ホワイトペーパー

組み込み AI アクセレーター（DRP-AI）

阿部　英明、IoT・インフラ事業本部エンタープライズ、コグニティブ・プロダクト部、ルネサス　エレクトロニクス株式会社

野瀬　浩一、IoT・インフラ事業本部エンタープライズ、コグニティブ・プロダクト部、ルネサス　エレクトロニクス株式会社

菊池　和貴、IoT・インフラ事業本部エンタープライズ、コグニティブ・プロダクト部、ルネサス　エレクトロニクス株式会社

2021 年 6 月

概要

近年のコンピューティング能力の目覚ましい進化により、AI(Artificial Intelligence)がクラウドサービスを皮切りに我々の生活に浸透し始めています。これに伴い、Gartner 社のレポート[1]によれば AI 半導体市場規模は 2019 年の$12Bilion から 2024 年には$43Bilion に拡大することが見込まれています。

新しいトレンドとして、クラウドに集中していた AI をエンドポイントへ実装する動きがあります。これはエンドポイント、例えば IoT 機器やロボットがよりスマートでリアルタイムに反応することを求められているためです。エンドポイントで求められる AI は人が行う視覚や聴覚といった知覚を代替えする深層・機械学習による推論処理です。

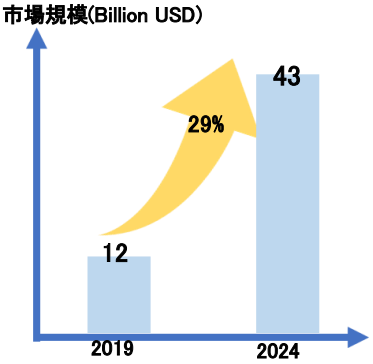

図1： AI 半導体市場規模推移

エンドポイントに AI を実装するには、大きく2つの課題をクリアする必要があります。1つ目は消費電力制限、2つ目は柔軟性です。十分な電力と冷却設備を備えられるクラウドと違い、エンドポイントでは駆動時間、発熱制限、コスト増の原因になる消費電力の制限が厳しく要求されます。消費電力を抑える手段として、特定の AI 処理に特化した専用ハードウェアを活用する方法があります。しかしながら、AI モデルは日々進化しているので、特定の AI に特化したハードウェアでは直ぐに陳腐化してしまいます。そこで、新しく開発される AI モデルに対応する柔軟性があるハードウェアの提供が必要になります。

ルネサスは長年培ったリコンフィギュアラブルプロセッサー技術をベースに、エンドポイントで求められるローパワーと柔軟性を兼ね備えた AI 推論処理を高速に処理する AI アクセラレータとして DRP-AI(Dynamically Reconfigurable Processor for AI)を開発しました。

[1] Graph created by Renesas based on Gartner Research, Source: Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, worldwide, Alan Priestley, 20 Apr 2020, Revenue Basis

DRP-AI アクセラレータの特徴

- AI 推論専用ハードウエアアクセラレータ
- HW(DRP-AI)と SW(DRP-AI トランスレータ)の協調による高い電力効率を実現
- DRP-AI トランスレータの継続的なアップデートにより AI モデルの拡張が可能

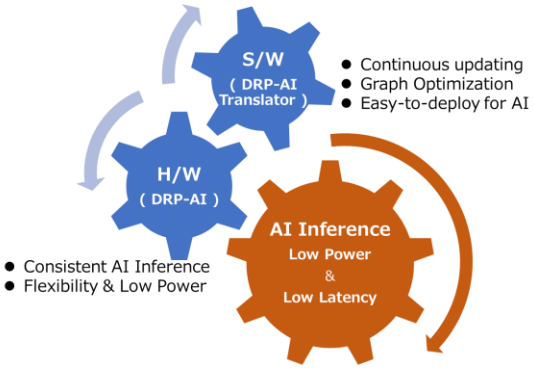

図 2: DRP-AI の特徴

AI 推論専用のハードウェアでありながら、ルネサス独自のダイナミックリコンフィギュアラブル技術を活用した柔軟性と高速処理および高い電力効率を実現しました。この柔軟なハードウェアに、ユーザーが簡単に AI モデルを DRP-AI の性能を最大限引き出すように最適化された状態で実装できるように、DRP-AI トランスレータを提供します。DRP-AI トランスレータにより出力される実行ファイルは、外部メモリに複数個配置することができます。これによりシステムとして複数の AI モデルをダイナミックに切り替えて使用することも可能です。また、DRP-AI トランスレータの継続的なアップデートにより、ハードウェアの変更なしで新しく開発される AI モデルにも対応可能です。

