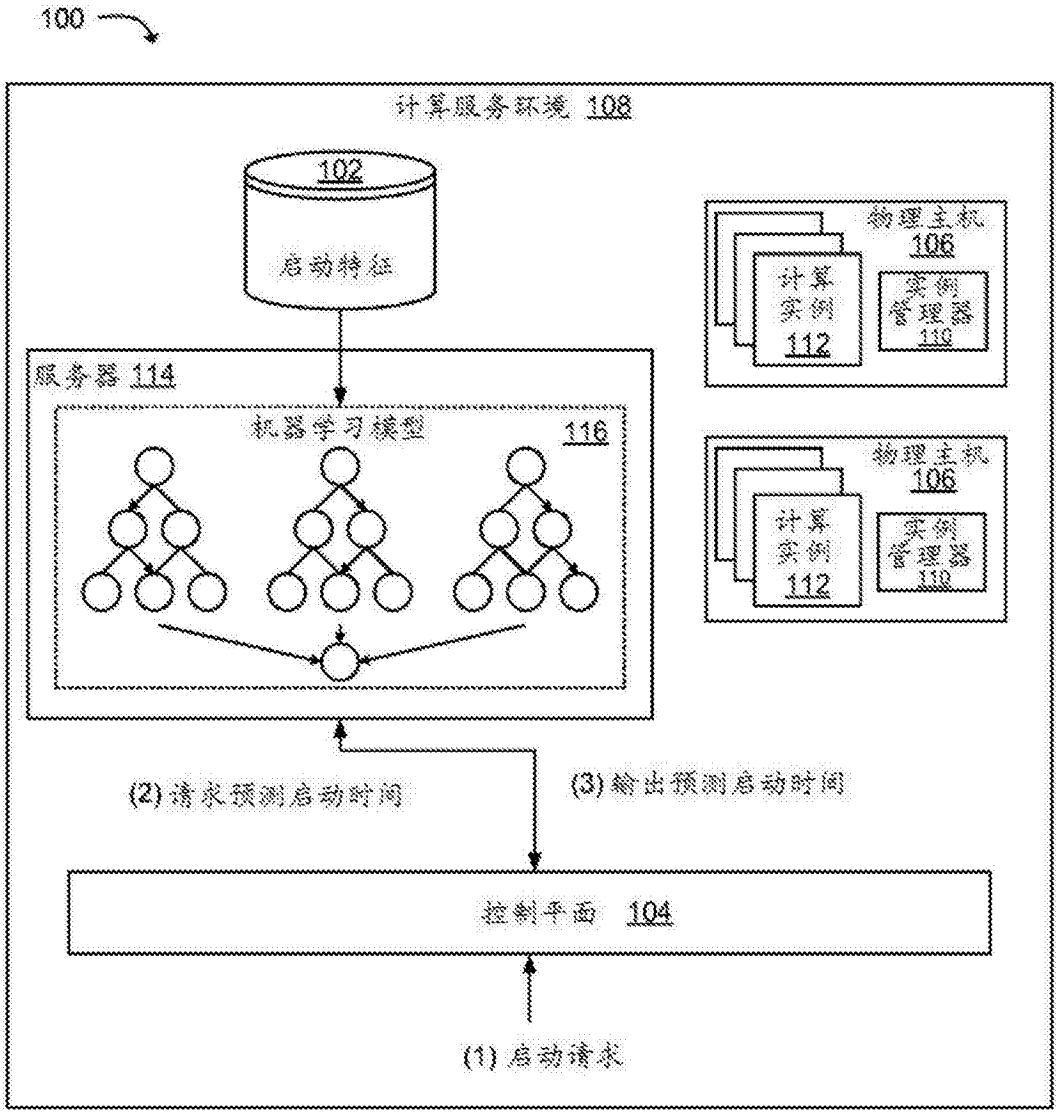

100
计算服务环境 108
102
启动特征
物理主机
106
计算
实例
112
实例
管理器
110
服务器 114
机器学习模型
116
物理主机
106
计算
实例
112
实例
管理器
110
(2)请求预测启动时间
(3)输出预测启动时间
控制平面 104
(1)启动请求

图1

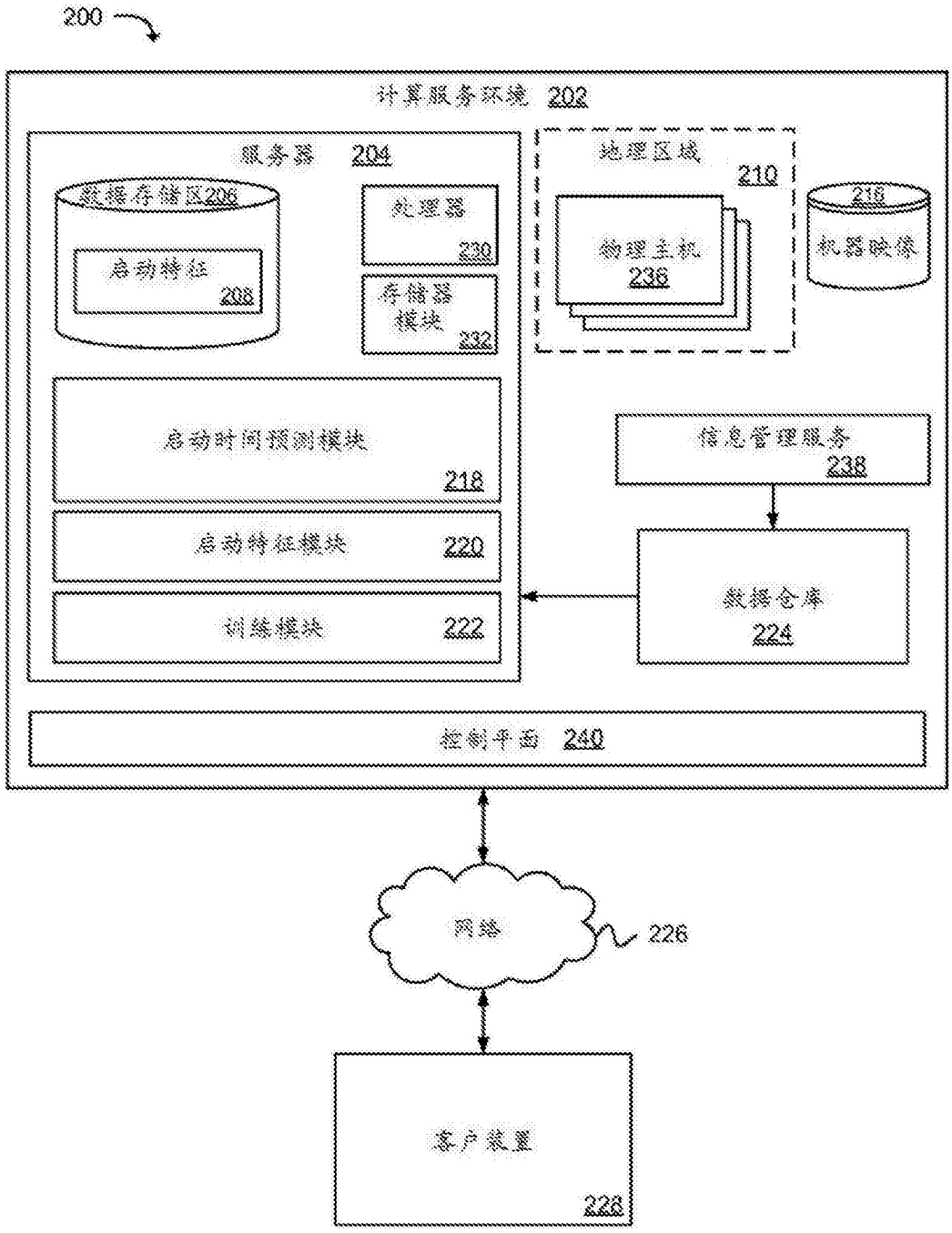

200
计算服务环境 202
服务器 204
数据存储区206
启动特征 208
处理器 230
存储器模块 232
地理区域 210
物理主机 236
216
机器映像
启动时间预测模块 218
启动特征模块 220
训练模块 222
信息管理服务 238
数据仓库 224
控制平面 240
网络 226
客户装置 228

图2

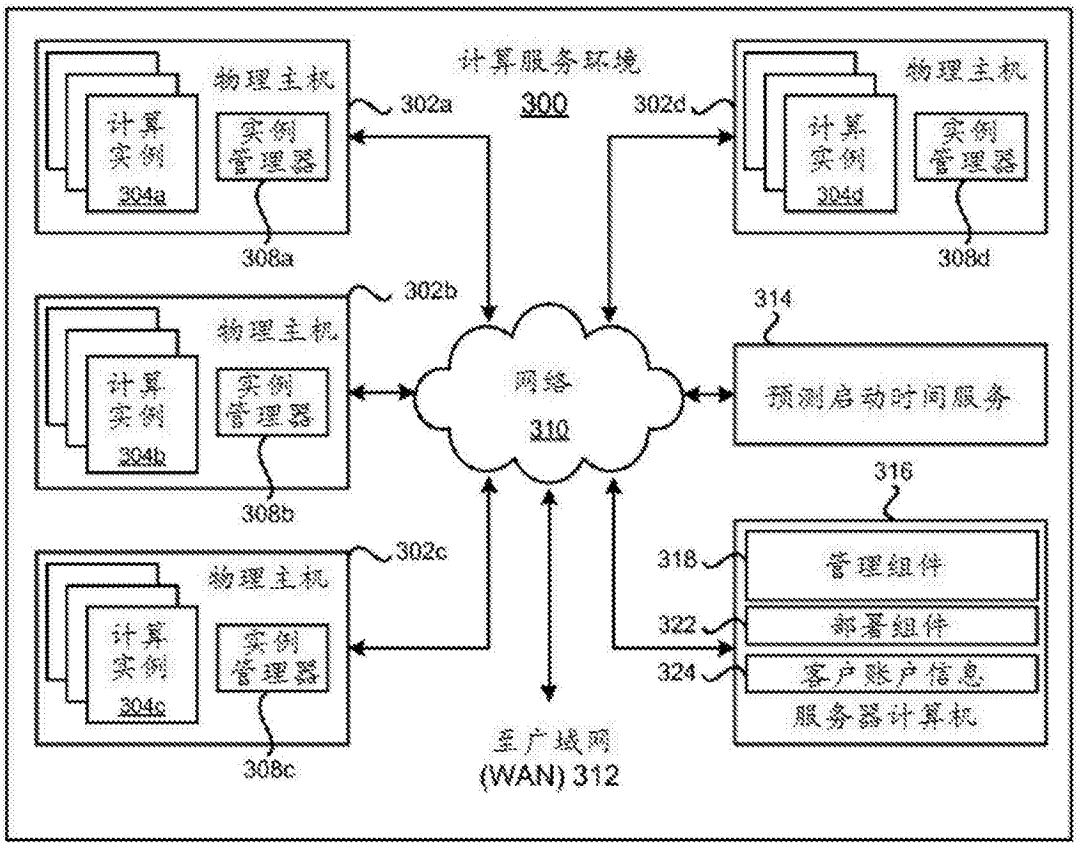

计算服务环境
300
物理主机
计算实例
304a
实例管理器
308a
302a
物理主机
计算实例
304d
实例管理器
308d
302d
物理主机
计算实例
304b
实例管理器
308b
302b
网络
310
314
预测启动时间服务
316
物理主机
计算实例
304c
实例管理器
308c
302c
318
管理组件
322
部署组件
324
客户账户信息
服务器计算机
互广域网
(WAN) 312

