Arkusz danych

NVIDIA L4 Tensor Core GPU

Przełomowy uniwersalny akcelerator do efektywnego przetwarzania wideo, AI i grafiki.

Przyspiesz prace związane z wideo, AI i grafiką

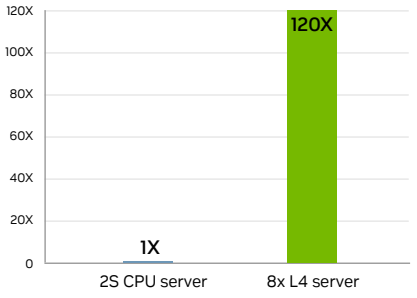
Procesor graficzny NVIDIA Ada Lovelace L4 Tensor Core zapewnia uniwersalne przyspieszenie i efektywność energetyczną dla aplikacji wideo, sztucznej inteligencji, wirtualnych komputerów stacjonarnych i aplikacji graficznych w przedsiębiorstwie, w chmurze i na krawędzi. Dzięki platformie AI firmy NVIDIA i podejściu typu full-stack, L4 jest zoptymalizowany pod kątem wnioskowania na dużą skalę dla szerokiego zakresu aplikacji AI, w tym rekomendacji, głosowych asystentów awatarów AI, generatywnej sztucznej inteligencji, wyszukiwania wizualnego i automatyzacji centrów kontaktowych, aby zapewnić najlepiej spersonalizowane doświadczenie.

Jako najbardziej wydajny akcelerator NVIDIA do głównego użytku, serwery wyposażone w L4 zapewniają do 120 razy wyższą wydajność wideo AI i 2,7 razy większą wydajność generatywnej sztucznej inteligencji w porównaniu z rozwiązaniami CPU, a także ponad 4 razy większą wydajność graficzną niż poprzednia generacja procesorów graficznych. Wszechstronność i energooszczędna, jednoslotowa, niskoprofilowa obudowa NVIDIA L4 sprawiają, że idealnie nadaje się ona do wdrożeń na całym świecie, w tym w lokalizacjach brzegowych.

Dane techniczne	
FP32	30.3 teraFLOPs
TF32 Tensor Core	120 teraFLOPS*
FP16 Tensor Core	242 teraFLOPS*
BFLOAT16 Tensor Core	242 teraFLOPS*
FP8 Tensor Core	485 teraFLOPs*
INT8 Tensor Core	485 TOPs*
Pamięć GPU	24GB
Przepustowość pamięci GPU	300 GB/s
NVENC | NVDEC | Dekodery JPEG	2 | 4 | 4
Maksymalna moc obliczeniowa cieplna (TDP)	72W
kształt obudowy	1-slot low-profile, PCIe
Złącze	PCIe Gen4 x16 64GB/s
Opcje serwera	Partner and NVIDIA-Certified Systems with 1–8 GPUs

* Shown with sparsity. Specifications 1/2 lower without sparsity.

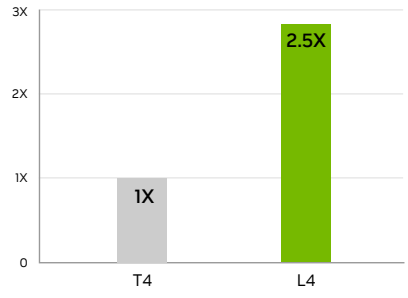
Do 120 x wyższa wydajność wideo AI



Wydajność wideo L4 AI

Zmierzone osiągi: porównanie 8x L4 z serwerem 2S Intel 8362 CPU, obejmujące końcową drogę przetwarzania wideo z dekodowaniem CV-CUDA®, wstępnym przetwarzaniem, wnioskowaniem (SegFormer), przetwarzaniem końcowym oraz kodowaniem, NVIDIA® TensorRT™ 8.6 w porównaniu z pipeline'em wyłącznie na CPU korzystającym z OpenCV 4.7 i wnioskowaniem PyTorch.

