(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111199275 B
(45)授权公告日 2023.04.28

(21)申请号 201811383562.3

(22)申请日 2018.11.20

(65)同一申请的已公布的文献号
申请公布号 CN 111199275 A

(43)申请公布日 2020.05.26

(73)专利权人 上海登临科技有限公司
地址 201203 上海市浦东新区盛夏路570号901室

(72)发明人 王平 孙洁

(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
专利代理师 王勇 李科

(51)Int.Cl.
G06N 3/063(2023.01)
G06N 5/04(2023.01)
G06F 15/78(2006.01)

(56)对比文件
WO 2018107383 A1,2018.06.21
EP 3373210 A1,2018.09.12
CN 107818367 A,2018.03.20
CN 107918794 A,2018.04.17
CN 107003989 A,2017.08.01
郭文生;李国和.人工神经网络在并行计算机集群上的设计研究.计算机应用与软件.2010,(05),全文.

审查员 王黎明

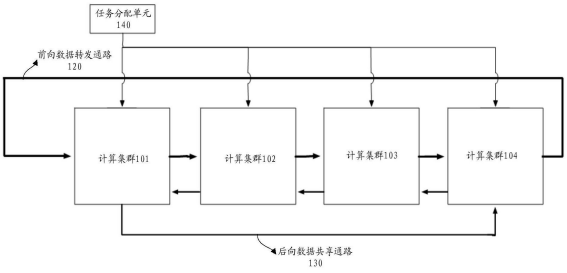
权利要求书2页 说明书10页 附图3页

(54)发明名称

用于神经网络的片上系统

(57)摘要

本发明提供一种用于神经网络的片上系统。该片上系统包括多个计算集群、前向数据转发通路、后向数据共享通路和任务分配单元，其中，所述多个计算集群用于实现神经网络中输入神经元矩阵和权重矩阵的乘法操作，其中，每个计算集群包括本地的片上内存以及对应的片外内存；所述前向数据转发通路用于在所述多个计算集群之间转发输入神经元数据；所述后向数据共享通路用于在所述多个计算集群之间传递权重数据或计算结果；所述任务分配单元用于根据待计算的输入神经元矩阵规模确定每个计算集群的任务分配策略，从而为每个计算集群分配待执行矩阵乘法操作的输入神经元数据。本发明的片上系统能够提高资源利用率和运算效率。



CN 111199275 B

1.一种用于神经网络的片上系统,包括多个计算集群、前向数据转发通路、后向数据共享通路和任务分配单元,其中:

所述多个计算集群用于实现神经网络中输入神经元矩阵和权重矩阵的乘法操作,其中,每个计算集群包括本地的片上内存以及对应的片外内存;

所述前向数据转发通路用于在所述多个计算集群之间转发输入神经元数据;

所述后向数据共享通路用于在所述多个计算集群之间传递权重数据或计算结果;

所述任务分配单元用于根据待计算的输入神经元矩阵规模确定每个计算集群的任务分配策略,从而为每个计算集群分配待执行矩阵乘法操作的输入神经元数据。

2.根据权利要求1所述的片上系统,其特征在于,所述计算集群包括数据流控制模块、数据缓存模块、乘累加模块、数据传送模块、片上内存,其中:

所述数据缓存模块用于存储神经元数据、权重数据或计算结果数据;

所述乘累加模块用于实现所述输入神经元矩阵和对应权重矩阵的乘法运算;

所述数据流控制模块用于控制数据向所述数据缓存模块、所述乘累加模块、所述数据传送模块、和所述片上内存的加载;

所述数据传送模块用于向其他的计算集群转发神经元数据。

3.根据权利要求1或2所述的片上系统,其特征在于,所述后向数据共享通路由依次连接的多个转发器构成,其中,每个转发器对应于一个计算集群,用于将从其他转发器接收的权重数据或计算结果发送至对应的计算集群。

4.根据权利要求1或2所述的片上系统,其特征在于,所述任务分配单元还用于根据所述权重矩阵的规模或者所述多个计算集群的计算能力中的至少一项确定所述权重矩阵在所述多个计算集群的本地片上内存以及对应的片外内存的存储策略。

5.根据权利要求4所述的片上系统,其特征在于,在所述输入神经元矩阵为B×N×K,所述权重矩阵为K×M,有b个计算集群,每个计算集群的计算能力为k×m,N、M、K、k、m、b为任意正整数的情况下:

所述任务分配策略是,对每个计算集群并行分配$\left\lceil B/b \right\rceil$个输入神经元数据矩阵。

6.根据权利要求5所述的片上系统,其特征在于,在M≤m的情况下所述权重矩阵的存储策略是,将所述权重矩阵在每个计算集群对应的片外内存各存储一份或者将所述权重矩阵在一个计算集群对应的片外内存存储一份或者将所述权重矩阵平均分成多个子矩阵分别存储在每个计算集群对应的片外内存中。

7.根据权利要求6所述的片上系统,其特征在于,在将所述权重矩阵在一个计算集群对应的片外内存存储一份的情况下,执行矩阵乘法运算时,该计算集群从其对应的片外内存将所述权重矩阵加载到本地的片上内存,并经由所述后向数据共享通路向其余的计算集群传递所述权重矩阵。

8.根据权利要求6所述的片上系统,其特征在于,在将所述权重矩阵平均分成多个子矩阵分别存储在每个计算集群对应的片外内存中的情况下,执行矩阵乘法运算时,各计算集群从其对应的片外内存将所述权重矩阵加载到本地的片上内存,并经由所述后向数据共享通路向其余的计算集群传递所述权重矩阵。

