(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
2018 年 2 月 8 日 (08.02.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/024093 A1

(51) 国际专利分类号:
G06N 3/063 (2006.01)

(21) 国际申请号: PCT/CN2017/093159

(22) 国际申请日: 2017 年 7 月 17 日 (17.07.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610640111.8 2016年8月5日 (05.08.2016) CN

(71) 申请人: 上海寒武纪信息科技有限公司 (SHANGHAI CAMBRICON INFORMATION TECHNOLOGIES LTD.) [CN/CN]; 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。

(72) 发明人: 陈天石(CHEN, Tianshi); 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。 郭崎(GUO, Qi); 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。 杜子东(DU, Zidong); 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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, 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,

(54) **Title:** OPERATION UNIT, METHOD AND DEVICE CAPABLE OF SUPPORTING OPERATION DATA OF DIFFERENT BIT WIDTHS

(54) **发明名称**: 一种能支持不同位宽运算数据的运算单元、方法及装置

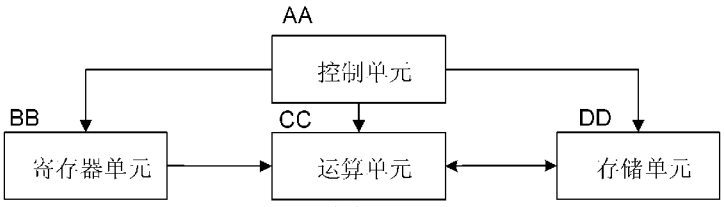

图 1

AA Control unit
BB Register unit
CC Operation unit
DD Storage unit

(57) **Abstract:** An operation unit, operation method and operation device for configuring a bit width of operation data involved in an operation by configuring a bit-width domain in a configuration instruction. The method comprises: upon executing an operation according to an instruction, firstly determining if an operator having a bit width identical to that of operation data indicated by an operand in the instruction is present; if yes, sending the operand directly to a corresponding operator; otherwise, generating an operator merging policy and merging multiple operators into a new operator according to the operator merging policy, such that a bit width of the new operator matches the bit width of the operand, and then sending the operand to the new operator; and enabling the operator acquiring the operand to execute a neural network operation/matrix operation/vector operation. The operation unit, operation method and operation device can support operations on operation data having different bit widths, thus realizing highly effective neural network operations, matrix operations and vector operations, while also reducing the quantity of operators and reducing hardware area at the same time.

WO 2018/024093 A1

[见续页]

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))。

(57) 摘要: 一种运算单元、运算方法及运算装置，通过配置指令中的位宽域来配置参与运算的运算数据位宽，在根据指令执行运算时，首先判断是否存在与指令中操作数所指示的运算数据位宽相同的运算器，如果是，将该操作数直接传给相应的运算器，否则，生成运算器合并策略，并根据运算器合并策略将多个运算器合并成一个新的运算器，以使该新的运算器的位宽符合该操作数的位宽，并将该操作数传给该新的运算器；再令获得该操作数的运算器执行神经网络运算/矩阵运算/向量运算。本运算单元、运算方法及运算装置能够支持不同位宽运算数据的运算，以实现高效的神经网络运算、矩阵运算及向量运算，同时，节省运算器的数量，减少硬件面积。

一种能支持不同位宽运算数据的运算单元、方法及装置

技术领域

本发明涉及计算机领域，尤其涉及一种运算单元、运算方法及运算装置，支持不同位宽运算数据的运算。

背景技术

人工神经网络（ANNs），简称神经网络（NNs），是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的。目前，神经网络在智能控制、机器学习等很多领域均获得长足发展。由于神经网络属于算法数学模型，其涉及大量的数学运算，因此如何快速、准确地执行神经网络运算是当前迫切需要解决的问题。其中，神经网络运算中各个参数在不同的层次中进行表示和运算时所需的位宽不同，使用不同位宽的运算器，可以减少实际的运算量，降低功耗；通过将低位宽的运算器合并成高位宽的运算器可以重复利用低位宽的运算器，减少运算器的数量，减少装置的面积。

发明内容

有鉴于此，本发明的目的在于提供一种运算单元、运算方法及运算装置，支持不同位宽运算数据的运算，以实现高效的神经网络运算、矩阵运算及向量运算。

本发明提供的运算单元、运算方法及运算装置，首先判断是否存在与运算数据位宽相同的运算器，如果是，将该运算数据直接传给相应的运算器，否则，生成运算器合并策略，并根据运算器合并策略将多个运算器合并成一个新的运算器，以使该新的运算器的位宽符合该运算数据

的位宽，并将该运算数据传给该新的运算器；再令获得该运算数据的运算器执行神经网络运算/矩阵运算/向量运算。

另外，本发明根据指令执行不同位宽运算数据的运算，指令采用了两种方式进行实现：一种为直接采用一条指令的方式，该指令中同时包括操作数和位宽域，运算单元能够直接根据该指令获取操作数和相应位宽的运算器，以执行相应的运算；另一种为采用两条指令的方式，运算单元先根据位宽配置指令获取或构造相应位宽的运算器，再根据运算指令获取操作数以执行相应的运算。

本发明具有以下有益效果：

1、本发明通过指令中的位宽域来指定运算数据的位宽，使得运算数据的位宽能够根据需要任意配置，针对某一位宽的运算数据，如果存在与该位宽符合的运算器，可直接调用该运算器执行运算，如果该运算数据的位宽过大，没有符合该位宽的运算器，可对多个较低位宽的运算器进行合并，以构造新的运算器，并利用新的运算器执行运算，能够支持不同位宽运算数据的运算，以实现高效的神经网络运算、矩阵运算及向量运算，同时，节省运算器的数量，减少硬件面积。

2、本发明采用高速暂存存储器，其能够实现对不同长度和不同位宽的运算数据（如：神经元、向量、矩阵）的存储。

附图说明

图 1 是本发明提供的运算装置的结构示意图。

图 2 是本发明提供的运算单元的结构示意图。

图 3 为本发明采用一条指令的方式执行运算的指令格式示意图。

图 4 是本发明的神经网络运算指令的格式示意图。

图 5 是本发明的矩阵-矩阵运算指令的格式示意图。

图 6 是本发明的向量-向量运算指令的格式示意图。

图 7 是本发明的矩阵-向量运算指令的格式示意图。

图 8 是本发明实施例的运算装置的结构示意图。

图 9 是本发明实施例中译码模块的结构示意图。

图 10 是本发明实施例的运算装置采用一条指令的方式执行运算的流程图。

图 11 是本发明采用两条指令的方式执行运算中位宽配置指令的格式示意图。

图 12 是本发明采用两条指令的方式执行运算中运算指令的格式示意图。

图 13 是本发明的神经网络位宽配置指令的格式示意图。

图 14 是本发明的神经网络运算指令的格式示意图。

图 15 是本发明的矩阵-矩阵位宽配置指令的格式示意图。

图 16 是本发明的矩阵-矩阵运算指令的格式示意图。

图 17 是本发明的向量-向量位宽配置指令的格式示意图。

图 18 是本发明的向量-向量运算指令的格式示意图。

图 19 是本发明的矩阵-向量位宽配置指令的格式示意图。

图 20 是本发明的矩阵-向量运算指令的格式示意图。

图 21 是本发明实施例的运算装置采用两条指令的方式执行运算的流程图。

具体实施方式

本发明公开了一种能支持不同位宽运算数据的运算单元、运算方法及运算装置，通过配置指令中的位宽域来配置参与运算的运算数据位宽，在根据指令执行运算时，首先判断是否存在与运算数据位宽相同的运算器，如果是，将该运算数据直接传给相应的运算器，否则，生成运算器合并策略，并根据运算器合并策略将多个运算器合并成一个新的运算器，以使该新的运算器的位宽符合该运算数据的位宽，并将该运算数据传给该新的运算器；再令获得该运算数据的运算器执行神经网络运算/矩阵运算/向量运算。本发明能够支持不同位宽运算数据的运算，以实现高效的神经网络运算、矩阵运算及向量运算，同时，节省运算器的数

量，减少硬件面积。

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 1 是本发明提供的运算装置的结构示意图，如图 1 所示，该运算装置包括：

