コグニティブ・チップ
コグニティブ・コンピューティング時代のハードウェア技術

さまざまな情報源からの大量データを統合し、瞬時に分析するコグニティブ・コンピューティングを実現する基本的アプローチは、大きく2通りに分けられます。一つ目は、汎用コンピューターなど、従来のノイマン型コンピューターを使ったソフトウェアによるアプローチです。そして二つ目は、脳内の神経ネットワークを模倣した新たなハードウェア・アーキテクチャーによるアプローチです。

本稿では、この新たなハードウェア・アーキテクチャーの実現を支える、次世代ハードウェア・コア技術について解説します。

1. 緒言

現在、コンピューターの世界は三つの大きな変曲点を迎えています。一つ目は米国のクイズ番組で自然言語による質問応答システムを実現した「IBM Watson」です。これは、コンピューターと人間の間に存在していた大きな一つの壁であった「言語」と「判断」において、コンピューターが人間に歩み寄る形でコグニティブ・コンピューティング時代の到来を示す大きなブレークスルーとなりました。二つ目はインターネット、ソーシャル・メディア、スマート・デバイス、各種センサーなどがもたらす世界規模の情報量の著しい増大です。こうしたビッグデータを処理し、活用することで、より高度なサービスの提供が可能になることが広く理解され始めました。三つ目は半導体の微細化の物理限界です。過去数十年にわたってスケーリング則に従って継続してきた現在の半導体の微細化技術が、間もなく物理的限界に到達しようとしています。

われわれを取り巻くこれらの大きな変化の中で、増大し続ける大量のデータを高度に処理するためには、コンピューティング・パワーを引き続き向上するための新たな技術開発に注力する必要があります。本稿では、今まさに到来しつつある新しいコグニティブ・コンピューティングの時代において、それを取り巻く環境や市場ニーズと、その実現を支えるコンピューター・ハードウェアの次世代基盤技術について解説します。

2. コグニティブ・コンピューティング時代のハードウェア技術

長い歴史を持つコンピューターの世代は三つの世代に分類できます。第1世代のコンピューターはデータを数えるための計算機として誕生し、第2世代ではOSやソフトウェアが作られ、プログラムによって動くコンピューターへと進化してきました。そして今、コンピューター自身が学習する第3世代へと移ろうとしています。

表1. Watsonと人間の脳の比較

3.55GHz	Clock	～10Hz
200kW	Power	20W
12000L(10racks)	Volume	1.2L
4×10^3 (bumps/cm^2)	Joint density	2×10^8 (synapses/cm^3)

コグニティブ・コンピューティングとは、この第3世代に相当するものです。米国のクイズ番組に挑戦したIBMの質問応答システムWatsonは、インターネットに接続されていない自己完結システムです。その内部には、約100万冊の本に匹敵する自然言語で書かれた15TBのメモリーと、2880個ものプロセッサー・コアが存在し、ラック10本からなるコンピューターとして200kWの電力を必要とします。この電力は人間の脳の約1万倍の消費電力に相当しますが、出題された問題を即座に解釈し、メモリー内の情報を分析し、短時間で最も適した解答を導き出す質問応答システムとして実現されました。Watsonの成功は、現在のITテクノロジーを結集すれば、人間の思考のアプローチとは全く異なる方法で曖昧な自然言語特有の問題をコンピューターが扱うことができるという展望をもたらしました。

しかしその一方で、現実の脳との間には、その消費電力、容積、および接続密度において大きなギャップが存在することも明らかとなりました（**表1**）。つまり、ビッグデータから短期間に必要な情報を抽出し意思決定を行うためには、従来型のシステムでは電力、ハードウェア、ソフトウェアなどの膨大なリソースを駆使したデータ処理が必要であり、さらなる大規模化には電力削減やコスト面での改善が課題となります。この問題を抜本的に解決するためにはソフトウェア的なアプローチだけでなく、ハードウェア・アーキテクチャーの観点からのアプローチが必要となります。

