プレスリリース

2024 年 2 月 19 日

慶應義塾大学
科学技術振興機構（JST）

アナログコンピュート・イン・メモリ回路で Transformer と CNN のハイブリッド処理を世界で初めて実現
― 対従来比 10 倍の演算エネルギー効率を達成し、AI の環境負荷を低減 ―

慶應義塾大学理工学部電気情報工学科の吉岡健太郎専任講師は、エッジコンピューティングの普及に伴い、より身近なデバイスへの人工知能（AI）応用を促進するため、深層ニューラルネットワーク（DNN）、特に Transformer 処理の高効率な推論を実現する高精度かつ省エネルギーなコンピュート・イン・メモリ（CIM）回路を開発しました。

本研究では、従来の CIM が抱えていた Transformer の推論に必要な演算精度を実現するために、データ格納、演算、アナログ－デジタル（A/D）変換を 1 つのメモリセルに集積した「容量再構成型 CIM（CR-CIM）」構造を提案しました。この構造によって、アナログ CIM で初めて Transformer 処理に必要な演算精度を達成しつつ、消費電力 1 W あたりの処理速度が 818 TOPS（兆回／秒）と非常に高い電力効率を実現しました。また畳み込みニューラルネットワーク（CNN）処理を行う際は、同等の演算精度を持つ従来技術と比べ 10 倍のエネルギー効率となる 4094 TOPS/W を達成しました。

本研究成果は、エッジコンピューティングや AI の分野で、電力効率と処理速度の両面で効率的な AI ハードウェアの開発に貢献します。また将来的にはより多くの人々が大規模言語モデル（LLM）といった AI サービスを利用しやすくなると期待されます。

研究成果の詳細は、米国時間 2 月 18 日から開催されている「ISSCC2024（国際固体素子回路会議）」にて発表されました。ISSCC は集積回路に関するオリンピックとも称される、最難関国際会議です。

1. 本研究のポイント

・**Transformer アーキテクチャへの対応：**Transformer※1 は、その柔軟性とタスク精度により、LLM を始め近年 AI 分野で広く採用されています。しかし、Transformer は高い演算精度を要求するため、従来の CIM 回路※2 では処理が困難でした。本研究で開発された新型 CIM 回路は、革新的なアナログ演算機構を採用することで、アナログ機構を用いる CIM として世界で初めて Transformer 処理に成功し、高演算精度と省エネルギーを両立させました。

・**高精度かつ省エネルギーな演算機構：**本 CIM 回路では、演算時とデータ読み出し時に回路内部の素子を効率的に流用する「容量再構成型 CIM」を採用しています。この構造により、高精度な演算能力と省エネルギーを同時に達成しました。これにより、Transformer と CNN の両方を効率的に実行する CIM を実現しました。

・**AI アクセシビリティの拡大：**本研究は、エッジデバイス※3 での AI 演算の効率化を目指しています。研究の進展に伴い、スマートフォンや自動車などの身近な電池駆動デバイスでも、ハードウェアの効率化を通じて、LLM などの高度な AI 処理が可能になります。これにより、より多くの人々が AI 技術に容易にアクセスできるようになることが期待されます。

2. 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。

現在の AI ハードウェアでは、メモリと演算器間のデータ通信がボトルネックとなっています。この問題への対応策として、CIM 計算機構が注目されています。CIM は、メモリ内で演算を行うことでデータ通信を削減し、従来の AI ハードウェアのボトルネックを解消します。

デジタル演算は誤差が生じない利点がありますが、演算回路の電力が大きくなってしまうのが問題です。そのため、電力と面積効率を重視するエッジデバイスに向けた CIM では、高い電力効率が得られるアナログ演算を用いて演算回路を実装します。アナログ演算回路はデジタル演算と比較し、データを連続値で処理することができるため、より少ない電力で高速に演算が可能です。しかしながら、アナログ演算は画像処理に用いられる CNN では十分な演算精度を提供しますが、より高い演算精度が求められる Transformer では不足しているとされています。さらに、最近では Transformer と CNN を組み合わせたアーキテクチャも登場しており、従来の CIM 技術では、これらの多様な機械学習の処理内容に対応するのが難しい現状があります。

