全ハードウエア形高速ニューロプロセッサの設計

苫米地　宣　裕

Design of Full-Hardware High-Speed Neuro-Processors

Nobuhiro TOMABECHI

Abstract

This paper shows the design of the high speed neuro-processors in which 200～1,000 neurons are integrated and all of the arithmetic operations and the signal transmission are realized by hardware technology. The sizes of the processors become wafer scale. It is concluded that a neuro-processor consisted with 500 neurons may be implemented on a single wafer.

Keywords: full-hardware/high-speed/neuro-processor/design

1. まえがき

ニューラルネットワークのリアルタイム制御など高速処理への応用が期待されている[1]。通常，ニューラルネットワーク機能は汎用プロセッサを用いてソフトウエアによって実現されているが，高速処理を実現するためには，ハードウエアで構成されたニューロン(実ニューロン)を搭載した専用プロセッサが必要となる。これまで，数十ないし数百個の実ニューロンを搭載したプロセッサの開発が報告されているが[2]-[5]，より一層多くのニューロンを集積したプロセッサが望まれている。また，ニューロプロセッサではニューロン間の相互配線が大きなチップ面積となるため，信号伝送を時分割で行うことが多く，その分プロセッサの演算速度が低下する。

本論文では，200～1,000 個の実ニューロンを搭載し，しかも，演算機能や信号伝送機能は時分割することなしに，すべてハードウエアで構成するニューロプロセッサの設計を行っている[6]-[7]。これらのプロセッサは WSI の規模となる。結論として，現在の技術水準で，ニューロンを 500 個搭載したプロセッサが 1 枚のウエーハー上に構成できる可能性のあることが分かった。

なお，WSI の製造においては，欠陥の増加による歩留り低下が最大の問題となる。その対策として，冗長構成を導入して欠陥救済を行う方法が考えられるが，本論文では，第 1 段階として，欠陥救済を行わない場合について論じている。

2. 仕様の設定

本論文では，リアルタイム制御などに用いる高速実行型のニューロプロセッサを対象とする。その仕様を表 1 に示している。高速性を達成するために，演算機能や信号伝送機能は時分割することなしに，すべてハードウエアで実現することとしている。また，学習機能はなしとしている。このとき，学習は他の汎用プロセッサを用いて行い，そこで得られたシナプスの重み係数を本プロセッサ内部の重みメモリに入力するという使い方をする。表 1 に示す 3 層フィードフォワードニューロプロセッサの動作は次式で表される。但し，y_j^k, $w_{i,j}^k$, h_j^k, $f(\ )$, および，N は，それぞれ，第 k 層第 j 番目のニューロンの出力，第 k 層第 j 番目のニューロンの第 i 番目の入力に対する重み，第 k 層第 j 番目のニューロンのしきい値，シグモイド関数，および，一つの層のニューロンの数を表している。

$$y_j^k = f\left(\sum_{i=1}^{N} w_{i,j}^k y_i^{k-1} - h_j^k\right) \quad k=1,2, \quad j=1,2,\ldots,N \tag{1}$$

なお，$k=1$ は中間層，$k=2$ は出力層に対応し，y_i^0 は入力層 i 番目の出力を表している。

表 1　ニューロプロセッサの仕様

ソフトウエア/ハードウエア	ハードウエア
アナログ/ディジタル	ディジタル
時分割/全ハードウエア	全ハードウエア
形式	3 層フィードフォワード型
ニューロンの数　N	中間層・出力層とも 100～500
ニューロン間の配線	完全配線
1 ニューロンのシナプスの数	各層のニューロン数に同じ
入出力データの精度	16 ビット
重みメモリのビット数	16 ビット
素子	CMOS
学習機能	なし(シミュレーションで学習)

