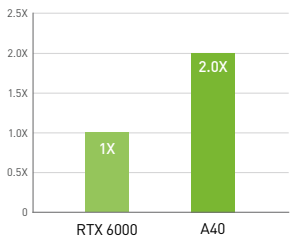
NVIDIA A40

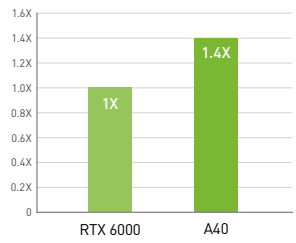
Wydajny procesor graficzny dla centrum danych do obliczeń wizualnych

NVIDIA A40 przyspiesza najbardziej wymagające obciążenia obliczeń wizualnych z centrum danych, łącząc najnowsze rdzenie RT, rdzenie Tensor i rdzenie CUDA® architektury NVIDIA Ampere 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 w celu realizacji najbardziej zaawansowanych profesjonalnych obciążeń wizualizacyjnych.

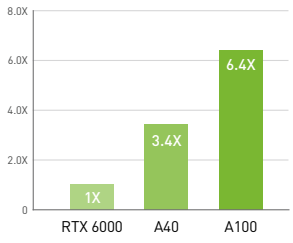
Do 2x większa wydajność renderowania[1]



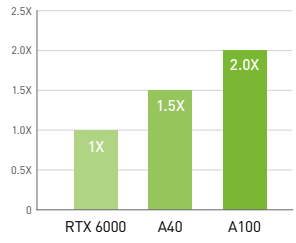
Do 40% szybsza wydajność grafiki[1]



Do 3x szybsze szkolenie AI[2]



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



SPECYFIKACJE

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 licencjonowania wirtualnego procesora graficznego**
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**

* Włączono ustrukturyzowana regularyzację rzadkości
** A40 jest domyślnie skonfigurowany do wirtualizacji z wyłączonymi złączami wyświetlacza fizycznego. Wyjścia wyświetlacza można włączyć za pomocą narzędzi oprogramowania zarządzającego.

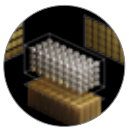
Spojrzenie na architekturę NVIDIA Ampere

RDZENIE CUDA OPARTE NA ARCHITEKTURZE NVIDIA AMPERE

Podwójna prędkość przetwarzania operacji zmiennoprzecinkowych o pojedynczej precyzji (FP32) i zwiększona wydajność energetyczna zapewniają znaczny wzrost wydajności w procesach graficznych i obliczeniowych, takich jak złożone projektowanie wspomagane komputerowo 3D (CAD) i inżynieria wspomagana komputerowo (CAE).

RDZENIE RT DRUGIEJ GENERACJI

Dzięki nawet dwukrotnie większej przepustowości w porównaniu z poprzednią generacją oraz możliwości jednoczesnego śledzenia promieni z funkcjami cieniowania lub odszumiania, rdzenie RT drugiej generacji zapewniają ogromne przyspieszenie w przypadku takich obciążeń, jak fotorealistyczne renderowanie treści filmowych, ocena projektów architektonicznych i wirtualne prototypowanie projekty produktów. Technologia ta przyspiesza także renderowanie rozmycia ruchu spowodowanego śledzeniem promieni, zapewniając szybsze rezultaty i większą dokładność wizualną.



RDZENIE TENSOROWE TRZECIEJ GENERACJI

Precyzja Tensor Float 32 (TF32) zapewnia do 5 razy większą przepustowość uczenia w porównaniu z poprzednią generacją, aby przyspieszyć szkolenie modeli AI i nauki o danych bez żadnych zmian w kodzie. Sprzętowa obsługa rzadkości strukturalnej zapewnia nawet dwukrotnie większą przepustowość wnioskowania. Rdzenie Tensor wprowadzają również sztuczną inteligencję do grafiki dzięki funkcjom takim jak superpróbkowanie głębokiego uczenia się (DLSS), odszumianie AI i ulepszona edycja dla wybranych aplikacji.

