



NVIDIA เปิดตัว DGX GH200 AI Supercomputer

อีกระดับของ AI ซูเปอร์คอมพิวเตอร์เชื่อมต่อซูเปอร์ชิป Grace Hopper 256 ตัวเข้ากับ GPU ขนาดใหญ่ 1-exaflop, 144TB สำหรับโมเดลขนาดยักษ์ที่ขับเคลื่อน Generative AI, ระบบแนะนำ และการประมวลผลข้อมูล

TAIPEI, Taiwan—COMPUTEX—May 30, 2023—NVIDIA ประกาศเปิดตัว AI ซูเปอร์คอมพิวเตอร์หน่วยความจำขนาดใหญ่ระดับใหม่ — ซูเปอร์คอมพิวเตอร์ NVIDIA [DGX™](#) ที่ขับเคลื่อนโดย NVIDIA [GH200 Grace Hopper Superchips](#) และ [NVIDIA® NVLink® Switch System](#) — สร้างขึ้นเพื่อเปิดใช้งานการพัฒนาโมเดลขนาดยักษ์, โมเดลภาษารุ่นต่อไปสำหรับ Generative AI, ระบบผู้แนะนำ และปริมาณงานการวิเคราะห์ข้อมูล

พื้นที่หน่วยความจำที่ใช้ร่วมกันขนาดใหญ่ของ NVIDIA DGX GH200 ใช้เทคโนโลยีการเชื่อมต่อระหว่าง NVLink กับ NVLink Switch System เพื่อรวม GH200 Superchips จำนวน 200 ตัวเข้าด้วยกัน ทำให้สามารถทำงานเป็น GPU เดียวได้ สิ่งนี้มอบประสิทธิภาพระดับ 1 exaflop และมีแบนด์วิดธ์ของหน่วยความจำที่ 144 เทราไบต์ที่ใช้ร่วมกัน — หน่วยความจำมากกว่าเกือบ 500 เท่าของ NVIDIA DGX A100 รุ่นก่อนหน้าที่เปิดตัวในปี 2020

Jensen Huang ผู้ก่อตั้งและ CEO ของ NVIDIA กล่าวว่า “Generative AI, โมเดลภาษาขนาดใหญ่ และระบบผู้แนะนำคือเครื่องมือดิจิทัลของเศรษฐกิจยุคใหม่ AI ซูเปอร์คอมพิวเตอร์ DGX GH200 ผสานรวมเทคโนโลยีคอมพิวเตอร์และเครือข่ายเร่งความเร็วขั้นสูงสุดของ NVIDIA เพื่อขยายขอบเขตของ AI”

เทคโนโลยี NVIDIA NVLink ขยาย AI ในวงกว้าง

ซูเปอร์ชิป GH200 ช่วยลดความจำเป็นในการเชื่อมต่อ PCIe CPU-to-GPU แบบดั้งเดิมโดยการรวม [CPU NVIDIA Grace™](#) ที่ใช้ Arm เข้ากับ [NVIDIA H100 Tensor Core GPU](#)

ในแพ็คเกจเดียวกันโดยใช้การเชื่อมต่อระหว่างชิปด้วย [NVIDIA NVLink-C2C](#)

สิ่งนี้จะเพิ่มแบนด์วิดธ์ระหว่าง GPU และ CPU ได้ถึง 7 เท่า เมื่อเทียบกับเทคโนโลยี PCIe ล่าสุด ลดการใช้พลังงานในการเชื่อมต่อมากกว่า 5 เท่า และให้บล็อกการสร้าง GPU สถาปัตยกรรม Hopper ขนาด 600GB สำหรับซูเปอร์คอมพิวเตอร์ DGX GH200

DGX GH200 เป็นซูเปอร์คอมพิวเตอร์เครื่องแรกที่ใช้ Grace Hopper Superchips กับ NVIDIA NVLink Switch System ซึ่งเป็นการเชื่อมต่อโครงข่ายใหม่ที่ช่วยให้ GPU ทั้งหมดในระบบ DGX GH200 ทำงานร่วมกันเป็นหนึ่งเดียว ระบบรุ่นก่อนหน้ามีให้สำหรับ GPU แปรตัวเท่านั้น ที่จะรวมกับ NVLink เป็น GPU เดียวโดยไม่ลดทอนประสิทธิภาพ

สถาปัตยกรรม DGX GH200 ให้แบนด์วิดธ์ NVLink มากกว่ารุ่นก่อนหน้าถึง 48 เท่า มอบพลังของซูเปอร์คอมพิวเตอร์ AI ขนาดใหญ่ด้วยความเรียบง่ายในการเขียนโปรแกรม GPU ตัวเดียว

เครื่องมือวิจัยใหม่สำหรับผู้บุกเบิก AI

Google Cloud, Meta และ Microsoft เป็นหนึ่งในกลุ่มแรก ๆ ที่คาดว่าจะสามารถเข้าถึง DGX GH200 เพื่อสำรวจความสามารถสำหรับปริมาณงาน AI เชิงสร้างสรรค์ NVIDIA ยังตั้งใจที่จะให้การออกแบบ DGX GH200 เป็นต้นแบบให้กับผู้ให้บริการคลาวด์และไฮเปอร์สเกลอื่น ๆ

