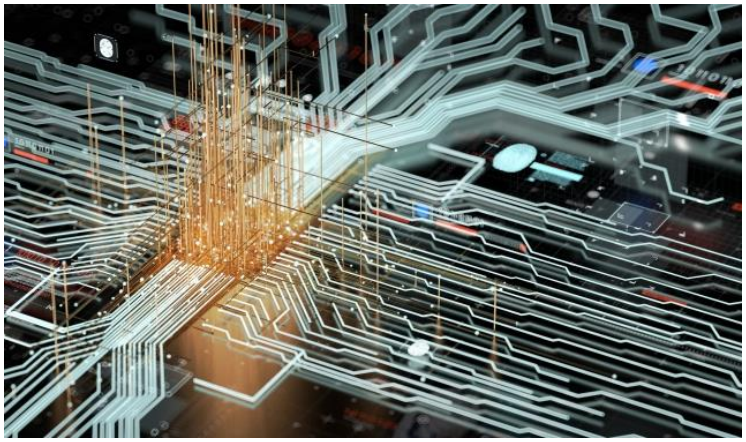
広帯域幅メモリ FPGA を使用したメモリ帯域幅障壁の解消

はじめに：



Virtex Ultrascale+広帯域幅メモリ（HBM）FPGA デバイスのリリースにより、メモリバウンドアプリケーションの全く新しい領域を開き、電力効率の良い FPGA アクセラレーションに利点をもたらします。最近のトレンドは、CPU を超えるメモリ帯域幅の大幅な利点のため、さまざまなメモリバウンドアプリケーションは高性能の並列浮動小数点演算を必要としないにもかかわらず GPU システムに向けられています。しかし、同様の外部メモリ帯域幅を備えた FPGA の出現により柔軟性と内部メモリ帯域幅が大幅に向上し、これらの問題に対するよりカスタマイズされたエネルギー効率の高いアクセラレートソリューションが可能になりました。

VU37P は、ザイリンクス Virtex Ultrascale+ HBM シリーズの最大規模のデバイスです。このデバイスは、Xilinx 3D Stacked Silicon Interconnect を使用して複数の FPGA ダイをスタックします。これには、2 つの 4GB HBM Gen2 DRAM ダイとともに 2 つの 4GB HBM Gen2 DRAM ダイが同じパッケージに含まれており、パッケージ内ウェーハ間の膨大な帯域幅を可能にします。単一デバイス内での大規模な並列処理機能と大規模なメモリ帯域幅のこの結合により、従来のメモリバウンドアプリケーションでのアクセラレートの要求、拡大をもたらす可能性があります。

Alpha Data 社　ADM-PCIE-9H7

Alpha Data ADM-PCIE-9H7 は、市場で最初の Virtex Ultrascale+ VU37P ボードです。このボードは、280 万の構成可能なロジックセル、60 MB の非常に柔軟なオンチップ（キャッシュ）メモリ、9024 DSP タイル（500 GFLOPS を超える倍精度のパフォーマンスが可能）、460 GB/s メモリ帯域幅、ホストメモリ（またはデュアル OpenCAPI 25Gx8）との Gen3x16 PCIe 接続、および他の FPGA ボードまたは 100G イーサネットネットワークへの接続に使用可能な 48x 25Gb/s リンクを備えています。

このホワイトペーパーでは、実世界のアプリケーションでこのボードの潜在的なパフォーマンスを評価するために、3 つのケーススタディを調査します。それは、多次元 FFT、マージソート、マトリックス乗算です。

多次元 FFT の実装：

FPGA は、多くの航空宇宙・防衛システムでの FFT 実装に最適なデバイスであり、リアルタイムレーダー、ソナー、および通信システムに実装されています。これは、デュアルポートオンチップメモリの非常に高い帯域幅を活用して、アプリケーションに適したビット幅に簡単にカスタマイズできる FFT バタフライ処理エンジン間でデータをバッファリン

株式会社ミッシュインターナショナル　〒190-0004 東京都立川市柏町 4-56-1
TEL: 042-538-7650　https://www.mish.co.jp　E-mail: sales@mish.co.jp

