Press Release

千葉工業大学

IRCN International Research Center for Neurointelligence

wpi

研究成果

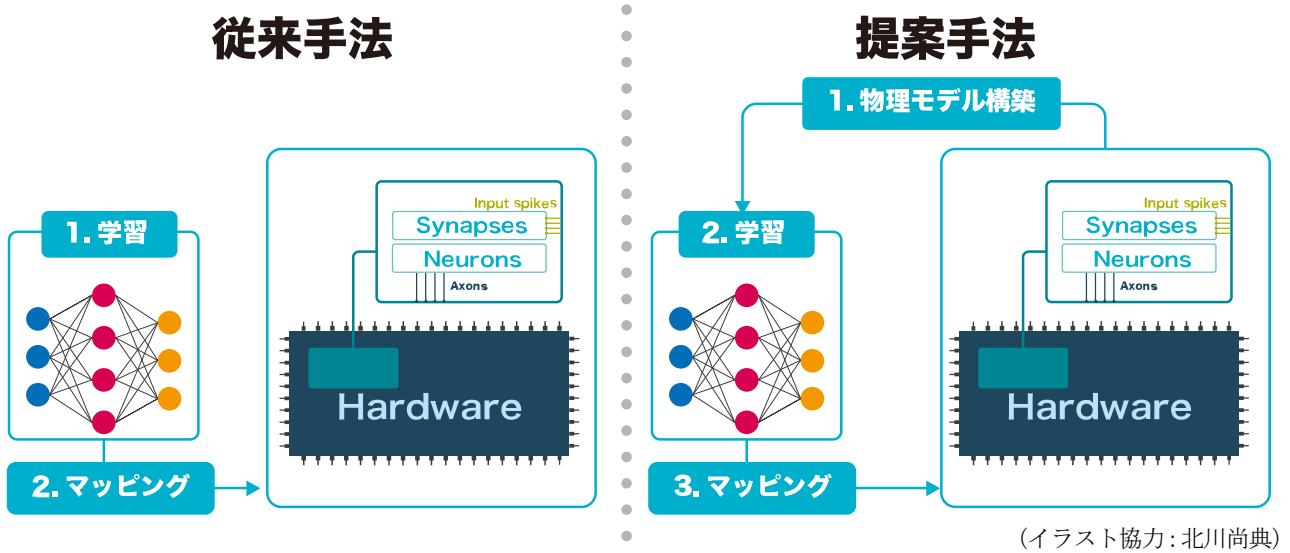
プロジェクト

2025 年 11 月 5 日

報道機関　各位

千葉工業大学 数理工学研究センター

アナログインメモリ計算回路の非理想的特性を取り込む ODE ベース学習手法を開発
―実用規模の学習に成功―



（イラスト協力：北川尚典）

［ 発表者 ］

・酒見 悠介（千葉工業大学 数理工学研究センター 上席研究員／東京大学国際高等研究所ニューロインテリジェンス国際研究機構（WPI-IRCN） 連携研究者）

・岡本 有司（京都大学 大学院医学研究科 特定助教）

・森江 隆（九州工業大学 大学院生命体工学研究科 特任教授・名誉教授）

・信川 創（千葉工業大学 情報変革科学部 教授）

・細見 岳生（日本電気株式会社　セキュアシステムプラットフォーム研究所　シニアプロフェッショナル）

・合原 一幸（東京大学国際高等研究所ニューロインテリジェンス国際研究機構（WPI-IRCN）エグゼクティブ・ディレクター／東京大学 特別教授・名誉教授／千葉工業大学 数理工学研究センター 主席研究員）

キーワード：アナログインメモリ計算、アナログ回路、物理モデル、スパイキングニューラルネットワーク、エッジ AI

[概要]

