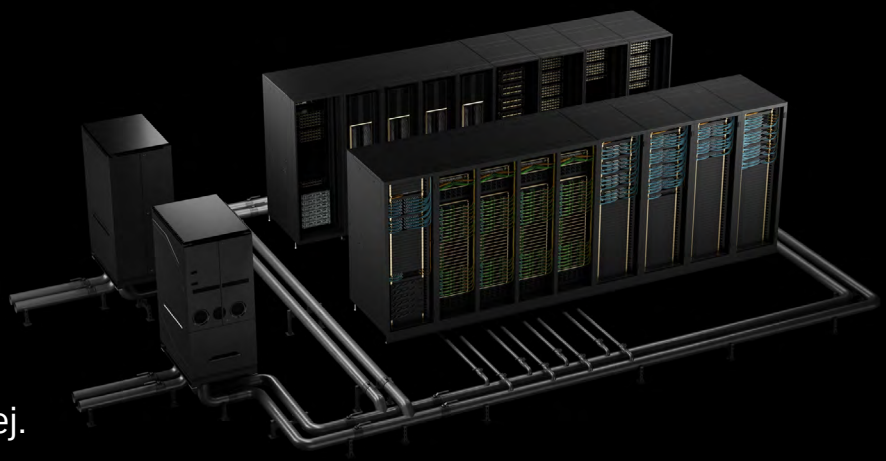
Karta katalogowa



NVIDIA Blackwell

Silnik nowej rewolucji przemysłowej.

Przełamywanie barier w przyspieszonym obliczaniu

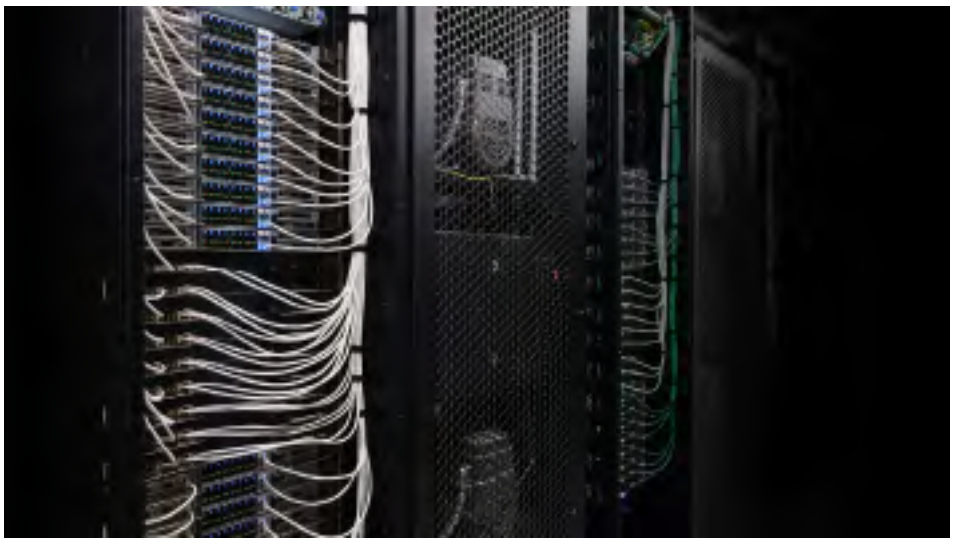
Architektura NVIDIA Blackwell wprowadza przełomowe innowacje dla generatywnej AI i przyspieszonego obliczania. Włączenie drugiej generacji Transformera, wraz z szybszym i szerszym interfejsem **NVIDIA NVLink™**, wprowadza centra danych w nową erę, oferując wielokrotnie większą wydajność w porównaniu do poprzedniej generacji architektury. Dalsze postępy w technologii NVIDIA Confidential Computing podnoszą poziom bezpieczeństwa dla inferencji LLM w czasie rzeczywistym na dużą skalę bez kompromisów w wydajności. Nowy silnik dekompresji Blackwell w połączeniu z bibliotekami Spark RAPIDS™ zapewnia niezrównaną wydajność baz danych, wspierając aplikacje analityki danych. Wielorakie ulepszenia Blackwell opierają się na generacjach technologii przyspieszonego obliczania, definiując kolejną fazę generatywnej AI, oferując niezrównaną wydajność, efektywność i skalę.

