エッジAIアクセラレータ、最小限の電力で60 TOPSを実現

2024年7月3日、 Gina Roosが 投稿

https://www.electronicproducts.com/edge-ai-accelerator-delivers-60-tops-at-minimal-power/

EdgeCortix のエネルギー効率に優れたエッジ AI アクセラレータ 「SAKURA-II」は、LLM や LVM を含む最新の生成系AI タスクをエッジで処理

スタートアップ企業であるEdgeCortix株式会社は最近、エッジで生成系AI (GenAI) ワークロードを最小限の消費電力と低遅延で処理するように設計された次世代の「SAKURA-IIエッジAIアクセラレータ」を発表しました。このプラットフォームは、同社の第二世代の「Dynamic Neural Accelerator (DNA)」アーキテクチャと組み合わさることにより、業界で最も困難とされている生成系AIタスクに対応できます。

日本に本社を置くEdgeCortixは、2022年に第一世代のAIアクセラレータであるSAKURA-Iを発表し、同社によると、SAKURA-IIはリアルタイムエッジアプリケーション向けのGPUベースの競合AI推論ソリューションと比較して、１ワット当たりのパフォーマンスが10倍以上優れているとのことです。また、MERAコンパイラソフトウェアフレームワークのオープンソースのリリースも発表しました。

同社は、第一世代のシリコンを使用している顧客からのフィードバックを活かして、スパース計算、高度な電力管理、混合精度のサポート、新しいリシェイパー・エンジンなどの新機能を追加し、活性化関数や変化するAIモデル環境への対応という点で、将来を見据えた最新のアーキテクチャを提供します。また、複数のフォームファクターで１ワットあたりのパフォーマンスが向上するだけでなく、メモリ容量も増加しました。

SAKURA-II エッジAI アクセラレータ (出典: EdgeCortix株式会社)

EdgeCortixのCEO兼創業者であるサキャシンガ・ダスグプタ氏は、以下のように述べています。

「SAKURA-IIアクセラレータまたはコプロセッサは、標準消費電力8Wで毎秒60TOPSの性能や混合精度のサポート、内蔵メモリ圧縮機能を備えた魅力的なソリューションです。従来のAIモデルを実行する場合でも、エッジで最新の生成系AIソリューションを実行する場合でも、このアクセラレータは最も柔軟で電力効率に優れたアクセラレータの1つです。」

SAKURA-II は、製造、インダストリー 4.0、セキュリティ、ロボティックス、航空宇宙、通信業界における大規模言語モデル（LLM）、大規模ビジョンモデル（ LVM）、マルチモーダル・トランスフォーマーベースのアプリケーションなどの複雑なタスクを処

理できます。Llama 2、Stable Diffusion、DETR、ViTなどの数十億のパラメータ・モデルを、標準的な8Wの電力エンベロープ内で管理することもできます。

一言で言えば、SAKURA-II　アクセラレータは、生成系AIと低レイテンシのリアルタイムデータストリーミングに最適化されています。優れたエネルギー効率（他のソリューションの2倍以上のAIコンピュート利用率と言われています）複雑なビジョンや生成系AIワークロードを処理するための最大32GBの大容量DRAM、他のAIアクセラレータよりも最大4倍広いDRAM帯域幅、超高効率モードのための高度な電力管理を実現します。また、メモリ帯域幅を削減するスパース計算、ホストCPUの負荷を最小限に抑える新しい統合テンソルリシェイパーエンジン、任意の活性化関数、F32に近い精度を実現するソフトウェア対応の混合精度も追加されています。

EdgeCortixのソリューションは、より多くのAI処理をデータ作成の現場に移すことで、データ転送のコスト、電力、時間を削減することを目的としています。エッジAIアクセラレーター・プラットフォームは、「最新世代のAIモデルの飛躍的な成長」による2つの大きな課題に対処しているとダスグプタ氏は述べています。最初の課題は、これらの「急激に成長するモデル」による計算需要の増加と、それに伴うハードウェアコストの上昇です。

スマートシティ、ロボティックス、エッジ環境における航空宇宙産業など、どの分野であっても、ソリューションの導入コストや運用コストは極めて重要であると彼は付け加えました。

第二の課題は、より電力効率の高いシステムを構築する方法です。ダスグプタ氏は以下のように述べています。「残念ながら、今日のAIモデルの大部分は、電力消費と二酸化炭素排出量の両面で、はるかに多くの電力を消費しています。では、電力、重量、サイ

ズに制約のあるエッジ環境で、よりエネルギー効率の高いソフトウェアとハードウェアの組み合わせでシステムを構築するにはどうすればいいのでしょうか？それが私たちの企業としての原動力となります。」

