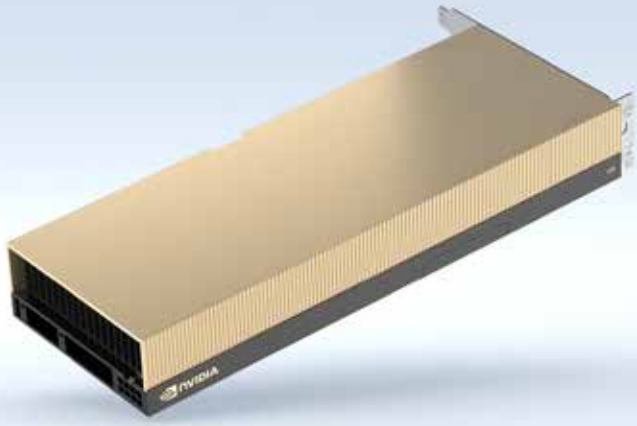
NVIDIA A30 TENSOR CORE GPU

WSZECHSTRONNE PRZYSPIESZENIE OBLICZEŃ DLA GŁÓWNYCH SERWERÓW KORPORACYJNYCH



Wnioskowanie AI i podstawowe obliczenia dla każdego przedsiębiorstwa

Procesor graficzny NVIDIA A30 Tensor Core to najbardziej wszechstronny procesor graficzny głównego nurtu, przeznaczony do wnioskowania AI i głównych obciążeń korporacyjnych. Oparty na architekturze NVIDIA Ampere i technologii Tensor Core, obsługuje szeroki zakres precyzji matematycznych, zapewniając pojedynczy akcelerator przyspieszający każde obciążenie.

Zbudowane z myślą o wnioskowaniu AI na dużą skalę, te same zasoby obliczeniowe mogą szybko przeszkolić modele AI za pomocą TF32, a także przyspieszyć aplikacje obliczeniowe o wysokiej wydajności (HPC) przy użyciu rdzeni Tensor FP64. Wieloinstancyjne procesory graficzne (MIG) i rdzenie Tensor FP64 łączą się z szybką przepustowością pamięci wynoszącą 933 gigabajtów na sekundę (GB/s) przy niskiej mocy 165 W, a wszystko to działa na karcie PCIe optymalnej dla serwerów głównego nurtu.

Połączenie rdzeni Tensor Core trzeciej generacji i MIG zapewnia bezpieczną jakość usług przy różnorodnych obciążeniach, a wszystko to jest zasilane przez wszechstronny procesor graficzny umożliwiający elastyczne centrum danych. Wszechstronne możliwości obliczeniowe A30 w dużych i małych obciążenia pracą zapewniają maksymalną wartość dla głównych przedsiębiorstw.

A30 jest częścią kompletnego rozwiązania NVIDIA dla centrum danych, które obejmuje elementy składowe sprzętu, sieci, oprogramowania, bibliotek oraz zoptymalizowanych modeli i aplikacji AI firmy NGC™. Stanowi najpotężniejszą, kompleksową platformę AI i HPC dla centrów danych, umożliwiając naukowcom dostarczanie wyników w świecie rzeczywistym i wdrażanie rozwiązań do produkcji na dużą skalę.

SPECYFIKACJE SYSTEMU

Pik FP64	**5.2TF**
Pik FP64 Tensor Core	**10.3 TF**
Pik FP32	**10.3 TF**
TF32 Tensor Core	**82 TF | 165 TF***
BFLOAT16 Tensor Core	**165 TF | 330 TF***
Pik FP16 Tensor Core	**165 TF | 330 TF***
Pik INT8 Tensor Core	**330 TOPS | 661 TOPS***
Pik INT4 Tensor Core	**661 TOPS | 1321 TOPS***
Silniki multimedialne	**1 optical flow accelerator (OFA)** **1 JPEG decoder (NVJPEG)** **4 Video decoders (NVDEC)**
Pamięć GPU	**24GB HBM2**
Przepustowość pamięci GPU	**933GB/s**
Łączność	**PCIe Gen4: 64GB/s** **Third-gen NVIDIA® NVLINK® 200GB/s****
Kształt obudowy	**2-slot, full height, full length (FHFL)**
Maksymalna moc obliczeniowa cieplna (TDP)	**165W**
GPU z wieloma instancjami (MIG)	**4 MIGs @ 6GB each** **2 MIGs @ 12GB each** **1 MIGs @ 24GB**
Obsługa oprogramowania wirtualnego GPU (vGPU).	**NVIDIA AI Enterprise for VMware** **NVIDIA Virtual Compute Server**

