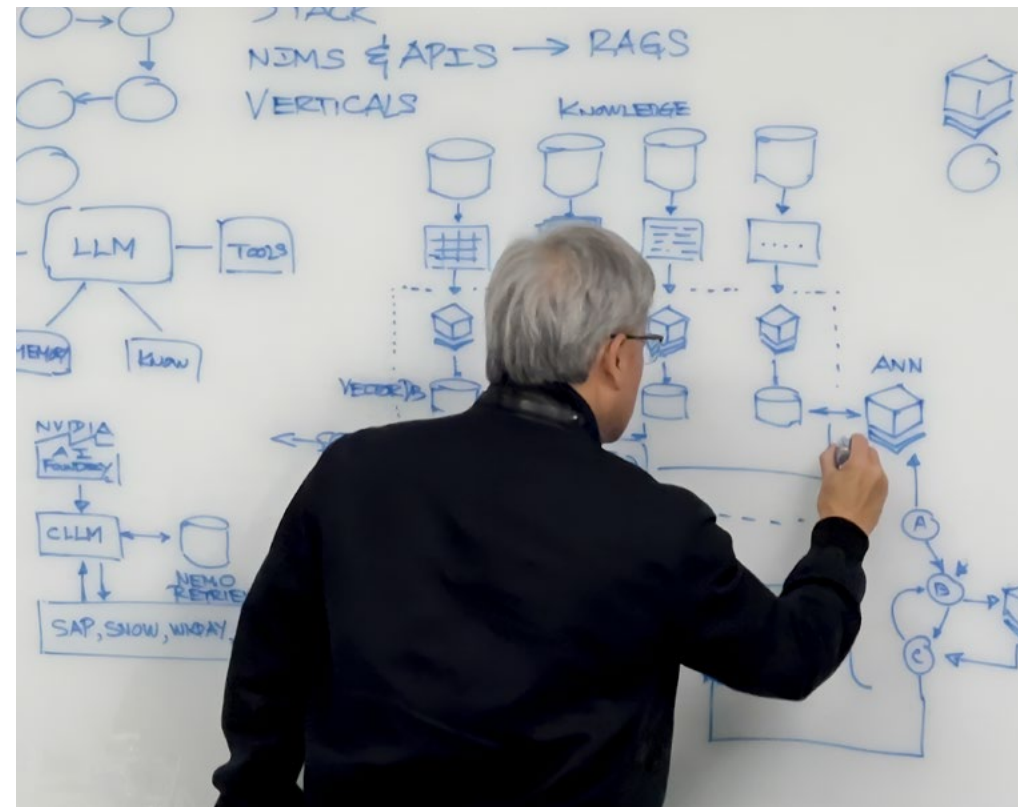




GTC 2024
Highlights



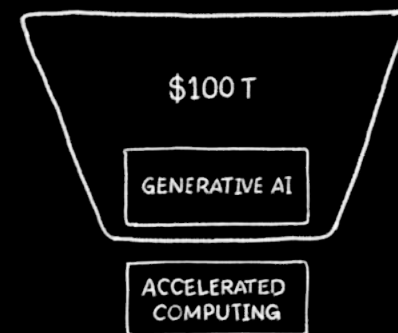


**“GTC is a front-row seat to
what’s happening in AI.”**

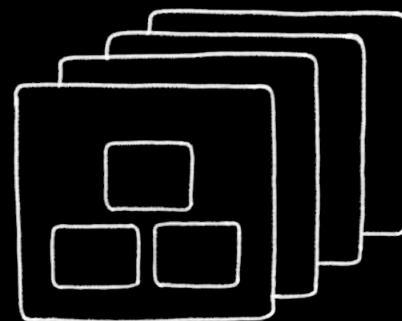
Bloomberg



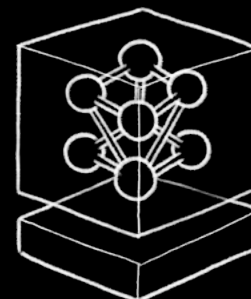
A New Industrial Revolution



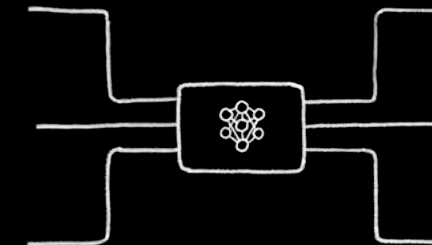
NEW INDUSTRY



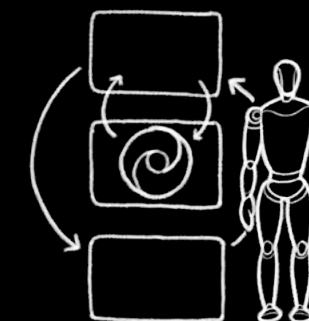
BLACKWELL
PLATFORM



NIMs



NEMO AND
NVIDIA AI FOUNDRY

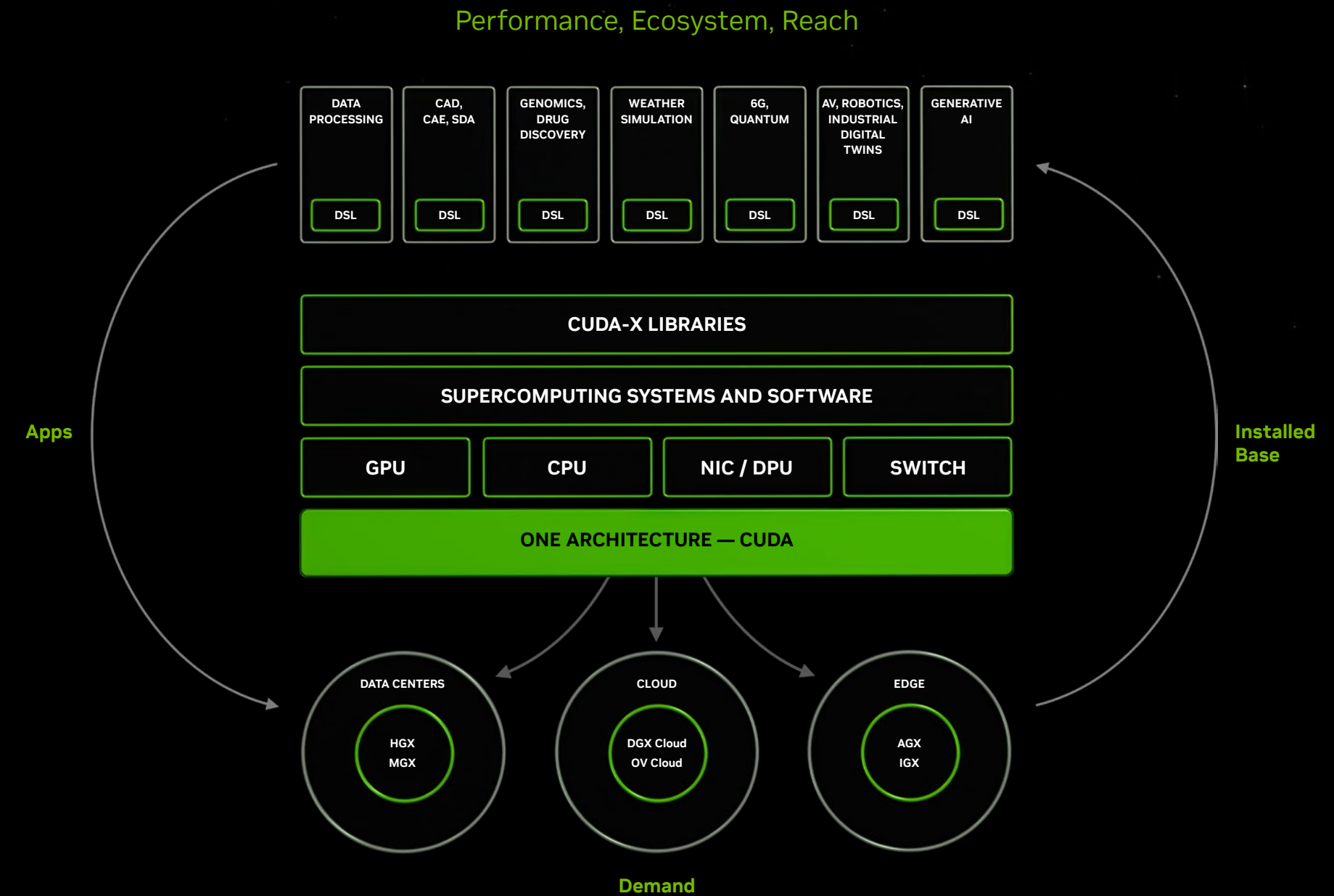


OMNIVERSE AND
ISAAC ROBOTICS

“NVIDIA’s moat is its software and ecosystem.”

