(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113254070 B
(45)授权公告日 2024.01.02

(21)申请号 202010082643.0

(22)申请日 2020.02.07

(65)同一申请的已公布的文献号
申请公布号 CN 113254070 A

(43)申请公布日 2021.08.13

(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号

(72)发明人 何军 尹莉 吴雪君

(74)专利代理机构 北京成创同维知识产权代理有限公司 11449
专利代理师 李镇江

(51)Int.Cl.
G06F 9/30(2006.01)
G06F 9/50(2006.01)
G06F 15/163(2006.01)

(56)对比文件
CN 101593097 A,2009.12.02
CN 104781803 A,2015.07.15
US 2010153681 A1,2010.06.17
US 2012110559 A1,2012.05.03
US 2019294570 A1,2019.09.26
苏文;王焕东;台运方;王靖.面向云计算的多核处理器存储和网络子系统优化设计.高技术通讯.2013,(第04期),全文.

审查员 王敏

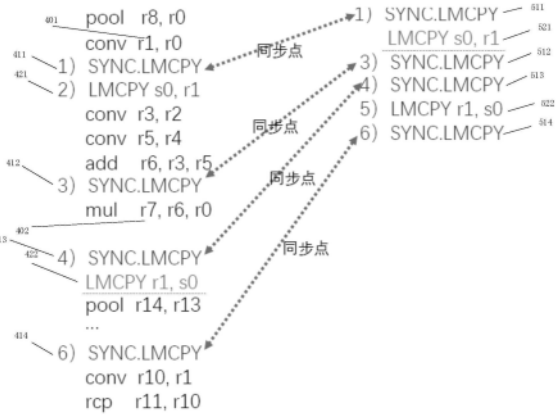
权利要求书5页 说明书22页 附图10页

(54)发明名称

加速单元、片上系统、服务器、数据中心和相关方法

(57)摘要

本公开提供了一种加速单元、片上系统、服务器、数据中心和相关方法。所述加速单元包括主核和副核,其中,所述主核包括:第一片上内存;主核定序器,用于对接收的第一跨核复制指令进行译码;主核存储器复制引擎,用于从第一片上内存的第一地址获取第一操作数,并将获取的第一操作数复制到副核的第二片上内存中的第二地址;所述副核包括:所述第二片上内存;副核定序器,用于对接收的第二跨核复制指令进行译码;副核存储器复制引擎,用于从第二片上内存的第二地址获取第一操作数,并将获取的第一操作数复制回第一片上内存中的第一地址。本公开实施例在片上内存中的操作数需要转移的情况下,提高操作数转移效率,提高加速单元性能。



CN 113254070 B

1.一种加速单元,包括加速单元主核和加速单元副核,其中,

所述加速单元主核包括:

第一片上内存;

主核定序器,用于对接收的第一跨核复制指令进行译码,所述第一跨核复制指令指示将第一操作数从所述第一片上内存上的第一地址复制到所述加速单元副核的第二片上内存的第二地址;

主核存储器复制引擎,用于接收并执行译码后的第一跨核复制指令,从而从所述第一片上内存的所述第一地址获取第一操作数,并将获取的第一操作数复制到所述第二片上内存中的所述第二地址;

所述加速单元副核包括:

所述第二片上内存;

副核定序器,用于对接收的第二跨核复制指令进行译码,所述第二跨核复制指令指示将第一操作数从所述第二片上内存上的第二地址复制回所述第一片上内存的第一地址;

副核存储器复制引擎,用于接收并执行译码后的第二跨核复制指令,从而从所述第二片上内存的所述第二地址获取第一操作数,并将获取的第一操作数复制回所述第一片上内存中的所述第一地址。

2.根据权利要求1所述的加速单元,其中,

所述加速单元主核还包括第一寄存器和第二寄存器,分别用于存放所述第一地址和所述第二地址,所述第一跨核复制指令指示将第一寄存器中的第一地址作为跨核复制的源地址,将第二寄存器中的第二地址作为跨核复制的目的地地址,从而使所述主核存储器复制引擎将所述第一地址中的第一操作数转移到所述第二地址;

所述加速单元副核还包括第三寄存器和第四寄存器,分别用于存放所述第二地址和所述第一地址,所述第二跨核复制指令指示将第三寄存器中的第二地址作为跨核复制的源地址,将第四寄存器中的第一地址作为跨核复制的目的地地址,从而使所述副核存储器复制引擎将所述第二地址中的第一操作数转移回所述第一地址。

3.根据权利要求1所述的加速单元,其中,所述加速单元主核还包括主核指令缓存器,用于接收并缓存所述第一跨核复制指令并提供给所述主核定序器;所述加速单元副核还包括副核指令缓存器,用于接收并缓存所述第二跨核复制指令并提供给所述副核定序器。

4.根据权利要求3所述的加速单元,其中,

所述主核指令缓存器用于在接收并缓存所述第一跨核复制指令之前,接收并缓存第一主核同步原语指令,用于所述主核与所述副核的同步;所述主核定序器用于对缓存的所述第一主核同步原语指令进行译码,并将译码后的第一主核同步原语指令发送给所述加速单元主核外的命令处理器;

所述副核指令缓存器用于接收并缓存与所述第一主核同步原语指令对应的第一副核同步原语指令、与所述第一跨核复制指令对应的第一哑跨核复制指令;所述副核定序器用于对缓存的所述第一副核同步原语指令、和第一哑跨核复制指令进行译码,并将译码后的第一副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第一主核同步原语指令、和所述第一副核同步原语指令,进行所述主核与所述副核的同步,并忽略译码后的第一哑跨核复制指令。

5.根据权利要求4所述的加速单元,其中,

所述主核指令缓存器是在接收并缓存所述第一操作数在待执行指令序列中第二操作数出现之前最后一次被使用的指令之后,接收并缓存第一主核同步原语指令的,其中,该第二操作数是所述待执行指令序列执行时会引起所述第一片上内存溢出的操作数。

6.根据权利要求5所述的加速单元,其中,

所述主核指令缓存器在接收并缓存所述第二操作数在待执行指令序列中第一次出现的指令之前,接收并缓存第二主核同步原语指令;所述主核定序器用于对缓存的所述第二主核同步原语指令进行译码,并将译码后的第二主核同步原语指令发送给所述加速单元主核外的命令处理器;

所述副核指令缓存器在接收并缓存第一哑跨核复制指令之后,接收并缓存与所述第二主核同步原语指令对应的第二副核同步原语指令;所述副核定序器用于对缓存的所述第二副核同步原语指令进行译码,并将译码后的第二副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第二主核同步原语指令和所述第二副核同步原语指令,进行所述主核与所述副核的同步。

7.根据权利要求6所述的加速单元,其中,

所述主核指令缓存器在接收并缓存所述第二操作数在待执行指令序列中最后一次出现的指令之后,顺序接收第三主核同步原语指令和与所述第二跨核复制指令对应的第二哑跨核复制指令;所述主核定序器用于对缓存的所述第三主核同步原语指令和第二哑跨核复制指令进行译码,并将译码后的第三主核同步原语指令发送给所述加速单元主核外的命令处理器,并忽略译码后的第二哑跨核复制指令;

所述副核指令缓存器在接收并缓存第二跨核复制指令之前、接收并缓存第二副核同步原语指令之后,接收并缓存第三副核同步原语指令;所述副核定序器用于对缓存的所述第三副核同步原语指令进行译码,并将译码后的第三副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第三主核同步原语指令、和所述第三副核同步原语指令,进行所述主核与所述副核的同步。

8.根据权利要求7所述的加速单元,其中,

所述主核指令缓存器在接收并缓存所述第一操作数在所述待执行指令序列中所述第二操作数最后一次出现之后第一次被使用的指令之前,接收并缓存第四主核同步原语指令;所述主核定序器用于对缓存的所述第四主核同步原语指令进行译码,并将译码后的第四主核同步原语指令发送给所述加速单元主核外的命令处理器;

所述副核指令缓存器在接收并缓存第二跨核复制指令之后,接收并缓存与所述第四主核同步原语指令对应的第四副核同步原语指令;所述副核定序器用于对缓存的所述第四副核同步原语指令进行译码,并将译码后的第四副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第四主核同步原语指令和所述第四副核同步原语指令,进行所述主核与所述副核的同步。

9.根据权利要求1所述的加速单元,其中,所述加速单元副核为多个,第二寄存器中的第二地址是多个加速单元副核中选定加速单元副核的第二片上内存上的第二地址。

10.根据权利要求9所述的加速单元,其中,所述选定加速单元副核是根据所述多个加速单元副核的每一个到所述加速单元主核的距离,从所述多个加速单元副核中选出的。

11.一种加速单元,包括加速单元主核、第一加速单元副核和第二加速单元副核,其中,

所述加速单元主核包括:

第一片上内存;

主核定序器,用于对接收的第一跨核复制指令进行译码,所述第一跨核复制指令指示将待转移操作数集合的第一部分从所述第一片上内存上的第一地址集合复制到所述第一加速单元副核的第二片上内存的第二地址集合,将待转移操作数集合的第二部分从所述第一片上内存上的第三地址集合复制到所述第二加速单元副核的第三片上内存的第四地址集合;

主核存储器复制引擎,用于接收并执行译码后的第一跨核复制指令,从而从所述第一片上内存的所述第一地址集合获取所述第一部分,并将获取的所述第一部分复制到所述第一加速单元副核的第二片上内存中的所述第二地址集合,且从所述第一片上内存的所述第三地址集合获取所述第二部分,并将获取的所述第二部分复制到所述第二加速单元副核的第三片上内存中的所述第四地址集合;

所述第一加速单元副核包括:

所述第二片上内存;

第一副核定序器,用于对接收的第二跨核复制指令进行译码,所述第二跨核复制指令指示将所述第一部分从所述第二片上内存的第二地址集合复制回所述第一片上内存上的第一地址集合;

第一副核存储器复制引擎,用于接收并执行译码后的第二跨核复制指令,从而从所述第二片上内存的所述第二地址集合获取所述第一部分,并将获取的第一部分复制回所述第一片上内存中的所述第一地址集合;

所述第二加速单元副核包括:

所述第三片上内存;

第二副核定序器,用于对接收的第三跨核复制指令进行译码,所述第三跨核复制指令指示将所述第二部分从所述第三片上内存的第四地址集合复制回所述第一片上内存上的第三地址集合;

第二副核存储器复制引擎,用于接收并执行译码后的第三跨核复制指令,从而从所述第三片上内存的所述第四地址集合获取所述第二部分,并将获取的第二部分复制回所述第一片上内存中的所述第三地址集合。

12.根据权利要求11所述的加速单元,其中,所述加速单元主核还包括:第一首地址寄存器和第一尾地址寄存器,分别用于存放所述第一地址集合的首地址和尾地址;第二首地址寄存器和第二尾地址寄存器,分别用于存放所述第二地址集合的首地址和尾地址;第三首地址寄存器和第三尾地址寄存器,分别用于存放所述第三地址集合的首地址和尾地址;第四首地址寄存器和第四尾地址寄存器,分别用于存放所述第四地址集合的首地址和尾地址;

所述第一跨核复制指令指示将第一片上内存上的、第一首地址寄存器中的首地址到第一尾地址寄存器中的尾地址之间的第一部分取出,复制到第二片上内存上的、第二首地址寄存器中的首地址到第二尾地址寄存器中的尾地址之间,且将第一片上内存上的、第三首地址寄存器中的首地址到第三尾地址寄存器中的尾地址之间的第二部分取出,复制到第三

片上内存上的、第四首地址寄存器中的首地址到第四尾地址寄存器中的尾地址之间;

所述第一加速单元副核还包括:第五首地址寄存器和第五尾地址寄存器,分别用于存放所述第二地址集合的首地址和尾地址;第六首地址寄存器和第六尾地址寄存器,分别用于存放所述第一地址集合的首地址和尾地址;

所述第二跨核复制指令指示将第二片上内存上的、第五首地址寄存器中的首地址到第五尾地址寄存器中的尾地址之间的第一部分取出,复制回第一片上内存上的、第六首地址寄存器中的首地址到第六尾地址寄存器中的尾地址之间;

所述第二加速单元副核还包括:第七首地址寄存器和第七尾地址寄存器,分别用于存放所述第四地址集合的首地址和尾地址;第八首地址寄存器和第八尾地址寄存器,分别用于存放所述第三地址集合的首地址和尾地址;

所述第三跨核复制指令指示将第三片上内存上的、第七首地址寄存器中的首地址到第七尾地址寄存器中的尾地址之间的第二部分取出,复制回第一片上内存上的、第八首地址寄存器中的首地址到第八尾地址寄存器中的尾地址之间。

13.根据权利要求11所述的加速单元,其中,所述加速单元主核还包括主核指令缓存器,用于接收并缓存所述第一跨核复制指令并提供给所述主核定序器;所述第一加速单元副核还包括第一副核指令缓存器,用于接收并缓存所述第二跨核复制指令并提供给所述第一副核定序器;所述第二加速单元副核还包括第二副核指令缓存器,用于接收并缓存所述第三跨核复制指令并提供给所述第二副核定序器。

14.一种片上系统(SoC),包含根据权利要求1-13中任一个所述的加速单元。

15.一种服务器,包括:

根据权利要求1-13中任一个所述的加速单元;

存储器,存储有计算机可执行指令;

调度单元,在执行存储器中存储的计算机可执行指令的过程中,确定要由所述加速单元执行的待执行指令序列,并将所述待执行指令序列分配给所述加速单元。

16.一种数据中心,包括根据权利要求15所述的服务器。

17.一种加速单元跨核复制方法,所述加速单元包括加速单元主核与加速单元副核,所述方法包括:

利用所述加速单元主核,对接收的第一跨核复制指令进行译码并执行,从而从所述加速单元主核上的第一片上内存的第一地址获取第一操作数,并将获取的第一操作数复制到所述加速单元副核的第二片上内存中的第二地址;

利用所述加速单元副核,对接收的第二跨核复制指令进行译码并执行,从而从所述加速单元副核上的第二片上内存的第二地址获取第一操作数,并将获取的第一操作数复制回所述加速单元主核的第一片上内存中的第一地址。

18.一种加速单元跨核复制方法,所述加速单元包括加速单元主核、第一加速单元副核、第二加速单元副核,所述加速单元主核包括第一片上内存,所述第一加速单元副核包括第二片上内存,所述第二加速单元副核包括第三片上内存,所述方法包括:

利用所述加速单元主核,对接收的第一跨核复制指令进行译码并执行,从而从所述第一片上内存的所述第一地址集合获取待转移操作数集合的所述第一部分,并将获取的所述第一部分复制到所述第一加速单元副核的第二片上内存中的所述第二地址集合,且从所述

第一片上内存的所述第三地址集合获取所述待转移操作数集合的第二部分,并将获取的所述第二部分复制到所述第二加速单元副核的第三片上内存中的所述第四地址集合;

利用所述第一加速单元副核,对接收的第二跨核复制指令进行译码并执行,从而从所述第二片上内存的所述第二地址集合获取所述第一部分,并将获取的第一部分复制回所述第一片上内存中的所述第一地址集合;

利用所述第二加速单元副核,对接收的第三跨核复制指令进行译码并执行,从而从所述第三片上内存的所述第四地址集合获取所述第二部分,并将获取的第二部分复制回所述第一片上内存中的所述第三地址集合。

