(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
2024 年 2 月 15 日 (15.02.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/032783 A1

(51) 国际专利分类号:
G06F 9/48 (2006.01)

(21) 国际申请号: PCT/CN2023/112645

(22) 国际申请日: 2023 年 8 月 11 日 (11.08.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202210964273.2 2022年8月11日 (11.08.2022) CN

(71) 申请人:北京有竹居网络技术有限公司(**BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.**) [CN/CN]; 中国北京市平谷区林荫北街13号信息大厦802室, Beijing 101299 (CN)。

(72) 发明人:笪禹(**DA, Yu**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。
李涛(**LI, Tao**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。罗海钊(**LUO, Haizhao**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。张永肃(**ZHANG, Yongsu**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。佘开锐(**SHE, Kairui**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。张宇(**ZHANG, Yu**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。王剑(**WANG, Jian**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。

(74) 代理人: 北京市金杜律师事务所 (**KING & WOOD MALLESONS**); 中国北京市朝阳区东三环中路1号环球金融中心办公楼东楼20层, Beijing 100020 (CN)。

(54) **Title:** TASK SCHEDULING METHOD AND ELECTRONIC DEVICE

(54) 发明名称: 任务调度的方法和电子设备

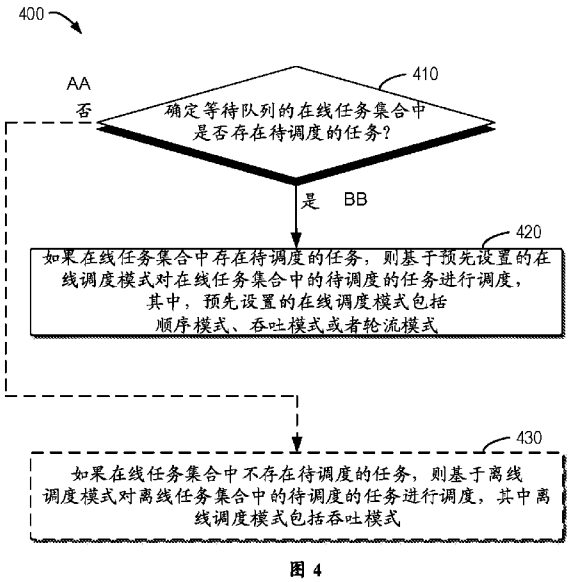

图 4

410 Determine whether a task to be scheduled is present in an online task set of a waiting queue

420 If a task to be scheduled is present in the online task set, schedule, on the basis of a preset online scheduling mode, the task to be scheduled in the online task set, wherein the preset online scheduling mode comprises a sequential mode, a throughput mode or an alternate mode

430 If a task to be scheduled is not present in the online task set, schedule, on the basis of an offline scheduling mode, a task to be scheduled in an offline task set, wherein the offline scheduling mode comprises the throughput mode

AA No

BB Yes

(57) **Abstract:** The embodiments of the present disclosure relate to a task scheduling method and an electronic device. The method comprises: determining whether a task to be scheduled is present in an online task set of a waiting queue; and if a task to be scheduled is present in the online task set, scheduling, on the basis of a preset online scheduling mode, the task to be scheduled in the online task set, wherein the preset online scheduling mode is a sequential mode, a throughput mode or an alternate mode. In this way, by means of the embodiments of the present disclosure, an online task can be preferentially scheduled, the time delay requirement of the online task is ensured, and the online task is scheduled on the basis of a preset online scheduling mode, such that the method can be suitable for different scenarios, thereby ensuring effective utilization of resources.

WO 2024/032783 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, MU, 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, SC, 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))。

(57) 摘要: 本公开的实施例涉及一种任务调度的方法和电子设备。该方法包括：确定等待队列的在线任务集合中是否存在待调度的任务；如果在线任务集合中存在待调度的任务，则基于预先设置的在线调度模式对在线任务集合中的待调度的任务进行调度，其中预先设置的在线调度模式为顺序模式、吞吐模式或者轮流模式。以此方式，本公开的实施例能够优先调度在线任务，保证在线任务的时延需求，并且基于预先设置的在线调度模式来对在线任务进行调度，能够适用于不同的场景，确保资源的有效利用。

任务调度的方法和电子设备

本申请要求于2022年8月11日提交中国国家知识产权局、申请号为202210964273.2、发明名称为“任务调度的方法和电子设备”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

本公开的实施例主要涉及计算机技术领域，并且更具体地，涉及任务调度的方法、装置、电子设备、计算机可读存储介质和计算机程序产品。

背景技术

随着技术的发展，对算力的处理需求和处理能力也在逐渐提升。例如，在人工智能（Artificial Intelligence，AI）领域中的某个AI计算任务可能需要多个处理单元来对数据的不同部分分别进行处理。

在处理大量任务的场景中，如何对任务进行充分且有效地调度，是当前需要解决的问题之一。

发明内容

根据本公开的示例实施例，提供了一种任务调度的技术方案，能够保证在线任务被优先处理，并且能够适应于各种不同的场景，实现了资源的有效利用。

在本公开的第一方面，提供了一种任务调度的方法，包括：确定等待队列的在线任务集合中是否存在待调度的任务；以及如果所述在线任务集合中存在所述待调度的任务，则基于预先设置的在线调度模式对所述在线任务集合中的所述待调度的任务进行调度，其中所述预先设置的在线调度模式包括顺序模式、吞吐模式或者轮流模式。

在本公开的第二方面，提供了一种电子设备，包括：至少一个处理单元；至少一个存储器，至少一个存储器被耦合到至少一个处理单元并且存储用于由至少一个处理单元执行的指令，该指令当由至少一个处理单元执行时使得电子设备执行动作，动作包括：确定等待队列的在线任务集合中是否存在待调度的任务；如果所述在线任务集合中存在所述待调度的任务，则基于预先设置的在线调度模式对所述在线任务集合中的所述待调度的任务进行调度，其中所述预先设置的在线调度模式包括顺序模式、吞吐模式或者轮流模式。

在本公开的第三方面，提供了一种任务调度的装置，包括：确定模块，被配置为确定等待队列的在线任务集合中是否存在待调度的任务；以及调度模块，被配置为：如果所述在线任务集合中存在所述待调度的任务，则基于预先设置的在线调度模式对所述在线任务集合中的所述待调度的任务进行调度，其中所述预先设置的在线调度模式包括顺序模式、吞吐模式或者轮流模式。

在本公开的第四方面，提供了一种计算机可读存储介质，该计算机可读存储介质具有在其上存储的机器可执行指令，该机器可执行指令在由设备执行时使该设备执行根据本公开的第一方面所描述的方法。

本公开的第五方面，提供了一种计算机程序产品，包括计算机可执行指令，其中计

算机可执行指令在被处理器执行时实现根据本公开的第一方面所描述的方法。

本公开的第六方面，提供了一种电子设备，包括：处理电路，被配置为执行根据本公开的第一方面所描述的方法。

提供该内容部分是为了以简化的形式来介绍一系列概念，它们在下文的具体实施方式中将被进一步描述。该内容部分不旨在标识本公开的关键特征或必要特征，也不旨在限制本公开的范围。本公开的其它特征将通过以下的描述变得容易理解。

附图说明

结合附图并参考以下详细说明，本公开各实施例的上述和其他特征、优点及方面将变得更加明显。在附图中，相同或相似的附图标注表示相同或相似的元素，其中：

图 1 示出了本公开实施例能够被应用于其中的场景的示意图；

图 2 示出了根据本公开的一些实施例的系统的示例架构的示意图；

图 3 示出了根据本公开的一些实施例的等待队列的示意图；

图 4 示出了根据本公开的一些实施例的示例过程的流程图；

图 5 示出了根据本公开的一些实施例的示例过程的流程图；

图 6 示出了根据本公开的实施例的示例装置的框图；以及

图 7 示出了可以用来实施本公开的实施例的示例设备的框图。

具体实施方式

下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施例，然而应当理解的是，本公开可以通过各种形式来实现，而且不应该被解释为限于这里阐述的实施例，相反提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是，本公开的附图及实施例仅用于示例性作用，并非用于限制本公开的保护范围。

随着芯片/加速器的处理需求和能力的快速提升，系统（或模块）内部往往存在多个处理引擎。处理引擎一般负责进行具体的计算工作，例如单个计算任务（如 AI 任务）可以使用多个处理引擎来对数据的不同部分分别进行计算。

诸如 AI 等技术在互联网业务中被广泛应用，为了提高硬件资源等的利用率，一种可行的方式是混合部署，也就是说在硬件计算资源上同时部署在线任务和离线任务。在线任务一般负责处理与当前用户请求相关的计算，并且要求响应时间短。离线任务一般属于非用户请求相关的计算，并且不需要进行迅速响应，但是一般计算量较大、占用资源多。在线任务和离线任务混合部署后会使用相同的处理模块，共同使用模块内部的多个处理引擎。但是目前没有有效的方式来对在线任务和离线任务进行调度。

