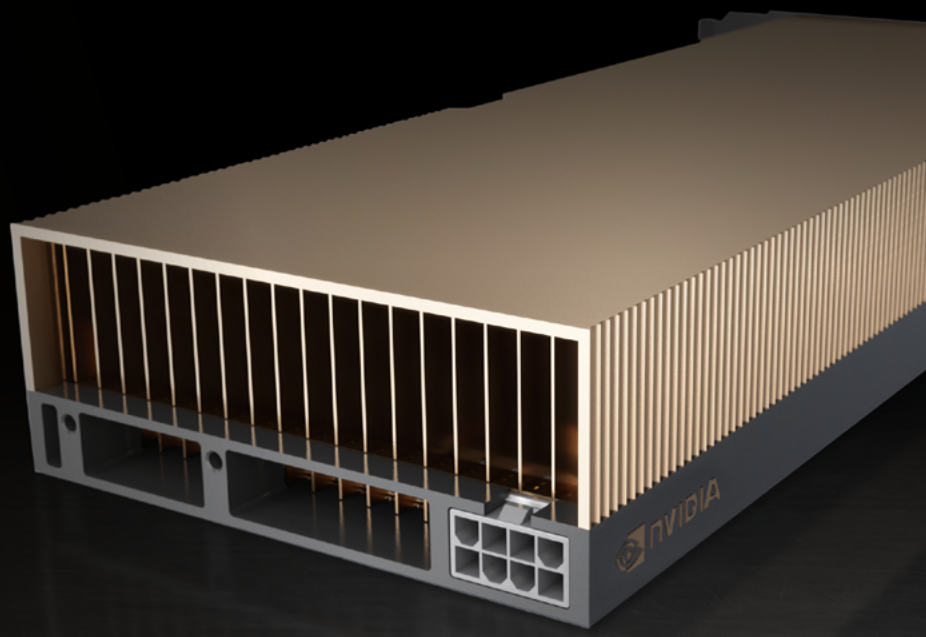
ARKUSZ DANYCH

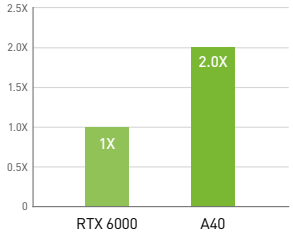


NVIDIA A40

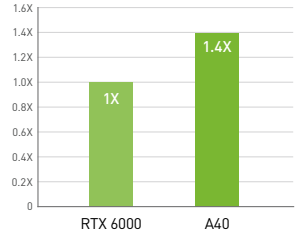
Potężny GPU do centrów danych dla obliczeń wizualnych

GPU NVIDIA A40 przyspiesza najbardziej wymagające obciążenia obliczeniowe związane z wizualizacją w centrach danych, łącząc najnowszą architekturę NVIDIA Ampere, rdzenie RT, rdzenie Tensor i rdzenie CUDA® z 48 GB pamięci graficznej. Od potężnych wirtualnych stacji roboczych dostępnych z dowolnego miejsca po dedykowane węzły renderujące, NVIDIA A40 wprowadza technologię NVIDIA RTX™ nowej generacji do centrum danych, zapewniając wsparcie dla najbardziej zaawansowanych obciążeń wizualizacji profesjonalnej.

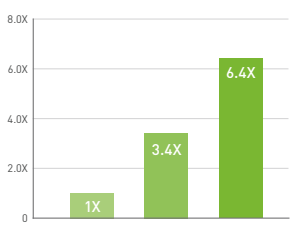
Do 2x większa wydajność renderowania[1]



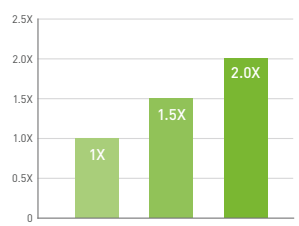
Do 40% szybsza wydajność grafiki[1]



Do 3x szybsze szkolenie AI[2]



Do 50% większa wydajność HPC z pojedynczą precyzją (FP32)[2]



* Włączona rzadkość strukturalna
** A40 jest domyślnie skonfigurowany do wirtualizacji z wyłączonymi fizycznymi złączami wyświetlacza. Wyjścia wyświetlacza można włączyć za pomocą narzędzi do zarządzania oprogramowaniem.

