ソリューション概要

マシンラーニング
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）によるマシンラーニング（ML）モデルの最適化

BERT（Bidirectional Encoder Representations from Transformers）モデルのスループットは、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 を使用すると、前世代と比較してパフォーマンスが2倍から3倍向上します。[1、2]

このソリューション概要では、融合最適化という手法を導入するためのプロキシーモデルとして、レイヤー数12、隠れ層の次元768、ヘッド数12、シーケンス長128（トークンサイズ）の標準 BERT モデルを使用しています。

概 要

BERT（Bidirectional Encoder Representations from Transformers）は、自然言語処理（NLP）で広く使用されている ML モデルとその技術です。登場以来、BERT は膨大な数の NLP タスクのレコードを更新するために利用されてきました。また、実際のコアバウンド・アプリケーションでも優れたパフォーマンスを発揮します。

BERT は、検索、機械翻訳、マンマシン・インタラクション、その他の NLP タスクを対象に、多岐にわたるユーザーシナリオで採用されています。BERT のパフォーマンスが向上すればアプリケーションのユーザー・エクスペリエンスに直接影響するだけでなく、1秒当たりのクエリー数（QPS）におけるスループット・レートが上昇するため、エンジニアは、モデルを最適化してパフォーマンスを向上させるさまざまな方法が検討されてきました。

例えばテンセントの StarLake Lab では、高度なクラウド・コンピューティング、人工知能（AI）、セキュリティー、ストレージ、ネットワーク・テクノロジーを調査し、データセンターのパフォーマンスを向上させると同時に、データセンターの総保有コスト（TCO）を削減するソリューションを研究しています。テンセントの Machine Learning Platform Department（MLPD）は、AI プラットフォームの中核部門として、同社のインターネットやテクノロジー・ビジネス全体でイノベーションを推進する取り組みを常時行っています。MLPD は、コンピューター・ビジョン、音声認識、グラフ計算、NLP など、幅広い分野に及ぶ研究開発を行っており、開発されたソリューションは、ソーシャルメディア、パーソナライズド広告、ゲーミング AI、レコメンド機能、検索の主要なシナリオに広く使われています。これらのさまざまな技術領域のアプリケーションにおいて、BERT が重要な役割を果たしています。

インテルは、テンセントの MLPD や StarLake ラボラトリーと緊密に協力し、第4世代インテル® Xeon® スケーラブル・プロセッサーの内蔵アクセラレーターであるインテル® AMX を活用した BERT 推論の最適化に取り組んでいます。インテル® AMX 内蔵の第4世代インテル® Xeon® スケーラブル・プロセッサーを搭載したシステムで BERT モデルを実行したところ、BERT モデルでの [INT8] のスループットが2倍、[BF16] のスループットが3倍に向上する可能性があることが実証されました。[1、2] インテル® AMX とソフトウェアの最適化を組み合わせた統合ソリューションにより、テンセントは一貫したサービス体験を提供する機能や TCO を最適化する機能の進化を目指しています。

テンセントのソーシャル・アプリケーション最適化

テンセントのソーシャル・アプリケーションは、世界で10億人を超えるアクティブユーザーをつないでいます。テンセントのソーシャル・アプリケーションの中で最も人気のあるものは2011年にリリースされ、2018年には世界最大のスタンドアロン・モバイル・アプリになりました。このアプリにはメッセージ交換、ネット通話、一対多の動画配信、オンラインゲーム、ビデオ会議などの機能が揃い、さまざまな用途で使えることから「中国での生活はこれだけで完結する」と言われています。さらに、写真共有や動画共有、位置情報共有機能も備えています。

このアプリのユーザーの大部分は、メッセージや記事、ミニプログラム、ショート動画、音楽などの人気コンテンツを探す際に、アプリに搭載された検索エンジンを使用する傾向にあります。検索エンジンにとっての主な課題は、いかに大規模なクエリーを処理し、検索結果に素早く対応できるかです。このような課題のソリューションとして、インテル® AMX を利用した深い最適化により TCO を削減すると同時に、この検索エンジンで使用していた既存の汎用インフラストラクチャーを活用し、アプリケーションの検索体験全体の底上げができると考えられています。

融合最適化

従来の融合最適化は、BERT ベースのモデルの12層のレイヤーを1つの大規模なオペレーション（op、演算処理）に融合するという FP32 のソリューションを通じて実現されていました。現在は、インテル® AMX によりさらに深い融合最適化に必要な機能が提供されています。

MatMul や BatchMatMul のオペレーションは、BERT 側の時間を大量に消費する可能性があるため、MatMul と BatchMatMul を最適化することがパフォーマンス向上のポイントとなります。過去の経験から計算処理やメモリーアクセスを削減すれば、パフォーマンスを最適化できます。モデルから不要なオペレーションを削除すれば、計算量も命令の数も減らせます。また複数のオペレーションを1つにまとめると、アクセスしたデータを次に使用するまでキャッシュ内に保持できるので、メモリーアクセスを減らせます。

図1に示すように、テンセントの MLPD とインテルは、この考え方に基づいて特に時間のかかるプロセスを「Fused BERT op」という1つの大きなオペレーションに変更しました。

