(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110533176 A
(43)申请公布日 2019.12.03

(21)申请号 201810513248.6

(22)申请日 2018.05.25

(71)申请人 北京深鉴智能科技有限公司
地址 100083 北京市海淀区王庄路1号院四号楼1706室

(72)发明人 方绍峡 于谦 王俊斌 隋凌志

(74)专利代理机构 北京展翼知识产权代理事务所(特殊普通合伙) 11452
代理人 张阳

(51)Int.Cl.
G06N 3/063(2006.01)
G06F 12/0875(2016.01)

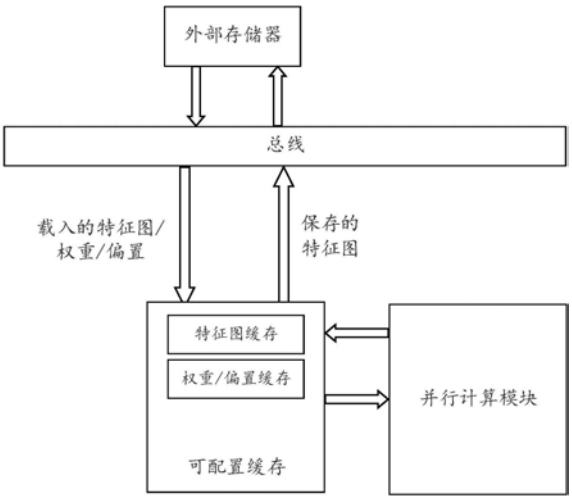
权利要求书2页 说明书8页 附图5页

(54)发明名称

用于神经网络计算的缓存装置及其相关计算平台

(57)摘要

公开了一种用于神经网络计算的缓存装置及其相关计算平台。所述缓存装置包括:可动态配置的片上缓存;以及缓存配置控制器,用于控制所述可动态配置的片上缓存针对神经网络的特定层以不同的比例缓存权重数据和特征图数据。由此,能够在片上缓存总量一定的情况下,通过适应神经网络算法不同阶段的变化,达到最佳的缓存分配比,从而最大化缓存利用率,提高实际计算性能,同时维持较简单的硬件结构。



CN 110533176 A

1.一种用于神经网络计算的缓存装置,包括:

可动态配置的片上缓存;以及

缓存配置控制器,用于控制所述可动态配置的片上缓存针对神经网络的特定层以不同的比例缓存权重数据和特征图数据。

2.如权利要求1所述的缓存装置,其中,所述可动态配置的片上缓存同时用作片上的输入缓存和输出缓存。

3.如权利要求2所述的缓存装置,其中,所述可动态配置的片上缓存是用于神经网络计算的计算平台的唯一片上缓存。

4.如权利要求1所述的缓存装置,其中,针对神经网络的特定层,所述可动态配置的片上缓存具有被分配为固定用于缓存权重数据的第一部分以及被分配为固定用于缓存特征图数据的第二部分。

5.如权利要求4所述的缓存装置,其中,所述可动态配置的片上缓存包括多个缓存单元,其中,在所述缓存配置控制器的控制下,每个缓存单元针对神经网络的特定层被规定为仅缓存权重数据的权重缓存单元或仅缓存特征图数据的特征图缓存单元。

6.如权利要求5所述的缓存装置,其中,所述多个缓存单元包括存储容量相同的三个或以上的缓存单元。

7.如权利要求5所述的缓存装置,还包括:

分别与其对应的缓存单元相连接的读命令选择器单元和写命令选择器单元,所述读命令选择器单元基于所述缓存配置控制器的控制指令选择允许针对权重还是特征图的读取请求通过,所述写命令选择器单元基于所述缓存配置控制器的控制指令选择允许针对权重还是特征图的写入请求通过。

8.如权利要求7所述的缓存装置,还包括:

连接至每个写命令选择器单元中的一个输入的写特征图缓存请求分配器,用于根据所述缓存配置控制器的控制指令将接收到的不同特征图写入请求调度至对应的写选择器单元;

连接至每个写命令选择器单元中的另一个输入的写权重缓存请求分配器,用于根据所述缓存配置控制器的控制指令将不同的权重写入请求调度至对应的写选择器单元;

连接至每个读命令选择器单元中的一个输入的读特征图缓存请求分配器,用于根据所述缓存配置控制器的控制指令将接收到的不同特征图读取请求调度至对应的读选择器单元;以及

连接至每个读命令选择器单元中的另一个输入的读权重缓存请求分配器,用于根据所述缓存配置控制器的控制指令将接收到的不同权重读取请求调度至对应的读选择器单元。

9.如权利要求8所述的缓存装置,还包括:

读数据通路复制单元,用于将从所述片上缓存返回的每份读取数据送回至读特征图缓存请求分配器或读权重缓存请求分配器以进行数据请求匹配,匹配的读取数据被返回发起对应读取请求的模块。

10.一种用于神经网络计算的计算平台,包括:

如权利要求1-9中任一项所述的缓存装置,所述缓存装置从外部存储器读取当前计算所需的特征图数据和权重数据;以及

并行计算模块，用于对从所述缓存装置中读取的读取特征图数据和权重数据的进行高并行度的卷积计算操作，并将计算结果存储回所述缓存装置。

11.如权利要求10所述的计算平台，其中，基于所述神经网络的特定层，预先确定所述片上缓存用于缓存权重数据和特征图数据的比例。

12.如权利要求11所述的计算平台，其中，所述并行计算模块将所述计算结果缓存至所述片上缓存中用于缓存特征图数据的部分。

13.如权利要求12所述的计算平台，其中，所述并行计算模块将所述片上缓存中用于缓存特征图数据的部分中缓存不下的部分计算结果直接存储至所述外部存储器。

14.如权利要求10所述的计算平台，其中，所述并行计算模块至少部分由FPGA、GPU或ASIC实现。

15.一种用于卷积神经网络计算的方法，包括：

获取针对神经网络特定层的片上缓存分配指令；

使用如权利要求1-9中任一项所述的缓存装置或如权利要求10-14中任一项所述的计算平台将特征图数据和权重数据从外部存储器中读取到所述片上缓存中，其中所述片上缓存以所述片上缓存分配指令所规定的比例缓存所述特征图数据和所述权重数据；