为了至少部分地解决上述技术方案中的缺陷，本公开的实施例提供了一种任务调度的技术方案，通过优先调度在线任务集合中的任务，保证在线任务的时延需求。并且基于预先设置的在线调度模式来对在线任务进行调度，能够适用于不同的场景，能够充分利用处理单元的处理资源，提升任务处理效率。

图 1 示出了本公开的实施例能够被应用于其中的场景 100 的示意图。如图 1 所示，用于对任务进行处理的系统 110 可以包括多个处理引擎，例如图 1 中所示的引擎 1 至引擎 N。该系统 110 能够获取输入数据和命令 102，并利用引擎 1 至引擎 N 进行处理，以得到输出数据 104。

在一些实施例中，输入数据和命令 102 可以包括多个任务，例如包括多个在线任务和/或多个离线任务。

在本公开的实施例中，术语“处理引擎”可以被称为“引擎”、“处理单元”、“处理模块”或其他名称等，本公开对此不限定。作为示例，在下文的描述中，以“处理单元”作为非限制示例进行阐述。相应地，用于进行任务调度和处理的系统可以包括多个处理单元。可理解，多个处理单元可以被实现为以下任一种：多个独立的硬件设备、单个硬件设备内的多个硬件单元（如多个处理器）、多张外围组件互联高速（Peripheral Component Interconnect Express，PCIE）卡、单张 PCIE 卡的多颗芯片、单颗芯片内的多个模块或多个单元等。本公开对此不限定。

本公开的实施例中，术语“接收到的任务”和“到达的任务”在一些场景下可以互换使用。本公开的实施例中，术语“任务”可以被称为子任务、线程或其他名称，本公开对此不限定。在一些示例中，任务可以包括 AI 任务，例如 AI 训练、AI 推理等 AI 计算任务。

图 2 示出了根据本公开的一些实施例的系统 200 的示例架构的示意图。如图 2 所示，系统 200 包括多个处理单元 210 以及调度器 220。多个处理单元 210 包括处理单元 210-1 至处理单元 210-N。调度器 220 包括队列管理器 221、配置管理器 222、等待队列 223、运行队列 224 和完成队列 225。

示例性地，队列管理器 221 可以被配置为将接收到任务添加到等待队列 223 中。队列管理器 221 还可以基于多个处理单元 210 的状态等将等待队列 223 中的部分或全部任务调度到运行队列 224 中。例如，队列管理器 221 可以基于多个处理单元 210 的忙/闲信息以及等待任务所需要的处理单元的数量等，综合确定合适的任务以调度到运行队列 224 中。

在本公开的实施例中，将任务添加到队列中也可以被称为将任务置入队列中等，也可以被称为其他方式，本公开对此不限定。

示例性地，配置管理器 222 可以被配置为从运行队列 224 中提取任务，并处理所提取的任务。例如，配置管理器 222 可以分析所提取的任务对应的指令并配置对应的一个或多个处理单元 210 开始工作。配置管理器 222 还可以将已完成的任务添加到完成队列 225 中，并从运行队列 224 中将已完成的任务删除。

示例性地，队列管理器 221 还可以被配置为将完成队列 225 中的任务进行后处理。例如，后处理可以包括释放处理单元的资源等。

本公开的实施例中，调度器 220 的队列管理器 221 可以将接收到的任务添加到等待队列 223 中。示例性地，队列管理器 221 可以基于任务的属性和/或调度器 220 的调度模式等将任务添加到等待队列 223 中。可选地，调度模式可以包括顺序模式、吞吐模式、轮流模式等，关于调度模式的实施例将在下文中进行详细阐述。

在本公开的一些实施例中，为了满足混合部署的需求，等待队列 223 可以包括在线任务队列集合和离线任务队列集合。图 3 示出了根据本公开的一些实施例的等待队列 223 的示意图。如图 3 所示，等待队列 223 包括在线任务集合 310 和离线任务集合 320。

在线任务集合 310 包括与多个优先级对应的多个在线任务子集。示例性地，在线任务集合 310 包括与优先级 1 至优先级 M1 分别对应的在线任务子集 310-1 至在线任务子集 310-M1，其中 M1 为正整数。可理解，如果 M1=1，则可以被理解为在线任务集合 310

未被划分为与优先级相关的在线任务子集。在一些示例中，可以假设优先级 1 至优先级 M1 表示优先级从低到高，即优先级 M1 具有最高的优先级。应理解，在另一些示例中，也可以假设优先级 1 具有最高的优先级，本公开对此不限定。

在一些实施例中，在线任务子集 310-1 至在线任务子集 310-M1 中的每个在线任务子集可以包括紧急（urgent）任务队列和多个资源匹配任务队列。示例性地，在线任务子集 310-i1 可以对应与优先级 i1（$1 \leqslant i1 \leqslant M1$），并且在线任务子集 310-i1 的紧急任务队列中的任务相对于多个资源匹配任务队列中的任务而言，紧急度更高，需要被优先处理。另外，在线任务子集 310-i1 的多个资源匹配任务队列对应于所需的处理单元的数量。

以在线任务子集 310-1 为例，其包括紧急任务队列 311 和多个资源匹配任务队列 312-1 至 312-N。示例性地，资源匹配任务队列 312-j 表示其中的任务需要的处理单元的数量为 j，其中 $1 \leqslant j \leqslant N$。

离线任务集合 320 包括与多个优先级对应的多个离线任务子集。示例性地，离线任务集合 320 包括与优先级 1 至优先级 M2 分别对应的离线任务子集 320-1 至离线任务子集 320-M2，其中 M2 为正整数。可理解，如果 M2=1，则可以被理解为离线任务集合 320 未被划分为与优先级相关的离线任务子集。在一些示例中，可以假设优先级 1 至优先级 M2 表示优先级从低到高，即优先级 M2 具有最高的优先级。应理解，在另一些示例中，也可以假设优先级 1 具有最高的优先级，本公开对此不限定。另外可理解，M1 和 M2 可以是基于在线任务和离线任务分别被设置的，两者可以相等或不相等，本公开对此不限定。

在一些实施例中，离线任务子集 320-1 至离线任务子集 320-M2 中的每个离线任务子集可以包括紧急任务队列和多个资源匹配任务队列。示例性地，离线任务子集 320-i2 可以对应与优先级 i2（$1 \leqslant i2 \leqslant M2$），并且离线任务子集 320-i2 的紧急任务队列中的任务相对于多个资源匹配任务队列中的任务而言，紧急度更高，需要被优先处理。另外，离线任务子集 320-i2 的多个资源匹配任务队列对应于所需的处理单元的数量。

以离线任务子集 320-1 为例，其包括紧急任务队列 321 和多个资源匹配任务队列 322-1 至 322-N。示例性地，资源匹配任务队列 322-j 表示其中的任务需要的处理单元的数量为 j，其中 $1 \leqslant j \leqslant N$。

以此方式，本公开实施例中的任务被分为三个层级的优先级，在线任务的优先级高于离线任务的优先级。在所有的在线任务（或离线任务）中，不同的任务子集具有不同的优先级。在特定优先级的在线任务子集（或离线任务子集）中，紧急队列中任务的优先级高于资源匹配队列中任务的优先级。

应注意，图 3 所示的等待队列 223 仅是示意，本领域技术人员可以在此基础上进行修改得到其他方案，例如，三个层级中第一层级为在线和离线，第二层级为紧急和多个资源匹配队列，第三层级为不同优先级。本公开对此不限定。

应注意的是，图 3 所示的等待队列 223 仅是本公开的实施例的示例。在实际的任务调度的过程中，等待队列 223 中某些任务子集可能为空，某个队列可能为空，具体取决于实际待处理的任务。在一些示例中，可以取决于调度模式，如下面的实施例中较为详细的阐述。

示例性地，调度器 220 的队列管理器 221 可以持续接收新任务，而不需要等到等待队列 223（或其中的任务子集）为空之后再接收新任务。这样能够确保实时性，避免因

任务接收所造成的时延，如此能够保证对任务的处理效率。

在一些实施例中，调度器 220（如队列管理器 221）可以接收新的任务，并基于该任务的属性将其添加到等待队列 223 的对应位置。示例性地，任务的属性可以包括以下中的一项或多项：在线或离线、优先级、是否紧急、所需的处理单元数量等。

举例而言，对于新到达的任务，可以先确定是在线任务还是离线任务，然后在根据优先级确定对应的在线任务子集或离线任务子集，最后再根据是否紧急和/或所需的处理单元数量将该任务添加到对应的队列中。例如，紧急的任务添加到紧急任务队列中，需要 1 个处理单元的任务添加到与数量 1 对应的资源匹配任务队列中，...，需要 n 个处理单元的任务添加到与数量 n 对应的资源匹配任务队列中，...。

在一些实施例中，调度器 220（如队列管理器 221）可以接收新的任务，并基于调度器 220 的调度模式以及该任务的属性将其添加到等待队列 223 的对应位置。示例性地，调度模式可以包括顺序模式、吞吐模式和轮流模式。任务的属性可以包括以下中的一项或多项：在线或离线、优先级、是否紧急、所需的处理单元数量等。