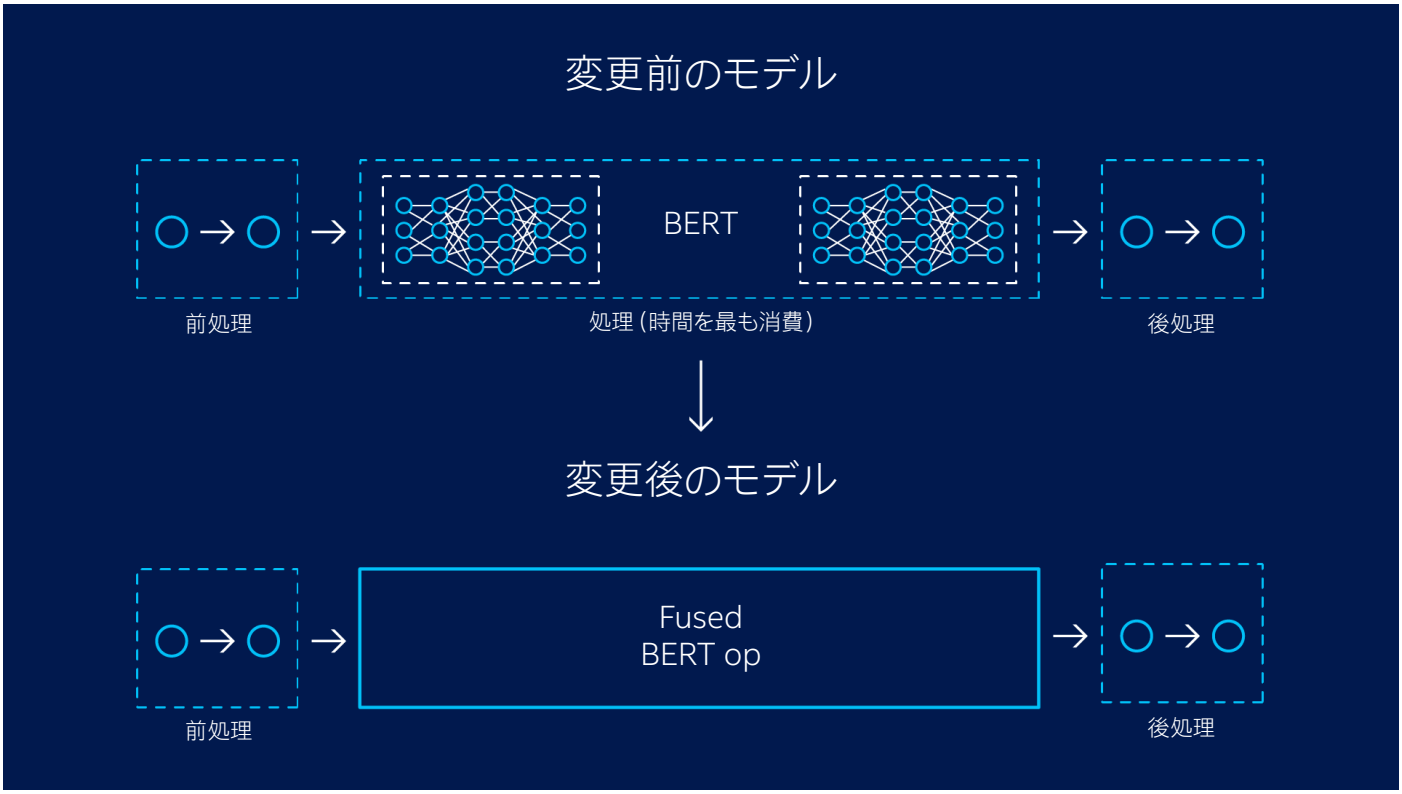

図1. 複数のオペレーションを1つの「Fused BERT op」にマージして BERT を最適化

「Fused BERT op」は、次の方法で最適化を実現します。

- QKV MatMul（QKV：Query/クエリー、Key/キー、Value/値）の入力が等しいことに着目し、これらのオペレーションの重みを1つの大きな重み行列にまとめ、1つの大きなQKV MatMulにマージしました。これらの重みが1つにマージされた後は、それぞれの重みと出力のメモリーが連続しなくなります。そのため、QKVの出力を次のオペレーションの入力として使用する場合は適切な「ストライド（間隔）」を設定する必要があります。この最適化フローを図2に示します。

