



(21) 申请号 201780095094.7

(22) 申请日 2017.09.21

(65) 同一申请的已公布的文献号
申请公布号 CN 111149090 A

(43) 申请公布日 2020.05.12

(85) PCT国际申请进入国家阶段日
2020.03.20

(86) PCT国际申请的申请数据
PCT/EP2017/073854 2017.09.21

(87) PCT国际申请的公布数据
W02019/057281 EN 2019.03.28

(73) 专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72) 发明人 塔尔·霍洛维茨 乌里·韦泽
吴祖光 罗会斌 约尼·乔克伦

(74) 专利代理机构 北京三高永信知识产权代理有限公司 11138

专利代理师 颜晶

(51) Int.Cl.

G06F 9/48 (2006.01)

G06F 9/38 (2018.01)

G06N 3/063 (2023.01)

(56) 对比文件

TW 201706871 A, 2017.02.16

US 2017124025 A1, 2017.05.04

US 2017103314 A1, 2017.04.13

US 2006248317 A1, 2006.11.02

US 7584342 B1, 2009.09.01

WO 9526531 A1, 1995.10.05

R. Sernec 等. Multithreaded systolic/SIMD DSP array processor-MUS2DAP”, R. Sernec. 1997 IEEE Workshop on Signal Processing Systems. SiPS 97 Design and Implementation formerly VLSI Signal Processing. 1997, 第448-457页.

审查员 张诗伯

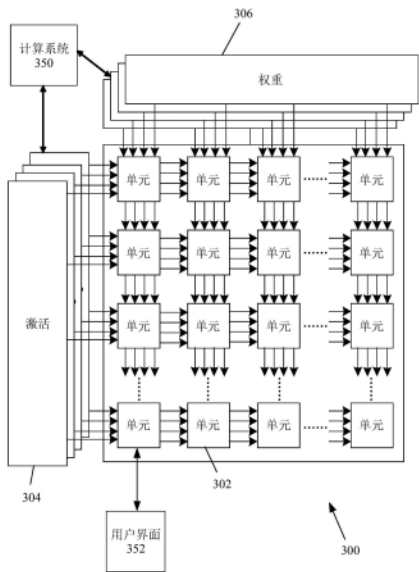
权利要求书2页 说明书11页 附图7页

(54) 发明名称

多线程脉动阵列

(57) 摘要

提供了一种多线程脉动阵列300, 包括: 多个处理元件302, 分别包括处理器308, 其中, 每个处理元件302用于: 从相应的第一输入源304接收多个第一输入, 从相应的第二输入源306接收多个第二输入, 其中, 多个第一输入和多个第二输入布置为与多个线程相对应的多对; 为处理器的每个操作周期调度多个线程中的线程; 对该线程执行计算操作。



1. 一种多线程脉动阵列(300), 其特征在于, 包括:
多个处理元件(302), 分别包括处理器(308), 其中, 每个所述处理元件(302)用于:
从相应的第一输入源(304)接收多个第一输入;
从相应的第二输入源(306)接收多个第二输入,
其中, 所述多个第一输入和所述多个第二输入布置为与多个线程相对应的多对;
为所述处理器的每个操作周期调度所述多个线程中的线程;
对所述线程执行计算操作, 其中, 所述处理器(308)用于: 当相应线程的值中的至少一个是非影响值时, 并行执行针对所述线程的所述计算操作和针对其他线程中的每个相应线程的跳过操作, 所述非影响值为不会显著影响所述处理器(308)计算的值的输入值; 或者, 当相应线程的值中的至少一个是影响值时, 并行执行针对所述线程的所述计算操作和针对其他线程中的每个相应线程的停顿操作。
2. 根据权利要求1所述的多线程脉动阵列(300), 其特征在于, 所述多个处理元件(302)布置为二维阵列, 从而将第一处理元件(302)的多个第一输出(316)作为多个第一输入(312)提供给第二相邻处理元件(302), 将所述第一处理元件的多个第二输出(318)作为多个第二输入(314)提供给第三相邻处理元件。
3. 根据权利要求1所述的多线程脉动阵列(300), 其特征在于, 根据所述线程的可用非影响值执行所述调度。
4. 根据权利要求1所述的多线程脉动阵列(300), 其特征在于, 所述计算操作由计算组件(308A)执行, 所述跳过操作由移动路径组件(308B)执行, 所述停顿操作由停顿组件(308C)执行, 其中, 所述计算组件(308A)的所述计算操作的执行独立于所述移动路径组件(308B)的所述跳过操作和所述停顿组件(308C)的所述停顿操作。
5. 根据前述权利要求1至4中的任一项所述的多线程脉动阵列(300), 其特征在于, 所述处理器(308)用于当所述多个线程包括非影响值时, 并行处理针对所述多个线程的多个跳过操作。
6. 根据前述权利要求1至4中任一项所述的多线程脉动阵列(300), 其特征在于, 所述处理器(308)用于针对相应的第一输入和/或第二输入, 对每个包括非影响值的线程执行跳过操作。
7. 根据前述权利要求1至4中的任一项所述的多线程脉动阵列(300), 其特征在于, 所述处理器(308)用于将神经网络的不同层分配给不同线程, 根据相应层的带宽在所述不同线程之间进行选择, 其中, 所述第一输入表示所述神经网络的激活输入, 所述第二输入表示所述神经网络的权重。
8. 根据前述权利要求1至4中的任一项所述的多线程脉动阵列(300), 其特征在于, 所述多个处理元件(302)分别包括: 至少一个第一延迟缓冲器(604A), 当相应处理元件的所述处理器(308)处理所述线程时存储所述多个第一输入的数据; 至少一个第二延迟缓冲器(604B), 当相应处理元件的所述处理器(308)处理所述线程时存储所述多个第二输入的数据。
9. 根据前述权利要求2所述的多线程脉动阵列(300), 其特征在于, 所述处理器(308)用于将非影响值输入从所述多个第一输入(312)移动到所述多个第一输出(316), 从所述多个第二输入(314)到所述多个第二输出(318)。

10. 根据权利要求9所述的多线程脉动阵列(300), 其特征在于, 移动所述非影响值输入而不计算所述非影响值输入的乘法。

11. 根据前述权利要求1至4中的任一项所述的多线程脉动阵列(300), 其中, 当所述多个第一输入和所述多个第二输入具有影响值时, 对执行所述计算操作的计算组件(308A)的时钟进行门控。

12. 根据前述权利要求1至4中的任一项所述的多线程脉动阵列(300), 其特征在于, 所述多线程脉动阵列(300)用于根据所述多个线程提供的数据计算多个矩阵乘法。

13. 根据前述权利要求1至4中的任一项所述的多线程脉动阵列(300), 其特征在于, 所述计算操作的结果为以下之一: 分别存储在所述多个处理元件本地, 并且作为输出移动到第二和第三相邻处理元件; 用作根据相应的第一输入和相应的第二输入进行下一次计算的部分和输入。

14. 一种操作多线程脉动阵列的方法, 其特征在于, 包括:

针对所述多线程脉动阵列的多个处理元件中的每个处理元件, 执行:

从相应的第一输入源(304)接收多个第一输入(202);

从相应的第二输入源(306)接收多个第二输入(202),

其中, 所述多个第一输入和所述多个第二输入布置为与多个线程相对应的多对;

为每个操作周期调度所述多个线程中的线程(204);

对所述线程执行计算操作(206), 其中, 当相应线程的值中的至少一个是非影响值时, 并行执行针对所述线程的所述计算操作和针对其他线程中的每个相应线程的跳过操作, 所述非影响值为不会显著影响处理器计算的值的输入值; 或者, 当相应线程的值中的至少一个是影响值时, 并行执行针对所述线程的所述计算操作和针对其他线程中的每个相应线程的停顿操作。

