(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 8 月 22 日 (22.08.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/169293 A1

(51) 国际专利分类号:
G06N 3/063 (2023.01)

(21) 国际申请号: PCT/CN2023/132743

(22) 国际申请日: 2023 年 11 月 20 日 (20.11.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310114640.4 2023年2月15日 (15.02.2023) CN

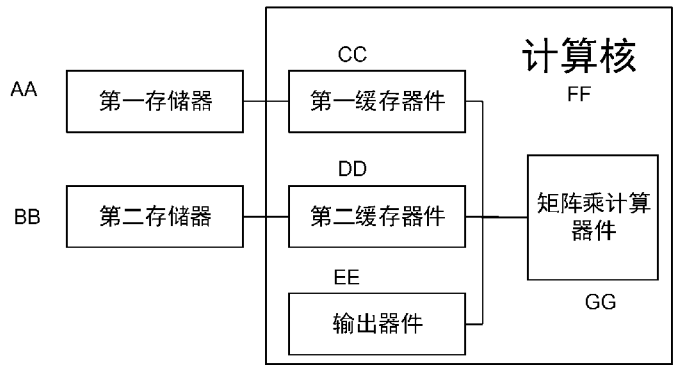
(71) 申请人: 苏州元脑智能科技有限公司 (**SUZHOU METABRAIN INTELLIGENT TECHNOLOGY CO., LTD.**) [CN/CN]; 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。

(72) 发明人:李仁刚(**LI, Rengang**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。杨宏斌(**YANG, Hongbin**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。董刚(**DONG, Gang**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。蒋东东(**JIANG, Dongdong**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。曹其春(**CAO, Qichun**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。胡克坤(**HU, Kekun**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。

(74) 代理人: 北京康信知识产权代理有限责任公司 (**KANGXIN PARTNERS , P.C.**); 中国北京市海淀区知春路甲48号盈都大厦A座16层, Beijing 100098 (CN)。

(54) **Title:** COMPUTING CORE, ACCELERATOR, COMPUTING METHOD AND APPARATUS, DEVICE, NON-VOLATILE READABLE STORAGE MEDIUM, AND SYSTEM

(54) 发明名称: 计算核、加速器、计算方法、装置、设备、非易失性可读存储介质及系统



AA First memory
BB Second memory
CC First buffer device
DD Second buffer device
EE Output device
FF Computing core
GG Matrix multiplication device

图 1

(57) **Abstract:** Disclosed in the present application are a computing core, an accelerator, a computing method and apparatus, a device, a non-volatile readable storage medium, and a system in the technical field of computers. According to the present application, parallel computing can be performed to obtain matrix multiplication results of N rows of data or N columns of data in a first matrix and N columns of data or N rows of data in a second matrix, so that N final matrix multiplication results may be obtained at one time, and the computing efficiency and speed are improved. The computing core does not need to temporarily store an intermediate result, and a resource-on-chip is saved. According to the present application, after the matrix multiplication results are obtained, which memory the matrix multiplication results are stored into can be determined according to a participation manner in which the matrix multiplication results participate in each round of computing in a next matrix multiplication operation, and hence, a storage format of the matrix multiplication results in the memory is consistent with an output format of the matrix multiplication results when participating in computing. Data is conveniently read in sequence in a continuous computing process, and matrix transposition does not need to be carried out. Therefore, the time overhead of accessing a memory can be reduced, and the efficiency is improved.

WO 2024/169293 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: 本申请公开了计算机技术领域内的一种计算核、加速器、计算方法、装置、设备、非易失性可读存储介质及系统。本申请能够并行计算第一矩阵中的N个行数据或N个列数据与第二矩阵中的N个列数据或N个行数据的矩阵乘结果，因此可一次性得到N个最终矩阵乘结果，提高了计算效率和速度；计算核也无需暂存中间结果，节约了片上资源。在得到矩阵乘结果后，本申请能够按照该矩阵乘结果参与下一矩阵乘操作中的每一轮计算时的参与方式确定将该矩阵乘结果存入哪个存储器，由此可实现：使矩阵乘结果在存储器中的存储格式与其参与计算时的输出格式一致，便于在连续的计算过程中顺序读取数据，而无需进行矩阵转置，因此还能降低访存的时间开销，提高效率。

计算核、加速器、计算方法、装置、设备、非易失性可读存储介质及系统

相关申请的交叉引用

本申请要求于 2023 年 2 月 15 日提交中国专利局，申请号为 2023101146404，申请名称为“计算核、加速器、计算方法、装置、设备、介质及系统”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，特别涉及一种计算核、加速器、计算方法、装置、设备、非易失性可读存储介质及系统。

背景技术

目前，对于图神经网络中的矩阵乘，现有的计算器件可以采用矩阵分块操作进行矩阵乘操作，但此方式需要将大量的中间结果暂存到计算器件之外，计算器件在进行后续计算时再读回至片内进行累加，增加了许多额外的数据搬移操作，会浪费计算资源，降低计算速度。

因此，如何提高图神经网络中矩阵乘操作的计算速度和计算效率，是本领域技术人员需要解决的问题。

发明内容

有鉴于此，本申请的目的在于提供一种计算核、加速器、计算方法、装置、设备、非易失性可读存储介质及系统，以提高图神经网络中矩阵乘操作的计算速度和计算效率。其方案如下：

第一方面，本申请提供了一种计算核，包括：第一缓存器件、第二缓存器件、矩阵乘计算器件和输出器件；

第一缓存器件，用于根据接收到的计算指令从外部的第一存储器中加载参与图神经网络中的当前矩阵乘操作的第一矩阵中的数据；

第二缓存器件，用于根据计算指令从外部的第二存储器中加载参与当前矩阵乘操作的第二矩阵中的数据；

矩阵乘计算器件，用于按照计算指令设定的并行数 N 并行计算第一矩阵中的 N 个行数据或 N 个列数据与第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果；

输出器件，用于暂存矩阵乘结果，在矩阵乘结果的数据量满足预设写条件时，将矩阵乘结果写入计算指令设定的第一存储器或第二存储器；

其中，第二存储器的传输带宽大于第一存储器；若矩阵乘结果参与下一矩阵乘操作中的每一轮计算时需以遍历方式参与计算，则计算指令设定有第二存储器；否则，计算指令设定有第一存储器。

可选地，第一缓存器件被设置为：在矩阵乘计算器件计算矩阵乘结果的过程中，从第一存储器中加载参与当前矩阵乘操作的第一矩阵中的剩余未加载数据；

相应地，第二缓存器件被设置为：在矩阵乘计算器件计算矩阵乘结果的过程中，从第二存储器中加载参与当前矩阵乘操作的第二矩阵中的剩余未加载数据。

可选地，第一缓存器件利用乒乓方式加载第一矩阵中的数据。

可选地，第二缓存器件利用先入先出队列加载第二矩阵中的数据。

可选地，矩阵乘计算器件包括：至少一个计算单元，计算单元利用乘法器和累加器构成且每次计算的数据量等于 N。

第二方面，本申请提供了一种加速器，包括：控制器、第一存储器、传输带宽大于第一存储器的第二存储器、以及前文的计算核；

其中，控制器与第一存储器、第二存储器、计算核分别连接；计算核与第一存储器、第二存储器分别连接；

控制器，用于将图神经网络中的矩阵乘操作的计算指令发送至第一存储器、第二存储器和计算核；

第一存储器，用于根据计算指令以直接内存访问 DMA 方式获取并存储参与当前矩阵乘操作的第一矩阵；

第二存储器，用于根据计算指令以 DMA 方式获取并存储参与当前矩阵乘操作的第二矩阵；

计算核，用于根据计算指令计算第一矩阵和第二矩阵的矩阵乘结果。

可选地，计算核有多个。

可选地，控制器还用于：控制多个计算核执行同一矩阵乘操作。

可选地，控制器被设置为：在控制多个计算核执行同一矩阵乘操作时，使每一计算核在同一计算周期内计算第一矩阵中的相同行数据或相同列数据与第二矩阵中的不同列数据或不同行数据的矩阵乘结果。

可选地，控制器被设置为：在控制多个计算核执行同一矩阵乘操作时，使每一计算核在同一计算周期内分别计算第一矩阵中的不同列数据或不同行数据与第二矩阵中的相同列数据或相同行数据的矩阵乘结果。

可选地，控制器还用于：控制不同计算核在同一计算周期内执行不同矩阵乘操作。

可选地，计算指令用于实现图神经网络中的任意矩阵乘操作。

可选地，计算核包括：第一缓存器件、第二缓存器件、矩阵乘计算器件和输出器件；

第一缓存器件，用于根据接收到的计算指令从外部的第一存储器中加载参与图神经网络中的当前矩阵乘操作的第一矩阵中的数据；

第二缓存器件，用于根据计算指令从外部的第二存储器中加载参与当前矩阵乘操作的第二矩阵中的数据；

矩阵乘计算器件，用于按照计算指令设定的并行数 N 并行计算第一矩阵中的 N 个行数据或 N 个列数据与第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果；

输出器件，用于暂存矩阵乘结果，在矩阵乘结果的数据量满足预设写条件时，将矩阵乘结果写入计算指令设定的第一存储器或第二存储器；

其中，第二存储器的传输带宽大于第一存储器；若矩阵乘结果参与下一矩阵乘操作中的每一轮计算时需以遍历方式参与计算，则计算指令设定有第二存储器；否则，计算指令设定有第一存储器。

