ソリューション概要

データセンター
Myrtle スパース RNN アクセラレーション・ソリューション

データセンター向け 音声ワークロードのアクセラレーション

インテル® FPGA に実装された Myrtle MAU アクセラレーター・コアが、データセンター・インフラストラクチャーの最適化、コスト削減、パフォーマンスの向上を実現

インテル® FPGA プログラマブル・アクセラレーション・カード D5005

概 要

コールセンターの自動化や翻訳サービスなどの音声アプリケーションは、業界の違いを問わず、さまざまなビジネスのサービスを改善し、運用コストを削減していますが、一方でデータセンターのリソース需要を高める要因にもなっています。音声テキスト変換（Speech-to-text）およびその他の再帰型ニューラル・ネットワーク（RNN）ワークロードには、膨大な量の演算能力とメモリーが必要とされるため、データセンターのボトルネックを引き起こし、エネルギー需要の増大、コストの上昇などにつながります。

こういったワークロードをより効率的に処理するために、GPU または FPGA ベースのアクセラレーターをクラウドやオンプレミスのデータセンターに実装します。GPU は早い時期からグラフィックス・アクセラレーション向けのソリューションを提供しましたが、インテル® FPGA PAC D5005（インテル® FPGA プログラマブル・アクセラレーション・カード D5005）といった新しい FPGA ベースのアクセラレーターは、ソフトウェア実装のみの場合と比べて 165 倍の同時音声チャネルに対応するなど、音声アプリケーションで優れた効果をもたらします。[1]

マシンラーニングの高速化において業界をリードする専門企業の 1 つである Myrtle は、NVIDIA* Tesla* V100 GPU に匹敵するコンピューティング・パフォーマンスを達成すると同時に、消費電力を 6 分の 1 に削減するインテル® FPGA PAC D5005 搭載向けの RNN ソリューションを開発しました。このソリューションは、レイテンシーや応答の遅延を、顧客対面型のインタラクティブな音声サービスなどのリアルタイム音声アプリケーションに不可欠なレベル（29 分の 1）にまで低減します。[1]

Myrtle のソリューションは、音声合成、音声の文字起こし、機械翻訳、またはその他の RNN ワークロードに依存している企業や実装を計画している企業に対し、データセンター・インフラストラクチャーの総保有コストを削減しつつ、高まる需要ピークに対応し、より高いレベルで自動化を導入でき、新たな収益源となるサービスを迅速かつコスト効率よく追加する能力を生み出すという、スマートな方法を提供します。

このソリューション概要では音声に焦点を当てていますが、ゲノム解析や金融アプリケーションなど、その他の RNN ワークロードにも Myrtle アクセラレーション・ソリューションはその効果を発揮します。

myrtle.ai

ビジネス上の課題：高まるデータセンター需要への対応

音声認識、自然言語処理、音声テキスト変換といった高度な音声関連サービスに対する需要は、消費者ユーザーと企業ユーザーの両方で高まっています。こういったRNNワークロードは、データセンターにおけるディープラーニング推論ワークロード全体の29%を占め[2]、データセンターに対して極めて高い要件を課しています。

2013年、Googleは、1日当たりわずか3分程度の音声検索を利用するユーザーの演算要件を満たすために、データセンターの数を2倍にする必要があると予測しました。[2] データセンターに関しては、Facebookの研究者も、「将来の需要の大部分は、ディープラーニング推論関連のワークロードから発生する」と予測しています。[3]

ディープラーニング推論モデルに対する需要が増大すると同時に、さまざまなモデルがより洗練され、要件は厳しくなり、より高い演算能力とメモリーが必要とされるようになります。RNNモデルは、数百万から数十億のパラメーターを含めるように拡張できるため、演算コストも電力コストも急速に増大します。さらにイノベーションにより、もっと大規模な新しいRNNモデルの継続的な導入が後押しされ、最新モデルに最適化されたハードウェアでさえ、すぐに不十分で非効率的なものとなってしまいます。

