Inline Text Wrapping Picture

北京邮电大学

硕士研究生学位论文开题报告

学 号: 2016110107

姓 名: 张大旺

学 院: 信息与通信工程学院

专业(领域): 信息与通信工程

研究方向: 多媒体与网络大数据

导师姓名: 别志松

攻 读 学 位: 工学硕士

2018年1月11日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 基于TTA的神经网络处理器架构设计 | | |
| 选题来源 | 其他 | 论文类型 | 应用研究 |
| 开题日期 | 2018-01-11 | 开题地点 | 教三819 |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  1.1 研究目的：  基于神经网络的人工智能近年取得了突破性进展，正在深刻改变人类的生产和生活方式，是世界各国争相发展的战略制高点。  神经网络作为实现人工智能任务的有效算法之一，已经在各种应用场景获得广泛的应用。从云端到移动端，不同应用场景也对神经网络的计算能力提出了不同的需求。  神经网络的广泛应用离不开核心计算芯片。目前的主流通用计算平台包括通用处理器CPU以及图形处理器GPU，但是由于CPU的计算规模过小，以及GPU的功耗过高，所以发展神经网络的专用处理器的需求日益强烈。  传输触发结构体系(TTA, Transport Triggered Architecture)的核心思想是利用数据传输来触发相应功能单元的具体操作。TTA结构将寄存器单元也作为一种特殊的基本单元，它有效地减少了寄存器堆的设计压力，成为一种非常适合于专用处理器领域的处理器架构。此外，TTA架构有着功能单元的灵活性以及可扩展性强等一系列优点，作为神经网络的处理器设计架构，也是非常有意义的。  1.2研究意义：  目前许多人工智能的产品都是采用通用处理器或者专用集成电路（ASIC, Application Specific Integrated Circuit）的实现方式，前者虽然灵活性较好，但是在对实时性或者功耗要求较高的场合并不适合，后者对于某一种算法或者网络，这种实现方式在功耗和性能上可以做到最佳，但是现在的产品通常是多个算法集中在一个设备上，使得这种实现方式的设计成本和功耗猛增，设计周期变长，灵活性很差。  而采用专用指令集处理器（ASIP, Application Specific Instruction Set Processor）实现方式，由于专用指令集处理器是针对某一算法或领域进行裁剪和优化，以满足性能、面积、功耗等约束的处理器，所以在功能与性能之间取得了一个平衡点。  本课题《基于TTA的神经网络处理器架构设计》旨在设计了一种基于TTA架构的专用指令集处理器，以同时满足在神经网络应用中的功能与性能需求。  1.3国内外研究现状和发展趋势：  对于应用神经网络的硬件，目前，国内外基于ASIC的研究较多，如何在硬件上实现算法的并行计算是研究的热点，而针对于神经网络处理器的设计则较少。  在国内，中国科学院计算技术研究所提出了寒武纪系列神经网络处理器，深鉴科技也提出了网络压缩以及专用处理器架构设计相结合的新型处理器结构。  1.4主要参考文献  [1] Han S, Kang J, Mao H, et al. ESE: Efficient Speech Recognition Engine with Sparse LSTM on FPGA[J]. 2017.  [2] 刘俊, 谢憬, 王琴. 基于TTA技术的专用处理器设计[J]. 微电子学与计算机, 2009, 26(11):161-164.  [3] 李杰. 基于TTA架构的ASIP设计与应用[D]. 西安电子科技大学, 2014.  [4] 徐争莉. 基于TTA的LTE符号级处理过程的研究[D]. 北京邮电大学, 2013.  [5] Teittinen J, Hiienkari M, Žliobaitė I, et al. A 5.3 pJ/op approximate TTA VLIW tailored for machine learning[J]. Microelectronics Journal, 2017, 61:106-113.  [6] Li H, Fan X, Jiao L, et al. A high performance FPGA-based accelerator for large-scale convolutional neural networks[C]// International Conference on Field Programmable Logic and Applications. IEEE, 2016:1-9.  [7] 朱学亮, 柴志雷, 钟传杰,等. 基于FPGA的图像卷积IP核的设计与实现[J]. 微电子学与计算机, 2011, 28(6):188-192.  [8] 方睿, 刘加贺, 薛志辉,等. 卷积神经网络的FPGA并行加速方案设计[J]. 计算机工程与应用, 2015, 51(8):32-36.  [9] Yue H, Shen L, Dai K, et al. A TTA-Based ASIP Design Methodology for Embedded Systems[J]. Journal of Computer Research & Development, 2006, 43(4):752-758.  [10] 陆志坚. 基于FPGA的卷积神经网络并行结构研究[D]. 哈尔滨工程大学, 2013.  [11] 朱礼波. 基于TTA技术的多功能可配置DSP处理器设计[D]. 上海交通大学, 2008. | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  2.1 研究的具体内容  对于神经网络的应用而言，其实现方式目前主要有三种。第一种，采用通用处理器，如CPU, GPU, DSP等，通过软件编程的方式实现。这种方式具有很高的灵活性以及较短的上市时间，但由于通用处理器的设计是面向通用，具有高性能以及高灵活性，但是在一些对实时性以及功耗要求比较高的场合，这种实现方式并不合适。第二种，采用专用集成电路（ASIC, Application Specific Integrated Circuit）的实现方式，对于某一种网络或者算法，生成其固定的物理版图。对于某一种网络或者算法，这种实现方式可以在功耗以及性能上达到最佳，但是如果想使用多个网络或者算法，只能将这些网络或者算法的物理版图独立的集中在一个设备上，使得这种实现方式的设计成本和功耗猛增，设计周期变长，灵活性很差。第三种，采用专用指令集处理器（ASIP, Application Specific Instruction Set Processor）实现，由于专用指令集处理器是针对于某一算法或领域进行裁剪和优化以满足性能、面积、功耗等约束的处理器。因此，它既具有ASIC的高性能又具有通用处理器的灵活性，同时还能够有效地缩短设计周期，降低设计风险。  随着 ASIP 技术的发展，其设计流程也产生了很多变化，但是大体上可以分为5个步骤：应用需求分析、体系结构选择、指令集设计、代码综合和硬件综合，其中体系结构选择在整个 ASIP 设计过程中至关重要，将直接影响到系统的性能。目前，主流的体系结构有如下几种。  第一种是复杂指令集结构体系（CISC, Complex Instruction Set Computer），CISC 结构采用微码状态机进行设计，一条汇编指令通常包含若干条微码指令，因此，CISC 结构的一条汇编指令可以执行复杂的功能，具有很高的执行效率，但这也使得 CISC 结构的硬件设计变得十分复杂。  第二种是精简指令集结构体系（RISC, Reduced Instruction Set Computer），在 20世纪七八十年代，研究人员通过对大量应用程序进行分析发现，CISC 指令集中只有20%的指令使用频率最大，约占运行时间的80%，针对这种情况，人们研究出了 RISC 结构。RISC 结构指令集只包含那些使用频率最大的指令，其他指令则通过这些指令编程实现，RISC 结构的显著特点就是硬件结构简单，开发周期短。  第三种是超标量结构体系（Superscalar），RISC 结构虽然硬件实现简单，但是运行效率不高，为了增加运行效率，必须开发指令级并行性，Superscalar 结构系统应运而生。Superscalar 结构体系可以同时执行多条指令，采用硬件的方式检测同时执行的指令间的相关性，以保证程序正确无误地运行。由于采用硬件的方式检测指令间的相关性，显著加大了硬件开销，因此只有在代码兼容性问题成为首要考虑因素时，才会选择这种结构体系进行设计。  第四种是超长指令字结构体系（VLIW, Very Long Instruction Word），VLIW 结构是通过编译器调度，将数据不相关的若干条指令打包成一条长指令执行，从而实现指令级并行性，显然，这种方式的硬件开销相比 Superscalar 结构大大降低，但是，具有指令调度功能的编译器设计成密度也成为一个不可忽略的问题。  