最小限のコード変更で CPU 上で超高速な Python* データサイエンスと AI パフォーマンスを実現

この記事は、Intel Tech.Decoded で公開されている「Deliver Blazing-fast Python Data Science and AI Performance on CPUs—with Minimal Code Changes」の日本語参考訳です。

IDC の業界アナリストは、2024 年までに世界のマシンラーニング（ML）市場は総額 306 億米ドルに成長し、年平均成長率は 43% に達すると予測しています [1]。また、同年には 143 ゼタバイトのデータが世界中で作成、取得、コピー、消費されると予測しています [1]。

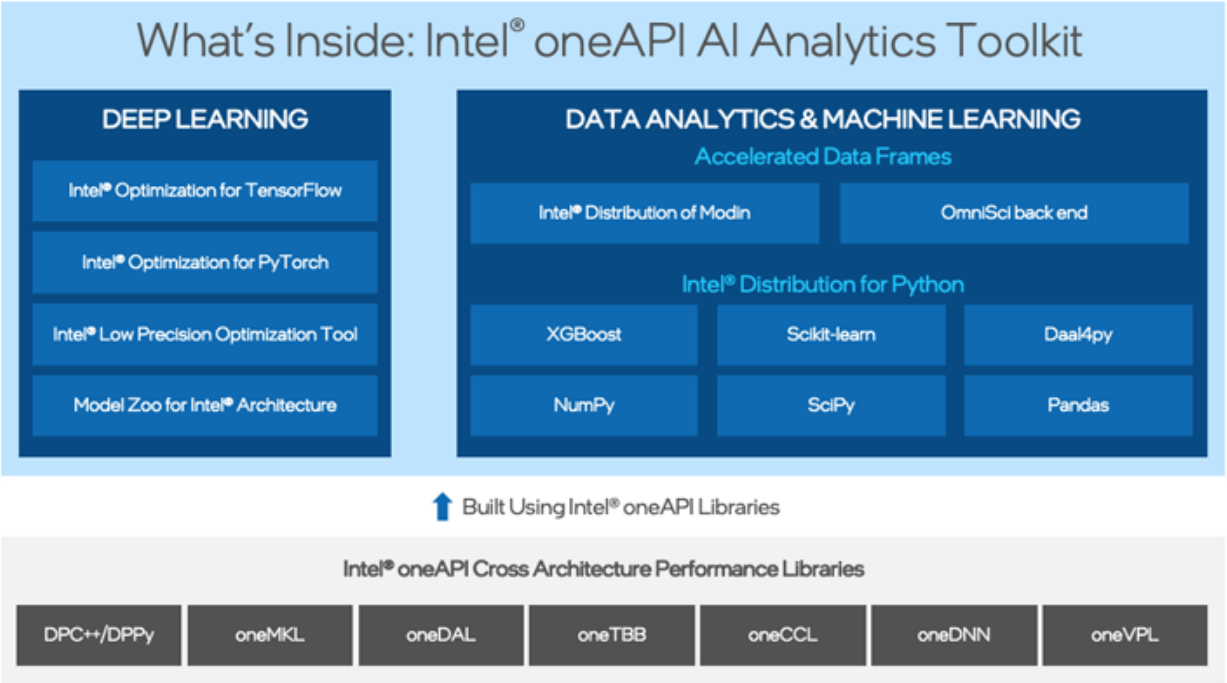
エンドツーエンドの AI ワークフロー全体でこのデータを管理するハードウェア環境は多様化しており、特定のユースケースを管理するためユニークなアクセラレーターが市場に登場しています。最近の Evans Data Corporation の報告では、開発者の 40% が複数の種類のプロセッサー、プロセッサー・コア、コプロセッサーを使用するヘテロジニアス・システムをターゲットにしていると指摘しています [2]。

AI、ML、およびディープラーニング（DL）開発者にとってこれは柔軟な配布方法であり、ハイパフォーマンスなアプリケーションを開発することが求められる、魅力的なビジネスチャンスです。1 つのハードウェア・アーキテクチャーですべてをまかなうことは困難です。

このチャンスを逃さないためには、さまざまなアーキテクチャーで最高の Python* パフォーマンスを実現することが重要です。Python* は、強力かつスケーラブルで使いやすい言語ですが、超高速なパフォーマンス向けには設計されていません。多くの開発者は、TensorFlow*、PyTorch*、scikit-learn* などのフレームワークやライブラリー向けのベンダー固有の最適化を利用して、選択したハードウェアで必要な速度を実現していますが、新しいハードウェアが登場したり、市場が発展して新たな種類のハードウェアへの展開が必要になったときに、どのようにして同様の結果を得ることができるのか疑問に思っています。