3. アーキテクチャ設計

式 (1) に基づくニューロプロセッサのシステム構成を図 1 に，1 個のニューロンの回路構成を図 2 に示してい

平成 9 年 12 月 15 日受理
八戸工業大学 電気工学科 教授

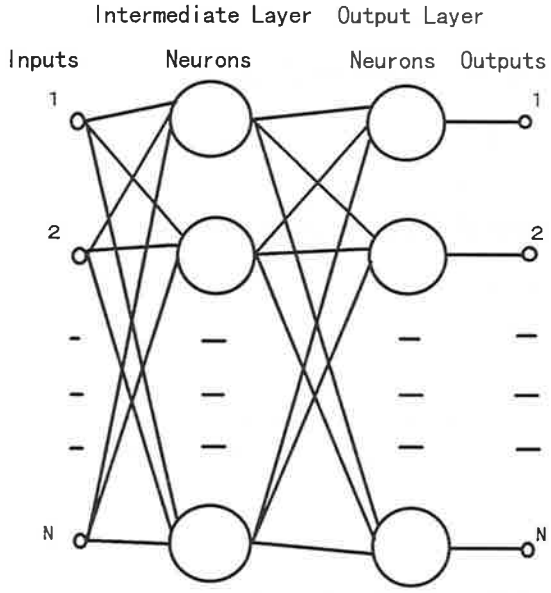

図1　ニューロプロセッサの全体構成

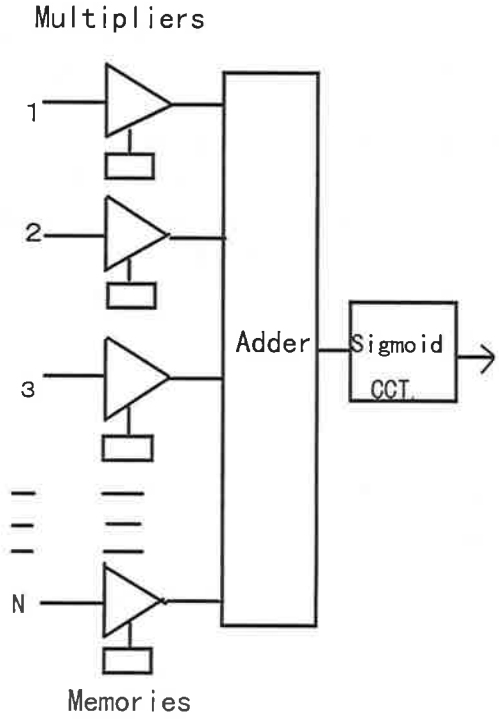

図2　ニューロンの回路構成

る。本論文では，ニューロン内の各回路を以下のように構成する。

①　乗算器

負数の演算が補数のままで実行できるBaugh-Wooleyの乗算器を用いる。本乗算器は，入力信号をnビット×nビットとすると，$n(n-1)+3 \fallingdotseq n^2$個の全加算器（但し，ANDゲート1個付）で構成される。図3に，$(a_4\ a_3\ a_2\ a_1\ a_0)\times(b_4\ b_3\ b_2\ b_1\ b_0)$の構成例を示している[8]。

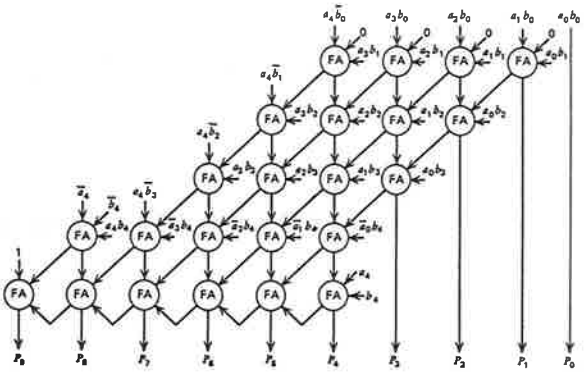

図3　乗算器の構成

②　重みメモリ

プロセッサの外部からデータを入力するのに好都合となるようシフトレジスタで構成し，すべてのレジスタを直列接続して用いる。以下，重みメモリは乗算器に含めて取り扱う。

③　加算器

N＝100～500個という多数のデータをできるだけ少ないハードウエアと少ない段数で加算する点が問題となる。本論文では，2入力加算器（但し，各入力が16ビット）を用いて，2個づつのデータを加算していく方法，いわゆる加算器の2進木構造を用いることとする。本構成に必要な2入力加算器の数N_Aは次式で与えられる。

