NVIDIA Tăng Cường Phiên Bản Hopper, Nền Tảng Tính Toán Trí Tuệ Nhân Tạo Hàng Đầu Thế Giới

HGX H200 Systems và Các Phiên Bản Đám Mây Sẽ Sớm Xuất Hiện Từ Các Nhà Sản Xuất Máy Chủ Và Nhà Cung Cấp Dịch Vụ Đám Mây Hàng Đầu Thế Giới

VIETNAM—November 17, 2023—Hôm nay, NVIDIA vừa công bố nền tảng tính toán trí tuệ nhân tạo (AI) hàng đầu thế giới NVIDIA HGX™ H200. Dựa trên kiến trúc NVIDIA Hopper™, nền tảng này sử dụng GPU Tensor Core NVIDIA H200 với bộ nhớ cấp tiến để xử lý lượng dữ liệu khổng lồ cho các nhiệm vụ trí tuệ nhân tạo sáng tạo và tính toán hiệu suất cao.

NVIDIA H200 là GPU đầu tiên trang bị HBM3e - bộ nhớ nhanh hơn và lớn hơn để đẩy nhanh quá trình tăng tốc trí tuệ nhân tạo sáng tạo và mô hình ngôn ngữ lớn, đồng thời thúc đẩy tính toán khoa học cho các nhiệm vụ tính toán hiệu suất cao (HPC). Với HBM3e, NVIDIA H200 cung cấp 141GB bộ nhớ tại tốc độ 4,8 terabytes mỗi giây, gần gấp đôi dung lượng và 2,4 lần băng thông so với tiền bối NVIDIA A100.

Các hệ thống được trang bị NVIDIA H200 từ các nhà sản xuất máy chủ hàng đầu thế giới và các nhà cung cấp dịch vụ đám mây dự kiến sẽ bắt đầu xuất hàng vào quý II năm 2024.

"Để tạo ra trí tuệ với trí tuệ nhân tạo sáng tạo và các ứng dụng HPC, lượng dữ liệu lớn phải được xử lý một cách hiệu quả ở tốc độ cao bằng bộ nhớ GPU lớn và nhanh", Ian Buck, Phó Chủ tịch hyperscale và HPC tại NVIDIA nói. "Với NVIDIA H200, nền tảng siêu

máy tính trí tuệ nhân tạo từ đầu đến cuối hàng đầu của ngành đã nhanh hơn để giải quyết một số thách thức quan trọng nhất trên thế giới."

Sự Đổi Mới Và Sự Nhảy Vọt Về Hiệu Suất Vĩnh Viễn

Kiến trúc NVIDIA Hopper mang lại một sự nhảy vọt về hiệu suất không giới hạn so với người tiền nhiệm và tiếp tục đặt ra một mức tiêu chuẩn mới thông qua việc cải thiện phần mềm liên tục với H100, bao gồm việc ra mắt gần đây các thư viện mã nguồn mở mạnh mẽ như NVIDIA TensorRT™-LLM.

Việc giới thiệu H200 sẽ dẫn đến những bước nhảy vọt về hiệu suất tiếp theo, bao gồm gần gấp đôi tốc độ suy luận trên Llama 2, một mô hình LLM với 70 tỷ tham số, so với H100. Dự kiến sẽ có thêm sự lãnh đạo về hiệu suất và cải thiện với H200 thông qua các bản cập nhật phần mềm trong tương lai.

Các Dạng Hình Dạng Của NVIDIA H200

NVIDIA H200 sẽ có sẵn trong các bo mạch máy chủ NVIDIA HGX H200 với cấu hình bốn và tám lõi, tương thích với cả phần cứng và phần mềm của các hệ thống HGX H100. Nó cũng có sẵn trong NVIDIA GH200 Grace Hopper™ Superchip với HBM3e, được công bố vào tháng Tám.

Với những tùy chọn này, H200 có thể triển khai trong mọi loại trung tâm dữ liệu, bao gồm trên cơ sở, đám mây, đám mây lai và điểm cuối. Hệ sinh thái toàn cầu của đối tác sản xuất máy chủ của NVIDIA - bao gồm ASRock Rack, ASUS, Dell Technologies, Eviden, GIGABYTE, Hewlett Packard Enterprise, Ingrasys, Lenovo, QCT, Supermicro, Wistron và Wiwynn - có thể cập nhật hệ thống hiện có của họ với H200.

Amazon Web Services, Google Cloud, Microsoft Azure và Oracle Cloud Infrastructure sẽ là những nhà cung cấp dịch vụ đám mây đầu tiên triển khai các phiên bản dựa trên H200 bắt đầu từ năm sau, cùng với CoreWeave, Lambda và Vultr.

Được hỗ trợ bởi NVIDIA NVLink™ và NVSwitch™, HGX H200 cung cấp hiệu suất cao nhất trên nhiều nhiệm vụ ứng dụng khác nhau, bao gồm đào tạo và suy luận LLM cho các mô hình lớn vượt quá 175 tỷ tham số.

Một bộ HGX H200 tám lõi cung cấp hơn 32 petaflops tính toán đào tạo FP8 và tổng cộng 1,1TB bộ nhớ băng thông cao cho hiệu suất cao nhất trong trí tuệ nhân tạo sáng tạo và ứng dụng HPC.

Khi kết hợp với CPU NVIDIA Grace™ với kết nối siêu nhanh NVLink-C2C, H200 tạo ra GH200 Grace Hopper Superchip với HBM3e - một mô-đun tích hợp được thiết kế để phục vụ các ứng dụng HPC và AI quy mô khổng lồ.

Tăng Tốc Trí Tuệ Nhân Tạo Với Phần Mềm Toàn Diện Của NVIDIA

Nền tảng tính toán tăng cường của NVIDIA được hỗ trợ bởi các công cụ phần mềm mạnh mẽ giúp các nhà phát triển và doanh nghiệp xây dựng và tăng tốc các ứng dụng sẵn sàng cho sản xuất từ trí tuệ nhân tạo đến HPC. Điều này bao gồm bộ phần mềm NVIDIA AI Enterprise dành cho các nhiệm vụ như phát triển phần mềm phát biểu, hệ thống gợi ý và suy luận quy mô lớn.

Sẵn Sàng

NVIDIA H200 sẽ có sẵn từ các nhà sản xuất hệ thống toàn cầu và các nhà cung cấp dịch vụ đám mây hàng đầu bắt đầu từ quý II năm 2024.

Xem bài phát biểu đặc biệt của Buck vào ngày 13 tháng 11 lúc 6 giờ sáng giờ Thái Bình Dương để tìm hiểu thêm về GPU Tensor Core NVIDIA H200.

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 industrial digitalization across markets. 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

Certain statements in this press release including, but not limited to, statements as to: the benefits, performance, specifications, impact and availability of the NVIDIA HGX H200 and NVIDIA Hopper architecture; the processing requirements to create intelligence with generative AI and HPC applications; the ease for partner server makers to update H100-based systems with H200; and the first cloud service providers expected to deploy H200-based instances are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; 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.