Kluczowe oferty

- NVIDIA GB200 NVL72
- NVIDIA HGX B200

NVIDIA GB200 NVL72

Napędza nową erę obliczeń



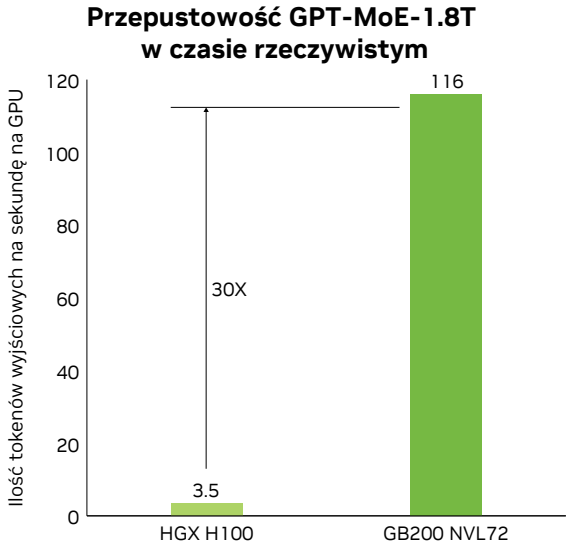
Odkrywanie możliwości inferencji dużych modeli w czasie rzeczywistym z trylionem parametrów

NVIDIA GB200 NVL72 łączy 36 procesorów Grace i 72 GPU Blackwell w konstrukcji z chłodzeniem cieczą, zorganizowanej na poziomie racka, połączonych za pomocą NVIDIA NVLink. Działa jako jeden, ogromny GPU i zapewnia 30 razy szybszą inferencję dużego modelu językowego (LLM) z trylionem parametrów w czasie rzeczywistym.

Superchip NVIDIA GB200 Grace Blackwell jest kluczowym elementem GB200 NVL72, łączącym dwa wysokowydajne GPU Blackwell z procesorem NVIDIA Grace za pomocą interfejsu NVLink-C2C.

Inferencja LLM w czasie rzeczywistym

GB200 NVL72 wprowadza zaawansowane możliwości oraz drugą generację Transformer Engine, umożliwiającą AI FP4. Ta innowacja była możliwa dzięki nowej generacji Tensor Cores, wprowadzającej nowe mikroskali-formaty, zapewniające wysoką precyzję i większą przepustowość. Dodatkowo, GB200 NVL72 korzysta z NVLink i chłodzenia cieczą, tworząc jeden, masywny rack z 72 GPU, który może pokonać wąskie gardła komunikacyjne.



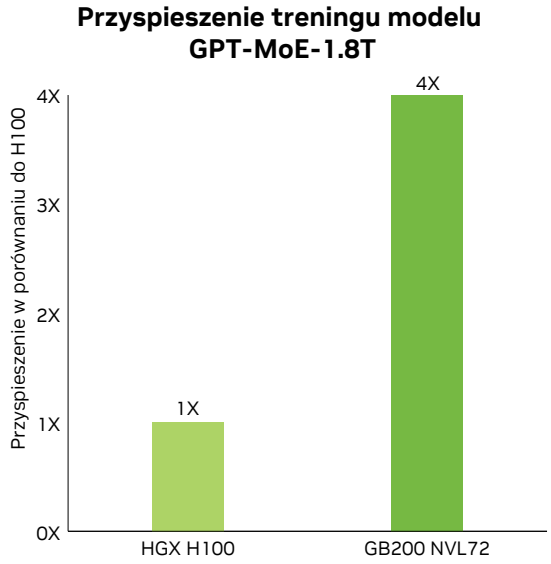
Przewidywana wydajność, może ulec zmianie. Inferencja dużych modeli językowych (LLM) i efektywność energetyczna: opóźnienie token do tokena (TTL) = 50 ms w czasie rzeczywistym, opóźnienie pierwszego tokena (FTL) = 5 s, wejście 32 768, wyjście 1 024. Porównanie: NVIDIA HGX™ H100 skalowane przez InfiniBand (IB) versus GB200 NVL72.

