News Release
2025.5.30

NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）
シャープ株式会社

映像データ処理用 AI デバイス向け高位合成ツールを開発しました
―回路設計期間の大幅短縮とデバイスでの電力効率 40 倍以上を確認―

NEDOは「省エネAI半導体及びシステムに関する技術開発事業」（以下、本事業）の一環として、エッジ領域に適したAI半導体デバイスの早期実現を目指した開発を進めています。このたび、シャープ株式会社（以下、シャープ）は、エッジコンピューティングにおけるAI映像データ処理の普及拡大を目的とした、AIデバイス向け高位合成ツール（以下、本ツール）を開発し、オープンソースソフトウェア（OSS）として本日公開しました。

本ツールは、映像データ処理用AIアルゴリズムが記述されたPython（パイソン）コードから、集積回路であるFPGAなど、ハードウェアの回路図となるRTLコードを、短時間で自動生成するソフトウェアです。専門技術者でなくても映像データ処理用AIデバイスの設計が可能となり、開発期間を大幅に短縮できます。また、本ツールで生成されたRTLコードを実装したFPGAで実証したところ、GPUを搭載したエッジ端末に比べ、40倍以上の高い電力効率が得られることを確認しました。

図1　本ツールの利用イメージ

1. 背景

パソコンやスマートフォンなどの情報処理端末の高度化やICT／IoT社会の進展、AI活用の拡大などによるネットワーク上のデータ転送量の急増に伴い、ネットワーク強化やデータセンター増設が進められていますが、これらを運営するための電力確保が課題となっています。その解決策の一つに、情報処理をクラウド側とエッジ側で分散的に行う分散コンピューティングがあります。しかし、AI活用における一般的なエッジ端末に搭載されている処理デバイスであるGPU※1の消費電力が大きいため、分散による消費電力低減効果

が薄れてしまう問題があります。この問題は、エッジ端末内の処理デバイスを、GPUから、FPGA※2やLSIなどの消費電力の低い専用デバイスに置き換えることで解消可能ですが、いずれのデバイスも急速に開発が進む高度なAI処理に対応できないケースがあるほか、専用デバイス開発には専門技術者による回路設計と開発期間が必要です。

このような状況の下、NEDOは2023年度からエッジコンピューティングの産業応用に取り組んでおり、その一環として、本事業※3をシャープと共同で進めてきました。

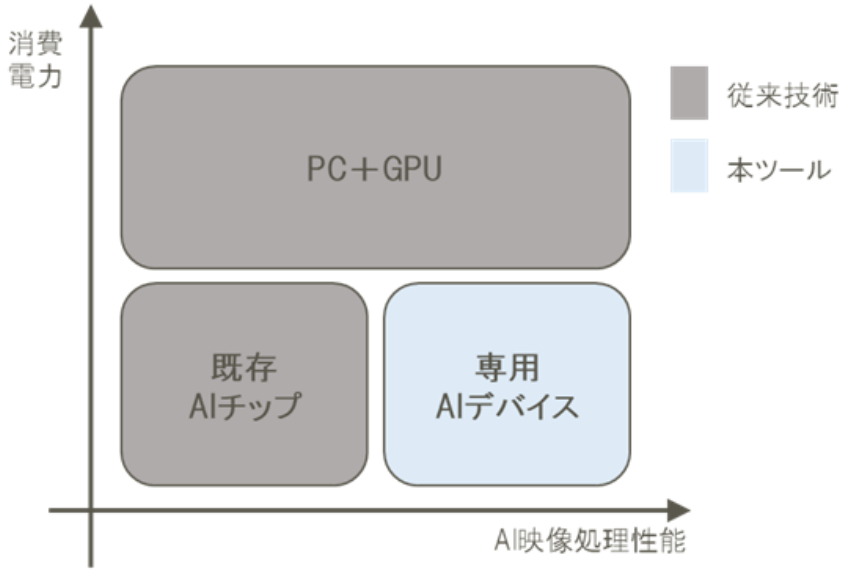

図2　本ツールが対象とするデバイスのイメージ

2. 今回の成果

本事業においてシャープは、テレビなどのディスプレイ製品開発で長年培ってきた映像処理に関するソフトウェアやハードウェアの技術、ノウハウを基に、本ツールを開発し、OSS※4として本日一般公開※5しました。AI開発における主流プログラミング言語であるPythonコードを読み込ませることで、RTLコード※6を自動的に短時間で生成します。生成されたRTLコードをFPGAに実装することで、AI専用処理デバイスとして利用可能になるほか、設計回路ツールとして、映像データ処理向けAI専用LSIの開発にも利用できます。専門技術者による設計作業が軽減され、開発プロセスの効率化と期間短縮化が図れることから、画像認識や映像内の物体検出、あるいは超解像などAI映像処理を利用した端末やアプリケーション開発での利用が期待され、エッジコンピューティングにおけるAI処理の普及拡大に貢献します。

