

ニューロモルフィックシステムと物理デバイス Neuromorphic systems and physical devices

森江 隆
Takashi MORIE

九州工業大学 大学院生命体工学研究科
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology

〒 808-0196 北九州市若松区ひびきの 2-4
2-4 Hibikino, Wakamatsu-ku, Kitakyushu 808-0196

e-mail: morie@brain.kyutech.ac.jp

分類番号: 3.8, 3.7

本稿では、人工知能（AI）に関連して頻繁に聞く「ニューロモルフィックシステム」や「スパイクニューロンモデル」について、その意味、歴史、現在の位置づけを紹介します。また、それらの実現に向けてどのような処理モデルが提案され、どのようなデバイスが求められているかを解説します。AI に対して応用物理の材料・デバイス分野がどのように貢献できるのか、本稿がそれを考えるヒントになれば幸いです。

1. ニューロモルフィックとは？

近年の人工知能（Artificial Intelligence: AI）ブームの波に乗ってしばしば耳にする「ニューロモルフィック」とは、1980年代に提唱された概念で、もともと、脳の神経回路における非線形な電氣的振る舞いを電子回路（集積回路）でまねて、脳の情報処理機能の解明に役立てるとともに、効率のよい情報処理システムを構築しようとする考え方です。ただし、以下に詳細に説明しますように、広く応用が進展している深層学習ベースの AI では脳の神経細胞とその結合回路を非常に単純化したモデル（アナログニューロン）が使用されるのに対し、ニューロモルフィックの立場では、神経細胞がスパイクパルスを出力することなど、より忠実に振る舞いをまねるモデル（スパイクニューロンなど）を用いることが特徴で、互いに異なる立場にあります。ニューロモルフィックに似たような用語として、ニューロインスパイアド（neuro-inspired）^{†1}があります。前者はより忠実に生体の神経回路や脳をまねようとする立場であり、後者はそれらにヒントは得るけれども必ずしも忠実にまね

る必要はないという立場で、後者のほうがより緩い定義といえます。なお、「ニューラルネットワーク」という用語は、関連学会の名称にも含まれるように、本来は上記の全てを包含する広い概念を指しますが、上記の単純化されたモデル（人工ニューラルネットワーク）とその応用分野を指して用いられることも多いようです。

2. アナログニューロンモデルとスパイクニューロンモデル

さて、近年の AI 技術の中心になっている深層学習は脳の仕組みをまねているといわれますが、実際には生体の脳をどの程度忠実にまねているのでしょうか？

生体の神経細胞（ニューロン）は、細胞体内にカリウムやナトリウムなどのイオンがチャネルを通して出入りすることで、膜電位が変化します。膜電位とイオン電流の関係は非線形な特性をもっており、あるしきい値を超えると「スパイク」と呼ばれる非常に幅の狭い膜電位のパルスを出力します。これを生理学用語で「発火（fire）」と呼びます。ニューロン同士はシナプスと呼ばれる結合部を介して接続されており、ほかのニューロンからスパイクが入力されると、シナプスでの結合重みに応じた影響（シナプス後電位）がニューロン内部で生成

^{†1}ニューロモルフィックとニューロインスパイアド これらの用語に対応して最近脳をまねるという意味を強調するためにブレインモルフィックやブレインインスパイアドという用語も用いられますが、同じ意味合いです。

