NVIDIA MGX Mang Đến Nhà Sản Xuất Hệ Thống Kiến Trúc Mô-đun Đáp Ứng Nhu Cầu Tăng Tốc Đa Dạng Trong Trung Tâm Dữ Liệu Toàn Cầu

QCT và Supermicro Là Hai Trong Những Đối Tác Đầu Tiên Sử Dụng Đặc Điểm Kỹ Thuật Máy Chủ Cho Hơn 100 Cấu Hình Hệ Thống Nhằm Tăng Tốc Công Việc Trí Tuệ Nhân Tạo, HPC và Omniverse

TAIPEI, Taiwan—COMPUTEX—May 30, 2023—Để đáp ứng nhu cầu tăng tốc đa dạng của các trung tâm dữ liệu trên toàn thế giới, NVIDIA hôm nay giới thiệu đặc điểm kỹ thuật máy chủ NVIDIA MGX™ , mang đến cho các nhà sản xuất hệ thống một kiến trúc tham chiếu mô-đun để nhanh chóng và tiết kiệm chi phí xây dựng hơn 100 biến thể máy chủ phù hợp với nhiều ứng dụng trí tuệ nhân tạo, tính toán hiệu năng cao và Omniverse.

ASRock Rack, ASUS, GIGABYTE, Pegatron, QCT và Supermicro sẽ áp dụng MGX, giúp giảm chi phí phát triển lên đến ba phần tư và rút ngắn thời gian phát triển xuống hai phần ba chỉ trong sáu tháng.

"Các doanh nghiệp đang tìm kiếm nhiều lựa chọn tăng tốc tính toán khi xây dựng các trung tâm dữ liệu đáp ứng nhu cầu kinh doanh và ứng dụng cụ thể của họ," Kaustubh Sanghani, Phó Chủ tịch sản phẩm GPU của NVIDIA nói. "Chúng tôi tạo ra MGX để giúp các tổ chức khởi động trí tuệ nhân tạo doanh nghiệp, đồng thời tiết kiệm cho họ lượng thời gian và tiền bạc đáng kể."

Với MGX, các nhà sản xuất bắt đầu với một kiến trúc hệ thống cơ bản được tối ưu hóa cho tính toán tăng tốc cho khung máy chủ của họ, sau đó chọn GPU, DPU và CPU. Các biến thể thiết kế có thể đáp ứng các tải công việc độc đáo, chẳng hạn như HPC, khoa học dữ liệu, mô hình ngôn ngữ lớn, tính toán cạnh, đồ họa và video, trí tuệ nhân tạo doanh nghiệp, và thiết kế và mô phỏng. Nhiều nhiệm vụ như đào tạo trí tuệ nhân tạo và 5G có thể được thực hiện trên cùng một máy, trong khi việc nâng cấp lên các thế hệ phần cứng tương lai có thể diễn ra một cách dễ dàng. MGX cũng dễ dàng tích hợp vào các trung tâm dữ liệu đám mây và doanh nghiệp.

Hợp Tác Với Các Nhà Lãnh Đạo Công Nghiệp

QCT và Supermicro sẽ là những nhà cung cấp đầu tiên trên thị trường, với các thiết kế MGX xuất hiện vào tháng Tám. Hệ thống ARS-221GL-NR của Supermicro, được công bố hôm nay, sẽ bao gồm vi xử lý siêu chip Grace CPU của NVIDIA, trong khi hệ thống S74G-2U của QCT, cũng được công bố hôm nay, sẽ sử dụng vi xử lý siêu chip GH200 Grace Hopper Superchip của NVIDIA.

Ngoài ra, SoftBank Corp. dự định triển khai nhiều trung tâm dữ liệu quy mô lớn trên khắp Nhật Bản và sử dụng MGX để phân bổ tài nguyên GPU động giữa các ứng dụng trí tuệ nhân tạo sáng tạo và 5G.

"Khi trí tuệ nhân tạo sáng tạo lan tỏa trong kinh doanh và lối sống của người tiêu dùng, xây dựng hạ tầng phù hợp với chi phí phù hợp là một trong những thách thức lớn nhất của các nhà khai thác mạng," Junichi Miyakawa, Chủ tịch kiêm CEO tại SoftBank Corp. nói. "Chúng tôi tin rằng NVIDIA MGX có thể đối phó với các thách thức đó và cho phép sử dụng đa năng trí tuệ nhân tạo, 5G và nhiều hơn nữa dựa trên yêu cầu công việc thời gian thực."

