NVIDIA GTC 2021

HIGHLIGHTS

> “I'M STUNNED BY THE BOLDNESS OF WHAT NVIDIA REVEALED”

MAD MONEY

“NVIDIA’S GTC: THE NEAR-TERM FUTURE OF ADVANCED AI”

DATAMATION

210,000
REGISTRATIONS

14,000,000
KEYNOTE VIEWS

6,000
PRESS ARTICLES

1,600
TALKS

900
TOP UNIVERSITIES

110
PARTNER SPONSORSHIPS

"NVIDIA HAS EVOLVED INTO A COMPANY THAT PROVIDES THE CORE TECHNOLOGY RESHAPING INDUSTRY AND SOCIETY"

DIGINOMICA

12 of 12

TOP IT COMPANIES

15 of 15

TOP E-COMMERCE COMPANIES

18 of 20

TOP CAR MAKERS

17 of 20

TOP TELCOS

9 of 10

TOP AEROSPACE & DEFENSE COMPANIES

10 of 10

TOP PHARMACEUTICAL COMPANIES

"WORLD-LEADING AI RESEARCH AT THE FOREFRONT OF THIS YEAR'S NVIDIA GTC"

VENTURE BEAT

Yoshua Bengio
University of Montreal
Quebec AI Institute

Yann LeCun
Facebook
New York University

Geoffrey Hinton
University of Toronto
Google
Vector Institute

Daphne Koller
Insitro
Coursera
Stanford

Jürgen Schmidhuber
Dalle Molle Institute for
AI Research

Raquel Urtasun
University of Toronto

Alvy Ray Smith
Pixar
Altamira

Abhay Parasnis
Adobe

Kim Libreri
Epic Games

Rommie Amaro
University of
California, San Diego

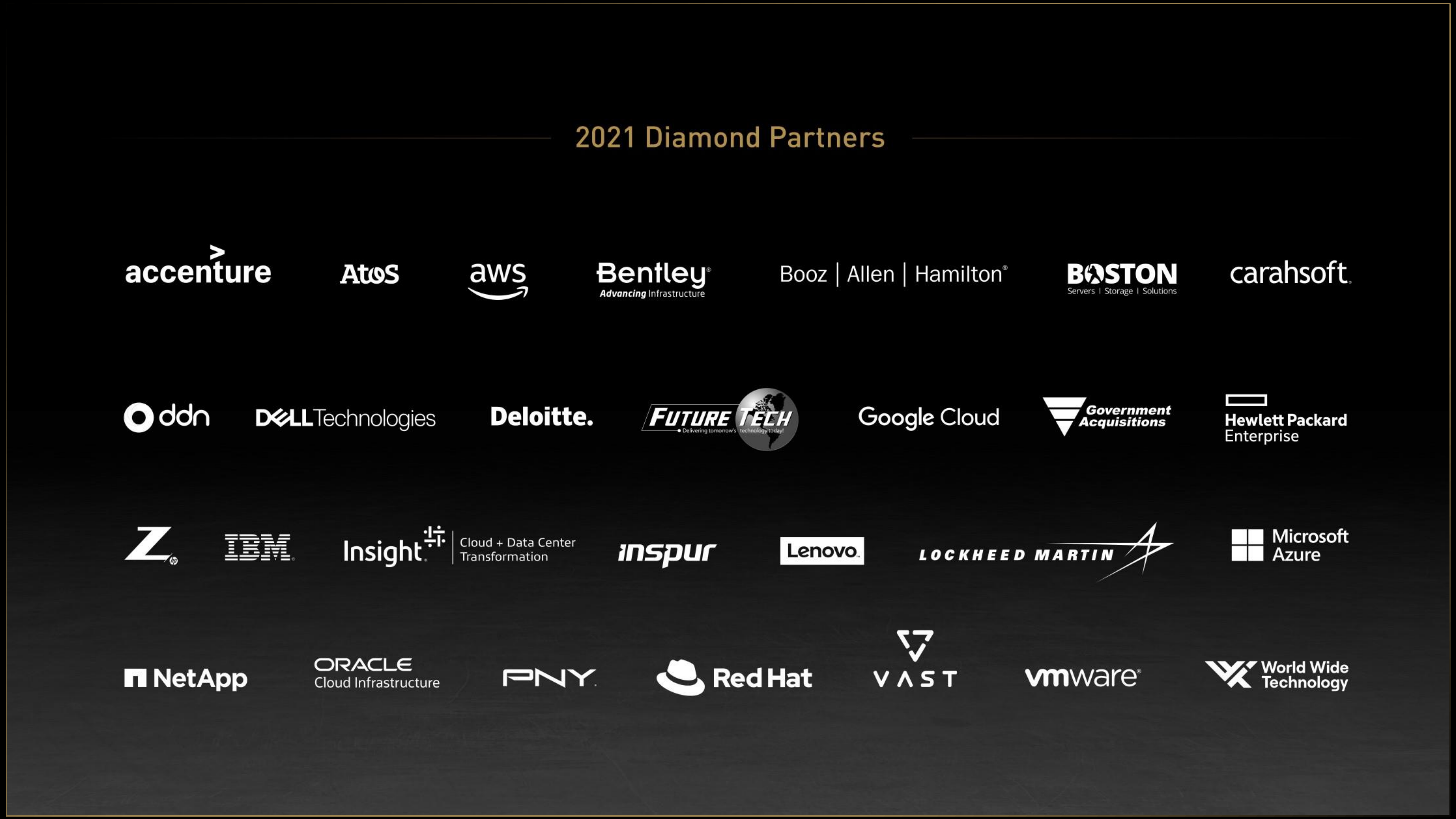
Soumith Chintala
Facebook

Rose Yu
University of
California, San Diego

Scientists, researchers, developers, and creators are using NVIDIA to do amazing things. It's a community of more than 2.5 million developers and 7,500 startups creating thousands of applications for accelerated computing. GTC is where we celebrate and learn from their works, extend the frontiers of computing, and discover the future together.

“GTC IS WHERE THE LARGEST SHOWCASE OF THE AI ECOSYSTEM IS ON DISPLAY”

ENDERLE GROUP



2021 Diamond Partners

accenture · AtoS · aws · Bentley Advancing Infrastructure · Booz | Allen | Hamilton · BOSTON Servers | Storage | Solutions · carahsoft.

ddn · DELL Technologies · Deloitte. · FUTURE TECH · Google Cloud · Government Acquisitions · Hewlett Packard Enterprise

Z · IBM · Insight Cloud + Data Center Transformation · inspur · Lenovo · LOCKHEED MARTIN · Microsoft Azure

NetApp · ORACLE Cloud Infrastructure · PNY · Red Hat · VAST · vmware · World Wide Technology

GTC brings together a vast ecosystem of the world’s most important and innovative technology companies. Our more than 100 sponsors include the largest cloud and IT services providers, system builders, enterprise software companies, and the hottest startups.

"NVIDIA IS DEVELOPING THE DATACENTER ECOSYSTEM OF CHOICE FOR THE 4TH TECTONIC SHIFT IN COMPUTING"

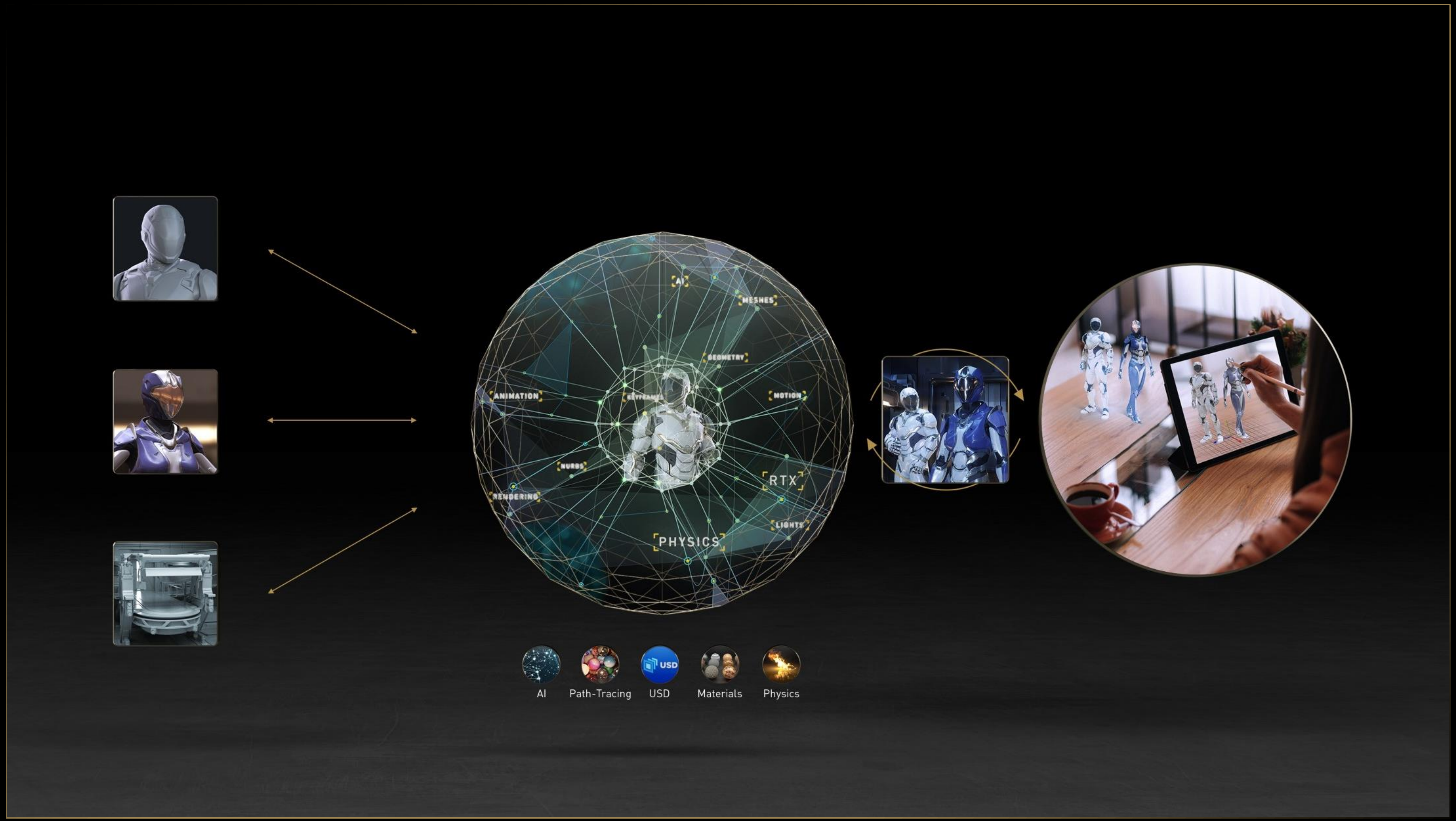
JEFFERIES

Accelerated Computing is the Path Forward
AI is Software that Writes Software
Data Center is the New Unit of Computing
AI-on-5G Kickstarts the 4th Industrial Revolution
Autonomous Systems in Real and Virtual Worlds

There are powerful forces shaping the world's industries. Accelerated computing that we pioneered has supercharged scientific discovery, while providing the computer industry a path forward. AI has seen incredible advances. With NVIDIA GPUs, computers learn and software writes software that no human can. Software is now composed of microservices that scale across the data center — treating the data center as a single-unit of computing. 5G and AI will kick-start the 4th industrial revolution. And "the metaverse" will deliver a virtual world that is a digital twin of ours.

“OMNIVERSE ALLOWS THE ENTIRE PRODUCTION PROCESS TO BE SIMULATED WITH PHOTO-REALISTIC DETAILS, AND WITH PHYSICAL PROPERTIES”

WIRED



Omniverse is a platform to create and simulate shared virtual worlds. It’s made up three major parts: Omniverse Nucleus connects designers so they can collaborate in real time on 3D production pipeline elements like modeling, layout, shading, animation, lighting, special effects, and rendering. The second part of Omniverse is the composition, rendering, and animation engine that simulates the virtual world with accurate physics and materials, and path-traced graphics. And with NVIDIA CloudXR, people can portal into Omniverse with VR, and AIs can teleport out with AR.