举例而言，调度器 220 对在线任务的在线调度模式可以为顺序模式、吞吐模式或者轮流模式的至少一种，调度器 220 对离线任务的离线调度模式可以为吞吐模式。

在一些示例中，对于新到达的任务是在线任务，则可以基于调度模式和任务的属性将任务添加到等待队列 223 的对应位置。具体而言，如果调度模式为顺序模式，则基于该在线任务的优先级，将其添加到与优先级对应的在线任务子集的紧急任务队列中。也就是说，在顺序模式时，不使用资源匹配队列，所有的在线任务都在紧急任务队列中，可选地，在顺序模式时该紧急任务队列可以被设置为先进先出。具体而言，如果调度模式为轮流模式，则可以不考虑任务属性中的是否紧急，而基于任务的优先级和所需的处理单元数量，将任务添加到等待队列 223 的对应位置。具体而言，如果调度模式为吞吐模式，则基于任务的属性将其添加到等待队列 223 的对应位置，如前述实施例中的相关描述。

在一些示例中，对于新到达的任务是离线任务，则基于任务的属性将其添加到等待队列 223 的对应位置，如前述实施例中的相关描述。

在本公开的一些实施例中，调度器 220 对在线任务和离线任务的调度模式可以是相同或不同的。在一些示例中，调度器 220 的调度模式可以为吞吐模式，且对在线任务和离线任务的调度时保持不变。在一些示例中，调度器 220 对在线任务的调度模式为顺序模式或者轮流模式，对离线任务的调度模式为吞吐模式，那么调度器 220 在调度在线任务和离线任务之间会对调度模式进行切换。例如，当调度器开始调度离线任务时切换到吞吐模式，随后调度在线任务时再切换回顺序模式或者轮流模式。

以此方式，本公开的实施例中在调度离线任务时使用吞吐模式，由于离线任务对时延几乎没有要求，通过吞吐模式能够最大限度地实现资源的充分利用，如此能够提升资源利用率，进而提高整体的处理效率。

图 4 示出了根据本公开的一些实施例的示例过程 400 的流程图。应当理解，过程 400 还可以包括未示出的附加框和/或可以省略所示出的某些框。本公开的范围在此方面不受限制。示例性地，图 4 所示的过程 400 可以由前述图 2 的系统 200 来执行，例如由调度器 220 来执行。

在框 410，确定等待队列的在线任务集合中是否存在待调度的任务。在框 420，如

果在线任务集合中存在待调度的任务，则基于预先设置的在线调度模式对在线任务集合中的待调度的任务进行调度，其中预先设置的在线调度模式包括顺序模式、吞吐模式或者轮流模式。

以此方式，本公开的实施例能够优先调度在线任务，保证在线任务的时延需求，并且基于预先设置的在线调度模式来对在线任务进行调度，能够适用于不同的场景，确保资源的有效利用。

本公开的实施例中，可以基于任务的实际场景，来设置在线调度模式。这样本公开实施例的系统具有更大的灵活性，能够针对各种不同场景的任务调度，实现各种场景下的资源分配需求。

可选地或附加地，如图 4 所示，在框 430，如果在线任务集合中不存在待调度的任务，则基于离线调度模式对离线任务集合中的待调度的任务进行调度，其中离线调度模式包括吞吐模式。

在一些实施例中，等待队列可以包括在线任务集合和离线任务集合，可以确定在线任务集合是否为空。可理解，在线任务集合为空表示该在线任务集合中不存在待调度（或待处理）的任务。相反，在线任务集合不为空表示该在线任务集合中存在待调度（或待处理）的任务。

在一些实施例中，对于到达的任务（或接收到的任务），队列管理器 221 可以基于任务的属性，或者基于任务的属性和调度模式，将任务添加到等待队列的对应位置。

在一些示例中，如果到达的任务是在线任务并且调度模式是顺序模式，那么可以将该任务添加到在线任务集合的与该任务的优先级对应的紧急任务队列中。可选地，如果调度模式是顺序模式，那么确定到达的任务的优先级，进而确定与该优先级对应的在线任务子集，并将该任务添加到对应的在线任务子集的紧急任务队列中，例如图 3 中所示的紧急任务队列 311。以此，在线任务集合中仅有紧急任务队列被用于任务调度（例如为非空），而其余的资源匹配任务队列都为空，不被用于任务调度。

在一些示例中，如果到达的任务是在线任务并且调度模式是轮流模式，那么可以进一步基于该任务的优先级和任务所需的处理单元的数量，将该任务添加到与该任务的优先级对应的在线任务子集中、与该数量对应的资源匹配任务队列中。可选地，如果调度模式是轮流模式，那么到达的任务仅会被添加到资源匹配任务队列中，例如图 3 中所示的多个资源匹配任务队列 312-1 至 312-N。以此，在线任务集合中仅有资源匹配任务队列被用于任务调度（例如为非空），而其余的紧急任务队列都为空，不被用于任务调度。

在一些示例中，如果到达的任务是在线任务并且调度模式是吞吐模式，那么可以进一步基于任务的优先级、是否紧急、所需的处理单元的数量，将该任务添加到在线任务集合的与其优先级对应的在线任务子集中。具体而言，如果到达的任务是紧急的在线任务，则将该任务添加到与其优先级对应的在线任务子集中的紧急任务队列中。如果到达的任务是非紧急的在线任务，则基于该任务所需的处理单元的数量，将该任务添加到与其优先级对应的在线任务子集中的、与其所需的处理单元的数量对应的资源匹配任务队列中。

在一些示例中，如果到达的任务是离线任务，那么可以进一步基于任务的优先级、是否紧急、所需的处理单元的数量，将该任务添加到离线任务集合的与其优先级对应的离线任务子集中。具体而言，如果到达的任务是紧急的离线任务，则将该任务添加到与

其优先级对应的离线任务子集中的紧急任务队列中。如果到达的任务是非紧急的离线任务，则基于该任务所需的处理单元的数量，将该任务添加到与其优先级对应的离线任务子集中的、与其所需的处理单元的数量对应的资源匹配任务队列中。

在一些实施例中，顺序模式表示对在线任务集合中的待调度的任务按照时间顺序依次进行调度。具体而言，队列管理器 221 可以基于顺序模式从等待队列 223 中选择要被调度的任务，并添加到运行队列 224 中。从而配置管理器 222 能够对运行队列 224 中的任务进行处理。

举例而言，如果预先设置的在线调度模式为顺序模式，那么在等待队列的在线任务集合中的每个任务都具有自己的时间戳，该时间戳可以表示任务到达的时间或者表示任务被添加到等待队列的时间等，本公开对此不限定。示例性地，在存在空闲的处理单元并进一步进行任务调度时，可以基于时间戳来确定顺序。如此，最早到达的在线任务将被最先调度，或者可以理解为是“先进先出”机制。

举例而言，如果预先设置的在线调度模式为顺序模式，那么，可以确定优先权最高的非空在线任务子集，对该优先权最高的非空在线任务子集中的任务按照时间顺序进行调度。

在一些实施例中，轮流模式表示对在线任务集合的多个在线任务子集中的待调度的任务进行轮流调度。具体而言，队列管理器 221 可以基于轮流模式从等待队列 223 中选择要被调度的任务，并添加到运行队列 224 中。从而配置管理器 222 能够对运行队列 224 中的任务进行处理。

举例而言，如果预先设置的在线调度模式为轮流模式，那么在等待队列的在线任务集合中的每个任务都具有自己的时间戳，该时间戳可以表示任务到达的时间或者表示任务被添加到等待队列的时间等，本公开对此不限定。

举例而言，如果预先设置的在线调度模式为轮流模式，那么，可以确定优先权最高的非空在线任务子集，对该优先权最高的非空在线任务子集中的任务按照多个资源匹配任务队列轮流的顺序进行调度。可选地，轮流的顺序可以是对应的数量从大到小，或者从小到大的顺序，本公开对此不限定。

示例性地，在存在空闲的处理单元并进一步进行任务调度时，可以先调度与数量 1 对应的在线任务子集中的一个任务，再调度与数量 2 对应的在线任务子集中的一个任务，...，再调度与数量 N 对应的在线任务子集中的一个任务。或者示例性地，在存在空闲的处理单元并进一步进行任务调度时，可以先调度与数量 N 对应的在线任务子集中的一个任务，再调度与数量 N-1 对应的在线任务子集中的一个任务，...，再调度与数量 1 对应的在线任务子集中的一个任务。如此，能够适用于保证公平的场景，实现对于多个在线任务子集中的任务的公平调度。

在一些实施例中，吞吐模式表示先基于优先级对多个任务子集的紧急任务队列中的待调度的任务进行调度，如果多个紧急任务队列为空，则基于优先级对多个任务子集的资源匹配任务队列中的待调度的任务进行调度。具体而言，队列管理器 221 可以基于吞吐模式从等待队列 223 中选择要被调度的任务，并添加到运行队列 224 中。从而配置管理器 222 能够对运行队列 224 中的任务进行处理。

