ソリューション概要

ディープラーニング・モデル
第４世代インテル® Xeon® スケーラブル・プロセッサー

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）で Alibaba Cloud Address Purification の AI 推論性能を拡張

インテル® AMX を組み込んだ第４世代インテル® Xeon® スケーラブル・プロセッサーが、エンドツーエンドの推論パフォーマンスを前世代と比べて 2.48 倍高速化。[1]

ディープラーニング（DL）は、人工知能（AI）の中でも重要な技法の１つとして、コンピューター・ビジョン（CV）、自然言語処理（NLP）、レコメンドシステムなど、幅広い分野で利用されるようになっています。しかし、データの爆発的な増加と、ますます複雑化する DL モデルとが相まって、本番環境での推論実行にさまざまな課題が生じていることも否定できません。ユーザーが期待するのは、最適化されたハードウェア、ソフトウェア、アルゴリズムによる、パフォーマンスの向上と総コストの削減です。DL 推論を最適化することで、ユーザーは複雑化する DL モデルを採用し、レイテンシーは同等レベルに維持したまま、精度を高めることができます。

Alibaba Cloud では、アドレス浄化サービスのパフォーマンスを向上させるべく、AI 専用マシンラーニング・プラットフォーム（PAI）と技術研究開発機構 DAMO アカデミーの NLP チームがインテルとの協働を開始しました。第４世代インテル® Xeon® スケーラブル・プロセッサーにインテル® AMX が加わり、最適化ツールと併用することで、エンドツーエンドの推論実行を前世代のプラットフォームと比べて最大 2.48 倍に高速化しています。[1]

Alibaba Cloud Address Purification

アドレス浄化とは、郵送物の宛先住所の標準化、訂正、検証を自動化するプロセスで、輸送 / 物流、e コマース、小売、金融サービスと、幅広い業種で利用されます。Alibaba Cloud Address Purification は、アドレスの効率化 / 標準化を提供する Algorithm as a Service（AaaS）であり、Alibaba Cloud の膨大なアドレス収集をベースに、DAMO Academy の NLP チームが開発しました。[2] エンドツーエンド・パフォーマンスの高速化は、Alibaba Cloud を利用するお客様にとってのビジネス成果に直結します。この AaaS は、クローズドループで制御できるワンストップのサービス・プラットフォームとして、住所データを処理します。NLP アルゴリズムを使用し、業務システムに登録されている住所データの訂正、入力補完、正規化、構造化、ラベル付けを行います。20 タイプ以上のアドレスサービスを展開し[3]、パブリック、プライベート、ハイブリッドとクラウドの形態を問わず導入が可能です。Alibaba Cloud は、次のようなことを目的としています。

- データ洗浄からモデル推論まで、多様なワークロードを総合的に評価して、プラットフォームのワンストップ・パフォーマンスを加速する。
- 既存のハードウェア・リソースを効果的に活用し、パブリック、プライベート、ハイブリッドと形態の異なるクラウド環境に広がる顧客サーバーのリソースをフル活用して、ハードウェアにかかるコストを削減する。

Alibaba のサービスを最適化するインテルのテクノロジー

Transformer による双方向のエンコード表現（BERT）モデルは、AI プログラムが複数の意味を持つテキスト内の語句からコンテキストを読み取れるようにする、自然言語処理（NLP）に用いるディープラーニング手法の１つです。Alibaba では BERT モデルを Address Purification サービスの検索モジュールとして採用し[4]、マルチタスクのベクトル再現と精細な分類処理に使用しています。インテルは、このソリューションのパフォーマンスを飛躍的に加速できる、多様なソリューションを提供しています。

インテル® AMX

第 4 世代インテル® Xeon® スケーラブル・プロセッサーにはインテル® AMX と呼ばれるアクセラレーターが組み込まれ、これが Alibaba Cloud Address Purification サービスの傑出したパフォーマンス、コスト効率、拡張性を実現します。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、レコメンドシステムから、自然言語処理、小売の e コマース・ソフトウェア・ソリューションと、幅広い DL ユースケースへの導入が可能です。