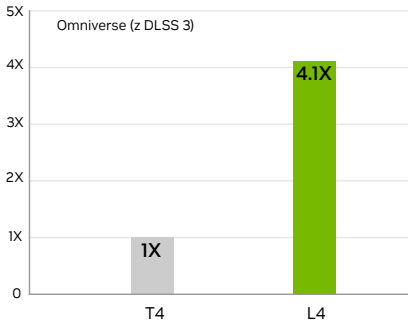
2,5 x wyższa wydajność generatywnej AI



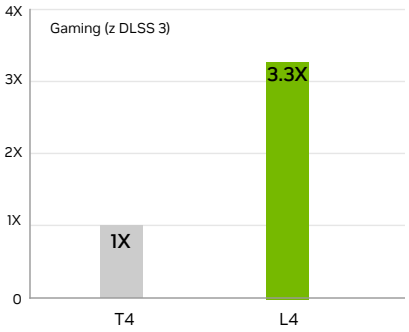
L4 Generacyjna wydajność AI

Zmierzone osiągi: L4 vs T4 w generowaniu obrazów, 512x512 stabilna dyfuzja v2.1, FP16, TensorRT 8.5.2.

Ponad 4 x wyższa wydajność renderowania w czasie rzeczywistym



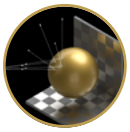
Ponad 3 x wyższa wydajność śledzenia promieni



Wizualna wydajność obliczeniowa L4

Zmierzone osiągi:
Renderowanie w czasie rzeczywistym:** Wydajność NVIDIA Omniverse™ w renderowaniu w czasie rzeczywistym w rozdzielczości 1080p i 4K z NVIDIA Deep Learning Super Sampling (DLSS) 3.
Śledzenie promieni:** Średnia wydajność gier dla tytułów AAA wspierających śledzenie promieni i DLSS 3.

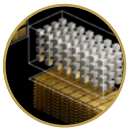
Odkryj przełomy architektury NVIDIA Ada Lovelace



Rdzenie RT trzeciej generacji

NVIDIA uczyniła śledzenie promieni w czasie rzeczywistym rzeczywistością dzięki wynalezieniu rdzeni RT, które są rdzeniami obliczeniowymi na GPU, zaprojektowanymi specjalnie do radzenia sobie z obliczeniami intensywnymi w zakresie renderowania śledzenia promieni.

Trzecia generacja rdzeni RT w architekturze Ada Lovelace ma dwukrotnie większą przepustowość przecięcia promieni-trójkąt, co zwiększa wydajność RT-TFLOP o ponad 2 razy. Technologia NVIDIA Shader Execution Reordering (SER) poprawia wydajność o ponad 3 razy, umożliwiając głębokie immersyjne doświadczenia w wirtualnych światach oraz niespotykaną dotąd wydajność w zakresie graficznych aplikacji AI i gier w chmurze.



Rdzenie Tensor czwartej generacji

Rdzenie Tensor w architekturze Ada Lovelace są zaprojektowane w celu przyspieszania transformacyjnych technologii AI, takich jak inteligentne chatboty, generatywna AI, przetwarzanie języka naturalnego (NLP), wizja komputerowa oraz NVIDIA DLSS 3. Rdzenie Tensor Ada Lovelace uruchamiają rzadkość strukturalną i precyzję zmiennoprzecinkową 8-bitową (FP8), co pozwala na osiągnięcie do 4 razy wyższej wydajności wnioskowania w porównaniu do poprzedniej generacji. FP8 zmniejsza obciążenie pamięci w porównaniu do większych precyzji i znacznie przyspiesza przepustowość AI.

Zaawansowane przyspieszenie wideo i AI wizji

Dzięki zoptymalizowanej stosowanej technologii AV1, NVIDIA L4 podnosi przyspieszenie wideo i AI wizji na wyższy poziom, tworząc szeroki wachlarz nowych możliwości zastosowań, takich jak transkodowanie wideo w czasie rzeczywistym, strumieniowanie, wideokonferencje, rozszerzona rzeczywistość (AR), wirtualna rzeczywistość (VR) oraz AI wizji.

Z czterema dekoderami wideo i dwoma enkoderami wideo, w połączeniu z formatem wideo AV1, serwery L4 mogą obsługiwać ponad 10002 jednoczesnych strumieni wideo i oferować ponad 120 razy większą wydajność end-to-end w przetwarzaniu wideo AI w porównaniu do rozwiązań CPU.3 Dodatkowo, cztery dekodery JPEG przyspieszają aplikacje wymagające mocy obliczeniowej wizji komputerowej.