Najważniejsze cechy NVIDIA GB200 NVL72

- 36 procesorów NVIDIA Grace™
- 72 GPU NVIDIA Blackwell
- Do 17 terabajtów (TB) pamięci LPDDR5X z korekcją błędów (ECC)
- Obsługa do 13,5 TB pamięci HBM3E
- Do 30,5 TB szybkiego dostępu do pamięci
- Domena NVLink: 130 TB/s niskolatencyjnej komunikacji GPU

Masowe szkolenie na dużą skalę

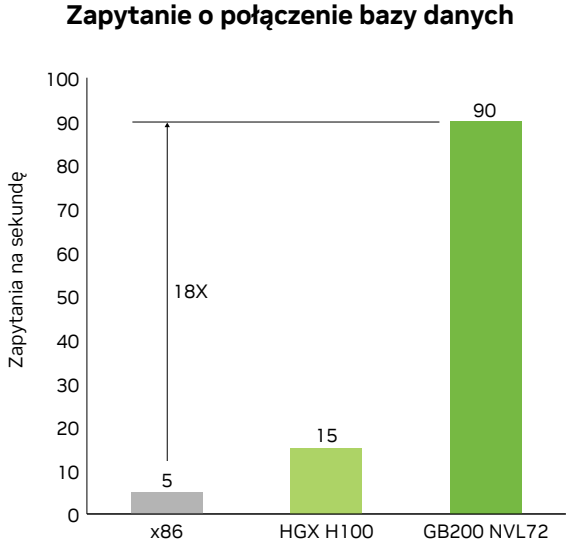
GB200 NVL72 zawiera szybszą drugą generację Transformer Engine, wyposażoną w precyzję FP8 (8-bitowe punkty zmiennoprzecinkowe), co umożliwia aż 4-krotne przyspieszenie treningu dużych modeli językowych na dużą skalę. To przełomowe rozwiązanie uzupełnia piątą generację NVLink, która zapewnia 1,8 TB/s interkonektu GPU-do-GPU, sieci InfiniBand oraz oprogramowanie NVIDIA Magnum IO™.



Trening GPT-MoE-1.8T — 4096x HGX H100 skalowany przez IB vs. 456x GB200 NVL72 skalowany przez IB. Rozmiar klastra: 32 768.

Przetwarzanie danych

Bazy danych odgrywają kluczową rolę w obsłudze, przetwarzaniu i analizie dużych ilości danych dla przedsiębiorstw. GB200 NVL72 wykorzystuje wysokopasmową pamięć, NVLink-C2C oraz dedykowane silniki dekompresji w architekturze NVIDIA Blackwell, aby przyspieszyć kluczowe zapytania bazodanowe aż 18 razy w porównaniu do CPU, zapewniając 5 razy lepszy całkowity koszt posiadania (TCO).

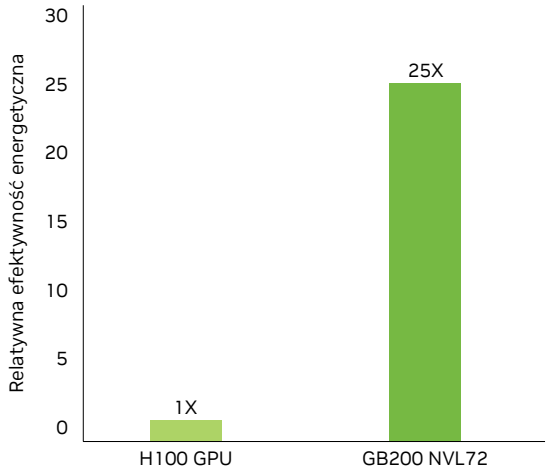


Przewidywana wydajność, może ulec zmianie. Przepustowość zapytań łączenia baz danych w porównaniu: GB200 NVL72, 72x H100 oraz 72x CPU x86.

Infrastruktura energooszczędna

Szafy z chłodzeniem cieczą GB200 NVL72 redukują ślad węglowy i zużycie energii centrum danych. Chłodzenie cieczą zwiększa gęstość obliczeń, zmniejsza powierzchnię zajmowaną na podłodze i wspiera wysokoprzepustową, niskolatencyjną komunikację GPU w dużych architekturach domen NVLink. W porównaniu do infrastruktury NVIDIA H100 chłodzonej powietrzem, GB200 NVL72 oferuje 25 razy większą wydajność przy tej samej mocy, jednocześnie zmniejszając zużycie wody.

