(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2023 年 6 月 8 日 (08.06.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/098614 A1

(51) 国际专利分类号:
G06F 9/455 (2018.01)

(21) 国际申请号: PCT/CN2022/134647

(22) 国际申请日: 2022 年 11 月 28 日 (28.11.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111450334.5 2021年11月30日 (30.11.2021) CN

(71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 蔡灏旻 (CAI, Haomin); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 敬锐 (JING, Rui); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 雷钟凯 (LEI, Zhongkai); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 卢景晓 (LU, Jingxiao); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (CN)。

(54) **Title:** CLOUD INSTANCE CAPACITY EXPANSION/REDUCTION METHOD AND RELATED DEVICE THEREFOR

(54) 发明名称: 一种云实例的扩缩容方法及其相关设备

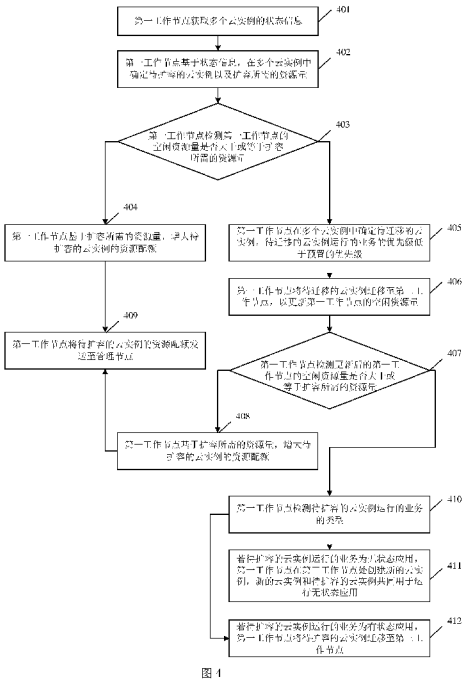

图 4

401 A first worker node acquires state information of a plurality of cloud instances

402 The first worker node determines, on the basis of the state information and from among the plurality of cloud instances, a cloud instance to be subjected to capacity expansion and a resource quantity that is required for capacity expansion

403 The first worker node detects whether the quantity of idle resources of the first worker node is greater than or equal to the resource quantity that is required for capacity expansion

404 The first worder node increases, on the basis of the resource quantity that is required for capacity expansion, the resource quota of the cloud instance to be subjected to capacity expansion

405 The first worker node determines, from among the plurality of cloud instances, a cloud instance to be migrated, wherein the priority of a service that is run by the cloud instance to be migrated is lower than a preset priority

406 The first worker node migrates, to a second worker node, the cloud instance to be migrated, so as to update the quantity of the idle resources of the first worker node

407 The first worker node detects whether an updated quantity of idle resources of the first worker node is greater than or equal to the resource quantity that is required for capacity expansion

408 The first worker node increases, on the basis of the resource quantity that is required for capacity expansion, the resource quota of the cloud instance to be subjected to capacity expansion

409 The first worker node sends, to a master node, the resource quota of the cloud instance to be subjected to capacity expansion

410 The first worker node detects the type of a service that is run by the cloud instance to be subjected to capacity expansion

411 If the service that is run by the cloud instance to be subjected to capacity expansion is a stateless application, the first worker node creates a new cloud instance at a third worker node, wherein the new cloud instance and the cloud instance to be subjected to capacity expansion are jointly used for running the stateless application

412 If the service that is run by the cloud instance to be subjected to capacity expansion is a stateful application, the first worker node migrates, to the third worker node, the cloud instance to be subjected to capacity expansion

(57) **Abstract:** Provided in the present application are a cloud instance capacity expansion/reduction method and a related device therefor, which can ensure that a service that is run by a cloud instance is not interrupted when the resource quota of the cloud instance is increased or decreased. The method of the present application comprises: a first worker node acquiring state information of a plurality of cloud instances; the first worker node determining, on the basis of the state information and from among the plurality of cloud instances, a cloud instance to be subjected to capacity expansion and a resource quantity that is required for capacity expansion; and

WO 2023/098614 A1

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

if the quantity of idle resources of the first worker node is greater than or equal to the resource quantity that is required for capacity expansion, the first worker node increasing, on the basis of the resource quantity that is required for capacity expansion, the resource quota of the cloud instance to be subjected to capacity expansion.

(57) 摘要：本申请提供一种云实例的扩缩容方法及其相关设备，可在增大或减小云实例的资源配额时，确保云实例运行的业务不会中断。本申请的方法包括：第一工作节点获取多个云实例的状态信息；第一工作节点基于状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量；若第一工作节点的空闲资源量大于或等于扩容所需的资源量，第一工作节点基于扩容所需的资源量，增大待扩容的云实例的资源配额。

一种云实例的扩缩容方法及其相关设备

本申请要求于 2021 年 11 月 30 日提交中国专利局，申请号为 202111450334.5，发明名称为“一种云实例的扩缩容方法及其相关设备”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

本申请涉及云技术领域，尤其涉及一种云实例的扩缩容方法及其相关设备。

背景技术

随着技术的飞速发展，云服务系统的规模越来越大。云服务系统通常包含多个工作节点（worker）和管理节点（master），其中，每个工作节点上部署有多个容器（docker），管理节点可对所有容器进行集中管理。

目前，云服务系统以 kubernetes 作为容器的管理标准，可对容器实现编排部署、灰度升降级、自动扩缩容等功能。在自动伸缩功能中，kubernetes 可支持两种自动两种自动伸缩方法，分别为垂直 pod（容器组）自动伸缩（vertical pod autoscale，VPA）和水平 pod 自动伸缩（horizontal pod autoscale，HPA）。在 VPA 中，管理节点可基于某个工作节点中 pod 的资源占用率计算出 pod 的资源配额推荐值，并发送至该工作节点。当该工作节点创建 pod 时，可基于该推荐值为 pod 设置新的资源配额（例如，增大 pod 的资源配额或减小 pod 的资源配额，相当于扩容或缩容）。

前述过程中，由于修改 pod 的资源配额的时间点，只能在创建 pod 的时候。当需要修改某个 pod 的资源配额时，工作节点只能先释放该 pod，并在重新创建该 pod 时才能实现资源配额的修改，这样会导致该 pod 运行的业务中断。

发明内容

本申请实施例提供了一种云实例的扩缩容方法及其相关设备，可在增大或减小云实例的资源配额时，确保云实例运行的业务不会中断。

本申请实施例的第一方面提供了一种云实例的扩缩容方法，该方法应用于云服务系统，云服务系统包含多个工作节点，也其中一个工作节点进行示意性介绍，并称该工作节点为第一工作节点，该方法包括：

第一工作节点部署有多个云实例，例如，多个云实例可以为多个容器，又如，多个云实例可以为多组容器，再如，多个云实例可以为多个虚拟机，还如，多个云实例可以为多组虚拟机等等。第一工作节点可获取这多个云实例的状态信息，每个云实例的状态信息可用于指示该云实例所运行的业务的状态。

得到多个云实例的状态信息后，第一工作节点可基于这多个云实例的状态信息，对这多个云实例逐个进行分析，从而在这多个云实例中确定待扩容的云实例以及对带扩容的云实例进行扩容所需的资源量。

第一工作节点可确定第一工作节点的空闲资源量，并检测第一工作节点的空闲资源量是否大于或等于扩容所需的资源量，若第一工作节点的空闲资源量大于或等于扩容所需的资源

量，说明第一工作节点的空闲资源是充足的，故第一工作节点可直接对待扩容的云实例直接进行扩容处理，即基于扩容所需的资源量，增大待扩容的云实例的资源配额，从而实现云实例的扩容。

从上述方法可以看出：第一工作节点在获取多个云实例的状态信息后，可基于这部分状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量。若第一工作节点的空闲资源量大于或等于扩容所需的资源量，第一工作节点可基于扩容所需的资源量，增大待扩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例扩容机制，第一工作节点自行确定待扩容的云实例后，可通过实时修改 cgroup 配置，以增大待扩容的云实例的资源配额，这对于待扩容的云实例运行的业务而言，是不感知的，故不会导致待扩容的云实例运行的业务中断。

在一种可能的实现方式中，第一工作节点基于状态信息，在多个云实例中确定待扩容的第一云实例以及扩容所需的扩容所需的资源量包括：在多个云实例中，第一工作节点将状态信息满足预置的扩容条件的云实例确定为待扩容的云实例；第一工作节点基于待扩容的云实例的状态信息，确定扩容所需的扩容所需的资源量。前述实现方式中，对于多个云实例中的任意一个云实例，第一工作节点可检测该云实例的状态信息是否满足预置的扩容条件，若满足，则确定需要进行扩容，即将该云实例确定为待扩容的云实例，若不满足，则确定该云实例不需要进行扩容，结束操作。在将该云实例确定为待扩容的云实例后，还可基于该云实例的状态信息准确计算出该云实例扩容所需的资源量。

在一种可能的实现方式中，云服务系统还包含第二工作节点，该方法还包括：若第一工作节点的空闲资源量小于扩容所需的资源量，第一工作节点在多个云实例中确定待迁移的云实例，待迁移的云实例运行的业务的优先级低于预置的优先级；第一工作节点将待迁移的云实例迁移至第二工作节点，以更新第一工作节点的空闲资源量；若更新后的第一工作节点的空闲资源量大于或等于扩容所需的资源量，第一工作节点基于扩容所需的资源量，增大待扩容的云实例的资源配额。前述实现方式中，若第一工作节点的空闲资源量小于该云实例扩容所需的资源量，说明第一工作节点的空闲资源是不足的，第一工作节点可在第一工作节点的多个云实例中，确定至少一个待迁移的云实例，这些待迁移的云实例运行的业务的优先级低于预置的优先级（即这些云实例所运行的业务往往优先级较低），故可将这些待迁移的云实例迁移至第二工作节点，那么，第一工作节点中被分配至这些待迁移的云实例的资源被释放，成为新的空闲资源，从而更新了第一工作节点的空闲资源量（即增大了第一工作节点的空闲资源量）。然后，第一工作节点可检测更新后的第一工作节点的空闲资源量是否大于或等于扩容所需的资源量，若更新后的第一工作节点的空闲资源量大于或等于该云实例（待扩容的云实例）扩容所需的资源量，说明更新后的第一工作节点的空闲资源是充足的，可利用这部分资源直接对该云实例进行扩容，故第一工作节点可基于该云实例扩容所需的资源量，增大该云实例的资源配额，从而实现云实例的扩容。

在一种可能的实现方式中，云服务系统还包含第三工作节点，该方法还包括：若更新后的第一工作节点的空闲资源量小于扩容所需的资源量，第一工作节点检测待扩容的云实例运行的业务的类型；若待扩容的云实例运行的业务为无状态应用，第一工作节点在第三工作节点处创建新的云实例，新的云实例和待扩容的云实例共同用于运行无状态应用；若待扩容的

