Arkusz danych

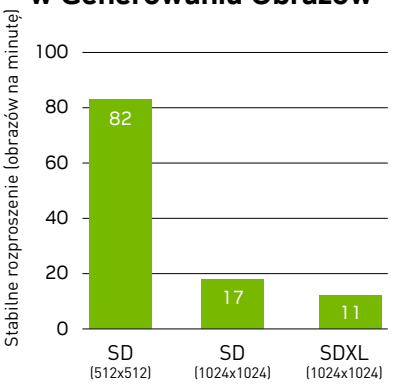
NVIDIA L40S

Bezprecedensowa wydajność AI i grafiki dla centrum danych.

Generatywna AI napędza transformacyjne zmiany, otwierając nowe możliwości dla przedsiębiorstw w każdej branży. Aby przekształcać się za pomocą AI, przedsiębiorstwa potrzebują większych zasobów obliczeniowych, większej skali oraz szerokiego zestawu możliwości, aby sprostać wymaganiom coraz bardziej różnorodnych i złożonych obciążeń.

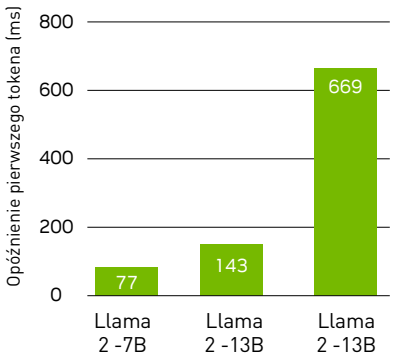
GPU NVIDIA L40S to najsilniejszy uniwersalny GPU dla centrum danych, zapewniający kompleksowe przyspieszenie dla nowej generacji aplikacji wspierających AI – od generatywnej AI, wnioskowania LLM, treningu i dostrajania małych modeli po grafikę 3D, renderowanie i aplikacje wideo.

Generatywna AI w Generowaniu Obrazów



Zmierzone Osiągi: 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 Dużych Modeli Językowych (LLM)



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

Napędzany architekturą NVIDIA Ada Lovelace

Rdzenie Tensor czwartej generacji

Sprzętowe wsparcie dla strukturalnej rzadkości oraz zoptymalizowany format TF32 zapewniają natychmiastowe zyski wydajności dla szybszego szkolenia modeli AI i nauki o danych. Przyspiesz możliwości grafiki wspieranej AI za pomocą DLSS, umożliwiając zwiększenie rozdzielczości z lepszą wydajnością w wybranych aplikacjach.

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

Specyfikacja

SPECYFIKACJE PRODUKTU

Całkowite zużycie energii	350 W domyślnie 350 W maksymalnie
Rozwiązanie termiczne	Pasywne
Mechaniczny format obudowy	Pełny profil, pełna długość (FHFL) 10,5", dwuslotowy
Taktowanie GPU	Bazowe: 1065 MHz Boost: 2520 MHz
Stany wydajności	P0, P8
VBIOS	Rozmiar pamięci EEPROM: 8 Mbit UEFI: Obsługiwane
Interfejs PCI Express	PCI Express Gen4 x16 Obsługiwane odwracanie linii i polaryzacji
Wieloinstancyjny GPU (MIG)	Nie obsługiwane
NVIDIA® NVLink®	Nie obsługiwane
Zero Power	Nie obsługiwane
Złącza	Jedno dodatkowe złącze zasilania PCIe 16-pin Cztery złącza VESA® DisplayPort®
Waga	Płyta: 1052 g (bez wspornika i przedłużaczy) Wspornik z wkrętami: 20 g Ulepszony prosty przedłużacz: 35 g Długi przedłużacz offsetowy: 48 g Prosty przedłużacz: 32 g

SPECYFIKACJE PAMIĘCI

Taktowanie pamięci	9001 MHz
Typ pamięci	GDDR6
Rozmiar pamięci	48 GB
Szerokość magistrali pamięci	384 bits
Maksymalna przepustowość pamięci	864 GB/s