The Edge Singapore

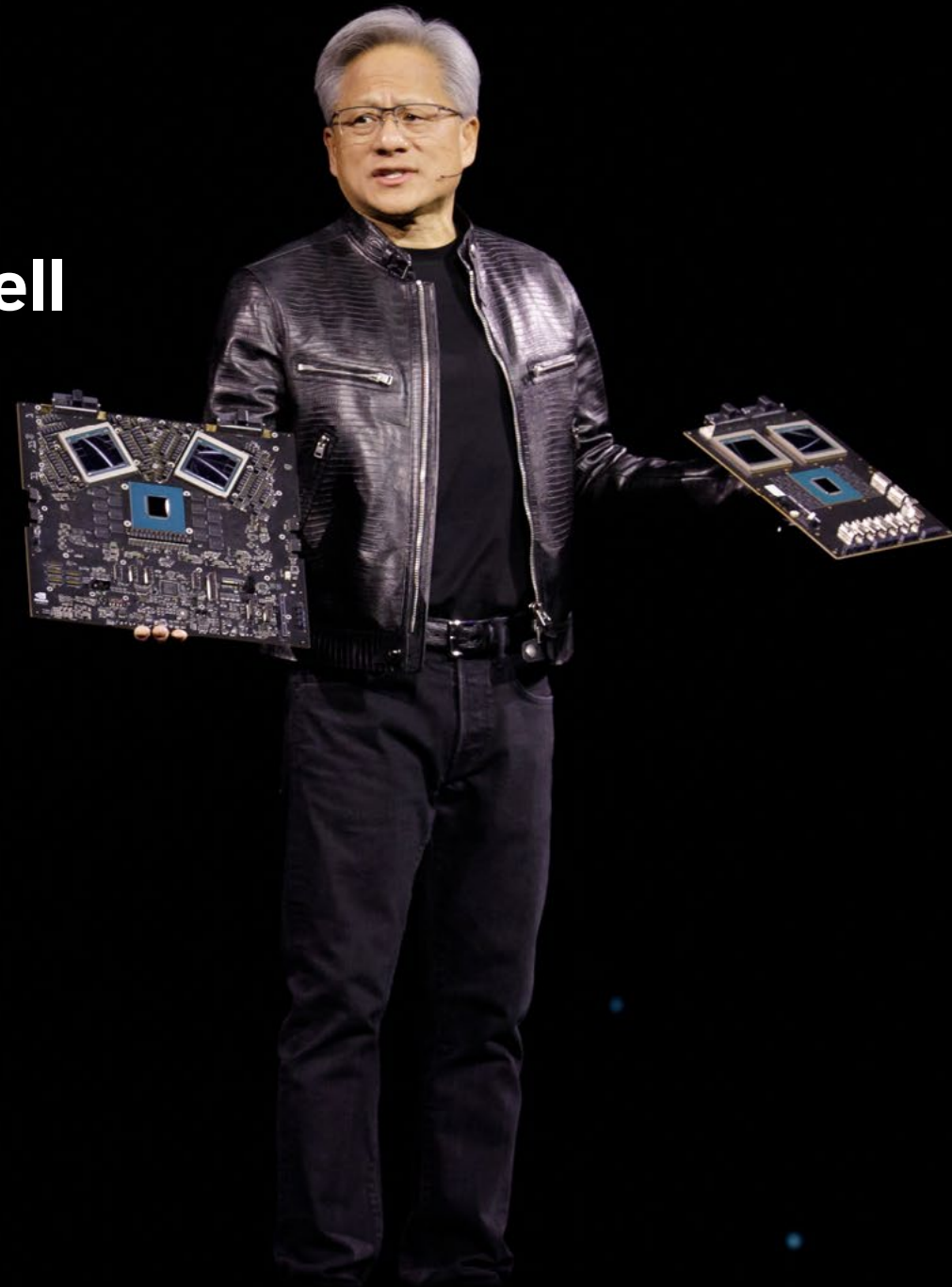
The global NVIDIA ecosystem now approaches 5 million developers—capping a record-setting year for new developers. Forty thousand companies have worked with NVIDIA. There are now more than 3,300 GPU-accelerated applications. More than 1,600 generative AI companies are building on NVIDIA.



“NVIDIA’s new Blackwell chip is key to the next stage of AI.”

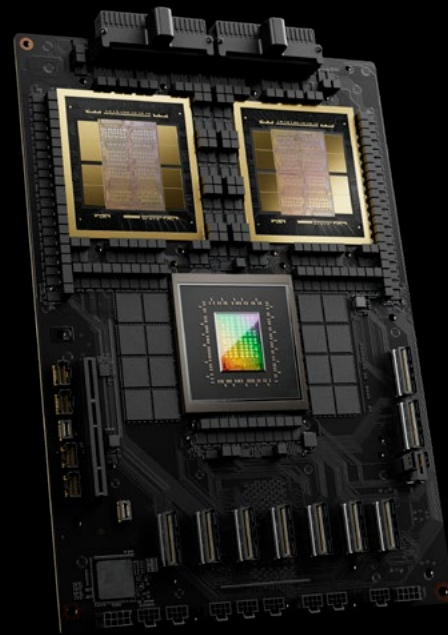
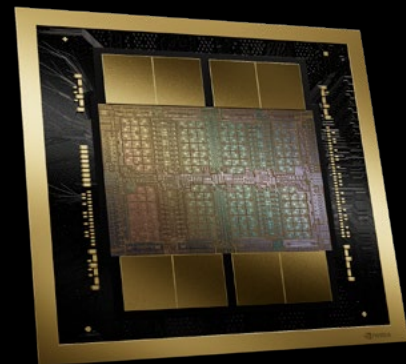
Bloomberg

The NVIDIA Blackwell platform has arrived—enabling organizations everywhere to build and run real-time generative AI on trillion-parameter large language models at up to 25X less cost and energy consumption than its predecessor.



“Meet Blackwell, the new GPU for the AI era.”

Engineering.com



NVIDIA Blackwell Platform Arrives to Power a New Era of Computing

March 18, 2024

GTC—Powering a new era of computing, NVIDIA today announced that the NVIDIA Blackwell platform has arrived—enabling organizations everywhere to build and run real-time generative AI on trillion-parameter large language models at up to 25x less cost and energy consumption than its predecessor.

The Blackwell GPU architecture features six transformative technologies for accelerated computing, which will help unlock breakthroughs in data processing, engineering simulation, electronic design automation, computer-aided drug design, quantum computing and generative AI—all emerging industry opportunities for NVIDIA.

“For three decades we’ve pursued accelerated computing, with the goal of enabling transformative breakthroughs like deep learning and AI,” said **Jensen Huang, founder and CEO of NVIDIA**. “Generative AI is the defining technology of our time. Blackwell is the engine to power this new industrial revolution. Working with the most dynamic companies in the world, we will realize the promise of AI for every industry.”

Among the many organizations expected to adopt Blackwell are Amazon Web Services, Dell Technologies, Google, Meta, Microsoft, OpenAI, Oracle, Tesla and xAI.

Sundar Pichai, CEO of Alphabet and Google: “Scaling services like Search and Gmail to billions of users has taught us a lot about managing compute infrastructure. As we enter the AI platform shift, we continue to invest deeply in infrastructure for our own products and services, and for our Cloud customers. We are fortunate to have a longstanding partnership with NVIDIA, and look forward to bringing the breakthrough capabilities of the Blackwell GPU to our Cloud customers and teams across Google, including Google DeepMind, to accelerate future discoveries.”

