ソリューション概要

人工知能(AI)
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)

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)による AI ワークロードの高速化

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 により AI 機能が進化し、前世代と比較して推論と学習処理のパフォーマンスが 3 倍 ~ 10 倍向上 [1]

AI パイプラインの最適化

企業はさまざまなシナリオに人工知能(AI) を適用してメリットを得ることができます。書籍や映画のレコメンドサービスから、大規模な e コマースサイトを動かす小売デジタル・ソフトウェア、チャットボットや機械翻訳の自然言語処理(NLP) まで、適用範囲は多岐にわたります。しかし、複雑な環境や膨大なデータセットを理解し、以前は解明できなかった問題を解決するという、AI を価値あるツールにしている特性に秘められた、ビジネスにさらなる変革をもたらす可能性は計り知れません。ある調査によると、2025 年までには、新たにリリースされるエンタープライズ・アプリケーションの 90% に AI 機能が組み込まれるようになると見込まれています。[2]

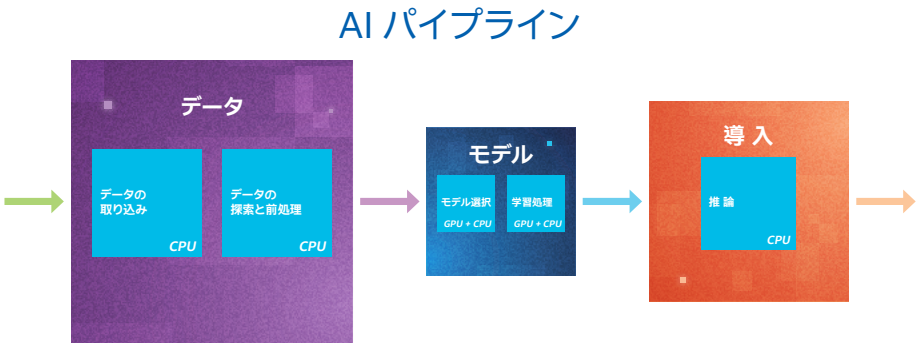

図 1. AI パイプライン内の AI ワークロードとプロセッサー・アクティビティー

AI パイプライン最適化のために導入すべきは、内蔵 AI アクセラレーターの 1 つであるインテル® アドバンスト・マトリクス・エクステンション (インテル® AMX) を搭載した第 4 世代インテル® Xeon® スケーラブル・プロセッサーです。インテル® AMX は、AI アプリケーションにおける CPU の用途として最も知られている推論と、比較的機能の多い学習処理とのバランスを意図して設計されました(図 1 を参照)。[3] インテル® Xeon® スケーラブル・プロセッサーは、データセンターで AI 推論ワークロードを実行しているプロセッサー・ユニットの(インストール・ベースで)70% を占めており、新たな AI 導入にインテル® AMX 搭載の第 4 世代インテル® Xeon® スケーラブル・プロセッサーを選択することは、AI ワークロードを高速化する効率的かつコスト効率の高いアプローチとなります。[4]

内蔵アクセラレーターの実例

インテル® ディープラーニング・ブースト (インテル® DL ブースト)搭載の第 3 世代インテル® Xeon® スケーラブル・プロセッサーを AI に導入すれば、IT チームはすぐにも顧客のサービスレベル・アグリーメント (SLA) を満たすことが可能ですが、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・プロセッサーを使用すると、状況はさらに一変します。

図 2 は、インテル® AMX により PyTorch でリアルタイム推論を実行したパフォーマンスが前世代と比較して最大 5.7 倍 ~ 10 倍に向上した結果を示しています。また図 3 は、インテル® AMX により PyTorch で学習処理を実行したパフォーマンスが前世代と比較して最大 3.5 倍 ~ 10 倍に向上した結果を示しています。[5] このパフォーマンスの結果から、インテル® AMX によって顧客満足度の向上を促進できることが分かります。すでに使い慣れているソリューションに、大幅なパフォーマンス向上を実現する内蔵アクセラレーターのインテル® AMX が統合されたことで、AI アプリケーションに最適な CPU の選択に迷うことはなくなります。

