Live From Taipei: NVIDIA CEO Unveils Gen AI Platforms for Every Industry

The COMPUTEX keynote showed new systems, software and services — many powered by Grace Hopper superchips — to harness generative AI, the most transformative technology of our time.

Author: Rick Merritt

In his first live keynote since the pandemic, NVIDIA founder and CEO Jensen Huang today kicked off the COMPUTEX conference in Taipei, announcing platforms that companies can use to ride a historic wave of generative AI that's transforming industries from advertising to manufacturing to telecom.

"We're back," Huang roared as he took the stage after years of virtual keynotes, some from his home kitchen. "I haven't given a public speech in almost four years — wish me luck!"

Speaking for nearly two hours to a packed house of some 3,500, he described accelerated computing services, software and systems that are enabling new business models and making current ones more efficient.

"Accelerated computing and AI mark a reinvention of computing," said Huang, whose travels in his hometown over the past week have been tracked daily by local media.

In a demonstration of its power, he used the massive 8K wall he spoke in front of to show a text prompt generating a theme song for his keynote, singable as any karaoke tune. Huang, who occasionally bantered with the crowd in his native Taiwanese, briefly led the audience in singing the new anthem.

"We're now at the tipping point of a new computing era with accelerated computing and AI that's been embraced by almost every computing and cloud company in the world," he said, noting 40,000 large companies and 15,000 startups now use NVIDIA technologies with 25 million downloads of CUDA software last year alone.

Grace Hopper powers big-memory supercomputers for gen AI. Modular reference architecture enables 100+ accelerated server variations. WPP, NVIDIA create digital ad content engine in Omniverse. SoftBank, NVIDIA build 5G, gen AI data centers in Japan. Networking technology accelerates Ethernet-based AI clouds. NVIDIA ACE for Games breathes life into characters with gen AI. Electronics manufacturers worldwide embrace NVIDIA AI.

Grace Hopper powers big-memory supercomputers for gen AI. Modular reference architecture enables 100+ accelerated server variations. WPP, NVIDIA create digital ad content engine in Omniverse. SoftBank, NVIDIA build 5G, gen AI data centers in Japan. Networking technology accelerates Ethernet-based AI clouds. NVIDIA ACE for Games breathes life into characters with gen AI. Electronics manufacturers worldwide embrace NVIDIA AI.

Grace Hopper powers big-memory supercomputers for gen AI.

Modular reference architecture enables 100+ accelerated server variations.

WPP, NVIDIA create digital ad content engine in Omniverse.

SoftBank, NVIDIA build 5G, gen AI data centers in Japan.

Networking technology accelerates Ethernet-based AI clouds.

NVIDIA ACE for Games breathes life into characters with gen AI.

Electronics manufacturers worldwide embrace NVIDIA AI.

Grace Hopper powers big-memory supercomputers for gen AI. Modular reference architecture enables 100+ accelerated server variations. WPP, NVIDIA create digital ad content engine in Omniverse. SoftBank, NVIDIA build 5G, gen AI data centers in Japan. Networking technology accelerates Ethernet-based AI clouds. NVIDIA ACE for Games breathes life into characters with gen AI. Electronics manufacturers worldwide embrace NVIDIA AI.

Grace Hopper powers big-memory supercomputers for gen AI.

Modular reference architecture enables 100+ accelerated server variations.

WPP, NVIDIA create digital ad content engine in Omniverse.

SoftBank, NVIDIA build 5G, gen AI data centers in Japan.

Networking technology accelerates Ethernet-based AI clouds.

NVIDIA ACE for Games breathes life into characters with gen AI.

Electronics manufacturers worldwide embrace NVIDIA AI.

Grace Hopper powers big-memory supercomputers for gen AI.

Modular reference architecture enables 100+ accelerated server variations.

WPP, NVIDIA create digital ad content engine in Omniverse.

SoftBank, NVIDIA build 5G, gen AI data centers in Japan.

Networking technology accelerates Ethernet-based AI clouds.

NVIDIA ACE for Games breathes life into characters with gen AI.

Electronics manufacturers worldwide embrace NVIDIA AI.

