(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
2017 年 7 月 27 日 (27.07.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/124647 A1

(51) 国际专利分类号:
G06F 17/16 (2006.01)

(21) 国际申请号: PCT/CN2016/078546

(22) 国际申请日: 2016 年 4 月 6 日 (06.04.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610037535.5 2016 年 1 月 20 日 (20.01.2016) CN

(71) 申请人: **北京中科寒武纪科技有限公司 (CAMBRICON TECHNOLOGIES CO., LTD.)** [CN/CN]; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。

(72) 发明人: **陈云霁 (CHEN, Yunji)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 **张潇 (ZHANG, Xiao)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 **刘少礼 (LIU, Shaoli)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 **陈天石 (CHEN, Tianshi)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。

(74) 代理人: **中科专利商标代理有限责任公司 (CHINA SCIENCE PATENT & TRADEMARK AGENT LTD.)**; 中国北京市海淀区西三环北路 87 号 4-1105 室, Beijing 100089 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))。

(54) Title: MATRIX CALCULATION APPARATUS

(54) 发明名称 : 一种矩阵计算装置

图 1

AA REGISTER UNIT BB MATRIX COMPUTING UNIT CC STORAGE UNIT

(57) Abstract: A matrix calculation apparatus, comprising a storage unit, a register unit, and a matrix computing unit, matrices being stored in the storage unit and the stored addresses of matrices being stored in the register unit, and the matrix computing unit acquiring a matrix address in the register unit on the basis of a matrix computing command, then acquiring the corresponding matrix in the storage unit on the basis of the matrix address, and implementing matrix calculation on the basis of the acquired matrix in order to obtain matrix calculation results. The apparatus temporarily stores matrix data used in a calculation in a high-speed temporary memory, such that data of different widths can be supported more flexibly and effectively during a matrix calculation process, thus enhancing the execution performance of tasks involving a large amount of matrix calculation.

(57) 摘要: 一种矩阵计算装置，包括存储单元、寄存器单元和矩阵运算单元，存储单元中存储有矩阵，寄存器单元中存储有矩阵存储的地址矩阵运算单元根据矩阵运算指令在寄存器单元中获取矩阵地址，然后，根据该矩阵地址在存储单元中获取相应的矩阵，接着，根据获取的矩阵进行矩阵运算，得到矩阵运算结果。所述装置将参与计算的矩阵数据暂存在高速暂存存储器上，使得矩阵运算过程中可以更加灵活有效地支持不同宽度的数据，提升包含大量矩阵计算任务的执行性能。

WO 2017/124647 A1

一种矩阵计算装置

技术领域

本发明涉及一种矩阵运算装置，用于根据矩阵运算指令执行矩阵运算，能够很好地解决当前计算机领域越来越多的算法包含大量矩阵运算的问题。

背景技术

当前计算机领域有越来越多的算法涉及到矩阵运算，以人工神经网络算法为例，多种神经网络算法中都含有大量的矩阵运算。在神经网络中，输出神经元的运算表达式为 y=f(wx+b)，其中 w 是矩阵，x、b 是矩阵，计算输出矩阵 y 的过程为矩阵 w 与矩阵 x 相乘，加上矩阵 b，然后对得到的矩阵进行激活函数运算（即对矩阵中的每个元素进行激活函数运算）。因此，矩阵运算成为目前各种计算装置在设计之初都需要考虑的一个重要问题。

在现有技术中，一种进行矩阵运算的已知方案是使用通用处理器，该方法通过通用寄存器堆和通用功能部件来执行通用指令，从而执行矩阵运算。然而，该方法的缺点之一是单个通用处理器多用于标量计算，在进行矩阵运算时运算性能较低。而使用多个通用处理器并行执行时，通用处理器之间的相互通讯又有可能成为性能瓶颈。

在另一种现有技术中，使用图形处理器（GPU）来进行矩阵计算，其中，通过使用通用寄存器堆和通用流处理单元执行通用 SIMD 指令来进行矩阵运算。然而，上述方案中，GPU 片上缓存太小，在进行大规模矩阵运算时需要不断进行片外数据搬运，片外带宽成为了主要性能瓶颈。

