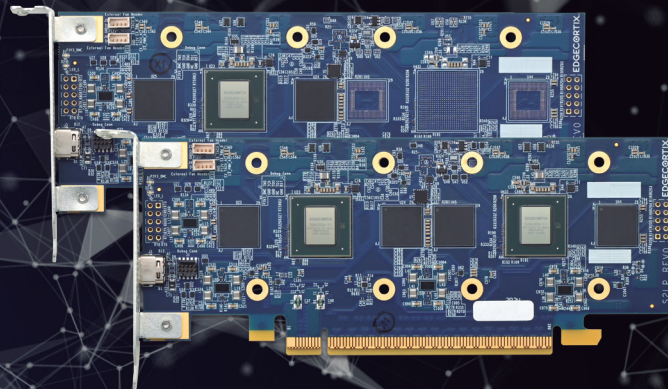




# EDGE CORTIX<sup>®</sup> SAKURA<sup>™</sup> PCIe カード

Energy-Efficient Edge AI:  
Vision to Generative AI



## エッジAI推論のための高性能PCIeフォームファクタ

SAKURA-II PCIeカードは、最大120TOPSの高性能エッジAIアクセラレータソリューションで、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モデルを実行できるように設計されています。

EdgeCortex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

**生成AIに最適**：標準的な10Wまたは20Wの電力エンベロップ以内で、Llama 2、Stable Diffusion、DETR、ViTのような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CPUの負荷を最小限

**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

**ロープロファイルPCIeカード**：高性能なエッジAIサーバーと製品にとって最適な選択

## 技術仕様

### フォームファクタ

ロープロファイル、シングルスロットPCIe (x16)

### 温度範囲

-20C to 85C

### DRAM帯域幅

68 GB/sec

### SINGLE SAKURA-II

#### 性能

60 TOPS (INT8)  
30 TFLOPS (BF16)

#### 消費電力

10W (typical)

#### 推論

PCI Gen 3.0 x8

#### オンボードDRAM

16GB (2 banks of 8GB LPDDR4)

### DUAL SAKURA-II

#### 性能

120 TOPS (INT8)  
60 TFLOPS (BF16)

#### 消費電力

20W (typical)

#### 推論

PCI Gen 3.0 x8/x8  
(bifurcated)

#### オンボードDRAM

32GB (4 banks of 8GB LPDDR4)



# Fast and Easy Model Porting and System Integration

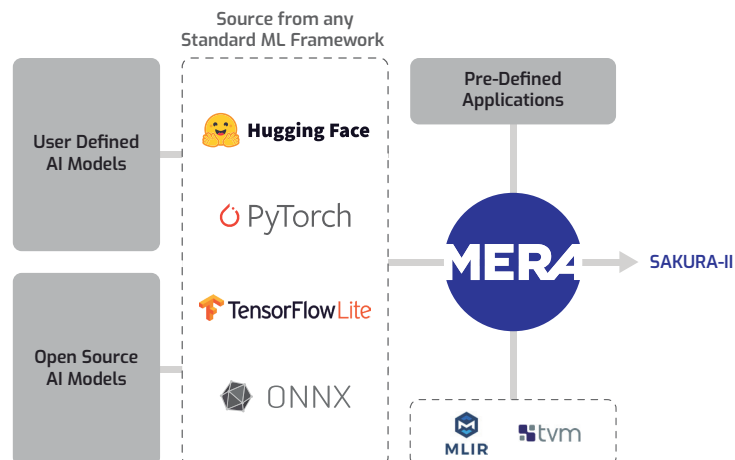
MERAは、使い慣れたニューラルネットワークモデルのワークフローでモデリングから展開まで、エッジAI推論のためのスタック全体を提供し、既存システムとの容易な統合をサポートすることで、市場投入までの時間を短縮します。

