News Release
2025.7.7

NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）
国立大学法人東北大学
株式会社アイシン

世界初、CMOS／スピントロニクス融合技術を活用した エッジ AI 向け実証チップの開発に成功しました —従来比 50 倍以上のエネルギー効率改善を実証システムで確認—

NEDOは「省エネAI半導体及びシステムに関する技術開発事業」（以下、本事業）において、エッジ領域に適した高性能かつ省エネルギーな人工知能（AI）半導体デバイスの早期実現を目指して、開発を進めてきました。このたび、国立大学法人東北大学と株式会社アイシンは、磁気抵抗メモリ（MRAM）を大容量搭載したエッジ領域向け「CMOS／スピントロニクス融合AI半導体」を開発しました。OSやアプリの起動用途とメインメモリ用途を兼ねた内蔵メモリとしてCMOS／スピントロニクス融合技術を活用した、エッジAI向けアプリケーションプロセッサ搭載チップの開発は世界初となります。

開発した実証チップは、動作時および待機時電力の大幅低減、起動時間の短縮が可能であり、実証チップを搭載した実証システムの検証において、従来比でエネルギー効率50倍以上、起動時間30分の1以下の改善効果を確認しました。

今後は、車載機器やその他幅広い分野での応用技術開発を進めます。

図1　本事業が目指す多様な社会実装

1. 概要

近年、情報処理に用いるデバイスの高度化、AIなどを用いる、さまざまな産業の創出とその基礎となるビッグデータの活用や、5Gなどの情報通信技術・インフラ整備により、ネットワーク上のデータ量が爆発的に増加しています。このため、これまでのサーバー集約型からネットワークの末端（エッジ）側へ、情報処理とそれに必要な電力を分散していくことが不可欠です。しかしエッジ側では、供給できる電力や、装置の大きさ、利用環境などにさまざまな制約があるため、エッジ用途に適したデバイスの早期実現が重要と考えられます。

このような背景の下、NEDOでは、本事業※1を実施し、この一環として東北大学、アイシン、日本電気株式会社と共同で、CMOS／スピントロニクス融合技術※2によるAI処理半導体の設計効率化と実証およびその応用技術に関する研究開発を進めてきました。CMOSは金属酸化膜を用いた半導体トランジスタ構造の一種で、携帯電話など広く用いられているメインチップの基盤技術となっています。また、スピントロニクス技術から生まれた磁気抵抗メモリ素子であるMRAM※3は、電源を切ってもデータが保持される不揮発の特徴を持っており、CMOS半導体製造プロセスに融合して形成することが可能となっています。

本事業において、東北大学はMRAMを用いた自動設計環境の構築や設計ツールの高度化を行い、大容量MRAMを搭載したAIアクセラレータ※4を開発しました。アイシンはこのAIアクセラレータと、アプリケーションプロセッサ※5、大容量MRAM、および周辺回路を統合して、画像認識などの機能を実現できる実証チップ（図2）を開発しました。

大容量の不揮発性MRAMを実証チップに内蔵したことにより、起動（BOOT※6）時にかかる時間を極めて短くできるほか、メモリへの書き込み／読み出しにかかる時間や電力を大幅に削減できるなど、これまでと一線を画した特徴が得られ、さまざまなAI応用の場面でこれまでにないユーザー体験やメリットを生み出すことが想定されます。

開発した実証チップを搭載した実証システム（図3）にて実測を行い、以下の効果を確認しました。

・電源ONからOS起動を経て最初のAI処理完了までのエネルギー効率が従来比で50倍以上

・電源ONからOS起動するまでの時間が従来比で30分の1以下

なお、今回のリリースは既報ニュースリリース※7の続報となるものです。

2. 今回の成果

(1)エッジAI向け実証チップの開発

現在、小規模から大規模まで多くのAIシステムでは、アプリケーションプロセッサを内蔵したチップを用いたシステムが採用されています。このようなチップでは、BOOT用外付けメモリ（FLASHメモリ※8）、プロセッサのメインメモリ用外付けメモリ（DRAM）、および内蔵メモリ（SRAM）を備える構成が標準的で、ファーメモリ・コンピューティング構造※9となっています。

