(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
2022 年 3 月 31 日 (31.03.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/062981 A1

(51) 国际专利分类号:
G06F 9/50 (2006.01)

(21) 国际申请号: PCT/CN2021/118436

(22) 国际申请日: 2021 年 9 月 15 日 (15.09.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011001012.8 2020年9月22日 (22.09.2020) CN

(71) 申请人: 中兴通讯股份有限公司 (ZTE CORPORATION) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。

(72) 发明人: 童遥 (TONG, Yao); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 王海新 (WANG, Haixin); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。

(74) 代理人: 北京天昊联合知识产权代理有限公司 (TEE&HOWE INTELLECTUAL PROPERTY ATTORNEYS); 中国北京市东城区东长安街1号东方广场东方经贸城西一办公楼5层1, 6-12室 崔利梅、梅丹丹, Beijing 100738 (CN)。

(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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, 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。

(54) **Title:** RESOURCE SCHEDULING METHOD AND SYSTEM, ELECTRONIC DEVICE, AND COMPUTER-READABLE STORAGE MEDIUM

(54) **发明名称:** 资源调度方法和系统、电子设备及计算机可读存储介质

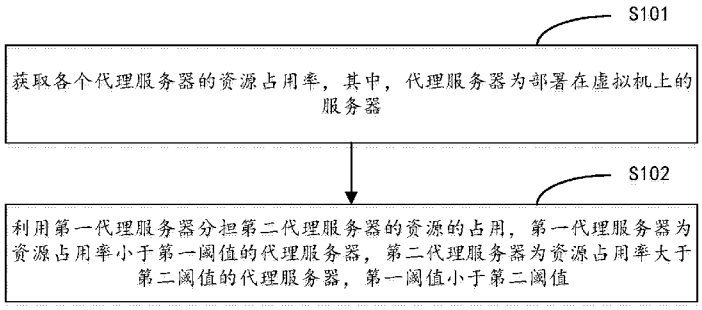

图 1

S101 Obtain resource occupancy rates of proxy servers, wherein the proxy servers are servers deployed on a virtual machine

S102 Share resource occupancy of a second proxy server by using a first proxy server, the first proxy server being a proxy server of which the resource occupancy rate is less than a first threshold, the second proxy server being a proxy server of which the resource occupancy rate is greater than a second threshold, and the first threshold being less than the second threshold

(57) **Abstract:** The present disclosure relates to the technical field of communications, and provides a resource scheduling method, comprising: obtaining resource occupancy rates of multiple proxy servers, the multiple proxy servers being deployed on a virtual machine; and sharing resource occupancy of at least one second proxy server by using at least one first proxy server, the resource occupancy rate of each of the at least one first proxy server being less than a first threshold, the resource occupancy rate of each of the at least one second proxy server being greater than a second threshold, and the first threshold being less than the second threshold. Also disclosed in the present disclosure are a resource scheduling system, an electronic device, and a computer-readable storage medium.

WO 2022/062981 A1

[见续页]

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, 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, 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))。

(57) 摘要：本公开涉及通信技术领域，并提供了一种资源调度方法，包括：获取多个代理服务器的资源占用率，多个代理服务器部署在虚拟机上；以及，利用至少一个第一代理服务器分担至少一个第二代理服务器的资源的占用，至少一个第一代理服务器中的每个第一代理服务器的资源占用率小于第一阈值，至少一个第二代理服务器中的每个第二代理服务器的资源占用率大于第二阈值，且第一阈值小于第二阈值。本公开还公开了一种资源调度系统、电子设备及计算机可读存储介质。

资源调度方法和系统、电子设备及计算机可读存储介质

本公开要求在 2020 年 9 月 22 日提交中国专利局、申请号为202011001012.8 的中国专利申请的优先权，该申请的全部内容通过引用结合在本公开中。

技术领域

本公开实施例涉及通信技术领域。

背景技术

随着互联网的普及和宽带基础设施建设的发展，用户对网络视频的需求也有很大增长。各网络视频网站已经覆盖 94%以上的互联网用户，与资讯、邮箱、即时通信（Instant Messaging，IM）等成为互联网的基础应用之一。

发明内容

本公开实施例的一个方面提供一种资源调度方法，包括：获取多个代理服务器的资源占用率，其中，多个代理服务器部署在虚拟机上；以及，利用至少一个第一代理服务器分担至少一个第二代理服务器的资源的占用；其中，至少一个第一代理服务器中的每个第一代理服务器的资源占用率小于第一阈值，至少一个第二代理服务器中的每个第二代理服务器的资源占用率大于第二阈值，且第一阈值小于第二阈值。

本公开实施例的另一个方面提供一种资源调度系统，包括调度服务器和多个代理服务器，多个代理服务器部署在虚拟机上；其中，调度服务器被配置为：获取多个代理服务器的资源占用率；以及，利用至少一个第一代理服务器分担至少一个第二代理服务器的资源的占用；其中，至少一个第一代理服务器中的每个第一代理服务器的资源占用率小于第一阈值，至少一个第二代理服务器中的每个第二代理服务器的资源占用率大于第二阈值，且第一阈值小于第二阈值。

本公开实施例的再一个方面提供一种电子设备，包括：至少一个处理器；以及，与至少一个处理器通信连接的存储器；其中，存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能够执行本公开实施例提供的资源调度方法。

本公开实施例的又一个方面提供一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现本公开实施例提供的资源调度方法。

附图说明

图 1 为资源调度方法的流程示意图。

图 2 为资源调度方法的流程示意图。

图 3 为资源调度系统的模块结构示意图。

图 4 为资源调度系统应用在云视频服务的示例图。

图 5 为图 4 对应的流程示意图。

图 6 为电子设备的结构示意图。

具体实施方式

为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合附图对本公开的实施例进行详细的阐述。然而，本领域的普通技术人员可以理解，在本公开实施例中，为了使读者更好地理解本申请而提出了许多技术细节。但是，即使没有这些技术细节和基于以下各实施例的种种变化和修改，也可以实现本申请所要求保护的技术方案。以下各个实施例的划分是为了描述方便，不应对本公开的具体实现方式构成任何限定，各个实施例在不矛盾的前提下可以相互结合相互引用。

