新規アナログ回路技術を用いたインメモリ演算回路の研究

吉岡健太郎 慶応義塾大学理工学部 専任講師

1 研究背景

深層学習は，画像処理，自然言語処理，音声認識など，さまざまなタスクでその活用範囲を広げています．しかし，深層学習には膨大な演算量が必要であり，特にエッジデバイスにおいては，電力効率と処理速度の両面で効率的な AI ハードウェアが求められています．

一方で現状のデジタル回路ベースの AI ハードウェア(AI HW)は低消費電力化の限界に達しており，最適化・改良では 10 倍の電力削減といったブレイクスルーにたどり着くのは困難です．主たる課題は２つのボトルネックです：１）メモリ移動電力で性能が律速するノイマンボトルネック(図 1)、２）演算回路は既に最適化されており一層の改善は困難．メモリと演算電力それぞれが課題として立ちはだかり，これらを同時に解決する技術が必要です．

我々は２つのボトルネックを打破するため，アナログ演算を用いたメモリと演算器を密結合した Compute-In-Memory (CIM)アーキテクチャに関する研究を行います(図 1B)．CIM アーキテクチャによって重みデータを保持したメモリの極近傍で演算を行うことで，従来回路に対しメモリ移動電力を大幅に低減します．それに加え，分散した演算器の結果を集約するために電荷ベースアナログ演算器を構成し演算電力の大幅低減に挑みます．これらメモリと演算電力の大幅低減により，デジタルベースの AI ハードウェアに対し消費エネルギーを 10 倍改善するブレイクスルーを実現します．またアナログ信号をデジタル演算結果へ変換する過程の誤差によって演算精度が律速される課題を解決するため，本研究では演算器アレーにアナログ-デジタル（AD）変換器を内包する CIM 回路，容量再構成型 CIM（CR-CIM）を提案します．具体的には演算に用いる容量を再利用し高精度 AD 変換器を実現し，従来研究よりも一桁高い演算精度を追加コストなく実現します．

