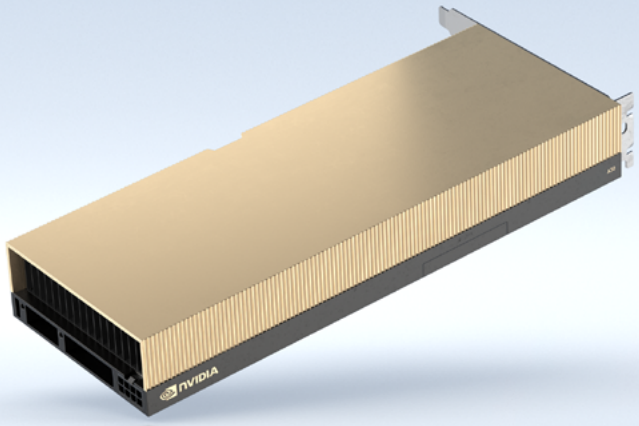
NVIDIA A30
GPU TENSOR CORE
WSZECHSTRONNE PRZYSPIESZENIE OBLICZEŃ DLA SERWERÓW KORPORACYJNYCH GŁÓWNEGO NURTU



AI WNIOSKOWANIE I OBCIĄŻENIA GŁÓWNEGO NURTU DLA KAŻDEGO PRZEDSIĘBIORSTWA

GPU NVIDIA A30 Tensor Core to najbardziej wszechstronny GPU do obliczeń głównego nurtu, przeznaczony do wnioskowania AI i obciążeń przedsiębiorstw. Napędzany technologią Tensor Core architektury NVIDIA Ampere, obsługuje szeroki zakres precyzji obliczeniowej, zapewniając pojedynczy akcelerator przyspieszający każde obciążenie.

Zbudowany z myślą o skalowalnym wnioskowaniu AI, ten sam zasób obliczeniowy może szybko ponownie trenować modele AI z TF32, a także przyspieszać aplikacje obliczeń o wysokiej wydajności (HPC) korzystając z rdzeni Tensor FP64. Wieloinstancyjny GPU (MIG) i rdzenie Tensor FP64 łączą się z szybką przepustowością pamięci wynoszącą 933 gigabajty na sekundę (GB/s) w niskim limicie mocy wynoszącym 165W, wszystko na karcie PCIe zoptymalizowanej dla serwerów głównego nurtu.

Kombinacja rdzeni Tensor trzeciej generacji i MIG oferuje wysoką jakość usług w różnych obciążeniach, zasilaną wszechstronnym GPU umożliwiającym elastyczne centrum danych. Wszechstronne możliwości obliczeniowe A30 w obciążeniach dużych i małych przynoszą maksymalną wartość dla przedsiębiorstw głównego nurtu.

A30 jest częścią kompleksowego rozwiązania centrum danych NVIDIA, które obejmuje komponenty sprzętowe, sieciowe, oprogramowanie, biblioteki oraz zoptymalizowane modele i aplikacje AI z NGC™. Reprezentując najsilniejszą platformę AI i HPC end-to-end dla centrów danych, umożliwia badaczom dostarczanie rzeczywistych wyników i wdrażanie rozwiązań 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**

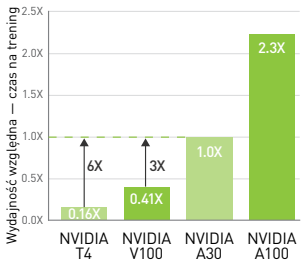
* Z rzadkością

** Most NVLink dla maksymalnie dwóch procesorów graficznych.

Niezwykła wydajność w różnych obciążeniach

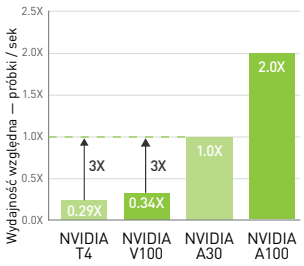
Szkolenie AI - do 3X większa przepustowość niż V100 i 6X wyższa niż T4[1]

BERT - Duże szkolenie wstępne [Znormalizowane]



