(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 8 月 3 日 (03.08.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/142091 A1

(51) 国际专利分类号:
H04W 72/04 (2009.01)

(21) 国际申请号: PCT/CN2022/075123

(22) 国际申请日: 2022 年 1 月 29 日 (29.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(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: 张 龙 (**ZHANG, Long**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 郑 明 (**ZHENG, Ming**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 何世明 (**HE, Shiming**); 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: 北 京 龙 双 利 达 知 识 产 权 代 理 有 限 公 司 (**LONGSUN LEAD IP LTD.**); 中国北京市海淀区北清路 81 号院二区 3 号楼 8 层 801-1室, Beijing 100094 (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,

(54) **Title:** COMPUTING TASK SCHEDULING APPARATUS, COMPUTING APPARATUS, COMPUTING TASK SCHEDULING METHOD AND COMPUTING METHOD

(54) 发明名称: 计算任务调度装置、计算装置、计算任务调度方法和计算方法

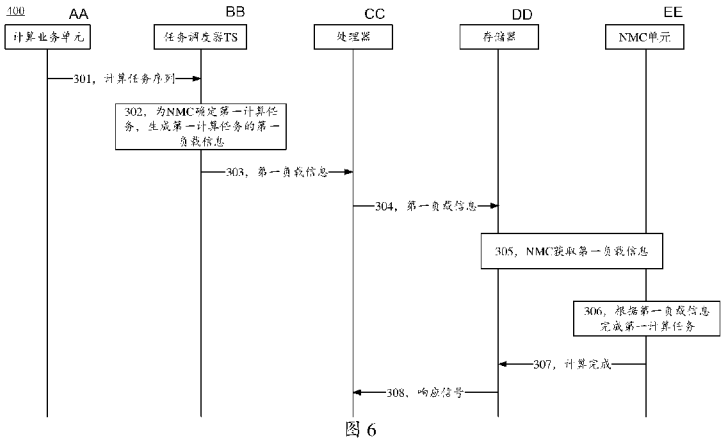

图 6

301 Computing task sequence
302 Determine a first computing task for an NMC, and generate first load information of the first computing task
303, 304 First load information
305 The NMC acquires the first load information
306 Complete the first computing task according to the first load information
307 Complete computing
308 Response signal
AA Computing service unit
BB Task scheduler (TS)
CC Processor
DD Memory
EE NMC unit

(57) **Abstract:** Provided in the present application are a computing task scheduling apparatus, a computing apparatus, a computing task scheduling method and a computing method. The computing task scheduling apparatus comprises: a task scheduler, which is used for determining a first computing task for a first computing unit, and generating load information of the first computing task, which load information is used for defining the first computing task; and a processor, which is used for receiving the load information from the

WO 2023/142091 A1

[见续页]

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, 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))。

task scheduler, and storing the load information in a first address in a memory, so as to allocate the first computing task to the first computing unit, wherein the first address is a reserved address of the first computing unit. The processor is coupled to at least one of the memory and the first computing unit by means of a bus; the first computing unit is tightly coupled to the memory, and the tight coupling does not require any bus; and the first computing unit can access the memory at a speed higher than that of accessing the memory via the bus. The technical solution can reduce the overheads of data transmission, and reduces the complexity of computing.

(57) 摘要：本申请提供了一种计算任务调度装置、计算装置、计算任务调度方法和计算方法，该计算任务调度装置，包括：任务调度器，用于为第一计算单元确定第一计算任务，生成第一计算任务的负载信息，负载信息用于定义第一计算任务；处理器，用于从任务调度器接收负载信息，将负载信息存储至存储器中的第一地址以将第一计算任务分配给第一计算单元，第一地址为第一计算单元的预留地址，其中，处理器与存储器和第一计算单元中的至少一个通过总线耦合，第一计算单元与存储器紧密耦合，紧密耦合无需经过任何总线，且第一计算单元能够以高于总线接入的速度接入存储器。该技术方案能够降低数据传输的开销，降低计算的复杂度。

计算任务调度装置、计算装置、计算任务调度方法和计算方法

技术领域

本申请涉及计算机技术领域，尤其涉及一种计算任务调度装置、计算装置、计算任务调度方法和计算方法。

背景技术

在通用计算机系统中，冯诺依曼或者哈佛架构为计算与存储分离架构。计算需要的数据需从外存加载到计算核内存，即缓存内，计算完成需要从核内存回到外存，使得计算过程中的数据传输功耗增加。

为了降低数据传输的功耗，可以采用计算融合技术、近存计算或存内计算技术。计算融合技术将多步计算通过融合的方式进行计算从而可以减少与外存的交互，但计算融合需要计算核内有一定大小的高速缓存且需要精细的软件切分管理，实现复杂度较高。近存计算在存储器附近完成计算，存内计算在存储器内部直接计算，从而可以降低数据传输的功耗，但近存计算或存内计算一般需要新增相应的计算指令，且针对不同的硬件平台，适配方案不统一，从而复杂度较高。

因此，如何在保证降低数据传输的功耗的同时，降低计算的复杂度，成为需要解决的技术问题。

发明内容

本申请提供一种存储计算装置和存储计算方法，以期降低数据传输的开销，降低计算的复杂度。

第一方面，提供了一种计算任务调度装置，包括：任务调度器，用于为第一计算单元确定第一计算任务，生成所述第一计算任务的负载信息，所述负载信息用于定义所述第一计算任务；处理器，用于从所述任务调度器接收所述负载信息，将所述负载信息存储至存储器中的第一地址以将所述第一计算任务分配给所述第一计算单元，所述第一地址为所述第一计算单元的预留地址，其中，所述处理器与所述存储器和所述第一计算单元中的至少一个通过总线耦合，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线接入的速度接入所述存储器。

该技术方案中，任务调度器可以为第一计算单元确定计算任务，从而软件模块可以无需感知该第一计算单元的计算能力，从而可以降低软件复杂度，此外，处理器采用访问存储器的方式将计算任务的负载信息存储至存储器中的固定地址中，以将计算任务分配给第一计算单元，第一计算单元与存储器紧密耦合，从而可以快速完成任务计算，无需在处理器与第一计算单元之间通过特定总线或接口传输任务调度信息，降低了数据传输的功耗与时延。

结合第一方面，在第一方面的一种实现方式中，所述处理器具体用于：通过直接存储访问 DMA 控制器将所述负载信息存储至所述第一地址。

该技术方案中，处理器通过复用现有的 DMA 技术实现上述有益效果，通过该 DMA 将负载信息存储器中的第一地址中，以将该第一计算任务分配给第一计算单元，从而可以节省系统开销，提升计算效率。

结合第一方面，在第一方面的一种实现方式中，所述任务调度器为系统软件或应用软件之外的专用任务调度器。

应理解，该任务调度器可以为专用于任务调度的硬件任务调度器。

结合第一方面，在第一方面的一种实现方式中，所述任务调度器还用于：接收来自所述系统软件或所述应用软件的计算任务序列，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务。

应理解，该计算任务序列中可以包括一个或多个计算任务，该第一计算任务可以是一个计算任务，也可以是多个计算任务，本申请实施例对此不予限定。

结合第一方面，在第一方面的一种实现方式中，所述任务调度器还用于：在所述计算任务序列中为第二计算单元确定第二计算任务；将所述第二计算任务调度至第二计算单元；其中，所述第二计算单元包括所述处理器、图像处理单元、人工智能 AI 处理单元、数字信号处理器或专用逻辑电路中的至少一个，所述第二计算单元和所述存储器通过总线耦合。

该技术方案中，当任务调度器在计算任务序列中确定该第二计算任务不适合第一计算单元完成计算时，可以将该第二计算任务调度至第二计算单元中，该第二计算单元与存储器通过总线耦合。例如，该第二计算任务可以是不适合进行近存计算、存内计算或存算一体的其他任务。

结合第一方面，在第一方面的一种实现方式中，所述任务调度器具体用于：根据计算列表，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务，其中，所述计算列表包括所述第一计算单元支持的计算任务类型。

应理解，该计算列表可以用链表等进行代替。

在一些实施例中，该计算列表可以进行更新。例如，当第一计算单元支持的计算类型发生改变时，可以将改变后的计算类型加入计算列表中，以完成对计算列表的更新。或者，当该任务调度器和处理器应用到其他的计算单元时，可以将该计算单元支持的计算类型加入该计算列表中。从而可以提升系统的兼容性。

