3A-05

TCI を用いた 3 次元積層型 DNN 向け アクセラレータ SNACC の設計と評価

坂本龍一† 高田遼† 石井潤† 近藤正章† 中村宏† 大久保徹以‡ 小島拓也‡ 天野英晴‡
東京大学† 慶應義塾大学‡

1 はじめに

近年，組込みシステム向けに高電力効率なディープニューラルネットワーク(DNN)向けアクセラレータの開発が重要になっている．畳込み層の演算に着目した Eyeriss[1]や，全結合層の省電力化に着目した EIE[2]などがある．また，DaDianNao[3]はオンチップに eDRAM を用いることで Off-Chip メモリへのアクセスを抑制し，高い電力効率を達成している．

しかしながら，これらの従来研究では消費電力削減のために畳込み層などの特定のネットワーク構造向けに最適化したアクセラレータや，データアクセス削減のためにネットワーク構造に手を加える研究が多い．対象とするネットワーク構造が限られる可能性もあり，進化を続ける DNN の多様なネットワーク構造を扱うには柔軟性が課題となる．それに対して，我々は，組込みシステム向けに高電力効率で多様なネットワーク構造に対応できる柔軟性を持ち，かつ TCI (ThruChip Interface)を用いた 3 次元積層が可能な DNN アクセラレータ SNACC を開発している．本稿では，3 次元積層型アクセラレータのスケーラビリティ評価として，4 コア構成の LSI チップ実装をもとに，積層する LSI の枚数とデータ転送バンド幅を変えた場合のエネルギー効率をシミュレーションにより評価する．

2 アクセラレータのアーキテクチャ

本研究では，マイクロコントローラと SIMD 型積和演算器を主な構成要素とするコアを複数搭載したマルチコアアクセラレータを開発している．4 コア構成のアクセラレータを図 1 に示す．各コアは，命令メモリ(inst)，ストリームバッファ(sbuf)，データメモリ(dmem)，ルックアップテーブル(lut)，データ出力用メモリ(omem)の 5 つのメモリを持つ．演算結果をコア間で共有する必要があるため，出力用メモリの omem はコア間で共有する．

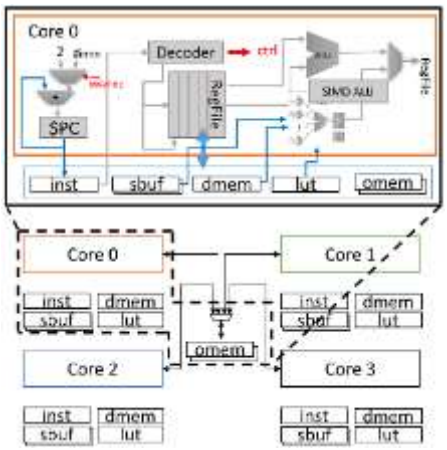

図 1 アーキテクチャの概要

2.1 コアのアーキテクチャ

コアは回路規模の小さなマイクロコントローラと SIMD 型積和演算器から構成される．マイクロコントローラは 16 ビット固定長の命令セットにより動作する．命令長が短いために命令デコーダや制御回路も単純化でき，小型で高電力効率なコントローラとなっている．

命令セットは，論理・算術演算，load/store 命令，分岐命令などの基本的な命令を含み，本研究のアクセラレータは汎用的な動作が可能であるため，様々なネットワーク構成に柔軟に対応できる．また，DNN の演算処理を高速化するための命令もいくつか追加されている．具体的には，ダブルバッファの切り替え制御命令やダイレクトメモリアクセス(DMA)発行命令，SIMD 型積和演算器の制御命令などである．特にマルチサイクルのカスタム SIMD 算術命令を定義しており，DNN の積和演算を行う際の制御オーバーヘッドを軽減する．具体的な制御オーバーヘッドとして，処理対象データにアクセスするためのアドレス計算や，ループの制御，条件分岐などの処理があげられる．本研究のアクセラレータでは，これらの処理と SIMD 型積和演算の動作シーケンスをハードウェアで実装し，マルチサイクルのカスタム SIMD 算術命令に集約している．これは，汎用命令セットでソフトウェア実装するのに比べ，CNN の識別高速化と消費電力削減の両方に効果がある．

2.2 SIMD 型積和演算器

SIMD 型積和演算器の基本構成を図 2 に示す．SIMD 型積和演算器は 16 ビット長データ 4 並列で

The Design and Implementation of 3D Stacked DNN Accelerator with TCI
Ryuichi Sakamoto†, Ryo Takada†, Jun Ishii†, Masaaki Kondo†, Hiroshi Nakamura†, Tetsui Ohkubo‡, Takuya Kojima‡, Hideharu Amano‡
The University of Tokyo†
Keio University‡