多线程脉动阵列

背景技术

[0001] 本发明在其一些实施例中涉及矩阵乘法,更具体但非排他地,涉及用于一个或多个矩阵的乘法的系统和方法。

[0002] 矩阵和向量乘法是各种处理方法中使用的基本运算,例如,用于进行神经网络推导和/或训练神经网络。目前正寻求计算高效的系统和/或方法,从而使用合理数量的计算资源(即,存储器和/或处理硬件)在合理的时间内快速执行矩阵乘法。例如,神经网络硬件加速器需要具有高数据吞吐量和高效的能耗,才能基于大量不同的输入对不同的神经网络进行处理。

发明内容

[0003] 本发明的目的在于提供一种多线程脉动阵列,以及用于操作该多线程脉动阵列的装置、方法、计算机程序产品和系统。

[0004] 通过独立权利要求的特征实现前述和其他目的。进一步的实施形式在从属权利要求、具体说明和附图中显而易见。

[0005] 根据第一方面,一种多线程脉动阵列包括:多个处理元件,分别包括处理器,其中,每个所述处理元件用于:从相应的第一输入源接收多个第一输入,从相应的第二输入源接收多个第二输入,其中,所述多个第一输入和所述多个第二输入布置为与多个线程相对应的多对,为所述处理器的每个操作周期调度所述多个线程中的线程,对所述线程执行计算操作。

[0006] 根据第二方面,一种操作多线程脉动阵列的方法包括:针对所述多线程脉动阵列的多个处理元件中的每个处理元件,执行:从相应的第一输入源接收多个第一输入,从相应的第二输入源接收多个第二输入,其中,所述多个第一输入和所述多个第二输入布置为与多个线程相对应的多对,为每个操作周期调度所述多个线程中的线程,对所述线程执行计算操作。

[0007] 当输入数据表现出例如如现实生活中的神经网络所表现出的高权重稀疏性和/或高输入稀疏性时,与其他实现(例如,单线程脉动阵列)相比,多线程脉动阵列提高了处理元件的利用率。

[0008] 不同于一次只处理一个矩阵运算(例如,如单线程脉动阵列实现所执行的),多线程脉动阵列的计算资源还用于其他矩阵运算,而不是对当前矩阵运算执行无意义的运算,从而提高了计算效率。

[0009] 与神经网络处理系统的其他实现相比,例如,在数据吞吐量和/或能耗效率的提高方面,将多线程脉动阵列纳入神经网络处理系统中提高了基于大量的不同输入执行多个不同的神经网络的效率。

[0010] 在第一和第二方面的另一实施形式中,所述多个处理元件布置为二维阵列,从而将第一处理元件的多个第一输出作为多个第一输入提供给第二相邻处理元件,将所述第一处理元件的多个第二输出作为多个第二输入提供给第三相邻处理元件。

[0011] 在第一和第二方面的另一实施形式中,根据所述线程的可用非影响值执行所述调度,其中,所述非影响值为不会显著影响所述处理器计算的值的输入值。

[0012] 在第一和第二方面的另一实施形式中,所述处理器用于:当相应线程的值中的至少一个是非影响值时,并行执行针对所述线程的所述计算操作和针对其他线程中的每个相应线程的跳过操作;或者,当相应线程的值中的至少一个是影响值时,并行执行针对所述线程的所述计算操作和针对其他线程中的每个相应线程的停顿操作。

[0013] 通过并行处理,多线程脉动阵列的计算性能得以提高,这是因为并行执行两个或更多线程所花费的时间大约是各线程依次执行(例如,由单线程脉动阵列实现执行)时所需时间的一半或更短。

[0014] 在第一和第二方面的另一实施形式中,所述计算操作由计算组件执行,所述跳过操作由移动路径组件执行,所述停顿操作由停顿组件执行,其中,所述计算组件的所述计算操作的执行独立于所述移动路径组件的所述跳过操作和所述停顿组件的所述停顿操作。

[0015] 在第一和第二方面的另一实现形式中,所述处理器用于当所述多个线程包括非影响值时,并行处理针对所述多个线程的多个跳过操作。

[0016] 在第一和第二方面的另一实现形式中,所述处理器用于针对相应的第一输入和/或第二输入,对每个包括非影响值的线程执行跳过操作。

[0017] 在第一和第二方面的另一实施形式中,所述处理器用于将神经网络的不同层分配给不同线程,根据相应层的带宽在所述不同线程之间进行选择,其中,所述第一输入表示所述神经网络的激活输入,所述第二输入表示所述神经网络的权重。

[0018] 所述多线程脉动阵列的计算效率随着变化的神经网络保持在较高水平。

[0019] 与不考虑和/或适应带宽差异的其他实现(例如,单线程脉动阵列实现)相比,所述多线程脉动阵列适应输入和/或权重(例如,全连接层和/或卷积层等神经网络的不同网络层的输入和/或权重)的不同带宽要求,从而提高了计算效率和/或能源利用率。例如,对于大多数现实生活的神经网络,单线程脉动阵列实现一次只能处理一个网络层,因而在等待权重和/或输入数据时,其没有得到充分的利用,和/或,阵列速度太慢,无法实现所需的性能。因此,与所述适应带宽的多线程脉动阵列相比,单线程脉动阵列无法匹配神经网络加速器存储器子系统带宽,从而导致计算性能和/或能源利用率降低。

[0020] 在第一和第二方面的另一实施形式中,所述多个处理元件分别包括:至少一个第一延迟缓冲器,当相应处理元件的所述处理器处理所述线程时存储所述多个第一输入的数据;至少一个第二延迟缓冲器,当相应处理元件的所述处理器处理所述线程时存储所述多个第二输入的数据。

[0021] 在第一和第二方面的另一实施形式中,所述处理器用于将非影响值输入从所述多个第一输入移动到所述多个第一输出,从所述多个第二输入到所述多个第二输出。

[0022] 在第一和第二方面的另一实施形式中,移动所述非影响值输入而不计算所述非影响值输入的乘法。

[0023] 在第一和第二方面的另一实施形式中,当所述多个第一输入和所述多个第二输入具有影响值时,对执行所述计算操作的所述计算组件的时钟进行门控。

[0024] 针对非影响值关闭所述处理器的电源(可选地,关闭所述处理器的计算组件的电源),从而减少能耗。如本文所讨论的,对于非影响值不执行计算操作。

[0025] 在第一和第二方面的另一实施形式中,所述多线程脉动阵列用于根据所述多个线程提供的数据计算多个矩阵乘法。

[0026] 在第一和第二方面的另一实施形式中,所述计算操作的结果为以下之一:分别存储在所述多个处理元件本地,并且作为输出移动到所述第二和第三相邻处理元件;用作根据相应的第一输入和相应的第二输入进行下一次计算的部分和输入。

[0027] 除非另有定义,否则本文所用的所有技术和科学术语都具有与本发明普通技术人员公知的含义相同的含义。尽管与本文描述的那些类似或等同的方法和材料可以用于本发明的实施例的实践或测试中,下面将对示例性的方法和/或材料进行描述。在有冲突的情况下,以专利说明书及其定义为准。另外,材料、方法和示例仅是说明性的,并不用于进行必要的限制。

附图说明

[0028] 此处仅作为示例,结合附图描述了本发明一些实施例。现在具体结合附图,需要强调的是所示的项目作为示例,为了说明性地讨论本发明的实施例。这样,根据附图说明,如何实践本发明实施例对本领域技术人员而言是显而易见的。

[0029] 在附图中:

[0030] 图1为用于帮助理解本发明一些实施例的标准脉动阵列的示例性实现的示意图;

[0031] 图2为根据本发明一些实施例的多线程脉动阵列的操作方法的流程图;

[0032] 图3A为根据本发明一些实施例的包括多个处理元件的多线程脉动阵列的框图,该处理元件均设计用于并行执行多个操作;

[0033] 图3B为根据本发明一些实施例的提供多线程脉动阵列的一个处理元件的细节的框图;

[0034] 图4为根据本发明一些实施例的示出了标准单线程脉动阵列和本文所述的多线程脉动阵列的计算效率之间的比较的示意图;