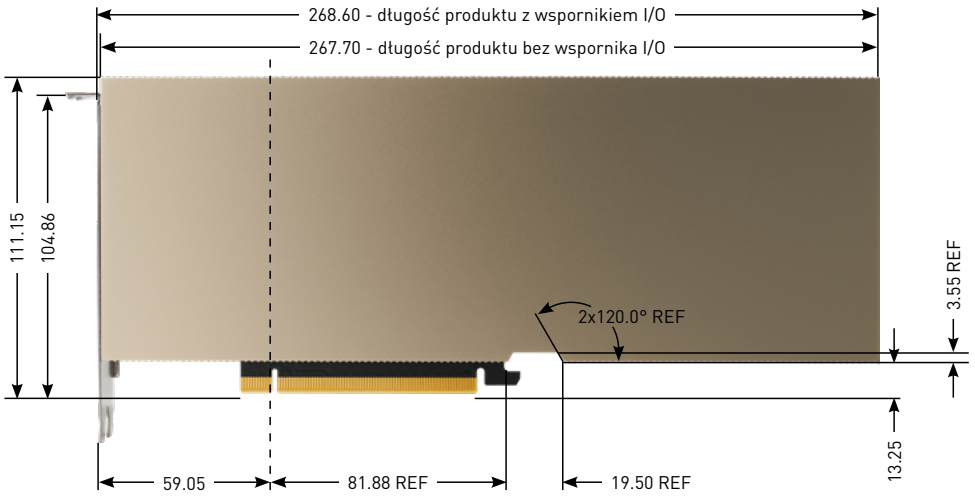
SPECYFIKACJE OPROGRAMOWANIA

Obsługa SR-IOV	Obsługiwane: 32 VF (funkcji wirtualnych)
Adres BAR (fizyczna funkcja)	BAR0: 16 MiB BAR1: 64 GiB (tryb wyłączonego wyświetlacza; domyślny) BAR1: 8 GiB (tryb włączonego wyświetlacza, 8 GB BAR1) BAR1: 256 MiB (tryb włączonego wyświetlacza, 256 MB BAR1) BAR3: 32 MiB
Adres BAR (funkcja wirtualna)	Tryb wyłączonego wyświetlacza (domyślny): • BAR0: 8 MiB (32 VF × 256 KiB) • BAR1: 64 GiB, 64-bit (32 VF × 2 GiB) • BAR3: 1 GiB, 64-bit (32 VF × 32 MiB) Tryby włączonego wyświetlacza: Rozmiary BAR VF nie mają zastosowania w trybach włączonego wyświetlacza.
Przerwania sygnalizowane poprzez wiadomość (MSI)	MSI-X: Obsługiwane MSI: Nieobsługiwane
Przekazywanie ARI	Obsługiwane
Wsparcie sterownika	Linux: R535TRD1 lub nowszy Windows: R535TRD1 lub nowszy
Uruchamianie zabezpieczeń	Obsługiwane
Firmware CEC	v2.0134 lub nowszy
NVFlash	Wersja 5.814 lub nowsza
Wsparcie dla NVIDIA® CUDA®	CUDA 12.2 lub nowsza
Wsparcie oprogramowania Virtual GPU	Obsługuje vGPU 16.1 (R535 GA6) lub nowszy: NVIDIA Virtual Compute Server Edition
Kod klasy PCI	0x03 – Kontroler wyświetlacza
Kod podklasy PCI	0x02 – Kontroler 3D
Wsparcie ECC	Włączone (domyślnie)
SMBus (adres 8-bitowy)	0x9E (zapis), 0x9F (odczyt)
Adres I2C pamięci EEPROM FRU IPMI	0x50 (7-bit), 0xA0 (8-bit)
Zarezerwowane adresy I2C	0xAA, 0xAC, 0xA0, 0x40
Bezpośredni dostęp do SMBus	Obsługiwany
Interfejs SMBPBI (SMBus Post-Box Interface)	Obsługiwany

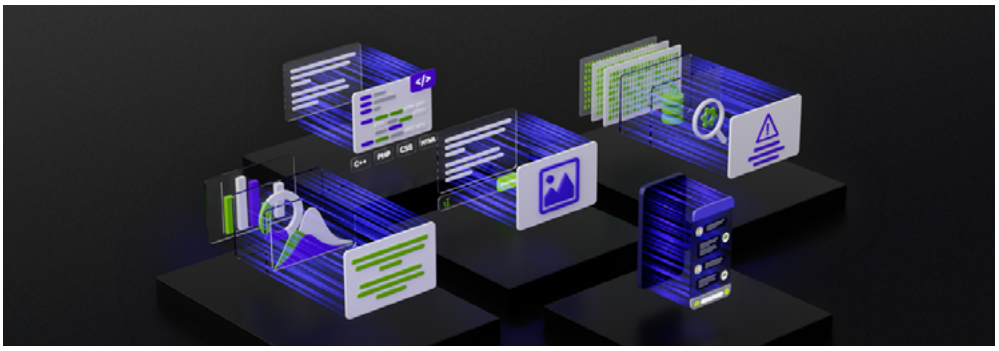
Uwaga:
[1]Notacja KiB, MiB i GiB podkreśla "potęgę dwóch" tych wartości.
Zatem,
- 256 KiB = 256 x 1024
- 16 MiB = 16 x 1024[2]
- 64 GiB = 64 x 1024[3]