作为示例，下面描述基于吞吐模式来调度在线任务集合中的任务的实施例。如上所述，在线任务集合包括多个在线任务子集，多个在线任务子集对应于多个不同的优先级，

并且述多个在线任务子集中每个在线任务子集包括紧急任务队列和多个资源匹配任务队列。

可理解，基于吞吐模式对在线任务集合中的任务进行调度可以包括：如果在线任务集合中的紧急任务队列不为空，则基于优先级对多个在线任务子集的紧急任务队列中的任务进行调度；如果在线任务集合中的紧急任务队列为空，则基于优先级对多个在线任务子集的资源匹配任务队列中的任务进行调度。

图 5 示出了根据本公开的一些实施例的对在线任务集合中的任务进行调度的示例过程 500 的流程图。在框 510，对在线任务集合的紧急任务队列中的任务进行调度。在框 520，响应于紧急任务队列中不存在未被调度的任务，基于多个处理单元的状态，对在线任务集合的多个资源匹配任务队列中的任务进行调度。

在一些实施例中，在线任务集合包括具有多个优先级的多个紧急任务队列。示例性地，可以按照多个优先级从高到低的顺序，依次地对多个紧急任务队列中的每个紧急任务队列中的任务依次进行调度。

举例而言，可以确定优先权最高的紧急任务队列是否为空，如果不为空，则对该优先权最高的紧急任务队列中的任务进行调度。如果优先权最高的紧急任务队列为空，则确定优先权次高的紧急任务队列是否为空。如此，便可以按照优先级的顺序对多个紧急任务队列依次进行调度。

举例而言，多个紧急任务队列包括具有第一优先级的第一紧急任务队列，第一紧急任务队列中包括多个紧急任务，那么对多个紧急任务可以按照时间顺序进行调度。如此，针对某个紧急任务队列，可以按照时间顺序进行调度。

在一些实施例中，如果在线任何集合中的所有紧急任务队列都为空，则可以基于优先级对多个任务子集的资源匹配任务队列中的待调度的任务进行调度。具体而言，可以基于与多个优先级以及基于当前空闲的处理单元的数量进行调度。示例性地，可以确定多个处理单元中状态为空闲的处理单元的第一数量；确定多个在线任务子集中的第一目标资源匹配任务队列，其中第一目标资源匹配任务队列为与第一数量对应的资源匹配任务队列中的具有最高优先级的非空队列；并对第一目标资源匹配任务队列中的待调度的任务进行调度。以此方式，在不存在紧急任务的情况下，可以优先调度与第一数量匹配的任务。

示例性地，如果多个在线任务子集中与第一数量对应的资源匹配任务队列都为空，则确定多个在线任务子集中的第二目标资源匹配任务队列和第三目标资源匹配任务队列，其中第二目标资源匹配任务队列为与第二数量对应的资源匹配任务队列中的具有最高优先级的非空队列，第三目标资源匹配任务队列为与第三数量对应的资源匹配任务队列中的具有最高优先级的非空队列，并且第二数量与第三数量之和等于第一数量。随后可以对第二目标资源匹配任务队列中的待调度的任务和第三目标资源匹配任务队列中的待调度的任务进行调度。以此方式，在不存在与第一数量匹配的任务时，可以优先调度具有较大数量需求的任务，这样能够充分地利用空闲的处理单元，例如可以通过组合的方式将空闲的处理单元尽可能地用满，这样能够提升资源利用率。

示例性地，多个在线任务子集中的每个资源匹配任务队列具有对应的权重。可选地，针对多个在线任务子集中的除第一目标资源匹配任务队列之外的其余的每个资源匹配任务队列，将对应的权重增加预设步进值。在一些示例中，如果多个在线任务子集中的第

一资源匹配任务队列的对应的权重达到预设阈值，则将第一资源匹配任务队列中的任务都添加到第一紧急任务队列中，其中第一资源匹配任务队列和第一紧急任务队列属于同一个在线任务子集。以此方式，针对未被调度的资源匹配任务队列中的任务，可以增加权重，从而在多次未被调度之后添加到紧急任务队列，这样能够避免该在线任务长时间未被调度，如此能够保证一定的公平性。

在一些实施例中，如果在线任务集合中不存在未被调度的在线任务，即在线任务集合为空，则可以对离线任务集合中的任务进行调度。具体而言，可以基于吞吐模式来对离线任务集合中的离线任务进行调度。示例性地，关于吞吐模式可以参照前述结合对在线任务的调度方式，为了简化示例，这里不再重复。

在本公开的一些实施例中，对离线任务的调度可以采样慢启动的机制。具体而言，考虑到任务的连续性，即队列管理器 221 可以连续接收新的任务。那么，本公开中在对离线任务进行调度时，可以尝试接收新的在线任务，这样能够保证在线任务的时延要求。

在一些实施例中，如果在线任务集合中为空，此时可以不立即调度离线任务，相反可以先尝试接收新的任务。如果成功接收到新的在线任务，则可以对在线任务进行调度。如果未成功接收到新的在线任务（如接收到离线任务或未接收到任何任务），则可以对离线任务进行调度。

示例性地，可以尝试接收新的在线任务；如果通过第一次尝试未接收到新的在线任务，则对离线任务集合中的第一预定数量的任务进行调度；以及如果通过第二次尝试未接收到新的在线任务，则对离线任务集合中的第二预定数量的任务进行调度，其中第二预定数量大于第一预定数量。

可理解，如果某一次尝试接收到新的在线任务，则对接收到的在线任务进行调度，如此能够保证在线任务的时延要求。如果下一次尝试仍未接收到新的在线任务，则可以增加调度的离线任务的数量，直到预定最大数量（例如 8 或其他值）。

举例而言，如果第一次尝试接收没有在线任务，则可以调度 1 个离线任务。如果第二次尝试接收没有在线任务，则可以调度 2 个离线任务。如果第三次尝试接收没有在线任务，则可以调度 4 个离线任务。如果第四次尝试接收没有在线任务，则可以调度 8 个离线任务。应注意，这里给出的调度的离线任务的数量仅是示意，不能解释为对本公开的实施例的限制。

示例性地，对多个离线任务进行调度时，可以是基于吞吐模式的。举例而言，第一预定数量的任务包括基于优先级顺序而确定的、离线任务集合的紧急任务列表中的待调度任务。如果离线任务集合中的所有紧急任务列表都为空，则基于优先级调度离线任务集合的多个离线任务子集中的资源匹配任务队列中的离线任务。

对离线任务进行调度的离线调度模式为吞吐模式。在一些实施例中，在某次尝试接收到新的在线任务时，可以继续对在线任务进行调度。可选地，如果在线调度模式不是吞吐模式，此时可以将调度模式切换为在线调度模式，例如顺序模式或轮流模式，如此能够实现在不同的调度模式之间的自动切换。

以上结合图 1 至图 5 描述了本公开的实施例的任务调度的方案。本公开的实施例中，等待队列 223 可以被设置为三个层级，如此能够适用于各种不同的场景，例如可以基于场景来设置在线调度模式，从而保证在不同场景下的资源分配需求。

应理解，在本公开的实施例中，“第一”，“第二”，“第三”等只是为了表示多个对

象可能是不同的，但是同时不排除两个对象之间是相同的，不应当解释为对本公开实施例的任何限制。

还应理解，本公开的实施例中的方式、情况、类别以及实施例的划分仅是为了描述的方便，不应构成特别的限定，各种方式、类别、情况以及实施例中的特征在符合逻辑的情况下，可以相互结合。

还应理解，上述内容只是为了帮助本领域技术人员更好地理解本公开的实施例，而不是要限制本公开的实施例的范围。本领域技术人员根据上述内容，可以进行各种修改或变化或组合等。这样的修改、变化或组合后的方案也在本公开的实施例的范围内。

还应理解，上述内容的描述着重于强调各个实施例之前的不同之处，相同或相似之处可以互相参考或借鉴，为了简洁，这里不再赘述。

图 6 示出了根据本公开的一些实施例的示例装置 600 的示意框图。装置 600 可以通过软件、硬件或者两者结合的方式实现。在一些实施例中，装置 600 可以被实现为电子设备。在一些实施例中，装置 600 可以包括如图 2 中所示的调度器 220。

如图 6 所示，装置 600 包括确定模块 610 和调度模块 620。确定模块 610 被配置为确定等待队列的在线任务集合中是否存在待调度的任务。调度模块 620 被配置为：如果在线任务集合中存在待调度的任务，则基于预先设置的在线调度模式对在线任务集合中的待调度的任务进行调度，其中预先设置的在线调度模式为顺序模式、吞吐模式或者轮流模式。

可选地或附加地，调度模块 620 还可以被配置为：如果在线任务集合中不存在待调度的任务，则基于离线调度模式对离线任务集合中的待调度的任务进行调度，其中离线调度模式为吞吐模式。