そこで、インテル® oneAPI AI アナリティクス・ツールキット（英語）の出番です。

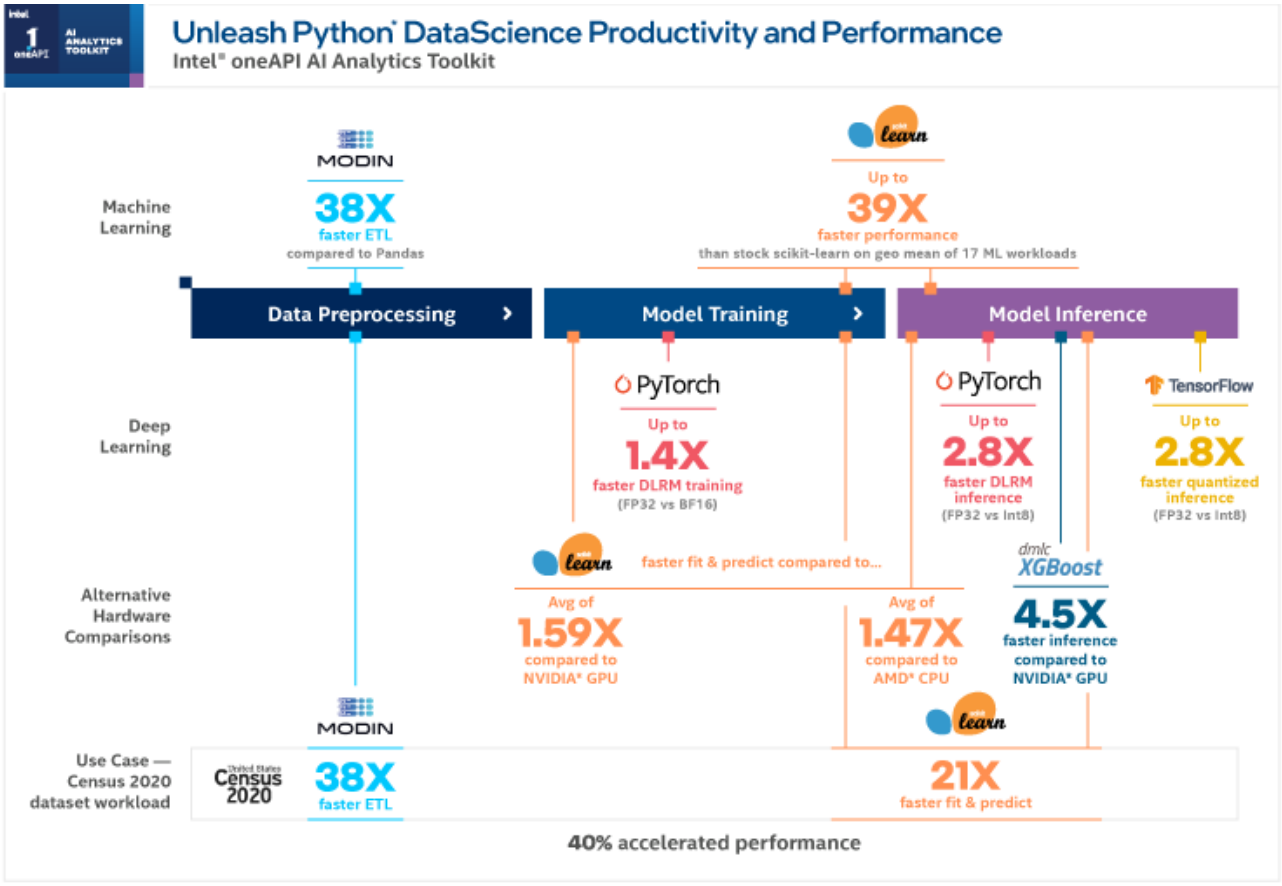
インテル® oneAPI クロスアーキテクチャー・パフォーマンス・ライブラリーをベースに構築されたこのドメイン固有のツールキットは、Python* エコシステムでエンドツーエンドのデータサイエンスと AI ワークフローを高速化します。現在一般的に使用されているインテル・ハードウェア・アーキテクチャーだけでなく、将来的に業界標準となるものでも、前処理から ML/DL のトレーニングと推論に至るまで、最大限のパフォーマンスを引き出します。使い慣れた Python* フレームワークでドロップインのアクセラレーションを提供することで、開発コストを最小限に抑えることが目的です。これにより、データ・サイエンティストや開発者は、独自のプログラム環境に限定されたり、新しいハードウェア・プラットフォームに対応するたびに新しいソフトウェア API を採用することなく、自信を持って開発に取り組むことができます。



