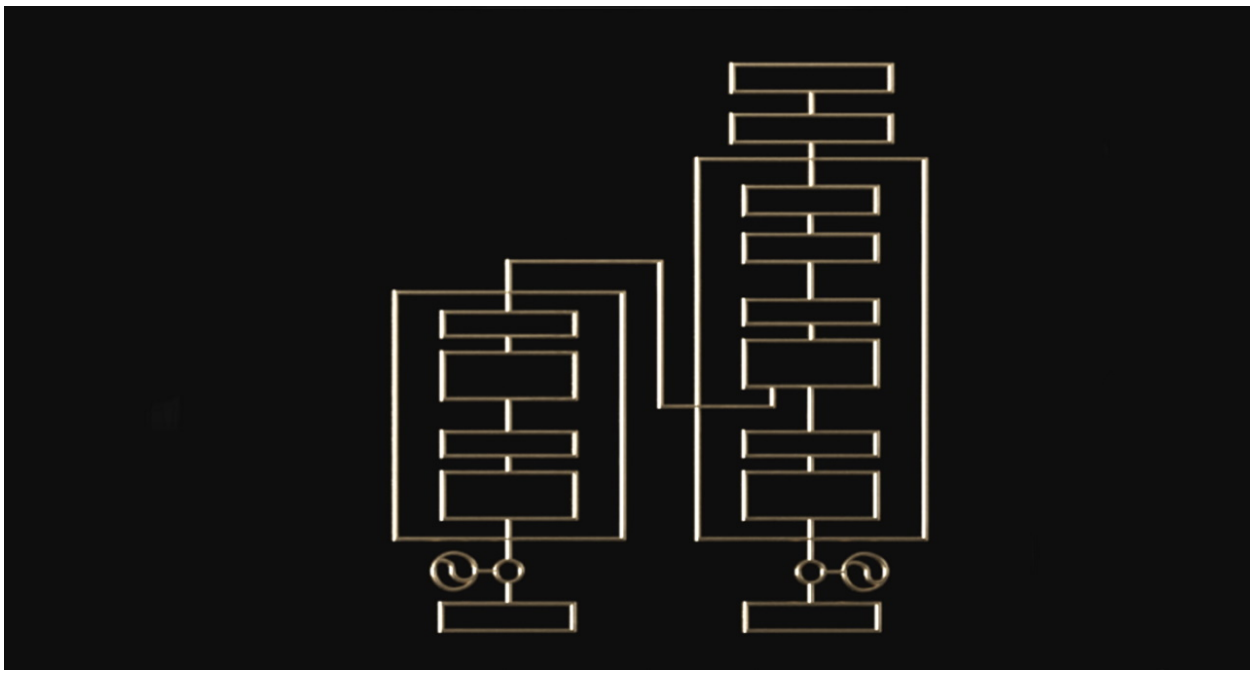
What Is a Transformer Model?

blogs.nvidia.com/blog/what-is-a-transformer-model

March 25, 2022

A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence.

March 25, 2022 by Rick Merritt



If you want to ride the next big wave in AI, grab a transformer.

They're not the shape-shifting toy robots on TV or the trash-can-sized tubs on telephone poles.

So, What's a Transformer Model?

A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence.

Transformer models apply an evolving set of mathematical techniques, called attention or self-attention, to detect subtle ways even distant data elements in a series influence and depend on each other.

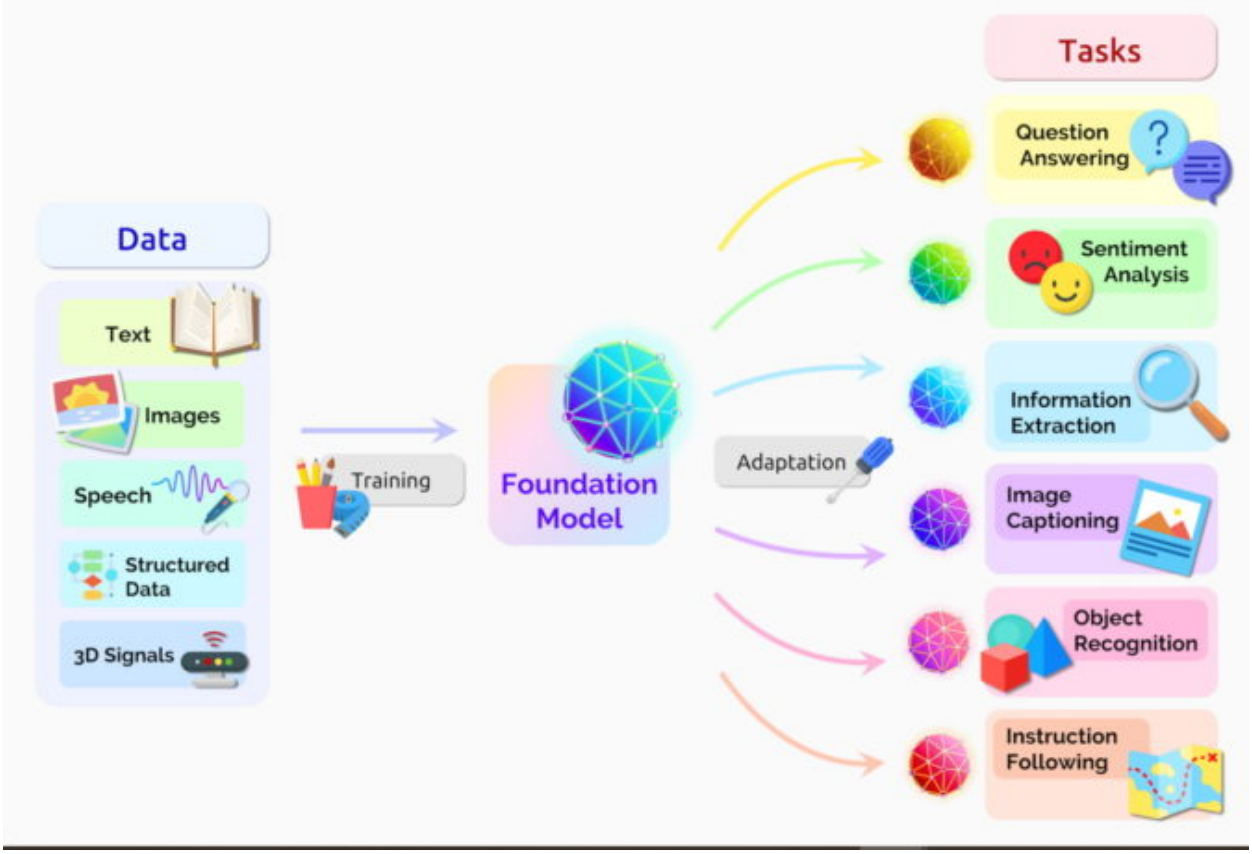
First described in a 2017 paper from Google, transformers are among the newest and one of the most powerful classes of models invented to date. They're driving a wave of advances in machine learning some have dubbed transformer AI.

Stanford researchers called transformers “foundation models” in an August 2021 paper because they see them driving a paradigm shift in AI. The “sheer scale and scope of foundation models over the last few years have stretched our imagination of what is possible,” they wrote.

What Can Transformer Models Do?

Transformers are translating text and speech in near real-time, opening meetings and classrooms to diverse and hearing-impaired attendees.

They’re helping researchers understand the chains of genes in DNA and amino acids in proteins in ways that can speed drug design.



Transformers, sometimes called foundation models, are already being used with many data sources for a host of applications.

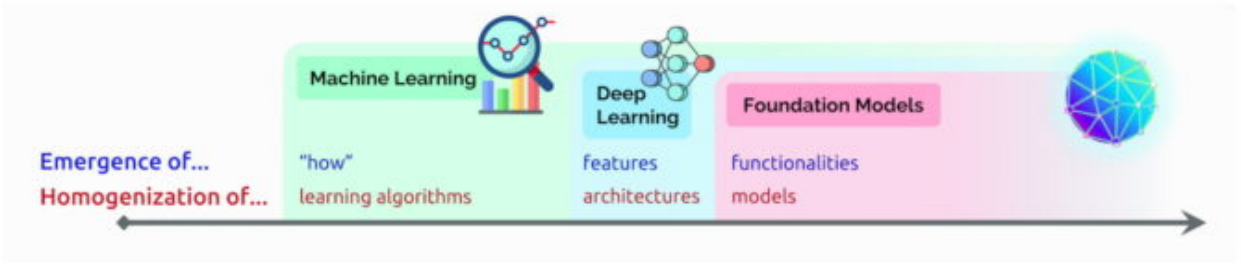
Transformers can detect trends and anomalies to prevent fraud, streamline manufacturing, make online recommendations or improve healthcare.

People use transformers every time they search on Google or Microsoft Bing.

The Virtuous Cycle of Transformer AI

Any application using sequential text, image or video data is a candidate for transformer models.

That enables these models to ride a virtuous cycle in transformer AI. Created with large datasets, transformers make accurate predictions that drive their wider use, generating more data that can be used to create even better models.



Stanford researchers say transformers mark the next stage of AI's development, what some call the era of transformer AI.

"Transformers made self-supervised learning possible, and AI jumped to warp speed," said NVIDIA founder and CEO Jensen Huang in his keynote address this week at GTC.

Transformers Replace CNNs, RNNs

Transformers are in many cases replacing convolutional and recurrent neural networks (CNNs and RNNs), the most popular types of deep learning models just five years ago.

Indeed, 70 percent of arXiv papers on AI posted in the last two years mention transformers. That's a radical shift from a 2017 IEEE study that reported RNNs and CNNs were the most popular models for pattern recognition.

No Labels, More Performance

Before transformers arrived, users had to train neural networks with large, labeled datasets that were costly and time-consuming to produce. By finding patterns between elements mathematically, transformers eliminate that need, making available the trillions of images and petabytes of text data on the web and in corporate databases.

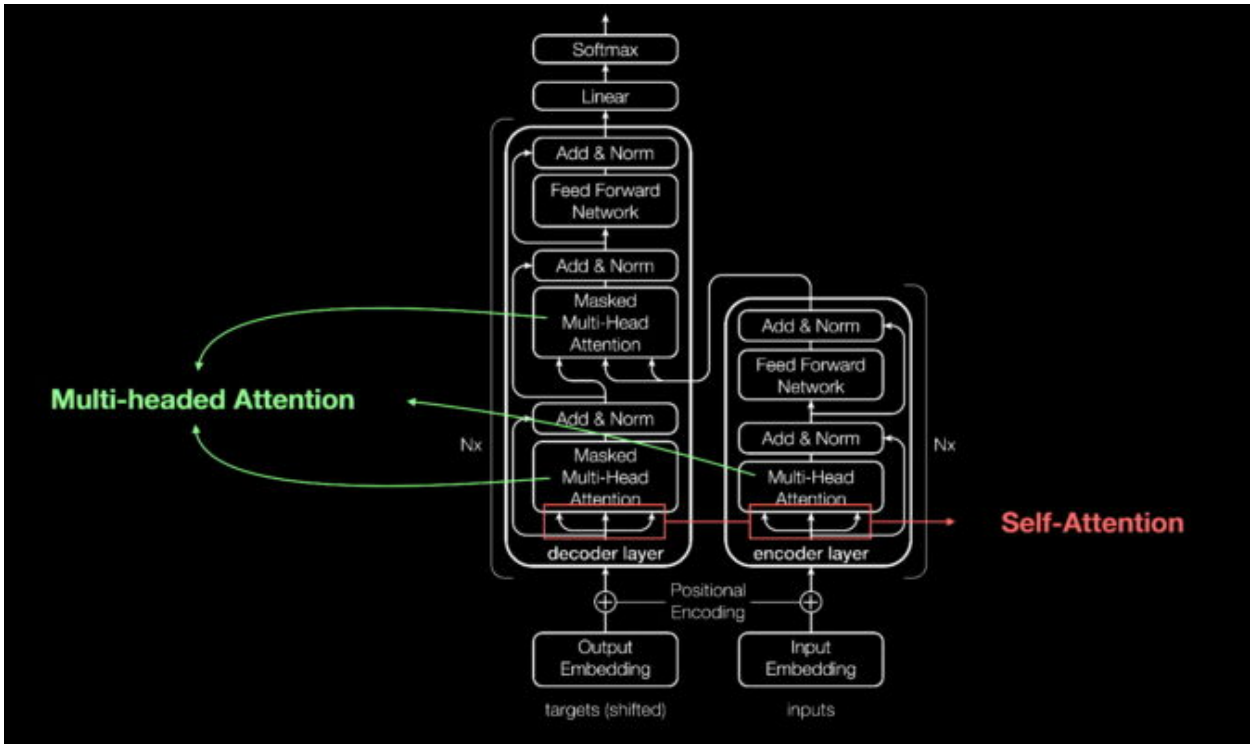
In addition, the math that transformers use lends itself to parallel processing, so these models can run fast.

