アナログメモリスタが拓く次世代AI計算アーキテクト

技術解説

Zhuo Diao*

Next-Generation Analog AI Edge Computing Architectures Enabled by Analog Memristors

Key Words: in-memory computing, AI electronics, memristor

はじめに

近年，生成AIの急速な発展は，単なる対話型ボットの枠を超え，自律的にタスクを遂行するAIエージェントへと進化しつつある。このような高い汎用性を持つAIは，製造業や科学研究を含む多様な分野における自動化を加速させており，「第四次産業革命」を支える中核技術として位置づけられている。この潮流の中で，AI計算基盤を支える半導体産業，とりわけメモリ分野は2024年より急激な成長を遂げている。実際に，2025年には半導体市場全体の指標であるSOX指数が約40%上昇する中で，最も顕著な伸びを示したのはメモリ関連企業であった。例えば，Micron Technologyは200%以上の株価上昇を記録し，キオクシアにおいて10倍近くの株価増加が観測されている。これらの動向は，単なる投機的な資金流入ではなく，AI時代におけるメモリの戦略的重要性が市場によって再評価された結果と捉えられる。

技術の観点でこの背景を解釈すると，AI計算における構造的なボトルネックの存在に由来する。従来，計算性能の向上は主にプロセッサの演算能力の強化によって達成されてきたが，大規模AIモデルにおいては，演算そのものよりもデータ転送に伴うコストが支配的となりつつある。この問題への対処として，プロセッサとメモリ間の帯域不足を緩和する高帯域メモリ（High Bandwidth Memory, HBM）と，電気配線の抵抗・容量成分によるRC遅延を克服する光インターコネクト技術への投資が急拡大している。HBMは三次元積層構造と広帯域インターフェースにより従来のDDRメモリと比較して大幅な帯域向上を実現し，市場規模は2025年の約350億ドルから2028年には1,000億ドル規模へと拡大すると予測されている。光インターコネクト分野では，波長分割多重（WDM）により極めて高い帯域密度を実現できることから，Ayar Labsが2024年にAMD・Intel・NVIDIA・TSMCなどから総額1億5,500万ドルの資金調達を実施，Lightmatterも累計8億5,000万ドルの資金調達により企業価値44億ドルに達するなど，大規模投資が進行している。

しかし，HBMも光インターコネクトも，演算とメモリを物理的に分離したまま「いかに速くデータを運ぶか」を最適化するアプローチであり，フォン・ノイマン型アーキテクチャの根本的な限界を解消するものではない。業界では現在，AI計算のボトルネックはプロセッサ性能の不足ではなく，より本質的な二つの限界，すなわち「Computing Wall（演算の壁）」と「Memory Wall（メモリの壁）」に起因するという認識が共有されつつある。HBMや光インターコネクトはMemory Wallを緩和する有効な手段ではあるものの，演算とメモリが分離されているかぎり，データ転送に伴うエネルギーと遅延のオーバーヘッドは本質的に残り続ける。

このような状況において注目されているのが，演算と記憶を同一素子内で実行する「Computing-in-Memory（CIM）」パラダイム[1]である。CIMはメモリセル自体を演算素子として利用することでデータ転送を最小化し，高いエネルギー効率と並列性を実現する新しい計算アーキテクチャである。特に，連続値の抵抗状態を利用して重みを表現できるアナログメモリスタは，大規模ニューラルネットワークの行

* Zhuo DIAO
1995年10月生まれ
大阪大学大学院 基礎工研究科 システム創成専攻博士後期課程（2024年）
現在、大阪大学大学院 基礎工学研究科 システム創成専攻　助教　工学博士
TEL：06-6850-6302
FAX：06-6850-6302
E-mail: diao.zhuo.es@osaka-u.ac.jp

