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．

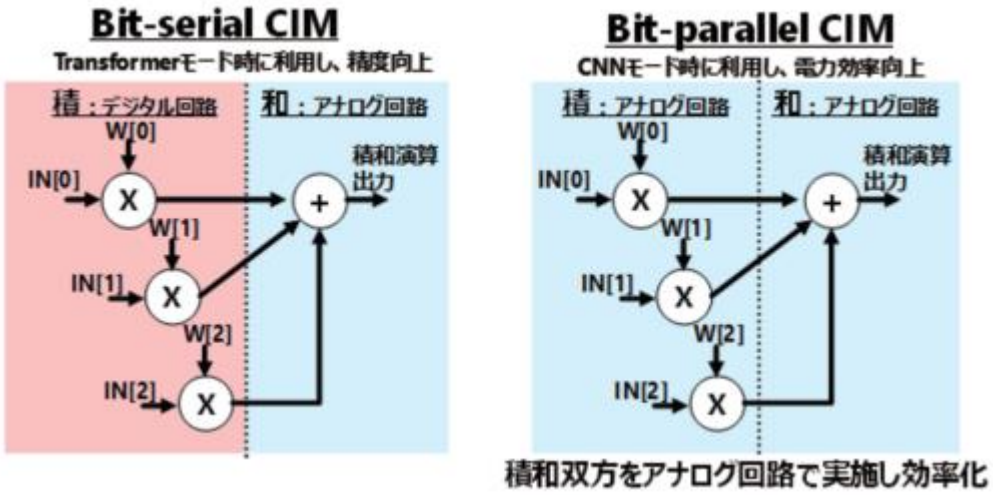

図 4 CNN モードで採用する Bit-parallel CIM 動作における積和演算計算のイメージ．積・和ともにアナログ回路で計算することで，さらに演算エネルギーを低減する．

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．

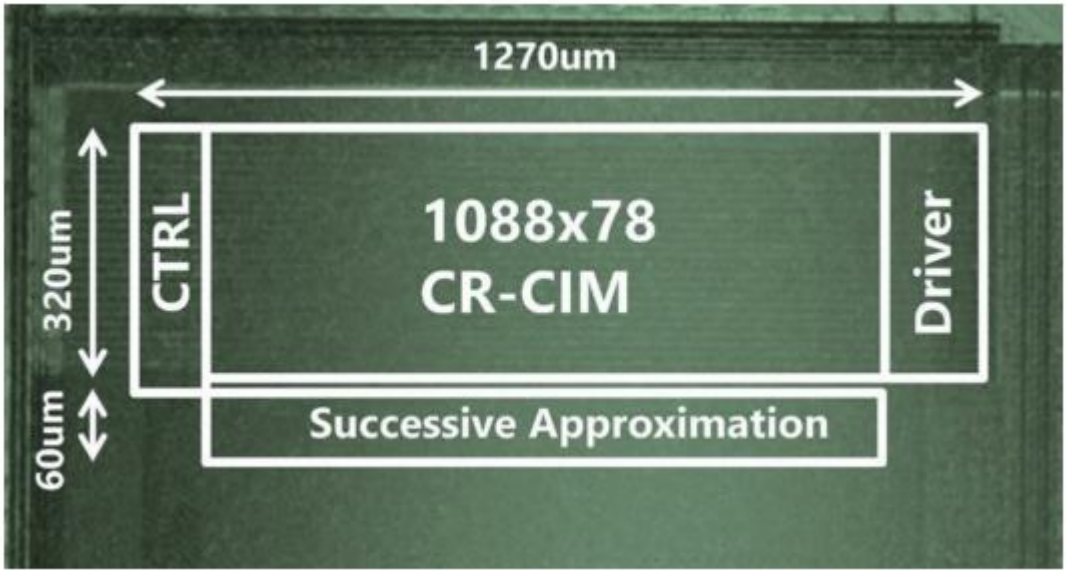

図 5 65nm CMOS で試作した CIM チップ

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）．

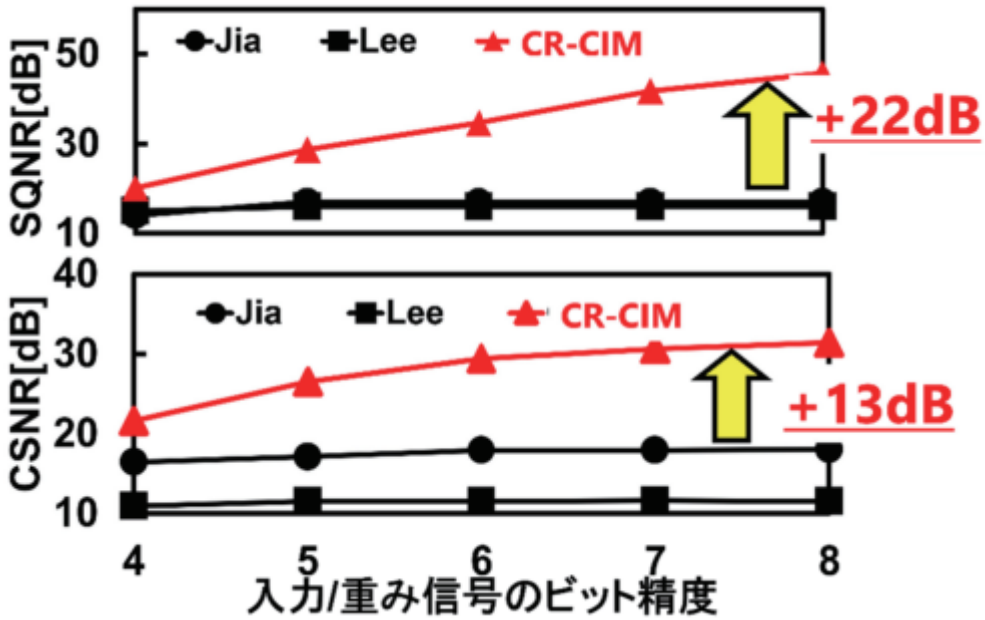

図 6 従来 CIM 研究（Jia[2]，Lee[3]）と本 CR-CIM と演算精度（SQNR，CSNR）の比較．CR-CIM によって A/D 変換器の高精度化を達成し，従来研究よりも大幅に高い演算精度を実現した．

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