所述并行计算模块读取针对所述特定层的多个单次卷积计算操作所需的特征图数据和权重数据以进行高并行度的卷积计算操作；以及

所述并行计算模块将所述卷积计算操作的计算结果缓存回所述片上缓存。

用于神经网络计算的缓存装置及其相关计算平台

技术领域

[0001] 本发明涉及硬件架构领域，尤其涉及一种用于神经网络计算的缓存装置及其相关计算平台与实现方法。

背景技术

[0002] 神经网络(Neural Network)近年来成为图像识别领域的研究热点。经过训练后的神经网络模型，可以用于图像分类、物体识别与显著性检测等诸多领域。近年来神经网络模型呈现计算规模增加、复杂度提升的趋势，利用传统的CPU平台，已无法满足其实用性需求。因此，利用FPGA、GPU、ASIC等高并行度异构计算平台进行神经网络加速器设计成为新的研究热点。

[0003] 在典型的神经网络处理器的设计里，特征图(Featuremap)、权重(Weights)、偏置(Bias)、中间特征图结果、最终特征图结果通常分别存储于不同的片上存储中。采用分立的片上存储形式尽管设计非常简洁，但是对于神经网络计算会导致整体效率偏低。

[0004] 因此，仍然需要一种能够优化神经网络计算的相关方案。

发明内容

[0005] 为了解决上述至少一个问题，本发明提出了一种新的可动态配置的片上缓存方案，其能够在片上缓存总量一定的情况下，通过适应神经网络算法不同阶段的变化，达到最佳的缓存分配比，从而最大化缓存利用率，提高实际计算性能，同时维持较简单的硬件结构。

[0006] 根据本发明的一个方面，提出了一种用于神经网络计算的缓存装置，包括：可动态配置的片上缓存；以及缓存配置控制器，用于控制所述可动态配置的片上缓存针对神经网络的特定层以不同的比例缓存权重数据和特征图数据。由此，能够通过灵活调整缓存分配比，适应神经网络算法不同阶段的变化，提高整体计算性能。

[0007] 优选地，可动态配置的片上缓存可以同时用作片上的输入缓存和输出缓存。可动态配置的片上缓存是用于神经网络计算的计算平台的唯一片上缓存。由此，能够通过合理调度，进一步实现对片上缓存的最大利用率，并简化片上硬件结构。

[0008] 针对神经网络的特定层，所述可动态配置的片上缓存可以具有被分配为固定用于缓存权重数据的第一部分以及被分配为固定用于缓存特征图数据的第二部分。优选地，可动态配置的片上缓存可以包括多个缓存单元，其中，在所述缓存配置控制器的控制下，每个缓存单元针对神经网络的特定层被规定为仅缓存权重数据的权重缓存单元或仅缓存特征图数据的特征图缓存单元。例如，多个缓存单元可以包括存储容量相同的三个或以上的缓存单元。由此，能够通过引入多个缓存单元来简单实现针对片上缓存的比例分配。

[0009] 优选地，缓存装置还可以包括分别与其对应的缓存单元相连接的读命令选择器单元和写命令选择器单元，所述读命令选择器单元基于所述缓存配置控制器的控制指令选择允许针对权重还是特征图的读取请求通过，所述写命令选择器单元基于所述缓存配置控制

器的控制指令选择允许针对权重还是特征图的写入请求通过。由此，通过简单引入选择器，来实现对每个缓存单元缓存内容的灵活切换。

[0010] 优选地，缓存装置还可以包括：连接至每个写命令选择器单元中的一个输入的写特征图缓存请求分配器，用于根据所述缓存配置控制器的控制指令将接收到的不同特征图写入请求调度至对应的写选择器单元；连接至每个写命令选择器单元中的另一个输入的写权重缓存请求分配器，用于根据所述缓存配置控制器的控制指令将不同的权重写入请求调度至对应的写选择器单元；连接至每个读命令选择器单元中的一个输入的读特征图缓存请求分配器，用于根据所述缓存配置控制器的控制指令将接收到的不同特征图读取请求调度至对应的读选择器单元；以及连接至每个读命令选择器单元中的另一个输入的读权重缓存请求分配器，用于根据所述缓存配置控制器的控制指令将接收到的不同权重读取请求调度至对应的读选择器单元。由此，通过以相对简单的分配器结构实现对请求的合理分配。

[0011] 优选地，缓存装置还可以包括：读数据通路复制单元，用于将从所述片上缓存返回的每份读取数据送回至读特征图缓存请求分配器或读权重缓存请求分配器以进行数据请求匹配，匹配的读取数据被返回发起对应读取请求的模块，由此方便对读取数据的分配。

[0012] 根据本发明的另一个方面，提出了一种用于神经网络计算的计算平台，包括：如权上任一项所述的缓存装置，所述缓存装置从外部存储器读取当前计算所需的特征图数据和权重数据；以及并行计算模块，用于对从所述缓存装置中读取的读取特征图数据和权重数据的进行高并行度的卷积计算操作，并将计算结果存储回所述缓存装置。由此，通过对缓存装置的优化而进一步提升系统效率。优选地，可以基于所述神经网络的特定层，预先确定所述片上缓存用于缓存权重数据和特征图数据的比例，以适应神经网络算法中随层数加深而发生的数据比例变化。

[0013] 优选地，并行计算模块可以将计算结果缓存至片上缓存中用于缓存特征图数据的部分。由此实现对片上缓存的输入和输出复用。进一步地，并行计算模块可以将所述片上缓存中用于缓存特征图数据的部分中缓存不下的部分计算结果直接存储至所述外部存储器，以应对输出特征图变大到无法缓存的少数情况。

[0014] 优选地，并行计算模块至少部分由FPGA、GPU或ASIC实现。

[0015] 根据本发明的又一个方面，提出了一种用于卷积神经网络计算的方法，包括：获取针对神经网络特定层的片上缓存分配指令；使用如任一项所述的缓存装置或包括该装置的计算平台将特征图数据和权重数据从外部存储器中读取到所述片上缓存中，其中所述片上缓存以所述片上缓存分配指令所规定的比例缓存所述特征图数据和所述权重数据；所述并行计算模块读取针对所述特定层的多个单次卷积计算操作所需的特征图数据和权重数据以进行高并行度的卷积计算操作；以及所述并行计算模块将所述卷积计算操作的计算结果缓存回所述片上缓存。由此通过提升缓存方案效率来实现高效的卷积神经网络计算。

