NEURO4システムの高速化の検討と評価

3K－3

牧田淳子　坪田浩乃　清水優子　田村俊之　田中健一　久間和生
三菱電機（株）　先端技術総合研究所

１．はじめに

高速並列処理システムであるNEURO4システムの最適化を検討したので報告する。

NEURO4は、ニューロアルゴリズムを高速に解くために開発された、12個の演算ユニットを有する最高性能1.2GFLOPSのSIMD(Single Instruction stream Multiple Data stream) 型マイクロプロセッサである(1)。このプロセッサを4個搭載したNEURO4ボードが開発され、33MHz動作時に最高性能3.17GFLOPSが達成されている(2)。

NEURO4ボードとホスト計算機を標準バス（VMEbus）で接続することで、低コスト高性能並列処理システム（NEURO4システム）を実現することができる。このような高速並列演算システムの性能を活かすためには、いかに効率よくデータを供給できるかが技術的に大きな問題となっている。

本稿では、まず、NEURO4システムの概要について述べた後、今回行った高速化手法とその評価結果について報告する。

２．NEURO4システム概要

NEURO4ボードは、1枚ごとに標準のVMEバスインターフェースを有し、同時に最大15枚までホスト計算機と接続することが可能である(2)。今回はホスト計算機として、VMEバスインターフェースを有するALPHAチップ搭載のCPUカード(AXPvme)

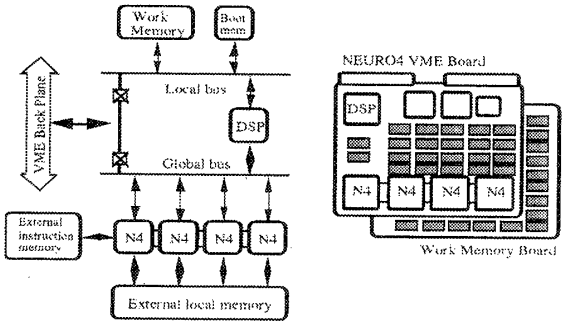

図１．NEURO4ボード構成

Improvement of Neuro4 System Performance,
Junko MAKITA, Hirono TSUBOTA, Masako SHIMIZU, Toshiyuki TAMURA, Ken'ichi TANAKA, Kazuo KYUMA, Mitsubishi Electric Corporation Advanced Technology R&D Center

を用いたシステムを構築した。また、各ボード上には、4個のNEURO4チップが搭載されている。各チップには24ビットの浮動小数点数の積和器が搭載されているので、1サイクルに48（12個×4チップ）回の積和演算が同時に実行可能である。また、これらの演算器が効率よく演算できるように、ホスト計算機から転送されたデータをダイナミックに再配置する機能を実現している汎用のDSPが搭載されている（図１）。

３．NEURO4システムのメモリ階層

NEURO4システムのメモリ階層を図２に示す。高速にアクセス可能なものから順に、PUごとにある1Kワードの内部ローカルメモリ（NEURO4プロセッサ内）、同じく16Kワードの外部ローカルメモリ

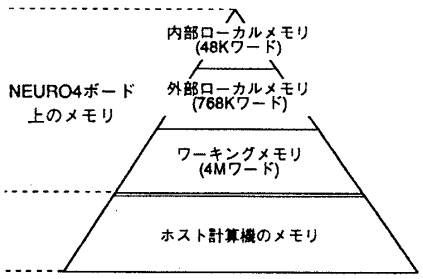

図２．NEURO4システムのメモリ階層

（NEURO4プロセッサ外）、DSPのローカルバスに接続されているワークメモリ（4Mワード）とホスト計算機のメモリという階層になる。この階層間で最も大きなギャップはホスト計算機とNEURO4ボード間にあり、転送レートは最高で約10Mbyte/秒である。一方、NEURO4ボードの演算性能は最高で、3.17GFLOPSであるために、1回の演算に2つの入力データと1つの出力データが必要だとすると、ピーク性能が達成されている時には、全演算器トータルで、3.17G＊3（≒9.4G）ワード/秒のデータの入出力が必要である。すべてのデータがホスト計算機から直接入出力されるとすると、ホスト計算機とボード間は、10M/4（≒2.5M）ワード/秒が最高レートであるので、この場合は、演算に必要な量の3800分の1しかデータが入出力されないことになる。