[0035] 图5为根据本发明一些实施例的多线程脉动阵列的一个处理元件的示例性架构的示意图;

[0036] 图6为根据本发明一些实施例的多线程脉动阵列的示例性架构的示意图,包括用于每个处理元件的每个输入的相应的延迟队列(可选地,基于FIFO的实现)。

具体实施方式

[0037] 本发明在其一些实施例中涉及矩阵乘法,更具体地,涉及一种或多种矩阵的乘法的系统和方法。但本发明不限于此。

[0038] 本发明一些实施例的一方面涉及用于通过多线程脉动阵列的每个处理元件执行多个可用线程中的线程的计算操作(可选地,乘法和加法)的方法、系统和/或装置。所述多脉动阵列的每个处理元件包括分别接收多个第一输入和多个第二输入的一个或多个处理器。第一和第二输入成对布置,每对对应一个线程。以操作周期为单位,调度多个线程中的线程,使得所述多线程脉动阵列的相应的处理元件的每个处理器执行计算操作。

[0039] 根据所述线程的可用影响(例如,非零)值来执行调度。对具有非影响(例如,不可用数据,零)值的其他非调度线程执行跳过操作,从而将所述非影响值作为输出提供(可选

地,提供给一个或多个相邻处理元件)。执行所述跳过操作,而不执行计算操作。对其他具有影响值的非调度线程执行停顿操作。所述计算操作与所述跳过操作和所述停顿操作并行且独立地执行。多个跳过操作并行执行。

[0040] 当输入数据表现出例如如现实生活中的神经网络所表现出的高权重稀疏性和/或高输入稀疏性时,与其他实现(例如,结合图1所描述的单线程脉动阵列)相比,多线程脉动阵列提高了处理元件的利用率。不同于一次只处理一个矩阵运算(例如,如单线程脉动阵列实现所执行的),多线程脉动阵列的计算资源还用于其他矩阵运算,而不是对当前矩阵运算执行无意义的运算,从而提高了计算效率。

[0041] 与神经网络处理系统的其他实现相比,例如,在数据吞吐量和/或能耗效率的提高方面,将多线程脉动阵列纳入神经网络处理系统中提高了基于大量的不同输入执行多个不同的神经网络的效率。

[0042] 在详细解释本发明的至少一个实施例之前,应当理解,本发明的应用并不一定限于以下描述的和/或附图和/或示例中示出的部件和/或方法的构造和布置的细节。本发明可以具有其他实施例或者可以各种方式实践或执行。

[0043] 本发明可以是系统、方法和/或计算机程序产品。该计算机程序产品可以包括其上具有用于使处理器执行本发明各方面的计算机可读程序指令的计算机可读存储介质。

[0044] 该计算机可读存储介质可以是有形设备,该有形设备可以保留和存储供指令执行设备使用的指令。计算机可读存储介质可以是例如但不限于电子存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或前述的任何合适的组合。

[0045] 此处描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络(例如,互联网、局域网、广域网和/或无线网络)下载到外部计算机或外部存储设备。

[0046] 计算机可读程序指令可以全部在用户计算机上或者部分在用户计算机上作为独立软件包执行,或者部分在用户计算机上且部分在远程计算机上或全部在远程计算机或服务器上执行。在后一种情况下,远程计算机可以通过任何类型的网络(包括局域网(LAN)或广域网(WAN))连接到用户计算机,或者可以与外部计算机建立连接(例如,通过Internet服务提供商的Internet)。在一些实施例中,包括例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)的电子电路可以通过利用计算机可读程序指令的状态信息来执行计算机可读程序指令来实现电子电路的个性化,以便执行本发明的各个方面。

[0047] 这里参考根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图图示和/或框图图示描述了本发明的各方面。将理解的是,流程图图示和/或框图的每个框以及流程图图示和/或框图中的各框的组合可以通过计算机可读程序指令来实现。

[0048] 图中的流程图和框图阐述了根据本发明各个实施例的系统、方法以及计算机程序产品的可能实现方式的结构、功能以及操作。就此而言,流程图或框图中的每个块可以表示模块、区段或部分指令,其包括用于实施指定逻辑功能的一个或多个可执行指令。在一些替代实施方式中,方框中指出的功能可以不按图中指出的顺序发生。例如,事实上,连续示出的两个方块可以几乎同时执行,或者有时候,方块可以按照相反的顺序执行,这取决于所涉及的功能。也应注意的是,框图和/或流程图图示中的每个框以及框图和/或流程图图示中的各框的组合可以由基于专用硬件的系统实现,该系统执行指定的功能或动作,或者执行

专用硬件和计算机指令的组合。

[0049] 应注意的是,本文描述的术语“矩阵”有时可以与术语“向量”互换。

[0050] 如本文所使用的,术语“非影响值”是指输入(到相应处理元件中)的值,其在执行计算操作时不会显著影响相应处理元件计算的值。例如,零值、接近零的值和/或根据要求(例如,定义的范围、规则集,统计公差要求)定义的其他值,这些值在执行计算操作(例如,乘法和加法)时不会显著影响其他值。例如,零输入值乘以其他输入值将产生零输出值。零输入值不会影响输出值,因为零输出值不受计算操作的影响。如本文另外详细描述的,零输出值由跳过操作(在本文描述)和计算操作两者产生,这使得计算操作变得多余。

[0051] 如本文所使用的,术语“影响值”是指当执行计算操作时对输出值产生重大影响的输入值。例如,非零值和/或根据要求定义的其他值,这些值在执行计算操作时会显著影响其他值。非零输入值会影响一个或多个输出值,因为第一和/或第二输入值受计算操作影响。例如,一个非零值(如远非零的值)乘以其他非零值会显著影响输出值。输出值明显不同于输入值中的任一个。

[0052] 现在参考图1,图1为用于帮助理解本发明一些实施例的标准脉动阵列100的示例性实现的示意图。标准脉动阵列100包括设计为执行矩阵和/或向量乘法的硬件,例如,作为用于加速基于神经网络的应用的神经网络硬件加速器。与本文所描述的多线程脉动阵列相反,标准脉动阵列100是一种单线程脉动阵列实现。由于所描述的标准脉动阵列100的低效操作,因此,与本文所描述的多线程脉动阵列相比,平均利用率和/或功耗效率相对较低。

[0053] 脉动阵列100包括布置成二维阵列的多个单元(为了清楚起见,标记一个单元102作为示例)。激活输入组件104沿着脉动阵列100的第一维度提供神经网络的激活输入。权重输入组件106沿着脉动阵列100的第二维度提供神经网络的权重输入。每个相应的单元102计算由该相应的单元接收的激活输入和权重输入的乘积。激活输入沿脉动阵列100的第二维度进行移动,而权重输入沿脉动阵列100的第二维度进行移动。

[0054] 发明人发现,实际上,例如在处理现实生活中的神经网络时,标准脉动阵列100的利用率、计算效率和/或能耗效率较低。发明人发现,由于标准脉动阵列100是针对密集的矩阵和/或向量而设计的,所以,现实生活中的神经网络(即,稀疏矩阵和/或稀疏向量)中存在的高度的权重和/或输入稀疏性会导致标准脉动阵列100的利用率降低。与本文所述的根据输入和/或权重稀疏性操作的多线程脉动阵列相比,标准脉动阵列100无法根据输入和/或权重稀疏性进行适配。

[0055] 此外,发明人发现,不同的网络层(例如,全连接层和/或卷积层)需要不同的输入带宽和/或权重以进行高效的计算,这是标准脉动阵列100所无法满足的。标准脉动阵列100的性能和/或能耗无法匹配神经加速器存储器子系统带宽。对于大多数现实生活中的神经网络,标准脉动阵列100一次只能处理一个神经网络层,因而在等待权重和/或激活输入数据时,脉动阵列100没有得到充分的利用,和/或,脉动阵列100可能速度太慢,无法实现所需的性能。