インテル® oneAPI AI アナリティクス・ツールキットでドロップインのアクセラレーションを実現

インテル® oneAPI AI アナリティクス・ツールキットを利用することで、インテル・アーキテクチャーに移行した後も使い慣れたフレームワークや Python* ライブラリーを使い続けることができます。最小限のコード変更、あるいはコードを全く変更しなくても、ドロップインのアクセラレーションを利用できます。

このツールキットを利用することで、ほとんど手間をかけずに CPU 上でワークフローを高速化できることを示すため、多くの一般的なライブラリーを対象とした 5 つのベンチマークを実施しました。



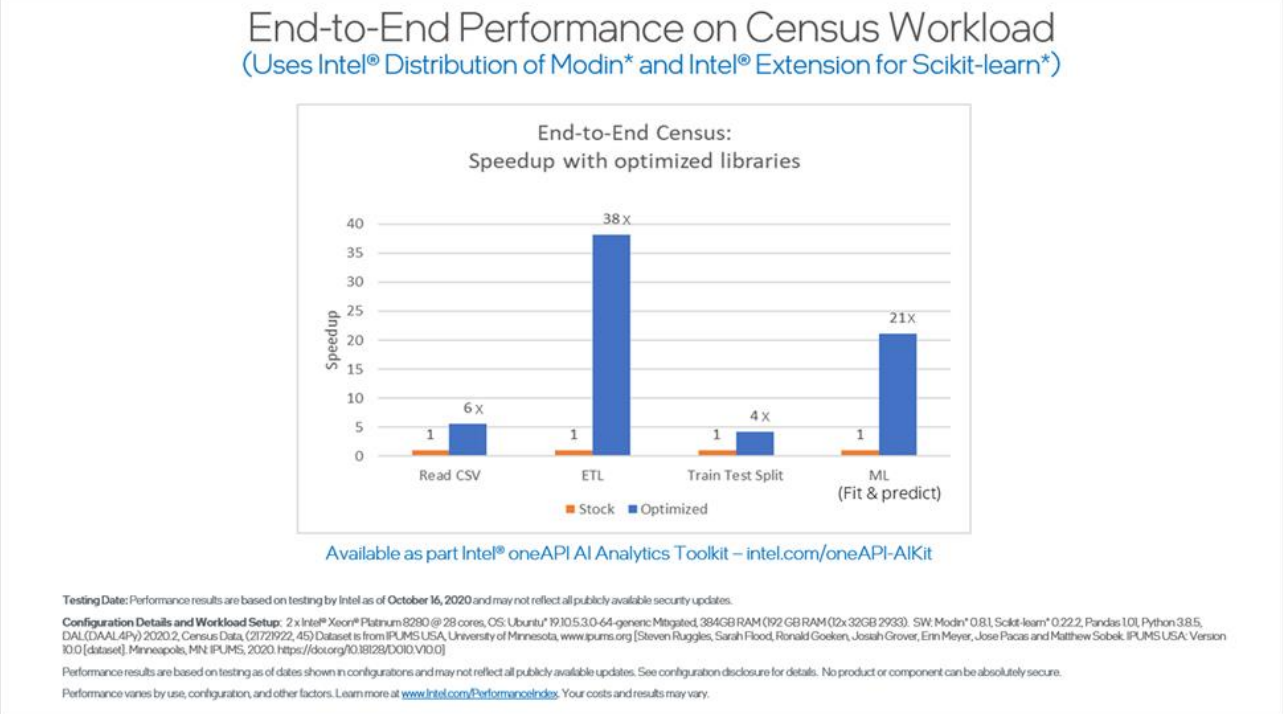
See all benchmarks configurations - https://techdecoded.intel.io/resources/deliver-blazing-fast-python-data-science-and-ai-performance-on-cpus-with-minimal-code-changes
Each performance claim and configuration data is available in the body of the article listed under sections 1, 2, 3, 4 and 5.
Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly availableupdates. See backup for configuration details. No product or component can be absolutely secure.
Intel technologies may require enabled hardware, software or service activation.

データ前処理、トレーニングのモデル化、推論のモデル化を含む、AI パイプライン全体の例を見てみましょう。

ベンチマーク 1: 国勢調査セットのエンドツーエンドのパフォーマンス

業界標準の国勢調査データセットのベンチマークでは、CPU アーキテクチャー上でインテル® oneAPI AI アナリティクス・ツールキットを使用して実行したところ、優れた結果が得られました。この結果を得るため、IPUMS.org からの 50 年分の国勢調査データを使用して、学歴を基に所得を予測するモデルをトレーニングしました。

