公開

研 究 終 了 報 告 書

「積層型 AI チップの低電力高効率アーキテクチャ」

研究期間： 2021 年 10 月～2023 年 3 月
研 究 者： 柴 康太

1. 研究のねらい

サイバー空間とフィジカル空間を高度に融合させたシステムにより実現される Society 5.0 では、多種多様な IoT デバイスから得られる大量データを高効率・リアルタイムにエッジで知的情報処理することが求められる。エッジで高効率・リアルタイムに情報処理をするためには、低電力・低レイテンシな AI 推論処理チップが必要不可欠である。その実現に向けたボトルネックは外部メモリアクセスである。加算演算と比較して DRAM アクセスには 3～4 桁高いエネルギーを必要とする。この課題を解決するために SRAM チップを演算チップの直上に三次元積層することで DRAM と比較しエネルギーが 1 桁低減する。三次元積層によりデータの移動に伴うエネルギーが削減され、DRAM を SRAM に置き換えることでメモリアクセスのエネルギーが削減されるためである。

本研究ではこのような低エネルギーなメモリアクセスを提供する三次元積層 SRAM の低レイテンシ性やランダムアクセス性といった特徴に着目し、AI のプルーニングを活用することでさらなる低電力化を図った。プルーニングはニューラルネットワークの中で重要度の低いパラメータを取り除き重要度の高いパラメータのみを残すことで、推論精度をほとんど下げることなくパラメータ数を削減する技術である。プルーニングを活用しパラメータ数を削減することでメモリアクセスの回数が低減しメモリアクセスに伴う消費電力が削減される。しかし、プルーニングされたニューラルネットワークでは処理する疎行列の構造が不規則になることでランダムアクセスが頻発し DRAM アクセスにおけるレイテンシが増大する。そこでランダムアクセス可能な三次元積層 SRAM を活用した推論処理ハードウェアを開発し、プルーニングされたニューラルネットワークも効率良く処理することを目指した。

このように本研究は低電力・低レイテンシな AI 推論処理チップの実現に向けて、低エネルギーな三次元積層 SRAM の低レイテンシ性やランダムアクセス性を活用し、AI のプルーニングと組み合わせることで従来にない低電力と低レイテンシを実現することをねらいとしたものである。

2. 研究成果

（1）概要

三次元積層 SRAM とプルーニングを活用した低電力・低レイテンシな AI 推論処理チップの実現に向けた以下の 3 つの研究をおこなった。

まずは疎行列のデータ圧縮技術の確立(A)である。プルーニングはニューラルネットワークの不要なパラメータを取り除き必要なパラメータのみを残すことでパラメータ数を削減することができる。しかし、ランダムプルーニングされたニューラルネットワークの行列構造は不規則な疎行列である。疎行列を記憶するためには非ゼロ値に加えて非ゼロ値の場所を示すインデックス情報が必要であるが、不規則な構造のためインデックス情報を記憶するためのオーバー

ヘッドが大きくデータ容量とデータ移動量の削減効果が制限される。N:M プルーニングとして知られる M 個の要素を持つ区分行列の中に N 個の非ゼロパラメータを持つようにしたプルーニングアルゴリズムは行列構造の規則性を維持することができ、低データ容量化につながる。しかし、プルーニングアルゴリズムに制限がかかることで推論精度の劣化が生じる。そこで、本研究ではランダムプルーニングされた不規則な疎行列を N:M 構造の疎行列に並び替えることで高い推論精度と低データ容量を両立する手法を研究した。

次に圧縮疎行列の高効率処理チップの開発(B)である。A の技術で圧縮された重み疎行列を効率良く処理できれば、低電力な推論処理ハードウェアが実現される。その実現に向けて圧縮データ処理に必要なメモリアクセスを効率良くおこなうことが鍵であるが、疎行列が並び替えられているため疎行列との演算に対応するデータを持ってくるときのメモリアクセスがランダム化する。そこで、ランダムアクセス可能な三次元積層 SRAM を用いた圧縮データ処理アーキテクチャを考案した。DRAM の代わりに三次元積層 SRAM を活用することで低電力なハードウェアが実現されることを示す。また、圧縮データ処理に対応したテストチップを 0.18μm CMOS プロセスで試作し、評価をおこなった。