Andy Jassy, president and CEO of Amazon: “Our deep collaboration with NVIDIA goes back more than 13 years, when we launched the world’s first GPU cloud instance on AWS. Today we offer the widest range of GPU solutions available anywhere in the cloud, supporting the world’s most technologically advanced accelerated workloads. It’s why the new NVIDIA Blackwell GPU will run so well on AWS and the reason that NVIDIA chose AWS to co-develop Project Ceiba, combining NVIDIA’s next-generation Grace Blackwell Superchips with the AWS Nitro System’s advanced virtualization and ultra-fast Elastic Fabric Adapter networking, for NVIDIA’s own AI research and development. Through this joint effort between AWS and NVIDIA engineers, we’re continuing to innovate together to make AWS the best place for anyone to run NVIDIA GPUs in the cloud.”

Michael Dell, founder and CEO of Dell Technologies: “Generative AI is critical to creating smarter, more reliable and efficient systems. Dell Technologies and NVIDIA are working together to shape the future of technology. With the launch of Blackwell, we will continue to deliver the next-generation of accelerated products and services to our customers, providing them with the tools they need to drive innovation across industries.”

Demis Hassabis, cofounder and CEO of Google DeepMind: “The transformative potential of AI is incredible, and it will help us solve some of the world’s most important scientific problems. Blackwell’s breakthrough technological capabilities will provide the critical compute needed to help the world’s brightest minds chart new scientific discoveries.”

Mark Zuckerberg, founder and CEO of Meta: “AI already powers everything from our large language models to our content recommendations, ads, and safety systems, and it’s only going to get more important in the future. We’re looking forward to using NVIDIA’s Blackwell to help train our open-source Llama models and build the next generation of Meta AI and consumer products.”

Satya Nadella, executive chairman and CEO of Microsoft: “We are committed to offering our customers the most advanced infrastructure to power their AI workloads. By bringing the GB200 Grace Blackwell processor to our datacenters globally, we are building on our long-standing history of optimizing NVIDIA GPUs for our cloud, as we make the promise of AI real for organizations everywhere.”

Sam Altman, CEO of OpenAI: “Blackwell offers massive performance leaps, and will accelerate our ability to deliver leading-edge models. We’re excited to continue working with NVIDIA to enhance AI compute.”

Larry Ellison, chairman and CTO of Oracle: “Oracle’s close collaboration with NVIDIA will enable qualitative and quantitative breakthroughs in AI, machine learning and data analytics. In order for customers to uncover more actionable insights, an even more powerful engine like Blackwell is needed, which is purpose-built for accelerated computing and generative AI.”

Elon Musk, CEO of Tesla and xAI: “There is currently nothing better than NVIDIA hardware for AI.”

Named in honor of David Harold Blackwell—a mathematician who specialized in game theory and statistics, and the first Black scholar inducted into the National Academy of Sciences—the new architecture succeeds the NVIDIA Hopper™ architecture, launched two years ago.

Key Features to Fuel Accelerated Computing and Generative AI

TensorRT-LLM—AI technologies, which together enable AI training and real-time LLM inference for models scaling up to 100x faster.

Full Chip—Packed with 208 billion transistors, Blackwell-architecture GPUs are manufactured using a 4nm process with two-reticle limit GPU dies connected by 10 TB/second chip-to-chip link into a single, unified die.

Transformer Engine—Fueled by new micro-tensor scaling support and NVIDIA’s advanced dynamic range, the Transformer Engine is integrated into NVIDIA TensorRT™-LLM and NeMo Megatron frameworks. Blackwell will support double the model sizes with new 4-bit floating point AI inference capabilities.

NVLink—To accelerate performance for multitrillion-parameter and mixture-of-experts AI models, the latest NVLink™ delivers groundbreaking 1.8TB/s bidirectional throughput per GPU, ensuring seamless high-speed networking up to 576 GPUs for the most complex LLMs.

Blackwell-powered GPUs include a dedicated engine for reliability, availability and serviceability. Additionally, the new architecture adds capabilities at the chip level to utilize AI-based preventative maintenance to run diagnostics and forecast system uptime and improves resiliency for massive-scale AI deployments to run uninterrupted for years at a time and to reduce operating costs.

Advanced confidential computing capabilities protect AI models and customer data without compromising performance. Confidential interface encryption protocols, which are critical for privacy-sensitive industries like healthcare and financial services, are also supported.

BlueField™-3—A dedicated decompression engine supports the latest formats, accelerating database queries to deliver 10x performance in data analytics and data science. In the coming years, data processing, on which companies spend tens of billions annually, will be increasingly GPU-accelerated.

GB200 NVL72—Blackwell Superchip connects two NVIDIA B200 Tensor Core GPUs to the NVIDIA Grace CPU over a 900GB/s chip-to-chip interconnect.

Quantum-X800—In addition, GB200-powered systems can be connected with the NVIDIA Quantum-X800 InfiniBand and Spectrum-X, also announced today, which deliver advanced networking at speeds up to 800Gb/s.

GB200 NVL72—A component of the NVIDIA GB200 NVL72, a multi-node, liquid-cooled, rack-scale system for the most demanding AI workloads. It combines 36 Grace Blackwell Superchips, which include 72 Blackwell GPUs and 36 Grace CPUs, with 36 Grace CPUs generation NVLink. Additionally, GB200 NVL72 includes NVIDIA BlueField™-3 data processing units to enable on-chip, composable storage, zero-trust security and GPU compute elasticity in hyperscale AI clouds. The GB200 NVL72 reduces cost and energy consumption by up to 25x.

GB200 NVL72—A single GPU with 1.4 exaflops of AI performance and 30TB of fast memory, and is a building block for the newest generation of AI systems.

GB200 NVL72—A server board that links eight B200 GPUs through NVLink to support x86-based generative AI platforms, offering speeds up to 400Gb/s through the NVIDIA Quantum-2 InfiniBand and Spectrum-X Ethernet networking.

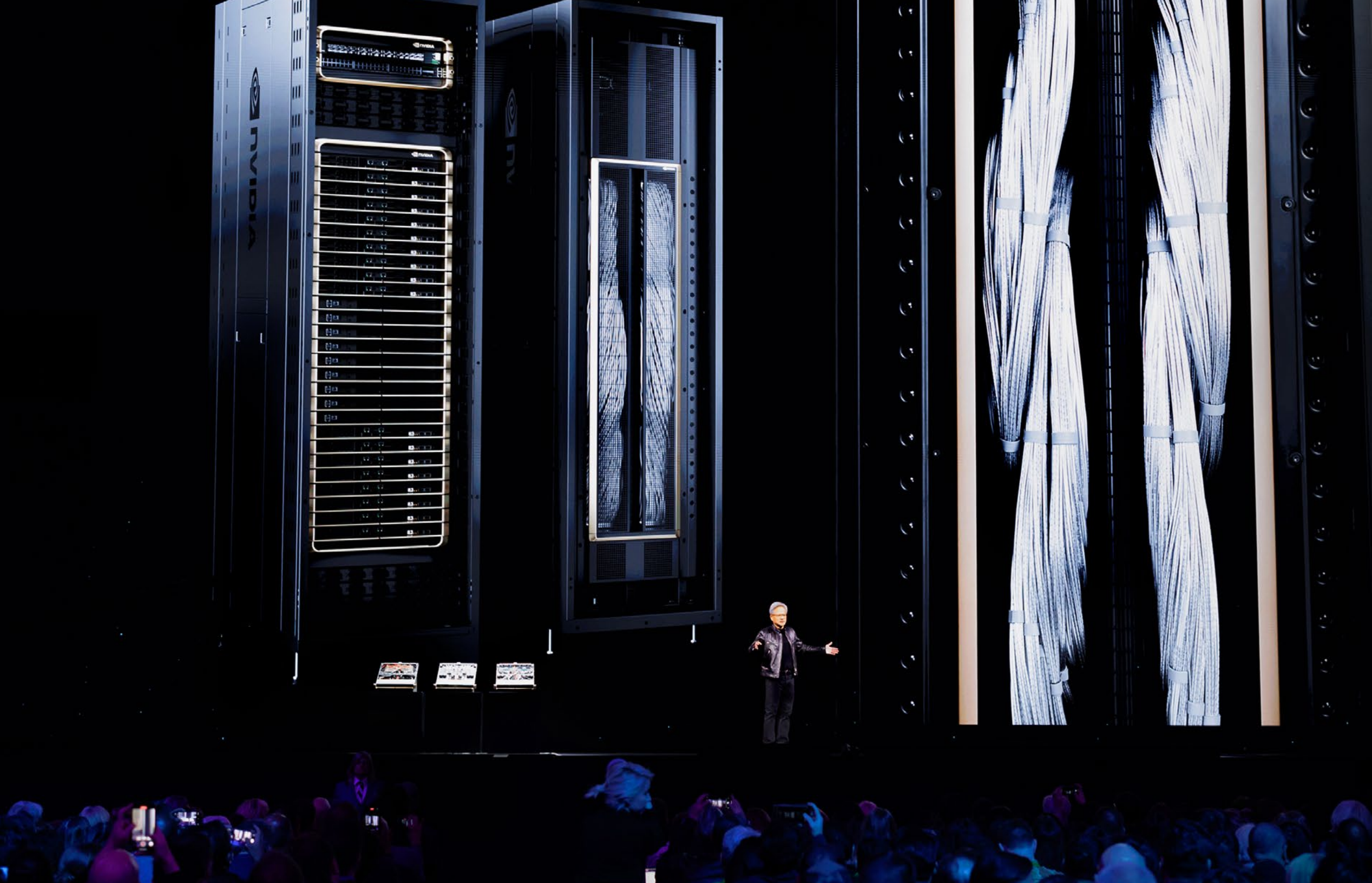
Blackwell Partners

Blackwell Superchips will be available from partners starting later this year.

