PRESS RELEASE

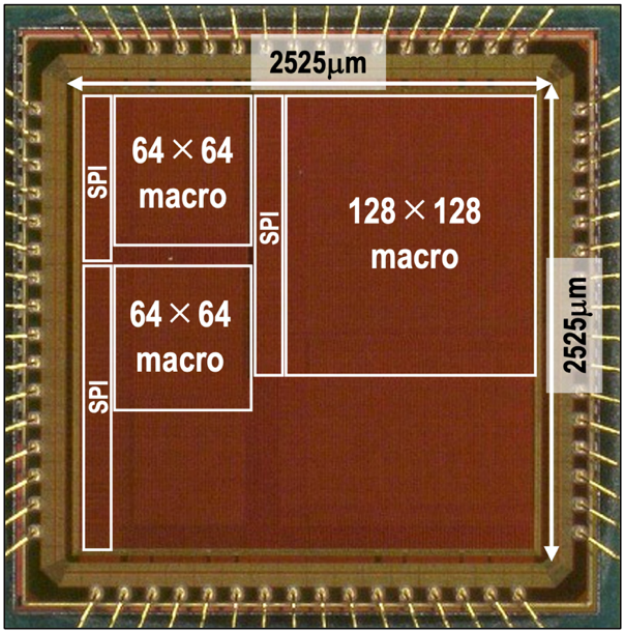
東京大学
科学技術振興機構（JST）

開発コストを 1/40 に削減する AI プロセッサの新方式を開発

――新規に必要なフォトマスクは 1 枚のみ、低コストと低電力動作を両立――

発表のポイント

- ◆ AI プロセッサの開発コストを 1/40 に削減する新方式半導体集積回路実装方法を開発。
- ◆ 配線 1 層のみのカスタマイズで、特定 AI 処理に応じた AI プロセッサを実現可能。
- ◆ 低コストと低電力性能を両立する世界で初めての方式であり、IoT や AR/VR 応用に好適。



新方式 AI プロセッサの 40nm プロセスでの試作品写真

発表概要

東京大学大学院工学系研究科の小菅敦丈 講師、Jaewon Shin（ジェウォン シン） 大学院生、濱田基嗣 特任教授らによる研究グループは、JST 戦略的創造研究推進事業 さきがけの助成のもと、低い開発コストと低電力性能を両立した新規ストラクチャード ASIC（注 1）型 AI プロセッサを開発しました。半導体製造において大部分を占めるフォトマスク（注 2）の開発コストを 1/40 に削減しつつ、既存の低電力性能に特化した AI プロセッサと同等の電力効率で処理できます。ウェアラブル IoT 応用におけるバイタル信号解析や音声認識に好適です。

スマートウォッチや AR/VR 機器において AI 機能を搭載することで、高度なバイタル解析による QOL の向上や機器操作性の向上によりユーザーエクスペリエンスの向上が期待されています。一方、こうした IoT 機器は小型軽量動作を追求するためバッテリ駆動であり、かつ安価であることが求められます。これまで低電力動作を追求するため AI プロセッサが世界中で研究開発されていますが、いずれもフォトマスク開発にかかる 10 億円単位の開発コストが高い障壁となり IoT デバイスへの採用が困難でした。低電力性能を追求するほどタスクに特化するた

め汎用性がなくなり、半導体の設計データであるフォトマスクを使い回すことができなくなります。フォトマスクの開発製造は非常に高額であり、この開発コストを回収するためにはチップ単価が極めて高額になることから､安価なデバイスの実現が難しいという問題がありました。