云实例运行的业务为有状态应用，第一工作节点将待扩容的云实例迁移至第三工作节点。前述实现方式中，若更新后的第一工作节点的空闲资源量小于该云实例（待扩容的云实例）扩容所需的资源量，说明更新后的第一工作节点的空闲资源是不足的，第一工作节点可先检测该云实例运行的业务（也可以理解为该云实例运行的应用）的类型，以基于该云实例运行的业务类型进行相应的处理：若该云实例运行的业务为无状态应用，第一工作节点可申请在第三工作节点处创建新的云实例，那么，该云实例和新的云实例可共同运行该云实例原先所运行的业务，也就相当于实现了扩容。若该云实例运行的业务为有状态应用，节点代理可将该云实例迁移至在第三工作节点，其中，第三工作节点的空闲资源量大于或等于该云实例扩容所需的资源量，那么，在将该云实例迁移至第三工作节点后，可增大该云实例的资源配额，也就相当于完成了扩容。

在一种可能的实现方式中，状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。可见，云实例的状态信息与云实例的业务逻辑相关。

在一种可能的实现方式中，预置的扩容条件包括以下至少一种：资源占用率大于或等于预置的第一资源占用率、负载程度大于或等于预置的第一负载程度和业务成功率小于预置的第一业务成功率。在 VPA 中，仅根据云实例的资源占用率来检测云实例是否需要进行扩容，无法深入了解业务需求，需要多次检测才能精准确定云实例是否需要扩容，耗费了较长的时间在检测上。前述实现方式可基于云实例的状态信息来检测云实例是否需要扩容，由于云实例的状态信息包含云实例的资源占用率、负载程度以及业务成功率等信息，故云实例的状态信息与云实例的业务逻辑相关，更能体现业务需求，故工作节点基于云实例的状态信息，可实时感应业务需求，并基于业务需求准确检测云实例是否需要扩容，可有效减少检测的次数，从而缩短检测时长。

在一种可能的实现方式中，将待扩容的云实例迁移至第三工作节点的方式为冷迁移或热迁移。

在一种可能的实现方式中，云服务系统还包含管理节点，第一工作节点增大待扩容的云实例的资源配额之后，该方法还包括：第一工作节点将待扩容的云实例的资源配额发送至管理节点。前述实现方式中，第一工作节点对待扩容的云实例的资源配额进行增大后，可将增大后的该云实例的资源配额发送至管理节点，故管理节点和第一工作节点可同步该云实例的资源配额，使得全局（即整个云服务系统）资源配置信息准确一致。

本申请实施例的第二方面提供了一种云实例的缩容方法，该方法应用于云服务系统，云服务系统包含第一工作节点，该方法包括：

第一工作节点部署有多个云实例，例如，多个云实例可以为多个容器，又如，多个云实例可以为多组容器，再如，多个云实例可以为多个虚拟机，还如，多个云实例可以为多组虚拟机等等。第一工作节点可获取这多个云实例的状态信息，每个云实例的状态信息可用于指示该云实例所运行的业务的状态。

得到多个云实例的状态信息后，第一工作节点可基于这多个云实例的状态信息，对这多个云实例逐个进行分析，从而在这多个云实例中确定待缩容的云实例。

确定待缩容的云实例后，第一工作节点可确定待缩容的云实例中的非空闲资源和空闲资源，并释放待缩容的云实例的空闲资源，并计算这部分被释放的资源大小，即确定被释放的

待缩容的云实例的空闲资源量。那么，第一工作节点可基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额，从而实现云实例的缩容。

从上述方法可以看出：在获取多个云实例的状态信息后，第一工作节点可基于这些状态信息，在多个云实例中确定待缩容的云实例。然后，第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量。最后，第一工作节点可基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例缩容机制，第一工作节点自行确定待缩容的云实例后，可通过实时修改 cgroup 配置，以减小待缩容的云实例的资源配额，这对于待缩容的云实例运行的业务而言，是不感知的，故不会导致待缩容的云实例运行的业务中断。

在一种可能的实现方式中，第一工作节点基于状态信息，在多个云实例中确定待缩容的云实例包括：在多个云实例中，第一工作节点将状态信息满足预置的缩容条件的云实例确定为待缩容的云实例。前述实现方式中，对于多个云实例中的任意一个云实例，第一工作节点可检测该云实例的状态信息是否满足预置的缩容条件，若满足，则确定需要进行缩容，即将该云实例确定为待缩容的云实例，若不满足，则确定该云实例不需要进行缩容，结束操作。

在一种可能的实现方式中，状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。可见，云实例的状态信息与云实例的业务逻辑相关。

在一种可能的实现方式中，预置的缩容条件包括以下至少一种：资源占用率小于预置的第二资源占用率、负载程度小于预置的第二负载程度和业务成功率大于或等于预置的第二业务成功率。在 VPA 中，仅根据云实例的资源占用率来检测云实例是否需要进行缩容，无法深入了解业务需求，需要多次检测才能精准确定云实例是否需要缩容，耗费了较长的时间在检测上。前述实现方式可基于云实例的状态信息来检测云实例是否需要缩容，由于云实例的状态信息包含云实例的资源占用率、负载程度以及业务成功率等信息，故云实例的状态信息与云实例的业务逻辑相关，更能体现业务需求，故工作节点基于云实例的状态信息，可实时感应业务需求，并基于业务需求准确检测云实例是否需要缩容，可有效减少检测的次数，从而缩短检测时长。

在一种可能的实现方式中，云服务系统还包含管理节点，第一工作节点基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额之后，该方法还包括：第一工作节点将待缩容的云实例的资源配额发送至管理节点。前述实现方式中，第一工作节点对该云实例的资源配额进行减小后，可将减小后的该云实例的资源配额发送至管理节点，故管理节点和第一工作节点可同步该云实例的资源配额，使得全局（即整个云服务系统）资源配置信息准确一致。

本申请实施例的第三方面提供了一种工作节点，工作节点作为第一工作节点，第一工作节点设置于云服务系统中，第一工作节点部署有多个云实例，第一工作节点包括：获取模块，用于获取多个云实例的状态信息；第一确定模块，用于基于状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量；第一调整模块，用于若第一工作节点的空闲资源量大于或等于扩容所需的资源量，基于扩容所需的资源量，增大待扩容的云实例的资源配额。

从上述工作节点可以看出：第一工作节点在获取多个云实例的状态信息后，可基于这部分状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量。若第一工作节点

的空闲资源量大于或等于扩容所需的资源量，第一工作节点可基于扩容所需的资源量，增大待扩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例扩容机制，第一工作节点自行确定待扩容的云实例后，可通过实时修改 cgroup 配置，以增大待扩容的云实例的资源配额，这对于待扩容的云实例运行的业务而言，是不感知的，故不会导致待扩容的云实例运行的业务中断。

在一种可能的实现方式中，第一确定模块，用于：在多个云实例中，第一工作节点将状态信息满足预置的扩容条件的云实例确定为待扩容的云实例；基于待扩容的云实例的状态信息，确定扩容所需的扩容所需的资源量。

在一种可能的实现方式中，云服务系统还包含第二工作节点，第一工作节点还包括：第二确定模块，用于若第一工作节点的空闲资源量小于扩容所需的资源量，在多个云实例中确定待迁移的云实例，待迁移的云实例运行的业务的优先级低于预置的优先级；第一迁移模块，用于将待迁移的云实例迁移至第二工作节点，以更新第一工作节点的空闲资源量；第二调整模块，用于若更新后的第一工作节点的空闲资源量大于或等于扩容所需的资源量，基于扩容所需的资源量，增大待扩容的云实例的资源配额。

在一种可能的实现方式中，云服务系统还包含第三工作节点，第一工作节点还包括：检测模块，用于若更新后的第一工作节点的空闲资源量小于扩容所需的资源量，第一工作节点检测待扩容的云实例运行的业务的类型；创建模块，用于若待扩容的云实例运行的业务为无状态应用，在第三工作节点处创建新的云实例，新的云实例和待扩容的云实例共同用于运行无状态应用；第二迁移模块，用于若待扩容的云实例运行的业务为有状态应用，将待扩容的云实例迁移至第三工作节点。

在一种可能的实现方式中，状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

在一种可能的实现方式中，预置的扩容条件包括以下至少一种：资源占用率大于或等于预置的第一资源占用率、负载程度大于或等于预置的第一负载程度和业务成功率小于预置的第一业务成功率。

在一种可能的实现方式中，前述的迁移为冷迁移或热迁移。

在一种可能的实现方式中，云服务系统还包含管理节点，第一工作节点还包括：反馈模块，用于将待扩容的云实例的资源配额发送至管理节点。

本申请实施例的第四方面提供了一种工作节点，工作节点作为第一工作节点，第一工作节点设置于云服务系统中，第一工作节点部署有多个云实例，第一工作节点包括：获取模块，用于获取多个云实例的状态信息；确定模块，用于基于状态信息，在多个云实例中确定待缩容的云实例；释放模块，用于释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量；调整模块，用于基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。

从上述工作节点可以看出：在获取多个云实例的状态信息后，第一工作节点可基于这些状态信息，在多个云实例中确定待缩容的云实例。然后，第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量。最后，第一工作节点可基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。基于前述过程可知，本申请

提供了一种新的云实例缩容机制，第一工作节点自行确定待缩容的云实例后，可通过实时修改 cgroup 配置，以减小待缩容的云实例的资源配额，这对于待缩容的云实例运行的业务而言，是不感知的，故不会导致待缩容的云实例运行的业务中断。

在一种可能的实现方式中，确定模块，用于在多个云实例中，将状态信息满足预置的缩容条件的云实例确定为待缩容的云实例。

在一种可能的实现方式中，状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

在一种可能的实现方式中，预置的缩容条件包括以下至少一种：资源占用率小于预置的第二资源占用率、负载程度小于预置的第二负载程度和业务成功率大于或等于预置的第二业务成功率。

在一种可能的实现方式中，云服务系统还包含管理节点，第一工作节点还包括：反馈模块，用于将待扩容的云实例的资源配额发送至管理节点。

本申请实施例的第五方面提供了一种工作节点，该工作节点包括存储器和处理器；存储器存储有代码，处理器被配置为执行代码，当代码被执行时，工作节点执行如第一方面、第一方面的任意一种可能的实现方式、第二方面或第二方面的任意一种可能的实现方式所述的方法。

本申请实施例的第六方面提供了一种计算机存储介质，计算机存储介质存储有一个或多个指令，指令在由一个或多个计算机执行时使得一个或多个计算机实施如第一方面、第一方面的任意一种可能的实现方式、第二方面或第二方面的任意一种可能的实现方式所述的方法。

本申请实施例的第七方面提供了一种计算机程序产品，计算机程序产品存储有指令，指令在由计算机执行时，使得计算机实施如第一方面、第一方面的任意一种可能的实现方式、第二方面或第二方面的任意一种可能的实现方式所述的方法。

本申请实施例中，在获取多个云实例的状态信息后，第一工作节点可基于这些状态信息，在多个云实例中确定待缩容的云实例。然后，第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量。最后，第一工作节点可基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例缩容机制，第一工作节点自行确定待缩容的云实例后，可通过实时修改 cgroup 配置，以减小待缩容的云实例的资源配额，这对于待缩容的云实例运行的业务而言，是不感知的，故不会导致待缩容的云实例运行的业务中断。