结合第一方面，在第一方面的一种实现方式中，所述负载信息包括如下信息中的至少一种：数据地址；数据维度；或控制命令字。

应理解，该负载信息还可以包括用于计算任务的其他信息等。

结合第一方面，在第一方面的一种实现方式中，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

第二方面，提供了一种计算装置，包括：存储器；第一计算单元，用于从所述存储器中的第一地址获取负载信息，并根据所述负载信息完成第一计算任务，其中，所述负载信息用于定义所述第一计算任务，所述第一地址为所述第一计算单元的预留地址；其中，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线接入的速度接入所述存储器；所述存储器和所述第一计算单元中的至少一个通过总线耦合至处理器。

该技术方案中，第一计算单元从存储器中获取负载信息，且第一计算单元与存储器紧密耦合，从而可以降低计算所需的系统开销，提升计算效率。

结合第二方面，在第二方面的一种实现方式中，所述负载信息包括如下信息中的至少一种：数据地址；数据维度；或控制命令字。

结合第二方面，在第二方面的一种实现方式中，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

结合第二方面，在第二方面的一种实现方式中，所述存储器具体用于：在直接存储访问 DMA 控制器的操作下在所述第一地址写入所述负载信息。

该技术方案中，存储器可以通过 DMA，在第一地址中写入负载信息。从而可以节省总线开销。

第三方面，提供一种计算任务调度方法，包括：任务调度器为第一计算单元确定第一计算任务，生成所述第一计算任务的负载信息，所述负载信息用于定义所述第一计算任务；处理器从所述任务调度器接收所述负载信息，将所述负载信息存储至存储器中的第一地址以将所述第一计算任务分配给所述第一计算单元，所述第一地址为所述第一计算单元的预留地址，其中，所述处理器与所述存储器和所述第一计算单元中的至少一个通过总线耦合，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线接入的速度接入所述存储器。

结合第三方面，在第三方面的一种实现方式中，所述将所述负载信息存储至存储器中的第一地址以将所述第一计算任务分配给所述第一计算单元，包括：通过直接存储访问 DMA 控制器将所述负载信息存储至所述第一地址以将所述第一计算任务分配给所述第一计算单元。

结合第三方面，在第三方面的一种实现方式中，所述任务调度器为系统软件或应用软件之外的专用任务调度器。

结合第三方面，在第三方面的一种实现方式中，所述任务调度器为第一计算单元确定第一计算任务，包括：所述任务调度器接收来自所述系统软件或所述应用软件的计算任务序列，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务。

结合第三方面，在第三方面的一种实现方式中，所述方法还包括：在所述计算任务序列中为第二计算单元确定第二计算任务；将所述第二计算任务调度至第二计算单元；其中，所述第二计算单元包括所述处理器、图像处理单元、人工智能 AI 处理单元、数字信号处理器或专用逻辑电路中的至少一个，所述第二计算单元和所述存储器通过总线耦合。

结合第三方面，在第三方面的一种实现方式中，所述在所述计算任务序列中为所述第一计算单元确定所述第一计算任务，包括：根据计算列表，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务，其中，所述计算列表包括所述第一计算单元支持的计算任务类型。

结合第三方面，在第三方面的一种实现方式中，所述负载信息包括如下信息中的至少一种：数据地址；数据维度；或控制命令字。

结合第三方面，在第三方面的一种实现方式中，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

第四方面，提供一种计算方法，包括：第一计算单元从存储器中的第一地址获取负载信息，并根据所述负载信息完成第一计算任务，其中，所述负载信息用于定义所述第一计算任务，所述第一地址为所述第一计算单元的预留地址；其中，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线

接入的速度接入所述存储器；所述存储器和所述第一计算单元中的至少一个通过总线耦合至处理器。

结合第四方面，在第四方面的一种实现方式中，所述负载信息包括如下信息中的至少一种：数据地址；数据维度；或控制命令字。

结合第四方面，在第四方面的一种实现方式中，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

结合第四方面，在第四方面的一种实现方式中，所述方法还包括：所述存储器在直接存储访问 DMA 控制器的操作下在所述第一地址写入所述负载信息。

第五方面，提供一种计算机可读存储介质，包括：所述存储介质中存储有计算机程序或指令，当所述计算机程序或指令被通信装置执行时，使得如第三方面及其任一种可能的实现方式中所述的计算任务调度方法被执行，或者，使得如第四方面及其任一种可能的实现方式中所述的计算方法被执行。

第六方面，提供一种计算机程序产品，当所述计算机程序产品在计算机上运行时，使得如第三方面及其任一种可能的实现方式中所述的计算任务调度方法被执行，或者，使得如第四方面及其任一种可能的实现方式中所述的计算方法被执行。

第七方面，提供一种计算系统，包括如第一方面及其任一种可能的实现方式中所述的任务调度装置和第二方面及其任一种可能的实现方式中所述的计算装置。

附图说明

图 1 是本申请实施例提供的一种计算装置的示意性框图。

图 2 是本申请实施例提供的一种近存计算装置的示意性框图。

图 3 是本申请实施例提供的一种计算任务调度装置的示意性框图。

图 4 是本申请实施例提供的一种根据计算列表确定目标计算任务的示意图。

图 5 是本申请实施例提供的另一种计算任务调度装置的示意性框图。

图 6 是本申请实施例提供的一种计算任务调度方法的示意性流程图。

图 7 是本申请实施例提供的另一种计算任务调度方法的示意性流程图。

具体实施方式

下面将结合附图，对本申请中的技术方案进行描述。

在通用计算机系统中，冯诺依曼或者哈佛架构均为计算与存储分离架构。计算需要的数据需从外存加载到计算核内，计算完成需要从核内存回到外存。在当前神经网络快速发展的时代，大多加速硬件采用冯诺依曼架构。而神经网络的计算特点为计算密集型同时也是数据密集型，计算核内具有高数据并行的计算资源，对带宽需求非常大。所以在整体计算的功耗分解中往往数据传输的功耗开销相比计算功耗开销还要高。

为了降低数据传输的功耗开销，可以采用计算融合的技术方案，即：将多步计算通过融合的方式进行计算从而可以减少与外存的交互。计算融合能够有效缓解带宽压力，起到降低传输功耗开销的目的。但计算融合需要计算核内有一定大小的高速缓存(如 static random-access memory，SRAM)，另外计算融合需要精细的软件切分管理，实现复杂度较高。

除了计算融合之外，还可以采用近存计算（near-memory computing，NMC）或存内计

算（in-memory computing，IMC）技术来降低数据传输的功耗开销。近存计算和存内计算都是一种聚焦在存储器的新体系结构技术方向，通过在存储器附近计算或者存储器内部直接计算，从而突破冯诺依曼架构的限制，从而解决了数据传输的功耗开销。近存计算把存储器和计算处理器紧密耦合一起，通过较短的导线降低数据传输的延迟和功耗，从而提高系统能效。随着制造工艺和封装技术发展，将计算逻辑与存储堆叠构建混合计算存储。而存内计算是计算直接在内存阵列中完成，减少了计算处理器与内存的数据传输。但近存或存内计算技术受限于计算特点与存储计算硬件设计复杂度。

一般情况下，近存计算或存内计算需要新增相应的计算指令，且针对不同的硬件平台，适配方案不统一，从而集成复杂度较高。有鉴于此，本申请实施例提供一种存储计算装置和存储计算方法，该技术方案能够在保证数据传输的功耗较低的情况下，进一步降低实现的复杂度。

在介绍本申请的技术方案之前，首先结合图 1-2 介绍一下普通计算与近存计算或存内计算的区别。

图 1 是本申请实施例提供的一种普通计算装置的示意性框图。如图 1 所示，该装置 100a 中，存储器 110 将需要计算的数据通过总线写入缓存（buffer）120 中，普通计算单元 130 在进行计算时，该普通计算单元 130 进行读缓存，从缓存 120 中获取需要计算的数据，然后该普通计算单元 130 完成计算操作，并将计算结果写入缓存 120 中，通过总线将数据从缓存 120 中写入存储器 110。在普通计算单元 130 完成计算的过程中，缓存 120 需要进行多次读写，且存储器 110 需要通过总线与缓存 120 进行多次交互，使得系统总线开销较大。

图 2 是本申请实施例提供的一种近存计算装置的示意性框图。如图 2 所示，该装置 100b 中，近存计算单元 150 可以位于存储器 140 外部，与存储器 140 紧密耦合，从而该近存计算单元 150 在进行计算时，可以不通过总线与存储器 140 进行交互，而是通过物理导线或电路连线进行数据交互，由于近存计算单元 150 与存储器 140 紧密耦合在一起，二者距离较近，传输数据的物理导线或电路连线较短，从而可以降低数据在近存计算单元和存储器传输的时延与功耗，也降低了总线开销。