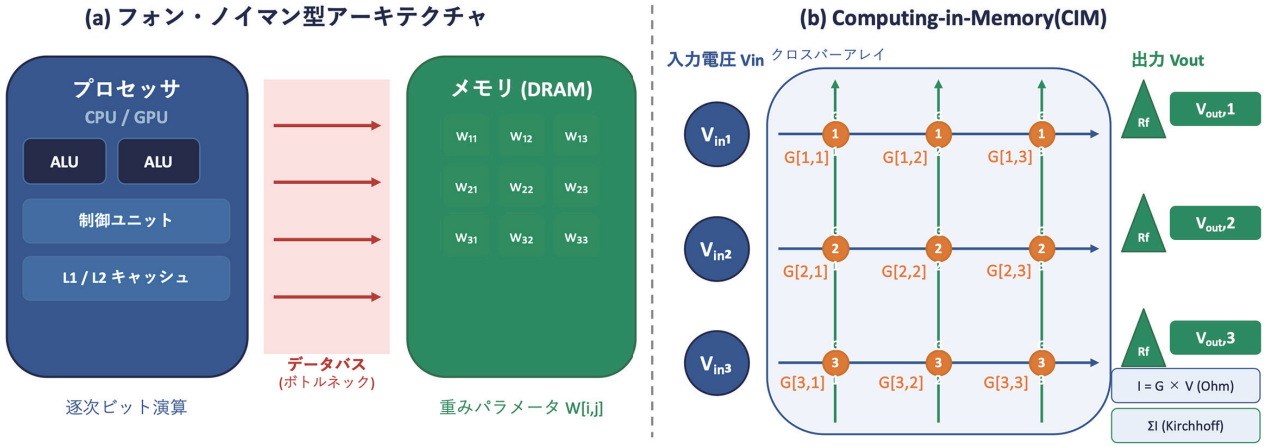

図1　フォン・ノイマン型アーキテクチャ vs メモリ内計算（CIM）

列演算を物理的に並列実行可能なデバイスとして有望視されている。本稿では，Computing WallおよびMemory Wallという二つの制約を背景に，それらをアーキテクチャレベルで根本から克服するアプローチとしてのCIMパラダイムと，その中核デバイスであるアナログメモリスタについて解説する。

ムーアの法則の限界（Computing Wall）

デジタル動作を筆頭とする集積回路の性能向上を長年支えてきたムーアの法則は，「トランジスタ数が約2年ごとに倍増する」という経験則として知られる。このスケーリングにより，半導体は高集積化と高性能化を実現してきた。しかし現在では，ゲート長が数ナノメートルに到達し，量子トンネル効果や熱ゆらぎ，リーク電流の増大といった物理的制約が顕在化しており，微細化による性能向上は限界に近づきつつある。こうしたプロセッサ性能向上の鈍化と同時に，AI処理への計算需要は指数関数的に増大しており，供給と需要の乖離が深刻化している。人工ニューラルネットワーク（ANN）の主要演算である行列ベクトル積は，デジタル回路との親和性が低い。デジタル計算では数値はビット列として処理され，一つの乗算でさえ複数段の論理演算を必要とする。大規模ニューラルネットワークでは数億～数千億規模のパラメータに対する行列ベクトル積を繰り返し実行するため，計算量が爆発的に増大し，計算資源およびエネルギー消費の観点で大きな負担となる。すなわちComputing Wallとは，「微細化の物理的限界によるプロセッサ性能向上の頭打ち」と「AIが要求する計算量の急増」という二重の制約が重なることで生じる壁である。

フォン・ノイマンボトルネック（Memory Wall）

現代の計算機の多くはフォン・ノイマン型アーキテクチャを採用しており，演算器（CPU/GPU）と記憶装置（DRAM）が物理的に分離されている（図1a）。この構造では，演算に必要なデータをメモリから読み出し，計算後に再び書き戻すというプロセスが繰り返される。このとき問題となるのが，データ転送に伴うエネルギーと遅延の大きさである。一般に，32ビットのデータに対する浮動小数点演算1回のエネルギーは数pJ程度であるのに対し，同じデータをDRAMから読み出すエネルギーは数十～数百pJに達するとされる。すなわち，演算そのものよりもデータ移動の方が1～2桁大きなエネルギーを消費する。さらに，DRAMアクセスのレイテンシは数十～数百ナノ秒に及び，高速な演算器と比較して大きな時間的ボトルネックとなる。AIの推論および学習では，数億から数千億に及ぶ重みパラメータを繰り返し読み出す必要があるため，このデータ転送コストが支配的となる。実際，GPUベースのAIアクセラレータでは，消費電力の大部分が演算ではなくメモリアクセスおよびデータ移動に費やされていると報告されている。この「Memory Wall」は，プロセッサの演算性能が指数関数的に向上する一方で，メモリ帯域やレイテンシの改善がそれに追従できないことに起因する。結果として，演算器は十分なデータ供給を受けられず待機状態に陥り，システム全体の性能が制限される。キャッシュ階層の多段化や，演算器とメモリを同一チップに集積するオンチップ統

