



(12) 发明专利

(10) 授权公告号 CN 110348564 B

(45) 授权公告日 2021. 07. 09

(21) 申请号 201910500982.3

(22) 申请日 2019.06.11

(65) 同一申请的已公布的文献号
申请公布号 CN 110348564 A

(43) 申请公布日 2019.10.18

(73) 专利权人 中国人民解放军国防科技大学
地址 410073 湖南省长沙市开福区砚瓦池
正街47号

(72) 发明人 王蕾 窦强 邓宇 郭莎莎
王树泉 杨智杰 李石明 聂子凯
康子扬 田烁 曲连华

(74) 专利代理机构 湖南兆弘专利事务所(普通
合伙) 43008
代理人 谭武艺

(51) Int.Cl.

G06N 3/04 (2006.01)
G06N 3/063 (2006.01)
G06N 5/04 (2006.01)

(56) 对比文件

US 2018075338 A1, 2018.03.15
US 2018075339 A1, 2018.03.15

审查员 黄睿

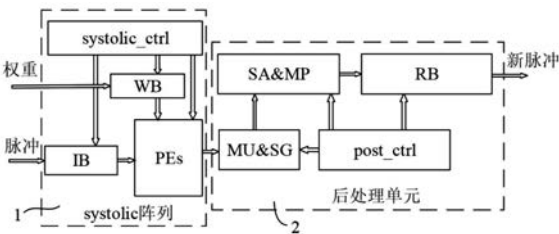
权利要求书2页 说明书6页 附图3页

(54) 发明名称

基于systolic阵列的SCNN推理加速装置、处理器及计算机设备

(57) 摘要

本发明公开了一种基于systolic阵列的SCNN推理加速装置、处理器及计算机设备,本发明的SCNN推理加速装置包括systolic阵列和后处理单元,systolic阵列包括脉冲缓存、权值缓存、systolic控制器、处理单元阵列,用于脉冲与权值点积,后处理单元包括膜电位更新与脉冲产生模块、脉冲累积与最大池化模块、结果缓存、后处理机制控制器。本发明能高效利用片上数据、减少片外访存压力,能够对SCNN卷积层、池化层推理过程进行加速,具有计算密度大、吞吐率高、支持流水并行的特点,能够减少不必要的片外数据搬移,最大限度的利用片上数据,实现对卷积层、池化层推理的硬件加速。



1. 一种基于systolic阵列的SCNN推理加速装置,其特征在于,包括systolic阵列(1)和后处理单元(2),所述systolic阵列(1)包括:

脉冲缓存(IB),用于输入脉冲缓存;

权值缓存(WB),用于权值缓存;

systolic控制器(systolic_ctrl),用于控制systolic阵列行为;

处理单元阵列(PEs),用于脉冲与权值点积;

所述后处理单元(2)包括:

膜电位更新与脉冲产生模块(MU&SG),用于膜电位更新及脉冲产生;

脉冲累积与最大池化模块(SA&MP),用于同一神经元的脉冲缓存以及池化操作;

结果缓存(RB),用于结果脉冲缓存;

后处理机制控制器(post_ctrl),用于控制后处理机制行为;

所述systolic控制器(systolic_ctrl)的控制输出端分别与脉冲缓存(IB)、权值缓存(WB)以及处理单元阵列(PEs)相连,所述脉冲缓存(IB)、权值缓存(WB)的输出端分别与处理单元阵列(PEs)的输入端相连,所述处理单元阵列(PEs)的输出端与后处理单元(2)的输入端相连;所述膜电位更新与脉冲产生模块(MU&SG)、脉冲累积与最大池化模块(SA&MP)、结果缓存(RB)依次相连,且所述膜电位更新与脉冲产生模块(MU&SG)、脉冲累积与最大池化模块(SA&MP)、结果缓存(RB)的控制端均与后处理机制控制器(post_ctrl)相连,所述膜电位更新与脉冲产生模块(MU&SG)根据处理单元阵列(PEs)的输出结果进行膜电位更新及脉冲产生,所述脉冲累积与最大池化模块(SA&MP)对膜电位更新与脉冲产生模块(MU&SG)输出的同一神经元的脉冲缓存以及池化操作,所述结果缓存(RB)用于将脉冲累积与最大池化模块(SA&MP)的输出缓存输出。