图3

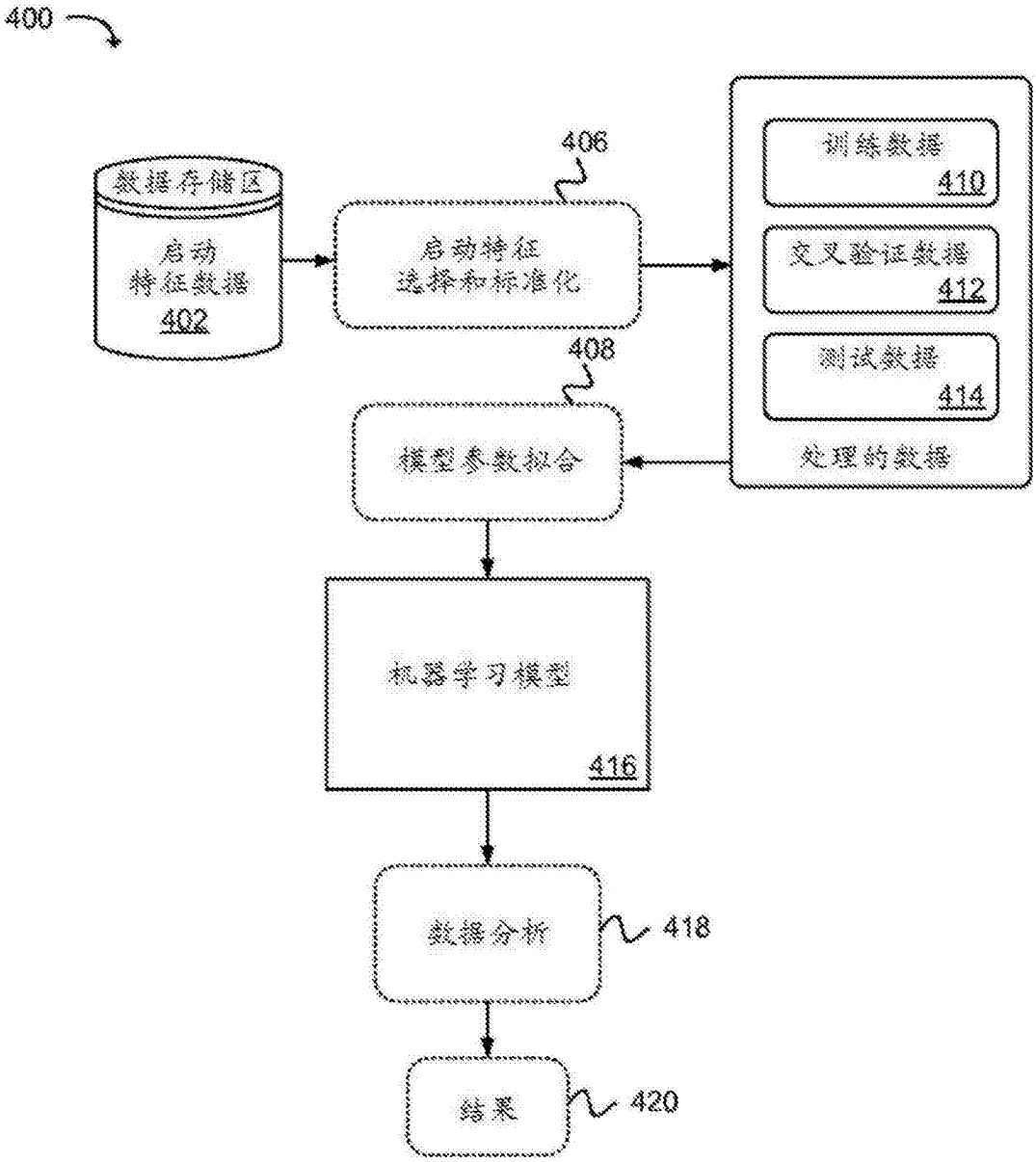

400
数据存储区
启动
特征数据
402
406
启动特征
选择和标准化
训练数据
410
交叉验证数据
412
测试数据
414
处理的数据
408
模型参数拟合
机器学习模型
416
数据分析
418
结果
420

图4

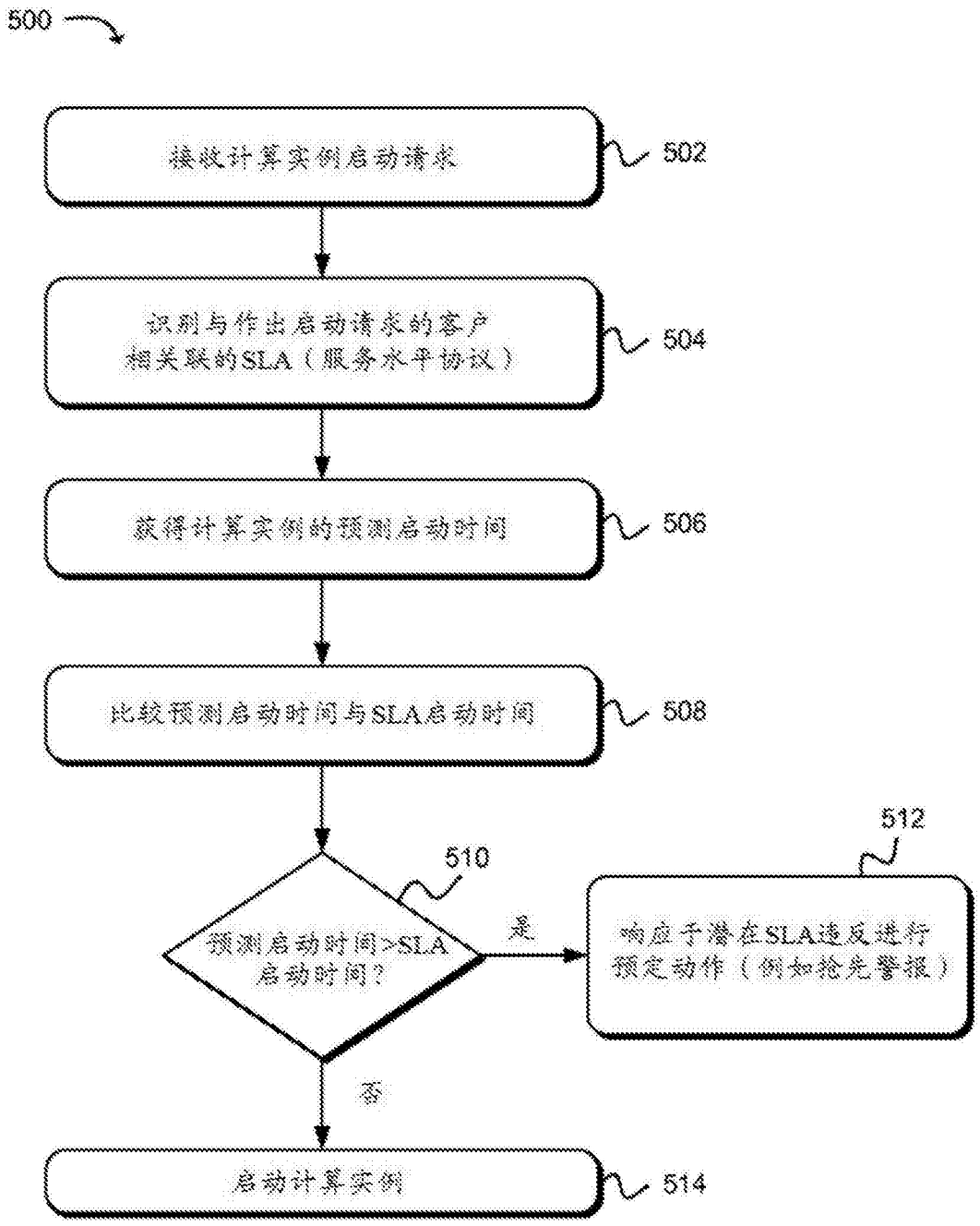

500
接收计算实例启动请求
502
识别与作出启动请求的客户相关联的SLA（服务水平协议）
504
获得计算实例的预测启动时间
506
比较预测启动时间与SLA启动时间
508
510
预测启动时间>SLA启动时间？
是
512
响应于潜在SLA违反进行预定动作（例如抢先警报）
否
启动计算实例
514

图5

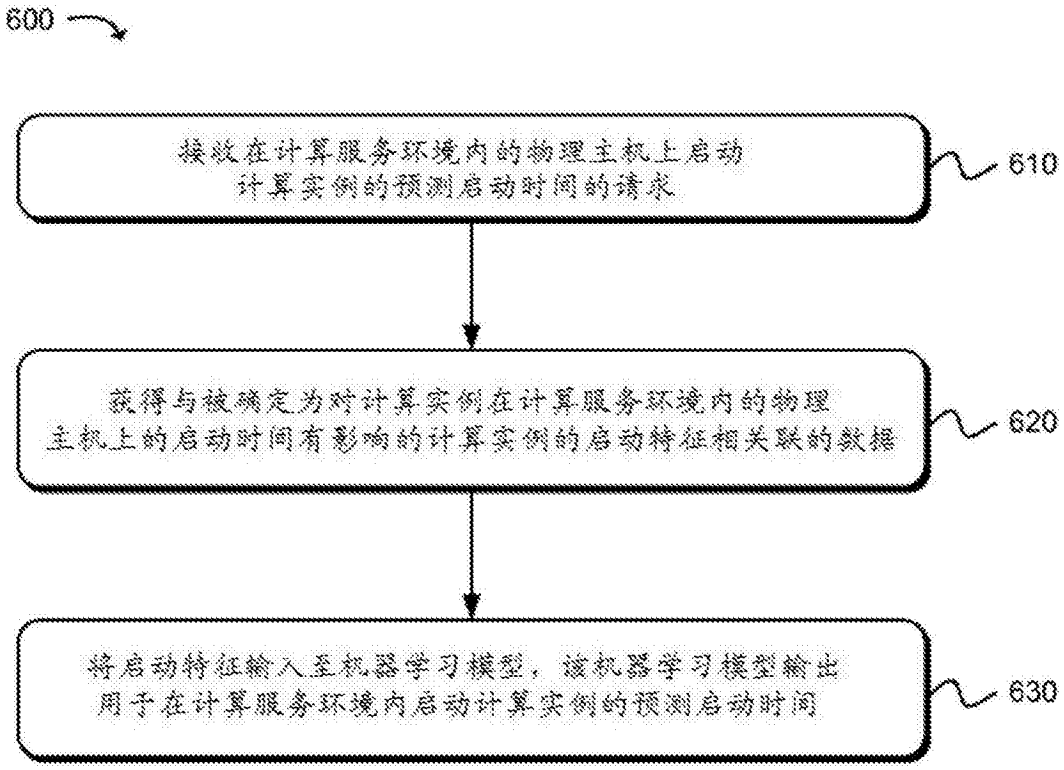

600
接收在计算服务环境内的物理主机上启动计算实例的预测启动时间的请求
610
获得与被确定为对计算实例在计算服务环境内的物理主机上的启动时间有影响的计算实例的启动特征相关联的数据
620
将启动特征输入至机器学习模型，该机器学习模型输出用于在计算服务环境内启动计算实例的预测启动时间
630