9.根据权利要求4所述的片上系统，其特征在于，在所述输入神经元矩阵为B×N×K，所述权重矩阵为K×M，有b个计算集群，每个计算集群的计算能力为k×m，N、M、K、k、m、b为任意正整数，并且M≥b×m的情况下：

所述任务分配策略是，对每个计算集群并行分配$\left\lceil B/b \right\rceil$个输入神经元矩阵；

所述权重矩阵的存储策略是，将所述权重矩阵根据所述多个计算集群的计算能力分割为多个子矩阵并将该多个子矩阵分布于所述多个计算集群对应的片外内存。

10.根据权利要求1或2所述的片上系统，其特征在于，所述前向数据转发通路在第一方向上依次串联所述多个计算集群，以形成传递输入神经元数据的环路，所述后向数据共享通路在第二方向上依次串联所述多个计算集群，以形成传递权重数据或计算结果的环路。

11.一种电子设备，包括权利要求1至10任一项所述的片上系统。

用于神经网络的片上系统

技术领域

[0001] 本发明涉及人工智能技术领域，尤其涉及一种用于神经网络的片上系统。

背景技术

[0002] 人工智能技术在近些年来得到了迅猛的发展，在全世界范围内得到了广泛的关注，无论是工业界还是学术界都开展了人工智能技术的研究工作，目前，人工智能技术已经渗透至视觉感知、语音识别、辅助驾驶、智能家居、交通调度等各个领域。

[0003] 深度神经网络是人工智能领域具有较高发展水平的感知模型之一，其通过建立模型来模拟人类大脑的神经连接结构，通过多个变换阶段分层对数据特征进行描述，为图像、视频和音频等大规模数据处理任务带来突破性进展。深度神经网络模型是一种运算模型，由大量节点通过网状互连结构构成，这些节点被称为神经元(或称输入数据)。每两个节点间连接强度都代表通过该连接信号在两个节点间的系数，即权重，与人类神经网络中的记忆相对应。

[0004] 针对输入神经元数据和权重矩阵的矩阵乘法是神经网络推理及训练应用中的典型运算，其运算电路和相关的数据输入输出通路是性能与功耗优化的重点，基本特征是矩阵乘单元的规模越大，功耗效果越好，但是其缺陷是在应用的浅层(即输入神经元数据规模较大，权重规模较小的情况)，性能会由于输出的分量数不够多而浪费计算资源。相比之下，越小的矩阵乘单元规模，计算资源的利用率越高，但是反复的调度，数据频繁访问(例如，对片外DRAM访问、片上通用寄存器堆、共享存储器)则使其功耗效率更低。

[0005] 现有的用于神经网络处理的片上系统主要存在三类架构，第一类是片上集中存储的单处理器网络，这类单处理器网络架构简单，但在浅层处理效率较低，并且数据搬运能耗较高；第二类是对称的多处理器网络，在这类架构中，多处理器之间无数据通信，每个处理器单元的存储有限，内存换入换出频繁，从而导致能耗更高；第三类是片上网络级联的架构形式，在这类架构中，由于神经网络中不同层之间性能不均衡，导致处理器的处理效率受限于性能较差的层，从而浪费片上处理资源。

[0006] 因此，为了将神经网络推向更广泛应用，例如，智能穿戴、智能机器人、自动驾驶以及模式识别等领域，需要对现有技术进行改进，以提高神经网络数据处理的效率、降低运行功耗并提升计算资源利用率。

发明内容

[0007] 本发明的目的在于克服上述现有技术的缺陷，提供一种用于神经网络的片上系统，通过改进片上系统的处理器架构和资源调度方式来提高计算能效比。

[0008] 根据本发明的第一方面，提供一种用于神经网络的片上系统。该片上系统包括多个计算集群、前向数据转发通路、后向数据共享通路和任务分配单元，其中：

[0009] 所述多个计算集群用于实现神经网络中输入神经元矩阵和权重矩阵的乘法操作，其中，每个计算集群包括本地的片上内存以及对应的片外内存；

[0010] 所述前向数据转发通路用于在所述多个计算集群之间转发输入神经元数据；

[0011] 所述后向数据共享通路用于在所述多个计算集群之间传递权重数据或计算结果；

[0012] 所述任务分配单元用于根据待计算的输入神经元矩阵规模确定每个计算集群的任务分配策略，从而为每个计算集群分配待执行矩阵乘法操作的输入神经元数据。

[0013] 在一个实施例中，所述计算集群包括数据流控制模块、数据缓存模块、乘累加模块、数据传送模块、片上内存，其中：

[0014] 所述数据缓存模块用于存储神经元数据、权重数据或计算结果数据；

[0015] 所述乘累加模块用于实现所述输入神经元矩阵和对应权重矩阵的乘法运算；

[0016] 所述数据流控制模块用于控制数据向所述数据缓存模块、所述乘累加模块、所述数据传送模块、和所述片上内存的加载；

[0017] 所述数据传送模块用于向其他的计算集群转发神经元数据。

[0018] 在一个实施例中，所述后向数据共享通路由依次连接的多个转发器构成，其中，每个转发器对应于一个计算集群，用于将从其他转发器接收的权重数据或计算结果发送至对应的计算集群。

[0019] 在一个实施例中，所述任务分配单元还用于根据所述权重矩阵的规模或者所述多个计算集群的计算能力中的至少一项确定所述权重矩阵在所述多个计算集群的本地片上内存以及对应的片外内存的存储策略。

