NVIDIA to Bring AI to Every Industry, CEO Says

From AI training to deployment, semiconductors to software libraries, systems to cloud services, NVIDIA CEO Jensen Huang outlined how a new generation of breakthroughs will be put at the world's fingertips.

Author: Brian Caulfield

ChatGPT is just the start.

With computing now advancing at what he called "lightspeed," NVIDIA founder and CEO Jensen Huang today announced a broad set of partnerships with Google, Microsoft, Oracle and a range of leading businesses that bring new AI, simulation and collaboration capabilities to every industry.

"The warp drive engine is accelerated computing, and the energy source is AI," Huang said in his keynote at the company's GTC conference. "The impressive capabilities of generative AI have created a sense of urgency for companies to reimagine their products and business models."

In a sweeping 78-minute presentation anchoring the four-day event, Huang outlined how NVIDIA and its partners are offering everything from training to deployment for cutting-edge AI services. He announced new semiconductors and software libraries to enable fresh breakthroughs. And Huang revealed a complete set of systems and services for startups and enterprises racing to put these innovations to work on a global scale.

Huang punctuated his talk with vivid examples of this ecosystem at work. He announced NVIDIA and Microsoft will connect hundreds of millions of Microsoft 365 and Azure users to a platform for building and operating hyperrealistic virtual worlds. He offered a peek at how Amazon is using sophisticated simulation capabilities to train new autonomous warehouse robots. He touched on the rise of a new generation of wildly popular generative AI services such as ChatGPT.

And underscoring the foundational nature of NVIDIA's innovations, Huang detailed how, together with ASML, TSMC and Synopsis, NVIDIA computational lithography breakthroughs will help make a new generation of efficient, powerful 2-nm semiconductors possible.

The arrival of accelerated computing and AI come just in time, with Moore's Law slowing and industries tackling powerful dynamics —sustainability, generative AI, and digitalization, Huang said. "Industrial companies are racing to digitalize and reinvent into software-driven tech companies — to be the disruptor and not the disrupted," Huang said.

Acceleration lets companies meet these challenges. "Acceleration is the best way to reclaim power and achieve sustainability and Net Zero," Huang said.

GTC, now in its 14th year, has become one of the world's most important AI gatherings. This week's conference features 650 talks from leaders such as Demis Hassabis of DeepMind , Valeri Taylor of Argonne Labs , Scott Belsky of Adobe , Paul Debevec of Netflix , Thomas Schulthess of ETH Zurich and a special fireside chat between Huang and Ilya Sutskever, co-founder of OpenAI, the creator of ChatGPT .

More than 250,000 registered attendees will dig into sessions on everything from restoring the lost Roman mosaics of 2,000 years ago to building the factories of the future, from exploring the universe with a new generation of massive telescopes to rearranging molecules to accelerate drug discovery , to more than 70 talks on generative AI.

NVIDIA's technologies are fundamental to AI, with Huang recounting how NVIDIA was there at the very beginning of the generative AI revolution. Back in 2016 he hand-delivered to OpenAI the first NVIDIA DGX AI supercomputer — the engine behind the large language model breakthrough powering

ChatGPT.

Launched late last year, ChatGPT went mainstream almost instantaneously, attracting over 100 million users, making it the fastest-growing application in history. “We are at the iPhone moment of AI,” Huang said.

NVIDIA DGX supercomputers, originally used as an AI research instrument, are now running 24/7 at businesses across the world to refine data and process AI, Huang reported. Half of all Fortune 100 companies have installed DGX AI supercomputers.

“DGX supercomputers are modern AI factories,” Huang said.

Deploying LLMs like ChatGPT are a significant new inference workload, Huang said. For large-language-model inference, like ChatGPT, Huang announced a new GPU — the H100 NVL with dual-GPU NVLink.

Based on NVIDIA’s Hopper architecture, H100 features a Transformer Engine designed to process models such as the GPT model that powers ChatGPT. Compared to HGX A100 for GPT-3 processing, a standard server with four pairs of H100 with dual-GPU NVLink is up to 10x faster.

“H100 can reduce large language model processing costs by an order of magnitude,” Huang said.

Meanwhile, over the past decade, cloud computing has grown 20% annually into a $1 trillion industry, Huang said. NVIDIA designed the Grace CPU for an AI- and cloud-first world, where AI workloads are GPU accelerated. Grace is sampling now , Huang said.

NVIDIA’s new superchip, Grace Hopper, connects the Grace CPU and Hopper GPU over a high-speed 900GB/sec coherent chip-to-chip interface. Grace Hopper is ideal for processing giant datasets like AI databases for recommender systems and large language models, Huang explained.

“Customers want to build AI databases several orders of magnitude larger,” Huang said. “Grace Hopper is the ideal engine.”