3. 研究内容・成果

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 1)。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 2)。

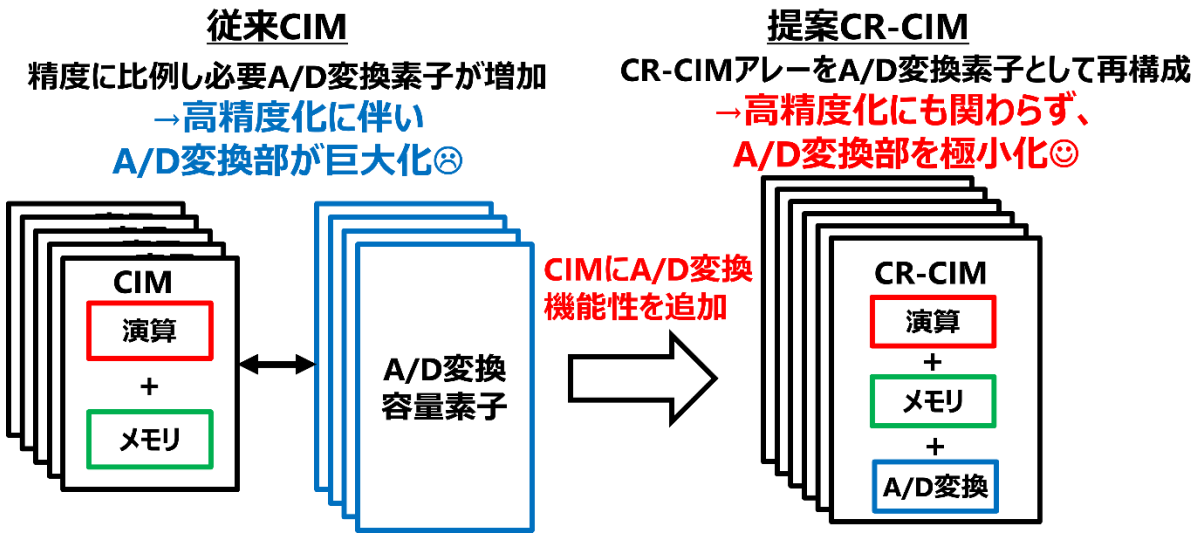

図 1 CR-CIM のコンセプト。CR-CIM メモリセルがデータ記憶、演算、そして A/D 変換 3 つの機能を備えるため、A/D 変換回路を高精度化しつつ面積増加を最小限に抑えることが可能。