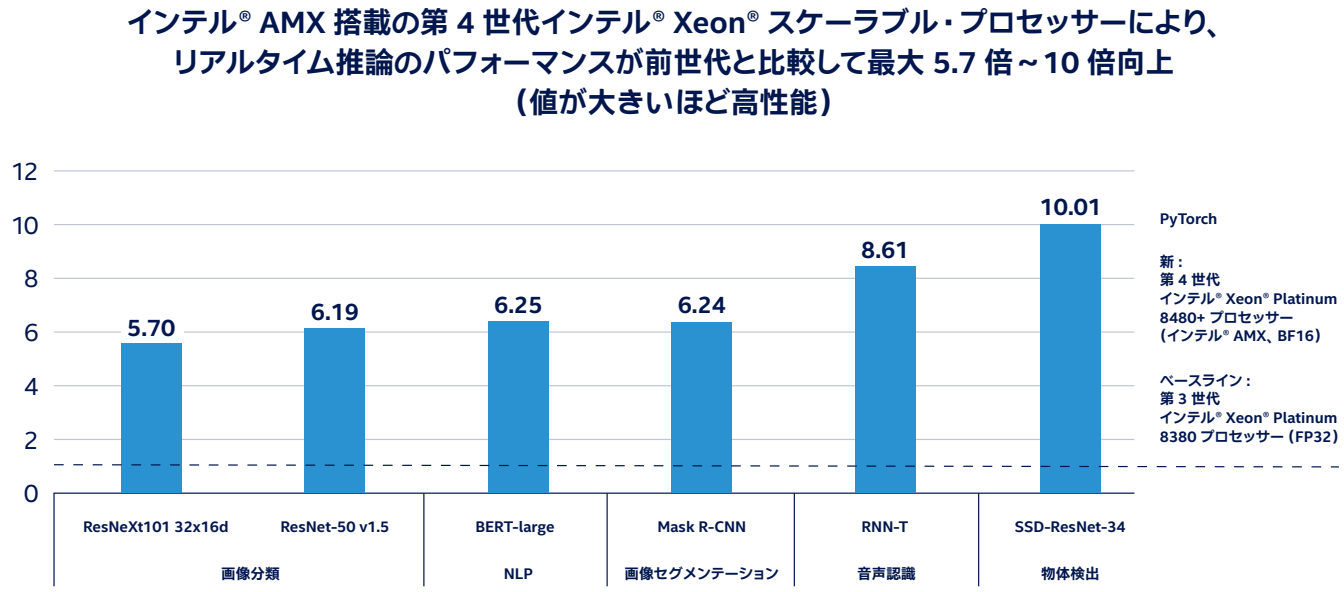

図 2. インテル® AMX 搭載の第 4 世代インテル® Xeon® スケーラブル・プロセッサーにより、PyTorch のリアルタイム推論パフォーマンスが向上[5]