こうした問題に対し、企業はデータセンターをスケールアウトしてインフラストラクチャーやコンピューティング・ハードウェアを追加することも可能ですが、それには相当なコストがかかり、電力消費も許容範囲を超えてしまう可能性があります。別のアプローチとして、企業が独自の特定用途向け集積回路（ASIC）を構築する方法が考えられますが、このオプションには18カ月以上の時間と数百万米ドルを超える費用がかかり、さらなる進化は期待できません。製品が導入される頃には、マシンラーニング・アプリケーションはさらに進化して、システム・アーキテクチャーも変更不可能な回路も最適とは言えなくなり、時代遅れとなってしまうのは目に見えています。

こういった課題とそれに関わる問題こそ、再構成可能で、消費電力当たり性能の高い、低レイテンシーかつ低コストなRNNワークロードとその他のマシンラーニング・ワークロードに対応できるテクノロジー・ソリューションを多くの企業が求めている理由です。

ソリューション：インテル® FPGA搭載向けの高度に最適化されたアクセラレーター

Myrtleはディープラーニング分野での長年にわたる経験を活かして、独自のアクセラレーター・コアであるMAUを開発しました。これは、今日最も演算負荷の高いRNNワークロードの一部を処理できるように、また今後生じる要求に応じて拡張できるように最適化されています。

各MAUアクセラレーター・コアは、非構造のスパース行列演算を実行するハイパフォーマンス・コンピューティングに最適化されています。コアには、マシンラーニングの音声アルゴリズムに必要なRNN演算、双方向LSTMの点方向オペレーション、非線形性をサポートする機能が備わっています。複数のコアを組み合わせることで多様な行列乗算が可能となり、異なるニューラル・ネットワーク記述を対象としてアクセラレーター・グリッドを柔軟に構成できます。

Myrtleでは、特定のRNNモデルや音声テキスト変換といったワークロードを分析した後に、MAUアクセラレーター・コアのネットワークを構成して、インテル® FPGA PAC D5005ボードの機能と特長を最大限に活用します。

インテル® Xeon® プロセッサーを基盤とするサーバーに実装されたインテル® FPGA PAC D5005は、異なるコンピューティング・エンジン（CPUおよびFPGA）を備えたヘテロジニアスなコンピューティング・プラットフォームを形成します。これによって、ワークロードをパーティションに分割し最適化できるため、負荷の高いRNN演算処理をインテル® FPGA PAC D5005上で実行することで、CPUは自身の最適なワークロードに集中できるようになります。

テスト結果から、Myrtleとインテル® FPGA PAC D5005を組み合わせたソリューションは、NVIDIA* Tesla* V100 GPUのパフォーマンスに匹敵する、4,000を超えるリアルタイム音声チャネルに1チップで対応できると示されています。28のリアルタイム音声チャネルに対応するインテル® Xeon® プロセッサーのみを実装[4]した場合と比較すると、インテル® FPGA PAC D5005でのRNN処理はリアルタイム・チャネル数が165倍になり、同一のサーバー・ソケット・インフラストラクチャーでの処理能力が飛躍的に向上します。