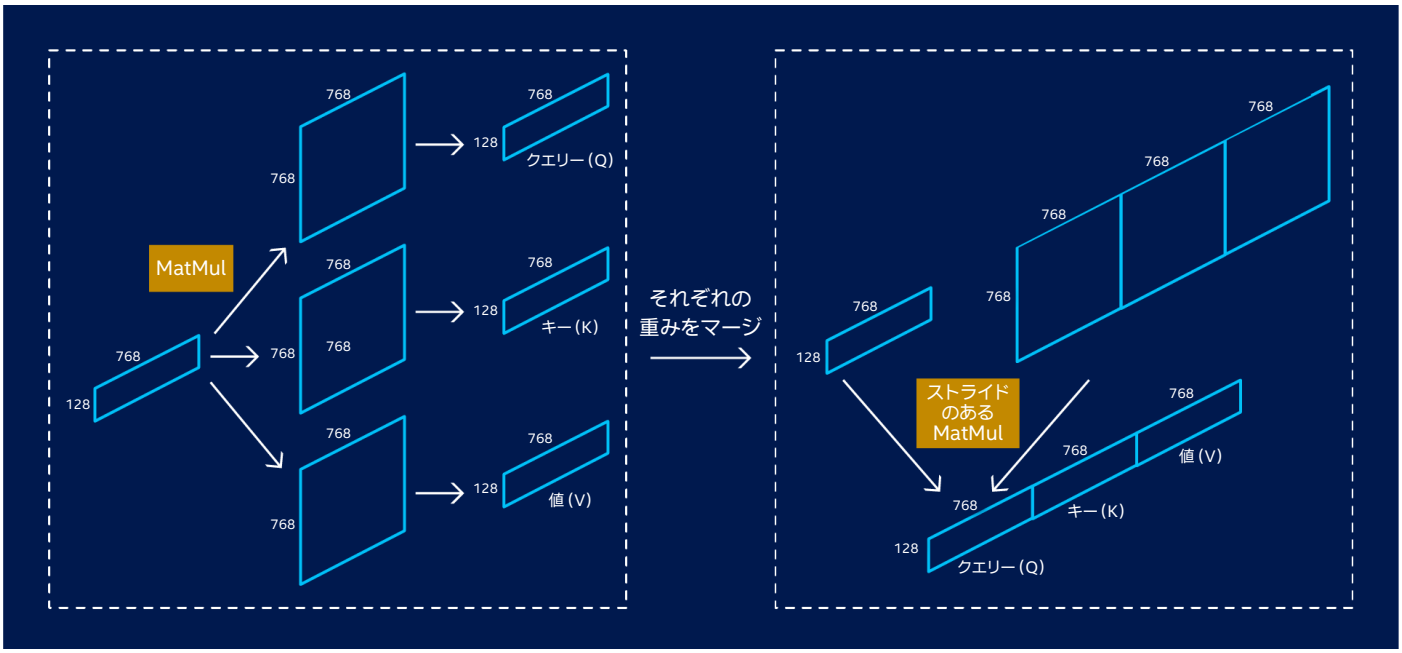

図2. QKV MatMul最適化フローチャート

- 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）がストライド（間隔）を持ったBatchMatMulをサポートするので、BatchMatMulの前後の置き換えオペレーション（op）が削除できます。これにより、大量のメモリーアクセスと計算処理が節約できます。この最適化フローを図3に示します。

