# 一种基于脉动阵列的二值复数神经网络卷积计算加速装置与方法

## 技术领域

[0001] 本发明涉及人工智能领域，具体涉及一种基于脉动阵列的二值复数神经网络卷积计算加速装置与方法。

## 背景技术

[0002] 近年来，随着人工智能的飞速发展，深度神经网络（Deep Neural Networks，DNN）在诸多领域都有着广泛应用，如目标检测、人脸识别、图像视频超分处理、自动驾驶以及语音识别等。DNN算法具有存储密集和计算密集两大特性，因为其中包含了大量的权重数据和特征图数据之间的卷积操作，使得目前的DNN模型无法在边缘侧充分地训练或者推理。

[0003] 在进行卷积操作时，一般会将一张特征图的卷积操作转换成多个小型的矩阵乘法；如果采用传统的计算系统，即从内存中读取特征图矩阵数据和卷积核权重矩阵数据，完成计算后写回存储器，再继续下一批数据的计算，这种方法会导致数据复用率低，内存访问次数多、耗时久，计算效率低等问题。所以在人工智能芯片领域多采用脉动阵列的计算方式，以权重固定型脉动阵列完成矩阵乘法为例，其中A为输入特征图矩阵，B为卷积核数据矩阵，C为输出特征图矩阵，计算方法为：首先将权重矩阵数据在脉动阵列中自上而下传播，传播完成后将权重数据固定于每个脉动阵列单元中；再将输入特征图矩阵数据按行依次间隔一个周期输入脉动阵列，并自左向右传播；输入数据与权重数据进行乘法计算，乘法结果与上一级计算单元的部分和输出进行累加计算，所得新的部分和结果向下传播给下一级计算单元；最后一级计算单元将最终的部分和结果传入累加器中。

[0004] 二值复数神经网络有效地结合二值神经网络和复数神经网络，其中每层的输入特征图数据、权重数据和输出特征图数据都使用二值化的复数值来表示，即{1+i,1-i,-1+i,-1-i}中某一个，在进行卷积计算时，原先的全精度乘法计算也相应地映射为单比特的同或运算（xnor）外加一个计“1”（popcount）操作，在数据存储时，使用双比特替代32比特的单精度浮点数或16比特的半精度浮点，该网络具有二值神经网络的计算效率、低硬件开销、高能效比、高鲁棒性、高精度等多种优势。

[0005] 假设二值复数为激活输入，二值复数为权重，是全精度复数输出。二值复数的点积运算遵循复数计算规则，形式如下式，其中，，， ∈ {+1，-1}。一个二值复数点积运算需要结合4个二值点积和2个额外的实数加法。

## 发明内容

[0006] 为解决深度神经网络高硬件开销、计算效率低、无法在资源受限的边缘侧设备中部署等问题，实现低功耗、低成本、高能效比的边缘智能设备的目的，本发明采用如下的技术方案：

一种基于脉动阵列的二值复数神经网络卷积计算加速装置，装置中包括两个移位寄存器阵列、针对二值复数设计的基本运算单元PE组成的脉动阵列、相邻PE之间的寄存器和一个带累加器的SRAM存储单元。其中基本运算PE包含了权重数据输入接口（in\_d）、权重数据输出接口（out\_c）、特征图数据输入接口（in\_a）、特征图数据输出接口（out\_a）、上级基本运算单元PE部分和输入接口（in\_b）、部分和输出接口（out\_b）、控制信号输入接口、控制信号输出接口、存储权重数据的寄存器C1和C2、进位保留加法器（Carry Saved Adder，CSA）、用于二值复数popcount(xnor)计算的BC\_POPC(XNOR)模块、四个二选一数字选择器。

[0007] 进一步地，阐述加速装置各组件之间的电路连接关系，脉动阵列中的第一行基本运算单元的权重数据输入接口（in\_d）用于接收来自片上存储的卷积核权重矩阵；第一列基本运算单元特征图数据输入接口（in\_a）与左侧移位寄存器阵列连接，用于接收输入特征图矩阵；最后一行基本运算单元的部分和输出接口（out\_b）与底侧移位寄存器阵列连接，移位寄存器阵列将各列基本运算单元的输出结果进行打拍同步后送往带累加器的SRAM存储单元；每个基本运算单元的权重数据输出接口（out\_d）通过寄存器打一拍与同一列下一行的基本运算单元的权重数据输入接口（in\_d）连接，部分和接口（out\_b）与同一列下一行的基本运算单元的部分和输入接口（in\_b）直接连接，特征图数据输出接口（out\_a）和控制信号输出端口通过寄存器打一拍后与同一行下一列的基本运算单元的特征图数据输入接口（in\_a）和控制信号输入端口连接，故每个基本运算单元仅与相邻基本运算单元进行数据交互。