在另一种现有技术中，使用专门定制的矩阵运算装置来进行矩阵计算，其中，使用定制的寄存器堆和定制的处理单元进行矩阵运算。然而，

目前已有的专用矩阵运算装置受限于寄存器堆，不能够灵活地支持不同长度的矩阵运算。

综上所述，现有的不管是片上多核通用处理器、片间互联通用处理器（单核或多核）、还是片间互联图形处理器都无法进行高效的矩阵运算，并且这些现有技术在处理矩阵运算问题时存在着代码量大，受限于片间通讯，片上缓存不够，支持的矩阵规模不够灵活等问题。

发明内容

（一）要解决的技术问题

本发明的目的在于，提供一种矩阵运算装置，解决现有技术中存在的受限于片间通讯、片上缓存不够、支持的矩阵长度不够灵活等问题。

（二）技术方案

本发明提供一种矩阵运算装置，用于根据矩阵运算指令执行矩阵运算，包括：

存储单元，用于存储矩阵；

寄存器单元，用于存储矩阵地址，其中，矩阵地址为矩阵在存储单元中存储的地址；

矩阵运算单元，用于获取矩阵运算指令，根据矩阵运算指令在寄存器单元中获取矩阵地址，然后，根据该矩阵地址在存储单元中获取相应的矩阵，接着，根据获取的矩阵进行矩阵运算，得到矩阵运算结果。

（三）有益效果

本发明提供的矩阵运算装置，将参与计算的矩阵数据暂存在高速暂存存储器上（Scratchpad Memory），使得矩阵运算过程中可以更加灵活有效地支持不同宽度的数据，提升包含大量矩阵计算任务的执行性能，本发明采用的指令具有精简的格式，使得指令集使用方便、支持的矩阵长度灵活。

附图说明

图 1 是本发明提供的矩阵运算装置的结构示意图。

图 2 是本发明提供的指令集的格式示意图。

图 3 是本发明实施例提供的矩阵运算装置的结构示意图。

图 4 是本发明实施例提供的矩阵运算装置执行矩阵点积指令的流程图。

图 5 为本发明实施例提供的矩阵运算装置进行幂乘法求解不可规约矩阵最大特征值对应特征向量的流程图。

图 6 为本发明实施例提供的矩阵运算单元的结构示意图。

具体实施方式

本发明提供一种矩阵计算装置，包括存储单元、寄存器单元和矩阵运算单元，存储单元中存储有矩阵，寄存器单元中存储有矩阵存储的地址矩阵运算单元根据矩阵运算指令在寄存器单元中获取矩阵地址，然后，根据该矩阵地址在存储单元中获取相应的矩阵，接着，根据获取的矩阵进行矩阵运算，得到矩阵运算结果。本发明将参与计算的矩阵数据暂存在高速暂存存储器上，使得矩阵运算过程中可以更加灵活有效地支持不同宽度的数据，提升包含大量矩阵计算任务的执行性能。

图 1 是本发明提供的矩阵运算装置的结构示意图，如图 1 所示，矩阵运算装置包括：

存储单元，用于存储矩阵，在一种实施方式中，该存储单元可以是高速暂存存储器，能够支持不同大小的矩阵数据；本发明将必要的计算数据暂存在高速暂存存储器上（Scratchpad Memory），使本运算装置在进行矩阵运算过程中可以更加灵活有效地支持不同宽度的数据。

寄存器单元，用于存储矩阵地址，其中，矩阵地址为矩阵在存储单元中存储的地址；在一种实施方式中，寄存器单元可以是标量寄存器堆，提供运算过程中所需的标量寄存器，标量寄存器不只存放矩阵地址，还存放有标量数据。当涉及到矩阵与标量的运算时，矩阵运算单元不仅要

从寄存器单元中获取矩阵地址，还要从寄存器单元中获取相应的标量。

矩阵运算单元，用于获取矩阵运算指令，根据矩阵运算指令在所述寄存器单元中获取矩阵地址，然后，根据该矩阵地址在存储单元中获取相应的矩阵，接着，根据获取的矩阵进行矩阵运算，得到矩阵运算结果，并将矩阵运算结果存储于存储单元中。矩阵运算单元包含包括矩阵加法部件、矩阵乘法部件、大小比较部件、非线性运算部件和矩阵标量乘法部件，并且，矩阵运算单元为多流水级结构，其中，矩阵加法部件和矩阵乘法部件处于第一流水级，大小比较部件处于第二流水级，非线性运算部件和矩阵标量乘法部件处于第三流水级。这些单元处于不同的流水级，当连续串行的多条矩阵运算指令的先后次序与相应单元所在流水级顺序一致时，可以更加高效地实现这一连串矩阵运算指令所要求的操作。矩阵运算单元负责装置的所有矩阵运算，包括但不限于矩阵加法操作、矩阵加标量操作、矩阵减法操作、矩阵减标量操作、矩阵乘法操作、矩阵乘标量操作、矩阵除法（对位相除）操作、矩阵与操作和矩阵或操作，矩阵运算指令被送往该运算单元执行。

