

一种业务分流方法、装置及系统

相关申请的交叉引用

本申请要求在 2022 年 02 月 14 日提交中华人民共和国知识产权局、申请号为 202210132528.9、申请名称为“一种数据处理方法和计算机”的中国专利申请的优先权，其全部内容通过引用结合在本申请中；本申请要求在 2022 年 06 月 13 日提交中华人民共和国知识产权局、申请号为 202210667595.0、申请名称为“一种业务分流方法、装置及系统”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及云计算技术领域，尤其涉及一种业务分流方法、装置及系统。

背景技术

负载均衡是一种根据负载均衡算法，把业务流量分散到一个实例集群中的可用服务器的技术。负载均衡通过管理进入实例集群的业务流量，从而使网络访问者获得尽可能好的网络体验。

目前，负载均衡策略包括负载均衡算法、预设分流比例等，其中，负载均衡算法主要用于选择分流的实例集群，例如轮询业务对应的实例集群，将业务请求分配到负载较小的实例集群上进行处理，或者，选择与触发业务请求的用户设备所在地距离较近实例集群来响应该业务请求。预设分流比例是指，按照业务对应的多个实例集群的预设比例进行分流。

然而，当业务流量突发，或者，业务对应的多个实例集群的资源可用量不一致时，选择负载最小的实例集群的或者按照预设流量比例进行分流，只能实现粗粒度的分流，可能导致分配至实例集群的业务请求不能得到及时处理，在流量突发场景中加重该情况的恶化，降低网络访问者的体验。

发明内容

本申请提供一种业务分流方法、装置及系统，用于实现细粒度精准分流，保证业务请求能够得到及时处理。

第一方面，本申请提供一种业务分流方法，该方法可应用于云计算系统，该系统至包括云管理平台节点、多个资源池和分流节点，其中，每个资源池包括多个主机用于为多个租户运行业务，每个主机上运行至少一个实例，每个资源池内可创建有目标租户的目标业务对应的实例集群，实例集群内的实例用于运行该目标业务；该方法包括：云管理平台节点获取目标业务关联的每个资源池的可用资源信息，其中，每个资源池的可用资源信息指示该资源池在一个时间段（记为目标时间段）内可提供给目标业务的实例数量；云管理平台节点基于每个资源池的可用资源信息生成目标业务在目标时间段内的分流规划；分流节点按照该分流规划将所述目标业务在目标时间段内的业务流量分配至一个或多个资源池。

上述技术方案中，云管理平台节点基于目标业务关联的资源池的可用资源信息，确定目标业务在目标时间段内的分流规划，从而可以预先确定出不同时间段内目标业务的业务

流量分配入各资源池的业务流量，有助于更好的实现流量分配规划，避免分配入实例集群的业务流量不超过其所在资源池的承载能力，尽可能保证分配入实例集群的业务请求能够得到及时处理，实现细粒度分流，减少或避免业务流量突发时业务请求大规模拥塞的情况。

5 在一种可能的实现方式中，每个资源池的可用资源信息指示：该资源池在目标时间段内空闲实例的数量，空闲实例为该资源池在目标时间段内未被其他租户占用的。

上述技术方案中，相对于负载均衡策略而言，本申请不再着眼于实例集群的负载，而是确定资源池在目标时间段内空闲实例的数量，云管理平台节点基于该空闲实例的数量制定分流规划，如此，不论业务对应的多个实例集群的资源可用量是否一致，相对于负载均衡等策略均可实现一种更细粒度的流量分配。

10 在一种可能的实现方式中，分流规划指示在目标时间段内每个资源池可承接目标业务的业务流量的最大值，或所述目标业务的业务流量在所述时间段内被分配入每个资源池比例。

上述技术方案中，云管理平台节点基于各资源池的可用资源信息确定分流规划，指示每个资源池可承接目标业务的业务流量的最大值，避免分配入实例集群的业务流量不超过其所在资源池的最大承载能力。当分流规划指示流量比例时，可以与前述的负载均衡策略中按照预设流量比例进行分流的分流节点兼容，提高本申请技术方案的适用性。

在一种可能的实现方式中，分流规划指示目标业务的业务流量在目标时间段内被分配入每个资源池优先级信息。

20 上述技术方案中，基于资源池的优先级信息可调整将目标业务的业务流量分配入资源池的顺序，使流量分配的方式更加灵活。

在一种可能的实现方式中，该方法还包括：云管理平台节点获取目标业务的资源需求信息，该资源需求信息指示目标业务在目标时间段内的业务流量预测值；云管理平台节点基于目标业务的资源需求信息和每个资源池的可用资源信息生成分流规划。

25 上述技术方案中，基于目标业务的资源需求信息和每个资源池的可用资源信息生成分流规划，可实现计算资源与业务需求的协同调度，实现更细粒度的分流。

在一种可能的实现方式中，在目标时间段内被分配入每个资源池的目标业务的业务流量之和等于业务流量预测值，且在目标时间段内被分配入每个资源池的目标业务的业务流量不超过该资源池可承载的目标业务的业务流量的最大值。

30 在一种可能的实现方式中，分流规划满足成本策略，成本策略包括在目标时间段内被分配入每个资源池的目标业务的业务流量所产生的费用最低。

上述技术方案中，提供给目标租户一种新的成本策略，当采用成本策略确定分流规划时，还可以有效降低目标租户的成本费用。

35 在一种可能的实现方式中，在目标时间段内被分配入每个资源池的目标业务的业务流量所产生的费用最低包括实例费用和通信费用；实例费用指示在所述时间段内被分配入每个资源池的目标业务的业务流量所需的实例的租赁费用，通信费用指示所述分流节点与每个资源池传输目标业务的业务流量所需的通信费用。

在一种可能的实现方式中，每个资源池可承接目标业务的业务流量的最大值基于资源池在目标时间段内可提供给目标业务的实例集群的实例数量及目标业务的资源转换信息确定的，资源转换信息指示处理所述目标业务的单个业务请求所需的实例的数量。

40 第二方面，本申请实施例还提供了一种计算装置，该装置具有实现上述第一方面或第

一方面任一可能的实现方式中云管理平台节点的功能，有益效果可以参见第一方面的描述
此处不再赘述。所述功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述
硬件或软件包括一个或多个与上述功能相对应的模块。在一个可能的设计中，所述装置的
结构中包括获取模块、确定模块，可选的，还可以包括通信模块，这些模块可以执行上述
5 第一方面方法示例中的相应功能，具体参见方法示例中的详细描述，此处不做赘述。

第三方面，本申请实施例还提供了一种计算设备，所述计算设备包括处理器和存储器，
还可以包括通信接口，所述处理器执行所述存储器中的程序指令执行上述第一方面或第
一方面任一可能的实现方式提供的方法。所述存储器与所述处理器耦合，其保存确定分流规
划的过程中必要的程序指令和数据。所述通信接口，用于与其他设备进行通信，如发送分
10 流规划。

第四方面，本申请提供了一种计算设备集群，该计算设备集群包括至少一个计算设备。
每个计算设备包括存储器和处理器。至少一个计算设备的处理器用于访问所述存储器中的
代码以执行第一方面或第一方面的任意一种可能的实现方式提供的方法。

第五方面，本申请提供了一种计算机可读存储介质，所述计算机可读存储介质被计算
设备执行时，所述计算设备执行前述第一方面或第一方面的任意可能的实现方式中提供
15 的方法。该存储介质中存储了程序。该存储介质包括但不限于易失性存储器，例如随机访问
存储器，非易失性存储器，例如快闪存储器、硬盘（hard disk drive, HDD）、固态硬盘（solid
state drive, SSD）。

第六方面，本申请提供了一种计算机程序产品，所述计算设备程序产品包括计算机指
令，在被计算设备执行时，所述计算设备执行前述第一方面或第一方面的任意可能的实现
20 方式中提供的方法。该计算机程序产品可以为一个软件安装包，在需要使用前述第一方面
或第一方面的任意可能的实现方式中提供的方法的情况下，可以下载该计算机程序产品并
在计算设备上执行该计算机程序产品。