此外，本申请实施例中，在获取多个云实例的状态信息后，第一工作节点可基于这些状态信息，在多个云实例中确定待缩容的云实例。然后，第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量。最后，第一工作节点可基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例缩容机制，第一工作节点自行确定待缩容的云实例后，可通过实时修改 cgroup 配置，以减小待缩容的云实例的资源配额，这对于待缩容的云实例运行的业务而言，是不感知的，故不会导致待缩容的云实例运行的业务中断。

附图说明

图 1 为 VPA 的一个示意图；

图 2 为本申请实施例提供的云服务系统的一个结构示意图；

图 3 为本申请实施例提供的云服务系统的另一结构示意图；

图 4 为本申请实施例提供的云实例的扩容方法的一个流程示意图；

图 5 为本申请实施例提供的云实例的缩容方法的一个流程示意图；

图 6 为本申请实施例提供的工作节点的一个结构示意图；

图 7 为本申请实施例提供的工作节点的另一个结构示意图；

图 8 为本申请实施例提供的工作节点的另一个结构示意图。

具体实施方式

本申请实施例提供了一种云实例的扩缩容方法及其相关设备，可在增大或减小云实例的资源配额时，确保云实例运行的业务不会中断。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换，这仅仅是描述本申请的实施例中对相同属性的对象在描述时所采用的区分方式。此外，术语“包括”和“具有”并他们的任何变形，意图在于覆盖不排他的包含，以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它单元。

随着技术的飞速发展，云服务系统的规模越来越大。云服务系统通常包含多个工作节点和管理节点，其中，每个工作节点上部署有多个云实例，管理节点可对所有云实例进行集中管理。

为了便于说明，以云实例为容器（docker）进行介绍。目前，云服务系统以 kubernetes 作为容器的管理标准，可对容器实现编排部署、灰度升降级、自动扩缩容等功能。在自动伸缩功能中，kubernetes 可支持两种自动两种自动伸缩方法，分别为 VPA 和 HPA。如图 1 所示（图 1 为 VPA 的一个示意图），在 VPA 中，管理节点可基于某个工作节点中 pod 的资源占用率计算出 pod 的资源配额推荐值，并发送至该工作节点。当该工作节点创建 pod 时，可基于该推荐值为 pod 设置新的资源配额，例如，增大 pod 的资源配额（也就是增大分配给 pod 的资源量，即对 pod 进行扩容）或减小 pod 的资源配额（也就是减小分配 pod 的资源量，即对 pod 进行缩容）。

前述过程中，由于修改 pod 的资源配额的时间点，只能在创建 pod 的时候。当需要修改某个 pod 的资源配额时，工作节点只能先释放该 pod，并在重新创建该 pod 时才能实现资源配额的修改，这样会导致该 pod 运行的业务中断。

为了解决上述问题，本申请实施例提供了一种云实例的扩缩容方法，该方法可应用于如图 2 所示的云服务系统中（图 2 为本申请实施例提供的云服务系统的一个结构示意图），该系统所适用的云场景包含公有云、私有云、混合云等场景，该系统包含管理节点和多个工作节点，管理节点用于集中管理这多个工作节点，下文将分别对管理节点和工作节点进行介绍：

管理节点，通常为单独的一台物理服务器（也可以称为网络设备）。管理节点内设置有多

个功能模块，分别为应用程序接口（application programming interface，API）、调度器（scheduler）、主控器（controller-manager）以及数据库（data base，DB）。其中，API可理解为云服务系统的接口，用户可通过 API 在工作节点上创建云实例。调度器可用于调度云实例到合适的节点上，调度器往往是可替换的组件，其形式可根据不同厂商的要求自行进行设置，此处不做限制。主控制器用于实现云服务系统中各个工作节点以及各个云实例的资源管理等功能。数据库用于存储云服务系统中的配置信息，例如，各个云实例的资源配额等等。

工作节点，通常也为单独的一台物理服务器。基于虚拟化技术，工作节点的操作系统（operating system）之上可部署有节点代理（kubelet）以及多个云实例。其中，操作系统用于实现工作节点的硬件资源管理。节点代理作为工作节点上的代理进程，用于管理工作节点上的所有云实例，包括生命周期的管理，资源配置等等。云实例可通过多种形式呈现，例如，一个云实例可以为一个虚拟机（virtual machine，VM），又如，一个云实例可以为一个容器（docker）。再如，一个云实例可以为一组虚拟机。还如，一个云实例可以为一组容器（也可以称为一个 pod）等等。可以理解的是，工作节点自身包含一定的硬件资源（计算资源、存储资源和网络资源等等），节点代理可为每个云实例设置一个资源配额，每个云实例的资源配额即为分配至该云实例的资源量（包含计算资源量、存储资源量和网络资源量等等），且节点代理可管理和调整所有云实例的资源配额。

需要说明的是，如图 3 所示（图 3 为本申请实施例提供的云服务系统的另一结构示意图），对于某个工作节点而言，该工作节点的节点代理设置有伸缩接口（scaleAPI）供该工作节点的云实例调用，且在该工作节点的节点代理和该工作节点的云实例之间构建了双向通道，故二者可实现相互通信。具体地，该工作节点的云实例可自行根据自身的状态信息判断是否需要进行扩缩容，在确定需要进行扩缩容后，向该工作节点的节点代理发起自动伸缩请求，以使得该工作节点的节点代理基于该请求对该工作节点的云实例进行扩缩容。

值得注意的是，在节点代理中增加伸缩接口及实现，这部分需要侵入式修改节点代理功能，或者开发成插件。节点代理与云实例之间的双向通道，无需侵入式修改，只需配置相关网络通路即可。

为了进一步理解前述扩缩容的过程，下文将分别从云实例的扩容和云实例的缩容这两个方面对该过程做进一步的介绍。首先对云实例的扩容过程进行说明，图 4 为本申请实施例提供的云实例的扩容方法的一个流程示意图，需要说明的是，该方法可应用于前述图 2 或图 3 所示的云服务系统，该方法的执行主体可为云服务系统中多个工作节点的任意一个工节点，下文将该工作节点称为第一工作节点。如图 4 所示，该方法包括：

401、第一工作节点获取多个云实例的状态信息。

402、第一工作节点基于状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量。

本实施例中，第一工作节点上部署有多个云实例，对于任意一个云实例而言，该云实例可周期性地获取自身的状态信息，以基于自身的状态信息判定是否需要进行扩容。其中，该周期通常是秒级的时间段。

进一步地，该云实例的状态信息可包含以下至少一种：该云实例的资源占用率、该云实

例的负载程度和该云实例的业务成功率。

更进一步地，该云实例的资源占用率可包含以下至少一种：该云实例的中央处理器（central processing unit，CPU）使用率、该云实例的内存使用量、该云实例的存储每秒进行读写操作的次数（input/output operations per second，IOPS）以及该云实例的网络IOPS等等。

更进一步地，该云实例的负载程度可包含以下至少一种：该云实例的任务响应时间、该云实例的任务处理时延、该云实例的数据库指标、该云实例的消息队列指标以及该云实例的任务队列长度等等。

更进一步地，该云实例的业务成功率可包含以下至少一种：该云实例的任务完成率以及该云实例的消息传输成功率等等。

具体地，该云实例可通过以下方式确定是否需要进行扩容：

该云实例可检测自身的状态信息是否满足预置的扩容条件，若满足，则确定需要进行扩容，即将自身确定为待扩容的云实例，若不满足，则确定不需要进行扩容，结束操作。

其中，该云实例的状态信息满足预置的扩容条件可包含以下至少一种情况：该云实例的资源占用率大于或等于预置的第一资源占用率、该云实例的负载程度大于或等于预置的第一负载程度以及该云实例的业务成功率小于预置的第一业务成功率。需要说明的是，预置的第一资源占用率可理解为达到扩容需求的资源占用率阈值，预置的第一负载程度可理解为达到扩容需求的负载程度阈值，预置的第一业务成功率可理解为达到扩容需求的业务成功率阈值，这三个阈值的大小可根据实际需求进行设置，此处不做限制。

例如，设该云实例的状态信息为该云实例的任务响应时间，相应的，预置的第一负载程度则为预置的响应时间阈值。并且，该云实例的任务响应时间为 3S，预置的响应时间阈值为1S，由此可见，该云实例的任务相应时间大于预置的响应时间阈值，故该云实例可确定需要进行扩容。又如，设该云实例的状态信息为该云实例的内存使用量，相应的，预置的第一资源占用率则为预置的内存使用量阈值。并且，该云实例的内存使用量为 8G，预置的内存使用量阈值为 1G，由此可见，该云实例的内存使用量大于预置的内存使用量阈值，故该云实例可确定需要进行扩容等等。

若该云实例确定需要进行扩容，该云实例可基于自身的状态信息，来确定自身扩容所需的资源量。例如，该云实例可基于该云实例的任务队列长度，来确定自身扩容所需的计算资源量、存储资源量和网络资源量等等。

403、第一工作节点检测第一工作节点的空闲资源量是否大于或等于扩容所需的资源量。

该云实例确定需要进行扩容以及自身扩容所需的资源量后，可向第一工作节点的节点代理发起扩容请求。节点代理接收到来自该云实例的扩容请求后，可解析该请求，从而确定该云实例为待扩容的云实例以及该云实例扩容所需的资源量。

那么，节点代理可检测第一工作节点的空闲资源量（即第一工作节点本地未被使用的资源量）是否大于或等于该云实例扩容所需的资源量，若第一工作节点的空闲资源量大于或等于该云实例扩容所需的资源量，则执行步骤 404，若第一工作节点的空闲资源量小于该云实例扩容所需的资源量，则执行步骤 405。

404、若第一工作节点的空闲资源量大于或等于扩容所需的资源量，第一工作节点基于扩

容所需的资源量，增大待扩容的云实例的资源配额。

若第一工作节点的空闲资源量大于或等于该云实例扩容所需的资源量，说明第一工作节点的空闲资源是充足的，可利用这部分资源直接对该云实例进行扩容，故节点代理可基于该云实例扩容所需的资源量，增大该云实例的资源配额。例如，设该云实例运行的业务量较大，致使该云实例扩容所需的内存量为 5G，原先该云实例的内存配额为 1G（即原先分配至该云实例的内存量），节点代理可通过修改 cgroup 配置，以使得修改后的该云实例的内存配额为 6G。

405、若第一工作节点的空闲资源量小于扩容所需的资源量，第一工作节点在多个云实例中确定待迁移的云实例，待迁移的云实例运行的业务的优先级低于预置的优先级。

406、第一工作节点将待迁移的云实例迁移至第二工作节点，以更新第一工作节点的空闲资源量。

若第一工作节点的空闲资源量小于该云实例扩容所需的资源量，说明第一工作节点的空闲资源是不足的，节点代理可尝试解决资源不足的问题，即节点代理可在第一工作节点的多个云实例中，确定至少一个待迁移的云实例，这些待迁移的云实例运行的业务的优先级低于预置的优先级（即这些云实例所运行的业务往往优先级较低），故可将这些待迁移的云实例迁移至第二工作节点，那么，第一工作节点中被分配至这些待迁移的云实例的资源被释放，成为新的空闲资源，从而更新了第一工作节点的空闲资源量（即增大了第一工作节点的空闲资源量）。