加速单元、片上系统、服务器、数据中心和相关方法

技术领域

[0001] 本公开涉及芯片领域,更具体而言,涉及一种加速单元、片上系统、服务器、数据中心和相关方法。

背景技术

[0002] 在大数据时代,神经网络得到了广泛的应用。神经网络中,各神经网络节点的大量运算(例如卷积、池化等)利用传统的CPU处理非常低效,因此开发了专门的加速单元,专门为人工智能神经网络而设计,用于加速神经网络的运算,解决传统芯片在神经网络运算时效率低下的问题。这些加速单元内部往往有多个核,每个核分别有片上内存。每个核可以并行执行同一个模型任务,这样就具有完全一样的指令和初始化权重数据,一次可以同时执行多笔的推理任务。每个核也可以具有不同的计算指令序列、初始化权重数据和输入,以执行不同的模型任务。各个核之间分工合作,大大提高了加速单元的处理能力。

[0003] 在传统的加速单元核的片上内存分配算法中,由于片上内存有限,不可能让所有操作数都驻在片上内存。这种情况下,一般会将片上内存放不下的操作数放到核外的多个核共享的存储器中,等需要的时候读回片上内存。从核内向核外共享存储器转移数据的效率很低,造成现有加速单元性能下降。

发明内容

[0004] 有鉴于此,本公开实施例旨在在核内片上内存中的操作数需要转移的情况下,提高操作数转移效率,提高加速单元性能。

[0005] 为了达到该目的,根据本公开的一方面,提供了一种加速单元,包括加速单元主核和加速单元副核,其中,所述加速单元主核包括:第一片上内存;主核定序器,用于对接收的第一跨核复制指令进行译码,所述第一跨核复制指令指示将第一操作数从所述第一片上内存上的第一地址复制到所述加速单元副核的第二片上内存的第二地址;主核存储器复制引擎,用于接收并执行译码后的第一跨核复制指令,从而从所述第一片上内存的所述第一地址获取第一操作数,并将获取的第一操作数复制到所述第二片上内存中的所述第二地址;所述加速单元副核包括:所述第二片上内存;副核定序器,用于对接收的第二跨核复制指令进行译码,所述第二跨核复制指令指示将第一操作数从所述第二片上内存上的第二地址复制回所述第一片上内存的第一地址;副核存储器复制引擎,用于接收并执行译码后的第二跨核复制指令,从而从所述第二片上内存的所述第二地址获取第一操作数,并将获取的第一操作数复制回所述第一片上内存中的所述第一地址。

[0006] 可选地,所述加速单元主核还包括第一寄存器和第二寄存器,分别用于存放所述第一地址和所述第二地址,所述第一跨核复制指令指示将第一寄存器中的第一地址作为跨核复制的源地址,将第二寄存器中的第二地址作为跨核复制的目的地地址,从而使所述主核存储器复制引擎将所述第一地址中的第一操作数转移到所述第二地址;所述加速单元副核还包括第三寄存器和第四寄存器,分别用于存放所述第二地址和所述第一地址,所述第

二跨核复制指令指示将第三寄存器中的第二地址作为跨核复制的源地址,将第四寄存器中的第一地址作为跨核复制的目的地地址,从而使所述副核存储器复制引擎将所述第二地址中的第一操作数转移回所述第一地址。

[0007] 可选地,所述加速单元主核还包括主核指令缓存器,用于接收并缓存所述第一跨核复制指令并提供给所述主核定序器;所述加速单元副核还包括副核指令缓存器,用于接收并缓存所述第二跨核复制指令并提供给所述副核定序器。

[0008] 可选地,所述主核指令缓存器用于在接收并缓存所述第一跨核复制指令之前,接收并缓存第一主核同步原语指令,用于所述主核与所述副核的同步;所述主核定序器用于对缓存的所述第一主核同步原语指令进行译码,并将译码后的第一主核同步原语指令发送给所述加速单元主核外的命令处理器;所述副核指令缓存器用于接收并缓存与所述第一主核同步原语指令对应的第一副核同步原语指令、与所述第一跨核复制指令对应的第一哑跨核复制指令;所述副核定序器用于对缓存的所述第一副核同步原语指令、和第一哑跨核复制指令进行译码,并将译码后的第一副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第一主核同步原语指令、和所述第一副核同步原语指令,进行所述主核与所述副核的同步,并忽略译码后的第一哑跨核复制指令。

[0009] 可选地,所述主核指令缓存器是在接收并缓存所述第一操作数在待执行指令序列中第二操作数出现之前最后一次被使用的指令之后,接收并缓存第一主核同步原语指令的,其中,该第二操作数是所述待执行指令序列执行时会引起所述第一片上内存溢出的操作数。

[0010] 可选地,所述主核指令缓存器在接收并缓存所述第二操作数在待执行指令序列中第一次出现的指令之前,接收并缓存第二主核同步原语指令;所述主核定序器用于对缓存的所述第二主核同步原语指令进行译码,并将译码后的第二主核同步原语指令发送给所述加速单元主核外的命令处理器;所述副核指令缓存器在接收并缓存第一哑跨核复制指令之后,接收并缓存与所述第二主核同步原语指令对应的第二副核同步原语指令;所述副核定序器用于对缓存的所述第二副核同步原语指令进行译码,并将译码后的第二副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第二主核同步原语指令和所述第二副核同步原语指令,进行所述主核与所述副核的同步。

[0011] 可选地,所述主核指令缓存器在接收并缓存所述第二操作数在待执行指令序列中最后一次出现的指令之后,顺序接收第三主核同步原语指令和与所述第二跨核复制指令对应的第二哑跨核复制指令;所述主核定序器用于对缓存的所述第三主核同步原语指令和第二哑跨核复制指令进行译码,并将译码后的第三主核同步原语指令发送给所述加速单元主核外的命令处理器,并忽略译码后的第二哑跨核复制指令;所述副核指令缓存器在接收并缓存第二跨核复制指令之前、接收并缓存第二副核同步原语指令之后,接收并缓存第三副核同步原语指令;所述副核定序器用于对缓存的所述第三副核同步原语指令进行译码,并将译码后的第三副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第三主核同步原语指令、和所述第三副核同步原语指令,进行所述主核与所述副核的同步。

[0012] 可选地,所述主核指令缓存器在接收并缓存所述第一操作数在所述待执行指令序列中所述第二操作数最后一次出现之后第一次被使用的指令之前,接收并缓存第四主核同步原语指令;所述主核定序器用于对缓存的所述第四主核同步原语指令进行译码,并将译

码后的第四主核同步原语指令发送给所述加速单元主核外的命令处理器;所述副核指令缓存器在接收并缓存第二跨核复制指令之后,接收并缓存与所述第四主核同步原语指令对应的第四副核同步原语指令;所述副核定序器用于对缓存的所述第四副核同步原语指令进行译码,并将译码后的第四副核同步原语指令发送给所述加速单元副核外的命令处理器按照所述第四主核同步原语指令和所述第四副核同步原语指令,进行所述主核与所述副核的同步。

[0013] 可选地,所述加速单元副核为多个,第二寄存器中的第二地址是多个加速单元副核中选定加速单元副核的第二片上内存上的第二地址。

[0014] 可选地,所述选定加速单元副核是根据所述多个加速单元副核的每一个到所述加速单元主核的距离,从所述多个加速单元副核中选出的。

[0015] 根据本公开的一方面,提供了一种加速单元,包括加速单元主核、第一加速单元副核和第二加速单元副核,其中,所述加速单元主核包括:第一片上内存;主核定序器,用于对接收的第一跨核复制指令进行译码,所述第一跨核复制指令指示将待转移操作数集合的第一部分从所述第一片上内存上的第一地址集合复制到所述第一加速单元副核的第二片上内存的第二地址集合,将待转移操作数集合的第二部分从所述第一片上内存上的第三地址集合复制到所述第二加速单元副核的第三片上内存的第四地址集合;主核存储器复制引擎,用于接收并执行译码后的第一跨核复制指令,从而从所述第一片上内存的所述第一地址集合获取所述第一部分,并将获取的所述第一部分复制到所述第一加速单元副核的第二片上内存中的所述第二地址集合,且从所述第一片上内存的所述第三地址集合获取所述第二部分,并将获取的所述第二部分复制到所述第二加速单元副核的第三片上内存中的所述第四地址集合;所述第一加速单元副核包括:所述第二片上内存;第一副核定序器,用于对接收的第二跨核复制指令进行译码,所述第二跨核复制指令指示将所述第一部分从所述第二片上内存的第二地址集合复制回所述第一片上内存上的第一地址集合;第一副核存储器复制引擎,用于接收并执行译码后的第二跨核复制指令,从而从所述第二片上内存的所述第二地址集合获取所述第一部分,并将获取的第一部分复制回所述第一片上内存中的所述第一地址集合;所述第二加速单元副核包括:所述第三片上内存;第二副核定序器,用于对接收的第三跨核复制指令进行译码,所述第三跨核复制指令指示将所述第二部分从所述第三片上内存的第四地址集合复制回所述第一片上内存上的第三地址集合;第二副核存储器复制引擎,用于接收并执行译码后的第三跨核复制指令,从而从所述第三片上内存的所述第四地址集合获取所述第二部分,并将获取的第二部分复制回所述第一片上内存中的所述第三地址集合。

[0016] 可选地,所述加速单元主核还包括:第一首地址寄存器和第一尾地址寄存器,分别用于存放所述第一地址集合的首地址和尾地址;第二首地址寄存器和第二尾地址寄存器,分别用于存放所述第二地址集合的首地址和尾地址;第三首地址寄存器和第三尾地址寄存器,分别用于存放所述第三地址集合的首地址和尾地址;第四首地址寄存器和第四尾地址寄存器,分别用于存放所述第四地址集合的首地址和尾地址;所述第一跨核复制指令指示将第一片上内存上的、第一首地址寄存器中的首地址到第一尾地址寄存器中的尾地址之间的第一部分取出,复制到第二片上内存上的、第二首地址寄存器中的首地址到第二尾地址寄存器中的尾地址之间,且将第一片上内存上的、第三首地址寄存器中的首地址到第三尾

地址寄存器中的尾地址之间的第二部分取出，复制到第三片上内存上的、第四首地址寄存器中的首地址到第四尾地址寄存器中的尾地址之间；所述第一加速单元副核还包括：第五首地址寄存器和第五尾地址寄存器，分别用于存放所述第二地址集合的首地址和尾地址；第六首地址寄存器和第六尾地址寄存器，分别用于存放所述第一地址集合的首地址和尾地址；所述第二跨核复制指令指示将第二片上内存上的、第五首地址寄存器中的首地址到第五尾地址寄存器中的尾地址之间的第一部分取出，复制回第一片上内存上的、第六首地址寄存器中的首地址到第六尾地址寄存器中的尾地址之间；所述第二加速单元副核还包括：第七首地址寄存器和第七尾地址寄存器，分别用于存放所述第四地址集合的首地址和尾地址；第八首地址寄存器和第八尾地址寄存器，分别用于存放所述第三地址集合的首地址和尾地址；所述第三跨核复制指令指示将第三片上内存上的、第七首地址寄存器中的首地址到第七尾地址寄存器中的尾地址之间的第二部分取出，复制回第一片上内存上的、第八首地址寄存器中的首地址到第八尾地址寄存器中的尾地址之间。

[0017] 可选地，所述加速单元主核还包括主核指令缓存器，用于接收并缓存所述第一跨核复制指令并提供给所述主核定序器；所述第一加速单元副核还包括第一副核指令缓存器，用于接收并缓存所述第二跨核复制指令并提供给所述第一副核定序器；所述第二加速单元副核还包括第二副核指令缓存器，用于接收并缓存所述第三跨核复制指令并提供给所述第二副核定序器。

[0018] 根据本公开的另一方面，提供了一种片上系统，包含根据如上所述的加速单元。

[0019] 根据本公开的另一方面，提供了一种服务器，包括：如上所述的加速单元；存储器，存储有计算机可执行指令；调度单元，在执行存储器中存储的计算机可执行指令的过程中，确定要由所述加速单元执行的待执行指令序列，并将所述待执行指令序列分配给所述加速单元。

[0020] 根据本公开的另一方面，提供了一种数据中心，包括如上所述的服务器。

[0021] 根据本公开的另一方面，提供了一种加速单元跨核复制方法，所述加速单元包括加速单元主核与加速单元副核，所述方法包括：利用所述加速单元主核，对接收的第一跨核复制指令进行译码并执行，从而从所述加速单元主核上的第一片上内存的第一地址获取第一操作数，并将获取的第一操作数复制到所述加速单元副核的第二片上内存中的第二地址；利用所述加速单元副核，对接收的第二跨核复制指令进行译码并执行，从而从所述加速单元副核上的第二片上内存的第二地址获取第一操作数，并将获取的第一操作数复制回所述加速单元主核的第一片上内存中的第一地址。

[0022] 根据本公开的一方面，提供了一种加速单元跨核复制方法，所述加速单元包括加速单元主核、第一加速单元副核、第二加速单元副核，所述方法包括：利用所述加速单元主核，对接收的第一跨核复制指令进行译码并执行，从而从所述第一片上内存的所述第一地址集合获取所述第一部分，并将获取的所述第一部分复制到所述第一加速单元副核的第二片上内存中的所述第二地址集合，且从所述第一片上内存的所述第三地址集合获取所述第二部分，并将获取的所述第二部分复制到所述第二加速单元副核的第三片上内存中的所述第四地址集合；利用所述第一加速单元副核，对接收的第二跨核复制指令进行译码并执行，从而从所述第二片上内存的所述第二地址集合获取所述第一部分，并将获取的第一部分复制回所述第一片上内存中的所述第一地址集合；利用所述第二加速单元副核，对接收的第

三跨核复制指令进行译码并执行,从而从所述第三片上内存的所述第四地址集合获取所述第二部分,并将获取的第二部分复制回所述第一片上内存中的所述第三地址集合。

[0023] 本公开实施例中,在待执行指令序列的适当位置加入第一跨核复制指令,加速单元主核接收到第一跨核复制指令后,对接收的第一跨核复制指令进行译码并执行,从而从所述加速单元主核上的第一片上内存的第一地址获取第一操作数,并将获取的第一操作数复制到所述加速单元副核的第二片上内存中的第二地址,从而完成了可能会溢出的第一操作数向加速单元副核的转移。另外,分配第二跨核复制指令给加速单元副核执行。在需要从加速单元副核取回第一操作数时,加速单元副核对接收的第二跨核复制指令进行译码并执行,从而从所述加速单元副核上的第二片上内存的第二地址获取第一操作数,并将获取的第一操作数复制回所述加速单元主核的第一片上内存中的第一地址。通过上述过程,实现了当加速单元主核中第一片上内存上存储的操作数可能溢出时,将要溢出的操作数转移到加速单元副核,并在需要时及时从加速单元副核取回的目的。相对于现有技术将片上内存的数据复制到片外共享存储器的方式,本公开实施例由于在片上内存间移动数据的效率要高于片上内存的数据移动到片外共享存储器的效率,提高了加速单元的性能。

附图说明

[0024] 通过参考以下附图对本公开实施例的描述,本公开的上述以及其它目的、特征和优点将更为清楚,在附图中:

[0025] 图1是本公开一个实施例所应用的数据中心的结构图;

[0026] 图2是本公开一个实施例的数据中心中一个服务器的内部结构图;

[0027] 图3是根据本公开一个实施例服务器内部的调度单元和加速单元的连接关系图;

[0028] 图4是根据本公开一个实施例的加速单元核的内部结构图;

[0029] 图5是根据本公开一个实施例的在主核和副核之间执行数据复制的示意图;

[0030] 图6A示出了本公开一个实施例中的待执行指令序列;

[0031] 图6B示出了现有技术中由于执行时会产生溢出因而要将一部分操作数移到共享存储器而编译时增加的指令;

[0032] 图6C示出了本公开实施例中在待执行指令序列中加入的跨核复制时主核用到的指令、和副核用到的指令;

[0033] 图7A-B示出了根据本公开实施例的跨核复制方法的交互流程图;