PODSTAWOWE CECHY

Architektura GPU	**Architektura NVIDIA Ampere**
Pamięć GPU	**48 GB GDDR6 z ECC**
Przepustowość pamięci	**696 GB/s**
Interfejs wzajemny	**NVIDIA® NVLink® 112.5 GB/s (dwukierunkowy)[3] PCIe Gen4: 64GB/s**
Rdzenie CUDA oparte na architekturze NVIDIA Ampere	**10,752**
Rdzenie RT drugiej generacji firmy NVIDIA	**84**
Trzecia generacja NVIDIA Rdzenie Tensorowe	**336**
Szczyt FP32 TFLOPS (bez tensora)	**37.4**
Szczytowy TFLOPS tensora FP16 z akumulacją FP16	**149.7 | 299.4***
Szczyt TF32 Tensor TFLOPS	**74.8 | 149.6***
Wydajność rdzenia RT TFLOPS	**73.1**
Szczytowy tensor BF16 TFLOPS z akumulacją FP32	**149.7 | 299.4***
Szczyt INT8 Tensor TOPS Szczyt INT 4 Tensor TOPS	**299.3 | 598.6*** **598.7 | 1,197.4***
Forma	**Podwójne gniazdo 4,4" (wys.) x 10,5" (dł.).**
Porty wyświetlacza	**3x DisplayPort 1.4**; Obsługuje NVIDIA Mosaic i Quadro® Sync[4]**
Maksymalne zużycie energii	**300 W**
Złącze zasilania	**8-pin CPU**
Rozwiązanie termiczne	**Bierne**
Obsługa oprogramowania wirtualnego procesora graficznego (vGPU).	**NVIDIA vPC/vApps, wirtualna stacja robocza NVIDIA RTX, wirtualny serwer obliczeniowy NVIDIA**
Obsługiwane profile vGPU	**Zobacz Przewodnik po licencjonowaniu wirtualnych GPU.**
NVENC | NVDEC	**1x | 2x (w tym dekodowanie AV1)**
Bezpieczny i wyważony rozruch ze sprzętowym źródłem zaufania	**Tak**
NEBS gotowy	**Poziom 3**
Oblicz interfejsy API	**CUDA, DirectCompute, OpenCL™, OpenACC®**
API graficzne	**DirectX 12.07[5], Shader Model 5.17[5], OpenGL 4.68[6], Vulkan 1.18[6]**
Wsparcie MIG	**Nie**

Specyfikacja

SPECYFIKACJE PRODUKTU

Całkowite zużycie energii	300 W
Rozwiązanie termiczne	Pasywne
Mechaniczny format obudowy	Pełny profil, pełna długość (FHFL) 10,5", dwuslotowy
Taktowanie GPU	Bazowe: 1305 MHz Boost: 1740 MHz
VBIOS	Rozmiar pamięci EEPROM: 8 Mbit UEFI: Obsługiwane
Interfejs PCI Express	PCI Express 4.0 ×16 Obsługiwane odwracanie linii i polaryzacji
Zero Power	Nie obsługiwane
Gotowość NEBS:	Obsługiwana
Złącza i gniazda zasilania	Jedno dodatkowe złącze zasilania CPU 8-pin
Waga	Płyta: 990 g (bez wspornika i przedłużaczy) Wspornik z wkrętami: 20 g Długi przedłużacz offsetowy: 48 g Prosty przedłużacz: 32 g

SPECYFIKACJE PAMIĘCI

Taktowanie pamięci	7251 MHz
Typ pamięci	GDDR6
Rozmiar pamięci	48 GiB[1,2]
Szerokość magistrali pamięci	384 bits
Maksymalna przepustowość pamięci	Do 696 GiB/s[1]

Uwaga:
[1] Notacja GiB podkreśla „potęgę dwójki" wartości. Zatem, 48 GiB = 48 x 1024[3]
[2] Pojemność pamięci DRAM obejmuje pamięć GPU dostępną dla aplikacji, wszelkie przestrzenie parametrów potrzebne przez sterownik NVIDIA oraz redundancję ECC (jeśli ECC jest włączone).