Microsoft Azure and Oracle Cloud Infrastructure will be among the first cloud service providers to offer Blackwell-powered AI services, as will NVIDIA Cloud Partner program companies Applied Digital, CoreWeave, Crusoe, IBM Cloud and others. AWS will also provide Blackwell-based cloud services and infrastructure, including Indosat Ooredoo Hutchinson, Oracle EU Sovereign Cloud, the Oracle US, UK, and Australian Government Clouds, Scaleway, Singtel, Northern Data, Yotta Data Services, Shakti Cloud and YTL Power International.

Blackwell will also be available on NVIDIA DGX™ Cloud, an AI platform co-engineered with leading cloud service providers that gives customers direct access to the infrastructure and software needed to build and deploy advanced generative AI models. NVIDIA Cloud Infrastructure plan to host new NVIDIA Grace Blackwell-based instances later this year.

Blackwell-powered servers are expected to be available from a wide range of partners, including Dell Enterprise, Lenovo and Supermicro are expected to deliver a wide range of servers based on Blackwell-powered GPUs, as are Aivres, ASRock Rack, ASUS, Eviden, Foxconn, GIGABYTE, Inventec, Pegatron, QCT, Wistron, Wiyynn and ZT Systems.



“NVIDIA moved the ball forward with the latest iteration of its speedy NVLink technology.”

HPCWire

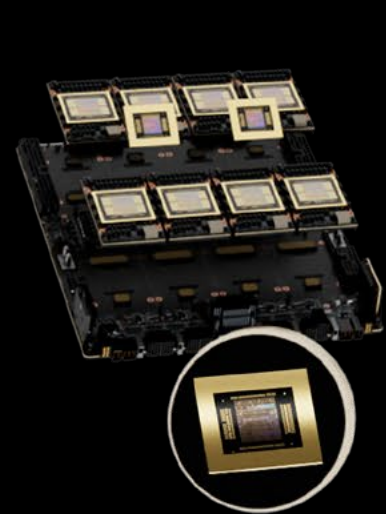
To accelerate performance for multitrillion-parameter and mixture-of-experts AI models, the latest iteration of NVIDIA NVLink delivers groundbreaking 1.8TB/s bidirectional throughput per GPU. This ensures seamless high-speed communication among up to 576 GPUs for the most complex LLMs.



**“The new AI platform
could be a game changer.”**

Barron's

Generative AI is the defining technology of our time. Blackwell is the engine to power this new industrial revolution. Working with the most dynamic companies in the world, we'll realize the promise of AI for every industry.



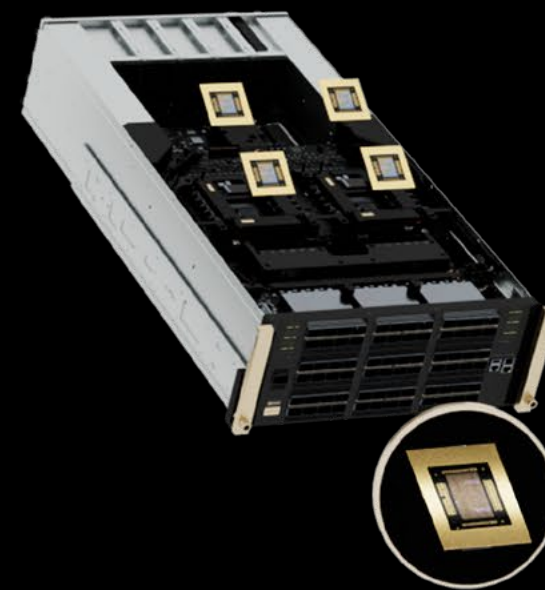
HGX B100



NVLink Switch



GB200 Superchip
Compute Node



Quantum X800 Switch
ConnectX-8 SuperNIC



Spectrum X800 Switch
BlueField-3 SuperNIC

“Widespread adoption anticipated.”

VentureBeat

Among the many organizations expected to adopt Blackwell are Amazon Web Services, Dell Technologies, Google, Meta, Microsoft, OpenAI, Oracle, Tesla, and xAI.