Deep Learning Super Sampling (DLSS)

NVIDIA DLSS 3 to rewolucyjny przełom w grafice zasilanej AI, który znacznie poprawia wydajność renderowania. Napędzany nowymi rdzeniami Tensor czwartej generacji oraz akceleratorem NVIDIA Optical Flow (OFA) w L4, DLSS 3 wykorzystuje AI do tworzenia dodatkowych wysokiej jakości klatek dla obciążeń związanych z grafiką.

Gotowy do Wirtualizacji

Dzięki ulepszeniom nowej generacji w oprogramowaniu NVIDIA wirtualnego GPU (vGPU) oraz 1,5 razy większej pamięci GPU niż w poprzedniej generacji, L4 zwiększa wydajność stacji roboczych o 1,7 razy dla średnio- i wysokowydajnych przepływów pracy projektowych działających na NVIDIA RTX™ Virtual Workstation (vWS) oraz przyspiesza aplikacje produktywności działające na NVIDIA Virtual PC (vPC).

Wydajność i bezpieczeństwo Centrum Danych

NVIDIA L4 jest zoptymalizowana do całodobowych operacji w centrach danych przedsiębiorstw i jest zaprojektowana, zbudowana, dokładnie testowana i wspierana przez NVIDIĘ oraz partnerów w celu uzyskania maksymalnej wydajności, trwałości i bezpieczeństwa. L4 oferuje bezpieczny rozruch z technologią podstawy zaufania, co zapewnia dodatkową warstwę bezpieczeństwa dla centrów danych.

1. FP8 L4 w porównaniu do FP16 T4.

2. 8x L4 z kodowaniem AV1 w niskiej latencji przy ustawieniach P1 w 720p30.

3. Porównanie wydajności 8x L4 z serwerem 2S Intel 8362 CPU: końcowa droga przetwarzania wideo z pre- i postprzetwarzaniem CV-CUDA, dekodowaniem, wnioskowaniem (SegFormer), kodowaniem, TRT 8.6 w porównaniu do pipeline'u wyłącznie na CPU korzystającego z OpenCV.

Specyfikacja

SPECYFIKACJE PRODUKTU

Całkowite zużycie energii	72 W domyślnie 72 W maksymalnie 40 W minimalnie
Rozwiązanie termiczne	Pasywne
Mechaniczny format obudowy	HHHL-SS (niski profil, połowa długości, jednoslotowy)
Taktowanie GPU	Bazowe: 795 MHz Boost: 2040 MHz
VBIOS	Rozmiar pamięci EEPROM: 16 Mbit UEFI: Obsługiwane
Sterowniki	Linux: R525 lub nowszy Windows: R525 lub nowszy
Interfejs PCI Express	Fizyczne 16 linii PCIe PCIe Gen4 x16, x8; Gen3 x16 Obsługiwane odwracanie linii i polaryzacji
Zero Power	Nie obsługiwane
Stany wydajności	P0, P8
Waga	Płyta: 270 g (bez wspornika) Wspornik (pełny profil) z wkrętami: 14 g Wspornik (niski profil) z wkrętami: 9 g

SPECYFIKACJE PAMIĘCI

Taktowanie pamięci	6251 MHz
Typ pamięci	GDDR6
Rozmiar pamięci	24 GB
Szerokość magistrali pamięci	192 bits
Maksymalna przepustowość pamięci	300 GB/sec