第七方面，本申请还提供一种计算机芯片，芯片与存储器相连，芯片用于读取并执行
25 存储器中存储的软件程序，执行前述第一方面或第一方面的任意可能的实现方式中提供
的方法。

上述第二方面至第七方面中任一实现方式的有益效果请参见第一方面的描述，此处不
再赘述。

附图说明

- 图 1 为一种区域和可用区的关系示意图；
- 图 2 为本申请实施例提供的一种可能的网络架构示意图；
- 图 3 为本申请实施例提供的另一种可能的网络架构示意图；
- 图 4 为本申请实施例提供的业务分流方法的流程示意图；
- 35 图 5 为本申请实施例提供的一种计算装置的结构示意图；
- 图 6 为本申请实施例提供的一种计算设备的结构示意图；
- 图 7 为本申请实施例提供的一种计算设备集群的结构示意图。

具体实施方式

为方便理解,首先对本申请实施例中的部分技术用语进行解释说明。

(1) 区域 (Region) 和可用区 (AZ, Availability Zone) 用于描述数据中心的位置。区域指数据中心所在的区域,可以是地区 (例如华北)、也可以是城市 (例如北京、天津)。

5 可用区指机房、站点所在的物理区域,具有能耗、网络所独立的特点。

参见图 1 所示,一个 Region 通常包含多个低时延互联的 AZ,用于同区域容灾备份和负载均衡等场景与服务。比如,在北京 (Region) 有两个机房,这两个机房就是两个可用区,它们拥有独立的水电、网络等。一个 Region 中的多个 AZ 间通过数据中心内部网络 (高速光纤) 相连,以满足租户跨 AZ 构建高可用性系统的需求。不同 region 之间需要通过公网互联。

10

(2) 实例,可指用于运行业务的软件程序或计算资源,还可以称为计算实例、计算模块、算法实例等。

(3) 虚拟机 (Virtual Machine, VM),是指通过软件模拟的具有完整硬件系统功能的、运行在一个完全隔离环境中的完整计算机系统。在实体计算机 (简称实体机) 中能够完成的工作在虚拟机中都能够实现。在实体机中创建虚拟机时,需要将实体机的部分硬件资源 (如 CPU、内存、硬盘等) 作为虚拟机的硬件资源。每个虚拟机都有独立的操作系统,可以像使用实体机一样对虚拟机进行操作。在实际应用中,一台物理主机可以通过虚拟化技术虚拟出多台虚拟机。虚拟机也可称为云服务器 (Elastic Compute Service, ECS)、弹性实例 (不同的云服务提供商有不同的叫法)。

15

(4) 容器,是实现操作系统虚拟化的一种途径,可以让用户在资源受到隔离的进程中运行应用程序及其依赖关系。

20

(5) 裸金属服务器,是指可以为租户提供专属化云资源的物理服务器,即主机 (host)。

图 2 为本申请实施例提供一种可能的云计算系统的架构示意图。该云计算系统包括云管理平台 1、一个或多个资源池 (图 2 示出三个资源池 20、21、22,但不限于于此)。

首先介绍资源池。通常,每个资源池包括多个主机 (Host),每个主机上运行一个或多个实例。示例性地,同一个 Region 内使用同一个资源池,即一个资源池可包括一个 Region 内运行的全部实例。本申请实施例中,实例可以是虚拟机、容器、主机等不同粒度的计算资源的任意一种或多种组合。

25

租户可在特定的一个或多个资源池购买实例。对应的,一个或多个资源池内可创建目标租户的目标业务的实例集群,如图 2 中,资源池 20 内创建有目标租户的目标业务的实例集群 200、资源池 21 内创建有目标租户的目标业务的实例集群 210、资源池 22 内创建有目标租户的目标业务的实例集群 220。实例集群内包括目标租户在该资源池租赁的实例,可包括虚拟机、容器、裸金属服务器中的一种或多种,用于运行目标业务,如图 2 右侧所示的,用户通过用户设备 101、102、103 等发送过来的业务流量。其中,每个实例集群内的实例可包括虚拟机、容器、裸金属服务器中的一种或多种。为便于说明,如下以实例为虚拟机为例来介绍。

30

在实际应用中,每个资源池可包括多种规格的虚拟机,目标租户从资源池租赁虚拟机时,可指定虚拟机的一种或多种规格。其中,不同规格的虚拟机可具有不同的硬件配置,硬件配置包括但不限于: CPU 核数、内存容量、磁盘空间、带宽等。表 1 示例性地示出了某资源池内的几种虚拟机规格及相应的收费标准。应注意,表 1 仅为举例,不构成对实际

40

应用的限定。

表 1

虚拟机规格	CPU 核	内存容量	磁盘空间	机时费用 (1h)
VM_1	1	1.7GB	50GB	\$0.12
VM_2	2	3.5GB	250GB	\$0.24
VM_3	4	7GB	1000GB	\$0.48
VM_4	8	15GB	2000GB	\$0.96

应注意，不同资源池内包括的虚拟机的规格可以完全相同，或者不完全相同，或者完全不同。其中，相同规格的虚拟机在不同的资源池可能具有不同的收费标准，这与资源池本身的运营成本相关。另外应注意，目标租户在不同资源池中租赁的虚拟机的规格可能是不同的，比如，实例集群 200 内的虚拟机的规格为 VM_1 和 VM_2，实例集群 210 内的虚拟机的规格为 VM_2，实例集群 220 内的虚拟机的规格为 VM_1。

在资源池内，除租赁给租户的虚拟机（即被纳入实例集群内的虚拟机）之外，还包括空闲的虚拟机，可以理解为未被其他租户占用或未出售的虚拟机，可用于资源池中已有实例集群的扩容。通常，目标租户不会购买固定数量的虚拟机，换言之，实例集群 200、210、220 内所包括的虚拟机的数量是不固定的，可通过弹性伸缩技术进行扩容、缩容。弹性伸缩技术，是指根据业务需求，自动增加或自动减少虚拟机的数量，如在接入的业务流量增多导致算力不足时，增加实例集群中的虚拟机数量，即将资源池中空闲的虚拟机纳入实例集群内，直至业务流量下降，算力富余后减少实例集群中的虚拟机数量，即将实例集群中的部分空闲的虚拟机释放回资源池中。

如下介绍云管理平台 1，云管理平台 1 包括监控系统 10、云管理平台节点 11、分流节点 12。

监控系统 10 包括监控模块 113 和数据库 114。

其中，监控模块 113，用于对资源池的性能指标进行监控，包括各种物理资源和虚拟资源的监控，如资源池中各实例集群内的 VM 数量、资源池中不同规格的空闲 VM 数量等。监控模块 113 负责实时或周期性的采集资源池的各项性能指标的值。数据库 114，用于存储监控模块 113 采集到的数据。可选的，数据库 114 还可用于存储其他信息，如各资源池包括的各种 VM 的规格及相应的收费标准（参见表 1）、目标业务对应的各实例集群、实例集群所在的资源池（目标业务关联的资源池）、及在资源池中指定的 VM 规格等的映射关系，等等，本申请对此不做限定。

云管理平台节点 11，用于执行本申请实施例提供的业务分流方法，示例性的，云管理平台节点 11 用于获取目标业务关联的每个资源池的可用资源信息；根据目标业务关联的每个资源池的可用资源信息，确定目标业务在未来一个时间段（记为目标时间段）内的分流规划。可选的，云管理平台节点 11 还用于提供一种配置界面，目标租户可在该配置界面配置业务参数，如目标业务关联的资源池、在资源池中用于运行目标业务的虚拟机的规格、应用于实例集群的弹性伸缩策略、用于指定分流规划的成本策略（下文会进行介绍，此处不做赘述）等，这些业务参数可存储于数据库 114 中。