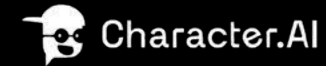


Google Cloud



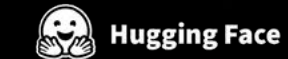
A D E P T

AI21 labs



cohere

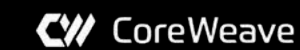
essential AI

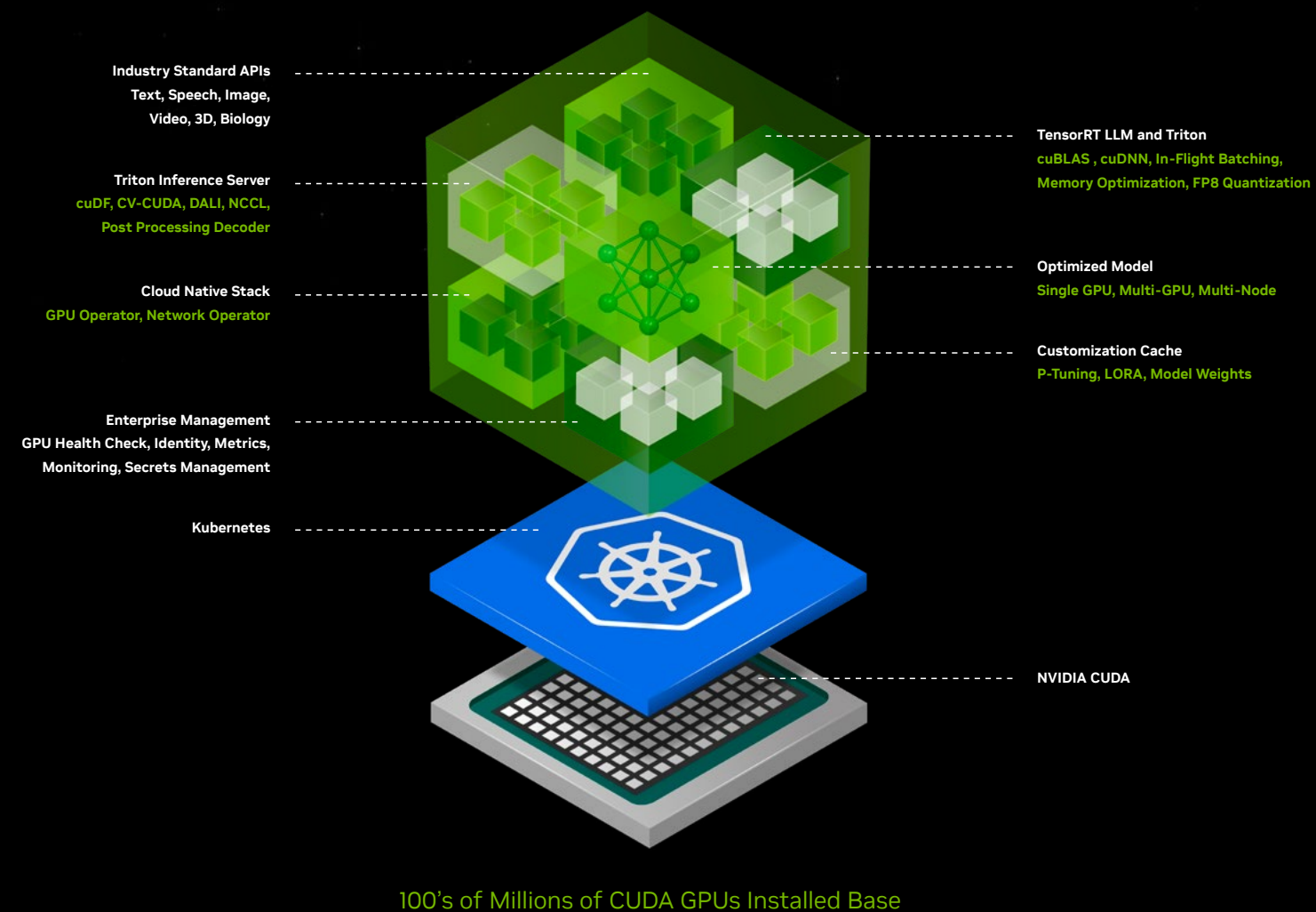


Inflection



together.ai





**“NVIDIA launches NIM to
make it smoother to deploy
AI models into production.”**

TechCrunch

NVIDIA Inference Microservices are a new way to package and deliver AI software. The curated selection of microservices adds a new layer to NVIDIA's full-stack computing platform. This layer connects the AI ecosystem of model developers, platform providers, and enterprises with a standardized path to run custom AI models.



“NVIDIA has virtually recreated the entire planet—and now it wants to use its digital twin to crack weather forecasting for good.”

TechRadar



To help combat the \$140 billion in economic losses due to extreme weather brought on by climate change, we announced the Earth-2 digital twin cloud platform for simulating and visualizing weather and climate at unprecedented scale.

**“NVIDIA packages inference
to deliver generative AI
for healthcare.”**

Electronics Weekly

The new suite of NIM for healthcare offers advanced imaging, natural language and speech recognition, and digital biology generation, prediction, and simulation.

“NVIDIA is working to bring AI robots to life.”

Barron's

We announced a collection of robotics pretrained models, libraries, and reference hardware. We also announced Project GR00T, a general-purpose foundation model for humanoid robots, designed to further our work driving breakthroughs in robotics and embodied AI.



