SIGGRAPH Special Address: NVIDIA CEO Brings Generative AI to LA Show

Speaking to thousands of developers and graphics pros, Jensen Huang announces updated GH200 Grace Hopper Superchip, NVIDIA AI Workbench, updates NVIDIA Omniverse with generative AI.

Author: Brian Caulfield

As generative AI continues to sweep an increasingly digital, hyperconnected world, NVIDIA founder and CEO Jensen Huang made a thunderous return to SIGGRAPH, the world's premier computer graphics conference.

"The generative AI era is upon us, the iPhone moment if you will," Huang told an audience of thousands Tuesday during an in-person special address in Los Angeles.

News highlights include the next-generation GH200 Grace Hopper Superchip platform , NVIDIA AI Workbench — a new unified toolkit that introduces simplified model tuning and deployment on NVIDIA AI platforms — and a major upgrade to NVIDIA Omniverse with generative AI and OpenUSD .

The announcements are about bringing all of the past decade's innovations — AI, virtual worlds, acceleration, simulation, collaboration and more — together.

"Graphics and artificial intelligence are inseparable, graphics needs AI, and AI needs graphics," Huang said, explaining that AI will learn skills in virtual worlds, and that AI will help create virtual worlds.

Five years ago at SIGGRAPH, NVIDIA reinvented graphics by bringing AI and real-time ray tracing to GPUs. But "while we were reinventing computer graphics with artificial intelligence, we were reinventing the GPU altogether for artificial intelligence," Huang said.

The result: increasingly powerful systems such as the NVIDIA HGX H100, which harnesses eight GPUs — and a total of 1 trillion transistors — that offer dramatic acceleration over CPU-based systems.

"This is the reason why the world's data centers are rapidly transitioning to accelerated computing," Huang told the audience. "The more you buy, the more you save."

To continue AI's momentum, NVIDIA created the Grace Hopper Superchip, the NVIDIA GH200, which combines a 72-core Grace CPU with a Hopper GPU, and which went into full production in May.

Huang announced that NVIDIA GH200, which is already in production, will be complemented with an additional version with cutting-edge HBM3e memory.

He followed up on that by announcing the next-generation GH200 Grace Hopper superchip platform with the ability to connect multiple GPUs for exceptional performance and easily scalable server design.

Built to handle the world's most complex generative workloads, spanning large language models, recommender systems and vector databases, the new platform will be available in a wide range of configurations.

The dual configuration — which delivers up to 3.5x more memory capacity and 3x more bandwidth than the current generation offering — comprises a single server with 144 Arm Neoverse cores, eight petaflops of AI performance, and 282GB of the latest HBM3e memory technology.

Leading system manufacturers are expected to deliver systems based on the platform in the second quarter of 2024.

To speed custom adoption of generative AI for the world's enterprises, Huang announced NVIDIA AI Workbench. It provides developers with a unified, easy-to-use toolkit to quickly create, test and fine-tune generative AI models on a PC or workstation — then scale them to virtually any data center, public cloud or NVIDIA DGX Cloud .

AI Workbench removes the complexity of getting started with an enterprise AI project. Accessed through a simplified interface running on a local system, it allows developers to fine-tune models from popular repositories such as Hugging Face, GitHub and NGC using custom data. The models can then be shared easily across multiple platforms.

While hundreds of thousands of pretrained models are now available, customizing them with the many open-source tools available can be challenging and time consuming.

"In order to democratize this ability, we have to make it possible to run pretty much everywhere," Huang said.

With AI Workbench, developers can customize and run generative AI in just a few clicks. It allows them to pull together all necessary enterprise-grade models, frameworks, software development kits and libraries into a unified developer workspace.

"Everybody can do this," Huang said.

Leading AI infrastructure providers — including Dell Technologies, Hewlett Packard Enterprise, HP Inc., Lambda, Lenovo and Supermicro — are embracing AI Workbench for its ability to bring enterprise generative AI capability to wherever developers want to work — including a local device.

Huang also announced a partnership between NVIDIA and startup Hugging Face , which has 2 million users, that will put generative AI supercomputing at the fingertips of millions of developers building large language models and other advanced AI applications.

Developers will be able to access NVIDIA DGX Cloud AI supercomputing within the Hugging Face platform to train and tune advanced AI models.

"This is going to be a brand new service to connect the world's largest AI community to the world's best training and infrastructure," Huang said.

In a video, Huang showed how AI Workbench and ChatUSD bring it all together: allowing a user to start a project on a GeForce RTX 4090 laptop and scale, seamlessly to a workstation, or the data center as it grows more complex.

Using Jupyter Notebook, a user can prompt the model to generate a picture of Toy Jensen in space. When the model provides a result that doesn't work, because it's never seen Toy Jensen, the user can fine-tune the model with eight images of Toy Jensen and then prompt it again to get a correct result.