分流节点 12，为目标业务的流量接口，用于接收目标业务的业务流量，如图 2 中右侧所示，用户设备可以是普通计算机 101、平板电脑 102、智能手机 103 等。用户可通过用户设备发送业务请求。分流节点 12，还用于按照云管理平台节点 11 生成的分流规划，将

目标业务在目标时间段内的业务流量分配至多个资源池。本申请中，业务流量可以以业务请求的数量来度量。举例来说，分流节点 12 基于分流规划，将目标业务在目标时间内 2000 个业务请求发送至实例集群 200，将目标业务在目标时间内其他的 3000 个业务请求发送至实例集群 210。示例性地，分流节点 12 可通过网络（如因特网）与资源池通信。

5 在一种可选的实施方式中，分流节点 12 可位于目标业务关联的一个资源池中，以节省分流节点 12 与该资源池中的实例集群之间的网络通信费用。比如，目标租户的目标业务对应于实例集群 200、实例集群 210 和实例集群 220，则该目标业务的关联资源池包括资源池 20、资源池 21 和资源池 22，目标业务的分流节点 12 可以位于资源池 20 中，或者资源池 21 中，或者资源池 22 中。参见图 3 所示，假设目标业务的分流节点 12 位于资源池 20 中，如此，当分流节点 12 向实例集群 200 发送业务流量时，不需要通过网络传输，从而节省两者之间的通信费用。

应注意，本申请中的用户与租户是不同的，租户为业务方，用户为使用业务方提供的服务的一方。举例来说，假设业务方提供某 app，如该 app 为搜索引擎，则用户为使用该搜索引擎的一方，在用户设备上运行搜索引擎触发的业务请求可以是网页获取请求等。又如该 app 为视频软件，用户为使用该视频软件的一方，在用户设备上运行该视频软件触发的业务请求可以是获取视频请求等等。再如该 app 为会议软件，用户可在用户设备上运行该会议软件，触发的业务请求可以是创建会议请求、加入会议请求等。对应的，目标业务对应的实例集群内的虚拟机可用于处理目标业务的业务请求，比如，针对创建会议请求，创建会议资源并生成会议链接，将会议链接等信息置于会议响应中返回给分流节点 12，由分流模型 13 返回给对应的用户。

需要说明的是，图 2 或图 3 所示的系统架构仅为示意，实际应用中，云计算系统可以包括相对图 2 或 3 更多或更少的节点，本申请实施例对此不做限定。

本申请提供一种业务分流方法，在制定分流规划时，基于目标业务关联的各资源池处理能力制定分流规划，尽可能保证分配至实例集群的业务请求能够得到及时处理，实现细粒度的分流。

接下来结合附图 4，以本申请实施例提供的业务分流方法应用于图 3 所示的系统为例，对该方法进行详细说明。为便于说明，仍以虚拟机为例进行介绍，下文中的虚拟机均可以替换为实例，假设图 3 中，目标租户的目标业务对应于实例集群 200、实例集群 210 和实例集群 220。为便于描述，下文中将“目标租户的目标业务”简称为“目标业务”。

30 图 4 为本申请实施例提供的一种业务分流方法所对应的流程示意图。如图 4 所示，该方法包括如下步骤：

步骤 401，云管理平台节点 11 获取目标业务关联的每个资源池的可用资源信息。

35 目标业务关联的资源池是指目标业务对应的实例集群所在的资源池，比如，图 3 中，目标业务对应于实例集群 200、实例集群 210、实例集群 220，则目标业务关联的资源池包括资源池 20、资源池 21 和资源池 22。

其中，每个资源池的可用资源信息，用于指示该资源池在目标时间段内可提供给目标业务的虚拟机的数量（记为目标数量）。示例性的，该目标数量为，该资源池内可供目标业务对应的实例集群扩容的虚拟机的数量。

40 例如，该目标数量是指，资源池在目标时间段内目标规格的空闲虚拟机的数量。其中，目标规格指示目标业务在该资源池中指定的虚拟机规格。举例来说，参见表 2，表 2 示例

性示出了目标业务与对应的实例集群、资源池及虚拟机规格的映射关系。

表 2

业务	实例集群	资源池	虚拟机规格
目标业务	实例集群 200	资源池 20	VM_1
	实例集群 210	资源池 21	VM_1
	实例集群 220	资源池 22	VM_1

其中，资源池 20 在目标时间段可提供给目标业务的虚拟机的数量为，资源池 20 在目标时间段内规格为 VM_1 的空闲虚拟机的数量。同理，资源池 21 在目标时间段可提供给目标业务的虚拟机的数量为，资源池 21 在目标时间段内规格为 VM_1 的空闲虚拟机的数量。

可选的，当资源池内多个业务（包括目标业务）指定的虚拟机规格相同时，该目标数量可以是基于资源池在目标时间段内目标规格的空闲虚拟机的数量和该资源池内目标业务对应的占比确定。比如，结合表 2，假设资源池 20 内，目标业务（如业务 A）和业务 B 对应的目标规格均为 VM_1，假设资源池 20 内，目标业务对应的占比为 2/3，若资源池 20 在目标时间段内规格为 VM_1 的空闲虚拟机的数量为 30 台，则资源池 20 在目标时间段内可提供给目标业务的虚拟机的数量为 $30 \times 2/3 = 20$ 台。其中，目标业务对应的占比可以是预设值也可以是基于资源池 20 在目标时间段之前，目标业务对应的实例集群的规模与业务 B 对应的实例集群的规模之比，其中，实例集群的规模可以由实例集群所包含的虚拟机的数量表示。

值得注意的是，目标业务在资源池内还可以对应多种指定的虚拟机规格。本申请对此不做限定，另外，目标业务在不同资源池中指定的虚拟机规格可以相同也可以不同，具体不做限定。表 3 示例性示出了目标业务的另一种映射关系。

表 3

业务	实例集群	资源池	虚拟机规格
目标业务	实例集群 200	资源池 20	VM_1、VM_2
	实例集群 210	资源池 21	VM_2
	实例集群 220	资源池 22	VM_1、VM_3

结合表 3，资源池 20 在目标时间段可提供给目标业务的虚拟机的数量为，资源池 20 在目标时间段内规格为 VM_1 的空闲虚拟机和规格为 VM_2 的空闲虚拟机的数量之和。比如，资源池 20 在目标时间段内规格为 VM_1 的空闲虚拟机的数量为 20 台，规格为 VM_2 的空闲虚拟机的数量为 30 台，则资源池在目标时间段内可提供给目标业务的虚拟机的数量为 50 台。同理，资源池 21 在目标时间段可提供给目标业务的虚拟机的数量为，资源池 21 在目标时间段内规格为 VM_2 的空闲虚拟机的数量。资源池 22 在目标时间段可提供给目标业务的虚拟机的数量为，资源池 22 在目标时间段内规格为 VM_1 的空闲虚拟机和规格为 VM_3 的空闲虚拟机的数量之和。

可选的，当资源池内多个业务（包括目标业务）指定的虚拟机规格有部分或完全重叠时，该目标数量可以是基于资源池在目标时间段内目标规格的空闲虚拟机的数量和该资源池内目标业务对应的占比确定。比如，结合表 3，资源池 20 内目标业务（如业务 A）对应的目标规格为 VM_1 和 VM_2，假设资源池 20 内业务 B 对应的目标规格均为 VM_1，假

设资源池 20 内，目标业务对应的占比为 $\frac{2}{3}$ ，若资源池 20 在目标时间段内规格为 VM_1 的空闲虚拟机的数量为 30 台，以及资源池 20 在目标时间段内规格为 VM_2 的空闲虚拟机的数量为 15 台，则资源池 20 在目标时间段内可提供给目标业务的虚拟机的数量（即目标数量）为 $30 \times \frac{2}{3} + 15 = 35$ 台。再比如，假设资源池 20 内目标业务（如业务 A）和业务 B 对应的目标规格均为 VM_1 和 VM_2，假设资源池 20 内，业务 A 和业务 B 的占比为 2:1，若资源池 20 在目标时间段内规格为 VM_1 的空闲虚拟机和规格为 VM_2 的空闲虚拟机的数量之和为 60 台，资源池 20 在目标时间段内可提供给目标业务的虚拟机的数量（即目标数量）为 $60 \times \frac{2}{3} = 40$ 台，则资源池 20 在目标时间段内可提供给目标业务的虚拟机的数量（即目标数量）为 40 台。