图6

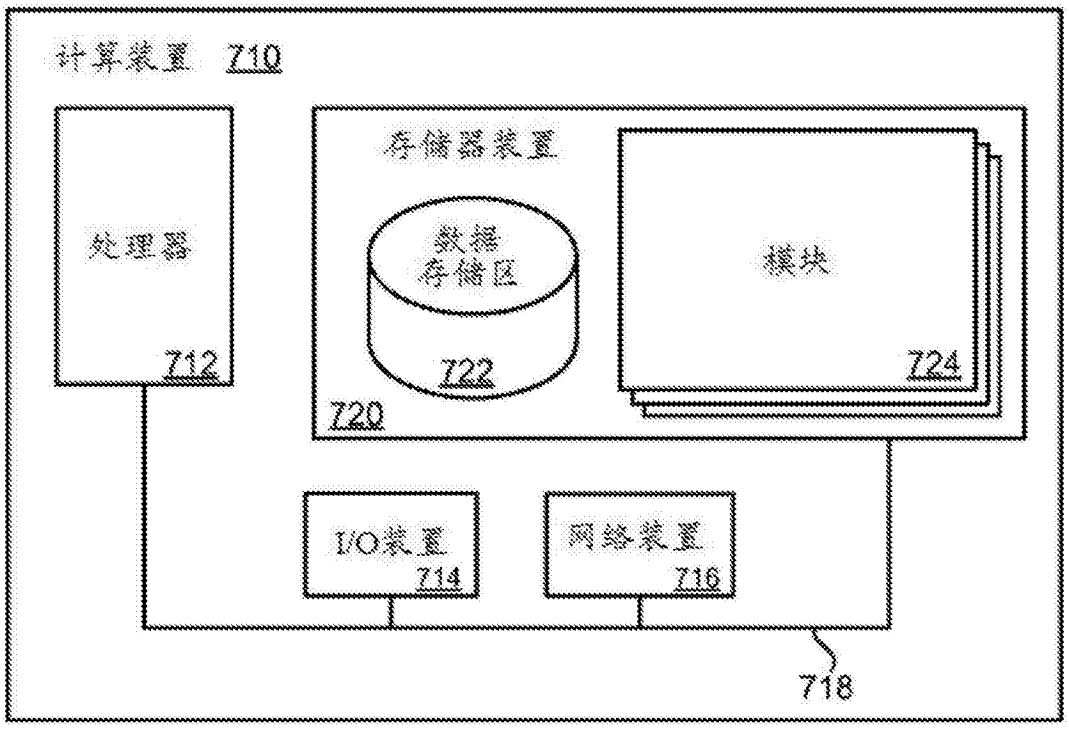

计算装置 710
处理器
712
存储器装置
数据存储区
722
720
模块
724
I/O装置
714
网络装置
716
718

图7

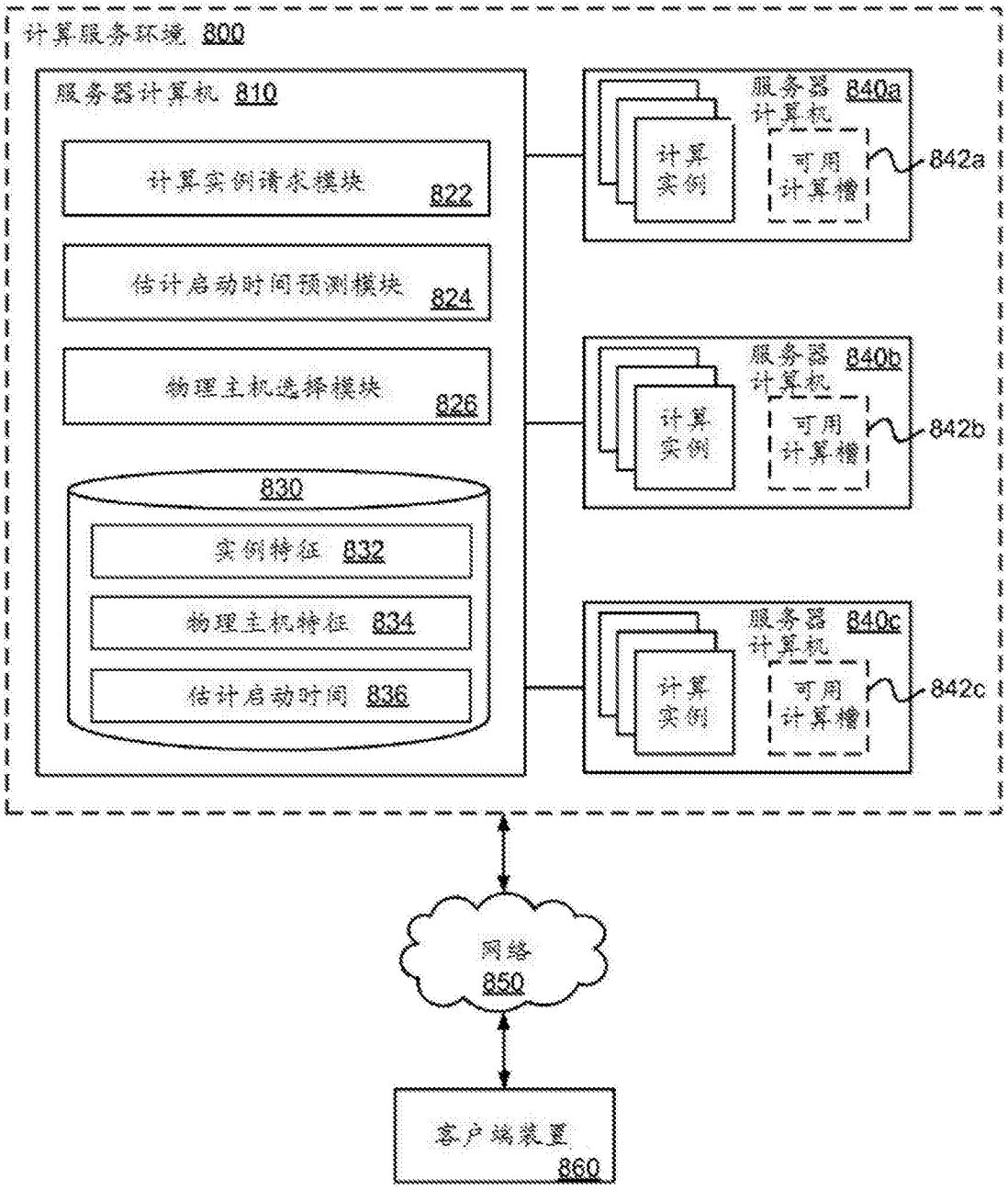

计算服务环境 800
服务器计算机 810
计算实例请求模块 822
估计启动时间预测模块 824
物理主机选择模块 826
830
实例特征 832
物理主机特征 834
估计启动时间 836
服务器计算机 840a
计算实例
可用计算槽
842a
服务器计算机 840b
计算实例
可用计算槽
842b
服务器计算机 840c
计算实例
可用计算槽
842c
网络 850
客户端装置 860

