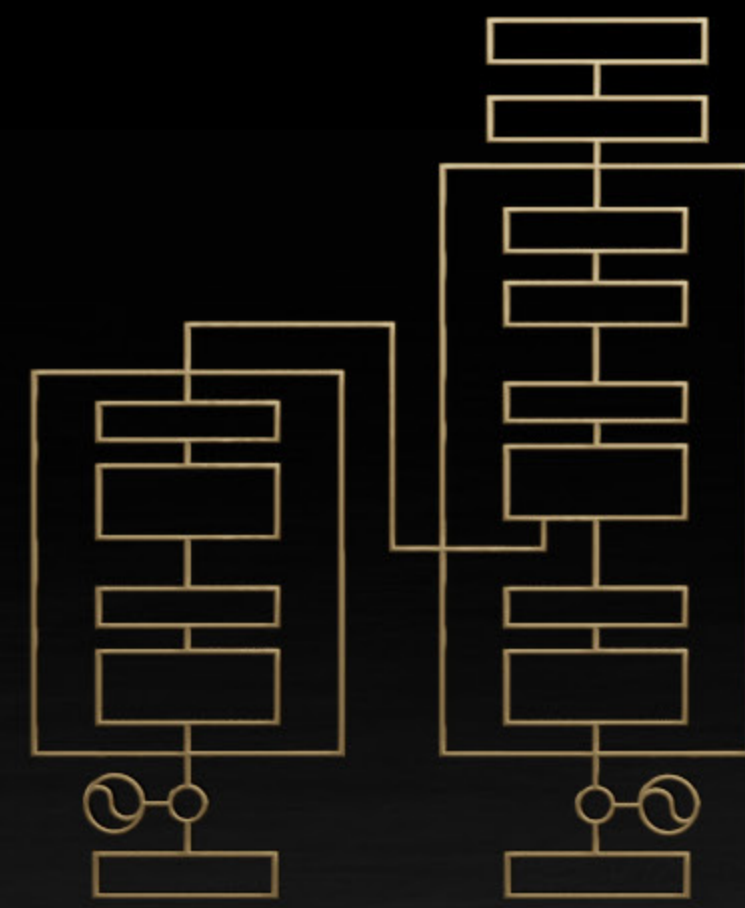




Accelerated Computing
is the Path Forward



AI is Software that
Writes Software



Data Center is the
New Unit of Computing



AI-on-5G Kickstarts the
4th Industrial Revolution



Autonomous Systems in
Real and Virtual Worlds



NEW NVIDIA TECHNOLOGIES

Omniverse
Isaac

Megatron
Drug Discovery
Quantum Computing

Jarvis
Merlin
Maxine
Morpheus
NVIDIA AI

DRIVE

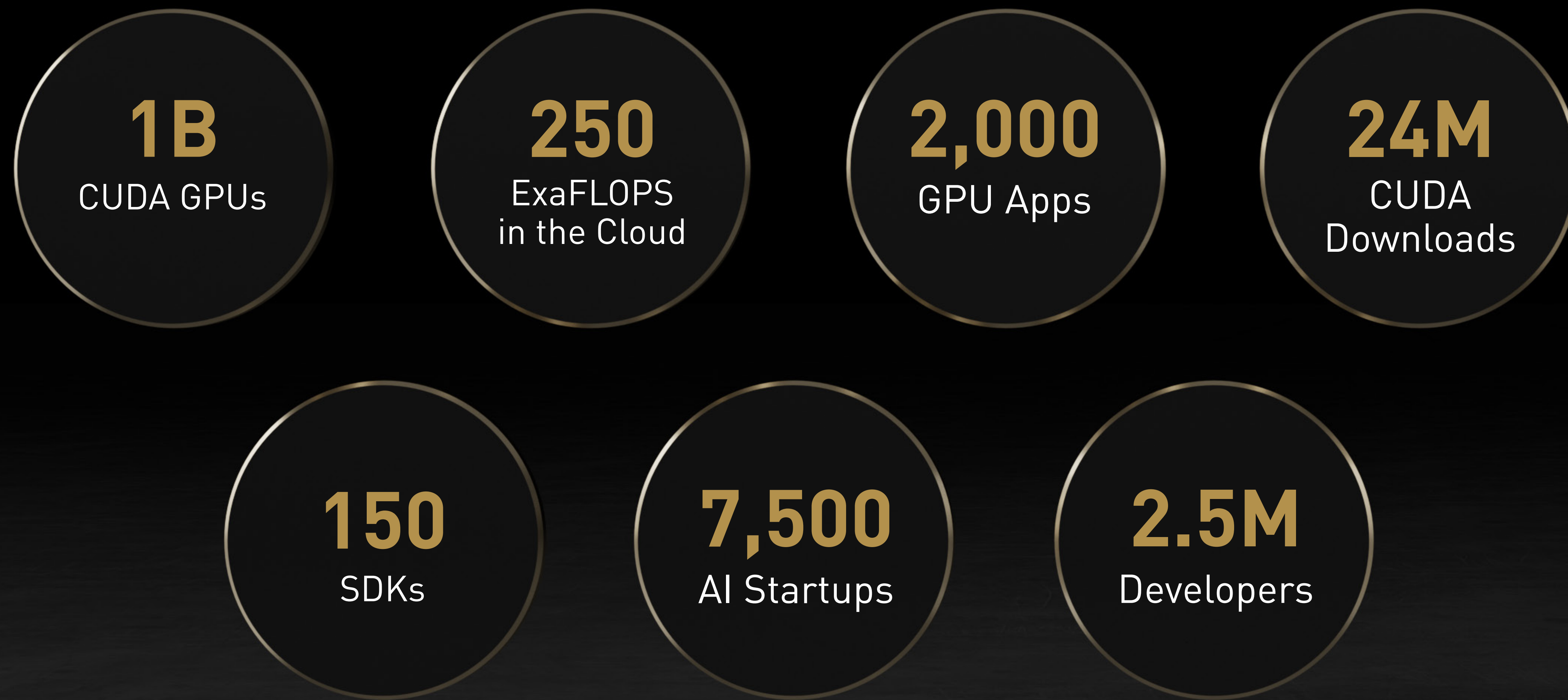
RTX

DGX
Grace
BlueField
DOCA

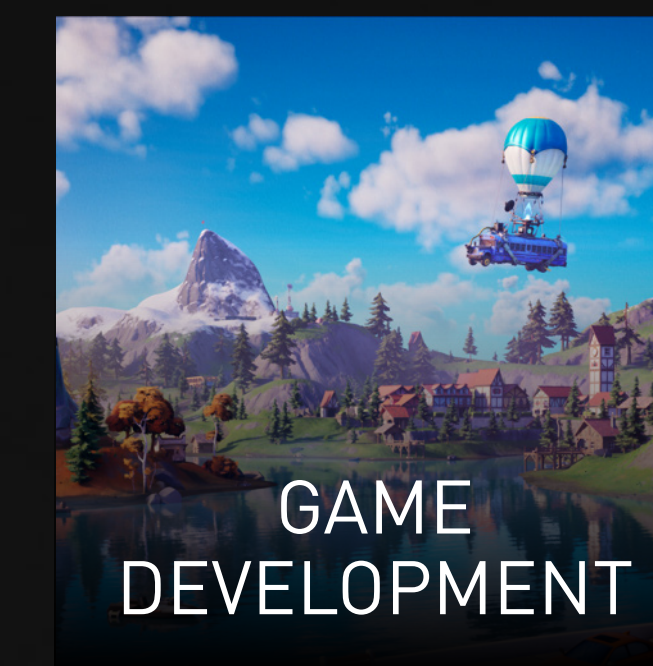
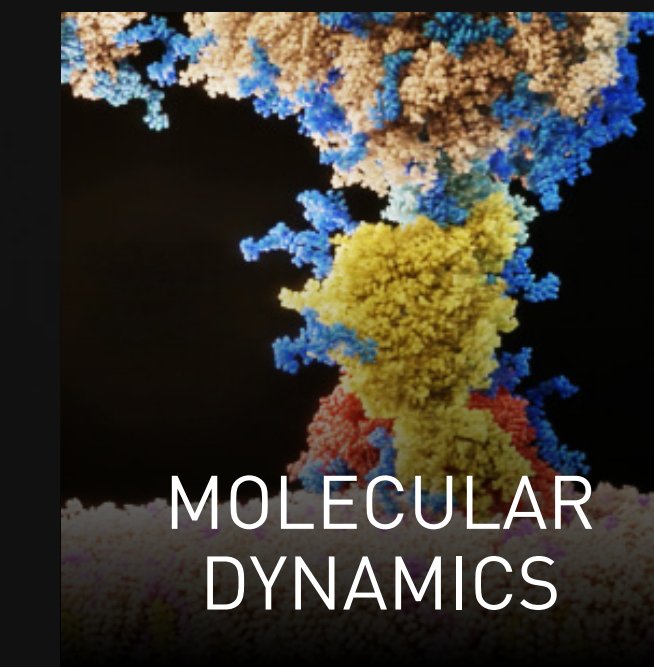
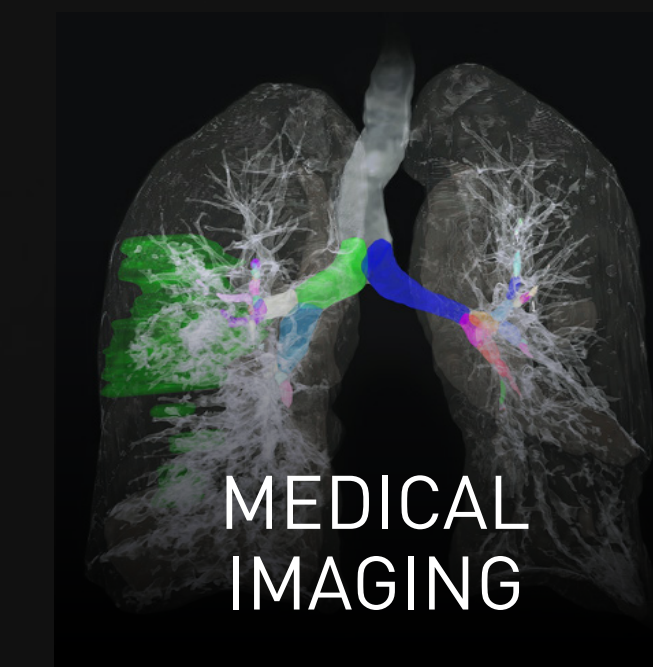
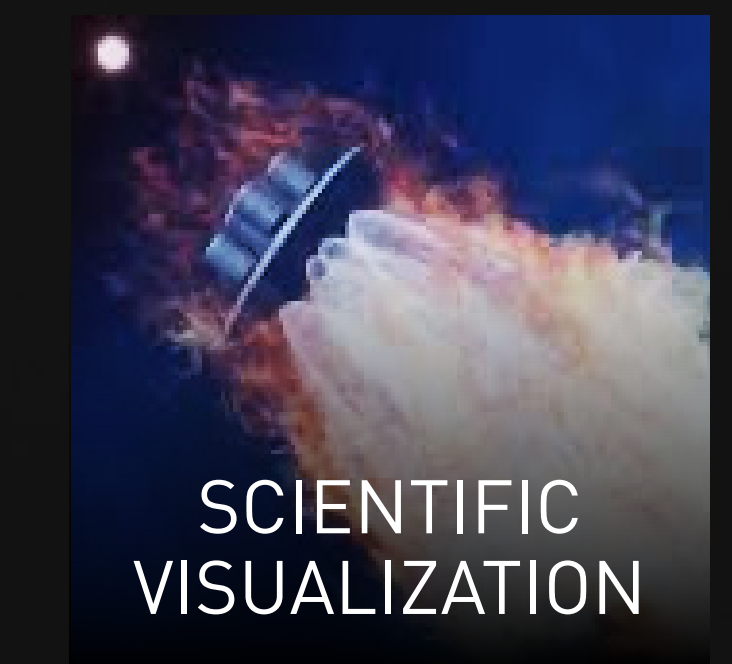
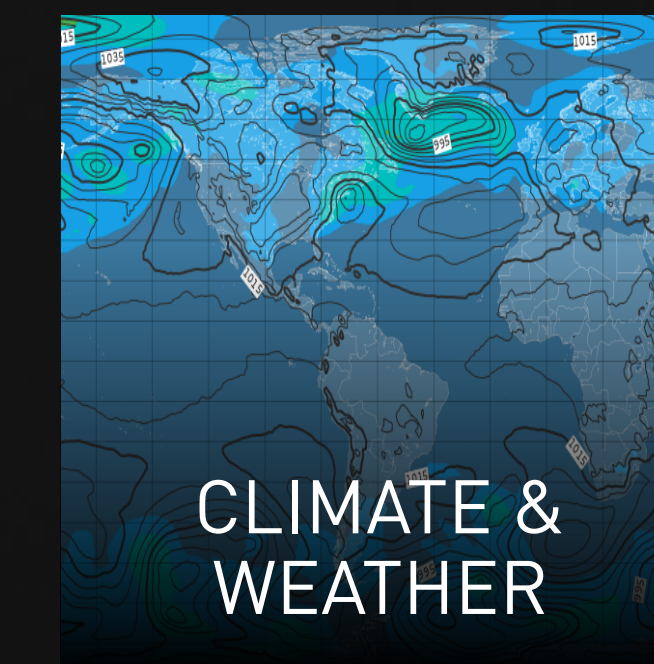
EGX
5G

