ソリューション概要

セキュリティー強化
インテル® Xeon® スケーラブル・プロセッサー

コンフィデンシャル・コンピューティングで AI 推論を高速化

Fortanix は、インテル® アクセラレーター・エンジンのワークロードを最適化することで、第 4 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーを使用して、クラウドでの AI 推論の安全性と高速化を支援します。

コンフィデンシャル AI では、セキュリティーとパフォーマンスの両方が重要です。第 4 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーは、AI 推論の安全性と高速化を支援する設計になっています。インテル® アクセラレーター・エンジンはインテル® Xeon® スケーラブル・プロセッサー・ファミリーに内蔵された専用アクセラレーターであり、今日急成長を遂げているワークロードの多くに、パフォーマンスと電力効率のメリットをもたらします。

インテル® ソフトウェア・ガード・エクステンションズ（インテル® SGX）とインテル® アドバンスト・マトリクス・エクステンション（インテル® AMX）の両方を使用する Fortanix Runtime Encryption® (RTE) プラットフォーム上のワークロードでは、TensorFlow Resnet50 を実行する場合のパフォーマンスが最大 7.57 倍[1]、Bert-Large を実行する場合のパフォーマンスが最大 5.26 倍向上します。[2]

インテル® AMX などのアクセラレーターにより、AI 推論ワークロードは、インテル® SGX などのハードウェアベースのセキュリティーと組み合わされた場合でも、優れたパフォーマンスを発揮できます。

コンフィデンシャル・コンピューティングがクラウドベース AI の安全性を支援

Fortanix RTE はインテル® SGX によるコンフィデンシャル・コンピューティングを使用して、OS やそのほか実行中のプロセスにプレーンテキストのアプリケーション・コードを露出することなく、暗号化されたデータの汎用演算を可能にします。インフラストラクチャーが侵害された場合、また悪意のある組織内の人間がルートパスワードを保持している場合でも、アプリケーションは暗号により保護されます。

コンフィデンシャル・コンピューティングでは、ほかのソフトウェア、コラボレーター、クラウド・プロバイダーなどに機密データを露出することなく使用する方法で、インサイトの導出や AI モデルのトレーニングが可能です。これにより、「機密性が高い」「規制されている」などの理由で以前は分析やその他の目的のために有効化できなかったデータを使用した業務改革が可能になります。

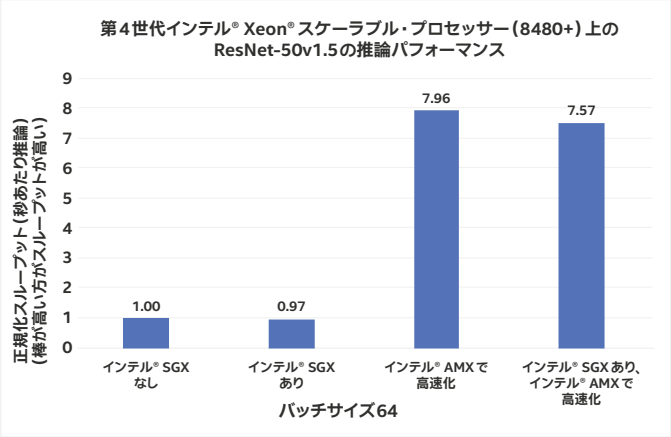

図1. インテル® AMX およびインテル® SGX 搭載の第4世代インテル® Xeon® スケーラブル・プロセッサーを使用した ResNet50 推論ワークロードのパフォーマンス[1]

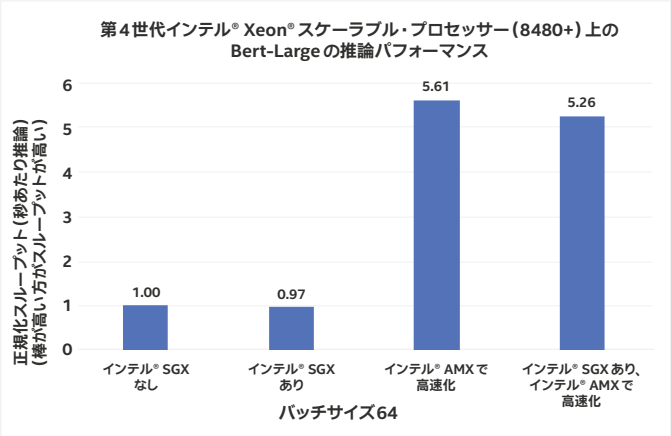

図2. インテル® AMX およびインテル® SGX 搭載の第4世代インテル® Xeon® スケーラブル・プロセッサーを使用した Bert-Large 推論ワークロードのパフォーマンス[2]

