

一种提供云服务的方法、装置及设备

技术领域

本发明属于云计算技术领域，尤其涉及一种提供云服务的方法、装置以及设备。

背景技术

云计算可以将基础硬件设施进行统一的虚拟化和池化，将这些虚拟化资源组成虚拟机（virtual Machine, VM）。其中，虚拟化的资源包括 CPU、内存、网络、存储等。通过这种虚拟化技术，可以在一台物理机上虚拟化出多台相互隔离的虚拟机，云服务供应商把这些虚拟机售卖给租户。通过配置不同的内存配比、不同磁盘、不同类别的网卡、不同的 CPU 硬件等规格，云服务商可以提供不同性能的计算产品，满足租户不同的需求。

发明内容

本申请提出了一种提供云服务的方法、装置、设备、计算机可读存储介质以及计算机程序产品，能有效提升虚拟机购买成功率和客户体验。

第一方面，本申请提出了一种提供云服务的方法。该方法包括：接收终端发送的业务请求，该业务请求中包括该用户所需的资源总量信息；向终端发送一种或多种云服务方案，该一种或多种云服务方案中的至少一种云服务方案中包括多个虚拟机，所述一种或多种云服务方案中虚拟机的资源总量满足所述业务请求中的资源总量。

本方法无需客户指定具体虚拟机的规格和数量，而是根据用户输入需要的资源总量，由云平台直接为用户提供一种或多种云服务方案（云服务方案中包含多台虚拟机）以满足用户需求。有效避免了现有技术中购买虚拟机时由于库存不足导致购买失败的问题。

可选的，该云服务方案中包括用户需要购买的虚拟机规格和以及相应数量。

第一方面一种可能的设计，该资源总量信息包括下述信息的至少一种：CPU 总量信息和内存总量信息。

CPU 总量信息和内存总量信息可以有效描述用户的业务请求中的资源总量信息，云平台根据用户输入的 CPU 总量信息和内存总量信息为用户提供满足该需求的云服务方案。

第一方面一种可能的设计，该业务请求中还包括资源性能信息，其中，该一种或多种云服务方案中虚拟机的资源性能满足该业务请求中的资源性能信息。

资源性能信息用于描述资源的性能特性，用户可以根据自身需求对资源的性能进行进一步限定，从而购买到更适合的虚拟机。

第一方面一种可能的设计，该资源性能信息包括底层资源性能参数，其中该底层资源性能参数包括下述至少一种：CPU 型号、CPU 频率、内存型号、内存频率、GPU 参数、网络性能以及存储读写性能。

上述列举的底层资源性能参数可以用于限定用户输入的资源总量中的 CPU 和内存的具体性能，能够帮助用户更精准的筛选出满足资源参数的虚拟机。

第一方面一种可能的设计，该资源性能信息包括下述信息的至少一种：虚拟机能力类型和业务场景类型，其中，该虚拟机能力类型用于指示虚拟机的计算、存储和网络能力，该业务场景类型用于指示虚拟机适用的业务场景。

用户还可以通过选择或输入虚拟机能力类型和业务场景类型来对资源的性能做进一步限定，虚拟机能力类型和业务场景类型更加易懂，能帮助用户更加直观的确定自己所需的资源

性能信息。

第一方面一种可能的设计，该 CPU 总量信息包括下述信息的至少一种：CPU 总量、CPU 总量的最大值、CPU 总量的最小值；该内存总量信息包括下述信息的至少一种：内存总量、内存总量的最大值、内存总量的最小值。

用户可以通过 CPU 总量、CPU 总量的最大值、CPU 总量的最小值、内存总量、内存总量的最大值、内存总量的最小值等输入信息，更加灵活的限定所需的资源总量，云平台可根据这些信息提供更多种的云服务方案的可能。

第一方面一种可能的设计，该业务请求中还包括用户指定的价格信息，该一种或多种云服务方案的价格满足该用户指定的价格信息。

云平台可根据用户指定的价格信息，筛选出符合用户理想价位的云服务方案，能提升用户的购买体验。

第一方面一种可能的设计，根据该用户从所述一种或多种云服务方案选择的目标云服务方案，向该用户提供云服务。

用户可以在终端上选中一种云服务方案，云平台基于该选中的方案为用户提供相应的购买服务。

可选的，用户还可以对选中的云服务方案进行修改。

第二方面，本申请提供了一种提供云服务的装置。该装置包括的多个功能模块，用于实现第一方面或第一方面的任意可能设计提供的方法的不同步骤。

第三方面，本申请提供一种设备，该设备包括处理器和存储器。该处理器执行该存储器存储的指令，使得该设备执行上述第一方面或者第一方面的各种可能设计提供的方法。

本申请提供一种设备，该设备包括处理器和存储器。该处理器执行该存储器存储的指令，使得该设备部署第二方面提供的装置。

第四方面，本申请提供一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当设备的处理器执行该指令时，该设备执行上述第一方面或者第一方面的各种可能设计提供的方法。

本申请提供一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当设备的处理器执行该指令时，该服务器实现第二方面提供的装置。

第五方面，本申请提供一种计算机程序产品，该计算机程序产品包括指令，该指令存储在计算机可读存储介质中。设备的处理器可以从计算机可读存储介质读取该指令；该处理器执行该指令，使得该设备执行上述第一方面或者第一方面的各种可能设计提供的方法。

