学習機能を搭載した連想記憶アナログニューラルネットワークLSIに関する研究

有馬　裕

学習機能を搭載した連想記憶アナログ
ニューラルネットワークLSIに関する研究

有馬　　裕

目次

第1章　序論　1

1.1 序 1

1.2 本研究の背景 4

1.3 本研究の目的と意義 10

1.4 本論文の構成 13

参考文献 16

第2章　ニューロ連想メモリーデバイスの高集積化　20

2.1 序 20

2.2 連想メモリーデバイス 21

2.2.1 連想メモリーの実現方式 21

2.2.2 ニューロ連想メモリーデバイスに要求されるシナプス精度 25

2.3 ニューロ連想メモリーの高集積化 29

2.3.1 ニューラルネットワーク表現回路の方式比較 29

2.3.2 アナログ集積回路の問題点 34

2.3.3 素子特性バラツキの連想性能への影響 42

2.4 学習機能のチップ実装による素子特性バラツキ補償能力 43

2.5 まとめ 47

参考文献 49

第3章　学習機能を搭載したニューラルネットワークの高集積化　50

3.1 序 50

3.2 連想記憶ニューラルネットワークの機能モデル 50

3.3 学習機能を備えたニューロン回路 53

3.4 学習機能を備えたシナプス回路 55

3.4.1 回路構成 55

3.4.2 シナプス荷重値修正の非線形特性 58

3.5 学習機能搭載ニューラルネットワークの回路構成と制御フロー 61
3.6 学習性能と動作マージン 62
3.6.1 回路の簡略化に伴う学習ルール近似表現の影響 62
3.6.2 電源電圧変動に対するシナプス回路動作マージン 70
3.7 連想記憶ニューラルネットワークLSIの素子微細化トレンド 72
3.8 まとめ 74
参考文献 75

第4章　学習機能搭載ニューロチップ 76
4.1 序 76
4.2 125ニューロン・10Kシナプス集積ニューロチップ 77
4.2.1 シナプス回路 77
4.2.2 ニューロン回路 82
4.2.3 チップ構成 84
4.3 336ニューロン・28Kシナプス集積ニューロチップ 86
4.3.1 シナプス回路 86
4.3.2 ニューロン回路 89
4.3.3 チップ構成 91
4.4 400ニューロン・40Kシナプス集積ニューロチップ 92
4.4.1 シナプス回路 93
4.4.2 ニューロン回路 96
4.4.3 チップ構成 96
4.5 オンチップ学習機能評価 98
4.6 まとめ 103
参考文献 105

第5章　マルチチップ拡張機能搭載ニューロチップ 107
5.1 序 107
5.2 マルチチップ拡張接続による規模拡張方式 108
5.3 マルチチップ拡張機能搭載ニューロチップ 110

5.3.1 BNUアーキテクチャーチップ構成 110
5.3.2 マルチチップ拡張性能評価 111
5.4 18チップ拡張接続ニューロボード 119
5.4.1 ボード構成 119
5.4.2 学習能力評価 122
5.5 まとめ 126
参考文献 127

第6章　高速リフレッシュ機能搭載ニューロチップ 128
6.1 序 128
6.2 荷重値リフレッシュ方式 129
6.3 マクロリフレッシュ機能搭載ニューロチップ 133
6.3.1 リフレッシュ制御回路 134
6.3.2 リフレッシュ専用サブネットワーク 136
6.3.3 スタティックシナプス回路 138
6.3.4 リフレッシュ制御フロー 141
6.3.5 リフレッシュ機能評価 142
6.4 マクロリフレッシュ方式の有効性 143
6.5 各種不揮発性連想記憶ニューラルネットワークLSIの微細化トレンド . 146
6.6 まとめ 148
参考文献 149

第7章　総括 151

付　章　時分割規模拡張方式 156

謝辞 158

研究発表リスト 159

第1章
序論

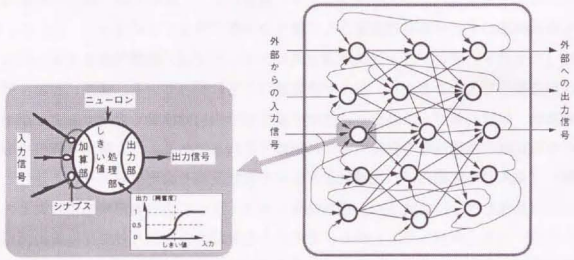
1.1 序

今日のコンピュータは、1970年代以来、半導体集積回路技術の進展に伴うハードウエア性能の向上と蓄積された膨大なソフトウエア資産によって極めて高い経済性が実現され、あらゆる機器の中央情報処理装置として様々な分野で普及するに至った。しかし、現在主流の（ストアド）プログラム逐次処理方式コンピュータは、処理手続きを明示するプログラムが必要不可欠なことから、我々が無意識のうちに行っている、連想、認識、組み合わせ最適化、経験に基づく予測など、処理手続を明示的に効率良く記述することが困難な、いわゆる直観的情報処理を表現するのに極めて不向きである。ここ数十年間の半導体集積回路デバイスの高性能化は、加減剰余算などの数値演算や論理演算処理の高速化とメモリーの大容量化を中心に実現され、逐次処理に基づく既存の情報処理の処理時間と装置コストの低減に大きく貢献した。しかし、そのような量的高性能化だけでは直観的情報処理や学習に基ずく柔軟な知識情報処理を実現する質的進展に直接結び付けることが困難なことから、新たな原理に基づく情報処理方式の実用化に対する期待が高まっている。

そこで近年、脳の情報処理様式に着目して直観的情報処理や柔軟な学習機能を工学的に実現しようとするニューラルネットワーク技術に関する研究が盛んに行われるようになり、現在では実用化研究の段階に入りつつある[1][2][3][4]。その中でニューラルネットワークのダイナミックスに基づく連想メモリーは、学習によって記憶構造が自動的に形成され高速な連想が実現できることから高度な知識情報処理装置のキーデバイスとしてその実用化が期待されている。

ニューラルネットワークは、生体脳内の神経回路網を単純化した機能モデルで、図1.1aに示したニューロンと呼ばれる機能単位が図1.1bのように多数相互接続されて構成された信号処理回路網である。各ニューロンでは、固有の信号伝達率を有したシナプスと呼ばれる接続端を介して他のニューロンからの信号が多数入力され、その入力信号の総和が各ニューロン固有のしきい値を超えた場合に、ニューロンの活性状態を示す信号を出力する単純な非線形演算処理が随時行われている。ニューラルネットワークにおいて取り扱う情報

は、複数のニューロンに分散して表現され、それぞれが非同期並列に処理実行されるので、極めて冗長な情報処理が実現され、部分的な機能不良や入力情報の曖昧さに対して高いロバスト性が発揮される。また、ニューラルネットワークの冗長な機能構成は、各シナプスの信号伝達効率を局所情報に基づいて修正することによって所望の信号処理機構を次第に自己組織化する、いわゆる学習機能を簡単な手続きで実現することができる[5]。この学習機能によってニューラルネットワークは、処理手続を明示的に記述することが困難な、連想やパターン認識などの直観的情報処理を効果的に実現することが可能となる。



a．ニューロンの機能　　b．ニューラルネットワーク構成例

図1.1　ニューラルネットワークの機能構成

ニューラルネットワークを実現する方法としては、その機能モデルを現行のコンピュータを用いて直列的逐次処理に焼き直して実行する、いわゆる計算機シミュレーション手法がニューラルネットワークモデルの実験・研究などに多く用いられている[6][7]。しかし、ニューラルネットワークをシミュレーションするためには、その冗長な処理機構の特徴によって極めて膨大な量の演算を実行する必要があり、数千ニューロン・数百万シナプスを超える実用規模[6][7]のニューラルネットワークを生体脳並の時間（数十～数百ms）で処理実行することは、高度の並列処理が実行できる高価なスーパーコンピュータを除いて現行のコンピュータでは到底困難である。

そこで、1980年代半ば頃から、ニューラルネットワークを高速に処理実行することを目

的とした、ニューラルネットワーク専用の半導体集積回路（ニューロチップ）に関する研究が盛んに行われるようになった[6][7]。現在までに様々な回路方式のニューラルネットワークLSIが提案・試作されているものの未だ実用化されているものは少ない。特に大規模な信号フィードバックがある連想記憶ニューラルネットワークに関しては、情報のコード化により離散時間で演算処理を行う従来のバイナリデジタル回路方式ではネットワークの緩和過程を繰り返し演算で表現する必要があり演算処理量が膨大となる結果、数百から数千程度の処理並列化では実用的なネットワーク規模の連想メモリーデバイスを実用化するのに不十分である。一方、アナログ回路方式によればニューラルネットワークの機能表現に関し次の優れた点を有し、少ない素子数で高い情報密度の処理回路が構成でき演算機能とメモリー機能を一体とした完全並列処理回路構成によって高い集積度と共に極めて高い演算処理速度が実現できる。

- シナプス結合演算（積演算）を極めて少ない素子数で実現できる。
- ニューロンの入力総和演算（多項和演算）をワイヤー接続（電流加算）で実現できる。
- ニューロンの非線形変換を極めて少ない素子数で実現できる。
- 高集積に伴う大規模並列処理により高速処理が実現できる。
- 並列処理構成により信号のフィードバックを実時間で表現できる。
- アナログ信号に基づく演算により演算処理当たりの消費電力が極めて少ない。

これらの優位性によって、アナログ回路方式のニューロチップはデジタル回路方式と比べて数百倍以上の高集積および高速処理性能を実現することができる[6][7]。しかし、更に大規模なアナログニューラルネットワークLSIを実用化するためには、次に上げるアナログ集積回路特有の問題点を克服する必要がある。

▲ 素子特性のバラツキが演算性能に強く影響し演算精度を高くすることが困難。

▲ 素子特性のバラツキは素子の微細化・集積化が進むに伴い増大する。

▲ アナログ値を電圧で表現するので電源電圧に関する動作マージンが極めて低い。

▲ アナログ値の長時間安定保持が困難。

本研究論文は、これらアナログ集積回路特有の問題を克服するために学習機能をチップ上に実装する研究を行い、実際に試作した学習機能実装アナログニューラルネットワークLSIによってその基本性能を実証した[8][9][10]、一連の研究をまとめたものである。学習機能をチップ上に実装することは、連想メモリーとしての記憶過程が高速に実行可能になると同時に、オンチップ学習機能による素子特性バラツキの補償を実現して素子の微細化に伴う連想性能の低下を防ぎ、高集積で大規模な連想記憶アナログニューラルネットワークLSIを実現可能とする。

本章では、続く第1.2節において本研究の背景を述べ、第1.3節で本研究の目的と意義を明らかにし、最後に第1.4節で本論文の構成について述べる。

1.2 本研究の背景

AT&T Bell研の W.SchockleyとW.BrattainそしてJ.Bardeenが1948年に点接触型トランジスターを発明して以来、半導体集積回路技術は飛躍的発展を遂げ、今なお進展を続けている。1959年には、J.Hoerniが、シリコン結晶表面を酸化し、その一部を取り除いて露出したシリコンへ不純物を拡散させる、いわゆるシリコンプレーナトランジスタ形成技術を開発した。この技術をもとに1961年、R.Noyce はシリコンプレーナ集積回路の製造に成功した。この技術により電子回路の信頼性と生産性は飛躍的に向上することになる[11]。

1971年にインテルは、12mm^{2}のシリコン片に約2300個のトランジスターを集積した、演算処理単位が4bitのマイクロプロセッサー4004を開発した。その後インテルは、8bitのマイクロプロセッサー8008（1972年）、8080（1974年）、8085（1977年）、16bitのマイクロプロセッサー8086（1979年）、80286（1982年）、32bitのマイクロプロセッサーi386（1985年）、i486（1989年）と着実に発展させ、現在では163mm^{2}のシリコン片に310万個のトランジスターを集積したマイクロプロセッサーPentiumを量産出荷するに至っている。これらのマイクロプロセッサーは、様々な家電製品や多くのパーソナルコンピュータの中央演算処理装置として用いられている。

一方、半導体メモリーは、1970年にインテルのMOSトランジスタによる1KbitDRAM

1103 [12] が市場に登場して以来、ほぼ3年に4倍の容量増加を達成し、現在では16MbitDRAMが量産出荷、64MbitDRAMがサンプル出荷されている。

シリコン半導体集積回路の製造技術は、この約20年間に、約1万倍の高集積化と、ほぼ千倍の高速化を実現させた。この飛躍的な集積回路技術の進展によって、大規模で冗長な回路構成が必要なニューラルネットワークでさえ、1チップに数百ニューロン、数万シナプスを集積できるようになった[8][9][10]。

半導体集積回路によるニューラルネットワークのLSI化に関する研究は、1980年代半ばから米国を中心に、数十ニューロン規模の基本機能検証用テストチップの試作から始まった。今までに試作発表された主なニューラルネットワークLSI（ニューロチップ）を表1.1にまとめる。

表1.1 発表された主なニューロチップ

番号	発表者	研究機関	ﾆｭｰﾛﾝ数	ｼﾅﾌﾟｽ数	ﾒﾓﾘﾌﾟﾚｲ数(精度)	使用プロセス	発表年	参考文献
1	Thakoor, A. et al.	ｼﾞｪｯﾄ推進研(JPL)	0	1024	nMOSｼﾞｽﾀ可変（1bit）	3μCMOS	1985	[13]
2	Howard, R. et al.	A T & T ベル研	22	484	ｱﾅﾛｸﾞVSI固定結合	2μCMOS	1986	[14]
3	Sage, J. et al.	M I T ﾘｿｰｽ研	13	169	壁化膜中の電荷可変（10階調）	CCD-MNOS	1986	[15]
4	Sivilotti, M. et al.	ｶﾘﾌｫﾙﾆｱ工科大	22	484	ｱﾅﾛｸﾞｺﾝﾊﾟｱ（3値）Hebb則学習	4μNMOS	1986	[16]
5	Graf, H. et al.	A T & T ベル研	256	65536	固定結合（数MΩ）	2.5μCMOS	1986	[17]
6	Thakoor, A. et al.	ｼﾞｪｯﾄ推進研(JPL)	0	1600	ｱﾓﾙﾌｧｽSiH WO3酸化膜元薄膜	1μα-SiH薄膜	1987	[18]
7	Alspector, J. et al.	ベルコア	6	15	ﾚｼﾞｽﾀ（5bit）BMFｵﾝﾁｯﾌﾟ学習	2μCMOS	1987	[19]
8	Graf, H. et al.	A T & T ベル研	54	2916	2bitSRAMﾉﾝｺﾞｱｽﾀ（3値）	2.5μCMOS	1988	[20]
9	Murray, A. et al.	ｴｼﾞﾝﾊﾞﾗ大	8	64	ｼﾌﾄﾚｼﾞｽﾀﾊﾞﾗﾝｽ変調方式（5bit）	3μCMOS	1988	[21]
10	秋山 他	慶応大・ﾎﾝﾀﾞ研究ｾﾝﾀｰ大	1	1	ｽｲｯﾁﾄﾝｼﾞｽﾀ(8bit)	3μCMOS	1988	[22]
11	Mead, C.	ｶﾘﾌｫﾙﾆｱ工科大	2304	6912	人工網膜 固定結合、MOSｽｲｯﾁ	CMOS	1988	[23]
12	Boahen, K. et al.	ｼｮﾝﾎﾟｷﾝｸﾞ大	32	448	BAM、ﾃﾞｼﾞﾀﾙﾚｼﾞｽﾀ（2値）	CMOS	1989	[25]
13	Holler, M. et al.	インテル	64	10240	EEPROM ｽﾄｰﾁﾝｸﾞｹﾞｰﾄ（6bit）	1μEEPROM	1989	[26]
14	土屋 他	富士通	1	0	外部RAM（16bit）	2.3μBiCMOS	1989	[27]
15	平井 他	筑波大・日立	6	84	ﾊﾟﾙｽ変調完全ﾃﾞｼﾞﾀﾙ方式（6bit）	1.2μCMOS G/A	1989	[28]
16	Mueller, P. et al.	ﾍﾟﾝｼﾙﾍﾞﾆｱ大	5	0	アナログ回路	3μCMOS	1989	[29]
17	Mueller, P. et al.	ﾍﾟﾝｼﾙﾍﾞﾆｱ大	0	512	ﾄﾞﾘﾌﾄﾚｼﾞｽﾀ(8bit)アナログ演算	3μCMOS	1989	[29]
18	Graf, H. et al.	A T & T ベル研	256	32768	SRAM（1bit）～8192（4bit）可変	0.9μCMOS	1990	[30]
19	森下 他	松下	64	768	ﾅﾅﾍﾞﾝﾀ（8bit）外部RAM→DACﾃﾞｼﾞﾀﾙﾌｨｰﾄﾞﾊﾞｯｸ	2μBiCMOS	1990	[31]
20	柴田 他	日立	288	18432	SRAM(16bit) BP学習 ｺﾝﾋﾟｭｰﾀｼﾐｭﾚｰﾀ	0.8μCMOS G/A	1990	[32]
21	有馬 他	三菱	125	10000	ﾅﾅﾍﾞﾝﾀ（4bit） ｵﾝﾁｯﾌﾟ学習	1μCMOS	1990	[8]
22	江口 他	リコー	1	64	ﾊﾟﾙｽ変調完全ﾃﾞｼﾞﾀﾙ方式（7bit）	1.5μCMOS G/A	1990	[33]
23	Chiang, A. et al.	M I T ﾘﾝｶｰﾝ研	14	2016	ｼﾌﾄﾚｼﾞｽﾀ(6bit) MDAC	3μCCD	1990	[34]
24	Griffin, M. et al.	A S I	64	262144	SRAM（8bit）2.7ｺｱﾝﾁｯﾌﾟﾄﾞﾗﾞﾃﾞｼﾞﾀﾙ方式	0.8μCMOS	1991	[35]
25	Boser, B et al.	A T & T ベル研	8	4096	ﾅﾅﾍﾞﾝﾀ（6bit）外部RAM→DACﾃﾞｼﾞﾀﾙﾌｨｰﾄﾞﾊﾞｯｸ	0.9μCMOS	1991	[36]
26	有馬 他	三菱	336	28224	ﾅﾅﾍﾞﾝﾀ（5bit）ｵﾝﾁｯﾌﾟ学習 ﾏﾙﾁﾁｯﾌﾟ拡張	1μCMOS	1991	[9]
27	有馬 他	三菱	400	40000	ﾅﾅﾍﾞﾝﾀ（5bit）ｵﾝﾁｯﾌﾟ学習 ﾏﾙﾁﾁｯﾌﾟﾌﾟﾛｾｯｻｰ	0.8μCMOS	1992	[10]
28	内村 他	N T T	13	832	SRAM(8bit) 完全ﾃﾞｼﾞﾀﾙ方式	0.8μCMOS	1992	[37]
29	島 他	東芝	24	0	ｱﾅﾛｸﾞ回路方式	0.8μCMOS	1992	[38]
30	島 他	東芝	0	576	SRAM(8bit)	0.8μCMOS	1992	[38]
31	中平 他	松下	64	192	SRAM(3bit) 完全ﾃﾞｼﾞﾀﾙ方式	1.2μCMOS G/A	1993	[39]
32	Park, C. et al.	インテル・Nestor	1024	262144	EPROM(5bit)完全ﾃﾞｼﾞﾀﾙ方式	0.8μCMOS	1993	[40]
33	近藤 他	三菱	1000	12000	SRAM(2-4bitfloat)完全ﾃﾞｼﾞﾀﾙ	0.5μCMOS	1994	[41]
34	森江 他	N T T	9	81	ﾅﾅﾍﾞﾝﾀ(14bit) BP学習	1.3μCMOS	1994	[42]
35	相原 他	N T T	384	12288	SRAM(16bit)完全ﾃﾞｼﾞﾀﾙ方式	0.5μCMOS	1995	[43]

図1.2はニューロチップに集積されたニューロン数の年次推移を示している。これらのチップは様々なネットワークモデルを表現しているので単純に比較できないが、集積ニューロン数は、概ね、4年に10倍のペースで高集積化が進んでいる。また、集積シナプス数は図1.3に示すように、ほぼ3年に10倍の速さで集積化が推移してきた。この集積度の推移は、メモリーデバイスにおける3年に4倍の集積化推移と比べて極めて急速な進展に見えるが、これはニューロチップにおける回路構成が、未だ十分に最適化されていない段階にあると解釈する方が妥当である。

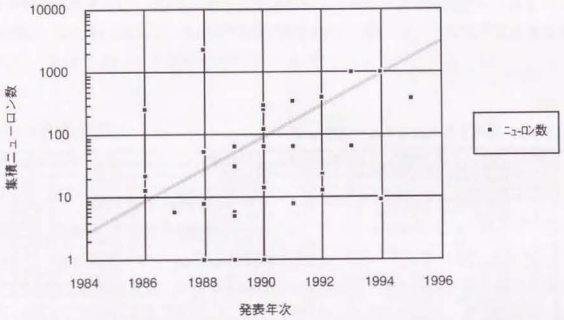


図1.2　試作された主なニューロチップの集積ニューロン数

今までに試作されたニューロチップは、その使用方法に制限を与えるシナプス荷重値の設定機能によって分類することができる。図1.3には、シナプス荷重値の設定方式によって、固定型（□）と可変型（△）そして学習型（○）の三種類にマークを分類して表示している。

ニューロチップの研究・開発が始まった当初は、固定された荷重値のシナプスを搭載したニューロチップが試作された。AT&T Bell研のR.Howard（1986年）[14]らは、アモルファスSiの抵抗値で2種類の荷重値を表現する484シナプス、22ニューロンのアナログニューロチップを、また、同Bell研のH.Graf（1986年）[17]らはアモルファスSi薄膜による3値

の荷重値を表現できる65536シナプス、256ニューロンのアナログニューロチップを試作した。これら試作されたチップによって、半導体集積回路で構成されたニューラルネットワークが高速に処理実行できることが確認された。

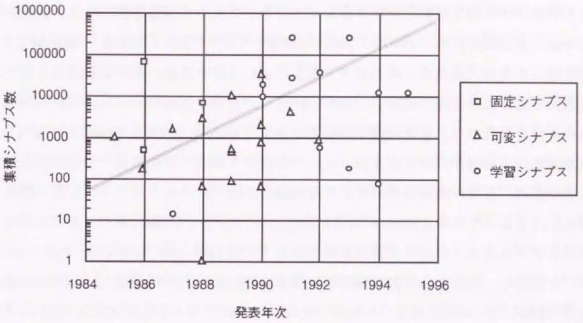


図1.3　試作された主なニューロチップの集積シナプス数

次に、ニューラルネットワークの処理内容をチップ製造後に変更できるようにする為に、シナプス荷重値を自由に設定できるようにした、いわゆる可変シナプス型のニューロチップが多く試作されるようになる。そしてシナプス荷重値を表現する方法として、高集積化に優れたキャパシターの蓄積電荷量でアナログ値を表現する方法が多く提案されている。MIT Lincoln Lab.のJ.Sage（1986年）[15]らはMNOSのCCDにより10階調程度の荷重値を設定できる169シナプス、13ニューロンのアナログニューロチップを試作した。また、松下電子の森下（1990年）[31]らはキャパシターの蓄積電荷量で荷重値を表現し、定期的に外部のSRAMに格納したデジタルデータのD/A変換値でリフレッシュする768シナプス、32ニューロンのアナログニューロチップを試作した。Bell研のBoser（1991年）[36]らも同様の4096シナプス、8ニューロンのアナログチップを試作している。これらキャパシターの蓄積電荷量でシナプス荷重値を表現する方式では、DRAMと同様に、電荷の微小な洩れによる蓄積電荷量の変動を復元するために定期的なリフレッシュ操作が必要である。

そこで、インテルのM.Holler（1989年）[26]らは長時間安定に荷重値を保持する為に、EEPROMと同様にフローティングゲートに電荷を蓄積して、6bit程度の荷重値を表現できる10240シナプス、64ニューロンのアナログニューロチップを開発し、ETANNと名付けて商品化した。また、リフレッシュ操作を必要としない、レジスタやSRAMで荷重値を保持し、アナログ回路で演算を実行するニューロチップも多く提案されている。Bell研のH.Graf（1988年）[20],（1990年）[30]らは2bitのメモリで3値（-1,0,1）の荷重値を表現する2916シナプス、54ニューロンのアナログニューロチップと、その発展版として、4bit精度まで可変にできる32768(2bit）～8192(4bit）シナプス、256ニューロンのアナログニューロチップを試作した。また、MIT Lincoln Lab.のA.Chiang（1990年）[34]らは6ビットのCCDで表現する2016シナプス、16ニューロンのアナログニューロチップを試作した。

その他に、全ての演算処理をデジタル回路で表現したニューロチップも多く提案されている。エジンバラ大のA.Murray（1988年）[21]らは符号付き4bitのシフトレジスタで表現するシナプスをもったパルス密度表現の64シナプス、8ニューロンのデジタルニューロチップを試作し、筑波大の平井（1989年）[28]らは完全デジタル回路によるパルス密度変調方式の84シナプス、6ニューロンのデジタルニューロチップを日立と共同で試作した。また、NTTの内村（1992年）[37]らは、デジタル回路方式の並列化に伴う消費電力増大を低減する機能を備えた、832シナプス、13ニューロンのバイナリデジタル回路方式のニューロチップを試作した。これら可変シナプス型のニューロチップはコンピュータなどで算出した荷重値をロードして使うために、いろいろなシナプス荷重値のニューラルネットワークを手軽に表現するのに適している。しかし、ロードする荷重値を算出するためには高速のコンピュータを用いても、長大な時間を費やさなければならず、頻繁に荷重値を修正することが必要な応用には適さない問題がある。

そこで最近では、シナプス荷重値を算出するのためのコンピュータを不要にして、ニューロチップ自身で高速に学習が実行できる、学習機能を実装した学習シナプス型ニューロチップの研究が盛んに行われるようになった。ニューロチップ開発の初期にも小規模ではあるが学習機能をチップに内蔵する試みが成されている。CaltechのM.Sivilotty（1986年）[16]らはHebbの学習則が実行できる、フリップフロップで構成された3値の484シナプス、22ニューロンのチップを試作し、ベルコアのJ.Alspector（1987年）[19]らはボルツマンマシンの近似学習則が実行できるレジスタで構成された5bitの15シナプス、6ニューロンのチップを試作している。1990年代に入り多くの研究機関で学習シナプス型のニューロチップ

が開発されてたが、多くの場合、デジタル回路により学習機能を実現しており、数万シナプス規模の高集積な学習機能実装ニューロチップは本研究の以前には実現されなかった。

今までに開発されたニューロチップは、演算処理の回路表現方式に関して、大まかに3方式に分類することができる。即ち、バイナリデジタル回路方式、パルス（密度）変調方式、そしてアナログ回路方式である。それぞれの方式ごとに一長一短があり一概に優劣を付けることは困難であるが、バイナリデジタル回路方式では、既存の回路資産を生かして多様な機能を高精度に比較的自由に造ることが可能である反面、機能表現に多くの素子が必要で回路面積が大きくなることと大規模な素子の並列動作に伴う消費電力増加の問題がある。また、パルス変調方式は回路の簡略化で面積の問題は緩和されるものの、情報の時間方向への拡張に伴う演算速度の低化と並列動作による消費電力の問題が残る。一方、アナログ回路方式はニューラルネットワーク特有の多入力積和演算や非線形変換回路を小数の素子で表現でき、高集積化とそれに伴う大規模並列処理によって極めて高速な処理が実現できるものの、演算精度に素子特性のバラツキが大きく影響することから、演算精度を高くすることが難しい問題がある。図1.4には、今までに試作された主なニューロチップとボードについての処理速度（CPS：Connections Per Second）と演算精度（階調bit）を示している。速度性能についてはアナログ回路方式ニューロチップが極めて高い性能を実現しているものの、デジタル回路方式ニューロチップは極めて高い演算精度を実現している。

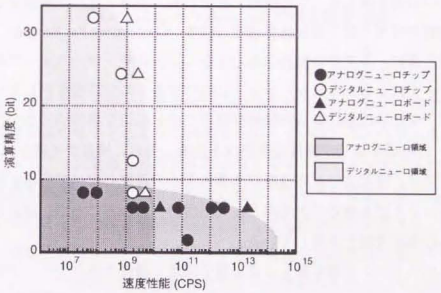


図1.4　演算精度と処理速度

このような回路表現方式ごとの特徴から、ニューロチップの有効な適用分野を回路方式ごとに次のように考えることができる。バイナリデジタル回路方式のニューロチップは、ネットワーク構成や機能モデルを自由にプログラムできる汎用性と、その学習課程で微分処理が必要なことから高い演算精度が強く要求されるEBP（Error Back-Propagation）学習モデル[44]などによる弁別処理への適用に於て、その優位性が発揮されると考えられる。また、パルス変調方式のニューロチップは、情報が一定期間に分散表現される特徴によって非同期処理が自然に表現でき、フィードバック信号を効率良く再現することが期待される。一方、アナログ回路方式によるアナログニューロチップは、その高い集積度とそれに伴う大規模な並列処理による高速性能によって、演算精度よりも大規模な回路網表現と高速演算処理が強く要求されるホップフィールドモデル[45]やボルツマンマシン[46]などのフィードバック結合型連想記憶ニューラルネットワークを実時間で表現するのに適している。

1.3 本研究の目的と意義

ニューラルネットワークに基づく情報処理を工学的に実現することを目的とした、ニューロチップやニューロボードの研究開発における現在の最も重要な課題は、表現できるネットワークサイズの大規模・高集積化と演算処理の高速化である。図1.5は、ニューラルネットワークの性能を表わす二つの指標、すなわち、ネットワーク規模を示すシナプス結合数と処理時間の目安となる結合演算速度（CPS：Connections Per Second）に関して、ヒトの大脳の例と共に、今までに試作された主なニューロチップ、ニューロボードそして現行のコンピュータによるシミュレーションについて各々の定位置を示している。

生体脳では、生化学反応による神経細胞膜のイオン転送変調作用に基づき、数100mV、数ms オーダーの膜電位信号が神経回路網内を伝播し並列に情報が処理されている[47]。高等生物の脳では、膜電位の反応速度が半導体素子の速度と比べて桁違いに遅いにも関わらず、その違いを上回る極めて大規模な並列処理が行われることで、現在のコンピュータを遥かに超える演算速度を実現している。

一方、現在のシリコン半導体素子による電子回路で構成されたニューラルネットワークの情報処理は、数V、数ns オーダーで信号が処理されるので、生体脳と比べて消費エネルギーは数桁大きいものの、信号処理速度は約6桁程度速い特長を有している。しかし、

PC（Personal Computer）やEWS（Engineering Work Station）などマイクロプロセッサーとメモリーデバイスで構成された現行のコンピュータでは、演算処理の並列度が1から数十程度と少なく、スーパーコンピュータですら大脳の一万分の一程度の速度性能でしかない。

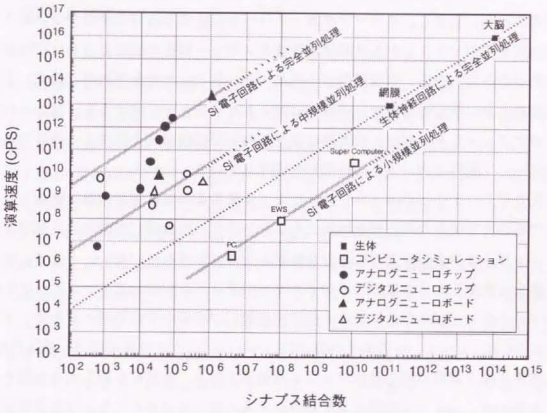


図1.5 ニューラルネット規模と処理速度

そこで、ニューラルネットワーク演算専用に開発されたニューロチップでは、一般に、演算処理の並列度を高めることによって高速化を図っている。ニューラルネットワークの機能表現には、シナプス荷重値を保持するためのメモリー機能とシナプスおよびニューロンの演算処理機能とが密接な関係のもと多数必要で、メモリー機能と演算処理機能との回路面積のバランスによって半導体集積回路における最適な並列度が決まる。つまりデジタル回路の場合、演算処理機能の回路面積はメモリー機能の回路面積の数百から数千倍以上と極めて大きく、一つの演算回路が多数のメモリーに格納されたシナプス荷重値を逐次処理する方が、演算処理速度とシナプスの集積規模を共に向上するために有効であることか

ら、数十から数百程度の中規模な並列処理構成のデジタルニューロチップが多く開発される結果となった。デジタル回路は、その極めて高い再現特性によって逐次的な繰り返し処理を行っても演算精度が劣化しない特徴によって設計の自由度は極めて高く、処理の並列度は消費電力とチップ面積にのみ制限される。

一方、アナログ回路の場合は、処理段数が増すほど演算精度が低下するので逐次処理には不向きである。しかし、ニューラルネットワークに必要な演算機能は少ない素子数でアナログ回路表現ができ、しかも演算処理機能とメモリー機能の回路面積がほぼ同等の大きさで実現できることから、演算回路とメモリーを一体とした完全並列処理構成にすることで高集積化と共に演算処理速度を飛躍的に向上させることが可能である。試作された多くのアナログニューロチップは、大規模な完全並列処理回路構成を採用し、現行のコンピュータと比べて7桁以上高いコスト性能比を実現している。

しかしながら、アナログニューロチップでさえ現在の半導体集積回路の微細化レベルでは、その極めて単純化された機能表現にもかかわらず、生体脳に匹敵する大規模なニューラルネットワークを表現するのに十分な集積レベルに到達していない。そこで、集積回路素子の更なる微細化と、多数のアナログニューロチップを接続してネットワーク規模を拡張するマルチチップ拡張技術による大規模高集積化が重要な研究課題となるが、アナログ回路方式のLSIにとって、回路素子の微細化とマルチチップの拡張接続は、素子特性バラツキの増大に伴う演算精度と動作マージンの低下を招き、実用化を阻む大きな障害となる。

そこで我々は、ニューロ連想メモリーデバイスの実用化に不可欠な大規模連想記憶アナログニューラルネットワークLSIの実現を目的として、このアナログ集積回路の問題点を克服すべく学習機能をチップ上に搭載する検討を行い、実際に学習機能搭載ニューロチップを試作・評価して基本性能を実証した。すなわち本研究は、ニューラルネットワークLSIに関する研究において次の位置付けと意義を有している。

○学習機能のオンチップ実装によりアナログ集積回路の大規模高集積化に伴う問題点を克服できる可能性を明らかにした。

○連想メモリーを構成するアナログニューラルネットワークLSIを試作し、その連想性能を評価した。

○ニューロ連想メモリーを構成するニューラルネットワークLSIにおいて学習機能を実

装しながら最高レベルの集積規模および処理速度を実現した。

○シリコンMOS集積回路によるニューラルネットワークLSIの高性能化が、素子の微細化に伴い進展できる限界を予測した。

1.4 本論文の構成

本論文の構成と研究の流れを図1.6に示す。本論文中ではニューラルネットワークのダイナミックスに基づく連想メモリーを、他の方式と区別するためにニューロ連想メモリーと称する。本研究論文では、まず第2章と第3章において連想記憶アナログニューラルネットワークLSIの高集積化に関して克服すべき課題と解決手段について述べ、第4章から第6章の3つの章で実際に試作した3種類の連想メモリーを構成するニューラルネットワークLSIについてその概要と評価結果について各々述べている。

本章に続く第2章では、連想メモリーデバイスを実現する各種回路方式の性能比較を行い、その高集積化・高速化にはアナログ回路によるニューラルネットワーク機能表現が優れていることを述べる。そして、アナログ集積回路の問題点として素子の微細化に伴ない増大する素子特性バラツキが連想性能へ与える影響を明らかにした上で、ニューラルネットワークの学習による自己組織化機能が素子特性バラツキを補償し、この問題を克服する手段として有効であることを示す。また、学習機能を実装することでアナログニューラルネットワークLSIが素子の微細化に伴い高性能化できる限界を予測し、その可能性を示す。

第3章では、学習機能をチップ上に高集積に実装するための回路構成とその制御方式について述べる。まず、考案した学習機能を備えたニューロン回路とシナプス回路の構成について述べた後、学習機能の高集積化のために採用した簡略化学習ルールと非線形なシナプス荷重値修正特性が学習性能へ与える影響について評価し、これらの機能的制限に関しても補償機能が働くことで十分な学習性能が実現できることを明らかにする。また、シナプス荷重値表現で採用した非線形な飽和領域で動作する回路構成が、電源電圧の変動に対して高い動作マージンを確保できることを、回路シミュレーションにより示す。

第4章では、第3章で述べた学習回路構成を採用して実際に試作した、3種類の学習機能搭載ニューロチップについて、その概要と評価した学習および連想性能について述べ、1チップに数百ニューロン、数万シナプスを集積して、それらの完全並列処理により、

10^{12}CPS以上の高速演算を実現できたことを示す。

第5章では、大規模化を実現するために開発した、複数のニューロチップを接続して回路規模を拡張できる、ニューロン機能分散表現方式とその回路構成について説明し、その回路構成を搭載したマルチチップ拡張機能搭載ニューロチップについて、その概要と拡張性能の評価結果について述べる。また、そのチップを最大18個まで搭載して拡張接続できるボードの試作概要と、そのボードで実際に評価した1000ニューロン、100万シナプスの大規模ニューラルネットワークの学習能力について述べる。拡張接続評価によって、数百チップまでの拡張接続が可能であることが見積もられることを示す。

第6章では、高集積化のために採用したシナプス荷重値のダイナミックストレージ方式の問題点である、キャパシターの蓄積電荷リークによる荷重値の変動を高速に修復するために開発したマクロリフレッシュ方式について説明し、その回路構成を採用して試作した、高速リフレッシュ機能搭載ニューロチップの概要と、そのリフレッシュ機能の評価結果について述べる。

第7章で、本研究で得られた一連の結果について総括する。

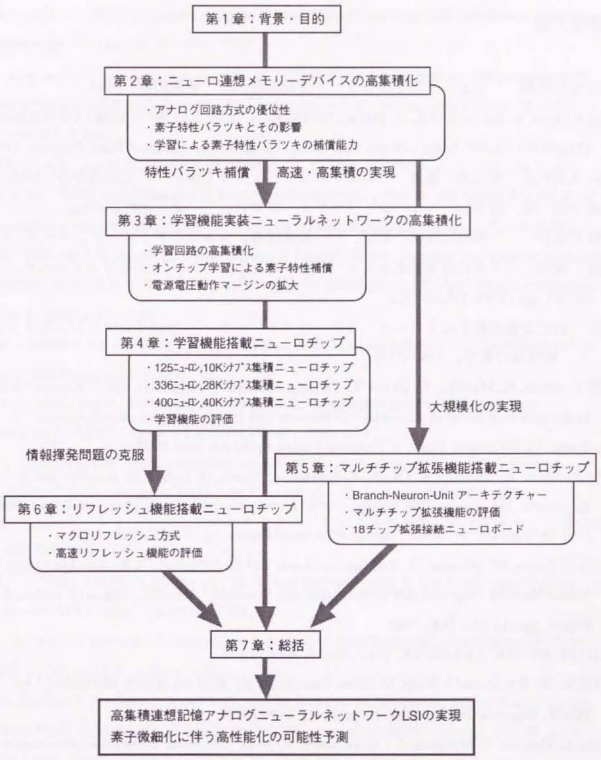


図1.6 研究論文の構成と流れ

参考文献

[1] 麻生英樹、”ニューラルネットワーク情報処理、” 産業図書、1988。

[2] J. Hertz, A. Krogh, and R. G. Palmer, "INTRODUCTION TO THE THEORY OF NEURAL COMPUTATION," Addison-Wesly Publishing Company, The Advanced Book Program, 1991.