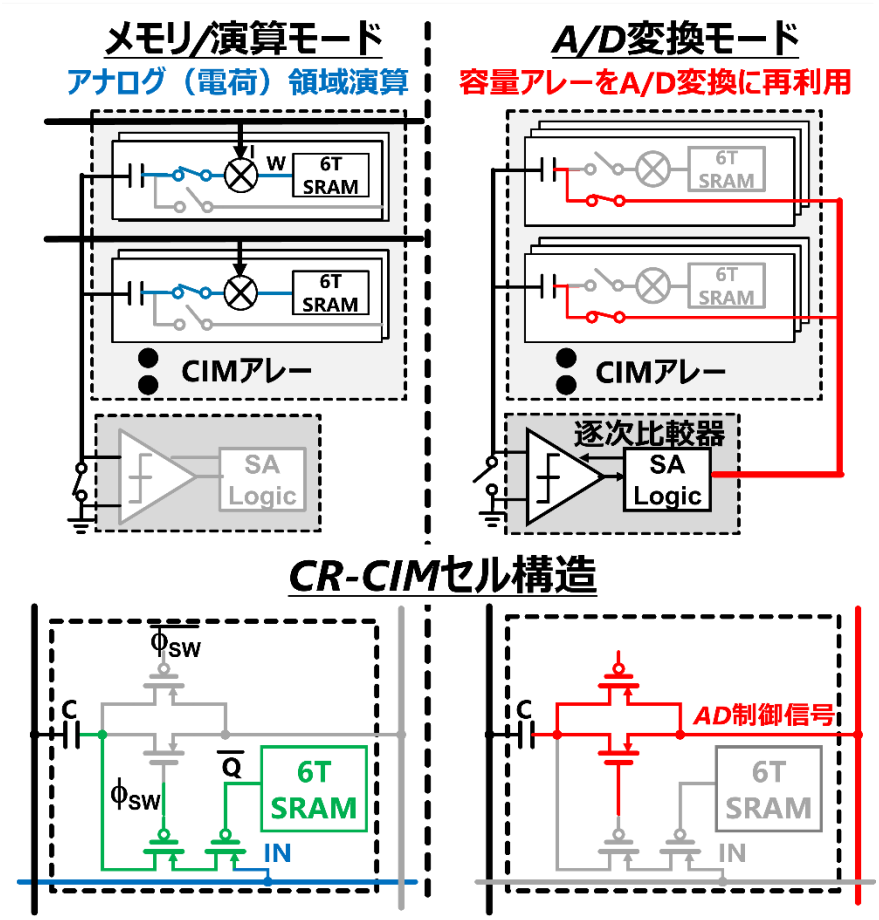

図 2 CR-CIM の詳細動作とメモリセル構造。CR-CIM は省トランジスタ構造でありながら、データ記憶、演算、そして A/D 変換 3 つの機能の集約を実現した。

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

Transformer と CNN は、深層学習の代表的なアーキテクチャであり、それぞれに得意とするタスクがあります。Transformer は自然言語処理、CNN は画像認識などのタスクに適しています。また、Transformer と CNN を混ぜたアーキテクチャは、音声認識などに適しており、多くの機械学習タスクに対応するには、Transformer と CNN の両方を実行可能なアクセラレータが必要です。しかし、Transformer と CNN では演算精度要求が異なるため、双方の要求を満たすアクセラレータの設計は困難です。

そこで、本研究グループの CR-CIM では、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 の双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した。具体的には、CNN 動作時には、従来のアナログ CIM の積和演算では和部分のみアナログで実施していましたが、掛け算もアナログ領域で実施する Bit-parallel 動作により、演算効率を 5 倍向上させました（図 3）。これは演算精度を犠牲にしますが、CR-CIM では CNN には十分すぎる精度を達成しているため、Bit-parallel 動作を行ってもアルゴリズム精度の劣化はほとんど見られません。

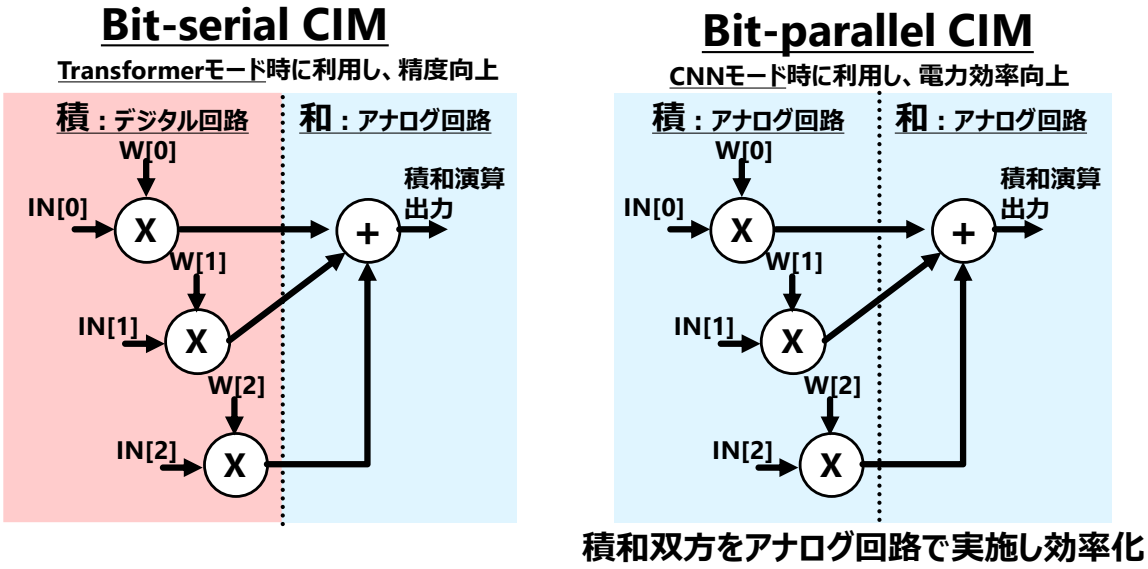

図 3 CNN モードで採用する Bit-parallel CIM 動作における積和演算計算のイメージ。積・和ともにアナログ回路で計算することで、さらに演算エネルギーを低減する。

