
NVIDIA
AI DAYS 2021 / PRAGUE
AI-WORKFLOW ECOSYSTEM
RALPH HINSCHE / DECEMBER 2021

AGENDA

Introduction

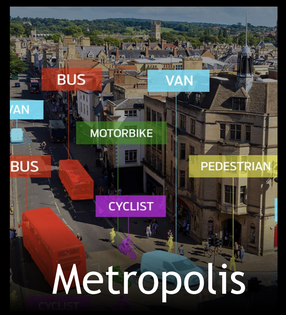
„OMNIVERSE“

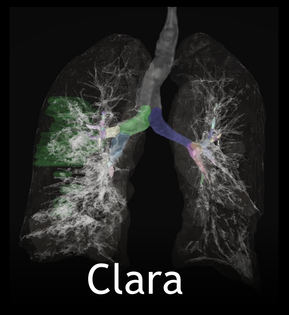
„TRITON“

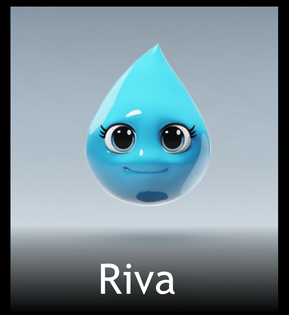
“NVidia AI Enterprise”

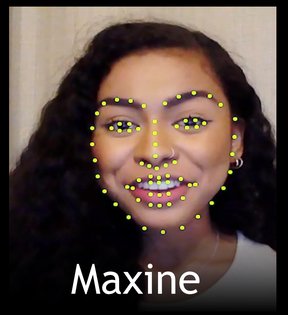
„Avatar / Replicator“

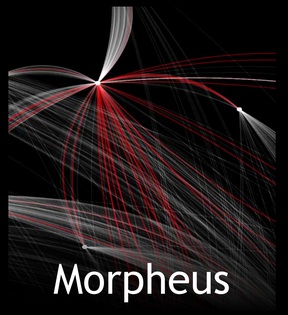
NVIDIA HAS OPTIMIZED AND CREATED THE END-TO-END WORKFLOW OF AI

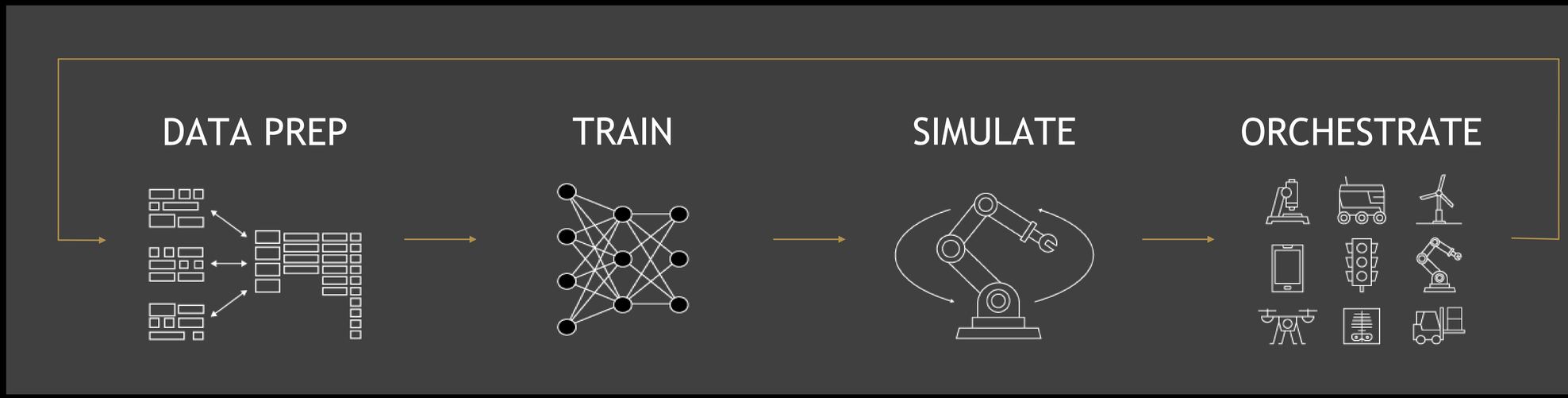












“THE MOST IMPORTANT THING TO UNDERSTAND ABOUT NVIDIA IS THAT IT IS NOT A HARDWARE COMPANY, AND NOT A SOFTWARE COMPANY. IT IS A COMPANY THAT INTEGRATES BOTH.”

STRATECHERY

“The magic of accelerated computing comes from the combination of CUDA, the acceleration libraries of algorithms that speed up applications, and the distributed computing systems and software that scale processing across entire data centers.

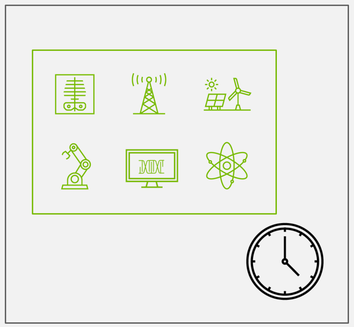
“We have been advancing CUDA and the ecosystem for 15 years and counting. We optimize across the ‘full stack,’ iterating between GPU, acceleration libraries, systems, and applications, continuously, all the while expanding the reach of our platform by adding new application domains that we accelerate.”

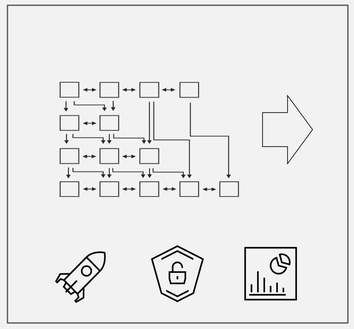
A NEW APPROACH TO ENTERPRISE AI

Almost half of AI projects never make it to production[1]

Increasing Complexity

32% of IT professionals rate increasing data complexity & data silos as a top barrier to AI adoption[2]





Lack of tools & platforms for AI

28% of IT professionals rate lack of infrastructure a top barrier to AI adoption[2]

An integrated solution for streamlined development and deployment

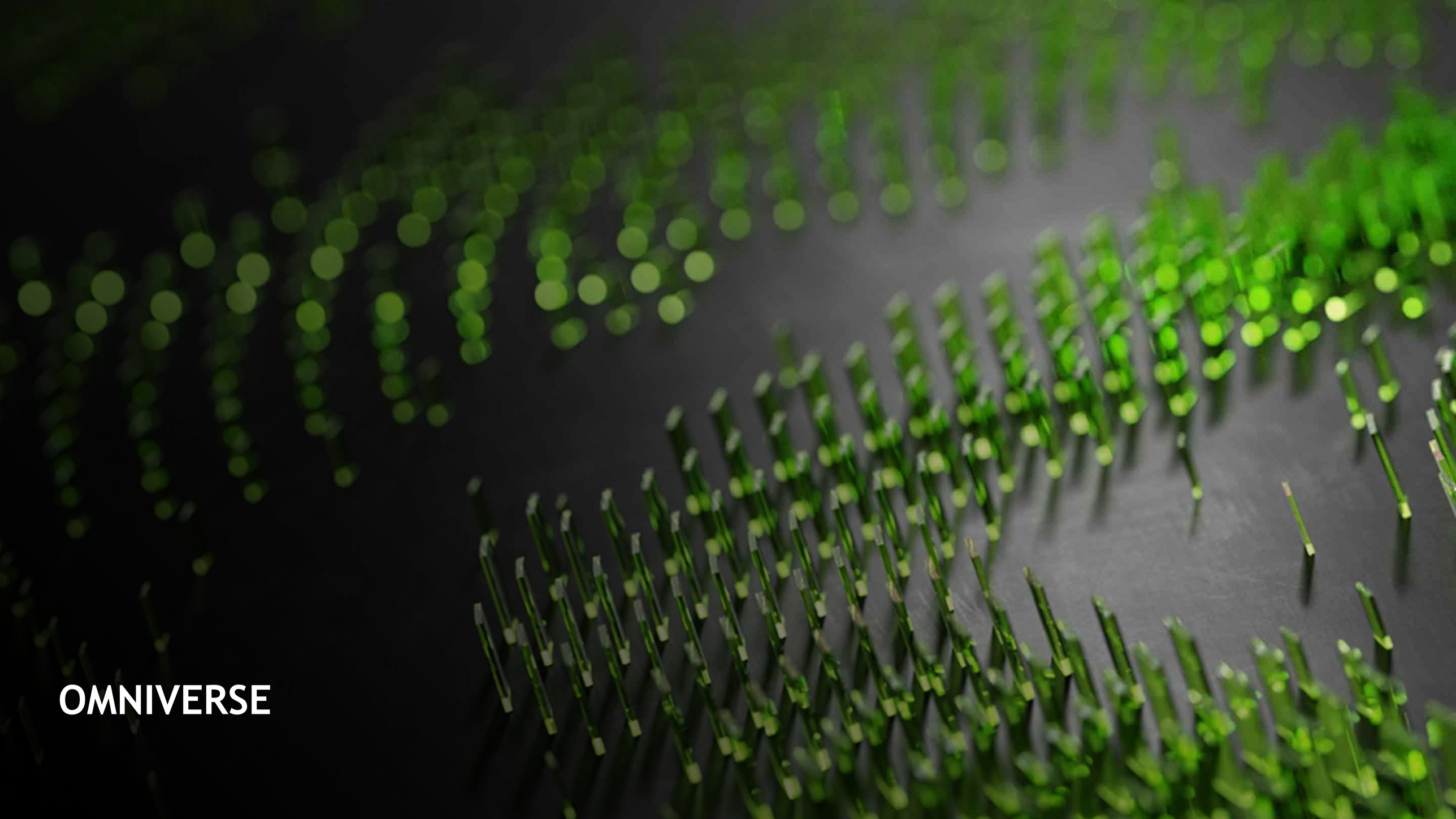
Process Automation | Conversational AI | Image Analysis | Existing Applications