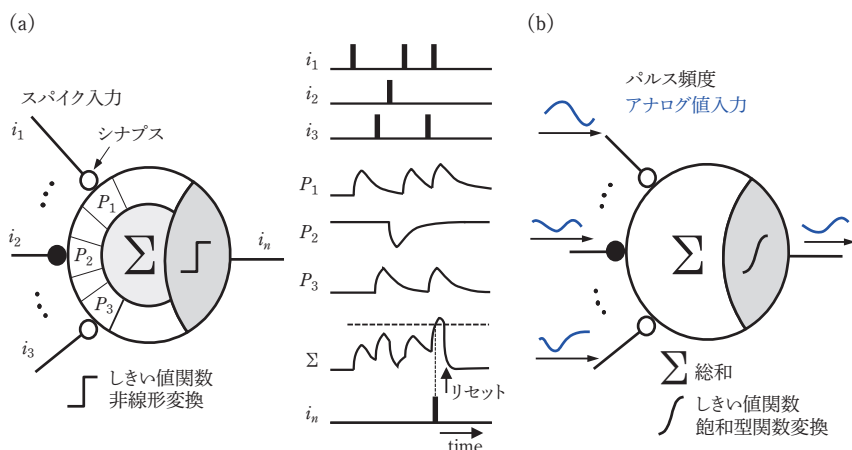


図1 (a)スパイクニューロンモデル（積分発火型）と、(b)発火頻度を表現するアナログニューロンモデル。

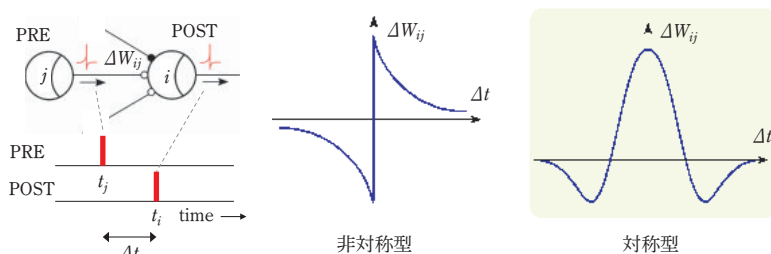


図2 STDPの概念図。前段ニューロン j (PRE) と後段ニューロン i (POST) の間のシナプス結合重みの変化 ΔW_{ij} が、それぞれの出力スパイクのタイミング差 Δt に依存するという関係を表しています。非対称型では PRE スパイクが POST スパイクよりも先にシナプス部に到着すると結合重みが増加し、逆であれば減少します。これにより因果関係を学習するといわれています。一方、対称型はスパイク時間差の絶対値に依存し、相関関係を学習するといわれています。

され、時空間的に加算されます。

イオンチャネルを含むニューロンの内部もシナプス部も、詳細にみると極めて非線形な特性をもっていますが、単純化すると、図1(a)に示すように、スパイクパルス i_k ($k=1,2,3, \dots$) によるシナプス後電位 P_k が特定の時定数で減衰する期間に、多くの入力スパイクで加算 (Σ) されて、発火しきい値を超えるとスパイクを出力するという単純化されたモデル（積分発火型モデル）が考えられます。このようにスパイク信号を考慮したニューロンモデルを「スパイクニューロン」と呼び、それを用いて構成したネットワークをスパイクニューラルネットワーク (Spiking Neural Network: SNN) と呼びます。これには生体への忠実度に応じたさまざまなモデルが提案されています。

一方で、神経細胞の発火スパイクパターンを観測すると、非常にランダムで1つひとつのスパイクに情報が乗っているよ

うに思えないことから、スパイクの発火頻度に意味がある（情報が乗っている）と考えたのが、図1(b)に示すように、「発火頻度モデル」または「アナログニューロンモデル」であり、発火頻度を表現するアナログ値で入出力を表し、シナプスも単純な重み付けによる積和演算で表現します。これが現在の深層学習で一般的に用いられているモデルで、人工ニューラルネットワーク (Artificial Neural Network: ANN) と呼ぶことがあります。

脳の仕組みをまねて、知的な情報処理を行わせようとする試みは、歴史的には上記の発火頻度モデル（アナログニューロンモデル）が、その単純性のために1980年代の第2次AI（ニューロ）ブームまで盛んに研究されました。しかし、脳の中のカオスを研究していた研究者らは1980年代からすでにスパイクパターンによる情報表現の重要性に気付いていました。その後、多くの研究者がスパイク1個1個に意味がありそうだと気付いたのは、スパイクタイミングでシナプス強度が変化するスパイクタイミング依存シナプス可塑性 (Spike-Timing Dependent Synaptic Plasticity: STDP)¹⁾ (図2) が発見されたことが大きな契機といえます。

