神经网络并行MIMD处理器的研究及实现

钱 艺[1]，王 沁[2]，吴 巍[2]，刘金龙[2]

(1. 泰山学院信息科学技术系 山东 泰安 271000; 2. 北京科技大学信息工程学院 北京 海淀区 100083)

【摘要】为了能高速地实现多种神经网络，拓展神经网络在工业控制中的实时性、嵌入式应用，设计了一种多指令多数据流(MIMD)的通用型神经网络处理器(APP)。处理器的处理单元组之间、处理单元组与乘累加协处理器之间均可以并行执行任务、处理单元组与其他存储器之间可以并行通信。在FPGA上仿真验证了处理器的功能，并实现了用于轧辊偏心在线控制的BP网络和用于字符识别的Hopfield网络等两种不同的拓扑结构。实验数据表明，该体系结构具有较高的并行性，其性能优于其他常见的通用型实现手段。

关 键 词 通用; 多指令多数据流; 神经网络; 处理器

中图分类号 TP183 **文献标识码** A

Realization of a Neural Network Parallel MIMD Processor

QIAN Yi[1], WANG Qin[2], WU Wei[2], and LIU Jin-long[2]

(1. Department of Engineering and Computer Science of Taishan University Taian Shandong 271000;

2. Information Engineering School in University of Science and Technology of Beijing Haidian Beijing 100083)

Abstract To expand the realtime and embedded application of the neural network in industry control, general purpose neural netowrk processor (APP) based on multiple instruction stream and multiple data stream (MIMD) is paoposed. The tasks among PEGs, PEG, and MAC coprocessor can be processed in parallel, and also communications among PEGs and other memorys can be carried out in paralled. The processor has been simulated on FPGA and is used to implement two different neural network: the BP network used in roll eccentricity online control and the Hopfield network used in character recogonition. The simulation result shows that the performance of the proposed APP is better than that of other general methods for neural network implementition.

Key words general purpose; MIMD; neural network; processor

人工神经网络(artificial neural network，ANN)已经广泛应用于工业控制领域，生产技术的进步对ANN的实现技术在实时性、嵌入式等方面不断提高要求。通用神经网络处理器针对大多数ANN算法而设计，既能够以较小的硬件开销获得较高的处理速度，又具有软件编程的灵活性，是实现ANN算法的重要手段。

目前多数神经网络处理器采用SIMD结构[1-2]，片内多个处理单元的数据能够并行从而获得较快的速度。但是这些处理器缺乏灵活性，运行时间受ANN模型的影响很大。当生产工艺、控制条件发生变化而要求ANN的算法随之变化时，处理器的运行时间则会大大降低，甚至不能实现已经发生改变的算法。相比之下，MIMD结构既可以执行相同的指令段作为SIMD结构来用，也可以同时运行多个任务，具有更好的灵活性和实时性，适于通用型的神经网络硬件实现。目前，许多大型的神经计算机系统采用MIMD结构[3-4]来构成多处理器阵列，由于其价格昂贵体积较大不适于工业控制的嵌入式应用。

为了能高速实现多种神经网络，拓展神经网络在工业控制中的实时性、嵌入式应用，本文在分析ANN算法并行性的基础上设计了通用型神经网络处理器APP。

1 ANN算法并行性的分析

ANN算法是典型的并行算法，它的硬件实现均采用并行处理技术来提高处理器的并行度，进而达到缩短运行时间的目的。开发ANN算法并行性的等级按照并行粒度的大小分为微任务级、神经元级、任务级等，在任务级开发神经网络算法的并行性，其通用性较好[5]。本文从这方面对传统的ANN算法

收稿日期：2007－07－16; 修回日期：2007－12－05

作者简介：钱 艺(1975－)，女，博士，主要从事人工智能、计算机体系结构方面的研究.

