Datasheet

NVIDIA L4 Tensor コア GPU

ビデオ、AI、グラフィックスを効率的に実現する
飛躍的進化を遂げたユニバーサル アクセラレータ

ビデオ、AI、グラフィックスのワークロードを加速

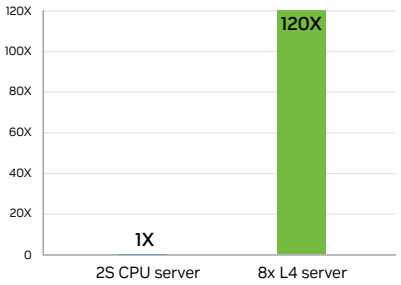
NVIDIA Ada Lovelace L4 Tensor コア GPU は、企業、クラウド、およびエッジでの ビデオ、AI、仮想デスクトップおよびグラフィックス アプリケーションにユニバーサルなアクセラレーションとエネルギー効率を提供します。NVIDIA の AI プラットフォームとフルスタック アプローチにより、L4 はレコメンデーション、音声ベースの AI アバターアシスタント、ジェネレーティブ AI、画像検索、コンタクト センター自動化など、幅広い AI アプリケーションの大規模な推論用に最適化され、最高のパーソナライズされた体験を実現します。

メインストリーム向けの最も効率的な NVIDIA アクセラレーターである L4 を搭載したサーバーは、CPU ソリューションよりも最大 120 倍高速な AI ビデオ 性能と 2.7 倍のジェネレーティブ AI での性能向上、および前世代の GPU よりも 4 倍以上のグラフィックス パフォーマンスを発揮します。NVIDIA L4 のシングル スロットの薄型フォーム ファクターは汎用性とエネルギー効率に優れ、エッジ ロケーションを含むグローバル展開に最適です。

製品仕様	
FP32	30.3 teraFLOPs
TF32 Tensor コア	120 teraFLOPS*
FP16 Tensor コア	242 teraFLOPS*
BFLOAT16 Tensor コア	242 teraFLOPS*
FP8 Tensor コア	485 teraFLOPs*
INT8 Tensor コア	485 TOPs*
GPU メモリ	24GB
GPU メモリ帯域幅	300 GB/s
NVENC | NVDEC | JPEG デコーダー	2 | 4 | 4
最大熱設計電力 **(TDP)**	72W
フォームファクター	1スロット ロープロファイル, PCIe
インターコネクト	PCIe Gen4 x16 64GB/s
サーバー オプション	1-8基 GPU搭載のパートナーおよびNVIDIA-Certified Systems対応システム

*スパース性を使用、スパース を使わない場合は 1/2 低くなります。

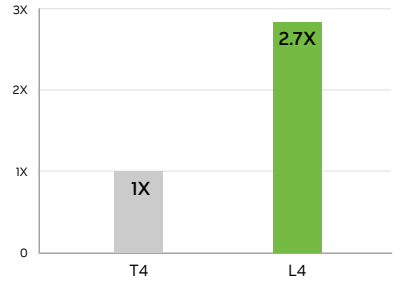
最大 120 倍高速なAI ビデオ



L4 AI Video Performance

パフォーマンス測定: 8x L4 vs 2S Intel 8362 CPU サーバーの比較、CV-CUDA® デコード、前処理、推論 (SegFormer)、後処理、エンコード、NVIDIA® TensorRT™ 8.6 vs OpenCV を使用した CPU のみのパイプラインによるエンドツーエンドのビデオ パイプライン 4.7、PyTorch