Transformers now dominate popular performance leaderboards like SuperGLUE, a benchmark developed in 2019 for language-processing systems.

How Transformers Pay Attention

Like most neural networks, transformer models are basically large encoder/decoder blocks that process data.

Small but strategic additions to these blocks (shown in the diagram below) make transformers uniquely powerful.



A look under the hood from a presentation by Aidan Gomez, one of eight co-authors of the 2017 paper that defined transformers.

Transformers use positional encoders to tag data elements coming in and out of the network. Attention units follow these tags, calculating a kind of algebraic map of how each element relates to the others.

Attention queries are typically executed in parallel by calculating a matrix of equations in what's called multi-headed attention.

With these tools, computers can see the same patterns humans see.

Self-Attention Finds Meaning

For example, in the sentence:

She poured water from the pitcher to the cup until it was full.

We know "it" refers to the cup, while in the sentence:

She poured water from the pitcher to the cup until it was empty.

We know “it” refers to the pitcher.

“Meaning is a result of relationships between things, and self-attention is a general way of learning relationships,” said Ashish Vaswani, a former senior staff research scientist at Google Brain who led work on the seminal 2017 paper.

“Machine translation was a good vehicle to validate self-attention because you needed short- and long-distance relationships among words,” said Vaswani.

“Now we see self-attention is a powerful, flexible tool for learning,” he added.

How Transformers Got Their Name

Attention is so key to transformers the Google researchers almost used the term as the name for their 2017 model. Almost.

“Attention Net didn’t sound very exciting,” said Vaswani, who started working with neural nets in 2011.

.Jakob Uszkoreit, a senior software engineer on the team, came up with the name Transformer.

“I argued we were transforming representations, but that was just playing semantics,” Vaswani said.

The Birth of Transformers

In the paper for the 2017 NeurIPS conference, the Google team described their transformer and the accuracy records it set for machine translation.

Thanks to a basket of techniques, they trained their model in just 3.5 days on eight NVIDIA GPUs, a small fraction of the time and cost of training prior models. They trained it on datasets with up to a billion pairs of words.

“It was an intense three-month sprint to the paper submission date,” recalled Aidan Gomez, a Google intern in 2017 who contributed to the work.

“The night we were submitting, Ashish and I pulled an all-nighter at Google,” he said. “I caught a couple hours sleep in one of the small conference rooms, and I woke up just in time for the submission when someone coming in early to work opened the door and hit my head.”

It was a wakeup call in more ways than one.

“Ashish told me that night he was convinced this was going to be a huge deal, something game changing. I wasn’t convinced, I thought it would be a modest gain on a benchmark, but it turned out he was very right,” said Gomez, now CEO of startup Cohere that’s providing a language processing service based on transformers.

A Moment for Machine Learning

Vaswani recalls the excitement of seeing the results surpass similar work published by a Facebook team using CNNs.

“I could see this would likely be an important moment in machine learning,” he said.

A year later, another Google team tried processing text sequences both forward and backward with a transformer. That helped capture more relationships among words, improving the model’s ability to understand the meaning of a sentence.

Their Bidirectional Encoder Representations from Transformers (BERT) model set 11 new records and became part of the algorithm behind Google search.

Within weeks, researchers around the world were adapting BERT for use cases across many languages and industries “because text is one of the most common data types companies have,” said Anders Arpteg, a 20-year veteran of machine learning research.

Putting Transformers to Work

Soon transformer models were being adapted for science and healthcare.

DeepMind, in London, advanced the understanding of proteins, the building blocks of life, using a transformer called AlphaFold2, described in a recent Nature article. It processed amino acid chains like text strings to set a new watermark for describing how proteins fold, work that could speed drug discovery.

AstraZeneca and NVIDIA developed MegaMolBART, a transformer tailored for drug discovery. It’s a version of the pharmaceutical company’s MolBART transformer, trained on a large, unlabeled database of chemical compounds using the NVIDIA Megatron framework for building large-scale transformer models.

Reading Molecules, Medical Records

“Just as AI language models can learn the relationships between words in a sentence, our aim is that neural networks trained on molecular structure data will be able to learn the relationships between atoms in real-world molecules,” said Ola Engkvist, head of molecular AI, discovery sciences and R&D at AstraZeneca, when the work was announced last year.

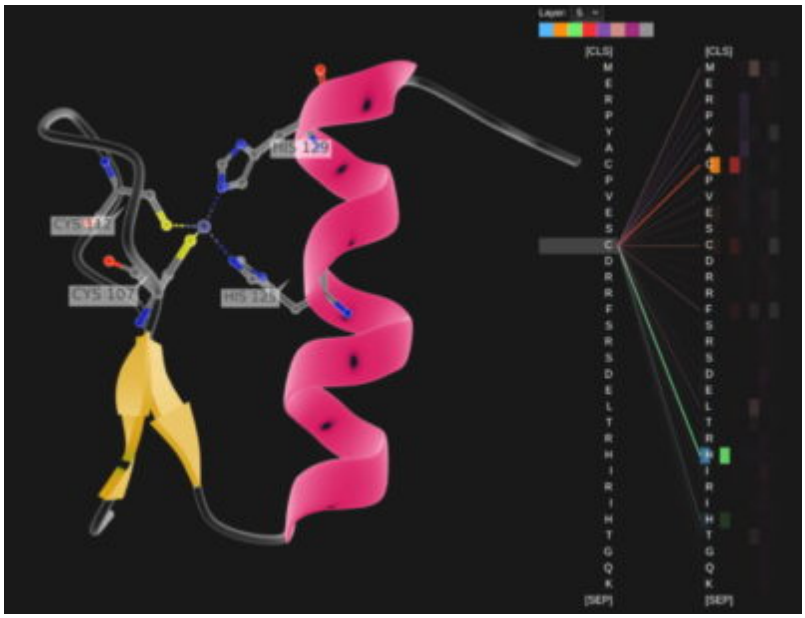
Watch Video At: https://youtu.be/wBgpMf_KQVw

Separately, the University of Florida's academic health center collaborated with NVIDIA researchers to create GatorTron. The transformer model aims to extract insights from massive volumes of clinical data to accelerate medical research.

Transformers Grow Up

Along the way, researchers found larger transformers performed better.

For example, researchers from the Rostlab at the Technical University of Munich, which helped pioneer work at the intersection of AI and biology, used natural-language processing to understand proteins. In 18 months, they graduated from using RNNs with 90 million parameters to transformer models with 567 million parameters.



Rostlab researchers show language models trained without labeled samples picking up the signal of a protein sequence.

The OpenAI lab showed bigger is better with its Generative Pretrained Transformer (GPT). The latest version, GPT-3**,** has 175 billion parameters, up from 1.5 billion for GPT-2.

With the extra heft, GPT-3 can respond to a user's query even on tasks it was not specifically trained to handle. It's already being used by companies including Cisco, IBM and Salesforce.

Tale of a Mega Transformer

NVIDIA and Microsoft hit a high watermark in November, announcing the Megatron-Turing Natural Language Generation model (MT-NLG) with 530 billion parameters. It debuted along with a new framework, NVIDIA NeMo Megatron, that aims to let any business create its own billion- or trillion-parameter transformers to power custom chatbots, personal assistants and other AI applications that understand language.

MT-NLG had its public debut as the brain for TJ, the Toy Jensen avatar that gave part of the keynote at NVIDIA's November 2021 GTC.

"When we saw TJ answer questions — the power of our work demonstrated by our CEO — that was exciting," said Mostofa Patwary, who led the NVIDIA team that trained the model.

Creating such models is not for the faint of heart. MT-NLG was trained using hundreds of billions of data elements, a process that required thousands of GPUs running for weeks.

"Training large transformer models is expensive and time-consuming, so if you're not successful the first or second time, projects might be canceled," said Patwary.

"Megatron helps me answer all those tough questions Jensen throws at me," TJ said at GTC 2022.

Trillion-Parameter Transformers

Today, many AI engineers are working on trillion-parameter transformers and applications for them.

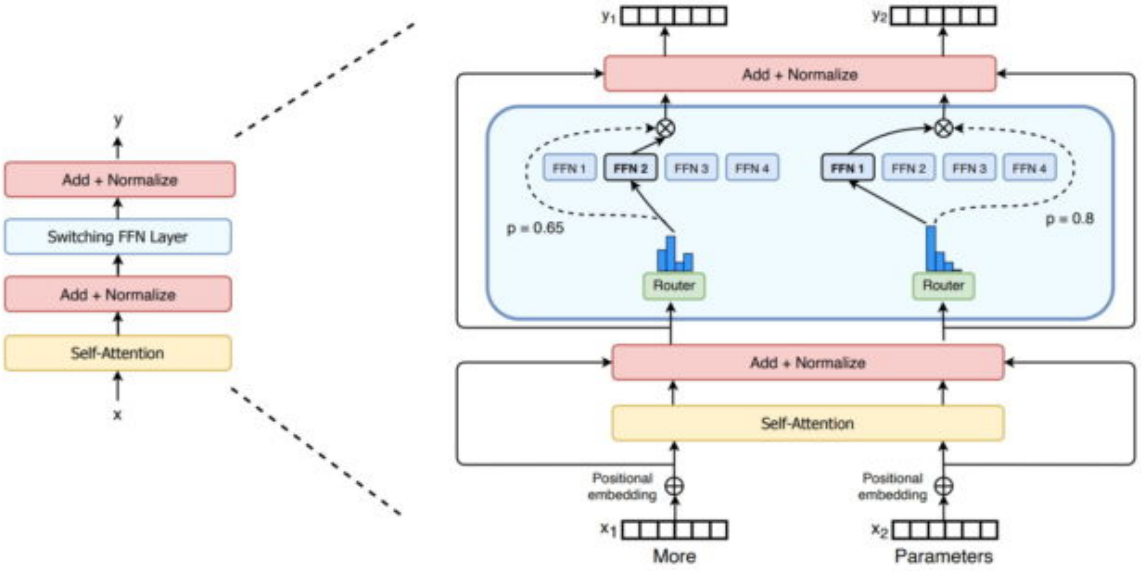
"We're constantly exploring how these big models can deliver better applications. We also investigate in what aspects they fail, so we can build even better and bigger ones," Patwary said.

To provide the computing muscle those models need, our latest accelerator — the NVIDIA H100 Tensor Core GPU — packs a Transformer Engine and supports a new FP8 format. That speeds training while preserving accuracy.

With those and other advances, "transformer model training can be reduced from weeks to days" said Huang at GTC.

MoE Means More for Transformers

Last year, Google researchers described the Switch Transformer, one of the first trillion-parameter models. It uses AI sparsity, a complex mixture-of experts (MoE) architecture and other advances to drive performance gains in language processing and up to 7x increases in pre-training speed.



The encoder for the Switch Transformer, the first model to have up to a trillion parameters.