可选地，第一缓存器件被设置为：在矩阵乘计算器件计算矩阵乘结果的过程中，从第一存储器中加载参与当前矩阵乘操作的第一矩阵中的剩余未加载数据；

相应地，第二缓存器件被设置为：在矩阵乘计算器件计算矩阵乘结果的过程中，从第二存储器中加载参与当前矩阵乘操作的第二矩阵中的剩余未加载数据。

可选地，第一缓存器件利用乒乓方式加载第一矩阵中的数据。

可选地，第二缓存器件利用先入先出队列加载第二矩阵中的数据。

可选地，矩阵乘计算器件包括：至少一个计算单元，计算单元利用乘法器和累加器构成且每次计算的数据量等于 N。

第三方面，本申请提供了一种计算方法，包括：

针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；

将计算指令发送至前文的加速器，以使加速器对图神经网络进行加速计算。

可选地，将计算指令发送至前文的加速器，包括：

按照图神经网络中不同矩阵乘操作的计算顺序将相应计算指令按序发送至加速器。

可选地，还包括：

根据加速器中的计算核的资源配置确定并行数 N 的取值。

可选地，根据加速器中的计算核的资源配置确定并行数 N 的取值，包括：

基于计算核的资源配置构建约束公式：$Count \geq N3/2$、$BW \geq DW \times f \times N2$；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW 表示计算核的数据位宽，f 为计算核的计算时钟频率；

基于约束公式确定 N 的取值。

第四方面，本申请提供了一种计算装置，包括：

生成模块，用于针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；

发送模块，用于将计算指令发送至前文的加速器，以使加速器对图神经网络进行加速计算。

可选地，发送模块被设置为：

按照图神经网络中不同矩阵乘操作的计算顺序将相应计算指令按序发送至加速器。

可选地，还包括：

配置模块，用于根据加速器中的计算核的资源配置确定并行数 N 的取值。

可选地，配置模块被设置为：

基于计算核的资源配置构建约束公式：Count≥N3/2、BW≥DW×f×N2；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW 表示计算核的数据位宽，f 为计算核的计算时钟频率；

基于约束公式确定 N 的取值。

第五方面，本申请提供了一种电子设备，包括：

存储器，用于存储计算机程序；

处理器，用于执行计算机程序，以实现前述公开的计算方法。

第六方面，本申请提供了一种非易失性可读存储介质，用于保存计算机程序，其中，计算机程序被处理器执行时实现前述公开的计算方法。

第七方面，本申请提供了一种计算系统，包括：服务器以及至少一个前述公开的加速器。

通过以上方案可知，本申请提供了一种计算核，包括：第一缓存器件、第二缓存器件、矩阵乘计算器件和输出器件；第一缓存器件，用于根据接收到的计算指令从外部的第一存储器中加载参与图神经网络中的当前矩阵乘操作的第一矩阵中的数据；第二缓存器件，用于根据计算指令从外部的第二存储器中加载参与当前矩阵乘操作的第二矩阵中的数据；矩阵乘计算器件，用于按照计算指令设定的并行数 N 并行计算第一矩阵中的 N 个行数据或 N 个列数据与第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果；输出器件，用于暂存矩阵乘结果，在矩阵乘结果的数据量满足预设写条件时，将矩阵乘结果写入计算指令设定的第一存储器或第二存储器；其中，第二存储器的传输带宽大于第一存储器；若矩阵乘结果参与下一矩阵乘操作中的每一轮计算时需以遍历方式参与计算，则计算指令设定有第二存储器；否则，计算指令设定有第一存储器。

可见，本申请提供的计算核能够并行计算第一矩阵中的 N 个行数据或 N 个列数据与第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果，因此可一次性得到 N 个最终矩阵乘结果，提高了计算效率和速度；计算核也无需暂存中间结果，节约了片上资源。在得到矩阵乘结果后，本申请能够按照该矩阵乘结果参与下一矩阵乘操作中的每一轮计算时的参与方式确定将该矩阵乘结果存入哪个存储器，可选的，当矩阵乘结果参与下一矩阵乘操作中的每一轮计算时需以遍历方式参与计算，则将该矩阵乘结果存入第二存储器；否则，该矩阵乘结果存入第一存储器。由此可实现：使矩阵乘结果在存储器中的存储格式与其参与计算时的输出格式一致，便于在连续的计算过程中顺序读取数据，也无需进行矩阵转置，因此还能降低访存的时间

开销，提高效率。在满足预设写条件时，再写入矩阵乘结果，可降低写入次数，提升写入效率和带宽利用率。

相应地，本申请提供的一种加速器、计算方法、装置、设备、非易失性可读存储介质及系统，也同样具有上述技术效果。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

图 1 为本申请公开的一种计算核示意图；

图 2 为本申请公开的一种计算单元结构示意图；

图 3 为本申请公开的一种加速器结构示意图；

图 4 为本申请公开的另一种加速器结构示意图；

图 5 为本申请公开的第三种加速器结构示意图；

图 6 为本申请公开的一种计算方法流程图；

图 7 为本申请公开的一种数据并行输入示意图；

图 8 为本申请公开的一种 4*4*4 的 PE 阵列示意图；

图 9 本申请公开的一种写入数据连续存储示意图；

图 10 为本申请公开的一种矩阵乘计算流程示意图；

图 11 为本申请公开的一种电子设备示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

目前，现有的计算器件可以采用矩阵分块操作进行矩阵乘操作，但此方式需要将大量的中间结果暂存到计算器件之外，计算器件在进行后续计算时再读回至片内进行累加，增加了许多额外的数据搬移操作，会浪费计算资源，降低计算速度。为此，本申请提供了一种方案，能够提高图神经网络中矩阵乘操作的计算速度和计算效率。

参见图 1 所示，本申请实施例公开了一种计算核，包括：第一缓存器件、第二缓存器件、矩阵乘计算器件和输出器件。

其中，第一缓存器件，被设置为根据接收到的计算指令从外部的第一存储器中加载参与图神经网络中的当前矩阵乘操作的第一矩阵中的数据；第二缓存器件，被设置为根据计算指令从外部的第二存储器中加

载参与当前矩阵乘操作的第二矩阵中的数据；矩阵乘计算器件，被设置为按照计算指令设定的并行数 N 并行计算第一矩阵中的 N 个行数据或 N 个列数据与第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果；输出器件，被设置为暂存矩阵乘结果，在矩阵乘结果的数据量满足预设写条件时，将矩阵乘结果写入计算指令设定的第一存储器或第二存储器。其中，一个行数据为：矩阵中的一行数据；一个列数据为：矩阵中的一列数据。

其中，第二存储器（如 HBM）的传输带宽大于第一存储器（如 DDR）；若矩阵乘结果参与下一矩阵乘操作中的每一轮计算时需以遍历方式参与计算，则计算指令设定有第二存储器；否则，计算指令设定有第一存储器。HBM（High Bandwidth Memory）是一种适用于高带宽的 CPU/GPU 内存芯片。DDR（Double Data Rate，双倍速率同步动态随机存储器）相比与 HBM，适用带宽较低。

其中，计算指令用于实现图神经网络中的任一矩阵乘操作，其中的并行数 N 的取值根据计算核的资源配置确定。可选的，基于计算核的资源配置构建约束公式：Count≥N3/2、BW≥DW×f×N2；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW 表示计算核的数据位宽，f 为计算核的计算时钟频率；基于约束公式确定 N 的取值。

在一种可选的实施方式中，第一缓存器件被设置为：在矩阵乘计算器件计算矩阵乘结果的过程中，从第一存储器中加载参与当前矩阵乘操作的第一矩阵中的剩余未加载数据；相应地，第二缓存器件被设置为：在矩阵乘计算器件计算矩阵乘结果的过程中，从第二存储器中加载参与当前矩阵乘操作的第二矩阵中的剩余未加载数据。由此便无需等待上一操作完成，就能提前进行数据加载，计算器件也不会出现空闲状态，能持续进行数据计算，提高了计算资源利用率和系统计算效率，降低任务延迟。

在一种可选的实施方式中，第一缓存器件利用乒乓方式加载第一矩阵中的数据。第一缓存器件可以为 URAM（Ultra（超高效） RAM）。

在一种可选的实施方式中，第二缓存器件利用先入先出队列加载第二矩阵中的数据。可选的，第二缓存器件利用第一先入先出队列缓存第二矩阵中的数据，利用第二先入先出队列缓存矩阵乘结果。

在一种可选的实施方式中，矩阵乘计算器件包括：至少一个计算单元，计算单元利用乘法器和累加器构成且每次计算的数据量等于 N。请参见图 2，任一个计算单元可以包括至少一个乘法器组，一个乘法器组中包括两个乘法器；并且，一个乘法器组的输出连接累加器。如图 2 所示，每次计算的数据量为 4，故有 4 个乘法器组。如有需要，乘法器组及其输出所连累加器可灵活增设或删除，由此可实现计算器件的灵活扩展和调整。图 2 中的量化截取用于实现：数据位宽的量化与截取，如果数据位宽为 8，那么乘法结果的位宽就为 16，再经过累加，数据位宽会更大，此时需要将最终结果的数据位宽截取为 8 位。

其中，每次计算的数据量等于 N=行并行数=列并行数，可以使参与计算的行/列矩阵在内存的存储格式一致，便于图神经网络中每步计算结果的行优先/列优先转存（左右矩阵）变换，使得在连续的计算过程中可以顺序访存。

本实施例提供的计算核能够并行计算第一矩阵中的 N 个行数据或 N 个列数据与第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果，因此可一次性得到 N 个最终矩阵乘结果，提高了计算效率和速度；计算核也无需暂存中间结果，节约了片上资源。

本申请在得到矩阵乘结果后，能够按照该矩阵乘结果参与下一矩阵乘操作中的每一轮计算时的参与方式确定将该矩阵乘结果存入哪个存储器，可选的，当矩阵乘结果参与下一矩阵乘操作中的每一轮计算时需以遍历方式参与计算，则将该矩阵乘结果存入第二存储器；否则，该矩阵乘结果存入第一存储器。由此可

实现：使矩阵乘结果在存储器中的存储格式与其参与计算时的输出格式一致，便于在连续的计算过程中顺序读取数据，也无需进行矩阵转置，因此还能降低访存的时间开销，提高效率。而在满足预设写条件时，再写入矩阵乘结果，可降低写入次数，提升写入效率和带宽利用率。可见，本实施例能够提高图神经网络中矩阵乘操作的计算速度和计算效率。

