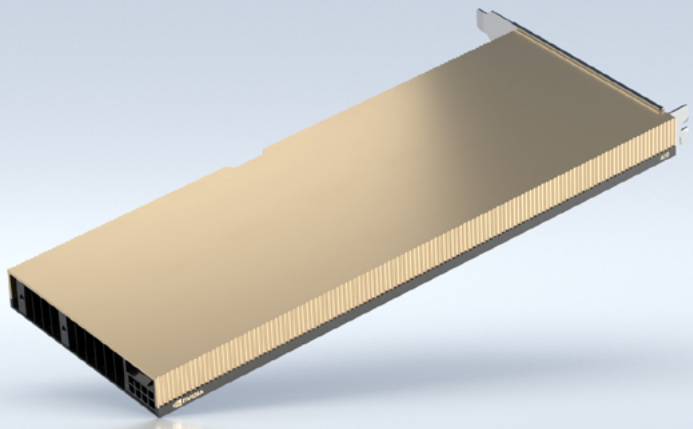
NVIDIA A10

AKCELEROWANA GRAFIKA I WIDEO Z AI DLA GŁÓWNYCH SERWERÓW KORPORACYJNYCH

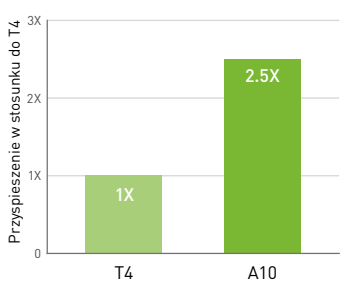


Wzbogacenie aplikacji graficznych i wideo o potężne AI

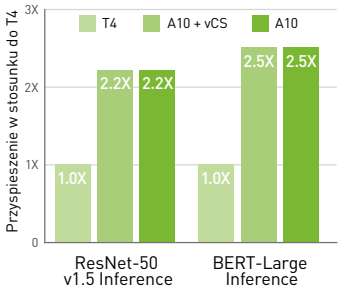
GPU NVIDIA A10 Tensor Core w połączeniu z oprogramowaniem NVIDIA RTX Virtual Workstation (vWS) dostarcza zaawansowaną grafikę i wideo wspieraną AI do serwerów korporacyjnych głównego nurtu, oferując rozwiązania, których potrzebują projektanci, inżynierowie, artyści i naukowcy, aby sprostać dzisiejszym wyzwaniom. Zbudowany na najnowszej architekturze NVIDIA Ampere, A10 łączy rdzenie RT drugiej generacji, rdzenie Tensor trzeciej generacji oraz nowe mikroprocesory strumieniujące z 24 gigabajtami (GB) pamięci GDDR6 – wszystko w limicie mocy wynoszącym 150W – zapewniając wszechstronne grafiki, rendering, AI i wydajność obliczeniową. Od wirtualnych stacji roboczych dostępnych w dowolnym miejscu na świecie, poprzez węzły renderujące, po centra danych obsługujące różnorodne obciążenia, A10 jest zbudowany, aby dostarczać optymalną wydajność w formacie PCIe o pełnej wysokości i długości oraz jednokrotnej szerokości.

NVIDIA A10 jest wspierany w ramach systemów certyfikowanych przez NVIDIĘ™, w lokalnym centrum danych, w chmurze i na brzegu sieci. A10 bazuje na bogatym ekosystemie frameworków AI z katalogu NVIDIA NGC™, bibliotekach CUDA-X™, ponad 2,3 milionach deweloperów oraz ponad 1800 aplikacjach zoptymalizowanych pod GPU, aby pomóc przedsiębiorstwom w rozwiązywaniu najważniejszych wyzwań związanych z ich działalnością.