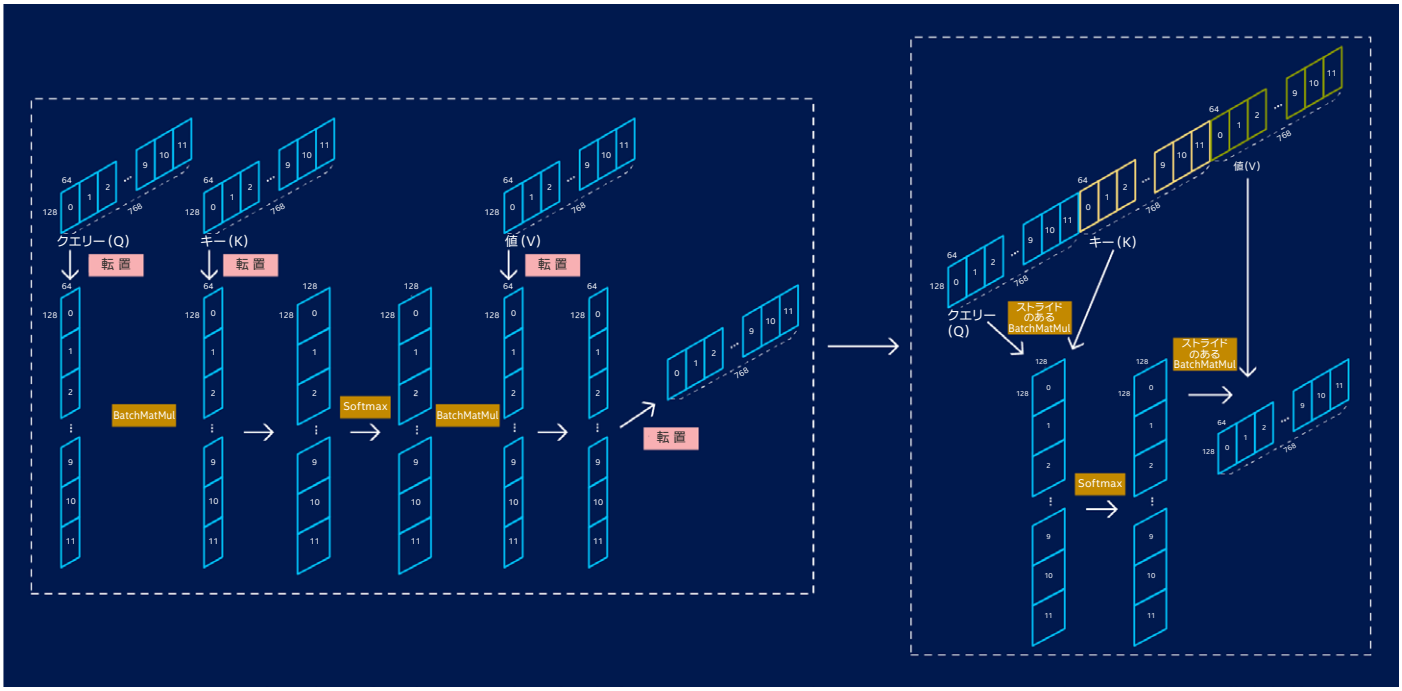

図3. BatchMul最適化フローチャート

- インテル® oneDNNは、BiasAddと特定の後処理（OutputScale、Sum、Relu、Gelu、Tanhなど）でMatMulをサポートするため、特定のオペレーションをMatMulのプリミティブに統合することができます。これらのオペレーション（op）を統合すると、キャッシュの使用効率が向上します。つまり、関連するデータを複数のタスクによって呼び出せるよう、キャッシュ内のデータをすぐに利用できる状態で保持できます。

Feature Dense 最適化

NLP のタスクでは、通常、フィーチャーやデータの長さが異なるため、バッチを形成するために大量のパディングを挿入する必要があります。その結果、不要な計算オーバーヘッドが大量に発生します。インテルはテンセントの MLPD と共同で、BERT モデル向けに Feature Dense 最適化ソリューションを開発しました。この最適化により計算オーバーヘッドが削除され、タスクのパフォーマンスが大幅に向上します。また、バッチサイズが大きいほどパフォーマンスが向上します。図 4 では、BERT モデルの一部に絞り、この最適化がどのように機能するかの概要を示しています。

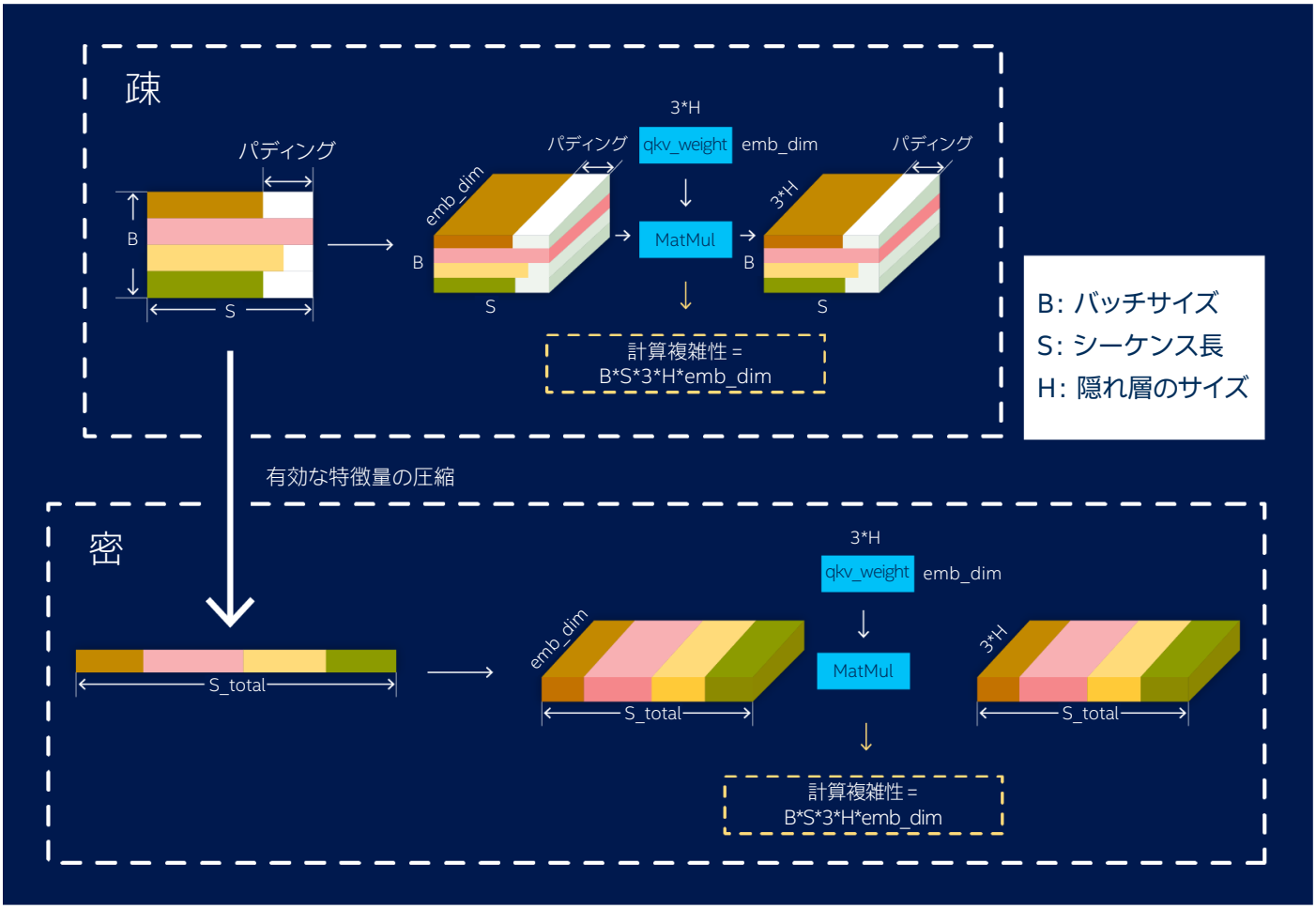

図 4. Feature Dense 最適化ソリューション

図 4 に示すように、Feature Dense 最適化ソリューションは、入力からパディングを排除し、各バッチのデータを 1 つずつつなげて、1 次元のデータを形成します。埋め込み、QKV MatMul、その後の MatMul オペレーション（op：図では省略）などの後続オペレーションは、圧縮されたデータを使うので、図 4 のように計算が大幅にシンプルになります。

