Karta katalogowa

NVIDIA HGX H100 i HGX H200

Wiodąca na świecie platforma obliczeniowa AI.

Stworzony specjalnie dla AI i obliczeń wysokowydajnych

AI, złożone symulacje i ogromne zbiory danych wymagają wielu GPU z ekstremalnie szybkimi połączeniami oraz w pełni przyspieszonego stosu oprogramowania. Platforma superkomputerowa NVIDIA HGX™ AI łączy pełną moc procesorów graficznych NVIDIA, NVIDIA NVLink™, sieci NVIDIA oraz w pełni zoptymalizowane stosy oprogramowania AI i HPC, aby zapewnić najwyższą wydajność aplikacji i przyspieszyć czas uzyskiwania informacji.

Niezrównana platforma przyspieszonego obliczania end-to-end

NVIDIA HGX H200 łączy procesory graficzne **H200 Tensor Core** z szybkimi interkonektami, tworząc najpotężniejsze serwery na świecie. Konfiguracje do ośmiu GPU zapewniają bezprecedensowe przyspieszenie, z aż 1,1 terabajtami (TB) pamięci GPU i 38 terabajtami na sekundę (TB/s) łącznej przepustowości pamięci. To w połączeniu z oszałamiającą wydajnością 32 petaFLOPS tworzy najpotężniejszą przyspieszoną platformę serwerową do skalowania dla AI i HPC na świecie.

Zarówno HGX H200, jak i HGX H100 obejmują zaawansowane opcje sieciowe — o prędkości do 400 gigabitów na sekundę (Gb/s) — z wykorzystaniem NVIDIA Quantum-2 InfiniBand i **Spectrum™-X** Ethernet dla najwyższej wydajności AI. HGX H200 i HGX H100 zawierają również jednostki przetwarzania danych **NVIDIA® BlueField®-3** (DPU), które umożliwiają sieci chmurowe, kompozytową pamięć masową, bezpieczeństwo o zerowej ufności i elastyczność obliczeń GPU w hiperskalowych chmurach AI.

Wnioskowanie głębokiego uczenia: Wydajność i wszechstronność

AI rozwiązuje szeroki wachlarz wyzwań biznesowych, korzystając z równie szerokiego wachlarza sieci neuronowych. Doskonały akcelerator wnioskowania AI musi nie tylko dostarczać najwyższą wydajność, ale także zapewniać wszechstronność potrzebną do przyspieszenia działania tych sieci w dowolnej lokalizacji, którą wybiorą klienci, od centrum danych po brzeg sieci.

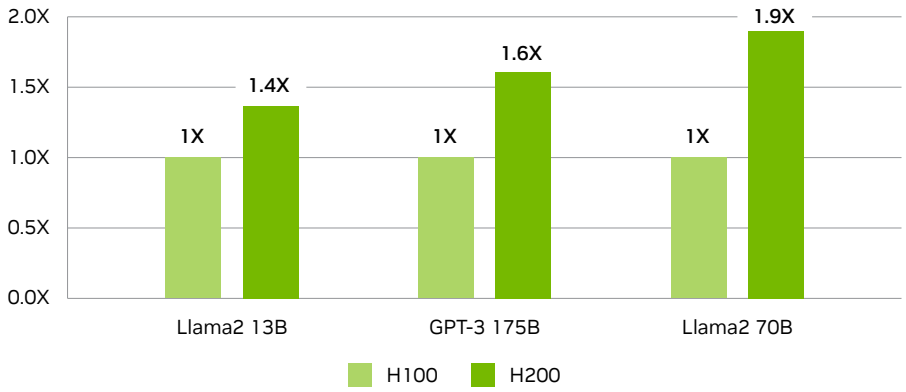
HGX H200 i HGX H100 rozwijają przewodnictwo NVIDIA w zakresie wnioskowania rynkowego.