グでき、実装効率が高いためです。大規模なデバイスでの 1 次元 FFT の性能は、IO 帯域幅によって制限される場合がありますが、多次元 FFT は各段階でデータを使用します。HBM アーキテクチャは結果を保存し、チップパッケージ内で効率的にコーナーターンできるためこのタスクに非常に適しています。FPGA を使用すると、FFT 計算用の非常に効率的なアーキテクチャを構築できます。特殊な乗算ロジック（DSP タイル）は、FFT の各ステージに Radix-2 または Radix-4 バタフライコアを実装するための非常に効率的な積和ハードウェアを提供します。ステージ間で、デュアルポートブロック RAM は、読み取りと書き込みを同時に行うことができるバッファーを提供します。 これにより、乗算ハードウェアを完全に利用して、すべてのステージでデータを継続的に実行できるパイプライン実装が可能になります。Figure 1 は、FPGA に実装されたパイプライン FFT の一般的な構造を示しています。

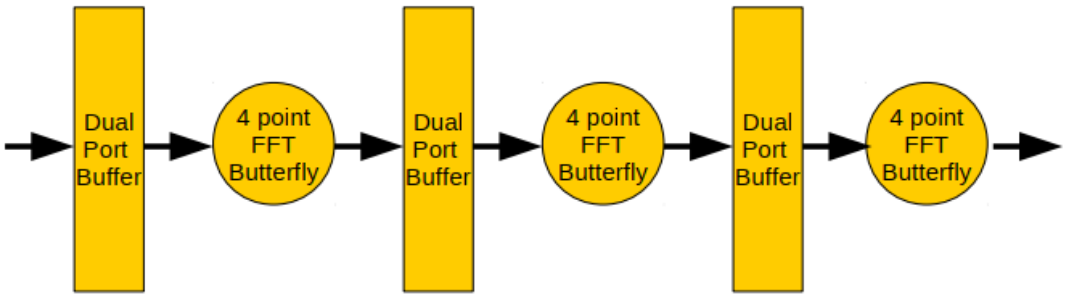


Figure 1　FFT のパイプライン実装

ほとんどの FPGA デザインツールで利用可能な IP コアを使用して、FPGA デザインの効率的な実装をすぐに入手できます。したがって、非常に特殊な要件を満たす必要がない限り、これを最初から実装する必要はありません。Vivado のザイリンクス IP コアは、固定および単精度の浮動小数点実装を提供し、400MHz を超えるクロックレートで実行できます。これらは、さまざまなサイズと構造に構成できます。8192x8192x8192 の複雑な単精度 3D FFT は 4GB の HBM RAM を占有し、8GB VU37P パーツでの効率的なダブルバッファリングを可能にするため、このペーパーでは 8192 ポイントの実装に焦点を当てます。ザイリンクス FFT IP コアは、約 100 個の DSP タイルを使用してクロックサイクルごとに 1 サンプル、26 個の単精度 GFLOPS の効果的なパフォーマンスで、単精度 8k FFT のパイプライン実装を提供できます。これは FPGA のごく一部しか占有しないため、3D FFT アクセラレータの設計では、これらの多くが並列に実装され、HBM メモリとこれらのコアとの間のデータ転送を効率的に実装するのにも役立ちます。

チップ内の並列実装は、処理性能と利用可能なメモリ帯域幅の両方を完全に活用するために不可欠です。HBM メモリは DDR4 ベースであるため、連続した長いバーストを介してアクセスするのが最適であり、ランダムアクセスでデータアクセスがポート幅未満の場合パフォーマンスが低下します。また、メモリは 32 個の独立した並列 256 ビット幅 AXI ポートで構成され、各ポートはチップ内のスイッチロジックを介して HBM アドレス空間全体にアクセスできます。HBM メモリは DRAM デバイスの並列スタックでもあるため、効率的に実現するには、少なくとも 256 ビット幅のアクセスが必要になります。Figure 2 は、一般的な HBM スイッチング構造を示しています。

株式会社ミッシュインターナショナル　〒190-0004 東京都立川市柏町 4-56-1
TEL: 042-538-7650　https://www.mish.co.jp　E-mail: sales@mish.co.jp