图8

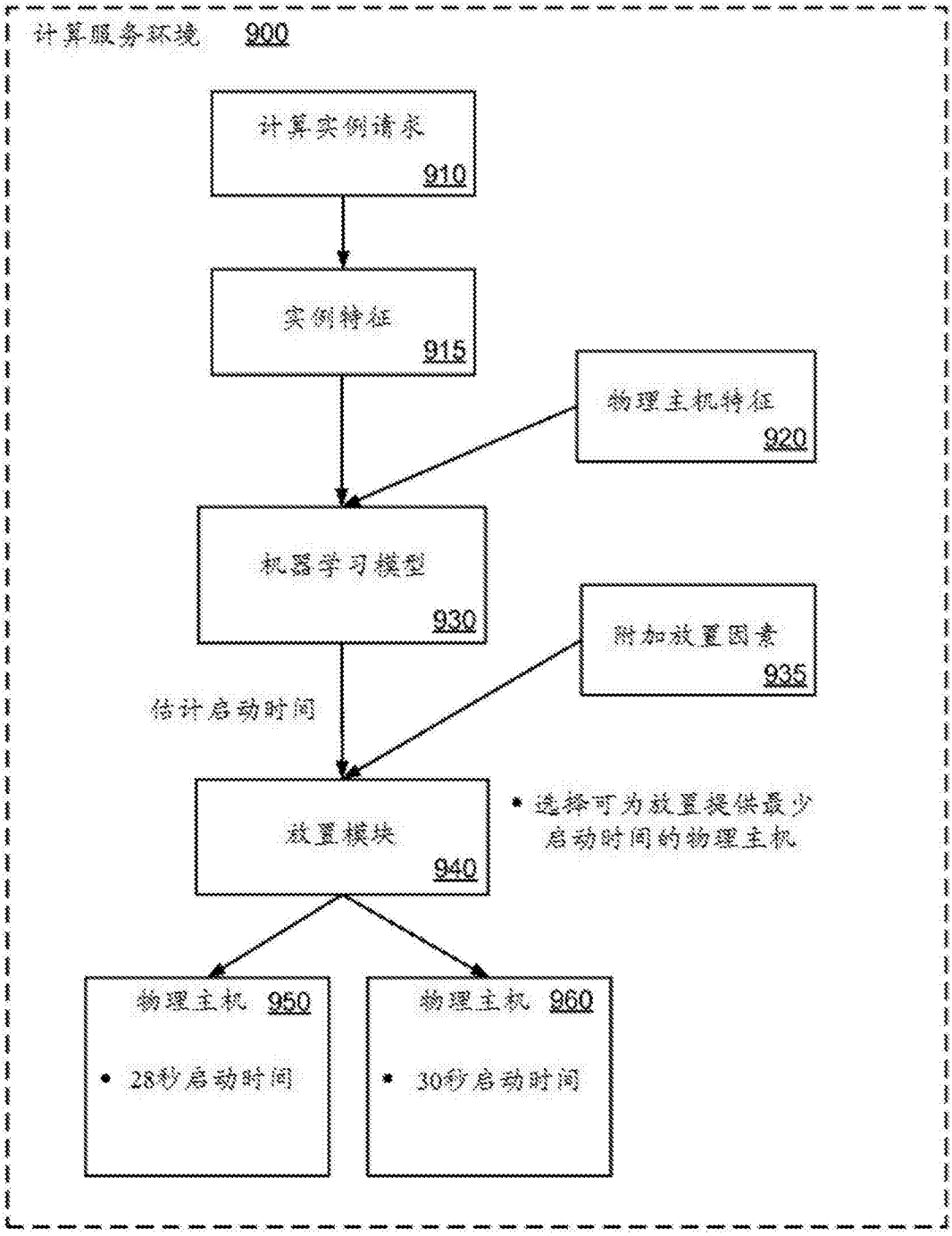

计算服务环境 900
计算实例请求 910
实例特征 915
物理主机特征 920
机器学习模型 930
附加放置因素 935
估计启动时间
放置模块 940
* 选择可为放置提供最少启动时间的物理主机
物理主机 950
• 28秒启动时间
物理主机 960
• 30秒启动时间

图9

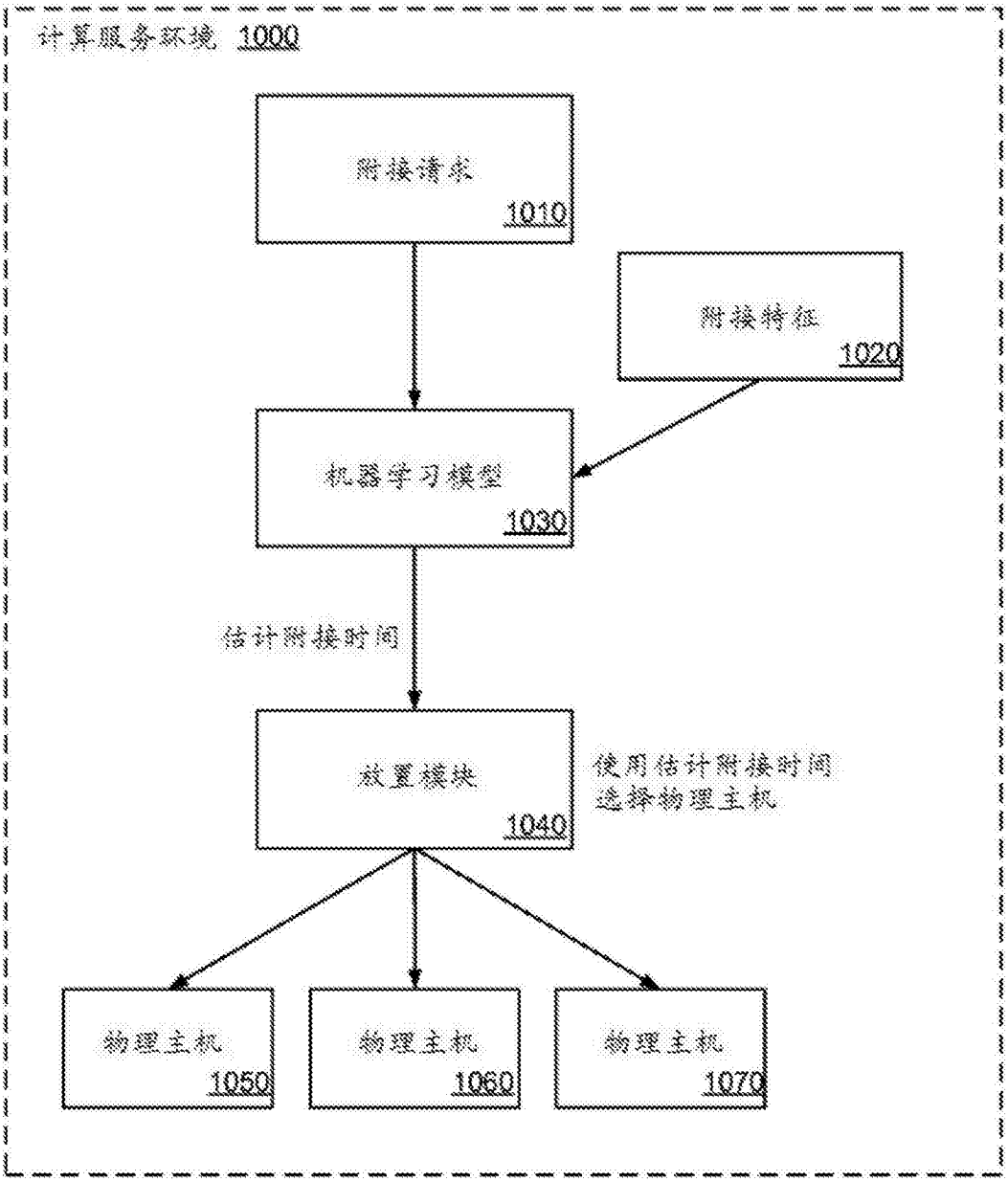

计算服务环境 1000
附接请求
1010
附接特征
1020
机器学习模型
1030
估计附接时间
放置模块
1040
使用估计附接时间
选择物理主机
物理主机
1050
物理主机
1060
物理主机
1070

图10

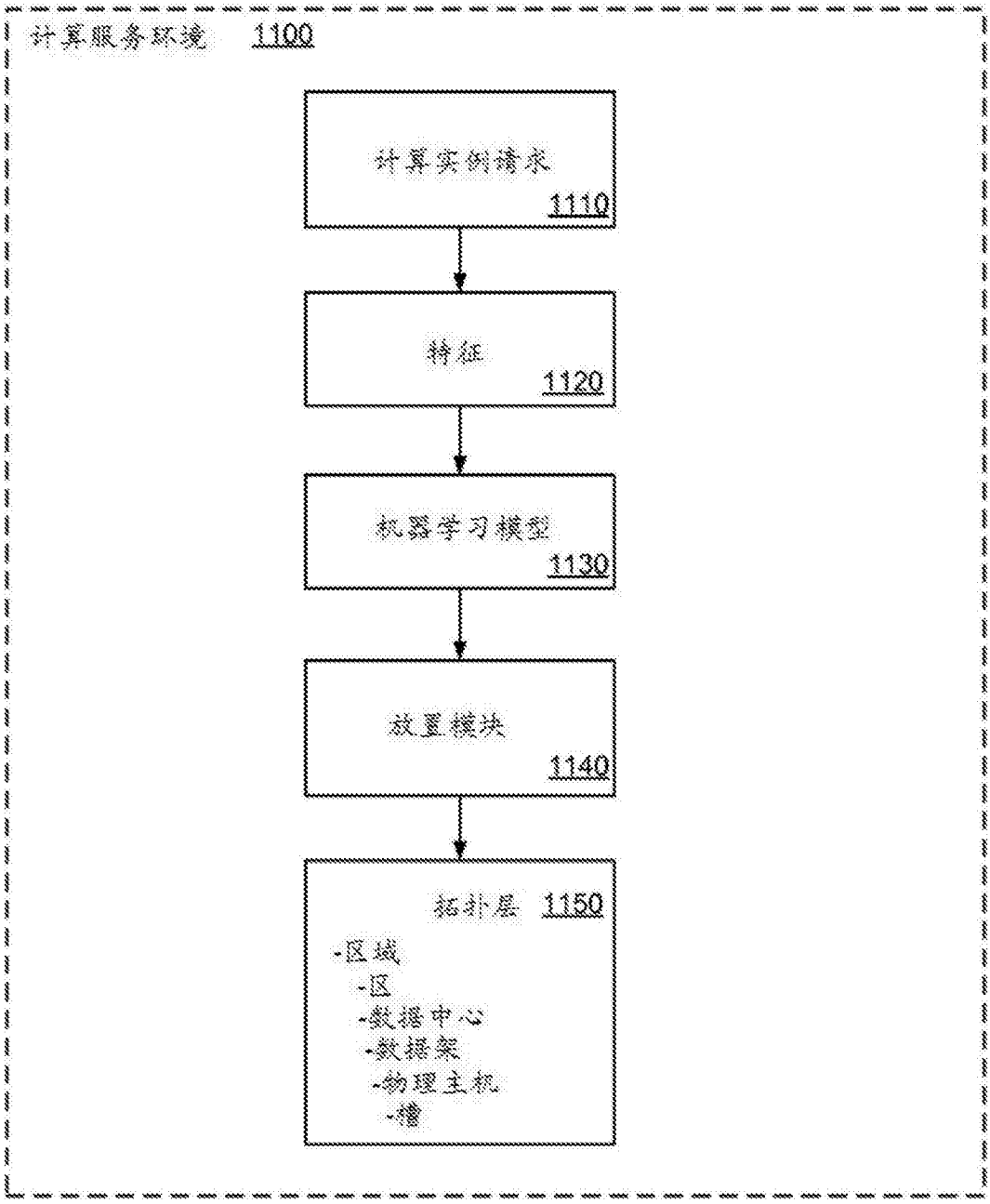

计算服务环境 1100
计算实例请求
1110
特征
1120
机器学习模型
1130
放置模块
1140
拓扑层 1150
-区域
-区
-数据中心
-数据架
-物理主机
-槽