DRP-AI アクセラレータの H/W 構成

- DRP (Dynamically Reconfigurable Processor) : programable H/W core
- AI-MAC (multiply-and-accumulate)
- DMAC (Direct Memory Access Controller) :

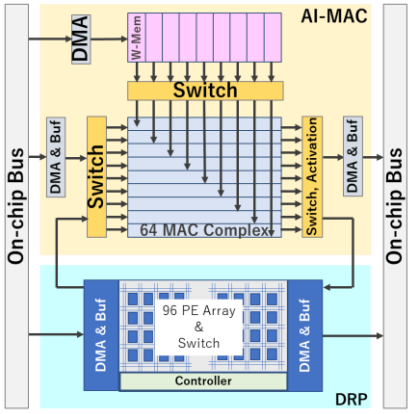

図 3: DRP-AI H/W 構成

DRP-AI は、AI-MAC と DRP(リコンフィギュアラブルプロセッサー)により構成されています。AI-MAC は内部スイッチでデータフローを最適化することにより、畳み込み層や全結合層の演算を効率的に処理することができます。DRP は画像の前処理や AI モデルの Pooling 層等の複雑な処理を、動的にハードウェアの構成を変更することで柔軟でかつ高速に処理することができます。この役割の違う2つのパーツを連携させることで、一貫した AI モデルを実行することができます。DRP-AI トランスレータが自動的に AI モデルの各処理を AI-MAC と DRP に振り分けます。したがって、ユーザーはハードウェアを意識することなく簡単に DRP-AI を使いこなすことができます。

DRP-AI トランスレータ

- 学習済み ONNX モデルを、DRP-AI に最適化した実行ファイルを生成するツール
- AI モデルのグラフ構造を最適化することで、メモリアクセスの最小化、演算の効率化を実現
- 継続的なアップデートにより、多様な AI モデルへの拡張性

DRP-AI トランスレータは、多様な AI フレームワークに左右されない ONNX フォーマットをベースとした学習済モデルから、DRP-AI に最適化した実行ファイルを生成するツールです。DRP-AI トランスレータの処理内容は、

1. AI モデルを処理する各オペレーションのスケジューリング
2. 1で決められたスケジュールの各オペレーション移行時に発生するメモリアクセス時間等のオーバーヘッドを隠蔽
3. ネットワークにおけるグラフ構造の最適化（Layer fusion、DRP & AI-MAC 処理振り分け）等

DRP-AI トランスレータを使うことで、ユーザーは ONNX 形式の AI モデルから DRP-AI のハードウェア構成を意識することなく自動的に最適化された AI モデルを DRP-AI に実装することが可能になります。そしてドライバーを通してコールするだけで簡単に高性能な AI モデルを実行できます。

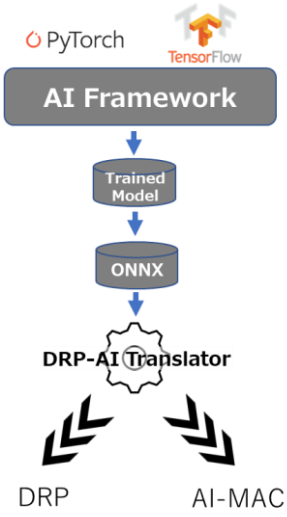

図 4: DRP-AI トランスレータによる AI モデル実装フロー

高い電力効率を実現するアーキテクチャー

- データの再利用による外部メモリ通信量の削減
- ゼロデータ入力を活用した低電力制御
- オペレーションフローのスケジューリング管理

データの再利用による外部メモリ通信量の削減

AI アクセラレータの消費電力は、莫大な行列演算による電力だけでなく、アクセラレータ内および外部メモリのデータ移動による電力成分も大きくなってきています。さらに、**図 5: AI モデルのデータ構成**のように、画像サイズやモデルに応じて、weight/input/output に関するデータ量の割合が変わり電力ボトルネック要因が変化します。そのため、様々な形状やサイズの AI モデルを包括的に低電力化するためには、すべてのデータに対するメモリアクセス量を低減する必要があります。