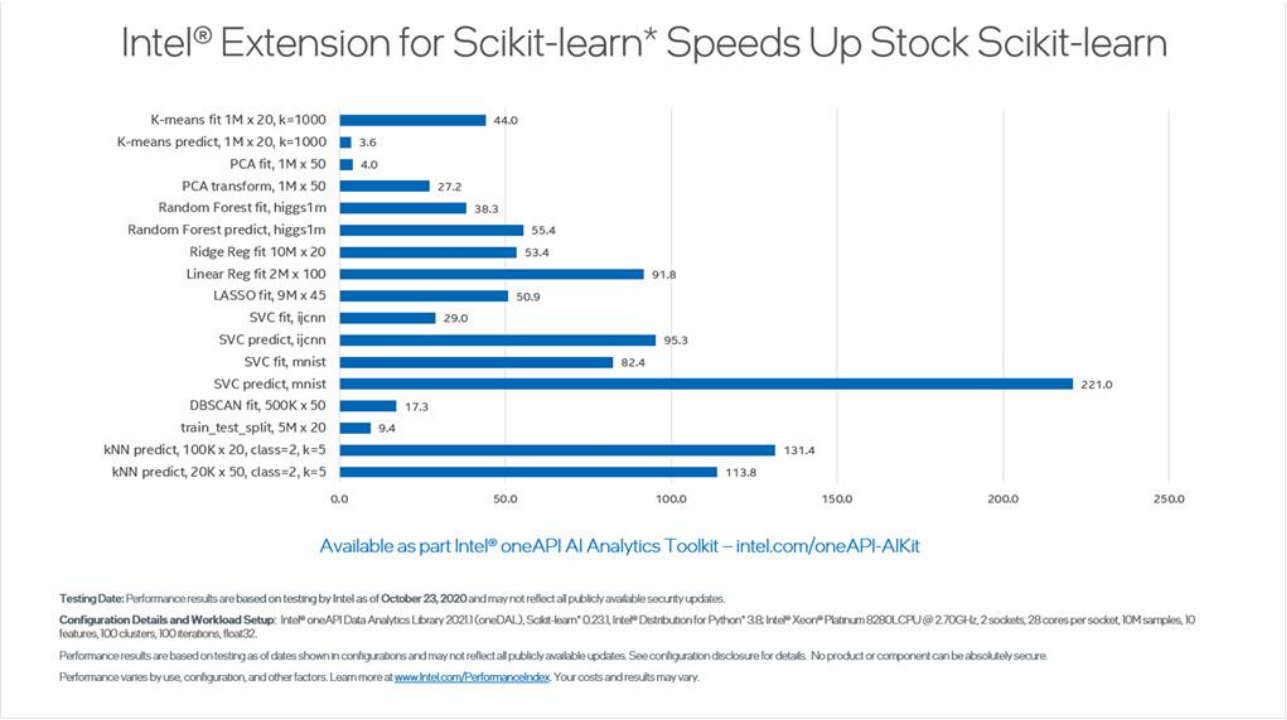
インテル® oneAPI AI アナリティクス・ツールキットは、標準ライブラリーと比較すると、モデルの実行速度を大幅に向上します。以下のベンチマークでは、データサイエンスのパイプライン全体で大幅な向上が見られ、ETL は 38 倍、リッジ回帰によるマシンラーニングの予測と適合は 21 倍に向上しています。