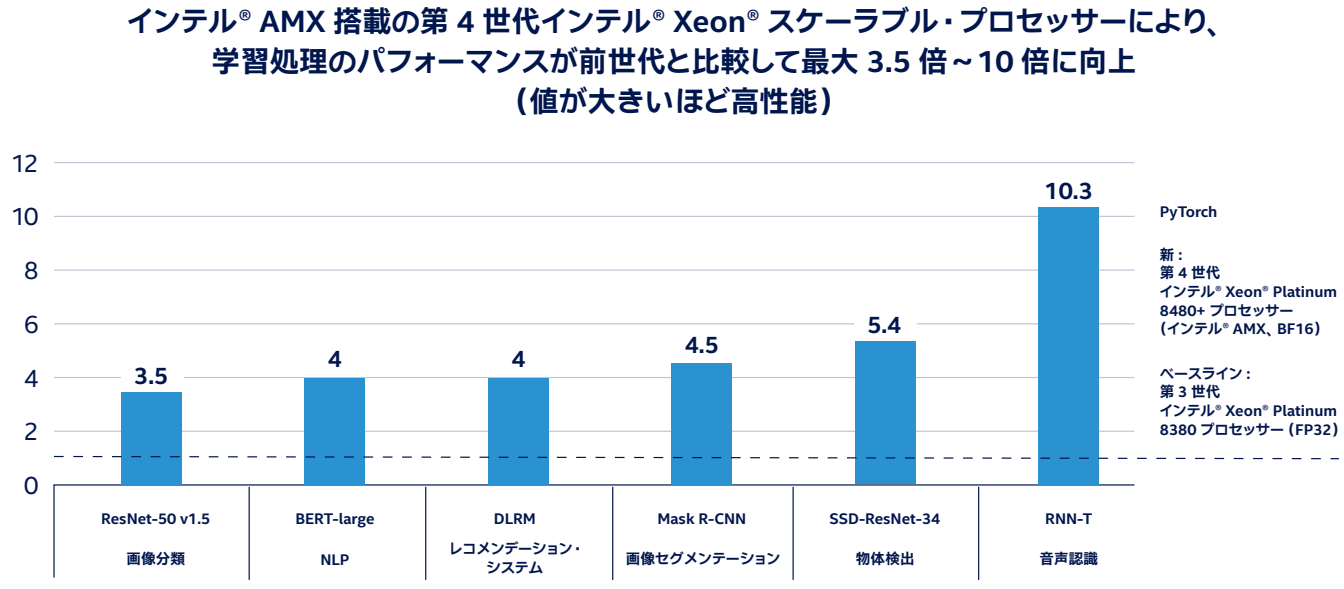

図 3. インテル® AMX 搭載の第 4 世代インテル® Xeon® スケーラブル・プロセッサーにより、PyTorch の学習処理パフォーマンスが向上[5]

図 4 は、初代のインテル® Xeon® スケーラブル・プロセッサーから世代を追うごとに増加したコア数よりも高い比率で、インテル® AMX がパフォーマンス向上を実現していることを示しています。

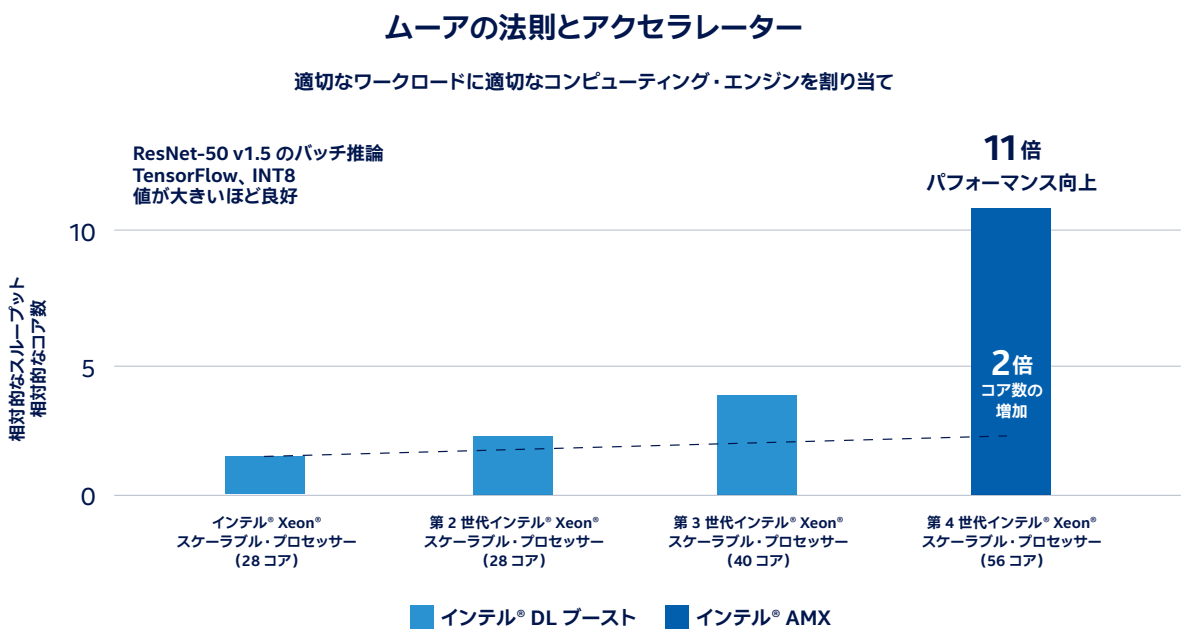

図 4. 初代のインテル® Xeon® スケーラブル・プロセッサーをベースラインとして、これまでの世代とは異なりインテル® AMX が実現するパフォーマンスの伸びは非線形[6]