需要说明的是，由于第一工作节点中待迁移的云实例所运行的业务往往优先级较低，故待迁移的云实例的迁移方式可以为冷迁移，即第一工作节点的节点代理向管理节点发调度请求，管理节点基于该调度请求，可在除第一工作节点之外的其余工作节点中，挑选一个工作节点作为迁移的目的地，即第二工作节点。然后，管理节点可通知第二工作节点的节点代理创建新的云实例，并控制新的云实例重新运行第一工作节点中待迁移的云实例所运行的业务。最后，管理节点可通知第一工作节点的节点代理释放待迁移的云实例，以使得分配至待迁移的云实例的资源被释放，从而更新第一工作节点的空闲资源量。

407、第一工作节点检测更新后的第一工作节点的空闲资源量是否大于或等于扩容所需的资源量。

更新了第一工作节点的空闲资源量后，节点代理可检测更新后的第一工作节点的空闲资源量是否大于或等于扩容所需的资源量，若更新后的第一工作节点的空闲资源量大于或等于该云实例扩容所需的资源量，则执行步骤 408，若更新后的第一工作节点的空闲资源量小于该云实例扩容所需的资源量，则执行步骤 410。

408、若更新后的第一工作节点的空闲资源量大于或等于扩容所需的资源量，第一工作节点基于扩容所需的资源量，增大待扩容的云实例的资源配额。

若更新后的第一工作节点的空闲资源量大于或等于该云实例扩容所需的资源量，说明更新后的第一工作节点的空闲资源是充足的，可利用这部分资源直接对该云实例进行扩容，故节点代理可基于该云实例扩容所需的资源量，增大该云实例的资源配额。

409、第一工作节点将待扩容的云实例的资源配额发送至管理节点。

节点代理对该云实例的资源配额进行增大后，可将增大后的该云实例的资源配额发送至管理节点，故管理节点和第一工作节点可同步该云实例的资源配额，使得全局（即整个云服

务系统）资源配置信息准确一致。

410、若更新后的第一工作节点的空闲资源量小于扩容所需的资源量，第一工作节点检测待扩容的云实例运行的业务的类型。

411、若待扩容的云实例运行的业务为无状态应用，第一工作节点在第三工作节点处创建新的云实例，新的云实例和待扩容的云实例共同用于运行无状态应用。

412、若待扩容的云实例运行的业务为有状态应用，第一工作节点将待扩容的云实例迁移至第三工作节点。

若更新后的第一工作节点的空闲资源量小于该云实例扩容所需的资源量，说明更新后的第一工作节点的空闲资源是不足的，节点代理则需要再次解决资源不足的问题，节点代理可先检测该云实例运行的业务（也可以理解为该云实例运行的应用）的类型，以基于该云实例运行的业务类型进行相应的处理：

若该云实例运行的业务为无状态应用，节点代理可申请在第三工作节点处创建新的云实例，那么，该云实例和新的云实例可共同运行该云实例原先所运行的业务，也就相当于实现了扩容。具体地，第一工作节点的节点代理确定该云实例运行的业务为无状态应用，可向管理节点向管理节点发调度请求，管理节点基于该调度请求，可在除第一工作节点之外的其余工作节点中，挑选一个工作节点，即第三工作节点。然后，管理节点可通知第三工作节点的节点代理创建新的云实例，并控制新的云实例用于运行第一工作节点中该云实例所运行的业务，由于该业务为无状态应用，对于不同的云实例而言，哪个云实例来运行的效果是相似的，故新的云实例和该云实例虽然位于不同的同坐节点上，但可共同承担该业务，相当于完成了扩容。

若该云实例运行的业务为有状态应用，节点代理可将该云实例迁移至在第三工作节点，其中，第三工作节点的空闲资源量大于或等于该云实例扩容所需的资源量，那么，在将该云实例迁移至第三工作节点后，可增大该云实例的资源配额，也就相当于完成了扩容。具体地，该云实例迁移至第三工作节点的方式既可以是热迁移，也可以是冷迁移。下文将分别对两种方式进行介绍：（1）热迁移的方式为：第一工作节点的节点代理确定该云实例运行的业务为有状态应用，可向管理节点向管理节点发调度请求，管理节点基于该调度请求，可在除第一工作节点之外的其余工作节点中，挑选一个工作节点作为迁移的目的地，即第三工作节点。然后，管理节点可通知第三工作节点的节点代理创建新的云实例，并保持新的云实例的业务状态和该云实例的业务状态一致，以控制新的云实例继续运行第一工作节点中该云实例所运行的业务。最后，管理节点可通知第一工作节点的节点代理释放该云实例，由于第三工作节点的节点代理可令新的云实例的资源配额大于第一工作节点中该云实例的资源配额，故相当于完成了扩容。（2）冷迁移的方式为：第一工作节点的节点代理确定该云实例运行的业务为有状态应用，可向管理节点向管理节点发调度请求，管理节点基于该调度请求，可在除第一工作节点之外的其余工作节点中，挑选一个工作节点作为迁移的目的地，即第三工作节点。然后，管理节点可通知第三工作节点的节点代理创建新的云实例，并控制新的云实例重新运行第一工作节点中该云实例所运行的业务。最后，管理节点可通知第一工作节点的节点代理释放该云实例，由于第三工作节点的节点代理可令新的云实例的资源配额大于第一工作节点中该云实例的资源配额，故相当于完成了扩容。

应理解，本实施例仅以第一工作节点的其中一个云实例进行示意性说明，第一工作节点的其余云实例也可执行如同该云实例所执行的操作，即对于第一工作节点的每一个云实例，均可执行如步骤 401 至步骤 412 所述的操作，此处不再赘述。

本申请实施例中，第一工作节点在获取多个云实例的状态信息后，可基于这部分状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量。若第一工作节点的空闲资源量大于或等于扩容所需的资源量，第一工作节点可基于扩容所需的资源量，增大待扩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例扩容机制，第一工作节点自行确定待扩容的云实例后，可通过实时修改 cgroup 配置，以增大待扩容的云实例的资源配额，这对于待扩容的云实例运行的业务而言，是不感知的，故不会导致待扩容的云实例运行的业务中断。

进一步地，在 VPA 中，若需要增大某个云实例的资源配额，工作节点只能在云实例重建的时候才能修改云实例的资源配额，但是云实例重建往往需要等待特定的时机，例如，被驱逐、被杀掉等等。这部分时间通常不可控，导致增大云实例的资源配额所需的时长过大，而本申请实施例不需要在云实例重建的时候去修改云实例的资源配额，可实时修改云实例的资源配额，从而有效缩短增大云实例的资源配额所需的时长。

更进一步地，在 VPA 中，若需要增大某个云实例的资源配额，工作节点只能在云实例重建的时候才能修改云实例的资源配额，但是云实例重建和业务启动往往需要较长的时间，导致增大云实例的资源配额所需的时长过大，而本申请实施例不需要在云实例重建的时候去修改云实例的资源配额，可实时修改云实例的资源配额，从而有效缩短增大云实例的资源配额所需的时长。

更进一步地，在 VPA 中，仅根据云实例的资源占用率来检测云实例是否需要进行扩容，无法深入了解业务需求，需要多次检测才能精准确定云实例是否需要扩容，耗费了较长的时间在检测上。本申请实施例可基于云实例的状态信息来检测云实例是否需要扩容，由于云实例的状态信息包含云实例的资源占用率、负载程度以及业务成功率等信息，故云实例的状态信息与云实例的业务逻辑相关，更能体现业务需求，故工作节点基于云实例的状态信息，可实时感应业务需求，并基于业务需求准确检测云实例是否需要扩容，可有效减少检测的次数，从而缩短检测时长。

以上对云实例的扩容过程所进行的详细说明，以下将对云实例的缩容过程进行介绍。图 5 为本申请实施例提供的云实例的缩容方法的一个流程示意图，需要说明的是，该方法可应用于前述图 2 或图 3 所示的云服务系统，该方法的执行主体可为云服务系统中多个工作节点的任意一个工节点，下文将该工作节点称为第一工作节点。如图 5 所示，该方法包括：

501、第一工作节点获取多个云实例的状态信息。

502、第一工作节点基于状态信息，在多个云实例中确定待缩容的云实例。

本实施例中，第一工作节点上部署有多个云实例，对于任意一个云实例而言，该云实例可周期性地获取自身的状态信息，以基于自身的状态信息判定是否需要进行缩容。

进一步地，该云实例的状态信息可包含以下至少一种：该云实例的资源占用率、该云实例的负载程度和该云实例的业务成功率。

更进一步地，该云实例的资源占用率可包含以下至少一种：该云实例的 CPU 使用率、该

云实例的内存使用量、该云实例的存储 IOPS 以及该云实例的网络 IOPS 等等。

更进一步地，该云实例的负载程度可包含以下至少一种：该云实例的任务响应时间、该云实例的任务处理时延、该云实例的数据库指标、该云实例的消息队列指标以及该云实例的任务队列长度等等。

更进一步地，该云实例的业务成功率可包含以下至少一种：该云实例的任务完成率以及该云实例的消息传输成功率等等。

具体地，该云实例可通过以下方式确定是否需要进行缩容：

该云实例可检测自身的状态信息是否满足预置的缩容条件，若满足，则确定需要进行缩容，即将自身确定为待缩容的云实例，若不满足，则确定不需要进行缩容，结束操作。

其中，该云实例的状态信息满足预置的缩容条件可包含以下至少一种情况：该云实例的资源占用率小于预置的第二资源占用率、该云实例的负载程度小于预置的第二负载程度以及该云实例的业务成功率大于或等于预置的第二业务成功率。需要说明的是，预置的第二资源占用率可理解为达到缩容需求的资源占用率阈值，预置的第二负载程度可理解为达到缩容需求的负载程度阈值，预置的第二业务成功率可理解为达到缩容需求的业务成功率阈值，这三个阈值的大小可根据实际需求进行设置，此处不做限制。

503、第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量。

该云实例确定需要进行缩容后，可将自身的空闲资源（即该云实例未使用的资源，该云实例运行业务所占用的资源则为该云实例的非空闲资源）释放，并在完成释放后，计算自身的空闲资源量（即该云实例所释放的资源的大小）。

需要说明的是，当云实例确定自身运行的业务下降需要进行缩容时，云实例可以尝试如下方式进行缩容：（1）对于系统级语言，内存由业务自行管理，大部分内存通过动态分配与释放能及时回收。还有一部分以内存池的方式进行管理，可以在业务量小的时候缩小内存池，业务量大的时候扩大内存池；（2）对于带有垃圾回收的高级语言，内存回收由高级语言的运行时进行内存回收，但是高级语言的内存回收和业务量通常不同步，这时可以由业务主动调用强制垃圾回收来释放内存。