在一些实施例中，该近存计算单元 150 可以用存内计算单元代替，该存内计算单元可以位于存储器 140 内部，例如，该存内计算单元可以嵌入存储器 140 内部，作为存储器的一部分，也就是说存储器具备计算能力，该存内计算单元可以通过物理导线或电路连线与存储器交互，该存内计算单元还可以不通过读写协议直接读取存储器内部的数据以完成计算，而无需通过总线，从而可以节省总线开销。

该近存计算单元 150 和存储器 140，还可以用存算一体单元代替，此时，该存算一体单元既可以存储数据，又可以完成计算，从而可以节省计算与存储之间的总线开销，也可以降低数据传输的时延与功耗。

下文将结合图 3 至图 7 详细介绍本申请实施例中的技术方案。

图 3 是本申请实施例提供的一种计算任务调度装置的示意性框图。如图 3 所示，该装置 200 可以包括计算业务单元 210、任务调度器 220、处理器 230、存储器 240 和近存计算单元 250。可选地，该装置 200 中的任务调度器 220、处理器 230 可以位于一个芯片，如片上系统（SoC）内。存储器 240 和近存计算单元 250 可以位于另一个芯片内。

其中，计算业务单元 210 位于业务调度层，属于软件模块，例如，该计算业务单元

210 可以是系统软件，也可以是应用软件；任务调度器 220 属于硬件调度器，处理器 230、存储器 240 和近存计算单元 250 均属于硬件器件。处理器 230 可以运行所述系统软件或应用软件以执行计算或处理任务，处理器还可以与其他硬件设备进行交互，如发送/接收数据或指令等。存储器 240 可以用于存储数据并能够被其他硬件设备，如处理器 230 访问。近存计算单元 250 可以包括计算电路，用于执行计算任务，该计算任务可以不同于处理器 230 执行的计算任务。

示例性地，计算业务单元 210 将编译好的计算任务序列发送至任务调度器 220；任务调度器 220 对该计算任务序列进行解析，并确定计算任务是否可以进行近存计算，当确定目标计算任务可以进行近存计算时，任务调度器 220 调用近存计算负载生成函数生成第一计算任务的负载信息，并将该第一计算任务的负载信息调度至处理器 230；该处理器 230（例如，CPU）将该第一计算任务的负载信息存储至存储器中的第一地址。可选地，该第一地址是预留地址，用于该处理器 230 与近存计算单元 250 之间的所述负载信息交互，近存计算单元 250 可以访问该第一地址，以获取目标计算任务的负载信息，之后近存计算单元 250 根据该负载信息完成计算，并将计算的结果存储至存储器 240 中。

具体地，任务调度器 220 可以为近存计算单元确定第一计算任务，生成第一计算任务的负载信息，该负载信息用于定义第一计算任务。

示例性地，该任务调度器为系统软件或应用软件之外的专用任务调度器。即该任务调度器为装置 200 中专门用于调度计算任务的硬件任务调度器。

处理器 230 可以从任务调度器 220 接收负载信息，将负载信息存储至存储器 240 中的第一地址以将第一计算任务分配给近存计算单元 250，第一地址为近存计算单元 250 的预留地址，其中，处理器 230 与存储器 240 和近存计算单元 250 中的至少一个通过总线耦合，近存计算单元 250 与存储器 240 紧密耦合，所述紧密耦合无需经过任何总线，且近存计算单元 250 能够以高于总线接入的速度接入存储器 240。

该实施例中，紧密耦合为近存计算耦合。

应理解，该负载信息用于定义第一计算任务，可以理解为，该负载信息中的内容是计算第一计算任务需要的内容，可以用于近存计算单元完成该第一计算任务。

该存储器 240 中的第一地址为近存计算单元 250 的预留地址，即该存储器 240 中为该近存计算单元 250 预留了一块区域，该区域中可以存储有近存计算单元 250 计算所需的负载信息。该近存计算单元 250 可以访问该第一地址，以获取该负载信息，从而根据该负载信息完成第一计算任务。

在一种可能的实现方式中，处理器 230 与存储器 240 通过总线耦合，存储器 240 与近存计算单元 250 紧密耦合，即近存计算单元 250 与存储器交互数据无需通过任何总线，且该近存计算单元接入存储器 240 的速度高于通过总线接入存储器 240 的速度。例如，二者可以通过物理导线或电路连线交互，从而无需通过总线，可以节省总线开销，进而降低数据传输的时延和功耗。在另一种可能的实现方式中，当处理器 230 与存储器 240 通过总线耦合，且处理器 230 与近存计算单元 250 不通过总线耦合时，处理器 230 可以通过直接存储访问 DMA 控制器将负载信息存储至第一地址，近存计算单元 250 从该第一地址中获取该负载信息。处理器 230 还可以通过配置寄存器等方式将负载信息调度至近存计算单元 250，例如，处理器 230 将负载信息写入寄存器中。近存计算单元 250 读取该寄存器，从寄存器中获取该负载信息并存入第一地址，以完成计算。该寄存器可以与处理器 230 位于

同一个芯片内，如 SoC 内。

该任务调度器可以在计算任务序列中为近存计算单元确定第一计算任务。

具体地，该任务调度器确定第一计算任务可以是根据该第一计算任务的类型进行确定的。例如，该任务调度器中可以预先存储有计算类型，该计算类型可以是预先设置的一项或多项，例如，该计算类型可以包括矩阵类计算、循环计算等等，该计算类型可以处于一个计算列表中或链表中。

示例性地，任务调度器根据计算列表，在计算任务序列中为近存计算单元 250 确定第一计算任务。具体地，当一个计算任务的计算类型处于计算列表中包括的计算类型时，可以确定该计算任务为第一计算任务。该第一计算任务可以是是一个计算任务，也可以是多个计算任务，本申请实施例对此不予限定。

参见图 4，图 4 是本申请实施例提供的一种根据计算列表确定第一计算任务的示意图。如图 4 所示，该计算列表中可以包括计算任务的计算类型 A、计算类型 B、计算类型 C、计算类型 D 等等，该计算任务序列可以包括计算任务一（计算类型为 A）、计算任务二（计算类型为 C）、计算任务三（计算类型为 E）、计算任务四（计算类型为 F）等等。

任务调度器可以预先存储有该计算列表，当任务调度器接收到计算业务单元发送的计算任务序列后，可以根据计算任务序列中的计算任务的类型是否包括在计算列表中来确定目标计算任务。继续参见图 4，计算任务序列中的计算任务一和计算任务二的计算类型包括在计算列表中，则可以确定计算任务一和计算任务二为第一计算任务。

应理解，该计算类型可以是与近存计算单元相关的，例如，该计算类型可以是该近存计算单元支持的计算类型。示例性地，该近存计算单元支持的近存计算的类型为矩阵类计算，则该计算类型可以包括该矩阵类计算，或者，当该计算类型不包括该矩阵类计算的类型时，可以将该矩阵类计算的类型添加至计算类型中，以完成对该计算类型的更新。

在一些实施例中，该计算类型还可以由计算任务单元发送至任务调度器的。该第一计算任务的负载信息可以包括但不限于：数据地址；数据维度；控制命令字等。其中，该数据地址可以是用于指示该数据在存储器中存放的地址；该数据维度用于指示该数据的维度信息，例如，行数、列数，按照行优先存储、按照列优先存储等，该数据维度还可以包括数据类型，该数据类型可以是浮点型、整型等；该控制命令字可以是用于控制该第一计算任务的计算类型，例如，乘法、加法、乘加等。

该处理器可以包括但不限于：中央处理器（central processing unit，CPU）、图形处理器（graphics processing unit，GPU）、神经网络处理器（neural-network processing unit，NPU）等。

该处理器将第一计算任务的负载信息调度至近存计算单元可以通过以下几种方式：

方式一：

处理器通过直接存储访问（direct memory access，DMA）控制器将该负载信息存储至第一地址，近存计算单元访问该第一地址，从而可以获取该负载信息。

例如，处理器向 DMA 发送指令，该指令中可以包括负载信息的源地址和目的地址，DMA 控制器根据处理器的指令，通过 DMA 写信号将负载信息从源地址搬运至目的地址中，也即将负载信息传输至存储器中。该技术方案中，处理器可以向 DMA 控制器发送指令，将该负载信息利用已有的 DMA 机制传输至存储器的第一地址中，从而可以无需重新设计，降低了设计复杂度。

在这种情况下 DMA 控制器在处理器的控制下，通过 DMA 写信号将该负载信息传输至存储器中的固定地址中，该固定地址可以是近存计算单元的预留地址，该近存计算单元可以从该固定地址中获取该负载信息并进行解析，并根据该负载信息完成计算，并将计算的结果写入存储器中，然后存储器返回 DMA 响应信号至 DMA 控制器，DMA 控制器将 DMA 传输完成的消息或指令传输至处理器。