酒見悠介、岡本有司、森江隆、信川創、細見岳生、合原一幸による研究チームは、超低消費電力で AI 推論が可能な次世代ハードウェアであるアナログインメモリ計算を効率的に動作させる新しい学習手法の開発に成功しました。この学習手法の肝は、アナログハードウェアの動作を常微分方程式（ordinary differential equation: ODE）で記述し、それを直接学習させることにあります。これにより、従来は学習アルゴリズムの中に取り込めなかったアナログ回路特有の複雑な非線形特性を取り込むことができます。しかし、この物理モデルの学習は計算負荷が高く実用的な規模の学習モデルを学習させることができませんでした。そこて（イラスト協力：北川尚典）differentiable spike-time discretization (DSTD)とよぶ新しい学習手法を開発し、学習効率を 1000 倍程度高めることで実用規模とされる 8 層の畳み込みニューラルネットワークの学習に成功しました。学習の結果、従来は学習性能を低減させるため、「非理想的」と考えられていた実際のアナログ回路の複雑性が、むしろ学習性能を高める場合も確認できました。さらに、回路設計とそのシミュレーションも実施し、実回路での同手法の有効性の検証も行いました。本手法は、従来は回避すべき存在であったアナログ回路の非理想的特性を活用することができるため、アナログインメモリ計算回路の設計常識を大きく覆す可能性があります。この成果は、2025 年 8 月 18 日に査読付き国際学術雑誌「Advanced Intelligent Systems」で公開されました。

■ 背景

この 10 年余りの間に、AI は、従来は人間が行っていた複雑な仕事を、高速かつ正確にこなすことができるようになりました。しかし、現在の AI は、消費される膨大な電力が大きな課題となっています。特に、ロボットなどのよりユーザーに近いところでの AI 動作（エッジ AI）では、供給可能な電力量が限られるため、電力効率の高さが AI の性能に直結します。

電力効率を高める最も直接的な方法の一つが専用ハードウェアを用いることです。特に、**アナログインメモリ計算（analog in-memory computing: AIMC)**は、GPU 比 1000 倍もの電力効率を発揮するとして学術界のみでなく産業界からも大きな注目を集めています。しかし、AIMC はアナログ回路をベースとするため、アナログ回路特有の複雑さが設計を困難にしています。この複雑さとして、アナログデバイスの非線形性やばらつきなどが挙げられます。これらによって、ソフトウェアで学習された AI モデルと、ハードウェア上の動作に齟齬が発生し、

結果として AI 性能の低下（認識精度や予測精度など)を引きおこします。そのため、これらの複雑性は「非理想的特性」として認識されてきました。

これらの「非理想的特性」に対処する方法の一つとして **hardware-aware training (HAT)**が知られています。これは、AIMC の特性を学習モデルに組み込むことで、ソフトウェア上で学習されたモデルと、実際のハードウェア上での動作を近づけることを目的にします。この手法は、一定の効果が得られていますが、ハードウェア動作を近似的にしか表現することができず、一般的な深層学習モデルと齟齬が大きくなれば、その違いを学習モデルに反映することが困難でした。

■ 内容

本研究では、AIMC 回路の極めて複雑な動作をより正確に学習モデルに取り込むために、AIMC 回路の動作を常微分方程式（ordinary differential equation: ODE)で記述し、それを学習モデルとして扱うアプローチを提案しました。これにより、学習モデルの動作とハードウェアの動作がほぼ一致することが期待できます。実際に扱ったニューロンモデルは、膜電位$v(t)$が以下のように時間発展するモデルです。

$$\frac{dv}{dt}(t) = \sum_i w_i \left(1 - \frac{v(t)}{E^i_{\mathrm{rev}}}\right) \theta(t - t_i) \tag{1}$$

ここで、$\theta(\cdot)$はステップ関数です。このニューロンモデルへの入力は時刻t_iに到達するパルス(スパイク)であり、ニューロンモデルの出力は$v(1)$として得られます。また、この最終膜電位は、このニューロン自身の出力スパイク時刻へと変換されます。E^i_{rev}は AIMC 回路の非理想的特性に相当し、実際　$|E^i_{\mathrm{rev}}| \to \infty$で、出力は$v(1) = \sum_i w_i(1 - t_i)$となり、通常の深層学習モデルにおける積和演算と一致させることができます。