传统的网络视频网站的视频服务都是基于物理机部署的，通过前端业务代理服务器进行负载调度。然而，各网络视频网站的资源（如带宽、中央处理器（Central Processing Unit，CPU）、内存等）使用情况是波动的，高峰时可能存在部分物理机的资源被占用过多的情况，导致网站的服务质量无法得到有效保证，影响用户体验。

本公开实施例的第一实施方式涉及一种资源调度方法，通过获取各个代理服务器的资源占用率，其种，代理服务器为部署在虚拟机上的服务器；利用第一代理服务器分担第二代理服务器的资源的占用，第一代理服务器为资源占用率小于第一阈值的代理服务器，第二代理服务器为资源占用率大于第二阈值的代理服务器，第一阈值小于第二阈值。通过虚拟化技术重新组合代理服务器的资源，利用资源占用率较低的代理服务器分担资源占用率较高的代理服务器，可以实现资源的灵活分配，从而保证每个代理服务器的资源占用率保持在合理范围内，有效保证网站的服务质量，提高用户体验。

应当说明的是，第一实施方式提供的资源调度方法的执行主体可以为各个代理服务器连接的服务端，服务端可以用独立的服务器或者是多个服务器组成的服务器集群来实现。可选地，服务端可以与接入服务器连接，接入服务器与客户端连接，用于处理客户端发起的请求。

第一实施方式提供的资源调度方法的具体流程如图 1 所示，具体包括以下步骤 S101 和步骤 S102。

在步骤 S101 中，获取各个代理服务器的资源占用率，代理服务器为部署在虚拟机上的服务器。

代理服务器是虚拟机通过虚拟化技术从物理服务器中虚拟出来的服务器，物理服务器为代理服务器的宿主机，一台物理服务器可以对应一台以上虚拟出来的代理服务器。

资源可以包括代理服务器的带宽、计算（包括 CPU 或图形处理器（Graphics Processing Unit，GPU））、存储（包括内存）资源等，而资源占用率是指资源被占用的比例。计算资源和存储资源可统称为处理资源，表示代理服务器的处理能力。应当说明的是，在第一实施方式中，代理服务器的资源可以指以上资源中至少一个资源。

可选地，代理服务器的资源为带宽资源，代理服务器为用于提供云视频服务的服务器。

在步骤 S102 中，利用第一代理服务器分担第二代理服务器的资源的占用，第一代理服务器为资源占用率小于第一阈值的代理服务器，

第二代理服务器为资源占用率大于第二阈值的代理服务器，第一阈值小于第二阈值。

第一阈值和第二阈值可以根据实际经验或实际情况进行设置，例如第一阈值可以为 20%、30%或 40%等，第二阈值可以为 70%、80%或 90%等。

在一个具体的例子中，在利用第一代理服务器分担第二代理服务器的资源的占用后，若还存在剩余的第一代理服务器，则回收剩余的第一代理服务器。由于第一代理服务器为资源占用率较低的代理服务器，因此回收剩余的第一代理服务器，可以减少代理服务器的数量，节省虚拟资源。可选地，也可以将剩余的第一代理服务器进行迁移，用到其它租户中。

可选地，对资源占用率在第一阈值与第二阈值之间的代理服务器可不作处理，亦可以在第一代理服务器不足以分担第二代理服务器的资源的占用时，使用该部分代理服务器分担第二代理服务器的资源。可选地，若在利用第一代理服务器分担第二代理服务器资源的占用后，仍存在部分第二代理服务器的资源的占用未被分担，则新建至少一个代理服务器，利用新建的代理服务器分担该部分第二代理服务器的资源的占用。在新建代理服务器时，可以优先在第二代理服务器的临近位置进行新建，例如优先在与第二代理服务器同一宿主（物理服务器）进行新建，或者在与第二代理服务器宿主附近的物理服务器进行新建。

可选地，在利用第一代理服务器分担第二代理服务器的资源的占用之前，还可以先增加第二代理服务器的资源；而利用第一代理服务器分担第二代理服务器的资源的占用，则包括若增加后的第二代理服务器的资源占用率仍大于第二阈值，再利用第一代理服务器分担增加资源后的第二代理服务器的资源的占用。

具体地，可以通过虚拟化技术在租户配额允许的范围内先增加第二代理服务器的资源。例如，若某一第二代理服务器的配额为 20M 的带宽，但当前只有 10M，则可以先将该第二代理服务器带宽配置为 20M 的带宽，再获取增加资源后该第二代理服务器的资源占用率；若增加资源后的资源占用率小于或等于第二阈值，则将该代理服务器从

第二代理服务器的列表中去除；若增加资源后的资源占用率仍大于第二阈值，则利用第一代理服务器分担该第二代理服务器的资源的占用。

可选地，在利用第一代理服务器分担第二代理服务器的资源的占用时，可以随机选取一个第一代理服务器分担第二代理服务器的资源的占用。

在一个具体的例子中，在步骤 S102 之前，第一实施方式中的资源调度方法还包括：以第一代理服务器的资源占用率从低到高的顺序，对各第一代理服务器进行排序；在步骤 S102 中，利用第一代理服务器分担第二代理服务器的资源的占用，包括：选择排序结果中前 n 个第一代理服务器用于分担第二代理服务器的资源占用，n 为正整数。

例如，将第一代理服务器的资源占用率从低到高排成一个队列，从队列的头部（即资源占用率较低）选取第一代理服务器分担第二代理服务器的资源的占用。

可以理解的是，资源占用率是随时波动的，通过选择资源占用率较低的第一代理服务器分担第二代理服务器的资源的占用，可以使代理服务器更加充分地应对资源的占用，保证服务的质量。另外，选择排序结果中前 n 个第一代理服务器可以包括优先选择排序结果前 n 个第一代理服务器，即在选择用于分担的第一代理服务器时，优先选择排序结果前 n 个第一代理服务器，在排序结果前 n 个第一代理服务器不足以分担第二代理服务器的资源的占用时，可以选择除排序结果前 n 个第一代理服务器之外的第一代理服务器来分担。