做了归纳来作为设计APP的依据，如表1所示。

表1　ANN算法总结

任务	主要公式	运算类型
计算加权和	$p=\sum w_{ij}x_i-\theta$	乘、加
计算激活函数值	斜面函数、阶跃函数、S型函数、高斯函数、墨西哥草帽函数等	LUT、线性分段等
训练权值 Delta	$w_{ij}(t+1)=w_{ij}(t)+\alpha o_i(t)o_j(t)$	乘、加
训练权值 Hebb	$w_{ij}(t+1)=w_{ij}(t)+\alpha o_i(t)(y_j-o_j(t))$	乘、加
计算能量函数	$E_p=\frac{1}{2}\sum(y_{pj}-o_{pj})^2$	乘、加
判断稳定性	比较数值的大小	逻辑

从任务并行的角度来看，表1中的任务如果分别属于不同的运算部件，而且任务之间不存在数据相关性，那么就可以开发任务并行性。如计算能量函数的任务与神经元计算加权和的任务没有数据相关性，虽然都是使用乘加部件，但是如果它们恰好不在同一个处理单元上，那么两个任务就可以并行。另外，神经元如果被映射到不同的处理单元上，那么也可以并行执行任务，特别是在循环迭代中，这种并行能够使开销显著减小。但是神经元被划分到不同的处理单元中，就会增加通信开销，所以存在处理单元之间的并行通信的问题。

最后，乘加运算部件在这些任务中占有最多的运算量，该部件的运行效率会影响整个程序的运行时间；乘加运算部件与其他运算部件之间在没有数据相关性时均可以开发任务并行性。

通过以上分析，可以得出结论：处理器应该具有表1中的计算功能，体系结构的设计重点是实现ANN算法的任务级并行性，从而达到缩短程序运行时间的目的。

2　APP处理器的设计

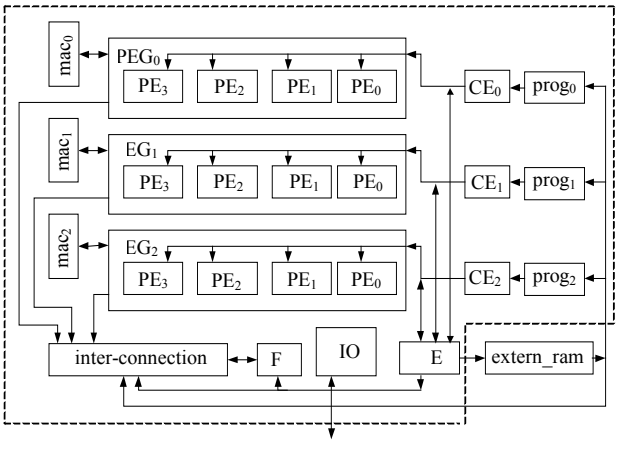

图1　APP处理器的体系结构

APP处理器的体系结构如图1所示，有3个并行特点使其能够高速实现ANN算法：采用基于分布式存储的MIMD结构，处理单元组PEG是独立编程的，可以并行执行多个任务；每个PEG附带一个乘累加MAC协处理器，使MAC运算有较高的效率，而且能够与PEG中的其他操作并行进行；可以并行访问多个存储器的数据通路，inter-connection模块能够减少通信开销。

2.1　总体设计

图1中处理单元组PEG_0、PEG_1、PEG_2分别由CE_0、CE_1、CE_2控制。每个PEG有独立的运算部件ALU、本地存储器LM、寄存器堆regfile以及数据通路，由各自的prog模块独立编程，能够灵活地开发PEG之间的任务并行性。PEG还能以8位为单位进行分割，每个PEG分割为4个结构相对独立的8位或16位PE。所以当操作数是32、16、8位时，APP一次可以最多运算3、6、12个神经元，可以增强低精度数据的并行性。