[3] 久間和生、中山高 編著、”ニューロコンピュータ工学、” 工業調査会、1992。

[4] 合原一幸 編著、”ニューロ・ファジイ・カオス、” オーム社、1993。

[5] 甘利俊一、”神経回路網の数理、” 産業図書、システムサイエンスシリーズ、1978。

[6] ”実用システムに適用始まるニューラル・ネットワーク、” 日経コンピュータ、no.247, pp.76-94, 1991年2月.

[7] ”新たな展開期を迎えるニューラル・ネットワーク、” 日経インテリジェントシステム、別冊1992夏号、1992年7月.

[8] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," in Symp. VLSI Circuits, Digest of Technical Papers, pp.63-64, June 1990.

[9] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," in ISSCC, Digest of Technical Papers, pp.182-183, Feb. 1991.

[10] Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," in ISSCC, Digest of Technical Papers, pp.132-133, Feb. 1992.

[11] SCIENTIFIC AMERICAN, Vol.7, No.11, Nov., 1977.

[12] W. W. Regitz, and J. Karp, "A Three-Transistor-Cell, 1024 bit, 500NS MOSRAM," in ISSCC, Digest of Technical Papers, pp.42-43, Feb. 1970.

[13] A. Thakoor, A. Moopenn, J. Lambe, and S. Khanna, "Electronic hardware implementation of neural networks," Applied Optics, Vol.26, No.23, pp.5085-5092, Dec., 1987.

[14] R. Howard, D. Schwartz, J. Denker, R. Epworth, H. Graf, W. Hubbard, L. Jackel, B. Straughn, and D. Tennant, "An Associative Memory Based on an Electronic Neural network Architecture," IEEE Trans, Electron Devices, Vol.34, pp.1553-1556, 1987.

[15] J. P. Sage, K. Yhompson, and R. S. Withers, "An Artificial Neural Network Integrated Circuit Based on MNOS/CCD principles," Neural Networks for Computing, AIP Conference Proceedings, No.151. pp.381-385, 1986.

[16] M. A. Sivilotti, M. R. Emerling, and C. A. Mead, "VLSI Architecture for Implementation of Neural Networks," Neural Networks for Computing, AIP Conference Proceedings, No.151. pp.408-413, 1986.

[17] H. Graf, L. Jackel, R. Howard, B. Straughn, J. Denker, W. Hubbard, D. Tennant, and D. Schwartz, "VLSI Implementation of a neural network memory with several hundred of neurons," Neural Networks for Computing, AIP Conference Proceedings, No.151. pp.414-419, 1986.

[18] A. Thakoor, J. L. Lamb, A. Moopenn, and J. Lumbe, "Binary Synaptic Connections Based on memory Switching in α-Si:H," Neural Networks for Computing, AIP Conference Proceedings, No.151. pp.151-158, 1986.

[19] J. Alspector, and R. Allen, "A Neuromorphic VLSI Learning System," Advanced Research in VLSI, MIT Press, pp.313-349, 1987.

[20] H. P. Graf, and P. DeVenguar, "A CMOS Associative Memory Chip Based on Neural Networks," ISSCC, Digest of Technical Papers, Feb., 1987.

[21] A. Murray, and A. Amith, "Asynchronous VLSI Neural Networka Using Pulse-Stream Arithmetic," IEEE, Journal of Solid-State Circuits, Vol.23, No.3, pp.688-697, 1988.

[22] 秋山、”ガウシャンマシンとそのアナログ／デジタル専用アーキテクチャ、” 電子情報通信学会技報、CPSY88-16、1988.

[23] C. A. Mead, and M. A. Mahowald, "A Sillicon Model of Early Visual Processing," Neural Networks, Vol.1, No.1, pp.91-97, 1988.

[24] J. Mann, R. Lippmann, B. Berger, and J. Raffel, "A Self-Organizing Neural Network Chip," IEEE, CICC, pp.10.3.1-10.3.5, 1988.

[25] K. Boahen, P.Pouliquen, A. Andreou, and R. Jenkins, "A Heteroassociative Memory Using Current-Mode MOS Analog VLSI Circuits," IEEE Trans, Circuits and Systems, Vol.36, No.5, pp.747-755, 1989.

[26] M. Holler, S. Tam, H. Castro, and R. Benson," An electricaly trai-nable artificial neural network (ETANN) with 10240 Floating gate synapses," Proc. of IJCNN-89, Vol.2, pp.191-196, 1989.

[27] 土屋、杉浦、岩本、吉沢、加藤、浅川、"世界で初めて商品化されるニューロチップ、" 日経マイクロデバイス、no.45, pp.123-129, 1989年.

[28] Y. Hirai, K. Kamada, M. Yamada, and M. Ooyama, "A Digital Neuro-Chip with Unlimited Connectivity for Large Scale Neural Networka," IJCNN, Digest of Technical Papers, Vol.2, pp.163-169, 1989.

[29] P. Mueller, "A General Purpose Analog Neural Computer," IJCNN, Digest of Technical Papers, Vol.2, 1989.

[30] H. P. Graf, and D. Henderson, "A Reconfigurable CMOS Neural Network", ISSCC, Digest of Technical Papers, pp.144-145, Feb. 1990.

[31] T. Morishita, Y. Tamura, and T. Otsuki, "A BiCMOS analog neural network with dynamically updated weights," ISSCC, Digest of Tecnical papers, pp.142-143, Feb., 1990.

[32] 柴田、安永、大山、益田、柳生、浅井、山田、坂口、橋本、" 高速学習型ニューロWＳＩのシステム設計、" 電子情報通信学会技報、CPSY90-71、pp.49-56、1990.

[33] 江口、古田、堀口、樗木、" 学習機能を有するパルス密度型ニューロンモデルとそのハード化、" 電子情報通信学会技報、CPSY90-73、pp.63-70、1990.

[34] A. Chiang, R. Mountain, J. Reinold, J. LaFranchise, J. Gregory, and G. Lincoln, "A Programmable CCD Signal Processor," ISSCC, Digest of Tecnical papers, pp.146-147, Feb., 1990.

[35] M.Griffin, G. Tahara, K. Knorpp, R. Pinkham, and B. Riley, "An 11-million transistor neural network execution engine," ISSCC, Digest of Tecnical papers, pp.180-181, Feb., 1991.

[36] B. Boser, and E. Sackinger, "An Analog Neural Net-work Processor with Programmable Network Topology," ISSCC, Digest of Technical Papers, pp.184-185, Feb., 1991.

[37] K. Uchimura, O. Saito, and Y. Amemiya, "An 8G Connections-per-Second 54mW Digital Neural Network Chip with Low-power Chain-Reaction Architecture," ISSCC, Digest of Tecnical papers, pp.134 -135, Feb., 1992.

[38] T. Shima, T. Kimura, Y. Kamatani, T. Itakura, Y. Fujita, and T. Iida, "Neuro Chips with On-Chip BackProp and/or Hebbian Learning," ISSCC, Digest of Tecnical papers, pp.138-139, Feb., 1992.

[39] H. Nakahira, S. Sakiyama, M. Maruyama, K. Hasegawa, T. Kousa, S. Maruno, Y. Shimeki, T. Satonaka, and Y. Nagano, "A Digital Neuroprocessor Using Quantizer Neurons," in Symp.

VLSI Circuits, Digest of Technical Papers, pp. 35-36, 1993.

[40] C. Park, K. Buckmann, J. Diamond, U. Santoni, S. C. The, M. Holler, M. Glier, C. L. Scofield, L. Nunez, and J. Cole, "A 40,000 Pattern/sec Recognition Accelerator with learning Capability," Hot Chips V, pp.7.3.1-7.3.10, Aug., 1993.

[41] Y. Kondo, Y. Koshiba, Y. Arima, M. Murasaki, T. Yamada, H. Amishiro, H. Shinohara, and H. Mori,"A 1.2GFLOPS Neural Network Chi p Exhibiting Fast Convergence," in ISSCC, Digest of Technical Papers, pp.218 -219, Feb. 1994.

[42] T. Morie, and Y. Amemiya, "An All-Analog Expandable Neural Network LSI On-Chip Backpropagation Learning," IEEE, Journal of Solid-State Circuits, Vol.29, No.9, pp.1086-1093, Sep., 1994.

[43] K. Aihara, O. Fujita, K. Uchimura, "A Sparse Memory-Access Neural Network Engine with 96 Parallel Data-Driven Processing Units," in ISSCC, Digest of Technical Papers, pp.72-73, Feb. 1995.

[44] D.E. Rumelhart, and J. L. McClelland, eds., "Learning inter-nal representation by error propergation,"Parallel distributed processing, MIT press, pp.318-362, 1986.

[45] J.J. Hopfield,"Neural networks and Physical Systems with Emergent Collective Computational Abilities," Proc. Nat. Acad. Sci. USA, 79, pp.2445-2558,1982.

[46] D H.Ackley, G E.Hinton, and T J.Sejnowski, "A Learning Algorithm for Boltzmann Machines," Cogunitive Science, Vol.9, No.1, pp.147-169, Jan-Mar, 1985.

[47] 伊藤薫、" 脳と神経の生物学、" 培風館、1982。

第2章
ニューロ連想メモリーデバイスの高集積化

2.1 序

生体脳の優れた情報処理様式である直観的知識情報処理を工学的に再現するためには、知識を獲得するための学習（記憶）処理と蓄積された知識情報を自由に取り出すための連想処理を効率良く実行できる連想メモリーデバイスが必要不可欠である。連想や学習（記憶）は、現行のコンピュータにおけるメモリーからのデータ読み出しや書き込みに相当して、直観的知識情報処理の基本となる機能であり、それを担う連想メモリーの高性能化は直観的知識情報処理装置の実用化において極めて重要な課題の一つである。

本章では、ニューラルネットワークLSIで実現される連想メモリーデバイスが柔軟性とコスト性能比で極めて優れていることを示すと共に、アナログ回路方式が連想メモリーを構成するニューラルネットワークLSIの実現に関して高集積・高速性に優れていることを示す。そして、アナログ回路の大規模高集積化において問題となる、素子の微細化に伴って増大する素子特性バラツキと、それによる連想性能の劣化について明らかにし、その問題を克服するために着目した学習機能のオンチップ実装による素子特性バラツキの自動補償機能について述べる。

まず、第2.2 節では、連想メモリー機能を実現する手段として、現行のコンピュータによるアルゴリズム表現方法とCAM（Content Addressable Memory）デバイス、そしてニューロ連想メモリーデバイスの各々の特徴を比較し、ニューラルネットワークLSIによるニューロ連想メモリーがその柔軟性とコスト性能比において極めて優れていることを示す。そして、計算機シミュレーションによって得られたニューロ連想メモリーデバイスに要求されるシナプス精度について述べる。

次に、第2.3 節では、連想記憶ニューラルネットワークLSIの実現には、アナログ回路がその高集積化に最も優れていることを示すと供に、その高精度化を実現するために克服すべき問題点を明らかにする。素子の微細化が進むに従って、アナログ集積回路チップ内の素子特性バラツキが増大し、それが連想性能へ及ぼす影響について計算機シミュレーションにより評価した結果について述べ、この問題の重大性を示す。

そして続く第2.4節において、アナログニューラルネットワークLSIの高精度・高集積化を阻むこの問題点を克服するために着目した、チップ上に学習機能を搭載することで期待される、ニューラルネットワークの自己組織化機能による素子特性バラツキや非線形特性変動等の不良因子の補償能力ついて、計算機シミュレーションにより評価した結果について述べ、十分な補償能力が期待できることを明らかにする。

2.2 連想メモリーデバイス

2.2.1 連想メモリーの実現方式

従来のコンピュータに用いられているメモリーは、データ格納アドレスを指定して記憶データを読み出しあるいは書き込む機能を担っている。従って、ユーザーは予めメモリーに格納されたデータとそのアドレスの関係を何らかの方法で管理する必要がある。それに対して連想メモリーは、入力されたデータの信号パターンに最も近い記憶データを出力する（自己想起）かあるいは、その入力データから連想されるべき、予め記憶した期待データを出力する（相互想起）、いわゆるパターン連想機能を備えている。連想メモリーはデータの内容による記憶データの読み出しができることによって、ユーザがメモリー内のデータ格納場所に関する情報を一切取り扱う必要がなく、それを管理するための手続が不要になることで情報処理手続き設計の基本的考え方をメモリー駆動型からデータ駆動型へと、繁雑なアロケーション等の環境設定を必要としない、あるいは処理全体を見渡し設定する必要がない、より直接的な処理表現を自由に組み合わせることが可能になる[1]。

従来から高機能コンピュータのキャッシュメモリー等に使われているCAM（Content Addressable Memory）は、格納したデータと入力データを比較する回路を内蔵したことによって、 入力データと記憶データのキーパターンとを比較して一致した内容の記憶データもしくはそのアドレスを出力する、一種の連想機能を実現しているデバイスである。CAMには、入力データの一部をマスクして一致比較する機能を有しているので、入力パターンにマスクするビットの数が入力パターンに許される曖昧さに対応し、マスクビットの位置によって自己想起や相互想起の連想を実現することが可能である。CAMでは、入力データと記憶データとの比較回路と一致検出回路を多数備えて処理を並列化することに

よって高速連想処理を実現している[2][3][4]。

そこで、連想メモリー機能を実現する手段として次の三種類のデバイスによる連想性能について比較を行った。すなわち、現行のコンピュータによるアルゴリズム表現方法を実現するストアドプログラム駆動のマイクロプロセッサー[i486]とCAMデバイス[4]、そしてニューラルネットワークLSI（ニューロチップ）[5]によるニューロ連想メモリーデバイスに関して比較した。

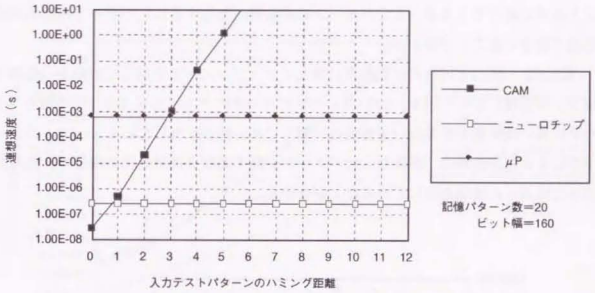


図2.1　連想処理速度比較

図2.1 は、記憶データおよび入力データの幅が160bitで、記憶パターン数が20の場合の連想速度に関して、マイクロプロセッサー（μP）とCAMそしてニューロチップを用いて実行した各々の場合について示している。マイクロプロセッサーによる場合は、入力データと全ての記憶パターンとのハミング距離を算出して最も近い記憶パターンを選択するプログラムを実行させることを想定した。このアルゴリズムによれば記憶データの序列化などの事前処理が不要となり、記憶データの柔軟な設定が可能となる。また、CAMによる場合は、入力データーの曖昧度がマスクするビット位置に依らないようにするために、全ての記憶データに関してビットローテーションした159×20のパターンを記憶データに追加して同時に比較する方法によって、入力データに対するハミング距離毎に並列処理でき

る構成を想定している。この方式によってCAMは、その容量を犠牲にするもののマスクビット位置に依らない柔軟な連想処理を高速化することができる。ニューロチップによる場合は、相互結合型のニューラルネットワークのダイナミックスによる連想処理を想定している。各々の性能を算出するために参考にした実際に試作された各デバイスの製造技術レベルを合わせるために、ニューロチップは0.8μmCMOSに換算した。

マイクロプロセッサーによるプログラム処理に比べニューロチップによる連想処理は、3桁以上高速であり入力パターンの曖昧度には影響しない。一方、CAMによる連想処理は入力パターンの曖昧度が極めて少ない、ハミング距離が1以下の場合、ニューロチップより高速に実行できるが、入力パターンの曖昧度が増加するにしたがって連想処理速度が急速に低下することが分かる。

図2.2は、図2.1と同様の連想処理に関してチップコストを考慮して縦軸を（処理速度×チップ面積）として示す。CAMは、コンピュータのキャッシュメモリーの様な入力データに高い曖昧度を許さない連想処理に関して高い性能を有している。一方、ニューロチップによる連想処理は、曖昧な入力パターンを許す柔軟な連想処理に関して、他の方式に比べて桁違いの性能を有していることが分かる。

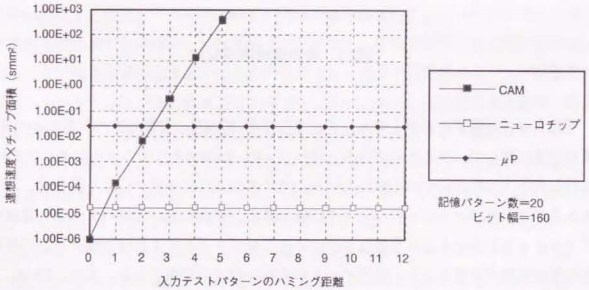


図2.2　コスト性能比較（その1）

図2.3は、マイクロプロセッサーとニューロチップによる連想処理の記憶パターン数に

対するコスト性能比を示している。記憶データ幅は160bitとして、記憶容量のみ変化させた場合を見積もっている。ニューロチップによる場合のコスト性能比は、記憶容量がニューラルネットワークの回路規模に比例するものの処理速度はその並列性によりほぼ一定であるので、記憶容量に比例したコスト性能を実現できる。一方、マイクロプロセッサーによる場合には、処理時間が記憶容量に比例して増加し、加えて記憶データを保持する為のメモリーも容量に比例して増加することから、コスト性能比は記憶容量の二乗に比例して低下する。

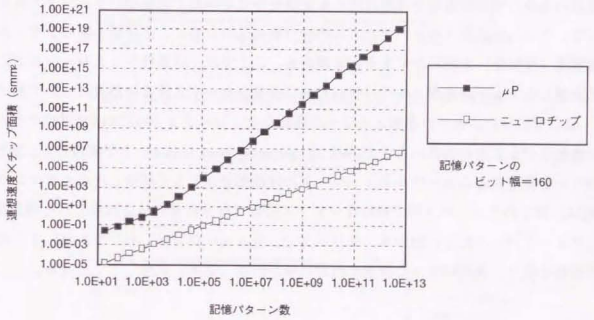


図2.3　コスト性能比較（その2）

ヒトの大脳規模の10^{11}ニューロンのニューラルネットワークに対応する記憶容量10^{10}～10^{11}の場合では、ニューロチップによるコスト性能がマクロプロセッサーによる場合より10桁程度優れていることが見積もられた。このことから、ニューラルネットワークLSIにおけるニューラルネットワークダイナミックスに基づく連想機能の実現が、直観的知識処理のキーデバイスとなる連想メモリーとして、その柔軟性とコスト性能比において極めて優れていると言える。

2.2.2 ニューロ連想メモリーデバイスに要求されるシナプス精度

ニューロ連想メモリーデバイスの機能性能は、集積されたニューロン数およびシナプス数によって記憶容量が決まり、それら機能回路の演算処理速度によって連想速度が決まる。また、シナプス荷重値の精度（分解能）はニューラルネットワーク規模と共に学習（記憶）に関する性能を左右する。シナプス荷重値の表現精度によって回路方式や回路面積が変わるので、シナプス精度はニューロ連想メモリーデバイスの設計において、学習（記憶）性能のみならず記憶容量や連想速度へも影響を与える極めて重要なパラメータである。従って、ニューロ連想メモリーデバイスの設計最適化のために、必要最小限のシナプス荷重値精度（分解能）を明らかにする必要がある。ここでは、計算機シミュレーションによって評価したニューロ連想メモリーデバイスに必要なシナプス荷重値精度について述べる。

ここで行ったニューロ連想メモリーの機能シミュレーションはC言語でプログラムした機能モデルを次の条件のもとでEWS（Engineering Work Station）上で実施した。連想メモリーを実現するニューラルネットワークの機能モデルとして採用したボルツマンマシン[6]は、全てのニューロン間で対称なシナプス結合（自分自身の結合は無し）の相互接続（フィードバック結合）型のネットワークで、各ニューロンは”０”または”１”の二つの状態を取り、次式に従った確率で出力状態が活性（発火）状態”１”となる。

$$P(Si=1)=1/(1+\exp(-ui/T)) \qquad (2.1)$$

$$ui=\Sigma_j Wij \cdot Sj \qquad (2.2)$$

ここで、uiはニューロンiの入力総和値（内部活性値）、Tはシステムの”温度”と呼ばれるパラメータで、Tの値が大きいほどしきい値付近の活性化（発火）確率の変化が緩やかになる。図2.4にニューロンの活性化（発火）確率特性とシナプス精度（階調）の例を示す。各シナプスの荷重値は-64から+64の間の値をとり、その間を等分割して取りうる値の数を階調数としてその対数（底は２）値をシナプスのbit精度（分解能）と称している。式2.1中の”温度”パラメータTは大きくなるほどニューロンの活性化（発火）確率の非線形特性傾きが緩やかとなる。

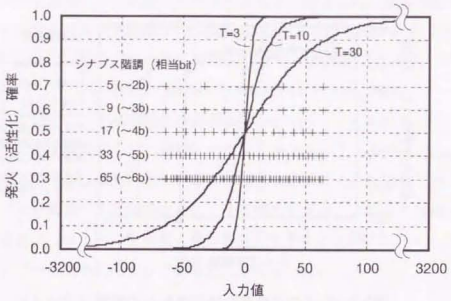


図2.4 ニューロンの非線形変換特性およびシナプス精度

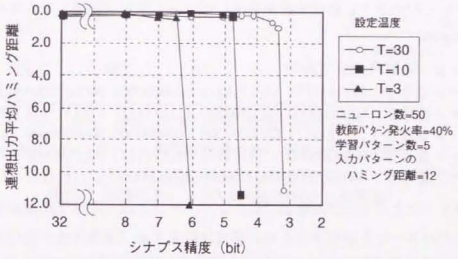


図2.5 シナプス精度の連想性能への影響（その１）

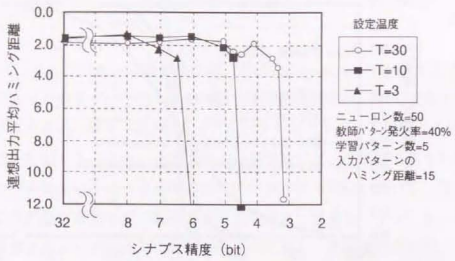


図2.6　シナプス精度の連想性能への影響（その２）

図2.5と図2.6はシナプス精度の連想性能へ与える影響を示している。ニューロン数が50のニューロ連想メモリー（隠れニューロン無し）に発火率が40%（50ニューロンの内20ニューロンが"1"）の５つのパターンを学習（記憶）した場合の連想性能がシナプスのbit精度によってどのように変化するかを示している。学習（記憶）した５パターンの各々は次に示すように、

ニューロン番号（12345678910.....20.....30.....40.....50）
パターン番号１：1111111111 1111111111 0000000000 0000000000 0000000000
パターン番号２：0000000000 0000000000 1111111111 1111111111 0000000000
パターン番号３：1111111111 0000000000 0000000000 0000000000 1111111111
パターン番号４：0000000000 1111111111 1111111111 0000000000 0000000000
パターン番号５：0000000000 0000000000 0000000000 1111111111 1111111111

他の２パターンとハミング距離20で別の２パターンとはハミング距離40を取るように選んだ。５つのパターンを記憶するための学習は収束する（連想性能が変化しない状態）まで十分な回数（5 x 100～500回）を行い、”温度”パラメータＴは3,10,30の各々の場合で一定としアニーリング（焼なまし）は行っていない。アニーリング（焼なまし）を適切に行えば学習性能および連想性能共に高まる[6]と思われるが、制御の繁雑さを避けると共に処理時間の短縮のために実際の連想メモリーデバイスでは”温度”パラメータＴを一定と

して動作させるほうが望ましい。従って、ここでのシミュレーション評価では実際のデバイスでの使用条件を想定して”温度”パラメータＴを一定とした。連想性能の評価では、記憶したパターンからハミング距離が12（図2.5）あるいは15（図2.6）になるようにランダムにノイズを加えた各記憶パターン毎に100種のテストパターンを生成して入力した。合計500のテストパターンを入力して連想出力されたパターンの、期待パターンからのハミング距離の平均値によって連想性能の指標とした。

図2.5と図2.6に示す結果から、”温度”パラメータＴが一定の条件下では、あるシナプス精度を境に連想性能は急激に劣化することが分かる。また、”温度”パラメータＴが高いほどシナプス荷重値の下限bit精度は低くなる傾向がある。これらの評価結果からシナプス荷重値の精度は、ニューロンの発火確率特性がT=3のほぼロジスティック関数に近い場合ですら、6～7bit精度あれば十分であることが分かった。

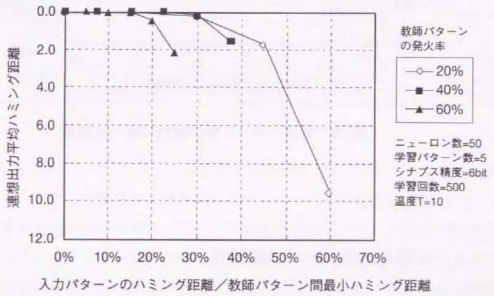


図2.7　記憶パターン発火率の連想性能への影響

図2.7には同一のニューロ連想メモリーに記憶させるパターンの発火率を20%, 40%, 60%と変えた場合の連想性能の変化を、また図2.8には学習する記憶パターンの数を3～8まで変えた場合の連想性能の変化を示す。これらの結果は図2.5と図2.6での条件が特殊な場合でないことを示している。ニューラルネットワークの規模として、ここでは50ニュー

ロンの場合についてシミュレーション評価を行ったが、ニューロン数を増加させると、記憶できるパターン数も増加することから連想性能を評価するために必要なシミュレーションに要する時間が指数関数的に増加し、現状のコンピュータによる現実的な時間での評価は困難であった。

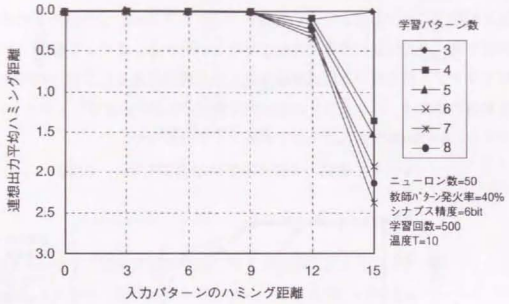


図2.8　記憶パターン数の連想性能への影響

2.3 ニューロ連想メモリーの高集積化

2.3.1 ニューラルネットワーク表現回路の方式比較

半導体集積回路によるニューラルネットワーク機能モデルの表現方式は、第１章1.2節で述べた通り、シナプスの機能表現あるいはニューロンの機能表現に関する回路方式によって大まかに三つに分類することができる。即ち、バイナリデジタル回路方式とパルス変調方式そしてアナログ回路方式である。この中でパルス変調方式は、数値情報を時間方向への拡張し各ノードの信号をバイナリコード化しないことで、各種演算をＯＲやＡＮＤな

どの簡単な論理回路で表現できることから、バイナリデジタル回路方式より高集積化が可能であるが、数値情報を時間方向へ拡張した分、処理速度が遅くなる欠点がある。つまり、パルス変調方式による処理時間とチップ面積の積で表わされるコスト性能比は、バイナリデジタル回路方式と同程度あるいはそれ以下であることから、以後、本論文中ではパルス変調方式はバイナリデジタル回路方式の一種として取り扱うことにする。

実際に試作されたニューロチップ[7][8][9]を参考にして、バイナリデジタル回路方式とアナログ回路方式とで表現した場合の素子数と回路面積それに処理速度について、シナプス機能とニューロン機能の各々の場合について、表2.1 と表2.2 に各々まとめた。

表2.1 シナプス機能表現における回路方式別特徴

荷重値記憶方式	アナログ回路	デジタル回路	デジタル回路	デジタル回路
シナプス演算方式	アナログ回路	アナログ回路	デジタル回路	デジタル回路
素子数	49	629	26,574	48,070
回路面積	3,025 μm^2	48,533 μm^2	3,150,000 μm^2	11,228,000 μm^2
面積比	1	16	1,041	3,712
演算時間	1 ns	1 ns	40 ns	20 ns
演算精度	1 bit x 6 bit MPY Kirchhoff ADD	1 bit x 6 bit MPY Kirchhoff ADD	8 bit x 16 bit int. MPY 32 bit int. ADD	24 bit float MPY 24 bit float ADD

シナプス回路に関しては、シナプス荷重値の記憶保持方式とシナプスの演算方式、すなわち荷重値と入力値との積算の表現方式について４つの場合を表2.1 に示している。供にアナログ回路の場合と荷重値保持がデジタル回路で演算がアナログ回路の場合は第６章で述べるアナログニューロチップ[7]のダイナミックシナプス回路とスタティックシナプス回路の値を参考にしている。また、供にデジタル回路の場合は、整数の演算精度を持ったデジタルニューロチップ[8]と浮動小数点の演算精度を持ったデジタルニューロチップ[9]を参考にしているが、[9]のチップは0.5μmCMOSテクノロジで試作されているため0.8μmに換算している。

また、表2.2 に示すニューロン回路に関しては、入力総和演算と非線形変換供にアナログ回路の場合を前述のアナログニューロチップ[7]を参考に、デジタル回路の場合をSRAMによる非線形変換テーブル方式を採用している前述のデジタルニューロチップ[9]を参考にした。

表2.2 ニューロン機能表現における回路方式別特徴

入力総和演算方式	アナログ回路	デジタル回路
非線形変換方式	アナログ回路	デジタル回路
素子数	103	48,737
回路面積	60,500 μm^2	11,267,000 μm^2
面積比	1	186
演算時間	40 ns	20 ns
演算精度	～8 bit	16 bit

シナプス機能を表現するのに必要な素子数は、アナログ回路方式とバイナリデジタル回路方式との比が約1対500～1000で、回路面積は回路の複雑さに伴う配線領域の増大により、その比がさらに開いて約1対1000～3700となる。また、ニューロン機能表現に関する回路面積比は約1対180である。従って、シナプス数がニューロン数の二乗に比例する全結合ニューラルネットワークを仮定した場合、同一サイズのチップに集積できるネットワーク規模は約1000～3700対1となり、アナログ回路方式は集積度に関してバイナリデジタル回路方式に比べ千倍以上の高集積化が実現されている。また、アナログ回路方式はシナプス荷重積や非線形変換などの演算機能を極めて少ない素子数で表現できることに加え、多数の信号出力端を配線で接続するだけで電流加算が全ての接続ノードに対して並列に実行できるキルヒホフアダーを構成できることから、大規模な並列処理回路構成を容易に表現することができる。その結果アナログ回路方式は、処理速度に関しても、回路面積と消費電力の制限で大規模な並列処理化が困難なバイナリデジタル回路方式に比べ、その並列度の違いに比例して、数千倍の処理速度を実現できる特長がある。

一方、演算精度に関しては、アナログ回路方式が高々6 bit程度であるのに対して、バイナリデジタル回路方式では原理的に任意のビット幅で演算器を作ることができるので、極めて高い演算精度が要求される処理にはバイナリデジタル回路方式のみが対応可能である。従って、バイナリデジタル回路方式は様々な機能モデルに対して比較的自由に機能表現することが可能である特長を有している。また、バイナリデジタル回路方式はそのスイッチ素子に基ずく回路構成により、素子特性のバラツキや電源電圧や環境温度などの変動に対して極めて高い動作マージンを実現できる特長を有しており、素子の微細化による高性能化を可能にした。この高い動作マージンは、第1章で述べた通り情報処理装置の進展に大きく貢献し、現在の殆どのLSIではバイナリデジタル回路方式が採用されている。

表2.1 および表2.2 で示した実際に試作されたニューロチップによる比較結果は、それどれのチップで実現されている演算精度が異なり、用途によっては対等な比較とならないことから、ここでニューロ連想メモリーを実現する場合を想定した演算精度での回路方式比較を行う。前節で述べた通りニューロ連想メモリーに必要なシナプス精度は6bit程度であることから、前述のデジタルニューロチップ[8]を参考にしてそのシナプス演算精度を6bit×1bitに換算して前述のアナログニューロチップ[7]と回路規模および回路面積を比較した結果を表2.3に示す。その場合、シナプス回路の面積比は約1対300、ニューロン回路の面積比は約1対10となる。この結果を基に0.8 μ mCMOSプロセスを用いた場合のニューロ連想メモリーデバイスのチップ面積と処理速度（CPS: Connections Per Second）見積もった結果を図2.9と図2.10に示す。

表2.3 連想メモリー表現に必要な演算精度の場合における回路方式比較

	シナプス回路		ニューロン回路	
荷重値記憶方式	アナログ回路	デジタル回路	アナログ回路	デジタル回路
シナプス演算方式	アナログ回路	デジタル回路	アナログ回路	デジタル回路
素子数	49	8,309	103	15,350
回路面積	3,025 μm^2	984,923 μm^2	60,500 μm^2	525,000 μm^2
面積比	1	326	1	8.7
演算時間	1 ns	40 ns	40 ns	40 ns
演算精度	1 bit x 6 bit MPY Kirchhoff ADD	1 bit x 6 bit int. MPY 16 bit int. ADD	分解能～8 bit コンパレータ等	16 bit int. ADD 8 bit x 256 Table

