計算機アーキテクチャ151－11
（２００３．１．３１）

ニューラルネットワークのMIMDへの実装評価に関する一考察

三ヶ島敬宏　吉田秀樹

ニューラルネットワークをハードウエアで実現するに当たり、非同期な入力信号を受信できる多入力のニューロ素子を設計しなければならない。本研究でのニューロ素子は、FIFOを備えたPEとして実現し、MIMDシステムを試作した。速度向上比はPE数３個から９個の範囲でリニアとなり、速度向上比８未満の性能を観察した。

A study of the implementation of the neural network in the MIMD architecture

MIKASHIMA, Takahiro and YOSHIDA, Hideki

The purpose of our study is to evaluate the MIMD architecture, which was comprised of nine PEs (processing elements). A PE was asynchronously received the time course of data from other eight PEs via a FIFO (First In First Out) buffer, and fulfilled a part of the task of the neural network. The speed-up ratio of this trial system was linearly increased and approached eight when the number of PEs varied from three to nine.

はじめに

良く知られたAmdahl則[1]は並列プロセッサシステムの速度向上について否定的な見解を示している。これに対し、最も成功した並列システムは、やはり生物の持つ脳であるといえよう。例えばヒトの場合では、並列システムの　PE　（ Processing Element ） に相当する素子は100億とも云われるニューロンである。これは、例えPEの性能を最小限に抑えることになったとしても、各PE間の独立性を高めたシステムや、通信ボトルネック[2]の低減を工夫したシステムの有望性を示唆するものである。

脳の情報処理を模倣した計算機システムとしては、ニューラルネットワーク[3]が挙げられる。ニューラルネットワークは学習能力[4]、汎化能力[5]を備えており、非線形な写像問題[6]を解決する手段として広く利用されている。ニューラルネットワークは、汎用計算機上でシミュレーションとして実現することも可能である。この場合の問題点は、実現するニューロ素子数nの増大に伴い、積和演算がn^2オーダで増大してしまい、入力が時系列となった場合には、リアルタイムな応答が難しくなる事である。この解決手段は専用のハードウエアを製作することであるが、今度は、チップ内のニューロ素子数の上限を当初から決定しておかなければならなくなる。更にチップパッケージの物理的な制約から、I/Oピン数の上限は500本程度と見積もれる。即ち、外部から、ハードワイヤで、500入力を超える接続を可能にした半導体ニューロ素子は、目下、現実的ではない。

我々は、後述のMIMD（ Multiple Instruction stream Multiple Data stream ）システム （ Fig.1の黒塗り部を参照 ） を使用して、学習後の （ 重み係数決定後の ） ニューラルネットワーク演算を実現することを試みた。提案した　MIMD　の構成は、特定目的[7]の為に試作したシステムではあるが、ニューラルネットワークとトポロジーが近似していることに着目して実装を試みた。システムに時系列データを入力した際に、速度向上比とスループットの観点から性能評価を実施した。

北見工業大学情報システム工学科
Department of Computer Sciences, Kitami Institude of Technology

方法

我々は本研究に先立ち、高い時間分解能と周波数分解能とを両立する周波数解析手法を提案[7]し、音響をリアルタイムに分析するシステムを試作した。それは先ず、入力信号を１オクターブ帯域の帯域通過フィルターに通すことで前処理を行う。得られた出力波形について、波形の極大値から極小値までの時間、および極小値から極大値までの時間を計測し、周波数に相当する情報を算出するものである。こうして概周期波形のみならず非定常波についても、観察する各１オクターブ帯域の中から、最も支配的な周波数相当成分を算出する手法を提案した。可聴帯域についてシステムを設計した場合には、例えば、64Hz-128Hz、128Hz-256Hz、256Hz-512Hz、512Hz-1024Hz、1024Hz-2048Hz、2048Hz-4096Hz、4096Hz-8192Hz、8192Hz-16384Hzの８チャンネルについてフィルターバンクを構成することが提案される。得られる出力は、毎時、要素数８個を上限とするベクトルとなり、可聴音響の時間-周波数特性を近似する。ここで時間分解能は、算出された周波数の周期の２分の１となる。システムを設計する際の問題点は、非同期に出力される８チャンネル分のデータを、取りこぼすことなく収集できる様にする点であった。

