# 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

本发明涉及人工智能芯片领域，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

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。

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。

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。

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

由于全精度深度神经网络中浮点数的乘加计算硬件开销大、计算效率低，二值神经网络精度受损严重等问题，此类神经网络均不便于在边缘智能设备中部署。二值复数神经网络虽然硬件开销小、精度高，但是在硬件部署过程中也在诸多问题，如存储资源和计算资源利用率不足、卷积核尺寸与脉动阵列行尺寸不相同时需要软件处理权重矩阵数据等。为解决上述问题，并实现低功耗、高能效比的边缘智能设备，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系统装置，系统装置中包括一个脉动阵列控制器、三个移位寄存器阵列、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累加器。脉动阵列尺寸为，即每行包括16个基本运算单元PE，每列包括16个基本运算单元。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。

所述脉动阵列控制器与左侧移位寄存器阵列和顶侧移位寄存器阵列连接，分别传输特征图输入矩阵数据和卷积核权重矩阵数据。

所述脉动阵列控制器可以动态即时处理不同尺寸的卷积核的权重矩阵数据，其中卷积核尺寸可以为、、、。

所述顶侧移位寄存器阵列与脉动阵列的第一行基本运算单元连接；左侧移位寄存器阵列与脉动阵列的第一列基本运算单元连接；底侧移位寄存器阵列与脉动阵列的最后一行基本运算单元和累加器连接；

所述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。

所述加速系统装置内部的数据传输通过握手机制来控制，每组数据信号都伴随一组握手信号，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。

所述移位寄存器阵列包括16组不同尺寸的移位寄存器，每组移位寄存器延迟数呈递增或递减趋势。

所述基本运算单元内部，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BC\_POPC(XNOR)模块可在一个周期内并行完成二值复数卷积层16个通道的“乘法”计算，包括四个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；计算过程遵循复数的乘法法则，特征图数据的高16位送往第一和第三个popc(xnor)计算单元，特征图数据的低16位送往第二和第四个popc(xnor)计算单元，权重数据的的高16位送往第一和第四个popc(xnor)计算单元，权重数据的的低16位送往第二和第三个popc(xnor)计算单元。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，得到输出数据的高16位，将第三个与第四个popc(xnor)计算单元的计算结果求和，得到输出数据的低16位；将输出数据的高16位与低16位拼接后送往进位保留加法器CSA。其中popc(xnor)计算单元用于实现popcount(xnor)计算，将两个16比特数据按位同或并逐位相加得到结果中“1”的个数，再将“1”的个数乘以2并减去通道数，即得到计算结果。

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，脉动阵列控制器获取多个卷积核权重数据和输入特征图数据；

S2，基本运算单元PE纵向传播并保存权重矩阵数据，包括如下步骤：

S21，权重数据输入接口中握手信号握手成功后，脉动阵列控制器动态即时处理权重矩阵数据，并将矩阵数据送往顶侧移动寄存器阵列进行分组打拍；

S22，权重矩阵数据进入脉动阵列，在控制信号PROP的控制下，基本运算单元PE通过双缓存结构纵向传播并将保留权重数据在寄存器中；

S3，基本运算单元PE横向传播输入特征图数据，并对输入的特征图矩阵数据和保存的权重矩阵数据进行popcount(xnor)和累加操作，包括如下步骤：

S31，特征图数据输入接口中握手信号握手成功后，脉动阵列控制器将特征图数据送往左侧移位寄存器阵列进行分组打拍；

S32，特征图输入数据按行依次进入脉动阵列，并沿着行方向在脉动阵列中传播；

S33，特征图数据横向传播的同时，在控制信号PROP的控制下，基本运算单元PE将寄存器C1或者C2保存的权重数据与特征图数据送往BC\_POPC(XNOR)模块进行popcount(xnor)计算，之后将计算结果送往进位保留加法器CSA，与部分和输入接口（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；

S4，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至累加器进行保存；

所述S1中的每个权重矩阵数据和每个输入特征图矩阵数据的元素个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。每个元素为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。