下面对本申请实施例提供的一种加速器进行介绍，下文描述的一种加速器与上文描述的一种计算核可以相互参照。

本申请实施例公开了一种加速器，包括：控制器、第一存储器、传输带宽大于第一存储器的第二存储器、以及前文的计算核。

其中，控制器与第一存储器、第二存储器、计算核分别连接；计算核与第一存储器、第二存储器分别连接。控制器，被设置为将图神经网络中的矩阵乘操作的计算指令发送至第一存储器、第二存储器和计算核；第一存储器，被设置为根据计算指令以直接内存访问 DMA(Direct Memory Access)方式获取并存储参与当前矩阵乘操作的第一矩阵；第二存储器，被设置为根据计算指令以 DMA 方式获取并存储参与当前矩阵乘操作的第二矩阵；计算核，被设置为根据计算指令计算第一矩阵和第二矩阵的矩阵乘结果。

其中，第二存储器（如 HBM）的传输带宽大于第一存储器（如 DDR），且以遍历方式参与计算的矩阵存入第二存储器，可使以遍历方式参与计算的矩阵的读取速度更快，能够提升计算速度。

其中，计算核包括：第一缓存器件、第二缓存器件、矩阵乘计算器件和输出器件；第一缓存器件能够在矩阵乘计算器件计算矩阵乘结果的过程中，从第一存储器中加载参与当前矩阵乘操作的第一矩阵中的剩余未加载数据；相应地，第二缓存器件能够在矩阵乘计算器件计算矩阵乘结果的过程中，从第二存储器中加载参与当前矩阵乘操作的第二矩阵中的剩余未加载数据。由此便无需等待上一操作完成，就能提前进行数据加载，计算器件也不会出现空闲状态，能持续进行数据计算，提高了计算资源利用率和系统计算效率，降低任务延迟。

参见图 3 所示，加速器中可设有多个计算核；相应地，需设置调度核来使控制器对这些计算核进行调度和控制。

在一种可选的实施方式中，控制器还被设置为：控制多个计算核执行同一矩阵乘操作。在一种可选的实施方式中，控制器被设置为：在控制多个计算核执行同一矩阵乘操作时，使每一计算核在同一计算周期内计算第一矩阵中的相同行数据或相同列数据与第二矩阵中的不同列数据或不同行数据的矩阵乘结果。例如：在同一计算周期（同一时间段）内，计算核 1 计算第一矩阵中第一行与第二矩阵中第一列的乘法结果，计算核 2 计算第一矩阵中第一行与第二矩阵中第二列的乘法结果，由此可实现：第一矩阵中第一行数据在多个计算核上的复用，可避免数据的重复读取，也能够提升计算效率。如图 3 所示，S 矩阵中的行数据 1（第 1 行数据）被计算核 1 和计算核 2 复用。

在一种可选的实施方式中，控制器被设置为：在控制多个计算核执行同一矩阵乘操作时，使每一计算核在同一计算周期内分别计算第一矩阵中的不同列数据或不同行数据与第二矩阵中的相同列数据或相同行数据的矩阵乘结果。例如：在同一计算周期（同一时间段）内，计算核 1 计算第一矩阵中第一行与第二矩阵中第一列的乘法结果，计算核 2 计算第一矩阵中第二行与第二矩阵中第一列的乘法结果，由此可实现：第二矩阵中第一行数据在多个计算核上的复用，可避免数据的重复读取，也能够提升计算效率。如图 4 所示，遍历矩阵被计算核 1 和计算核 2 复用。

在一种可选的实施方式中，控制器还被设置为：控制不同计算核在同一计算周期内执行不同矩阵乘操

作。例如：在同一计算周期（同一时间段）内，计算核 1 计算第一矩阵中第一行与第二矩阵中第一列的乘法结果，计算核 2 计算第三矩阵中第一行与第四矩阵中第一列的乘法结果，由此可实现：第一矩阵与第二矩阵的矩阵乘操作、第三矩阵与第四矩阵的矩阵乘操作同时执行，能够提高计算效率。如图 5 所示，计算核 1 计算矩阵 A 与矩阵 B 的乘法结果，计算核 2 计算矩阵 C 与矩阵 D 的乘法结果。

由于需要连续遍历整个遍历矩阵中的数据，因此可使用高带宽的 HBM 实现遍历矩阵的数据缓存；而使用单端口带宽低的 DDR4 缓存 S 矩阵中少量几行/列数据。使用高、低带宽的存储器除了分别给遍历矩阵和 S 矩阵提供合理的带宽之外，由于可在计算的同时继续加载后续数据，因此能够隐藏访存时间开销，提高系统整体吞吐效率。如图 3 所示，只需要加载 S 矩阵中少量几行/列数据于片上缓存，而后使用 FIFO(First In First Out,先进先出)流式加载遍历矩阵，便可实现 S 矩阵中的数据复用。也即：S 矩阵中的数据被多个计算核共用，既可以按照数据内存组织格式依次写回数据，又可避免共享外存接口的访问冲突。

在一种可选的实施方式中，计算指令用于实现图神经网络中的任意矩阵乘操作。图神经网络中的主要运算为：特征矩阵 F 与相应的小波矩阵 Wav、权重矩阵 W 相乘，用公式表示为：Wav×(F×W)。如该公式所示，特征矩阵 F 在与 W 相乘时，F 为乘号左边的左矩阵，而 F×W 的乘积与 Wav 相乘时，该乘积为乘号右边的右矩阵，这种位置变换就需要将 F×W 的乘积按照该乘积在下一次乘法操作中的位置提前确定该乘积是行优先存储和列优先存储。当前一次矩阵乘结果需要存为行优先矩阵用于下一次计算时，前一次矩阵乘过程为：保持左矩阵中 N 行数据不动，遍历右矩阵的所有列，得到 N 行乘所有列的结果后；再次计算 N+1~2N 行乘所有列的结果，直至计算完成；此过程在按序生成行优先矩阵的同时，对左矩阵的行数据进行了数据重用。当前一次矩阵乘结果需要存为列优先矩阵用于下一次计算时，前一次矩阵乘过程为：保持右矩阵中 N 列数据不动，遍历左矩阵的所有行，得到 N 列乘所有行的结果后；再次计算 N+1~2N 列乘所有行的结果，直至计算完成；此过程在按序生成列优先矩阵的同时，对右矩阵的列数据进行了数据重用。

其中，任一个计算核包括：第一缓存器件、第二缓存器件、矩阵乘计算器件和输出器件；矩阵乘计算器件包括：至少一个计算单元，该计算单元利用乘法器和累加器构成且每次计算的数据量等于 N。每次计算的数据量等于 N=行并行数=列并行数，可以使参与计算的行/列矩阵在内存的存储格式一致，便于图神经网络中每步计算结果的行优先/列优先转存（左右矩阵）变换，使得在连续的计算过程中可以顺序访存。通过灵活设置 N 的取值可调整计算加速比。可选的，并行数 N 的取值根据计算核的资源配置确定。例如：基于计算核的资源配置构建约束公式：Count≥N3/2、BW≥DW×f×N2；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW 表示计算核的数据位宽，f 为计算核的计算时钟频率；基于约束公式确定 N 的取值。

可见，本实施例提供了一种加速器，利用该加速器能够避免矩阵转置的访存和时间开销，实现计算过程中的数据重用，提高图神经网络中矩阵乘操作的计算速度和计算效率，还可实现计算器件的灵活并行扩展。

下面对本申请实施例提供的一种计算方法进行介绍，下文描述的一种计算方法与其他实施例可以相互参照。

参见图 6 所示，本申请实施例公开了一种计算方法，包括：

S601、针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置。

S602、将计算指令发送至前文的加速器，以使加速器对图神经网络进行加速计算。

在图神经网络中有很多节点，每一节点对应有一个特征矩阵。特征矩阵 F 与相应的小波矩阵 Wav、权重矩阵 W 相乘，用公式表示为：Wav×(F×W)。图神经网络中包括多个这样的乘法步骤，故需要按照图神经网络中不同矩阵乘操作的计算顺序确定相应计算指令的顺序，而后按序发送计算指令至加速器，以使加速器按序进行图神经网络中的矩阵乘操作。在一种可选的实施方式中，将计算指令发送至前文的加速器，包括：按照图神经网络中不同矩阵乘操作的计算顺序将相应计算指令按序发送至加速器。

在一种可选的实施方式中，根据加速器中的计算核的资源配置确定并行数 N 的取值。在一种可选的实施方式中，根据加速器中的计算核的资源配置确定并行数 N 的取值，包括：基于计算核的资源配置构建约束公式：Count≥N3/2、BW≥DW×f×N2；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW 表示计算核的数据位宽，f 为计算核的计算时钟频率；基于约束公式确定 N 的取值。可选的，求解满足 Count≥N3/2、BW≥DW×f×N2 的最大 N 值作为并行数。

可见，本实施例提供了一种计算方法，能够避免矩阵转置的访存和时间开销，实现计算过程中的数据重用，提高图神经网络中矩阵乘操作的计算速度和计算效率，还可实现计算器件的灵活并行扩展。

下面对本申请实施例提供的一种计算装置进行介绍，下文描述的一种计算装置与其他实施例可以相互参照。

本申请实施例公开了一种计算装置，包括：

生成模块，被设置为针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；

发送模块，被设置为将计算指令发送至前文的加速器，以使加速器对图神经网络进行加速计算。

在一种可选的实施方式中，发送模块被设置为：

按照图神经网络中不同矩阵乘操作的计算顺序将相应计算指令按序发送至加速器。

在一种可选的实施方式中，还包括：

配置模块，被设置为根据加速器中的计算核的资源配置确定并行数 N 的取值。

在一种可选的实施方式中，配置模块被设置为：

基于计算核的资源配置构建约束公式：Count≥N3/2、BW≥DW×f×N2；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW 表示计算核的数据位宽，f 为计算核的计算时钟频率；

基于约束公式确定 N 的取值。

其中，关于本实施例中各个模块、单元更加详细的工作过程可以参考其他实施例中公开的相应内容，在此不再进行赘述。

可见，本实施例提供了一种计算装置，能够避免矩阵转置的访存和时间开销，实现计算过程中的数据重用，提高图神经网络中矩阵乘操作的计算速度和计算效率，还可实现计算器件的灵活并行扩展。