在实际应用中，为方便管理，同一租户针对同一业务在不同资源池中通常指定同一种规格的虚拟机，如下以目标租户在各资源池指定的用于运行目标业务的虚拟机的规格均相同，如以表 2 为例进行说明。需要说明的是，表 2 仅为一种示例，并不构成对本申请实施例涉及的映射关系的限定，如表 2 还可以不包括实例集群这一列，或者，表 2 可以拆分为多个列表存储，本申请对此均不做限定。

接下来介绍如何确定各资源池在目标时间段内可提供给目标业务的虚拟机的数量（或目标数量）。

以一个资源池如资源池 20 为例，示例性地，该预测过程可以包括：基于资源池 20 在一段历史时间（记为第一历史窗口）内目标规格的空闲虚拟机的数量，来预测资源池 20 在目标时间段内目标规格的空闲虚拟机的数量。

示例性的，第一历史窗口可以是目标时间段之前一段相邻的时间段。比如，目标时间段为某一天的 2:00—2:01，第一历史窗口可以是同一天的 1:30—2:00。示例性地，结合表 2，当达到 2:00 时，基于资源池 20 在该日 1:30—2:00 内的多个采样数据，来预测当天 2:00—2:01 内规格为 VM_1 的空闲虚拟机的数量。其中多个采样数据为资源池 20 在当天 1:30—2:00 内的不同时刻规格为 VM_1 的空闲虚拟机的数量。比如，采样间隔为 5 分钟，则多个采样数据可包括资源池 20 分别在该日的 1:30、1:35、1:40、...、2:00 时刻，规格为 VM_1 的空闲虚拟机的数量。类似的，当达到 2:01 时，基于资源池 20 在该日 1:31—2:01 内的多个采样数据，来预测当天 2:01—2:02 内规格为 VM_1 的空闲虚拟机的数量。当达到 2:02 时，基于资源池 20 在该日 1:32—2:02 内的多个采样数据，来预测当天 2:02—2:03 内规格为 VM_1 的空闲虚拟机的数量。其中，第一历史窗口内的采样数据可以从数据库 114 获取，也可以是从其他设备中获取，示例性地，这些采用数据可以是监控模块 113 采集的，本申请实施例对此均不做限定，下文类似之处不再赘述。上述方式也可以称为实时预测，通常应用于对预测精度要求较高或资源池中空闲虚拟机的数量不具有周期规律的场景中。

再示例性的，第一历史窗口为目标时间段的历史同期窗口，如当资源池 20 中目标规格的空闲虚拟机的数量具有周期规律时，第一历史窗口可以是目标时间段的历史同期窗口。比如，当资源池 20 中目标规格的空闲虚拟机的数量具有日内规律时，假设目标时间段为 2024.04.10 的 0:00—24:00，第一历史窗口可以是 2024.04.09 的 0:00—24:00，或第一历史窗口可以是 2024.04.08 的 0:00—24:00 等等。其中，第一历史窗口与目标时间段间隔 k 天，k 取正整数。再比如，当资源池 20 中目标规格的空闲虚拟机的数量具有周内规律时，假设目标时间段为 2024.04.25 的 0:00—24:00，第一历史窗口可以是 2024.04.18 的 0:00—24:00，或第一历史窗口可以是 2024.04.11 的 0:00—24:00 等等。其中，资源池中目标规格的空闲

虚拟机的数量还可能具有月内规律等等不同时间长度内的变化周期，具体不做限定。该预测方式通常应用于资源池中目标规格的空闲虚拟机的数量具有周期规律的场景，可以在保证准确性的基础上，减少算力开销。

需要说明的是，上述示例仅为举例，第一历史窗口的长度、采样间隔、预测间隔、输出数据间隔（可等于目标时间段的长度）、目标时间段的长度等还可以是其他值，具体不做限定。另外需要说明的是，在执行预测时，第一历史窗口的数量可以是一个或多个，如该多个历史窗口可包括目标时间段之前一个或多个相邻时间段，和/或，一个或多个目标时间段的历史同期窗口。该多个第二历史窗口的时间长度可以相同也可以不同或不完全相同，本申请对此也不做限定。比如，目标时间段为 2024.04.25 的 0:00—24:00 时，则对应的多个第一历史窗口可以是 2024.04.18 的 0:00—24:00 和 2024.04.11 的 0:00—24:00，或 2024.04.18 的 0:00—24:00 和 2024.04.24 的 23:00—24:00 等等。另外，前述的预测方式仅为示例，具体不做限定，任何预测方法均适用于本申请实施例。

通过上述方式，分别获取目标业务关联的每个资源池的可用资源信息。其中，上述预测过程可以由云管理平台节点 11 执行，也可以由其他设备（如监控模块 113）执行，具体不做限定，若有其他设备执行，则云管理平台节点 11 从其他设备获取目标业务关联的每个资源池的可用资源信息。假设云管理平台节点 11 获取的目标业务关联的各资源池的可用资源信息如下列表 4 所示。

表 4

业务	时间段	资源池	虚拟机规格	可用资源信息
目标业务	目标时间段	资源池 20	VM_1	100VM
		资源池 21	VM_1	200VM
		资源池 22	VM_1	40VM

应注意，表 4 仅为示意，本申请对资源池的可用资源信息的存储形式及内容不做限定。步骤 402，云管理平台节点 11 获取目标业务的资源需求信息。

目标业务的资源需求信息可用于指示目标业务在目标时间段内的业务流量预测值。示例性地，该预测过程可以包括：基于目标业务在一段历史时间（记为第二历史窗口）内的业务流量数据，来预测目标业务在目标时间段内的业务流量，所得的值即为业务流量预测值。该预测过程可以由云管理平台节点 11 执行。或者，也可以由其他设备（如监控模块 113）执行，如此，云管理平台节点 11 可以从其他设备获取目标业务的资源需求信息。

类似的，第二历史窗口可以是目标时间段之前一段相邻的时间段。或者，目标时间段的历史同期窗口，如目标业务的业务流量具有周期规律时，如日内规律、周内规律等等。

需要说明的是，第二历史窗口的长度、采样间隔、预测间隔、输出数据间隔（可等于目标时间段的长度）、目标时间段的长度等还可以是其他值，具体不做限定。另外需要说明的是，在执行一次预测时，第二历史窗口的数量可以是一个或多个，如多个历史窗口可包括目标时间段之前相邻的一个或多个时间窗口，和/或，一个或多个目标时间段的历史同期窗口，其中，多个第二历史窗口的时间长度可以相同也可以不同或不完全相同，本申请对此不做限定。具体参见前述对基于第一历史窗口预测资源池在目标时间段内空闲的虚拟机数量的相关描述，此处不再赘述。

步骤 403，云管理平台节点 11 生成目标业务在目标时间段内的分流规划。
如下介绍两种确定分流规划的方式：

方式一：基于目标业务关联的每个资源池的可用资源信息，生成目标业务在目标时间段内的分流规划。

该分流规划可指示, 在目标时间段内, 每个资源池可承载目标业务的业务流量的最大值。

5 示例性地,可基于目标业务的资源转换信息,将该资源池在目标时间段内可提供给目标业务的虚拟机数量转换为资源池可承载目标业务的业务流量的最大值。其中,目标业务的资源转换信息用于指示处理目标业务的单个业务请求所需的计算资源数目。其中的计算资源可以是虚拟机或 CPU 核等不同粒度的计算资源,如目标业务的资源转换信息用于指示:运行目标业务的一个业务请求所需的 CPU 核数或 VM 数目等。