同样地，还可以以第二代理服务器的资源占用率从高到低的顺序，对各个第二代理服务器进行排序，利用第一代理服务器优先分担资源占用率较高的第二代理服务器的资源的占用。

根据第一实施方式中的资源调度方法，通过获取虚拟机上代理服务器的资源占用率，利用资源占用率较低的代理服务器分担资源占用率较高的代理服务器，通过虚拟化技术重新组合代理服务器的资源，可以实现整体资源的灵活分配，从而保证每个代理服务器的资源占用率保持在一个合理的范围内，有效保证服务器（网站）的服务质量，提高用户体验。

本公开实施例的第二实施方式涉及一种资源调度方法，第二实施方式与第一实施方式大致相同，主要区别在于：在第二实施方式中，还对第二代理服务器的资源占用需求进行预测，根据资源占用需求的情况进行代理服务器资源占用的分担。

第二实施方式提供的资源调度方法的具体流程如图 2 所示，具体包括以下步骤 S201-步骤 S204。

在步骤 S201 中，获取各个代理服务器的资源占用率，代理服务器为部署在虚拟机上的服务器。

步骤 S201 与第一实施方式中的步骤 S101 相同，具体可参见第一实施方式中的相关描述，这里不再赘述。

在步骤 S202 中，预测每一第二代理服务器的资源占用需求。

资源占用需求是指代理服务器可能出现的资源占用比例，例如某一代理服务器在高峰时的资源占用率为 90%，非高峰时的资源占用率为 20%，则该代理服务器的资源占用需求为 20%-90%。

可选地，可以根据各第二代理服务器的资源占用率的历史数据预测其资源占用需求，例如选取一个月内的资源占用率的历史数据进行预测，具体选取的历史数据的时间范围可以根据实际需要进行设置，此处不做具体限制。

在步骤 S203 中，以第二代理服务器的资源差值从大到小的顺序，对各第二代理服务器进行排序，资源差值为资源占用需求最大值与资源占用需求平均值的差值。

可选地，资源占用需求最大值可以为历史数据中的最大值，亦可以为历史数据在高峰时的平均值，还可以根据历史数据作一定修改后得到；资源占用需求平均值可以为历史数据的平均值，亦可以为高峰时的历史数据和非高峰时的历史数据的加权平均，具体可以根据实际需要进行具体设置，此处不做具体限制。

在步骤 S204 中，利用第一代理服务器分担排序结果中前 m 个第二代理服务器的资源的占用，m 为正整数。

应当理解的是，当资源差值较大时，表示第二代理服务器可能会出现较高的资源占用率，需要保证这些第二代理服务器的资源的占

用被有效分担。在利用第一代理服务器分担排序结果中前 m 个第二代理服务器的资源的占用时，可以是利用第一代理服务器优先分担排序结果中前 m 个第二代理服务器的资源的占用；在第一代理服务器足于分担排序结果中前 m 个第二代理服务器的资源的占用时，也可以利用第一代理服务器分担其它第二代理服器的资源的占用。

在一个具体的例子中，可以将资源占用需求的平均值大于第二阈值的代理服务器作为第二代理服务器，从而使第二代理服务器的划分更加稳定。

可选地，可以根据资源占用需求最大值对第二代理服务器进行排序，利用第一代理服务器优先分担资源占用需求最大值较大的第二代理服务器的资源的占用。

应当说明的是，由于资源可以是多种，因此实际应用中可以根据资源进行分类，再根据分类后的代理服务器运用第一实施方式和/或第二实施方式中的方法进行资源的调度。为了更加清楚地说明本公开提供的资源调度方法，以下以将代理服务器分为六类、应用场景为云视频服务（即注重带宽资源）为例进行说明。

队列 1：代理服务器带宽资源占用率过低且处理资源占用率正常（带宽资源占用率小于 20%，且 CPU、内存、存储平均资源占用率在 20%-80%），可作为第一代理服务器，作为其它代理服务器分担的资源池。

调度方法：将代理服务器按带宽资源占用率升序排列，优先选择带宽资源占用率低的代理服务器进行分担。

队列 2：代理服务器带宽资源占用率正常且处理资源占用率稳定（带宽资源占用率在 20%-80%，且 CPU、内存、存储平均资源占用率在 20%-80%）。

调度方法：对此队列不做处理。

队列 3：代理服务器带宽资源占用率过高且处理资源占用率正常（带宽资源占用率超过 80%，且 CPU、内存、存储平均资源占用率在 20%-80%），可作为第二代理服务器。

调度方法：可选择队列 1 中的代理服务器为其分担，否则尝试

在允许范围增加代理服务器的带宽资源，或者，向服务端申请实施虚拟机迁移，或者，在队列 3 代理服务器的临近位置创建新的代理服务器进行分担。

队列 4：代理服务器资源占用率过高且处理资源占用率过高（带宽资源占用率超过 80%，且 CPU、内存、存储平均资源占用率超过 80%），可作为第二代理服务器。

调度方法：在配额允许范围内增加代理服务器的带宽资源，使分配结束后代理服务器的资源占用情况接近队列 6 中代理服务器。若带宽资源增加后代理服务器的资源占用率仍过高，则根据历史数据对每个代理服务器的带宽、CPU、内存、存储资源占用需求进行预测，将代理服务器按预测最大值（资源占用需求最大值）与预测平均值（资源占用需求平均值）的差值降序排列，然后从队列 1 中选择资源占用率最低的代理服务器进行分担，仍以接近队列 6 中代理服务器的资源占用情况为目标。如队列 1 中无合适的代理服务器，则申请在代理服务器的临近位置建立新的代理服务器进行分担，将分担后的新代理服务器加入队列 1，原代理服务器加入队列 6。

