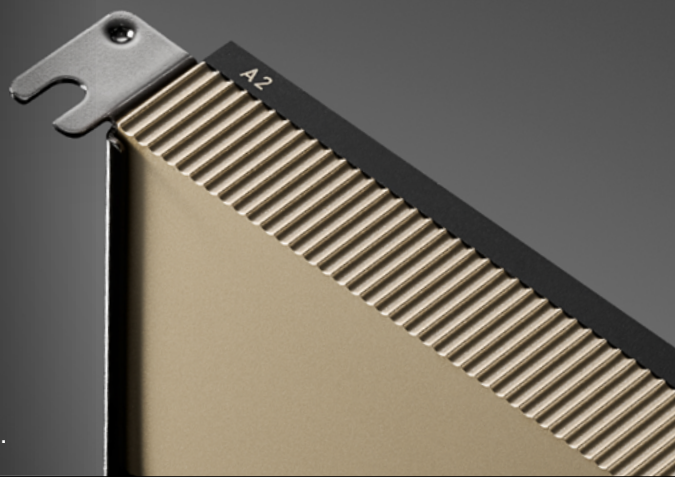
ARKUSZ DANYCH

KARTA NVIDIA A2 Z RDZENIAMI TENSOROWYMI

Karta graficzna podstawowej klasy, dostarczająca sztuczną inteligencję NVIDIA do dowolnego serwera.



Wszechstronne wnioskowanie na poziomie podstawowym

Procesor graficzny NVIDIA A2 Tensor Core zapewnia wnioskowanie na poziomie podstawowym przy niskim poborze mocy, niewielkich rozmiarach i wysokiej wydajności dla **NVIDIA AI** na brzegu sieci. Będąca niskoprofilową kartą PCIe Gen4 i niską konfigurowalną moc cieplną (TDP) wynoszącą 40–60 W (W), **A2** zapewnia adaptacyjne przyspieszenie wnioskowania dla każdego serwera.

Wszechstronność, niewielkie rozmiary i mała moc **A2** przewyższają wymagania dotyczące wdrożeń brzegowych na dużą skalę, umożliwiając natychmiastową modernizację istniejących podstawowych serwerów procesorowych w celu obsługi wnioskowania. Serwery przyspieszane procesorami graficznymi A2 zapewniają wyższą wydajność w porównaniu z procesorami CPU i wydajniejsze wdrożenia inteligentnej analizy wideo (IVA) niż poprzednie generacje procesorów graficznych – a wszystko to w przystępnej cenie.

Systemy z certyfikatem **NVIDIA™** wyposażone w procesory graficzne A2 i sztuczną inteligencję NVIDIA, w tym silnik wnioskowania **NVIDIA Triton™**, zapewniają przełomową wydajność na brzegu, w centrum danych i w chmurze.

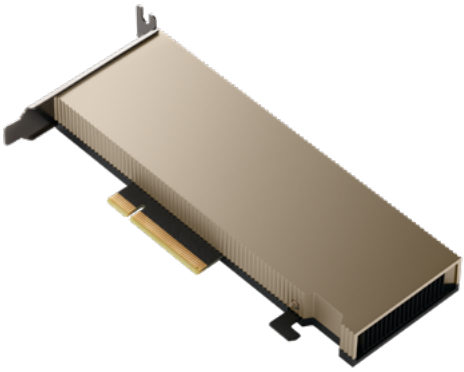
Zapewniają, że aplikacje obsługujące sztuczną inteligencję będą wdrażane z mniejszą liczbą serwerów i mniejszym poborem prądu, co skutkuje łatwiejszymi implementacjami, szybszym wglądem w wyniki i znacznie niższymi kosztami wejściowymi i operacyjnymi.