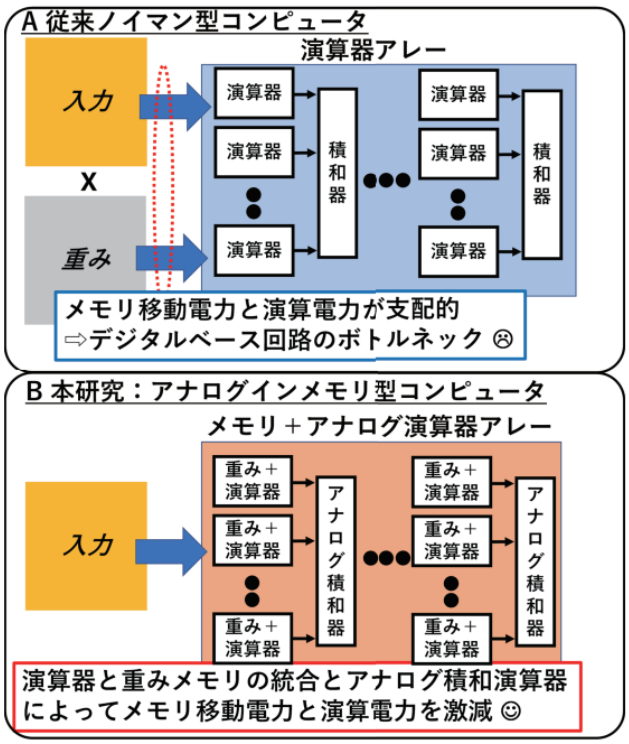

図 1 本研究で開発するインメモリアクセラレータ

公益財団法人電気通信普及財団
研究調査助成報告書 第39号 2024年度

2 研究内容

2-1 CR-CIM 構造による演算高精度化

Transformer モデルの推論を実現するためには，アナログ演算に使用する回路素子の精度と，アナログ値をデジタル信号に変換する A/D 変換器の精度を向上させる必要があります．しかし，従来の CIM 技術では，精度向上に伴いA/D変換器の素子数が増加し,A/D変換部分が巨大になってしまうという問題がありました．これが，アナログ演算を利用した CIM での Transformer 推論が難しい主な理由です．
本研究では，データ格納用メモリ，演算，A/D 変換素子を 1 つのメモリセルに統合した新しい「容量再構成型 CIM（CR-CIM)」構造を提案し，このボトルネックを解消しました（図 2)．CR-CIM では，演算時に使用される回路素子を A/D 変換にも再利用することで，精度向上に伴う A/D 変換部分の素子増加を最小限に抑えることができます．この革新により，Transformer 推論に必要な演算精度を達成し，具体的には 10 ビットの A/D 変換機能を備えつつ，従来の研究[1]に比べてトランジスタ数を 60%削減しました（図 3)．

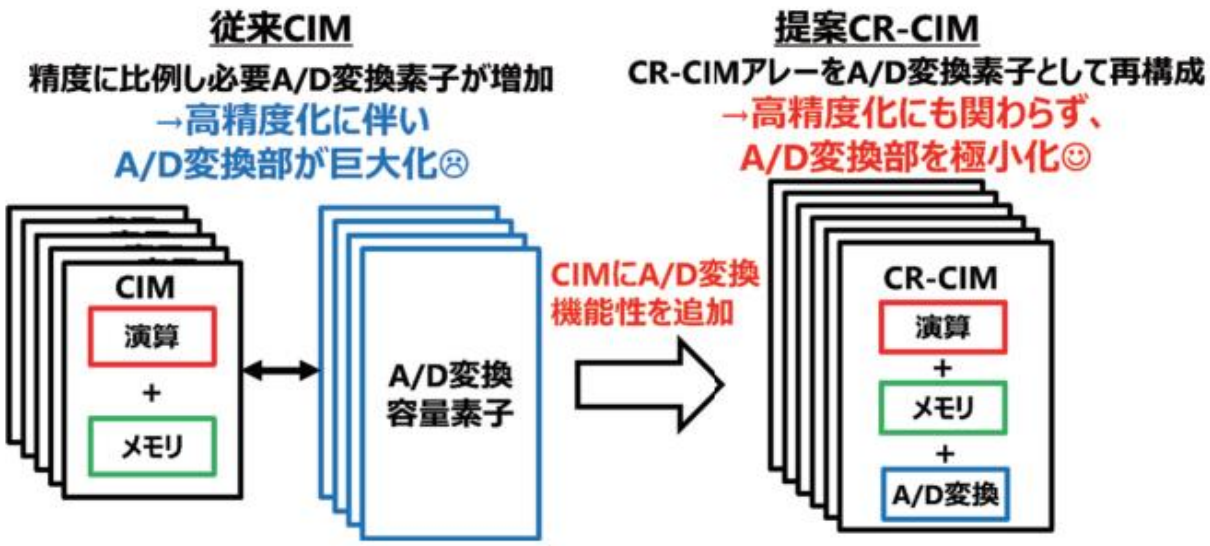

図 2 CR-CIM のコンセプト．CR-CIM メモリセルがデータ記憶，演算，そして A/D 変換 3 つの機能を備えるため，A/D 変換回路を高精度化しつつ面積増加を最小限に抑えることが可能.