Hyperion
Atlan
Orin

NVIDIA IS A COMPUTING PLATFORM



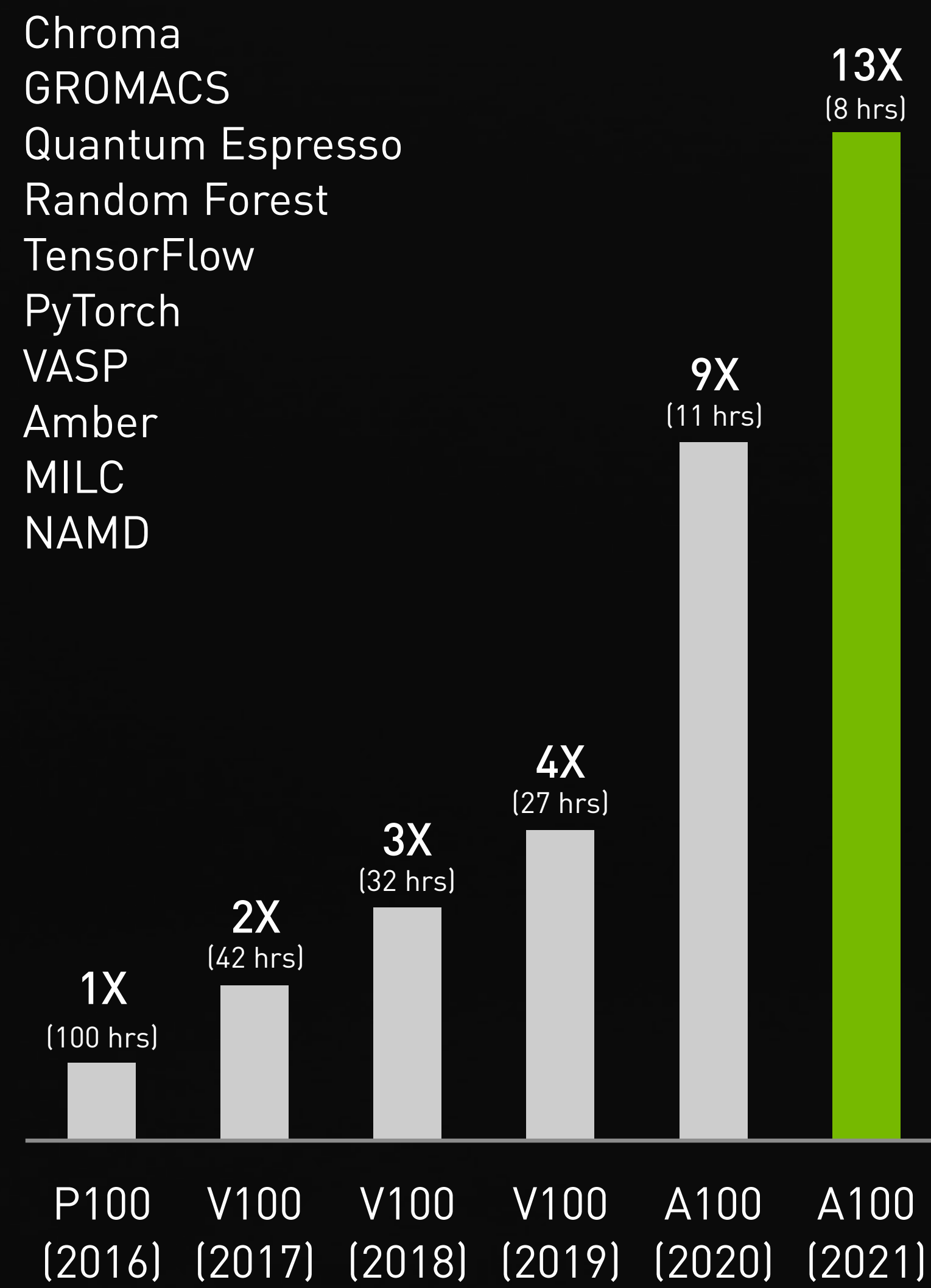
AN INSTRUMENT OF SCIENCE



GIVING SCIENTISTS A TIME MACHINE

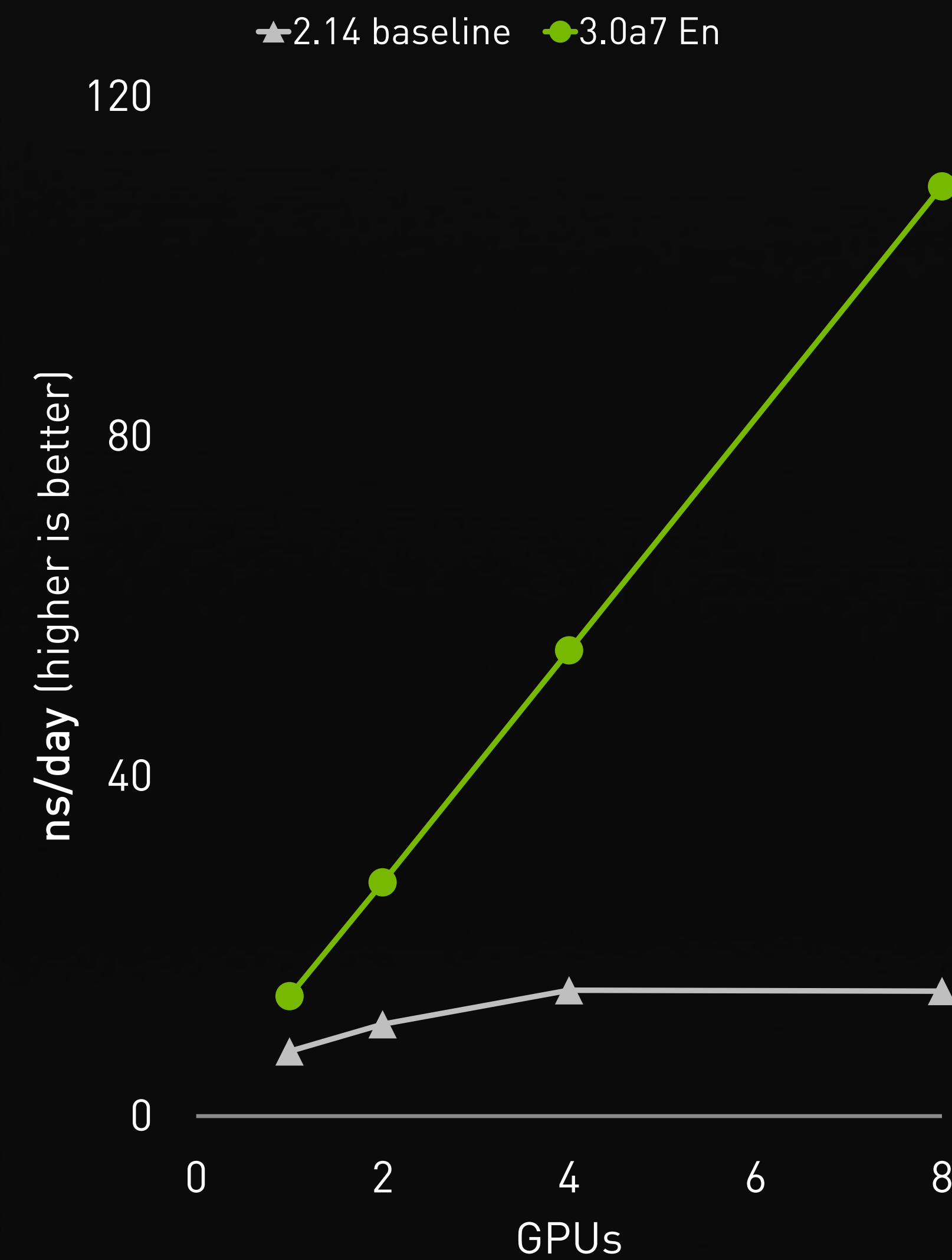
FULL-STACK OPTIMIZATION

13x in 5 years



MULTI-GPU MULTI-NODE SCALE

NAMD Version 3.0



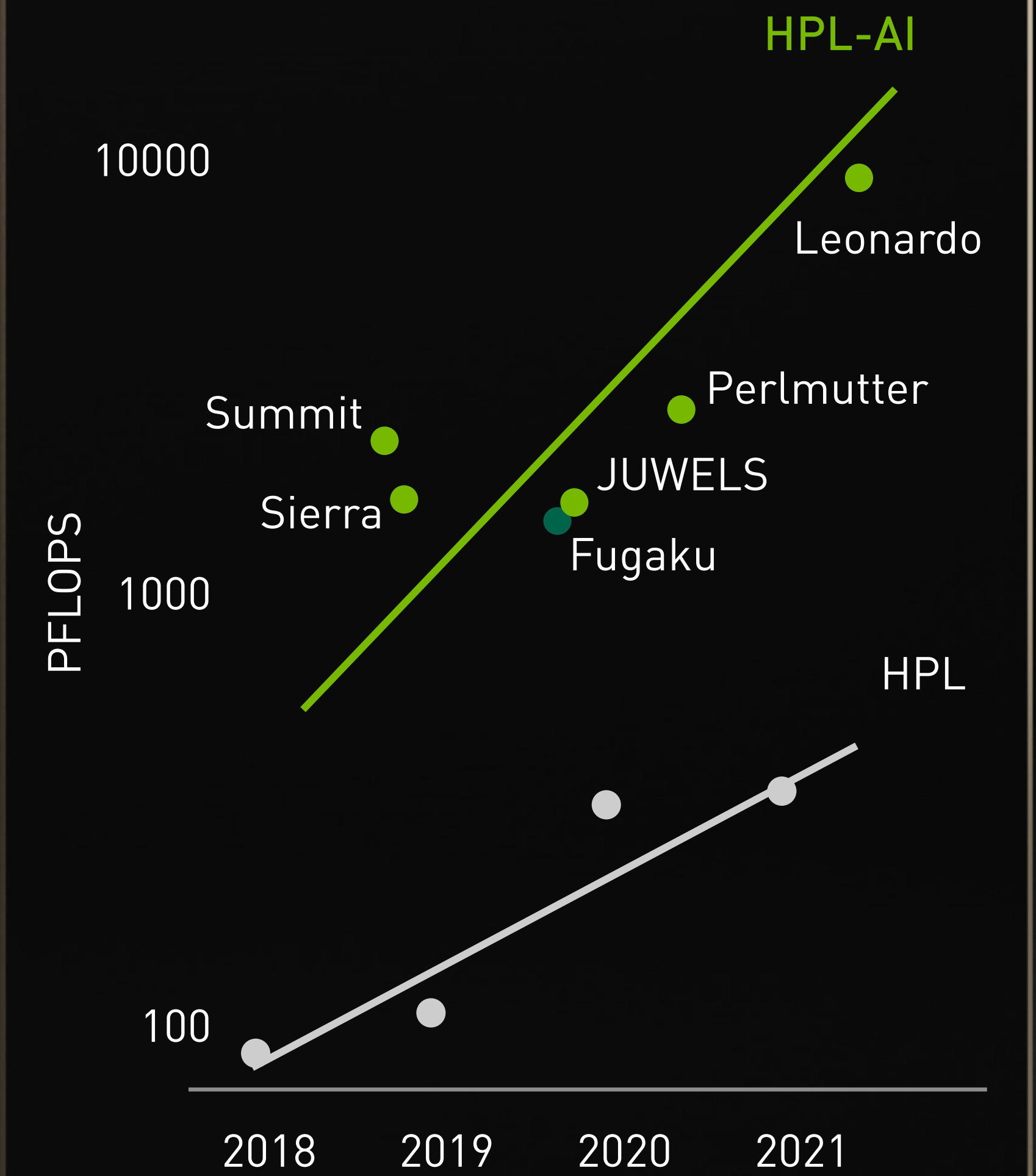
LARGEST AI + MD SIMULATION

305 Million Atoms



EXASCALE HPC AI

HPL-AI and HPL Performance



FOR THE DA VINCIS OF OUR TIME

SCIENTISTS AT GTC

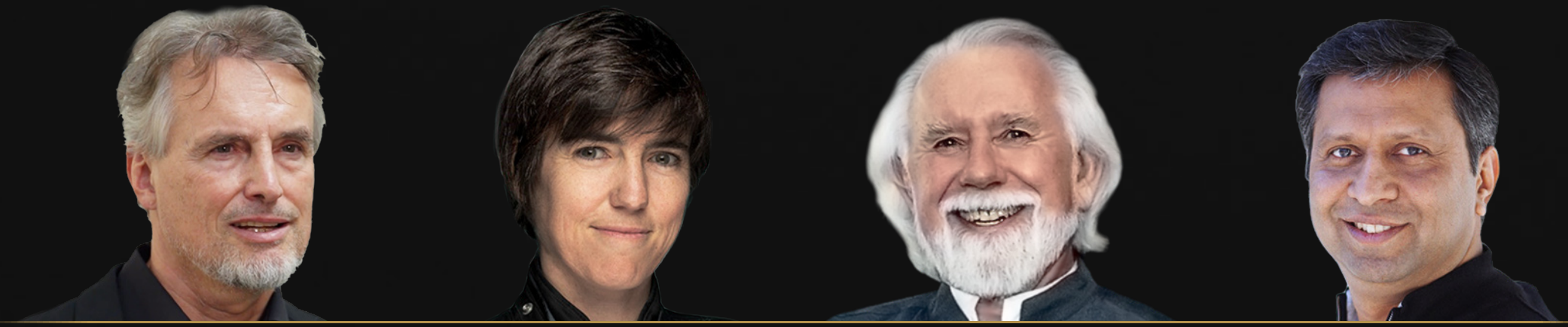