[0008] 进一步地，阐述加速装置内部的数据传输机制，加速装置内部通过握手机制来控制数据的传输，一组握手信号包括valid信号和ready信号，在数据的交互过程中，当valid和ready信号同时高有效时，数据在时钟上升沿传输。

[0009] 进一步地，阐述移位寄存器阵列内部的连接关系，左侧移位寄存器阵列模块的输入可分为握手信号接口、控制信号接口、输入数据信号接口；将输入信号分为16组分别送往脉动阵列的各行的第一列基本运算单元，每32比特的数据信号对应一组握手信号和1比特控制信号；其中第一组输入信号通过一个寄存器打一拍后与第一行第一列的基本运算单元特征图数据输入接口连接，第二组输入信号通过两个寄存器打一拍后与第二行第一列的基本运算单元特征图数据输入接口连接，依次类推，第16组输入信号通过16个寄存器打16拍后与第16行第一列的基本运算单元特征图数据输入接口连接，该移位寄存器阵列可以保证每组输入数据依次延迟一个周期后进入脉动阵列。底侧移位寄存器阵列模块结构与左侧移位寄存器阵列模块结构类似，输入可分为16组握手信号接口和输出数据信号，输入出为一个握手信号组和输出数据信号组，第一列最后一行的基本运算单元的部分和接口（out\_b）经过16个移位寄存器打16拍后与第一组握手信号和输出数据信号接口连接，第二列最后一行的基本运算单元的部分和接口（out\_b）经过15个移位寄存器打15拍后与第二组握手信号和输出数据信号接口连接，以此类推，该移位寄存器阵列可以保证每组输入数据打拍后同步进入带累加器的SRAM存储器中。

[0010] 进一步地，阐述基本运算单元内部的连接关系，在基本运算单元内部，特征图数据输入接口（in\_a）与特征图数据输出接口（out\_a）直接连接；权重数据输入接口（in\_d）经过双缓存结构通过数字选择器Mux3与权重数据输出接口（out\_c）连接，双缓存结构由数字选择器Mux0、Mux1和寄存器C1、C2组成；数字选择器Mux2的输入端分别与寄存器C1、C2连接，选择信号端与控制信号PROP连接，输出端与BC\_POPC(XNOR)模块连接；BC\_POPC(XNOR)模块的另一个输入数据端与特征图数据输入接口（in\_a）连接，权重数据和特征图输入数据进行popcount(xnor)计算后输出；进位保留加法器的两个输入端分别与BC\_POPC(XNOR)模块的输出端和上级基本运算单元PE部分和输入接口（in\_b）连接，输出端与部分和输出接口（out\_b），将上级基本运算单元部分和计算结果与BC\_POPC(XNOR)模块的输出结果累加后输出；

所述双缓冲结构主要用于缓冲输入数据便于后续的计算，且通过双缓冲可以在当前计算周期未完成的情况下将先前的计算结果从脉动阵列输出，实现了不间断计算，极大的提高了计算效率。针对两条数据通路，我们在电路的输入端和输出端各添加两个数字选择器Mux0、Mux1和Mux3，将控制信号PROP作为三个数字选择器的选择信号，当PROP为高电平时，该结构利用寄存器C1所在通路进行存储和传播权重数据，当PROP为低电平时，该结构利用寄存器C2所在通路进行存储和传播权重数据。

所述基本运算单元PE内部通过数字选择器Mux2以及控制信号PROP来选择寄存器C1、C2中哪一个寄存器的权重数据用于计算，当PROP为高电平时，选择寄存器C2中存储的权重数据送往BC\_POPC(XNOR)模块进行二值复数乘法计算，当PROP为低电平时，选择寄存器C1中存储的权重数据送往BC\_POPC(XNOR)模块进行二值复数乘法计算。

[0011] 进一步地，所述BC\_POPC(XNOR)模块中含有四个16比特的popc(xnor)计算单元、两个加法器，特征图数据输入接口、权重数据输入接口以及特征图数据输出接口。来自输入接口的特征图数据和权重数据均为32比特，进入BC\_POPC(XNOR)模块后首先分别分解为高16位与低16位两个部分，其中高16位代表多通道的二值复数的实部数据的集和，低16位代表多通道的二值复数的虚部数据的集和；计算过程遵循复数的乘法法则，特征图数据的实部分别送往第一和第三个popc(xnor)计算单元，特征图数据的虚部分别送往第二和第四个popc(xnor)计算单元，同理分配权重数据的实部和虚部。完成popcount(xnor)计算后，将第一个与第二个popc(xnor)计算单元的计算结果求差，得到特征图输出数据的实部，将第三个与第四个popc(xnor)计算单元的计算结果求和，得到特征图输出数据的虚部。其中popc(xnor)计算单元用于实现popcount(xnor)计算，即将两个16比特数据按位同或后得到中间计算结果，再将中间结果通过并行结构按位相加得到中间结果中“1”的个数，作为最终结果输出。