Efektywność energetyczna



Przewidywana wydajność, może ulec zmianie. Oszczędność energii w porównaniu do 65 racks z ośmioma HGX H100 chłodzonym powietrzem versus 1 rack GB200 NLV72 chłodzony cieczą, przy równoważnej przepustowości inferencji w czasie rzeczywistym GPT MoE 1.8T.

NVIDIA HGX B200

Napędzając centrum danych w nową erę przyspieszonego obliczania

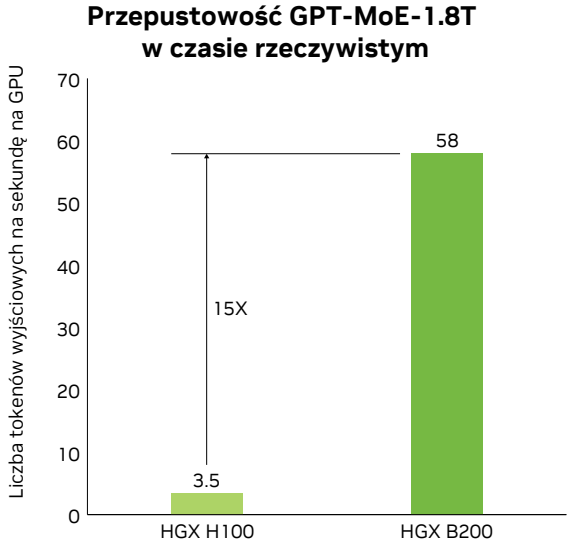
NVIDIA HGX® B200 napędzając centrum danych w nową erę przyspieszonego obliczania i generatywnej AI, integruje GPU NVIDIA Blackwell z szybkim interfejsem, aby podnosić wydajność AI na dużą skalę. Jako wiodąca platforma x86 do skalowania przyspieszeń, z możliwością osiągnięcia do 15 razy szybszej inferencji w czasie rzeczywistym, 12 razy niższych kosztów i 12 razy mniejszego zużycia energii, HGX B200 jest stworzony dla najbardziej wymagających obciążeń AI, analizy danych i obliczeń wysokiej wydajności (HPC).

Najważniejsze cechy HGX B200

- 8 GPU NVIDIA Blackwell
- Do 1,4 terabajta (TB) pamięci HBM3E
- 1800 GB/s NVLink między GPU za pośrednictwem układu NVSwitch™
- 15 razy szybsza inferencja dużych modeli językowych (LLM) w czasie rzeczywistym
- 3 razy szybsza wydajność treningu

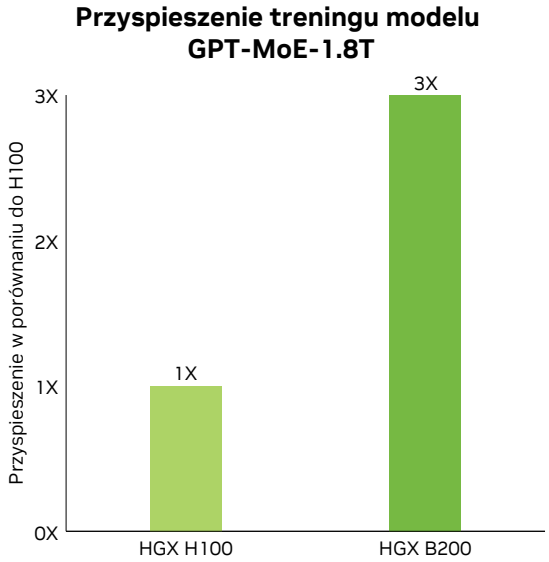
Inferencja w czasie rzeczywistym dla kolejnej generacji dużych modeli językowych

HGX B200 osiąga nawet 15 razy wyższą wydajność inferencji w porównaniu do poprzedniej generacji NVIDIA Hopper™ dla masowych modeli takich jak GPT MoE 1.8T. Druga generacja Transformer Engine wykorzystuje niestandardową technologię Tensor Core NVIDIA Blackwell w połączeniu z innowacjami TensorRT™-LLM i frameworku NVIDIA NeMo™, co przyspiesza inferencję modeli LLM i modeli typu mixture-of-experts (MoE).