他方で、このアプローチには学習負荷を増大させる課題があります。本研究で扱った ODE 型学習モデル（式（１））では、通常の深層学習モデルと比べて入力次元分だけ、計算が複雑です。これを数学的に表すと、通常の信号学習モデルの計算複雑性が$O(N_{\mathrm{in}}N_{\mathrm{out}})$であるところ、この物理モデルの計算複雑性は$O(N_{in}^2 N_{\mathrm{out}})$となります。ここで、$N_{\mathrm{in}}$は入力次元、$N_{\mathrm{out}}$は出力次元です。現在一般に用いられているようなネットワーク規模を考えると、計算負荷は、およそ 1000 倍、もしくはそれ以上に計算を要します。この問題により、従来研究 [Sakemi2022]では、中間層二層程度の小規模なモデルの学習が限界でした。

本論文では、学習にかかる計算を劇的に少なくする手法 **differentiable spike-time discretization (DSTD)**を開発しました（図１）。これは、ある層への入力信号を、連続時間表現 t_iから、離散時刻表現$(s_{i1}, s_{i2}, \dots, s_{iM})$へと変換するものです。ここで、$t_i$は i 番目のニューロ

ンからの入力を表しており、s_{im}は、i 番目のニューロンからの入力が、時刻 T_mにおいて存在するかどうかを表します。実際の変換は以下を用いました。

$$s_{im} = \max\left(0, \frac{\Delta_\tau - |T_m - t_j|}{\Delta_\tau}\right) \tag{2}$$

ここで、Δ_τは離散時間幅（$T_m - T_{m-1}$）です。

DSTD には三つの優れた特徴があります。一つ目は、DSTD により、一つの層の計算複雑性が$O(N_{\text{in}}^2 N_{\text{out}})$から$O(MN_{\text{in}}N_{\text{out}})$に改善されることです。ここで、$N_{\text{in}}$は入力次元、$N_{\text{out}}$は出力次元、$M$は離散時間のステップ数です。ステップ数が大きいと計算量が爆発しますが、実験により M=10 程度で十分な精度を満たせることがわかりました。二つ目は、DSTD の変換則(式（２）)は微分可能であることです。そのため、時間離散化しても、通常の自動微分ツール(PyTorch など)が使用でき容易に学習できます。三つ目は、スパイク時間を離散時間表現に近似変換した後は、解析解をもちいて解析解を用いて時間発展を計算でき、連続時間表現の出力スパイク時刻を正確に計算できることです。そして、ステップ数Mが十分に大きいと、厳密な ODE に一致します。論文ではこの性質を$O(M^{-2}|E_{\text{rev}}|^{-1})$として、定理として厳密な証明を与えました。

図 2 に膜電位の時間発展の例を示しています。非理想的特性がつよくなる（$|E_i^{\text{rev}}| = 1$)ほど DSTD による近似精度が下がりますが、離散時間ステップ数Mが 4 においても、終状態が正確に計算されていることがわかります。また、図 3 においては、離散時間ステップ数Mを大きくしたときの時間発展の様子を示しています。離散時間ステップ数Mが 10 もあれば、ほとんど誤差なく計算ができていることがわかります。なお、学習プロセスにおいても、離散時間ステップ数Mが 10 もあれば十分な学習性能が得られることが実験的にわかりました（論文参照）。

図 4 に、DSTD の計算効率を示しています。DSTD により、計算速度が２０倍、メモリ効率が１００倍向上することがわかります。メモリ効率が高くなれば多くのデータを一度に処理できるため、あわせて 1000 倍以上（理論上）の効率性が見込めます。