相应的，存储器在 DMA 控制器的操作下在第一地址写入负载信息。

该技术方案中，处理器通过已有的 DMA 机制将负载信息调度至近存计算单元，从而处理器无需通过近存计算专用指令将负载信息单独发送至近存计算单元，从而实现了处理器与近存计算单元的解耦，节省了总线开销。此外，由于复用现有的 DMA 机制，可以降低设计的复杂度。

方式二:

处理器可以通过配置寄存器等方式将负载信息调度至近存计算单元。例如，负载信息调度至近存计算单元中第一地址。

示例性地，处理器将负载信息写入片上寄存器中，近存计算单元读取该寄存器，从寄存器中获取该负载信息，以完成计算。该片上寄存器可以位于 SoC 上。

应理解，由于该近存计算单元 250 可以在存储器 240 附近完成计算，该近存计算单元 250 可以通过物理导线或电路连线与存储器 240 完成交互，无需通过系统总线，从而可以降低数据传输的时延与功耗，节省系统总线开销，从而提高系统计算效率。

进一步地，该技术方案中，任务调度器可以对计算任务进行解析以确定计算任务是否支持近存计算，从而计算业务单元无需感知计算单元的计算能力，此外，处理器采用访问存储器的方式将计算任务的负载信息存储至存储器中的第一地址中，以将计算任务分配给近存计算单元，近存计算单元与存储器紧密耦合，从而可以快速完成任务计算，无需在处理器与近存计算单元之间通过特定总线或接口传输任务调度信息，从而降低了总线开销。

在一些实施例中，当存储器和近存计算单元更换时，更换后的近存计算单元支持的计算类型可能会发生改变，此时，可以将新增加的计算类型添加至上述计算列表中，以完成计算列表的更新。

该技术方案中，任务调度器中的预设计算类型可以针对不同的近存计算单元和存储器进行更新，使得任务调度器的适配性更好，进一步提高了兼容性。

在另一些实施例中，任务调度器还可以用于在计算任务序列中确定第二计算任务，当任务调度器确定计算任务序列的第二计算任务不适合进行近存计算时，可以将该计算任务调度至其他的第二计算单元中，该第二计算单元计算该第二计算任务，其中，该第二计算单元可以与存储器通过总线耦合，该第二计算单元可以是包括所述处理器、图像处理单元、人工智能（artificial intelligence，AI）处理单元、数字信号处理器或专用逻辑电路中的至少一个。

图 4 是本申请实施例提供的另一种计算任务调度装置的示意性框图。如图 4 所示，该装置 300 可以包括计算业务单元 210、任务调度器 220、处理器 230、存储器 240 和存内计算单元 260。

其中，该存内计算单元 260 可以处于存储器 240 中，例如，该存内计算单元 260 可以嵌入存储器 240 内部，作为存储器的一部分，该存内计算单元 260 在存储器 240 内部可以通过更短的物理导线或电路连线与存储器 240 进行数据交互，或者，该存内计算单元 260

可以直接读取存储器 240 中的数据，而无需通过总线，从而可以节省总线开销，可以快速完成数据计算和数据传输。

示例性地，计算业务单元 210 将编译好的计算任务序列发送至任务调度器 220；任务调度器 220 对该计算任务序列进行解析，并确定目标计算任务是否可以进行存内计算，当确定目标计算任务可以进行存内计算时，任务调度器 220 调用存内计算负载生成函数生成目标计算任务的负载信息，并将该负载信息调度至处理器 230；该处理器 230（例如，CPU）将该目标计算任务的负载信息调度至存储器中的第一地址，存内计算单元 260 访问该第一地址以获取负载信息，存内计算单元 260 根据该负载信息对该目标计算任务进行计算，并将计算的结果存储至存储器 240 中。

应理解，对于计算业务单元 210、任务调度器 220、处理器 230 的相关描述可以参见前文，为了简洁，不再赘述。

该实施例中，紧密耦合可以为存内计算耦合。

应理解，该存内计算单元 260 和存储器 240 可以用存算一体单元代替，此时，紧密耦合为存算一体耦合。

该技术方案中，任务调度器可以对计算任务进行解析以确定计算任务是否支持存内计算，从而计算业务单元无需感知计算单元的的计算能力，从而降低了软件的复杂度。此外，处理器采用访问存储器的方式将计算任务的负载信息存储至存储器中的固定地址中，以将计算任务分配给存内计算单元，存内计算单元与存储器紧密耦合，从而可以快速完成任务计算，无需在处理器与存内计算单元之间通过特定总线或接口传输任务调度信息，从而可以降低系统总线开销。

图 5 是本申请实施例提供的一种计算任务调度方法的示意性流程图。如图 5 所示，该方法 400 可以包括步骤 301 至步骤 308。

301，计算业务单元将计算任务序列发送至任务调度器 TS。相应的，任务调度器接收该计算任务序列。

其中，该计算任务序列可以一个计算任务，也可以是多个计算任务序列，本申请实施例对该计算任务序列中包括的计算任务的数量不做限定。

该计算业务单元可以是系统软件，也可以是应用软件。

在一些实施例中，该计算业务单元可以将该计算任务序列编译后，发送至任务调度器。

302，任务调度器为近存计算单元确定第一计算任务，生成第一计算任务的第一负载信息。

应理解，该任务调度器可以对编译后的计算任务序列进行解析，在确定第一计算任务后，调用负载生成函数生成该第一计算任务对应的近存计算负载信息。

应理解，该第一计算任务是任务调度器确定可以进行近存计算的计算任务。

该任务调度器确定第一计算任务的方式可以是根据该第一计算任务的计算类型确定的。例如，任务调度器可以确定该第一计算任务的计算类型是否属于预设的计算类型。

其中，任务调度器中可以预先保存有计算类型，例如，该计算类型可以是矩阵类计算，如矩阵与矩阵乘、矩阵与向量乘；该计算类型还可以是循环计算、向量卷积运算等等。

在一些实施例中，该计算类型还可以是计算业务单元发送至任务调度器的。

该预设的计算类型可以处于一个列表中、或链表中等。

例如，该计算类型处于列表中，该列表可以是计算类型列表，该计算类型列表中可以

包括计算类型 A、计算类型 B、计算类型 C，当任务调度器解析该计算任务序列，当计算任务序列中包括计算类型 A、或计算类型 B 或计算类型 C 的第一计算任务时，均可确定该第一计算任务适合进行近存计算，从而该任务调度器可以调用负载生成函数生成负载信息。

在另一些实施例中，任务调度器可以对该预设的计算类型进行更新，例如，针对不同的存储器和近存计算单元，其支持的近存计算的类型可能是不同的，当存储器和近存计算单元更换时，在其支持的近存计算的目标类型不包括在预设的计算类型的情况下，可以将该目标类型添加至预设的计算类型中，如将该目标类型添加至计算列表中，以完成该计算类型的更新，使得任务调度器的适配性更好，从而可以提升兼容性。

在另一些实施例中，任务调度器除了根据第一计算任务的计算类型确定该第一计算任务之外，还可以进一步根据第一计算任务的数据维度确定该第一计算任务，以确定该第一计算任务是否适合近存计算。例如，可以根据数据维度确定数据量大小（例如，行数乘以列数），当数据量大于预设值时，可以确定该第一计算任务适合近存计算，否则，该第一计算任务不适合近存计算。又如，当根据数据维度确定该第一计算任务的数据类型（例如，浮点型）与近存计算单元支持的数据类型（例如，浮点型）一致时，可以确定该第一计算任务适合近存计算，否则，该第一计算任务不适合近存计算。

在一些实施例中，任务调度器在确定了第一计算任务的计算类型属于预设的计算类型后，进一步确定该第一计算任务的数据维度不适合进行近存计算，则该任务调度器可以将该第一计算任务调度至其他计算核进行正常计算。或者，在确定了第一计算任务的计算类型属于预设的计算类型后，进一步确定该目标计算任务的数据维度适合进行近存计算，则该任务调度器可以调用负载生成函数生成负载信息，并将该负载信息调度至处理器。

在另一些实施例中，当计算任务序列中的第二计算任务不属于预设的计算类型时，说明该第二计算任务不适合近存计算，例如，该第二计算任务为控制流、激活函数等计算任务，则该任务调度器可以将该计算任务调度至第二计算单元进行正常计算，从而计算业务单元无需感知计算单元是否支持近存计算，进而降低了软件实现的复杂度。例如，该第二计算单元可以是上述处理器、图像处理单元、AI 处理单元、数字信号处理器或专用逻辑电路中的至少一个，其中，该第二计算单元可以和存储器通过总线耦合。