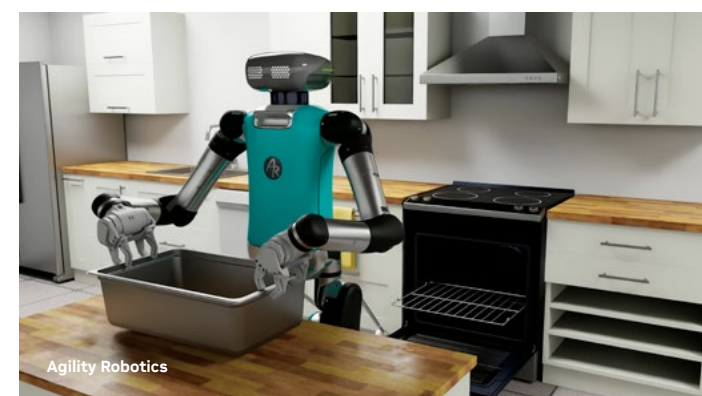
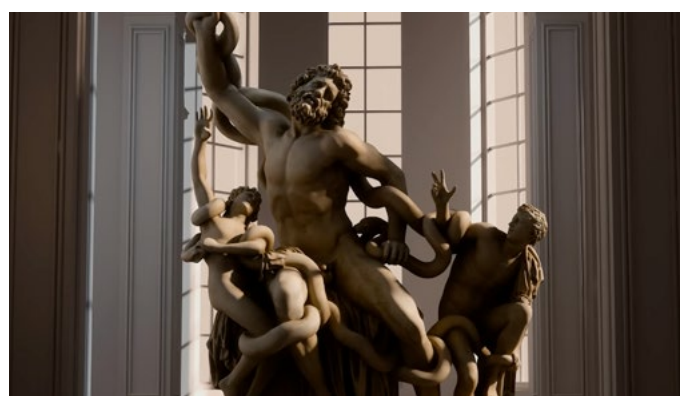
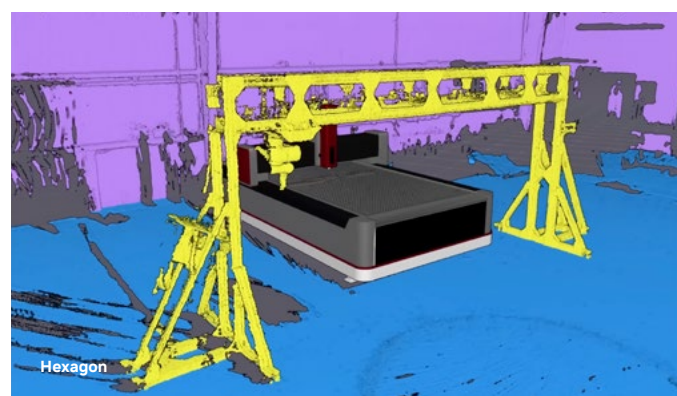
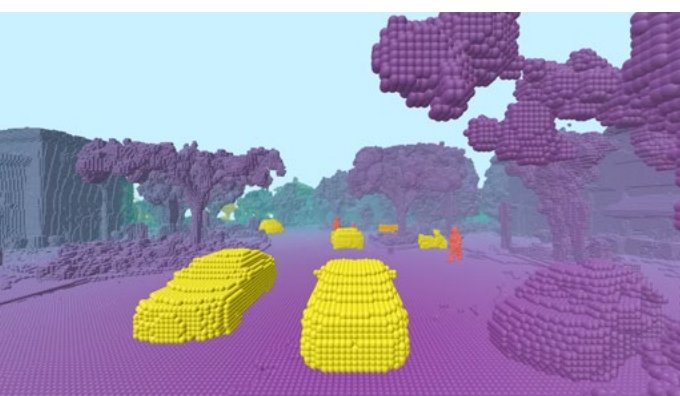
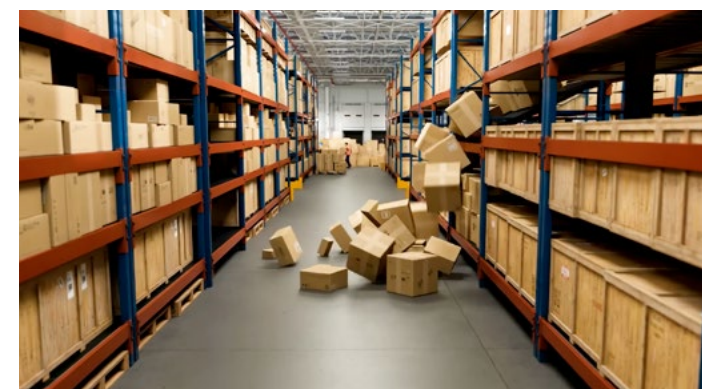
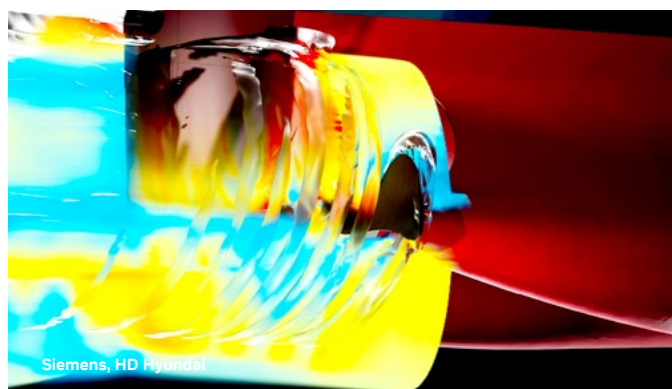
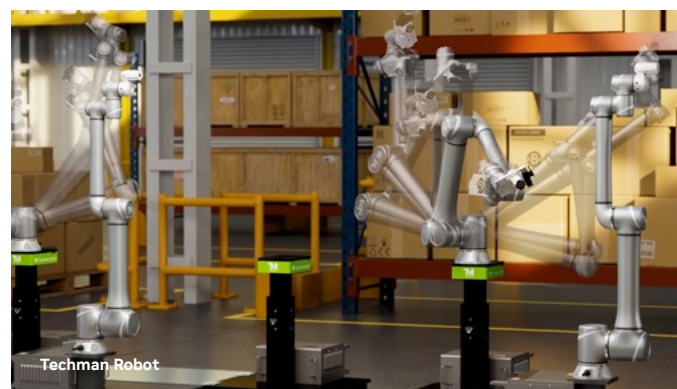
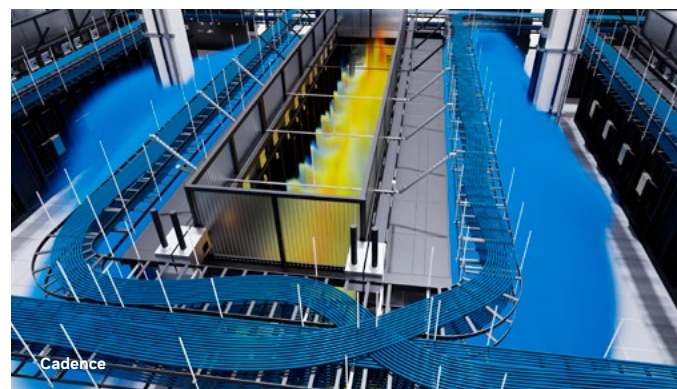
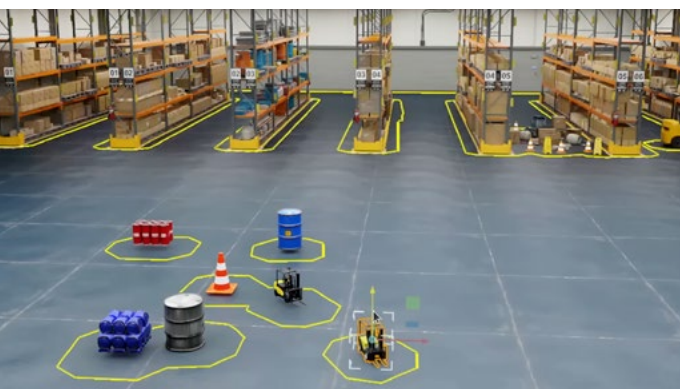
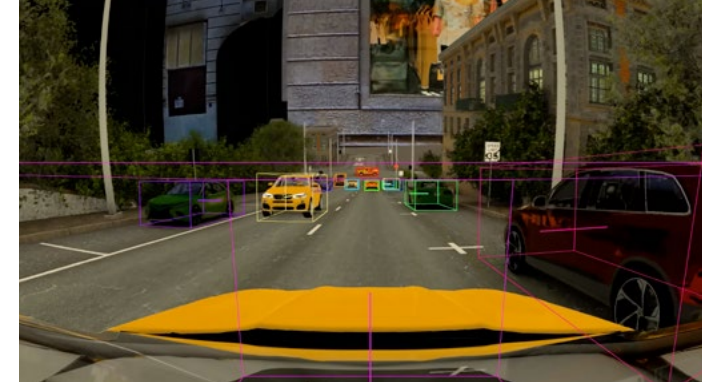
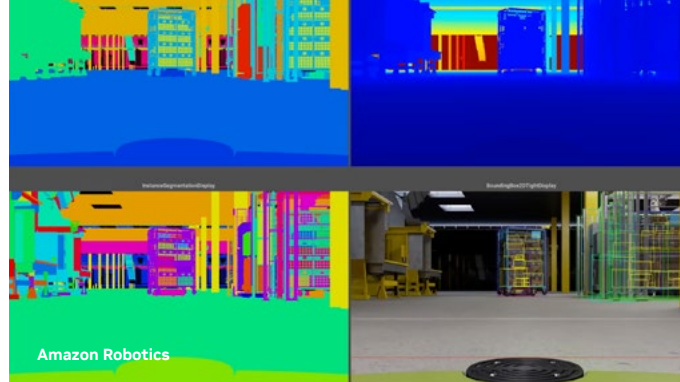
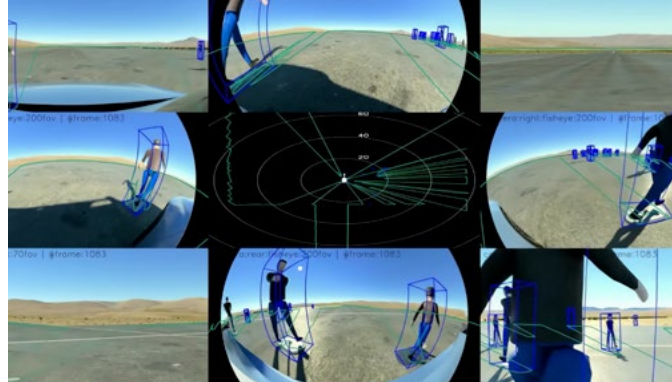
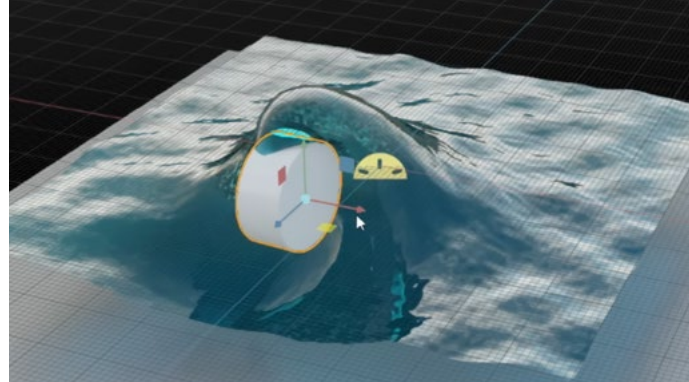
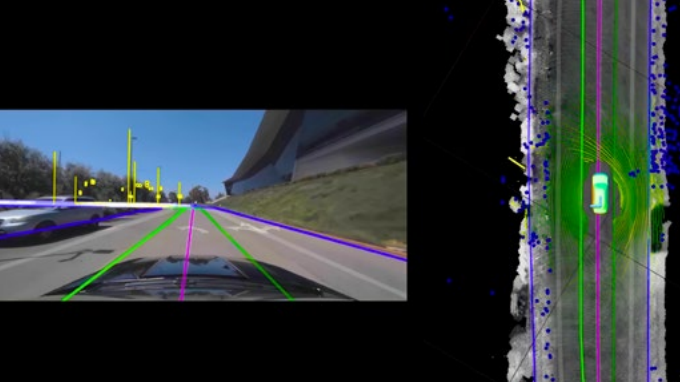
**“NVIDIA Omniverse
Cloud APIs will elevate
digital twins for a new
industrial revolution.”**

VentureBeat

NVIDIA Omniverse Cloud will be available as APIs, extending the reach of the world's leading platform for creating industrial digital twin applications and workflows across the entire ecosystem of software makers.

Siemens





What Analysts Said...

“Move over
Taylor Swift...”

Bernstein

“The soothsayer of Santa Clara
did not disappoint.”

Cantor

“[NVIDIA] sits on the
cusp of an entirely new wave
of demand.”

UBS

“The leader in AI showcasing innovation
across the full stack
of accelerated computing.”

Cowen

“No one else in the industry
can match this capability.”

Melius

“[NVIDIA]’s platform expansion
is remarkable...period!”

Wells Fargo

What Press Said...

“NVIDIA’s CEO painted a vision of AI turbocharging computing power.”

The New York Times

“NVIDIA expects to win outsized chunk of data center spending.”

Bloomberg

“Event was a daunting reminder of the speed at which NVIDIA is moving.”

Financial Times

“With Blackwell GPUs, AI gets cheaper and easier, competing with NVIDIA gets harder.”

The Next Platform

“The NVIDIA frenzy over artificial intelligence has come to this: AI Woodstock.”

The Wall Street Journal

“GTC can serve as a preview of where the entire field is going.”

Fast Company





GTC

The Conference for the Era of AI

SAN JOSE McENERY CONVENTION CENTER

#GTC24

*"GTC exists to inspire
the world on the
art-of-the-possible."*



Explore and Share

@nvidia #gtc24

A large crowd of people is gathered at a conference. The background features bright green walls and several large digital displays showing various content, including a detailed image of Earth from space. People are wearing lanyards and some have backpacks. The atmosphere appears busy and professional.