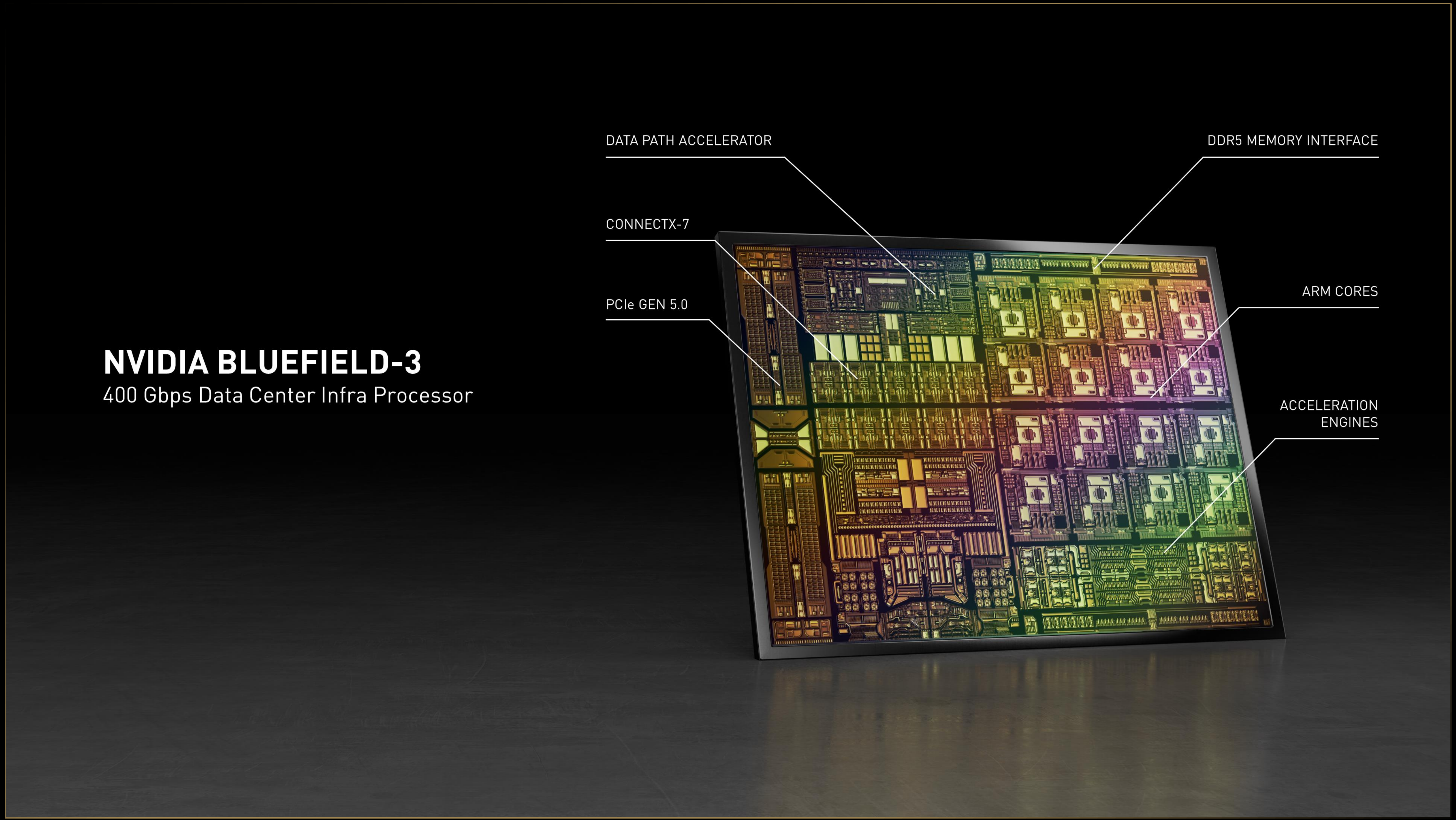
“BMW’S VIRTUAL FACTORY USES AI TO HONE THE ASSEMBLY LINE”

WIRED

BMW produces over 2 million cars a year, most of them from custom orders. How BMW builds its cars is the company’s core technology. BMW is creating its future factory in Omniverse. It’s designed completely in digital and simulated from beginning to end. Using Omniverse, BMW has created a digital twin factory where robots and humans can work together to build the ultimate driving machine.

“NVIDIA SCALES THE CLOUD WITH A REIMAGINED DATA PROCESSING UNIT”

ALL ABOUT CIRCUITS



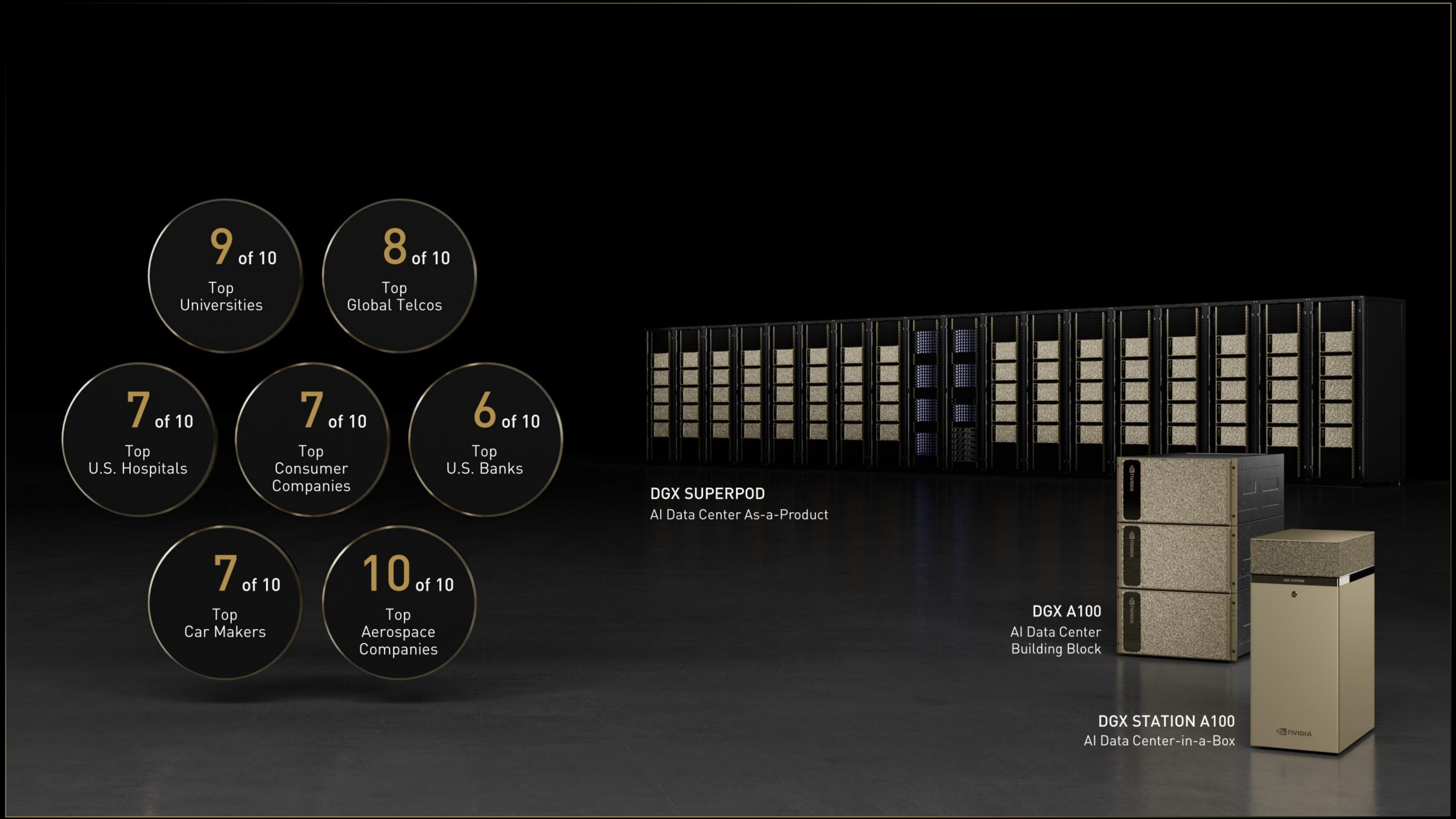
As data centers take on more work, a mountain of infrastructure software continues to grow. Applications and services should be the largest workloads in the data center — not infrastructure. The answer is a new type of chip for data center infrastructure processing.

The NVIDIA BlueField-2 DPU can isolate the infrastructure from the applications, and offload and accelerate the networking, storage, and security. And the NVIDIA DOCA SDK simplifies offloading to BlueField’s accelerators and programmable engines.

We’re just getting started with BlueField-2, but at GTC we announced BlueField-3, the world’s first 400 Gbps networking chip.

“NVIDIA CONTINUES TO EXPAND ITS LINE OF DGX APPLIANCES AND CLUSTERS FOR AI COMPUTING”

CRN



At the beginning of the big bang of modern AI, we recognized the need to create a new kind of computer for a new way of developing software.

This computer will need new chips, new system architecture, new ways to network, new software, and new methodologies and tools. It all comes together as DGX — a computer for AI.

The DGX A100 is a building block that contains 5 petaflops of computing and super-fast storage and networking to feed it. DGX Station is an AI data-center-in-a-box designed for workgroups. And for intensive AI research and development, DGX SuperPOD is a fully integrated, fully network-optimized, AI-data-center-as-a-product.

“THE DAY HAS COME WHEN YOU CAN RENT YOUR OWN MINI SUPERCOMPUTER”

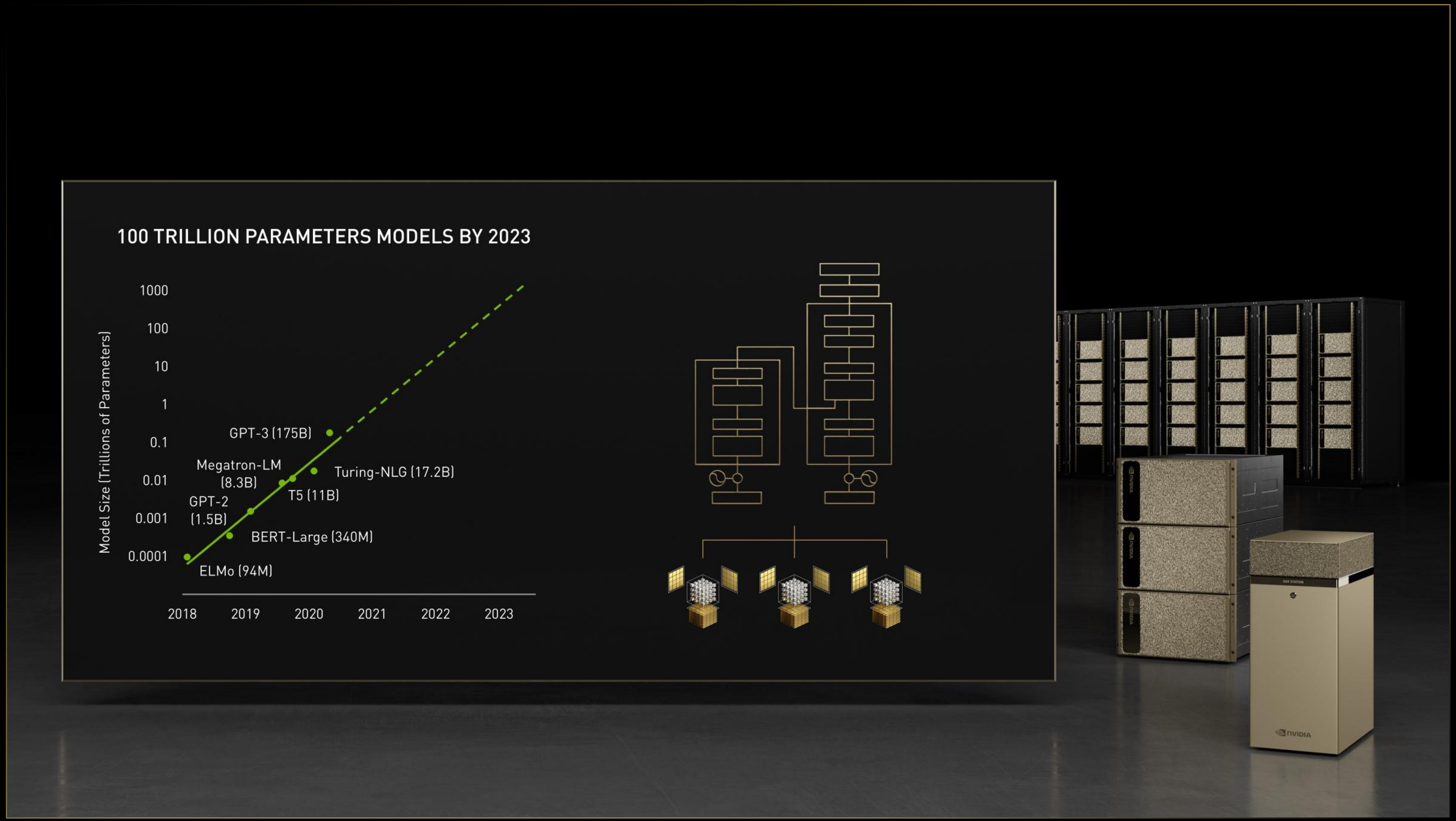
VENTURE BEAT

The new DGX Station 320G can train large models at lightning speed — 320GB of super-fast HBM2e memory connected to four A100 GPUs delivers over 8TB per second of memory bandwidth. It would take 40 CPU servers to achieve this.

