参考資料

2023 年 4 月 26 日
東京工業大学

エネルギー最小点で動作する AI 半導体（ニューラルネットワーク・アクセラレータ）技術の開発に成功

－モバイルエッジ高性能 AI 技術－

【要点】

○エネルギー最小点における SRAM 動作と、超低電圧リテンションによるパワーゲーティングの両方を実現できる新たな SRAM 技術を用いて、プロセッシング・イン・メモリ（PIM）型のニューラルネットワーク（NN）アクセラレータのマクロを開発。

○動作時電力を 99%、待機時電力を 84%削減し、推論のエネルギー効率の飛躍的な増大に成功。全結合層において 65 TOPS/W の高いエネルギー効率を実現可能。

○本技術は、将来のスマート社会で重要となるモバイルエッジデバイスに搭載可能な低消費電力・高性能 AI 技術として期待。

【概要】

東京工業大学 科学技術創成研究院 未来産業技術研究所の菅原聡准教授と工学院 電気電子系の塩津勇作博士後期課程大学院生（研究当時）らは、エネルギー最小点（**EMP**、用語 1）動作によって動作時電力を 99%削減し、また、パワーゲーティング（PG）によって重みデータを失うことなく待機時電力を 84%削減できる、プロセッシング・イン・メモリ（PIM）型のニューラルネットワーク（NN）**アクセラレータ**（用語 2）の**マクロ**（用語 3）を開発した。本技術は CMOS のみで構成できる。この PIM 型マクロのメモリ部には EMP で動作し、超低電圧リテンション（ULVR）による PG が可能な新しい SRAM 技術を用いた（**ULVR-SRAM**、用語 4）。ULVR-SRAM の導入によって可能となる EMP 動作を用いることで推論のエネルギー効率（**TOPS/W**、用語 5）は大幅に向上し、また、許容される積和（MAC）演算の並列数も大幅に増やすことができることから、演算能力（TOPS）も大きく向上できる。本マクロを用いて MAC 演算の適切な並列化を行い EMP 動作させることで、通常電圧動作時や従来技術に比べて、演算能力が同じであれば、1/10 程度の消費電力で済み、また、消費電力が同じであれば、10 倍程度の演算能力を実現できる。全結合層を用いたベンチマークから、この EMP による推論では、65 TOPS/W もの高いエネルギー効率を示した。本技術は将来のスマート社会で重要となる**モバイルエッジ**（用語 6）デバイスに搭載可能な低消費電力・高性能 AI アクセラレータ技術となる。

本成果は IEEE（米国電気電子学会）の「*IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*」誌に 12 月 22 日付けで掲載された。

●背景

将来のスマート社会では、スマートフォンなどのスマートモバイルエッジ階層のコンピューティングシステムに AI 技術を導入することが期待されており、これまで以上に動作時電力および待機時電力がともに低く、エネルギー効率の高いニューラルネットワーク（NN）アクセラレータが要求されている。NN アクセラレータには様々な構成方法があるが、processing-in-memory（PIM）型のハードウェアは、メモリアレイの中または近傍に演算ユニットを配置することで、バスを介したデータ転送を用いることなく、データの処理が可能なことから、NN アクセラレータの高性能化に有効である。特に、SRAM を用いた PIM 型 NN アクセラレータは現状の CMOS 技術で実装が可能なため、応用上極めて重要である。

NN アクセラレータのエネルギー効率を向上させるためには、エネルギー最小点（EMP）となる駆動電圧（V_{EMP}）を用いた推論動作が極めて効果的になる。EMP 動作には許容される積和（MAC）演算の並列数を増大させ、演算能力を向上させるという効果もある。また、モバイルエッジ応用ではパワーゲーティング（PG）による待機時電力の削減も必須であるが、従来の 6T セルを SRAM に用いた場合、EMP 動作および PG を実現することはできない。特殊なセル構成を用いて EMP 動作または PG のいずれかを実現できても、2 つを同時に実装した例はない。

最近、本研究グループは、通常電圧下（V_{DD}）では従来の SRAM と同様の高性能 SRAM 動作を実現し、0.2 V 程度の超低電圧（V_{UL}）でもデータを失うことなく保持できる超低電圧リテンション SRAM（ULVR-SRAM）セルを提案した（図 1）。ULVR-SRAM では、V_{UL} でデータ保持（ULVR）を行うことで実質的な PG を実現できるため、SRAM であっても待機時電力の大幅な削減が可能となる。さらに、ULVR-SRAM では、このセルの特徴から EMP 動作も実現できる可能性がある。すなわち、電源電圧を EMP となる電圧 V_{EMP} まで低減させて SRAM 動作させることで、エネルギー効率を最大化できると考えられる。