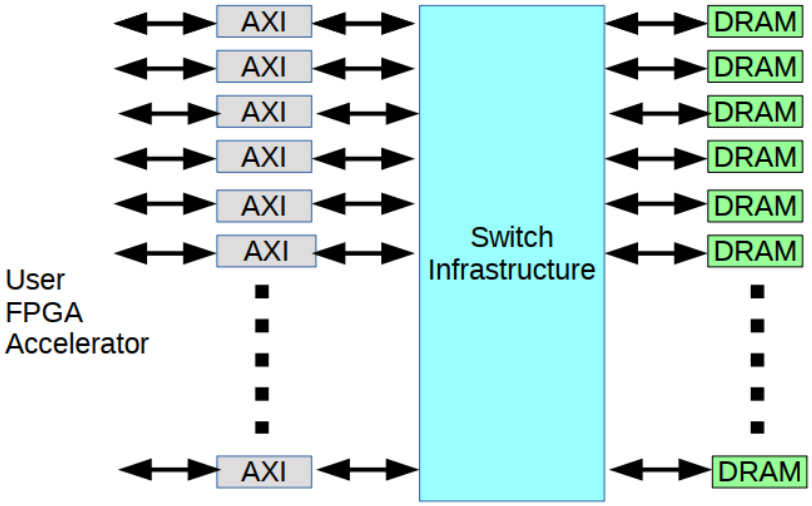


Figure 2　広帯域メモリアクセス構造

ここでは、メモリアクセスと演算性能の比率が重要です。単一の FFT コアでは、26 GFLOPS のパフォーマンスに一致する 3.2 GB/s の連続的な読み取りパフォーマンスが必要です。処理されたデータを書き戻すには、同じ帯域幅が必要です。単一の AXI ポートは、この読み取り帯域幅の最大 4 倍を提供し、それぞれポートが必要な読み取りと書き込みを想定すると、メモリ帯域幅の考慮事項に基づいて、デバイスは最大 64 コアを並行してサポートできます。

多次元 FFT では、データアクセスパターンはデータがメモリに格納される方法に必ずしも適合しない場合があります。3D FFT では、各次元で順番に FFT を実行します。最初のパスで FFT 入力を読み取ると、X 方向は効率的ですが、DRAM から Y 方向と Z 方向の FFT のデータを直接読み取るには、帯域幅の少なくとも 75%を無駄にする非効率的な 64 ビット幅、バースト長 1 の転送が必要です。この解決策は、FPGA 内の内部 RAM でデータのコーナーターンを実行することです。これは、別のアクセラレータコアを使用して、メモリからの読み取り中、FFT の書き戻し中、または FFT 転送の間に実行できます。このペーパーでは、最初のオプションについて検討します。

株式会社ミッシュインターナショナル　〒190-0004 東京都立川市柏町 4-56-1
TEL: 042-538-7650　https://www.mish.co.jp　E-mail: sales@mish.co.jp

