40 Years on, PAC-MAN Recreated with AI by NVIDIA Researchers

GameGAN, a generative adversarial network trained on 50,000 PAC-MAN episodes, produces a fully functional version of the dot-munching classic without an underlying game engine.

May 22, 2020 by ISHA SALIAN

Share

Forty years to the day since PAC-MAN first hit arcades in Japan, and went on to munch a path to global stardom, the retro classic has been reborn, delivered courtesy of AI.

Trained on 50,000 episodes of the game, a powerful new AI model created by NVIDIA Research, called NVIDIA GameGAN, can generate a fully functional version of PAC-MAN — without an underlying game engine. That means that even without understanding a game's fundamental rules, AI can recreate the game with convincing results.

GameGAN is the first neural network model that mimics a computer game engine by harnessing generative adversarial networks, or GANs. Made up of two competing

neural networks, a generator and a discriminator, GAN-based models learn to create new content that's convincing enough to pass for the original.

"This is the first research to emulate a game engine using GAN-based neural networks," said Seung-Wook Kim, an NVIDIA researcher and lead author on the project. "We wanted to see whether the AI could learn the rules of an environment just by looking at the screenplay of an agent moving through the game. And it did."

As an artificial agent plays the GAN-generated game, GameGAN responds to the agent's actions, generating new frames of the game environment in real time. GameGAN can even generate game layouts it's never seen before, if trained on screenplays from games with multiple levels or versions.

This capability could be used by game developers to automatically generate layouts for new game levels, as well as by AI researchers to more easily develop simulator systems for training autonomous machines.

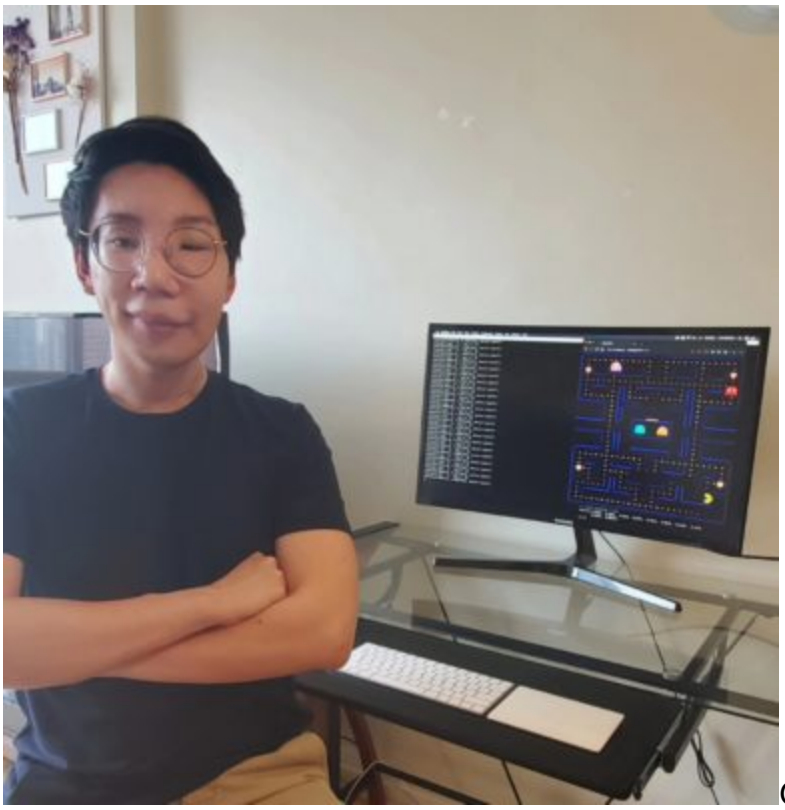
"We were blown away when we saw the results, in disbelief that AI could recreate the iconic PAC-MAN experience without a game engine," said Koichiro Tsutsumi from BANDAI NAMCO Research Inc., the research development company of the game's publisher BANDAI NAMCO Entertainment Inc., which provided the PAC-MAN data to train GameGAN. "This research presents exciting possibilities to help game developers accelerate the creative process of developing new level layouts, characters and even games."

We'll be making our AI tribute to the game available later this year on AI Playground, where anyone can experience our research demos firsthand.

AI Goes Old School

PAC-MAN enthusiasts once had to take their coins to the nearest arcade to play the classic maze chase. Take a left at the pinball machine and continue straight past the air hockey, following the unmistakable soundtrack of PAC-MAN gobbling dots and avoiding ghosts Inky, Pinky, Blinky and Clyde.

In 1981 alone, Americans inserted billions of quarters to play 75,000 hours of coin-operated games like PAC-MAN. Over the decades since, the hit game has seen versions for PCs, gaming consoles and cell phones.

 Game Changer: NVIDIA Researcher Seung-Wook Kim and his collaborators trained GameGAN on 50,000 episodes of PAC-MAN.
The GameGAN edition relies on neural networks, instead of a traditional game engine, to generate PAC-MAN's environment. The AI keeps track of the virtual world, remembering what's already been generated to maintain visual consistency from frame to frame.

No matter the game, the GAN can learn its rules simply by ingesting screen recordings and agent keystrokes from past gameplay. Game developers could use such a tool to automatically design new level layouts for existing games, using screenplay from the original levels as training data.

With data from BANDAI NAMCO Research, Kim and his collaborators at the NVIDIA AI Research Lab in Toronto used NVIDIA DGX systems to train the neural networks on the PAC-MAN episodes (a few million frames, in total) paired with data on the keystrokes of an AI agent playing the game.

The trained GameGAN model then generates static elements of the environment, like a consistent maze shape, dots and Power Pellets — plus moving elements like the enemy ghosts and PAC-MAN itself.

It learns key rules of the game, both simple and complex. Just like in the original game, PAC-MAN can't walk through the maze walls. He eats up dots as he moves around, and when he consumes a Power Pellet, the ghosts turn blue and flee. When PAC-MAN exits the maze from one side, he's teleported to the opposite end. If he runs into a ghost, the screen flashes and the game ends.

Since the model can disentangle the background from the moving characters, it's possible to recast the game to take place in an outdoor hedge maze, or swap out PAC-MAN for your favorite emoji. Developers could use this capability to experiment with new character ideas or game themes.

It's Not Just About Games

Autonomous robots are typically trained in a simulator, where the AI can learn the rules of an environment before interacting with objects in the real world. Creating a simulator is a time-consuming process for developers, who must code rules about how objects interact with one another and how light works within the environment.

Simulators are used to develop autonomous machines of all kinds, such as warehouse robots learning how to grasp and move objects around, or delivery robots that must navigate sidewalks to transport food or medicine.

GameGAN introduces the possibility that the work of writing a simulator for tasks like these could one day be replaced by simply training a neural network.

Suppose you install a camera on a car. It can record what the road environment looks like or what the driver is doing, like turning the steering wheel or hitting the accelerator. This data could be used to train a deep learning model that can predict what would happen in the real world if a human driver — or an autonomous car — took an action like slamming the brakes.

"We could eventually have an AI that can learn to mimic the rules of driving, the laws of physics, just by watching videos and seeing agents take actions in an environment," said Sanja Fidler, director of NVIDIA's Toronto research lab. "GameGAN is the first step toward that."

NVIDIA Research has more than 200 scientists around the globe, focused on areas such as AI, computer vision, self-driving cars, robotics and graphics.

GameGAN is authored by Fidler, Kim, NVIDIA researcher Jonah Philion, University of Toronto student Yuhao Zhou and MIT professor Antonio Torralba. The paper will be

presented at the prestigious Conference on Computer Vision and Pattern Recognition in June.

PAC-MAN™ & ©BANDAI NAMCO Entertainment Inc.

Categories: Deep Learning | Research

Tags: Artificial Intelligence | NVIDIA Research

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

What's a DPU?

...And what's the difference between a DPU, a CPU, and a GPU?

May 20, 2020 by KEVIN DEIERLING

Share

What's a DPU?

Of course, you're probably already familiar with the Central Processing Unit or CPU. Flexible and responsive, for many years CPUs were the sole programmable element in most computers.

More recently the GPU, or graphics processing unit, has taken a central role. Originally used to deliver rich, real-time graphics, their parallel processing capabilities make them ideal for accelerated computing tasks of all kinds.

That's made them the key to artificial intelligence, deep learning, and big data analytics applications.

Over the past decade, however, computing has broken out of the boxy confines of PC and servers — with CPUs and GPUs powering sprawling new hyperscale data centers.

These data centers are knit together with a powerful new category of processors. The DPU, or data processing unit, has become the third member of the data centric accelerated computing model."This is going to represent one of the three major pillars of computing going forward," NVIDIA CEO Jensen Huang said during a talk earlier this month.

"The CPU is for general purpose computing, the GPU is for accelerated computing and the DPU, which moves data around the data center, does data processing."

What's a DPU?

Data Processing Unit
Industry-standard, high-performance, software-programmable multi-core CPU
High-performance network interface
Flexible and programmable acceleration engines

So What Makes a DPU Different?

A DPU is a new class of programmable processor that combines three key elements. A DPU is a system on a chip, or SOC, that combines:
An industry standard, high-performance, software programmable, multi-core CPU, typically based on the widely-used Arm architecture, tightly coupled to the other SOC components

A high-performance network interface capable of parsing, processing, and efficiently transferring data at line rate, or the speed of the rest of the network, to GPUs and CPUs

A rich set of flexible and programmable acceleration engines that offload and improve applications performance for AI and Machine Learning, security, telecommunications, and storage, among others.

All these DPU capabilities are critical to enable an isolated, bare-metal, cloud-native computing that will define the next generation of cloud-scale computing.

DPUs: Incorporated into SmartNICs

The DPU can be used as a stand-alone embedded processor, but it's more often incorporated into a SmartNIC, a network interface controller that's used as a key component in a next generation server.

Other devices that claim to be DPUs miss significant elements of these three critical capabilities that are fundamental to claiming to answer the question: What is a DPU?

DPUs can be used as a stand-alone embedded processor, but they're more often incorporated into a SmartNIC, a network interface controller that's used as a key component in a next generation server.
For example, some vendors use proprietary processors that don't benefit from the rich development and application infrastructure offered by the broad Arm CPU ecosystem.

Others claim to have DPUs but make the mistake of focusing solely on the embedded CPU to perform data path processing.

DPUs: A Focus on Data Processing

This isn't competitive and doesn't scale, because trying to beat the traditional x86 CPU with a brute force performance attack is a losing battle. If 100 Gigabit/sec packet processing brings an x86 to its knees, why would an embedded CPU perform better?

Instead the network interface needs to be powerful and flexible enough to handle all network data path processing. The embedded CPU should be used for control path initialization and exception processing, nothing more.

At a minimum, there are 10 capabilities the network data path acceleration engines need to be able to deliver:

- Data packet parsing, matching, and manipulation to implement an open virtual switch (OVS)
- RDMA data transport acceleration for Zero Touch RoCE
- GPU-Direct accelerators to bypass the CPU and feed networked data directly to GPUs (both from storage and from other GPUs)
- TCP acceleration including RSS, LRO, checksum, etc
- Network virtualization for VXLAN and Geneve overlays and VTEP offload
- Traffic shaping “packet pacing” accelerator to enable multi-media streaming, content distribution networks, and the new 4K/8K Video over IP (RiverMax for ST 2110)
- Precision timing accelerators for telco Cloud RAN such as 5T for 5G capabilities
- Crypto acceleration for IPSEC and TLS performed inline so all other accelerations are still operation
- Virtualization support for SR-IOV, VirtIO and para-virtualization
- Secure Isolation: root of trust, secure boot, secure firmware upgrades, and authenticated containers and application life cycle management

These are just 10 of the acceleration and hardware capabilities that are critical to being able to answer yes to the question: “What is a DPU?”

So what is a DPU? This is a DPU:

Many so-called DPUs focus solely on delivering one or two of these functions.

The worst try to offload the datapath in proprietary processors.

While good for prototyping, this is a fool's errand, because of the scale, scope, and breadth of data center.

Additional DPU-Related Resources

- Defining the SmartNIC: What is a SmartNIC and How to Choose the Best One
- Best Smart NICs for Building the Smart Cloud: PART I
- Welcome to the DPU-Enabled Data Revolution Era
- Accelerating Bare Metal Kubernetes Workloads, the Right Way
- Mellanox Introduces Revolutionary SmartNICs for Making Secure Cloud Possible
- Achieving a Cloud Scale Architecture with SmartNICs
- Provision Bare-Metal Kubernetes Like a Cloud Giant!

Categories: Data Center | Explainer | Networking

Tags: Data Science | GPU

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

NVIDIA Xavier Achieves Industry First with Expert Safety Assessment

World's first autonomous driving system-on-a-chip meets toughest safety standards, according to TÜV SÜD.

May 20, 2020 by GARY HICOK

Attaining the highest levels of safety takes years of hard engineering work and investment. Now, autonomous vehicle developers can achieve it with a single system-on-a-chip.

The NVIDIA Xavier SoC passed the final assessment for safety product approval by TÜV SÜD, one of the most knowledgeable and stringent safety assessment bodies in the industry.

TÜV SÜD has determined that the chip meets ISO 26262 random hardware integrity of ASIL C and a systematic capability of ASIL D for process — the strictest standard for functional safety.

NVIDIA Xavier, the world's first processor for autonomous driving, is the most complex SoC the safety agency has assessed in its 150-year history.

As part of a three-step approach, TÜV SÜD has previously assessed the process to develop Xavier as well as the SoC's architecture. This current assessment completes the last step to show that Xavier SoC meets all applicable requirements of ISO 26262.

NVIDIA is working with the entire industry to ensure the safe deployment of autonomous vehicles. It participates in standardization and regulation bodies worldwide, including the International Organization for Standardization (ISO), the Society of Automotive Engineers (SAE), the Institute of Electrical and Electronics Engineers (IEEE), the United Nations Economic Commission of Europe (UNECE), the National Highway Traffic Safety Administration (NHTSA), the Association for Standardization of Automation and Measuring Systems (ASAM) and the European Association of Automotive Suppliers (CLEPA).

By working with these groups — and having our technology reviewed by them — we're able to share our expertise while also delivering a robust AI computing system for the entire industry.

It Takes a Village

The TÜV SÜD assessment spans multiple disciplines in Xavier's development. The audit reviewed 1,400 internal work products across a range of cross-functional teams, all contributing to the most complex SoC ever assessed.