2. 根据权利要求1所述的基于systolic阵列的SCNN推理加速装置,其特征在于,所述处理单元阵列(PEs)包括以systolic阵列形式组织的多个处理单元(PE),每一个处理单元(PE)包括脉冲寄存器(spike_reg)、权重寄存器(weight_reg)、零值判断器、控制寄存器(control_reg)、累积寄存器(accumulate_reg)、乘法器、选择器(MUX)和输出寄存器(res_reg),上一级处理单元(PE)的脉冲输入激励(pre_spike)输入脉冲寄存器(spike_reg)并直接作为脉冲(spike)输出、权重输入激励(pre_weight)输入权重寄存器(weight_reg)并直接作为权重(weight)输出,零值判断器对脉冲寄存器(spike_reg)中的值进行判断并在脉冲寄存器(spike_reg)中的值为0时写控制寄存器(control_reg),在控制寄存器(control_reg)为1时乘法器将权重输入激励(pre_weight)、脉冲输入激励(pre_spike)进行乘法运算并输出至累积寄存器(accumulate_reg),且选择器(MUX)在控制信号(switch_in)的控制下选择将上一级处理单元(PE)的输出(pre_res)或累积寄存器(accumulate_reg)中的值写入输出寄存器(res_reg)并通过输出寄存器(res_reg)输出。

3. 根据权利要求1所述的基于systolic阵列的SCNN推理加速装置,其特征在于,所述膜电位更新与脉冲产生模块(MU&SG)包括:

历史膜电位存储器(RAM1),用于存储历史膜电位,且历史膜电位存储(RAM)每个块(bank)对应同一行中一个处理单元(PE);

当前膜电位寄存器(current_m_reg),用于缓存systolic阵列(1)输出的膜电位增量;

加法器,用于将systolic阵列(1)计算的膜电位增量与当前拍数对应的处理单元(PE)

的历史膜电位进行累加得到加和值；

累加膜电位寄存器(next_m_reg),用于存储加法器累加得到的加和值；

比较器(CMP),用于将累加膜电位寄存器(next_m_reg)保存的加和值与膜电位阈值进行比较,若加和值大于膜电位阈值,则以膜电位复位值(RESET_VALUE)更新历史膜电位存储器(RAM1)中存储的历史膜电位,并完成脉冲点火,输出脉冲“1”;若加和值小于膜电位阈值,则以加和值更新历史膜电位存储器(RAM1)中存储的历史膜电位,输出脉冲“0”；

脉冲生成器(spike_gen),用于根据比较器(CMP)的输出脉冲生成脉冲信号。

4.根据权利要求1所述的基于systolic阵列的SCNN推理加速装置,其特征在于,所述脉冲累积与最大池化模块(SA&MP)包括:

脉冲信号存储器(RAM2),用于按照指定的长度批量存储膜电位更新与脉冲产生模块(MU&SG)输出的脉冲信号；

最大池化层,用于将脉冲信号存储器(RAM2)中的各批脉冲信号进行逻辑或运算输出。

5.一种处理器,包括处理器本体,其特征在于,该处理器本体中设有权利要求1~4中任意一项所述的基于systolic阵列的SCNN推理加速装置。

6.一种计算机设备,其特征在于,该计算机设备中设有权利要求1~4中任意一项所述的基于systolic阵列的SCNN推理加速装置。

基于systolic阵列的SCNN推理加速装置、处理器及计算机设备

技术领域

[0001] 本发明涉及图像实时处理的脉冲卷积神经网络(SCNN)硬件加速引擎技术,具体涉及一种基于systolic阵列的SCNN推理加速装置、处理器及计算机设备。

背景技术

[0002] 人工神经网络是当前人工智能领域研究的热点。主流的方法是采用部分数据集对人工神经网络进行预先的训练(训练过程),而后将训练好的人工神经网络应用于实际的工作场景,利用其进行实际的图像分类、识别等操作(推理过程)。卷积神经网络(CNN)、脉冲神经网络(SNN)都属于人工神经网络。脉冲神经网络(SNN)主要通过模仿人脑神经元工作过程实现计算并行,减少功耗。传统的方法在训练SNN时存在网络性能弱,训练收敛慢的缺点,而卷积神经网络(CNN)直接转SNN的方法能够有效解决这一问题,该方法最终得到的SNN称为脉冲卷积神经网络(SCNN)。卷积神经网络(CNN)采用卷积核(Kernel)对图像进行特征识别,通过非线性函数过滤次要信息。在实际过程中还采用稀疏化方法减少神经元数目(池化操作)来提高计算性能。脉冲神经网络则通过对神经元间脉冲影响神经元内部膜电位更新,模仿生物神经元抑制兴奋机制,完成计算处理。在实际工作中,卷积神经网络(CNN)通过导数对层间误差进行修正,采用BP算法进行训练。脉冲卷积神经网络(SCNN)本身理论基于生物神经元工作原理,采用脉冲作为内部激励进行传导,本身无法求导。因此,采用拟生物训练算法进行训练。但是,该系列算法本身在训练上效率不高,收敛速度慢,最终得到的脉冲神经网络图像分类精度不高。针对这一情况,研究界有两种方案。一种是采用更加先进的拟生物的训练算法,进一步提高脉冲神经网络的训练效率,但目前进展甚微,另一种就是采用DNN转脉冲卷积神经网络(SCNN)的方式,通过间接方式保证脉冲卷积神经网络的训练效率,提高脉冲卷积神经网络的性能,以便高效的获取能够实际应用的脉冲神经网络。针对第二种方案,如何实现针对脉冲卷积神经网络(SCNN)进行硬件加速,尤其是采用同步电路对SCNN推理过程进行硬件加速,目前仍然是一项亟待解决的关键技术问题。