基于上述任意实施例，下面提供一种用于图神经网络中矩阵乘操作的计算器件的设计方案。如图 7 所示，通过 3 个参数 kernel 长度、行并行数 R、列并行数 C 设计构成计算器件的 PE（Process Element，计算单元）阵列，且 kernel 长度=行并行数 R=列并行数 C。在图 7 中，矩阵中的一行数据或一列数据被划分为 7 个 K 大小的数据块，图 7 所示的 PE 阵列可同时计算输入给其的数据。也就是说，当 kernel 长度=行并行数 R=列并行数 C=8 时，同时输入 8 行数据和 8 列数据至 PE 阵列，使得 PE 阵列同时计算 8 行数据和 8 列数据的乘法结果。需要说明的是，图 7 示意的 7 个 K 大小的数据块仅为示例，在实际实现时，数据块的个数应为 2 的指数，即：一行数据或一列数据被划分为 2 的指数个 K 大小的数据块。

在本实施例中，kernel 长度、行并行数 R、列并行数 C 根据单一 PE 的 DSP 资源和访存带宽来设置；其中，DSP 使用数量为：CountDSP = K*R*(C/2)；K 为 kernel 长度，R 为行并行数，C 为列并行数；其访存带宽为：BWtraverse= DW*K*C*f；DW 为每个数据的位宽，K 为 kernel 长度，C 为列并行数，f 为 DSP 运行时钟频率。例如，数据为 int8，K=C=16，f 为 400M/s 时，BWtraverse= 8bit*16*16*400M/s=100GB/s。由此可确定并行数的取值。由于 kernel 长度=行并行数 R=列并行数 C，因此在实际实现时，上述公式可用 Count≥N3/2、BW≥DW×f×N2 代替，N=K=R=C。

在本实施例中，kernel 为参与计算的最小单位，单一 PE 采用乘法器和加法器结构，该结构的好处是：能够并行加速计算过程。如果行的长度为 L，那么整行/整列顺序计算的时间为 L，而使用图 2 所示结构，用时为 log2(L)。

在一种示例中，4*4*4 的 PE 阵列如图 8 所示。在图 8 中，每个 PE 中都存在一组数据的复用。

其中，每次计算的数据量等于 N=行并行数=列并行数，可以使参与计算的行/列矩阵在内存的存储格式一致，便于图神经网络中每步计算结果的行优先/列优先转存（左右矩阵）变换，使得在连续的计算过程中可以顺序访存。例如：4 行×4 列的 A 矩阵和 4 行×N 列的 B 矩阵相乘时，在 R=C=4 时，针对 A 矩阵的 1-4 行和 B 矩阵的 1-4 列数据，每时钟周期取 K 个数据，经过完整行/列遍历后，得到结果矩阵中 1-4 行 1-4 列的数据；然后再取 B 矩阵 5-8 列数据，和 A 矩阵 1-4 行数据再次遍历后，得到结果矩阵中 1-4 行 5-8 列的数据，如此循环，直到结果矩阵 1-4 行数据计算完成，再切换 A 矩阵 5-8 行数据继续进行遍历；这样结果矩阵就和 A 矩阵的数据排列一致，从而保证每次矩阵计算时输入、输出矩阵格式的一致性。本例属于行优先矩阵，列优先矩阵类似。可选的，结果矩阵的存储顺序如图 9 所示。

本实施例提供的计算核能够并行计算第一矩阵中的 N 个行数据或 N 个列数据与第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果，因此可一次性得到 N 个最终矩阵乘结果，提高了计算效率和速度；计算核也无需暂存中间结果，节约了片上资源。

请参见图 10，按照本实施例实现的矩阵乘操作流程包括：

（1）执行重用矩阵加载指令，从 DDR 读取 stationary 矩阵（S 矩阵，重用矩阵，即该矩阵相对于遍历矩阵来说，数据可重复使用）的一组行/列数据到 URAM，由于 URAM 的存储深度通常较大，可以在当前组计算过程中提前将下一组数据读入进行乒乓缓存；

（2）执行遍历矩阵加载指令，从 HBM 依次读取遍历矩阵的所有列/行数据到 FIFO；

（3）执行矩阵乘计算指令，PE 阵列进行计算；

（4）计算完成一组数据后，执行结果写回指令，根据指令中定义的地址通过总线 burst（突发）传输将结果写到对应的外部 HBM/DDR4；

（5）复位 PE 阵列中的累加器，以便进行下一次计算；

重复（1）-（4）步，直到 S 矩阵所有组数据计算完成。

其中，通过总线 burst 传输将结果写到对应的外部 HBM/DDR4，可以减少写入次数。也即：凑够一个 burst 长度，向外部存储器写一次。例如：如果当前加速器为基于 FPGA 的系统，其中所用的 AXI4（Advanced eXtensible Interface 4，一种总线协议）的 burst 长度为 4KB，而对于 int8 数据类型，需要凑够 4096 个数据向部存储器写一次。

可见，本实施例使用定量设计方法，根据片上资源及外部可提供的访存带宽，得出 PE 阵列的组织形式及规模；使用行/列优先矩阵的生成方法生成图神经网络下一步计算所需格式的特征矩阵；使用完整的硬件

加速系统及方法进行数据加载、遍历计算及结果写回；可灵活扩展的多 PE 阵列系统架构用于单个或多个矩阵乘的并行计算加速；将操作指令拆分为数据加载、计算和结果写回指令，提前进行数据加载，保证 PE 阵列连续不间断工作。该方案可灵活适配硬件资源，实现指令周期性流水处理，有效提升计算和访存效率，完成图神经网络中各级矩阵的灵活变换和连续高效计算。

下面对本申请实施例提供的一种电子设备进行介绍，下文描述的一种电子设备与其他实施例可以相互参照。

参见图 11 所示，本申请实施例公开了一种电子设备，包括：

存储器 1101，被设置为保存计算机程序；

处理器 1102，被设置为执行计算机程序，以实现上述任意实施例公开的方法。

在本实施例中，处理器执行存储器中保存的计算机程序时，可以实现以下步骤：针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；将计算指令发送至前文的加速器，以使加速器对图神经网络进行加速计算。

在本实施例中，处理器执行存储器中保存的计算机程序时，可以实现以下步骤：按照图神经网络中不同矩阵乘操作的计算顺序将相应计算指令按序发送至加速器。

在本实施例中，处理器执行存储器中保存的计算机程序时，可以实现以下步骤：根据加速器中的计算核的资源配置确定并行数 N 的取值。

在本实施例中，处理器执行存储器中保存的计算机程序时，可以实现以下步骤：基于计算核的资源配置构建约束公式：Count≥N3/2、BW≥DW×f×N2；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW 表示计算核的数据位宽，f 为计算核的计算时钟频率；基于约束公式确定 N 的取值。

可选的，本申请实施例还提供了一种服务器来作为上述电子设备。该服务器，可以包括：至少一个处理器、至少一个存储器、电源、通信接口、输入输出接口和通信总线。其中，存储器被设置为存储计算机程序，计算机程序由处理器加载并执行，以实现前述任一实施例公开的计算方法中的相关步骤。

本实施例中，电源被设置为服务器上的各硬件设备提供工作电压；通信接口能够为服务器创建与外界设备之间的数据传输通道，其所遵循的通信协议是能够适用于本申请技术方案的任意通信协议，在此不对其进行限定；输入输出接口，被设置为获取外界输入数据或向外界输出数据，其接口类型可以根据应用需要进行选取，在此不进行限定。

另外，存储器作为资源存储的载体，可以是只读存储器、随机存储器、磁盘或者光盘等，其上所存储的资源包括操作系统、计算机程序及数据等，存储方式可以是短暂存储或者永久存储。

其中，操作系统被设置为管理与控制服务器上的各硬件设备以及计算机程序，以实现处理器对存储器中数据的运算与处理，其可以是 Windows Server、Netware、Unix、Linux 等。计算机程序除了包括能够被设置为完成前述任一实施例公开的计算方法的计算机程序之外，还可以包括能够被设置为完成其他特定工作的计算机程序。数据除了可以包括虚拟机等数据外，还可以包括虚拟机的开发商信息等数据。

可选的，本申请实施例还提供了一种终端来作为上述电子设备。该终端可以包括但不限于智能手机、平板电脑、笔记本电脑或台式电脑等。

通常，本实施例中的终端包括有：处理器和存储器。

其中，处理器可以包括一个或多个处理核心，比如 4 核心处理器、8 核心处理器等。处理器可以采用 DSP(Digital Signal Processing，数字信号处理)、FPGA(Field－Programmable Gate Array，现场可编程

门阵列)、PLA(Programmable Logic Array，可编程逻辑阵列)中的至少一种硬件形式来实现。处理器也可以包括主处理器和协处理器，主处理器是被设置为对在唤醒状态下的数据进行处理的处理器，也称CPU(Central Processing Unit，中央处理器)；协处理器是被设置为对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器可以在集成有 GPU(Graphics Processing Unit，图像处理器)，GPU被设置为负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器还可以包括 AI(Artificial Intelligence，人工智能)处理器，该AI处理器被设置为处理有关机器学习的计算操作。

存储器可以包括一个或多个非易失性可读存储介质，该非易失性可读存储介质可以是非暂态的。存储器还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。本实施例中，存储器至少被设置为存储以下计算机程序，其中，该计算机程序被处理器加载并执行之后，能够实现前述任一实施例公开的由终端侧执行的计算方法中的相关步骤。另外，存储器所存储的资源还可以包括操作系统和数据等，存储方式可以是短暂存储或者永久存储。其中，操作系统可以包括 Windows、Unix、Linux等。数据可以包括但不限于应用程序的更新信息。