したがって、PIM の SRAM 部に ULVR-SRAM を用いることで、PG と EMP 動作が可能な NN アクセラレータを実現できると考えられ、従来の SRAM 技術では到達できない高いエネルギー効率を実現できる可能性がある。本研究では、EMP で動作し、PG が可能な ULVR-SRAM 技術を新たに開発し、これを用いた PIM 型 NN アクセラレータのマクロ技術を開発した。

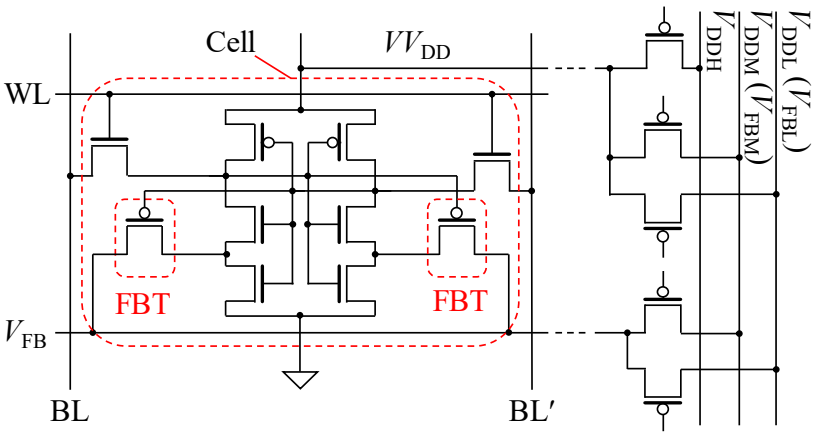

図 1　提案した ULVR-SRAM セルの回路構成

●研究成果

開発した ULVR-SRAM は①通常電圧 V_{DD}（=1.2V）での高性能 SRAM 動作、②EMP となる電圧 V_{EMP}（=0.4V）を用いた最大エネルギー効率 SRAM 動作、③超低電圧 V_{UL}（=0.2 V）を用いたデータ保持（ULVR）による実質的な PG の 3 モード動作を実現できる。

本研究では、はじめに、この 3 モード動作を実現できる ULVR-SRAM セルの設計技術の開発を行った。トランジスタの特性バラツキを考慮して、設計指標には各動作モードにおけるノイズマージンと、ULVR モードにおけるリーク電流を用いた。最適設計された ULVR-SRAM セルを用いて PIM 型 NN アクセラレータ・マクロの設計を行った。今回、NN のアーキテクチャには 2 値化ニューラルネットワーク（BNN）を用いた（本技術は他の多ビットの NN アーキテクチャにも応用できる）。

図 2 に開発した PIM 型 BNN アクセラレータ（BNA）マクロの構成とレイアウトを示す。このマクロ 1 つで 256 b の入力ベクトルに対する積和演算が可能である。この BNA マクロは重みデータとバイアスデータを格納するための ULVR-SRAM を有し、これらの同時読み出しによって効率的に MAC 演算を実行し、アクティベーション判定を行うことができる。

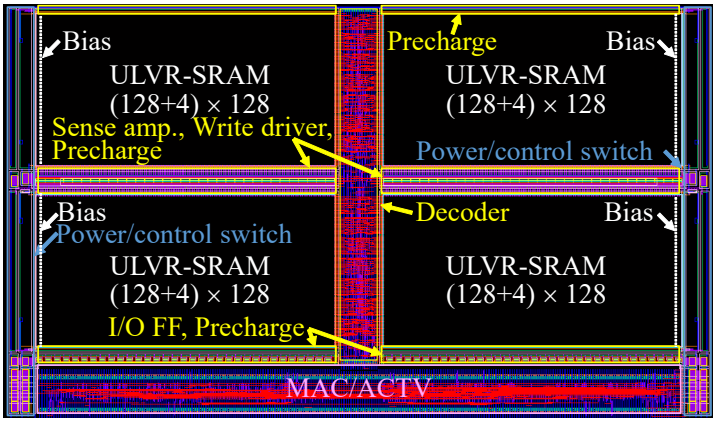

図 2　PIM 型 BNA マクロのレイアウト

開発した BNA マクロの PG 性能について調べ、待機電力削減効果を評価した。この BNA マクロの PG ではメモリセル部は V_{UL}=0.2 V でデータ保持を行い、その他の回路は電源遮断を行う。この PG によって、待機時電力を 84%削減できる（クロックゲーティング時のスタンバイ状態からの比較）。これは従来の 6T セルの SRAM を用いた場合の待機時電力と比較すると 94%の削減率となる。