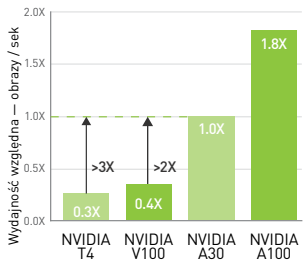
Wnioskowanie AI — do 3 x większa przepustowość w porównaniu z V100 przy konwersacji AI[2] w czasie rzeczywistym

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



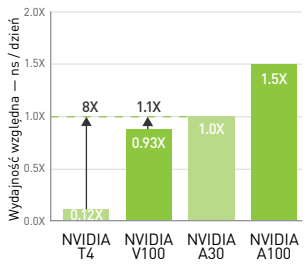
Wnioskowanie AI — ponad 3 x większa przepustowość niż T4 przy klasyfikacji obrazu w czasie rzeczywistym[3]

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



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

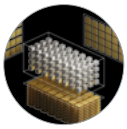
LAMPY [Znormalizowane]



Przełomowe innowacje

ARCHITEKTURA NVIDIA AMPERE

Niezależnie od tego, czy używasz MIG do podziału GPU A30 na mniejsze instancje, czy NVIDIA NVLink do łączenia wielu GPU w celu przyspieszenia większych obciążeń, A30 z łatwością radzi sobie z różnorodnymi potrzebami przyspieszenia, od najmniejszych zadań po największe obciążenia wielowęzłowe. Wszechstronność A30 oznacza, że menedżerowie IT mogą maksymalnie wykorzystać potencjał każdego GPU w swoim centrum danych przy użyciu serwerów głównego nurtu przez całą dobę.



RDZENIE TENSOR TRZECIEJ GENERACJI

NVIDIA A30 zapewnia 165 teraFLOPS (TFLOPS) wydajności w uczeniu głębokim TF32. To oznacza 20 razy większą przepustowość treningową AI i ponad 5 razy wyższą wydajność wnioskowania w porównaniu do GPU NVIDIA T4 Tensor Core. W przypadku obliczeń o wysokiej wydajności (HPC) A30 osiąga 10,3 TFLOPS wydajności, co oznacza niemal 30 procentowy wzrost w porównaniu do GPU NVIDIA V100 Tensor Core.

NASTĘPNA GENERACJA NVLINK

NVIDIA NVLink w A30 zapewnia 2 razy wyższą przepustowość w porównaniu do poprzedniej generacji. Dwa GPU A30 PCIe można połączyć za pomocą mostka NVLink, aby uzyskać 330 TFLOPS wydajności w uczeniu głębokim.

WIELONSTANCYJNY GPU (MIG)

GPU A30 można podzielić na maksymalnie cztery instancje GPU, całkowicie izolowane na poziomie sprzętowym, z własną pamięcią o dużej przepustowości, pamięcią podręczną i rdzeniami obliczeniowymi. MIG zapewnia deweloperom dostęp do przełomowego przyspieszenia dla wszystkich ich aplikacji. A administratorzy IT mogą oferować odpowiednio dostosowane przyspieszenie GPU dla każdego zadania, optymalizując wykorzystanie zasobów i rozszerzając dostęp dla każdego użytkownika i aplikacji.

HBM2

Z maksymalnie 24 GB pamięci o dużej przepustowości (HBM2), A30 zapewnia 933 GB/s przepustowości pamięci GPU, co jest optymalne dla różnorodnych obciążeń AI i HPC w serwerach głównego nurtu.

ZRZADKOŚĆ STRUKTURALNA

Sieci AI mają miliony do miliardów parametrów. Nie wszystkie te parametry są potrzebne do dokładnych prognoz, a niektóre mogą być przekształcone w zera, co sprawia, że modele stają się „rzadkie" bez utraty dokładności. Rdzenie Tensor w A30 mogą zapewnić do 2 razy wyższą wydajność dla modeli rzadkich. Chociaż cecha rzadkości bardziej korzystnie wpływa na wnioskowanie AI, może również poprawić wydajność treningu modeli.

