Arkusz danych

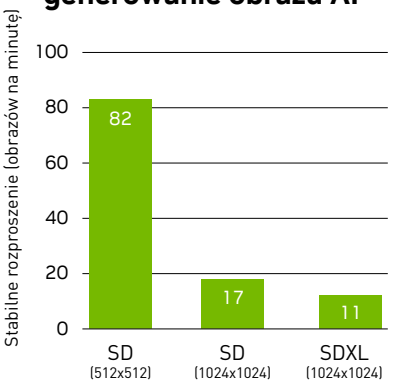
NVIDIA L40S

Niezrównana wydajność sztucznej inteligencji i grafiki dla centrum danych.

Generatywna sztuczna inteligencja napędza zmiany transformacyjne, otwierając nową granicę możliwości dla przedsiębiorstw ze wszystkich branż. Aby dokonać transformacji za pomocą sztucznej inteligencji, przedsiębiorstwa potrzebują większych zasobów obliczeniowych, większej skali i szerokiego zestawu możliwości, aby sprostać wymaganiom stale rosnącego zestawu różnorodnych i złożonych obciążeń.

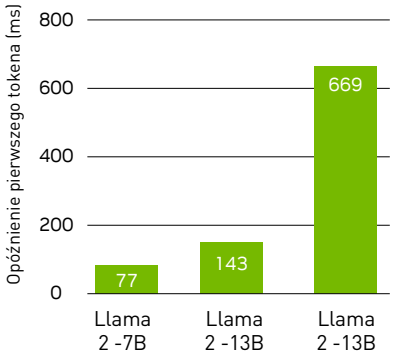
Procesor graficzny NVIDIA L40S to najpotężniejszy uniwersalny procesor graficzny dla centrów danych, zapewniający kompleksowe przyspieszenie dla aplikacji obsługujących sztuczną inteligencję nowej generacji – od gen AI, wnioskowania LLM, szkolenia i dostrajania małych modeli po grafikę 3D , renderowania i aplikacji wideo.

Generatywne generowanie obrazu AI



Measured performance; NVIDIA L40S Stable Diffusion v2.1, TRT 8.6.1, BS:1, FP16 | Stable Diffusion XL 1.0, TRT 8.6.1, BS:1, FP16

Wnioskowanie z modelu dużego języka (LLM).



Measured performance; NVIDIA L40S Llama 2-7B/13B/70B, ISL=2048, OSL=128, BS=1: FP8.

Przyspiesz obciążenia nowej generacji

NVIDIA AI Enterprise

- Generatywna sztuczna inteligencja
- Wnioskowanie LLM
- Dostrajanie LLM i szkolenie na małych modelach
- NVIDIA Omniverse™ Enterprise
- Renderowanie i grafika 3D
- Przesyłanie strumieniowe i treści wideo

Oparty na architekturze NVIDIA Ada Lovelace

Rdzenie tensorowe czwartej generacji

Obsługa sprzętu zapewniająca rzadkość strukturalną i zoptymalizowany format TF32 zapewnia natychmiastowy wzrost wydajności w celu szybszego szkolenia modeli sztucznej inteligencji i nauki o danych. Przyspiesz możliwości graficzne wspomagane sztuczną inteligencją za pomocą DLSS, aby zwiększyć rozdzielczość i lepszą wydajność w wybranych aplikacjach.

Rdzenie RT trzeciej generacji

Większa przepustowość oraz możliwości jednoczesnego śledzenia promieni i cieniowania poprawiają wydajność śledzenia promieni, przyspieszając renderowanie projektów produktów oraz procesów związanych z architekturą, inżynierią i konstrukcją. Zobacz realistyczne projekty w akcji dzięki przyspieszanemu sprzętowo rozmyciu ruchu i oszałamiającym animacjom w czasie rzeczywistym.

Transformer Engine

Transformer Engine radykalnie przyspiesza wydajność sztucznej inteligencji i poprawia wykorzystanie pamięci zarówno na potrzeby uczenia, jak i wnioskowania. Wykorzystując moc rdzeni Tensor czwartej generacji Ada Lovelace, Transformer Engine inteligentnie skanuje warstwy sieci neuronowych architektury transformatorowej i automatycznie przekształca precyzję między FP8 a FP16, aby zapewnić większą wydajność sztucznej inteligencji oraz przyspieszyć szkolenie i wnioskowanie.

Gotowe dla Centrum Danych

Procesor graficzny L40S jest zoptymalizowany do pracy w korporacyjnym centrum danych 24 godziny na dobę, 7 dni w tygodniu i został zaprojektowany, zbudowany, przetestowany i wspierany przez firmę NVIDIA, aby zapewnić maksymalną wydajność, trwałość i czas pracy. Procesor graficzny L40S spełnia najnowsze standardy dla centrów danych, jest zgodny z systemem budowania sprzętu sieciowego (NEBS) poziom 3 i oferuje bezpieczny rozruch z technologią root of trust, zapewniając dodatkową warstwę bezpieczeństwa dla centrów danych.

Specyfikacja techniczna	
Architektura GPU	NVIDIA Ada Lovelace Architecture
Pamięć GPU	48GB GDDR6 z ECC
Przepustowość pamięci	864GB/s
Interfejs połączenia	PCIe Gen4 x16: 64 GB/s dwukierunkowe
Rdzenie CUDA® oparte na architekturze NVIDIA Ada Lovelace	18,176
Rdzenie NVIDIA RT trzeciej generacji	142
Rdzenie Tensorowe NVIDIA czwartej generacji	568
TFLOPS wydajności rdzenia RT	209
FP32 TFLOPS	91.6
TF32 Tensor Core TFLOPS	183 I 366*
BFLOAT16 Tensor Core TFLOPS	362.05 I 733*
FP16 Tensor Core	362.05 I 733*
FP8 Tensor Core	733 I 1,466*
Szczytowe INT8 Tensor TOPS	733 I 1,466*
Szczytowe INT4 Tensor TOPS	733 I 1,466*
Typ obudowy	4.4” (H) x 10.5” (L), dual slot
Złącze monitora	4x DisplayPort 1.4a
Maksymalny pobór mocy	350W
Złącze zasilania	16-pin

Specyfikacja techniczna	
Rozwiązanie termiczne	Pasywne
Obsługa oprogramowania wirtualnego procesora graficznego (vGPU)	Tak
Obsługiwane profile vGPU	Zobacz przewodnik licencjonowania wirtualnego procesora graficznego
NVENC I NVDEC	3x I 3x (obejmuje kodowanie i dekodowanie AV1)
Bezpieczny rozruch z rootem zaufania	Tak
NEBS Ready	Level 3
Obsługa MIG	Nie
Obsługa NVIDIA® NVLink®	Nie

* Z rzadkością

Gotowy, aby zacząć?

Aby dowiedzieć się więcej o NVIDIA L40S, odwiedź stronę **www.nvidia.com/l40s**

© 2024 NVIDIA CORPORATION AND AFFILIATES. ALL RIGHTS RESERVED. NVIDIA, THE NVIDIA LOGO, CUDA, HGX, NVLINK, AND OMNIVERSE ARE TRADEMARKS AND/OR REGISTERED TRADEMARKS OF NVIDIA CORPORATION AND AFFILIATES IN THE U.S. AND OTHER COUNTRIES. OTHER COMPANY AND PRODUCT NAMES MAY BE TRADEMARKS OF THE RESPECTIVE OWNERS WITH WHICH THEY ARE ASSOCIATED. 3110647. FEB24

FORMAT | NVIDIA