该第一负载信息可以用于定义该第一计算任务，从而近存计算单元可以根据该第一负载信息计算该第一计算任务。

该负载信息可以包括数据地址、数据维度、控制命令字等，对于该负载信息的具体描述可以参见前文中的相关描述，此处不再详述。

应理解，该负载信息还可以包括其他进行数据计算所需要的信息。

303，任务调度器将第一负载信息调度至处理器。相应的，处理器接收该第一负载信息。

其中，该处理器可以是 CPU、GPU、NPU 等等，该处理器还可以是 CPU、GPU、NPU 中的一个或多个计算核或计算单元，本申请实施例对此不予限定。

304，处理器将该第一负载信息调度至存储器。

在一种可能的实现方式中，处理器通过 DMA 控制器将第一负载信息调度至存储器。具体地，处理器向 DMA 控制器发送指令，DMA 控制器根据该指令通过 DMA 写信号将负载信息传输至存储器中的固定地址中，例如，DMA 控制器通过该指令利用总线将负载

信息从源地址搬运至目的地址。

其中，该存储器中预留了专门用于存储近存计算相关的信息的预留地址，即第一地址，该预留地址中可以用于存储该第一负载信息，从而近存计算单元可以访问该预留地址，以获取该第一负载信息。

这样，处理器通过已有的 DMA 机制实现负载信息的下发，可以无需增加近存计算专用指令，从而处理器无需单独通过总线下发调度信息，从而可以节省系统总线开销，提升计算效率。此外，由于复用现有的 DMA 机制，可以降低设计的复杂度。

在其他的实施例中，处理器还可以通过其他方式将该负载信息调度至近存计算单元中，例如，通过配置寄存器等方式该负载信息调度至近存计算单元中。例如，处理器将负载信息写入寄存器中，近存计算单元读取该寄存器，从寄存器中获取该负载信息，以完成计算。

305，近存计算单元从存储器中获取该第一负载信息。

示例性地，近存计算单元可以从存储器中的预留地址中获取该第一负载信息，由于近存计算单元位于存储器附近，访问存储器无需经过总线，因而可以提升数据传输的时延和功耗。

在其他的实施例中，该近存计算单元还可以通过近存计算指令获取该负载信息。

应理解，该近存计算单元还可以通过其他方式获取该负载信息。

306，近存计算单元根据该第一负载信息完成第一计算任务。

示例性地，该第一负载信息定义了地址 A 中的矩阵 A1 与地址 B 中的矩阵 B1 进行矩阵乘运算，则该近存计算单元可以根据该第一负载信息指示存储器完成相应的计算。

307，近存计算单元将计算完成的信息发送至存储器。

其中，近存计算单元在完成计算后，将计算结果写入存储器中，也即将计算完成的信息发送至存储器。

308，存储器向处理器发送响应信号。

示例性地，存储器接收到近存计算完成的信息或指令后，可以向 DMA 控制器发送 DMA 响应信号，该 DMA 控制器接收到 DMA 响应信号后，向处理器发送 DMA 传输完成的信号或指令。

基于本申请实施例，近存计算单元可以在存储器附近完成计算，从而可以降低数据传输的时延与功耗，从而提高系统计算效率。

该技术方案中，任务调度器可以对计算任务进行解析以确定计算任务是否支持近存计算，从而计算业务单元无需感知近存计算单元的计算能力，从而降低了软件复杂度。此外，任务调度器中的预设存储计算类型可以针对不同的硬件平台进行更新，进一步提高了兼容性。进一步地，处理器可以通过已有的 DMA 机制将近存计算需要的负载信息调度至近存计算单元，从而无需新增加近存计算指令用于调度计算任务，因此，可以节省总线开销，提升计算效率。

图 6 是本申请实施例提供的另一种计算任务调度方法的示意性流程图。如图 6 所示，该方法 500 可以包括步骤 401 至步骤 407。

401，计算业务单元将计算任务序列发送至任务调度器 TS。相应的，任务调度器接收该计算任务序列。

402，任务调度器为存内计算单元确定第三计算任务，生成第三计算任务的第二负载信息。

403，任务调度器将第二负载信息调度至处理器。相应的，处理器接收该第二负载信息。

404，处理器将第二负载信息调度至存储器。

应理解，步骤 401 至步骤 404 可以参见步骤 301 至步骤 304 的相关描述，为了简洁，不再赘述。

405，存内计算单元从存储器中获取该第二负载信息。

示例性地，该存储器中可以为存内计算单元预留了专门用于存储存内计算相关的信息的预留地址，该预留地址中可以用于存储该第二载信息，从而存内计算单元可以访问该预留地址，以获取该第二负载信息。

406，存内计算单元根据该第二负载信息完成第三计算任务。

407，存储器向处理器发送响应信号。

应理解，步骤 406 至步骤 407 可以参见步骤 306 至步骤 307 的相关描述，为了简洁，不再赘述。

在另一些实施例中，该存内计算单元和存储器还可以用存算一体单元进行代替。

基于本申请实施例，存内计算单元可以在存储器内部完成计算，从而无需新增存内计算指令，节省了总线开销，可以降低数据传输的时延与功耗，从而提高系统计算效率。

该技术方案中，任务调度器可以对计算任务进行解析以确定计算任务是否支持存内计算，从而计算业务单元无需感知存内计算单元支持的计算任务的计算类型，降低了软件复杂度。此外，任务调度器中的预设计算类型可以针对不同的存内计算单元和存储器进行更新，进一步提高了兼容性。进一步地，处理器可以通过已有的 DMA 机制将存内计算需要的负载信息调度至存内计算单元，无需通过总线单独传输存内计算指令，从而可以节省总线开销，提升计算效率。

本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质中存储有计算机指令，当该计算机指令在计算机上运行时，使得如前文中任一项所述的计算任务调度方法被执行。

本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质中存储有计算机指令，当该计算机指令在计算机上运行时，使得如前文中任一项所述的计算方法被执行。

本申请实施例还提供了一种计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述相关步骤，以实现上述实施例中的计算任务调度方法。

本申请实施例还提供了一种计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述相关步骤，以实现上述实施例中的计算方法。

本申请实施例还提供了一种计算系统，包括如前文中任一项所述的计算任务调度装置和计算装置。

另外，本申请的实施例还提供一种装置，这个装置具体可以是芯片，组件或模块，该装置可包括相连的处理器和存储器；其中，存储器用于存储计算机执行指令，当装置运行时，处理器可执行存储器存储的计算机执行指令，以使芯片执行上述各方法实施例中的计算任务调度方法或计算方法。

其中，本实施例提供的计算任务调度装置、计算装置、计算机可读存储介质、计算机程序产品或芯片均用于执行上文所提供的对应的方法，因此，其所能达到的有益效果可参考上文所提供的对应的方法中的有益效果，此处不再赘述。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

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。

所述功能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（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。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权利要求书

1. 一种计算任务调度装置，其特征在于，包括:

任务调度器，用于为第一计算单元确定第一计算任务，生成所述第一计算任务的负载信息，所述负载信息用于定义所述第一计算任务;

处理器，用于从所述任务调度器接收所述负载信息，将所述负载信息存储至存储器中的第一地址以将所述第一计算任务分配给所述第一计算单元，所述第一地址为所述第一计算单元的预留地址，其中，所述处理器与所述存储器和所述第一计算单元中的至少一个通过总线耦合，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线接入的速度接入所述存储器。

2. 根据权利要求 1 所述的装置，其特征在于，所述处理器具体用于:

通过直接存储访问 DMA 控制器将所述负载信息存储至所述第一地址。

3. 根据权利要求 1 或 2 所述的装置，其特征在于，所述任务调度器为系统软件或应用软件之外的专用任务调度器。

4. 根据权利要求 3 所述的装置，其特征在于，所述任务调度器还用于:

接收来自所述系统软件或所述应用软件的计算任务序列，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务。

5. 根据权利要求 4 所述的装置，其特征在于，所述任务调度器还用于:

在所述计算任务序列中为第二计算单元确定第二计算任务;

将所述第二计算任务调度至第二计算单元;

其中，所述第二计算单元包括所述处理器、图像处理单元、人工智能 AI 处理单元、数字信号处理器或专用逻辑电路中的至少一个，所述第二计算单元和所述存储器通过总线耦合。

6. 根据权利要求 4 所述的装置，其特征在于，所述任务调度器具体用于:

根据计算列表，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务，其中，所述计算列表包括所述第一计算单元支持的计算任务类型。

7. 根据权利要求 1-6 中任一项所述的装置，其特征在于，所述负载信息包括如下信息中的至少一种:

数据地址; 数据维度; 或控制命令字。

8. 根据权利要求 1-7 中任一项所述的装置，其特征在于，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

9. 一种计算装置，其特征在于，包括:

存储器;