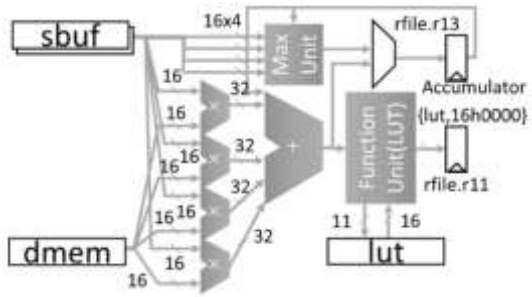

図 2　SIMD 型積和演算器

演算を行うことができ，実行可能な演算はテーブルルックアップ付きの積和演算と MAX 演算である．処理対象データはレジスタファイルを介さずに，sbuf と dmem から直接演算器に供給され，データバスは 64 ビット幅である．ルックアップテーブル(lut)はニューラルネットワークの活性化関数に利用する．DNN アクセラレータの先行研究では活性化関数に ReLU 関数のみをサポートするものもあるが，汎用性の観点からルックアップテーブルによる実装を採用した．これら積和演算と MAX 演算の結果は予め定められた汎用レジスタに書き込まれる．

2.3　TCI を用いた 3 次元積層

SNACC は積層する LSI の枚数を変えることによって性能を変化させることが可能であり，要求される電力や性能の仕様に対して柔軟に対応が可能である．図 3 に 3 次元積層を用いたアクセラレータ LSI の概要を示している．一番下にはアクセラレータ全体の資源を管理するための汎用のプロセッサがあり，ライブラリや OS が動作する．アクセラレータ LSI は最大 3 枚まで積層可能であり，LSI 間は TCI によって接続される．

我々はルネサス 65nm SOTB プロセスを用いてこれらのチップをテープアウトした．チップのサイズは 3mm x 6mm であり，4 つのコアと 68KB の SRAM が搭載されている．動作周波数は 50MHz であり，電源は 0.55V で動作可能である．

3　スケーラビリティ評価

本章では，スケーラビリティ評価としてコア数を変えた場合の 7 層 CNN アプリケーションの性能評価を示す．また，Off-chip メモリに対する DMA データ転送バンド幅もパラメータとする．結果を図 4 に示す．横軸はコア数を示し縦軸は電力効率を示している．本評価ではコア数を 1 コアから 16 コア，DMA バンド幅を 10MB/s から 3.6GB/s とした．これらより，データ転送バンド幅が増大するほどエネルギー効率が改善することが確認できる．実際にはデータ転送バンド幅を高くするには相応の電力コストが生じ性能が悪化する可能性があるが，今回の評価では考慮できていない．データ転送時間との兼ね合いを踏まえ

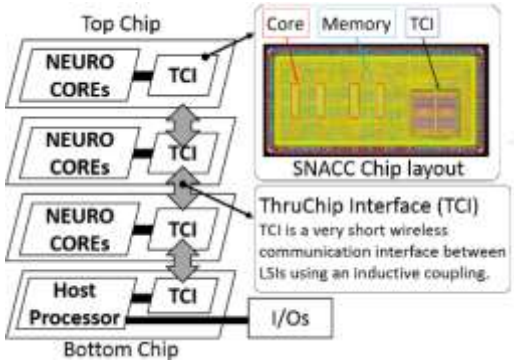

図 3　3 次元積層型アクセラレータ構成

ると，エネルギー効率を最大化する望ましい構成は 8 コア構成・DMA データ転送バンド幅が 500MB/s であることが分かった．

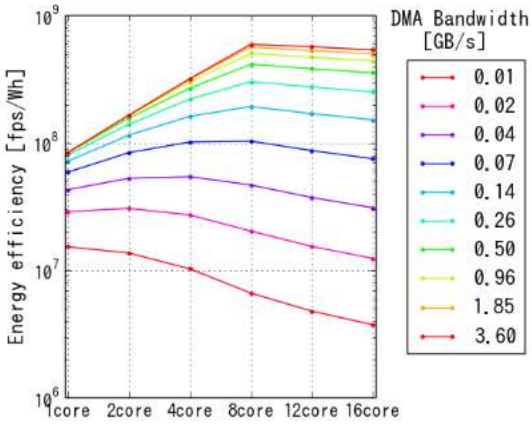

図 4　コア数とバンド幅を変えた場合の電力効率

4　まとめ

本稿では，高電力効率かつプログラマブルな動作が可能な DNN 向けアクセラレータのアーキテクチャを紹介した．さらに，コア数とバンド幅を変化させた場合の電力効率の評価について示した．

謝辞

本研究は JSPS 科研費基盤研究（S）　25220002 の助成によるものである．

参考文献

[1] Chen et al.: Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks, 2016 IEEE International Solid-State Circuits Conference (ISSCC), IEEE, pp. 262{263 (2016).

[2] Han et al.: EIE: Efficient Inference Engine on Compressed Deep Neural Network, Proceedings of the 43rd International Symposium on Computer Architecture, ISCA '16, Piscataway, NJ, USA, IEEE Press, pp.243-254 (2016).

[3] Chen et al.: Dadiannao: A machine-learning supercomputer, Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture, IEEE Computer Society, pp. 609-622 (2014).