AI Tools and Frameworks

Management & Orchestration

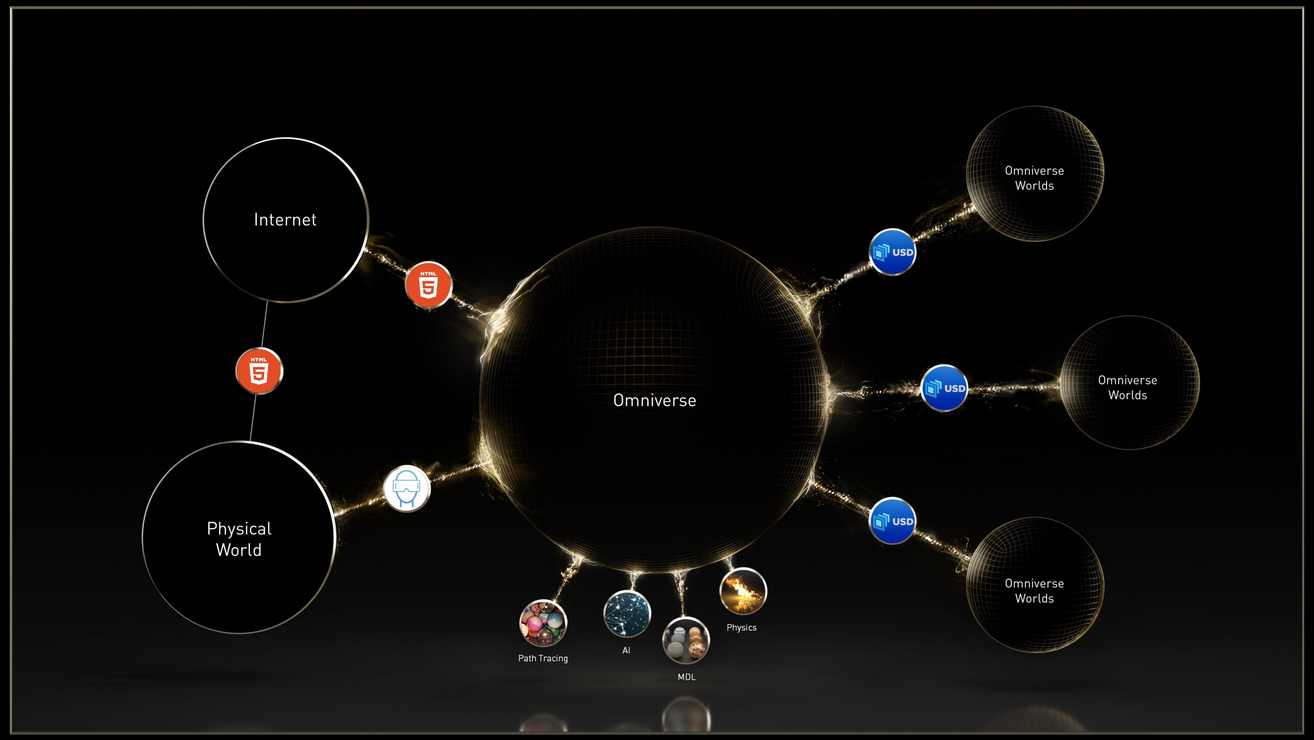
Accelerated Mainstream Systems

Source:1. Gartner “P-19019 AI in Organizations”, Claudia Ramos, Erick Brethenoux, 2020;
2. IBM, "Global AI Adoption 2021


OMNIVERSE

“NVIDIA IS LAUNCHING PLATFORMS, SOFTWARE AND COMPUTING TOOLS TO HELP ENABLE THE FUTURE OF VIRTUAL WORLDS”

BARRON’S



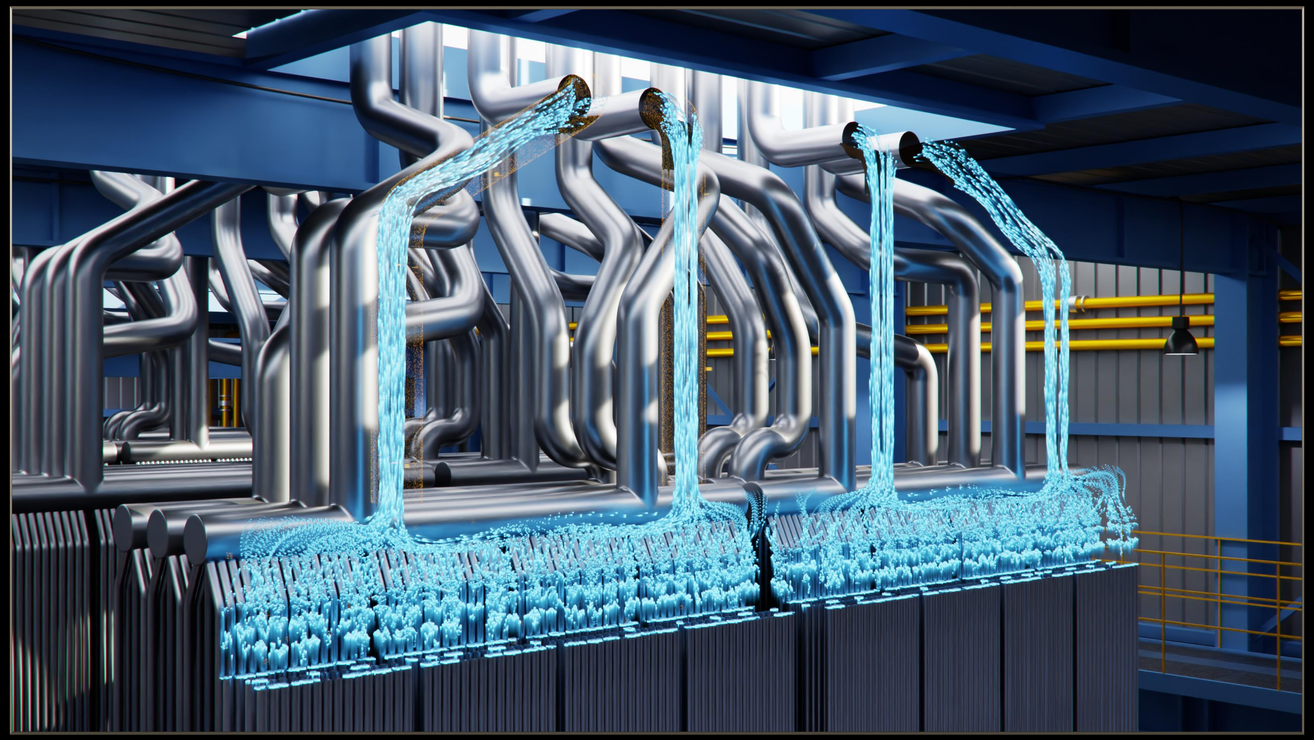
“The internet is about to leap into 3D. The resulting virtual worlds can obey the laws of physics, or not. There can be AI robots and the avatars of friends. We will jump from one world to another like we do on the web with hypertext.

“We built Omniverse for the builders of these worlds. Some worlds will be built for gathering and games. But a great many will be built by scientists, creators, and companies. Virtual worlds will crop up like websites today.

“Omniverse has been downloaded 70,000 times by designers in 500 companies. And TIME magazine recently highlighted Omniverse as one of the best inventions of 2021.”