Przewidywana wydajność, może ulec zmianie. Opóźnienie token do tokena (TTL) = 50 ms w czasie rzeczywistym, opóźnienie pierwszego tokena (FTL) = 5 s, długość wejściowej sekwencji = 32 768, długość wyjściowej sekwencji = 1 028. Porównanie wydajności na GPU: 8-krotne HGX H100 z chłodzeniem powietrznym versus 1x HGX B200 z chłodzeniem powietrznym.Wydajność treningu na wyższym poziomie

Druga generacja Transformer Engine, wykorzystująca FP8 i nowe precyzje, umożliwia 3-krotnie szybszy trening dużych modeli językowych, takich jak GPT MoE 1.8T. To przełomowe rozwiązanie jest wspierane przez piątą generację NVLink z 1,8 TB/s interkonektu GPU-do-GPU, układ NVSwitch, sieci InfiniBand oraz oprogramowanie NVIDIA Magnum IO. Razem zapewniają one efektywną skalowalność dla przedsiębiorstw i rozległych klastrów obliczeniowych GPU.

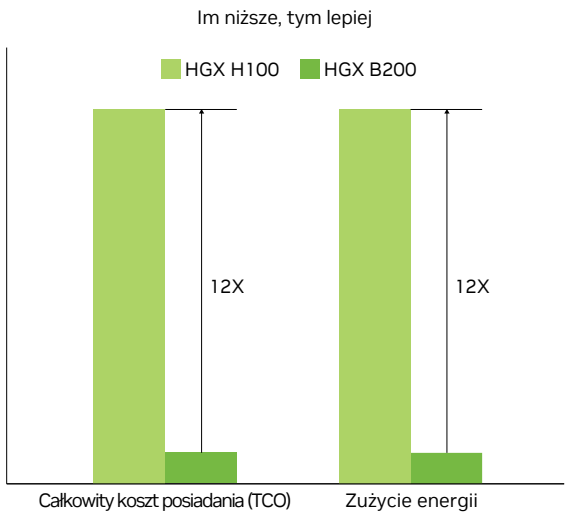


Przewidywana wydajność, może ulec zmianie. Skala 32 768 GPU, klaster złożony z 4 096 węzłów HGX H100 chłodzonych powietrzem, sieć IB 400G, oraz klaster złożony z 4 096 węzłów HGX B200 chłodzonych powietrzem, sieć IB 400G.