300,000

Registrations

19,000

In-Person Attendees

1,100

Sessions

33M

Keynote Views

29,000

Press Articles

300

Partner Sponsorships



**“GTC: Spearheading
AI and accelerated
computing innovation.”**

siliconANGLE

Transforming AI Panel: the authors of the seminal research paper—*Attention Is All You Need*—that introduced the transformer neural network architecture came together at GTC.



 NVIDIA
GTC



**“NVIDIA’s GTC conference,
not surprisingly, was an
absolute whirlwind.”**

VentureBeat





“GTC 2024 was the single most important event in the history of the technology industry.”

SiliconANGLE theCUBE





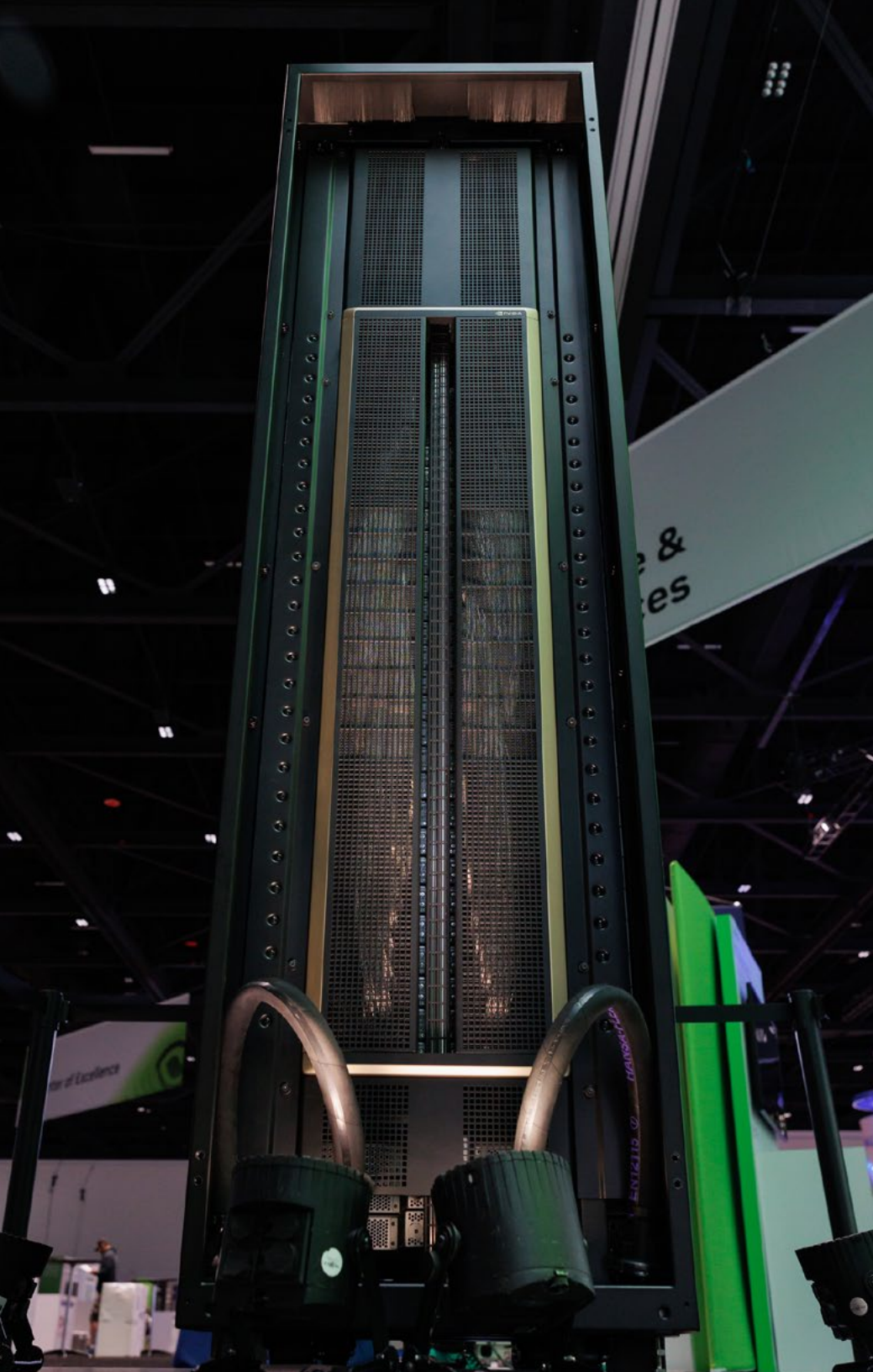


**“The Generative AI
Future Is Now.”**

HPCWire

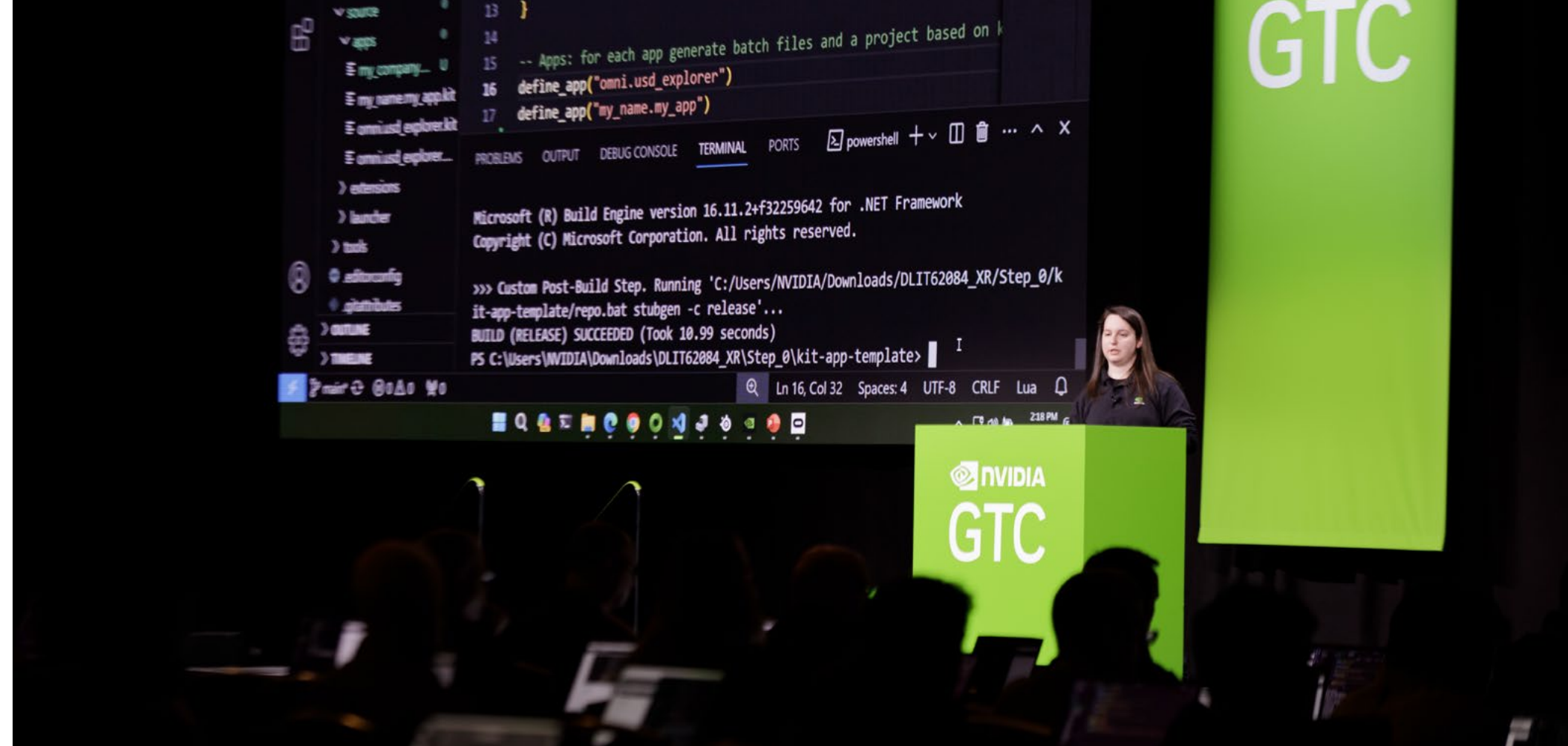




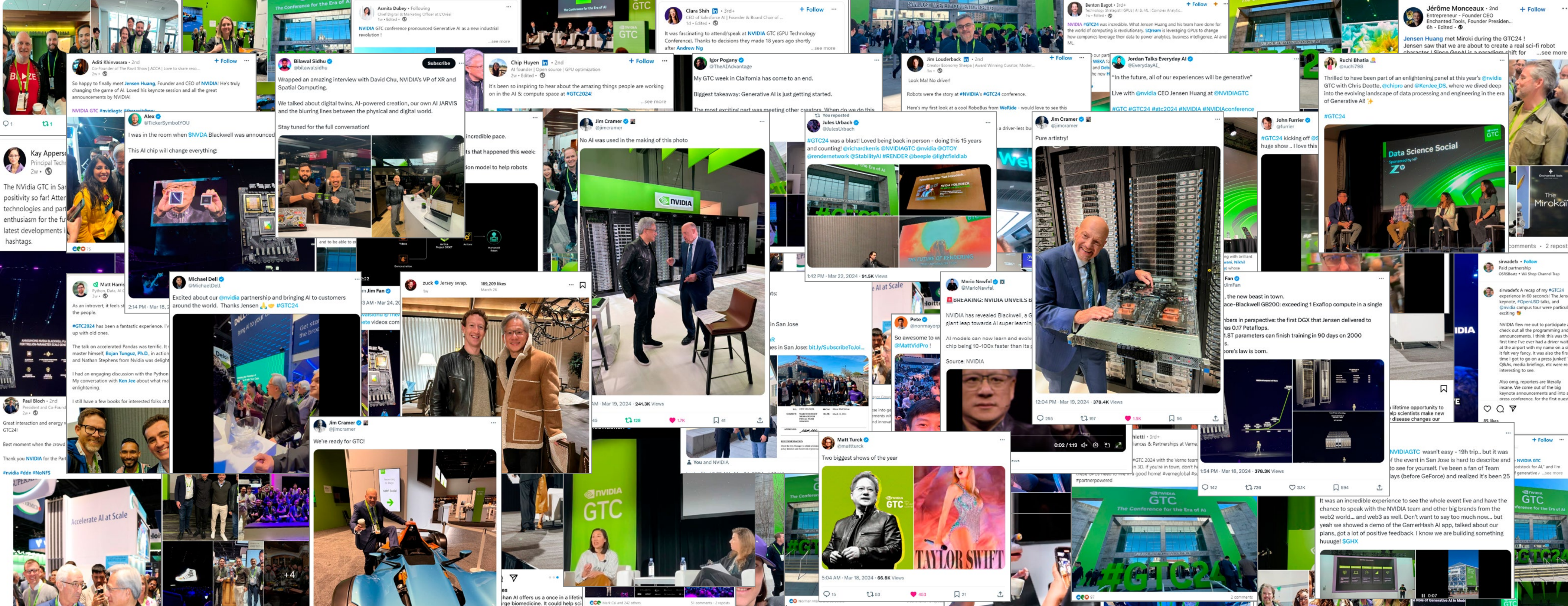










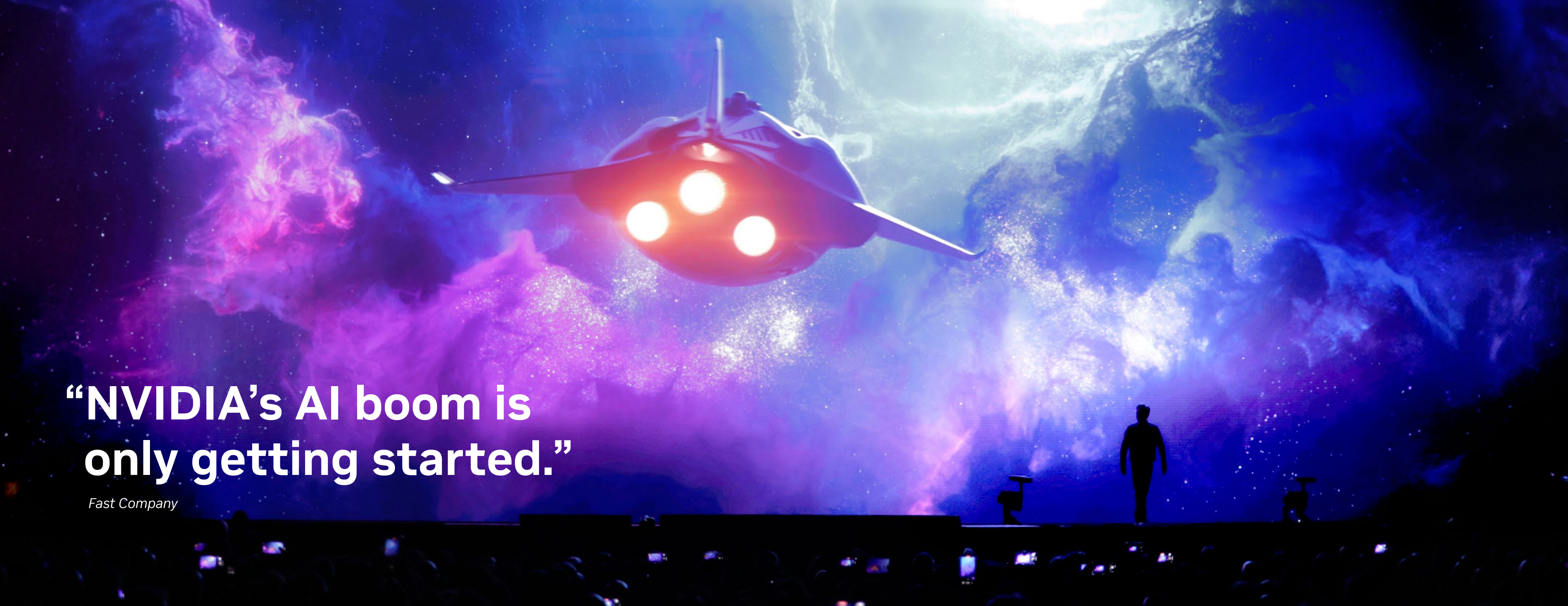




“For three decades we’ve pursued accelerated computing, with the goal of enabling transformative breakthroughs like deep learning and AI. Generative AI is the defining technology of our time.

Blackwell is the engine to power this new industrial revolution. Working with the most dynamic companies in the world, we will realize the promise of AI for every industry.”

Jensen Huang



**“NVIDIA’s AI boom is
only getting started.”**

Fast Company

Except for the historical information contained herein, certain matters in this presentation including, but not limited to, statements as to the benefits, features, availability, and impact of NVIDIA's products and technologies; NVIDIA powering a new industrial revolution; NVIDIA realizing the promise of AI for every industry while working with the most dynamic companies in the world; and third parties using or adopting of NVIDIA products, technologies and platforms, and the benefits and impacts thereof are forward-looking statements. These forward-looking statements and any other forward-looking statements that go beyond historical facts that are made in this presentation are subject to risks and uncertainties that may cause actual results to differ materially. Important factors that could cause actual