DGX Station plugs into a normal wall outlet like a big gaming rig, consumes just 1,500 watts, and is liquid-cooled to a silent 37 decibels.

“NVIDIA IS AGAIN DISRUPTING THE AI AND ACCELERATION MARKETS TO KEEP PACE WITH AI MODEL COMPLEXITY DOUBLING EVERY 2.5 MONTHS”

ROSENBLATT

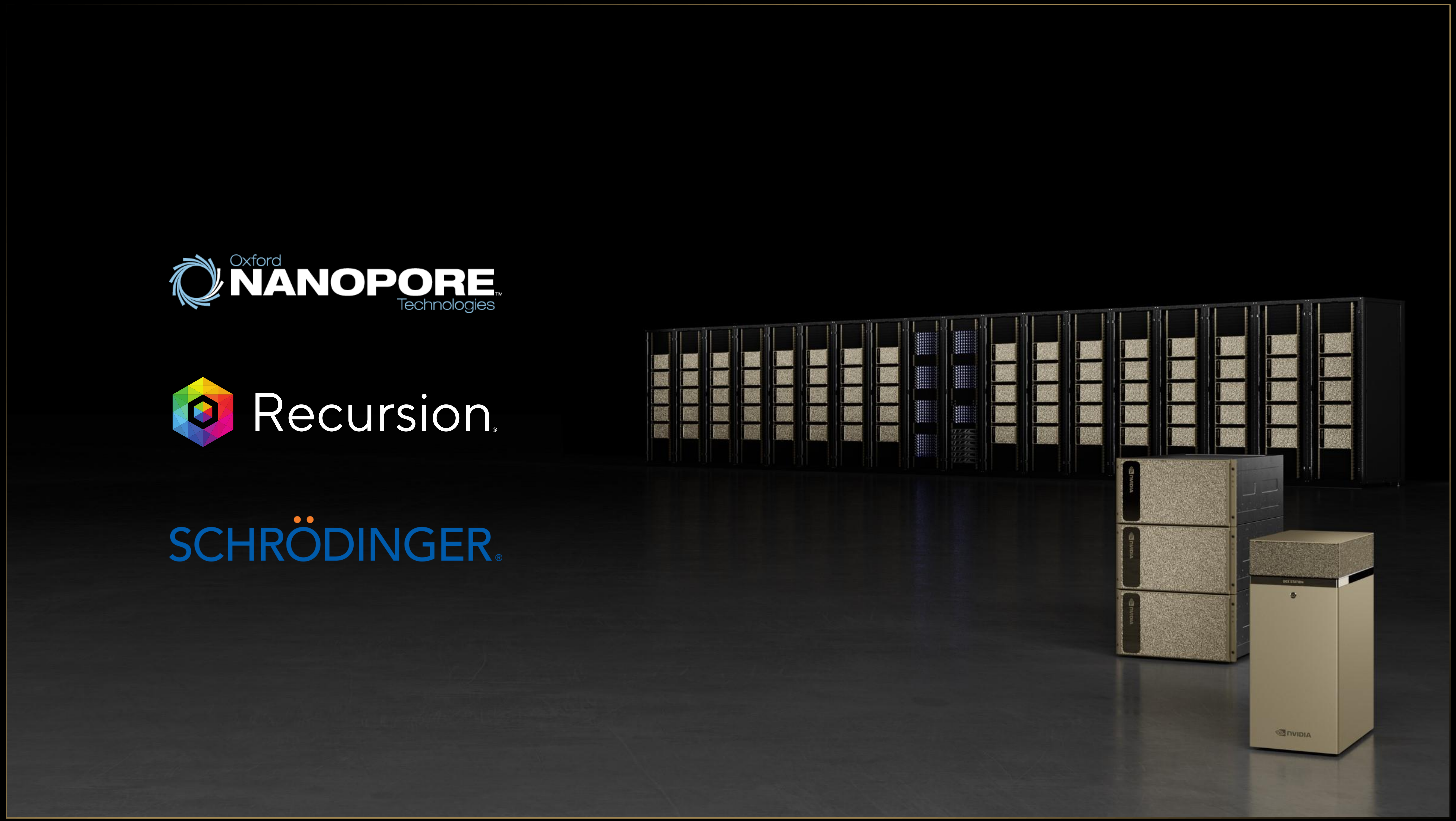


Transformers are incredibly large and ever-expanding models that have led to dramatic breakthroughs in natural language processing. NVIDIA Megatron trains giant Transformer models in parallel, dramatically speeding training time.

When it comes to inference, it’s equally powerful. On OpenAI’s giant GPT-3 model, it takes a dual-CPU server over a minute to respond to a single 128-word query. A DGX with Megatron Triton will respond within a second. And for 16 queries at the same time.

"NVIDIA AIMS CLARA HEALTHCARE AT DRUG DISCOVERY, IMAGING VIA DGX"

HPCWIRE

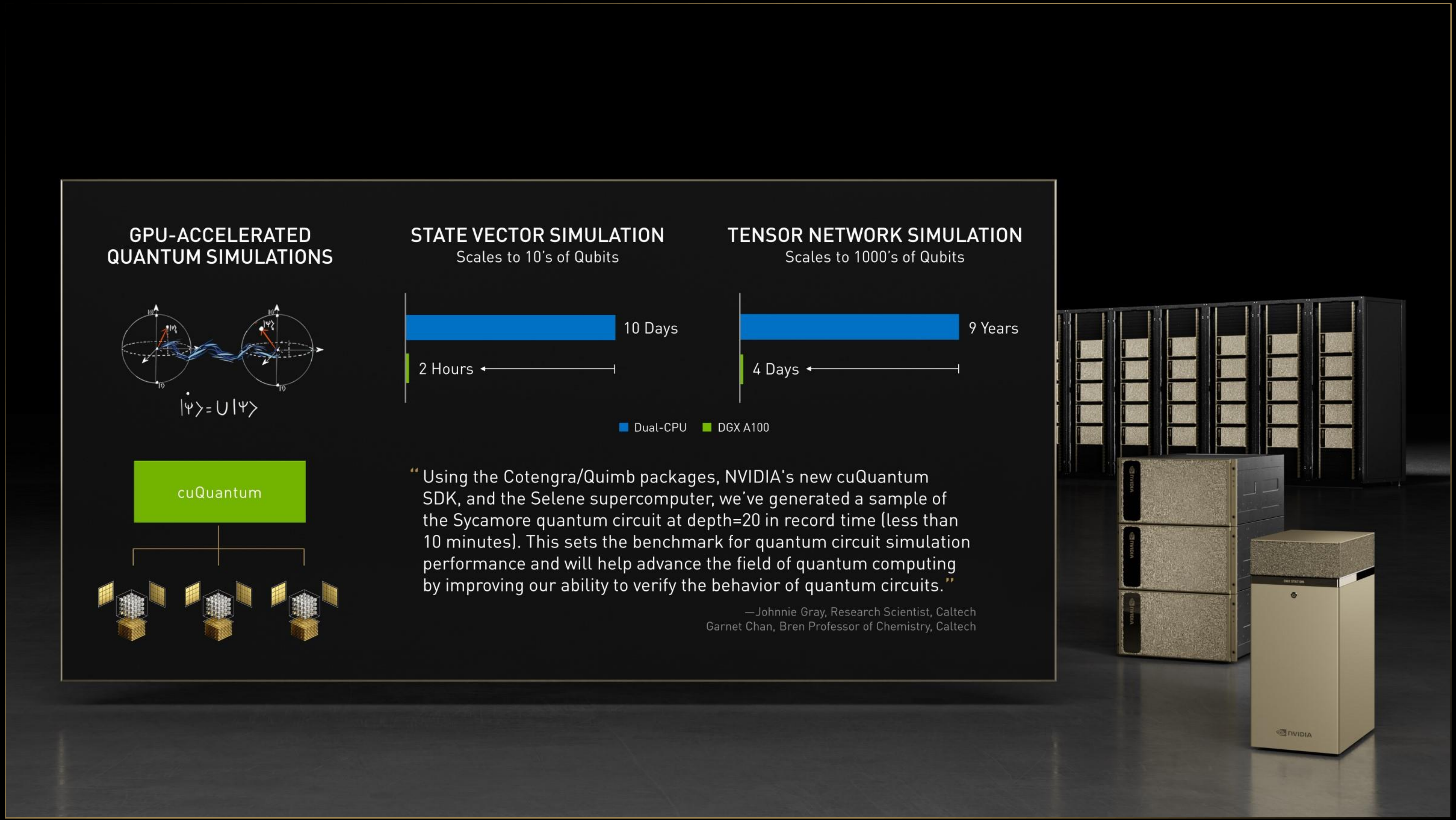


NVIDIA Clara Discovery is our suite of acceleration libraries created for computational drug discovery — from imaging, to quantum chemistry, to gene variant-calling, to using NLP to understand genetics, to using AI to generate new drug compounds.

Biotech innovators are developing incredible technologies using DGX. Oxford Nanopore trained AI models for its popular COVID-19 test on DGX. Schrödinger is accelerating its drug discovery on the system. And Recursion has deployed its own DGX SuperPOD to gain insight from huge biological and chemical datasets.

"NVIDIA SDK SIMULATES QUANTUM COMPUTING CIRCUITS ON GPU SYSTEMS"

SEARCH DATA CENTER



"Using the Cotengra/Quimb packages, NVIDIA's new cuQuantum SDK, and the Selene supercomputer, we've generated a sample of the Sycamore quantum circuit at depth=20 in record time (less than 10 minutes). This sets the benchmark for quantum circuit simulation performance and will help advance the field of quantum computing by improving our ability to verify the behavior of quantum circuits."

