次世代高電力効率 AI アクセラレータ(DRP-AI3)の紹介
～自律システム等の高度な AI を組み込みで処理可能に～

戸井 崇雄, 課長, エンベデッドプロセッシングプロダクトグループ エンベデッドプロセッシング第一事業部　プロダクトマネジメント第二部 第二課

下別府 正行, 課長, エンベデッドプロセッシングプロダクトグループ エンベデッドプロセッシング第一事業部 プロダクトマネジメント第二部 第三課

三上 顕太郎, 主任技師, ソフトウェア＆デジタライゼーショングループ ソフトウェア開発統括部 システムソリューション開発第一部 ＡＩソリューション開発課

野瀬 浩一, シニアプリンシパルプロダクトエンジニア,エンベデッドプロセッシング第一事業部 / エンベデッドプロセッシングプロダクトグループ

概要

少子高齢化に伴って労働人口が減少する中、工場、物流、医療、街中で稼働するサービスロボットやセキュリティカメラなど社会のさまざまな場面で、周辺環境の認識、行動の判断や動作制御など、高度な人工知能（AI）処理を含めた多様な種類のプログラムをリアルタイムに処理するシステムが必要となっています。特に、環境認識した結果を即座にロボットの行動に反映させるためには、システムを機器内に組み込み応答性を高める必要があります。しかしながら、発熱に対する制限が厳しい組み込み機器で高度な AI 処理を行うには、AI チップの低消費電力化がより一層求められています。

これらの要求に応えるため、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、このアクセラレータを搭載した AI 向け MPU として RZ/V シリーズを提供してきました。

今回は、さらなる AI の進化やロボット等のアプリケーションの高度化に対応するため、DRP-AI の次世代版として、従来比約 10 倍の電力効率を実現する DRP-AI3 を開発しました。本ホワイトペーパーでは、DRP-AI3 の主な開発技術の概要および実証結果により、発熱の問題を解決し、リアルタイム性の高い処理速度が実現、AI 搭載した製品の高性能化と低消費電力化が実現できることについて紹介します。

DRP-AI3 アクセラレータの特徴

1. AI モデル軽量化（枝刈り）に対応した H/W-S/W 協調により、従来比約 10 倍の電力効率を実現
   - 枝刈りモデルの高速・低電力化技術を導入した AI アクセラレータ(DRP-AI3)
   - RP-AI に適した枝刈りモデルを容易に生成し、H/W に最適実装するソフトウェア
2. DRP-AI, DRP, CPU が協調動作するヘテロジニアスアーキテクチャにより、AI 以外の多様なアルゴリズムも高速化