Yoshua Bengio
 University of Montreal
 Quebec AI Institute

Yann LeCun
 Facebook
 New York University

Geoffrey Hinton
 University of Toronto
 Google
 Vector Institute

Daphne Koller
 Insitro
 Coursera
 Stanford



Jürgen Schmidhuber
 Dalle Molle Institute for
 AI Research

Raquel Urtasun
 University of Toronto

Alvy Ray Smith
 Pixar
 Altamira

Abhay Parasnis
 Adobe



Kim Libreri
 Epic Games

Rommie Amaro
 University of
 California, San Diego

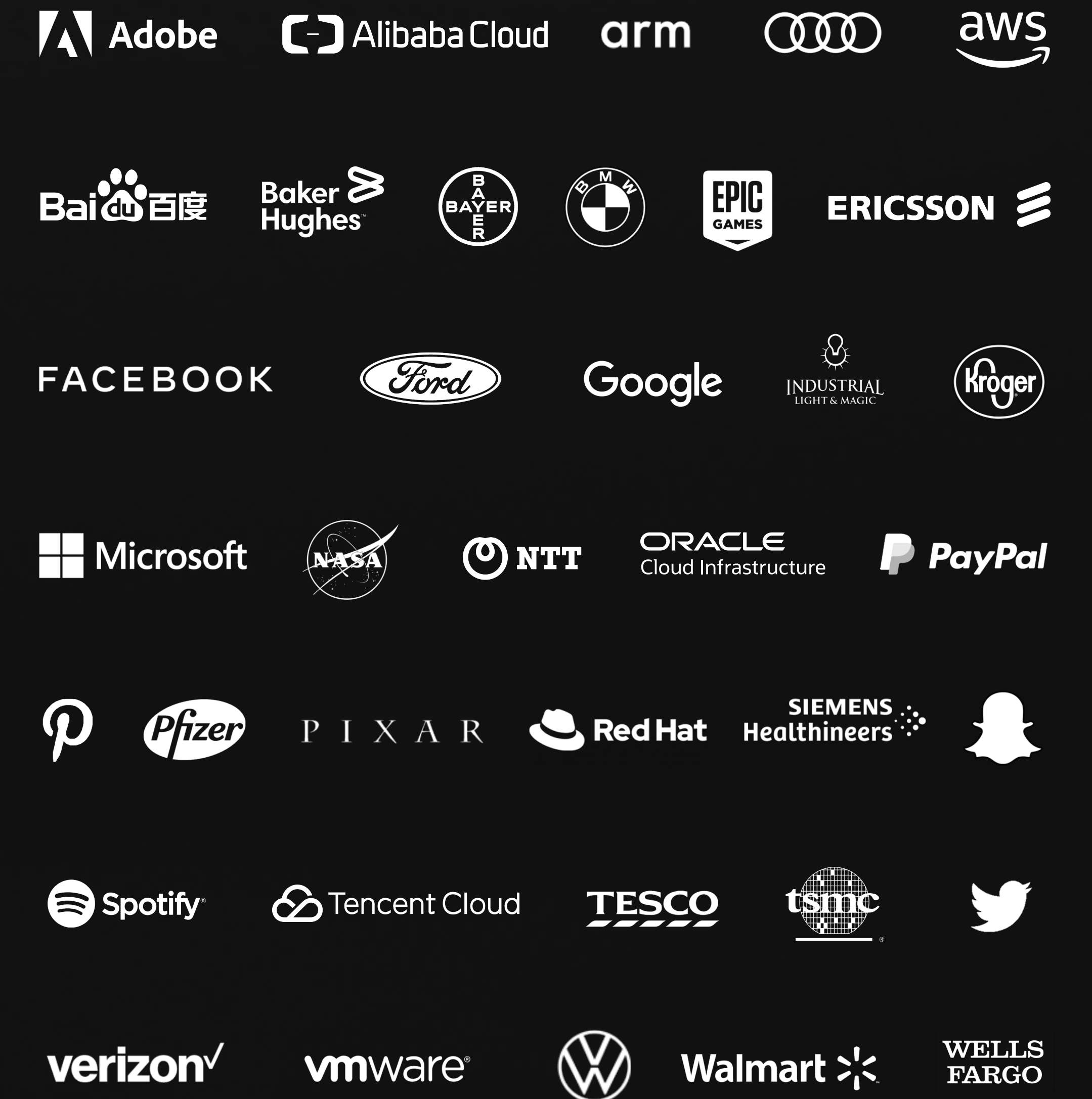
Soumith Chintala
 Facebook