第一计算单元，用于从所述存储器中的第一地址获取负载信息，并根据所述负载信息完成第一计算任务，其中，所述负载信息用于定义所述第一计算任务，所述第一地址为所述第一计算单元的预留地址;

其中，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线接入的速度接入所述存储器;

所述存储器和所述第一计算单元中的至少一个通过总线耦合至处理器。

10. 根据权利要求 9 所述的装置，其特征在于，所述负载信息包括如下信息中的至少一种：

数据地址；数据维度；或控制命令字。

11. 根据权利要求 9 或 10 所述的装置，其特征在于，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

12. 根据权利要求 9-11 中任一项所述的装置，其特征在于，所述存储器具体用于：

在直接存储访问 DMA 控制器的操作下在所述第一地址写入所述负载信息。

13. 一种计算任务调度方法，其特征在于，包括：

任务调度器为第一计算单元确定第一计算任务，生成所述第一计算任务的负载信息，所述负载信息用于定义所述第一计算任务；

处理器从所述任务调度器接收所述负载信息，将所述负载信息存储至存储器中的第一地址以将所述第一计算任务分配给所述第一计算单元，所述第一地址为所述第一计算单元的预留地址，其中，所述处理器与所述存储器和所述第一计算单元中的至少一个通过总线耦合，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线接入的速度接入所述存储器。

14. 根据权利要求 13 所述的方法，其特征在于，所述将所述负载信息存储至存储器中的第一地址以将所述第一计算任务分配给所述第一计算单元，包括：

通过直接存储访问 DMA 控制器将所述负载信息存储至所述第一地址以将所述第一计算任务分配给所述第一计算单元。

15. 根据权利要求 13 或 14 所述的方法，其特征在于，所述任务调度器为系统软件或应用软件之外的专用任务调度器。

16. 根据权利要求 15 所述的方法，其特征在于，所述任务调度器为第一计算单元确定第一计算任务，包括：

所述任务调度器接收来自所述系统软件或所述应用软件的计算任务序列，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务。

17. 根据权利要求 16 所述的方法，其特征在于，所述方法还包括：

在所述计算任务序列中为第二计算单元确定第二计算任务；

将所述第二计算任务调度至第二计算单元；

其中，所述第二计算单元包括所述处理器、图像处理单元、人工智能 AI 处理单元、数字信号处理器或专用逻辑电路中的至少一个，所述第二计算单元和所述存储器通过总线耦合。

18. 根据权利要求 16 所述的方法，其特征在于，所述在所述计算任务序列中为所述第一计算单元确定所述第一计算任务，包括：

根据计算列表，在所述计算任务序列中为所述第一计算单元确定所述第一计算任务，其中，所述计算列表包括所述第一计算单元支持的计算任务类型。

19. 根据权利要求 13-18 中任一项所述的方法，其特征在于，所述负载信息包括如下信息中的至少一种：

数据地址；数据维度；或控制命令字。

20. 根据权利要求 13-19 中任一项所述的方法，其特征在于，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

21. 一种计算方法，其特征在于，包括:

第一计算单元从存储器中的第一地址获取负载信息，并根据所述负载信息完成第一计算任务，其中，所述负载信息用于定义所述第一计算任务，所述第一地址为所述第一计算单元的预留地址;

其中，所述第一计算单元与所述存储器紧密耦合，所述紧密耦合无需经过任何总线，且所述第一计算单元能够以高于总线接入的速度接入所述存储器;

所述存储器和所述第一计算单元中的至少一个通过总线耦合至处理器。

22. 根据权利要求 21 所述的方法，其特征在于，所述负载信息包括如下信息中的至少一种:

数据地址; 数据维度; 或控制命令字。

23. 根据权利要求 21 或 22 所述的方法，其特征在于，所述紧密耦合包括近存计算耦合、存内计算耦合或存算一体耦合。

24. 根据权利要求 21-23 中任一项所述的方法，其特征在于，所述方法还包括:

所述存储器在直接存储访问 DMA 控制器的操作下在所述第一地址写入所述负载信息。

25. 一种计算机可读存储介质，其特征在于，包括: 所述存储介质中存储有计算机程序或指令，当所述计算机程序或指令被通信装置执行时，使得如权利要求 13-20 中任一项所述的计算任务调度方法被执行，或者，使得如权利要求 21-24 中任一项所述的计算方法被执行。

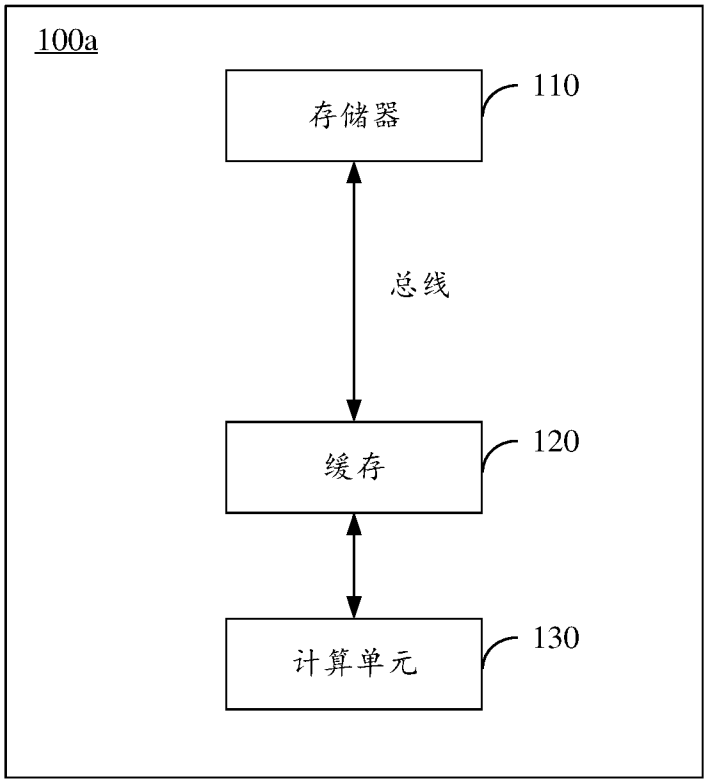

图 1

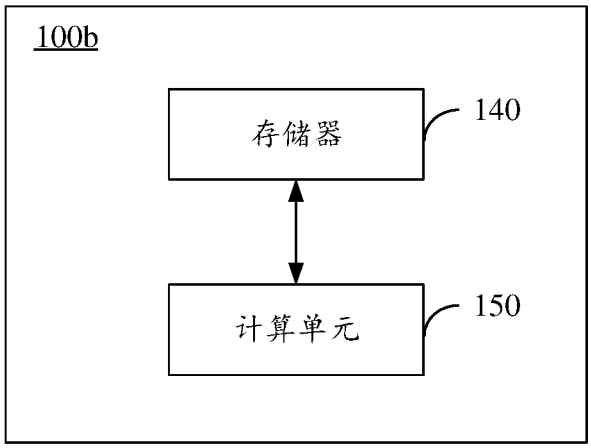

图 2

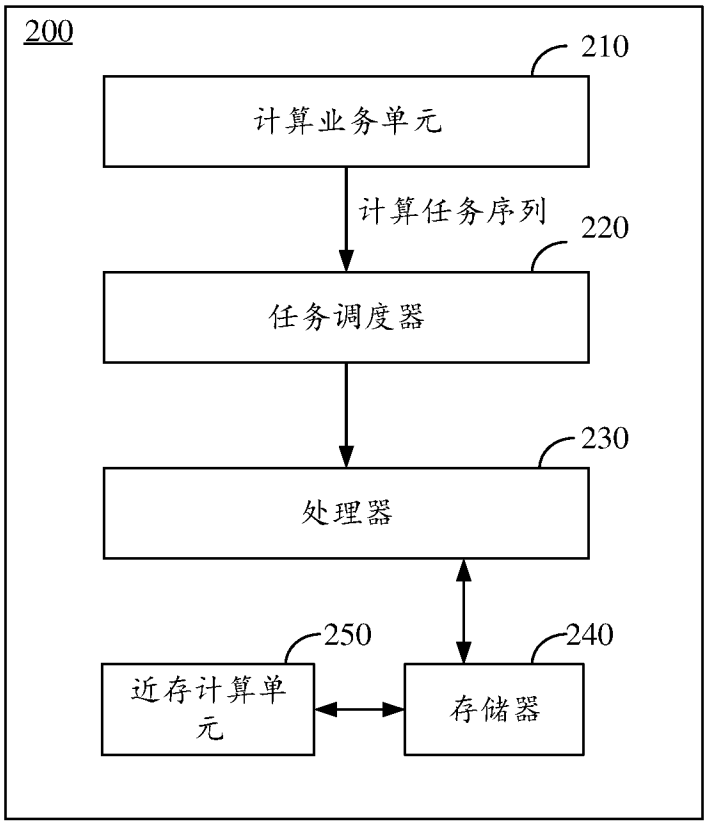
200
210
计算业务单元
计算任务序列
220
任务调度器
230
处理器
250
近存计算单元
240
存储器