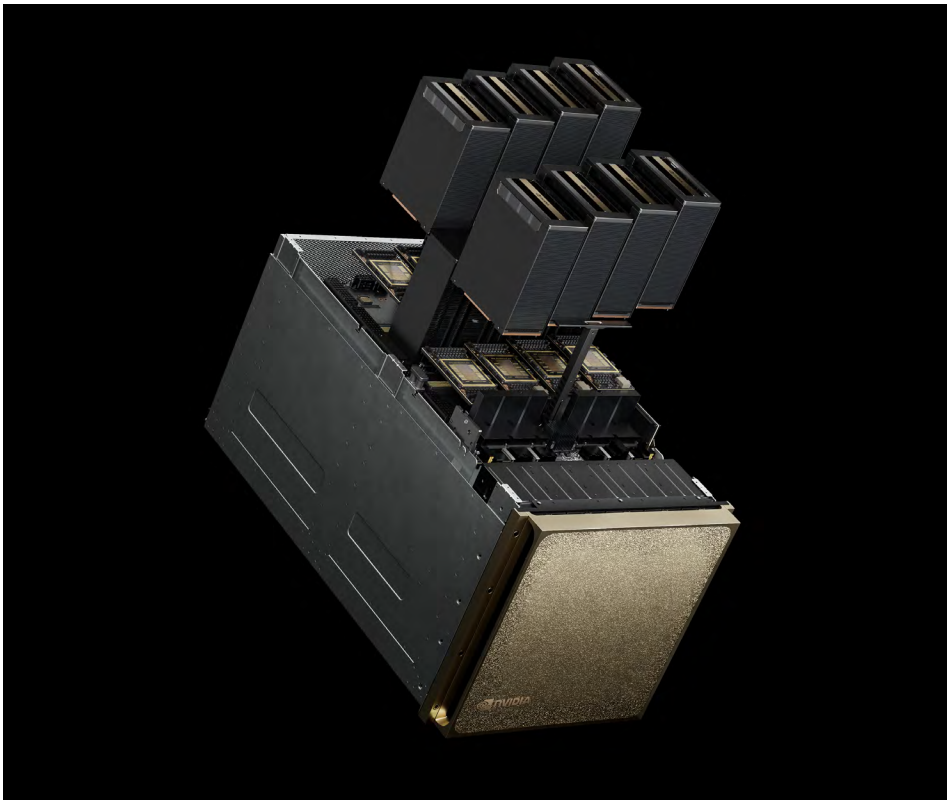
Zrównoważone obliczenia

Przyjęcie praktyk zrównoważonego rozwoju w centrach danych pozwala obniżyć ich ślad węglowy i zużycie energii oraz poprawić rentowność. Efektywność ta może zostać osiągnięta dzięki większej wydajności w obliczeniach przyspieszonych przy użyciu HGX. W zakresie wydajności inferencji dużych modeli językowych (LLM), HGX B200 zwiększa efektywność energetyczną 12-krotnie i obniża koszty o 12-krotnie w porównaniu do generacji Hopper.

12-krotnie niższe zużycie energii i całkowity koszt posiadania (TCO)



Przewidywana wydajność, może ulec zmianie. Opóźnienie token do tokena (TTL) = 50 ms w czasie rzeczywistym, opóźnienie pierwszego tokena (FTL) = 5 s, długość wejściowej sekwencji = 32 768, długość wyjściowej sekwencji = 1 028. Porównanie wydajności na GPU: 8-krotne HGX H100 chłodzonym powietrzem versus 1x HGX B200 chłodzonym powietrzem; Oszczędności TCO i energii; Przy równoważnej wydajności oszczędności TCO i energii dla 100 racków z 8 węzłami HGX H100 chłodzonym powietrzem w porównaniu do 8 racków HGX B200 chłodzonych powietrzem.



Specyfikacje techniczne[1]		
	GB200 NVL72	**HGX B200**
Blackwell GPUs | Grace CPUs	72 | 36	8 | 0
Rdzenie CPU	2 592 rdzeni Arm Neoverse V2	–
Całkowite FP4 Tensor Core	1 440 PFLOPS	144 PFLOPS
Całkowite FP8/FP6 Tensor Core	720 PFLOPS	72 PFLOPS
Całkowita szybka pamięć	Do 30 TB	Do 1,4 TB
Całkowita przepustowość pamięci	Do 576 TB/s	Do 62 TB/s
Całkowita przepustowość NVLink	130 TB/s	14,4 TB/s
Specyfikacja indywidualnego GPU Blackwell		
FP4 Tensor Core	20 PFLOPS	18 PFLOPS
FP8/FP6 Tensor Core	10 PFLOPS	9 PFLOPS
INT8 Tensor Core	10 POPS	9 POPS
FP16/BF16 Tensor Core	5 PFLOPS	4,5 PFLOPS
TF32 Tensor Core	2,5 PFLOPS	2,2 PFLOPS
FP32	80 TFLOPS	75 TFLOPS
FP64/FP64 Tensor Core	40 TFLOPS	37 TFLOPS
Pamięć GPU | Przepustowość	186GB HBM3E | 8TB/s	180GB HBM3E | 7.7TB/s
Multi-Instance GPU (MIG)	7	
Silnik dekompresji	Tak	
Dekoder	7 NVDEC[2] 7 nvJPEG	
Maksymalny projekt mocy termicznej (TDP)	Konfigurowalny do 1,200 W	Konfigurowalny do 1,000 W
Interkonekt	5. generacja NVLink: 1,8 TB/s PCIe Gen5: 128 GB/s	
Opcje serwera	Partner NVIDIA GB200 NVL72 i Systemy NVIDIA-Certified™ z 72 GPU	Partner NVIDIA HGX B200 i Systemy NVIDIA-Certified z 8 GPU

[1] Wszystkie liczby dotyczące Tensor Core z wyłączeniem FP64 z sparsity.
[2] Obsługiwane formaty zapewniają te przyspieszenia w porównaniu do GPU Tensor Core H100: 2× H.264, 1,25× HEVC, 1,25× VP9. Obsługa AV1 jest nowością w GPU Blackwell.