10 举例来说,以表 4 为例,假设目标业务的资源转换信息指示一个业务请求需要 0.1 个规格为 VM_1 的 VM,结合表 4 可确定,在目标时间段内目标业务关联的各资源池可承接目标业务的业务流量的最大值,如资源池 20 可承接的目标业务的业务流量最大值为 $100/0.1=1000$ 条业务请求。基于该方式,依次计算目标业务关联的每个资源池在目标时间段内可承接目标业务的业务流量最大值,参见表 5 所示。

表 5

业务	资源池	可用资源信息	可承接目标业务的业务流量最大值 (单位: 业务请求)
目标业务	资源池 20	100 VM	1000 个
	资源池 21	200 VM	2000 个
	资源池 22	300 VM	3000 个

如果目标业务的资源转换信息用于指示:运行目标业务的一个业务请求所需的 CPU 核数等计算资源,则可以结合资源池 20 的可用资源信息、目标业务的资源转换信息和虚拟机规格信息确定目标业务关联的每个资源池在目标时间段内可承接目标业务的业务流量最大值,其中,资源规格信息用于指示规格的虚拟机包括的计算资源数目,如表 1 所示。

20 举例来说, 目标业务的资源转换信息指示一个业务请求需要 0.1 个 CPU 核, 参见表 1 可确定规格为 VM_1 的虚拟机包括 1 个 CPU 核, 则结合表 5 可确定在目标时间段内, 资源池 20 可承接的目标业务的业务流量最大值为 $100 \times 1 / 0.1 = 1000$ 条业务请求。

方式二：基于目标业务的资源需求信息和目标业务关联的每个资源池的可用资源信息，生成目标业务在目标时间段内的分流规划。

25 该分流规划可指示, 在目标时间段内, 被分配入一个或多个资源池的目标业务的业务流量之和等于业务流量预测值, 且在所述时间段内被分配入每个资源池的目标业务的业务流量不超过该资源池可承载的目标业务的业务流量的最大值。换言之, 该分流规划指示在目标时间段内, 将目标业务的业务流量分配至哪些资源池, 以及分配至每个资源池的业务流量是多少。

可以理解,满足该条件的分流规划有多种,比如,继续结合表 5 的示例,假设在目标时间段内,目标业务的业务流量预测值为 1000 个业务请求,则分流规划可以指示将 1000 个业务请求分配入资源池 20,或者,分流规划指示将 500 个业务请求分配入资源池 20,将其余 500 个业务请求分配入资源池 21。或者,分流规划指示将 500 个业务请求分配入资源池 20,将其余 500 个业务请求分配入资源池 22,等等。在一种可选的实施方式中,云管理平台节点 11 可以基于成本策略选择一种分流规划,如该成本策略包括在目标时间段内

将预测的目标业务的业务流量分配入一个或多个资源池所产生的费用最低。

示例性的，该费用包括但不限于：虚拟机费用和通信费用。其中，虚拟机费用指示在目标时间段内，将预测的目标业务的业务流量分配入一个或多个资源池时，每个资源池用于处理被分配的业务流量所需的虚拟机的租赁费用之和。其中，虚拟机的租赁费用可按照机时计费， $\text{机时} = \text{机器数量} \times \text{机器被占用的时间}$ （如单位为：h），如表 1 示出了其中一个资源池内各规格的虚拟机的收费标准。本申请中，虚拟机的租赁费用可包括目标业务的业务流量分配入一个或多个资源池时，每个资源池用于处理被分配的业务流量所需的虚拟机的数量 $\times$ 目标时间段长度（/h） $\times$ 该资源池内该规格的虚拟机每机时的费用。应理解的是，不同资源池中同一规格的虚拟机可能具有不同的价格，这与资源池本身的运营成本（如水费、电费等）有关，比如，位于城市中心的资源池比位于城市郊区的资源池的运营成本高，相应的，运营成本高的资源池中的虚拟机的价格也就越高。

通信费用指示分流节点与每个资源池传输目标业务的业务流量所需的通信费用。示例性地，通信费用可包括，分流节点 12 将目标业务的业务流量发送至各资源池所需的带宽费用，以及各目标资源池返回的业务响应的带宽费用。值得注意的是，分流节点 12 与目标资源池之间的带宽费用可以是 0，如图 3 中，分流节点 12 将目标业务的业务请求分配入实例集群 200 时不需要带宽费用。

进一步，在一种可选的实施方式中，云管理平台节点 11 还可以基于约束条件确定分流规划，比如，约束条件包括但不限于：分流规划满足目标业务的服务等级协议（service level agreement, SLA）。其中，SLA 可以从实例集群的响应时延、抖动等指标进行评估，这里不做重点说明。理论上，租户会选择在满足 SLA 的资源池中部署业务，即目标业务的资源池均能够满足 SLA 的资源池，然而，当目标业务的 SLA 为发生变更时，可能存在目标业务所在的资源池不能满足目标业务的 SLA 的情况。在这种情况下，示例性地，可以由监控模块 113 实时或周期性测量各资源池中用于评估其服务质量所需的信息，从而计算该资源池的 SLA，并将计算的各资源池的 SLA 存储至数据库 114，令云管理平台节点 11 可以获取每个资源池的 SLA，并从目标业务关联的多个资源池中筛选出能够满足目标业务的 SLA 的资源池。另外，目标业务的 SLA 等信息可以由业务管理员在云管理平台节点 11 中进行配置更新，并将更新后的 SLA 的信息存储至数据库 114 中，或者也可以存储于其他设备中，具体不做限定。需要说明的，约束条件还可以包括其他条件，比如，不超过 AZ 或 Region 的最大连接数等，本申请对此不做限定。

在前述的两种方式中，分流规划可以按照绝对值格式输出，指示各资源池可承接目标业务的业务请求的数量。如在目标时间段内，资源池 20、资源池 21、资源池 22 可承接目标业务的业务流量最大值分别为 1000 个业务请求、2000 个业务请求、3000 个业务请求。或者，分流规划按照流量比例格式输出，指示目标业务的业务流量在目标时间段内被分配入每个资源池比例。比如，基于表 5 可确定，在目标时间段内，目标业务的业务流量分配入资源池 20、资源池 21、资源池 22 的比例为 1:2:3。当分流规划指示流量比例时，可以与前述的负载均衡策略中按照预设的流量比例进行分流的分流节点兼容，提高本申请技术方案的适用性。

可选的，该分流规划还可以指示目标业务的业务流量在目标时间段内被分配入每个资源池优先级信息。示例性地，目标业务的业务流量被优先分配入优先级较高的资源池中，当分配入该资源池的业务流量达到该资源池可承接的业务流量最大值时，再将目标业务的

业务流量分配入优先级次之的资源池，依此类推。

其中，资源池的优先级信息可根据各资源池处理目标业务的业务流量所需的费用来确定。比如，多个资源池的优先级信息可基于该多个资源池的费用比确定，示例性的，多个资源池的费用比可以是每个资源池处理单个业务请求的费用之比，资源池处理单个业务请求所需的费用越低则资源池的优先级越高。其中，虚拟机费用指处理单个业务请求所需的虚拟机的租赁费用。如果目标业务在一个资源池指定了多种虚拟机规格，则虚拟机费用可按照多种规格的虚拟机费用的加权值确定，比如，目标业务在资源池 20 内指定的虚拟机规格包括 VM_1 和 VM_2，其中，资源池 20 在目标时间段内可提供给目标业务的规格为 VM_1 的虚拟机为 30 台，可提供的规格为 VM_2 的虚拟机为 20 台，规格为 VM_1 的虚拟机每机时的费用为 a，规格为 VM_2 的虚拟机每机时的费用为 b，则平均每台虚拟机的费用为 $a*3/5+b*2/5$ 。再计算平均每台虚拟机可处理的业务请求的数量，从而计算资源池 20 内处理一条业务请求所需的虚拟机的平均租赁费用。