Xavier contains 9 billion transistors to process vast amounts of data, as well as thousands of safety mechanisms to address random hardware failures. Its MIPI CSI-2 and Gigabit Ethernet high-speed I/O connects Xavier to the largest array of lidar, radar and camera sensors of any chip ever built.

Inside the SoC, six types of processors — ISP (image signal processor), VPU (video processing unit), PVA (programmable vision accelerator), DLA (deep learning accelerator), CUDA GPU, and CPU — process 30 trillion operations per second.

Using any one of these components separately would require significant investment to achieve the same safety functionality as the complete Xavier SoC. By choosing Xavier, autonomous vehicle developers can meet the highest levels of safety with a single processor.

Raising the Bar

This milestone also marks Xavier as one of the first processors to meet the requirements of the latest ISO 26262 standard.

ISO 26262 is the definitive global standard for automotive functional safety — a system's ability to avoid, identify and manage failures. In 2018, the organization released the second edition of these standards to adapt to new vehicle technologies.

The standards cover the hardware itself as well as the processes that surround it — ensuring a product has been developed in a way that mitigates potential systematic and random hardware faults. That is, SoC development must not only avoid failures whenever possible, but also detect and respond to them when they cannot be avoided.

Under these standards, Xavier has been determined to meet the requirements for random hardware integrity of ASIL C and a systematic capability of ASIL D — the highest degree of safety integrity. ASIL refers to a component's automotive safety integrity level and classifies the ability to mitigate risk of hazard on a scale of A to D, A representing the lowest degree and D the highest.

By meeting these requirements, Xavier has demonstrated the ability to achieve the necessary complexity for high-performance compute, while also maintaining functional safety.

Completing this assessment is just the start of NVIDIA's journey to deliver safer and more efficient transportation. We continue to raise the bar for AI compute, ensuring safety in development and execution at every step.

Categories: Driving

Tags: Automotive | NVIDIA DRIVE

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

While the World Works from Home, NVIDIA's AV Fleet Drives in the Data Center

NVIDIA DRIVE Constellation enables high-fidelity, end-to-end simulation for development and validation of autonomous vehicles.

May 19, 2020 by ZVI GREENSTEIN

Share

As much of the world continues to conduct business from home, NVIDIA's autonomous test vehicles are hard at work in the cloud.

During the GTC 2020 keynote, NVIDIA CEO Jensen Huang demonstrated how NVIDIA DRIVE technology is being developed and tested in simulation. While physical testing is temporarily paused, the cloud-based NVIDIA DRIVE Constellation platform makes it possible to dispatch virtual vehicles in virtual environments to continue making great progress in self-driving technology.

In the video demonstration, a virtual NVIDIA BB8 test vehicle drives near NVIDIA headquarters in Silicon Valley, traveling through highways and urban streets — all in simulation. The 17-mile loop shows the NVIDIA DRIVE AV Software navigating the roadways, pedestrians and traffic in a highly accurate replica environment.

Data Center Proving Ground

NVIDIA DRIVE Constellation is a cloud-based simulation platform, designed from the ground up to support the development and validation of autonomous vehicles. The data center-based platform consists of two side-by-side servers.

The first server uses NVIDIA GPUs running DRIVE Sim software and generates the sensor output from the virtual car driving in a virtual world. The second server contains

the actual vehicle computer, processing the simulated sensor data running the exact same DRIVE AV and DRIVE IX software that's being deployed in the real car.

The driving decisions from the second server are fed back into the first, enabling real-time, bit-accurate, hardware-in-the-loop development and testing.

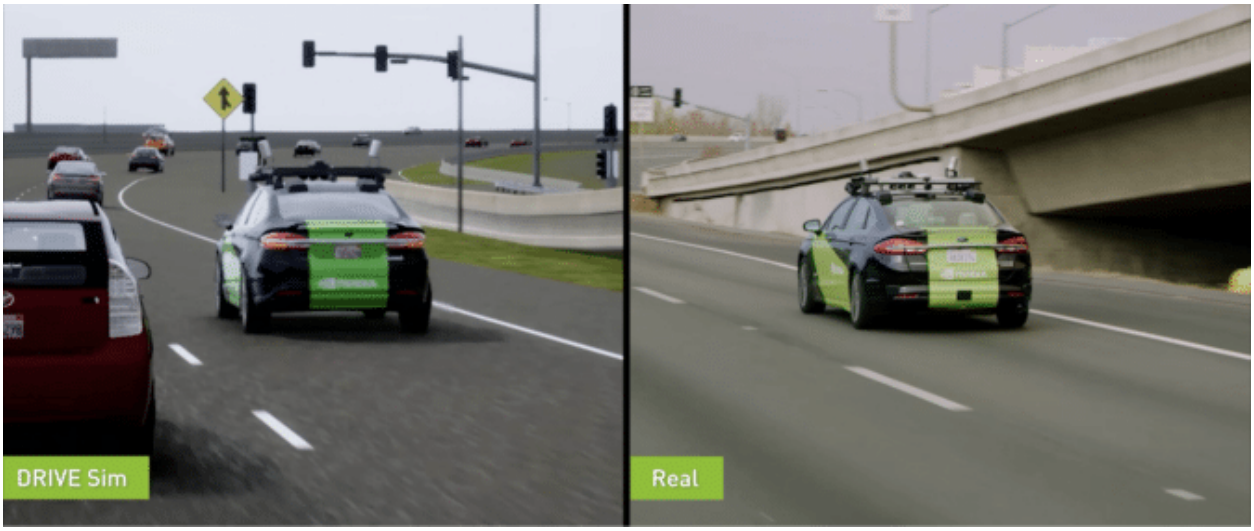
DRIVE Constellation is composed of two side-by-side servers enabling bit-accurate, hardware-in-the-loop testing.

The system is designed to be deployed in a data center as a scalable virtual fleet. This provides development engineers with a vehicle on demand, and gives them the ability to conduct testing at scale. It also makes it possible to consistently test rare and dangerous scenarios that are difficult or impossible to encounter in the real world.

Development and Testing from End to End

Building an autonomous vehicle requires testing at every level — starting at subsystems and continuing all the way to full vehicle integration tests. DRIVE Constellation enables this type of end-to-end development and testing for autonomous vehicles in simulation, similar to developing a physical car.

End-to-end tests ensure timing and performance accuracy as well as accurate modeling of the complex interdependency of different systems in autonomous vehicle software.



DRIVE Sim creates a digital twin of the real world to provide a realistic driving environment.
Achieving this level of fidelity at scale is a major undertaking. The environment, traffic behavior, sensor inputs and vehicle dynamics must appear, act and feed into the car computer just as they would in the real world.

This requires multiple GPUs to generate synthetic data in sync with precise timing. The vehicle software and hardware signals and interfaces must be replicated in simulation — and everything has to run in real time.

Simulating Silicon Valley

Comprehensive simulation starts with the environment. To accurately recreate the Silicon Valley driving loop, 3D Mapping, a member of the NVIDIA DRIVE ecosystem, scanned the roadways to within 5 centimeters of accuracy. The raw scanned data was then processed into a dataset format known as OpenDRIVE.

From there, NVIDIA developed a content creation pipeline to generate a highly accurate 3D environment using the NVIDIA Omniverse collaboration platform. The environment includes accurate road networks and roadmarks. Material properties are also applied to ensure it interacts with light rays, radio waves and lidar rays in the same way real sensors interact with the physical world.

DRIVE Sim allows end-to-end testing, including in-car visualization.

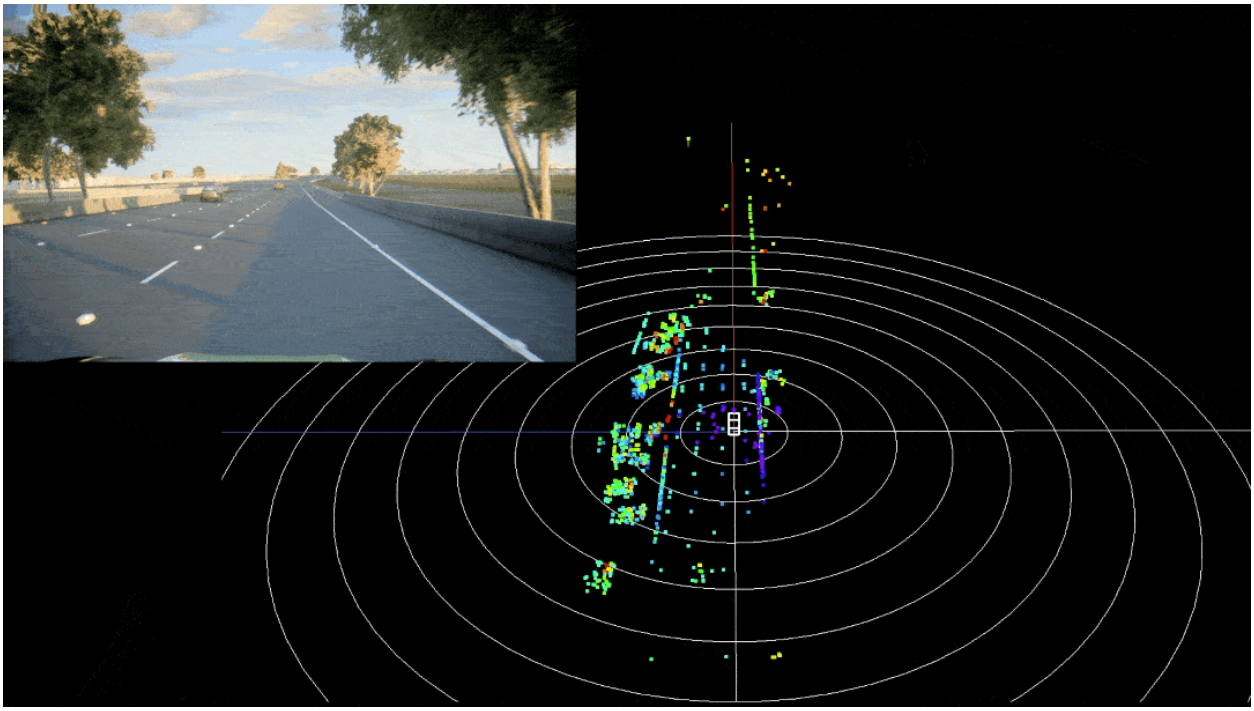
DRIVE Sim provides a wide range of light and weather conditions for testing AV software.

Recreating Sensor Data

With an accurate environment in place, high-fidelity development and testing next requires accurately generated sensor data. The sensor models include those typically found on an autonomous test vehicle, such as camera, lidar, radar and inertial measurement unit. DRIVE Sim provides a flexible sensor pipeline and APIs that allow configuring sensors to match real-world vehicle architectures.

For camera data, the image pipeline starts by rendering an HDR image that is warped according to the lens properties of the camera used on the vehicle. Exposure control, black and white level balancing, and color grading are applied to the image to match the sensor profile. Finally, the pixel data is converted to its native output format using a sensor-specific encoder.

In addition to camera models, DRIVE Sim provides physically based lidar and radar sensors using ray tracing. NVIDIA RTX GPUs enable DRIVE Sim to run highly computationally intensive radar and lidar models in real time.



DRIVE Constellation provides powerful RTX GPUs that allow real-time rendering of sensors using ray tracing. The scene above shows the combined returns of eight radars being rendered in real time for the AV stack.

Modeling Vehicle Behavior

Finally, vehicle models are critical for accurate simulation. As control signals — steering, acceleration and braking — are sent to the in-vehicle computer, the car must respond just as it would in the physical world.

To do so, the simulation platform must recreate motion properly, including details such as interaction with the road surface. Vehicle models in DRIVE Sim are handled using a plugin system with the included PhysX models or third-party vehicle dynamics models from NVIDIA DRIVE ecosystem partners such as Mechanical Simulation or IPG.

Vehicle dynamics also play a key role in accurate sensor data generation. As the vehicle operates, the position and pose of the vehicle changes significantly, affecting a sensor's viewpoint. For example, the forward-facing cameras will pitch downward when a car is braking. Modeling vehicle dynamics correctly is important to generating sensor data properly.

By accurately simulating each of these components — environment, sensors, vehicle dynamics — on a single, end-to-end platform, NVIDIA DRIVE Constellation and DRIVE Sim are critical pieces to a comprehensive development and testing pipeline. They enable NVIDIA and its partners to work toward safer and more efficient autonomous vehicles as physical fleets remain in the garage.

Categories: Driving

Tags: Automotive | GTC | NVIDIA DRIVE | Omniverse

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

COVID Caught on Camera: Startup's Sensors Keep Hospitals Safe

Whiteboard Coordinator deploys GPU-powered systems to fend off the coronavirus.

May 19, 2020 by RENEE YAO

Share

Andrew Gostine's startup aims to make hospitals more efficient, but when the coronavirus hit Chicago he pivoted to keeping them safer, too.

Gostine is a critical-care anesthesiologist at Northwestern Medicine's 105-bed Lake Forest hospital, caring for 60 COVID-19 patients. He's also the CEO of Whiteboard Coordinator Inc., a startup that had a network of 400 cameras and other sensors deployed across Northwestern's 10 hospitals before the pandemic.

After the virus arrived, "the hospital said it was having a hard time screening people coming in for COVID-19 using conventional temperature probes, and asked if we could help," he said.

Ten days later, the startup had thermal cameras linked to its network installed at 31 entrances to the hospitals. They detect about a dozen cases of fever in the 6,000 people coming through the doors each day.

The approach reduced lines waiting to get in. It also cut from four to one the number of people the hospital needed to post at each door.

Digital Window Protects Care Givers

About the same time, Northwestern asked Whiteboard for “a digital window” into COVID-19 rooms. They wanted to limit nurses’ exposure to the virus and reduce the need for the protective gear that’s now in high demand.

Whiteboard’s HIPAA-compliant thermal camera system can measure temperature to within +/- 0.3 °C on up to 36 people per video frame at a distance of nine meters.
So, the startup deployed another 400 cameras sporting night vision and microphones across the 10 hospitals. They use Whiteboard’s network of NVIDIA GPUs to transcode the video streams so they can be viewed securely on any hospital display.

“Nurses tell us the remote viewing is phenomenal. They report going into rooms less and consumption of protective gear is down. Our next challenge is using our computer-vision capabilities to track inventory of protective gear in real time,” he said.

The thermal cameras and patient monitors link to 36 NVIDIA RTX 2080 Ti GPUs. They handle transcoding and other algorithms to deliver low-latency feeds at 20 frames/second.