高い面積効率と CSNR の両方を達成するために設計された、提案された CR-CIM 回路の説明を以下で実施します。CR-CIM は、再構成可能なキャパシタアレイを利用して、2 つの機能を果たします。それは、電荷ベースの計算と、10b キャパシタ DAC（C-DAC）を使用した 10b ADC の動作です。大面積の C-DAC を CIM 計算キャパシタと共有することで、10b ADC の追加の面積オーバーヘッドを最小限に抑えています。計算フェーズでは、入力（IN）と 6T SRAM に格納された重みの積が CIM セルキャパシタに供給されます。ADC フェーズでは、CIM キャパシタが DAC フィードバック信号（DDAC[9:0]）に接続され、C-DAC アレイへの再構成が可能になります。DDAC[9]信号は 512 個の CIM セルに、DDAC[8]信号は 256 個の CIM セルに接続され、以下同様に、バイナリ C-DAC アレイを形成します。CR-CIM のキャパシタアレイの上部プレートはコンパレータに直接接続されており、逐次近似（SA）を使用して 10b AD コンバージョンを実現できます。
CR-CIM と類似のメカニズムが[1]で紹介されていますが、そのデザインにはいくつかの構造的な制限があります。まず、1b 入力信号の処理のみに制限され、ビット並列動作をサポートしていません。次に、ADC 参照電圧は専用の参照 CIM アレイ内で生成する必要があり、CIM の面積効率を大幅に低下させます。対照的に、我々の CR-CIM セルは PMOS パスゲートを採用し、マルチレベル入力信号をキャパシタに接続することで、ビット並列動作を可能にしています。さらに、DDAC/リセット共有構造により、リセットおよび ADC 動作に関与するトランジスタ数を最小限に抑えています。DDAC パスは、キャパシタのリセットにも再利用され、セル内のリセットスイッチを不要にしています。この設計により、コンパクトな 10T セルが実現されています。共有 DDAC/リセット[9:0]ノードは、リセット（RST）時にリセット電圧に接続され、ADC フェーズ中はグローバ

ルスイッチによって DDAC[9:0]に切り替えられます。CIM セルキャパシタは、カスタムの 1.5fF フリンジキャパシタで実現され、ポストレイアウトシミュレーションで 10b の線形性が確認されました。CR-CIM セルの面積は、65nm ロジックルールを使用して 2.3μm2 です。