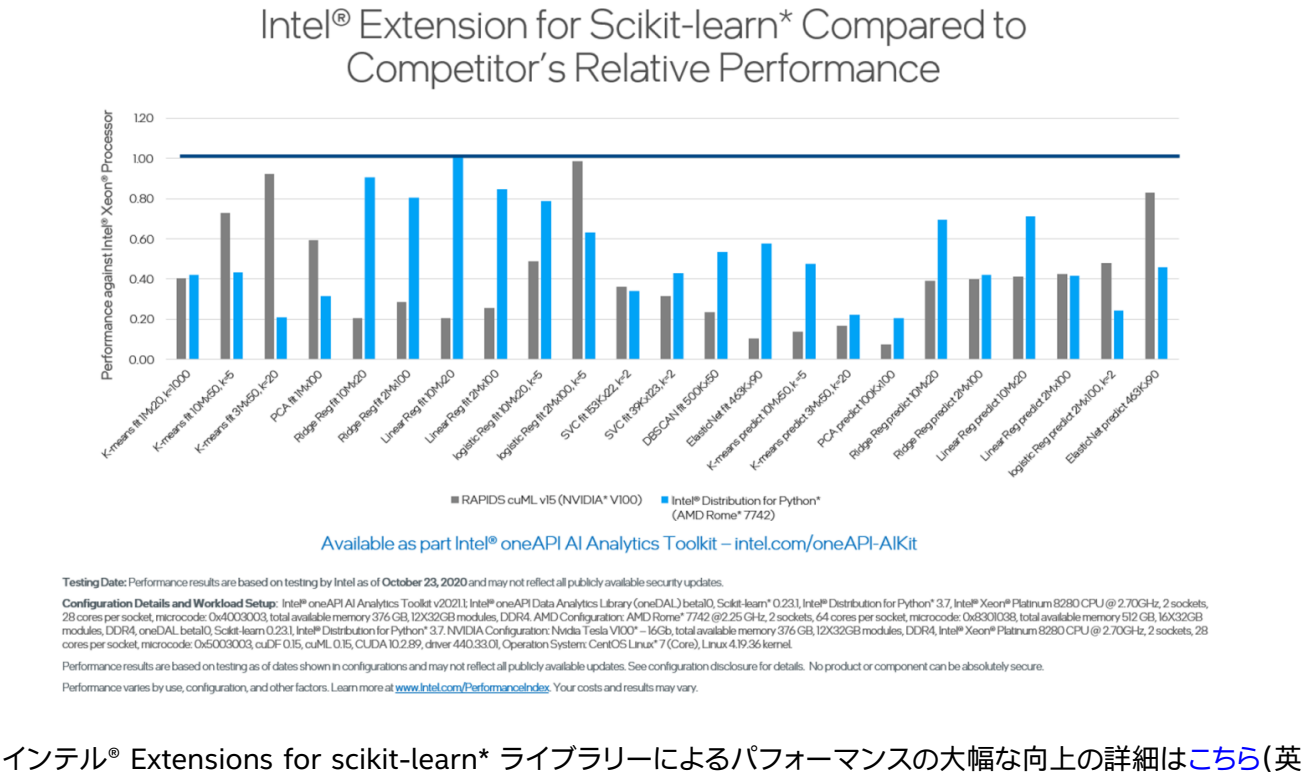
これらの結果を得るため、データの取り込みと ETL にはインテル® ディストリビューション for Modin を、モデルのトレーニングと予測にはインテル® Extensions for scikit-learn* を使用しました。また、CSV の読み込みパフォーマンスは 6 倍、トレーニング・テストの分割パフォーマンスは 4 倍に向上しました。

ベンチマーク 2: インテル® Extensions for scikit-learn* を使用したマシンラーニングのパフォーマンス

インテル® Extensions for scikit-learn* は、効率良いデータレイアウト、ブロッキング、マルチスレッド、ベクトル化を実現します。このベンチマークでは、標準バージョンと比較して、scikit-learn* アルゴリズムのパフォーマンスが 220 倍向上しました。



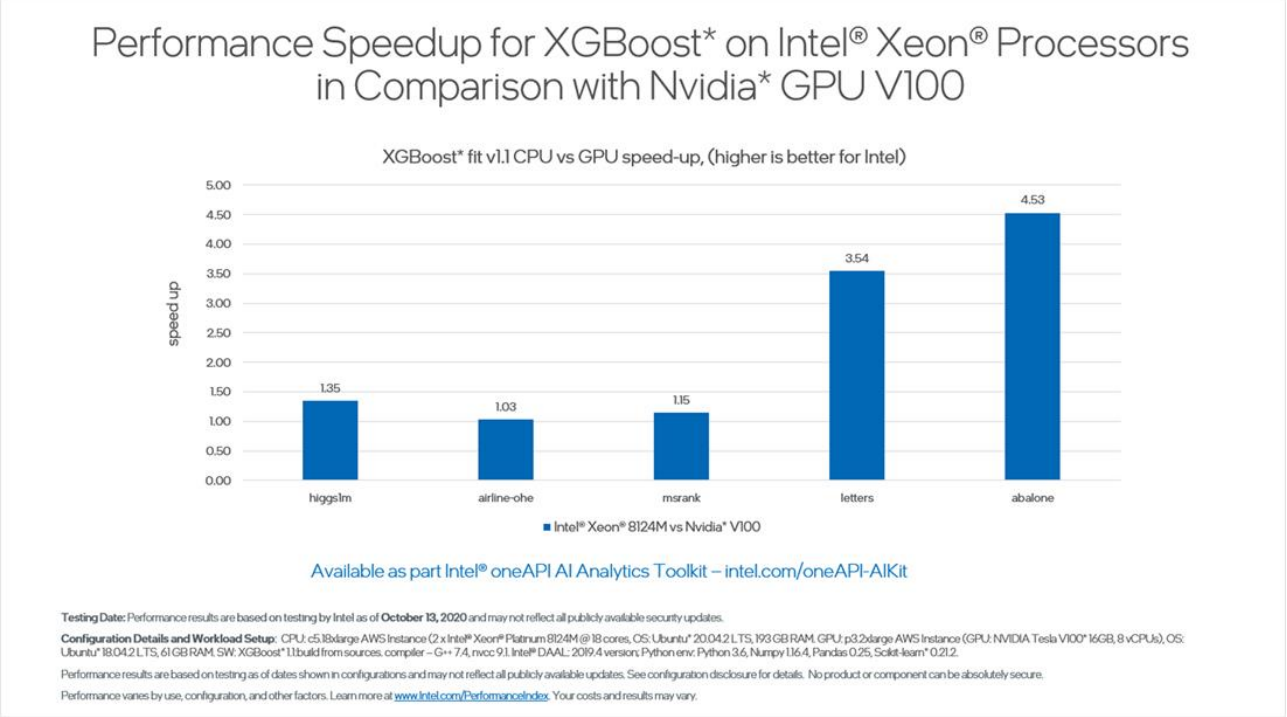
また、同じインテル® Extensions for scikit-learn* アルゴリズムを AMD EPYC* 7742 プロセッサー上で実行した場合よりもパフォーマンスが優れていました。ほとんどのパフォーマンス向上は、AMD のプロセッサーにはないインテル® アドバンスト・ベクトル・エクステンション 512（インテル® AVX-512）によってもたらされました。さらに、インテル® Extensions for scikit-learn* は一貫して NVIDIA* V100 GPU を上回りました。