第五种是传输触发结构体系（TTA, Transport Triggered Architecture），TTA 结构由 Corporaal 等人提出，其核心思想是利用数据传输来触发相应功能单元的具体操作来触发相应功能单元的具体操作。TTA架构可以看成VLIW的一个超集，我们把VLIW看成 SIMO(单指令多操作)类型的体系结构，那么 TTA则是 SIMT(单指令多传输)类型的体系结构。TTA相比于VLIW，将寄存器单元作为一个功能单元（function units），解决了VLIW读写寄存器带宽的瓶颈问题，同时采用触发结构，解决了VLIW的功能单元之间互联过于复杂的问题。  由于神经网络具有内存密集（memory intensive）的特性，因此采用VLIW并不合适，而采用TTA架构则可以缓解这一问题。因此本课题最终选择采用TTA架构来进行神经网络处理器的设计。  基于 TTA 结构的专用处理器设计主要体现在以下三方面的设计：  1) 指令集的设计。  2) 功能单元的设计。  3) 数据交换网络的设计。  由于TTA架构的指令格式统一，只有一种MOVE格式，因此难点以及重点在后两个方面，即如何设计针对于神经网络的专用功能单元以及如何设计数据交换网络。  对于功能单元，需要针对神经网络的架构进行单独设计，满足神经网络运算的通用性以及完备性。功能单元的设计的目的是计算代码中运算最为密集的一些操作，从而提升处理器的性能。因此，程序中包含的主要运算操作及数量决定了TTA结构中功能单元的种类及数量。根据操作的类型,设计者可以很快确定需要哪些功能单元。根据某种操作占总操作的百分比,设计者可以很快确定需要使用该类型的功能单元的数量。对于寄存器文件这种特殊的功能单元，需要分析其它功能单元需要的存取带宽，从而确定需要多少通用寄存器。在保证性能的前提下，尽可能的节约硬件开销，由此确定处理器寄存器文件的大小与数量。  通常情况下，数据交换网络会成为整个处理器的关键路径，因此，数据交换网络的设计是整个 TTA 处理器设计的重点，数据交换网络的数据传输速度将直接影响到整个处理器的处理性能。数据交互网络包括总线与接口，主要负责将不同的功能单元，寄存器单元联系起来。然而并不是每个功能单元和寄存器单元的输入输出都要连接到每条总线上。这样不仅会增加面积，而且增加输入输出接口电路上的扇出，降低信号的品质，使性能下降。所以在满足性能需求的前提下，可以减少输入输出接口的数目以及连接的总线数量。这也使得数据交换网络的设计的难度大大增加。  2.2研究目标和效果  在经过调研以后，对比多个TTA架构的开发工具集，最终选择了TTA协同设计环境(TTA-based Co-Design Environment, TCE)作为本课题的开发工具。TCE是是芬兰的坦佩雷科技大学（Tampere University of Technology）研发的一个面向TTA处理器的架构设计的工具集。TCE提供了半自动的处理器设计流程，支持设计空间探测。TTA协同设计环境设计、执行和验证为一体，提供了编译器和指令集仿真器等一些软件工具，为设计过程中的设计空间探测提供了极大的便利。  本课题《基于TTA的神经网络处理器架构设计》将使用TCE工具集，面向TTA处理器架构，设计出一套神经网络处理器的功能单元与数据交换网络结构，旨在满足灵活性的同时，在性能上也达到一定要求。  本课题最终的实现效果为，对于任意规模与结构的神经网络，都可以使用该处理器来进行运行。并在性能上可以满足应用时的需求。  2.3 拟解决的关键科学问题  （1）功能单元的设计。  功能单元的设计是本课题最基础与最重要的问题之一，功能单元设计的好坏将直接影响到并行度以及数据交换网络的复杂度。如何根据神经网络设计出通用、高效的功能单元，是本课题的难点之一。通用性指的是，对任意规模的网络层，该功能单元都可以通过与其它功能单元互联来进行实现；高效性指的是，功能单元的利用率需要达到一定的值，以免面积的浪费。  （2）数据交换网络设计。  数据交换网络提供处理器中各个单元交换数据的通道，它包含两种基本模块，Socket与总线。除了提供数据交换功能以外，总线还用于传输控制信号，比如源和目标寄存器的ID，功能单元锁存信号等。Socket提供了功能单元和寄存器文件与总线的连接，每个Socket可以连接到一条或多条总线以及某功能单元的一个或多个寄存器。每个Socket与每条总线都相连的方式成为全连接网络，它能简化总线的传输调度，但因为连接点会增加总线负载，延长全局周期时间，增加功耗，因此并不是一种高效的设计方式。尤其在神经网络这种大型网络中，采取全连接型的数据交换网络更不可取，因此如何设计高效的数据交换网络也是难点之一。 |