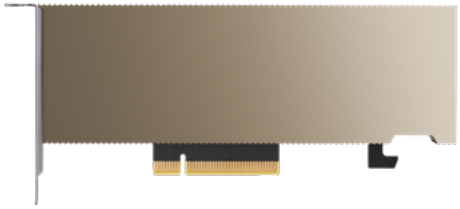
Do 20x wydajniejsze osiągi wnioskowania

Wnioskowanie AI jest wdrażane, aby uczynić życie konsumentów wygodniejszym dzięki doświadczeniom w czasie rzeczywistym i umożliwia im uzyskanie wglądu w biliony czujników i kamer brzegowych. W porównaniu z serwerami wyposażonymi wyłącznie w procesor, serwery zbudowane z procesorem graficznym NVIDIA A2 Tensor Core oferują do 20 razy większą wydajność, umożliwiając natychmiastową modernizację dowolnego serwera pod kątem obsługi nowoczesnych technologii sztucznej inteligencji.

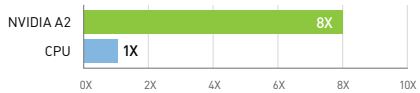
PODSTAWOWE CECHY

- Trzecia generacja rdzeni Tensor NVIDIA
- Druga generacja rdzeni RT
- Wsparcie dla rozproszonych struktur danych
- Sprzętowy Root-of-Trust
- Najlepsza wydajność sprzętowego transkodowania





Wizja komputerowa (EfficientDet-D0)

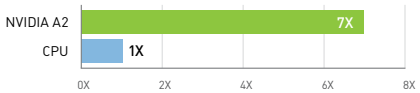


Przyspieszenie wnioskowania

Porównanie jednego procesora graficznego NVIDIA A2 Tensor Core z dwugniazdowym procesorem Xeon Gold 6330N

Konfiguracja systemu: CPU: HPE DL380 Gen10 Plus, 2S Xeon Gold 6330N @2.2GHz, 512GB DDR4 | Computer Vision: EfficientDet-D0 (COCO, 512x512) | TensorRT 8.2, Precision: INT8, BS:8 (GPU) | OpenVINO 2021.4, Precision: INT8, BS:8 (CPU)

Przetwarzanie języka naturalnego (BERT-Large)

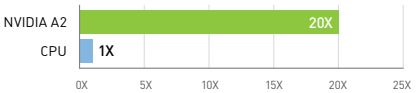


Przyspieszenie wnioskowania

Porównanie jednego procesora graficznego NVIDIA A2 Tensor Core z dwugniazdowym procesorem Xeon Gold 6330N

Konfiguracja systemu: CPU: HPE DL380 Gen10 Plus, 2S Xeon Gold 6330N @2.2GHz, 512GB DDR4 | NLP: BERT-Large (Sequence length: 384, SQuAD: v1.1) | TensorRT 8.2, Precision: INT8, BS:1 (GPU) | OpenVINO 2021.4, Precision: INT8, BS:1 (CPU)

Konwersja tekstu na mowę (Tacotron2 + Waveglow)



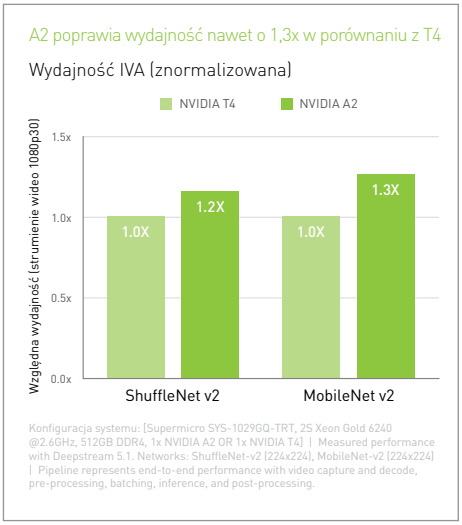
Przyspieszenie wnioskowania

Porównanie jednego procesora graficznego NVIDIA A2 Tensor Core z dwugniazdowym procesorem Xeon Gold 6330N

