製品概要

インテル® AI エンジン
第5世代インテル® Xeon® スケーラブル・プロセッサー・ファミリー

第5世代インテル® Xeon® スケーラブル・プロセッサーとインテル® AI エンジンが AI 処理全体のパフォーマンスを大幅に向上

65%
インテル® Xeon® プロセッサー上で
稼働するデータセンターの AI 推論の割合[1]

最大
14倍向上
インテル® AMX BF16内蔵の第5世代インテル® Xeon® プロセッサー使用時のリアルタイム・オブジェクト検出（検知）推論パフォーマンス（SSD-ResNet34）を第3世代インテル® Xeon® プロセッサーと比較した場合[2]

最大
9.9倍向上
インテル® AMX BF16内蔵第5世代インテル® Xeon® プロセッサー使用時のリアルタイム自然言語処理推論（BERT-large）パフォーマンスを第3世代インテル® Xeon® プロセッサーと比較した場合。また、同じ比較条件でワット当たりの性能は最大7.7倍向上[3]

最大
8倍向上
第5世代インテル® Xeon® プロセッサー使用時のバッチのレコメンデーション・システムの推論パフォーマンス（DLRM）を、第3世代インテル® Xeon® プロセッサーと比較した場合。また、同じ比較条件でワット当たりの性能は6.2倍向上[4]

AI のワークロードやユースケースは、データの前処理、昔からあるマシンラーニング（ML）をはじめ、自然言語処理や画像認識といったディープラーニングの用途まで広範囲に及びます。インテル® Xeon® スケーラブル・プロセッサーを使用することで、このような AI 処理全体のパフォーマンスを大幅に向上させることが可能です。各プロセッサーには、マシンラーニングやデータ分析、ディープラーニングなど、特定の AI ワークロード向けに最適化されたアクセラレーターが内蔵されています。

あらゆる用途の AI に対応

AI は今日のビジネスに欠かせないものとなっており、影響範囲は多種多様かつ非常に重要なワークロードにおよびます。マシンラーニングやディープラーニングは、コア・エンタープライズ・アプリケーションから自動音声応答機能に至るまで、ビジネスの基本的な構成要素になりつつあります。AI を大規模に運用できるかどうかは、データの前処理からトレーニング、導入までの長期的な開発プロセス次第です。各ステップにはそれぞれ独自の開発ツールチェーン、フレームワーク、ワークロードがあり、そのすべてにおいて特有のボトルネックが発生し、それぞれ異なる要因でコンピューティング・リソースに負荷をかけます。インテル® Xeon® スケーラブル・プロセッサーに内蔵されているアクセラレーターが、処理全体の実行環境を整え、AI パフォーマンスを全体的に向上させることが可能です。

要求が非常に厳しい新たなワークロードをサポートするための統合アクセラレーター、インテル® アクセラレーター・エンジン

第5世代インテル® Xeon® スケーラブル・プロセッサーは、汎用コンピューティングに優れているだけでなく、日々重要性の高まる AI ワークロードの多くをサポートする基盤としての役割を果たします。また、CPU のディープラーニング推論とトレーニングの高速化をはかるために設計された AI アクセラレーター、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）が内蔵されています。内蔵 AI アクセラレーターのおかげで、ディスクリート・アクセラレーターにかかる追加コストや複雑性を回避できます。最新世代のインテル® Xeon® プロセッサーは、パラメーター数が200億（20B）未満の大規模言語モデル（LLM）に適しており、ほとんどのお客様との合意サービス水準（SLA）を満たしています。[5] インテル® AMX は、トランスファー・ラーニング（転移学習）やファインチューニングにも優れているため、ハードウェア追加の必要なしに、わずか4分でモデルをトレーニングできます。 データセンターの推論の65% がインテル® Xeon® プロセッサーで実行されています。そのためお客様は、GPU インフラストラクチャーに移行する際の複雑性に煩わされることなく、汎用 AI 向けの既存アーキテクチャーの恩恵を受けられます。

第5世代インテル® Xeon® スケーラブル・プロセッサーとインテル® アクセラレーター・エンジンが導く未来のイノベーション

インテル® Xeon® プロセッサーを使用しているなら、オンプレミス、クラウド、エッジなど実行環境を問わず、内蔵のインテル® アクセラレーター・エンジンで、ビジネスを新たな高みに導くことが可能です。データ保護の強化やインフラストラクチャーの活用など、さまざまな利点を得られます。インテル® アクセラレーター・エンジンは、仮想 CPU と物理 CPU の使用率を向上させ、コア・ソリューション・ライセンスのコストを削減します。すなわち、これらの内蔵アクセラレーターは、アプリケーションのパフォーマンス向上、コスト削減、プラットフォーム・レベルの効率向上を実現します。

