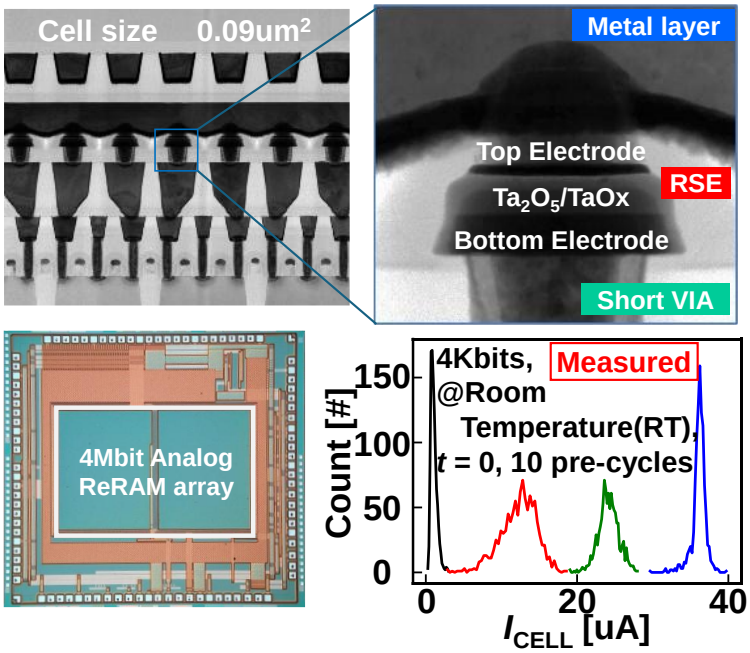
PRESS RELEASE

東京大学大学院工学系研究科

低電力エッジ AI 半導体、ReRAM CiM（Computation-in-Memory）の多値記憶による大容量化と 10 年記憶の両立に成功

発表のポイント

- ◆ 生成 AI などによる AI 推論のパラメータ数と AI 計算に必要なメモリ容量の増加が進む中、データ記憶と演算を一体化することで、AI 推論の低電力化が可能なアクセラレータ、ReRAM（抵抗変化型）で構成される CiM（Computation-in-Memory）を多値記憶によりメモリ容量を増加しつつ、10 年記憶を両立することに成功しました。
- ◆ CiM を構成するアナログ ReRAM を多値記憶により大容量化すると、パラメータ保持中にメモリ信頼性が劣化する問題があります。本研究では、メモリ信頼性の劣化に起因する AI 推論時の積和演算の変化を補正する手法を提案し、多値記憶による大容量化を実現しつつ、10 年間にわたる高い AI 推論精度を達成しました。
- ◆ モビリティ・ロボティクス・ヘルスケア・モバイル応用等では、AI 半導体の低電力化が急務です。本技術によりこれらのエッジ応用において低電力エッジ AI 半導体 CiM の利用が拡大し、GX（グリーントランスフォーメーション）へ貢献することが期待されます。



40nm TaO_xベースの多値アナログ ReRAM テストチップ

概要

東京大学大学院工学系研究科の竹内健 教授、松井千尋 特任准教授らによる研究グループはヌヴォトンテクノロジージャパン株式会社と共同で、モビリティ・ロボティクス・ヘルスケア・モバイルなどエッジ機器における AI 推論で期待される、低電力エッジ AI 半導体である ReRAM CiM の多値記憶による大容量化と 10 年記憶の両立に成功しました。CiM ではメモリと演算器を一体化することで、従来の課題であるメモリのデータ移動に使われる電力を大幅に抑えられま

す。生成 AI などにより AI 推論のパラメータ数と AI 計算に必要なメモリ容量の増加が進む中、CiM に多値記憶を採用することによりメモリ容量を増加することが可能になります。

CiM を構成する抵抗変化型メモリ（ReRAM）を多値記憶により大容量化した際には、パラメータ保持中にメモリ信頼性が劣化する問題があります。本技術では、データ保持時間をモニタする回路を導入し（図 1）、モニタしたデータ保持時間に基づいて、メモリ信頼性の劣化に起因する AI 推論時の積和演算の変化を補正する手法を提案しました。更に、メモリのエラーにロバストな AI 計算の活性化関数を採用することで、多値記憶による大容量化を実現しつつ、10 年間にわたる高い AI 推論精度を達成しました。

本研究成果は、2025 年 9 月 11 日にドイツ ミュンヘンで開催される IEEE European Solid-State Electronics Research Conference（ESSERC）にて口頭発表されました。