BF16/INT8 の最適化

上記のように FP32 の最適化により BERT のパフォーマンスが大幅に向上しますが、さらに改善が可能です。メモリーのアクセスをさらに減らせば、パフォーマンスがいっそう向上します。これは、計算中のデータサイズを小さくすることにより実現します。入力と重みを含む FP32 のデータは、計算の前に BF16/INT8 データに変換できます。

では、高いパフォーマンスを維持しながら BF16/INT8 計算をサポートするのはどのようなプラットフォームでしょうか？ 第 3 世代インテル® Xeon® スケーラブル・プロセッサー（Cooper Lake および Ice Lake）には、ベクトル乗算向けの VNNI（INT8）命令をサポートする、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）が内蔵されています。また Cooper Lake は BF16 の数値フォーマットもサポートしています。このため、FP32 最適化ソリューションを使って BF16 または INT8 を最適化できます。テストの結果では、BF16 や INT8 を最適化すると、FP32 ソリューションと比較してパフォーマンスが大幅に向上することが確認されています。

インテル® AMX を内蔵した第 4 世代インテル® Xeon® スケーラブル・プロセッサー

BF16/INT8の最適化を行った上で、さらに最適化が可能でしょうか? 可能です。インテル® AMX は、第 4 世代インテル® Xeon® スケーラブル・プロセッサーに内蔵されたアクセラレーターです。インテル® AMX は、より大きな2次元メモリーイメージのサブアレイ（部分配列）で表される、2次元レジスター（タイル）のセット、そしてタイルのオペレーション（op）を可能にするアクセラレーターを備えた64ビット・プログラミング・パラダイムを提供します。最初の実装は「tile matrix multiply unit」を表す TMUL ユニットです。

図5は、インテル® AMX アーキテクチャーの概念図です。インテル® アーキテクチャー・ホストは、アルゴリズム、メモリーブロック、ループ・インデックス、ポインター演算を担っています。タイルの読み込みと格納、アクセラレーター・コマンドは、マルチサイクル実行ユニットの TMUL に送られます。