Rozwiązanie End-to-End dla Przedsiębiorstw

GPU NVIDIA A30 Tensor Core – napędzany architekturą NVIDIA Ampere, sercem nowoczesnego centrum danych – jest integralną częścią platformy centrum danych NVIDIA. Zbudowany z myślą o uczeniu głębokim, HPC i analizie danych, platforma przyspiesza ponad 2000 aplikacji, w tym wszystkie główne frameworki uczenia głębokiego. Dodatkowo, NVIDIA AI Enterprise, kompleksowy, natywny w chmurze pakiet oprogramowania do AI i analizy danych, jest certyfikowany do uruchamiania na A30 w wirtualnej infrastrukturze opartej na hypervisorach z VMware vSphere. Umożliwia to zarządzanie i skalowanie obciążeń AI w hybrydowym środowisku chmurowym. Cała platforma NVIDIA jest dostępna wszędzie, od centrum danych po brzeg sieci, oferując zarówno radykalne zyski wydajnościowe, jak i możliwości oszczędności kosztów.

Specyfikacja

SPECYFIKACJE PRODUKTU

Całkowite zużycie energii	165 W
Rozwiązanie termiczne	Pasywne
Mechaniczny format obudowy	Pełny profil, pełna długość (FHFL) 10,5", dwuslotowy
Taktowanie GPU	Bazowe: 930 MHz Boost: 1440 MHz
VBIOS	Rozmiar pamięci EEPROM: 8 Mbit UEFI: Obsługiwane
Interfejs PCI Express	Fizyczne 8 linii PCIe PCIe Gen4 x8, x4; Gen3 x8 Obsługiwane odwracanie linii i polaryzacji
Stany wydajności	P0
Wieloinstancyjny GPU (MiG)	Obsługiwane (do czterech instancji)
Zero Power	Nie obsługiwane
Gotowość NEBS:	Obsługiwana
Złącza i gniazda zasilania	Jedno dodatkowe złącze zasilania CPU 8-pin
Waga	Płyta: 1240 gramów (bez wspornika, przedłużaczy i mostka) Mostek NVLink: 20,5 g Wspornik z wkrętami: 20 g Długi przedłużacz offsetowy: 64 g Prosty przedłużacz: 39 gramów

SPECYFIKACJE PAMIĘCI

Taktowanie pamięci	1215 MHz
Typ pamięci	HBM2
Rozmiar pamięci	24 GB
Szerokość magistrali pamięci	3072 bits
Maksymalna przepustowość pamięci	Do 933 GB/s

SPECYFIKACJE OPROGRAMOWANIA

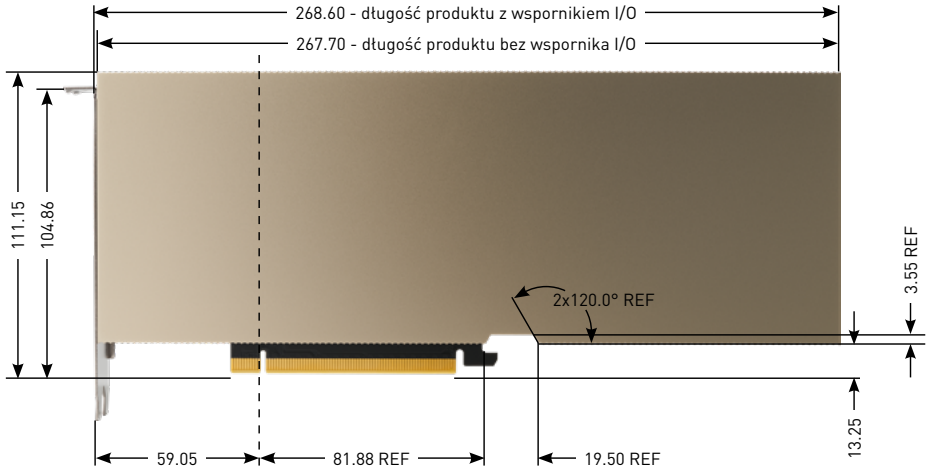
Obsługa SR-IOV	Obsługiwane: 8 VF (funkcji wirtualnych)
Adres BAR (fizyczna funkcja)	BAR0: 16 MiB[1] BAR1: 32 GiB[1] BAR3: 32 MiB[1]
Adres BAR (funkcja wirtualna)	BAR0: 2 MiB (256 KiB na VF)[1] BAR1: 32 GiB, 64-bit (4 GiB na VF)[1] BAR3: 256 MiB, 64-bit (32 MiB na VF)[1]
Przerwania sygnalizowane poprzez wiadomość (MSI)	MSI-X: Obsługiwane MSI: Nieobsługiwane
Multi-Instance GPU (MIG)	Obsługiwane
Przekazywanie ARI	Obsługiwane
Wsparcie sterownika	Linux: R460.65 lub nowszy Windows: R461.98 lub nowszy
Uruchamianie zabezpieczeń	Obsługiwane
Firmware CEC	v6.01 lub nowszy (dla kart z obsługą CEC)
Wsparcie dla NVIDIA® CUDA®	CUDA 11.3 lub nowsze
Wsparcie oprogramowania Virtual GPU	Obsługuje vGPU 13.0 lub nowszy: NVIDIA Virtual Compute Server Edition
Program Systemów Certyfikowanych przez NVIDIA	NVIDIA-Certified Systems™ w wersji 2.5 lub nowszy
NVIDIA AI Enterprise	Obsługiwane z VMware
NVIDIA® NGC-Ready™ Test Suite	Certyfikacja NGC-Next 2.2 lub nowsza
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	Nie obsługiwane
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).