本研究では低電力動作と低コストを両立するため、ストラクチャード ASIC 方式の新規 AI プロセッサを開発しました（図 1）。演算回路と配線をあらかじめ実装したチップを上層配線の途中まで製造しておき、VIA（注 3）1 層のみで特定の AI 処理に応じた AI プロセッサ回路を構成するビアプログラマブルニューロンアレー（Via-programmable Neuron Array）技術を開発しました。AI プロセッサの製造に必要なフォトマスク枚数を“VIA 層”1 枚に減らし低コスト化を実現しました。実現にあたっての技術課題は巨大な実装面積です。深層ニューラルネットワーク（注 4）を布線論理方式（注 5）で実装するため、実装する信号配線が膨大になり広大なチップ面積が必要となっていました。従来方式では半導体集積回路として製造可能な限界面積を大幅に超過していたため、実現できませんでした。そこで研究チームは新たに回路と信号配線を時分割で再利用し回路面積を削減する、ビットニューロン順次回路技術を開発しました。さらに深層ニューラルネットワークの重み係数を 16 ビット（65,536 種類）から 3 値（+1、-1、0 の 3 種類）に削減しながらも精度を保つ、関数選択的非線形ニューラルネットワーク（Function-Selective Non-linear Neural Network、FS-NNN）技術（注 6）も開発しました。重み係数を 16 ビットから 3 値にすることで必要な信号配線本数を削減しています。これらの技術を組み合わせた結果、信号配線本数を 1/1024 に削減し省面積化を実現、10mm^2 以下と IoT 用途として十分小さな回路面積でストラクチャード ASIC による AI 機能実装に成功しました。ウェアラブル IoT 機器のみならず、ドローン、自動車内エンタメ機器制御、AR/VR 機器への応用が期待されます。

本研究成果は、2025 年 2 月 18 日（米国太平洋時間）に、半導体集積回路分野で世界最高峰の国際会議である International Solid-State Circuits Conference（ISSCC）にて口頭発表されます。

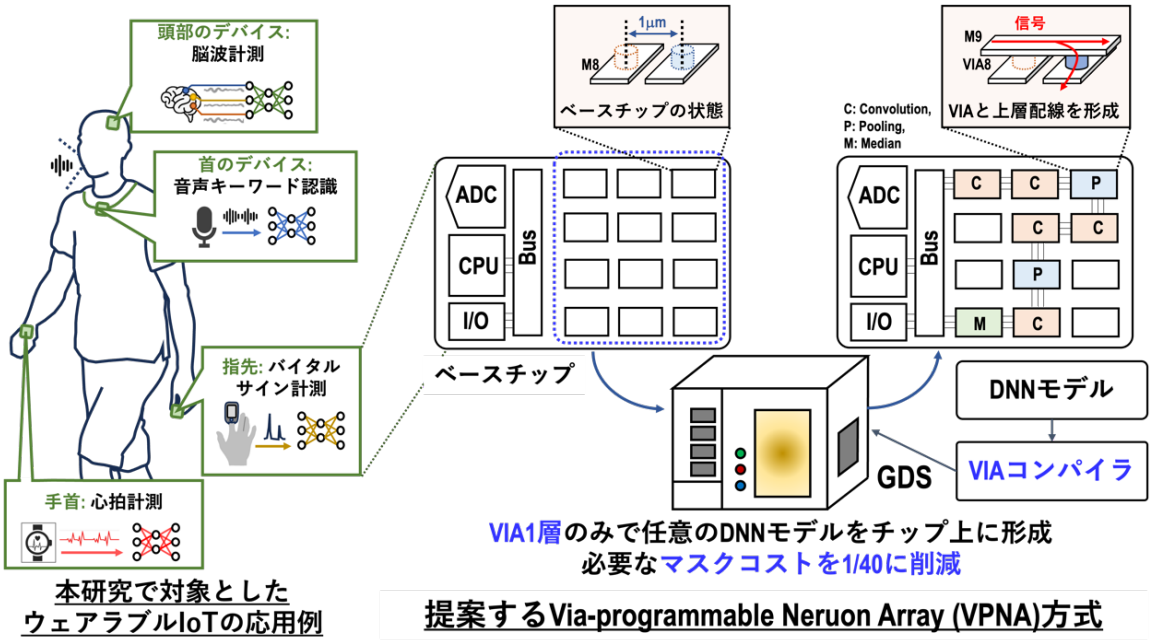


図 1：開発した新規 AI プロセッサの概要。低コストと低電力動作の両立に成功。
既存の低電力 AI プロセッサと同等の低電力動作を 1/40 のフォトマスク開発コストで実現。

発表内容

〈研究の背景〉

AI 技術は多くの産業に技術革新をもたらし、日常生活を変革すると期待されています。膨大な数のニューロンとシナプスを持つ深層ニューラルネットワークが技術の中核であり、シナプス接続を学習により最適化することでさまざまな能力を獲得しています。