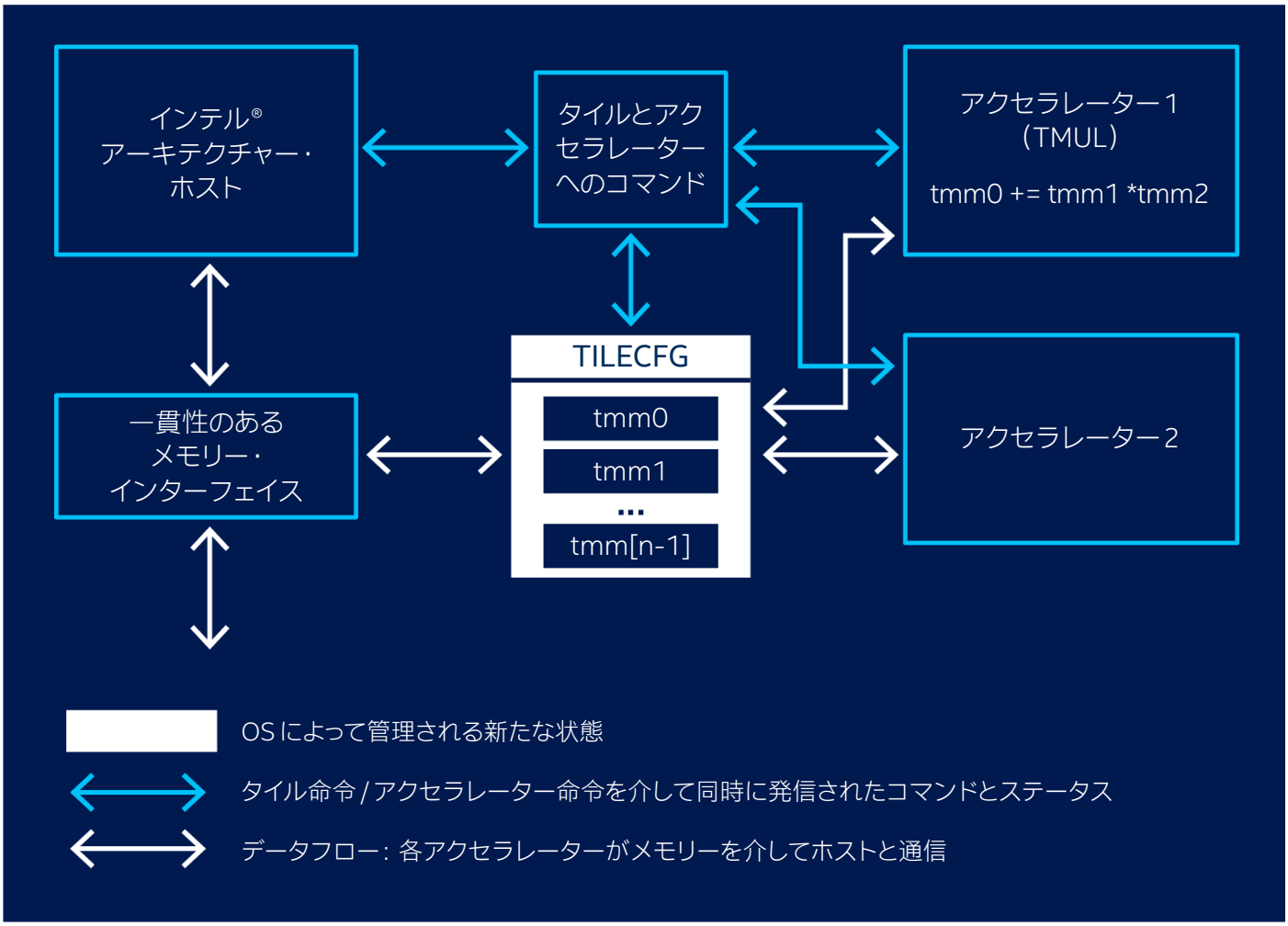

図 5. インテル® AMX アーキテクチャー

行列積（行列乗算）を実行するには、非常に複雑な計算が必要です。行列乗算のオペレーションが必要になるたびに複雑なプログラムを書くのでは、コストパフォーマンスが低下します。

インテル® oneDNN を使えば、行列計算の実行をシンプルにできます。作業に必要なのは、一部の後処理で MatMul プリミティブを呼び出して、いくつかのパラメーター（m、n、k、ストライド、データアドレスなど）を渡すことです。以前のインテル® マス・カーネル・ライブラリー（インテル® MKL）と同じようにインテル® oneDNN は、タイル・レジスター・ファイルの設定、メモリーからのデータ読み込み、後処理の行列乗算の実行、結果をメモリーに格納する処理、タイル・レジスター・ファイルの解放など、残りの作業を完了させます。インテル® oneDNN 経由でインテル® AMX を使う方法はプログラマーにとってわかりやすいので、プログラムをシンプルにできます。