在一些实施例中，终端还可包括有显示屏、输入输出接口、通信接口、传感器、电源以及通信总线。

下面对本申请实施例提供的一种非易失性可读存储介质进行介绍，下文描述的一种非易失性可读存储介质与其他实施例可以相互参照。

本申请实施例公开了一种非易失性可读存储介质，被设置为保存计算机程序，其中，计算机程序被处理器执行时实现前述实施例公开的计算方法。其中，非易失性可读存储介质作为资源存储的载体，可以是只读存储器、随机存储器、磁盘或者光盘等，其上所存储的资源包括操作系统、计算机程序及数据等，存储方式可以是短暂存储或者永久存储。

在本实施例中，处理器执行的计算机程序，可以实现以下步骤：针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；将计算指令发送至前文的加速器，以使加速器对图神经网络进行加速计算。

在本实施例中，处理器执行的计算机程序，可以实现以下步骤：按照图神经网络中不同矩阵乘操作的计算顺序将相应计算指令按序发送至加速器。

在本实施例中，处理器执行的计算机程序，可以实现以下步骤：根据加速器中的计算核的资源配置确定并行数N的取值。

在本实施例中，处理器执行的计算机程序，可以实现以下步骤：基于计算核的资源配置构建约束公式：Count≥N3/2、BW≥DW×f×N2；其中，Count 表示计算核的计算资源数量，BW 表示计算核的访存带宽，DW表示计算核的数据位宽，f为计算核的计算时钟频率；基于约束公式确定N的取值。

下面对本申请实施例提供的一种计算系统进行介绍，下文描述的一种计算系统与其他实施例可以相互参照。

本申请实施例公开了一种计算系统，包括：服务器以及至少一个前述公开的加速器。服务器用于针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；将计算指令发送至任一个加速器。接收到计算指令的加速器对图神经网络进行加速计算。

可见，本实施例提供了一种计算系统，能够避免矩阵转置的访存和时间开销，实现计算过程中的数据重用，提高图神经网络中矩阵乘操作的计算速度和计算效率，还可实现计算器件的灵活并行扩展。

本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其它实施例的不同之处，

各个实施例之间相同或相似部分互相参见即可。

结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程 ROM、电可擦除可编程 ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的非易失性可读存储介质中。

本文中应用了可选的个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在可选的实施方式及应用范围上均会有改变之处，综上，本说明书内容不应理解为对本申请的限制。

权 利 要 求 书

1、一种计算核，其特征在于，包括：第一缓存器件、第二缓存器件、矩阵乘计算器件和输出器件；

所述第一缓存器件，被设置为根据接收到的计算指令从外部的第一存储器中加载参与图神经网络中的当前矩阵乘操作的第一矩阵中的数据；

所述第二缓存器件，被设置为根据所述计算指令从外部的第二存储器中加载参与当前矩阵乘操作的第二矩阵中的数据；

所述矩阵乘计算器件，被设置为按照所述计算指令设定的并行数 N 并行计算所述第一矩阵中的 N 个行数据或 N 个列数据与所述第二矩阵中的 N 个列数据或 N 个行数据的矩阵乘结果；

所述输出器件，被设置为暂存所述矩阵乘结果，在所述矩阵乘结果的数据量满足预设写条件时，将所述矩阵乘结果写入所述计算指令设定的所述第一存储器或所述第二存储器；

其中，所述第二存储器的传输带宽大于所述第一存储器；若所述矩阵乘结果参与下一矩阵乘操作中的每一轮计算时需以遍历方式参与计算，则所述计算指令设定有所述第二存储器；否则，所述计算指令设定有所述第一存储器。

2、根据权利要求 1 所述的计算核，其特征在于，所述第一缓存器件被设置为：在所述矩阵乘计算器件计算所述矩阵乘结果的过程中，从所述第一存储器中加载参与当前矩阵乘操作的第一矩阵中的剩余未加载数据；

相应地，所述第二缓存器件被设置为：在所述矩阵乘计算器件计算所述矩阵乘结果的过程中，从所述第二存储器中加载参与当前矩阵乘操作的第二矩阵中的剩余未加载数据。

3、根据权利要求 1 所述的计算核，其特征在于，所述第一缓存器件利用乒乓方式加载所述第一矩阵中的数据。

4、根据权利要求 1 所述的计算核，其特征在于，所述第二缓存器件利用先入先出队列加载所述第二矩阵中的数据。

5、根据权利要求 1 至 4 任一项所述的计算核，其特征在于，所述矩阵乘计算器件包括：至少一个计算单元，所述计算单元利用乘法器和累加器构成且每次计算的数据量等于 N。

6、一种加速器，其特征在于，包括：控制器、第一存储器、传输带宽大于所述第一存储器的第二存储器、以及如权利要求 1 至 5 任一项所述的计算核；

其中，所述控制器与所述第一存储器、所述第二存储器、所述计算核分别连接；所述计算核与所述第一存储器、所述第二存储器分别连接；

所述控制器，被设置为将图神经网络中的矩阵乘操作的计算指令发送至所述第一存储器、所述第二存储器和所述计算核；

所述第一存储器，被设置为根据所述计算指令以直接内存访问 DMA 方式获取并存储参与当前矩阵乘操作的第一矩阵；

所述第二存储器，被设置为根据所述计算指令以 DMA 方式获取并存储参与当前矩阵乘操作的第二矩阵；

所述计算核，被设置为根据所述计算指令计算所述第一矩阵和所述第二矩阵的矩阵乘结果。

7、根据权利要求 6 所述的加速器，其特征在于，所述计算核有多个。

8、根据权利要求 7 所述的加速器，其特征在于，所述控制器还被设置为：控制多个计算核执行同一矩阵乘操作。

9、根据权利要求 8 所述的加速器，其特征在于，所述控制器被设置为：在控制多个计算核执行同一矩阵乘操作时，使每一计算核在同一计算周期内计算所述第一矩阵中的相同行数据或相同列数据与所述第二矩阵中的不同列数据或不同行数据的矩阵乘结果。

10、根据权利要求 8 所述的加速器，其特征在于，所述控制器被设置为：在控制多个计算核执行同一矩阵乘操作时，使每一计算核在同一计算周期内分别计算所述第一矩阵中的不同列数据或不同行数据与所述第二矩阵中的相同列数据或相同行数据的矩阵乘结果。

11、根据权利要求 7 所述的加速器，其特征在于，所述控制器还被设置为：控制不同计算核在同一计算周期内执行不同矩阵乘操作。

12、根据权利要求 6 所述的加速器，其特征在于，所述计算指令用于实现图神经网络中的任意矩阵乘操作。

13、一种计算方法，其特征在于，包括：

针对图神经网络中的每一矩阵乘操作生成计算指令；所述计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；

将所述计算指令发送至加速器，以使所述加速器对所述图神经网络进行加速计算；所述加速器为权利要求 6 至 12 任一项所述的加速器。

14、根据权利要求 13 所述的方法，其特征在于，所述将所述计算指令发送至加速器，包括：

按照所述图神经网络中不同矩阵乘操作的计算顺序将相应计算指令按序发送至所述加速器。

15、根据权利要求 13 所述的方法，其特征在于，还包括：

根据所述加速器中的计算核的资源配置确定并行数 N 的取值。

16、根据权利要求 15 所述的方法，其特征在于，所述根据所述加速器中的计算核的资源配置确定并行数 N 的取值，包括：

基于所述计算核的资源配置构建约束公式：$Count \geq N3/2$、$BW \geq DW \times f \times N2$；其中，Count 表示所述计算核的计算资源数量，BW 表示所述计算核的访存带宽，DW 表示所述计算核的数据位宽，f 为所述计算核的计算时钟频率；