ダスグプタ氏によると、同社の主な使命は、エッジ環境にクラウドレベルに近いパフォーマンスをもたらすと同時に、桁違いの電力効率を実現するソリューションを提供することだといいます。

ダスグプタ氏は、顧客にとってワット当たりのパフォーマンスは重要な要素であり、特にエッジ環境ではリアルタイム処理が重要な要素になると付け加えました。

エッジでのデータ

データセンターとエッジの状況を見ると、消費されるデータの大部分、特にエンタープライズデータはエッジで生成または処理されており、この傾向は今後も続くだろうとダスグプタ氏は述べました。

IDCによると、2025年までにエッジで生成されるデータは74ゼタバイトに達すると予想されています。この膨大な量のデータをエッジからクラウドに継続的に移動するには、電力と時間の両面でコストがかかります、と同氏は付け加えました。「AIの基本的な考え方は、データが作成される場所にいかに計算とインテリジェンスを持ち込むかということです。」

ダスグプタ氏によれば、EdgeCortixは、ソフトウェア・ファーストという設計理念と、データ転送コストを削減するための電力効率の高いハードウェアを組み合わせることで、これを実現したとのことです。

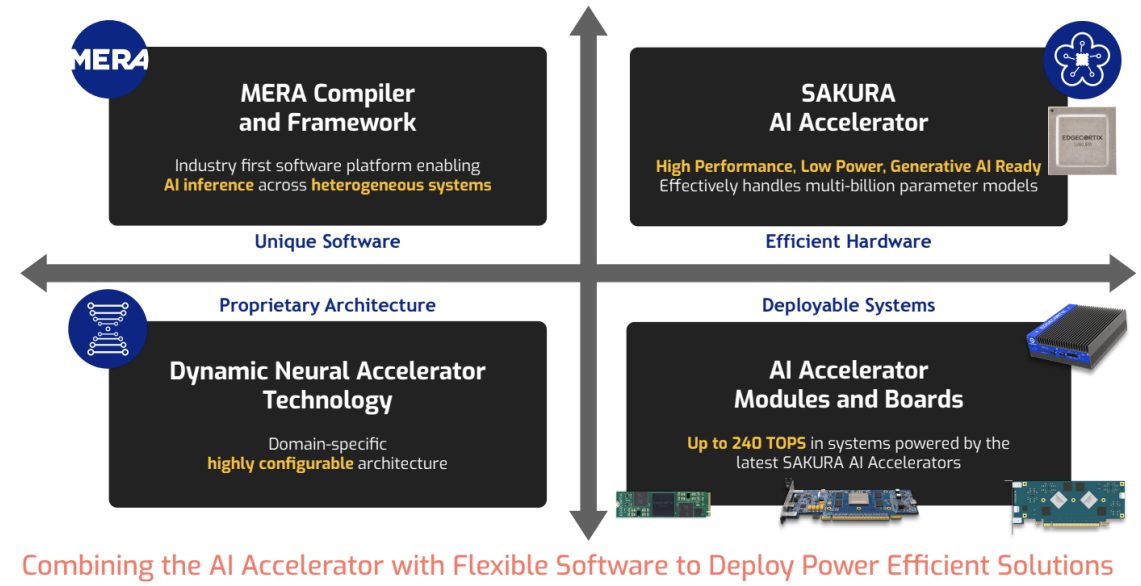
ダスグプタ氏は、この最新の製品は、アプリケーション全体のエッジにおける低電力要件という制約の中で、生成系AI環境だけでなく、数十億のパラメータを持つLLMやビジョンモデルにも対応していると述べました。最新の低消費電力生成系AIソリューションは、スマートシティ、スマートリテール、通信から、ロボティックス、工場現場、自動運転車、さらには軍事／航空宇宙に至るまで、さまざまな業界をターゲットにしています。

プラットフォーム

ソフトウェア駆動型の統合プラットフォームは、SAKURA　AIアクセラレータ、MERAコンパイラとフレームワーク、DNAテクノロジー、AIアクセラレータモジュールとボードで構成されており、最新の生成系AIと畳み込みモデルの両方をサポートしています。

柔軟性と電力効率を重視して設計されたSAKURA-IIは、高いメモリ帯域幅、高精度、コンパクトなフォームファクタを提供します。EdgeCortixの最新世代のランタイムで再構成可能なニューラル処理エンジンであるDNA-IIを活用することで、SAKURA-IIは、低レイテンシで複数のディープ・ニューラル・ネットワーク・モデルを同時に実行しながら、高い電力効率とリアルタイム処理能力を提供することができます。