通信费用指传输单个业务请求所需的通信费用，如分流节点 12 将业务流量发送至资源池所需的带宽费用，以及接收资源池返回的该业务请求的业务响应的带宽费用。值得注意的是，当分流节点 12 位于该资源池时，分流节点 12 与该资源池之间的带宽费用为 0，参见前述的说明，此处不再赘述。

需要说明的是，确定多个资源池的优先级信息的方式有多种，上述仅为举例，本申请实施例对此不做限定。

步骤 404，云管理平台节点 11 将分流规划发送给分流节点 12。

步骤 405，分流节点 12 按照分流规划，将目标业务在目标时间段内的业务流量分配至一个或多个资源池。

示例性的，分流规划指示在目标时间段（如 2024.04.09 的 16:00—16:01），资源池 20、资源池 21、资源池 22 可承接目标业务的业务流量最大值分别为 1000 个业务请求、2000 个业务请求、3000 个业务请求。分流节点 12 按照该分流规划，将目标业务在 2024.04.09 的 16:00—16:01 内的业务流量分配至一个或多个资源池。分配顺序可随机分配，或按照负载均衡轮询分配，或按预设顺序分配，或按上一次分流规划指示的多个资源池的优先级信息确定分配顺序。其中，分配入资源池 21 的业务请求的数量不超过 1000 个，分配入资源池 21 的业务请求的数量不超过 2000 个，分配入资源池 22 的业务请求的数量不超过 3000 个。

若该分流规划还指示目标业务对应的多个资源池的优先级排序为：资源池 20>资源池 21>资源池 22，则在 2024.04.09 的 16:00—16:01 段内，分流节点 12 首先将接收到的目标业务的业务流量优先分配入资源池 20，当分配入资源池 20 的业务流量达到 1000 条业务请求时，再将目标业务的业务流量分配入资源池 21，当分配入资源池 21 的业务流量达到 2000 条业务请求时，再将目标业务的业务流量分配入资源池 22，直至分配入资源池 22 的业务流量达到 3000 个。应注意，上述分流规划是一种预分配，分流节点 12 在目标时间段内实际接收到的目标业务的业务请求的数量并非定值，上述分流规划并非限定其数量为 6000（1000+2000+3000）个，如分流节点在目标时间段内接收到的目标业务的业务请求数量可能是 600 个，则分流节点 12 将接收到的 600 个业务请求分配入资源池 20 即可。下文类似之处不再赘述。

再示例性的，分流规划指示在目标时间段内，目标业务的业务流量分配入资源池 20

(或实例集群 200) 的业务请求的数量为 1000 个, 分配入资源池 21 (或实例集群 210) 的业务请求的数量为 500, 分配入资源池 22 (或实例) 的业务请求的数量为 400。分流节点 12 按照该分流规划, 将在目标时间段内接收到的业务请求分配至一个或多个资源池。

若该分流规划还指示目标业务对应的多个资源池的优先级排序为: 资源池 20 > 资源池 22 > 资源池 21, 则在 2024.04.09 的 16:00—16:01 段内, 分流节点 12 首先将接收到的目标业务的业务请求发送至资源池 20, 当发送至资源池 20 的业务请求的数量达到 1000 个时, 若继续接收到目标业务的业务请求, 再将接收到的目标业务的业务请求发送至资源池 22, 当发送至资源池 22 的业务请求的数量达到 400 个时, 若继续接收到目标业务的业务请求, 再将接收到的目标业务的业务请求发送至资源池 21, 直至数量达到 500。

需要说明的是, 步骤 404 为可选的步骤, 若云管理平台节点 11 和分流节点 12 位于同一计算设备时, 云管理平台节点 11 可不向分流节点 12 发送分流规划。另外, 上述在确定每个资源池在目标时间段内可提供给目标业务的实例集群的实例数量时, 仅包括资源池在目标时间段内可提供目标业务的实例集群扩容的虚拟机数量, 可选的, 还可以包括目标业务的实例集群本身所包括的虚拟机数量, 比如在目标时间段的起始时刻, 该实例集群内的虚拟机的数量。

基于上述设计, 云管理平台节点 11 基于目标业务关联的各资源池的可用资源信息, 确定目标业务在目标时间段内的分流规划, 从而可以预先确定出不同时间段内目标业务的业务流量分配入各资源池的业务流量, 有助于更好的实现流量分配规划, 避免分配入实例集群的业务流量不超过其所在资源池的承载能力, 可以实现计算资源与业务需求的协同调度, 实现细粒度分流, 减少或避免流量突发场景中业务请求大规模拥塞的情况。进一步, 提供给目标租户一种新的成本策略, 当采用成本策略确定分流规划时, 还可以有效降低目标租户的成本费用。

基于与方法实施例同一发明构思, 本申请实施例还提供了一种计算装置, 该计算装置用于执行上述图 4 或图 5 的方法实施例中云管理平台节点 11 执行的方法。如图 5 所示, 计算装置 500 包括获取模块 501、确定模块 502; 具体地, 在计算装置 500 中, 各模块之间通过通信通路建立连接。

获取模块 501, 用于获取目标业务关联的每个资源池的可用资源信息。具体实现方式请参见图 4 中的步骤 401 的描述此处不再赘述。

确定模块 502, 用于根据目标业务关联的每个资源池的可用资源信息, 确定目标业务在未来一个时间段内的分流规划。具体实现方式请参见图 4 中的步骤 402 的描述, 此处不再赘述。

在一种可能的实现方式中, 每个资源池的可用资源信息指示: 该资源池在所述时间段内空闲实例的数量, 所述空闲实例为该资源池在所述时间段内未被其他租户占用的。

在一种可能的实现方式中, 所述分流规划指示所述目标业务的业务流量在所述时间段内被分配入每个资源池优先级信息。

在一种可能的实现方式中, 所述获取模块 501, 还用于获取所述目标业务的资源需求信息, 所述资源需求信息指示所述目标业务在所述时间段内的业务流量预测值; 具体实现方式请参见图 4 中的步骤 402 的描述, 此处不再赘述。

在一种可能的实现方式中, 在所述时间段内被分配入每个资源池的所述目标业务的业务流量之和等于所述业务流量预测值, 且在所述时间段内被分配入每个资源池的所述目标

业务的业务流量不超过该资源池可承载的所述目标业务的业务流量的最大值。

在一种可能的实现方式中，所述分流规划满足成本策略，所述成本策略包括在所述时间段内被分配入每个资源池的所述目标业务的业务流量所产生的费用最低。

在一种可能的实现方式中，在所述时间段内被分配入每个资源池的所述目标业务的业务流量所产生的费用最低包括实例费用和通信费用；所述实例费用指示在所述时间段内被分配入每个资源池的所述目标业务的业务流量所需的实例的租赁费用，所述通信费用指示所述分流节点与每个资源池传输所述目标业务的业务流量所需的通信费用。

在一种可能的实现方式中，每个资源池可承接所述目标业务的业务流量的最大值基于所述资源池在所述时间段内可提供给所述目标业务的实例集群的实例数量及所述目标业务的资源转换信息确定的，所述资源转换信息指示处理所述目标业务的单个业务请求所需的实例的数量。

示例性的，接下来以计算装置 500 中的确定模块 502 为例，介绍确定模块 502 的实现方式。类似的，获取模块 501 的实现方式可以参考确定模块 502 的实现方式。

当通过软件实现时，确定模块 502 可以是运行在计算机设备上的应用程序或代码块。其中，计算机设备可以是物理主机、虚拟机、容器等计算设备中的至少一种。进一步地，上述计算机设备可以是一台或者多台。例如，确定模块 502 可以是运行在多个主机/虚拟机/容器上的应用程序。需要说明的是，用于运行该应用程序的多个主机/虚拟机/容器可以分布在相同的可用区（availability zone, AZ）中，也可以分布在不同的 AZ 中。用于运行该应用程序的多个主机/虚拟机/容器可以分布在相同的区域（region）中，也可以分布在不同的 region 中。其中，通常一个 region 可以包括多个 AZ。