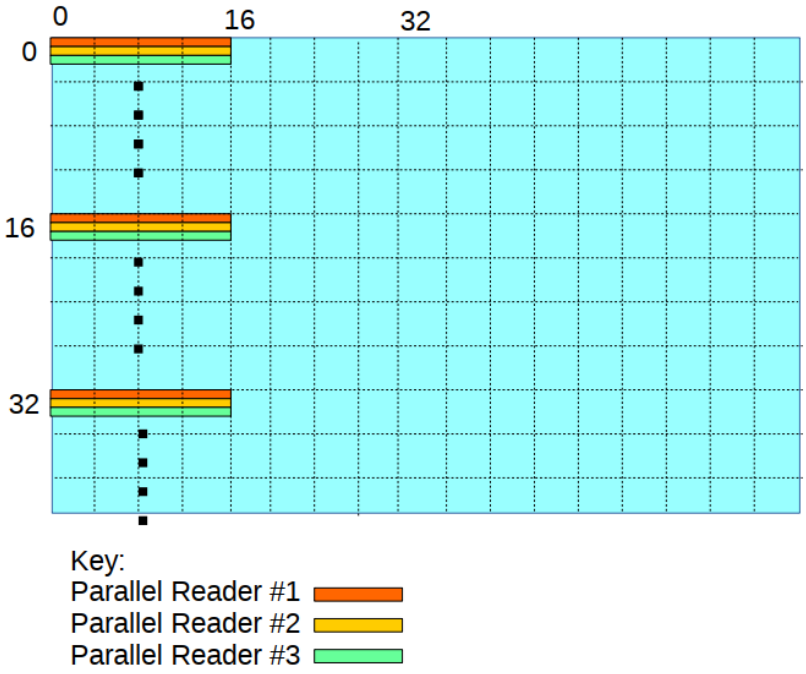


Figure 3　コーナーターニング FFT メモリからの読み取り

このケーススタディでは、各並列 FFT コアがメモリリーダーおよびメモリライターエンジンとペアになっています。メモリリーダーはまず、すべてのパラレルリーダーからのデータを集約し、コーナーターン変換後に出力するコーナーターンブロックを介してデータをプッシュします。X 方向の最初の FFT パスでは、このコーナーターンをバイパスできます。Figure 3 は、コーナーターンを使用した 16 個の並列コアのメモリアクセスパターンを示しています。最初のユニットは、最初の行（オレンジで表示）から 16 チャネル幅のスライスを読み取ります。この DMA 読み取りエンジンによる次の読み取りは、17 行目、33 行目などからです。2 番目の DMA エンジンは行 2,18,34 から読み取り、3 番目は 3,19,35 から読み取ります。コーナーターンブロックは 16 個のパラレル入力ストリームをすべて受信し、16 個のパラレルコーナーターンストリームとして出力します。各 FFT コアは、Y 方向（Z 方向、同じロジックで 16 行ごとに 16 リーダーのそれぞれから 1 サンプルを取得します）行間で異なるアドレスジャンプで動作します。）。

8192 の FFT サイズの選択は、この場合、HBM メモリと 8192x8192x8192 データセットのフットプリントの一致によって動機付けられました。コーナーターニングアプローチは、より大きな FFT サイズの範囲に適しています。メモリ帯域幅により、並列コアの数が事実上 64 個に制限されるため、4k より小さい FFT サイズは効率が低下します。128x128x128 などの小さい FFT サイズでは、FPGA 内のオンチップ SRAM メモリを使用して、FFT ステージ間でコーナーターニングをより効果的に実行できるため、これらのアプリケーションでは HBM の使用はそれほど重要ではありません。

このソリューションを実装する場合のリソースの制限は、FFT コア内のダブルバッファーとして使用される BlockRAM の可用性になります。VU37P には非常に大きなオンチップ SRAM メモリがありますが、その 80%はより大きな UltraRAM ブロックであり、FFT IP コアでは使用されていません。これにより、FFT コアの数が 48 に制限されますが、一般的な 250MHz システムクロックを使用して 0.75TFLOP（単精度）を超える持続動作が可能な 3D FFT コアが得られます。

株式会社ミッシュインターナショナル　〒190-0004 東京都立川市柏町 4-56-1
TEL: 042-538-7650　https://www.mish.co.jp　E-mail: sales@mish.co.jp

並列マージソート:

ソートおよび検索アルゴリズムは、多くのアプリケーションの基本的な構成要素です。浮動小数点演算の要件がなければ、それらはアクセラレータの候補として無視されます。これらのアルゴリズムの最適化に関する研究は、多くの場合、いくつかのヒューリスティックな手段を介して、比較操作の数を最小限に抑えることにのみ焦点を当てています。ただし、FPGA の実装では、比較操作が実際に処理コストと処理時間の重要な部分ではないことは明らかです。データをメモリから比較ユニットに移動したり、比較ユニットに戻したりする際に、はるかに多くのパフォーマンスとエネルギーが使用されます。したがって、比較ロジックとメモリの間に大きな帯域幅を持つ HBM 対応 FPGA は、より優れたソリューションを提供できるはずです。FPGA 内の Ultra RAM およびブロック RAM は、データキャッシングとデータフローのアプリケーション固有の制御を可能にし、高性能で低エネルギーのソリューションを提供するのにも役立ちます。