图 3

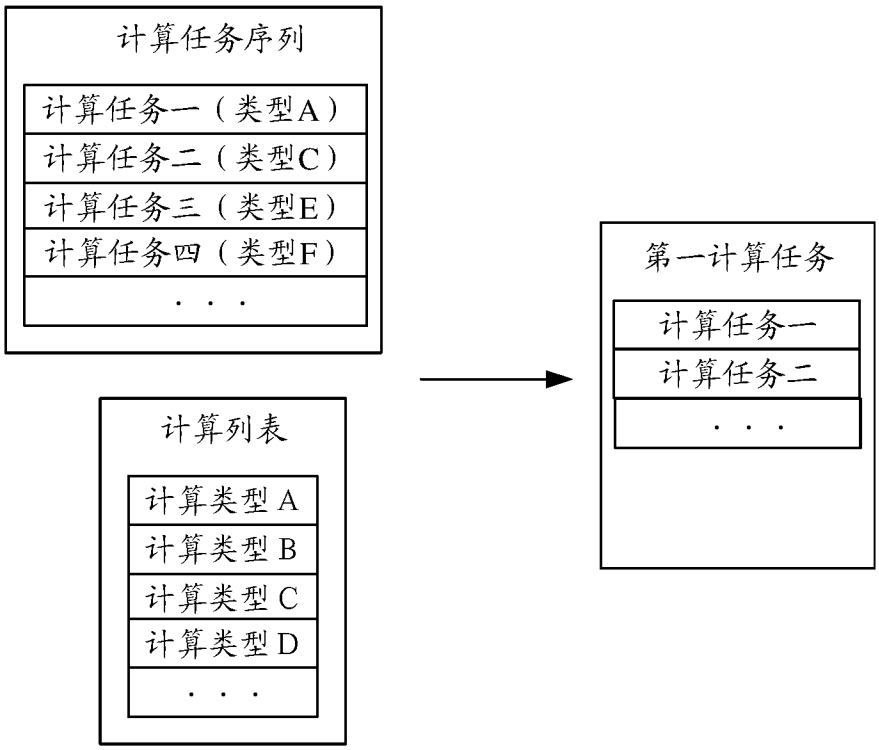
计算任务序列
计算任务一（类型A）
计算任务二（类型C）
计算任务三（类型E）
计算任务四（类型F）
· · ·
计算列表
计算类型 A
计算类型 B
计算类型 C
计算类型 D
· · ·
第一计算任务
计算任务一
计算任务二
· · ·

图 4

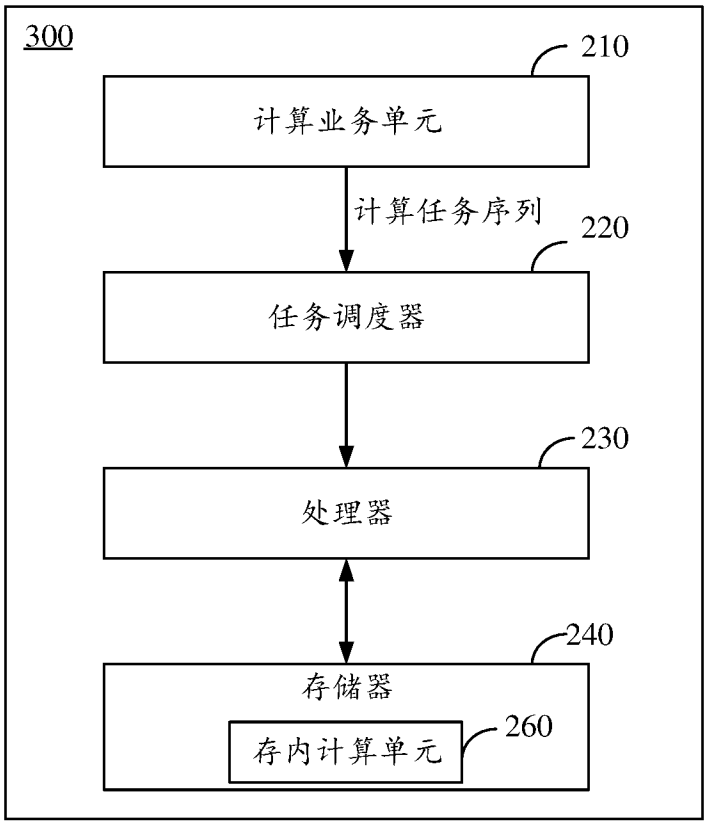
300
210
计算业务单元
计算任务序列
220
任务调度器
230
处理器
240
存储器
260
存内计算单元

图 5

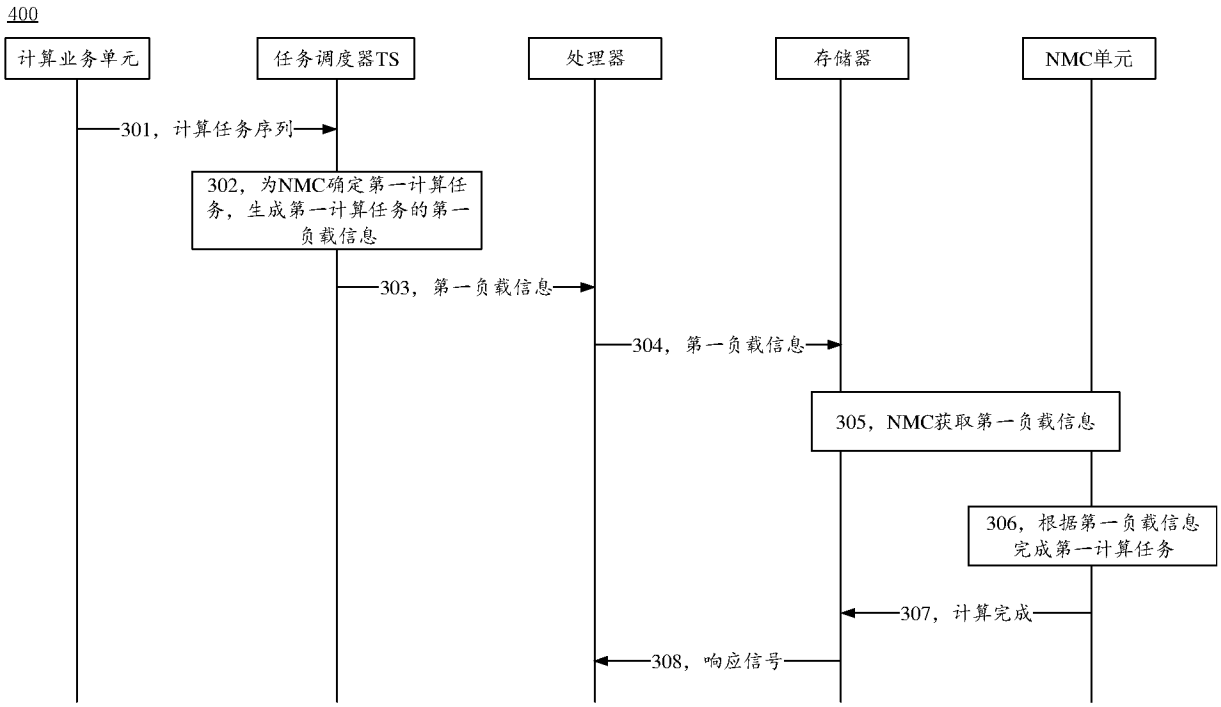
400
计算业务单元
任务调度器TS
处理器
存储器
NMC单元
301，计算任务序列
302，为NMC确定第一计算任务，生成第一计算任务的第一负载信息
303，第一负载信息
304，第一负载信息
305，NMC获取第一负载信息
306，根据第一负载信息完成第一计算任务
307，计算完成
308，响应信号