[0020] 在一个实施例中，在所述输入神经元矩阵为B×N×K，所述权重矩阵为K×M，有b个计算集群，每个计算集群的计算能力为k×m，N、M、K、k、m、b为任意正整数的情况下：

[0021] 所述任务分配策略是，对每个计算集群并行分配$\left\lceil B/b \right\rceil$个输入神经元数据矩阵。

[0022] 在一个实施例中，在M≤m的情况下所述权重矩阵的存储策略是，将所述权重矩阵在每个计算集群对应的片外内存各存储一份或者将所述权重矩阵在一个计算集群对应的片外内存存储一份或者将所述权重矩阵平均分成多个子矩阵分别存储在每个计算集群对应的片外内存中。

[0023] 在一个实施例中，在将所述权重矩阵在一个计算集群对应的片外内存存储一份的情况下，执行矩阵乘法运算时，该计算集群从其对应的片外内存将所述权重矩阵加载到本地的片上内存，并经由所述后向数据共享通路向其余的计算集群传递所述权重矩阵。

[0024] 在一个实施例中，在将所述权重矩阵平均分成多个子矩阵分别存储在每个计算集群对应的片外内存中的情况下，执行矩阵乘法运算时，各计算集群从其对应的片外内存将所述权重矩阵加载到本地的片上内存，并经由所述后向数据共享通路向其余的计算集群传递所述权重矩阵。

[0025] 在一个实施例中，在所述输入神经元矩阵为B×N×K，所述权重矩阵为K×M，有b个计算集群，每个计算集群的计算能力为k×m，N、M、K、k、m、b为任意正整数，并且M≥b×m的情况下：

[0026] 所述任务分配策略是，对每个计算集群并行分配$\left\lceil B/b \right\rceil$个输入神经元矩阵；

[0027] 所述权重矩阵的存储策略是，将所述权重矩阵根据所述多个计算集群的计算能力分割为多个子矩阵并将该多个子矩阵分布于所述多个计算集群对应的片外内存。

[0028] 在一个实施例中，所述前向数据转发通路在第一方向上依次串联所述多个计算集群，以形成传递输入神经元数据的环路，所述后向数据共享通路在第二方向上依次串联所述多个计算集群，以形成传递权重数据或计算结果的环路。

[0029] 根据本发明的第二方面，提供了一种电子设备。该电子设备包括本发明的片上系统。

[0030] 与现有技术相比，本发明的优点在于：针对神经网络推理应用中不同层的运算特征，提出了一种统一的多计算集群协调的片上系统架构，解决了单一运算单元在推理应用浅层计算效率低的问题；通过设计专用的数据转发通路与片上网络实现多计算集群之间的数据共享；根据输入神经元矩阵规模或权重矩阵规模将所需较重的带宽负载调度在本计算集群内部，而通过数据转发通路传递较轻的带宽负载，从而实现了局部访存的能耗优化。

附图说明

[0031] 以下附图仅对本发明作示意性的说明和解释，并不用于限定本发明的范围，其中：

[0032] 图1是根据本发明一个实施例的用于神经网络的片上系统的架构示意图；

[0033] 图2是根据本发明一个实施例的片上系统的计算集群的结构示意图；

[0034] 图3是根据本发明另一个实施例的片上系统的结构示意图。

具体实施方式

[0035] 为了使本发明的目的、技术方案、设计方法及优点更加清楚明了，以下结合附图通过具体实施例对本发明进一步详细说明。应当理解，此处所描述的具体实施例仅用于解释本发明，并不用于限定本发明。

[0036] 在本文示出和讨论的所有例子中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它例子可以具有不同的值。

[0037] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

[0038] 在本文的描述中，输入神经元数据是神经网络模型中节点数据，权重是指连接两个节点的系数，可通过训练获得，而数据通常泛指输入神经元数据、权重数据和计算结果等各类型数据，除非根据上下文另有所指。

[0039] 根据本发明的一个实施例，提供了一种用于神经网络处理的片上系统，参见图1所示，该系统包括多个计算集群(或称处理器集群)，示出了计算集群101、计算集群102、计算集群103以及计算集群104，前向数据转发通路120、后向数据共享通路130和任务分配单元140。

[0040] 计算集群101-104用于执行矩阵乘运算功能，可由一个或多个处理单元构成，例如，仅包含矩阵乘处理单元，或者包含矩阵乘处理单元和其他类型单元。每个计算集群可以具有相同或不同的电路结构，例如可以由ASIC或DSP等多种类型的电路结构实现，每个计算集群的计算能力可以相同或不相同。此外，每个计算集群具有其专有的片上存储器(在本文中也称为本地片上存储器)和片外存储器，其中片上存储器例如可采用SRAM或其他类型，片

外存储器例如可采用DDR颗粒或其他类型，本发明对此并不进行限制。

[0041] 前向数据转发通路120构成环形通路，用于在多个计算集群之间转发输入神经元数据，每个计算集群可将从外部(如片外内存)读取的神经元数据或接收的其他计算集群转发的神经元数据经由该前向数据转发通路120依次转发给与之连接的其他计算集群，从而神经元数据可在多个计算集群之间循环流动。

[0042] 后向数据共享通路130构成环形通路，用于在所述多个计算集群之间传递权重数据或矩阵乘的计算结果，每个计算集群可将从外部(如片外内存)读取的权重数据或接收的来自于其他计算集群的权重数据经由该后向数据共享通路130依次转发给与之连接的其他计算集群，从而使权重数据在多个计算集群之间循环流动。通过这种方式，各计算集群能够实现对片上存储器和片外存储器的访问。