図2.9には50ニューロン・2500シナプスを表現する場合で、アナログ保持・アナログ演算方式とデジタル保持・アナログ演算方式は50ニューロン・2500シナプスを集積し完全に並列処理ができるが、デジタル保持・演算方式では回路面積の都合で時分割処理を余儀なくされる。デジタル保持・演算方式では演算を実行するニューロン回路とシナプス回路の他に50ニューロンと2500シナプスの値を格納するメモリーが必要であり、集積するニューロン回路とシナプス回路の数によってチップサイズおよび演算速度が変化する。

図2.10には200ニューロン・40Kシナプスを表現する場合を示した。アナログ保持・アナログ演算方式はデジタル保持・アナログ演算方式よりチップ面積で約10分の1、デジタル保持・演算方式より演算速度で約500倍程度の優位性があることが見積もられた。デジタル演算方式における時間分割処理は、大規模な信号のフィードバックが存在する連想記

憶ニューラルネットワークを表現する場合、離散時間表現により状態緩和までの繰り返し演算が必要なことから、並列処理方式との実質的な連想速度の差を更に開くことになると予想される。

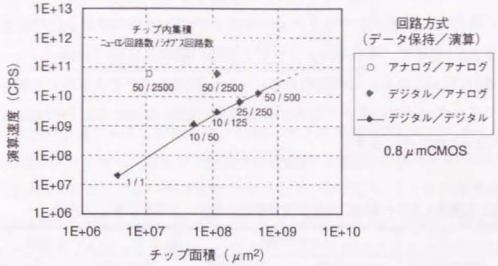


図2.9　ニューロ連想メモリー性能比較（その１）

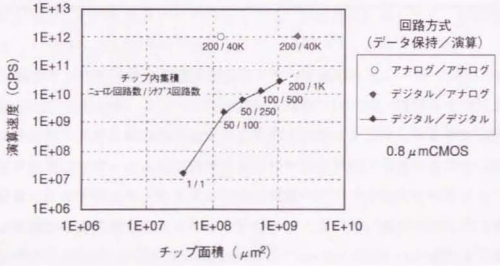


図2.10　ニューロ連想メモリー性能比較（その２）

2.3.2 アナログ集積回路の問題点

アナログ回路方式は連想記憶ニューラルネットワークの高集積化に関して極めて高い優位性があることが見積もられた。しかし、LSIレベルの微細素子によるアナログ集積回路は、その回路構成上、素子特性の素子間バラツキやパラメータ変動が演算結果に直接影響を及す特徴から、素子製造の不均一性や電源電圧・環境温度の変動などに対する十分な動作マージンと高い演算精度を確保することが難しい問題がある。特に、素子特性のバラツキは素子の微細化が進むにしたがって顕著に増大することから、アナログ集積回路の集積化を阻む大きな要因となる。ここでは、アナログ集積回路の高集積化を制限する要因について整理し、その問題点を明らかにする。

一般に、半導体集積回路の微細化限界は、機能の再現性とその安定性が確保できなくなる所で決まり、信号または電源電圧のゆらぎや熱ノイズなどに起因する動的な機能特性バラツキと、素子の製造過程で生じるドーズ量やデバイス形状のゆらぎに起因する静的な素子特性バラツキ、そしてデバイスの放熱限界から制限される消費電力などで見積もることができる。

まず最初に、信号と熱ノイズの比が素子の微細化に伴いどのように変化するかを調べる。ニューラルネットワークLSI内部における最も微弱な信号はシナプスの信号であり、その電圧レベルは一つのニューロンに接続されるシナプスの数Nsと電源電圧Vddで決まる。図2.11のアナログ回路構成例で示すように、ニューロンの入力信号は入力ノードに接続された全てのシナプス回路の出力電流が共通ノードで足し合わされ、抵抗RLによって電圧に変換され非線形変換器（コンパレータなど）に与えられる。従って、ニューロンの入力ダイナミックレンジをVddの3/5とすると、一つのシナプス当たりの信号電圧Ssは次式となる。

$$Ss = (3/5) \cdot Vdd / Ns \qquad (2.3)$$

0.8μmCMOSプロセスで実際に試作したアナログニューロチップ[7]では、Ns=200、Vdd=5Vなので、その時の信号電圧レベルSsは15mVである。

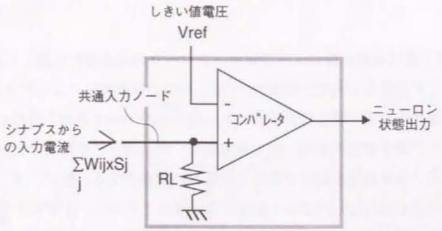
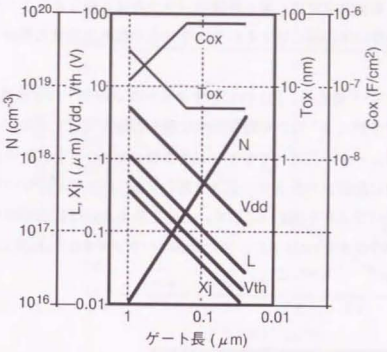


図2.11 ニューロンのアナログ回路構成

一方、ノイズの電圧レベルは熱ノイズエネルギーと放熱限界で規定される最大消費電力から導出される。室温での熱ノイズエネルギーは約4.0×10^{-21}Jで、放熱限界を10W/cm^{2}とし、反応時間を100psとすると、0.8μmCMOSのアナログニューロチップ[7]の例では、38Kシナプス/cm^{2}であることから一シナプス当たりの最大消費電流は53μAとなり、熱ノイズ電圧Snは約0.76μVと見積もられる。

そこで、素子の微細化が進むに従ってこれらの信号と熱ノイズの電圧レベルがどのように変化するかを見積もってみる。図2.12にMOS素子のスケーリング則を示す。酸化膜厚Toxはトンネル電流の増加を防ぐために50Å以下には薄くできず、不純物のドーズ量Nはデバイスチャネル内の電界を一定に保つために微細化に伴い増加させる必要がある。素子の微細化が図2.12のスケーリング則に従うとすれば、シナプスの信号電圧Ssは$1/k^2$でスケーリングされ、熱ノイズ電圧Snはシナプス数に逆比例するのでk^2で増加することになり、図2.13に示すように、微細化が進むに従ってノイズマージンが急速に縮小することが見積もられた。0.1μmレベルになると熱ノイズは、信号の約21%に及ぶことが予想される。



スケーリング係数

パラメータ		係数
電源電圧	Vdd	1／k
しきい値	Vth	1／k
ゲート長	L	1／k
接合深さ	Xj	1／k
酸化膜厚	Tox	～1
ドーズ量	N	k

図2.12 MOSトランジスタのスケーリング則

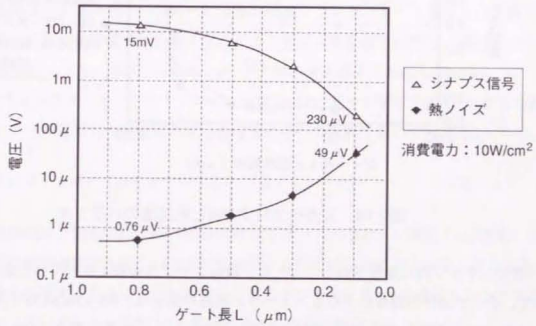


図2.13 シナプス信号と熱ノイズ

次に、素子の寸法やドーズ量の統計的ゆらぎで生じる、素子特性の静的バラツキが及ぼす影響について考察する。本論文中では、素子特性のバラツキは正規分布をとると仮定し、バラツキの程度を素子特性の平均値に対するバラツキ分布の標準偏差値の割合（%）で表わすこととする。

まず、ニューロン回路について考える。図2.11に示すニューロンのアナログ回路構成によって、素子特性のバラツキがニューロンの機能精度に最も影響を及ぼすものは、RLの抵抗値のバラツキとコンパレータの入力ペアトランジスタ間のVthのバラツキである。抵抗値のバラツキおよび近傍に配置されたトランジスタ間のVthのバラツキについて、測定によって得られた素子特性バラツキを図2.14に示す。ゲート長あるいは抵抗の幅を比較的大きくすれば、抵抗値のバラツキを1%以下に、またVthのバラツキを0.5%程度に押えられることが期待できる。

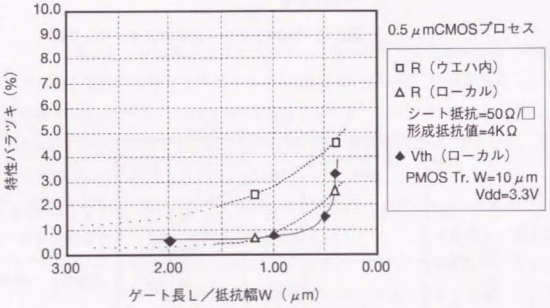


図2.14　入力ペアTr.のVthと抵抗値のバラツキ

一般に、チップ内に集積されるニューロン数はシナプスの数と比べてけた違いに少ないので、チップ全体の面積に占めるニューロン回路の割合は、0.8μmCMOSで試作したニューロチップ[7]の場合で約12%と比較的小さい少ない。しかもその割合は図2.15内に示すように素子の微細化が進むに従って更に少なくなる。従って、ニューロン回路に関しては、その精度を確保するために一部の部品（入力ペアトランジスタと抵抗RL）のみを相対的

に大きくしても、図2.15内に示すようにチップ全体面積増加に至らないと考えられる。

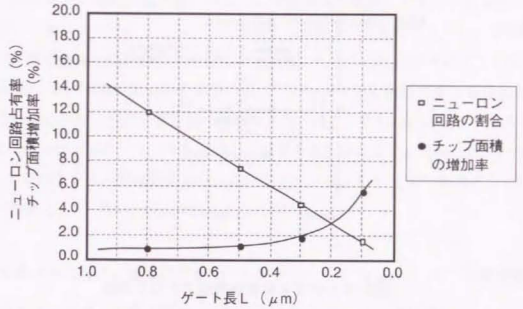


図2.15　ニューロン回路のバラツキを抑制した場合のチップ面積増加率

次にシナプス回路に関する素子特性バラツキの影響について考察する。図2.16に、シナプスの機能を表現するアナログ回路の例を示す。シナプスの基本的機能を表現するこのアナログ回路は、1個のキャパシターと1個のMOSトランジスターで構成されており、キャパシターに負の電荷量－Qを蓄えることでシナプス荷重値を保持し、MOSトランジスターを流れるソース・ドレイン電流Ioでシナプス荷重値Wを表現している。

MOSトランジスターのソース・ドレイン電流Ioは、ゲート電圧Vg（Vddからの電圧）がキャパシターの容量Cと蓄積電荷量Qとの積となって次式で表わせられる。

$$Io = 1/2 \cdot \varepsilon ox / tox \cdot \mu \cdot W/L \cdot (Q/C - Vth)^2 \qquad (2.4)$$

但し本回路構成は、電源電圧Vddの変動に対してソース・ドレイン電流Ioが変動しないように、常に飽和領域でトランジスタが動作する条件で用いることとする。また負のシナプス荷重値を表現するために、予め設定した正のオフセット値Wbを導入し、シナプス荷重値Wとソース・ドレイン電流Ioの関係を次式のとおり表現する。

$$W = -Wb + Io \qquad (2.5)$$

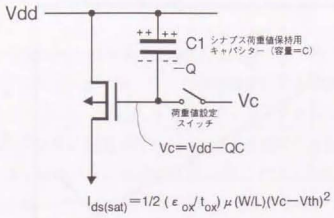


図2.16 シナプス荷重値表現アナログ回路

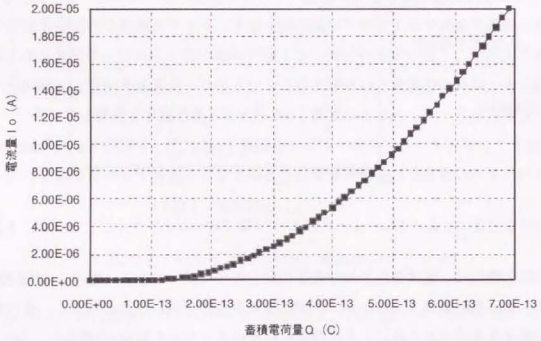


図2.17　蓄積電荷量と電流量の関係

このアナログ回路における動作精度に関する問題点は２つある。その一つは、図2.17に示すような、キャパシターに蓄積された電荷量Ｑとシナプス荷重値Ｗとの非線形な関係である。本回路構成では第3.6節で述べるように電源電圧の変動に対する十分な動作マージンを確保する為に、トランジスターを飽和領域で動作させていることから、蓄積電荷量Ｑとシナプス荷重値Ｗとは非線形な関係になる問題があることに注意する必要がある。

第二の問題点は、半導体集積回路素子の製造において生じる統計的ゆらぎに起因する素子の静的特性の不均一性が演算性能へ及ぼす影響である。0.5μmCMOSプロセスで製造した素子のＶth及びＩdsのバラツキを測定した結果を図2.18に示す。異なったゲート長ＬのＶthとＩdsのバラツキの測定結果から、各パラメータごとのバラツキを次のように見積もることができる。

トランジスターのしきい値電圧Ｖthのバラツキ	σ_{Vth}≒2.4%	（測定値）
トランジスターのゲート酸化膜厚ｔoxのバラツキ	σ_{tox}≒1.2%	
トランジスターのゲート長Ｌのバラツキ	σ_{L}≒2.4%	
トランジスターのゲート幅Ｗのバラツキ	σ_{W}≒0%	

ここで、σ_{W}はトランジスタのＷがＬと比べて十分大きいことから0%とした。また、σ_{L}及びσ_{tox}とキャパシター面積（～30μm^2）からの容量Ｃのバラツキがもとまり、ゲート電圧Ｖgのバラツキσ_{Vg}が見積もられる。

キャパシターの容量Ｃ及びゲート電圧Ｖgのバラツキ　σ_{C}≒σ_{Vg}≒1.3%

導出されたこれら0.5μmレベルの素子の特性バラツキが、図2.12のスケーリング則に基づき微細化された場合、シナプス荷重値を表現するＩdsのバラツキは図2.19に示す通りに見積もることができる。素子の微細化によるシナプス荷重値のバラツキは、0.2μmレベルから急速に増大する事が分かる。図中のΔＬはゲート長のバラツキを表わす。

次の節では、このシナプス荷重値のバラツキが連想性能に及ぼす影響について調べる。

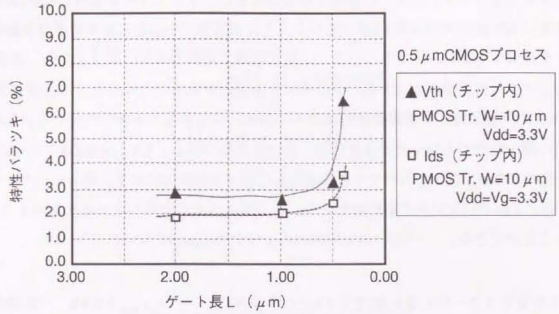


図2.18　MOSトランジスタの特性バラツキ

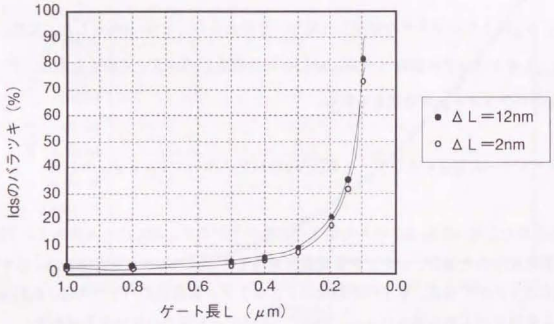


図2.19　微細化に伴う素子特性バラツキの増大

2.3.3 素子特性バラツキの連想性能への影響

計算機シュミレーションによって評価した素子特性のバラツキによって生じる連想性能の劣化を図2.20 に示す。このシミュレーションは2.2節で行ったのと同一の、50ニューロンの完全フィードバック結合型ニューラルネットワークに第2.2節で示したのと同一の5個のパターンを学習して記憶し、学習したパターンに対して5通りのハミング距離毎に各々100通りの入力テストパターンを生成して、自己想起による連想出力パターンの期待パターンからのハミング距離の平均値を用いてその連想性能を評価している。

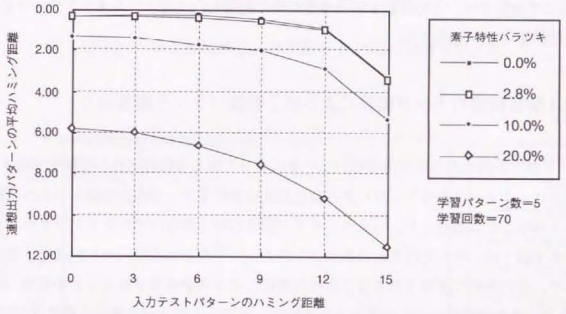


図2.20　素子特性バラツキの連想性能への影響

このシミュレーションでは静的な素子特性のバラツキについて0%, 2.8%, 10%, 20%の4通りについて評価した。また、信号ノイズで生じる電源電圧や信号のゆらぎによる動的な機能特性バラツキは、回路パターンの形状により特定の信号間に相関を持つと考えられるので、その機能モデルを一般化することが困難なことから、ニューロン状態の緩和期間で平滑化されると仮定し無いものとした。但し、熱ノイズに伴う動的な機能特性バラツキ

は正規分布による一般的表現が可能なので、その機能的な動的バラツキがニューロン回路のコンパレータのしきい値の変動に反映されると仮定し温度パラメータTを30と設定した。

このシミュレーションによれば、素子特性のバラツキが3%程度までは、ニューラルネットワークダイナミックスの構造安定性により殆ど連想性能の劣化が生じないことが分かる。しかし、それ以上のバラツキでは、入力パターンのハミング距離によらず、連想性能が徐々に劣化することが分かる。素子特性バラツキが10.0%の場合では、連想結果が平均ハミング距離で1から2程度悪化しており、もはや、記憶したパターンが完全に復元できない状態になっている。

このようにアナログ回路方式の場合、その素子特性のバラツキによって連想性能に無視できない影響を与えることが見積もられた。従って、この問題を克服しない限り、アナログ回路方式によるニューロ連想メモリーデバイスの実用化は困難となる。

そこで次節では、この問題を解決するために着目したニューラルネットワークの自己組織化機能について述べる。

2.4 学習機能のチップ実装による素子特性バラツキ補償能力

アナログ回路における素子特性バラツキによって生じる性能劣化の問題を克服する為に我々は、ニューラルネットワークの自己組織化機能をチップ内の回路レベルへ積極的に取り入れることに着目した。つまり、チップ固有の素子特性バラツキやパラメータ変動などの不良因子は、それを内在したチップ上のニューラルネットワークにおける学習処理によって、その過程で実現される自己組織化機能により自動補償されることを期待した。

また、その補償能力が十分強力である場合、更なる大胆な回路構成の簡略化が可能になり、学習機能を搭載することの面積デメリットをも吸収することができると考えた。

図2.21 は、チップ内の素子特性バラツキが18.8%の条件で、学習機能を実装したニューロチップとそうでない場合との連想性能の違を、前述の計算機シミュレーションにより評価した結果を示す。学習機能を実装することで、素子特性バラツキが無い場合とほぼ同等の性能が得られることが見積もられた。この学習シミュレーションでは、前述のシナプス荷重値の非線形特性と第3章で述べる学習回路の荷重修正非線形特性も考慮されており、この学習による自己補償能力は、素子特性のバラツキのみならず回路の非線形特性すら十分に補償できることが見積もられた。

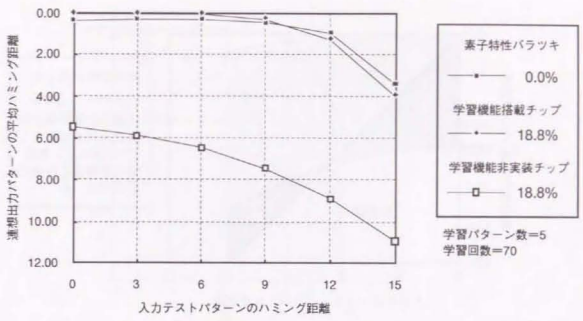


図2.21　学習機能実装／非実装チップ間の連想性能比較

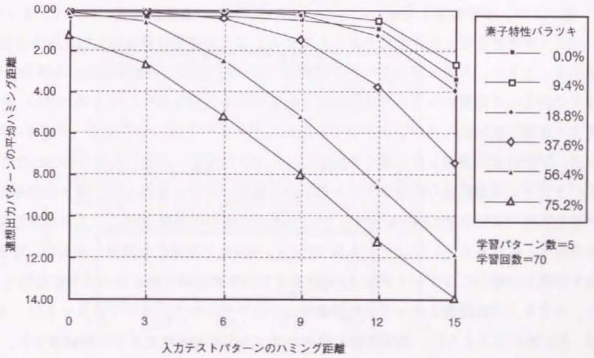


図2.22　オンチップ学習による素子特性バラツキ補償能力

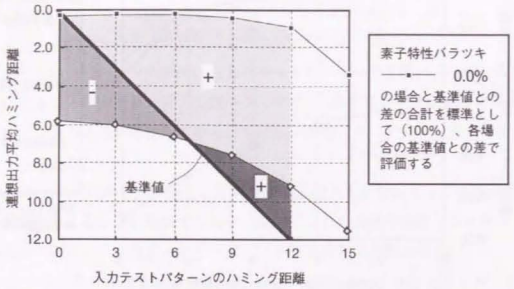


図2.23　連想性能の評価方法

図2.22 は、学習機能を実装したニューロチップにおける連想性能について、素子特性のバラツキ程度を変えた評価結果を示している。ここで連想性能を図2.23に示す方法で評価する。つまり、入力と出力のハミング距離（0～12）が等しい基準値線と各連想特性曲線とで囲まれる面積の大きさで連想性能を定量化する。素子特性バラツキが無い（0.0%）場合を連想性能100%として、図2.19で示した素子のゲート長と素子特性バラツキの関係を基に学習機能を搭載しない場合の図2.20および学習機能を搭載した場合の図2.22内の素子バラツキと連想性能の関係をゲート長と連想性能の関係に換算して、素子の微細化に伴う連想性能の変化を図2.24内に示す。ここで連想性能の劣化を-10%、つまり連想性能90%まで許すと仮定すると、にはゲート長が約0.4μm程度で微細化の限界となるが、学習機能を搭載した場合にはゲート長0.15μm程度までの微細化が可能となることが見積もられた。つまり、学習機能をチップ上に搭載することでアナログニューラルネットワークLSIは、図2.25に示すように、素子特性の静的バラツキが約30%程度まで自動補償でき、デジタルLSIの微細化限界[10][11]とほぼ同じ、0.15μmレベルまで微細化が可能であることが見積もられた。

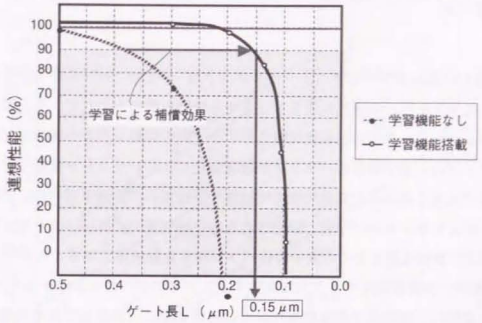


図2.24　素子の微細化に伴う連想性能の劣化

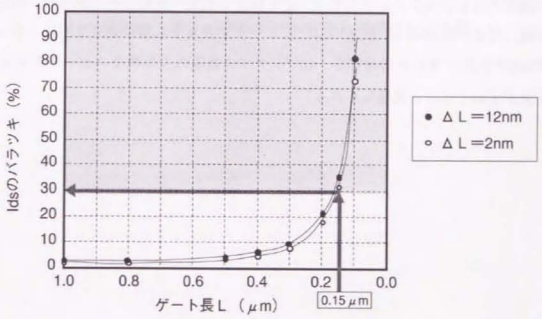


図2.25　オンチップ学習で補償できる素子特性バラツキの範囲

2.5 まとめ

高集積化に有効なアナログニューラルネットワークLSIにおける素子特性のバラツキが、ニューロ連想メモリーの連想性能を劣化させる問題を明らかにして、その解決策としてニューラルネットワークの自己組織化機能をチップ内の回路レベルへ積極的に取り入れることで、チップ内の素子特性のバラツキや環境変化によるパラメータ変動などの不良因子を、学習過程で実現される自己組織化機能で補償する技術の可能性を調べた。計算機シミュレーションによるオンチップ学習で期待される自己補償能力の評価によって、回路の非線形特性に加えて約30%程度までの素子特性バラツキを自動補償できることが明らかになった。

この結果、学習機能をチップ上に実装することでアナログニューラルネットワークLSIの欠点を克服し、高精度で高集積なニューロ連想メモリーを実現できる見通しが得られた。つまりニューロデバイスは、図2.26内に示すように、2000年頃には最小線幅0.15μmに微細化された半導体集積回路によって、1つのニューロチップで数千万シナプス規模を集積し、数百テラCPSの演算速度性能に達し、そのチップを数十個搭載したボードレベルで大脳に匹敵するレベルの演算速度が実現できると予測することができる。さらに多数のボードで構成されるシステムレベルでは、その数百倍程度の性能拡張が図られ、時分割規模拡張表現（付章）によって、従来のコンピュータで表現する場合より2桁ほど少ない、数千ギガByteの汎用メモリーの追加で、10^{14}シナプス規模の大規模なニューラルネットワークの表現が実現できると見積もられる。

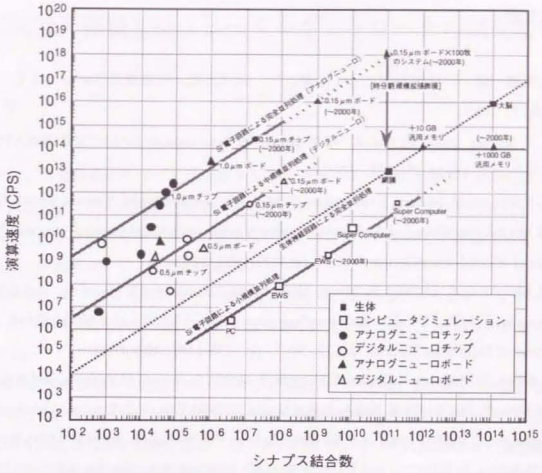


図2.26　ニューラルネット規模と処理速度の進展

参考文献

[1] 安西祐一郎、”認知科学と人工知能、” 共立出版、計算機科学／ソフトウエア技術講座17、1987。

[2] K. J. Schultz, and P. G. Gulak, "Architectures for large-capacity CAMs," INTEGRATION, the VLSI journal, 18, pp.151-171, 1995.

[3] T. Yamagata, M. Mihara, T.hamamoto, Y. Murai, T. Kobayashi, M. Yamada, and H. Ozaki, "A 288 kbit fully parallel content addressable memory using stacked capacitor cell structure," IEEE, Journal of Solid-State Circuits, Vol.27, pp.1927-1933, 1992.

[4] M. Motomura, J. Toyoura, K. Hirata, H. Ooka, H. Yamada, and T. Enomoto, "A 1.2 million transistor, 33MHz, 20 b dictionary search processor (DISP) ULSI with a 160 kb CAM," IEEE, Journal of Solid-State Circuits, Vol.25, No.5, pp.1158-1165, May,1990.

[5] Y.Arima, K.Mashiko, K.Okada, T.Yamada, A.Maeda, H.Notani, H.Kondoh, and S.Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," IEEE, Journal of Solid-State Circuits, Vol.26, No.11, pp.1637-1644, Nov.,1991.

[6] D H.Ackley, G E.Hinton, and T J.Sejnowski, "A Learning Algorithm for Boltzmann Machines," Cogunitive Science, Vol.9, No.1, pp.147-169, Jan-Mar, 1985.

[7] Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," IEEE, Journal of Solid-State Circuits, Vol.27, No.12, pp.1854-1861, Dec.,1992.

[8] M.Griffin, G. Tahara, K. Knorpp, R. Pinkham, and B. Riley, "An 11-million transistor neural network execution engine," ISSCC, Digest of Tecnical papers, pp.180-181, Feb., 1991.

[9] Y. Kondo, Y. Koshiba, Y. Arima, M. Murasaki, T. Yamada, H. Amishiro, H. Shinohara, and H. Mori,"A 1.2GFLOPS Neural Network Chip Exhibiting Fast Convergence," in ISSCC, Digest of Technical Papers, pp.218 -219, Feb. 1994.

[10] C. Mead, and L. Conway,"INTRODUCTION TO VLSI SYSTEMS," Addison-Wesley Publishing Company, Inc., 1980.

[11] 菅野、”デバイス動作の理論限界、” 電子情報通信学会誌、Vol.75、No.4、pp.326 - 332、April、1992。

第3章
学習機能を搭載したニューラルネットワークの高集積化

3.1 序

前章では、学習機能をチップ上に実装することでアナログ集積回路の欠点が克服でき、ニューロ連想メモリーの高精度化および高集積化が実現できる可能性を明らかにした。そこで本章では、学習機能をチップ上に高集積に実装するために提案した回路構成と、その学習性能および動作マージンについて述べる。

第3.2節では、ニューロ連想メモリー機能を実現するためのニューラルネットワーク機能モデルと、その学習アルゴリズムに対する半導体集積回路に適した近似表現の要点について述べる。

第3.3節では学習機能を実現するニューロン回路について述べ、第3.4節では高集積化のために考案した、チャージポンプ回路による荷重値修正回路と簡単な論理ゲートによる学習制御回路で構成される学習機能実装シナプス回路について、その回路構成と機能動作について述べる。第3.5節ではそれらの回路に基ずくニューラルネットワークの回路構成とその制御フローについて述べる。

第3.6節では、高集積化のために大胆に簡略化された、これら学習回路に関して、学習ルールの近似と回路の非線形特性が学習性能へ及ぼす影響について、計算機シミュレーションによる評価結果について述べる。また、電源電圧変動に対する動作マージンを拡大するために採用した、シナプス荷重値表現MOSトランジスターの飽和領域動作回路構成に対する効果を回路シミュレーション（SPICE [1]）によって評価した結果について述べる。

第3.7節では、連想記憶ニューラルネットワークLSIの素子微細化に伴う高性能化について予測する。

3.2 連想記憶ニューラルネットワークの機能モデル

ニューロ連想メモリーデバイスには任意のパターンを書き込む（記憶）機能と記憶し

たパターンを読み出す（連想）機能が必要である。フィードバック結合型のニューラルネットワークによって連想メモリー機能を実現することができるが、パターンを記憶するためには学習機能を実装する必要がある。半導体チップ上に学習機能を実装する場合、その高集積化のポイントは、より少ない信号配線と回路規模で学習回路を実現するために、ローカルな情報のみを使い比較的簡単で精度が要求されない学習アルゴリズムを採用することである。ボルツマンマシン（Boltzmann Machine）[2]の学習アルゴリズムは、まさにローカルな情報のみを使い比較的簡単な荷重値修正ルールで実現できることから半導体チップ上への学習機能実装に適している。

ボルツマンマシンは1983年にHintonとSejnowski [2]によって提案された、対称なシナプス荷重値（Wij＝Wji）を持つフィードバック結合（相互結合）型のニューラルネットワークで、ニューロンの出力状態が”１”（活性状態）になる確率Ｐ(Ｓi＝１)は次式で表わされる。

$$P(Si=1)=1/(1+\exp(-ui/T)) \qquad (3.1)$$

$$ui=\Sigma_j Wij \cdot Sj \qquad (3.2)$$

ここで、uiはニューロンiの入力総和値（内部活性値）、Ｔはシステムの”温度”と呼ばれるパラメータで、各ニューロンのしきい値は省略している。対称な結合と式3.1、3.2によるニューロンの確率的状態遷移規則によって、ニューラルネットワークの平衡状態の出現確率Ｐはボルツマン分布に従、ある状態αの出現確率$P(\alpha)$は次式で表わせる。

$$P(\alpha)=\exp(-E(\alpha)/T)/\Sigma_\beta \exp(-E(\beta)/T) \qquad (3.3)$$

ここで、$E(\alpha)$、$E(\beta)$は、各々平衡状態α、β状態における次式（リヤプノフ関数[3]）で表わせるニューラルネットワークのエネルギーである。

$$E=-\Sigma_{i<j} Wij \cdot Si \cdot Sj \qquad (3.4)$$

ボルツマンマシンの学習アルゴリズムは、1985年にAckley[4]らによって導出された。

そこで、外部との信号入出力を担う”可視ニューロン”に加えて、外部との直接的な信号のやり取りを行わない”隠れニューロン”を含んだ回路網についての学習アルゴリズムを導出したことによって、複雑な情報処理構造を獲得できる高度な学習ができるようになった。このボルツマンマシンの学習アルゴリズムは、ニューラルネットワークに与える教師情報の出現確率とボルツマンマシン内部で実現されるネットワーク平衡状態の確率分布との違いを表わす相対エントロピー（Kullbackのdivergence）Ｇを最小にするシナプス荷重値の修正規則から導出される。

$$G=\Sigma_I \{P(I) \cdot \Sigma_O (R(O|I) \cdot \log(R(O|I)/P(O|I)))\} \qquad (3.5)$$

$$\Delta Wij=\eta' \delta G/\delta Wij=\eta'/T \cdot (pij^+-pij^-) \qquad (3.6)$$

ここでＩ、Ｏは各々入力、出力ニューロンの変更状態を表し、Ｐ(I)は入力ニューロンの状態Ｉの出現確率、Ｐ(O|I)は入力と出力ニューロン状態が各々ＩとＯの出現確率で、Ｒ(O|I)は外部から与えられる教師情報の入力と出力ニューロン状態が各々ＩとＯの出現確率である。ΔWijはWijの修正量を表す。η'は正の定数、pij^+は可視ニューロンを固定した状態で回路網が平衡状態に達した状態でのニューロンｉとｊが供に”１”（活性状態）になる確率を表わす。また、pij^-は入力ニューロンのみを固定した状態で回路網が平衡状態に達した状態でのニューロンｉとｊが供に”１”（活性状態）になる確率を表わす。

ニューロ連想メモリーデバイスの記憶機能を実現するために、このボルツマンマシンの学習アルゴリズムを次のように修正して導入する。また、全てのニューロンは状態更新を同時に行える回路構成とした。これらの修正を行っても、連想メモリーの記憶（学習）機能は十分に実現されることは、試作したニューロチップによる学習機能評価で確認された[5][6][7]。

まず学習機能の高集積化のために、荷重値修正ルールは、回路規模が増大する中間値（０～１）表現の確率pij^+、pij^-の代りに２値（０、１）しかとらない各状態出力の積$Si^+ \times Sj^+$、$Si^- \times Sj^-$に置き換えた次式に修正する。

$$\Delta Wij=\eta (Si^+ \cdot Sj^+ - Si^- \cdot Sj^-) \qquad (3.7)$$

ここでηは正の定数で学習係数と称する。Si^{+}とSj^{+}は可視ニューロンを固定した状態で回路網が平衡状態に達した状態でのニューロンｉとｊの状態出力値。
Si^{-}とSj^{-}は入力ニューロンのみを固定した状態で回路網が平衡状態に達した状態でのニューロンｉとｊの状態出力値である。また、一時的なデータ保持回路を省くために次式のように時間分割でη（$Si^{+}\cdot Sj^{+}$）と$-\eta$（$Si^{-}\cdot Sj^{-}$）を実行する場合もある。

$$\Delta Wij=\pm\eta\ (Si^{\pm}\cdot Sj^{\pm}) \qquad (3.8)$$

次に、連想メモリーとして自己想起用にパターンを記憶するために、全ての可視ニューロンを出力ニューロンとして取り扱い、$Si^{-}\times Sj^{-}$を算出する過程でネットワークの初期値として学習パターンを用いる様に学習アルゴリズムを修正する。また、ネットワークの温度Ｔを低く一定に保った状態で前述の学習ルールを実行することで、単純に空間分布した記憶パターンを等しい出現確率として、想起用入力パターンからの距離のみで記憶パターンが想起される、最も基本的な連想メモリーを実現することができる。