本申请提供一种计算机程序产品，该计算机程序产品包括指令，该指令存储在计算机可读存储介质中。设备的处理器可以从计算机可读存储介质读取该指令；该处理器执行该指令，该设备实现第二方面提供的装置。

附图说明

图 1 是本申请实施例提供的一种网络架构举例示意图。

图 2 是本申请实施例提供的一种提供云服务的方法流程举例示意图。

图 3 (a) 是本申请实施例提供的一种需求采集界面的举例示意图。

图 3 (b) 是本申请实施例提供的一种需求采集界面的举例示意图。

图 3 (c) 是本申请实施例提供的一种需求采集界面的举例示意图。

图 3 (d) 是本申请实施例提供的一种需求采集界面的举例示意图。

图 3 (e) 是本申请实施例提供的一种需求采集界面的举例示意图。

图 3 (f) 是本申请实施例提供的一种需求采集界面的举例示意图。

图 4 是本申请实施例提供的一种购买方式选择界面举例示意图。

图 5 是本申请实施例提供的一种云平台生成该云服务方案的方法流程举例示意图。

图 6 是本申请实施例提供的一种云服务方案返回界面的举例示意图。

图 7 是本申请实施例提供的一种装置 210 的结构举例示意图。

图 8 是本申请实施例提供了一种计算设备 300 的结构举例示意图。

具体实施方式

本发明实施例描述的网络架构以及业务场景是为了更加清楚的说明本发明实施例的技术方案，并不构成对于本发明实施例提供的技术方案的限定，本领域普通技术人员可知，随着网络架构的演变和新业务场景的出现，本发明实施例提供的技术方案对于类似的技术问题，同样适用。需要说明的是，在本发明实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行描述。

为了便于理解本申请实施例，首先，对本申请涉及的部分术语进行解释说明。

云平台：可以基于云计算技术，提供与云服务相关的页面以供租户远程访问云服务。租户可通过预先注册的账号密码在终端登录云平台（例如 huaweicloud.com），并在登录成功之后，在访问页面选择并购买对应的云服务，例如虚拟机、容器、OBS 等服务。例如，云服务创建实例成功后，租户就可以像使用自己的本地 PC 或物理服务器一样，在云平台上使用实例。

实例：实例是云平台上提供计算服务的最小单位，不同的实例规格可以提供的计算能力不同，可用于不同的业务场景。例如，云平台上提供的虚拟机产品就是实例的一种。

规格：云服务商会在云平台上发布一系列的实例规格，实例规格定义了实例的基本属性：CPU 和内存（包括 CPU 型号、主频等）。实际中，实例还需要配合存储、镜像和网络类型，才可以确定一台实例的具体服务形态。这些规格是云服务商根据用户需求、资源量、销售情况等方面发布的。例如，huaweicloud.com 上提供的 s6.large.2 就是一种机型名称/规格名称，它对应着包含 2vCPUs、4GB 内存，处理器为 Intel Cascade Lake 2.6GHz 带宽为 0.2/1.5 Gbit/s 的实例。

算力：可用于衡量实例的计算能力，即单位时间内完成单位任务的数量。实例的算力的大小和很多方面相关，例如处理器、内存、网络等等。

资源池：指云计算架构中的虚拟资源层，实质上是在多道应用作业实例和底层的物理资源设备或设备集群之间进行时分和空分的调度。虚拟资源层包括：计算虚拟化、存储虚拟化和网络虚拟化。以计算虚拟化举例，通过在上层软件与裸机硬件之间插入的一层弹性计算资源管理和虚拟化软件，可以使得从上层软件视角，它仍然是独占底层的 CPU、内存以及输入输出（Input/Output, I/O）资源。

价格：不同规格的实例的价格通常不一样。

图 1 是本申请实施例提供的一种网络架构举例示意图。

如图 1 所示，图中包括：终端 001、网络 002、云平台 003、数据中心 004、资源池 005。云平台的用户可在终端 001 通过网络 002 访问云平台 003。

终端 001 可以是手机、电脑、平板等其它任何终端设备，可以通过网络 002 和云平台通信，实现和云平台 003 之间的通讯。

终端访问云平台的方式举例：例如，终端 001 设备上安装了云平台 003 的客户端的软件，用户通过该客户端实现对云平台 003 的访问。再例如，终端 001 通过浏览器直接访问云平台 003 提供的网页。

网络 002 可以是诸如互联网这类的公有网络、诸如机构内网的私有网络，或者是公有网络和私有网络共存的混合型网络等。从技术角度看，网络 130 可以包括有线网络和无线网络，诸如有线网络、WiFi 网络、移动通信网络（例如 3G、4G 和 5G）、广域网、局域网等其他网络。

云平台 003 基于一台或多台服务器 006（图中未示出）提供云服务，服务器 006 可以是机架式服务器或者刀片式服务器，本申请对服务器的具体形态不做限定。云平台 003 可设置在数据中心 004 中或云服务提供方的其他数据中心中。云平台 003 可以是基于公有云、私有云或混合云等方式提供相关云计算服务的平台。