A10 zapewnia do 2,5 razy większą wydajność wirtualnej stacji roboczej w porównaniu do T4[1]



A10 zapewnia do 2,5 x większą wydajność wnioskowania w porównaniu do T4[2]



NAJWAŻNIEJSZE CECHY

FP32	**31.2 TF**
TF32 Tensor Core	**62.5 TF | 125 TF***
BFLOAT16 Tensor Core	**125 TF | 250 TF***
FP16 Tensor Core	**125 TF | 250 TF***
INT8 Tensor Core	**250 TOPS | 500 TOPS***
INT4 Tensor Core	**500 TOPS | 1000 TOPS***
RT Cores	**72**
Encode / Decode	**1 encoder** **2 decoders** **(+AV1 decode)**
Pamięć GPU	**24 GB GDDR6**
Przepustowość pamięci GPU	**600 GB/s**
Łączność	**PCIe Gen4: 64 GB/s**
Kształt obudowy	**1-slot FHFL**
Maksymalna moc TDP	**150W**
Wsparcie oprogramowania vGPU	**NVIDIA vPC/vApps, NVIDIA RTX™ vWS, NVIDIA Virtual Compute Server (vCS)**
Bezpieczne i mierzone uruchamianie dzięki sprzętowemu Root of Trust	**TAK**
Gotowy NEBS	**Poziom 3**
Złącze zasilania	**PEX 8-pin**

*z rzadkością

Specyfikacja

SPECYFIKACJE PRODUKTU

Całkowite zużycie energii	150 W
Rozwiązanie termiczne	Pasywne
Mechaniczny format obudowy	Pełny profil, pełna długość (FHFL) 10,5", jednoslotowy
Identyfikatory urządzeń PCI	Identyfikator urządzenia: 0x2236 Identyfikator dostawcy: 0x10DE Identyfikator poddostawcy: 0x10DE Identyfikator podsystemu: 0x1482
Taktowanie GPU	Base: 885 MHz Boost: 1695 MHz
Stany wydajności	P0, P8
VBIOS	Rozmiar pamięci EEPROM: 8 Mbit UEFI: Obsługiwane
Interfejs PCI Express	PCI Express 4.0 ×16, x8; PCIe 3.0 ×16 Obsługiwane odwracanie linii i polaryzacji
Stany wydajności	P0, P8
Brak mocy (Zero Power)	Nie obsługiwane
Gotowość NEBS:	Obsługiwana
Złącza zasilania	Jedno dodatkowe złącze zasilania PCIe 8-pin
Waga	Płyta: 550 g (bez wspornika i przedłużaczy) Wspornik z wkrętami: 12 g Długi przedłużacz offsetowy: 64 g Prosty przedłużacz: 39 g

SPECYFIKACJE PAMIĘCI

Taktowanie pamięci	6251 MHz
Typ pamięci	GDDR6
Rozmiar pamięci	24 GB
Szerokość magistrali pamięci	384 bits
Maksymalna przepustowość pamięci	Do 600 GB/s

SPECYFIKACJE OPROGRAMOWANIA

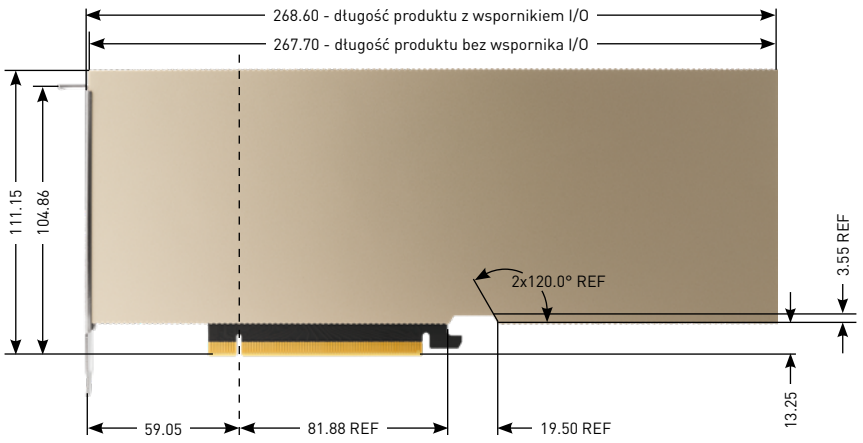
Obsługa SR-IOV	Obsługiwane: 32 VF (funkcji wirtualnych)
Adres BAR (fizyczna funkcja)	BAR0: 16 MiB[1] BAR1: 32 GiB[1] BAR3: 32 MiB[1]
Adres BAR (funkcja wirtualna)	BAR0: 8 MiB (256 KiB na VF)[1] BAR1: 64 GiB, 64-bit (2 GiB na VF)[1] BAR3: 1 GiB, 64-bit (32 MiB na VF)[1]
Wieloinstancyjny GPU (MIG)	Nie obsługiwane
Przekazywanie ARI	Obsługiwana
Wsparcie sterownika	Linux: R460.21 lub nowszy Windows: R460.57 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.x lub nowsze: • NVIDIA RTX Virtual Workstation (vWS) • NVIDIA Virtual PC (vPC)/Virtual Applications (vApps) • NVIDIA AI Enterprise • NVIDIA Virtual Compute Server (vCS)
NVIDIA® NGC-ReadyTM Test Suite	Certyfikacja NGC-Next 2.x lub nowsza
Kod klasy PCI	0x03 – Kontroler wyświetlacza
Kod podklasy PCI	0x02 – Kontroler 3D
Zdolność podstawowego urządzenia rozruchowego	Nie obsługiwane
Wsparcie ECC	Włączone (domyślnie); można wyłączyć za pomocą oprogramowania
SMBus (adres 8-bitowy)	0x9E (zapis), 0x9F (odczyt)
Zarezerwowane adresy I2C	0xAA, 0xAC
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]