在一些实施例中，顺序模式表示对在线任务集合中的待调度的任务按照时间顺序依次进行调度。

示例性地，调度模块 620 可以被配置为：如果到达的任务是在线任务并且在线调度模式是顺序模式，则将到达的任务添加到在线任务集合的与到达的任务的优先级对应的紧急任务队列中。可选地，紧急任务队列中的每个任务具有对应的时间戳。

在一些实施例中，轮流模式表示对在线任务集合的多个资源匹配任务队列中的待调度的任务进行轮流调度。

示例性地，调度模块 620 可以被配置为：如果到达的任务是在线任务并且在线调度模式是轮流模式，则基于到达的任务的优先级和所需要的处理单元的数量，将到达的任务添加到与到达的任务的优先级对应的在线任务子集中的、与数量对应的资源匹配任务队列中，其中多个资源匹配任务队列中不同的资源匹配任务队列对应不同的数量。

在一些实施例中，吞吐模式表示：先基于优先级对多个任务子集的紧急任务队列中的待调度的任务进行调度，如果多个紧急任务队列为空，则基于优先级对多个任务子集的资源匹配任务队列中的待调度的任务进行调度。

示例性地，任务子集包括在线任务集合中的多个在线任务子集，多个在线任务子集对应于多个不同的优先级，多个在线任务子集中每个在线任务子集包括紧急任务队列和多个资源匹配任务队列。调度模块 620 可以被配置为：确定多个处理单元中状态为空闲的处理单元的第一数量；确定多个在线任务子集中的第一目标资源匹配任务队列，第一目标资源匹配任务队列为与第一数量对应的资源匹配任务队列中的具有最高优先级的非

空队列；以及对第一目标资源匹配任务队列中的待调度的任务进行调度。

示例性地，调度模块 620 可以被配置为：如果多个在线任务子集中与第一数量对应的资源匹配任务队列都为空，则确定多个在线任务子集中的第二目标资源匹配任务队列和第三目标资源匹配任务队列，第二目标资源匹配任务队列为与第二数量对应的资源匹配任务队列中的具有最高优先级的非空队列，第三目标资源匹配任务队列为与第三数量对应的资源匹配任务队列中的具有最高优先级的非空队列，第二数量与第三数量之和等于第一数量；以及对第二目标资源匹配任务队列中的待调度的任务和第三目标资源匹配任务队列中的待调度的任务进行调度。

示例性地，多个在线任务子集中的每个资源匹配任务队列具有对应的权重，调度模块 620 还可以被配置为：针对多个在线任务子集中的除第一目标资源匹配任务队列之外的其余的每个资源匹配任务队列，将对应的权重增加预设步进值。

示例性地，调度模块 620 可以被配置为：如果多个在线任务子集中的第一资源匹配任务队列的对应的权重达到预设阈值，则将第一资源匹配任务队列中的任务都添加到第一紧急任务队列中，其中第一资源匹配任务队列和第一紧急任务队列属于同一个在线任务子集。

示例性地，调度模块 620 可以被配置为：如果到达的任务是离线任务，则基于到达的任务的优先级，将到达的任务添加到离线任务集合中与优先级对应的离线任务子集中。

示例性地，调度模块 620 可以被配置为：如果到达的任务为紧急的离线任务，则将到达的任务添加到与优先级对应的离线任务子集中的紧急任务队列中；如果到达的任务为非紧急的离线任务，则基于到达的任务所需处理单元的数量，将到达的任务添加到与优先级对应的离线任务子集中的与数量对应的资源匹配任务队列中。

在一些实施例中，调度模块 620 可以被配置为：尝试接收新的在线任务；如果通过第一次尝试未接收到新的在线任务，则对离线任务集合中的第一预定数量的任务进行调度；以及如果通过第二次尝试未接收到新的在线任务，则对离线任务集合中的第二预定数量的任务进行调度，其中第二预定数量大于第一预定数量。

示例性地，第一预定数量的任务包括基于优先级顺序而确定的、离线任务集合的紧急任务列表中的待调度任务。

图 6 的装置 600 能够用于实现上述结合图 4 至图 5 所述的过程，为了简洁，这里不再赘述。

本公开的实施例中对模块或单元的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时也可以有另外的划分方式，另外，在公开的实施例中的各功能单元可以集成在一个单元中，也可以是单独物理存在，也可以两个或两个以上单元集成为一个单元中。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。

图 7 示出了可以用来实施本公开的实施例的示例设备 700 的框图。应当理解，图 7 所示出的设备 700 仅仅是示例性的，而不应当构成对本文所描述的实现方式的功能和范围的任何限制。例如，可以使用设备 700 来执行上文描述的过程 400 和/或过程 500。

如图 7 所示，设备 700 是通用计算设备的形式。计算设备 700 的组件可以包括但不限于一个或多个处理器或处理单元 710、存储器 720、存储设备 730、一个或多个通信单元 740、一个或多个输入设备 750 以及一个或多个输出设备 760。处理单元 710 可以是实际或虚拟处理器并且能够根据存储器 720 中存储的程序来执行各种处理。在多处理器系

统中，多个处理单元并行执行计算机可执行指令，以提高计算设备 700 的并行处理能力。

计算设备 700 通常包括多个计算机存储介质。这样的介质可以是计算设备 700 可访问的任何可以获得的介质，包括但不限于易失性和非易失性介质、可拆卸和不可拆卸介质。存储器 720 可以是易失性存储器（例如寄存器、高速缓存、随机访问存储器（Random Access Memory，RAM））、非易失性存储器（例如，只读存储器（Read Only Memory，ROM）、电可擦除可编程只读存储器（Electrically Erasable Programmable Read Only Memory，EEPROM）、闪存）或它们的某种组合。存储设备 730 可以是可拆卸或不可拆卸的介质，并且可以包括机器可读介质，诸如闪存驱动、磁盘或者任何其他介质，其可以能够用于存储信息和/或数据（例如用于训练的训练数据）并且可以在计算设备 700 内被访问。

计算设备 700 可以进一步包括另外的可拆卸/不可拆卸、易失性/非易失性存储介质。尽管未在图 7 中示出，可以提供用于从可拆卸、非易失性磁盘（例如“软盘”）进行读取或写入的磁盘驱动和用于从可拆卸、非易失性光盘进行读取或写入的光盘驱动。在这些情况中，每个驱动可以由一个或多个数据介质接口被连接至总线（未示出）。存储器 720 可以包括计算机程序产品 725，其具有一个或多个程序模块，这些程序模块被配置为执行本公开的各种实现方式的各种方法或动作。

通信单元 740 实现通过通信介质与其他计算设备进行通信。附加地，计算设备 700 的组件的功能可以以单个计算集群或多个计算机器来实现，这些计算机器能够通过通信连接进行通信。因此，计算设备 700 可以使用与一个或多个其他服务器、网络个人计算机（Personal Computer，PC）或者另一个网络节点的逻辑连接来在联网环境中进行操作。

输入设备 750 可以是一个或多个输入设备，例如鼠标、键盘、追踪球等。输出设备 760 可以是一个或多个输出设备，例如显示器、扬声器、打印机等。计算设备 700 还可以根据需要通过通信单元 740 与一个或多个外部设备（未示出）进行通信，外部设备诸如存储设备、显示设备等，与一个或多个使得用户与计算设备 700 交互的设备进行通信，或者与使得计算设备 700 与一个或多个其他计算设备通信的任何设备（例如，网卡、调制解调器等）进行通信。这样的通信可以经由输入/输出（Input/Output，I/O）接口（未示出）来执行。

根据本公开的示例性实现方式，提供了一种计算机可读存储介质，其上存储有计算机可执行指令，其中计算机可执行指令被处理器执行以实现上文描述的方法。根据本公开的示例性实现方式，还提供了一种计算机程序产品，计算机程序产品被有形地存储在非瞬态计算机可读介质上并且包括计算机可执行指令，而计算机可执行指令被处理器执行以实现上文描述的方法。根据本公开的示例性实现方式，提供了一种计算机程序产品，其上存储有计算机程序，所述程序被处理器执行时实现上文描述的方法。

这里参照根据本公开实现的方法、装置、设备和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理单元，从而生产出一种机器，使得这些指令在通过计算机或其他可编程数据处理装置的处理单元执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，

这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上，使得在计算机、其他可编程数据处理装置或其他设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其他可编程数据处理装置、或其他设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本公开的多个实现的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

以上已经描述了本公开的各实现，上述说明是示例性的，并非穷尽性的，并且也不限于所公开的各实现。在不偏离所说明的各实现的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实现的原理、实际应用或对市场中的技术的改进，或者使本技术领域的其他普通技术人员能理解本文公开的各个实现方式。

权利要求书

1. 一种任务调度的方法，包括:

确定等待队列的在线任务集合中是否存在待调度的任务；以及

如果所述在线任务集合中存在所述待调度的任务，则基于预先设置的在线调度模式对所述在线任务集合中的所述待调度的任务进行调度，其中所述预先设置的在线调度模式包括顺序模式、吞吐模式或者轮流模式。

2. 根据权利要求 1 所述的方法，其中所述顺序模式表示对所述在线任务集合中的所述待调度的任务按照时间顺序依次进行调度。

3. 根据权利要求 2 所述的方法，还包括:

如果到达的任务是在线任务并且所述在线调度模式是所述顺序模式，则将所述到达的任务添加到所述在线任务集合的与所述到达的任务的优先级对应的紧急任务队列中。

4. 根据权利要求 3 所述的方法，其中所述紧急任务队列中的每个任务具有对应的时间戳。

5. 根据权利要求 1 所述的方法，其中所述轮流模式表示对所述在线任务集合的多个资源匹配任务队列中的所述待调度的任务进行轮流调度。

6. 根据权利要求 5 所述的方法，所述方法还包括:

如果到达的任务是在线任务并且所述在线调度模式是所述轮流模式，则基于所述到达的任务的优先级和所需要的处理单元的数量，将所述到达的任务添加到与所述到达的任务的优先级对应的在线任务子集中的、与所述所需要的处理单元的数量对应的资源匹配任务队列中，其中所述多个资源匹配任务队列中不同的资源匹配任务队列对应不同的数量。

7. 根据权利要求 1 所述的方法，其中所述吞吐模式表示：先基于优先级对多个任务子集的紧急任务队列中的待调度的任务进行调度，如果所述多个紧急任务队列为空，则基于优先级对所述多个任务子集的资源匹配任务队列中的待调度的任务进行调度。

8. 根据权利要求 7 所述的方法，其中所述任务子集包括所述在线任务集合中的多个在线任务子集，所述多个在线任务子集对应于多个不同的优先级，所述多个在线任务子集中每个在线任务子集包括紧急任务队列和多个资源匹配任务队列，

并且其中基于优先级对所述多个任务子集的资源匹配任务队列中的待调度的任务进行调度包括:

确定多个处理单元中状态为空闲的处理单元的第一数量;

确定所述多个在线任务子集中的第一目标资源匹配任务队列，所述第一目标资源匹配任务队列为与所述第一数量对应的资源匹配任务队列中的具有最高优先级的非空队列；以及

对所述第一目标资源匹配任务队列中的待调度的任务进行调度。

9. 根据权利要求 8 所述的方法，还包括:

如果所述多个在线任务子集中与所述第一数量对应的资源匹配任务队列都为空，则确定所述多个在线任务子集中的第二目标资源匹配任务队列和第三目标资源匹配任务队列，所述第二目标资源匹配任务队列为与第二数量对应的资源匹配任务队列中的具有最高优先级的非空队列，所述第三目标资源匹配任务队列为与第三数量对应的资源匹配任务队列中的具有最高优先级的非空队列，所述第二数量与所述第三数量之和等于所述第一数量；以及

对所述第二目标资源匹配任务队列中的待调度的任务和所述第三目标资源匹配任务队列中的待调度的任务进行调度。

10. 根据权利要求 8 所述的方法，其中所述多个在线任务子集中的每个资源匹配任务队列具有对应的权重，所述方法还包括：

针对所述多个在线任务子集中的除所述第一目标资源匹配任务队列之外的其余的每个资源匹配任务队列，将对应的权重增加预设步进值。

11. 根据权利要求 10 所述的方法，还包括：

如果所述多个在线任务子集中的第一资源匹配任务队列的对应的权重达到预设阈值，则将所述第一资源匹配任务队列中的任务都添加到第一紧急任务队列中，其中所述第一资源匹配任务队列和所述第一紧急任务队列属于同一个在线任务子集。

12. 根据权利要求 1 所述的方法，还包括：

如果所述在线任务集合中不存在待调度的任务，则基于离线调度模式对离线任务集合中的待调度的任务进行调度，其中所述离线调度模式包括所述吞吐模式。

13. 根据权利要求 12 所述的方法，其中对离线任务集合中的待调度的任务进行调度包括：

尝试接收新的在线任务；

如果通过第一次尝试未接收到新的在线任务，则对所述离线任务集合中的第一预定数量的任务进行调度；以及

如果通过第二次尝试未接收到新的在线任务，则对所述离线任务集合中的第二预定数量的任务进行调度，其中所述第二预定数量大于所述第一预定数量。

14. 根据权利要求 13 所述的方法，其中所述第一预定数量的任务包括基于优先级顺序而确定的、所述离线任务集合的紧急任务列表中的待调度任务。

15. 根据权利要求 12 所述的方法，还包括：

如果到达的任务是离线任务，则基于所述到达的任务的优先级，将所述到达的任务添加到所述离线任务集合中与所述优先级对应的离线任务子集中。

16. 根据权利要求 15 所述的方法，还包括：

如果所述到达的任务为紧急的离线任务，则将所述到达的任务添加到与所述优先级对应的所述离线任务子集中的紧急任务队列中；

如果所述到达的任务为非紧急的离线任务，则基于所述到达的任务所需处理单元的数量，将所述到达的任务添加到与所述优先级对应的所述离线任务子集中的与所述所需处理单元的数量对应的资源匹配任务队列中。

17. 一种电子设备，包括：

至少一个处理单元；

至少一个存储器，所述至少一个存储器被耦合到所述至少一个处理单元并且存储用于由所述至少一个处理单元执行的指令，所述指令当由所述至少一个处理单元执行时使得所述电子设备执行动作，所述动作包括：

确定等待队列的在线任务集合中是否存在待调度的任务；以及

如果所述在线任务集合中存在所述待调度的任务，则基于预先设置的在线调度模式对所述在线任务集合中的所述待调度的任务进行调度，其中所述预先设置的在线调度模式包括顺序模式、吞吐模式或者轮流模式。