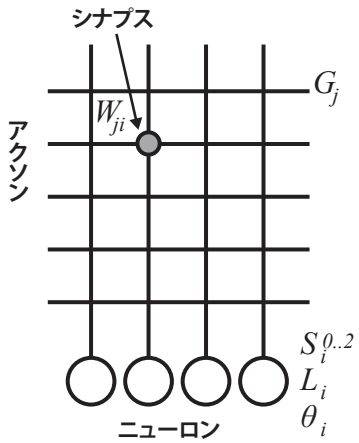
3. SyNAPSEプロジェクト

SyNAPSEプロジェクトは、米国国防省高等研究計画局の助成金を受けて進めているプロジェクトです。ここでは、本プロジェクトで開発した半導体チップの概要を説明します。

このチップは、神経細胞の情報伝達の仕組みをモデル化した積分発火型ニューロン・モデル[1]に基づく、新たな並列処理アーキテクチャーを採用しています。

図1にチップ・コアの基本構成図を示します。図中のアクソン、シナプス、ニューロンは神経細胞の名称ですが、実際に各部分が、それらの脳神経細胞を模倣した機能を担当しています。入力情報は並列スパイク信号の形式で、アクソンを通してシナプスに入力されます。シナプスでは結合の有無を判断し、選ばれた信号が各ニューロンに伝達されます。ニューロンで情報処理がなされ、出力が次段のアクソンによって伝達されます。

図1の表に示した項目は、ニューロンで処理する積分発火型ニューロン・モデルに基づく回路に使用される各パラメーターとその数値範囲を表しています。256個の各ニューロンは、それぞれ独立した1536のパラメーターと、1024個のアクソンおよびシナプスによる並列入力情報を既定の周期ごとに積算し、リーク情報を加えた結果を256の閾値（しきいち）と比較し、その結果を既定の周期ごとに出力します。25万6000個のシナプスは、SRAMアレーで、256個のニューロンは、デジ



Name	Description	Range
W_{ji}	Connection between axon j and neuron i	0,1
G_j	Axon type	0,1,2
$S_i^{0..2}$	Synapse values	-256 to 255
L_i	Leak	-256 to 255
θ_i	Threshold	1 to 256

図1. チップ・コア構成図

タルCMOS回路で実現しています。

図2にチップ写真を示します。2mm×3mmの面積で、1出力あたり45ピコ・ジュールの低消費電力での並列処理を実行します。本チップは、クロック同期のデジタル回路で構成されています。また、各パラメーターもバイナリーコードで表します。従って、ソフトウェアによるモデリング、シミュレーション環境とハードウェアとの一対一対応を可能にしています。

既知のアルゴリズムに基づく、ソフトウェアでのオフライン学習で事前に摘出されたパラメーターを、本チップにマップすることで、例えばPongゲームのような初歩的ビデオ・ゲームを**図2**に示したテスト・ボード上において自身で動作させることができます。

本プロジェクトでは、現在、大規模化に向けたソフトウェアを含む、統合された取り組みが進行中です。8月に発表された第2世代チップは、100万個のニューロンを有しています。これは基本的には、本章で説明したチップ・コアをチップ上に4000個搭載し、それらを2次元メッシュ・ネットワーク接続で構成したもので、54億個のトランジスターからなる大規模チップです。また、チップ間シームレス接続機能を有し、ボード上でのさらなるスケーラビリティーをサポートしています。並行して開発中のプログラミング言語、シミュレーターと併せて、新たなアーキテクチャーによるニューロシナプティック・システムに取り組んでいます。

4. ニューロモーフィック・チップ

本章では、自己学習機能を持ち、さらなる高密度化、低消費電力を目指したチップの取り組みを紹介します。

神経科学の分野ではスパイクタイミング依存可塑性学習則（STDP）[2]が研究されています。これは、ニューロン同士を結合するシナプスにおいて、前段ニューロンからアクソンを経由して入力されるスパイクと、後段ニューロンの発火スパイクとの時間的関係でシナプス結合荷重が変化する、脳での記憶の基礎現象です。