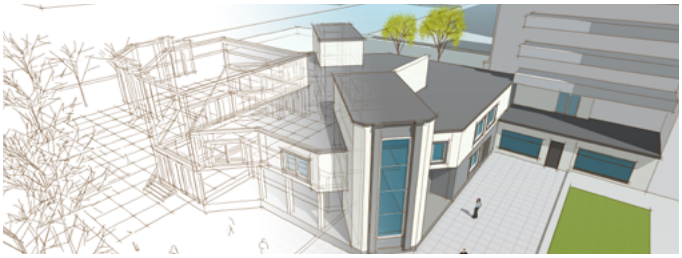
ŚRODOWISKOWE I NIEZAWODNOŚCIOWE SPECYFIKACJE

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).



Obciążenia związane z grafiką i wizualizacją



Procesor graficzny NVIDIA A10 to potężny akcelerator zaprojektowany specjalnie do obciążeń graficznych i wizualizacyjnych. Wykorzystując architekturę Ampere, łączy w sobie możliwości obliczeniowe o wysokiej wydajności z zaawansowanymi funkcjami graficznymi, dzięki czemu idealnie nadaje się do zadań takich jak renderowanie 3D, projektowanie wspomagane komputerowo (CAD) i aplikacje rzeczywistości wirtualnej (VR). A10 obsługuje technologię RTX firmy NVIDIA, która umożliwia śledzenie promieni w czasie rzeczywistym i grafikę wzmocnioną sztuczną inteligencją, zapewniając oszałamiająco realistyczny obraz i przyspieszone czasy renderowania. Dzięki znacznej przepustowości pamięci i rdzeniom CUDA, A10 zapewnia płynne i wydajne przetwarzanie złożonych zbiorów danych i skomplikowanych wizualizacji. Dodatkowo jego kompatybilność z pakietem profesjonalnych narzędzi i sterowników oprogramowania NVIDIA zapewnia bezproblemową integrację z istniejącymi przepływami pracy, co czyni go doskonałym wyborem dla profesjonalistów w takich dziedzinach, jak architektura, media i rozrywka oraz wizualizacja naukowa.