图 6

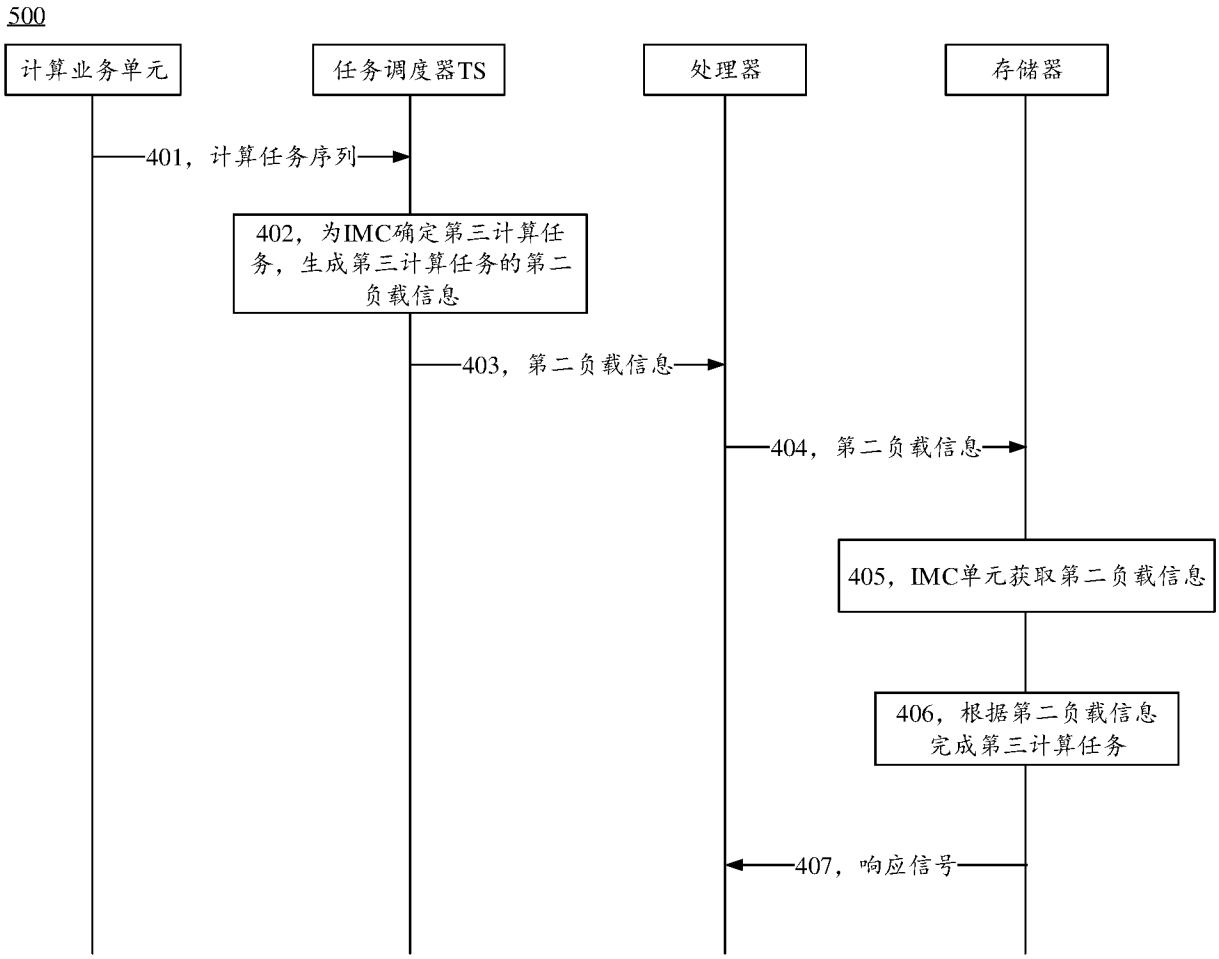
500
计算业务单元
任务调度器TS
处理器
存储器
401，计算任务序列
402，为IMC确定第三计算任务，生成第三计算任务的第二负载信息
403，第二负载信息
404，第二负载信息
405，IMC单元获取第二负载信息
406，根据第二负载信息完成第三计算任务
407，响应信号

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/075123

A. CLASSIFICATION OF SUBJECT MATTER

H04W 72/04(2009.01)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)

H04W; H04Q; 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; CNKI; 3GPP: 高于, 调度, 计算, 总线, 耦合, 负载, 存储器; calculation, address, schedul, memory, computing, load

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 111651253 A (CHINA UNITED NETWORK COMMUNICATIONS GROUP CO., LTD.) 11 September 2020 (2020-09-11) description, paragraphs [0032]-[0127], and figures 1-7	1-25
A	CN 101630053 A (HONGFUJIN PRECISION INDUSTRY (SHENZHEN) CO., LTD.) 20 January 2010 (2010-01-20)	1-25
A	CN 110678847 A (ADVANCED MICRO DEVICES, INC.) 10 January 2020 (2020-01-10) entire document	1-25
A	CN 111656335 A (MICRON TECHNOLOGY INC.) 11 September 2020 (2020-09-11) entire document	1-25
A	CN 110049130 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 23 July 2019 (2019-07-23) entire document	1-25
A	US 6738881 B1 (TEXAS INSTRUMENTS INC.) 18 May 2004 (2004-05-18) entire document	1-25
A	CA 3083316 A1 (COMCAST CABLE COMMUNICATIONS, LLC.) 11 December 2020 (2020-12-11) entire document	1-25

☐ 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
23 September 2022	**29 September 2022**
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/CN2022/075123

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	111651253	A	11 September 2020	None			
CN	101630053	A	20 January 2010	US	2010013113	A1	21 January 2010
CN	110678847	A	10 January 2020	KR	20200011958	A	04 February 2020
				EP	3631636	A1	08 April 2020
				US	2020379802	A1	03 December 2020
				WO	2018222522	A1	06 December 2018
				US	2018349145	A1	06 December 2018
				JP	2020522797	A	30 July 2020
CN	111656335	A	11 September 2020	US	2021181991	A1	17 June 2021
				EP	3746903	A1	09 December 2020
				KR	20200113264	A	06 October 2020
				CN	111656334	A	11 September 2020
				US	2019324928	A1	24 October 2019
				US	2019272119	A1	05 September 2019
				KR	20200111722	A	29 September 2020
				EP	3746902	A1	09 December 2020
				US	2021149600	A1	20 May 2021
CN	110049130	A	23 July 2019	None			
US	6738881	B1	18 May 2004	DE	69924475	D1	04 May 2005
				EP	1059589	A1	13 December 2000
CA	3083316	A1	11 December 2020	US	2020396760	A1	17 December 2020
				EP	3751776	A1	16 December 2020

国际检索报告

国际申请号
PCT/CN2022/075123

A. 主题的分类

H04W 72/04(2009.01)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)

H04W; H04Q; 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;CNKI;3GPP:高于，调度，计算，总线，耦合，负载，存储器；calculation，address，schedul，memory，computing，load

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
A	CN 111651253 A (中国联合网络通信集团有限公司) 2020年9月11日 (2020 - 09 - 11) 说明书第[0032]-[0127]段，图1-7	1-25
A	CN 101630053 A (鸿富锦精密工业深圳有限公司) 2010年1月20日 (2010 - 01 - 20)	1-25
A	CN 110678847 A (超威半导体公司) 2020年1月10日 (2020 - 01 - 10) 全文	1-25
A	CN 111656335 A (美光科技公司) 2020年9月11日 (2020 - 09 - 11) 全文	1-25
A	CN 110049130 A (北京邮电大学) 2019年7月23日 (2019 - 07 - 23) 全文	1-25
A	US 6738881 B1 (TEXAS INSTRUMENTS INC) 2004年5月18日 (2004 - 05 - 18) 全文	1-25
A	CA 3083316 A1 (COMCAST CABLE COMM LLC) 2020年12月11日 (2020 - 12 - 11) 全文	1-25

☐ 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
2022年9月23日	2022年9月29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 李冰 电话号码 86-(010)-62089557

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/075123

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	111651253	A	2020年9月11日	无			
CN	101630053	A	2010年1月20日	US	2010013113	A1	2010年1月21日
CN	110678847	A	2020年1月10日	KR	20200011958	A	2020年2月4日
				EP	3631636	A1	2020年4月8日
				US	2020379802	A1	2020年12月3日
				WO	2018222522	A1	2018年12月6日
				US	2018349145	A1	2018年12月6日
				JP	2020522797	A	2020年7月30日
CN	111656335	A	2020年9月11日	US	2021181991	A1	2021年6月17日
				EP	3746903	A1	2020年12月9日
				KR	20200113264	A	2020年10月6日
				CN	111656334	A	2020年9月11日
				US	2019324928	A1	2019年10月24日
				US	2019272119	A1	2019年9月5日
				KR	20200111722	A	2020年9月29日
				EP	3746902	A1	2020年12月9日
				US	2021149600	A1	2021年5月20日
CN	110049130	A	2019年7月23日	无			
US	6738881	B1	2004年5月18日	DE	69924475	D1	2005年5月4日
				EP	1059589	A1	2000年12月13日
CA	3083316	A1	2020年12月11日	US	2020396760	A1	2020年12月17日
				EP	3751776	A1	2020年12月16日