Superchip AI

GPU o architekturze Blackwell zawierają 208 miliardów tranzystorów i są produkowane przy użyciu własnego procesu TSMC 4NP. Wszystkie produkty Blackwell wyposażone są w dwa układy ograniczone maską (reticle-limited dies), połączone interfejsem chip-to-chip o przepustowości 10 TB/s, tworząc zintegrowany pojedynczy GPU.

2. generacja Transformer Engine

Druga generacja Transformer Engine wykorzystuje niestandardową technologię Tensor Core Blackwell w połączeniu z innowacjami NVIDIA TensorRT-LLM i NeMo Framework, aby przyspieszyć inferencję i trening dużych modeli językowych (LLMs) oraz modeli typu mixture-of-experts (MoE).

NVLink i NVLink Switch

Piąta generacja interfejsu NVIDIA NVLink umożliwia skalowanie do 576 GPU, odblokowując przyspieszoną wydajność dla wielo-trillionowych modeli AI o wielu parametrach. Układ NVIDIA NVLink Switch zapewnia 130 TB/s przepustowości GPU w jednym domenie NVLink z 72 GPU (NVL72) i osiąga czterokrotną efektywność przepustowości dzięki protokołowi NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ z obsługą FP8.

Silnik RAS

Blackwell wprowadza inteligentną odporność dzięki dedykowanemu silnikowi RAS (reliability, availability, and serviceability), który identyfikuje potencjalne usterki na wczesnym etapie, minimalizując przestoje. Zdolności NVIDIA do przewidywalnego zarządzania, oparte na AI, stale monitorują tysiące punktów danych w sprzęcie i oprogramowaniu, aby ocenić stan zdrowia systemu, przewidywać źródła przestojów i zapobiegać im.

Bezpieczna AI

Blackwell obejmuje NVIDIA Confidential Computing, które chroni wrażliwe dane i modele AI przed nieautoryzowanym dostępem, zapewniając silne zabezpieczenia sprzętowe. Blackwell jest pierwszym na rynku GPU z obsługą TEE-I/O, oferującym najbardziej wydajne rozwiązanie dla poufnych obliczeń, z hostami obsługującymi TEE-I/O i ochroną inline nad NVIDIA NVLink.

Silnik dekompresji

Silnik dekompresji Blackwell oraz możliwość dostępu do ogromnych ilości pamięci na CPU NVIDIA Grace przez szybki interfejs o przepustowości 900 GB/s w obie strony przyspieszają pełny proces obsługi zapytań bazodanowych, zapewniając najwyższą wydajność w analizie danych i naukach o danych. Obsługuje najnowsze formaty kompresji, takie jak LZ4, Snappy i Deflate.

Zrównoważone obliczenia

NVIDIA AI Enterprise to kompletna platforma programowa, która umożliwia dostęp do generatywnej AI dla każdego przedsiębiorstwa, zapewniając najszybszy i najbardziej efektywny czas pracy dla modeli podstawowych AI. Obejmuje **NVIDIA NIM™** — mikroserwisy inferencyjne, frameworki AI, biblioteki i narzędzia certyfikowane do uruchamiania na popularnych platformach centrów danych i systemach klasy NVIDIA-Certified Systems zintegrowanych z GPU NVIDIA.

W ramach NVIDIA AI Enterprise, NVIDIA NIM to zestaw łatwych w użyciu mikroserwisów inferencyjnych do przyspieszania wdrażania modeli podstawowych na dowolnej chmurze lub w centrum danych, pomagając jednocześnie w zachowaniu bezpieczeństwa danych. Firmy opierające swoje działania na AI polegają na bezpieczeństwie, wsparciu, zarządzalności i stabilności oferowanych przez NVIDIA AI Enterprise, co zapewnia płynne przejście od pilotażowego projektu do produkcji.

Razem z GPU NVIDIA Blackwell, NVIDIA AI Enterprise nie tylko upraszcza budowę platformy gotowej do AI, ale także przyspiesza czas uzyskania wartości.

Dowiedz się więcej o przepływach pracy AI z NVIDIA AI Enterprise, korzystając z praktycznych laboratoriów na NVIDIA Launchpad.

Gotowy, aby zacząć?

Aby dowiedzieć się więcej o NVIDIA Blackwell, odwiedź **nvidia.com/blackwell**

FORMAT
IMPOSSIBLE BECOMES REALITY