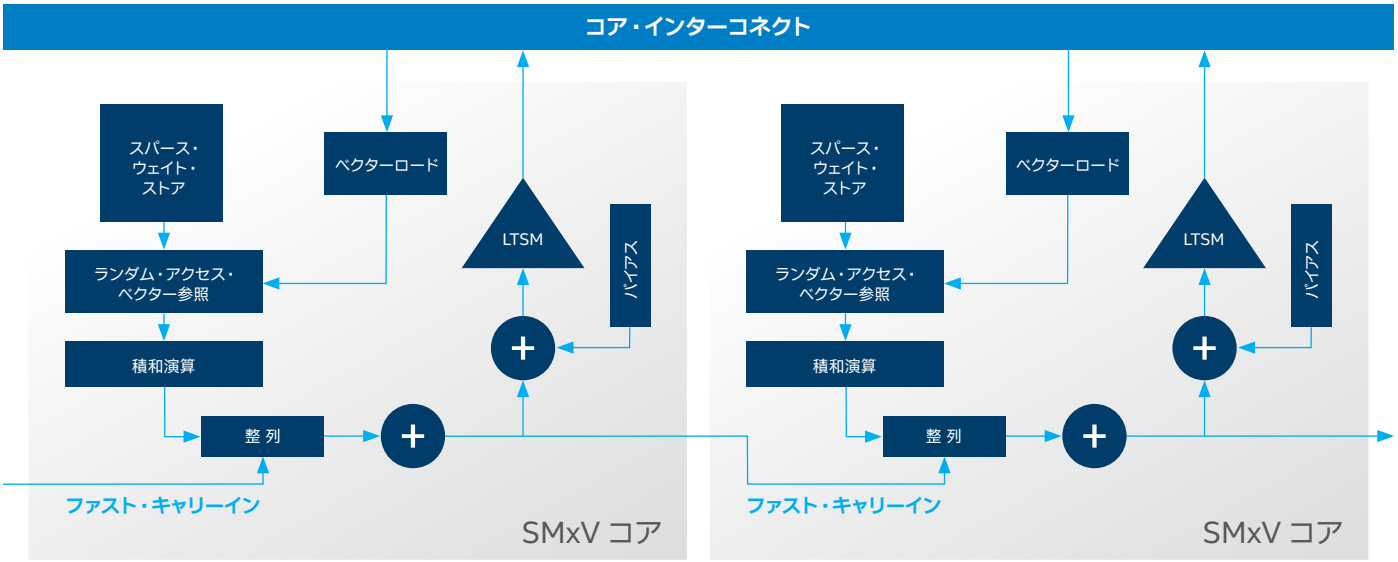


図1. インテル® FPGA PAC D5005向けに最適化済みの、同ボード上に配置されたMyrtleのMAUアクセラレーター・コア。

固定デバイスとは異なり、インテル® FPGA PAC D5005 ボードは動的な再構成ができるため、短時間でアップデートして最新の RNN モデルや数値の実行が可能になります。さらに、ASIC とは異なり、FPGA は低負荷時にはほかのタスクを実行するように再構成したり、変化するワークロードのニーズに合わせて新しいアルゴリズムで更新することも可能です。

インテル® FPGA PAC D5005 は、ドライバー、アプリケーション・プログラミング・インターフェイス、FPGA インターフェイス・マネージャーを含む共通の開発者向けインターフェイスを提供する、インテル® アクセラレーション・スタック（インテル® Xeon® CPU & FPGA 対応）を備えています。インテル® アクセラレーション・スタックは、業界最先端のオペレーティング・システムや仮想化およびオーケストレーション・ソフトウェアと連動し、ソフトウェア開発者向けの共通インターフェイスを提供し、迅速な収益化、管理の簡素化、成長するアクセラレーション・ワークロード・エコシステムへのアクセスを実現します。

Myrtle は、さまざまなパフォーマンス要件やコスト要件を満たすうえで、精度と拡張性を犠牲にすることなく、RNN モデルを圧縮する専門知識を備えており、その技術はデータセンター・アプリケーションだけでなく、エッジのリアルタイム・アプリケーションにも適用できます。これにより、モバイルデバイスや自動車、その他の新しいアプリケーションで急増している音声サービスの分野においても、多くの企業に成長の機会を提供しています。

ソリューション・コンポーネント

- Myrtle MAU アクセラレーター・コア、RNN 処理機能を備えた高性能のスパース線形代数アクセラレーター
- インテル® FPGA PAC D5005、データセンター向けの高性能 PCI Express* (PCIe*) ベースの FPGA アクセラレーション・カード
- 開発者の時間を節約するための共通インターフェイス、ドライバー、API、および FPGA インターフェイス・マネージャーを提供するインテル® アクセラレーション・スタック（インテル® Xeon® CPU & FPGA 対応）
- インテル® Xeon® プロセッサー搭載サーバー