18. 一种任务调度的装置，包括：

确定模块，被配置为确定等待队列的在线任务集合中是否存在待调度的任务；以及

调度模块，被配置为：如果所述在线任务集合中存在所述待调度的任务，则基于预先设

置的在线调度模式对所述在线任务集合中的所述待调度的任务进行调度，其中所述预先设置的在线调度模式包括顺序模式、吞吐模式或者轮流模式。

19. 一种计算机可读存储介质，其上存储有计算机程序，所述程序被处理器执行时实现根据权利要求 1 至 16 中任一项所述的方法。

20. 一种计算机程序产品，其上存储有计算机程序，所述程序被处理器执行时实现根据权利要求 1 至 16 中任一项所述的方法。

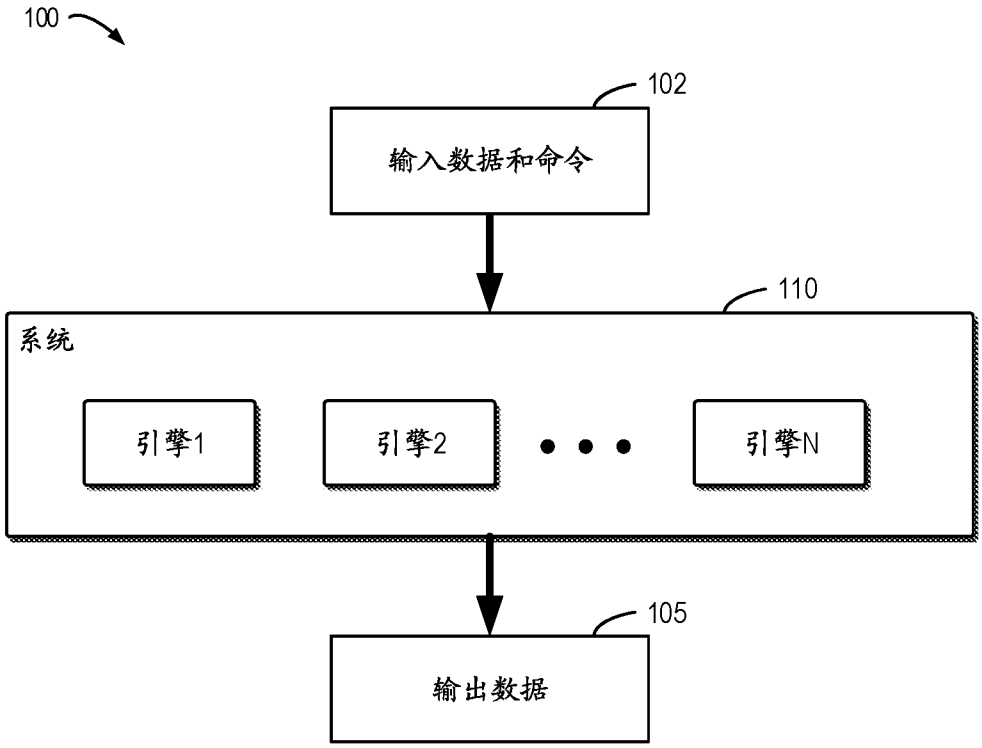

图 1

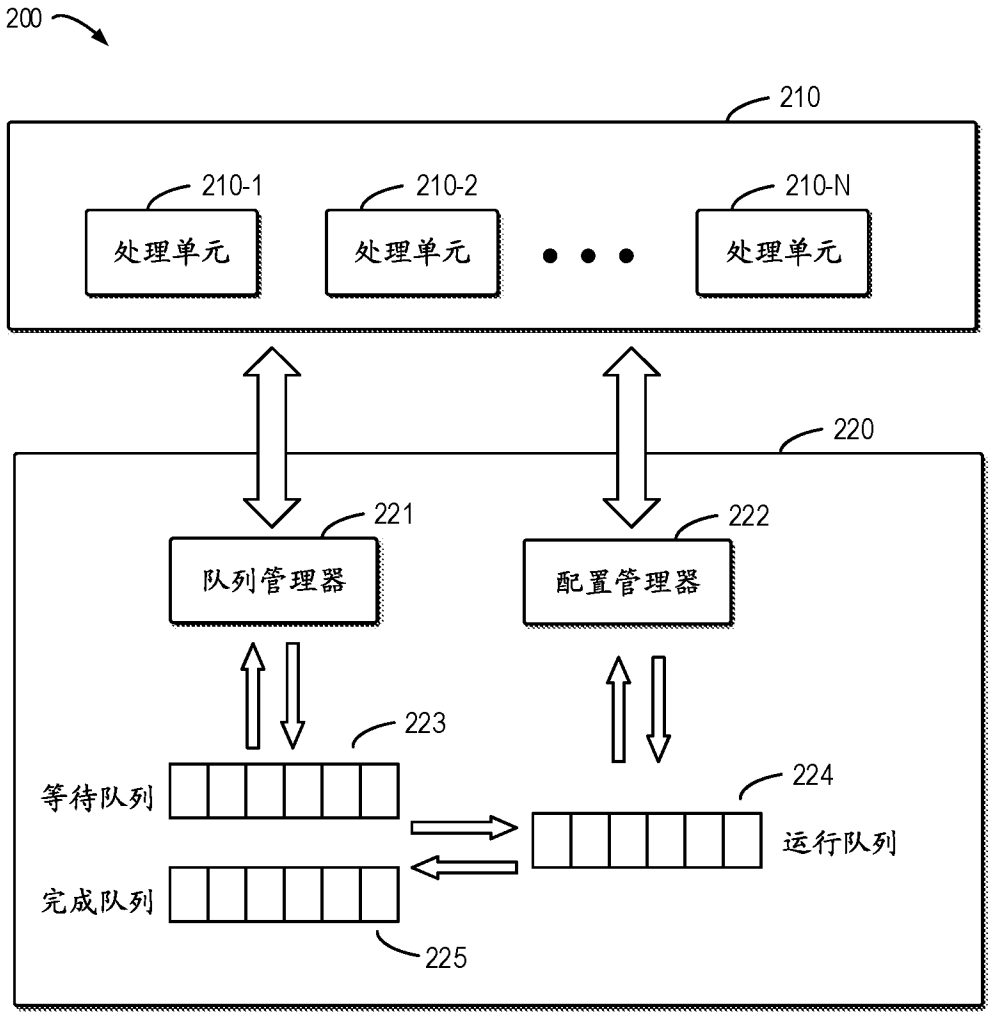
200
210
210-1
处理单元
210-2
处理单元
210-N
处理单元
220
221
队列管理器
222
配置管理器
223
等待队列
224
运行队列
完成队列
225

图 2

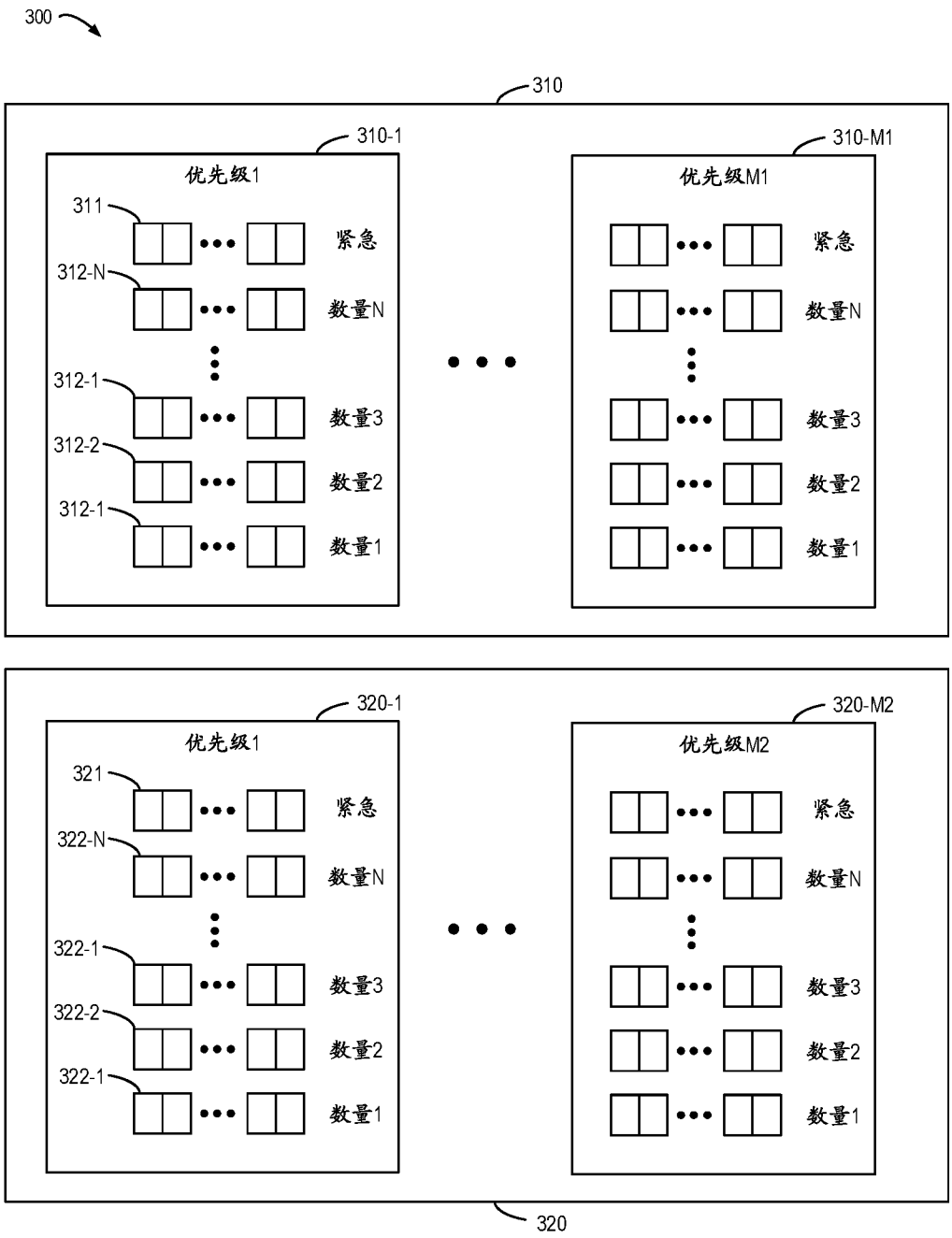
300
310
310-1
优先级1
311
紧急
312-N
数量N
312-1
数量3
312-2
数量2
312-1
数量1
310-M1
优先级M1
紧急
数量N
数量3
数量2
数量1
320-1
优先级1
321
紧急
322-N
数量N
322-1
数量3
322-2
数量2
322-1
数量1
320-M2
优先级M2
紧急
数量N
数量3
数量2
数量1
320

