2026 年 11 月 2 日

国立大学法人東北大学
株式会社アイシン

世界初、STT-MRAM を搭載した 16nm AI SoC を開発
―電源投入から 65.7 ミリ秒で物体検出 AI モジュールを高速起動、従来比 1/50 の低消費エネルギーで物体検出を実現―

【発表のポイント】

- 32MB 組込み STT-MRAM を搭載した 16nm FinFET AI SoC を開発し、DRAM・SSD などの外部メモリを必要としない AI 処理システムを実現しました。
- 電源投入から OS 起動までの時間を 65.7 ミリ秒（従来比約 39 倍高速化）、外部メモリからの重み転送などを不要とすることで物体検出などの AI 処理時間を約 1/10 に短縮し、システム全体で約 25 倍の高速化しました。
- AI 処理のエネルギー消費量を従来比 1/50 となる 50.7 mJ/frame の低消費電力化を実現しました。

【概要】

近年、生成 AI や画像認識などの AI 技術が、自動車、ロボット、カメラ、IoT 機器などのエッジデバイスへ急速に展開されています。一方、エッジ AI では、AI モデルの大容量化に伴うメモリへのアクセスや、外部メモリとのデータ転送が、処理速度やエネルギー効率を低下させる大きな課題となっています。また、電源投入時に OS や AI モデルを外部メモリから読み込む必要があるため、AI システムの起動に時間がかかるという課題もあります。

東北大学国際集積エレクトロニクス研究開発センター（CIES）の遠藤哲郎教授のチームと株式会社アイシンのチームは、これらの課題を解決するため、不揮発性メモリである STT-MRAM を AI SoC に組込んだ「近メモリコンピューティング（Near-Memory Computing：NMC）」型 AI SoC を開発しました。

開発した AI SoC は、AI モデルの重みデータだけでなく、OS のブートデータも不揮発性メモリである組込み STT-MRAM に格納することで、電源投入時に外部 SSD や DRAM から大量のデータを転送する必要をなくしました。さらに、8/16 ビットのハイブリッド乗算アーキテクチャや重み共有技術を採用することで、限られたオンチップ STT-MRAM 容量の中で Tiny YOLOv3 による物体検出

を実現しました。

その結果、電源投入から OS 起動までを 65.7 ミリ秒に短縮し、従来方式と比較して 38.59 倍の高速化を達成しました。また、AI 処理時間は 10.16 倍短縮され、システム全体としては 25.2 倍の時間短縮を実現しました。さらに、Tiny YOLOv3 による 1 フレームの物体検出に必要なエネルギーを 50.7 mJ/frame としました。

本研究成果は、2026 年 11 月 1 日～4 日、台湾新竹市で開催される半導体集積回路技術に関する主要な国際学会である米国 IEEE 主催の「IEEE Asian Solid-State Circuits Conference（A-SSCC）2026」で発表します。

なお、この成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の委託業務の結果得られたものです。

【詳細な説明】

研究の背景

AI を搭載したエッジ機器では、限られた電力・計算資源の中で高速に AI 処理を行うことが求められています。しかし、AI モデルの大容量化に伴い、プロセッサと外部メモリとの間で大量のデータを転送する必要が生じています。

特に、従来の AI SoC では、AI モデルの重みデータを Flash から DRAM へ転送してから AI アクセラレータで処理するため、データ転送に伴う遅延とエネルギー消費が発生します。また、電源投入時には DRAM の初期化や OS の読み込みが必要となるため、AI アプリケーションが実際に動作するまでに長い時間を要します。さらに、AI SoC 内部の揮発性メモリではリーク電力も課題となります。

そこで、本研究では、電源を切っても情報を保持でき、待機時のリーク電力が極めて小さい STT-MRAM に着目しました。STT-MRAM をプロセッサや AI アクセラレータの近くに配置し、データを必要な場所で処理する「近メモリコンピューティング（NMC）」を採用することで、外部メモリとのデータ転送そのものを削減することを目指しました。

今回の取り組み

今回開発した AI SoC では、16nm FinFET プロセスを用い、Arm® Cortex®-A53、デュアルコア CPU、Arm® Mali™-G31 GPU、AI アクセラレータ、オンチップ STT-MRAM などを集積しました（図 1）。オンチップ STT-MRAM を AI モデルの重みメモリおよび OS のブートメモリとして利用しています。

本研究の大きな特徴は、AI モデルだけでなく OS そのものをオンチップ STT-MRAM から起動することです。従来方式では、電源投入後に DRAM を初期化し、SSD から U-Boot や AI モデルのデータを DRAM へ転送する必要がありました。一方、今回の方式では、OS のブートに必要なデータと AI モデルの重みをオンチップ STT-MRAM に保持することで、DRAM の初期化や外部メモリ（DRAM や SSD）からのデータ転送を削減しました。その結果、従来方式では電源投入からカーネル起動まで 2,535.3 ミリ秒を要していたのに対し、今回の方式では 65.7 ミリ秒まで短縮されました。これは 38.59 倍の高速化に相当します。