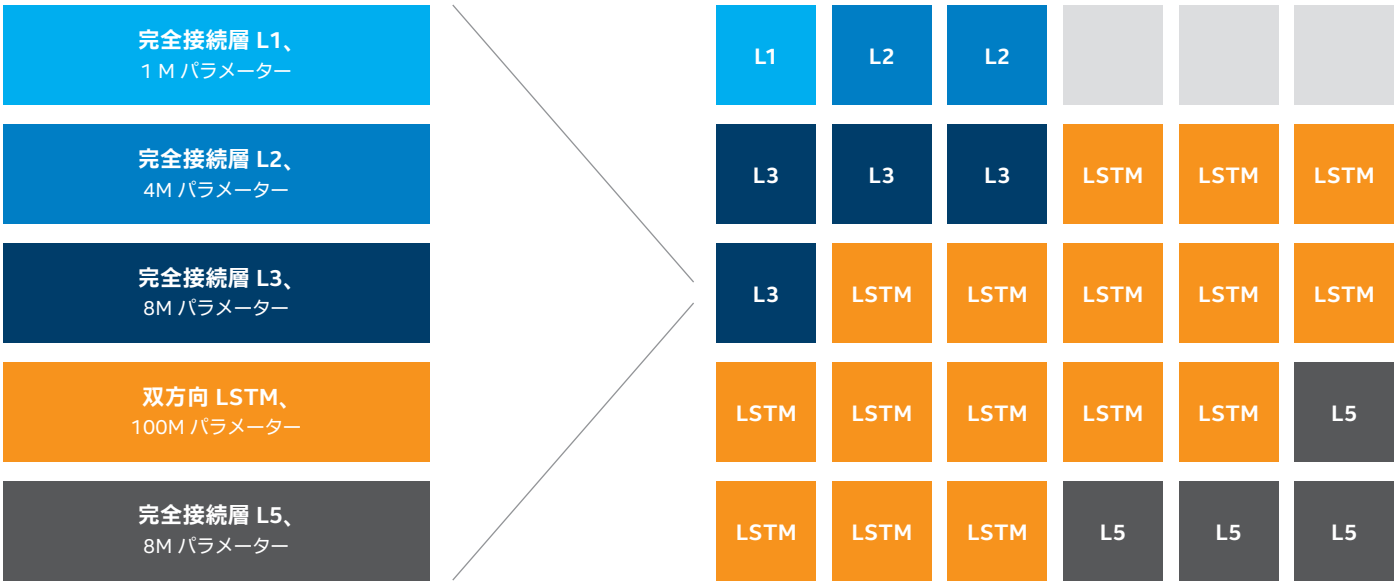


図 2. DeepSpeech 音声読み上げワークロード対応のインテル® FPGA PAC D5005 ボードに組み込まれた Myrtle の拡張可能な MAU アクセラレーター・コア。

導入事例 : FPGA ベースの音声テキスト変換アクセラレーション

Myrtle はインテルのネットワーク & カスタムロジック事業本部（NCLG）と連携して、高性能なインテル® FPGA PAC D5005 ボード上で動作する RNN 最適化音声認識アルゴリズムを開発しました。このソリューションは、実稼動システムに展開された RNN の代表である Baidu のオープンソースの音声テキスト変換モデル DeepSpeech を使ってテスト実行されています。このソリューションと、インテル® Xeon® プロセッサーのみを実装したソリューション、NVIDIA* Tesla* V100 GPU をベースとした標準 GPU 処理モデルが比較されました。

DeepSpeech ソリューションでは、Myrtle は 2 つの最適化手法（スパース性と量子化）を活用して、スループット性能の最適化、待ち時間の短縮、消費電力の効率化を実現しています。Myrtle の社内マシンラーニング・チームは、機能を DeepSpeech モデルに学習させ、モデルの精度をほとんど損なうことなく、FPGA での演算を効率化しました。

スパース性

スパース性は、一部の値を明示的にゼロに設定することで有効パラメーターの総数を減らしますが、この手法は、GPU や CPU では効率的に活用できません。FPGA ファブリック上に構築された適切なハードウェア・アーキテクチャーでスパース性を使用することにより、システムの有効な演算密度を大幅に高め、メモリー要件を軽減しています。これはパラメーターを保存する必要がないためです。FPGA プラットフォームの場合、ウェイトを完全にオンチップ RAM に保存できることになり、極めて高い帯域幅でモデル・パラメーターにアクセス可能な、電力効率の高いソリューションが実現されます。