The current COVID-19 uses don’t require AI today, but deep learning is a core part of Whiteboard’s system. “Eighty percent of what we do is computer vision, but we can integrate different sensors for different problems,” Gostine said.

A Sensory-Friendly Guardian for Hospitals

Whiteboard's system also supports Bluetooth and RFID sensors for a range of patient monitoring, inventory tracking, resource scheduling and security apps. One hospital increased the use of its operating rooms 27 percent while reducing its costs, thanks to the startup's OR scheduling system. It currently runs on an NVIDIA Jetson TX2 and is being upgraded to Quadro 4000 GPUs.

For use cases such as fever and mask detection, Whiteboard also plans to adopt NVIDIA Clara Guardian, an application framework that simplifies the deployment in hospitals of smart sensors with multi-mode AI. It is among 18 companies currently supporting Clara Guardian, software that runs on the NVIDIA EGX platform for AI computing on edge servers and embedded devices.

The pandemic spawned orders from 100 hospitals for Whiteboard's thermal cameras. The startup currently has at least one of its systems installed in a total of 22 hospitals.

"Our biggest problem is sourcing cameras and other hardware we need because supply chains are in disarray," Gostine said.

Seeking Better Surgery Outcomes with AI

Once the pandemic passes, the startup aims to employ AI to improve outcomes of surgical techniques used in the operating room. Long term, Whiteboard's value will come from its expanding AI algorithms and datasets, trained on NVIDIA V100 Tensor Core GPUs in Microsoft's Azure service, he said.

It's a big opportunity. Accenture predicts by 2026 the top 10 AI healthcare apps will generate a $150 billion market. It will span areas such as robotic surgery, virtual nursing assistants and automated workflows.

The startup's mission was born of Gostine's personal passion for making hospitals more modern and efficient.

"When I got to med school, I was frustrated because it seemed we lagged behind the internet era I grew up in," he said.

From Faxes to the Future

After graduating, he got an MBA and spent some time consulting with healthcare startups before his internship. Work with more than a dozen companies led to a position with a VC firm during his medical residency.

“It was like night and day. The venture world was thinking 10 years ahead, and I realized healthcare was really behind — we’re still using pagers and fax machines,” he said.

“There’s so much paperwork to get through every day, just so we can think about our patients. What we are doing at Whiteboard really stems from the frustrations I felt in my practice,” he added.

A series of chance encounters led him to three AI, software and medical experts who formed Whiteboard, a member of NVIDIA’s Inception program, which gives startups access to new technologies and other resources.

Whiteboard’s first product aimed to streamline OR scheduling, then it expanded into patient monitoring. Now the coronavirus has taken its networks all the way to the hospital’s front door.

Categories: Deep Learning

Tags: Artificial Intelligence | Computer Vision | COVID-19 | Inception | Jetson | Machine Learning | Medical Research and Healthcare | Metropolis | NVIDIA EGX | NVIDIA RTX | Quadro | Social Impact

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics
Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight
Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program
The Great AI Bake-Off: Recommendation Systems on the Rise
Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

Create at the Speed of Imagination with New Thin and Light Devices from Dell, HP and Microsoft

Over 75 RTX Studio laptops and mobile workstations powered by GeForce and Quadro RTX GPUs have launched in the last year.

May 19, 2020 by GERARDO DELGADO

Share

Creative workflows are getting more demanding. Project timelines continue to shrink. And with many people working remotely, a fast and reliable computer is more important than ever.

New NVIDIA-powered laptops and mobile workstations from Dell, HP and Microsoft give creators amazing choices to turn their imagination into actual creations.

These new systems launch just shy of the one-year anniversary of our introduction of NVIDIA Studio, a platform featuring dedicated drivers, performance-enhancing software development kits, and thin and light RTX Studio laptops purpose-built for creators.

Since then, we've worked with every major system manufacturer to expand the RTX Studio lineup and provide a wide range of choices that feature NVIDIA Quadro and GeForce RTX GPUs. In total, there are now 78 RTX Studio systems.

Dell-ightful RTX Studio Laptops and Mobile Workstations

Designed to be Dell's most powerful XPS laptop ever, the XPS 17 muscles through intensive creative projects and gaming alike, with up to NVIDIA GeForce RTX 2060 graphics. Thanks to a thin bezel design, it's the smallest 17-inch laptop, with similar dimensions to a typical 15-inch one.

Dell XPS 17

Packing this much performance into an RTX Studio laptop of this size requires a little engineering ingenuity to keep the system performing smoothly. Under the hood is a new proprietary thermal design that provides more overall airflow and higher sustained performance to fuel the most demanding projects.

Today, Dell announced it has reengineered its Precision workstation portfolio. These RTX Studio mobile workstations are designed to handle demanding workloads like 8K editing, 3D rendering, data analysis and CAD modeling. The Precision 5750 and 7000 series mobile workstations are ISV certified and feature Quadro RTX graphics.

Dell Precision 5750
The Dell Precision 5750 allows creators and engineers to see and do more with up to Quadro RTX 3000 graphics and a 16:10, four-sided InfinityEdge (up to HDR 400) display. For editors, engineers and scientists running intensive workloads, the Dell Precision 7550 and Dell Precision 7750 are available with up to Quadro RTX 5000 GPUs. They've been reengineered with more power and intelligent performance in an even smaller, lighter footprint.

A Laptop to ENVY

HP's new Create Ecosystem is empowering creators of all types. That starts with the HP ENVY 15 that'll be available later this month on HP.com.

HP ENVY 15

The ENVY 15 is an RTX Studio laptop that can be configured with up to a GeForce RTX 2060 GPU for the ultimate in creator performance. It also features an all-aluminum chassis with 82.8 percent screen-to-body ratio, up to a 4K OLED VESA DisplayHDR 400 True Black display with touch interface display, 10th gen Intel processors, and gaming-class thermals.

Creative pros will want to keep an eye out for HP's ZBook Studio and ZBook Create. Shipping later this year, they provide true mobility without compromise, thanks to Quadro and GeForce RTX GPUs, respectively. These systems also feature an 87 percent screen-to-body ratio and bring creators the first DreamColor display with all-day battery life.

Inside the Surface

Microsoft recently announced their most powerful laptop ever, the Surface Book 3. And it's being powered by Quadro RTX and GeForce GTX graphics. The Surface Book 3

has the power of a desktop, the versatility of a tablet and the freedom of a thin and light laptop in one beautifully designed device.

Microsoft Surface Book 3

In addition to NVIDIA graphics, the Surface Book 3 can be configured with 10th Generation Intel Core processors, up to 32GB of RAM, the fastest SSD Microsoft has ever shipped, a beautifully crisp, high-DPI PixelSense display, and up to 17.5 hours of battery life.

Surface Book 3 starts at $1,599 and will be available starting May 21.

GPU-Accelerated Exports in Adobe Premiere Pro

These new laptops will take advantage of over 200 NVIDIA GPU-accelerated creative and design applications, including one major addition released just yesterday.

Adobe Premiere Pro is helping content creators go from concept to completion faster with new GPU-accelerated exports. With NVIDIA encoder acceleration in Adobe Premiere Pro, editors can export high-resolution videos up to five times quicker than on CPU.

Adobe Premiere Pro, now with NVENC support

Creators can still take advantage of a limited time offer. Purchase a new RTX Studio laptop or desktop and both new and existing Adobe users get a free three-month subscription to Adobe Creative Cloud.

These new laptops, along with 10 recently announced RTX Studio laptops powered by new GeForce RTX SUPER GPUs, are powering the creative intersection between imagination and innovation.

Learn more about NVIDIA Studio and RTX Studio systems. And stay tuned for exciting announcements as the platforms continue to grow.

Categories: Laptops | Mobile | Workstation

Tags: GeForce | NVIDIA RTX | NVIDIA Studio | Quadro Mobile Workstations

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics
Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight
Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program
The Great AI Bake-Off: Recommendation Systems on the Rise
Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

Cut to the Video: Adobe Premiere Pro Helps Content Creators Work Faster with GPU-Accelerated Exports

With NVIDIA encoder acceleration in Adobe Premiere Pro, editors can export high-resolution videos up to 5x faster than on CPU.

May 19, 2020 by STANLEY TACK

Share

With more people working from home, video editors are being challenged to deliver content in new ways. Many are using footage shot at home or getting creative with stock footage to meet the demands for fresh content.

With the latest release of Adobe Premiere Pro, available today, creators can get new NVIDIA GPU-enhanced features that help them deliver high-quality content faster than ever.

Elevate Editing Workflows with GPU Acceleration

With the new Premiere Pro 14.2, video creators gain massive time-savings with new GPU-accelerated encoding. Adobe and NVIDIA have optimized Premiere Pro for the built-in NVIDIA hardware encoder on NVIDIA Quadro and GeForce GPUs.

The results are staggering. Editors can now export high-resolution videos up to 5x faster than with CPU alone by using the popular H.264 or H.265 / HEVC codecs. Less time exporting means more time for editing content, and quicker turnarounds on projects.

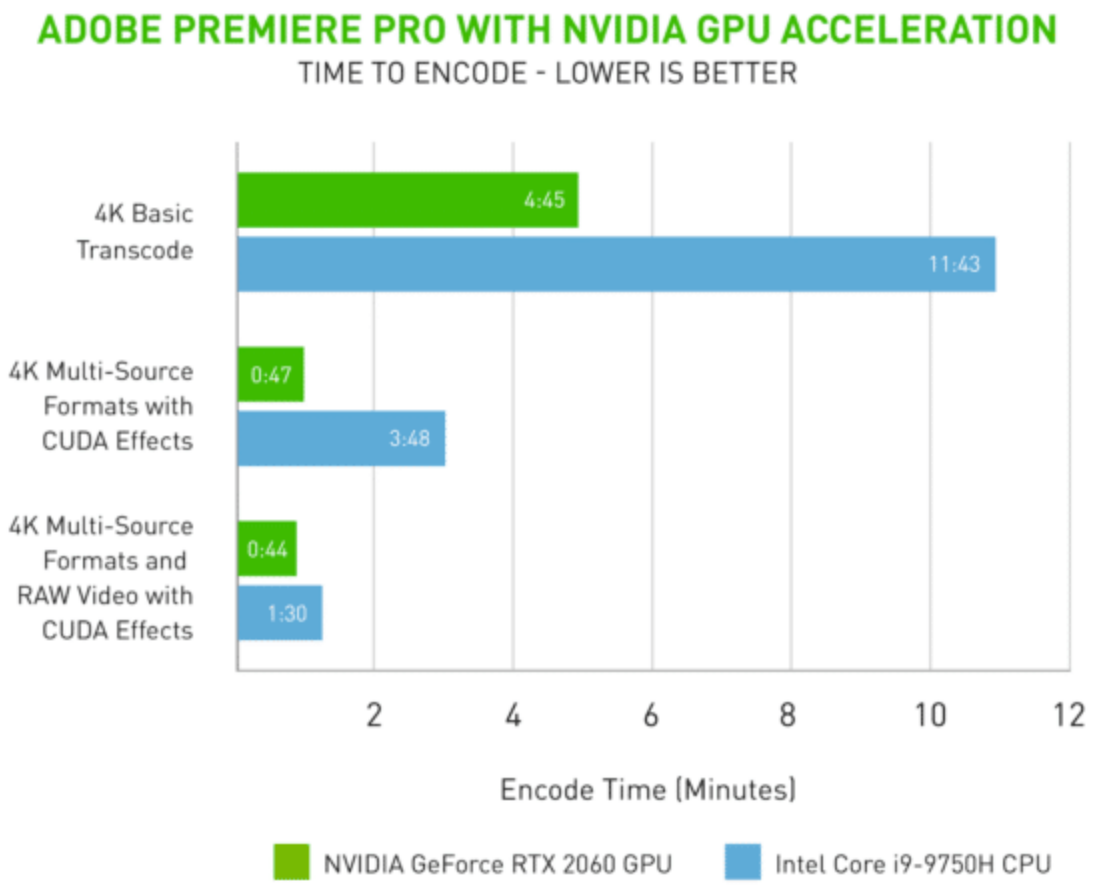
“I’m often required to export multiple versions of my videos. Sometimes to submit for approval, but mostly I just prefer to error check the final render instead of playing back the timeline,” said YouTuber Gerald Undone. “With NVENC integration into Premiere Pro, I can do this step in a third of the time, which should equate to dozens of hours saved by the end of the year.”

And thanks to enhancements in the encoder on our latest NVIDIA GeForce and Quadro RTX GPUs, encoding quality and efficiency are second to none.

For example, the music video below is three minutes and nine seconds long. With traditional software encoding using a Core i9-9750H laptop CPU, it takes 3:48 to export. By using the NVIDIA hardware encoder on a GeForce RTX 2060 Max-Q GPU, the export completes in one-fifth the time — a mere 47 seconds.

In addition to Premiere Pro, the NVIDIA hardware encoder speeds up video exports in Adobe Media Encoder, After Effects and Audition.

“With NVENC, our workflow has drastically improved,” said cinematographer Armando Ferreira. “We are able to playback higher resolutions in real time in our timeline and export up to 40 percent faster.”



Video encoding joins a growing list of Premiere Pro features enhanced by NVIDIA GPUs, including accelerated video effects using CUDA, and Auto Reframe with GPU-accelerated AI.

“These improvements are the result of years of collaboration between NVIDIA and Adobe to deliver high-quality applications and tools to creators,” said Manish Kulkarni, senior engineering manager at Adobe. “With new support for NVIDIA GPUs on Windows, exports are hardware accelerated leveraging the power of the GPU to make Premiere Pro more powerful and keep video creators productive and nimble.”

Also included in today’s release is support for Apple’s ProRes RAW in both Premiere Pro and After Effects. For the first time, video editors and motion graphics artists can import and edit ProRes RAW files in Windows with no need to transcode. This is accelerated by CUDA, available exclusively on NVIDIA GPUs.

Don’t yet have access to Premiere Pro? Get a free three-month subscription to Adobe Creative Cloud with the purchase of an RTX Studio laptop or desktop.

Find out more about the latest Premiere Pro release.

Categories: Pro Graphics

Tags: GeForce | NVIDIA RTX | NVIDIA Studio | Quadro

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA’s David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven ‘Double Robot’ Supports Remote Working

Post navigation

Amped Up: HPC Centers Ride A100 GPUs to Accelerate Science

Supercomputers put AI in the loop, moving into the exascale era with the NVIDIA Ampere architecture.