2000年頃から、スパイクタイミングを考慮することにより、時間領域での演算を組み込むことができ、より高度な情報処理ができると期待され、研究が活発化しました。しかし現状では、実用的なAI処理を実行するには、あえてSNNモデルを用いる必要はない、というのが多くのAI研究者の見方ようです。ただし、脳をより模倣した次世代の脳型AIについては、スパイク同士の同期性が重要であるとも考えられるため、SNNを用いたブレイクスルーが起こる可能性が高いと思われます。また、ほとんどのAI研究者の見方は、あくまでデジタル計算機での実行を想定したうえでのことです。SNNは、現状のデジタル計算機ではなく、新しい脳型アナログ計算機こそ、その真価が発揮できるものと思われます。

ハードウェア化の観点からのスパイク（パルス）を用いる利点は、

- ①スパイクが2値であり、ニューロン間にアナログ値を伝送する必要がないため、アンプではなく電力消費の少ないコンパレータ回路を用いることができる。そのため回路構

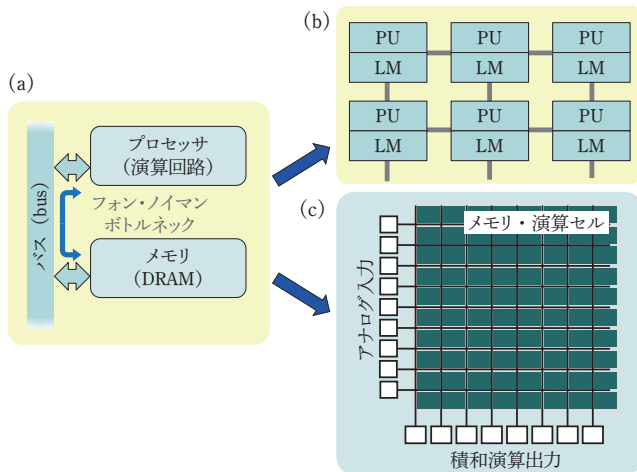


図3 (a), (b)フォン・ノイマンアーキテクチャ, (c)イン・メモリ計算モデル。

成が単純になり、低消費電力化が可能になること、

②スパイクの到達により演算を開始させる「イベントドリブン型」の演算ができ、通常のデジタル回路のようなクロック同期方式をとる必要がなくなり、回路構成の単純化と低消費電力化に適すること²⁾、

③時間領域動作すなわち過渡動作を利用することで、直流動作回路よりも低消費電力化が可能になること（第4章参照）、

④時間領域を利用することで、非線形な電圧/電流波形を利用して非線形変換を実行できること³⁾、

などが挙げられます。以上により、大規模で非線形なネットワーク回路が実現できます。

3. 現状のデジタル計算機の限界

現在のデジタル計算機は、図3(a)に示すように、演算を行うプロセッサと、データやプログラムを記憶するメモリ（主記憶装置）が別チップで構成されており、両者をバス配線で接続する、フォン・ノイマン（von Neumann）アーキテクチャと呼ばれる構成を取っています。そのプロセッサ/メモリが高速化・低消費電力化されたことから、バス部分での配線容量の充放電による通信帯域の制約や消費電力の増大が顕著になってきました。これはフォン・ノイマンボトルネックと呼ばれます。これを低減するために、図3(b)に示すように、GPU（Graphics Processing Unit）や専用のAIプロセッサではチップ上で小規模のプロセッサユニット（PU）やローカルメモリ（LM）を多数並列配置する構成が採用されています。さらにこの問題を解消するために図3(c)に示すように、メモリ上で演算を行うイン・メモリ計算（In-memory computing）と呼ば

