TESLA V100
パフォーマンス ガイド

生命科学アプリケーション

NVIDIA

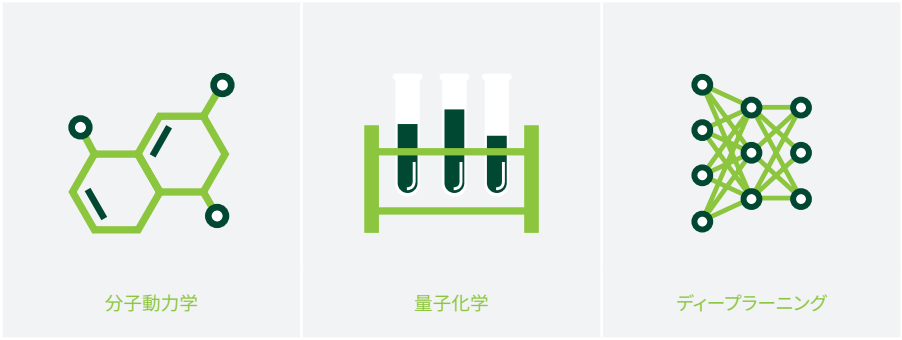
2017 年 11 月

TESLA V100 パフォーマンス ガイド

最新のハイ パフォーマンス コンピューティング (HPC) データセンターは、科学分野と工学分野における世界で最も重要な課題を解決する鍵を握っています。NVIDIA® Tesla® アクセラレーション コンピューティング プラットフォームは、最新のデータセンターで業界をリードするアプリケーションを実行し、HPC や AI のワークロードを高速化します。AI と HPC を組み合わせることで、科学の幅広い分野に対応し、科学技術イノベーションはこれまでにないスピードで進められています。Tesla V100 GPU は最新のデータセンターのエンジンです。少ないサーバーで画期的なパフォーマンスを実現できるため、コストを大幅に削減すると共に、短期間でインサイトが得られます。パフォーマンスが向上し、問題解決までの時間が短縮され、収益と生産性が大きく向上します。

あらゆる HPC データセンターにおいて、Tesla プラットフォームはメリットをもたらします。幅広い分野の 500 を超える HPC アプリケーション (上位 15 の HPC アプリケーションすべてを含む) と、主要なディープラーニング フレームワークのすべてが GPU 向けに最適化されています。

GPU アクセラレーション アプリケーションを使用する研究分野:

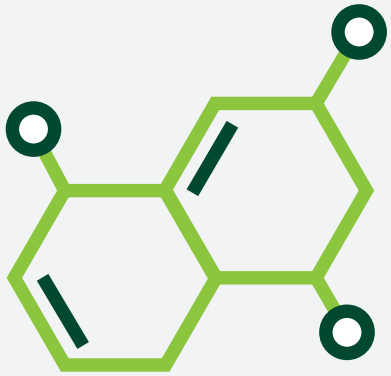


500 を超える HPC アプリケーションおよびすべてのディープラーニング フレームワークが GPU アクセラレーション対応です。

- GPU アクセラレーション アプリケーションの最新のカタログは、こちらをご覧ください。
  www.nvidia.com/teslaapps (英語)
- 幅広いアクセラレーション対応アプリケーションに対して、すばやく GPU で簡単な命令セットを実行する手順については、こちらをご覧ください。
  www.nvidia.com/gpu-ready-apps (英語)

TESLA V100 パフォーマンス ガイド

分子動力学



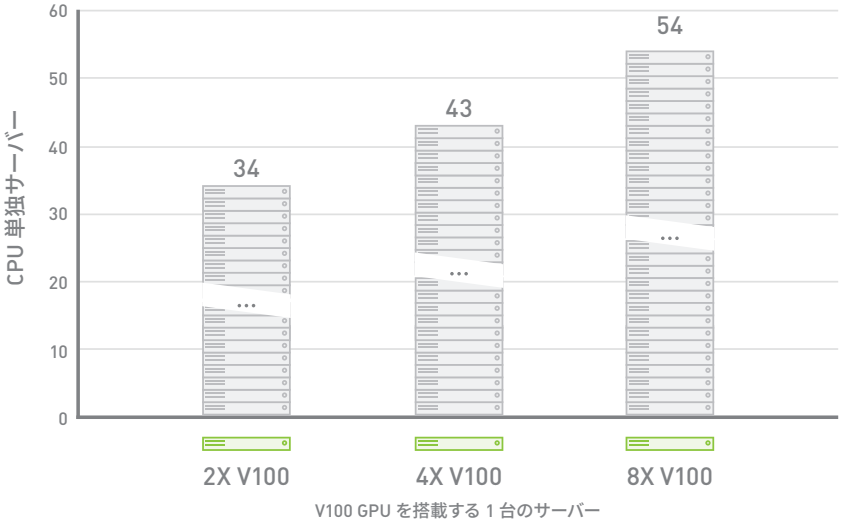
分子動力学 (MD) は、HPC データセンターのワークロードの大部分を占めています。主要な MD アプリケーションはすべて GPU アクセラレーションに対応しています。このため、科学者は、これまでの CPU 単独のシステムでは実行できなかったシミュレーションを実行することができるようになります。MD アプリケーションを実行する場合、Tesla V100 GPU を使用するデータセンターでは、サーバーとインフラの取得コストを最大 80% 節約できます。