For enterprises that need the ultimate in AI performance, he unveiled DGX GH200 , a large-memory AI supercomputer. It uses NVIDIA NVLink to combine up to 256 NVIDIA GH200 Grace Hopper Superchips into a single data-center-sized GPU.

The GH200 Superchip, which Huang said is now in full production , combines an energy-efficient NVIDIA Grace CPU with a high-performance NVIDIA H100 Tensor Core GPU in one superchip.

The DGX GH200 packs an exaflop of performance and 144 terabytes of shared memory, nearly 500x more than in a single NVIDIA DGX A100 320GB system. That lets developers build large language models for generative AI chatbots, complex algorithms for recommender systems , and graph neural networks used for fraud detection and data analytics.

Google Cloud, Meta and Microsoft are among the first expected to gain access to the DGX GH200, which can be used as a blueprint for future hyperscale generative AI infrastructure.

“DGX GH200 AI supercomputers integrate NVIDIA’s most advanced accelerated computing and networking technologies to expand the frontier of AI,” Huang told the audience in Taipei, many of whom had lined up outside the hall for hours before the doors opened.

NVIDIA is building its own massive AI supercomputer, NVIDIA Helios, coming online this year. It will use four DGX GH200 systems linked with NVIDIA Quantum-2 InfiniBand networking to supercharge data throughput for training large AI models.

The DGX GH200 forms the pinnacle of hundreds of systems announced at the event. Together, they’re bringing generative AI and accelerated computing to millions of users.

Zooming out to the big picture, Huang announced more than 400 system configurations are coming to market powered by NVIDIA’s latest Hopper , Grace, Ada Lovelace and BlueField architectures. They aim to tackle the most complex challenges in AI, data science and high performance computing.

To fit the needs of data centers of every size, Huang announced NVIDIA MGX , a modular reference architecture for creating accelerated servers. System makers will use it to quickly and cost-effectively build more than a hundred different server configurations to suit a wide range of AI, HPC and NVIDIA Omniverse applications.

MGX lets manufacturers build CPU and accelerated servers using a common architecture and modular components. It supports NVIDIA's full line of GPUs, CPUs, data processing units (DPUs) and network adapters as well as x86 and Arm processors across a variety of air- and liquid-cooled chassis.

QCT and Supermicro will be the first to market with MGX designs appearing in August. Supermicro's ARS-221GL-NR system announced at COMPUTEX will use the Grace CPU, while QCT's S74G-2U system, also announced at the event, uses Grace Hopper.

ASRock Rack, ASUS, GIGABYTE and Pegatron will also use MGX to create next-generation accelerated computers.

Separately, Huang said NVIDIA is helping shape future 5G and 6G wireless and video communications. A demo showed how AI running on Grace Hopper will transform today's 2D video calls into more lifelike 3D experiences, providing an amazing sense of presence.

Laying the groundwork for new kinds of services, Huang announced NVIDIA is working with telecom giant SoftBank to build a distributed network of data centers in Japan. It will deliver 5G services and generative AI applications on a common cloud platform.

The data centers will use NVIDIA GH200 Superchips and NVIDIA BlueField-3 DPUs in modular MGX systems as well as NVIDIA Spectrum Ethernet switches to deliver the highly precise timing the 5G protocol requires. The platform will reduce cost by increasing spectral efficiency while reducing energy consumption.

The systems will help SoftBank explore 5G applications in autonomous driving, AI factories, augmented and virtual reality, computer vision and digital twins. Future uses could even include 3D video conferencing and holographic communications.

Separately, Huang unveiled NVIDIA Spectrum-X , a networking platform purpose-built to improve the performance and efficiency of Ethernet-based AI clouds. It combines Spectrum-4 Ethernet switches with BlueField-3 DPUs and software to deliver 1.7x gains in AI performance and power efficiency over traditional Ethernet fabrics.

NVIDIA Spectrum-X , Spectrum-4 switches and BlueField-3 DPUs are available now from system makers including Dell Technologies, Lenovo and Supermicro.

Generative AI impacts how people play, too.

