Keynote Wrap-Up: NVIDIA CEO Unveils Next-Gen RTX GPUs, AI Workflows in the Cloud

Kicking off GTC, Jensen Huang unveils advances in natural language understanding, the metaverse, gaming and AI technologies impacting industries from transportation and healthcare to finance and entertainment.

Author: Brian Caulfield

New cloud services to support AI workflows and the launch of a new generation of GeForce RTX GPUs featured today in NVIDIA CEO Jensen Huang's GTC keynote, which was packed with new systems, silicon, and software.

"Computing is advancing at incredible speeds, the engine propelling this rocket is accelerated computing, and its fuel is AI," Huang said during a virtual presentation as he kicked off NVIDIA GTC .

Again and again, Huang connected new technologies to new products to new opportunities – from harnessing AI to delight gamers with never-before-seen graphics to building virtual proving grounds where the world's biggest companies can refine their products.

Driving the deluge of new ideas, new products and new applications: a singular vision of accelerated computing unlocking advances in AI, which, in turn will touch industries around the world.

Gamers and creators will get the first GPUs based on the new NVIDIA Ada Lovelace architecture .

Enterprises will get powerful new tools for high-performance computing applications with systems based on the Grace CPU and Grace Hopper Superchip . Those building the 3D internet will get new OVX servers powered by Ada Lovelace L40 data center GPUs . Researchers and computer scientists get new large language model capabilities with NVIDIA LLMs NeMo Service . And the auto industry gets Thor, a new brain with an astonishing 2,000 teraflops of performance .

Huang highlighted how NVIDIA's technologies are being put to work by a sweep of major partners and customers across a breadth of industries.

To speed adoption, he announced Deloitte, the world's largest professional services firm, is bringing new services built on NVIDIA AI and NVIDIA Omniverse to the world's enterprises .

And he shared customer stories from telecoms giant Charter, as well as General Motors in the automotive industry, the German railway system's Deutsche Bahn in transportation, The Broad Institute in medical research , and Lowe's in retail .

NVIDIA GTC, which kicked off this week, has become one of the world's most important AI gatherings, with 200+ speakers from companies such as Boeing , Deutsche Bank , Lowe's , Polestar , Johnson & Johnson , Kroger , Mercedes-Benz , Siemens AG , T-Mobile and US Bank . More than 200,000 people have registered for the conference.

First out of the blocks at the keynote was the launch of next-generation GeForce RTX 40 Series GPUs powered by Ada, which Huang called a "quantum leap" that paves the way for creators of fully simulated worlds.

Huang gave his audience a taste of what that makes possible by offering up a look at Racer RTX, a fully interactive simulation that's entirely ray traced, with all the action physically modeled.

Ada's advancements include a new Streaming Multiprocessor, a new RT Core with twice the ray-triangle intersection throughput, and a new Tensor Core with the Hopper FP8 Transformer Engine and 1.4 petaflops of Tensor processor power.

Ada also introduces the latest version of NVIDIA DLSS technology , DLSS 3, which uses AI to generate new frames by comparing new frames with prior frames to understand how a scene is changing. The result: boosting game performance by up to 4x over brute force rendering.

DLSS 3 has received support from many of the world's leading game developers, with more than 35 games and applications announcing support. "DLSS 3 is one of our greatest neural rendering inventions," Huang said.

Together, Huang said, these innovations help deliver 4x more processing throughput with the new GeForce RTX 4090 versus its forerunner, the RTX 3090 Ti. "The new heavyweight champ" starts at $1,599 and will be available Oct. 12.

Additionally, the new GeForce RTX 4080 is launching in November with two configurations.

The GeForce RTX 4080 16GB, priced at $1,199, has 9,728 CUDA cores and 16GB of high-speed Micron GDDR6X memory. With DLSS 3, it's twice as fast in today's games as the GeForce RTX 3080 Ti, and more powerful than the GeForce RTX 3090 Ti at lower power.

The GeForce RTX 4080 12GB has 7,680 CUDA cores and 12GB of Micron GDDR6X memory, and with DLSS 3 is faster than the RTX 3090 Ti, the previous-generation flagship GPU. It's priced at $899.

Huang also announced that NVIDIA Lightspeed Studios used Omniverse to reimagine Portal , one of the most celebrated games in history. With NVIDIA RTX Remix , an AI-assisted toolset, users can mod their favorite games, enabling them to up-res textures and assets, and give materials physically accurate properties.

Once more tying systems and software to broad technology trends, Huang explained that large language models, or LLMs, and recommender systems are the two most important AI models today.

Recommenders "run the digital economy," powering everything from e-commerce to entertainment to advertising, he said. "They're the engines behind social media, digital advertising, e-commerce and search."

And large language models based on the Transformer deep learning model first introduced in 2017 are now among the most vibrant areas for research in AI, and able to learn to understand human language without supervision or labeled datasets.

"A single pre-trained model can perform multiple tasks, like question answering, document summarization, text generation, translation and even software programming," Huang said.

Delivering the computing muscle needed to power these enormous models, Huang said the NVIDIA H100 Tensor Core GPU, with Hopper's next-generation Transformer Engine, is in full production, with systems shipping in the coming weeks.

"Hopper is in full production and coming soon to power the world's AI factories," Huang said.

Partners building systems include Atos, Cisco, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Lenovo and Supermicro. And Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure will be among the first to deploy H100-based instances in the cloud starting next year.