—Johnnie Gray, Research Scientist, Caltech
Garnet Chan, Bren Professor of Chemistry, Caltech

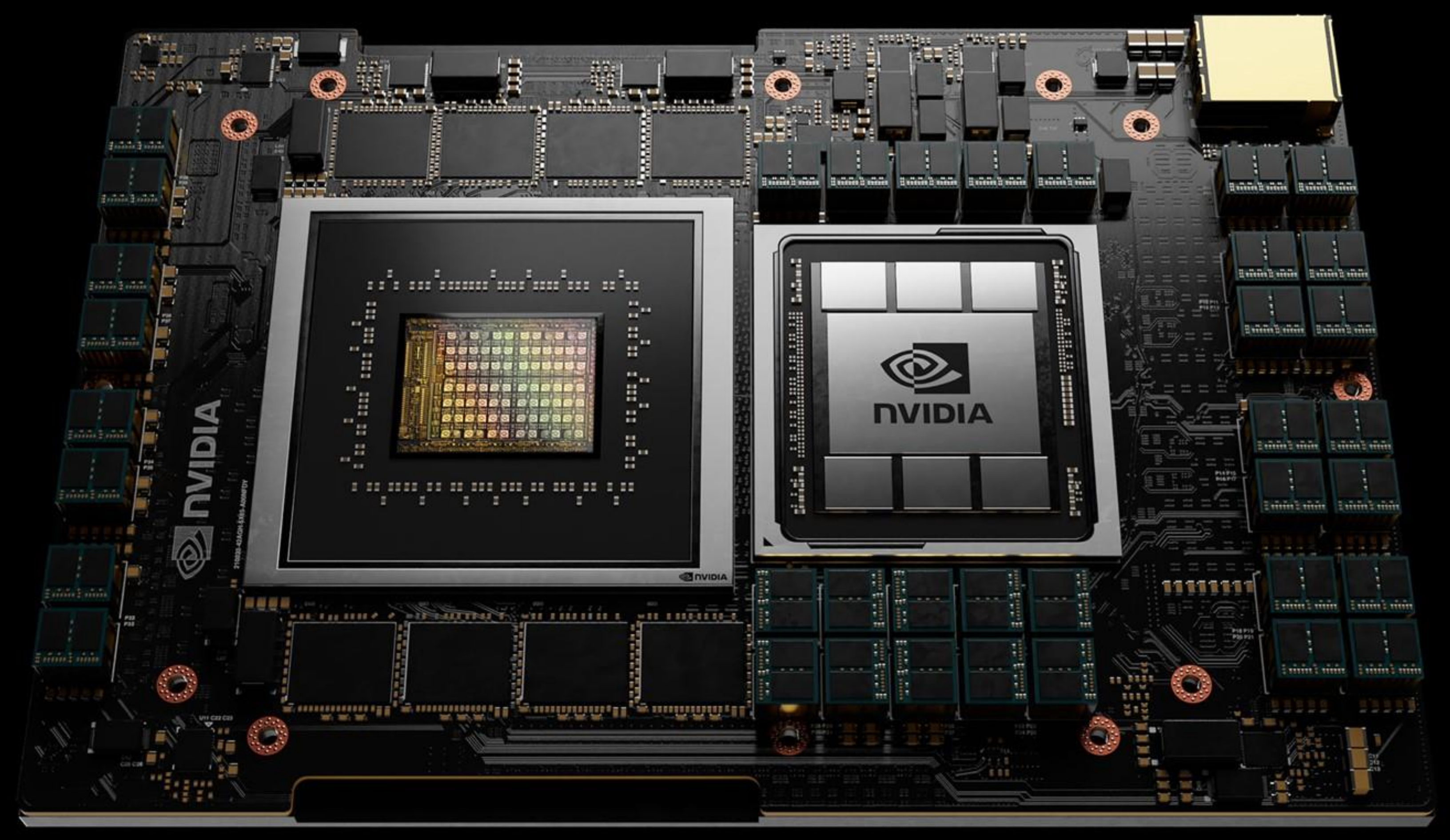
A large community of companies and labs around the world is doing research in quantum computers and algorithms. We're working with many of them.

We announced our cuQuantum acceleration library for simulating quantum circuits — optimized to scale to large GPU memories, multiple GPUs, and multiple DGX nodes.

The speed-up of cuQuantum on DGX is excellent. Running the cuQuantum Benchmark, state vector simulation takes 10 days on a dual-CPU server but only two hours on a DGX A100. We hope cuQuantum will do for quantum computing what cuDNN did for deep learning.

“NVIDIA UNVEILS GRACE, A HIGH-PERFORMANCE ARM SERVER CPU FOR USE IN BIG AI SYSTEMS”

ANAND TECH



We announced Grace, our first data center CPU, named after Grace Hopper, a computer scientist and U.S. Navy Rear Admiral, who in the 1950s pioneered computer programming.

Grace is Arm-based and purpose-built for accelerated computing applications with large amounts of data, such as AI.

Coupled with the GPU and DPU, Grace gives us the third foundational technology for computing, and the ability to rearchitect every aspect of the data center for AI.

"NVIDIA'S NOVEL CPU 'GRACE' WILL POWER THE WORLD'S MOST POWERFUL AI-CAPABLE SUPERCOMPUTER"

MARKETECHPOST

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre
Hewlett Packard Enterprise
NVIDIA
ALPS
CSCS
ETH zürich
Hewlett Packard Enterprise

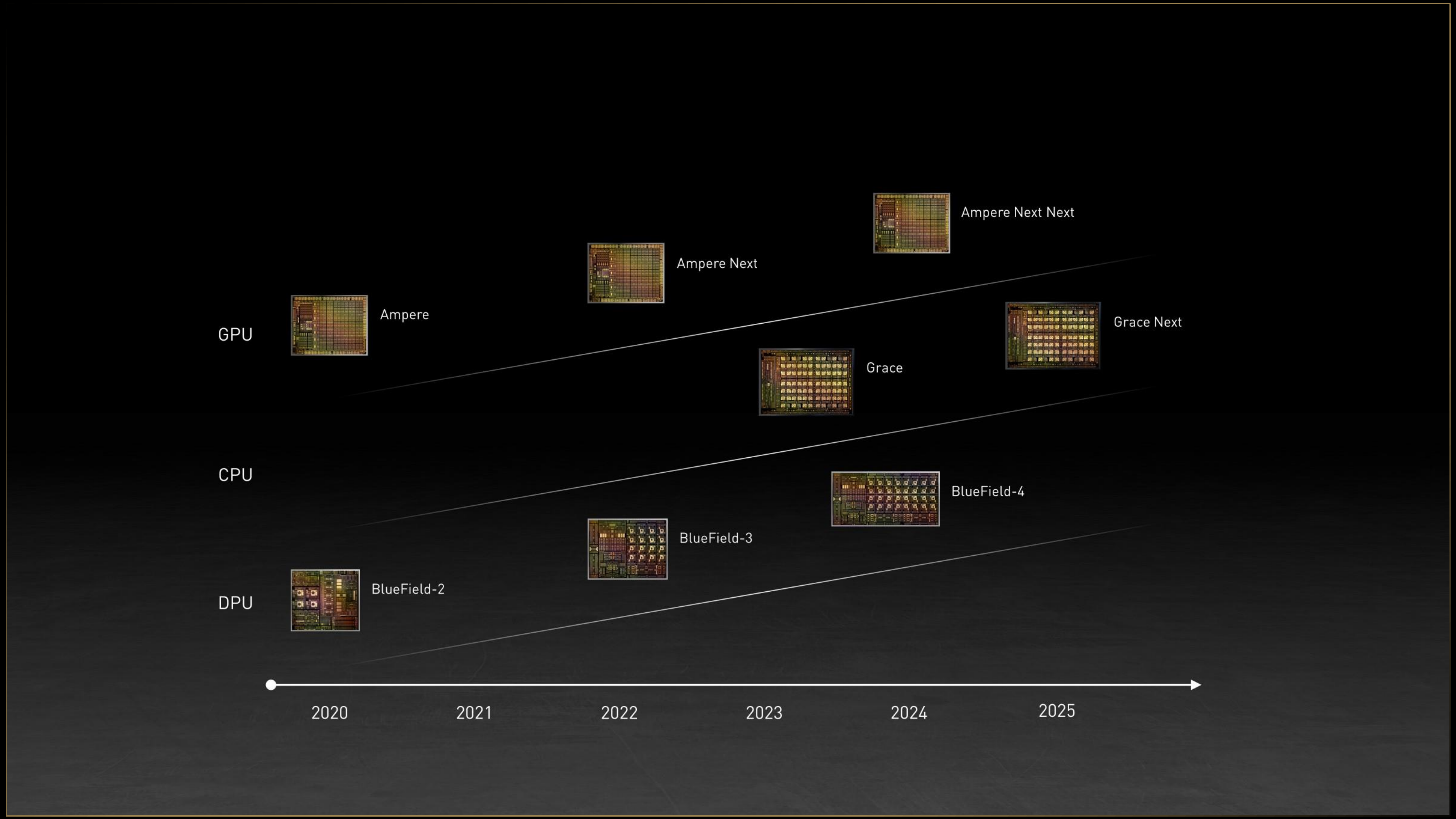
We announced that the Swiss National Supercomputing Centre will build a supercomputer powered by Grace and our next-generation GPU.

Alps will be capable of 20 exaflops for AI, 10x faster than the world's fastest supercomputer today.

The supercomputer will be used for whole-earth-scale weather and climate simulation, quantum chemistry, and quantum physics for the Large Hadron Collider. Alps will be built by HPE and is coming online in 2023.

"NVIDIA IS NOW SUPPLYING ALL MAJOR PROCESSORS: GPU, DPU, AND SOON CPU"

EVERCORE

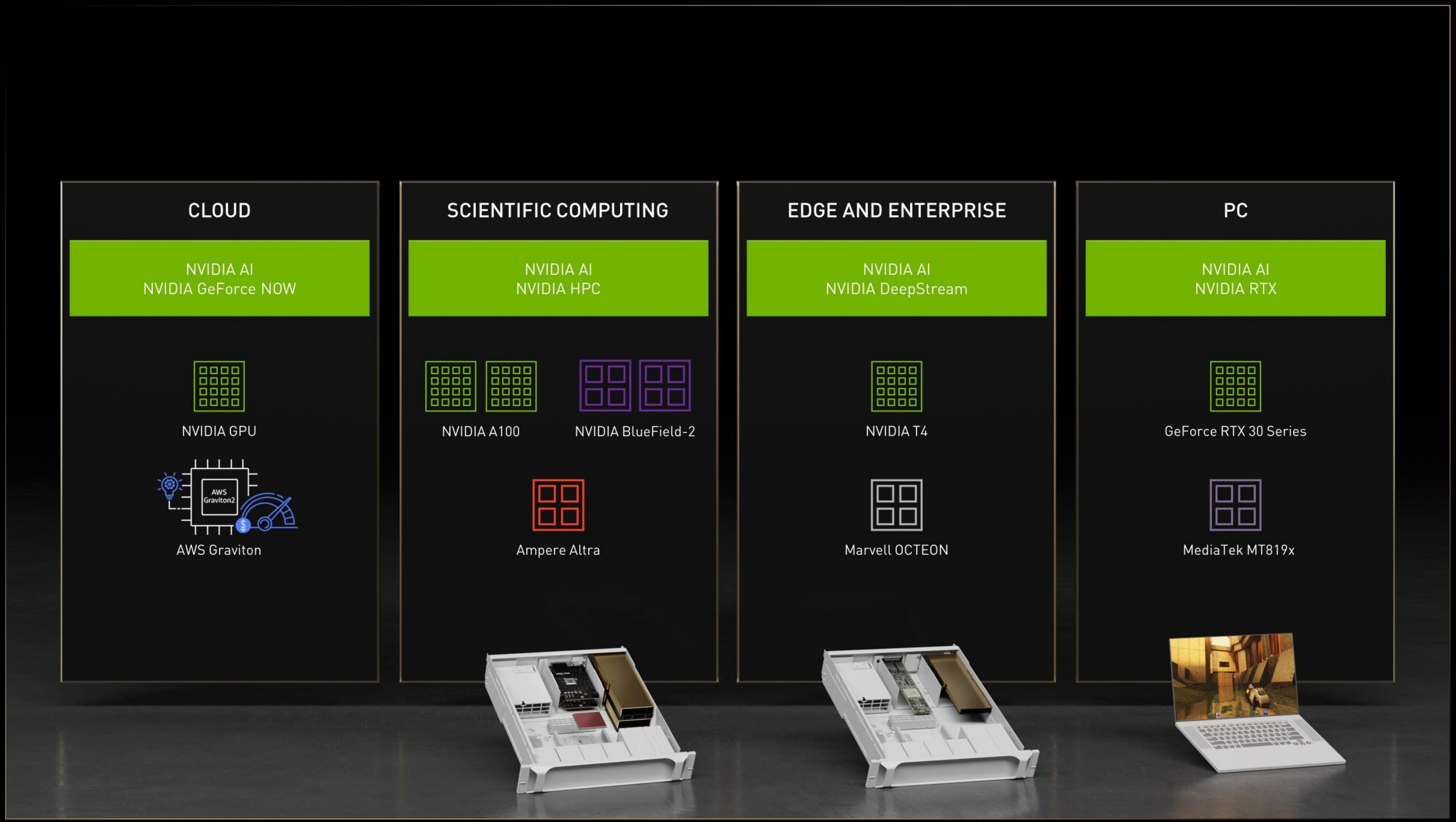


Our new data center roadmap is a rhythm consisting of three chips: CPU, GPU, and DPU. Each chip architecture has a two-year rhythm with likely a kicker in between.

One year will focus on x86 platforms. One year will focus on Arm platforms. Every year will see new exciting products from us. The NVIDIA architecture and platforms will support x86 and Arm — whichever customers and markets prefer.