Temperatura pracy otoczenia	0 °C do 50 °C
Temperatura pracy otoczenia (krótkoterminowa)[1]	-5 °C do 55 °C
Temperatura przechowywania	-40 °C do 75 °C
Wilgotność robocza (krótkoterminowa)[1]	5% do 93% wilgotności względnej
Wilgotność robocza	5% do 85% wilgotności względnej
Wilgotność przechowywania	5% do 95% wilgotności względnej
Średni czas między awariami (MTBF)	Nieuregulowane środowisko:[2] 2 502 369 godzin w temperaturze 35 °C Kontrolowane środowisko:[3] 3 270 359 godzin w temperaturze 35 °C

Notatki:
Specyfikacje w tej tabeli dotyczą wysokości do 6000 stóp.
[1] Okres nie dłuższy niż 96 godzin ciągłych, nie więcej niż 15 dni w roku.
[2] Pewne obciążenie środowiskowe z ograniczoną konserwacją (GF35).
[3] Brak obciążenia środowiskowego z optymalną eksploatacją i konserwacją (GB35).

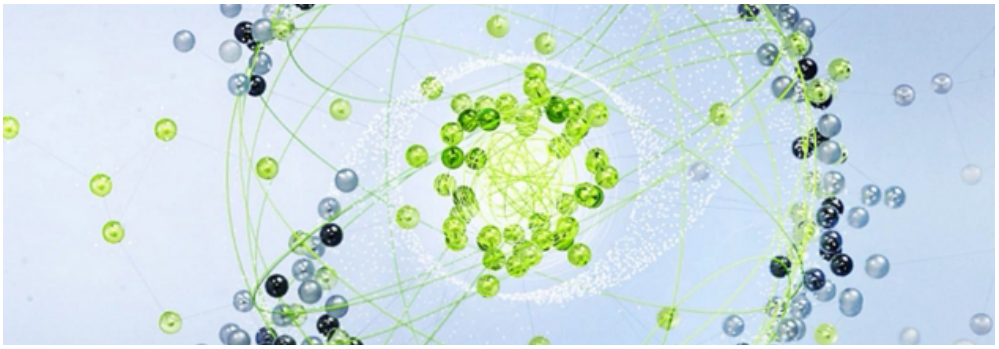


Generatywna AI i duże modele językowe (LLM)



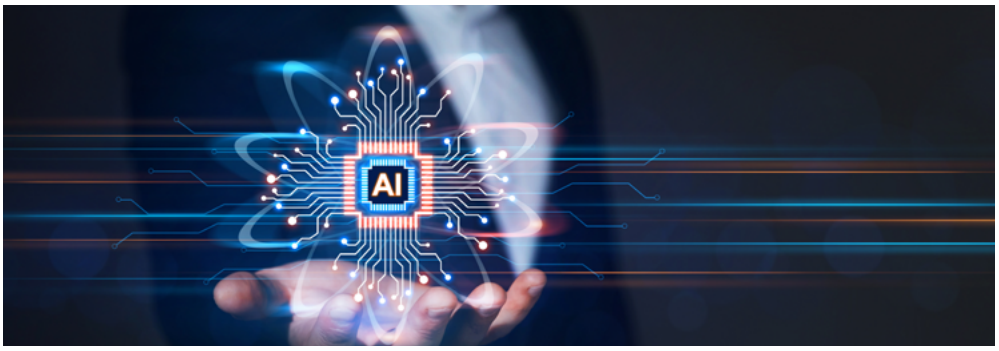
The NVIDIA L40S GPU is a powerful solution for NVIDIA Omniverse and 3D content creation, offering exceptional performance and versatility in data center environments. Built on the Ada Lovelace architecture, it features third-generation RT cores for enhanced real-time ray tracing and fourth-generation Tensor Cores that support AI-driven features, significantly improving the quality and speed of 3D workflows. As the engine of NVIDIA Omniverse in the data center, the L40S delivers stunning real-time ray tracing and AI-accelerated capabilities, making it ideal for extended reality (XR) and virtual production tasks. With 48GB of GDDR6 memory, it can handle complex 3D models, high-resolution textures, and large-scale simulations with ease, enabling creative professionals to work on intricate designs and render photorealistic scenes more efficiently. The L40S's support for Universal Scene Description (OpenUSD)-based 3D workflows within the Omniverse ecosystem enhances collaboration and streamlines production pipelines. Its performance in Omniverse applications is described as „stunning," positioning it as a top-tier solution for organizations looking to leverage cutting-edge technologies in virtual world creation, 3D visualization, and immersive content production.