これらの学習アルゴリズムの修正による学習性能あるいは連想性能への影響については、計算機シミュレーションによる評価結果を第3.6節で述べる。

3.3 学習機能を備えたニューロン回路

図3.1 に考案したニューロン回路[6]の機能構成を示す。共通入力ノードに接続されたシナプス回路からの出力電流（荷重化出力信号）がキルヒホフの法則に従って足し合わされ（キルヒホフアダー）、抵抗ＲＬによって電圧に変換されてコンパレータに入力される。コンパレータのもう一方の入力端子には、ニューロンのしきい値を表現する基準電圧Ｖrefがチップ外部から与えられる。ニューロンの状態遷移における確率過程は、しきい値電圧V_{θ}に一定の最大振幅を持ったノイズ信号を加えた電圧を基準電圧Ｖrefとして与えることで疑似的に表現している。加えるノイズの最大振幅がシステムの温度Ｔに対応する。従って、シミュレーテッドアニーリングは、ノイズの振幅を次第に減少させることで疑似的に表現できる。しかし、記憶パターンを等出現確率で記憶する用途の連想メモリーデバイスの場合、実際には基準電圧Ｖrefを一定として、自然に生じる熱ノイズによる成分以

外に積極的な交流成分の添加を必要としない。第４～６章で述べる試作したニューロ連想メモリーデバイスにおいて基準電圧Ｖrefは一定値に固定し連想性能を評価したが、十分な連想性能が得られている[5][6][7]。

コンパレータの出力はニューロンの内部活性値に対応した状態出力を表す。シフトレジスターＳＲ(T)には、本ニューロンの教師パターンデータが格納され、ＳＲ(P)には、本ニューロンの属性（隠れ又は入力・出力）データが格納される。本ニューロンの状態出力Ｓiは、ＮＯＲゲートとセレクターＳＥＬ１で制御され、ＳＲ(P)に格納された属性データと制御信号ＩselSとに従って、表3.1 に示す通り各々の属性によって各学習フェーズ毎に、内部活性値による状態か教師データかが選択され出力される。つまり制御信号ＩselSは、ニューロンの出力状態を教師データで固定するか内部活性値による状態にするかを選択制御する信号で、”Ｌ”の時にはＳＲ(T)に格納された教師データを”Ｈ”の時には内部活性値による状態を選択出力する。但しＳＲ(P)に格納されたニューロンの属性データが”Ｈ”で隠れニューロンを示している場合には、ＮＯＲゲートによりＩselSの信号にかかわらずセレクターＳＥＬ１は常に内部活性値による状態を選択出力する。また入力ニューロンに属性を指定する場合には、ＩselS信号を常に”Ｌ”固定する必要がある。この構成によって、各ニューロン毎に任意の属性（入力または出力、隠れ）と教師データ（”０”または”１”）を設定することができる。

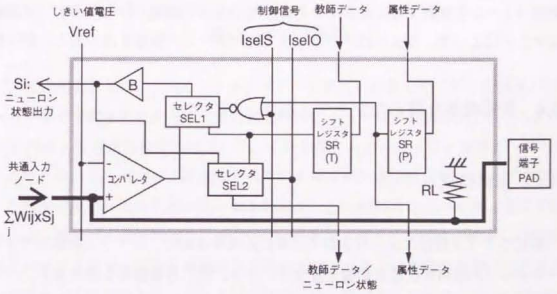


図3.1 ニューロン回路

セレクターＳＥＬ２は、各ニューロンの状態をモニターするのに使われる。制御信号ＩselOは通常”Ｈ”となっており、セレクターＳＥＬ２はシフトレジスタＳＲ(T)の出力を選択出力しているが、制御信号ＩselOを”Ｌ”にするとセレクターＳＥＬ２は、ニューロンの状態出力Ｓiを選択出力する。共通入力ノードに接続された信号端子ＰＡＤは、ニューロンの内部活性電圧値のモニター用あるいは、第５章で述べる回路網の拡張時にチップ間接続端子として用いられる。

表3.1 ニューロンの属性と出力状態

ニューロンの属性	学習フェーズ	SR(P) データ	IselS	ニューロンの状態出力
隠れ	+	H	L	内部活性値
	-		H	内部活性値
出力	+	L	L	SR(T) データ
	-		H	内部活性値
入力	+	L	L	SR(T) データ
	-		L	SR(T) データ

このニューロン回路構成によれば、教師パターンを各ニューロンのＳＲ(T)に格納した後、次節で述べるシナプス回路の学習制御信号とニューロンの基準電圧Ｖrefそして制御信号ＩselSを、学習フェーズに従って操作することによって、ニューロチップは与えられた教師パターンを学習することができる[6]。また全ての教師パターン毎にこの操作を繰り返すことによって、ニューロチップは全てのパターンの記憶を深めることが可能となる。

3.4 学習機能を備えたシナプス回路

3.4.1 回路構成

一般にシナプス数はニューロン数の二乗に比例するので、シナプス回路のサイズがニューロチップの集積度に最も影響を及ぼす。一方、自己補償機能を回路素子レベルに有効に働かせるためには、完全な並列処理構成が望ましく、各シナプス毎に並列に実行できる学習回路を備えることが必要である。そこでシナプス回路を極力小さくするために、演算

精度と線形性を犠牲にすると供に学習ルールを近似表現することで、素子数を少なくした。

図3.2 に考案したシナプス回路[7]を示す。シナプス荷重値は、第2.3節で述べた様に、キャパシターＣ１に蓄えられる電荷量で表現している。また対称なシナプス結合を効率よく表現するために、キャパシターＣ１に接続された二つのシナプス結合演算回路を備えて、双方向の結合を一つのシナプス回路で表現できる構成になっている。

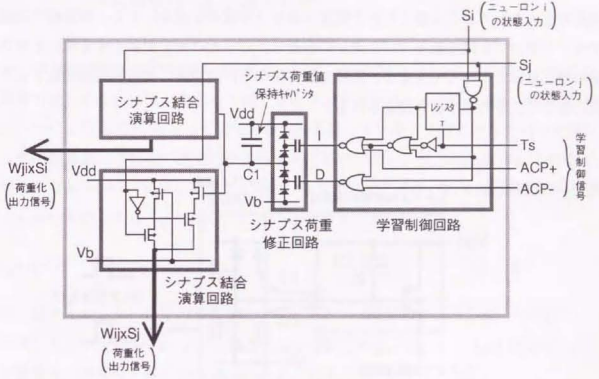


図3.2 シナプス回路

シナプス結合演算回路は、pMOSトランジスターを直列に接続した二つの定電流パス回路とインバータで構成されており、二つの電流パスはニューロンの状態入力信号Ｓj（またはＳi）に対応して相補的に活性化される。すなわち、入力ニューロン状態が”Ｈ”（活性状態）の時には、Ｃ１がゲートに接続されている電流パスが活性化し、シナプス荷重値に対応した電流が流れ、また入力ニューロン状態が”Ｌ”の時には、バイアス電圧Ｖｂが接続されている電流パスが活性化し、Wij×Ｓj＝０を表すバイアス電流に対応した電流が流れ出る。バイアス電流を流す定電流パス回路のトランジスター駆動能力を、シナプス荷重値を流す電流パス回路のトランジスターのほぼ半分に設定することで、Wij×Ｓj＝０を表すバイアス電流値をシナプス荷重値が最大の場合の電流量のほぼ中央値にすることができる。従って、このバイアス電流値より少ない電流は負の結合荷重を意味し、多い電流は

正の結合荷重を意味する。二つのシナプス結合演算回路は異なるニューロンの状態入力信号Ｓj、Ｓiが接続されており、各々Wij×ＳjとWji×Ｓiに対応する電流を出力する。

シナプス荷重値として保持されているキャパシターＣ１の蓄積電荷を修正する為の荷重修正回路は、図3.3内に示すように、直列に接続された二つのチャージポンプ回路から構成される。上方の入力端子Ｉにパルス信号が与えられると、Ｃ１に蓄えられた電荷は汲み出されて減少（負電荷は増加）するし、下方の入力端子Ｄにパルス信号が与えられると、電荷が注入されてＣ１に蓄えられた電荷は増加（負電荷は減少）する。荷重修正回路内のチャージポンプを構成駿キャパシターの容量は、Ｃ１との比を大きくする程シナプス荷重値の修正単位を小さくできるが、回路上に寄生する浮遊容量による修正感度低下による制限によって、その最小値が決定される。

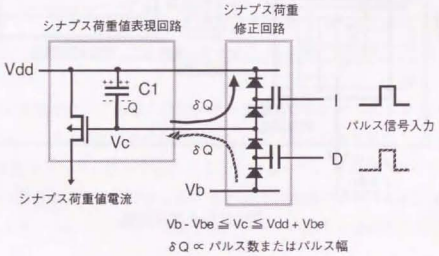


図3.3　荷重修正回路

シナプス荷重修正回路は、学習ルールに従って荷重値を修正するように学習制御回路により制御される。全シナプス回路に共通に与えられる学習制御信号Ｔｓ，ＡＣＰ＋，ＡＣＰ－は、想起を実行中など荷重値を修正しない時には全て”Ｈ”レベルに固定されている。学習時には、可視ニューロンが教師状態で固定される＋フェーズにおいて平衡状態に達したときＴｓに負のパルス信号を与えて$Si^{+}\times Sj^{+}$の反転信号をレジスタに保持し、続く入力ニューロンのみ教師状態で固定される－フェーズにおいて平衡状態に達した後ＡＣＰ＋，ＡＣＰ－にそれぞれ負のパルス信号が与えられることで、そのときのＳi

$^{-}\times Sj^{-}$の反転信号とレジスタに保持されている$Si^{+}\times Sj^{+}$の反転信号との論理演算結果によってＩまたはＤに負のパルス信号を出力する。この学習制御回路によって、式3.7で表されるボルツマンマシンの近似学習ルールが実現される。ここで学習係数ηは、各ＡＣＰに与えるパルス数またはパルス幅により表現することができる[7]。

3.4.2 シナプス荷重値修正の非線形特性

シナプス荷重値修正回路を構成するチャージポンプ回路は、パルス状の信号入力によって電荷を注入あるいは汲み出すことができるが、１回のパルスで修正できる電荷量はキャパシターＣ１の電圧Ｖcによって変化する特性を持っている。１回のパルス信号で修正される負の電荷量ΔＱは次の式で近似表現することができる。但し与えるパルス信号は、信号電位がＶdd→Gnd→Ｖddと変化する負のパルス形状で、そのパルス幅は回路の時定数より十分長いものとする。入力端子Ｉにパルス信号を与えた場合は、

$$\Delta Q(+)=C\ (Vb+Vt-Vc)\ /\ r \qquad (3.9)$$

ここでｒはキャパシターＣ１の容量Ｃとチャージポンプ回路のキャパシターの容量比とダイオードのしきい値電圧Ｖtによって決まる正の定数である。従って、式2.1 および2.2 によるり荷重値の増加修正量ΔＷ(+)は次式で表される。

$$\Delta W(+)=\varepsilon_{ox}/t_{ox}\mu W/L(Vc+(Vb+Vt-Vc)/r-Vth)(Vb+Vt-Vc)/r \qquad (3.10)$$

また、入力端子Ｄにパルス信号を与えた場合は、

$$\Delta Q(-)=C(Vt+Vc)/r \qquad (3.11)$$

$$\Delta W(-)=\varepsilon_{ox}/t_{ox}\mu W/L(Vc-(Vt+Vc)/r-Vth)(Vt+Vc)/r \qquad (3.12)$$

但しキャパシターＣ１の電圧Ｖcは、回路構成上次の範囲を超えることはできない。

$$-Vt \leqq Vc \leqq Vb+Vt \qquad (3.13)$$

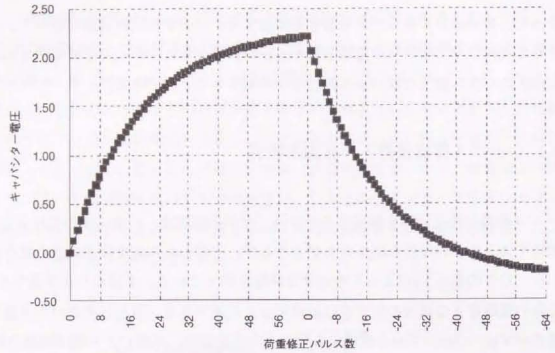


図3.4 荷重値修正回路特性（キャパシター電圧修正）

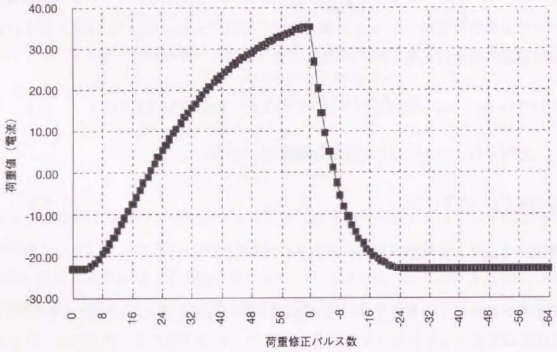


図3.5 荷重値修正回路特性（電流値修正）

図3.4 は、式3.9 と3.11 で表される本シナプス荷重修正回路の電荷量修正特性を示す。図3.5 は、式3.10 と3.12 で表される荷重値修正特性を示す。また、図3.6 には、第６章で述べるニューロチップで採用したシナプス回路についての、電荷量修正とシナプス荷重値修正に関するSPICEシミュレーション結果を示す。この回路シミュレーションでは回路上の浮遊容量や抵抗等をも考慮しているので、前述の近似表現による特性よりも、荷重値修正の感度がやや低下している。しかし本シナプス回路の荷重値修正特性は、式3.9 ～3.12 の近似表現によって十分に再現できることから、次節で述べる学習性能を評価した計算機シミュレーションでは、本シナプス回路の荷重値修正非線形特性をこれらの近似式によって表現している。

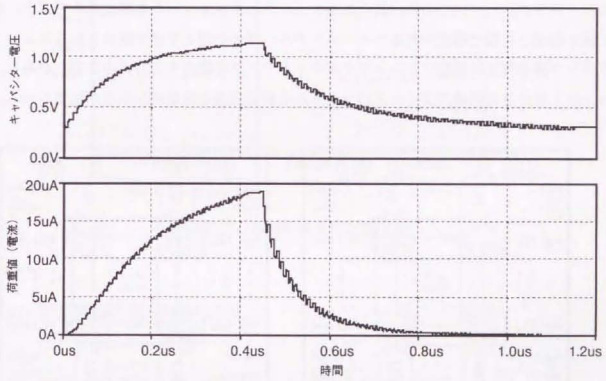
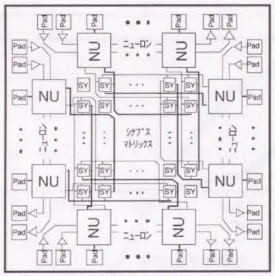


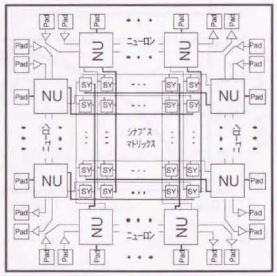
図3.6 シナプス荷重値修正特性

3.5 学習機能搭載ニューラルネットワークの回路構成と制御フロー

図3.7内に示すように、ニューロン回路はチップ外部との信号のやり取りを行う必要があるためチップの周辺部に配置され、シナプス回路はマトリックス状にチップの中央に配置され、相互に接続することでニューラルネットワークを構成する。ニューロン回路はシナプス回路より面積が大きいのでシナプス回路の２倍の高さにしてチップの４辺に配置することで、チップ内に集積するニューロン数とシナプス数の良好なバランスを保ちながら規則的な繰り返しレイアウトパターンを実現することができる。ニューロン回路とシナプス回路の接続方式は、図3.7内の左図に示す上辺と右辺あるいは左辺と下辺で構成されるニューロングループ内での自己結合型のニューラルネットワークを構成する方式と、右図に示す右辺と左辺で構成されるニューロングループと上辺と下辺で構成されるニューロングループ間を相互に接続するニューラルネットワークを構成する方式の２通りがある。ニューロン信号は実時間でフィードバックされるので高速な状態緩和処理が実現される。



自己結合ニューラルネットワーク
（２セット）



相互結合ニューラルネットワーク
（１セット）

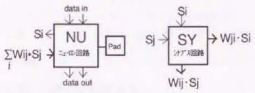


図3.7 ニューラルネットワーク回路構成

学習あるいは連想時には、各辺毎にニューロンへの属性あるいは教師データをシフトインし連想結果をシフトアウトする。学習制御信号は図3.8に示すような制御フローに従って、全てのニューロン回路またはシナプス回路へ共通に与えることで連想メモリーの学習（記憶）および連想が実行できる。

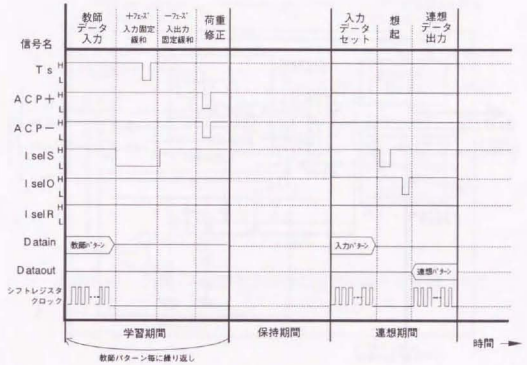


図3.8 各種制御信号制御フロー

3.6 学習性能と動作マージン

3.6.1 回路の簡略化に伴う学習ルール近似表現の影響

第3.4節で述べた通りシナプス回路において、式3.7 で示す学習ルールの簡略化や図3.3で示す非線形特性を持つ荷重修正回路を採用することは、簡単な回路構成で学習機能が実現でき、学習機能を搭載したニューラルネットワークLSIの高集積化に極めて有効である。一般に、線形な荷重修正回路は図3.9に示すようにフィードバック回路を設けることで実現することができるが、回路規模と消費電力が２倍以上に増大すると予想される。そこで、

高集積化に有効な学習ルールの簡略化や荷重修正回路の非線形特性などが学習あるいは連想性能へ及ぼす影響を調べておく必要がある。ここでは、素子の特性バラツキに加えて近似学習ルールと図3.5に示す荷重修正非線形特性の連想性能に及ぼす影響を計算機シミュレーションにより評価した結果について述べる。

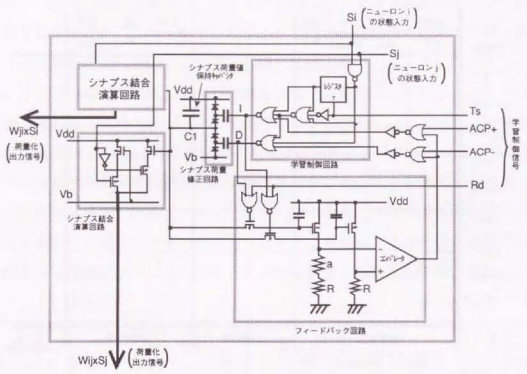


図3.9　リニアな荷重値修正特性を実現するシナプス回路

ここでのシミュレーションも、第２章で行ったのと同様に、50ニューロンのニューラルネットワークに第3.2節で述べた自己想起用の学習アルゴリズムで5個のパターンを学習して記憶した後、学習したパターンに対して生成したハミング距離毎に各々100通りの入力テストパターンによって自己想起した連想出力パターンの、期待学習パターンからのハミング距離の平均値によって連想性能を評価した。シナプス回路の荷重修正非線形特性は、式2.1 および式3.9 ～3.12をシミュレータの機能モデルに組み込むことで表現している。また素子特性バラツキは、シナプス荷重値Wと荷重修正量ΔWそしてニューロンのしきい値 V_θ に関して、予め各回路素子毎に乱数で生成した正規分布をとる固有のノイズ項を加えることで機能表現している。素子特性のバラツキは正規分布をとると仮定し、バラツキの程度は素子特性の平均値に対するバラツキ分布の標準偏差値の割合（%）で表わしている。

図3.10,図3.11,図3.12は、式3.7に従う近似学習ルールにおいて、素子特性バラツキが無く（0.0%）線形荷重修正と素子特性バラツキが18.8%で非線形（図3.5に示す）または線形の荷重修正による学習回数が70, 200, 500の場合を各々示している。

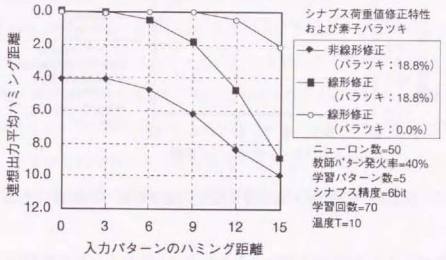


図3.10　シナプス荷重修正特性の連想性能への影響（その１）

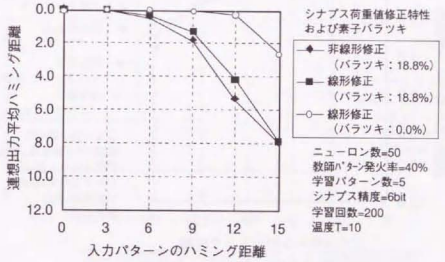


図3.11　シナプス荷重修正特性の連想性能への影響（その２）

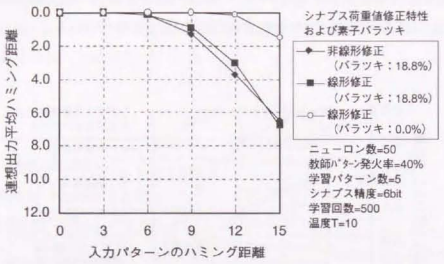


図3.12　シナプス荷重修正特性の連想性能への影響（その３）

これら温度パラメータTが10の場合、入力パターンのハミング距離が9以上になると素子特性バラツキが無い場合と18.8%の場合で連想性能に差が出るものの、学習が十分行われると荷重修正特性が線形と非線形の場合で連想性能に殆ど差が無くなることが分かった。

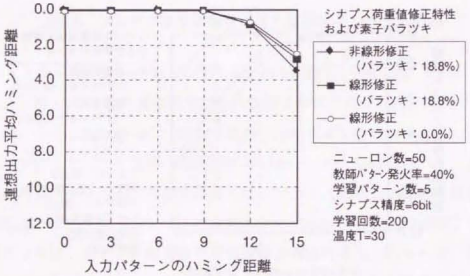


図3.13　シナプス荷重修正特性の連想性能への影響（その４）

図3.13では温度パラメータTを30と高くした場合の同様の評価結果を示している。温度パラメータTが高くなると素子特性バラツキの補償能力が高まることが分かる。

図3.14 は、学習過程における荷重修正ルールの違いによる連想能力の差異を示している。理想的な学習として、荷重修正ルールが$\Delta Wij=\eta\ (pij^{+}-pij^{-})$で素子特性バラツキが無い（0.0%）場合について評価した。そして簡略化した学習アルゴリズムについては、荷重修正ルールが$\Delta Wij=\eta\ (Si^{+}\cdot Sj^{+}-Si^{-}\cdot Sj^{-})$の場合と、$\Delta Wij(+)=\eta\ (Si^{+}\cdot Sj^{+})$と$\Delta Wij(-)=-\eta\ (Si^{-}\cdot Sj^{-})$とを時分割で実行する場合との２例について、素子特性バラツキが無い（0.0%）場合と18.8%の場合について、温度パラメータTが30の条件でシミュレーションを行った。$\Delta Wij(+)=\eta\ (Si^{+}\cdot Sj^{+})$と$\Delta Wij(-)=-\eta\ (Si^{-}\cdot Sj^{-})$とを時分割で実行する荷重修正ルールは、第４章で述べるニューロチップ[6]で採用したシナプス回路のように、学習回路をより簡単にすることができる。

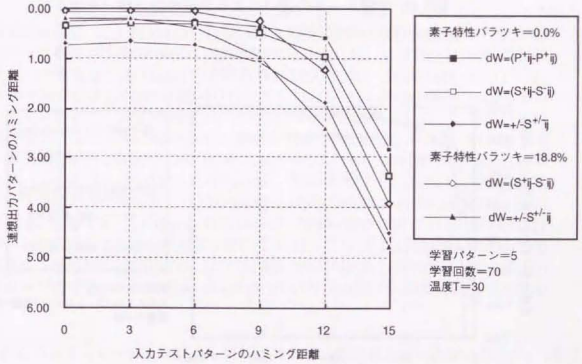


図3.14 学習ルールの違いによる連想性能比較

これらの評価により、荷重修正ルールが$\Delta Wij=\eta\ (Si^{+}\cdot Sj^{+}-Si^{-}\cdot Sj^{-})$の場合では素子特性バラツキが18.8% においても十分良好な理想の学習に近い性能を得れること

が分かった。また、荷重修正ルールが$\Delta Wij(+/-)=\eta\ (Si^{+/-}\cdot Sj^{+/-})$の場合ではやや性能が低下する。

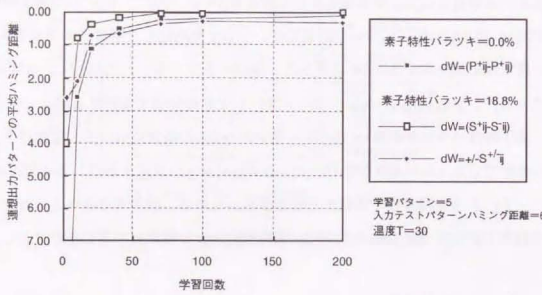


図3.15 学習ルールの違いによる学習性能比較（その１）

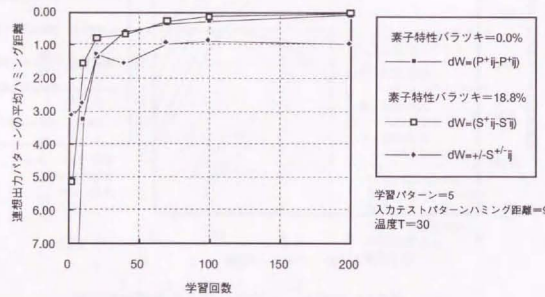


図3.16 学習ルールの違いによる学習性能比較（その２）

図3.15 と図3.16 は、各々の荷重値修正ルールによる学習において、荷重値修正操作の繰り返し回数による変化を示している。この評価によって学習の収束特性が分かる。入力テストパターンのハミング距離が6の場合を図3.15 に、ハミング距離が9の場合を図3.16 に示す。これらの評価により、荷重修正ルールが$\Delta Wij=\eta\ (Si^{+}\cdot Sj^{+}-Si^{-}\cdot Sj^{-})$の場合では、素子特性バラツキが18.8% にもかかわらず理想的な学習よりも収束が速いことが分かる。また、荷重修正ルールが$\Delta Wij(+/-)=\eta\ (Si^{+/-}\cdot Sj^{+/-})$の場合では、学習過程で認識率が振動し、収束する認識率は入力パターンのハミング距離が大きくなるほど悪くなる性質を示している。

次に、学習するパターン数による学習性能の違いを、理想学習と荷重修正ルールが$\Delta Wij=\eta\ (Si^{+}\cdot Sj^{+}-Si^{-}\cdot Sj^{-})$で素子特性バラツキが18.8% の場合について、入力パターンのハミング距離が6、9、12の場合について、各々図3.17 、図3.18 、図3.19に示す。ここで学習に用いたパターンは次の通りである。

```
ニューロン番号（12345678910.........20.........30.........40.........50）
パターン番号 1：1111111111 1111111111 0000000000 0000000000 0000000000
パターン番号 2：0000000000 0000000000 1111111111 1111111111 0000000000
パターン番号 3：1111111111 0000000000 0000000000 0000000000 1111111111
パターン番号 4：0000000000 1111111111 1111111111 0000000000 0000000000
パターン番号 5：0000000000 0000000000 0000000000 1111111111 1111111111
パターン番号 6：1111000000 1111000000 1111000000 1111000000 1111000000
パターン番号 7：0000111100 0000111100 0000111100 0000111100 0000111100
パターン番号 8：1100000011 1100000011 1100000011 1100000011 1100000011
パターン番号 9：0000011111 0000011111 0000000000 1111100000 1111100000
パターン番号10：1111100000 1111100000 0000000000 0000011111 0000011111
パターン番号11：0000000000 0000011111 1111111111 1111100000 0000000000
パターン番号12：1010101000 1010101000 1010101000 1010101000 1010101000
パターン番号13：0101010100 0101010100 0101010100 0101010100 0101010100
```

これらのシミュレーションによる評価によって、荷重修正ルールが$\Delta Wij=\eta\ (Si^{+}\cdot Sj^{+}-Si^{-}\cdot Sj^{-})$の場合、学習するパターンの数が増えるにしたがって牽引皿の大きさが、理想的学習に比べて小さくなる性質があることが分かった。また、チップ内素子の特性バラツキによる演算特性バラツキに加えて、高集積化のためにシナプス回路に採用した簡略化学習ルールや非線形特性を持つ荷重修正回路による理論的な学習アルゴリズムか

らのズレをも、チップ上に実装した学習機能により十分に自己補償されることが明らかになった。

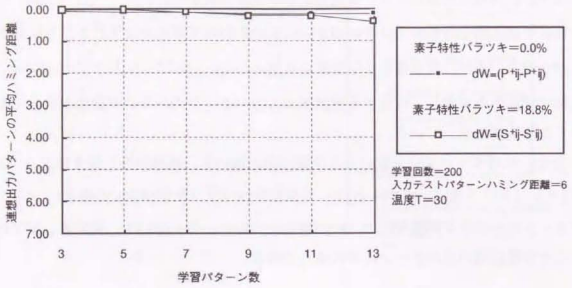


図3.17 学習能力の学習パターン数依存（その１）

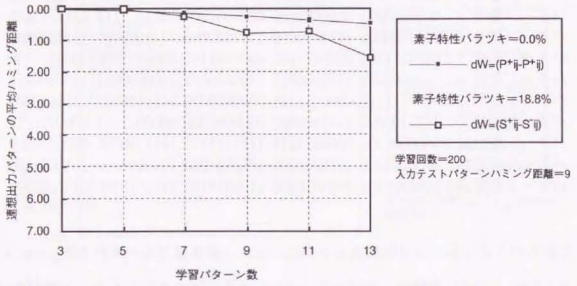


図3.18 学習能力の学習パターン数依存（その２）

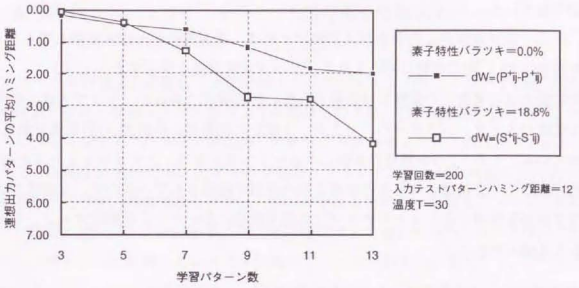


図3.19 学習能力の学習パターン数依存（その３）

3.6.2 電源電圧変動に対するシナプス回路動作マージン

ここで、アナログ回路における克服すべき課題の一つである、電源電圧の変動に対する動作マージンについて評価する。従来のアナログ回路方式ニューロチップにおける動作マージンの問題は、電源電圧の変動に伴うシナプス荷重値の変動が極めて大きいことに起因している。従来のアナログ方式シナプス回路は、シナプス荷重値の設定精度を確保する目的で大きなダイナミックレンジの線形特性領域を好んで使い、トランジスターは非飽和領域で動作させていた。その結果シナプス荷重値は電源電圧や環境温度の変動に対して強い感度を示し、従来のアナログ回路方式ニューロチップの動作マージンは実用には不十分であった。

この問題を回避する為に、第２章で述べたとおり、シナプス荷重値を表現するトランジスターは飽和領域で動作する回路構成（図2.11）を採用している。その効果についてSPICEシミュレーションにより評価した結果を図3.20 に示す。比較のために従来のアナログ方式シナプス回路[8]の場合も示している。従来のアナログ回路によれば、電源電圧の±10% 変動に対してシナプス荷重値は40% 近い変動を生じており、第２章で示した素子

特性バラツキが及ぼす連想性能への影響評価結果から、極めて大きな性能低下を招くことが予想される。

一方、我々が採用したシナプス回路によれば、電源電圧が±10% 変動してもシナプス荷重値は3% 以内の変動に抑えられる。シナプス荷重値を保持するキャパシターＣ１の蓄積電荷量は電源電圧の変動には影響されない回路構成であるが、シナプス結合演算回路内の電流パストランジスターのバックゲート電圧の変動がこの約3% の荷重値変動を生じさせている。しかし、3% 程度の変動はニューラルネットワークダイナミックスの構造安定性によって十分補償されることが第２章の評価で確認されているので、本回路方式のアナログ回路を採用したニューロチップは実使用環境において十分な動作マージンを確保できると見積もられる。

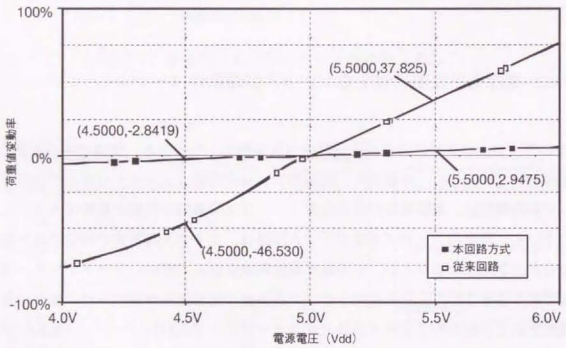


図3.20 シナプス回路動作マージン比較

3.7 連想記憶ニューラルネットワークLSIの素子微細化トレンド

本章で述べた学習機能を搭載した連想記憶アナログニューラルネットワークLSIに関して素子の微細化に伴う集積シナプス数および演算速度の進展を図3.21と図3.22に示す。比較のために、第２章で述べたデジタル保持・アナログ演算方式とデジタル保持・デジタル演算方式による場合と図3.9で示したアナログ保持・アナログ演算、線形荷重修正方式の場合とを合わせて示す。この見積もりでは、図2.12で示した素子の微細化トレンドに従い、またチップ面積が図3.21の上図に示すように微細化に伴って増大するバイルール[9]を仮定した。そしてチップ内に集積するシナプス回路はニューロン回路の２乗とした。バイナリデジタル回路によるニューロ連想メモリーLSIの場合はシナプス回路あたりの時分割表現シナプス数が100と1000の場合を示した。デジタル回路方式の場合、ニューロン一つあたりに接続されたシナプス数の増加に伴い加算bit幅が増加するのでチップ内に集積できるシナプス回路の数は、スケーリング係数をKとした場合、K^2に比例し、単位処理の遅延時間はKによらず一定となる。その結果、処理速度CPS（Connections Per Seconds）はK^2に比例し、0.15μmレベルで約260ギガCPSの処理速度に到達すると予想される。一方、本章で述べた学習機能を搭載した連想記憶アナログニューラルネットワークLSIの場合は、チップ内に集積できるシナプス回路の数はK^3に比例し、単位処理の遅延時間は$K^{0.5}$に比例して遅くなる結果、処理速度は$K^{2.5}$に比例し、素子の微細化が0.15μmレベルで、約2000万シナプスを集積し約200テラCPSの処理速度に到達することが見積もられた。また、それぞれの方式とも微細化に伴う電力密度は一定であるため、素子の微細化とチップ面積の増大による放熱限界に関する制限は無いものと考えられる。

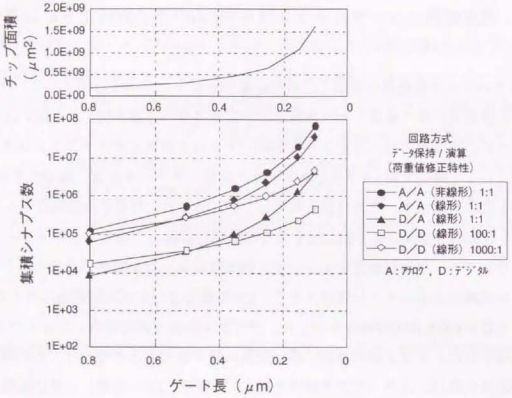


図3.21　各種ニューロ連想メモリーの微細化トレンド（その1）

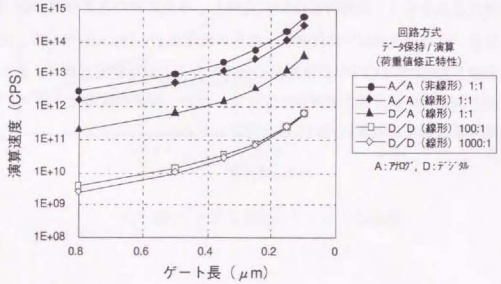


図3.22　各種ニューロ連想メモリーの微細化トレンド（その2）

3.8　まとめ

チップ上に学習機能を実装することにより実現される自己組織化機能は、チップ内素子の特性バラツキに加えて、高集積化のためにシナプス回路に採用した近似学習ルールや非線形特性を持つ荷重修正回路による学習機能表現上の制限をも、十分に補償できる能力を示すことが計算機シミュレーションによる評価によって明らかになった。また、シナプス荷重値を表現するトランジスターを飽和領域で動作させる回路構成を採用したことによって、電源電圧が±10%変動してもシナプス荷重値は3%以内の変動に抑えられることがSPICE [1]シミュレーションによって確認された。これら提案した学習回路構成によってアナログニューラルネットワークLSIは、学習機能をチップ上に備えながら高い集積度を実現でき、しかもその自己補償能力により高い連想性能と実使用環境において十分な動作マージンを確保でき、素子の微細化に伴って0.15μmレベルでは、1チップに数千万シナプスを集積して数百テラCPSを超える性能に達することが見積もられた。

参考文献

[1] L. W. Nagel, "A Computer Program to Simulate Semiconductor Circuits," U.C.Berkeley ERL Memo, No. ERL-M75/520, May 1975.