“FOR ARM-DRIVEN SUPERCOMPUTING, NVIDIA IS RIGHT ON TIME”

THE NEXT PLATFORM



CLOUD

NVIDIA AI
NVIDIA GeForce NOW

NVIDIA GPU

AWS Graviton

SCIENTIFIC COMPUTING

NVIDIA AI
NVIDIA HPC

NVIDIA A100

NVIDIA BlueField-2

Ampere Altra

EDGE AND ENTERPRISE

NVIDIA AI
NVIDIA DeepStream

NVIDIA T4

Marvell OCTEON

PC

NVIDIA AI
NVIDIA RTX

GeForce RTX 30 Series

MediaTek MT819x

NVIDIA can accelerate Arm’s adoption well beyond its success in mobile and embedded.

For the cloud, we announced a partnership with AWS to bring Graviton2 and NVIDIA GPUs together.

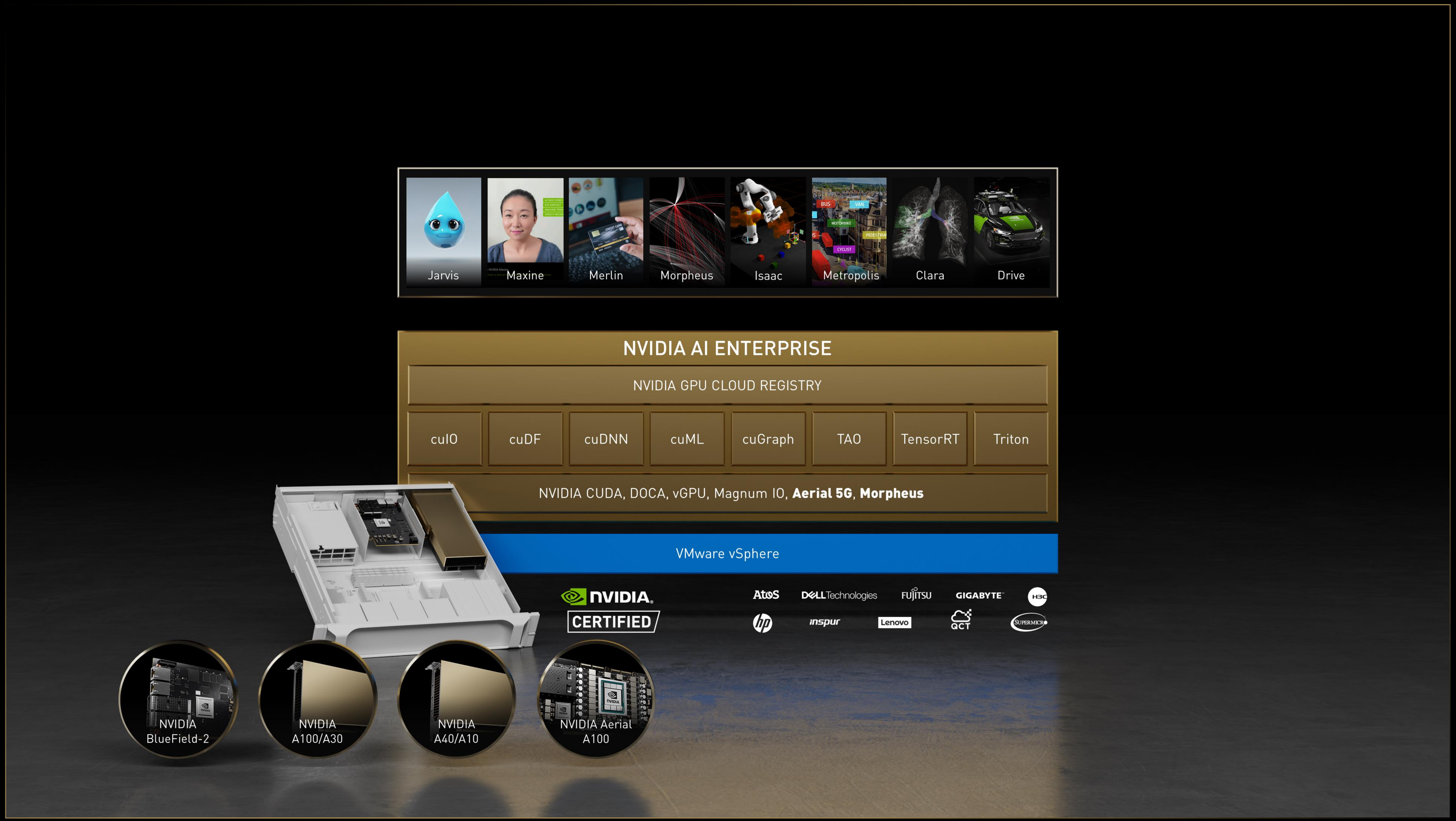
We’re partnering with Ampere Computing to create a scientific and cloud computing SDK and reference system.

We announced a partnership with Marvell to create an edge and enterprise computing SDK and reference system.

And we’re working with MediaTek to create a reference system and SDK to make excellent PCs and notebooks for Chrome OS and Linux.

"NVIDIA'S EFFORTS IN DRIVING AI ADOPTION IN ENTERPRISE SOLIDIFIES ITS POSITION AS A FULL-STACK COMPUTE PROVIDER"

GOLDMAN SACHS



Jarvis | Maxine | Merlin | Morpheus | Isaac | Metropolis | Clara | Drive

NVIDIA AI ENTERPRISE

NVIDIA GPU CLOUD REGISTRY

cuIO | cuDF | cuDNN | cuML | cuGraph | TAO | TensorRT | Triton

NVIDIA CUDA, DOCA, vGPU, Magnum IO, **Aerial 5G**, **Morpheus**

VMware vSphere

NVIDIA CERTIFIED

NVIDIA BlueField-2 | NVIDIA A100/A30 | NVIDIA A40/A10 | NVIDIA Aerial A100

Seventy percent of the world's enterprises run VMware. To bring NVIDIA AI to the enterprise market, we've working together to make VMware a top-notch platform for AI.

We announced NVIDIA EGX server with Aerial A100, the first 5G base station that is also a cloud-native, secure, AI edge data center. Google will support NVIDIA Aerial in the GCP cloud.

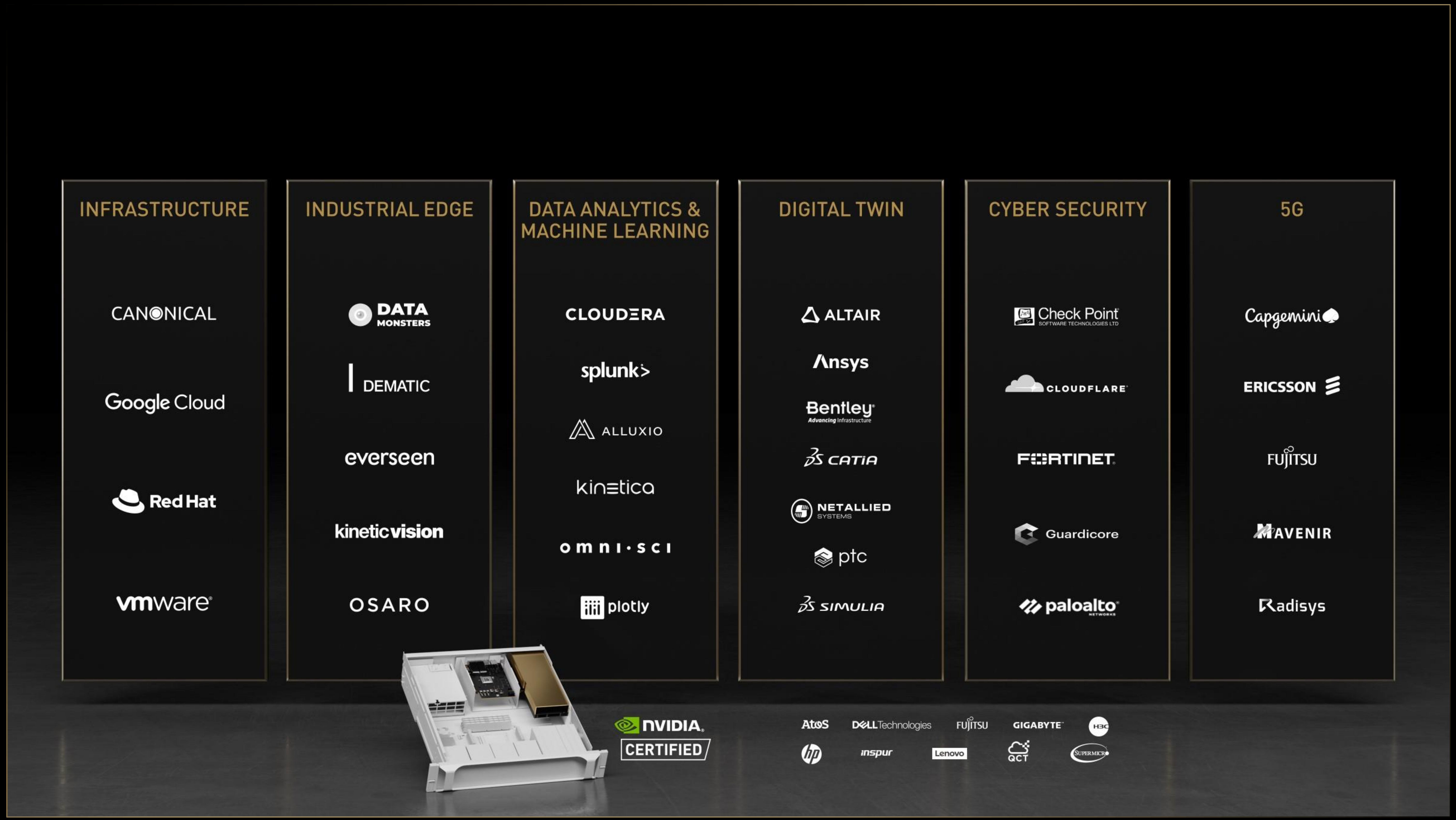
We also announced NVIDIA Morpheus — a data center security platform for real-time all-packet inspection.

And we now have NVIDIA AI Enterprise software so businesses can get direct-line support from NVIDIA.

NVIDIA AI on EGX with Aerial 5G — the AI computing platform for the enterprise and edge, the next wave of AI.

“FOR ENTERPRISES, THE MOST SIGNIFICANT SYSTEM ANNOUNCEMENT OUT OF GTC IS THE EGX REFERENCE PLATFORM”