マージソートは、通常 FPGA 実装で考慮されるアルゴリズムではありません。ただし、大規模なデータセットの場合、これは決定論的で効率的な並べ替えアルゴリズムであり、簡単に並列化できます。算術要件は最小限です。オンチップメモリの柔軟な構成により、非常に効率的な低電力ソリューションを提供できます。計算は最小限であるため、データの移動がこのアルゴリズムのパフォーマンスと電力の要件を支配するため、このアルゴリズムで HBM を使用することには大きな利点があります。ソートのマージは、データをソートするための分割統治アプローチです。確定的な複雑さ O(N $\log_2$ N)を持ち、ベストケースデータ(クイックソートなど)を使用するいくつかのヒューリスティックアルゴリズムよりも遅い場合がありますが、データ依存の問題はありません。また、並列化可能であるため、FPGA 実装により適しています。O(N $\log_2$ N)の複雑さは FFT アルゴリズムのそれに類似しており、前のケーススタディで説明したのと同様のデータフロー構造を使用し、パイプラインを通して log2 N の並列性を作成できます。これにより、ソート時間 O(N)は、データをメモリから読み取り、書き戻すことができるレートと効果的に一致します。

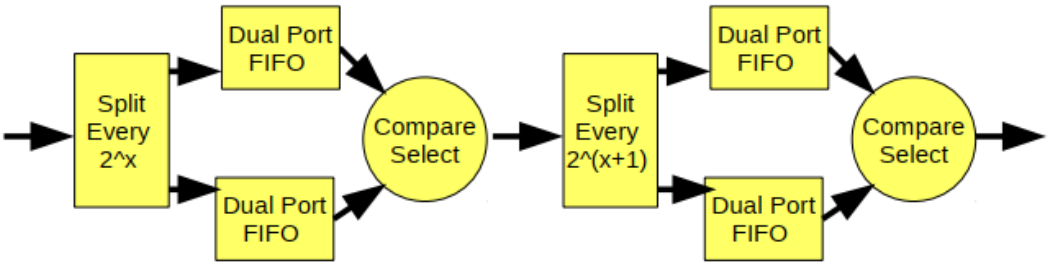


Figure 4　マージソートパイプラインの段階

Figure 4 は、マージソートパイプラインの 2 つの段階を示しています。各段階で、入力ストリームは 2 つの FIFO に分割され、比較演算を使用して最初に大きな出力を選択し、次の段階に掃出します。入力データストリームは、2 つの FIFO のいずれかに代替的に供給されます。最初の段階では、データはソートされていないため、代替要素が各 FIFO に送信されます。後の段階で、入力データは 2 つの長いセクションに分類されるため、この要素数の後に分割が行われます。各ステージで 1 つの FIFO から 2^x 要素が読み取られると、$2^{(x+1)}$ のソートされた要素が出力されるまで、残りのデータが他の FIFO から読み取られます。

比較および選択操作のための FPGA リソース要件は比較的最小限であり、必要なロジックセルはわずかです（280 万個が利用可能と考えられます）。最大のパフォーマンスを得るために、比較演算はハードコーディングされており、例では 64 ビットビッグエンディアン比較（テキスト文字順）ですが、より複雑な比較（たとえば、整数の 64 ビットリトルエンディアン、またはバイトごとの大文字と小文字を区別しない ASCII の比較）を簡単に実装することもできます。実行時に、ソフトウェアで選択可能な比較ユニットをわずかな追加ロジックコストで追加することもできます（必要な可能性のあるすべての比較を効果的に実装します）。データ幅の柔軟性は低く、データ要素は比較に使用される値とその

株式会社ミッシュインターナショナル　〒190-0004 東京都立川市柏町 4-56-1
TEL: 042-538-7650　https://www.mish.co.jp　E-mail: sales@mish.co.jp

他のデータ要素に分割されて、並べ替えに使用される必要があります。サンプルのケーススタディでは、8 バイト値は 8 バイトキーとペアになっており、比較には使用されませんが、参照データベースへの 64 ビットポインターである可能性があります。このアクセラレータでリソースを使用するコンポーネントは FIFO です。各段階で、FIFO の深さは 2^x 要素である必要があります。これにより、FPGA に収まるステージ数に事実上制限が設けられます。さらに、HBM デバイスでは、追加のステージで HBM デバイスのあるポートから別のポートにソートをマージできます。VU37P デバイスを使用して 16 バイト幅のデータレコード（8 バイトの比較可能な値、8 バイトのキー/ポインター）を選択し、分散 RAM、ブロック RAM、ウルトラ RAM から FIFO を構築して 20 の並列ステージを構築し、並列処理を可能にします。メモリポートの読み取り速度で、O（N）時間で 100 万を超える要素（16MB）をソートします。他の HBM ポートを使用して追加のステージを追加すると、さらに 8 種類のソートをパイプラインに配置でき、O（N）時間で 4GB のデータをソートできるようになります。