发明内容

[0003] 本发明要解决的技术问题:针对现有技术的上述问题,提供一种基于systolic阵列的SCNN推理加速装置、处理器及计算机设备,本发明采用同步电路实现,通过内部数据流共享减少片外(off-chip)数据搬移,能够支持片上膜电位更新、池化操作,具有计算密度大、并行度高、吞吐率高的特点。

[0004] 为了解决上述技术问题,本发明采用的技术方案为:

[0005] 一种基于systolic阵列的SCNN推理加速装置,包括systolic阵列和后处理单元,所述systolic阵列包括:

[0006] 脉冲缓存IB,用于输入脉冲缓存;

[0007] 权值缓存WB,用于权值缓存;

- [0008] systolic控制器systolic_ctrl,用于控制systolic阵列行为;
- [0009] 处理单元阵列PEs,用于脉冲与权值点积;
- [0010] 所述后处理单元包括:
- [0011] 膜电位更新与脉冲产生模块MU&SG,用于膜电位更新及脉冲产生;
- [0012] 脉冲累积与最大池化模块SA&MP,用于同一神经元的脉冲缓存以及池化操作;
- [0013] 结果缓存RB,用于结果脉冲缓存;
- [0014] 后处理机制控制器post_ctrl,用于控制后处理机制行为;
- [0015] 所述systolic控制器systolic_ctrl的控制输出端分别与脉冲缓存IB、权值缓存WB以及处理单元阵列PEs相连,所述脉冲缓存IB、权值缓存WB的输出端分别与处理单元阵列PEs的输入端相连,所述处理单元阵列PEs的输出端与后处理单元2的输入端相连;所述膜电位更新与脉冲产生模块MU&SG、脉冲累积与最大池化模块SA&MP、结果缓存RB依次相连且其控制端均与后处理机制控制器post_ctrl相连,所述膜电位更新与脉冲产生模块MU&SG根据处理单元阵列PEs的输出结果进行膜电位更新及脉冲产生,所述脉冲累积与最大池化模块SA&MP对膜电位更新与脉冲产生模块MU&SG输出的同一神经元的脉冲缓存以及池化操作,所述结果缓存RB用于将脉冲累积与最大池化模块SA&MP的输出缓存输出。
- [0016] 优选地,所述处理单元阵列PEs包括以systolic阵列形式组织的多个处理单元PE,每一个处理单元PE的包括脉冲寄存器spike_reg、权重寄存器weight_reg、零值判断器、控制寄存器control_reg、累积寄存器accumulate_reg、乘法器、选择器MUX和输出寄存器res_reg,上一级处理单元PE的脉冲输入激励pre_spike输入脉冲寄存器spike_reg并直接作为脉冲spike输出、权重输入激励pre_weight输入权重寄存器weight_reg并直接作为权重weight输出,零值判断器对脉冲寄存器spike_reg中的值进行判断并在脉冲寄存器spike_reg中的值为0时写控制寄存器control_reg,在控制寄存器control_reg为1时乘法器将权重输入激励pre_weight、脉冲输入激励pre_spike进行乘法运算并输出至累积寄存器accumulate_reg,且选择器MUX在控制信号switch_in的控制下选择将上一级处理单元PE的输出pre_res或累积寄存器accumulate_reg中的值写入输出寄存器res_reg并通过输出寄存器res_reg输出。
- [0017] 优选地,所述膜电位更新与脉冲产生模块MU&SG包括:
- [0018] 历史膜电位存储器RAM1,用于存储历史膜电位,且历史膜电位存储RAM每个块bank对应同一行中一个处理单元PE;
- [0019] 当前膜电位寄存器current_m_reg,用于缓存systolic阵列1输出的膜电位增量;
- [0020] 加法器,用于将systolic阵列1计算的膜电位增量与当前拍数对应的处理单元PE的历史膜电位进行累加得到加和值;
- [0021] 累加膜电位寄存器next_m_reg,用于存储加法器累加得到的加和值;
- [0022] 比较器CMP,用于将累加膜电位寄存器next_m_reg保存的加和值与膜电位阈值进行比较,若加和值大于膜电位阈值,则以膜电位复位值RESET_VALUE更新历史膜电位存储器RAM1中存储的历史膜电位,并完成脉冲点火,输出脉冲“1”;若加和值小于膜电位阈值,则以加和值更新历史膜电位存储器RAM1中存储的历史膜电位,输出脉冲“0”;
- [0023] 脉冲生成器spike_gen,用于根据比较器CMP的输出脉冲生成脉冲信号。
- [0024] 优选地,所述脉冲累积与最大池化模块SA&MP包括:

[0025] 脉冲信号存储器RAM2,用于按照指定的长度批量存储膜电位更新与脉冲产生模块MU&SG输出的脉冲信号;

[0026] 最大池化层,用于将脉冲信号存储器RAM2中的各批脉冲信号进行逻辑或运算输出。

[0027] 此外,本发明还提供一种处理器,包括处理器本体,该处理器本体中设有所述的基于systolic阵列的SCNN推理加速装置。

[0028] 此外,本发明还提供一种计算机设备,该计算机设备中设有所述的基于systolic阵列的SCNN推理加速装置。

[0029] 和现有技术相比,本发明具有下述优点:本发明基于systolic阵列的SCNN推理加速装置包括systolic阵列和后处理单元,systolic阵列包括脉冲缓存IB、权值缓存WB、systolic控制器systolic_ctrl、处理单元阵列PEs,用于脉冲与权值点积,后处理单元包括膜电位更新与脉冲产生模块MU&SG、脉冲累积与最大池化模块SA&MP、结果缓存RB、后处理机制控制器post_ctrl,采用systolic阵列计算神经元膜电位更新值,通过硬件模块实现批次池化,从而高效利用片上数据、减少片外访存压力,支持人工神经元模型中的集成与点火模型(IF),能够对SCNN卷积层、池化层推理过程进行加速,具有计算密度大、吞吐率高、支持流水并行的特点,本发明能够减少不必要的片外(off-chip)数据搬移,最大限度的利用片上数据,实现对卷积层、池化层推理的硬件加速。

附图说明

[0030] 图1为本发明实施例SCNN推理加速装置的基本结构示意图。

[0031] 图2为本发明实施例SCNN推理加速装置的详细结构示意图。

[0032] 图3为本发明实施例中处理单元的基本结构示意图。

[0033] 图4为本发明实施例中膜电位更新与脉冲产生模块MU&SG的结构示意图。

[0034] 图5为本发明实施例中脉冲累积与最大池化模块SA&MP的结构示意图。

具体实施方式

[0035] 如图1所示,本实施例基于systolic阵列的SCNN推理加速装置包括systolic阵列1和后处理单元2,systolic阵列1包括:

[0036] 脉冲缓存IB,用于输入脉冲缓存;

[0037] 权值缓存WB,用于权值缓存;

[0038] systolic控制器systolic_ctrl,用于控制systolic阵列行为;

[0039] 处理单元阵列PEs,用于脉冲与权值点积;

[0040] 后处理单元2包括:

[0041] 膜电位更新与脉冲产生模块MU&SG,用于膜电位更新及脉冲产生;

[0042] 脉冲累积与最大池化模块SA&MP,用于同一神经元的脉冲缓存以及池化操作;

[0043] 结果缓存RB,用于结果脉冲缓存;

[0044] 后处理机制控制器post_ctrl,用于控制后处理机制行为;

[0045] systolic控制器systolic_ctrl的控制输出端分别与脉冲缓存IB、权值缓存WB以及处理单元阵列PEs相连,脉冲缓存IB、权值缓存WB的输出端分别与处理单元阵列PEs的输

入端相连,处理单元阵列PEs的输出端与后处理单元2的输入端相连;膜电位更新与脉冲产生模块MU&SG、脉冲累积与最大池化模块SA&MP、结果缓存RB依次相连且其控制端均与后处理机制控制器post_ctrl相连,膜电位更新与脉冲产生模块MU&SG根据处理单元阵列PEs的输出结果进行膜电位更新及脉冲产生,脉冲累积与最大池化模块SA&MP对膜电位更新与脉冲产生模块MU&SG输出的同一神经元的脉冲缓存以及池化操作,结果缓存RB用于将脉冲累积与最大池化模块SA&MP的输出缓存输出。