このテストでは、Myrtle は精度をほとんど損なわずに MAU アクセラレーター・コアを非常に高いレベルのスパース性に調整し、ユーザー体験を低下させることなくスパース性のパフォーマンスを向上できることが示されています。

量子化

量子化は、浮動小数型の数値を幅の狭い整数型に変換することで、推論中に使用されるビット数を減らします。このテストでは、Myrtle は 8 ビット整数型に量子化できることを示しました。これにより、RNN モデルのサイズを 4 分の 1 に縮小し、算術密度をすぐに 4 倍高め、データ帯域幅の要件を 4 分の 1 に軽減しました。繰り返しになりますが、この手法は、精度をほとんど損なうことなく使用されています。

マシンラーニングのワークロードをアルゴリズム・アクセラレーターを協調設計するうえでの課題として扱うことで、Myrtle はシステム・パフォーマンスの実質的な向上を実証し、インテル® Xeon® プロセッサー搭載サーバーからインテル® FPGA PAC D5005 ボードへの大幅なアルゴリズム・オフロードを可能にしました。

	Myrtle MAU アクセラレーター	GPU	Myrtle MAU アクセラレーターのメリット
プラットフォーム	インテル® FPGA PAC D5005[a]	NVIDIA* Tesla* V100[b]	
周波数(MHz)	250	1530	
スパース性(%)	96	0	
量子化	8 ビット整数	16 ビット浮動小数	
バッチサイズ	1	256	
実効スループット (TOPS)	54	53.37[c]	同等
電力消費(W)	34.9	216	6 分の 1 の消費電力
消費電力当たり性能(実効 GOPS/W)	1547	247	6 倍の GOPS/W
1 秒のオーディオ入力当たりのレイテンシー(ms)	0.343	126	365 分の 1 のレイテンシー

図 3. テスト結果 : インテル® Xeon® プロセッサーとインテル® FPGA PAC D5005 で実行される Myrtle MAU アクセラレーターを組み合わせたソリューションは、NVIDIA* Tesla* V100 と比較して同等のスループットを処理しながら、電力を 6 分の 1、待ち時間を 365 分の 1 に短縮。

GPU アクセラレーション・モデルと比較した場合、FPGA はパフォーマンスとレイテンシー目標を同時に達成し、消費電力当たり性能が向上しています。GPU アクセラレーションでは、パフォーマンスの向上によってレイテンシー要件が犠牲となり、リアルタイム・アプリケーションで達成可能なパフォーマンスが低下します。システムが低レイテンシーに最適化されている場合の CPU と GPU の比較については、図 4 を参照してください。

	Myrtle MAU アクセラレーター	GPU	Myrtle MAU アクセラレーターのメリット
プラットフォーム	インテル® FPGA PAC D5005[a]	NVIDIA* Tesla* V100[b]	
周波数(MHz)	250	1530	
スパース性(%)	96	0	
量子化	8 ビット整数	16 ビット浮動小数	
バッチサイズ	1	1	
実効スループット (TOPS)	54	1.12	48 倍のスループット
電力消費(W)	34.9	191.8	5 分の 1 の消費電力
消費電力当たり性能(実効 GOPS/W)	1547	5.84	265 倍の GOPS/W
1 秒のオーディオ入力当たりのレイテンシー(ms)	0.343	10.1	29 分の 1 のレイテンシー

図 4. テスト結果 : GPU システムが低レイテンシーに最適化されている場合でも、インテル® FPGA PAC D5005 で実行されている MAU は、NVIDIA* Tesla* V100 と比較して 29 倍高速、スループットは 48 倍向上。

a. インテル® FPGA PAC D5005 カードの測定値 : インテル® Xeon® プロセッサー、16GB RAM (2800 MHz) x4、1TB M.2 PCIe* SSD、PRIME Z270-P* ボード、650 W PSU、Ubuntu* の構成で測定。
b. NVIDIA* Tesla* V100 の測定値 : Google* クラウド上の NVIDIA* Tesla* V100 インスタンスとインテル® Xeon® プロセッサー (2.30 GHz) x12、16GB RAM x4 の構成で測定。
c. 200 ms の短い入力時間で測定された 53.37 TOPS の最大スループット。1 秒の入力時間でレイテンシーを測定する場合、最大スループットは 23 TOPS に低下。