图11

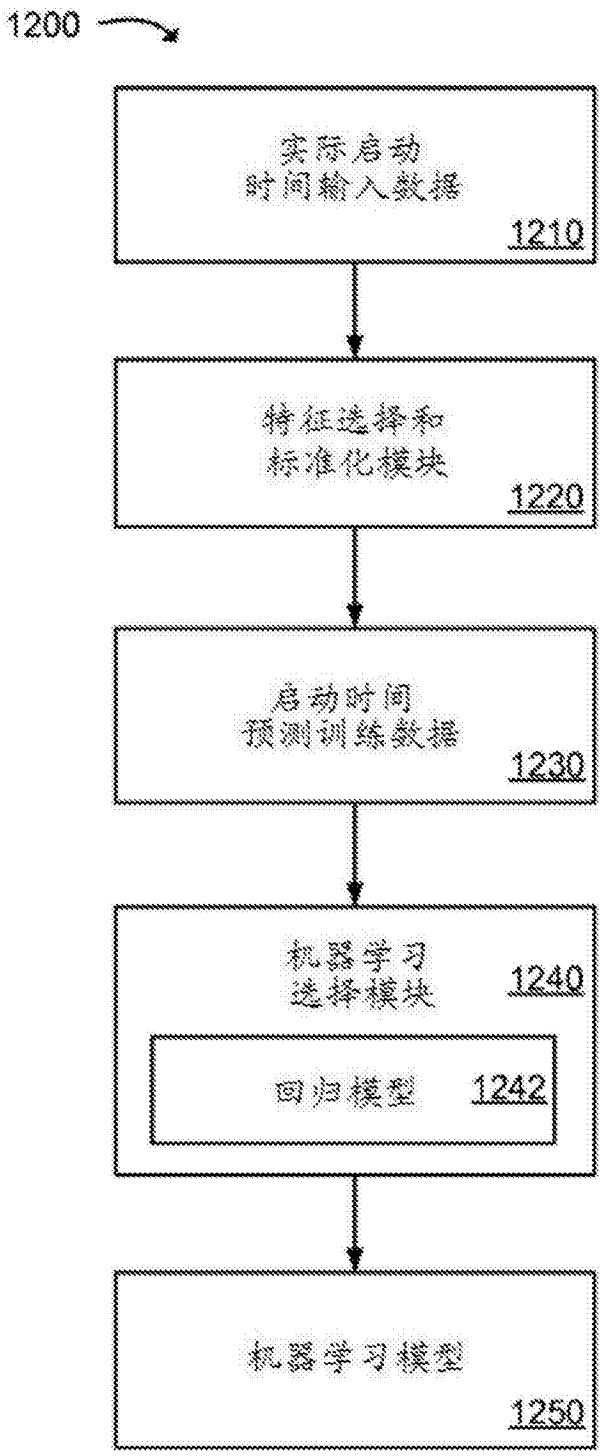

1200
实际启动
时间输入数据
1210
特征选择和
标准化模块
1220
启动时间
预测训练数据
1230
机器学习
选择模块
1240
回归模型
1242
机器学习模型
1250

图12

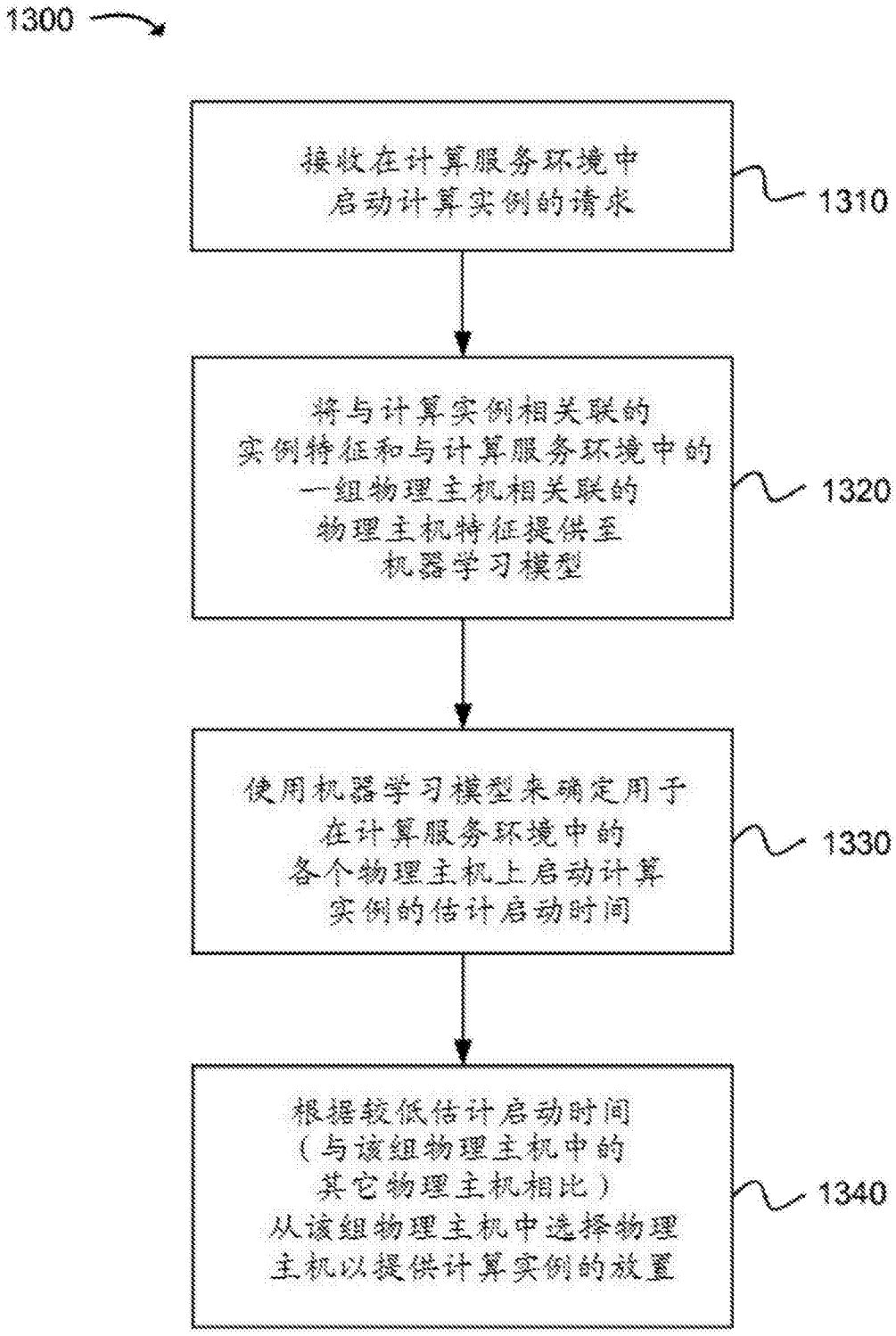

1300
接收在计算服务环境中
启动计算实例的请求
1310
将与计算实例相关联的
实例特征和与计算服务环境中的
一组物理主机相关联的
物理主机特征提供至
机器学习模型
1320
使用机器学习模型来确定用于
在计算服务环境中的
各个物理主机上启动计算
实例的估计启动时间
1330
根据较低估计启动时间
(与该组物理主机中的
其它物理主机相比)
从该组物理主机中选择物理
主机以提供计算实例的放置
1340

图13

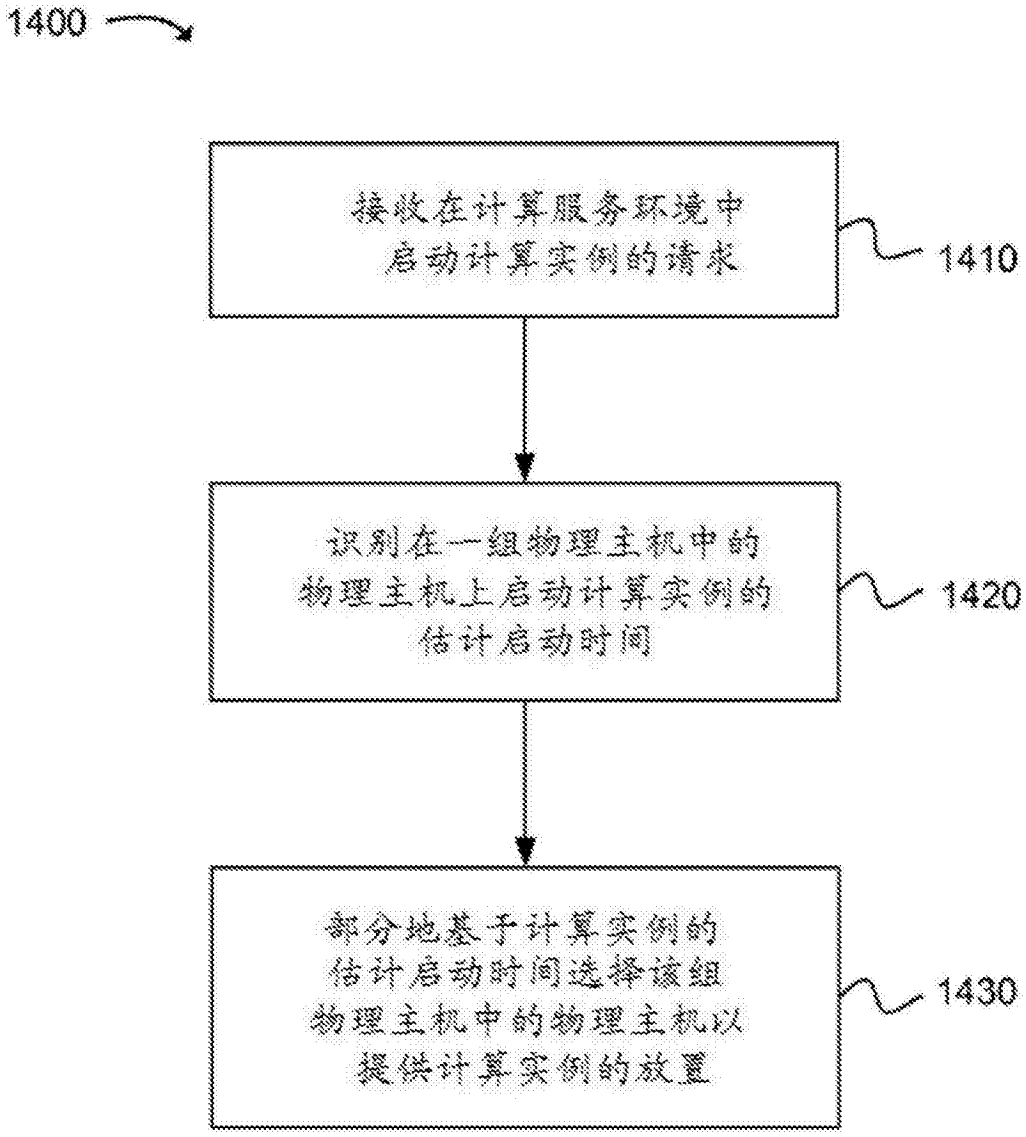

1400
接收在计算服务环境中启动计算实例的请求
1410
识别在一组物理主机中的物理主机上启动计算实例的估计启动时间
1420
部分地基于计算实例的估计启动时间选择该组物理主机中的物理主机以提供计算实例的放置
1430