本チップでは、このスパイクタイミング依存可塑性を有する自己学習機能をシリコン上に実現しました。シナプス結合荷重の変化は、新たに開発したナノ・スケール記憶素子の端子間に印加する電位差により、アナログ的に変化するコンダクタンスで実現しています。また、CMOSアナログ回路で非線形スパイク波形を生成しこの記憶素子に印加することで、脳神経のSTDPを模倣した機能を実現しています。

図3にチップの構成図を示します。シナプスは、ナノ・スケール記憶素子のクロスバー・アレーで構成されています。ある時点で任意のニューロンから出力されたスパイク信号は、ネットワーク回路を経由し接続先アクソン・ドライバーに届き、ここで非線形スパイク波形が生成され、

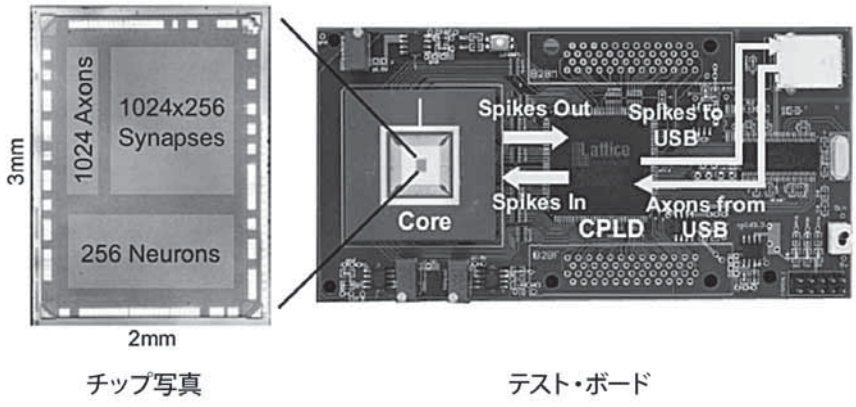


図2. チップ写真とテスト・ボード

そのアクソン・ライン上のシナプスに印加されます。また、別のニューロンでは、自身のスパイク発生時に、そのデンドライト・ドライバーが非線形スパイク波形を生成し、そのデンドライト・ライン上のシナプスに印加されます。その結果、この二つの非同期スパイク信号の相対的タイミングに依存したシナプスのコンダクタンスが更新されます。このようにスパイク信号による刺激を繰り返すことで、自己学習がオンラインで実行されます。

また、スパイク信号を生成するニューロンの機能は、積分発火型ニューロン・モデルに基づくアナログ回路で実現しています。スパイク信号を受けたシナプスはシナプス電位をニューロンに伝達します。このシナプス電位の時間的な総和は、ニューロンの内部電位として蓄積され、その電位が発火閾値を越えると、ニューロンから新たなスパイク信号が生成されます。自己学習と並行して連想認識処理も実行されます。例えばイメージ認識では、マトリックス上イメージ情報を入力すると、自己学習により記憶された類似するイメージに対応するユニークな組み合わせのニューロンが、頻繁にスパイクを出力します。このように、可塑性を有するナノ・スケール記憶素子、アナログ・スパイク生成回路、アナログ積分回路、デジタル・ネットワーク回路を組み合わせることで、脳神経細胞の動作を模倣した情報処理ハードウェアをシリコン上に実現しています。また、2端子構成のナノ・スケール記憶素子の不揮発性、CMOSプロセスとの親和性から、高集積度、大容量化、低消費電力化に向けての進展を可能にします。