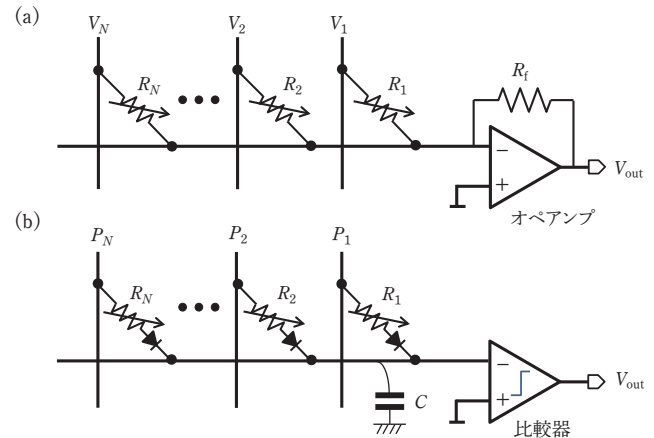


図4 積和演算を実現する回路方式：(a)電圧モード・アナログ方式、(b)時間領域アナログ方式。

れる手法が盛んに検討されるようになってきました⁴⁾。

一方で、ニューラルネットワークのアナログ回路実装においては、1980年代より図3(c)に示すような直交した配線の交点部分にメモリを配置したクロスバー構成で積和演算を行う方式が取られており、これはまさにイン・メモリ計算の手法にほかなりません。

4. AIハードウェアに必要な積和演算機能とそのアナログ回路での実装法

ここでは筆者らがSNNモデルにヒントを得て提案している、時間領域において積和演算を高効率で実行する回路構成を紹介し^{5,6)}。

通常のアナログ方式で積和演算を行う方法としてよく用いられるのは、図4(a)に示すように抵抗を用いた電圧モード直流回路方式です。オペアンプを用いると、抵抗の逆数（コンダクタンス）を重みとして、アナログ電圧入力に対して、積和演算した結果を電圧で正確に取り出すことができます。この方法は1984年の有名なHopfieldの論文ですでに用いられています⁷⁾。Hopfieldは、この回路で得られた積和演算結果をシグモイド関数（S字状の飽和型連続非線形関数）で処理したあと、入力にフィードバックすることで、連想メモリや最適化問題を解くことができることを示し、その後のニューラルネットワーク研究に大きな影響を与えました。しかし、この種の回路は、オペアンプの消費電力が大きくなってしまい、低消費電力動作をさせるために高抵抗を用いるとノイズの影響を受けやすいなどの課題があります。

そこで、筆者らは時間領域を用いるアナログ積和演算方式を提案しました。これは図4(b)に示すように、ステップ電圧またはパルス信号 P_k ($k=1,2,3,\dots,N$) を入力して、抵抗 -

容量 (RC) 回路の過渡応答を利用して電荷加算で積和演算を行い、比較器によりパルスに変換して出力する回路です。この場合、積和演算 1 回に消費されるエネルギーは、電圧を V 、入力数を N とすると CV^2/N となり、 C を小さくするほど低消費電力化が可能です。このためには、低容量配線技術や、低入力容量の回路が必要になります。一方で、抵抗 R は消費エネルギーには影響しませんが、時定数を決定します。時間領域で一定の時間分解能を得るために、ある程度の長い時定数 (μs オーダ) を必要とすることから、1 積和演算回路当たりの C が fF (10^{-15}F) オーダとすると、 R は $\text{G}\Omega$ ($10^9\Omega$) オーダ以上の高抵抗が必要になります。これは通常アナログ回路で用いられる抵抗値よりもはるかに高い抵抗で、これを可変かつ不揮発に保持しようとすると、高抵抗不揮発性アナログメモリ素子が必要になりますが、一般にはかなり難易度の高いデバイス開発になります。これが時間領域アナログ集積回路方式のデバイスレベルでの課題です。

なお、図 4(a) に示した回路方式は連続時間／連続状態での積和演算を実行できるので、シナプス機能を実現する方式として、アナログニューロンモデルにもスパイクングニューロンモデルにも適用できます。一方で、図 4(b) に示した回路方式はスパイクングニューロンモデルにヒントを得て時間領域で積和演算を行う方式で、アナログニューロンモデルの計算には適用できますが、しきい値処理が含まれるため、スパイクングニューロンモデルへの適用には制約があります。つまり、両方式は必ずしもアナログニューロンとスパイクングニューロンの両モデルと 1 対 1 に対応するものではないことに注意してください。