在使用云服务时，用户可以根据终端 001 上显示的界面在云平台 003 上购买和创建实例，实例可以包括虚拟机、容器、裸金属服务器等其他形式的计算或存储等形态的实例。数据中心 004 包括虚拟化的资源池 005，资源池 005 中包括虚拟机（例如 VM1、VM2、VM3）等各类虚拟化资源，云平台 002 可以访问到资源池 005 中各种资源的信息，从而对用户的购买行为做出反馈。其中，这些虚拟机可以部署在一个或多个数据中心中物理机上。

现有技术中，当资源池某种资源量小于用户请求的（最小）数量时，客户的购买（或创建）请求就会失败。以在 huaweicloud.com 上购买虚拟机举例，用户 A 请求购买 100 台 ecs.g6e.2xlarge 的实例，但是云平台查询到当前规格的实例在可用区的配额都不足 100 台时，用户 A 的本次购买就会失败。此时，用户不知道是否可以用其他实例机型替换当前不足部分的购买请求，或是进行多种实例规格的组合购买，以满足用户的业务需求。

针对上述问题，本申请实施例提供了一种在云平台上购买资源的方法，可根据用户的业务需求，对库存资源进行组合后提供给用户。其中，该业务需求中包括用户对底层资源的需求总量。本申请实施例提供的方法无需客户指定具体实例规格以及实例的数量，而是由云平台直接为用户提供购买组合（其中包含可用的实例的规格以及实例的数量）以满足用户需求，能提升购买成功率和客户体验。

图 2 是本申请实施例提供的一种在云平台上购买资源的方法流程举例示意图。图 2 中包括步骤 110 至步骤 130，其中步骤 120 还包括子步骤 121-步骤 123，具体的：

步骤 110、接收终端发送的业务请求。

服务器 005 接收终端 001 发送的业务请求，该业务请求中包括用户输入的业务需求，其中，服务器 005 是用于部署云平台 003 的计算设备。该业务请求中的业务需求是用户在云平台的用户界面（User interface, UI）中输入的，业务需求中包括用户所需的资源总量，这里的资源可以是任何在平台提供的服务资源。

资源总量中可以包括底层资源的总量，例如用户可以在需求采集界面中输入所需的 CPU 的总数量、内存大小总量。通常在云平台中，用户并不可以单独购买或使用 CPU、内存、网络、存储这类底层资源，这类底层资源需要基于诸如虚拟机或物理机才可以为用户提供相关服务。

一种可选的实施方式，除上述资源总量以外，该业务需求中还可以包括资源性能信息。该资源性能信息可用于对资源总量中的各类资源做进一步限定，云平台根据资源性能信息对资源池中的资源进行筛选，找到更贴合用户业务需求的资源，而不是仅根据输入的资源总量

为用户分配实例的规格以及实例的数量。

本申请中的实例可以指云平台上为用户提供计算服务的计算实例。换句话说，本申请中的实例可以是虚拟机。

可选的，上述资源性能信息可以是底层资源相关的性能参数，例如：CPU 型号、CPU 频率、内存型号、内存频率、本地磁盘类型这类，还可以包括 GPU 参数、网卡型号、网络性能参数、存储性能参数等。其中，GPU 参数可以诸如 GPU 的型号和内存大小等参数，网络性能可以诸如网络延时、网卡型号等参数，存储性能参数是诸如磁盘型号、IOPS (Input/Output Operations Per Second) 等参数。

图 3 (a) 是本申请实施例提供的一种需求采集界面的举例示意图，用户可以在图 3 (a) 的界面中输入底层资源相关的性能参数和资源总量。例如在“资源参数”一栏中，用户可以输入（或选择）了 CPU 的类型为 v2、CPU 主频为 2.6Ghz、内存主频为 2.8Ghz 作为底层资源的性能参数，用户还可以在该界面中（“资源总量”一栏）输入 100 个 vCPU 和 200GB 的内存作为需要购买的 CPU 总量和购买的内存总量。

需要说明的是，由于虚拟机来不直接感知物理 CPU，云计算中虚拟机的计算单元一般通过 vCPU 对象来呈现。虚拟机只看到虚拟化管理器 (Virtual Machine Manager, VMM) 呈现给它的 vCPU。在 VMM 中，每个 vCPU 对应一个 VMCS (Virtual-Machine Control Structure) 结构。对于一台物理机，可提供的 vCPU 总数 (逻辑处理器) = Socket 数 (物理 CPU 个数) x Core 数 (内核) x Thread 数 (超线程)。

可选的，底层资源性能参数用户可以选择性设置，即用户可以不进行选择或只选择其中一部分参数，例如用户可以只选择 CPU 型号，不设置其他参数。可选的，对于这里的单个参数项目，用户可以选择多个值，例如用户可以选择多个 CPU 型号或全部型号 (图中未示出)。

一种可能的实施方式，资源总量还可以描述为该资源最小购买总量和最大购买总量，例如图 3 (b) 中所示，“资源总量”一栏示出了 CPU 购买量的上限和下限：100 个 vCPU 和 110 个 vCPU、内存购买量的上限和下限：180GB 内存和 220GB 内存。可选的，用户还可以只选择上限或下限，例如当用户选择购买量下限时意味着只要当资源池库存满足下限就可以为用户创建成功相应实例。