SPECYFIKACJE OPROGRAMOWANIA

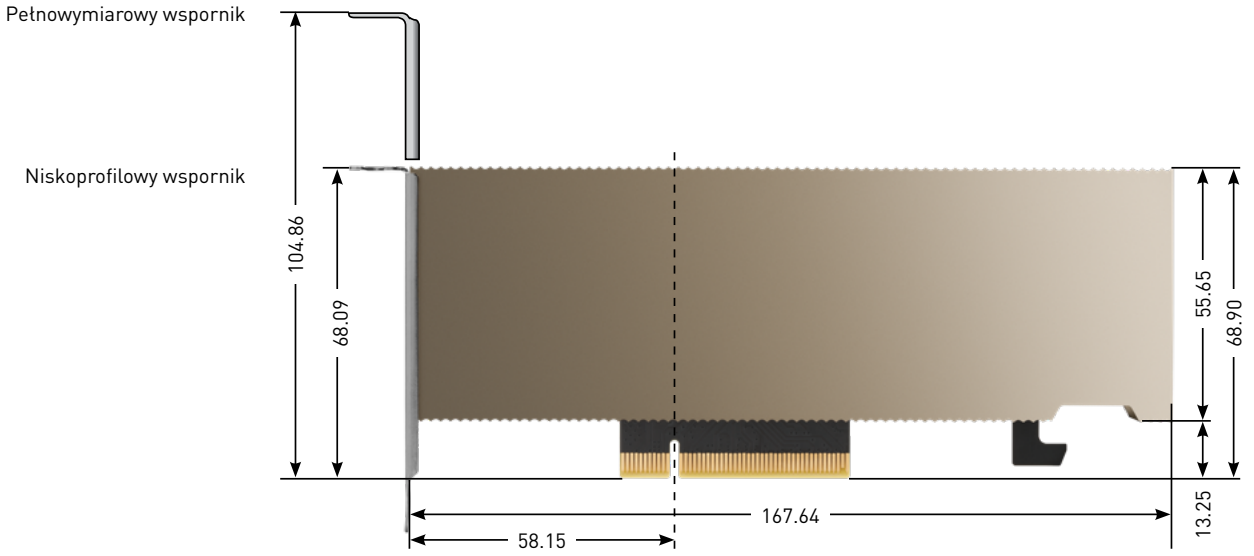
Obsługa SR-IOV	Obsługiwane: 32 VF (funkcji wirtualnych)
Adres BAR (fizyczna funkcja)	BAR0: 16 MiB[1] BAR1: 32 GiB[1] BAR3: 32 MiB[1]
Adres BAR (funkcja wirtualna)	BAR0: 8 MiB (256 KiB na VF)[1] BAR1: 64 GiB, 64-bit (2 GiB na VF)[1] BAR3: 1 GiB, 64-bit (32 MiB na VF)[1]
Przerwania sygnalizowane poprzez wiadomość (MSI)	MSI-X: Obsługiwane MSI: Nieobsługiwane
Multi-Instance GPU (MIG)	Nie obsługiwane
Przekazywanie ARI	Obsługiwane
Uruchamianie zabezpieczeń	Obsługiwane
Wsparcie dla NVIDIA® CUDA®	CUDA 12.0 lub nowsze
Wsparcie oprogramowania Virtual GPU	Obsługuje vGPU 15.2 lub nowszy
Wsparcie ECC	Włączone (domyślnie); można wyłączyć za pomocą oprogramowania
Kod klasy PCI	0x03 – Kontroler wyświetlacza
Kod podklasy PCI	0x02 – Kontroler 3D
SMBus (adres 8-bitowy)	0x9E (zapis), 0x9F (odczyt)
Adres I2C pamięci EEPROM FRU IPMI	0x50 (7-bit), 0xA0 (8-bit)
Bezpośredni dostęp do SMBus	Obsługiwany
Zarezerwowane adresy I2C	0xA0, 0xAA, 0xAC
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]

ŚRODOWISKOWE I NIEZAWODNOŚCIOWE SPECYFIKACJE

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).



Wnioskowanie AI

Procesor graficzny NVIDIA L4 Tensor Core to wydajne i wszechstronne rozwiązanie do obciążeń wnioskowania AI, oferujące znaczną poprawę wydajności w porównaniu do swojego poprzednika, T4. Zbudowany w oparciu o architekturę NVIDIA Ada Lovelace, L4 zawiera rdzenie Tensor czwartej generacji i rdzenie RT trzeciej generacji, dzięki czemu doskonale nadaje się do szerokiej gamy zastosowań AI. Dzięki 24 GB pamięci GDDR6 i energooszczędnej obudowie o mocy 72 W, L4 zapewnia do 2,7 razy większą wydajność generatywną AI niż poprzednia generacja. Doskonale sprawdza się w zadaniach wnioskowania AI w różnych dziedzinach, w tym w przetwarzaniu obrazu komputerowego, przetwarzaniu języka naturalnego i systemach rekomendacji. Silniki przetwarzania obrazu i wideo z akceleracją sprzętową, w tym możliwości kodowania/dekodowania AV1, sprawiają, że L4 jest szczególnie skuteczny w przypadku analiz wideo i transkodowania wykorzystujących sztuczną inteligencję. Jego jednogniazdowa, niskoprofilowa obudowa umożliwia łatwą integrację z głównymi serwerami, co czyni go idealnym wyborem dla organizacji chcących wdrożyć wnioskowanie AI na dużą skalę w centrach danych lub środowiskach przetwarzania brzegowego. Wszechstronność i wydajność L4 pozycjonuje go jako uniwersalny akcelerator wnioskowania AI, zdolny do obsługi różnorodnych obciążeń, od przesyłania strumieniowego wideo po odkrywanie leków.