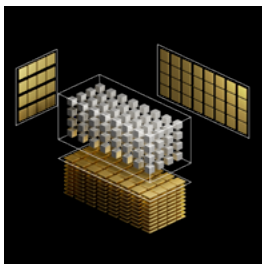
Konfiguracja systemu: CPU: HPE DL380 Gen10 Plus, 2S Xeon Gold 6330N @2.2GHz, 512GB DDR4 | Text-to-Speech: Tacotron2 + Waveglow end-to-end pipeline (input length: 128) | PyTorch 1.9, Precision: FP16, BS:1 (GPU) | PyTorch 1.9, Precision: FP32, BS:1 (CPU)

Wyższa wydajność IVA dla inteligencji brzegowej

Serwery wyposażone w A2 oferują do 1,3 razy większą wydajność w inteligentnych zastosowaniach brzegowych, w tym w inteligentnych miastach, produkcji i handlu detalicznym. Obsługujące obciążenia IVA zapewniają bardziej wydajne wdrożenia, oferując nawet 1,6 razy lepszą wydajność w stosunku do ceny i o dziesięć procent lepszą efektywność energetyczną w porównaniu z poprzednimi generacjami procesorów graficznych.

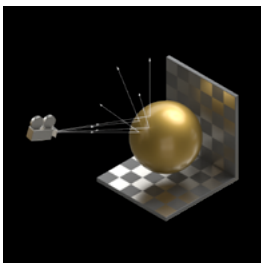


NVIDIA A2 zapewnia przełomową technologię NVIDIA Ampere - Innowacje w architekturze



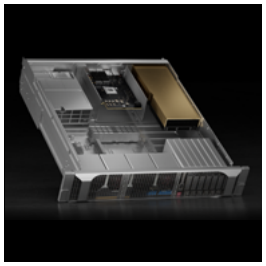
RDZENIE TENSOROWE TRZECIEJ GENERACJI

Rdzenie Tensorowe trzeciej generacji w A2 obsługują matematykę na liczbach całkowitych aż do INT4 i matematykę zmiennoprzecinkową do FP32, zapewniając wysoką wydajność uczenia i wnioskowania AI. Architektura NVIDIA Ampere obsługuje także funkcje TF32 i automatycznej mieszanej precyzji (AMP) firmy NVIDIA.



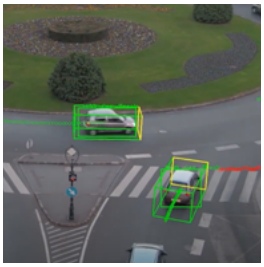
RDZENIE RT DRUGIEJ GENERACJI

A2 zawiera dedykowane rdzenie RT do śledzenia promieni, które umożliwiają przełomowe technologie z przełomową szybkością. Z nawet dwukrotnie większą przepustowością w porównaniu z poprzednią generacją i możliwością jednoczesnego śledzenia promieni z możliwością cieniowania lub usuwania szumów.



RDZEŃ BEZPIECZEŃSTWA ZAUFANIA

Zapewnienie bezpieczeństwa wdrożeń brzegowych i punktów końcowych ma kluczowe znaczenie dla operacji biznesowych przedsiębiorstwa. A2 oferuje bezpieczny rozruch dzięki uwierzytelnianiu zaufanego kodu i wzmocnionym zabezpieczeniom przed wycofywaniem w celu ochrony przed atakami złośliwego oprogramowania.



WYDAJNOŚĆ TRANSKODOWANIA SPRZĘTU

Wykładniczy rozwój aplikacji wideo wymaga skalowalnej wydajności w czasie rzeczywistym, co wymaga najnowszych możliwości sprzętowego kodowania i dekodowania. Procesory graficzne A2 wykorzystują dedykowany sprzęt do pełnego przyspieszenia dekodowania i kodowania wideo dla najpopularniejszych kodeków, w tym dekodowania H.265, H.264, VP9 i AV1.