[0012] 进一步地，阐述含累加器的SRAM存储器结构，该SRAM存储中含有写端口、读端口、双端口RAM、一至二数据分发器Demux、累加器adder、三个流水线队列。其中写端口中分为写地址输入接口、写数据输入接口、握手信号输入接口，读端口分为读地址输入接口、读数据输出接口、握手信号输入接口。写端口信号通过流水线队列pipeline0处理后与数据分发器Demux的输入端连接；Demux的两个输出端分别与双端口RAM写数据端口和累加器adder的操作数端口（op2）连接；累加器adder的操作数端口（op1）与双端口RAM的读数据接口连接，累加器的输出端口（sum）与双端口RAM的写数据端口连接。读数据端口通过流水线队列pipeline1与双端口RAM的读端口连接，双端口RAM中该读地址对应的读数据通过流水线队列pipeline2输出。

所述流水线队列pipeline0中包含译码电路，会对写地址进行译码，通过地址信号的第30比特，判断该组信号中写数据是直接覆盖该地址的原始数据还是与该地址的原始数据累加；

所述累加器adder会将需要累加的写数据与双端口RAM中写地址对应的原始数据进行累加，并将结果从数据输出端口（sum）输出、写回双端口RAM中写地址对应存储单元。

[0012] 基于上述加速装置，提出一种基于脉动阵列的二值复数神经网络卷积计算方法，包括如下步骤：

S1，获取多个卷积核权重数据和输入特征图数据；

S2，权重数据输入接口（in\_d）中握手信号握手成功后，在控制信号PROP的控制下，将权重数据通过权重输入通道分发到脉动阵列的各列基本运算单元PE，通过双缓存结构广播并保留在寄存器中；

S3，特征图数据输入接口（in\_a）中握手信号握手成功后，特征图数据通过左侧移位寄存器阵列分组打拍，沿着行方向在脉动阵列中向后传播，同时基本运算单元对保存的权重数据和输入的特征图数据进行popcount(xnor)计算和累加操作，包括如下过程：

S31，左侧移位寄存器阵列对特征图数据进行分组并打拍，输入数据按行依次延迟一个周期进入脉动阵列；

S32，第一列基本运算单元PE在接收到特征图输入数据时，在控制信号PROP的控制下，选择双缓存接口中某一个寄存器C1或者C2保存的权重数据与特征图数据进行popcount(xnor)计算，同时特征图输入数据通过特征图输出接口（out\_a）向同一行的下一列基本运算单元PE传播；

S33，BC\_POPC(XNOR)模块将计算结果送往进位保留加法器CSA，与上级基本运算单元PE的部分和输入接口（in\_b）的输入进行累加，并将新的部分和传输给同一列的下级基本运算单元PE；

S34，最后一行的基本运算单元PE将最终的部分和结果，通过底侧移位寄存器阵列打拍同步后传输给带累加器的SRAM存储器进行保存；

S4，带累加器的SRAM存储器根据写端口的地址信号判断是否进行累加，完成卷积计算后，通过读该SRAM存储器即可得到卷积的最终结果。

[0012] 进一步地，所述S1中的每个权重数据和每个特征图数据的数据量的组数分别对应于脉动阵列的每行基本运算单元PE的个数和每列基本运算单元PE的个数。每组数据为32比特，分别表示16个通道的二值复数的集合，其中高16位表示各通道的实部数据的集合，低16位表示各通道的虚部数据的集合。

[0013] 进一步地，所述S2中权重数据输入接口（in\_d）每次最多可输入16个卷积核的权重数据矩阵，如果卷积核个数大于16，则需要分批次进行卷积计算，并在带累加器的SRAM存储器中将部分和结果累加。

[0014] 进一步地，所述S3中二值复数的popcount(xnor)计算作用等同于整数或全精度浮点数的乘法计算，不过BC\_POPC(XNOR)模块仅使用了一系列的同或门和加法器，不仅节约了硬件开销，还提高了计算效率。

[0015] 本发明的优势和有益效果在于：

本发明的一种基于脉动阵列的二值复数神经网络卷积计算加速装置与方法，采用两比特的二值复数代替32比特的单精度浮点数或16比特的半精度浮点数，采用硬件开销小、计算简单的popcount(xnor)计算代替硬件开销大、计算复杂的浮点数乘法计算，并将多通道的二值复数合并为单通道进行计算，不仅节省了存储资源、减少了内存访问次数，还提高了计算效率；另外通过脉动阵列可以多次复用权重数据和输入特征图数据，一次内存访问，就可以完成多个卷积核的卷积计算，提高了加速装置的吞吐率和计算效率。综上所述在人工智能芯片领域采用该加速装置和计算方法，可以实现低功耗、低成本、高能效的边缘智能设备，解决了在资源受限的边缘侧设备中部署深度神经网络的难题。

