No Hang Ups With Hangul: KT Trains Smart Speakers, Customer Call Centers With NVIDIA AI

South Korea's leading mobile operator builds billion-parameter large language models trained with the NVIDIA DGX SuperPOD platform and NeMo framework.

September 20, 2022 by Angie Lee

Share

Reading Time: 3 mins

South Korea's most popular AI voice assistant, GiGA Genie, has conversed with 8 million people.

The AI-powered speaker from telecom company KT can control TVs, offer real-time traffic updates and complete a slew of other home-assistance tasks based on voice commands. It has mastered its conversational skills in the highly complex Korean language thanks to large language models (LLMs) — machine learning algorithms that can recognize, understand, predict and generate human languages based on huge text datasets.

Feedback

The company's models are built using the NVIDIA DGX SuperPOD data center infrastructure platform and the NeMo framework for training and deploying LLMs with billions of parameters.

The Korean language, known as Hangul, reliably shows up in lists of the world's most challenging languages. It includes four types of compound verbs, and words are often composed of two or more roots.

KT — South Korea's leading mobile operator with over 22 million subscribers — improved the smart speaker's understanding of such words by developing LLMs. And through integration with Amazon Alexa, GiGA Genie can converse with users in English, too.

"With transformer-based models, we've achieved significant quality improvements for the GiGA Genie smart speaker, as well as our customer services platform AI Contact Center, or AICC," said Hwijung Ryu, LLM development team lead at KT.

AICC is an all-in-one, cloud-based platform that offers AI voice agents and other customer service-related applications.

It can receive calls and provide requested information — or quickly connect customers to human agents for answers to more detailed inquiries. AICC without human intervention manages more than 100,000 calls daily across Korea, according to Ryu.

"LLMs enable GiGA Genie to gain better language understanding and generate more human-like sentences, and AICC to reduce consultation times by 15 seconds as it summarizes and classifies inquiry types more quickly," he added.

Training Large Language Models

Developing LLMs can be an expensive, time-consuming process that requires deep technical expertise and full-stack technology investments.

The NVIDIA AI platform simplified and sped up this process for KT.

"We trained our LLM models more effectively with NVIDIA DGX SuperPOD's powerful performance — as well as NeMo's optimized algorithms and 3D parallelism techniques," Ryu said. "NeMo is continuously adopting new features, which is the biggest advantage we think it offers in improving our model accuracy."

3D parallelism — a distributed training method in which an extremely large-scale deep learning model is partitioned across multiple devices — was crucial for training KT's LLMs. NeMo enabled the team to easily accomplish this task with the highest throughput, according to Ryu.

"We considered using other platforms, but it was difficult to find an alternative that provides full-stack environments — from the hardware level to the inference level," he added. "NVIDIA also provides exceptional expertise from product, engineering teams and more, so we easily solved several technical issues."

Using hyperparameter optimization tools in NeMo, KT trained its LLMs 2x faster than with other frameworks, Ryu said. These tools allow users to automatically find the best configurations for LLM training and inference, easing and speeding the

development and deployment process.

KT is also planning to use the NVIDIA Triton Inference Server to provide an optimized real-time inference service, as well as NVIDIA Base Command Manager to easily monitor and manage hundreds of nodes in its AI cluster.

"Thanks to LLMs, KT can release competitive products faster than ever," Ryu said. "We also believe that our technology can drive innovation from other companies, as it can be used to improve their value and create innovative products."

KT plans to release more than 20 natural language understanding and natural language generation APIs for developers in November. The application programming interfaces can be used for tasks including document summarization and classification, emotion recognition, and filtering of potentially inappropriate content.

Learn more about breakthrough technologies for the era of AI and the metaverse at NVIDIA GTC, running online through Thursday, Sept. 22.

Watch NVIDIA founder and CEO Jensen Huang's keynote address in replay below:

Categories: Data Center

Tags: Artificial Intelligence | Conversational AI | GTC 2022 | NVIDIA DGX | NVIDIA NeMo | Telecommunications

Comments for this thread are now closed

0 Comments

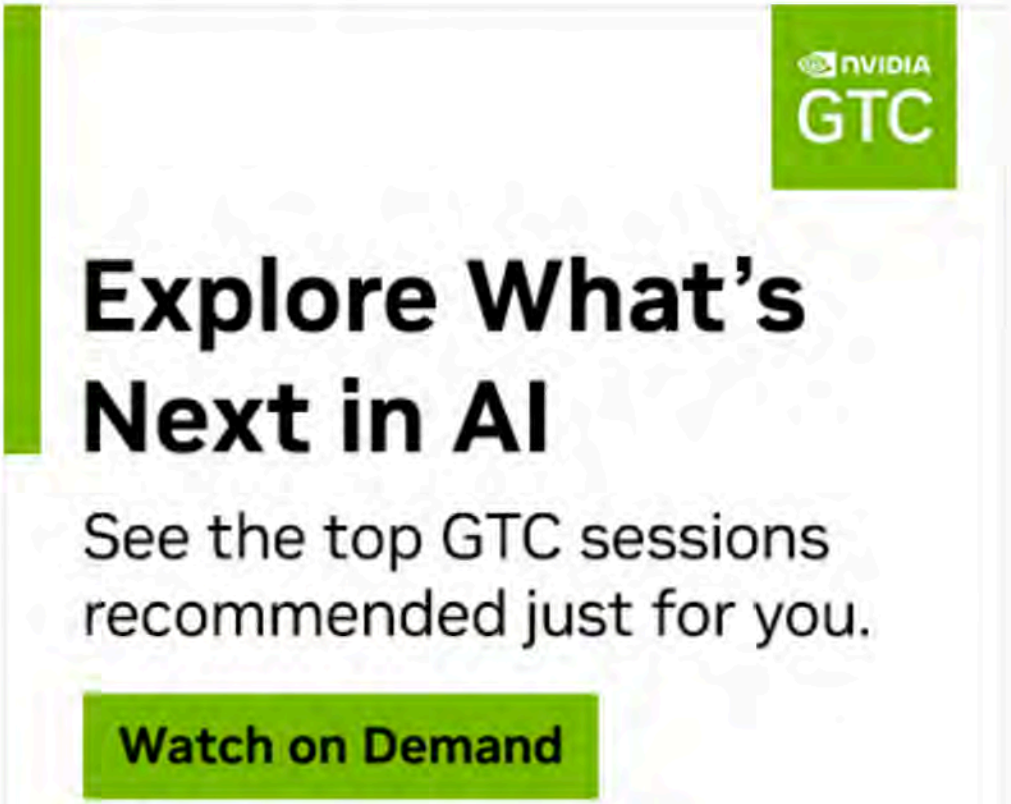
Login

Share

Best Newest Oldest

This discussion has been closed.