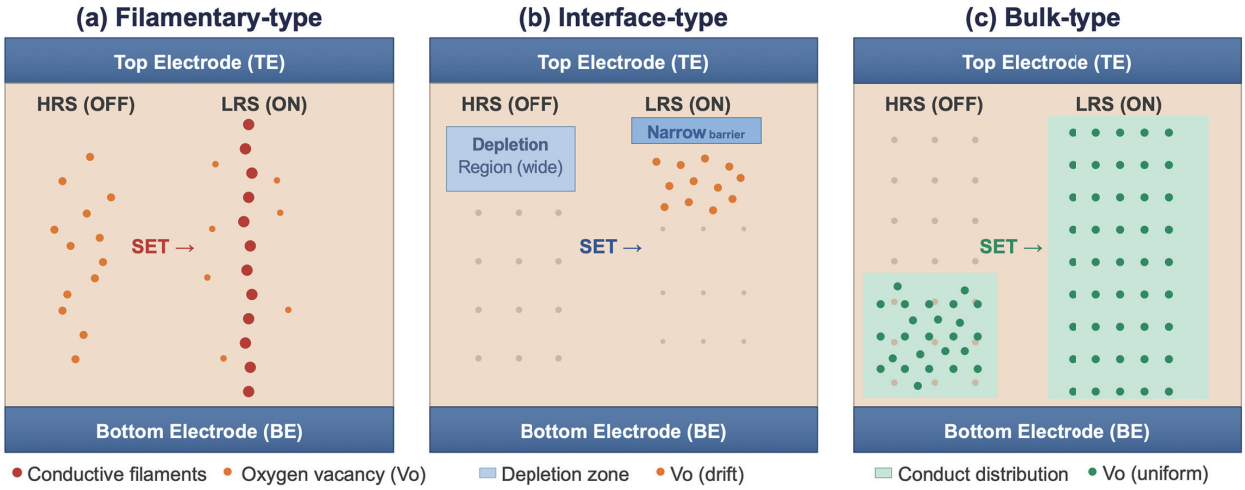

図２　メモリスタのスイッチングに関わる３種類のメカニズム

合はこの問題を緩和する手法として有効である。しかし，これらはいずれも「データを移動させる」という前提を維持したままでの最適化に過ぎず，演算とメモリが物理的・機能的に分離されている限り，データ転送に伴うエネルギーと遅延のオーバーヘッドは本質的に残り続ける。

Computing-in-Memory（CIM）パラダイム

上述の二つの壁を根本から打破するアプローチとして，世界的に注目されているのが「Computing-in-Memory（CIM）」パラダイムである。これは，メモリと演算を同一素子内で実行することにより，データ移動を最小化する計算アーキテクチャである。CIMでは，メモリ素子そのものに演算機能を担わせる点に本質的な特徴がある。その中核デバイスの一つがメモリスタである。

メモリスタは，1971年にLeon Chuaにより理論的に提唱された「第四の回路素子」であり，2008年にHP Labsの研究グループによってデバイスとして実証された[2]。その特長は，印加電圧や電流の履歴に応じて抵抗値が変化し，電源遮断後もその状態を保持する不揮発性にある。設定可能な抵抗値の逆数であるコンダクタンスGをニューラルネットワークの重みとして直接対応付けることが，CIMの基本的な発想である。

メモリスタを格子状に配置したクロスバーアレイは，CIMの物理的実装基盤となる（図1b）。行方向電極と列方向電極の交点に各素子を配置し，G[i,j]を重み行列Wの要素w[i,j]に対応させる。ここで行方向に入力電圧ベクトルVinを印加すると，各素子に流れる電流はオームの法則に従い，各列でキルヒホッフの電流則により自動的に加算される。列端の演算増幅器（帰還抵抗Rf）の出力は

$$Vout,k = -R_f \times \Sigma_j\, G[k,j] \times Vin,j$$

となり，これは行列ベクトル積W×Vinのk行目に対応する（負号は反転増幅器の構成に由来する）。このように，積和演算は個々の論理演算の逐次実行としてではなく，物理法則に基づくアナログ応答として並列的に実行される。したがって，デジタル計算における逐次ビット演算と比較して，演算レイテンシおよびエネルギー効率の両面で大幅な向上が期待される。さらに，演算がメモリ素子内で完結するためデータを外部に転送する必要がなく，低レイテンシな推論と高いエネルギー効率が実現できる。

