EDGECORTIX®
SAKURA™ AI アクセラレータ

Energy-Efficient Edge AI:
Vision to Generative AI

高性能かつ省電力のAI推論

SAKURA-IIは、業界をリードするエネルギー効率と低遅延で最新のビジョンおよび生成系AIモデルを実行するために設計された60TOPSの高性能エッジAIアクセラレータです。

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。

SAKURA®-IIは、複数のフォームファクタで利用でき、柔軟なシステム統合、容易な評価、迅速な市場投入を可能にします。

主な利点

生成系AIに最適：標準的な8Wの電力エンベロープ以内で、Llama 2、Stable Diffusion、DETR、ViT等の数十億のパラメータの生成系AIモデルをサポート

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

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

大容量DRAM ： 最大32GBのDRAMをサポートし、複雑なビジョンや生成系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

SAKURA-II Offering

シリコンデバイス

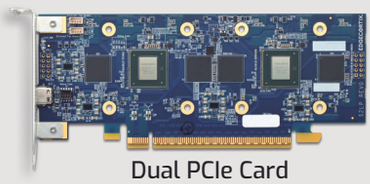
19 x 19 BGA

M.2 モジュール および PCIe カード

迅速な統合と市場投入までの時間を短縮するソリューション

M.2 2280 Key M Module

Single PCIe Card

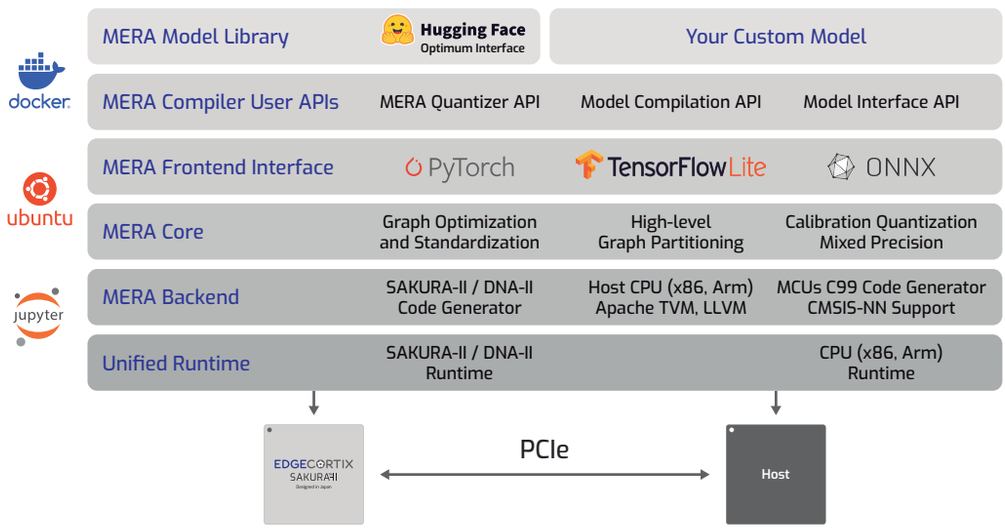


Dual PCIe Card

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コンパイラ とソフトウェアフレームワーク



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系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

技術仕様

性能[1]
60 TOPS (INT8)
30 TFLOPS (BF16)

DRAMサポート
Dual 64-bit LPDDR4X
(8/16/32GB total)

DRAM帯域幅
68 GB/sec

オンチップSRAM
20MB

演算効率
最大90%稼働率

温度範囲
-40C to 85C

消費電力
8W (typical)

パッケージ
19mm x 19mm BGA

注釈1. High utilization TOPS

SAKURA-IIの詳細はこちら

edgecortix.com/ja/sakura

SAKURA-II の主なマーケットセグメント

- 運輸／自動運転車
- 防衛／航空宇宙
- セキュリティ
- 5G 通信
- 拡張現実 (AR)と仮想現実 (VR)
- スマート・マニュファクチャリング／産業用ロボット
- スマートシティ
- スマートリテール
- ドローン＆ロボティクス