[0016] 由此，本发明通过提出的可动态配置的缓存方案，能够良好适应特征图和权重比例随层数增加而变化的这一事实，从而提升缓存使用率。进一步地，通过将片上缓存同时用作输入和输出缓存，能够使得缓存利用率最大化。上述比例变化可由多个缓存单元结合分配器和选择器的相对简单的硬件结构实现，从而能在高效利用缓存的同时保持硬件复杂度开销较小。

附图说明

[0017] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0018] 图1示出了现有的用于实现神经网络计算的专用硬件处理器的一个例子。

[0019] 图2示出了图1所示神经网络处理器中的典型神经网络计算数据流。

[0020] 图3给出了一个典型的深度卷积神经网络VGG-16数据量随层变化的曲线。

[0021] 图4示出了根据本发明一个实施例的用于神经网络计算的缓存装置的示意图。

[0022] 图5示出了根据本发明一个实施例的用于神经网络计算的计算平台的数据流示意图。

[0023] 图6示出了根据本发明一个实施例的用于卷积神经网络计算方法的示意性流程图。

[0024] 图7示出了缓存单元动态设置的示意图。

[0025] 图8示出了根据本发明一个实施例的缓存装置的具体实现。

具体实施方式

[0026] 下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

[0027] 长久以来，高并行计算在科学计算、天气模拟、生物模拟、分子力学模型、飞机制造和军事模拟等领域得到了充分的运用。近年来，随着深度学习热潮的持续发酵，用于神经网络，尤其是卷积神经网络(Convolutional Neutral Network，随后简称为CNN)的高并行计算实现方案更是得到了多方关注。

[0028] 典型的神经网络由一系列有序运行的层组成。例如，CNN神经网络由输入层、输出层和多个隐藏层串联组成。CNN的第一层读取输入值(也可称为输入特征图)，例如输入图像，并输出一系列的特征图。下面的层读取由上一层产生的特征图，并输出新的特征图。最后一个分类器输出该输入图像可能属于的每一类别的概率。

[0029] 这些层大致可分为带权重的层(如卷积层、全连接层、批量归一化层等)和不带权重的层(如池化层、ReLU层、Softmax层等)。在这其中，卷积层以一系列特征图作为输入，并以卷积内核卷积获得输出特征图。池化层通常与卷积层相连，用于输出每个特征图中的每个分区的最大值或平均值，由此通过亚采样降低计算量，同时保持某种程度的位移、尺度和形变不变性。一个CNN中可以包括卷积层和池化层之间的多个交替，由此逐步降低空间分辨率并增加特征映射的数量。随后可以连接至至少一个全连接层，通过应用于输入特征向量上的线性变换，得到包括多个特征值的一维向量输出。

[0030] 总体来说，带权重的层的操作可以表示为：

[0031] $Y=WX+b$，

[0032] 其中W为权重值，b为偏移量，X为输入激活值，Y为输出激活值。

[0033] 不带权重的层的操作可以表示为：

[0034] Y＝f(X)，

[0035] 其中f(X)为非线性函数。

[0036] 在此，“权重”(weights)指代隐藏层中的参数，例如用于继续卷积计算的卷积和W，从广义上理解可以包括偏移量b，是通过训练过程习得的数值，并且在推理时保持不变；特征值指代从输入层开始，每一层的输出由输入值和权重值通过运算得到，在各层之间传递的数值，也称为激活值。与权重值不同，特征值的分布会根据输入数据样本而动态变化。

[0037] 在使用CNN进行推理(例如，图像分类)之前，首先需要对CNN进行训练。通过训练数据的大量导入，确定神经网络模型各层的参数，例如权重和偏移量。

[0038] 现有的通用处理器(CPU)由于需要高通用性来处理各种不同的数据类型，并且其逻辑判断会引入大量的分支跳转和中断的处理。这些都使得CPU内部结构异常复杂，不适用于类型高度统一且相互无依赖的大规模数据的数据运算。因此，CNN的训练主要是在大型服务器上实现。而对于CNN推理，则通常会利用FPGA、GPU和ASIC等高并行度异构计算平台进行高并行度计算。在其中，专用的神经网络处理器设计成为神经网络领域里新的研究热点。

[0039] 图1示出了现有的用于实现神经网络计算的专用硬件处理器的一个例子。在现有的神经网络处理器设计中，通常将特征图、权重、偏置、中间特征图结果、最终特征图结果分别存储于不同的片上存储中。图2示出了图1所示神经网络处理器中的典型神经网络计算数据流。如图2所示，处理器通过总线从外部存储将特征图、权重、偏置数据加载至输入缓存。其中输入缓存由特征图缓存、权重缓存和偏置缓存三部分组成。计算单元从输入缓存读取出特征图、权重、偏置数据，并进行计算，将结果写入输出缓存。计算单元产生的中间结果将写入中间特征图结果缓存，最终结果将写入最终特征图结果缓存。中间结果可能会被计算单元再次读取并参与运算。最终特征图结果最终被读取并通过总线写回外部存储。

[0040] 由上可知，在现有的神经网络处理器计算过程中，特征图、权重、偏置、中间特征图结果、最终特征图结果分别存储于不同的片上存储中。采用分立的片上存储形式尽管设计非常简洁，但是对于神经网络计算会导致整体效率偏低。

[0041] 另外，深度神经网络算法一般由数个甚至数百个级联的层组成，随着层数的加深，特征图和权重/偏置数据量会有所变化。图3给出了一个典型的深度卷积神经网络VGG-16数据量随层变化的曲线。从图中可见，在浅层，特征图的数据量大，权重/偏置数据量小，而到了深层，特征图的数据量较小，权重/偏置数据量大。对用于特征提取的神经网络而言，还可能出现随着层数的加深，特征图数据量先变小再变大，权重/偏置数据量先变大再变小的情况。

[0042] 针对上述问题，本发明提出了一种可动态配置的片上缓存方案，其能够在片上缓存总量一定的情况下，通过适应神经网络算法不同阶段的变化，达到最佳的缓存分配比，从而最大化缓存利用率，提高实际计算性能，同时维持较简单的硬件结构。