[0034] 图8示出了根据本公开一个实施例的完成协同跨核复制的加速单元结构图。

具体实施方式

[0035] 以下基于实施例对本公开进行描述,但是本公开并不仅仅限于这些实施例。在下文对本公开的细节描述中,详尽描述了一些特定的细节部分。对本领域技术人员来说没有这些细节部分的描述也可以完全理解本公开。为了避免混淆本公开的实质,公知的方法、过程、流程没有详细叙述。另外附图不一定是按比例绘制的。

[0036] 在本文中使用以下术语。

[0037] 神经网络:一般指人工神经网络(Artificial Neural Network,简写为ANN),它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。

[0038] 加速单元:针对传统处理单元在一些专门用途的领域(例如,处理图像、处理神经网络的各种运算,等等)效率不高的情况,为了提高在这些专门用途领域中的数据处理速度而设计的处理单元,在本公开实施例中主要是为了加速神经网络模型的运算处理速度而设计的专门处理单元。

[0039] 调度单元:对加速单元进行调度、向各加速单元分配要执行的待执行指令序列的处理单元,它可以采用中央处理单元(CPU)、专用集成电路(ASIC)、现场可编程门阵列(FPGA)等多种形式。

[0040] 主核:加速单元内部分配用于执行待执行指令序列,从而完成一系列实体操作或运算的核。

[0041] 副核:加速单元内部不用来完成实体操作或运算,只与主核相配合,在主核的片上内存存储不下指令序列运行时所需的操作数时,为主核暂存存储不下的操作数,等到主核的片上内存能够容纳时再复制回主核的片上内存的核。注意,这里的主核和副核只是为了描述方便,其区分不是绝对的。对于某个待执行指令序列来说,分配用来执行该执行指令序列的核对于该执行指令序列来说就是主核,为主核暂时容纳操作数的核就是副核,但对于另外的待执行指令序列,可能刚好相反。

[0042] 片上内存:在主核或副核内单独使用,不能被共享的存储器。

[0043] 命令处理器:在加速单元和驱动该加速单元工作的调度单元之间的命令接口。命令处理器接收调度单元让加速单元执行的指令,将这些指令分给加速单元中的各个核去执行。另外,它还负责加速单元中各个核的同步。

[0044] 跨核复制:主核中的片上内存存储不下指令序列执行过程中用到的操作数,因而需要将一部分操作数先转移到副核的片上内存,等到需要时再转移回来的行为。

[0045] 跨核复制指令:本公开实施例中能将操作数从一个核的片上内存复制到另一个核的片上内存的指令,一般采取LMCPY X1,X2的形式,其中X1是要复制到的目的地地址,X2是要复制的操作数的源地址。

[0046] 操作数:操作数是运算符作用于的实体,是表达式中的一个组成部分,它规定了指令中进行数字运算的量。

[0047] 同步:确保核之间执行步调一致的行为。两个核执行指令的进度不一致,可能一个核执行完某一指令时另一个核正在执行另一指令,同步的意思就是等待另一个核执行完后,两个核一起开始执行各自后面的指令。

[0048] 同步原语指令:本公开实施例中一种用于核间同步的原语指令,一般写作SYNC.LMCPY,而且配对使用,即在两个同步核的执行的指令序列需要同步的位置都加一个相同的同步原语指令,用于在该位置两核同步。

[0049] 数据中心

[0050] 数据中心是全球协作的特定设备网络,用来在互联网网络基础设施上传递、加速、展示、计算、存储数据信息。在今后的发展中,数据中心也将会成为企业竞争的资产。随着数据中心应用的广泛化,人工智能等越来越多地应用到数据中心。而神经网络作为人工智能的重要技术,已经大量应用到数据中心大数据分析运算中。

[0051] 在传统的大型数据中心,网络结构通常如图1所示,即互连网络模型(hierarchical inter-networking model)。这个模型包含了以下部分:

[0052] 服务器140:各服务器140是数据中心的处理和存储实体,数据中心中大量数据的处理和存储都是由这些服务器140完成的。

[0053] 接入交换机130:接入交换机130是用来让服务器140接入到数据中心中的交换机。一台接入交换机130接入多台服务器140。接入交换机130通常位于机架顶部,所以它们也被称为机顶(Top of Rack)交换机,它们物理连接服务器。

[0054] 汇聚交换机120:每台汇聚交换机120连接多台接入交换机130,同时提供其他的服务,例如防火墙,入侵检测,网络分析等。

[0055] 核心交换机110:核心交换机110为进出数据中心的包提供高速的转发,为汇聚交换机120提供连接性。整个数据中心的网络分为L3层路由网络和L2层路由网络,核心交换机110为通常为整个数据中心的网络提供一个弹性的L3层路由网络。

[0056] 通常情况下,汇聚交换机120是L2和L3层路由网络的分界点,汇聚交换机120以下的是L2网络,以上是L3网络。每组汇聚交换机管理一个传送点(POD,Point Of Delivery),每个POD内都是独立的VLAN网络。服务器在POD内迁移不必修改IP地址和默认网关,因为一个POD对应一个L2广播域。

[0057] 汇聚交换机120和接入交换机130之间通常使用生成树协议(STP,Spanning Tree Protocol)。STP使得对于一个VLAN网络只有一个汇聚层交换机120可用,其他的汇聚交换机120在出现故障时才被使用(上图中的虚线)。也就是说,在汇聚交换机120的层面,做不到水平扩展,因为就算加入多个汇聚交换机120,仍然只有一个在工作。

[0058] 服务器

[0059] 由于服务器140才是数据中心真实的处理设备,图2示出了一个服务器140内部的结构框图。服务器140包括有总线连接的存储器210、调度单元集群270和加速单元集群280。调度单元集群270包括多个调度单元220。加速单元集群280包括多个加速单元230。加速单元在本公开实施例中主要是为了加速神经网络模型的运算处理速度而设计的专门处理单元,可以体现为专门为神经网络运算处理设计的处理单元、图形处理单元(GPU)、专用集成电路(ASIC)和现场可编程门阵列(FPGA)等。调度单元是对加速单元进行调度、向各加速单元分配要执行的待执行指令序列的处理单元,它可以采用中央处理单元(CPU)、专用集成电路(ASIC)、现场可编程门阵列(FPGA)等多种形式。

[0060] 传统的中央处理单元的架构设计,使得在架构中控制单元、存储单元占用了很大一部分空间,而计算单元占用的空间反而不足,因此其在逻辑控制方面十分有效,而在大规模并行计算方面则效率不够。因此,开发出了各种专门的加速单元,用来针对不同功能和不同领域的计算进行更有效的提高运算速度的处理。本发明提出的加速单元是专用于加速神经网络模型的运算处理速度的处理单元。它是采用数据驱动并行计算的架构,用于处理各神经网络节点的大量运算(例如卷积、池化等)的处理单元。由于各神经网络节点的大量运算(例如卷积、池化等)中的数据和中间结果在整个计算过程中紧密联系,会被经常用到,用现有的中央处理单元构架,由于中央处理单元的核内的内存容量很小,因此要大量频繁访问核外存储器,造成处理的低效。采用这种专用于加速神经网络模型的运算处理速度的加速单元,由于其每个核中具有适于神经网络计算用到的存储容量的片上内存,避免频繁访

问核外部的存储器,就能大大提高处理效率,提高计算性能。

[0061] 加速单元230要接受调度单元220的调度。如图2所示,存储器210中存储有各种神经网络模型,包括这些模型的节点和节点的权重数据等。这些神经网络模型当需要时被图2中的一个调度单元220部署到一个加速单元230。即,调度单元220可以通过指令的形式向加速单元230发送模型中的参数(如各节点的权重)在存储器210中的地址。加速单元230在实际使用该神经网络模型进行计算时,就会根据这些参数(例如权重)在存储器210中的地址,直接在存储器210中寻址这些参数,将其暂存在其片上内存中。加速单元230在实际使用该神经网络模型进行计算时,调度单元220还会将模型的输入参数通过指令的形式发送给加速单元230,暂存在加速单元230的片上内存中。这样,加速单元230就可以根据这些输入参数和模型中的参数(例如权重)进行推理计算。

[0062] 调度单元和加速单元的内部结构

[0063] 下面结合图3的调度单元220与加速单元230的内部结构图,具体说明调度单元220是如何调度加速单元230进行工作的。

[0064] 如图3所示,调度单元220内包含多个处理器核222和被多个处理器核222共享的高速缓存221。每个处理器核222包括取指令单元203、指令译码单元224、指令发射单元225、指令执行单元226。

[0065] 取指令单元223用于将要执行的指令从存储器210中搬运到指令寄存器(可以是图3示出的寄存器堆229中的一个用于存放指令的寄存器)中,并接收下一个取指地址或根据取指算法计算获得下一个取指地址,取指算法例如包括:根据指令长度递增地址或递减地址。

[0066] 取出指令后,调度单元220进入指令译码阶段,指令译码单元224按照预定的指令格式,对取回的指令进行解码,以获得取回的指令所需的操作数获取信息,从而为指令执行单元225的操作做准备。操作数获取信息例如指向立即数、寄存器或其他能够提供源操作数的软件/硬件。

[0067] 指令发射单元225位于指令译码单元224与指令执行单元226之间,用于指令的调度和控制,以将各个指令高效地分配至不同的指令执行单元226,使得多个指令的并行操作成为可能。

[0068] 指令发射单元225将指令发射到指令执行单元226后,指令执行单元226开始执行指令。但如果该指令执行单元226判断该指令应该是加速单元执行的,则将其转发到相应的加速单元执行。例如,如果该指令是一条神经网络推理(inference)的指令,指令执行单元226不再执行该指令,而是将该指令通过总线发送到加速单元230,由加速单元230执行。

[0069] 加速单元30内部包括多个核236(图3中示出了4个核,但本领域技术人员应当理解,加速单元230中也可以包含其它数目的核236)、命令处理器237、直接存储访问机制235、和总线通道231。

[0070] 总线通道231是指令从总线进出加速单元230的通道。

[0071] 直接内存访问(DMA,Direct Memory Access)机制235是一些计算机总线架构提供的功能,它能使数据从附加设备直接写入计算机主板的存储器上。这种方式相比于设备之间所有的数据传输都要通过调度单元的方式,大大提高了数据访问的效率。正是因为有这样的机制,加速单元230的核可以直接访问存储器210,读取神经网络模型中的参数(例如各

节点的权重)等,大大提高了数据访问效率。

[0072] 命令处理器237将由调度单元220发送至加速单元230的指令分配给核236执行。指令执行单元226将需要加速单元230执行的待执行指令序列发送给加速单元230。该待执行指令序列从总线通道231进入后,缓存在命令处理器237,由命令处理器237选择核236,将指令序列分配给其执行。本公开实施例主要是在命令处理器237中执行的。在本公开实施例中,命令处理器237在将待执行指令序列分配给核236执行之前,还在待执行指令序列中适当的位置加入跨核复制需要的一些指令,分配给主核,同时也给副核生成跨核复制需要的一些指令,分配给副核,在待执行指令序列实际执行时,这些指令与分配给主核的那些指令相配合,共同完成跨核复制。另外,命令处理器237还负责核236之间的同步操作。

[0073] 加速单元核

[0074] 图4是根据本公开一个实施例的加速单元核的内部结构图。

[0075] 在一个实施例中,如图4所示,加速单元核236包括张量引擎310、池化引擎320、存储器复制引擎330、定序器350、指令缓存器340、片上内存360、常数缓冲器370、寄存器堆380。

[0076] 寄存器堆380可以包括用于存储不同类型的数据和/或指令的多个寄存器,这些寄存器可以是不同类型的。例如,寄存器堆380可以包括:整数寄存器、浮点寄存器、状态寄存器、指令寄存器和指针寄存器等。寄存器堆380中的寄存器可以选用通用寄存器来实现,也可以根据实际需求采用特定的设计。

[0077] 命令处理器237分配给加速单元核236的指令序列首先进入指令缓存器340缓存。然后,定序器350从指令缓存器340中按照先进先出的顺序取指令,根据指令的性质分配给张量引擎310或池化引擎320执行。张量引擎310负责处理神经网络模型中的卷积和矩阵乘法等相关操作。池化引擎320负责处理神经网络模型中的池化操作。定序器350根据取出的指令是卷积、矩阵乘法还是池化等操作性质,决定将指令分配给张量引擎310还是池化引擎320。另外,由于本发明实施例中,命令处理器237会在待执行指令序列实际分配到核236之前,在待执行指令序列中加入跨核复制指令,如果定序器350发现从指令缓存器340取出的指令是一个跨核复制指令,就将其分给存储器复制引擎330处理。存储器复制引擎330是专门处理跨核的数据复制的单元。另外,为了顺利完成跨核复制,如后文详述,命令处理器237还可以会在待执行指令序列中一些位置加入同步原语指令。由于命令处理器237是负责核236之间同步的单元,如果定序器350发现从指令缓存器340中取出的指令是一个同步原语指令,其请求命令处理器237进行核间同步。如果命令处理器237从主核接收到一个主核同步原语指令,又从副核接收到一个副核同步原语指令,则命令处理器237按照这两个指令,使得主副核的处理同步。

[0078] 片上内存360是存储神经网络模型中的权重参数、以及神经网络模型实际使用时的输入和各种中间结果的核内存储器。常数缓冲器370是存储神经网络模型中除权重参数之外的其它常量参数(例如,神经网络模型中的超参)的缓冲器。如上所述,在调度单元220将神经网络模型预先配置在加速单元230的过程中,调度单元220通过指令的形式向加速单元230发送模型中的参数在存储器210中的地址。这些参数包括节点的权重和其它参数(例如超参)。对于权重,加速单元230在实际的神经网络模型运算时,将它从存储器210相应的位置取出,放在片上内存360中。对于其它参数,加速单元230在实际的神经网络模型运算

时,从存储器210相应的位置取出,放在常数缓冲器370中。另外,当实际开始推理(inference)的指令由命令处理器237分配给核236执行后,指令中的输入参数(给神经网络模型的输入)也存储在片上内存360。另外,当张量引擎310和池化引擎320进行卷积或池化运算后,得到的各种中间结果也存放在片上内存360中。

[0079] 如图5所示,核236可以分为主核2361和副核2362。主核2361是加速单元内部用于执行待执行指令序列,从而完成一系列实体操作或运算的核。副核2362是加速单元内部不用来完成实体操作或运算,只与主核2361相配合,在主核2361的片上内存存储不下指令序列运行时所需的操作数时,为主核2361暂存存储不下的操作数,等到主核2361的片上内存能够容纳时再复制回主核2361的片上内存的核。

[0080] 注意,这里的主核和副核只是为了描述方便,其区分不是绝对的。对于某个待执行指令序列来说,分配用来执行该执行指令序列的核对于该执行指令序列来说就是主核,为主核暂时容纳操作数的核就是副核,但对于另外的待执行指令序列,可能刚好相反。

[0081] 在主核2361的定序器350发现从主核2361的指令缓冲器340取出的一条指令是向副核2362的跨核复制指令时,主核2361的存储器复制引擎330与副核2362的存储器复制引擎330通信,从而通过副核的存储器复制引擎330将主核2361的第一片上内存3601的需要转移的操作数复制到副核2362的第二片上内存3602。

[0082] 在副核2362的定序器350发现从副核2362的指令缓冲器340取出的一条指令是向主核2361的跨核复制指令时,副核2362的存储器复制引擎330与主核2361的存储器复制引擎330通信,从而通过主核的存储器复制引擎330将主核2362的第二片上内存3602的需要转移的操作数复制到主核2361的第一片上内存3601。