Rose Yu
 University of
 California, San Diego

TALKS AT GTC

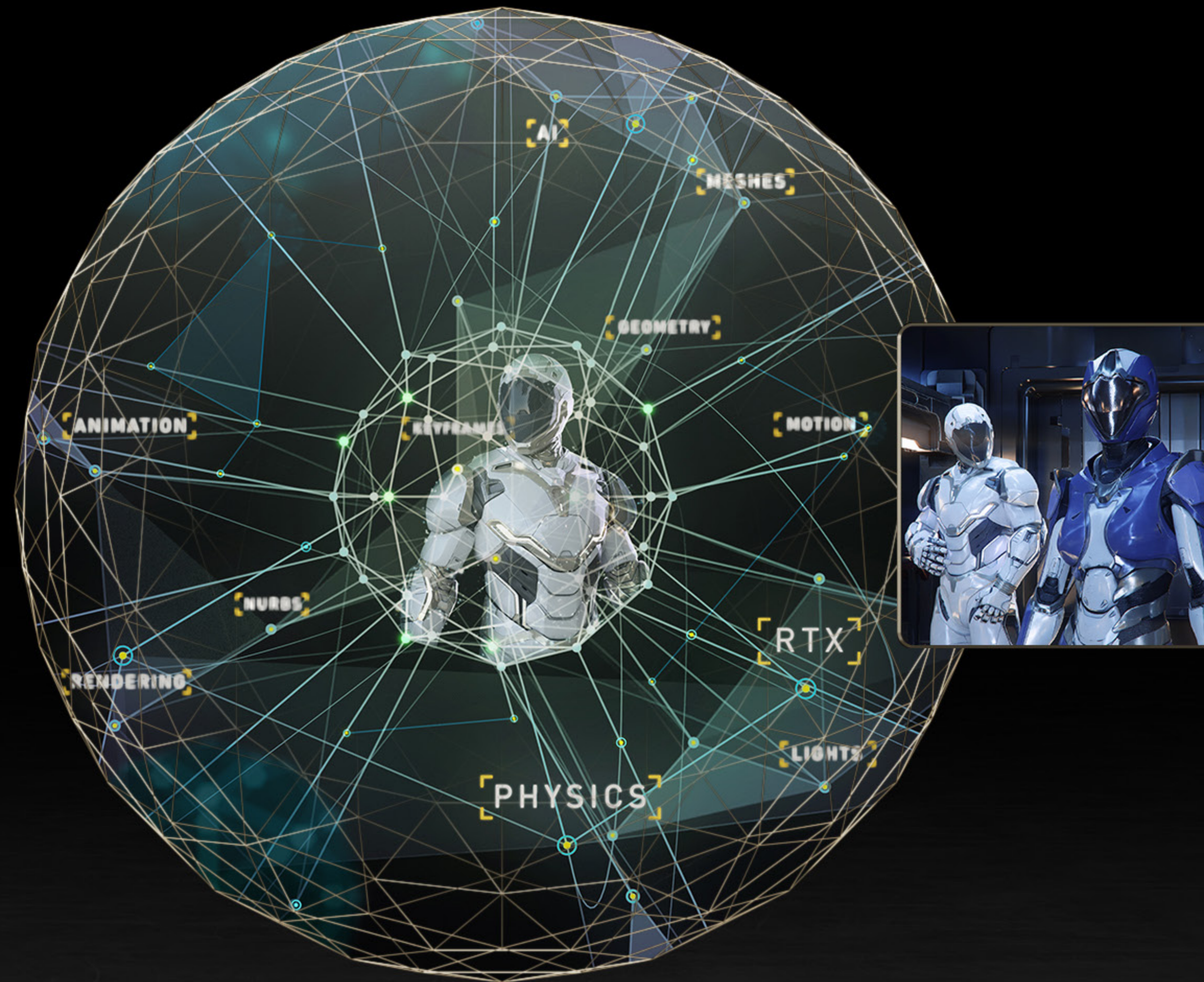
AI	5G	IOT & EDGE
QUANTUM COMPUTING	SPEECH NLU RECOMMENDERS	SELF-DRIVING CARS
CYBERSECURITY	DIGITAL TWINS	ROBOTICS

LEADERS AT GTC





NVIDIA OMNIVERSE



AI



Path-Tracing



USD

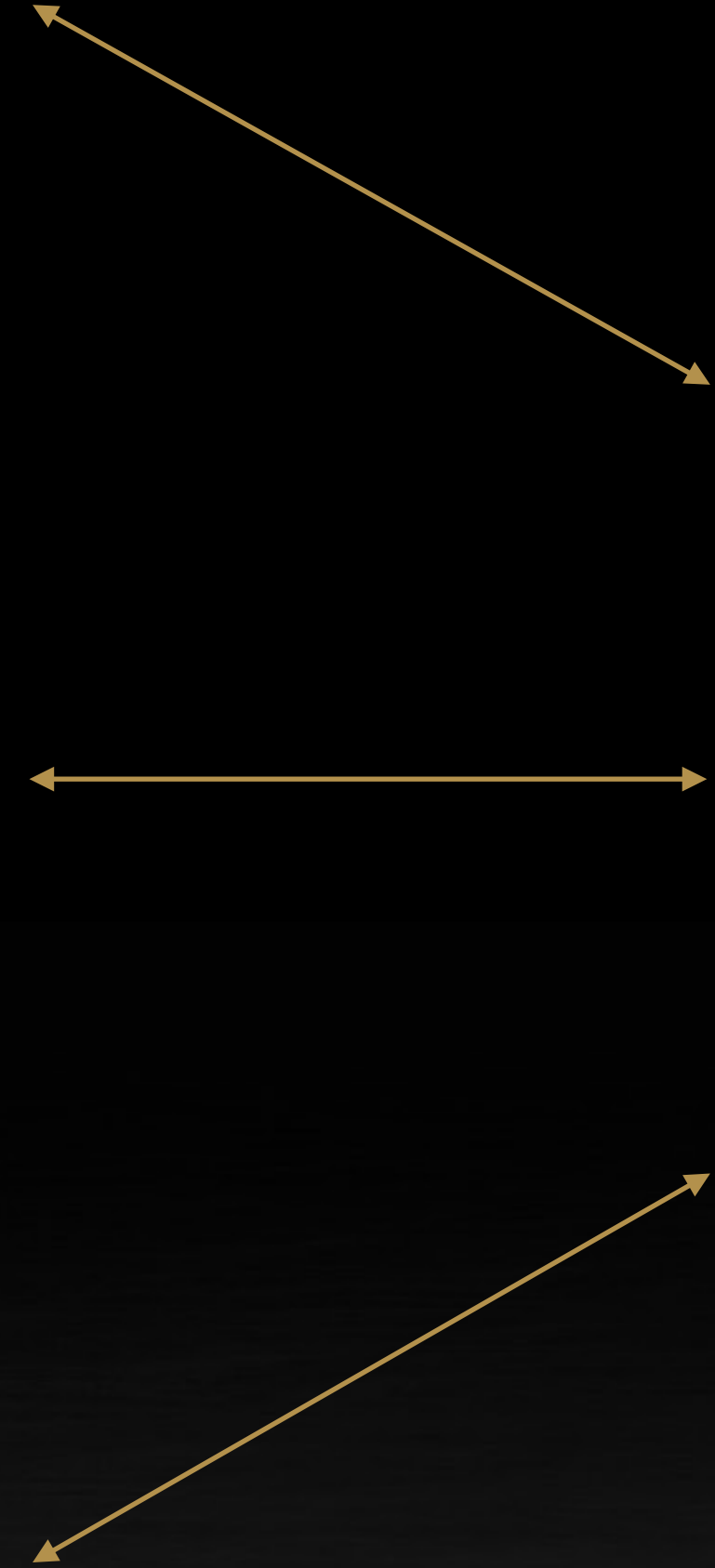


Materials



Physics

NVIDIA OMNIVERSE



AI



Path-Tracing



USD

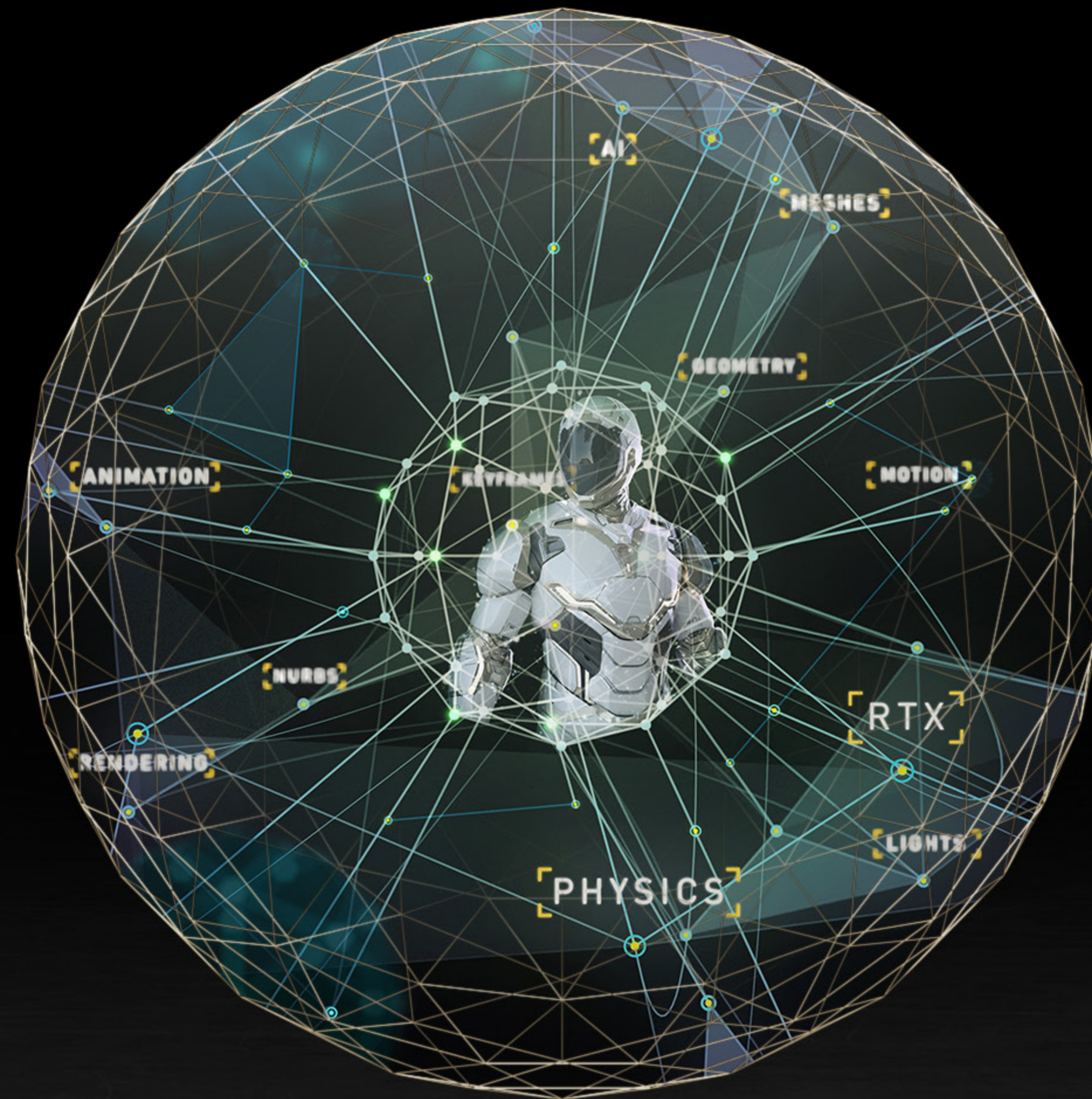
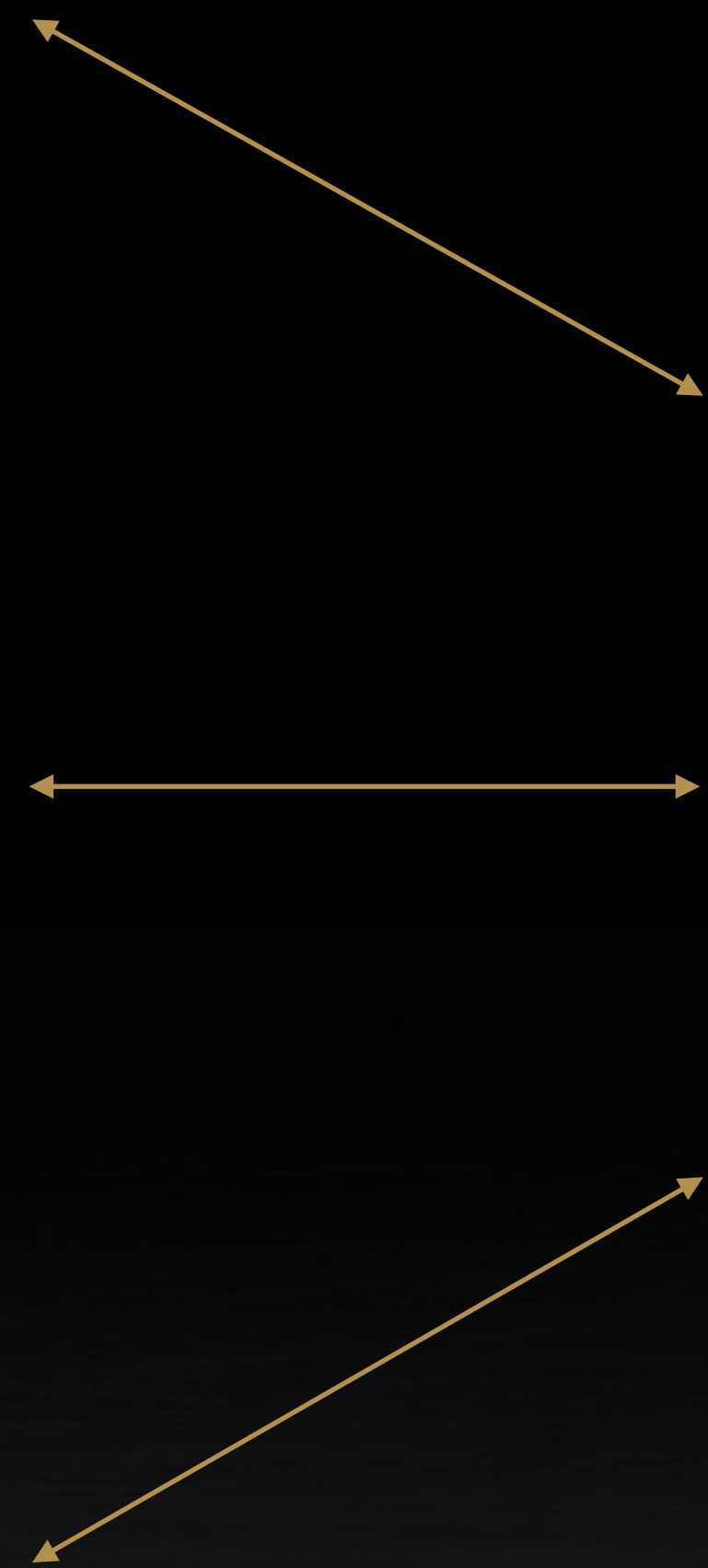
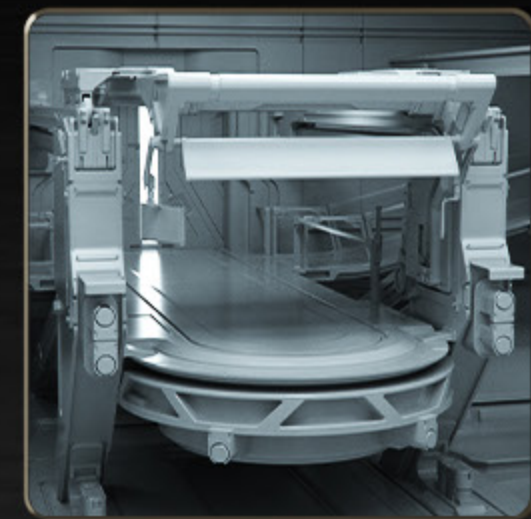