この DSTD の効率化によって、8 層の畳み込みニューラルネットワークの学習が 1 台の GPU (NVIDIA A100)で可能になりました。CIFAR-10 データセットを学習させた結果を図 5 に示しています。従来手法（緑線)では、非理想的特性が強い（$|E_{\text{rev}}|$が小さい)と、単調に認識性能が低下することがわかります。一方で、提案手法を用いると、非理想的特性が強く出ている場合($|E_{\text{rev}}| \sim 3$)において、性能が最大化しました。これは、従来では学習性能を下げる「非理想的」なアナログ回路の複雑な特性が、学習を工夫することで学習性能を向上させるように活用できることを示唆しています。

本研究ではさらに、この ODE 型学習モデルに基づく手法が、実際の AIMC 回路において有効に働くかどうかを検証するために、「非理想的特性」を含む AIMC 回路を設計しました（図６）。この回路のシミュレーション結果を図７に示しています。このシミュレーション結果から、提案手法を用いることによって、学習モデルと実際のハードウェアの動作との誤差を 1/20 に低減することを示しました。

本研究では、アナログハードウェアの複雑な特性を ODE として学習モデルに組み込むことで、従来「非理想的」と考えられてきた特性の悪影響を低減できるばかりか、場合によっては、学習性能を向上させる場合もあることを示しました。これは、非理想的特性を許容することで、AIMC 回路の設計を簡易化することを可能にし、より高集積化・低電力化することが期待できます。本研究は、非理想的特性を学習に組み込む新たな HAT 手法とみなせます。これまで、HAT 手法は、ばらつく可能性のあるパラメータを確率変数化したり、活性化関数をアナログハードウェアに則したものに差し替える probabilistic/precise modeling approach や、学習の推論・訓練ループに実ハードウェアを差し込む hardware-in-the-loop が知られていました[Sakemi2025]。本手法は、ハードウェアを ODE ベースによって正確にモデル化する、第三の HAT アプローチといえます。

本研究では、代表的な AIMC 形態における重要な非理想的特性を学習モデルに組み込むことに成功しました。しかし、AIMC においては、これ以外にも IR-drop や capacitive coupling などの重要な非理想的特性が存在します。これらについても対応可能に拡張することは将来的な課題です。また、本手法を適用するためには、物理モデルをある程度正確に構築することが必要です。この方法としては、回路シミュレーションから抽出する以外にも、実ハードウェアからのデータをもとに構築することも考えられます。

※１）アナログインメモリ計算（AIMC）
メモリ内で計算を行うことでデータ移動を最小化し、電力効率を高めるアナログ演算手法で、AI ハードウェアにおいては行列ベクトル積（$\boldsymbol{y = Wx}$）の計算に用いられる。一般的にクロスバーアレー構造を持ち、横配線と縦配線の交点に重みが配置される。この重みは抵抗素子などで物理的に実装される。

■ 参考文献

[Sakemi2022] Y. Sakemi, K. Morino, T. Morie, T. Hosomi and K. Aihara, "A Spiking Neural Network with Resistively Coupled Synapses Using Time-to-First-Spike Coding Towards Efficient Charge-Domain Computing," ISCAS, 2152-2156 (2022)

[Sakemi2025] Y. Sakemi, H. Awano, and T. Morie, “A Tutorial on Analog In-Memory Computing: Theory, Nonidealities, and Hardware-Aware Training”, IEICE Fundamentals (2025, in press)

■ 論文情報

雑誌名: Advanced Intelligent Systems

論文題目: Harnessing Nonidealities in Analog In-Memory Computing Circuits: A Physical Modeling Approach for Neuromorphic Systems

著者: Yusuke Sakemi、Yuji Okamoto、Takashi Morie、Sou Nobukawa、Takeo Hosomi、and Kazuyuki Aihara