可选的，上述资源性能信息还可以是用户所需的业务场景。在实际中，用户可能无法确认该选择什么底层资源的性能参数，因此本申请实施例还提供了根据业务场景的关键词来确认用户对资源性能的需求。例如，用于描述业务场景的关键词可以是：新手入门、网站、物联网、电商、数据库、机器学习、视频、游戏、内存数据库、图形渲染、大数据、超算等等。用户在用户界面中输入或选择相应业务场景关键词，云平台可以根据接收到的业务场景，筛选符合用户业务需求的资源库存。

图 3 (c) 是本申请实施例提供的一种需求采集界面的举例示意图，用户可以在图 3 (c) 的界面中选择业务场景和资源总量。其中，在“业务场景”一栏中，用户可以输入（或选择）“网站”或“机器学习”等符合业务场景的标签。

可选的，用户还可以同时选择多个业务场景的标签。例如，用户选择了“网站”、“电商”、“数据库”三个标签，云平台可以根据这三类标签对应的资源，筛选符合用户业务需求的实例库存。

可选的，上述资源性能信息还可以是用户所需的实例类型。这里的实例类型用于描述实例自身性能特点，诸如：通用型实例、内存密集型实例、高性能计算型实例、本地盘实例、计算加速型实例。不同的实例类型代表该系列在综合算力、网络宽带、网络延时、IOPS、存

储宽带、内存访问等方面能力的侧重点不同。例如华为云(huaweicloud.com)上的内存密集型实例的系列中包括一种搭载了自研华为鲲鹏 920 处理器及 25GE 智能高速网卡,提供最大 480GB 基于 DDR4 的内存实例和高性能网络,擅长处理大型内存数据集和高网络场景。

图 3 (d) 是本申请实施例提供的一种需求采集界面的举例示意图,用户可以在图 3 (d) 的界面中选择实例类型和资源总量。其中,在“实例类型”一栏中,用户可以输入(或选择)“内存密集型实例”或“大数据型实例”等实例类型的标签。可选的,用户还可以同时选择多个实例类型的标签。

在一种可能的实施方式中,前述资源性能信息可以任意形式组合,使得用户可以从实例类型、业务场景、底层资源性能参数等多个维度进行筛选。

需要说明的是,用户选择“业务场景”和“实例类型”各类标签在本质上也是用于描述底层资源的性能,只不过他们描述的是底层资源性能参数的一个组合形态。举个例子,“内存密集型实例”在云平台内部的处理中,可以用于代表底层资源性能参数满足如下条件的实例资源:(1) CPU 型号为:英特尔® 至强® 可扩展处理器(2) CPU 基频为 3.0Ghz(3) 最大网络收发包为:500 万或 1000 万 PPS(4) 内存类型:DDR4(5) 最大内网带宽为以下中的一种:40 Gbps、25 Gbps、17 Gbps(6) 网卡型号:华为自研 25GE 智能高速网卡。

一种可选的实施方式,该业务需求中还可以包括价格信息。这里的 price 信息可以是资源的总价或资源的单价,还可以表示为价格的上限或下限。

图 3 (e) 是本申请实施例提供的一种需求采集界面的举例示意图,用户在图

3 (e) 的界面中输入了资源的单价上限为“26.8 元/h”。一种可能的实施方式,这里的单价是一个最小单位的资源组合的价格,例如在图 3 (e) 的需求采集界面中,一个最小单位的资源组合可以满足“资源参数”限定条件的 1 个 vCPU 和 2GB 内存的价格之和。可选的,单价也可以指满足条件的实例的平均单价。

一种可选的实施方式,该业务请求中还可以包括筛选规则。这里的筛选规则可以包括:价格优先、性能优先、地域优先、自定义优先等规则。规则举例:价格优先指在满足用户输入的前述需求的前提下,云平台会优先返回价格较低的资源组合。同理,性能优先返回性能更优的实例组成的购买组合,地域优先返回位于同一地域或可用区的资源,自定义优先指云平台根据用户自定义的优先方式匹配满足用户业务需求的资源组合。值得注意的是,这里的“资源组合”可以组成云服务方案,换句话说,云服务方案包括资源组合。

图 3 (f) 是本申请实施例提供的一种需求采集界面的举例示意图,用户可以在图 3 (f) 的界面中选择“筛选规则”的标签:优选低价、优选高性能、优选同区域。可选的,用户可以选择其中一个或多个标签。

可选的,图 4 还提供了一种购买方式选择界面举例示意图。当用户按照本申请实施例的方法购买资源前,可以通过选择图 4 中示出的“算力包”的购买方式标签进入本实施例提出的购买方式界面。由于本申请实施例将实例的算力抽象成资源的性能和资源总量,“算力包”仅作为一种购买方式的名称举例,帮助读者更好理解本申请实施例中提出的方法。

步骤 120、根据业务请求,生成一种或多种资源云服务方案。

云平台 003 根据终端发送的业务请求,生成满足用户业务需求的一种或多种云服务方案其中,该请求中包括以下一项或多项内容:资源总量、资源性能参数、筛选规则、价格信息等,该资源组合中包括实例的规格和数量。一种可能的方式,至少一种云服务方案包括多台虚拟机。