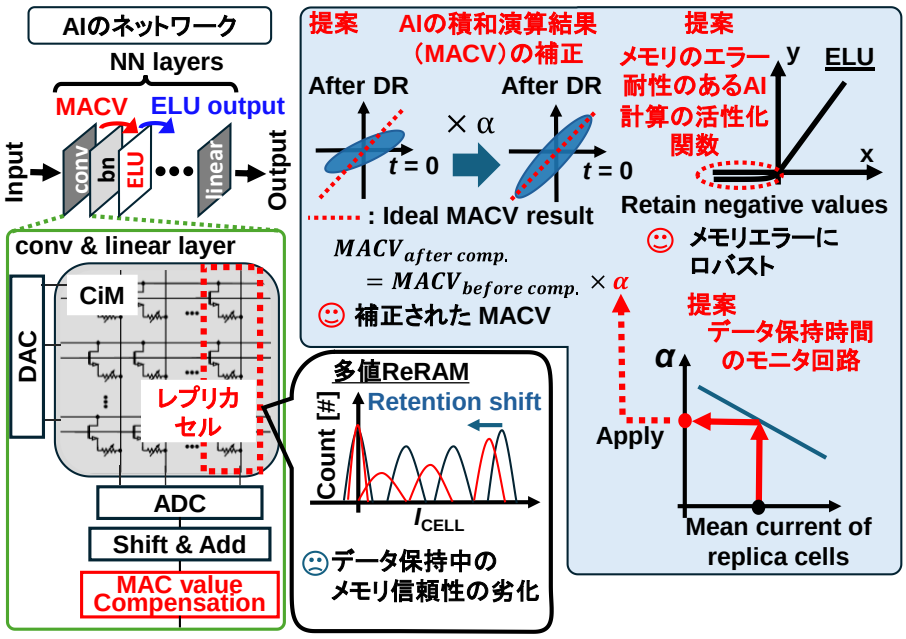


図 1：提案する低電力・高信頼・多値記憶 ReRAM CiM 回路システム

発表内容

GPU など AI の推論処理を実行する半導体に比べ、電力を 1/10 以下に抑えられると期待される CiM が注目されています。従来のコンピュータでは、データ処理を行うプロセッサとデータを記憶するメモリの間のデータの移動が電力・速度のボトルネックとなっており、フォン・ノイマン・ボトルネックと呼ばれています。ReRAM CiM はアナログ ReRAM（抵抗変化型メモリ）で構成されるため、データ処理と記憶を一体化することでフォン・ノイマン・ボトルネックを解決すると期待されています。

一方、生成 AI の発展に伴いニューラルネットワークは大規模化を続けており、そのパラメータ（重み）も増大しています。これに対応するため、膨大なパラメータを記憶する大容量のメモリが求められています。CiM を構成する ReRAM に 2 ビット以上を記憶する多値記憶を採用することで、CiM のメモリ容量を増加することが可能になります。しかし、多値記憶には、ReRAM メモリ中の伝導パスを構成する酸素欠陥の拡散により、10 年といった長期動作中に、メモリのデータにエラーが生じるという問題があります。CiM を構成する ReRAM のメモリエラーはニューラルネットワークのパラメータのエラーに相当し、AI 推論で重要となる積和演算値（MACV）の変動、ひいては AI 推論の精度劣化につながります（図 2）。