NVIDIA Omniverse i tworzenie treści 3D



Procesor graficzny NVIDIA L40S to potężny procesor NVIDIA Omniverse i tworzenia treści 3D, oferujący wyjątkową wydajność i wszechstronność w środowiskach centrów danych. Zbudowany w oparciu o architekturę Ada Lovelace, zawiera rdzenie RT trzeciej generacji zapewniające ulepszone śledzenie promieni w czasie rzeczywistym oraz rdzenie Tensor czwartej generacji, które obsługują funkcje oparte na sztucznej inteligencji, znacznie poprawiając jakość i szybkość przepływów pracy 3D. Jako silnik NVIDIA Omniverse w centrum danych, L40S zapewnia oszałamiające możliwości śledzenia promieni w czasie rzeczywistym i akceleracji sztucznej inteligencji, dzięki czemu idealnie nadaje się do zadań w rozszerzonej rzeczywistości (XR) i wirtualnej produkcji. Dzięki 48 GB pamięci GDDR6 z łatwością radzi sobie ze złożonymi modelami 3D, teksturami o wysokiej rozdzielczości i symulacjami na dużą skalę, umożliwiając kreatywnym profesjonalistom pracę nad skomplikowanymi projektami i wydajniejsze renderowanie fotorealistycznych scen. Obsługa przez monitor L40S procesów 3D opartych na uniwersalnym opisie scen (OpenUSD) w ekosystemie Omniverse usprawnia współpracę i usprawnia procesy produkcyjne. Jego wydajność w aplikacjach Omniverse określa się jako „oszałamiającą", co stawia go jako najwyższej klasy rozwiązanie dla organizacji chcących wykorzystać najnowocześniejsze technologie do tworzenia wirtualnego świata, wizualizacji 3D i produkcji treści immersyjnych.

Szkolenie i wnioskowanie AI (sztucznej inteligencji)



Procesor graficzny NVIDIA L40S to potężne rozwiązanie do obciążeń związanych ze szkoleniem i wnioskowaniem AI, oferujące wyjątkową wydajność i wszechstronność w środowiskach centrów danych. Zbudowany na architekturze Ada Lovelace, zawiera 18 176 rdzeni CUDA i 568 rdzeni Tensor czwartej generacji, zapewniając do 5 razy lepszą wydajność zmiennoprzecinkową pojedynczej precyzji (FP32) w porównaniu do A100. Jego zaawansowany silnik transformatorowy inteligentnie zarządza precyzją między FP8 a FP16, znacznie zwiększając wydajność sztucznej inteligencji zarówno na potrzeby uczenia, jak i wnioskowania modeli opartych na transformatorach. Dzięki 48 GB pamięci GDDR6 L40S może skutecznie obsługiwać złożone zadania AI i modele z dużymi językami. W przypadku szkolenia AI osiem procesorów graficznych L40S w głównym serwerze pozwala na 0,8-krotny wzrost wydajności w porównaniu z systemem 8-GPU A100 dla modeli MLPerf. W zadaniach wnioskowania L40S wykazuje imponujące możliwości, często dorównujące lub przekraczające wydajność A100 w różnych testach MLPerf. To sprawia, że L40S szczególnie dobrze nadaje się do wdrażania i uruchamiania wyrafinowanych modeli sztucznej inteligencji w środowiskach produkcyjnych, oferując organizacjom wydajne i wydajne rozwiązanie dla ich obciążeń związanych ze sztuczną inteligencją.