インテル® Extensions for scikit-learn* ライブラリーによるパフォーマンスの大幅な向上の詳細はこちら（英語）から確認できます。

ベンチマーク 3: インテルにより最適化された XGBoost で NVIDIA* GPU を上回る ML パフォーマンス

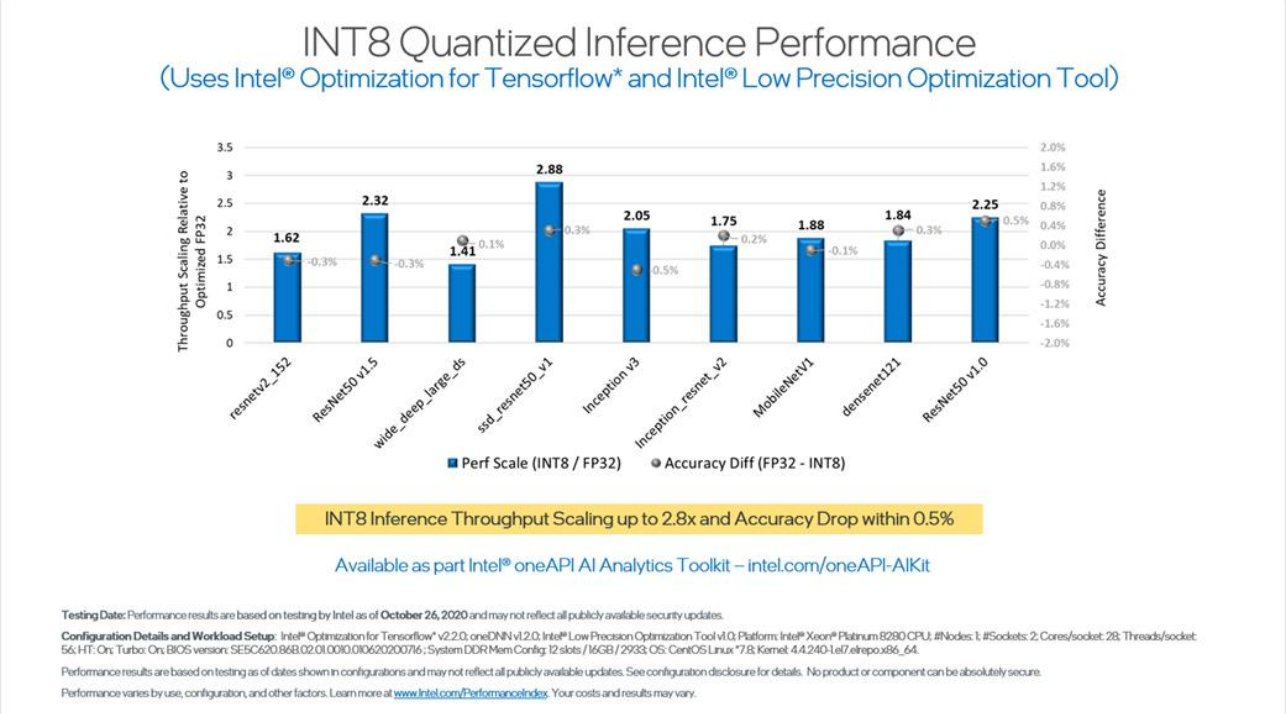
インテルは、オープンソース・プロジェクトである XGBoost の最適化に積極的に取り組んでいます。XGBoost は、Python* やその他のプログラミング言語に勾配ブースティング・フレームワークを提供するオープンソース・ライブラリーです。インテル® oneAPI AI アナリティクス・ツールキットに含まれるインテルにより Python* 向けに最適化された XGBoost ライブラリーを使用することで、NVIDIA* V100 GPU と比較して、インテル® Xeon® プロセッサー上で 4.5 倍高速に推論を実行できます。このグラフは、分類と回帰によく使用される複雑な勾配ブースティング・アルゴリズムにおいて、インテルの CPU が優れたパフォーマンスを達成するだけでなく、このようなワークロードを専用のアクセラレーターにオフロードするコストと労力を節約できることを示しています。



