行動する知性。
中央大学

2019年6月10日

機械学習向け、“Approximate コンピューティング”を採用した、高速低電力 ReRAM ストレージ実現に見通し

～エラーを許容して7倍高速化と90%消費エネルギーを低減～

学校法人　中央大学

概　　要

中央大学 理工学部 教授 竹内 健と機構助教 松井 千尋のグループは、機械学習を用いた応用に向けて、“Approximate コンピューティング”を採用した、高速低電力なReRAM（抵抗変化型メモリ）[注1] ストレージの実現に見通しをつけました。画像や音声の認識、SNS の分類、Web 広告のレコメンデーションなど統計的機械学習[注2]の応用では、多少のエラーが許容されます。それを活かして、メモリのデータマネジメントの簡略化や、読み出しや書き込みの動作条件を変える事で、ReRAM を用いたデータセンタ[注3]のストレージシステムに対して、従来のコンピューティングでは実現不可能な、7倍の高速化、90%の消費エネルギーの低減を実現しました。本技術により、将来のデータセンタのストレージが高速・低電力になるのみならず、次世代メモリ ReRAM の更なる微細化、大容量化が期待できます。

本研究成果は、新エネルギー・産業技術総合開発機構（NEDO）の委託事業である高効率・高速処理を可能とする AI チップ・次世代コンピューティングの技術開発プロジェクト／次世代コンピューティング技術の開発「イン不揮発性メモリ分散 Approximate コンピューティングの研究開発」において実施されたものです。

本研究成果は、2019年6月10日から6月14日に京都で開催された「IEEE Symposia on VLSI Technology」で発表されました。

＊＊

【研究者】　竹内　健　　中央大学理工学部　教授（電気電子情報通信工学科）
松井　千尋　　中央大学研究開発機構　機構助教

【発表（雑誌・学会）】

本研究成果は、2019年6月10日から6月14日に京都で開催された「IEEE Symposia on VLSI Technology」で発表されました。

論文名：Application-Induced Cell Reliability Variability-Aware Approximate Computing in TaOx-based ReRAM Data Center Storage for Machine Learning

【研究内容】

1．背景

従来のコンピュータでは、例えば大陸間弾道ミサイルの軌道を計算することに使われたように、厳密な計算が必要でした。そのような用途では、精度を落としたりエラーを許容したりすることは難しいとされてきました。また、トランジスタの微細化（ムーアの法則）により、いわば自動的に性能・電力・コストが向上できたため、精度を落とす必要もありませんでした。

一方、現在はムーアの法則が終焉しつつあるために、トランジスタの微細化による性能・電力・コストの向上は期待しづらくなっています。また、機械学習を使う画像認識・音声認識・自然言語処理といった分野では、従来のような厳密な計算は必要とされません。人間による認識も完璧ではありません。以上から、今後の機械学習を中心とする応用に対しては、ある程度のエラーを許容する本研究の Approximate コンピューティングは非常に有効であると考えました。

2．研究内容と成果

本研究では、次世代不揮発性メモリ ReRAM（抵抗変化型メモリ）に対して、『（最終的な推論結果の確度は落とさずに）処理やデータの精度を落とす・エラーを許容する Approximate コンピューティング』を適用しました。図1に示すように、不揮発性メモリでは、精度・信頼性とコスト、性能・電力等の間にトレードオフがあり、これらの要素の一方を向上させると、他方が悪化します。本研究の Approximate コンピューティングでは、このトレードオフのうち、精度・信頼性の制約を部分的に緩和する、つまり、精度を落としたり、エラーを許容したりすることで、性能・電力やコストの改善を図りました。

本研究ではまず、アプリケーションにより ReRAM セルの信頼性にばらつきが発生することを明らかにしました。一つのアプリケーション内で頻繁に書き換えられるデータとそうでないデータが混在するため、通常はウェアレベリング[注4]技術を用いて ReRAM セルの書き換え回数を平滑化します。ただし、ウェアレベリングにより余分な書き換えが発生するため、性能・電力の劣化をもたらします。更に余分な書き換えにより ReRAM はエラーが増加してしまいます。