インテル® AMX で AI 機能を高速化

インテル® AMX は、ディープラーニング・トレーニングと CPU 上の推論のパフォーマンスを向上させる新しい内蔵アクセラレーターであり、自然言語処理、推奨システム、イメージ認識などのワークロードに最適です。インテルは第4世代インテル® Xeon® スケーラブル・プロセッサーとインテル® AMX で AI 機能をさらに進化させています。推論やトレーニングのパフォーマンスも前世代のインテル® Xeon® スケーラブル・プロセッサーに比べて向上しています。[3]

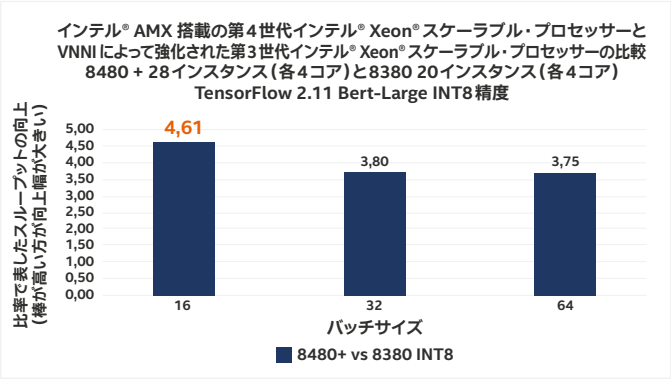

図3. 第3世代インテル® Xeon® スケーラブル・プロセッサーと第4世代インテル® Xeon® スケーラブル・プロセッサーの INT8 精度におけるセキュア DL 推論ワークロードのパフォーマンスの比較[3]

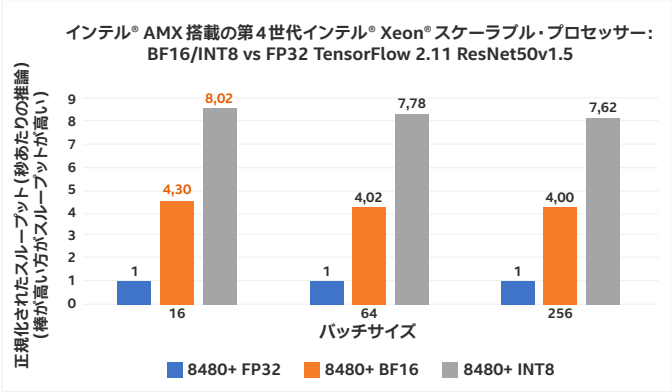

図4. 第4世代インテル® Xeon® スケーラブル・プロセッサー上の TensorFlow ResNet50 のさまざまなバッチサイズでのセキュア DL 推論ワークロードのパフォーマンス[4]

インテル® SGX とインテル® AMX を両方搭載した Fortanix RTE を使用した場合、第4世代インテル® Xeon® スケーラブル・プロセッサー上の INT8 精度における Bert-Large 推論ワークロードのパフォーマンスは、第3世代インテル® Xeon® スケーラブル・プロセッサー上と比較して最大 4.61 倍向上します。[3]

市場最多の内蔵アクセラレーターを手中に

第4世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーは、市販されている CPU で最も多くのアクセラレーターを内蔵しており、新たなワークロード、特に AI を活用するワークロードのパフォーマンスを高めるために役立ちます。

パフォーマンスの向上に加えて、第4世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーは、絶えず変化する脅威環境の中でデータを保護する高度なセキュリティー・テクノロジーを備え、ビジネス上のインサイトを導出する新たな機会をもたらします。インテル® SGX など、セキュリティー向けに設計されたハードウェア対応の機能を有効にした場合でも、インテル® AMX などのインテル® アクセラレーター・エンジンにより AI 推論ワークロードの卓越したパフォーマンスが実現します。