Fig.1の黒く塗った部分が設計したMIMDである。各PEはハードウエアで実現されたFIFO（ First In First Out ）を備えており、８個のPE（前段）が９個目のPE（次段）に、並列にハードワイヤ接続されている。本例には次段のニューロ素子が１個の場合を示したが、次段のニューロ素子は電気的特性の許す限り、必要数増設する事ができる。前段のニューロ素子も８個となってはいるが、これはFIFOを実現するCPLD（ Complex Programmable Logic Device ）の回路規模とパッケージの制約を受けている。一段当たり数千、数万のニューロ素子を実現するには、全入力データを時分割で転送する方式が現実的である。即ち、Fig.1の黒く塗った部分が初回に転送される８素子分のデータの流れを示している。次回にはFig.1の白抜きで示した８素子分に相当するデータを転送し、以下必要回数繰り返す。こうして全ての入力データを９個目のPEに転送させた後で、９個目のPEは演算結果を出力できる様になる。もし次段も８個のPEが設置されていると仮定すると、９個目のPEは自分自身の演算結果と、８台間隔で下行に配置されるはずのPEの演算結果とを、時分割方式で出力することになる。

Fig.2にCPLD上で試作したFIFOのブロック図を示した。S-RAM（ Static Random Access Memory ）を外付けする事で、大容量のFIFOを安価に実現した。８個のPE各々から送信されるのは1bitの情報であり、独立に入力バッファに蓄えられる。こうしてまとめて8bitsにした情報を、S-RAMの書き込みポインタが示すアドレスに記憶させた。もし出力バッファからのデータ読み出しが既に実施されていれば、書き込み動作に引き続き、読み出しポインタが示すアドレスからデータが読み出され、出力バッファが更新される。試作したFIFOを高速連続アクセス（最小アクセス間隔1μs）することはできないが、書き込み動作と読み出し動作を同時に実施することは勿論許される。ここで、PEとしては日立製作所製 H8/3048F（ 16MHz ）を使用し、CPLDとしてはXilinx社製XC95108-15PC84C(108マクロセル、2400ゲート)、S-RAMには1Mbits、アクセスタイム120nsの部品を使用した。

先ず、PE数に対する速度向上比を計測した。システムが実現するニューラルネットワークの規模は、前段840素子、2520素子、4200素子が各々同数の次段に接続される場合の３通りについて計測した。PEが１個の場合は、主記憶上に入力データと重みデータを確保し、ニューラルネットワークを逐次処理により実現した。PEが２個の場合は、前段に１個、次段に１個を配置しハードワイヤ接続した。以下同様にして、前段８個、次段１個、計９個になるまで続けた。各ニューロ素子の機能は、前段から16bitsのデータがシリアルに入力されるものとし、既に主記憶上に用意された重みデータ（16bits）との積和演算を実施した後、非線形な演算操作により16bitsに加工して、再びシリアルに出力するものと定義した。速度向上比の算出にあたっては、前段に入力されたデータが処理されて次段から出力されるまでの時間Ｔを計測した。特にＰＥが１個の場合には、ＰＥを複数使用した時と等価な作業量を、逐次処理により実施させた。尚、本研究では、ニューラルネットワークの学習動作は実施せず、学習済みの重み係数を主記憶上に初期設定して利用した。

次に、PE数を９個（前段８個、次段１個）に固定し、ニューラルネットワークの規模（ニューロ素子数）に対する速度向上比とスループットを計測した。スループットの算出にあたっても、前段に入力されたデータが処理されて次段から出力されるまでの時間Ｔを計測した。従ってスループットはＴ/２の逆数として表され、単位時間あたり（ここでは１秒当たり）に処理可能な入力パターン数を意味することになる。

結果

Fig.3にPE数に対する速度向上比を示した。細線は理想とされる線形速度向上を示している。実線がニューロ素子数として4200素子、破線が2520素子、点線が840素子の場合を示した。いずれの場合もPE数が２個の時の速度向上比には、PE数が１個の時に比べ、僅かに劣化が観られた。しかしながら、PE数が３個以上では、PE数が９個に到達するまでリニアな速度向上が観察された。これはPE数の増加に伴い、速度向上比が頭打ちとなるAmdahl則に反している。また、ニューロ素子数の増大に伴い、速度向上曲線の傾きが僅かに増大していた。

そこでFig.4にはニューロ素子数Ｍに対する速度向上比とスループットを重ねて示した。計測はPE数をＮ=９個に固定して行い、白丸が速度向上比の計測点を、菱形がスループットの計測点を示している。速度向上比はニューロ素子数の増大に伴い対数的に増大するが、Ｎ-１=８に漸近した。一方、スループットはニューロ素子数の増大に伴いほぼリニアに減少してしており、両者にトレードオフの関係が成立していた。