URL: https://advanced.onlinelibrary.wiley.com/doi/full/10.1002/aisy.202500351

DOI: 10.1002/aisy.202500351

発表日時: 2025 年 8 月 18 日

■ 謝辞

本研究の一部は、JST さきがけ JPMJPR22C5、国立研究開発法人 新エネルギー・産業技術総合開発機構（NEDO）の委託プロジェクト JPNP14004 、JSPS KAKENHI Grant Number 25K00148/JP20H05921、日本電気株式会社、セコム科学技術振興財団、JST Moonshot R&D Grant Number JPMJMS2021、Institute of AI and Beyond of UTokyo、International Research Center for Neurointelligence (WPI-IRCN) , The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo、Cross-ministerial Strategic Innovation Promotion Program (SIP), the 3rd period of SIP, grant no. JPJ012207 and JPJ012425 から助成を受けて行われました。

【本研究内容に関する問い合わせ】
千葉工業大学　数理工学研究センター　上席研究員
酒見　悠介
HP: https://sites.google.com/view/rcme-cit/
TEL: 047-478-0345
E-mail: yusuke.sakemi@p.chibakoudai.jp
【取材・大学広報関連に関する問い合わせ】
千葉工業大学　入試広報部
大橋　慶子
E-mail: ohashi.keiko@it-chiba.ac.jp

■ 添付資料

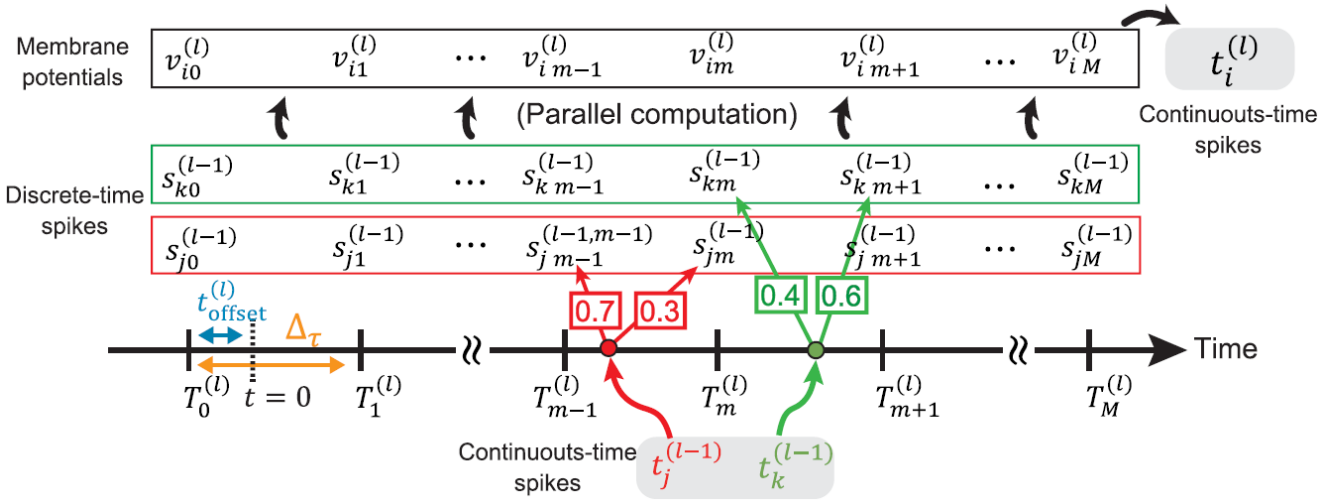


図 1 Differentiable spike-time discretization (DSTD)

連続時間表現の入力スパイクt_iを離散時刻表現$(s_{i1}, s_{i2}, \ldots, s_{iM})$に変換する。変換後は、厳密解を利用して高速な計算が可能になる。また、この変換は微分可能なため通常の誤差逆伝搬アルゴリズムが適用可能である。