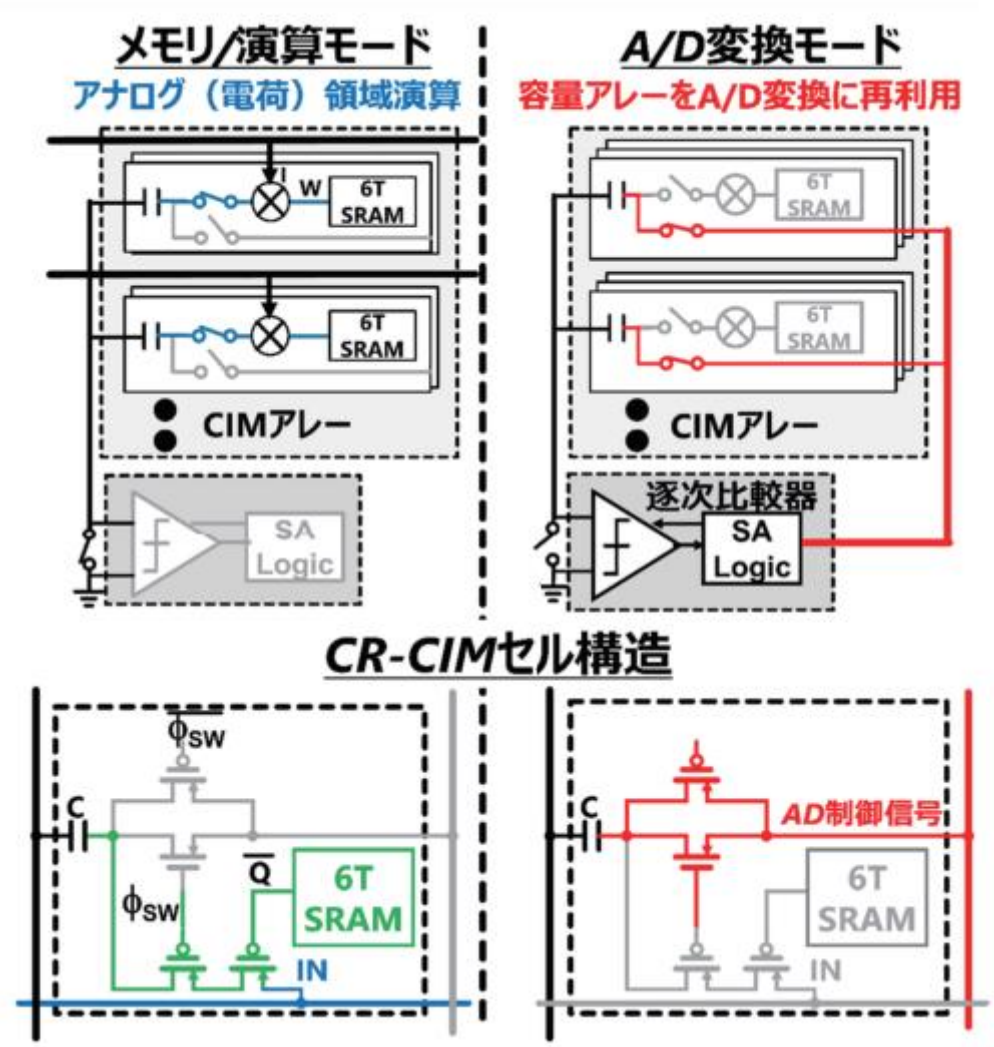

図 3　CR-CIM の詳細動作とメモリセル構造．CR-CIM は省トランジスタ構造でありながら，データ記憶，演算，そして A/D 変換 3 つの機能の集約を実現した．

2-2 Transformer と CNN のハイブリッド・アクセラレータの実現

Transformer は自然言語処理に，CNN は画像認識など，それぞれ得意とするタスクが異なる深層学習アーキテクチャです．Transformer と CNN を組み合わせたアーキテクチャは音声認識などに適しており，多様な機械学習タスクに対応するには両方を実行可能なアクセラレータが求められます．しかし，Transformer と CNN では演算精度要求が異なるため，両者の要求を満たすアクセラレータの設計は容易ではありません．
本研究の CR-CIM は，この課題に対する独自のアプローチを提案しています．CR-CIM は，CNN 動作時には低精度・高効率演算モード，Transformer 動作時には高精度演算モードで動作することで，Transformer と CNN の両方の要求を満たすハイブリッド・アクセラレータを実現します．

CNN 動作時には，従来のアナログ CIM の積和演算において和部分のみアナログで実施していた手法を発展させ，掛け算もアナログ領域で実施する Bit-parallel 動作を採用しました．これにより，演算効率を 5 倍に向上させることに成功しました（図 4）．Bit-parallel 動作は演算精度を犠牲にしますが，CR-CIM では CNN に十分な精度を達成しているため，アルゴリズム精度の劣化はほとんど見られません．

一方，Transformer 動作時には，高精度演算モードを使用することで，Transformer が要求する演算精度を確保します．CR-CIM は，メモリセル内でデータ記憶，演算，A/D 変換を行うことで，高精度な演算を実現しています．
CR-CIM のハイブリッド・アクセラレータとしての設計は，Transformer と CNN の両方を効率的に実行可能にし，多様な機械学習タスクに対応するための重要な一歩となります．今後，CR-CIM の技術をさらに発展させることで，より高度な深層学習アプリケーションの実現が期待されます．