存储单元，用于存储神经元/矩阵/向量，在一实施方式中，该存储单元可以是高速暂存存储器（Scratchpad Memory），能够支持不同长度和不同位宽的神经元/矩阵/向量数据，将必要的运算数据暂存在高速暂存存储器上，使本运算装置在进行神经网络运算以及矩阵/向量运算过程中可以更加灵活有效地支持不同长度和不同位宽的数据。高速暂存存储器可以通过各种不同存储器件（SRAM、eDRAM、DRAM、忆阻器、3D-DRAM 或非易失存储等）实现。

寄存器单元，用于存储神经元/矩阵/向量地址，其中：神经元地址为神经元在存储单元中存储的地址、矩阵地址为矩阵在存储单元中存储的地址、向量地址为向量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神经元/矩阵/向量地址，还存放有标量数据。当涉及到矩阵/向量与标量的运算时，运算单元不仅要从寄存器单元中获取矩阵/向量地址，还要从寄存器单元中获取相应的标量。

控制单元，用于控制装置中各个模块的行为。在一实施方式中，控制单元读取准备好的指令，进行译码生成多条微指令，发送给装置中的其他模块，其他模块根据得到的微指令执行相应的操作。

运算单元，用于获取指令，根据指令在寄存器单元中获取神经元/矩阵/向量地址，然后，根据该神经元/矩阵/向量地址在存储单元中获取相应的神经元/矩阵/向量，从而对该运算数据（神经元/矩阵/向量）执行运算。运算单元执行的运算包括但不限于：卷积神经网络正向运算操作、卷积神经网络训练操作、神经网络 Pooling 运算操作、full connection 神经网络正向运算操作、full connection 神经网络训练操作、batch normalization 运算操作、RBM 神经网络运算操作、矩阵-向量乘运算操

作、矩阵-矩阵加/减运算操作、向量外积（张量）运算操作、向量内积运算操作、向量四则运算操作、向量逻辑运算操作、向量超越函数运算操作、向量比较运算操作、求向量最大/最小值运算操作、向量循环移位运算操作、生成服从一定分布的随机向量运算操作。

运算单元在执行运算的过程中，根据指令中操作数所指示的运算数据的位宽，选择相应的一个或多个运算器以执行运算，其中，一个或多个运算器具有不同的位宽，例如，有的运算器支持 16 位的数据运算，有的运算器支持 32 位的数据运算，运算器实质上可以是向量乘法部件、累加部件和标量乘法部件等。如图 2 所示，运算单元包括判断子模块、运算器合并子模块和运算子模块；

判断子模块用于判断是否存在与该操作数所指示的运算数据位宽相同的运算器，如果是，将该操作数传给相应的运算器，否则，将运算器合并策略及该操作数传递给运算器合并子模块；

运算器合并子模块用于根据运算器合并策略将多个运算器合并成一个新的运算器，以使该新的运算器的位宽符合该操作数的位宽，并将该操作数传给该新的运算器。具体的，运算器合并策略是指优先选用较大位宽的运算器进行组合。当存在与所需位宽相同的运算器时，直接使用对应的运算器；若不存在，则选用比所需运算器位宽小且最为接近的可用的运算器进行组合。例如，可用的用于组合的运算器位宽分别为 8 位、16 位、32 位时，当所需的运算器的位宽为 32 位时，直接使用 32 位运算器；当所需的运算器的位宽为 64 位时，使用两个 32 位运算器进行合并；当所需的运算器的位宽为 48 位时，使用一个 32 位运算器和一个 16 位运算器进行合并；当所需的运算器的位宽为 40 位时，则选用一个 32 位运算器和一个 8 位运算器进行合并。

运算子模块用于令获得该操作数的运算器执行运算。

本发明的指令采用了两种方式进行实现：一种为直接采用一条指令的方式，该指令中同时包括操作数和位宽域，运算单元能够直接根据该指令获取操作数和相应位宽的运算器，以执行相应的运算；另一种为采用两条指令的方式，运算单元先根据位宽配置指令获取或构造相应位宽

的运算器，再根据运算指令获取操作数以执行相应的运算。

需要说明的是，本发明指令集采用 Load/Store 结构，运算单元不会对内存中的数据进行操作。本指令集采用超长指令字架构，通过对指令进行不同的配置可以完成复杂的神经网络运算，也可以完成简单的矩阵/向量运算。另外，本指令集同时采用定长指令，使得本发明的神经网络运算以及矩阵/向量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 3 示出了本发明采用一条指令的方式执行运算的指令格式示意图，如图 3 所示，指令包括至少一操作码和至少 3 个操作数和至少 2 个位宽域，其中，位宽域与在运算器中运算时操作数的种类数量相同；其中，操作码用于指示该运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的运算，操作数用于指示该运算指令的数据信息，位宽域用于指明对应操作数的位宽；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。

图 4 是本发明的神经网络运算指令的格式示意图，其为图 3 指令的实例化指令，如图 4 所示，神经网络运算指令包括至少一操作码和 16 个操作数和 4 个位宽域，其中，操作码用于指示该神经网络运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的神经网络运算，操作数用于指示该神经网络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，位宽域用于指明操作数在运算中所对应的位宽，同时，位宽域用于指明运算过程中所对应的运算器的位宽以及是否需要将低位宽运算器合并为高位宽运算器。

图 5 是本发明的矩阵-矩阵运算指令的格式示意图，其为图 3 指令的实例化指令，如图 5 所示，矩阵-矩阵运算指令包括至少一操作码和至少 4 个操作数和 2 个位宽域，其中，操作码用于指示该矩阵-矩阵运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的矩阵运算，操作数用于指示该矩阵-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，位宽域用于指明操作数在运算中所对应的位宽，同时，位宽域用于指明运算过程中所对应的运算器的位宽以及是否需要将低位宽运算器合并为高位宽运算器。

图 6 是本发明的向量-向量运算指令的格式示意图，其为图 3 指令的实例化指令，如图 6 所示，向量-向量运算指令包括至少一操作码和至少 3 个操作数和至少 2 个位宽域，其中，操作码用于指示该向量-向量运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的向量运算，操作数用于指示该向量-向量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，位宽域用于指明操作数在运算中所对应的位宽，同时，位宽域用于指明运算过程中所对应的运算器的位宽以及是否需要将低位宽运算器合并为高位宽运算器。

图 7 是本发明的矩阵-向量运算指令的格式示意图，其为图 3 指令的实例化指令，如图 7 所示，矩阵-向量运算指令包括至少一操作码和至少 6 个操作数和至少 3 个位宽域，其中，操作码用于指示该矩阵-向量运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的矩阵和向量运算，操作数用于指示该矩阵-向量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，位宽域用于指明操作数在运算中所对应的位宽，同时，位宽域用于指明运算过程中所对应的运算器的位宽以及是否需要将低位宽运算器合并为高位宽运算器。

图 8 是本发明一优选实施例的运算装置的结构示意图，如图 8 所示，该装置包括取指模块、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、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；如图 9 所示，该译码模块包括：指令接受模块、微指令生成模块、微指令队列、微指令发射模块；其中，指令接受模块负责接受从取指模块取得的指令；微指令译码模块将指令接受模块获得的指令译码成控制

各个功能部件的微指令；微指令队列用于存放从微指令译码模块发送的微指令；微指令发射模块负责将微指令发射到各个功能部件；

指令队列，用于顺序缓存译码后的指令，送往依赖关系处理单元；

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，直至依赖关系消除之后，提交指令。

重排序缓存，指令在执行过程中，同时也被缓存在该模块中，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，该指令将被提交。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；该重排序缓存里的指令起到占位的作用，当它包含的第一条指令存在数据依赖时，那么该指令就不会提交（释放）；尽管后面会有很多指令不断进入，但是只能接受部分指令（受重排序缓存大小控制），直到第一条指令被提交，整个运算过程才会顺利进行。