[0046] 脉冲缓存IB,权值缓存WB,结果缓存RB只是单纯的缓存,本实施例省略其详细介绍。systolic控制器systolic_ctrl与后处理机制控制器post_ctrl内部均实现计数器,但值一样,以数拍数的方式控制两大部分(systolic阵列1与后处理机制2)行为。

[0047] 如图2和图3所示,处理单元阵列PEs包括以systolic阵列形式组织的多个处理单元PE,每一个处理单元PE的包括脉冲寄存器spike_reg、权重寄存器weight_reg、零值判断器、控制寄存器control_reg、累积寄存器accumulate_reg、乘法器、选择器MUX和输出寄存器res_reg,上一级处理单元PE的脉冲输入激励pre_spike输入脉冲寄存器spike_reg并直接作为脉冲spike输出、权重输入激励pre_weight输入权重寄存器weight_reg并直接作为权重weight输出,零值判断器对脉冲寄存器spike_reg中的值进行判断并在脉冲寄存器spike_reg中的值为0时写控制寄存器control_reg,在控制寄存器control_reg为1时乘法器将权重输入激励pre_weight、脉冲输入激励pre_spike进行乘法运算并输出至累积寄存器accumulate_reg,且选择器MUX在控制信号switch_in的控制下选择将上一级处理单元PE的输出pre_res或累积寄存器accumulate_reg中的值写入输出寄存器res_reg并通过输出寄存器res_reg输出。

[0048] 本实施例中,处理单元PE有4个输入,3个输出。其中“pre_”标号所示输入分别表示来自前一处理单元PE的脉冲、权值以及结果;switch_in输入用于控制累加结果传出;三个输出用于向下一处理单元PE传递脉冲、权值以及结果。注意,由于脉冲为“0”或“1”序列,脉冲与权值的点积可以通过选择器与加法器实现。具体计算过程中,使用一个处理单元PE处理一组脉冲与权值的点积。由于需要多拍才能得到最终的累加值,所有点积中间结果均通过寄存器记录在处理单元内部。通过systolic控制器systolic_ctrl内部的计数器记录拍数。当处理单元完成累加拍数后,switch_in输入控制处理单元将累加结果传出。结果传出时需要按照一拍一个处理单元的方式传出,最终在最后一列的处理单元获得最终结果。处理单元PE以systolic阵列形式组织,同行共享脉冲数据流,同列共享权值数据流。

[0049] 如图4所示,膜电位更新与脉冲产生模块MU&SG包括:

[0050] 历史膜电位存储器RAM1,用于存储历史膜电位,且历史膜电位存储RAM每个块bank对应同一行中一个处理单元PE;

[0051] 当前膜电位寄存器current_m_reg,用于缓存systolic阵列1输出的膜电位增量;

[0052] 加法器,用于将systolic阵列1计算的膜电位增量与当前拍数对应的处理单元PE的历史膜电位进行累加得到加和值;

[0053] 累加膜电位寄存器next_m_reg,用于存储加法器累加得到的加和值;

[0054] 比较器CMP,用于将累加膜电位寄存器next_m_reg保存的加和值与膜电位阈值进行比较,若加和值大于膜电位阈值,则以膜电位复位值RESET_VALUE更新历史膜电位存储器RAM1中存储的历史膜电位,并完成脉冲点火,输出脉冲“1”;若加和值小于膜电位阈值,则以

加和值更新历史膜电位存储器RAM1中存储的历史膜电位,输出脉冲“0”;

[0055] 脉冲生成器spike_gen,用于根据比较器CMP的输出脉冲生成脉冲信号。

[0056] 本实施例支持人工神经元模型中的集成与点火模型(IF模型)。该IF模型主要包括膜电位更新、脉冲产生两个功能过程。算法1对应膜电位更新过程,算法2对应脉冲产生过程。其中 m 表示膜电位, T 表示膜电位阈值, Δm 表示脉冲与权值的点积(即膜电位增量),SUM表示膜电位更新值与历史膜电位的加和值。

[0057] 算法1:膜电位更新过程:

[0058] 首先初始化(initialization),包括:膜电位 m_j^l (初始为0),阈值 T (初始为 Thr),膜电位增量 Δm_j^l (初始为0),复位膜电位 R (初始为RESET_VALUE),下标 j 表示循环变量(初始为0),上标 l 表示神经元层层数。