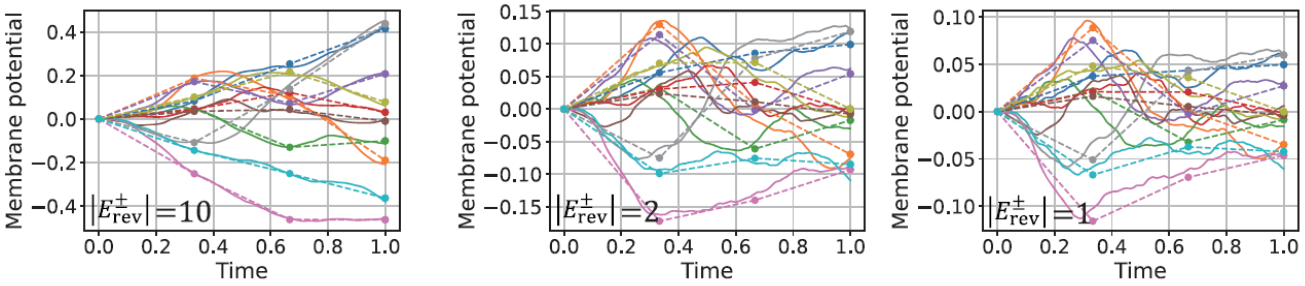


図 2　膜電位の時間発展の例 1

1000 個のランダムな入力スパイクを与えた時の膜電位の時間発展を異なる非理想的特性$|E_{\mathrm{rev}}|$の場合について示している。$|E_{rev}|$が小さいほど非理想的となる。実線は厳密な ODE の解であり、プロットしているものは DSTD による近似解である。

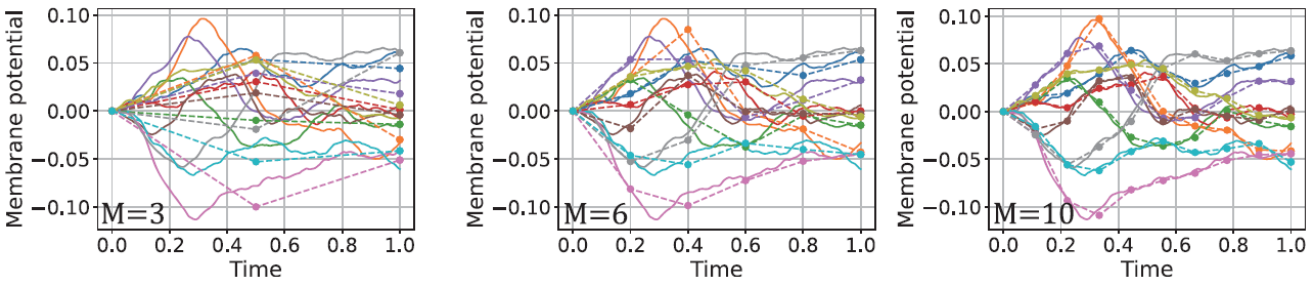


図３　膜電位の時間発展の例２

1000 個のランダムな入力スパイクを与えた時の膜電位の時間発展を異なる離散時間ステップ数 M の場合について示している。実線は厳密な ODE の解であり、プロットしているものは DSTD による近似解である。$|E_{\mathrm{rev}}| = 1$の極めて非理想的な状況下において、M=10 程度で、近似解が厳密解とかなり近くなっていることがわかる。

