WP505 (v1.1.1) 2020 年 9 月 29 日

Versal: 初の ACAP (Adaptive Compute Acceleration Platform)

このたび発表した Versal ACAP は、スカラー エンジン、適応型エンジン、およびインテリジェント エンジンを組み合わせた完全にソフトウェア プログラマブルなヘテロジニアス演算プラットフォームで、データセンター、有線ネットワーク、5G 無線、および先進運転支援システムなどのアプリケーションにおいて現在最速の FPGA インプリメンテーションに比べ最大 20 倍、現在最速の CPU インプリメンテーションに比べ 100 倍以上という劇的な性能向上を実現します。

概要

近年の技術的な課題により、これまでのように CPU スカラー プロセッシング ソリューションであらゆる用途に対応するのは困難となっており、業界は別のアプローチを検討することを余儀なくされています。この問題は非常に大規模なベクター プロセッシング (DSP、GPU) により部分的には解決できますが、メモリ帯域幅を効率よく柔軟に利用できないため、伝統的なスケーリングの課題に直面します。これまでの FPGA ソリューションはメモリ階層をプログラムできますが、ハードウェア フローが足かせとなってデータセンター市場など幅広いアプリケーションでの大規模な導入が阻まれています。

ACAP (Adaptive Compute Acceleration Platform) はこれら 3 つのエレメントをすべて組み合わせ、フレームワークから C、そして RTL レベルのコーディングまで幅広い抽象度に対応した新しいツール フローを提供することによってこの問題を解決します。このまったく新しいカテゴリのデバイスとして登場したザイリンクス Versal™ ACAP では、これら 3 つのプログラマブル エレメントで独自の特定用途向けアーキテクチャ (DSA) をカスタマイズできます。

はじめに

業界はこれまで CPU スカラー演算エンジンの微細化によってあらゆる用途に対応してきましたが、ここにきて半導体プロセスが技術的な課題に直面したことにより、このアプローチは立ちゆかなくなってきました。図 1 に示すように、半導体プロセスの微細化による周波数の向上が鈍化したことで、標準演算エレメントは並列性を高める方向へと舵を切りました [参照 1]。

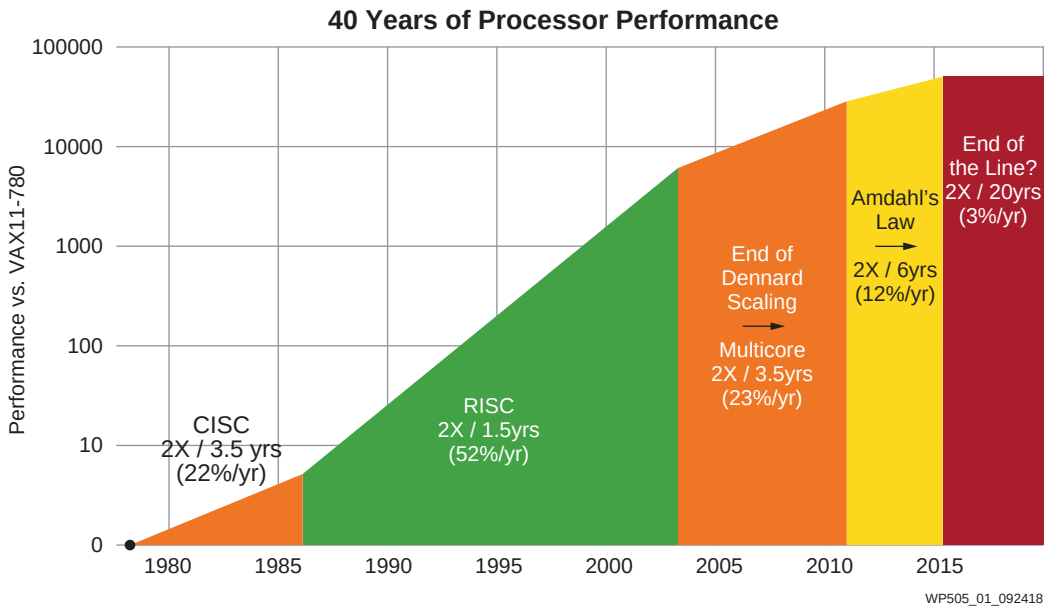

図 1: プロセッサ性能の変遷

この結果、半導体業界はベクター ベース プロセッシング (DSP、GPU) や完全な並列プログラマブル ハードウェア (FPGA) など、これまで一部の高性能分野で利用されていた特定分野向けアーキテクチャを代替アプローチとして検討するようになっています。そこで問題となるのが、「どのアーキテクチャがどのタスクに最適なのか」ということです。

- **スカラー プロセッシング エレメント** (例: CPU): さまざまな決定木と幅広いライブラリを使用する複雑なアルゴリズムは非常に効率よく実行できますが、性能のスケーリングに限界があります。
- **ベクター プロセッシング エレメント** (例: DSP、GPU): 並列化が可能な演算機能では高い効率を発揮しますが、メモリ階層の柔軟性に欠けるためにレイテンシおよび効率のペナルティがあり、適用範囲は限られます。
- **プログラマブル ロジック** (例: FPGA): 特定の演算機能に合わせてきめ細かくカスタマイズできるため、レイテンシの要求が厳しいリアルタイム アプリケーション (先進運転支援システムなど) や不規則なデータ構造 (ゲノム配列決定など) に最適です。ただしアルゴリズムを変更するには数時間かけてコンパイルする必要があります。

図 2 を参照してください。

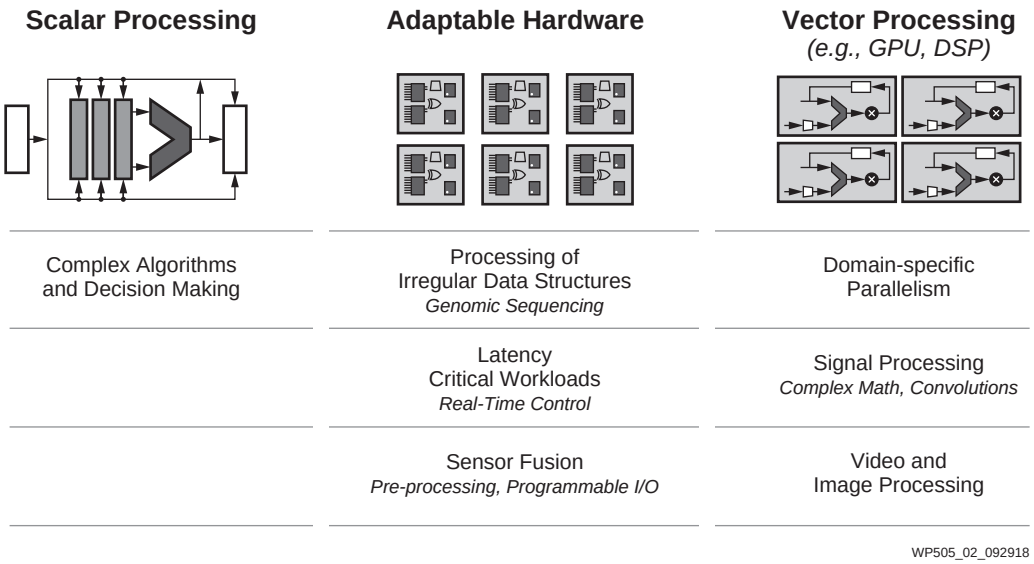

図 2: 演算エンジンの種類

この問題への回答として、ザイリンクスはこれら 3 つのエレメントを最高の形で組み合わせた革新的なヘテロジニアス演算アーキテクチャである ACAP (Adaptive Compute Acceleration Platform) を発表しました。ACAP は業界トップクラスのベクターおよびスカラー プロセッシング エレメントを次世代プログラマブル ロジック (PL) に密結合し、これらすべてを広帯域幅のネットワーク オン チップ (NoC) で接続することにより、これら 3 つのプロセッシング エレメントすべてに対するメモリ マップド アクセスを可能にしています。この密結合されたハイブリッド アーキテクチャでは、先に挙げた 3 つのホモジニアス実装に比べ、カスタマイズ性と性能が劇的に向上します (図 3 を参照)。