Extern_ram是片外存储器，存储神经网络的原始数据、运行结果、激活函数值以及指令代码等，用户指定extern_ram中的一个区域存储对应的激活函数的采样值，并且通过extern_ram的数据线进行多片扩展。F模块包含除法器、开方器，由于它们的硅面积较大而且在许多ANN算法中的使用频率不高，因此只做成一个模块供所有PEG访问。IO模块是PCI桥接口芯片PCI9054与处理器的接口模块，可以产生中断与上位机通信，从而使APP处理器具有在线学习的能力。CE协调各个PEG之间的数据通信，控制F模块、extern_ram、inter-connection以及IO接口模块的工作。

每个PEG在任务级编写程序，为了缩短程序总的运行时间，对各个处理单元分配任务时基于以下原则：将同层的神经元平均分配至各个处理单元中，做到计算负载的平衡，而且可以并行计算，多余的神经元在计算过程中动态分配到某个处理单元上，以求在循环迭代中使总的计算负载尽量平衡。

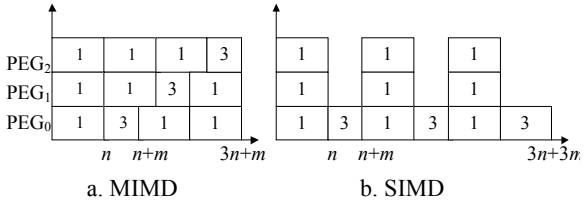

图2　Hopfield的任务时序图

以Hopfield网络为例，如果将神经元平均分配到3个PEG上，在回忆阶段有以下任务：计算本次迭代

的神经元值(任务1)；广播本次的计算结果到所有处理单元(任务2)，与前次的神经元值比较判断系统是否稳定(任务3)。任务2的时间可以忽略，任务1和任务3在APP上可以并行执行，时序如图2a所示。与图2b的SIMD结构实现相比，3次迭代之后总运行时间缩短了2 m。为了保持任务级的并行，APP必须避免数据相关，使得在处理任务3时改成每次只对三分之二的样本值进行比较，最终需要多比较一次才能确定系统是否稳定，但是这与上百次的迭代所耗费的时间相比是微不足道的。

2.2 MAC协处理器

ANN算法中存在大量的乘累加运算，根据前面的分析，可以与其他数据不相关的任务并行。神经网络处理器为了获得特定应用下的最大效率，通常将乘加运算部件与其他运算部件用特定的数据通路进行连接，在ANN算法改变时效率反而降低。大部分DSP则是采用超标量流水线技术达到乘加运算与其他运算并行的目的，控制比较复杂。

为了保证APP处理器在具备通用性的同时有较高的并行度，APP把乘加运算部件从PEG中分离出来作为MAC协处理器专门实现乘法、乘加、乘累加运算，可以与PEG中的其他操作并行。由于参与MAC运算的权值和神经元值都是存放在LM中的，所以数据通路设计为直接从PE的LM中读取操作数进行运算，计算结果直接写回存储器。CE_i对“MAC_1”和“MAC_2”两条连续的指令译码，得到A、B两个操作数在LM的相对地址、每次运算的地址增量、结果C在LM中的相对地址以及累加次数。SM读取这些初始值，计算当前A、B的地址并控制MAC协处理器。这样可以既避免中间暂存环节上的数据相关，缩短MAC的计算时间，又能使CE_i在MAC运行的同时继续取下一条指令，从而PEG同时可以进行其他操作。

整个乘累加过程的执行指令数为2+n(n为累加次数)，当n>>2时乘累加运算部件可以认为是满流水操作。相比之下，DSP的MAC指令只能执行乘加运算，如TMS32C5x重复执行n次才能得到最后的累加结果，之后还要再将结果存入存储器；而且要求参加运算的操作数之前先按地址增1或减1的顺序存放，至少增加了2 n个时钟周期的开销。