该云实例得到自身的空闲资源量后，则将自身的空闲资源量发送至第一工作节点的节点代理。

504、第一工作节点基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。

节点代理确定该云实例的空闲资源量后，可基于该云实例的空闲资源量，减小该云实例的资源配额。例如，设该云实例运行的业务量较小，致使该云实例的空闲内存量为 2G，原先该云实例的内存配额为 5G（即原先分配至该云实例的内存量），节点代理可通过修改 cgroup 配置，以使得修改后的该云实例的内存配额为 3G。

505、第一工作节点将待缩容的云实例的资源配额发送至管理节点。

节点代理对该云实例的资源配额进行减小后，可将减小后的该云实例的资源配额发送至管理节点，故管理节点和第一工作节点可同步该云实例的资源配额，使得全局（即整个云服务系统）资源配置信息准确一致。

应理解，本实施例仅以第一工作节点的其中一个云实例进行示意性说明，第一工作节点

的其余云实例也可执行如同该云实例所执行的操作，即对于第一工作节点的每一个云实例，均可执行如步骤 501 至步骤 502 所述的操作，此处不再赘述。

本申请实施例中，在获取多个云实例的状态信息后，第一工作节点可基于这些状态信息，在多个云实例中确定待缩容的云实例。然后，第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量。最后，第一工作节点可基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例缩容机制，第一工作节点自行确定待缩容的云实例后，可通过实时修改 cgroup 配置，以减小待缩容的云实例的资源配额，这对于待缩容的云实例运行的业务而言，是不感知的，故不会导致待缩容的云实例运行的业务中断。

进一步地，在减小待缩容的云实例的资源配额之前，可令待缩容的云实例自行释放空闲资源，从而确保对待缩容的云实例的资源配额进行修改时的成功率。

更进一步地，在 VPA 中，若需要减小某个云实例的资源配额，工作节点只能在云实例重建的时候才能修改云实例的资源配额，但是云实例重建往往需要等待特定的时机，例如，被驱逐、被杀掉等等。这部分时间通常不可控，导致减小云实例的资源配额所需的时长过大，而本申请实施例不需要在云实例重建的时候去修改云实例的资源配额，可实时修改云实例的资源配额，从而有效缩短减小云实例的资源配额所需的时长。

更进一步地，在 VPA 中，若需要减小某个云实例的资源配额，工作节点只能在云实例重建的时候才能修改云实例的资源配额，但是云实例重建和业务启动往往需要较长的时间，导致减小云实例的资源配额所需的时长过大，而本申请实施例不需要在云实例重建的时候去修改云实例的资源配额，可实时修改云实例的资源配额，从而有效缩短减小云实例的资源配额所需的时长。

更进一步地，在 VPA 中，仅根据云实例的资源占用率来检测云实例是否需要进行缩容，无法深入了解业务需求，需要多次检测才能精准确定云实例是否需要缩容，耗费了较长的时间在检测上。本申请实施例可基于云实例的状态信息来检测云实例是否需要缩容，由于云实例的状态信息包含云实例的资源占用率、负载程度以及业务成功率等信息，故云实例的状态信息与云实例的业务逻辑相关，更能体现业务需求，故工作节点基于云实例的状态信息，可实时感应业务需求，并基于业务需求准确检测云实例是否需要缩容，可有效减少检测的次数，从而缩短检测时长。

以上是对本申请实施例提供的云实例的缩容方法所进行的详细说明，以下将对本申请实施例提供的工作节点进行介绍。图 6 为本申请实施例提供的工作节点的一个结构示意图，如图 6 所示，该工作节点作为第一工作节点，第一工作节点设置于云服务系统中，第一工作节点部署有多个云实例，第一工作节点包括：

获取模块 601，用于获取多个云实例的状态信息；

第一确定模块 602，用于基于状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量；

第一调整模块 603，用于若第一工作节点的空闲资源量大于或等于扩容所需的资源量，基于扩容所需的资源量，增大待扩容的云实例的资源配额。

本申请实施例中，第一工作节点在获取多个云实例的状态信息后，可基于这部分状态信

息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量。若第一工作节点的空闲资源量大于或等于扩容所需的资源量，第一工作节点可基于扩容所需的资源量，增大待扩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例扩容机制，第一工作节点自行确定待扩容的云实例后，可通过实时修改 cgroup 配置，以增大待扩容的云实例的资源配额，这对于待扩容的云实例运行的业务而言，是不感知的，故不会导致待扩容的云实例运行的业务中断。

在一种可能的实现方式中，第一确定模块 602，用于：在多个云实例中，第一工作节点将状态信息满足预置的扩容条件的云实例确定为待扩容的云实例；基于待扩容的云实例的状态信息，确定扩容所需的扩容所需的资源量。

在一种可能的实现方式中，云服务系统还包含第二工作节点，第一工作节点还包括：第二确定模块，用于若第一工作节点的空闲资源量小于扩容所需的资源量，在多个云实例中确定待迁移的云实例，待迁移的云实例运行的业务的优先级低于预置的优先级；第一迁移模块，用于将待迁移的云实例迁移至第二工作节点，以更新第一工作节点的空闲资源量；第二调整模块，用于若更新后的第一工作节点的空闲资源量大于或等于扩容所需的资源量，基于扩容所需的资源量，增大待扩容的云实例的资源配额。

在一种可能的实现方式中，云服务系统还包含第三工作节点，第一工作节点还包括：检测模块，用于若更新后的第一工作节点的空闲资源量小于扩容所需的资源量，第一工作节点检测待扩容的云实例运行的业务的类型；创建模块，用于若待扩容的云实例运行的业务为无状态应用，在第三工作节点处创建新的云实例，新的云实例和待扩容的云实例共同用于运行无状态应用；第二迁移模块，用于若待扩容的云实例运行的业务为有状态应用，将待扩容的云实例迁移至第三工作节点。

在一种可能的实现方式中，状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

在一种可能的实现方式中，预置的扩容条件包括以下至少一种：资源占用率大于或等于预置的第一资源占用率、负载程度大于或等于预置的第一负载程度和业务成功率小于预置的第一业务成功率。

在一种可能的实现方式中，前述的迁移为冷迁移或热迁移。

在一种可能的实现方式中，云服务系统还包含管理节点，第一工作节点还包括：反馈模块，用于将待扩容的云实例的资源配额发送至管理节点。

图 7 为本申请实施例提供的工作节点的另一个结构示意图，如图 7 所示，该工作节点作为第一工作节点，第一工作节点设置于云服务系统中，第一工作节点部署有多个云实例，第一工作节点包括：

获取模块 701，用于获取多个云实例的状态信息；

确定模块 702，用于基于状态信息，在多个云实例中确定待缩容的云实例；

释放模块 703，用于释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量；

调整模块 704，用于基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。

本申请实施例中，在获取多个云实例的状态信息后，第一工作节点可基于这些状态信息，在多个云实例中确定待缩容的云实例。然后，第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量。最后，第一工作节点可基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额。基于前述过程可知，本申请提供了一种新的云实例缩容机制，第一工作节点自行确定待缩容的云实例后，可通过实时修改 cgroup 配置，以减小待缩容的云实例的资源配额，这对于待缩容的云实例运行的业务而言，是不感知的，故不会导致待缩容的云实例运行的业务中断。

在一种可能的实现方式中，确定模块 702，用于在多个云实例中，将状态信息满足预置的缩容条件的云实例确定为待缩容的云实例。

在一种可能的实现方式中，状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

在一种可能的实现方式中，预置的缩容条件包括以下至少一种：资源占用率小于预置的第二资源占用率、负载程度小于预置的第二负载程度和业务成功率大于或等于预置的第二业务成功率。

在一种可能的实现方式中，云服务系统还包含管理节点，第一工作节点还包括：反馈模块，用于将待扩容的云实例的资源配额发送至管理节点。

需要说明的是，上述装置各模块/单元之间的信息交互、执行过程等内容，由于与本申请方法实施例基于同一构思，其带来的技术效果与本申请方法实施例相同，具体内容可参考本申请实施例前述所示的方法实施例中的叙述，此处不再赘述。

图 8 为本申请实施例提供的工作节点的另一个结构示意图。如图 8 所示，该工作节点作为第一工作节点，第一工作节点设置于云服务系统中，第一工作节点部署有多个云实例，第一工作节点可以包括一个或一个以上中央处理器 801，存储器 802，输入输出接口 803，有线或无线网络接口 804，电源 805。

存储器 802 可以是短暂存储或持久存储。更进一步地，中央处理器 801 可以配置为与存储器 802 通信，在第一工作节点上执行存储器 802 中的一系列指令操作。

本实施例中，中央处理器 801 可以执行前述图 4 或图 5 所示实施例中第一工作节点所执行的操作，具体此处不再赘述。

本实施例中，中央处理器 801 中的具体功能模块划分可以与前述图 6 中所描述的获取模块、第一确定模块、第二调整模块、第二确定模块、第一迁移模块、第二调整模块、检测模块、创建模块、第二迁移模块和反馈模块等模块的划分方式类似，此处不再赘述；或，

中央处理器 801 中的具体功能模块划分可以与前述图 7 中所描述的获取模块、确定模块、释放模块、调整模块和反馈模块等模块的划分方式类似，此处不再赘述。

本申请实施例还涉及一种计算机存储介质，包括计算机可读指令，当所述计算机可读指令被执行时，实现如图 5 所示实施例中第一服务器所执行的步骤，或，实现如图 5 所示实施例中仲裁器所执行的步骤。

本申请实施例还涉及一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行如图 5 所示实施例中第一服务器所执行的步骤，或，实现如图 5 所示实施例中仲裁器所执行的步骤。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，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系统，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之间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，机械或其它的形式。

所述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或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。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。

另外，在本申请各个实施例中的各功能单元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。

所述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时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请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，该计算机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台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述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（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

权利要求

1. 一种云实例的扩容方法，其特征在于，所述方法应用于云服务系统，所述云服务系统包含第一工作节点，所述第一工作节点部署有多个云实例，所述方法包括：

所述第一工作节点获取所述多个云实例的状态信息；

所述第一工作节点基于所述状态信息，在所述多个云实例中确定待扩容的云实例以及扩容所需的资源量；

若所述第一工作节点的空闲资源量大于或等于所述扩容所需的资源量，所述第一工作节点基于所述扩容所需的资源量，增大所述待扩容的云实例的资源配额。

2. 根据权利要求 1 所述的方法，其特征在于，所述第一工作节点基于所述状态信息，在所述多个云实例中确定待扩容的第一云实例以及扩容所需的扩容所需的资源量包括：

在所述多个云实例中，所述第一工作节点将状态信息满足预置的扩容条件的云实例确定为待扩容的云实例；

所述第一工作节点基于所述待扩容的云实例的状态信息，确定扩容所需的扩容所需的资源量。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述云服务系统还包含第二工作节点，所述方法还包括：