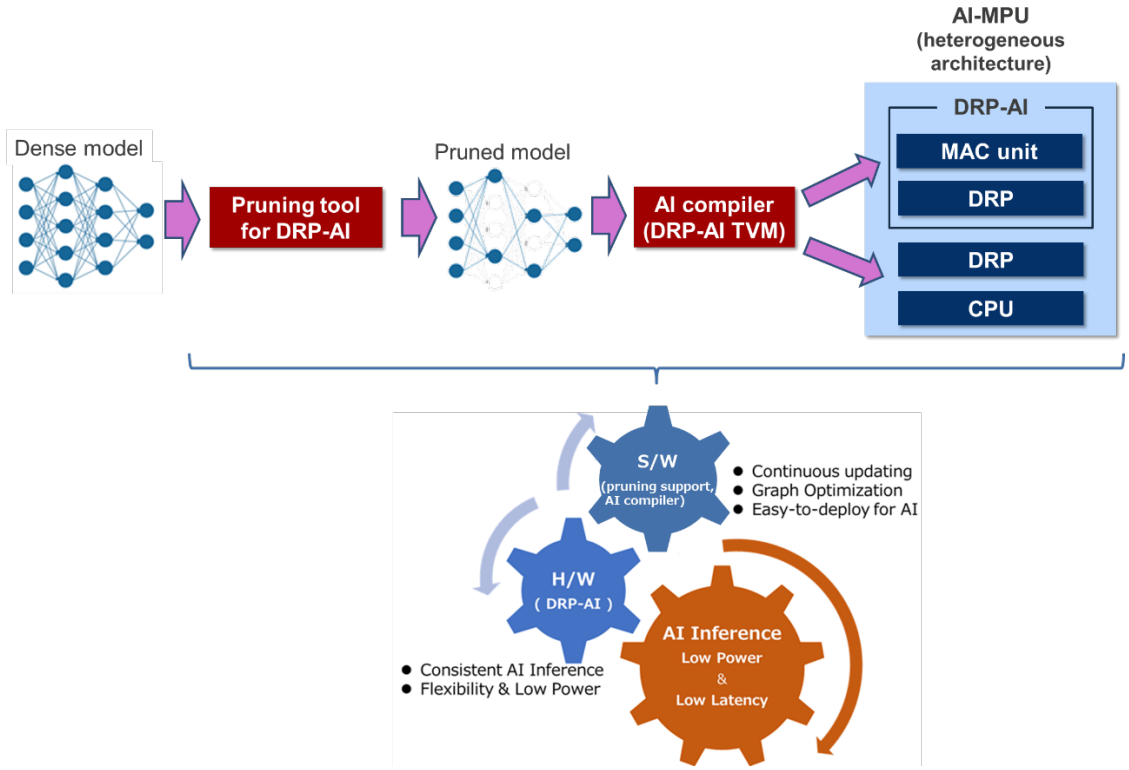


図 1. DRP-AI3 のハードウェア・ソフトウェア協調設計

枝刈り AI モデルの高速・低電力化ハードウェアの特長

- 主要な軽量化技術であるビット数低減(INT8)に加え、枝刈り技術にも対応したハードウェアアーキテクチャ
- DRP-AI の柔軟性を活かし、既存のハードウェアでは困難であったランダムな枝刈りモデルの高速化に対応
- 枝刈り適用前に比べて処理時間は最小 1/16、消費電力は約 1/8 まで削減可能

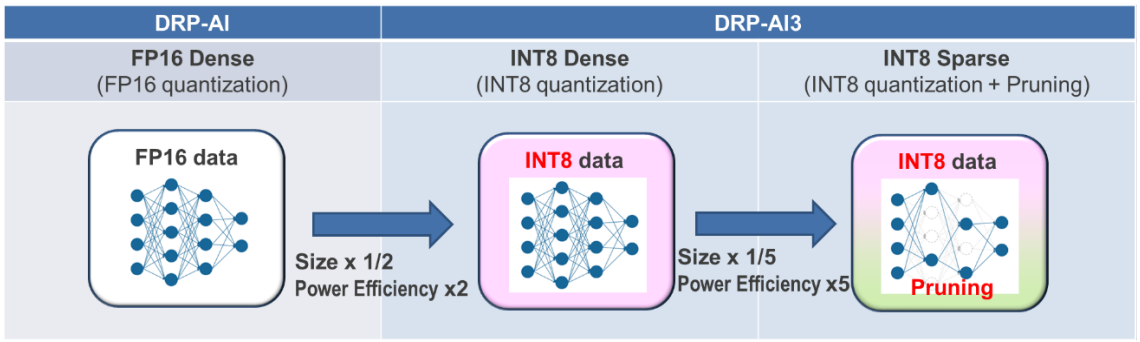


図 2　DRP-AI3 に適用した軽量化技術

DRP-AI3 は、主要な AI モデル軽量化手法に対応した高速・低電力化手法を導入しています。具体的には以下の軽量化手法に対応しています。

1) 量子化：ニューラルネットワークの重み情報(weight)と各レイヤの入出力データ(feature map)の低ビット化（従来の DRP-AI の 16 ビット浮動小数点演算から 8 ビット整数演算（INT8）に変更）

2) 枝刈り：認識精度に影響しない重み情報（枝）を 0 として計算をスキップする技術

理想的には、1）の量子化は、演算器の規模やデータアクセス量がビット数に応じて軽量になるため、従来の DRP-AI (16 ビット処理)に比べて約 2 倍以上の低電力化が期待できます。また、2）の枝刈りは、どの程度の重み情報を残せるかは AI モデルに依存しますが、たとえば 90%程度の枝刈りが実現できる場合、約 10 倍の高速化や低電力化が期待値となります。

しかしながら、現状の AI ハードウェアは、特に 2）枝刈りをした AI モデルを効率よく処理できないことが大きな課題となっています。AI ハードウェアは、ニューラルネットワークの大規模な積和行列演算を効率よく処理するため、多数の積和演算を同時に行う SIMD (Single Instruction Multiple Data) アーキテクチャになっていることが一般的です。一方、認識精度に影響しない重みの場所は行列内にランダムに存在しているため、並列積和演算の内部で一部の重みがゼロになっても、ゼロ以外の重みと一緒に並列計算が行われてしまうため、枝刈りによって計算量を減らすことはできませんでした(図 3)。