队列 5：代理服务器资源占用率过低且处理资源占用率过高（带宽资源占用率小于 20%，且 CPU、内存、存储平均资源占用率超过 80%），可作为第一代理服务器。

调度方法：将带宽预测需求最大值高于实际占有值的代理服务器进行降序排列，其余代理服务器加入队列 1，从队列 1 尾部选择物理位置临近的代理服务器对队列 1 头的代理服务器进行分担。如队列 5 中资源不足，从队列 1 中选择代理服务器进行分配。

队列 6：代理服务器资源占用率正常且处理资源占用率过高（即带宽资源占用率在 20%-80%，且 CPU、内存、存储平均资源占用率超过 80%）。

调度方法：根据历史数据对每个代理服务器的带宽占用需求进行预测，将代理服务器按带宽资源占用需求最大值和带宽资源占用需求平均值的差值进行升序排列，在队列 1 中依次选出带宽资源占用率低的代理服务器对这些代理服务器进行分担。如队列 1 中资源不足，

则申请在队列尾的至少一个代理服务器临近位置建立新的代理服务器分担负载，加入队列 1 继续分配。分配完毕后，如队列 1 中代理服务器仍有剩余资源，可将这些剩余资源与队列 4 和 6 中的代理服务器关联起来作为备用。对其余代理服务器可通过配置减少虚拟资源，或者对其余代理服务器进行迁移、关闭等操作。

根据第二实施方式中的资源调度方法，通过预测第二代理服务器的资源占用需求，利用第一代理服务器分担排序结果中资源差值较大的第二代理服务器的资源的占用，可以保证资源占用可能出现较大变化的第二代理服务器的资源，进一步优化了资源的调度，提高了代理服务器服务质量的稳定性。

此外，本领域技术人员可以理解，上面各种方法的步骤划分，只是为了描述清楚，实现时可以合并为一个步骤或者对某些步骤进行拆分，分解为多个步骤，只要包括相同的逻辑关系，都在本专利的保护范围内；对算法中或者流程中添加无关紧要的修改或者引入无关紧要的设计，但不改变其算法和流程的核心设计都在该专利的保护范围内。

本公开实施例的第三实施方式涉及一种资源调度系统 300，如图 3 所示，包括调度服务器 301 和多个代理服务器 302，代理服务器 302 为部署于虚拟机上的服务器。

调度服务器 301 被配置为：获取各个代理服务器的资源占用率；以及，利用第一代理服务器分担第二代理服务器的资源的占用，第一代理服务器为资源占用率小于第一阈值的代理服务器，第二代理服务器为资源占用率大于第二阈值的代理服务器，第一阈值小于第二阈值。

进一步地，调度服务器 301 还被配置为：以第一代理服务器的资源占用率从低到高的顺序，对各第一代理服务器进行排序，选择排序结果中前 n 个第一代理服务器用于分担第二代理服务器的资源的占用，n 为正整数。

进一步地，调度服务器 301 还被配置为：在存在剩余的第一代理服务器时，回收剩余的第一代理服务器。

进一步地，调度服务器 301 还被配置为：预测各第二代理服务

器的资源占用需求，以第二代理服务器的资源差值从大到小的顺序，对第二代理服务器进行排序，资源差值为资源占用需求最大值与资源占用需求平均值的差值，利用第一代理服务器分担排序结果中前 m 个第二代理服务器的资源的占用，m 为正整数。

进一步地，调度服务器 301 还被配置为：在存在第二代理服务器的资源的占用未被分担时，新建至少一个代理服务器，利用新建的代理服务器分担第二代理服务器的资源的占用。

进一步地，调度服务器 301 还被配置为：增加第二代理服务器的资源，若增加后的第二代理服务器的资源占用率仍大于第二阈值，则利用第一代理服务器分担增加资源后的第二代理服务器的资源的占用。

进一步地，资源为带宽资源，代理服务器为用于提供云视频服务的服务器。

进一步地，资源为处理资源，处理资源包括计算资源和存储资源。

请参考图 4，其为第三实施方式提供的资源调度系统应用在云视频服务的示例图，其中调度服务器 301 为图 4 中的监控服务器。具体地，在图 4 中，除了资源调度系统 300 的调度服务器 301（监控服务器）和代理服务器外，还包括客户端、接入服务器、目录服务器、内容服务器和编码转换服务器。物理服务器群为虚拟服务器群的宿主。客户端可以是计算机、电视机或手机等数字终端，通过无线网络、有线电视网络或互联网与接入服务器连接，用于获取云网络的视频服务。中心云由目录服务器和内容服务器组成，目录服务器为代理云提供视频对象的查找服务，内容服务器存有所有视频对象的完整备份。近客户端代理云由接入服务器、监控服务器、代理服务器和编码转换服务器组成。接入服务器作为客户端向代理云请求服务的接口，接收客户端的服务请求，向监控服务器获取可提供服务的代理服务器列表及各自运行状态，选择合适的代理服务器与客户端建立连接并提供服务。监控服务器用于收集各代理服务器中视频资源的分布信息，为客户端提供查找视频资源服务，利用本公开第一实施方式和第二实施方式提

供的资源调度方法，在收集各代理服务器运行状态信息的同时，为各代理服务器平衡负载，调度资源分配，即选择为客户端数据请求提供服务的服务器和调度相关虚拟资源；编码转换服务器由专门的物理服务器担当，用于将提供商或用户上传的不同码制的视频对象转换成符合云网络视频服务平台的码制的视频对象。

请参考图 5，其为图 4 在客户端获取视频服务时的流程示意图；具体的流程为：1、客户端通过互联网连接到云中的接入服务器，提交视频服务请求（包括视频对象）；2、接入服务器向监控服务器获取可为客户端提供服务的代理服务器列表；3、监控服务器搜索可为客户端提供服务的代理服务器，如代理云内无客户端请求的资源则向中心云的目录服务器查询资源，否则转步骤 6；4、中心云的目录服务器查找并选择存有客户所需数据的一个内容服务器；5、内容服务器将所需资源传送给代理云中的代理服务器（例如可以是第一代理服务器）；6、监控服务器根据掌握的代理服务器资源和运行状态信息，选择第一代理服务器为客户端服务，同时平衡各个代理服务器的资源占用率，并在完成选择后，将服务信息返回给接入服务器；7、接入服务器建立代理服务器与客户端的连接；8、代理服务器响应客户端的数据请求；9、客户端接收并缓冲一部分数据后开始播放，继续请求剩余数据。