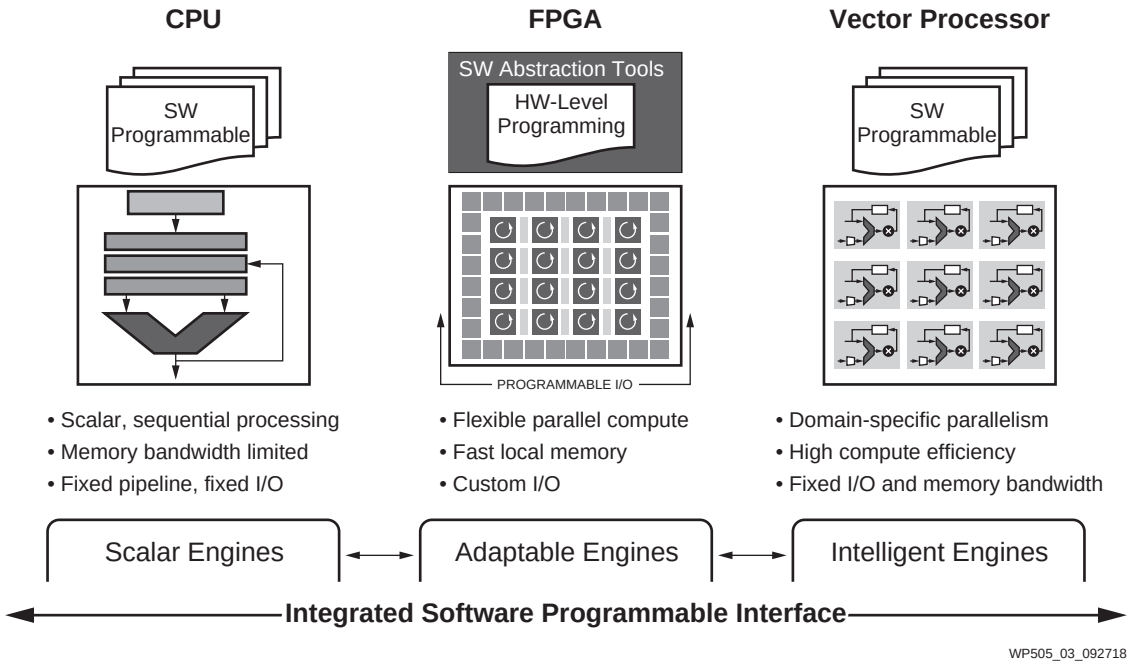

図 3: 3 種類のプログラマブル エンジンをヘテロジニアス統合

こうした劇的な性能の向上に伴い、ツール側にも使いやすさに配慮した同様の劇的な改良が必要となります。ACAP は RTL フローを必要とせず、すぐに扱えるよう設計されています。ACAP はネイティブなソフトウェア プログラミングをサポートしており、C ベースおよびフレームワーク ベースのデザイン フローが可能です。このデバイスは、DMA を統合したキャッシュ コヒーレントなホスト インターフェイス (PCIe® または CCIX テクノロジ)、NoC、および統合メモリ コントローラーで構成されるシェルを内蔵しており、RTL 設計の必要がありません。

新しい ACAP アーキテクチャは、使いやすさも劇的に改善されています。完全に統合された、メモリ マップド プラットフォームである ACAP は、統合型ツールチェーンによってプログラムできます。ザイリンクスのツールチェーンは、開発者の種類に応じていつくものエントリ方法をサポートしています。たとえば、AI 機械学習の推論などのアプリケーションはフレームワーク レベル (Caffe、TensorFlow など) でコーディングできる一方、アプリケーションによっては最適化済みのライブラリ (5G 無線用フィルターなど) を使用して C でコーディングすることもできます。ハードウェア開発の経験があれば、従来の RTL エントリ フローを利用して既存の RTL を ACAP に移植することもできます。

ここまで、伝統的な CPU ベースの演算モデルからの脱却を後押ししているニーズについて概観し、CPU 以外のアプローチについて詳しく見てきました。ここからは、業界初のヘテロジニアス演算プラットフォームであるザイリンクス Versal ACAP をご紹介します。

ACAP の最大の長所は次の 3 つにあります。

1. **ソフトウェア プログラマビリティ** — ソフトウェアにより抽象化したツールチェーンを使用して、最適化したアプリケーションを短時間で開発できます。
2. **アクセラレーション** — 人口知能、スマート NIC (Network Interface Card)、高密度ストレージ、5G 無線、自動運転車、AMR (Advanced Modular Radar)、およびテラビット光ネットワークなど幅広いアプリケーションを高速化します。
3. **動的適応型リコンフィギュレーション** — ミリ秒単位でハードウェアをリコンフィギュレーションし、別のワークロードをアクセラレーションできます。

初の ACAP: 並列ヘテロジニアス演算に最適化したハードウェアとソフトウェア

ACAP は次世代のスカラー エンジン、適応型エンジン、およびインテリジェント エンジンを組み合わせています。これら 3 つのエンジンは、全体で 1Tb/s を超える帯域幅を持つ NoC で接続され、メモリ マップド アクセスが可能です。NoC に加え、プログラマブル ロジック (および内蔵 RAM ブロック) によって実現する広大なメモリ帯域幅により、プログラマブルなメモリ階層を個々の演算タスクに合わせて最適化できます。これにより、一般的なキャッシュ ベースの演算ユニットに比べてレイテンシおよびその確定性が改善されます (図 4 参照)。

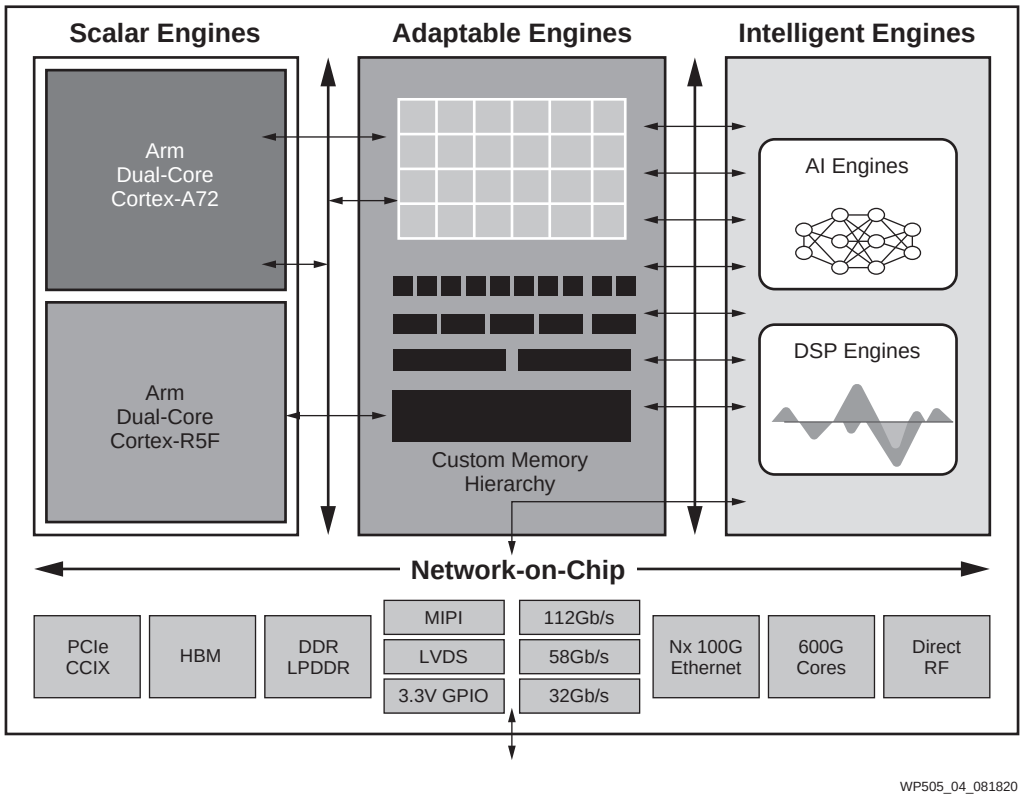

図 4: ザイリンクス Versal ACAP の機能ブロック図

スカラー エンジンはデュアルコア Arm® Cortex-A72 をベースにしており、ザイリンクスの従来世代の Arm Cortex-A53 コアに比べ、コアごとのシングル スレッド性能が 2 倍に向上しています。
先進のアーキテクチャと 7nm FinFET プロセスによる省電力効果により、従来の 16nm に比べ単位ワットあたり DMIPs も 2 倍に向上しています。ASIL-C 認証済みの(1) UltraScale+™ Cortex-R5F スカラー エンジンも 7nm に移行し、自動車分野でのザイリンクス デバイスの豊富な導入実績から得た知見に基づいて、さらに多くのシステムレベルの安全機能を追加しています。

適応型エンジンは、プログラマブル ロジックとメモリ セルを業界最速の次世代プログラマブル ロジックで接続することで構成されています。これらの構造はレガシ デザインをサポートするだけでなく、再プログラムすることによって特定の演算タスクに合わせてカスタマイズしたメモリ階層を構築できます。これにより、ザイリンクスのインテリジェント エンジンは同じ演算量なら最新の GPU および CPU に比べ高いサイクル効率およびメモリ帯域幅を大幅に向上させます。このことは、エッジ側でのレイテンシと消費電力の最適化、およびコア側での絶対性能の最適化に大きく役立ちます。

インテリジェント エンジンは、革新的な VLIW (Very Long Instruction Word) および SIMD (Single Instruction, Multiple Data) プロセッシング エンジンとメモリをアレイに配置し、これらすべてを数百 Tb/s のインターコネクトとメモリ帯域幅で相互接続しています。これにより、機械学習およびデジタル信号処理 (DSP) アプリケーションの性能が 5 ～ 10 倍に向上します。

