NVIDIA thông báo ra mắt siêu máy tính trí tuệ nhân tạo DGX GH200

Lớp mới của siêu máy tính trí tuệ nhân tạo kết nối 256 vi xử lý siêu chip Grace Hopper vào một GPU lớn, dung lượng 1 Exaflop, 144TB để cung cấp sức mạnh cho mô hình to lớn điều khiển trí tuệ nhân tạo sáng tạo, hệ thống gợi ý, xử lý dữ liệu.

TAIPEI, Taiwan—COMPUTEX—May 30, 2023—NVIDIA hôm nay thông báo về một lớp mới của siêu máy tính trí tuệ nhân tạo với bộ nhớ lớn — một siêu máy tính trí tuệ nhân tạo NVIDIA DGX™ được trang bị vi xử lý siêu chip Grace Hopper của NVIDIA® GH200 Grace Hopper Superchips và NVIDIA NVLink® Switch System — được tạo ra để thúc đẩy phát triển các mô hình to lớn cho ứng dụng ngôn ngữ trí tuệ nhân tạo sáng tạo, hệ thống gợi ý và công việc phân tích dữ liệu.

Với không gian bộ nhớ chia sẻ rộng lớn, NVIDIA DGX GH200 sử dụng công nghệ kết nối NVLink với Hệ thống chuyển mạch NVLink để kết hợp 256 vi xử lý siêu chip GH200, cho phép chúng hoạt động như một GPU đơn lẻ. Điều này cung cấp hiệu suất 1 exaflop và bộ nhớ chia sẻ 144 terabyte — gần 500 lần bộ nhớ so với thế hệ trước đó là NVIDIA DGX A100, được giới thiệu vào năm 2020.

"Trí tuệ nhân tạo sáng tạo, mô hình ngôn ngữ lớn và hệ thống gợi ý là động cơ kỹ thuật số của nền kinh tế hiện đại," Jensen Huang, người sáng lập và CEO của NVIDIA nói. "Siêu máy tính trí tuệ nhân tạo DGX GH200 tích hợp những công nghệ tính toán và mạng lưới tiên tiến nhất của NVIDIA để mở rộng ranh giới của trí tuệ nhân tạo."

Công nghệ NVLink của NVIDIA mở rộng quy mô trí tuệ nhân tạo

Vi xử lý siêu chip GH200 loại bỏ nhu cầu kết nối truyền thống từ CPU đến GPU thông qua PCIe bằng cách kết hợp NVIDIA Grace™ CPU dựa trên kiến trúc Arm với GPU NVIDIA H100 Tensor Core trong cùng một gói, sử dụng kết nối chip NVLink-C2C của NVIDIA. Điều này tăng băng thông giữa GPU và CPU lên gấp 7 lần so với công nghệ PCIe mới nhất, giảm tiêu thụ năng lượng giao tiếp gấp hơn 5 lần và cung cấp một khối xây dựng GPU kiến trúc Hopper có dung lượng 600GB cho siêu máy tính trí tuệ nhân tạo DGX GH200.

DGX GH200 là siêu máy tính đầu tiên ghép cặp vi xử lý siêu chip Grace Hopper với Hệ thống chuyển mạch NVLink của NVIDIA, một kết nối mới cho phép tất cả các GPU trong một hệ thống DGX GH200 hoạt động cùng nhau như một. Thế hệ trước chỉ cho phép kết hợp tám GPU với NVLink như một GPU mà không làm giảm hiệu suất.

Kiến trúc DGX GH200 cung cấp băng thông NVLink gấp 48 lần so với thế hệ trước, mang đến sức mạnh của một siêu máy tính trí tuệ nhân tạo to lớn với tính đơn giản khi lập trình một GPU duy nhất.

Một Công Cụ Nghiên Cứu Mới Cho Nhà Tiên Phong Trí Tuệ Nhân Tạo

Dự kiến Google Cloud, Meta và Microsoft sẽ là những đối tác đầu tiên có cơ hội trải nghiệm DGX GH200 để khám phá khả năng của nó đối với các công việc trí tuệ nhân tạo sáng tạo. NVIDIA cũng dự định cung cấp thiết kế DGX GH200 như một bản thiết kế cho các nhà cung cấp dịch vụ đám mây và những người tạo ra hạ tầng cực lớn khác để tùy chỉnh nó theo yêu cầu của họ.