5. 東京基礎研究所における取り組み

これまで述べてきたように、米国主導でのプロジェクトに参加してこの分野の進展に貢献することと並行して、東京基礎研究所独自の研究活動も進めています。

われわれが考えるこの分野の研究における今後の進むべき方向性を**図4**に示します。前世代の研究は個々のニューロンをどのように構成するか、また現世代の研究はシナプス結合をどのように集積するかに重点を置いていました。それに対して、次世代の研究テーマは、より集積化が進んだ状況を想定して、ニューロン間を大域的に結合するアクソンに焦点をあてて研究していくことが重要になると考えています。

このような観点から、米国主導で進められているSyNAPSEプロジェクトとは別に、IBM内では東京基礎研究所がリードする形で、2012年より東京大学との間で社会連携講座をスタートさせ、次世代情報処理システムの共同研究を始めています。この共同研究プロジェクトでは、次世代の計算機システムを念頭に、特にこれまで人間が得意としてきたコグニティブな情報処理のさらなる低消費電力化を目指すべく、インターコネクトに

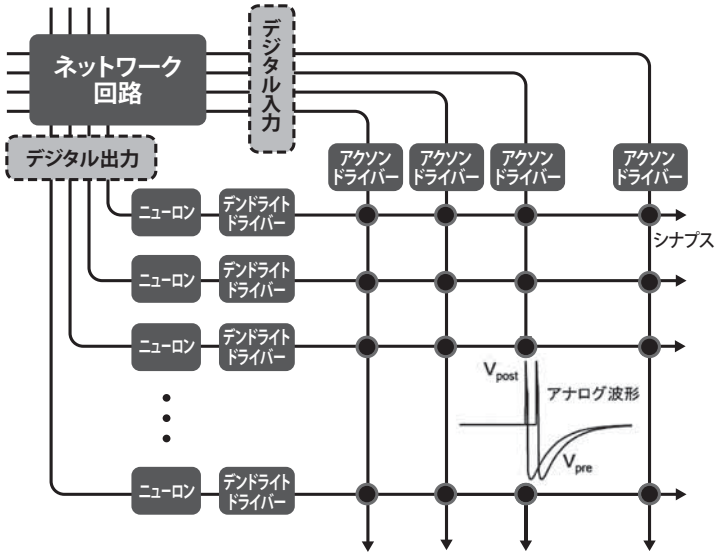


図3. チップ構成図

関する制約をアルゴリズムとデバイスの両方から攻めるアプローチを採っています。今年1月には、東京大学山上会館カンファレンス・ホールにて、「Bio-Inspired Energy-Efficient Information Systems」というタイトルでワークショップを行い、大学や企業のさまざまな専門分野から100名近くの参加者を集め、パネル討論などを含めて活発な議論を行いました。

人間の脳には、およそ10の10乗個のニューロンが存在しており、そのおのおのがアクソン・シナプス経由で複雑につながったネットワークを形成しています。このようなネットワークをそのまま現在のVLSIの技術で実現しようとすると、インターコネクトのボトルネックに直面します。脳は3次元的に集積されているにもかかわらず、現在のVLSIの配線はチップや基板上に2次元的に配置されていることが大きな要因の一つです。1次元的に入力と出力をそれぞれX方向とY方向に並べた場合、2次元的に配置された配線リソースを使い切ってしまいます。3次元実装を使ったとしても、メモリーのように局所的で規則的な配線をすることはできても、脳のようなより複雑に絡みあった配線を実現するのは容易ではありません。限られた配線リソースを時分割多重（例えば、パケット多重）で使ったとしても、パケットの局所的な集中により、全体の処理速度が大きく影響されます。

このような課題に対してわれわれは、アルゴリズムだけで問題を解決しようとしたり、ハードウェア・デバイスだけで問題を解決したりするのではなく、アルゴリズムとハードウェア・デバイスの両方の観点から研究を進めていくことが重要だと考えています。東京大学との共同研究においても、アルゴリズムからのアプローチとしては、ヘブ則ベースの認識アルゴリズムに対して、学習アルゴリズムとインターコネクト制約を同時に最適化することにより、密結合を前提としない疎なインターコネクトにおいても効率良く処理する方法を研究しています。また、デバイスからのアプローチとしては、より多くのノードと同時に結合ができる方式を検討しており、基礎的な研究課題ですが、波動を用いた情報伝送デバイスの研究を進めています。