インテル® AMX を内蔵した第 4 世代インテル® Xeon® スケーラブル・プロセッサーで動作する BERT の全体的なフローチャートを図 6 に示します。

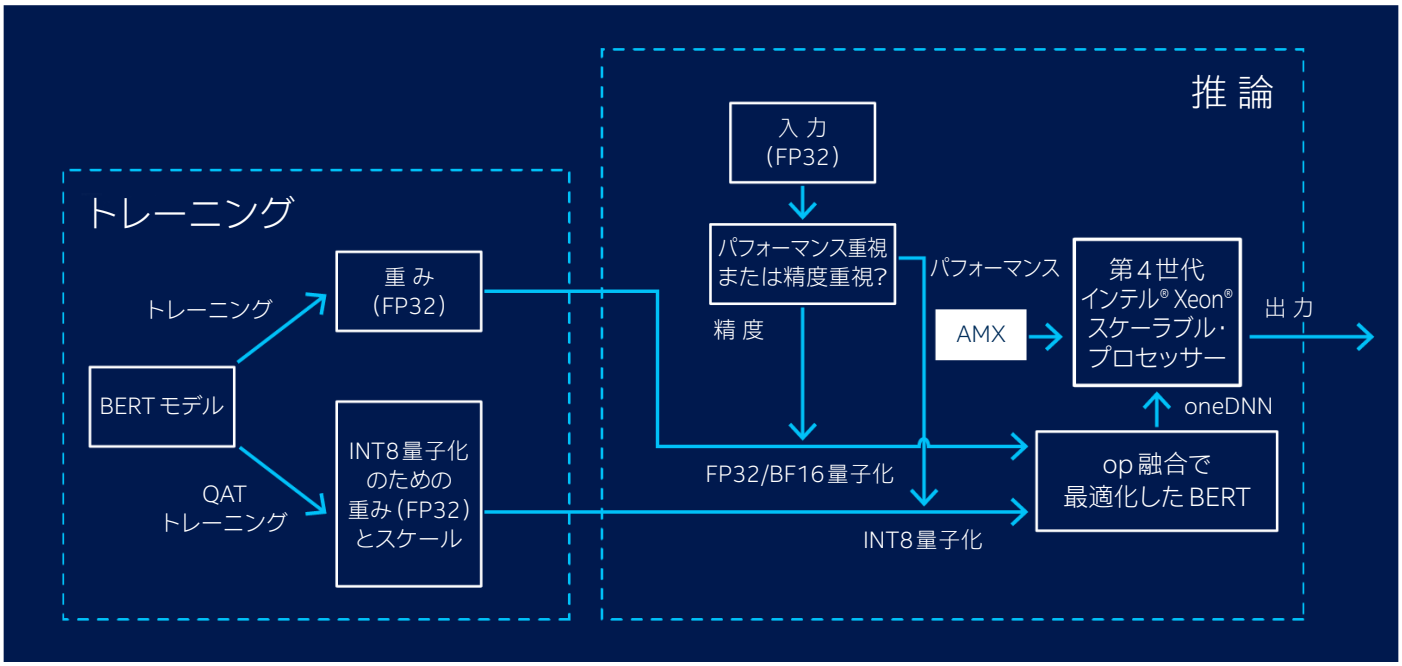

図6. 第４世代インテル® Xeon® スケーラブル・プロセッサー内蔵インテル® AMX 使用時の BERT 全体のフローチャート

パフォーマンスの比較

インテル® AMX により BERT のパフォーマンスを大幅に向上できます。プロセッサーの世代間でパフォーマンスの優位性を明らかにするには、さまざまなプラットフォームでパフォーマンスを比較する必要があります。インテルとの協業を通じて、テンセントの StarLake Lab は、x86マイクロアーキテクチャーに対する深い理解と、パフォーマンス・チューニングの経験を活用し、パフォーマンスの比較と最適化の作業に大きく貢献し、BERT テクノロジーを使用した際の第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・プロセッサーでのパフォーマンス検証で貴重なデータが得られました。

1つのソケットで複数の最適化インスタンスを実行し、各インスタンスのレイテンシーは一定のままでした。図7は、インテル® AMX 内蔵の第４世代インテル® Xeon® スケーラブル・プロセッサーを第3世代インテル® Xeon® スケーラブル・プロセッサーと比較すると、INT8と BF16の双方において、それぞれ2.05倍および3.02倍とシステム・パフォーマンスが大幅に向上したことを示しています。[1、2]

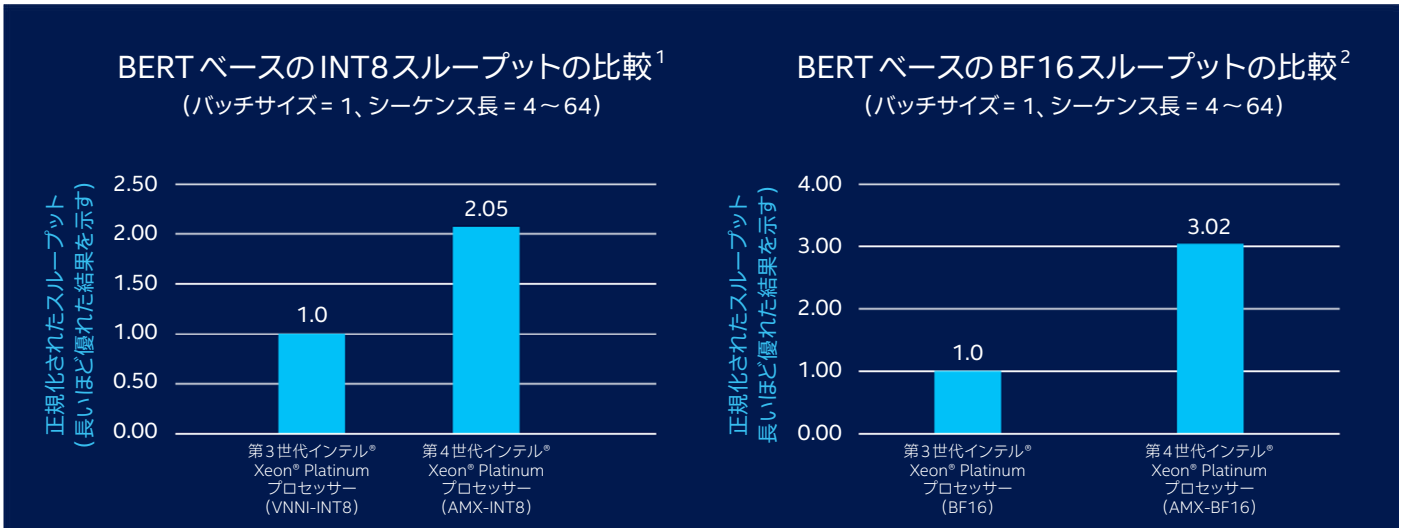

図7. インテル® AMX が内蔵された第４世代インテル® Xeon® スケーラブル・プロセッサーと前世代での BERT モデルのスループット比較

まとめ

インテル® AMXを内蔵した第4世代インテル® Xeon® スケーラブル・プロセッサーは、BF16/INT8 TMULコンピューティング・ユニットと、それに関連する命令により、行列乗算のパフォーマンスを大幅に向上させることが実証されました。インテルとテンセントは、インテル® AMXにより、前世代と比較してBERTモデルのスループットが2倍から3倍向上することを証明しました。[1,2] 現在、テンセントは最適化したBERTモデルを活用し、サービス体験の向上を実現しています。

intel XEON

1. BERT-base16 INT8スループットの比較

比較基準（Baseline）: Tencent TriRivers、BIOSバージョン1.08.00。OS：CentOS Linux release 8.5.2111、カーネル4.18.0-348.7.1.el8_5.x86_64、マイクロコード0xd000375。IRQバランス：有効。CPU：インテル® Xeon® Platinumプロセッサー、コアあたりのスレッド数2、ソケット数2、NUMAノード数2、プリフェッチャーL2 HW、L2 Adj.、DCU HW、DCU IP、ターボ有効、PPINs b59090a6f33f7966、b591426010edf86a。電源およびパフォーマンスのポリシー：パフォーマンス、TDP 270ワット、周波数ドライバーintel_pstate、周波数ガバナーperformance、周波数（MHz）2701、Max C-State 9、搭載メモリー960GB（15x64GB DDR4 3200MT/s [3200MT/s]）、Huge Pageサイズ2048kB、Transparent Huge Pages [always]。自動NUMAバランシング：有効。NIC：MT2892 Family [ConnectX-6 Dx] x2、デバイスx1、イーサネット・インターフェイスx1。ドライバーサマリー：111.8G INTEL SSDSCKHB12 x1。ワークロードおよびバージョン：INT8向けBERT最適化、コンパイラーGCC 8.5、ライブラリーインテル® oneDNN-master-0721。テスト実施日：2022年8月5日。