运算单元，该模块负责装置的所有的神经网络运算和矩阵/向量运算操作，包括但不限于：卷积神经网络正向运算操作、卷积神经网络训练操作、神经网络 Pooling 运算操作、full connection 神经网络正向运算操作、full connection 神经网络训练操作、batch normalization 运算操作、RBM 神经网络运算操作、矩阵-向量乘运算操作、矩阵-矩阵加/减运算操作、向量外积（张量）运算操作、向量内积运算操作、向量四则运算操作、向量逻辑运算操作、向量超越函数运算操作、向量比较运算操作、求向量最大/最小值运算操作、向量循环移位运算操作、生成服从一定分布的随机向量运算操作。运算指令被送往该运算单元执行，首先，运算单元判断是否有与指令中操作数对应的位宽域长度相同的运算器，如果

有，选用对应的运算器，如果没有，通过多个低位宽的运算器合并的方式构成所需位宽的运算器，然后，根据指令中操作码对运算数用选择的运算器进行对应的运算，得出相应的结果；

高速暂存存储器，该模块是数据专用的暂存存储装置，能够支持不同长度和不同位宽的数据；

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 10 是本发明实施例的运算装置采用一条指令的方式执行运算的流程图。如图 10 所示，过程包括：

S1，取指模块取出指令，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，在译码模块中，指令被送往指令接受模块。

S4，指令接受模块将指令发送到微指令生成模块，进行微指令生成。

S5，微指令生成模块从标量寄存器堆里获取指令的神经网络运算操作码和神经网络运算操作数，同时将指令译码成控制各个功能部件的微指令，送往微指令发射队列。

S6，在取得需要的数据后，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。

S7，依赖关系不存在后，该条神经网络运算以及矩阵/向量指令对应的微指令被送往运算单元等功能部件。

S8，运算单元根据所需数据的地址和大小从高速暂存存储器中取出需要的数据，然后判断是否有与指令中位宽域相同的运算器，如果有，则选用匹配的运算器完成指令对应的运算，如果没有，则通过将低位宽的运算器合并的方式组成一个所需位宽的运算器来完成指令对应的运算。

S9，运算完成后，将输出数据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 11 和图 12 示出了本发明采用两条指令的方式执行运算的指令格式示意图，其中，图 11 是位宽配置指令的格式示意图，位宽配置指令包括至少一操作码至少 2 个位宽域，用于指明下条运算指令所使用的运算器的位宽。图 12 是运算指令的格式示意图，运算指令包括至少一操作码至少 3 个操作数，其中，操作码用于指示该运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的运算，操作数用于指示该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。

图 13~14 是图 11~12 的实例化，其分别为神经网络位宽配置指令和神经网络运算指令的格式示意图，如图 13~14 所示，位宽配置指令包括至少一操作码至少 4 个位宽域，用于指明下条神经网络运算指令所使用的运算器的位宽。配置指令包括至少一操作码和 16 个操作数，其中，操作码用于指示该神经网络运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的神经网络运算，操作数用于指示该神经网络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。

图 15~16 是图 11~12 的实例化，其分别为矩阵-矩阵位宽配置指令和矩阵-矩阵运算指令的格式示意图，如图 15~16 所示，位宽配置指令包括至少一操作码至少 2 个位宽域，用于指明下条矩阵-矩阵运算指令所使用的运算器的位宽。矩阵-矩阵运算指令包括至少一操作码和至少 4 个操作数。其中，操作码用于指示该矩阵-矩阵运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的矩阵运算，操作数用于指示该矩阵-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。

图 17~18 是图 11~12 的实例化，其分别为向量-向量位宽配置指令和向量-向量运算指令的格式示意图，如图 17~18 所示，位宽配置指令包括至少一操作码至少 2 个位宽域，用于指明下条向量-向量运算指令所使用的运算器的位宽。向量-向量运算指令包括至少一操作码和至少 3 个操作数，其中，操作码用于指示该向量-向量运算指令的功能，运算单元通过

识别一个或多个操作码可进行不同的向量运算，操作数用于指示该向量-向量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。

图 19~20 是图 11~12 的实例化，其分别为矩阵-向量位宽配置指令和矩阵-向量运算指令的格式示意图，如图 19~20 所示，位宽配置指令包括至少一操作码至少 3 个位宽域，用于指明下条向量-向量运算指令所使用的运算器的位宽。矩阵-向量运算指令包括至少一操作码和至少 6 个操作数，其中，操作码用于指示该矩阵-向量运算指令的功能，运算单元通过识别一个或多个操作码可进行不同的矩阵和向量运算，操作数用于指示该矩阵-向量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。

图 21 是本发明实施例的运算装置采用两条指令的方式执行运算的流程图。如图 21 所示，过程包括：

步骤 S1，取指模块取出一条位宽配置指令，并将指令送往译码模块；

步骤 S2，译码模块对所述指令译码，并将所述指令送往指令队列；

步骤 S3，在译码模块，所述指令被送往指令接受模块；

步骤 S4，指令接收模块将所述指令发送到微指令译码模块，进行微指令译码；

步骤 S5，微指令译码模块将指令译码成控制运算单元选定指定位宽的运算器的微指令，发送到微指令发射队列；

步骤 S6，取指模块取出一条神经网络运算以及矩阵/向量指令，并将所述指令送往译码模块；

步骤 S7，译码模块对所述指令译码，并将所述指令送往指令队列；

步骤 S8，在译码模块中，所述指令被送往指令接受模块；

步骤 S9，指令接受模块将所述指令发送到微指令译码模块，进行微指令译码；

步骤 S10，微指令译码模块从标量寄存器堆里获取所述指令的神经网络运算操作码和神经网络运算操作数，同时将所述指令译码成控制各个功能部件的微指令，送往微指令发射队列；

步骤 S11，在取得需要的数据后，所述指令被送往依赖关系处理单

元；依赖关系处理单元分析所述指令与之前尚未执行完的指令在数据上是否存在依赖关系，如果存在，则所述指令需要在存储队列中等待至其与之前未执行完的指令在数据上不再存在依赖关系为止；

步骤 S12，将所述指令对应的微指令以及之前的指定运算器位宽的微指令送往运算单元；

步骤 S13，运算单元根据所需数据的地址和大小从高速暂存存储器中取出需要的数据;然后判断是否有与位宽指定指令中位宽域相同的运算器，如果有，则选用匹配的运算器完成所述指令对应的神经网络运算和/或矩阵/向量运算，如果没有，则通过将低位宽的运算器合并的方式组成一个所需位宽的运算器来完成所述指令对应的神经网络运算和/或矩阵/向量运算；

S14，运算完成后，将输出数据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。

综上所述，本发明公开了一种运算器位宽可配置的用于执行神经网络运算以及矩阵/向量运算的装置和方法，配合相应的指令，能够很好地解决当前计算机领域神经网络算法和大量矩阵/向量运算的问题，相比于已有的传统解决方案，本发明可以具有指令可配置、使用方便、运算器的位宽可以选择，多个运算器可以合并，并通过专用位宽配置指令和在运算指令上指定位宽域两种方式来实现运算器位宽的配置，支持的神经网络规模和矩阵/向量位宽和规模灵活、片上缓存充足，运算器可合并等优点。

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、一种运算单元，用于根据运算数据的位宽，选择相应的一个或多个运算器以执行运算，其中，所述一个或多个运算器具有不同的位宽，其特征在于，运算单元包括判断子模块、运算器合并子模块和运算子模块；

判断子模块用于判断是否存在与该运算数据位宽相同的运算器，如果是，将该运算数据传给相应的运算器，否则，将运算器合并策略及该运算数据传递给运算器合并子模块；

运算器合并子模块用于根据运算器合并策略将多个运算器合并成一个新的运算器，以使该新的运算器的位宽符合运算数据的位宽，并将该运算数据传给该新的运算器；

运算子模块用于令获得该运算数据的运算器执行运算。

2、根据权利要求 1 所述的运算单元，其特征在于，所述运算单元根据一指令执行运算，其中，所述指令包括：

操作码，用于指示该指令的运算类型；

操作数，用于作为运算数据或用于指示运算数据的存储地址；

位宽域，用于指示运算数据的位宽；

所述运算单元执行该指令，根据指令中的位宽域确定运算数据的位宽，并选择相应的运算器，然后将指令中的操作数传给相应的运算器，运算器根据操作数获取运算数据，并执行操作码所指示的运算。

3、根据权利要求 1 所述的运算单元，其特征在于，所述运算单元根据位宽配置指令和运算指令执行运算，所述位宽配置指令包括操作码和位宽域，所述运算指令包括操作码和操作数，其中，

