WP508 (v1.0) 2019 年 1 月 15 日

Samsung HBM2 搭載のザイリンクス HBM 対応 UltraScale+ デバイスで AI および データベース アプリケーションを強化

広帯域幅メモリ (HBM) を利用することで、データ移動やデータ アクセスが効率化され、ザイリンクス UltraScale+ デバイスの優れた演算能力を存分に活用できるようになります。

概要

これまで演算性能の不足がボトルネックとなっていたアプリケーション分野の多くは、ザイリンクスの UltraScale™ および UltraScale+™ デバイスや Versal™ ACAP などのヘテロジニアス コンピューティング アクセラレータの開発によって大幅に進歩しました。

一方、アクセラレータ ソリューションでは、DDR メモリから演算ユニットへのデータ移動によって速度が制限されます。広帯域幅メモリ (HBM) は、データ移動やデータ アクセスのボトルネックを軽減するのに役立ちます。ザイリンクスの HBM 対応デバイスは、データセンターの新たな可能性を開き、演算の高速化を次のレベルへと引き上げます。

このホワイトペーパーでは、深層学習とデータベース、2 つの高速化のユース ケースを例にとり、ザイリンクスが提供する Samsung HBM2 搭載コンピューティング アクセラレータ ソリューションによって、柔軟性、効率性、性能に優れたハイインパクトな演算高速化がどのように実現されるかについて説明します。

バックグランド

近年のヘテロジニアス コンピューティングの台頭は、ポスト「ムーアの法則」時代における演算集中型ワークロードの高速化を前進させせました。今日のデータセンターにおいて、ヘテロジニアス コンピューティングで高速化されるワークロードとしては、人工知能、ライブ ビデオのトランスコーディング、ゲノム解析などが一般的ですが、これらはほんの一例にすぎません。ザイリンクスの UltraScale+™ および次世代の Versal™ デバイスは、現代のデータセンターのワークロードにこれまでにない適応性と演算高速化をもたらします。

一方、DDR メモリ アーキテクチャは長い間それほど進化せず、演算高速化の技術革新のペースについていけませんでした。並列メモリ インターフェイスの帯域幅性能は過去 10 年で向上しているもののペースが遅く、今日の FPGA でサポートされている最大 DDR4 データ レートは、2008 年当時の DDR3 性能の 2 倍程度しかありません。対照的に FPGA の演算性能は 2008 年から約 8 倍に向上しており、さらに AI コアを搭載した Versal デバイスの市場展開によって、今後 2 年でさらなる飛躍的向上が見込まれています [参照 1]。つまり、メモリ帯域幅とメモリ容量は、データセンターで演算やメモリ帯域幅を多用するさまざまなワークロードを開発する際の主要な制限要因となります [参照 2]。図 1 を参照してください。

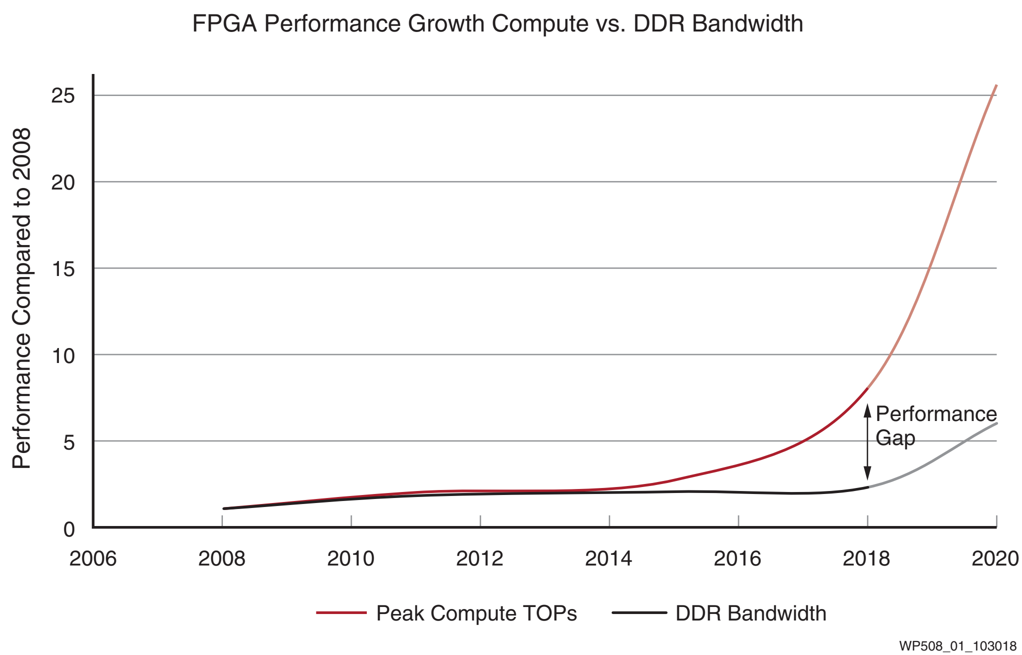

図 1: 演算能力と DDR 帯域幅の向上の比較

UltraScale+ ファミリの最新製品は、同じデバイス パッケージで広帯域幅メモリ (HBM) に対応しているため、メモリと演算間の性能ギャップを埋めることができます。

Samsung HBM

HBM は、縦に相互接続された DRAM チップのスタックとワイド (1024 ビット) インターフェイスを用いた高速 SiP (System in Package) メモリ技術であり、従来のワイヤ ボンディング ベースのパッケージを用いたメモリ製品よりもストレージ容量とデータ帯域幅を増やすことができます。Joint Electron Device Engineering Council (JEDEC) は 2013 年に最初の HBM 規格を採用し、2016 年 1 月に第 2 世代の HBM2 バージョンを業界規格として承認しました。図 2 を参照してください。