May 15, 2020 by DION HARRIS

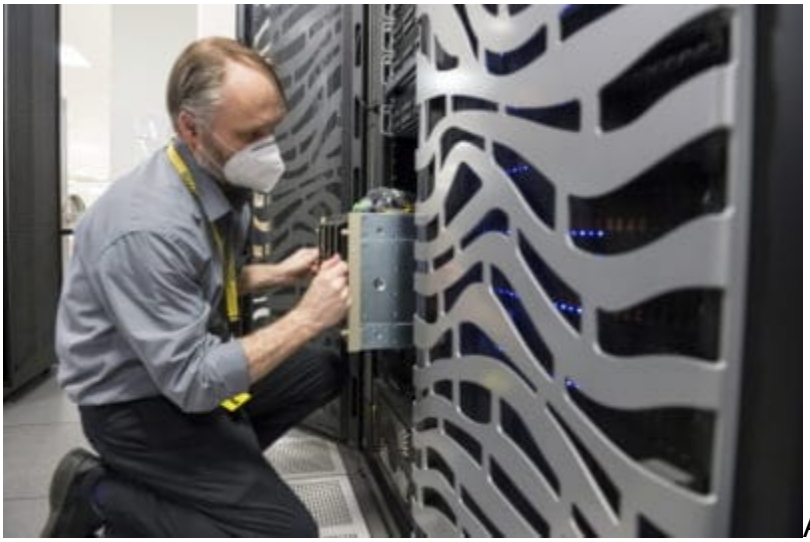
Share

Six supercomputer centers around the world are among the first to adopt the NVIDIA Ampere architecture. They'll use it to bring science into the exascale era in fields from astrophysics to virus microbiology.

The high performance computing centers scattered across the U.S. and Germany will use a total of nearly 13,000 A100 GPUs.

Together these GPUs pack more than 250 petaflops in peak performance for simulations that use 64-bit floating point math. For AI inference jobs that use mixed precision math and leverage the A100 GPU's support for sparsity, they deliver a whopping 8.07 exaflops.

Researchers will harness that horsepower to drive science forward in many dimensions. They plan to simulate larger models, train and deploy deeper networks, and pioneer an emerging hybrid field of AI-assisted simulations.

 Argonne deployed one of the first NVIDIA DGX-A100 systems. Photo courtesy of Argonne National Laboratory.
For example, Argonne's researchers will seek a COVID-19 vaccine by simulating a key part of a protein spike on a coronavirus that's made up of as many as 1.5 million atoms.

The molecule "is a beast, but the A100 lets us accelerate simulations of these subsystems so we can understand how this virus infects humans," said Arvind Ramanathan, a computational biologist at Argonne National Laboratory that will use a cluster of 24 NVIDIA DGX A100 systems.

In other efforts, "we will see substantial improvement in drug discovery by scanning millions and billions of drugs at a time. And we may see things we could never see before, like how two proteins bind to one another," he said.

A100 Puts AI in the Scientific Loop

"Much of this work is hard to simulate on a computer, so we use AI to intelligently guide where and when we will sample next," said Ramanathan.

It's part of an emerging trend of scientists using AI to steer simulations. The GPUs then will speed up the time to process biological samples by "at least two orders of magnitude," he added.

Across the country, the National Energy Research Scientific Computing Center (NERSC) is poised to become the largest of the first wave of A100 users. The center in Berkeley, Calif., is working with Hewlett Packard Enterprise to deploy 6,200 of the GPUs in Perlmutter, its pre-exascale system.

“Across NERSC’s science and algorithmic areas, we have increased performance by up to 5x when comparing a single V100 GPU to a KNL CPU node on our current-generation Cori system, and we expect even greater gains with the A100 on Perlmutter,” said Sudip Dosanjh, NERSC’s director.

Exascale Computing Team Works on Simulations, AI

A team dedicated to exascale computing at NERSC has defined nearly 30 projects for Perlmutter that use large-scale simulations, data analytics or deep learning. Some projects blend HPC with AI, such as one using reinforcement learning to control light source experiments. Another employs generative models to reproduce expensive simulations at high-energy physics detectors.

Two of NERSC’s HPC applications already prototyped use of the A100 GPU’s double-precision Tensor Cores. They’re seeing significant increases in performance over previous generation Volta GPUs.

Software optimized for the 10,000-way parallelism Perlmutter’s GPUs offer will be ready to run on future exascale systems, Christopher Daley, an HPC performance engineer at NERSC said in a talk at GTC Digital. NERSC supports nearly a thousand scientific applications in areas such as astrophysics, Earth science, fusion energy and genomics.

“On Perlmutter, we need compilers that support all the programming models our users need and expect — MPI, OpenMP, OpenACC, CUDA and optimized math libraries. The NVIDIA HPC SDK checks all of those boxes,” said Nicholas Wright, NERSC’s chief architect.

German Effort to Map the Brain

AI will be the focus of some of the first applications for the A100 on a new 70-petaflops system designed by France’s Atos for the Jülich Supercomputing Center in western Germany.

One, called Deep Rain, aims to make fast, short-term weather predictions, complementing traditional systems that use large, relatively slow simulations of the atmosphere. Another project plans to construct an atlas of fibers in the human brain, assembled with deep learning from thousands of high-resolution 2D brain images.

The new A100 system at Jülich also will help researchers push the edges of understanding the strong forces binding quarks, the sub-atomic building blocks of matter. At the macro scale, a climate science project will model the Earth’s surface and subsurface water flow.

“Many of these applications are constrained by memory,” said Dirk Pleiter, a theoretical physicist who manages a research team in applications-oriented technology development at Jülich. “So, what is extremely interesting for us is the increased memory footprint and memory bandwidth of the A100,” he said.

The new GPU’s ability to accelerate double-precision math by up to 2.5x is another feature researchers are keen to harness. “I’m confident when people realize the opportunities of more compute performance, they will have a strong incentive to use GPUs,” he added.

Data-Hungry System Likes Fast NVLink

Some 230 miles south of Jülich, the Karlsruhe Institute of Technology (KIT) is partnering with Lenovo to build a new 17-petaflops system that will pack 740 A100 GPUs on an NVIDIA Mellanox 200 Gbit/s InfiniBand network. It will tackle grand challenges that include:

- Atmospheric simulations at the kilometer scale for climate science
- Research to fight COVID-19, including support for Folding@home
- Explorations of particle physics beyond the Higgs boson for the Large Hadron Collider
- Research on next-generation materials that could replace lithium-ion batteries
- AI applications in robotics, language processing and renewable energy

“We focus on data-intensive simulations and AI workflows, so we appreciate the third-generation NVLink connecting the new GPUs,” said Martin Frank, director of KIT’s supercomputing center and a professor of computational science and math.

“We also look forward to the multi-instance GPU feature that effectively gives us up to 28 GPUs per node instead of four — that will greatly benefit many of our applications,” he added.

Just outside Munich, the computer center for the Max Planck Institute is creating with Lenovo a system called Raven-GPU, powered by 768 NVIDIA A100 GPUs. It will support work in fields like astrophysics, biology, theoretical chemistry and advanced materials science. The research institute aims to have Raven-GPU installed by the end of the year and is taking requests now for support porting applications to the A100.

Indiana System Counters Cybersecurity Threats

Finally, Indiana University is building Big Red 200, a 6 petaflops system expected to become the fastest university-owned supercomputer in the U.S. It will use 256 A100 GPUs.

Announced in June, it's among the first academic centers to adopt the Cray Shasta technology from Hewlett Packard Enterprise that others will use in future exascale systems.

Big Red 200 will apply AI to counter cybersecurity threats. It also will tackle grand challenges in genetics to help enable personalized healthcare as well as work in climate modeling, physics and astronomy.

Photo at top: Shyh Wang Hall at UC Berkeley will be the home of NERSC's Perlmutter supercomputer.

Categories: Data Center | Deep Learning | Supercomputing

Tags: Artificial Intelligence | COVID-19 | Customer Stories | GPU | High Performance Computing | Machine Learning | New GPU Uses | NVIDIA Ampere Architecture | NVIDIA DGX | Parallel Computing | Science

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

NVIDIA Bolsters NGC Private Registry with New Security Features

GPU-optimized container catalog expands AI toolkit offering and adds multi-architecture support.

May 15, 2020 by ADEL EL HALLAK

If data is the new oil, then AI is the refinery. The raw data enterprises create and gather is like black crude — it's real value shines after it's processed into gasoline and other consumable products.

AI transforms the vast deposits of data organizations possess to help them extract insights that improve customer experiences, build new business models, drive operational efficiencies and fuel competitive advantage.

NVIDIA's GPU computing platform, including NGC, NGC-Ready systems and NGC Support services, powers the AI refinery. NGC is the software hub that provides GPU-optimized frameworks, pre-trained models and toolkits to train and deploy AI in production.

And today we're expanding NGC to help developers securely build AI faster with toolkits and SDKs and share and deploy with a private registry.

NVIDIA AI Toolkits and SDKs Simplify Training, Inference and Deployment

NVIDIA AI toolkits provide libraries and tools to train, fine-tune, optimize and deploy pre-trained NGC models across a broad domain of industries and AI applications. They include:

- An AI-assisted annotation tool to help users label their datasets for training.
- A transfer learning toolkit to fine-tune pre-trained models with user data, saving the cost of training from scratch.
- Federated learning that preserves privacy by allowing users to collaborate and train AI models without sharing private data between clients.
- The NeMo toolkit to quickly build state-of-the-art models for speech recognition and natural language processing.
- The Service Maker toolkit that exposes trained models as a gRPC service that can be scaled and easily deployed on a Kubernetes service.

Models built using the toolkits can be integrated inside client applications and deployed in production with the software development kits offered by NVIDIA.
Leveraging TensorRT and the Triton Inference Server as foundational building blocks, the deployment SDKs span industries, including conversational AI with NVIDIA Jarvis, recommender systems with NVIDIA Merlin, medical imaging with NVIDIA Clara, video analytics with NVIDIA Metropolis and NVIDIA DeepStream, robotics with NVIDIA Isaac Sim and 5G acceleration with NVIDIA Aerial.

Safeguarding IP with NGC Private Registry

Custom models built using these AI toolkits are a company's highly guarded intellectual property. Organizations will want to ensure that they can be securely shared, modified and deployed into production environments.

That's why we've extended NGC with a private registry that allows organizations to sign, store, manage and share custom-built containers and models securely within teams across their enterprise.

The registry is equipped with several new security capabilities. Containers pushed to the private registry are automatically scanned for common vulnerabilities and exposures (CVEs). By flagging containers with critical or high CVEs, enterprises can remedy issues early in the development stages, ensuring secure software is released in production.

The private registry will also soon allow publishers and developers to sign their containers. This will give software users the confidence that they're consuming content that is authentic, intact and from a verified source.

As data scientists collaborate across their organizations and iterate on training, they produce many models, each with unique parameters. The NGC model-versioning system makes it easy to capture these unique variables, discover the appropriate models and, ultimately, deploy the right one into production.

Securely Deploy AI Models at the Edge

Enabling the smart everything revolution, AI models are increasingly being deployed at the edge, closer to the point of action. However, edge locations often don't have the IT personnel or security protocols present in a data center, which poses potential security challenges.

Announced yesterday, the EGX A100 has a confidential AI enclave that uses a new GPU security engine to load encrypted AI models, further preventing the theft of valuable IP. NGC private registry will enable developers to encrypt AI models and transfer them securely to the edge. Model encryption at rest and in transit from the private registry to the GPU memory reduces attacks on IP encoded in AI models running at the edge.

The private registry ultimately provides users with a cloud-hosted platform to secure their custom containers and models that they will deploy at the edge.

Run Seamlessly on x86, Arm and POWER Systems

AI workflows span the data center, cloud and the edge and may need to run on different CPU architectures. To enable flexibility, we've upgraded NGC with multi-architecture support.

This allows popular runtimes, including Docker, cri-o, containerD and Singularity, to automatically select and run an image variant that matches the system architecture, simplifying the deployment process across heterogeneous environments.

In addition to x86- and Arm-based containers, NGC now supports POWER architecture thanks to IBM bringing its Visual Insights software onto the registry. IBM Visual Insights makes computer vision with deep learning more accessible to business users, accelerating AI vision deployments and increasing productivity.

Get Started with NGC Today

Quickly build your AI solutions with GPU-optimized software tools from NGC, share and collaborate with the private registry, and securely deploy across on-premises, cloud and edge systems. Get started at ngc.nvidia.com.

Categories: Deep Learning | Software

Tags: Jarvis | Merlin | Metropolis | NGC | TensorRT

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

TensorFloat-32 in the A100 GPU Accelerates AI Training, HPC up to 20x

NVIDIA's Ampere architecture with TF32 speeds single-precision work, maintaining accuracy and using no new code.

May 14, 2020 by PARESH KHARYA

Share

As with all computing, you've got to get your math right to do AI well. Because deep learning is a young field, there's still a lively debate about which types of math are needed, for both training and inferencing.

In November, we explained the differences among popular formats such as single-, double-, half-, multi- and mixed-precision math used in AI and high performance computing. Today, the NVIDIA Ampere architecture introduces a new approach for improving training performance on the single-precision models widely used for AI.

TensorFloat-32 is the new math mode in NVIDIA A100 GPUs for handling the matrix math also called tensor operations used at the heart of AI and certain HPC applications. TF32 running on Tensor Cores in A100 GPUs can provide up to 10x speedups compared to single-precision floating-point math (FP32) on Volta GPUs. Combining TF32 with structured sparsity on the A100 enables performance gains over Volta of up to 20x.

Understanding the New Math

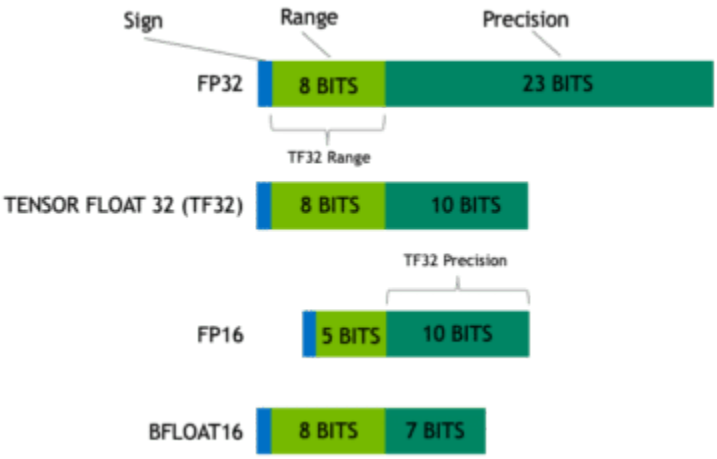
It helps to step back for a second to see how TF32 works and where it fits.