[2] G. E. Hinton, and T. J. Sejnowski, "Optimal Perceptual Inference," Prc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 448-453, 1983.

[3] M. A. Cohen, and S. Grossberg, "Absolute Stability of Global Pattern Formations and Parallel Memory Storage by Competitive Neural Networks," IEEE, Trans., System, Man, and Cybernetics, 13, pp.815-826, 1983.

[4] D H.Ackley, G E.Hinton, and T J.Sejnowski, "A Learning Algorithm for Boltzmann Machines," Cognitive Science, Vol.9, No.1, pp.147-169, Jan-Mar, 1985.

[5] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," IEEE, Journal of Solid-State Circuits, Vol.26, No.4, pp. 607-611, April, 1991.

[6] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," IEEE, Journal of Solid- State Circuits, Vol.26, No.11, pp. 1637-1644, Nov.,1991.

[7] Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," IEEE, Journal of Solid-State Circuits, Vol.27, No.12, pp.1854-1861, Dec.,1992.

[8] M. Holler, S. Tam, H. Castro, and R. Benson," An electricaly trai-nable artificial neural network (ETANN) with 10240 Floating gate synapses," Proc. of IJCNN-89, Vol.2, pp.191-196, 1989.

[9] Y. Tarui, and T. Tarui," New DRAM Princing Trends : The Bi Rule," IEEE Circuits & Devices, Vol.7, No.2, pp.44-45, March 1991.

第４章
学習機能搭載ニューロチップ

4.1 序

本章では、第３章で述べた回路構成に基づき実際に試作した３種類の、学習機能を備えたニューロチップについて述べる。表4.1には、試作した３つのニューロチップの諸元を示す。最初に、125ニューロンと10Kシナプスを集積したニューロチップ（NEURO1）を1.0μmCMOSプロセス技術を用いて1989年に試作し、1990年6月に発表[1]した。翌1991年2月には、336ニューロン、28Kシナプスと集積度を高めるとに加えて、ニューロチップ同士を接続してニューラルネットワークの規模を拡張できるマルチチップ拡張機能を実装したニューロチップ（NEURO2）を試作発表[2]した。更に、1992年2月には、0.8μmCMOSプロセス技術を用いて、400ニューロン、40Kシナプスを集積し、新たにシナプス荷重値の高速リフレッシュ機能を搭載したニューロチップ（NEURO3）を試作発表[3]した。

表4.1 試作したアナログニューロチップの諸元

チップ名称	NEURO1 [1]	NEURO2 [2]	NEURO3 [3]
ニューロン数	125	336	400
シナプス数	10,000	28,224	40,000
シナプス結合数	20,000	56,448	80,000
演算精度	6-bit	6-bit	6-bit
演算スピード	60-GCPS	1.1-TCPS	2-TCPS
学習スピード	4-GCUPS	28-GCUPS	80-GCUPS
消費電力	1.5-W (Max.)	3.0-W (Max.)	4.5-W (Max.)
チップサイズ	13.0mm×13.0mm	14.5mm×14.5mm	14.5mm×14.5mm
使用プロセス	1.0μmCMOS	1.0μmCMOS	0.8μmCMOS
搭載機能	ｵﾝﾁｯﾌﾟ学習機能	ｵﾝﾁｯﾌﾟ学習機能 ﾏﾙﾁﾁｯﾌﾟ拡張機能	ｵﾝﾁｯﾌﾟ学習機能 ﾏﾙﾁﾁｯﾌﾟ拡張機能 荷重値ﾘﾌﾚｯｼｭ機能
発表年・学会	'90 Symp. on VLSI Circuits	ISSCC '91 IJCNN '93	ISSCC '92

CPS:Connections Per Second
CUPS:Connections Update Per Second
ISSCC:International Solid-State Circuits Conference
IJCNN:International Joint Conference on Neural Networks

第4.2節、第4.3節および、第4.4節において、これら３つのニューロチップについて各々の回路構成概要を述べ、第4.5節において実チップによる学習機能評価の結果について述べる。但し、NEURO2に実装したマルチチップ拡張機能および、NEURO3の荷重値リフレッシュ機能に関しては第５章と第６章で各々詳しく述べることとし、本章においては、各々のチップにおけるオンチップ学習機能に関する内容のみを述べるに止める。

4.2 125ニューロン・10Kシナプス集積ニューロチップ（NEURO1）

本チップは、第３章で述べた、学習機能をチップ上に実装して回路素子のバラツキや非線形特性を補償するニューラルネットワークの自己組織化機能に着目して、回路構成を大胆に簡略化して高集積化を図る設計思想に基づいて設計試作した、最初のニューロチップ[1][4][5]である。従って、冗長な機能や最適化されていないパラメータが多くの部位で残されている。例えば、シナプス回路の荷重値リセット機能やニューロン状態のモニター専用出力バッファー機能は、実用上冗長であることが分かり、次のNEURO2チップ以降ではそれらの機能は削除された。また、ニューロンの入力ノードの容量および抵抗値を自由に調整できるようにチップの外部に接続する方式を採用したが、NEURO2チップ以降では抵抗をチップ内に作り込み、キャパシタや抵抗器の外付けを不要とした。このように、最初に試作したNEURO1チップは、後のニューロチップの設計最適化に多くの有益な情報を与えた。

続く4.2.1節と4.2.2節で、NEURO1チップを構成するシナプス回路とニューロン回路について述べ、4.2.3節でチップ構成について述べる。

4.2.1 シナプス回路

図4.1はNEURO1チップで採用したシナプス回路を示す。このシナプス回路は、対称なシナプス荷重値Wij＝Wji保持とその双方向シナプス結合演算機能Wij×SjとWji×Si、そしてシナプス荷重値修正機能±ΔW＝±η×Si×Sjを備えている。この回路構成によって、学習機能を備えた一つの双方向シナプスを表現することができる。この回路には、二つのニューロン状態信号SiとSjが入力され、回路内部で保持されているシナプス荷重値

Wij＝Wjiとの積算値Wij×SjとWji×Siに対応する二つの荷重化電流Ioが常時出力される。また、全シナプス回路に共通に与えられる学習制御信号Red, Acp, C+/-により、シナプス荷重値は学習則に従って修正される。

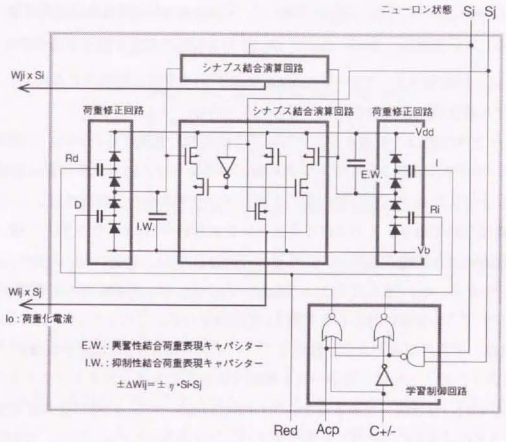


図4.1 シナプス回路（NEURO1）

図中のI.W.とE.W.は、各々、抑制性と興奮性のシナプス荷重を表現するキャパシターで、シナプス荷重値は、各々のキャパシターに蓄積される電荷量によって次のように表現される。

$$W_{+}=1/2\cdot\varepsilon ox/t ox\cdot\mu\cdot W/L\cdot(Q_{EW}/C_{EW}-Vth)^{2} \qquad (4.1)$$

但し、$Q_{EW}/C_{EW}<Vth$ の場合は $W_{+}=0$

$$W_{-}=1/2\cdot\varepsilon ox/t ox\cdot\mu\cdot W/L\cdot(Vb-Q_{IW}/C_{IW}-Vth)^{2} \qquad (4.2)$$

但し、$(Vb-Q_{I}/C_{I})<Vth$ の場合は $W_{-}=0$

$$Wij = Wji = W_{+} + W_{-} - Wb \qquad (4.3)$$

ここで、L＝1.0μm、W＝4.5μm、$C_{EW}=C_{IW}$＝0.5pF、Vth＝0.7V、Vb＝4～3V、Wb＝1/2・εox/tox・μ・W/L・$(Vb-Vth)^2$、Vt＝～0.4Vは荷重修正回路を構成するダイオードのしきい値電圧。また、Q_{EW}/C_{EW}またはQ_{IW}/C_{IW}で表される各キャパシターの電圧Vcは、回路構成上、Vb－Vt≦Vc≦Vdd＋Vtの範囲に制限されるので、Vbによってシナプス荷重値の最大値を設定することができる。

このシナプス回路は、対称なシナプス結合を効率良く表現するために、一組のI.W.とE.Wのシナプス荷重値表現キャパシターを二つのシナプス結合演算回路に接続して、双方向のシナプス結合を表現している。これらのシナプス結合演算回路は、三つの定電流パスから構成されており、入力されるニューロン状態信号Sj（またはSi）に従って外側の二つの電流パスと中央の電流パスが相補的に活性化する。すなわち、入力ニューロン状態が"発火"つまり、Sj（またはSi）＝"High"レベル時には、外側の二つの電流パスが活性化し、シナプス荷重値に対応した電流Ioが出力される。また、Sj（またはSi）＝"Low"時には、中央の電流パスが活性化して、バイアス電圧Vbで規定されるバイアス電流Wbが出力される。三つの定電流パスを構成する全てのトランジスターサイズは、L＝1.0μm、W＝4.5μmと同一であるので、Sj（またはSi）＝"Low"時に流れる電流は、Wij×Sj＝0（またはWji×Si＝0）を意味するバイアス値Wbとなる。即ち、このバイアス値より少ない電流値は負の荷重値を表し、多い電流は正の荷重値を表している。

I.W.とE.Wは各々、荷重修正回路に接続されており、それらの荷重修正回路は、学習制御回路から出力される制御信号を受けて各荷重値を修正する構成になっている。荷重修正回路は、直列に接続された二つのチャージポンプ回路から構成される。上方の入力端子（Rd，I）にパルス信号が与えられると、I.W.とE.Wに蓄えられた正電荷は、パルス数に比例して汲み出され減少する。また、下方の入力端子（D，Ri）にパルス信号が与えられると、正電荷がパルス数に比例して注入され、増加する。つまり、Rd，Riに与えるパルス信号は、各々の荷重値の絶対値を減少させ、D，Iに与えるパルス信号は、各々の絶対値を増加させる。各荷重修正回路内のキャパシタは0.05pFであり、一回のパルス信号で修正される電荷量は、飽和量の十分の一以下になる。

簡単な論理ゲートで構成される学習制御回路は、全シナプス回路に共通の学習制御信

号AcpとC+/-、それと二つのニューロン状態信号SiとSjによって、荷重修正回路へ与えるパルス信号DとIを生成する。信号C+/-は学習フェーズを規定し、"High"レベル時に＋学習フェーズを、"Low"レベル時に－学習フェーズを表す。一方、信号Acpに与えるパルス信号は、表4.2に示すように、SiとSjが共に活性状態"High"のときには、＋学習フェーズ時にAcpに与えるパルス信号は反転してIノードに伝えられ、－学習フェーズ時には反転パルス信号がDノードに伝えられる。

表4.2 シナプス荷重値修正信号（NEURO1）

学習フェーズ	Si·Sj	D	I
＋ C+/- = H	H	L	$\overline{Acp}$
	L	L	L
－ C+/- = L	H	$\overline{Acp}$	L
	L	L	L

従って、学習パターン毎に＋と－の学習フェーズを実行すれば、この学習制御回路は次式に従う荷重修正信号を発生することができる。

$$\triangle Wij = \kappa\ (Si^{+} \cdot Sj^{+} - Si^{-} \cdot Sj^{-}) \qquad (4.4)$$

ここでκは正値の係数で、その大きさはAcpに与えるパルス信号の数またはパルスの幅で表現することができる。また、SiとSjの肩に付いた＋／－は各学習フェーズを示す。

この荷重値修正規則はHebb則と反Hebb則とを組み合せた学習則であり、第２章で述べた連想メモリーの銘記手続きとしての、ボルツマンマシンの学習則[6]の修正近似則を実現している。荷重修正回路のRdとRiに直接伝えられるRedにパルス信号を与えることで各荷重の絶対値は減少し、パルス信号を連続して多数与えると荷重値をリセットすることができる。また、非学習時にはAcpを"High"レベルに固定しておく必要がある。

図4.2は試作したNEURO1チップ上のシナプス回路を顕微鏡で拡大した写真である。一つのシナプス回路の占有面積は100μm×100μmで、下部約三分の一にシナプス荷重値保持用のキャパシター、中央に荷重修正回路、上部中央に学習制御回路、上部の両側にシナプス結合演算回路が配置されている。各ニューロンの状態（軸策）信号とシナプス結合（樹状突起）信号は、各々のニューロンに対応した入出力に応じて、つまりSjとWji×SiまたはSiとWij×Sjのように、平行に配線することができチップの周辺に配置されている各ニ

ューロン回路へ接続されている。このようにシナプス回路はメモリーセルと同様に、整然とマトリックス状に配置することができる。

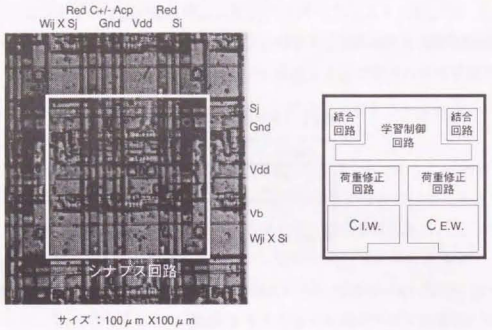


図4.2 シナプス回路（NEURO1）写真と機能構成

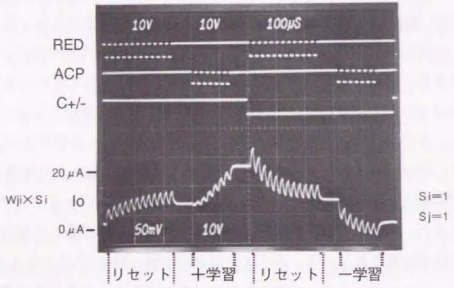


図4.3 シナプス波形（NEURO1）写真

図4.3は、シナプス回路の荷重値修正基本動作を示す実測波形写真である。これは、負荷抵抗５ＫΩを接続時のＳｉとＳｊが共に活性"High"時における学習制御信号と荷重化電流出力を示している。＋学習フェーズ時（Ｃ+/-＝"High"）にＡcpパルスによってシナプス荷重が増加し、－学習フェーズ時（Ｃ+/-＝"Low"）には、減少していることが確認された。出力波形上にＡcpまたはＲedパルス信号が容量結合によって伝搬しているのが観測されているが、一つの出力ノードに100以上のシナプス回路が並列に接続されている、ニューラルネットワーク構成時には、出力ノードの容量は相対的に極めて高くなるので、この種のノイズ振幅は極めて小さくなると考えられる。実際のニューロチップ上での学習評価において、目立った不具合は観測されなかった。

4.2.2 ニューロン回路

図4.4はNEURO1チップで採用したニューロン回路を示す。接続されている全てのシナプス回路からの荷重化電流出力は共通ノードで足し合わされ、ΣWij×Ｓjの電流はPADを介してチップ外部で接続された負荷抵抗により電圧に変換されて、コンパレータCompの入力端子に与えられる。

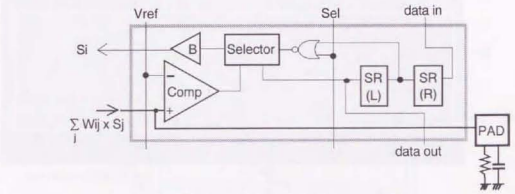


図4.4 ニューロン回路（NEURO1）

Compの他方の入力端子には、チップ外部から比較電圧Ｖrefが与えられ、本ニューロンのしきい値を表現している。従ってCompの出力は、本ニューロンへ入力される荷重化信号の合計である内部活性値が、しきい値Ｖrefを超えているか否かを示す状態信号を表す。右側のシフトレジスタＳＲ(R)には本ニューロンの属性を規定するデータが格納され、左側のシフトレジスタＳＲ(L)には学習時に教師データ、想起時に入力データが格納される。

ＳＲ(R)の入力には前のニューロンのＳＲ(L)の出力が接続されており、各ニューロンの属性データ及び教師データまたは入力データをチップ外部より交互にシフトインすることによって各ニューロン毎に所望の値を設定することができる。ニューロンの状態出力Ｓｉは、セレクタによって内部活性値による状態値かＳＲ(L)に格納された教師データ（想起時は入力データ）かを選択して出力される。そのセレクターの選択制御信号はNORゲートを介してＳＲ(R)信号とＳel信号を接続されている。この回路構成によってニューロン回路は、各ニューロン毎にＳＲ(R)とＳＲ(L)に格納するデータによって、各属性毎に異なった学習制御を実現している。Ｓel信号は学習フェーズを規定すると共に、想起時の可視ニューロンの初期状態をＳＲ(L)に格納されたデータで固定する期間を規定する。

表4.3 ニューロン状態出力（NEURO1）

ニューロン属性		入力ニューロン	出力ニューロン	隠れニューロン
SR (R) データ		L (Sel=L)	L	H
学習フェーズ	＋	SR (L) データ	SR (L) データ	内部活性値
	－	SR (L) データ	内部活性値	内部活性値

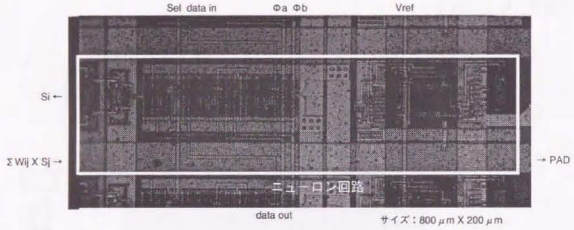


図4.5 ニューロン回路（NEURO1）写真

表4.3は各学習フェーズにおけるニューロン状態出力Ｓｉの選択内容を示す。ＳＲ(R)に格納された属性データが"High"の時は常に内部活性値による状態がＳｉとして出力され隠れニューロンとして機能する。また、属性データが"Low"の時は＋学習フェーズで教師データが、－学習フェーズで内部活性値による状態が出力され、出力ニューロンとして機能す

る。そして、属性データが"Low"でしかもＳel信号を常に"Low"に固定することで、常に教師データが出力される入力ニューロンとすることができる。

図4.5は試作したNEURO1チップ上のニューロン回路を顕微鏡で拡大した写真である。回路面積は800μm×200μmであり、右側にコンパレータ、中央にセレクター、NOR、シフトレジスタＳＲ(R)、ＳＲ(L)、左側に出力バッファーを配置している。

4.2.3 チップ構成

図4.6にNEURO1チップのブロック構成図を示す。チップ中央にシナプス回路をマトリックス状に50行50列を四つ配置している。シナプスマトリックスの右側と下側に25個ずつのニューロン回路ブロックをそれぞれ二つずつ配置している。このニューロンブロック単位にＳel信号とＶref信号ノードが共通に接続されている。シナプスマトリックスの左側には25個の入力専用ニューロンが配置されており、マトリックスの上側には、各ニューロンの状態をモニターするための、100-to-50のセレクタを介して50個の出力バッファーを配置している。各ニューロン回路とシナプス回路間の配線接続は、125ニューロンの完全フィードバック結合（入力ニューロン間の結合はない）のニューラルネットワークを構成した。

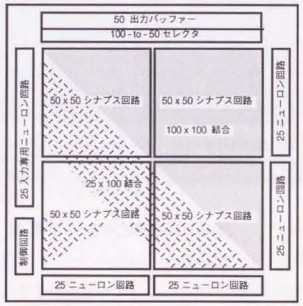
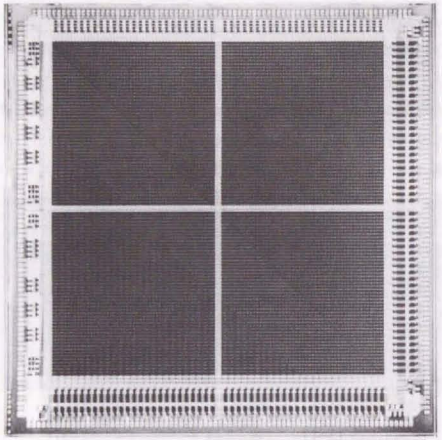


図4.6 NEURO1チップ構成

図4.7はNEURO1チップの顕微鏡拡大写真である。1.0μmCMOS、二層ポリSi、二層金属配線プロセス技術で試作した。各シナプスセルに二つのシナプス結合演算回路があるので、本チップに集積された10K個のシナプス回路は20K個の対称なシナプス結合を実現できる。従って、今回試作したチップでは2,500個のシナプス回路が未使用である。チップサイズは13.0mm×13.0mmで、281ピンのセラミックPGAパッケージに格納できる。消費電力は最大で1.5Wである。



チップサイズ：13.0mmX13.0 mm

図4.7 NEURO1チップ写真

4.3 336ニューロン・28Kシナプス集積ニューロチップ（NEURO2）

一チップに集積できるニューロンとシナプスの規模は半導体集積回路の微細化が進むに連れて増大することが期待できるが、通常のメモリーデバイスと違い、ニューロ連想メモリーデバイスではニューラルネットワークの規模を拡張する為の新たな技術が必要となる。通常のメモリーデバイスにおいては、複数のメモリーデバイスを拡張したい方向に応じて、アドレス信号を共通にしてデータI/O信号を並列化するか、データI/O信号を共通にしてアドレス信号を拡張するかの信号接続の違いによって、ワードビット方向の拡張あるいはワード深さ方向の拡張共に容易に実現することができる。しかしニューロ連想メモリーデバイスにおいては、これら二種類の規模拡張の内、ワードビット方向の拡張を実現するためにニューラルネットワークの規模拡張が必要になることから、新たな拡張技術が必要になる。

そこで、フィードバック全結合構造を保ったニューラルネットワークの規模拡張を効率良く実現できる、BNU（Branch-Neuron-Unit）アーキテクチャを新たに考案し、NEURO2チップで採用した[2][7][8]。このBNUアーキテクチャによるマルチチップ拡張機能は、同一ニューロンの機能を複数のチップに分散配置する構成によって、拡張接続するチップ数に依らず全ての動作速度は一定に保たれ、拡張接続したニューロチップの数に比例した速度性能向上を実現する。

BNUアーキテクチャおよびマルチチップ拡張機能に関しては次の第５章で詳しく述べることにし、本章ではNEURO2チップのオンチップ学習機能に関する回路構成を中心に述べるに止める。

4.3.1 シナプス回路

NEURO2チップで採用したシナプス回路は、NEURO1チップと同一のプロセス技術を採用しながらも、更なる高集積化を図っている。シナプス回路の高集積化のためにNEURO2チップからはシナプス荷重値を保持するキャパシターを一つに減らし、また荷重値のリセット機能を削除した。図4.8にNEURO2チップで採用したシナプス回路を示す。

シナプス荷重値はキャパシターＣ1に蓄えられた電荷量で表現し、二つのシナプス結合演算回路がＣ１の蓄積電荷量に比例して生じる電圧に規定される電流量、すなわち荷重化電流、Wij×Ｓj、Wji×Ｓiを出力する。シナプス結合演算回路はシナプス荷重値保持用キャパシターを一つにしたことに伴い、キャパシターＣ1の電圧で制御される電流パス回路の最大電流量をバイアス用電流パス回路の２倍にするために、MOSトランジスターのゲート幅Wを5.0μmにし、バイアス用回路のW＝2.5μmの２倍にしている。Ｃ1の蓄積電荷量は荷重修正回路によって修正され、荷重修正回路は学習則に従って荷重値を修正するように学習制御回路により制御される。全シナプス回路に共通な学習制御信号ＡCP+とＡCP-は、非学習時には共に"High"レベルに固定されており、学習時には＋フェーズではＡCP+に、－フェーズではＡCP-にそれぞれ負のパルス信号が与えられる。各学習パターン毎にこの信号操作によって、各シナプス荷重値は次の量だけ修正される。

$$\triangle Wij = \kappa\ (Si^{+} \cdot Sj^{+} - Si^{-} \cdot Sj^{-}) \qquad (4.5)$$

ここでκは正値の係数で、その大きさはＡcp+/-に与えるパルス信号の数またはパルスの幅で表現することができるが、第３章で述べた通り、シナプス荷重値によって非線形な特性を示す。Ｃ1の容量は0.6pF、荷重修正回路内のキャパシターの容量は0.03pFである。

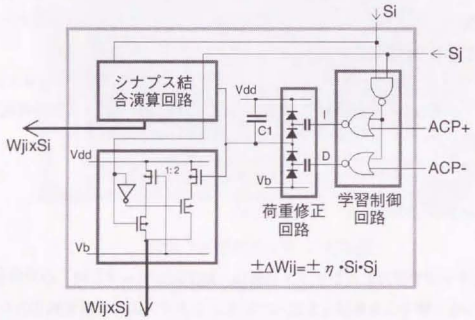


図4.8 シナプス回路（NEURO2）

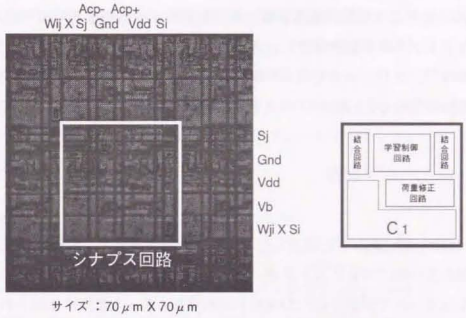


図4.9 シナプス回路（NEURO2）写真と機能構成

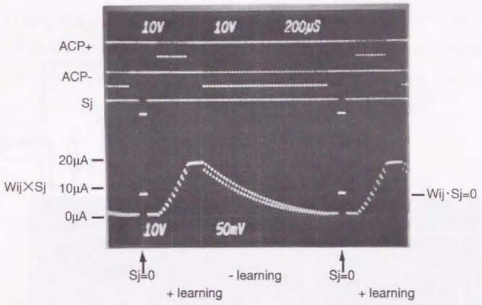


図4.10 シナプス波形（NEURO2）写真

図4.9は試作したNEURO2チップ上のシナプス回路を顕微鏡で拡大した写真である。一つのシナプス回路の占有面積は70μm×70μmで、下部約三分の一にシナプス荷重値保持用のキャパシター、中央に荷重修正回路、上部中央に学習制御回路、上部の両側にシナプ

ス結合演算回路が配置されている。

図4.10は、シナプス回路の荷重値修正基本動作を示す実測波形写真である。Acp+パルスによってシナプス荷重値が増加し、Acp-パルスによってシナプス荷重値が減少することが確認された。また、ニューロンの状態入力Sjが"Low"の時には、最大のシナプス荷重値のほぼ半分に当たる8μAのバイアス電流が流れていることも確認できる。

4.3.2 ニューロン回路

図4.11はNEURO2チップで採用したニューロン回路を示す。シナプス回路からの出力荷重化電流が太い線で示した入力共通ノードで足し合わされΣWij×Sjとなり、チップ内に第二ポリシリコンで形成された2.0KΩの抵抗RLによって電圧に変換されて、コンパレータCompに入力される。Compの他方の入力端子にはチップ外部から基準電圧Vrefが与えられ、ニューロンのしきい値を表現する。

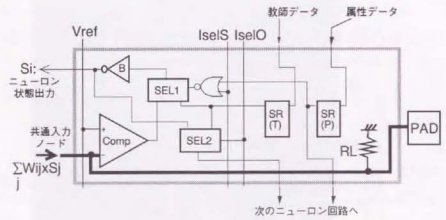


図4.11 ニューロン回路（NEURO2）

Compの出力はニューロンの内部活性値に対応した状態を表す。シフトレジスターＳＲ(T)には本ニューロンの教師データまたは想起用入力データが格納され、ＳＲ(P)には本ニューロンの属性データが格納される。NORゲートとセレクターSEL1は、本ニューロンの状態出力を制御する。即ち、ＳＲ(P)に格納された属性データと制御信号ＩselSとに従って、表4.4に示すように、本ニューロンの状態出力Siには、内部活性値による状態かＳＲ(T)

に格納されたデータかが出力される。この構成によって、各ニューロン毎に任意の属性と教師データを設定できる。なお、ＩselSは42ニューロン毎に共通で、この単位で入力か出力ニューロンかが区別できる。教師パターンを各ニューロンのＳＲ(T)に格納した後、シナプスの学習制御信号Acp+, Acp-とニューロンの制御信号ＩselSを、学習フェーズに従って操作することによって、ニューロチップは与えられた教師パターンを学習することができる。教師パターンを変えて、この操作を繰り返すことによって、ニューロチップの記憶は全ての教師パターンに関して深められる。

表4.4 ニューロンの属性と出力状態（NEURO2）

ニューロンの属性	学習フェーズ	SR(P)	IselS	ニューロンの状態出力
隠れ	+ -	H	L H	内部活性値 内部活性値
出力	+ -	L	L H	SR(T) 反転データ 内部活性値
入力	+ -	L	L L	SR(T) 反転データ SR(T) 反転データ

SEL2信号は各ニューロンの状態をモニターするのに使われる。全てのニューロン状態は制御信号ＩselOに従って、同時に次のニューロン回路のＳＲ(T)に格納された後、通常のシフトアウト動作によってチップ外部に出力することができる。太線に接続された信号端子PADは、ニューロンの内部活性値モニター用あるいは、マルチチップ拡張時にチップ間接続端子として用いられる。

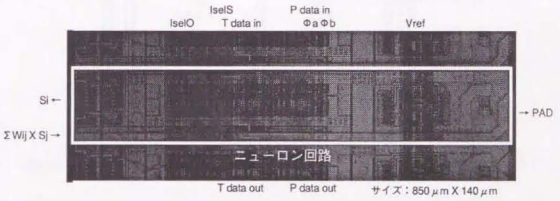


図4.12 ニューロン回路（NEURO2）写真

図4.12は試作したNEURO2チップ上のニューロン回路を顕微鏡で拡大した写真である。

回路面積は850μm×140μmであり、右側にコンパレータ、中央にセレクター、NOR、シフトレジスタＳＲ(T)、ＳＲ(P)、左側に出力バッファーを配置している。

4.3.3 チップ構成

図4.13にNEURO2チップのブロック構成図を示す。チップ中央にシナプス回路をマトリックス状に84行84列を四つ配置している。チップの各周辺に42個ずつのニューロン回路ブロックをそれぞれ二つずつ配置している。このニューロンブロック単位にＩselS信号とVref信号ノードが共通に接続されている。チップの四隅には制御信号や各種データI/Oバッファーが配置されている。

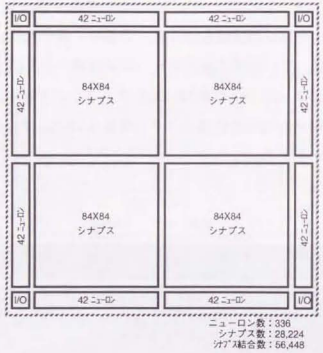
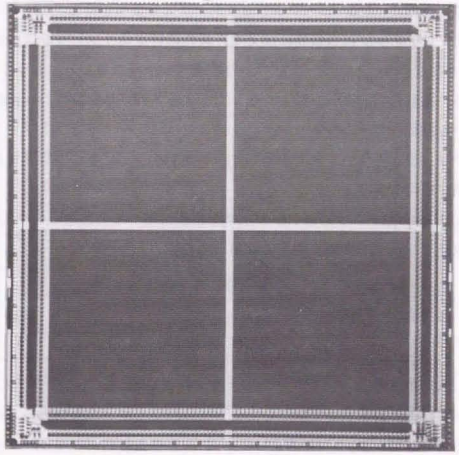


図4.13 NEURO2チップ構成

図4.14はNEURO2チップの顕微鏡拡大写真である。1.0μmCMOS、二層ポリＳｉ、二層金属配線プロセス技術で試作した。各シナプスセルに二つのシナプス結合演算回路があるので、本チップには28,224個のシナプス回路すなわち56,448個の対称なシナプス結合と

336個のニューロン回路を集積している。チップサイズは14.5mm×14.5mmで、393ピンのセラミックPGAパッケージに格納できる。消費電力は最大で3.0Wである。



チップサイズ：14.5mm×14.5mm

図4.14 NEURO2チップ写真

4.4 400ニューロン・40Kシナプス集積ニューロチップ（NEURO3）

ニューロチップの高集積化および高速処理を実現した、キャパシターにシナプス荷重値を保持する、所謂ダイナミックストレージ方式は、微量電荷のリークによって、学習後時間経過とともに荷重値が変動する問題がある。従ってダイナミックストレージ方式のニ

ューロチップを長時間安定に使用するためには、DRAMと同様に、リフレッシュ機能を設けることが必要となる。そこで、大規模なシナプスを集積したニューロチップに適した新しいリフレッシュ方式を考案し、その機能を搭載したNEURO3チップを試作した[3][9][10]。NEURO3チップは、0.8μmCMOS、２層ポリＳｉ、２層金属配線プロセス技術を用いて更なる高集積化を図り、400ニューロン、40Kシナプスを集積すると共に、オンチップ学習機能およびマルチチップ拡張機能に加えて、新たに荷重値リフレッシュ機能を実装した。

新たに考案したリフレッシュ方式は、従来の各シナプス毎にリフレッシュ操作を行う方式と違い、記憶したパターンごとに全シナプスを並列にリフレッシュ操作することを特徴としている。従って、この新リフレッシュ方式はニューラルネットワークが大規模になるほどリフレッシュ操作の並列度が増して、リフレッシュに要する時間を短くすることができる効果が期待できる。

本章ではNEURO3チップのオンチップ学習機能に関する回路構成を中心に述べるに止め、荷重値リフレッシュ機能に関しては第６章で詳しく述べる。

4.4.1　シナプス回路

図4.15はNEURO3チップで採用したシナプス回路を示す。このシナプス回路は学習制御回路以外、NEURO2チップで採用したシナプス回路と同一の回路構成であり、シナプス荷重値保持用キャパシターＣ1の容量は0.3pF、荷重修正回路内のキャパシターの容量は0.02pFである。MOSトランジスターのゲート長はＬは0.8μmで電流パス回路のゲート幅Ｗ＝4.0μm、バイアス用回路のＷ＝2.0μmである。

学習制御回路は、NEURO2における学習制御回路よりやや複雑になり、データのラッチ機能と学習制御信号Ｔsが新たに加わった。全シナプス回路に共通に与えられる学習制御信号Ｔs、Ａcp+、Ａcp-は、非学習時には共に"High"レベルに固定されている。学習時には、＋学習フェーズにおいてニューロンの状態が平衡状態に達したときＴsに負のパルス信号を与えて"Ｓi×Ｓj"の反転信号をレジスタLatchに保持し、－学習フェーズ時の平衡状態に達したところでＡcp+とＡcp-にそれぞれ負のパルス信号を与える。これによって、荷重値は次の量だけ修正される。

$$\Delta Wij = \eta\ (Si^{+} \cdot Sj^{+} - Si^{-} \cdot Sj^{-}) \qquad (4.6)$$

ここでηは正の値でＡcp+/-に与えるパルス数またはパルス幅により変化させることができる。$Si^{+} \times Sj^{+}$データのラッチ機能を加えたことによって、$Si^{+} \times Sj^{+}$と$Si^{-} \times Sj^{-}$が共に"１"の場合にΔWij＝０を正確に表現でき、第３章で評価したように、より安定した学習が実現されると期待される。