Thiết Kế Khác Nhau Cho Nhu Cầu Khác Nhau

Các trung tâm dữ liệu ngày càng cần đáp ứng yêu cầu về khả năng tính toán gia tăng và giảm lượng khí thải carbon để chống biến đổi khí hậu, đồng thời giữ chi phí ổn định.

Các máy chủ tính toán tăng tốc từ NVIDIA đã lâu đã cung cấp hiệu năng tính toán và hiệu suất năng lượng xuất sắc. Bây giờ, thiết kế mô-đun của MGX cho phép nhà sản xuất hệ thống đáp ứng một cách hiệu quả hơn các yêu cầu đặc thù về ngân sách, cung cấp điện, thiết kế nhiệt và cơ khí của từng khách hàng.

Nhiều Hình Thức Thiết Kế Đa Dạng

MGX hoạt động với nhiều hình thức thiết kế và tương thích với các thế hệ phần cứng NVIDIA hiện tại và tương lai, bao gồm:

- Khung máy chủ: 1U, 2U, 4U (làm mát bằng không khí hoặc chất lỏng)
- GPU: Toàn bộ danh mục GPU NVIDIA bao gồm các phiên bản mới nhất H100, L40, L4
- CPU: Vi xử lý siêu chip Grace CPU, vi xử lý siêu chip Grace Hopper GH200, CPU x86
- Mạng: Vi xử lý DPU BlueField®-3 của NVIDIA, bộ chuyển mạng ConnectX®-7

MGX khác biệt với NVIDIA HGX™ ở chỗ nó cung cấp khả năng tương thích linh hoạt, đa thế hệ với các sản phẩm NVIDIA để đảm bảo rằng các nhà xây dựng hệ thống có thể tái sử dụng các thiết kế hiện có và dễ dàng áp dụng các sản phẩm thế hệ tiếp theo mà không cần thiết kế lại đắt đỏ. Ngược lại, HGX được dựa trên một bo mạch cơ bản kết nối NVLink® cho phép mở rộng để tạo ra hệ thống trí tuệ nhân tạo và tính toán hiệu năng cao tối ưu.

Phần mềm để Đẩy Mạnh Tăng Tốc Hơn

Ngoài phần cứng, MGX được hỗ trợ bởi ngăn xếp phần mềm đầy đủ của NVIDIA, cho phép các nhà phát triển và doanh nghiệp xây dựng và tăng tốc các ứng dụng trí tuệ nhân tạo, tính toán hiệu năng cao và các ứng dụng khác. Điều này bao gồm NVIDIA AI Enterprise, lớp phần mềm của nền tảng trí tuệ nhân tạo của NVIDIA, với hơn 100 framework, mô hình được đào tạo trước và các công cụ phát triển để tăng tốc trí tuệ nhân tạo và khoa học dữ liệu cho phát triển và triển khai trí tuệ nhân tạo doanh nghiệp được hỗ trợ đầy đủ.

MGX tương thích với Dự án Máy tính Mở và các tủ máy chủ của Hiệp hội Công nghiệp Điện tử, để dễ dàng tích hợp vào các trung tâm dữ liệu doanh nghiệp và đám mây.

Xem NVIDIA đồng sáng lập và CEO Jensen Huang trình bày về đặc điểm kỹ thuật máy chủ MGX trong diễn thuyết chính của ông tại COMPUTEX 2023.

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

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features and availability of our products, collaborations, services and technologies, including the NVIDIA MGX server specification, Omniverse, NVIDIA GPUs including H100, L40 and L4, NVIDIA DPUs including BlueField-3, NVIDIA CPUs including x86 CPUs, NVIDIA HPCs, large language models, and edge computing, Grace CPU Superchip, GH200 Grace Hopper Superchip, chassis, ConnectX-7 network adapters, NVIDIA HGX, NVLink, NVIDIA AI Enterprise, and the NVIDIA AI platform; our collaborations with QCT, Supermicro, ASRock Rack, ASUS, GIGABYTE, Pegatron and SoftBank Corp., and the benefits, impact, performance and availability thereof; enterprises seeking more accelerated computing options when architecting data centers to meet their specific business and application needs; and data centers increasingly needing to meet requirements for growing compute capabilities and decreasing carbon emissions, while also keeping costs down 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.