所述操作码用于指示该指令的运算类型；

所述操作数用于作为运算数据或用于指示运算数据的存储地址；

所述位宽域用于指示指令中各个操作数的位宽；

所述运算单元依次执行位宽配置指令和运算指令，根据位宽配置指令中的位宽域确定运算指令中操作数的位宽，并选择相应的运算器，然

后将运算指令中的操作数传给相应的运算器，运算器根据操作数获取运算数据，并执行操作码所指示的运算。

4、根据权利要求 1 所述的运算单元，其特征在于，所述运算器合并策略为，合并一个或多个最接近运算数据位宽的运算器。

5、根据权利要求 1 所述的运算单元，其特征在于，所述操作数为运算数据或运算数据存储位置，所述运算器根据该操作数获得相应的运算数据后，执行运算。

6、根据权利要求 1 所述的运算单元，其特征在于，所述运算数据为向量、矩阵和神经元中的一种。

7、一种运算方法，用于根据运算数据的位宽，选择相应的一个或多个运算器以执行运算，其中，所述一个或多个运算器具有不同的位宽，其特征在于，方法包括：

S1，判断是否存在与该运算数据位宽相同的运算器，如果是，将该运算数据传给相应的运算器，然后执行步骤 S3，否则，生成运算器合并策略并执行步骤 S2；

S2，根据运算器合并策略将多个运算器合并成一个新的运算器，以使该新的运算器的位宽符合该运算数据的位宽，并将该运算数据传给该新的运算器；

S3，令获得该运算数据的运算器执行运算。

8、根据权利要求 7 所述的运算单元，其特征在于，所述运算单元根据一指令执行运算，其中，所述指令包括：

操作码，用于指示该指令的运算类型；

操作数，用于作为运算数据或用于指示运算数据的存储地址；

位宽域，用于指示运算数据的位宽；

所述运算单元执行该指令，根据指令中的位宽域确定运算数据的位宽，并选择相应的运算器，然后将指令中的操作数传给相应的运算器，运算器根据操作数获取运算数据，并执行操作码所指示的运算。

9、根据权利要求 7 所述的运算单元，其特征在于，所述运算单元根据位宽配置指令和运算指令执行运算，所述位宽配置指令包括操作码

和位宽域，所述运算指令包括操作码和操作数，其中，

所述操作码用于指示该指令的运算类型；

所述操作数用于作为运算数据或用于指示运算数据的存储地址；

所述位宽域用于指示指令中各个操作数的位宽；

所述运算单元依次执行位宽配置指令和运算指令，根据位宽配置指令中的位宽域确定运算指令中操作数的位宽，并选择相应的运算器，然后将运算指令中的操作数传给相应的运算器，运算器根据操作数获取运算数据，并执行操作码所指示的运算。

10、根据权利要求 7 所述的运算单元，其特征在于，所述运算器合并策略为，合并一个或多个最接近运算数据位宽的运算器。

11、根据权利要求 7 所述的运算方法，其特征在于，所述操作数为运算数据或运算数据存储位置，所述运算器根据该操作数获得相应的运算数据后，执行运算。

12、根据权利要求 7 所述的运算方法，其特征在于，所述运算数据为向量、矩阵和神经元中的一种。

13、一种运算装置，其特征在于，包括：

权利要求 1-5 任意一项所述的运算单元；

存储单元，用于存储所述运算数据；

寄存器单元，用于存储所述运算数据的地址；

控制单元，用于对运算单元、存储单元及寄存器单元进行控制，以使运算单元根据指令中的操作数访问寄存器单元，以获取运算数据的地址，并根据该运算数据的地址访问存储单元，以获取该运算数据，从而对该运算数据执行运算。

14、根据权利要求 13 所述的运算装置，其特征在于，所述存储单元为高速暂存存储器。

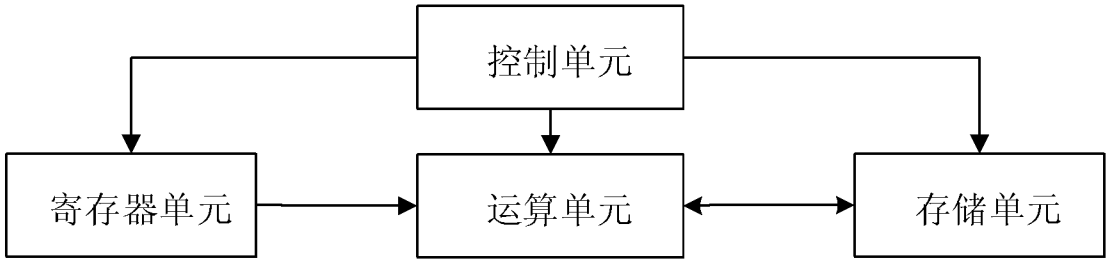

图 1

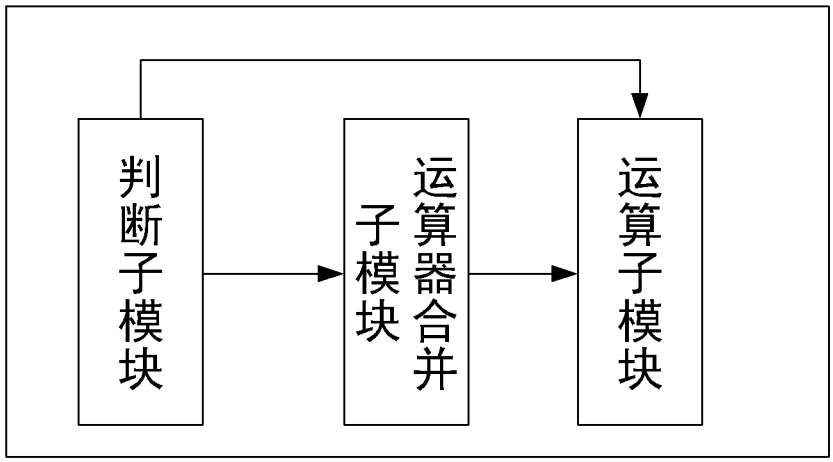

图 2

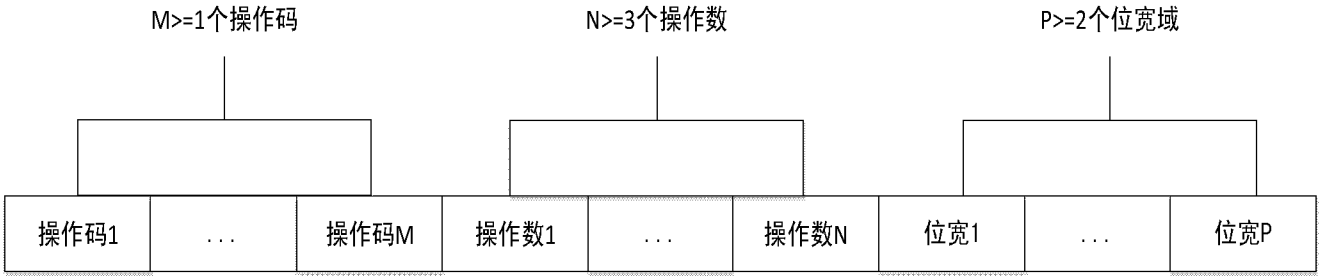
M>=1个操作码
N>=3个操作数
P>=2个位宽域
操作码1
. . .
操作码M
操作数1
. . .
操作数N
位宽1
. . .
位宽P