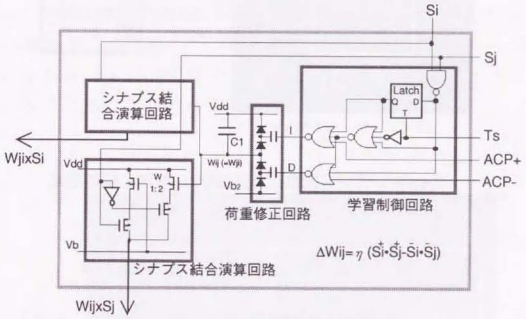


図4.15 シナプス回路（NEURO3）

図4.16は試作したNEURO3チップ上のシナプス回路を顕微鏡で拡大した写真である。一つのシナプス回路の占有面積は55μm×55μmで、下部にシナプス荷重値保持用のキャパシター、下部左側および右側に荷重修正回路、上部および中央に学習制御回路、中央の両側にシナプス結合演算回路が配置されている。

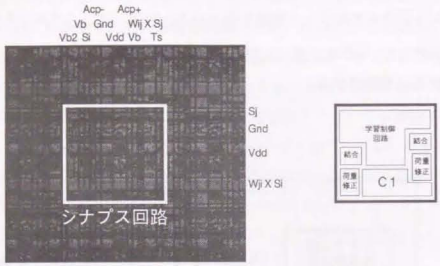


図4.16 シナプス回路（NEURO3）写真と機能構成

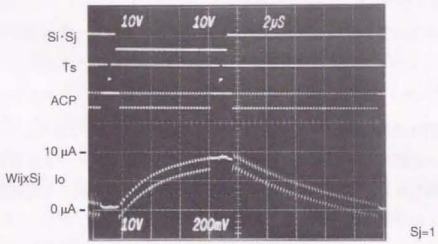


図4.17 シナプス波形（NEURO3）写真

図4.17は、シナプス回路の荷重値修正基本動作を示す実測波形写真である。Ts信号によってSi×Sjがラッチされ、その値とSi×Sjとの差にしたがって、Acp+およびAcp-の

共通信号Acpのパルス信号によりシナプス荷重値が増減することが確認された。最大のシナプス荷重値はVbが4.0Vの時に10μAと、一つのニューロンへ接続されるシナプス数が増加することに対応して少なくしている。

4.4.2 ニューロン回路

ニューロン回路構成はNEURO2チップと同一でありるので説明を省略する。図4.18は試作したNEURO3チップ上のニューロン回路を顕微鏡で拡大した写真である。回路面積は550μm×110μmであり、右側にコンパレータ、中央にセレクター、NOR、シフトレジスタSR(T)、SR(P)、左側に出力バッファーを配置している。

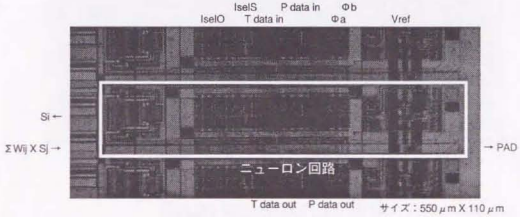


図4.18 ニューロン回路（NEURO3）写真

4.4.3 チップ構成

図4.19にNEURO3チップのブロック構成図を示す。チップの中央に8×33のスタティックシナプス回路ブロックを挟んで、100行100列のシナプス回路マトリックスが四つ配置されている。チップの各辺には50個ずつのニューロン回路ブロックをそれぞれ二つずつ配置し、このニューロンブロック単位にIselS信号とVref信号ノードが共通に接続されている。

チップの四隅には制御信号や各種データI/Oバッファーが配置されて、チップ上部中央と下部中央にはスタティックシナプス回路ブロックに接続されるニューロン回路ブロックが配置されているのいる。スタティックシナプス回路ブロックとそれに接続されたニューロン回路ブロックは荷重値リフレッシュ制御専用に用いられる。その構成内容および、リフレッシュ制御に関しては第6章で詳しく述べる。

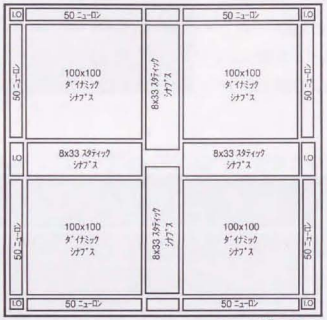
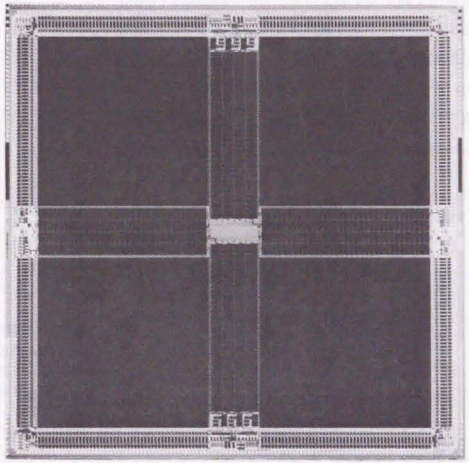


図4.19 NEURO3チップ構成

図4.20はNEURO3チップの顕微鏡拡大写真である。0.8μmCMOS、二層ポリSi、二層金属配線プロセス技術で試作した。各シナプスセルに二つのシナプス結合演算回路があるので、本チップには40K個のシナプス回路すなわち80K個の対称なシナプス結合と400個のニューロン回路を集積している。チップサイズは14.5mm×14.5mmで、消費電力は最大で4.5Wである。



チップサイズ：14.5mm×14.5mm

図4.20 NEURO3チップ写真

4.5 オンチップ学習機能評価

試作したNEURO1チップを用いて、ニューロチップのオンチップ学習能力を評価した[5]。図4.21は、15個の教師パターンを学習した後に、各教師パターンにノイズを加えたテストパターンを用いて連想した出力パターンの平均ハミング距離を示している。125ニューロンの内25個が入力ニューロンで、残り100個は出力ニューロンとするプロパティデータをセットしている。15個の教師パターンは各々125ニューロンの内25個だけが発火状態

"1"とし、各教師パターン間のハミング距離は20以上になるように選んだ。15個の教師パターンは順番に１回ずつ学習して、一通り巡回したら最初のパターンに戻り同じ順序で学習を繰り返す。この繰り返し回数を学習回数と表現している。15パターンを学習させた後には、自己想起型の連想操作を行い、十分にニューロンの状態が安定したところで、それを連想出力パターンとした。この学習能力評価の結果、15パターンを学習する場合、10回程度の学習でほぼ正しい連想結果が得られることが分かった。15パターンを10回学習するのに要する時間は、750μsである。

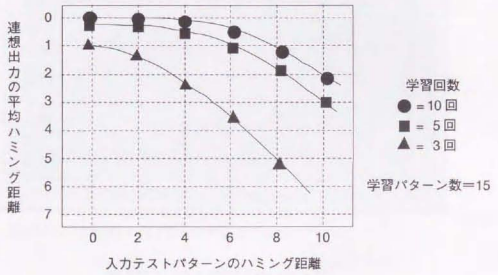


図4.21 学習能力（NEURO1）

図4.22はニューロンの状態収束時間を示している。これは前述の学習能力評価における入力パターンのハミング距離が各々2,4,6の場合の、出力パターンの期待値パターンからのハミング距離の推移を200ns毎に観察した結果である。当初のハミング距離が離れているほど収束に時間がかかっているが、1.5μs以内に収束していることが分かった。この評価においては、各ニューロンの入力ノードに信号ピンを介してチップの外部に1KΩの抵抗と100pFのキャパシターを接続している。このニューロンの収束時間は、ニューロンの入力ノードの容量を小さくすることで短くすることができる。実際、NEURO2チップ以

降では、キャパシターは特に設けず、入力ノードに寄生する約10pFのみで十分安定に、しかも高速に動作することが確認されている。

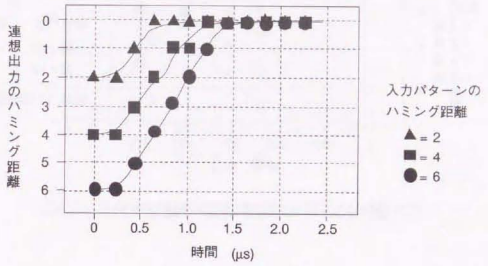


図4.22 ニューロン状態収束時間（NEURO1）

図4.23に学習したパターンの記憶保持時間を示す。試作したニューロチップで採用した、シナプス荷重値をキャパシターで保持する方式においては、素子の構造上避けることのできない微小な電荷のリークによって、荷重値が時間経過と共に変動する問題がある。そこで、この荷重値の変動が連想能力に及ぼす影響を評価した。ここでは、室温（～27℃）の環境温度で10パターンと15パターンの二通りの教師パターン数でそれぞれ10回学習した後に、ハミング距離４の入力パターンにおける連想出力の平均ハミング距離を学習完了直後からの100ms毎に評価した。学習するパターンの数によって記憶保持時間が異なり、10パターン学習時には300ms位まで正しく連想できるものの、15パターン学習時には100ms程度しか正しく連想できない。しかし、15パターンを10回学習するのに要する時間は750μsであり、記憶保持時間の１％以下の時間を定期的に学習に費やすことで、DRAMと同様に長時間安定してニューロ連想メモリーとして使用することが可能であると思われる。

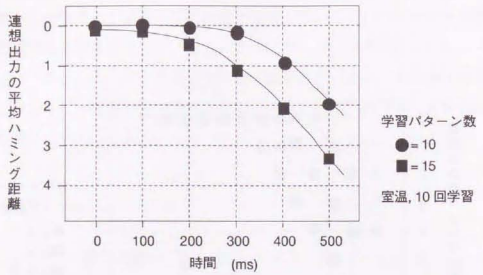


図4.23 データ保持時間（NEURO1）

図4.24はNEURO2チップを２個接続して、336ニューロンの全結合ニューラルネットワークを形成し、そこで学習可能なパターン数を評価した結果を示している。この評価では、二つのニューロチップに各々異なるしきい値電圧Vrefを与え、その電圧差ΔVrefを大きくすることで強制的に学習能力を低下させることで、学習能力を見積もる方法を用いた。336個のニューロンは全て出力ニューロンのプロパティデータをセットし、自己想起型の連想によって評価した。教師パターンは各々336ニューロンの内42個だけが発火状態"1"とし、各教師パターン間のハミング距離は42以上になるように選んだ。学習するパターン数が16,22,28,34の4通りについて、ハミング距離が各々0,8,16,24の場合のテストパターンを入力して連想された出力パターンの平均ハミング距離が0.4以上になる、チップ間のしきい値電圧差ΔVrefの最大値を測定した。図4.24内にプロットされたそれらの測定値による概挿線とΔVref=0との交点から、ほぼ正しく連想できる学習パターン数の限界を見積もることができる。すなわちこの学習評価の条件においては、入力パターンの曖昧度をハミング距離24まで許す場合には約36パターン、ハミング距離16の場合は約41パターン、ハミング距離8の場合には約45パターンまで学習によって記憶できることが見積もられた。試作したニューロチップによれば、ニューロン数の10%から13%に相当する数のパターンを記憶することができることが分かった[8]。

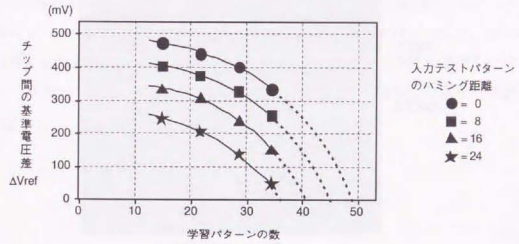


図4.24 ２チップ拡張接続時の学習能力（NEURO2）

図4.25は、NEURO2チップにおけるニューロンの反応速度を評価した波形写真である。写真内の上部の波形はニューロンの入力を制御する信号で、この信号が"High"になることでモニターしているニューロンに接続された別のニューロンの状態を発火状態にして、予め正の荷重値に設定しておいたシナプス回路を介して正の荷重化入力信号がモニターしているニューロンに入力され、その出力状態は発火状態に変化する。また、入力制御信号が"Low"に変化すると接続されたニューロンは非発火状態になり、荷重化入力信号が減すことでモニターニューロンは非発火状態に変化する。この入力制御信号と、写真内の中央部のニューロンの出力状態信号の変化時間の差により、ニューロンの反応時間が50ns以内であることが分かる。NEURO2チップには56,448個のシナプス結合が内蔵され、それらが並列に動作するので、NEURO2チップのシナプス結合演算処理速度は、1.129×10^{12} CPS（Connections Per Second）に相当することが分かった。また、NEURO3チップでのニューロン反応時間は40ns以内であり、80,000シナプス結合内蔵により、2.0×10^{12} CPSを達成している。NEURO2チップおよびNEURO3チップに関するその他の評価に関しては、第５章と第６章で各々述べる。

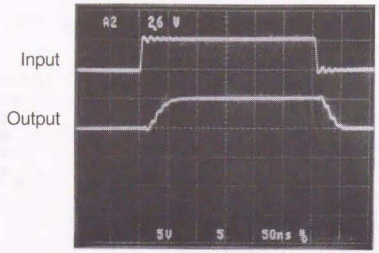


図4.25 ニューロン信号波形写真（NEURO2）

4.6 まとめ

本章では、チップ上に学習機能を実装することにより実現される自己組織化機能に着目した設計思想に基ずき、大胆に簡略化したアナログ回路方式のシナプスおよびニューロン回路を用いて実際に試作した3種類の、学習機能を備えたニューロチップについてその概要とオンチップ学習機能に関する評価結果について述べた。

最初に、1.0μmCMOSプロセス技術を用いて、125ニューロンと10Kシナプスを集積したニューロチップ（NEURO1）を1989年に試作した[1][4][5]。翌1991年2月には、336ニューロン、28Kシナプスと集積度を高めるとに加えて、ニューロチップ同士を接続してニューラルネットワークの規模を拡張できるマルチチップ拡張機能を実装したニューロチップ（NEURO2）を試作した[2][7][8]。更に、1992年2月には、0.8μmCMOSプロセス技術を用いて、400ニューロン、40Kシナプスを集積し、新たにシナプス荷重値の高速リフレッシュ機能を搭載したニューロチップ（NEURO3）を試作した[3][9][10]。

実際に試作したニューロチップによる学習機能評価によって、ニューロン数の10%か

ら13%に相当する数のパターンを記憶することができることが分かった[8]。また、学習したパターンの記憶保持時間は室温環境化で、数100msであることが分かり、学習に要する時間のほぼ100倍程度の期間、記憶が保持されることが分かった[5][10]。ニューロンの反応時間評価により、NEURO2チップのシナプス結合演算処理速度は1.1TCPS、NEURO3が2TCPSに相当することが分かり、世界で初めてテラオーダの処理速度を達成した[8][10]。各々のニューロチップは、学習機能を備えながら世界最高の集積度と最速の処理性能を達成した[5][8][10]。

参考文献

[1] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," in Symp. VLSI Circuits, Digest of Technical Papers, pp. 63-64, June 1990.

[2] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," in ISSCC, Digest of Technical Papers, pp. 182-183, Feb. 1991.

[3] Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," in ISSCC, Digest of Technical Papers, pp.132-133, Feb. 1992.

[4] 有馬裕，益子耕一郎，岡田圭介，山田強，前田敦，近藤晴房，茅野晋平，”125ニューロ,10,000シナプス搭載学習機能付きニューラルネットチップ、” 電子情報通信学会技術研究報告、CPSY90-72,ICD90-128、pp.57-62、1990年10月。

[5] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," IEEE, Journal of Solid-State Circuits, Vol.26, No.4, pp. 607-611, April, 1991.

[6] D H.Ackley, G E.Hinton, and T J.Sejnowski, "A Learning Algorithm for Boltzmann Machines," Cogunitive Science, Vol.9, No.1, pp.147-169, Jan-Mar, 1985.

[7] 村崎充弘，有馬裕，益子耕一郎，岡田圭介，山田強，前田敦，近藤晴房，茅野晋平，”336ニューロ,28Kシナプス搭載学習機能付きニューラルネットチップとBranch-Neuron-Unitアーキテクチャ” 電子情報通信学会技術研究報告、ICD91-108、pp79-85,1991年9月。

[8] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," IEEE, Journal of Solid- State Circuits, Vol.26, No.11, pp. 1637-1644, Nov.,1991.

[9] 有馬裕，村崎充弘，山田強，前田敦，篠原尋史，”リフレッシュ機能内蔵400ニューロ,40,000シナプス搭載アナログニューロチップ、” 電子情報通信学会技 術研究報告、ICD92-16、pp.31-38、1992年5月。

[10] Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," IEEE, Journal of Solid-State Circuits, Vol.27, No.12, pp.1854-1861, Dec.,1992.

第５章
マルチチップ拡張機能搭載ニューロチップ

5.1序

複数のメモリーデバイスを相互に接続してメモリー機能を拡張する技術は、メモリーデバイスの汎用性を高める為に極めて重要である。メモリーの機能拡張には、データのビット幅拡張と記憶容量の拡張の二種類があり、従来のメモリーデバイスでは各デバイス間のアドレス信号線とデータ信号線の接続形態を変えることで、二種の拡張を容易に実現することができた。しかし、ニューロ連想メモリーデバイスにおけるメモリー機能の拡張では、異なった接続形態をとる必要がある。つまり、ニューロ連想メモリーの記憶容量を拡張する為には、複数のニューロデバイスに学習させる記憶内容を分配する機能と、各デバイスの連想結果から一つの出力を選択する機能を設ける必要がある。但しこの付加機能は、メモリーコントローラー側に設ければ良く、並列に接続されるニューロ連想メモリーデバイス自体に機能変更の必要はない。一方、取り扱えるパターンデータのビット幅を拡張する為には、ニューラルネットワークの規模を拡張する必要がある。フィードバック全結合の構造を保ったままその規模を拡張する為には極めて多くのニューロンの入力信号線および出力信号線をチップ間で相互接続しなければならない。この膨大なチップ間配線数とそれに伴う電気特性劣化の問題により、実際に拡張接続できるチップの数は製造コストと動作マージン性能とにより制限される。広範な応用分野で利用できる汎用性の高いニューロ連想メモリーを実現するためには、この問題を解決する必要がある。

そこで、この問題に関わる困難性を緩和し、より大規模なマルチチップ拡張接続を可能にする、チップアーキテクチャーを新たに考案した。そして、そのアーキテクチャーをNEURO2チップ以降で採用し、マルチチップ拡張システムを実際に試作して拡張性能に関する評価を行った[1]。

本章では、新たに考案した拡張接続を容易に実現できるチップアーキテクチャーと、それを採用したマルチチップ拡張機能搭載ニューロチップについて述べる。本節に続く第5.2節で、複数のチップによる拡張接続方式について述べ、BNU（Branch-Neuron-Unit）アーキテクチャの概要と有効性を明らかにする。第5.3節ではBNUアーキテクチャを採用し

実際に試作した、マルチチップ拡張機能を搭載したニューロチップ[2]について、その概要と拡張性能の評価結果について述べる。第5.4節では、拡張接続した18個のニューロチップを搭載したニューロボード[3]の試作概要と、その1000ニューロン全結合規模のニューラルネットワークによる学習機能の評価結果について述べる。

5.2 マルチチップ拡張接続による規模拡張方式

ニューロ連想メモリーでは、取り扱うパターンのビット信号をニューロンの状態信号に対応させて連想処理を行うので、記憶パターンのビット幅を拡張するためには、ニューラルネットワークの規模を拡張する必要がある。また、ニューロ連想メモリーに記憶できるパターン数は、ニューラルネットワークの規模に比例して増大するので、より能力の高い大規模な連想メモリーの実現に対する要求は通常の応用において常に存在する。この要求に答えるためには、マルチチップ拡張機能を搭載したニューロチップの開発が不可欠である。

ここでは、マルチチップ拡張において最も複雑な、フィードバック全結合構造を保ったニューラルネットワーク規模拡張の方法について説明する。まず図5.1にJ. Alspectorらが提案[4]した、従来のマルチチップ拡張方式を示して、その問題点を明らかにする。

図5.1および図5.2では、各チップを構成するニューロンあるいはシナプスを必要最少数で表現しており、集積数に関して実際の例とは異なっている。

この従来の例では、二種類の構造のチップを用いる。すなわち、ニューロンとシナプスが相互に全結合されたニューラルネットワークを内蔵したニューラルネットワークチップとシナプスのみのネットワークを内蔵したシナプスネットワークチップである。この二種類のチップを各々3個づつ用いて拡張接続し、一つのニューロンに関してのチップ関接続関係を抽出して描いている。この従来の拡張接続方式には二つの問題点がある。その一つは、スピード性能が拡張接続するチップ数の増加に伴って低下することである。これは、一つのニューロン回路が駆動しなければならないシナプスの数は拡張接続するチップの数が増加するに伴い多くなることによるニューロン信号伝達時間の増大によって生じる。もう一つの問題は、チップ間の信号接続線数がニューロン数の2倍にも及ぶということである。

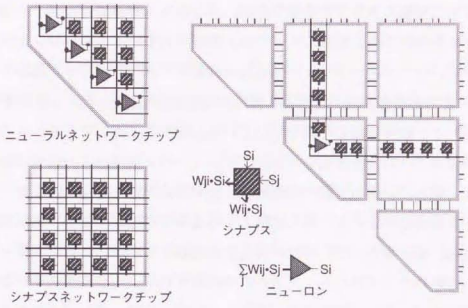


図5.1 従来の拡張接続構成例

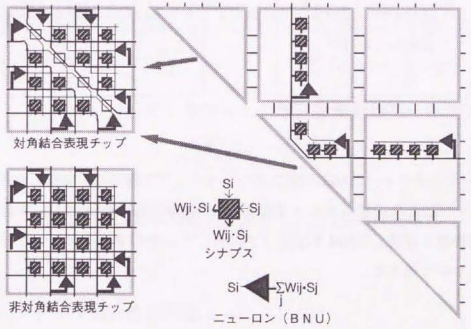


図5.2 新方式の拡張接続構成（BNUアーキテクチャ）

一方、図5.2には新たに考案した、拡張接続方式による接続構成を示している。この新方式においても二種類のチップが必要である。すなわち、対各結合表現チップと非対各結合表現チップである。この二種類のチップは、集積されているニューロンとシナプスの数と配置は同一で、シナプスマトリックスの対角線部分のみの配線だけが異なっている。図5.2には3個の非対角結合表現チップと3個の対角結合表現部分（1.5個の対角結合表現チップに相当）とを拡張接続した例を示している。抽出して描いてある、3つのニューロン回路は一つのニューロンを表現しており、各ニューロン回路は入力信号のみが共通に接続されている。従って、チップ間の接続線数は従来方式の半分である。また、この共通入力ノードは、拡張接続するチップ数に比例して容量成分が増大するが、抵抗成分は並列に接続される結果、減少するので、平均の時定数は拡張チップ数によらず一定である。更に、各ニューロン回路はチップ内のシナプスのみを駆動すればよく、拡張チップ数に依らず負荷は一定であるので、チップ間接続線に寄生する抵抗，容量，インダクタンスが無視できる値であれば、拡張接続するチップ数に依らずニューロンの反応時間は一定に保たれ、拡張チップ数の増加に伴う並列度の増加による速度性能の向上が見込まれる。

この拡張接続方式を可能にするチップアーキテクチャーは、複数のチップにニューロンの機能を分散させて配置させることで特徴づけられ、拡張接続したニューロン回路の構成が、あたかも枝分かれしているように見えることから、BNU（Branch-Neuron-Unit）アーキテクチャーと命名した[1]。

5.3 マルチチップ拡張機能搭載ニューロチップ

BNUアーキテクチャーはNEURO2以降の試作チップで採用し、実際に拡張接続したNEURO2チップにより、マルチチップ拡張機能の性能評価を行った[1]。本節ではマルチチップ拡張機能を搭載したNEURO2チップに関して、BNUアーキテクチャーとその拡張機能評価について述べる。

5.3.1 BNUアーキテクチャーチップ構成

NEURO2チップは、第4.3節で述べた通り、1.0μmCMOS、二層ポリSi、二層金属配

線プロセス技術を用いて、14.5mm×14.5mmのチップサイズに、28,224個のシナプス回路と336個のニューロン回路を集積している。BNUアーキテクチャーは、マトリックス状に配置されたシナプス回路ブロックの四辺に沿ってニューロン回路（BNU）が各辺一列に配置された構成によって特徴づけられる。このBNUアーキテクチャーにおけるこのチップ構成上の特徴は、拡張接続する場合に必要な二種類のチップを容易に作り分けることを可能にしている。つまり、対角結合表現チップと非対角結合表現チップの二種類は、図5.3で示す通り、シナプスマトリックスの対角部分の信号接続を一部変更するだけで作り分けることができる。この信号配線のみの修正による多品種製造方法は、ゲートアレイ製造手法[5]と全く同様で、CMOS集積回路を製造するために必要な全工程のパターン転写マスク十数枚の内、金属配線工程の数枚のパターン転写マスクのみを変更するだけで二種類のチップを開発することができた[1]。

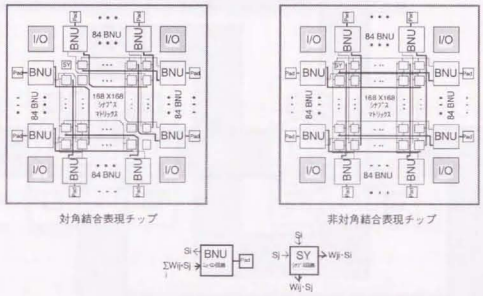


図5.3 対角／非対角結合表現のチップ内信号接続構成

5.3.2 マルチチップ拡張性能評価

2個のNEURO2チップを用いて、図5.4に示すチップ間信号接続によって構成した拡張ニューラルネットワーク評価用ボードを試作し、BNUアーキテクチャーによるマルチチ

ップ拡張性能を評価した。図5.5は、試作した評価用ボードをLSIテスターのプローブヘッドへ装着した状態を撮影した写真である。チップへ与える信号は全て、予め用意した学習用教師パターンと想起用入力パターンを、LSIテスター上にプログラムした学習処理時の制御シーケンスと想起処理時の制御シーケンスに従って順次与え、連想によって出力される信号をLSIテスターに取り込み、出力パターンと期待パターンとのハミング距離を算出して学習性能の評価を行った。LSIテスターを用いれば、VrefやVb,Vdd等の基準電圧や信号の周期等のパラメータを任意に設定することができるので、様々な条件での学習評価を効率良く実施することができた。

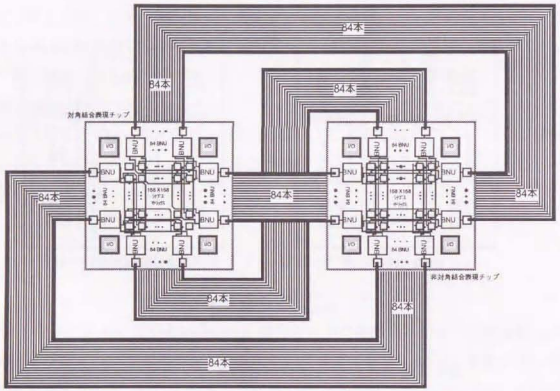


図5.4 2チップ拡張接続構成

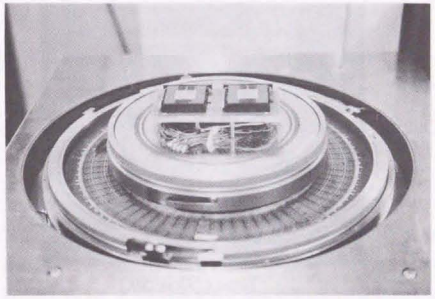


図5.5 チップ拡張接続評価ボード写真

図5.6と図5.7はニューロンの反応時間を示す波形写真である。図5.6と図5.7は各々、1チップの場合と2チップを拡張接続した場合のニューロン信号を、オシロスコープによってトレースした波形である。この波形を観測するために、予め学習操作によって全てのシナプス荷重値が正になるように学習した後、モニターするニューロンに入力される全てのシナプス回路を活性化するか否かを、IselSに与える入力信号を用いて制御している。モニターしているニューロンを"隠れニューロン"に設定し、その他のニューロンを"出力ニューロン"に設定しておき、IselS入力信号が"Low"のとき、全ての出力ニューロンが非発火状態に固定されるように入力データをセットし、IselS入力信号を"High"にして出力ニューロンが自由状態となったときに発火状態になるように、モニターニューロンを含むブロック以外の全てのニューロンのしきい値Vrefを低めに設定しておくことで、このようなニューロンの状態信号の波形が観測できる。各々のモニターニューロンは、84個接続配置されたニューロン回路ブロックの最後に位置するニューロン回路を用いている。また、2チップ拡張接続時には同一のニューロン機能を表現している2つのニューロン回路（BNU）の出力状態を各々モニターしている。

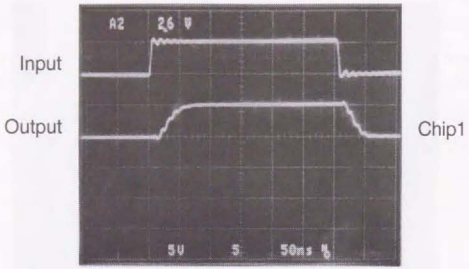
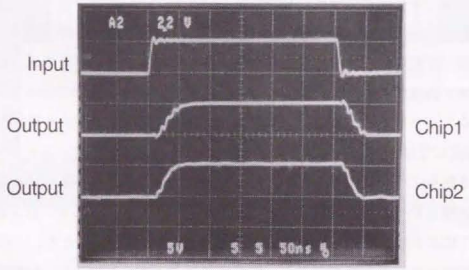


図5.6 ニューロン信号波形写真（１チップ）



56,000 ÷ 50ns = 1.12 X 10 ^{12} cps

図5.7 ニューロン信号波形写真（２チップ拡張）

IselS入力信号が変化してから出力信号が変化するまでの遅延時間には、NORとセレクターSEL1の反応時間、と各ニューロンの状態信号がシナプスへ伝わりその出力電流がモニターニューロンのコンパレータへ伝えられ反応する時間、そして、その出力信号がセレクターSEL2を介して出力バッファーを通過する時間が含まれている。これらの反応時間の内、NORやセレクターSEL1,2、バッファーに関する反応時間はシナプスを介したニューロンの反応時間に比べて十分に小さいので、この波形写真における入力信号からの出力信号の遅延時間をニューロンの反応時間、すなわちシナプス結合演算処理及び非線形変換に要した時間と見なすことができる。

これらの波形を観測した結果、１チップの場合も２チップ拡張の場合も、ニューロンの反応時間には差がなく、拡張接続に伴う速度性能の劣化が生じないことが２チップの拡張接続において確認された。また、どの場合においてもニューロンの反応時間は50ns以内であり、NEURO2チップのシナプス結合演算処理速度性能は1.1×10^{12} CPS （Connections Per Second）と、拡張接続した場合にも各チップの処理性能が保持されることが分かった。この拡張接続による速度性能劣化評価については、第5.4節で述べる18チップ拡張接続ニューロボードを用いた18チップ拡張接続時においても速度性能の劣化が無いことが確認されており、BNUアーキテクチャーの有効性の一つである、拡張接続によって速度性能の劣化を伴わない特長が示された。

拡張接続によって速度性能の劣化を伴わない特長によって、BNUアーキテクチャーにより拡張されたニューラルネットワークは、規模の大きさに関わらず一定の連想速度が実現できることになり、連想メモリーを取り扱う上での制限を緩和することが期待できる。

次に、拡張接続できるチップ数の限界値に関する見積もりについて述べる。図5.8は、図5.5で示した２個のNEURO2チップを拡張接続した評価用ボードを用いて、学習パターン数と２チップ間のしきい値電圧Vrefの差ΔVrefに対する学習能力の変化を評価した結果を示している。この評価結果は第4.5節で述べた学習可能なパターン数の見積りにも用ている。学習する教師パターンは各々336ニューロンの内84個だけが発火状態"1"とし、各教師パターン間のハミング距離は84以上になるように選び、自己想起型の連想によって評価した。学習するパターン数が16,22,28,34の4通りについて、ハミング距離が各々0,8,16,24の場合のテストパターンを入力して連想された出力パターンの平均ハミング距離が0.4以上になる、チップ間のしきい値電圧差ΔVrefの最大値を測定した。

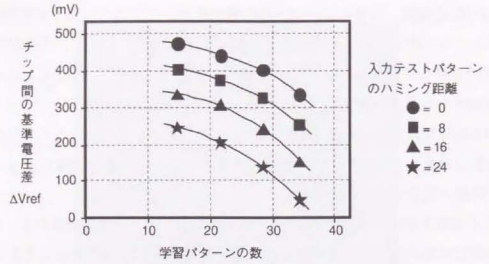


図5.8 拡張接続チップ間補償能力

図5.8の中に示されたチップ間のしきい値電圧差ΔVrefの値は、チップの学習過程で実現される補償能力を表していると考えることができる。この補償能力の評価結果から、拡張接続できるチップ数の限界を見積もることができる。

BNUアーキテクチャーにおいて拡張チップ数を制限する要因は、主に次の二つである。すなわち、チップ間の信号を接続する配線で生じる容量や抵抗、インダクタンスなどの寄生特性と、同一ニューロンを表現するBNU間でのしきい値の不一致である。図5.9は拡張接続されたチップ間およびチップ内の信号配線を、一つのニューロンに関して抽出して示している。C_{LV}とC_{LG}は、各々のチップ内BNUの入力ノードとVdd、GND間との寄生容量を示しており10pF以下である。またシナプス回路の出力電流は最大で14μAであり、BNU内の抵抗器R_Lは1.7KΩである。一方、Rc、Lc、Ccは各々、二つのBNUの拡張接続PAD間を接続した配線に寄生する、抵抗、インダクタンス、容量を示しており、一般的なボード上のでの配線における各々の値は、Rc<10Ω、Lc<100nH、Cc<10pFと見積もられる。これらの寄生回路により生じるAC的な影響は、数十nsのニューロン反応時間に対して、無視し得る程度に小さいので、ここではRcによって生じるDC的な影響にのみに注意を払うことにする。

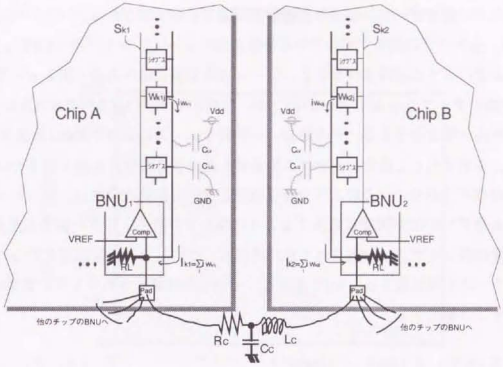


図5.9 拡張接続時の回路構成

チップの外部からPADを介してBNU内の抵抗器R_Lに流れる電流の最大値をI_{kn}とし、最小値をI_{km}とすると、接続されたBNU間の入力ノードの電位差ΔVcは、拡張接続するチップ数をMとして次式で表される。

$$\Delta Vc \leqq Rc \times (I_{kn} - I_{km}) / (2M)^{0.5} \qquad (5.1)$$

ここで、$(2M)^{0.5}$はM個のチップで拡張されたときの同一ニューロンを表現するBNUの数である。ΔVcは拡張接続するチップ数が増加するに従って減少するので、ΔVcの最悪値は2チップ拡張時であり、その値は次の通りになる。

$$\Delta Vc(worst) \leqq 10 \times (168 \times 14\mu A - 0) / 2 = 11.76mV \qquad (5.2)$$

一方、素子特性のバラツキなどによるBNU内のコンパレータのしきい値電圧の差は、8mV以内であるので、同一ニューロンを表現するBNUのしきい値電圧の違いは最大で約20mVとなる。この見積もられた最悪時のしきい値電圧差ΔV(worst)=20mVは、オンチッ