ダスグプタ氏は、レイテンシはAIモデルやアプリケーションによって異なるため、数値で示すのは難しいと述べています。ほとんどのアプリケーションは、より大きなモデルにもよりますが、10ミリ秒未満になり、場合によっては、1ミリ秒未満になることもあります。

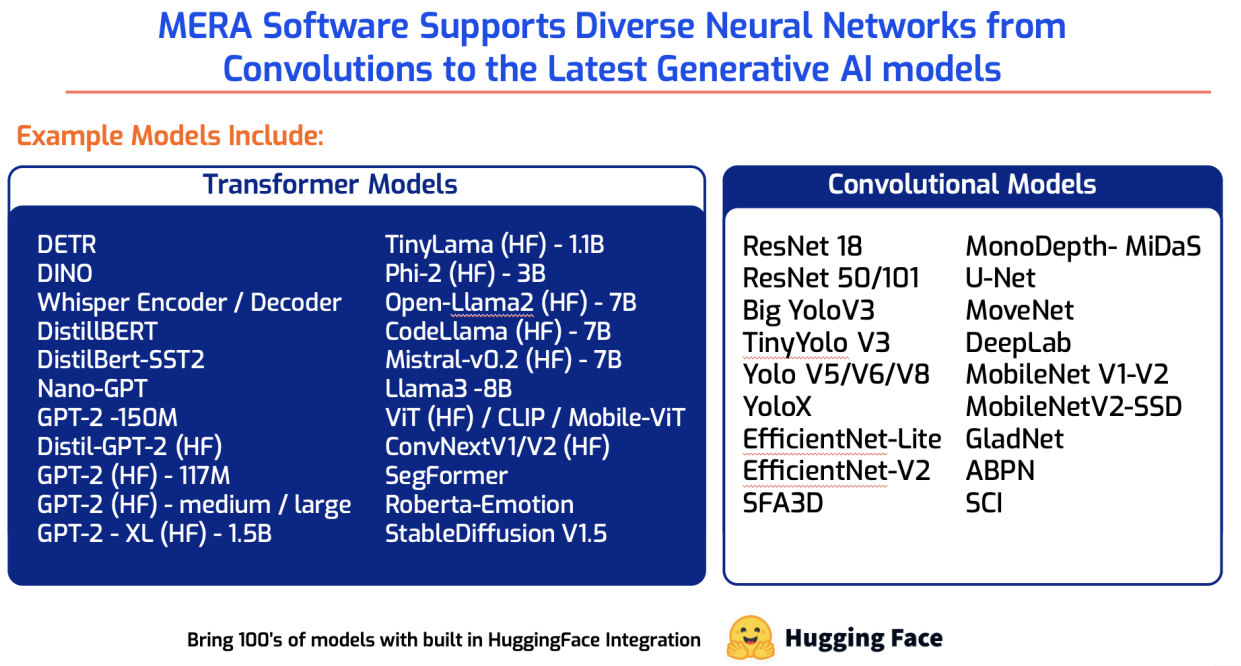


SAKURA-II ハードウェアおよびソフトウェア プラットフォームは、柔軟性、拡張性、電力効率を実現します。(出典 : EdgeCortix Inc.)

SAKURA-II プラットフォームは、同社のMERA ソフトウェア スイートを備え、異種コンパイラプラットフォーム、高度な量子化およびモデルキャリブレーション機能を備えています。このソフトウェア スイートには、PyTorch、TensorFlow Lite、ONNX などの開発フレームワークのネイティブサポートが含まれています。MERAの柔軟なホストからアクセラレータまでの統合ランタイムは、エッジにおいて、シングル、マルチチップ、マルチカードシステムにわたって拡張できます。これにより、AI推論が大幅に効率化され、導入時間が短縮されます。

さらに、MERA Model Libraryとの統合により、Hugging Face Optimum へのシームレスなインターフェイスが提供され、ユーザーは最新のトランスフォーマーモデルの広い範囲にアクセスできるようになります。これにより、トレーニングからエッジ推論へのスムーズな移行が保証されます。

「魅力的な要素のひとつは、Hugging Face との直接的なインターフェイスを作成する新しい MERA モデルライブラリです。これにより、お客様は互換性を心配することなく、多数の最新世代のトランスモデルを持ち込むことができます。」とダスグプタ氏は述べています。



MERA ソフトウェアは、畳み込みから最新の 生成系AI モデルまで、多様なニューラル ネットワークをサポートしています。(出典: EdgeCortix株式会社)