Deep Learning Inference

Procesor graficzny NVIDIA A30 jest specjalnie zoptymalizowany pod kątem głębokiego uczenia się, zaspokajając wymagające potrzeby obliczeniowe wdrożeń sztucznej inteligencji w centrach danych. Zbudowany w oparciu o wydajną architekturę Ampere, A30 jest wyposażony w dużą liczbę rdzeni CUDA i rdzeni Tensor, umożliwiając przetwarzanie sieci neuronowych z wyjątkową szybkością i dokładnością. Ten procesor graficzny doskonale radzi sobie z obciążeniami wnioskowania na dużą skalę w różnych branżach, od przetwarzania języka naturalnego i rozpoznawania obrazów po systemy rekomendacji i pojazdy autonomiczne. Obsługa obliczeń o mieszanej precyzji przez A30 zwiększa wydajność, równoważąc dokładność obliczeń z wydajnością, zapewniając szybkie wyniki wnioskowania bez uszczerbku dla precyzji modelu. Integracja z zestawem narzędzi do optymalizacji wnioskowania TensorRT firmy NVIDIA jeszcze bardziej usprawnia wdrażanie i maksymalizuje przepustowość, ułatwiając przedsiębiorstwom efektywne skalowanie aplikacji AI. Ogólnie rzecz biorąc, procesor graficzny NVIDIA A30 to solidne rozwiązanie dla przedsiębiorstw, które chcą przyspieszyć swoje możliwości wnioskowania w zakresie głębokiego uczenia się, zapewniając doskonałą wydajność i skalowalność w środowiskach opartych na sztucznej inteligencji.

Analiza danych o wysokiej wydajności

Procesor graficzny NVIDIA A30 ma zrewolucjonizować wysokowydajną analizę danych dzięki swoim solidnym możliwościom i wydajnej architekturze Ampere. Dostosowany do wymagających aplikacji wymagających dużej ilości danych, A30 jest wyposażony w bogactwo rdzeni CUDA i rdzeni Tensor, które wyróżniają się przyspieszaniem złożonych zadań analitycznych, takich jak przetwarzanie danych na dużą skalę, uczenie maszynowe i analityka predykcyjna. Duża przepustowość pamięci i obsługa technologii NVIDIA NVLink zapewniają szybki dostęp do ogromnych zbiorów danych i ich przetwarzanie, umożliwiając organizacjom szybkie uzyskiwanie informacji i podejmowanie świadomych decyzji. Wszechstronność A30 obejmuje obsługę obliczeń o mieszanej precyzji, optymalizując wydajność obliczeniową bez utraty dokładności, która jest kluczowa dla wydajnej obsługi różnorodnych obciążeń. Zintegrowany z pakietem narzędzi programowych NVIDIA, takimi jak RAPIDS do przyspieszanych przez procesor graficzny procesów analizy danych i bibliotekami CUDA-X, A30 upraszcza wdrażanie i skalowanie rozwiązań do analizy danych w chmurze hybrydowej i środowiskach lokalnych. Ostatecznie procesor graficzny NVIDIA A30 ustanawia nowy standard w zakresie wysokowydajnej analizy danych, umożliwiając przedsiębiorstwom wydobywanie przydatnych wniosków szybciej i skuteczniej niż kiedykolwiek wcześniej.