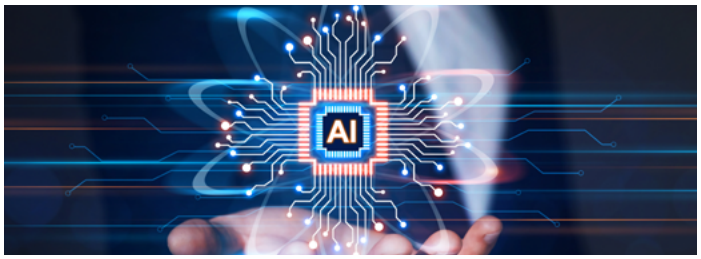
Obliczenia brzegowe i głównego nurtu

Procesor graficzny NVIDIA A40 to potężne rozwiązanie dla zaawansowanych aplikacji renderujących, wykorzystujące zaawansowaną architekturę Ampere, aby zapewnić wyjątkową wydajność i efektywność. Zaprojektowany specjalnie z myślą o wymagających obciążeniach związanych z renderowaniem w branżach takich jak media i rozrywka, architektura i projektowanie motoryzacyjne, A40 oferuje solidną gamę rdzeni CUDA i rdzeni Tensor. Taka konfiguracja umożliwia obsługę złożonych zadań renderowania 3D, śledzenia promieni w czasie rzeczywistym i grafiki wspomaganej sztuczną inteligencją z niezwykłą szybkością i precyzją. A40 obsługuje technologię RTX firmy NVIDIA, umożliwiając fotorealistyczne renderowanie i symulację oświetlenia, cieni i odbić w czasie rzeczywistym, usprawniając twórczy przepływ pracy i skracając czas wprowadzania produktów na rynek dla twórców i projektantów treści cyfrowych. Wysoka przepustowość pamięci zapewnia płynną obsługę dużych zbiorów danych i skomplikowanych szczegółów wizualnych, a zgodność z profesjonalnymi narzędziami programowymi firmy NVIDIA, takimi jak RTX Renderer i Omniverse, upraszcza integrację z istniejącymi potokami. Ogólnie rzecz biorąc, procesor graficzny NVIDIA A40 na nowo definiuje możliwości renderowania najwyższej klasy, oferując niezrównaną wydajność i wierność, dzięki czemu profesjonaliści mogą tworzyć oszałamiające wrażenia wizualne i przesuwać granice tworzenia treści cyfrowych.

Grafika i wizualizacja

Procesor graficzny NVIDIA L4 Tensor Core oferuje znaczący postęp w zakresie obciążeń graficznych i wizualizacyjnych, zapewniając ponad 4 razy wyższą wydajność w porównaniu do swojego poprzednika, T4. Zbudowany na architekturze Ada Lovelace, L4 wyposażony jest w rdzenie RT trzeciej generacji i technologię DLSS 3 opartą na sztucznej inteligencji, dzięki czemu może obsługiwać wymagające zadania, takie jak awatary oparte na sztucznej inteligencji, wirtualne światy NVIDIA Omniverse, gry w chmurze i wirtualne stacje robocze. Możliwości te pozwalają twórcom tworzyć w czasie rzeczywistym grafikę o kinowej jakości i niezwykle szczegółowe sceny, zapewniając wciągające wrażenia wizualne, które wcześniej były nieosiągalne w przypadku procesorów. Wszechstronność L4 rozciąga się na profesjonalne zastosowania wizualizacyjne, w tym projektowanie wspomagane komputerowo (CAD) i inżynierię wspomaganą komputerowo (CAE), co czyni go doskonałym wyborem dla projektantów i inżynierów. Dzięki energooszczędnej obudowie o mocy 72 W i niskoprofilowej obudowie z jednym gniazdem, L4 można łatwo zintegrować z głównymi serwerami, umożliwiając organizacjom wdrażanie wydajnych funkcji graficznych i wizualizacyjnych w centrach danych, lokalizacjach brzegowych i środowiskach chmurowych.