图 3

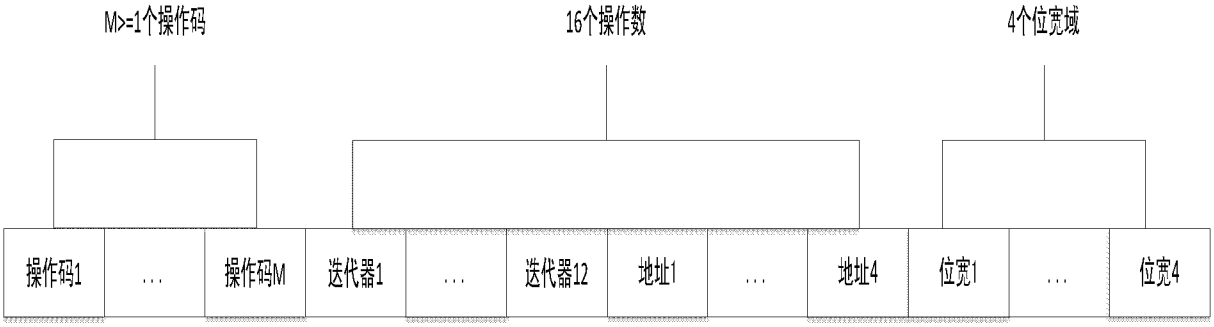
M>=1个操作码
16个操作数
4个位宽域
操作码1
. . .
操作码M
迭代器1
. . .
迭代器12
地址1
. . .
地址4
位宽1
. . .
位宽4

图 4

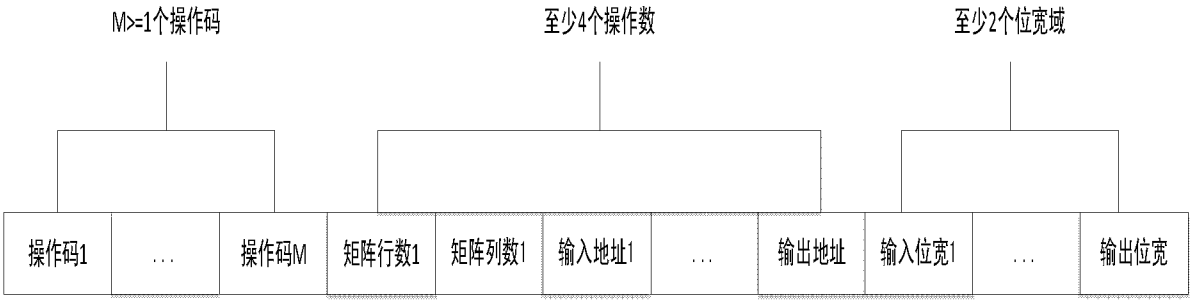
M>=1个操作码
至少4个操作数
至少2个位宽域
操作码1
. . .
操作码M
矩阵行数1
矩阵列数1
输入地址1
. . .
输出地址
输入位宽1
. . .
输出位宽

图 5

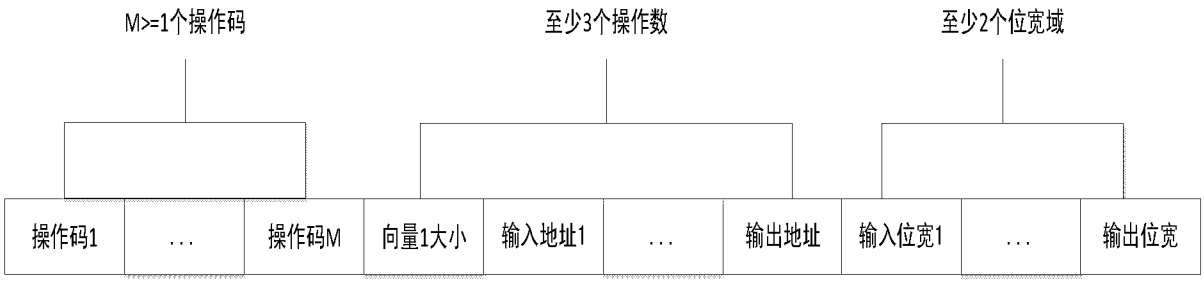
M>=1个操作码
至少3个操作数
至少2个位宽域
操作码1
...
操作码M
向量1大小
输入地址1
...
输出地址
输入位宽1
...
输出位宽

图 6

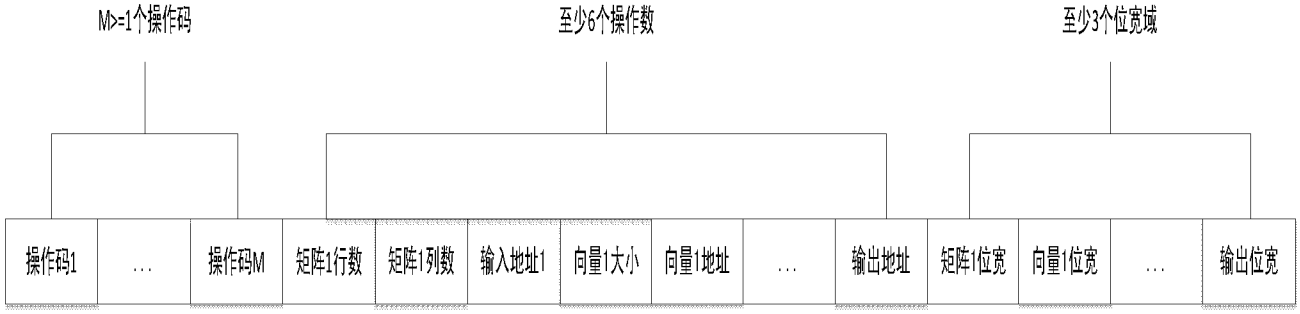
M>=1个操作码
至少6个操作数
至少3个位宽域
操作码1
...
操作码M
矩阵1行数
矩阵1列数
输入地址1
向量1大小
向量1地址
...
输出地址
矩阵1位宽
向量1位宽
...
输出位宽

图 7

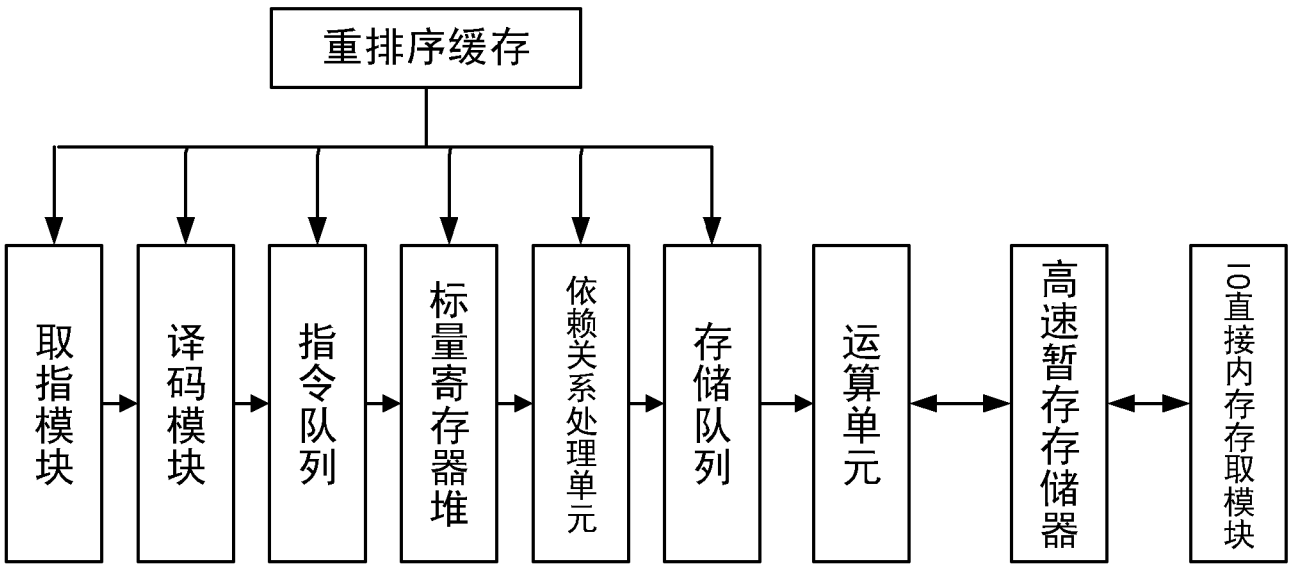
重排序缓存
取指模块
译码模块
指令队列
标量寄存器堆
依赖关系处理单元
存储队列
运算单元
高速暂存存储器
I/O直接内存存取模块