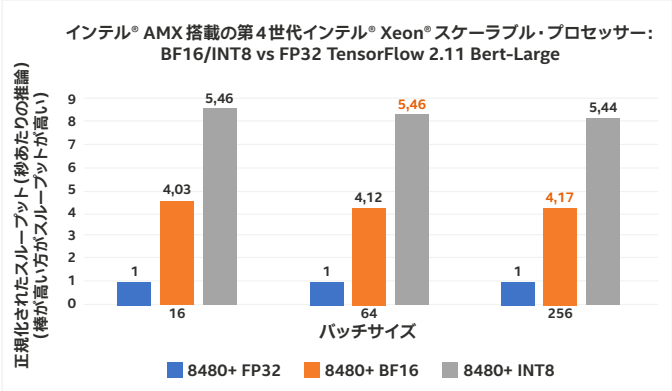

図5. 第4世代インテル® Xeon® スケーラブル・プロセッサー上の TensorFlow Bert-Large のさまざまなバッチサイズでのセキュア DL 推論ワークロードのパフォーマンス[5]

第４世代インテル® Xeon® スケーラブル・プロセッサーであれば、パフォーマンスとセキュリティーを両立できます。

詳細情報は、**intel.co.jp/xeonaccelerated/** および **fortanix.com/platform/runtime-encryption/（英語）** を参照してください。

1. インテル® SGX およびインテル® AMX を搭載した第４世代インテル® Xeon® スケーラブル・プロセッサー上で TensorFlow ResNet50推論ワークロードを実行した場合、パフォーマンスが最大7.57倍向上。以下の構成情報を参照してください。
2. インテル® SGX およびインテル® AMX を搭載した第４世代インテル® Xeon® スケーラブル・プロセッサー上で Bert-Large 推論ワークロードを実行した場合、パフォーマンスが最大5.26倍向上。以下の構成情報を参照してください。
3. INT8精度で Bert-Large 推論ワークロードを実行した場合のパフォーマンスは、インテル® SGX およびインテル® AMX を搭載した第４世代インテル® Xeon® スケーラブル・プロセッサー上では前世代と比較して最大4.61倍向上。以下の構成情報を参照してください。
4. インテル® SGX およびインテル® AMX を搭載した第４世代インテル® Xeon® スケーラブル・プロセッサー上で TensorFlow ResNet50推論ワークロードを実行した場合、FP32との比較で、INT8精度でのパフォーマンスが最大8.02倍、BF16精度でのパフォーマンスが最大4.30倍向上。以下の構成情報を参照してください。
5. インテル® SGX およびインテル® AMX を搭載した第４世代インテル® Xeon® スケーラブル・プロセッサー上で Bert-Large 推論ワークロードを実行した場合、FP32との比較で、INT8精度でのパフォーマンスが最大5.46倍、BF16精度でのパフォーマンスが最大4.17倍向上。以下の構成情報を参照してください。

構成の詳細

TEST-1：2022年11月21日にインテルが社内で実施したテストの結果。１ノード、インテル® Xeon® Platinum 8380 CPU @ 2.30GHz x2、40コア、HT 無効、ターボ有効、メモリー合計512GB（16x32GB DDR4 3200MT/s [run @3200MT/s]）、BIOS バージョン SE5C6200.86B.0022.D64.2105220049、ucode バージョン0xd000375、OS バージョン Ubuntu 22.04.1 LTS、カーネルバージョン6.0.6-060006-generic、Fortanix の安全なエンクレーブ内のワークロード/ベンチマーク・ディープラーニング推論、フレームワーク・バージョン TensorFlow 2.11、モデル名＆バージョン ResNet50v1.5/Bert-Large

TEST-2：2022年11月21日にインテルが社内で実施したテストの結果。１ノード、インテル® Xeon® Platinum 8480+ CPU @ 2.0GHz x2、56コア、HT 無効、ターボ有効、メモリー合計512GB（16x32GB DDR5 4800MT/s [run @4800MT/s]）、BIOS バージョン3A05、ucode バージョン 0x2b000070、OS バージョン Ubuntu 22.04.1 LTS、カーネルバージョン6.0.6-060006-generic、Fortanix の安全なエンクレーブ内のワークロード/ベンチマーク・ディープラーニング推論、フレームワーク・バージョン TensorFlow 2.11、モデル名＆バージョン ResNet50v1.5/Bert-Large

通知と免責事項

性能は、使用状況、構成、その他の要因によって異なります。詳細については Performance Index のサイトをご覧ください。

性能の測定結果は構成情報に記載された日付時点のテストに基づくものです。また、公開中のすべてのアップデートが適用されているとは限りません。構成の詳細については、補足資料を参照してください。

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。

実際のコストと結果は異なる場合があります。

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。

356372-001JA
JPN/2308/PDF/TAG/MKTG/SN