若所述第一工作节点的空闲资源量小于所述扩容所需的资源量，所述第一工作节点在所述多个云实例中确定待迁移的云实例，所述待迁移的云实例运行的业务的优先级低于预置的优先级；

所述第一工作节点将所述待迁移的云实例迁移至第二工作节点，以更新所述第一工作节点的空闲资源量；

若更新后的第一工作节点的空闲资源量大于或等于所述扩容所需的资源量，所述第一工作节点基于所述扩容所需的资源量，增大所述待扩容的云实例的资源配额。

4. 根据权利要求 3 所述的方法，其特征在于，所述云服务系统还包含第三工作节点，所述方法还包括：

若更新后的第一工作节点的空闲资源量小于所述扩容所需的资源量，所述第一工作节点检测所述待扩容的云实例运行的业务的类型；

若所述待扩容的云实例运行的业务为无状态应用，所述第一工作节点在所述第三工作节点处创建新的云实例，所述新的云实例和所述待扩容的云实例共同用于运行所述无状态应用；

若所述待扩容的云实例运行的业务为有状态应用，所述第一工作节点将所述待扩容的云实例迁移至第三工作节点。

5. 根据权利要求 2 所述的方法，其特征在于，所述状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

6. 根据权利要求 5 所述的方法，其特征在于，所述预置的扩容条件包括以下至少一种：资源占用率大于或等于预置的第一资源占用率、负载程度大于或等于预置的第一负载程度和业务成功率小于预置的第一业务成功率。

7. 根据权利要求 4 所述的方法，其特征在于，所述迁移为冷迁移或热迁移。

8. 根据权利要求 1 至 7 任意一项所述的方法，其特征在于，所述云服务系统还包含管理

节点，所述第一工作节点增大所述待扩容的云实例的资源配额之后，所述方法还包括：

所述第一工作节点将所述待扩容的云实例的资源配额发送至所述管理节点。

9. 一种云实例的缩容方法，其特征在于，所述方法应用于云服务系统，所述云服务系统包含第一工作节点，所述第一工作节点部署有多个云实例，所述方法包括：

所述第一工作节点获取所述多个云实例的状态信息；

所述第一工作节点基于所述状态信息，在所述多个云实例中确定待缩容的云实例；

所述第一工作节点释放所述待缩容的云实例的空闲资源，并确定被释放的所述待缩容的云实例的空闲资源量；

所述第一工作节点基于所述待缩容的云实例的空闲资源量，减小所述待缩容的云实例的资源配额。

10. 根据权利要求 9 所述的方法，其特征在于，所述第一工作节点基于所述状态信息，在所述多个云实例中确定待缩容的云实例包括：

在所述多个云实例中，所述第一工作节点将状态信息满足预置的缩容条件的云实例确定为待缩容的云实例。

11. 根据权利要求 9 或 10 所述的方法，其特征在于，所述状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

12. 根据权利要求 11 所述的方法，其特征在于，所述预置的缩容条件包括以下至少一种：资源占用率小于预置的第二资源占用率、负载程度小于预置的第二负载程度和业务成功率大于或等于预置的第二业务成功率。

13. 根据权利要求 9 至 12 任意一项所述的方法，其特征在于，所述云服务系统还包含管理节点，所述第一工作节点基于所述待缩容的云实例的空闲资源量，减小所述待缩容的云实例的资源配额之后，所述方法还包括：

所述第一工作节点将所述待缩容的云实例的资源配额发送至所述管理节点。

14. 一种工作节点，其特征在于，所述工作节点作为第一工作节点，所述第一工作节点设置于云服务系统中，所述第一工作节点部署有多个云实例，所述第一工作节点包括：

获取模块，用于获取所述多个云实例的状态信息；

第一确定模块，用于基于所述状态信息，在所述多个云实例中确定待扩容的云实例以及扩容所需的资源量；

第一调整模块，用于若所述第一工作节点的空闲资源量大于或等于所述扩容所需的资源量，基于所述扩容所需的资源量，增大所述待扩容的云实例的资源配额。

15. 根据权利要求 14 所述的工作节点，其特征在于，所述第一确定模块，用于：

在所述多个云实例中，所述第一工作节点将状态信息满足预置的扩容条件的云实例确定为待扩容的云实例；

基于所述待扩容的云实例的状态信息，确定扩容所需的扩容所需的资源量。

16. 根据权利要求 14 或 15 所述的工作节点，其特征在于，所述云服务系统还包含第二工作节点，所述第一工作节点还包括：

第二确定模块，用于若所述第一工作节点的空闲资源量小于所述扩容所需的资源量，在所述多个云实例中确定待迁移的云实例，所述待迁移的云实例运行的业务的优先级低于预置

的优先级；

第一迁移模块，用于将所述待迁移的云实例迁移至第二工作节点，以更新所述第一工作节点的空闲资源量；

第二调整模块，用于若更新后的第一工作节点的空闲资源量大于或等于所述扩容所需的资源量，基于所述扩容所需的资源量，增大所述待扩容的云实例的资源配额。

17. 根据权利要求 16 所述的工作节点，其特征在于，所述云服务系统还包含第三工作节点，所述第一工作节点还包括：

检测模块，用于若更新后的第一工作节点的空闲资源量小于所述扩容所需的资源量，所述第一工作节点检测所述待扩容的云实例运行的业务的类型；

创建模块，用于若所述待扩容的云实例运行的业务为无状态应用，在所述第三工作节点处创建新的云实例，所述新的云实例和所述待扩容的云实例共同用于运行所述无状态应用；

第二迁移模块，用于若所述待扩容的云实例运行的业务为有状态应用，将所述待扩容的云实例迁移至第三工作节点。

18. 根据权利要求 15 所述的工作节点，其特征在于，所述状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

19. 根据权利要求 18 所述的工作节点，其特征在于，所述预置的扩容条件包括以下至少一种：资源占用率大于或等于预置的第一资源占用率、负载程度大于或等于预置的第一负载程度和业务成功率小于预置的第一业务成功率。

20. 一种工作节点，其特征在于，所述工作节点作为第一工作节点，所述第一工作节点设置于云服务系统中，所述第一工作节点部署有多个云实例，所述第一工作节点包括：

获取模块，用于获取所述多个云实例的状态信息；

确定模块，用于基于所述状态信息，在所述多个云实例中确定待缩容的云实例；

释放模块，用于释放所述待缩容的云实例的空闲资源，并确定被释放的所述待缩容的云实例的空闲资源量；

调整模块，用于基于所述待缩容的云实例的空闲资源量，减小所述待缩容的云实例的资源配额。

21. 根据权利要求 20 所述的工作节点，其特征在于，所述确定模块，用于在所述多个云实例中，将状态信息满足预置的缩容条件的云实例确定为待缩容的云实例。

22. 根据权利要求 20 或 21 所述的工作节点，其特征在于，所述状态信息包括以下至少一种：资源占用率、负载程度和业务成功率。

23. 根据权利要求 22 所述的工作节点，其特征在于，所述预置的缩容条件包括以下至少一种：资源占用率小于预置的第二资源占用率、负载程度小于预置的第二负载程度和业务成功率大于或等于预置的第二业务成功率。

24. 一种工作节点，其特征在于，所述工作节点包括存储器和处理器；所述存储器存储有代码，所述处理器被配置为执行所述代码，当所述代码被执行时，所述工作节点执行如权利要求 1 至 13 任一所述的方法。

25. 一种计算机存储介质，其特征在于，所述计算机存储介质存储有一个或多个指令，所述指令在由一个或多个计算机执行时使得所述一个或多个计算机实施权利要求 1 至 13 任一所