同样，用于运行该应用程序的多个主机/虚拟机/容器可以分布在同一个虚拟私有云（virtual private cloud, VPC）中，也可以分布在多个 VPC 中。其中，通常一个 region 可以包括多个 VPC，而一个 VPC 中可以包括多个 AZ。

当通过硬件实现时，确定模块 502 中可以包括至少一个计算设备，如服务器等。或者，确定模块 502 也可以是利用专用集成电路（application-specific integrated circuit, ASIC）实现、或可编程逻辑器件（programmable logic device, PLD）实现的设备等。其中，上述 PLD 可以是复杂程序逻辑器件（complex programmable logical device, CPLD）、现场可编程门阵列（field-programmable gate array, FPGA）、通用阵列逻辑（generic array logic, GAL）或其任意组合实现。

确定模块 502 包括的多个计算设备可以分布在相同的 AZ 中，也可以分布在不同的 AZ 中。确定模块 502 包括的多个计算设备可以分布在相同的 region 中，也可以分布在不同的 region 中。同样，确定模块 502 包括的多个计算设备可以分布在同一个 VPC 中，也可以分布在多个 VPC 中。其中，所述多个计算设备可以是服务器、ASIC、PLD、CPLD、FPGA 和 GAL 等计算设备的任意组合。

本申请还提供一种计算设备 600。如图 6 所示，计算设备 600 包括：总线 602、处理器 604、存储器 606 和通信接口 608。处理器 604、存储器 606 和通信接口 608 之间通过总线 602 通信。计算设备 600 可以是服务器或终端设备。应理解，本申请不限定计算设备 600 中的处理器、存储器的个数。

总线 602 可以是外设部件互连标准（peripheral component interconnect, PCI）总线或扩展工业标准结构（extended industry standard architecture, EISA）总线等。总线可以分为

地址总线、数据总线、控制总线等。为便于表示，图6中仅用一条线表示，但并不表示仅有一根总线或一种类型的总线。总线602可包括在计算设备600各个部件（例如，存储器606、处理器604、通信接口608）之间传送信息的通路。

5 处理器604可以包括中央处理器(central processing unit, CPU)、图形处理器(graphics processing unit, GPU)、微处理器(micro processor, MP)或者数字信号处理器(digital signal processor, DSP)等处理器中的任意一种或多种。

存储器606可以包括易失性存储器(volatile memory)，例如随机存取存储器(random access memory, RAM)。处理器604还可以包括非易失性存储器(non-volatile memory)，例如只读存储器(read-only memory, ROM)、快闪存储器、机械硬盘(hard disk drive, HDD)或固态硬盘(solid state drive, SSD)。

存储器606中存储有可执行的程序代码，处理器604执行该可执行的程序代码以分别实现前述获取模块501、确定模块502的功能，从而实现业务分流方法。也即，存储器606上存有计算装置500用于执行本申请提供的业务分流方法的指令。

15 通信接口608使用例如但不限于网络接口卡、收发器一类的收发模块，来实现计算设备600与其他设备或通信网络之间的通信。

本申请实施例还提供了一种计算设备集群。该计算设备集群包括至少一台计算设备。该计算设备可以是服务器，例如是中心服务器、边缘服务器，或者是本地数据中心中的本地服务器。在一些实施例中，计算设备也可以是台式机、笔记本电脑或者智能手机等终端设备。

20 如图7所示，所述计算设备集群包括至少一个计算设备600。计算设备集群中的一个或多个计算设备600中的存储器606中可以存有相同的计算装置500用于执行本申请提供的业务分流方法的指令。

在一些可能的实现方式中，该计算设备集群中的一个或多个计算设备600也可以用于执行计算装置500用于执行本申请提供的业务分流方法的部分指令。换言之，一个或多个计算设备600的组合可以共同执行计算装置500用于执行本申请提供的业务分流方法的指令。

需要说明的是，计算设备集群中的不同的计算设备600中的存储器606可以存储不同的指令，用于执行计算装置500的部分功能。也即，不同的计算设备600中的存储器606存储的指令可以实现获取模块501、确定模块602中的一个或多个模块的功能。

30 本申请实施例还提供了一种包含指令的计算机程序产品。所述计算机程序产品可以是包含指令的，能够运行在计算设备上或被储存在任何可用介质中的软件或程序产品。当所述计算机程序产品在至少一个计算机设备上运行时，使得至少一个计算机设备执行上述图4或图5实施例中的云管理平台节点11所执行的业务分流方法，参见图4或图5各步骤的描述，此处不再赘述。

35 本申请实施例还提供了一种计算机可读存储介质。所述计算机可读存储介质可以是计算设备能够存储的任何可用介质或者是包含一个或多个可用介质的数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘)等。该计算机可读存储介质包括指令，所述指令指示计算设备执行上述图4或图5实施例中的云管理平台节点11所执行的业务分流方法，参见图40 4或图5各步骤的描述，此处不再赘述。

通过以上实施方式的描述, 所属领域的技术人员可以了解到, 为描述的方便和简洁, 仅以上述各功能模块的划分进行举例说明, 实际应用中, 可以根据需要而将上述功能分配由不同的功能模块完成, 即将装置的内部结构划分成不同的功能模块, 以完成以上描述的全部或者部分功能。

5 在本申请所提供的几个实施例中, 应该理解到, 所揭露的装置和方法, 可以通过其他的方式实现。例如, 以上所描述的装置实施例仅仅是示意性的, 例如, 模块或单元的划分, 仅仅为一种逻辑功能划分, 实际实现时可以有另外的划分方式, 例如多个单元或组件可以结合或者可以集成到另一个装置, 或一些特征可以忽略, 或不执行。另一点, 所显示或讨论的相互的耦合或直接耦合或通信连接可以是通过一些接口, 装置或单元的间接耦合或通信连接, 可以是电性, 机械或其他的形式。

10 作为分离部件说明的单元可以是或者也可以不是物理上分开的, 作为单元显示的部件可以是一个物理单元或多个物理单元, 即可以位于一个地方, 或者也可以分布到多个不同地方。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

15 另外, 在本申请各个实施例中的各功能单元(或模块)可以集成在一个处理单元中, 也可以是各个单元单独物理存在, 也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现, 也可以采用软件功能单元的形式实现。

20 集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用, 可以存储在一个可读取存储介质中。基于这样的理解, 本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来, 该软件产品存储在一个存储介质中, 包括若干指令用以使得一个设备(可以是单片机, 芯片等)或处理器(processor)执行本申请各个实施例方法的全部或部分步骤。而前述的存储介质包括: U 盘、移动硬盘、只读存储器(read only memory, ROM)、随机存取存储器(random access memory, RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

25 可选的, 本申请实施例中的计算机执行指令也可以称之为应用程序代码, 本申请实施例对此不作具体限定。

30 在上述实施例中, 可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时, 可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时, 全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中, 或者从一个计算机可读存储介质向另一个计算机可读存储介质传输, 例如, 所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包括一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质, (例如, 软盘、硬盘、磁带)、光介质(例如, DVD)、或者半导体介质(例如固态硬盘(Solid State Disk, SSD))等。

40 本申请实施例中所描述的各种说明性的逻辑单元和电路可以通过通用处理器, 数字信号处理器, 专用集成电路(ASIC), 现场可编程门阵列(FPGA)或其它可编程逻辑装置, 离散门或晶体管逻辑, 离散硬件部件, 或上述任何组合的设计来实现或操作所描述的功能。

通用处理器可以为微处理器,可选地,该通用处理器也可以为任何传统的处理器、控制器、微控制器或状态机。处理器也可以通过计算装置的组合来实现,例如数字信号处理器和微处理器,多个微处理器,一个或多个微处理器联合一个数字信号处理器核,或任何其它类似的配置来实现。