|  |
| --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  3.1 研究方法  针对本课题的研究内容与特性，结合自身的学习经验和深入广泛的调研，确定了以下研究方法。  （1）文献研究法。  文献研究法是根据一定的研究目的或课题，通过调查文献来获得资料，从而全面地、正确地了解掌握所要研究问题的一种方法。在开展课题的前期，需要更为深入的了解课题的广度以及难度，因此需要利用网上图书馆的文献资源，对相关领域进行调查以及研究。  （2）定性分析法。  定性分析法就是对研究对象进行“质”的方面的分析。在对课题内容有了一定的了解以后，需要对所做课题进行抽象与概括，进行去粗取精，分析课题的重点以及难点。  3.2 技术路线  本课题主要包括三个方面的内容：  （1）初始架构设计以及高级语言代码实现。  TCE是基于TTA架构的软硬件协同开发集，其在一个初始架构的基础上，对软件代码进行分析，生成适应于此架构的指令集。最终根据硬件描述的功能单元，生成最终的编程比特流文件。因此，初始架构以及相应的高级语言所编写的软件代码，是需要最初实现的。初始架构可以有一定的冗余度，这也大大减小的设计的难度。  （2）功能单元的设计。  功能单元是对算法有了一个的认识之后，将算法中的运算最为密集的部分使用功能单元来进行单独的运算。然而，这样的话，功能单元的颗粒度并不能确定。若功能单元的颗粒度过小，则系统架构的复杂性则会大大增加，若功能单元的颗粒度太大，则在灵活性上又会有很大的损失。所以功能单元的颗粒度大小的设计是需要不断尝试的，最终找到一个平衡灵活性以及复杂性的中间结果。  在功能单元的设计完成之后，还需要使用硬件描述语言（HDL）来进行硬件的描述，这样，TCE工具集才能生成最终的编程比特流文件。  （3）架构设计。  架构设计主要包括总线、输入输出接口、功能单元、立即数、寄存器单元。其中总线数量、各种功能单元数量、寄存器单元数量等可以根据性能文件简单更改即可。由输入输出接口与总线所组成的数据互联网络，则需要根据对算法以及功能单元的深入理解，定制出更简单的网络，直到满足设计约束的体系结构为止。  3.3 步骤  1. 使用高级语言写出串行运行代码。  2. 使用front-end编译器编译出串行的MOVE指令代码。使用一个冗余度比较大的系统架构，由仿真器仿真，得到性能文件。  3. 分析源代码，设计或更改功能单元；分析性能文件，更改系统架构或互联网络。  4. 使用功能单元代替源代码中的操作，使用back-end编译，得到性能文件。  5. 重复步骤3与4，直到找到满足需求的功能单元与系统架构的设计。  3.4 可行性分析  从步骤可以看出，TTA处理的开发实际上是一个迭代的过程，是根据具体的应用或者算法，去探索与其最匹配的处理器结构的过程。因此，当有了一个最初始的架构以后，我们就可以不断的迭代与优化。而最初始的架构的冗余性可以较大，这样寻找的难度也大大减小，所以在可行性上是可以满足的。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  本课题《基于TTA的神经网络处理器架构设计》着眼于神经网络的应用方向，旨在利用TTA架构的优越特点，解决神经网络应用中的灵活性问题，并在计算密集以及内存密集两个神经网络应用瓶颈上，采用功能单元的形式进行解决。  本课题的创新之处主要有以下两点。  （1）神经网络专用功能单元设计。  在以往的神经网络应用产品中，主要有使用ASIC设计来生产成品和使用通用处理器做为产品两种方式。使用ASIC设计的方式灵活性不足，若产品需要多种功能，则只能用面积换功能，极大的增加了成本以及功耗；使用通用处理器运算的方式，虽然灵活性足够，但是其功耗较高，在特定的场合中，并不能满足需求。  本课题设计的神经网络专用功能单元，旨在同时满足灵活性以及功耗两个方面。因为可以使用专用功能单元实现任意网络，因此在实现多种功能的同时，就不需要以面积换功能，保证了灵活性。同时，因为它是面向神经网络设计的专用功能单元，因此在功耗上也可以得到满足，从而在灵活性与功耗上达到一个平衡。  （2）基于TTA架构神经网络实现。  因为神经网络具有内存密集型特点，采用TTA架构，可以减少访存次数，从而起到降低处理器所需带宽的需求。  