[0083] 现有技术的片上内存容量不够时操作数转移策略

[0084] 下面结合图6A-B介绍现有技术的编译器编译时发现指令序列在执行时可能会片上内存不够时的操作数转移策略。

[0085] 图6A示出了一个待执行指令序列。当调度单元220的指令执行单元226判定该待执行指令序列是需要加速单元230执行的指令序列(例如,神经网络模型的推断指令序列)时,将该待执行指令序列发送给总线通道231,最后该指令序列进入命令处理器237。命令处理器237首先从前到后一一考查该指令序列中的每条指令,逐一确定是否该指令的执行会引起指令序列当前需要用到的操作数数量大于片上内存的最大允许存储操作数数量而产生溢出。当判断出某一条指令的执行时会产生上述溢出时,假设该指令新引入的操作数为第二操作数402。由于该操作数需要放入片上内存,因而需要挤出一个操作数放入加速单元中多个加速单元核共享的存储器(图3未示)。一般会在片上内存已有的操作数中,挑选代价最小的操作数挤出,挑选的算法例如线性扫描(linear scan)、贪婪分配因子(greedy allocator)算法等等。挑中的挤出的操作数为第一操作数401。在图6B中,第二操作数402是地址r7中的操作数,第一操作数401是地址r1中的操作数。

[0086] 然后,命令处理器237会在指令序列的第一位置410加入外存存储指令store MEM,r1,表示将挑中的地址r1中的第一操作数401移入外存,即加速单元中多个加速单元核共享的存储器(图3未示)。第一位置410是第二操作数402出现之前第一操作数401最后一次出现的指令之后。这是因为,在片上内存中,第二操作数402是用来替换第一操作数401的,要在第二操作数402需要使用之前将第一操作数401移除,所以要在第二操作数402出现之前找

到最后一条包含第一操作数401的指令,在它的后面将它转移到外存。由于这之后、第二操作数402出现之前,没有第一操作数401的指令,就不会影响指令序列的正常执行。

[0087] 然后,命令处理器237会在指令序列的第二位置420加入外存下载指令load r1,MEM,表示将地址r1中的第一操作数401移回片上内存。第二位置420是第二操作数402最后一次出现的指令后第一操作数401第一次出现的指令之前。因为,第二操作数402最后一次出现的指令后,第二操作数402已经不再使用,就可以将其从片上内存中去除,而第二操作数402最后一次出现的指令后第一操作数第一次出现的指令是该第一操作数401再一次被需要的位置,必须在该指令之前将第一操作数401召回,该指令才能顺利执行。

[0088] 命令处理器237加入指令后,图6A的指令序列变成图6B的指令序列。然后,命令处理器237将图6B的指令序列分配给加速单元核执行。

[0089] 本公开实施例中的用于跨核复制的相关指令的添加

[0090] 本公开实施例中,加速单元230的命令处理器237在从调度单元220接收到要由加速单元230执行的待执行指令序列(例如,涉及神经网络运算的指令)后,不是直接将其分配给核执行,而是在待执行指令序列的适当位置加入跨核复制所需要的各种主核执行语句(如图6C中的411、412、413、414、421、422),分配给主核2361执行,并为副核2362分配若干跨核复制所需要的各种副核执行语句(如图6C中的511、512、513、514、521、522),从而在待执行指令实际执行过程中,这些语句相互配合,共同完成跨核复制。

[0091] 下面结合图6C,详细讨论本公开实施例中命令处理器237发现指令序列在执行时可能会片上内存不够时添加用于跨核复制的相关指令的过程。图6C左侧的指令都是在待执行指令序列中添加的、要分配给加速单元主核2361执行的指令,图6C右侧的指令都是要分配给加速单元副核2362执行的指令。

[0092] 还以图6A的要执行的指令序列为例。命令处理器237与图6B中同样地,确定出会引起指令序列当前需要用到的操作数数量大于片上内存的最大允许存储操作数数量而产生溢出的指令,从而定位到指令序列执行时会引起第一片上内存3601溢出的第二操作数402,也与图6B中同样地确定出移出后代价最小的第一操作数401。第二操作数402仍然是地址r7中的操作数,第一操作数401仍然是地址r1中的操作数。

[0093] 然后,命令处理器237会在待执行指令序列的第一位置加入第一跨核复制指令421,即LMVPY s0,r1。该指令指示主核2363的寄存器堆380中的第一寄存器381作为源复制地址r1,其指示所述第一片上内存3601上要移到加速单元副核2362的第一操作数401的地址。该指令指示主核2363的寄存器堆380中的第二寄存器382作为目标复制地址s0,其指示所述第一操作数401要移到所述加速单元副核2362的第二片上内存3602上的地址。在主核2361实际执行这条指令时,从第一寄存器381中取出源复制地址r1,从第二寄存器382中取出目标复制地址s0,从所述第一片上内存3601中获取所述源复制地址r1中的第一操作数401,并将获取的第一操作数401复制到所述加速单元副核2362的第二片上内存3602中的所述目标复制地址s0。

[0094] 在一个实施例中,第一位置410可以是第二操作数402出现之前第一操作数401最后一次被使用的指令之后。这是因为,必须要在第二操作数402使用之前将第一操作数401从第一片上内存3601转移,从而第二操作数402才能放入第一片上内存410。所以,要在第二操作数402出现之前找到最后一条包含第一操作数401的指令,在它的后面将它转移到加速

单元副核2362的第二片上内存3602。由于这之后、第二操作数402出现之前，没有第一操作数401的指令，就不会影响指令序列的正常执行。在图6C中，第一位置410就是在r7第一次出现之前r1最后一次使用的指令conv r1,r0之后。

[0095] 命令处理器237可以在待执行指令序列中上述插入的第一跨核复制指令421之前，插入第一主核同步原语指令411，即SYNC.LMCPY。对称地，为副核2362分配一个第一副核同步原语指令511，即SYNC.LMCPY。

[0096] 同步的意义在于，确保核之间执行步调一致。两个核执行指令的进度不一致，可能一个核执行完某一指令时另一个核正在执行另一指令，同步的意思就是等待另一个核执行完后，两个核一起开始执行各自后面的指令。因此，本公开实施例中，由于插入了第一跨核复制指令421。在主核2361执行第一跨核复制指令421的复制时，副核2362如果正在执行其它的动作，可能造成复制的错误。因此，在前面插入第一主核同步原语指令411，为副核2362分配一个第一副核同步原语指令511。在同步时，第一主核同步原语指令411、第一副核同步原语指令511互相配合一起执行，这样执行完毕后，再执行第一跨核复制指令421时，副核2362也刚刚执行完第一副核同步原语指令511，不会被别的指令占用，因此跨核复制不会出错。

[0097] 命令处理器237可以在第一副核同步原语指令511的后面为副核2362分配一个第一哑跨核复制指令521，即图6C右边划虚线的LMCPY s0,r1，它并不是必要的，在执行时它被忽略，不会执行。它的作用是，在某些加速单元中要求，同步的实现机制必须保证多个加速单元核的同步原语指令跟着后面必须有相同数量的LMCPY指令，但有些加速单元中并没有这样的要求。因此，它是可选的指令。

[0098] 命令处理器237可以在所述第二操作数402在待执行指令序列中第一次出现的指令之前的位置，插入一个第二主核同步原语指令412，即SYNC.LMCPY。相应地，为副核2362分配一个第二副核同步原语指令512，即SYNC.LMCPY。

[0099] 在所述第二操作数402在待执行指令序列中第一次出现的指令之前的位置插入第二主核同步原语指令412，是为了在第二操作数402真正开始被使用之前，确保第一操作数401已经安全地复制到加速单元副核2362的第二片上内存3602了。操作数的复制需要一个过程。虽然，在第一位置已经放入第一跨核复制指令411，用于将第一操作数401复制到第二片上内存3602，但由于操作数的复制需要时间，有可能当指令序列执行到第二操作数402在待执行指令序列中第一次出现的指令之前的位置时，该复制仍未完成。需要说明的是，待执行指令序列由定序器350分配给张量引擎310、池化引擎320、存储器复制引擎330去执行。虽然它们是按照从前到后的顺序取出并分配给各引擎，但各引擎的执行通常是并行的，而每条指令随着其内容的不同，其执行时间不一定是一样的，这就导致出现了在几条指令之前的一个指令可能在几条指令之后仍未执行完的情况。通过给待执行指令序列添加第二主核同步原语指令412、为副核2362分配第二副核同步原语指令512，实现加速单元主核2361侧和加速单元副核2362侧的同步，这样，等到第二操作数402第一次出现的指令开始执行的时候，第一跨核复制指令已经保证执行完，不会出现第一操作数401没有复制完最后丢失的情况。

[0100] 命令处理器237可以所述第二操作数402在待执行指令序列中最后一次出现的指令之后，加入一个第三主核同步原语指令413和第二哑跨核复制指令422。第三主核同步原

语指令413,即SYNC.LMCPY,第二哑跨核复制指令422如图6C左边划横虚线的LMCPY r1,s0。同时,命令处理器237为副核2362生成一个第三副核同步原语指令513,即SYNC.LMCPY,并生成第二跨核复制指令522,即图6C右边的LMCPY r1,s0。

[0101] 之所以在所述第二操作数402在待执行指令序列中最后一次出现的指令之后,加入一个第三主核同步原语指令413和第二哑跨核复制指令422,是因为第二操作数402在指令序列中最后一次出现的指令之后,就不再使用该第二操作数402了,就可以安全地利用第二跨核复制指令522将第一操作数401从第二片上内存3602取回,将第二操作数402在第一片上内存3601中替换。这样,一旦之后又出现使用第一操作数401的指令,就不会影响其执行。在图6C中,在r7最后一次出现的指令mul r7,r6,r0之后,加入上述指令。

[0102] 第二跨核复制指令522,即图6C右边的LMCPY r1,s0,指示副核2362的寄存器堆380中的第三寄存器383作为源复制地址s0,其指示所述第二片上内存3602上要移回加速单元主核2361的第一操作数401的地址。该指令还指示副核2362的寄存器堆380中的第四寄存器384作为目标复制地址r1,其指示所述第一操作数401要移回所述加速单元主核2361的第一片上内存3601上的地址。在副核2362实际执行这条指令时,从第三寄存器383中取出源复制地址s0,从第四寄存器384中取出目标复制地址r1,从所述第二片上内存3602中获取所述源复制地址s0中的第一操作数401,并将获取的第一操作数401复制到所述加速单元主核2361的第一片上内存3601中的所述目标复制地址r1。

[0103] 命令处理器237为副核2362生成第二跨核复制指令522之前,为副核2362生成第三副核同步原语指令513,并且对称地,在待执行指令序列中也插入第三主核同步原语指令413,是因为,通过这两个同步原语指令确保主核2361和副核2362之间执行步调一致,即如果一个核执行完一个指令,另一个核没有执行完正在执行的指令,则等待另一个核执行完后,两个核一起开始执行各自后面的指令。由于如上所述,为副核2362生成了第二跨核复制指令522,在主核2362执行第二跨核复制指令522的复制时,主核2361如果正在执行其它的动作,可能造成复制的错误。因此,为副核2362生成第三副核同步原语指令513,在待执行指令序列中也插入第三主核同步原语指令413,从而在同步时,这两个同步原语指令互相配合一起执行,这样执行完毕后,再执行第二跨核复制指令522时,主核2361也刚刚执行完第三主核同步原语指令413,不会被别的指令占用,因此跨核复制不会出错。

[0104] 在第三主核同步原语指令413的后面加入的第二哑跨核复制指令422并不是必要的,在执行时它被忽略,不会执行。它的作用是,在某些加速单元中要求,同步的实现机制必须保证多个加速单元核的同步原语指令跟着后面必须有相同数量的LMCPY指令,但有些加速单元中并没有这样的要求。因此,它是可选的指令。

[0105] 命令处理器237可以在所述第一操作数401在所述待执行指令序列中所述第二操作数402最后一次出现之后第一次被使用的指令之前,加入第四主核同步原语指令414,即SYNC.LMCPY。相应地,为副核2362分配一个第四副核同步原语指令514,即SYNC.LMCPY。

[0106] 在指令序列中第二操作数402最后一次出现之后第一次被使用的指令之前加入第四主核同步原语指令414,是为了在第一操作数401又一次开始被使用之前,确保第一操作数401已经安全地复制回加速单元主核2361的第一片上内存3601了。操作数的复制需要一个过程。虽然,在如上所述,已经为副核2362生成了第二跨核复制指令,用于将第一操作数401复制回第一片上内存3601,但由于操作数的复制需要时间,有可能当指令序列执行到所

述第一操作数401在所述待执行指令序列中所述第二操作数402最后一次出现之后第一次被使用的指令的位置,该复制仍未完成。通过在加速单元主核2361侧和在加速单元副核2362侧都增加一个同步原语指令,就能避免这一问题。通过该指令,实现加速单元主核2361侧和加速单元副核2362侧的同步,这样,等到第一操作数401又一次出现的指令开始执行的时候,第二跨核复制指令已经保证执行完,不会出现第一操作数401没有复制完整就开始使用的情况。

[0107] 虽然在图5中只画出了一个副核2362,但本领域技术人员应当理解,所述副核2362也可以有多个。在多个副核2362的情况下,第一跨核复制指令用于将第一操作数从所述第一片上内存3601复制到多个副核2362中选定的副核2362的第二片上内存3602,第二跨核复制指令用于将第一操作数从所述选定的副核2362的第二片上内存3602复制回第一片上内存3601。选定副核2362的方法可以有多种。在一个实施例中,选定的副核2362是根据所述多个副核2362的每一个到所述主核的距离,从所述多个副核2362中选出的。这样选择副核2362的好处是,可以节省复制时的传输距离,从而提高跨核复制的效率。

[0108] 跨核复制的相关指令的实际执行

[0109] 本公开实施例的跨核复制,主要是通过在确定加速单元主核2361的第一片上内存3601有可能溢出时,将需要转移的操作数从第一片上内存3602转移到加速单元副核2362的第二片上内存3602,并在不会导致溢出的情况下,再将该操作数从第二片上内存3602转移回第一片上内存3601。因此,先描述一下上述主要实现过程的执行流程。

[0110] 如图5所示,加速单元主核2361的寄存器堆380包括第一寄存器381和第二寄存器382,分别用于存放所述第一地址和所述第二地址。加速单元主核2361的指令缓存器340接收到命令处理器237插入了各种完成跨核复制需要的指令的待执行指令序列后,主核定序器350从中顺序取出进行译码。当取出的是第一跨核复制指令421,即LMCPY s0,r1时,第一跨核复制指令421将第一寄存器381中的第一地址作为跨核复制的源地址r1,将第二寄存器382中的第二地址作为跨核复制的目的地地址s0。主核定序器350将译码后的该指令交给加速单元主核2361中的存储器复制引擎330执行。主核存储器复制引擎330接收并执行译码后的第一跨核复制指令421,从第一寄存器381中取出作为源地址r1的第一地址,从第二寄存器382取出作为目的地地址s0的第二地址,从而将所述源地址r1中的第一操作数转移到所述第二片上内存3602上的目的地地址s0。

[0111] 加速单元副核2362的寄存器堆380包括第三寄存器383和第四寄存器384,分别用于存放所述第二地址和所述第一地址。加速单元副核2362的指令缓存器340接收到命令处理器237分配的用于跨核复制的副核需执行指令后,副核定序器350从中顺序取出进行译码。当取出的是第二跨核复制指令522,即LMCPY r1,s0时,第二跨核复制指令522将第三寄存器383中的第二地址作为跨核复制的源地址s0,将第四寄存器384中的第一地址作为跨核复制的目的地地址r1。副核定序器350将译码后的该指令交给加速单元副核2362中的存储器复制引擎330执行。副核存储器复制引擎330接收并执行译码后的第二跨核复制指令522,从第三寄存器383中取出作为源地址s0的第二地址,从第四寄存器384取出作为目的地地址r1的第一地址,从而将所述源地址s0中的第一操作数转移回所述第一片上内存3601上的目的地地址r1。

