

论并行神经计算机系统设计的一些问题

戴葵 胡守仁

(国防科技大学计算机系 长沙 410073)

摘 要 Research of parallel neurocomputer system is one of major field of neural network realization technology. The design of neurocomputer system touches upon many problems, such as architecture, selection of processing elements, etc. According to our research results and practice experiment, some issues of parallel neurocomputer system design are presented in this paper. The discussion is helpful to the design of parallel neurocomputer system.

关键词 Neural Network, Neurocomputer, Neurocomputing.

神经计算机的研究是神经网络研究中的一项重要内容,神经计算机就是指根据神经网络结构及其计算特点,用电子器件、光学器件或分子/化学器件而构成的计算系统。神经计算机的研究主要集中在两个方面,即器件研究及系统构造。神经计算机的研究也称为神经网络实现技术的研究。

如果用 P 个物理处理单元去实现由 N 个神经元组成的神经网络,只要 $P < N$,就称其为神经网络的虚拟实现。在虚拟实现中软件的采用增加了其通用性和灵活性,神经网络虚拟实现主要是研究在现有的技术条件基础之上,如何将各种功能部件有效地组织起来,极大程度地支持神经网络的高性能仿真。

在用“虚拟”实现中,神经计算机的研究经历了神经网络软件模拟环境、神经计算加速器和并行神经计算机的过程。并行神经计算机是利用传统计算机技术和并行处理技术来设计实现的用于完成大规模神经网络模拟和高性能神经计算的并行神经计算系统。在并行神经计算机系统的设计中,同其它计算机系统的设计一样,主要从性能价格比的角度来衡量折衷,进行方案的选择,下面主要以性能价格比为基础出发点,根据我们的研究经验,论述在进行并行神经计算机系统设计时所必须考虑的一些问题,这些分析对并行神经计算机系统的设计大有裨益。

1 前端机和后端机的连接方式

当前并行神经计算机系统一般采用主辅机结构,后端机与前端机之间通常有三种连接方式,一种是后端机通过总线作为前端机的存储器映像设备;一种是通过前端机的标准外设接口(如 RS-232, DRV-11, SCSI 等等)后端机作为前端机的外部设

备;还有一种是后端机作为前端机的协处理器。

第一种连接中(图 1),后端机存储器地址空间是前端机存储器地址空间的一部分,可以使后端机和前端机结合更加紧密,前端机可以通过对存储器访问的方式直接对后端机进行访问,这一功能可以很容易通过存储器映像设备驱动软件实现。同样后端机也可以利用前端机的大部分资源,如图象处理卡,语音卡等等。

将后端机作为前端机的存储器映像设备的缺点是必然引起并行神经计算机系统存储总线的拥塞。后端机如果和前端机以及其它一些总线上连接设备频繁进行大量的信息交换,神经计算机系统的存储总线将成为系统瓶颈,这必然会降低前端机的性能。

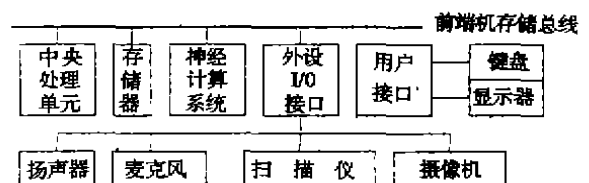


图 1 将后端机作为前端机的存储器映像设备

第二种连接(如图 2)的优点在于后端机可以与任何具有后端机标准外设接口的前端机相连,也就是说硬件的“可移植性”很好,但存在着一个严重的缺陷,因为前端机和后端机之间的关系可以看成是神经计算应用主程序调用神经计算应用子程序的关系,这种调用关系的实现是和系统所采用的操作系统以及前端机的设备驱动器紧密相关的,所以硬件的移植必然会带来软件的修改问题,而往往软件的修改比硬件接口的修改要复杂费事得多。

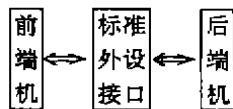


图2 后端机作为前端机的外部设备方法

在第三种方案(如图3)中,将后端机和其它外设、协处理器、文件服务器等设备通过高带宽局域网连接在一起,其优点在于后端机可以通过高速局域网访问其它网上设备,而不必利用前端机的存储总线,有效地降低了前端机的负载,同时大大增加了后端机直接访问信息的类型,这种方案可以通过分时的方式使一台后端机同时为多台前端机服务,同时由于软件的开发是基于网络的高层次,所以软件开发比上述第二种方案要简单得多。另外由于网络提供多层次的信息传输协议,可以大大提高后端机的接口能力。