## 附图说明

[0016] 图1为本发明的整体架构示意图。

[0017] 图2为本发明的脉动阵列的结构示意图。

[0018] 图3为本发明的左侧移位寄存器阵列的结构示意图。

[0019] 图4为本发明的基本运算单元PE的结构示意图。

[0020] 图5为本发明的BC\_POPC(XNOR)模块的结构示意图。

[0021] 图6为本发明的底侧移位寄存器阵列的机构示意图。

[0022] 图7为本发明的带累加器的SRAM存储器的结构示意图。

[0023] 图8为本发明方法中数据处理的流程示意图。

[0024] 图9为本发明方法中多通道二值复数合并为单通道的示意图。

[0025] 图10为本发明方法中卷积计算的示意图。

## 具体实施方法

[0024] 以下结合附图对本发明的具体实施方式进行详细说明。应当理解的是，此处所描述的具体实施方式仅用于说明和解释本发明，并不用于限制本发明。、

[0025] 为解决深度神经网络硬件开销高、计算效率低、无法在资源受限的边缘侧设备中部署等问题，本申请通过以下实施例公开了一种基于脉动阵列的二值复数神经网络卷积计算加速装置与方法，如图1所示，包括：

两个移位寄存器阵列、针对二值复数设计的基本运算单元PE组成的脉动阵列、相邻PE之间的寄存器和一个带累加器的SRAM存储单元，其中脉动阵列由基本运算单元PE，若输入特征图数据或卷积核权重数据大于16，则通过多次计算求和即可得到最终计算结果。

[0026] 为了解决深度神经网络中卷积计算存储密集和计算密集的问题，本文设计的加速装置采用了脉动阵列结构，如图2所示，每个基本运算单元PE仅与相邻的PE通过寄存器连接，也仅与相邻的PE进行数据的传递。

[0027] 权重输入通道将相应的权重数据输入到脉动阵列的第一行的PE基本单元，随后权重自上而下传输，将权重数据输入到每一个PE基本运算单元中的寄存器中并保存。

[0028] 数据输入通道将相应的输入特征图数据经过左侧移位寄存器阵列分组打拍后，依次进入脉动阵列的第一列的基本运算单元PE中，随后特征图数据从左向右传递，特征图数据陆续通过所在行所有列的PE。

[0029] 上述左侧移位寄存器阵列的结构如图3所示，寄存器阵列共有16行，每行寄存器的个数依次递增，打拍延迟周期数递增，即第1行有1个寄存器，第2行有2个寄存器，…，第16行有16个寄存器。通过该阵列可以保证同一张输入特征图的各组输入数据可以依次延迟一个周期进入脉动阵列的各行PE中，使得可以在同一列PE中完成一张卷积核的卷积计算。

[0029] 采用脉动阵列结构可以将输入数据多次复用，减少了内存访问次数，实现了高效的运算吞吐率，提高了计算效率。

[0030] 基本运算单元PE的结构如图4所示，在控制信号PROP的作用下，选择寄存器C1或者C2中保存的权重数据与输入的特征图数据进行popcount(xnor)计算和累加计算。每个基本运算单元包括一个权重数据输入接口（in\_d）、一个权重数据输出接口（out\_c）、一个特征图数据输入接口（in\_a）、一个特征图数据输出接口（out\_a）、一个上级基本运算单元PE部分和输入接口（in\_b）、一个部分和输出接口（out\_b）、一个控制信号输入接口、一个控制信号输出接口、两个存储权重数据的寄存器C1和C2、一个进位保留加法器（Carry Saved Adder，CSA）、一个用于二值复数popcount(xnor)计算的BC\_POPC(XNOR)模块、四个二选一数字选择器。其中所有接口除数据通道外，还包括一对握手信号通道，用于控制数据的发送与接收。

[0031] 基本运算单元PE中的两个存储权重数据的寄存器C1和C2组成了双缓存结构，该结构可以在保证在计算的同时进行权重数据的传递与存储。在PROP信号为低电平，权重数据从权重数据输入接口（in\_d）进入PE，通过数字选择器Mux0、Mux1、Mux3的选择后，经过寄存器C2并从权重数据输出接口（out\_c）输出给同一列的下一级PE，权重数据传递过程结束后，最后经过寄存器C2的权重数据被保存；之后输入特征图数据与高电平的控制信号PROP从特征图数据输入接口（in\_a）进入PE，特征图输入数据一方面通过特征图数据输出接口（out\_a）与PE间寄存器直接输出到同一行的下一级PE，另一方面在数字选择器Mux2的选择下与寄存器C2中保存的权重数据进行popcount(xnor)计算；因为此时控制信号为高电平，所以新的权重数据可以通过双缓存结构在寄存器C1中传递并存储，等待下一个输入特征图数据与低电平的控制信号PROP从特征图数据输入接口（in\_a）进入PE，即可直接进行下一个输入特征图的卷积计算；可见利用双缓存结构可以实现了不间断计算，极大的提高了计算效率。