[0059] 然后逐个神经元层进行遍历,针对当前的第 j 次循环变量:(1)将上一层的膜电位 m_{j-1}^l 加上膜电位增量 Δm_j^l 得到膜电位累积量 $SUM(m_j^l)$; (2)针对膜电位累积量 $SUM(m_j^l)$ 进行下述处理(working out the added result):如果膜电位累积量 $SUM(m_j^l) \geq$ 阈值 T ,则将本层的膜电位 m_{j-1}^l 复位为复位膜电位 R ;如果膜电位累积量 $SUM(m_j^l) <$ 阈值 T ,则将本层的膜电位 m_{j-1}^l 更新为膜电位累积量 $SUM(m_j^l)$,上一层的膜电位 m_{j-1}^l 加上膜电位增量 Δm_j^l 的结果;最终存储本层的膜电位 m_{j-1}^l (membrane potential)至内存RAM中的对应的bank中。

[0060] 本实施例中,算法1以systolic阵列1计算得来的膜电位增量同对应的历史膜电位($j-1$ cycle的膜电位)进行加和,得到加和值;若该加和值大于阈值,则以加和值更新当前历史膜电位;若该加和值小于阈值,则以复位膜电位更新当前历史膜电位;而后循环前进一步。

[0061] 算法2:脉冲产生过程:

[0062] 首先初始化(initialization),包括:脉冲 $spike_j^l$ (初始为0),阈值 T (初始为 Thr),下标 j 表示循环变量(初始为0),上标 l 表示神经元层层数;

[0063] 然后针对来自算法1(Algorithm 1)的膜电位累积量 $SUM(m_j^l)$:如果膜电位累积量 $SUM(m_j^l) \geq$ 阈值 T ,则产生脉冲“1”($spike_j^l < -1$);如果膜电位累积量 $SUM(m_j^l) <$ 阈值 T ,则产生脉冲“0”($spike_j^l < -0$);最终将生成的脉冲传递给脉冲累积与最大池化模块SA&MP。

[0064] 膜电位更新与脉冲产生模块MU&SG主要对相应的两个功能过程(膜电位更新、脉冲产生)进行支持。使用RAM对历史膜电位进行存储,每个块(bank)对应同一行中一个处理单元。将systolic阵列计算的膜电位增量与当前拍数对应的处理单元的历史膜电位进行累加,得到加和值。以此加和值与膜电位阈值进行比较。若加和值大于膜电位阈值,则以膜电位复位值(RESET_VALUE)更新历史膜电位,并完成脉冲点火,输出“1”;若加和值小于膜电位阈值,则以加和值更新历史膜电位,输出脉冲“0”。

[0065] 如图5所示,脉冲累积与最大池化模块SA&MP包括:

[0066] 脉冲信号存储器RAM2,用于按照指定的长度批量存储膜电位更新与脉冲产生模块MU&SG输出的脉冲信号;

[0067] 最大池化层,用于将脉冲信号存储器RAM2中的各批脉冲信号进行逻辑或运算输出。

[0068] 由于SNN本身理论设计,需要将同一像素点转换为多组“0”或“1”序列。在算法上,表示对同一神经元进行多次脉冲与权值的点积。若不进行同一神经元的脉冲累积直接进行

池化操作,势必造成多次片外数据(同一神经元脉冲与权值)搬移。本实施例提出在片上对同一神经元的脉冲进行累积,而后批次进行池化操作。注意,由于平均池化(avg-pooling)操作涉及“0”“1”以外的数据表示,在SNN中无法实现,故只设计最大池化(max-pooling)操作。此处,最大池化操作本质为逻辑或运算,即范围内有脉冲则输出“1”,否则为“0”。

[0069] 此外,本实施例还提供一种处理器,包括处理器本体,该处理器本体中设有本实施例前述的基于systolic阵列的SCNN推理加速装置。

[0070] 此外,本实施例还提供一种计算机设备,该计算机设备中设有本实施例前述的基于systolic阵列的SCNN推理加速装置。

[0071] 以上所述仅是本发明的优选实施方式,本发明的保护范围并不局限于上述实施例,凡属于本发明思路下的技术方案均属于本发明的保护范围。应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理前提下的若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