Analiza wideo i transkodowanie

Procesor graficzny NVIDIA A10 oferuje znaczne korzyści w zakresie analizy i transkodowania wideo, wykorzystując potężną architekturę Ampere i wyspecjalizowane możliwości przetwarzania wideo. Zawiera dedykowane kodery i dekodery sprzętowe, które umożliwiają wydajne przetwarzanie wielu strumieni wideo o wysokiej rozdzielczości jednocześnie, dzięki czemu idealnie nadaje się do zastosowań takich jak nadzór, moderowanie treści i inteligentna infrastruktura miejska, gdzie niezbędna jest analiza wideo w czasie rzeczywistym. Solidne rdzenie CUDA i rdzenie Tensor A10 przyspieszają algorytmy AI do analizy wideo, usprawniając zadania, takie jak wykrywanie obiektów, rozpoznawanie twarzy i śledzenie aktywności. W przypadku transkodowania wideo A10 zapewnia szybką, wysokiej jakości konwersję pomiędzy różnymi formatami wideo i rozdzielczościami, zapewniając niskie opóźnienia i utrzymanie wierności wizualnej na różnych urządzeniach i platformach. Integracja z kompleksowym pakietem oprogramowania firmy NVIDIA, w tym zestawem SDK do kodeków wideo i zestawem DeepStream SDK, zapewnia płynną optymalizację przepływu pracy i wdrażanie. Te funkcje sprawiają, że NVIDIA A10 jest niezbędnym narzędziem zwiększającym wydajność i wydajność w operacjach analizy wideo i transkodowania.

Wnioskowanie AI

Procesor graficzny NVIDIA A10 zmienia zasady gry w zakresie wnioskowania AI, oferując wyjątkową wydajność i efektywność przy wdrażaniu modeli AI w środowiskach produkcyjnych. Zbudowany w oparciu o zaawansowaną architekturę Ampere, charakteryzuje się dużą gęstością rdzeni CUDA i rdzeni Tensor nowej generacji, zaprojektowanych specjalnie w celu przyspieszenia zadań AI, takich jak wnioskowanie w sieci neuronowej. Dzięki temu A10 może zapewnić wysoką przepustowość i niskie opóźnienia, co jest kluczowe dla aplikacji AI działających w czasie rzeczywistym, takich jak rozpoznawanie obrazu i mowy, systemy autonomiczne i silniki rekomendacji. Jego możliwości obliczeniowe o mieszanej precyzji pozwalają zrównoważyć dokładność i wydajność, dzięki czemu modele AI są szybsze i wydajniejsze bez utraty precyzji. Ponadto A10 bezproblemowo integruje się z rozbudowanym ekosystemem oprogramowania AI firmy NVIDIA, w tym bibliotekami TensorRT i CUDA-X AI, upraszczając wdrażanie i optymalizację przepływów pracy AI. Te cechy sprawiają, że NVIDIA A10 jest idealnym rozwiązaniem dla przedsiębiorstw, które chcą wykorzystać pełny potencjał wnioskowania AI, zapewniając szybsze wyniki i stymulując innowacje w różnych branżach.

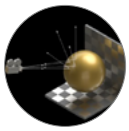
Główne obliczenia w przedsiębiorstwach

Procesor graficzny NVIDIA A10 to wszechstronna jednostka mocy zaprojektowana z myślą o podniesieniu poziomu głównego nurtu obliczeń korporacyjnych, zapewniając niezrównaną wydajność przy różnorodnych obciążeniach. Wykorzystując zaawansowaną architekturę Ampere, A10 zapewnia znaczną poprawę wydajności obliczeniowej, dzięki czemu idealnie nadaje się do analizy danych, infrastruktury wirtualnych pulpitów (VDI) i środowisk przetwarzania w chmurze. Bogate rdzenie CUDA i rdzenie Tensor umożliwiają przyspieszone przetwarzanie złożonych obliczeń, ułatwiając szybsze wyciąganie wniosków z dużych zbiorów danych i zwiększając wydajność modelu uczenia maszynowego. Szeroka przepustowość pamięci A10 zapewnia płynne zarządzanie zadaniami wymagającymi dużej ilości danych, podczas gdy technologia wirtualizacji NVIDIA umożliwia wielu użytkownikom jednoczesny dostęp do możliwości procesora graficznego, optymalizując wykorzystanie zasobów i redukując koszty operacyjne. Co więcej, płynna integracja A10 z kompleksowym ekosystemem oprogramowania NVIDIA, w tym CUDA, cuDNN i TensorRT, zapewnia kompatybilność i łatwość wdrożenia w istniejącej infrastrukturze IT. Cechy te wspólnie pozycjonują NVIDIA A10 jako kluczowy atut dla przedsiębiorstw, których celem jest zwiększenie mocy obliczeniowej, usprawnienie operacji i stymulowanie innowacji.