[0032] 本次设计的加速装置采用两比特的二值复数来表示32比特的单精度浮点数或16比特的半精度浮点数，所以原先复杂的浮点数乘法计算可以使用计算简单的popcount(xnor)计算来代替，并且由于二值数的各通道之间的独立性，故可将多通道的二值复数合并为单通道进行计算，进行popcount(xnor)计算的BC\_POPC(XNOR)模块结构如图5所示，包括一个特征图数据输入接口、一个权重数据接口、一个计算结果输出接口、四个16比特的popcount(xnor)计算单元、两个加法器。该模块输入接口与输出接口位宽均为32比特，可在一个周期内完成16个通道的二值复数的popcount(xnor)计算，且遵循复数乘法法则，将相应的部分求差或求和得到复数结果的实部与虚部。利用该计算模块不仅节省了存储资源、减少了内存访问次数，还提高了计算效率。

[0033] 完成上述popcount(xnor)计算后，将复数结果输出到累加器中，并与基本运算单元PE的上级基本运算单元PE部分和输入接口（in\_b）中的部分和数据进行复数加法计算，即实部与实部相加、虚部与虚部相加，得到该级PE的部分和数据，并从部分和输出接口（out\_b）输出给同一列的下级PE中。部分和数据在同一列的基本运算单元PE中不断累加，最终累加结果通过最后一行PE的部分和输出接口（out\_b）输出到底侧移位寄存器阵列。

[0034] 上述底侧移位寄存器阵列结构如图6所示，其结构与左侧移位寄存器阵列类似，共有16个寄存器队列，与脉动阵列中最后一行基本运算单元PE的部分和输出接口（out\_b）一一对应，分别对各列PE的最终部分和数据及其握手信号进行缓存打拍。该阵列的第一列包括16个寄存器，第二列包括15个寄存器，依次递减，第一列包括1个寄存器。之所以采用该结构，是因为脉动阵列中第一列的部分和最先完成计算，第二列相较第一列晚一个周期，依次类推，故通过该结构将同一批的卷积计算结果打拍同步后送往SRAM存储器中。

[0035] 与传统的SRAM存储器不同，本发明中的SRAM存储器是一种带累加器的SRAM存储器，结构如图7所示，包括一个写端口、一个读端口、一个双端口RAM、一个一至二数据分发器Demux、一系列累加器adder、三个流水线队列。其中写端口中分为写地址输入接口、写数据输入接口、握手信号输入接口，读端口分为读地址输入接口、读数据输出接口、握手信号输入接口。双端口RAM每次读取和写入的数据宽度等于比特，该宽度对应脉动阵列所有行或者所有列输入或输出的数据位宽之和。

[0036] 当需要写数据时，写地址、写数据与握手信号同时通过写端口进入流水线队列pipeline0，队列中会对写地址的第30位比特进行译码，如果为0，则直接向双端口RAM中写地址对应的存储单元写数据，如果为1，则需要先读取双端口RAM中写地址对应的存储单元中的保存数据，并与写数据同步送入一系列加法器中进行累加，最后将累加结果写回该存储单元。

[0037] 如图8所示，基于脉动阵列的二值复数神经网络卷积计算加速方法的工作流程，包括：

步骤S1，获取多通道合并后的权重核矩阵以及特征图数据矩阵，其中卷积核矩阵包括多个卷积核的多个通道的权重数据，特征图数据矩阵包括一个或多个特征图的多个通道的输入数据。

[0038] 步骤S2，在权重数据接口（in\_d）中握手信号握手成功后，在控制信号PROP的控制下，将权重数据通过权重输入通道分发到脉动阵列的各列基本运算单元PE，通过双缓存结构广播并保留在寄存器中；

[0039] 步骤S3，特征图数据输入接口（in\_a）中握手信号握手成功后，特征图数据通过左侧移位寄存器阵列分组打拍，沿着行方向在脉动阵列中向后传播，同时基本运算单元对保存的权重数据和输入的特征图数据进行popcount(xnor)计算和累加操作。

[0040] 步骤S4，将S3中的计算结果通过底侧移位寄存器阵列后存入带累加器的SRAM存储器，且根据写端口的地址信号判断是否进行累加，完成卷积计算后，通过读该SRAM存储器即可得到卷积的最终结果。