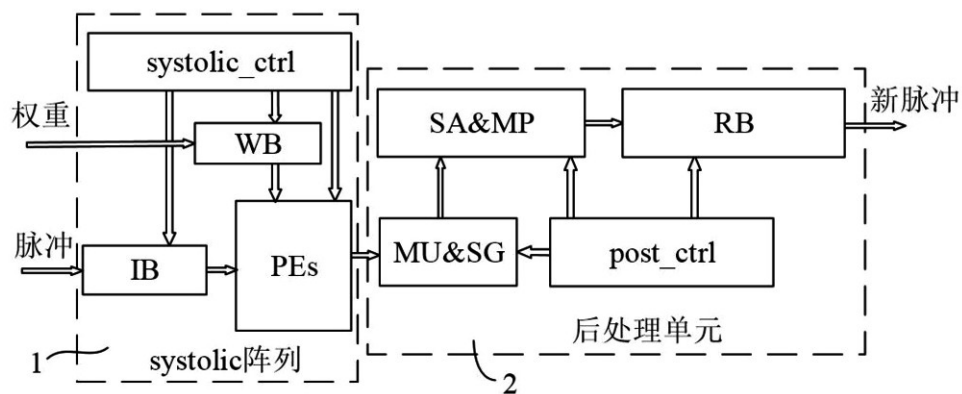


图 1

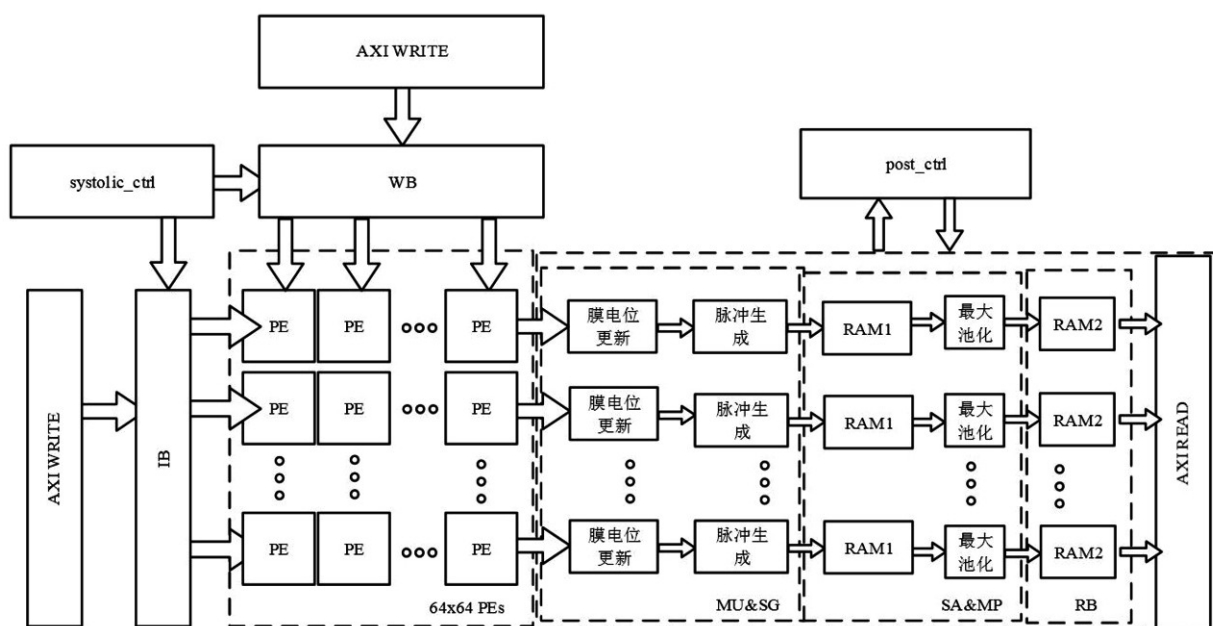


图 2

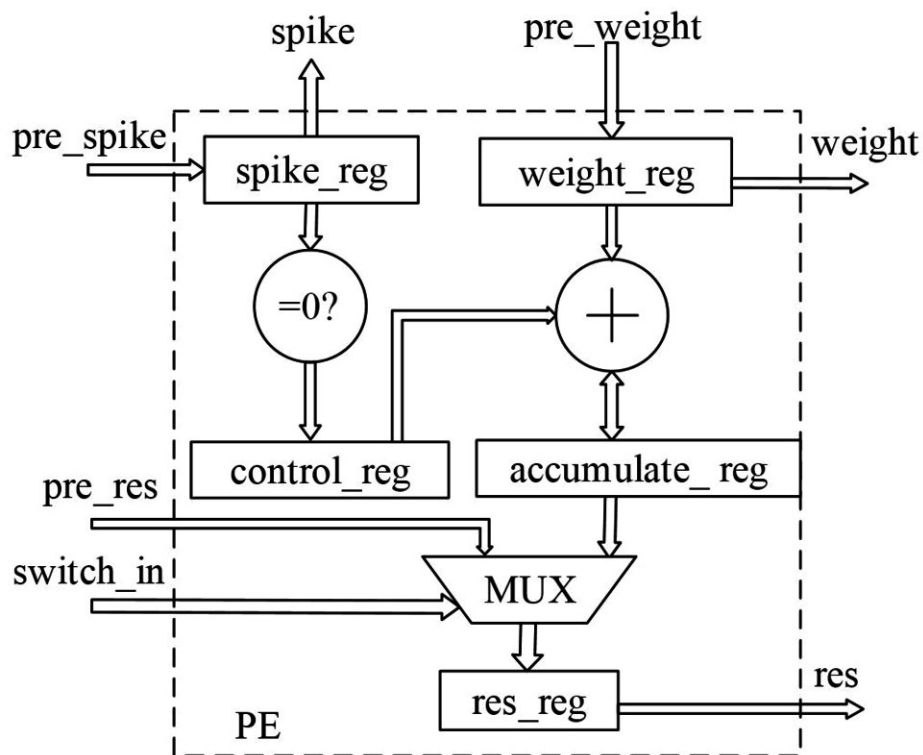