图 5 是本申请实施例给出的一种云平台生成该资源组合的方法流程举例示意图,包括步

骤 121-123，具体的介绍如下：

步骤 121、库存查询

对应于用户购买资源时输入的各类信息，云平台将每种实例的算力抽象成两类信息：性能类和数值类，每一台在云平台发布的实例都具有这两类参数，具体的：

性能类对应前述步骤中的资源性能信息，用于描述某种实例规格对应的底层硬件设施的固有能力，例如：CPU 型号、CPU 主频、CPU 睿频、内存型号、内存主频、本地盘的类型、GPU 的型号、网络性能等等。参见下表 1 所示的举例：

资源	属性	类别	值	含义
vCPU	频率	能力类	2.5Ghz	vCPU 主屏
vCPU	类型	能力类	v3	vCPU 型号
内存	频率	能力类	2.8Ghz	内存主屏
网络	时延	能力类	0.01ns	网络时延性能

数值类可以指物理机通过虚拟化技术拆分后的各类资源（计算、存储、网络），分配到不同的实例上的数量或大小。例如：vCPU 的个数，内存大小，本地盘大小和个数、带宽大小等等。参见下表 2 所示的举例：

资源	属性	类别	值	含义
vCPU	数量	数值类	2	vCPU 个数
内存	大小	数值类	4GB	内存大小
网络	带宽	数值类	0.75G/s	网络带宽

这里的带宽可以是基准带宽或最大带宽。

可选的，数值类还可以包括价格，例如对于上述表 1 和表 2 中的资源组合后的单价。参见表 3 所示：

	属性	类别	值	含义
价格	大小	数值类	1.4 元	单价

举例，单价指包括按包年包月或按量付费方式购买一个月或小时的价格。

可选的，CPU、内存和网络还可以基于基础值和阶梯倍数对外发布。参见下表 4 所示的举例。

资源	属性	类别	值	含义
vCPU	阶梯值	数值类	1, 2, 4, 6, 8, ...	vCPU 阶梯数值
内存	阶梯值	数值类	1, 2, 4, 6, 8, ...	内存阶梯数值
网络	阶梯值	数值类	0.1, 0.2, 0.35, 0.75...	网络阶梯数值

例如在表 4 中，vCPU 基础值为 1 个，接着是 2 个、4 个、6 个、8 个等等。内存的大小基础值为 1GB，接着是 2GB、4GB、6GB、8GB 等等。网络带宽的基础值为 0.1Gbit/s，接着为 0.2 Gbit/s、0.35 Gbit/s、0.75 Gbit/s 等等。

一种可能的实施方式，云平台首先根据业务请求中的性能类信息查询库存，找到符合条件的资源，然后再根据数值类信息进行匹配，最终满足业务请求的实例组合。

一种实施方式举例，云平台获取业务请求后，根据满足用户购买请求中的能力类信息作为的条件，查询当前库存，生成可用库存记录。例如，业务请求中要求了 CPU 主频为 2.5Ghz，内存主频为 2.8Gh，查询到符合条件的库存记录如表 5 所示：

机型ID	CPU总数	内存总量	单价
------	-------	------	----

048	3000	6000	1.2
049	2000	5000	2.8

表中关于机型和单价的详细内容请参加后文的描述。可选的，库存表中的内容还可以包括更多信息，例如资源所在的地域等。

步骤 122、库存匹配

云平台基于符合条件的库存记录，根据数值类信息进行匹配。一种可能的实施方式，包括：首先，获取查询到的库存记录（符合能力类的资源），再根据业务请求中的资源总量，建立库存匹配函数。然后，根据筛选规则确立目标函数，例如筛选规则可以为低价优先、性能优先或其他规则。最后，确定约束条件，求解优化模型，得到一种或多种资源组合方式。

以低价优先作为筛选规则，一种实施方式举例如下，具体的：

(1) 获取查询到的库存记录，将库存中的资源记为 $r = r_i(r_{i0}, r_{i1}, \dots, r_{im})$ ，将每一种实例在库存中的总额记为 C_i 。其中， i 代表某一种实例机型资源。 $\{0, 1, \dots, m-1\}$ 代表各个维度的资源标号，举例说明：0 代表 CPU、1 代表内存、2 代表网络带宽，则 $(r_{i0}, r_{i1}, \dots, r_{im})$ 用于表述前述各个维度的资源容量。

此外，记库存中的某种资源单价为 W_i ，即 i 对应机型的单份资源的价格，单份价格可以理解为云平台上发布的该机型最小容量的实例资源的价格。

一种可能的实施方式，同一种机型可以对应多个实例规格型号，它们通常具备的能力类参数大部分是相同的，但 CPU 数量和内存大小通常不一样。可选的，同一种机型可以具有相同的 CPU/内存配比的实例。

可选的，云平台发布的实例规格可以是上述单份资源的整数倍数形式，价格是上述单份资源价格的整数倍数形式。

可选的，云平台发布的实例规格可以不是上述单份资源的整倍数形式，而是以阶梯值的方式发布，并且每种资源的阶梯值可能不一样。例如，某种实例的 CPU 基础值为 2 个，接着是 4 个、8 个、16 个等等，网络带宽的基础值为 0.1 Gbit/s，接着为 0.2 Gbit/s、0.35 Gbit/s、0.75 Gbit/s 等。云平台发布实例的价格不是上述单价的整数倍数形式计费，或是以其他的价格计费规则。一种可能的实施方式，在本步骤中还要考虑这些阶梯值对可售资源的影响，而不仅仅是记录单份资源的价格。