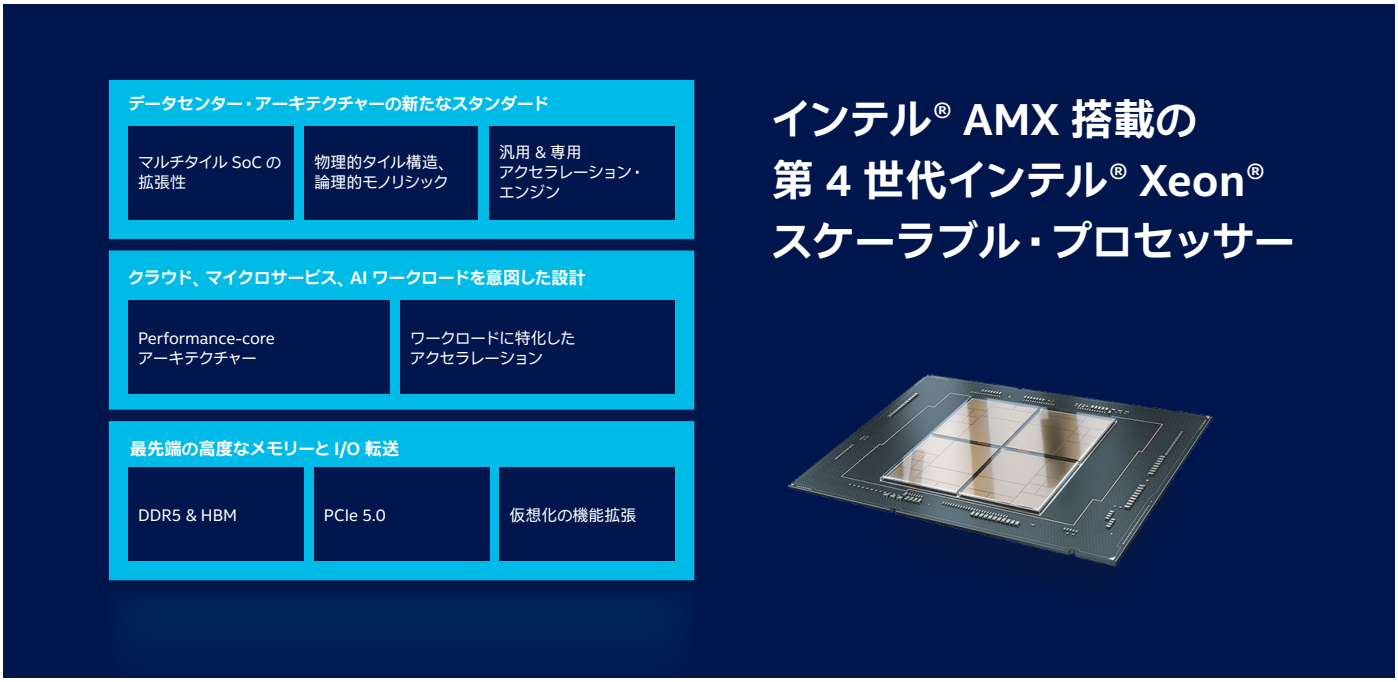

図 1. インテル® AMX 搭載の第 4 世代インテル® Xeon® スケーラブル・プロセッサー

図 2. インテル® AMX の全体像

Blade : 推論を最適化する汎用ツール

Alibaba Cloud Address Purification ソリューションでは、Alibaba Cloud マシンラーニング PAI チームが提供する、推論実行の最適化を図る汎用ツール「Blade」を採用し、アドレス浄化の推論パフォーマンスを最適化しています。Blade には数多くの最適化メソッドが統合され、計算グラフの最適化、インテル® oneAPI ディープ・ニューラル・ネットワーク (インテル® oneDNN) 最適化ライブラリー、BladeDISC コンパイラー、Blade ハイパフォーマンス演算子ライブラリー、インテルのカスタム・バックエンド、Blade の混合精度なども含まれます。