さらに、東京基礎研究所単独の研究テーマとして、自由空間通信を使ったインターコネクトの可能性も検討しています[3]。例えば、**図5**に示すように、Deep Believe Networkに代表される階層型のニューラルネットワーク構造に対して、モジュールもしくはボード間を、光自由空間伝送を用いることで、アクソンによる非局所的な伝送をうまく現在のVLSI技術に取り込むことを研

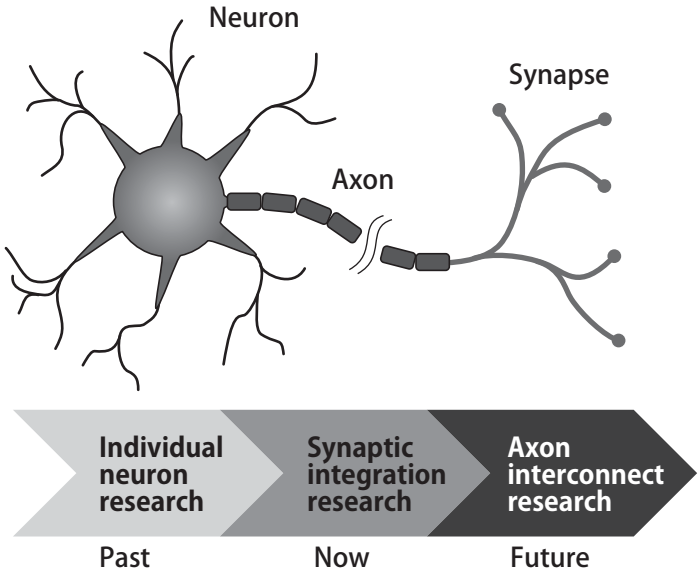


図4. コグニティブ・ハードウェア研究の過去、現在、未来

究しています。自由空間を伝送する光の向きをコントロールすることにより、ニューロン間の結合を変更することも可能となります。

このような機構を入れることにより、隣接モジュール間での信号伝送だけではなく、より離れたモジュール同士が直接信号をやり取りすることが可能になります。ニューロン数が増大した場合でも、離れたニューロン間を直接自由に結合する機構を入れることにより、隣接モジュール間での多段ホッピング経由でニューロンを結合する場合と比較して、実効バンド幅の効率的な活用が可能となると考えています。

いずれにしても、この分野は多角的な観点からのさまざまなアプローチが必要であり、日本の電子産業の復活のシナリオとしての面からも、日本独自の研究活動が望まれる分野であると考えられます。

[参考文献]
[1] A.L.Hodgkin and A.F.Huxley, “A quantitative description of membrane current and application to conduction and excitation in nerve,” J.Phys., vol.117, pp.500–544, 1952
[2] H.Markram, J.Lubke, M.Frotscher, and B.Sakmann, “Regulation of synaptic effi-cacy by coincidence of postsynaptic APs and EPSPs,” Science, vol.275, pp.213–215, 1997
[3] Y.Katayama, T.Yamane, and D.Nakano, “An Energy-Efficient Computing Approach by Filling the Connectome Gap,” The International Conference on Unconventional Computation and Natural Computation (UCNC), 2014

日本アイ・ビー・エム株式会社
東京基礎研究所
シニア・テクニカル・スタッフ・メンバー

細川 浩二
Kohji Hosokawa

日本IBM入社以来、半導体回路設計、メモリー製品開発に従事。マイクロ・エレクトロニクス部門のASICデザインセンター担当を経て、2012年より、東京基礎研究所、サンエンス・テクノロジー、半導体開発センター担当に就任。