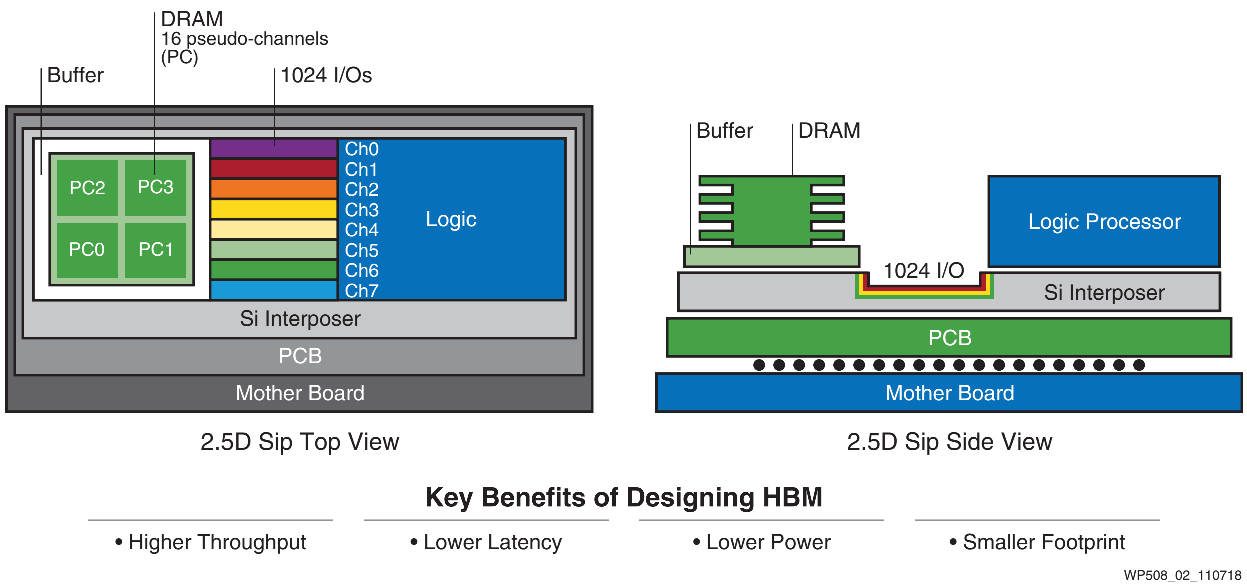

図 2: HBM および 2.5D の構造 (情報提供: Samsung)

HBM2 規格では、4 つまたは 8 つの 8Gb DRAM ダイを積み重ねることができます。これらのダイは、メモリ スタックと基礎をなす回路基板をつなぐ 2.5D シリコン インターポーザーに支えられています。一部のダイ (一般にメモリ) のみを積み重ね、それ以外のダイ (一般にプロセッサ) は積み重ねないマルチチップ パッケージを 2.5D デバイスと呼びます。

スタックは、同一のパッケージに複数追加できます。複数のダイを積み重ね、回路基板上でそれらのスタックを互いに近づけると、従来のマルチチップ DRAM 製品と比べて HBM メモリ パッケージの平面フットプリントを大幅に減らすことができます。HBM 技術により、デバイス間の信号移動距離が短くなるため、システム性能も向上します。距離の短縮は、所定量のデータを伝送するのに必要なエネルギーの削減にもつながります。図 3 を参照してください。

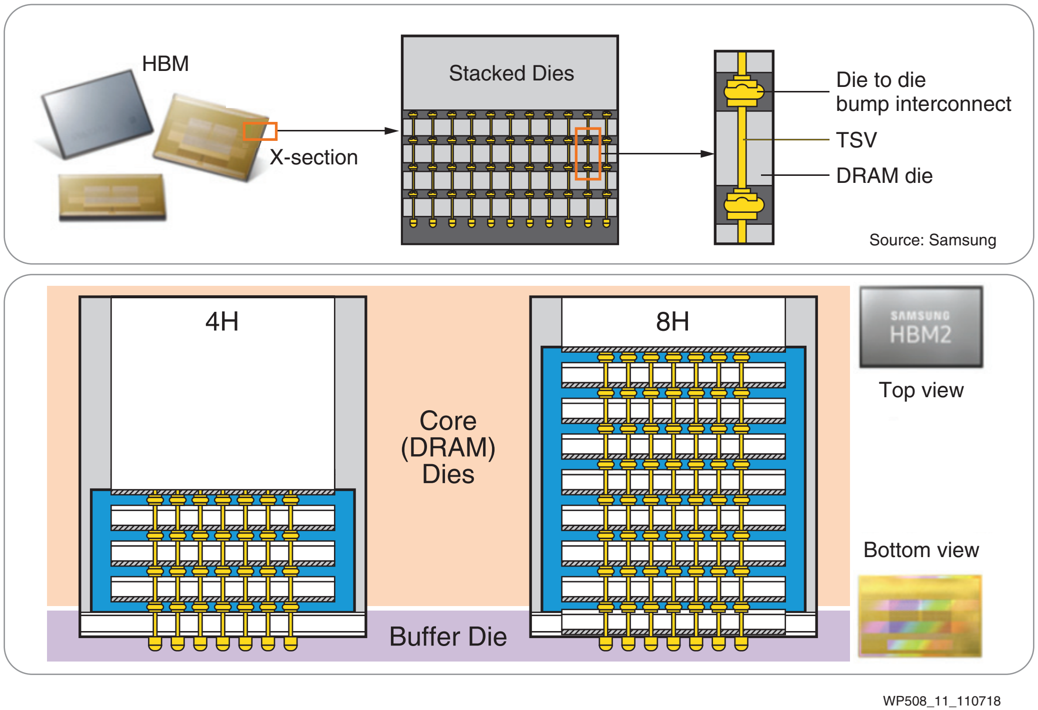

図 3: HBM による積層ダイ技術 (情報提供: Samsung)