インテル® AMT とは

インテル® AMX は、第 4 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 (DL) 学習処理 / 推論ワークロードの最適化を可能にする内蔵アクセラレーターです。第 4 世代インテル® Xeon® スケーラブル・プロセッサーはインテル® AMX を使用することで、汎用的なコンピューティングの最適化と AI ワークロードの間の素早い切り替えが可能になります。例えば市街地での走行を得意とする車が、突然 F1 レースのパフォーマンスに切り替わる状況を想像してみてください。このようなタイプの柔軟性が、第 4 世代インテル® Xeon® スケーラブル・プロセッサーには備わっています。開発者は、インテル® AMX の命令セットのメリットを活かして AI 機能をコーディングすることも、プロセッサーの命令セット・アーキテクチャー (ISA) を使用して AI 以外の機能をコーディングすること可能です。インテルは、oneAPI の DL エンジンであるインテル® oneAPI ディープ・ニューラル・ネットワーク (oneDNN) ライブラリーを、AI アプリケーションに広く利用されているオープンソース・ツール (TensorFlow、PyTorch、PaddlePaddle、ONNX など) に統合しました。

インテル® AMX アーキテクチャー

インテル® AMX のアーキテクチャーは、2 つのコンポーネントで構成されています (図 5 を参照)。

- 1 つ目のコンポーネントはタイルです。タイルは 8 つの 2 次元レジスターで構成されており、各タイルのサイズは 1 キロバイトです。タイルには大きなデータチャンクが格納されます。
- 2 つ目のコンポーネントは、タイル行列乗算 (TMUL) です。TMUL はタイルに接続されたアクセラレーター・エンジンであり、AI の行列乗算を実行します。

図5. 2Dレジスターファイル（タイル）とTMULで構成されるインテル® AMXアーキテクチャー

インテル® AMXでは、AIワークロードに必要な行列乗算で、INT8とBF16の2つのデータ型をサポートしています。

- INT8は、AIでよく使用される単精度浮動小数点形式であるFP32の精度が不要な場合に、推論に使用されるデータ型です。INT8データ型はFP32よりも精度が低い分、コンピューティング・サイクル当たりに処理できるINT8演算数が多くなります。
- BF16は、ほとんどの学習処理に十分な精度を提供するデータ型です。必要に応じて、推論の精度を高めることもできます。

この新しいタイル型アーキテクチャーにより、インテル® AMXは前世代と比較してパフォーマンスを大幅に向上させます。インテル® アドバンスト・ベクトル・エクステンション 512 ベクトル・ニューラル・ネットワーク・インストラクション（インテル® AVX-512 VNNI）を実行する第3世代インテル® Xeon® スケーラブル・プロセッサーと比較して、インテル® AMXを実行する第4世代インテル® Xeon® スケーラブル・プロセッサーでは、1サイクル当たり256回のINT8演算ではなく、2,048回のINT8演算を実行できます。また、図6に示すように、1サイクル当たり64回のFP32演算ではなく、1,024回のBF16演算を実行することも可能です。[7]