“NVIDIA’S OMNIVERSE IS MAKING IT EASIER TO CREATE ULTRA-REALISTIC VIRTUAL SPACES FOR REAL-WORLD PURPOSES”

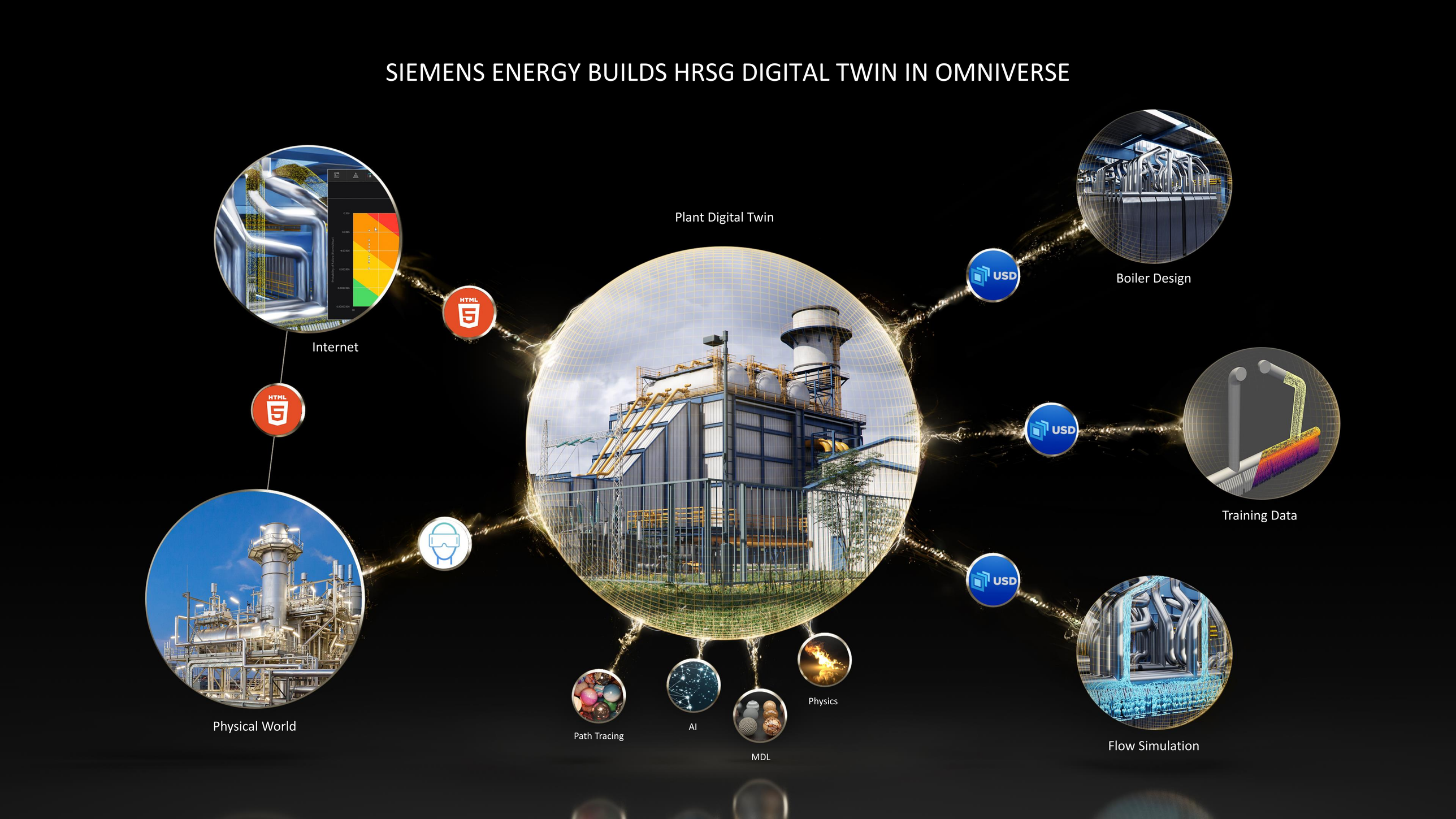
TIME



“Omniverse is our virtual world simulation engine. Robots, AV fleets, warehouses, factories, industrial plants, and cities will be created, trained, and operated as Omniverse digital twins.

“Siemens Energy is using Omniverse to create an industrial digital twin of a power plant.”

SIEMENS ENERGY BUILDS HRSG DIGITAL TWIN IN OMNIVERSE



“OMNIVERSE EMERGES AS HOME OF THE DIGITAL TWINS”

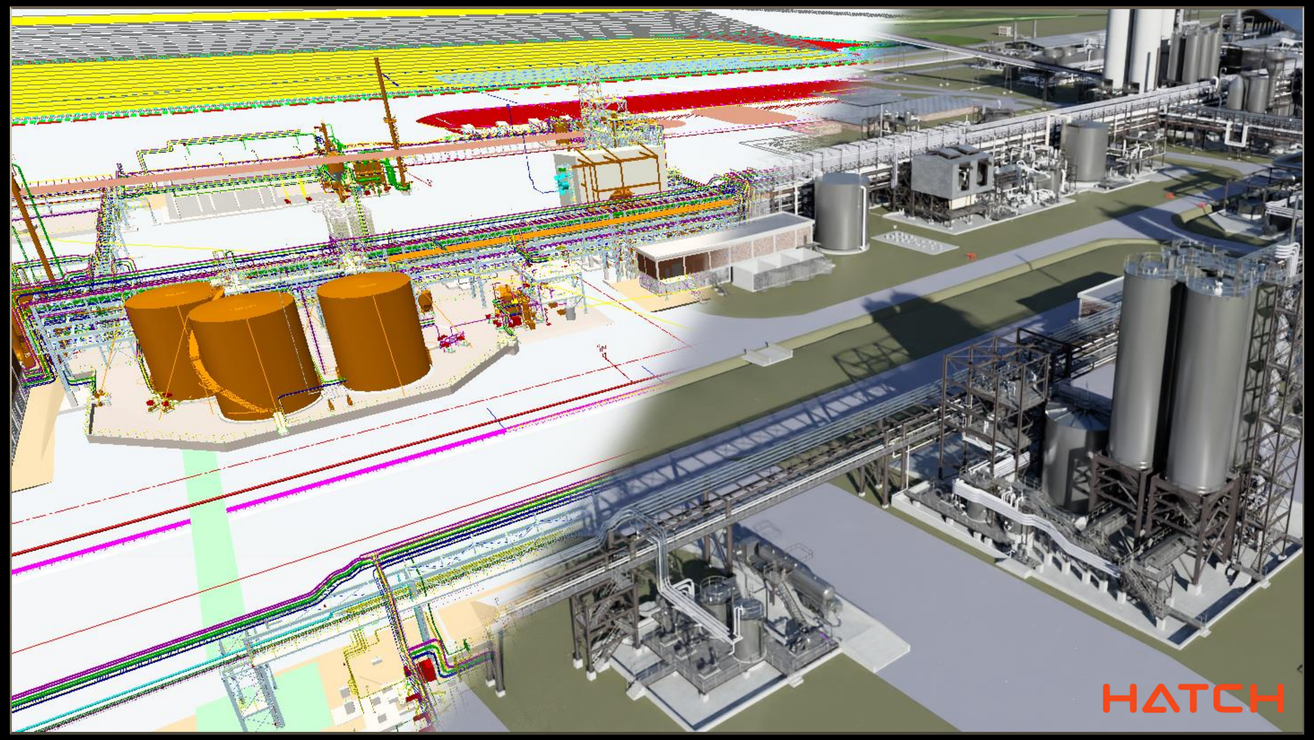
DIGITAL ENGINEERING

“BMW produces one vehicle per minute. Each has roughly 25,000 parts. There are 5 million parts on the factory floor at any time.

“At GTC Spring, BMW showed us how they are building a digital twin of their Regensburg factory. They have since expanded to three other factories totaling 10 million square meters.”