最後に疎行列処理コンピュテーションインメモリ(CIM) SRAM の開発(C)である。三次元積層 SRAM に搭載されている SRAM チップで行列演算が可能となれば、演算チップ−三次元積層 SRAM 間のデータ移動量をさらに削減することができ低電力な推論処理を実現できる。SRAM チップで行列演算に有効な手段が CIM であり、メモリ内部で演算をおこなうことで高いエネルギー効率やスループットを達成できる。しかし、従来提案された CIM SRAM は疎行列の演算処理に対応していないため、データ容量やデータ移動量の増大に繋がる。そこで、本研究では N:M 構造の規則性に着目し疎行列処理に対応した CIM SRAM を設計し、高いエネルギー効率と高いスループットを達成することを示す。

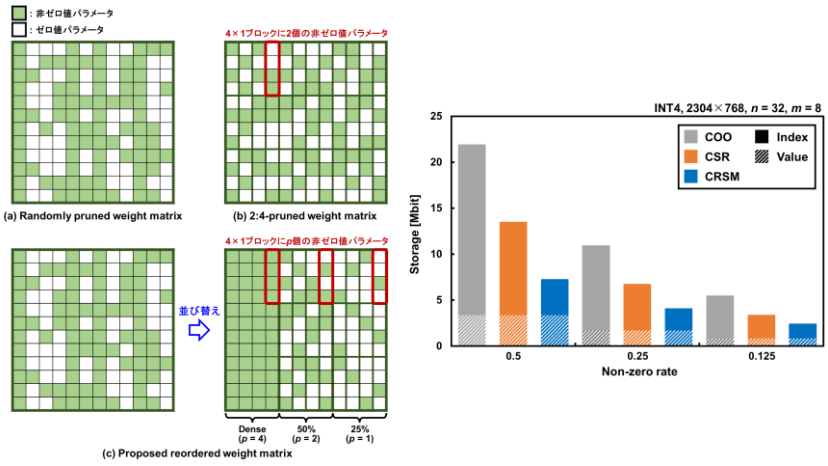
（2）詳細

研究テーマ A「疎行列のデータ圧縮技術の確立」

プルーニングはニューラルネットワークの不要なパラメータを取り除き必要なパラメータのみを残すことでパラメータ数を削減することができる。しかし、ランダムプルーニングされたニューラルネットワークの行列構造は不規則な疎行列である。疎行列を記憶するためには非ゼロ値に加えて非ゼロ値の場所を示すインデックス情報が必要であるが、不規則な構造のためインデックス情報を記憶するためのオーバーヘッドが大きくデータ容量とデータ移動量の削減効果が制限される。N:M プルーニングとして知られる M 個の要素を持つ区分行列の中に N 個の非ゼロパラメータを持つようにしたプルーニングアルゴリズムは行列構造の規則性を維持することができ、低データ容量化につながる。しかし、プルーニングアルゴリズムに制限がかかることで推論精度の劣化が生じる。そこで、本研究ではランダムプルーニングされた不規則な疎行列を N:M 構造の疎行列に並び替えることで高い推論精度と低データ容量を両立する手法を研究した。

ランダムプルーニングされたニューラルネットワークの学習後の重み疎行列に対して並び替えをおこなうため、プルーニングアルゴリズムに制約がかからず自由にプルーニングができるため汎用性が高いことがメリットである。さらに、並び替えはエッジでの推論処理の前にクラ