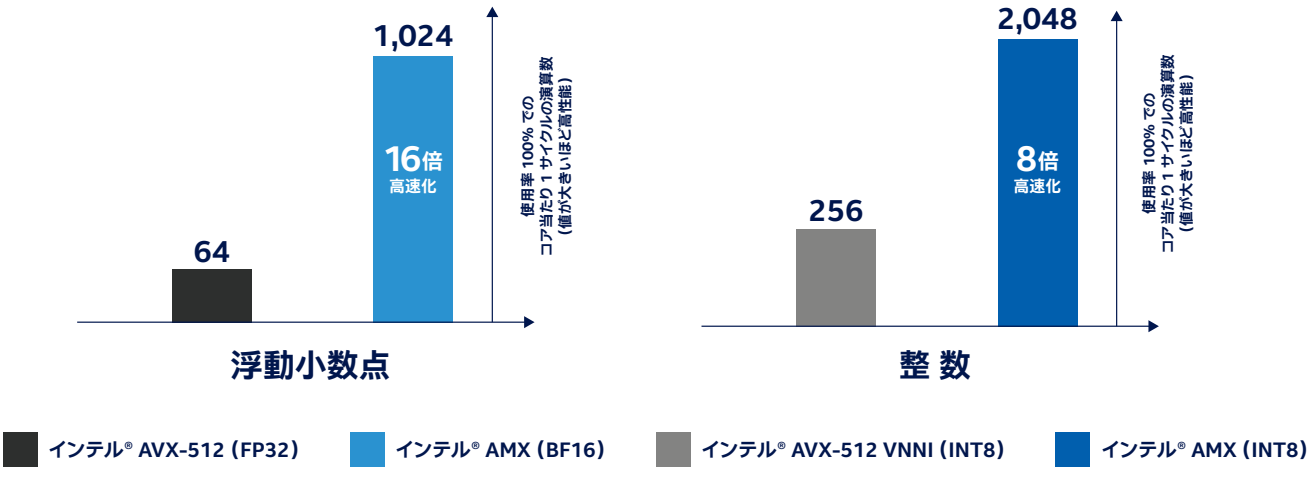

図6. インテル® AMXは、INT8/BF16データ型で、インテル® AVX-512VNNIよりも高いパフォーマンスを発揮[7]

AI のユースケース

インテル® AMX 搭載の第 4 世代インテル® Xeon® スケーラブル・プロセッサーは、幅広い DL ユースケースに導入できます。

レコメンドシステム
おすすめの映画や本の紹介、ターゲット広告の表示など、カスタマイズされたエンドユーザー体験を提供。リアルタイムのユーザー行動シグナルを反映し、時間や場所などのほぼリアルタイムのコンテキスト機能を備えた、DL ベースのレコメンドシステムを構築します。

自然言語処理(NLP)
言語推論やマシンラーニング (ML) を含む NLP アプリケーションは、2026 年までに世界の市場規模が 806 億 8,000 万ドルに達すると予測されており[8]、企業がセンチメント分析、チャットボット、機械翻訳などのさまざまな機能を採用 / 拡張するために不可欠となっています。

小売業の e コマース・ソフトウェア・ソリューション
AI に最適化された PyTorch や TensorFlow などのフレームワークに加え、DL 推論 / 学習処理を駆使し、トランザクション時間を最小限に抑え、ピーク需要に対応することで、収益の拡大と同時に、格別の顧客体験を提供します。

インテル® AMX の利用開始

インテル® AMX によってパフォーマンスを向上させるために必要な作業はほとんどありません。これは、インテル® oneAPI ディープ・ニューラル・ネットワーク (oneDNN) でデフォルトのフレームワークが最適化されているためです。Windows/Linux オペレーティング・システム、カーネルベースの仮想マシン (KVM)、広く利用されているハイパーバイザーは、インテル® AMX 命令セットを公開しています。TensorFlow や PyTorch といったオープンソースのフレームワークでは、INT8/BF16 演算が自動的に最適化されます。インテル® ディストリビューションの OpenVINO™ ツールキットを活用することで、コーディングの知識がほとんどなくても、AI 推論を自動化、最適化、調整し、実行できます。開発者が行う必要があるのは、インテル® ニューラル・コンプレッサーを使用して、学習処理モデルを INT8 データ型に量子化することだけです。

インテル® Xeon® スケーラブル・プロセッサーによる AI の高速化

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・プロセッサーに移行することで、まだ眠っている AI の可能性をビジネスで活用できます。すでにデータセンターで広く導入されているインテル® Xeon® スケーラブル・プロセッサーの幅広い基盤の上に構築される、高速化された全く新しい行列乗算演算で、AI 学習処理 / 推論の卓越したパフォーマンスを体験してください。

インテルの AI とインテル® AMX の詳細については、https://www.intel.co.jp/ai/ をご覧ください。

[1] https://www.intel.com/processorclaims/(英語) : 4th Generation Intel® Xeon® Scalable Processors の [A16] および [A17] を参照。結果は異なる場合があります。

[2] 「Top Artificial Intelligence (AI) Predictions For 2020 From IDC and Forrester」(Forbes、2019 年 11 月)。https://www.forbes.com/sites/gilpress/2019/11/22/top-artificial-intelligence-ai-predictions-for-2020-from-idc-and-forrester/#4fef9821315a/(英語)

[3] 「With AMX, Intel Adds AI/ML Sparkle to Sapphire Rapids」(The Next Platform、2021 年 8 月)。https://www.nextplatform.com/2021/08/19/with-amx-intel-adds-ai-ml-sparkle-to-sapphire-rapids/(英語)

[4] AI 推論ワークロードを実行するデータセンター・サーバーを対象に世界全体のインストール・ベースでインテルが実施した市場モデリングに基づく(2021 年 12 月時点)。