基于所述约束公式确定 N 的取值。

17、一种计算装置，其特征在于，包括：

生成模块，被设置为针对图神经网络中的每一矩阵乘操作生成计算指令；所述计算指令中设有并行数 N 以及相应矩阵乘操作的矩阵乘结果的写入位置；

发送模块，被设置为将所述计算指令发送至加速器，以使所述加速器对所述图神经网络进行加速计算；所述加速器为权利要求 6 至 12 任一项所述的加速器。

18、一种电子设备，其特征在于，包括：

存储器，被设置为存储计算机程序；

处理器，被设置为执行所述计算机程序，以实现如权利要求 13 至 16 任一项所述的方法。

19、一种非易失性可读存储介质，其特征在于，被设置为保存计算机程序，其中，所述计算机程序被处理器执行时实现如权利要求 13 至 16 任一项所述的方法。

20、一种计算系统，其特征在于，包括：服务器以及至少一个如权利要求 6 至 12 任一项所述的加速器。

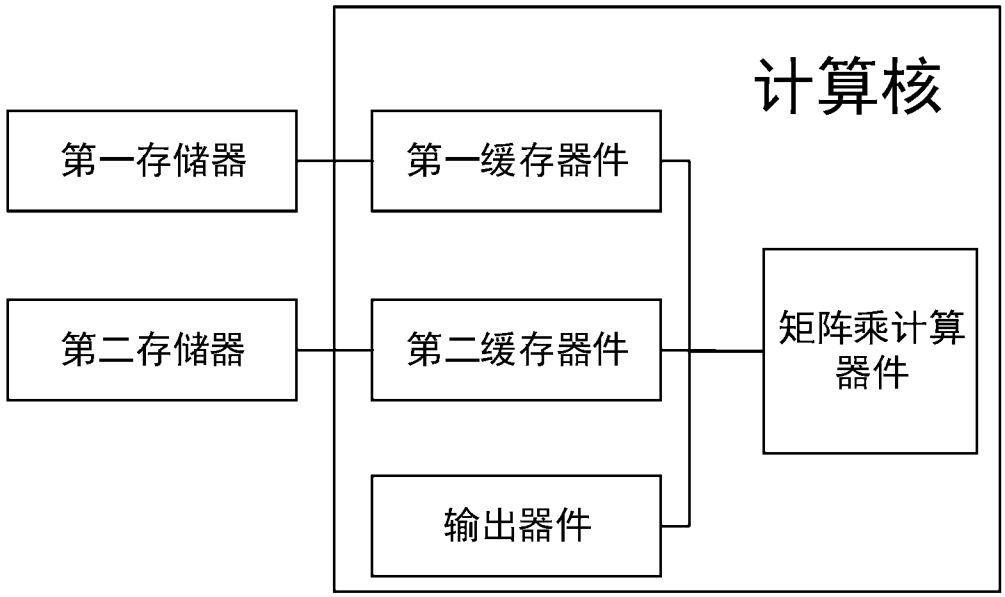

计算核
第一存储器
第一缓存器件
第二存储器
第二缓存器件
矩阵乘计算器件
输出器件

图 1

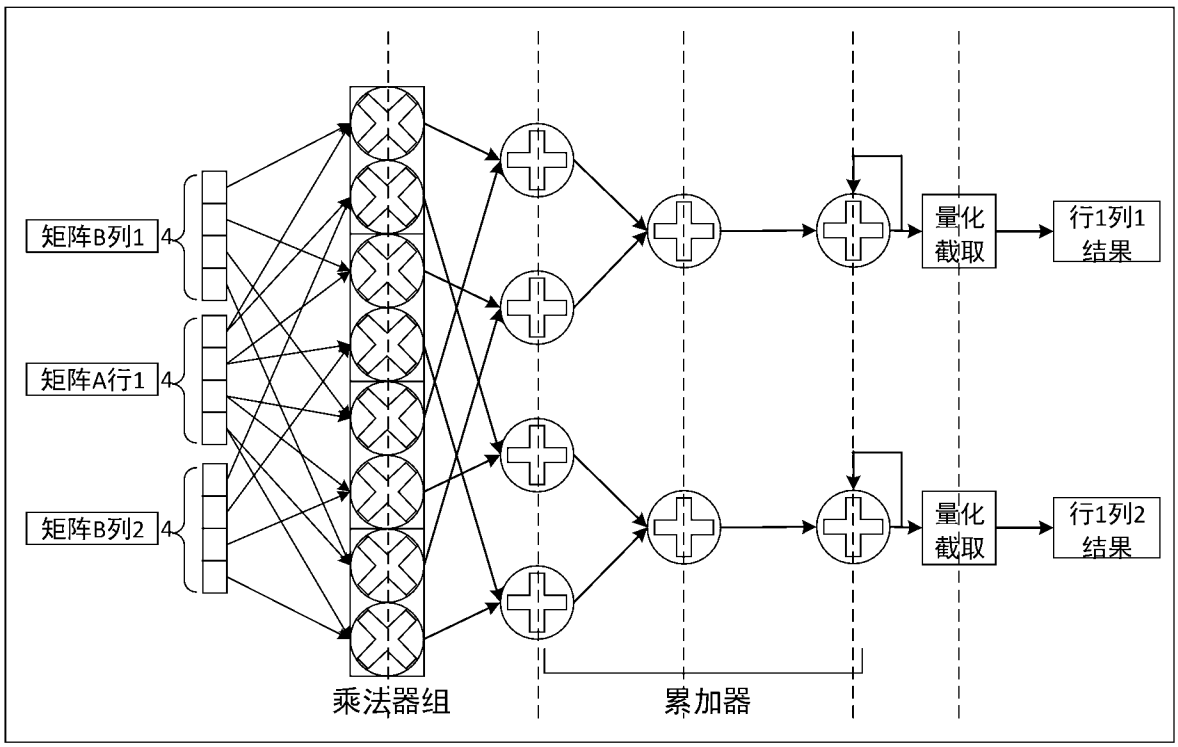

矩阵B列1
4
矩阵A行1
4
矩阵B列2
4
乘法器组
累加器
量化截取
行1列1结果
量化截取
行1列2结果

图 2

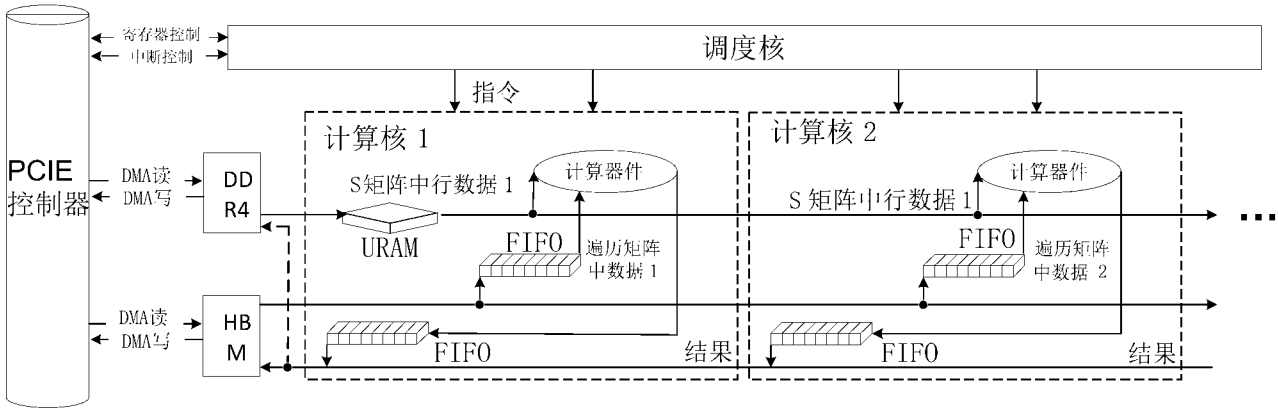

寄存器控制
中断控制
调度核
指令
PCIE
控制器
DMA读
DMA写
DD
R4
计算核 1
S矩阵中行数据 1
计算器件
URAM
FIFO
遍历矩阵
中数据 1
计算核 2
S 矩阵中行数据 1
计算器件
FIFO
遍历矩阵
中数据 2
DMA读
DMA写
HB
M
FIFO
结果
FIFO
结果

图 3

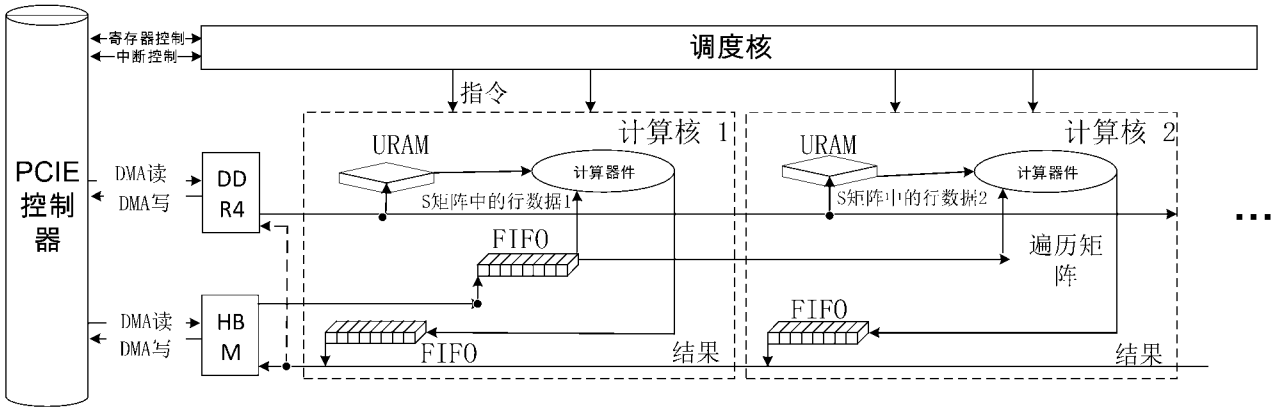

寄存器控制
中断控制
调度核
指令
PCIE
控制器
DMA读
DMA写
DD
R4
计算核 1
URAM
计算器件
S矩阵中的行数据1
FIFO
遍历矩阵
DMA读
DMA写
HB
M
FIFO
结果
计算核 2
URAM
计算器件
S矩阵中的行数据2
FIFO
结果