根据本发明的一种实施方式，矩阵运算装置还包括：指令缓存单元，用于存储待执行的矩阵运算指令。指令在执行过程中，同时也被缓存在指令缓存单元中，当一条指令执行完之后，如果该指令同时也是指令缓存单元中未被提交指令中最早的一条指令，该指令将背提交，一旦提交，该条指令进行的操作对装置状态的改变将无法撤销。在一种实施方式中，指令缓存单元可以是重排序缓存。

根据本发明的一种实施方式，矩阵运算装置还包括：指令处理单元，用于从指令缓存单元获取矩阵运算指令，并对该矩阵运算指令进行处理后，提供给所述矩阵运算单元。其中，指令处理单元包括：

取指模块，用于从指令缓存单元中获取矩阵运算指令；

译码模块，用于对获取的矩阵运算指令进行译码；

指令队列，用于对译码后的矩阵运算指令进行顺序存储，考虑到不同指令在包含的寄存器上有可能存在依赖关系，用于缓存译码后的指令，当依赖关系被满足之后发射指令。

根据本发明的一种实施方式，矩阵运算装置还包括：依赖关系处理单元，用于在矩阵运算单元获取矩阵运算指令前，判断该矩阵运算指令与前一矩阵运算指令是否访问相同的矩阵，若是，将该矩阵运算指令存储在一存储队列中，待前一矩阵运算指令执行完毕后，将存储队列中的该矩阵运算指令提供给所述矩阵运算单元；否则，直接将该矩阵运算指令提供给所述矩阵运算单元。具体地，矩阵运算指令访问高速暂存存储器时，前后指令可能会访问同一块存储空间，为了保证指令执行结果的正确性，当前指令如果被检测到与之前的指令的数据存在依赖关系，该指令必须在存储队列内等待至依赖关系被消除。

根据本发明的一种实施方式，矩阵运算装置还包括：输入输出单元，用于将矩阵存储于存储单元，或者，从存储单元中获取矩阵运算结果。其中，输入输出单元可直接存储单元，负责从内存中读取矩阵数据或写入矩阵数据。

根据本发明的一种实施方式，用于本发明装置的指令集采用 Load/Store 结构，矩阵运算单元不会对内存中的数据进行操作。本指令集采用精简指令集架构，指令集只提供最基本的矩阵运算操作，复杂的矩阵运算都由这些简单指令通过组合进行模拟，使得可以在高时钟频率下单周期执行指令。另外，本指令集同时采用定长指令，使得本发明提出的矩阵运算装置在上一条指令的译码阶段对下一条指令进行取指。

图 2 是本发明提供的指令集的格式示意图，如图 2 所示，矩阵运算指令包括一操作码和至少一操作域，其中，操作码用于指示该矩阵运算指令的功能，矩阵运算单元通过识别该操作码可进行不同的矩阵运算，操作域用于指示该矩阵运算指令的数据信息，其中，数据信息可以是立即数或寄存器号，例如，要获取一个矩阵时，根据寄存器号可以在相应的寄存器中获取矩阵起始地址和矩阵长度，再根据矩阵起始地址和矩阵长度在存储单元中获取相应地址存放的矩阵。

指令集包含有不同功能的矩阵运算指令：

矩阵乘向量指令（MMV），根据该指令，装置从高速暂存存储器的指定地址取出指定大小的矩阵数据和向量数据，在矩阵运算单元中进行

矩阵乘向量的乘法运算，并将计算结果写回至高速暂存存储器的指定地址；值得说明的是，向量可以作为特殊形式的矩阵（只有一行元素的矩阵）存储于高速暂存存储器中。

向量乘矩阵指令（VMM），根据该指令，装置从高速暂存存储器的指定地址取出制定长度的向量数据和矩阵数据，在矩阵运算单元中进行向量乘矩阵的乘法运算，并将计算结果写回至高速暂存存储器的指定地址；值得说明的是，向量可以作为特殊形式的矩阵（只有一行元素的矩阵）存储于高速暂存存储器中。