倍精度行列乗算：

線形代数ライブラリは多くの HPC アプリケーションの中心であり、行列乗算は最も一般的に使用される演算の 1 つです。単純な実装では、行列乗算は O（N^3）ですが、メモリ帯域幅の要件は O（N^2）です。FPGA は、各乗数ユニットの行と列のデータをそのユニットのローカルにキャッシュできる固定サイズの N ユニットを効率的に実装できます。ローカルメモリと処理リソースは N のサイズを制限します。値を大きくするとメモリ帯域幅の制限を克服できますが、値を小さくすると柔軟性が高まり、より多くのマトリックスサイズをより効率的にサポートできます。HBM デバイスは FPGA マトリックス乗算コアのより良い利用を可能にします。通常、固定乗算サイズのニーズは、反復ループ内での加算、転置、スケーリングなどの単純なメモリバウンド操作と組み合わされ、HBM 帯域幅とマルチポート構造により、これらの追加操作が同じデータで可能になります。そして、データセットをデバイス上でローカルに保持します。
O（N^3）よりも優れたパフォーマンスを実現できる行列乗算アルゴリズムがいくつかありますが、これらは特定のデータセットに対して安定性の問題があるか、ほとんどの実用的な用途に適さない他の計算の複雑さを持っています。したがって、ほとんどのアルゴリズムは、未処理の O（N^3）実装を使用する傾向があります。O（N^3）計算の場合、O（N^2）メモリアクセスのみが必要なため、これは必ずしもメモリバウンドではありません。したがって、メモリアクセスに対する計算の比率は O（N）であるため、メモリバウンドの問題を回避するために N を増やすと状況が改善されます。ただし、これを実現するには、適切な行と列のデータを各乗算ユニットの近くにキャッシュする必要があります。CPU システムではこれにより、小さなマトリックスサイズで非常に高速なパフォーマンスが得られ、キャッシュミスが支配的な大きなマトリックスサイズでパフォーマンスが低下します。FPGA 実装では、デュアルポートメモリにより、現在のマトリックス計算が進行している間に次のマトリックス操作データをロードできるダブルバッファードキャッシュの実装が可能になります。アルゴリズム内でデータを再利用する O（N）要素があるため、メモリからキャッシュへの書き込みは計算よりもはるかに短い時間で済み、キャッシュの依存関係が回避されます。FPGA マトリックス乗算コアは、N 個の浮動小数点乗算ユニットで構成できます。各ユニットには、最初のマトリックスの列と 2 番目の行列の行を含む 2 つのローカルキャッシュメモリがあります。これら 2 つのベクトルの乗算を反復処理すると、NxN の結果ごとに部分積が生成され、N-1 倍精度浮動小数点加算器を備えた加算器ツリーパイプラインを使用して、他の乗算ユニットからのすべての積と合計できます。Figure 5 は、行列乗算のためのこのシストリック配列構造を示しています。キャッシュラインの読み取りからマトリックス製品要素の出力までのレイテンシは非常に長く、倍精度の乗算レイテンシに $\log_2$（N）

倍の加算レイテンシを加えたものです。ただし、N^2 乗算に比べて、この時間は小さくなり、最後のデータの読み取りが終了した直後に次の行列乗算を開始できるため、パフォーマンスにとってレイテンシは重要ではない場合があります。