[0112] 通过上述过程,就大体实现了在加速单元主核2361的第一片上内存3601有可能溢

出时,将需要转移的第一操作数从第一片上内存3602转移到加速单元副核2362的第二片上内存3602,并适当时再将该操作数从第二片上内存3602转移回第一片上内存3601的跨核复制过程。

[0113] 上述过程仅是一个粗略的跨核复制的过程。实际上,为了保证跨核复制过程中的各种数据不会丢失,或者为了保证在执行跨核复制时相关的指令已经执行完,不会发生复制错误的情况,还要考虑到接收到各种同步原语指令时对同步原语指令的执行。下面详细描述对各种同步原语指令的执行过程。

[0114] 主核指令缓存器340在接收并缓存所述第一跨核复制指令421之前,可以接收并缓存第一主核同步原语指令411,即SYMC.LMCPY,用于加速单元主核2361与加速单元副核2362的同步。所述主核定序器350对缓存的所述第一主核同步原语指令411进行译码,并将译码后的第一主核同步原语指令411发送给命令处理器237。

[0115] 副核指令缓存器340接收并缓存与所述第一主核同步原语指令411对应的第一副核同步原语指令511,即SYMC.LMCPY,并且也接收并缓存与所述第一跨核复制指令421对应的第一哑跨核复制指令521。副核定序器350对缓存的所述第一副核同步原语指令511、和第一哑跨核复制指令521进行译码,并将译码后的第一副核同步原语指令511发送给命令处理器237。这样,命令处理器237就不仅接收到了译码后的第一主核同步原语指令411和译码后的第一副核同步原语指令511。根据这两个指令,就可以加速单元主核2361与加速单元2362的同步。根据主核同步原语指令411和第一副核同步原语指令511进行同步的意义在上文命令处理器237在待执行指令序列中插入上述指令的描述中已经提到,故不赘述。对于第一哑跨核复制指令521,如前所述,其没有实际意义,因此不执行它。

[0116] 主核指令缓存器340是在接收并缓存所述第一操作数r1在待执行指令序列中第二操作数r7出现之前最后一次被使用的指令conv r1,r0之后,接收并缓存第一主核同步原语指令411的。之所以这样做的意义,在上文命令处理器237在待执行指令序列中插入上述指令的描述中已经提到。

[0117] 主核指令缓存器340在接收并缓存所述第二操作数r7在待执行指令序列中第一次出现的指令mul r7,r6,r0之前,接收并缓存第二主核同步原语指令422,即SYMC.LMCPY。在mul r7,r6,r0之前接收第二主核同步原语指令422的原因,在上文命令处理器237在待执行指令序列中插入上述指令的描述中已经提到。然后,主核定序器350对缓存的所述第二主核同步原语指令422进行译码,并将译码后的第二主核同步原语指令422发送给命令处理器237。

[0118] 副核指令缓存器340在接收并缓存第一哑跨核复制指令521之后,接收并缓存与所述第二主核同步原语指令412对应的第二副核同步原语指令512。副核定序器350对缓存的第二副核同步原语指令512进行译码,并将译码后的第二副核同步原语指令512发送给命令处理器237。命令处理器237既接收到了上述第二主核同步原语指令412,又接收到了上述第二副核同步原语指令512,就可以根据这两个指令进行所述主核与所述副核的同步。根据第二主核同步原语指令412和第二副核同步原语指令512进行同步的意义,在上文命令处理器237在待执行指令序列中插入上述指令的描述中已经提到,故不赘述。

[0119] 主核指令缓存器340在接收并缓存所述第二操作数r7在待执行指令序列中最后一次出现的指令mul r7,r6,r0之后,顺序接收第三主核同步原语指令413和与所述第二跨核

复制指令522对应的第二哑跨核复制指令422。第三主核同步原语指令413即SYNC.LMCPY,第二哑跨核复制指令422即LMCPY r1,s0。在所述第二操作数在待执行指令序列中最后一次出现的指令之后,顺序接收这些指令的意义,在上文命令处理器237在待执行指令序列中插入上述指令的描述中已经提到,故不赘述。

[0120] 主核定序器350对缓存的所述第三主核同步原语指令413和第二哑跨核复制指令422进行译码,并将译码后的第三主核同步原语指令413发送给命令处理器237。由于第二哑跨核复制指令422是不执行的,因此可以忽略。

[0121] 副核指令缓存器340在接收并缓存第二跨核复制指令522之前、接收并缓存第二副核同步原语指令512之后,接收并缓存第三副核同步原语指令513。副核定序器350对缓存的所述第三副核同步原语指令513进行译码,并将译码后的第三副核同步原语指令513发送给命令处理器237。此时,命令处理器237既接收到了第三主核同步原语指令413,又接收到了第三副核同步原语指令513,就可以根据这两个指令进行加速单元主核2361与加速单元副核2362的同步。该同步的意义在上文命令处理器237在待执行指令序列中插入上述指令的描述中已经提到。

[0122] 主核指令缓存器340在接收并缓存所述第一操作数r1在所述待执行指令序列中所述第二操作数r7最后一次出现之后第一次被使用的指令conv r10,r1之前,接收并缓存第四主核同步原语指令414,即SYNC.LMCPY。在所述第一操作数r1在所述待执行指令序列中所述第二操作数r7最后一次出现之后第一次被使用的指令conv r10,r1之前接收并缓存的原因,在上文命令处理器237在待执行指令序列中插入上述指令的描述中已经提到,故不赘述。

[0123] 主核定序器350对缓存的所述第四主核同步原语指令414进行译码,并将译码后的第四主核同步原语指令414发送给命令处理器237。

[0124] 副核指令缓存器340在接收并缓存第二跨核复制指令522之后,接收并缓存与所述第四主核同步原语指令414对应的第四副核同步原语指令514,即SYNC.LMCPY。副核定序器350对缓存的第四副核同步原语指令514进行译码,并将译码后的第四副核同步原语指令514发送给命令处理器237。命令处理器237接到了第四主核同步原语指令414,又接到了第四副核同步原语指令514,就可以按照这两个指令进行加速单元主核2361与加速单元副核2362的同步。同步的意义在待执行指令序列中插入上述指令的描述中已经提到,故不赘述。

[0125] 以上部分侧重于从对跨核复制的影响力方面,对各种与跨核复制相关的指令的执行进行描述,其不对应于时间的顺序。为了更清楚各指令在时间上执行的先后顺序,下面参考图7A-B描述本公开实施例中跨核复制的相关指令的执行过程。

[0126] 如图7A所示,调度单元220将需要加速单元230执行的待执行指令序列发送给加速单元230,经总线通道231进入命令处理器237。命令处理器237在待执行指令序列的适当位置加入跨核复制所需要的各种主核执行语句(如图6C中的411、412、413、414、421、422),分配给主核2361执行,并为副核2362分配若干跨核复制所需要的各种副核执行语句(如图6C中的511、512、513、514、521、522),分配给副核2362执行。

[0127] 主核2361的指令缓存器340接收并缓存命令处理器237加入了跨核复制所需要的各种主核执行语句的待执行指令序列。主核2361的定序器350从该指令缓存器340按照指令缓存器340先进先出的顺序逐一取出其中的指令,译码,并决定分配给主核2361的张量引擎

310,还是池化引擎320,还是存储器复制引擎330执行。

[0128] 副核2362的指令缓存器340接收并缓存命令处理器237发送来的跨核复制所需要的各种副核执行语句。副核2362的定序器350从该指令缓存器340按照指令缓存器340先进先出的顺序逐一取出其中的指令,译码,并决定分配给副核2362的张量引擎310,还是池化引擎320,还是存储器复制引擎330执行。

[0129] 主核2361的指令缓存器340接收并缓存所述第一操作数401在该待执行指令序列中第二操作数402出现之前最后一次被使用的指令conv r1,r0之后,接收并缓存第一主核同步原语指令411。主核2362的定序器350对缓存的所述第一主核同步原语指令411进行译码,并将译码后的第一主核同步原语指令411发送给所述主核2361外的命令处理器237。

[0130] 副核2362的指令缓存器340接收并缓存第一副核同步原语指令511。副核2362的定序器350对缓存的所述第一副核同步原语指令511进行译码,并将译码后的第一副核同步原语指令511发送给所述副核2362外的命令处理器237。

[0131] 命令处理器237根据译码后的第一主核同步原语指令411和第一副核同步原语指令511进行主核2361和副核2362的同步。

[0132] 接着,主核2361的指令缓存器340接收并缓存待执行指令序列中插入的第一跨核复制指令421,即LMCPY s0,r1。主核2361的定序器350对缓存的第一跨核复制指令421进行译码,发送给存储器复制引擎330。所述第一跨核复制指令421指示将第一寄存器381中的第一地址作为跨核复制的源地址r1,将第二寄存器382中的第二地址作为跨核复制的目的地地址s0。主核存储器复制引擎330将作为所述源地址的第一地址中的第一操作数转移到作为所述目的地地址的第二地址。

[0133] 与此同时,副核2362的指令缓存器340接收并缓存待执行指令序列中插入的第一哑跨核复制指令521,即图6C右边画虚线的LMCPY s0,r1。副核2362的定序器350对缓存的第一哑跨核复制指令521进行译码,发现其是哑跨核复制指令,对其不作处理。

[0134] 接着,主核2361的指令缓存器340在接收并缓存所述第二操作数402在待执行指令序列中第一次出现的指令mul r7,r6,r0之前,接收并缓存第二主核同步原语指令412。主核2362的定序器350对缓存的所述第二主核同步原语指令412进行译码,并将译码后的第二主核同步原语指令412发送给所述主核2361外的命令处理器237。

[0135] 副核2362的指令缓存器340接收并缓存第二副核同步原语指令512。副核2362的定序器350对缓存的所述第二副核同步原语指令512进行译码,并将译码后的第二副核同步原语指令512发送给所述副核2362外的命令处理器237。

[0136] 命令处理器237根据译码后的第二主核同步原语指令412和第二副核同步原语指令512进行主核2361和副核2362的同步。

[0137] 接着,如图7B所示,主核2361的指令缓存器340在接收并缓存所述第二操作数402在待执行指令序列中最后一次出现的指令之后,接收第三主核同步原语指令413。主核2362的定序器350对缓存的所述第三主核同步原语指令413进行译码,并将译码后的第三主核同步原语指令413发送给所述主核2361外的命令处理器237。

[0138] 副核2362的指令缓存器340接收并缓存第三副核同步原语指令513。副核2362的定序器350对缓存的所述第三副核同步原语指令513进行译码,并将译码后的第三副核同步原语指令513发送给所述副核2362外的命令处理器237。

[0139] 命令处理器237根据译码后的第三主核同步原语指令413和第三副核同步原语指令513进行主核2361和副核2362的同步。

[0140] 接着,副核2362的指令缓存器340接收并缓存待执行指令序列中插入的第二跨核复制指令522,即LMCPY r1,s0。副核2362的定序器350对缓存的第二跨核复制指令522进行译码,发送给存储器复制引擎330。所述第二跨核复制指令522指示将第三寄存器383中的第二地址作为跨核复制的源地址s0,将第四寄存器384中的第一地址作为跨核复制的目的地地址r1。副核存储器复制引擎330将作为所述源地址的第二地址中的第一操作数转移回作为所述目的地地址的第一地址。这样,实现了当主核2361的第一片上内存3601的存储空间恢复时暂时复制到副核2362的第二片上内存3602的操作数的及时取回。

[0141] 与此同时,主核2362的指令缓存器340接收并缓存待执行指令序列中插入的第二哑跨核复制指令422,即图6C左边画虚线的LMCPY r1,s0。副核2362的定序器350对缓存的第二哑跨核复制指令422进行译码,发现其是哑跨核复制指令,对其不作处理。

[0142] 接着,主核2361的指令缓存器340在接收并缓存所述第一操作数401在所述待执行指令序列中所述第二操作数402最后一次出现之后第一次被使用的指令conv r10,r1之前,接收第四主核同步原语指令414。主核2362的定序器350对缓存的所述第四主核同步原语指令414进行译码,并将译码后的第四主核同步原语指令414发送给所述主核2361外的命令处理器237。

[0143] 副核2362的指令缓存器340接收并缓存第四副核同步原语指令514。副核2362的定序器350对缓存的所述第四副核同步原语指令514进行译码,并将译码后的第四副核同步原语指令514发送给所述副核2362外的命令处理器237。

[0144] 命令处理器237根据译码后的第四主核同步原语指令414和第四副核同步原语指令514进行主核2361和副核2362的同步。

[0145] 协同跨核复制

[0146] 以上仅仅描述了将加速单元主核2361的第一片上内存3601中可能会溢出的操作数复制到单个加速单元副核2362的第二片上内存3602并在适当时拷贝回的过程。但实际上,有可能出现协同跨核复制的情况。协同跨核复制是指,将加速单元主核的第一片上内存中的多个操作数复制到多个加速单元副核的片上内存并在适当时拷贝回的过程。它有利于提高多个加速单元副核之间的协同存储效率。尤其是,其在需要复制的操作数比较多的情况下,很有可能在一个加速单元副核上未必存在足够的容纳这些操作数的存储空间,需要利用多个加速单元副核的存储空间一起工作才能满足要求。因此,该实施例提高了复制成功的概率,也提高了加速单元的总体存储空间利用率。

[0147] 在协同跨核复制的一个实施例中,要从加速单元主核2361转移出的操作数包括两部分操作数,即第一部分和第二部分,其中,第一部分和第二部分各自可能都包含一个或多个操作数,第一部分要从加速单元主核2361的第一片上内存3601转移到第一加速单元副核2362的第二片上内存3602,第二部分要从加速单元主核2361的第一片上内存3601转移到第二加速单元副核2363的第三片上内存3603。虽然上述实施例以要转移的操作数包括两个部分来举例,但本领域技术人员都知道,要转移的操作数也可以包括三个、四个或更多的部分,以转移到三个、四个或更多的加速单元副核的片上内存,其原理与上述包括两个部分的情形是相同的。因此,下面仅示例性描述操作数包括二个部分的情况下的加速单元结构、以

及工作原理。对于操作数包括三个、四个或更多的部分的情况,可以以此类推。

[0148] 操作数包括二个部分的情况下,待转移操作数集合分成第一部分和第二部分。

[0149] 第一部分包括一个或多个操作数,存储在加速单元主核2361的第一片上内存3601上的第一地址集合的存储空间,将要复制到第一加速单元副核2362的第二片上内存3602的第二地址集合的存储空间。例如,第一地址集合为1700-17FF,其中存储了256个操作数,第二地址集合为3800-38FF,也可以存储256个操作数。这样,可以正好地将第一地址集合中存储的操作数拷贝到第二地址集合的存储空间。

[0150] 第二部分包括一个或多个操作数,存储在加速单元主核2361的第一片上内存3601上的第三地址集合的存储空间,将要复制到第二加速单元副核2363的第三片上内存3603的第四地址集合的存储空间。例如,第三地址集合为1800-181F,其中存储了32个操作数,第四地址集合为4000-401F,也可以存储32个操作数。这样,可以正好地将第三地址集合中存储的操作数拷贝到第四地址集合的存储空间。

