Datasheet

NVIDIA L40S

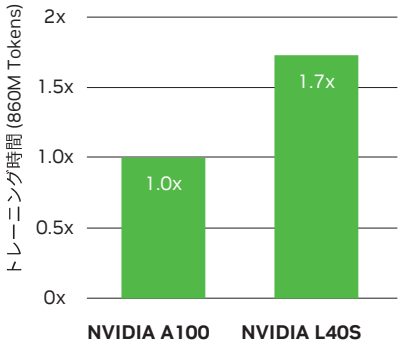
データセンター向けの比類のない AI およびグラフィックスのパフォーマンス

生成 AI は変革を促進し、あらゆる業界の企業に新たな機会のフロンティアをもたらします。 AI で変革するには、企業はより多くのコンピューティング リソース、より大規模なスケール、そして増え続ける多様で複雑なワークロードの要求を満たす幅広い機能セットを必要とします。
NVIDIA L40S GPU は、データセンター向けの最も強力なユニバーサル GPU であり、生成 AI やモデルのトレーニングと推論から 3D グラフィックス、レンダリング、ビデオ アプリケーションに至るまで、次世代の AI 対応アプリケーションにエンドツーエンドのアクセラレーションを提供します。

次世代のワークロードを加速する

- 生成 AI
- 大規模言語モデル (LLM) トレーニングと推論
- NVIDIA Omniverse™ Enterprise
- レンダリングと 3D グラフィックス
- ストリーミングとビデオコンテンツ

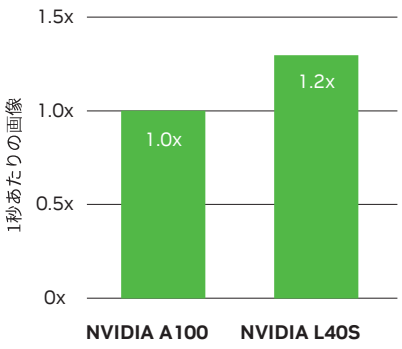
AI トレーニング



GPT-408 LoRA (相対性能)

Fine-tuning LoRA (GPT-40B, GPT-175B): global train batch size: 128 (sequences), seq-length: 256 (tokens). NVIDIA HGX™ A100 (8x A100) vs. Two systems with 4x L40S GPUs. Performance based on prerelease build, subject to change.

生成 AI



Stable Diffusion, 512x512 (相対性能)

Stable Diffusion v2.1. Relative speedup for 512 x 512 resolution image generation. NVIDIA HGX A100 (8x A100) vs. Two systems with 4x L40S GPUs. Performance based on prerelease build, subject to change.

NVIDIA Ada Lovelace アーキテクチャ搭載

第4世代 Tensor コア

構造的スパース性と最適化された TF32 フォーマットのハードウェア サポートにより、すぐに使用できるパフォーマンスの向上が実現し、AI およびデータ サイエンス モデルのトレーニングが高速化されます。 DLSS を使用して AI で強化されたグラフィックス機能を高速化し、解像度をアップスケールし、選択したアプリケーションでのパフォーマンスを向上させます。

RYOYO

第3世代 RT コア

スループットの向上とレイトレーシングとシェーディングの同時実行機能により、レイトレーシングのパフォーマンスが向上し、製品設計やアーキテクチャ、エンジニアリング、建設のワークフローのレンダリングが高速化されます。 ハードウェア アクセラレーションによるモーション ブラーと見事なリアルタイム アニメーションを使用して、実際に動作している本物のようなデザインをご覧ください。

Transformer Engine

Transformer Engine は AI のパフォーマンスを劇的に加速し、トレーニングと推論の両方でのメモリ使用率を向上させます。 Ada Lovelace 第 4 世代 Tensor コアのパワーを利用して、Transformer Engine はトランスフォーマー アーキテクチャのニューラル ネットワークのレイヤーをインテリジェントにスキャンし、FP8 と FP16 の精度の間で自動的にリキャストすることで、より高速な AI パフォーマンスを実現し、トレーニングと推論を加速します。

データセンターに対応

L40S GPU は、エンタープライズ データ センターの 24 時間年中無休の運用に最適化されており、NVIDIA によって設計、構築、テスト、サポートが行われ、最大のパフォーマンス、耐久性、稼働時間を保証します。 L40S GPU は最新のデータセンター標準を満たしており、Network Equipment-Building System (NEBS) レベル 3 に対応しており、ルート オブ トラスト テクノロジーによるセキュア ブートを備えており、データセンターに追加のセキュリティ層を提供します。

技術仕様	
GPU アーキテクチャ	NVIDIA Ada Lovelace アーキテクチャ
GPU メモリー	48GB GDDR6 ECC付き
メモリー帯域幅	864GB/s
インターコネクトインタフェース	PCIe Gen4 x16: 64GB/s 双方向
NVIDIA Ada Lovelace アーキテクチャベースの CUDA® コア	18,176
NVIDIA 第3世代 RT コア	142
NVIDIA 第4世代 Tensor コア	568
RT コア性能 TFLOPS	209
FP32 TFLOPS	91.6
TF32 Tensor コア TFLOPS	183 I 366*
BFLOAT16 Tensor コア TFLOPS	362.05 I 733*
FP16 Tensor コア	362.05 I 733*
FP8 Tensor コア	733 I 1,466*
ピーク INT8 Tensor TOPS	733 I 1,466*
ピーク INT4 Tensor TOPS	733 I 1,466*
フォームファクター	4.4" (H) x 10.5" (L), デュアルスロット
ディスプレイポート	4x DisplayPort 1.4a
最大消費電力	350W
電源コネクター	16-pin

菱洋エレクトロ株式会社
ソリューション事業本部
ソリューション第 5 ビジネスユニット

【お問い合わせ】
03-3546-6211
nvidia_ws_info@ryoyo.co.jp