不难发现，第三实施方式为与第一实施方式及第二实施方式相对应的系统实施方式，第三实施方式可与第一实施方式及第二实施方式互相配合实施。第一实施方式及第二实施方式中提到的相关技术细节在第三实施方式中依然有效，为了减少重复，这里不再赘述。相应地，第三实施方式中提到的相关技术细节也可应用在第一实施方式及第二实施方式中。

值得一提的是，第三实施方式中所涉及到的各模块均为逻辑模块，在实际应用中，一个逻辑单元可以是一个物理单元，也可以是一个物理单元的一部分，还可以以多个物理单元的组合实现。此外，为了突出本公开的创新部分，第三实施方式中并没有将与解决本公开所提出的技术问题关系不太密切的单元引入，但这并不表明第三实施方

式中不存在其它的单元。

本公开实施例的第四实施方式涉及一种电子设备，如图6所示，包括：至少一个处理器402；以及，与至少一个处理器402通信连接的存储器401；存储器401存储有可被至少一个处理器402执行的指令，指令被至少一个处理器402执行，以使至少一个处理器402能够执行上述的资源调度方法。

存储器401和处理器402采用总线方式连接，总线可以包括任意数量的互联的总线和桥，总线将一个或多个处理器402和存储器401的各种电路连接在一起。总线还可以将诸如外围设备、稳压器和功率管理电路等之类的各种其他电路连接在一起，这些都是本领域所公知的，因此，本文不再对其进行进一步描述。总线接口在总线和收发机之间提供接口。收发机可以是一个元件，也可以是多个元件，比如多个接收器和发送器，提供用于在传输介质上与各种其他装置通信的单元。经处理器402处理的数据通过天线在无线介质上进行传输，进一步，天线还接收数据并将数据传送给处理器402。

处理器402负责管理总线和通常的处理，还可以提供各种功能，包括定时，外围接口，电压调节、电源管理以及其他控制功能。而存储器401可以被用于存储处理器402在执行操作时所使用的数据。

本公开实施例第五实施方式涉及一种计算机可读存储介质，其上存储有计算机程序。计算机程序被处理器执行时实现本公开提供的任一个资源调度方法。

即，本领域技术人员可以理解，实现上述实施方式中的方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序存储在一个存储介质中，包括至少一个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（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。

本领域的普通技术人员可以理解，上述各实施例是实现本公开

的具体实施例，而在实际应用中，可以在形式上和细节上对其作各种改变，而不偏离本公开的精神和范围。

权利要求

1. 一种资源调度方法，包括：

获取多个代理服务器的资源占用率，其中，所述多个代理服务器部署在虚拟机上；以及

利用至少一个第一代理服务器分担至少一个第二代理服务器的资源的占用；其中，所述至少一个第一代理服务器中的每个第一代理服务器的资源占用率小于第一阈值，所述至少一个第二代理服务器中的每个第二代理服务器的资源占用率大于第二阈值，且所述第一阈值小于所述第二阈值。

2. 根据权利要求 1 所述的资源调度方法，其中，

在利用所述至少一个第一代理服务器分担所述至少一个第二代理服务器的资源的占用之前，所述方法还包括：根据资源占用率从低到高的顺序，对所述至少一个第一代理服务器进行排序，得到第一排序结果；以及

利用所述至少一个第一代理服务器分担所述至少一个第二代理服务器的资源的占用，包括：

确定所述第一排序结果中的前 n 个第一代理服务器，所述 n 为正整数；以及

利用所述前 n 个第一代理服务器分担所述至少一个第二代理服务器的资源的占用。

3. 根据权利要求 2 所述的资源调度方法，在利用所述前 n 个第一代理服务器分担所述至少一个第二代理服务器的资源的占用之后，还包括：

响应于确定所述至少一个第一代理服务器的数量大于所述 n，回收所述至少一个第一代理服务器中的除所述前 n 个第一代理服务器之外的其它第一代理服务器。

4. 根据权利要求 1 所述的资源调度方法，其中，

在利用所述至少一个第一代理服务器分担所述至少一个第二代理服务器的资源的占用之前，所述方法还包括：

预测所述至少一个第二代理服务器的资源占用需求；以及

根据资源差值从大到小的顺序，对所述至少一个第二代理服务器进行排序，得到第二排序结果；其中，所述第二代理服务器的资源差值为所述第二代理服务器的资源占用需求最大值与所述第二代理服务器的资源占用需求平均值的差值；以及