DIGINOMICA



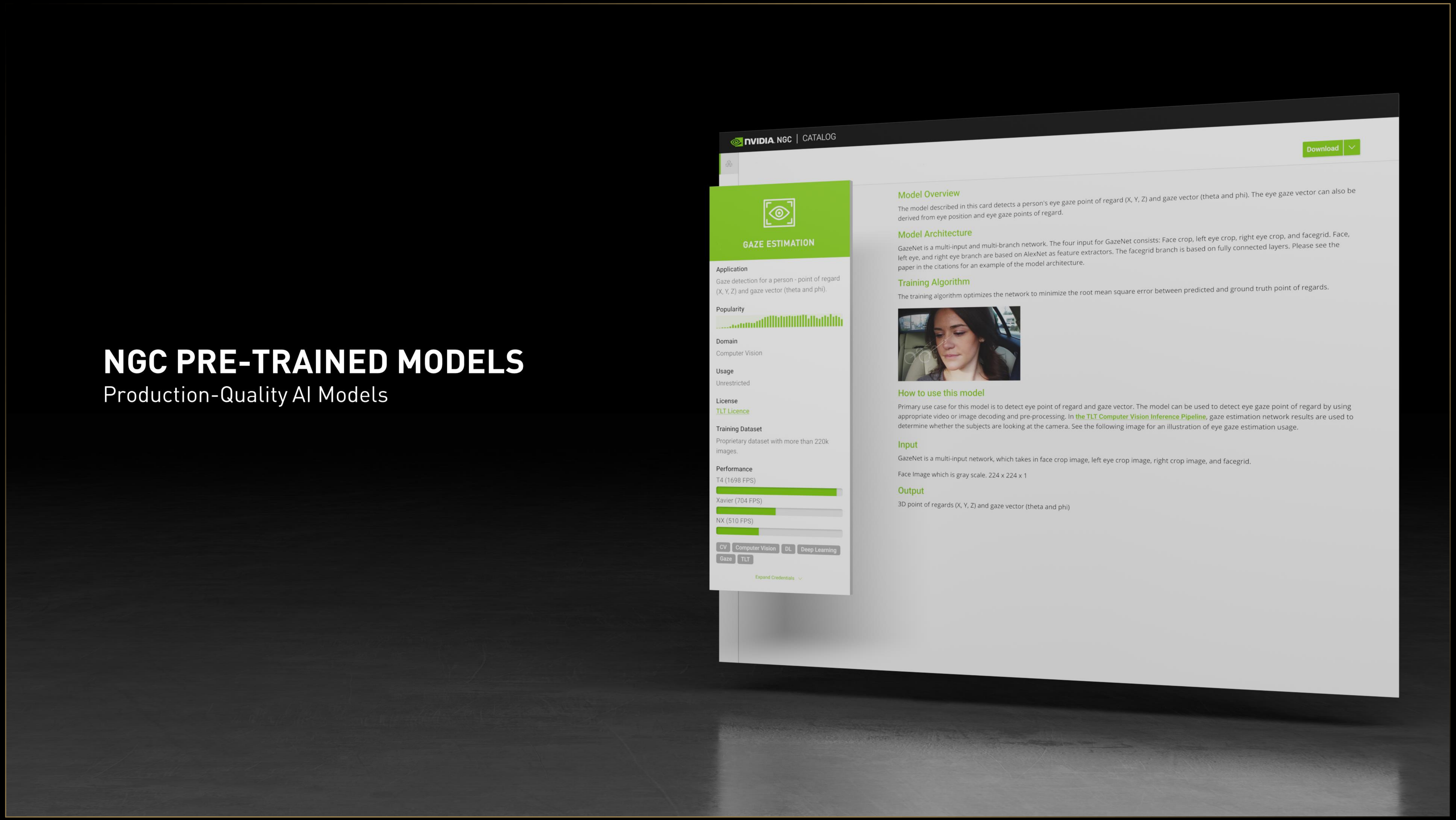
INFRASTRUCTURE	INDUSTRIAL EDGE	DATA ANALYTICS & MACHINE LEARNING	DIGITAL TWIN	CYBER SECURITY	5G
CANONICAL	DATA MONSTERS	CLOUDERA	ALTAIR	Check Point	Capgemini
Google Cloud	DEMATIC	splunk>	Ansys	CLOUDFLARE	ERICSSON
Red Hat	everseen	ALLUXIO	Bentley	FORTINET	FUJITSU
vmware	kineticvision	kinetica	CATIA	Guardicore	MAVENIR
	OSARO	omni·sci	NETALLIED	paloalto	Radisys
		plotly	ptc		
			SIMULIA		

The IT ecosystem has been hungry for an AI computing platform that is enterprise and edge ready. NVIDIA AI on EGX with Aerial 5G is the foundation of what the IT ecosystem has been waiting for.

We are supported by leaders from all across the IT industry, from systems, infrastructure software, storage and security, data analytics, industrial edge solutions, manufacturing design and automation, to 5G infrastructure.

“SPEED UP AI DEVELOPMENT BY OVER 10X WITH NVIDIA PRE-TRAINED MODELS”

SMART INDUSTRY



NGC PRE-TRAINED MODELS
Production-Quality AI Models

GAZE ESTIMATION

Application
Gaze detection for a person - point of regard (X, Y, Z) and gaze vector (theta and phi).

Popularity

Domain
Computer Vision

Usage
Unrestricted

License
TLT Licence

Training Dataset
Proprietary dataset with more than 220k images.

Performance
T4 (1698 FPS)
Xavier (704 FPS)
NX (510 FPS)

CV Computer Vision DL Deep Learning Gaze TLT

Model Overview

The model described in this card detects a person's eye gaze point of regard (X, Y, Z) and gaze vector (theta and phi). The eye gaze vector can also be derived from eye position and eye gaze points of regard.

Model Architecture

GazeNet is a multi-input and multi-branch network. The four input for GazeNet consists: Face crop, left eye crop, right eye crop, and facegrid. Face, left eye, and right eye branch are based on AlexNet as feature extractors. The facegrid branch is based on fully connected layers. Please see the paper in the citations for an example of the model architecture.

Training Algorithm

The training algorithm optimizes the network to minimize the root mean square error between predicted and ground truth point of regards.

How to use this model

Primary use case for this model is to detect eye point of regard and gaze vector. The model can be used to detect eye gaze point of regard by using appropriate video or image decoding and pre-processing. In the TLT Computer Vision Inference Pipeline, gaze estimation network results are used to determine whether the subjects are looking at the camera. See the following image for an illustration of eye gaze estimation usage.

Input

GazeNet is a multi-input network, which takes in face crop image, left eye crop image, right crop image, and facegrid.

Face Image which is gray scale. 224 x 224 x 1

Output

3D point of regards (X, Y, Z) and gaze vector (theta and phi)

NVIDIA has invested billions in developing AI, but it goes well beyond systems and apps. We’ve packaged up pre-trained AI models that developers can easily integrate with their own apps. They’re production quality, trained by experts, and will continue to improve over time.

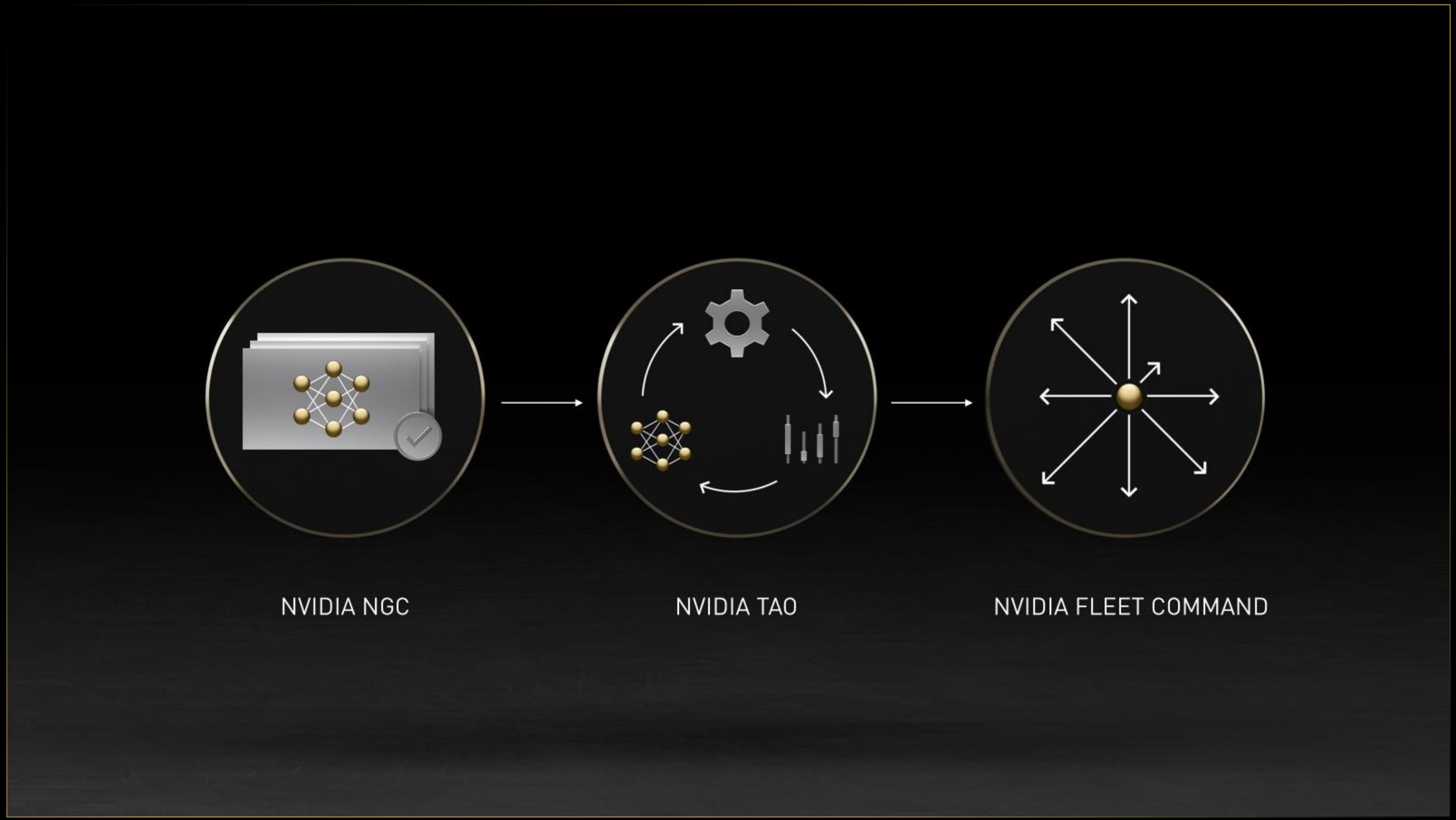
NVIDIA Jarvis is our state-of-the-art framework for conversational AI. Jarvis interacts in about 100 milliseconds — faster than the blink of an eye.

NVIDIA Merlin is our framework for recommender systems — the most important machine learning pipeline today.

And NVIDIA Maxine, our framework for enabling our avatars in virtual worlds, is already transforming video conferencing.

“NVIDIA SIMPLIFIES THE AI WORKFLOW WITH NVIDIA TAO AND FLEET COMMAND TO MAKE THE TRIP SHORTER AND LESS COSTLY”

ANALYTICS INDIA



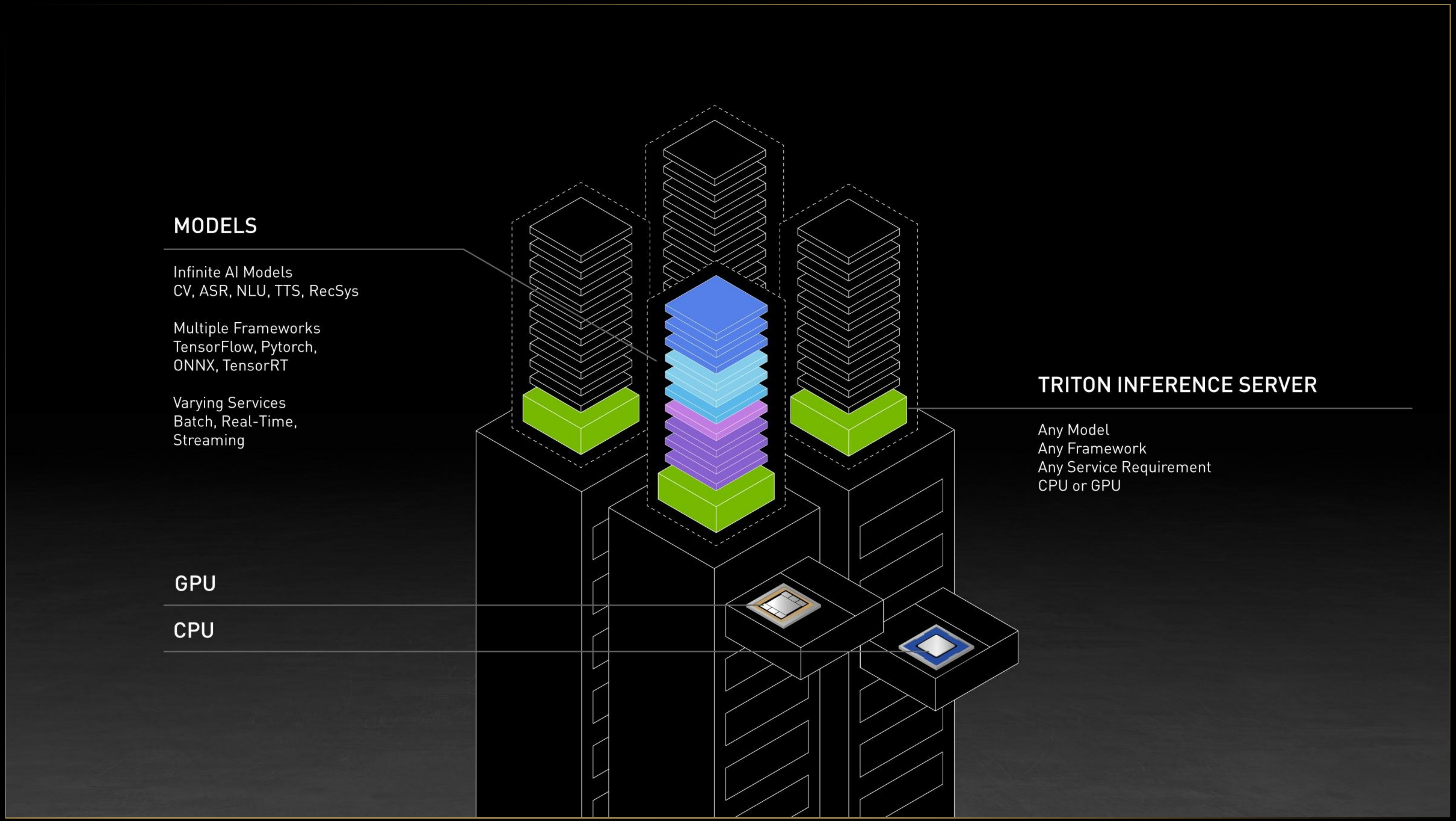
NVIDIA pre-trained models are state of the art and meticulously trained, but there is an infinite diversity of application domains, environments, and specializations.