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16个通道的权重数据矩阵，如果该卷积层的卷积核个数或卷积核通道数大于16，则需要分批次进行卷积计算，并在累加器中将部分和结果累加。

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。

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系统装置与方法，在基本运算单元PE中采用双缓存结构，实现了不间断计算；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，并创新地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，还提高了计算资源的利用率；另外本发明中的脉动阵列控制器可以动态即时处理不同尺寸卷积核的权重矩阵数据，避免了软件处理权重数据的时间消耗，提高了系统装置的计算效率。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、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

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例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发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图1为本发明的整体架构示意图。

图2为本发明的脉动阵列的结构示意图。

图3为本发明的基本运算单元PE的结构示意图。

图4为本发明的BC\_POPC(XNOR)模块的结构示意图。

图5为本发明中脉动阵列加速二值复数神经网络卷积计算的示意图。

图6为本发明中卷积核尺寸等于情况下，多通道二值复数合并为单通道的示意图。

图7为本发明中卷积核尺寸小于情况下，多通道二值复数合并为单通道的示意图。

图8为本发明方法中卷积计算的流程示意图。

## 具体实施方法

下面将结合本发明实施例中的附图，对本发明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发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

本发明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系统装置与方法，为使本发明的目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。本发明的加速系统装置整体架构如图1所示，包括：

一个脉动阵列控制器、三个移位寄存器阵列、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之间的寄存器和累加器，其中脉动阵列由基本运算单元PE组成。若输入特征图或卷积核通道数大于16，则需要通过多次计算求和即可得到最终计算结果。

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。采用该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。

基本运算单元PE的结构如图3所示，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。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

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。

所述卷积计算加速系统装置中各模块之间的数据交互、基本运算单元PE之间的数据交互过程都受到握手机制的控制。

本发明中的加速系统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。注意到如果每个存储单元仅用来存储两比特的二值复数，每个基本运算单元每次仅进行单通道二值复数的popcount(xnor)计算，势必会造成存储资源和计算资源的利用率不足等问题。所以基于二值复数神经网络各通道之间的数据独立性，本发明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4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popc(xnor)计算单元、两个加法器。其中popc(xnor)计算单元将两个16比特数据按位同或并逐位相加得到结果中“1”的个数，再将“1”的个数乘以2并减去通道数，即得到popcount(xnor)计算结果。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提高了存储资源利用率、减少了内存访问次数，还提高了计算效率。

本发明中的脉动阵列控制器不仅可以控制所述权重矩阵数据和特征图输入矩阵数据的数据流，还具有动态即时处理权重矩阵数据的功能，主要用于应对二值复数神经网络中卷积核尺寸不固定的应用场景，例如卷积核尺寸可能为、、、、等尺寸中的一种。若卷积核权重矩阵尺寸小于，脉动阵列控制器将动态即时处理权重矩阵数据；若卷积核权重矩阵尺寸等于，脉动阵列控制器不需要处理权重矩阵数据；若卷积核权重矩阵尺寸大于，则需要将权重矩阵拆分成一个或多个尺寸为的权重矩阵和零个或一个尺寸小于的权重矩阵。如图5所示，当脉动阵列控制器需要传输卷积核权重矩阵尺寸（或特征图输入矩阵尺寸）等于时，权重矩阵（或特征图输入矩阵尺寸）可以一一对应分布在脉动阵列的各行，例如图5中的F[0,1]~E[15,1]、F[0,0]~F[15,0]、W[0,0]~W[15,0]、W[0,1]~W[15,1]等。而当脉动阵列控制器需要传输卷积核权重矩阵尺寸（或特征图输入矩阵尺寸）小于时，权重矩阵（或特征图输入矩阵尺寸）不能完全分布满脉动阵列的各行，例如卷积核权重矩阵尺寸等于时，如果不经过动态即时处理，仅W[0,0]~W[8,0]、W[0,1]~W[8,1]等为有效数据，W[9,0]~W[15,0]、W[9,1]~W[15,1]等为无效数据，默认为32比特的全零数据。当权重数据和输入数据均为32比特的全零数据时，在进行popcount(xnor)计算后的得到的中间结果为，而预期值应该为，这就会导致二值复数卷积计算的最终结果会出现问题。