然而，本申请实施例在下文的一个举例中，提供了一种实例规格是单份资源的整数倍的理想情况作为举例，但不代表不申请提出的方法不需要考虑阶梯值和对实例计费方式的问题。

(2) 根据业务请求中的资源总量，建立库存匹配函数。一种可能实施方法举例， $\text{HighR} = (\text{HighR}_0, \text{HighR}_1, \dots, \text{HighR}_m)$ 和 $\text{lowR} = (\text{lowR}_0, \text{lowR}_1, \dots, \text{lowR}_m)$ 表示满足用户请求上限的函数和满足用户请求下限的函数， R 代表用户请求算力资源量，0 至 m 代表在各个算力维度上请求的资源量。记满足用户购买请求，在总库存 $r = r_i$ 中申请的某一种实例机型资源的数额为 N_i 。

(3) 确立目标函数。一种可能的实施方式，目标函数可以是优化函数 $\min \sum_{i=0}^n N_i \cdot W_i$ ，即最终购买的总价最低。需要说明的是，这里的目标函数可以是服务器默认的规则，也可以是用户选择的筛选规则。

(4) 确定约束条件，求解优化模型。一种可能的实施方式，约束条件可以从 $r = r_i$ 中申请的资源份额 N_i 能满足用户的请求 R 。可选的，用户的请求 R 可以包括上限 highR 和下限 LowR 。一种实施方式举例，若用户还配置了价格最大值 P ，则约束条件可以表达为：

$$\begin{cases} \text{high}R_j \geq \sum_{i=0}^n N_i * r_{ij} \geq \text{low}R_j \quad \forall j \in \{0,1,\dots,m-1\} \\ \sum_{i=0}^R N_i * W_i < P \\ N_i \leq C_i, \quad N_i \text{ 为整数} \end{cases}$$

为了进一步介绍求解上述优化模型的过程，这里以CPU和内存两个维度的底层资源作为用户“算力包”购买请求的内容，作为举例加以说明：

a. 服务器接收到用户A的购买请求为：最小需要100个CPU和300GB内存，其中CPU类型要求为Intel Cascade Lake。

b. 通过查询库存，发现符合CPU类型要求为Intel Cascade Lake 的库存中有三种机型资源，分别如下：

机型ID为001的CPU和内存1:2配比的资源40份，单价是1。

机型ID为002的CPU和内存1:4配比的资源50份，单价是1.5。

机型ID为003的CPU和内存1:2配比的资源100份，单价是0.5。

此时带入上述求解模型：

i=3, 代表有库存中有三种符合要求的机型资源。m=2, 代表请求的2种资源，分别是CPU和内存。各库存的单价表示为： $W_0=1$ ， $W_1=1.5$ ， $W_2=0.5$ 。

用户A的购买请求为 $r=r_i(r_{i0}, r_{i1}, r_{i2})$ ，其中 r_0 代表库存001的容量， r_{00} 代表库存001在CPU维度的资源， r_{01} 代表库存001在内存维度的资源。 r_1 代表库存002的容量， r_{10} 代表库存002在CPU维度的资源， r_{11} 代表库存002在内存维度的资源。 r_2 代表库存003的容量， r_{20} 代表库存003在CPU维度的资源， r_{21} 代表库存003在内存维度的资源。

用户A请求R中包含资源下限为R(100,300)，即 $\text{low}R = R(100,300)$ 。

c. 目标函数为： $\min \sum_{i=0}^n N_i * W_i$ ，求解的 N_i 为从上述三种机型资源中分别取出的份数。

e. 约束条件为：

$$\begin{cases} N_0 * r_0 + N_1 * r_1 + N_2 * r_2 \geq R(100,300) \\ N_0 \leq 40, N_1 \leq 50, N_2 \leq 100 \end{cases}$$

f、通过整数规划问题求解防范进行求解，例如使用割平面求解方法可以求解出，最优解是 $N_0=0$ ， $N_1=30$ ， $N_2=100$ 。通过上述模型的建立，可以使用整数规划中常见解法进行求解最优解。

步骤123、生成一种或多种资源组合

根据优化模型求解的结果，可以得到一种或多种资源组合结果，结果中包括实例的机型和数量。

一种可能的实施方式，实例的机型和数量就是返回终端的实例规格和数量。即同一种机型只对应一种实例规格，机型和实例规格的关系请参照前文的描述。

一种可能的实施方式，这里的实例的机型和数量不是返回终端的实例规格和数量。即同一种机型对应多种实例规格。例如，华为云上c6.large.2、c6.xlarge.2、c6.2xlarge.2、c6.3xlarge.2这四种规格是同一种机型的不同规格。它们具有相同的能力类参数(例如CPU型号)和相同的CPU/内存配比(例如1:2)，但CPU数量和内存大小不相同，分别为：2vCPUs|4GB、4vCPUs|8GB、8vCPUs|16GB、12vCPUs|24GB。一种可能的实施方式，在求解到机型和数量(指单份资源或最小可售单位资源的数量)后，还需要进一步分配机型对应的具体实例规格和数量，本申请不对这部分分配的方式做任何限定。