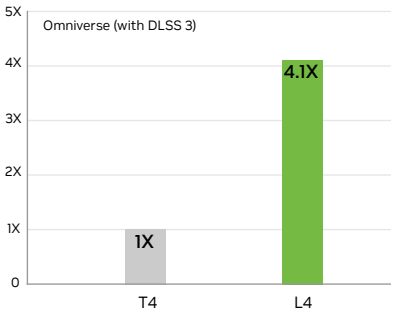
2.7倍のジェネレーティブAI 性能向上



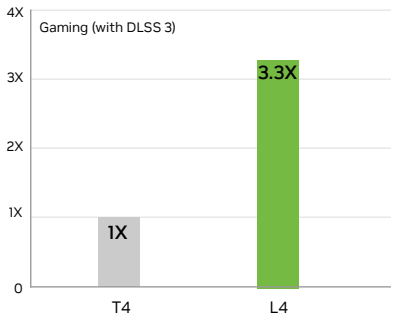
L4 Generative AI Performance

予測されるパフォーマンス: L4 対 T4 画像生成、512x512 安定拡散 v2.1、FP16、TensorRT 8.5.2

4 倍以上のリアルタイム レンダリング性能



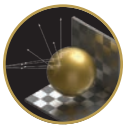
3 倍以上のレイ トレーシング 性能



L4 Visual Computing Performance

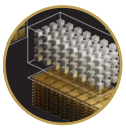
パフォーマンス測定:
リアルタイム レンダリング: NVIDIA ディープラーニング スーパー サンプリング (DLSS) 3 による 1080p および 4K でのリアルタイム レンダリングによる NVIDIA Omniverse™ パフォーマンス
レイトレーシング: レイトレーシングと DLSS 3 をサポートする AAA タイトルのゲーム パフォーマンスの幾何平均

NVIDIA Ada Lovelace アーキテクチャのブレークスルー



第3世代 RT コア

NVIDIA は、RT コアの発明により、リアルタイムでのレイ トレーシングを現実のものにしました。RT コアは、パフォーマンス集約型のレイ トレーシング レンダリングに対応するために特別に設計された GPU 上の処理コアです。
Ada Lovelace の第 3 世代 RT コアは、レイ-トライアングルの交差スループットが 2 倍になり、RT-TFLOP のパフォーマンスが 2 倍以上向上します。NVIDIA Shader Execution Reordering (SER) は、パフォーマンスを 3 倍以上向上させ、仮想世界での没入型体験と、AI ベースのニューラル グラフィックスとクラウド ゲームで前例のない生産性を実現します。



第4世代 Tensor コア

Ada Lovelace アーキテクチャの Tensor コアは、インテリジェント チャットボット、ジェネレーティブ AI、自然言語処理 (NLP)、コンピューター ビジョン、NVIDIA DLSS 3 などの革新的な AI テクノロジを加速するように設計されています。

Ada Lovelace Tensor コアは、構造化されたスパース性と 8 ビット浮動小数点 (FP8) 前世代[1]より最大 4 倍高い推論パフォーマンスの精度を実現、FP8 は、より大きな精度と比較してメモリ負荷を軽減し、AI スループットを劇的に加速します。

高度なビデオおよびビジョン AI アクセラレーション

最適化された AV1 スタックにより、NVIDIA L4 はビデオとビジョン AI アクセラレーションを次のレベルに引き上げ、リアルタイム ビデオ トランスコーディング、ストリーミング、ビデオ会議、拡張現実 (AR)、仮想現実 (VR)、およびビジョン AIを加速。4 つのビデオ デコーダーと 2 つのビデオ エンコーダーを AV1 ビデオ フォーマットと組み合わされ、L4 サーバーは 1,000 [2]以上のコンカレントビデオをホストとCPU ソリューション[3]の 120 倍以上の AI ビデオ エンドツーエンド パイプライン性能を発揮。さらに 4 つの JPEG デコーダーが、コンピュータービジョンで大きなパワーを必要とするアプリケーションをさらに高速化します。

Deep Learning Super Sampling (DLSS)

NVIDIA DLSS 3 は、レンダリング パフォーマンスを大幅に向上させる、AI を活用したグラフィックスにおける画期的なブレークスルーです。新しい第 4 世代の Tensor コアと L4 上の NVIDIA オプティカル フロー アクセラレータ (OFA) を搭載した DLSS 3 は、AI を使用して、グラフィックス ベースのワークロード用の追加の高品質フレームを作成します。

仮想環境に対応

次世代の NVIDIA 仮想 GPU (vGPU) ソフトウェアの改良と前世代の 1.5 倍の GPU メモリにより、L4 は NVIDIA RTX™ 仮想ワークステーション (vWS) で実行されるミッドエンドからハイエンドの設計ワークフローでワークステーションのパフォーマンスを 1.7 倍向上させます。また、NVIDIA 仮想 PC (vPC) で実行される生産性アプリケーションを高速化します。

データセンターの効率とセキュリティ

NVIDIA L4 は、24 時間/7日のエンタープライズ データ センター運用向けに最適化されており、最大のパフォーマンス、耐久性、セキュリティを実現するために、NVIDIA とパートナーによって設計、構築、広範なテスト、サポートが行われています。L4 は、ルート オブ トラスト テクノロジーを使用したセキュア ブートを特徴としており、データ センターに追加のセキュリティ レイヤーを提供します。

