DOI:10.19651/j.cnki.emt.2211543

基于次级缓存的 SDRAM 调度策略的研究*

杜忠文[1,2] 李庚霖[1,2] 蒋 菡[1,2] 褚江恒[1,2] 伍 俊[2,3]
(1.重庆邮电大学光电工程学院 重庆 400065;2.中国科学院重庆绿色智能技术研究院跨尺度制造技术重庆市重点实验室 重庆 400722;3.中国科学院大学重庆学院 重庆 400714)

摘 要:针对卷积神经网络算法 FPGA 硬件加速器存在的内存带宽瓶颈,提出了一种基于次级缓存的行重组调度策略。通过分析 SDRAM 存储器的性能、FPGA 硬件加速原理和内存带宽瓶颈,建立了次级缓存机制。该机制可服务于加速过程中堆叠的访问请求,通过合并相同 Bank/Row 的访问请求,减少 Active 和 Precharge 操作的额外开销。实验测试结果表明,在 SC-RR 调度策略下,存储器的访存时间减少 32.87%,功耗降低 31.71%,有效带宽利用率提高到 91.3%。在性能相近的情况下,硬件资源消耗减少 83.8%,满足了设计要求。

关键词:卷积神经网络;FPGA;硬件加速;SDRAM;SC-RR
中图分类号:TP333 **文献标识码**:A **国家标准学科分类代码**:510.4030

Research on SDRAM scheduling strategy based on secondary cache

Du Zhongwen[1,2] Li Genglin[1,2] Jiang Han[1,2] Chu Jiangheng[1,2] Wu Jun[2,3]
(1. School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
2. Chongqing Key Laboratory of Cross-scale Manufacturing Technology, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400722, China;
3. Chongqing College, University of Chinese Academy of Sciences, Chongqing 400714, China)

Abstract: Aiming at the memory bandwidth bottleneck of FPGA hardware accelerator of convolutional neural network algorithm, this paper proposes a Secondary Cache-Row Recombination (SC-RR) based on secondary cache. By analyzing the performance of SDRAM memory, FPGA hardware acceleration principle and memory bandwidth bottleneck, a secondary cache mechanism is established. This mechanism can serve the stacked access requests during the acceleration process, reducing the additional overhead of Active and Precharge operations by merging access requests from the same Bank/Row. The experimental test results show that under the SC-RR scheduling strategy, the memory access time is reduced by 32.87%, the power consumption is reduced by 31.71%, and the effective bandwidth utilization is increased to 91.3%. In the case of similar performance, hardware resource consumption is reduced by 83.8%, which meets the design requirements.

Keywords: convolutional neural networks;FPGA;hardware acceleration;SDRAM;SC-RR

0 引 言

FPGA 因其具有高灵活、高性能、低功耗和开发周期短等优势广泛应用于卷积神经网络算法(convolutional neural network,CNN)硬件加速领域[1-12]。目前,卷积神经网络算法的 FPGA 硬件加速器的主要瓶颈之一是内存带宽不足,即 FPGA 与存储器之间的通信效率较低[13-15]。

目前,在卷积神经网络算法的 FPGA 硬件加速研究领域中。李沙沙等[2]通过数据复用,减少了对片外存储器的访问次数[2-4]。武世雄等[5]基于权重参数对内存访问请求进行重排序。该方法能够以较大突发长度进行传输,提高了有效带宽利用率[5-6]。陈浩敏等[9]在不损失准确率或准确率损失较小的情况下,用位数较低的定点数代替全精度浮点数,减少了片上内存的使用和对片外存储器的访存次数[7-9]。Zhang 等[10]提出 roofline 模型解决了计算吞吐量与内存带宽不匹配的问题。Lee 等[11]提出了基于 CNN 的 SR 的硬件高效数据流,实现了最低的内存使用量和最高的 PE 利用率。然而,目前的卷积神经网络算法 FPGA 硬件加速器主要研究加速器的计算性能,较少详细分析访存调度策略对内存带宽的影响。