Materials



Physics

NVIDIA OMNIVERSE



AI



Path-Tracing



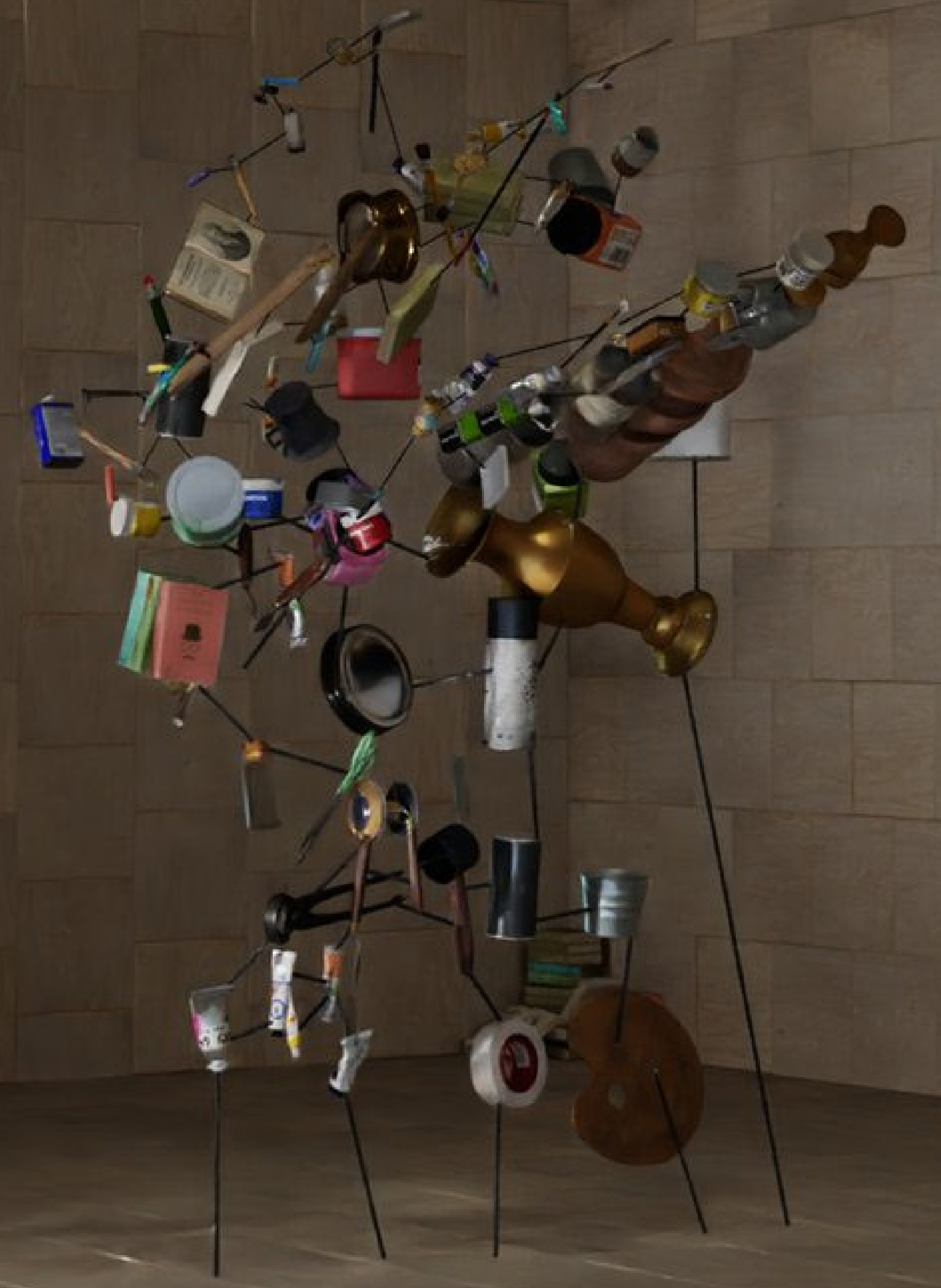
USD



Materials



Physics

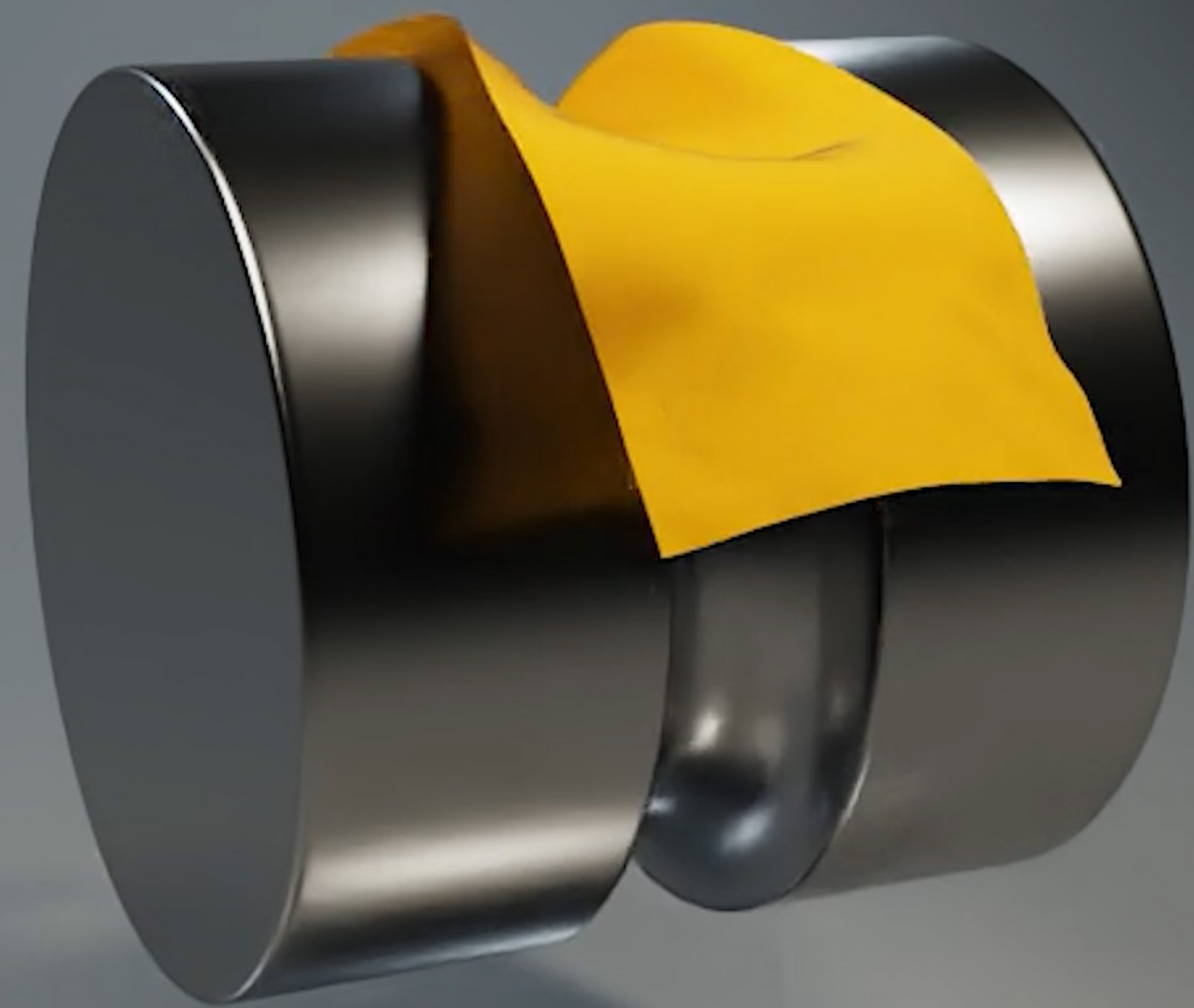


COLLABORATING AND SIMULATING IN OMNIVERSE

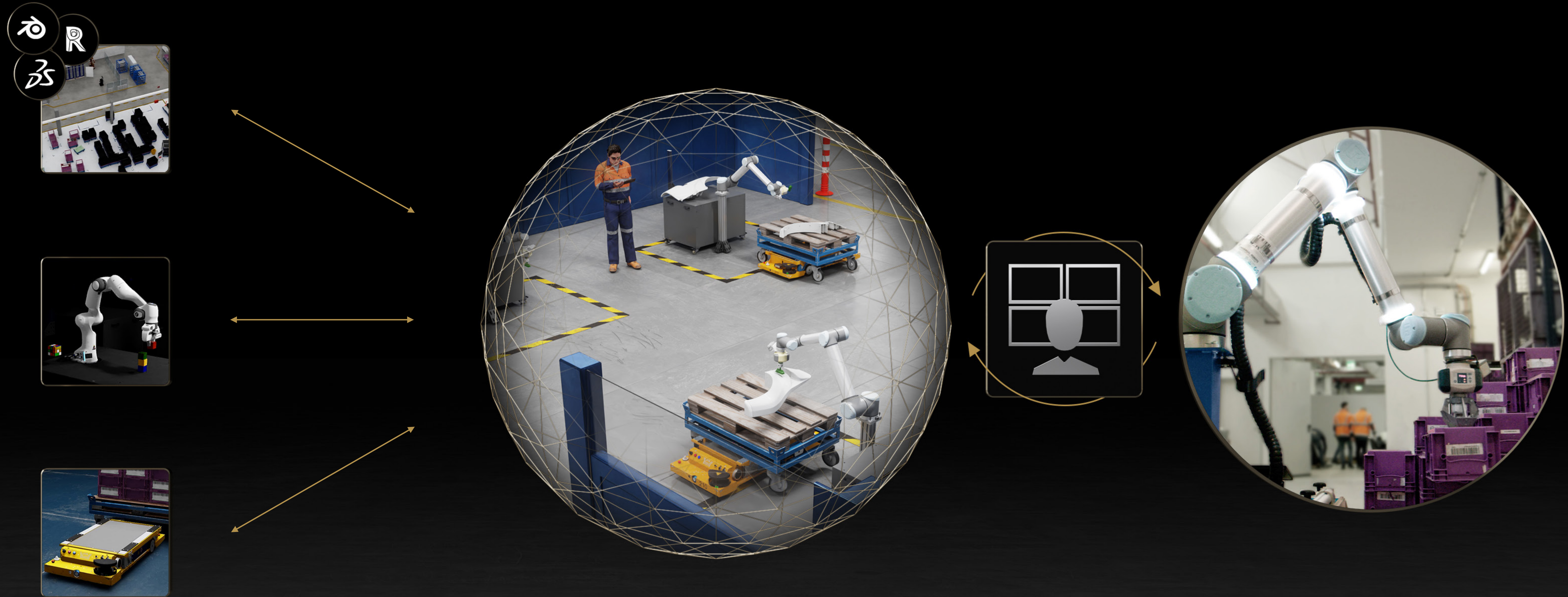


10M DESIGNERS | 20M CREATORS | 1M SCIENTISTS | 2M DEVELOPERS | 40M ENGINEERS

Image Courtesy of Industrial Light & Magic.
© and TM Lucasfilm Ltd. All Rights Reserved.