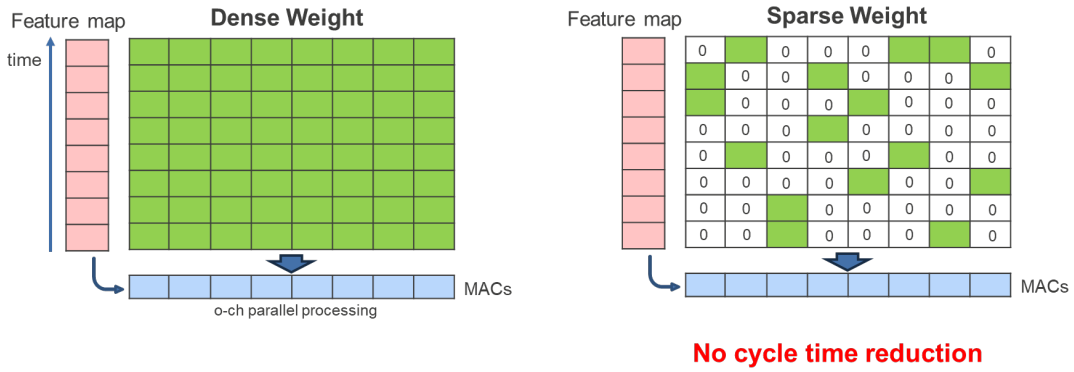


図 3　一般的な並列アーキテクチャによる軽量化（枝刈り）モデルの処理

AI ハードウェアで実用化されている枝刈り対応手法としては、並列性を損なわない範囲で値をゼロにする（たとえば重み行列の列単位で 0 にする）という構造的枝刈り手法 (structured pruning)が知られています。しかしながら、本来ランダムに存在する認識精度に影響しない重み情報とは大きく条件が異なるため、枝刈り率を高くとることができませんでした。その他の手法としては、重み行列内で隣接した 2 つの重みのうちいずれか 1 つを選択して演算するという手法があります[(1)]。この方式も重みの情報量を最大でも 1/2 までしか削減することができないため、高速化や低電力化の効果は限定的になってしまいます。

そこでルネサスでは、よりランダムな枝刈りにおいても柔軟に演算をスキップできる手法として、フレキシブル N:M 枝刈り手法を開発しました。

この技術の基本コンセプトは、図 4 にある通り、元の重み行列を M 行ごとの重み行列グループに分割し、各グループで有意な重みのみを抽出した小さな N 行の重み行列グループに再構築し、新たな重み行列グループに対して並列演算を行うものです。このとき、重み行列グループごとに N の値を自由に切り替えて演算サイクル数を調整できる機能を DRP-AI 内に新たに搭載したことで、図 4 にあるように実際の AI モデル内で局所的に変化する枝刈り率に対しても、最適な演算スキップ処理が可能になりました。また、N を細かく変化できることで、重み行列全体の枝刈り率も細かく設定できるようになり、ユーザが必要とする消費電力や動作速度、認識精度に応じて最適な枝刈り処理が可能になりました。

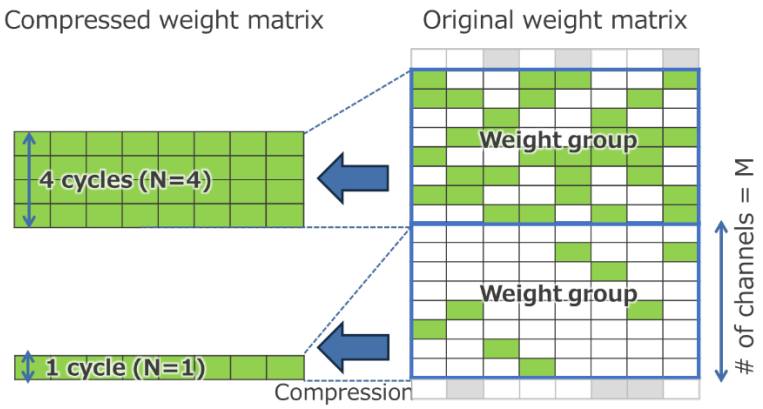


図４　DRP-AI3 による軽量化（枝刈り）モデルの圧縮

また、実際の AI モデルを対象に、枝刈り方式と認識精度との関係を分析することで、精度と面積、電力増加のトレードオフを考慮した最適なパラメータ値を特定し、ハードウェアアーキテクチャに反映いたしました。

本技術により、AI モデルの処理サイクル数を最小で 1/16 まで削減するとともに、消費電力も最小で約 1/8 にすることが可能になり、従来の AI アクセラレータ構成よりも大幅に処理効率が向上しました（図 5）。これにより、従来の AI プロセッサの課題であった発熱の問題が解消され、ロボットや小型の AI 機器の内部に冷却機構無しで実装する事ができるようになります。