そこで本研究では、ウェアレベリングをしないことで ReRAM セルの書き換え回数にばらつきを発生させ、頻繁に書き換えられるセルと比較して、典型的なセルのエラー率を 1/400 に削減されることを確認しました（図2）。

次に、ReRAM を用いたストレージのシステム・回路・デバイスの協調設計(SCDCD)プラットフォーム（図3）には、ReRAM デバイスをモデル化し、提案するシステム・回路・デバイス技術（表1）を実装しました。入力のアプリケーションワークロード特性に合わせて、提案するコントローラ技術を用い、システムレベルの性能、エネルギー、セルの書き換え回数を出力します。

6つのコントローラ技術のうち、第1のウェアレベリング排除により、ReRAM を用いたストレージの性能が 33%向上することを明らかにしました（図 4）。通常はウェアレベリング技術を用いて ReRAM セ

ルの書き換え回数を平滑化するところ、本研究ではストレージ内に書き換え回数のばらつきを発生させています。

第2に、典型的なエラーを持つセルを訂正ターゲットとするエラー訂正技術（図 5）を提案しました。第1のウェアレベリング排除により、ReRAM ストレージ内には、書き換え回数が多くエラーが発生したセルと、書き換え回数が少なく発生するエラーが少ないセルが混在することになります。そこで、典型的なエラーをターゲットとすることで訂正能力を弱くすることができます。結果として復号に要する時間を削減できるため、ストレージ性能が最大で 85%向上します。さらに、パリティに要するメモリセルが削減できるため、チップ面積が 8.5%低減します。本技術の Approximate コンピューティングでは、全てのエラーを訂正しないため少数のエラーが残りますが、前述のように機械学習を用いた画像認識・音声認識・自然言語処理といった分野では問題になりません。

その他、表 1 に示した 3-6 の技術（3：書き込みと書き込みの間に緩和時間を設けてエラー削減する技術、4：書き込みのストレスと読み出しのストレスが重なることを回避する技術、5：読み出しエラーを許容する高速な読み出し方式、6：書き込みエラーを許容する高速・低電力な書き込み方式）を合わせて用いることで、ReRAM ストレージは最大で性能が 7.0 倍高速となり、消費エネルギーは 90%低減することを確認しました。

3. 今後の展開

今後はメモリのみならず、データ処理（AI アクセラレータ）、分散処理、ネットワーク等にも Approximate コンピューティングを適用し、コンピュータシステムを構成するハード・ソフトの全体を最適化します。これにより、機械学習応用に向けた、次世代コンピュータプラットフォームを世界に先駆けて構築することを目標としています。

●本研究は、以下の委託事業によって実施されました。
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）
「高効率・高速処理を可能とする AI チップ・次世代コンピューティングの技術開発プロジェクト／次世代コンピューティング技術の開発」
研究課題名：「イン不揮発性メモリ分散 Approximate コンピューティングの研究開発」