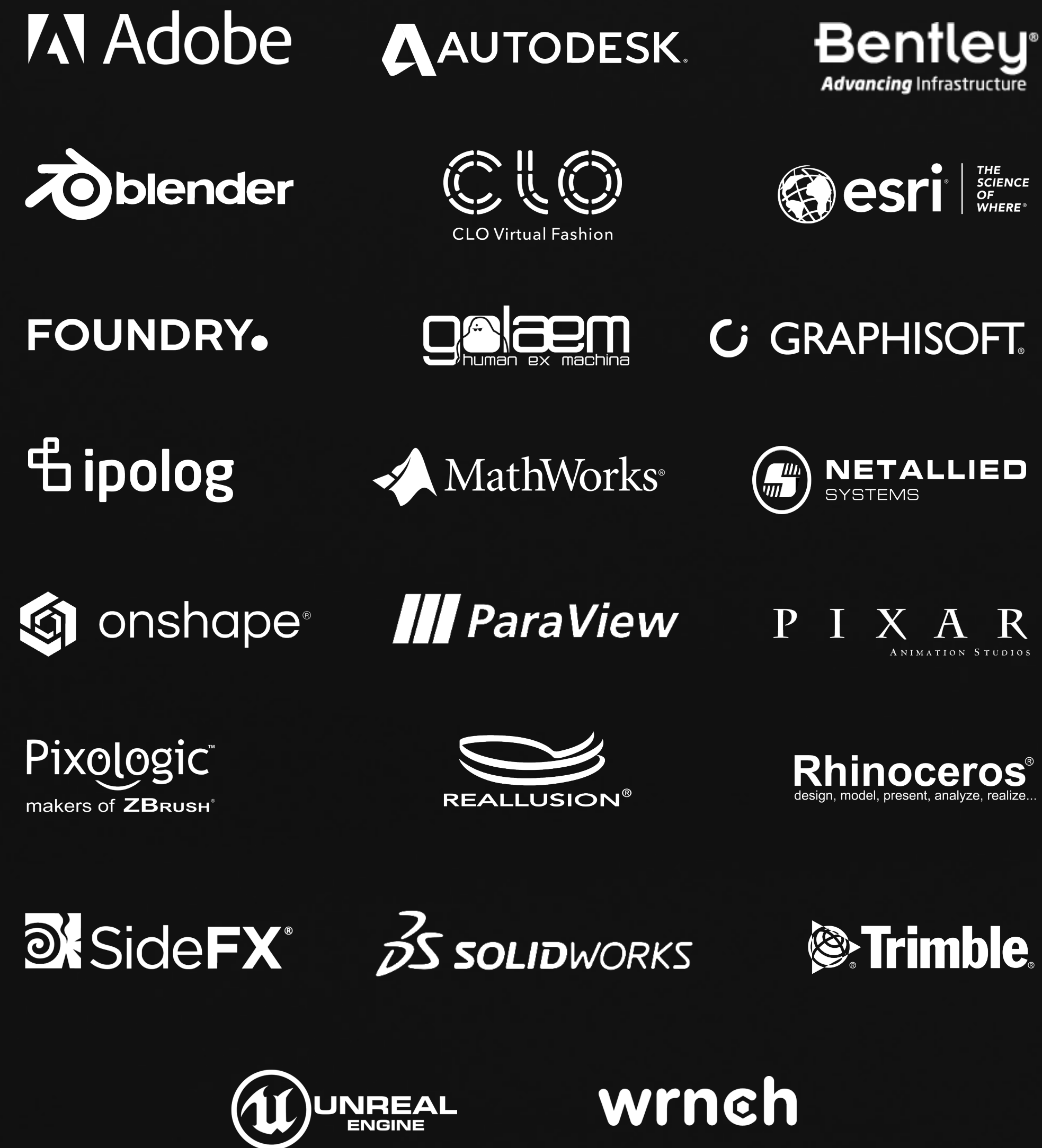
NVIDIA ISAAC DIGITAL TWIN IN OMNIVERSE



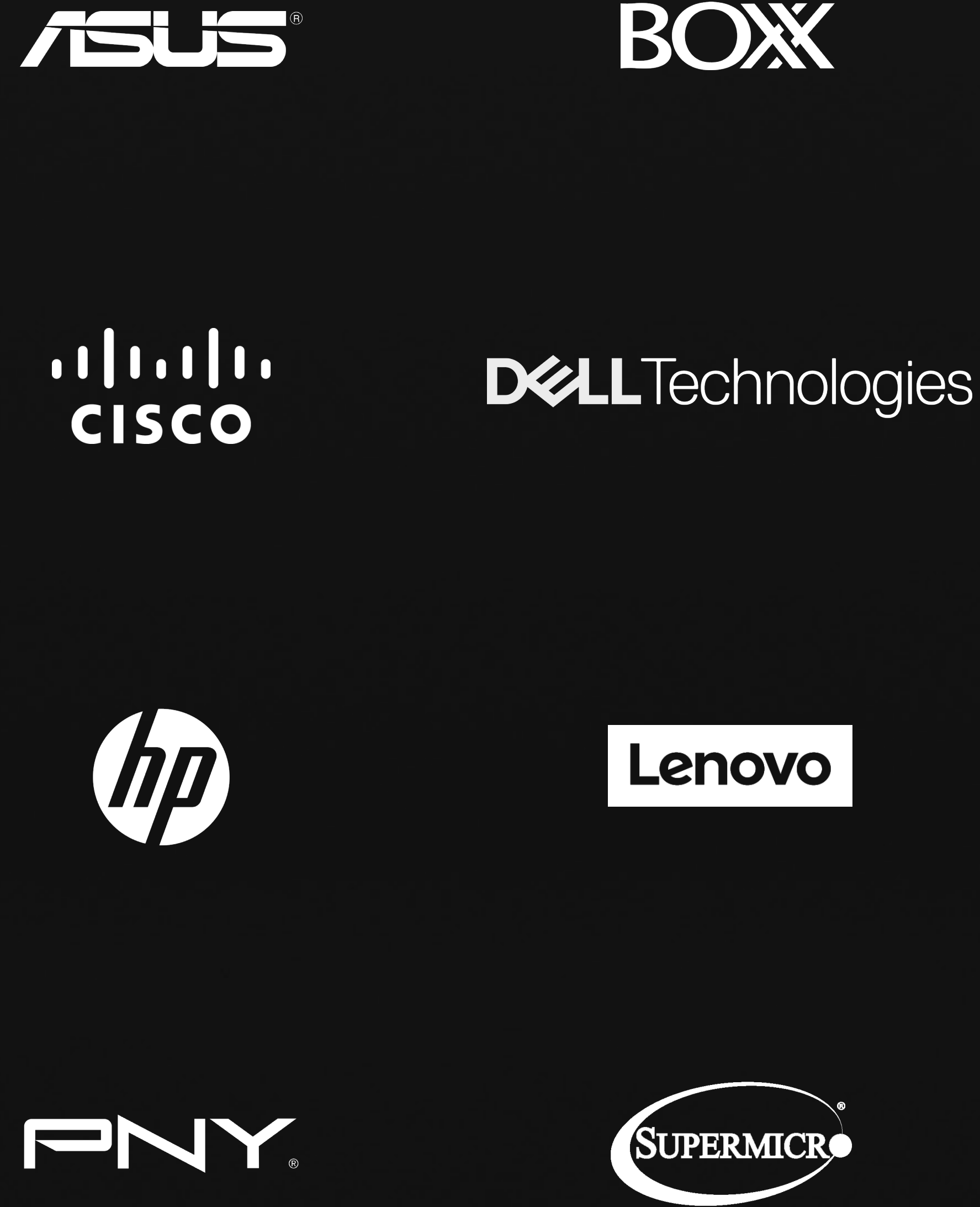


CONNECTING AND CREATING WITH OMNIVERSE

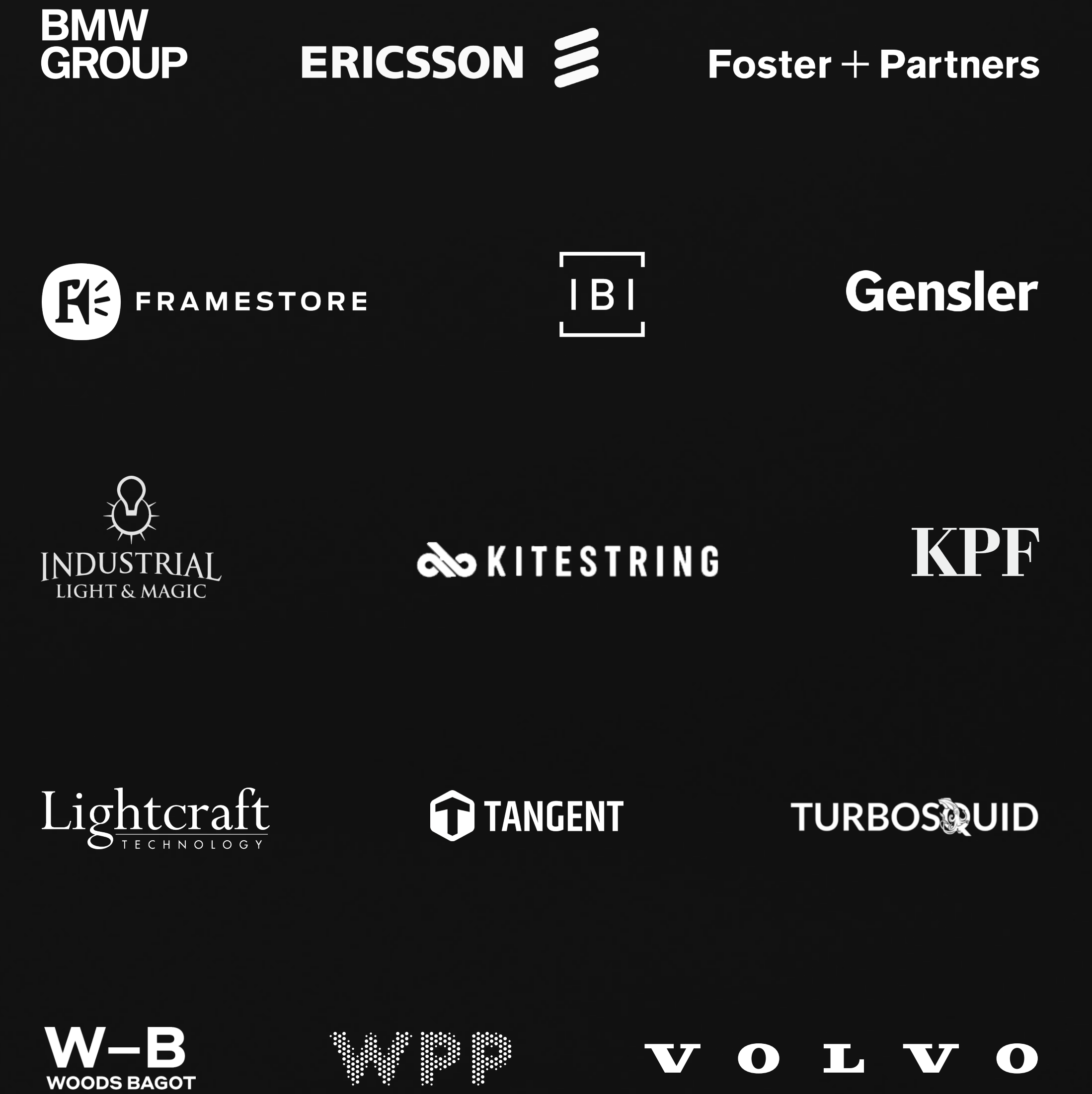
SOFTWARE



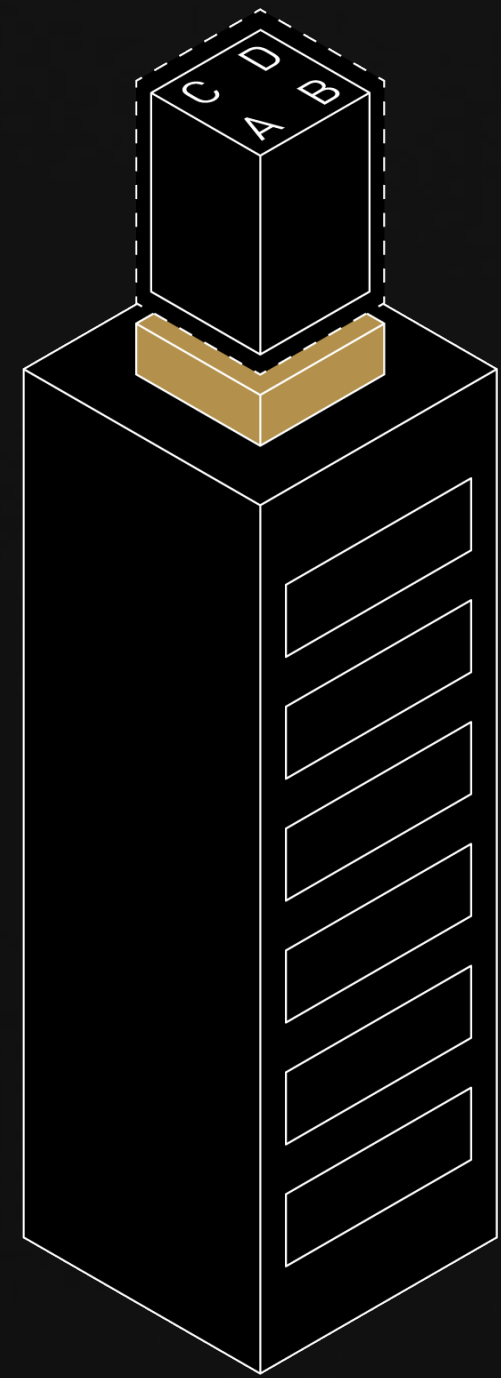