NVIDIA TAO is a framework that encapsulates the entire workflow to customize AI models.

Fleet Command is a cloud-native platform for securely operating and orchestrating AI across a distributed fleet of computers — it was purpose-built for operating AI at the edge.

“NVIDIA TRITON INFERENCE SERVER SUPPORTS DEPLOYMENT OF MACHINE LEARNING MODELS IN PRODUCTION AND COMMERCIAL ENVIRONMENTS”

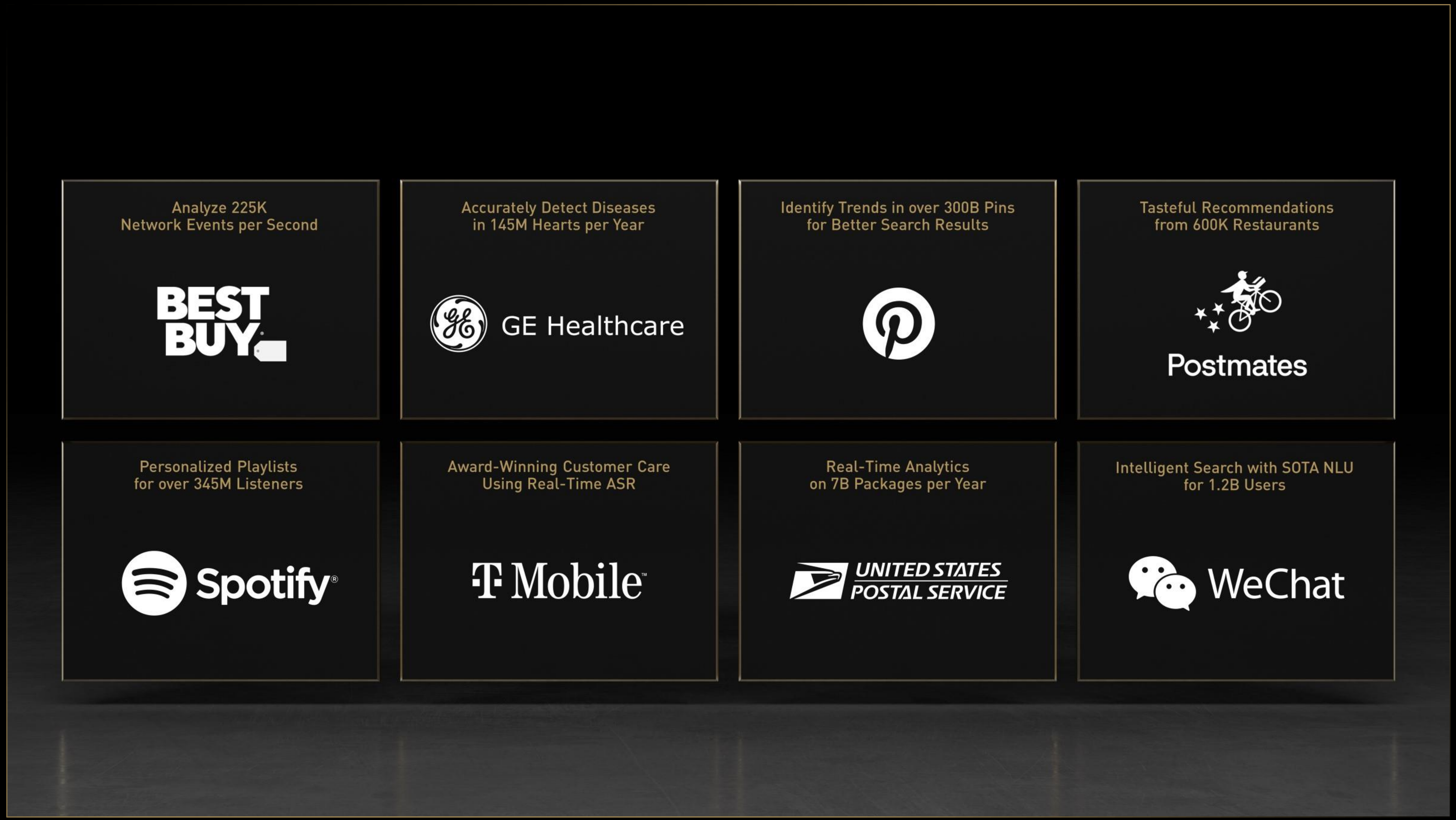
ELECTRONIC DESIGN



NVIDIA Triton is our inference server, offering insight from the continuous streams of data coming into an EGX server or cloud instance. Triton is a model scheduling and dispatch engine that can handle just about anything thrown at it: any AI model, from any framework, with controllable service requirements. Triton schedules on multiple generations of NVIDIA GPUs and on x86 CPUs and scales with Kubernetes.

“NVIDIA IS MAKING AI MORE ACCESSIBLE TO TRADITIONAL ENTERPRISES”

DATA CENTER KNOWLEDGE



Analyze 225K Network Events per Second — BEST BUY

Accurately Detect Diseases in 145M Hearts per Year — GE Healthcare

Identify Trends in over 300B Pins for Better Search Results

Tasteful Recommendations from 600K Restaurants — Postmates

Personalized Playlists for over 345M Listeners — Spotify

Award-Winning Customer Care Using Real-Time ASR — T Mobile

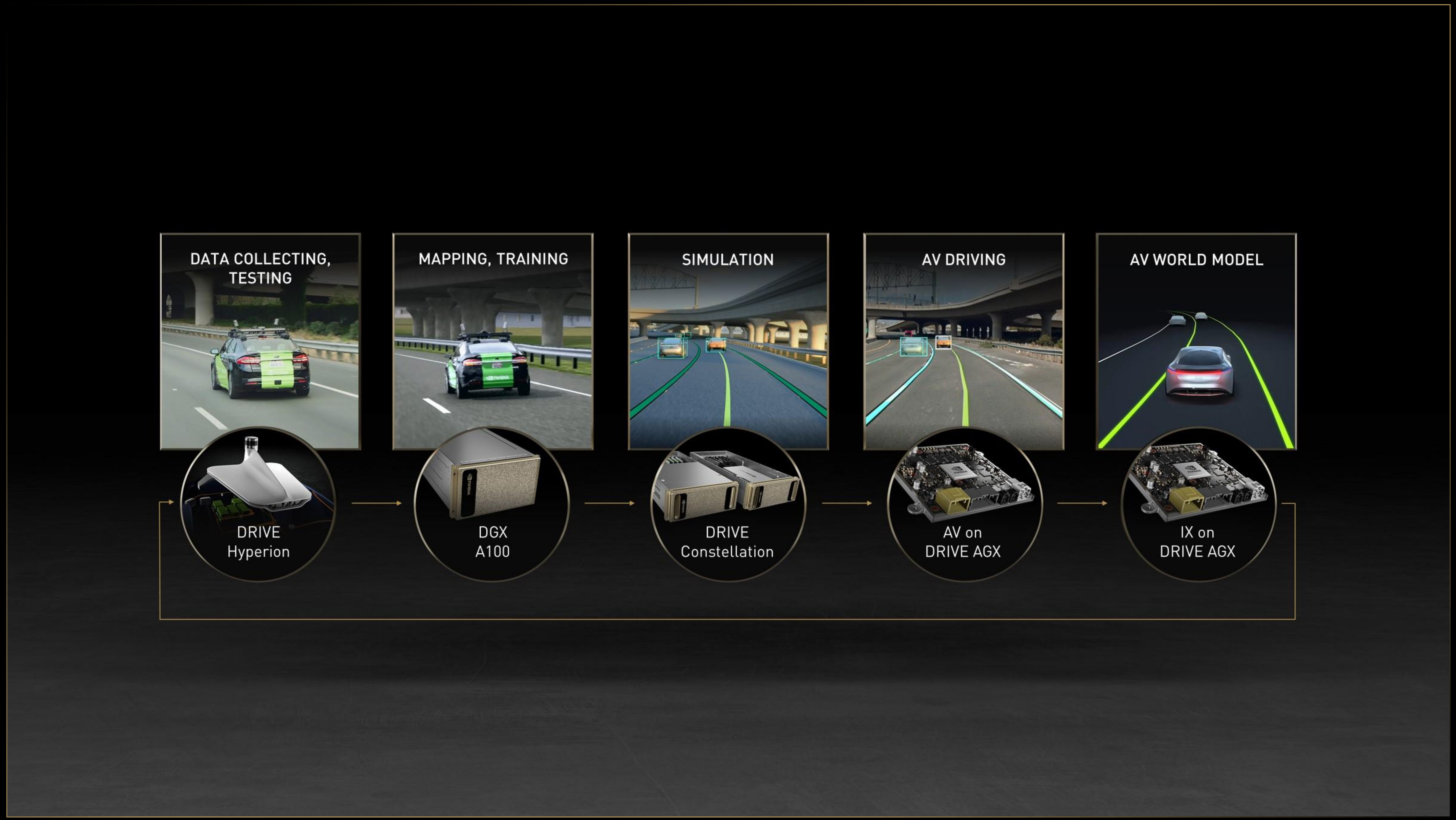
Real-Time Analytics on 7B Packages per Year — UNITED STATES POSTAL SERVICE

Intelligent Search with SOTA NLU for 1.2B Users — WeChat

We provide NVIDIA’s AI capabilities to our ecosystem of millions of developers around the world. Thousands of companies are building their most important services on NVIDIA AI.

“NVIDIA’S AUTO BUSINESS TAKES A FRONT SEAT AT GTC”