Math formats are like rulers. The number of bits in a format's exponent determines its range, how large an object it can measure. Its precision — how fine the lines are on the

ruler — comes from the number of bits used for its mantissa, the part of a floating point number after the radix or decimal point.

A good format strikes a balance. It should use enough bits to deliver precision without using so many it slows processing and bloats memory.

The chart below shows how TF32 is a hybrid that strikes this balance for tensor operations.



TF32 strikes a balance that delivers performance with range and accuracy.
TF32 uses the same 10-bit mantissa as the half-precision (FP16) math, shown to have more than sufficient margin for the precision requirements of AI workloads. And TF32 adopts the same 8-bit exponent as FP32 so it can support the same numeric range.

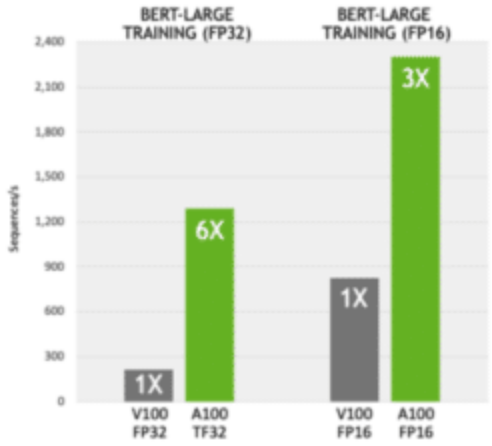
The combination makes TF32 a great alternative to FP32 for crunching through single-precision math, specifically the massive multiply-accumulate functions at the heart of deep learning and many HPC apps.

Applications using NVIDIA libraries enable users to harness the benefits of TF32 with no code change required. TF32 Tensor Cores operate on FP32 inputs and produce results in FP32. Non-matrix operations continue to use FP32.

For maximum performance, the A100 also has enhanced 16-bit math capabilities. It supports both FP16 and Bfloat16 (BF16) at double the rate of TF32.
Employing Automatic Mixed Precision, users can get a further 2x higher performance with just a few lines of code.

TF32 Is Demonstrating Great Results Today

Compared to FP32, TF32 shows a 6x speedup training BERT, one of the most demanding conversational AI models. Applications-level results on other AI training and HPC apps that rely on matrix math will vary by workload.



BERT Large Training (FP32 & FP16) measures Pre-Training phase, uses PyTorch including (2/3) Phase1 with Seq Len 128 and (1/3) Phase 2 with Seq Len 512, V100 is DGX1 Server with 8xV100, A100 is DGX A100 Server with 8xA100, A100 uses TF32 Tensor Core for FP32 training

To validate the accuracy of TF32, we used it to train a broad set of AI networks across a wide variety of applications from computer vision to natural language processing to recommender systems. All of them have the same convergence-to-accuracy behavior as FP32.

That's why NVIDIA is making TF32 the default on its cuDNN library which accelerates key math operations for neural networks. At the same time, NVIDIA is working with the open-source communities that develop AI frameworks to enable TF32 as their default training mode on A100 GPUs, too.

In June, developers will be able to access a version of the TensorFlow framework and a version of the PyTorch framework with support for TF32 on NGC, NVIDIA's catalog of GPU-accelerated software.

"TensorFloat-32 provides a huge out-of-the-box performance increase for AI applications for training and inference while preserving FP32 levels of accuracy," said Kemal El Moujahid, director of Product Management for TensorFlow.

"We plan to make TensorFloat-32 supported natively in TensorFlow to enable data scientists to benefit from dramatically higher speedups in NVIDIA A100 Tensor Core GPUs without any code changes," he added.

"Machine learning researchers, data scientists and engineers want to accelerate time to solution," said a spokesperson for the PyTorch team. "When TF32 is natively integrated

into PyTorch, it will enable out-of-the-box acceleration with zero code changes while maintaining accuracy of FP32 when using the NVIDIA Ampere architecture-based GPUs."

TF32 Accelerates Linear Solvers in HPC

HPC apps called linear solvers — algorithms with repetitive matrix-math calculations — also will benefit from TF32. They're used in a wide range of fields such as earth science, fluid dynamics, healthcare, material science and nuclear energy as well as oil and gas exploration.

Linear solvers using FP32 to achieve FP64 precision have been in use for more than 30 years. Last year, a fusion reaction study for the International Thermonuclear Experimental Reactor demonstrated that mixed-precision techniques delivered a speedup of 3.5x for such solvers using NVIDIA FP16 Tensor Cores. The same technology used in that study tripled the Summit supercomputer's performance on the HPL-AI benchmark.

To demonstrate the power and robustness of TF32 for linear system solvers, we ran a variety of tests in the SuiteSparse matrix collection using cuSOLVER in CUDA 11.0 on the A100. In these tests, TF32 delivered the fastest and most robust results compared to other Tensor Core modes, including FP16 and BF16.

Beyond linear solvers, other domains in high performance computing make use of FP32 matrix operations. NVIDIA plans to work with the industry to study the application of TF32 to more use cases that rely on FP32 today.

Where to Go for More Information

To get the big picture on the role of TF32 in our latest GPUs, watch the keynote with NVIDIA founder and CEO Jensen Huang. To learn even more, register for webinars on mixed-precision training or CUDA math libraries or read a detailed article that takes a deep dive into the NVIDIA Ampere architecture.

TF32 is among a cluster of new capabilities in the NVIDIA Ampere architecture, driving AI and HPC performance to new heights. For more details, check out our blogs on:

- Our support for sparsity, driving up to 50 percent improvements for AI inference.
- Double-precision Tensor Cores, speeding HPC simulations up to 2.5x
- Multi-instance GPU (MIG), supporting up to 7x in GPU productivity gains.
- Or, see the web page describing the NVIDIA A100 GPU.

Categories: Explainer

Tags: Artificial Intelligence | GTC 2020 | High Performance Computing | Machine Learning | NVIDIA Ampere Architecture

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

Omniverse Delivers Interactivity and Collaboration for Early Access Customers

May 14, 2020 by RICHARD KERRIS

Share

NVIDIA Omniverse, a computer graphics and simulation platform that makes it possible for artists to work seamlessly in real time across software applications on premises or around the world via the cloud, is now available for early access customers in the architecture, engineering and construction (AEC) market.

NVIDIA CEO Jensen Huang made the announcement during the GTC 2020 keynote, where he previewed an update of Omniverse highlighting simulation and real-time GPU rendering. He showcased “Marbles,” a playable game environment displaying real-time physics with dynamic lighting and rich physically based materials, as well as the platform’s latest “AEC Experience” feature set, which provides seamless connectivity between CAD applications with real-time visualization.

Real-Time Collaboration for Real-Time Creativity

Creating visual effects, architectural visualizations and manufacturing designs typically requires multiple people collaborating across teams, remote work locations and various customer sites for reviews. 3D assets take shape using an assortment of software tools.

But seamless data transfers across applications have long been the challenge for millions of artists, designers, architects, engineers and developers globally. Using Pixar’s Universal Scene Description and NVIDIA RTX technology, Omniverse allows people to easily work with applications and collaborate simultaneously with colleagues and customers, wherever they may be.

For years, collaborating in visual effects has been through exporting and importing large datasets and scene files. In 2010, Lucasfilm and Sony Pictures Imageworks joined together to create Alembic, an open computer graphics interchange framework that

helped multiple studios ensure consistency and accuracy when working on the same project.

Pixar developed USD to provide the interchange of assets and enable collaboration on 3D scenes that may be intricately composed from many elemental assets. With a single scene graph and consistent API, USD delivers a rich toolset for reading, writing, editing and rapidly previewing 3D geometry and shading.

Omniverse benefits from the flexibility and consistency of the USD interchange format and builds upon it with synchronized workflows. Entire studio teams around the world can collaborate in real time as they create, with version control support needed to ensure production stays on track.

In Omniverse, the Portals connection module unites top industry tools into a collaborative space for users to work seamlessly on real-time modeling, shading, animation, lighting, visual effects and rendering, introducing incredible new opportunities for creativity and production.

Leeza SOHO, Beijing by Zaha Hadid Architects.

Omniverse View with RTX: New Class of Renderer Blends Real-Time Speed and Offline Quality

Up until now, there have been two types of renderers. Real-time rendering is geared toward producing images at 30 or 60 frames per second and adheres to the lowest

device targeted for use. Offline rendering focuses on delivering photorealistic final images or scenes that take hours per frame to render with a CPU. To achieve top speeds, many corners often get cut — from simplifying geometry to baking lighting and normal maps — and this can reduce image quality.

To overcome this, Omniverse introduces a new type of rendering with Omniverse View. This module is accelerated by multiple NVIDIA RTX GPUs and built for extreme scalability on arrays of GPUs to provide high-quality, real-time output, even on very large scenes.

Omniverse View displays the 3D content aggregated from different applications inside Omniverse, or directly in the 3D application being used. It's also designed to support commercial game engines such as Unreal Engine and Unity as well as offline renderers.

Software Partners Empower Real-Time, Collaborative Creativity

Many industry software leaders are integrating their applications into Omniverse so artists can work collaboratively through the 3D creative process.

Omniverse's Portals is made possible by using Omniverse Kit, a software development kit that enables RTX View capability to be directly integrated into the software partner's application interface. This gives their product an ultra-high-quality, real-time, ray-traced application viewport.

Since Omniverse is a software-defined platform, vendor applications can also benefit from NVIDIA RTX acceleration or other technologies, like PhysX, as soon as new features become available. NVIDIA has been working closely with companies like Epic Games, Autodesk, Pixar, Adobe, Trimble, McNeel & Associates, and Teradici, with other partner announcements coming soon.

"Epic Games has had a long-standing partnership with NVIDIA," said Miles Perkins, business development at Epic Games. "Unreal Engine continues to set the standard for the world's most open and advanced real-time 3D tool for creators, and with NVIDIA's Omniverse platform we are seamlessly connected, enabling an even broader set of third-party products to work with our technology providing cutting-edge photorealistic content, new collaborative workflows, and immersive virtual worlds for games, enterprise, or any discipline dependent on visualization."

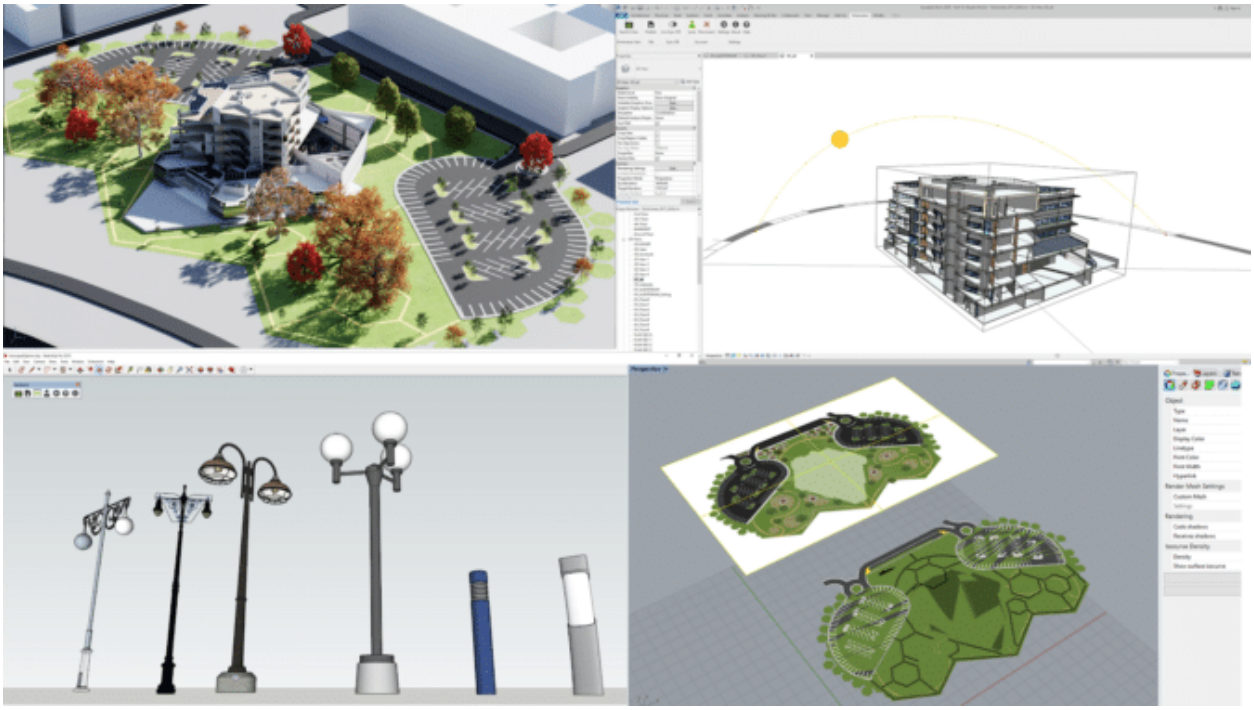
"We've been working with NVIDIA's Omniverse team and our mutual customers in AEC who have been testing and providing real-world feedback as this pioneering platform is being developed," said Amy Bunszel, senior vice president of Design and Creation Products at Autodesk. "This technology can give our customers access to immersive,

interactive and collaborative experiences across industries. So we're thrilled that even more customers will have early access with the expansion of the program announced today at the GTC keynote."

"NVIDIA Omniverse is a new environment that integrates with Rhino and extends our customers' ability to distribute 3D content," said Scott Davidson, vice president of marketing at McNeel & Associates. "We look forward to partnering more with NVIDIA and continuing to support this technology as it develops."

"We are thrilled to partner with the NVIDIA Omniverse team and leverage this technology to increase the pace of collaboration for our customers," said Prakash Iyer, senior vice president of Software Architecture and Strategy at Trimble Inc. "This tool is very promising for SketchUp customers and the AEC industry."

"Teradici is thrilled to enable Omniverse 3D creators across M&E, AEC and manufacturing industries to remotely access the NVIDIA RTX platform through Cloud Access Software, powered by PCoIP technology," said Ziad Lammam, vice president of product management at Teradici.



Real-time collaboration: Revit, Sketchup, Rhino and Omniverse View.

Early Access Customers Provide Real-World Testing

During the GTC keynote, Huang shared how some of our customers across various industries are working with the early access version of Omniverse. Industrial Light & Magic is exploring Omniverse in its advanced visual effect productions.

