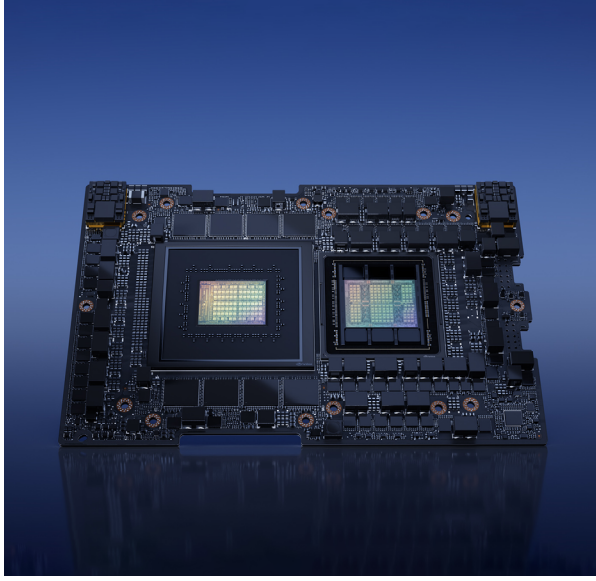
Karta katalogowa



NVIDIA GH200 Grace Hopper Superchip

Przełomowy procesor do wielkoskalowych aplikacji AI i HPC

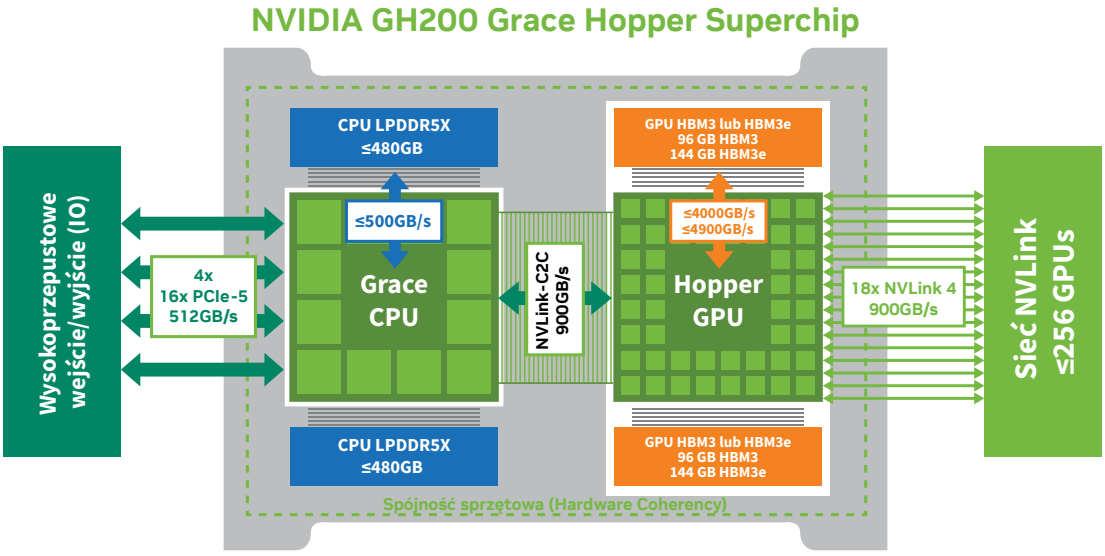
Najbardziej wszechstronna platforma obliczeniowa na świecie.

Architektura NVIDIA Grace Hopper™ łączy niezwykłą wydajność GPU NVIDIA Hopper™ z wszechstronnością CPU NVIDIA Grace™ w jednym superchipie, połączonych wysokoprzepustowym, spójnym z pamięcią interfejsem NVIDIA® NVLink®-C2C Chip-2-Chip (C2C).