Versal ポートフォリオは、これらの演算機能をさまざまな配分で組み合わせたデバイスをラインナップしています (表 1 参照)。

表1: Versal ポートフォリオのデバイス、市場、および主な特長

Versal ポートフォリオ	主な市場	主な特長
Versal AI コア	データセンター、無線	最も多くのインテリジェントエンジンを内蔵
Versal AI エッジ	オートモーティブ、無線、放送、A&D	熱エンベロープを 5W に抑え、インテリジェント エンジンの数と電力効率を最適化
Versal AI RF	無線、A&D、有線	ダイレクト RF コンバーター、SD-FEC
Versal プライム	データセンター、有線	シェルを内蔵したベースライン プラットフォーム
Versal プレミアム	有線、データセンター、A&D、テスト/計測	最も多くの適応型エンジン、112G SerDes および統合された 600G IP を内蔵したプレミアム プラットフォーム
Versal HBM	データセンター、有線、テスト/計測	プレミアム プラットフォームに HBM を追加

ザイリンクスの ACAP はベクター、スカラー、および適応型ハードウェア エレメントを統合することにより、次の 3 つの強力な利点をもたらします。

- 「ソフトウェア プログラマビリティ」
- 「ヘテロジニアス アクセラレーション」
- 「適応性」

1. https://japan.xilinx.com/news/press/2018/availability-of-automotive-xa-zynq-ultrascale-plus-mpsoc.html

ソフトウェア プログラマビリティ

適応型のシリコンによって適応型アクセラレーションが実現

Versal ACAP は、ソフトウェアで容易にプログラムできる適応型アクセラレーション ハードウェアを備えています。これらのヘテロジニアス エンジンにより、あらゆる種類のソフトウェア アプリケーションを最大限に高速化できます。インテリジェント エンジンは、機械学習および一般的な古典的 DSP アルゴリズムを高速化します。適応型エンジンに含まれる次世代プログラマブル ロジックは、並列化が可能なアルゴリズムを高速化します。その他のアプリケーションには、マルチコア CPU の包括的な内蔵演算リソースで対処します。Versal デバイス全体は、ハードウェアの専門知識がなくてもソフトウェアを用いて容易にプログラムできるように設計されています (図 5 参照)。

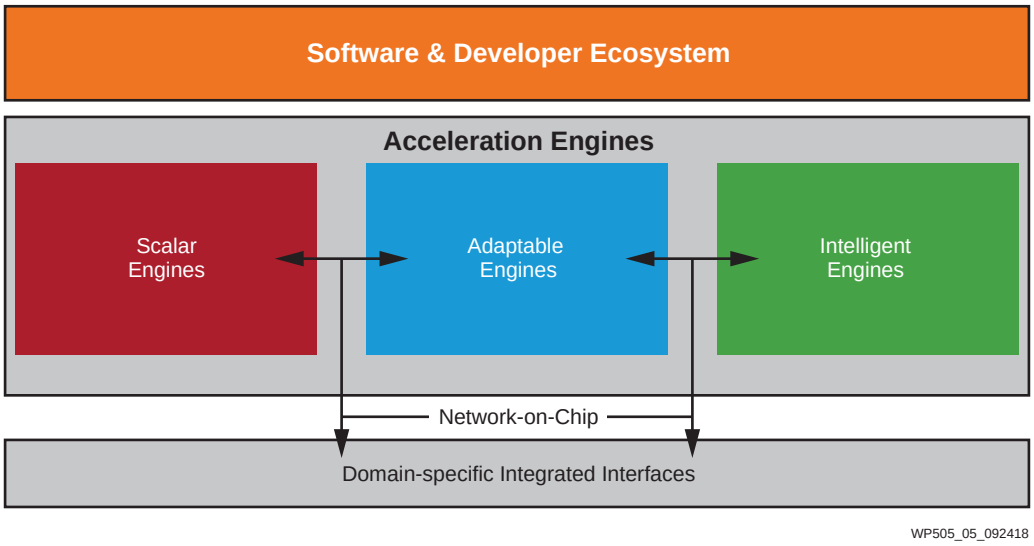

図 5: Versal ACAP の最上位概念図

- データ/AI サイエンティストは標準ソフトウェア フレームワークで構築したアプリケーションを展開して数桁の高速化が可能です。
- ソフトウェア アプリケーション開発者は、ハードウェアの専門知識がなくてもザイリンクスの統合型ソフトウェア開発環境を使用してあらゆるソフトウェア アプリケーションを高速化できます。
- ハードウェア設計者はこれまで同様 Vivado® Design Suite を使用した設計が可能で、Versal プラットフォームの統合 I/O インターフェイスおよび NoC インターコネクトを使用して開発期間を短縮できます。

図 6 を参照してください。

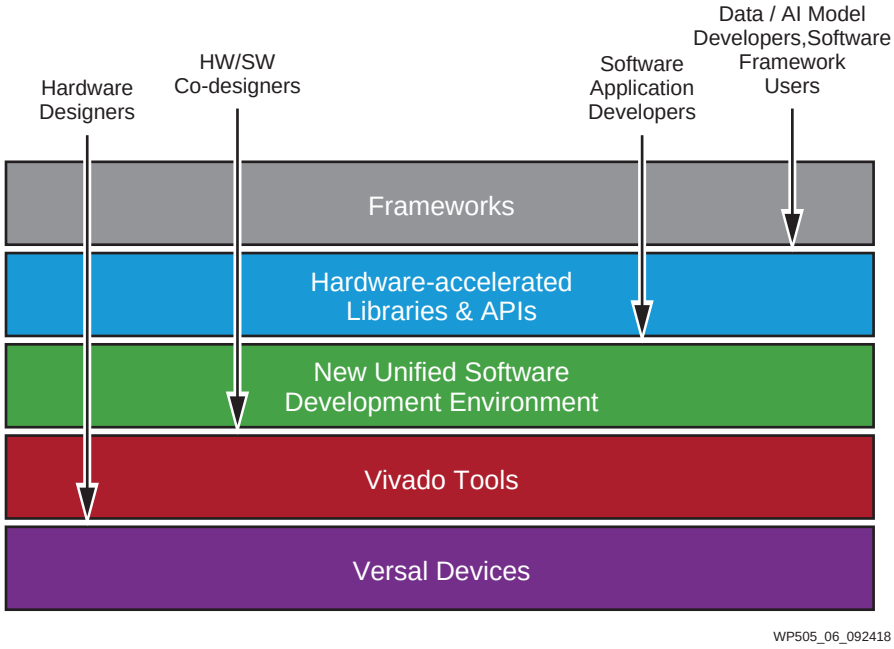

図 6: 開発者別に見た Versal プラットフォームのソフトウェア プログラマビリティ

専用ハードウェアにより使いやすさとアプリケーション効率が向上

オフチップ インターフェイスには、適応型のインターフェイス ロジックによって簡単にアクセスできます。これには、外部ホスト プロセッサへの標準インターフェイスも含まれます。データセンター アプリケーションでは、一般的にソフトウェア アプリケーションはエンベデッド マイクロプロセッサではなくホスト CPU で実行されます。ホスト CPU と Versal プラットフォームのプログラマブル リソースを接続するインターフェイスをシェルと呼びます。この統合シェルには、規格に完全準拠した CCIX (アクセラレータ向けのキャッシュ コヒーレント なインターコネクト) または PCIe Gen4x16 ホスト インターフェイス、DMA コントローラー、キャッシュ コヒーレンシ メモリ、統合メモリ コントローラー、高度な機能安全、およびセキュリティ機能が内蔵されています。

NoC が提供するメモリ マップド インターフェイスにより、各ハードウェア コンポーネントとソフト IP モジュールの相互アクセス、およびソフトウェアからのアクセスが容易です。NoC は標準化されたスケーラブルなハードウェア フレームワークを提供し、ヘテロジニアス エンジンとインターフェイス ロジック間における効率的な通信を可能にします。

ヘテロジニアス アクセラレーション

近年では、CPU よりもプログラマブル ロジック (FPGA) やベクター ベース (DSP、GPU) の実装に性能面での優位性があることが実証されていますが、ACAP アーキテクチャの真価は、Versal ACAP の複数の演算エレメントを利用して密結合型の演算モデルを構築した場合に最大限に発揮されます。この場合、1+1+1 は 3 を大きく上回ります。