1.T4 の FP16 と比較した L4 の FP8

2. 720p30 で 8x L4 AV1 低遅延 P1 プリセット エンコード

3. 8基の L4 と 2S Intel 8362 CPU サーバーのパフォーマンス比較: CV-CUDA の前処理と後処理、デコード、推論 (SegFormer)、エンコード、TRT 8.6 と OpenCV を使用した CPU のみのパイプラインによるエンドツーエンドのビデオ パイプライン

ワークロードを効率的かつ持続的に加速

NVIDIA L4 は、NVIDIA データ センター プラットフォームに不可欠な要素です。AI、ビデオ、仮想ワークステーション、グラフィックス、シミュレーション、データ サイエンス、データ分析向けに構築されたこのプラットフォームは、3,000 を超えるアプリケーションを高速化し、データセンターからエッジ、クラウドに至るまで、あらゆる場所で大規模に利用可能であり、劇的なパフォーマンスの向上とエネルギー効率の両方を実現します。

AI とビデオがより普及するにつれて、効率的で費用対効果の高いコンピューティングに対する需要がこれまで以上に高まっています。NVIDIA L4 Tensor コア GPU は、最大 120 倍優れた AI ビデオ パフォーマンスを提供し、従来の CPU ベースのインフラストラクチャと比較して最大 99% 優れたエネルギー効率と低い総所有コストを実現します。これにより、企業はラック スペースを削減し、二酸化炭素排出量を大幅に削減しながら、データ センターをより多くのユーザーに拡張できます。2 メガワット (MW) のデータ センターで CPU から NVIDIA L4 に切り替えることで節約されるエネルギーは、1 年間で 2,000 世帯以上の電力に相当し、10 年間[4] [5]で成長する 172,000 本の木のカーボン オフセットに匹敵します

エンタープライズ対応: AI ソフトウェアが開発と展開を合理化

企業による AI の採用が進んでおり、この新しい時代に対応できるエンド ツー エンドの AI 対応インフラストラクチャを必要としています。NVIDIA AI Enterprise は、エンドツーエンドのクラウドネイティブな AI およびデータ分析ソフトウェアのスイートであり、すべての組織が AI を活用できるように最適化されており、エンタープライズ データ センターからクラウドまで、どこにでも展開が可能です。またAI プロジェクトを順調に進めるためのグローバルでのエンタープライズ サポートにも対応します。

4. 8x L4 と 2S Intel 8362 CPU サーバーの比較: CV-CUDA の前処理と後処理、デコード、推論 (SegFormer)、エンコード、TRT 8.6 と OpenCV 4.7、PyT を使用した CPU のみのパイプラインによるエンドツーエンドのビデオ パイプライン 推論

5. 1.677MW の節約で設定した EPA 計算の結果 www.epa.gov/energy/greenhouse-gas-equivalencies-calculator

AI の開発と展開を合理化するために最適化された NVIDIA AI Enterprise には、一般的なデータ センター プラットフォームと、NVIDIA L4 Tensor コア GPU を備えた主流の NVIDIA-Certified Systems™ で認定されている、実績のあるオープンソースのコンテナーとフレームワークが含まれています。これにはサポートが含まれており、組織はオープン ソースの透明性と、AI の専門家と IT 管理者の両方に対する AI の専門知識を備えたグローバルな NVIDIA エンタープライズ サポートの保証を得ることができます。

NVIDIA AI Enterprise ソフトウェアは、NVIDIA L4 Tensor コア GPU に追加されるライセンスであり、ほぼすべての組織がトレーニング、推論、データ サイエンスで最高のパフォーマンスを発揮する AI を利用できるようにします。NVIDIA AI Enterprise と NVIDIA L4 を組み合わせることで、AI 対応プラットフォームの構築が簡素化され、AI の開発と展開が加速され、パフォーマンス、セキュリティ、およびスケーラビリティが提供されて、洞察をより迅速に収集し、ビジネス価値をより早く実現できます。

NVIDIA LaunchPad経由での **NVIDIA AI Enterprise labs** 無料ハンズオンで、L4 で実行できるすべての AI ワークロードについて学ぶことができます。

始める準備はできましたか？

NVIDIA L4 Tensor コア GPU についての詳細は: **www.nvidia.com/ja-jp/data-center/l4**

NVIDIA