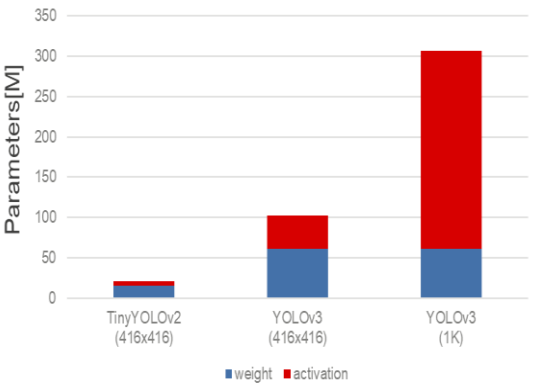

図 5: AI モデルのデータ構成

外部メモリアクセスを削減する有効な方法として、DRP-AI は一度入力されたデータを AI-MAC 内部で効率的に再利用する技術を採用しています。

たとえば 3x3 フィルタを用いた畳み込み演算では、1 画素のデータは 9 回のフィルタ演算に使われます。GPGPU などでの高並列演算手法として広く用いられている im2col 手法では、GPU に入力する前処理として、画像データを行列演算順に全て展開します。このとき、1 画素のデータ情報が 9 回あらわれるため、データ数が 9 倍に増えてしまいます。そのため、消費電力の増加や通信帯域の増加を引き起こします。一方で、AI-MAC は MAC 演算器に対応したレジスタへ取り込んだデータを隣のレジスタへシフトすることで、データを再利用することができます。具体的なフローを**図 6: データ再利用説明図**を使って説明します。

1. 外部メモリに格納されたデータ(青)を AI-MAC のバッファーへ取り込む(図 7-左)
2. バッファーからレジスタへデータ(青)を転送(図 7-中央)
3. そのレジスタのデータを使い対応する MAC 演算器で演算(図 7-右)
4. 一つ下のレジスタへデータ(青)をシフト(データ再利用)

この構成を採用することにより外部メモリ、内部バッファーから AI-MAC へのデータロード回数が GPU に比べ、最大 1/9 倍まで削減することができます。その結果、データ移動に必要な電力・通信帯域が大幅に削減されています。

また、AI-MAC では入力データ入力だけではなく、出力や重み情報それぞれに対しても、データの再利用を行い、外部メモリとのアクセスを 1 桁以上削減できます。

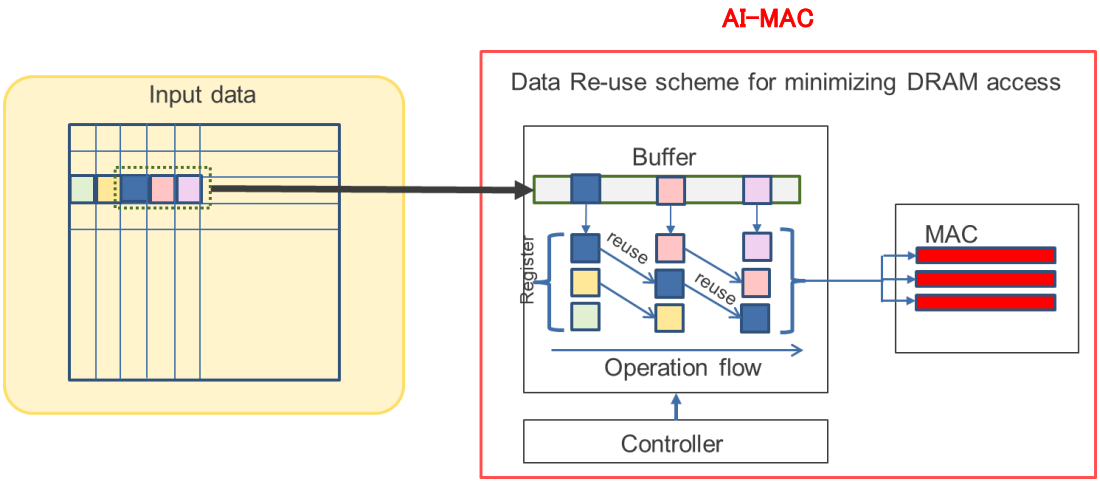

図 6: データ再利用説明図

ゼロデータ入力を活用した低電力制御

AI モデル演算における特徴の一つとして、重みデータや各レイヤの入出力データに「ゼロ」の値が入る割合が高い（いわゆるスパース化）ということがあります。

たとえば**図 7: AI モデルのゼロデータ比率**にあるように、画像認識モデルでは、全レイヤの平均 50%以上の入出力データがゼロ値となっています。これは多くの AI モデルにおいて、積和演算の結果がマイナスになるものをすべてゼロに置き換える活性化関数(ReLU)が使われている事などによります。

