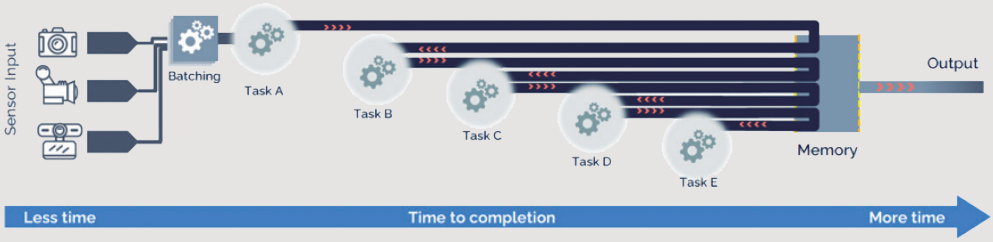
EDGECORTIX®
Dynamic Neural Accelerator®
Run-time Reconfigurable Neural Network IP

EdgeCortix Dynamic Neural Accelerator-II (DNA-II) は、あらゆるホスト・プロセッサと組み合わせることができる高効率でパワフルなニューラル・ネットワークIPコアです。
DNA-IIは演算要素間のランタイム再構成可能なインターコネクトによって優れた並列性と効率を実現し、畳み込みネットワークとトランスフォーマーネットワークの両方をサポートしており、さまざまなエッジAIアプリケーションに最適です。DNA-IIは、1K MACからスケーラブルなパフォーマンスで、幅広いターゲット・アプリケーションとSoCの実装をサポートします。

DNA-II の効率性

通常、TOPS（Tera Operations Per Second）は、コンピューティングユニットが完全に並列化された最適な状態で見積もられます。AIモデルが他のベンダーのハードウェアにマッピングされると、並列度が低下し、達成可能なTOPSは、謳われているピーク性能のほんの一部にまで低下します。EdgeCortixは、特許取得済み のランタイムで再構成可能なデータパスアーキテクチャを使用して、DNAエンジン間のデータパスを再構成 し、より優れた並列性とオンチップ・メモリ帯域幅の削減を実現します。

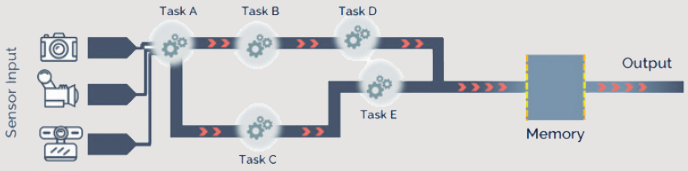
産業用IPコアの典型的なAI推論フロー



IPコアの非効率性

- バッチ処理により処理速度が低下
- リソースの再利用が多いため消費電力が高い
- コンピュート利用率の低下により効率が低下

ランタイムで再構成可能なデータパスを備えたDNA

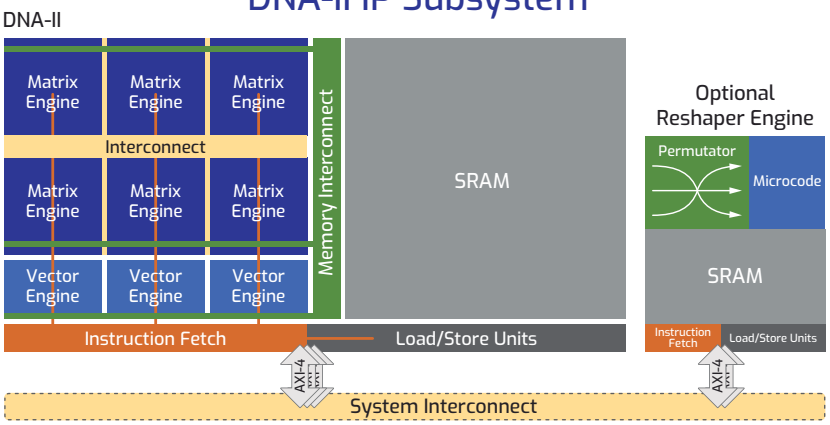


DNAデータパスの利点

- はるかに高い稼働率と効率性
- タスクとモデルの並列処理により処理速度が大幅に向上
- エッジAIのユースケースに適した超低消費電力

MERAソフトウェアとコンパイラフレームワークは、同じDNAエンジンで複数のモデルが同時に実行されている場合でも、ニューラルネットワークのタスクをスケジューリングする際の計算順序とリソース割り当てを最適化し、より良い性能、より高いエネルギー効率、より低いレイテンシを実現します。

DNA-II IP Subsystem



MERAを用いたDNA IPの主な利点

カスタマイズ可能なIP実装：対象の実装とプロセス技術に応じたご希望の性能（TOPS）,サイズ、消費電力

生成系AIに最適：LLMやLVMを含む一般的な生成系AIモデルをサポート

効率的なAI演算：非常に高いAI演算利用率を達成し、卓越したエネルギー効率を実現

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：ハードウェアアクセラレーションによる近似が適応性を向上

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

効率的なメモリ利用：効率的なメモリ圧縮により、メモリトータル量を削減し、DRAM帯域幅を最適化

電力管理：高度な電力管理により、超高効率モードを実現。DNAコアは、消費電力と性能のトレードオフに合わせて最適化

柔軟な計算ブロック：計算ブロックは、32x32から64x64まで最適化されたシストリックアレイを使用しています。独立したベクトル・ユニットにより、ニューラルネットワークの活性化とスケーリングが可能

オプションのテンソル変換IP：統合されたテンソル変換エンジンにより、ホストCPUの負荷を最小化

DNA IPは業界をリードする性能と電力を実現

下表は、12nm TSMC FinFETプロセスに基づくIP機能の例を示しています。

IPは、特定のユーザー要件に合わせてカスタマイズできます。

Example DNA IP Implementations					
Configuration	1	2	3	4	5
MACs	1K	2K	4K	8K	12K
Matrix Engines	1	2	4	2	3
Vector Engines	0.5	1	2	1	1
Channels	32	32	32	64	64
Memory (MB)	2	2.5	4	5	6
TOPS (@ 800MHz*)	1.6	3.3	6.6	13.1	19.7

*DNA-IIは100MHzのクロック周波数をサポート

提出物

- カスタマイズ可能なマルチコアオプションのDNA IPソースコード
- MERAソフトウェアとコンパイラフレームワーク
- 定義済みモデルの大規模セット
- テストベンチ例
- ドキュメント一式
- テクニカルサポート
- オプションのテンソル変換IP

EdgeCortix DNA IPとMERAを使用した開発

MERAは、DNA IPによるディープニューラルネットワークグラフのコンパイルとAI推論を可能にするコンパイラとソフトウェアフレームワークです。オープンソースのApache TVMコンパイラ・フレームワークのビルトインサポートにより、DNA IPに事前トレーニングされたディープニューラルネットワークのビットストリームを展開するために必要なツール、API、コードジェネレータ、ランタイムを提供します。MERAは、PyTorch、TensorFlow、TensorFlow Lite、ONNXなどのツールを使用したモデル開発ワークフローをサポートしています。

DNA AI IPコアの詳細はこちら

MERAコンパイラとソフトウェアフレームワークの詳細はこちら