次に、推論動作時の性能評価を行った。図 3 に BNA マクロにおける EMP 動作の解析結果を示す。電源電圧（図中の V_{DDM}）の減少とともに平均動作時電力（P_{avg}）と動作周波数（f_m）は減少し、0.4 V で消費エネルギー（E_{cyc}）が最小となる。すなわち、V_{EMP} は 0.4 V となる。この V_{EMP} の動作点では通常電圧動作（1.2 V）と比較して動作周波数 f_m は 1/10 となるが、動作時電力は大幅に削減され、1/100 になる（99%減）。この特徴は後述する MAC 演算の並列化に極めて有効となる。また、同図には BNA における推論のエネルギー効率 η も示している。η は V_{EMP} である 0.4 V で最大となり、65 TOPS/W と極めて高い値を実現できる。

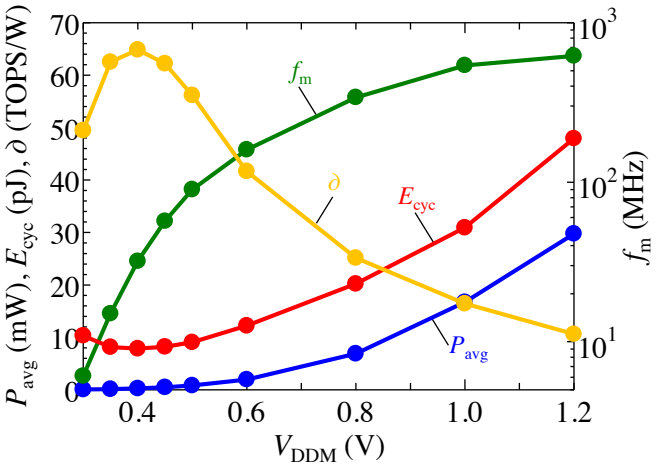

図 3　消費エネルギー E_{cyc}、平均動作時電力 P_{avg}、動作周波数 f_m、エネルギー効率 η の電源電圧 V_{DDM} 依存性

任意のサイズ・形状のネットワークは複数の BNA マクロを用いて構成できる。図 4 に BNA マクロを用いた 1024 ノードの全結合層の構成例を示す。$M_{uv}^{(j)}$ が 1 つの BNA マクロを表し、この場合では 16 個のマクロと追加の加算器で容易に全結合層を構成できる。以下、図中の v 軸方向にある複数の BNA マクロを同時に MAC 演算することをノード内並列化（INP）と呼び、u 軸方向にある複数の BNA マクロを同時に MAC 演算することをレイヤ内並列化（ILP）と呼ぶ。それぞれの並列化による並列数を N_{INP}、N_{ILP} とすれば、全体の並列数 N_P は N_{INP} と N_{ILP} によって与えられる。1 つのマクロ当たり MAC 演算ユニットを 1 つ有する場合、N_P=16 となる。さらに、1 マクロ当たり MAC 演算ユニットを増やすことで、N_P は大きくとれる（開発した BNA マクロでは MAC 演算ユニットの増設は容易である）。