3. LSI 試作と評価

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 4）。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。

従来アナログ CIM に比べると、量子化雑音比（SQNR※4）は 22 dB 高く、演算精度（CSNR※5）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 5）。

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。

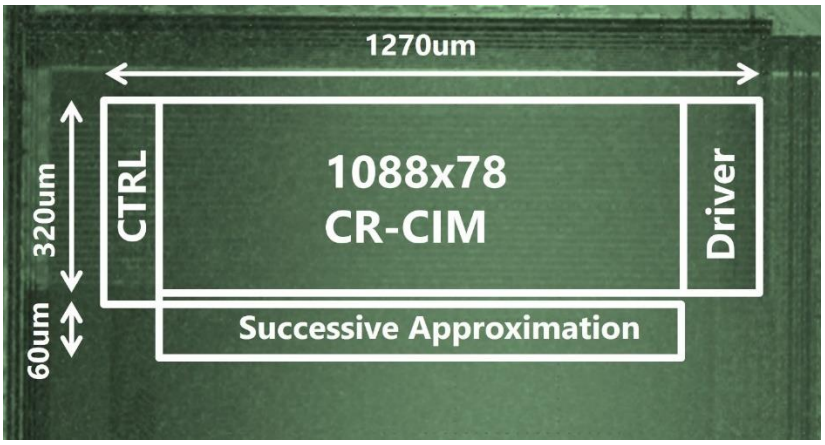

図 4 65 nm CMOS プロセスにて試作した 1088x78 アレーCR-CIM のチップ写真

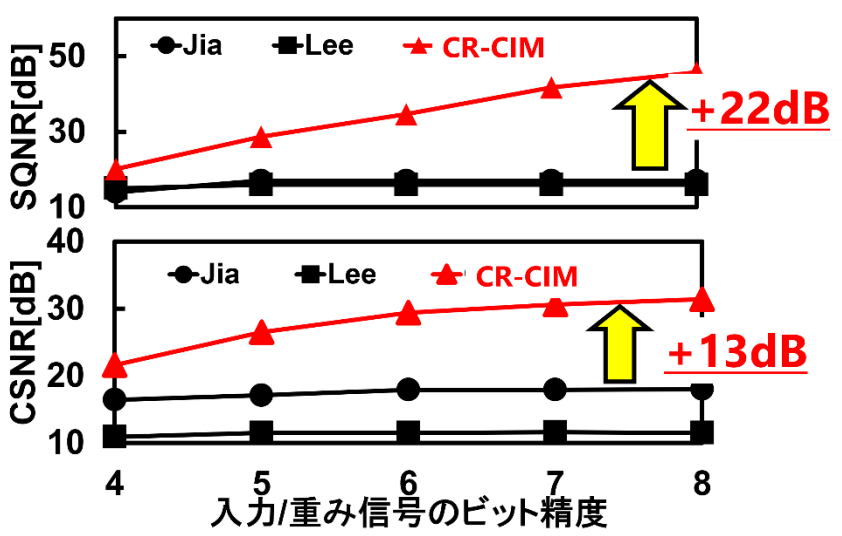

図 5 従来 CIM 研究（Jia[2]、Lee[3]）と本 CR-CIM の演算精度（SQNR, CSNR）の比較。CR-CIM によって A/D 変換器の高精度化を達成し、従来研究よりも大幅に高い演算精度を実現した。

4. 今後の展開

・エッジコンピューティングの分野では、データ処理の高速化とエネルギー消費の削減が重要な課題です。この技術を活用することにより、自動運転車、スマートシティ、モバイルデバイスなど、より身近な分野への拡大が可能となり、人々の AI 利用を促進することを目指しています。