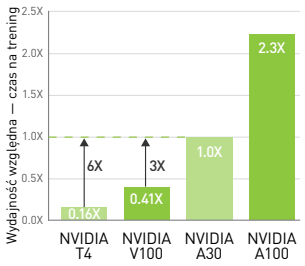
* Z rzadkością

** Most NVLink dla maksymalnie dwóch procesorów graficznych.

Niesamowita wydajność w różnych obciążeniach

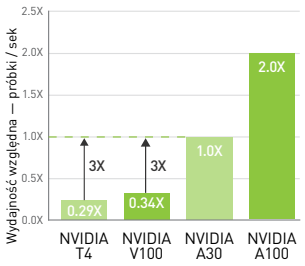
Szkolenie AI - do 3X większa przepustowość niż V100 i 6X wyższa niż T4[1]

BERT - Duże szkolenie wstępne [Znormalizowane]



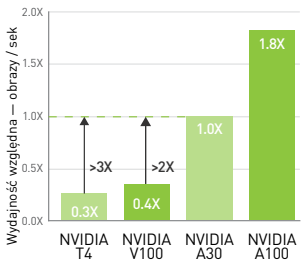
Wnioskowanie AI — do 3 x większa przepustowość w porównaniu z V100 przy konwersacji AI[2] w czasie rzeczywistym

BERT - Duże wnioskowanie [znormalizowane] Przepustowość dla opóźnienia < 1 Oms



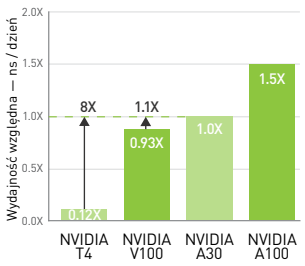
Wnioskowanie AI — ponad 3 x większa przepustowość niż T4 przy klasyfikacji obrazu w czasie rzeczywistym[3]

RN50 v1.5 - Wnioskowanie [znormalizowane] Przepustowość dla opóźnienia <7 ms



HPC - do 1,1X większa przepustowość niż V100 i 8X większa niż T4[4]

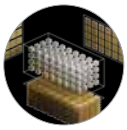
LAMPY [Znormalizowane]



Przełomowe innowacje

ARCHITEKTURA NVIDIA AMPERE

Niezależnie od tego, czy używasz MIG do podziału procesora graficznego A30 na mniejsze instancje, czy NVIDIA NVLink do łączenia wielu procesorów graficznych w celu przyspieszenia większych obciążeń, A30 z łatwością poradzi sobie z potrzebami akceleracji o różnej wielkości, od najmniejszego zadania po największe obciążenie wielowęzłowe. Wszechstronność A30 oznacza, że menedżerowie IT mogą przez całą dobę maksymalizować użyteczność każdego procesora graficznego w swoim centrum danych z głównymi serwerami.

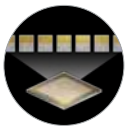


RDZENIE TENSOROWE TRZECIEJ GENERACJI

NVIDIA A30 zapewnia 165 teraflopów (TFLOPS) wydajności głębokiego uczenia się TF32. To 20 razy większa przepustowość szkolenia AI i ponad 5 razy większa wydajność wnioskowania w porównaniu z procesorem graficznym NVIDIA T4 Tensor Core. W przypadku komputerów HPC procesor A30 zapewnia wydajność na poziomie 10,3 TFLOPS, czyli prawie 30 procent więcej niż procesor graficzny NVIDIA V100 Tensor Core.

NVLINK NOWEJ GENERACJI

NVIDIA NVLink w A30 zapewnia 2X wyższą przepustowość w porównaniu do poprzedniej generacji. Za pomocą mostka NVLink można połączyć dwa procesory graficzne A30 PCIe, aby zapewnić wydajność głębokiego uczenia na poziomie 330 TFLOP.



WIELOINSTANCYJNY GPU (MIG)

Procesor graficzny A30 można podzielić na aż cztery instancje GPU, w pełni izolowane na poziomie sprzętowym z własną pamięcią o dużej przepustowości, pamięcią podręczną i rdzeniami obliczeniowymi. MIG zapewnia programistom dostęp do przełomowego przyspieszenia dla wszystkich ich aplikacji. Administratorzy IT mogą zaoferować akcelerację GPU odpowiedniej wielkości do każdego zadania, optymalizując wykorzystanie i rozszerzając dostęp dla każdego użytkownika i aplikacji.