And Grace Hopper, which combines NVIDIA's Arm-based Grace data center CPU with Hopper GPUs, with its 7x increase in fast-memory capacity, will deliver a "giant leap" for recommender systems, Huang said. Systems incorporating Grace Hopper will be available in the first half of 2023.

The next evolution of the internet, called the metaverse, will be extended with 3D, Huang explained. Omniverse is NVIDIA's platform for building and running metaverse applications.

Here, too, Huang explained how connecting and simulating these worlds will require powerful, flexible new computers. And NVIDIA OVX servers are built for scaling out metaverse applications.

NVIDIA's 2nd-generation OVX systems will be powered by Ada Lovelace L40 data center GPUs, which are now in full production, Huang announced.

In today's vehicles, active safety, parking, driver monitoring, camera mirrors, cluster and infotainment are driven by different computers. In the future, they'll be delivered by software that improves over time, running on a centralized computer, Huang said.

To power this, Huang introduced DRIVE Thor, which combines the transformer engine of Hopper, the GPU of Ada, and the amazing CPU of Grace.

The new Thor superchip delivers 2,000 teraflops of performance, replacing Atlan on the DRIVE roadmap, and providing a seamless transition from DRIVE Orin, which has 254 TOPS of performance and is currently in production vehicles. Thor will be the processor for robotics, medical instruments, industrial automation and edge AI systems, Huang said.

Bringing NVIDIA's systems and silicon, and the benefits of accelerated computing, to industries around the world, is a software ecosystem with more than 3.5 million developers creating some 3,000 accelerated apps using NVIDIA's 550 software development kits, or SDKs, and AI models, Huang announced.

And it's growing fast. Over the past 12 months, NVIDIA has updated more than 100 SDKs and introduced 25 new ones.

"New SDKs increase the capability and performance of systems our customers already own, while opening new markets for accelerated computing," Huang said.

Large language models "are the most important AI models today," Huang said. Based on the transformer architecture, these giant models can learn to understand meanings and languages without supervision or labeled datasets, unlocking remarkable new capabilities.

To make it easier for researchers to apply this "incredible" technology to their work, Huang announced the Nemo LLM Service, an NVIDIA-managed cloud service to adapt pretrained LLMs to perform specific tasks.

To accelerate the work of drug and bioscience researchers, Huang also announced BioNeMo LLM, a service to create LLMs that understand chemicals, proteins, DNA and RNA sequences.

Huang announced that NVIDIA is working with The Broad Institute, the world's largest producer of human genomic information, to make NVIDIA Clara libraries, such as NVIDIA Parabricks, the Genome Analysis Toolkit, and BioNeMo, available on Broad's Terra Cloud Platform.

Huang also detailed NVIDIA Omniverse Cloud , an infrastructure-as-a-service that connects Omniverse applications running in the cloud, on premises or on a device.

New Omniverse containers – Replicator for synthetic data generation, Farm for scaling render farms, and Isaac Sim for building and training AI robots – are now available for cloud deployment, Huang announced.

Omniverse is seeing wide adoption, and Huang shared several customer stories and demos:

Lowe's, which has nearly 2,000 retail outlets, is using Omniverse to design, build and operate digital twins of their stores;

Charter, a $50 billion dollar telecoms provider, and interactive data analytics provider HeavyAI, are using Omniverse to create digital twins of Charter's 4G and 5G networks;

GM is creating a digital twin of its Michigan Design Studio in Omniverse where designers, engineers and marketers can collaborate.

Shifting from virtual worlds to machines that will move through their world, robotic computers "are the newest types of computers," Huang said, describing NVIDIA's second-generation processor for

robotics, Orin, as a homerun.

To bring Orin to more markets, he announced the Jetson Orin Nano , a tiny robotics computer that is 80x faster than the previous super-popular Jetson Nano.

Jetson Orin Nano runs the NVIDIA Isaac robotics stack and features the ROS 2 GPU-accelerated framework, and NVIDIA Iaaac Sim, a robotics simulation platform, is available on the cloud.

And for robotics developers using AWS RoboMaker, Huang announced that containers for the NVIDIA Isaac platform for robotics development are in the AWS marketplace .

Most of the world's internet traffic is video, and user-generated video streams will be increasingly augmented by AI special effects and computer graphics, Huang explained.

"Avatars will do computer vision, speech AI, language understanding and computer graphics in real time and at cloud scale," Huang said.

To enable new innovations at the intersection of real-time graphics, AI and communications possible, Huang announced NVIDIA has been building acceleration libraries like CV-CUDA , a cloud runtime engine called UCF Unified Computing Framework, Omniverse ACE Avatar Cloud Engine , and a sample application called Tokkio for customer service avatars.

And to speed the adoption of all these technologies to the world's enterprises, Deloitte, the world's largest professional services firm, is bringing new services built on NVIDIA AI and NVIDIA Omniverse to the world's enterprises, Huang announced.

He said that Deloitte's professionals will help the world's enterprises use NVIDIA application frameworks to build modern multi-cloud applications for customer service, cybersecurity, industrial automation, warehouse and retail automation and more.

Huang ended his keynote by recapping a talk that moved from outlining new technologies to product announcements and back — uniting scores of different parts into a singular vision.

"Today, we announced new chips, new advances to our platforms, and, for the very first time, new cloud services," Huang said as he wrapped up. "These platforms propel new breakthroughs in AI, new applications of AI, and the next wave of AI for science and industry."

Original URL: https://blogs.nvidia.com/blog/2022/09/20/keynote-gtc-nvidia-ceo/