Kluczowe cechy

NVIDIA HGX H100 i HGX H200

- Silnik Transformer
- Czwarta generacja NVIDIA NVLink
- NVIDIA Confidential Computing
- NVIDIA Multi-Instance GPU (MIG)
- Instrukcje DPX

Do 2X wyższa wydajność wnioskowania LLM

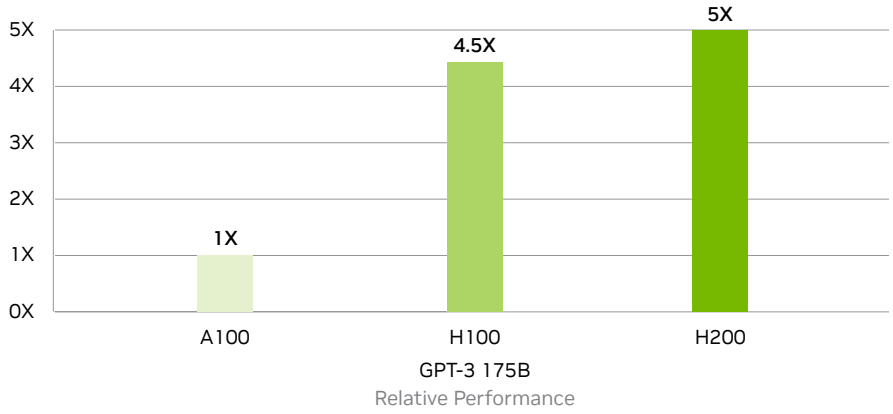


Wstępne specyfikacje. Mogą ulec zmianie.
Llama 2 13B: ISL 128, OSL 2K | Przepustowość | H100 SXM 1x GPU BS 64 | H200 SXM 1x GPU BS 128.
GPT-3 175B: ISL 80, OSL 200 | x8 H100 SXM GPU BS 64 | x8 H200 SXM GPU BS 128.
Llama 2 70B: ISL 2K, OSL 128 | Przepustowość | H100 SXM 1x GPU BS 8 | H200 SXM 1x GPU BS 32.

Trening Głębokiego Uczenia: Wydajność i Skalowalność

Procesory graficzne NVIDIA H200 i H100 wyposażone są w silnik Transformer z precyzją FP8, który zapewnia do 5 razy szybszy trening w porównaniu do poprzedniej generacji GPU dla dużych modeli językowych. Połączenie czwartej generacji NVLink, które oferuje 900 GB/s przepustowości między GPU, PCIe Gen5 oraz oprogramowania NVIDIA Magnum IO™ zapewnia efektywną skalowalność, od małych przedsiębiorstw po ogromne zintegrowane klastry GPU. Te postępy infrastrukturalne, działające w tandemie z pakietem oprogramowania NVIDIA AI Enterprise, czynią HGX H200 i HGX H100 wiodącą platformą obliczeniową AI na świecie.

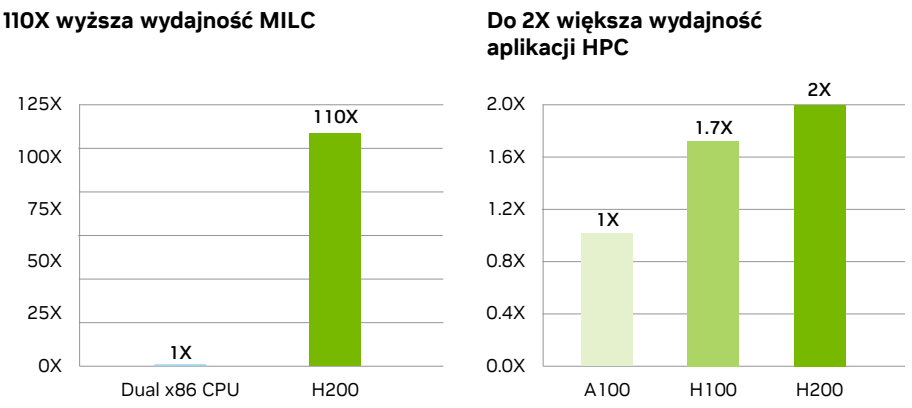
Do 5 razy szybszy trening w Skali



Prognozowana wydajność, może ulec zmianie.
Trening GPT-3 175B na klastrze GPU Tensor Core NVIDIA A100: sieć NVIDIA Quantum InfiniBand, klaster H100: sieć NVIDIA Quantum-2 InfiniBand.

Przyspiesz wydajność HPC

Przepustowość pamięci jest kluczowa dla aplikacji obliczeń wysokowydajnych, ponieważ umożliwia szybszy transfer danych i redukuje złożone wąskie gardła przetwarzania. Dla aplikacji HPC intensywnie korzystających z pamięci, takich jak symulacje, badania naukowe i sztuczna inteligencja, wyższa przepustowość pamięci H200 zapewnia, że dane mogą być efektywnie dostępne i przetwarzane, co prowadzi do uzyskania wyników do 110 razy szybciej w porównaniu do CPU.