The latest version of DGX features eight NVIDIA H100 GPUs linked together to work as one giant GPU. “NVIDIA DGX H100 is the blueprint for customers building AI infrastructure worldwide,” Huang said, sharing that NVIDIA DGX H100 is now in full production.

H100 AI supercomputers are already coming online.

Oracle Cloud Infrastructure announced the limited availability of new OCI Compute bare-metal GPU instances featuring H100 GPUs.

Additionally, Amazon Web Services announced its forthcoming EC2 UltraClusters of P5 instances, which can scale in size up to 20,000 interconnected H100 GPUs.

This follows Microsoft Azure’ s private preview announcement last week for its H100 virtual machine, ND H100 v5.

Meta has now deployed its H100-powered Grand Teton AI supercomputer internally for its AI production and research teams.

And OpenAI will be using H100s on its Azure supercomputer to power its continuing AI research.

Other partners making H100 available include Cirrascale and CoreWeave , both which announced general availability today. Additionally, Google Cloud, Lambda , Paperspace and Vultr are planning to offer H100.

And servers and systems featuring NVIDIA H100 GPUs are available from leading server makers including Atos, Cisco, Dell Technologies, GIGABYTE, Hewlett Packard Enterprise, Lenovo and Supermicro.

And to speed DGX capabilities to startups and enterprises racing to build new products and develop AI strategies, Huang announced NVIDIA DGX Cloud , through partnerships with Microsoft Azure, Google Cloud and Oracle Cloud Infrastructure to bring NVIDIA DGX AI supercomputers “to every company, from a browser.”

DGX Cloud is optimized to run NVIDIA AI Enterprise , the world’s leading acceleration software suite for end-to-end development and deployment of AI. “DGX Cloud offers customers the best of NVIDIA AI and the best of the world’s leading cloud service providers,” Huang said.

NVIDIA is partnering with leading cloud service providers to host DGX Cloud infrastructure, starting with Oracle Cloud Infrastructure. Microsoft Azure is expected to begin hosting DGX Cloud next quarter, and the service will soon expand to Google Cloud and more.

This partnership brings NVIDIA’s ecosystem to cloud service providers while amplifying NVIDIA’s scale and reach, Huang said. Enterprises will be able to rent DGX Cloud clusters on a monthly basis, ensuring they can quickly and easily scale the development of large, multi-node training workloads.

To accelerate the work of those seeking to harness generative AI, Huang announced NVIDIA AI Foundations , a family of cloud services for customers needing to build, refine and operate custom LLMs and generative AI trained with their proprietary data and for domain-specific tasks.

AI Foundations services include NVIDIA NeMo for building custom language text-to-text generative models ; Picasso, a visual language model-making service for customers who want to build custom models trained with licensed or proprietary content ; and BioNeMo, to help researchers in the $2 trillion drug discovery industry.

Adobe is partnering with NVIDIA to build a set of next-generation AI capabilities for the future of creativity.

Getty Images is collaborating with NVIDIA to train responsible generative text-to-image and text-to-video foundation models.

Shutterstock is working with NVIDIA to train a generative text-to-3D foundation model to simplify the creation of detailed 3D assets.

And NVIDIA announced Amgen is accelerating drug discovery services with BioNeMo. In addition, Alchemab Therapeutics, AstraZeneca, Evozyne, Innophore and Insilico are all early access users of BioNemo.

BioNeMo helps researchers create, fine-tune and serve custom models with their proprietary data, Huang explained.

Huang also announced that NVIDIA and Medtronic , the world’s largest healthcare technology provider, are partnering to build an AI platform for software-defined medical devices. The partnership will create a common platform for Medtronic systems, ranging from surgical navigation to robotic-assisted surgery.

And today Medtronic announced that its GI Genius system, with AI for early detection of colon cancer, is built on NVIDIA Holoscan, a software library for real-time sensor processing systems, and will ship around the end of this year.

“The world’s $250 billion medical instruments market is being transformed,” Huang said.

To help companies deploy rapidly emerging generative AI models, Huang announced inference platforms for AI video, image generation, LLM deployment and recommender inference . They combine NVIDIA’s full stack of inference software with the latest NVIDIA Ada, Hopper and Grace Hopper processors — including the NVIDIA L4 Tensor Core GPU and the NVIDIA H100 NVL GPU , both launched today.

• NVIDIA L4 for AI Video can deliver 120x more AI-powered video performance than CPUs, combined with 99% better energy efficiency.

• NVIDIA L40 for Image Generation is optimized for graphics and AI-enabled 2D, video and 3D image generation.

• NVIDIA H100 NVL for Large Language Model Deployment is ideal for deploying massive LLMs like ChatGPT at scale.

• And NVIDIA Grace Hopper for Recommendation Models is ideal for graph recommendation models, vector databases and graph neural networks.