MD における TESLA プラットフォームと V100 の主な特長

- V100 を搭載するサーバーは、HOOMD-Blue、Amber などのアプリケーションにおいて、CPU サーバー最大 54 台分に匹敵
- 主要な MD アプリケーションは 100% GPU アクセラレーション対応
- FFT、BLAS などの主要な数学ライブラリ
- GPU あたり最大 15.7 TFLOPS の単精度演算能力
- GPU あたり最大 900 GB/秒のメモリ帯域幅

該当するすべてのアプリケーションは、こちらからご確認ください。
www.nvidia.com/molecular-dynamics-apps (英語)

HOOMD-Blue Performance Equivalence
1 台の GPU サーバーと同等の CPU 単独サーバー



CPU サーバー: Dual Xeon E5-2690 v4 @ 2.6 GHz、GPU サーバー: 同じ CPU サーバーに NVIDIA® Tesla® V100 for PCIe を搭載 | NVIDIA CUDA® バージョン 9.0.145 | データセット: Microsphere | 同等の CPU ノードを求めるために、最大 8 CPU ノードの測定ベンチマークを使用。その後、9 ノード以上に線形スケーリング。

HOOMD-BLUE
粒子力学パッケージは GPU 向けにゼロから記述されている

バージョン
2.1.6

アクセラレーション機能
CPU バージョンと GPU バージョンを提供

スケーラビリティ
マルチ GPU、マルチノード

関連情報
http://codeblue.umich.edu/hoomd-blue/index.html (英語)

AMBER におけるパフォーマンス比較
1 台の GPU サーバーと同等の CPU 単独サーバー

CPU サーバー: Dual Xeon E5-2690 v4 @ 2.6 GHz、GPU サーバー: 同じ CPU サーバーに NVIDIA® Tesla® V100 for PCIe を搭載 | NVIDIA CUDA® バージョン: 9.0.103 | データセット: PME-Cellulose_NVE | 同等の CPU ノードを求めるために、最大 8 CPU ノードの測定ベンチマークを使用。その後、9 ノード以上に線形スケーリング。

AMBER
生体分子力学をシミュレートするプログラム スイート

バージョン
16.8

アクセラレーション機能
PMEMD 陽溶媒および GB。陽溶媒/陰溶媒、REMD、aMD

スケーラビリティ
マルチ GPU、シングルノード

関連情報
http://ambermd.org/gpus (英語)

TESLA V100 パフォーマンス ガイド

量子化学

量子化学 (QC) シミュレーションは新薬や新原料の発見にとって重要であり、HPC データセンターのワークロードの多くの部分を占めています。今日の主要 QC アプリケーションの 60% は GPU アクセラレーション対応です。QC アプリケーションを実行する場合、Tesla V100 GPU を使用するデータセンターのワークロードでは、サーバーとインフラの取得コストを 30% 以上節約できます。

QC における TESLA プラットフォームと V100 の主な特長

- V100 を搭載するサーバーは、VASP などのアプリケーションにおいて、CPU サーバー最大 5 台分に匹敵
- 主要な QC アプリケーションの 60% が GPU アクセラレーション対応
- FFT、BLAS などの主要な数学ライブラリ
- GPU あたり最大 7.8 TFLOPS の倍精度演算能力
- 大規模データセットに対して最大 32 GB のメモリ容量

該当するすべてのアプリケーションは、こちらからご確認ください。
www.nvidia.com/quantum-chemistry-apps

VASP におけるパフォーマンス比較
1 台の GPU サーバーと同等の CPU 単独サーバー

V100 GPU を搭載する 1 台のサーバー

CPU サーバー: Dual Xeon E5-2690 v4 @ 2.6 GHz、GPU サーバー: 同じ CPU サーバーに NVIDIA® Tesla® V100 for PCIe を搭載 | NVIDIA CUDA® バージョン: 9.0.103 | データセット: Si-Huge | 同等の CPU ノードを求めるために、最大 8 CPU ノードの測定ベンチマークを使用。その後、9 ノード以上に線形スケーリング。