表 2 に、各種市場における Versal ACAP デバイスの利点をまとめます。

表2: Versal ACAP とターゲット市場

市場	ベンチマーク	対 CPU 比	対 GPU 比	対 FPGA 比	説明
データセンター	画像認識 (推論) - レイテンシ制約なし	43X	2X	5X	GoogLeNet v1 (バッチ サイズ無制限)
	画像認識 (推論) - レイテンシ制約 2ms	N/A	8X	5X	GoogLeNet v1 (< 2ms) CPU の下限レイテンシは 5ms
	リスク分析	89X	N/A	>1X	金利スワップのバリュー アット リスク (VaR) (Maxeler 社データ)
	ゲノミクス	90X	N/A	>1X	ヒト遺伝子の解析結果 (Edico Genome 社データ)
	Elasticsearch	91X	N/A	>1X	1TB のデータでレイテンシを 1/91 に削減 (BlackLynx 社データ)
5G 無線	16x16 5G リモート無線	N/A	N/A	>5X	5G リモート無線の無線帯域幅が 5 倍以上に拡大
	ビーム フォーミング	N/A	N/A	>5X	5 倍以上の演算性能
A&D レーダー	DSP TMAC	N/A	N/A	>5X	>27 TMAC
	アルゴリズム反復時間	N/A	N/A	>100X	ソフトウェア プログラマブルなインテリジェント エンジンを数分でコンパイル
オートモーティブ	低レイテンシの推論 (<2ms)	N/A	3x	15X	ResNet50 (バッチ =1) 低レイテンシが要求される安全系 ADAS/自動運転にも AI エンジンはスケーラブルに対応
	エンクロージャ タイプ	1	2	4	<10W、20W、30W、およびトランク設置型のエンクロージャをすべて効率よくカバーできるのは ACAP ポートフォリオのみ
有線	暗号化ネットワーク トラフィック	N/A	N/A	4X	ネットワーキングおよび暗号化 IP を ACAP に統合することで、数 Tb/s のシングルチップ インプリメンテーションが可能

データセンター人工知能: 機械学習の推論アクセラレーション

現代社会に人工知能が浸透するにつれ、より高い演算効率を求めることが半導体業界におけるイノベーションの原動力となっています。しかしホモジニアスな実装で最大限の効率を達成するのは困難です。この分野において、ベクター プロセッシングとプログラマブル ハードウェアの密結合は高い価値をもたらします。

これまで演算ユニットの精度 (FP32、FP16、INT16、INT8 など) についてはさかんに議論されてきましたが、ネットワークの種類によってメモリ階層の要求が大きく異なることについてはあまり注意が払われておらず、最新の AI 推論エンジンの多くはネットワークの種類が変わると効率が大きく低下するという問題を抱えています。たとえば、現在最先端の機械学習推論エンジンでピーク性能を達成しようとすると 4 つの HBM メモリ (7.2Tb/s の外部メモリ帯域幅) が必要ですが、キャッシュ ベースのメモリ階層は動作効率が約 25 ～ 30% にとどまっており、リアルタイム アプリケーションではレイテンシの不確実性が大きな問題となります。
この問題は、FPGA の大規模並列ロジックによって実現するプログラマブルなメモリ階層をネットワークの種類に合わせてきめ細かく最適化し、インテリジェント エンジンによって実行されるベクター プロセッシングと組み合わせることによって解決できます。

たとえば GoogLeNet を Versal プラットフォームに実装すると、レイテンシの制約がないアプリケーションでは現在最先端の Skylake Platinum CPU(2) の 43 倍、現在最先端の GPU [参照 2] の約 3 倍という圧倒的なスループットが得られると同時に、消費電力は大幅に削減されます (図 7 参照)。

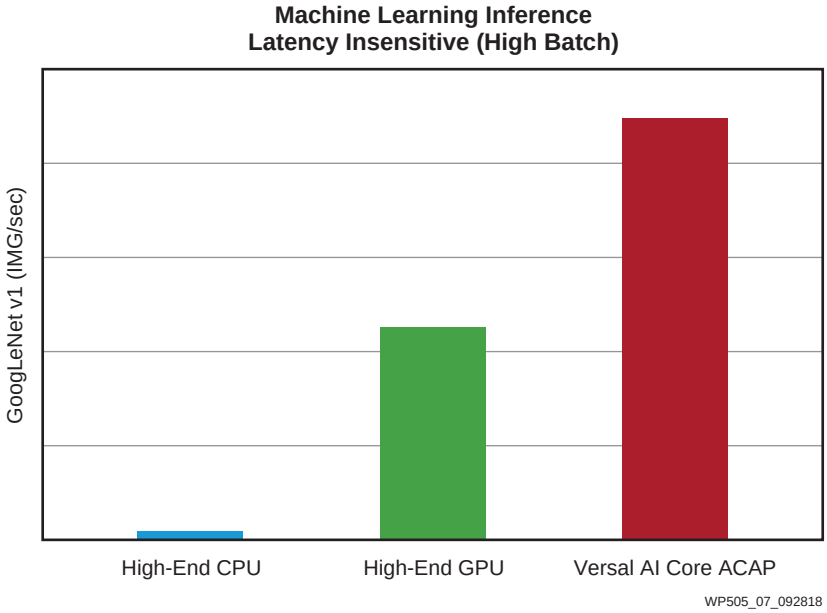

図 7: GoogLeNet の性能 (レイテンシ <7ms) = ハイエンド CPU の 43 倍[1, 2]

1. Xeon Platinum 8124 Skylake、c5.18xlarge AWS インスタンス上で測定。Intel Caffe: https://github.com/intel/caffe
2. V100 のデータは Nvidia 社の技術概要『Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services』より引用。

ニューラル ネットワークのデータセンター アプリケーションが進歩を続ける中、複数のニューラル ネットワークがチェーン接続されるようになり、低レイテンシのニューラル ネットワーク性能に対する要求が大幅に高まっています。たとえば話し言葉のリアルタイム翻訳では、音声からテキストへの変換、自然言語処理、レコメンダ システム、テキストから音声への変換、そして音声合成が必要となります [参照 2]。つまりこのアプリケーションでは、全体のレイテンシ バジェットにおけるニューラル ネットワークのレイテンシが 5 倍になります。

2. Xeon Platinum 8124 Skylake、c5.18xlarge AWS インスタンス、Canonical、Ubuntu 16.04LTS、AMD64 Xenial イメージ (ビルド日: 2018-08-14)、Intel Caffe、Git バージョン: a3d5b02、`run_benchmark.py` (修正なし)。

リアルタイム アプリケーションの数が増大の一途をたどる中、データセンター カスタマーにとっては、将来のニーズに合わせて拡張可能なテクノロジを選択することが重要となります。現在、次の 2 つのトレンドが台頭しています。

- ソフトウェアの設計効率を高めるために確定的レイテンシの重要性が高まっています [参照 3]。
- これまで以上に複雑な相互作用のモデル化 (ヒューマン コンピューター インタラクション、金融取引)、および自動車や産業用など安全系アプリケーションの重要性の増大により、ニューラル ネットワークのレイテンシ要件が厳しさを増しています。

これら 2 つの要件を満たすにはバッチ処理をなくす必要があり、そうするとキャッシュ ベースのメモリ階層が固定された CPU および GPU ベースのソリューションは性能が大幅に低下してしまいます。CPU はハイエンド製品でもレイテンシ 5ms が限界であり、ハイエンド GPU さえもレイテンシ 7ms 未満では性能が大きく低下します。レイテンシ 2ms で許容可能な性能を達成できるのは Versal ACAP のみです (図 8 参照)。

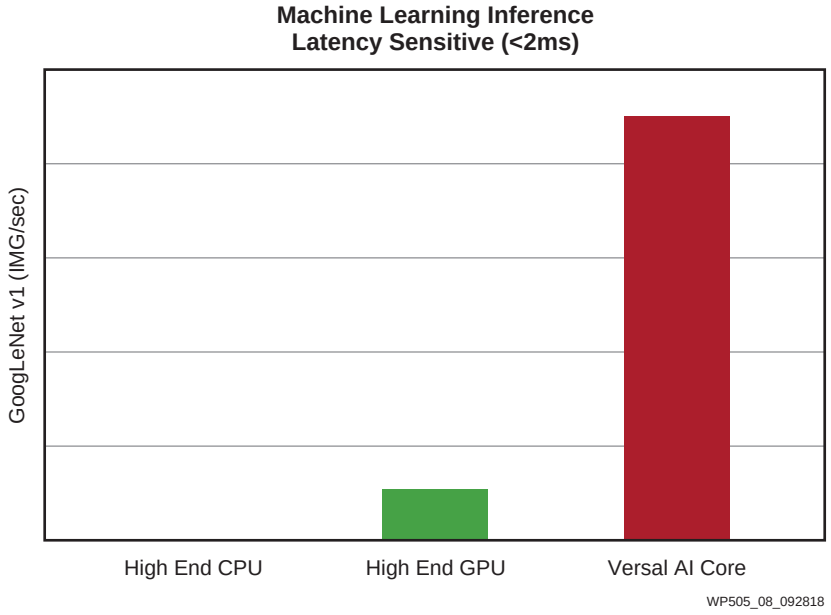

図 8: GoogLeNet のリアルタイム性能 (レイテンシ <2ms) = ハイエンド GPU (Nvidia) の 8 倍[1, 2]

1. Xeon Platinum 8124 Skylake、c5.18xlarge AWS インスタンス上で測定。Intel Caffe: https://github.com/intel/caffe
2. V100 のデータは Nvidia 社の技術概要『Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services』より引用。