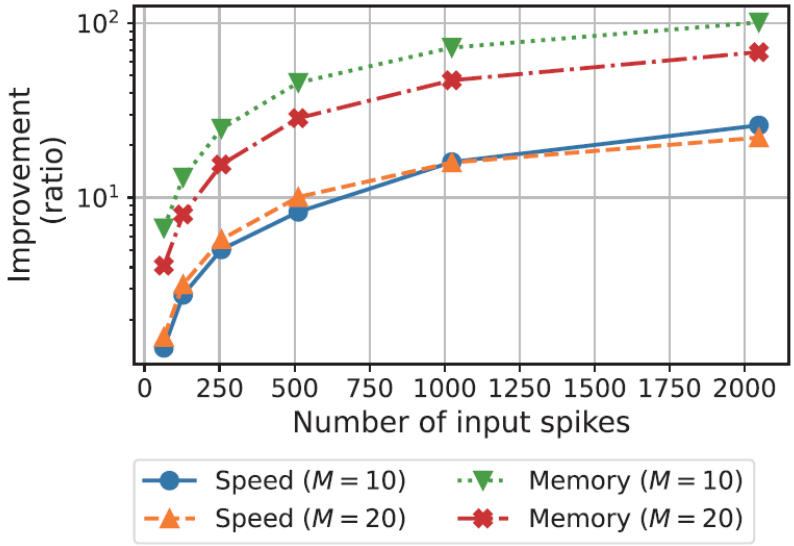


図４　DSTD による計算効率の向上

1000 個のニューロンで構成される単層ニューラルネットワークにおける計算速度の向上と、メモリ効率の向上を、異なる離散ステップ数 M について示している。横軸は、入力スパイク数(入力次元)である。

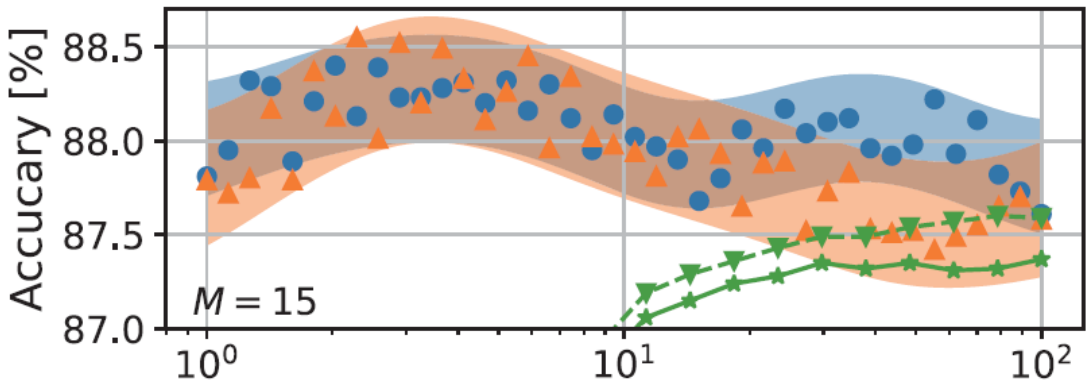


図５　非理想的特性の学習性能への影響

8 層畳み込みニューラルネットワークによる CIFAR-10 の学習結果を示している。横軸は非理想的特性 $|E_{\mathrm{rev}}|$であり、小さいほど非理想的であり、大きいほど理想的（通常の深層学習モデルと等価）となる。従来手法（緑線)では、非理想的になるほど、認識性能が低下するが、提案手法（青/橙プロット)では、非理想的特性が強いときに性能が最大化している。