[0043] 在一个实施例中，可对每个计算集群用于数据共享的片上内存资源统一编址，并对片外内存资源也统一编址，编址位包括用于标识片外内存和片上内存的部分、用于标识所选择的计算集群的部分以及用于标识具体的片外内存地址或片上内存地址的部分。例如，统一的地址位数是35bit，其中，最高位bit34用于标识是片上或者片外，bit33和bit32用于选择计算集群，对于4个计算集群的情况，采用2bit可选择任一个对应的计算集群，低位32bits，即bit31-bit0可表示4G的地址。参见表1所示。

[0044] 表1:比特位标识

[0045]

比特位编号	作用
bit34	用于指示片外内存或片外内存
Bit33-bit32	用于选择计算集群
Bit31-bit0	用于指示具体片上内存地址或片外内存地址

[0046] 由表1可以看出，通过这种方式，每个计算集群能够访问片上的4G存储空间以及片外的4G存储空间。

[0047] 任务分配单元140用于根据待计算任务的需求和每个计算集群的计算能力等，确定多个计算集群的任务分配策略以及片上、片外存储策略。任务分配单元140可由片上系统的软件模块实现。关于任务分配策略和片上、片外的存储策略将在下文进一步介绍。

[0048] 图2示出了根据本发明一个实施例的一个计算集群的结构图，该计算集群包括数据流控制模块210、数据缓存模块220、乘累加模块230、数据传送模块240、片上内存250、转发器260。

[0049] 数据流控制模块210与数据缓存模块220、乘累加模块230和数据传送模块240以及转发器260具有通信连接(其中未示出与转发器260的连接)，可接收来自于任务分配单元的任务分配策略和片上、片外内存的存储策略，并根据这些策略控制以及任务执行情况控制数据(包含神经元数据、权重或矩阵乘结果等)在该计算集群中各模块之间的传递以及与该计算集群外部的数据交互。例如，包括但不限于控制数据缓存模块220从片上内存250选择数据，控制转发器260从其他计算集群的转发器接收权重数据，控制从该计算集群外部向数据缓存模块220加载数据，并进而控制传递给乘累加模块230，或者在乘累加模块230执行完矩阵乘法运算之后，控制神经元数据传递给数据传送模块240，并进而控制数据传送模块

240向后续的计算集群传递该神经元数据等。

[0050] 数据缓存模块220用于缓存多种类型的数据。例如，包括但不限于待执行矩阵乘法运算的权重数据、输入神经元数据、乘累加模块230的计算结果等。

[0051] 乘累加模块230用于执行权重矩阵和输入神经元矩阵的乘法操作，其可包括一个或多个矩阵乘处理单元，以快速处理不同规模的矩阵乘法运算。

[0052] 数据传送模块240用于构成前向数据转发通路，其与计算集群内部的数据流控制模块240具有通信连接，并且与其他的计算集群也具有通信连接，以将数据传递给其他的计算集群。例如，传递给其他计算集群的数据缓存模块。

[0053] 片上内存250，即该计算集群的本地片上内存，用于存储多种类型的数据，例如，神经元数据或权重数据等。

[0054] 转发器260用于构成后向数据共享通路，可从外部内存加载权重数据、接收来自于其他计算集群的权重数据或者将权重数据转发给其他计算集群(例如通过与其他计算集群的转发器交互实现)或者接收来自于其他计算集群的矩阵乘法结果将其存入片上内存250或者将矩阵乘法结果转发给其他计算集群。

[0055] 为了图示清晰起见，在图2中没有示出数据流控制模块210与片上内存250和转发器260之间的连接关系。对于本领域的普通技术人员来说，为了实现本发明的功能，这种连接关系是可以理解的。而且，对于本领域的普通技术人员来说，也可以不局限于图2所示出部件以及部件之间的连接关系，可以根据系统的需要和目的，增加、删除某些部件，以及改变部件之间的连接关系。

[0056] 结合图2的计算集群结构，当片上系统包括多个计算集群时，在数据流控制模块210的控制下，可形成前向数据转发通路和后向数据共享通路。

[0057] 例如，前向数据转发通路由数据缓存模块220、乘累加模块230、数据传送模块240以及其他计算集群中的这些模块构成，即神经元数据可依次转发至数据缓存模块220、乘累加模块230、数据传送模块240以及其他计算集群的数据缓存模块、乘累加模块、数据传送模块。又如，前向数据转发通路由数据缓存模块220、数据传送模块240以及其他计算集群的数据缓存模块的、乘累加模块和数据传送模块构成，在这种情况下，某些神经元数据可不参与乘累加运算，而直接转发给其他的计算集群。

[0058] 例如，后向数据共享通路包括转发器260和其他计算集群中的转发器构成，即权重数据或矩阵乘的计算结果由转发器260发送给与其连接的计算集群的转发器。

[0059] 应理解的是，图2中各模块之间的连接关系仅用于示意，在实际应用中，本领域技术人员可作适当的变型，例如，乘累加模块230的计算结果也可暂存于数据缓存模块220，在适当时机再转发给其他的计算集群。

[0060] 图3是根据本发明另一个实施例的片上系统，该系统与图1所示的系统类似，也包括计算集群301-304、前向数据转发通路320、后向数据共享通路430和任务分配单元(未示出)，但是与图1相比，用于构成后向数据共享通路的转发器布置在计算集群的外部，并示出了后向数据共享通路430。后向数据共享通路430由多个转发器连接构成，转发器与计算集群一一对应，可与对应的计算集群进行数据交互，分别标记为转发器401、转发器402、转发器403和转发器404。

