第 3 世代インテル® Xeon® スケーラブル・プロセッサー上で優れたマシンラーニング・パフォーマンスを実現

この記事は、Intel Tech.Decoded に公開されている「Superior Machine Learning Performance on the Latest Intel® Xeon® Scalable Processors」の日本語参考訳です。

インテルがデータ・サイエンティストに必要なパフォーマンスと使いやすさを提供

第 3 世代インテル® Xeon® スケーラブル・プロセッサーは、人工知能 (AI)、クラウド・コンピューティング、セキュリティーなど、さまざまな分野を強化します。アプリケーションが最新のハードウェアを容易に活用できるように、一連のソフトウェア・ツール、ライブラリー、フレームワークを最適化するインテルの取り組みは、素晴らしい結果をもたらしました。この記事では、ポピュラーな scikit-learn* マシンラーニング (ML) ライブラリーとインテル® Extension for scikit-learn* (英語) に注目します。

インテルは以前、たった 2 行のコードを変更するだけで、第 2 世代インテル® Xeon® スケーラブル・プロセッサーが NVIDIA や AMD のプロセッサーを上回るパフォーマンスを発揮することを実証しました。

ここでは、インテル® Extension for scikit-learn* が最新の第 3 世代インテル® Xeon® スケーラブル・プロセッサーにおいて、旧世代と比較して 1.09 ~ 1.63 倍、NVIDIA DGX* A100 との比較では 0.65 ~ 7.23 倍、そして AMD EPYC* (開発コード名 Milan) との比較では 0.61 ~ 2.63 倍のスピードアップを実現することを紹介します。

インテル® Extension for scikit-learn*

インテル® Extension for scikit-learn* (旧称: daal4py) は、標準の scikit-learn* パッケージのドロップイン置換機能を備えています。通常の scikit-learn* の import の前に 2 行のコードを追加するだけで、パフォーマンスの最適化を利用できます。

```
from sklearnex import patch_sklearn
patch_sklearn()

# the start of the user's code
from sklearn.cluster import DBSCAN
...
```

インテル® Extension for scikit-learn* は、インテルによる最新のディープラーニング (DL) とマシンラーニング (ML) の最適化統合パッケージを提供する、インテル® oneAPI AI アナリティクス・ツールキット (AI キット) (英語) に含まれます。インテル® oneAPI AI アナリティクス・ツールキットは、次のディストリビューション・チャネルからダウンロードできます: Docker* コンテナー、YUM、APT、および Anaconda。インテル® Extension for scikit-learn* コンポーネント (単体) は、PyPI または Conda Forge からダウンロードできます。

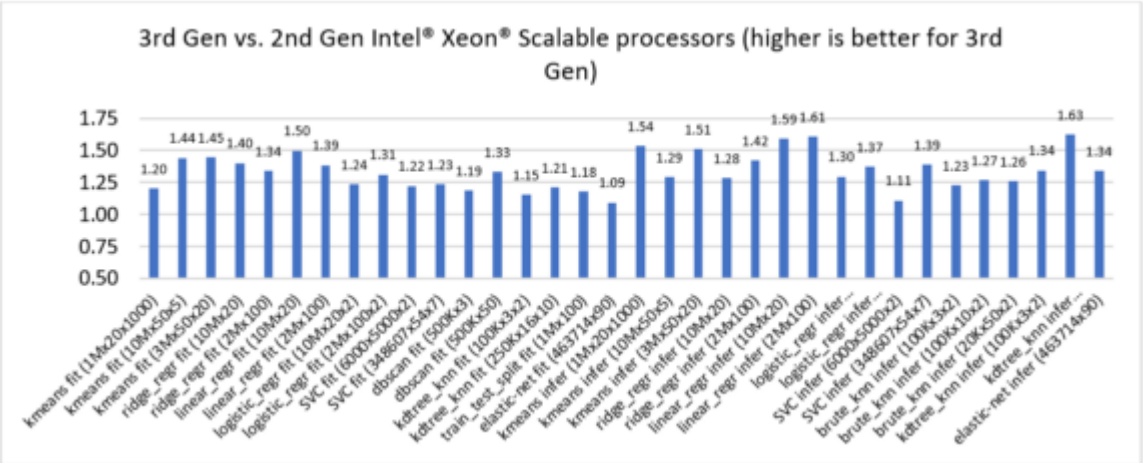
```
pip install scikit-learn-intelex

conda install scikit-learn-intelex -c conda-forge
```

インテル® Extension for scikit-learn* は、インテル® oneAPI データ・アナリティクス・ライブラリー (インテル® oneDAL) を使用して高速化します。インテル® oneDAL は、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) など、最新のベクトル命令をすべて有効にします。また、キャッシュフレンドリーなデータ・ブロッキング、インテル® oneAPI マス・カーネル・ライブラリー (インテル® oneMKL) の高速な BLAS 操作、インテル® oneAPI スレッディング・ビルディング・ブロック (インテル® oneTBB) のスケーラブルなマルチスレッド処理を使用します。

