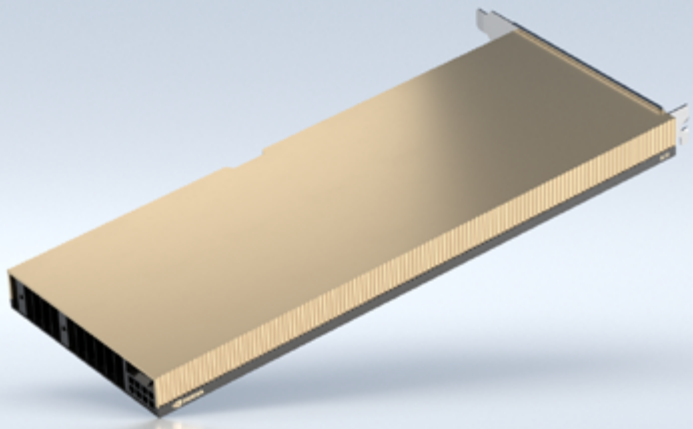
NVIDIA A10

AKCELERACJA GRAFIKI I WIDEO DZIĘKI SZTUCZNEJ INTELIGENCJI DLA GŁÓWNYCH SERWERÓW KORPORACYJNYCH

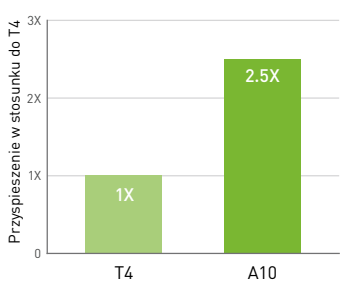


Wzbogać aplikacje graficzne i wideo z potężną sztuczną inteligencją

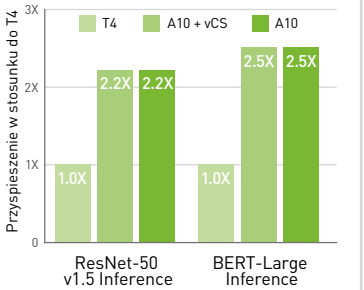
Procesor graficzny **NVIDIA A10 Tensor Core** łączy się z oprogramowaniem **NVIDIA RTX Virtual Workstation** (vWS), aby zapewnić mainstreamową grafikę i wideo wraz z usługami AI na głównych serwerach korporacyjnych, dostarczając rozwiązania potrzebne projektantom, inżynierom, artystom i naukowcom, aby sprostać dzisiejszym wyzwaniom. Zbudowany w oparciu o najnowszą architekturę **NVIDIA Ampere**, **A10** łączy w sobie rdzenie RT drugiej generacji, rdzenie Tensor trzeciej generacji i nowe mikroprocesory do przesyłania strumieniowego z 24 gigabajtami (GB) pamięci GDDR6 – a wszystko to w zakresie mocy 150 W – dla wszechstronnej grafiki, renderowania, AI i dużej wydajności obliczeniowej. Od wirtualnych stacji roboczych dostępnych w dowolnym miejscu na świecie, węzły renderujące po centra danych obsługujące różne obciążenia, **A10** został zbudowany tak, aby zapewniać optymalną wydajność PCIe w obudowie o pojedynczej szerokości, pełnej wysokości i pełnej długości.

NVIDIA A10 jest obsługiwana w ramach systemów z certyfikatem NVIDIA™, w lokalnym centrum danych, w chmurze i na krawędzi. **NVIDIA A10** opiera się na bogatym ekosystemie frameworków AI z katalogu NVIDIA NGC™, bibliotek CUDA-X™, ponad 2,3 miliona programistów i ponad 1800 aplikacji zoptymalizowanych pod kątem procesora graficznego, aby pomóc przedsiębiorstwom stawić czoła najbardziej krytycznym wyzwaniom w ich biznesie.

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]



SPECYFIKACJA

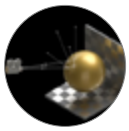
FP32	**31.2 TF**
TF32 Tensor Core	**62.5 TF | 125 TF***
BFLOAT16 Tensor Core	**125 TF | 250 TF***
FP16 Tensor Core	**125 TF | 250 TF***
INT8 Tensor Core	**250 TOPS | 500 TOPS***
INT4 Tensor Core	**500 TOPS | 1000 TOPS***
RT Cores	**72**
Encode / Decode	**1 encoder** **2 decoders** **(+AV1 decode)**
Pamięć 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

Spojrzenie na architekturę NVIDIA Ampere

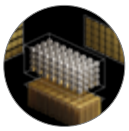
ARCHITEKTURA NVIDIA AMPERE RDZENIE CUDA

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



DRUGA GENERACJA RDZENI RT

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.

24 GB GDDR6

Ultraszybka pamięć GDDR6 zapewniająca przepustowość 600 GB/s do renderowania, analizy danych, symulacji inżynieryjnych i innych obciążeń intensywnie wykorzystujących pamięć GPU.

PCIE EXPRESS GEN 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 technologią NVIDIA GPUDirect® dla urządzeń obsługujących wideo, zapewniając potężne rozwiązanie do transmisji na żywo. A10 jest również wstecznie kompatybilny z PCI Express Gen 3, co zapewnia elastyczność wdrażania.

WYDAJNOŚĆ I BEZPIECZEŃSTWO CENTRUM DANYCH

Wyposażona w jednogniazdową, pełnowymiarową i energooszczędną konstrukcję, NVIDIA A10 jest kompatybilna z szeroką gamą serwerów światowych producentów OEM. NVIDIA A10 zapewnia bezpieczne i mierzone uruchamianie ze sprzętową technologią root-of-trust, która gwarantuje, że oprogramowanie sprzętowe nie zostanie naruszone ani uszkodzone.

Procesor graficzny NVIDIA A10 Tensor Core jest idealny do popularnych zastosowań graficznych i wideo ze sztuczną inteligencją. Rdzenie RT drugiej generacji i rdzenie Tensor trzeciej generacji wzbogacają aplikacje graficzne i wideo o potężną sztuczną inteligencję przy TDP 150 W dla serwerów głównego nurtu.

NVIDIA A10 łączy się również z oprogramowaniem wirtualnych procesorów graficznych NVIDIA (vGPU), aby przyspieszyć wiele obciążeń w centrach danych — od bogatych w grafikę VDI, przez wysokowydajne wirtualne stacje robocze, po sztuczną inteligencję — w łatwo zarządzanej, bezpiecznej i elastycznej infrastrukturze, którą można skalować w celu dostosowania do zasobów wymagania.

STRUKTURA DEEP LEARNING

mxnet

PYTORCH

Apache 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