步骤 130、向终端提供上述生成的云服务方案，至少一个云服务方案中包括多个虚拟机。

云平台向终端提供上述生成的云服务方案，至少一个云服务方案中包括多个虚拟机，上述的云服务方案中虚拟机的资源总量满足业务请求中的资源总量信息。上述的“提供”可以是发送，换句话说，即云平台向终端发送上述生成的云服务方案。一种可能的实施方式，终端接收到服务器发送的返回界面。

图 6 是本申请提供的一种返回界面的举例示意图。在图 6 界面中，用户选择了“网站”和数据库“作为业务场景，用户输入了 CPU 购买量的上限和下限：100 个 vCPU 和 110 个 vCPU、内存购买量的上限和下限：180GB 内存和 220GB 内存作为需求的资源总量。一种可能的实施方式，用户在点击“确认”按钮后，云平台会返回匹配的云服务方案的结果，其中云服务方案包括前述的资源组合，包括实例的规格和数量。如图 6 所示，界面中包括了两种资源购买组合：组合 1 包括 15 台 s2 型实例和 10 台 k3 型实例、组合 2 包括 40 台 k2 型实例和 10 台 m3 型实例，参考价格分别是 78.7 元/小时。一种可能的实施方式，这两种资源组合分别是云平台推荐的价格较低的两种购买组合，其中第二种组合的性能较第一种更优。

可选的，用户还可以图 6 中返回的购买组合进一步修改，例如：用户可以点击进入资源组合的详细信息界面，修改其中的实例规格和数量，并根据修改后的资源组合下单购买。

本申请实施例还提供一种装置 210，该装置可以部署在云平台 003 上，用于实现本申请实施例中提供的在云平台 003 上购买资源的方法。

本申请实施例还提供一种提供云服务的装置 210，该装置部署在云平台 003 上。装置 210 可以包含一个或多个模块（图中未示出），可以实现提供云服务的功能。图 7 是本申请实施例提供的一种装置 210 的结构举例示意图。图中环境信息预测装置 210 包括接收模块 211、发送模块 212、服务模块 213。

接收模块 211，用于接收终端发送的业务请求，该业务请求中包括用户所需的资源总量信息；

发送模块 212，向终端提供一种或多种云服务方案，该一种或多种云服务方案中的至少一种云服务方案中包括多个虚拟机，该一种或多种云服务方案中虚拟机的资源总量满足该业务需求中的资源总量。

一种可能的实施方式，该资源总量信息包括下述信息的至少一种：CPU 总量信息和内存总量信息。

一种可能的实施方式，所述业务需求中还包括资源性能信息，其中，所述一种或多种云服务方案中虚拟机的资源性能满足所述业务需求中的资源性能信息。

一种可能的实施方式，所述资源性能信息包括底层资源性能参数，其中所述底层资源性能参数包括下述至少一种：CPU 型号、CPU 频率、内存型号、内存频率、GPU 参数、网络性能以及存储读写性能。

可选的，所述资源性能信息包括下述信息的至少一种：虚拟机能力类型和业务场景类型，其中，所述虚拟机能力类型用于指示虚拟机的计算、存储和网络能力，所述业务场景类型用于指示虚拟机适用的业务场景。

一种可能的实施方式，所述 CPU 总量信息包括下述信息的至少一种：CPU 总量、CPU 总量的最大值、CPU 总量的最小值；

一种可能的实施方式，所述内存总量信息包括下述信息的至少一种：内存总量、内存总量的最大值、内存总量的最小值。

可选的，所述业务请求中还包括所述用户指定的价格信息，所述一种或多种云服务方案的价格满足所述用户指定的价格信息。

服务模块 213，用于根据用户从所述一种或多种云服务方案中选择的目标云服务方案，向用户提供相应的云服务。

图 8 是本申请实施例提供了一种计算设备 300 的结构举例示意图，该设备可以是任何计算设备，可以用于实现由云平台 003 上的服务器 006 实现的功能，云平台的用户可以通过终端设备访问该计算设备 300 购买云平台上的服务，换句话说云平台可以基于计算设备 300 提供本申请实施例中的云服务。

如图 8 所示，计算设备 300 包括总线 301、处理器 302、存储器 303 和接口 304。处理器 302、存储器 303 之间通过总线 301 通信。

总线 301 用以连接处理器 302、存储器 303 和接口 304 以及其他可能的组件，可以包括地址总线、数据总线、控制总线、扩展总线、局部总线中的一种或多种。为便于表示，图中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。总线 301 应用外设部件互连标准（peripheral component interconnect, PCI）总线或扩展工业标准结构（extended industry standard architecture, EISA）总线等标准。

处理器 302 可以为中央处理器（Central Processing Unit, CPU），此外，还可以有图形处理器（graphics processing unit, GPU）、神经网络处理器（Neural network Processing Unit, NPU）、FPGA（Field Programmable Gate Array）等其他类型处理器芯片中的一种或多种。处理器 302 可以执行本申请实施例中提供云服务的方法的步骤。