[0151] 如图8所示,协同跨核复制模式下的加速单元至少包括加速单元主核2361、第一加速单元副核2362和第二加速单元副核2363。

[0152] 加速单元主核2361包括主核张量引擎310、主核池化引擎320、主核存储器复制引擎330、主核定序器350、主核指令缓存器340、第一片上内存3601、主核常数缓冲器370、主核寄存器堆380。其结构与图5中主核2361的结构基本是相同的,故不赘述。不同的是,主核寄存器堆380包括第一首地址寄存器3801、第一尾地址寄存器3802、第二首地址寄存器3803、第二尾地址寄存器3804、第三首地址寄存器3805、第三尾地址寄存器3806、第四首地址寄存器3807、第一尾地址寄存器3808。第一首地址寄存器3801、第一尾地址寄存器3802分别用于存放第一地址集合的首地址和尾地址。第二首地址寄存器3803和第二尾地址寄存器3804分别用于存放所述第二地址集合的首地址和尾地址。第三首地址寄存器3805和第三尾地址寄存器3806分别用于存放所述第三地址集合的首地址和尾地址。第四首地址寄存器3807和第四尾地址寄存器3808分别用于存放所述第四地址集合的首地址和尾地址。

[0153] 第一加速单元副核2362包括第一副核张量引擎311、第一副核池化引擎321、第一副核存储器复制引擎331、第一副核主核定序器351、第一副核指令缓存器341、第二片上内存3602、第一副核常数缓冲器371、第一副核寄存器堆380’。其结构与图5中副核2362的结构基本是相同的,故不赘述。不同的是,第一副核寄存器堆380’包括第五首地址寄存器3809、第五尾地址寄存器3810、第六首地址寄存器3811、第六尾地址寄存器3812。第五首地址寄存器3809和第五尾地址寄存器3810分别用于存放所述第二地址集合的首地址和尾地址。第六首地址寄存器3811和第六尾地址寄存器3812分别用于存放所述第一地址集合的首地址和尾地址。

[0154] 与第一加速单元副核2362类似地,第二加速单元副核2363包括第二副核张量引擎312、第二副核池化引擎322、第二副核存储器复制引擎332、第二副核主核定序器352、第二副核指令缓存器342、第三片上内存3603、第二副核常数缓冲器372、第二副核寄存器堆380”。由于各个部分的功能与第一加速单元副核2362各个部分的功能类似,故不赘述。不同的是,第二副核寄存器堆380”包括第七首地址寄存器3813、第七尾地址寄存器3814、第八首地址寄存器3815、第八尾地址寄存器3816。第七首地址寄存器3813和第七尾地址寄存器3814分别用于存放所述第四地址集合的首地址和尾地址。第八首地址寄存器3815和第八尾

地址寄存器3816分别用于存放所述第三地址集合的首地址和尾地址。

[0155] 首先,与前述实施例类似,加速单元230的命令处理器237在从调度单元220接收到要由加速单元230执行的待执行指令序列(例如,涉及神经网络运算的指令)后,不是直接将其分配给核执行,而是在待执行指令序列的适当位置加入协同跨核复制所需要的各种主核执行语句,分配给加速单元主核2361执行,并为第一加速单元副核2362和第二加速单元副核2363分配若干跨核复制所需要的各种副核执行语句,从而在待执行指令实际执行过程中,这些语句相互配合,共同完成协同跨核复制。

[0156] 命令处理器237根据预设的规则,根据待执行指令序列确定出加速单元主核2361的第一片上内存3601中需要移出的待转移操作数集合,并根据第一加速单元副核2362的第二片上内存3602的剩余存储容量、第二加速单元副核2363的第三片上内存3603的剩余存储容量,确定出需要转移到第二片上内存3602的第一部分、需要转移到第三片上内存3603的第二部分。然后,将第一部分存储在第一片上内存3601上的第一地址集合的首地址和尾地址分别放在第一首地址寄存器3801和第一尾地址寄存器3802中,将第一部分将要存储在第二片上内存3602上的第二地址集合的首地址和尾地址分别放在第二首地址寄存器3803和第二尾地址寄存器3804中。将第二部分存储在第一片上内存3601上的第三地址集合的首地址和尾地址分别放在第三首地址寄存器3805和第三尾地址寄存器3806中,将第二部分将要存储在第三片上内存3603上的第四地址集合的首地址和尾地址分别放在第四首地址寄存器3807和第四尾地址寄存器3808中。

[0157] 然后,将第一部分存储在第一加速单元副核2362上的第二片上内存3602上的第二地址集合的首地址和尾地址分别放在第五首地址寄存器3809和第五尾地址寄存器3810中,将第二部分将要拷贝回加速单元主核2361的第一片上内存3601的第一地址集合的首地址和尾地址分别放在第六首地址寄存器3811和第六尾地址寄存器3812中。将第二部分存储在第二加速单元副核2363上的第三片上内存3603上的第四地址集合的首地址和尾地址分别放在第七首地址寄存器3813和第七尾地址寄存器3814中,将第二部分将要存储在加速单元主核2361上的第一片上内存3601的第三地址集合的首地址和尾地址分别放在第八首地址寄存器3815和第八尾地址寄存器3816中。

[0158] 命令处理器237在待执行指令序列中加入的协同跨核复制所需要的主核执行语句包括第一跨核复制指令。第一跨核复制指令指示将待转移操作数集合的第一部分从所述第一片上内存3601上的第一地址集合复制到所述第一加速单元副核2362的第二片上内存3602的第二地址集合,将待转移操作数集合的第二部分从所述第一片上内存3601上的第三地址集合复制到所述第二加速单元副核2362的第三片上内存3603的第四地址集合。命令处理器237在待执行指令序列中加入的协同跨核复制所需要的主核执行语句还可以包括类似图6C左边的同步原语指令的同步原语指令。

[0159] 此外,命令处理器237还分别为第一加速单元副核2362和第二加速单元副核2363生成协同跨核复制所需要的第一加速单元副核执行语句、第二加速单元副核执行语句。

[0160] 协同跨核复制所需要的第一加速单元副核执行语句包括第二跨核复制指令。第二跨核复制指令指示将第一部分从所述第二片上内存3602的第二地址集合复制回所述第一片上内存3601上的第一地址集合。此外,协同跨核复制所需要的第一加速单元副核执行语句还包括类似图6C右边的同步原语指令的同步原语指令。

[0161] 协同跨核复制所需要的第二加速单元副核执行语句包括第三跨核复制指令。第三跨核复制指令指示将第二部分从所述第三片上内存3603的第四地址集合复制回所述第一片上内存3601上的第三地址集合。此外,协同跨核复制所需要的第二加速单元副核执行语句还包括类似图6C右边的同步原语指令的同步原语指令。

[0162] 在实际执行时,加速单元主核2361的主核指令缓存器340接收到命令处理器237插入了各种完成协同跨核复制需要的指令的待执行指令序列后,主核定序器350从中顺序取出进行译码。发现取出的是第一跨核复制指令。第一跨核复制指令指示将第一片上内存上3601的、第一首地址寄存器3801中的首地址到第一尾地址寄存器3802中的尾地址之间的第一部分取出,复制到第二片上内存3602上的、第二首地址寄存器3803中的首地址到第二尾地址寄存器3804中的尾地址之间,且将第一片上内存3601上的、第三首地址寄存器3805中的首地址到第三尾地址寄存器3806中的尾地址之间的第二部分取出,复制到第三片上内存3603上的、第四首地址寄存器3807中的首地址到第四尾地址寄存器3808中的尾地址之间。

[0163] 主核定序器350将译码后的该指令交给加速单元主核2361中的存储器复制引擎330执行。主核存储器复制引擎330接收并执行译码后的第一跨核复制指令421,将第一片上内存上3601的、第一首地址寄存器3801中的首地址到第一尾地址寄存器3802中的尾地址之间的第一部分取出,复制到第二片上内存3602上的、第二首地址寄存器3803中的首地址到第二尾地址寄存器3804中的尾地址之间,且将第一片上内存3601上的、第三首地址寄存器3805中的首地址到第三尾地址寄存器3806中的尾地址之间的第二部分取出,复制到第三片上内存3603上的、第四首地址寄存器3807中的首地址到第四尾地址寄存器3808中的尾地址之间。

[0164] 第一加速单元副核2362的第一副核指令缓存器341接收到命令处理器237分配的用于跨核复制的第一副核需执行指令后,第一副核定序器351从中顺序取出进行译码。发现取出的是第二跨核复制指令。所述第二跨核复制指令指示将第二片上内存3602上的、第五首地址寄存器3809中的首地址到第五尾地址寄存器3810中的尾地址之间的第一部分取出,复制回第一片上内存3601上的、第六首地址寄存器3811中的首地址到第六尾地址寄存器3812中的尾地址之间。

[0165] 第一副核定序器351将译码后的该指令交给第一加速单元副核2362中的第一副核存储器复制引擎331执行。第一副核存储器复制引擎331接收并执行译码后的第二跨核复制指令,从而将第二片上内存3602上的、第五首地址寄存器3809中的首地址到第五尾地址寄存器3810中的尾地址之间的第一部分取出,复制回第一片上内存3601上的、第六首地址寄存器3811中的首地址到第六尾地址寄存器3812中的尾地址之间。

[0166] 第二加速单元副核2363的第二副核指令缓存器342接收到命令处理器237分配的用于跨核复制的第二副核需执行指令后,第二副核定序器352从中顺序取出进行译码。发现取出的是第三跨核复制指令。所述第三跨核复制指令指示将第三片上内存3603上的、第七首地址寄存器3813中的首地址到第七尾地址寄存器3814中的尾地址之间的第二部分取出,复制回第一片上内存3601上的、第八首地址寄存器3815中的首地址到第八尾地址寄存器3816中的尾地址之间。

[0167] 第二副核定序器352将译码后的该指令交给第二加速单元副核2363中的第二副核存储器复制引擎332执行。第二副核存储器复制引擎332接收并执行译码后的第三跨核复制

指令,从而将第三片上内存3603上的、第七首地址寄存器3813中的首地址到第七尾地址寄存器3814中的尾地址之间的第二部分取出,复制回第一片上内存3601上的、第八首地址寄存器3815中的首地址到第八尾地址寄存器3816中的尾地址之间。

[0168] 主核定序器350可能从主核指令缓存器340接收到同步原语指令,第一副核定序器351也可能从第一副核指令缓存器341接收到同步原语指令,第二副核定序器352也可能从第二副核指令缓存器342接收到同步原语指令,主核定序器350、第一副核定序器351、第二副核定序器352接收到这些同步原语指令后的处理与上面结合图6C所述的主副核定序器接收到同步原语指令后的处理相同,故不赘述。

[0169] 通过上述过程,就大体实现了在加速单元主核2361的第一片上内存3601有可能溢出时,将第一片上内存3601上待转移的第一部分操作数从加速单元主核2361的第一片上内存3601转移到第一加速单元副核2362的第二片上内存3602,将待转移的第二部分操作数要从加速单元主核2361的第一片上内存3601转移到第二加速单元副核2363的第三片上内存3603,并适当时再将第一部分和第二部分转移回第一片上内存3601。

[0170] 本公开实施例的商业价值

[0171] 本公开实施例在核内片上内存中的操作数可能溢出因此需要转移的情况下,大大了提高操作数转移效率,提高了加速单元性能。实验表明,这样的加速单元用于神经网络模型计算的速度快了一倍左右,大大提高了神经网络的推理速度,具有良好的市场前景。

[0172] 需要领会,以上所述仅为本公开的优选实施例,并不用于限制本公开,对于本领域技术人员而言,本说明书的实施例存在许多变型。凡在本公开的精神和原理之内所作的任何修改、等同替换、改进等,均应包含在本公开的保护范围之内。

[0173] 应该理解,本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同或相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于方法实施例而言,由于其基本相似于装置和系统实施例中描述的方法,所以描述的比较简单,相关之处参见其他实施例的部分说明即可。

[0174] 应该理解,上述对本说明书特定实施例进行了描述。其它实施例在权利要求书的范围内。在一些情况下,在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是可以的或者可能是有利的。

[0175] 应该理解,本文用单数形式描述或者在附图中仅显示一个的元件并不代表将该元件的数量限于一个。此外,本文中被描述或示出为分开的模块或元件可被组合为单个模块或元件,且本文中被描述或示出为单个的模块或元件可被拆分为多个模块或元件。

[0176] 还应理解,本文采用的术语和表述方式只是用于描述,本说明书的一个或多个实施例并不应局限于这些术语和表述。使用这些术语和表述并不意味着排除任何示意和描述(或其中部分)的等效特征,应认识到可能存在的各种修改也应包含在权利要求范围内。其他修改、变化和替换也可能存在。相应的,权利要求应视为覆盖所有这些等效物。

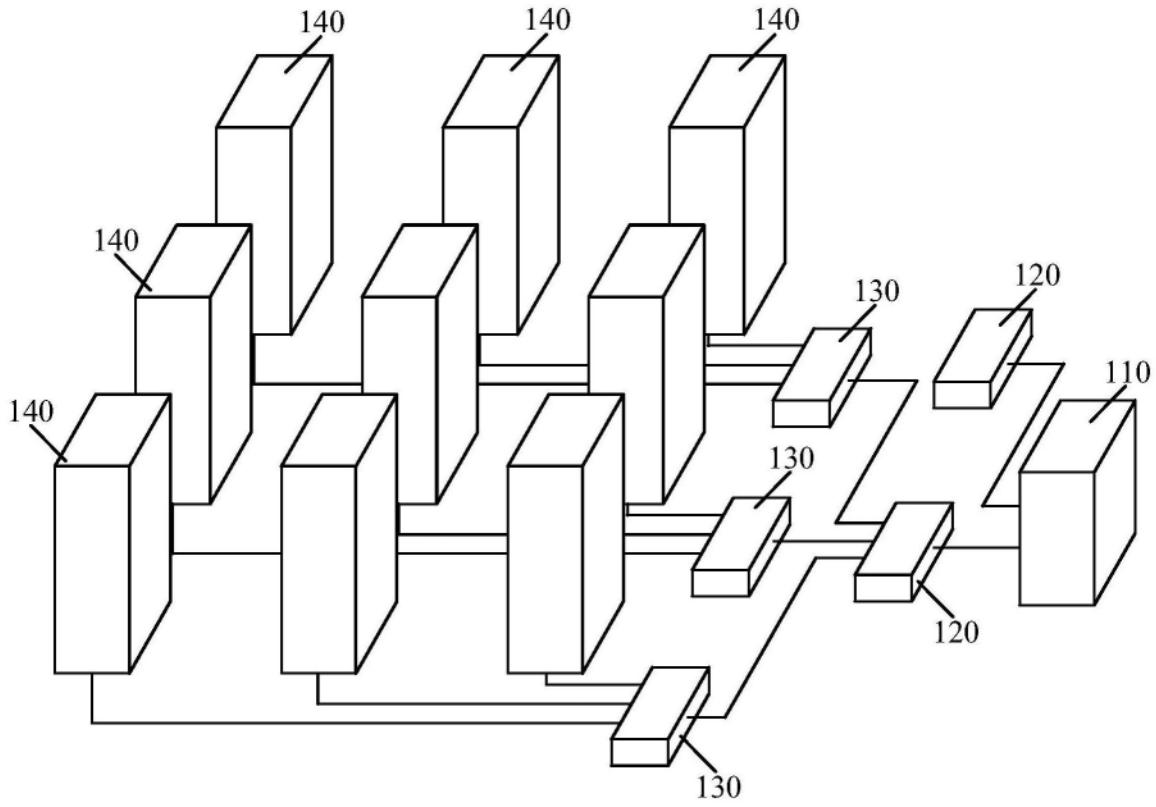


图1

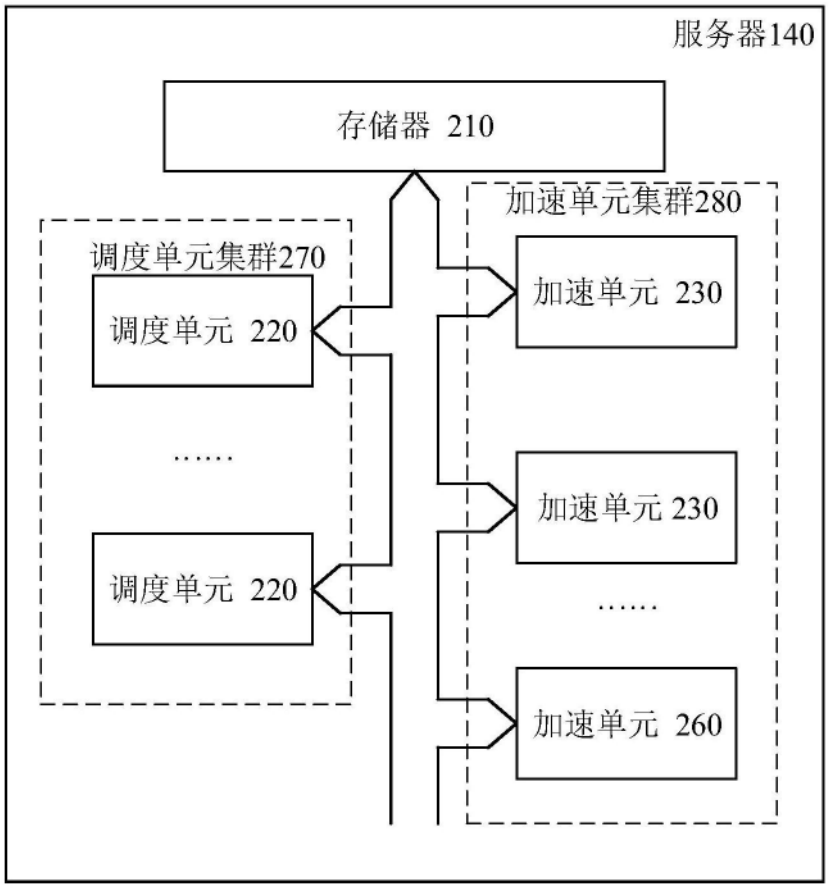

服务器140
存储器 210
加速单元集群280
调度单元集群270
加速单元 230
调度单元 220
……
加速单元 230
调度单元 220
……
加速单元 260