Google Cloud is the first cloud service provider to offer L4 to customers with the launch of its new G2 virtual machines, available in private preview today. Google is also integrating L4 into its Vertex AI model store.

Unveiling a second cloud service to speed unprecedented simulation and collaboration capabilities to enterprises, Huang announced NVIDIA is partnering with Microsoft to bring NVIDIA Omniverse Cloud, a fully managed cloud service, to the world's industries .

"Microsoft and NVIDIA are bringing Omniverse to hundreds of millions of Microsoft 365 and Azure users," Huang said, also unveiling new NVIDIA OVX servers and a new generation of workstations powered by NVIDIA RTX Ada Generation GPUs and Intel's newest CPUs optimized for NVIDIA Omniverse .

To show the extraordinary capabilities of Omniverse, NVIDIA's open platform built for 3D design collaboration and digital twin simulation, Huang shared a video showing how NVIDIA Isaac Sim, NVIDIA's robotics simulation and synthetic generation platform, built on Omniverse, is helping Amazon save time and money with full-fidelity digital twins.

It shows how Amazon is working to choreograph the movements of Proteus, Amazon's first fully autonomous warehouse robot, as it moves bins of products from one place to another in Amazon's cavernous warehouses alongside humans and other robots.

Illustrating the scale of Omniverse's reach and capabilities, Huang dug into Omniverse's role in digitalizing the $3 trillion auto industry . By 2030, auto manufacturers will build 300 factories to make 200 million electric vehicles, Huang said, and battery makers are building 100 more megafactories. "Digitalization will enhance the industry's efficiency, productivity and speed," Huang said.

Touching on Omniverse's adoption across the industry, Huang said Lotus is using Omniverse to virtually assemble welding stations. Mercedes-Benz uses Omniverse to build, optimize and plan assembly lines for new models. Rimac and Lucid Motors use Omniverse to build digital stores from actual design data that faithfully represent their cars.

Working with Idealworks, BMW uses Isaac Sim in Omniverse to generate synthetic data and scenarios to train factory robots. And BMW is using Omniverse to plan operations across factories worldwide and is building a new electric-vehicle factory, completely in Omniverse, two years before the plant opens, Huang said.

Separately. NVIDIA today announced that BYD, the world's leading manufacturer of new energy vehicles NEVs, will extend its use of the NVIDIA DRIVE Orin centralized compute platform in a broader range of its NEVs.

Enabling semiconductor leaders such as ASML, TSMC and Synopsis to accelerate the design and manufacture of a new generation of chips as current production processes near the limits of what physics makes possible, Huang announced NVIDIA cuLitho , a breakthrough that brings accelerated computing to the field of computational lithography.

The new NVIDIA cuLitho software library for computational lithography is being integrated by TSMC, the world's leading foundry, as well as electronic design automation leader Synopsys into their software, manufacturing processes and systems for the latest-generation NVIDIA Hopper architecture GPUs.

Chip-making equipment provider ASML is working closely with NVIDIA on GPUs and cuLitho, and plans to integrate support for GPUs into all of their computational lithography software products. With lithography at the limits of physics, NVIDIA's introduction of cuLitho enables the industry to go to 2nm and beyond, Huang said.

"The chip industry is the foundation of nearly every industry," Huang said.

Companies around the world are on board with Huang's vision.

Telecom giant AT&T; uses NVIDIA AI to more efficiently process data and is testing Omniverse ACE and the Tokkio AI avatar workflow to build, customize and deploy virtual assistants for customer service and its employee help desk.

American Express, the U.S. Postal Service, Microsoft Office and Teams, and Amazon are among the 40,000 customers using the high-performance NVIDIA TensorRT inference optimizer and runtime, and NVIDIA Triton, a multi-framework data center inference serving software.

Uber uses Triton to serve hundreds of thousands of ETA predictions per second.

And with over 60 million daily users, Roblox uses Triton to serve models for game recommendations, build avatars, and moderate content and marketplace ads.

Microsoft, Tencent and Baidu are all adopting NVIDIA CV-CUDA for AI computer vision. The technology, in open beta, optimizes pre- and post-processing, delivering 4x savings in cost and energy.

Wrapping up his talk, Huang thanked NVIDIA's systems, cloud and software partners, as well as researchers, scientists and employees.

NVIDIA has updated 100 acceleration libraries, including cuQuantum and the newly open-sourced CUDA Quantum for quantum computing, cuOpt for combinatorial optimization, and cuLitho for computational lithography, Huang announced.

The global NVIDIA ecosystem, Huang reported, now spans 4 million developers, 40,000 companies and 14,000 startups in NVIDIA Inception.

"Together," Huang said. "We are helping the world do the impossible."

Original URL: https://blogs.nvidia.com/blog/2023/03/21/gtc-keynote-spring-2023/