“BENTLEY, ESRI ANNOUNCE MAJOR INTEGRATIONS WITH NVIDIA OMNIVERSE”

GEO WEEK NEWS

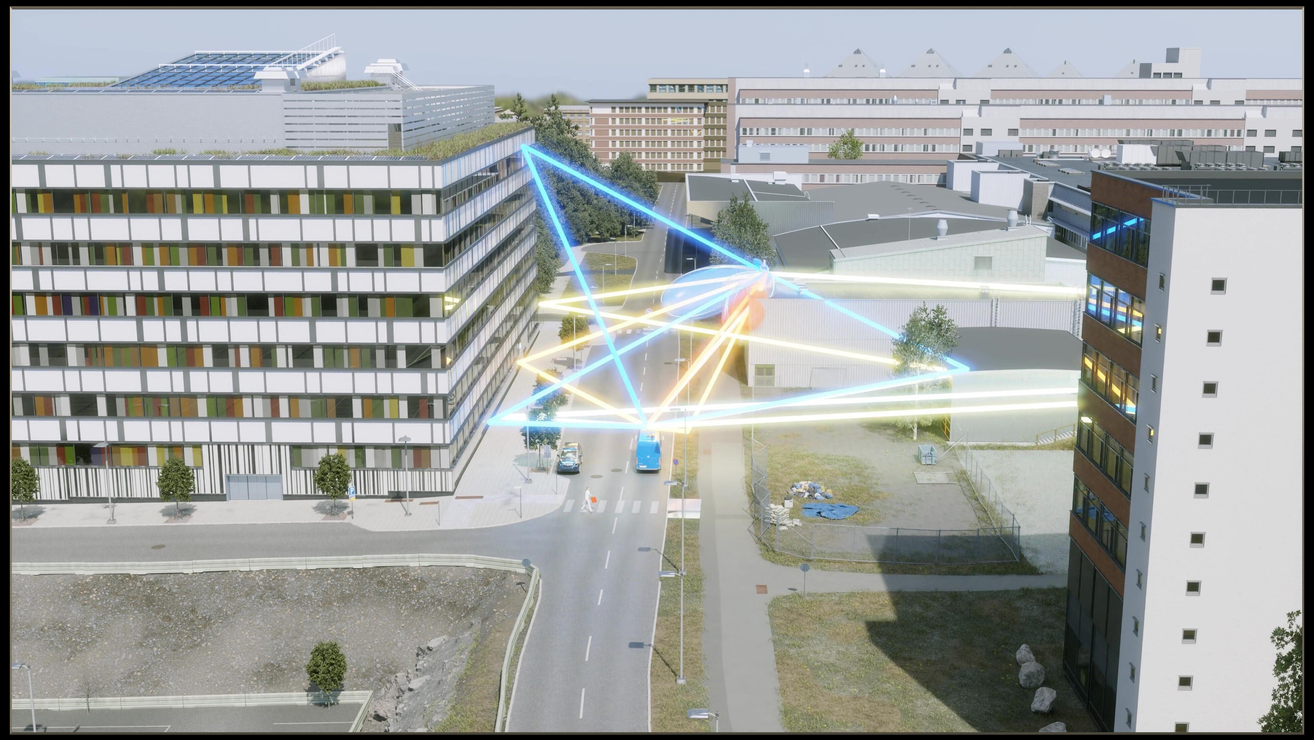


“Bentley is not just connecting to Omniverse, they’re building their digital platform on it.

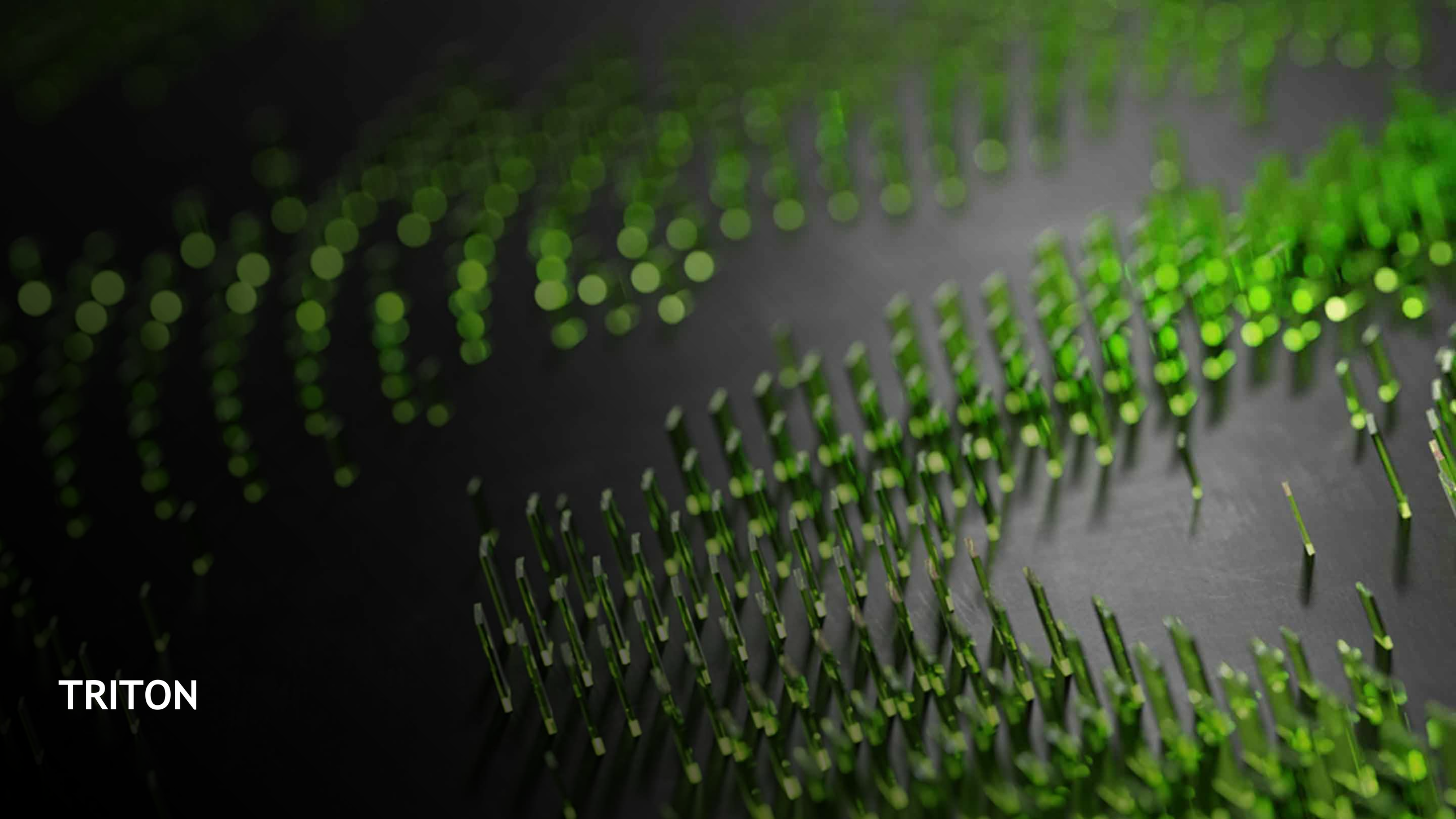
“Bentley is used by 90% of engineering firms and has nearly 2 million users of Bentley iTwin.”

“THE OMNIVERSE IS NVIDIA’S ‘METAVERSE FOR ENGINEERS,’ AND ERICSSON IS USING IT TO SIMULATE RECEPTION FOR 5G NETWORKS”