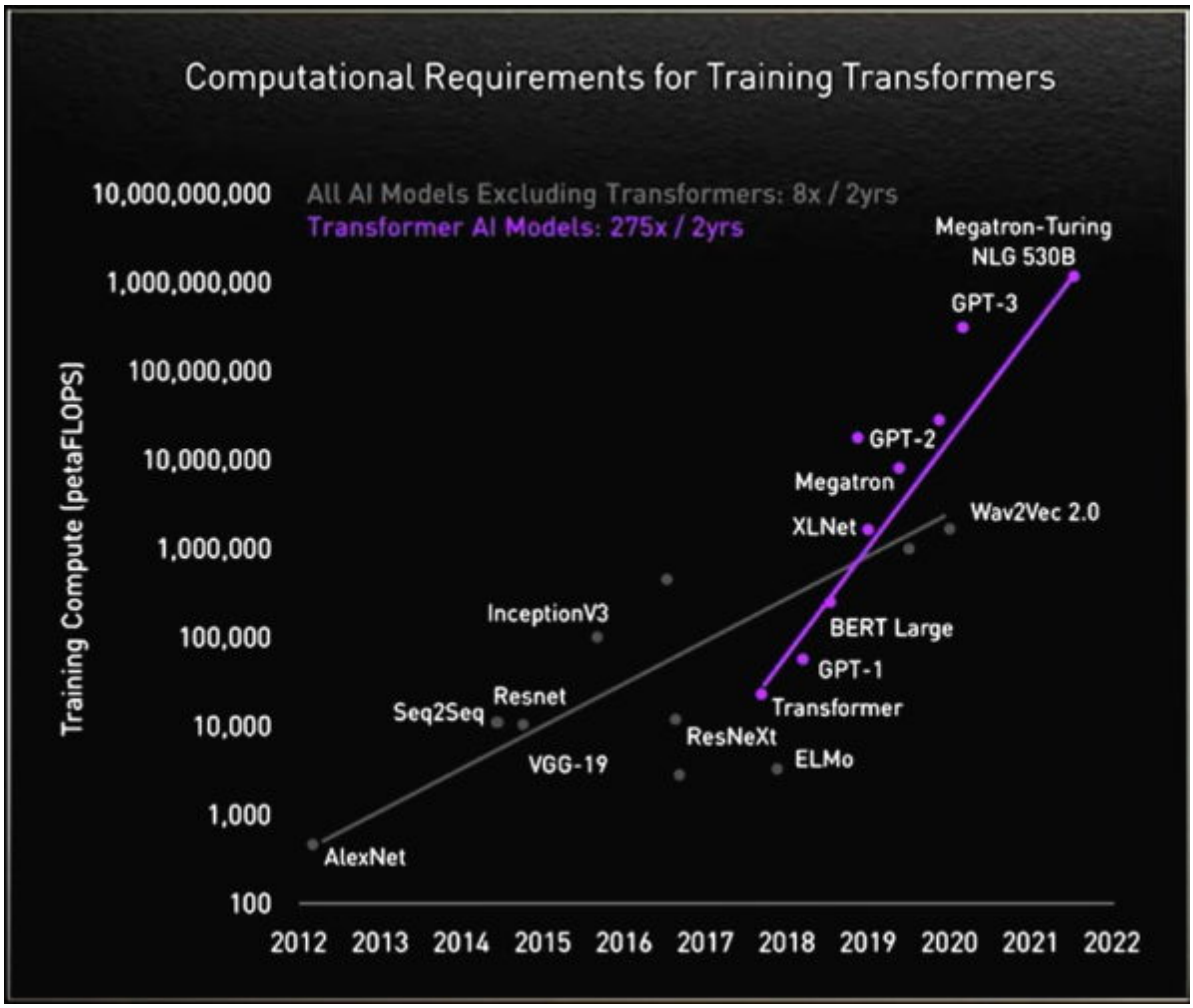
For its part, Microsoft Azure worked with NVIDIA to implement an MoE transformer for its Translator service.

Tackling Transformers' Challenges

Now some researchers aim to develop simpler transformers with fewer parameters that deliver performance similar to the largest models.

"I see promise in retrieval-based models that I'm super excited about because they could bend the curve," said Gomez, of Cohere, noting the Retro model from DeepMind as an example.

Retrieval-based models learn by submitting queries to a database. "It's cool because you can be choosy about what you put in that knowledge base," he said.



In the race for higher performance, transformer models have grown larger.

The ultimate goal is to “make these models learn like humans do from context in the real world with very little data,” said Vaswani, now co-founder of a stealth AI startup.

He imagines future models that do more computation upfront so they need less data and sport better ways users can give them feedback.

“Our goal is to build models that will help people in their everyday lives,” he said of his new venture.

Safe, Responsible Models

Other researchers are studying ways to eliminate bias or toxicity if models amplify wrong or harmful language. For example, Stanford created the Center for Research on Foundation Models to explore these issues.

“These are important problems that need to be solved for safe deployment of models,” said Shrimai Prabhumoye, a research scientist at NVIDIA who’s among many across the industry working in the area.

“Today, most models look for certain words or phrases, but in real life these issues may come out subtly, so we have to consider the whole context,” added Prabhumoye.

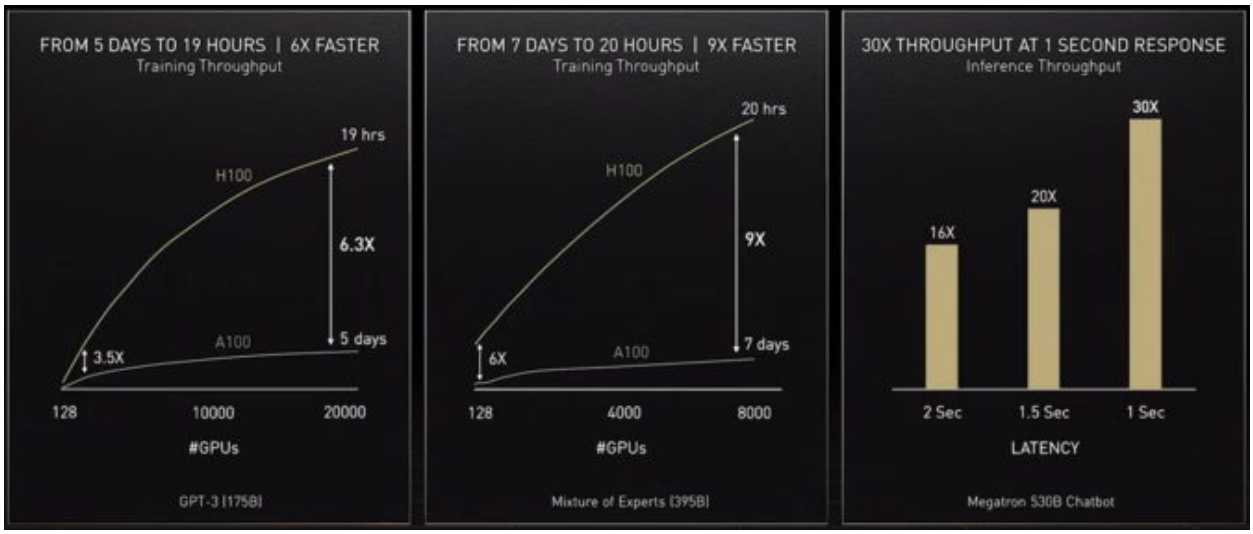
“That’s a primary concern for Cohere, too,” said Gomez. “No one is going to use these models if they hurt people, so it’s table stakes to make the safest and most responsible models.”

Beyond the Horizon

Vaswani imagines a future where self-learning, attention-powered transformers approach the holy grail of AI.

“We have a chance of achieving some of the goals people talked about when they coined the term ‘general artificial intelligence’ and I find that north star very inspiring,” he said.

“We are in a time where simple methods like neural networks are giving us an explosion of new capabilities.”



Transformer training and inference will get significantly accelerated with the NVIDIA H100 GPU.

Learn more about transformers on the NVIDIA Technical Blog.

Post navigation

Take Control This GFN Thursday With New Stratus+ Controller From SteelSeries

Plus, drop into a new season of ‘Fortnite,’ explore a unique quest for members in ‘MapleStory’ and stream six new titles this week.

March 24, 2022 by GeForce NOW Community

GeForce NOW gives you the power to game almost anywhere, at GeForce quality. And with the latest controller from SteelSeries, members can stay in control of the action on Android and Chromebook devices.

This GFN Thursday takes a look at the SteelSeries Stratus+, now part of the GeForce NOW Recommended program.

And it wouldn't be Thursday without new games, so get ready for six additions to the GeForce NOW library, including the latest season of *Fortnite* and a special in-game event for *MapleStory* that's exclusive for GeForce NOW members.

The Power to Play, in the Palm of Your Hand

GeForce NOW transforms mobile phones into powerful gaming computers capable of streaming PC games anywhere. The best mobile gaming sessions are backed by recommended controllers, including the new Stratus+ by SteelSeries.

Take control of how you play with the new SteelSeries Stratus+.

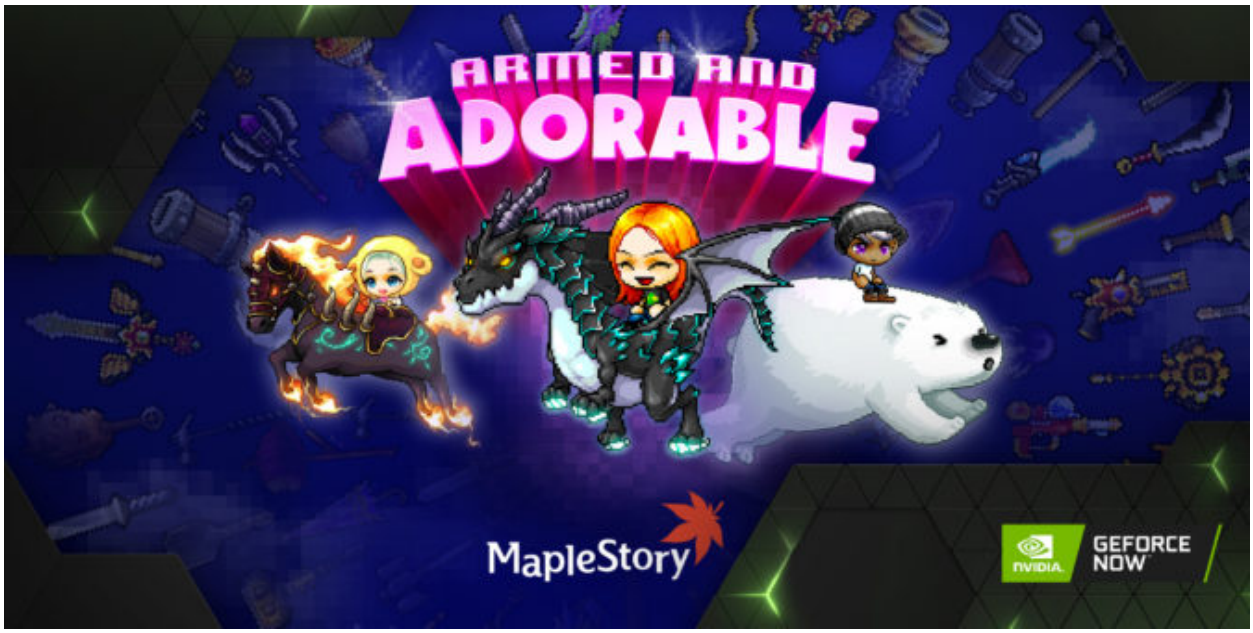
The Stratus+ wireless controller combines precision with comfort, delivering a full console experience on a mobile phone and giving a competitive edge to Android and Chromebook gamers. Gamers can simply connect to any Android mobile or Chromebook device with Bluetooth Low Energy and play with a rechargeable battery that lasts up to 90 hours. Or they can wire in to any Windows PC via USB connection.

The controller works great with GeForce NOW's RTX 3080 membership. Playing on select 120Hz Android phones, members can stream their favorite PC games at up to 120 frames per second.

SteelSeries' line of controllers is part of the full lineup of GeForce NOW Recommended products, including optimized routers that are perfect in-home networking upgrades.

Get Your Game On

This week brings the start of *Fortnite* Chapter 3 Season 2, "Resistance." Building has been wiped out. To help maintain cover, you now have an overshield and new tactics like sprinting, mantling and more. Even board an armored battle bus to be a powerful force or attach a cow catcher to your vehicle for extra ramming power. Join the Seven in the final battle against the IO to free the Zero Point. Don't forget to grab the Chapter 3 Season 2 Battle Pass to unlock characters like Tsuki 2.0, the familiar foe Gunnar and The Origin.



Adventure and rewards await on this exclusive GeForce NOW quest.

Nexon, maker of popular global MMORPG *MapleStory*, is launching a special in-game quest — exclusive to GeForce NOW members. Level 30+ Maplers who log in using GeForce NOW will receive a GeForce NOW quest that grants players a Lil Boo Pet, and a GeForce NOW Event Box that can be opened 24 hours after acquiring. But hurry – this quest is only available March 24-April 28.

And since GFN Thursday means more games every week. This week includes open-ended, zombie-infested sandbox *Project Zomboid*. Play alone or survive with friends thanks to multiplayer support across persistent servers.

Finally, a game that proves you can learn valuable skills by watching TV. Won't your mother be proud?

Feeling zombie shy? That's okay, there's always something new to play on GeForce NOW. Here's the complete list of six titles coming this week:

- *Highrise City* (New release on Steam)
- *Fury Unleashed* (Steam)
- *Power to the People* (Steam and Epic Games Store)
- *Project Zomboid* (Steam)
- *Rugby 22* (Steam)
- *STORY OF SEASONS: Pioneers of Olive Town* (Steam)

Finally, the release timing for *Lumote: The Mastermote Chronicles* has shifted and will join GeForce NOW at a later date.