“Omniverse has really shown its versatility by seamlessly integrating with the typical DCC applications used for visual effects, along with embracing open standards for truly portable assets by leveraging USD and integrating MaterialX with MDL for a collaborative experience that was never possible before,” said Francois Chardavoine, vice president of Technology at Lucasfilm & Industrial Light & Magic. “It’s now easier than ever to leverage Omniverse for high-end visual effects workflows.”

Millennium Falcon used with permission. Rendered with Omniverse View.
NVIDIA has also been working with Lockheed Martin, Foster & Partners and Volvo Cars.

Volvo Cars is testing Omniverse in its research and development workflows. “We immediately saw the opportunity for real-time collaboration for our design workflow using Omniverse,” said Mattias Wikenmalm, senior visualization expert at Volvo Cars. “It’s something we have been striving for throughout our efforts to optimize our design and development process.”

Additional customers are putting Omniverse to the test in areas such as visual effects, AR/VR, manufacturing, architectural design and robotics. Each of these verticals has workflows that rely on collaboration, yet were stuck with the traditional process of exporting and importing data. As USD gains wider adoption across the applications, Omniverse will solve key challenges for creative companies worldwide.

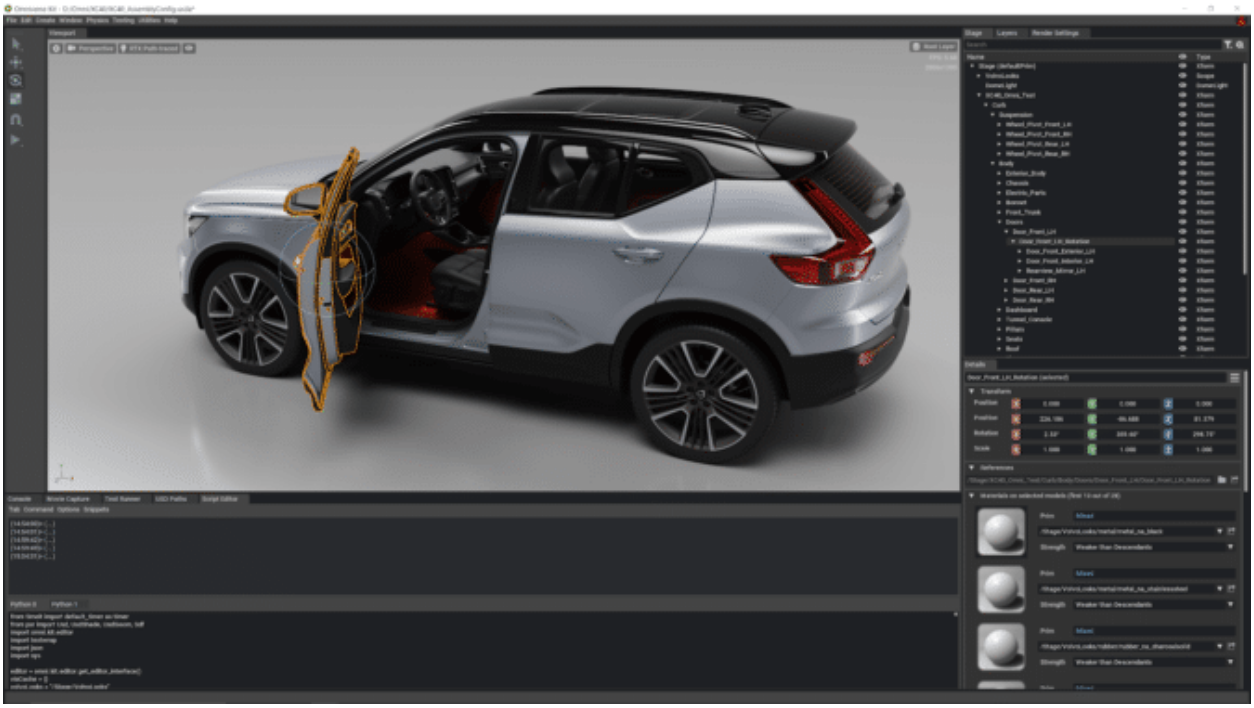


Image courtesy of Volvo Cars.

A New Experience for AEC on RTX Server

The Omniverse early access program is now available to customers that purchase the RTX Server configuration for AEC. They'll have the ability to be a part of the Omniverse AEC Experience program, which provides the full platform with Portal connections to Autodesk Revit, McNeel Rhino and Trimble SketchUp.


NVIDIA RTX Server

What’s Next

Omniverse development will continue with early access customers and ISVs. If you’re interested in integrating your applications into Omniverse or would like to learn more about how it could fit into your production workflow, visit NVIDIA Omniverse Developer Site.

Categories: Pro Graphics

Tags: 3D | NVIDIA RTX | Omniverse

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

NVIDIA CloudXR Cuts the Cord for VR, Raises the Bar for AR

NVIDIA CloudXR 1.0 SDK enables high-fidelity VR and AR streaming across 5G and enterprise networks.

May 14, 2020 by GREG JONES

Share

-

Power up your XR displays and 5G devices because NVIDIA is taking streaming to the next level.

With the announcement today of the NVIDIA CloudXR 1.0 software development kit, we're bringing major advancements to streaming augmented reality, mixed reality and virtual reality content — collectively known as XR — over 5G, Wi-Fi and other high-performance networks.

With the NVIDIA CloudXR platform, any end device — including head-mounted displays (HMDs) and connected Windows and Android devices — can become a high-fidelity XR display capable of showcasing professional-quality graphics.

CloudXR is built on NVIDIA RTX GPUs and the CloudXR SDK to allow streaming of immersive AR, MR or VR experiences from anywhere, whether from the data center, cloud or at the edge. And with NVIDIA GPU virtualization software, CloudXR scales efficiently allowing multiple users or tenants to securely share GPU resources.

Window to an XR World

From architecture to retail, NVIDIA CloudXR is bringing innovation to many industries as 5G networks roll out around the world. By streaming XR experiences from GPU-powered edge servers, companies can expand mobile access to graphics-intensive applications and content, enabling immersive, responsive XR experiences that can be enjoyed on a remote client.

Whether through an HMD, smartphone or tablet, NVIDIA CloudXR accelerates professional XR experiences to power design reviews, speed collaboration and heighten creative productivity.

Additionally, NVIDIA CloudXR early access partners including ZeroLight, VMware, The GRID Factory, Theia Interactive, Luxion KeyVR, ESI Group, PresenZ and PiXYZ have tested CloudXR with a suite of apps and are thrilled with the results. Their customers get access to the highest quality visuals, all from a lightweight mobile XR device.

Tethered or Not, Here XR and 5G Come

The NVIDIA CloudXR SDK is created for telecommunications platforms, enterprise data centers, consumer platforms and next-generation display devices to deliver graphics-rich XR content.

The SDK consists of powerful tools and APIs packaged in three core components:

1. CloudXR Server Driver: Server-side binaries and libraries
2. CloudXR Client App: Operating system-specific sample application
3. CloudXR Client SDK: OS-specific binaries and libraries

CloudXR fits seamlessly with NVIDIA RTX Servers to deliver the richest immersive experiences, and NVIDIA Quadro Virtual Workstation gives the flexibility to enable users to scale with demand.

NVIDIA Partners Bring Mobile XR to the World

NVIDIA is collaborating with Ericsson and Qualcomm Technologies to bring a unique 5G VR solution to market.

Qualcomm Technologies' latest reference design HMD is powered by the Qualcomm Snapdragon XR2 Platform, the world's first 5G-enabled XR device that drives all on-device processing workloads. The high-performance 5G network of Ericsson connects the HMD to the edge by delivering high-speed, low-latency and reliable wireless connectivity.

Combining NVIDIA RTX graphics with CloudXR and GPU virtualization, Qualcomm Technologies' Boundless XR client optimizations and Ericsson's network have yielded an unparalleled ability to deliver boundless XR over 5G.

Learn more about the bundle, which is available now.

Drive XR to the Next Level

Additional information is available for telecommunications providers and developers who register for access to the CloudXR 1.0 SDK.

Categories: Cloud | Mobile | Pro Graphics

Tags: 5G | GTC 2020 | NVIDIA RTX | Quadro | Virtual Reality

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

NVIDIA CEO Introduces NVIDIA Ampere Architecture, NVIDIA A100 GPU in News-Packed 'Kitchen Keynote'

New GPU architecture features as NVIDIA announces major new software applications, new hardware systems and a partnership with BMW.

May 14, 2020 by BRIAN CAULFIELD

Share

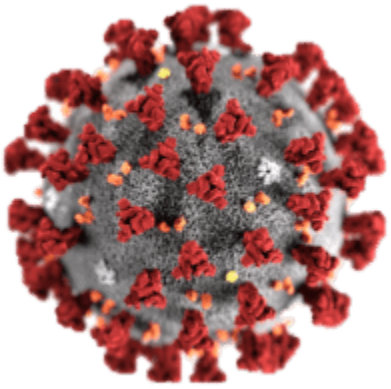
NVIDIA today set out a vision for the next generation of computing that shifts the focus of the global information economy from servers to a new class of powerful, flexible data centers.

In a keynote delivered in nine simultaneously released episodes recorded from the kitchen of his California home, NVIDIA founder and CEO Jensen Huang discussed NVIDIA's recent Mellanox acquisition, new products based on the company's much-awaited NVIDIA Ampere GPU architecture and important new software technologies.

Original plans for the keynote to be delivered live at NVIDIA's GPU Technology Conference in late March in San Jose were upended by the coronavirus pandemic.

Huang kicked off his keynote on a note of gratitude.

"I want to thank all of the brave men and women who are fighting on the front lines against COVID-19," Huang said.



Structura Biotechnology, the University of Texas at Austin and the National Institutes of Health have reconstructed the 3D structure of COVID-19's spike protein.
NVIDIA, Huang explained, is working with researchers and scientists to use GPUs and AI computing to treat, mitigate, contain and track the pandemic. Among those mentioned:

- Oxford Nanopore Technologies has sequenced the virus genome in just seven hours.
- Plotly is doing real-time infection rate tracing.
- Oak Ridge National Laboratory and the Scripps Research Institute have screened a billion potential drug combinations in a day.
- Structura Biotechnology, the University of Texas at Austin and the National Institutes of Health have reconstructed the 3D structure of the virus's spike protein.

NVIDIA also announced updates to its NVIDIA Clara healthcare platform aimed at taking on COVID-19.

"Researchers and scientists applying NVIDIA accelerated computing to save lives is the perfect example of our company's purpose — we build computers to solve problems normal computers cannot," Huang said.

At the core of Huang's talk was a vision for how data centers, the engine rooms of the modern global information economy, are changing, and how NVIDIA and Mellonox, acquired in a deal that closed last month, are together driving those changes.

"The data center is the new computing unit," Huang said, adding that NVIDIA is accelerating performance gains from silicon, to the ways CPUs and GPUs connect, to the full software stack, and, ultimately, across entire data centers.

Systems Optimized for Data Center-Scale Computing

That starts with a new GPU architecture that's optimized for this new kind of data center-scale computing, unifying AI training and inference, and making possible flexible, elastic acceleration.

NVIDIA A100, the first GPU based on the NVIDIA Ampere architecture, providing the greatest generational performance leap of NVIDIA's eight generations of GPUs, is also built for data analytics, scientific computing and cloud graphics, and is in full production and shipping to customers worldwide, Huang announced.

Eighteen of the world's leading service providers and systems builders are incorporating them, among them Alibaba Cloud, Amazon Web Services, Baidu Cloud, Cisco, Dell Technologies, Google Cloud, Hewlett Packard Enterprise, Microsoft Azure and Oracle.

The A100, and the NVIDIA Ampere architecture it's built on, boost performance by up to 20x over its predecessors, Huang said. He detailed five key features of A100, including:

- More than 54 billion transistors, making it the world's largest 7-nanometer processor.
- Third-generation Tensor Cores with TF32, a new math format that accelerates single-precision AI training out of the box. NVIDIA's widely used Tensor Cores are now more flexible, faster and easier to use, Huang explained.
- Structural sparsity acceleration, a new efficiency technique harnessing the inherently sparse nature of AI math for higher performance.
- Multi-instance GPU, or MIG, allowing a single A100 to be partitioned into as many as seven independent GPUs, each with its own resources.
- Third-generation NVLink technology, doubling high-speed connectivity between GPUs, allowing A100 servers to act as one giant GPU.

The result of all this: 6x higher performance than NVIDIA's previous generation Volta architecture for training and 7x higher performance for inference.

NVIDIA DGX A100 Packs 5 Petaflops of Performance

NVIDIA is also shipping a third generation of its NVIDIA DGX AI system based on NVIDIA A100 — the NVIDIA DGX A100 — the world's first 5-petaflops server. And each DGX A100 can be divided into as many as 56 applications, all running independently.

 The U.S. Department of Energy's Argonne National Laboratory will use DGX A100's AI and computing power to better understand and fight COVID-19.
This allows a single server to either "scale up" to race through computationally intensive tasks such as AI training, or "scale out," for AI deployment, or inference, Huang said.

Among initial recipients of the system are the U.S. Department of Energy's Argonne National Laboratory, which will use the cluster's AI and computing power to better understand and fight COVID-19; the University of Florida; and the German Research Center for Artificial Intelligence.

A100 will also be available for cloud and partner server makers as HGX A100.

A data center powered by five DGX A100 systems for AI training and inference running on just 28 kilowatts of power costing $1 million can do the work of a typical data center with 50 DGX-1 systems for AI training and 600 CPU systems consuming 630 kilowatts and costing over $11 million, Huang explained.

"The more you buy, the more you save," Huang said, in his common keynote refrain.

Need more? Huang also announced the next-generation DGX SuperPOD. Powered by 140 DGX A100 systems and Mellanox networking technology, it offers 700 petaflops of AI performance, Huang said, the equivalent of one of the 20 fastest computers in the world.

The next-generation DGX SuperPOD delivers 700 petaflops of AI performance.
NVIDIA is expanding its own data center with four DGX SuperPODs, adding 2.8 exaflops of AI computing power — for a total of 4.6 exaflops of total capacity — to its SATURNV internal supercomputer, making it the world's fastest AI supercomputer.

Huang also announced the NVIDIA EGX A100, bringing powerful real-time cloud-computing capabilities to the edge. Its NVIDIA Ampere architecture GPU offers third-generation Tensor Cores and new security features. Thanks to its NVIDIA Mellanox ConnectX-6 SmartNIC, it also includes secure, lightning-fast networking capabilities.