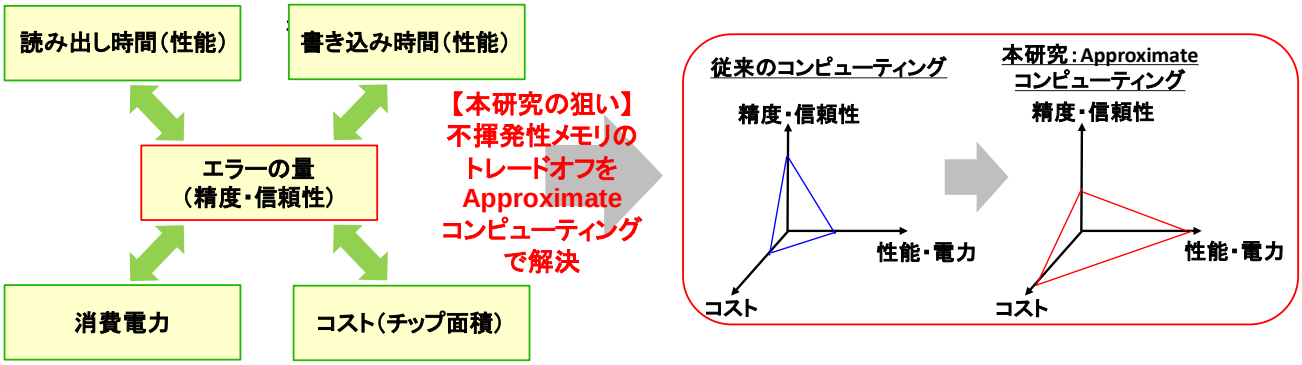

図1．不揮発性メモリのトレードオフ（左図）と、本研究の Approximate コンピューティング（右図）

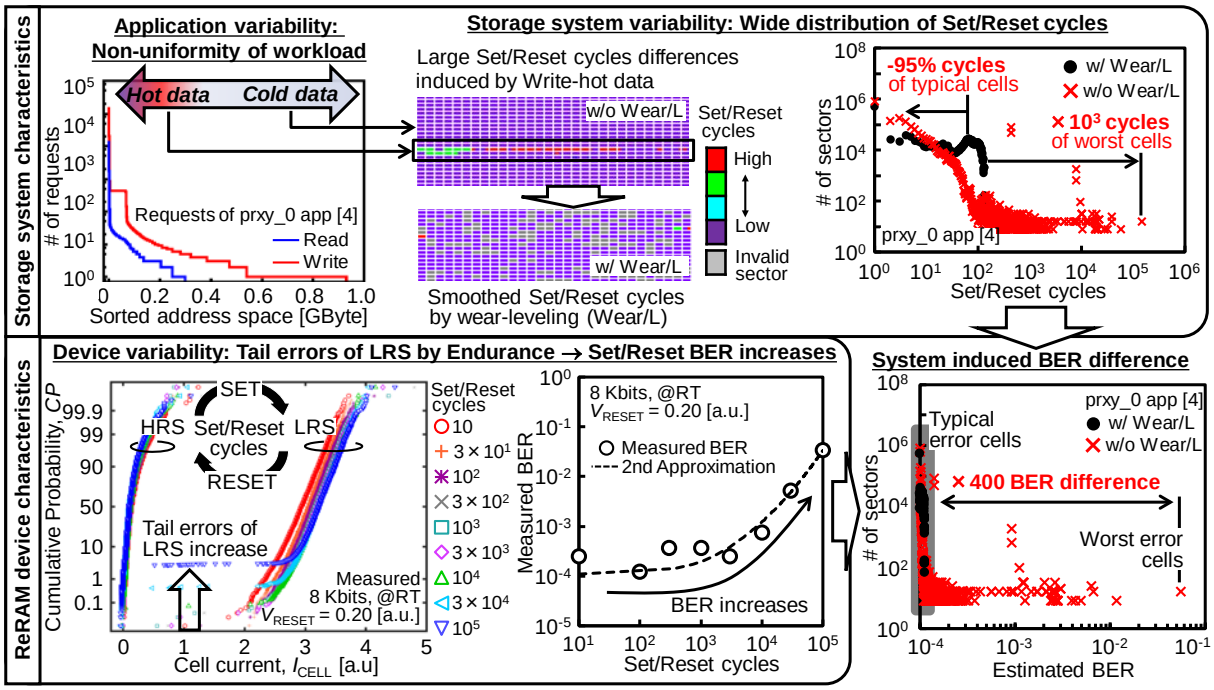

図2．ストレージシステムおよびデバイスが引き起こす ReRAM セルエラーのばらつき。アプリケーションそのものが不均一なアクセスを有するため、従来、ウェアレベリング技術により ReRAM セルの書き換え回数の平滑化を行う。本研究ではウェアレベリングを行わないことにより ReRAM セルの書き換え回数にばらつきを発生させ、最悪なエラーを有するセルと比較して典型的なセルのエラーが 1/400に低減することを確認した。

