

インテル® アドバンスド・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きました。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。

データセンター・アーキテクチャーの新たなスタンダード

- マルチタイル SoC の拡張性
- 物理的タイル構造、論理的モノリシック
- 汎用 & 専用 アクセラレーション・エンジン

クラウド、マイクロサービス、AI ワークロードを意図した設計

- Performance-core アーキテクチャー
- ワークロードに特化したアクセラレーション

最先端の高度なメモリ & I/O 転送

- DDR5 & HBM
- PCIe 5.0
- 仮想化の機能拡張

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

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

インテル® AMX

新たに追加された内蔵アクセラレーション・エンジン

第2世代インテル® Xeon® スケーラブル・プロセッサ	第3世代インテル® Xeon® スケーラブル・プロセッサ	第4世代インテル® Xeon® スケーラブル・プロセッサ
インテル® ディープラーニング・ブースト (初実装) インテル® アドバンスド・ベクトル・エクステンション 512 (インテル® AVX-512) (VNNI/INT8)	インテル® DL ブースト インテル® AVX-512 : VNNI/INT8 (CPX/ICX) & BFloat16 (CPX)	インテル® AMX INT8 & BFloat16 対応 インテル® AVX-512 (VNNI/INT8)

主なメリット

- 広範なハードウェア (専用のシリコン / タイル、マトリクス乗算命令セット / TMUL) とソフトウェア (関連市場を横断したフレームワーク、ツールキット、ライブラリー) の最適化による、インテル® Xeon® スケーラブル・プロセッサに搭載された AI アクセラレーションの機能拡張
- インテル® AMX でサポートするデータ型 : INT8 (推論)、BFloat16 (学習処理 / 推論)

