製品概要

アクセラレーション
FPGA

# インテル® Stratix® 10 NX FPGA

## AIに最適化されたFPGAで高帯域幅、低レイテンシーのAIアクセラレーションを実現

**最大**
**143TOPS(INT8)**
**286TOPS(INT4)[1]**

インテル® Stratix® 10 NX FPGAは、高性能な人工知能(AI)を統合したカスタム・ハードウェアを実装するために必要な機能を、独自に組み合わせて提供します。その主な機能は次のとおりです。

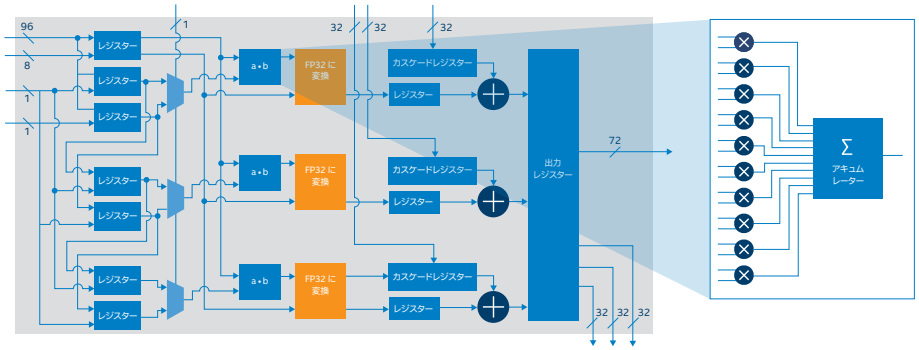
- **高性能 AI Tensorブロック**
  - 最大143TOPS(INT8)/286TOPS(INT4)を実現、AIワークロードでワット当たり1～2TOPSの演算処理[1]
  - AIによるカスタム・ワークロードを実現するプログラマブルなハードウェア
- **豊富なニア・コンピューティング・メモリー**
  - 多様なメモリー階層を組み込み、モデルの永続性を実現
  - 最大512GB/秒のパッケージ内メモリー帯域幅を備えた、最大16GBの広帯域幅メモリー(HBM)を統合
- **高帯域幅ネットワーキング**
  - 最大668GB/秒の接続帯域幅を実現する最大57.8GのPAM4トランシーバー
  - 最大12のハード100Gイーサネット MAC/PCS/FEC接続
  - 高い柔軟性とカスタマイズ可能なインターコネクトにより、複数のノードに拡張が可能

これらの3つの機能により、インテル® Stratix® 10 NX FPGAは、再構成可能なカスタム機能に加え、極めて高密度なコンピューティング、メモリー帯域幅、複数のノードへの拡張性が求められる、低レイテンシーで大規模なAIモデルへと移行するトレンドに、独自の方法で対処できます。

### AI Tensorブロックの導入:画期的な高密度コンピューティングを実現

インテル® Stratix® 10 NX FPGAファブリックには、「AI Tensorブロック」と呼ばれるAIに最適化された新しいタイプの演算ブロックが含まれます。各ブロックに3つのドット積ユニットが備わり、ユニットそれぞれに乗算器x10、アキュムレーターx10、つまり1つのAI Tensorブロックに合計30の乗算器とアキュムレーターが実装されることになります。AI Tensorブロックのアーキテクチャーは、多岐にわたるAIコンピューティングでも汎用的な行列x行列演算またはベクトルx行列演算に合わせて調整されており、行列のサイズが小さい場合でも大きい場合でも効率的に機能するよう設計されています。

**AI Tensorブロックの概略図**




性能やベンチマーク結果について、さらに詳しい情報をお知りになりたい場合は、http://www.intel.com/benchmarks/(英語)を参照してください。

AI Tensor ブロックの乗算器は、INT8 と INT4 を基本精度とし、共有指数サポート・ハードウェアを介して 16 ビットのブロック浮動小数点（Block FP16）と 12 ビットのブロック浮動小数点（Block FP12）の数値形式をサポートします。すべての加算または累算は INT32 または IEEE754 単精度浮動小数点（FP32）精度で実行でき、複数の AI Tensor ブロックをカスケード接続することで、大規模な行列演算も実行可能です。インテル® Stratix 10 NX FPGA は、最大 143TOPS/TFLOPS（INT8/Block FP16）または 286TOPS/TFLOPS（INT4/Block FP12）を達成できると推定されています。[1]

## 低レイテンシーかつ大規模モデルを実現するマルチノード・ソリューションで、AI 機能を拡張

### 自然言語処理

- 音声認識
- 音声合成

### セキュリティー

- ディープ・パケット・インスペクション
- 輻輳制御の識別
- 不正検出

### リアルタイム・ビデオ解析

- コンテンツ認識
- ビデオの前処理 / 後処理

## インテル® Stratix® 10 NX FPGA – 主な特長

| 主な特長 | |
|---|---|
| AI Tensor ブロック | AI 演算に最適化された AI Tensor ブロックは、標準的なインテル® Stratix® 10 FPGA の DSP ブロックに比べて、推定で最大 15 倍の INT8 スループットを実現し[1]、高スループットの AI 推論アプリケーションに必要な高密度コンピューティングを実装します。 |
| 3D スタックの HBM2 高帯域幅 DRAM をパッケージ内に統合 | 統合されたメモリースタックにより、大規模で永続的な AI モデルをオンチップに格納でき、その結果、低レイテンシーと高メモリー帯域幅を実現し、大規模なモデルにおけるメモリーの制約によるパフォーマンスの課題を解消します。 |
| トランシーバー・データ・レート | 最大 57.8G の PAM4 トランシーバーを搭載したインテル® Stratix® 10 NX FPGA は、マルチノード設計の制約要因となっている接続帯域幅の要件を緩和または排除して、マルチノードの AI 推論ソリューションを実装できる拡張性と柔軟性を実現します。また、インテル® Stratix® 10 NX FPGA は、PCI Express*（PCIe*）Gen3 x16、10/25/100G イーサネット MAC（メディア・アクセス・コントロール）/PCS（フィジカル・コーディング・サブレイヤー）/FEC（順方向エラー訂正）などのハード IP も内蔵しています。これらのトランシーバーにより、市場の要件に適応できる拡張性と柔軟性を備えた接続ソリューションを提供します。 |

## 詳細情報

インテル® Stratix® 10 NX FPGA の詳細については、http://www.intel.co.jp/stratix10nx/ を参照してください。

[1] インテル社内での推定値に基づきます。

テストは、特定システムでの特定テストにおけるコンポーネントのパフォーマンスを測定しています。ハードウェア、ソフトウェア、システム構成などの違いにより、実際の性能は掲載された性能テストや評価とは異なる場合があります。購入を検討される場合は、ほかの情報も参考にして、パフォーマンスを総合的に評価することをお勧めします。性能やベンチマーク結果について、さらに詳しい情報をお知りになりたい場合は、http://www.intel.com/benchmarks/（英語）を参照してください。

インテルのテクノロジーを使用するには、対応したハードウェア、ソフトウェア、またはサービスの有効化が必要となる場合があります。絶対的なセキュリティーを提供できる製品またはコンポーネントはありません。

結果は推定またはシミュレーションに基づいています。実際のコストや結果は異なる場合があります。



ss-1121-2.0/JP