[0056] 另外,发明人发现,由于标准脉动阵列100一次仅能够处理一个矩阵运算,因此,标准脉动阵列100的计算资源不能再用于其他矩阵运算,而只是用于对当前矩阵运算执行无意义的运算,例如,乘以零。此外,标准脉动阵列100不能区分具有需要计算的有意义的值(例如,非零值)的矩阵和/或向量元素和具有不需要计算操作(例如,由于乘以零为零)的无

意义的值(例如,零值)的元素。相反,本文所描述的多线程脉动阵列设计为调度其计算资源以处理有意义的运算(例如,导致非零值的运算),同时将计算资源从无意义的运算(例如,导致零值的运算)中转移。

[0057] 现在参考图2,图2为根据本发明一些实施例的多线程脉动阵列的操作方法的流程图。还参考图3A,图3A为根据本发明一些实施例的包括多个处理元件302的多线程脉动阵列300的框图,该处理元件302均设计用于并行执行针对多个线程的多个操作。还参考图3B,图3B为根据本发明一些实施例的提供多线程脉动阵列300的一个处理元件302的细节的框图。结合图2所描述的方法可以由结合图3A所描述的多线程脉动阵列300和结合图3B所描述的处理元件302来实现。

[0058] 现在参考图3A,多线程脉动阵列300可以实现为硬件组件、软件组件或硬件和软件的组合(例如,硬件可以包括存储在数据存储设备中以由一个或多个处理器执行的代码指令)。多线程脉动阵列300可以实现为与另一计算系统350通信的外部组件(例如,独立单元)、与计算系统350集成的组件(例如,插入计算系统现有端口中的硬件加速器卡)和/或集成在另一个计算系统350中(例如,设计和构建为另一个系统的组成部分,例如,设计为处理阵列的特征)。

[0059] 多线程脉动阵列300可以设计为向计算系统350提供计算高效的矩阵乘法服务(例如,在减少处理时间方面)。例如,计算系统350可以执行神经网络的代码指令,多线程脉动阵列300在执行用于训练神经网络矩阵乘法和/或通过训练的神经网络计算结果时提供减少的处理时间。

[0060] 可以与多线程脉动阵列300通信的示例性计算系统350例如包括以下中的一项或多项:服务器、计算云、计算机网络、虚拟机(例如,管理程序、虚拟服务器)、单个计算设备(例如,客户端、移动设备)和并行布置的一组计算设备。

[0061] 多线程脉动阵列300包括第一输入源304,当计算神经网络的矩阵时,该第一输入源304向多线程脉动阵列300的第一组处理元件302中的每个成员提供多个第一输入,例如,激活输入。多线程脉动阵列300包括第二输入源306,当计算神经网络的矩阵时,该第二输入源306向多线程脉动阵列300的第二组处理元件302中的每个成员提供多个第二输入,例如,权重。第一输入源304和第二输入源306可以例如由正在执行计算的计算系统350来实现。例如,第一输入源304和第二输入源306可以通过总线、局域网或其他通信信道将输入提供给相应的处理元件302。

[0062] 多线程脉动阵列300设计为根据各线程提供的数据计算多个矩阵乘法。

[0063] 处理元件302可以布置为例如二维正方形阵列、二维矩形阵列、二维三角形阵列或其他架构,其可以例如根据执行的应用(例如,神经网络)而变化。

[0064] 多线程脉动阵列300可以与用户界面352(直接地和/或通过其他计算系统350间接地)进行通信,该用户界面352向用户呈现数据和/或包括用于输入数据的机制,例如,触摸屏、显示器、键盘、鼠标、手写笔、语音激活软件和麦克风中的一个或多个。

[0065] 现在参考图3B,每个处理元件302包括一个或多个处理器308,实现为例如中央处理器(CPU)、图形处理单元(GPU)、现场可编程门阵列(FPGA)、数字信号处理器(DSP)、专用集成电路(ASIC)、定制电路和/或专用硬件加速器。处理器308可以实现为单个处理器、多核处理器和/或布置为用于并行执行的处理器集群(其可以包括同构和/或异构处理器)。应注意

的是,处理器308可以设计为以硬件方式实现作为代码存储在存储器310中的一个或多个特征。

[0066] 处理元件302的以下组件可以与处理器308通信,和/或可以实现为处理器308的组件(例如,集成在其中):

[0067] *计算组件308A,其对所选线程执行计算操作,如本文所述。

[0068] *移动路径组件308B,其执行跳过操作,如本文所述。

[0069] *停顿组件308C,其执行停顿操作,如本文所述。

[0070] 计算组件308A的执行独立(即,并行)于移动路径组件308B和/或停顿组件308C。通过组件并行执行,多线程脉动阵列的计算性能得以提高,这是因为并行执行两个或更多线程所花费的时间大约是各线程依次执行(例如,由单线程脉动阵列实现执行)时所需时间的一半或更短。

[0071] 组件308A-C可以实现为一个或多个处理器308的集成特征,例如,实现为硬件特征和/或在由处理器308执行时实现相应特征的代码指令。组件308A-C可以实现为与处理器308集成和/或与处理器308通信的不同组件,例如,设计为执行由相应组件定义的各个特征的硬件和/或代码指令。

[0072] 处理元件302可包括存储有可由处理器308执行的代码指令的存储器310,例如,随机存取存储器(RAM)和/或只读存储器(ROM)。存储器310可以存储中间计算结果,和/或可以存储其他数据。

[0073] 处理元件302可以包括用于从第一输入源304接收输入的第一输入接口312和用于从第二输入源306接收输入的第二输入接口314,例如总线接口和/或网络接口。第一输入接口312的第一输入和第二输入接口314的第二输入成对布置,每对在本文中称为“线程”。

[0074] 处理元件302可以包括求和寄存器,例如,每个线程对应一个求和寄存器。求和寄存器可以存储在存储器310和/或其他数据存储设备中。

[0075] 处理元件302可以包括分别用于第一和第二输入的移动路径。

[0076] 可选地,可以为每个线程实现隔离的输入/输出端口(即,接口312-318的端口)。隔离端口提高了数据吞吐量。

[0077] 第一输入接口312和/或第二输入接口314可以与一个或多个延迟缓冲器相关联,该延迟缓冲器存储从相应的第一输入源304和第二输入源306接收的数据。每个输入可以与各自相应的延迟缓冲器相关联,和/或从第一输入源304或第二输入源306接收的一组输入可以与相应的组合延迟缓冲器相关联。当处理器308处理其他接收到的数据时(即,在一操作周期中),延迟缓冲器可以存储从相应的第一输入接口312和/或第二输入接口314接收的数据。

[0078] 处理元件302可以包括第一输出接口316,该第一输出接口316将来自一处理元件302的输出提供给相邻的处理元件,可选地,提供给另一处理元件的第一输入接口。处理元件302可以包括第二输出接口318,该第二输出接口318将来自一处理元件302的输出提供给另一相邻处理元件,可选地,提供给另一处理元件的第二输入接口。应注意的是,处理元件302可以例如根据多线程脉动阵列的总体设计和/或根据由多线程脉动阵列执行的计算操作,包括附加输出。

[0079] 现在回到图2,为了解释的清楚和简单,结合多线程脉动阵列300的多个处理元件

中的一个处理元件302来讨论结合图2所描述的方法。多线程脉动阵列300的其他处理元件302以类似的方式操作。

[0080] 202,接收多线程输入。如本文所讨论的,每个线程包括一个第一输入和与之配对的一个第二输入。从第一输入源304(可选地,通过第一输入接口312)接收多个第一输入,例如,神经网络的激活输入。从第二输入源306(可选地,通过第二输入接口314)接收多个第二输入,例如,神经网络的权重。