2.3 片内数据通路

处理单元间的互连网络是并行处理系统中的关键部件，其功能的强弱直接影响通信开销的多少[7]。APP处理器中存在extern_ram与PEG之间、F与PEG之间、PEG与其他PEG的LM之间的通信。其中由于神经元之间的广泛互连，PEG与其他LM之间的通信是大量存在的，如在同一时钟周期PEG_0读LM_1、PEG_1读LM_2、PEG_2读LM_0，如果采用单一的广播方式，则这段通信开销为3个PEG分别读取外部LM的总和，从而成为通信的瓶颈。

本文设计了inter-connection模块作为APP处理器的内部数据通路，如图3所示。

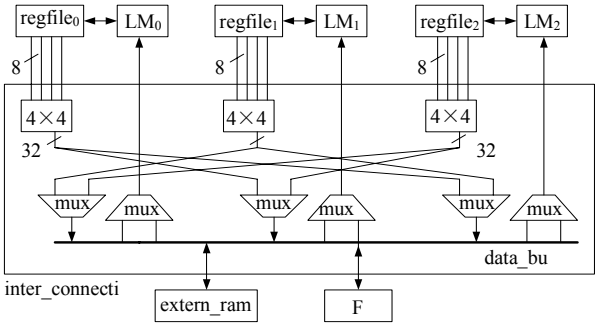

图3 inter-connection模块

通过控制各个交叉开关及多路选通器的工作状态，使互连网络完成extern_ram与PEG、F与PEG、PEG与其他LM之间的广播通信以及并行的点对点通信，具有极大的灵活性。4×4交叉开关能使处理器在数据长度为8位或16位时把一个8位或16位的数据传送到对应的LM中。当3个PEG并行地对其他LM通信时，比总线方式下的通信时间减少了三分之一。

3 实验及结果分析

目前APP处理器的功能在FPGA上通过了仿真验证并分别实现了两种结构完全不同的ANN算法。仿真平台是基于PCI的FPGA板卡，其中桥接口芯片是PCI9054，FPGA器件是Xilinx的XC2VP20。

3.1 实际运行时间的比较及分析

用APP实现的BP算法是用于带钢轧制过程中轧辊偏心信号的在线检测。网络拓扑结构为：8×10×1，激活函数为Sigmoid函数，32位数据位长。每40 μs可得到一个输出值，完全满足采样间隔100 ms的时间要求。因此，在轧辊偏心控制中使用APP处理器，还可以进一步提高采样频率，从而提高偏心控制的精度。在APP上实现了一个离散的25个神经元全互连的Hopfield网络，数据长度为8位，网络结构为单层全互联，可以联想和记忆3个不同的字符。

与工业控制中常用的PC机、DSP等实现手段比较，APP的运行时间最短，具有较好的实时性，而且BP网络和Hopfield网络分别代表神经网络中的两种典型结构，可以说APP处理器具有较好的通用性。由于设计思路不同，它的实现速度仍然不能与专用

的FPGA相比，具体数据如表2所示。

表2　在不同芯片上实现ANN算法的运行时间比较

比较对象	频率/MHz	BP/ms (1次迭代)	Hopfield/ms
APP	75	0.040 0	0.23
TMS320C54	100	0.220 0	0.31
Pentium*	2 800	0.500 0	0.67
文献[8]	50	0.004 6	—

注：*是根据C语言所编程序经10 000次迭代所得到的平均值。

3.2　性能指标的比较与分析

APP处理器基于0.25 μm的CMOS工艺设计，使用Cadence公司的PKS工具，综合后的时钟频率为75 MHz，采用RISC指令系统5级流水线，指令周期和MAC周期均为13 ns，当12个PE并行执行指令时，它的处理能力将达到最大值900 MIPS。

通常情况下，MCPS(million connection per second)是衡量神经网络处理器性能的重要指标。APP与其他实现手段的MCPS值在3种数据长度下进行比较，如表3所示。