ソリューションの価値：精度を犠牲にしない高パフォーマンスと低レイテンシー

インテル® FPGA PAC D5005 ボードに Myrtle MAU アクセラレーター・コアを実装することで、音声サービスを提供する企業にとって、次のように重要なメリットが得られると、テストにより示されています。

- **効率的なパフォーマンス：**このソリューションは、NVIDIA* Tesla* V100 GPU と比較して、同等のスループットと、6 倍の消費電力当たり性能を実現し、結果的に電力消費と運用コストの削減につながります。[1]
- **低レイテンシー：**このソリューションでは、NVIDIA* Tesla* V100 GPU に比べてレイテンシーを 29 分の 1 に低減します。[1] 低レイテンシー処理では、GPU の実効スループットは劇的に低下しましたが、MAU アクセラレーターは高性能と低レイテンシーを同時に実現しました。これにより、エンドユーザーは目立った遅延を意識することなく、マルチチャネルのインタラクティブな音声サービスを利用できるようになります。また、実施可能なコスト範囲内で、サービスの大規模な展開も可能です。
- **精 度：**精度低下は 0.23% 未満に収まり、非常に高いパフォーマンス・レベルが達成されました。これにより、エンドユーザー体験を損なうことなく、プラットフォームのパフォーマンスを飛躍的に向上できます。
- **拡張性：**このソリューションは、ソフトウェアのみの実装と比べて、165 倍の RNN 処理に対応できます。[1] これにより、音声チャネル当たりの運用コストが大幅に削減され、桁違いに高い収益性につながります。さらに、サーバー・インフラストラクチャーの要件が大幅に軽減されるため、これまでは物理的なスペースとインフラストラクチャーの制限によって現実的ではなかったマルチチャネル音声ソリューションのオンプレミス配置が可能になります。
- **柔軟性：**FPGA は動的な再構成が可能であるため、今後のニーズに応じてソリューションを継続的に更新および最適化していくことができます。設置済みハードウェアの耐用年数の延長にもつながり、急速に進化するマシンラーニングの分野では特に重要なメリットです。

インテル® FPGA PAC D5005 ボードを既存のインテル® Xeon® プロセッサー搭載サーバー構成に追加すると、サーバーの処理能力は全体で 10 ~ 80 倍向上し、インテル® Xeon® プロセッサーの負荷を軽減し別の機能に割り当てることができるようになります。[5] GPU ベースの実装に比べて、インテル® FPGA PAC D5005 の実装は、運用コストを 45% 削減。[6] 低レイテンシー重視のアプリケーションでは、インテル® FPGA PAC D5005 によりスループットが 48 倍になり、GPU ベースの実装で同等のパフォーマンスを実現するには何倍もの GPU とサーバーが必要になることを考えると、設備投資の大幅なコスト削減が期待できます。

図 5. サーバー・インフラストラクチャーの要件が軽減されることにより、TCO の大幅な削減と、これまで現実的ではなかったソリューションのオンプレミス配置を実現。

まとめ

レイテンシー、電力、コストの厳しい要件を満たすために、音声テキスト変換やその他の RNN ワークロードには、ハードウェアとソフトウェアが密に結合したソリューションが求められます。従来の CPU ベースまたは GPU ベースのソリューションと比べると、インテル® FPGA PAC D5005 ボード上で実行される Myrtle の MAU アクセラレーター・コアは、比較的少ないサーバーで音声アプリケーションを運用でき、インフラストラクチャー・コスト、運用コストを削減すると同時に、データセンターにおける電力と設置フロアのさらに厳しい制約にも対応することが可能です。また、Myrtle の拡張可能なソリューションはサーバー容量を解放するため、多くの企業が、高まる需要ピークを処理し、より高いレベルの自動化を導入して、新たな収益源となるサービスを迅速かつ高いコスト効率で展開できるようになります。