メモリスタの動作原理

メモリスタにおいて，抵抗値を不揮発的に制御する抵抗スイッチング機構は，材料系やデバイス構造に応じて異なる固体物性に基づく。一般に，その動作機構はフィラメント型（filamentary），界面型（interface），バルク型（bulk）の三種類に分類され，これらの違いはデバイス特性を大きく左右する（図2）。ここで，メモリスタは上部電極（top electrode: TE）と下部電極（bottom electrode: BE）に挟まれた構造を持ち，電圧印加により高抵抗状態（HRS）と低抵抗状態（LRS）の間を可逆的に遷移する。一般に，HRSからLRSへの遷移はSET過程，LRSからHRSへの遷移はRESET過程と呼ばれる。これらの基本

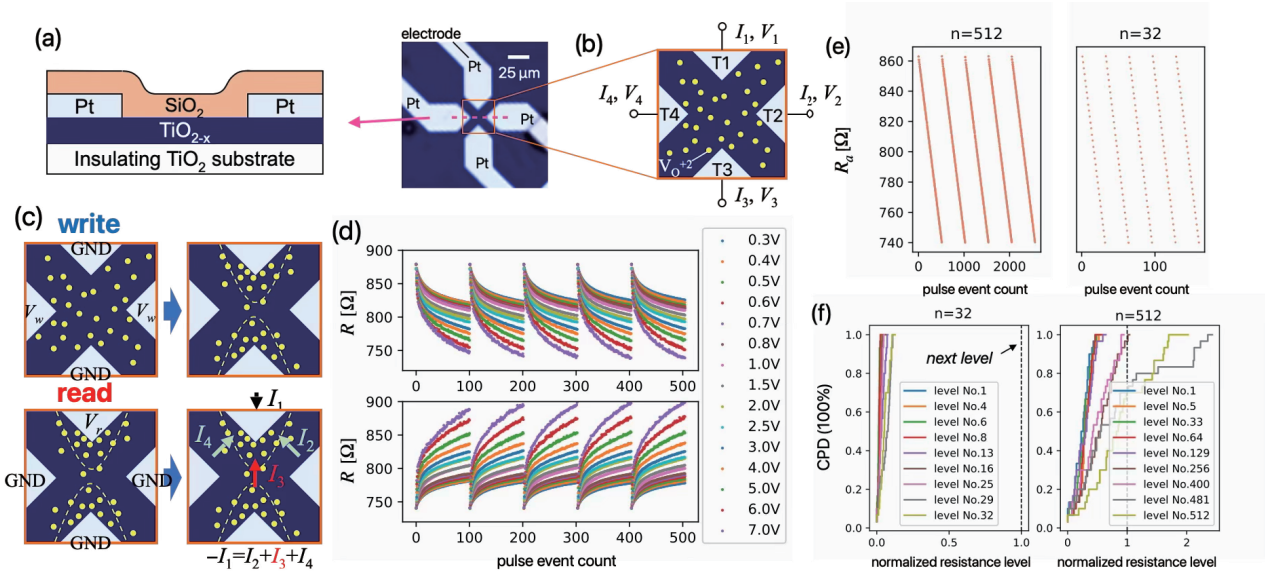

図3 (a) 四端子TiO_{2-x}デバイスの顕微鏡像と断面図。(b) 抵抗スイッチング領域における酸素空孔（Vo）の分布と、各端子の電流・電圧の定義。(c) 書き込み・読み出し時の電圧印加条件と測定電流の説明図。点線部分は酸素空孔の高濃度領域を示す。(d) 同一異なる電圧パルスを100回印加を繰り返したときの抵抗スイッチング特性。(e) 32, 512段階（n = 32, 512）に抵抗を多値分割した結果と (f) 累積確率分布（CPD）。著者らの既報論文（CC BY 3.0 Deedライセンス）より掲載 [4]。