この結果、ACAP ベースのソリューションが持つ独自のプログラマブルなメモリ階層は、機械学習の推論において最高の性能を達成するだけでなく、レイテンシの削減と確定性が要求される今後のアプリケーションにもこれまで以上にスケーラブルに対応できます。

データセンター SmartNIC

ネットワーク インターフェイス カード (NIC) はもともと、単純なコネクティビティとして開発されました。その後、各種ネットワーク アクセラレーション (暗号化、ハイパーバイザー ネットワーキング オフロード、バーチャル スイッチング) の機能を取り込むことにより、「SmartNIC」へと変化してきました。Amazon 社の Annapurna プロジェクトでは、すべてのハイパーバイザー機能を CPU からオフロードすることにより、CPU サイクルの 100% を収益に関係する演算に集中させることに成功しています。

今後さらに SmartNIC が進歩していくと、3 つの新しい利点が生まれてくるとザイリンクスは考えています。それは、データセンターのイーサネット ロジック全体に対してワークロードを動的に分散、拡張できること、リコンフィギュレーション可能なアクセラレーション プールを利用して任意の演算機能を高速化できること (クラウド リソースの最大活用)、そしてネットワーク データ プレーンで演算機能をインライン実行できることです。

ザイリンクス Versal ACAP デバイスでは、ベクター ベースとプログラマブル ロジックのハイブリッド演算エンジンを利用して NIC 機能を統合でき、これらすべてがザイリンクスの豊富なネットワーキング IP および業界トップクラスの SerDes (次世代 NIC-TOR (Top of Rack) リンクに向けたシングル チャネル 112G SerDes を含む) によってサポートされます。

しかも、これらの NIC リソースはワークロードの種類に合わせて動的にリコンフィギュレーションまたは再展開が可能です。

表3: データセンター NIC の種類

	説明	機能	例
タイプ 1	基本的なコネクティビティ NIC	• 基本的なオフロード (チェックサム、LSO、RSS) • SR-IOV (Single Root I/O Virtualization) • 一部のトンネル オフロード (VXLAN、GRE0)	• Fortville • ConnectX • NetExtreme
タイプ 2	ネットワーク アクセラレーション用 SmartNIC	• 暗号化/復号化 (IPSec) • バーチャル スイッチ オフロード (OVS など) • プログラマブルなトンネル タイプ	• **ザイリンクス タイプ 2** • LiquidIO • Annapurna • Innova
タイプ 3	ネットワーク演算アクセラレーション用 SmartNIC	• インライン機械学習 • インライン ビデオ トランスコード • データベース分析 • ストレージ (圧縮、暗号化、重複排除)	• **ザイリンクス タイプ 3** • MSFT (NIC+FPGA)

データセンター ストレージ アクセラレーション

FPGA は長い間、ストレージ ドライブ内でエラー訂正やライト レベリングに使用されてきました。変化の速いフラッシュ テクノロジの世界では、FPGA の柔軟な I/O が設計再利用の面で特に重要な役割を果たしています。また、現在のデータベース検索/アクセラレーション アプライアンスの多くは、ドライブのすぐ横で FPGA ベースのアクセラレーションを実行することで大きな成果をあげています(効率は演算エレメントをドライブのすぐ横に配置することで最大化する)。

ACAP アーキテクチャなら、ドライブおよび DB アクセラレーション ベンダーは既に FPGA が使用されているドライブ内部に機械学習の演算エレメントを直接追加できるため、データセンター内でのデータ移動 (およびそれに伴うレイテンシ、消費電力、および運用コスト) を 1/10 に削減できます。

5G 無線通信

無線加入者による帯域幅への要求はとどまるところを知らず、無線業界では「10 年で 10 倍」という猛烈な速度でイノベーションが進んでいます。2020 年のオリンピック開催に合わせ、業界は第 5 世代移動通信技術 (5G) の一般向けデモンストレーションを開始する予定です。これらの初期実装の大半は、ザイリンクスの既存デバイス (特に量産実績の豊富な 16nm RFSoC デバイス) で構築されることになります。この RFSoC デバイスには、次の 3 つの重要な利点があります。

- ダイレクト RF サンプリング ADC および DAC を内蔵
- LDPC およびターボ SD-FEC (Soft-Decision Forward Error) 訂正符号ブロックを内蔵
- 16nm FinFET プロセス テクノロジによる電力効率に優れた DSP

ただし量産立ち上げには 2 つの課題を解決する必要があります。1 つは低コストでスペクトラムを拡大していくこと、そしてもう 1 つは機械学習の推論技術を無線に追加し、ビーム ステアリング アルゴリズムと加入者ハンドオフ アルゴリズムの改良、および自己修復型ネットワークの実現を図っていくことです。

従来は、ベクター DSP ベースの ASIC 実装にすることでコスト削減を図る無線ベンダーもありました。Versal ACAP は、インテリジェント エンジンの内蔵によりシングル チップで 5 倍の TMAC を実現できるため、ASIC と FPGA の従来のコスト格差が大幅に縮小します (図 9 参照)。

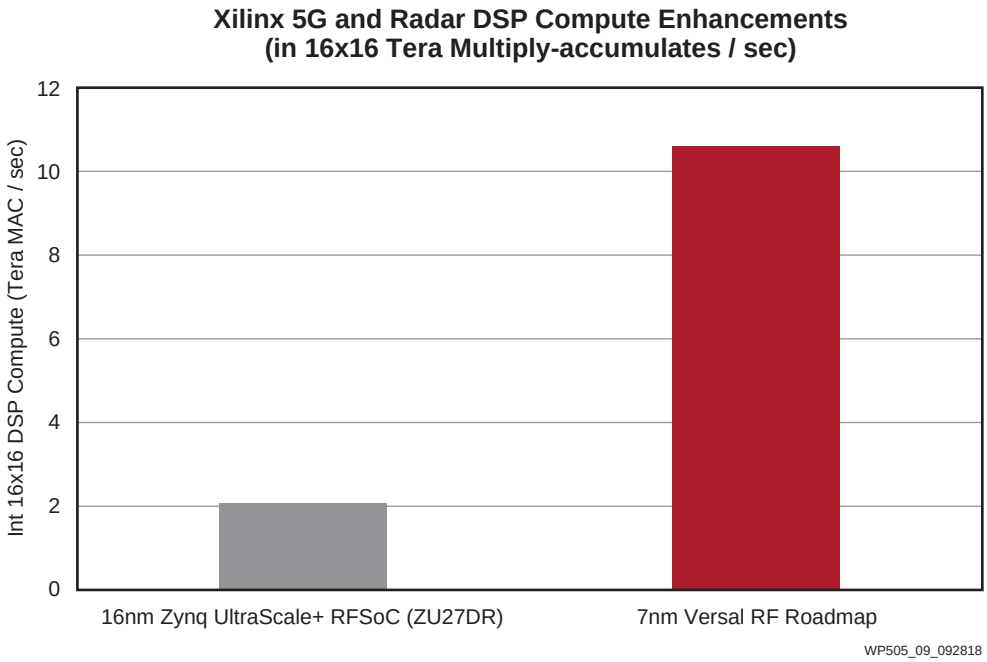

図 9: ザイリンクスの RF 演算のロードマップ

この結果、16nm Zynq® UltraScale+ RFSoC デバイス 1 つで実装できるのが 200MHz 16x16 のリモート無線ユニット (RRU) であるのに対し、7nm Versal デバイスのロードマップでは 800MHz 16x16 の RRU を完全に実装できます (図 10 参照)。

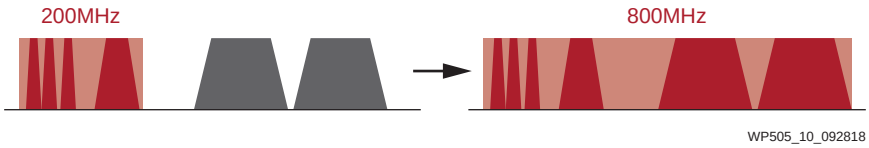

図 10: 16nm と 7nm の無線デバイスがシングル チップでサポートできるスペクトラム

ACAP ベースの Versal ポートフォリオは、電力効率に優れた機械学習 (フレームワーク レベルのデザイン フロー) を追加しているという点において、他に類を見ないデバイスとなっています。このテクノロジにより、ビーム ステアリングおよび加入者ハンドオフのアルゴリズムは従来のプログラム定義によるアルゴリズムの 2 倍に向上し、理論上の限界の 85% まで到達します (図 11 参照)。