利用所述至少一个第一代理服务器分担所述至少一个第二代理服务器的资源的占用，包括：

确定所述第二排序结果中的前 m 个第二代理服务器，所述 m 为正整数；以及

利用所述至少一个第一代理服务器分担所述前 m 个第二代理服务器的资源的占用。

5. 根据权利要求 1 所述的资源调度方法，在利用所述至少一个第一代理服务器分担所述至少一个第二代理服务器的资源的占用之后，还包括：

响应于确定所述至少一个第二代理服务器的数量大于所述 m，新建至少一个第三代理服务器；以及

利用所述至少一个第三代理服务器分担所述至少一个第二代理服务器中的除所述前 m 个第二代理服务器之外的其它第二代理服务器的资源的占用。

6. 根据权利要求 1 所述的资源调度方法，其中，

在利用所述至少一个第一代理服务器分担所述至少一个第二代理服务器的资源的占用之前，还包括：增加所述至少一个第二代理服务器的资源；以及

利用所述至少一个第一代理服务器分担所述至少一个第二代理服务器的资源的占用，包括：

响应于确定增加资源后的第二代理服务器的资源占用率仍大于

所述第二阈值，利用所述至少一个第一代理服务器分担所述增加资源后的第二代理服务器的资源的占用。

7. 根据权利要求 1 至 6 中任一项所述的资源调度方法，其中，所述资源为带宽资源，以及所述多个代理服务器用于提供云视频服务。

8. 根据权利要求 1 至 6 中任一项所述的资源调度方法，其中，所述资源为处理资源，且所述处理资源包括计算资源和存储资源。

9. 一种资源调度系统，包括调度服务器和多个代理服务器，所述多个代理服务器部署在虚拟机上；其中，所述调度服务器被配置为：

获取所述多个代理服务器的资源占用率；以及

利用至少一个第一代理服务器分担至少一个第二代理服务器的资源的占用；其中，所述至少一个第一代理服务器中的每个第一代理服务器的资源占用率小于第一阈值，所述至少一个第二代理服务器中的每个第二代理服务器的资源占用率大于第二阈值，且所述第一阈值小于所述第二阈值。

10. 一种电子设备，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行根据权利要求 1 至 8 任一项所述的资源调度方法。

11. 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现根据权利要求 1 至 8 任一项所述的资源调度方法。

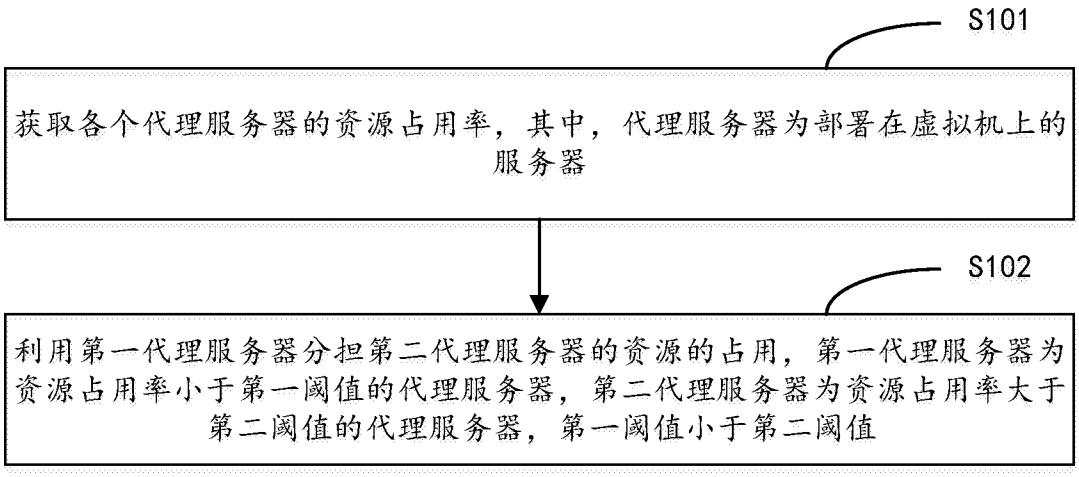
S101
获取各个代理服务器的资源占用率，其中，代理服务器为部署在虚拟机上的服务器
S102
利用第一代理服务器分担第二代理服务器的资源的占用，第一代理服务器为资源占用率小于第一阈值的代理服务器，第二代理服务器为资源占用率大于第二阈值的代理服务器，第一阈值小于第二阈值

图 1

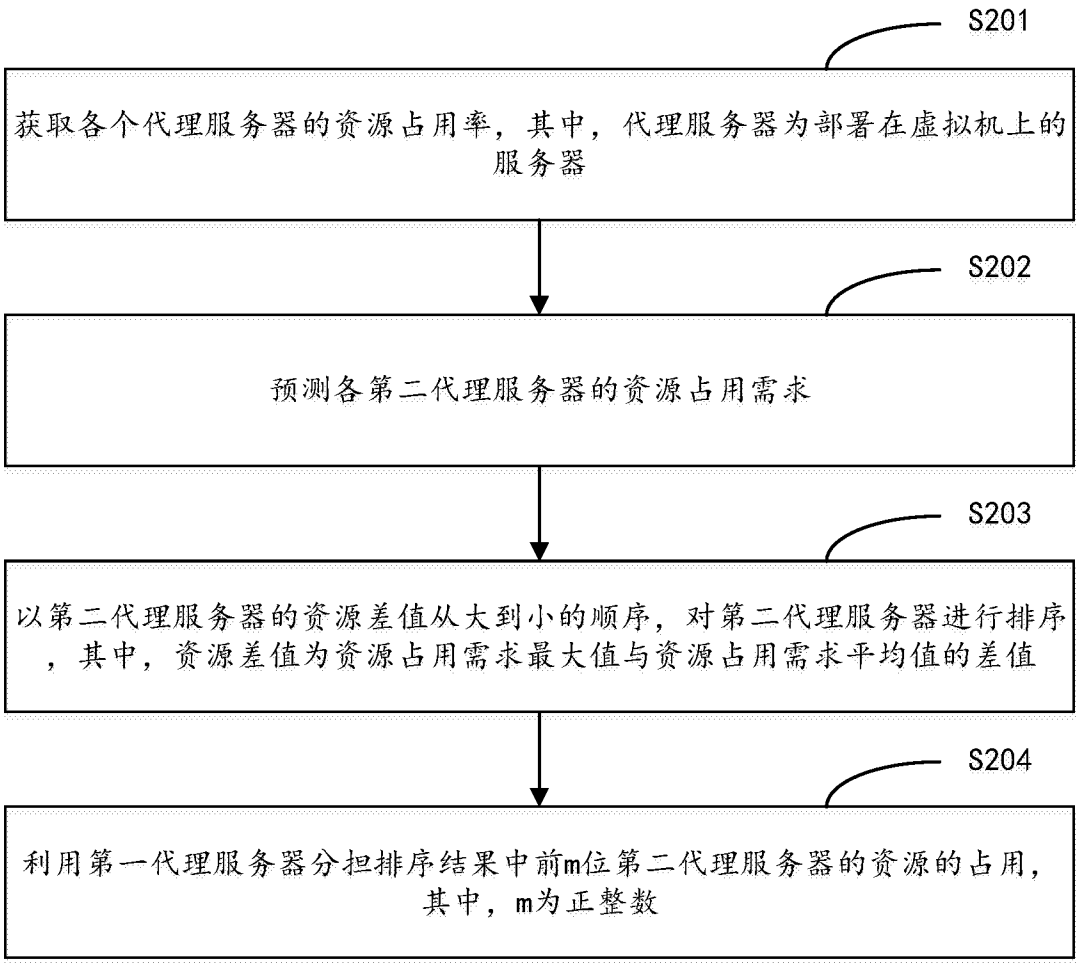
S201
获取各个代理服务器的资源占用率，其中，代理服务器为部署在虚拟机上的服务器
S202
预测各第二代理服务器的资源占用需求
S203
以第二代理服务器的资源差值从大到小的顺序，对第二代理服务器进行排序，其中，资源差值为资源占用需求最大值与资源占用需求平均值的差值
S204
利用第一代理服务器分担排序结果中前m位第二代理服务器的资源的占用，其中，m为正整数