图 3

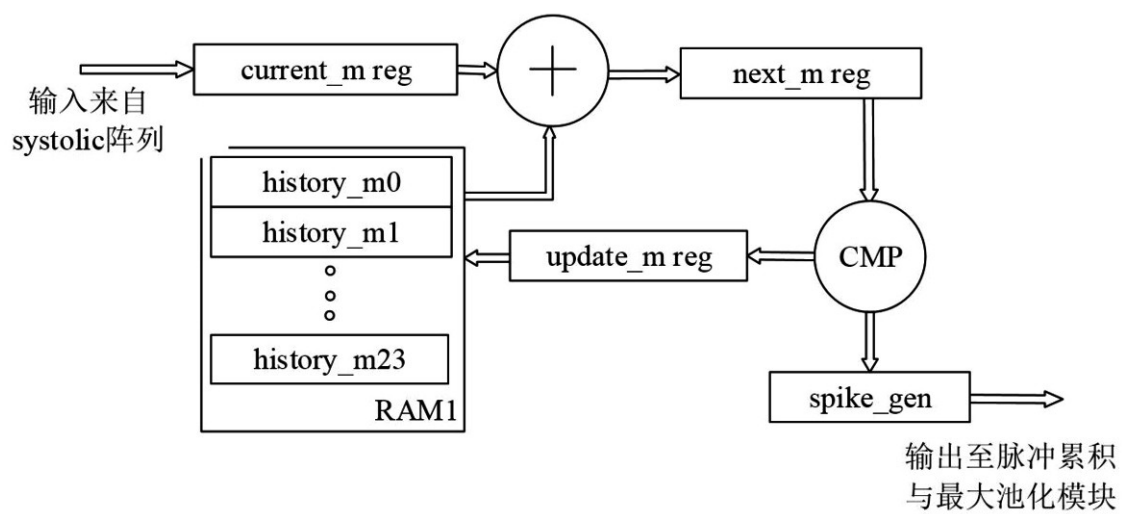


图 4

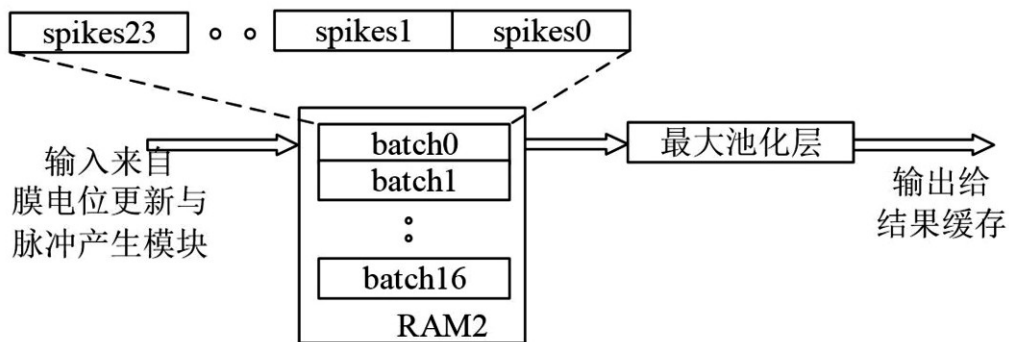


图 5