ダスグプタ氏は、新たなイノベーションを取り入れたソフトウェアファースト（ソフトウェア 2.0）アーキテクチャを構築することで、アプリケーションに応じてCPUやGPUを使用する汎用システムと比較して、１ワットあたりのピーク性能を１桁または２桁向上させることができたと述べています。

ダスグプタ氏は以下のように述べています。

「エッジの制約された環境内のアプリケーションに対して、高い精度（FP32の 99%）を維持できます。特に最新の生成系AIモデルを使用したマルチモーダル型アプリケーシ

ョンによって駆動される、遅延に敏感でリアルタイム性が重要なアプリケーションに関しては、ワットあたりの性能が大幅に向上するだけでなく、効率も向上し、はるかに高速になります。そして最後に、1ドルあたりのパフォーマンスという点で、他の競合ソリューションと比較して大きなアドバンテージを維持しながら、運用コストをより低く抑えることができます。」

これがエッジAIアクセラレーターの要件となり、最新のSAKURA製品の設計に反映されています、とダスグプタ氏は述べています。

SAKURA-IIの詳細

SAKURA-II は、８ビット整数 (INT8) で最大60TOPS のパフォーマンスと、１秒あたり30兆回の16ビット浮動小数点演算 (TFLOPS) を実現できるほか、次世代 AIタスクを処理するための組み込みの混合精度もサポートしています。より広いDRAM帯域幅は、LLMやLVMの性能に対するより高い需要に対応します。

SAKURA-IIの新機能には、超高効率モード用のオンチップ・パワーゲーティングやパワーマネジメント機能、複雑なデータの並び替えをオンチップで管理し、ホストCPUの負荷を最小限に抑えて効率的なデータ処理を実現する専用のテンソルリシェイパーエンジンなど、高度な電力管理機能も含まれます。

ダスグプタ氏によれば、主要なアーキテクチャの革新には、大量のモデル、特に数十億パラメータのLLMを動かすのに必要なメモリ量を削減するために、メモリフットプリントの最適化をネイティブにサポートするスパース計算が含まれ、これは性能と消費電力に大きな影響を与えるとのことです。

内蔵された高度な電力管理メカニズムにより、アプリケーションの実行中にデバイスのさまざまな部分をオフにし、電力と性能のバランスをとることができるため、60TOPSすべてを必要としないアプリケーションやモデルではワットあたりの性能が向上すると、ダスグプタ氏は述べています。

「我々はまた、ハードウェア自体に新しいリシェイパー・エンジンという形で専用IPを追加しました。これは大規模なテンソル演算を処理できるように設計されています」とダスグプタ氏。この専用エンジンをオンチップに搭載することで、消費電力が改善するとともに、レイテンシがさらに短縮されると同氏は付け加えました。

ダスグプタ氏は、アクセラレータ・アーキテクチャがSAKURA-IIの性能の重要な構成要素であると述べました。「当社は、特にGPUの観点から、競合他社と比較して、半導体上のトランジスタの利用率がはるかに高くなっています。通常、平均して２倍以上の利用率で、ワットあたりの性能が大幅に向上しています。」

SAKURA-IIはまた、ハードウェア上の任意の活性化関数のサポートも追加しました。ダスグプタ氏はこれを将来を見据えたメカニズムと呼んでおり、新しいタイプの任意の活性化関数が登場しても、ハードウェアを変更することなくユーザーに拡張できます。