そこで、ホスト計算機から転送されたデータを最大限に再利用するためのキャッシュ機能を実現し

た。

4．NEURO4のキャッシュ機構

今回は演算に使用するデータを、NEURO4ボード上に常駐させてホストが必要とするときのみホストに転送する方式とした。このために、外部ローカルメモリを常駐場所（キャッシュ）とするキャッシュ機構を実現し、不要なデータ転送（図3．破線矢

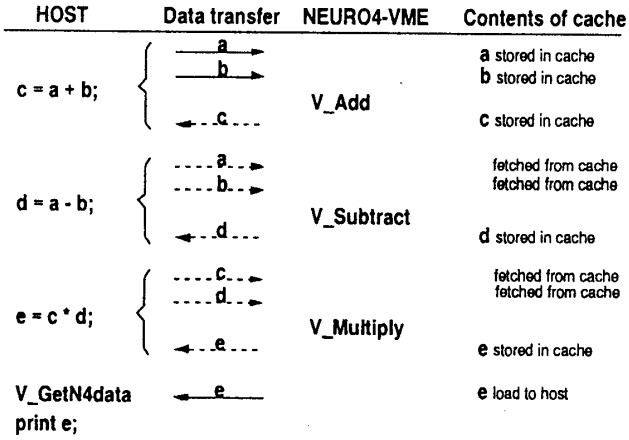

図3．NEURO4のキャッシュ機構

印）を削除した。例えば図3のような3種類の演算を行う場合、ホストからNEURO4へのデータ転送は、1回目の演算(c=a+b)時には必要であるが、それ以降は不要である。また、演算結果を途中でホストへもどす必要もない。これらを削除することで、9回必要であったデータ転送が3回に削減される。

NEURO4で演算実行に使用されるデータは、転送時にホスト計算機上のアドレスおよびサイズをキーとしてテーブルに登録する。データ転送時には、転送済みかどうかをチェックし、キャッシュに常駐していないデータだけを転送することで、不要なデータ転送を回避する。変数名ではなくアドレスで管理することで、ソースプログラム上完全に独立なスコープに属するデータも本キャッシュ機構の対象とすることができる。

また、キャッシュ機構をS/Wで実現しているため、フレキシブルなサイズのデータのキャッシュ管理を実現している。

ライブラリの速度評価結果（V_Add）

system / cache / vector size	axpvme + NEURO4	
	cache ON	cache OFF
1,008	1.37	5.65
5,040	1.50	16.27
10,032	1.73	28.53
30,000	2.44	80.08
120,000	5.87	319.75
261,840	11.16	711.05

単位 (sec)

表1．キャッシュの効果

5．評価結果

NEURO4システムでは、各種のプログラム開発ツールやライブラリの開発が完了している(3,4)。今回は、これらのライブラリ関数を用いて評価を行った。

2つのベクトルの各要素同士の加算を行うベクトル関数(V_Add)を1000回実行した場合の時間を、ベクトルサイズを変えながら、キャッシュをONした時とOFFした時について測定した。その結果を、表1に示す。表から、ベクトルサイズが大きくなるほどキャッシュの有効性が上がることがわかる。
キャッシュ機能の実現により、約256K次元のベクトルの加算で60倍以上の高速化を達成した。

また、NEURO4システムとEWS(Sparc Station 20)とを比較すると、ニューラルネットの代表的な学習アルゴリズムであるバックプロパゲーションのライブラリ関数で約80倍、ベクトル内積のライブラリ関数で約17倍の性能が得られている。

6．まとめ

ニューロプロセッサ（NEURO4）システムの高速化手法、及び評価結果について述べた。キャッシュ機構の実現で、ベクトル加算ライブラリ実行時で約60倍の高速化を実現した。今後は、制御関数や演算関数の最適化を行い、更に高速化をすすめる。

参考文献

1) Y.Kondo,"A 1.2GFLOPS Neural Network Chip Exhibiting Fast Convergence", ISSCC Digest of Technical Papers, Vol.37, pp.218-219, 1994.

2)"汎用ニューロボード", 三菱電機技報・Vol 69.No.1, pp32, 1995

3)牧田他. 1.2GFLOPSニューロチップ用S/Wシミュレータの開発 情報処理学会第50回全国大会論文集, 4B-7, (1995)

4)坪田他. 1.2GFLOPSニューロチップ用S/W開発支援環境の開発 情報処理学会第50回全国大会論文集, 4B-8, (1995)

5)Y. Kondo et al, "Silicon VLSI Neural Network Chips for Real-Time Neural Applications", Proc. Int. Conf. Artificial Neural Networks (Paris), vol. A9, pp. 7-12, 1995.