With the cloud making new ways to play PC games across your devices possible, we've got a question that may get you a bit nostalgic this GFN Thursday. Let us know your answer on Twitter:

Post navigation

Orchestrated to Perfection: NVIDIA Data Center Grooves to Tune of Millionfold Speedups

March 24, 2022 by Michael Kagan

The hum of a bustling data center is music to an AI developer's ears — and NVIDIA data centers have found a rhythm of their own, grooving to the swing classic "Sing, Sing, Sing" in this week's GTC keynote address.

The lighthearted video, created with the NVIDIA Omniverse platform, features Louis Prima's iconic music track, re-recorded at the legendary Abbey Road Studios. Its drumming, dancing data center isn't just for kicks — it celebrates the ability of NVIDIA data center solutions to orchestrate unprecedented AI performance.

Watch Video At: https://youtu.be/DFKdU6AIseI

Cutting-edge AI is tackling the world's biggest challenges — but to do so, it needs the most advanced data centers, with thousands of hardware and software components working in perfect harmony.

At GTC, NVIDIA is showcasing the latest data center technologies to accelerate next-generation applications in business, research and art. To keep up with the growing demand for computing these applications, optimization is needed across the entire computing stack, as well as innovation at the level of distributed algorithms, software and systems.

Performance growth at the bottom of the computing stack, based on Moore's law, can't keep pace with the requirements of these applications. Moore's law, which predicted a 2x growth in computing performance every other year, has yielded to Huang's law — that GPUs will double AI performance every year.

Advancements across the entire computing stack, from silicon to application-level software, have contributed to an unprecedented million-x speedup in accelerated computing in the last 20 years. It's not just about faster GPUs, DPUs and CPUs. Computing based on neural network models, advanced network technologies and distributed software algorithms all contribute to the data center innovation needed to keep pace with the demands of ever-growing AI models.

Through these innovations, the data center has become the single unit of computing. Thousands of servers work seamlessly as one, with NVIDIA Magnum IO software and new breakthroughs like the NVIDIA NVLink Switch System unveiled at GTC combining to link advanced AI infrastructure.

Orchestrated to perfection, an NVIDIA-powered data center will support innovations that are yet to be even imagined.

Developing a Digital Twin of the Data Center

The GTC video performance showcases a digital twin NVIDIA is building of its own data centers — a virtual representation of the physical supercomputer that NVIDIA designers and engineers can use to test new configurations or software builds before releasing updates to the physical system.

In addition to enabling continuous integration and delivery, a digital twin of a data center can be used to optimize operational efficiency, including response time, resource utilization and energy consumption.

Digital twins can help teams predict equipment failures, proactively replace weak links and test improvement measures before applying them. They can even provide a testing ground to fine-tune data centers for specific enterprise users or applications.

Applicable across industries and applications, digital twin technology is already being used as a powerful tool for warehouse optimizations, climate simulations, smart factory development and renewable energy planning.

In NVIDIA's data center digital twin, viewers can spot flagship technologies including NVIDIA DGX SuperPOD and EGX-based NVIDIA-Certified systems with BlueField DPUs and InfiniBand switches. The performance also features a special appearance by Toy Jensen, an application built with Omniverse Avatar.

The visualization was developed in NVIDIA Omniverse, a platform for real-time world simulation and 3D design collaboration. Omniverse connects science and art by bringing together creators, developers, engineers and AIs across industries to work together in a shared virtual world.

Omniverse digital twins are true to reality, accurately simulating the physics and materials of their real counterparts. The realism allows Omniverse users to test out processes, interactions and new technologies in the digital space before moving to the physical world.

Every factory, neighborhood and city could one day be replicated as a digital twin. With connected sensors powered by edge computing, these sandbox environments can be continuously updated to reflect changes to the corresponding real-world assets or systems. They can help develop next-generation autonomous robots, smart cities and 5G networks.

A digital twin can learn the laws of physics, chemistry, biology and more, storing this information in its computing brain.

Just as kingdoms centuries ago sent explorers to travel the world and return with new knowledge, edge sensors and robots are today's explorers for digital twin environments. Each sensor brings new observations back to the digital twin's brain, which consolidates the data, learns from it and updates the autonomous systems within the virtual environment. This collective learning will tune digital twins to perfection.

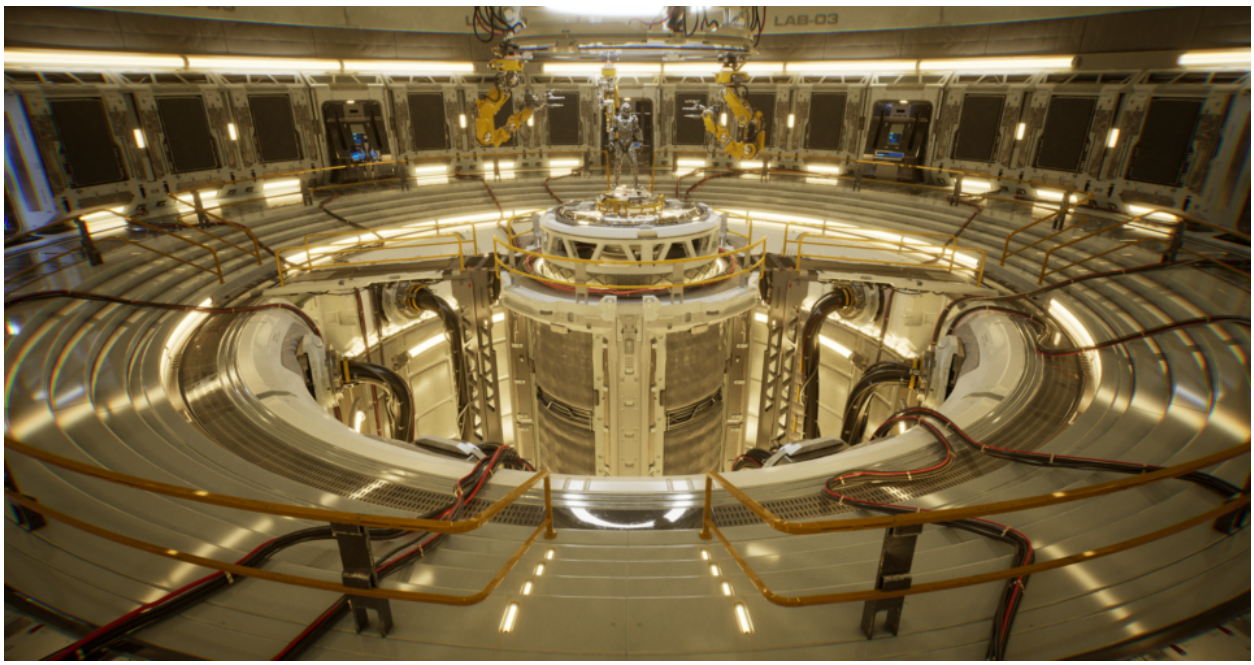
Hear about the latest innovations in AI, accelerated computing and virtual world simulation at GTC, streaming online through March 24. Register free and learn more about data center acceleration in the session replay, "How to Achieve Millionfold Speedups in Data Center Performance." Watch NVIDIA founder and CEO Jensen Huang's keynote address below:

Watch Video At: https://youtu.be/39ubNuxnrK8

Post navigation

What Is Path Tracing?

March 23, 2022 by Brian Caulfield



Turn on your TV. Fire up your favorite streaming service. Grab a Coke. A demo of the most important visual technology of our time is as close as your living room couch.

Propelled by an explosion in computing power over the past decade and a half, path tracing has swept through visual media.

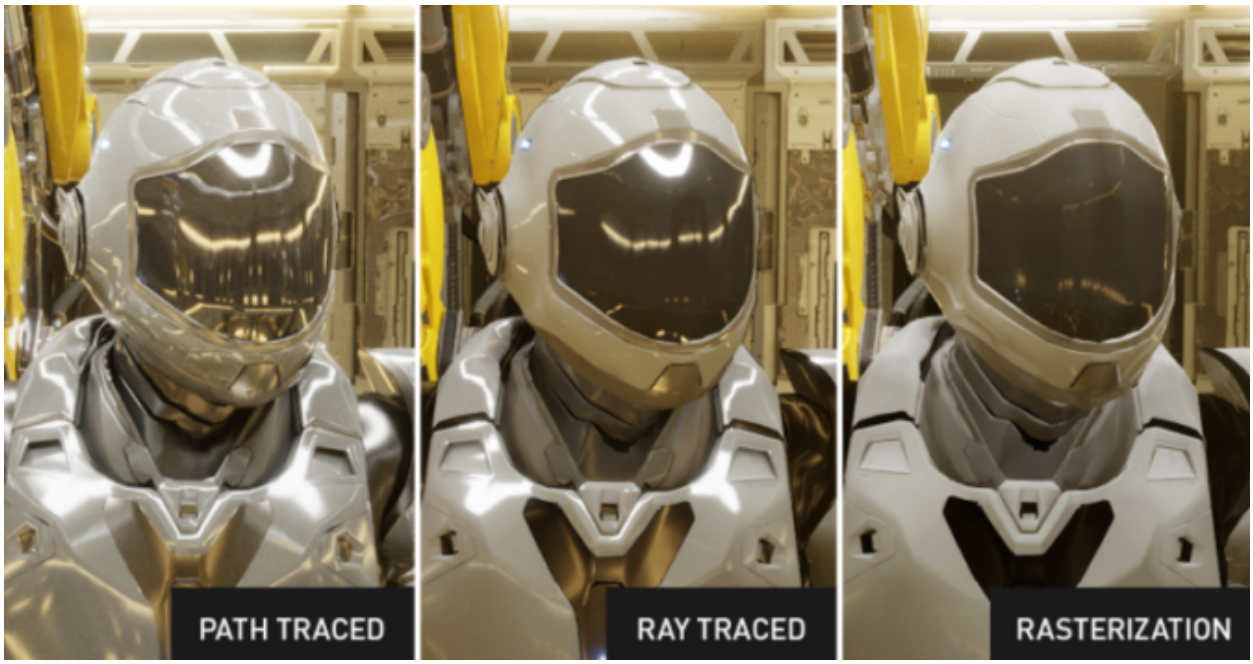
It brings big effects to the biggest blockbusters, casts subtle light and shadow on the most immersive melodramas and has propelled the art of animation to new levels.

More's coming.

Path tracing is going real time, unleashing interactive, photorealistic 3D environments filled with dynamic light and shadow, reflections and refractions.

So what is path tracing? The big idea behind it is seductively simple, connecting innovators in the arts and sciences over the span half a millennium.

What's the Difference Between Rasterization and Ray Tracing?



First, let's define some terms, and how they're used today to create interactive graphics — graphics that can react in real time to input from a user, such as in video games.

The first, rasterization, is a technique that produces an image as seen from a single viewpoint. It's been at the heart of GPUs from the start. Modern NVIDIA GPUs can generate over 100 billion rasterized pixels per second. That's made rasterization ideal for real-time graphics, like gaming.

Ray tracing is a more powerful technique than rasterization. Rather than being constrained to finding out what is visible from a single point, it can determine what is visible from many different points, in many different directions. Starting with the NVIDIA Turing architecture,

NVIDIA GPUs have provided specialized RTX hardware to accelerate this difficult computation. Today, a single GPU can trace billions of rays per second.