パフォーマンス・リーダーシップ

第 2 世代および第 3 世代インテル® Xeon® スケーラブル・プロセッサー上でインテル® Extension for scikit-learn* のいくつかの ML アルゴリズムのパフォーマンスを比較したところ、トレーニングと推論で 1.09 ~ 1.63 倍のスピードアップが見られました (図 1)。



パート 1: 第 2 世代と比較した第 3 世代インテル® Xeon® スケーラブル・プロセッサーのパフォーマンス向上

競合製品のパフォーマンスと比較するため、第 3 世代インテル® Xeon® スケーラブル・プロセッサーと最新の NVIDIA DGX* A100 と AMD EPYC* (開発コード名 Milan) プロセッサーを比較しました。新しい第 3 世代インテル® Xeon® スケーラブル・プロセッサーは、さまざまな ML アルゴリズムにおいてパフォーマンスのリーダーシップを示し、NVIDIA DGX* A100 との比較では 0.65 ~ 7.23 倍のスピードアップ (図 2)、AMD EPYC* (開発コード名 Milan) との比較では 0.61 ~ 2.63 倍のスピードアップ (図 3) を達成しました。

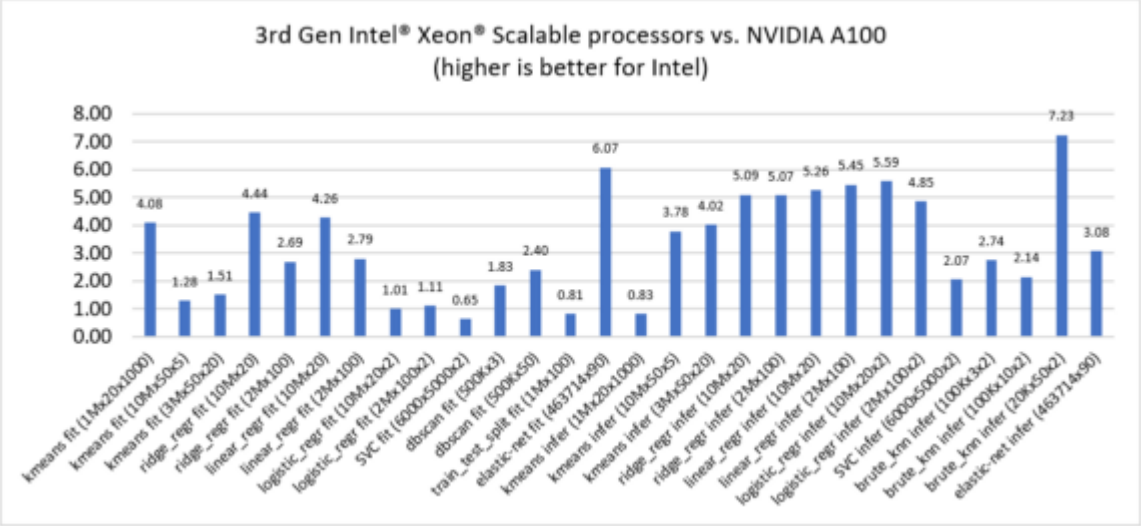


図 2. NVIDIA DGX* A100 (RAPIDS cuML 使用) と比較した第 3 世代インテル® Xeon® スケーラブル・プロセッサー (インテル® Extension for scikit-learn* を使用) のスピードアップ

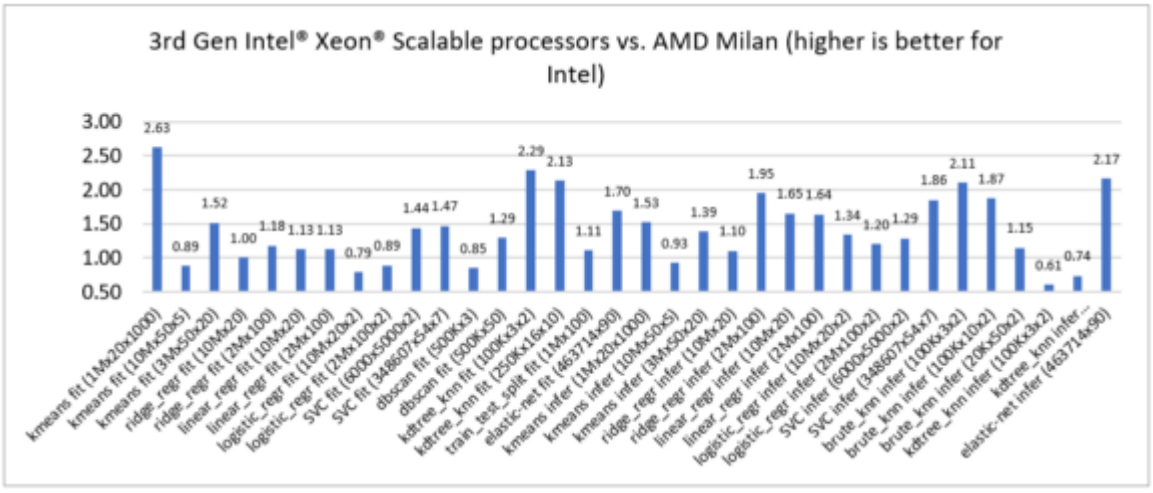


図 3. AMD EPYC* (開発コード名 Milan) と比較した第 3 世代インテル® Xeon® スケーラブル・プロセッサーのスピードアップ (どちらもインテル® Extension for scikit-learn* を使用)