矩阵乘标量指令（VMS），根据该指令，装置从高速暂存存储器的指定地址取出指定大小的矩阵数据，从标量寄存器堆的指定地址中取出指定大小的矩阵数据，在矩阵运算单元中进行标量乘矩阵的乘法运算，并将计算结果写回至高速暂存存储器的指定地址，需要说明的是，标量寄存器堆不仅存储有矩阵的地址，还存储有标量数据。

张量运算指令（TENS），根据该指令，装置从高速暂存存储器的两个指定地址取出分别取出指定大小的两块矩阵数据，在矩阵运算单元中对两矩阵数据进行张量运算，并将计算结果写回至高速暂存存储器的指定地址。

矩阵加法指令（MA），根据该指令，装置从高速暂存存储器的两个指定地址取出分别取出指定大小的两块矩阵数据，在矩阵运算单元中对两矩阵进行加法运算，并将计算结果写回至高速暂存存储器的指定地址。

矩阵减法指令（MS），根据该指令，装置从高速暂存存储器的两个指定地址取出分别取出指定大小的两块矩阵数据，在矩阵运算单元中对两矩阵进行减法运算，并将计算结果写回至高速暂存存储器的指定地址。

矩阵检索指令（MR），根据该指令，装置从高速暂存存储器的指定地址取出指定大小的向量数据，从高速暂存存储器的指定地址取出指定大小的矩阵数据，在矩阵运算单元中，该向量是索引向量，输出的向量中的第 i 个元素是以索引向量的第 i 个元素作为索引，在矩阵的第 i 列中

找到的数，该输出向量写回至高速暂存存储器的指定地址。

矩阵加载指令（ML），根据该指令，装置从指定外部源地址载入指定大小的数据至高速暂存存储器的指定地址。

矩阵存储指令（MS），根据该指令，装置将高速暂存存储器的指定地址的指定大小的矩阵数据存至外部目的地址处。

矩阵搬运指令（MMOVE）。根据该指令，装置将高速暂存存储器的指定地址的指定大小的矩阵数据存至高速暂存存储器的另一指定地址处。

为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明进一步详细说明。

图 3 是本发明实施例提供的矩阵运算装置的结构示意图，如图 3 所示，装置包括取指模块、译码模块、指令队列、标量寄存器堆、依赖关系处理单元、存储队列、重排序缓存、矩阵运算单元、高速暂存器、IO 内存存取模块；

取指模块，该模块负责从指令序列中取出下一条将要执行的指令，并将该指令传给译码模块；

译码模块，该模块负责对指令进行译码，并将译码后指令传给指令队列；

指令队列，考虑到不同指令在包含的标量寄存器上有可能存在依赖关系，用于缓存译码后的指令，当依赖关系被满足之后发射指令；

标量寄存器堆，提供装置在运算过程中所需的标量寄存器；

依赖关系处理单元，该模块处理处理指令与前一条指令可能存在的存储依赖关系。矩阵运算指令会访问高速暂存存储器，前后指令可能会访问同一块存储空间。为了保证指令执行结果的正确性，当前指令如果被检测到与之前的指令的数据存在依赖关系，该指令必须在存储队列内等待至依赖关系被消除。

存储队列，该模块是一个有序队列，与之前指令在数据上有依赖关系的指令被存储在该队列内直至存储关系被消除；

重排序缓存，指令在执行过程中，同时也被缓存在给模块中，当一

条指令执行完之后，如果该指令同时也是重排序缓存中未被提交指令中最早的一条指令，该指令将背提交。一旦提交，该条指令进行的操作对装置状态的改变将无法撤销；

矩阵运算单元，该模块负责装置的所有矩阵运算，包括但不限于矩阵加法操作、矩阵加标量操作、矩阵减法操作、矩阵减标量操作、矩阵乘法操作、矩阵乘标量操作、矩阵除法（对位相除）操作、矩阵与操作和矩阵或操作，矩阵运算指令被送往该运算单元执行；