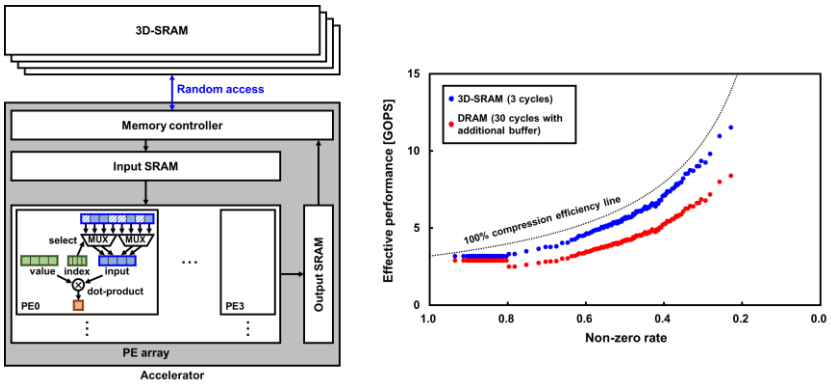
ウドでおこなわれるため、並び替えの処理にエッジのハードウェアリソースを必要としないメリットもある。並び替えの方法は各列をスパース率の順に並び替え、その後に M 個の区分行列の中に N 個の非ゼロパラメータを持つように行の並び替えをおこなう。この並び替えによって疎行列構造に規則性が生まれ、非ゼロ値のほかに必要なインデックス情報のデータ容量が削減される。この並び替えられた疎行列を記憶するための新たなフォーマットとして圧縮並び替え行列格納(CRSM: Compressed Reordered Sparse Matrix)形式を提案した。従来用いられた圧縮行格納(CSR: Compressed Sparse Row)形式や座標(COO: Coordinate)形式と比較し、データ容量を 63%削減できることを示した。さらに、プルーニングレートに対してデータ容量もスケール可能であることを示した。



研究テーマ B「圧縮疎行列の高効率処理チップの開発」

A の技術で圧縮された重み疎行列を効率良く処理できれば、低電力な推論処理ハードウェアが実現される。その実現に向けて圧縮データへのメモリアクセスを効率良くおこなうことが鍵である。課題はメモリアクセスがランダム化することである。従来の COO 形式や CSR 形式では比較的シーケンシャルなインデックス情報を持つのに対し、CSRM 形式では疎行列を並び替えたことでランダムなインデックス情報を持つ。従って、疎行列との演算に対応するデータを持ってくるときのメモリアクセスがランダム化するため、従来の DRAM ではレイテンシとオンチップの必要メモリ容量が増大する。そこで、ランダムアクセス可能な三次元積層 SRAM を用いた圧縮データ処理アーキテクチャを考案した。三次元積層 SRAM を活用することでデータの読み出しが容易におこなわれ、一度データが読み出されると高い規則性を活かしてオンチップでは単純な積和演算がおこなわれ高いスループットと高いエネルギー効率を達成できる。

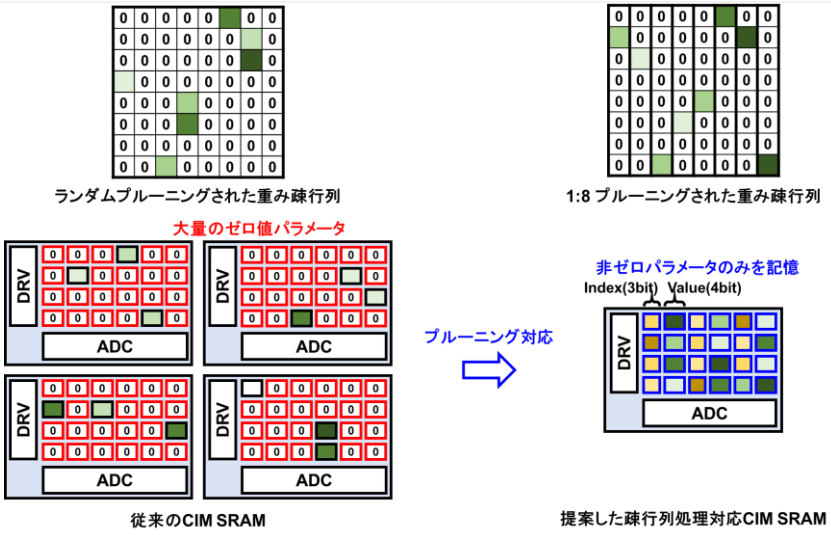
圧縮データ処理に対応したテストチップを 0.18 μ m CMOS プロセスで試作した。テストチップ上に三次元積層 SRAM をエミュレートし、実際の三次元積層モジュールとして正しく機能することを実機で確認した。DRAM の代わりに三次元積層 SRAM を活用することでハードウェアのスループットが 40%改善されることを示した。測定結果やシミュレーション結果に基づき、従来の三次元積層 SRAM を用いた推論処理ハードウェアと比較してエネルギー効率が 2 倍改善されることを示した。