動作はすべてのスイッチング機構に共通するが，その物理的起源や空間分布は各方式によって大きく異なる。

フィラメント型は最も広く研究されている方式であり，電圧印加により酸素空孔（oxygen vacancy: Vo）や金属イオンが移動し，局所的にナノスケールの導電経路（conductive filament）が形成・断裂することで抵抗状態が切り替わる（図2a）。この方式は高速動作と高いON/OFF比を実現できる一方で，フィラメント形成が確率的に生じるため，その位置や形状にばらつきが不可避である。その結果，中間抵抗状態の精密制御が困難であり，アナログ多値動作には課題が残る。

界面型では，電極／酸化物界面における電気的障壁の変調を利用する（図2b）。電場によりキャリアの捕獲・放出が起こると，界面の障壁幅や高さが変化する。この障壁変調により電流の流れやすさが連続的に制御され，それに伴って抵抗値が連続的に変調される。このため，フィラメント型と比較して滑らかなアナログ特性と高い再現性が得られる。一方で，伝導変化が界面近傍に局在するため，ON/OFF比が比較的小さいことや，温度や外部環境の影響を受けやすい点が課題である。

バルク型[3]は，材料内部全体にわたるイオン分布の変化により伝導が変調される方式である（図2c）。フィラメントのような局所的経路形成や界面障壁に依存せず，デバイス全体にわたって均一な伝導変化が生じるため，本質的に高い線形性と優れた多値制御性を実現できる。この特性はニューラルネットワークにおける重み更新精度に直結するため，CIM用途において特に重要である。一方で，スイッチング速度や動作電圧，および材料設計の難しさといった課題は依然として残されている。それにもかかわらず，高精度アナログ演算の観点から有望なアプローチの一つであり，本研究で扱うデバイスもこのバルク型に分類される。

バルク型四端子 TiO_{2-x}メモリスタデバイスの作製と特性評価

我々のグループでは，Voを導入したTiO_{2-x}薄膜上にPt電極4本（T1～T4）を平面配置した四端子メモリスタデバイスを作製し，従来のフィラメント型とは異なる「バルクスイッチング機構」に基づく抵抗変調特性を実証した[4]（図3a）。活性層中には正に帯電した酸素空孔（Vo）が存在し，その二次元空間分布が内部状態変数として機能する（図3b）。外部電場に応じてVoがドリフト・拡散することで電極近傍の空孔ドメイン面積が連続的に変化し，それに対応してデバイスの抵抗値がアナログ的に変調される[5]。書き込み（Write）動作では，T1–T3間に電場を印加

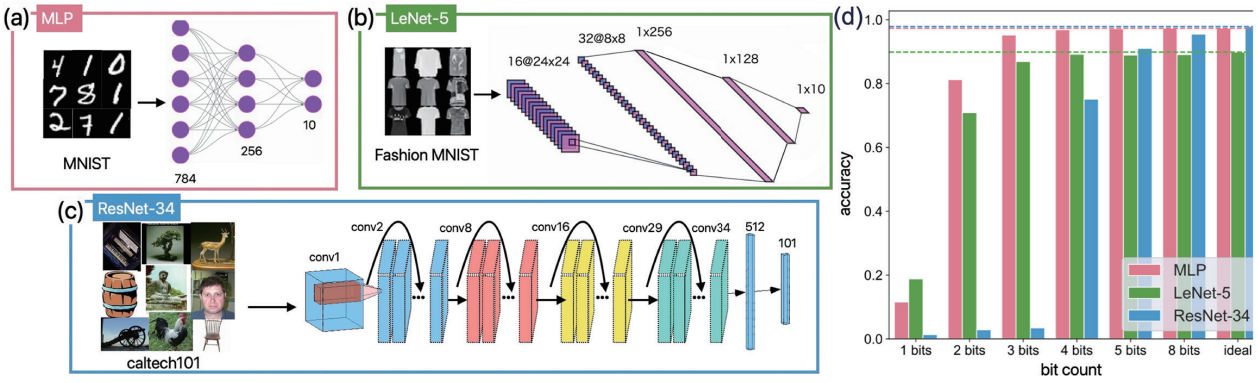

図4　四端子TiO_{2-x}デバイスにおいてる画像認識精度評価。ニューラルネットワークの構造として、(a) MLP、(b) LeNet-5、(c) ResNet-34 の3種類を使用した場合の結果を示す。(d) 異なるビット精度でAI計算に用いたときの精度比較結果。著者らの既報論文(CC BY 3.0 Deedライセンス)より掲載 [4]。