HBM と、メモリ帯域幅を増やすための HBM の高度な TSV 技術、マイクロスケールのインターコネクト、多数の I/O により、最も近い競合製品、つまりグラフィックス カードに用いられる GDDR (Graphics Double Data Rate) メモリよりも優れた性能が実現します。コンポーネント レベルでは、1 つの Samsung HBM キューブで最大 307GB/s のデータ帯域幅を提供でき、データ伝送速度は GDDR5 チップの約 10 倍に達します。システム レベルでは、HBM は GDDR ベースのソリューションと比べて 80% の消費電力で約 3 倍のスループットを実現するだけでなく、貴重な回路基板のスペースも節約します。図 4 を参照してください。

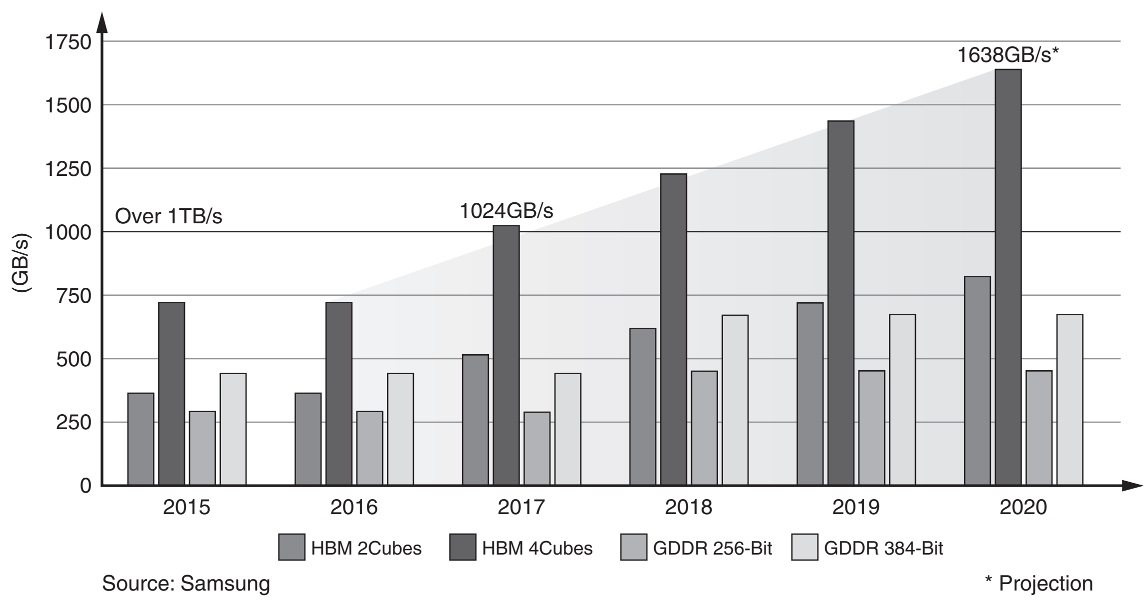

図 4: HBM と GDDR の帯域幅の比較

Virtex UltraScale+ デバイスによるメモリ帯域幅の向上

HBM 搭載の Virtex® UltraScale+ FPGA (VU31P、VU33P、VU35P、VU37P) [参照 2] により、メモリ帯域幅は大幅に向上しました。Samsung HBM2 メモリ スタックを 2 つ使用すると、メモリ帯域幅は最大 460GB/s に達します。1 つまたは 2 つの HBM2 スタックと、さまざまなサイズの FPGA ロジック (最大 285 万のロジック セル) および DSP (最大 9,024 の DSP48E スライスで 28.1 のピーク INT8 TOP を実現) を組み合わされた、新しい HBM 対応 UltraScale+ デバイス ファミリによって、アプリケーションに最適な演算能力とメモリ帯域幅/容量の組み合わせが選択できるようになります。

ザイリンクスは HBM 対応 Virtex UltraScale+ FPGA に加え、Alveo™ データセンター アクセラレータ カード ファミリの新製品として Alveo U280 カードのリリースを発表しています [参照 3]。

Alveo U280 カードはザイリンクスの 16nm UltraScale+ アーキテクチャをベースに構築されており、8GB の Samsung High Bandwidth Memory (HBM2) を搭載しています。これにより、データベースの検索と分析や機械学習による推論などのメモリ負荷が大きいアプリケーションを最大限に高速化します。U280 アクセラレータ カードには、既存のサーバー インターコネクト インフラストラクチャを利用するための CCIX サポートが組み込まれています。これにより、今後の CCIX 対応プロセッサを使用して、キャッシュ コヒーレントな共有メモリに高帯域幅かつ低レイテンシでアクセスできるようになります。

HBM 対応の UltraScale+ デバイスは、メモリがボトルネックとなる各種のアプリケーションにとってメリットがあります。このホワイトペーパーでは、ディープ ニューラル ネットワークとデータベース高速化に関するケース スタディを例にとって、HBM 対応 UltraScale+ デバイスのメリットを紹介します。

ケース スタディ 1

言語翻訳用 AI の高速化

現在、さまざまなクラウド アプリケーションによってリアルタイムの自動言語翻訳サービスが提供されています。このようなサービスは機械翻訳と呼ばれ、ニューラル ネットワーク ベースの機械学習メソッドを利用して 2 つの言語間でセンテンスを翻訳します。[1]

今日の商用自動翻訳サービスには、エンコーダー /デコーダー アーキテクチャが使用されています。マシンを利用して翻訳タスクを実行すると同時に、ワード エンベディングと呼ばれるプロセスを通して両言語の単語が高次元ベクターとして表現されます。これにより、単語間の関係を定量的にモデル化し、ベクターに反映させることができます。エンコード機能とデコード機能の実行には、再帰ニューラル ネットワーク、たたみ込みニューラル ネットワーク、アテンション ベース モデルなどの構造が用いられます。

最近の調査では、アテンション ベースのネットワークを言語翻訳に利用するだけで、業界最高の正確性を達成できることがわかりました [参照 4]。この調査報告書で説明されているアテンション メカニズム (スケール ドット積のアテンション) は、2 つの行列乗算とその他の関数 (scale、mask、softmax) で構成されます。