ベンチマーク 4: 精度の低下を最小限に抑えた INT8 で量子化された推論パフォーマンス

近年、ディープラーニングの推論ワークロードのパフォーマンスを向上させるため、低精度の量子化を使用するケースが増えています。しかし、高速化の代償として精度が低下します。インテル® oneAPI AI アナリティクス・ツールキットに新たに追加されたインテル® Low Precision Optimization Tool を使用することで、わずかな精度低下で推論スループットを最大 2.8 倍に向上できます。それを可能にしているのが、ツールに導入されている精度に応じた自動チューニングです。

以下のベンチマークでは、インテル® Low Precision Optimization Tool で FP32 から INT8 へ変換して、インテル® Optimization for TensorFlow* で推論を実行しています。これらのツールは、高速なパフォーマンスを実現するインテル® DL ブースト・テクノロジーをサポートしています。

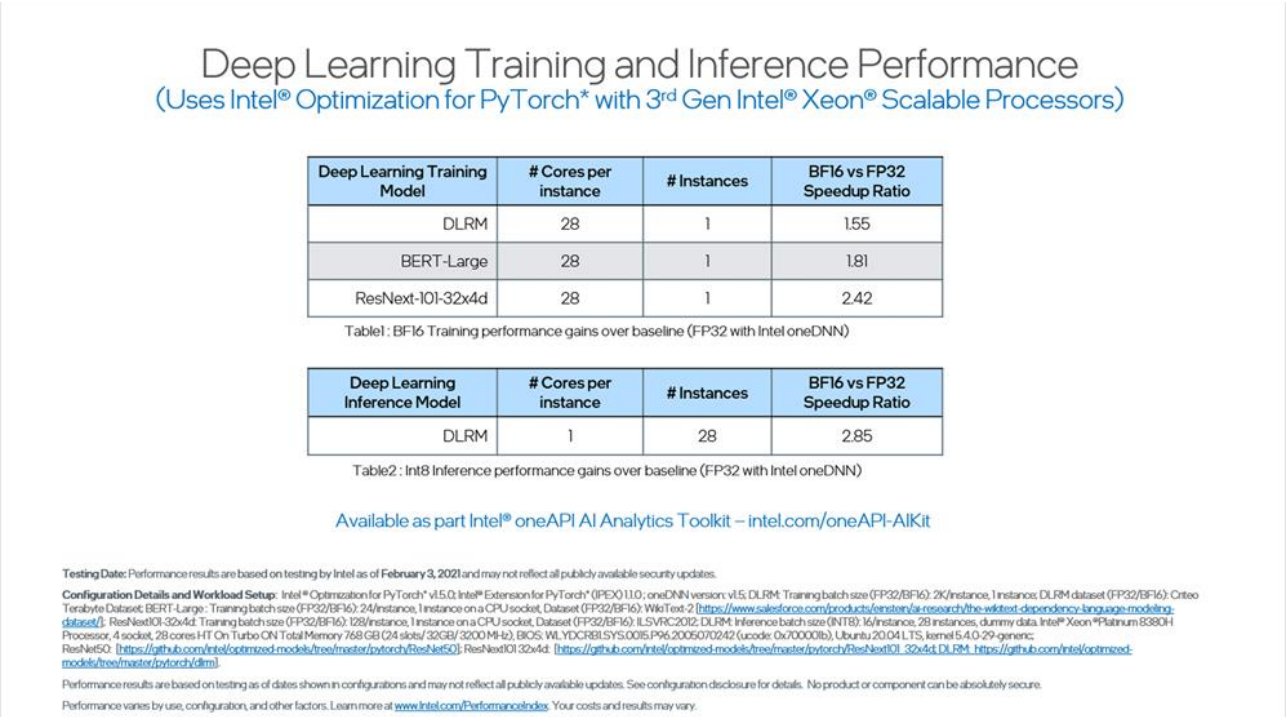


ベンチマーク 5: 第 3 世代インテル® Xeon® スケーラブル・プロセッサーでインテル® Optimization for PyTorch* を使用したディープラーニングのトレーニングと推論のパフォーマンス

インテル® oneAPI AI アナリティクス・ツールキットは、インテルにより最適化されたライブラリーを使用して、CPU アーキテクチャー上で PyTorch* のパフォーマンスを向上します。