在其他的指令集结构体系中，运算都是指令触发的，数据存放在内存中，在运算触发后，从内存中取数据，运算结束后，将结果写回内存中。而在TTA架构中，运算时传输的边沿效应，当功能单元的触发寄存器传入数据时，则开始运算。这样的好处在于，数据不需要传回内存，可以在功能单元中暂存。所以，在对多个数进行相加时，传统架构，需要多次的把结果写回到内存中，而TTA架构则只需要把最终的结果写回。通过设计数据传输网络，则可以把这一优势放大，从而减小访存次数，从而缓解神经网络内存密集的问题。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**  5.1与本项目相关的研究工作积累基础  通过研究生阶段的学习以及实践，我对神经网络以及IC设计都有了一定的了解以及认识。在研究生阶段的工作中，我曾经参与过《便携式测试设备微型测控模块》中的硬件部分的工作，使我对FPGA以及硬件设计都有了一定的认识。在随后的《基于FPGA的Alexnet网络实现》项目中，我参与并完成了神经网络并行计算的设计，其中包括并行计算设计、访存优化、流水线的设计等。使我对如何设计神经网络的并行结构有了深一步的认识。但由于在FPGA实现的神经网络属于ASIC的开发方式，在应用中并不能很好的满足需求，因此，产生了研究工作的初衷和重点，神经网络处理器的架构设计。  5.2 已具备的实验条件  本课题来自于信息与信号处理实验室，本实验室长期从事硬件的设计与开发项目，近期与互联网公司又开展了神经网络的应用项目。因此，无论在硬件设计还是在神经网络应用方面，都有着一定的积累以及经验，为神经网络处理器设计的课题开展打下了基础。  TCE是基于TTA架构设计和编程定制处理器的工具集，该工具集提供了从高级语言程序到可综合RTL以及并行程序二进制文件的完全可重定向的协同设计流程。TCE最佳的支持环境为Linux系统。本实验室已具有一台高性能装有TCE工具集的服务器，可以满足课题需求。  同时，本实验室又采购了Xilinx公司的KCU1500加速开发套件，该套件是超大规模应用开发人员的一个很好的起点。 KCU1500加速开发套件适用于计算密集型的应用，因此为神经网络处理器的开发提供了一个计算资源充足的硬件开发平台。 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 时间 | 研究内容 | 预期效果 |
| 2017/12-2018/02 | TTA架构与神经网络基础知识调研 | 对TTA架构与神经网络有基本认识 |
| 2018/02-2018/05 | 熟悉TCE工具集并设计出初始架构，完成高级语言的编写 | 完成高级语言软件部分的编写，生成最一份性能文件 |
| 2018/05-2018/07 | 专用功能单元的设计 | 设计出专用功能单元 |
| 2018/08-2018/11 | 针对数据交换网络进行设计与优化 | 数据交换网络完成优化，提升性能 |
| 2018/11-2019/1 | 对不同网络进行实现 | 可以实现不同网络 |
| 2019/1-2019/3 | 论文撰写与修订 | 完成论文的撰写 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 单位名称 | 职务 |
| 别志松 | 副教授 | 北京邮电大学 | 组长 |
| 王思野 | 讲师 | 北京邮电大学 | 成员 |
| 龚萍 | 副教授 | 北京邮电大学 | 成员 |
| 林雪红 | 副教授 | 北京邮电大学 | 成员 |
| 李永华 | 副教授 | 北京邮电大学 | 成员 |
|  |  |  |  |
| 导师意见： | | | | |
| 针对深度神经网络设计专用处理器是目前的一个研究热点，选题基于TTA架构优化设计针对深度神经网络的ASIP，选题具有理论意义和应用价值。研究内容和技术路线合理，同意开题。 | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | |
| 开题报告小组意见： | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | |
| 学院意见（签章）： | | | | |
| 负责人：  日期： 年 月 日 | | | | |