图 3

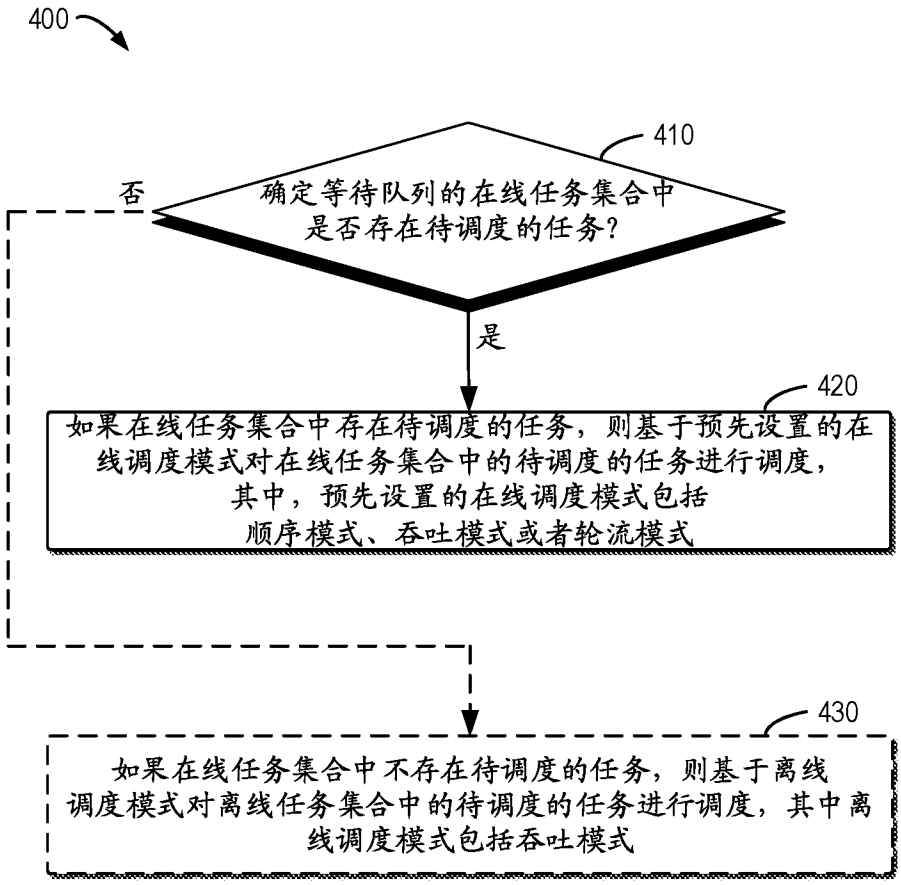
400
410
确定等待队列的在线任务集合中
是否存在待调度的任务?
否
是
420
如果在线任务集合中存在待调度的任务，则基于预先设置的在
线调度模式对在线任务集合中的待调度的任务进行调度，
其中，预先设置的在线调度模式包括
顺序模式、吞吐模式或者轮流模式
430
如果在线任务集合中不存在待调度的任务，则基于离线
调度模式对离线任务集合中的待调度的任务进行调度，其中离
线调度模式包括吞吐模式

图 4

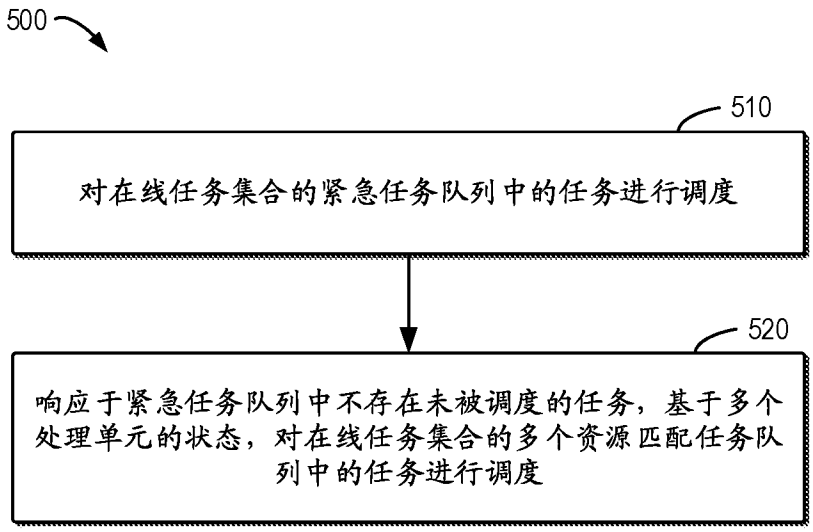
500
510
对在线任务集合的紧急任务队列中的任务进行调度
520
响应于紧急任务队列中不存在未被调度的任务，基于多个处理单元的状态，对在线任务集合的多个资源匹配任务队列中的任务进行调度

图 5

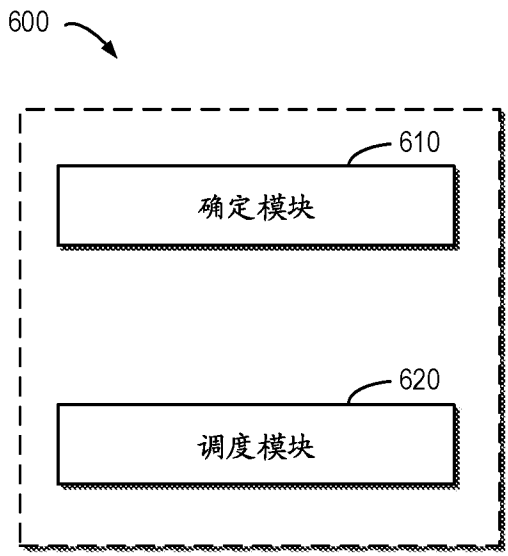
600
610
确定模块
620
调度模块

图 6

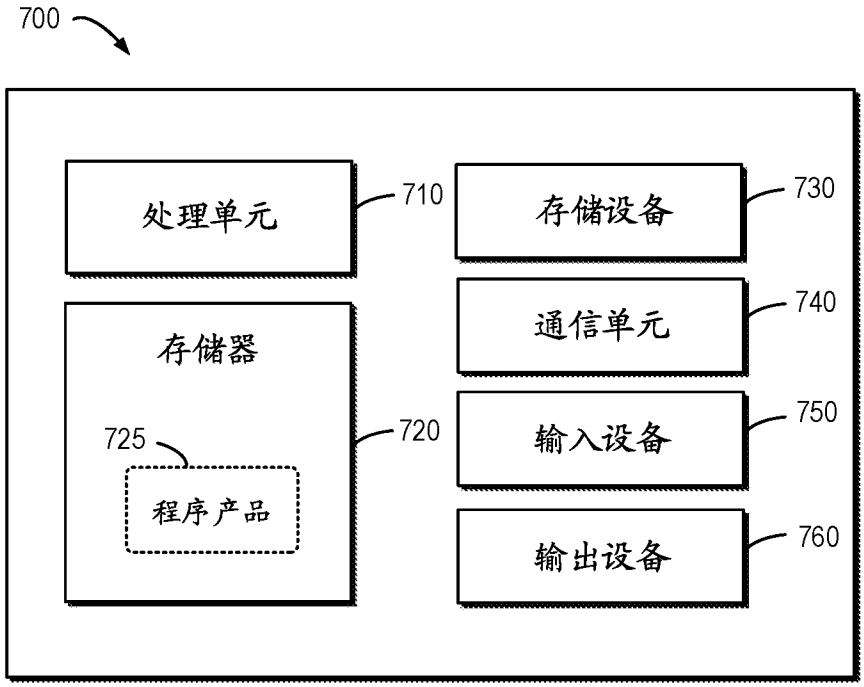
700
处理单元
710
存储设备
730
通信单元
740
存储器
720
725
程序产品
输入设备
750
输出设备
760

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/112645

A. CLASSIFICATION OF SUBJECT MATTER

G06F9/48(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

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, ENTXT, VEN, CNKI, IEEE: 任务, 调度, 在线, 离线, 顺序, 轮流, 吞吐, task, schedule, online, offline, sequence, turn, throughput

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 107977268 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 01 May 2018 (2018-05-01) claims 1-14, and description, paragraphs 62-101	1-20
A	CN 112130963 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 25 December 2020 (2020-12-25) entire document	1-20
A	CN 113282381 A (ZHONGKE CAMBRICON TECHNOLOGY CO., LTD.) 20 August 2021 (2021-08-20) entire document	1-20
A	CN 114579279 A (ALIBABA (CHINA) CO., LTD.) 03 June 2022 (2022-06-03) entire document	1-20

☐ 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
13 October 2023	**23 October 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**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/112645

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 107977268 A	01 May 2018	US 2019114202 A1	18 April 2019
CN 112130963 A	25 December 2020	HK 40035776 A0	21 May 2021
CN 113282381 A	20 August 2021	None	
CN 114579279 A	03 June 2022	None	

国际检索报告

国际申请号
PCT/CN2023/112645

A. 主题的分类

G06F9/48(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

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, ENTXT, VEN, CNKI, IEEE: 任务，调度，在线，离线，顺序，轮流，吞吐，task, schedule, online, offline, sequence, turn, throughput

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 107977268 A (北京百度网讯科技有限公司) 2018年5月1日 (2018 - 05 - 01) 权利要求1-14，说明书第62-101段	1-20
A	CN 112130963 A (腾讯科技（深圳）有限公司) 2020年12月25日 (2020 - 12 - 25) 全文	1-20
A	CN 113282381 A (中科寒武纪科技股份有限公司) 2021年8月20日 (2021 - 08 - 20) 全文	1-20
A	CN 114579279 A (阿里巴巴（中国）有限公司) 2022年6月3日 (2022 - 06 - 03) 全文	1-20

☐ 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10月13日	2023年10月23日

ISA/CN的名称和邮寄地址	受权官员
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	林婉娟 电话号码 (+86) 010-53961343

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/112645

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	107977268	A	2018年5月1日	US	2019114202	A1	2019年4月18日
CN	112130963	A	2020年12月25日	HK	40035776	A0	2021年5月21日
CN	113282381	A	2021年8月20日		无		
CN	114579279	A	2022年6月3日		无		