表3　各芯片的MCPS值

名称	频率/MHz	性能(MCPS)		
		8 bit	16 bit	32 bit
MA16[9]	25	—	400	—
NM6403[10]	50	1200	200	50
TMS320C54	100	—	100	—
Kestrel	50	400	—	—
Pentium*	2 800	—	—	7.78
文献[8]	50	—	1 000	—
APP	75	900	450	225

FPGA的MCPS值虽然较高，但是仅限于一种网络，其通用性不足。TMS320C54虽然频率较高，但是只有一个乘累加器，且受自身的MAC指令的限制，影响了并行度。MA16与NM6403都是SIMD体系结构的神经网络处理器，与APP在3种数据长度下的MCPS值进行综合比较，APP的处理能力更具有通用性和较高的并行度。

4　结　论

针对ANN算法具有任务级并行性的特点，本文设计了基于MIMD的通用神经网络处理器结构，并在FPGA上进行了功能仿真，实现了两种典型的神经网络算法，性能指标和实际运算时间均优于其他常见的通用型实现手段。由于该处理器具有并行度高、运行速度快、通用性好的特点，因此适于实时性要求较高的嵌入式应用场合，也可以多片互连构成并行神经计算机系统。它的可编程性及通用性也为使用者开发新的ANN控制器、缩短设计周期和快速投入生产带来方便。目前APP处理器仍处于研究的初级阶段，需要用多种ANN算法对其性能进行验证，以便在硬件及软件方面进一步优化改进。

参　考　文　献

[1] DIAS F M, ANTUNES A, MOTA A M. Artificial neural networks: A review of commercial hardware[J]. Engineering Applications of Artificial Intelligence, 2004, 17(8): 945-952.

[2] MCBADER S, LEE P, SARTORIN A. The impact of modern FPGA architectures on neural hardware: a case study of the TOTEM neural processor[C]//2004 IEEE International Joint Conference on Neural Networks. Budapest: IEEE, 2004: 3149-54.

[3] BONIFACE Y. A parallel simulator to build distributed neural algorithms[C]//Proceedings of the International Joint Conference on Neural Networks. Washington: IEEE, 2001: 2399-2403.

[4] ABBAS H M. Performance of the Alex AVX-2 MIMD architecture in learning the Net Talk database[J]. IEEE Transactions on Neural Networks, 2004, 15(2): 5-14.

[5] 戴　葵. 神经网络实现技术[M]. 长沙: 国防科技大学出版社, 1998: 23-24.
DAI Kui. Implementation technology of neural network[M]. Changsha: National University of Defense Technology Press, 1998: 23-24.

[6] 王　运, 黄大贵, 杨天文. 开放式数控切割机神经网络误差补偿研究[J]. 电子科技大学学报, 2007, 36(2): 305-308.
WANG Yun, HUANG Da-gui, YANG Tian-wen. Study on error compensation for open CNC cutting machine based on neural network[J]. Journal of University of Electronic Science and Technology of China, 2007, 36(2): 305-308.

[7] 韩　亮, 李　莺, 张　馨, 等. 高性能可重构DSP处理器的数据通路设计[J]. 电子科技大学学报, 2005, 34(2): 194-197.
HAN Liang, LI Ying, ZHANG Xin, et al. Data path design in high performance reconfigurable DSP processor[J]. Journal of University of Electronic Science and Technology of China, 2005, 34(2): 194-197.

[8] WANG Qin, LI A, LI Zhan-cai. A design and implementation of reconfigurable architecture for neural networks based on systolic arrays[C]//Proceedings of the 3rd International Symposium on Neural Networks. Chengdu: Springer-Verlag, 2006: 1328-1333.

[9] ZELL A. Simulation of artificial neural networks on parallel computer architectures[J]. Systems Analysis Modelling Simulation, 1999, 35(4): 483-519.

[10] ROCA D, MILA B, RANDON E. Design of a parallel neural processor[C]//Proceedings of the IEEE International Caracas Conference on Devices, Circuits and Systems. Margarita: IEEE, 1998: 109-112.

编　辑　漆　蓉