[0043] 图4示出了根据本发明一个实施例的用于神经网络计算的缓存装置的示意图。如图4所示，该缓存装置400包括可动态配置的片上缓存410和缓存配置控制器420，后者用于控制片上缓存410针对神经网络的特定层以不同的比例缓存权重数据和特征图数据。在此，不同比例可以指代针对神经网络的特定层，片上缓存410具有被分配为用于缓存权重数据的第一部分以及被分配为用于缓存特征值数据的第二部分。上述第一部分和第二部分的大

小和位置在针对同一层的计算过程中保持不变。而对于不同的层，则可以根据特征值和权重的相关比例，灵活调整片上缓存410的分配比例。在此，权重数据指代广义上的权重数据，包括神经网络的权重和偏置参数。

[0044] 在一个实施例中，可动态配置的片上缓存410同时用作片上的输入缓存和输出缓存，优选地，可以是用于神经网络计算的计算平台的唯一片上缓存。由此，并行计算单元计算出的输出特征图可以回存至相应的片上缓存。

[0045] 图5示出了根据本发明一个实施例的用于神经网络计算的计算平台的数据流示意图。如图所示，本发明的缓存装置可以包括在用于进行神经网络计算的计算平台中。该计算平台可以是专门用于神经网络推理的神经网络处理器。在进行推理计算的过程中，本发明的可动态配置片上缓存可以从外部存储器读取当前计算所需的特征图数据和权重数据，例如，经由计算平台的总线进行上述读取。在一个实施例中，可以基于神经网络的特定层，预先确定片上缓存用于缓存权重数据和特征图数据的比例，并基于上述比例进行读取。随后，并行计算模块可以从片上缓存中获取计算所需的特征图数据和权重数据，并将计算结果存回片上缓存。在一个实施例中，该并行计算模块至少部分由FPGA、GPU或ASIC实现。优选地，该并行计算模块可以完全由FPGA或ASIC实现。更优选地，包括该并行计算模块的神经网络计算平台可完全由ASIC实现，并经由外部存储器读取神经网络计算所需的特征图和权重数据并缓存在其上如前所述的缓存装置中。

[0046] 相应地，图6示出了根据本发明一个实施例的用于卷积神经网络计算方法的示意性流程图。在步骤S610，获取针对神经网络特定层的片上缓存分配指令。在步骤S620，使用本发明的缓存装置或是包括该缓存装置的计算平台将特征图数据和权重数据从外部存储器中读取到可动态配置片上缓存中。该片上缓存以片上缓存分配指令所规定的比例缓存特征图数据和权重数据。在步骤S630，并行计算模块读取针对所述特定层的多个单次卷积计算操作所需的特征图数据和权重数据以进行高并行度的卷积计算操作。随后在步骤S640，并行计算模块可以将卷积计算操作的计算结果缓存回片上缓存。

[0047] 上述并行计算模块进行卷积计算的结果是输出特征图，因此可以将其缓存回片上缓存中当前用于缓存特征图数据的部分。例如在层融合场景中，上述输出特征图数据可被看作是中间特征图，并行计算模块可以读取这些中间特征图以进行下一层的卷积计算。在其他场景中，上述输出特征图数据可由片上缓存存回外部存储器，片上缓存于是可以读取其他数据以进行后续计算。

[0048] 在目标识别应用中，特征图的数据量会随着层数的加深而越来越小，因此将输出特征图存回片上缓存中当前用于缓存特征图数据的部分的操作中通常不会出现数据溢出的状态。在例如需要生成图像的应用中，由于特征图的数据量会随着层数的加深而先减小后增大，因此在增大过程中可能出现片上缓存中当前用于缓存特征图数据的部分缓存不下返回的输出特征图。这时，可以将缓存不下的部分或是全部的输出特征图直接存储至所述外部存储器。

[0049] 在一个实施例中，片上缓存的可动态配置特性可以通过设置多个缓存单元来确定。每个缓存单元针对神经网络的特定层，可以被规定仅用于缓存特征图数据或用于缓存权重数据。换句话说，可以针对不同的层将每个缓存单元视为权重缓存或特征图缓存。在一个实施例中，多个缓存单元可以包括存储容量相同的三个或以上的缓存单元。图7示出了缓

存单元动态设置的示意图。如图所示，片上缓存例如可由四个缓存单元(例如，四片RAM)组成。特征图和权重缓存的空间分配可以根据神经网络算法不同层的情况而进行动态配置。在其他实施例中，各缓存单元之间的大小也可以不同，本发明对此不做限制。

[0050] 在具体的实现中，各缓存单元还可以各自连接有与其相对应的读命令选择器单元和写命令选择器单元。读命令选择器单元基于所述缓存配置控制器的控制指令选择允许针对权重还是特征图的读取请求通过，写命令选择器单元基于所述缓存配置控制器的控制指令选择允许针对权重还是特征图的写入请求通过。由此，针对每个缓存单元，在同一时刻(或是神经网络特定层的计算过程中)，其写入端仅能写入权重或是特征图中的一种，相应地，其读取端也仅能读取权重或是特征图中的一种。

[0051] 优选地，还可以引入相应的读写请求分配器来实现对上述读写命令选择器单元的控制。因此在一个实施例中，本发明的缓存装置还可以包括写特征图缓存请求分配器、写权重缓存请求分配器、读特征图缓存请求分配器以及读权重缓存请求分配器。

[0052] 写特征图缓存请求分配器连接至每个写命令选择器单元中的一个输入，用于根据所述缓存配置控制器的控制指令将接收到的不同特征图写入请求调度至对应的写选择器单元。写权重缓存请求分配器连接至每个写命令选择器单元中的另一个输入，用于根据所述缓存配置控制器的控制指令将不同的权重写入请求调度至对应的写选择器单元。读特征图缓存请求分配器连接至每个读命令选择器单元中的一个输入，用于根据所述缓存配置控制器的控制指令将接收到的不同特征图读取请求调度至对应的读选择器单元。读权重缓存请求分配器连接至每个读命令选择器单元中的一个输入，用于根据所述缓存配置控制器的控制指令将接收到的不同权重读取请求调度至对应的读选择器单元。

[0053] 对于读取的特征图和权重数据，本发明的缓存装置还可以包括读数据通路复制单元，用于将从所述片上缓存返回的每份读取数据送回至读特征图缓存请求分配器或读权重缓存请求分配器以进行数据请求匹配，匹配的读取数据被返回发起对应读取请求的模块。