图2

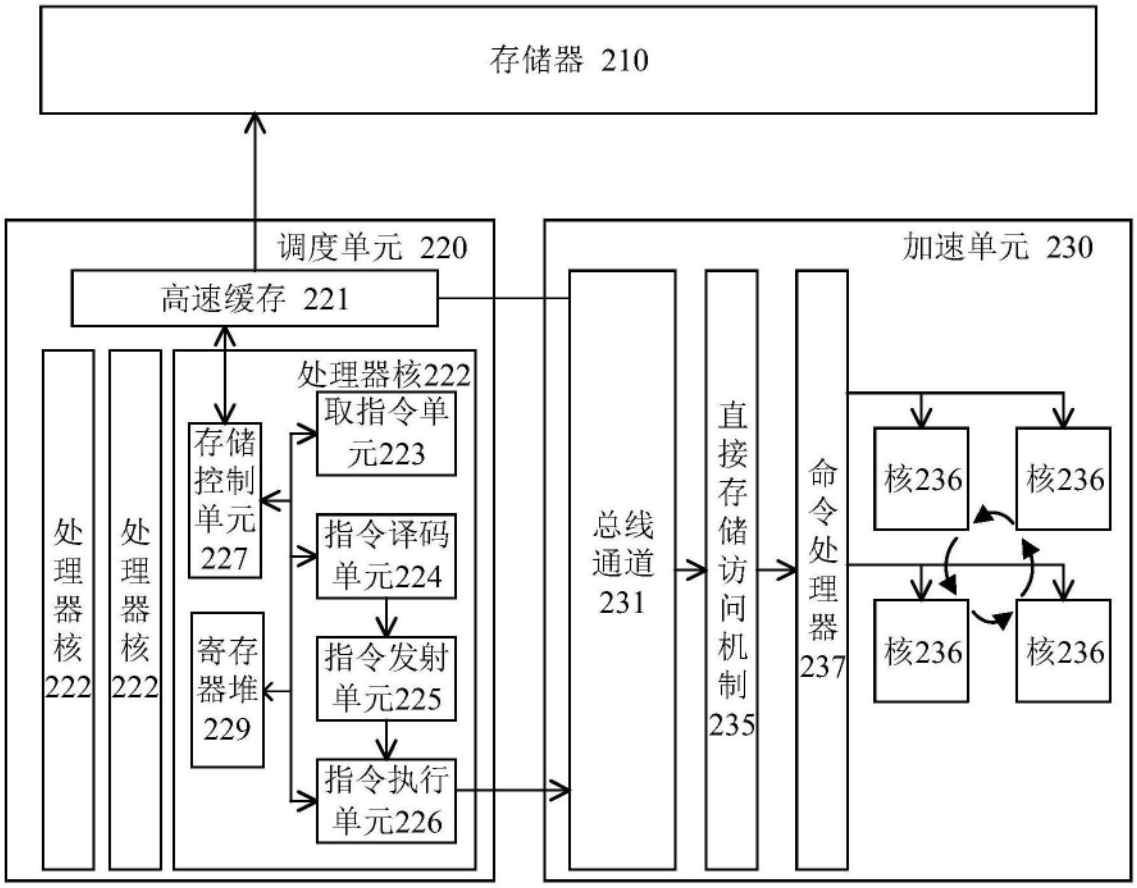

存储器 210
调度单元 220
高速缓存 221
处理器核222
处理器核222
处理器核222
取指令单元223
存储控制单元227
指令译码单元224
寄存器堆229
指令发射单元225
指令执行单元226
加速单元 230
总线通道231
直接存储访问机制235
命令处理器237
核236
核236
核236
核236

图3

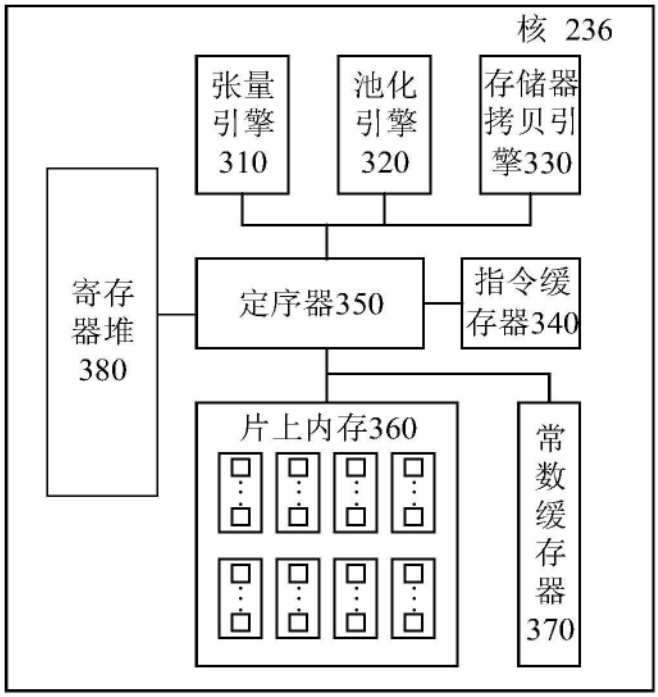

核 236
张量引擎310
池化引擎320
存储器拷贝引擎330
寄存器堆380
定序器350
指令缓存器340
片上内存360
常数缓存器370

图4

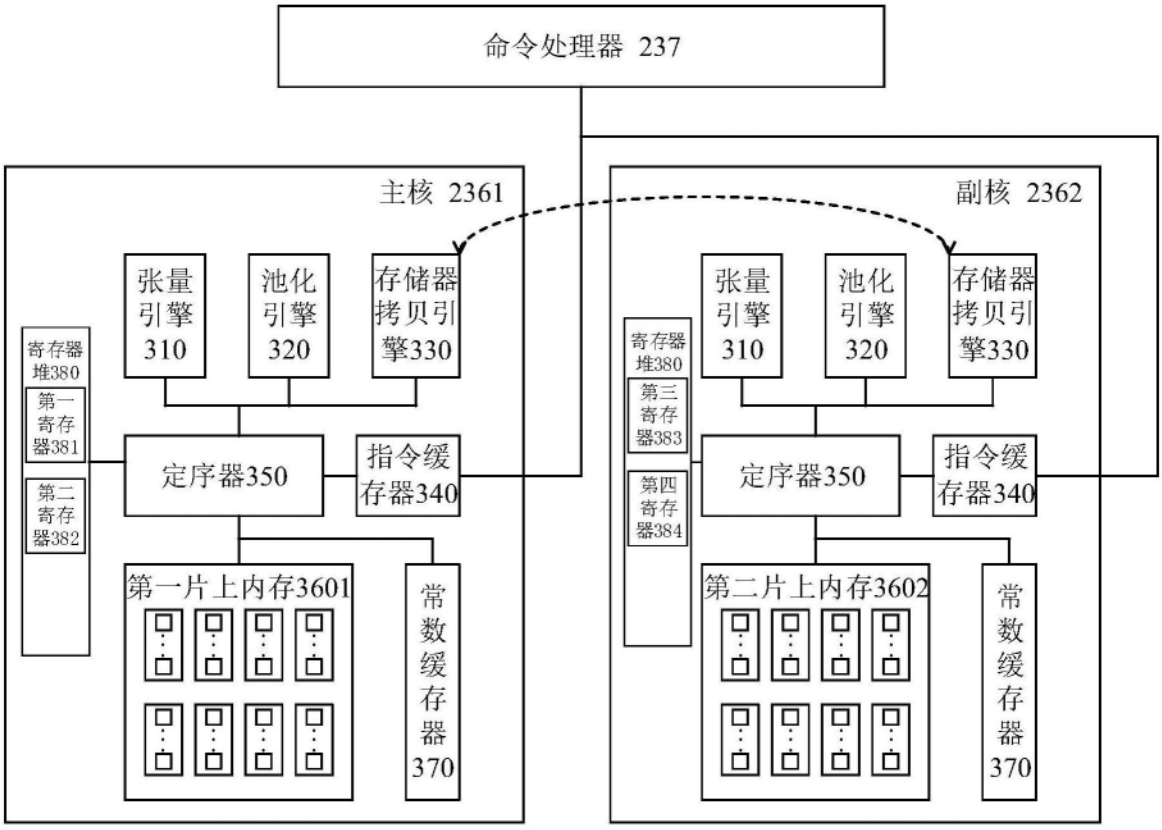

命令处理器 237
主核 2361
副核 2362
张量引擎310
池化引擎320
存储器拷贝引擎330
寄存器堆380
第一寄存器381
第二寄存器382
第三寄存器383
第四寄存器384
定序器350
指令缓存器340
第一片上内存3601
第二片上内存3602
常数缓存器370

图5

```
pool   r8, r0
conv   r1, r0
conv   r3, r2
conv   r5, r4
add    r6, r3, r5
mul    r7, r6, r0

pool   r14, r13
...
conv   r10, r1
rcp    r11, r10
```

图6A

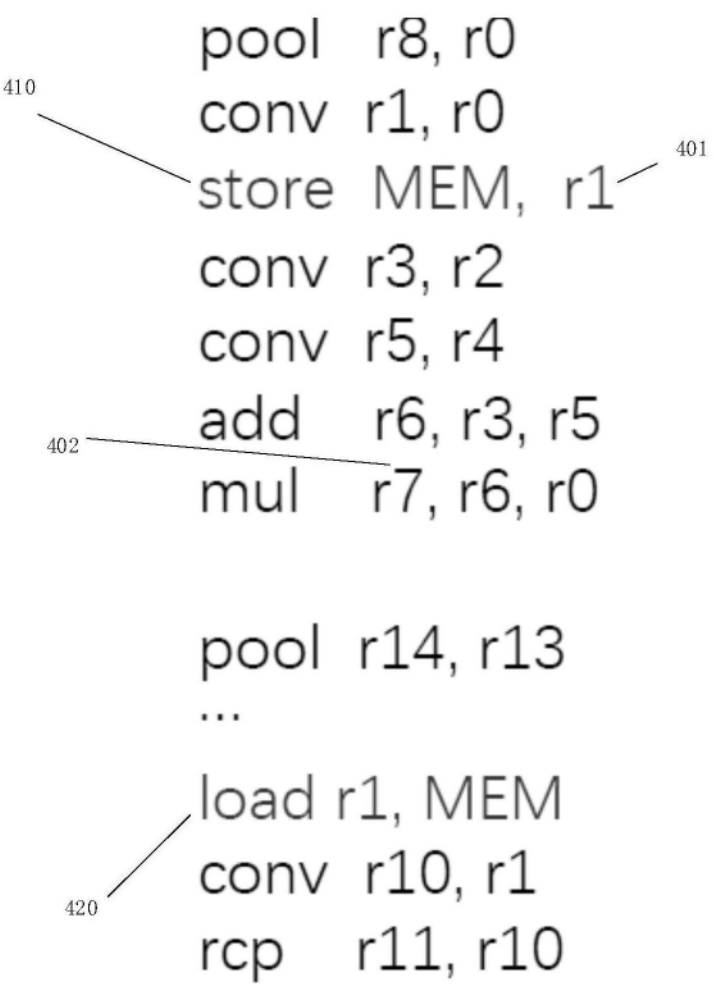

pool r8, r0
410
conv r1, r0
store MEM, r1
401
conv r3, r2
conv r5, r4
add r6, r3, r5
402
mul r7, r6, r0
pool r14, r13
...
load r1, MEM
420
conv r10, r1
rcp r11, r10

图6B

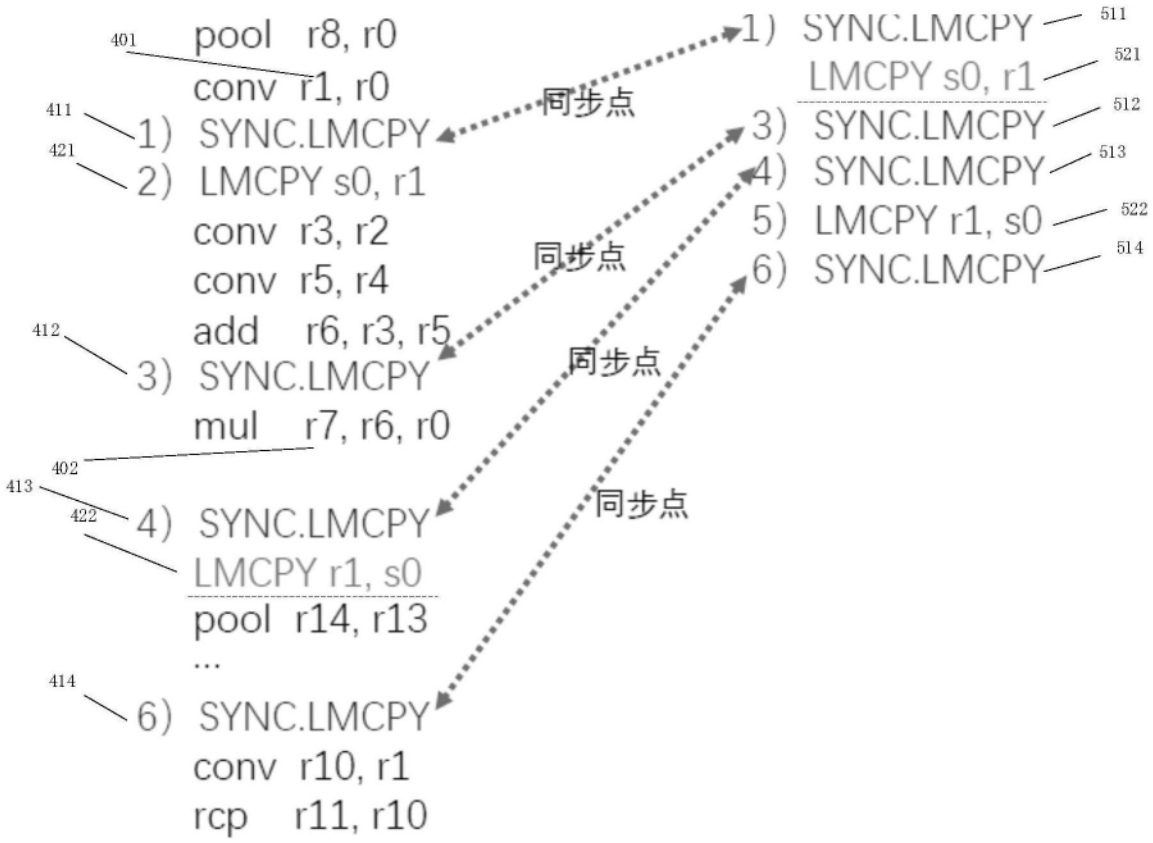

401
pool r8, r0
conv r1, r0
411
1) SYNC.LMCPY
421
2) LMCPY s0, r1
conv r3, r2
conv r5, r4
add r6, r3, r5
412
3) SYNC.LMCPY
mul r7, r6, r0
402
413
422
4) SYNC.LMCPY
LMCPY r1, s0
pool r14, r13
...
414
6) SYNC.LMCPY
conv r10, r1
rcp r11, r10
同步点
同步点
同步点
同步点
1) SYNC.LMCPY
511
LMCPY s0, r1
521
3) SYNC.LMCPY
512
4) SYNC.LMCPY
513
5) LMCPY r1, s0
522
6) SYNC.LMCPY
514