Subscribe Privacy Do Not Sell My Data DISQUS



All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

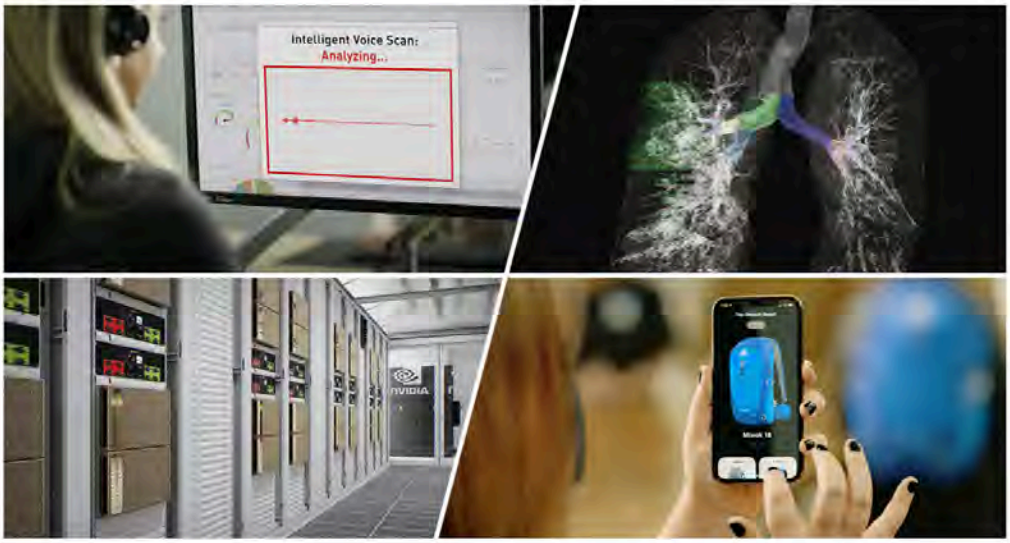
NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

New NVIDIA DGX System Software

and Infrastructure Solutions Supercharge Enterprise AI

Adept AI, Hyundai Motor Group, KT and the University of Wisconsin-Madison are among the latest innovators to deliver AI built on DGX systems, Base Command and DGX SuperPOD.

September 20, 2022 by Tony Paikeday



Share

Reading Time: 3 mins

At GTC today, NVIDIA unveiled a number of updates to its DGX portfolio to power new breakthroughs in enterprise AI development.

NVIDIA DGX H100 systems are now available for order. These infrastructure building blocks support NVIDIA's full-stack enterprise AI solutions.

With 32 petaflops of performance at FP8 precision, NVIDIA DGX H100 delivers a leap in efficiency for enterprise AI development. It offers 3x lower total cost of ownership and 3.5x more energy efficiency compared to the previous generation.

New NVIDIA Base Command software, which simplifies and speeds AI development, powers every DGX system — from single nodes to DGX SuperPODs.

Also unveiled was NVIDIA DGX BasePOD — the evolution of DGX POD — which makes enterprise data-center AI deployments simpler and faster for IT teams to acquire, deploy and manage.

Many of the world's AI leaders are building technological breakthroughs — from self-driving cars to voice assistants — using NVIDIA DGX systems and software, and the pace of innovation is not slowing down.

New NVIDIA Base Command Features

NVIDIA Base Command provides enterprise-grade orchestration and cluster management, and it now features a full software stack for maximizing AI developer productivity, IT manageability and workload performance.

The workflow management features of Base Command now include support for on-premises DGX SuperPOD environments, enabling businesses to gain centralized control of AI development projects with simplified collaboration for project teams, and integrated monitoring and reporting dashboards.

Base Command works with the NVIDIA AI Enterprise software suite, which is now included with every DGX system. The NVIDIA AI software enables end-to-end AI development and deployment with supported AI and data science tools, optimized frameworks and pretrained models.

Additionally, it offers enterprise-workflow management and MLOps integrations with DGX-Ready Software providers Domino Data Lab, Run.ai, Weights & Biases and NVIDIA Inception member Rescale. It also includes libraries that optimize and accelerate compute, storage and network infrastructure — while ensuring maximized system uptime, security and reliability.

New DGX BasePOD Reference Architecture

DGX BasePOD provides a reference architecture for DGX systems that incorporates design best practices for integrating compute, networking, storage and software.

Customers are already using NVIDIA DGX POD to power the development of a broad range of enterprise applications. DGX BasePOD builds on the success of DGX POD with new industry solutions targeting the biggest AI opportunities, including natural language processing, healthcare and life sciences, and fraud detection.

Delivered as fully integrated, ready-to-deploy offerings through the NVIDIA Partner Network, DGX BasePOD solutions range in size, from two to hundreds of DGX systems, with certified high-performance storage from NVIDIA DGX storage technology partners including DDN, Dell, NetApp, Pure Storage, VAST Data and WEKA.

Leaders Power AI Breakthroughs With DGX Systems

Enterprises around the world choose NVIDIA DGX systems to power their most advanced AI workloads. Among the AI innovators developing mission-critical AI capabilities on DGX A100 systems:

- ML research and product lab **Adept** is building an AI teammate powered by a large language model prototyped on NVIDIA DGX Foundry, and then scaled with NVIDIA A100 GPUs and NVIDIA Megatron on Oracle Cloud Infrastructure.
- **Hyundai Motor Group** is using a 40-node DGX SuperPOD to explore hyperscale AI workloads.
- Telecom company **KT** is developing a LLM with around 200 billion parameters for a variety of Korean-language applications, including the GiGA Genie smart speaker, using the NVIDIA NeMo Megatron framework, NVIDIA DGX SuperPOD and NVIDIA Base Command software.
- **The University of Wisconsin–Madison** is [illegible] AI to medical imaging

- **The University of Wisconsin-Madison** is quickly bringing AI to medical imaging devices using NVIDIA DGX systems with the Flywheel research platform and the NVIDIA Clara healthcare application framework. Using the NVIDIA Federated Learning Application Runtime Environment, or NVIDIA FLARE, in collaboration with other hospitals, the university is securely training AI models on DGX systems for medical imaging, annotation and classification.

Learn more about the AI breakthroughs powered by NVIDIA DGX systems by watching NVIDIA founder and CEO Jensen Huang's GTC keynote in replay. And join the GTC session, "Designing Your AI Center of Excellence," with Charlie Boyle, vice president of DGX systems at NVIDIA.

Categories: Data Center

Tags: Artificial Intelligence | Data Science | GTC 2022 | NVIDIA DGX | NVIDIA NeMo

Load Comments

NVIDIA GTC

Explore What's Next in AI

See the top GTC sessions recommended just for you.

Watch on Demand

All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

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.

September 20, 2022 by Brian Caulfield

Share

Reading Time: 8 mins

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.

A 'Quantum Leap': GeForce RTX 40 Series GPUs

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.

NVIDIA CEO Jensen Huang launched the next-generation GeForce RTX 40 Series GPUs

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.

NVIDIA Lightspeed Studios used Omniverse to reimagine Portal, one of the most celebrated games in history

Powering AI Advances, H100 GPU in Full Production

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.

Weaving Together the Metaverse, L40 Data Center GPUs in Full Production

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.

Thor for Autonomous Vehicles, Robotics, Medical Instruments and More

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.

3.5 Million Developers, 3,000 Accelerated Applications

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.

New Services for AI, Virtual Worlds

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.