存储器 303 可以包括随机存取存储器（random access memory, RAM）、只读存储器（read-only memory, ROM）、快闪存储器（flash EPROM）、机械硬盘（hard disk drive, HDD）或固态硬盘（solid state drive, SSD）等任何可以用来存储并保存信息的其他存储器、存储设备、存储介质中的一种或多种。存储器 303 中存储有可执行的程序代码，该可执行的程序代码可通过处理器 302 执行以实现本申请实施例中提供云服务的方法的步骤。

接口 304 可以为网络接口卡（Networking Interface Card, NIC）或其他类型的收发机设备、主机总线适配器（Host Bus Adaptor, HBA）等。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分可互相参考，每个实施例重点说明的都是与其他实施例的不同之处，部分实施例描述简略的地方可以参考其他实施例的描述。

以上为本申请的具体实现方式。应当理解，以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的范围。

权 利 要 求 书

1. 一种提供云服务的方法，其特征在于，所述方法包括：
接收终端发送的业务请求，所述业务请求中包括所述用户所需的资源总量信息；
向终端提供一种或多种云服务方案，所述一种或多种云服务方案中的至少一种云服务方案中包括多个虚拟机，所述一种或多种云服务方案中虚拟机的资源总量满足所述业务请求中的资源总量信息。
2. 根据权利要求1所述的方法，其特征在于，所述资源总量信息包括下述信息的至少一种：CPU 总量信息和内存总量信息。
3. 根据权利要求1或2所述的方法，其特征在于，所述业务请求中还包括资源性能信息，其中，所述一种或多种云服务方案中虚拟机的资源性能满足所述业务请求中的资源性能信息。
4. 根据权利要求3所述的方法，其特征在于，所述资源性能信息包括底层资源性能参数，其中所述底层资源性能参数包括下述至少一种：CPU 型号、CPU 频率、内存型号、内存频率、GPU 参数、网络性能以及存储读写性能。
5. 根据权利要求3或4所述的方法，其特征在于，所述资源性能信息包括下述信息的至少一种：虚拟机能力类型和业务场景类型，其中，所述虚拟机能力类型用于指示虚拟机的计算、存储和网络能力，所述业务场景类型用于指示虚拟机适用的业务场景。
6. 根据权利要求2至5任一项所述的方法，其特征在于，所述CPU 总量信息包括下述信息的至少一种：CPU 总量、CPU 总量的最大值、CPU 总量的最小值；所述内存总量信息包括下述信息的至少一种：内存总量、内存总量的最大值、内存总量的最小值。
7. 根据权利要求1至6任一项所述的方法，其特征在于，所述业务请求中还包括所述用户指定的价格信息，所述一种或多种云服务方案的价格满足所述用户指定的价格信息。
8. 根据权利要求1至7任一项所述的方法，其特征在于，所述方法还包括：根据所述用户从所述一种或多种云服务方案选择的目标云服务方案，向所述用户提供云服务。
9. 一种提供云服务的装置，其特征在于，所述装置包括：
接收模块，用于接收终端发送的业务请求，所述业务请求中包括所述用户所需的资源总量信息；
发送模块，向终端提供一种或多种云服务方案，所述一种或多种云服务方案中的至少一种云服务方案中包括多个虚拟机，所述一种或多种云服务方案中虚拟机的资源总量满足所述业务请求中的资源总量信息。
10. 根据权利要求9所述的装置，其特征在于，所述资源总量信息包括下述信息的至少一种：CPU 总量信息和内存总量信息。
11. 根据权利要求9或10所述的装置，其特征在于，所述业务请求中还包括资源性能信息，其中，所述一种或多种云服务方案中虚拟机的资源性能满足所述业务请求中的资源性能信息。
12. 根据权利要求11所述的装置，其特征在于，所述资源性能信息包括底层资源性能参数，其中所述底层资源性能参数包括下述至少一种：CPU 型号、CPU 频率、内存型号、内存频率、GPU 参数、网络性能以及存储读写性能。
13. 根据权利要求11或12所述的装置，其特征在于，所述资源性能信息包括下述信息的

至少一种：虚拟机能力类型和业务场景类型，其中，所述虚拟机能力类型用于指示虚拟机的计算、存储和网络能力，所述业务场景类型用于指示虚拟机适用的业务场景。

14. 根据权利要求 10 至 13 任一项所述的装置，其特征在于，
所述 CPU 总量信息包括下述信息的至少一种：CPU 总量、CPU 总量的最大值、CPU 总量的最小值；

所述内存总量信息包括下述信息的至少一种：内存总量、内存总量的最大值、内存总量的最小值。

15. 根据权利要求 9 至 14 任一项所述的装置，其特征在于，所述业务请求中还包括所述用户指定的价格信息，所述一种或多种云服务方案的价格满足所述用户指定的价格信息。

16. 根据权利要求 9 至 15 所述的装置，其特征在于，所述装置还包括：
服务模块，用于根据所述用户从所述一种或多种云服务方案中选择的目标云服务方案，向所述用户提供云服务。

17. 一种设备，其特征在于，所述设备包括：存储器和处理器；所述处理器用于执行所述存储器存储的指令，使得所述设备执行权利要求 1 至 8 任一项所述的方法。

18. 一种非易失性存储介质，其特征在于，所述非易失性存储介质包括指令，所述指令被计算设备执行时，使所述计算设备执行权利要求 1 至 8 任一项所述的方法。