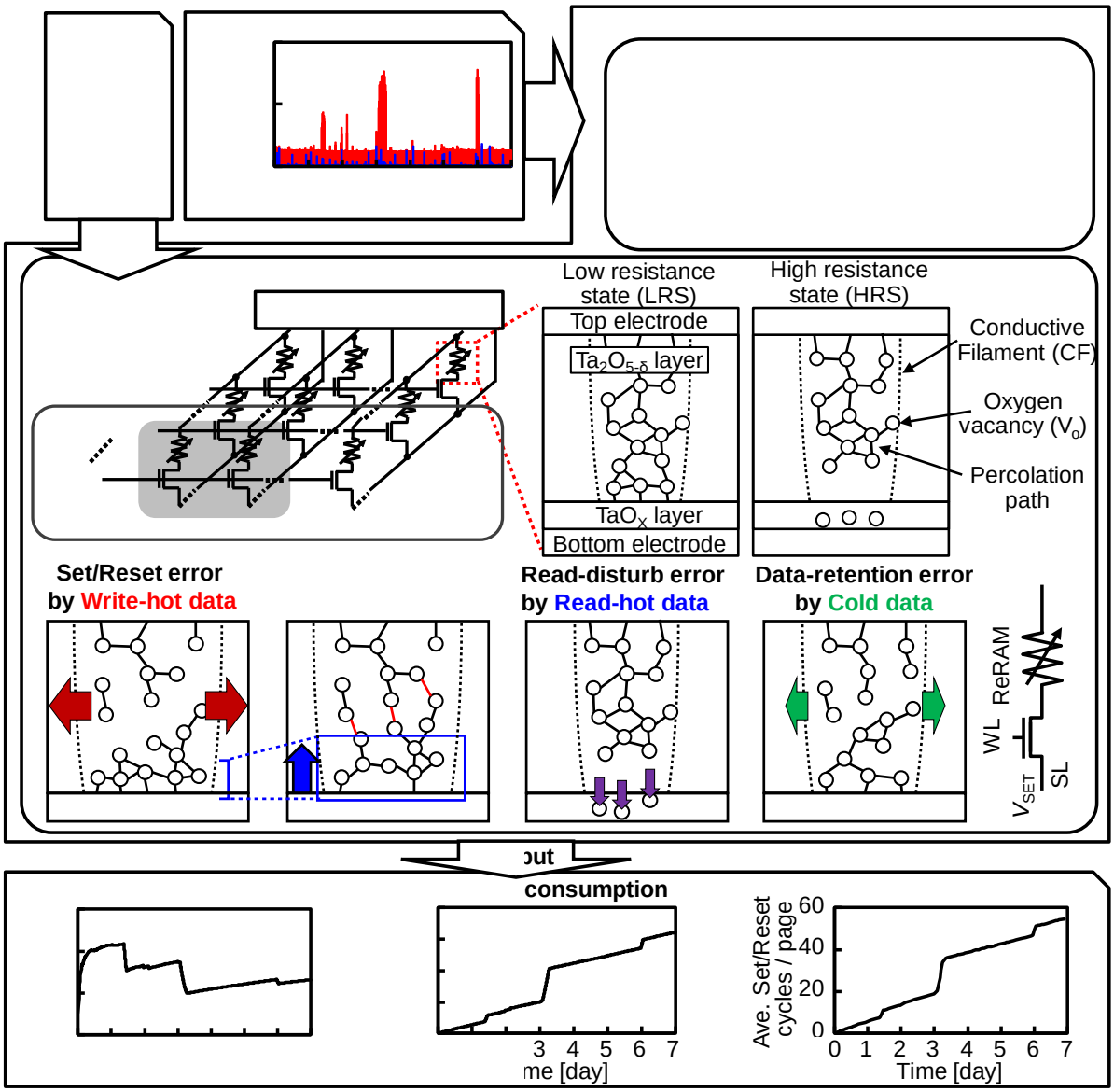

図3. システム・回路・デバイスの協調設計（SCDCD）プラットフォーム。

Hierarchy	Operation	Conv. computing	Proposed V-AC Strategy	Technique	Data characteristic *Write-hot*	*Read-hot*	*Cold*
System	Wear-leveling (Wear/L)	w/ Wear/L	I	w/o Wear/L	✓	✓	
	ECC	Worst-error target (35-bit correction)	II	Typical-error target (5-bit correction)	✓	✓	
	Data management	NA	III	Interval-assured Write	✓		
		NA	IV	Dynamic Dual-stress Scrubbing	✓	✓	
Circuit	Read	NA	V	Adaptive Read		✓	✓
Device	Set/Reset	Verify	VI	w/o Verify	✓		
		NA		Lower V_{SET}/V_{RESET}	✓		

表1. 提案する Approximate コンピューティング向け ReRAM ストレージコントローラ技術。異なる特性を持つデータに向けて、ストレージの異なる階層（システム・回路・デバイス）において 6 技術を提