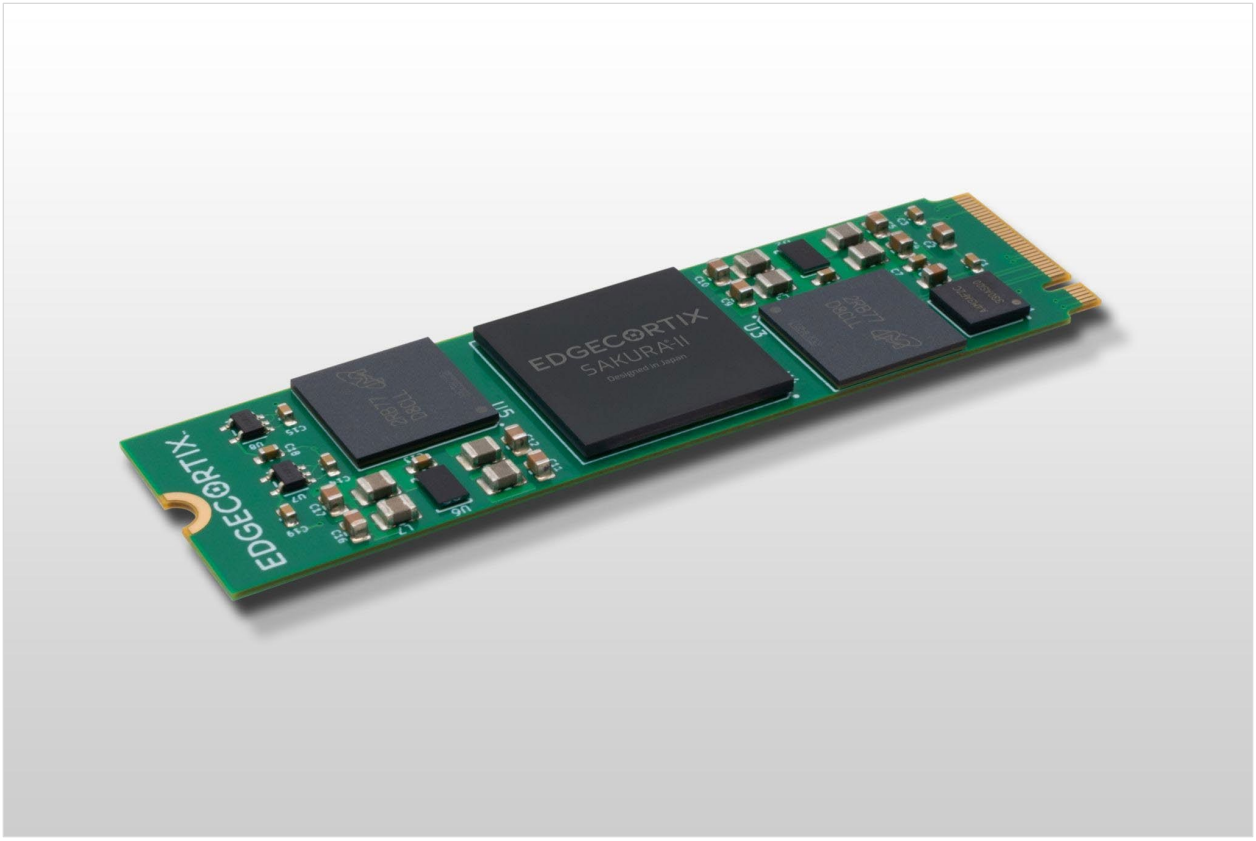
また、ソフトウェアとハードウェア上で混合精度のサポートも提供しており、パフォーマンスと精度のトレードオフが可能となっています。アプリケーションに応じて、モデルの一部を高精度で実行し、他の部分を低精度で実行することが、マルチモーダルの場合には重要になるとダスグプタ氏は述べました。

SAKURA-IIは、顧客のさまざまな要件を満たすために、複数のフォームファクタを用意しています。これには、19×19mm BGAパッケージのスタンドアロンデバイス、単一

デバイスを備えたM.2モジュール、および、最大4つのデバイスを備えたPCIeカードが含まれます。M.2モジュールは８GBまたは16GBのDRAMを搭載し、スペースに制約のあるアプリケーション向けに設計されています。一方、シングル (16GB DRAM) およびデュアル (32GB DRAM) PCIeカードは、エッジ サーバーアプリケーションを対象としています。

SAKURA-IIは、M.2モジュールのフォームファクターで、スペースに制約のある環境に対応し、x86とArmの両方のシステムをサポートし、従来のビジョンモデルだけでなく、数十億パラメータモデルをサポートすることで性能を発揮する、とダスグプタ氏は述べました。

「最新世代の製品は、非常に強力なコンパイラとソフトウェア・スタックをサポートしており、当社のコプロセッサをエッジ間におけるさまざまな種類の異種環境で動作する既存のx86またはArmシステムと組み合わせることができます。」



SAKURA-II M.2 モジュール (出典: EdgeCortix株式会社)

統合プラットフォームは、消費電力が 50 W 未満の 4 つのデバイスを備えた単一のPCIe カードで、最大 240 TOPS の高コンピューティング性能も提供します。

ダスグプタ氏は、SAKURA-IIでは電力が以前のレベルに維持されているため、顧客は1ワットあたりはるかに高い性能を実現することができると述べました。消費電力は、最も複雑なAIモデルでも通常約８Wで、一部のアプリケーションではさらに低くなると同氏は述べました。

SAKURA-IIは、スタンドアロンデバイスとして、DRAM容量の異なる2種類のM.2モジュール（8GBと16GB）、およびシングルデバイスとデュアルデバイスのロープロファイルPCIeカードを提供する予定です。お客様は、2024年後半の出荷に向けてM.2モジ

ュールとPCIeカードを予約できます。アクセラレータ、M.2モジュール、PCIeカードは、事前注文が可能です。