マルチヘッド アテンション構造は、並列スケール ドット積のアテンションの倍数と、さまざまな入力予測で形成されます。この構造とフィードフォワード ネットワークを使用して、全体的な言語翻訳モデルのデコーダーとエンコーダーが構成されます [参照 4]。図 5 を参照してください。

1. Amazon Translate:
https://aws.amazon.com/blogs/aws/introducing-amazon-translate-real-time-text-language-translation/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+AmazonWebServicesBlog+%28Amazon+Web+Services+Blog%29

Microsoft Azure: https://www.microsoft.com/en-us/translator/business/machine-translation/

Google Machine Translation Service:
https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+blogspot/gJZg+(Google+AI+Blog)

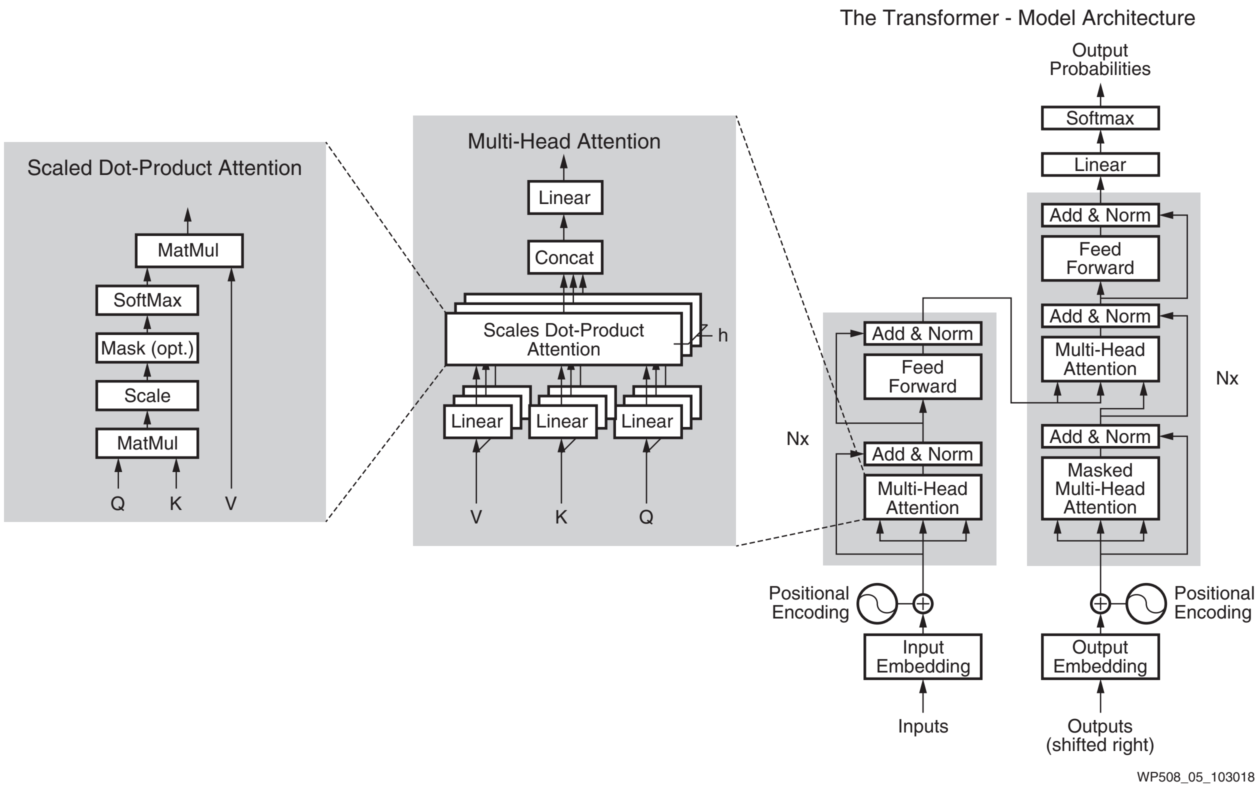

図 5: Transformer: アテンション ベースの言語翻訳用ニューラル ネットワーク モデル [参照 4]

図 5 に示すように、アテンション ベースの言語翻訳モデルで主に演算が集中するのは、スケール ドット積のアテンションとフィードフォワード ネットワークに使用するドット積の演算です。また、演算を効率化するため、これらのドット積は一般に、行列乗算にグループ化されます。従来のたたみ込みニューラル ネットワークは、データ移動のフットプリントを減らすために重みパラメーターをスペース全体で存分に再利用できる一方、このアテンション ベースのモデルは入力スペースにパラメーターをほとんど再利用しないため、行列乗算を処理するためのメモリ帯域幅の要件が非常に高くなります。

Transformer の解析モデルは、業界最先端のアテンション ベースのニューラル ネットワークとして言語翻訳に利用されますが、このモデルは UltraScale+ デバイスに実装するよう構築およびマッピングされています。FPGA 実装のアーキテクチャは、DSP シストリック アレイを使用して行列乗算を実行します。中間アクティベーション データはオンチップ URAM に格納されるため、アクティベーションのためにプロセッサとオフチップ メモリ間で頻繁にデータを移動する必要はありません。エンベディング ディクショナリのワード エンベディングと重みパラメーターはすべて HBM または DDR に格納されます。

さまざまなセンテンス長 (L) とセンテンス数 (バッチ サイズとも呼ばれる) (B) の英語/ドイツ語間の翻訳タスクを解析すると、DDR に接続する UltraScale+ デバイスと HBM 対応 UltraScale+ デバイスにおけるレイテンシとスループットのトレードオフがわかります。このケース スタディでは、256 の 16 ビット入力と 32 の 16 ビット出力によるシストリック アレイを使用し、DDR と HBM のどちらの実装でも 710MHz の UltraScale+ デバイス上で実行することを前提としています。図 6 に、スループット対レイテンシの詳細なグラフを示します。