案した。

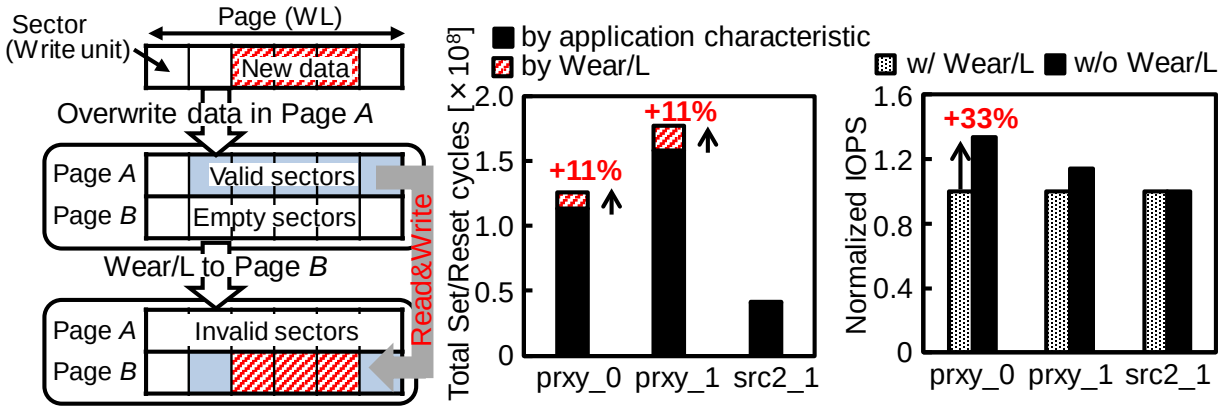

図4. 第1の技術、ウェアレベリング排除。ReRAM ストレージの性能が最大 33%高速化した。

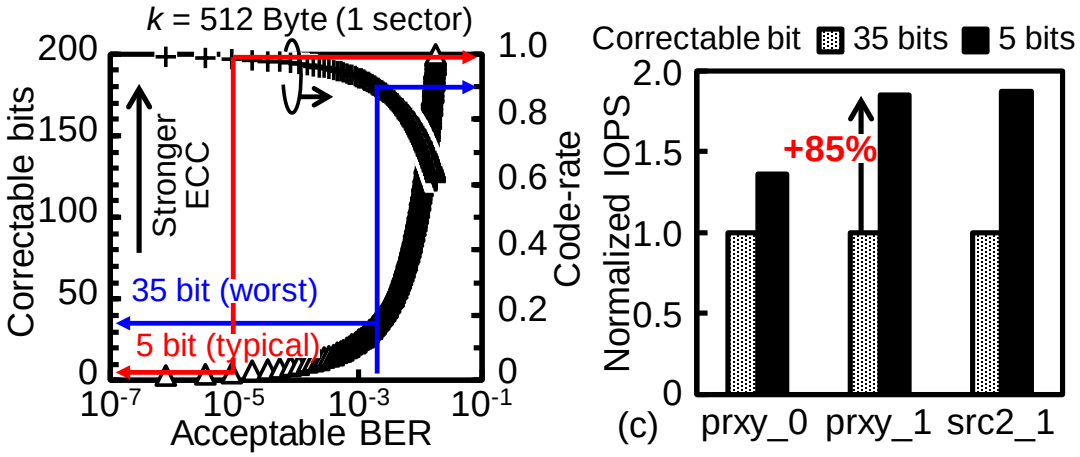

図5. 第2の技術、典型的なエラーを持つセルをターゲットとしてエラー訂正。ストレージ性能が最大で 85%高速化した。

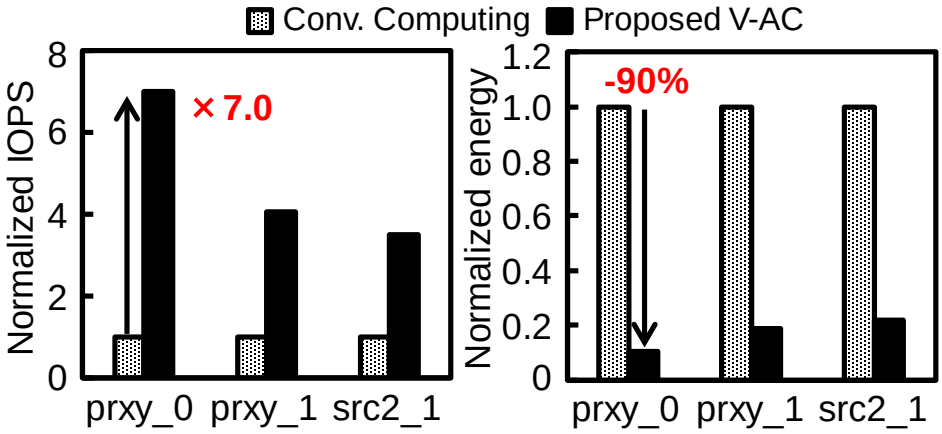

図6. 提案する 6 技術を用いた ReRAM ストレージの性能およびエネルギー。

【お問い合わせ先】

＜研究に関すること＞
竹内 健　（タケウチ　ケン）
中央大学理工学部　教授（電気電子情報通信工学科）
TEL： 03-3817-7374
E-mail: takeuchi@takeuchi-lab.org

＜広報に関すること＞
加藤　裕幹　　（カトウ　ユウキ）
中央大学　研究支援室
TEL　03-3817-1602，FAX　03-3817-1677
E-mail: k-shien@tamajs.chuo-u.ac.jp

【用 語 解 説】

注1）ReRAM（抵抗変化型メモリ）
電流を流すことで抵抗値が変化する材料を記憶素子として用いた次世代の半導体メモリ。ReRAM は電源を落としてもデータを保持できる不揮発性の新しいメモリで、データの読み書きが高速で消費電力も少ないのが特長。

注2）統計的機械学習
データから統計的方法にもとづいてコンピュータを用いて解析し、有用なルール、規則性、知識などを抽出し判断を行う技術。

注3）データセンタ
SNS やインターネットを使ったサービスを行うためにサーバーやストレージ、ネットワーク機器などの IT 機器を設置・運用する施設。

注4）ウェアレベリング
ReRAM などの不揮発性メモリのセルは書き換え回数が増えるとエラーが増大するため、セル間の書き換え回数を均一化し、発生するエラーを低減する技術。