图 8

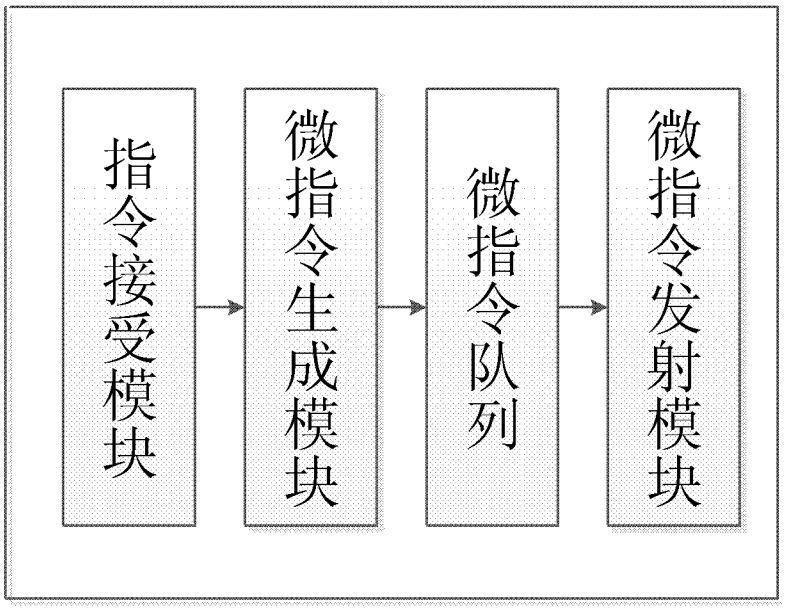

图 9

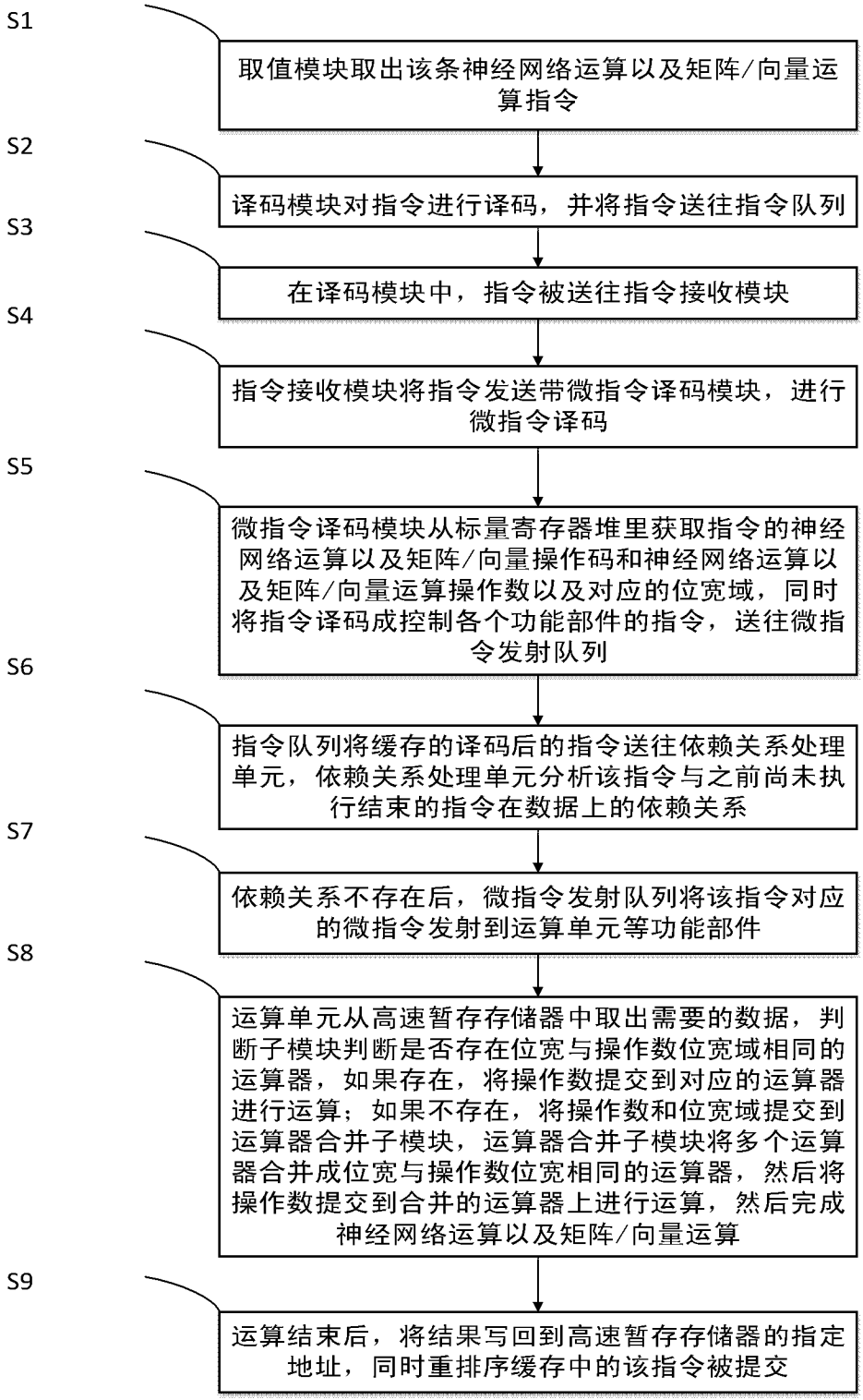
S1
取值模块取出该条神经网络运算以及矩阵/向量运算指令
S2
译码模块对指令进行译码，并将指令送往指令队列
S3
在译码模块中，指令被送往指令接收模块
S4
指令接收模块将指令发送带微指令译码模块，进行微指令译码
S5
微指令译码模块从标量寄存器堆里获取指令的神经网络运算以及矩阵/向量操作码和神经网络运算以及矩阵/向量运算操作数以及对应的位宽域，同时将指令译码成控制各个功能部件的指令，送往微指令发射队列
S6
指令队列将缓存的译码后的指令送往依赖关系处理单元，依赖关系处理单元分析该指令与之前尚未执行结束的指令在数据上的依赖关系
S7
依赖关系不存在后，微指令发射队列将该指令对应的微指令发射到运算单元等功能部件
S8
运算单元从高速暂存存储器中取出需要的数据，判断子模块判断是否存在位宽与操作数位宽域相同的运算器，如果存在，将操作数提交到对应的运算器进行运算；如果不存在，将操作数和位宽域提交到运算器合并子模块，运算器合并子模块将多个运算器合并成位宽与操作数位宽相同的运算器，然后将操作数提交到合并的运算器上进行运算，然后完成神经网络运算以及矩阵/向量运算
S9
运算结束后，将结果写回到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 10