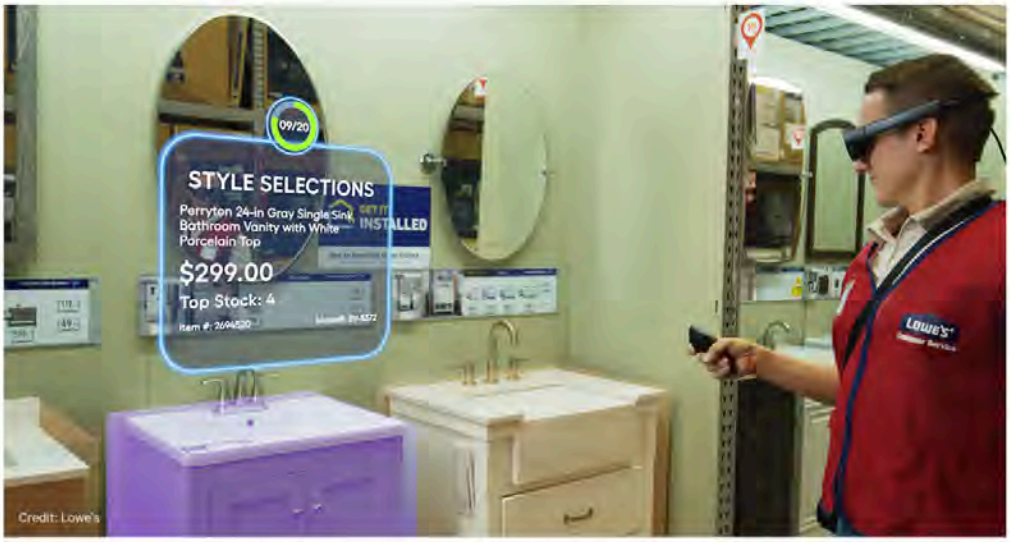
NVIDIA is working with The Broad Institute, the world's largest producer of human genomic information, to make NVIDIA Clara libraries available on Broad's Terra Cloud Platform.

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.



Home improvement retailer Lowe's is using Omniverse to design, build and operate digital twins of their stores.

New Jetson Orin Nano for Robotics

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.

New Tools for Video, Image Services

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.

Deloitte to Bring AI, Omniverse Services to Enterprises

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.

Just Getting Started

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."

Categories: Corporate | Data Center | Deep Learning | Driving | Gaming

Tags: Conversational AI | GeForce | GTC | GTC 2022 | Metaverse | NVIDIA DRIVE | NVIDIA DRIVE Sim | NVIDIA NeMo | NVIDIA RTX | Omniverse | Rendering

Load Comments

Next in AI

See the top GTC sessions recommended just for you.

Watch on Demand

All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

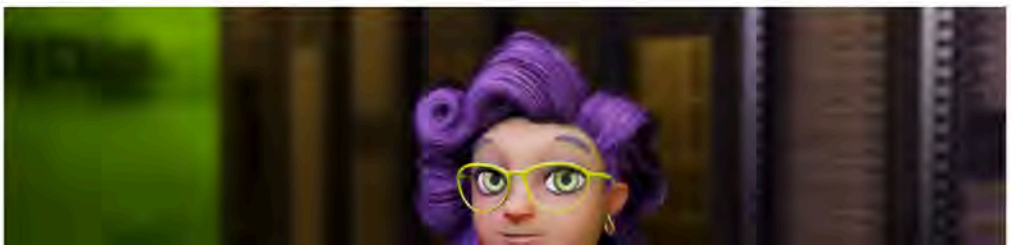
Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

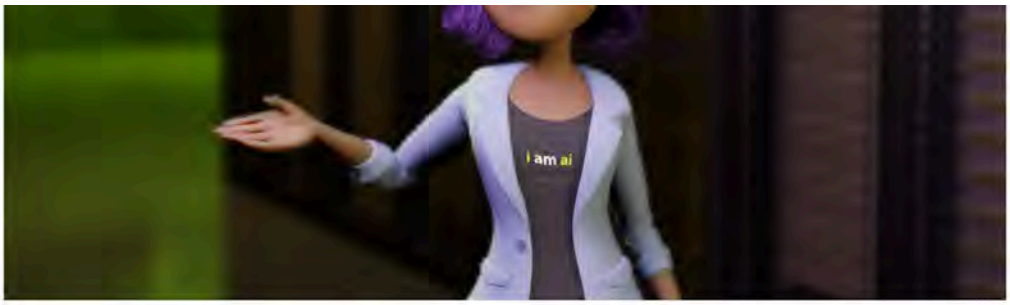
NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

NVIDIA Omniverse ACE Enables Easier, Faster Deployment of Interactive Avatars

GTC demo showcases AI-powered avatar responding to natural speech and making intelligent recommendations.

September 20, 2022 by Stephanie Rubenstein





Share

Reading Time: 3 mins

Meet Violet, an AI-powered customer service assistant ready to take your order.

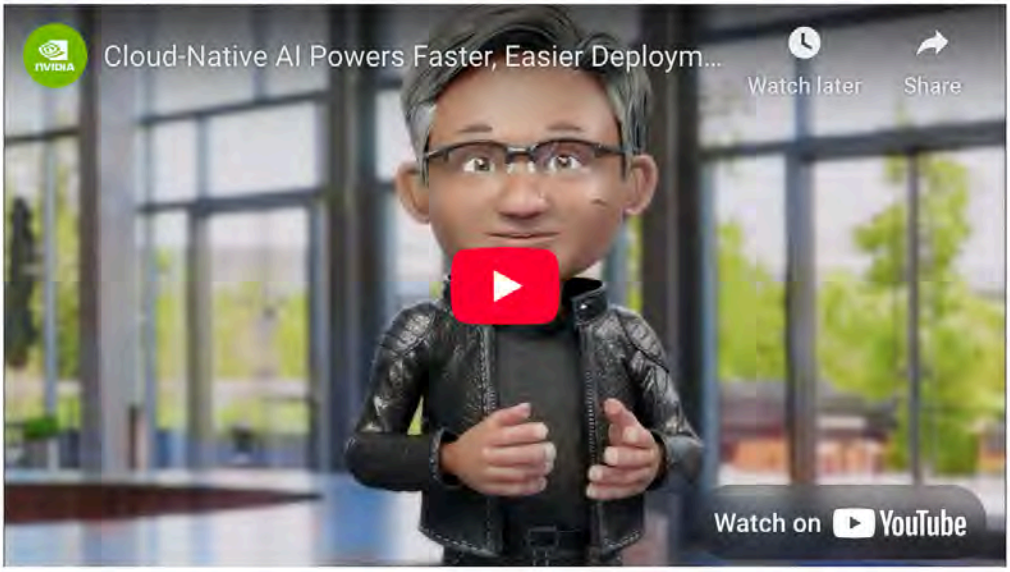
Unveiled this week at GTC, Violet is a cloud-based avatar that represents the latest evolution in avatar development through NVIDIA Omniverse Avatar Cloud Engine (ACE), a suite of cloud-native AI microservices that make it easier to build and deploy intelligent virtual assistants and digital humans at scale.

To animate interactive avatars like Violet, developers need to ensure the 3D character can see, hear, understand and communicate with people. But bringing these avatars to life can be incredibly challenging, as traditional methods typically require expensive equipment, specific expertise and time-consuming workflows.

The Violet demo showcases how Omniverse ACE eases avatar development, delivering all the AI building blocks necessary to create, customize and deploy interactive avatars. Whether taking restaurant orders or answering questions about the universe, these AI assistants are easily customizable for virtually any industry, and can help organizations enhance existing workflows and unlock new business opportunities.