$$N_A = N-1 \tag{2}$$

2入力加算器が直列接続される段数Sは次式で与えられる。但し，INT[X]は，X以上の値となる最小の整数を表している。

$$S = \mathrm{INT}[\log_2(N-1)] \tag{3}$$

2入力加算器の具体的回路は，リップルキャリアダーとする。リップキャリーアダーはキャリーの伝播に時間がかかるという問題を有するが，多数の加算器を用いて2進木構造とした場合は最終段のアッダー1個分のキャリー伝播時間が問題となるだけある。

④　シグモイド関数回路

いくつかの直線，または，いくつかの2次曲線により近似する[9]。ここでは5個の直線で近似することとする。このとき，直線を$y=a_ix+b_i(i=1,2,\ldots,5)$と表わすと，シグモイド関数回路は，$x$の値の範囲を判定する5個の比較器，$a_i, b_i$の選択ゲート，および，$(a_ix+b_i)$の計算回路で構成される。なお，しきい値のメモリレジスタとしきい値の減算器はシグモイド関数回路に含めて取り扱う。

4. レイアウト設計

(1) フロアプラン

① ニューロプロセッサ全体のフロアプラン

図4に示している。100～500個のニューロンが縦方向に並ぶので，ニューロン1個のレイアウトは，可能な限り横長に設計する必要がある。

② ニューロン1個のフロアプラン

図5に示している。レイアウトを横長にするため，すべての乗算器を横方向に1列に配置している。入力バスラインはバラバラにして，各々の乗算器に入力される配線だけをそれぞれの乗算器の直前に配置している。2入力加算器は，乗算器2個ごとに1個挿入する。また，2入力加算器2個に対して1個の2入力加算器を付加していく。2入力加算器の入出力配線は，16ビットをまとめないで，1ビットごとに全加算器(以下，FAという)とFAの間を通過させる。

(2) レイアウト設計

レイアウト設計規則は，Mead-Conwayの設計規則[10]をCMOSに拡張した規則[11]を用いる。図6に，ANDゲート1個付きFAのレイアウトを示している。但し，λは基準寸法を表している。FAを元にした各回路の形状を表2に示している。図7は，ニューロン1個のレイアウトを示している。但し，回路の種類ごとにまとめた寸法を示している。結局，ニューロプロセッサの縦の長さh，横の長さl，および，面積A_Pは次のようになる。