[0054] 图8示出了根据本发明一个实施例的缓存装置的具体实现。如上所述，缓存装置800可以包括写特征图缓存请求分配器(WDF)1、写权重缓存请求分配器(WDWB)2、读特征图缓存请求分配器(RDF)3、读权重缓存请求分配器(RDWB)4、多个读命令选择器单元6、读数据通路复制单元7、片上RAM组8、多个写命令选择器单元10和缓存配置控制器(BC)11。

[0055] 片上RAM组8包括4片RAM，并用作可动态配置的片上缓存。读命令选择器单元6和写命令选择器单元10的数量与片上RAM组8中包含的RAM数量相同，并且多个读命令选择器单元6构成读命令通路选择器集合5，多个写命令选择器单元10构成写命令通路选择器集合9。

[0056] 如下将对上述缓存装置的工作流程进行描述。

[0057] 在准备进行针对神经网络某一层的计算之前，缓存配置控制器11按特定比例将片上RAM组8划分为两部分(如果与前一层相同，也可保持该特定比例不变)，并产生对应的信号控制读命令通路选择器集合5和写命令通路选择器集合9，以确保其包含的每个读命令选择器单元6和写命令选择器单元10的两条通路中只有一条通路可放行。该组控制信号也优选地同时作为写特征图缓存请求分配器1、写权重缓存请求分配器2、读特征图缓存请求分配器3和读权重缓存请求分配器4的参考信号输入。

[0058] 当一个或多个写特征图请求到达写特征图缓存请求分配器1时，写特征图缓存请求分配器1负责处理各个请求的优先级，并根据缓存配置控制器11信息，将请求调度至不同

的写命令选择器单元10。写命令选择器单元10根据缓存配置控制器11信息给予该命令放行，并最终写入片上RAM组(8)对应的RAM块。相应地，当一个或多个写权重请求到达写权重缓存请求分配器2时，写权重缓存请求分配器2负责处理各个请求的优先级，并根据缓存配置控制器11信息，将请求调度至不同的写命令选择器单元10。写命令选择器单元10根据缓存配置控制器11信息给予该命令放行，并最终写入片上RAM组8对应的RAM块。在此，权重不仅包括用作卷积核的权重参数，还包括偏置。

[0059] 当一个或多个读特征图请求到达读特征图缓存请求分配器3时，读特征图缓存请求分配器3负责处理各个请求的优先级，并根据缓存配置控制器11信息，将请求调度至不同的读命令选择器单元6，并保留尚未返回读数据的所有读请求ID。读命令选择器单元6根据缓存配置控制器11信息给予该命令放行，并最终送至片上RAM组8对应的RAM块读端口。相应地，当一个或多个读权重请求到达读权重缓存请求分配器4时，读权重缓存请求分配器4负责处理各个请求的优先级，并根据缓存配置控制器11信息，将请求调度至不同的读命令选择器单元6，并保留尚未返回读数据的所有读请求ID。读命令选择器单元6根据缓存配置控制器11信息给予该命令放行，并最终送至片上RAM组8对应的RAM块读端口。

[0060] 所有从RAM块返回的读数据及原始请求ID信号经过读数据通路复制单元7分别送回读特征图缓存请求分配器3和读权重缓存请求分配器4进行读数据和读请求匹配，并将完成匹配的读数据返回发起读请求模块。

[0061] 应该理解的是，针对读请求和写请求的处理可以根据情况同时或相继进行，本发明对此不做限制。虽然如上主要结合针对卷积神经网络来描述本发明的动态缓存方案，但本领域技术人员应该理解的是，本发明的硬件架构适用于各类神经网络计算计算场景，尤其适用于深度神经网络的应用场景。

[0062] 上文中已经参考附图详细描述了根据本发明的可配置缓存方案、相关计算平台及实现方法。

[0063] 由此，本发明通过提出的可动态配置的缓存方案，能够良好适应特征图和权重比例随层数增加而变化的这一事实，从而提升缓存使用率。进一步地，通过将片上缓存同时用作输入和输出缓存，能够使得缓存利用率最大化。上述比例变化可由多个缓存单元结合分配器和选择器的相对简单的硬件结构实现，从而能在高效利用缓存的同时保持硬件复杂度开销较小。

[0064] 此外，根据本发明的方法还可以实现为一种计算机程序或计算机程序产品，该计算机程序或计算机程序产品包括用于执行本发明的上述方法中限定的上述各步骤的计算机程序代码指令。

[0065] 或者，本发明还可以实施为一种非暂时性机器可读存储介质(或计算机可读存储介质、或机器可读存储介质)，其上存储有可执行代码(或计算机程序、或计算机指令代码)，当所述可执行代码(或计算机程序、或计算机指令代码)被电子设备(或计算设备、服务器等)的处理器执行时，使所述处理器执行根据本发明的上述方法的各个步骤。

[0066] 本领域技术人员还将明白的是，结合这里的公开所描述的各种示例性逻辑块、模块、电路和算法步骤可以被实现为电子硬件、计算机软件或两者的组合。

[0067] 附图中的流程图和框图显示了根据本发明的多个实施例的系统和方法的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程

序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标记的功能也可以以不同于附图中所标记的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0068] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