インテルのカスタム・バックエンドを Blade に統合

インテルのカスタム・バックエンド[5] は、Blade のソフトウェア・バックエンドとして、量子化とスパース化の面から推論モデルのパフォーマンスを加速します。このカスタム・バックエンドでは大きく分けて 3 レベルの最適化を行い、初めに基本的なキャッシュ戦略を適用してメモリーを最適化し、次にグラフ融合を最適化、最後に演算子レベルで、カスタムのスパースカーネルを含む効率的な演算子ライブラリーを構築します。

図 3. インテルのカスタム・バックエンドの構造

インテル® Xeon® スケーラブル・プロセッサーでは第 2 世代以降、INT8 量子化を目的にベクトル・ニューラル・ネットワーク命令 (VNNI) が実装され、INT8 データ型ベースの AI パフォーマンスを最適化する、モデル量子化ソリューションに広く採用されるようになりました。

インテル® AMX は、インテル® oneDNN を活用することで、INT8 機能を大幅に拡張します。インテル® AMX ベースの INT8 量子化では、VNNI をはるかに上回るモデル性能の向上が可能です。図 4 は、インテル® AMX が機能する仕組みを示しています。

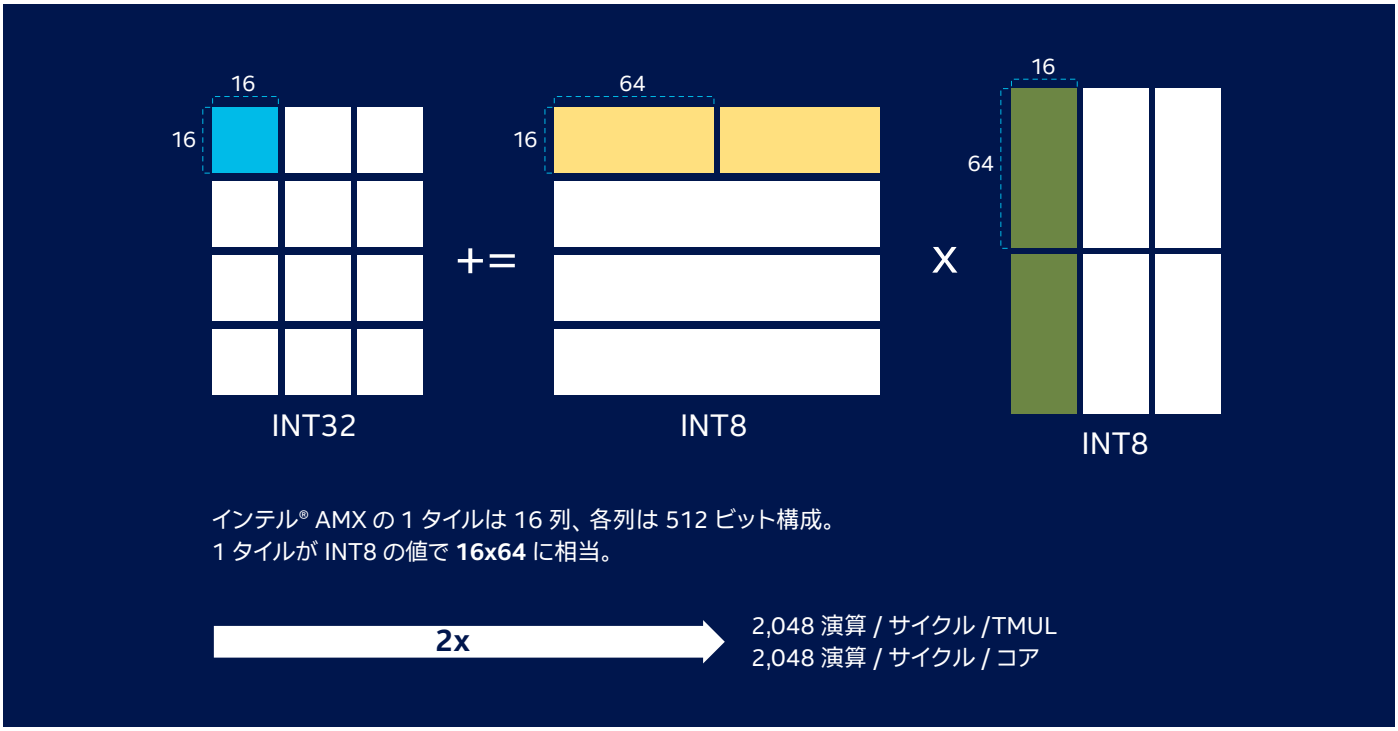

図 4. インテル® アドバンスト・マトリクス・エクステンション (インテル® AMX) の機能

パフォーマンスと最適化の拡張

Alibaba Cloudとインテルはさらに、推論パフォーマンスの向上を目的にアドレス浄化モデルをチューニング。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、前世代よりも PAI のパフォーマンスは 2.48 倍向上します。[1] インテル® AMX ベースのカスタム・バックエンドは、形状サイズが固定された(10x53)、4 レイヤーの BERT モデルを最適化することで、この高速化を実現しました(図 5 を参照)。

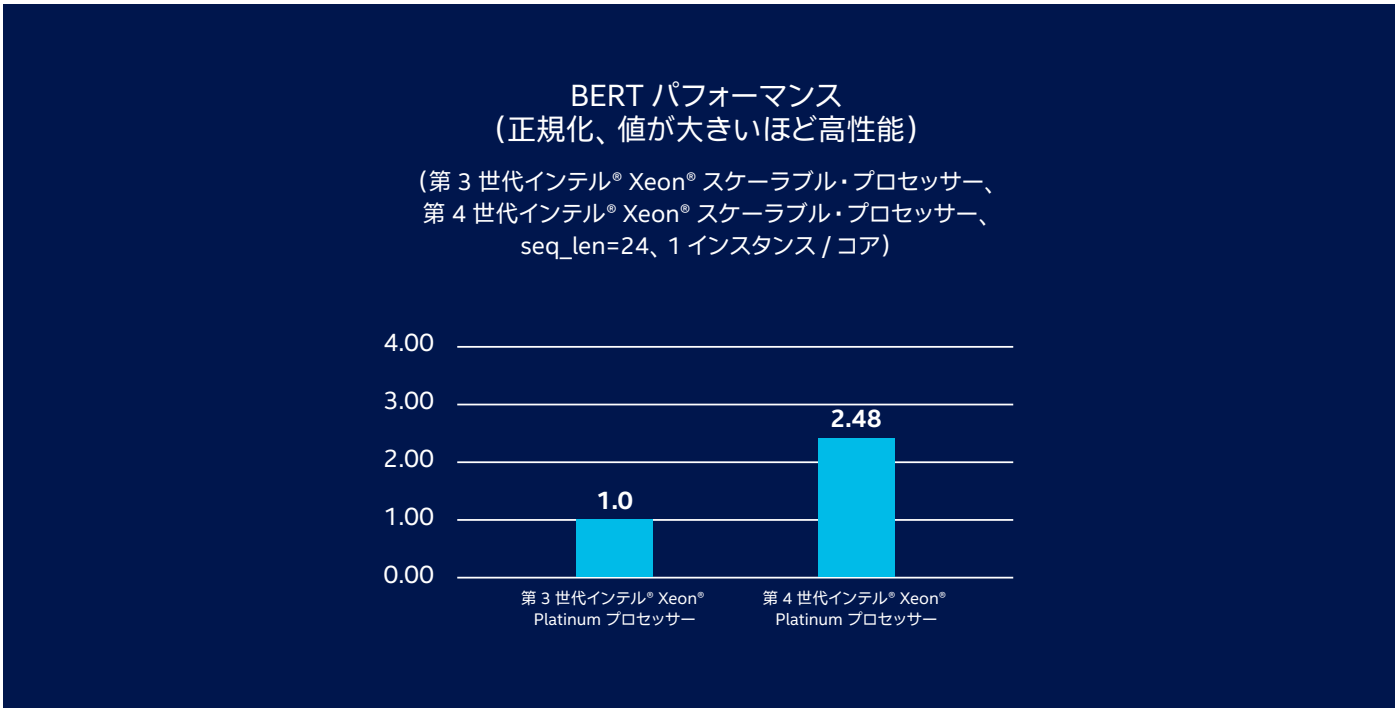

図 5. BERT モデルの推論パフォーマンス[1]