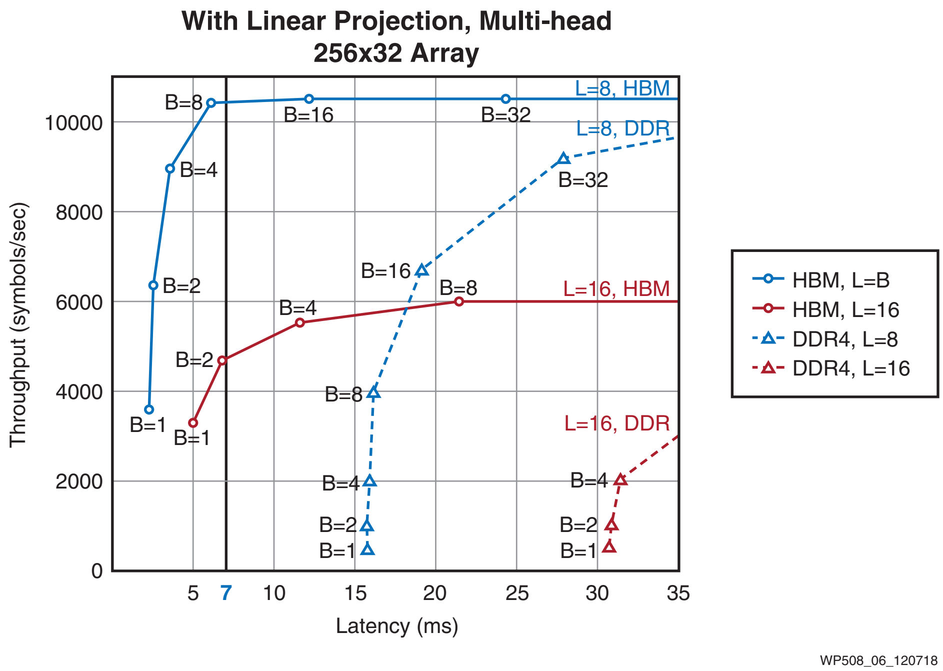

図 6: HBM および DDR4 搭載の UltraScale+ デバイスを使用した言語翻訳の性能解析

4 チャネル実装の最小レイテンシは、センテンス長 8 の場合に約 15.7ms で、スループットは 508 シンボル/秒になります。1 つの DDR チャネルはワード エンベディング データへのアクセスに使用され、残り 3 つの DDR チャネルは重みのアップロードに使用されます。HBM 対応デバイスの場合、センテンス長 8 の最小レイテンシは 2.2ms で、DDR インターフェイスより 7 倍以上小さくなっています。DDR インターフェイスはいずれのセンテンス長でもレイテンシを 7ms 未満にできないのに対し、HBM 対応デバイスは、センテンス長 8 で 6.1ms のレイテンシ (スループットは 10,419 シンボル/秒)、センテンス長 16 で 6.8ms のレイテンシ (スループットは 4,682 シンボル/秒) を実現します。

汎用行列演算ライブラリを使用した AI の高速化

前述の機械翻訳の例と同様、現代のほぼすべてのディープ ニューラル ネットワーク (深層学習) における主な演算は、行列乗算の形式で実行されます。

データセンターで広く運用されている一般的な深層学習アプリケーションには、機械翻訳以外にも、画像/動画解析、ウェブ検索に使用される検索ランキング システム、広告配置に使用される推奨システム、コンテンツ/フィードの推奨、音声認識、自然言語処理などがあります。

ザイリンクスでは、幅広い深層学習アプリケーションのサポート向上を目的として、SDAccel™ 開発環境でサポートされるザイリンクス デバイス上での行列演算を高速化するための汎用行列演算 (GEMX) ライブラリを開発しました。このライブラリには、エンジン ライブラリ、ホスト コード コンパイラ、アプリケーション/システム構築環境の 3 つのコンポーネントが含まれています。エンジン ライブラリは、C++ テンプレートのセットと BLAS のような関数インターフェイスで構成され、FPGA での行列演算を実現します。ホスト コード コンパイラは、ホスト コードの行列関数呼び出しをコンパイルし、FPGA で行列演算をトリガーするための命令シーケンスにします。構築環境は GNU make フローを使用して、FPGA とホスト コード イメージの生成プロセスを自動化します。これによりユーザーは、システムのさまざまな側面 (FPGA プラットフォーム、FPGA イメージで実装される各種エンジンなど) をコンフィギュレーションすることもできます。GEMX エンジンの設計の詳細は、https://github.com/Xilinx/gemx を参照してください。

GEMX エンジンの入力行列はどちらも DDR メモリからのフィードですが、GEMX のスループットは DDR インターフェイスの帯域幅によって決まります。次に示す解析は、DDR4 に接続した UltraScale+ デバイスを使用した GEMX と、HBM 対応 UltraScale+ デバイスを使用した GEMX の性能を比較したものです。この解析は、メモリ帯域幅をフル活用した状態をモデル化しており、GEMX への入力として 32x32x128 の行列を使用することを前提としています。(図 7 に示す) 結果から、HBM 対応デバイスは、DDR の 4 つのチャネルを使用した場合と比べて GEMX の性能を約 3.6 倍向上できることがわかります。