SPECYFIKACJE OPROGRAMOWANIA

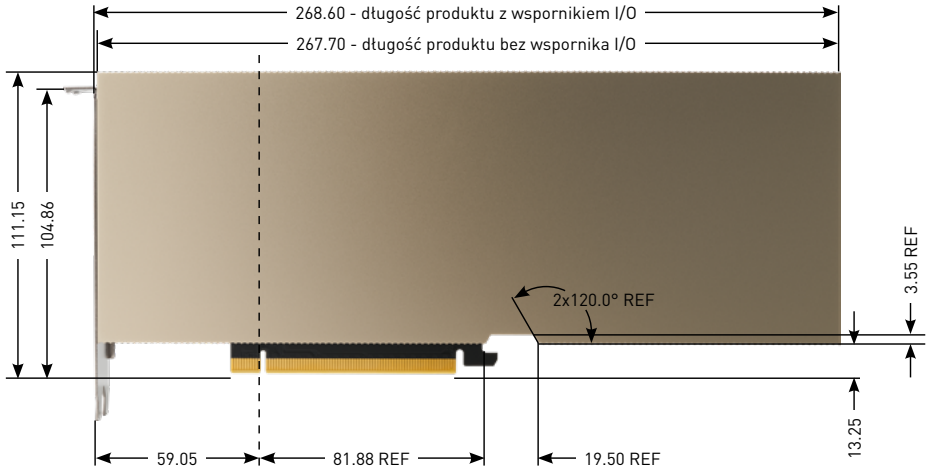
Obsługa SR-IOV	Obsługiwane: 32 VF (funkcji wirtualnych)
Adres BAR (fizyczna funkcja)	BAR0: 16 MiB BAR1: 64 GiB (tryb wyłączonego wyświetlacza; domyślny) BAR1: 8 GiB (tryb włączonego wyświetlacza, 8 GB BAR1) BAR1: 256 MiB (tryb włączonego wyświetlacza, 256 MB BAR1) BAR3: 32 MiB
Adres BAR (funkcja wirtualna)	Tryb wyłączonego wyświetlacza (domyślny): • BAR0: 8 MiB (32 VF x 256 KiB) • BAR1: 64 GiB, 64-bit (32 VF x 2 GiB) • BAR3: 1 GiB, 64-bit (32 VF x 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
Multi-Instance GPU (MIG)	Nie obsługiwane
Przekazywanie ARI	Obsługiwane
Wsparcie sterownika	R460.16 lub nowszy
Uruchamianie zabezpieczeń	Obsługiwane
Firmware CEC	v5.01 lub nowszy (dla kart z obsługą CEC)
Wsparcie dla NVIDIA® CUDA®	CUDA 11.2 lub nowsze
Wsparcie oprogramowania Virtual GPU	Obsługuje vGPU 12.0 lub nowszy
NVIDIA® NGC-Ready™ Test Suite	Certyfikacja NGC-Next 2.x lub nowsza
Tryby pracy	Tryb wyłączonego wyświetlacza (domyślny) Tryb włączonego wyświetlacza, 8 GiB BAR1 Tryb włączonego wyświetlacza, 256 MiB BAR1
Wsparcie ECC	Włączone (domyślnie); można wyłączyć za pomocą oprogramowania
Kod klasy PCI	0x03 – Kontroler wyświetlacza
Kod podklasy PCI	0x02 – Kontroler 3D
Zdolność podstawowego urządzenia rozruchowego	Nieobsługiwane w żadnym z trybów pracy
SMBus (adres 8-bitowy)	0x9E (zapis), 0x9F (odczyt)
Bezpośredni dostęp do SMBus	Obsługiwany
Zarezerwowane adresy I2C	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).



Infrastruktura Wirtualnych Pulpitów (VDI)