[0081] 每个线程可以表示一个矩阵运算。

[0082] 输入可以针对每个线程实现为单独输入(例如,队列),或者实现为两个或更多个(例如,所有)线程共享的公共输入(例如,队列)。

[0083] 204,调度多个线程中的线程。如结合动作206所描述的,调度线程以执行计算操作。

[0084] 计算操作的示例包括:乘法和加法(MULADD)。

[0085] 当多个线程满足执行计算操作的要求时,可以根据优先级要求,例如基于循环、随机选择、层次结构、一组规则和/或优先级排名(例如,得分),从其他满足要求的线程中选择线程。

[0086] 根据操作周期执行调度。可选地,针对每个操作周期执行调度。在每个操作周期中,一个计算操作(如结合动作208所描述的),可选地,与一个或多个跳过操作(如结合动作210所描述的)和/或停顿操作(如结合动作212所描述的)并行执行。可替代地,在每个操作周期中,多个跳过操作和/或停顿操作可以并行执行。每个操作可以是一个任务的组成部分,例如,单个任务可以包括多个操作。

[0087] 根据线程的可用影响值执行调度。可用数据具有定义的值,和/或者可以进行处理而不额外进行延迟。

[0088] 当第一输入和第二输入中的至少一个对于线程不可用时,调度另一个线程。例如,由于存储带宽要求,提供输入的另一进程尚未完成操作(例如,在计算系统350上执行的操作)时,和/或当值经历传输延迟时,值可能不可用。

[0089] 可选地,处理元件执行用于训练神经网络的计算和/或使用训练的神经网络计算结果。对于神经网络应用,第一输入表示神经网络的激活输入,第二输入表示神经网络的权重。可选地,将神经网络的不同层分配给不同的线程。处理元件根据相应层的带宽在不同线程之间进行选择。所述多线程脉动阵列的计算效率随着变化的神经网络保持在较高水平。

[0090] 与不考虑和/或适应带宽差异的其他实现(例如,单线程脉动阵列实现)相比,所述多线程脉动阵列适应输入和/或权重(例如,全连接层和/或卷积层等神经网络的不同网络层的输入和/或权重)的不同带宽要求,从而提高了计算效率和/或能源利用率。例如,对于大多数现实生活的神经网络,单线程脉动阵列实现一次只能处理一个网络层,因而在等待权重和/或输入数据时,其没有得到充分的利用,和/或,阵列速度太慢,无法实现所需的性能。因此,与所述适应带宽的多线程脉动阵列相比,单线程脉动阵列无法匹配神经网络加速器存储器子系统带宽,从而导致计算性能和/或能源利用率降低。

[0091] 可选地,206,针对线程执行计算操作,例如,计算激活输入和相应的权重输入的乘积。该计算操作可以由相应处理元件302的乘法和加法组件308A执行。

[0092] 可选地,当第一输入和第二输入(即,线程)具有非影响值时,对执行计算操作的计

算组件308A的时钟进行门控。针对非影响值关闭所述处理器的电源和/或计算组件的电源，从而减少能耗。如本文所讨论的，对于非影响值不执行计算操作。

[0093] 可选地，计算操作的结果例如通过存储器310存储在相应的处理元件302本地。替代地或附加地，结果作为输出移动到相邻的处理元件。每个相应的相邻处理元件将所移动的输出用作根据相应的第一输入和相应的第二输入进行下一次计算的部分和输入。

[0094] 应注意的是，当一个或两个输入均为非影响值时，不执行计算操作。通过对重要的影响值执行计算操作并避免对不可用和/或非影响值执行计算操作，多线程脉动阵列的整体效率和/或性能得以提高（即，处理资源的利用率得到提高），和/或，例如，与标准单线程脉动阵列相比，单位矩阵运算的功耗降低了。

[0095] 替代地或附加地，208，针对相应的第一输入和/或第二输入，对每个包括非影响值的线程执行跳过操作。跳过操作可以由相应处理元件302的跳过组件308B执行。

[0096] 通过将非影响值输入从第一输入（例如，通过接口312）移动到第一输出（例如，通过接口316），以及从第二输入（例如，通过接口314）移动到第二输出（例如，通过接口318）来执行跳过操作。对所述非影响值进行移动而不执行影响值输入的计算操作。

[0097] 可选地，当多个线程包括非影响值时，可以并行执行多个跳过操作。

[0098] 替代地或附加地，210，执行停顿操作。当相应线程的第一和第二输入的值中的至少一个为影响值时，当有多个线程满足计算操作的处理要求，执行停顿操作。对未选中进行计算操作的处理的线程执行停顿操作。

[0099] 停顿操作可以由相应处理元件302的停顿组件308C执行。由停顿操作所停顿的值可以在本地存储，例如，存储在存储器310中，和/或在本文所描述的一个或多个延迟队列中。

[0100] 212，在恰当的时候，根据多线程脉动阵列中相应处理元件的位置，和/或根据多线程脉动阵列的架构设计，将输出（即，计算操作的结果和/或被跳过的数据）作为输入提供给一个或多个相邻处理元件。

[0101] 可选地，对于结合图3A所描述的情况，在二维正方形多线程脉动阵列300中，将相应处理元件302的第一输出（例如，通过第一输出接口316）作为第一输入提供给第二相邻处理元件（例如，通过第二相邻处理元件的第一输入接口312）。将第一处理元件的第二输出（例如，通过第二输出接口318）作为第二输入提供给第三邻近处理元件（例如，通过第三邻近处理元件的第二输入接口314）。

[0102] 214，对结合动作202-212所描述的一个或多个特征进行迭代。可选地，以操作周期为单位进行迭代。

[0103] 现在参考图4，图4为根据本发明一些实施例的示出了标准单线程脉动阵列和本文所述的多线程脉动阵列的计算效率之间的比较的示意图。为了解释的简单和清楚，对多线程脉动阵列的双线程实现进行描述。跳过操作402表示为“Bps”。乘法和加法运算404表示为“MAC”。箭头406表示单线程脉动阵列执行任务1（408）和执行任务2（410）的示例。总共需要8个操作周期才能将任务1（即，MAC、Bps、Bps、MAC）和任务2（即，Bps、MAC、Bps、MAC）定义的操作依次处理完毕。箭头412A-B表示多线程脉动阵列的双线程实施执行任务1和任务2的示例。箭头412A表示多线程脉动阵列的双线程实现中的第一线程T1的执行。箭头412B表示多线程脉动阵列的双线程实现中的第二线程T2的执行。MAC操作与Bps操作并行执行，或者Bps

操作与另一个Bps操作并行执行,因此,并行执行任务1和任务2定义的操作总共需要5个操作周期(与单线程脉动阵列相比,处理时间大约为其的一半)。

[0104] 现在参考图5,图5为根据本发明一些实施例的多线程脉动阵列的一个处理元件502的示例性架构的示意图。可以基于结合处理元件502所描述的架构来实现多线程脉动阵列的其他处理元件。

[0105] 处理元件502的仲裁组件504识别包括可用数据的线程。如本文所讨论的,每个线程包括一对输入,例如,一组激活输入506中的激活输入和一组权重输入508中的权重输入。仲裁组件504从标识为包括可用数据的各线程中调度一个线程,并将所调度的线程的输入数据引导(和/或提供)至MAC组件510以进行处理。如本文所述,仲裁组件504例如根据非影响数据来确定其他线程的哪些输入数据(可选地,作为一个或多个其他相邻处理元件的输入数据)跳至输出端口512。仲裁组件504和/或MAC组件510可以实现为处理元件502的处理器的组件和/或特征。

[0106] 如本文所述,输入端口506和508可以包括FIFO队列(first in first out)。输出端口512可以包括寄存器。

[0107] 应注意的是,对于每个线程,存在用于将输入(例如,激活和权重)移动到相邻单元(例如,位于处理元件502的右侧和下方的单元)的单独的通道。

[0108] 现在参考图6,图6为根据本发明一些实施例的多线程脉动阵列602的示例性架构的示意图,包括用于每个处理元件606的每个输入的相应的延迟队列604A-b(可选地,基于FIFO的实现)。应注意的是,多线程脉动阵列的 3×3 处理元件606阵列实现仅出于示例目的而示出,不应理解为大小限制。