Software for the Most Important Applications in the World Today

Huang also announced NVIDIA GPUs will power major software applications for accelerating three critical usages: managing big data, creating recommender systems and building real-time, conversational AI.

These new tools arrive as the effectiveness of machine learning has driven companies to collect more and more data. "That positive feedback is causing us to experience an exponential growth in the amount of data that is collected," Huang said.

To help organizations of all kinds keep up, Huang announced support for NVIDIA GPU acceleration on Spark 3.0, describing the big data analytics engine as “one of the most important applications in the world today.”

Built on RAPIDS, Spark 3.0 shatters performance benchmarks for extracting, transforming and loading data, Huang said. It’s already helped Adobe Intelligent Services achieve a 90 percent compute cost reduction.

Key cloud analytics platforms — including Amazon SageMaker, Azure Machine Learning, Databricks, Google Cloud AI and Google Cloud Dataproc — will all accelerate with NVIDIA, Huang announced.

“We’re now prepared for a future where the amount of data will continue to grow exponentially from tens or hundreds of petabytes to exascale and beyond,” Huang said.

Huang also unveiled NVIDIA Merlin, an end-to-end framework for building next-generation recommender systems, which are fast becoming the engine of a more personalized internet. Merlin slashes the time needed to create a recommender system from a 100-terabyte dataset to 20 minutes from four days, Huang said.

And he detailed NVIDIA Jarvis, a new end-to-end platform for creating real-time, multimodal conversational AI that can draw upon the capabilities unleashed by NVIDIA’s AI platform.

Huang highlighted its capabilities with a demo that showed him interacting with a friendly AI, Misty, that understood and responded to a sophisticated series of questions about the weather in real time.

Huang also dug into NVIDIA's swift progress in real-time ray tracing since NVIDIA RTX was launched at SIGGRAPH in 2018, and he announced that NVIDIA Omniverse, which allows "different designers with different tools in different places doing different parts of the same design," to work together simultaneously is now available for early access customers.

Autonomous Vehicles

Autonomous vehicles are one of the greatest computing challenges of our time, Huang said, an area where NVIDIA continues to push forward with NVIDIA DRIVE.

NVIDIA DRIVE will use the new Orin SoC with an embedded NVIDIA Ampere GPU to achieve the energy efficiency and performance to offer a 5-watt ADAS system for the front windshield as well as scale up to a 2,000 TOPS, level-5 robotaxi system.

Now automakers have a single computing architecture and single software stack to build AI into every one of their vehicles.

"It's now possible for a carmaker to develop an entire fleet of cars with one architecture, leveraging the software development across their whole fleet," Huang said.

The NVIDIA DRIVE ecosystem now encompasses cars, trucks, tier one automotive suppliers, next-generation mobility services, startups, mapping services, and simulation.

And Huang announced NVIDIA is adding NVIDIA DRIVE RC for managing entire fleets of autonomous vehicles to its suite of NVIDIA DRIVE technologies.

BMW has selected NVIDIA Isaac robotics to power its factories.

Robotics

NVIDIA also continues to push forward with its NVIDIA Isaac software-defined robotics platform, announcing that BMW has selected NVIDIA Isaac robotics to power its factories.

BMW's 30 factories around the globe build one vehicle every 56 seconds: that's 40 different models, each with hundreds of different options, made from 30 million parts flowing in from nearly 2,000 suppliers around the world, Huang explained.

BMW joins a sprawling NVIDIA robotics global ecosystem that spans delivery services, retail, autonomous mobile robots, agriculture, services, logistics, manufacturing and healthcare.

In the future, factories will, effectively, be enormous robots. "All of the moving parts inside will be driven by artificial intelligence," Huang said. "Every single mass-produced product in the future will be customized."

Categories: Accelerated Analytics | Cloud | Corporate | Data Center | Deep Learning | Networking | Software

Tags: COVID-19 | GPU | Jarvis | Merlin | NVIDIA Ampere Architecture | NVIDIA Clara | NVIDIA DGX | NVIDIA DRIVE | Omniverse | Robotics | SATURNV | TensorRT

LOAD COMMENTS

What's a Recommender System?

Deep learning based recommender systems are driving the growth of online giants; now, with the NVIDIA Merlin application framework and GPU acceleration they're becoming accessible to the rest of us.

May 14, 2020 by BRIAN CAULFIELD

Search and you might find.

Spend enough time online, however, and what you want will start finding you just when you need it.

This is what's driving the internet right now.

They're called recommender systems, and they're among the most important applications today.

That’s because there is an explosion of choice and it’s impossible to explore the large number of available options.

If a shopper were to spend just one second each swiping on their mobile app through the two billion products available on one prominent ecommerce site, it would take 65 years — almost an entire lifetime — to go through their entire catalog.

This is one of the major reasons why the Internet is now so personalized, otherwise it’s simply impossible for the billions of Internet users in the world to connect with the products, services, even expertise — among hundreds of billions of things — that matter to them.

They might be the most human, too. After all, what are you doing when you go to someone for advice? When you’re looking for feedback? You’re asking for a recommendation.

Now, driven by vast quantities of data about the preferences of hundreds of millions of individual users, recommender systems are racing to get better at doing just that.

Sign up now for the NVIDIA Deep Learning Institute workshop.

The internet, of course, already knows a lot of facts: your name, your address, maybe your birthplace. But what the recommender systems seek to learn better, perhaps, than the people who know you are your preferences.

Key to Success of Web's Most Successful Companies

Recommender systems aren't a new idea. Jussi Karlgren formulated the idea of a recommender system, or a "digital bookshelf," in 1990. Over the next two decades researchers at MIT and Bellcore steadily advanced the technique.

The technology really caught the popular imagination starting in 2007, when Netflix — then in the business of renting out DVDs through the mail — kicked off an open competition with a $1 million prize for a collaborative filtering algorithm that could improve on the accuracy of Netflix's own system by more than 10 percent, a prize that was claimed in 2009.

Over the following decade, such recommender systems would become critical to the success of Internet companies such as Netflix, Amazon, Facebook, Baidu and Alibaba.

Virtuous Data Cycle

And the latest generation of deep-learning powered recommender systems provide marketing magic, giving companies the ability to boost click-through rates by better targeting users who will be interested in what they have to offer.

Now the ability to collect this data, process it, use it to train AI models and deploy those models to help you and others find what you want is among the largest competitive advantages possessed by the biggest internet companies.

It's driving a virtuous cycle — with the best technology driving better recommendations, recommendations which draw more customers and, ultimately, let these companies afford even better technology.

That's the business model. So how does this technology work?

Collecting Information

Recommenders work by collecting information — by noting what you ask for — such as what movies you tell your video streaming app you want to see, ratings and reviews you've submitted, purchases you've made, and other actions you've taken in the past

Perhaps more importantly, they can keep track of choices you've made: what you click on and how you navigate. How long you watch a particular movie, for example. Or which ads you click on or which friends you interact with.

All this information is streamed into vast data centers and compiled into complex, multidimensional tables that quickly balloon in size.

They can be hundreds of terabytes large — and they're growing all the time.

That's not so much because vast amounts of data are collected from any one individual, but because a little bit of data is collected from so many.

In other words, these tables are sparse — most of the information most of these services have on most of us for most of these categories is zero.

But, collectively these tables contain a great deal of information on the preferences of a large number of people.

And that helps companies make intelligent decisions about what certain types of users might like.

Content Filtering, Collaborative Filtering

While there are a vast number of recommender algorithms and techniques, most fall into one of two broad categories: collaborative filtering and content filtering.

Collaborative filtering helps you find what you like by looking for users who are similar to you.

So while the recommender system may not know anything about your taste in music, if it knows you and another user share similar taste in books, it might recommend a song to you that it knows this other user already likes.

Content filtering, by contrast, works by understanding the underlying features of each product.

So if a recommender sees you liked the movies "You've Got Mail" and "Sleepless in Seattle," it might recommend another movie to you starring Tom Hanks and Meg Ryan, such as "Joe Versus the Volcano."

Those are extremely simplistic examples, to be sure.

Data as a Competitive Advantage

In reality, because these systems capture so much data, from so many people, and are deployed at such an enormous scale, they're able to drive tens or hundreds of millions of dollars of business with even a small improvement in the system's recommendations.

A business may not know what any one individual will do, but thanks to the law of large numbers, they know that, say, if an offer is presented to 1 million people, 1 percent will take it.

But while the potential benefits from better recommendation systems are big, so are the challenges.

Successful internet companies, for example, need to process ever more queries, faster, spending vast sums on infrastructure to keep up as the amount of data they process continues to swell.

Companies outside of technology, by contrast, need access to ready-made tools so they don't have to hire whole teams of data scientists.

If recommenders are going to be used in industries ranging from healthcare to financial services, they'll need to become more accessible.

GPU Acceleration

This is where GPUs come in.

NVIDIA GPUs, of course, have long been used to accelerate training times for neural networks — sparking the modern AI boom — since their parallel processing capabilities let them blast through data-intensive tasks.

But now, as the amount of data being moved continues to grow, GPUs are being harnessed more extensively. Tools such as RAPIDS, a suite of software libraries for accelerating data science and analytics pipelines much more quickly, so data scientists can get more work done much faster.

And NVIDIA's just announced Merlin recommender application framework promises to make GPU-accelerated recommender systems more accessible still with an end-to-end pipeline for ingesting, training and deploying GPU-accelerated recommender systems.

These systems will be able to take advantage of the new NVIDIA A100 GPU, built on our NVIDIA Ampere architecture, so companies can build recommender systems more quickly and economically than ever.

Our Recommendation? Learn How to Build Intelligent Recommendation Systems

The NVIDIA Deep Learning Institute offers instructor-led, hands-on training on the fundamental tools and techniques for building highly effective recommender systems. Taught by an expert, this in-depth, 8-hour-long workshop instructs participants in how to:

- Build a content-based recommender system using the open-source cuDF library and Apache Arrow
- Construct a collaborative filtering recommender system using alternating least squares and CuPy
- Design a wide and deep neural network using TensorFlow 2 to create a hybrid recommender system
- Optimize performance for training and inference using large, sparse datasets
- Deploy a recommender model as a high-performance web service

Earn a DLI certificate to demonstrate subject-matter competency and accelerate your career growth. Take this workshop at GTC or request a workshop for your organization.

Read more about NVIDIA Merlin, NVIDIA's application framework for deep recommender systems.

Featured image credit: © Monkey Business – stock.adobe.com.

Categories: Corporate | Deep Learning | Explainer | Software

Tags: Artificial Intelligence | GTC 2020 | Merlin

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics
Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight
Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program
The Great AI Bake-Off: Recommendation Systems on the Rise
Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

Artemis Supercomputer on the Hunt for Deeper Understanding of Genomics

Group 42's supercomputer, powered by NVIDIA DGX systems, fuels national genome program to enhance

understanding of UAE citizens' genomes, improve healthcare and fight COVID-19.

May 14, 2020 by MARC DOMENECH

Share

Ever wondered what you're really made of?

Your genome is a unique genetic code that determines your characteristics. It's a specific combination of DNA molecules that makes you *you*.

Studying the entire genetic code of an individual or a group of individuals can help us gain a better understanding of diseases, enable precision medicine and power pharmacogenomics — how genes affect a person's response to drugs.

As part of a national project launched by Abu Dhabi's Department of Health, Group 42 is harnessing its Artemis supercomputer to decode the human genome and improve patient care. Powered by NVIDIA GPUs, Artemis is the 26th fastest system in the world.

G42, based in Abu Dhabi, develops and deploys holistic and scalable AI and cloud computing offerings. Through its Inception Institute of Artificial Intelligence, it carries out fundamental research on AI, big data and machine learning.

Building a world-class AI supercomputer normally takes six months or longer. In just three weeks, G42 designed, built and deployed Artemis with NVIDIA, using the DGX SuperPOD reference architecture and Mellanox AI networking fabric.

The Population Genome Program

Built with 81 NVIDIA DGX systems, Artemis can deliver a total of 7.2 petaflops of double-precision HPL performance and run workloads 120x faster than G42's previous system.

Now the supercomputer is being put to work on the Population Genome Program. This national effort aims to enhance scientific understanding of Abu Dhabi citizens' genomes and improve healthcare in the country.

Till now, the understanding of genetic variation in the Arab population has been a challenge due to the lack of a high-quality Emirati reference genome. The Population Genome Program will enrich available data by producing a reference genome specific to citizens of the United Arab Emirates.

The program aims to be the first of its kind in the world to then use this as a baseline and incorporate the genomic data into healthcare management processes.

"Embracing innovation and providing a comprehensive healthcare programme in the Emirate of Abu Dhabi remains at the forefront of our priorities. Two of the world's most exciting technologies — DNA sequencing and AI — will come together in this project," explained H.E. Sheikh Abdulla Bin Mohamed Al Hamed, Chairman of Department of Health-Abu Dhabi, in a press statement.

Accelerating Processing of Genomic Data

In the first phase of the program, the genomes of 10,000 individuals are set to be tested. To ensure the highest throughput and accurate analysis, both short-read and long-read genome sequencing platforms will be used, leveraging G42's collaboration with BGI and Oxford Nanopore — two global genome sequencing leaders.

Anonymized DNA samples will first be collected and processed using Oxford Nanopore PromethION sequencers. These devices, which contain embedded NVIDIA GPU technology to enable AI at the edge, will help to accelerate the processing of genomic data.

The processed data will be supplied, in a graphical format, to Artemis for AI-powered analysis, with support from NVIDIA Parabricks software to support their population analysis.

The final results will be provided to the research and medical community to help deliver more effective patient care. This could include more advanced treatments for conditions such as cancer, schizophrenia, autism, and cardiovascular and neuronal diseases.

"With NVIDIA's GPU technology we're able to provide a highly optimized AI platform for the national Population Genome Program and accelerate data processing," said Min S. Park, director of Genome Programs at G42. "This collaboration supports our goals of

developing a program for personalized care across the UAE, bringing experts, data and technology together for improving patient care.”

Combatting COVID-19

G42 is also using its supercomputing prowess in the battle against COVID-19, having recently established a new detection laboratory in Masdar City, Abu Dhabi. This facility can, on a daily basis, support tens of thousands of real-time reverse transcription polymerase chain reaction (RT-PCR) tests. These tests detect the presence of the SARS-CoV-2 virus in samples taken from patients.