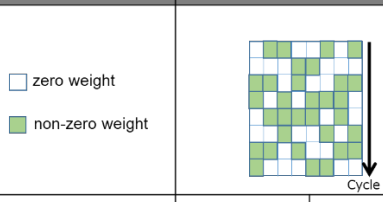
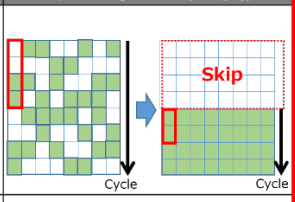
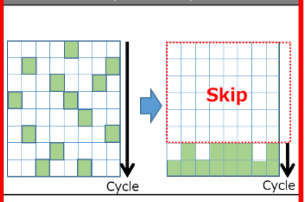
	General AI accelerators		50% pruning arch. (ref [1])	This work (DRP-AI)
				
Pruning structure	structured	unstructured	unstructured	unstructured
pruning rate	0~30%	0~90%	0 / 50%	0 ~ 93%
weight data size	~2/3x	~1/10x	~5/8x	~1/10x
Cycle time	~2/3x	1x	1/2x	~1/16x

図５　アクセラレータごとの枝刈りモデル処理性能の比較

枝刈りモデルの生成・実装ソフトウェアの特長

- DRP-AI3 のアーキテクチャの特徴を基に、どこを枝刈りすると効率よく性能向上できるかを考慮した、ハード・ソフト協調設計
- DRP-AI Extension Pack (枝刈りツール) を開発
- DRP-AI TVM (INT8 量子化ツールとコンパイラ)を開発

枝刈りは前述の通り認識精度に影響しない重み情報（枝）を 0 として計算をスキップする方法です。AI モデルの中で枝刈りを行う箇所が多いほど高速化や低電力化が見込めますが、認識精度は低下する傾向があります。

認識精度の低下を抑えつつ枝刈り率を向上させるため、一般的に枝刈りには図 6 に示すように、初期学習（Training）後、枝刈り箇所を選択して枝刈り（Pruning）を行い、枝刈り後の重み情報を使用して再学習（Retraining）を行うフローが適用されます。再学習には枝刈りを行うことによる認識精度の低下を抑制する効果が期待できます。