採用事例: インテル® Xeon® スケーラブル・プロセッサーによる現実世界のアクセラレーション

Tencent Cloud: インテル® Xeon® スケーラブル・プロセッサーにより、リアルタイムの音声合成を実現。

ストーリーを読む(英語) ›

Gunpowder: 第4世代インテル® Xeon® スケーラブル・プロセッサー搭載 Google Cloud C3インスタンスで、レンダリング・パフォーマンスの高速化を実現。

ストーリーを読む ›

インテル® アドバンスト・マトリクス・エクステンションによるディープラーニングの高速化

インテル® AMX は、第 5 世代インテル® Xeon® スケーラブル・プロセッサーでディープラーニングのトレーニングや推論を行うためのインテルの最先端テクノロジーです。インテル® AMX は、自然言語処理、レコメンデーション・システム、画像認識のようなワークロードに最適で、AMX BF16 内蔵の第5世代インテル® Xeon® プロセッサーは、第3世代と比較して、リアルタイム・オブジェクト分類推論パフォーマンスが最大7.2倍、ワット当たりの性能が5.3倍向上します。[6]

またインテル® AMX は、AI モデルのワークロードを向上させるだけでなく、多くのお客様がすでに運用しているプラットフォームの SLA を満たします。第5世代インテル® Xeon® スケーラブル・プロセッサーは、5段階のターボレシオを追加し、ハイパフォーマンス・コンピューティングや AI などのベクトル演算、マトリックス演算に親和性のあるワークロードに対してターボ周波数を向上させることが可能です。

インテル® AMX は、CPU コアのインテル® アドバンスト・ベクトル・エクステンション512 (インテル® AVX-512) と比較すると、スループット (Ops/Cycle) が高く、行列乗算のパフォーマンスが向上しています。[7] そのため、ディープラーニングのトレーニングのワークロードがより迅速に完了するだけでなく、多くのお客様がすでにビジネスで運用しているプラットフォームの SLA を満たすことが可能です。

自然言語処理と生成 AI をサポート

インテル® AMX 内蔵の第5世代インテル® Xeon® スケーラブル・プロセッサーは、ハードウェアの追加を必要することなく、自然言語処理のパフォーマンスが大幅に向上します。インテルの各種ライブラリーは TensorFlow や PyTorch 向けに最適化されているため、開発者は、内蔵 AI アクセラレーターの利点をすぐに利用できます。開発者は、異なるハードウェア環境からコードを移植するという、時間とコストがかかる可能性のあるプロセスを、より簡単に行えるようにもなります。

インテル® AMX 内蔵の第5世代インテル® Xeon® スケーラブル・プロセッサーは、ディープラーニングの推論とトレーニングを高速化することで、総保有コスト (TCO) の均衡を保ちつつ SLA を満たします。リアルタイムのユーザー行動シグナルに加え、時間や場所のようなコンテキスト (状況) を考慮するディープラーニング・ベースのレコメンダー・システムで活用されています。

第5世代インテル® Xeon® プロセッサーは、人間中心のコンテンツを模倣する生成 AI モデルの実行も可能で、大規模言語モデルや Text-to-image 生成 (テキストから画像への生成) をサポートします。生成 AI により特化したタスクについては、専用のインテル® Gaudi® AI アクセラレーター、インテル® データセンター GPU、その他のハードウェア・コンポーネントを使用して、CPU の機能拡張が可能です。

インテル® AVX-512 で ML を高速化

インテル® Xeon® プロセッサーは、ウェブサイト向け SSL 暗号のハッシュ化、巨大なデータベースの処理、そして製薬研究やチップ設計、F1 エンジンのシミュレーションを可能にします。

数世代にわたり改良されてきたインテル® AVX-512 は、インテル® Xeon® スケーラブル・プロセッサーの各クロックサイクルにおいてより多くのオペレーションを可能にするだけでなく、並列処理アプリケーションのパフォーマンス向上も実現しています。インテル® AVX-512の命令セット・アーキテクチャー (ISA) には、AI、HPC、ネットワーキング、ストレージといった多様なワークロードのパフォーマンス強化のために構築された拡張機能を備えています。

新世代のプロセッサーのターボ・パフォーマンスについては、ターボレシオが4段階から5段階に増えています。そのため、インテル® AMX やインテル® AVX-512 を活用すると、特定の HPC や AI ワークロードのターボ周波数が向上します。

少ないステップで処理を高速化

第5世代インテル® Xeon® スケーラブル・プロセッサーに内蔵されているインテル® AVX-512は、スマートで美しい数学を多用して、通常のコンピューティング作業を凝縮・結合・融合し、ステップの数を減らします。基本要素の例として、5回のクロックサイクルを要する「3x3x3x3x3」の演算をCPUに命令するとします。あるいはCPUが1回のサイクルで処理できるよう、3^5（3の5乗）の命令にすることもできます。インテル® AVX-512は、このようなロジックを採用し、AIで最も困難な作業を含む数百ものワークロードに特化した作業に応用しています。