图 4

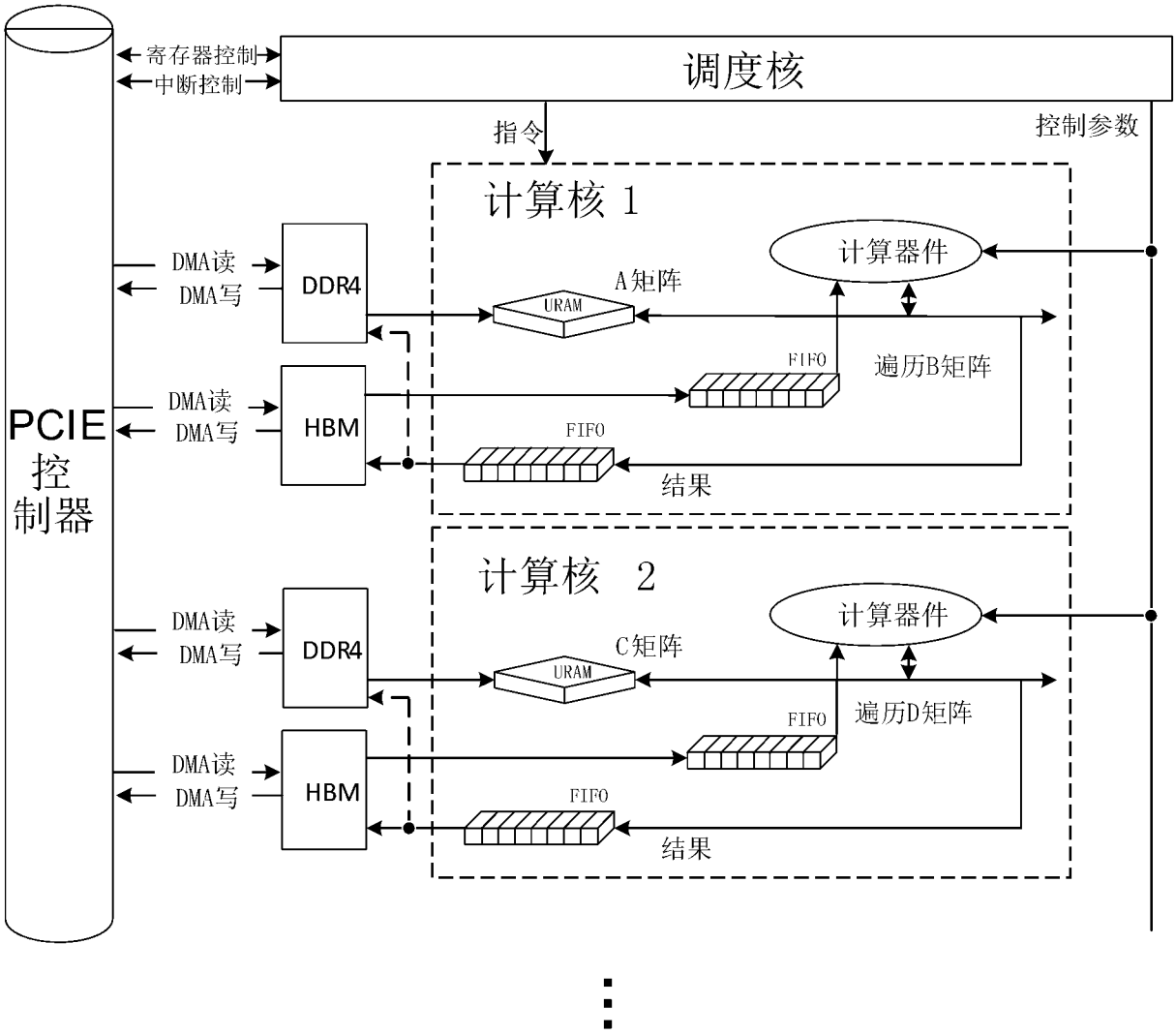

寄存器控制
中断控制
调度核
指令
控制参数
计算核 1
计算器件
DMA读
DMA写
DDR4
A矩阵
URAM
FIFO
遍历B矩阵
PCIE
控
制器
DMA读
DMA写
HBM
FIFO
结果
计算核 2
计算器件
DMA读
DMA写
DDR4
C矩阵
URAM
FIFO
遍历D矩阵
DMA读
DMA写
HBM
FIFO
结果

图 5

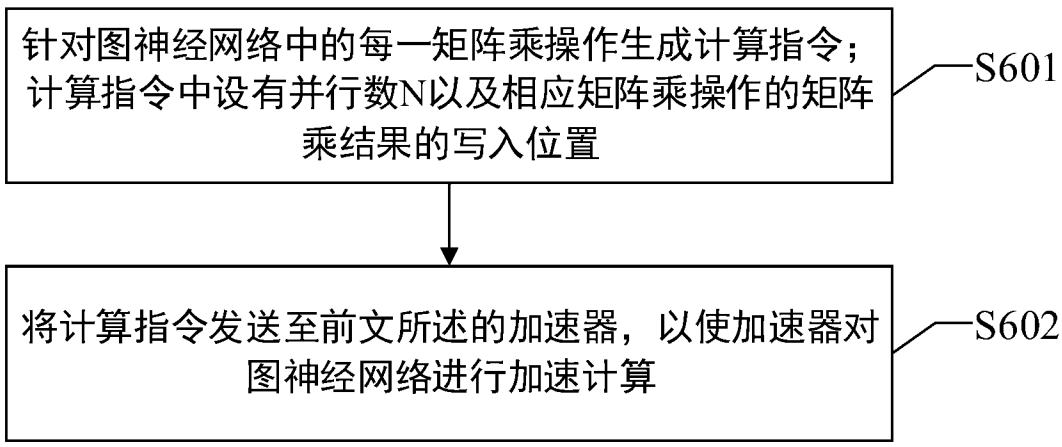

针对图神经网络中的每一矩阵乘操作生成计算指令；计算指令中设有并行数N以及相应矩阵乘操作的矩阵乘结果的写入位置
S601
将计算指令发送至前文所述的加速器，以使加速器对图神经网络进行加速计算
S602

图 6

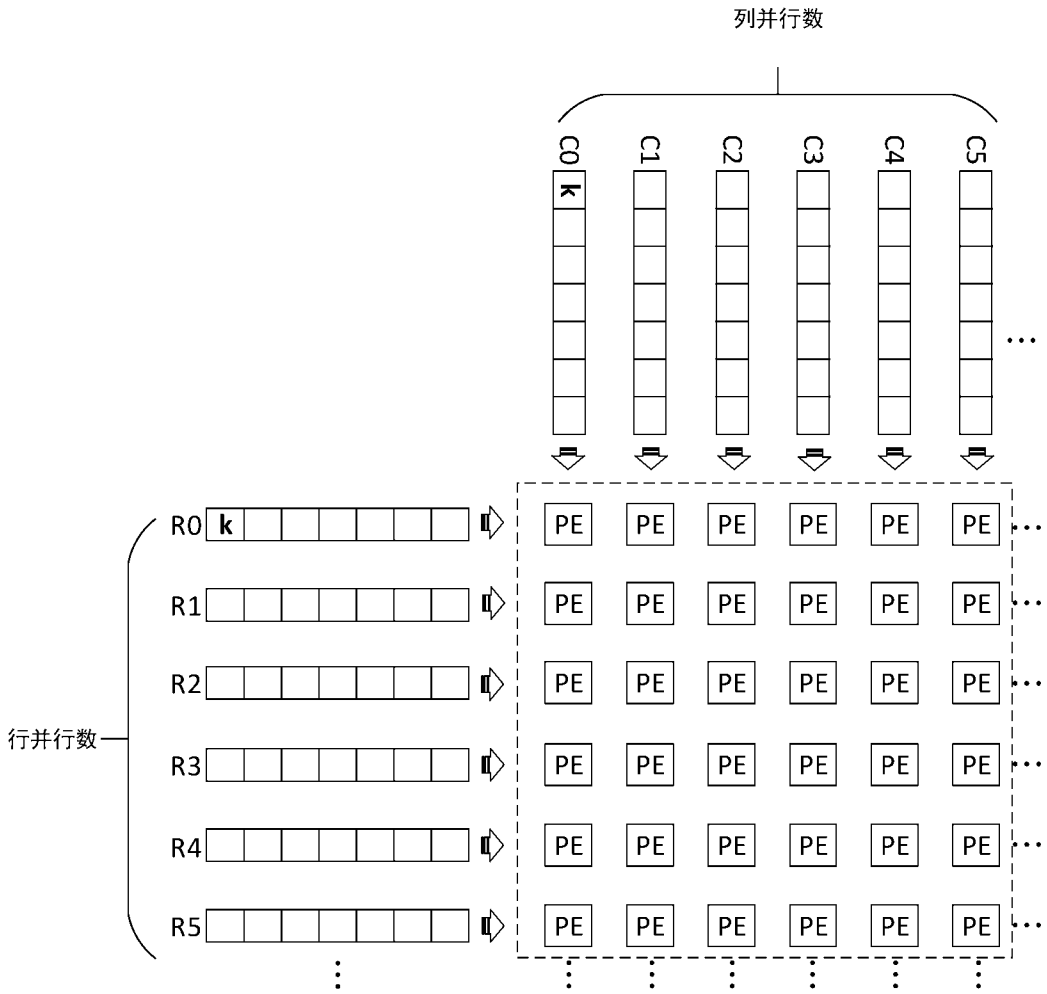


图 7

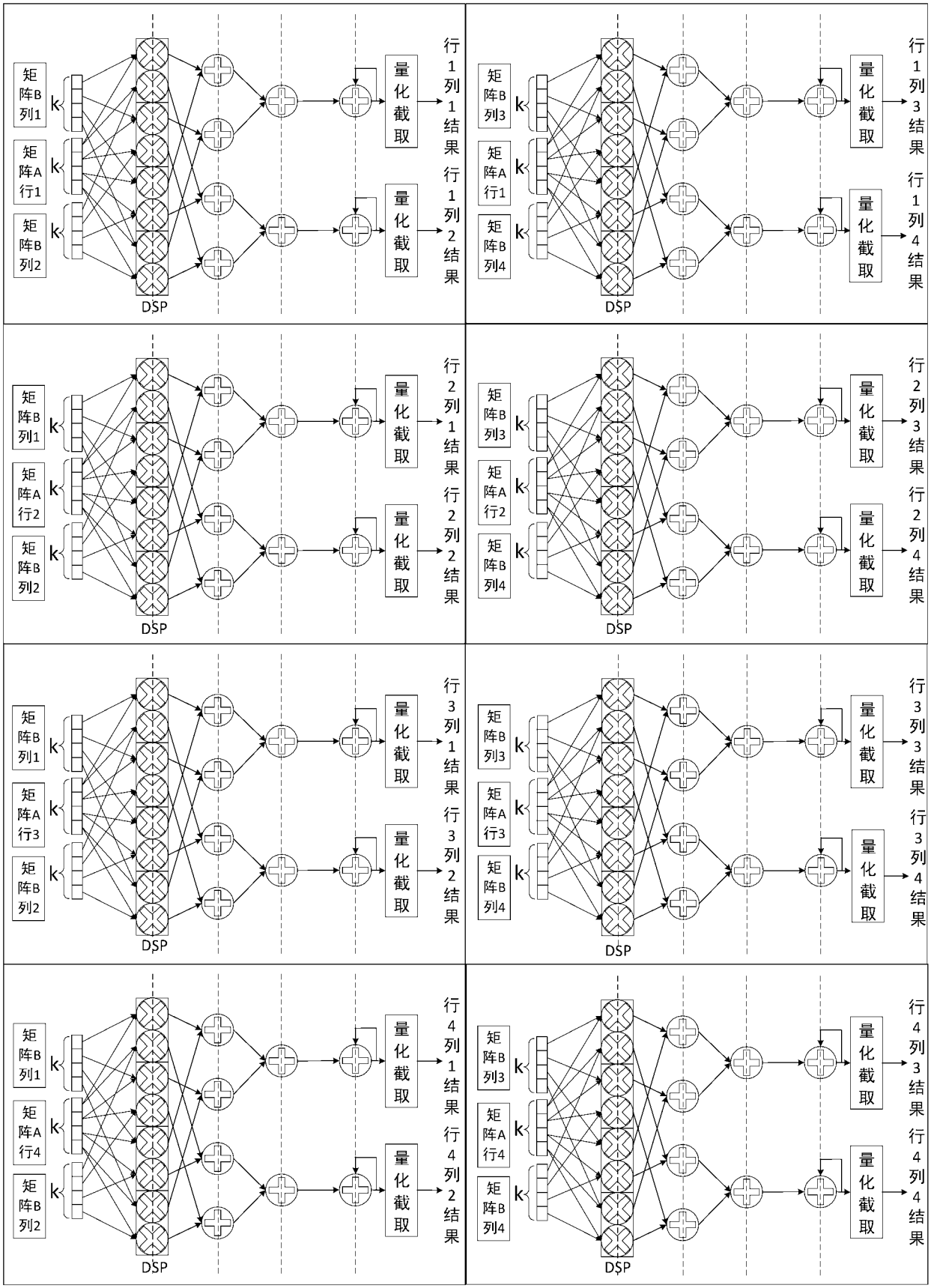

矩阵B列1
矩阵A行1
矩阵B列2
矩阵B列3
矩阵B列4
k
DSP
量化截取
行1列1结果
行1列2结果
行1列3结果
行1列4结果
矩阵A行2
行2列1结果
行2列2结果
行2列3结果
行2列4结果
矩阵A行3
行3列1结果
行3列2结果
行3列3结果
行3列4结果
矩阵A行4
行4列1结果
行4列2结果
行4列3结果
行4列4结果