VENTUREBEAT

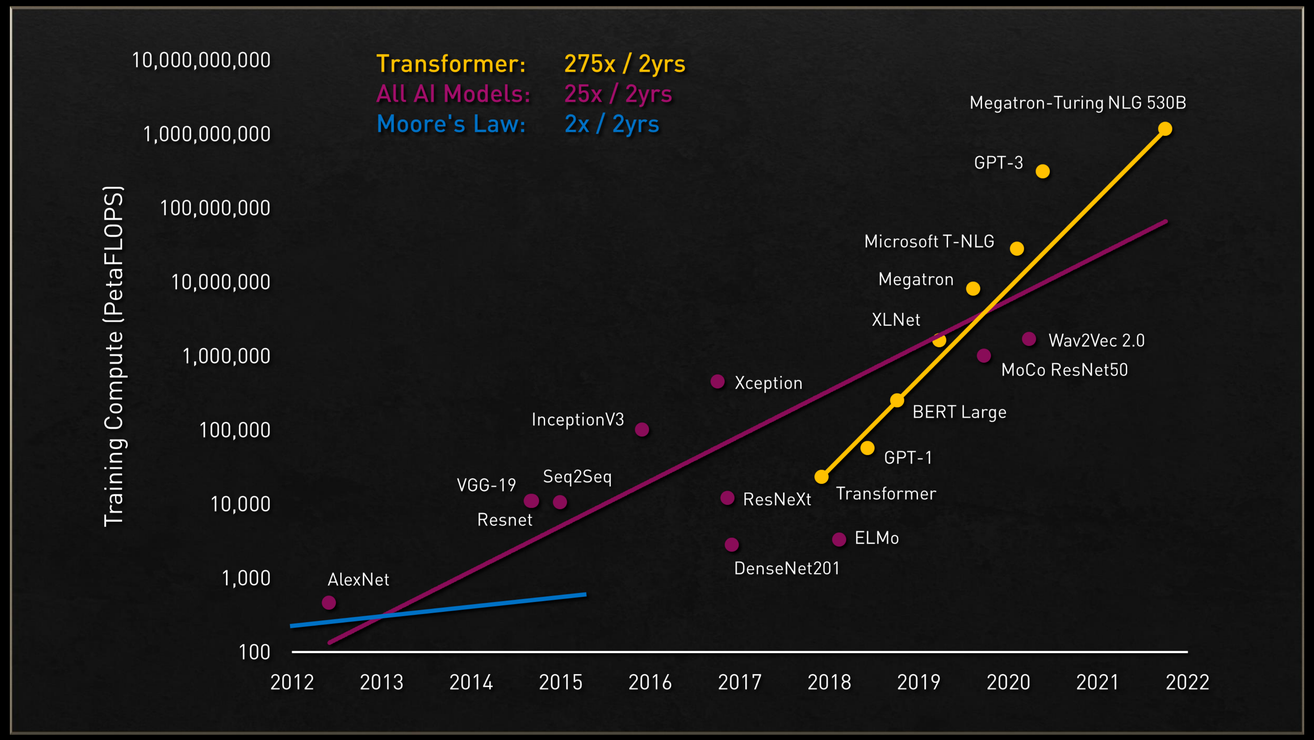


“Ericsson is using Omniverse to build a digital twin of an entire city to configure, operate, and continuously optimize their fleet of 5G antennas and radios.”


TRITON

“BRINGING LARGE LANGUAGE MODEL CAPABILITIES TO ENTERPRISES IS THE FOCUS OF NVIDIA’S NEW NEMO MEGATRON FRAMEWORK”

ENTERPRISE AI

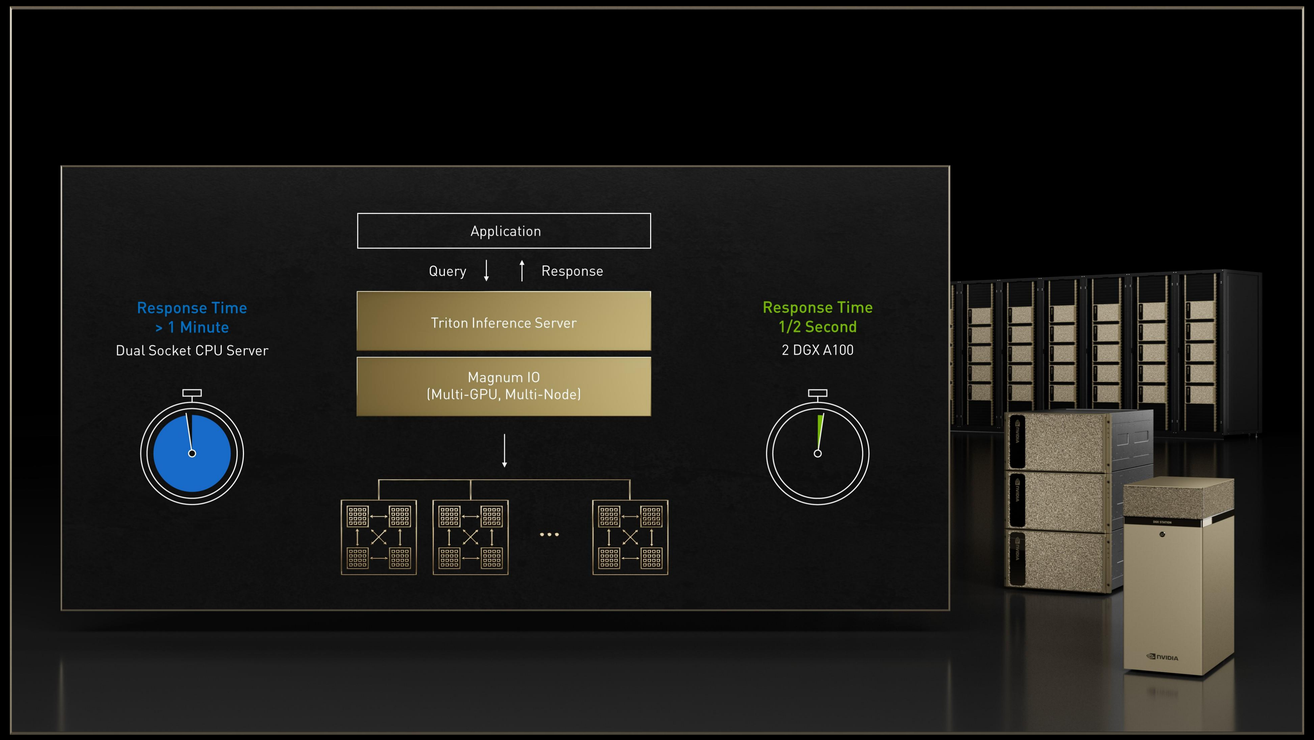


“Transformers have sped up language model training dramatically. No longer limited by human data labelling, giant self-supervised transformers benefit from the troves of digital knowledge on the internet.

“We created NeMo Megatron, a framework dedicated to training billion- and trillion-parameter speech and language models. Now any company can train state-of-the-art large language models.”

“NVIDIA HAS BOLSTERED ITS TRITON INFERENCE SERVER ”

THE NEXT PLATFORM



“Once trained, the inference response time on a large language model has to be sufficiently fast to be useful. On a high-end dual Xeon Platinum CPU server, inferencing Megatron 530B takes over a minute. For many applications, that’s unusable.

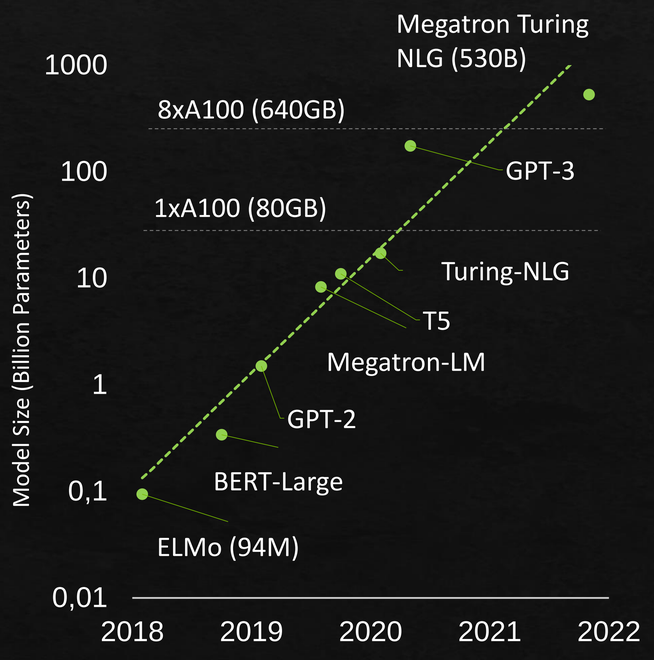
“We created the world’s first distributed inferencing engine. NVIDIA Triton now does distributed processing across multiple GPUs and multiple nodes. And Triton can also do inferencing on both DL and ML models, including widely used tree models like XGBoost, Random Forest, and LightGBM.”