高速暂存器，该模块是矩阵数据专用的暂存存储装置，能够支持不同大小的矩阵数据；

IO 内存存取模块，该模块用于直接访问高速暂存存储器，负责从高速暂存存储器中读取数据或写入数据。

图 4 是本发明实施例提供的矩阵运算装置执行矩阵乘向量指令的流程图，如图 4 所示，执行矩阵乘向量指令的过程包括：

S1，取指模块取出该条矩阵乘向量指令，并将该指令送往译码模块。

S2，译码模块对指令译码，并将指令送往指令队列。

S3，在指令队列中，该矩阵乘向量指令需要从标量寄存器堆中获取指令中五个操作域所对应的标量寄存器里的数据，包括输入向量地址、输入向量长度、输入矩阵地址、输出向量地址、输出向量长度。

S4，在取得需要的标量数据后，该指令被送往依赖关系处理单元。依赖关系处理单元分析该指令与前面的尚未执行结束的指令在数据上是否存在依赖关系。该条指令需要在存储队列中等待至其与前面的未执行结束的指令在数据上不再存在依赖关系为止。

S5，依赖关系不存在后，该条矩阵乘向量指令被送往矩阵运算单元。矩阵运算单元根据所需数据的地址和长度从高速暂存器中取出需要的矩阵和向量数据，然后在矩阵运算单元中完成乘法运算。

S6，运算完成后，将结果写回至高速暂存存储器的指定地址，同时重排序缓存中的该指令被提交。

图 5 为本发明实施例提供的矩阵运算装置进行幂乘法求解不可规约矩阵最大特征值对应特征向量的流程图，如图 5 所示，包括：

step1：设置任意 2-范数为 1 的起始向量 x；

step2：通过 IO 指令将起始向量 x 和待求解矩阵 A 分别存至装置内向量专用高速暂存存储器和矩阵专用高速暂存存储器的指定地址处；

step3：执行矩阵乘向量指令（MMV），从高速暂存存储器中将 A 和 x 读出，在矩阵运算单元中将矩阵 A 与向量 x 相乘，将得到的向量归一化之后与 x 比较，当差值仍大于阈值时，把 x 赋成该矩阵，返回 step2；

step4：得到的向量即不可规约矩阵 A 的最大特征值对应的特征向量，将输出向量存至指定地址，通过 IO 指令存至片外地址空间。

图 6 为本发明实施例提供的矩阵运算单元的结构示意图，如图 6 所示，矩阵运算单元包含包括矩阵加法部件、矩阵乘法部件、矩阵标量乘法部件和非线性运算部件等，另外，矩阵运算单元为多流水级结构，其中，矩阵乘法部件和矩阵标量乘法部件处于流水级 1，矩阵加法部件处于流水级 2，非线性运算部件处于流水级 3。这些单元处于不同的流水级，当连续串行的多条矩阵运算指令的先后次序与相应单元所在流水级顺序一致时，可以更加高效地实现这一连串矩阵运算指令所要求的操作。

综上所述，本发明提供矩阵运算装置，并配合相应的精简指令集架构，能够很好地解决当前计算机领域越来越多的算法包含大量矩阵运算的问题，相比于已有的传统解决方案，本发明可以具有指令集精简、使用方便、支持的矩阵规模灵活、片上缓存充足等优点。本发明可以用于多种包含大量矩阵运算的计算任务，包括目前表现十分出色的人工神经网络算法的反相训练和正向预测，以及传统的如求解不可规约矩阵最大特征值的幂乘法的数值计算方法。

以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权利要求

1、一种矩阵运算装置，用于根据矩阵运算指令执行矩阵运算，其特征在于，包括：

存储单元，用于存储矩阵；

寄存器单元，用于存储矩阵地址，其中，所述矩阵地址为矩阵在所述存储单元中存储的地址；

矩阵运算单元，用于获取矩阵运算指令，根据矩阵运算指令在所述寄存器单元中获取矩阵地址，然后，根据该矩阵地址在存储单元中获取相应的矩阵，接着，根据获取的矩阵进行矩阵运算，得到矩阵运算结果。