Procesor graficzny NVIDIA A40 stanowi znaczący postęp w infrastrukturze wirtualnych pulpitów (VDI), oferując solidne możliwości dostosowane do wymagań nowoczesnych środowisk pracy zdalnej. Zbudowany w oparciu o architekturę Ampere firmy NVIDIA, A40 łączy w sobie dużą liczbę rdzeni CUDA i zaawansowane funkcje graficzne, aby zapewnić wyjątkową wydajność i wierność wizualną dla wirtualnych pulpitów. Obsługuje wielu jednoczesnych użytkowników, uzyskując z łatwością dostęp do grafiki o wysokiej rozdzielczości i aplikacji wymagających dużej mocy obliczeniowej, co czyni go idealnym rozwiązaniem dla branż takich jak finanse, opieka zdrowotna i projektowanie, gdzie kluczowa jest bezproblemowa zdalna współpraca i bezpieczne przetwarzanie danych. Wydajne możliwości kodowania i dekodowania A40 zapewniają płynne przesyłanie strumieniowe i responsywność użytkownika, zwiększając produktywność i elastyczność zdalnych pracowników. Integracja z technologiami wirtualizacyjnymi NVIDIA, w tym NVIDIA GRID i VMware vSphere, umożliwia efektywną alokację zasobów i zarządzanie nimi, optymalizację infrastruktury IT i redukcję kosztów operacyjnych. Ogólnie rzecz biorąc, procesor graficzny NVIDIA A40 wyznacza nowy standard dla rozwiązań VDI, umożliwiając organizacjom dostarczanie wysokiej wydajności wirtualnych komputerów stacjonarnych, które mogą konkurować z tradycyjnymi konfiguracjami stacji roboczych, przy jednoczesnym zachowaniu bezpieczeństwa i skalowalności w rozproszonych środowiskach pracy.

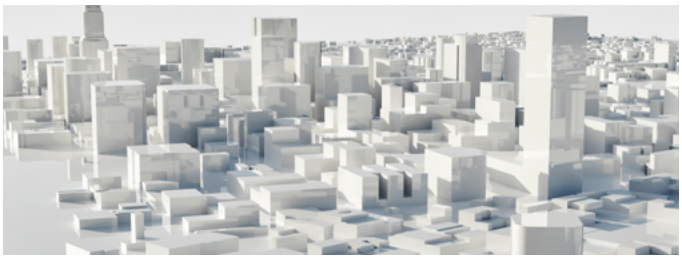
High-Performance Computing (HPC)

ProcesorgraficznyNVIDIAA40topotężnerozwiązaniezaprojektowane z myślą o wyniesieniu środowisk obliczeniowych o dużej wydajności (HPC) na nowy poziom wydajności i efektywności. Zbudowany w oparciu o architekturę Ampere firmy NVIDIA, A40 jest wyposażony w znaczną liczbę rdzeni CUDA i rdzeni Tensor, zoptymalizowanych do obsługi złożonych symulacji naukowych, obliczeniowej dynamiki płynów, modelowania molekularnego i innych obciążeń HPC wymagających dużej ilości danych z niezrównaną szybkością i dokładnością. Wysoka przepustowość pamięci i obsługa technologii NVLink umożliwiają bezproblemową komunikację danych i wydajne skalowanie na wielu procesorach graficznych, umożliwiając badaczom i naukowcom skuteczne radzenie sobie z większymi i bardziej złożonymi problemami obliczeniowymi. Zdolność A40 do obliczeń o mieszanej precyzji zwiększa wydajność obliczeniową bez uszczerbku dla precyzji, co czyni go idealnym wyborem do przyspieszania zadań związanych ze sztuczną inteligencją i uczeniem maszynowym w ramach przepływów pracy HPC. Zintegrowany z platformą obliczeń równoległych i bibliotekami CUDA firmy NVIDIA, A40 upraszcza tworzenie i wdrażanie zoptymalizowanych rozwiązań programowych, umożliwiając organizacjom osiąganie przełomów w badaniach naukowych, symulacjach inżynieryjnych i nie tylko. Krótko mówiąc, procesor graficzny NVIDIA A40 ustanawia nowy standard wydajności i skalowalności w HPC, zapewniając moc obliczeniową potrzebną do napędzania innowacji i odkryć w różnych dziedzinach nauki i przemysłu.