[0061] 下文将结合图3介绍任务分配策略和片上、片外存储策略以及对应的数据处理过

程。

[0062] 在一个实施例中，根据待计算矩阵的规模(包括输入神经元矩阵规模和/或权重矩阵的规模)或者每个处理器的计算能力来确定每个计算集群需要执行的任务，以及计算过程中的片上、片外存储策略，以通过选择不同方案，最大程度的实现计算资源的高效利用和最小的数据传输。

[0063] 例如，假设输入神经元数据矩阵规模为B×N×K，表示有B个N×K(N为行维数，K为列维数)的输入神经元数据矩阵，有多个权重矩阵，权重矩阵规模为K×M(K为行维数，M为列维数)，并且，为便于说明，假设总共b个计算集群的计算能力相同，均为k×m(即一次执行矩阵乘运算时，权重矩阵的行维数为k，列维数为m)，其中，N、M、K、k、m为任意正整数。下面分别介绍权重规模小、权重规模大两种情况下任务分配策略和片上、片外存储策略。

[0064] 1)、权重规模较小的情况

[0065] 例如，如果M≤m，这种计算典型存在于图像识别应用的较浅层中，通常M值较小、K也较小，因此权重矩阵的规模K×M也较小。

[0066] 在这种情况下，子任务分配策略为，采用输入神经元矩阵并行的方式，为每个计算集群并行分配$\left\lceil B/b \right\rceil$个待计算的输入神经元矩阵。

[0067] 在一个实施例中，权重矩阵的存储策略为：将权重矩阵在每个计算集群对应的片外内存各存一份，执行矩阵乘运算时，各计算集群将权重矩阵从片外内存加载到本地片上内存，这样在运行推理中所有输入神经元矩阵和权重矩阵都由计算集群本地处理，而计算集群之间无须进行数据通信，在这种情况下，后向数据共享通路不作为。通过这种方式，能够降低访问延迟和访问能耗。

[0068] 在另一个实施例中，权重矩阵的存储策略是，将权重矩阵平均分配各个计算集群的片内内存中，执行矩阵乘运算时，每个计算集群中乘累加模块从本地片上内存加载权重矩阵，而通过后向数据共享通路获得其他计算集群的权重矩阵。

[0069] 为便于理解，结合图3所示的片上系统，下表2示意了不同时刻计算集群的行为，表3示意了不同时刻转发器的行为。具体地，以每个计算集群并行分配$\left\lceil B/b \right\rceil$个输入神经元矩阵为例，分别标记为$\left\lceil B/b \right\rceil 1$、$\left\lceil B/b \right\rceil 2$、$\left\lceil B/b \right\rceil 3$、$\left\lceil B/b \right\rceil 4$，将权重矩阵也平均分为四个子权重矩阵，标记为权重部分1-4，分别分配给计算集群301-304，在时刻T0，计算集群301执行神经元矩阵$\left\lceil B/b \right\rceil 1$和权重部分1的矩阵乘法运算，并且计算集群对应的转发器401从计算集群402读取权重部分2，在时刻T1，计算集群301执行神经元矩阵$\left\lceil B/b \right\rceil 1$和权重部分2的矩阵乘法操作，其他计算集群和对应的转发器的行为类似，可参见表2和表3。

[0070] 表2:不同时刻的计算集群行为

[0071]

	计算集群 301	计算集群 302	计算集群 303	计算集群 304
初始状态	存储权重部分 1	存储权重部分 2	存储权重部分 3	存储权重部分 4
时刻 T0	处理$\lceil B/b \rceil$1和权重部分 1	处理$\lceil B/b \rceil$2和权重部分 2	处理$\lceil B/b \rceil$3和权重部分 3	处理$\lceil B/b \rceil$4和权重部分 4
时刻 T1	处理$\lceil B/b \rceil$1和权重部分 2	处理$\lceil B/b \rceil$2和权重部分 3	处理$\lceil B/b \rceil$3和权重部分 4	处理$\lceil B/b \rceil$4和权重部分 1
时刻 T2	……			

[0072] 表3:不同时刻转发器行为

[0073]

	转发器 401	转发器 402	转发器 403	转发器 404
时刻 T0	从计算集群 402 读	从计算集群 403 读	从计算集群 404 读	从计算集群 401 读

[0074]

	取权重部分 2	取权重部分 3	取权重部分 4	取权重部分 1
时刻 T1	从计算集群 402 读取权重部分 3	从计算集群 403 读取权重部分 4	从计算集群 404 读取权重部分 1	从计算集群 401 读取权重部分 2
时刻 T2	从计算集群 402 读取权重部分 4	从计算集群 403 读取权重部分 1	从计算集群 404 读取权重部分 2	从计算集群 401 读取权重部分 3
时刻 T3	……			

[0075] 由表2和表3可以看出,每个计算集群在执行矩阵乘法运算的同时,其对应的转发器可经由后向数据共享通路从其它计算集群读取后续时刻将参加矩阵乘法运算的权重数据,从而使权重数据在各个转发器之间流动,以供计算集群在需要时加载,可加载到数据缓存模块、片上内存等。通过这种方式,能够控制权重数据在计算集群之间流动,从而提高计算集群的资源利用率以及矩阵乘法的运算效率。