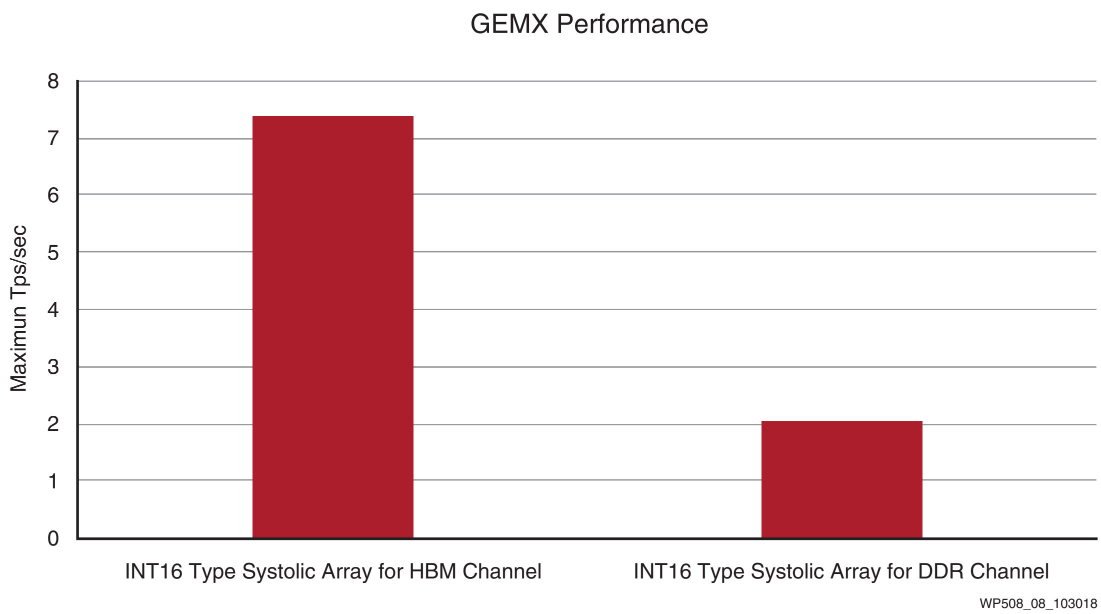

図 7: GEMX の性能

ケース スタディ 2

データベースの性能向上

現代のデータベース システムは、大量の非構造化データを低レイテンシで処理する必要があります。最新の x86 ソリューションは、プロセッサ内で大容量の内部メモリ (キャッシュ) や並列処理ユニット (AVX) までもサポートできます。ただし、CPU ベースの実装では、高まり続ける性能要件にそれでも対応できず、大半のケースで費用対効果の目標を達成できていません。GPU ベースの実装も普及が進んでいますが、高い性能を実現するには大量のバッチ処理と消費電力を伴います。

大容量の Samsung HBM を搭載したザイリンクス デバイスであれば、必要なメモリ データ幅、チャネル数、カーネルごとに必要なメモリ量をユーザーが選択できます。プログラマブル ロジックは並列処理エンジンと HBM メモリへの高速アクセスを組み合わせて、主要なデータベース演算 (Gather、Group、Indirect Gather、Fused Gather Group、Segmented Gather など) を高速化できます。

図 8 は、ホストが受け取ったクエリに対して、データベース演算要求を FPGA ベースに送信する一般的なユース ケースを示しています。この後、接続された HBM に FPGA 演算エンジンが高帯域幅かつ高速でアクセスして、高速化された性能でデータベース演算を実行します。演算結果はホストに返され、クエリ要求は終了します。

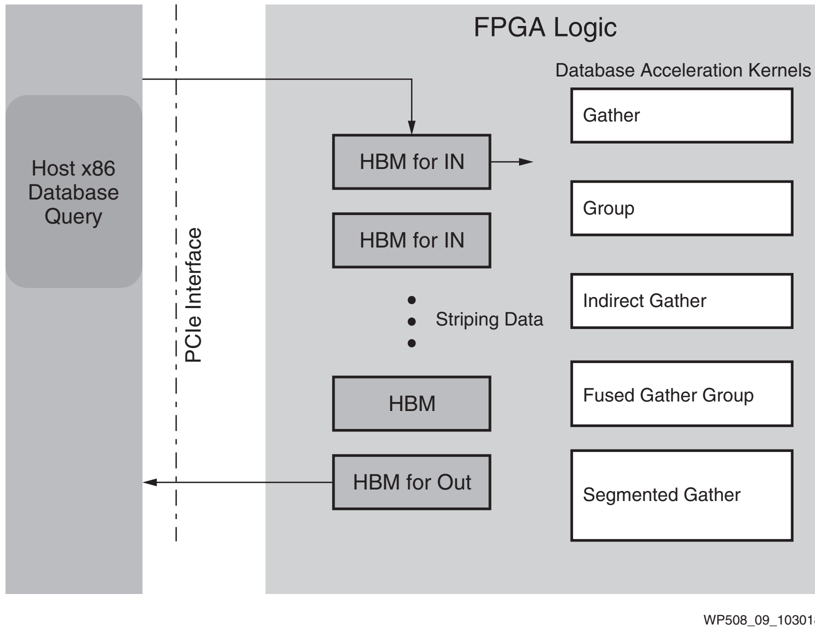

図 8: HBM 対応 UltraScale+ デバイスでのデータベース ユース ケースの例

一般的なデータベース演算は多くのメモリ アクセス (主にランダムなメモリ読み出し/書き込み処理) を必要とします。HBM 上のデータベース演算の解析では、性能を測定するために Gather と呼ばれるデータベース プリミティブ演算の 1 つが使用されます。Gather 演算では、ランダムな少量データへのアクセスが何回も実行されます。図 9 に、演算ユニット (並列処理エンジン) と疑似チャネルの数を変えて測定した、Gather 演算の性能向上を示します。