CCKS2021 中国語 NLP アドレス相関タスクを実行し、モデルの検証を行いました。浮動小数点 32 ビット (FP32)ベースの最適化ではスコア 78.72、INT8 ベースの最適化ではスコア 78.85 の評価となっています(値が大きいほど高性能)。[6]

まとめ

Alibaba Cloud は、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・プロセッサーを導入し、アドレス浄化サービスの AI 推論を最適化しました。エンドツーエンド・パフォーマンスの高速化は、輸送 / 物流、e コマース、エネルギー、小売、金融と幅広い業種にわたる、Alibaba のサービスを利用する顧客にとってのビジネス成果に直結します。またインテル® AMX は、Alibaba が独立型 GPU などの専用アクセラレーターを導入していたとしたら発生するであろうと考えられるオーバーヘッドも低減します。Alibaba では、内蔵アクセラレーターを使用することで、アドレス浄化サービスにおける総保有コスト (TCO) の制御が可能になりました。

インテルと Alibaba は、新たな DL モデルのエンドツーエンド・パフォーマンスを拡大するために、顧客との連携を強化し、ソフトウェアとハードウェアの統合を最適化しています。最終的な目標は、DL モデルのパフォーマンスを加速し、インテルのテクノロジーに備わる価値を最大限まで引き出すことです。インテルはさらに業界パートナーとの協力関係を深め、AI テクノロジーの導入と実装に貢献していきたいと考えています。

[1] システム構成 : 基準 : インテルが実施した 2022 年 10 月 19 日時点のテスト結果。1 ノード、第 3 世代インテル® Xeon® Platinum プロセッサー x2、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、総メモリー容量 256GB(16 スロット x16GB、3,200MT/s [動作周波数 @3,200MHz])、WLYDCRB1.SYS.0029.P30.2209011945、0xd00037b、CentOS Linux 8、4.18.0-305.12.1.el8_4.x86_64、GCC 8.5.0、NLP ツールキット v0.3、Pytorch 1.11、BERT-mini、INC 1.13、Transformer 4.18.0、1 インスタンス / コア、BS=32、seq_len=24、データ型 : INT8。

比較対象の新システム -1 : インテルが実施した 2022 年 10 月 19 日時点のテスト結果。1 ノード、第 4 世代インテル® Xeon® Platinum プロセッサー x2、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 256GB(16 スロット x16GB、4,800MHz [動作周波数 @4,800MHz])、EGSDCRB1.SYS.0090.D03.2210040200、0x2b0000c0、CentOS Stream 8、5.15.0-spr.bkc.pc.8.8.5.x86_64、GCC 8.5.0、NLP ツールキット v0.3、Pytorch 1.11、BERT-mini、INC 1.13、Transformer 4.18.0、1 インスタンス / コア、BS=32、seq_len=24、データ型 : INT8。

[2] Alibaba Cloud、「Address Normalization」https://cn.aliyun.com/product/addresspurification/addrp/(中国語)

[3] Alibaba Cloud、「What is Address Normalization?」https://help.aliyun.com/document_detail/169746.html(中国語)

[4] Devlin ほか、「BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding」。ACL Anthology、2019 年 6 月。https://aclanthology.org/N19-1423/(英語)

[5] GitHub、「Intel Neural Compressor」https://github.com/intel/neural-compressor/commits/inc_with_engine/(英語)

[6] Alibaba Cloud、「Intel Innovation Master Cup' Deep Learning Challenge Track 3: CCKS2021 Chinese NLP Address Correlation Task」、2021 年 11 月。https://tianchi.aliyun.com/competition/entrance/531901/introduction/(中国語)

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

356217-001JA
JPN/2308/PDF/SE/MKTG/TK