Grafika i wizualizacja

Procesor graficzny NVIDIA L40S oferuje wyjątkowe możliwości w zakresie obciążeń graficznych i wizualizacyjnych, co czyni go potężnym rozwiązaniem do profesjonalnych zastosowań w takich dziedzinach, jak projektowanie wspomagane komputerowo (CAD), produkcja wirtualna i wizualizacja naukowa. Zbudowany na architekturze Ada Lovelace, zawiera rdzenie RT trzeciej generacji, które znacznie zwiększają wydajność śledzenia promieni w czasie rzeczywistym, zapewniając oszałamiającą wierność wizualną i fotorealistyczne renderowanie. 48 GB pamięci GDDR6 modelu L40S pozwala z łatwością obsługiwać złożone modele 3D, tekstury o wysokiej rozdzielczości i duże zbiory danych, umożliwiając profesjonalistom pracę nad skomplikowanymi projektami i wizualizacjami bez wąskich gardeł wydajności. Rdzenie Tensor czwartej generacji obsługują funkcje graficzne wzmocnione sztuczną inteligencją, takie jak DLSS (Deep Learning Super Sampling), które mogą zwiększyć wydajność i jakość obrazu w obsługiwanych aplikacjach. W połączeniu z oprogramowaniem NVIDIA RTX Virtual Workstation (vWS), L40S może zasilać wirtualne stacje robocze o wysokiej wydajności z centrum danych, zapewniając elastyczny dostęp do wymagających aplikacji graficznych z dowolnego urządzenia. To sprawia, że NVIDIA L40S jest doskonałym wyborem dla organizacji, które chcą ulepszyć swoje możliwości wizualizacji, poprawić efektywność przepływu pracy i dostarczać wysokiej jakości treści wizualne w różnych branżach.

Kodowanie wideo i strumieniowanie

Procesor graficzny NVIDIA L40S oferuje wyjątkowe możliwości w zakresie kodowania wideo i przesyłania strumieniowego, co czyni go potężnym rozwiązaniem do strumieniowego przesyłania transmisji, produkcji wideo i transkrypcji. Zbudowany na architekturze Ada Lovelace, L40S posiada trzy silniki kodowania i dekodowania wideo, co znacznie zwiększa jego zdolność do jednoczesnej obsługi wielu strumieni wideo wysokiej jakości. Kluczowym postępem jest dodanie obsługi kodowania i dekodowania AV1, która zapewnia przełomową wydajność i niższy całkowity koszt posiadania dla twórców treści i platform przesyłania strumieniowego. Ta funkcja pozwala na uzyskanie wyższej jakości wideo przy niższych przepływnościach, z korzyścią zarówno dla dostawców treści, jak i użytkowników końcowych. L40S może obsługiwać ponad 1000 jednoczesnych strumieni wideo AV1 w rozdzielczości 720p30 do zastosowań mobilnych, co czyni go idealnym rozwiązaniem dla usług przesyłania strumieniowego i sieci dostarczania treści. Potężne przyspieszenie sprzętowe w połączeniu z 48 GB pamięci GDDR6 umożliwia wydajne przetwarzanie złożonych obciążeń wideo, w tym transkodowanie w czasie rzeczywistym i tworzenie treści w wysokiej rozdzielczości. Organizacjom zajmującym się transmisją strumieniową na żywo, usługami wideo na żądanie lub produkcją wirtualną NVIDIA L40S zapewnia wydajność i wszechstronność niezbędną do spełnienia wymagań współczesnego tworzenia i dystrybucji treści wideo.

Rdzenie RT trzeciej generacji

Zwiększona przepustowość oraz jednoczesne możliwości śledzenia promieni i cieniowania poprawiają wydajność śledzenia promieni, przyspieszając renderowanie dla projektów produktów, architektury oraz przepływów pracy w inżynierii i budownictwie. Zobacz realistyczne projekty w akcji dzięki sprzętowemu przyspieszeniu rozmycia ruchu i oszałamiającym animacjom w czasie rzeczywistym.

Silnik Transformatorów

Silnik Transformatorów znacznie przyspiesza wydajność AI i poprawia wykorzystanie pamięci zarówno podczas treningu, jak i wnioskowania. Wykorzystując moc rdzeni Tensor czwartej generacji architektury Ada Lovelace, Silnik Transformatorów inteligentnie skanuje warstwy architektury sieci neuronowych transformatorów i automatycznie przekształca je między precyzjami FP8 a FP16, aby zapewnić szybszą wydajność AI i przyspieszyć trening oraz wnioskowanie.

Gotowy do Centrum Danych

GPU L40S jest zoptymalizowany do pracy w centrum danych przedsiębiorstw 24/7 i zaprojektowany, zbudowany, przetestowany oraz wspierany przez NVIDIĘ, aby zapewnić maksymalną wydajność, trwałość i dostępność. GPU L40S spełnia najnowsze standardy centrum danych, jest zgodny z poziomem NEBS Level 3 i oferuje bezpieczny rozruch z technologią podstawy zaufania, co zapewnia dodatkową warstwę bezpieczeństwa dla centrów danych.

Gotowy, aby zacząć?

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

FORMAT | NVIDIA