Then with AI Workbench, the new model can be deployed to an enterprise application.

In a further step to accelerate the adoption of generative AI, NVIDIA announced the latest version of its enterprise software suite, NVIDIA AI Enterprise 4.0 .

NVIDIA AI Enterprise gives businesses access to the tools needed to adopt generative AI, while also offering the security and API stability required for large-scale enterprise deployments.

Offering new foundation applications and services for developers and industrial enterprises to optimize and enhance their 3D pipelines with the OpenUSD framework and generative AI , Huang announced a major release of NVIDIA Omniverse, an OpenUSD-native development platform for building, simulating, and collaborating across tools and virtual worlds.

He also announced NVIDIA's contributions to OpenUSD, the framework and universal interchange for describing, simulating and collaborating across 3D tools. Updates to the Omniverse platform include advancements to Omniverse Kit — the engine for developing native OpenUSD applications and extensions — as well as to the NVIDIA Omniverse Audio2Face foundation app and spatial-computing

capabilities .

Cesium, Convai, Move AI, SideFX Houdini and Wonder Dynamics are now connected to Omniverse via OpenUSD.

And expanding their collaboration across Adobe Substance 3D, generative AI and OpenUSD initiatives, Adobe and NVIDIA announced plans to make Adobe Firefly — Adobe's family of creative generative AI models — available as APIs in Omniverse.

Omniverse users can now build content, experiences and applications that are compatible with other OpenUSD-based spatial computing platforms such as ARKit and RealityKit. Huang announced a broad range of frameworks, resources and services for developers and companies to accelerate the adoption of Universal Scene Description, known as OpenUSD , including contributions such as geospatial data models, metrics assembly and simulation-ready, or SimReady , specifications for OpenUSD. Huang also announced four new Omniverse Cloud APIs built by NVIDIA for developers to more seamlessly implement and deploy OpenUSD pipelines and applications.

ChatUSD — Assisting developers and artists working with OpenUSD data and scenes, ChatUSD is a large language model (LLM) agent for generating Python-USD code scripts from text and answering USD knowledge questions.

RunUSD — a cloud API that translates OpenUSD files into fully path-traced rendered images by checking compatibility of the uploaded files against versions of OpenUSD releases, and generating renders with Omniverse Cloud.

DeepSearch — an LLM agent enabling fast semantic search through massive databases of untagged assets.

USD-GDN Publisher — a one-click service that enables enterprises and software makers to publish high-fidelity, OpenUSD-based experiences to the Omniverse Cloud Graphics Delivery Network (GDN) from an Omniverse-based application such as USD Composer , as well as stream in real time to web browsers and mobile devices.

These contributions are an evolution of last week's announcement of NVIDIA's co-founding of the Alliance for OpenUSD along with Pixar, Adobe, Apple and Autodesk.

Providing more computing power for all of this, Huang said NVIDIA and global workstation manufacturers are announcing powerful new RTX workstations for development and content creation in the age of generative AI and digitization.

The systems, including those from BOXX, Dell Technologies, HP and Lenovo, are based on NVIDIA RTX 6000 Ada Generation GPUs and incorporate NVIDIA AI Enterprise and NVIDIA Omniverse Enterprise software.

Separately, NVIDIA released three new desktop workstation Ada Generation GPUs — the NVIDIA RTX 5000 , RTX 4500 and RTX 4000 — to d eliver the latest AI, graphics and real-time rendering technology to professionals worldwide.

Huang also detailed how, together with global data center system manufacturers, NVIDIA is continuing to supercharge generative AI and industrial digitization with new NVIDIA OVX featuring the new NVIDIA L40S GPU, a powerful, universal data center processor design.

The powerful new systems will accelerate the most compute-intensive, complex applications, including AI training and inference, 3D design and visualization, video processing and industrial digitalization with the NVIDIA Omniverse platform.

More innovations are coming, thanks to NVIDIA Research.

At the show's Real Time Live Event, NVIDIA researchers will demonstrate a generative AI workflow that helps artists rapidly create and iterate on materials for 3D scenes, using text or image prompts to

generate custom textured materials faster and with finer creative control.

And NVIDIA Research also demo'd how AI can take video conferencing to the next level with new 3D features. NVIDIA Research recently published a paper demonstrating how AI could power a 3D video-conferencing system with minimal capture equipment.

The production version of Maxine, now available in NVIDIA Enterprise, allows professionals, teams, creators and others to tap into the power of AI to create high-quaity audio and video effects, even using standard microphone and webcams. Watch Huang's full special address at NVIDIA's SIGGRAPH event site . where there are also details of labs, presentations and more happening throughout the show.

Original URL: https://blogs.nvidia.com/blog/2023/08/08/siggraph-2023-special-address/