Spojrzenie na architekturę NVIDIA Ampere

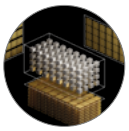
ARCHITEKTURA NVIDIA AMPERE RDZENIE CUDA

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



RDZENIE RT DRUGIEJ GENERACJI

Z nawet 2-krotnie większą przepustowością w porównaniu do poprzedniej generacji i możliwością jednoczesnego uruchamiania śledzenia promieni (ray tracing) z cieniowaniem lub technologii usuwania szumów, rdzenie RT drugiej generacji zapewniają ogromne przyspieszenie 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, zapewniając szybsze rezultaty z większą dokładnością wizualną.



RDZENIE TENSOR TRZECIEJ GENERACJI

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

24 GB GDDR6

Ultraszybka pamięć GDDR6, dostarczająca 600 GB/s przepustowości do renderowania, nauki o danych, symulacji inżynierskich i innych obciążeń intensywnie korzystających z pamięci GPU.

PCIE EXPRESS GEN 4

PCI Express Gen 4 podwaja przepustowość PCIe Gen 3, poprawiając prędkości transferu danych z pamięci CPU w przypadku zadań wymagających dużej ilości danych, takich jak AI, nauka o danych i projektowanie 3D. Szybsza wydajność PCIe przyspiesza również bezpośrednie transfery pamięci GPU (DMA), zapewniając szybszą komunikację wejścia/wyjścia danych wideo między GPU a NVIDIA GPUDirect® dla urządzeń obsługujących wideo, co stanowi potężne rozwiązanie do transmisji na żywo. A10 jest również wstecznie kompatybilny z PCI Express Gen 3, co zapewnia elastyczność wdrożenia.

WYDAJNOŚĆ I BEZPIECZEŃSTWO CENTRUM DANYCH

Dzięki jednoslotowej konstrukcji o pełnej wysokości i długości oraz energooszczędnej konstrukcji, NVIDIA A10 jest kompatybilna z szeroką gamą serwerów od globalnych producentów OEM. NVIDIA A10 obejmuje bezpieczny i mierzalny rozruch z technologią sprzętowej podstawy zaufania, co zapewnia, że oprogramowanie sprzętowe nie jest manipulowane ani uszkodzone.

GPU NVIDIA A10 Tensor Core jest idealny do zaawansowanej grafiki i wideo z AI. Rdzenie RT drugiej generacji i rdzenie Tensor trzeciej generacji wzbogacają aplikacje graficzne i wideo o potężne AI w 150W TDP dla serwerów głównego nurtu.

NVIDIA A10 łączy się również z oprogramowaniem NVIDIA wirtualnego GPU (vGPU), aby przyspieszyć różne obciążenia centrów danych — od bogatego w grafiki VDI po wysokowydajne wirtualne stacje robocze do AI — w łatwej w zarządzaniu, bezpiecznej i elastycznej infrastrukturze, którą można skalować, aby zaspokoić potrzeby zasobów.

STRUKTURA DEEP LEARNING

mxnet

PYTORCH

Spark

TensorFlow

RTX DO ZASTOSOWAŃ PROFESJONALNYCH

AUTODESK REVIT

creo

CATIA

Rhinoceros design, model, present, analyze, realize...

SOLIDWORKS

SIEMENS

Aby dowiedzieć się więcej na temat procesora graficznego NVIDIA A10 Tensor Core, odwiedź stronę **www.nvidia.com/a10**

1 Test run on a server with 2x Xeon Gold 6154 3.0GHz (3.7GHz Turbo), NVIDIA RTX vWS software, VMware ESXi 7 U2, host/guest driver 461.33. | SPECviewperf 2020 Subtest, and HD 3dsmax-07 composite.
2 BERT Large inference NVIDIA TensorRT7.2, Seq Length =128, batch size =128; NGC Container: 21.02-py3 | ResNet-50 v1.5: NVIDIA TensorRT7.2, INT8 precision batch size = 128 NGC Container: 20.12-py3 | NVIDIA A10 with vCS software, VMware ESXi 7 U2 host/guest driver 461.33

FORMAT | NVIDIA