5. 新しい計算モデル・リザーバー計算と物理デバイス

ニューラルネットワークモデルには、画像認識でよく用いられている畳み込みネットワークのように入力層から出力層に一方方向に処理が進んでいくフィードフォワード型のネットワークだけでなく、後段の処理結果が前段に戻ってくるフィードバック型のネットワークもあり、リカレントニューラルネットワーク (Recurrent Neural Network: RNN) とも呼ばれます。これは時系列処理に有効で音声認識などで威力を発揮しています。RNN も 1980 年代からさまざまなモデルが提案されてきましたが、現在では LSTM (Long Short-Term Memory) が有名です。しかし、複雑で人工的なモデルであることから、デジタル計算機以外での実装はかなり困難です。

それに対して、近年、注目されている RNN モデルが、リザーバー計算 (Reservoir computing) モデル⁸⁾ です。その最も簡単な例は、図 5 に示すように、リザーバーと呼ばれるランダムなリカレント (回帰) 結合を有するネットワーク部分に、入力

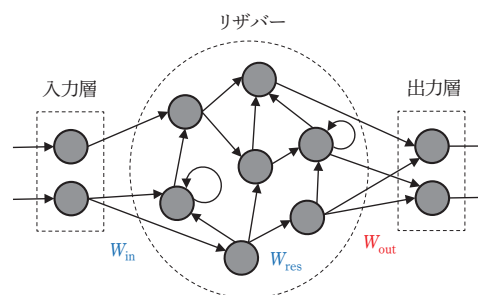


図 5 リザーバー計算モデル。

層と出力層の結合を付加したもので、学習はリザーバーと出力層間の結合荷重でのみ行われます。リザーバー部については、ユニット構成や結合に一定の制約条件がありますが、非線形で時間遅延のある処理（時間発展、すなわちダイナミクスがあること）が行われればよく、学習機能も不要なことからさまざまな実装法が可能で、自然界の物理現象を利用した物理リザーバーの提案も行われています。

リザーバー計算モデルは、その単純な構成にもかかわらず、脳の皮質のモデルとしても議論されており、単に時系列予測器という実用的な面だけでなく、脳機能の一部を実現できる可能性を秘めている有望なモデルです。したがって、このモデルの実現はデバイス・材料分野から脳型計算機構築に貢献できる大きな礎になると期待されます。なお、本技術については以降の号で詳しく説明がなされます。

6. 脳型処理モデルと物理デバイス（海馬を中心とする脳型モデルに必要とされる物理デバイス）

現在の深層学習ベースの AI は統計的機械学習手法の一種と見なすことができ、大量のデータから万人共通の認識機能を実現する技術です。しかし、人はその誕生以来さまざまな経験を重ねて、その人ならではの記憶を脳内に形成しています。この能力を AI に付与することで、個々人の経験と記憶をサポートする次世代 AI を開発することができると考えられます。実際に、最先端の AI 研究では、既存 AI モデルをプロセッサとみなしたときに、メモリに相当する装置をシステムに付加する提案⁹⁾ もされています。

一方で、ブレインモルフィックの立場から、筆者らはこれを脳の機能的な構造をまねることで実現したいと考え、脳の記憶機能をつかさどる海馬とその周辺部位（嗅内皮質）のモデル化とハードウェア化を進めています¹⁰⁾。これには、外界の情報をエピソード記憶¹²⁾として符号化・保持・統合・想起する機能が必要です。自分が今どこにいるかを「場所細胞」のネットワークで表現し、そこで何かの出来事（イベント）が起