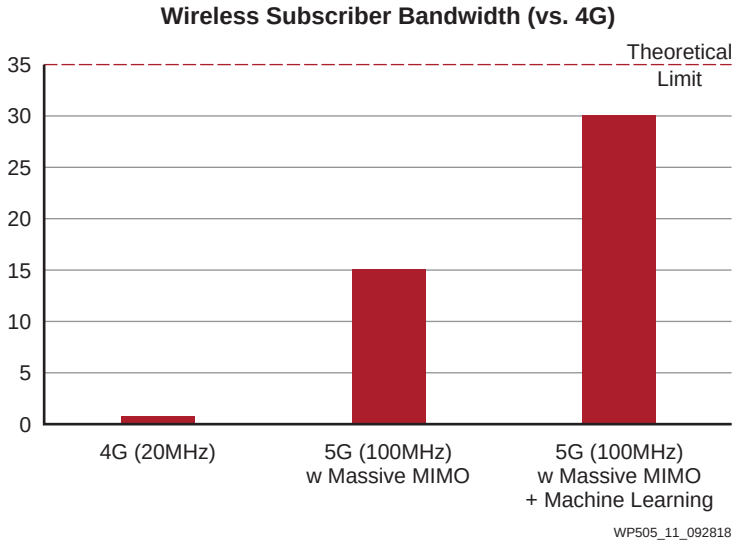

図 11: 無線帯域幅の最適化と理論上の限界

ザイリンクスはダイレクト RF サンプリング ADC/DAC、統合型 SD-FEC 符号、高密度のベクター ベース DSP、およびフレームワークでプログラム可能な機械学習推論エンジンという 4 つの重要なクノロジすべてをワンチップに統合した業界唯一のベンダーで、これによって業界初の完全な 5G RoC (Radio-on-Chip) を実現しています。

航空宇宙/防衛

FPGA の 大規模並列 DSP 機能は、これまで長期にわたって多くの防衛レーダー実装のバックボーンとして採用されてきました。ところが近年、ADC の技術革新が進んだことにより ADC のサンプル レートは数百 GSPS にも達しており、DSP にも相応の高性能化が要求されるようになっています。

強力なベクター ベースの DSP エンジンと AI 機械学習を組み合わせることによって、航空宇宙および防衛産業では AMR (Advanced ModularRadar) などの画期的な新製品が登場しています。これらの製品では高周波の波長を使用するため、アンテナの間隔を考慮すると極小のフォーム ファクターが求められます。ザイリンクスは、数 Tb/s のアンテナ帯域幅および最大 17TMAC の INT24、または 24TFLOPS の 32 ビット単精度浮動小数点 DSP を 1 つのパッケージに統合したデバイスを提供しています。

先進運転支援システム (ADAS)

ザイリンクスには、自動車、航空宇宙、衛星、医療、および商用ネットワーキング システムなど熱制約の大きいシステムで高い信頼性を実現してきた長年の実績があります。ザイリンクスのテクノロジはシングル イベント アップセット (SEU) の影響を軽減しながら最大 125°C の温度で動作するように設計されています。マシン ビジョンおよび機械学習への積極的な取り組みもあり、信頼性と品質に定評のあるそのテクノロジは、先進運転支援システム (ADAS) および将来の自動運転技術に最適な選択肢となっています。これまで、ザイリンクスは自動車業界に向けて 1 億 5000 万個を超える FPGA および SoC を出荷しており、ADAS アプリケーションだけでもその数は 5000 万を超えています。ザイリンクスにとって、過去 2 年間で最も急成長しているのが自動車市場です。

ザイリンクスは、電力効率に優れたデュアルコア Cortex-R5F 内蔵スカラー エンジン、プログラマブル I/O、および低レイテンシのインテリジェント AI エンジンを組み合わせた Versal ACAP をスケーラブルに提供します。特に AI エンジンは、現在の FPGA をベースにした ASIL-C 認証済み[3] ADAS ソリューションに比べ 15 倍の INT8 機械学習性能を発揮し、電力効率と機能安全を重視した AI 活用型の自動運転ソリューションを実現します。さらに、無線経由 (OTA) のハードウェア アップデートによりデバイス全体を再プログラムでき、フィールドにおけるシステムの柔軟性が向上することも、顧客にとっての大きな付加価値となります。また、センサーの種類を変更する場合もザイリンクスのプログラマブル I/O であれば、ASSP や GPU のリスピンにかかる時間およびコストが生じず、ベンダーにとっての柔軟性と適応性が向上します (図 12 参照)。

3. https://japan.xilinx.com/news/press/2018/availability-of-automotive-xa-zynq-ultrascale-plus-mpsoc.html

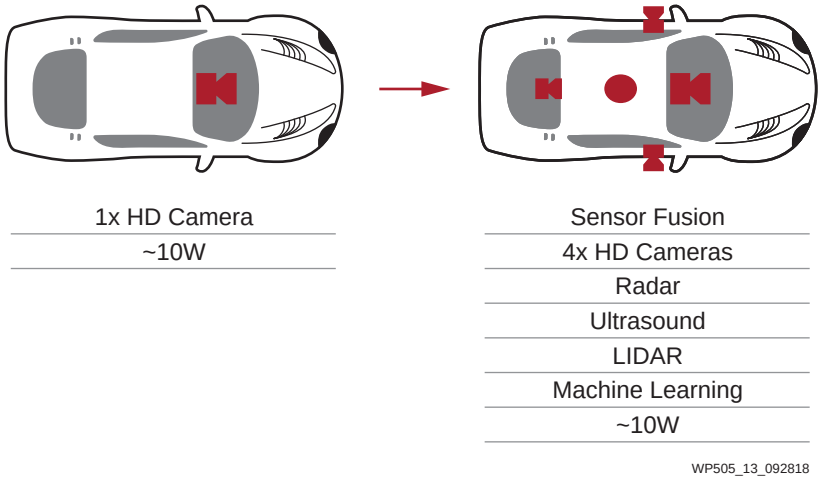

図 12: 少ない消費電力でセンサー フュージョンが可能な ACAP デバイス

技術革新がめざましい自動車市場では、フロントガラスに設置される 5 ～ 10W の小型カメラから車室内に設置される 20 ～ 30W のセントラル モジュール、さらにはトランク内に設置される 100W 超の水冷スーパーコンピューターまで幅広いプラットフォームを同じプログラミング モデルでサポートし、コードの移植性とスケーラビリティを備えたプロセッシング デバイス ポートフォリオを選択することが重要です (表 4 参照)。

表 4: ザイリンクスと競合他社の車載向け製品の幅 (同一プログラミング モデル)

	(10W) インテリジェント エンドポイント (例: フロント カメラ)	(20W) セントラル モジュール (基本、パッシブ冷却)	(30W) セントラル モジュール (高度、強制空冷)	(100W+) トランク内スーパーコンピューター (水冷)
ザイリンクス	●	●	●	●
Nvidia		○	●	●
Intel MobilEye	●			

レイテンシが処理性能の特に重要な要素であることは、自動車の走行速度を考えるとわかります。たとえば 60MPH (100km/h) の場合、ADAS システム間で応答時間が数ミリ秒違うだけでシステムの効果に大きな影響が及びます。自動運転技術が今後発展してくると、複数のニューラル ネットワークを直列にチェーン接続して複雑なタスクを実行することも必要になってきます。そうなると、バッチ サイズの大きい GPU 実装は不利になります。ザイリンクスの Versal AI エッジ シリーズは、小さいバッチ サイズでもきわめて高い動作効率が得られるように最適化されています (図 13 参照)。

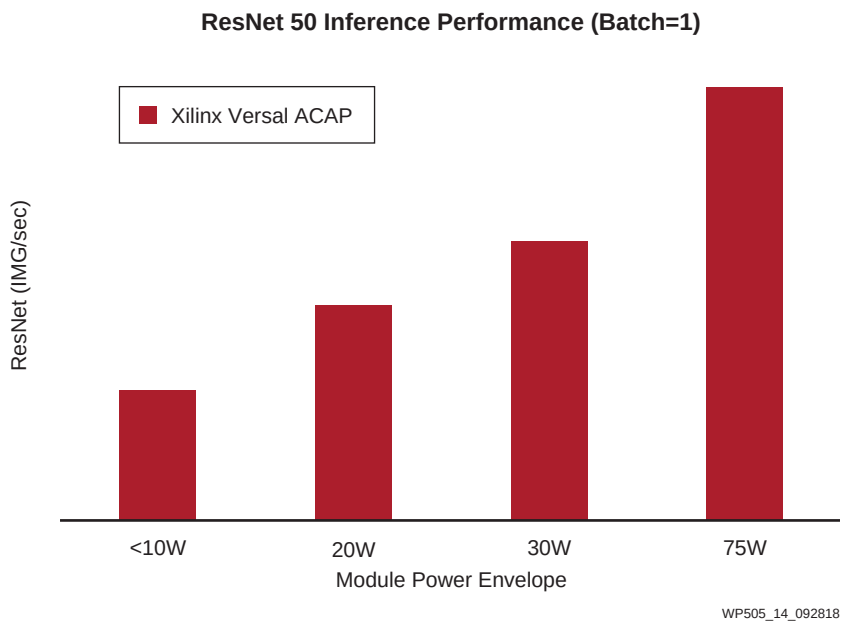

図 13: 低レイテンシの安全系に幅広く対応する Versal ポートフォリオ

現在の自動車に搭載される ADAS/AD システムは、高解像度カメラをより多く使用するようになっています。必要な演算性能はピクセル数に比例するため、HD カメラからの画像 (1080x1920) を処理するには、データセンターで一般的な画像 (224x224) を処理するよりもはるかに高い演算性能が要求されます。ザイリンクス Versal デバイスは演算効率が高く、今後のさらなる高解像度化にもスケーラブルに対応できるという強みがあります。