IoT 用途においても AI を活用した新たなアプリケーションが日々研究されています。代表例がウェアラブル IoT で、常時バイタルサインを AI で解析しモニタリングすることで病気の早期発見につながることが研究で明らかにされています。AR/VR 機器では AI 機能を搭載し高機能なマシンインターフェースをユーザーに提供することで、より良いユーザーエクスペリエンスを実現できます。

IoT 用途における課題は低消費電力と低コストの両立です。IoT 機器は一般に、小型軽量動作を追求するためバッテリ駆動であり、かつ安価であることが求められます。これまでも低電力動作を追求するため AI プロセッサが世界中で研究開発されていますが、いずれも 100 億円単位の開発コストが生じるため IoT デバイスへの採用が困難でした。低電力性能を追求するほどタスクに特化するため汎用性がなくなり、開発コストを回収するためにはチップ単価が極めて高額になってしまうためです。本研究グループにおいても、人の大脳を真似た布線論理方式を採用した新規 AI プロセッサを開発してきました。2023 年には新たなマシンインターフェースとして期待される音声コマンド認識 AI を題材とし、高い精度、少ないチップ実装面積、低い消費電力すべてを同時に実現するため、新方式のアルゴリズム-回路協調最適化技術を開発しました。これにより、1 チップで 16 層の深層ニューラルネットワークを布線論理方式で実装することができ、152.8μW（マイクロワット）での推論に成功しました（関連情報：プレスリリース①）。また、当時の最先端 AI プロセッサと比較し、消費電力を 1/2552 以下に削減することができました。一方で、この技術には、音声コマンド認識に特化しており製造後は他用途に転用できないという弱点がありました。さらに、用途ごとに特化した AI プロセッサを開発するためには、高額なフォトマスクを全て再開発し直す必要があり、プロセッサ 1 チップあたりのコストが非常に高額になってしまうという問題がありました。

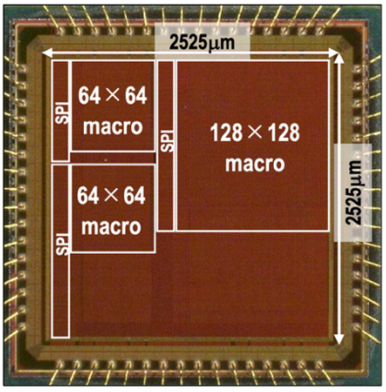
〈研究の内容〉

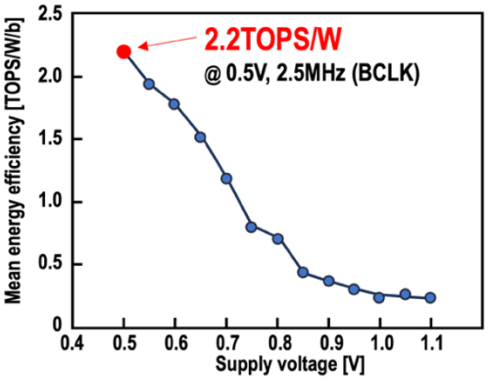
本研究では低電力動作と低コストを両立するため、ストラクチャード ASIC 方式を用いた新規 AI プロセッサを開発しました。演算回路と配線をあらかじめ実装したベースチップと呼ばれるチップを上層配線の途中まで製造しておき、上層配線のみで任意の AI モデル処理になるように回路を再構成します。ベースチップは汎用的に多数のタスクに転用できます。結果、このベースチップの開発コストは多くのユーザーとアプリケーションにて賄われるため、他の汎用プロセッサ同様低く抑えることが可能です。本研究では、VIA 配線 1 層のみで特定の AI 処理に応じた AI プロセッサ回路を構成する VIA-programmable Neuron Array 技術を開発しました。配線のカスタマイズを全て VIA 1 層のみで行うことで、AI プロセッサの製造に必要なフォトマスク枚数を“VIA”1 枚に減らし低コスト化を実現できます。従来の AI プロセッサでは数十枚のフォトマスクが必要であったところを 1 枚に減らすことができ、開発コストも同様に削減されます。

ストラクチャード ASIC は古くからある技術ですが、これまで AI 用途には適用されてきませんでした。膨大な配線が必要となり、チップ面積が製造不可能なほど巨大になってしまうためです。そこで新たに回路と信号配線を時分割で再利用し回路面積を削減する、ビットニューロン順次回路技術により信号線本数を 1/1024 に削減することで面積を削減し、10mm^2 以下と IoT