述的方法。

26. 一种计算机程序产品，其特征在于，所述计算机程序产品存储有指令，所述指令在由计算机执行时，使得所述计算机实施权利要求 1 至 13 任意一项所述的方法。

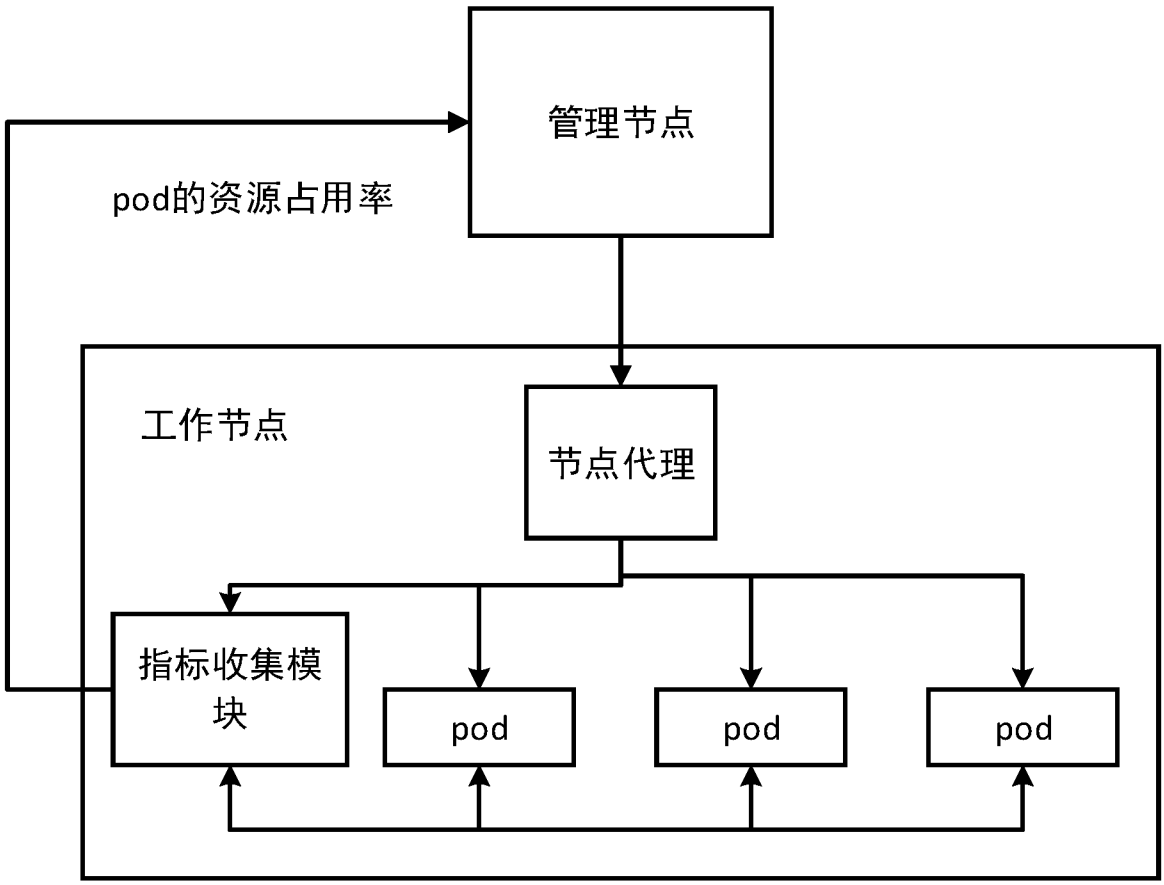

图 1

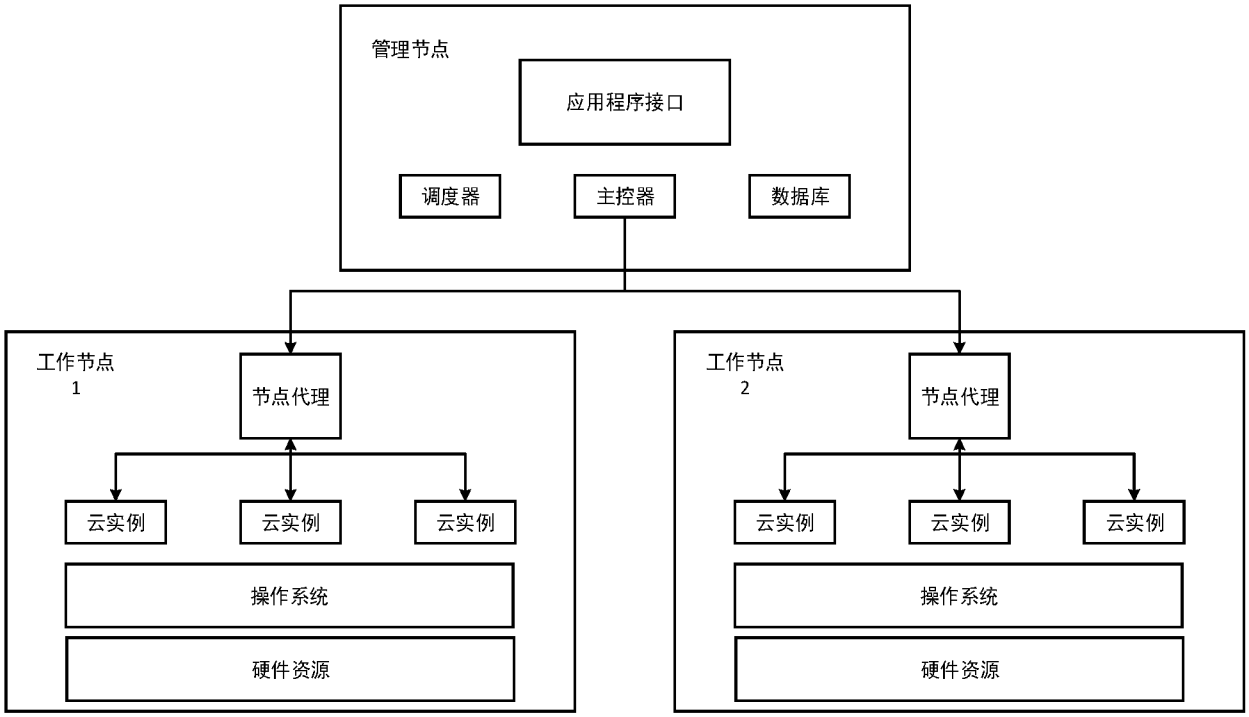
管理节点
应用程序接口
调度器
主控器
数据库
工作节点
1
节点代理
云实例
云实例
云实例
操作系统
硬件资源
工作节点
2
节点代理
云实例
云实例
云实例
操作系统
硬件资源

图 2

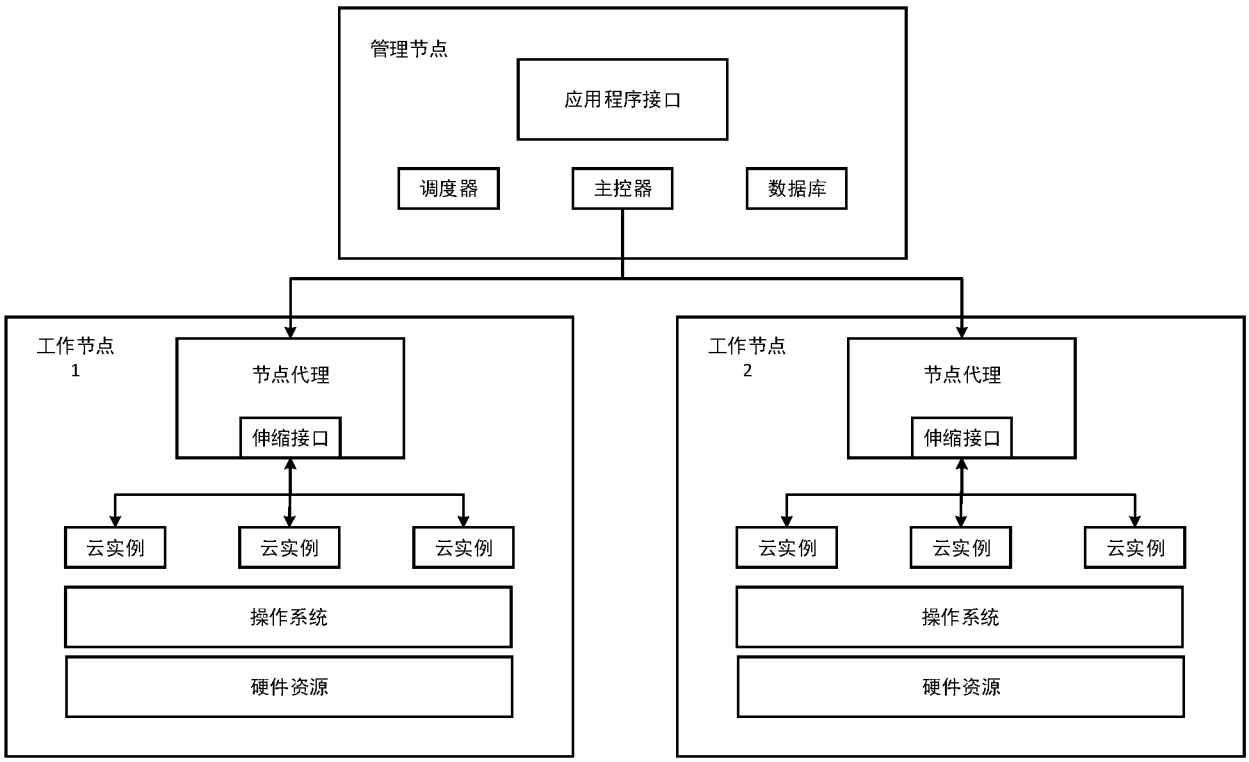
管理节点
应用程序接口
调度器
主控器
数据库
工作节点
1
节点代理
伸缩接口
云实例
云实例
云实例
操作系统
硬件资源
工作节点
2
节点代理
伸缩接口
云实例
云实例
云实例
操作系统
硬件资源

图 3

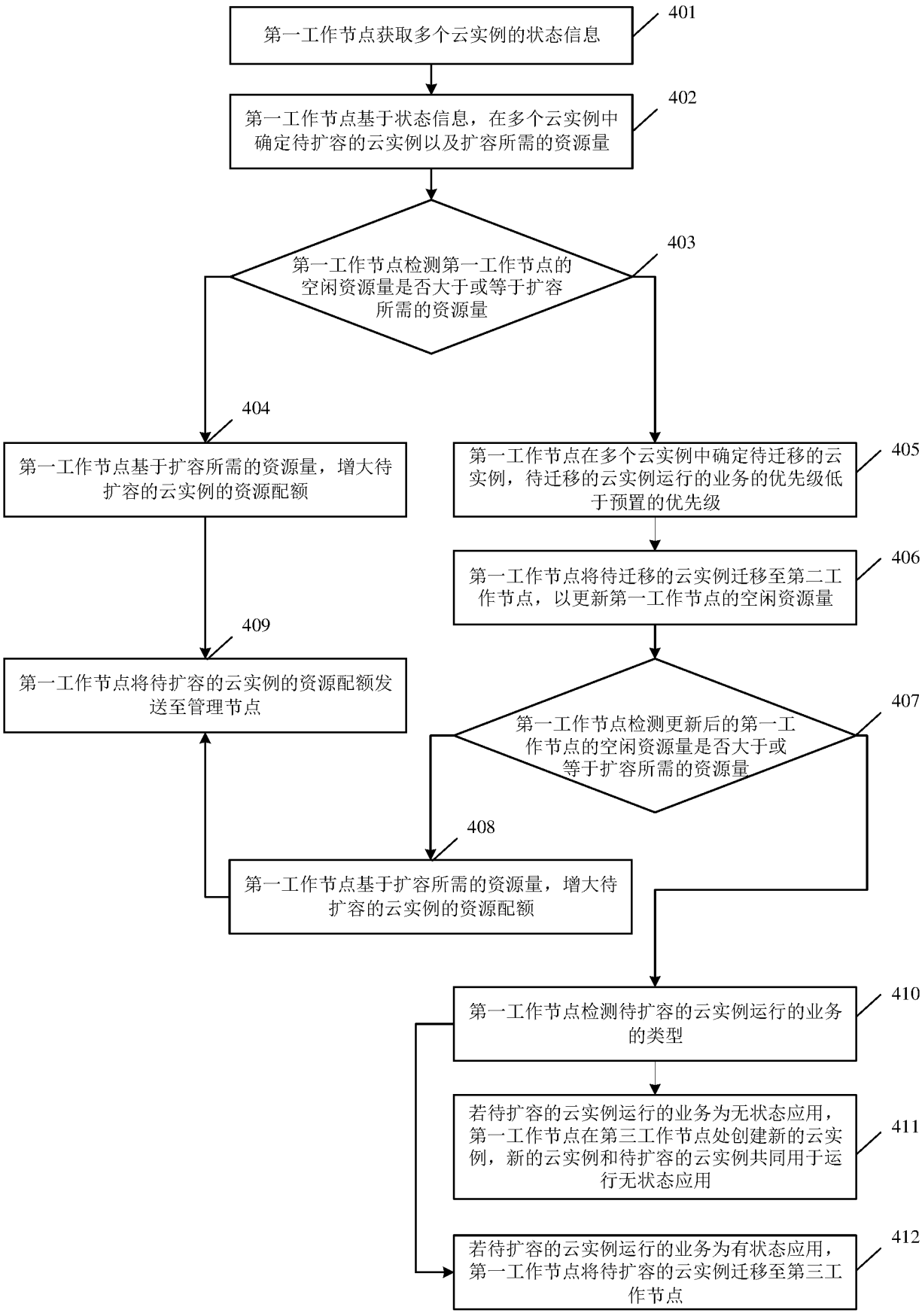
第一工作节点获取多个云实例的状态信息
401
第一工作节点基于状态信息，在多个云实例中确定待扩容的云实例以及扩容所需的资源量
402
第一工作节点检测第一工作节点的空闲资源量是否大于或等于扩容所需的资源量
403
404
第一工作节点基于扩容所需的资源量，增大待扩容的云实例的资源配额
405
第一工作节点在多个云实例中确定待迁移的云实例，待迁移的云实例运行的业务的优先级低于预置的优先级
406
第一工作节点将待迁移的云实例迁移至第二工作节点，以更新第一工作节点的空闲资源量
407
第一工作节点检测更新后的第一工作节点的空闲资源量是否大于或等于扩容所需的资源量
408
第一工作节点基于扩容所需的资源量，增大待扩容的云实例的资源配额
409
第一工作节点将待扩容的云实例的资源配额发送至管理节点
410
第一工作节点检测待扩容的云实例运行的业务的类型
411
若待扩容的云实例运行的业务为无状态应用，第一工作节点在第三工作节点处创建新的云实例，新的云实例和待扩容的云实例共同用于运行无状态应用
412
若待扩容的云实例运行的业务为有状态应用，第一工作节点将待扩容的云实例迁移至第三工作节点