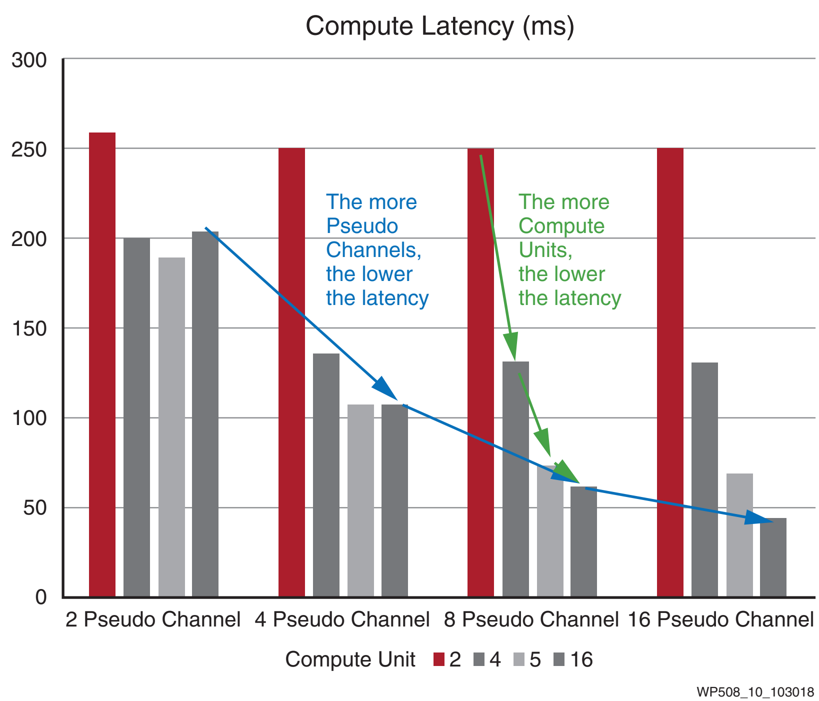

図 9: データベース演算のレイテンシ解析

図 9 に示すように、演算ユニット、疑似チャネル、またはその両方の数を増やすと、レイテンシが向上します。たとえば、8 つの疑似チャネルを使用した場合、演算ユニットの数が増えるほどレイテンシは低下します (緑色の線)。同様に、演算ユニットを 16 個使用した場合も、疑似チャネルの数が増えるほどレイテンシは低下します (青色の線)。

考察

高帯域幅によるリアルタイム アプリケーションの演算能力の向上

「ケース スタディ 1」では、高いメモリ帯域幅を必要とするニューラル ネットワークの例について説明しました。

RNN (再帰ニューラル ネットワーク) または MLP (マルチレイヤー パーセプトロン) では、1 バイトのパラメーター (INT8 演算を想定) を使用するのに必要な演算は乗算と加算 の 2 回のみです。これは、CNN (たたみ込みニューラル ネットワーク) とはまったく対照的です。CNN は理論上、各レイヤーの 1D、2D、さらには 3D スペース全体で重みパラメーターを幅広く再利用できるため、演算強度 (1 バイトのデータを使用するのに必要な算術演算) が非常に高くなります。この場合、最高の演算効率を達成するには、演算カーネルにデータをフィードするデータ レートまたはメモリ帯域幅が演算能力と一致している必要があります。

ザイリンクスの UltraScale+ デバイスや Versal ACAP など、現在最高の演算アクセラレータは数十～数百の INT8 TOP を容易に実現できます。理論上、正確に一致するメモリ帯域幅量は数十テラバイト/秒以上必要であり、これは非常に限られたオンチップ SRAM でしか実現できません。このメモリ帯域幅要件を容易にする理想的なアプローチは、バッチ処理を使用することです。バッチ処理のタスクはまとめて実行できると同時に、タスクのバッチ間で同じパラメーターを共有できます。このため演算強度が増し、それにより演算処理 1 回あたりのデータ移動の集中度が減ります。バッチ処理のトレードオフは、「ケース スタディ 1」の図 6 に示すようにレイテンシが増えることです。したがって、レイテンシが重要となる多くのタスク (リアルタイム翻訳など) の場合、大量のバッチ処理は選択肢となりません。

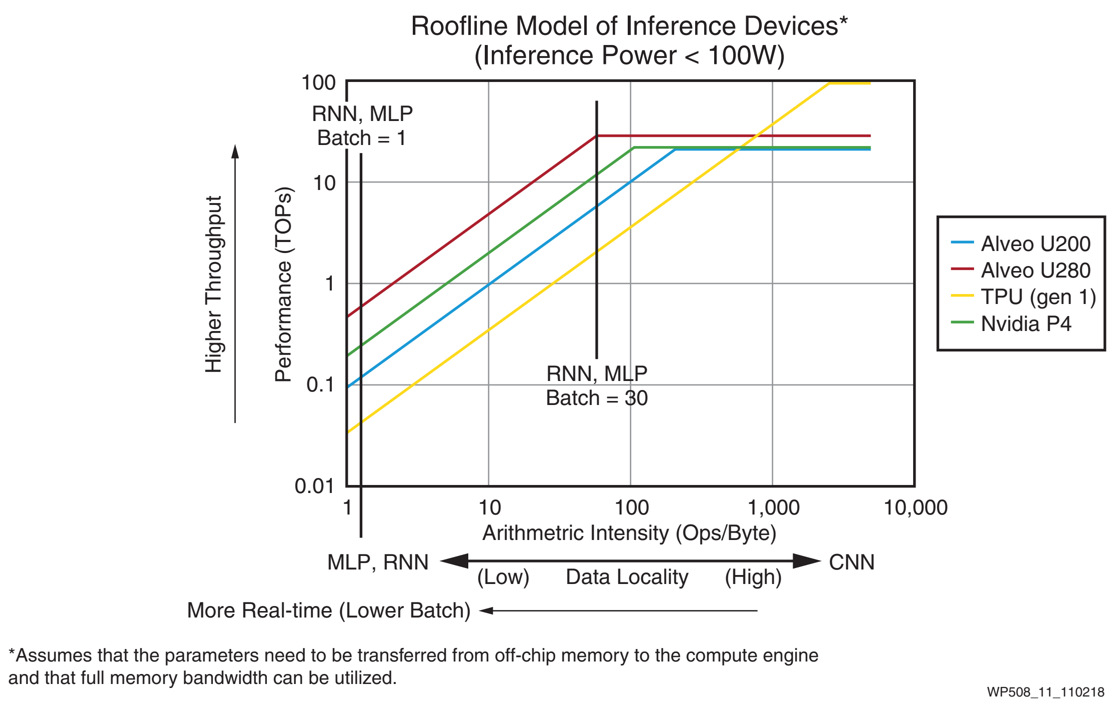

図 10: ザイリンクス UltraScale+ およびその他のデバイスのルーフライン モデル

図 10 に、機械学習の推論ワークロードに対応できる、消費電力 100W 以下の各種デバイスの理論上のルーフライン モデルを示します。この図では、メモリのルーフラインが、所定の演算強度におけるデバイスの理論的なピーク性能を示しています。リアルタイム (低バッチ、低レイテンシ) の RNN および MLP ユースケースで、ザイリンクスの HBM 対応 Alveo U280 アクセラレータ カードがあらゆる推論デバイスの中で最高の性能を提供できることは明らかです。