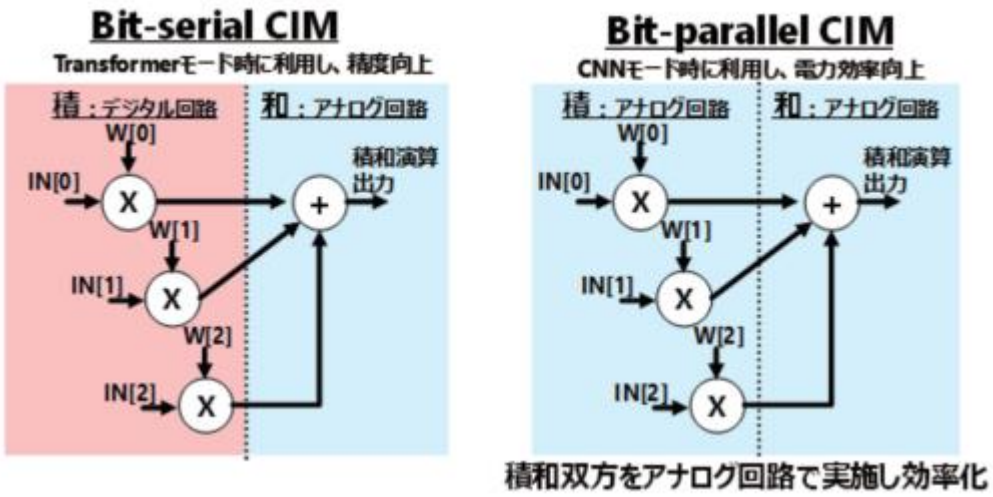

図 4 CNN モードで採用する Bit-parallel CIM 動作における積和演算計算のイメージ．積・和ともにアナログ回路で計算することで，さらに演算エネルギーを低減する．

また Bit-serial 計算を可能にするリソース効率の高いマルチビット行ドライバを提案します。より低い CSNR が許容される CNN モードでは、アナログマルチビット入力を利用してビット並列計算を容易にし、さらなる効率を達成します。従来の ACIM 設計では、ドライバの実現のために 10 以上の正確な基準電圧を利用しています。このような実装では、複数の基準電圧を生成するために低抵抗の R ラダーが必要となるため、電力のオーバーヘッドが増大します。この問題に対処するため、我々は単一の基準電圧を用いた 5b ドライバを提案し、C-DAC で補強しています。しかし、ACIM で単純な電荷再分配 C-DAC を利用すると、大きな電圧エラーが発生してしまいます。ドライバから見た容量性負荷は、集約された行の重み(SW)に基づいて変動するためです。異なる SW による容量性負荷の変動を緩和するために、動的負荷補償(DLC)回路を導入しています。DLC は、SW に反比例して負荷コンデンサを追加し、重み構成に関係なく一定の負荷容量を維持します。DLC 回路は、性能と面積のバランスを最適化するために 4b の解像度で設計されており、初期のエラーマージンを 26 から 1.2-LSB まで減少させています。さらにドライバのオーバーヘッドを削減するために、SW はスケジューリング段階で事前に計算されて保存され、MOM コンデンサはコアメタル層のすべての利用可能な層を占有して、コンデンサ密度を最大化しています。これにより、ドライバアレイの大きさは約 10 個の CIM 列の大きさに抑えられています。

2-3 Transformer 用低雑音回路の設計

ソフトウェアとアナログの協調設計によって、トランスフォーマーモードの電力効率を向上させるアプローチを提案します。典型的なトランスフォーマーアーキテクチャは、2 つの主要なレイヤーで構成されています。1 つは、画像パッチからの特徴ベクトルの重み付き混合を担当するアテンションレイヤーであり、もう 1 つは、その後に続くマルチレイヤーパーセプトロン(MLP)レイヤーです。シミュレーションに基づくと、アテンションレイヤーに必要な CSNR は、MLP レイヤーに必要な CSNR よりも 10dB 低いことがわかりました。我々は、読み出し精度と消費電力のトレードオフを可能にする CSNR ブースト(CB)技術を提案しています。アクティブなトランスフォーマーレイヤーに応じて CB 技術を適応的に活性化することで、ADC の電力を 30%改善しています。CB が有効になっている場合、最後の 3 つの SA 比較に 6 倍の多数決が適用され、CSNR が 5.5dB 改善されますが、電力と変換時間がそれぞれ 1.9 倍と 2.5 倍に増加します。