## MERAツール

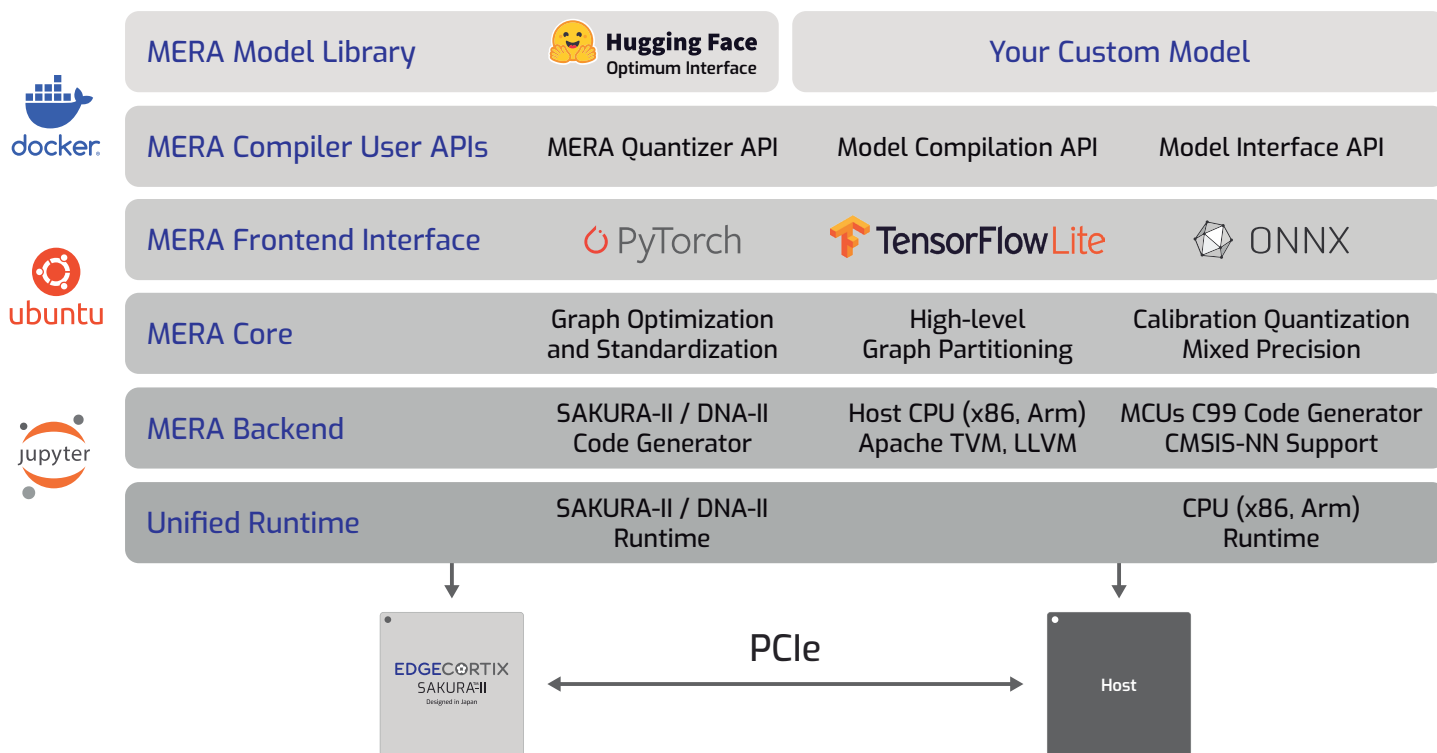
- Hugging Face、PyTorch、TensorFlow Lite、またはONNXを使用したソースモデル
- PythonまたはC++を使用して設計を統合し、カスタマイズ
- MERAフロントエンドはApache TVMとMLIRをサポートするオープンソース

## 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

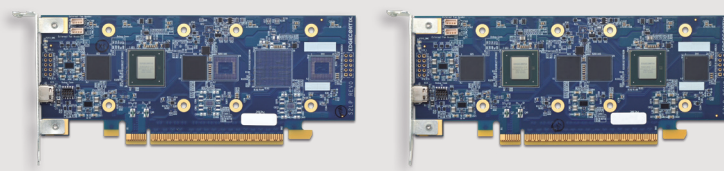


## MERAコンパイラとソフトウェアフレームワーク



PCIeカードの事前予約は  
こちら

[edgecortex.com/en/pre-order-sakura](https://edgecortex.com/en/pre-order-sakura)



© 2024 EdgeCortex Inc. All Rights Reserved. | EdgeCortex, Dynamic Neural Accelerator, and SAKURA are registered trademarks of EdgeCortex, Inc. All other products are the trademarks or registered trademarks of their respective holders. | May 2024: A4