柔軟な並列メモリ チャネル アクセスによるメモリと演算リソースの使用効率向上

HBM 対応 UltraScale+ デバイスの利点の 1 つは、HBM メモリ コントローラー [参照 2] に組み込まれた柔軟なアドレス指定機能によって、HBM チャネルを柔軟に並列使用できることです。HBM チャネルのアクセスと使用は完全にリコンフィギュレーション可能かつ再プログラム可能で、FPGA ロジックによって容易にアクセスできます。図 11 に示すように、ケース スタディ 2 で説明したデータベースのユース ケースを拡張すると、オリジナルのデータベース ファイルを複数の HBM チャネルに分け、並列処理演算ユニットにさまざまなデータベース演算を実行させて、同期を必要とせずにデータベースの複数部分に独立かつ並列してアクセスできます。この機能を使用すれば、バッチ処理も同期も必要とせずに、すべての演算ユニットがリアルタイムでデータベース クエリに対応できるため、リアルタイム ユース ケースにおける演算のレイテンシと HBM の使用効率を向上させることができます。

「ケース スタディ 1」の機械翻訳の例でも、2 つのチャネルをワード エンベディング アクセス用、残りのチャネルを高帯域幅の重みパラメーター転送用に分けることで、この機能が利用されています。

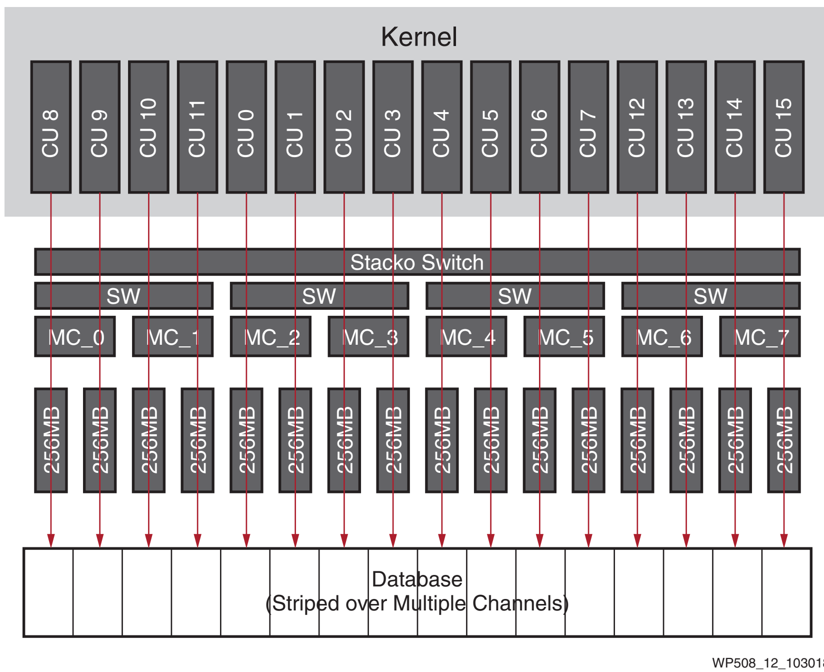

図 11: 柔軟な並列メモリ チャネル アクセス

まとめ

AI やデータベース高速化などのユース ケースにおける Samsung HBM の重要性は説明したとおりですが、HBM から大きなメリットが得られるデータセンター ワークロードがほかにも多数あります。たとえば、SmartNIC のユース ケースはザイリンクスのホワイトペーパーで昨年紹介され [参照 2]、ザイリンクスのパートナーも HBM 対応デバイスのアクセラレーション FFT を使用したユース ケースを公開しました [参照 5]。

コンピューティング デバイスの技術革新と共に、データセンターではヘテロジニアス コンピューティングの高速化の傾向が進んでおり、演算ユニットの近くに接続された HBM のような高性能メモリ システムの必要性は切迫しています。

HBM 対応 UltraScale+ デバイスには、ほかにはない価値があります。深層学習やデータベース高速化を例にとってみても、ザイリンクスの HBM 対応 UltraScale+ デバイスが新たな可能性を開き、演算高速化を次のレベルへと引き上げることは明らかです。

参考資料

1. ザイリンクス ホワイトペーパー :『Versal: 業界初の Adaptive Compute Acceleration Platform (ACAP)』(WP505: 英語版、日本語版)
2. ザイリンクス ホワイトペーパー :『Virtex UltraScale+ HBM FPGA: メモリ性能の革新的向上』(WP485: 英語版、日本語版)
3. ザイリンクス: 『Alveo 製品選択ガイド』
4. Vaswani, A. ほか共著: 『Attention Is All You Need』、v0.4、2018 年 7 月 18 日。Google, Inc. Cornell University Library (arXiv.org) から抜粋。
5. McCormick, A. C. 著: 『Breaking Memory Bandwidth Barriers Using High Bandwidth Memory FPGA』、v1.0、2018 年 5 月 15 日。Alpha Data Parallel Systems Ltd. (ftp.alpha-data.com) から抜粋。

改訂履歴

次の表に、この文書の改訂履歴を示します。

日付	バージョン	内容
2019 年 1 月 15 日	1.1	「Virtex UltraScale+ デバイスによるメモリ帯域幅の向上」、「言語翻訳用 AI の高速化」、および図 6 を更新。
2018 年 11 月 12 日	1.0	初版

免責事項

自動車用のアプリケーションの免責条項

この資料に関するフィードバックおよびリンクなどの問題につきましては、jpn_trans_feedback@xilinx.com まで、または各ページの右下にある [フィードバック送信] ボタンをクリックすると表示されるフォームからお知らせください。いただきましたご意見を参考に早急に対応させていただきます。なお、このメール アドレスへのお問い合わせは受け付けておりません。あらかじめご了承ください。