有線通信

現在、すべてのインターネット トラフィックは多くのザイリンクス FPGA を経由して転送されています。これまで長い間、FPGA はネットワーク事業者のニーズの変化に合わせてネットワーク ハードウェアを適応させるための「グルー ロジック」として使用されてきました。ザイリンクスは、業界をリードする最先端の 112G SerDes テクノロジにより、新しいプロトコルおよび難易度の高い光、銅線ケーブル、およびバックプレーン規格、さらには標準化前の PCI Express® Gen5 など最新の 58G PAM4 および 32G NRZ プロトコルを業界で初めて実装することに成功しています。豊富な IP ポートフォリオは標準インターフェイスの統合を容易にし、コストおよび消費電力を削減してきました。ザイリンクスは幅広い IP を提供しており、これらを自由に組み合わせることで、ハードウェア レベルでの差別化が可能です。

常に新しい機能を必要とするネットワーク事業者は、短時間でコーディングでき、フィールドでアップデートできる適応型ハードウェアを採用することで、従来の ASSP に比べ競争力を高めることができます。

ザイリンクスの Versal ACAP は次世代 600G 波長計画に準拠した IP をかつてないレベルで統合しており、イーサネットおよび OTN 規格の 10G、25G、50G、および 100G SerDes レートを完全にサポートしています。これには次のものが含まれます。

- IEEE Std 1588 タイムスタンプ (誤差 ±1ns)、eCPRI、および TSN をサポートした 10/25/40/50/100GE MAC/PCS/FEC
- 600G FlexE コア (10G チャネルへのチャネライゼーションが可能) および高密度 400GE/200GE/100GE MAC/PCS/FEC
- MACSEC と IPSEC、およびバルク AES-GCM 暗号化をサポートした 600G ワイヤ レート暗号化エンジン
- PAM4 レーン用 FEC を統合した 600G Interlaken
- DOCSIS ケーブル LDPC アプリケーション用 SD-FEC

こうした SerDes の大幅な強化によって、次のものが実現します。

- OTN およびエッジ ルーター アプリケーション向けのシングル チップ 1.0Tb/s+ ネットワーク ライン カード (商用 ASSP よりも高い柔軟性を同等の消費電力で達成)
- シングル チップ 2.4Tb/s+ 暗号化 DCI (Data Center Interconnect) ラックマウント型ネットワーク アプライアンス (RU ごとに複数のインスタンス (図 14 参照)。

- 加入者ごとにトンネルを暗号化し、ビジネス/一般家庭用向けに高度なサービスを提供する 400Gb/s+ CMTS (Cable Modem Termination Systems)

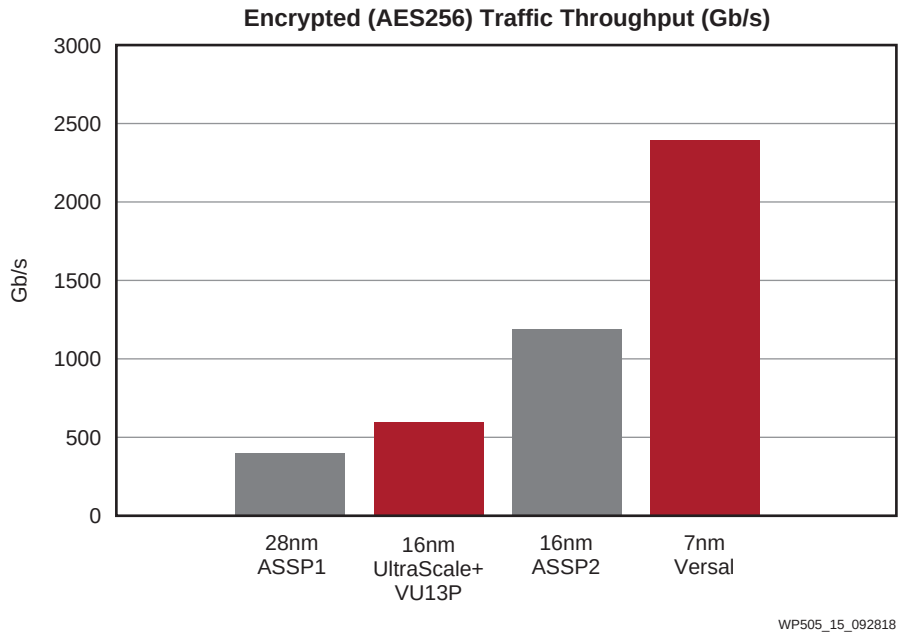

図 14: 有線通信: 暗号化済みデータセンター トラフィックのシングル チップ性能[1, 2]

1. Microsemi 社 DIGI-G4 OTN ASSP: https://www.microsemi.com/product-directory/multi-service-otn-processors/4227-pm5990-digi-g4
2. Microsemi 社 DIGI-G5 OTN ASSP: https://www.microsemi.com/product-directory/multi-service-otn-processors/5056-pm6010-digi-g5-otn-processo

適応性

プログラマブル ロジック テクノロジの最大の利点の 1 つは、ハードウェアをフィールドでアップグレードできることにあります。このため、プログラマブル ロジックは現在、4G 無線、光ネットワーク、自動運転車などの製品に広く採用されています。

ザイリンクス Versal ACAP は、C やフレームワーク レベル インターフェイスなどのより高い抽象度をサポートすると同時に、パーシャル リコンフィギュレーションを 8 倍高速化してカーネルの入れ替えにかかる時間を大幅に短縮するなど、フィールドでのアップグレード機能が改善されています。

適応型ハードウェア

以前から、FPGA はデザインをフィールドで変更できることを最大の価値としてきました。バグの修正、アルゴリズムの最適化、あるいはまったく新しい機能の追加ができるなど、プログラマブル ロジックはほかの半導体デバイスにはない柔軟性を備えています。

ザイリンクス Versal ACAP はこのコンセプトをさらに発展させ、コンフィギュレーション時間をほぼ 1 桁高速化することにより、パーシャル ビットストリームの動的な置き換えをミリ秒単位で完了できるようにするなど、ソフトウェアの機敏性を備えたハードウェアとなっています。

プログラマブルなメモリ階層

Versal ACAP において、適応型ハードウェアは ACAP アーキテクチャの新しい機能の効率を最適化する上で補完的な役割を果たします。

プログラマブル ロジックの最大の利点の 1 つは、メモリ階層をリコンフィギュレーションして特定の演算ワークロードに最適化できることにあります。たとえば、画像認識に特化したニューラル ネットワークだけを比べても、イメージごとの演算量とメモリ フットプリントはアルゴリズムによって大きく異なります。メモリ階層がプログラマブルであると、サポートするネットワークに合わせてプログラマブル ロジックを調整し、演算効率を最適化できます。

このため、Versal ACAP でベクター プロセッシングとプログラマブル ロジックを組み合わせてニューラル ネットワークをインプリメントすると、最先端の GPU でメモリ階層を固定してベクター プロセッシングを実装した場合に比べ、ほぼ 2 倍の演算効率を達成できます (図 15 参照)。