所述脉动阵列控制器的动态即时处理操作包括：判断卷积核权重尺寸、即时处理并传输无效数据、传输有效数据。例如卷积核权重矩阵尺寸等于时，经过动态即时处理，W[9,0]~W[15,0]、W[9,1]~W[15,1]等不再为全零数据，而是0x00FF\_00FF，F[9,1]~F[15,1]、F[9,0]~F[15,0]依旧为全零数据，此时四个popc(xnor)计算单元中计算结果均为0，不会对有效数据的卷积计算结果造成影响。该动态即时处理操作只需要一个计数器和比较器即可完成，且不额外消耗卷积计算运行时间。

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系统中的多通道二值复数合并为单通道方法进行说明。

作为实施例1，如图6所示，以卷积核权重矩阵为例，其原始二值复数卷积核的个数为16，通道数为16，卷积核的尺寸为，首先将每个卷积核的权重矩阵分解为16个通道的二值复数实部和16个通道的二值复数虚部，并按列排布，如B[0,0]，B[1,0]，…，B[15,0]为第一个卷积核的权重矩阵；将实部或虚部16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16比特数，并将两个 16比特的数合并为一个32比特的数，其中实部合并得到的16比特数占高16位。如图6中，将像素点B[0,0]在32个通道中的二值化数合并在一起得到W[0,0]，依此类推得到W[1,0]，W[2,0]，直至W[15,0]，同理得到其余卷积核的权重矩阵。

作为实施例2，如图7所示，以卷积核权重矩阵为例，其原始二值复数卷积核的个数为16，通道数为16，卷积核的尺寸为，合并原理与实施例1相同，最终的合并结果是尺寸为的矩阵。该权重矩阵通过脉动阵列控制器送往脉动阵列的同时需要进行处理，并补充为的权重矩阵，补充部分的数据均为0x00FF\_00FF。

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，脉动阵列控制器获取多通道合并后的16个卷积核权重数据和输入特征图数据。

步骤S2，脉动阵列控制器动态即时处理权重矩阵数据，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，将权重数据通过顶侧移位寄存器阵列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；

步骤S3，脉动阵列控制器传输特征图输入数据，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)和累加操作。

步骤S4，将S3中的卷积计算结果通过底侧移位寄存器阵列后存入累加器。

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。

作为实施例3，如果5所示，原始二值复数特征图矩阵和卷积核权重矩阵的尺寸均为，通道数为16，卷积核个数为16，卷积核尺寸为。首先将多通道二值复数的特征图矩阵和权重矩阵分别合并为单通道的矩阵数据，矩阵尺寸压缩为，该合并过程可在主处理器完成。其次将卷积核权重矩阵按列存储在脉动阵列中，由于卷积核尺寸等于，故不需要进行动态即时处理。接着将特征图矩阵通过左侧移位寄存器阵列送入脉动阵列，特征图数据在脉动阵列中沿着行方向不断向前传播，同时与基本运算单元中存储的权重数据进行popcount(xnor)和累加操作，并将所得部分和沿着列方向向下级PE传播。经过16个周期，第一列PE完成了第一个特征图矩阵完成对第一个权重矩阵的卷积计算，再经过16个周期，第15列PE完成了第一个特征图矩阵完成对第15个权重矩阵的卷积计算。最后将第一个特征图矩阵对所有卷积核的卷积计算结果经过底侧移位寄存器阵列打拍同步后，送入累加器暂存。

作为实施例4，原始二值复数特征图矩阵和卷积核权重矩阵的尺寸均为，通道数为16，卷积核个数为16，卷积核尺寸为。首先将多通道二值复数的特征图矩阵和权重矩阵分别合并为单通道的矩阵数据，矩阵尺寸压缩为，该合并过程可在主处理器完成。其次将卷积核权重矩阵按列存储在脉动阵列中，由于卷积核尺寸小于，故需要进行动态即时处理。经过脉动阵列控制器对卷积核权重矩阵进行动态即时处理后，后续的卷积计算过程与实施例3相同。