图 2

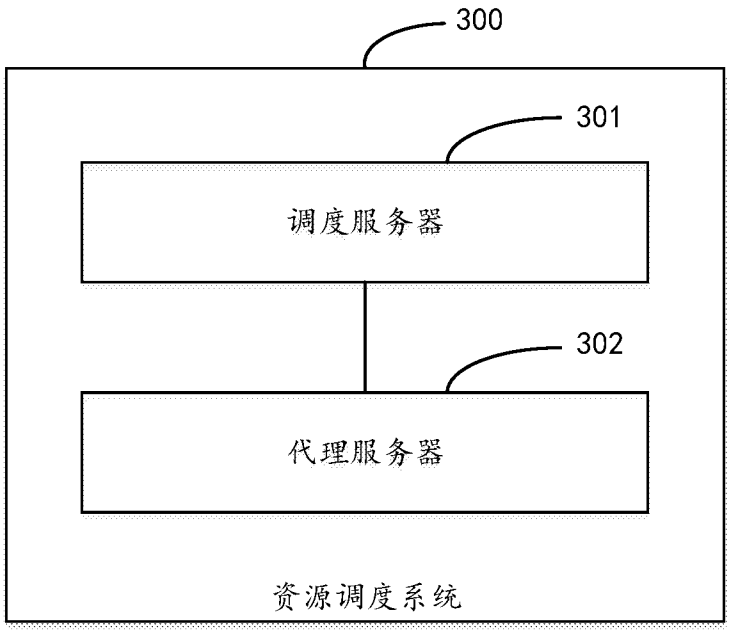
300
301
调度服务器
302
代理服务器
资源调度系统

图 3

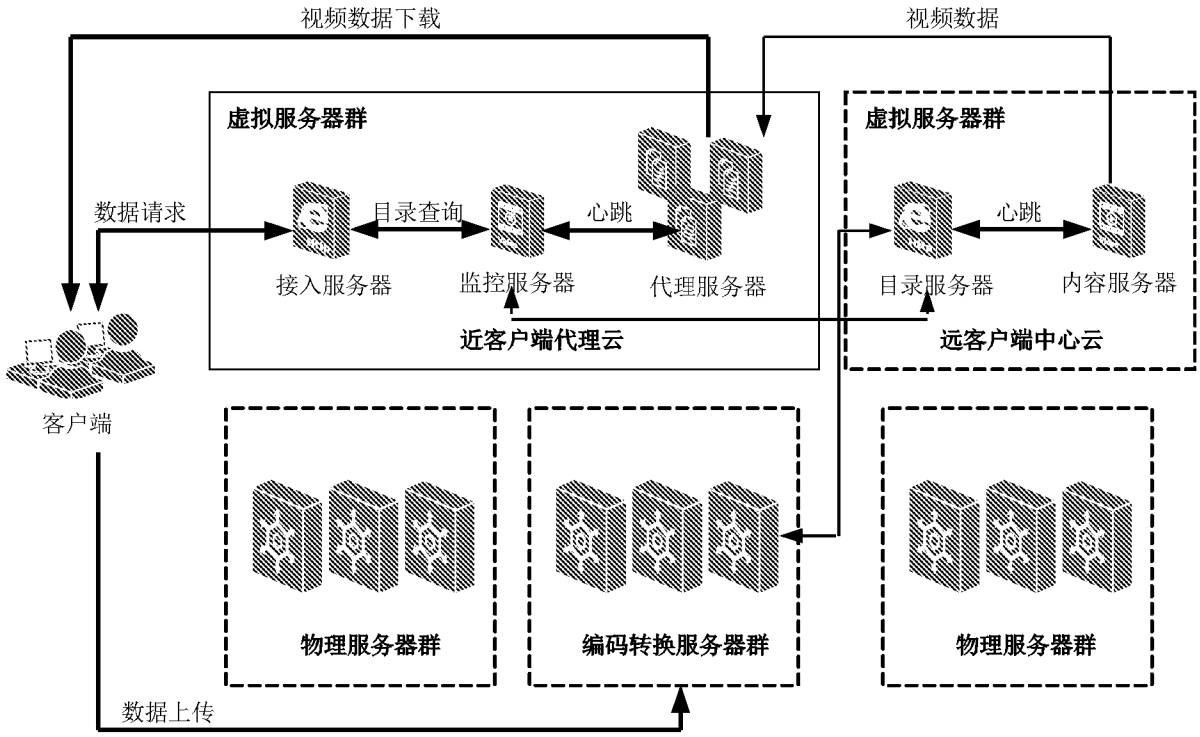
视频数据下载
视频数据
虚拟服务器群
虚拟服务器群
数据请求
目录查询
心跳
心跳
接入服务器
监控服务器
代理服务器
目录服务器
内容服务器
近客户端代理云
远客户端中心云
客户端
物理服务器群
编码转换服务器群
物理服务器群
数据上传