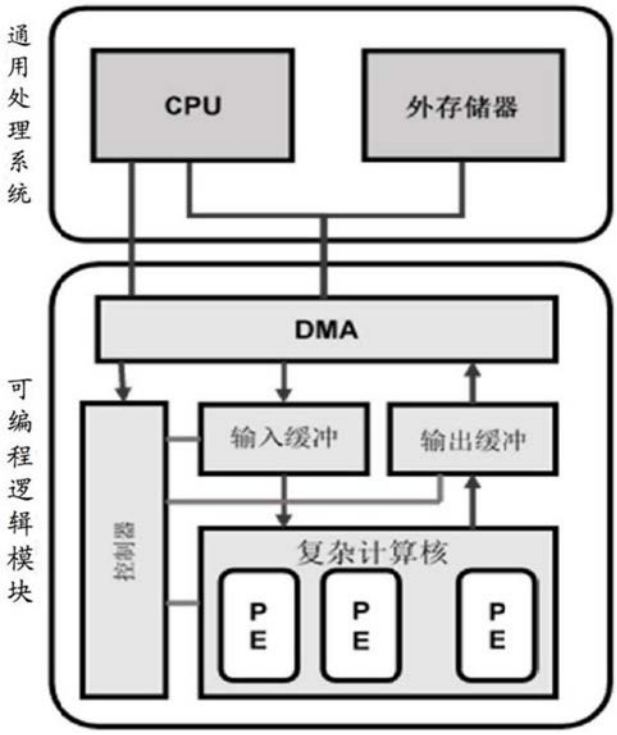

通用处理系统
CPU
外存储器
可编程逻辑模块
DMA
输入缓冲
输出缓冲
控制器
复杂计算核
PE
PE
PE

图1

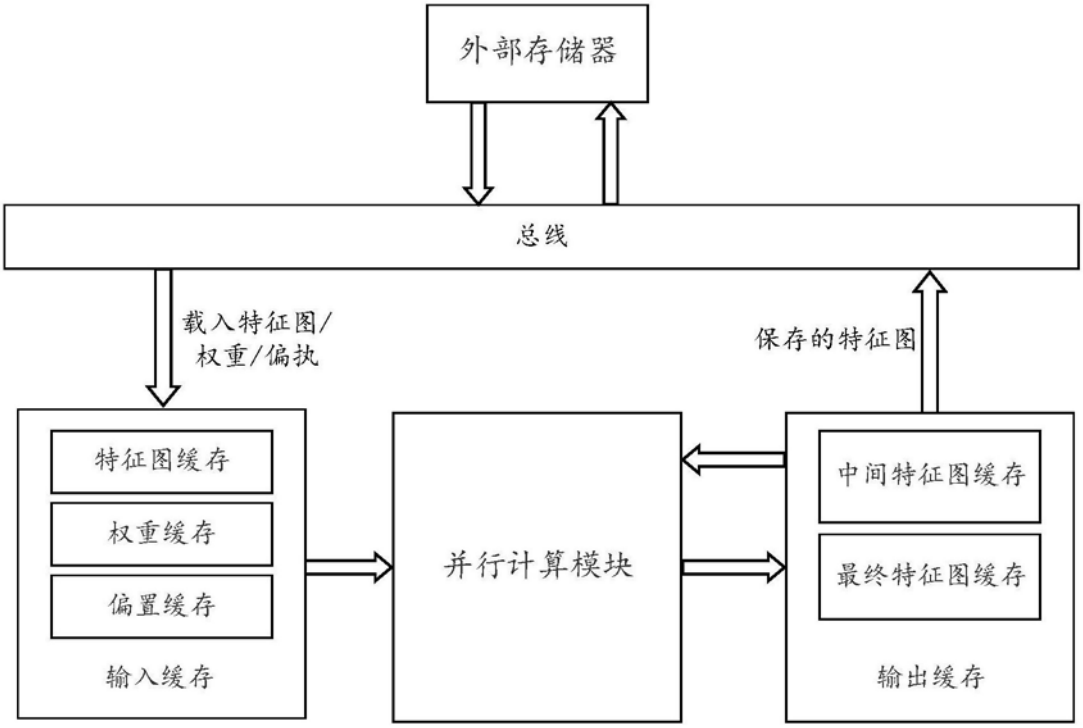

外部存储器
总线
载入特征图/
权重/偏执
保存的特征图
特征图缓存
权重缓存
偏置缓存
输入缓存
并行计算模块
中间特征图缓存
最终特征图缓存
输出缓存

图2

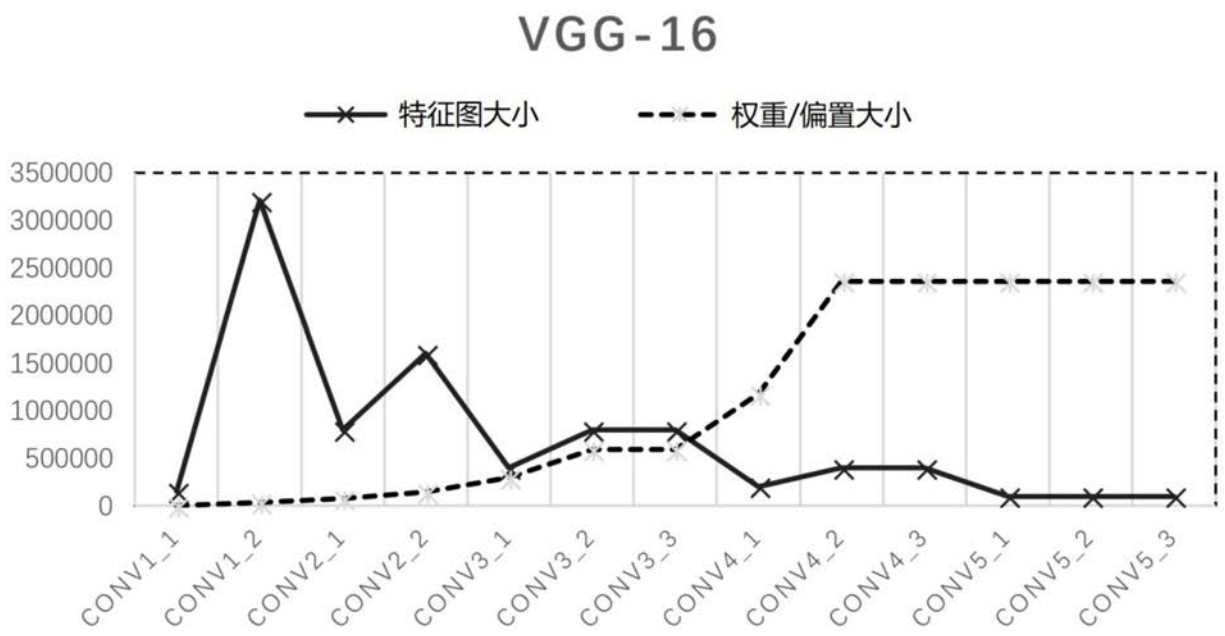

VGG-16
特征图大小
权重/偏置大小
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
CONV1_1
CONV1_2
CONV2_1
CONV2_2
CONV3_1
CONV3_2
CONV3_3
CONV4_1
CONV4_2
CONV4_3
CONV5_1
CONV5_2
CONV5_3