3 研究成果

本 CIM は TSMC 社の 65 nm プロセスでチップを設計・試作し，Transformer モードで最大 1.2 TOPS，ピーク電力効率 818 TOPS/W を達成しました（図 5）．また CNN モードで最大 6 TOPS，ピーク電力効率 4094 TOPS/W を達成し，これは同等の演算精度を達成する[2]に対し 10 倍高い電力効率です．

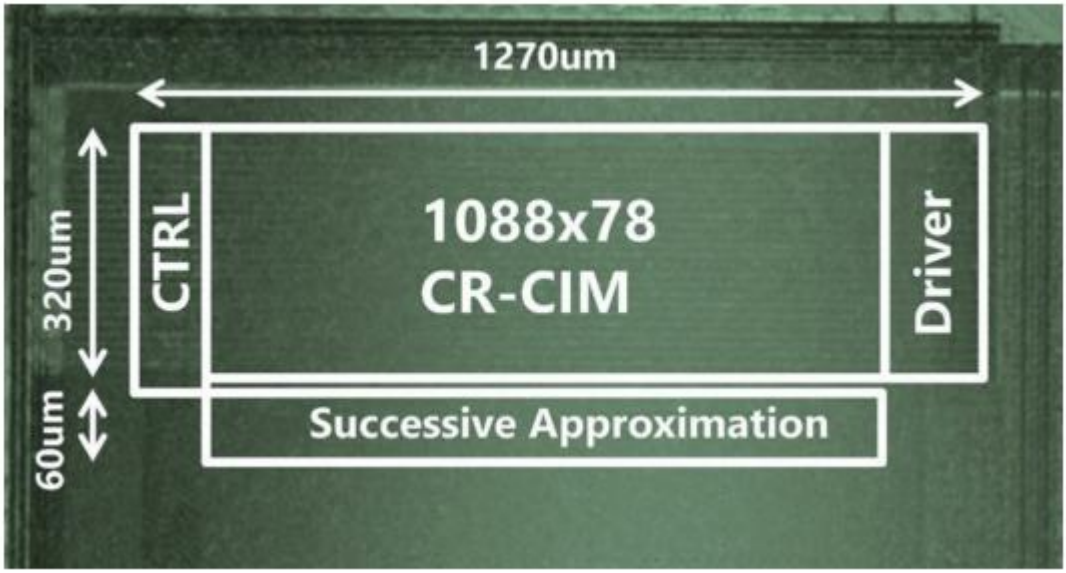

図 5 65nm CMOS で試作した CIM チップ

従来アナログ CIM に比べると，量子化雑音比（SQNR）は 22 dB 高く，演算精度（CSNR）は 13 dB 高い性能を達成しました．これにより，高効率ながら Transformer に十分な計算精度を達成しました（図 6）．