新規比較対象（New）: Intel Corporation ArcherCity、BIOSバージョンEGSDCRB1.SYS.0090.D03.2210040200。OS：CentOS Linux 8、カーネル5.16.0、マイクロコード0x2b0000c0。IRQバランス：有効。CPU：インテル® Xeon® Platinumプロセッサー、コアあたりのスレッド数2、ソケット数2、NUMAノード数2、プリフェッチャーL2 HW、L2 Adj.、DCU HW、DCU IP、ターボ有効、PPINs 31461920530bed98、3143931fcf7f6036。電源およびパフォーマンスのポリシー：パフォーマンス、TDP 350ワット、周波数ドライバーintel_pstate、周波数ガバナーperformance、周波数（MHz）2494、Max C-State 9、搭載メモリー512GB（16x32GB< 仕様外 >4800MT/s [4800MT/s]）、Huge Pageサイズ2048kB、Transparent Huge Pages [always]。自動NUMAバランシング：有効。NIC：イーサネット・コントローラーI225-LM x1、QSFP向けイーサネット・コントローラーE810-C x1。ドライバーサマリー：349.3G INTEL SSDPE21K375GA x1、1.5T INTEL SSDPEDMD016T4 x1、1.9T INTEL SSDPEKNW020T8 x1。ワークロードおよびバージョン：INT8向けBERT最適化、コンパイラーGCC 8.5、ライブラリー インテル® oneDNN-master-0721。テスト実施日：2022年10月19日。

2. BERT-base16 BF16スループットの比較

比較基準（Baseline）: Tencent QinghaiLake、BIOSバージョン1.02.00。OS：Red Hat Enterprise Linux 8.2（Ootpa）、カーネル4.18.0-193.el8.x86_64、マイクロコード0x7002502。IRQバランス：有効。CPU：インテル® Xeon® Platinumプロセッサー、コアあたりのスレッド数2、ソケット数4、NUMAノード数4、プリフェッチャーL2 HW、L2 Adj.、DCU HW、DCU IP、ターボ有効、PPINs 07ab90bc7f220116、07be7bc039fedc8f、07abcfbe92ff7f9b、07aba8bff01f278d。電源およびパフォーマンスのポリシー：パフォーマンス、TDP 175ワット、周波数ドライバーacpi-cpufreq、周波数ガバナーperformance、周波数（MHz）2695、Max C-State 9、搭載メモリー1536GB（24x64GB DDR4 3200MT/s [3200MT/s]）、Huge Pageサイズ2048kB、Transparent Huge Pages [always]。自動NUMAバランシング：有効。NIC：イーサネット・インターフェイスx1。ドライバーサマリー：447.1G SSSTC ER2-GD480 x1、894.3G Micron_5100_MTFD x1。ワークロードおよびバージョン：BF16向けBERT最適化、コンパイラーGCC 8.3、ライブラリー インテル® oneDNN-master-0721。テスト実施日：2022年8月8日。

新規比較対象（New）: Intel Corporation ArcherCity、BIOSバージョンEGSDCRB1.SYS.0090.D03.2210040200。OS：CentOS Linux 8、カーネル5.16.0、マイクロコード0x2b0000c0。IRQバランス：有効。CPU：インテル® Xeon® Platinumプロセッサー、コアあたりのスレッド数2、ソケット数2、NUMAノード数2、プリフェッチャーL2 HW、L2 Adj.、DCU HW、DCU IP、ターボ有効、PPINs 31461920530bed98、3143931fcf7f6036。電源およびパフォーマンスのポリシー：パフォーマンス、TDP 350ワット、周波数ドライバーintel_pstate、周波数ガバナーperformance、周波数（MHz）2552、Max C-State 9、搭載メモリー512GB（16x32GB< 仕様外 >4800MT/s [4800MT/s]）、Huge Pageサイズ2048kB、Transparent Huge Pages [always]。自動NUMAバランシング：有効。NIC：イーサネット・コントローラーI225-LM x1、QSFP向けイーサネット・コントローラーE810-C x1。ドライバーサマリー：349.3G INTEL SSDPE21K375GA x1、1.5T INTEL SSDPEDMD016T4 x1、1.9T INTEL SSDPEKNW020T8 x1。ワークロードおよびバージョン：BF16向けBERT最適化、コンパイラーGCC 8.5、ライブラリー インテル® oneDNN-master-0721。テスト実施日：2022年10月19日。

356379-001JA
JPN/2309/PDF/TAG/MKTG/SN