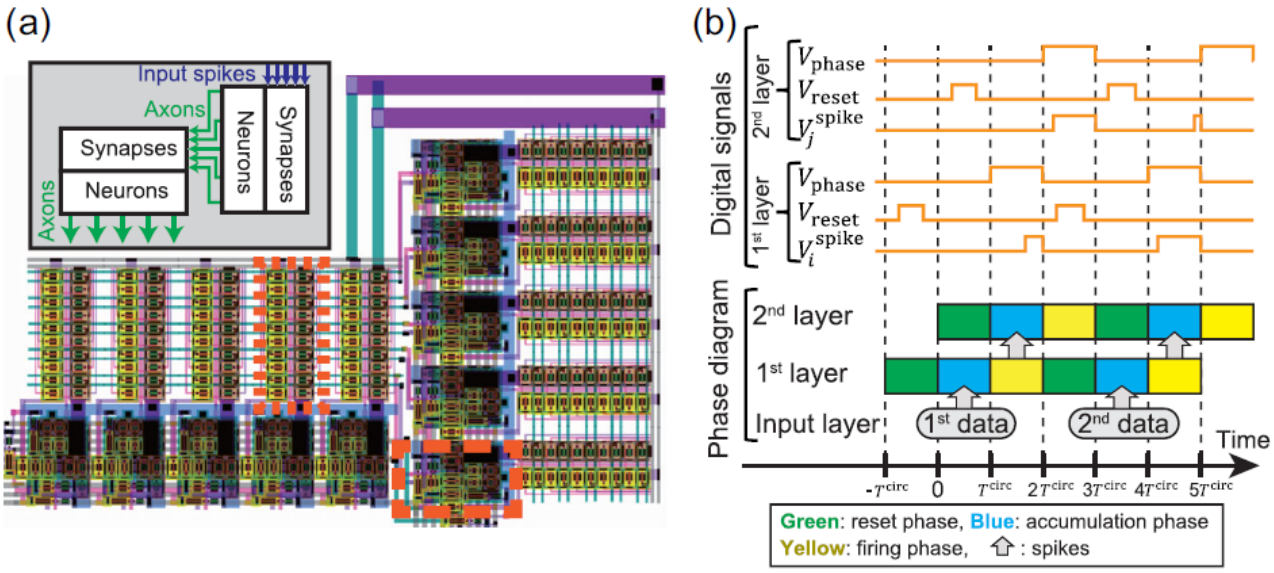


図６　設計した AIMC 回路

(a) 回路のレイアウト図。オープンソースなプロセスである SkyWater SKY130 を用いて設計した。隠れ層が 1 層 (5-5-5)のみの小規模回路。各層はシナプスブロックとニューロンブロックで構成されており、シナプスブロックに強い非理想的特性（$|E_{rev}| \sim 2$)がある。(b) 各層は、reset phase、入力スパイクを受け付ける accumulation phase、スパイクを出力する firing phase を持ち、パイプライン動作が可能である。

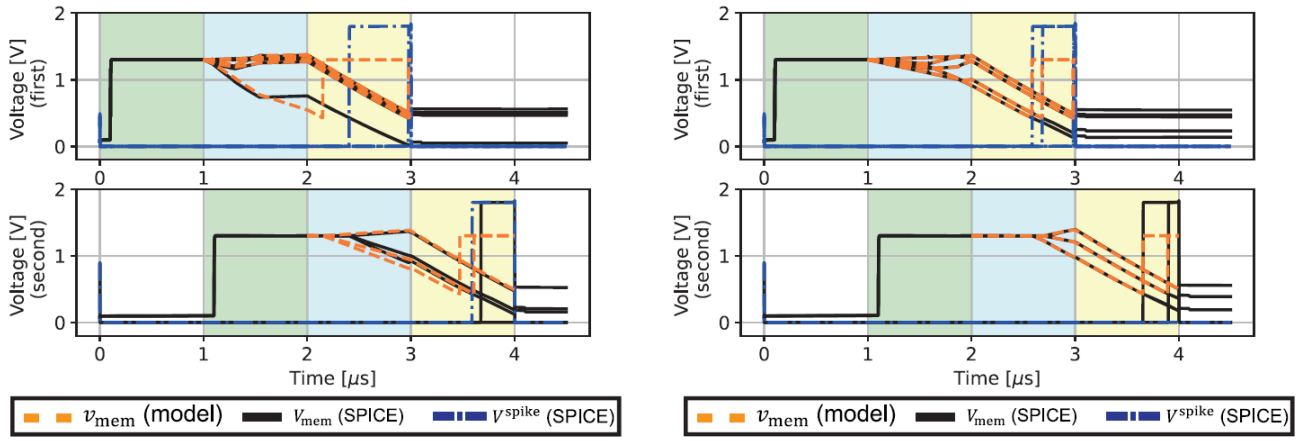


図7　回路シミュレーションの結果

（左）従来手法の場合のシミュレーション結果。上パネルは隠れ層、下のパネルは出力層の膜電位の時間発展を示している。モデルの結果（破線橙色)と、回路シミュレーションの結果（実線黒線)の結果が大きく異なっていることがわかる。（右）同シミュレーションを提案手法を用いて重みのマッピングを行った時のシミュレーション結果。モデルの結果（破線橙色)と、回路シミュレーションの結果（実線黒線)の結果がほとんど一致していることがわかる