DRP-AI では、演算の要素単位で入力にゼロが入ることを事前に検知し、無駄な動作を行わないようスイッチングする技術を導入することで、不要な演算電力を削減しています。

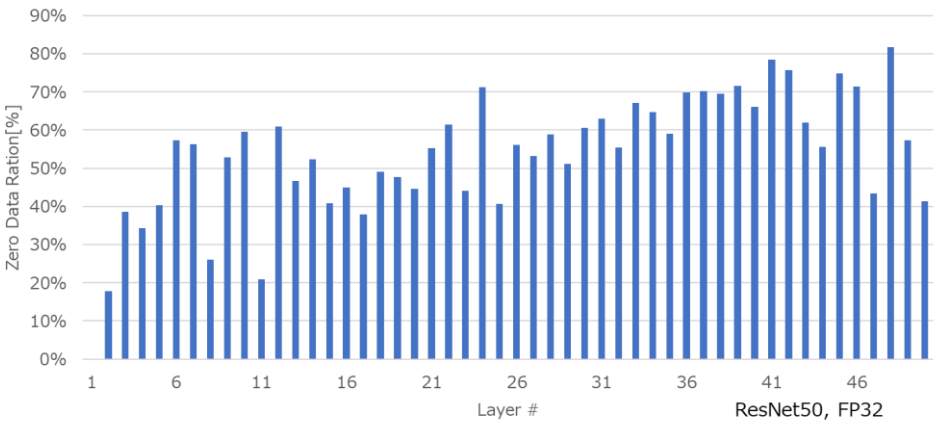

図 7: AI モデルのゼロデータ比率

オペレーションフローのスケジューリング

前述したデータ再利用技術に加え、外部データアクセスや MAC 演算処理の順序とタイミングの最適化が効率的な AI 演算を実現する上で不可欠です。言い換えれば、オペレーションフローのスケジューリングを行うことで、DRP-AI の性能を最大限に引き出すことができます。一つの例を説明します。外部メモリから AI-MAC 内のバッファーに直接書き込む構成となっているが、外部メモリアクセスには一定の時間が必要となります。AI-MAC 動作中に次のレイヤの重み情報を先読みしてバッファーへ格納するように外部メモリアクセスタイミングをスケジューリングすることで、外部メモリアクセスに掛かる時間を隠蔽することができます。このような事例は、他にも内部メモリアクセスや内部演算あらゆる処理のタイミングで発生し、スケジューリングを行うことで各処理間の無駄な待ち時間や電力の発生を避けることができます。この最適化されたスケジューリングを DRP-AI トランスレータが自動で生成するので、ユーザーは簡単に DRP-AI を扱うことができます。

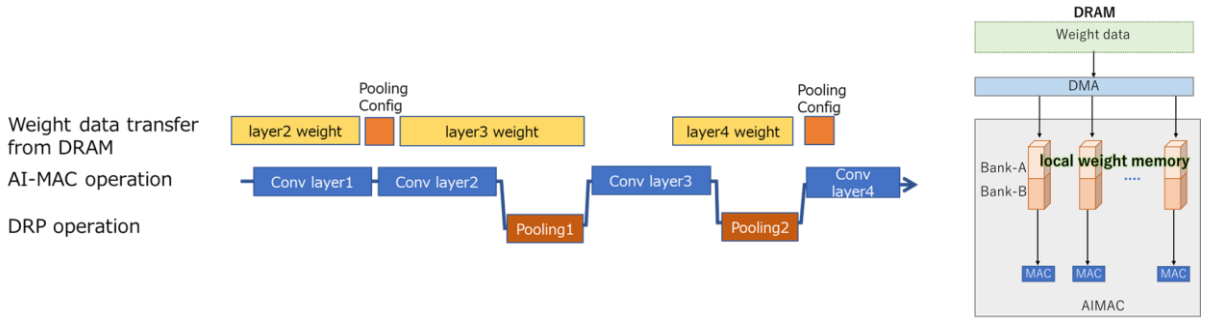

図 8: オペレーションフロースケジューリングの例

実測

これまで説明してきた高い電力効率を実現するアーキテクチャを採用した DRP-AI のテストビークルに TinyYolov2[2]を実装し、サーモグラフィーでデバイスの表面温度を計測する実験を行いました。

実験では DRP-AI の性能をわかりやすく表現するため、市販の GPU を比較対象として同一の条件で AI を実行させました。その結果が**図 9: AI 実行時のデバイス表面温度**です。