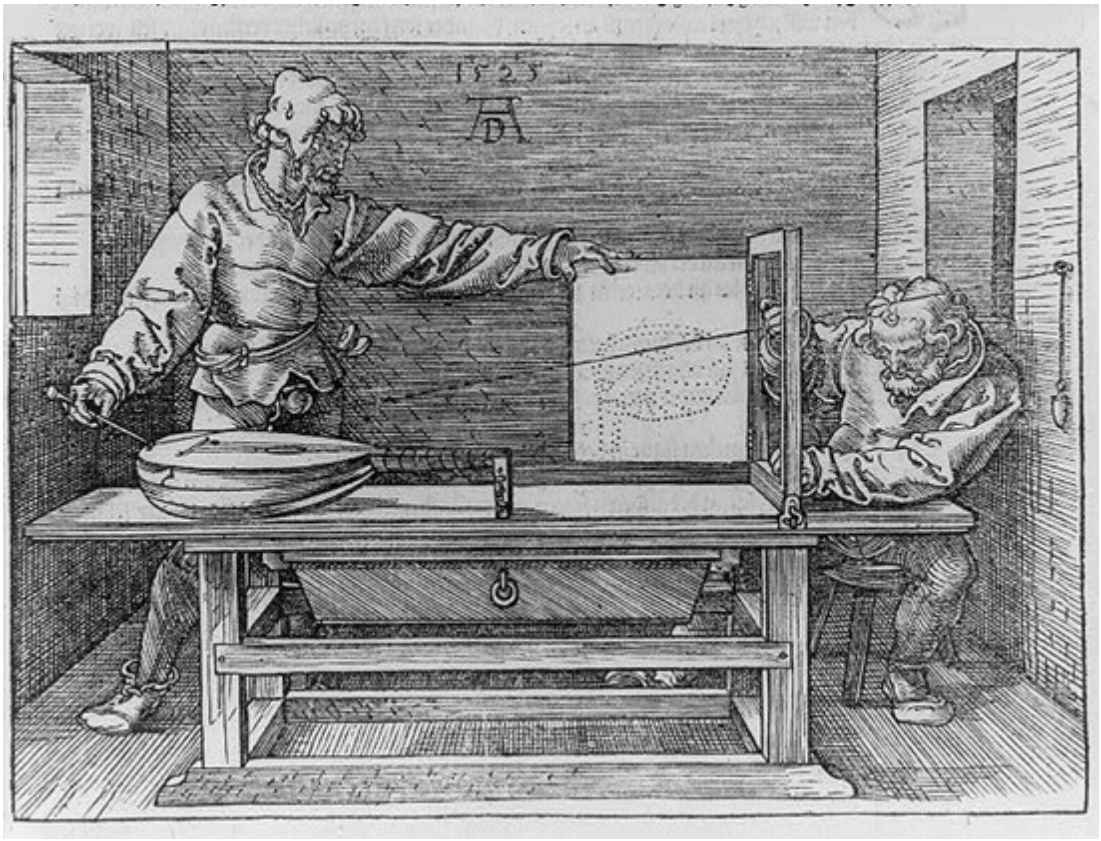
Being able to trace all of those rays makes it possible to simulate how light scatters in the real world much more accurately than is possible with rasterization. However, we still must answer the questions, how will we simulate light and how will we bring that simulation to the GPU?

What's Ray Tracing? Just Follow the String

To better answer that question, it helps to understand how we got here.

David Luebke, NVIDIA vice president of graphics research, likes to begin the story in the 16th century with Albrecht Dürer — one of the most important figures of the Northern European Renaissance — who used string and weights to replicate a 3D image on a 2D surface.

Dürer made it his life's work to bring classical and contemporary mathematics together with the arts, achieving breakthroughs in expressiveness and realism.



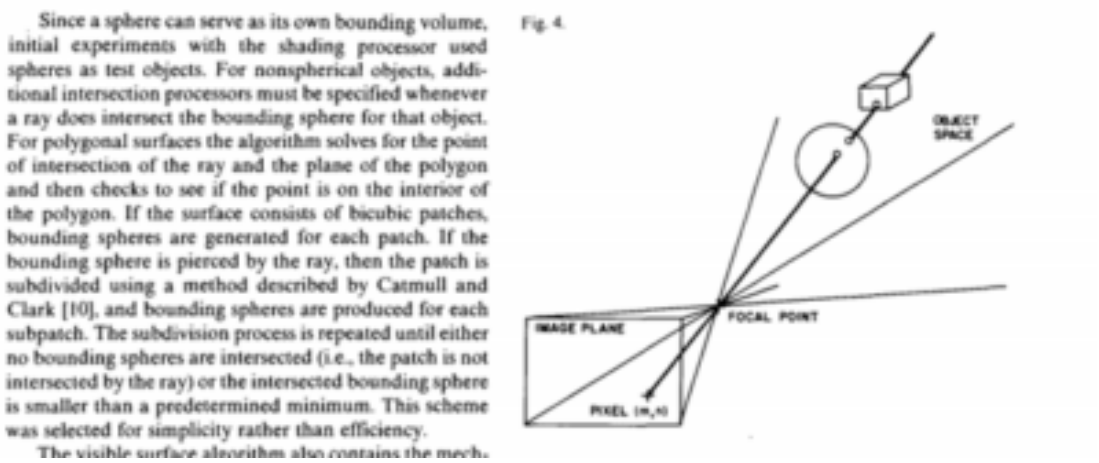
The string's the thing: Albrecht Dürer was the first to describe what's now known as "ray tracing," a technique for creating accurate representations of 3D objects on a 2D surfaces in Underweysung der Messung (Nuremberg, 1525),f15

In 1525 with *Treatise on Measurement,* Dürer was the first to describe the idea of ray tracing. Seeing how Dürer described the idea is the easiest way to get your head around the concept.

Just think about how light illuminates the world we see around us.

Now imagine tracing those rays of light backward from the eye with a piece of string like the one Dürer used, to the objects that light interacts with. That's ray tracing.

Ray Tracing for Computer Graphics



Since a sphere can serve as its own bounding volume, initial experiments with the shading processor used spheres as test objects. For nonspherical objects, additional intersection processors must be specified whenever a ray does intersect the bounding sphere for that object. For polygonal surfaces the algorithm solves for the point of intersection of the ray and the plane of the polygon and then checks to see if the point is on the interior of the polygon. If the surface consists of bicubic patches, bounding spheres are generated for each patch. If the bounding sphere is pierced by the ray, then the patch is subdivided using a method described by Catmull and Clark [10], and bounding spheres are produced for each subpatch. The subdivision process is repeated until either no bounding spheres are intersected (i.e., the patch is not intersected by the ray) or the intersected bounding sphere is smaller than a predetermined minimum. This scheme was selected for simplicity rather than efficiency.

The visible surface algorithm also contains the mech-

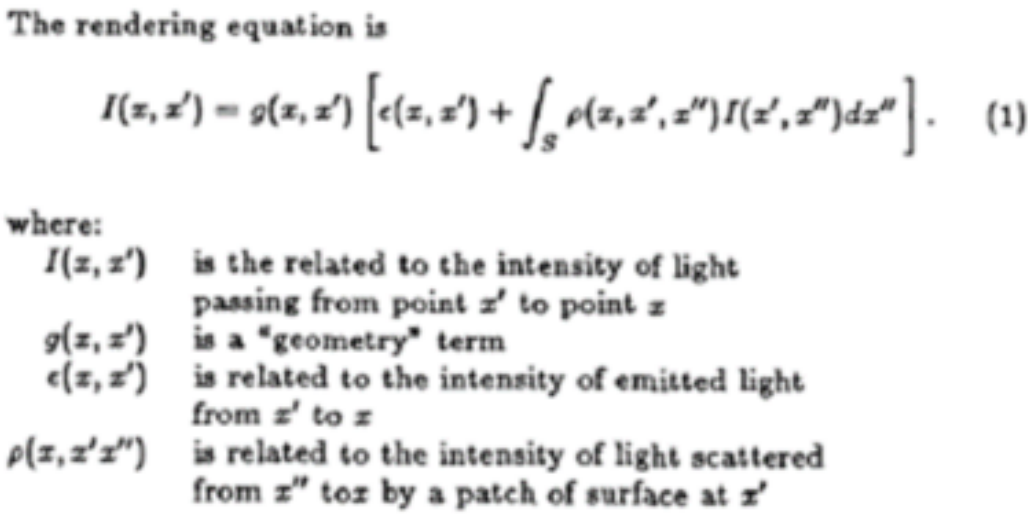
Fig. 4.

Turner Whitted's 1979 paper, "An improved illumination model for shaded display," jump-started a ray-tracing renaissance.

In 1969, more than 400 years after Dürer's death, IBM's Arthur Appel showed how the idea of ray tracing could be brought to computer graphics, applying it to computing visibility and shadows.

A decade later, Turner Whitted was the first to show how this idea could capture reflection, shadows and refraction, explaining how the seemingly simple concept could make much more sophisticated computer graphics possible. Progress was rapid in the following few years.

In 1984, Lucasfilm's Robert Cook, Thomas Porter and Loren Carpenter detailed how ray tracing could incorporate many common filmmaking techniques — including motion blur, depth of field, penumbras, translucency and fuzzy reflections — that were, until then, unattainable in computer graphics.



The rendering equation is

$$I(x,x') = g(x,x')\left[\epsilon(x,x') + \int_S \rho(x,x',x'')I(x',x'')dx''\right]. \quad (1)$$

where:

$I(x,x')$	is the related to the intensity of light passing from point x' to point x
$g(x,x')$	is a "geometry" term
$\epsilon(x,x')$	is related to the intensity of emitted light from x' to x
$\rho(x,x'x'')$	is related to the intensity of light scattered from x'' toz by a patch of surface at x'

Jim Kajiya's 1986 paper, "The Rendering Equation," not only outlined an elegant, physics-based equation for describing how light moves around in a scene, it outlined an efficient way to put it to work.

Two years later, CalTech professor Jim Kajiya's crisp, seven-page paper, "The Rendering Equation," connected computer graphics with physics by way of ray tracing and introduced the path-tracing algorithm, which makes it possible to accurately represent the way light scatters throughout a scene.

What's Path Tracing?

In developing path tracing, Kajiya turned to an unlikely inspiration: the study of radiative heat transfer, or how heat spreads throughout an environment. Ideas from that field led him to introduce *the rendering equation*, which describes how light passes through the air and scatters from surfaces.

The rendering equation is concise, but not easy to solve. Computer graphics scenes are complex, with billions of triangles not being unusual today. There's no way to solve the rendering equation directly, which led to Kajiya's second crucial innovation.

Kajiya showed that statistical techniques could be used to solve the rendering equation: even if it isn't solved directly, it's possible to solve it along the paths of individual rays. If it is solved along the path of enough rays to approximate the lighting in the scene accurately, photorealistic images are possible.

And how is the rendering equation solved along the path of a ray? Ray tracing.

The statistical techniques Kajiya applied are known as Monte Carlo integration and date to the earliest days of computers in the 1940s. Developing improved Monte Carlo algorithms for path tracing remains an open research problem to this day; NVIDIA researchers are at the

forefront of this area, regularly publishing new techniques that improve the efficiency of path tracing.

By putting these two ideas together — a physics-based equation for describing the way light moves around a scene — and the use of Monte Carlo simulation to help choose a manageable number of paths back to a light source, Kajiya outlined the fundamental techniques that would become the standard for generating photorealistic computer-generated images.

His approach transformed a field dominated by a variety of disparate rendering techniques into one that — because it mirrored the physics of the way light moved through the real world — could put simple, powerful algorithms to work that could be applied to reproduce a large number of visual effects with stunning levels of realism.

Path Tracing Comes to the Movies

In the years after its introduction in 1987, path tracing was seen as an elegant technique — the most accurate approach known — but it was completely impractical. The images in Kajiya's original paper were just 256 by 256 pixels, yet they took over 7 hours to render on an expensive mini-computer that was far more powerful than the computers available to most other people.

But with the increase in computing power driven by Moore's law — which described the exponential increase in computing power driven by advances that allowed chipmakers to double the number of transistors on microprocessors every 18 months — the technique became more and more practical.

Beginning with movies such as 1998's *A Bug's Life*, ray tracing was used to enhance the computer-generated imagery in more and more motion pictures. And in 2006, the first entirely path-traced movie, *Monster House*, stunned audiences. It was rendered using the Arnold software that was co-developed at Solid Angle SL (since acquired by Autodesk) and Sony Pictures Imageworks.

The film was a hit — grossing more than $140 million worldwide. And it opened eyes about what a new generation of computer animation could do. As more computing power became available, more movies came to rely on the technique, producing images that are often indistinguishable from those captured by a camera.

The problem: it still takes hours to render a single image and sprawling collections of servers — known as "render farms" — are running continuously to render images for months in order to make a complete movie. Bringing that to real-time graphics would take an extraordinary leap.

What Does This Look Like in Gaming?

For many years, the idea of path tracing in games was impossible to imagine. While many game developers would have agreed that they would want to use path tracing if it had the performance necessary for real-time graphics, the performance was so far off of real time that path tracing seemed unattainable.

Yet as GPUs have continued to become faster and faster, and now with the widespread availability of RTX hardware, real-time path tracing is in sight. Just as movies began incorporating some ray-tracing techniques before shifting to path tracing — games have started by putting ray tracing to work in a limited way.

Right now a growing number of games are partially ray traced. They combine traditional rasterization-based rendering techniques with some ray-tracing effects.

So what does path traced mean in this context? It could mean a mix of techniques. Game developers could rasterize the primary ray, and then path trace the lighting for the scene.

Rasterization is equivalent to casting one set of rays from a single point that stops at the first thing they hit. Ray tracing takes this further, casting rays from many points in any direction. Path tracing simulates the true physics of light, which uses ray tracing as one component of a larger light simulation system.