图3

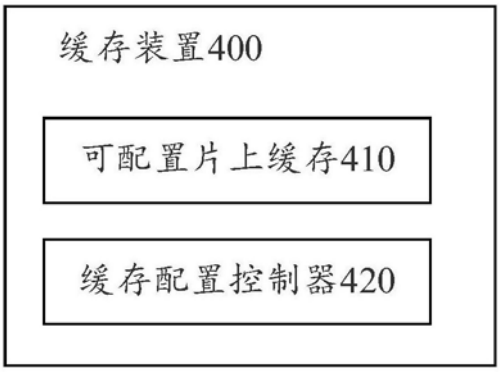

缓存装置400
可配置片上缓存410
缓存配置控制器420

图4

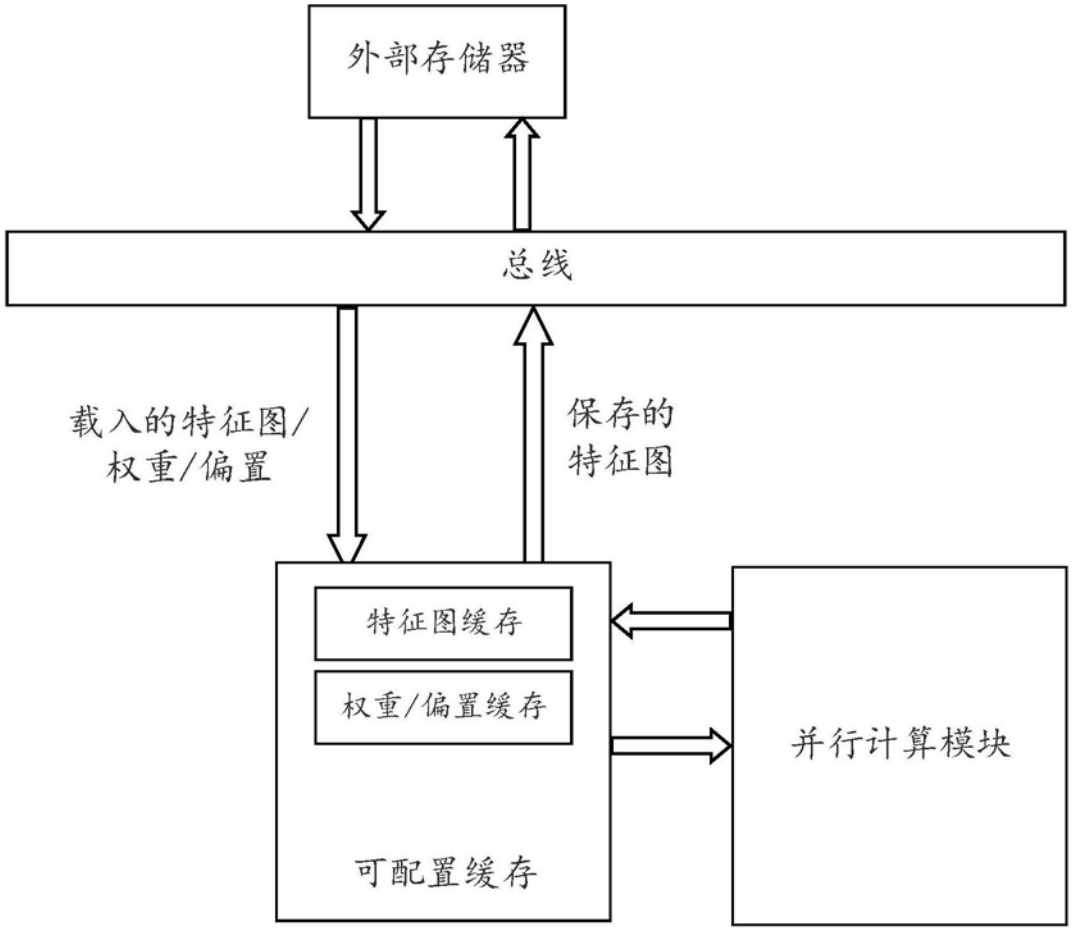


图5

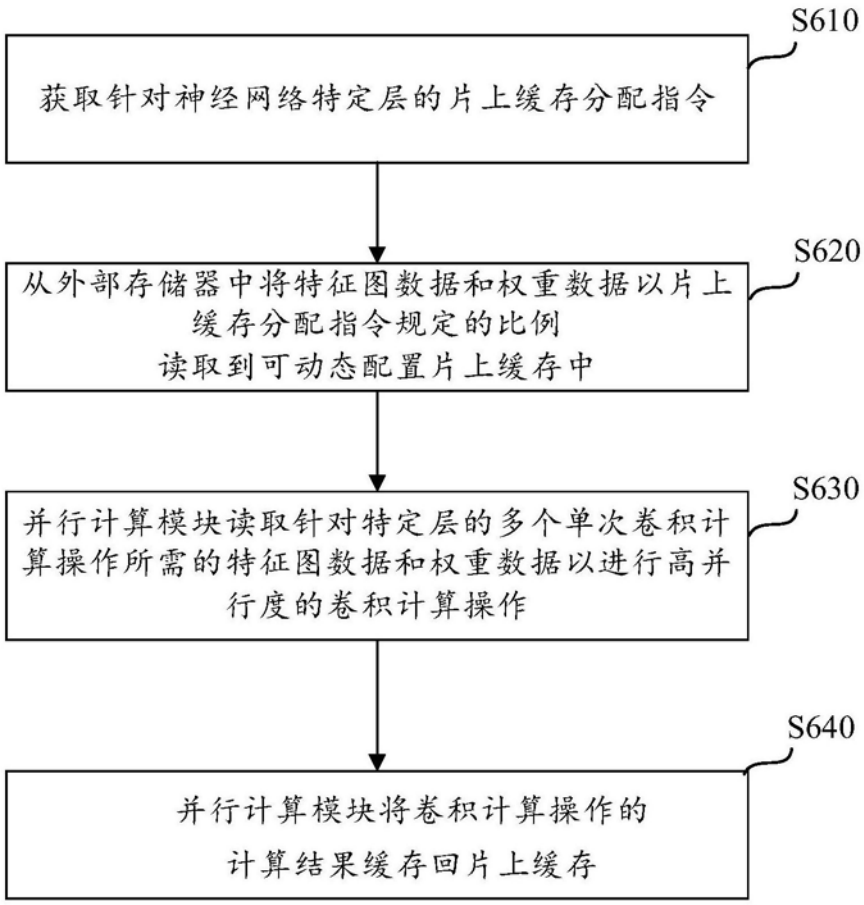

S610
获取针对神经网络特定层的片上缓存分配指令
S620
从外部存储器中将特征图数据和权重数据以片上缓存分配指令规定的比例读取到可动态配置片上缓存中
S630
并行计算模块读取针对特定层的多个单次卷积计算操作所需的特征图数据和权重数据以进行高并行度的卷积计算操作
S640
并行计算模块将卷积计算操作的计算结果缓存回片上缓存