图 4

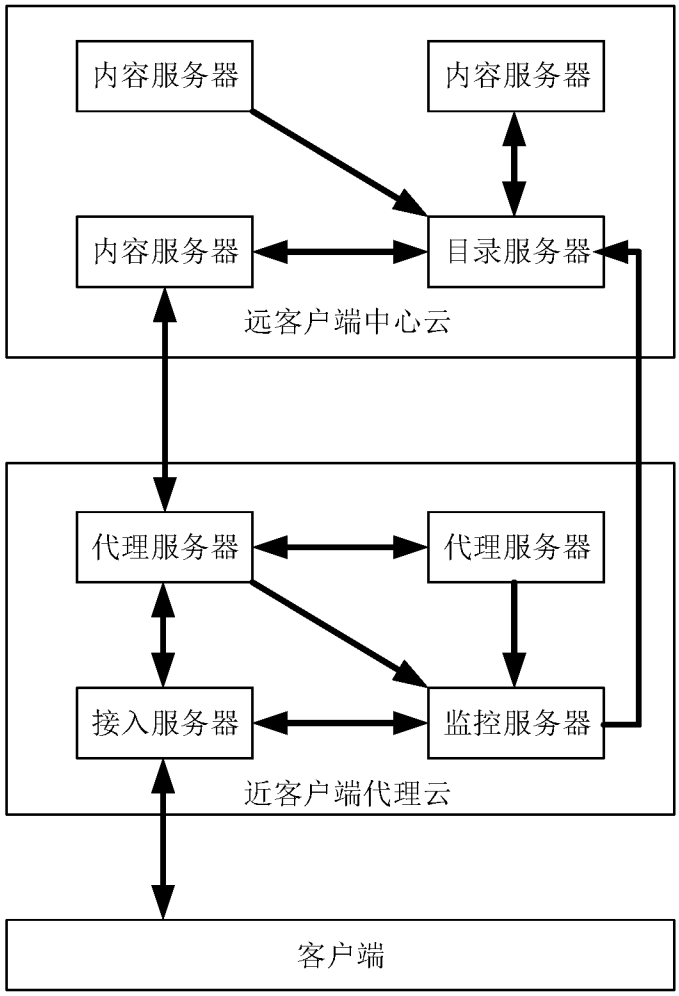
内容服务器
内容服务器
内容服务器
目录服务器
远客户端中心云
代理服务器
代理服务器
接入服务器
监控服务器
近客户端代理云
客户端

图 5

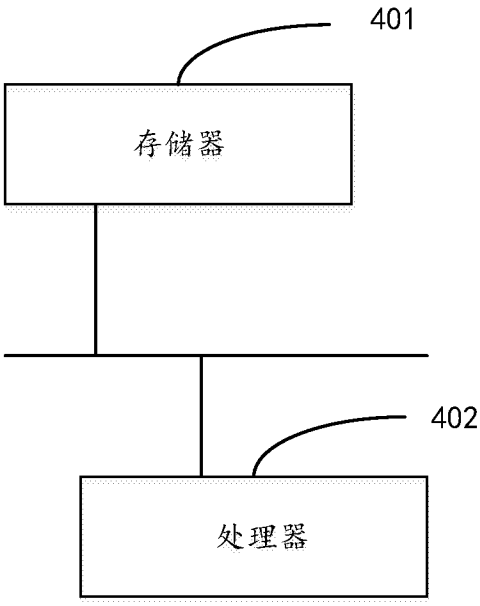
401
存储器
402
处理器

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/118436

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/50(2006.01)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)

G06F; H04L

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)

WPI, EPODOC, CNKI, CNPAT, IEEE, GOOGLE: 资源, 调度, 代理, 服务器, 负载, 占用, 阈值, resource, schedule, agent, server, load, occupancy, threshold

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 110597631 A (GUANGZHOU HUADUO NETWORK TECHNOLOGY CO., LTD.) 20 December 2019 (2019-12-20) description paragraphs 34-63, figure 2	1-11
A	CN 105553721 A (LANGCHAO ELECTRONIC INFORMATION INDUSTRY CO., LTD.) 04 May 2016 (2016-05-04) entire document	1-11
A	CN 108924139 A (HANGHZOU ANHENG INFORMATION TECHNOLOGY CO., LTD.) 30 November 2018 (2018-11-30) entire document	1-11
A	US 2019220369 A1 (EMC IP HOLDING COMPANY L.L.C.) 18 July 2019 (2019-07-18) entire document	1-11

☐ 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
"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
17 November 2021	**14 December 2021**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/118436

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 110597631 A	20 December 2019	None	
CN 105553721 A	04 May 2016	None	
CN 108924139 A	30 November 2018	None	
US 2019220369 A1	18 July 2019	CN 110058966 A	26 July 2019

国际检索报告

国际申请号
PCT/CN2021/118436

A. 主题的分类

G06F 9/50(2006.01)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)

G06F; H04L

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））

WPI, EPODOC, CNKI, CNPAT, IEEE, GOOGLE: 资源, 调度, 代理, 服务器, 负载, 占用,阈值, resource, schedule, agent, server, load, occupancy, threshold

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 110597631 A (广州华多网络科技有限公司) 2019年12月20日 (2019-12-20) 说明书第34-63段, 附图2	1-11
A	CN 105553721 A (浪潮电子信息产业股份有限公司) 2016年5月4日 (2016-05-04) 全文	1-11
A	CN 108924139 A (杭州安恒信息技术股份有限公司) 2018年11月30日 (2018-11-30) 全文	1-11
A	US 2019220369 A1 (EMC IP HOLDING COMPANY LLC) 2019年7月18日 (2019-07-18) 全文	1-11

☐ 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

“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
2021年11月17日	2021年12月14日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 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 王艳臣 电话号码 86-(10)-53961435

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/118436

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 110597631 A	2019年12月20日	无	
CN 105553721 A	2016年5月4日	无	
CN 108924139 A	2018年11月30日	无	
US 2019220369 A1	2019年7月18日	CN 110058966 A	2019年7月26日