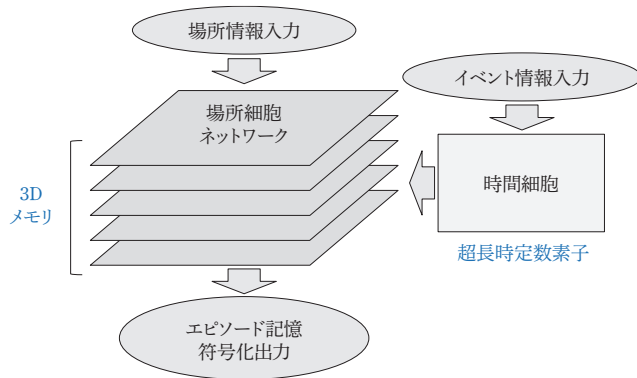


図6 脳型記憶・海馬モデル。

こったときに、感覚系から入力された情報に応答して、イベント間の時間遅れを検出する「時間細胞」を導入して、図6に示すような構成で、エピソード記憶を符号化・記憶するハードウェアの構築をまずは目指しています。それには、高密度の3次元積層（メモリ）技術や、秒オーダ以上の超長時定数を実現するデバイスなどが必要です。

7. むすび

情報系のAI研究者は主にアルゴリズムで知能を実現しようとしますが、それはとすると、正確だけれども効率の悪い既存のデジタル計算機というハードウェア上でしか実現できない可能性を有しています。一方で、脳科学者は、脳の神経回路の制約の中でその機能を説明するために、工学的には簡単に実現できる機能であっても、あえて複雑なモデルを考えることがあります。実用的で人間に迫るような新たなAIの開発には、これらのどちらか一方だけでは不十分です。脳の構造にヒントを得てそのエッセンスを取り出してモデルとアルゴリズムを考案し、それに適した構造をもったハードウェアAIを構築する必要があります。そのうえ、これらのデバイス開発は一

定の仕様を与えられて開発を進めればよいというものではなく、AI処理モデルやアルゴリズムとデバイス特性との妥協を重ねつつ開発することが必要です。そのためにはデバイス研究とAI研究を並行して、かつ協力して推進する必要があります。現状では、情報系のAI研究者がデバイスの勉強をするよりも、デバイス研究者がAIの勉強をするほうが現実的です。デバイス研究者の皆さんのAI研究への参入を大いに期待します。

文 献

- 1) G.Q. Bi and M.-M. Poo: J. Neurosci. **18**, 10464 (1998).
- 2) P.A. Merolla, J.V. Arthur, R. Alvarez-Icaza, A.S. Cassidy, J. Sawada, F. Akopyan, B.L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S.K. Esser, R. Appuswamy, B. Taba, A. Amir, M.D. Flickner, W.P. Risk, R. Manohar, and D.S. Modha: Science **345**, 668 (2014).
- 3) T. Morie, J. Fukakoshi, M. Nagata, and A. Iwata: IEICE T. Fund. Electr. **E82-A**, 356 (1999).
- 4) S.K. Gonugondla, M. Kang, Y. Kim, M. Helm, S. Eilert, and N. Shanbhag: IEEE Int. Solid-State Circuits Conf. (ISSCC), pp. 490-491 (2018).
- 5) Q. Wang, H. Tamukoh, and T. Morie: <https://arxiv.org/abs/1810.06819> (2018).
- 6) M. Yamaguchi, G. Iwamoto, H. Tamukoh, and T. Morie: <https://arxiv.org/abs/1902.07707> (2019).
- 7) J.J. Hopfield: Proc. Natl. Acad. Sci. USA. **81**, 3088 (1984).
- 8) 田中剛平：電子情報通信学会誌 **102**, 108 (2019).
- 9) A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S.G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. P. Badia, K.M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis: Nature **538**, 471 (2016).
- 10) M. Kawauchi, K. Takada, K. Tateno, and T. Morie: Proc. of 28th Annual Conf. of JNNS, 158 (2018).

(2019年3月3日 受理)



森江 隆 (もりえ たかし)

九州工業大学大学院生命体工学研究科教授。1981年大阪大学大学院理学研究科物理学専攻博士前期課程修了。日本電信電話株式会社（NTT）研究所研究員を経て、97年広島大学工学部助教授。02年より現職。博士（工学）。応用物理学会、日本神経回路学会、電子情報通信学会、IEEE等会員。NTT研究所時代より脳型ハードウェアの研究に従事。

^{†2}エピソード記憶 いつ、どこで、どのような出来事が起こったかという個人の体験に関する記憶のことです。