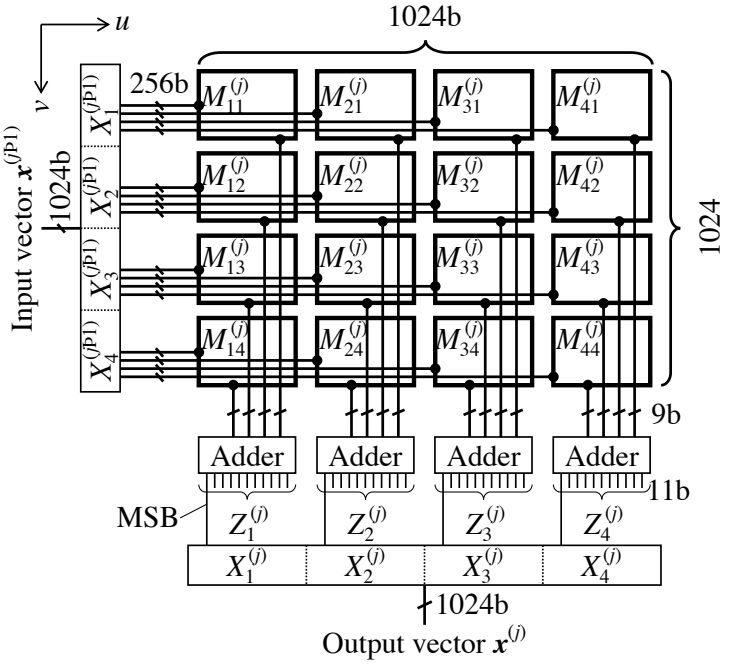

図 4　開発した BNA マクロを用いた全結合層の構成例

図5に並列化の効果を示す。この図は平均動作時電力（P_{FCL}^{MP}）、演算能力（$TOPS_{FCL}^{MP}$）、エネルギー効率（η_{FCL}^{MP}）のネットワーク層数 m 依存性である。全結合層1層あたりのニューロン数は 1024 である。黒線が通常電圧（V_{DD}=1.2 V）における推論モード（$INFER^{Norm}$ モード）の場合で総並列数 N_P を 1 としたとき、青線、緑線、赤線はそれぞれ EMP 電圧（V_{EMP}=0.4 V）にける推論モード（$INFER^{EMP}$ モード）で N_P=1、16、128 とした場合を示している。N_P=1 の場合に現れているように、EMP による演算能力の劣化の効果は消費電力の削減率より小さい。したがって、MAC 演算の並列化によって高性能化を実現できる。N_P=16 の場合では、EMP 動作を用いることで 1.2 V 動作時と同程度の演算性能を保ちながら、P_{avg} を 1.2 V 動作時の約 1/10 にできる。また、N_P=128 にすると（1 マクロあたり 8 並列）、P_{avg} は 1.2 V 動作時と同程度であるが、演算量を 10 倍程度に増大できる。また、EMP 動作時の演算効率（η_{FCL}^{MP}）はどの場合でも概ね 65 TOPS/W 程度と非常に高い。以上から、ULVR-SRAM をメモリ部に用いた PIM 型 NN アクセラレータは、待機時電力の削減効果を大きくとれ、また、エネルギー効率に優れ、消費電力の削減および演算能力の高性能化を実現できる。

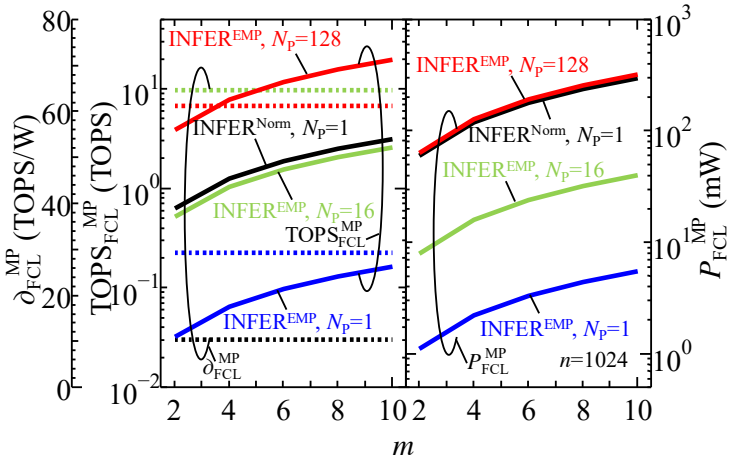

図5　全結合層の平均動作時電力 P_{FCL}^{MP}、演算能力 $TOPS_{FCL}^{MP}$、エネルギー効率 η_{FCL}^{MP}

●社会的インパクト

将来のスマート社会では、AI 技術をクラウドだけでなく、モバイルエッジ階層にまでその応用範囲を拡張することが重要・必要となってくる。モバイルエッジ階層で AI 技術を実現できれば、クラウドを介した通信のレイテンシの問題を解消することが可能となり、また、クラウドとの非接続状態でも使用できる。モバイルエッジ階層への応用では、演算能力とエネルギー効率が高く、低消費電力な AI システムが必要不可欠となり、このためには、これまで以上に高性能なアクセラレータが要求される。本技術は、新しい SRAM 技術（ULVR-SRAM）を用いて構成した PIM 型 NN アクセラレータのマクロ技術であり、以下のような特徴を持つ：(i) モバイルエッジ応用に必須の低消費電力化と高性能化を実現できる、(ii) マクロで構成されているため、回路設計者が目的に合わせて任意のサイズ・形態の NN を構成できる、(iii) CMOS コンパチブルな技術である。したがって、モバイルエッジデバイスへの搭載も可能であり、将来のスマート社会における AI 技術の 1 つの担い手となる可能性がある。