用途として十分小さな回路面積でストラクチャード ASIC による AI 機能実装に成功しました。さらに任意の深層ニューラルネットワークを入力として受け取り、VIA の配置情報に半自動で変換し半導体設計図面に仕立て上げる、VIA コンパイル技術も開発しました。設計エンジニアの工数もかからないため、開発コストをさらに削減することができます。また、VIA 個数をさらに削減するため、通常広範な正負の値を有する深層ニューラルネットワークの重み係数を 16 ビット（65,536 種類）から 3 値（+1、-1、0 の 3 種類）に削減する Function Selective Nonlinear Neural Network（FS-NNN）技術も新たに開発しました。正の重み係数はどのような値であっても+1 に、負の重み係数も同様に-1 に簡略化しています。そのまま簡略表現を適用すると AI の認識精度が劣化するため、再学習技術を適用しネットワーク構造と各ニューロンの非線形関数を最適化することで認識精度の劣化を防いでいます。

40nm（ナノメートル） CMOS プロセスにて 3mm×3mm のチップを試作したところ、0.5V 電源電圧にて深層ニューラルネットワーク全体で 2.2TOPS/W（消費電力 1W あたりの処理速度）の電力効率を確認しました（図 2）。1mW（ミリワット）という微小電力で、ウェアラブル IoT 応用で求められる脳波解析、心電図解析、音声認識などの AI タスクを処理することができます。脳波解析タスクで比較したところ、従来 AI プロセッサ（ISSCC’23）（注 7）と同程度の高い電力効率を実現しながら、開発コストを 1/40 に削減することに成功しました。また再利用可能で低コストな半導体である FPGA（注 8）と比較しても、8.4 倍高い電力効率を実現しています。





	FPGA [TBioCAS’19]	ASIC [ISSCC’23]	本成果
フォトマスク枚数	-	40枚 (1)	1枚 (1/40)
脳波解析AI 電力効率	108.1 µJ/inf. (1)	23.2 µJ/inf. (1/4.6)	12.8 µJ/inf. (1/8.4)

図 2：試作チップ写真（左上）と性能評価結果（右上、下）

最先端 AI プロセッサ(ISSCC’23)と比較し、同程度の高い電力効率と 1/40 以下のフォトマスク開発コストを実現。

〈今後の展望〉

開発した新方式の AI プロセッサは低コストと低電力動作の両立だけでなく、VIA コンパイラ技術と組み合わせることで、Python などの高位プログラミング言語から、短い設計期間で AI プロセッサの製造図面にまで変換できることが特徴です。短期間に機能更新を繰り返す AI アプリケーションに最適といえます。今後、マシンビジョン、設備点検自動化、物流倉庫、無人店舗など、多くのエッジ AI アプリケーションへ展開することを目指しています。

〇関連情報：

「記事①：システムデザイン研究センター　小菅敦丈 講師が「MIT Technology Review Japan Innovators Under 35」を受賞されました」(2021/12/22)
https://www.t.u-tokyo.ac.jp/topics/foe/topics/setnws_20211221112322095760 1664.html

「記事②：若手研究者紹介：小菅 敦丈 講師」(2023/5/2)
https://www.t.u-tokyo.ac.jp/topics/tp2023-05-08-069

「プレスリリース①：音声コマンド認識 AI の電力を 3 桁削減、新方式 AI プロセッサを開発 ―乾電池 1 本で 2 年以上連続動作、ドローンやロボットへの応用に期待―」(2023/6/9)
https://www.t.u-tokyo.ac.jp/press/pr2023-06-09-002

発表者

東京大学大学院工学系研究科附属システムデザイン研究センター
小菅　敦丈　講師
濱田　基嗣　特任教授
Jaewon Shin（ジェウォン　シン）　博士課程
澄川　玲維　修士課程：研究当時
Dong Zhu Li（ドンジュ　リ）　修士課程：研究当時