5 本申请实施例中所描述的方法或算法的步骤可以直接嵌入硬件、处理器执行的软件单元、或者这两者的结合。软件单元可以存储于RAM存储器、闪存、ROM存储器、EPROM存储器、EEPROM存储器、寄存器、硬盘、可移动磁盘、CD-ROM或本领域中其它任意形式的存储媒介中。示例性地,存储媒介可以与处理器连接,以使得处理器可以从存储媒介中读取信息,并可以向存储媒介存写信息。可选地,存储媒介还可以集成到处理器中。

10 处理器和存储媒介可以设置于ASIC中。

 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

15 尽管结合具体特征及其实施例对本申请进行了描述,显而易见的,在不脱离本申请的精神和范围的情况下,可对其进行各种修改和组合。相应地,本说明书和附图仅仅是所附权利要求所界定的本申请的示例性说明,且视为已覆盖本申请范围内的任意和所有修改、变化、组合或等同物。显然,本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的范围。这样,倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内,则本申请也意图包括这些改动和变型在内。

20

权 利 要 求

- 1.一种业务分流方法，其特征在于，所述方法应用于云计算系统，所述系统至少包括云管理平台节点、多个资源池和分流节点，其中，每个资源池包括多个主机用于为多个租户运行业务，每个主机上运行至少一个实例，每个资源池内创建有目标租户的目标业务对应的实例集群，所述实例集群内的实例用于运行所述目标业务；
- 5 所述方法包括：
- 所述云管理平台节点获取所述目标业务关联的每个资源池的可用资源信息，其中，每个资源池的可用资源信息指示该资源池在一个时间段内可提供给所述目标业务的实例集群的实例数量；
- 10 所述云管理平台节点基于每个资源池的可用资源信息生成所述目标业务在所述时间段内的分流规划；
- 所述分流节点按照所述分流规划将所述目标业务在所述时间段内的业务流量分配至所述多个资源池。
- 2.如权利要求 1 所述的方法，其特征在于，每个资源池的可用资源信息指示：该资源池在所述时间段内空闲实例的数量，所述空闲实例为该资源池在所述时间段内未被其他租户占用的。
- 15 3.如权利要求 1 或 2 所述的方法，其特征在于，所述分流规划指示在所述时间段内每个资源池可承接所述目标业务的业务流量的最大值，或所述目标业务的业务流量在所述时间段内被分配入每个资源池比例。
- 20 4.如权利要求 3 所述的方法，其特征在于，所述分流规划指示所述目标业务的业务流量在所述时间段内被分配入每个资源池优先级信息。
- 5.如权利要求 1-4 任一项所述的方法，其特征在于，所述方法还包括：
- 所述云管理平台节点获取所述目标业务的资源需求信息，所述资源需求信息指示所述目标业务在所述时间段内的业务流量预测值；
- 25 所述云管理平台节点基于每个资源池的可用资源信息生成所述目标业务在所述时间段内的分流规划，包括：
- 所述云管理平台节点基于所述资源需求信息和每个资源池的可用资源信息生成所述分流规划。
- 6.如权利要求 5 所述的方法，其特征在于，在所述时间段内被分配入每个资源池的所述目标业务的业务流量之和等于所述业务流量预测值，且在所述时间段内被分配入每个资源池的所述目标业务的业务流量不超过该资源池可承载的所述目标业务的业务流量的最大值。
- 30 7.如权利要求 5 或 6 所述的方法，其特征在于，所述分流规划满足成本策略，所述成本策略包括在所述时间段内被分配入每个资源池的所述目标业务的业务流量所产生的费用最低。
- 35 8.如权利要求 7 所述的方法，其特征在于，在所述时间段内被分配入每个资源池的所述目标业务的业务流量所产生的费用最低包括实例费用和通信费用；所述实例费用指示在所述时间段内被分配入每个资源池的所述目标业务的业务流量所需的实例的租赁费用，所述通信费用指示所述分流节点与每个资源池传输所述目标业务的业务流量所需的通信费

用。

9.如权利要求 3-8 任一项所述的方法,其特征在于,每个资源池可承接所述目标业务的业务流量的最大值基于所述资源池在所述时间段内可提供给所述目标业务的实例集群的实例数量及所述目标业务的资源转换信息确定的,所述资源转换信息指示处理所述目标业务的单个业务请求所需的实例的数量。

10.一种业务分流装置,其特征在于,所述装置包括:

获取模块,用于获取所述目标业务关联的每个资源池的可用资源信息,其中,每个资源池的可用资源信息指示该资源池在一个时间段内可提供给所述目标业务的实例集群的实例数量;

处理模块,用于基于每个资源池的可用资源信息生成所述目标业务在所述时间段内的分流规划。

11.如权利要求 10 所述的装置,其特征在于,每个资源池的可用资源信息指示:该资源池在所述时间段内空闲实例的数量,所述空闲实例为该资源池在所述时间段内未被其他租户占用的。

12.如权利要求 10 或 11 所述的装置,其特征在于,所述分流规划指示所述目标业务的业务流量在所述时间段内被分配入每个资源池优先级信息。

13.如权利要求 10-12 任一项所述的装置,其特征在于,所述获取模块,还用于获取所述目标业务的资源需求信息,所述资源需求信息指示所述目标业务在所述时间段内的业务流量预测值;

所述处理模块在基于所述资源需求信息和每个资源池的可用资源信息生成所述分流规划。

14.如权利要求 13 所述的装置,其特征在于,在所述时间段内被分配入每个资源池的所述目标业务的业务流量之和等于所述业务流量预测值,且在所述时间段内被分配入每个资源池的所述目标业务的业务流量不超过该资源池可承载的所述目标业务的业务流量的最大值。

15.如权利要求 13 或 14 所述的装置,其特征在于,所述分流规划满足成本策略,所述成本策略包括在所述时间段内被分配入每个资源池的所述目标业务的业务流量所产生的费用最低。

16.如权利要求 15 所述的装置,其特征在于,在所述时间段内被分配入每个资源池的所述目标业务的业务流量所产生的费用最低包括实例费用和通信费用;所述实例费用指示在所述时间段内被分配入每个资源池的所述目标业务的业务流量所需的实例的租赁费用,所述通信费用指示所述分流节点与每个资源池传输所述目标业务的业务流量所需的通信费用。

17.如权利要求 12-16 任一项所述的装置,其特征在于,每个资源池可承接所述目标业务的业务流量的最大值基于所述资源池在所述时间段内可提供给所述目标业务的实例集群的实例数量及所述目标业务的资源转换信息确定的,所述资源转换信息指示处理所述目标业务的单个业务请求所需的实例的数量。

18.一种云计算系统,其特征在于,所述系统至少包括云管理平台节点、多个资源池和分流节点,其中,每个资源池包括多个主机用于为多个租户运行业务,每个主机上运行至少一个实例,每个资源池内创建有目标租户的目标业务对应的实例集群,所述实例集群内

的实例用于运行所述目标业务；

所述云管理平台节点，用于获取所述目标业务关联的每个资源池的可用资源信息，其中，每个资源池的可用资源信息指示该资源池在一个时间段内可提供给所述目标业务的实例集群的实例数量；基于每个资源池的可用资源信息生成所述目标业务在所述时间段内的分流规划；

所述分流节点，用于按照所述分流规划将所述目标业务在所述时间段内的业务流量分配至所述多个资源池。

19.一种计算设备，其特征在于，所述装置包括处理器和存储器；

所述存储器用于存储计算机执行指令；

所述处理器用于执行所述存储器所存储的计算机执行指令，以使所述计算设备执行如权利要求 1-9 中任一项所述的方法。

20.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储程序，所述程序被处理器调用时，权利要求 1-9 任一所述的方法被执行。

21.一种包含指令的计算机程序产品，其特征在于，当所述指令被计算机设备集群运行时，使得所述计算机设备集群执行如权利要求的 1-9 中任一项所述的方法。