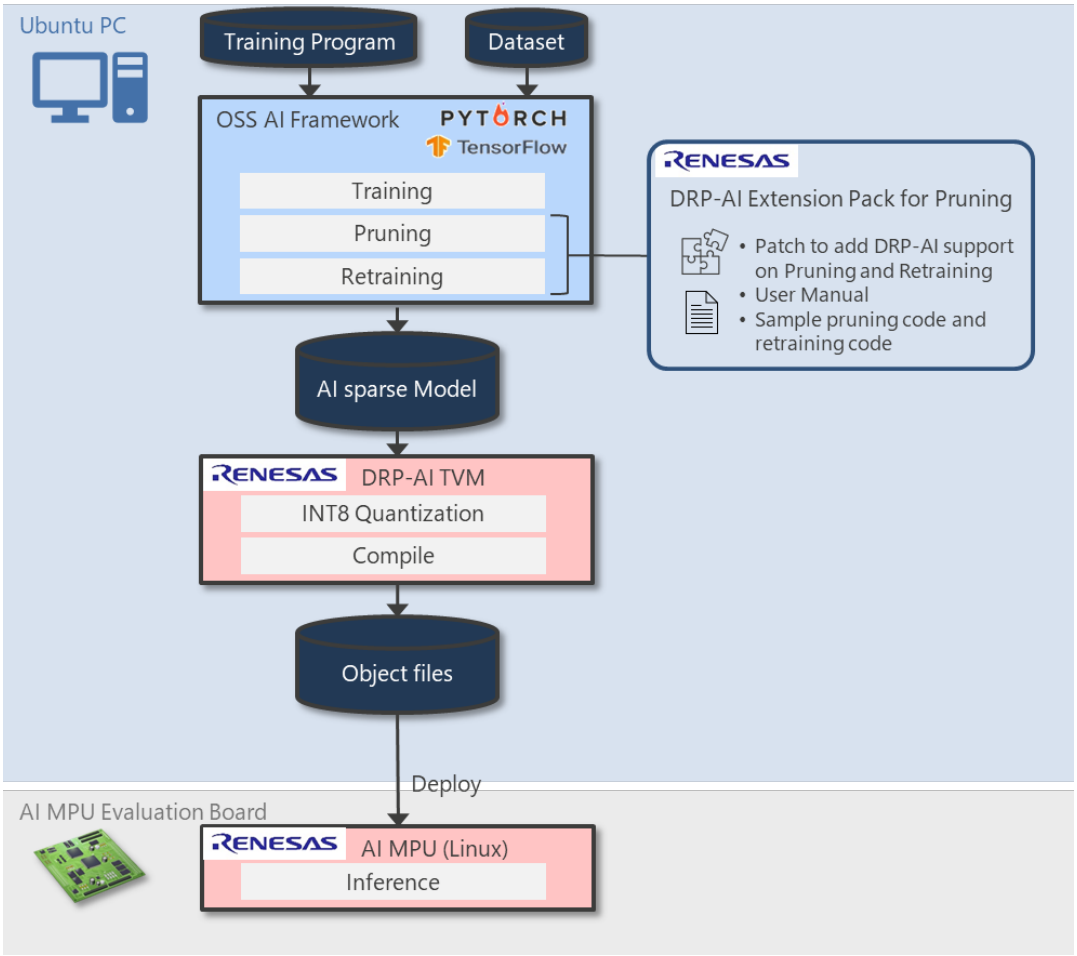


図 6 DRP-AI3 の AI モデル軽量化・実装フロー

このフローの中で枝刈り箇所の選択は非常に重要な要素です。一般的に初期学習後、認識精度への影響が少ない、重みの絶対値が小さい箇所や学習時の勾配が小さい箇所などが枝刈り箇所として選択されます。ルネサスではこれらの要素に加え、前述した DRP-AI3 の枝刈りのハードウェアアーキテクチャの制約を満たすように枝刈り箇所を選択する枝刈りツール（DRP-AI Extension Pack）を開発しました。ユーザは枝刈り率を指定するだけで、DRP-AI3 の特徴である「フレキシブル N:M 枝刈り」を適用することができます。

さらに枝刈りの導入を容易にするため、上記ツールは OSS AI フレームワーク（Pytorch、Tensorflow）のライブラリ形式で提供され、ユーザが保有している学習用のスクリプトに数行追加するだけで枝刈りと再学習（図 6 の Pruning と Retraining）を実現できます。

また、枝刈りツールで生成した枝刈り後の AI モデルを DRP-AI TVM で変換することで INT8 量子化とコンパイルを同時に実行できます。DRP-AI3 で実行可能なオブジェクトファイルに関しても簡単に生成することが可能です。

ここで DRP-AI TVM は学習済み AI モデルをルネサスの AI MPU で実行可能な形式に変換するツールです。OSS の ML コンパイラフレームワークである「Apache TVM」をベースに開発されており、AI モデルの各レイヤの処理のうち、DRP-AI3 で実行可能なオペレーションを DRP-AI3 へ、DRP-AI3 で実行できないオペレーションを CPU へ割り振って演算を行うことが可能です。このように複数のプロセッサを併用して演算処理を行うことをヘテロジニアス・コンピューティングと言います。DRP-AI3 は強力な AI アクセラレータですが、実行可能なオペレーションには限りがあります。ヘテロジニアス・コンピューティングの仕組みを導入することで、対応可能な AI モデルの種類（AI モデルカバレッジ）を大幅に拡張することができます。

このようにルネサスではハード・ソフト協調設計により、DRP-AI3 のハードウェアアーキテクチャを最大限活用し枝刈り効率を高めつつ、ユーザの枝刈り導入の手間を最小にする DRP-AI Extension Pack や、ヘテロジニアス・コンピューティングにより AI モデルカバレッジの大幅な拡張を実現できる DRP-AI TVM といったソフトウェア環境を提供し、UX 改善を推進しています。

DRP-AI, DRP, CPU が協調動作するヘテロジニアスアーキテクチャの特長

- AI アクセラレータ, DRP, CPU の連携によるマルチスレッド&パイプライン処理
- DRP(動的再構成可能な wired logic hardware)によるロボットアプリの低ジッタ・高速化

サービスロボットなどでは、周辺環境の認識などにより高度な AI 処理が求められてきています。一方で、ロボットの行動判断や制御においては、AI を用いないアルゴリズムベースの処理も同時に必要になっています。しかしながら、現在の組み込みプロセッサ（CPU）においては、これら多様な処理をリアルタイムに行う十分な性能が無いことが課題になっています。そこでルネサスは、動的再構成プロセッサ(DRP)と AI アクセラレータ(DRP-AI)および CPU との協調動作が可能なヘテロジニアスアーキテクチャ技術を開発し、この課題を解決しました。