プ学習により実現される、図5.8に示されたしきい値電圧差ΔVrefに対する補償能力の評価結果から、十分なマージンをもって補償されることが予想される。

ここで、各チップの内部で実行される学習操作によって、チップ間のBNUしきい値電圧差ΔVがどのように補償されるかを、シナプス荷重値の修正過程の例をもって説明する。図5.9内左側のチップChip A内のBNU1の方が、右側のチップChip B内のBNU2よりもしきい値電圧が高い場合を考える。その場合、一学習フェーズにおいてBNU1は発火しにくくなるので、本来発火した場合にシナプス荷重値を減少させる修正を施すはずのシナプス荷重値W_{k1j}は修正されない。このような荷重値修正量の自動的な調整は、同一ニューロンを表現する全てのBNUが同時に発火するようになるまで続き、しきい値電圧差を補正できる電流量は各シナプスに分散されて保持される。このしきい値電圧補正のために調整された各シナプスの電流量をΔi_{wkj}とすると、そのときに補正されるしきい値電圧ΔVは次式で表すことができる。

$$\Delta V = R_L \times \Sigma_j (\Delta i_{wkj}) / (2M)^{0.5} \qquad (5.3)$$

ここでΔi_{wkj}を平均値$\langle \Delta i_{wkj} \rangle$に置き換えると、ニューロンの発火率をFrとして次のように表現できる。

$$\Delta V = R_L \times 168 \times Fr \times \langle \Delta i_{wkj} \rangle / (2M)^{0.5} \qquad (5.4)$$

この式を用いて、M=2、R_L=1.7KΩ、Fr=25%として、図5.8内にプロットされた各評価点毎のΔVrefの値をΔVとしたときの$\langle \Delta i_{wkj} \rangle$の値を求めた後に、ΔV=20mVの場合のMの値を算出することで最悪時のしきい値電圧差ΔV(worst)=20mVを補正することができる限界の最大拡張チップ数を求めることができる。その換算結果を図5.10に示す。

図中の横軸は、テスト用入力パターンの曖昧度を、学習する教師パターン間の最小ハミング距離に対する入力テストパターンのハミング距離の割合で示している。また縦軸は拡張接続するチップ数を示し、ハッチングで示された領域が補償機能が十分に働き安定に動作することが期待できる拡張チップ数である。なおこの結果は、記憶するパターンの数をニューロン数の10%とし、記憶パターンは25%のビットが"1"である場合である。

この拡張性能評価によって、実際の拡張システム構築で問題になるBNUの特性バラツキやチップ間接続線に寄生する抵抗、容量、インダクタンスなどの不良因子は、各チップ

の学習過程で実現される自己補償能力によってある程度吸収されることが分かった。その結果、数百から500チップ程度までの拡張接続が可能であることが見積もられた。

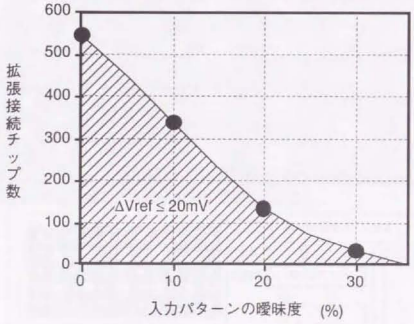


図5.10 拡張性能予測（補償可能領域）

5.4 18チップ拡張接続ニューロボード

BNUアーキテクチャーによるマルチチップ拡張機能性能を、実際の中規模ニューラルネットワークで確かめるために、拡張接続した18個のNEURO2チップを搭載したニューロボードを試作し、学習機能と速度性能について評価を行った[3]。本節では、本ボードの概要と8チップ拡張時および18チップ拡張時における学習能力の評価結果について述べる。

5.4.1 ボード構成

試作したニューロボードの構成図と写真を、図5.11と図5.12に示す。表5.1にこのニューロボードの諸元をまとめる。このボードには、18個のNEURO2チップを拡張接続して

構成した、1,008ニューロン、1,016,064シナプスのニューラルネットワークが実装されており、それらのニューロチップを制御するための制御機能と、学習用教師パターンおよび入力パターンデータ、そして連想出力パターンデータを一時格納するための3.5MByteのSRAMを搭載している。3.5MByteのSRAMには、学習用および想起用のパターンデータを各々1024パターンづつ格納することができる。学習パターン数や学習係数、学習回数、想起パターン数等の制御に必要な値はパラメータレジスターに格納される。このレジスターと制御ロジックは、29個のFPGA（Field Programmable Gate Array device）によって実現している。

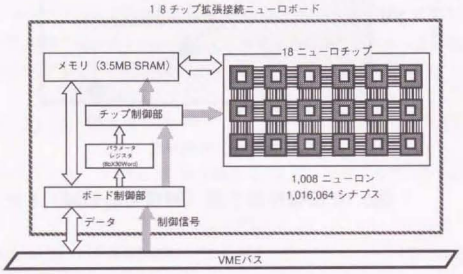


図5.11 １８チップ拡張接続ニューロボード構成

表5.1 １８チップ拡張接続ニューロボード諸元

ニューロン数	1,008ニューロン
シナプス数	1,016,064シナプス
演算速度	20 × 10^{12} CPS
学習速度	500 × 10^{9} CUPS
学習信号ﾒﾓﾘ容量	1,008ﾆｭｰﾛﾝ × 1,024 ﾊﾟﾀｰﾝ
想起信号ﾒﾓﾘ容量	1,008ﾆｭｰﾛﾝ × 1,024 ﾊﾟﾀｰﾝ
ボードサイズ	33.7 cm × 55.0 cm
消費電力	50W（Max.）

CPS：Connections Per Second
CUPS：Connections Update Per Second

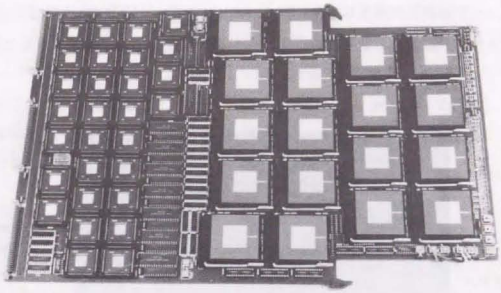


図5.12 １８チップ拡張接続ニューロボード写真

図5.13 １８チップ拡張接続ニューロボード評価装置写真

このニューロボードは、VMEインターフェース機能を備えており、EWSに接続して制御することが可能である。次に述べる評価は、図5.13の写真に示した評価システムにおいて実施した。図5.14は、18チップ拡張接続時におけるニューロンの反応時間を示す波形写真である。前節で述べた2チップ拡張接続の場合と同様の方法でニューロン信号波形を観測している。2チップ拡張接続時と同じく、ニューロンの反応時間には変化を生じな

いことが確認された。ニューロンの反応時間には、50ns以内であるので、本ニューロボードのニューロ演算処理速度は、1,016,064÷50ns＝20.32×10^{12}CPS に相当し、BNUアーキテクチャにより拡張したチップ数に比例して速度性能を容易に増強できることが確認された。

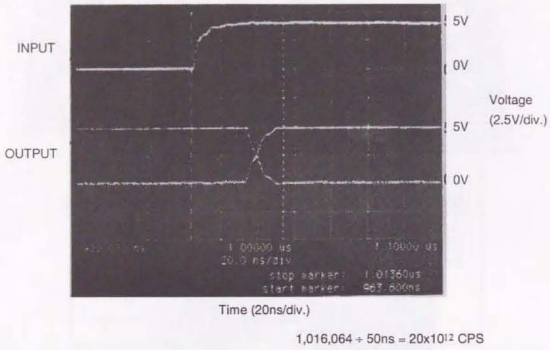


図5.14 ニューロン信号波形写真（１８チップ拡張）

5.4.2 学習能力評価

本ニューロボードはニューロチップの着脱が可能であり、8チップで構成される672ニューロン全結合ネットワークと18チップで構成される1008ニューロン全結合ネットワークの２種類の拡張ネットワークに対して、学習能力を評価した。

8チップ拡張ネットワークでは、504個のニューロンを入力ニューロンとし、168個のニューロンを出力ニューロンと設定して、相互想起型の連想によって学習能力を評価した。学習用の教師パターンは、504ビットの内60ビットだけをランダムに選択して"1"にして発火率を12%にし、各教師パターン間のハミング距離が120以上になるように、自動生成するプログラムにより発生させた。出力ニューロンは1つの教師パターンの対して4ニューロンを割り当てた。テスト用入力パターンは、教師パターンと同じ発火率で、教師パターンから一定のハミング距離分だけ異なるパターンを自動発生させた。

図5.15は、8チップ拡張ネットワークにおける学習するパターン数に関して、学習回数が10,20,30回の場合の認識率を示している。ここで認識率は、教師パターンからのハミング距離が20のランダムに自動生成した500個のテストパターンを入力して連想した結果の正答率を意味している。評価の結果、約150個のパターンを、98%以上の認識率で学習することができることが分かった。

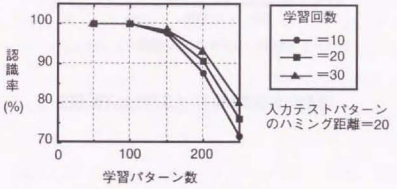


図5.15 学習能力の記憶パターン数依存性（８チップ拡張）

図5.16は、8チップ拡張ネットワークにおいて100個のパターンを学習した場合の連想能力の評価結果を示している。認識率は、入力パターンのハミング距離が70位以上で急激に低下することが分かった。

18チップ拡張ネットワークの評価では、入力ニューロン数を840個、出力ニューロン数を168個と設定して、相互想起型連想を実行した。学習用の教師パターンは、840ビットの内100ビットだけをランダムに選択して"1"にして発火率を12%にし、各教師パターン間のハミング距離が200以上になるように自動生成した。

図5.17は、18チップ拡張ネットワークにおける学習するパターン数に対する認識率の変化の評価結果を示している。テスト用の入力パターンとして、教師パターンからのハミング距離が30の、ランダムに自動生成した500パターンを用いて連想評価した。評価の結果、約250個のパターンを30回学習することで、97%以上の認識率で連想できることが分

かった。

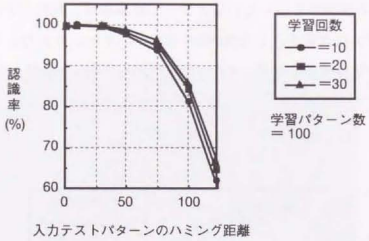


図5.16 連想能力（８チップ拡張）

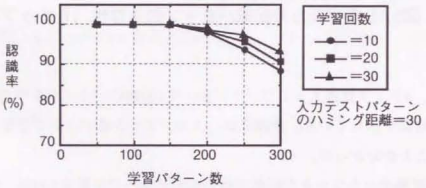


図5.17 学習能力の記憶パターン数依存性（１８チップ拡張）

図5.18は、18チップ拡張ネットワークにおいて150個のパターンを学習した場合の連想能力の評価結果を示している。認識率は、入力パターンのハミング距離が120位以上で急激に低下することが分かった。

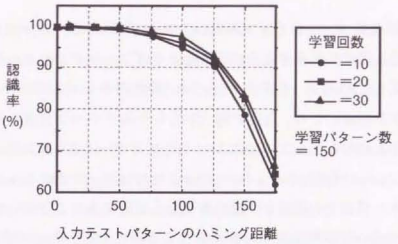


図5.18 連想能力（１８チップ拡張）

アナログニューロボードを使用して8チップ拡張ネットワークと18チップ拡張ネットワークの学習能力を評価した結果、拡張接続するニューロチップの数に比例して、記憶できるパターン数が増加することが確認できた。また、連想処理に必要な時間は拡張チップ数に依らず一定に保たれており、BNUアーキテクチャーが実際の中規模なニューラルネットワークにおいても有効であることが示された。

5.5 まとめ

本章では、複数のニューロチップを用いてニューラルネットワーク規模を拡張できるマルチチップ拡張機能を実現する為に行ったチップアーキテクチャーの研究について述べた。新たに考案したBNUアーキテクチャーは、複数のチップに同一のニューロン機能を分散して配置する構成により、速度の劣化がないマルチチップ拡張を可能にし接続配線数を半減させることができた。このBNUアーキテクチャーによるマルチチップ拡張機能は、実際に試作した2チップ拡張ネットワークと8または18チップ拡張ネットワークにおいて評価され、数百チップまでの拡張が性能の劣化なく実現できることが分かった[1][3]。

また、マルチチップ拡張接続時の性能劣化の原因となる、チップ関素子特性のバラツキによるニューロン回路特性の不一致や、チップ間接続線間に寄生する抵抗、容量、インダクタンスなどの不良因子は、各チップの学習機能によって自動的に補償されることが確認された[1]。BNUアーキテクチャーは、オンチップ学習による自己補償機能が拡張接続した回路網にまで効果的に働く構成を実現したことによって、拡張できるチップ数を大幅に改善することができた。

参考文献

[1] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," IEEE, Journal of Solid- State Circuits, Vol.26, No.11, pp. 1637-1644, Nov.,1991.

[2] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," in ISSCC, Digest of Technical Papers, pp. 182-183, Feb. 1991.

[3] M. Murasaki, Y. Arima, and H.Shinohara,"A 20 Tera-CPS Analog Neural Network Board," in IJCNN, Digest of Technical Papers, pp. 3027-3030, Oct. 1993.

[4] J. Alspector, and R. Allen, "A Neuromorphic VLSI Learning System," Advanced Research in VLSI, MIT Press, pp.313-349, 1987.

[5] 例えば、S. Muroga, "VLSI SYSTEM DESIGN," John Wiley & Sons, Inc., New York, 1982.

第6章
高速リフレッシュ機能搭載ニューロチップ

6.1 序

シナプス荷重値をキャパシターの蓄積電荷量で表現するダイナミックストレージ方式は、多値の荷重値を高集積に保持できることに加え、電荷量の高速な修正が可能なことから、オンチップ学習機能の表現に適している。しかし、キャパシターに保持された電荷量は、微小な電荷のリークによって変動する問題がある。先に試作した2つのニューロチップNEURO1[1]とNEURO2[2]はリフレッシュ機能を搭載しておらず、学習後約300msで認識率が数％程度低下することが分かっている[3]。

一方、従来から多値のシナプス荷重値を高集積に保持する方式として、不揮発性メモリーのEEPROMと同様な、フローティングゲートに荷重値を保持する方法[4]が提案されている。しかし、フローティングゲートによる保持方式は、蓄積電荷量を修正するのに数十μs以上の時間がかかることや、修正回数が十万回程度に制限されることなどから、オンチップ学習機能を表現するのにあまり適していない。また、フローティングゲートへは、ホットキャリアー等の高エネルギー電荷が注入される結果、蓄積電荷量の変動を完全に抑えることは不可能で、フローティングゲートによるデータの保持期間は、2値の場合には10年程度保証できるものの、多値のシナプス荷重値の保持期間は階調分の一以下になる。つまり、シナプス荷重値が30階調の場合、フローティングゲートによるデータの保持期間は、わずか4ヵ月程度となり、実用上不十分である。高階調のシナプス荷重値を実用的な期間不揮発に保持することは、現状の半導体集積回路技術では極めて困難である。

そこで、オンチップ学習機能搭載ニューロチップの高集積化と高速化を実現したダイナミックストレージ方式の特長を活かしつつシナプス荷重値の揮発性の問題を解決する為に、DRAMと同様にリフレッシュ機能を設けることを検討した。但しニューロチップの場合、2値状態のDRAMメモリーデータのリフレッシュと違い、数十階調程度の精度で各シナプス荷重値をリフレッシュする必要があり、より高度なリフレッシュ機能が要求される。

本章では、シナプス荷重値の高速リフレッシュ方式と、その方式を採用して試作したリフレッシュ機能搭載ニューロチップについて述べる。第6.2節では、考案したマクロリ

フレッシュ方式の概要を説明し、第6.3節で、試作したリフレッシュ機能搭載ニューロチップNEURO3[5]のリフレッシュ制御回路構成とリフレッシュ機能の評価結果について述べる。第6.4節では、マクロリフレッシュ方式の有効性について述べる。

6.2 荷重値リフレッシュ方式

従来から提案されているリフレッシュ方式[6][7]は、図6.1に示すように基本的にはDRAMのリフレッシュ方法と同様である。但し、多値のシナプス荷重値をリフレッシュするためには、DRAMの場合と異なり、センスアンプとリフレッシュドライバーの間にA/Dコンバータと D/Aコンバータが挿入される。一度、バイナリーデジタルの離散値に変換することで、離散値間の半値幅までの変動を補正することができる。従って、多値のシナプス荷重値を高速にリフレッシュする為には、高速なA/D，D/Aコンバータが不可欠である。高速なA/D，D/Aコンバータ回路は、回路面積と消費電力の律速から、チップ内に多数設けることが困難で、リフレッシュ回路の並列化によるリフレッシュ高速化には限界がある。特に、大規模なニューロチップにおいては、従来のリフレッシュ方式の適用が難しくなる。

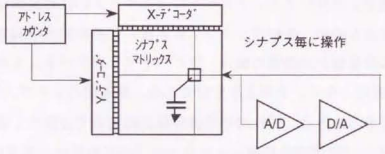


図6.1 シナプス荷重値リフレッシュの従来方式

そこで、新しいリフレッシュ方式を考案し、それを採用したリフレッシュ機能搭載ニューロチップNEURO3を実際に試作した[5]。この新しいリフレッシュ方式は、記憶した

パターン毎に全シナプスを並列に操作することを特徴としており、将来のさらに大規模なニューロチップにも適用可能である。ここでは、シナプス毎にリフレッシュ操作を行う従来方法を、ミクロ・リフレッシュ方式と呼び、新たに考案したパターン毎に全シナプス並列にリフレッシュ操作を実行する方式をマクロ・リフレッシュ方式と対照的に表現して呼ぶ[8]。

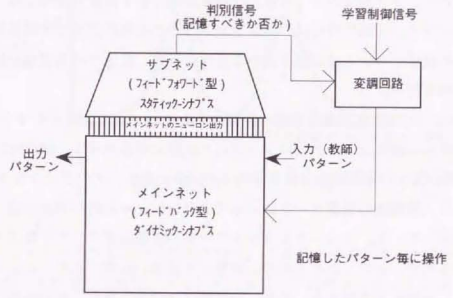


図6.2 シナプス荷重値高速リフレッシュ（マクロリフレッシュ）方式

マクロリフレッシュ方式では、図6.2に示すように、本来のニューラルネットワークであるメイン・ネットワークとリフレッシュ制御専用の階層型ニューラルネットワーク構造のサブ・ネットワークと学習制御信号変調回路を導入する。サブネットワークのシナプスはバイナリデジタルデータ保持方式のスタティック型回路で構成されており、エラーバックプロパゲーションの学習則を実行する。一方、ダイナミックストレージ方式のシナプスで構成された、フィードバック全結合型ニューラルネットワーク構造のメインネットワークは、第3章で述べた通り、ボルツマンマシンの近似学習則を実行する。サブネットワークの入力にはメインネットワークのニューロン状態信号を入力信号として与える。与えられた教師パターンをメインネットワークが学習するのと同時に、サブネットワークはその

入力パターンが記憶するべきパターンか否かを判別できるように学習する。つまり、メインネットワークが与えられた教師パターンを学習しているときに、サブネットワークは、その出力の期待値を”記憶すべき状態”を意味する信号として学習し、次にランダムパターンを入力して出力期待値を”記憶すべきでない状態”として学習する。これらの学習をメインネットワークの教師パターンごとに交互に繰り返すことで、学習後、サブネットワークの出力は、メインネットワークの状態が学習によって得た記憶すべき状態か、または学習した事がない記憶する必要がない状態かを判別する信号となる。サブネットワークのシナプスはスタティック型なので、学習した情報は再学習するまで保持され、メインネットワークのリフレッシュ操作時には、サブネットワークの判別信号で学習制御信号を変調し、記憶すべき状態については記憶を深める操作を行い、記憶すべきでない状態については忘却する操作を行う。

マクロリフレッシュ方式の動作原理を説明するために、メインネットワークのポテンシャルエネルギーを便宜上、一次元にマップした状態に対応させて、図6.3に示した。

図6.3の上側の図中の実線は、学習直後のメインネットワークのポテンシャルエネルギーを示しており、学習後の電荷リークによってポテンシャルの谷の形状が浅く変形した状態を、破線で示している。ニューラルネットワークの状態はポテンシャルエネルギーが低い方へ遷移するので、ポテンシャルの谷底が想起され易い状態であり、ニューラルネットワークが記憶した状態を意味する。また、谷が深ければ深いほど、その谷底の状態は深く銘記されている事になり、記憶をリフレッシュする事は、ぼやけた谷を深くして、谷の形状を元の形に復元する操作に対応している。

図6.3の下側の図に、サブネットワークの学習後の入出力特性を示す。メインネットワークのポテンシャルエネルギーが低いほどサブネットワークの出力ニューロンが発火する確率が高い。サブネットワークの出力が1に近いほどその入力状態、すなわちメインネットワークの状態が、記憶すべき状態である事を意味している。そこで、サブネットワークの出力が発火した場合には、メインネットワークのポテンシャルが深くなるよう、全シナプス回路で同時に次式の荷重値修正則を実行する。

$$\Delta Wij = \eta \times Si \times Sj \qquad (6.1)$$

この荷重値修正則に従って全シナプス同時に修正すれば、その状態の近傍を中心に、図6.4の上側の図中に示した下向きの矢印の如く、ポテンシャルエネルギーが修正される。

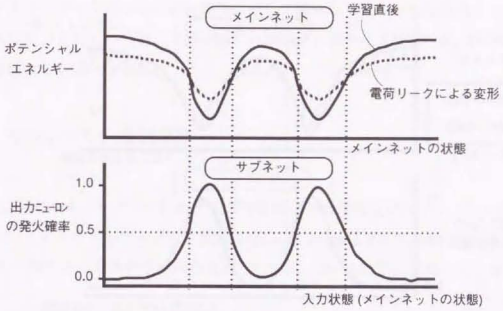


図6.3 各ネットワークの特性

またサブネットワークの出力が非発火の場合、つまり、メインネットワークの状態が記憶すべきでないと判断した場合、メインネットワークのポテンシャルが浅くなるように、全シナプス回路で同時に次式の荷重値修正則を実行する。

$$\Delta Wij = -\eta \times Si \times Sj \qquad (6.2)$$

この荷重値修正則に従って全シナプス同時に修正すれば、その状態の近傍を中心に、図6.4の上側の図中に示した上向きの矢印の如く、ポテンシャルエネルギーが修正される。これらの操作を各想起状態毎に繰り返す事によって、破線のポテンシャルは、元のポテンシャル形状に回復することが期待される。このリフレッシュ方式では、記憶を呼び起こすことでその記憶が忘却されるのを防いでいるので、忘却が進行する前に、このリフレッシュ操作を行うことができれば、記憶を完全にリフレッシュすることができる。

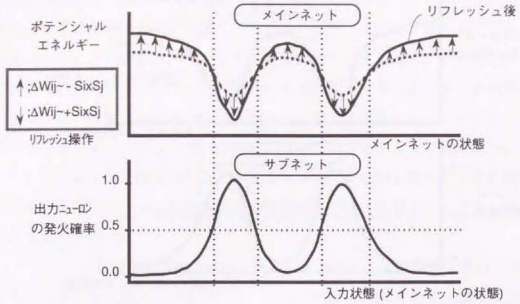


図6.4 マクロリフレッシュのメカニズム

6.3 マクロリフレッシュ機能搭載ニューロチップ

0.8μmCMOS、2層ポリSi、2層金属配線プロセス技術を用いて試作したニューロチップNEURO3は、400個のニューロン回路と40K個のシナプス回路を集積すると共に、オンチップ学習機能およびマルチチップ拡張機能に加えて、新たにマクロリフレッシュ方式を採用し荷重値リフレッシュ機能を実装した。NEURO3チップの回路構成は、既に第4.4節で述べたように、チップの中央にメインネットワークの100行100列ダイナミックシナプス回路マトリックスが4つ配置されており、それらに挟まれた位置にサブネットワークの33行8列スタティクシナプス回路マトリックスが4つ配置されている。メインネットワークのニューロン回路はチップの各辺に沿って100個ずつ配置されており、これら4つのニューロンブロックは、制御信号ＩselＳ、ＩselＯが共通で、各ニューロン内のSR(T)とSR(P)が隣のニューロン間で直列に接続され、各ブロックにそれぞれ2本の100段シフトレジスターを構成している。チップの四隅には、制御信号ＩselＳ、ＩselＯのドライバーや入出

力信号用I/Oバッファーそしてメインネットワークのニューロン状態初期値を設定するための疑似ランダムパターン発生器が備えられている。チップ中央の上部と下部には変調回路とサブネットワークのニューロン回路、そして隠れニューロンから出力ニューロンへのシナプス回路がそれぞれ6個ずつ配されている。チップサイズは14.5mm×14.5mmで、消費電力は最大で4.5Wである。

6.3.1 リフレッシュ制御回路

図6.5にメインネットワークに対する学習制御信号の学習及びリフレッシュ時の制御概要を示す。メインネットワークの学習実行時には、Mode信号を"Low"に固定する。その場合、ACP_に与えるパルス信号は何ら変調されずにACP+とACP-に反転されて伝達され、

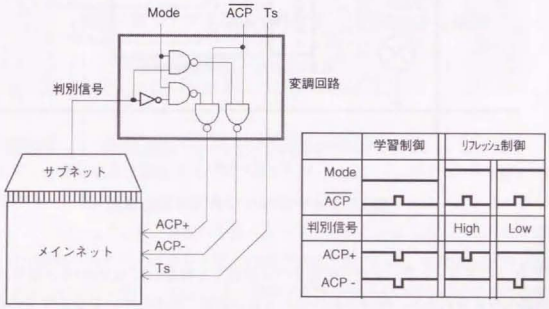


図6.5 学習／リフレッシュ制御信号

各ダイナミックシナプス回路はボルツマンマシンの近似学習則に従って各荷重値を修正する。

リフレッシュ操作時には、Mode信号を"High"に固定し、Ts信号を"Low"に固定する。その場合、ACP_信号はサブネットワークから出力される判別信号によって変調される。サブネットワークが" 記憶すべき状態" と判断した場合、判別信号は"High"となり、

ACP_のパルス信号はACP+にのみ伝達される。また、サブネットワークが" 記憶すべきでない状態" と判断した場合、判別信号は"Low"となり、ACP_のパルス信号はACP-にのみ伝達される。これらのACP+/-信号のサブネットワーク判別信号による変調によって、図6.6で示した各ダイナミックシナプス回路では図6.7に示すごとく、記憶を深めるかまたは忘却するように各荷重値を同時に修正することができる。

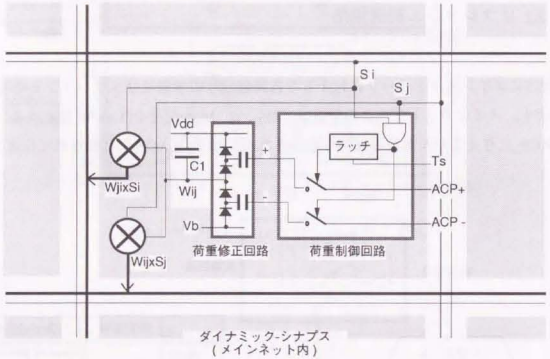
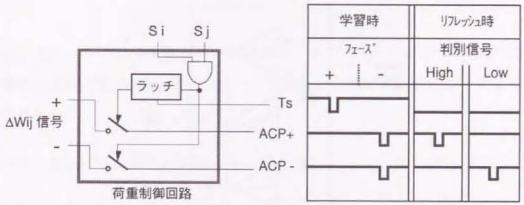


図6.6 ダイナミックシナプス回路構成

つまり、サブネットワークが" 記憶すべき状態" と判断してACP+パルス信号が全シナプス回路へ与えられると、接続されたニューロンが共に発火しているシナプスの荷重値のみが一定量増強され、その状態のポテンシャルエネルギーがより低くなる方向に修正される。また、サブネットワークが" 記憶すべきでない状態" と判断した場合、ACP-にパルス信号が伝達され、接続されたニューロンが共に発火しているシナプスの荷重値のみが一定量弱められ、その状態のポテンシャルエネルギーが高くなる方向へ修正される。

マクロリフレッシュ方式は、図6.6に示す通り、シナプス回路に何らリフレッシュ専用の付加回路を設けることなく、図6.5内に示す簡単な論理回路で構成される学習制御信号変調回路とサブネットワークを追加することによって機能表現することができる。この回

路構成上の特徴による有効性は、第6.4節で述べる。



学習時: $\Delta Wij = S^{+}ixS^{+}j - S^{-}ixS^{-}j$

リフレッシュ時: $\Delta Wij = +SixSj$; 判別信号="記憶すべき"

$\Delta Wij = -SixSj$; 判別信号="記憶すべきでない"

図6.7 各シナプス回路での学習／リフレッシュ制御

6.3.2 リフレッシュ専用サブネットワーク

リフレッシュ制御専用に搭載されたサブネットワークは、入力層400、中間層6、出力層4の三層フィードフォワード型回路網で、400個の入力ニューロンはメインネットワークのニューロンに対応しており、6個の中間ニューロン回路と4個の出力ニューロン回路、そして合計1068個のスタティックシナプス回路で構成されている。中間ニューロン回路は図6.8に示すように、三つのコンパレータで構成され三つの出力信号Sh0,Sh1,Sh2で四つの状態を表し、中間の状態をMSh_信号で表している。図6.9に中間ニューロン回路の顕微鏡拡大写真を示す。出力ニューロン回路は一つのコンパレータで構成され、出力状態は2値である。各ニューロン回路内のコンパレータはメインネットワークのニューロン回路内のコンパレータと同一である。サブネットワークはその入力状態、すなわちメインネットワークの状態が、記憶すべき状態か否かを判別するために、エラーバックプロパゲーション学習の近似則を実行することができる。サブネットワークの判別信号には、4つの出力ニューロンの状態出力と、記憶すべき状態を学習するときに設定した期待パターンとの一致検出回路の出力を用いている。

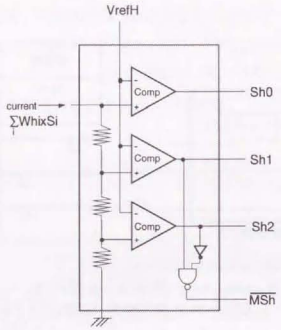


図6.8 ニューロン回路（サブネット中間層）

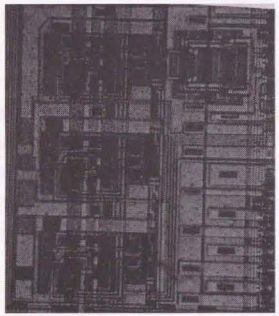


図6.9 ニューロン回路（サブネット中間層）写真

6.3.3 スタティックシナプス回路

図6.10に入力ニューロンと中間ニューロン間のスタティックシナプス回路を示す。シナプス荷重値は4-bitのUp/Douwnカウンターで表現し、学習制御回路によって次の近似学習則が実現される。

$$\Delta Whi = \Sigma\ (Tk - Sk) \times Si \times MSh \times SgnWkh \qquad (6.3)$$

ここで、Whiはｉ番目の入力ニューロンからｈ番目の中間ニューロンへのシナプス荷重値。Tkはｋ番目の出力ニューロンに対する教師状態。Skはｋ番目の出力ニューロン状態。Siはｉ番目の入力ニューロン状態。MShはｈ番目の隠れニューロンの状態が中間値であることを示す信号。SgnWkhはｈ番目の中間ニューロンからｋ番目の出力ニューロン経のシナプス荷重値の符号信号。Σはｋに関して行われる。このスタティックシナプス回路のサイズは195×165μm^2である。図6.11に顕微鏡拡大写真を示す。

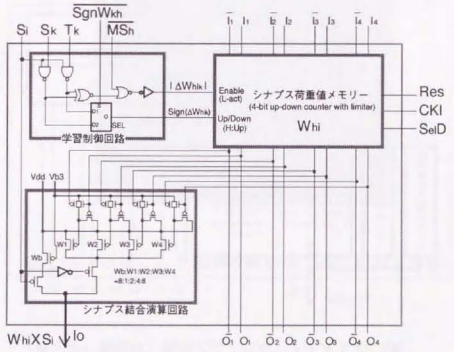
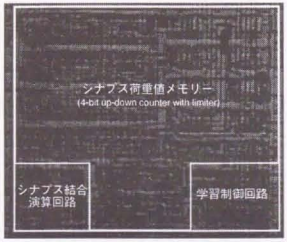


図6.10 スタティックシナプス回路（入力層→中間層）



サイズ：195μm×165μm

図6.11 スタティックシナプス回路（入力層→中間層）写真

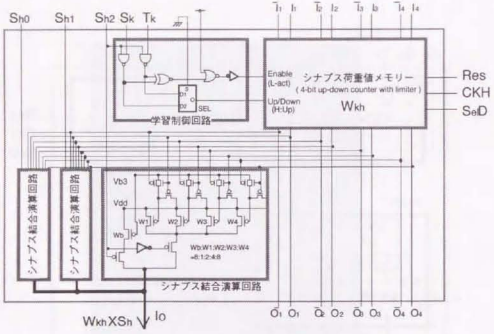


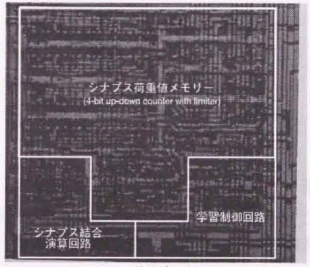
図6.12 スタティックシナプス回路（中間層→出力層）

図6.12には中間ニューロンから出力ニューロンへのシナプス回路を示す。シナプス結合回路は、中間ニューロンの４つの状態に対する各Sh0,Sh1,Sh2信号に対応して３つ備えられている。また学習用入力信号は、MSh＝"Low"、SgnWkh＝"High"にそれぞれ固定することによって、次の学習則が実行される。

$$\Delta Wkh = (Tk - Sk) \times Sh \qquad (6.4)$$

ここで、Wkhはｈ番目の中間ニューロンからｋ番目の出力ニューロンへのシナプス荷重値。Ｓhはｈ番目の中間ニューロン状態。このスタティックシナプス回路のサイズは195×190 μm^2である。図6.13に顕微鏡拡大写真を示す。

図6.14は、スタティックシナプス回路の荷重値修正に関する基本的機能を示す波形観測写真を示す。



サイズ：195μm×190μm

図6.13 スタティックシナプス回路（中間層→出力層）写真

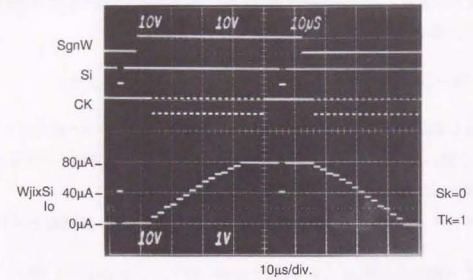


図6.14 スタティックシナプス波形写真

6.3.4 リフレッシュ制御フロー

図6.15に各種制御信号の制御フローを示す。図中のＩselR信号はニューロンの入力信号Datainに与える信号をチップ外部から教師もしくは入力パターンを与える入力ピンかチップに内蔵されている疑似乱数パターン発生回路の出力かを選択する信号で、”Low”時に疑似乱数パターン発生回路が選択される。リフレッシュ操作時にはMode信号を”High”として入力としてランダムパターンをセットし、ニューロン状態の緩和期間を経てTs信号を”Low”としたままACP反転信号を加える。このリフレッシュ操作は通常の連想時に通常の入力パターンを用いて操作してもよい。但しその場合は、あまり連想されないパターンが生じると、そのパターンの忘却が進むことになる。このように出現頻度によって記憶の強さが自動的に調整される本リフレッシュ方式の特徴は、情報を整理する一つの手段として活用できる可能性がある。

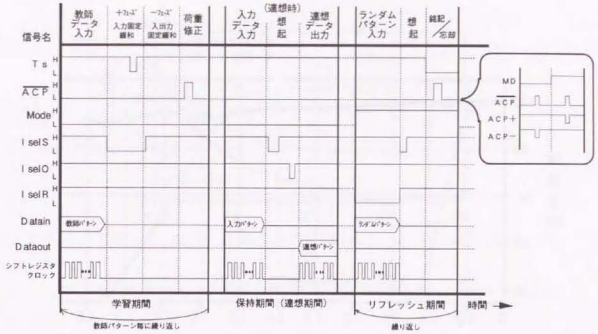


図6.15 各種制御信号制御フロー

6.3.5 リフレッシュ機能評価

図6.16は、NEURO3チップを用いて実際に評価した、学習情報の保持能力に関する評価結果を示す。これは学習後の経過時間に対する認識率の変化を示している。この測定は、27℃の室温において行われ、２パターンをそれぞれ100回学習後、各教師パターンの25%のニューロンにランダムノイズを加えた入力パターンを使って想起を実行させた。破線はリフレッシュ操作を行わない場合と300ms毎にリフレッシュを実施した場合であり、実線が200ms毎にリフレッシュを実行した場合を示している。一回のリフレッシュ操作には20μsを要している。この結果、200ms毎にリフレッシュを実行すれば、約87%の認識率で数秒の間保持できることが分かった。この値は十分とは言いがたいが、サブネットワークの学習能力の向上とリフレッシュスケジュールの最適化によって更に向上させることができると考えられる。