TRITON MULTI-GPU MULTI-NODE INFERENCE

Fast Inference on Giant Transformer Models

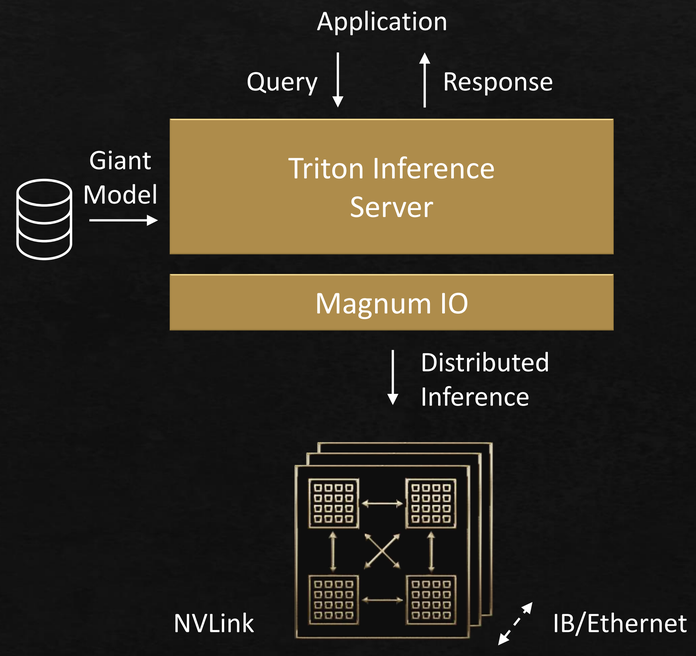
MODELS GROWING EXPONENTIALLY

Cannot Fit Single GPU



TRITON DISTRIBUTED INFERENCE

Scales To Trillion Parameter Models

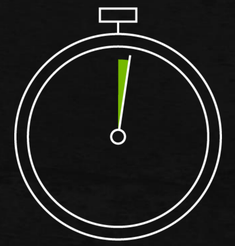


REAL-TIME GIANT MODEL INFERENCE

Megatron Turing NLG — 530B

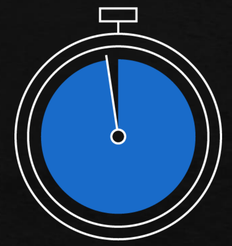
Response Time
1/2 Second

2 DGX A100



Response Time
> 1 Minute

Dual Socket CPU Server

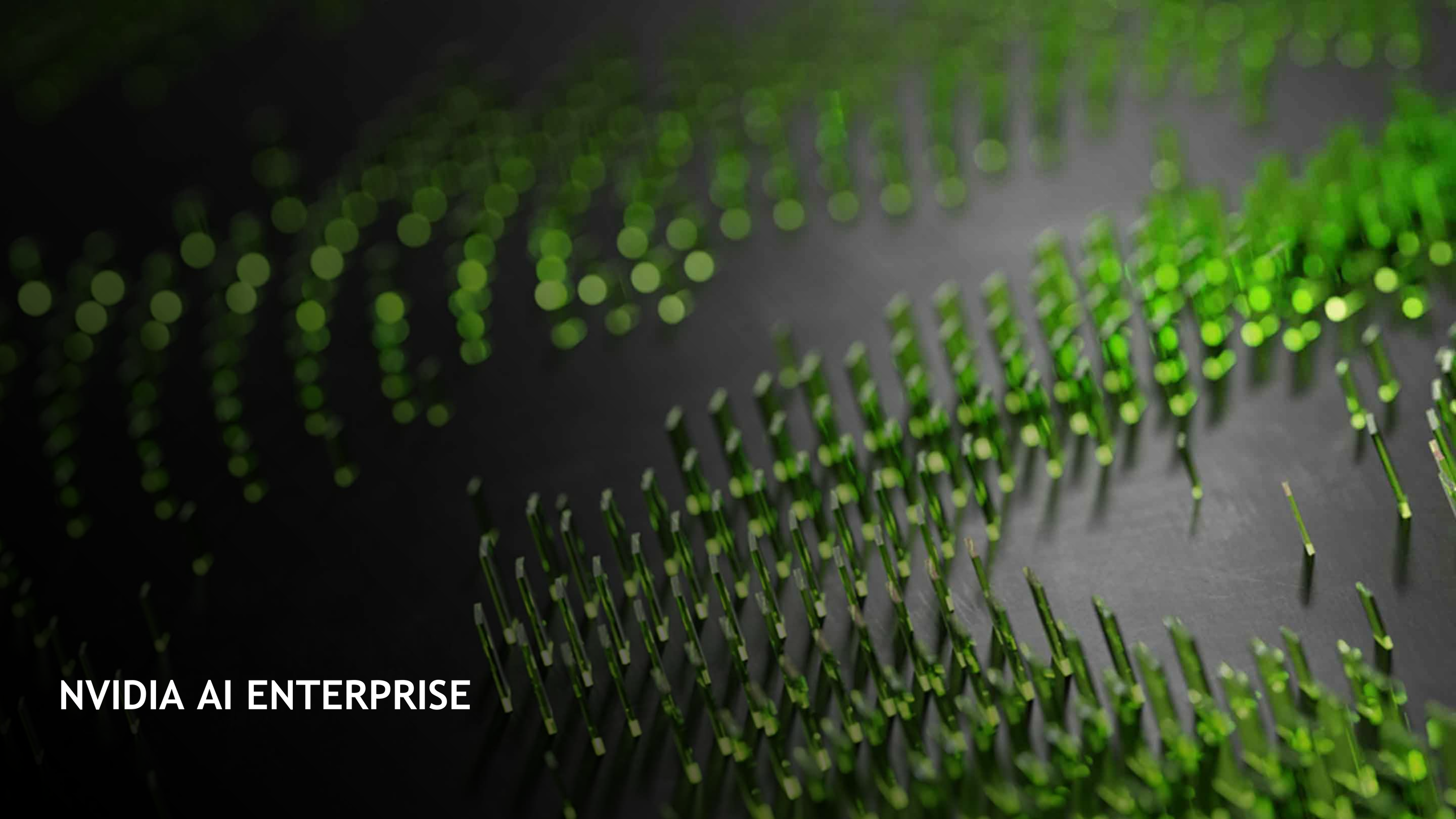


Input sequence length=128 tokens (average of 102 words), Output sequence length=8 tokens (average of 6 words)
GPU: DGX-A100-80GB, Batch size=1, FP16, FasterTransformer 4.0, Triton 2.15
CPU: Xeon Platinum 8280 2S, Up to 1TB/socket System memory, Batch size=1, FP32, TensorFlow

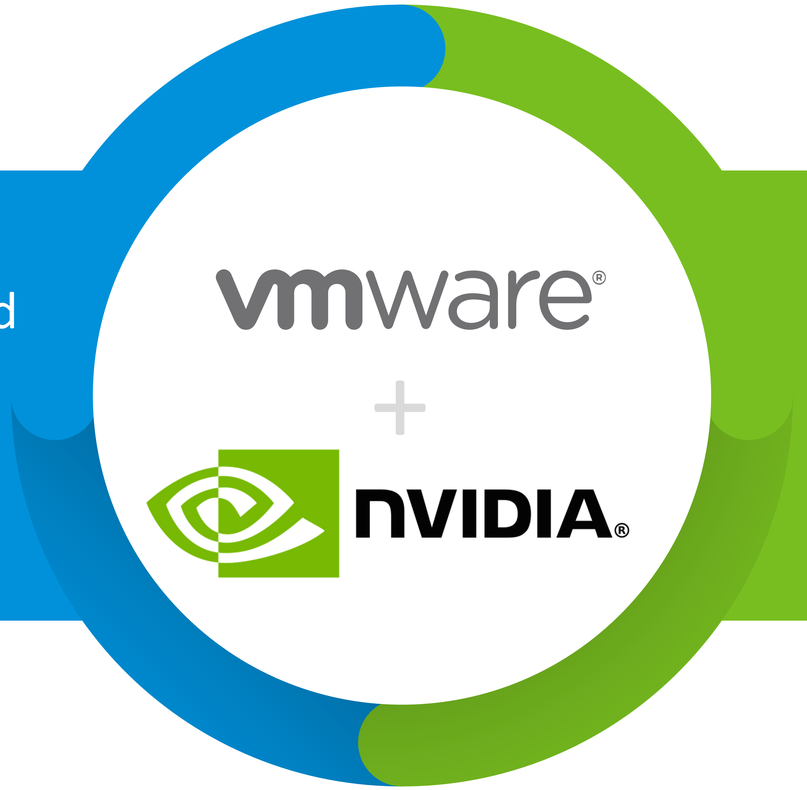
“LIVE CAPTIONING AND TRANSCRIPTION IN MICROSOFT TEAMS JUST GOT BETTER WITH MICROSOFT AZURE AND NVIDIA AI”

NEWSCOLLIGENS

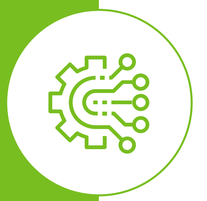
“Video conferencing is the most important app for most of us today. Microsoft Teams has over 200 million active users. We’re delighted to work with Microsoft to develop live captioning across 28 languages.”