2、根据权利要求 1 所述的矩阵运算装置，其特征在于，还包括：

指令缓存单元，用于存储待执行的矩阵运算指令。

3、根据权利要求 2 所述的矩阵运算装置，其特征在于，还包括：

指令处理单元，用于从所述指令缓存单元获取矩阵运算指令，并对该矩阵运算指令进行处理后，提供给所述矩阵运算单元。

4、根据权利要求 3 所述的矩阵运算装置，其特征在于，所述指令处理单元包括：

取指模块，用于从所述指令缓存单元中获取矩阵运算指令；

译码模块，用于对获取的矩阵运算指令进行译码；

指令队列，用于对译码后的矩阵运算指令进行顺序存储。

5、根据权利要求 1 所述的矩阵运算装置，其特征在于，还包括：

依赖关系处理单元，用于在所述矩阵运算单元获取矩阵运算指令前，判断该矩阵运算指令与前一矩阵运算指令是否访问相同的矩阵，若是，则等待前一矩阵运算指令执行完毕后，将该矩阵运算指令提供给所述矩阵运算单元；否则，直接将该矩阵运算指令提供给所述矩阵运算单元。

6、根据权利要求 5 所述的矩阵运算装置，其特征在于，当该矩阵运算指令与前一矩阵运算指令访问相同的矩阵时，所述依赖关系处理单元将该矩阵运算指令存储在一存储队列中，待前一矩阵运算指令执行完

毕后，将存储队列中的该矩阵运算指令提供给所述矩阵运算单元。

7、根据权利要求 1 所述的矩阵运算装置，其特征在于，所述存储单元还用于存储所述矩阵运算结果。

8、根据权利要求 6 所述的矩阵运算装置，其特征在于，还包括：

输入输出单元，用于将矩阵存储于所述存储单元，或者，从所述存储单元中获取矩阵运算结果。

9、根据权利要求 6 所述的矩阵运算装置，其特征在于，所述存储单元为高速暂存存储器。

10、根据权利要求 1 所述的矩阵运算装置，其特征在于，所述矩阵运算指令包括一操作码和至少一操作域，其中，所述操作码用于指示该矩阵运算指令的功能，操作域用于指示该矩阵运算指令的数据信息。

11、根据权利要求 1 所述的矩阵运算装置，其特征在于，所述矩阵运算单元包含包括矩阵加法部件、矩阵乘法部件、矩阵标量乘法部件和非线性运算部件。

12、根据权利要求 11 所述的矩阵运算装置，其特征在于，所述矩阵运算单元为多流水级结构，其中，所述矩阵乘法部件和矩阵标量乘法部件处于第一流水级，矩阵加法部件处于第二流水级，非线性运算部件处于第三流水级。

图 1

操作码	寄存器或立即数	寄存器/立即数	...