以上实施例仅用于说明本发明的一种应用方案，用于帮助理解本发明的原理和设计思想，而非对本发明作任何形式上的限制；但是,凡熟悉本专业的技术人员在不脱离本发明技术方案范围内，利用以上所揭示的技术内容而做出的些许更动、修饰与演变的等同变化，均为本发明的等效实施例；同时,凡依据本发明的实质技术对以上实施例所作的任何等同变化的更动、修饰与演变等，均仍属于本发明的技术方案的保护范围之内。

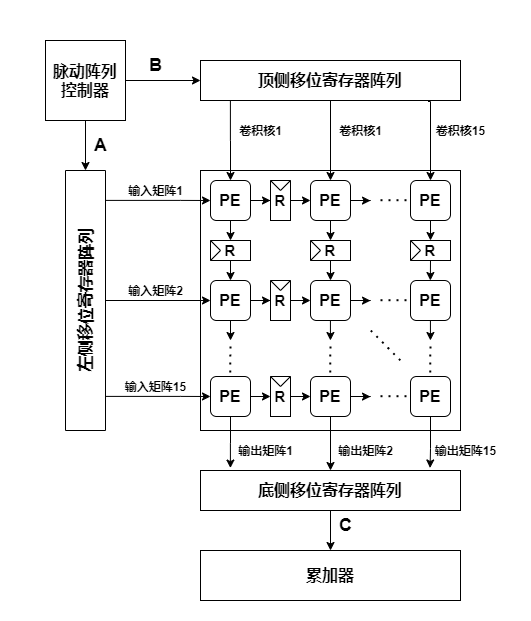


图1

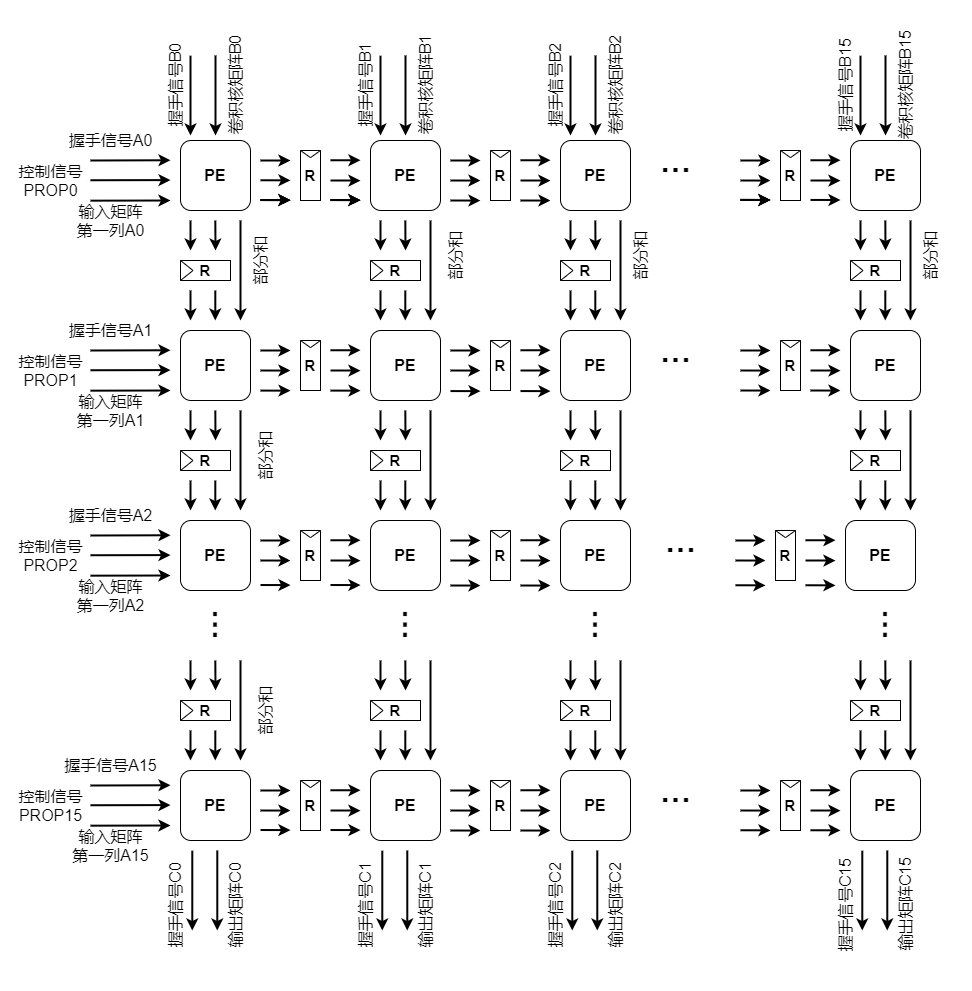


图2

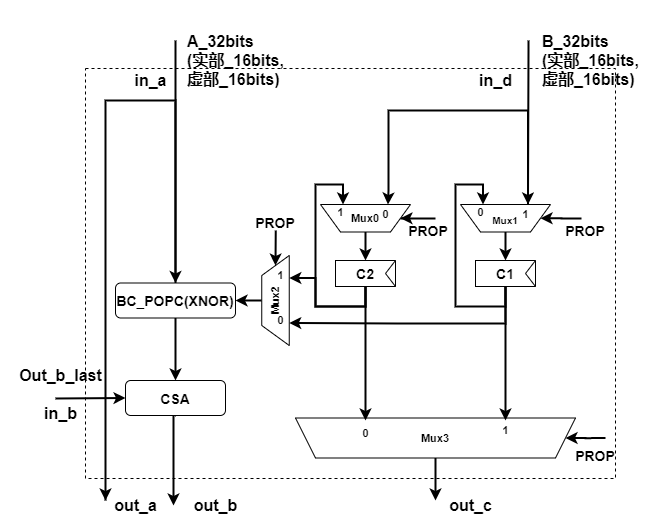


图3

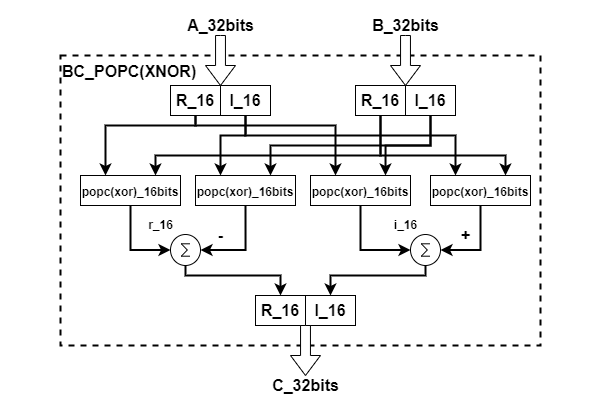