"Xây dựng các mô hình sáng tạo tiên tiến đòi hỏi các phương pháp đổi mới về cơ sở hạ tầng trí tuệ nhân tạo," Mark Lohmeyer, Phó Chủ tịch Trách nhiệm tính toán tại Google Cloud nói. "Tính quy mô và bộ nhớ chia sẻ mới của vi xử lý siêu chip Grace Hopper giải quyết các hạn chế quan trọng trong trí tuệ nhân tạo quy mô lớn, và chúng tôi mong đợi khám phá khả năng của nó cho Google Cloud và các sáng kiến trí tuệ nhân tạo sáng tạo của chúng tôi."

"Khi các mô hình trí tuệ nhân tạo ngày càng lớn hơn, chúng cần hạ tầng mạnh mẽ có thể mở rộng để đáp ứng nhu cầu ngày càng tăng," Alexis Björlin, Phó Chủ tịch Cơ sở hạ tầng, Hệ thống Trí tuệ nhân tạo và Nền tảng Tăng tốc tại Meta nói. "Thiết kế Grace Hopper của NVIDIA nhằm cung cấp cho các nhà nghiên cứu khả năng khám phá các phương pháp mới để giải quyết những thách thức lớn nhất của họ."

"Việc huấn luyện các mô hình trí tuệ nhân tạo lớn truyền thống là một công việc tốn tài nguyên và thời gian," Girish Bablani, Phó Chủ tịch Công nghệ Cơ sở hạ tầng Azure tại Microsoft nói. "Tiềm năng của DGX GH200 để làm việc với các bộ dữ liệu có dung lượng terabyte sẽ cho phép các nhà phát triển tiến hành nghiên cứu tiên tiến với quy mô lớn hơn và tốc độ nhanh hơn."

Siêu Máy Tính Trí Tuệ Nhân Tạo NVIDIA Helios Mới Nâng Cao Nghiên Cứu và Phát Triển

NVIDIA đang xây dựng siêu máy tính trí tuệ nhân tạo dựa trên DGX GH200 để cung cấp sức mạnh cho các nhóm nghiên cứu và phát triển của mình.

Với tên gọi NVIDIA Helios, siêu máy tính này sẽ bao gồm bốn hệ thống DGX GH200. Mỗi hệ thống sẽ được kết nối với mạng lưới NVIDIA Quantum-2 InfiniBand để tăng tốc độ truyền dữ liệu cho việc huấn luyện các mô hình trí tuệ nhân tạo lớn. Helios sẽ bao gồm 1.024 vi xử lý siêu chip Grace Hopper và dự kiến sẽ đi vào hoạt động vào cuối năm.

Được tích hợp hoàn chỉnh và xây dựng cho các mô hình lớn

Siêu máy tính trí tuệ nhân tạo DGX GH200 bao gồm phần mềm NVIDIA để cung cấp một giải pháp đầy đủ cho các công việc trí tuệ nhân tạo và phân tích dữ liệu lớn. Phần mềm NVIDIA Base Command™ cung cấp quản lý luồng công việc trí tuệ nhân tạo, quản lý cụm đạt tiêu chuẩn doanh nghiệp, thư viện tăng tốc tính toán, lưu trữ và cơ sở hạ tầng mạng lưới và phần mềm hệ thống được tối ưu hóa để chạy các công việc trí tuệ nhân tạo.

Ngoài ra, còn có NVIDIA AI Enterprise, là lớp phần mềm của nền tảng trí tuệ nhân tạo NVIDIA. Nó cung cấp hơn 100 framework, mô hình được huấn luyện trước và công cụ phát triển để tối ưu hóa quá trình phát triển và triển khai các ứng dụng trí tuệ nhân tạo sử dụng mô hình sáng tạo, thị giác máy tính, trí tuệ nhân tạo giọng nói và nhiều hơn nữa.

Sản phẩm NVIDIA DGX GH200 dự kiến sẽ có sẵn vào cuối năm.

Xem cuộc trò chuyện của ông Huang về siêu máy tính trí tuệ nhân tạo NVIDIA DGX GH200 trong bài diễn thuyết chính tại COMPUTEX.

About NVIDIA

Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling the creation of the industrial metaverse. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

# # #

For further information, contact:

Melody Tu
NVIDIA Asia-Pacific
(65) 9355 1454
metu@nvidia.com

Inez Lim
CIZA Concept
(65) 9756 8877
inezlimjie@ciza.com

changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.