Kompletne portfolio

NVIDIA dostarcza listę z pełną gamą serwerów certyfikowanych dla swoich rozwiązań, w tym wyposażonych w procesory graficzne Ampere Tensor Core jako silnik wnioskowania napędzający sztuczną inteligencję NVIDIA. Procesory graficzne NVIDIA A2 Tensor Core rozszerzają o wnioskowanie na poziomie podstawowym w niskoprofilowej obudowie do portfolio AI NVIDIA, które obejmuje już procesory graficzne A100 i A30 Tensor Core. A100 charakteryzuje się najwyższą wydajnością wnioskowania w każdej skali, a A30 zapewnia optymalną wydajność wnioskowania dla głównych serwerów. Procesory graficzne NVIDIA A2, NVIDIA A30 i NVIDIA A100 Tensor Core zapewniają wiodącą wydajność wnioskowania na brzegu, w centrach danych i w chmurze.

Zoptymalizowane oprogramowanie i usługi dla przedsiębiorstw

Przedsiębiorstwo oparte na sztucznej inteligencji NVIDIA

NVIDIA AI Enterprise, kompleksowy pakiet oprogramowania do sztucznej inteligencji i analizy danych, natywnie działający w chmurze, posiada certyfikat do działania na platformie NVIDIA A2 w infrastrukturze wirtualnej opartej na hypervisorze z VMware vSphere. Umożliwia to zarządzanie i skalowanie obciążeń AI i wnioskowania w środowisku chmury hybrydowej.

Specyfikacja

SPECYFIKACJE PRODUKTU

Całkowite zużycie energii	60 W domyślnie 60 W maksymalnie 40 W minimalnie
Rozwiązanie termiczne	Pasywne
Mechaniczny format obudowy	HHHL-SS (niski profil, połowa długości, jednoslotowy)
Taktowanie GPU	Bazowe: 1440 MHz Maksymalne przyspieszenie: 1770 MHz
VBIOS	Rozmiar pamięci EEPROM: 16 Mbit UEFI: Obsługiwane
Interfejs PCI Express	Fizyczne 8 linii PCIe PCIe Gen4 x8, x4; Gen3 x8 Obsługiwane odwracanie linii i polaryzacji
Stany wydajności	P0, P8
Zero Power	Nie obsługiwane
Gotowość NEBS:	Obsługiwana
Waga	Płyta: 260 g (bez wspornika) Wspornik (pełny profil) z wkrętami: 14 g Wspornik (niski profil) z wkrętami: 9 g

SPECYFIKACJE PAMIĘCI

Taktowanie pamięci	6251 MHz
Typ pamięci	GDDR6
Rozmiar pamięci	16 GB
Szerokość magistrali pamięci	128 bits
Maksymalna przepustowość pamięci	200 GB/sec