Watch the video below to see Violet interact with users, respond to speech prompts

and make intelligent recommendations:



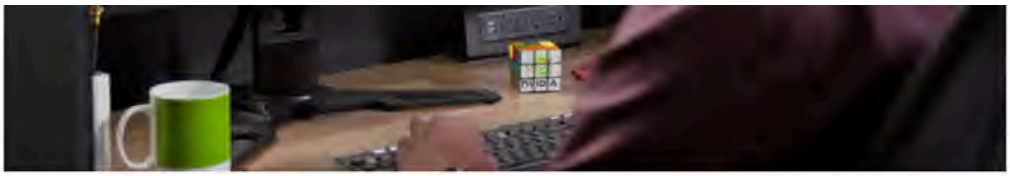
How Omniverse ACE Brings Violet to Life

The demo showcases Violet as a fully rigged avatar with basic animation. To create Violet, NVIDIA's creative team used the company's Unified Compute Framework, a fully accelerated framework that enables developers to combine optimized and accelerated microservices into real-time AI applications. UCF helped the team build a graph of microservices for Violet that were deployed in the cloud.

Omniverse ACE powers the backend of interactive avatars, essentially acting as Violet's brain. Additionally, two reference applications are built on ACE: NVIDIA Tokkio and NVIDIA Maxine.

Violet was developed using the Tokkio application workflow, which enables interactive avatars to see, perceive, converse intelligently and provide recommendations to enhance customer service, both online and in places like restaurants and stores.

NVIDIA Maxine delivers a suite of GPU-accelerated AI software development kits and cloud-native microservices for deploying AI features to enhance real-time video communications. Maxine integrates the NVIDIA Riva SDK's real-time automatic speech recognition and text-to-speech capabilities with real-time "live portrait" photo animation and eye contact features, which enable better communication and understanding.



With UCF, NVIDIA's creative team built a graph of microservices for Violet that were deployed in the cloud.

Latest Microservices Expand Possibilities for Avatars

The demo with Violet highlights how developers of digital humans and virtual assistants can use Omniverse ACE to accelerate their avatar development workflows. Omniverse ACE also delivers microservices that enable developers to access the best NVIDIA AI technology, with no coding required.

Some of the latest microservices include:

- **Animation AI:** Omniverse Audio2Face simplifies animation of a 3D character to match any voice-over track, helping users animate characters for games, films or real-time digital assistants.
- **Conversational AI:** Includes the NVIDIA Riva SDK for speech AI and NVIDIA NeMo Megatron framework for natural language processing, allowing developers to quickly build and deploy cutting-edge applications that deliver high-accuracy, expressive voices and respond in real time.

AI Avatars Deliver New Transformations Across Industries

The AI avatars that ACE enables will enhance interactive experiences in industries such as gaming, entertainment, transportation and hospitality.

Leading professional-services company Deloitte has worked with NVIDIA to help enterprises deploy transformative applications. At GTC, Deloitte announced that new hybrid-cloud offerings for NVIDIA AI and NVIDIA Omniverse services and platforms, including Omniverse ACE, will be added to the existing Deloitte Center for AI Computing.

NVIDIA Omniverse ACE brings all three versions of Violet to life, from customer service Violet (left), to intelligent Violet (right) to Ultra Violet (center).

"Cloud-based AI models and services are opening up new ways for digital humans to make people feel more connected, and today's interaction with Violet in the NVIDIA GTC keynote shows a glimpse into the future of AI-powered avatars," said Vladimir Mastilović, vice president of digital humans technology at Epic Games. "We are delighted to see NVIDIA Omniverse ACE using MetaHumans in Unreal Engine 5 to make it even easier to deploy engaging high-fidelity 3D avatars."

NVIDIA Omniverse ACE will be available to early-access partners starting later this year, along with the Tokkio reference application for simplified customer-service avatar implementation.

Learn more about Omniverse ACE by joining this session at GTC, and explore all the technologies that go into the creation and animation of realistic, interactive digital humans.

Customers can request the hands-on, web-based Tokkio demo.

Developers and partners can sign up to be notified when ACE is available.

And catch up on the latest announcements from the GTC keynote by NVIDIA founder and CEO Jensen Huang:

Categories: Pro Graphics

Tags: Artificial Intelligence | GTC 2022 | NVIDIA NeMo | Omniverse

Load Comments

Explore What's Next in AI

See the top GTC sessions recommended just for you.

Watch on Demand

All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

New NVIDIA Maxine Cloud-Native Architecture Delivers Breakthrough Audio and Video Quality at Scale

Early-access AI microservices deliver premium-quality communications in the cloud.

September 20, 2022 by Delilah Liu

Share

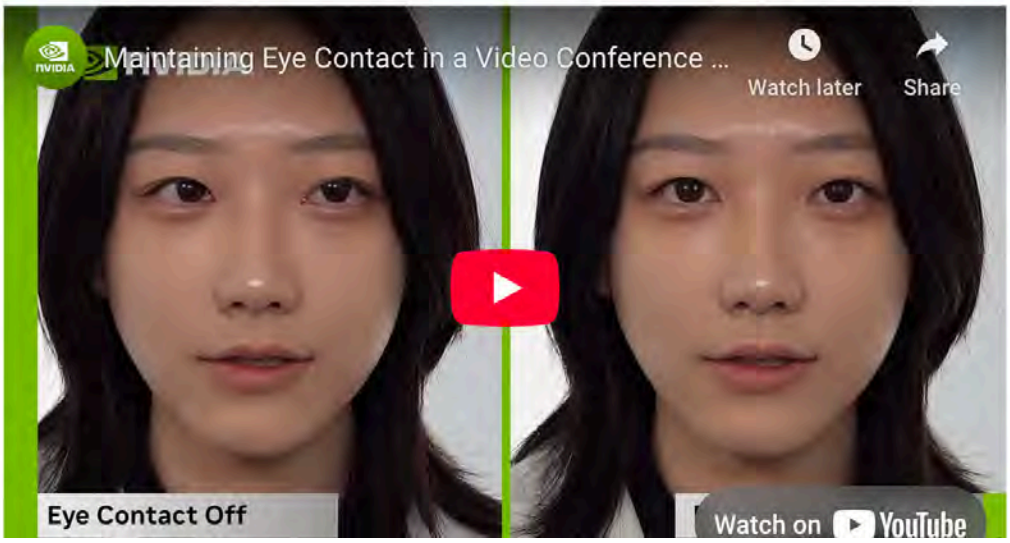
Reading Time: 3 mins

The latest release of NVIDIA Maxine is paving the way for real-time audio and video communications. Whether for a video conference, a call made to a customer service center, or a live stream, Maxine enables clear communications to enhance virtual interactions.

NVIDIA Maxine is a suite of GPU-accelerated AI software development kits (SDKs) and cloud-native microservices for deploying optimized and accelerated AI features that enhance audio, video and augmented-reality (AR) effects in real time.

And with Maxine's state-of-the-art models, end users don't need expensive gear to improve audio and video. Using NVIDIA AI-based technology, these high-quality effects can be achieved with standard microphones and camera equipment.

At GTC, NVIDIA announced the re-architecture of Maxine for cloud-native microservices, with the early-access release of Maxine's audio-effects microservice. Additionally, new Maxine SDK features were unveiled, including Speaker Focus and Face Expression Estimation, as well as the general availability of Eye Contact. NVIDIA Maxine now also includes enhanced versions of existing SDK features.



Maxine Goes Cloud Native

Maxine's cloud-native microservices allow developers to build real-time AI applications. Microservices can be independently managed and deployed seamlessly in the cloud, accelerating development timelines.