[5] PyTorch モデルのパフォーマンス構成。**PT-NLP BERT-large : 8480 :** 1 ノード、インテル® Xeon® Platinum 8480 プロセッサー x2 を搭載した製品出荷前のプラットフォーム、総メモリー容量 1,024GB(16 スロット /64GB/DDR5-4800)、ucode 0x2b0000a1、インテル® ハイパースレッディング・テクノロジー(インテル® HT テクノロジー)有効、インテル® ターボ・ブースト・テクノロジー有効、CentOS Stream 8、5.15.0、1TB インテル® SSDSC2KW256G8(PT)/Samsung SSD 860 EVO(TF)x1、BERT-large、Inf : SQuAD1.1 (seq len=384)、bs=1(インスタンス当たり 4 コア)、bs=n(インスタンス当たり 1 ソケット)、bs : FP32=1,56、インテル® AMX BF16=1,16、インテル® AMX INT8=1,56、Trg : Wikipedia 2020/01/01 (seq len=512)、bs : FP32=28、インテル® AMX BF16=56(1 インスタンス、1 ソケット)、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**8380 :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2、総メモリー容量 1,024GB(16 スロット /64GB/DDR4-3200)、ucode 0xd000375、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 22.04 LTS、5.15.0-27-generic、インテル® SSDSC2KG960G8 x1、BERT-large、Inf : SQuAD1.1 (seq len=384)、bs=1(インスタンス当たり 4 コア)、bs=n(インスタンス当たり 1 ソケット)、bs : FP32=1,56、INT8=1,56、Trg : Wikipedia 2020/01/01 (seq len=512)、bs : FP32=28、インテル® AMX BF16=56(1 インスタンス、1 ソケット)、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**PT-DLRM : 8480 :** 1 ノード、インテル® Xeon® Platinum 8480 プロセッサー x2 を搭載した製品出荷前のプラットフォーム、総メモリー容量 1,024GB(16 スロット /64GB/DDR5-4800)、ucode 0x2b0000a1、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、CentOS Stream 8、5.15.0、1TB インテル® SSDSC2KW256G8(PT)/Samsung SSD 860 EVO(TF)x1、DLRM、推論 : bs=n(インスタンス当たり 1 ソケット)、bs : FP32=128、インテル® AMX BF16=128、インテル® AMX INT8=128、学習処理 bs : fp32/ インテル® AMX BF16=32K(1 インスタンス、1 ソケット)、テラバイトの Criteo データセット、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**8380 :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2、総メモリー容量 1,024GB(16 スロット /64GB/DDR4-3200)、ucode 0xd000375、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 22.04 LTS、5.15.0-27-generic、インテル® SSDSC2KW256G8(PT)/Samsung SSD 860 EVO(TF)x1、DLRM、推論 : bs=n(インスタンス当たり 1 ソケット)、bs : FP32=128、INT8=128、学習処理 bs : FP32=32K(1 インスタンス、1 ソケット)、テラバイトの Criteo データセット、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**PT-ResNet-34 : 8480 :** 1 ノード、インテル® Xeon® Platinum 8480 プロセッサー x2 を搭載した製品出荷前のプラットフォーム、総メモリー容量 1,024GB (16 スロット /64GB/DDR5-4800)、ucode 0x2b0000a1、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、CentOS Stream 8、5.15.0、1TB インテル® SSDSC2KW256G8 (PT) /Samsung SSD 860 EVO (TF) x1、SSD-ResNet-34、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,112、インテル® AMX BF16=1,112、インテル® AMX INT8=1,112、学習処理 bs : FP32/ インテル® AMX BF16=224 (1 インスタンス、1 ソケット)、Coco 2017、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**8380 :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2、総メモリー容量 1,024GB (16 スロット /64GB/DDR4-3200)、ucode 0xd000375、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 22.04 LTS、5.15.0-27-generic、インテル® SSDSC2KG960G8 x1、SSD-ResNet-34、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,112、INT8=1,112、学習処理 bs : FP32=224 (1 インスタンス、1 ソケット)、Coco 2017、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**PT-ResNet-50 : 8480 :** 1 ノード、インテル® Xeon® Platinum 8480 プロセッサー x2 を搭載した製品出荷前のプラットフォーム、総メモリー容量 1,024GB (16 スロット /64GB/DDR5-4800)、ucode 0x2b0000a1、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、CentOS Stream 8、5.15.0、1TB インテル® SSDSC2KW256G8 (PT) /Samsung SSD 860 EVO (TF) x1、ResNet-50 v1.5、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,64、インテル® AMX BF16=1,64、インテル® AMX INT8=1,116、学習処理 bs : FP32、インテル® AMX BF16=128 (1 インスタンス、1 ソケット)、ImageNet (224 x 224)、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**8380 :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2、総メモリー容量 