1刻みよりも遥かに速い8刻みを採用

インテル® AVX-512の「512」とは、命令が1回のクロックサイクルにおけるCPU処理のビット数を増やすという第2の方法を意味しています。40年前の16ビットPCの登場は非常に印象的でしたが、その後すぐに32ビットマシンに取って代わりました。今日のスマートフォンは64ビットで動作します。ビット数は、CPUが1回のクロックサイクルごとにアドレス指定できるレジスターの数（CPUがデータを保持するメモリースロット）を示しています。名前が示すように、インテル® AVX-512はレジスター数が512ビットに拡張されています。アプリケーションがインテル® AVX-512を活用する際、レジスターの数を単に拡張するだけで、CPUが基本とする64ビットの速度よりも最大で8倍高速に動作させられることになります。96まで数える際に、8、16、24・・・と数えるか、1、2、3・・・と数えるかの違いです。

より少ない電力でより強力なAIを実行するエンジン

インテル® AIエンジン内蔵のインテル® Xeon® スケーラブル・プロセッサーは、必要なハードウェア・リソースが少なくて済むため、より強力で電力効率に優れたソリューションでAIワークロードの実行が可能になります。

アクセラレーター・エンジンを内蔵したインテル® Xeon® スケーラブル・プロセッサーは、TCOの削減や、要求の厳しいAIワークロードの投資収益率（ROI）の向上など、ワークロードの結果改善に貢献しています。

インテル® Xeon® プロセッサーにより、実質的に自動で動く高速AI

インテル® Xeon® スケーラブル・プロセッサーのAIアクセラレーターは、CPUの命令セット・アーキテクチャー（ISA）に組み込まれています。これは、利用可能なソフトウェアが準備されていて、それらを活用できることを意味しています。インテルのソフトウェア・エンジニアは、オープンソースのAIツールチェーンを常に最適化し、成果物をコミュニティに還元しています。例えばTensorFlow 2.9には、最適化されたインテル® oneAPIディープ・ニューラル・ネットワーク・ライブラリー（インテル® oneDNN）が最初から付属しています。最新版のTensorFlowをダウンロードすると、自動的にインテル® Optimization for TensorFlowが使用可能です。

データ・サイエンティストや開発者は、AI処理に使用されるアプリケーションのために、オープンソースで無料のインテル® ディストリビューションやライブラリー、開発環境をダウンロードして、インテル® Xeon® スケーラブル・プロセッサー用のISAに内蔵されているすべてのアクセラレーターを利用できます。データ・サイエンティストやAIの開発者は、ツールをコーディングし直し、インテル® AVX-512用に再コンパイルする必要がなくなります。なぜなら、それらはすでにインテル® AVX-512用に最適化されているからです。

今日の企業や組織は、より高いワークロード・パフォーマンスをインフラストラクチャーから引き出し、電力効率の向上とコストの削減を行う必要があります。インテル® Xeon® スケーラブル・プロセッサーに統合された専用のインテル® AIエンジンは、ビジネスにとって最も重要なAIワークロードを最大限活用するのに役立ちます。

ビジネスにとって最も重要なAIワークロードのために、インテル® アクセラレーター・エンジン内蔵のインテル® Xeon® スケーラブル・プロセッサーを通じて、どんなことを実現できるかについては、詳細情報も参照してください。

詳細情報

インテル® Xeon® スケーラブル・プロセッサー上の AI とディープラーニング ›

インテル® アドバンスト・ベクトル・エクステンション 512（インテル® AVX-512）›

インテル® AI アナリティクス・ツールキット（英語）›

インテル® ハードウェア上およびソフトウェア上での開発（英語）›

AI やマシンラーニングのために最適化されたインテル® Xeon® プロセッサーを活用して、クラウドや自社のインフラストラクチャーでの AI ワークロードのアクセラレーションを始めてみませんか。

詳細情報（英語）›

intel xeon

1. 2022年12月時点での AI 推論ワークロードを実行するデータセンター・サーバーの、インテルによる世界のインストール・ベース市場モデリングに基づく。
2. intel.com/processorclaims/（英語）：5th Gen Intel Xeon® Scalable processors（A21）を参照してください。実際の結果は異なる場合があります。
3. intel.com/processorclaims/（英語）：5th Gen Intel Xeon® Scalable processors（A19）を参照してください。実際の結果は異なる場合があります。
4. intel.com/processorclaims/（英語）：5th Gen Intel Xeon® Scalable processors（A20）を参照してください。実際の結果は異なる場合があります。
5. 2023年12月時点でのインテル社内モデリングに基づく。
6. intel.com/processorclaims/（英語）：5th Gen Intel Xeon® Scalable processors（A22）を参照してください。実際の結果は異なる場合があります。
7. https://edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/（英語）の Session Benchmark #41と#42を参照してください。実際の結果は異なる場合があります。

358671-001JA
JPN/2401/PDF/TAG/MKTG/SI