Prognozowana wydajność, może ulec zmianie.
Zbiór danych HPC MILC NERSC Apex Medium | HGX H200 4-GPU | podwójny Sapphire Rapids 8480
Aplikacje HPC - CP2K: zbiór danych H2O-32-RI-dRPA-96points | GROMACS: zbiór danych STMV | ICON: zbiór danych r2b5 | MILC: zbiór danych NERSC Apex Medium | Chroma: zbiór danych HMC Medium | Quantum Espresso: zbiór danych AUSURF112 | 1x H100 | 1x H200.

Przyspieszanie HGX z siecią NVIDIA

Centrum danych stało się nową jednostką obliczeniową, a sieć odgrywa integralną rolę w skalowaniu wydajności aplikacji w tym obszarze. W połączeniu z NVIDIA Quantum InfiniBand, HGX zapewnia wydajność i efektywność klasy światowej, co zapewnia pełne wykorzystanie zasobów obliczeniowych.

Dla centrów danych AI w chmurze, które wdrażają Ethernet, HGX najlepiej działa z platformą sieciową NVIDIA Spectrum-X, która zapewnia najwyższą wydajność AI przez Ethernet. Obejmuje przełączniki Spectrum-X i jednostki przetwarzania danych BlueField-3, co zapewnia optymalne wykorzystanie zasobów i izolację wydajności, dostarczając stałe, przewidywalne wyniki dla tysięcy równoczesnych zadań AI na każdym poziomie. Spectrum-X umożliwia zaawansowane zarządzanie wieloma użytkownikami w chmurze oraz bezpieczeństwo o zerowej ufności. Jako projekt referencyjny, NVIDIA zaprojektowała Israel-1, hiperskalowy superkomputer generatywnej AI zbudowany na serwerach Dell PowerEdge XE9680 opartych na platformie NVIDIA HGX 8-GPU, jednostkach przetwarzania danych BlueField-3 oraz przełącznikach Spectrum-4.

Specyfikacje techniczne				
	HGX H200 4-GPU	**HGX H200 8-GPU**	**HGX H100 4-GPU**	**HGX H100 8-GPU**
Format	4x NVIDIA H200 SXM	8x NVIDIA H200 SXM	4x NVIDIA H100 SXM	8x NVIDIA H100 SXM
Rdzeń Tensor FP8*	16 PFLOPS	32 PFLOPS	16 PFLOPS	32 PFLOPS
Rdzeń Tensor INT8*	16 POPS	32 POPS	16 POPS	32 POPS
Rdzeń Tensor FP16/BFLOAT16*	8 PFLOPS	16 PFLOPS	8 PFLOPS	16 PFLOPS
Rdzeń Tensor TF32*	4 PFLOPS	8 PFLOPS	4 PFLOPS	8 PFLOPS
FP32	270 TFLOPS	540 TFLOPS	270 TFLOPS	540 TFLOPS
FP64	140 TFLOPS	270 TFLOPS	140 TFLOPS	270 TFLOPS
Rdzeń Tensor FP64	270 TFLOPS	540 TFLOPS	270 TFLOPS	540 TFLOPS
Pamięć	564 GB HBM3e	1.1 TB HBM3e	320 GB HBM3	640 GB HBM3
Przepustowość łączy GPU	19 GB/s	38 GB/s	13 GB/s	27 GB/s
NVLink	Czwarta generacja	Czwarta generacja	Czwarta generacja	Czwarta generacja
NVSwitch™	N/A	Trzecia generacja	N/A	Trzecia generacja
Przepustowość NVSwitch GPU do GPU	N/A	900 GB/s	N/A	900 GB/s
Całkowita przepustowość agregowana	3.6 TB/s	7.2 TB/s	3.6 TB/s	7.2 TB/s

* Przy rozrzedzeniu.

Gotowy, aby zacząć?

Aby dowiedzieć się więcej o NVIDIA HGX, odwiedź **www.nvidia.com/hgx**

FORMAT
IMPOSSIBLE BECOMES REALITY

NVIDIA