图6C

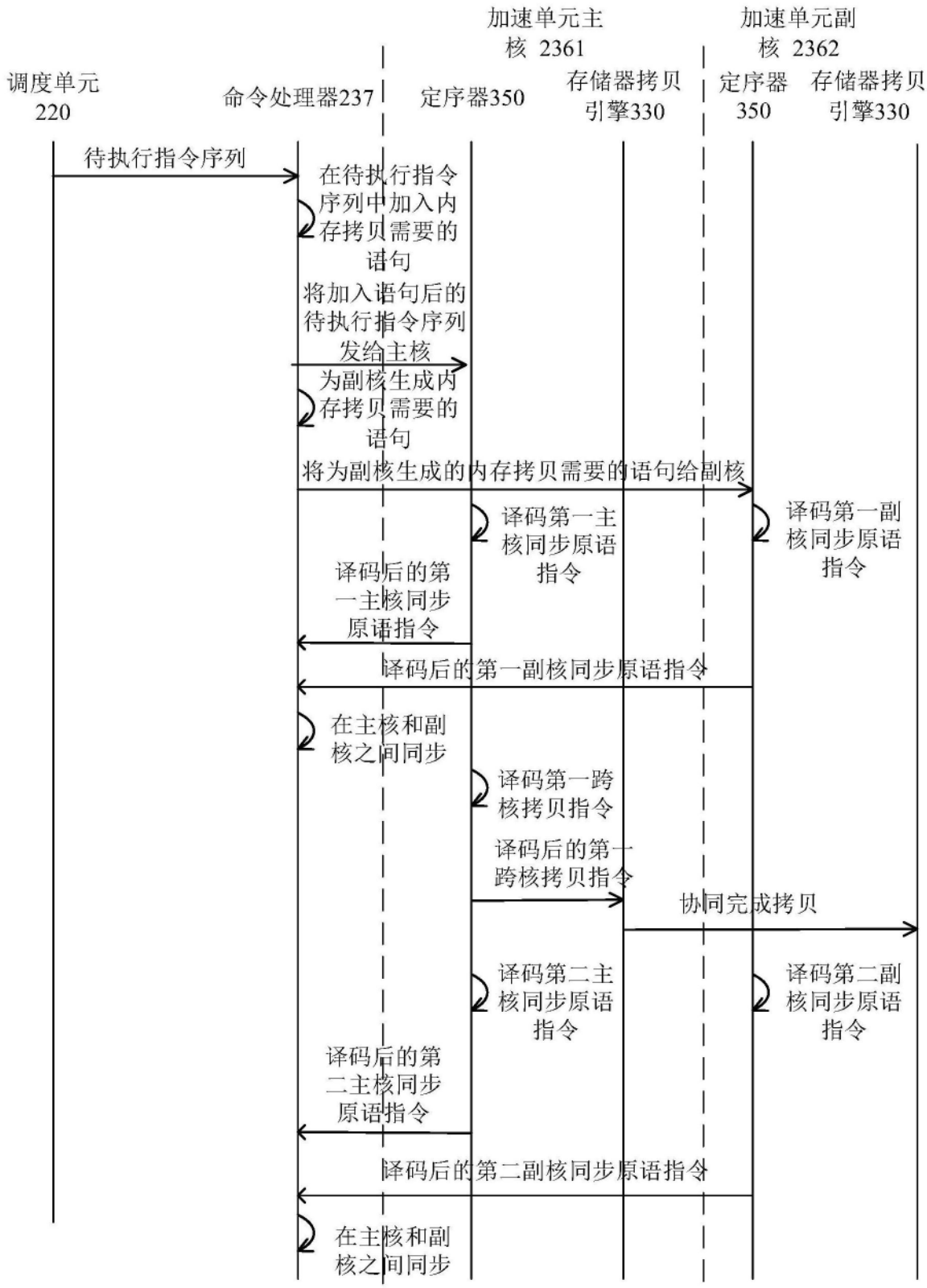

调度单元 220
命令处理器237
定序器350
加速单元主核 2361
存储器拷贝引擎330
定序器 350
加速单元副核 2362
存储器拷贝引擎330
待执行指令序列
在待执行指令序列中加入内存拷贝需要的语句
将加入语句后的待执行指令序列发给主核
为副核生成内存拷贝需要的语句
将为副核生成的内存拷贝需要的语句给副核
译码第一主核同步原语指令
译码第一副核同步原语指令
译码后的第一主核同步原语指令
译码后的第一副核同步原语指令
在主核和副核之间同步
译码第一跨核拷贝指令
译码后的第一跨核拷贝指令
协同完成拷贝
译码第二主核同步原语指令
译码第二副核同步原语指令
译码后的第二主核同步原语指令
译码后的第二副核同步原语指令
在主核和副核之间同步

图7A

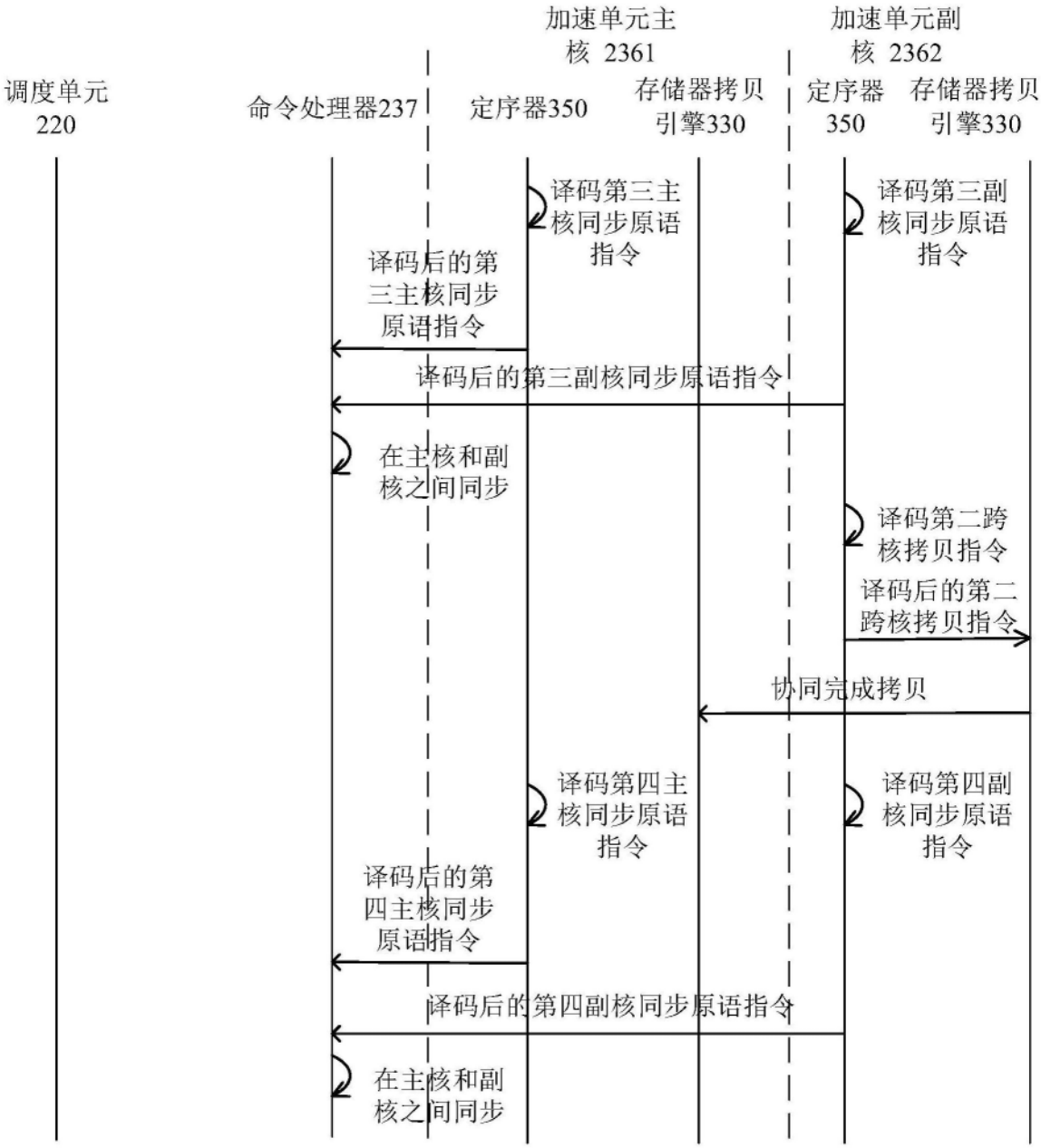

调度单元
220
命令处理器237
定序器350
加速单元主
核 2361
存储器拷贝
引擎330
定序器
350
加速单元副
核 2362
存储器拷贝
引擎330
译码第三主
核同步原语
指令
译码第三副
核同步原语
指令
译码后的第
三主核同步
原语指令
译码后的第三副核同步原语指令
在主核和副
核之间同步
译码第二跨
核拷贝指令
译码后的第二
跨核拷贝指令
协同完成拷贝
译码第四主
核同步原语
指令
译码第四副
核同步原语
指令
译码后的第
四主核同步
原语指令
译码后的第四副核同步原语指令
在主核和副
核之间同步

图7B

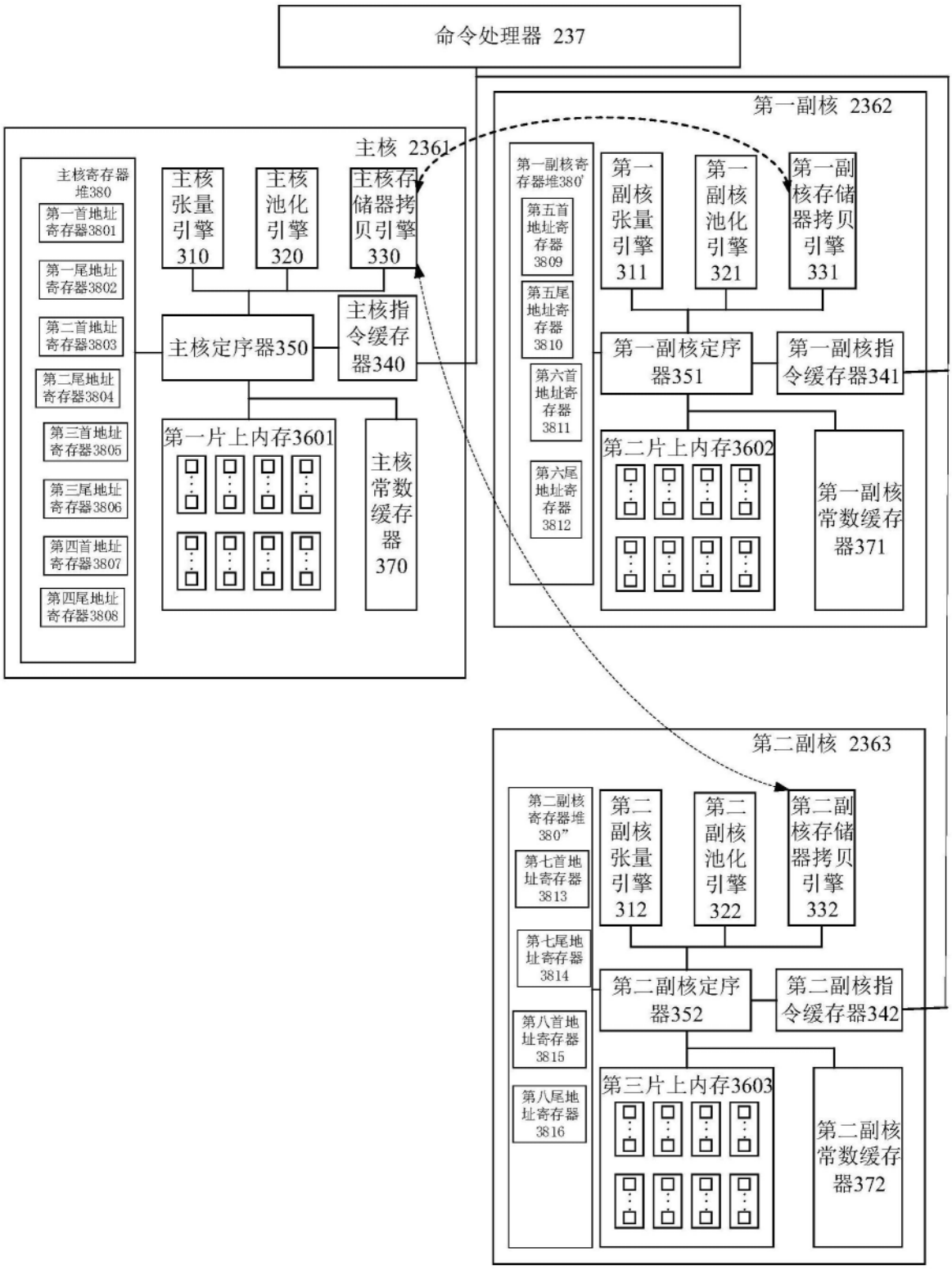

命令处理器 237
主核 2361
主核寄存器堆380
第一首地址寄存器3801
第一尾地址寄存器3802
第二首地址寄存器3803
第二尾地址寄存器3804
第三首地址寄存器3805
第三尾地址寄存器3806
第四首地址寄存器3807
第四尾地址寄存器3808
主核张量引擎310
主核池化引擎320
主核存储器拷贝引擎330
主核定序器350
主核指令缓存器340
第一片上内存3601
主核常数缓存器370
第一副核 2362
第一副核寄存器堆380'
第五首地址寄存器3809
第五尾地址寄存器3810
第六首地址寄存器3811
第六尾地址寄存器3812
第一副核张量引擎311
第一副核池化引擎321
第一副核存储器拷贝引擎331
第一副核定序器351
第一副核指令缓存器341
第二片上内存3602
第一副核常数缓存器371
第二副核 2363
第二副核寄存器堆380"
第七首地址寄存器3813
第七尾地址寄存器3814
第八首地址寄存器3815
第八尾地址寄存器3816
第二副核张量引擎312
第二副核池化引擎322
第二副核存储器拷贝引擎332
第二副核定序器352
第二副核指令缓存器342
第三片上内存3603
第二副核常数缓存器372

图8