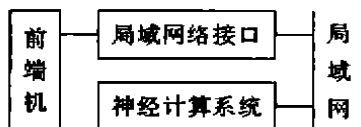


图3 后端机作为协处理器方案

总之,在并行神经计算机系统设计中,如果前端机系统比较简单,把后端机作为前端机的存储器映像设备是很好的选择;当前端机系统很复杂,系统需要多种其它外部设备,这时选择第三种方案比较好。

2 前端机和后端机的交互协议

当前在并行神经计算机实现中,一般是在后端机上运行的神经网络计算过程看作是在前端机上运行程序的一个子过程,后端机的任务非常简单,仅仅根据前端机发出的命令完成相应的神经计算。可以从这里得到一个很简单的前端机和后端机交互协议:“前端机调用后端机,后端机不调用前端机”。

前端机和后端机之间的这种交互协议使得在并行神经计算机系统设计中前端机和后端机的接口设计变得比较简单,而且可以很容易对整个神经计算过程进行管理。但是一定要注意时间冲突的处理,前端机是完全的主设备,后端机是完全的从设备。

3 并行神经计算机系统的模拟规模

并行神经计算机系统的模拟规模是指并行神经计算机所能够模拟的神经元和连接数目,这是衡量一台并行神经计算机计算能力的一个很重要的指标,也是在并行神经计算机系统设计中必须考虑的问题。

一般来说,表示神经网络中一个神经元的信息需要多个浮点数,而表示神经元之间的连接信息至

少需要一个浮点数。并行神经计算机的模拟规模是和并行神经计算机系统的内存容量紧密相关的,内存能够存储的神经网络信息越多,其模拟规模就越大,但是系统的造价就会越高,如果不很好地组织存储神经网络信息,并行神经计算机在进行神经计算时访问神经网络信息的效率就会很低,这必然会降低并行神经计算机系统的计算性能。在并行神经计算机系统的设计中,对其模拟规模应该从以下几个方面去考虑:①并行神经计算机系统的模拟规模应该与其所能支持的应用紧密联系;②必须要很好地组织神经网络信息,避免出现冗余信息,尽量提高对神经网络信息的访问速度;③采用“虚拟存储”的概念通过辅存实现大规模神经网络信息的存储。

4 并行神经计算机系统的通用性

并行神经计算机系统的通用性是指它所能实现的神经网络模型种类,有的可以模拟“所有”(大部分)神经网络模型,有的可以模拟某些特殊的神经网络模型。并行神经计算机模拟神经网络模型的种类越多,其通用性越好,应用范围就越广。

一般来说,并行神经计算机系统通用性的好坏直接与其设计的难易程度紧密联系,同样也和其整体性能密切联系。当前神经网络研究者提出了许多种神经网络模型,这些神经网络模型的拓扑结构和计算过程各不相同,它们在不同体系结构的并行神经计算机上实现,所能得到的性能也会不同。所以如果要求并行神经计算机系统具有很好的性能,同时获得较高的计算性能,其体系结构应该是可变的。

并行神经计算机系统具有良好的通用性带来的另一个问题是其系统软件设计的困难和系统管理的困难。为了完成高性能神经计算,希望并行神经计算机系统具有变结构的特点,同样要求软件具有较好的适应能力,软件不仅能够适应不同神经计算过程描述的需要,还要根据硬件体系结构对要实现的神经计算进行不同的任务分配,以及对神经计算过程进行不同的管理。

5 并行神经计算机系统的计算速度

高性能是并行神经计算机系统设计追求的一个重要目标。就多数神经网络模型来说,每个神经元的处理功能极其简单,所以在并行神经计算机系统设计中首先遇到的问题是用什么方法或器件来实现神经元的功能,可以从两个方面来考虑这个问题。

第一可以自行设计出能够完成神经元计算功能的处理单元,这种处理单元很简单,但自行设计并行神经计算机的处理单元是一件复杂的工作,必然会使并行神经计算机系统的研制周期加长,同时开发