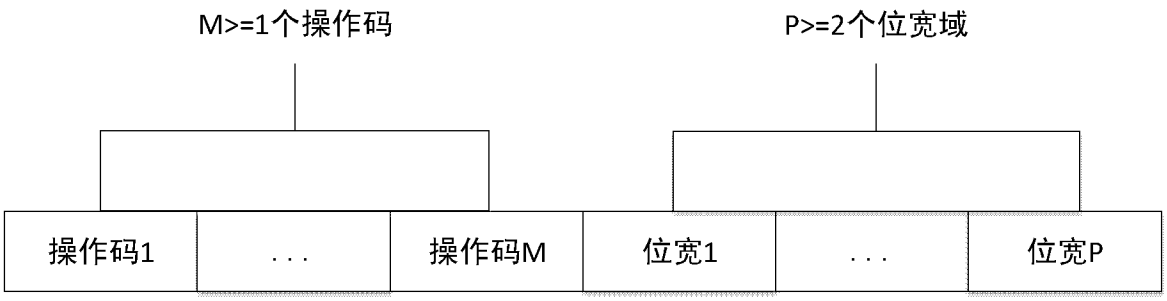

图 11

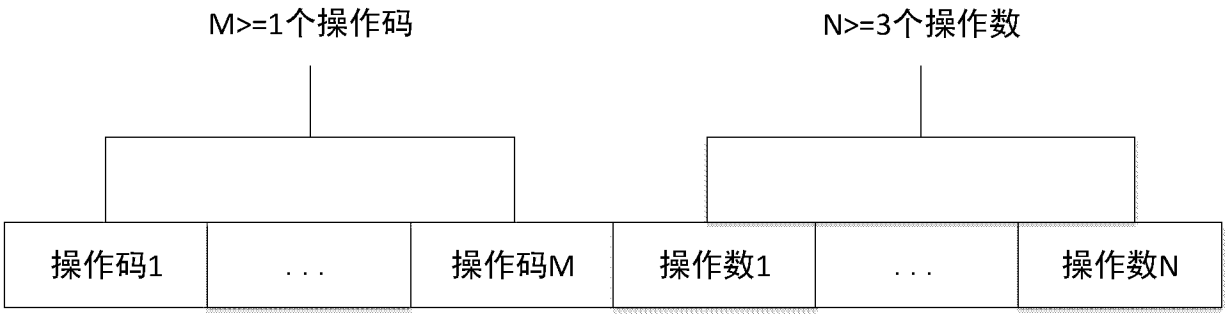

图 12

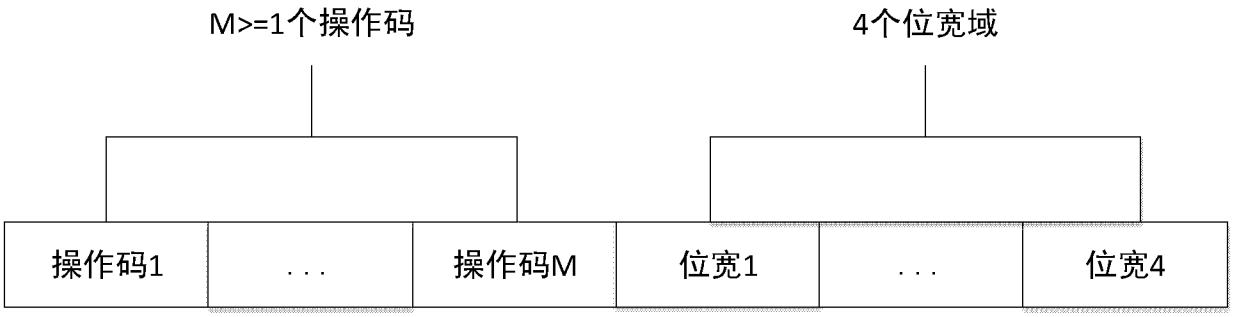

图 13

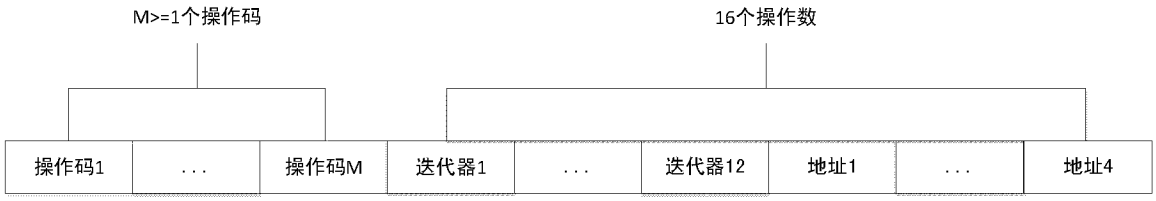

图 14

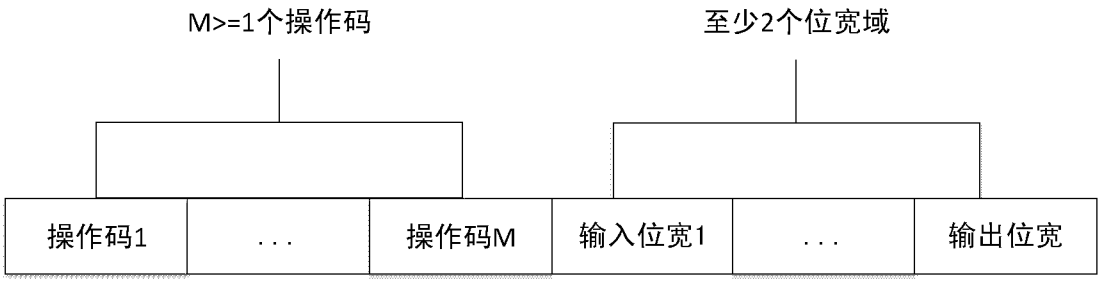

图 15

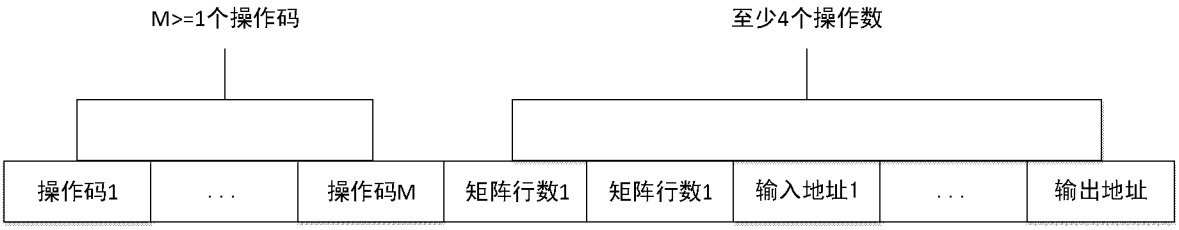

图 16

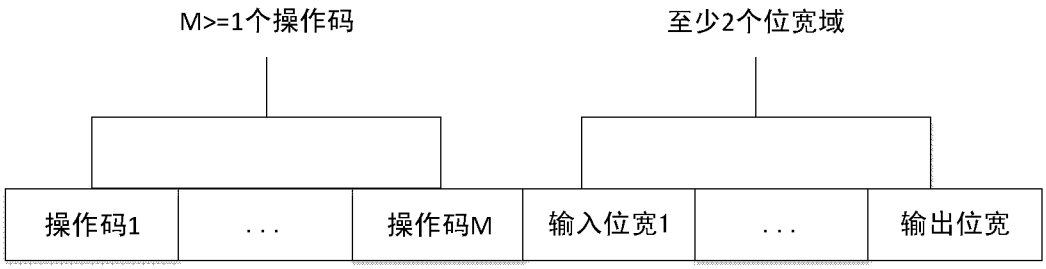

图 17

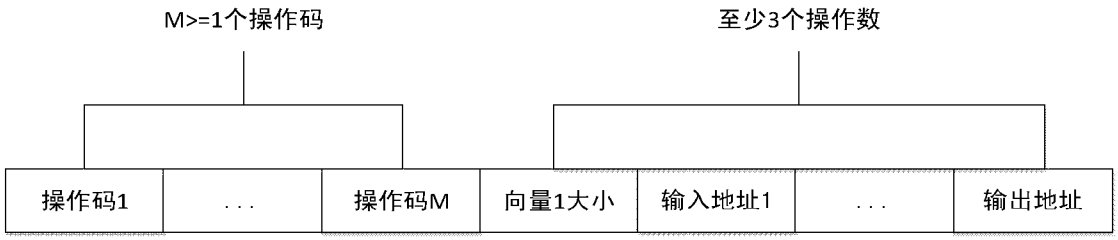

图 18

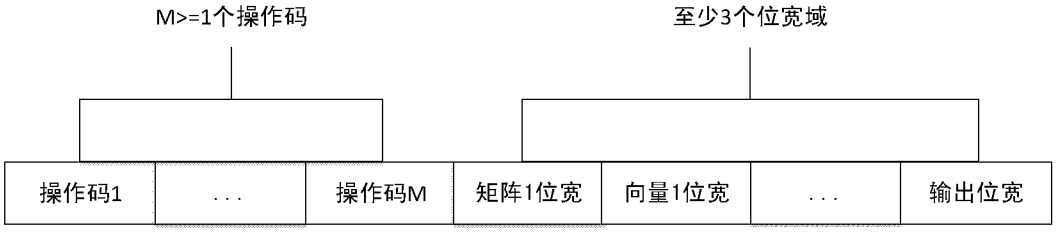
M>=1个操作码
至少3个位宽域
操作码1
. . .
操作码M
矩阵1位宽
向量1位宽
. . .
输出位宽