收稿日期:2022-09-28
*基金项目:重庆英才创新领军人才项目(CQYC201903020)、重庆市杰出青年基金(cstc2019jcyjjqX0017)项目资助

本文针对 FPGA 加速器的内存带宽瓶颈进行了研究，提出了一种基于次级缓存的 SDRAM 行重组调度策略(secondary cache-row recombination，SC-RR)。该调度策略采用增加次级缓存的方式，将加速过程中堆叠的访问请求存储在次级缓存模块中，通过重新排序合并相同 Bank/Row 的访问请求，减少 Active 和 Precharge 操作带来的额外开销。通过 FPGA 硬件平台测试，详细分析了 SC-RR 调度策略下，SDRAM 访存过程的有效带宽的利用率、访存调度时间和存储器的功耗。

1　SDRAM 性能及 FPGA 硬件加速器分析

1.1　SDRAM 性能分析

因为 SDRAM 和 DDRx SDRAM 拥有类似的存储结构和访问机制，因此本文使用 SDRAM 代表其他类型的 SDRAM 存储器[16]。

1)存储器功耗分析

存储系统的功耗主要分为两个部分：动态功耗和静态功耗[17]。本节在量化分析了 SDRAM 芯片内部功耗的基础上，计算了开页和闭页策略下 SDRAM 访存调度的功耗[18]。SDRAM 访存操作具体功耗如图 1 所示。其中，e 是标准功耗，即完成单 Burst 写访问消耗的能量。深褐和浅褐 WR 矩形分别代表开页策略下，SDRAM 执行写操作消耗的能量以及 SDRAM 执行 Precharge 和 Active 操作消耗的能量；深褐和浅褐 ACT-PRE 矩形分别代表闭页策略下，SDRAM 执行写操作消耗的能量以及 SDRAM 执行 Precharge 和 Active 操作消耗的能量。横坐标分别表示相同的 Bank/Row 地址下，FPGA 对 SDRAM 分别进行传输长度为 1、2、3、4 的单独和连续突发访问消耗的能量，其中 C 表示连续访问。

当 FPGA 访问存储器内部存储单元时，单 Burst 突发访问的 ACT-PRE 消耗的能量占比超过 60%。在开页策略下，SDRAM 可以连续写入多个相同 Bank/Row 地址的访问队列，并且不会产生额外的 ACT-PRE 能量消耗。因此，在开页策略的 ACT-PRE 周期内，读取或写入两倍的数据量不会使能量消耗增加一倍，而只会增加大约 32%。并且，随着访问数据量的增加，该比例会进一步降低。

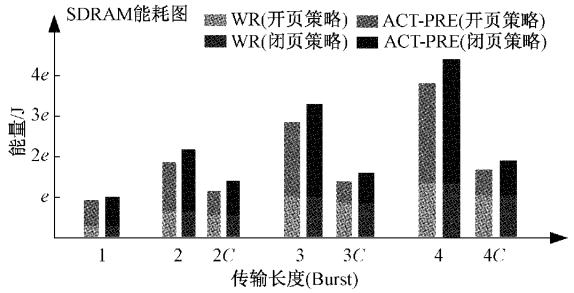

图 1　内存访问功耗

2)存储器有效带宽利用率分析

FPGA 在闭页策略下，对 SDRAM 进行传输长度为 1 突发写访问，具体流程如下：先执行 Active 命令激活需要访问的 Bank 和 Row；然后激活访问 Col 并写入数据；最后执行 Precharge 命令将打开的 Bank 和 Row 关闭。因此，在闭页策略下，FPGA 只能访问一次 SDRAM。而在开页策略下，FPGA 对 SDRAM 进行写访问会在 Active 激活 Bank 和 Row 并写入数据后，预判下一个访问请求。如果下一个访问请求的 Bank/Row 地址相同，则允许 FPGA 对 SDRAM 连续访问。反之，SDRAM 执行 Precharge 命令。因此，开页策略下，FPGA 可以连续访问 SDRAM，内存带宽有效利用率也更高。