研究テーマ C「疎行列処理対応 CIM SRAM の設計」

三次元積層 SRAM に搭載されている SRAM チップで行列演算が可能となれば、演算チップ−三次元積層 SRAM 間のデータ移動量をさらに削減できる。例えば、演算チップがメモリ a の行列とメモリ b の行列の乗算結果をメモリ c に書き込むよう指示して SRAM チップで演算が完了すれば、演算チップ−三次元積層 SRAM 間のデータ移動は命令に必要なデータだけとなる。SRAM チップでの行列演算に有効な手段が CIM であり、メモリの内部でデータの記憶だけでなくデータの演算もおこなうことで高いエネルギー効率やスループットを達成できる。しかし、従来提案された CIM SRAM は疎行列の演算処理に対応していないため、データ容量やデータ移動量の増大に繋がる。

そこで、本研究では疎行列の演算処理に対応した CIM SRAM を提案した。疎行列の演算処理に N:M 構造の規則性を活用することで、疎行列対応のオーバーヘッドを最小限に留めた。従来はゼロ値パラメータを含めすべてのパラメータを SRAM に記憶していたのに対し、本研究ではインデックス情報とともに非ゼロ値パラメータのみを記憶することで必要なメモリ容量を 78%削減した。



3.　今後の展開

本研究の中核を担ったのが三次元積層 SRAM である。近年では AI アクセラレータに加えて、市販の CPU にも搭載され計算性能の向上に寄与している。一方で、過去半世紀に渡って続いてきたムーアの法則に従った SRAM のスケールは止まりつつあり最先端プロセスでは 1 世代でビットセル面積が 5%しか削減されていない。このような状況を鑑みるとこれまで以上に三次元積層 SRAM を用いた大容量化へのニーズが高まることが予想されるだろう。本研究ではその三次元積層 SRAM の活用方法をいち早く研究したものであり、今後も三次元積層 SRAM の研究が続くと考えられる。

本研究では提案した並び替え疎行列がデータ容量を削減し、三次元積層 SRAM を活用することで高効率にその行列を処理できることが示された。一方で、並び替えには大量の計算が必要であるために並び替えに対応可能なプルーニングレートは 75%に留まった。より少ない計算量で並び替えが完了するアルゴリズムが開発されることでさらに高いプルーニングレートにも対応しさらなるデータ容量削減を達成することができると考えられる。

本研究で提案した CIM SRAM は世界で初めて細粒度の疎行列演算に対応したものである。これまで CIM SRAM では対応できなかったプルーニングされたニューラルネットワークの処理も可能となることでプロセススケーリングに頼らずプルーニングレートでスケール可能なスループットやエネルギー効率の向上が達成された。このように今後は SRAM スケーリングに頼らない SRAM の性能向上が重要な鍵となるだろう。例えば、通常の SRAM ではシーケンシャルアクセスに特化した SRAM や広帯域アクセスに特化した SRAM が登場しており、汎用 SRAM と比較して低電力なメモリアクセスを実現している。CIM SRAM ではあるモデルの処理に特化した SRAM や本研究の疎行列処理に特化した SRAM が高い性能を発揮している。このようにこれまでの汎用 SRAM から領域特化型 SRAM に転換することで設計コストは高くなるがスケーリングに頼らない性能改善を獲得することができるため、アプリケーションを意識した SRAM の研究が発展すると考えられる。

4.　自己評価

当初はテーマ A と B をさらに発展させる方向で計画を進めていたが、テーマ A での成果を活用することで CIM SRAM の課題を解決できそうであることに気が付きテーマ C の研究を立ち上げた。その結果として、研究開始当初は予想していなかった良い方向に研究を進めることができた。当初予定から 1 年早く研究終了となってしまったが、本年度までは計画通りの成果が出たことと今後の研究の基盤となる結果も出たことで十分に満足している。

本研究では今後重要となる三次元積層 SRAM というメモリデバイスの活用方法について研究した。プロセススケールによる大容量化が難しくなる中で SRAM の特徴を活用したデータの低容量化を提案し、疎行列処理対応 CIM SRAM を世界に先駆けて提案した。このようにプロセススケーリングに依存しない SRAM に記憶させるデータの低容量化技術と CIM SRAM の疎行列処理対応技術が今後の SRAM 技術の発展に寄与していくと考えている。