Generatywna AI



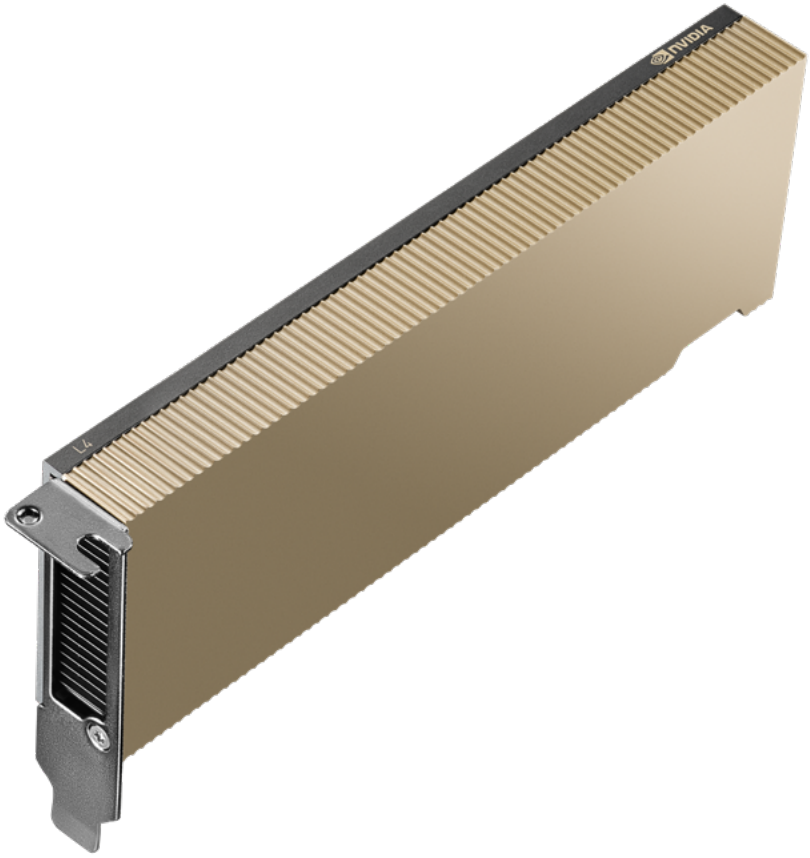
Procesor graficzny NVIDIA L4 Tensor Core oferuje znaczące postępy w zakresie generatywnych obciążeń AI, zapewniając do 2,7 razy wyższą wydajność w porównaniu do swojego poprzednika, NVIDIA T4. Zbudowany na architekturze Ada Lovelace, L4 jest wyposażony w rdzenie Tensor czwartej generacji i 24 GB pamięci GDDR6, dzięki czemu może obsługiwać większe i bardziej złożone modele generatywnej sztucznej inteligencji. Ta zwiększona pojemność pamięci pozwala na generowanie obrazu do rozdzielczości 1024x768, co nie było możliwe w przypadku procesora graficznego T4. Wszechstronność L4 sprawia, że doskonale nadaje się do szerokiej gamy generatywnych zastosowań sztucznej inteligencji, w tym generowania tekstu na obraz, awatarów zasilanych sztuczną inteligencją i zadań związanych z przetwarzaniem języka naturalnego 3 . Jego energooszczędna konstrukcja, działająca w zakresie mocy 72 W, czyni go atrakcyjną opcją do wdrożeń na dużą skalę w centrach danych i środowiskach przetwarzania brzegowego. Połączenie wydajności, wydajności i wszechstronności L4 sprawia, że jest to potężne rozwiązanie dla organizacji, które chcą przyspieszyć swoje przepływy pracy związane z generatywną sztuczną inteligencją, przy jednoczesnym zachowaniu opłacalności i zrównoważonego rozwoju.

Przyspieszanie obciążeń efektywnie i zrównoważenie

NVIDIA L4 jest integralną częścią platformy centrum danych NVIDIA. Stworzona z myślą o AI, wideo, wirtualnych stacjach roboczych, grafice, symulacjach, nauce o danych oraz analizie danych, platforma przyspiesza ponad 3,000 aplikacji i jest dostępna wszędzie w skali, od centrum danych po brzegi sieci i chmurę, dostarczając zarówno znaczące zyski wydajnościowe, jak i możliwości zwiększenia efektywności energetycznej.