相应的软件系统也很困难。所以这种方法在一般的并行神经计算机系统中不是一件可取的事情,除非要设计大规模并行神经计算机系统。

另外,可以考虑采用已有的微处理器芯片作为并行神经计算机系统的处理单元,如 Intel80386 和 Intel80486 芯片、数字信号处理芯片 TMS320 系列等。这样可以避免设计实现神经元功能的处理单元所带来的麻烦,当然这种方法或许会带来硬件资源的某种程度浪费,因为现有的微处理器芯片的功能都很强,有些功能在神经网络计算中并不一定有用,但是正是由于这些微处理器芯片具有很强的处理能力,可以很好地“虚拟”实现神经网络中神经元的计算。同时这些微处理器的配套软件很丰富,相应的并行神经计算机系统的系统软件和应用软件的开发也很容易,所以这是一种很可取的设计方案,当前的多数并行神经计算机系统的设计几乎都是采用了这种方法。

在进行微处理器芯片选取时,首先主要考虑的应该是微处理器芯片的数字处理能力,各种不同的微处理器芯片具有不同的数字处理能力,比如 Intel386/486 微处理器芯片具有很强的浮点处理能力,TMS320 系列的微处理器芯片具有很快的乘加运算速度,这必须要和神经网络计算所用到的基本操作和并行神经计算机系统面向的应用紧密结合考虑。

另外要考虑的是微处理器芯片的通信接口能力,因为通常是通过微处理器对存储器的访问或微处理器芯片之间的通信来实现神经元之间互连的,所以微处理器的存储器访问速度或微处理器之间的通信速度是支持高性能实现神经元之间互连的决定因素,如 Intel80486 支持猝发方式的数据传送,Transputer 支持微处理器之间的直接通信,这些微处理器芯片都是神经计算机系统设计的一种很好的选择。

6 并行神经计算机系统的计算精度

计算精度是指神经计算在并行神经计算机上实现时,其结果可以达到的精确程度,并行神经计算机的计算精度与其模拟能力密切相关,比如并行神经计算机系统支持的计算精度是 8 位,由于不能支持浮点数表示,那么该系统就有可能不能实现多数 BP 神经网络模型。一般来说,神经计算机系统的计算精度至少应该是 16 位。同时并行神经计算机的计算精度和其性能也有很密切的联系。

7 并行神经计算机系统的软件接口能力

由于神经网络属于多学科交叉领域,许多神经网络研究和开发工作者并不是计算机程序编写的专

家,也不可能是并行神经计算机系统的管理员,所以在并行神经计算机系统中,一个很重要的设计部分就是并行神经计算机系统的软件设计。

并行神经计算机系统的软件不仅要能够很好地控制其运行,而且还要为其应用软件的开发提供很好的高层次接口,这样的软件系统才能够被用户接受,才能够具有广阔的推广应用前景。

软件的设计还必须与硬件设计紧密结合,相互联系,相互分工。这样设计出的并行神经计算机系统才具有较高的神经计算性能。

8 并行神经计算机系统的适应能力

在并行神经计算机系统的设计中,应该尽可能使并行神经计算机具有很好的“可移植”性,它包括软件的可移植性和硬件的可移植性。设计一台并行神经计算机系统要花费很多的人力物力,而当前微电子技术和计算机技术发展如此迅速,以至于今天是先进的东西到明天就过时了。所以并行神经计算机系统的设计者必须要有发展的眼光。如果并行神经计算机系统具有很好的“可移植”性,那么一旦相关技术发展,所设计的并行神经计算机系统只需对相应软件或硬件作少量修改就有可以适应相关技术的发展,提高系统的计算性能。

并行神经计算机系统的设计还要考虑到系统的可扩充性,具有良好的可扩充性可以使其适应不同的应用规模,并行神经计算机系统还应具有多种接口能力,神经网络的应用领域非常广泛,包括图形图象处理、语音处理、传感器信息处理等等,神经网络计算所面临的信息也是丰富多彩,所以一个好的并行神经计算机系统应该考虑到提供多种与外界信息的接口,使其得到更广泛的应用。

参考文献

- [1] 戴葵,通用高速变结构并行神经计算机 NeuroC-I 的研制计划,内部资料,1992
- [2] 戴葵,通用高速变结构并行神经计算机 NeuroC-I 的研究方案,内部资料,1993
- [3] 戴葵,黄光奇,王意洁,通用高速变结构并行神经计算机 NeuroC-I 软件研制文档,内部资料,1993
- [4] 戴葵,张春元,通用高速变结构并行神经计算机 NeuroC-I 硬件研制文档,内部资料,1993
- [5] 胡守仁,余少波,戴葵,《神经网络导论》,国防科技大学出版社,1993
- [6] Hecht.Nielson, Performance Limits of Optical, Electrooptical, and Electronic Neurocomputers, SPIE Vol. 634 Optical and Hybrid Computing, 1986