SYSTEMS



PIONEERS

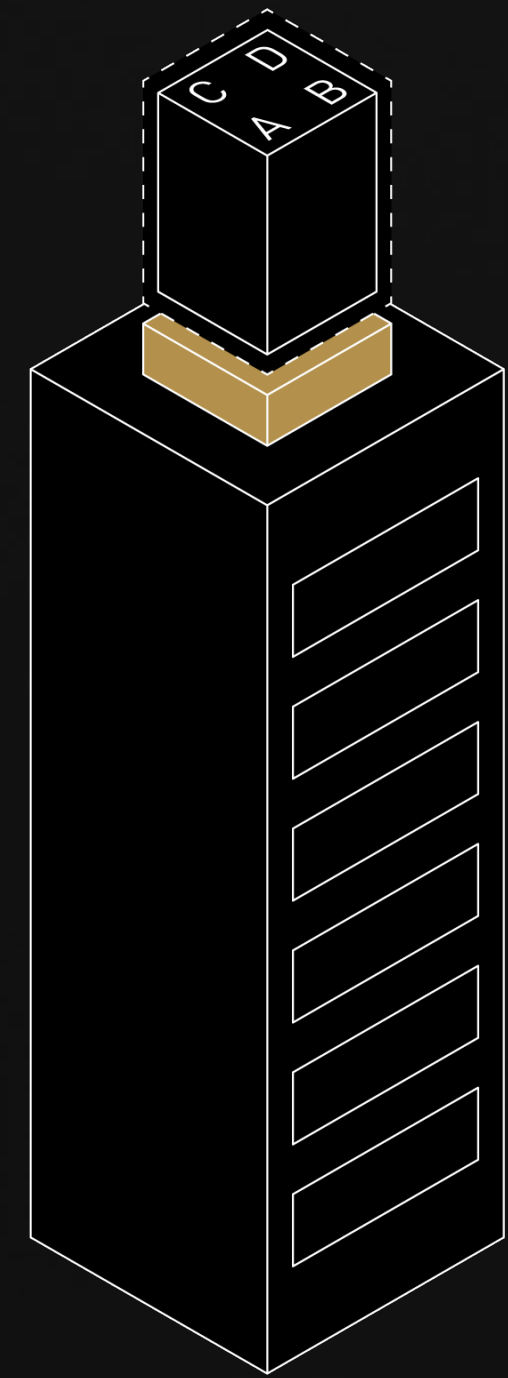


DATA CENTER IS THE NEW UNIT OF COMPUTING

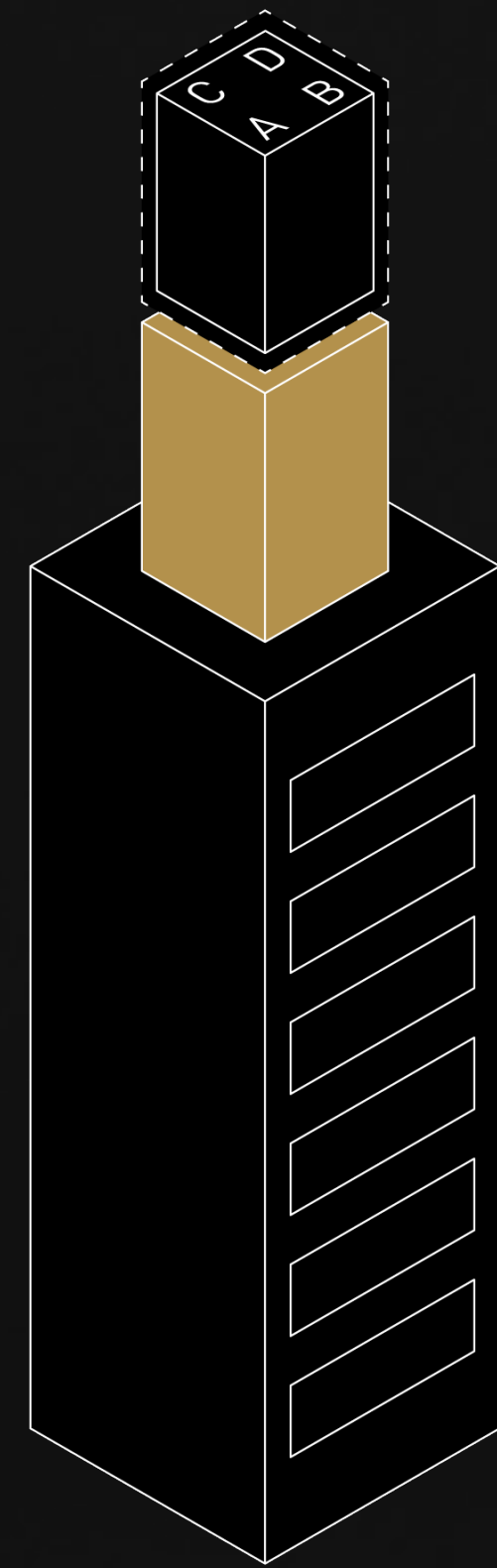


Monolithic

DATA CENTER IS THE NEW UNIT OF COMPUTING

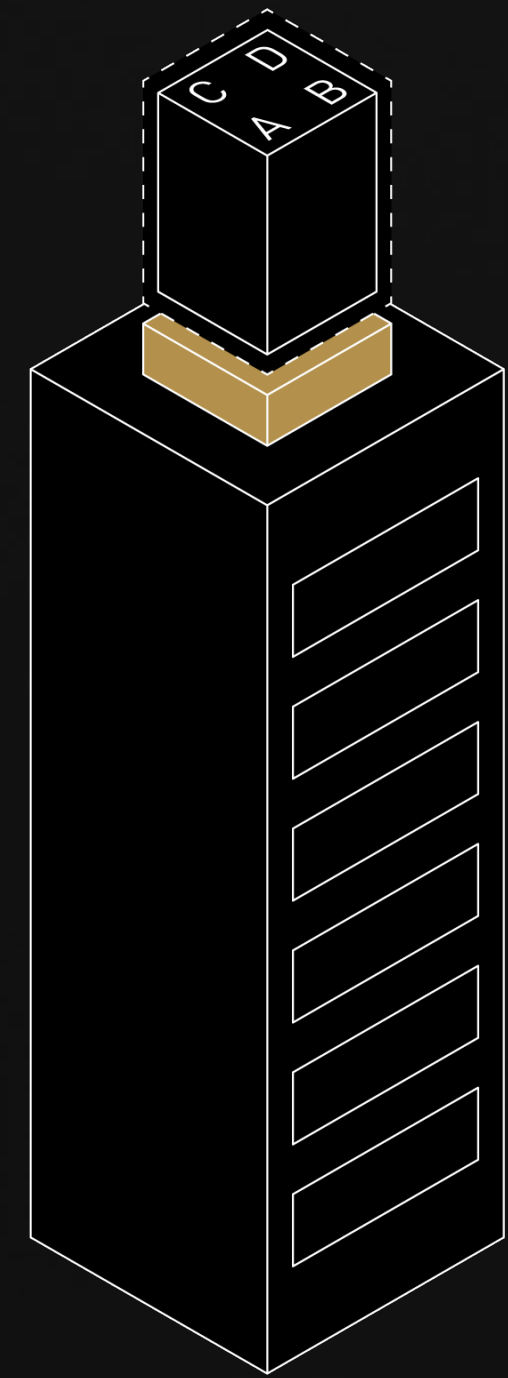