This would mean all lights in a scene are sampled stochastically — using Monte Carlo or other techniques — both for direct illumination, to light objects or characters, and for global illumination, to light rooms or environments with indirect lighting.

To do that, rather than tracing a ray back through one bounce, rays would be traced over multiple bounces, presumably back to their light source, just as Kajiya outlined.

A few games are doing this already, and the results are stunning.

Microsoft has released a plugin that puts path tracing to work in *Minecraft*.

Quake II, the classic shooter — often a sandbox for advanced graphics techniques — can also be fully path traced, thanks to a new plugin.

Watch Video At: https://youtu.be/p7RniXWvYhY

There's clearly more to be done. And game developers will need to know customers have the computing power they need to experience path-traced gaming.

Gaming is the most challenging visual computing project of all: requiring high visual quality and the speed to interact with fast-twitch gamers.

Expect techniques pioneered here to spill out to every aspect of our digital lives.

What's Next?

As GPUs continue to grow more powerful, putting path tracing to work is the next logical step.

For example, armed with tools such as Arnold from Autodesk, V-Ray from Chaos Group or Pixar's Renderman — and powerful GPUs — product designers and architects use ray tracing to generate photorealistic mockups of their products in seconds, letting them collaborate better and skip expensive prototyping.

As GPUs offer ever more computing power, video games are the next frontier for ray tracing and path tracing.

In 2018, NVIDIA announced NVIDIA RTX, a ray-tracing technology that brings real-time, movie-quality rendering to game developers.

NVIDIA RTX, which includes a ray-tracing engine running on NVIDIA Ampere and Turing architecture GPUs, supports ray-tracing through a variety of interfaces.

And NVIDIA has partnered with Microsoft to enable full RTX support via Microsoft's new DirectX Raytracing (DXR) API.

Since then, NVIDIA has continued to develop NVIDIA RTX technology, as more and more developers create games that support real-time ray tracing.

Minecraft even includes support for real-time path tracing, turning the blocky, immersive world into immersive landscapes swathed with light and shadow.

Thanks to increasingly powerful hardware, and a proliferation of software tools and related technologies, more is coming.

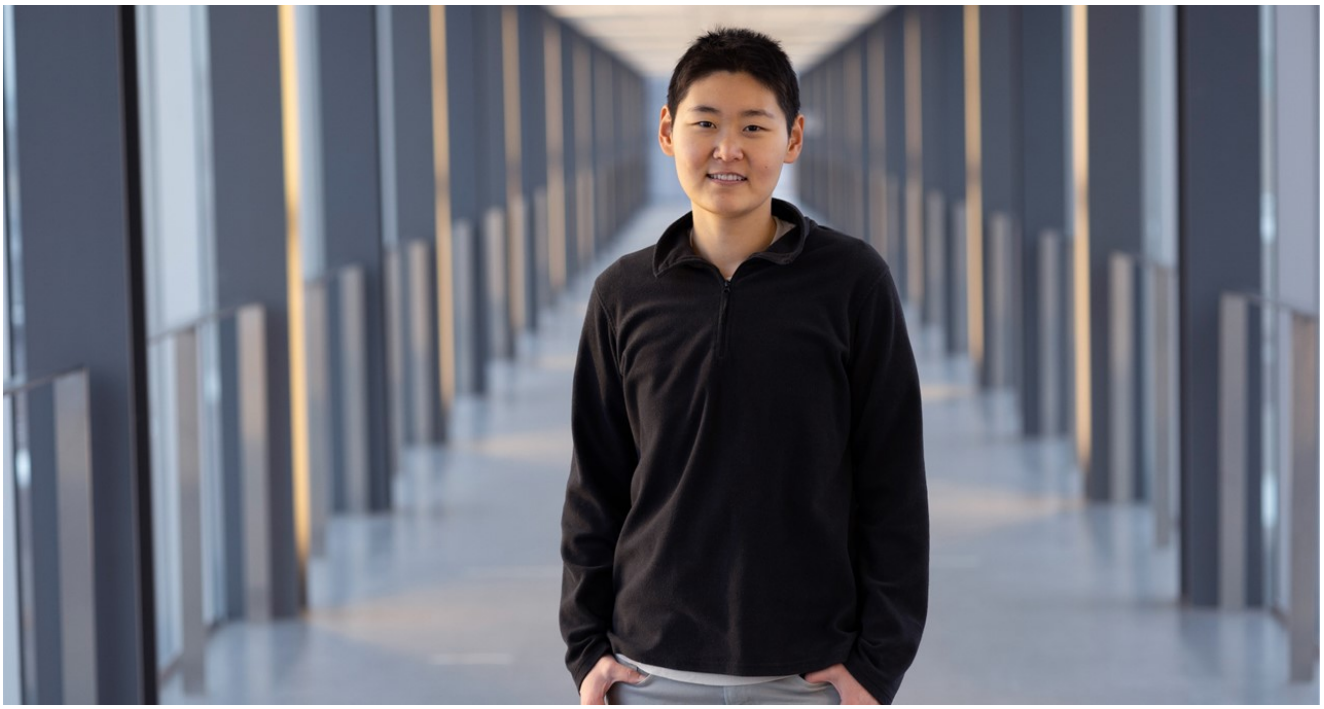
As a result, digital experiences — games, virtual worlds and even online collaboration tools — will take on the cinematic qualities of a Hollywood blockbuster.

So don't get too comfy. What you're seeing from your living room couch is just a demo of what's to come in the world all around us.

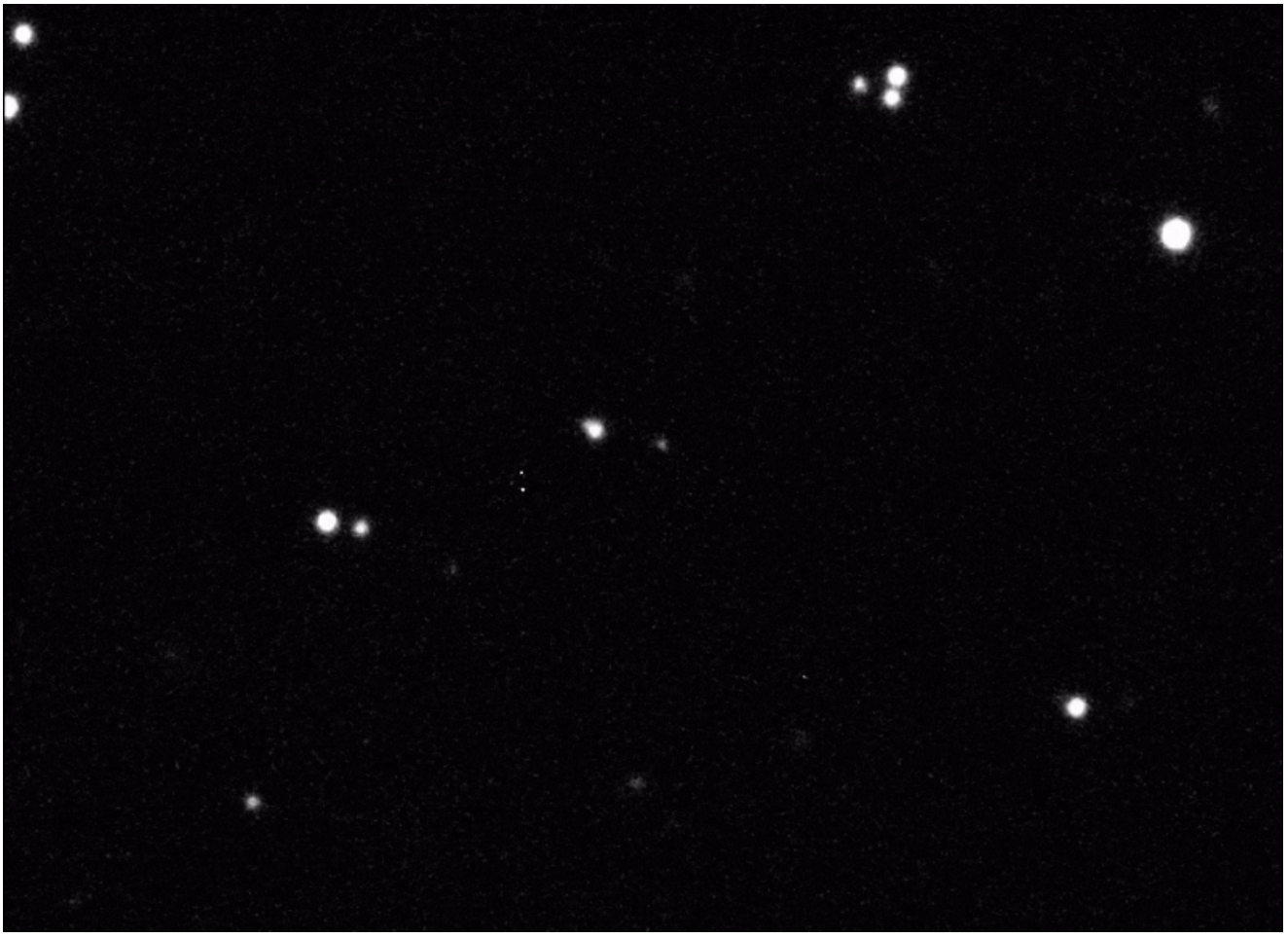
- *Learn more about the NVIDIA RTX Path Tracing SDK.*
- *Watch our "Research Advances Toward Real-time Path Tracing" session on demand.*
- *For a deep dive into ray tracing, dig into Ray Tracing Gems and Ray Tracing Gems II.*
- *Check out "Physically Based Rendering: From Theory to Implementation," by Matt Pharr, Wenzel Jakob and Greg Humphreys. It offers both mathematical theories and practical techniques for putting modern photorealistic rendering to work.*
- *Access NVIDIA resources for Game Development.*
- *Learn more about path tracing on the NVIDIA Technical Blog.*

This article has been updated to reflect the correct date for the publication of Albrecht Dürer's Treatise on Measurement.

All NVIDIA News

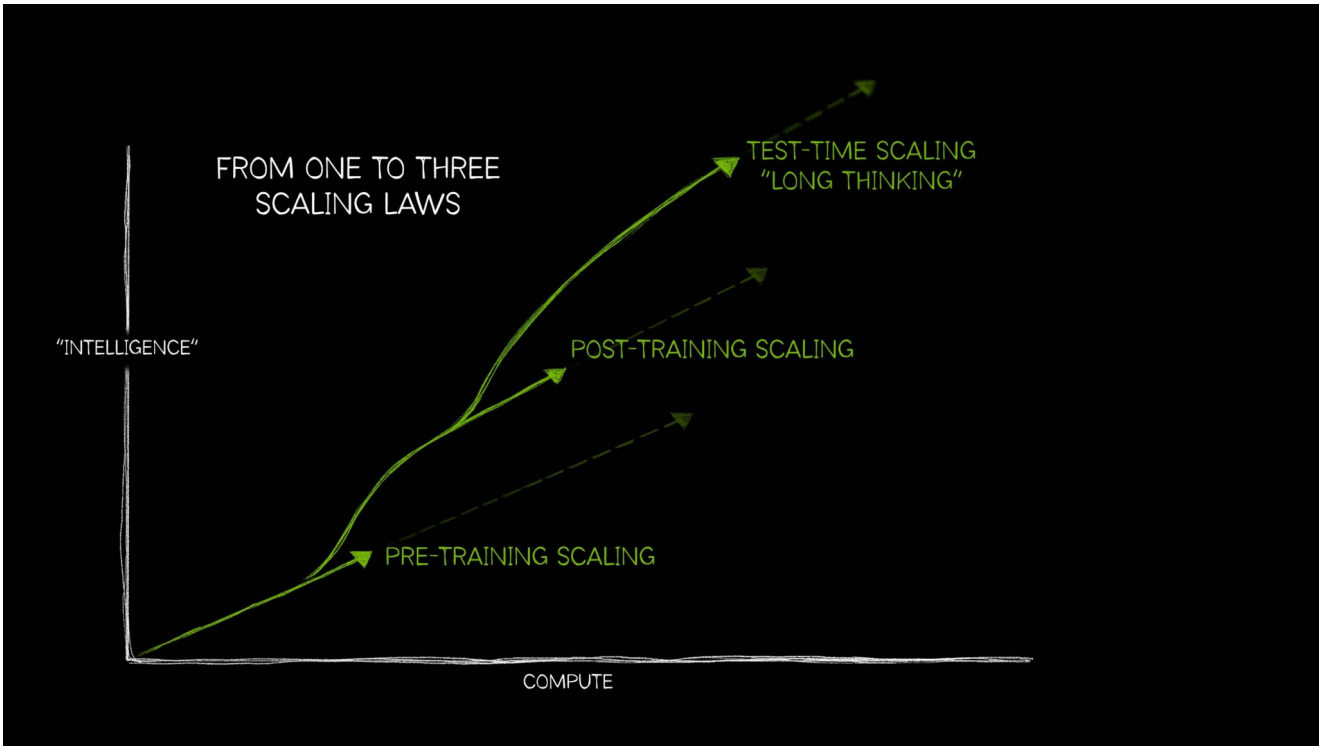


All Systems Go: NVIDIA Engineer Takes NIMble Approach to Innovation

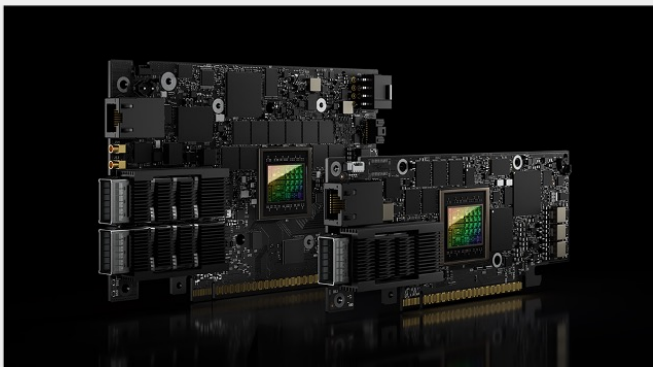


Physicists Tap James Web Space Telescope to Track New Asteroids and City-Killer Rock