考察

(i) 何故、PE数が２個の時に速度向上が望めなかったのか？

本研究では、PE数を２個にした時の並列計算機の性能は、単体の逐次計算機の性能未満であると云った悲観的な結果を得た。FIFOを導入したのは、特別な通信プロトコルを必要とせず、主記憶をアクセスする手間でデータ転送が実現できるからである。しかしながら一度に1bitの情報しか転送できない本試作機の仕様により、通信ボトルネックの低減には充分な効果を発揮できなかった。本機は特定用途に向けられた設計である。本機をニューラルネットワークの様な汎用性のある使い方をした場合には、一度に16bits程度のデータ転送が要求され、速度向上に反映されなかった。本実験では、前段より16bitsのデータがシリアルに送信されてくるので、PEはシリアル-パラレル変換処理（通信時間）に無視できないCPU時間を費やしたものと考えられる。

(ii) 何故、PE数が３個以上ではリニアな速度向上が観察されたのか？

速度向上比を増大させるには、並列演算の割合を増やす事、通信処理より演算処理自体の割合を増やす事、適切な負荷分散により休止中のPEを減少させる事が努力目標に掲げられる。本試作機に使用したFIFOは最大８個のPEから入力を受信できる設計であった為に、FIFOに接続されたPE数は１個から８個の範囲で変化しても通信時間は一定となる。また本FIFOへのアクセスは、連続読み出し間隔を1μs以上、連続書き込み間隔を1μs以上あけることが条件ではあるが、FIFOへのアクセスに要する手間は、主記憶のアクセスに要する手間と等価である。負荷分散についても、ニューラルネットワークの特性上、各PE素子での作業量は完全に等価であることから、最適な負荷分散が実現されていたことになる。以上から、通信時間一定、負荷は常時均等であり、PE数の増加は直ちに理想的な速度向上比の増加につながったと考えられる。

(iii) 何故、ニューロ素子の増加は速度向上につながったのか？

前段と次段のニューロ素子数が共にｎ個で等しいニューラルネットワークでは、ｎ個の入力信号に重み係数を乗ずる為のｎ^{2}回の積演算が実施されることになる。本実験では前段に8個のPEを配置して、これを次段の１個のPEへデータ転送する構成であった為、前段の全てのPEからデータを収集するにはｎ/８回に時分割してデータを転送しなければならなかった。一方、次段の１個のPEはｎ/８個分のニューロ素子の働きを兼ねていた為、積演算はｎ^2/８回必要であった。即ち、ニューロ素子数nが増加すれば、並列演算処理の手間がｎ^{2}のオーダで増大した分、相対的にシリアル-パラレル

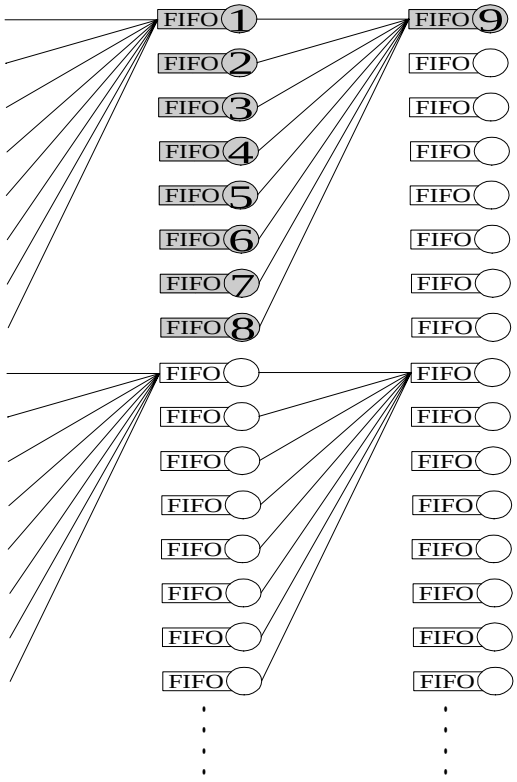

Fig. 1 The filled structure shows the proposed computational architecture embedded in the neural network. The eight processing elements (PEs) were connected to the FIFO of the nineth PE. The connection with more than eight PEs was realized by the time-division manner.

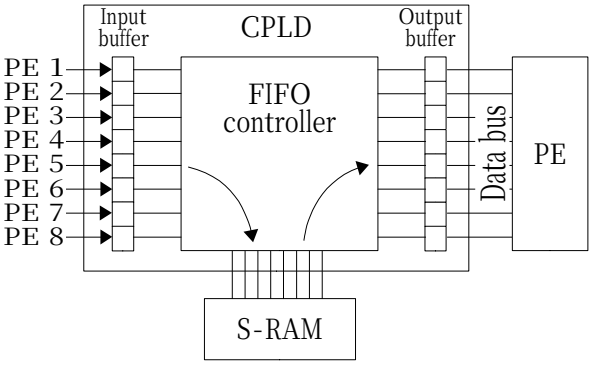

Fig. 2 The block diagram of the FIFO controller constructed by a Complex Programmable Logic Device. Each one-bit data from eight PEs was buffered asynchronously, and was stored in a Static Ramdom Access Memory sequentially. An output data was sent to the data bus via the attached output buffer after the aforementioned storing process if required.