日本アイ・ビー・エム株式会社
東京基礎研究所
シニア・テクニカル・スタッフ・メンバー

片山 泰尚
Yasunao Katayama

日本IBM入社以来、半導体物理、回路設計、メモリーシステム、情報理論、光インターコネクト、ミリ波システムなどの研究に従事。2012年より、東京基礎研究所、サイエンス&テクノロジー、サーキット&システムズ担当。IBMアカデミー・オブ・テクノロジー・メンバー。IEEEシニアメンバー。

日本アイ・ビー・エム株式会社
東京基礎研究所
シニア・テクニカル・スタッフ・メンバー

折井 靖光
Yasumitsu Orii

日本IBM入社以来、大型コンピューター、ノートブック、ハードディスク・ドライブの実装技術に従事。2009年より、東京基礎研究所へ異動、3次元積層デバイスプロジェクトをリード。2012年、サイエンス&テクノロジー担当、ハードウェアの基礎研究を統括。IBMアカデミー・オブ・テクノロジー・メンバー。エレクトロニクス実装学会理事。

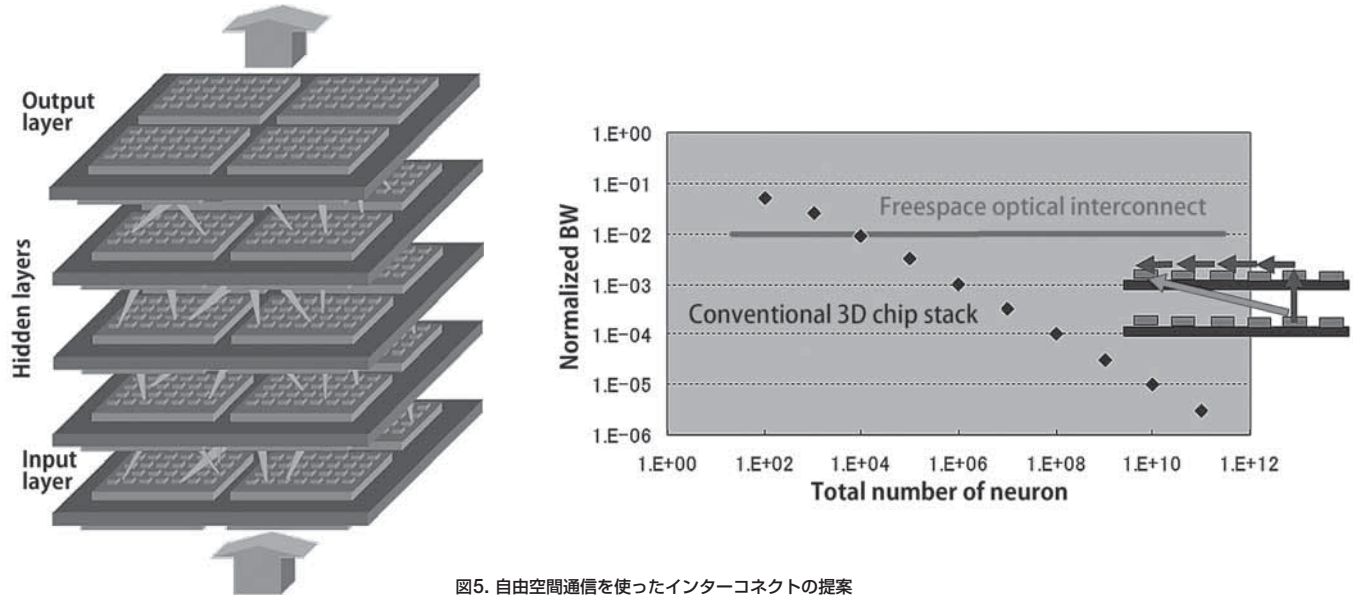


図5. 自由空間通信を使ったインターコネクトの提案