GeForce NOW Welcomes Warner Bros. Games to the Cloud With ‘Batman: Arkham’ Series



How Scaling Laws Drive Smarter, More Powerful AI



Safety First: Leading Partners Adopt NVIDIA Cybersecurity AI to Safeguard Critical Infrastructure

Post navigation

NVIDIA Showcases Novel AI Tools in DRIVE Sim to Advance Autonomous Vehicle Development

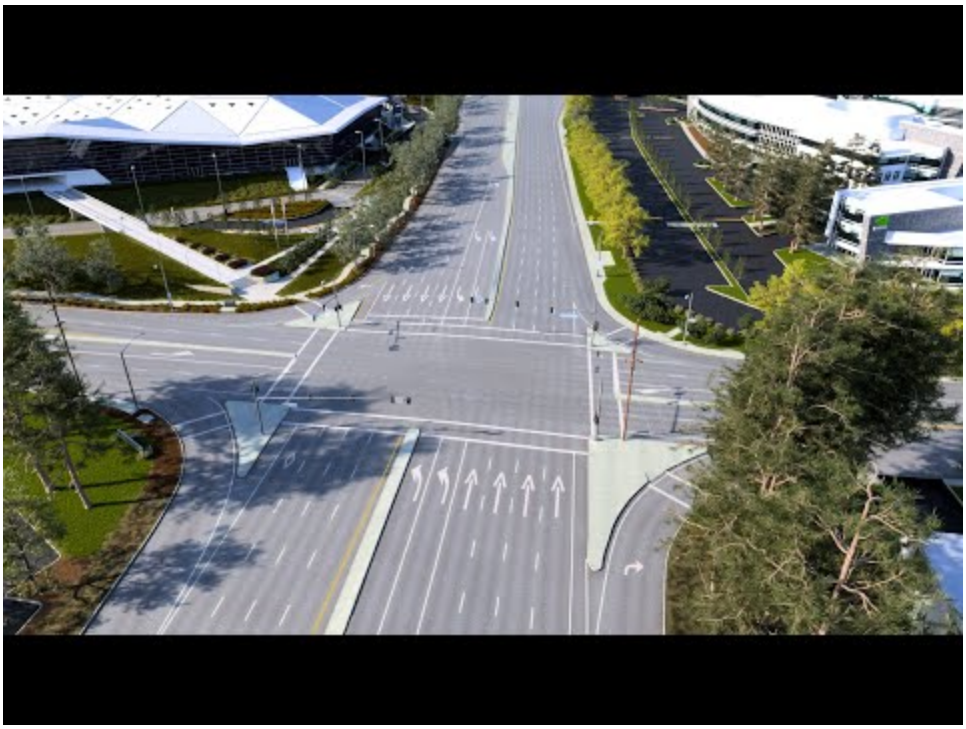
Breakthroughs by NVIDIA Research demonstrate the power of Omniverse digital twin to reconstruct real world scenarios in simulation.

March 23, 2022 by Matt Cragun

Autonomous vehicle development and validation require the ability to replicate real-world scenarios in simulation.

At GTC, NVIDIA founder and CEO Jensen Huang showcased new AI-based tools for NVIDIA DRIVE Sim that accurately reconstruct and modify actual driving scenarios. These tools are enabled by breakthroughs from NVIDIA Research that leverage technologies such as NVIDIA Omniverse platform and NVIDIA DRIVE Map.

Huang demonstrated the methods side-by-side, showing how developers can easily test multiple scenarios in rapid iterations:



Watch Video At: https://youtu.be/RVFIDEuNtt0

Once any scenario is reconstructed in simulation, it can act as the foundation for many different variations — from changing the trajectory of an oncoming vehicle, or adding an obstacle to the driving path — giving developers the ability to improve the AI driver.

However, reconstructing real-world driving scenarios and generating realistic data from it in simulation is a time- and labor-intensive process. It requires skilled engineers and artists, and even then, can be difficult to do.

NVIDIA has implemented two AI-based methods to seamlessly perform this process: virtual reconstruction and neural reconstruction. The first replicates the real-world scenario as a fully synthetic 3D scene, while the second uses neural simulation to augment real-world sensor data.

Both methods are able to expand well beyond recreating a single scenario to generating many new and challenging scenarios. This capability accelerates the continuous AV training, testing and validation pipeline.

Virtual Reconstruction

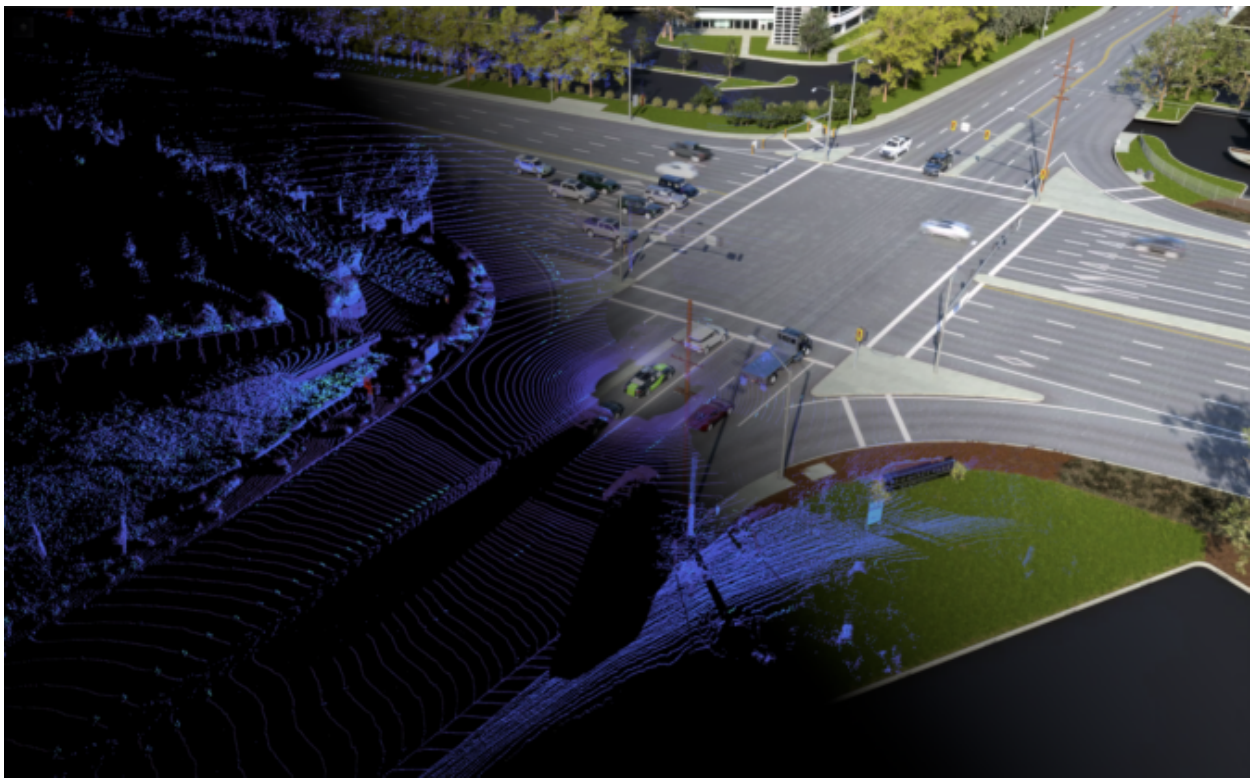
In the keynote video above, an entire driving environment and set of scenarios around NVIDIA's headquarters are reconstructed in 3D using NVIDIA DRIVE Map, Omniverse and DRIVE Sim.

With DRIVE Map, developers have access to a digital twin of a road network in Omniverse. Using tools built on Omniverse, the detailed map is converted into a drivable simulation environment that can be used with NVIDIA DRIVE Sim.

With the reconstructed simulation environment, developers can recreate events, like a close call at an intersection or navigating a construction zone, using camera, lidar and vehicle data from real-world drives.

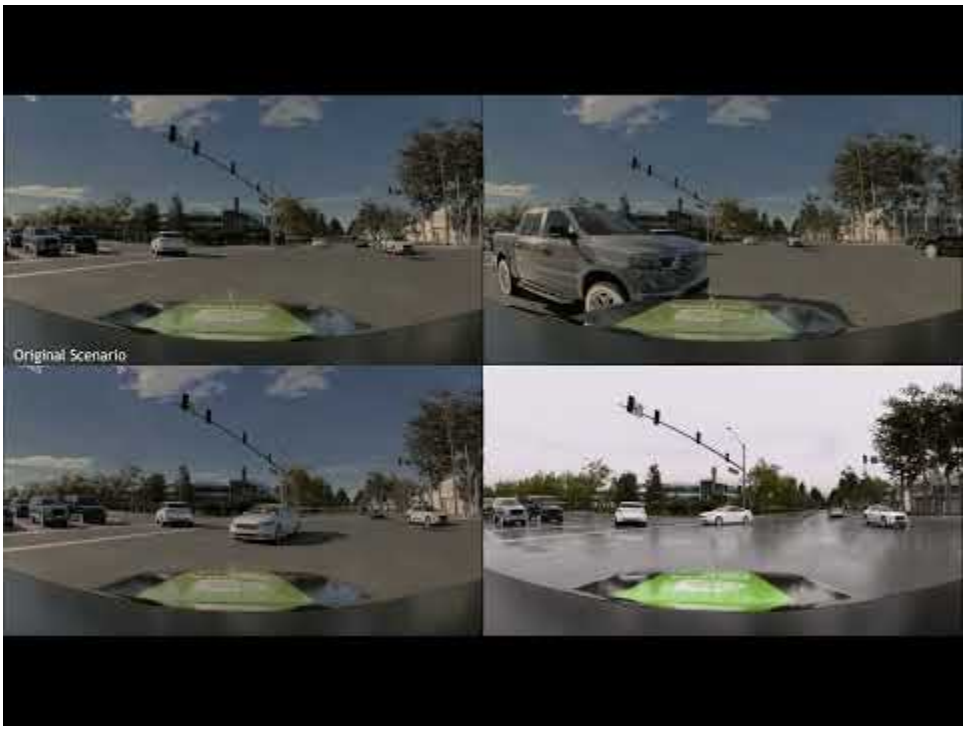
The platform's AI helps reconstruct the scenario. First, for each tracked object, an AI looks at camera images and finds the most similar 3D asset available from the DRIVE Sim catalog and color that most closely matches the color of the object from the video.

Finally, the actual path of the tracked object is recreated; however, there are often gaps because of occlusions. In such cases, an AI-based traffic model is applied to the tracked object to predict what it would have done and fill in the gaps in its trajectory.



Camera and lidar data from real drives are used with AI to reconstruct scenarios.

Virtual reconstruction enables developers to find potentially challenging situations to train and validate the AV system with high-fidelity data generated by physically based sensors and AI behavior models that can create many new scenarios. Data from the scenario can also train the behavior model.