NVIDIA NVLink-C2C to interfejs o dużej przepustowości, niskim opóźnieniu i spójności pamięci, przeznaczony dla superchipów. To serce GH200, zapewniające do 900 GB/s łącznej przepustowości, czyli 7 razy więcej niż PCIe Gen5 używane w systemach przyspieszających. NVLink-C2C umożliwia aplikacjom nadrezerwację pamięci GPU i bezpośrednie korzystanie z pamięci CPU Grace z wysoką przepustowością. Z do 480 GB pamięci LPDDR5X CPU na superchipie GH200, GPU ma dostęp do 7 razy więcej szybkiej pamięci niż HBM3, albo prawie 8 razy więcej niż HBM3e, w zależności od konfiguracji pamięci. GH200 można łatwo wdrożyć w standardowych serwerach do obsługi różnych zadań inferencyjnych, analizy danych oraz innych obciążeń obliczeniowych i pamięciowych. Dodatkowo, GH200 może być połączony z systemem NVIDIA NVLink Switch System, z do 256 GPU połączonych NVLink, mających dostęp do aż 144 TB pamięci przy wysokiej przepustowości.

Kluczowe cechy

- 72-rdzeniowy procesor NVIDIA Grace CPU
- GPU NVIDIA H100 Tensor Core
- Do 480 GB pamięci LPDDR5X z korekcją błędów (ECC)
- Obsługa do 96 GB HBM3 lub do 144 GB HBM3e
- Do 624 GB szybkiej pamięci dostępnej z najwyższą przepustowością
- NVLink-C2C: 900 GB/s spójnej pamięci



Moc i wydajność z procesorem Grace.

Procesor NVIDIA Grace zapewnia dwukrotnie wyższą wydajność na wat w porównaniu z konwencjonalnymi platformami x86-64 i jest najszybszym CPU dla centrów danych opartym na Arm®. Został zaprojektowany z myślą o wysokiej wydajności w pojedynczych wątkach, dużej przepustowości pamięci i doskonałych możliwościach przesyłu danych. Procesor NVIDIA Grace łączy 72 rdzenie Neoverse V2 ARMv9 z do 480 GB pamięci LPDDR5X klasy serwerowej z ECC. Ten design osiąga optymalny balans między przepływem danych, efektywnością energetyczną, pojemnością i kosztem. W porównaniu do ośmiokanałowego systemu DDR5, pamięć LPDDR5X w procesorze Grace zapewnia do 53% większą przepustowość przy jednej ósmej mocy na gigabajt na sekundę.

Wydajność i szybkość z GPU Hopper H100

GPU NVIDIA H100 Tensor Core to dziewiąta generacja GPU dla centrów danych firmy NVIDIA, oferująca dziesięciokrotny skok wydajności w zastosowaniach AI i HPC względem poprzedniej generacji, NVIDIA A100 Tensor Core. Oparty na nowej architekturze Hopper, H100 zawiera liczne innowacje:

- Czwarte generacje Tensor Cores wykonują szybsze obliczenia macierzowe na jeszcze szerszym zakresie zadań AI i HPC.
- Nowy Transformer Engine pozwala na przyspieszenie treningu AI do 9 razy i inferencji AI do 30 razy w porównaniu z poprzednią generacją GPU.
- Secure Multi-Instance GPU (MIG) dzieli GPU na odizolowane, odpowiednio dopasowane instancje, maksymalizując jakość usług (QoS) dla mniejszych obciążeń.