Uczenie głębokie i AI (sztuczna inteligencja)

Procesor graficzny NVIDIA A40 przoduje w przyspieszaniu aplikacji głębokiego uczenia się i sztucznej inteligencji, wykorzystując najnowocześniejszą architekturę Ampere, aby zapewnić niezrównaną wydajność i efektywność. Zaprojektowany specjalnie pod kątem obciążeń AI, A40 może pochwalić się dużą gęstością rdzeni CUDA i rdzeni Tensor, umożliwiając przetwarzanie ogromnych ilości danych i złożonych modeli sieci neuronowych z wyjątkową szybkością i precyzją. Dzięki temu idealnie nadaje się do szkolenia wielkoskalowych modeli sztucznej inteligencji, wykonywania zadań wnioskowania i prowadzenia przełomowych badań w takich dziedzinach, jak przetwarzanie języka naturalnego, wizja komputerowa i systemy autonomiczne. Obsługa obliczeń o mieszanej precyzji w A40 dodatkowo optymalizuje wydajność, równoważąc dokładność obliczeń z wydajnością, aby przyspieszyć przebieg procesów uczenia się i wnioskowania. Zintegrowany z kompleksowym ekosystemem oprogramowania NVIDIA, w tym CUDA, cuDNN i TensorRT, A40 usprawnia rozwój i wdrażanie aplikacji AI, skracając czas uzyskania wglądu i zwiększając produktywność badaczy i analityków danych. Ogólnie rzecz biorąc, procesor graficzny NVIDIA A40 wyznacza nowy standard głębokiego uczenia się i akceleracji sztucznej inteligencji, umożliwiając organizacjom odblokowanie nowych możliwości w zakresie innowacji i odkryć.

High-End Rendering

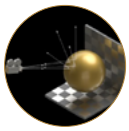


Procesor graficzny NVIDIA A40 to potężne rozwiązanie dla zaawansowanych aplikacji renderujących, wykorzystujące zaawansowaną architekturę Ampere, aby zapewnić wyjątkową wydajność i efektywność. Zaprojektowany specjalnie z myślą o wymagających obciążeniach związanych z renderowaniem w branżach takich 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.

Zajrzyj do wnętrza architektury NVIDIA Ampere

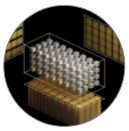
ARCHITEKTURA NVIDIA AMPERE RDZENIE CUDA

Przetwarzanie z podwójną szybkością dla operacji zmiennoprzecinkowych pojedynczej precyzji (FP32) oraz poprawiona efektywność energetyczna zapewniają znaczne zyski wydajności w zadaniach graficznych i obliczeniowych, takich jak złożony projekt wspomagany komputerowo (CAD) oraz inżynieria wspomagana komputerowo (CAE).



RDZENIE RT DRUGIEJ GENERACJI

Z nawet 2-krotnie większą przepustowością w porównaniu do poprzedniej generacji oraz możliwością jednoczesnego uruchamiania śledzenia promieni (ray tracing) z cieniowaniem lub technologią usuwania szumów, rdzenie RT drugiej generacji zapewniają ogromne przyspieszenia dla obciążeń takich jak fotorealistyczne renderowanie treści filmowych, oceny projektów architektonicznych i wirtualne prototypowanie projektów produktów. Ta technologia przyspiesza również renderowanie rozmycia ruchu z ray tracingiem, co prowadzi do szybszych wyników z większą dokładnością wizualną.



RDZENIE TENSOR TRZECIEJ GENERACJI

Precyzja Tensor Float 32 (TF32) zapewnia do 5 razy wyższą przepustowość treningową w porównaniu do poprzedniej generacji, co przyspiesza szkolenie modeli AI i nauki o danych bez potrzeby wprowadzania zmian w kodzie. Sprzętowe wsparcie dla strukturalnej rzadkości (structural sparsity) umożliwia do podwojenia przepustowości dla wnioskowania. Rdzenie Tensor wprowadzają również AI do grafiki, oferując takie funkcje jak superpróbkowanie z wykorzystaniem głębokiego uczenia (DLSS), usuwanie szumów AI oraz ulepszoną edycję dla wybranych aplikacji.

48 GB PAMIĘCI GDDR6 Z NVLINK

Ultraszybka pamięć GDDR6, rozszerzalna do 96 GB z NVLink[3], zapewnia naukowcom zajmującym się danymi, inżynierom i profesjonalistom kreatywnym dużą pamięć niezbędną do pracy z ogromnymi zestawami danych i obciążeniami, takimi jak nauka o danych i symulacje.