BOOT用外付けFLASHメモリは、他のメモリに比べバス帯域※10が狭く、ランダムアクセスができないため、起動時に外付けFLASHメモリの内容を順次内蔵SRAMや外付けDRAMにコピーするプロセスが必須であり、小規模なAIエッジシステムを構築する際にも長い起動時間がかかることが課題でした。

今回東北大学とアイシンで開発した実証チップ（図2）では、台湾積体電路製造（TSMC社）のMRAM混載に対応した次世代16nmFinFETプロセス、MRAMマクロ（実証チップ内に32Mbyte搭載）を活用して開発し、アプリケーションプロセッサにArm® Cortex®-A53デュアルコア、東北大学国際集積エレクトロニクス研究開発センターにて開発したAIアクセラレータを搭載しています。内部メモリと重みメモリ※11にMRAMを用いることでニアメモリ・コンピューティング構造※12を実現し、外付けFLASHメモリのバス帯域不足の解消や、アプリ

ケーションプロセッサ上のソフトウエアが起動する際に必要とされていた多くのプロセスを削減しました。また、エッジシステムに適した、高速起動に必要となるコンパクトOSも同時に開発し、OSを実証チップ内部メモリ（MRAM）に内蔵することを実現しました。

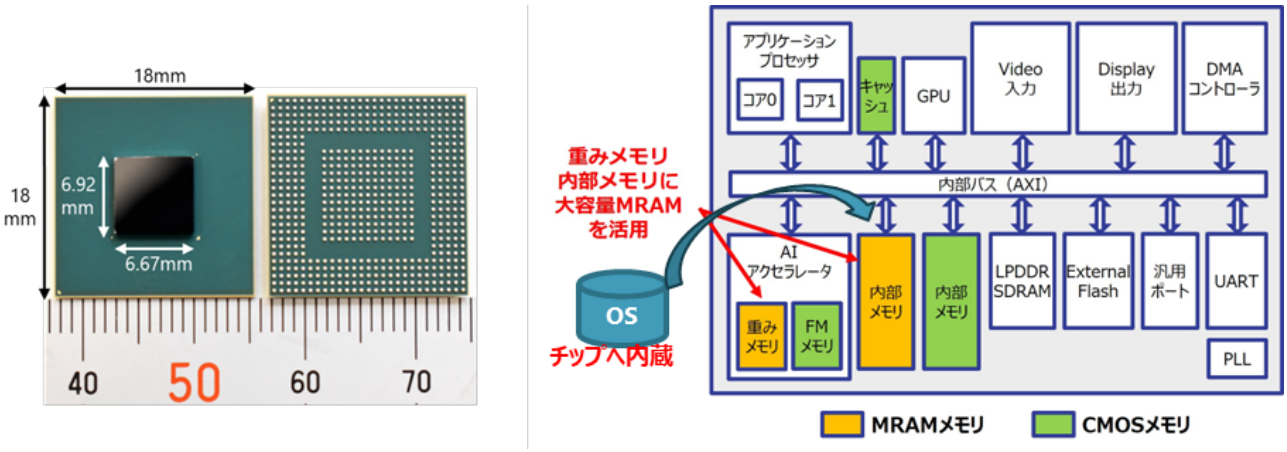

図2　開発した実証チップとブロック図

（2）低消費電力AIアクセラレータの開発

今回設計した実証チップには、東北大学国際集積エレクトロニクス研究開発センターが開発・設計してきた低消費電力AIアクセラレータを搭載しています。本アクセラレータでは、主要メモリにMRAMを適用することで、従来のSRAMと比較して面積効率と電力効率のどちらにも良い特徴があるため、待機電力や動作電力を大幅に削減できます。さらに、不揮発化により重みメモリへのロード時間を削減でき、AI処理システム全体の高速起動が実現可能になります。