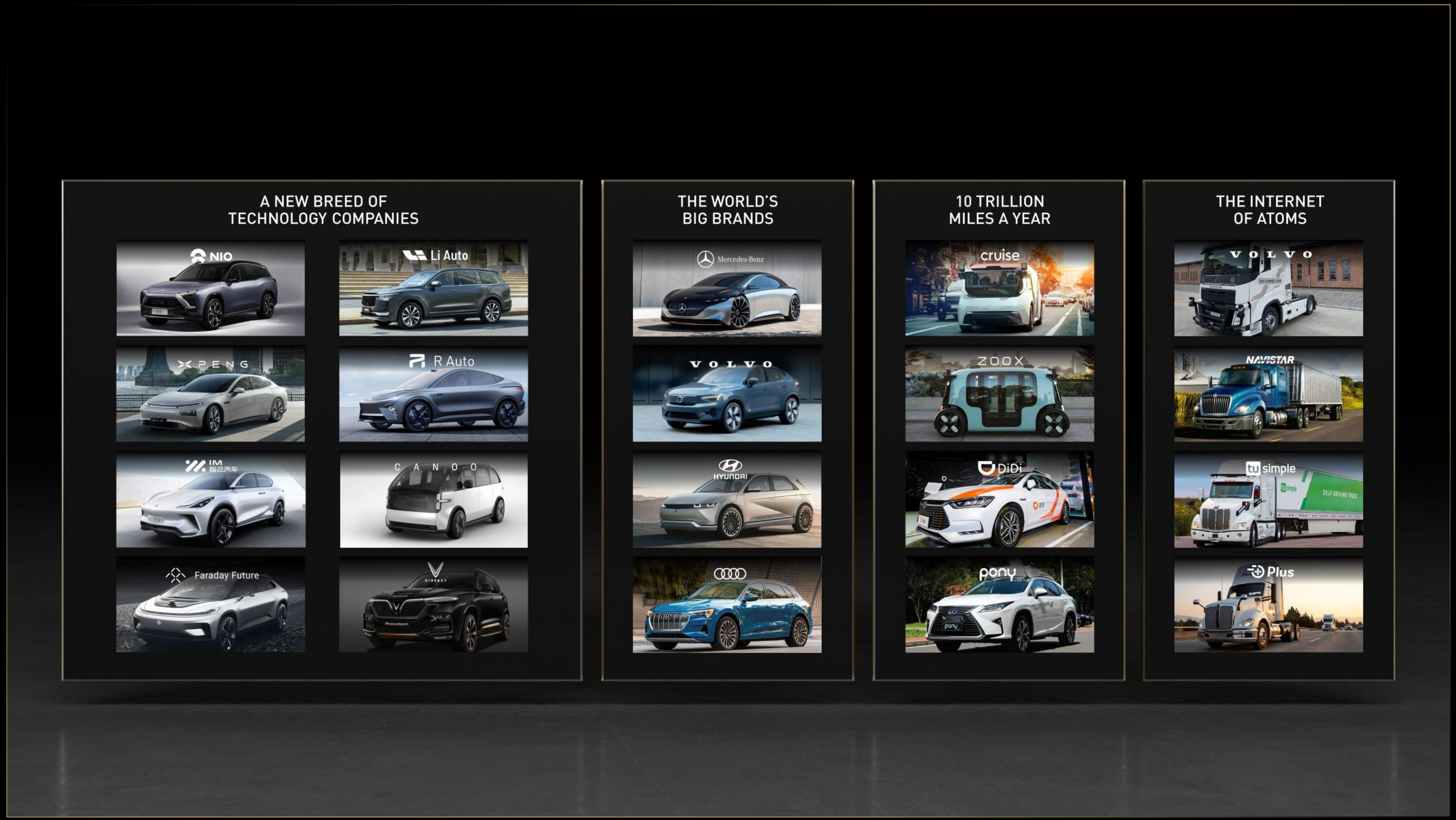
FORBES



DRIVE is our end-to-end AV platform and service — from the chips and computers, to the driving software, the simulator and digital twin, fleet command operations, to road testing — all of it adhering to the highest functional safety and cybersecurity standards. Our current chip, Orin, can process in one central computer the cluster, infotainment, passenger interaction AI, and confidence view. And the 8th generation Hyperion car platform — which includes reference sensors, AV and central computers, 3D ground-truth data recorders, networking, and software — offers car companies a full AV development platform.

“AGREEMENTS EXPECTED IN THE COMING WEEKS MARK A REBOOT OF GROWTH FOR NVIDIA’S AUTOMOTIVE BUSINESS”

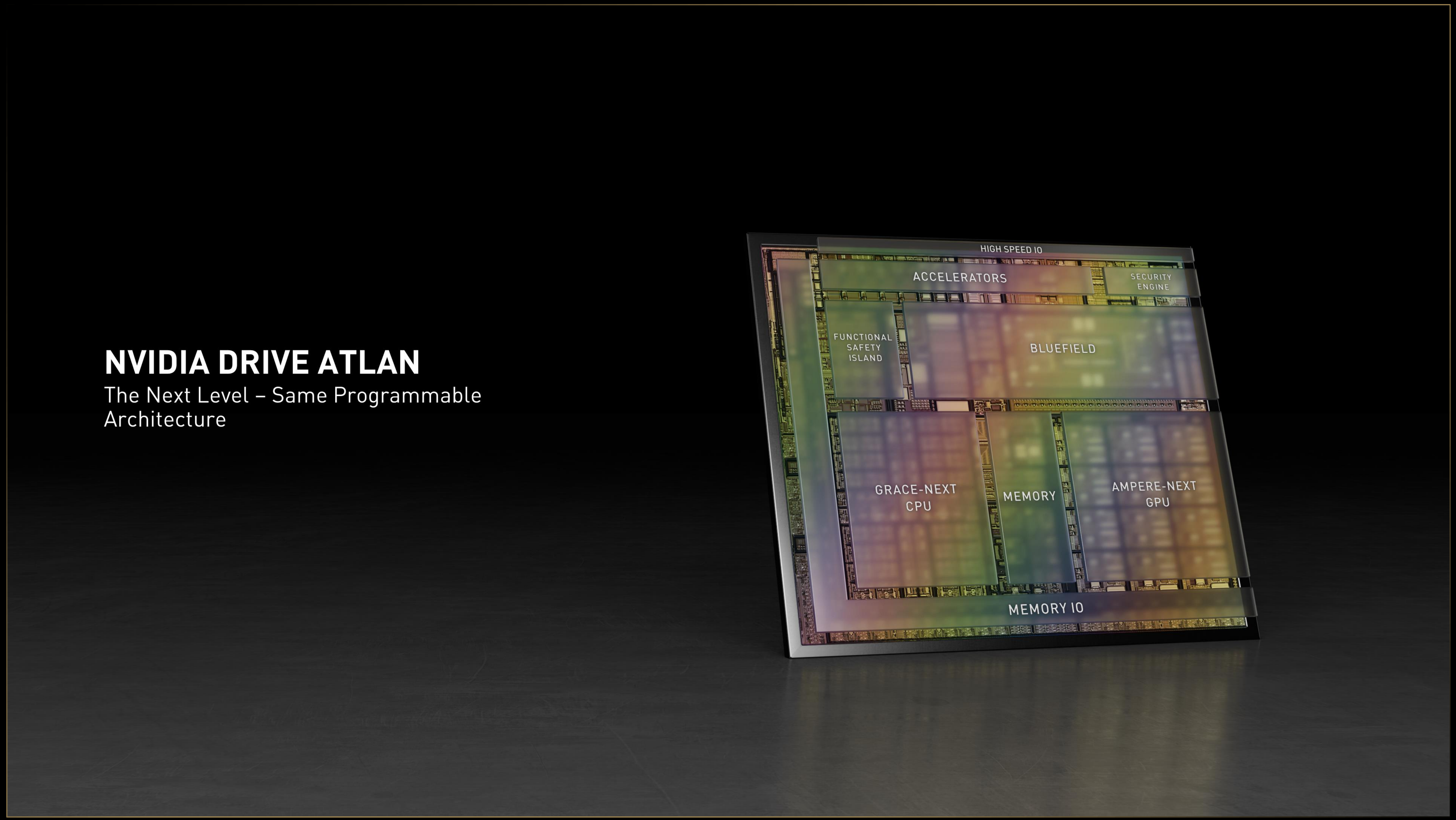
REUTERS



The transportation industry is quickly becoming a technology industry. Future cars are going to be completely programmable computers and business models are going to be software driven. Car companies will offer software services for the life of the car. From the latest EV and robotaxi companies to the world’s biggest auto brands, the leaders have adopted NVIDIA DRIVE Orin as their computing platform.

“THE SHOWSTOPPER WAS THE UNVEILING OF NVIDIA’S SUPER-POWERFUL NEW SOC DRIVE ATLAN”

FORBES



NVIDIA DRIVE ATLAN
The Next Level – Same Programmable Architecture

To achieve higher autonomy in more conditions, sensor resolutions will continue to increase. There will be more of them. AI models will get more sophisticated. There will be more redundancy and safety functionality. We’re going to need all of the computing we can get.

NVIDIA DRIVE Atlan will be 1,000 TOPS on one chip — more than the total compute in most L5 robotaxis today. Atlan is a technical marvel — fusing all of NVIDIA’s technologies in AI, auto, robotics, safety, and BlueField-secure data centers.

“NVIDIA IS BUILDING A GIANT VIRTUAL ‘METAVERSE’ OF THE WORLD, WITH ‘DIGITAL TWINS’ OF CARS, CITIES, AND PEOPLE”

THE INDEPENDENT

The NVIDIA DRIVE Digital Twin is used throughout the development of a self-driving car — for HD map reconstruction, synthetic data generation, new scenario simulations, hardware-in-the-loop simulations, and release validation. It can replay unfamiliar scenarios experienced by a car, and allow a teleoperator to pilot a car remotely. DRIVE Sim, the engine of Drive Digital Twin, will be available for the community this summer.

“THE LEVEL OF PLATFORM INNOVATION IS MIND BOGGLING”

ROSENBLATT

GEFORCE | OMNIVERSE | cuQUANTUM | MORPHEUS | JARVIS | MERLIN | MAXINE | CLARA | METROPOLIS | AERIAL | ISAAC | DRIVE

APPLICATION FRAMEWORKS

NVIDIA AI

NVIDIA GPU CLOUD REGISTRY

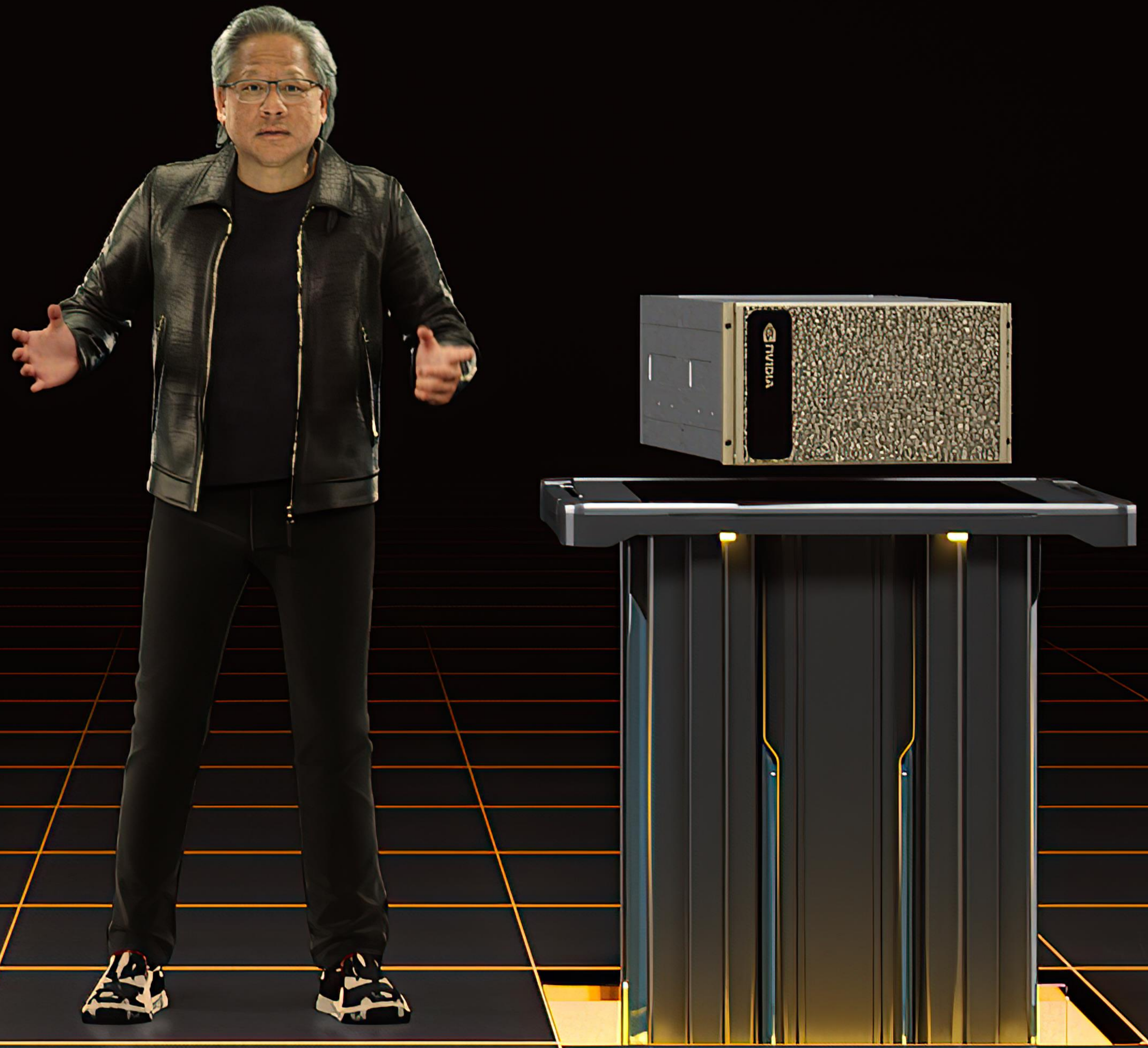
cuIO | cuDF | cuDNN | cuML | cuGraph | TAO | TensorRT | Triton

NVIDIA CUDA, DOCA, vGPU, Magnum IO

PLATFORM SOFTWARE

CHIPS & SYSTEMS

> "NVIDIA REINVENTS ITSELF EVERY SINGLE YEAR. WE ARE GOING TO CALL NVIDIA 'THE GOAT,' THAT IS, THE GREATEST OF ALL TIME"
>
> MAD MONEY



NVIDIA.COM/GTC