1,024GB (16 スロット /64GB/DDR4-3200)、ucode 0xd000375、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 22.04 LTS、5.15.0-27-generic、インテル® SSDSC2KG960G8 x1、ResNet-50 v1.5、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,64、インテル® AMX BF16=1,64、インテル® AMX INT8=1,116、学習処理 bs : FP32=128 (1 インスタンス、1 ソケット)、ImageNet (224 x 224)、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**PT-RNN-T : 8480 :** 1 ノード、インテル® Xeon® Platinum 8480 プロセッサー x2 を搭載した製品出荷前のプラットフォーム、総メモリー容量 1,024GB (16 スロット /64GB/DDR5-4800)、ucode 0x2b0000a1、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、CentOS Stream 8、5.15.0、1TB インテル® SSDSC2KW256G8 (PT) /Samsung SSD 860 EVO (TF) x1、Resnext101 32x16d、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,64、インテル® AMX BF16=1,64、インテル® AMX INT8=1,116、ImageNet、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**8380 :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2、総メモリー容量 1,024GB (16 スロット /64GB/DDR4-3200)、ucode 0xd000375、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 22.04 LTS、5.15.0-27-generic、インテル® SSDSC2KG960G8 x1、Resnext101 32x16d、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,64、INT8=1,116、ImageNet、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**PT-ResNext101 : 8480 :** 1 ノード、インテル® Xeon® Platinum 8480 プロセッサー x2 を搭載した製品出荷前のプラットフォーム、総メモリー容量 1,024GB (16 スロット /64GB/DDR5-4800)、ucode 0x2b0000a1、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、CentOS Stream 8、5.15.0、1TB インテル® SSDSC2KW256G8 (PT) /Samsung SSD 860 EVO (TF) x1、Resnext101 32x16d、bs=n (インスタンス当たり 1 ソケット)、推論 : bs : FP32=1,64、インテル® AMX BF16=1,64、インテル® AMX INT8=1,116、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**8380 :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2、総メモリー容量 1,024GB (16 スロット /64GB/DDR4-3200)、ucode 0xd000375、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 22.04 LTS、5.15.0-27-generic、インテル® SSDSC2KG960G8 x1、Resnext101 32x16d、bs=n (インスタンス当たり 1 ソケット)、推論 : bs : FP32=1,64、INT8=1,116、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**PT-MaskRCNN : 8480 :** 1 ノード、インテル® Xeon® Platinum 8480 プロセッサー x2 を搭載した製品出荷前のプラットフォーム、総メモリー容量 1,024GB (16 スロット /64GB/DDR5-4800)、ucode 0x2b0000a1、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、CentOS Stream 8、5.15.0、インテル® SSDSC2KW256G8 (PT) /Samsung SSD 860 EVO (TF) x1、MaskRCNN、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,112、インテル® AMX BF16=1,112、学習処理 bs : FP32/ インテル® AMX BF16=112 (1 インスタンス、1 ソケット)、Coco 2017、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。**8380 :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2、総メモリー容量 1,024GB (16 スロット /64GB/DDR4-3200)、ucode 0xd000375、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 22.04 LTS、5.15.0-27-generic、インテル® SSDSC2KG960G8 x1、MaskRCNN、推論 : bs=1 (インスタンス当たり 4 コア)、bs=n (インスタンス当たり 1 ソケット)、bs : FP32=1,112、インテル® AMX BF16=1,112、学習処理 bs : FP32=112 (1 インスタンス、1 ソケット)、Coco 2017、フレームワーク : https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66/、ModelZoo : https://github.com/IntelAI/models/tree/spr-launch-public/ 、PT : 1.13、IPEX : 1.13、oneDNN : v2.7 (2022 年 10 月 24 日に実施したインテル社内テストで測定)。推論 : ResNet-50 v1.5 : ImageNet (224 x 224)、SSD ResNet-34 : Coco 2017 (1200 x 1200)、BERT-large : SQuAD1.1 (seq len=384)、Resnext101 : ImageNet、Mask RCNN : COCO 2017、DLRM : テラバイトの Criteo データセット、RNNT : LibriSpeech。学習処理 : ResNet-50 v1.5 : ImageNet (224 x 224)、SSD ResNet-34 : COCO 2017、BERT-large : Wikipedia 2020/01/01 (seq len=512)、DLRM : テラバイトの Criteo データセット、RNNT : LibriSpeech、Mask RCNN : COCO 2017。