[0041] 以下结合具体实例，对本申请的一种基于脉动阵列的二值复数神经网络卷积计算加速方法进行说明。

[0042] 首先设定原始二值复数特征图矩阵的通道数为32，原始二值复数卷积核权重矩阵的通道数也为32，其中前16个通道为实部，后16个通道为虚部，特征图矩阵的维度为，权重矩阵的维度为；将32个通道的每个像素点对应的1比特数按照通道顺序合并为1个通道的32比特数，过程如图9所示，例如将像素点A00在32个通道中的二值化数合并在一起得到F00，依此类推得到F01，F02，直至F33。同理得到多个权重矩阵数据，本次实例仅展示W、G两个权重矩阵。

[0042] 将卷积核权重矩阵按列存储在脉动阵列中，如图10所示，权重矩阵W数据依次存储在第一列基本运算单元PE中，权重矩阵G数据依次存储在第二列基本运算单元PE中，等等。合并后的特征图矩阵进入左侧移位寄存器阵列，对每个元素打拍后进入脉动阵列，例如F00进入第一行第一列PE后一个周期F02才进入第二行第一列PE，F01进入第二行第一列PE后一个周期F03才进入第三行第一列PE。因为特征图数据在脉动阵列中沿着行方向不断向前传播，进行popcount(xnor)和累加计算并将所得部分和沿着列方向不断向下级PE传播，所以在经过16个周期，第一列PE完成了第一个特征图矩阵完成对权重矩阵W的卷积计算；再经过一个周期，第一列PE完成了第二个特征图矩阵完成对权重矩阵W的卷积计算，同时第二列PE完成了第一个特征图矩阵完成对权重矩阵G的卷积计算。

[0042] 按照上述过程，不断计算，因为第一列PE的卷积计算比最后一列PE的卷积计算提前16个周期完成，为了让它们存储在SRAM存储器的同一行中，所以需要将最后一行基本运算单元PE的计算结果通过底侧移位寄存器阵列，分别打拍延迟，最后同步输入到SRAM存储器中。

[0042] 伴随权重矩阵和特征图矩阵的获取，同时还有目标存储地址的获取，该目标地址即为卷积计算结果写入带累加器的SRAM存储器的存储地址，该地址存储在队列中，等待卷积计算结果及其有效信号通过底侧移位寄存器队列，后同步送往存储单元中。同时带累加器的SRAM存储器会对目标地址进行译码，判断是否与目标地址的原保存数据进行累加。

[0041] 以上实例说明本发明的一种应用方案，只是用于帮助理解本发明的原理和设计思想，而非对其限制；本领域的普通技术人员应当理解：依据本发明的设计思想，在具体实施方式和应用范围上都可以做成适当修改或替换，而这些修改或者替换，都不能使其技术或方案的本质脱离本发明的原理和设计思想。