เพื่อให้พวกเขาสามารถปรับแต่งเพิ่มเติมสำหรับโครงสร้างพื้นฐานของตนได้

"การสร้างโมเดล Generative ขั้นสูงต้องใช้แนวทางที่เป็นนวัตกรรมสำหรับโครงสร้างพื้นฐาน AI" Mark Lohmeyer รองประธานฝ่ายประมวลผลของ Google Cloud กล่าว "สเกล NVLink ใหม่และหน่วยความจำที่เข้าร่วมกันของ Grace Hopper Superchips ช่วยแก้ปัญหาคอขวดที่สำคัญใน AI ขนาดใหญ่ และเราหวังว่าจะได้สำรวจความสามารถของ Google Cloud และโครงการริเริ่มด้าน Generative AI ของเรา"

"เมื่อโมเดล AI มีขนาดใหญ่ขึ้น

พวกเขาต้องการโครงสร้างพื้นฐานที่มีประสิทธิภาพซึ่งสามารถปรับขนาดเพื่อตอบสนองความต้องการที่เพิ่มขึ้น" Alexis Björlin รองประธานฝ่ายโครงสร้างพื้นฐาน ระบบ AI และแพลตฟอร์มเร่งความเร็วที่ Meta กล่าว "การออกแบบ Grace Hopper ของ NVIDIA จะช่วยให้นักวิจัยสามารถสำรวจแนวทางใหม่ ๆ เพื่อแก้ปัญหาค่าความท้าทายที่ยิ่งใหญ่ที่สุดของพวกเขา"

"การฝึกอบรมโมเดล AI ขนาดใหญ่เป็นงานที่ต้องใช้ทรัพยากรและเวลามาก" Girish Bablani รองประธานองค์กรของ Azure Infrastructure ที่ Microsoft กล่าว "ศักยภาพของ DGX GH200 ในการทำงานกับชุดข้อมูลขนาดใหญ่ไปข้างหน้าจะช่วยให้นักพัฒนาสามารถทำการวิจัยขั้นสูงในระดับที่ใหญ่ขึ้นและเร่งความเร็วได้"

NVIDIA Helios Supercomputer เพื่อความก้าวหน้าในการวิจัยและพัฒนา

NVIDIA กำลังสร้าง AI ซูเปอร์คอมพิวเตอร์ที่ใช้ DGX GH200

ของตัวเองเพื่อขับเคลื่อนการทำงานของนักวิจัยและทีมพัฒนา

NVIDIA Helios ซูเปอร์คอมพิวเตอร์จะมีระบบ DGX GH200 สีชุด แต่ละเครือข่ายจะเชื่อมต่อกับเครือข่าย [NVIDIA Quantum-2 InfiniBand](#) เพื่อเพิ่มปริมาณการประมวลผลข้อมูลสำหรับการฝึกอบรมโมเดล AI ขนาดใหญ่ Helios จะรวม Grace Hopper Superchips 1,024 ตัวและคาดว่าจะออนไลน์ได้ภายในสิ้นปีนี้

ครบวงจรและสร้างขึ้นตามวัตถุประสงค์สำหรับโมเดลขนาดยักษ์

ซูเปอร์คอมพิวเตอร์ DGX GH200 ประกอบด้วยซอฟต์แวร์ NVIDIA

เพื่อมอบโซลูชันแบบพูลสแตกแบบครบวงจรสำหรับปริมาณงาน AI และการวิเคราะห์ข้อมูลที่ใหญ่ที่สุด

ซอฟต์แวร์ [NVIDIA Base Command™](#) ให้การจัดการเวิร์กโฟลว์ AI

การจัดการคลัสเตอร์ระดับองค์กรไลบรารีที่เร่งการประมวลผลการจัดเก็บข้อมูลและโครงสร้างพื้นฐานเครือข่าย และซอฟต์แวร์ระบบที่ปรับให้เหมาะสมสำหรับการเรียกใช้ปริมาณงาน AI

นอกจากนี้ยังมี [NVIDIA AI Enterprise](#) ซึ่งเป็นเลเยอร์ซอฟต์แวร์ของแพลตฟอร์ม NVIDIA AI

มีเฟรมเวิร์กมากกว่า 100

เฟรมเวิร์กโมเดลที่ผ่านการฝึกอบรมและเครื่องมือการพัฒนาเพื่อปรับปรุงการพัฒนาและการปรับใช้ AI การผลิตรวมถึง Generative AI, คอมพิวเตอร์วิทัศน์, Speech AI และอื่น ๆ

ความพร้อมในการใช้งาน

ซูเปอร์คอมพิวเตอร์ NVIDIA DGX GH200 คาดว่าจะพร้อมใช้งานภายในสิ้นปีนี้

รับชม Huang พูดคุยเกี่ยวกับซูเปอร์คอมพิวเตอร์ NVIDIA DGX GH200 ระหว่างการ[ปราศรัยที่](#)

[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.

© 2023 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, DGX, NVIDIA Base Command, NVIDIA Grace, NVIDIA Hopper and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.