[6] **INT8 測定のソフトウェア構成 :** TensorFlow ResNet-50 v1.5、推論 : BS=116 (INT8)、ソケット当たり 1 インスタンス。oneDNN v2.7、インテルが最適化した TensorFlow 2.10。2022 年 10 月 24 日(第 3 世代および第 4 世代インテル® Xeon® スケーラブル・プロセッサー)、2022 年 7 月 19 日(第 2 世代および初代のインテル® Xeon® スケーラブル・プロセッサー) に実施したインテル社内テストで測定。**ハードウェア構成 : 第 4 世代インテル® Xeon® スケーラブル・プロセッサーのハードウェア構成(測定済み) :** 2 ソケットのインテル® Xeon® Platinum 8480 プロセッサー(56 コア、350W 熱設計電力(TDP)) を搭載した製品出荷前のプラットフォーム、DDR5 総メモリー容量 1TB(8 チャネル /64GB/4,800MHz)、BKC 01 を使用、インテル® AMX(INT8、BF16) を使用、CentOS Stream 8、インテル® AMX カーネル (5.15)、測定値は異なる場合があります。**第 3 世代インテル® Xeon® スケーラブル・プロセッサーのハードウェア構成(測定済み) :** 1 ノード、インテル® Xeon® Platinum 8380 プロセッサー x2 (40 コア /2.30GHz、270W TDP)、DDR4 総メモリー容量 1TB (8 スロット /64GB/3,200MHz)、ucode 0xd0002f2、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、Ubuntu 20.04.2 LTS (Focal Fossa)、5.4.0-73-generic、インテル® SSDSC2CW480A3 OS ドライブ x1。**第 2 世代インテル® Xeon® スケーラブル・プロセッサーのハードウェア構成(測定済み) :** 1 ノード、2 ソケットのインテル® Xeon® Platinum 8280 プロセッサー(28 コア)、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、総メモリー容量 384GB (12 スロット /32GB/2,933MHz)、BIOS : SE5C620.86B.02.01.0013.121520200651 (ucode : 0x500320a)、CentOS Stream 8、4.18.0-383.el8.x86_64。**インテル® Xeon® スケーラブル・プロセッサーのハードウェア構成(測定済み) :** 1 ノード、2 ソケットのインテル® Xeon® Platinum 8180 プロセッサー(28 コア)、インテル® HT テクノロジー有効、インテル® ターボ・ブースト・テクノロジー有効、総メモリー容量 384GB (12 スロット /32GB/2,666MHz)、BIOS : SE5C620.86B.0X.01.0117.021220182317 (ucode : 0x2006b06)、Ubuntu 20.04.2 LTS、5.4.0-73-generic。

[7] CPU 使用率 100% を想定した、コア当たりの 1 サイクルの行列乗算演算 + 累積演算のピーク・アーキテクチャー能力に基づく。2021 年 8 月時点。完全なワークロードと構成の詳細については、https://www.intel.com/PerformanceIndex/(英語) : Architecture Day 2021 を参照。結果は異なる場合があります。

[8] 世界の NLP 市場規模の出典 :「Natural Language Processing (NLP) Market Size, Share & COVID-19 Impact Analysis, By Deployment (On-Premises, Cloud, Hybrid), By Enterprise Size (SMEs, and Large Enterprises), By Technology (Interactive Voice Response (IVR), Optical Character Recognition (OCR), Text Analytics, Speech Analytics, Classification and Categorization), By Industry Vertical (Healthcare, Retail, High Tech, and Telecom, BFSI) and Regional Forecast, 2021-2028」(Fortune Business Insights、2021 年 6 月)。https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933#(英語)

インテル株式会社

〒100-0005 東京都千代田区丸の内 3-1-1
http://www.intel.co.jp/

2023 年 8 月

353107-001JA
JPN/2308/PDF/SE/MKTG/TK