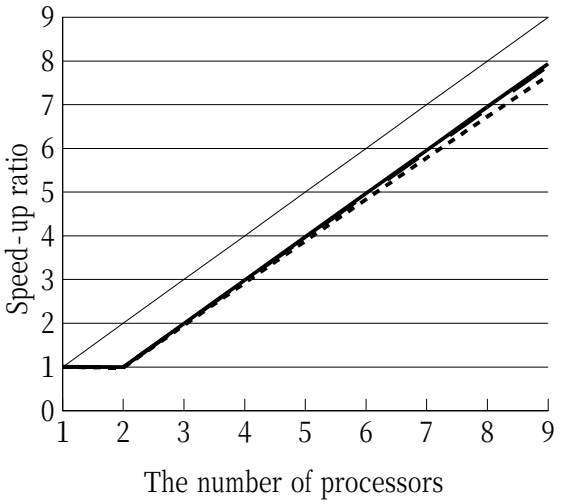

Fig. 3 The speed-up ratio vs the number of processors, where the proposed architecture realized 840, 2520, and 4200 neural elements. The thin line refers the linear speed-up.

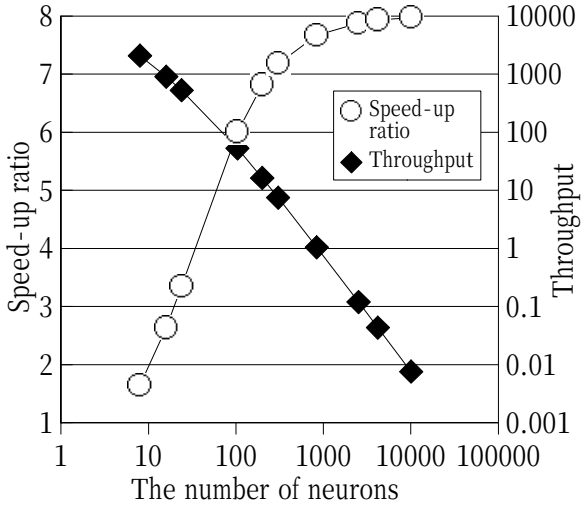

Fig. 4 The circles and diamonds show the speed-up ratio and throughput as a function of the number of neurons M, respectively, with the proposed architecture by using nine processors (N=9). Although the speed-up ratio approached N-1 (M→∞), the speed-up and throughput had a trade-off.

変換処理（通信処理）の手間は減少したとみなされる。以上から、ニューロ素子の増加も速度向上比の増加に寄与したと考えられる。

(iv) FIFO結合型のMIMDシステムの有効性について

プロセッサの目覚ましい性能向上が実現する中、本実験で使用したPEは決して単体での計算速度が著しく早い分類には入らない。しかしながら、脳を模倣したシステムを構成する際には、最小限度の機能を備えたPEを密結合するアプローチは正攻法であると考える。実際、本実験からは、システム全体としてみても小規模ではあったが、リニアな速度向上を達成した。脳がシステムとして破綻しないのと同様に、本試作機もMIMDシステムの将来性について楽観的な一性能評価結果を示した。残念なことには、速度向上比とスループットはトレードオフの関係にあり、仮に毎秒100件のパターン認識をリアルタイムに実現するには、本試作機では、一段当たりのニューロ素子数は100個程度に制限されることがFig.4からわかる。尤も、これはPEとFIFOの動作速度の向上によって改善できる値ではある。

まとめ

PE間をFIFOで接続したMIMDシステムに、ニューラルネットワークを実装した。速度向上比はPE数３個から９個の範囲でリニアとなり、速度向上比８未満の性能を観察した。

謝辞

本研究は（財）北海道科学技術総合振興センターからの助成を受けて実施された。

参考文献

[1] Amdahl G (1967) Proc. 1967 AFIPS Conf., Vol. 30, pp.483.
[2] 近山隆 (1989) 数理科学 27, 7, 26-33.
[3] Rosenblatt F (1961) Principles of neurodynamics, Spartan.
[4] Rumelhart DE, Hinton GE and Williams RJ (1986) Nature 323, 533-536.
[5] Fukushima K (1984) Biological Cybernetics 50, 105-113.
[6] Dayhoff RE and Dayhoff J E (1988) IEEE SCAMC Proceedings, 271-275.
[7] 吉田秀樹 (1999) 特許第2988914号、周波数解析装置及びその方法.

EOF