High-Performance Computing (HPC)

Procesor graficzny NVIDIA A30 stanowi znaczący postęp w dziedzinie obliczeń o dużej wydajności (HPC), zaprojektowanych w celu zapewnienia niezrównanej mocy obliczeniowej i wydajności w szerokim zakresie zadań obliczeniowych. Zbudowany w oparciu o wydajną architekturę Ampere, A30 zawiera znaczną liczbę rdzeni CUDA i rdzeni Tensor, zoptymalizowanych do obsługi złożonych symulacji naukowych, analiz numerycznych i obliczeń wymagających dużej ilości danych z niezwykłą szybkością i dokładnością. Wysoka przepustowość pamięci i obsługa technologii NVIDIA NVLink umożliwiają bezproblemową komunikację pomiędzy procesorami graficznymi i innymi komponentami systemu, zwiększając ogólną wydajność i skalowalność systemu. Solidne możliwości obliczeniowe A30 sprawiają, że idealnie nadaje się do przyspieszania aplikacji w takich dziedzinach, jak fizyka, chemia, prognozowanie pogody i dynamika molekularna, gdzie krytyczne znaczenie ma szybkie przetwarzanie danych i symulacja. Integracja z platformą obliczeń równoległych i bibliotekami CUDA firmy NVIDIA zapewnia kompatybilność i ułatwia opracowywanie zoptymalizowanych rozwiązań programowych, umożliwiając badaczom i inżynierom skuteczne radzenie sobie z większymi i bardziej złożonymi problemami. Podsumowując, procesor graficzny NVIDIA A30 to potężne rozwiązanie dla środowisk HPC, oferujące niezrównaną wydajność i niezawodność w celu napędzania innowacji i odkryć naukowych.

Szkolenie AI (AI Training)

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.

ZOPTYMALIZOWANE OPROGRAMOWANIE I USŁUGI DLA PRZEDSIĘBIORSTW

NVIDIA CERTIFIED

KAŻDY FRAMEWORK UCZENIA GŁĘBOKIEGO

mxnet PYTORCH Apache Spark TensorFlow

PONAD 2000 APLIKACJI PRZYSPIESZONYCH 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 Fazy 2 = 512, zestaw danych = rzeczywisty, kontener NGC™ = 21.03, 8x GPU: T4 (FP32, BS=8, 2) | V100 PCIE 16GB (FP32, BS=8, 2) | A30 (TF32, BS=8, 2) | A100 PCIE 40GB (TF32, BS=54, 8) | rozmiary wsadu wskazane są dla Fazy 1 i Fazy 2 odpowiednio.

[2] NVIDIA® TensorRT®, Precyzja = INT8, Długość sekwencji = 384, Kontener NGC 20.12, Opóźnienie <10 ms, Zestaw danych = syntetyczny; 1x GPU: A100 PCIe 40GB (BS=8) | A30 (BS=4) | V100 SXM2 16GB (BS=1) | T4 (BS=1).

[3] TensorRT, Kontener NGC 20.12, Opóźnienie <7 ms, Zestaw danych = syntetyczny,; 1x GPU: T4 (BS=31, INT8) | V100 (BS=43, mieszana precyzja) | A30 (BS=96, INT8) | A100 (BS=174, INT8).

[4] Zestaw danych: ReaxFF/C, FP64 | 4x GPU: T4, V100 PCIE 16GB, A30.

FORMAT | NVIDIA