（3）エネルギー効率改善およびOS起動時間短縮の実証

今回アイシンは実証チップのアーキテクチャ設計を行い、アプリケーションプロセッサのBOOT用途とメインメモリ用途を兼ねた内蔵メモリとしてMRAMを採用し、大容量チップに内蔵しました。このように大容量MRAMを配置してエッジAI向けアプリケーションプロセッサを搭載したチップの開発は、世界初となります。内蔵メモリにMRAMを採用することにより、起動時間の短縮と外付けメモリの削減が可能になり、チップの小面積化および低消費エネルギー化が図れます。また、チップに内蔵可能なコンパクトOSも同時に開発、OSを実証チップ内部メモリに内蔵することで、従来課題となっていたOS起動時間の大幅な短縮化が図れます。この不揮発性メモリ技術を活用したシステムでは、車載システムで課題になっている暗電流[※13]もゼロにすることが可能となります。

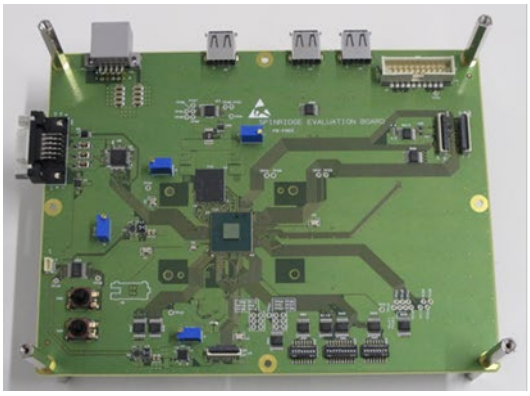

図3　実証環境

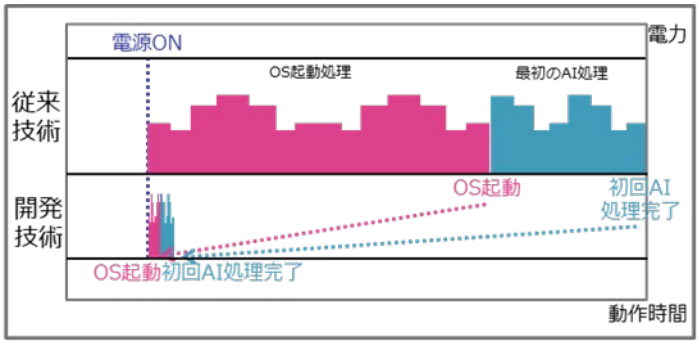

図4　電源ONから最初のAI完了までの処理

今回チップの外付けにBOOTメモリとメインメモリ（DRAM）で構成された従来技術と、エネルギー効率およびOS起動時間短縮の効果を検証するため、実証チップを搭載した開発技術の実証システム（図3）での効果実証を行いました。エネルギー効率実証では、CMOS／スピントロニクス融合技術の活用効果として、電源ONからOS起動を経て最初のAI処理が完了するまで（図4）のエネルギー効率の実証を行い、従来技術4.7624Jに対し開発技術は0.0942Jと50倍以上の改善効果を確認しました。またOS起動時間は従来技術2535.3msに対し開発技術は65.7msと30分の1以下の起動時間短縮効果を確認しました。

3. 今後の予定

NEDOは、今後もエッジ領域での分散コンピューティング実現に向けて取り組み、AI処理でのエネルギー効率改善とその早期社会実装を通して、高度なエッジコンピューティングの実現を目指します。

東北大学は、本研究成果をさらに発展させて、エッジAI領域含め半導体の設計効率向上と、低消費電力AI半導体技術の高度化に資する研究開発を推進します。

アイシンは、本研究成果をいち早く実用化につなげるため、車載機器やその他幅広い分野の応用製品開発に取り組みます。

【注釈】

※1　本事業

事　業　名：省エネAI半導体及びシステムに関する技術開発事業／AIエッジコンピューティングの産業応用加速のための設計技術開発／CMOS／スピントロニクス融合技術によるAI処理半導体の設計効率化と実証、及び、その応用技術に関する研究開発