The Audio Effects microservice, available in early access, contains four state-of-the-art audio features:

- **Background Noise Removal:** Removes several common background noises using AI models, while preserving the speaker's natural voice.
- **Room Echo Removal:** Removes reverberations from audio using AI models, restoring clarity of a speaker's voice.
- **Audio Super Resolution:** Improves audio quality by increasing the temporal resolution of audio signal. It currently supports upsampling from 8 kHz to 16 kHz and from 16 kHz to 48 kHz.
- **Acoustic Echo Cancellation:** Cancels real-time acoustic device echo from the input-audio stream, eliminating mismatched acoustic pairs and double-talk. With AI-based technology, more effective cancellation is achieved than with traditional digital signal processing.

Pexip, a leading provider of enterprise video conferencing and collaboration solutions, is using NVIDIA AI technologies to take virtual meetings to the next level with advanced features for the modern workforce.

"With Maxine's move to cloud-native microservices, it will be even easier to combine NVIDIA's advanced AI technologies with our own unique server-side architecture," said Eddie Clifton, senior vice president of Strategic Alliances at Pexip. "This allows our teams at Pexip to deliver an enhanced experience for virtual meetings."

Sign up for early access.

Explore Enhanced Features of SDKs

Maxine offers three GPU-accelerated SDKs that reinvent real-time communications with AI: audio, video and AR effects.

The audio effects SDK delivers multi-effect, low-latency, AI-based audio-quality enhancement algorithms. **Speaker Focus**, available in early access, is a new feature that separates the audio tracks of foreground and background speakers, making each voice more intelligible. Additionally, the **Audio Super Resolution** SDK feature has been updated with enhanced quality.

The video effects SDK creates AI-based video effects with standard webcam input. The **Virtual Background** feature, which segments a person's profile and applies AI-powered background removal, replacement or blur, has been updated with enhanced temporal stability.

And the AR SDK provides AI-powered, real-time 3D face tracking and body pose estimation based on a standard web camera feed. Latest features include:

- **Eye Contact:** Simulates eye contact by estimating and aligning gaze with the

camera.

- **Face Expression Estimation**: Tracks the face and infers what expression is presented by the subject.

The following AR features have been updated:

- **Body Pose Estimation**: Predicts and tracks 34 key points of the human body in 2D and 3D — now with support for multi-person tracking.
- **Face Landmark Tracking**: Recognizes facial features and contours using 126 key points. Tracks head pose and facial deformation due to head movement and expression — in three degrees of freedom in real time — now with Quality mode to achieve even higher-quality tracking.
- **Face Mesh**: Represents a human face with a 3D mesh with up to 3,000 vertices and six degrees of freedom — now includes 3D morphable models from the USC Institute of Creative Technologies.

Try out the Maxine SDKs. To directly experience Maxine's effects, download the NVIDIA Broadcast App.

Experience State-of-the-Art Effects With the Power of AI

Maxine SDKs and microservices provide a suite of low-latency AI effects that can be integrated with existing customer infrastructures. Developers can tap into cutting-edge AI capabilities with Maxine, as the technology is built on the NVIDIA AI platform and has world-class pretrained models for users to create, customize and deploy premium audio- and video-quality features.

Maxine is also part of the NVIDIA Omniverse Avatar Cloud Engine, a collection of cloud-based AI models and services for developers to build, customize and deploy interactive avatars. Maxine's customizable cloud-native microservices allow for independent deployment into AI-effects pipelines. Maxine can be deployed on premises, in the cloud or at the edge.

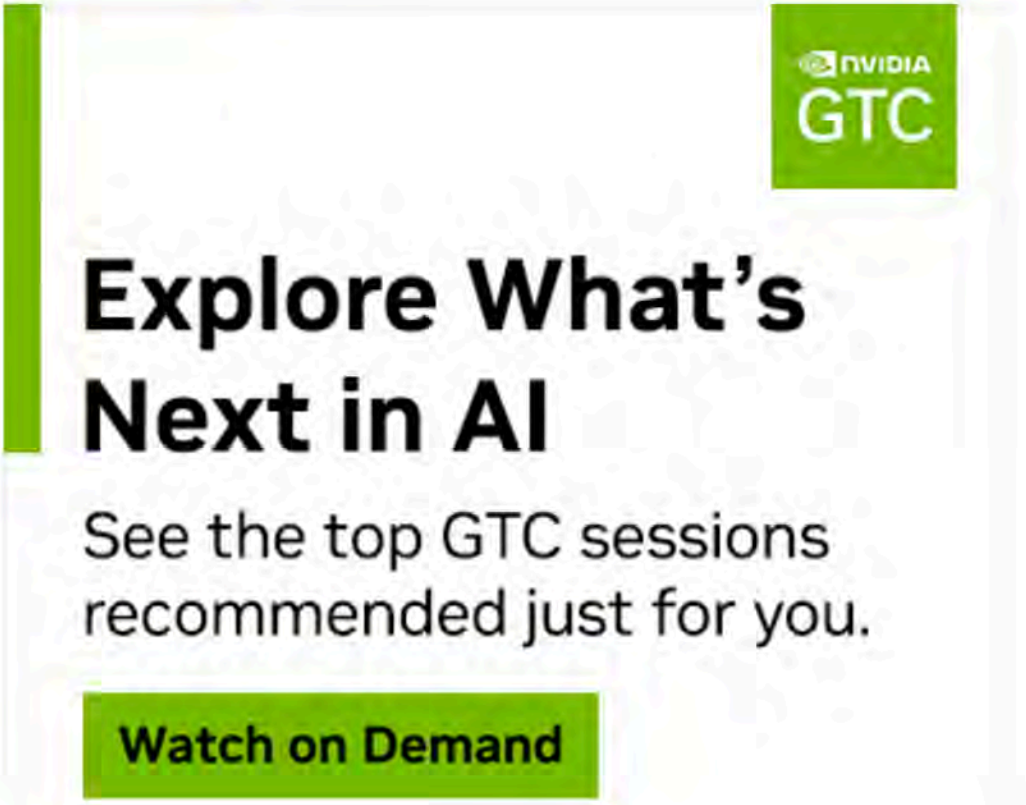
Learn more about NVIDIA Maxine and other technology breakthroughs by watching the GTC keynote by NVIDIA founder and CEO Jensen Huang:

Categories: Pro Graphics

Categories: Pro Graphics

Tags: Artificial Intelligence | GTC 2022 | NVIDIA Maxine | Video Streaming

Load Comments



All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

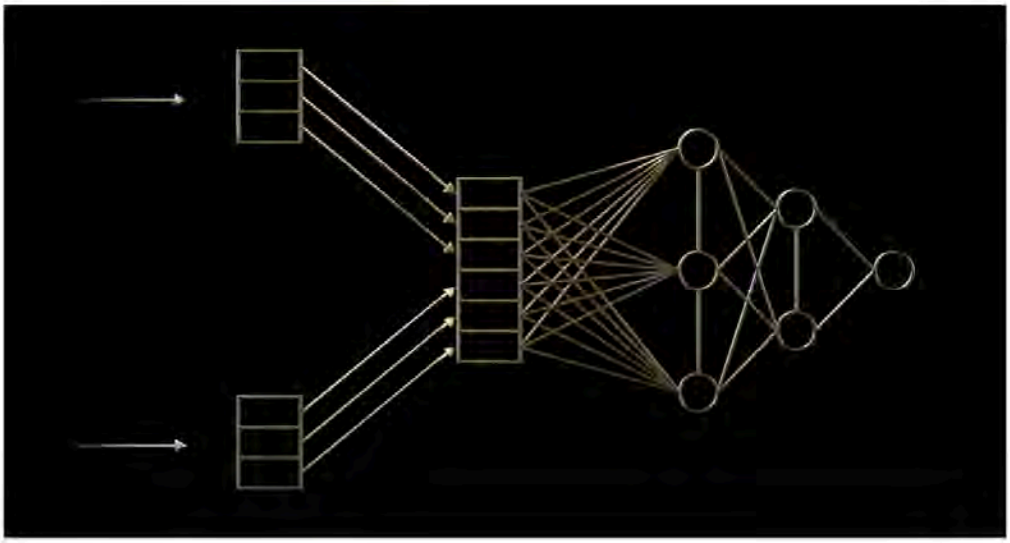
Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

Why the New NVIDIA Grace Hopper Superchip Is Ideal for Next-Gen Recommender Systems

Performance of the massive AI models that help users personalize the internet will hit new levels of accuracy with the Grace Hopper Superchip.

