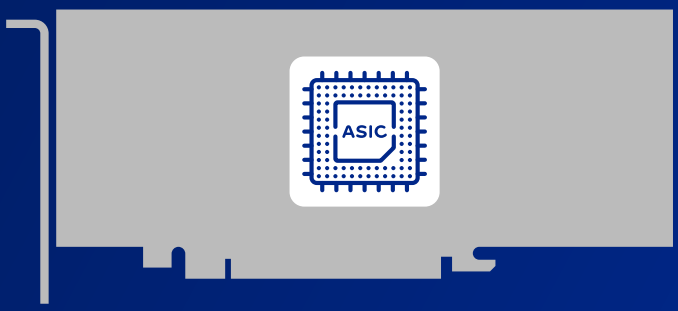
EDGECORTIX

Deep Learning Accelerator

SAKURA™
Dynamic Neural Accelerator
フレームワーク



TSMC 12nm FinFET
AI Performance of 40 TOPs @ 800 MHz

EdgeCortix® SAKURA: 高効率なエッジAIコプロセッサ

EdgeCortix SAKURAは、TSMC 12nm FinFETコプロセッサ（アクセラレータ）で、業界最高レベルの計算効率とレイテンシを実現します。SAKURAは、40TOPSのシングルコアである**Dynamic Neural Accelerator® (DNA)IP**という、全ての計算エンジンをつなぐ再構成可能なデータパスを内蔵した当社の独自のニューラルプロセッシングエンジンを搭載しています。DNAは新製品であるSAKURA　AIコプロセッサにおいて、優れたTOPSを維持しながら、超低レイテンシで複数のディープニューラルモデルを実行することが可能です。この独自の特性は、SoCの処理速度、エネルギー効率、寿命を強化する鍵となり、TCO(総保有コスト)においても大きな利益をもたらします。DNA IPは、特にストリーミングデータや高解像度データの推論に最適化されています。

SAKURAの特性が活かせる産業分野

自律走行車 / 輸送車・防衛とセキュリティ・5G通信
・VR/AR・スマート製造・スマートシティ・スマートリテール・ロボット工学

省電力かつ低レイテンシのAI推論を必要とするあらゆる分野に最適です。

KEY FEATURES

最大	INT8 推論	省電力
40 TOPS @ 800 MHz	(99% of FP32 精度)	PCI-E デバイス 10-15W TDP

PCIe x16 dev.
2022冬頃販売予定

ハードウェア・アーキテクチャの概要

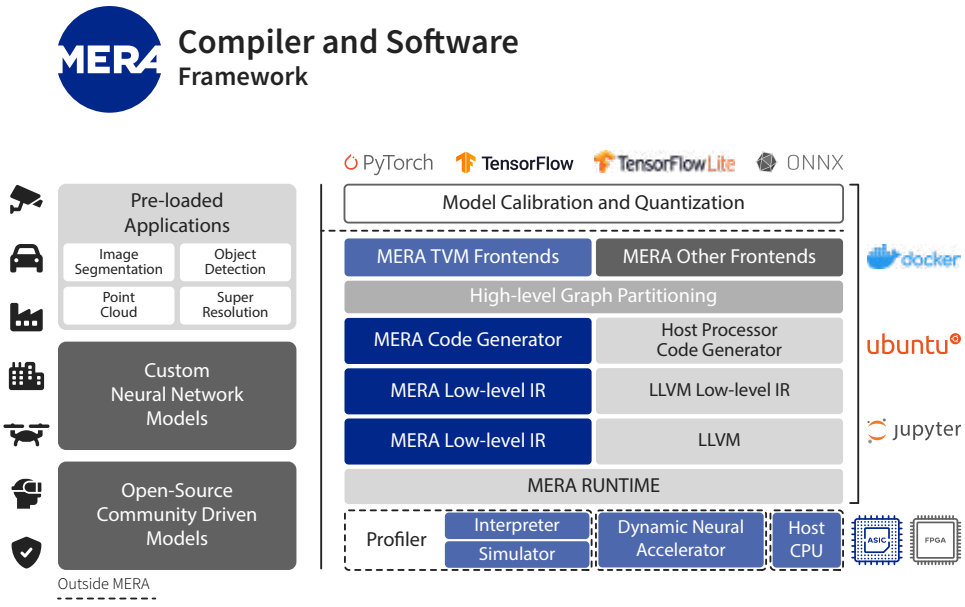
- 最大40TOPS (シングルチップ) / 200TOPS(マルチチップ)
- PCI-eデバイスのTDP @10W～15W
- 代表的なモデルの消費電力：～5W
- 2x64 LPDDR4x：16 GB
- PCIe Gen 3 ：最大16 GB/s の帯域幅
- 2つのフォームファクタ： デュアルM.2 / ロープロファイルPCIe
- ランタイムで再構成可能なデータパス