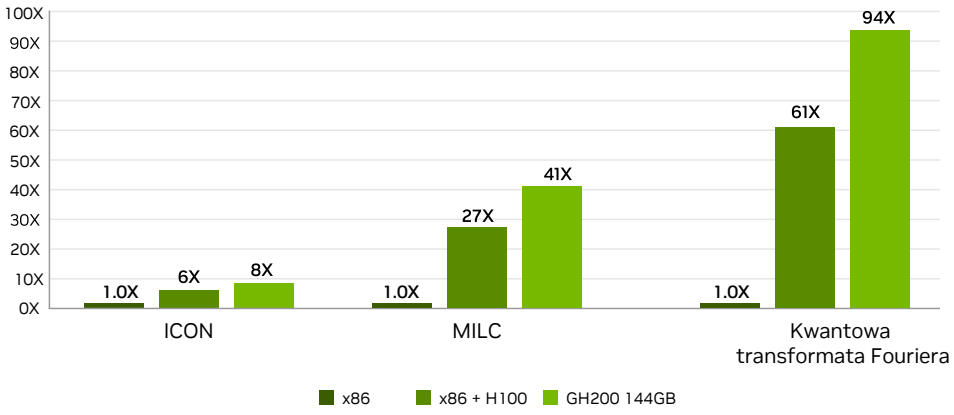
Siła spójnej pamięci

Pamięć NVLink-C2C zwiększa produktywność deweloperów, wydajność i ilość dostępnej pamięci GPU. Wątki CPU i GPU mogą jednocześnie i transparentnie korzystać zarówno z pamięci CPU, jak i GPU, co pozwala skupić się na algorytmach, zamiast na ręcznym zarządzaniu pamięcią. Spójność pamięci oznacza, że deweloperzy mogą przesyłać jedynie potrzebne dane, bez konieczności migracji całych stron pamięci. Umożliwia to także lekkie, niskopoziomowe synchronizacje między wątkami GPU i CPU, dzięki obsłudze natywnych operacji atomowych. Czwarta generacja NVLink umożliwia dostęp do pamięci peer z poziomu bezpośrednich loadów, store'ów i operacji atomowych, co pozwala na większe rozwiązywanie problemów w obliczeniach przyspieszonych.

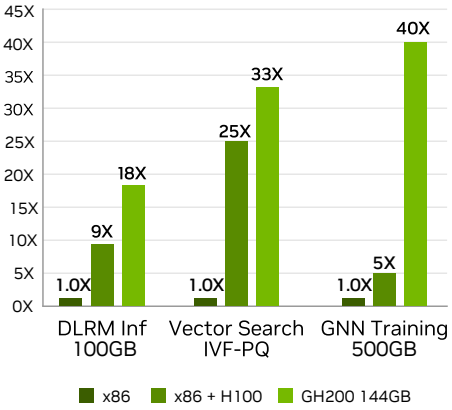
Przewaga w obciążeniach HPC i AI

Superchip GH200 Grace Hopper jest pierwszą prawdziwie heterogeniczną platformą przyspieszoną dla obciążeń HPC i AI. Umożliwia uruchamianie dowolnej aplikacji, korzystając z obu GPU i CPU, zapewniając najprostszy i najbardziej produktywny model programowania heterogenicznego do tej pory, co pozwala naukowcom i inżynierom skupić się na rozwiązywaniu najważniejszych problemów. W przypadku zadań inferencyjnych, superchip GH200 łączy się z technologiami sieci NVIDIA, aby zapewnić najlepszy TCO (całkowity koszt posiadania) dla rozwiązań skalowalnych, obsługując większe zbiorniki danych, bardziej złożone modele i nowe obciążenia, korzystając z do 624 GB szybkiej pamięci. Superchip NVIDIA GH200 w konfiguracji podwójnej (dual-GH200) zawiera dwa superchipy Grace Hopper, połączone w pełni NVLink, co zapewnia 288 GB HBM3e i 1,2 TB szybkiej pamięci, umożliwiając obsługę zarówno obciążeń obliczeniowych, jak i pamięciowych na jeszcze większą skalę.

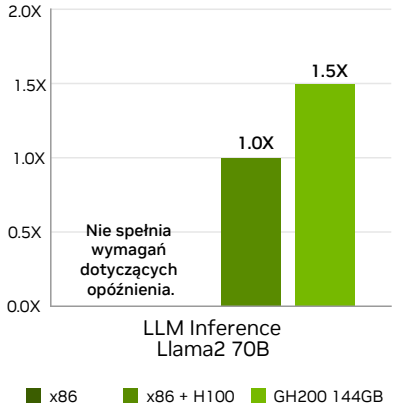
Wydajność GH200 w zastosowaniach HPC



Wydajność GH200 w zastosowaniach AI



Wydajność GH200 w obciążeniach Large Language Models (LLM)