图14

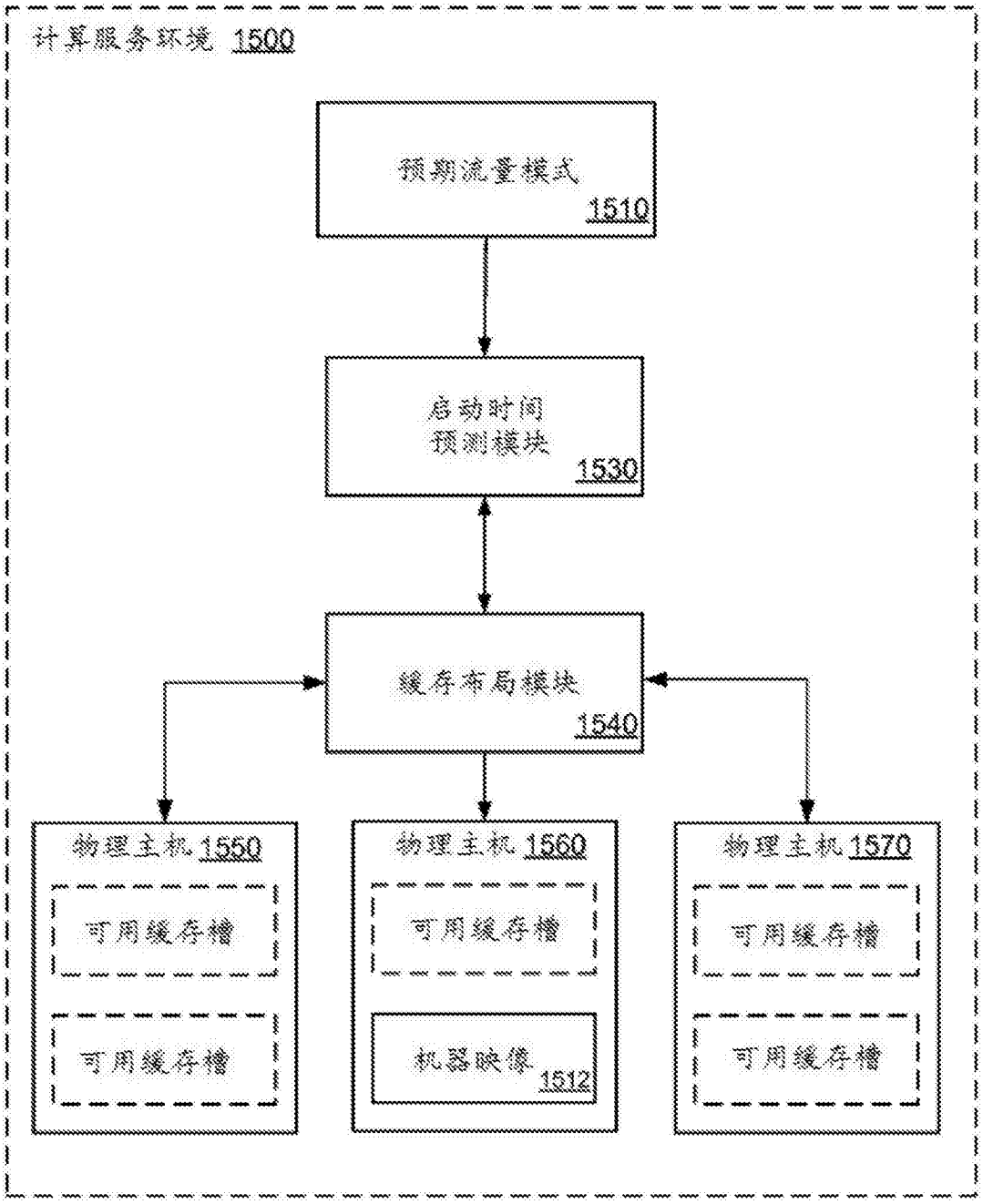

计算服务环境 1500
预期流量模式 1510
启动时间预测模块 1530
缓存布局模块 1540
物理主机 1550
可用缓存槽
可用缓存槽
物理主机 1560
可用缓存槽
机器映像 1512
物理主机 1570
可用缓存槽
可用缓存槽

图15

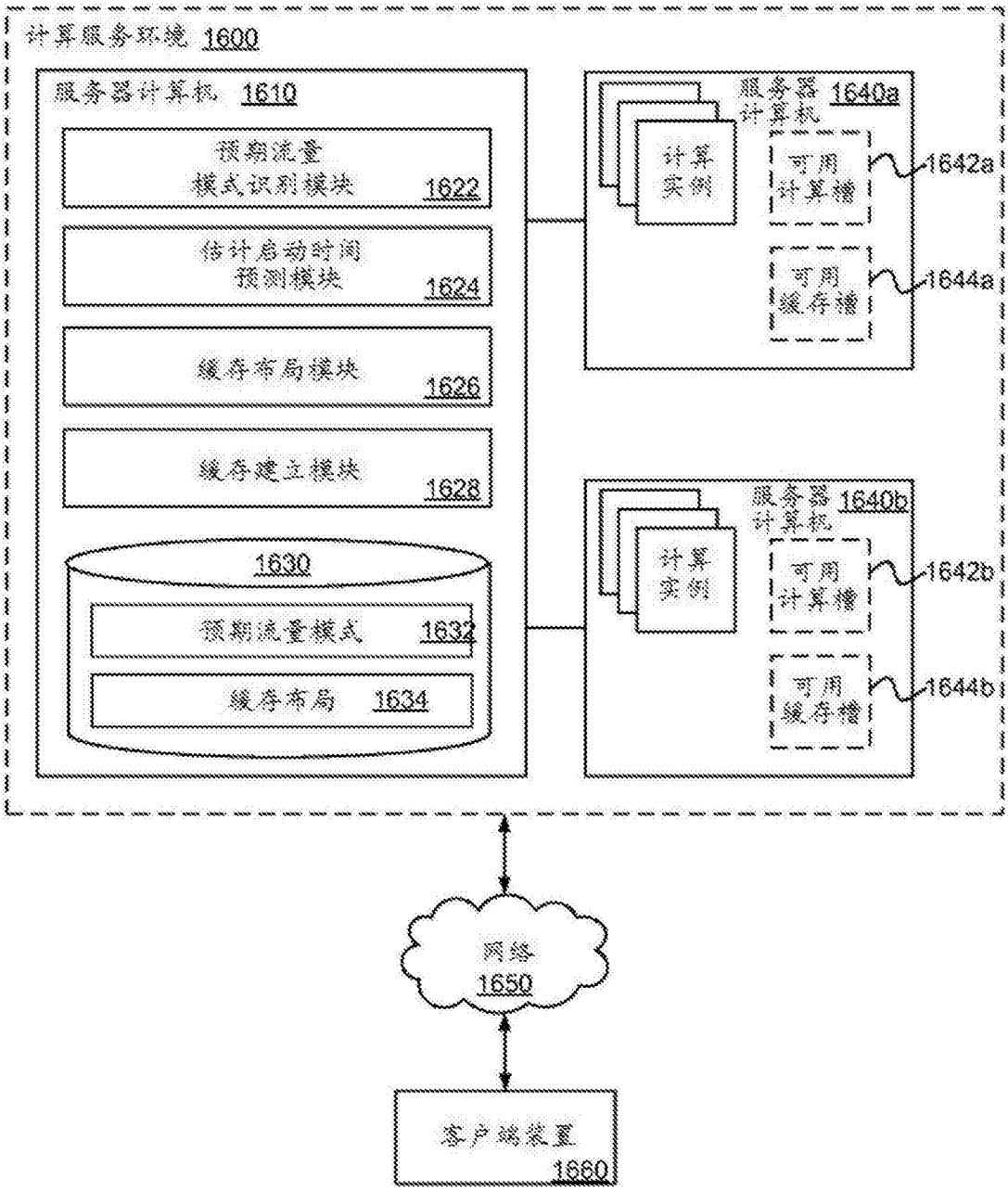

计算服务环境 1600
服务器计算机 1610
预期流量模式识别模块 1622
估计启动时间预测模块 1624
缓存布局模块 1626
缓存建立模块 1628
1630
预期流量模式 1632
缓存布局 1634
服务器计算机 1640a
计算实例
可用计算槽
1642a
可用缓存槽
1644a
服务器计算机 1640b
计算实例
可用计算槽
1642b
可用缓存槽
1644b
网络 1650
客户端装置 1660

图16

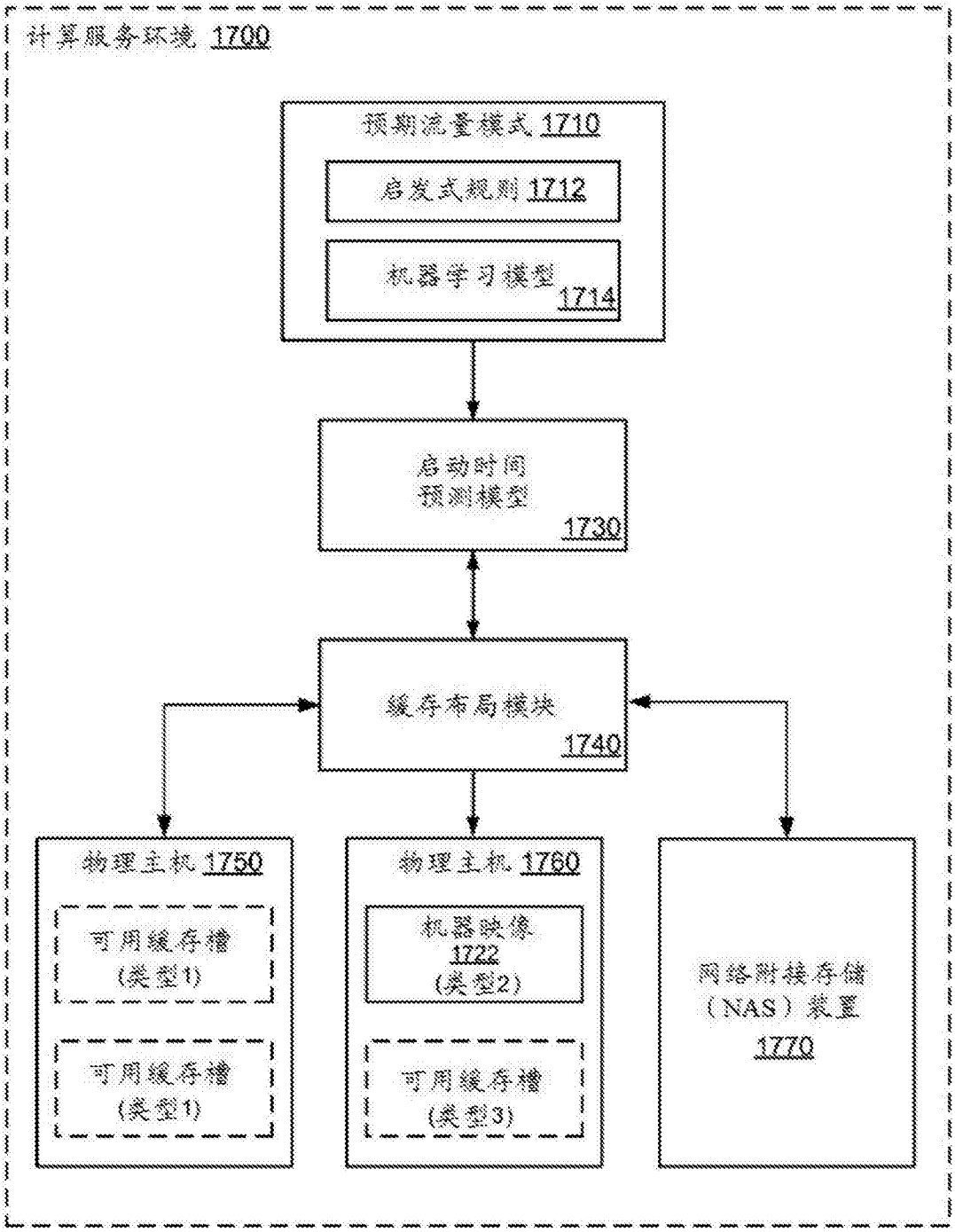

计算服务环境 1700
预期流量模式 1710
启发式规则 1712
机器学习模型 1714
启动时间预测模型 1730
缓存布局模块 1740
物理主机 1750
可用缓存槽 (类型1)
可用缓存槽 (类型1)
物理主机 1760
机器映像 1722 (类型2)
可用缓存槽 (类型3)
网络附接存储 (NAS) 装置 1770