图6

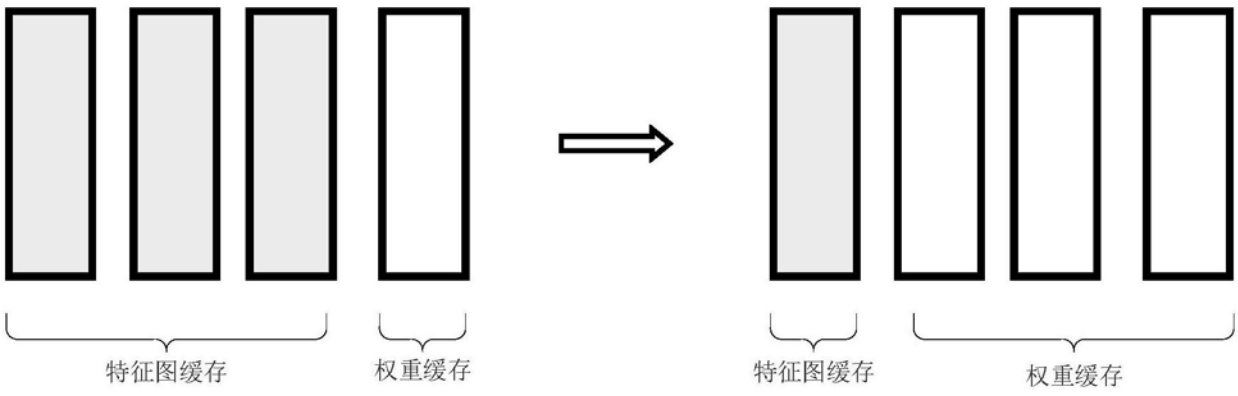

特征图缓存
权重缓存
特征图缓存
权重缓存

图7

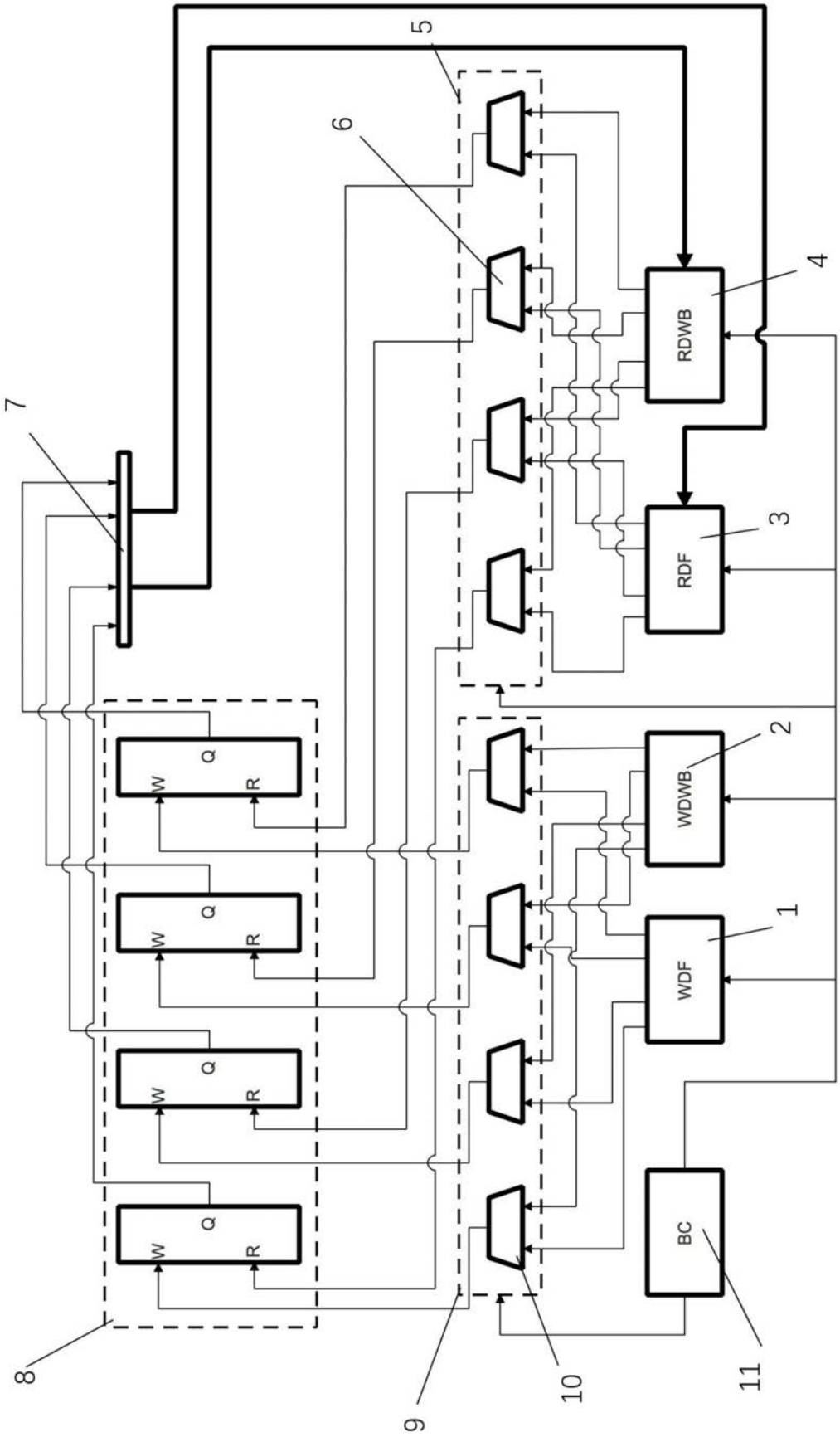


图8