* Porównanie: 2S Xeon Platinum 8480+, Xeon Platinum 8480+ i H100 Tensor Core GPU oraz NVIDIA GH200 144GB: ICON (poziomy QUBICC 191, promieniowanie 80 km), MILC (APEX Medium), Transformata Fouriera kwantowa, inferencja DLRM 100GB, wyszukiwanie wektorów (batch = 10 000 | zapytania = 10 000 na ponad 85 mln wektorów IVF-PQ), GNN (GraphSage OGB-100M, zbiór danych paperów), Llama2 70B (batch = 64 - GH200, 96 - 2x H100s | precyzja = FP8 | TensorRT-LLM — przepustowość na GPU).

Wyniki mogą ulec zmianie.

Pełne wsparcie platformy NVIDIA

Superchip NVIDIA GH200 Grace Hopper rozszerza istniejący, duży i zróżnicowany ekosystem procesorów Arm 64-bit. Te same kontenery, binaria aplikacji i systemy operacyjne, które działają na innych produktach Arm, uruchamiają się na Grace Hopper bez modyfikacji — tylko szybciej.

Dla klientów chcących korzystać z i rozwijać oprogramowanie NVIDIA, Superchip NVIDIA Grace Hopper jest obsługiwany przez pełny stos oprogramowania NVIDIA, w tym platformy NVIDIA HPC, NVIDIA AI i NVIDIA Omniverse™.

Specyfikacja produktu

Funkcja	GH200	GH200 NVL2
Ilość rdzeni CPU	72 Arm Neoverse V2 cores	144 Arm Neoverse V2 cores
L1 cache	64KB i-cache + 64KB d-cache	
L2 cache	1MB per core	
L3 cache	114MB	228MB
Częstotliwość bazowa | pojedyncza instrukcja na wszystkie rdzenie (SIMD)	3.1GHz | 3.0GHz	
Wielkość pamięci LPDDR5X	480GB 120GB, 240GB	960GB 240GB, 480GB
Przepustowość pamięci	Do 384GB/s Do 512GB/s	Do 768GB/s Do 1024GB/s
PCIe links	Do 4x PCIe x16 (Gen5)	Do 8x PCIe x16 (Gen5)

Funkcja	GH200	GH200 NVL2
FP64	34 teraFLOPS	68 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	134 teraFLOPS
FP32	67 teraFLOPS	134 teraFLOPS
TF32 Tensor Core	989 teraFLOPS* | 494 teraFLOPS	1979 teraFLOPS* | 990 teraFLOPS
BFLOAT16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS	3958 teraFLOPS* | 1979 teraFLOPS
FP16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS	3958 teraFLOPS* | 1979 teraFLOPS
FP8 Tensor Core	3,958 teraFLOPS* | 1,979 teraFLOPS	7916 teraFLOPS* | 3958 teraFLOPS
INT8 Tensor Core	3,958 TOPS* | 1,979 TOPS	7916 TOPS* | 3958 TOPS
Rozmiar wysokowydajnej pamięci (HBM)	96GB HBM3 | 144GB HBM3e	Up to 288GB HBM3e
Przepustowość pamięci	Do 4TB/s | Up to 4.9TB/s	Do 9.8TB/s
NVIDIA NVLink-C2C (pasmo CPU-GPU)	900 GB/s	900 GB/s
Zasilanie	Konfigurowalna od 450 do 1000 W (pamięć + CPU + GPU)	Konfigurowalna od 900 do 2000 W (pamięć + CPU + GPU)
Rozwiązanie termiczne	Chłodzenie powietrzem lub cieczą	

* Z zastosowaniem sparsity

Gotowy, aby zacząć?

Aby dowiedzieć się więcej o superchipie NVIDIA Grace Hopper, odwiedź: **nvidia.com/grace-hopper-superchip/**

Aby uzyskać więcej informacje, prosimy o kontakt z naszym przedstawicielem handlowym NVIDIA na: **FORMAT.COM.PL**

FORMAT
IMPOSSIBLE BECOMES REALITY

NVIDIA