事業期間：2022年度～2024年度

事業概要：省エネAI半導体及びシステムに関する技術開発事業

https://www.nedo.go.jp/activities/ZZJP_100254.html

※2　CMOS／スピントロニクス融合技術

電子には、電荷だけでなく、スピンと呼ばれる磁気的な性質があります。半導体において電子が持つ電荷の流れを制御してさまざまな機能を引き出す技術をエレクトロニクスと呼びますが、磁気をもたらすスピンの性質も利用するエレクトロニクスの分野を「スピントロニクス」と呼びます。技術開発が進むCMOS技術において極めて重要な問題となっている待機時電力をスピントロニクスの不揮発性を活用して大幅に削減し、加えてスピントロニクスの高い面積効率を生かし、スピントロニクス素子の微細化を通して電力効率10倍以上の改善を実現可能とする技術です。

※3 MRAM

Magnetoresistive Random Access Memoryの略で、磁化の方向で情報を記憶する不揮発性メモリです。1ns程度の高速な磁化反転速度により高速動作が可能であるとともに、原子移動がないために書き換え耐性が高く、他の不揮発性メモリにはない優位性を有しています。

※4 AIアクセラレータ

AIの計算処理を高速化するために設計されたハードウエアのことを指します。従来のCPUやGPUよりも高速にAIの計算を行い、AIアプリケーションにおけるコストを大幅に低減します。

※5 アプリケーションプロセッサ

スマートフォンやタブレット端末などに内蔵されているマイクロプロセッサの一つで、コンピューター機能においてCPUとしてOSやアプリの実行を担当するものです。

※6 BOOT

コンピューターシステムの電源投入時、あるいはシステムのリセット後、OSなどの基本的なシステムソフトウエアをメインメモリ（主記憶）に展開し、ユーザープログラムを実行できるようにするまでの処理の流れを指します。

※7 既報ニュースリリース

「大容量MRAMを搭載したエッジ領域向け「CMOS／スピントロニクス融合AI半導体」により従来比10倍以上の電力効率をシステム動作シミュレーションで確認」

https://www.nedo.go.jp/news/press/AA5_101787.html

※8 FLASHメモリ

データの書き込みと消去ができ、電源を切ってもデータが保持される特徴を持つ半導体メモリのことです。SDカード、USBメモリ、SSDなど、多くの電子機器に部品として搭載されています。

※9 ファーメモリ・コンピューティング構造

演算回路とプログラムおよびデータが格納されているメモリを離れた位置に配置する構造のことで、大容量メモリや複数の異なるメモリ配置が可能となり、また使用するメモリを複数のモジュールで共有することができ、システムの共通化が可能になります。しかしこの構造の要になるメモリと演算器をつなぐバス（配線）が、高性能化と低消費電力化を阻むボトルネックになってしまう欠点を抱えています。

※10 バス帯域

バスとは複数の回路や装置、機器間を結び、データなどのやりとりのために共有される伝送路（配線）のことで、1秒間に扱えるデータ量を幅（本数）とクロック（周波数）の掛け算により帯域として表現します。帯域の数値が大きいほど大量のデータを伝送でき、これを帯域が広いと言います。

※11 重みメモリ

ニューラルネットワークにおいて入力値の重要性、貢献度を数値化して表したものを格納するメモリのことです。ニューラルネットワークでは重みとバイアスを調整することで学習を進めます。バイアスは入力値を一定の範囲に偏らせるために用いるもので、重みはその入力値ごとに決められ、その入力値の価値を決めるもので推論や学習に使用されます。

※12 ニアメモリ・コンピューティング構造

演算回路とプログラムおよびデータが格納されているメモリを極限まで近傍に配置する構造のことで、メモリアクセスがシステム性能上のボトルネックとなることを解消できます。イン・メモリ・コンピューティング（In Memory Computing：IMC）構造とも言います。

※13 暗電流

自動車がエンジンOFFの駐車状態においてバッテリーに流れる待機電流のことで、主には揮発性メモリのデータを保持するため消費されています。バッテリーから常時消費されるため、バッテリー上がりの原因になることがあります。