SDRAM 访存时序如图 2 所示，其中 ACT、WR、RD、PRE、tCL、Data 分别表示：Active 激活 Bank 和 Row；写命令和写入的数据；读命令；Precharge 关闭 Bank 和 Row；写访问缓存时间；读出或写入数据。One Burst Wr 和 One Burst Rd 分别对应闭页策略下的 FPGA 对 SDRAM 的单次写访问和读访问。其中 ACT 和 PRE 占用了超过 80% 的有效带宽，带宽有效利用率分别是 14.2% 和 10%。Four Burst Wr 和 Six Burst Wr 分别表示开页策略下，FPGA 连续突发写访问 4 次和 6 次 SDRAM 的时序。其中连续 4 次突发写访问的有效带宽利用率能提高到 40%，连续 6 次突发访问有效带宽利用率能进一步提高到 50%。Six Burst Rd 和 Wr and Rd 分别表示 FPGA 连续 6 次突发读访问 SDRAM 和 FPGA 连续对 SDRAM 进行多次读写交叉访问的时序。它们的带宽有效利用率都达到了 40%。可以

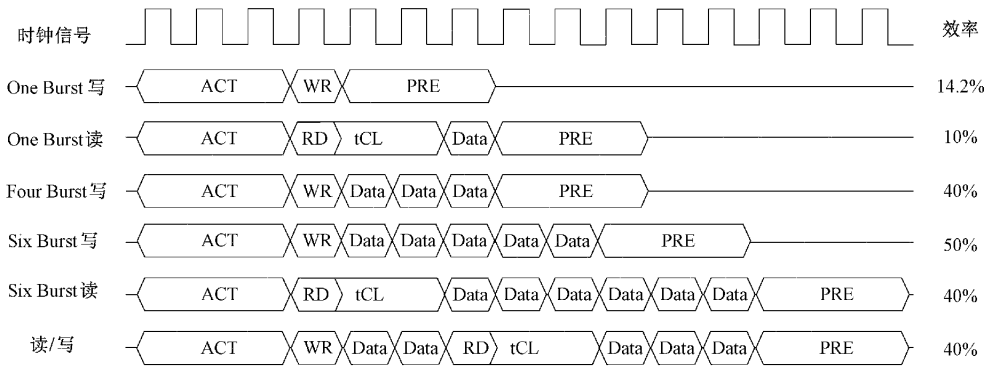

图 2　SDRAM 访存时序

看出,在开页策略下,FPGA 对 SDRAM 连续访问次数越多,有效带宽有效利用率越高。

1.2 FPGA 硬件加速器分析

1) FPGA 硬件加速原理

FPGA 硬件加速原理是利用 FPGA 的并行性实现对数据的多通道并行处理和流水线处理,从而实现对算法的硬件加速[12]。FPGA 硬件加速原理如图 3 所示。FPGA 并行处理:利用 FPGA 并行性,将电路模块复制 n 份,可以同时处理 n 数据。FPGA 流水线设计思想:利用 FPGA 的并行性,将数据处理流程对应的硬件结构分割,每个硬件结构负责单一的功能,如流水线一般独立工作。因此 FPGA 可以同时进行数据读取、计算、搬移等多个 CPU 不同阶段的操作,从而实现算法的硬件加速。

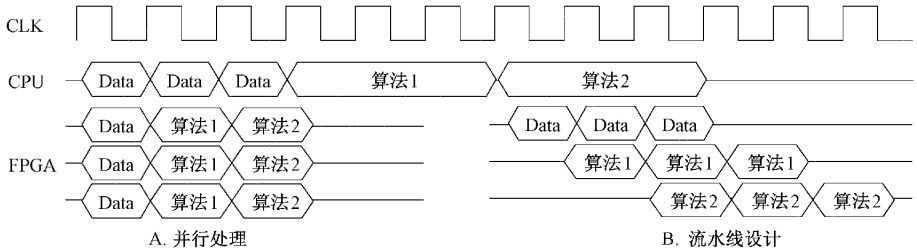

图 3 FPGA 硬件加速原理

在 FPGA 硬件加速过程中,同时实现并行处理数据和流水线处理数据是可行的[19]。这将进一步提高 FPGA 对数据的处理速度。因此,FPGA 能实现对卷积神经网络算法的硬件加速。