VASP
ab-initio 量子力学的分子力学 (MD) シミュレーションを実行するためのパッケージ

バージョン
5.4.4

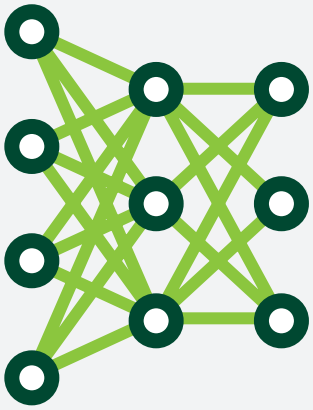
アクセラレーション機能
RMM-DIIS、Blocked Davidson、K-points、Exact-Exchange

スケーラビリティ
マルチ GPU、マルチノード

関連情報
www.nvidia.com/vasp (英語)

TESLA V100 パフォーマンス ガイド

ディープラーニング



ディープラーニングは、数年前まで解決が難しいと思われていた科学、企業、消費者の重要な問題を解決しています。主要なディープラーニングフレームワークはすべて NVIDIA GPU 向けに最適化されているため、データサイエンティストや研究者は自らの業務に人工知能を活用することができます。Tesla V100 GPU を使用するデータセンターは、ディープラーニングのトレーニングと推論のフレームワークの実行において、サーバーとインフラの取得コストを最大 85% 削減できます。

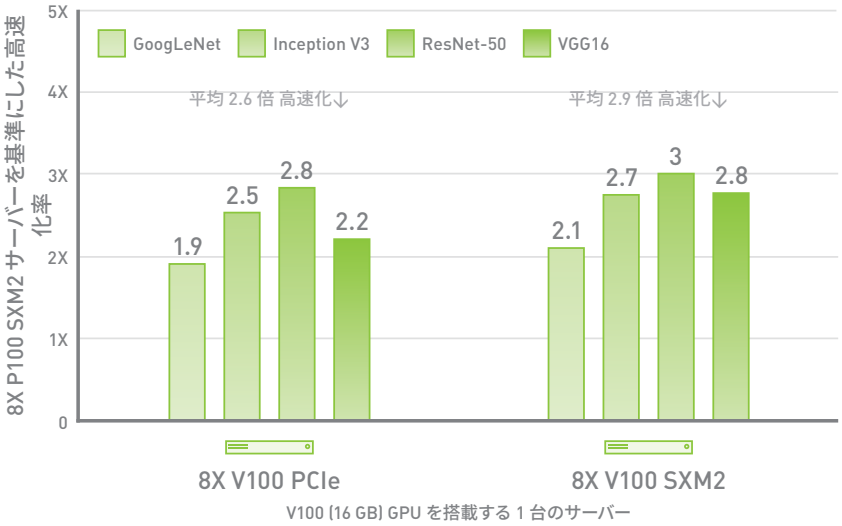
ディープラーニング トレーニングにおける TESLA プラットフォームと V100 の主な特長

- P100 と比較して Tesla V100 では Caffe、TensorFlow、CNTK が 3 倍高速化
- 主要なディープラーニング フレームワークのすべてが GPU アクセラレーションに対応
- 最大 125 TFLOPS の混合精度演算
- 最大 32 GB のメモリ容量と最大 900 GB/秒のメモリ帯域幅

該当するすべてのアプリケーションは、こちらからご確認ください。
www.nvidia.com/deep-learning-apps (英語)

Caffe ディープラーニング フレームワーク

8X V100 GPU サーバーと 8X P100 GPU サーバーのトレーニング比較

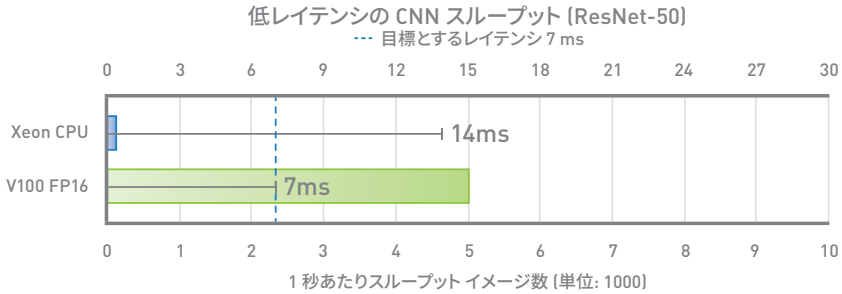


V100 (16 GB) GPU を搭載する 1 台のサーバー

CPU サーバー: Dual Xeon E5-2698 v4 @ 3.6 GHz。GPU サーバーは表記のとおり | Ubuntu 14.04.5 | CUDA バージョン: CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | ドライバー 384.66 | データセット: ImageNet | バッチ サイズ: GoogleNet 192、Inception V3 96、ResNet-50 64 (P100 SXM2) および 128 (Tesla P100)、VGG16 96