当社の DRP-AI テストビークルの表面温度が市販の GPU に比べて明らかに温度が低いことがおわかりいただけるでしょう。

DRP-AI テストビークルのデバイス表面温度はヒートシンクなしで 40.9° でした。この時の TinyYolov2 実行フレームレートは 42fps[3][4]です。一方、市販の GPU はヒートシンクが付いているにも関わらず 79.0° でした。その時の fps は DRP-AI と同等以下でした。

[2] Tiny Yolov2: https://pjreddie.com/darknet/yolov2/
[3] AI推論部のみの値
[4] デバイスセッティングの違いによりAI推論性能は製品毎に異なります

この実際のユースケースに近い実験において、DRP-AI 搭載製品が実用レベルの AI を実行しても温度制約の厳しいエンドポイント製品の量産に十分に耐えられる製品であることを理解いただけたことを期待します。

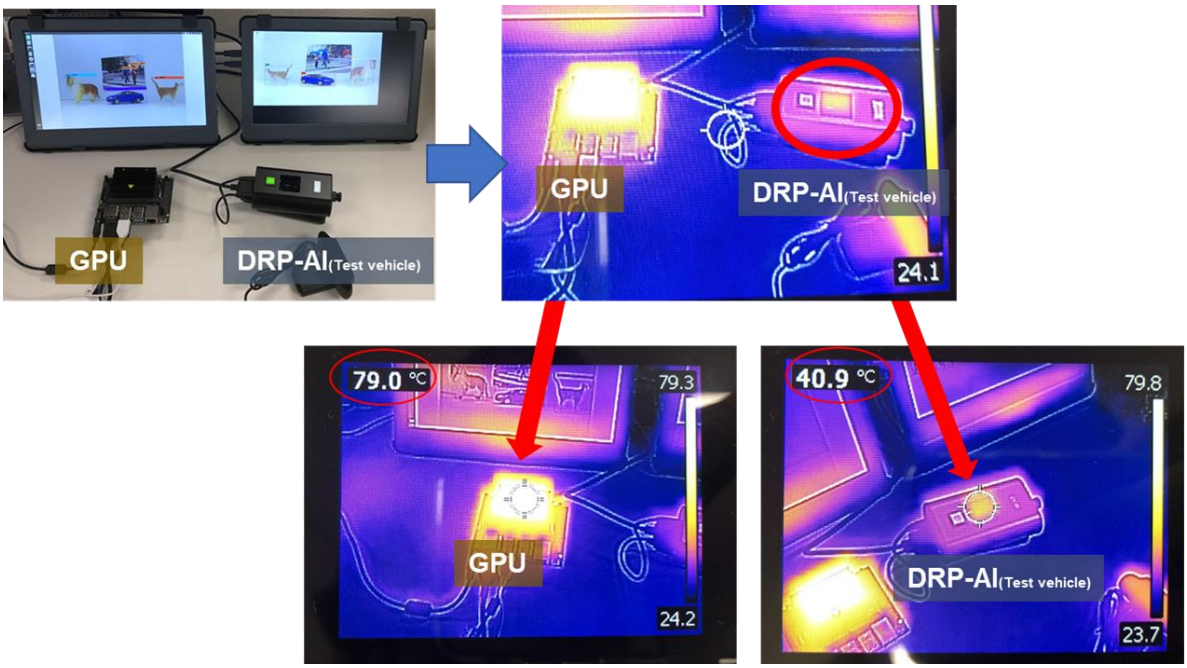

図 9: AI 実行時のデバイス表面温度

結論

ルネサスは、エンドポイントで求められるローパワーと柔軟性を兼ね備えた AI 推論処理を高速に処理する AI アクセラレータとして DRP-AI（Dynamically Reconfigurable Processor for AI）を開発しました。この優れた AI アクセラレータを搭載した MPU 製品をスケーラブルに展開していきます。これにより、エンドポイント製品をスマートでかつリアルタイムな反応をする AI 搭載に貢献いたします。

詳細情報

- RZ/V2M ルネサス独自 AI 専用アクセラレータ「DRP-AI」と 4K 対応イメージシグナルプロセッサ（ISP）を搭載し、組み込み機器におけるリアルタイムな人・物体認識を実現するビジョン AI 向け ASSP
- RZ/V2L ルネサス独自の AI 専用アクセラレータ「DRP-AI」と 1.2GHz Dual コア Arm® Cortex®-A55 CPU、3D グラフィックス、ビデオコーデックエンジン搭載の汎用マイクロプロセッサ