NVIDIA AI ENTERPRISE

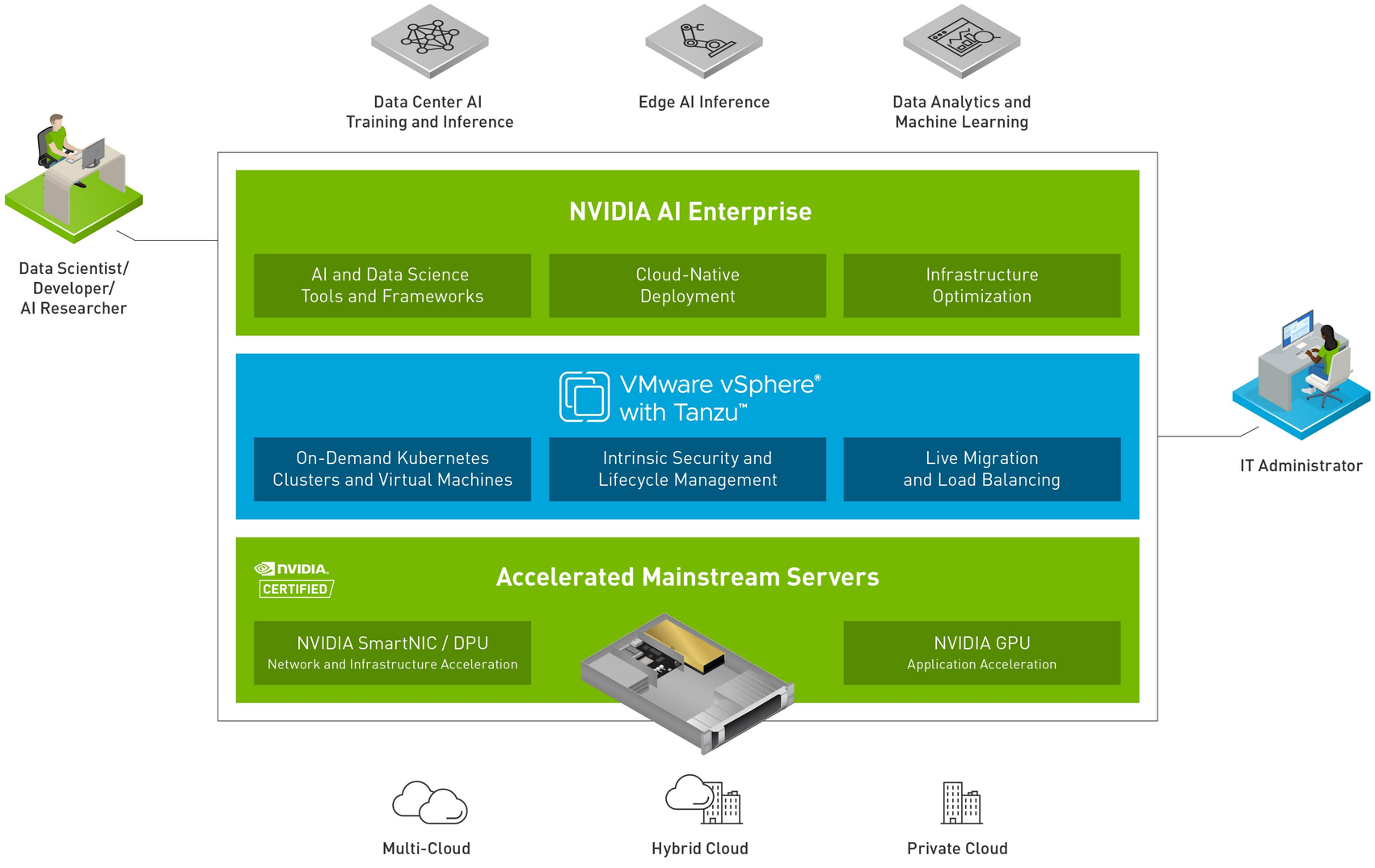
Delivering next gen hybrid cloud infrastructure for next gen apps