图17

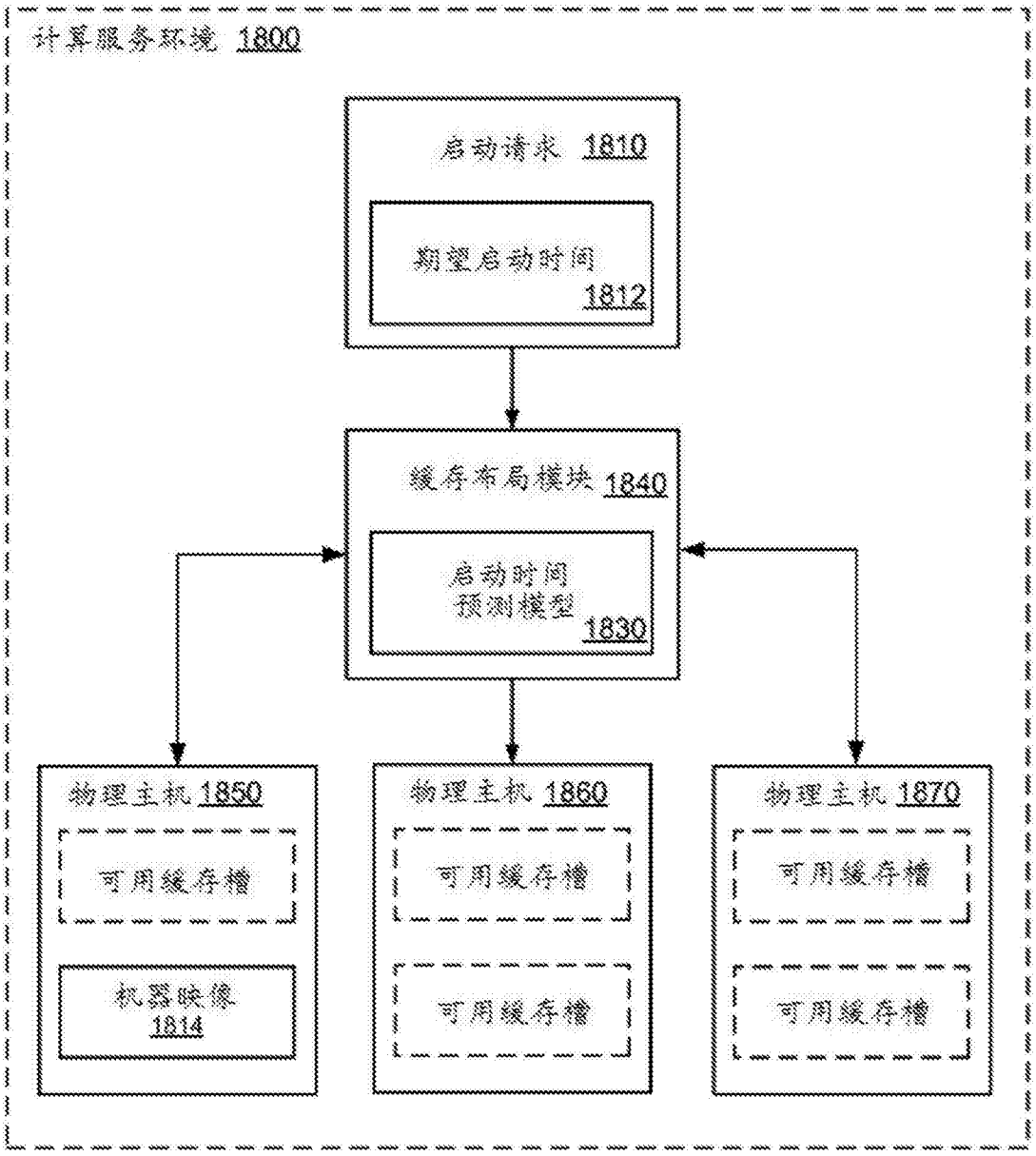

计算服务环境 1800
启动请求 1810
期望启动时间 1812
缓存布局模块 1840
启动时间预测模型 1830
物理主机 1850
可用缓存槽
机器映像 1814
物理主机 1860
可用缓存槽
可用缓存槽
物理主机 1870
可用缓存槽
可用缓存槽

图18

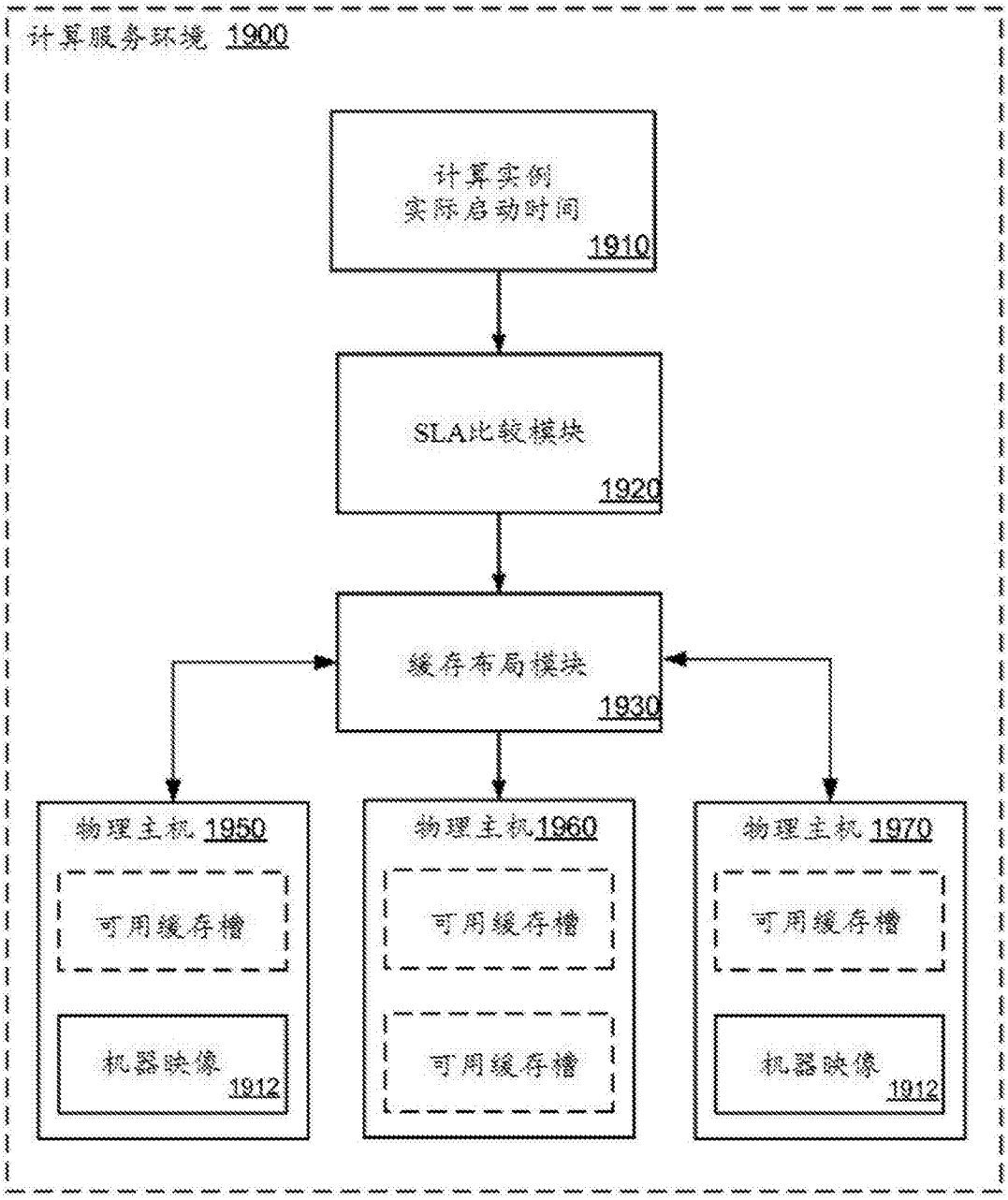

计算服务环境 1900
计算实例
实际启动时间
1910
SLA比较模块
1920
缓存布局模块
1930
物理主机 1950
可用缓存槽
机器映像
1912
物理主机 1960
可用缓存槽
可用缓存槽
物理主机 1970
可用缓存槽
机器映像
1912