SPECYFIKACJE OPROGRAMOWANIA

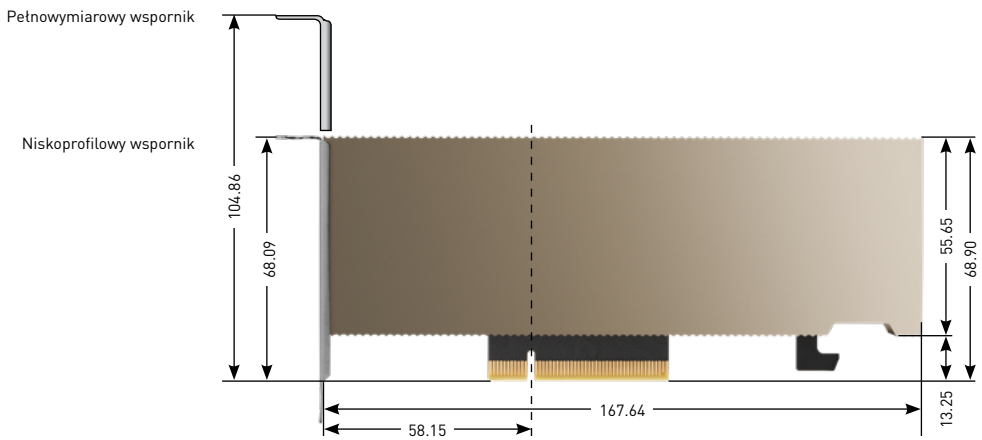
Obsługa SR-IOV	Obsługiwane: 16 VF (funkcji wirtualnych)
Adres BAR (fizyczna funkcja)	BAR0: 16 MiB[1] BAR1: 16 GiB[1] BAR3: 32 MiB[1]
Adres BAR (funkcja wirtualna)	BAR0: 4 MiB (256 KiB na VF)[1] BAR1: 32 GiB, 64-bit (2 GiB na VF)[1] BAR3: 512 MiB, 64-bit (32 MiB na VF)[1]
Przerwania sygnalizowane poprzez wiadomość (MSI)	MSI-X: Obsługiwane MSI: Nieobsługiwane
Przekazywanie ARI	Obsługiwane
Wsparcie sterownika	R470.82 lub nowszy
Uruchamianie zabezpieczeń	Obsługiwane
Firmware CEC	v6.07 lub nowszy (dla kart z obsługą CEC)
Wsparcie dla NVIDIA® CUDA®	CUDA 11.1 lub nowsze
Wsparcie oprogramowania Virtual GPU	Obsługuje vGPU 14.0 lub nowsze
Program Systemów Certyfikowanych przez NVIDIA	NVIDIA-Certified Systems™ w wersji 2.5 lub nowszy
Kod klasy PCI	0x03 – Kontroler wyświetlacza
Kod podklasy PCI	0x02 – Kontroler 3D
Wsparcie ECC	Włączone (domyślnie); można wyłączyć za pomocą oprogramowania
SMBus (adres 8-bitowy)	0x9E (zapis), 0x9F (odczyt)
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).



Inteligentne miasta

Procesor graficzny NVIDIA A2 odgrywa kluczową rolę w zastosowaniach inteligentnych miast, zapewniając solidną moc obliczeniową niezbędną do przetwarzania i analiz danych w czasie rzeczywistym. Ten kompaktowy, ale wydajny procesor graficzny jest dostosowany do obliczeń brzegowych, umożliwiając efektywne wdrożenie w różnorodnych środowiskach miejskich. Niski pobór mocy i wysoka wydajność sprawiają, że idealnie nadaje się do zadań takich jak zarządzanie ruchem, gdzie może przetwarzać ogromne ilości danych wideo z kamer monitorujących w celu optymalizacji przepływu ruchu i zmniejszenia zatorów. Dodatkowo A2 może zwiększyć bezpieczeństwo publiczne poprzez rozpoznawanie twarzy w czasie rzeczywistym i wykrywanie anomalii, zapewniając szybką reakcję na potencjalne zagrożenia. Jego zastosowanie w systemach monitorowania środowiska pozwala również na analizę jakości powietrza i poziomu hałasu, przyczyniając się do zdrowszych warunków życia w miastach. Ogólnie rzecz biorąc, NVIDIA A2 przyspiesza wdrażanie inteligentnej infrastruktury, napędzając ewolucję inteligentniejszych i bardziej responsywnych miast.

Handel detaliczny

Procesor graficzny NVIDIA A2 odgrywa kluczową rolę w rewolucjonizowaniu aplikacji detalicznych poprzez ulepszanie funkcjonalności opartych na sztucznej inteligencji, takich jak wizja komputerowa, analiza klientów i zarządzanie zapasami. Dzięki zaawansowanym funkcjom sztucznej inteligencji A2 może zapewniać analizę wideo w czasie rzeczywistym na potrzeby nadzoru w sklepie, umożliwiając skuteczniejsze zapobieganie stratom i analizę zachowań klientów. Sprzedawcy detaliczni mogą wykorzystać tę technologię do personalizacji doświadczeń zakupowych poprzez dynamiczne oznakowanie cyfrowe i promocje dostosowane do danych demograficznych i wzorców zachowań kupujących. Dodatkowo, solidna moc obliczeniowa A2 ułatwia automatyzację śledzenia zapasów i zarządzania nimi, zapewniając optymalny poziom zapasów i zmniejszając prawdopodobieństwo wyczerpania zapasów lub sytuacji nadmiernych zapasów. Integrując procesory graficzne NVIDIA A2, sprzedawcy detaliczni mogą osiągnąć znaczną poprawę wydajności operacyjnej, zadowolenia klientów i ogólnej wydajności biznesowej.