Extending AI to every enterprise in the data center, cloud and edge

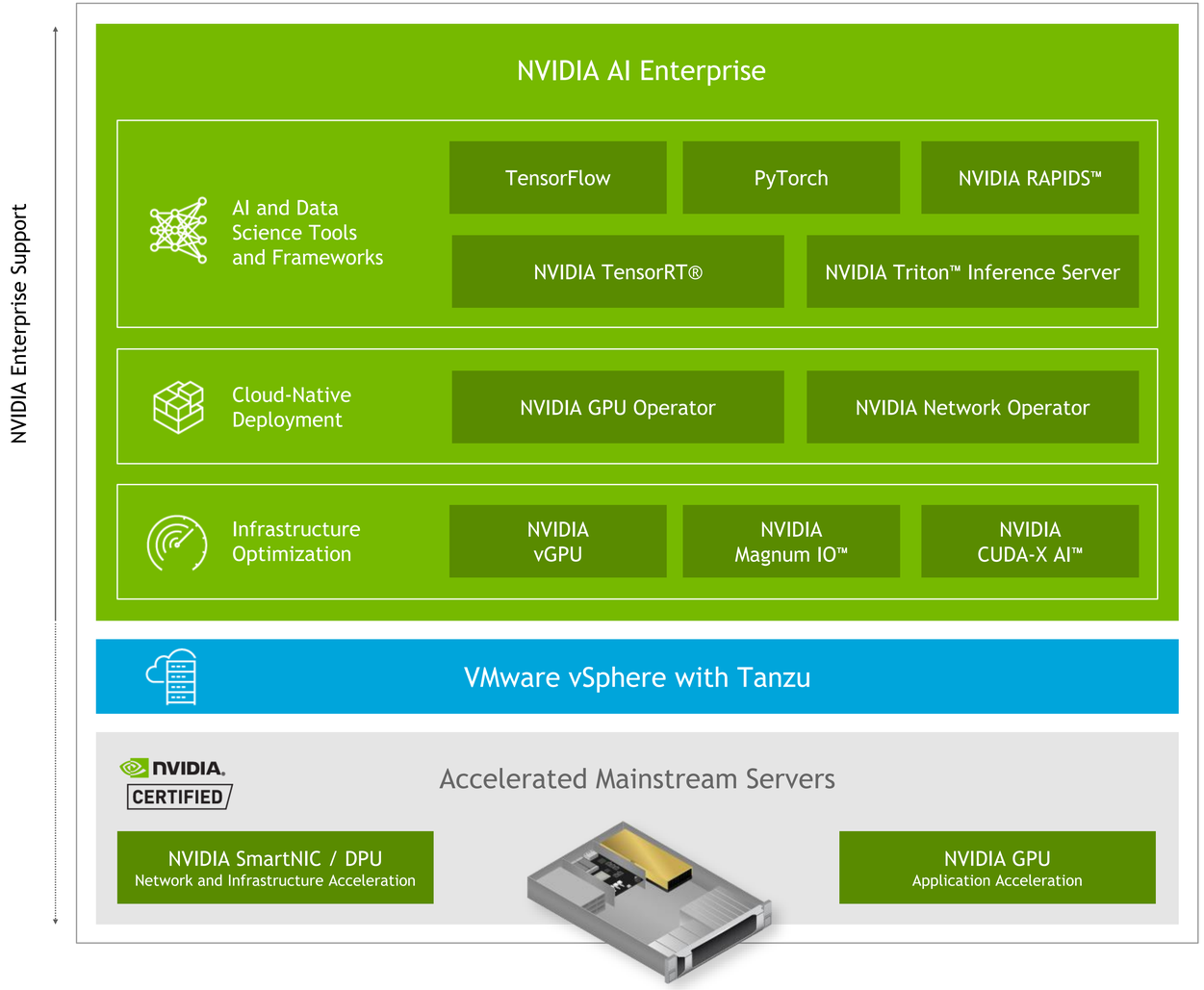


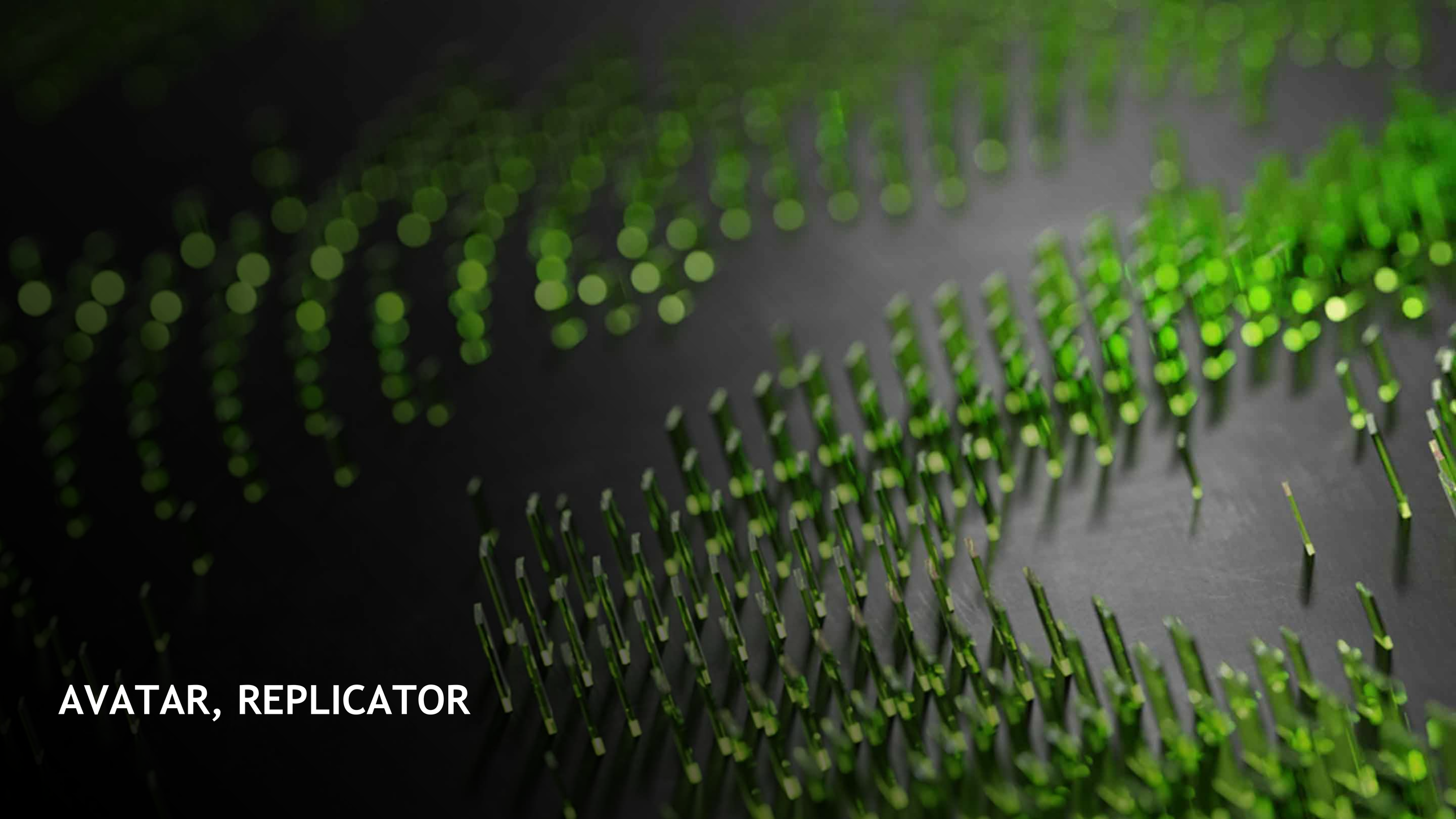
AI-READY ENTERPRISE PLATFORM



NVIDIA AI ENTERPRISE SOFTWARE SUITE

Exclusively Certified on VMware vSphere 7




AVATAR, REPLICATOR

“NVIDIA’S NEW AI-POWERED AVATARS ARE PRETTY INCREDIBLE”

VRSCOUT

“To help developers create interactive characters with Omniverse that can see, speak, converse on a wide range of subjects, and understand naturally spoken intent, we created Omniverse Avatar.

“Omniverse Avatar connects Riva speech AI, computer vision, natural language understanding, recommendation engines, and simulation technologies into a real-time conversational AI robot. A demo of the technology featured ‘Toy Jensen’ fielding questions on climate change, methods to find exoplanets, and the role of proteins in human cells.”

“NVIDIA’S SPECTACULAR DEMO OF ITS AVATAR CREATION TOOLS”

KULTURE GEEK

“Project Maxine, another example of an Omniverse Avatar application, can add state-of-the-art video and audio features to virtual collaboration and content creation applications.

“This demo shows a woman speaking English on a video call in a noisy cafe, but she can be heard clearly without background noise. As she speaks, her words are transcribed and translated in real time into French, German, and Spanish. Maxine also simulates eye contact by estimating and aligning gaze with the camera.”

“OMNIVERSE REPLICATOR GOES FURTHER. IT’S NOT A 3D MODELING ENGINE — IT’S A SYNTHETIC DATA GENERATION ENGINE”

DIGITAL TRENDS

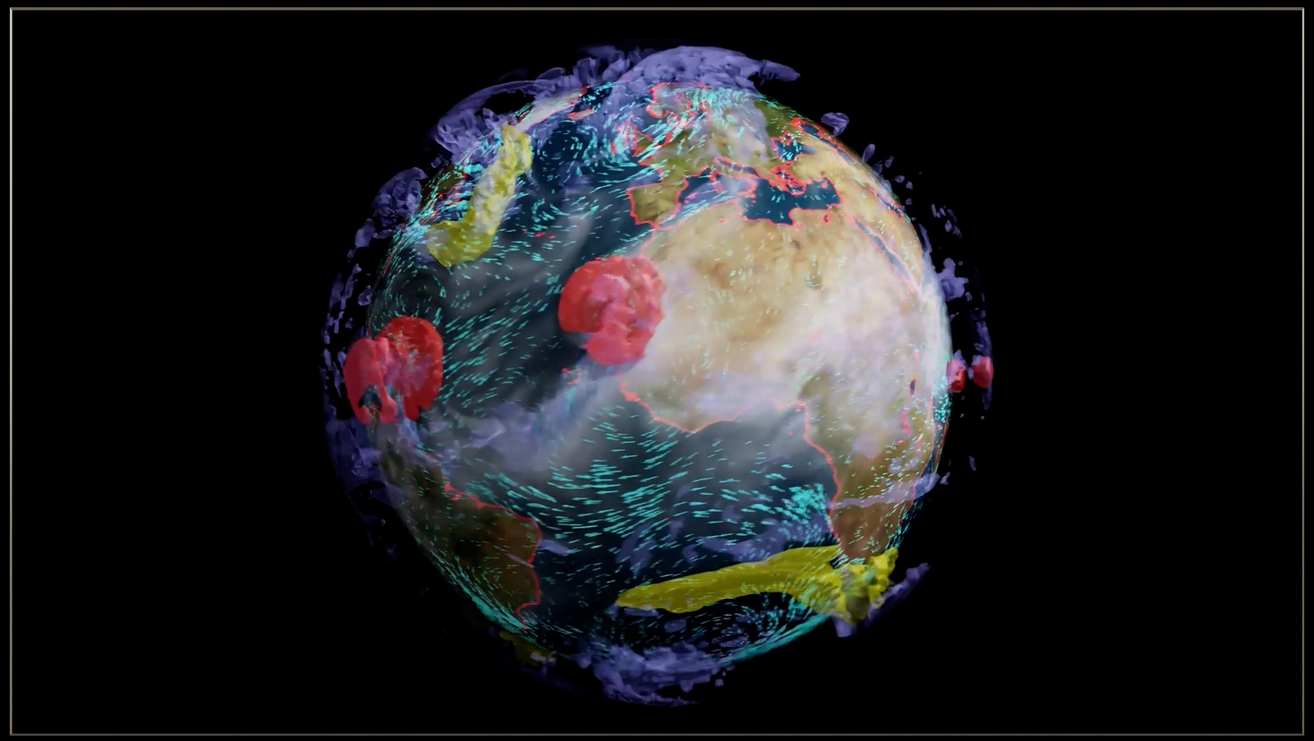
“To help developers to create the huge amounts of data needed to train AI, we’re introducing Omniverse Replicator, a synthetic-data-generation engine for training deep neural networks.

“Replicator has two applications: one for NVIDIA DRIVE Sim, a virtual world for hosting the digital twin of autonomous vehicles, and another for NVIDIA Isaac Sim, a virtual world for the digital twin of manipulation robots.

“Replicator can label ground truth in ways that humans cannot — tracking moving objects across sensors, velocity, distance, occlusion, and severe weather conditions. It is accurate and low cost. And it augments data where we have known gaps.”

“NVIDIA CREATES DIGITAL TWIN OF EARTH TO BATTLE CLIMATE CHANGE”

VENTUREBEAT



“The past seven years are on track to be the seven warmest on record. We’re experiencing extreme weather — historic droughts, unprecedented heatwaves, intense hurricanes, violent storms, and catastrophic floods.

“For the first time, we have the technology to do ultra-high-resolution climate modeling, to jump to lightspeed and predict changes in regional extreme weather decades out.

“At GTC, we revealed plans to build the world’s most powerful AI supercomputer dedicated to predicting climate change. Named Earth-2, or E-2, the system would create a digital twin of Earth in Omniverse.”


“MAYBE THE COOLEST COMPANY IN THE WORLD”
LibertyRPF Substack Blog


NVIDIA.COM/GTC
NVIDIA