[0076] 2)权重规模较大的情况

[0077] 如果M≥b×m,在这种情况下,权重矩阵规模较大,神经元矩阵规模较小,计算集群无法一次完成输入神经元矩阵和权重矩阵的乘法运算。在这种情况下,仍然可对每个计算集群并行分配$\left\lceil B/b \right\rceil$个待计算矩阵,而将权重矩阵切割成若干个较小的矩阵分布于不同的计算集群,例如,计算集群101分配的子矩阵为K×[0,m-1],计算集群102分配的子矩阵为K×[m,2m-1],依此类推。通过这种方式,在执行矩阵乘法操作时,较大的通信带宽,即规模大的权重仍将保持在本地,而神经元数据则可从片上内存(例如SDRAM)中一次读取,并借用计算集群中前向数据转发通路传播至片上系统的其他计算集群。通过这种方式,仅矩阵乘的结果或称中间计算结果通过后向数据共享通路写回至片上共享内存,其他访存均发生在计算集群内部。

[0078] 仍结合图3,下表4和表5分别示意了不同时刻处理器的行为和转发器的行为。具体

地，仍以每个计算集群并行分配$\left\lceil B/b \right\rceil$个输入神经元矩阵为例，分别标记为$\left\lceil B/b \right\rceil 1$、$\left\lceil B/b \right\rceil 2$、$\left\lceil B/b \right\rceil 3$、$\left\lceil B/b \right\rceil 4$，而将一个权重矩阵分割为四个子权重矩阵，标记为权重部分1-4，分别分配给计算集群301-304，在执行神经元矩阵和该一个权重矩阵的乘法运算时，需要对拼接四个子权重矩阵的运算结果。

[0079] 在此示例中，在时刻T0，计算集群301执行神经元矩阵$\left\lceil B/b \right\rceil 1$和权重部分1的矩阵乘法运算，计算集群302执行神经元矩阵$\left\lceil B/b \right\rceil 2$和权重部分2的矩阵乘法运算；在时刻T1，计算集群301执行神经元矩阵$\left\lceil B/b \right\rceil 4$和权重部分1的矩阵乘法运算，计算集群302执行神经元矩阵$\left\lceil B/b \right\rceil 1$和权重部分2的矩阵乘法运算；在时刻T2，计算集群301对应的转发器401从计算集群402读取$\left\lceil B/b \right\rceil 1$与权重部分2的结果，在时刻T3，转发器401从计算集群403读取$\left\lceil B/b \right\rceil 1$与权重部分3的结果，依次类推，当计算集群获得针对一个权重矩阵的多个子权重矩阵的结果之后，通过拼接，即可获得神经元矩阵和权重矩阵的运算结果，其他计算集群和对应的转发器的行为类似，可参见表4和表5。

[0080] 表4：不同时刻处理器行为

[0081]

	计算集群 301	计算集群 302	计算集群 303	计算集群 304
初始状态	存储权重部分 1	存储权重部分 2	存储权重部分 3	存储权重部分 4
时刻 T0	处理$\left\lceil B/b \right\rceil 1$和权重部分 1	处理$\left\lceil B/b \right\rceil 2$和权重部分 2	处理$\left\lceil B/b \right\rceil 3$和权重部分 3	处理$\left\lceil B/b \right\rceil 4$和权重部分 4
时刻 T1	处理$\left\lceil B/b \right\rceil 4$和权重部分 1	处理$\left\lceil B/b \right\rceil 1$和权重部分 2	处理$\left\lceil B/b \right\rceil 2$和权重部分 3	处理$\left\lceil B/b \right\rceil 3$和权重部分 4
时刻 T2	……			

[0082] 表5：不同时刻转发器行为

[0083]

	转发器 401	转发器 402	转发器 403	转发器 404
时刻 T0，T1	无作为	无作为	无作为	无作为
时刻 T2	从计算集群 402	从计算集群 403	从计算集群 404	从计算集群 401

[0084]

	读取 $\lfloor B/b \rfloor 1$ 与权重部分 2 的结果	读取 $\lfloor B/b \rfloor 2$ 与权重部分 3 的结果	读取 $\lfloor B/b \rfloor 3$ 与权重部分 4 的结果	读取 $\lfloor B/b \rfloor 4$ 与权重部分 1 的结果
时刻 T3	从计算集群 403 读取 $\lfloor B/b \rfloor 1$ 与权重部分 3 的结果	从计算集群 404 读取 $\lfloor B/b \rfloor 2$ 与权重部分 4 的结果	从计算集群 401 读取 $\lfloor B/b \rfloor 3$ 与权重部分 1 的结果	从计算集群 402 读取 $\lfloor B/b \rfloor 4$ 与权重部分 2 的结果
时刻 T4	……			

[0085] 由表4和表5可以看出，每个计算集群在执行矩阵乘法运算时，对应的转发器可经由后向数据共享通路从其它计算集群读取前续时刻的矩阵乘法运算的结果，从而使计算结果在各个转发器之间顺序流动，以供计算集群在需要进行拼接。

[0086] 应理解的是，上述权重数据和计算结果在各处理器之间的流动的时序并不是固定的，可根据数据处理的规模、乘累加模块的计算能力以及数据缓存模块和片上内存的容量，控制数据在各个模块之间的传递顺序。例如，每个计算集群的 $\left\lceil B/b \right\rceil$ 个待计算矩阵并不一定是全部处理完再将计算结果通过后向数据共享通路转发，也有可能是处理一部分矩阵后就转发一部分矩阵的计算结果。此外，尽管表4和表5示意的是一个权重矩阵的处理过程，对于多个权重矩阵的情况与此类似，只需依次处理即可。