results to differ materially include: global economic conditions; NVIDIA's reliance on third parties to manufacture, assemble, package and test NVIDIA's products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA's existing product and technologies; market acceptance of NVIDIA's products or NVIDIA's partners' products; design, manufacturing or software defects; changes in consumer preferences and demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA's products or technologies when integrated into systems and other factors.

NVIDIA has based these forward-looking statements largely on its current expectations and projections about future events and trends that it believes may affect its financial condition, results of operations, business strategy, short-term and long-term business operations and objectives, and financial needs. These forward-looking statements are subject to a number of risks and uncertainties, and you should not rely upon the forward-looking statements as predictions of future events. The future events and trends discussed in this presentation may not occur and actual results could differ materially and adversely from those anticipated or implied in the forward-looking statements. Although NVIDIA believes that the expectations reflected in the forward-looking statements are reasonable, the company cannot guarantee that future results, levels of activity, performance, achievements or events and circumstances reflected in the forward-looking statements will occur. Except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances. For a complete discussion of factors that could materially affect NVIDIA's financial results and operations, please refer to the reports NVIDIA files from time to time with the SEC, including NVIDIA's most recent Annual Report on Form 10-K, Quarterly Reports on Form 10-Q, and Current Reports on Form 8-K. Copies of reports NVIDIA files with the SEC are posted on NVIDIA's website and are available from NVIDIA without charge.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements within are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.