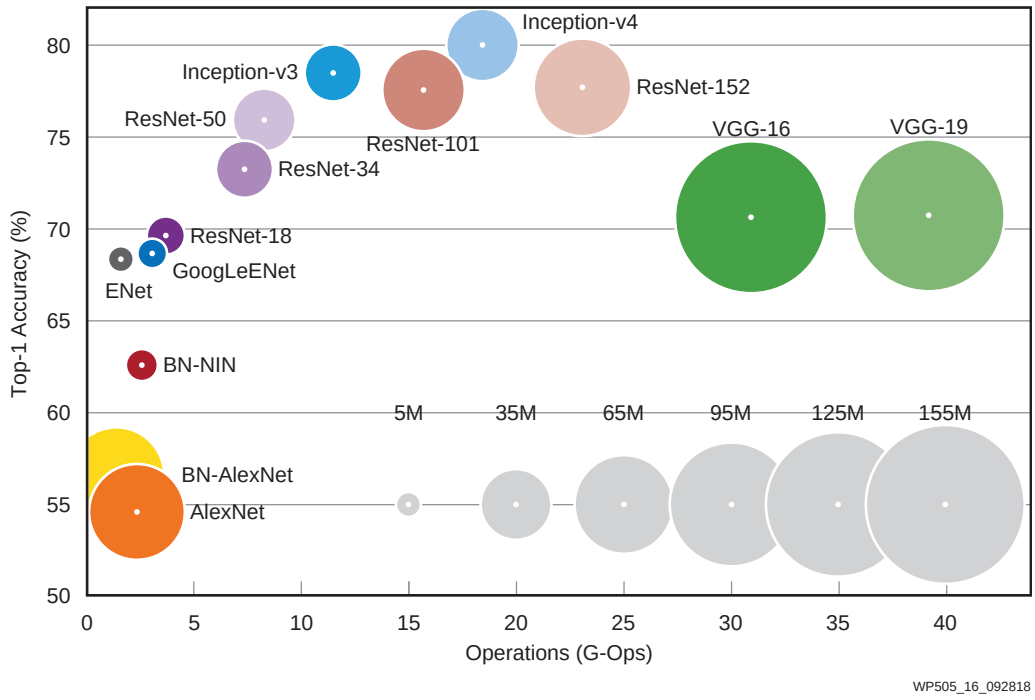

図 15: ニューラル ネットワークごとのメモリ フットプリントと演算量

ダイナミック リコンフィギュレーション

適応型エンジンは 1 ミリ秒未満でパーシャル リプログラミングが可能なため、コスト重視のリアルタイム アプリケーションでは、1 つのプログラマブル ハードウェアに複数のロジック ファンクションを多重化することでプログラマブル デバイスの利点を引き出すことができます。このため、データセンターではこれまで CPU で実行していた多くの機能を Versal ACAP デバイスで実行できるようになり、その機能の幅は GPU などのベクトル プロセッサをはるかにしのぎます(図 16 参照、[参照 4])。

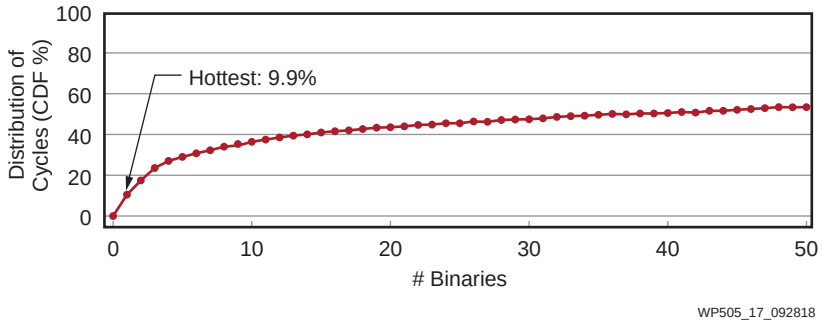

図 16: データセンターのワークロードは多様に分布しており、「キラー アプリケーション」は存在しない (Kanev)

まとめ

近年の技術的な課題により、ホモジニアスな CPU スカラー プロセッシング ソリューションであらゆる用途に対応するのは困難となっており、業界は別のアプローチを検討することを余儀なくされています。この問題は、ベクター プロセッシング (DSP、GPU) により部分的には解決できますが、メモリ帯域幅の利用効率が低いため、伝統的なスケーリングの課題に直面します。従来の FPGA ソリューションはメモリ階層をプログラムできますが、ハードウェア フローが導入の足かせとなっています。

ACAP (Adaptive Compute Acceleration Platform) はこれら 3 つのエレメントをすべて組み合わせ、フレームワークから C、そして RTL レベルのコーディングまで幅広い抽象度に対応した新しいツール フローを提供することによってこの問題を解決します。

ACAP アーキテクチャは、プログラマブル ロジックだけを使用した場合に比べ、はるかに強力な機能を実現します。プログラマブル ロジックとベクター プロセッシング エレメントをヘテロジニアス統合することにより、データセンター、無線ネットワーキング、先進運転支援、および有線通信などのアプリケーションで演算性能が破壊的に向上します。

データセンターでは、強力な AI 機械学習演算、先進のネットワーキング、および暗号化 IP を組み合わせてまったく新しいタイプの適応型演算アクセラレーション エンジンおよび SmartNIC が実現します。

既製の人工知能機械学習推論に高密度 DSP およびダイレクト RF サンプリング ADC/DAC を組み合わせると、内製の DSP ベース ASIC に比べて 5G 無線のスループットが 2 倍に向上し、先進運転支援システム (ADAS) アプリケーションにおける LIDAR、レーダー、および視覚センサーのセンサー フュージョンをシングル チップで実現できます。

ザイリンクス Versal ACAP デバイス ポートフォリオの詳細は、ザイリンクスのウェブサイトを参照してください。
https://japan.xilinx.com/products/silicon-devices/acap/versal.html.

参考資料

1. J. Hennessy, D. Patterson 『Computer Architecture: A Quantitative Approach』(第 6 版、2019)。
2. Nvidia: Nvidia AI Inference Platform: Giant Leaps in Performance and Efficiency for AI Services, from the Data Center to the Network's Edge (2018)。2018 年、nvidia.com から取得。
3. N. Jouppi, C. Young, N. Patil, et al.: In-Datacenter Performance Analysis of a Tensor Processing Unit™。International Symposium on Computer Architecture (ISCA 2017) にて発表。2018 年、arxiv.org から取得。
4. S. Kanev, J. Darago, K. Hazelwood, et al.: Profiling a warehouse-scale computer (2015)。2018 年、google.com から取得。

関連情報

1. H. Esmaeilzadeh, E. Blem, R. St. Amant, et al.: Dark Silicon and the End of Multicore Scaling。International Symposium on Computer Architecture (ISCA 2011) にて発表。2018 年、gatech.edu から取得。
2. M. Horowitz: Scaling Power and the Future of CMOS。第 20 回 International Conference on VLSI Design (VLSID 2005) にて発表。2018 年、semanticscholar.org から取得。
3. A. Putnam: Large-Scale Reconfigurable Computing in a Microsoft Datacenter。IEEE Hot Chips 26 Symposium (2014) にて発表。2018 年、microsoft.com から取得。

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
2020 年 9 月 29 日	1.1.1	誤字の修正。
2019 年 9 月 23 日	1.1	「5G 無線通信」を更新。
2018 年 10 月 2 日	1.0	初版

免責事項

本通知に基づいて貴殿または貴社 (本通知の被通知者が個人の場合には「貴殿」、法人その他の団体の場合には「貴社」。以下同じ) に開示される情報 (以下「本情報」といいます) は、ザイリンクスの製品を選択および使用することのためにのみ提供されます。適用される法律が許容する最大限の範囲で、(1) 本情報は「現状有姿」、およびすべて受領者の責任で (with all faults) という状態で提供され、ザイリンクスは、本通知をもって、明示、黙示、法定を問わず (商品性、非侵害、特定目的適合性の保証を含みますがこれらに限られません)、すべての保証および条件を負わない (否認する) ものとします。また、(2) ザイリンクスは、本情報 (貴殿または貴社による本情報の使用を含む) に関係し、起因し、関連する、いかなる種類・性質の損失または損害についても、責任を負わない (契約上、不法行為上 (過失の場合を含む)、その他のいかなる責任の法理によるかを問わない) ものとし、当該損失または損害には、直接、間接、特別、付随的、結果的な損失または損害 (第三者が起こした行為の結果被った、データ、利益、業務上の信用の損失、その他あらゆる種類の損失や損害を含みます) が含まれるものとし、それは、たとえ当該損害や損失が合理的に予見可能であったり、ザイリンクスがそれらの可能性について助言を受けていた場合であったとしても同様です。ザイリンクスは、本情報に含まれるいかなる誤りも訂正する義務を負わず、本情報または製品仕様のアップデートを貴殿または貴社に知らせる義務も負いません。事前の書面による同意のない限り、貴殿または貴社は本情報を再生産、変更、頒布、または公に展示してはなりません。一定の製品は、ザイリンクスの限定的保証の諸条件に従うこととなるので、https://japan.xilinx.com/legal.htm#tos で見られるザイリンクスの販売条件を参照してください。IP コアは、ザイリンクスが貴殿または貴社に付与したライセンスに含まれる保証と補助的条件に従うことになります。ザイリンクスの製品は、フェイルセーフとして、または、フェイルセーフの動作を要求するアプリケーションに使用するために、設計されたり意図されたりしていません。そのような重大なアプリケーションにザイリンクスの製品を使用する場合のリスクと責任は、貴殿または貴社が単独で負うものです。https://japan.xilinx.com/legal.htm#tos で見られるザイリンクスの販売条件を参照してください。

自動車用のアプリケーションの免責条項

オートモーティブ製品 (製品番号に「XA」が含まれる) は、ISO 26262 自動車用機能安全規格に従った安全コンセプトまたは余剰性の機能 (「セーフティ設計」) がない限り、エアバッグの展開における使用または車両の制御に影響するアプリケーション (「セーフティ アプリケーション」) における使用は保証されていません。顧客は、製品を組み込むすべてのシステムについて、その使用前または提供前に安全を目的として十分なテストを行うものとします。セーフティ設計なしにセーフティ アプリケーションで製品を使用するリスクはすべて顧客が負い、製品の責任の制限を規定する適用法令および規則にのみ従うものとします。

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。