動的再構成プロセッサ(DRP)は、図 7 にあるように、チップ内の演算器の回路接続構成を処理内容に応じて動作クロックごとに動的に切り替えながらアプリケーションを実行可能で、必要な演算回路だけが動作するため、CPU 処理よりも消費電力が小さく、高速化も可能になります。さらに、キャッシュミス等による外部メモリアクセスが頻繁に起こり、性能が低下してしまう CPU に比べて、DRP は必要なデータパスを先にハードウェア上に構築できるため、メモリアクセスによる性能低下や動作速度のばらつき（ジッタ）が少ないという特徴もあります。

また、アルゴリズムが変わるごとに回路接続情報も切り替える Dynamic Loading 機能も有しているため、多数のアルゴリズムを処理する必要があるロボットアプリケーションなどにおいても、限られたハードウェアリソース内で処理することが可能となっています。

DRP は特に、並列化やパイプライン化が性能向上に直結する画像認識などのストリーミング処理で大きな効果を発揮します。一方、ロボットの行動判断や制御などのプログラムでは、周辺環境の変化に応じて条件や処理内容を細かく変えながら処理する必要があり、これは DRP のようなハードウェア処理よりも CPU のソフトウェア処理の方が適している場合があり、適材適所に処理を分配し、協調動作することも重要となります。

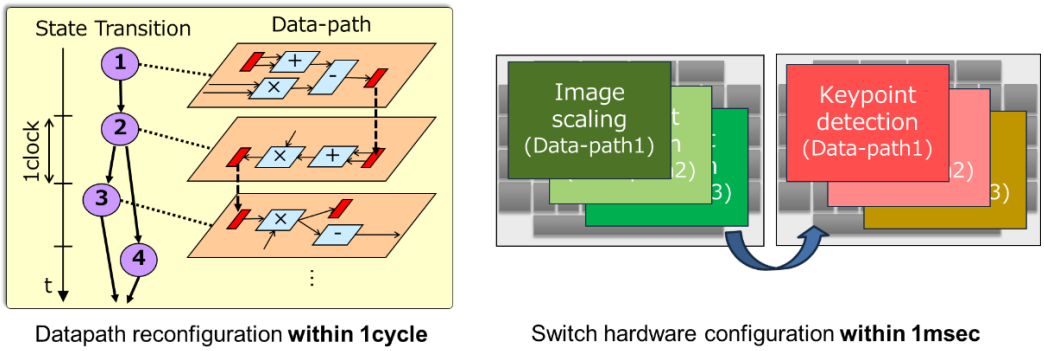


図 7 柔軟性の高い動的再構成プロセッサ(DRP)の特徴

MPU と AI アクセラレータ(DRP-AI)のアーキテクチャの概要を図 8 に示します。ロボットアプリケーションは、AI による画像認識と AI を使わない判断や制御のアルゴリズムを高度に併用します。そのため、AI 処理向けの DRP(DRP-AI)と非 AI アルゴリズム用の DRP を搭載した構成を採用することで、ロボットアプリケーションのスループットが大幅に向上します。