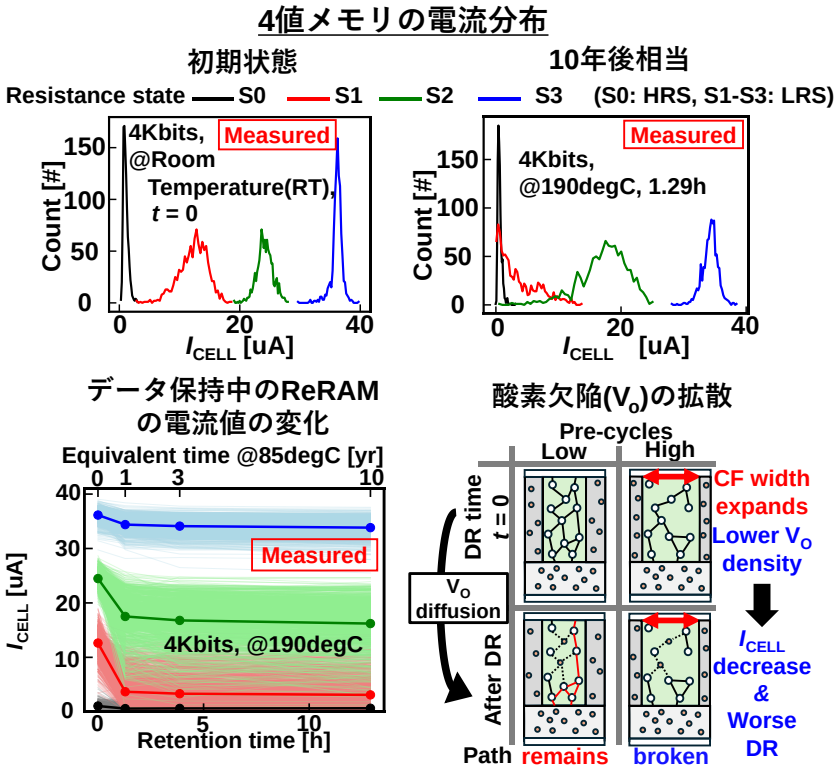


図 2：多値アナログ ReRAM のデータ保持中の信頼性の劣化（高温での加速実験による実験結果）および酸素欠陥（V_O）拡散の物理モデル

本技術ではデータ保持時間をモニタする回路を導入し、モニタしたデータ保持時間に基づいて、メモリ信頼性の劣化に起因する AI 推論時の積和演算の変化を補正する手法を提案しました。更に、メモリのエラーにロバストな AI 計算の活性化関数（ELU）を採用することで積和演算の変化を精緻に補正することに成功しました（図 3）。AI のパラメータのうち上位ビットを 1 ビット記憶の 2 値メモリセル、下位ビットを多値メモリセルに記憶するハイブリッド構造を採用することで、従来の 2 値記憶の CiM に比べて、多値記憶による大容量化を実現しつつ、10 年間にわたる高い AI 推論精度を達成しました（図 4）。

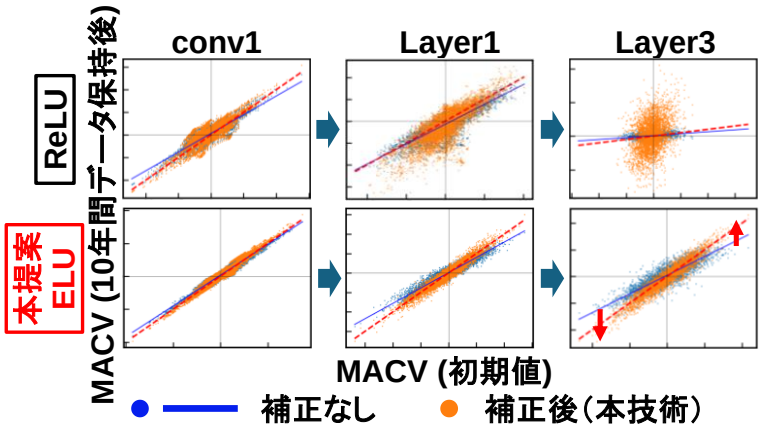


図 3：本技術による AI の積和演算値（MACV）補正の実験結果

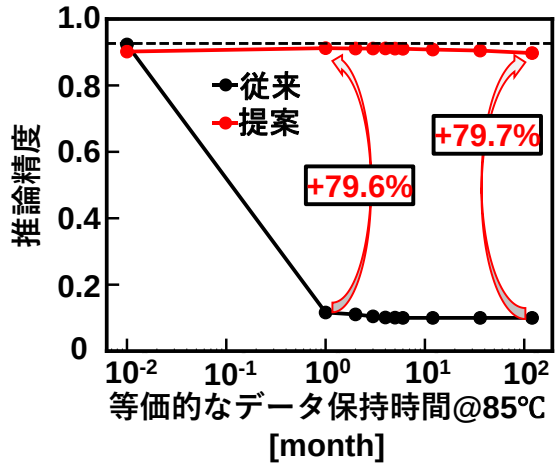


図 4：本技術による AI の推論時間の実験結果。高温で加速実験を行い 10 年間動作を確認。

本技術により CiM の多値化によるメモリ容量の大容量化と 10 年間の高信頼性が両立することが可能になり、モビリティ・ロボティクス・ヘルスケア・モバイル応用等への低電力エッジ AI 半導体 CiM の利用が拡大すると期待されます。そしてエッジ AI の低電力化により、GX（グリーントランスフォーメーション）へ貢献することが期待されます。

〇関連情報：

「世界最高水準の低消費電力化を実現する AI 半導体向け「脳型情報処理回路」を開発」（2018/6/18、NEDO ニュースリリース）

https://www.nedo.go.jp/news/press/AA5_100977.html

発表者・研究者等情報

東京大学大学院工学系研究科
附属システムデザイン研究センター
竹内　健　教授
兼：電気系工学専攻　教授

電気系工学専攻
松井　千尋　特任准教授
三澤　奈央子　学術専門職員
平田　佑亮　研究当時：修士課程
山内　堅心　研究当時：修士課程

学会情報

学会名：European Solid-State Electronics Research Conference (ESSERC)
会　期：2025 年 9 月 8 日～11 日（中央ヨーロッパ夏時間）
題　名：Adaptive Oxygen Vacancy Diffusion Compensation in MLC Intermediate States for over 10-year Data-retention of TaOx ReRAM Analog CiM Array
著者名：Yusuke Hirata*, Kenshin Yamauchi, Naoko Misawa, Chihiro Matsui, Ken Takeuchi
URL：https://www.esserc2025.org

研究助成

本研究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）の助成業務（JPNP23015）の支援により実施されました。

問合せ先

（研究内容については発表者にお問合せください）

東京大学大学院工学系研究科
教授　竹内　健（たけうち　けん）

東京大学大学院工学系研究科 広報室