[0109] 每个处理元件606包括多个第一输入(集中示出)和多个第二输入610(集中示出)。每个输入608和610可以与相应的延迟队列604A-B相关联,或者可以为多个输入实现公共延迟队列。延迟队列604可以实现为用于一组第一输入608的公共队列和用于一组第二输入610的另一个队列604,或者每个第一输入608和第二输入610单独与自己的队列604相关联。

[0110] 在每个处理周期中(例如,如本文所述,并行执行多个操作时),每个相应处理元件606的处理器分析其相应的输入608和610,并为MAC操作的执行调度一个线程,如本文所述。其他未调度的具有可用且有意义的的数据的线程的数据被停顿在相应队列604中,直到当前执行的MAC操作完成为止。

[0111] 如本文所述,每个处理元件606包括相应的MAC组件612和存储器614。

[0112] 在研究下文附图和详细描述之后,本发明的其它系统、方法、特征和优点对于本领域技术人员来说是或变得显而易见的。希望所有这些其它系统、方法、特征和优点包含在本说明书中,在本发明的范围内,并且受所附权利要求的保护。

[0113] 对本发明各个实施例的描述只是为了说明的目的,而这些描述并不旨在穷举或限于所公开的实施例。在不脱离所描述的实施例的范围和精神的情况下,本领域技术人员可以清楚理解许多修改和变化。相比于市场上可找到的技术,选择此处使用的术语可最好地解释本实施例的原理、实际应用或技术进步,或使本领域其他技术人员理解此处公开的实施例。

[0114] 期望在本申请专利的有效期中,将开发出许多相关处理器,术语“处理器”的范围意图先验地包含所有此类新技术。

[0115] 如本文所使用的,术语“约”是指 $\pm 10\%$ 。

[0116] 术语“包括”、“包含”、“具有”以及其变化形式表示“包含但不限于”。这个术语包括了术语“由……组成”以及“本质上由……组成”。

[0117] 短语“主要由…组成”意指组成物或方法可以包含额外成分和/或步骤,但前提是所述额外成分和/或步骤不会实质上改变所要求的组成物或方法的基本和新颖特性。

[0118] 除非上下文中另有明确说明,此处使用的单数形式“一个”和“所述”包括复数含义。例如,术语“化合物”或“至少一个化合物”可以包含多个化合物,包含其混合物。

[0119] 此处使用的词“示例性的”表示“作为一个例子、示例或说明”。任何“示例性的”实施例并不一定理解为优先于或优越于其他实施例,和/或并不排除其他实施例特点的结合。

[0120] 本文使用的词语“可选地”表示“在一些实施例中提供且在其他实施例中未提供”。本发明的任意特定的实施例可以包含多个“可选的”特征,除非这些特征相矛盾。

[0121] 在整个本申请案中,本发明的各种实施例可以范围格式呈现。应理解,范围格式的描述仅为了方便和简洁起见,并且不应该被解释为对本发明范围的固定限制。因此,对范围的描述应被认为是已经具体地公开所有可能的子范围以及所述范围内的个别数值。例如,对例如从1到6的范围的描述应视为已明确公开了子范围,例如从1到3、从1到4、从1到5、从2到4、从2到6、从3到6等,以及该范围内的单个数字,例如1、2、3、4、5和6。无论范围的广度如何,这都适用。

[0122] 当此处指出一个数字范围时,表示包含了在指出的这个范围内的任意所列举的数字(分数或整数)。短语“在第一个所指示的数和第二个所指示的数范围内”以及“从第一个所指示的数到第二个所指示的数范围内”和在这里互换使用,表示包括第一个和第二个所指示的数以及二者之间所有的分数和整数。

[0123] 应了解,为简洁起见,在单独实施例的上下文中描述的本发明的某些特征还可以组合提供于单个实施例中。相反地,为简洁起见在单个实施例的上下文中描述的本发明的各个特征也可以单独地或以任何合适的子组合或作为本发明的任何合适的其它实施例提供。在各个实施例的上下文中描述的某些特征不应视为这些实施例的基本特征,除非在没有这些元素的情况下实施例无效。

[0124] 此处,本说明书中提及的所有出版物、专利和专利说明书都通过引用本说明书结合在本说明书中,同样,每个单独的出版物、专利或专利说明书也具体且单独地结合在此。此外,对本申请的任何参考的引用或识别不可当做是允许这样的参考在现有技术中优先于本发明。就使用节标题而言,不应该将节标题理解成必要的限定。