48 GB PAMIĘCI GDDR6 NVLINK

Ultraszybka pamięć GDDR6, skalowalna do 96 GB za pomocą NVLink3, zapewnia analitykom danych, inżynierom i kreatywnym profesjonalistom dużą pamięć niezbędną do pracy z ogromnymi zbiorami danych i obciążeniami, takimi jak analiza danych i symulacje.

PCIE EXPRESS GENERACJI 4

PCI Express Gen 4 podwaja przepustowość PCIe Gen 3, poprawiając prędkość przesyłania danych z pamięci procesora w przypadku zadań wymagających dużej ilości danych, takich jak sztuczna inteligencja, analiza danych i projektowanie 3D. Szybsza wydajność PCIe przyspiesza także transfery bezpośredniego dostępu do pamięci GPU (DMA), zapewniając szybszą komunikację wejścia/wyjścia danych wideo pomiędzy procesorem graficznym a GPUDirect® dla urządzeń obsługujących wideo, zapewniając potężne rozwiązanie do transmisji na żywo. A40 jest wstecznie kompatybilny z PCI Express Gen 3, co zapewnia elastyczność wdrażania.

WYDAJNOŚĆ I BEZPIECZEŃSTWO CENTRUM DANYCH

Wyposażona w dwa gniazda, energooszczędną konstrukcję, NVIDIA A40 jest nawet 2 razy bardziej energooszczędna niż poprzednia generacja i kompatybilna z szeroką gamą serwerów światowych producentów OEM. NVIDIA A40 obejmuje bezpieczne i mierzone uruchamianie ze sprzętową technologią root-of-trust, która gwarantuje, że oprogramowanie sprzętowe nie zostanie naruszone ani uszkodzone.

Procesor graficzny NVIDIA A40 zapewnia najnowocześniejsze możliwości obliczeń wizualnych, w tym śledzenie promieni w czasie rzeczywistym, akcelerację AI i elastyczność obsługi wielu obciążeń, aby przyspieszyć głębokie uczenie się, analizę danych i obciążenia obliczeniowe. Wirtualne stacje robocze wyposażone w technologię NVIDIA A40 i NVIDIA RTX Virtual Workstation (vWS) oraz oprogramowanie NVIDIA Virtual Compute Server korzystają z szeroko zakrojonych testów w szerokiej gamie aplikacji branżowych i profesjonalnego oprogramowania w celu zapewnienia optymalnej wydajności i stabilności.

RAMY GŁĘBOKIEGO NAUCZANIA

mxnet

PYTORCH

Apache Spark

TensorFlow

RTX DO ZASTOSOWAŃ 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 Rendering and Graphics tests run on 2x Xeon Gold 6126 2.6GHz (3.7GHz Turbo). 256GB system memory. NVIDIA Driver 461.09. Rendering test: Iray 2020.1, Render time of NVIDIA Endeavor scene. Graphics test: SPECviewperf 2020 Subtest, 4K medical-03 Composite | 2 AI and HPC tests run on AMD EPYC 7742@2.25GHz (3.4GHz Turbo). 512GB system memory. NVIDIA Driver 460.14. AI Training: BERT pre-training throughput. PyTorch (2/3) Phase 1 and (1/3) Phase 2. Precision FP32 for RTX 6000 and TF32 for A40 and A100. Sequence length for Phase 1 = 128. Phase 2 = 512. Single Precision HPC: NAMD version 3.0a7, stmv_nve_cuda; Precision=FP32; ns/day, CUDA Version: 11.1.74 | 3 Connecting two NVIDIA A40 cards with NVLink to scale performance and memory capacity to 96 GB is only possible if your application supports NVLink technology. Please contact your application provider to confirm their support for NVLink. | 4 Quadro Sync II card sold separately. Mosaic supported on Windows 10 and Linux. | 5 GPU supports DX 12.0 API, Hardware Feature Level 12 + 1. | 6 Product is based on a published Khronos specification and is expected to pass the Khronos conformance testing process when available. Current conformance status can be found at www.krhonos.org/conformance

FORMAT | NVIDIA