今回開発した本ツールは、図3のデータ処理フローを用いており、次の機能や特長を持っています。

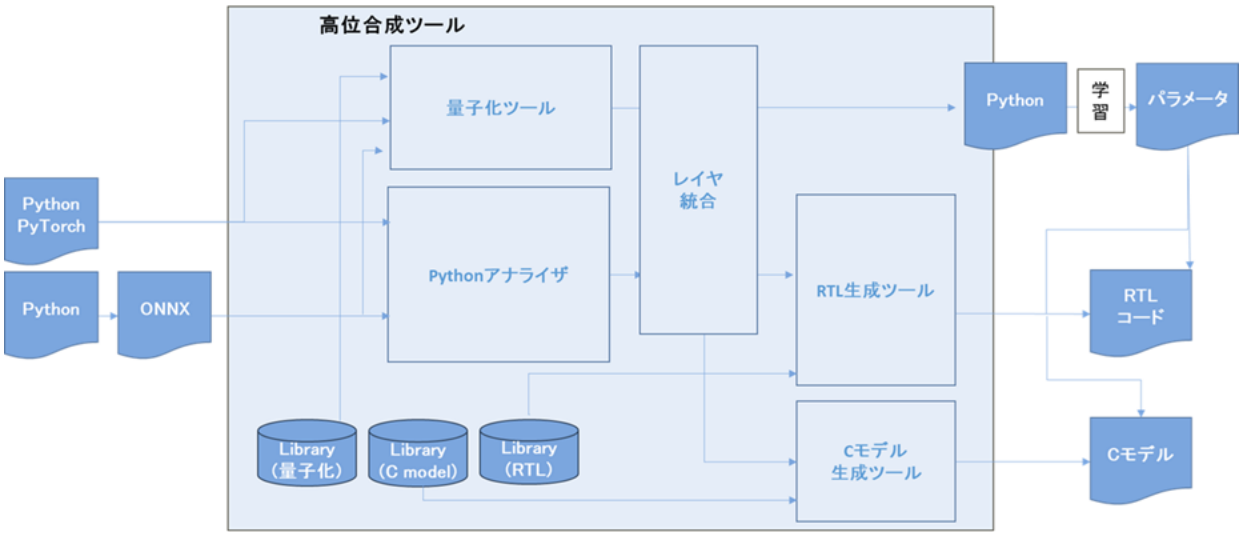

図3　本ツールでのデータ処理フロー

(1)高位合成ツールの機能

入力ファイルであるPythonコードに関して、AI開発において利用頻度の高いPyTorch※7やONNX※8に対応できるようフレームワークを拡張し、多様なAI映像処理アルゴリズムからのRTLコード生成を可能としました。複数の演算処理を一つの処理に統合するレイヤ統合機能も搭載し、演算処理の効率化や高速化を実現します。また、ハードウェア開発の効率化のため、代表的なプログラミング言語の一つであるC言語で、RTLコードの機能検証が行えるCモデル生成機能も搭載しました。さらに、専用デバイスでAI処理を行う際に必要となる学習済みパラメータを得るための、量子化対応したAIモデルもPythonコードで生成します。

(2)RTLコードの特長

本ツールが生成するRTLコードには、本事業で開発した回路技術を多数組み込むことで、AI映像処理回路の消費電力削減を実現します。

1点目は、映像データがライン単位で順番に伝送されることを活かし、従来方式では必要だったフレームメモリを使用せず、ラインメモリで効率的に処理を行うフレームメモリレス構造を開発しました。この開発により、消費電力を削減し、データ入出力に必要な時間を短縮します(図4)。

2点目は、本ツール用に、畳み込み演算などの代表的なAI演算に特化した4次元での処理に対応する演算器構成を開発しました。これにより、映像データの特性である2次元配列に最適な情報処理が可能となり、演算性能が飛躍的に向上し、複雑な計算を効率よく高速で実行できるため、リアルタイムでの処理が可能となります。

3点目は、隣接するラインメモリの間に境界バッファを設定しました。並列処理時に発生する重複処理が効果的に削減され、処理効率が向上することで、全体のパフォーマンスも改善します。

4点目は、演算を簡略化し、エッジ端末内の計算リソースを抑えるために搭載されている量子化ツールにおいて、小数点位置の変化に対応した固定小数点演算回路を開発しました。浮動小数点演算に近い精度でのデータ処理が可能となるため、高度なAI映像処理にも対応します。