Dynamic Neural Accelerator® IP

- シングルコアで+24K MAC @ 800 MHz
- INT8とバッチサイズ1に最適化
- 大容量オンチップメモリ： 20 MB
- ソフトウェアで定義された複数の並列度を利用することで コンピューティング使用率を最大化
- Yolov3やYolov5、ポイントクラウド処理ベースのAIなど、 要求の厳しいワークロードで超低レイテンシ(4ms未満) を実現

商品説明

EdgeCortix SAKURA AIコプロセッサは、異種コンパイラや当社のソフトウェアフレームワークであるEdgeCortix MERAに対応しており、パブリックpipリポジトリからインストールすることで、業界標準のフレームワークで開発した標準またはカスタムCNN（畳み込みニューラルネットワーク）をシームレスにコンパイルおよび実行することが可能です。MERAは、Apache TVMを搭載しており、SAKURAのDNA AIエンジンを用いて、ディープニューラルネットワークのグラフコンパイルと推論をシームレスに行うためにシンプルなAPIを提供します。



また、プロファイリングツール、コードジェネレータ、ランタイムを提供し、簡単なキャリブレーションと量子化ステップを経て、事前学習済みのディープニューラルネットワークをデプロイすることができます。MERAは、PytorchやTensorflowLiteなどの深層学習フレームワークで直接量子化されるモデルをサポートしています。

機能詳細について

多様なオペレータサポート

- 標準畳み込みおよびDepthwise畳み込み
- ストライドとダイレーション
- 対称/非対称パディング
- 最大値プーリング、平均プーリング
- ReLU、ReLU6、LeakyReLU、H-SwishおよびH-Sigmoid
- アップサンプリングとダウンサンプリング
- 残差接続、分岐など

GPUへのドロップイン置換

- PythonおよびC++のインターフェース
- PyTorchとTensorFlow-liteをネイティブサポート
- 再トレーニング不要
- 高解像度入力に対応

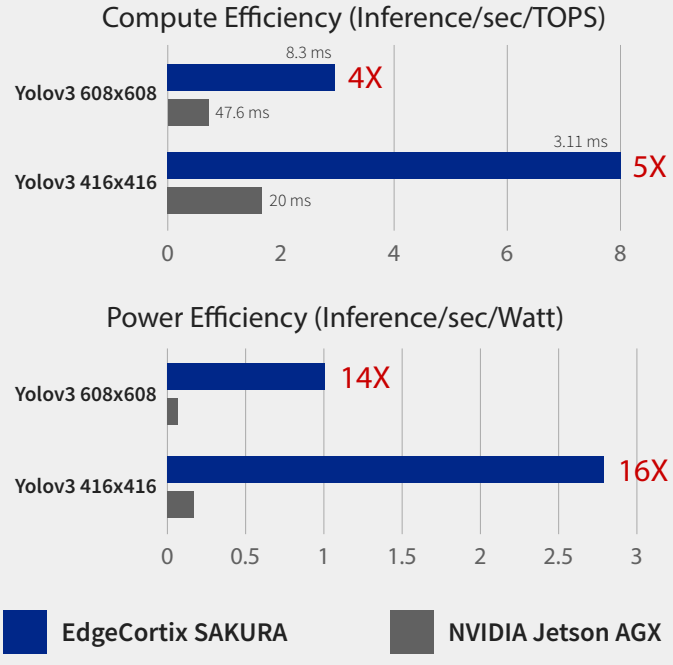
INT8ビット量子化

- トレーニング後のキャリブレーションと量子化
- 高精度を維持

シミュレータの内蔵

- SAKURAデバイスを使用せず、x86環境で推論のシミュレートが可能
- 様々な条件下での推論レイテンシとスループットを推定

10倍以上のエネルギー効率



* バッチサイズ1、MAXNモード(30W)でのNvidiaの結果
** ベースラインはいずれもYolov3 608x608
*** SAKURAは10WTDP

詳しくはこちら edgecortix.com

EDGECORTIX