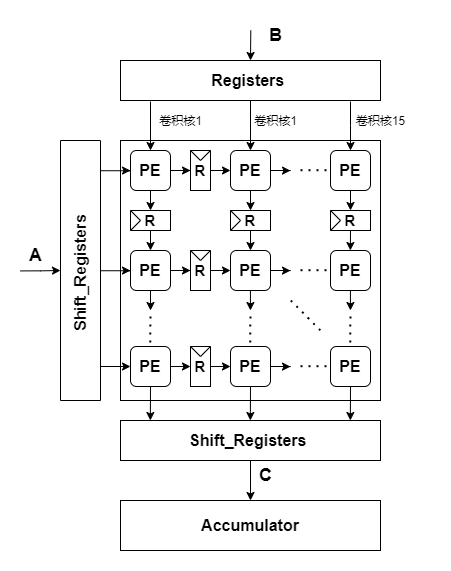


图1

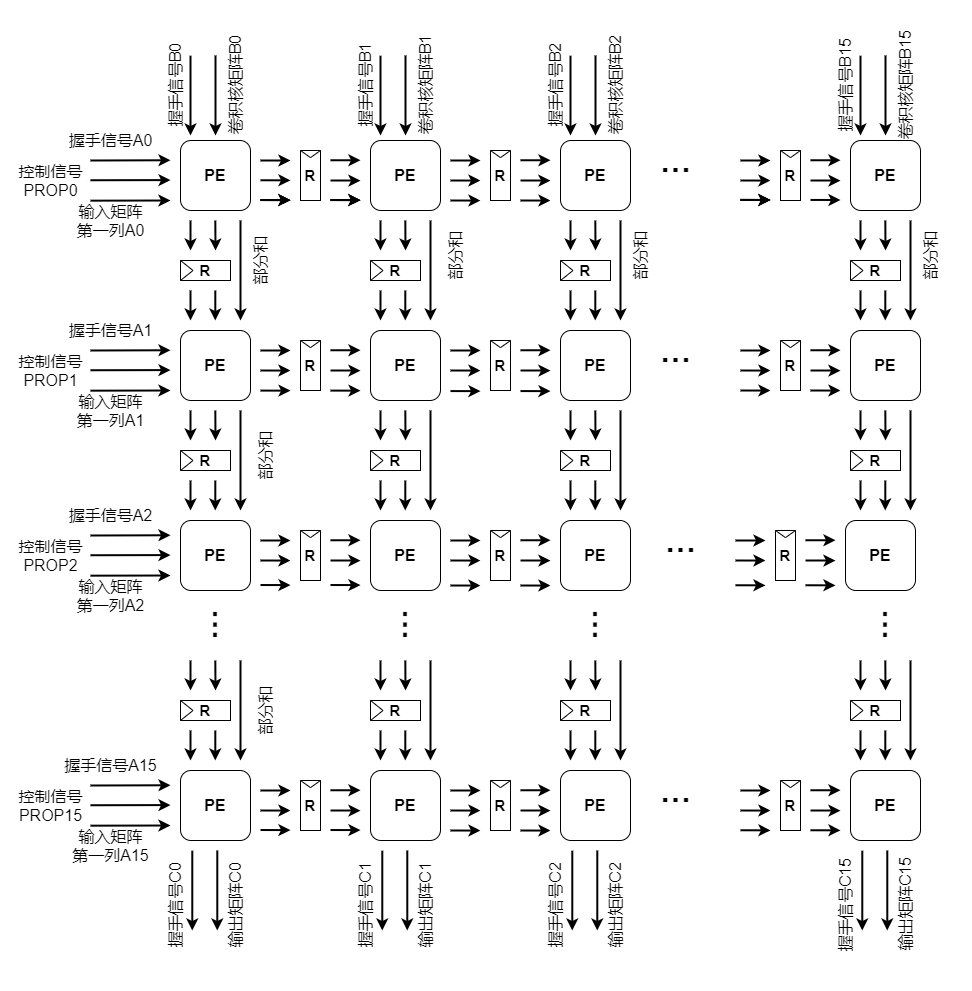


图2

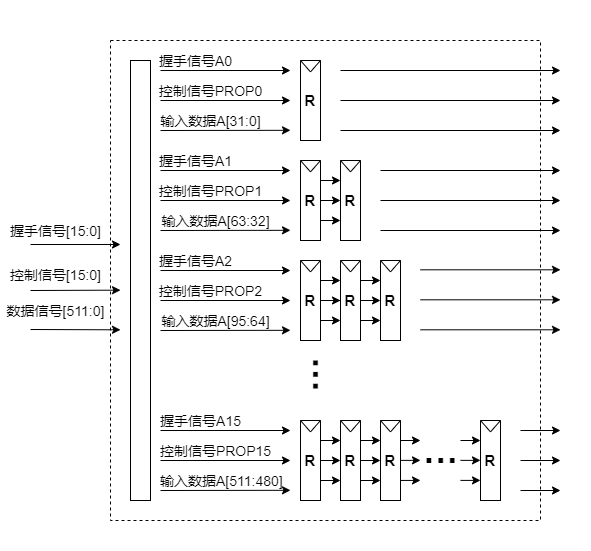


图3

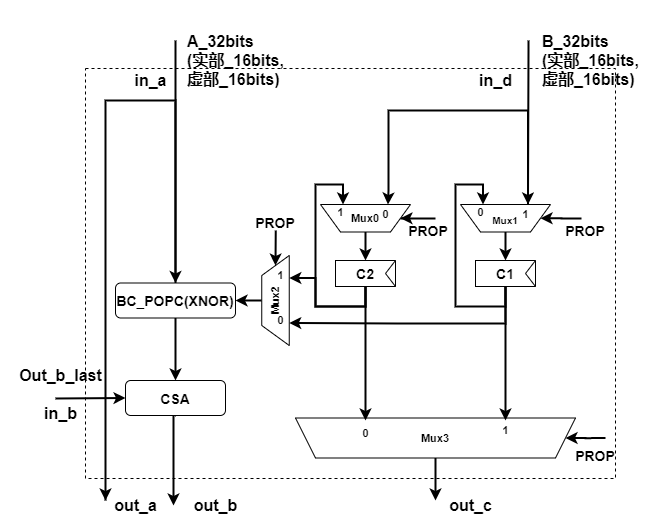


图4

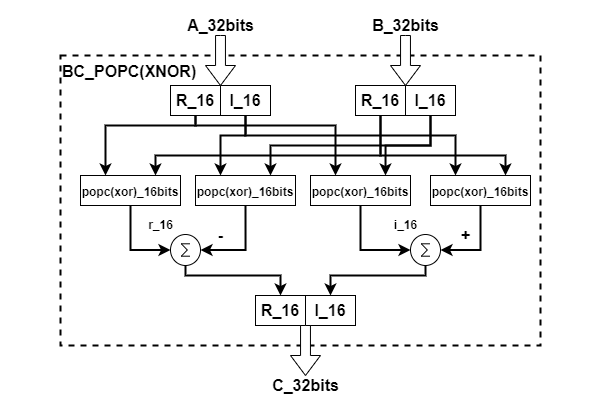


图5