September 20, 2022 by Paresh Kharya



Share

Reading Time: 3 mins

Recommender systems, the economic engines of the internet, are getting a new turbocharger: the NVIDIA Grace Hopper Superchip.

Every day, recommenders serve up trillions of search results, ads, products, music and news stories to billions of people. They're among the most important AI models of our time because they're incredibly effective at finding in the internet's pandemonium the pearls users want.

These machine learning pipelines run on data, terabytes of it. The more data recommenders consume, the more accurate their results and the more return on investment they deliver.

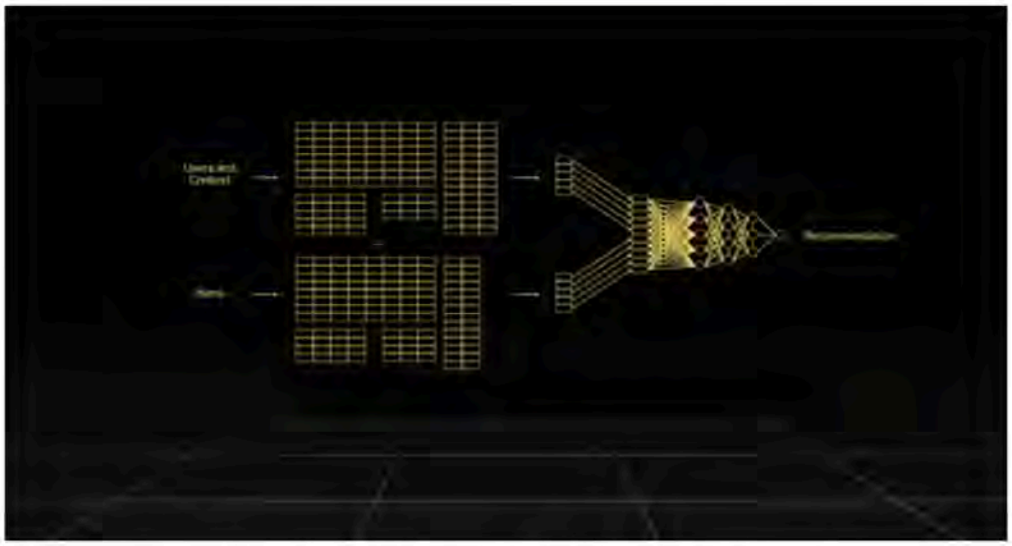
To process this data tsunami, companies are already adopting accelerated computing to personalize services for their customers. Grace Hopper will take their advances to the next level.

GPUs Drive 16% More Engagement

Pinterest, the image-sharing social media company, was able to move to 100x larger recommender models by adopting NVIDIA GPUs. That increased engagement by

16% for its more than 400 million users.

"Normally, we would be happy with a 2% increase, and 16% is just a beginning," a software engineer at the company said in a recent blog. "We see additional gains — it opens a lot of doors for opportunities."



Recommenders consume tens of terabytes of embeddings, data tables that provide context for making accurate predictions.

The next generation of the NVIDIA AI platform promises even greater gains for companies processing massive datasets with super-sized recommender models.

Because data is the fuel of AI, Grace Hopper is designed to pump more data through recommender systems than any other processor on the planet.

NVLink Accelerates Grace Hopper

Grace Hopper achieves this because it's a superchip — two chips in one unit, sharing a superfast chip-to-chip interconnect. It's an Arm-based NVIDIA Grace CPU and a Hopper GPU that communicate over NVIDIA NVLink-C2C.

What's more, NVLink also connects many superchips into a super system, a computing cluster built to run terabyte-class recommender systems.

NVLink carries data at a whopping 900 gigabytes per second — 7x the bandwidth of PCIe Gen 5, the interconnect most leading edge upcoming systems will use.

That means Grace Hopper feeds recommenders 7x more of the embeddings — data tables packed with context — that they need to personalize results for users.

More Memory, Greater Efficiency

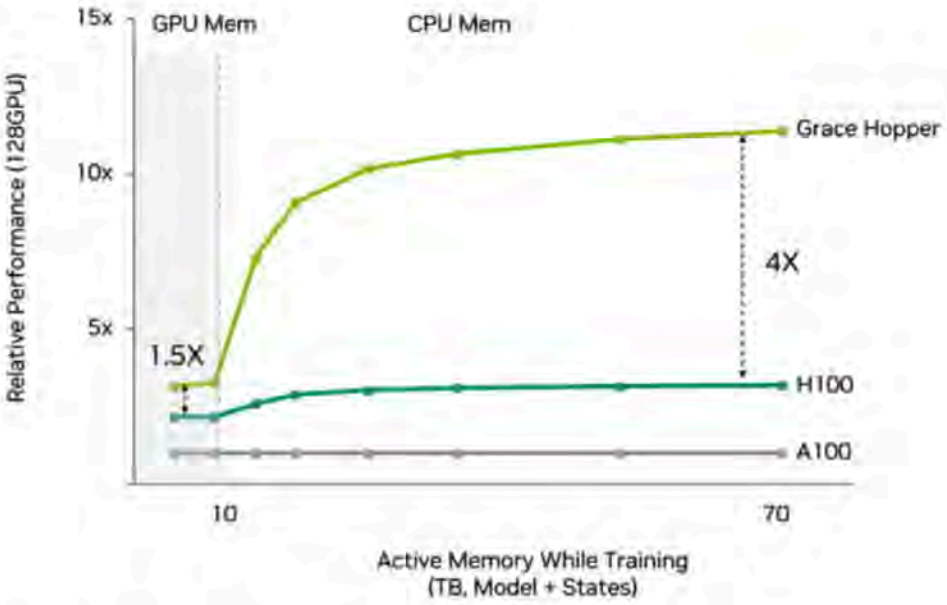
The Grace CPU uses LPDDR5X, a type of memory that strikes the optimal balance of bandwidth, energy efficiency, capacity and cost for recommender systems and other demanding workloads. It provides 50% more bandwidth while using an eighth of the power per gigabyte of traditional DDR5 memory subsystems.

Any Hopper GPU in a cluster can access Grace's memory over NVLink. It's a feature other demanding workloads. It provides 50% more bandwidth while using an eighth of the power per gigabyte of traditional DDR5 memory subsystems.

Any Hopper GPU in a cluster can access Grace's memory over NVLink. It's a feature of Grace Hopper that provides the largest pools of GPU memory ever.

In addition, NVLink-C2C requires just 1.3 picojoules per bit transferred, giving it more than 5x the energy efficiency of PCIe Gen 5.