発表学会

学会名：International Solid-State Circuits Conference (ISSCC)
会　期：2025 年 2 月 16 日～20 日
（論文配布は 2 月 15 日、発表は 2 月 18 日 11:50-12:05。いずれも米国太平洋時間。）
題　名：A Via-Programmable DNN-Processor Fabrication Toward 1/40th Mask Cost
著者名：Jaewon Shin, Rei Sumikawa, Dong Zhu Li, Mototsugu Hamada, Atsutake Kosuge*

研究助成

本研究成果は、主として、以下の事業・研究領域・研究課題によって得られました。
科学技術振興機構(JST)戦略的創造研究推進事業 個人型研究(さきがけ)(課題番号:JPMJPR21B4)
研究領域：「情報担体とその集積のための材料・デバイス・システム」(研究総括：若林 整　東京科学大学 総合研究院　教授)
研究課題：「デバイス・システム協調による超低電圧布線論理型 AI プロセッサ」
研究代表者：小菅 敦丈（東京大学 大学院工学系研究科　講師）

用語解説

（注 1）ストラクチャード ASIC：ユーザーが任意の回路機能を最小限の上層配線で実現するため、メモリやプロセッサ、アナログデジタル変換器やクロック発生回路、入出力インターフェースなどの汎用機能をあらかじめ組み込んだ下地半導体基板に対して、さらにユーザーの設計を反映するゲートと呼ばれる基本素子回路を多数敷き詰めたゲートアレーを備えたもの。ゲートに対して配線接続を切り替えることで、任意の機能を有する回路を実装できる。フルカスタムで設計製造する ASIC に比べると設計可能な回路に制約があり実装面積も大きくなるが、コストも設計製造期間も短期間で済む。

（注 2）フォトマスク：半導体製造工程で使用されるもので、半導体設計図面を半導体ウエハーに転写するための原版。微細プロセスノードになる程使用するマスク枚数が増え、さらにマスク 1 枚あたりの製造コストも高くなる。

（注 3）VIA：半導体集積回路における多層メタル配線をつなぐ層間配線のこと。

（注 4）深層ニューラルネットワーク：脳の仕組みを模した AI モデルの 1 つであり、多数のニューロンとシナプスからなる層を多段に重ね、シナプスの係数を計算により最適化することで所望の認知機能を獲得する。

（注 5）布線論理方式：演算器同士を物理的に結線し、結線を組み替えることで、プログラムの命令を実行する方式。汎用プロセッサと異なり、命令や各種データのメモリへの格納が原則不要であり、高速かつ低消費電力であるという特徴がある。

（注 6）関数選択的非線形ニューラルネットワーク（Function-Selective Non-linear Neural Network、FS-NNN）技術：本研究で開発した新たなニューラルネットワーク技術であり、重み係数を+1、-1、0 の 3 種類のみしか使用しない代わりに、各ニューロンにあらかじめ決めていた 4 種類の非線形関数から選択的に割り当てることで精度劣化を防いでいる。ニューラルネットワークの学習によってどの非線形関数を採用するかが決定される。

（注 7）ISSCC：International Solid-State Circuits Conference の略称であり、米国電気電子学会 固体回路素子分科会（IEEE Solid-State Circuit Society）が主催する最高峰のフラグシップ学会である。ここでは先行研究の以下の論文を指す。
C.Tsai *et al.*, ″SciCNN: A 0-Shot-Retraining Patient-Independent Epilepsy-Tracking SoC,″ in ISSCC, pp. 488-490, Feb. 2023

（注 8）FPGA：製造後に購入者や設計者が構成を設定できるようにした集積回路であり、多数の再構成可能ロジックと再構成可能信号配線からなる。ここでは FPGA を用いた AI プロセッサに関する以下の先行研究を指す。
H. Elhosary *et al.*, ″Low-Power Hardware Implementation of a Support Vector Machine Training and Classification for Neural Seizure Detection,″ in *IEEE TBioCAS,* vol. 13, no. 6, pp. 1324-1337, Dec. 2019.

問合せ先

（研究内容については発表者にお問合せください）

東京大学 大学院工学系研究科
講師　小菅　敦丈（こすげ　あつたけ）

〈報道に関すること〉
東京大学 大学院工学系研究科 広報室

科学技術振興機構 広報課

〈JST 事業に関すること〉
科学技術振興機構 戦略研究推進部 グリーンイノベーショングループ
安藤　裕輔（あんどう　ゆうすけ）