さらに、AI アクセラレータには、画像認識で広く利用されている Tiny YOLOv3 を実装しました。限られたオンチップメモリ容量を有効活用するため、8 ビットと 16 ビットを組み合わせたハイブリッド乗算アーキテクチャを開発しました。16×16、16×8、8×8 ビットの乗算を組み合わせることで、16 ビットの精度を維持しながら、AI モデルの重みデータ量を 17.7MB から 9.36MB へ 53%削減しました。一方、物体検出精度を示す mAP は 0.325 から 0.316 への低下にとどまりました。また、4×4 のタイル内で同じカーネルの重みを複数のクロックサイクルで再利用する重み共有技術を導入し、データ移動量とローカル帯域幅を削減しました。

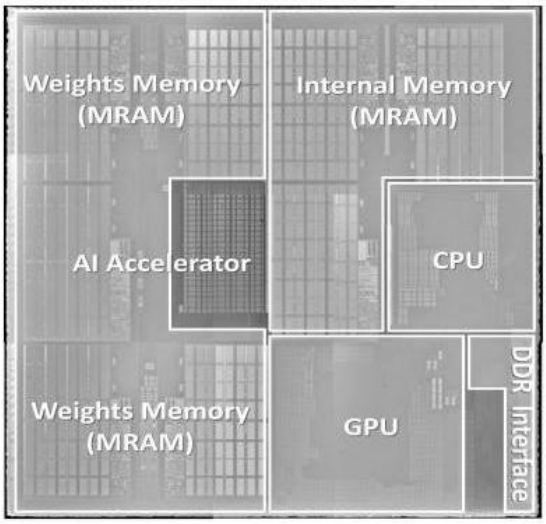
これらの技術により、従来方式では AI 処理前に SSD から DRAM へ 12MB の重みデータを転送する必要があったのに対し、開発した AI SoC では STT-MRAM から必要な重みを直接読み出して処理できるようになりました。その結果、AI 処理時間は 594.6 ミリ秒から 58.5 ミリ秒へ短縮され、10.16 倍の高速化を実現しました。システムレベルの実証では、Tiny YOLOv3 による 1 フレームの推論エネルギーとして 50.7 mJ/frame を達成しました（図 2）。

今後の展開

今回の成果は、「メモリをプロセッサの近くに置く」だけでなく、「AI に必要なデータと OS そのものをチップ内に保持する」ことで、エッジ AI システムの高速起動と低エネルギー化を同時に実現できることを示したものです。

今後は、本技術をさらに発展させ、AI 処理に必要となるメモリ容量、演算性能、エネルギー効率の向上を図るとともに、自動車、ロボット、産業機器、センシング機器など、瞬時起動と低消費電力が求められるエッジ AI システムへの応用を目指します。

特に、不揮発性メモリを活用した AI SoC は、待機時の電力削減だけでなく、電源投入直後から AI 処理を開始できることから、将来の省電力 AI・常時稼働型エッジコンピューティングを支える基盤技術として期待されます。



	Specifications
Technology	TSMC 16 nm
CPU	Dual-core ARM Cortex-A53 L1 Instruction/data Cache (32 KB/32 KB) L2 cache (256 KB)
Video Interface	Video input resolution 4096×4096 Display output resolution 1920×1080
AI Accelerator	Tiny YOLOv3 Feature map memory (SRAM): 2.4 MB Weight memory (MRAM): 20 MB
GPU	ARM Mali-G31
Internal Memory	SRAM: 32 KB, MRAM: 12 MB
DDR Interface	LPDDR5-IF, 16 bit, 3200 Mbps