HBM2

Dzięki maksymalnie 24 GB pamięci o dużej przepustowości (HBM2), A30 zapewnia przepustowość pamięci GPU na poziomie 933 GB/s, optymalną dla różnorodnych obciążeń AI i HPC na głównych serwerach.

NIEDOSTĘPNOŚĆ STRUKTURALNA

Sieci AI mają od milionów do miliardów parametrów. Nie wszystkie z tych parametrów są potrzebne do dokładnych przewidywań, a niektóre można przekształcić w zera, dzięki czemu modele będą „rzadkie" bez utraty dokładności. Rdzenie Tensor w A30 mogą zapewnić do 2X wyższą wydajność w przypadku rzadkich modeli. Chociaż funkcja rzadkości łatwiej przynosi korzyści w zakresie wnioskowania AI, może również poprawić wydajność uczenia modeli.inference, it can also improve the performance of model training.

Kompleksowe rozwiązanie dla przedsiębiorstw

Procesor graficzny NVIDIA A30 Tensor Core — oparty na architekturze NVIDIA Ampere, będący sercem nowoczesnego centrum danych — stanowi integralną część platformy centrum danych NVIDIA. Platforma stworzona z myślą o głębokim uczeniu się, HPC i analizie danych przyspiesza działanie ponad 2000 aplikacji, w tym wszystkich głównych platform głębokiego uczenia się. Ponadto NVIDIA AI Enterprise, kompleksowy, natywny w chmurze pakiet oprogramowania do sztucznej inteligencji i analizy danych, posiada certyfikat do działania na platformie A30 w infrastrukturze wirtualnej opartej na hypervisorze z VMware vSphere. Umożliwia to zarządzanie i skalowanie obciążeń AI w środowisku chmury hybrydowej. Kompletna platforma NVIDIA jest dostępna wszędzie, od centrum danych po brzeg, zapewniając zarówno ogromny wzrost wydajności, jak i możliwości oszczędności.

ZOPTYMALIZOWANE OPROGRAMOWANIE I USŁUGI DLA PRZEDSIĘBIORSTW

NVIDIA CERTIFIED

STRUKTURA KAŻDEGO GŁĘBOKIEGO NAUCZANIA

mxnet PYTORCH APACHE Spark TensorFlow

PONAD 2000 APLIKACJI AKCELEROWANYCH PRZEZ GPU

- Altair nanoFluidX
- Altair ultraFluidX
- AMBER
- ANSYS Fluent
- DS SIMULIA Abaqus
- GAUSSIAN
- GROMACS
- NAMD
- OpenFOAM
- VASP
- WRF

Aby dowiedzieć się więcej na temat GPU NVIDIA A30 Tensor Core, odwiedź stronę **www.nvidia.com/a30**

[1] BERT-Large Pre-Training (9/10 epok) Faza 1 i (1/10 epok) Faza 2, długość sekwencji dla fazy 1 = 128 i faza 2 = 512, zbiór danych = rzeczywisty, pojemnik NGC™ = 21,03,8x GPU: T4 (FP32, BS=8, 2) | V100 PCIE 16 GB (FP32, BS=8, 2) | A30 (TF32, BS=8, 2) | A100 PCIE 40 GB (TF32, BS=54, 8) | wskazane wielkości partii dotyczą odpowiednio fazy 1 i fazy 2

[2] NVIDIA® TensorRT®, precyzja = INT8, długość sekwencji = 384, kontener NGC 20.12, opóźnienie <10 ms, zbiór danych = syntetyczny; 1x procesor graficzny: A100 PCIe 40 GB (BS=8) | A30 (BS=4) | V100 SXM2 16 GB (BS=1) | T4 (BS=1)

[3] TensorRT, kontener NGC 20.12, opóźnienie <7 ms, zbiór danych=syntetyczny,; 1x GPU: T4 (BS=31, INT8) | V100 (BS=43, mieszane precyzja) | A30 (BS=96, INT8) | A100 (BS=174, INT8)

[4] Zbiór danych: ReaxFF/C, FP64 | 4x karta graficzna: T4, V100 PCIE 16 GB, A3

FORMAT NVIDIA