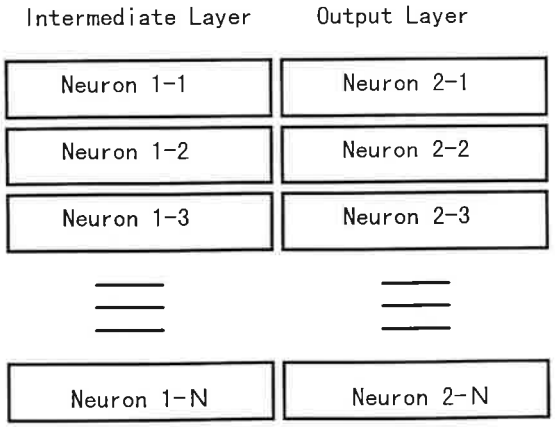

図4 ニューロプロセッサのフロアプラン

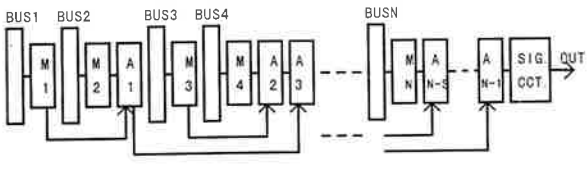

図5 ニューロンのフロアプラン

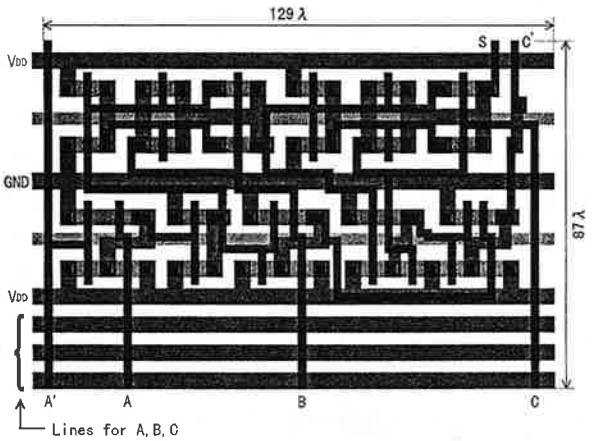

図6 FAのレイアウト

表2 各回路の形状とチップ面積

回　路	形状 縦(λ)×横(λ)	チップ面積 (λ^2)
FA	129×87	1.12×10^4
乗算器	2,064×1,392	2.88×10^6
重みメモリ	41×66	2.71×10^3
2入力加算器	2,064×87	1.80×10^5
シグモイド回路	2,064×2,349	4.85×10^6
入力バスライン	2,064×(112×N)	$2.31\,N\times10^5$

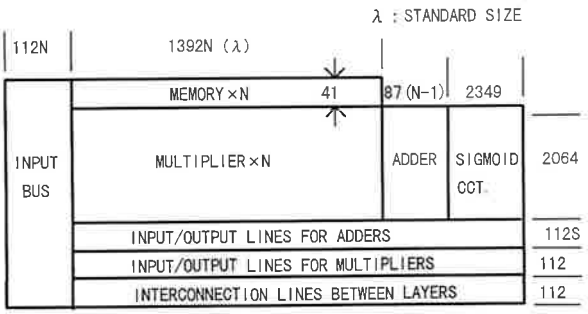

図7 ニューロンのレイアウト

$$h=(112S+2329)N \quad (\lambda)$$
$$l=2(1591N+2262) \quad 〃$$
$$A_P=2N(112S+2329)(1591N+2262)\ (\lambda^2) \quad (4)$$

例として，$N=100$ととると，縦横比は，縦311,300対横322,724≒1対1となり，これはWSIに適した形状といえる。式(4)より，h, lともにNにほぼ比例するので，h対l≒1対1の関係は，Nの値によらず成り立つ。従って，ここに示したレイアウトは，任意のNの値に対して有効であるということができる。

また，$N=100$の場合，ニューロプロセッサの面積A_Pは$1.00\times10^{11}(\lambda^2)$となるが，これは，基準VLSI[12]のチップ面積の33倍となり，本プロセッサがWSIに近い規模であることを示している。

5. 考　　察

式(4)より，ニューロプロセッサの面積 A_P は N^2 で増大する。図8に，全ニューロン数 $T(=2N)$ 対 A_P の関係を示している。今，利用可能なウエーハーの最大寸法を10インチ，$\lambda=0.2\,\mu$ とすると，WSIの最大面積 A_M は，$(25.4/(0.2\times10^{-4}))^2/2=8.06\times10^{11}(\lambda^2)$ となる。図8より，$A_P=A_M$ となる T の値を求めると約500であることが分かる。よって，本論文で得られた結果を次のようにまとめることができる。

［結論］ ニューロンを500個搭載したプロセッサが1枚のウエーハー上に集積できる可能性がある。

WSIの製造では，製造欠陥の増加による歩留りの低下が最大の問題となる。その対策として，冗長構成を導入して欠陥救済を行うことが考えられる[7]。この点については別報にて論ずる。