これらの回路技術により、さまざまなAI映像処理アプリケーションにおいて、高精度かつ高速な演算を低消費電力で実現することが可能となりました。

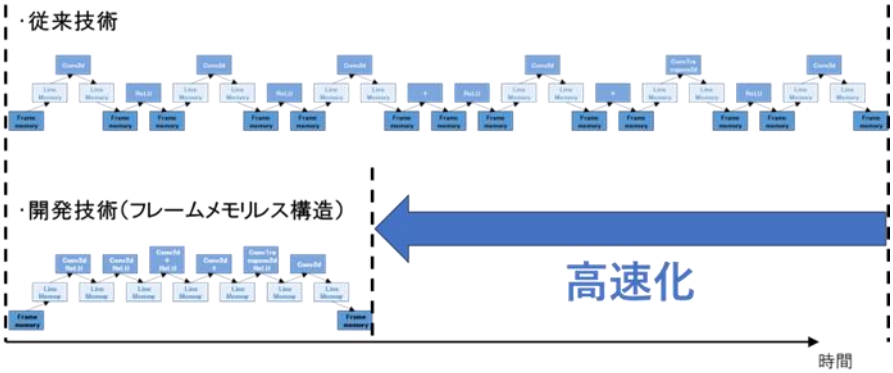

図4　フレームメモリレス構造による処理時間短縮のイメージ

本ツールを利用して、4K映像から8K映像への超解像処理を行うRTLコードを生成し、FPGAに実装したところ、1ワット（W）あたりの情報処理速度が、GPU搭載PCでは0.0079TOPS／W※9であったのに対し、0.374TOPS／Wの結果が得られ、電力効率※10が40倍以上に向上することを確認※11しました(図5)。また、生成期間も、専門技術者による開発では6週間かかるのに対し、本ツールでは約5分で完了※12しました。

	TOPS値	消費電力（実測）（W）	TOPS／W	電力効率
GPU搭載PC	2.90 （ = 3840 x 2160 x 22.5 x 7779 x 2 ）	365	0.0079	1.0 （基準）
高位合成ツール 生成回路 （20nmプロセスFPGA実装）	3.87 （ =1920 x 2160 x 60 x 7779 x 2 ）	10.35	0.374	47.1 （倍）

図5　CNN超解像※13モデルにおける電力効率比較

シャープでは、今回開発した本ツールをOSSとして一般公開しました。AI映像データ処理を活かした製品やサービスの実用化を目的に、エッジ端末での実証を効率的に行いたい開発者など、さまざまなユーザーによる専用デバイスを利用したエッジコンピューティングの産業応用加速とそれに伴う消費電力の低減が期待されます。

3. 今後の予定

NEDOは、今後もエッジ領域で活用可能となる、高度なデータ処理を高速かつ効率的に実現するAI半導体およびシステムの開発を進めると共に、性能を最大限に発揮することができるチップ設計を短期間に実現する設計技術の開発を実施することで、高度なエッジコンピューティングの実現を目指します。

シャープは、今回一般公開した高位合成ツールの認知拡大に向けた情報発信を行うことで、大学や研究機関、企業での利用を促進し、医療や製造、インフラ保全など、さまざまな分野でのAI映像処理の普及拡大に取り組みます。

【注釈】

※1 GPU

Graphics Processing Unitの略。画像や映像データ処理に特化したプロセッサですが、近年はAIの深層学習や機械学習の分野でも利用されています。

※2 FPGA

Field-Programmable Gate Arrayの略。ユーザーが自由にプログラムできるプロセッサであり、RTLコードで、用途に合わせた設計が可能です。

※3 本事業

事 業 名：省エネAI半導体及びシステムに関する技術開発事業／AIエッジコンピューティングの産業応用加速のための設計技術開発／映像データリアルタイム処理用AIデバイス高位合成ツールの研究開発

事業期間：2022年度～2024年度

事業概要：省エネAI半導体及びシステムに関する技術開発事業 https://www.nedo.go.jp/activities/ZZJP_100254.html

※4 OSS

Open Source Softwareの略。無料または廉価で配布され、改変が認められたソフトウェアのことです。

※5 一般公開

OSS公開ページ https://corporate.jp.sharp/8k5g/8klab/ai-edge_report_202505.html

※6 RTLコード

Register Transfer Levelコードの略。ハードウェアの動作を記述したコードで、ハードウェア開発や回路のシミュレーションなどに利用されます。

※7 PyTorch

Python向け機械学習ライブラリのことです。効率的なプログラミングを目的に利用されています。

※8 ONNX

Open Neural Network Exchangeの略。機械学習のフォーマットの一つです。

※9 TOPS／W

Tera Operations Per Second Per Wattの略で、消費電力1Wあたりの演算処理量を表します。数値が大きいほど少ない消費電力で多くの演算処理を行っていることになるため、同じ計算量を処理する場合、消費電力が低くなります。

※10 電力効率

複数のものを比較するために、その中の一つを基準として、消費電力1Wあたりの情報処理量（TOPS／W）を規格化した相対値。単位は、基準に対する倍率となります。

※11 40倍以上に向上することを確認

AI処理では、パラメータ数7779でConv2Dが6層のネットワーク構造を持つAI超解像モデルを当社独自の実験環境で駆動し比較しています。

※12 約5分で完了

パラメータ数7779でConv2Dが6層のネットワーク構造を持つAI超解像モデルのRTLコード生成で比較しています。

※13 CNN超解像

Convolutional Neural Network超解像の略。画像や動画など視覚データの深層学習を利用した超解像処理のことです。