ACT-X での研究を通じて、同じ領域内の様々な研究者が苦労の末に成功されているのを目の当たりにして私にとって大きなモチベーションとなった。また、領域内でハードウェアを研究する研究者と出会い、半導体チップの応用について議論する機会を得ることができた。この議論

を通じて想像もしていなかった応用方法やニーズを知ることができ、私自身にとって貴重な経験となった。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：3件（投稿中: 1件）

1. K. Shiba, M. Okada, A. Kosuge, M. Hamada, and T. Kuroda, “A 7-nm FinFET 1.2-TB/s/mm^2 3D-Stacked SRAM Module with 0.7-pJ/b Inductive Coupling Interface Using Over-SRAM Coil and Manchester-encoded Synchronous Transceiver,” *IEEE Journal of Solid-State Circuits (JSSC)*, 2023, in press.
7nm FinFET プロセスを用いて三次元積層 SRAM モジュールを試作した。積層チップ間の通信は無線でおこなわれ、SRAM ブロック直上に配置したコイルと同期式送受信回路を用いることで低電力かつ広帯域な三次元積層 SRAM を実現した。将来必要となるメモリ容量やメモリ帯域を議論し本研究で提案した三次元積層 SRAM モジュールが要求を満たしていることを示した。
2. K. Shiba, A. Kosuge, M. Hamada, and T. Kuroda, “Crosstalk Analysis and Countermeasures of High-Bandwidth 3D-Stacked Memory Using Multi-Hop Inductive Coupling Interface,” *IEICE Transactions on Electronics*, 2023, vol. E106-C, no. 7, in press.
広帯域な三次元積層メモリデバイスのクロストークの解析と対策をおこなった。クロストークの解析をおこない主要なクロストーク源を 2 個に特定するとともに、それぞれの対策として短絡コイルと 8 形コイルを提案した。クロストークの削減効果をシミュレーションで確認し、面積当たりのメモリ帯域が 4 倍に改善されることを示した。
3. K. Shiba, M. Okada, A. Kosuge, M. Hamada, and T. Kuroda, “A 12.8-Gb/s 0.5-pJ/b Encoding-Less Inductive Coupling Interface Achieving 111-GB/s/W 3D-Stacked SRAM in 7-nm FinFET,” *IEEE Solid-State Circuits Letters (SSC-L)*, 2023, vol. 6, pp. 65-68.
7nm FinFET プロセスを用いて三次元積層 SRAM モジュールを試作した。積層チップ間の通信は無線でおこなわれ、SRAM ブロック直上に配置したコイルと符号化不要な同期式送受信回路を用いることで低電力かつ広帯域な三次元積層 SRAM を実現した。先行研究と比較して競争力のあるエネルギー効率で高い面積効率が達成されることを示した。

（2）特許出願

研究期間全出願件数: 0件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

主要な学会発表

1. 柴康太, 小菅敦丈, 濱田基嗣, 黒田忠広, “[招待講演] 三次元積層 SRAM と近接場無線接続技術,” *電子情報通信学会(IEICE) 集積回路研究会(ICD) メモリ技術と集積回路技術一般*, Apr. 2022.

2. K. Shiba, M. Okada, A. Kosuge, M. Hamada, and T. Kuroda, “Polyomino: A 3D-SRAM-Centric Architecture for Randomly Pruned Matrix Multiplication with Simple Rearrangement Algorithm and x0.37 Compression Format,” *IEEE Interregional NEWCAS Conference (NEWCAS)*, June 2022.
3. K. Shiba, M. Okada, A. Kosuge, M. Hamada, and T. Kuroda, “A 7-nm FinFET 1.2-TB/s/mm^2 3D-Stacked SRAM with an Inductive Coupling Interface Using Over-SRAM Coils and Manchester-Encoded Synchronous Transceivers,” *IEEE Hot Chips 34 Symposium (HCS)*, Aug. 2022.

解説記事

4. 柴康太, 小菅敦丈, 濱田基嗣, 黒田忠広, “近接場無線接続技術を用いた三次元積層SRAM,” *エレクトロニクス実装学会誌*, vol. 25, no. 6, pp. 549-555, Sep. 2022. (解説記事)

受賞

5. IEEE Solid-State Circuits Society (SSCS) Predoctoral Achievement Award, 2023.