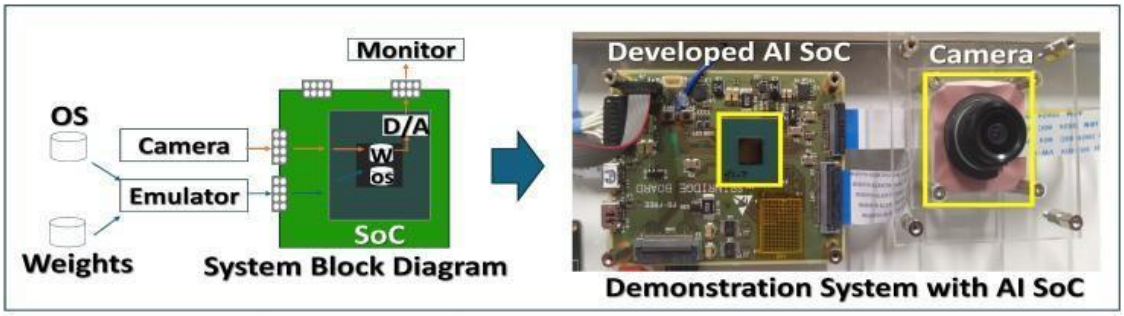


図 1 開発した AI SoC を用いたダイ写真および
キーメトリックサマリーテーブル・デモンストレーションシステム

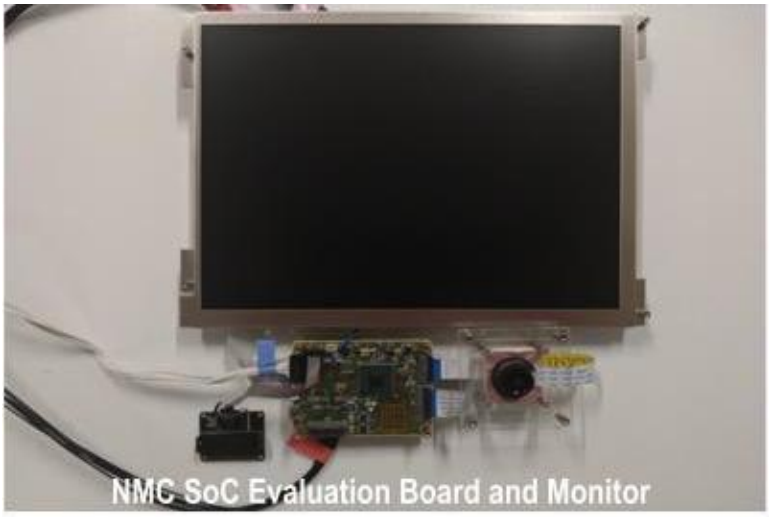


図 2 組み込み STT-MRAM ベースの NMC 評価システム

【謝辞】

この成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の委託業務の結果得られたものです。

【用語説明】

注1.　STT-MRAM（Spin-Transfer-Torque Magnetic Random Access Memory）：磁性体の磁化状態を利用して情報を記憶する不揮発性メモリ。電源を切ってもデータを保持できるため、待機時のリーク電力を抑えられるほか、高速な読み書きが可能。本研究では、AI モデルの重みデータと OS のブートデータを STT-MRAM に格納。

注2.　近メモリコンピューティング（Near-Memory Computing：NMC）：プロセッサや AI アクセラレータの近くにメモリを配置し、データの移動距離を短縮することでデータ転送に伴う消費電力を削減するコンピューティング方式。本研究では、AI SoC 内の STT-MRAM から AI モデルの重みを直接読み出すことで、データ転送に伴う遅延とエネルギー消費を削減。

注3.　Tiny YOLOv3：画像中に存在する物体の種類と位置をリアルタイムに検出する「YOLO（You Only Look Once）」を小型化した AI モデル。本研究では、エッジ機器への搭載を想定した AI 処理の実証に利用。

注4.　AI SoC（System on Chip）：CPU、GPU、AI アクセラレータ、メモリ、各種インターフェースなど、システムを構成する複数の機能を 1 つの半導体チップ上に集積したもの。

【論文情報】

タイトル：A 65.7ms Quick-instant-on and 50.7mJ/frame Tiny YOLOv3-based Object-detection 16nm-AI-SoC with 32MB-embedded-MRAM and Hybrid-multiplication-architecture

著者：Tetsuo Endoh[1, *], Takanobu Naruse[2], Tao Li[1], Shoji Ikeda[1], Mitsuo Yasuhira[1], Hiroshi Yamada[2]

[1]Center for Innovative Integrated Electronic Systems, Tohoku University

[2]AISIN Corporation

*東北大学国際集積エレクトロニクス研究開発センター　センター長・教授　遠藤 哲郎

株式会社アイシン　成瀬 峰信

東北大学国際集積エレクトロニクス研究開発センター　准教授　李 奕

東北大学国際集積エレクトロニクス研究開発センター　副センター長・教授　池田 正二

東北大学国際集積エレクトロニクス研究開発センター　学術研究員　安平 光雄

株式会社アイシン　山田 洋

掲載誌：IEEE Asian Solid-State Circuits Conference (A-SSCC) 2026

URL：https://www.a-sscc2026.org/

【問い合わせ先】
（研究内容およびセンターの活動について）
東北大学国際集積エレクトロニクス研究開発センター
センター長・教授　遠藤 哲郎
TEL: 022-796-3410

（その他の事項について）
東北大学国際集積エレクトロニクス研究開発センター
支援室長　三上 洋一
TEL: 022-796-3410
FAX: 022-796-3432
Email: support-office*cies.tohoku.ac.jp
（*を@に置き換えてください）

株式会社アイシン
広報部　企画・広報 G
TEL: 0566-24-8232