2)内存带宽限制

图 4 为 FPGA 访问 SDRAM 示意图,其中深灰矩形表示 FPGA 片上缓存数据,浅灰矩形表示 SDRAM 内部存储数据。研究发现,FPGA 硬件加速过程快,而访存调度过程慢。在 FPGA 硬件加速计算的过程中,计算结果先缓存在片上存储单元上,然后保存到片外存储器上。因为 SDRAM 内存带宽有限,FPGA 加速器的计算结果不能同步访存到 SDRAM。来不及存储到 SDRAM 的计算结果会消耗大量的片上存储单元,而 FPGA 的片上存储单元是有限的。因此为了避免 FPGA 的片上资源被无限制的消耗,FPGA 硬件加速器的计算性能上限会受到片外存储器的内存带宽限制。

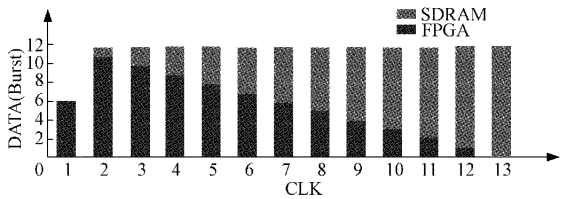

图 4 FPGA 访问 SDRAM 示意图

2 SC-RR 调度策略

2.1 SDRAM 控制器结构

如图 5 所示为 SDRAM 控制器的基本结构,SDRAM 控制器包含物理层设计和传输层设计。物理层为 SDRAM_Enable 模块,实现 SDRAM 存储器的物理接口驱动。传输层包含 3 个模块,其中数据接收模块接收上位机的访问请求和数据;数据通道模块实现数据缓存和跨时钟域信号传输;仲裁调度模块实现对访问请求的仲裁调度。

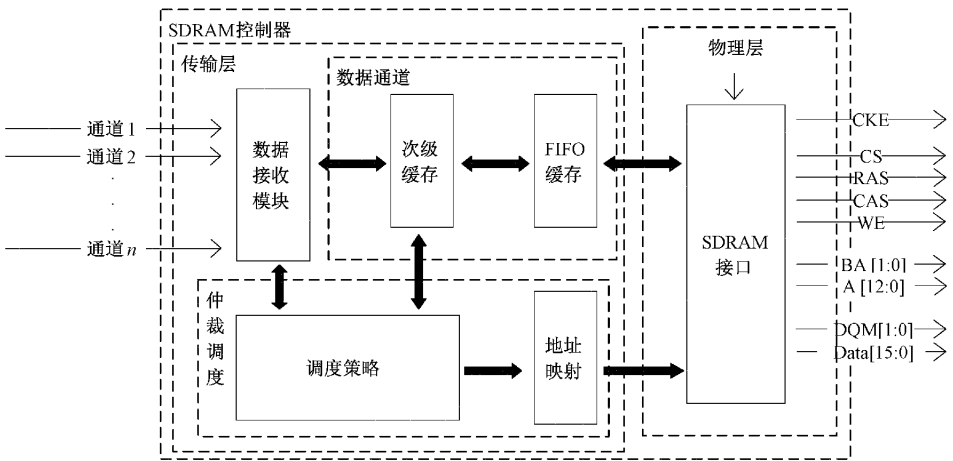

图 5 SDRAM 控制器硬件结构

2.2　次级缓存基本思想

在第 1.1 节中研究了存储器的性能。从带宽、访问时间和能耗分析来看，在开页策略下，SDRAM 连续写入相同 Bank/Row 的访问请求，能减少大量的 ACT 和 PRE 操作，提高有效带宽的利用率、减少访存调度时间和功耗。在 1.2 节中分析内存带宽瓶颈发现，存储器内存带宽是有限的。但提高带宽的有效利用率，可以在有限的带宽限制下，传输更多的数据。从而提高 FPGA 硬件加速器的计算性能上限。因此，提出了次级缓存的基本思想，希望能够将堆叠的访问请求进行缓存，并进行预判重组，提高有效带宽的利用率。