・大規模 AI の環境負荷は、無視できない問題です。本技術の実用化によって、AI の環境負荷を低減し、持続可能な AI の実現を目指します。本研究グループの CIM は、従来のアナログ CIM に比べて、演算精度を犠牲にすることなく、エネルギー効率を大幅に向上させることができます。そのため、本 CIM を大規模 AI に適用することで、エネルギー消費を抑えることが期待されます。

・今回開発した CIM は、CNN と Transformer の両方のモデルを効果的に推論可能なハイブリッド・アクセラレータを実現する最初のステップです。将来的には、この回路技術をさらに進化させることで、演算精度とエネルギー効率をより一層高めることが期待されます。また、大規模言語モデルなどのより大きなモデルの推論処理も目指しています。

5. 本プロジェクトについて

本研究成果は、科学技術振興機構（JST） 戦略的創造研究推進事業 チーム型研究（CREST）「信頼される AI システムを支える基盤技術」研究領域（研究総括：相澤 彰子）「D3-AI: 多様性と環境変化に寄り添う分散機械学習基盤の創出」（研究代表者：高前田 伸也、グラント番号：JPMJCR21D2）の事業・研究課題の助成により得られました。

＜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

＜原論文情報＞

国際学会名：International Solid-State Circuits Conference（ISSCC）2024

タイトル：*A 818-4094TOPS/W Capacitor-Reconfigured CIM Macro for Unified Acceleration of CNNs and Transformers*

著者と所属： Kentaro Yoshioka

Keio University

＜用語説明＞

※1 Transformer：自然言語処理を始めとする AI 分野で広く用いられているモデル構造。繰り返しの層を持たず、代わりに「自己注意」機構を使用して入力データ間の関係性を捉えることで、文脈の理解や予測に優れた性能を発揮する。特に文章生成、機械翻訳、要約、画像処理などのタスクで高い精度を示している。

※2 CIM（コンピュート・イン・メモリ）回路：演算とメモリを同じ回路上で行うことで、データ転送に関わる時間とエネルギー消費を大幅に削減する技術。その上でアナログ CIM 回路は演算もアナログ領域で行うことで、さらなる演算電力削減を達成する技術。CIM 技術はエッジコンピューティングやモバイルデバイス向けに研究が進んでいる。

※3 エッジデバイス：データ処理をクラウドではなく、スマートフォン、センサー、監視カメラ、自動車などのデバイス上で直接行うシステム。これにより、レイテンシー（遅延）を低減し、データプライバシーを保護することが可能となる。これらのデバイスはデータを収集し、分析してアクションを起こし、IoT（モノのインターネット）アプリケーションにおける重要な役割を担う。

※4 SQNR（Signal-to-Quantization-Noise Ratio）：信号処理で使われる用語で、信号の強さと A/D 変換によって混入される量子化ノイズとの比率を示す。この場合は、CIM の A/D 変換回路がどれほど高精度化しているかを示す。

※5 CSNR（Compute Signal-to-Noise-Ratio）：CIM 計算の精度を測る指標で、アナログ演算や A/D 変換によって混入される雑音がどれほど低く抑えられているかを示す。

・研究内容についてのお問い合わせ先
慶應義塾大学 理工学部 電気情報工学科　専任講師　吉岡　健太郎（よしおか　けんたろう）
E-mail：kyoshioka47[at]keio.jp

・本リリースの配信元
慶應義塾広報室（望月）
TEL：03-5427-1541　FAX：03-5441-7640
E-mail：m-pr[at]adst.keio.ac.jp　https://www.keio.ac.jp/

科学技術振興機構 広報課
TEL：03-5214-8404　FAX：03-5214-8432
E-mail：jstkoho[at]jst.go.jp

・JST 事業に関すること
科学技術振興機構 戦略研究推進部 ICT グループ　前田　さち子（まえだ　さちこ）
TEL：03-3512-3526　FAX：03-3222-2066
E-mail：crest[at]jst.go.jp