图19

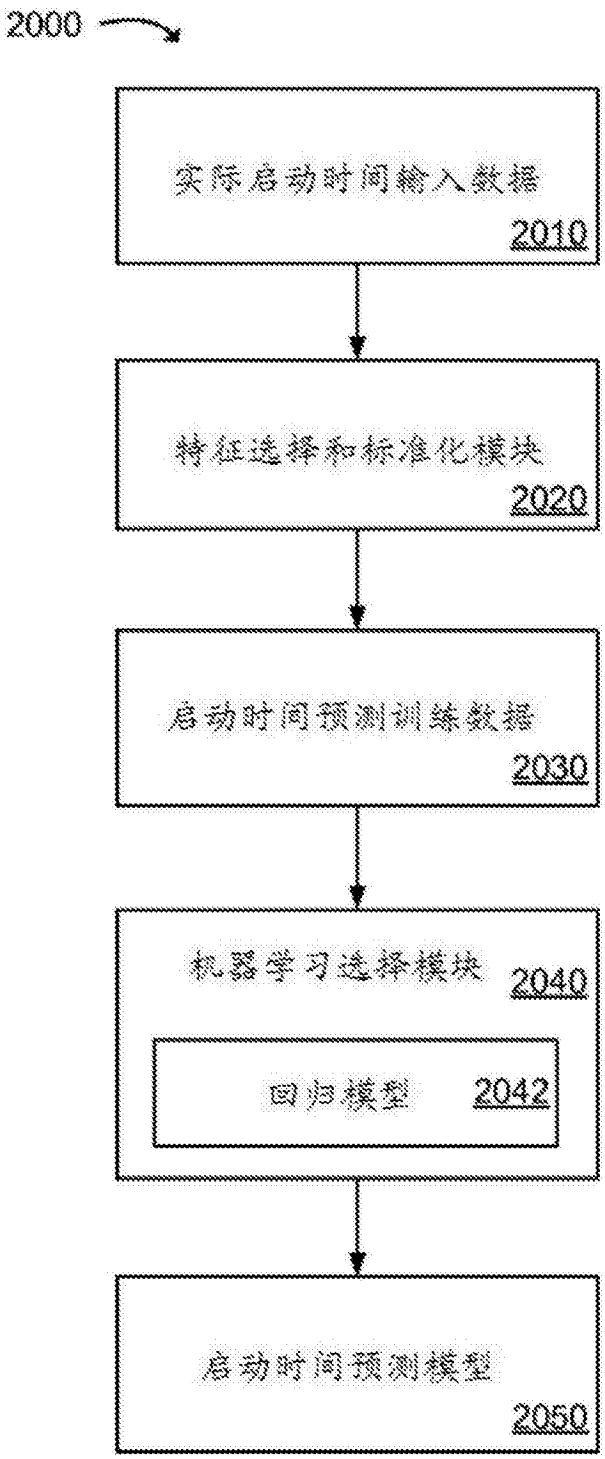

2000
实际启动时间输入数据
2010
特征选择和标准化模块
2020
启动时间预测训练数据
2030
机器学习选择模块
2040
回归模型
2042
启动时间预测模型
2050

图20

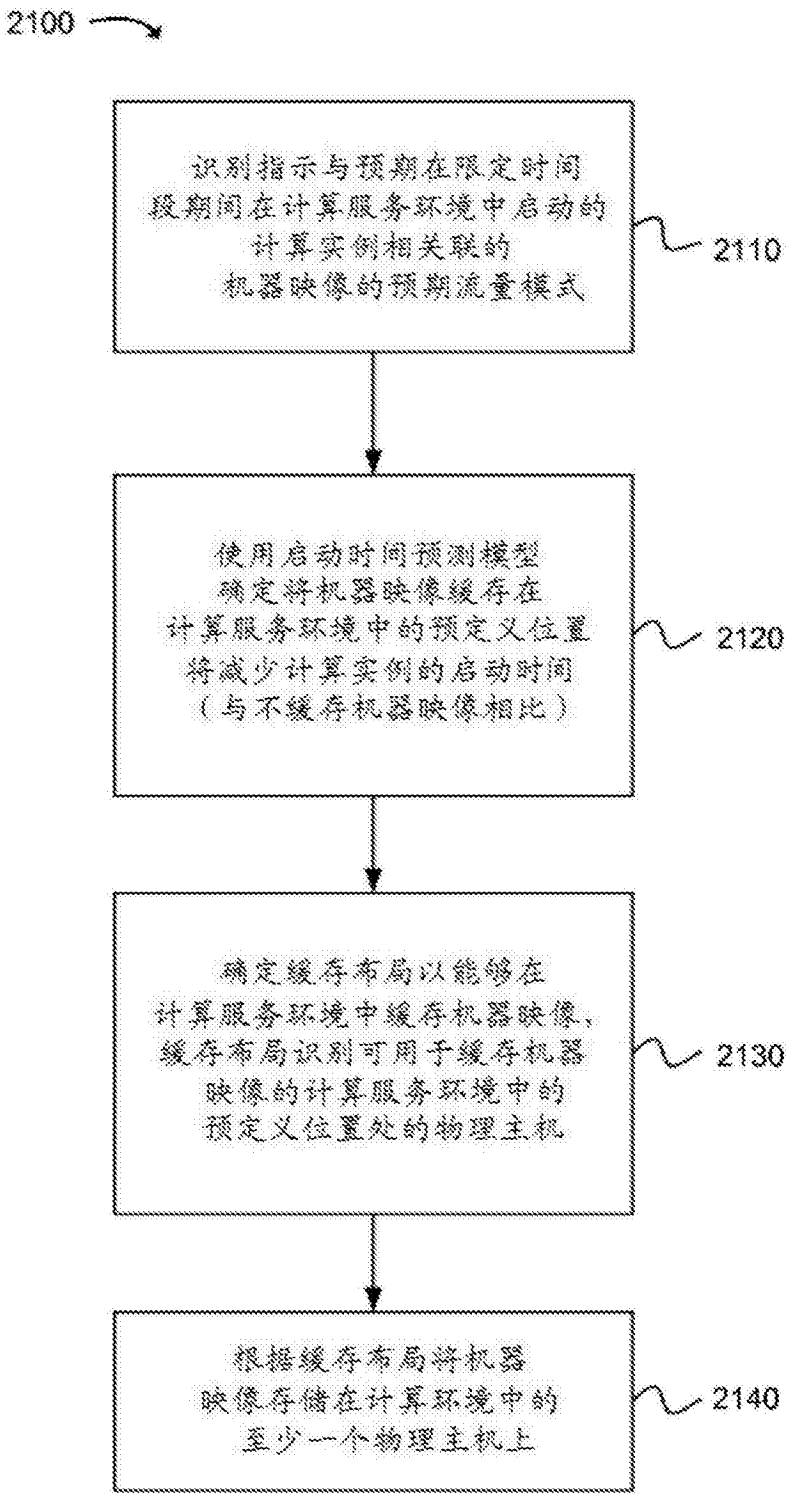

2100
识别指示与预期在限定时间段期间在计算服务环境中启动的计算实例相关联的机器映像的预期流量模式
2110
使用启动时间预测模型确定将机器映像缓存在计算服务环境中的预定义位置将减少计算实例的启动时间（与不缓存机器映像相比）
2120
确定缓存布局以能够在计算服务环境中缓存机器映像，缓存布局识别可用于缓存机器映像的计算服务环境中的预定义位置处的物理主机
2130
根据缓存布局将机器映像存储在计算环境中的至少一个物理主机上
2140

图21

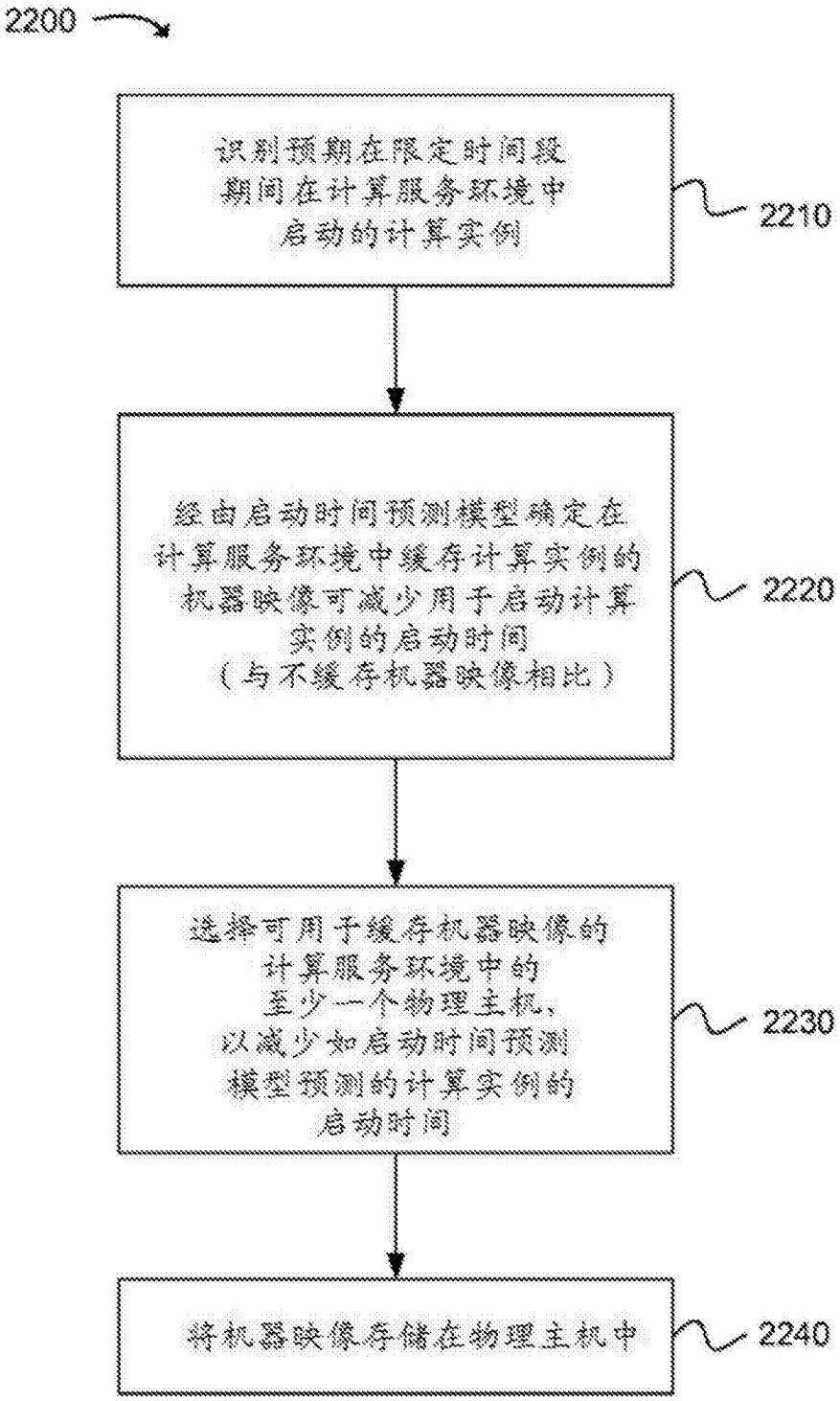

2200
识别预期在限定时间段期间在计算服务环境中启动的计算实例
2210
经由启动时间预测模型确定在计算服务环境中缓存计算实例的机器映像可减少用于启动计算实例的启动时间（与不缓存机器映像相比）
2220
选择可用于缓存机器映像的计算服务环境中的至少一个物理主机，以减少如启动时间预测模型预测的计算实例的启动时间
2230
将机器映像存储在物理主机中
2240

图22