图 8

第1行第1列	第1行第2列	第1行第3列	第1行第4列	第2行第1列	第2行第2列	第2行第3列	第2行第4列
第3行第1列	第3行第2列	第3行第3列	第3行第4列	第4行第1列	第4行第2列	第4行第3列	第4行第4列
第1行第5列	第1行第6列	第1行第7列	第1行第8列	第2行第5列	第2行第6列	第2行第7列	第2行第8列
第3行第5列	第3行第6列	第3行第7列	第3行第8列	第4行第5列	第4行第6列	第4行第7列	第4行第8列

⋮

图 9

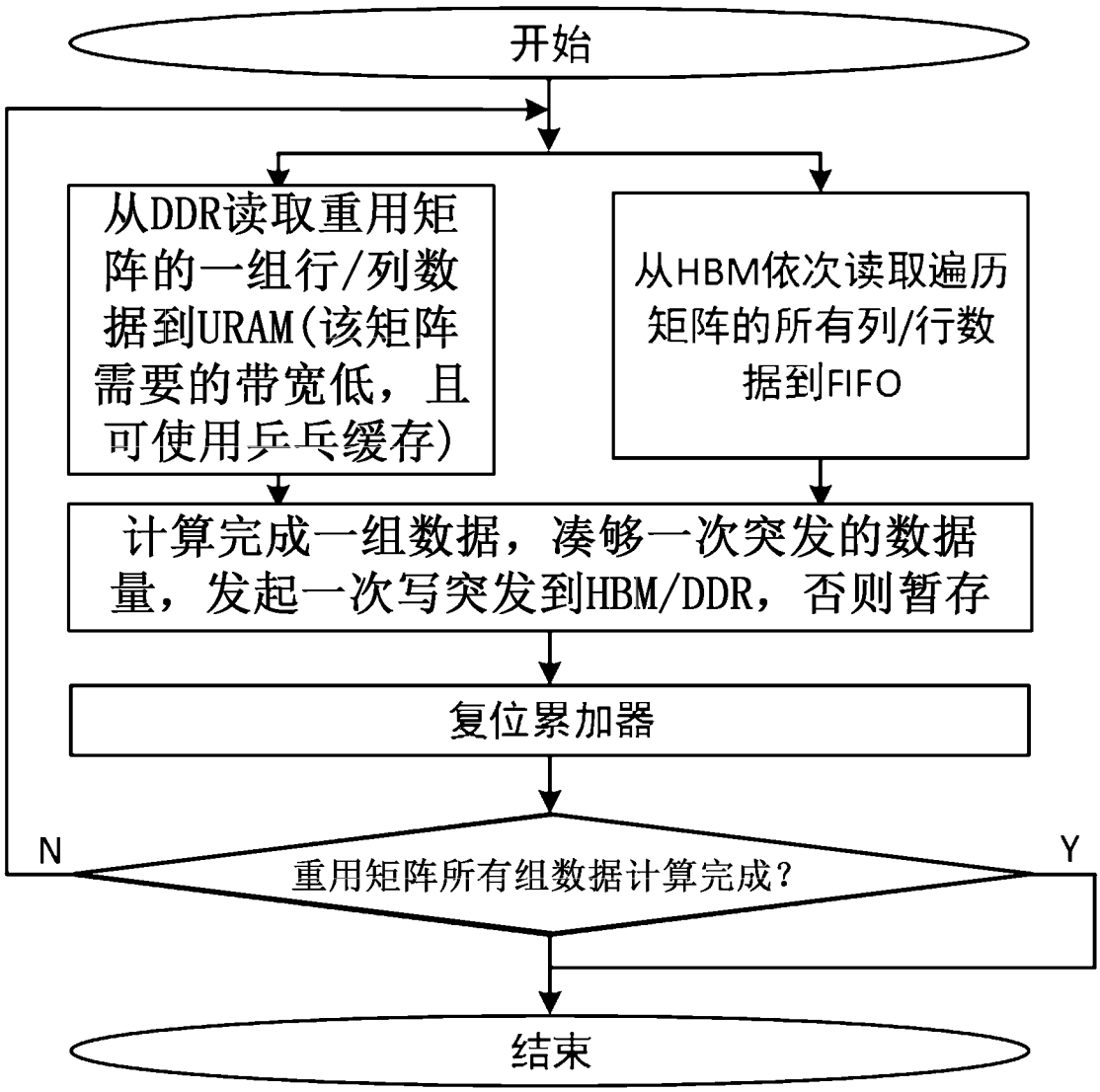


图 10

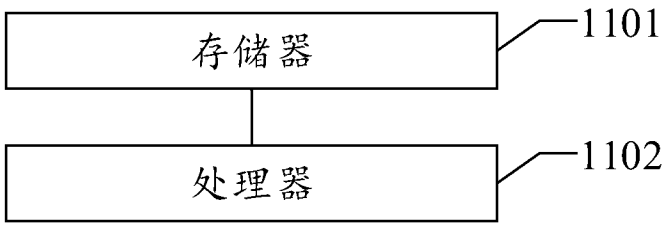


图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/132743

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/063(2023.01)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: G06N

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)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 加速, 神经网络, 图, 矩阵, 乘, 缓冲, 缓存, 存储器, 带宽, 位宽, 并行, 遍历, 复用, 循环, 迭代, accelerator, CNN, graph, buffer, cache, memory, storage, bit, width, parallel, traversal, reuse, iteration

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 115860080 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 28 March 2023 (2023-03-28) claims 1-20	1-20
A	CN 106445471 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 22 February 2017 (2017-02-22) description, paragraphs [0026]-[0065], and figures 1-6	1-20
A	CN 111967582 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 20 November 2020 (2020-11-20) entire document	1-20
A	CN 113947200 A (ZHUHAI SPACETOUCH TECHNOLOGY CO., LTD.) 18 January 2022 (2022-01-18) entire document	1-20
A	US 2022188613 A1 (THE GEORGE WASHINGTON UNIVERSITY) 16 June 2022 (2022-06-16) 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
17 January 2024	**22 January 2024**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Authorized officer Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 113918120 A (GUANGDONG OPPO MOBILE COMMUNICATIONS CO., LTD.) 11 January 2022 (2022-01-11) entire document	1-20
A	CN 109522125 A (ZHENGZHOU YUNHAI INFORMATION TECHNOLOGY CO., LTD.) 26 March 2019 (2019-03-26) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/132743

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 115860080 A	28 March 2023	None	
CN 106445471 A	22 February 2017	US 2018107630 A1	19 April 2018
CN 111967582 A	20 November 2020	None	
CN 113947200 A	18 January 2022	None	
US 2022188613 A1	16 June 2022	None	
CN 113918120 A	11 January 2022	None	
CN 109522125 A	26 March 2019	None	

国际检索报告

国际申请号
PCT/CN2023/132743

A. 主题的分类

G06N 3/063(2023.01)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: G06N

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））

VEN，CNABS，CNTXT，WOTXT，EPTXT，USTXT，CNKI，IEEE：加速，神经网络，图，矩阵，乘，缓冲，缓存，存储器，带宽，位宽，并行，遍历，复用，循环，迭代，accelerator，CNN，graph，buffer，cache，memory，storage，bit，width，parallel，traversal，reuse，iteration

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
PX	CN 115860080 A (苏州浪潮智能科技有限公司) 2023年3月28日 (2023 - 03 - 28) 权利要求1-20	1-20
A	CN 106445471 A (北京百度网讯科技有限公司) 2017年2月22日 (2017 - 02 - 22) 说明书第[0026]-[0065]段、图1-6	1-20
A	CN 111967582 A (苏州浪潮智能科技有限公司) 2020年11月20日 (2020 - 11 - 20) 全文	1-20
A	CN 113947200 A (珠海普林芯驰科技有限公司) 2022年1月18日 (2022 - 01 - 18) 全文	1-20
A	US 2022188613 A1 (THE GEORGE WASHINGTON UNIVERSITY) 2022年6月16日 (2022 - 06 - 16) 全文	1-20
A	CN 113918120 A (OPPO广东移动通信有限公司) 2022年1月11日 (2022 - 01 - 11) 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
2024年1月17日	2024年1月22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	受权官员 董洪梅 电话号码 (+86) 010-53961528

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 109522125 A (郑州云海信息技术有限公司) 2019年3月26日 (2019 - 03 - 26) 全文	1-20

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/132743

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 115860080 A	2023年3月28日	无	
CN 106445471 A	2017年2月22日	US 2018107630 A1	2018年4月19日
CN 111967582 A	2020年11月20日	无	
CN 113947200 A	2022年1月18日	无	
US 2022188613 A1	2022年6月16日	无	
CN 113918120 A	2022年1月11日	无	
CN 109522125 A	2019年3月26日	无	