してVoをドリフトさせ，空孔ドメインを連続的に変調する（図3c）。点線で示される高濃度Vo領域が電場に応じて拡大・縮小することで，抵抗値が任意の中間値に設定される。読み出し（Read）動作はT2–T4の分離経路を用い，複数の電流経路が交差する構造において各端子の電流をキルヒホッフの電流則（KCL）で連立することにより抵抗値を算出する。最大の特長は，書き込みパス（T1–T3間）と読み出しパス（T2–T4間）を電気的に分離した構成にある。これにより，バルク型の抵抗スイッチングに関与するVoの配置を二次元の分布として，決定論的かつ精細に制御できる利点を持つ。

フィラメント型に見られる局所的な導電路形成に依存しない二次元バルク応答であるため，急峻なスイッチングや確率的な破壊挙動が抑制され，優れたcycle-to-cycle 再現性が得られる。図3dは，異なる振幅の書き込み電圧（0.3～7.0 V）を同一デバイスに100回ずつ繰り返し印加し，5サイクル分の抵抗スイッチング特性を示している。各電圧条件において抵抗変化曲線はほぼ重なっており，サイクル間のばらつきがピアソン相関係数0.998以上という値からも極めて小さいことが確認できる。また，印加電圧の大きさに応じて抵抗変化量を連続的に調整できることも示されており，これがアナログ多値制御の基礎となっている。

これらの優れたサイクル耐性と連続的な抵抗変調特性を活かして，32段階（5 bit相当）および512段階（9 bit相当）の多値化を拡張し，動作を確認した（図3e）。図3fは，32レベルと512レベルのそれぞれについて，一部のレベルを抽出して累積確率分布（CPD）でレベルごとのばらつきを評価した結果である。ここで注目すべきは，レベル数と設定精度の間に存在するトレードオフである。32レベル（5 bit）では各レベルへの書き込み誤差が小さく安定した設定が可能であるが，表現できる数値の量子化誤差は相対的に大きい。一方，512レベル（9 bit）では量子化誤差が大幅に低減され数値精度が向上するものの，各レベルの分離度が低下しイオンの熱拡散やランダムなドリフト成分の影響による書き込み誤差が増大する。すなわち，「多値数（分解能）を増やして数値精度を上げる」ほど「各レベルへの設定精度は下がる」という固有のトレードオフが存在する。本デバイスでは，このトレードオフを実測した書き込み誤差の確率分布としてデータ化し，異なるビット精度の条件下でのニューラルネットワーク演算精度を予測するシミュレーションフレームワークに組み込んだ。

AI推論精度への実証

構築したシミュレーションフレームワークを用いて，本四端子TiO_{2-x}メモリスタをアナログ計算素子として用いた場合のAI推論性能を評価した。重みをメモリスタの抵抗値として表現するCIMアーキテクチャを想定し，プログラムレベル数（すなわちビット精度）が推論精度に与える影響を体系的に解析した。解析では1 bit（2レベル）から8～9 bit（256～512レベル）までの異なる精度条件を設定し，モデル規模の異なる複数のニューラルネットワークを対象とした。また，デバイス誤差が存在しない理想条

件（Ideal）も同時に算出することで，実デバイスに起因する精度劣化を明確に比較した。

評価には規模の異なる3種類のAI画像認識タスクとモデルを用いた。AIは認識すべき対象が複雑になるほど，より高い特徴抽出能力をもつネットワーク構造を必要とし，それに伴って必要な計算量も飛躍的に増大する。第一は，28×28ピクセルの手書き数字画像（0〜9）を10クラスに分類するMNISTであり，2層の全結合層からなる軽量な多層パーセプトロン（MLP, 約20万パラメータ）を用いた。第二は，Tシャツ・バッグ・靴など10種類の衣類・服飾品画像を認識するFashion-MNISTであり，畳み込み層2段と全結合層2段からなるLeNet-5（約8万パラメータ）を適用した。LeNet-5はMLPよりパラメータ数が少ないが，畳み込み演算により画像の空間的特徴を効率的に抽出できるため，より複雑な認識タスクに対応できる構造となっている。第三は，動植物・乗り物・日用品など101カテゴリの自然物体を256×256カラー画像から認識する高難度ベンチマークCaltech101であり，34層の畳み込みブロックを持つ深層CNN ResNet-34（約2,130万パラメータ）を使用した。モデルが深く・大規模になるほど多くの計算層での処理が必要となるため，計算量は大幅に増加する。