[0087] 本发明的片上系统，针对神经网络推理应用中不同层的运算特征，提出了一种统一协调的多处理器架构，每个计算集群有自己的较大的存储，能够高效的访问自己的存储，从而解决第一类架构的浅层处理效率较低并且数据搬运能耗较高的问题，同时解决了第二类架构的存储有限带来的问题。此外，通过协调调度处理任务和选择不同的存储策略，本发明能够适用神经网络中不同层的运算特点，从而减轻第三类架构的性能不均衡问题。在软硬件协同方面，本发明基于应用在软件层通过任务划分的方式和非一致性的内存存储策略，使得运算层所需较重的带宽负载集中在本计算集群内部，通过片上互联网络传输较轻的带宽负载，从而实现局部访存的能耗优化。

[0088] 本发明在人工智能推理领域提高了计算能效比，尤其适合在高性能推理需求，例如，数据中心、无人驾驶等应用场景。

[0089] 本发明的片上系统可应用了各种电子设备，例如、移动设备、嵌入式电子设备、智能计算处理设备、机器人等，可应用于文字处理、语音识别与处理、多国语言翻译、图像识别、生物特征识到、智能控制等领域。

[0090] 需要说明的是，虽然上文按照特定顺序描述了各个步骤，但是并不意味着必须按照上述特定顺序来执行各个步骤，实际上，这些步骤中的一些可以并发执行，甚至改变顺

序,只要能够实现所需要的功能即可。

[0091] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0092] 计算机可读存储介质可以是保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以包括但不限于电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。