●今後の展開

本研究で採用した PIM 型ハードウェアはメモリアレイの中もしくは近傍に MAC ユニットを配置することで、メモリアレイから取り出したデータを、バスを介したデータ転送を用いることなく直接処理することができる。PIM ハードウェアのこの特徴は消費電力の低減に有効なだけでなく、バスによる制約を受けることなく MAC 演算を効率的に並列化できることから、演算能力の向上にも極めて有用である。これらの効果に加えて、本研究で PIM に導入を行った ULVR-SRAM は、従来の SRAM では実現が困難な PG と EMP 動作の両方を実現できる。したがって、ULVR-SRAM を用いて構成した PIM 型 NN アクセラレータはモバイルエッジの AI 技術として要求される動作時・待機時消費電力を低く抑え、演算能力・エネルギー効率の高いアクセラレータ技術となる。今後は、本マクロ技術をたたみ込み層などにも応用していく。簡単な解析からは本技術をたたみ込み層に応用することで 200 TOPS/W 以上の高いエネルギー効率が得られることがわかっている。また、ユーザーフレンドリなマクロベースの設計技術を確立していく。

【用語説明】

(1) **エネルギー最小点（EMP）**：CMOS ロジックシステムで消費される全エネルギーは回路動作に基づく動作時エネルギーと、トランジスタのリークによって生じる待機時エネルギーからなる。これらはそれぞれ異なる電圧依存性を有することから、全エネルギーは特定の電圧で最小値をとる。この全エネルギーが最小となる動作点が EMP である。EMP を動作電圧とすることで、エネルギー効率を最大化できる。

(2) **アクセラレータ**：特定の処理を高速化できるハードウェアまたは演算システム。特に、AI システムでは、各種ニューラルネットワーク（NN）の推論処理に特化した専用のアクセラレータが用いられることも多い。AI 半導体の分野では、NN アクセラレータや AI アクセラレータとも呼ばれる。

(3) **マクロ**：特定の機能を持つ回路ブロック。例えば、SRAM では比較的小さな記憶容量（数 kB 程度）のマクロを作っておき、これを複数個組み合わせることで任意のサイズのメモリを実現できる。

(4) **TOPS、TOPS/W**：TOPS は tera operation per second の略で、1 秒当たりの演算数を意味する。TOPS/W は TOPS を電力で割った指標で、エネルギー当たりの演算数に相当し、エネルギー効率を表す。これらの指標は NN アクセラレータの性能評価の指標として用いられる。

(5) **ULVR-SRAM**：超低電圧でデータ保持（リテンション）できる SRAM のことで、本研究者らによって提案された。動作モードの切り替えが可能な特殊なインバータを用いて構成され、通常の電圧では従来の SRAM と同等の高性能 SRAM 動作を実現し、0.2 V 程度の超低電圧では、記憶データを失うことなくリテンションを行い、待機時電力を大幅に削減できる。この超低電圧リテンションに用いる動作モードを応用することでエネルギー最小点（EMP）での SRAM 動作が可能となる。詳細は以下の論文を参照：H. Yoshida, Y. Shiotsu, D. Kitagata, S. Yamamoto, and S. Sugahara, *IEEE Open J. Circuits Syst.*, **2**, pp. 520-533 (2021).

(6) **モバイルエッジ**：将来のスマート社会におけるコンピューティングシステムは、クラウド、エッジ、モバイルエッジ、ウェアラブルといった階層から構成されると予想されている。モバイルエッジはスマートフォンなどのスマートモバイルデバイスを中心とするコンピューティング階層で、モバイル環境下におけるコンピューティングの中心的な役割を担う。

【論文情報】

掲載誌：*IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*
論文タイトル：Binarized Neural Network Accelerator Macro Using Ultralow-Voltage Retention SRAM for Energy Minimum-Point Operation
著者：Yusaku Shiotsu, and Satoshi Sugahara
DOI：10.1109/JXCDC.2022.3225744

【問い合わせ先】

東京工業大学 科学技術創成研究院 未来産業技術研究所 准教授
菅原 聡
Email: sugahara@isl.titech.ac.jp
TEL: 045-924-5456　FAX: 045-924-5456

東京工業大学 科学技術創成研究院 未来産業技術研究所 博士研究員
塩津 勇作
Email: y.shiotsu@isl.titech.ac.jp
TEL: 045-924-5456　FAX: 045-924-5456

【取材申し込み先】

東京工業大学 総務部 広報課
Email: media@jim.titech.ac.jp
TEL: 03-5734-2975 FAX: 03-5734-3661