インテルは Facebook と協力して、PyTorch* オープンソース・プロジェクトの最適化に貢献しています。また、自動型変換とレイアウト変換によりトレーニングと推論のパフォーマンスを大幅に向上する PyTorch* 拡張ライブラリーを提供しています。

インテル® DL ブーストと BFloat16 テクノロジーを利用した最適化は、DLRM モデルのトレーニング・パフォーマンスを最大 1.55 倍向上します。一方で、インテル® DL ブーストと INT8 による最適化は、DLRM モデルの推論パフォーマンスを、FP32 のパフォーマンスと比較して最大 2.85 倍向上します。このベンチマークは、インテルのハードウェアの進化とソフトウェアの最適化が、新しい推奨モデルやコンピューター・ビジョン・ワークロードにおいて優れたパフォーマンスを発揮することを明確に示しています。



Deep Learning Training and Inference Performance
(Uses Intel® Optimization for PyTorch* with 3rd Gen Intel® Xeon® Scalable Processors)

Deep Learning Training Model	# Cores per instance	# Instances	BF16 vs FP32 Speedup Ratio
DLRM	28	1	1.55
BERT-Large	28	1	1.81
ResNext-101-32x4d	28	1	2.42

Table1 : BF16 Training performance gains over baseline (FP32 with Intel oneDNN)

Deep Learning Inference Model	# Cores per instance	# Instances	BF16 vs FP32 Speedup Ratio
DLRM	1	28	2.85

Table2 : Int8 Inference performance gains over baseline (FP32 with Intel oneDNN)

Available as part Intel® oneAPI AI Analytics Toolkit – intel.com/oneAPI-AIKit

Testing Date: Performance results are based on testing by Intel as of February 3, 2021 and may not reflect all publicly available security updates.

Configuration Details and Workload Setup: Intel® Optimization for PyTorch* v1.5.0; Intel® Extension for PyTorch* (IPEX) 1.1.0.; oneDNN version: v1.5; DLRM: Training batch size (FP32/BF16): 2K/instance, 1 instance; DLRM dataset (FP32/BF16): Criteo Terabyte Dataset; BERT-Large : Training batch size (FP32/BF16): 24/instance, 1 instance on a CPU socket, Dataset (FP32/BF16): WikiText-2 [https://www.salesforce.com/products/einstein/ai-research/the-wikitext-dependency-language-modeling-dataset/]; ResNext101-32x4d: Training batch size (FP32/BF16): 128/instance, 1 instance on a CPU socket, Dataset (FP32/BF16): ILSVRC2012; DLRM: Inference batch size (INT8): 16/instance, 28 instances, dummy data. Intel® Xeon ®Platinum 8380H Processor, 4 socket, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 3200 MHz), BIOS: WLYDCR81.SYS.0015.P96.2005070242 (ucode: 0x700001b), Ubuntu 20.04 LTS, kernel 5.4.0-29-generic; ResNet50: [https://github.com/intel/optimized-models/tree/master/pytorch/ResNet50]; ResNext101 32x4d: [https://github.com/intel/optimized-models/tree/master/pytorch/ResNext101_32x4d; DLRM: https://github.com/intel/optimized-models/tree/master/pytorch/dlrm].

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex. Your costs and results may vary.

インテル® oneAPI AI アナリティクス・ツールキットを始めよう

インテル® oneAPI AI アナリティクス・ツールキットは、最適化された Python* ライブラリー、ディープラーニング・フレームワーク、軽量な並列データフレームを提供します。これらはすべて oneAPI ライブラリーを使用して構築されており、インテル・ハードウェアでエンドツーエンドのデータサイエンスと AI ワークフローのパフォーマンスを最大限に引き出します。このツールキットを使用することで、データ・サイエンティストと AI 開発者は、シームレスな相互運用性とハイパフォーマンスを提供する、最新のインテルによる DL/ML の最適化を単一のリソースで実現できます。

インテル® oneAPI AI ツールキットは無料でダウンロードできます。AI アプリケーションでクロスアーキテクチャーの Python* パフォーマンスを実現してください。

インテル® oneAPI AI ツールキットをダウンロード(英語)して、この記事で紹介したインテルにより最適化されたフレームワークとライブラリーをぜひお試しください。

関連情報

- インテル® AI アナリティクス・ツールキット製品ページ(英語)
- GitHub* のインテル® AI アナリティクス・ツールキットのサンプルコード(英語)
- インテル® AI アナリティクス・ツールキットのサポートフォーラム(英語)

製品とパフォーマンス情報

[1] 実際の性能は利用法、構成、その他の要因によって異なります。詳細は、www.Intel.com/PerformanceIndex(英語)を参照してください。