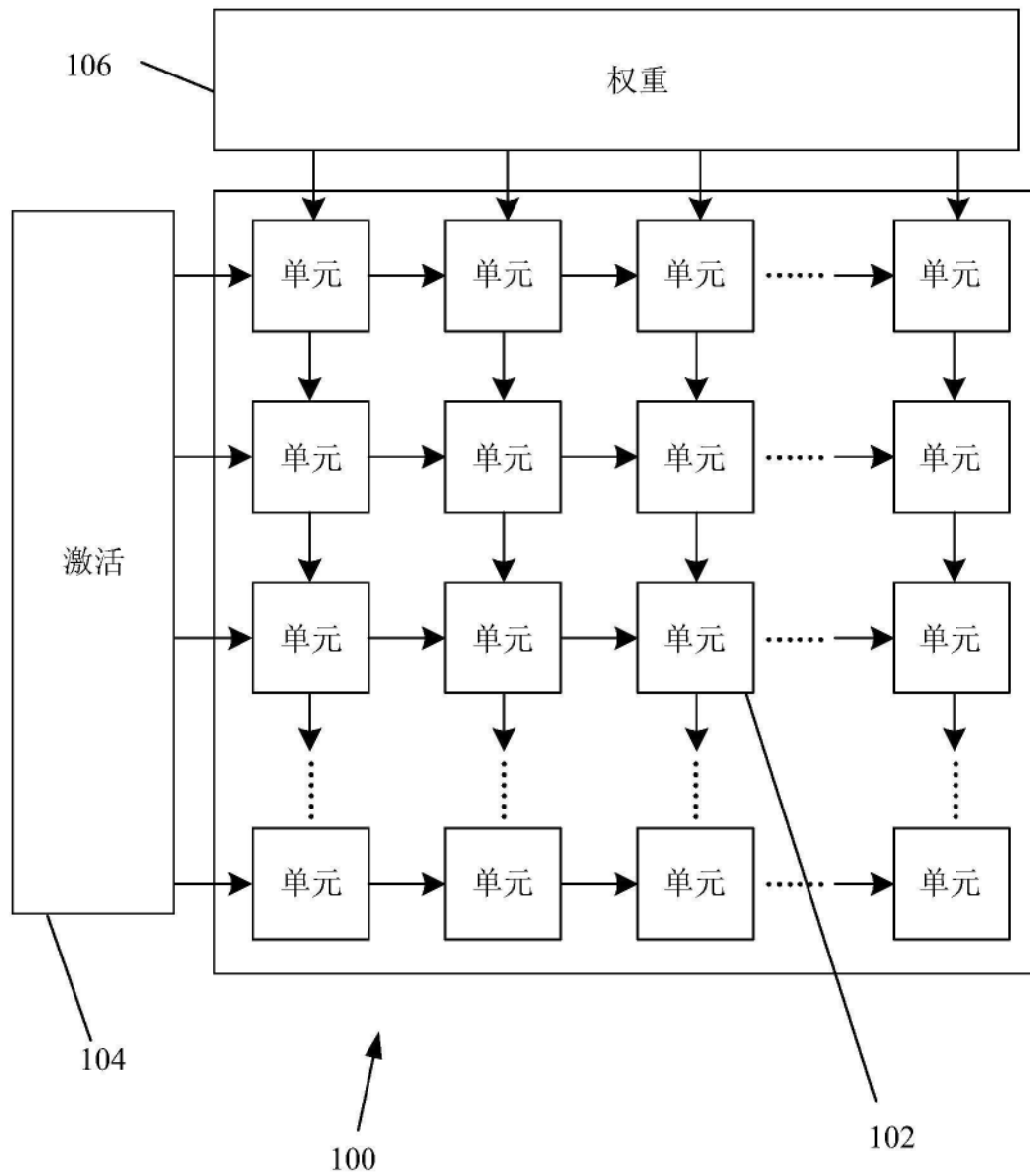


图1

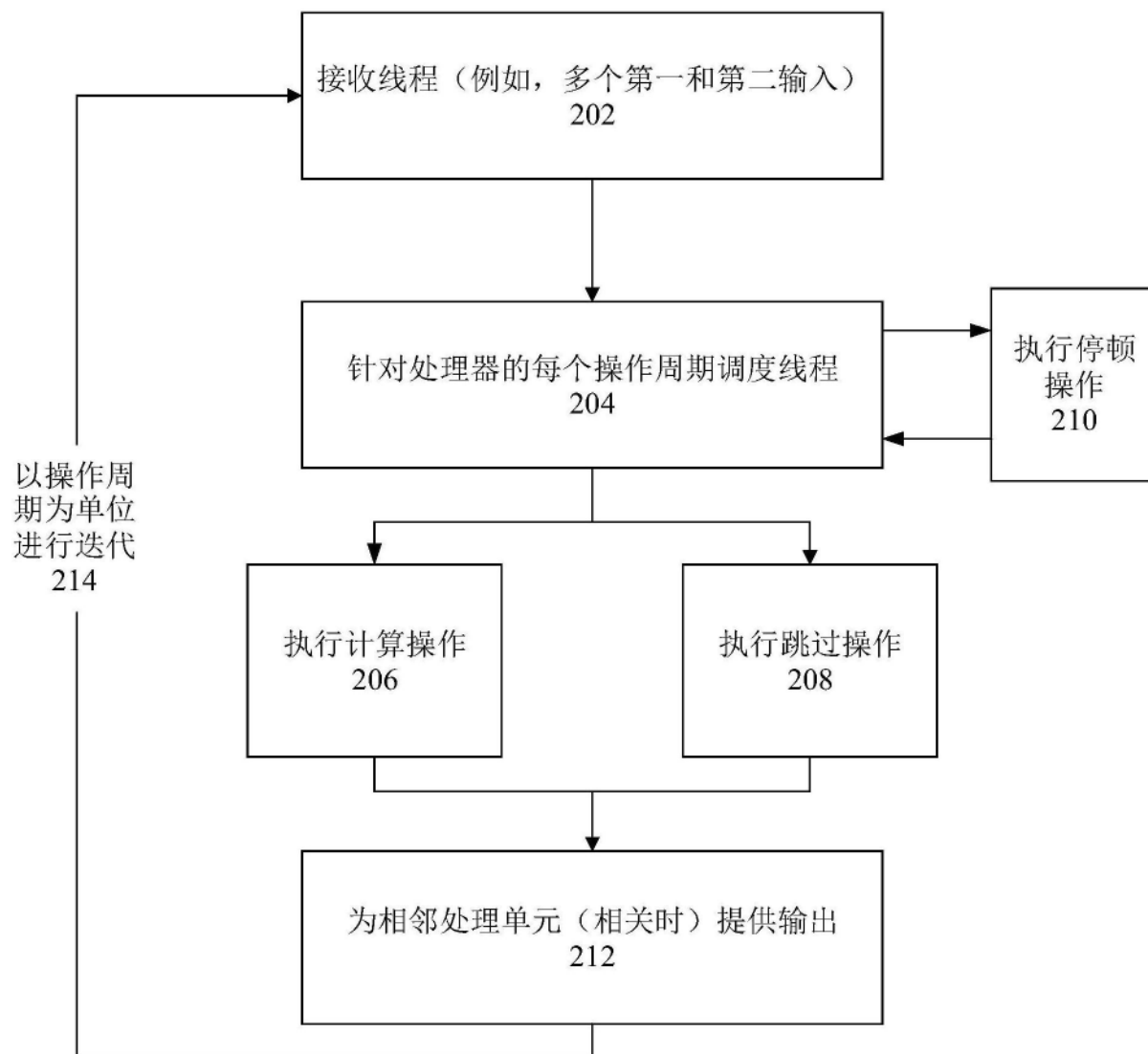


图2

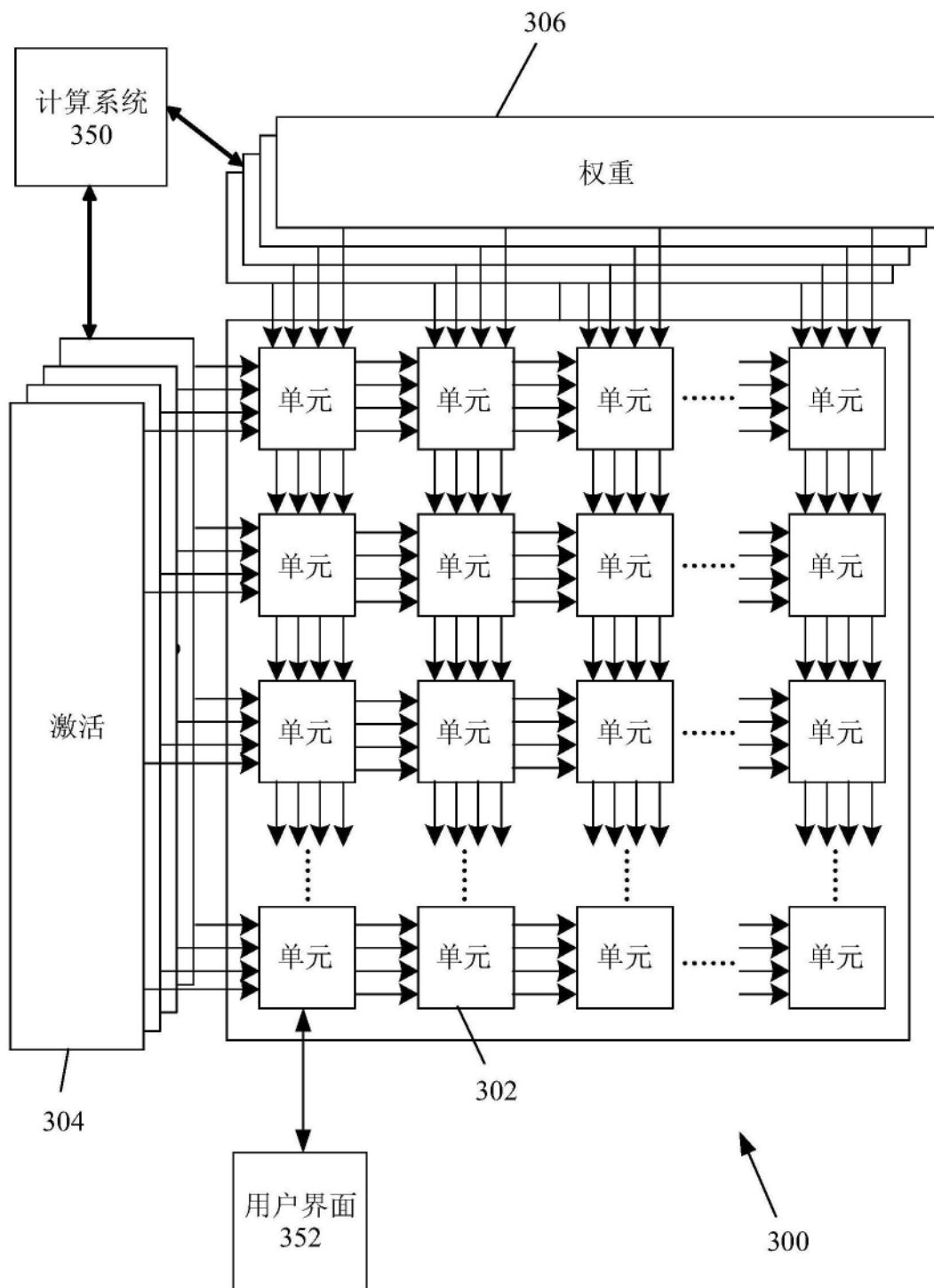