图4

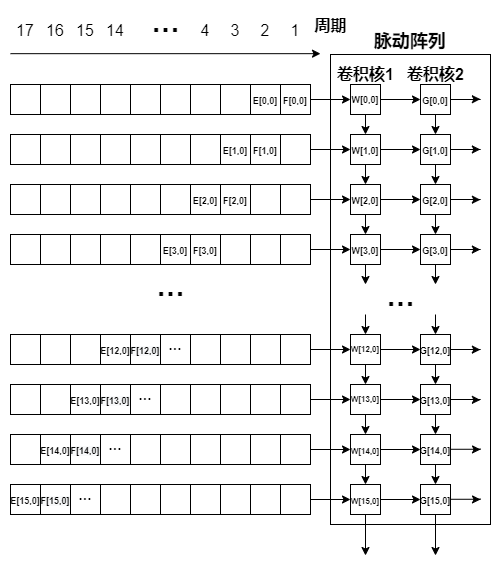


图5

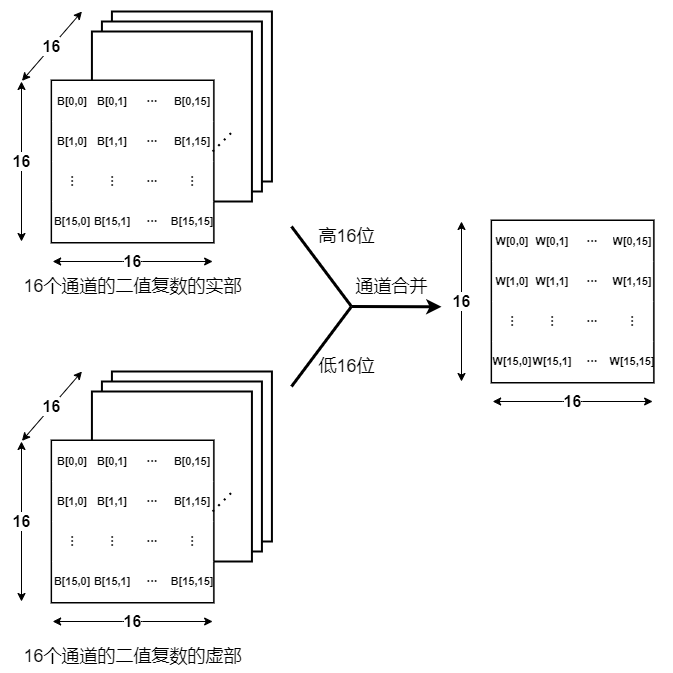


图6

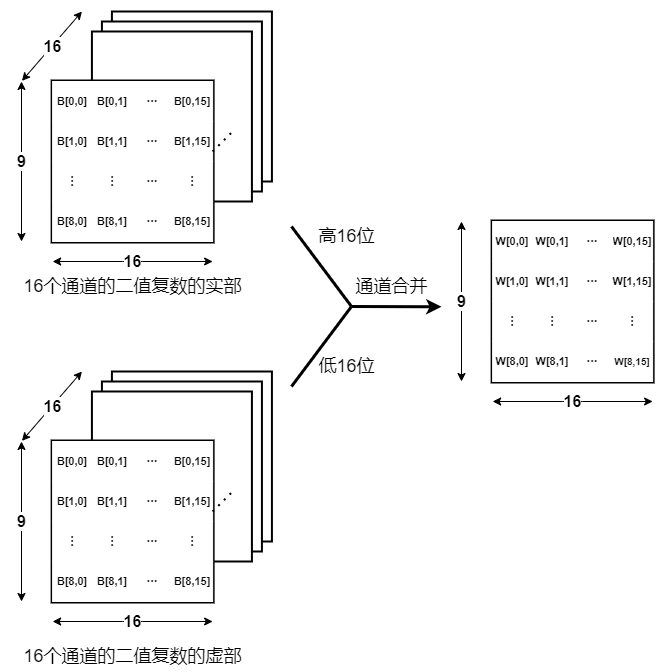


图7

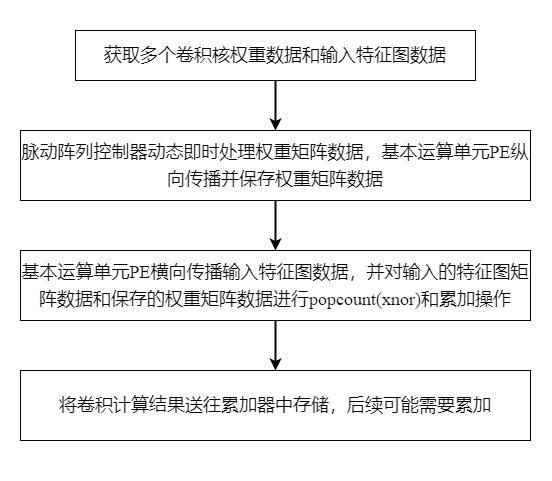


图8