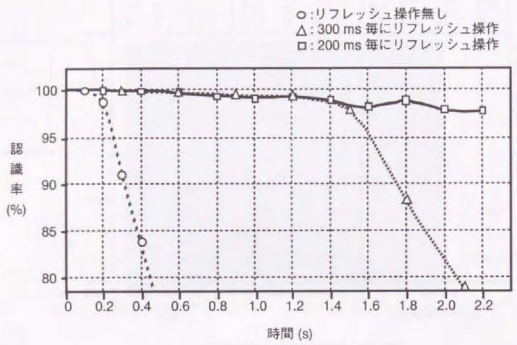


図6.16 シナプス荷重値保持特性

6.4 マクロリフレッシュ方式の有効性

マクロリフレッシュ方式は、各記憶パターン毎に全シナプスが並列に荷重値修正操作を実行することから、高速なリフレッシュ処理が実現される特長がある。また、本方式を実行する為に、シナプス回路内にリフレッシュ用の付加回路を設ける必要がなく、ニューラルネットワークの規模が大きくなるに連れて、サブネットおよび学習制御信号変調回路の付加回路の面積占有率は小さくなる特長がある。これらの特長は、ニューラルネットワークの規模が大きくなればなるほど有効に働くと考えられる。次に、これらのマクロリフレッシュ方式における有効性に関して見積もった結果について述べる。

図6.17は、リフレッシュ機能を搭載する場合の、シナプス規模に対するチップ面積の増加率を示す。従来のミクロリフレッシュ方式では、A/D,D/Aコンバータの他に、シナプスマトリックスの中から目的のシナプスを選択するためのX-Yデコーダと、各シナプス回路内にスイッチ用素子と縦横に走る信号線を設ける必要がある。

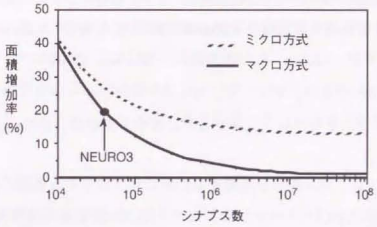


図6.17 面積増加率の比較

デコーダの面積はシナプス数の平方根に比例するが、シナプス内のスイッチ素子と信号線で占められる面積はシナプス数に比例して増大する。従って、従来方式における面積の増加率はシナプス数の増加に伴いほぼ10%増の値で飽和することが見積もられる。

一方、マクロリフレッシュ方式で必要な回路は、サブネットワークと学習制御信号変調回路のみである。サブネットワークの回路規模はその大部分を占める入力から中間へのシナプスの規模 $n \times h$ に比例する。ここで、nはメインネットワークのニューロン数つまりサブネットワークの入力ニューロン数で、hはサブネットワークの中間ニューロン数、サブネットワークの出力ニューロン数は1とする。サブネットワークの中間ニューロン数hはメインネットが安定に記憶できる状態の数によって次式で規定される。

$$O(n \times h / \log(n \times h^2)) \geqq O(n) \qquad (6.5)$$

上式の左辺はサブネットワークが弁別可能なパターン数の下限[9]であり、右辺は学習後のメインネットワークのとりうる安定な状態数[10][11]である。nが十分大きい場合、サブネットワークのhは2log n程度で十分なので、サブネットワークの回路規模は $n \times \log n$ に比例することが見積もられる。従って、マクロリフレッシュ機能搭載によるチップ面積の増加率は図6.17に示すように、メインネットワークのシナプス数の増加に従って非常に小さくすることができると予想される。

次に、リフレッシュに要する時間を見積もる。従来のリフレッシュ方式ではセンスア

ンプとリフレッシュドライバーの間にA/DコンバータとD/Aコンバータが挿入され、リフレッシュ時間はA/D,D/Aコンバータの反応時間が大きく影響する。ここで、リフレッシュされるべきシナプス荷重値の要求精度を500mV／5bitとした場合、0.8μmCMOSプロセスによるA/Dコンバータが ～20ns, ～0.5×0.5mm^2 ,～60mW 、 D/Aコンバータが ～5ns,～0.3×0.3mm^2, ～40mW であり、デコーダ、センスアンプ、データラッチなどの遅延を考慮して、一つのシナプスをリフレッシュするのに要する時間は、約50ns程度であることが見積もられる。

全シナプスのリフレッシュ時間を短縮するためにリフレッシュ処理の並列化が望まれるが、高速のA/D,D/Aコンバータセットは１セット約100mW程度の消費電力が必要であり、リフレッシュ回路に許される消費電力の制限から、高度な並列処理化が困難である。従って、この従来方式によれば全シナプスのリフレッシュに要する時間は、図6.18内の破線で示すように、シナプス数に比例して増大することになる。例えば、リフレッシュ回路に許される消費電力を600mWとした場合、800万シナプス規模で想起可能な期間に対する学習期間の割合が50%を割ってしまう。これは、ニューロチップの高集積化を進める上で大きな障害となる。

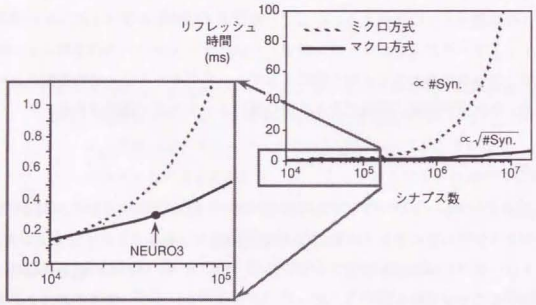


図6.18 リフレッシュ時間の比較

一方、マクロリフレッシュ方式の場合、リフレッシュ操作がメインネットワークが記

憶しているパターン毎に実施されることから、リフレッシュに必要な時間は、記憶パターン数に比例して増加し、メインネットワークが記憶できるパターン数は、ニューロン数に比例し、ニューロン数はシナプス数の平方根に比例することから、マクロリフレッシュ方式によるリフレッシュ時間は、図6.18内の実線で示すように、メインネットワークのシナプス数の平方根に比例することが期待される。マクロリフレッシュ方式によれば、より大規模なニューラルネットワークにおいて、その高速リフレッシュの効果が顕著になることが見積もられた。

6.5 各種不揮発性連想記憶ニューラルネットワークLSIの微細化トレンド

ここで、シナプス荷重値を保持する手段としてキャパシターの代りにフローティングゲートにした場合の素子微細化に伴うニューロ連想メモリーデバイスの性能トレンドを見積もる。フローティングゲートに精度良く電荷を出し入れするために幾つかの回路構成が考案されている[12][13]。書き込み制御電圧に対して負の1/k電圧をk倍の容量でフローティングゲートにカップリングさせることで、書き込み時のフローティングゲートの電位を常に一定電圧（0V）にして電荷注入精度を高めたCCBS（Coupling-Charge Balancing Scheme）[12]方式の場合、キャパシターとほぼ同程度の面積で実現できる。また、フローティングゲートの電荷注入部分と保持部分との間に高抵抗を挿入することで、パルス信号を使って少量の電荷を少しずつ注入できるRCS（Resistance Connecting Structur）[13]の場合は、フローティングゲートの電荷注入部分と保持部分との面積比を100倍以上とる必要があることから、高集積化には向かない。キャパシターに保持する場合とそれにリフレッシュ機能を設けた場合、そして、CCBSとRCS回路方式によるフローティングゲートの場合について、各々、１チップに集積できるシナプス数（ニューロン数=シナプス数$^{0.5}$を含む）と演算速度（CPS）とのゲート長の微細化に伴う推移を図6.19と図6.20に示す。
CCBS方式によるフローティングゲートを用いれば高集積と高速処理性能で優れているが、その実用化には、次の問題を解決する必要がある。集積規模が高まり記憶できるパターン数が増すほど学習過程での電荷書換回数が増大し電荷注入膜の劣化が生じ保持性能が低することで安定に学習によって記憶できるパターン数が制限される問題がある。現状のフローティングゲート書き換え可能回数は10～100万回程度で、高々1000パターンを数百回程

度学習すると保持特性は急速に劣化するものと予想される。また、書き込み速度が数10μsとキャパシターの場合より4～5桁遅い特性は付章で述べる時分割規模拡張表現を高速に実行できない問題がある。フローティングゲートを用いてシナプス荷重値を保持する方式は、学習速度が極めて遅く荷重値修正回数に制限があることから、オンチップで頻繁に学習操作が必要な汎用の連想メモリーデバイスとしては適さないと思われる。

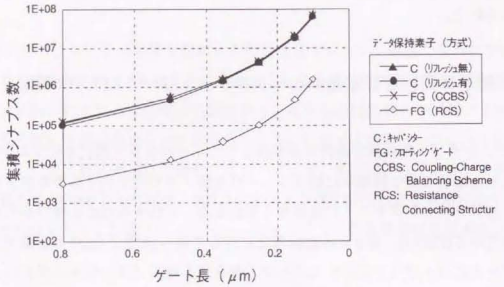


図6.19　連想記憶アナログニューロLSIの微細化トレンド（その１）

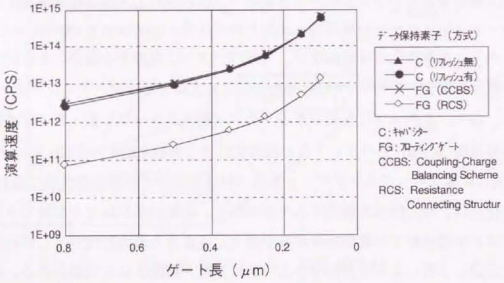


図6.20　連想記憶アナログニューロLSIの微細化トレンド（その２）

6.6 まとめ

リフレッシュ機能を搭載しないニューロチップにおいては、チップ外部に教師情報を保持するメモリーを設け、そのメモリに保持した教師パターンを用いて定期的に学習を繰り返すことで、高い認識率を維持することが可能である。しかし、学習するパターン数が増大するに伴って、想起可能な期間に対する学習期間の割合が無視できなくなり、多くのパターンを記憶できる大規模なニューラルネットワークほど、この問題が顕著になる。現状の記憶保持時間と学習速度によれば、１万ニューロン、１億シナプス規模のニューラルネットワークで想起可能期間の割合が50%を割ることになる。また、記憶したいパターンの近傍に分布した類似パターンの集合で、学習パターンが構成される場合では、記憶する代表パターンの数に比べて学習するパターン数が増大し、学習時間が極めて長くる問題が生じる。

そこで、このような問題を解決できる、大規模なニューロチップに対して有効な、シナプス荷重値リフレッシュ方式を開発し、本方式を採用したアナログニューロチップを試作し、基本的動作を確認した[8]。新に考案したリフレッシュ方式は、シナプス毎に荷重値をリフレッシュ操作を行う従来の方式とは異なり、記憶したパターン毎に全シナプス並列にリフレッシュ操作する方式により、大規模なニューラルネットワークにおいても高速なリフレッシュが可能であることが見積もられた[8]。また、このリフレッシュ方式は、リフレッシュ制御専用のサブ・ネットワークを導入し、その判別信号によって学習制御信号を変調するという簡単な回路構成によって機能表現ができ、各シナプス回路にはリフレッシュ専用の付加回路を必要としないことから、シナプス回路の占有率が高い大規模なニューロチップほど、リフレッシュ機能搭載時の回路面積増加率を小さく押えられる効果があることが見積もられた[8]。

参考文献

[1] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," in Symp. VLSI Circuits, Digest of Technical Papers, pp. 63-64, June 1990.

[2] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," in ISSCC, Digest of Technical Papers, pp. 182-183, Feb. 1991.

[3] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," IEEE, Journal of Solid-State Circuits, Vol.26, No.4, pp. 607-611, April, 1991.

[4] M. Holler, S. Tam, H. Castro, and R. Benson," An electricaly trai-nable artificial neural network (ETANN) with 10240 Floating gate synapses," Proc. of IJCNN-89, Vol.2, pp.191-196, 1989.

[5] Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," in ISSCC, Digest of Technical Papers, pp.132-133, Feb. 1992.

[6] T. Morishita, Y. Tamura, and T. Otsuki, "A BiCMOS analog neural network with dynamically updated weights," ISSCC, Digest of Tecnical papers, pp.142-143, Feb., 1990.

[7] B. Boser, and E. Sackinger, "An Analog Neural Net-work Processor with Programmable Network Topology," ISSCC, Digest of Technical Papers, pp.184-185, Feb., 1991.

[8] Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," IEEE, Journal of Solid-State Circuits, Vol.27, No.12, pp.1854-1861, Dec.,1992.

[9] S. Akaho, and S. Amari," On the Capacity of Three Layer Networks," Proc. of IJCNN-90, Vol.3, p1-6, 1990.

[10] D.J. Amit, H. Cutfreund, and H. Sompolinsky,"Spin-Glass Models of Neural Networks," Phys. Rev., A2, pp.1007-1018, August, 1985.

[11] D. J. Amit, H.Gutfreund, and H. Sompolinsky," Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks," Physical Review Letters, Vol. 55, No. 14, pp. 1530-

1533, Sep., 1985.

[12] Kyu-hyoun-Kim, and Kwyro Lee, "A True Non-Volatile Analog Memory Cell using Coupling-Carge Balancing," ISSCC, Digest of Tecnical papers, pp.268-269, Feb., 1996.

[13] T. Morie, O.Fujita, and K. Uchimura, "Self-Learnig Analog Neural Network LSI with High-Resolution Non-Volatile Analog Memory and a Partially-Serial Weight-Update Architecture," IEICE, Transactions on Electronics, Vol. E80-C, No.7, pp.990-995, July 1997.

第7章
総括

本論文は、学習機能を搭載した連想記憶ニューラルネットワークLSIの高集積化および高速化を検討し、実際にニューロチップを試作すると共にマルチチップ拡張システムを構築して、大規模ニューロ連想メモリーが高い信頼性で実現できることを実証した、一連の研究結果をまとめたものである。本章では、この研究で得られた結果を要約する。

第2章では、連想記憶ニューラルネットワークの高集積化と高速化に適したアナログ回路の優位性と問題点を明らかにした上で、チップ上に学習機能を実装することでニューラルネットワークLSIの微細化に伴う素子特性バラツキの問題を克服できる可能性を示した。その要点は次のとおり。

（1）アナログ回路は、ニューラルネットワークの大規模並列処理を表現するのに適しており、演算機能に関してデジタル回路表現と比べて約300倍の高集積化と約500倍の高速化が実現できることが、実際に試作されたニューロチップの比較で確認された。また、50ニューロンの連想記憶ニューラルネットワークを計算機シミュレーションで評価した結果、連想メモリーに必要なシナプス精度は6～7bit程度であることを確認した。

（2）アナログニューラルネットワークLSIの連想性能は、素子特性バラツキが3%を超えると徐々に低化しはじめ、ゲート長が0.3～0.4μmレベルになると急速に劣化することを計算機シミュレーションによって確認した。

（3）チップ上に学習機能を設けて回路素子レベルに自動補償機能が働く回路構成にした場合、素子特性バラツキが30%程度までは、連想性能の顕著な劣化を招かないことを計算機シミュレーションによって確認した。

（4）オンチップ学習による自動補償機能によって、アナログニューラルネットワークLSIは、最小線幅0.15μm程度まで微細化が可能であることが見積もられた。

（5）チップ上に学習機能を設けることで自動補償機能が働くことが確認できた結果、自動補償が期待される機能部位に関して、大胆な回路の簡略化が許され、学習機能自体の高集積化を図る有効な回路設計方針を得ることができた。

第3章では、自動補償機能を考慮した学習機能の高集積化回路を提案し、大胆に簡略化された学習回路でも自動補償機能によって十分な学習性能が実現できることを明らかにした。また、提案したシナプス回路は電源電圧に対する動作マージンを大幅に改善できることを示した。要点は次のとおり。

（1）ニューロンの属性と教師データを任意に設定できる学習機能を備えたニューロン回路を提案した。

（2）アナログ量のシナプス荷重値をキャパシターの蓄積電荷量で表現し、チャージポンプ回路による荷重値修正回路とその修正指示パルス信号を発生する簡単な論理回路で構成された学習回路によって構成される、極めてコンパクトな学習機能を備えたシナプス回路を提案した。

（3）提案した学習回路は、その高集積化のために採用した簡略化学習ルールや荷重修正回路の非線形特性などの機能制限に加え、約30%のチップ内素子特性バラツキをも自動補償でき、十分な学習性能と連想性能を実現できることを計算機シミュレーションによって確認した。

（4）提案したシナプス荷重値表現回路は、トランジスターを飽和領域で動作させる回路構成によって、電源電圧が±10% 変動してもシナプス荷重値は3% 以内の変動に抑えられることを回路シミュレーションによって確認した。

（5）提案した学習回路構成によって連想記憶アナログニューラルネットワークLSIは、0.15μmレベルで1チップに数千万シナプスを集積し数百テラCPSの処理速度に到達できることが見積もられた。

第4章では、実際に試作した学習機能搭載ニューロチップの概要と評価結果をまとめ、その連想メモリー性能や演算速度を明らかにした。要点は次のとおり。

（1）1.0μmCMOSプロセス技術を用いて、125ニューロンと10Kシナプスを集積した学習機能搭載ニューロチップ（NEURO1）を試作した。

（2）1.0μmCMOSプロセス技術を用いて、ニューロチップ同士を接続してニューラルネットワークの規模を拡張できるマルチチップ拡張機能を実装した、336ニューロン、28Kシナプスの学習機能搭載ニューロチップ（NEURO2）を試作した。

（3）0.8μmCMOSプロセス技術を用いて、シナプス荷重値の高速リフレッシュ機能を

実装した、400ニューロン、40Kシナプスの学習機能搭載ニューロチップ（NEURO3）を試作した。

（4）試作したニューロチップによる学習機能評価によって、ニューロン数の10%から13%に相当する数のパターンを記憶することができることを確認した。

（5）室温における学習パターンの保持時間評価によって、記憶保持時間は、数100msであることが分かり、学習に要する時間のほぼ100倍程度の期間、記憶が保持されることを確認した。

（6）ニューロンの出力信号波形観測によってニューロンの反応時間を評価した結果、NEURO2チップのシナプス結合演算処理速度は1.1×10^{12}CPS、NEURO3が2×10^{12}CPSに相当することを確認した。

第5章では、大規模回路網を実現するために、複数のニューロチップを接続して容易に規模拡張が図れる、ニューロン機能分散表現マルチチップ拡張方式を提案し、実際に試作したマルチチップ拡張機能搭載ニューロチップとその拡張接続システムによって評価した結果をまとめ、その拡張性能を明らかにした。要点は次のとおり。

（1）同一のニューロン機能を各チップで分散表現する構成を用いることにより、接続配線数を半減するとともに拡張接続による速度性能の劣化を招かない特長を有した、マルチチップ拡張方式とそのチップアーキテクチャーを提案した。

（2）試作したマルチチップ拡張機能搭載ニューロチップを用いて実際に2チップ拡張ネットワークを構築して拡張性能を評価した結果、チップ間の素子特性バラツキによるニューロン回路特性の不一致や、チップ間接続線間に寄生する抵抗、容量、インダクタンスなどの不良因子は、各チップの学習機能によって自動的に補償されることが確認され、数百チップまでの拡張が可能であることが見積もられた。

（3）ニューロンの信号波形を観測して、拡張接続するチップ数によらず、各ニューロンの反応時間が一定であったことから、拡張接続によるスピード性能の劣化を生じないことが確認された。その結果、拡張システムの演算速度性能は、拡張接続するチップ数に比例して向上することが見積もられた。

（4）18チップまで搭載可能なニューロボードを試作した。実際に18チップを拡張接続して構成された1000ニューロン、100万シナプスのニューラルネットワークで連想性能を評価し、20×10^{12}CPSの演算速度と数百パターンを学習できることを確

認した。

第6章では、高集積化のために採用したシナプス荷重値のダイナミックストレージ方式の揮発問題を克服するために、大規模なニューラルネットワークに有効な高速リフレッシュ方式を提案し、その回路構成を採用して試作した高速リフレッシュ機能搭載ニューロチップの概要とその評価結果についてまとめた。要点は次のとおり。

（1）記憶したパターン毎に全シナプスを並列にリフレッシュ操作することで、大規模なニューラルネットワークほど高速な荷重値リフレッシュが可能な、マクロリフレッシュ方式を提案した。

（2）マクロリフレッシュ方式は、リフレッシュに要する時間が記憶したパターン数に比例することから、シナプス数に比例する従来のリフレッシュ方式に対して、シナプス数の0.5乗に比例した時間でリフレッシュできることが期待でき、大規模なニューラルネットワークに有効であることが見積もられた。

（3）マクロリフレッシュ方式は、リフレッシュ制御専用のサブ・ネットワークと学習制御信号の変調回路を付加するだけで実現することができ、各シナプスには何ら回路を付加する必要が無いことから、シナプス回路の占有率が高い大規模なニューロチップほど、リフレッシュ機能搭載時の回路面積増加率を小さく押えられる効果があることが見積もられた。

（4）試作したリフッレシュ機能搭載ニューロチップのリフレッシュ性能を評価した結果、200msごとにリフレッシュ操作を繰り返した場合に10倍以上の保持時間延長ができることが確認された。

これら一連の研究によって、高集積化および大規模並列処理に優れたアナログニューラルネットワークLSIは、学習機能をチップ上に実装することで素子特性バラツキを補償し動作マージンを高められることが実証され、従来のデジタル集積回路と同様に、半導体集積回路の更なる微細化に対しても十分に高い信頼性と拡張性を確保でき、大規模な連想記憶ニューラルネットワークを構築できる見通しを得ることができた。しかし、今回実現した連想メモリーの機能モデルは極めて単純なものであり、今後、生体脳に匹敵する高度な機能を工学的に再構築しニューラルネットワーク技術を実用化するためにはまだ多くの課題が残されている。

本研究では連想メモリーの機能として極めて単純な静的挙動にのみに着目したが、高度な連想メモリー機能を実現するためにはダイナミックな連想挙動を実現できるカオス現象を再現する必要がある。そのためにニューロン回路にカオス振動子を組み込むなどの新しい機能回路に関する研究を進める必要がある。また、学習機能についても同時性を考慮した動的な機能を強化する必要がある。そして、多数の連想メモリーで構成され、それらが自立的に相互作用しながら高度な直感的情報処理を実現する脳型コンピュータの構築を目指して、連想メモリーに基づくニューロコンピュータアーキテクチャに関しても、大規模で高速なニューラルネットワークを実際に構築・動作させて構成的研究を進める必要がある。

生体脳の工学的再現に関する研究が構成的手法にのみその成功の可能性があるとすれば、ニューラルネットワークの機能モデルをより大規模により高速に実現し安定に動作させる学習機能を搭載したアナログニューラルネットワークLSI技術は、脳の研究分野における有効な道具になると確信する。

付章
時分割規模拡張方式

生体脳と比較すると極めて単純なニューラルネットワークの機能集積に過ぎないものの、学習機能を実装して素子の特性バラツキを補償する集積回路構成を採用することで、最小線幅0.15μmのプロセス技術が実用化される西暦2000年頃には、一チップに約2000万シナプスを集積し200×10^{12}CPSの処理性能を持つアナログニューロチップが実現できると予想される。そのチップを約50個搭載したボードレベルでは大脳の処理速度に匹敵する10^{16}CPSに到達することができる。しかしながら、表現できるニューラルネットワークの規模は大脳の規模に比べて５桁も少ない10^{9}シナプス程度である。そこで、複数のボードを用いたシステムレベルの規模拡張を図った場合、消費電力や信号通信遅延等の制限により数千チップ（～100ボード）までの拡張が限界と考えられるが、その場合でも大脳より３桁程度少ない規模に留まる。ヒトの大脳レベルの生体脳と比べて、半導体集積回路によるニューラルネットワークは、演算速度に関して十分な性能に到達できるものの、集積度に関しては未だかなりの隔たりが残されている。

そこで、半導体集積回路によるニューラルネットワークの高速性能を活かし、実質的に規模を拡張できる、時分割規模拡張方式を検討する。本章ではその概要を述べると共にその方式により演算速度を４桁程度犠牲にし従来のデジタルメモリーを付加することで任意の規模拡張が可能であることを示す。

ここで、拡張表現したいニューラルネットワークは多数の密に結合した小部分ニューラルネットワークで構成されているものとし、全ニューロン数をN、全シナプス数をYとし、小部分ニューラルネットワークの平均ニューロン数をｎ、シナプス数はn^2とする。また、ニューロン数と記憶できるパターン数との比をａとし、１パターン当たりの学習に必要な最小時間をＬ、連想に必要な最小時間をＴとする。

大規模なニューラルネットワークを時分割に拡張表現する場合は、密に結合された小部分ニューラルネットワークが分割する最小粒度となる。ここで、ハードウエアで同時に機能表現できるシナプス数をｙとすると、時分割数Ｄは、

$D=Y/y$　　である。但し、ｙはn^2の整数倍と仮定した。

また、同時に表現できる小部分ニューラルネットワークの数はy/n^2である。時分割処理によってニューラルネットワークの異なった部分を逐次表現するためには、毎回、表現するニューラルネットワーク部分に対応するシナプス荷重値を設定し直す必要がある。その設定に必要な処理時間分が時分割拡張表現のオーバーヘッドとなる。すなわち、同時に表現する全ての小部分ニューラルネットに対応する記憶パターンを学習によって記憶するのに必要な時間がそのオーバーヘッドとなる。

学習は各小部分ニューラルネットワーク毎に並列に実行できるので、学習に必要な時間は$a \times n \times L$となり、演算速度Sは拡張表現しない場合の　$S0 = y/T$　に対して、

$S = y/(a \times n \times L + T)$　　と低下することになる。

また、時分割表現するためには各小部分ニューラルネットワーク毎の記憶パターンを保持するメモリー機能が別途必要になる。そのメモリー容量Mは、

$M = D \times y/n^2 \times a \times n \times n = Y \times a$　　となる。

ここで、$S0 = 10^{18}$CPS、$L = 3T \times 30$、$a = 0.1$、$n = 10^3$　とすると、

$S = O(10^{14})$ CPS、　$M = O(10^{12})$ Byte　となる。

時分割で表現する部分ニューラルネットワーク毎に学習によって記憶パターンを設定する、この時分割規模拡張表現方式は、各シナプス荷重値を直接設定する方式と比べて、第6章で述べたように、より少ない時間で実質的に目的のニューラルネットワークを形成することができ、時分割表現のオーバーヘッドを少なくすることができる。また、別途必要となるメモリー容量は、小部分ニューラルネットワーク毎にそのシナプス荷重値を保持する場合と比べて、シナプス荷重値のビット精度をBとした場合の$M = Y \times B$より、約二桁程度少なくすることができる。但し、時分割表現のオーバーヘッド量はnに比例して増加することに注意する必要がある。

謝辞

本論文を結ぶにあたり、終始懇切なるご指導と御鞭撻を賜った東京大学大学院工学系研究科計数工学専攻 合原一幸 助教授に深く感謝致します。

また、本論文をまとめるに際し、詳細な御検討と貴重な御指示を賜りました、東京大学大学院工学系研究科機械情報工学専攻 吉澤修治 教授、同大学先端科学技術研究センター 岡部洋一 教授、南谷崇 教授、同大学大学院工学系研究科計数工学専攻 石川正俊 助教授、に厚く御礼申し上げます。

本研究遂行にあたってご指導と御鞭撻を賜り、また本論文作成の機会を与えて頂くと共に激励頂いた三菱電機株式会社先端技術総合研究所ニューロ応用技術部部長 久間和生 博士、同社半導体事業本部半導体営業企画部パワーＰＣプロジェクトプロジェクトマネージャー 茅野晋平 博士には心から感謝致します。

本研究の遂行、および論文の作成にあたり、直接御指導頂き、数々の御教示を頂いた三菱電機株式会社先端技術総合研究所ニューロ応用技術部先端LSI設計技術グループマネージャー 小守伸史 博士、同社先端技術総合研究所ニューロ応用技術部フォトニックLSI技術グループマネージャー 太田淳 博士、同社鎌倉製作所CCV事業推進センター 森伯郎 参事、同社半導体事業本部液晶事業推進部開発部 篠原尋史 課長、同社ULSI開発研究所システムLSI先端回路開発部第２グループマネージャー 益子耕一郎 博士、同社システムLSI事業化推進センターDTVプロジェクト受像端末グループマネージャー 岡田圭介 博士、に心から感謝致します。

また、本論文における数々の実験とその分析、解析にご協力頂き、有益な御討論と御指摘をしていただいた応用地質株式会社関西事業本部技術センター 村崎充弘 氏、三菱電機株式会社先端技術総合研究所ニューロ応用技術部 近藤由和 博士、小柴優一 氏、同社システムLSI事業化推進センター 近藤晴房 博士、同社ULSI開発研究所 野谷宏美 氏、前田敦 氏、同社半導体基盤技術統括部 山田強 氏に心から感謝致します。また、同社先端技術総合研究所ニューロ応用技術部の各位に御礼申し上げます.

最後に、本研究遂行における妻 昭子の理解と協力に感謝します。

研究発表リスト

［本論文に関連する論文］

1 Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," IEEE, Journal of Solid-State Circuits, Vol.26, No.4, pp. 607-611, April, 1991.

2 Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S. Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," IEEE, Journal of Solid- State Circuits, Vol.26, No.11, pp. 1637-1644, Nov.,1991.

3 Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40K Synapses," IEEE, Journal of Solid-State Circuits,, Vol.27, No.12, pp.1854-1861, Dec.,1992.

［その他の研究論文］

1 M. Okabe, M. Tatsuki, Y. Arima, T. Hirao, and Y. Kuramitsu,"Design for Reducing Alpha-Particle-Induced Soft Errors in ECL Logic Circuitry," IEEE, Journal of Solid-State Circuits,, Vol.24, No.5, pp.1397-1403, May 1989.

2 Y. Kondo, Y. Koshiba, Y. Arima, M. Murasaki, T. Yamada, H. Amishiro, H. Mori, and K. Kyuma,"A 1.2GFLOPS Neural Network Chip for High-Speed Neural Network Servers," IEEE, Journal of Solid-State Circuits, Vol.31, No.6, pp.860-864, June,1996.

3 S. Komori, Y. Arima, Y. Kondo, H. Tsubota, K.Tanaka, and K. Kyuma,"A 3.2GFLOPS Neural Network Accelerator," IEICE, Transactions on Electronics, Vol.E80-C, No.7, pp.859-867, July,1997.

［国際会議発表］

1 Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano,"A Self-Learning Neural Network Chip with 125 Neurons and 10K Self-Organization Synapses," in Symp. VLSI Circuits, Digest of Technical Papers, pp. 63-64, June 1990.

2 Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Notani, H. Kondoh, and S.

Kayano,"A 336 Neuron, 28K Synapse, Self-Learning Neural Network Chip with Branch-Neuron-Unit Architecture," in ISSCC, Digest of Technical Papers, pp. 182-183, Feb. 1991.

3 K. Mashiko, Y. Arima, M. Murasaki, and S. Kayano,"Silicon Implementation of Self-Learning Neural Networks," in ISCAS, Digest of Technical Papers, pp. 1279-1282, J une 1991.

4 Y. Arima, M. Murasaki, T. Yamada, A. Maeda, and H. Shinohara,"A Refreshable Analog VLSI Neural Network Chip with 400 Neurons and 40 K Synapses," in ISSCC, Digest of Technical Papers, pp.132-133, Feb. 1992.

5 M. Murasaki, Y. Arima, and H.Shinohara,"A 20 Tera-CPS Analog Neural Network Board," in IJCNN, Digest of Technical Papers, pp. 3027-3030, Oct. 1993.

6 Y. Kondo, Y. Koshiba, Y. Arima, M. Murasaki, T. Yamada, H. Amishiro, H. Shinohara, and H. Mori,"A 1.2GFLOPS Neural Network Chi p Exhibiting Fast Convergence," in ISSCC, Digest of Technical Papers, pp.218 -219, Feb. 1994.

7 Y. Kondo, Y. Koshiba, Y. Arima, M. Murasaki, T. Yamada, H. Amishiro, H. Mori, and K. Kyuma,"A Digital Neural Network Chip Exhibiting Fast Convergence," in IIZUKA, Digest of Technical Papers, pp. 563-564, Aug. 1994.

［研究会報告］

1 有馬裕，益子耕一郎，岡田圭介，山田強，前田敦，近藤晴房，茅野晋平，”125ニューロ,10,000シナプス搭載学習機能付きニューラルネットチップ、”電子情報通信学会技術研究報告、CPSY90-72,ICD90-128、pp.57-62、1990年10月。

2 村崎充弘，有馬裕，益子耕一郎，岡田圭介，山田強，前田敦，近藤晴房，茅野晋平，”336ニューロ,28Kシナプス搭載学習機能付きニューラルネットチップとBranch-Neuron-Unitアーキテクチャ”電子情報通信学会技術研究報告、ICD91-108, pp.79-85、1991年9月。

3 有馬裕，村崎充弘，山田強，前田敦，篠原尋史，”リフレッシュ機能内蔵400ニューロ,40,000シナプス搭載アナログニューロチップ、”電子情報通信学会技 術研究報告、ICD92-16、pp.31-38、1992年5月。

4 有馬裕，”学習機能搭載ニューロチップ、”電気学会技術報告、第602号、赤外線知能化技術の動向、pp.31-34、電気学会、1996年6月。

［雑誌等解説記事］

1 有馬裕、益子耕一郎、”ニューロコンピュータへの挑戦、第6,7回-Si-LSIニューロチップ(1),(2)、pp.88-89、pp.120-121、”電子材料、工業調査会、1990年2月、3月。

2 有馬裕、”学習機能付きニューラルネットチップ、”コンピュータ.シミュレーション、Vol.1-4、pp.50-54、コンピュータエンジニアリング社、1990年12月。

3 有馬裕、”ニューロチップは実用レベルへ、”エレクトロニクス、第36巻8号、pp.58-62、オーム社、1991年8月。

4 有馬裕、村崎充弘、山田強、前田敦、”学習機能付きニューロチップ、”三菱電機技報、Vol.66, No.2, pp.20-23、三菱電機技報社、1992年2月。

5 有馬裕、”学習機能付きニューロチップ、”トリプルＡ、No.47、pp.4-7、三菱電機、1992年11月。

6 有馬裕、益子耕一郎、”VLSIニューロチップ”、第5章、久間和生、中山高、編著、ニューロコンピュータ工学、工業調査会、1992年2月。

7 有馬裕、”ニューロＬＳＩ、”BREAKE THROUGH、1993年10月号、pp.9-11、リアライズ社、1993年10月。

8 有馬裕、田中健一、太田淳、森伯郎、久間和生、”ＶＬＳＩニューロチップ最近の話題、”システム制御情報学会誌第38巻第8号、pp.423-429、システム制御情報学会、1994年8月。

9 有馬裕、近藤由和、小柴優一、森伯郎、久間和生、”ＶＬＳＩニューロチップ、”三菱電機技報、Vol.68, No.5, pp.696-701、三菱電機技報社、1994年8月。

10 K. kyuma, J. Ohta, and Y.Arima,"Comparison between Optical and Electical Implementation of Neural Networks", International Conference on Solid State Device and Materials, August 23-26, 1994, Pacifico Yokohama, Yokohama, Japan

11 Y. Kondo, Y. Koshiba, Y. Arima, T. Yamada, H. Amishiro, H. Mori, and K. Kyuma,"A Digital Neural Network Chip for High-Speed Neural Network Servers," IEEE, Denshi Tokyo, No. 33, pp. 197-200,1994.

12 小柴優一、近藤由和、有馬裕、森伯郎、久間和生、”超高速ニューロプロセッサーの特徴と応用、”pp.97-102、電子材料、工業調査会、1995年1月。