图3A

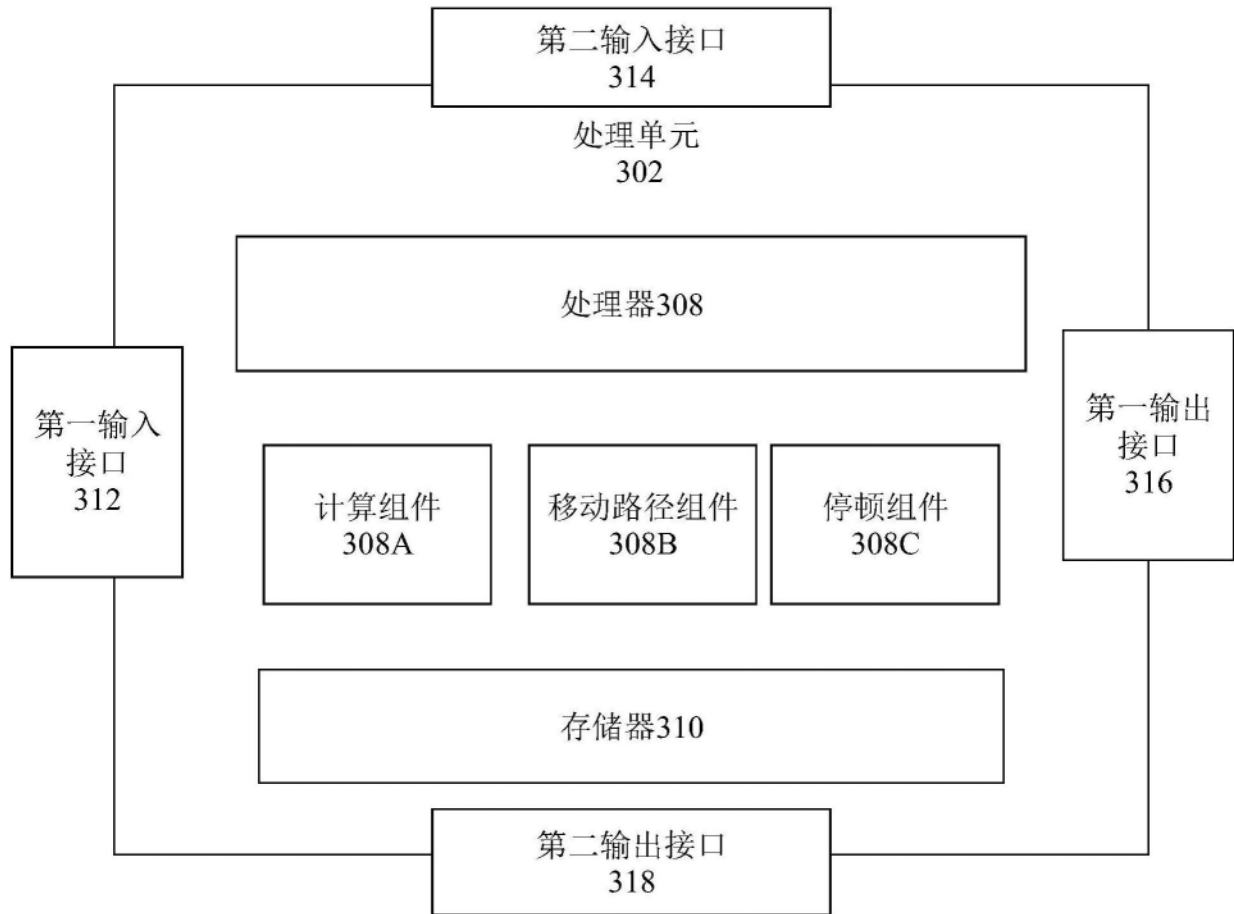


图3B

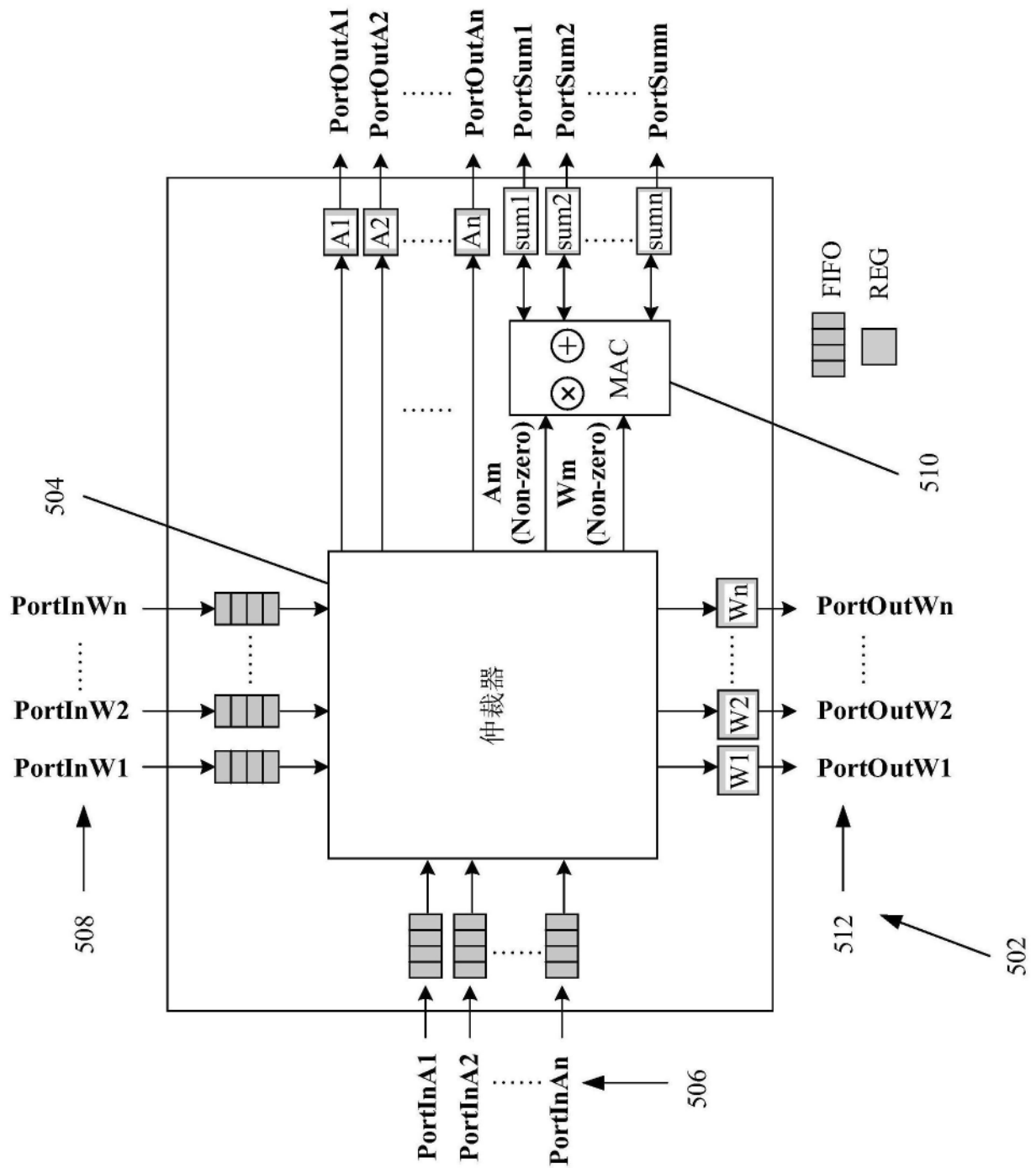


图5