图 2

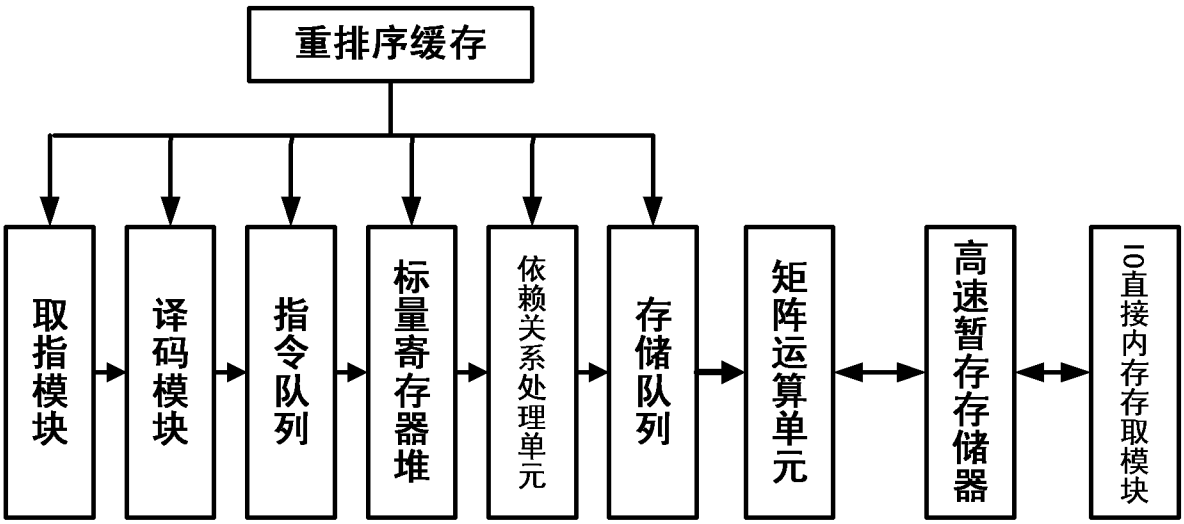

图 3

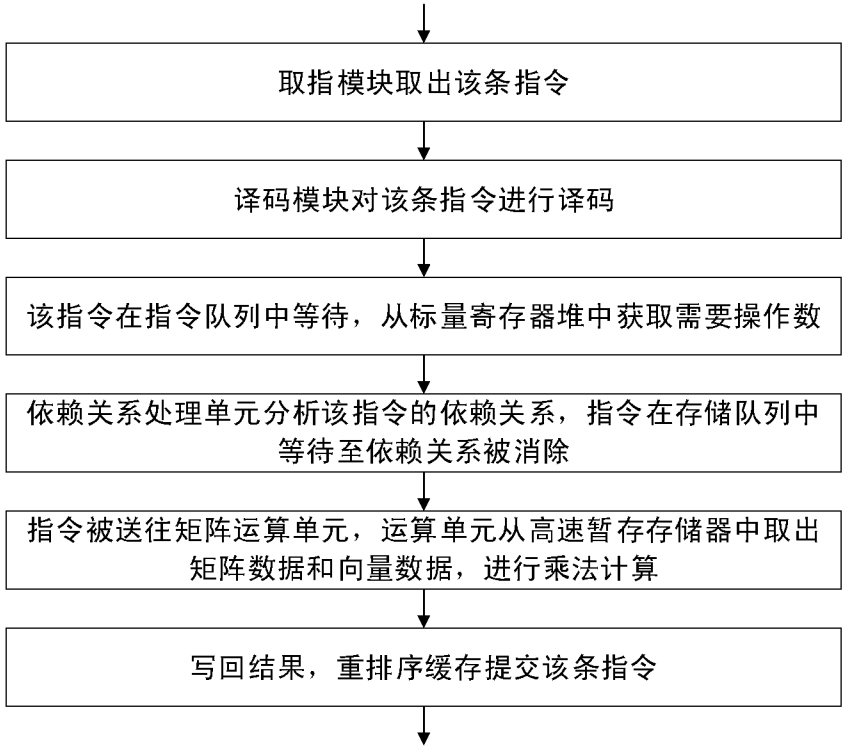
取指模块取出该条指令
译码模块对该条指令进行译码
该指令在指令队列中等待，从标量寄存器堆中获取需要操作数
依赖关系处理单元分析该指令的依赖关系，指令在存储队列中等待至依赖关系被消除
指令被送往矩阵运算单元，运算单元从高速暂存存储器中取出矩阵数据和向量数据，进行乘法计算
写回结果，重排序缓存提交该条指令

图 4

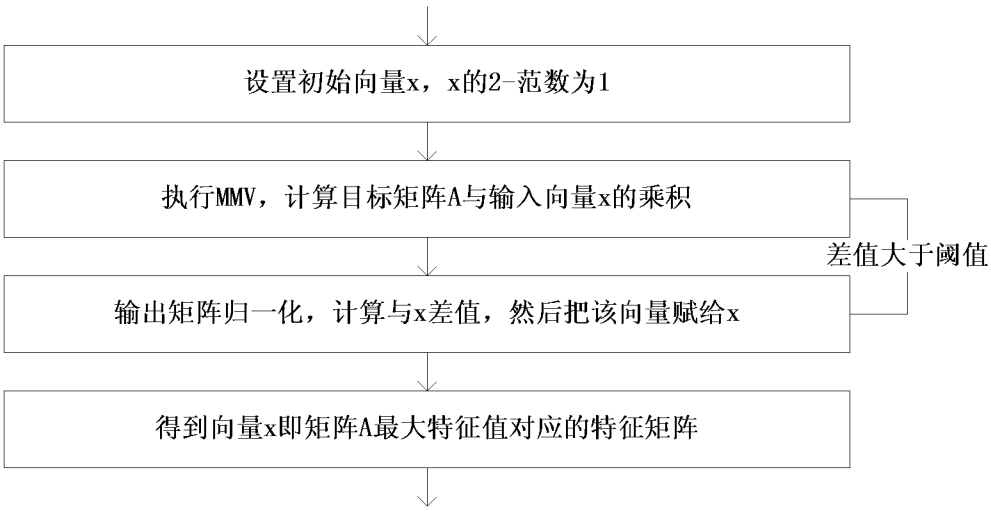
设置初始向量x，x的2-范数为1
执行MMV，计算目标矩阵A与输入向量x的乘积
差值大于阈值
输出矩阵归一化，计算与x差值，然后把该向量赋给x
得到向量x即矩阵A最大特征值对应的特征矩阵