対象のワークロード / 用途

- 画面認識
- 機械翻訳
- 自然言語処理 (NLP)
- メディア分析
- レコメンドシステム
- 強化学習
- メディア処理 / 配信

もたらされる効果

- 前世代のインテル® Xeon® スケーラブル・プロセッサと比べて、AI / ディープラーニングの推論や学習処理ワークロードのパフォーマンスを大幅に向上

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。

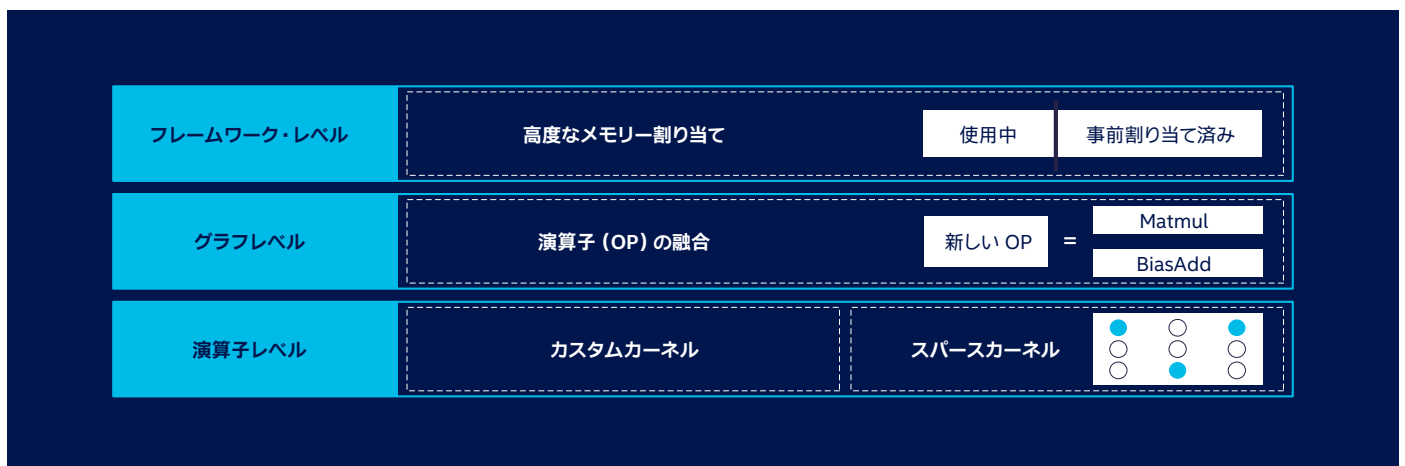


図3. インテルのカスタム・バックエンドの構造

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。

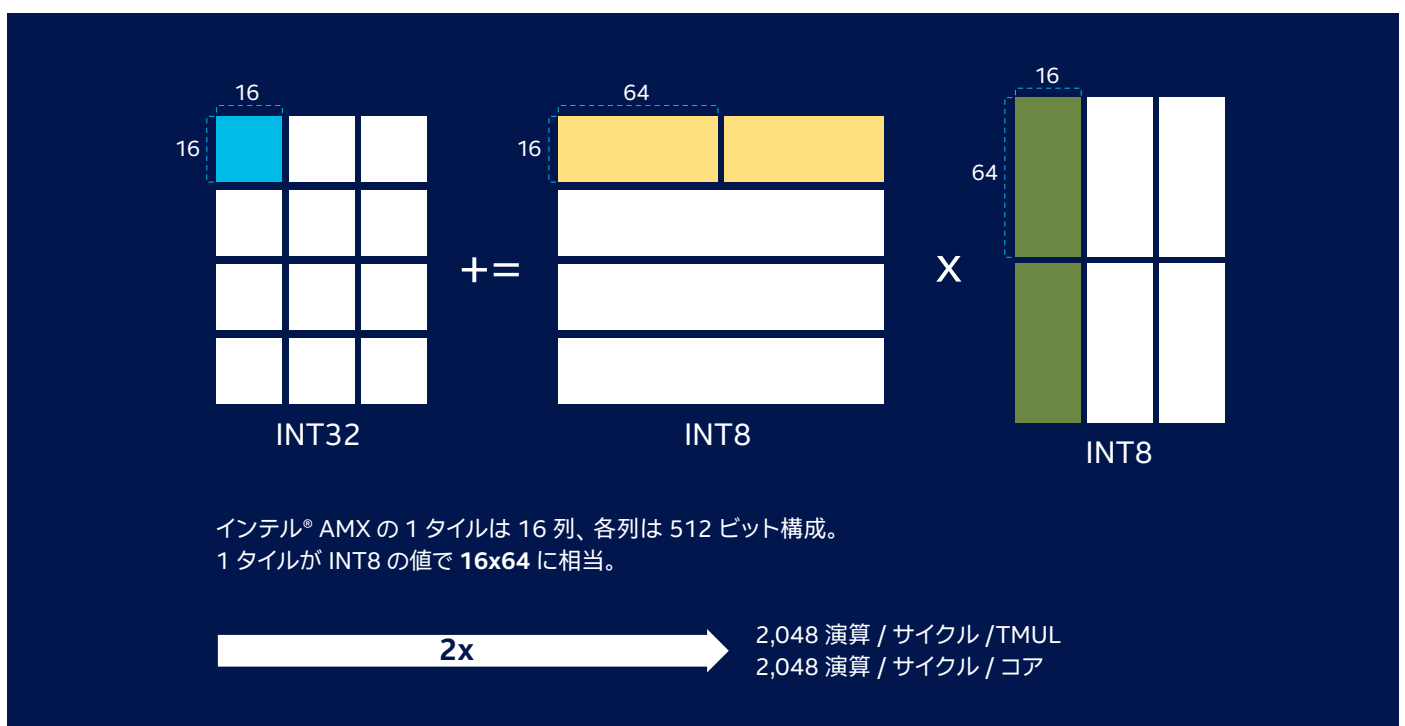


図4. インテル® アドバンスド・マトリクス・エクステンション (インテル® AMX) の機能

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)。

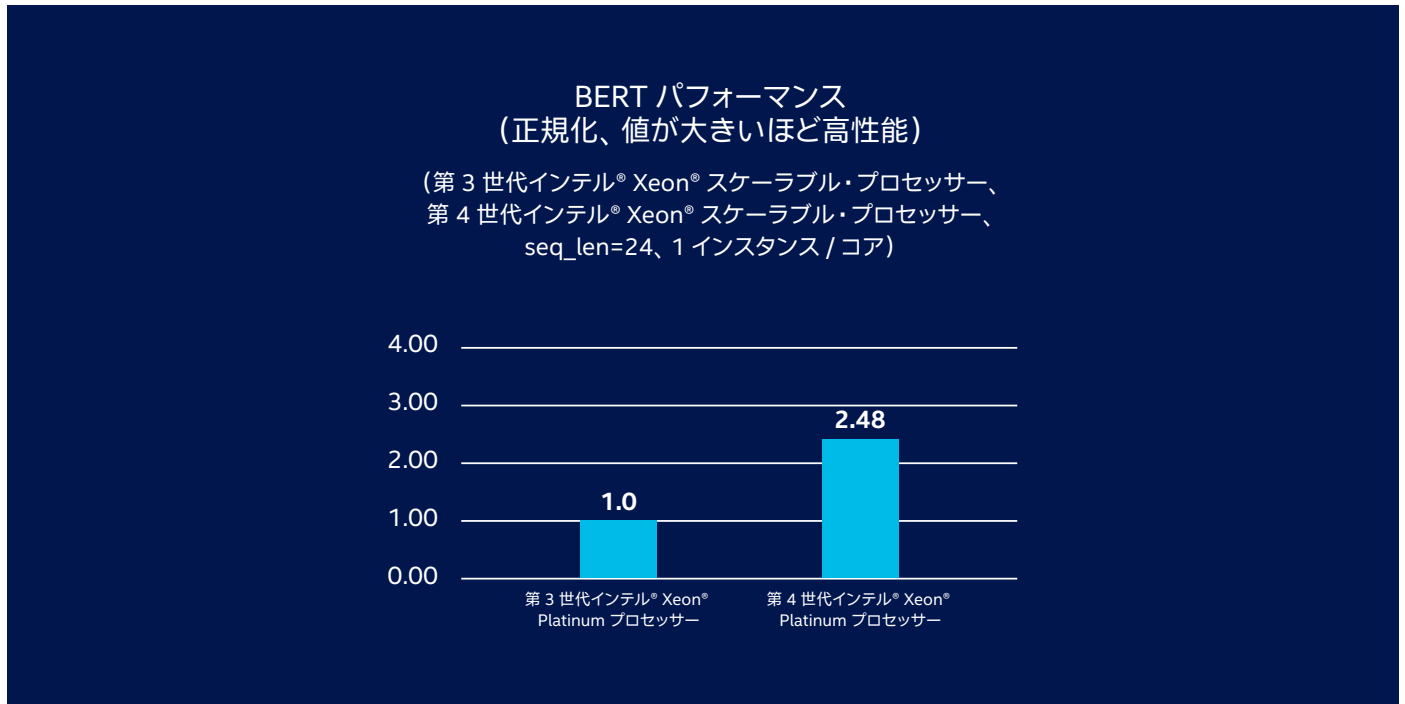


図5. BERTモデルの推論パフォーマンス¹

CCKS2021 中国語 NLP アドレス関連タスクを実行し、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 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 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)

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、<https://www.intel.com/PerformanceIndex/> (英語) を参照してください。

性能の測定結果は、構成に示されている日付時点のテストに基づいています。また、現在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。

実際のコストや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。

インテルは、明示されているか否かにかかわらず、いかなる保証もいたしません。ここにいる保証には、商品適格性、特定目的への適合性、および非侵害性の黙示の保証、ならびに履行の過程、取引の過程、または取引での使用から生じるあらゆる保証を含みますが、これらに限定されるわけではありません。

インテルは、サードパーティーのデータについて管理や監査を行っていません。ほかの情報も参考にしてデータの正確さを評価してください。

Intel、インテル、Intel ロゴ、その他のインテルの名称やロゴは、Intel Corporation またはその子会社の商標です。

その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

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 Intel Corporation. 無断での引用、転載を禁じます。

2023年8月

356217-001JA

JPN/2308/PDF/SE/MKTG/TK