图 4

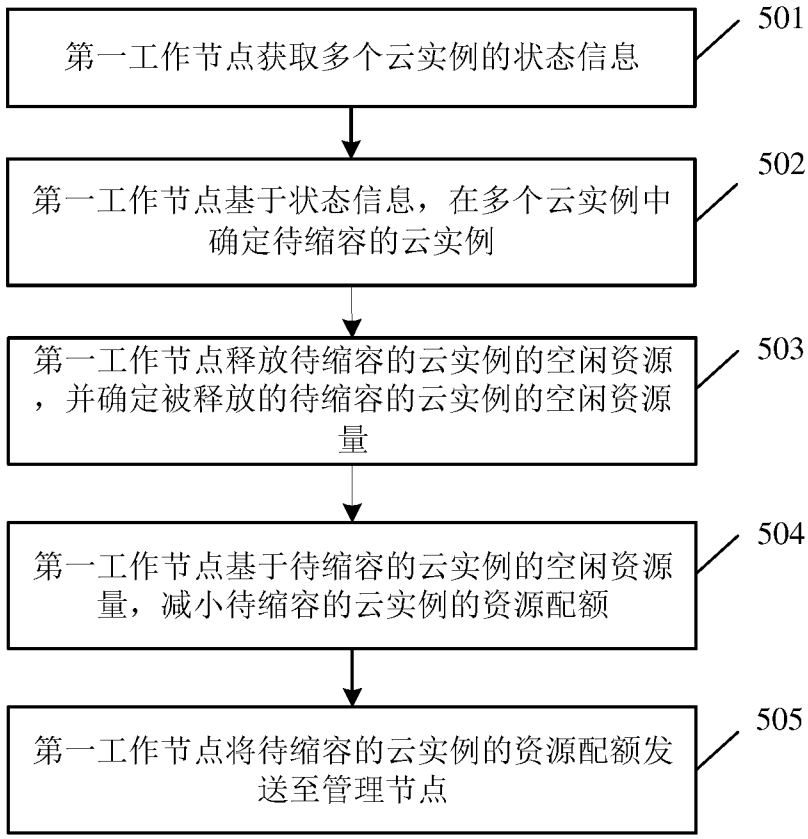
第一工作节点获取多个云实例的状态信息
501
第一工作节点基于状态信息，在多个云实例中确定待缩容的云实例
502
第一工作节点释放待缩容的云实例的空闲资源，并确定被释放的待缩容的云实例的空闲资源量
503
第一工作节点基于待缩容的云实例的空闲资源量，减小待缩容的云实例的资源配额
504
第一工作节点将待缩容的云实例的资源配额发送至管理节点
505

图 5

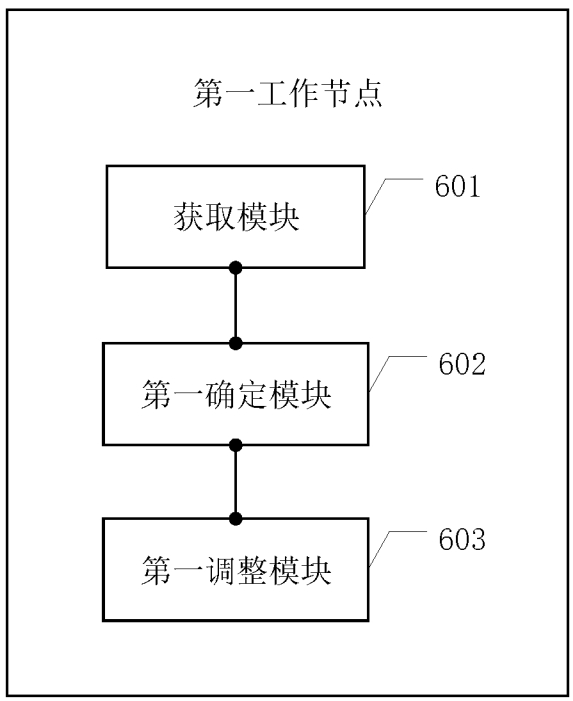
第一工作节点
获取模块
601
第一确定模块
602
第一调整模块
603

图 6

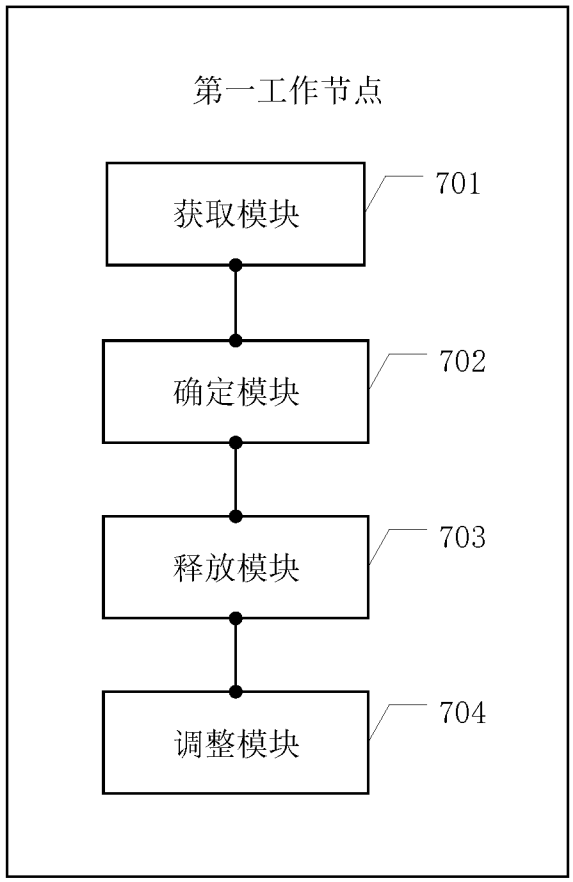
第一工作节点
获取模块
701
确定模块
702
释放模块
703
调整模块
704

图 7

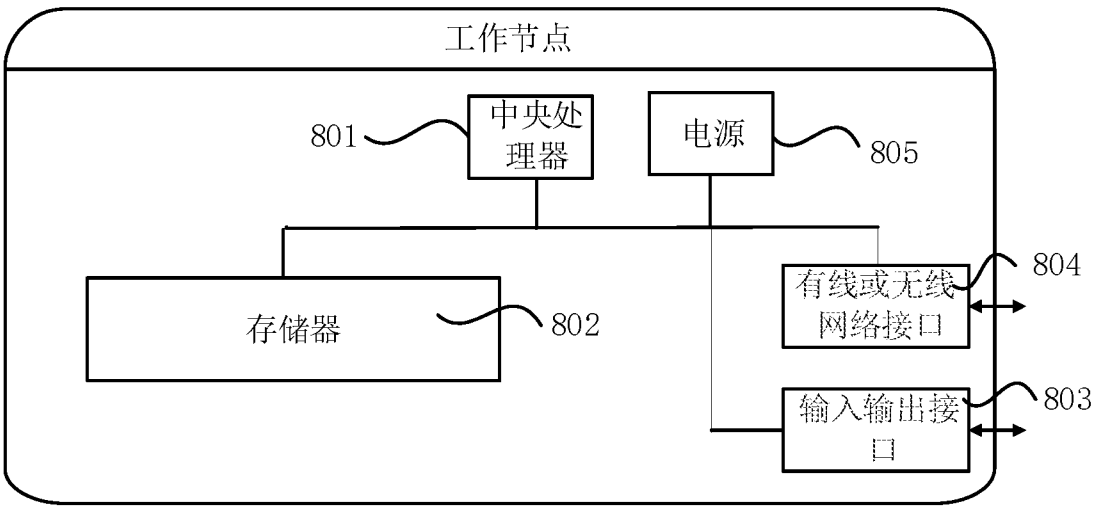
工作节点
801
中央处理器
电源
805
存储器
802
有线或无线网络接口
804
输入输出接口
803

图 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/134647

A. CLASSIFICATION OF SUBJECT MATTER

G06F9/455(2018.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; WPABS; DWPI; USTXT; WOTXT; EPTXT: 实例, 容器, 扩容, 扩大, 增大, 增加, 缩容, 缩小, 减小, 减少, 调整, 修改, 改变, 资源, 重启, pod, instance, container, capacity, expand, increase, add, shrink, reduce, adjust, modify, change, resource, restart

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113037794 A (MASHANG CONSUMER FINANCE CO., LTD.) 25 June 2021 (2021-06-25) description, paragraphs [0073]-[0163]	1-26
A	CN 112199194 A (GUANGZHOU HUYA TECHNOLOGY CO., LTD.) 08 January 2021 (2021-01-08) entire document	1-26
A	CN 113268310 A (SINA.COM TECHNOLOGY (CHINA) CO., LTD.) 17 August 2021 (2021-08-17) entire document	1-26
A	CN 113395178 A (JUHAOKAN TECHNOLOGY CO., LTD.) 14 September 2021 (2021-09-14) entire document	1-26

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
06 February 2023	**23 February 2023**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/134647

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 113037794 A	25 June 2021	None	
CN 112199194 A	08 January 2021	None	
CN 113268310 A	17 August 2021	None	
CN 113395178 A	14 September 2021	CN 113395178 B	09 December 2022

国际检索报告

国际申请号
PCT/CN2022/134647

A. 主题的分类

G06F9/455(2018.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC：G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS;CNTXT;CNKI;WPABS;DWPI;USTXT;WOTXT;EPTXT：实例，容器，扩容，扩大，增大，增加，缩容，缩小，减小，减少，调整，修改，改变，资源，重启，pod，instance，container，capacity，expand，increase，add，shrink，reduce，adjust，modify，change，resource，restart

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 113037794 A (马上消费金融股份有限公司) 2021年6月25日 (2021-06-25) 说明书第[0073]-[0163]段	1-26
A	CN 112199194 A (广州虎牙科技有限公司) 2021年1月8日 (2021-01-08) 全文	1-26
A	CN 113268310 A (新浪网技术（中国）有限公司) 2021年8月17日 (2021-08-17) 全文	1-26
A	CN 113395178 A (聚好看科技股份有限公司) 2021年9月14日 (2021-09-14) 全文	1-26

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2023年2月6日	2023年2月23日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

李维

电话号码 (+86) 0512-88996023

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/134647

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 113037794 A	2021年6月25日	无	
CN 112199194 A	2021年1月8日	无	
CN 113268310 A	2021年8月17日	无	
CN 113395178 A	2021年9月14日	CN 113395178 B	2022年12月9日