图 6 所示为次级缓存结构。SC 为次级缓存模块，能缓存访问请求的命令、地址和传输数据缓存和传输数据长度等信息。FIFO 缓存模块用于实现跨时钟域信号传输，其 FIFO 深度为 512。MUX 模块根据 SC-RR 调度策略，将次级缓存模块中的访问请求缓存到 FIFO 模块，将 FIFO 模块的缓存数据传输 SDRAM。Time-cnt 负责记录次级缓存模块的等待时间。SABT 模块缓存了次级缓存模块的 Bank/Row 地址和等待时间。

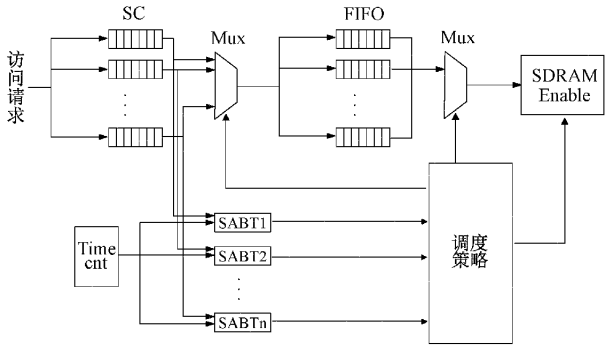

图 6　次级缓存结构

2.3　SC-RR 调度策略

1）SC-RR 调度基本思想

In-Order 调度策略根据访问请求的先后顺序，依次对 SDRAM 进行访问[18]。虽然这种调度策略几乎所有的内存控制器都在使用，但其带宽有效利用率较低。为了提高带宽的有效利用率，在次级缓存思想的基础上，本节提出了 SC-RR 调度策略。该策略的具体过程为：首先对次级缓存模块的访问请求进行预判，将其中相同 Bank/Row 地址的访问请求队列合并，按照 In-Order 调度进行重新排序。然后在开页策略下，通过 FIFO 实现对 SDRAM 的连续访问。图 7 为 In-Order 调度策略和 SC-RR 调度策略的时序。

2）SC-RR 调度流程

SC-RR 调度策略包含两个部分：对次级缓存模块的调度和对 FIFO 模块的调度。次级缓存的 SC-RR 调度是对 Cop_CC 和 Cop_ABT 两个子模块进行判断。Cop_CC 负责判断 FIFO 缓存的 3 个状态：畅通、准入和排队状态。Cop_ABT 负责对次级缓存中所有访问的命令、地址、传输长度和等待时间进行判断，并将结果反馈给 MUX、Cop_BT、SC 和 SABT 模块。Cop_BT 模块负责判断 FIFO 和 SDRAM 的状态，并将结果反馈给 FIFO、FBT 和 SDRAM_Enable 子模块。