Myrtle とインテルにより、数々の企業が、コストを削減し、お客様に提供する音声サービスの品質を向上し、その範囲を拡大しながら、新たに登場する音声アプリケーション向けマシンラーニングの最新技術がもたらすメリットを取り入れる柔軟性も維持できるようになっています。

Myrtle について

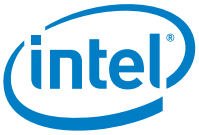
Myrtle は、英国ケンブリッジに拠点を置くテクノロジー企業です。FPGA 上で実行されるディープラーニング推論を高速化する強力な最先端のソフトウェアを開発しています。データセンター向け音声アプリケーションに最適化された実装の開発において、業界でも有数のエキスパート企業です。同社の音声推論コードベースおよびモデルは、業界主導の機械学習ベンチマークへの取り組みである MLPerf コンソーシアムで、エッジおよびデータセンターの新しいハードウェアのベンチマークとして採用されています。

関連情報

Myrtle の MAU アクセラレーターは、インテル® PAC に展開できます。ソリューションおよびビジネス上のメリットの詳細については、Myrtle（stratix_eval@myrtle.ai）にお問い合わせください。

インテル® FPGA ソリューションの詳細については、こちらを参照してください。

インテル® プログラマブル・アクセラレーション・カードとアクセラレーション・スタックの詳細については、こちらを参照してください。



インテル ® テクノロジーの機能と利点はシステム構成によって異なり、対応するハードウェアやソフトウェア、またはサービスの有効化が必要となる場合があります。実際の性能はシステム構成によって異なります。絶対的なセキュリティーを提供できるコンピューター・システムまたはデバイスはありません。詳細については、各システムメーカーまたは販売店にお問い合わせいただくか、http://www.intel.co.jp/iot/ を参照してください。

本ドキュメントで参照される性能結果は、2019 年第 2 四半期、インテル ® Stratix® 10 FPGA 上での DeepSpeech の実行に基づきます。これらのテストは Myrtle によって実施されました。

[1] M. Ashby、C. Baaij、P. Baldwin、M. Bastiaan、O. Bunting、A. Cairncross ほか、「Exploiting unstructured sparsity on next-generation datacenter hardware.」 https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/myrtlle-unstructured-sparsity-wp.pdf（英語）

[2] N. P. Jouppi、C. Young、N. Patil、D. Patterson、G. Agrawal、R. Bajwa、S. Bates、S. Bhatia、N. Boden、A. Borchers ほか、「In-Datacenter Performance Analysis of a Tensor Processing Unit」、2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture（ISCA）。IEEE、2017 年、pp. 1–12。

[3] J. Park、M. Naumov、P. Basu、S. Deng、A. Kalaiah、D. Khudia、J. Law、P. Malani、A. Malevich、S. Nadathur ほか、「Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications」、arXiv preprint arXiv: 1811.09886、2018 年。

[4] DeepSpeech ワークロードは、デュアルソケットのインテル ® Xeon® 6140M Gold プロセッサー（2.30 GHz）、リアルタイム音声チャネル x28、有効 0.327 TOPS で実行。インテル ® FPGA PAC D5005（54.0 TOPS）は 165 倍のスループットを出力。

[5] DeepSpeech 1 については、Myrtle はインテル ® FPGA PAC D5005 上で実行するニューラル・ネットワークの 98.7% を高速化。CPU 上で実行した残りの 1.2% により、CPU の演算処理が 83 分の 1 にオフロードされたことが示されています。

[6] サーバー + インテル ® FPGA PAC D5005 構成の推定コストをサーバーのみ構成のコストの 1.5 倍として、サーバー・インフラストラクチャーを 10 分の 1 に縮小した場合のコストに基づき、資本コストを当初費用の 15% と算出。NVIDIA* Tesla* V100 またはインテル ® FPGA PAC D5005 搭載の同等数のサーバーに基づき、運用コストは消費電力に比例、サーバーのみのシステムの消費電力を 200W と想定し算出。サーバー +GPU の構成では 416W、サーバー + インテル ® FPGA PAC D5005 の構成では 235 W。

341254-001JA