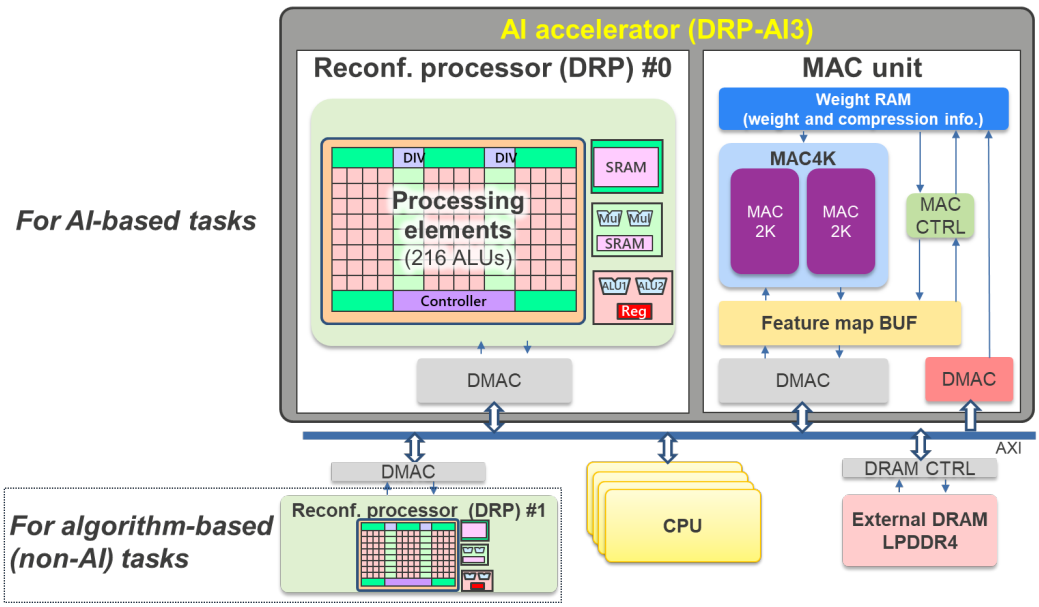


図 8　DRP-AI3 ベースのヘテロジニアスアーキテクチャの構成

評価結果

(1) AI モデル処理性能の評価

本技術を搭載したテストチップを試作し、AI アクセラレータの積和演算処理性能としては、最大 8 TOPS（1 秒間に 8 兆回の積和演算が可能）を実現しています。さらに、枝刈り処理した AI モデルについては、枝刈り量に応じて演算サイクル数を削減できるため、枝刈り前のモデルに換算すると最大 80TOPS の処理性能に相当する AI モデル処理性能を実現しております（注 1）。これは、従来の DRP-AI 比約 80 倍の処理性能と大幅な性能向上となっており、急速な AI の進化に十分追随できるものとなっております(図 9)。

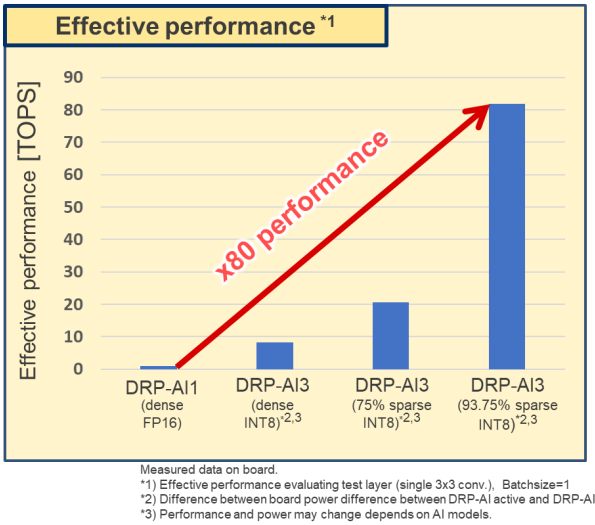


図９　DRP-AI のピーク性能実測比較

一方、AI 処理が高速化されるにつれ、AI の前後処理などに代表される AI を使わないアルゴリズムベースの画像処理の処理時間が相対的にボトルネックになりつつあります。AI-MPU においては、画像処理プログラムの一部を DRP にオフロード処理することで、システム全体の処理時間の向上に寄与しています (図 10) 。

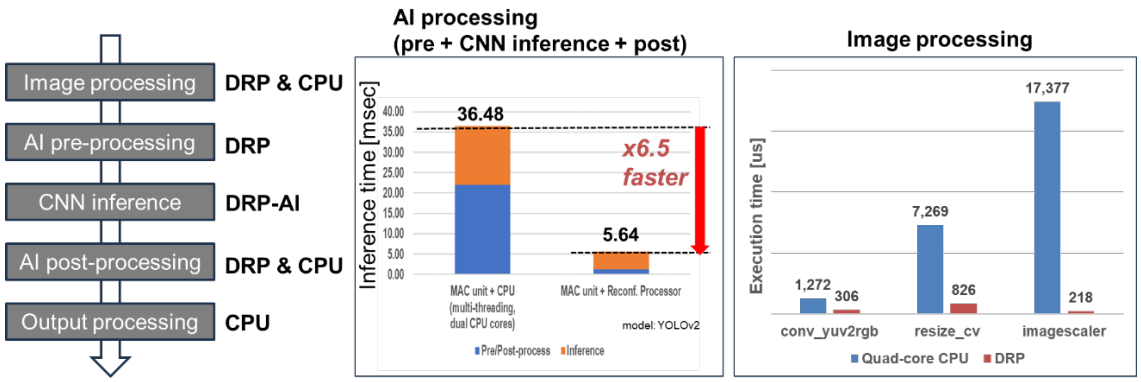


図１０　ヘテロジニアスアーキテクチャによる画像認識処理の高速化

電力効率については、AI アクセラレータ単独の性能評価において、理論最大性能としては約 23TOPS/W, 主要な AI モデル動作時においても世界トップレベルの電力効率（1 ワット当たり約 10TOPS）を実証しました (図 11) 。