The overall result is recommenders get a further up to 4x more performance and greater efficiency using Grace Hopper than using Hopper with traditional CPUs (see chart below).



All the Software You Need

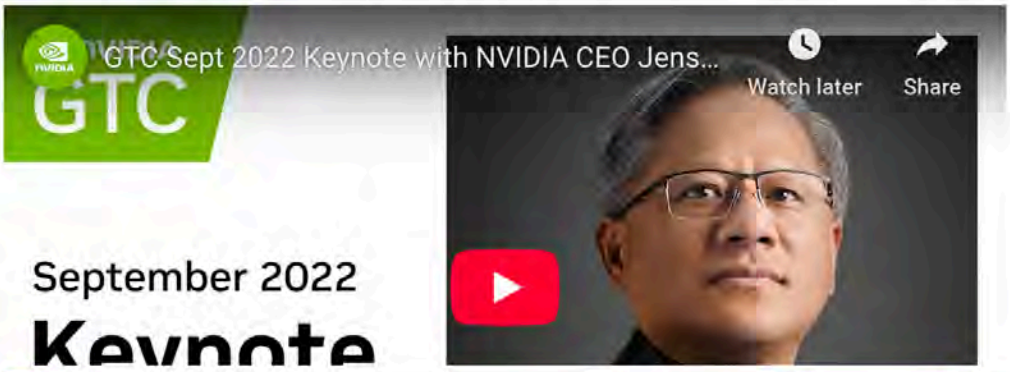
The Grace Hopper Superchip runs the full stack of NVIDIA AI software used in some of the world's largest recommender systems today.

NVIDIA Merlin is the rocket fuel of recommenders, a collection of models, methods and libraries for building AI systems that can provide better predictions and increase clicks.

NVIDIA Merlin HugeCTR, a recommender framework, helps users process massive datasets fast across distributed GPU clusters with help from the NVIDIA Collective Communications Library.

Learn more about Grace Hopper and NVLink in this technical blog. Watch this GTC session to learn more about building recommender systems.

You can also hear NVIDIA CEO and co-founder Jensen Huang provide perspective on recommenders here or watch the full GTC keynote below.



NVIDIA GTC

Explore What's Next in AI

See the top GTC sessions recommended just for you.

Watch on Demand

All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

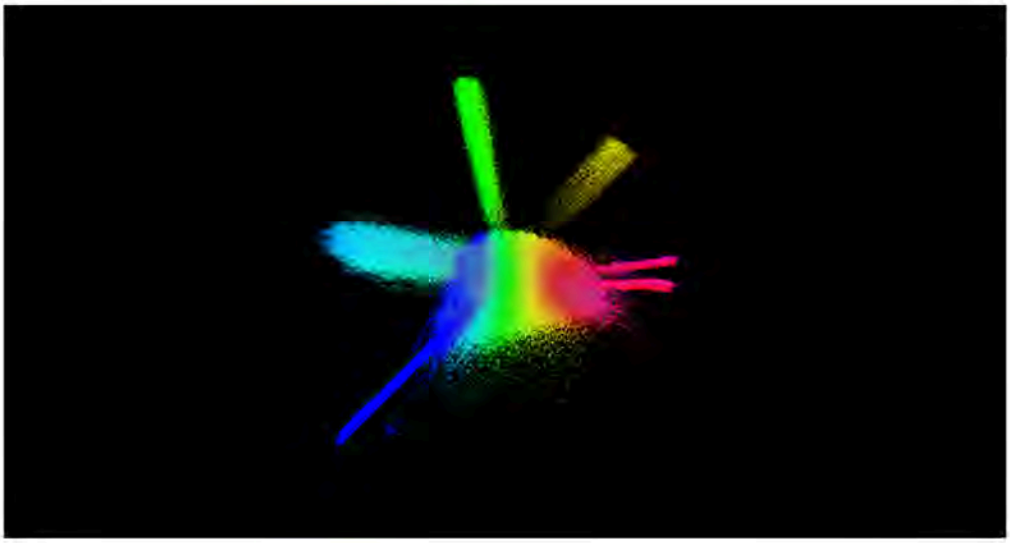
Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

NVIDIA Expands Large Language Models to Biology

Leading pharma companies, biotech startups and pioneering biology researchers are developing AI applications with the NVIDIA BioNeMo LLM service and framework to generate, predict and understand biomolecular data.

September 20, 2022 by Abraham Stern



Share

Reading Time: 4 mins

As scientists probe for new insights about DNA, proteins and other building blocks of life, the NVIDIA BioNeMo framework — announced today at NVIDIA GTC — will accelerate their research.

NVIDIA BioNeMo is a framework for training and deploying large biomolecular language models at supercomputing scale — helping scientists better understand disease and find therapies for patients. The large language model (LLM) framework will support chemistry, protein, DNA and RNA data formats.

It's part of the NVIDIA Clara Discovery collection of frameworks, applications and AI models for drug discovery.

Just as AI is learning to understand human languages with LLMs, it's also learning the languages of biology and chemistry. By making it easier to train massive neural networks on biomolecular data, NVIDIA BioNeMo helps researchers discover new patterns and insights in biological sequences — insights that researchers can connect to biological properties or functions, and even human health conditions.

NVIDIA BioNeMo provides a framework for scientists to train large-scale language models using bigger datasets, resulting in better-performing neural networks. The framework will be available in early access on NVIDIA NGC, a hub for GPU-optimized software.

In addition to the language model framework, NVIDIA BioNeMo has a cloud API service that will support a growing list of pretrained AI models.

BioNeMo Framework Supports Bigger Models, Better Predictions

Scientists using natural language processing models for biological data today often train relatively small neural networks that require custom preprocessing. By adopting BioNeMo, they can scale up to LLMs with billions of parameters that capture information about molecular structure, protein solubility and more.

BioNeMo is an extension of the NVIDIA NeMo Megatron framework for GPU-accelerated training of large-scale, self-supervised language models. It's domain specific, designed to support molecular data represented in the SMILES notation for chemical structures, and in FASTA sequence strings for amino acids and nucleic acids.

"The framework allows researchers across the healthcare and life sciences industry to take advantage of their rapidly growing biological and chemical datasets," said Mohammed AlQuraishi, founding member of the OpenFold Consortium and assistant professor at Columbia University's Department of Systems Biology. "This makes it easier to discover and design therapeutics that precisely target the molecular signature of a disease."

BioNeMo Service Features LLMs for Chemistry and Biology

For developers looking to quickly get started with LLMs for digital biology and chemistry applications, the NVIDIA BioNeMo LLM service will include four pretrained language models. These are optimized for inference and will be available under early access through a cloud API running on NVIDIA DGX Foundry.

- **ESM-1**: This protein LLM, based on the state-of-the-art ESM-1b model published by Meta AI, processes amino acid sequences to generate representations that can be used to predict a wide variety of protein properties and functions. It also improves scientists' ability to understand protein structure.
- **OpenFold**: The public-private consortium creating state-of-the-art protein modeling tools will make its open-source AI pipeline accessible through the BioNeMo service.
- **MegaMolBART**: Trained on 1.4 billion molecules, this generative chemistry model can be used for reaction prediction, molecular optimization and de novo

molecular generation.

- **ProtT5**: The model, developed in a collaboration led by the Technical University of Munich's RostLab and including NVIDIA, extends the capabilities of protein LLMs like Meta AI's ESM-1b to sequence generation.