W miarę jak AI i wideo stają się coraz bardziej powszechne, zapotrzebowanie na wydajne i opłacalne obliczenia rośnie jak nigdy dotąd. GPU Tensor Core NVIDIA L4 oferują do 120 razy lepszą wydajność wideo AI, co przekłada się na do 99 procent lepszą efektywność energetyczną oraz niższy całkowity koszt posiadania w porównaniu do tradycyjnej infrastruktury opartej na procesorach CPU. Pozwala to przedsiębiorstwom na zmniejszenie potrzebnej przestrzeni rackowej oraz znaczne obniżenie ich śladu węglowego, jednocześnie umożliwiając skalowanie centrów danych dla znacznie większej liczby użytkowników.

Energia zaoszczędzona przez przejście z CPU na NVIDIA L4 w centrum danych o mocy 2 megawatów (MW) może zasilać ponad 2,000 domów przez rok lub odpowiadać kompensacji węglowej 172,000 drzew rosnących przez 10 lat.[4,5]



[4] Porównanie serwerów z procesorem 8x L4 i 2S Intel 8362 CPU: kompleksowy potok wideo z przetwarzaniem wstępnym i końcowym CV-CUDA, dekodowaniem, wnioskowaniem (SegFormer), kodowaniem, TRT 8.6 w porównaniu z potokiem wykorzystującym wyłącznie procesor przy użyciu OpenCV 4.7, PyT wnioskowanie.

[5] Wyniki z kalkulatora EPA przy zastosowaniu oszczędności 1,677 MW. www.epa.gov/energy/greenhouse-gas-equivalency-calculator

Gotowe dla przedsiębiorstw: Oprogramowanie AI ułatwia rozwój i wdrożenie

Adopcja AI w przedsiębiorstwach stała się już powszechna, a organizacje potrzebują infrastruktury gotowej do AI, która zapewni im przyszłość w tej nowej erze. NVIDIA AI Enterprise to kompleksowy, natywny w chmurze zestaw oprogramowania do AI i analizy danych, zoptymalizowany, aby pomóc każdej organizacji odnosić sukcesy w AI i certyfikowany do wdrożenia wszędzie, od centrum danych przedsiębiorstwa po chmurę. Zestaw ten oferuje globalne wsparcie dla przedsiębiorstw, aby zapewnić, że projekty AI pozostaną na właściwej ścieżce.

NVIDIA AI Enterprise jest zoptymalizowane w celu uproszczenia rozwoju i wdrażania AI, a jego składniki obejmują sprawdzone kontenery i frameworki oparte na otwartym kodzie, które są certyfikowane do działania na popularnych platformach centrum danych oraz standardowych systemach certyfikowanych przez NVIDIĘ™ z GPU Tensor Core L4. Ponieważ wsparcie jest wliczone, organizacje uzyskują transparentność otwartego źródła oraz pewność globalnego wsparcia przedsiębiorstw NVIDIA z ekspercką wiedzą w zakresie AI zarówno dla praktyków AI, jak i administratorów IT.

Oprogramowanie NVIDIA AI Enterprise jest dodatkiem licencyjnym dla GPU Tensor Core L4, co czyni AI dostępnym dla niemal każdej organizacji, oferując najwyższą wydajność w zakresie treningu, wnioskowania i nauki o danych. NVIDIA AI Enterprise wraz z NVIDIA L4 upraszcza budowę platformy gotowej do AI, przyspiesza rozwój i wdrażanie AI oraz zapewnia wydajność, bezpieczeństwo i skalowalność, aby szybciej gromadzić informacje i osiągać wartość biznesową.

Dowiedz się, jakie obciążenia AI możesz uruchamiać na L4 dzięki darmowym, praktycznym laboratoriom NVIDIA AI Enterprise w ramach NVIDIA LaunchPad.

Gotowy, aby zacząć?

Aby dowiedzieć się więcej na temat NVIDIA L4 Tensor Core GPU, odwiedź stronę: **www.nvidia.com/l4**

FORMAT | NVIDIA