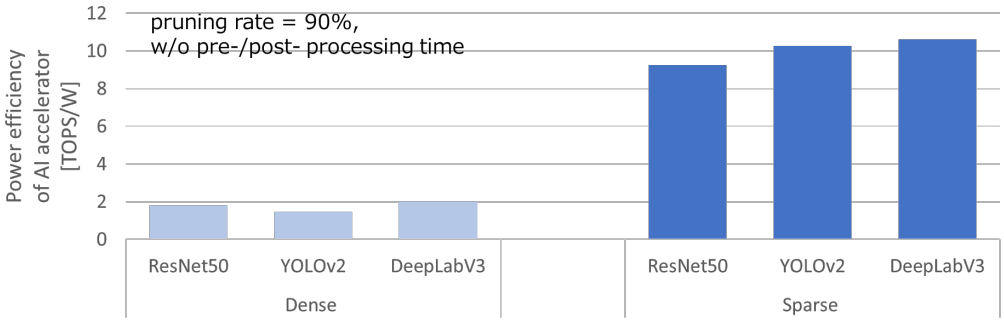


図 1 1　実 AI モデルの電力効率

また、試作チップを搭載した評価ボードにおいては、同じ AI リアルタイム処理を行った場合、ファンを搭載した他社製品と同程度の温度で、ファンなしで AI 処理を行うことが可能であることを示しました。(図 12)

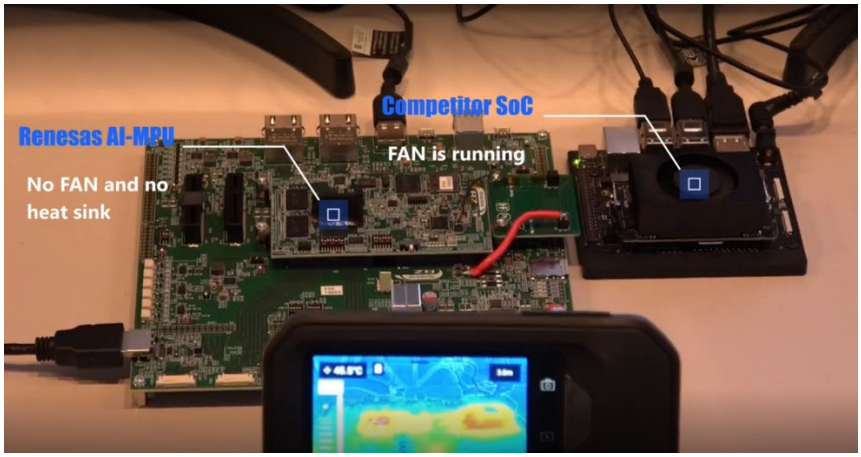


図 12 ファン不要な DRP-AI テストボードと GPU(ファン付き)との発熱比較評価

（２）ロボットアプリケーションへの適用事例

たとえば代表的なロボットアプリの一つである SLAM（Simultaneously Localization And Mapping）では、AI 処理による環境認識と並行し、ロボット位置認識のための複数のプログラム処理が必要になる複雑な構成になっていますが、DRP により瞬時にプログラムを切り替えながら実行し、さらに AI アクセラレータや CPU との並列動作により、CPU 単独動作に比べて約 17 倍の高速動作や、消費電力を 1/12 程度に削減できることを実証しました。

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独自の AI アクセラレータである DRP-AI（Dynamically Reconfigurable Processor for AI）について、軽量化 AI モデルの処理機能を進化させた DRP-AI3 を開発し、従来比 10 倍(10TOPS/W)の電力効率を実現しました。この優れた AI アクセラレータを搭載した MPU 製品（RZ/V シリーズ）をスケーラブルに展開していきます。

ルネサスは、今後ますます高性能化が進む AI の進化に対応した製品をタイムリーに供給し、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システムの実現に貢献いたします。

より詳細な技術内容については、2024 年 2 月 18 日より開催された半導体回路の国際トップ学会である ISSCC 2024（International Solid-State Circuits Conference 2024）で発表しております。

参考文献

[1] “NVIDIA Jetson AGX Orin Series tech. brief”, 2022
https://www.nvidia.com/content/dam/en-zz/Solutions/gtcf21/jetson-orin/nvidia-jetson-agx-orin-technical-brief.pdf

関連情報

- RZ/V2H: 高効率 AI 推論と高速リアルタイム制御を実現できる４コアビション AI プロセッサ
- DRP-AI : 高い AI 推論性能と低消費電力を両立したルネサス独自の AI アクセラレータ

本社所在地

〒 135-0061 東京都江東区豊洲 3-2-24
（豊洲フォレシア）
https://www.renesas.com

お問合せ窓口

弊社の製品や技術、ドキュメントの最新情報、最寄りの営業お問合せ窓口に関する情報などは、弊社ウェブサイトをご覧ください。
http://www.renesas.com/contact/

Doc Number: R01WP0022JJ0100