Watch Video At: https://youtu.be/47YFEDoDMNg

Neural Reconstruction

The other approach relies on neural simulation rather than synthetically generating the scene, starting with real sensor data then modifying it.

Sensor replay — the process of playing back recorded sensor data to test the AV system's performance — is a staple of AV development. This process is open loop, meaning the AV stack's decisions don't affect the world since all of the data is prerecorded.

A preview of neural reconstruction methods by NVIDIA Research turn this recorded data into a fully reactive and modifiable world — as in the demo, when the originally recorded van driving past the car could be reenacted to swerve right instead. This revolutionary approach allows closed-loop testing and full interaction between the AV stack and the world it's driving in.

The process starts with recorded driving data. AI identifies the dynamic objects in the scene and removes them to create an exact replica of the 3D environment that can be rendered from new views. Dynamic objects are then reinserted into the 3D scene with realistic AI-based behaviors and physical appearance, accounting for illumination and shadows.

Watch Video At: https://youtu.be/o34qZaFVERU

The AV system then drives in this virtual world and the scene reacts accordingly. The scene can be made more complex through augmented reality by inserting other virtual objects, vehicles and pedestrians which are rendered as if they were part of the real scene and can physically interact with the environment.

Every sensor on the vehicle, including camera and lidar, can be simulated in the scene using AI.

Watch Video At: https://youtu.be/rba73bUgAFQ

A Virtual World of Possibilities

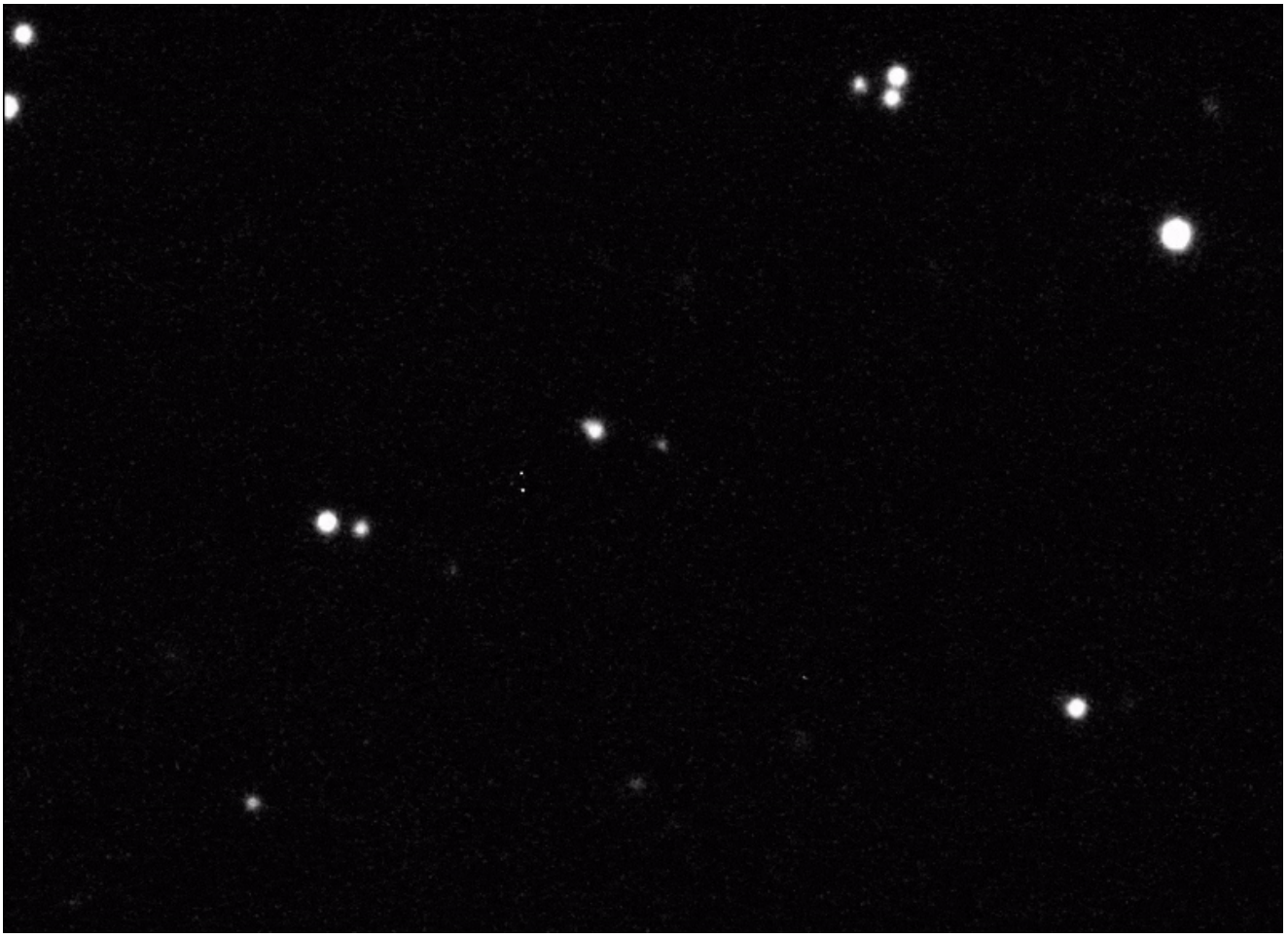
These new approaches are driven by NVIDIA's expertise in rendering, graphics and AI.

As a modular platform, DRIVE Sim supports these capabilities with a foundation of deterministic simulation. It provides the vehicle dynamics, AI-based traffic models, scenario tools and a comprehensive SDK to build any tool needed.

With these two powerful new AI methods, developers can easily move from the real world to the virtual one for faster AV development and deployment.

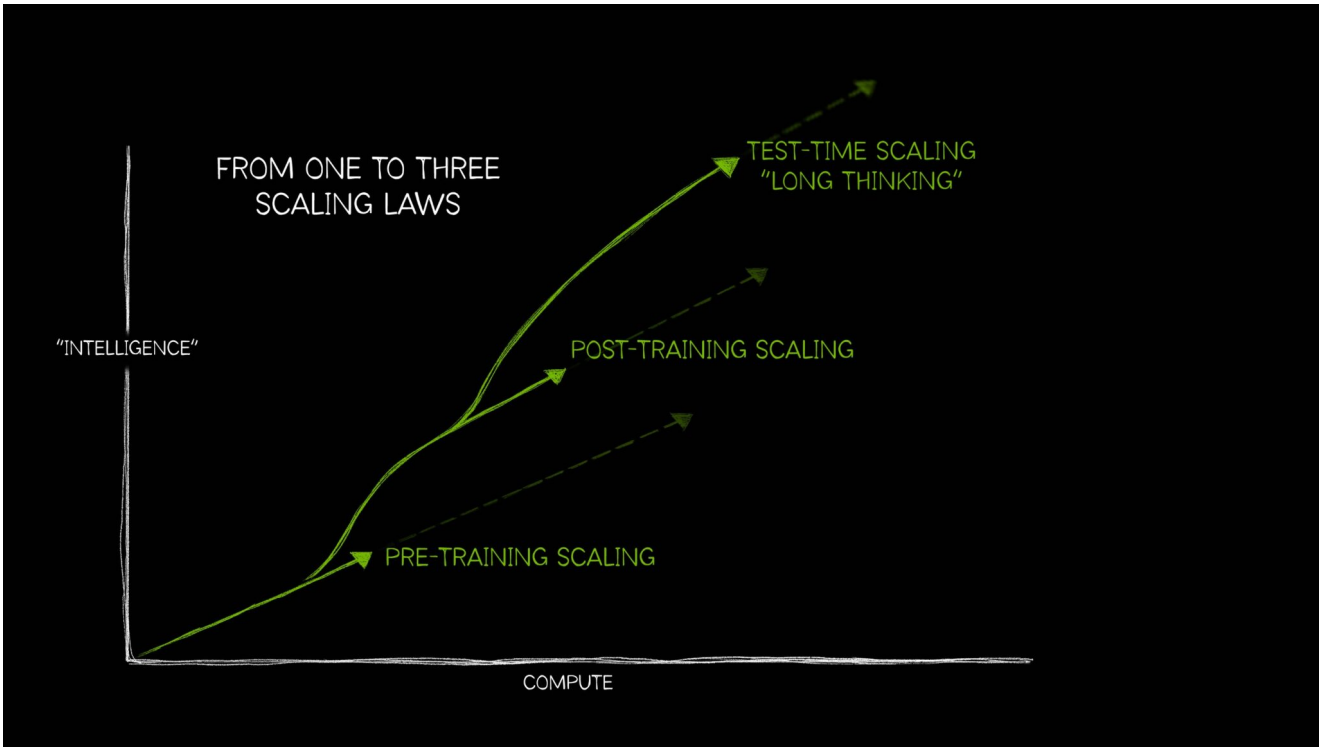
All NVIDIA News

All Systems Go: NVIDIA Engineer Takes NIMble Approach to Innovation

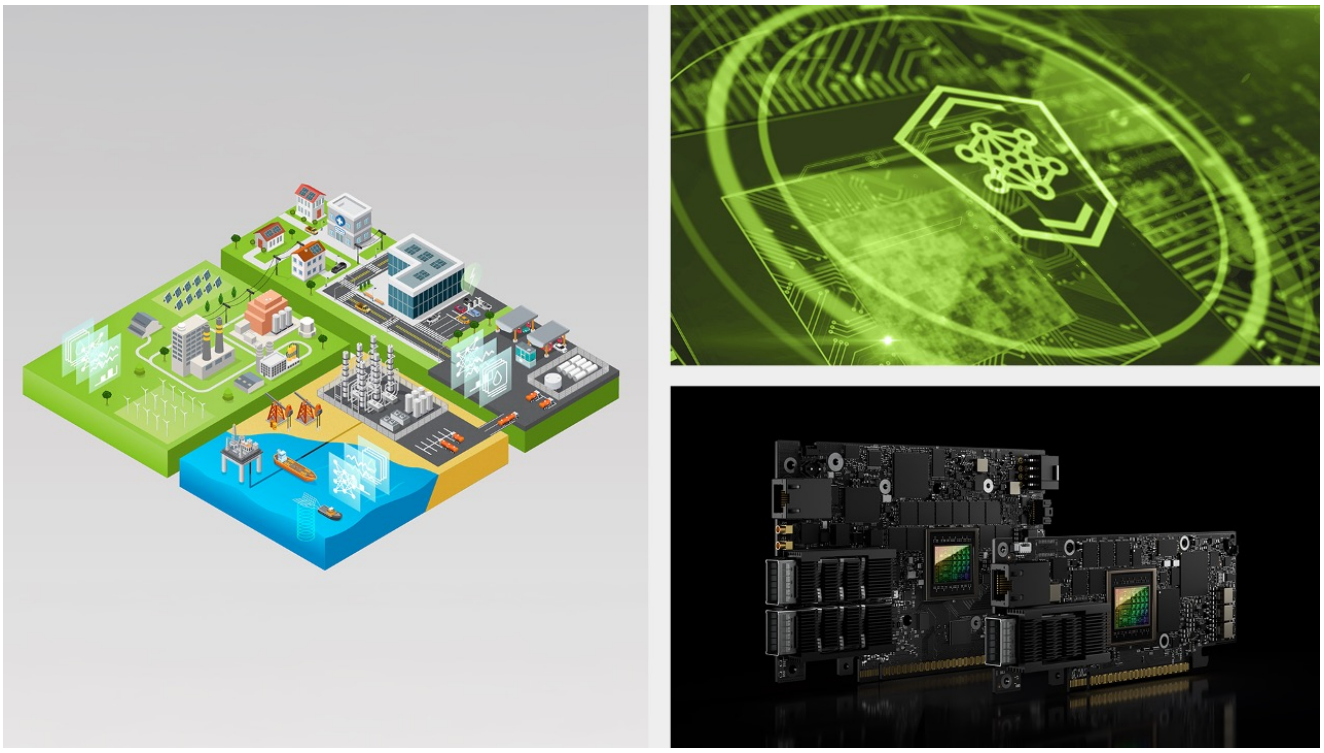


Physicists Tap James Web Space Telescope to Track New Asteroids and City-Killer Rock

GeForce NOW Welcomes Warner Bros. Games to the Cloud With ‘Batman: Arkham’ Series



How Scaling Laws Drive Smarter, More Powerful AI



Safety First: Leading Partners Adopt NVIDIA Cybersecurity AI to Safeguard Critical Infrastructure

Post navigation