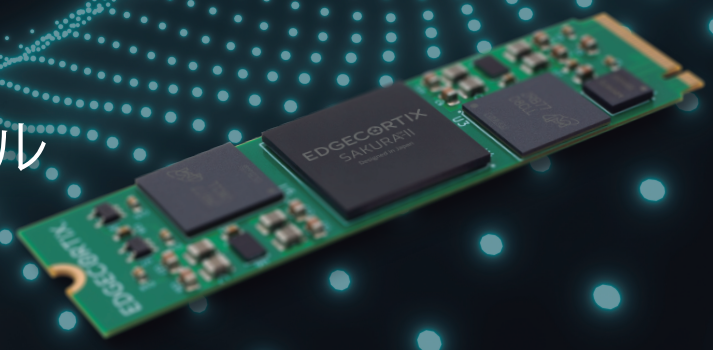
EDGECORTIX®
SAKURA™-II M.2 モジュール

Energy-Efficient Edge AI:
Vision to Generative AI



高性能で小型フォームファクターのエッジAI推論

SAKURA-IIM.2モジュールは、60TOPSの高性能エッジAIアクセラレータで、最新のビジョンおよび生成AIモデルを、業界をリードするエネルギー効率と低遅延で実行できるように設計されています。

EdgeCortixのMERAコンパイラとソフトウェアフレームワークは、アプリケーションに依存しない方法で、最新のAI推論モデルを迅速かつ容易に展開するための堅牢なプラットフォームを提供します。

主な利点

小型フォームファクター：M.2はスペースに制約のある設計に最適

生成AIに最適：標準的な10Wの電力エンベロープ以内で、Llama 2、Stable Diffusion、DETR、ViTのような数十億のパラメータモデルをサポート

メモリ帯域幅の強化：競合するAIアクセラレータと比較して最大4倍のDRAM帯域幅を確保し、LLMとLVMの優れたパフォーマンスを保証

効率的なAI演算：他のソリューションと比較して2倍以上のAI演算利用率を達成し、卓越したエネルギー効率を実現

大容量DRAM：最大16GBのDRAMをサポートし、複雑なビジョンや生成AIのワークロードを効率的に処理

リアルタイム データ ストリーミング：バッチサイズ 1 で低遅延オペレーションに最適化

任意の活性化関数のサポート：専用のハードウェア機能による近似の関数が適応性を向上

高度な精度：ソフトウェア対応の混合精度でFP32に近い精度を実現

効率的なデータ処理：統合されたテンソル変換処理機能により、ホストのCPU負荷を最小限

スパース計算：メモリ使用量を削減し、DRAM帯域幅を最適化

電力管理：高度な電力管理で超高効率モードを実現

技術仕様

性能[1]
60 TOPS (INT8)
30 TFLOPS (BF16)

推論
PCI Gen 3.0 x4

DRAM 帯域
68 GB/sec

消費電力
10W (typical)

モジュールの高さ
D6 (3.2mm top, 1.5mm bottom)

温度範囲
-20C to 85C

フォームファクタ
M.2 2280 Key M

オンボードDRAM
16GB (2 banks of 8GB LPDDR4X)
8GB (2 banks of 4GB LPDDR4X)

注釈1. High utilization TOPS

MERA Fast and Easy Model Porting and System Integration

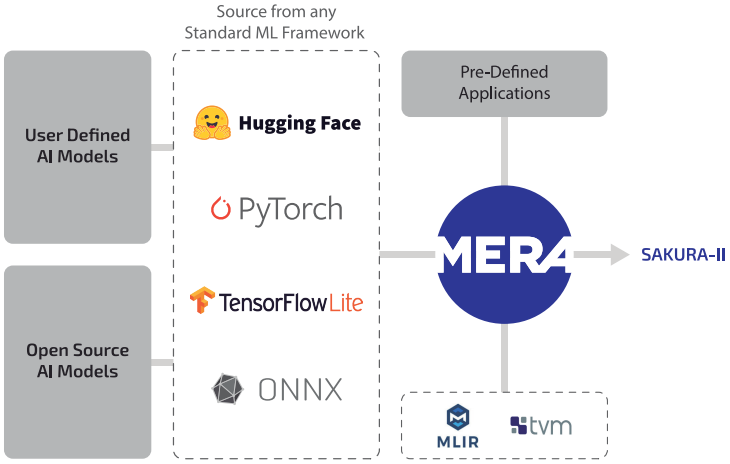
MERAは、使い慣れたニューラルネットワークモデルのワークフローでモデリングから展開まで、エッジAI推論のためのスタック全体を提供し、既存システムとの容易な統合をサポートすることで、市場投入までの時間を短縮します。

MERAツール

- Hugging Face、PyTorch、TensorFlow Lite、またはONNXを使用したソースモデル
- PythonまたはC++を使用して設計を統合し、カスタマイズ
- MERAフロントエンドはApache TVMとMLIRをサポートするオープンソース

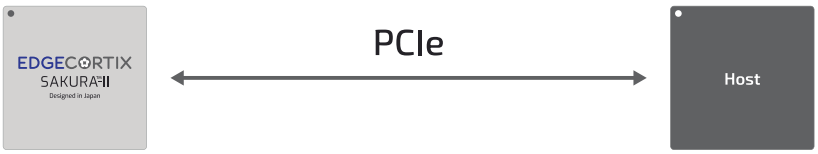
モデルリソース

- Model Zoo：事前にトレーニングされ、最適化されたAI推論モデル
- Llama-2、Stable Diffusion、Whisper、DETR、DistillBert、DINO、ViTなどの一般的な生成AIモデルをサポート
- トレーニング後のモデルのキャリブレーションと量子化



MERAコンパイラ とソフトウェアフレームワーク

MERA Model Library	Hugging Face Optimum Interface	Your Custom Model	
MERA Compiler User APIs	MERA Quantizer API	Model Compilation API	Model Interface API
MERA Frontend Interface	PyTorch	TensorFlow Lite	ONNX
MERA Core	Graph Optimization and Standardization	High-level Graph Partitioning	Calibration Quantization Mixed Precision
MERA Backend	SAKURA-II / DNA-II Code Generator	Host CPU (x86, Arm) Apache TVM, LLVM	MCUs C99 Code Generator CMSIS-NN Support
Unified Runtime	SAKURA-II / DNA-II Runtime		CPU (x86, Arm) Runtime



M.2モジュールの事前予約はこちら

edgecortix.com/ja/pre-order-sakura