Huang announced NVIDIA Avatar Cloud Engine (ACE) for Games , a foundry service developers can use to build and deploy custom AI models for speech, conversation and animation. It will give non-playable characters conversational skills so they can respond to questions with lifelike personalities that evolve.

NVIDIA ACE for Games includes AI foundation models such as NVIDIA Riva to detect and transcribe the player's speech. The text prompts NVIDIA NeMo to generate customized responses animated with NVIDIA Omniverse Audio2Face .

Huang described how NVIDIA and Microsoft are collaborating to drive innovation for Windows PCs in the generative AI era.

New and enhanced tools, frameworks and drivers are making it easier for PC developers to develop and deploy AI. For example, the Microsoft Olive toolchain for optimizing and deploying GPU-accelerated AI models and new graphics drivers will boost DirectML performance on Windows PCs with NVIDIA GPUs.

The collaboration will enhance and extend an installed base of 100 million PCs sporting RTX GPUs with Tensor Cores that boost performance of more than 400 AI-accelerated Windows apps and games.

Generative AI is also spawning new opportunities in the $700 billion digital advertising industry.

For example, WPP, the world's largest marketing services organization, is working with NVIDIA to build a first-of-its kind generative AI-enabled content engine on Omniverse Cloud .

In a demo, Huang showed how creative teams will connect their 3D design tools such as Adobe Substance 3D, to build digital twins of client products in NVIDIA Omniverse. Then, content from generative AI tools trained on responsibly sourced data and built with NVIDIA Picasso will let them quickly produce virtual sets. WPP clients can then use the complete scene to generate a host of ads, videos and 3D experiences for global markets and users to experience on any web device.

"Today ads are retrieved, but in the future when you engage information much of it will be generated — the computing model has changed," Huang said.

With an estimated 10 million factories, the $46 trillion manufacturing sector is a rich field for industrial digitalization.

"The world's largest industries make physical things. Building them digitally first can save billions," said Huang.

The keynote showed how electronics makers including Foxconn Industrial Internet, Innodisk, Pegatron, Quanta and Wistron are forging digital workflows with NVIDIA technologies to realize the vision of an entirely digital smart factory.

They're using Omniverse and generative AI APIs to connect their design and manufacturing tools so they can build digital twins of factories. In addition, they use NVIDIA Isaac Sim for simulating and testing robots and NVIDIA Metropolis , a vision AI framework, for automated optical inspection.

The latest component, NVIDIA Metropolis for Factories , can create custom quality-control systems, giving manufacturers a competitive advantage. It's helping companies develop state-of-the-art AI applications.

For example, Pegatron — which makes 300 products worldwide, including laptops and smartphones — is creating virtual factories with Omniverse, Isaac Sim and Metropolis. That lets it try out processes in a simulated environment, saving time and cost.

Pegatron also used the NVIDIA DeepStream software development kit to develop intelligent video applications that led to a 10x improvement in throughput.

Foxconn Industrial Internet, a service arm of the world's largest technology manufacturer, is working with NVIDIA Metropolis partners to automate significant portions of its circuit-board quality-assurance inspection points.

In a video, Huang showed how Techman Robot, a subsidiary of Quanta, tapped NVIDIA Isaac Sim to optimize inspection on the Taiwan-based giant's manufacturing lines. It's essentially using simulated robots to train robots how to make better robots.

In addition, Huang announced a new platform to enable the next generation of autonomous mobile robot (AMR) fleets. Isaac AMR helps simulate, deploy and manage fleets of autonomous mobile robots.

A large partner ecosystem — including ADLINK, Aetina, Deloitte, Quantiphi and Siemens — is helping bring all these manufacturing solutions to market, Huang said.

It's one more example of how NVIDIA is helping companies feel the benefits of generative AI with accelerated computing.

"It's been a long time since I've seen you, so I had a lot to tell you," he said after the two-hour talk to enthusiastic applause.

To learn more, check out the highlights of NVIDIA’s week at Computex 2023 in a four-minute video here . And watch the full keynote below

Original URL: https://blogs.nvidia.com/blog/2023/05/28/computex-keynote-generative-ai/