低レイテンシ CNN 推論パフォーマンス

低レイテンシで強力なスループットと優れた効率を発揮



1 秒あたりスループット イメージ数 (単位: 1000)

システム構成: シングルソケット Xeon E2690 v4 @ 3.5 GHz およびシングル NVIDIA® Tesla® V100 GPU (TensorRT 3 RC を実行) 対 Intel DL SDK beta 2 | Ubuntu 14.04.5 | CUDA バージョン: 7.0.1.13 | CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | ドライバー 384.66 | 精度: CPU FP32、NVIDIA Tesla V100 FP16

CAFFE
カリフォルニア大学バークレー校で開発された有名な GPU アクセラレーション ディープラーニング フレームワーク

バージョン
1.0

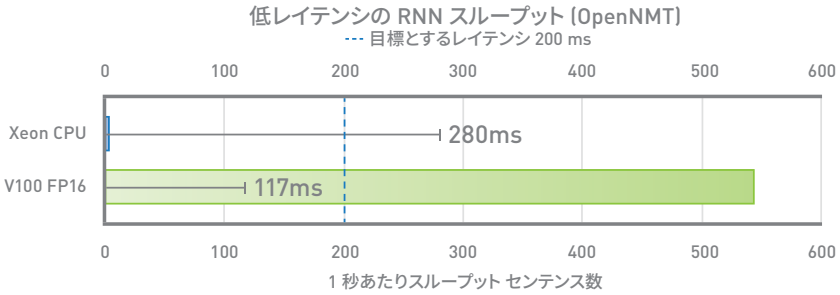
アクセラレーション機能
フル フレームワーク アクセラレーション

スケーラビリティ
マルチ GPU

関連情報
caffe.berkeleyvision.org

低レイテンシ RNN 推論パフォーマンス

低レイテンシで強力なスループットと優れた効率を発揮



システム構成: シングルソケット Xeon E2690 v4 @ 3.5 GHz およびシングル NVIDIA® Tesla® V100 GPU (TensorRT 3 RC を実行) 対 Intel DL SDK beta 2 | Ubuntu 14.04.5 | CUDA バージョン: 7.0.1.13 | CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | ドライバー 384.66 | 精度: CPU FP32、NVIDIA Tesla V100 FP16

TESLA V100 製品仕様

	用 NVIDIA Tesla V100 PCIe ベースのサーバー	NVLink 最適化サーバー用 NVIDIA Tesla V100
倍精度演算能力	最大 7 TFLOPS	最大 7.8 TFLOPS
単精度演算能力	最大 14 TFLOPS	最大 15.7 TFLOPS
ディープラーニング	最大 112 TFLOPS	最大 125 TFLOPS
NVIDIA NVLink™ 相互接続帯域幅	-	300 GB/秒
PCIe x 16 相互接続帯域幅	32 GB/秒	32 GB/秒
CoWoS HBM2 積層メモリ容量	32 GB	32 GB
CoWoS HBM2 積層メモリ帯域幅	900 GB/秒	900 GB/秒

前提と免責

GPU アクセラレーション対応の主要なアプリケーションの割合は、i360 レポート「HPC Support for GPU Computing」に掲載された主要 50 アプリケーションの一覧に基づいて表記されています。
スループットとコスト削減は、当該分野のアプリケーション ベンチマークが等しい計算サイクルを実行するときのワークロード プロファイルに基づいて算出されています。
http://www.intersect360.com/industry/reports.php?id=131 (英語)
1 GPU ノードに相当する CPU ノードの数は、GPU ノード アプリケーションの高速化とマルチ CPU ノードのスケーリング パフォーマンスのラボ パフォーマンス結果を使用して算出されています。たとえば、分子力学アプリケーション HOOMD-Blue の場合、GPU ノード アプリケーションは 37.9 倍高速化されています。CPU ノードを 8 ノード クラスターにスケーリングすると、合計システム出力は 7.1 倍になります。したがって、スケーリング係数は 8 ÷ 7.1 = 1.13 です。1 GPU ノードのパフォーマンスに相当する CPU ノードの数は、37.9 (GPU ノード アプリケーションの高速化) x 1.13 (CPU ノードのスケーリング係数) = 43 ノードになります。