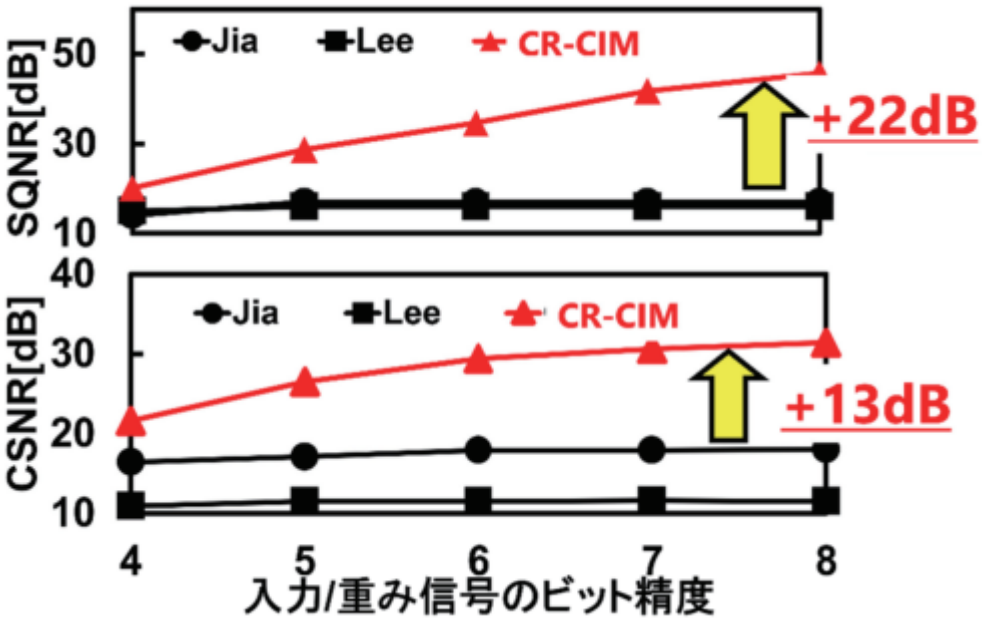

図 6 従来 CIM 研究（Jia[2]，Lee[3]）と本 CR-CIM と演算精度（SQNR，CSNR）の比較．CR-CIM によって A/D 変換器の高精度化を達成し，従来研究よりも大幅に高い演算精度を実現した．

表 1 性能まとめ

	This work		[3] JSSC 2020	[4] ISSCC 2023	[5] VLSI 2021	[6] JSSC 2023
CIM type	Charge (MAC)		Charge	Charge	Charge	Charge
Process	65nm		65nm	12nm	28nm	28nm
Supply Voltage[V]	0.6-1.1		0.85-1.2	0.5/0.85	0.6-0.9	0.8
Array Size	10KB		72KB	128KB	36KB	16KB
Bit Precision	4-8b		1-8b	1-8b	1-5b	1-16b
Application	CNN	Trans-former	CNN	CNN	CNN	CNN
Peak TOPS Normalized to 1-b	6	1.2	2.1	6.4	6.1	1.31
Peak TOPS/mm^2 Normalized to 1-b	12.5	2.5	0.6	N.A.	12	27.7
Peak TOPS/W Norm. to 1-b	4094*	818*	400	4534	5796	383
Peak TOPS/W Norm. to 65nm**	4094	818	400	837	2496	165
ADC bit	8	10	8	8	8	8
SQNR [dB]	26.7	45.3	22	N.A.	17.5	N.A.
CSNR [dB]	16.8	31.3	17	N.A.	10.5	N.A.
CIFAR-10 Accuracy	91.7	95.8	92.4	N.A.	91.1	N.A.

*Include power of CIM core, driver, clock, ADC but do not include IO circuits
**Assuming a linear reduction in power relative to CMOS process

アルゴリズムでは Vision Transformer (ViT-S)モデルを用いた際に，CIFAR10 データセットにて 95 %と高い正答率を確認しました．また，高効率な CNN モード使用時は Resnet-20 モデルで同データセットにて 91 %の精度を確認しました．

【参考文献】

[1] C. Y. Yao et al, “A Fully Bit-Flexible Computation in Memory Macro Using Multi-Functional Computing Bit Cell and Embedded Input Sparsity Sensing,” IEEE JSSC, vol. 58, no. 5, pp.1487-1495, May 2023

[2] H. Jia et al, “A Programmable Heterogeneous Microprocessor Based on Bit-Scalable In-Memory Computing,” IEEE JSSC, vol. 55, no. 9, pp. 2609–2621, Sept. 2020

[3] J. Lee et al, “Fully Row/Column-Parallel In-memory Computing SRAM Macro employing Capacitor-based Mixed-signal Computation with 5-b Inputs,” IEEE Symp. on VLSI Circuits, 2021

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
A 818 - 4094 TOPS/W Capacitor-Reconfigured CIM Macro for Unified Acceleration of CNNs and Transformers	IEEE ISSCC	2024/2 月
OSA-HCIM: On-The-Fly Saliency-Aware Hybrid SRAM CIM with Dynamic Precision Configuration	IEEE ASP-DAC	2024/1 月