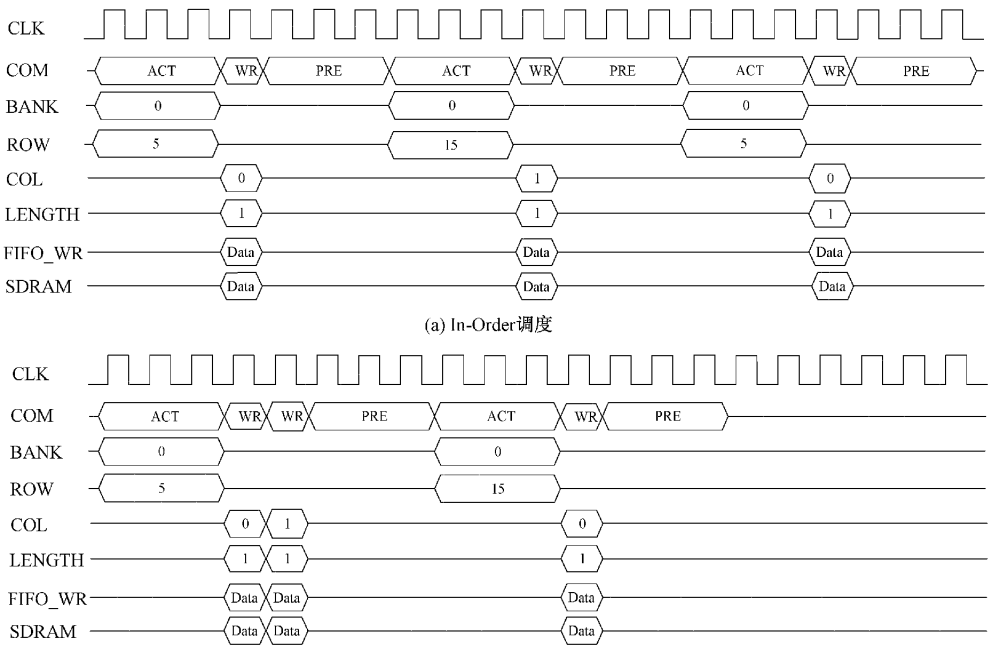

图 7　In-Order 与 SC-RR 时序

如图 8 所示,SC-RR 调度流程分 3 个步骤:1) FIFO 忙碌现状分析:比较 CC,若 $CC<4$,表示 FIFO 模块处于畅通状态,访问请求跳过次级缓存模块直接写入 FIFO 缓存模块;若 $4\leqslant CC<5$,准入状态,表示 FIFO 允许访问请求队列进入,次级缓存模块的访问请求队列经过 SC-RR 调度缓存到 FIFO 中;若 $CC=5$,FIFO 模块处于排队等待状态,所有访问请求写入次级缓存模块排队等待。2)次级缓存调度:若存在读访问请求,则读请求先行;若存在多个读访问请求,则执行 In-Order 调度;然后比较所有次级缓存的访问请求队列,将相同 Bank/Row 地址数最多的次级缓存队列合并;按照 In-Order 调度策略缓存到 FIFO 模块中。3)FIFO 缓存模块调度:FIFO 模块由前后两个状态机控制,前状态机控制 SC 的访问请求队列写入 FIFO 模块,后状态机控制 FIFO 的访问请求队列写入 SDRAM。当 SDRAM 处于空闲状态时,SC-RR 判断 FIFO 缓存模块空满状态和传输状态;按照 In-Order 调度策略,将 FIFO 缓存的访问请求队列写入 SDRAM。

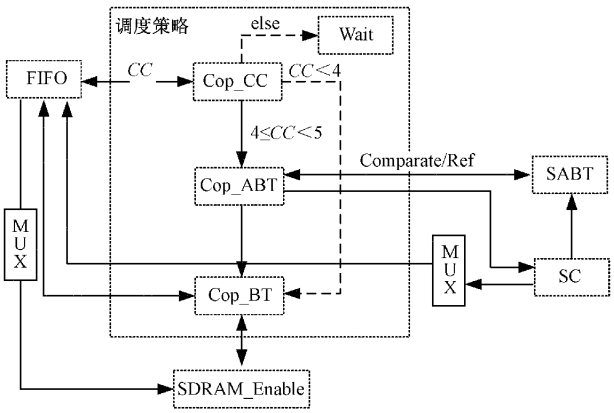

图 8 SC-RR 调度流程

3 SC-RR 调度策略测试与分析

3.1 测试方案设计

实验采用 Intel 公司的 Cyclone 10 LP 芯片搭建 FPGA 硬件平台,在实验平台搭建 CNN 硬件加速器仿真模型。本文根据文献[2]和[5]所述 CNN 硬件加速器设计方案,在 FPGA 硬件平台上搭建 CNN 硬件加速器等效模型进行测试。将 50 帧图像导入模型,进行运算,运用不同的调度策略实现 FPGA 和 SDRAM 的数据传输。