In addition, G42 is involved in the production of COVID-19 diagnostic kits, the supply of thermal sensors and, working in coordination with local and international health authorities, assisting in the creation of effective prevention and detection protocols to contain the virus.

“Technology will play a crucial role in curbing the spread of the coronavirus and the superior computing capability of Artemis can help in many ways — from rapid vaccine development, where computer simulations may replace manual experiments and reduce the development time of a vaccine, to mapping and predicting trends in the outbreak, as well as predicting virus mutations,” said Peng Xiao, CEO of G42.

Image credit: Gerd Altmann

Categories: Deep Learning | Supercomputing

Tags: Artificial Intelligence | COVID-19 | GTC 2020 | Medical Research and Healthcare | NVIDIA DGX | Social Impact

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics
Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight
Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program
The Great AI Bake-Off: Recommendation Systems on the Rise
Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

Double-Precision Tensor Cores Speed High-Performance Computing

Simulations and iterative solvers get FP64 math boosts of up to 2.5x with the NVIDIA Ampere architecture.

May 14, 2020 by GEETIKA GUPTA

Share

What you can see, you can understand.

Simulations help us understand the mysteries of black holes and see how a protein spike on the coronavirus causes COVID-19. They also let designers create everything from sleek cars to jet engines.

But simulations are also among the most demanding computer applications on the planet because they require lots of the most advanced math.

Simulations make numeric models visual with calculations that use a double-precision floating-point format called FP64. Each number in the format takes up 64 bits inside a computer, making it one the most computationally intensive of the many math formats today's GPUs support.

As the next big step in our efforts to accelerate high performance computing, the NVIDIA Ampere architecture defines third-generation Tensor Cores that accelerate FP64 math by 2.5x compared to last-generation GPUs.

That means simulations that kept researchers and designers waiting overnight can be viewed in a few hours when run on the latest A100 GPUs.

Science Puts AI in the Loop

The speed gains open a door for combining AI with simulations and experiments, creating a positive-feedback loop that saves time.

First, a simulation creates a dataset that trains an AI model. Then the AI and simulation models run together, feeding off each other's strengths until the AI model is ready to deliver real-time results through inference. The trained AI model also can take in data from an experiment or sensor, further refining its insights.

Using this technique, AI can define a few areas of interest for conducting high-resolution simulations. By narrowing the field, AI can slash by orders of magnitude the need for thousands of time-consuming simulations. And the simulations that need to be run will run 2.5x faster on an A100 GPU.

With FP64 and other new features, the A100 GPUs based on the NVIDIA Ampere architecture become a flexible platform for simulations, as well as AI inference and training — the entire workflow for modern HPC. That capability will drive developers to migrate simulation codes to the A100.

Users can call new CUDA-X libraries to access FP64 acceleration in the A100. Under the hood, these GPUs are packed with third-generation Tensor Cores that support DMMA, a new mode that accelerates double-precision matrix multiply-accumulate operations.

Accelerating Matrix Math

A single DMMA job uses one computer instruction to replace eight traditional FP64 instructions. As a result, the A100 crunches FP64 math faster than other chips with less work, saving not only time and power but precious memory and I/O bandwidth as well.

We refer to this new capability as Double-Precision Tensor Cores. It delivers the power of Tensor Cores to HPC applications, accelerating matrix math in full FP64 precision.

Beyond simulations, HPC apps called iterative solvers — algorithms with repetitive matrix-math calculations — will benefit from this new capability. These apps include a wide range of jobs in earth science, fluid dynamics, healthcare, material science and nuclear energy as well as oil and gas exploration.

To serve the world's most demanding applications, Double-Precision Tensor Cores arrive inside the largest and most powerful GPU we've ever made. The A100 also packs more memory and bandwidth than any GPU on the planet.

The third-generation Tensor Cores in the NVIDIA Ampere architecture are beefier than prior versions. They support a larger matrix size — 8x8x4, compared to 4x4x4 for Volta — that lets users tackle tougher problems.

That's one reason why an A100 with a total of 432 Tensor Cores delivers up to 19.5 FP64 TFLOPS, more than double the performance of a Volta V100.

Where to Go to Learn More

To get the big picture on the role of FP64 in our latest GPUs, watch the keynote with NVIDIA founder and CEO Jensen Huang. To learn more, register for the webinar or read a detailed article that takes a deep dive into the NVIDIA Ampere architecture.

Double-Precision Tensor Cores are among a battery of new capabilities in the NVIDIA Ampere architecture, driving HPC performance as well as AI training and inference to new heights. For more details, check out our blogs on:

- Multi-Instance GPU (MIG), supporting up to 7x in GPU productivity gains.
- TensorFloat-32 (TF32), a format, speeding up AI training and certain HPC jobs up to 20x.
- Our support for sparsity, accelerating math throughput 2x for AI inference.
- Or, see the web page describing the A100 GPU.

Categories: Explainer

Tags: GPU Computing | GTC 2020 | High Performance Computing | New GPU Uses | NVIDIA Ampere Architecture | Supercomputing

LOAD COMMENTS

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

Word Power: Conversational AI Rewrites the Textbook on Success

Across industries, companies are parlaying text and speech into business results with deep learning.

May 14, 2020 by SID SHARMA

When startup Kensho was acquired by S&P Global for $550 million in March 2018, Georg Kucsko felt like a kid in a candy store.

The head of AI research at Kensho and his team had one of Willy Wonka's golden tickets dropped into their laps: S&P's 100,000 hours of recorded and painstakingly transcribed audio files.

The dataset helped Kensho build Scribe, considered the most accurate voice recognition software in the finance industry. It transcribes earning calls and other business meetings fast and at low cost, helping extend S&P's coverage by 1,500 companies and earning kudos from the company's CEO in his own quarterly calls.

“We used these transcripts to train speech-recognition models that could do the work faster — that was a new angle no one had thought of. It allowed us to drastically improve the process,” said Kucsko.

It’s one example among many of the power of conversational AI.

What the Buzz Is All About

There are lots of reasons why conversational AI is the talk of the town.

It can turn speech into text that’s searchable. It morphs text into speech you can listen to hands-free while working or driving.

As it gets smarter, it’s understanding more of what it hears and reads, making it even more useful. That’s why the word is spreading fast.

Conversational AI is perhaps best known as the language of Siri and Alexa, but high-profile virtual assistants share the stage with a growing chorus of agents.

Businesses are using the technology to manage contracts. Doctors use it to take notes during patient exams. And a laundry list of companies are tapping it to improve customer support.

Covering the Waterfront of Words

“There is a huge surface area of conversations between buyers and sellers that we can and should help people navigate,” said Gabor Angeli, an expert in conversational AI at Square Inc., who described his company’s work in a session at GTC Digital.

Deloitte uses conversational AI in its dTrax software that helps companies manage complex contracts. For instance, dTrax can find and update key passages in lengthy agreements when regulations change or when companies are planning a big acquisition. The software, which runs on NVIDIA GPUs, won a smart-business award from the Financial Times in 2019.

China’s largest insurer, Ping An, already uses conversational AI to sell insurance. It’s a performance-hungry application that runs on GPUs because it requires a lot of intelligence to gauge a speaker’s mood and emotion.

In healthcare, Nuance provides conversational AI software, trained with NVIDIA GPUs and software, that most radiologists use to make transcriptions and many other doctors use to document patient exams.

Voca.ai deploys AI models on NVIDIA GPUs because they slash latency on inference jobs in half compared to CPUs. That's key for its service that automates responses to customer support calls from as many as 10 million people a month for one of its largest users.

Framing the Automated Conversation

The technology is built on a broad software foundation of conversational AI libraries, all accelerated by GPUs. The most popular ones get lots of "stars" on the GitHub repository, the equivalent of "likes" on Facebook or bookmarks in a browser. They include:

- Huggingface, 26.1k stars
- Fast.ai, 17.8k stars
- spaCy, 16.3k stars
- Kaldi, 8.7k stars
- DeepPavlov, 4.2k stars
- ESPnet, 2.2k stars

To make it easier to get started in conversational AI, NVIDIA provides a growing set of software tools, too.

Kensho and Voca.ai already use NVIDIA NeMo to build state-of-the-art conversational AI algorithms. These machine- and deep-learning models can be fine-tuned on any company's data to deliver the best accuracy for its particular use case.

When NVIDIA announced NeMo last fall, it also released Jasper, a 54-layer model for automatic speech recognition that can lower word error rates to less than 3 percent. It's one of several models optimized for accuracy, available from NGC, NVIDIA's catalog for GPU-accelerated software.

Say Hello to Jarvis, the Valet of Conversational AI

Today we're rolling out NVIDIA Jarvis, an application framework for building and deploying AI services that fuse vision, speech and language understanding. The services can be deployed in the cloud, in the data center or at the edge.

Jarvis includes deep-learning models for building GPU-accelerated conversational AI applications capable of understanding terminology unique to each company and its customers. It includes NeMo to train these models on specific domains and customer data. The models can take advantage of TensorRT to minimize latency and maximize throughput in AI inference tasks.

Jarvis services can run in 150 milliseconds on an A100 GPU. That's far below the 300ms threshold for real-time application and the 25 seconds it would take to run the same models on a CPU.

Jarvis Is Ready to Serve Today

Kensho is already testing some of the tools in Jarvis.

"We are using NeMo a lot, and we like it quite a lot," said Kucsko. "Insights from NVIDIA, even using different datasets for training at scale, made crucial insights for us," he said.

For Kensho, using such tools is a natural next step in tuning the AI models inside Scribe. When Kensho was developing the original software, NVIDIA helped train those models on one of its DGX SuperPOD systems.

"We had the data and they had the GPUs, and that led to an amazing partnership with our two labs collaborating," Kucsko said.

"NVIDIA GPUs are indispensable for deep learning work like that. For anything large scale in deep learning, there's pretty much not another option," he added.

Categories: Deep Learning

Tags: Artificial Intelligence | GTC 2020 | Inference | Jarvis | Machine Learning | New GPU Uses | NVIDIA DGX | TensorRT

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics
Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight
Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program
The Great AI Bake-Off: Recommendation Systems on the Rise
Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

NVIDIA DRIVE Gets Amped: Scalable Platform Moves to NVIDIA Ampere Architecture

Automakers can now deploy a single scalable architecture and single software stack achieving up to 2,000 TOPS.

May 14, 2020 by DANNY SHAPIRO

Share

With the introduction of the NVIDIA Ampere architecture, the NVIDIA DRIVE platform is expanding driving capabilities from an entry-level ADAS solution all the way to a level 5 robotaxi system.

In his virtual GTC keynote, NVIDIA founder and CEO Jensen Huang announced the expansion of the DRIVE AGX platform, leveraging new variants of the upcoming Orin system-on-a-chip (SoC) and new NVIDIA Ampere GPUs. With a single architecture, manufacturers can deploy a high-performance AI system to make every vehicle in their lineup software-defined.

This newly expanded range starts at an NCAP 5-star ADAS system and runs all the way to a DRIVE AGX Pegasus robotaxi platform. The latter features two Orin SoCs and two NVIDIA Ampere GPUs to achieve an unprecedented 2,000 trillion operations per second, or TOPS — more than 6x the performance of the previous platform.

The current generation of DRIVE AGX delivers capabilities that scale from level 2+ automated driving to level 5 fully autonomous driving with different combinations of Xavier SoCs and Turing-based GPUs. DRIVE AGX Xavier delivers 30 TOPS of performance and the NVIDIA DRIVE AGX Pegasus platform processes up to 320 TOPS to run multiple redundant and diverse deep neural networks for real-time perception, planning and control.

With the introduction of the NVIDIA Ampere GPU, and the upcoming Orin processor family featuring its integrated cores, we're able to deliver compute for everything that moves, raising the DRIVE platform even higher while adding an entry-level ADAS offering.

Single Scalable Architecture

Based on customer requests, the new DRIVE AGX family now begins with a single Orin SoC variant that sips just five watts of energy and delivers 10 TOPS of performance.

Automakers have generally developed one computer system for ADAS systems and a different one for higher levels of automated driving, however, the development of multiple systems has become cost prohibitive.

With a single platform, developers can leverage one architecture to more easily develop autonomous driving technology across all their market segments. And since the DRIVE platform is software-defined and based on the large CUDA developer community, it can easily and constantly benefit from over-the-air updates.

Driving Performance Further

While the DRIVE AGX family is extending to entry levels of autonomy, the NVIDIA Ampere architecture is pushing performance even higher with the next-generation Pegasus robotaxi platform.

With two Orin SoCs and two NVIDIA Ampere-based GPUs delivering 2,000 TOPS, the platform is capable of handling higher resolution sensor inputs and more advanced autonomous driving DNNs required for full self-driving robotaxi operation.

The architecture offers the largest leap in performance within the eight generations of NVIDIA GPUs — boosting performance by up to 6x.

The Orin family of SoCs will begin sampling next year and be available for automakers starting production in late 2022, laying the foundation for the next-generation of the programmable, software-defined NVIDIA DRIVE AGX lineup.

Categories: Driving

Tags: Automotive | GTC 2020 | NVIDIA Ampere Architecture | NVIDIA DRIVE

ALL NVIDIA NEWS

Perfect Pairing: NVIDIA's David Luebke on the Intersection of AI and Graphics

Vision of AI: Startup Helps Diabetic Retinopathy Patients Retain Their Sight

Scaling New Heights: Surge in Remote Work Fuels NVIDIA Cloud Service Provider Program

The Great AI Bake-Off: Recommendation Systems on the Rise

Office Ready? Jetson-Driven 'Double Robot' Supports Remote Working

Post navigation

- CUDA-X
- Autonomous Machines
- Data Center
- Deep Learning and AI
- Design and Visualization
- Healthcare
- High Performance Computing
- Self-Driving Cars
- Gaming & Entertainment
- NGC

- DGX Systems
- DRIVE PX
- GeForce RTX 20-Series
- NVIDIA Virtual GPU
- Jetson
- Quadro
- SHIELD TV
- Tesla
- T4 Enterprise Server

- NVIDIA Developer
- Developer News
- Developer Blog
- Developer Forums
- Open Source Portal
- Training
- GPU Tech Conference
- CUDA

- NVIDIA Partner Network
- Careers
- Contact Us
- Security
- Communities
- RSS Feeds
- Email Signup
- Privacy Center
- Share Your Story Idea

EXPLORE OUR REGIONAL BLOGS AND OTHER SOCIAL NETWORKS

- Legal Info

- Contact Us

USA - United States