[0093] 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。

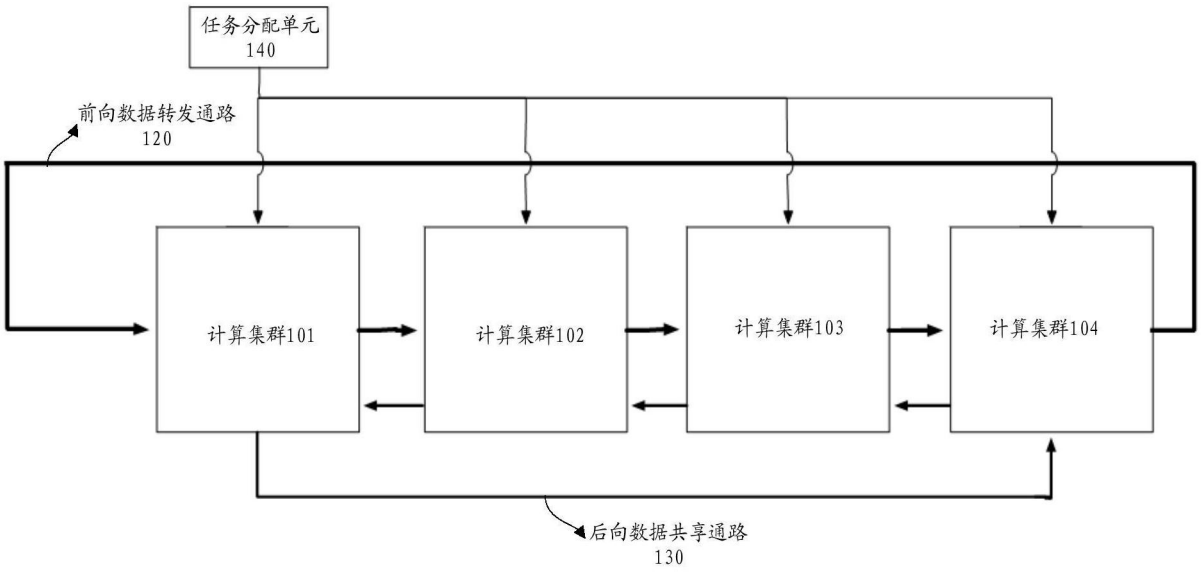

任务分配单元
140
前向数据转发通路
120
计算集群101
计算集群102
计算集群103
计算集群104
后向数据共享通路
130

图1

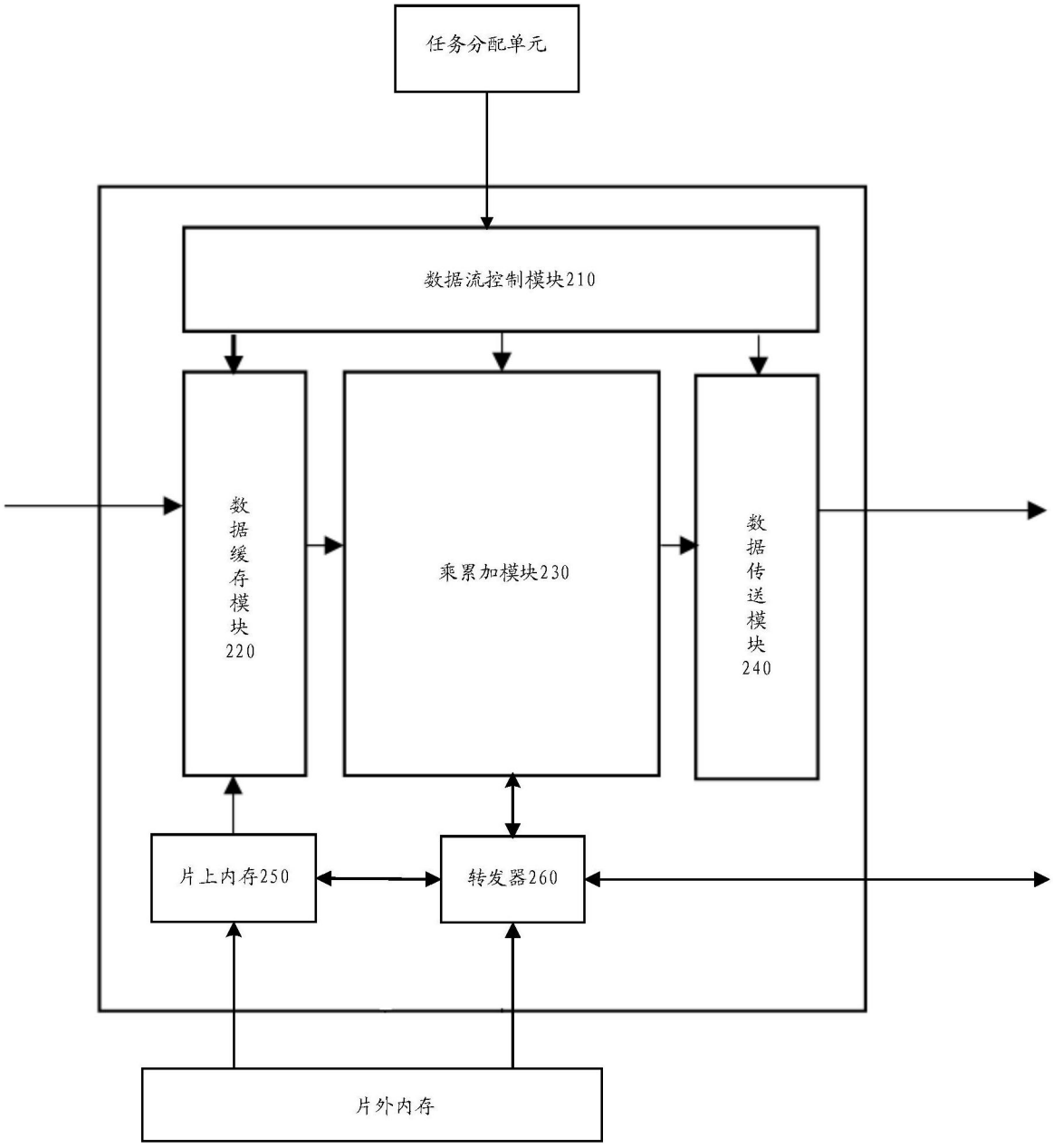

任务分配单元
数据流控制模块210
数据缓存模块220
乘累加模块230
数据传送模块240
片上内存250
转发器260
片外内存

图2

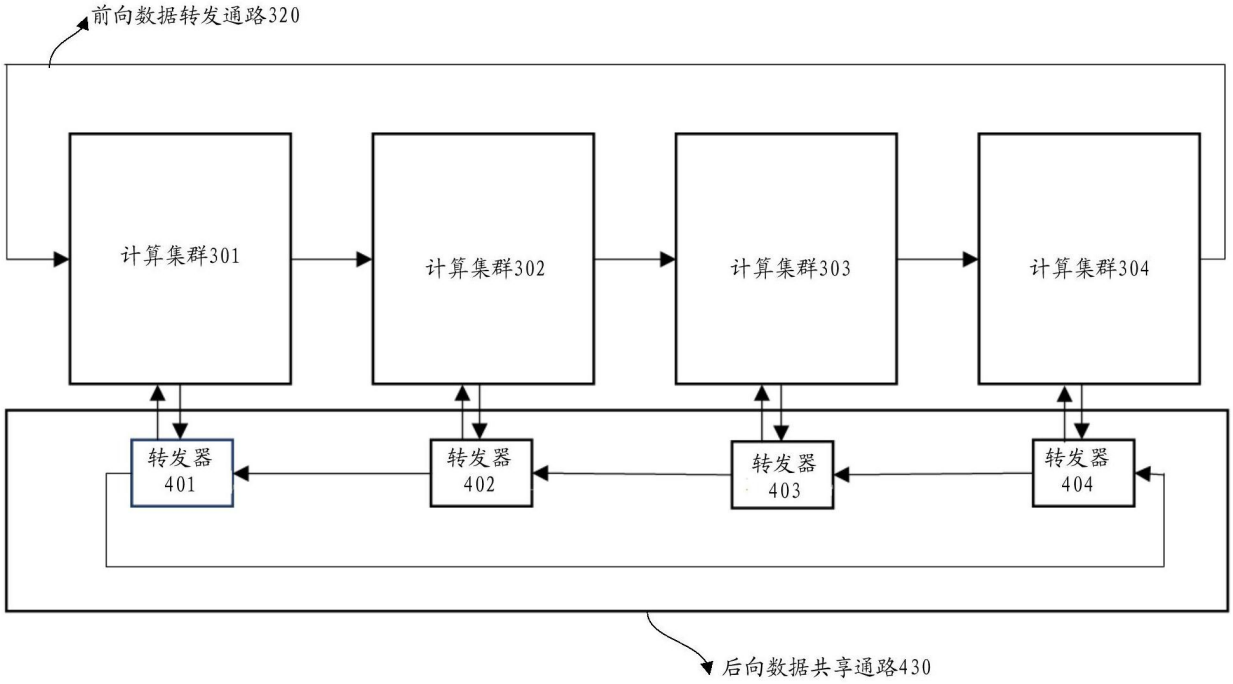

前向数据转发通路320
计算集群301
计算集群302
计算集群303
计算集群304
转发器 401
转发器 402
转发器 403
转发器 404
后向数据共享通路430

图3