3.2 硬件资源消耗

本文根据文献[2]和[5]中的 CNN 硬件加速器设计原理和参数,设计 Model-5 和 Model-8 等效仿真模型。CNN 硬件加速器模型的硬件资源消耗量如表 1 所示。发现,Model-5 模型的整体硬件资源的消耗比 Model-8 模型多。

发现文献[2]缺少对 SDRAM 调度策略的详细阐述。因此,根据文献[2]的 CNN 硬件加速器运行时间和 DDR 空闲时间,设计了 Model-5 等效调度策略。SC-RR 调度策略的次级缓存模块,相比 Model-5 增加了少量的逻辑单元和寄存器单元消耗和较多的内存空间资源消耗。本文根据文献[5]设计了 Model-8 等效调度策略,先将需要重排序的访问请求缓存,然后进行重排序。不同调度策略在 FPGA 实验平台上的硬件资源消耗量如表 2 所示。其中,SC-RR 调度策略消耗的硬件资源为 Model 的 1.405 倍,Model-8 消耗的硬件资源为 SC-RR 调度策略的 6.177 倍。

表 1 仿真模型的 FPGA 资源消耗

MODEL	LUT	FF	BRAM	DSP	LUTRAM
Model-5	95 366	178 139	225	835	8 330
Model-8	39 397	42 183	260	172	7 238

表 2 调度策略资源消耗

Strategy	Logic element	Register	Memory
Model-5	1 331	548	41 384
Model-8	2 198	1 229	359 296
SC_RR	1 427	743	58 165

3.3 测试结果分析

基于 Model-5 模型对 50 帧 289×386 阵列规模的红外图像进行了加速处理,分别用 Model-5 和 SC-RR 调度策略实现 FPGA 到 SDRAM 的数据传输。在 Model-5 和 SC-RR 调度策略下,Model-5 等效模型平均处理一帧图像数据,SDRAM 的访存时间、功耗和有效带宽利用率,如表 3 所示。SC-RR 调度策略下,访存时间约为 6.69 ms 比 Model-5 减少 32.87%。功耗为 3.92 万标准功耗比 Model-5 降低了 31.71%。有效带宽利用率高于 Model-5 调度策略,达到了 91.3%。

表 3 Model-5 模型的帧测试数据

调度策略	访存时间/μs	功耗(e)	有效带宽利用率/%
Model-5	9 965	10 1521	63.2
SC_RR	6 689	69 329	91.3

基于 Model-8 模型对 50 帧 289×386 阵列规模的红外图像进行了加速处理,分别用 Model-8 和 SC-RR 调度策略实现 FPGA 到 SDRAM 的数据传输,记录两种调度策略的访存时间、功耗和有效带宽利用率。在 Model-8 和 SC-RR 调度策略下,Model-8 等效模型平均处理一帧图像数据的访存时间、功耗和有效带宽利用率如表 4 所示。SC-RR 调度策略下,SDRAM 的访存时间约为 3.86 ms 略高于 Model-8 调度策略访存时间。功耗为 3.87 万标准功耗略高于 Model-8 调度策略功耗。有效带宽利用率为 91.7% 略低于 Model-8 调度策略。

分析不同的加速器模型的调度策略测试结果发现,

表 4　Model-8 的帧测试数据

调度策略	访存时间/μs	功耗(e)	有效带宽利用率/%
Model-8	3 782	35 893	93.4
SC_RR	3 862	38 712	91.7

SC-RR 调度策略能有效降低文献[2]中 SDRAM 的访存时间、功耗，并提高有效带宽利用率。SC-RR 调度策略在总体性能与文献[5]相近的情况下，降低 83.8%的硬件资源消耗。

4　结　　论

本文基于 FPGA 硬件平台设计一种针对卷积神经网络硬件加速器的 SDRAM 调度策略，通过次级缓存模块存储访问请求，并对这些访问请求进行行重组。实验结果表明，在不同类型的 CNN 硬件加速器中，SC-RR 调度策略都能保持较低的访存时间，功耗和较高的有效带宽利用率。未来，我们希望能在 CNN 加速器的硬件系统设计上改进，进一步打破 CNN 硬件加速器的瓶颈。

参考文献

[1] CHEN Y H, FAN C P, ROBERT C. Prototype of low complexity CNN hardware accelerator with FPGA-based PYNQ platform for dual-mode biometrics recognition[J]. 2020 International SoC Design Conference, 2020: 189-190.