Produkcja

NVIDIA A2, część linii zaawansowanych procesorów graficznych NVIDIA, w coraz większym stopniu staje się kamieniem węgielnym w zastosowaniach produkcyjnych, wykorzystując swoje potężne możliwości sztucznej inteligencji i uczenia maszynowego w celu zwiększenia wydajności i precyzji. W zautomatyzowanych systemach kontroli jakości solidna moc obliczeniowa A2 umożliwia analizę obrazów produktów w czasie rzeczywistym, identyfikację defektów z dużą dokładnością i zapewnienie stałych standardów jakości. Co więcej, w ramach konserwacji predykcyjnej A2 przetwarza ogromne ilości danych z czujników, aby przewidzieć awarie sprzętu, zanim one wystąpią, redukując przestoje i koszty konserwacji. Jego zastosowanie w systemach zrobotyzowanych zwiększa automatyzację, gdzie zdolność procesora graficznego do przetwarzania złożonych algorytmów i wykonywania zadań głębokiego uczenia się pozwala robotom wykonywać skomplikowane procesy produkcyjne z precyzją i szybkością. Integrując kartę NVIDIA A2 z procesami produkcyjnymi, firmy mogą osiągnąć znaczny postęp w zakresie produktywności, zapewnienia jakości i wydajności operacyjnej.

Wnioskowanie na krawędzi (Edge Inference)

Procesor graficzny NVidia A2 jest idealnym rozwiązaniem do zastosowań związanych z wnioskowaniem brzegowym ze względu na równowagę pomiędzy wydajnością energetyczną i wydajnością. Zaprojektowany specjalnie do obciążeń AI, A2 jest wyposażony w architekturę Ampere, zawierającą zaawansowane rdzenie tensorowe, które przyspieszają głębokie uczenie się i zadania AI. Kompaktowa obudowa i niskie zużycie energii sprawiają, że nadaje się do wdrożenia w środowiskach brzegowych, gdzie kluczowa jest przestrzeń i efektywność energetyczna. A2 umożliwia przetwarzanie w czasie rzeczywistym dużych ilości danych blisko źródła, redukując opóźnienia i wykorzystanie przepustowości w porównaniu z wnioskowaniem opartym na chmurze. Dzięki temu idealnie nadaje się do zastosowań takich jak pojazdy autonomiczne, inteligentne miasta i przemysłowy Internet Rzeczy, gdzie niezbędne jest szybkie podejmowanie decyzji i analityka w czasie rzeczywistym. Dodatkowo A2 obsługuje szeroką gamę frameworków i narzędzi AI, upraszczając integrację i wdrażanie modeli AI na krawędzi.

SYSTEMY Z CERTYFIKATEM NVIDIA

Systemy z certyfikatem NVIDIA z NVIDIA A2 łączą przyspieszenie obliczeniowe i szybką, bezpieczną sieć z systemami wiodących partnerów NVIDIA w konfiguracjach sprawdzonych pod kątem optymalnej wydajności, niezawodności i skali. Dzięki systemom z certyfikatem NVIDIA przedsiębiorstwa mogą śmiało wybierać rozwiązania sprzętowe zoptymalizowane pod kątem wydajności, aby obsługiwać przyspieszone obciążenia obliczeniowe — od komputerów stacjonarnych, przez centra danych, aż po urządzenia brzegowe.

NVIDIA CERTIFIED

Dowiedz się więcej

Aby dowiedzieć się więcej o procesorze graficznym NVIDIA A2 Tensor Core, odwiedź stronę **nvidia.com/a2.**

FORMAT | NVIDIA