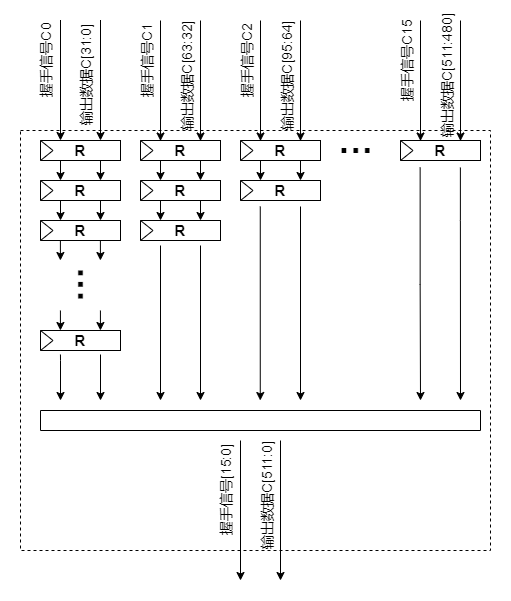


图6

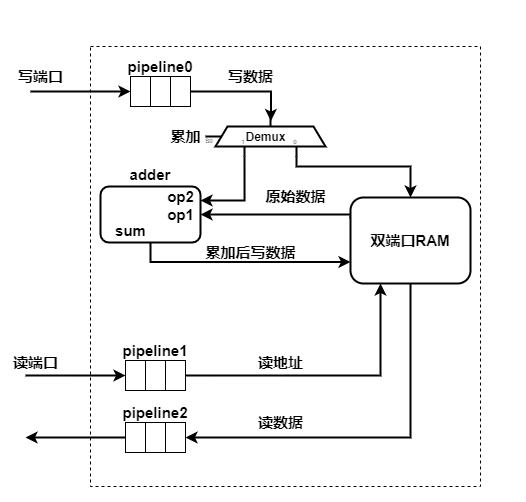


图7

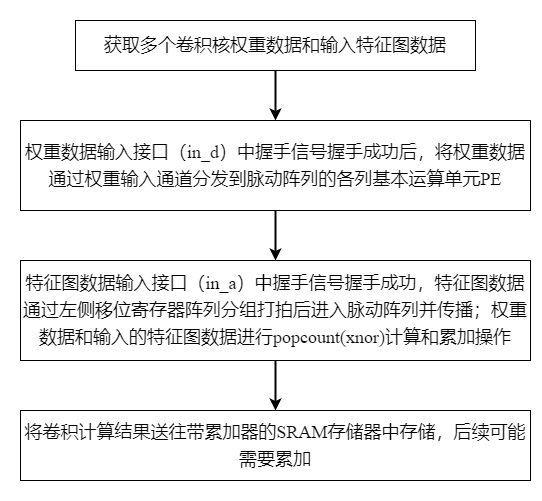


图8

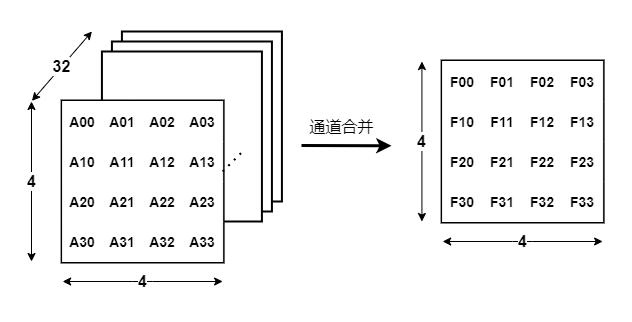


图9

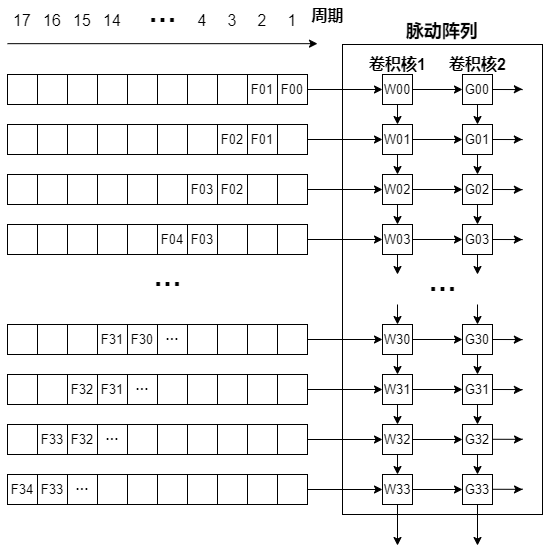


图10