[2] 李沙沙，李夏禹，刘珊珊，等. 一种基于 FPGA 的通用卷积神经网络加速器的设计与实现[J]. 复旦学报(自然科学版)，2022，61(1)：69-76，84.

[3] MA Y, CAO Y, VRUDHULA S, et al. Optimizing the convolution operation to accelerate deep neural networks on FPGA[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2018, 26(7): 1354-1367.

[4] MOTAMEDI M, GYSEL P, AKELLA V, et al. Design space exploration of FPGA-based deep convolutional neural networks[J]. 2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC), 2016: 575-580.

[5] 武世雄，高巍，尹震宇，等. 面向 ZYNQ SoC 的卷积神经网络加速器研究[J]. 小型微型计算机系统，2022，1-8.

[6] 谢坤鹏，卢冶，靳宗明，等. FAQ-CNN：面向量化卷积神经网络的嵌入式 FPGA 加速框架[J]. 计算机研究与发展，2022，59(7)：1409-1427.

[7] 梁修壮，倪伟. FPGA 加速器深度卷积神经网络优化计算方法[J]. 计算机仿真，2022，39(5)：314-318.

[8] 夏琪迪，颜秉勇，周家乐，等. 基于异构 FPGA 的目标检测硬件加速器架构设计[J]. 华东理工大学学报(自然科学版)，2021，47(6)：706-715.

[9] 陈浩敏，姚森敬，席禹，等. YOLOv3-tiny 的硬件加速设计及 FPGA 实现[J]. 计算机工程与科学，2021，43(12)：2139-2149.

[10] ZHANG C, LI P, SUN G, et al. Optimizing FPGA-based accelerator design for deep convolutional neural networks[C]. Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Pages 161-170, ACM, 2015.

[11] LEE S, JOO S, AHN H K, et al. CNN acceleration with hardware-efficient dataflow for super-resolution[J]. In IEEE Access, 2020(8): 187754-187765.

[12] 吴宇航，何军. 基于 FPGA 加速的行为识别算法研究[J]. 电子测量技术，2022，45(13)：25-32.

[13] 吴艳霞，梁楷，刘颖，等. 深度学习 FPGA 加速器的进展与趋势[J]. 计算机学报，2019，42(11)：2461-2480.

[14] 刘腾达，朱君文，张一闻. FPGA 加速深度学习综述[J]. 计算机科学与探索，2021，15(11)：2093-2104.

[15] 彭宇，姬森展，于希明，等. 语义分割网络的 FPGA 加速计算方法综述[J]. 仪器仪表学报，2021，42(9)：1-12.

[16] 王树争. SDRAM 的发展历程[J]. 电脑知识与技术，2017，13(15)：213-214.

[17] CHANDRASEKAR K, AKESSON B, GOOSSENS K. Run-time power-down strategies for real-time SDRAM memory controllers[J]. In Proceedings of the 49th Annual Design Automation Conference (DAC' 12), 2012: 988-993.

[18] LEE Y S, HAN A T. Task parallelism-aware deep neural network scheduling on multiple hybrid memory cube-based processing-in-Memory[J]. IEEE Access, 2021, 9: 68561-68572.

[19] LIU S L, FAN H X, FERIANC M, et al. Toward full-stack acceleration of deep convolutional neural networks on FPGAs[J]. In IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(8): 3974-3987.

作者简介

杜忠文，硕士研究生，主要研究方向为基于 FPGA 的 QSPI 通讯系统设计。

E-mail：942333266@qq.com

伍俊(通信作者)，博士研究生，高级工程师，主要研究方向为红外图像处理系统设计，红外目标检测及红外图像增强算法的应用研究。

E-mail：wujun@cigit.ac.cn