インテルの最先端のデータセンター向けプロセッサー

第 3 世代インテル® Xeon® スケーラブル・プロセッサーは、インテル® ディープラーニング・ブースト (インテル® DL ブースト) (英語) テクノロジーによる AI アクセラレーションを内蔵した柔軟なアーキテクチャーに加え、その他多くの機能が強化されています。

- **高速なメモリー。**1 ソケットあたりのメモリーチャネル数は 6 から 8 に、メモリーの最大周波数は 2933MHz から 3200MHz に増加しました。その結果、DRAM のメモリー帯域幅は最大で 1.45 倍になりました。データ・アナリティクスのワークロードは、多くの処理をインメモリーで行う必要があるため、DRAM 依存であることが多く、第 3 世代インテル® Xeon® スケーラブル・プロセッサーは、これらのワークロードを大幅に改善します。
- **より多くのコア。**第 3 世代インテル® Xeon® スケーラブル・プロセッサーは、1 ソケットあたり最大 40 コアを搭載しており、優れたマルチスレッド・データ処理を実現します。

- **最新のマイクロアーキテクチャー。**1 サイクルあたりの命令数 (IPC) は 4 から 5 に向上し、プロセッサー・コアの実行ポート数は 8 から 10 に増えました。さらに、シングルコアのパフォーマンスを向上する AVX512 BITALG、AVX512 VBMI2 などの新しい命令が追加されました。
- **大きなキャッシュ。**インテル® Xeon® Platinum 8380 プロセッサーは 60MB のラストレベル・キャッシュ (LLC) を搭載おり、これはインテル® Xeon® Platinum 8280L (38.5MB) と比較して 58% 増です。L2 キャッシュは 1 コアあたり 1MB から 1.25MB へ、L1 キャッシュは 1 コアあたり 32KB から 48KB へ増えました。一部の ML アルゴリズムは、キャッシュに格納されたデータの処理に大半の時間を費やすため、キャッシュの改善はパフォーマンスに大きく影響します。
- **新しいレベルのセキュリティー。**ML アルゴリズムは機密データを処理することが多いため、新しい第 3 世代インテル® Xeon® スケーラブル・プロセッサーでは、インテル® ソフトウェア・ガード・エクステンション (インテル® SGX) によるきめ細かな制御が可能なハードウェアベースのメモリー暗号化を提供しています。

インテル® Extension for scikit-learn* による最適化と第 3 世代インテル® Xeon® スケーラブル・プロセッサーの組み合わせは、ML とデータ・アナリティクスのワークロードにおいて優れたパフォーマンスをもたらします。これにより、エンタープライズ・アプリケーションを単一のアーキテクチャー上で実行し、混在するワークロードの総所有コストを最適化し、革新的なソリューションを素早く市場に投入することができます。

ハードウェアとソフトウェアのベンチマーク構成

すべてインテルによるテスト。

Platform	Model	Parameters	Testing date
3rd Gen Intel Xeon Scalable processors	3rd Gen Intel Xeon Platinum 8380 processor	2 sockets; 40 cores per socket; HT: on; Turbo: on; RAM: 512GB (16 slots / 32GB / 3200MHz)	3/19/2021
2nd Gen Intel Xeon Scalable processors	2nd Gen Intel Xeon Platinum 8280L processor	2 sockets; 28 cores per socket; HT: on; Turbo: on; RAM: 384GB (12 slots/ 32GB / 2933 MHz)	2/5/2021
AMD Milan	AMD EPYC™ 7763	AMD EPYC™ 7763 64-Core: 2 sockets; 64 cores per socket; HT: on; Turbo: on; RAM: 512GB (16 slots / 32GB / 3200MHz)	3/8/2021
NVIDIA A100	NVIDIA A100, AMD EPYC™ 7742	NVIDIA A100 Tensor (DGX-A100); AMD EPYC™ 7742 64-Core: 2 sockets; 64 cores per socket; HT: on; Turbo: on; RAM: 512GB (16 slots / 32GB / 3200MHz)	2/4/2021

Software	**CPU workloads**	**GPU workload**
Python	3.7.9	3.7.9
Scikit-learn	Sklearn 0.24.1	-
Intel® Extension for Scikit-learn	2021.2.2	-
NVIDIA RAPIDS	-	RAPIDS 0.17
CUDA Toolkit	-	CUDA 11.0.221

製品および性能に関する情報

[1] 性能は、使用状況、構成、その他の要因によって異なります。詳細については、www.Intel.com/PerformanceIndex/ (英語) を参照してください。