图 19

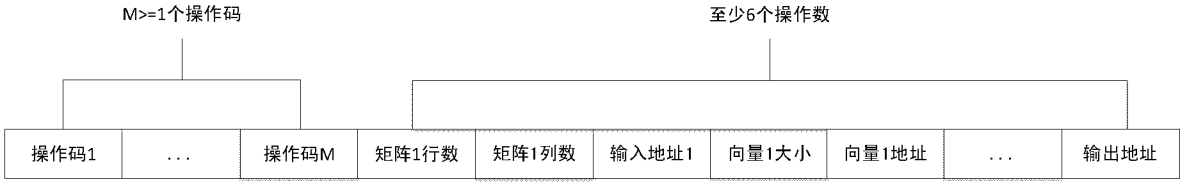
M>=1个操作码
至少6个操作数
操作码1
. . .
操作码M
矩阵1行数
矩阵1列数
输入地址1
向量1大小
向量1地址
. . .
输出地址

图 20

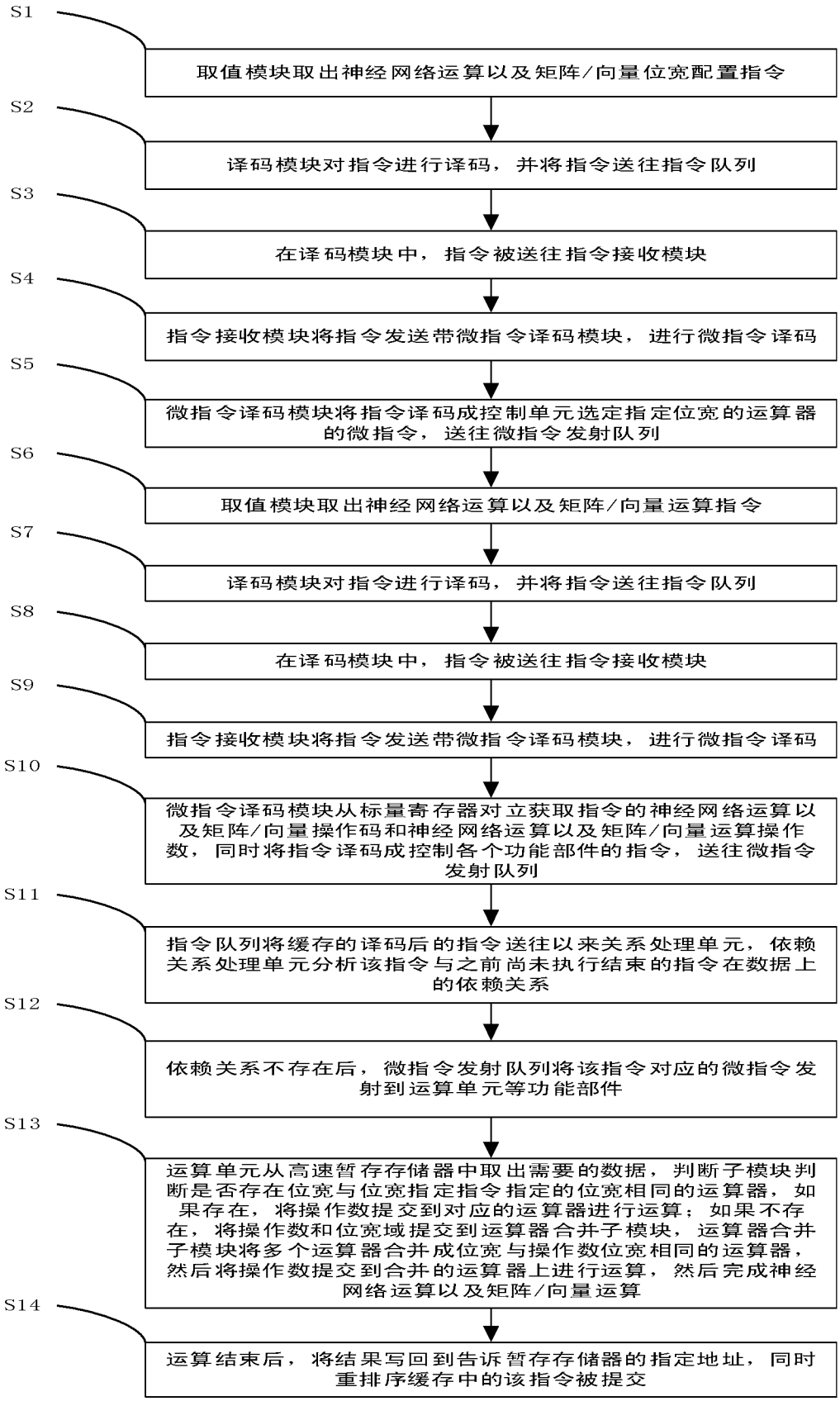
S1
取值模块取出神经网络运算以及矩阵/向量位宽配置指令
S2
译码模块对指令进行译码，并将指令送往指令队列
S3
在译码模块中，指令被送往指令接收模块
S4
指令接收模块将指令发送带微指令译码模块，进行微指令译码
S5
微指令译码模块将指令译码成控制单元选定指定位宽的运算器的微指令，送往微指令发射队列
S6
取值模块取出神经网络运算以及矩阵/向量运算指令
S7
译码模块对指令进行译码，并将指令送往指令队列
S8
在译码模块中，指令被送往指令接收模块
S9
指令接收模块将指令发送带微指令译码模块，进行微指令译码
S10
微指令译码模块从标量寄存器对立获取指令的神经网络运算以及矩阵/向量操作码和神经网络运算以及矩阵/向量运算操作数，同时将指令译码成控制各个功能部件的指令，送往微指令发射队列
S11
指令队列将缓存的译码后的指令送往以来关系处理单元，依赖关系处理单元分析该指令与之前尚未执行结束的指令在数据上的依赖关系
S12
依赖关系不存在后，微指令发射队列将该指令对应的微指令发射到运算单元等功能部件
S13
运算单元从高速暂存存储器中取出需要的数据，判断子模块判断是否存在位宽与位宽指定指令指定的位宽相同的运算器，如果存在，将操作数提交到对应的运算器进行运算；如果不存在，将操作数和位宽域提交到运算器合并子模块，运算器合并子模块将多个运算器合并成位宽与操作数位宽相同的运算器，然后将操作数提交到合并的运算器上进行运算，然后完成神经网络运算以及矩阵/向量运算
S14
运算结束后，将结果写回到告诉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 21

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/093159

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/063 (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)

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)

CNPAT, WPI, EPODOC, CNKI: bit width, corresponding, operation code, bit width domain, operation, same, combination, code

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 102012876 A (ZTE CORP.), 13 April 2011 (13.04.2011), description, paragraphs [0024]-[0048], and figures 1-5	1-14
A	CN 102238348 A (SHANGHAI HUAHONG INTEGRATED CIRCUIT CO., LTD.), 09 November 2011 (09.11.2011), the whole document	1-14
A	CN 103188487 A (LEADCORE TECHNOLOGY CO., LTD.), 03 July 2013 (03.07.2013), the whole document	1-14

☐ 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
20 September 2017 (20.09.2017)	**17 October 2017 (17.10.2017)**
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 **WANG, Xiaofei** Telephone No.: (86-10) **62413918**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/093159

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102012876 A	13 April 2011	None	
CN 102238348 A	09 November 2011	None	
CN 103188487 A	03 July 2013	None	

国际检索报告

国际申请号 PCT/CN2017/093159

A. 主题的分类

G06N 3/063(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)

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））

CNPAT，WPI，EPODOC，CNKI：位宽，运算，相应，相同，合并，操作码，位宽域，operation，same，combination，code

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 102012876 A (中兴通讯股份有限公司) 2011年 4月 13日 (2011-04-13) 说明书第[0024]-[0048]段及附图1-5	1-14
A	CN 102238348 A (上海华虹集成电路有限责任公司) 2011年 11月 9日 (2011-11-09) 全文	1-14
A	CN 103188487 A (联芯科技有限公司) 2013年 7月 3日 (2013-07-03) 全文	1-14

☐ 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
2017年 9月 20日	2017年 10月 17日
ISA/CN的名称和邮寄地址 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 传真号 (86-10)62019451	受权官员 王晓飞 电话号码 (86-10)62413918

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/093159

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 102012876 A	2011年 4月 13日	无	
CN 102238348 A	2011年 11月 9日	无	
CN 103188487 A	2013年 7月 3日	无	