NVIDIA Mengumumkan Superkomputer AI DGX GH200

Kelas Baru Superkomputer AI Menghubungkan 256 Superchip Grace Hopper Ke dalam 1-Exaflop, 144TB GPU Masif untuk Model Raksasa yang Mendukung AI Generatif, Sistem Rekomendasi, dan Pemrosesan Data

TAIPEI, Taiwan—COMPUTEX—May 30, 2023—NVIDIA hari ini mengumumkan kelas baru dari superkomputer AI bermemori besar — NVIDIA DGX™ yang ditenagai oleh NVIDIA® GH200 Grace Hopper Superchips dan NVIDIA NVLink® Switch System — dibuat untuk pengembangan model raksasa generasi mendatang demi aplikasi bahasa AI generatif, sistem pemberi rekomendasi, dan beban kerja analitik data.

Ruang shared memory NVIDIA DGX GH200 yang sangat besar menggunakan teknologi interkoneksi NVLink dengan NVLink Switch System untuk menggabungkan 256 superchip GH200, yang membuatnya bisa bekerja sebagai GPU tunggal. Ini memberikan kinerja 1 exaflop dan memori berbagi 144 terabyte — hampir 500x memori lebih banyak daripada NVIDIA DGX A100 generasi sebelumnya, yang diperkenalkan pada tahun 2020.

“Generative AI, model bahasa besar, dan sistem pemberi rekomendasi adalah mesin digital ekonomi modern,” kata Jensen Huang, pendiri dan CEO NVIDIA. “Superkomputer DGX GH200 AI mengintegrasikan teknologi komputasi terakselerasi dan jaringan tercanggih NVIDIA untuk memperluas batas AI.”

Teknologi NVIDIA NVLink Memperluas AI dalam Skala Besar

Superchip GH200 menghapus kebutuhan akan koneksi PCIe CPU-ke-GPU tradisional dengan menggabungkan CPU NVIDIA Grace™ berbasis Arm dengan GPU NVIDIA H100 Tensor Core dalam paket yang sama, menggunakan interkoneksi chip NVIDIA NVLink-C2C. Hal ini meningkatkan bandwidth antara GPU dan CPU hingga 7x dibandingkan dengan teknologi PCIe terbaru, memangkas konsumsi daya interkoneksi lebih dari 5x, dan menyediakan blok penyusun GPU arsitektur Hopper 600GB untuk superkomputer DGX GH200.

DGX GH200 adalah superkomputer pertama yang memasangkan Grace Hopper Superchips dengan NVIDIA NVLink Switch System, interkoneksi baru yang memungkinkan semua GPU dalam sistem DGX GH200 bekerja bersama sebagai satu kesatuan. Sistem generasi sebelumnya hanya menyediakan delapan GPU untuk dipadukan dengan NVLink sebagai satu GPU tanpa mengurangi performa.

Arsitektur DGX GH200 menyediakan bandwidth NVLink 48x lebih banyak daripada generasi sebelumnya, menghadirkan kekuatan superkomputer AI yang masif dengan kemudahan memprogram satu GPU.

Alat Riset Baru untuk Pionir AI

Google Cloud, Meta, dan Microsoft adalah yang pertama diprediksi mendapatkan akses ke DGX GH200 untuk mengeksplorasi kemampuannya untuk beban kerja AI generatif. NVIDIA juga bermaksud menyediakan desain DGX GH200 sebagai cetak biru untuk penyedia layanan cloud dan hyperscaler lainnya sehingga mereka dapat menyesuaikannya lebih lanjut untuk infrastruktur mereka.

“Membangun model generatif lanjutan membutuhkan pendekatan inovatif untuk infrastruktur AI,” kata Mark Lohmeyer, wakil presiden Compute di Google Cloud. “Skala NVLink baru dan memori bersama dari Grace Hopper Superchips mengatasi hambatan utama dalam AI skala besar dan kami berharap dapat mengeksplorasi kemampuannya untuk Google Cloud dan inisiatif AI generatif kami.”

“Saat model AI tumbuh lebih besar, mereka membutuhkan infrastruktur yang kuat yang dapat disesuaikan untuk memenuhi permintaan yang meningkat,” kata Alexis Björlin, wakil presiden Infrastruktur, Sistem AI, dan Platform yang Dipercepat di Meta. “Rancangan Grace Hopper NVIDIA tampaknya memberi para peneliti kemampuan untuk mengeksplorasi pendekatan baru demi memecahkan tantangan terbesar mereka.”

“Melatih model AI besar merupakan tugas yang memerlukan sumber daya dan waktu secara intensif,” kata Girish Bablani, wakil presiden perusahaan Infrastruktur Azure di Microsoft. “Potensi DGX GH200 untuk bekerja dengan kumpulan data berukuran terabyte akan memungkinkan pengembang untuk melakukan penelitian lanjutan pada skala yang lebih besar dan kecepatan yang terakselerasi.”

Superkomputer NVIDIA Helios Baru untuk Majukan Riset dan Pengembangan
NVIDIA sedang membangun superkomputer AI berbasis DGX GH200 miliknya sendiri untuk mendukung pekerjaan para peneliti dan tim pengembangannya.

Dinamakan NVIDIA Helios, superkomputer ini akan menampilkan empat sistem DGX GH200. Masing-masing akan saling terhubung dengan jaringan NVIDIA Quantum-2 InfiniBand untuk meningkatkan throughput data dalam melatih model AI berukuran besar. Helios akan menyertakan 1.024 Grace Hopper Superchips dan diperkirakan online pada akhir tahun ini.

Terintegrasi Sepenuhnya dan Dibangun Khusus untuk Model Raksasa
Superkomputer DGX GH200 menyertakan perangkat lunak NVIDIA untuk menyediakan solusi turnkey, full-stack untuk beban kerja AI dan analitik data terbesar. Perangkat lunak NVIDIA Base Command™ menyediakan manajemen alur kerja AI, manajemen klaster tingkat perusahaan, pustaka yang mempercepat komputasi, penyimpanan, dan infrastruktur jaringan, serta perangkat lunak sistem yang dioptimalkan untuk menjalankan beban kerja AI.

Juga tersedia NVIDIA AI Enterprise, lapisan perangkat lunak platform NVIDIA AI. Ini menyediakan lebih dari 100 kerangka kerja, model yang telah dilatih sebelumnya, dan alat pengembangan untuk merampingkan pengembangan dan penerapan AI produksi termasuk AI generatif, visi komputer, AI ucapan, dan lainnya.

Ketersediaan

Superkomputer NVIDIA DGX GH200 diharapkan akan tersedia pada akhir tahun ini.

Tonton Huang membahas superkomputer NVIDIA DGX GH200 selama pidato utamanya di 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

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features and availability of our products, services and technologies, including NVIDIA Grace Hopper Superchips and supercomputer, NVIDIA DGX and DGX GH200, NVLink including the NVLink Switch System and NVLink interconnect technology, DGX H100, NVIDIA Grace CPU, H100 Tensor Core GPU, Helios supercomputer, Quantum-2 InfiniBand, NVIDIA Base Command and NVIDIA AI Enterprise; our collaborations with Google Cloud, Meta and Microsoft and the benefits, impact, performance, features and availability thereof; generative AI, recommender systems and data analytics being engines of the modern economy, requiring unprecedented scale, speed and efficiency; and NVIDIA's intention to provide the DGX GH200 design as a blueprint to cloud service providers and other hyperscalers so they can further customize it for their infrastructure 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.