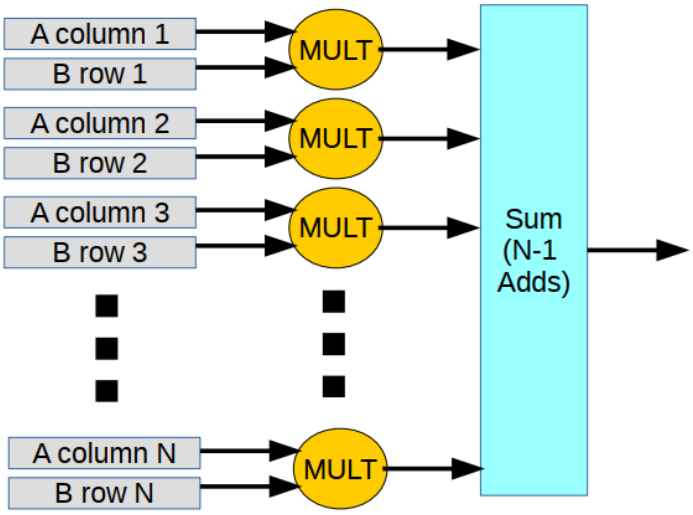


Figure 5　行列乗算シストリックアレイ

ローカルキャッシュは、ブロック RAM またはウルトラ RAM を使用して FPGA に実装されます。これにより、これらのブロックにサイズ制限が設定され、2 のべき乗のサイズになります。N の 2 のべき乗以外の並列化では、これによりメモリ使用効率が低下します。データをダブルバッファリングし、外部メモリからのデータロードを前のデータフレームのマトリックス操作と並行して実行できるようにするには、キャッシュのサイズが少なくとも 2N である必要があります。ブロック RAM コンポーネントは、当然 512 個のディープメモリーにマップされるため、最大 256 ユニットの並列化をサポートします。256 から 512 に移動すると、並列化 N に比例して RAM 要件が増加するだけでなく、512 の並列化を超えて移動する場合と同様に、各ローカルバッファーサイズを 2 倍にする必要があります。この制限を念頭に置いて、使用される行列乗算コアは、可能な限り多くの VU37P を満たすようにスケールアップされました。512 の N を使用した場合、必要な計算リソースは 50%未満でしたが、RAM の 50%以上が必要でした。1024 の並列化までのスケーリングは不可能と思われますが、並列 704x704 倍精度マトリックス乗算器は適合します。250MHz でクロックされた場合、これは約 350 GFLOPS のパフォーマンスになります。

これまでに説明した行列乗算の実装は、メモリ帯域幅が制限ではないため、非 HBM FPGA でも同等に機能します。HBM を使用する利点は、行列乗算演算を他の行列演算と組み合わせると明らかになります。行列乗算コアは 2 つのマトリックスから読み取り、1 つのマトリックスに書き戻すため、最大 3 つの HBM AXI ポートを使用してメモリにアクセスできます。これにより、他の多くのポートが自由に並行して演算を実行できます。1 つの便利な操作は、要素の追加による行列要素です。これは、より大きなマトリックスを乗算するための分割統治アルゴリズムの一部として使用でき、コアがその固定サイズよりもはるかに大きなマトリックスを処理できるようにします。コアは、未使用の行と列をゼロに設定することにより、より小さいマトリックスを処理することもできます。ただし、計算時間は完全な行列サイズと同じ時間かかるため、これは比較的非効率的です。

将来の作業では、単一の大きなマトリックス乗算コアの代わりに複数の小さなコアを使用することの利点と欠点を検討します。これは、HBM へのマルチポートアクセスにより HBM パーツに簡単に実装できます。これは、追加のマトリックス加算操作だけでなく、マトリックススカラー乗算、マトリックスベクトル乗算、マトリックス転置、および要素ご

とのレシプリコルまたは平方根など、大規模な Matrix-Matrix 製品オペレーションの HPC アプリケーションループでよく使用される他のいくつかのマトリックス操作と組み合わせることができます。これらの操作の一部をパイプライン化してメモリ要件を削減できます。たとえば、メモリに書き戻す前に行列積の平方根を計算します。

結論：

このホワイトペーパーでは、Alpha Data ADM-PCIE-9H7 アクセラレータカードで実行される Xilinx Virtex Ultrascale+ HBM VU37P デバイスでの 3 つのアプリケーションケーススタディの実装について説明しました。従来の FPGA は、FFT 実装に非常に効率的なメカニズムを提供しますが、大規模な多次元ワークロードの場合、広帯域幅の並列 HBM インターフェイスは、特に効率的なデータのコーナーターニングを可能にする FPGA オンボードメモリの柔軟性と組み合わせた場合、これらの大きなデータセットを処理するための完璧なバッファーを提供します。並列マージソートは、HBM FPGA デバイスをターゲットにすることもできます。VU37P の大規模な FPGA サイズ、特に Ultra RAM の大規模な提供により、最大 20 の並列ステージでのソートパイプラインの非常に効率的な実装が可能になり、100 万を超える要素のメモリ読み取り速度でソートされます。HBM では、追加の並列ソートステージをパイプラインで動作させ、ソートサイズを数桁拡張できます。大規模な並列行列乗算コアはほとんどの大規模 FPGA に実装できますが、これらの実装は特にメモリ帯域幅の影響を受けません。ただし、HBM を使用すると 1 つ以上の行列乗算コアと、より基本的な低パフォーマンスの追加やスケーリングなどのメモリバウンドオペレーターを簡単に組み合わせて、より複雑な処理ループを実行しながらデータセットをアクセラレータに保持できます。最大 8GB のサイズの同じ作業データセットへのこのマルチポートパラレルアクセスにより、同じ FPGA デザインにアクセラレータの多くの組み合わせを実装できます。たとえば、行列乗算と FFT 処理の組み合わせには多くのアプリケーションがあります。