图 5

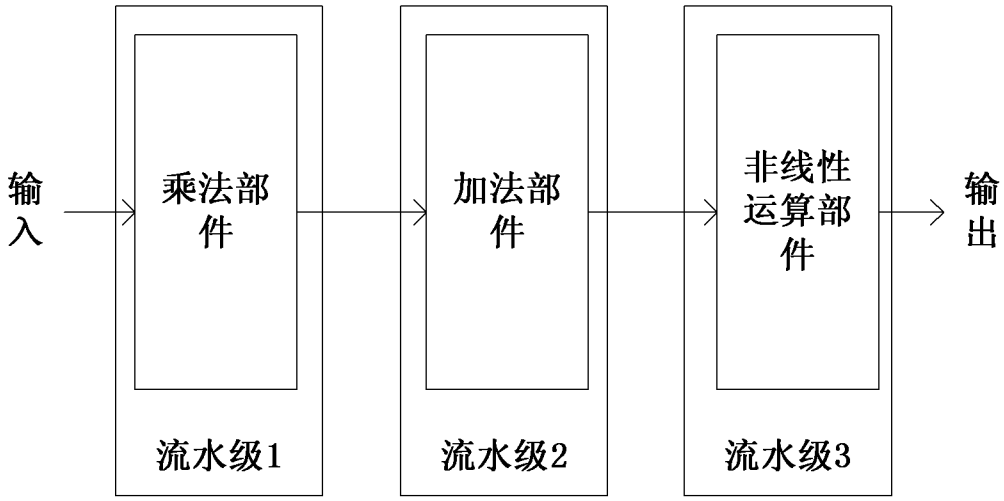

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/078546

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/16 (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)

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: matrix, operation?, comput+, storage, memory, register?, address, instruction

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 102750127 A (TSINGHUA UNIVERSITY), 24 October 2012 (24.10.2012), description, paragraphs [0035]-[0049], and figures 2	1-12
A	CN 103678257 A (SHANGHAI JIAO TONG UNIVERSITY), 26 March 2014 (26.03.2014), the whole document	1-12
A	CN 102360344 A (XI'AN JIAOTONG UNIVERSITY), 22 February 2012 (22.02.2012), the whole document	1-12
A	CN 101620524 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY), 06 January 2010 (06.01.2010), the whole document	1-12
A	JPH 04365170 A (MITSUBISHI ELECTRIC CORPORATION), 17 December 1992 (17.12.1992), the whole document	1-12

☐ 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
29 September 2016 (29.09.2016)	**31 October 2016 (31.10.2016)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **ZHAO, Ting** Telephone No.: (86-10) **62414434**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/078546

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102750127 A	24 October 2012	None	
CN 103678257 A	26 March 2014	None	
CN 102360344 A	22 February 2012	None	
CN 101620524 A	06 January 2010	None	
JPH 04365170 A	17 December 1992	None	

国际检索报告

国际申请号
PCT/CN2016/078546

A. 主题的分类

G06F 17/16(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）)

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: 矩阵, 运算, 计算, 存储, 寄存器, 地址, 指令, matrix, operation?, comput+, storage, memory, register?, address, instruction

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 102750127 A (清华大学) 2012年 10月 24日 (2012-10-24) 说明书第[0035]-[0049]段、图2	1-12
A	CN 103678257 A (上海交通大学) 2014年 3月 26日 (2014-03-26) 全文	1-12
A	CN 102360344 A (西安交通大学) 2012年 2月 22日 (2012-02-22) 全文	1-12
A	CN 101620524 A (中国人民解放军国防科学技术大学) 2010年 1月 6日 (2010-01-06) 全文	1-12
A	JPH 04365170 A (MITSUBISHI ELECTRIC CORPORATION) 1992年 12月 17日 (1992-12-17) 全文	1-12

☐ 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
2016年 9月 29日	2016年 10月 31日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	赵婷
传真号 (86-10)62019451	电话号码 (86-10)62414434

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/078546

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	102750127	A	2012年 10月 24日	无	
CN	103678257	A	2014年 3月 26日	无	
CN	102360344	A	2012年 2月 22日	无	
CN	101620524	A	2010年 1月 6日	无	
JPH	04365170	A	1992年 12月 17日	无	