In the future, researchers using the BioNeMo LLM service will be able to customize the LLM models for higher accuracy on their applications in a few hours — with fine-tuning and new techniques such as p-tuning, a training method that requires a dataset with just a few hundred examples instead of millions.

Startups, Researchers and Pharma Adopting NVIDIA BioNeMo

A wave of experts in biotech and pharma are adopting NVIDIA BioNeMo to support drug discovery research.

- **AstraZeneca** and NVIDIA have used the Cambridge-1 supercomputer to develop the MegaMolBART model included in the BioNeMo LLM service. The global biopharmaceuticals company will use the BioNeMo framework to help train some of the world's largest language models on datasets of small molecules, proteins and, soon, DNA.
- Researchers at the **Broad Institute of MIT and Harvard** are working with NVIDIA to develop next-generation DNA language models using the BioNeMo framework. These models will be integrated into Terra, a cloud platform co-developed by the Broad Institute, Microsoft and Verily that enables biomedical researchers to share, access and analyze data securely and at scale. The AI models will also be added to the BioNeMo service's collection.
- The **OpenFold** consortium plans to use the BioNeMo framework to advance its work developing AI models that can predict molecular structures from amino acid sequences with near-experimental accuracy.
- **Peptone** is focused on modeling intrinsically disordered proteins — proteins that lack a stable 3D structure. The company is working with NVIDIA to develop versions of the ESM model using the NeMo framework, which BioNeMo is also based on. The project, which is scheduled to run on NVIDIA's Cambridge-1 supercomputer, will advance Peptone's drug discovery work.
- **Evozyne**, a Chicago-based biotechnology company, combines engineering and deep learning technology to design novel proteins to solve long-standing challenges in therapeutics and sustainability.

"The BioNeMo framework is an enabling technology to efficiently leverage the power of LLMs for data-driven protein design within our design-build-test cycle," said Andrew Ferguson, co-founder and head of computation at Evozyne. "This will have an immediate impact on our design of novel functional proteins, with applications in human health and sustainability."

"As we see the ever-widening adoption of large language models in the protein space, being able to efficiently train LLMs and quickly modulate model architectures is becoming hugely important," said Istvan Redl, machine learning lead at Peptone, a biotech startup in the NVIDIA Inception program. "We believe that these two engineering aspects — scalability and rapid experimentation — are exactly what the BioNeMo framework could provide."

Sign up for early access to the NVIDIA BioNeMo LLM service or BioNeMo

framework. For hands on-experience with the MegaMolBART chemistry model in BioNeMo, request a free lab from NVIDIA LaunchPad on training and deploying LLMs.

Discover the latest in AI and healthcare at GTC, running online through Thursday, Sept. 22. Registration is free.

Watch the GTC keynote address by NVIDIA founder and CEO Jensen Huang below:

Main image by Mahendra awale, licensed under CC BY-SA 3.0 via Wikimedia Commons

Categories: Deep Learning

Tags: Artificial Intelligence | Genomics | GTC 2022 | Healthcare and Life Sciences | NVIDIA Clara | NVIDIA NeMo | Science | Social Impact

Load Comments

NVIDIA GTC

Explore What's Next in AI

See the top GTC sessions recommended just for you.

Watch on Demand

All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

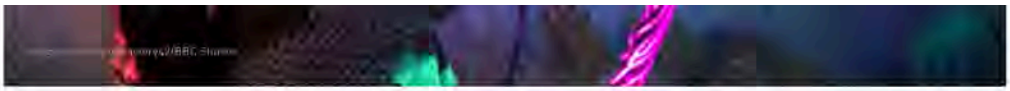
Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

NVIDIA Introduces Open-Source Project to Accelerate Computer Vision Cloud Applications

CV-CUDA combines accelerated image pre- and post-processing algorithms and tools to process 10x the number of images at the same cost.

September 20, 2022 by Michael Boone



Share

Reading Time: 2 mins

Promising to help process images faster and more efficiently at a vast scale, NVIDIA introduced CV-CUDA, an open-source library for building accelerated end-to-end computer vision and image processing pipelines.

The majority of internet traffic is video. Increasingly, this video will be augmented by AI special effects and computer graphics.

To add to this complexity, fast-growing social media and video-sharing services are experiencing growing cloud computing costs and bottlenecks in their AI-based imaging processing and computer vision pipelines.

CV-CUDA accelerates AI special effects such as relighting, reposing, blurring backgrounds and super resolution.

NVIDIA GPUs already accelerate the inference portion of AI computer vision pipelines. But pre- and post-processing using traditional computer vision tools gobble up time and computing power.

CV-CUDA gives developers more than 50 high-performance computer vision algorithms, a development framework that makes it easy to implement custom kernels and zero-copy interfaces to remove bottlenecks in the AI pipeline.

gobble up time and computing power.

CV-CUDA gives developers more than 50 high-performance computer vision algorithms, a development framework that makes it easy to implement custom kernels and zero-copy interfaces to remove bottlenecks in the AI pipeline.

The result is higher throughput and lower cloud-computing costs. CV-CUDA can process 10x as many streams on a single GPU.

All this helps developers move much faster when tackling video content creation, 3D worlds, image-based recommender systems, image recognition and video conferencing.

Video content creation platforms must process, enhance and moderate millions of video streams daily and ensure mobile-based users have the best experience running their apps on any phone.

- For those building 3D worlds or metaverse applications, CV-CUDA is anticipated to enable tasks to help build or extend 3D worlds and their components.
- In image understanding and recognition, CV-CUDA can significantly speed up the pipelines running at hyperscale, allowing mobile users to enjoy sophisticated and responsive image recognition applications.
- And in video conferencing, CV-CUDA can support sophisticated augmented reality-based features. These features could involve complex AI pipelines requiring numerous pre- and post-processing steps.

CV-CUDA accelerates pre- and post-processing pipelines through hand-optimized CUDA kernels and natively integrates into C/C++, Python and common deep learning frameworks, such as PyTorch.

Featured image credit: Factory42/BBC Studios

Categories: Corporate | Data Center

Tags: GTC | GTC 2022 | Open Source

Load Comments

NVIDIA GTC

Explore What's Next in AI

See the top GTC sessions

recommended just for you.

Watch on Demand

All NVIDIA News

The More You Buy, the More You Make

The Supercomputer Designed to Accelerate Nobel-Worthy Science

RTX on Deck: The GeForce NOW Native App for Steam Deck Is Here

Run LLMs on AnythingLLM Faster With NVIDIA RTX AI PCs

NVIDIA's Bartley Richardson on How Teams of AI Agents Provide Next-Level Automation

Growing Range of Researchers, Scientists Adopt NVIDIA

Scientists Adopt NVIDIA cuQuantum and CUDA-Q

The NVIDIA platform is increasingly used to help design quantum algorithms, simulations and classical-quantum computers.

September 20, 2022 by Jin-Sung Kim

cuQuantum and CUDA-Q

The NVIDIA platform is increasingly used to help design quantum algorithms, simulations and classical-quantum computers.

September 20, 2022 by Jin-Sung Kim

Share

Reading Time: 4 mins

Quantum's Software Foundation

Laying the software foundation for quantum computers is the focus of the QIR Alliance that Heim chairs.

The industry group oversees work on a common interface between quantum programming languages and targeted quantum computers. QIR enables developers to use tools like CUDA-Q to build apps that run on hybrid quantum computers to