6. む　す　び

本論文では，ニューロンを200～1,000個搭載し，しかも，すべての演算機能をハードウエアで構成する高速ニューロプロセッサの設計を行った。その結果，現在の技術水準で，ニューロンを500個搭載したプロセッサが1枚のウエーハー上に集積できる可能性があることが分かった。

今後，学習回路を組み込んだ場合について設計を行う予定である。

また，WSIの製造では，製造欠陥の増加による歩留りの低下が最大の問題となるが，この問題に対しては，冗長構成を導入して欠陥救済を行う方法を検討しており，別途報告する予定である。

なお，本研究は平成8年度文部省科学研究費（基盤研究C）の補助を一部受けたことを付記する。

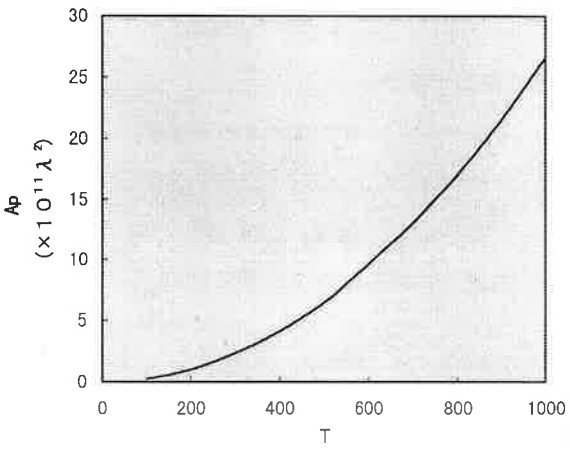

図8　全ニューロン数 T 対 A_P の関係

文　　献

［1］ 岩田，雨宮，“ニューラルネットワークLSI,” 電子情報通信学会，1995.

［2］ H. Kato, H. Yoshizawa, et. al., “A parallel neurocomputer architecture towards billion connection update per second,” Proc. *IJCNN* **'90**, pp. 47-50, 1990.

［3］ M. Yasunaga, N. Masuda, et. al., “Design fabrication and evaluation of a 5-inch wafer scale neural network LSI composed of 576 digital neurons,” Proc. *IJCNN* **'90**, pp. 527-535, 1990.

［4］ M. Griffin, G. Tahara et. al., “An 11-million transistor neural network execution engine,” *ISSCC* **'91**, pp. 180-181, 1991.

［5］ Y. Kondo, Y. Koshiba et. al., “A 1.2GFLOPS neural network chip exhibiting fast convergence,” *ISSCC* **'94**, pp. 218-219, 1994.

［6］ 苫米地宣裕，“フィードフォワードニューロWSIの階層冗長構成，” 電子情報通信学会1997年総合大会講演論文集，p. D-2-20, 1997.

［7］ 苫米地宣裕，“冗長化によるWSI規模ニューロプロセッサの構成法，” 電子情報通信学会技術研究報告，vol. ICD97-30, pp. 57-64, May 1997.

［8］ Kai Hwang, “Computer Arithmetic,” John Wisely & Sons, Inc., 1979.

［9］ 林原香織他，“シグモイド関数の連続性/離散性とニューラルネットワークのマシン能力について，” 電子情報通信学会論文誌D-II, vol. J73-D-II, no. 8, pp. 1220-1226, Aug. 1990.

［10］ C. Mead and L. Conway, “Introduction to VLSI systems,” Addison-Wesley Pub., 1980.

［11］ 松山泰男，富沢　孝，“VLSI設計入門，” 共立出版，1983.

［12］ 苫米地宣裕，金沢正治，“WSI規模高速FFTプロセッサの冗長化，” 電子情報通信学会論文誌D-I, vol. J80-D-I, no. 1, pp. 110-120, Jan. 1997.