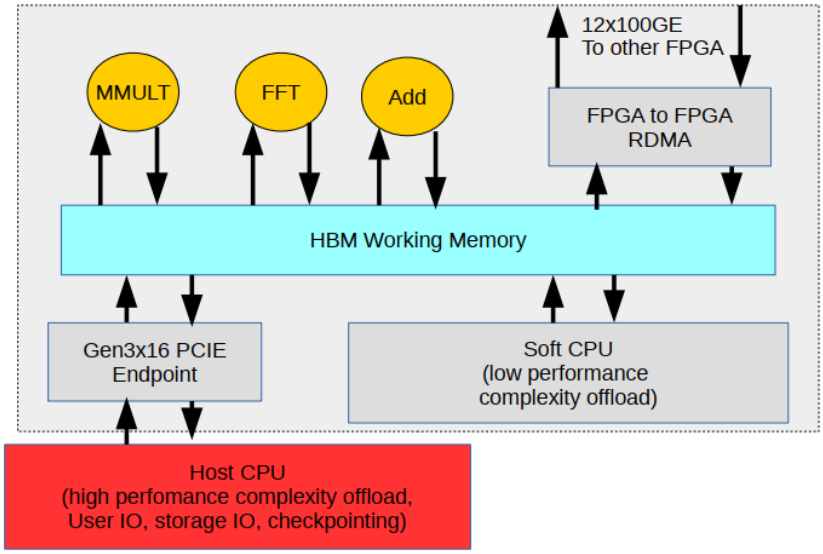


Figure 6　異種コンピューティングアクセラレーションノード

複数のヘテロジニアスアクセラレーションコアを備えたコンピューティングアクセラレーションノードのこのアーキテクチャのアイデアは、ほぼ汎用のメモリとデータフロー中心のパラダイムに拡張できます。アプリケーションの作業

株式会社ミッシュインターナショナル　〒190-0004 東京都立川市柏町 4-56-1
TEL: 042-538-7650　https://www.mish.co.jp　E-mail: sales@mish.co.jp

データは HBM メモリに保持され、処理はこれを中心に動作します。Figure 6 は、そのような構造の 1 つを示しています。FPGA には、さまざまなアクセラレータとワークモジュールが含まれています。シンプルで複雑な計算アクセラレータは、HBM メモリ上で直接動作します。制御およびモードの複雑なタスクは、オンチップソフト CPU、または PCIe を介したホスト CPU によって処理できます。ノード間通信は並行して動作し、クラスター内の異なる FPGA 間で 1.2TB/s の IO 帯域幅を介してデータをシフトします。

Alpha Data 社について

Alpha Data 社は、1993 年に設立され、計算集約型アプリケーションをターゲットとした最先端の FPGA ソリューションを提供しており、FPGA アクセラレータのマーケットリーダとして市場を牽引しています。主な製品は VPX、XMC、PMC、PCI、CompactPCI、PCIExpress、VXS、VME などの A/D, D/A, FPGA ボードや CameraLink 等のデジタル I/F を搭載した信号処理ボードです。ボーイング、ロックウェルコリンズ、JPL、ロッキードマーチン、モトローラ、BAE などのミリタリ向けやデータセンター向けに広く採用されています。Alpha Data 社の詳細については、https://www.alpha-data.com/を参照してください。

株式会社ミッシュインターナショナル 〒190-0004 東京都立川市柏町 4-56-1
TEL: 042-538-7650 https://www.mish.co.jp E-mail: sales@mish.co.jp