図4に，重み転送（Weight Transfer）方式で異なるビット精度を用いたときの各モデルの推論精度を示す。重み転送とは，まず通常のGPUを使ってAIモデルの学習を完了させ，その後，学習によって得られた重みパラメータ（各ニューロン間の結合強度を表す数値）をメモリスタの抵抗値として書き込む手法である。すなわち，学習フェーズはデジタル計算機上で行い，推論フェーズのみをメモリスタCIMに担わせることで，低電力・高速なアナログ行列演算を活用する。図中の破線（ピンク・緑・水色）はそれぞれのモデルをデジタルGPUで学習した際の精度上限を表している。GPU精度に近い推論性能を実現するために必要なビット精度はモデル規模によって大きく異なり，最も軽量なMLPでは3 bitで95.2%，LeNet-5では4 bitで86.9%，そしてResNet-34では8 bitで95.5%が必要であった。モデルが大規模になるほど要求ビット精度が高まる理由は，ニューラルネットワークが何十・何百もの計算層を積み重ねた構造をもち，各層での小さな誤差が次の層へ引き継がれ，層が深くなるほど誤差が雪だるま式に積み重なるためである。本デバイスの優れたcycle-to-cycle耐性が大半の抵抗範囲における量子化誤差を低水準に保つため，書き込み誤差が増大するにもかかわらず量子化誤差の低減効果がそれを上回り，多値数を最大化することがANN精度の向上に有効であることが確認された。これらの知見から，高度なAIを用いた社会実装に向けて，メモリスタの計算精度を向上させることの重要性が改めて示された。

エッジAIへの展望

現在のAI処理は，データセンターにおけるGPUクラスタを用いたクラウド処理が主流である。しかし，スマートフォン，IoTセンサ，自動運転車，医療診断デバイスなどへの応用拡大に伴い，インターネットを経由しないエッジ側のみでAIを実行する「エッジ推論」への需要が急速に高まっている。クラウド依存型アーキテクチャでは，通信レイテンシ，消費電力，およびプライバシーリスクといった本質的制約が存在し，特にミリ秒オーダーの応答が求められるリアルタイム処理には適用が困難である。エッジ推論における最大の課題は，計算遅延と電力消費の両立である。従来のニューラルネットワークアクセラレータでは，処理時間および消費電力の大部分が演算そのものではなく，メモリと演算ユニット間のデータ転送に起因する。このデータ移動は帯域制約によりレイテンシを増大させると同時に，大きなエネルギーを消費するため，リアルタイム応答と低消費電力が求められるエッジデバイスにおいて本質的なボトルネックとなる。これに対し，メモリスタCIMは重みを抵抗値として不揮発的に保持し，演算をメモリ内部で直接実行することで，データ移動を原理的に排除する。この特性により，従来比で1〜2桁の消費電力削減が可能となり，数mW以下での推論動作が期待される。さらに，不揮発性により電源遮断後も重み情報が保持されるため，即時起動・即時推論が可能となる点も実用上重要である。

本研究で示したバルク型四端子TiO_{2-x}メモリスタは，512段階（9 bit）の多値制御と優れたcycle-to-cycle再現性を実現しており，エッジAI実装に必要な精度要件を満たす。特に，ResNet-34規模の深層CNNにおいてもGPU相当の推論性能が得られるという結果は，クラウドAIに匹敵する認識能力をエ

ッジデバイス上で実現可能であることを示唆している。今後は，大規模クロスバーアレイへの集積化と，デバイス固有の非理想性を考慮したCIMアルゴリズムの開発が鍵となる。これらが進展すれば，クラウドベースのAIインフラのみならず，ウェアラブル医療診断，自律移動ロボット，スマートセンサネットワークといった多様な応答性が問われるエッジAI応用において，CIMアーキテクチャの実用化が加速すると期待される。

参考文献

[1] S. Ambrogio et al., *Nature*, **558**, 60–67 (2018)
[2] R. S. Williams et al., *Nature*, **453**, 80–83 (2008)
[3] Y.Wu et al., *Adv. Mater*., **35**, 2305465 (2023)
[4] Z. Diao et al., *Nanoscale Horizons*, **10**, 780–790 (2025)
[5] R. Miyake et al., *ACS Appl. Electron. Mater*., 4, 5, 2326–2336 (2022)