PCIE EXPRESS GEN 4

PCI Express Gen 4 podwaja przepustowość PCIe Gen 3, poprawiając prędkości transferu danych z pamięci CPU dla zadań wymagających dużej ilości danych, takich jak AI, nauka o danych i projektowanie 3D. Szybsza wydajność PCIe przyspiesza również bezpośrednie transfery pamięci GPU (DMA), co umożliwia szybszą komunikację wejścia/wyjścia danych wideo między GPU a GPUDirect® dla urządzeń obsługujących wideo, co stanowi potężne rozwiązanie do transmisji na żywo. A40 jest również wstecznie kompatybilny z PCI Express Gen 3, co zapewnia elastyczność wdrożenia.

WYDAJNOŚĆ I BEZPIECZEŃSTWO CENTRUM DANYCH

Dzięki konstrukcji o podwójnej szerokości i efektywności energetycznej, NVIDIA A40 jest do 2 razy bardziej energooszczędna niż poprzednia generacja i w pełni kompatybilna z szeroką gamą serwerów od globalnych producentów OEM. NVIDIA A40 obejmuje bezpieczny i mierzalny rozruch z technologią sprzętowej podstawy zaufania, co zapewnia, że oprogramowanie sprzętowe nie zostanie zmodyfikowane ani uszkodzone.

GPU NVIDIA A40 zapewnia nowoczesne możliwości obliczeń wizualnych, w tym śledzenie promieni w czasie rzeczywistym, przyspieszenie AI oraz elastyczność w obsłudze wielu obciążeń, co pozwala na przyspieszenie zadań związanych z uczeniem głębokim, nauką o danych oraz obliczeniami. Wirtualne stacje robocze zasilane przez NVIDIA A40 oraz oprogramowanie NVIDIA RTX Virtual Workstation (vWS) i NVIDIA Virtual Compute Server korzystają z obszernego testowania w szerokim zakresie zastosowań przemysłowych i profesjonalnego oprogramowania, co zapewnia optymalną wydajność i stabilność.

KAŻDY FRAMEWORK UCZENIA GŁĘBOKIEGO

mxnet

PYTORCH

Apache Spark

TensorFlow

RTX DLA APLIKACJI PROFESJONALNYCH

Adobe Premiere Pro

SOLIDWORKS

PLM Software SIEMENS NX

AUTODESK ARNOLD

REDSHIFT

AUTODESK VRED

KeyShot

UNREAL ENGINE

blender

octanerender

V-RAY

Dowiedz się więcej

Aby dowiedzieć się więcej na temat procesora graficznego NVIDIA A40, odwiedź stronę **www.nvidia.com/a40**

[1] Testy renderowania i grafiki przeprowadzono na dwóch procesorach Xeon Gold 6126 2.6GHz (3.7GHz Turbo). Pamięć systemowa 256 GB. Sterownik NVIDIA 461.09. Test renderowania: Iray 2020.1, czas renderowania sceny NVIDIA Endeavor. Test grafiki: SPECviewperf 2020 Subtest, 4K medical-03 Composite.
[2] Testy AI i HPC przeprowadzono na AMD EPYC 7742@2.25GHz (3.4GHz Turbo). Pamięć systemowa 512 GB. Sterownik NVIDIA 460.14. Szkolenie AI: przepustowość pre-treningu BERT. PyTorch (2/3) Faza 1 i (1/3) Faza 2. Precyzja FP32 dla RTX 6000 oraz TF32 dla A40 i A100. Długość sekwencji dla Fazy 1 = 128, Faza 2 = 512. HPC w precyzji pojedynczej: NAMD wersja 3.0a7, stmv_nve_cuda; Precyzja = FP32; ns/dzień, wersja CUDA: 11.1.74.
[3] Połączenie dwóch kart NVIDIA A40 za pomocą NVLink w celu zwiększenia wydajności i pojemności pamięci do 96 GB jest możliwe tylko wtedy, gdy Twoja aplikacja obsługuje technologię NVLink. Proszę skontaktować się z dostawcą aplikacji, aby potwierdzić ich wsparcie dla NVLink.
[4] Karta Quadro Sync II sprzedawana oddzielnie. Mosaic jest obsługiwany na systemach Windows 10 i Linux.
[5] GPU wspiera interfejs API DX 12.0, Hardware Feature Level 12 + 1.
[6] Produkt oparty jest na opublikowanej specyfikacji Khronosa i oczekuje się, że przejdzie proces testowania zgodności Khronosa, gdy stanie się dostępny. Aktualny status zgodności można znaleźć na stronie www.khronos.org/conformance.

FORMAT | NVIDIA