Monolithic

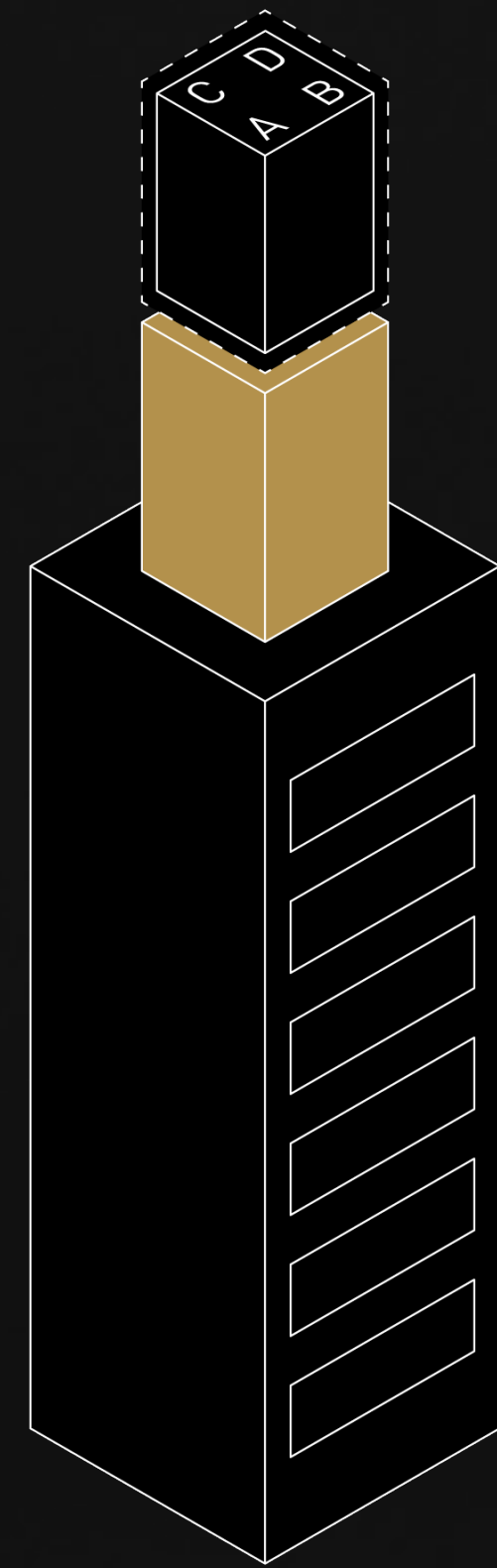


Software-Defined
Data Center

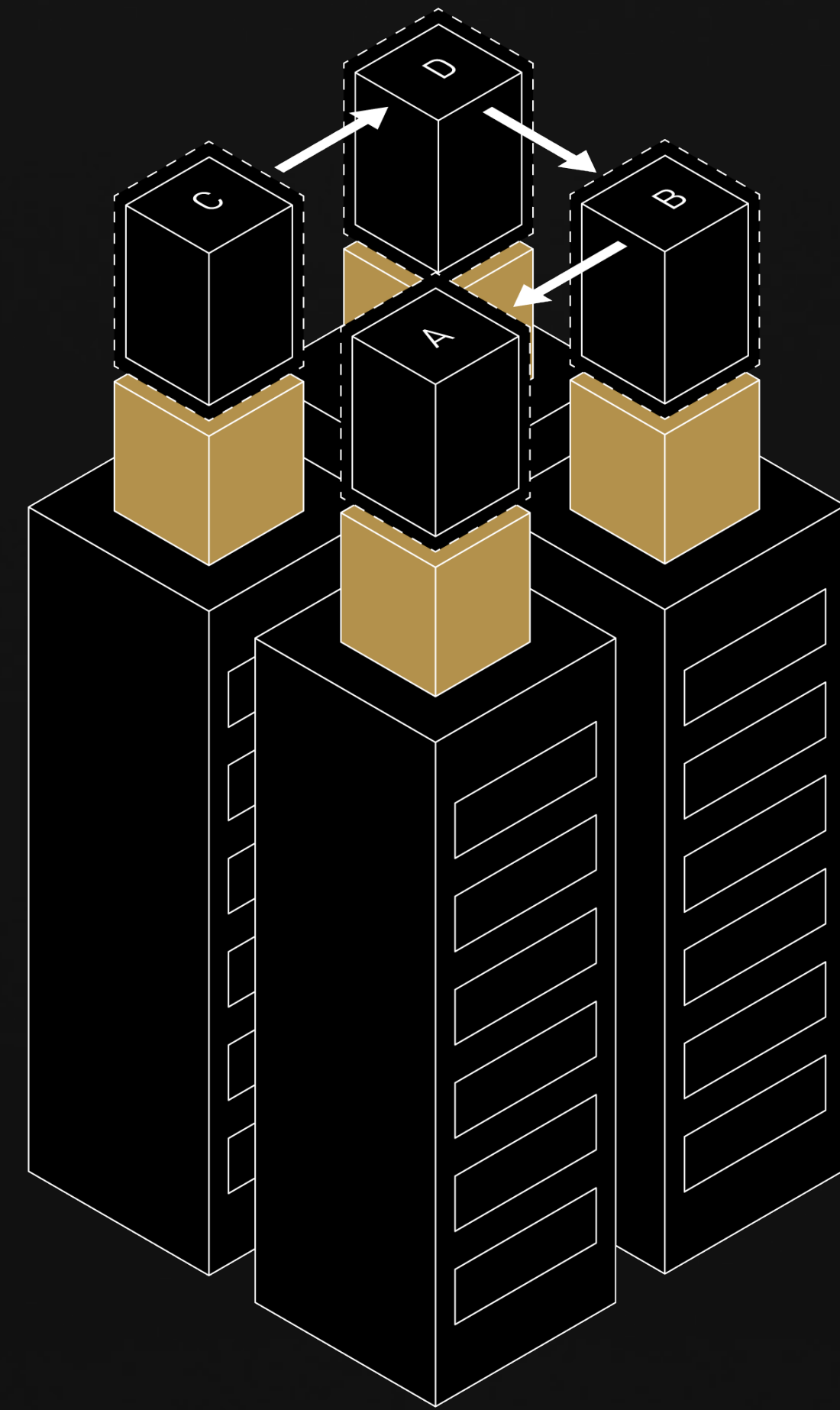
DATA CENTER IS THE NEW UNIT OF COMPUTING



Monolithic

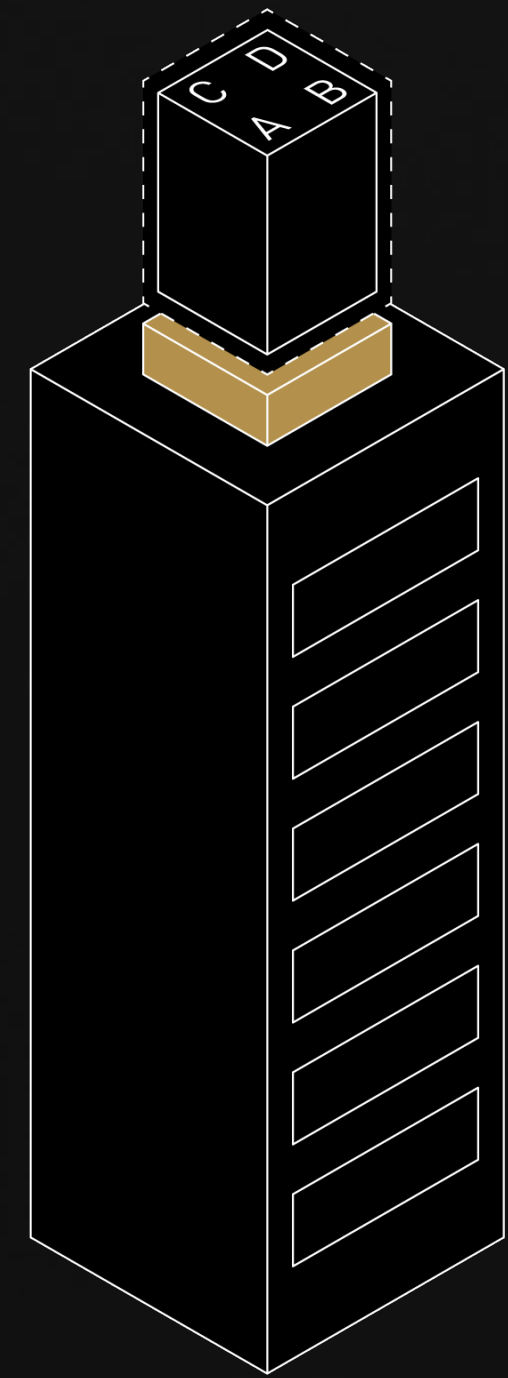


Software-Defined
Data Center

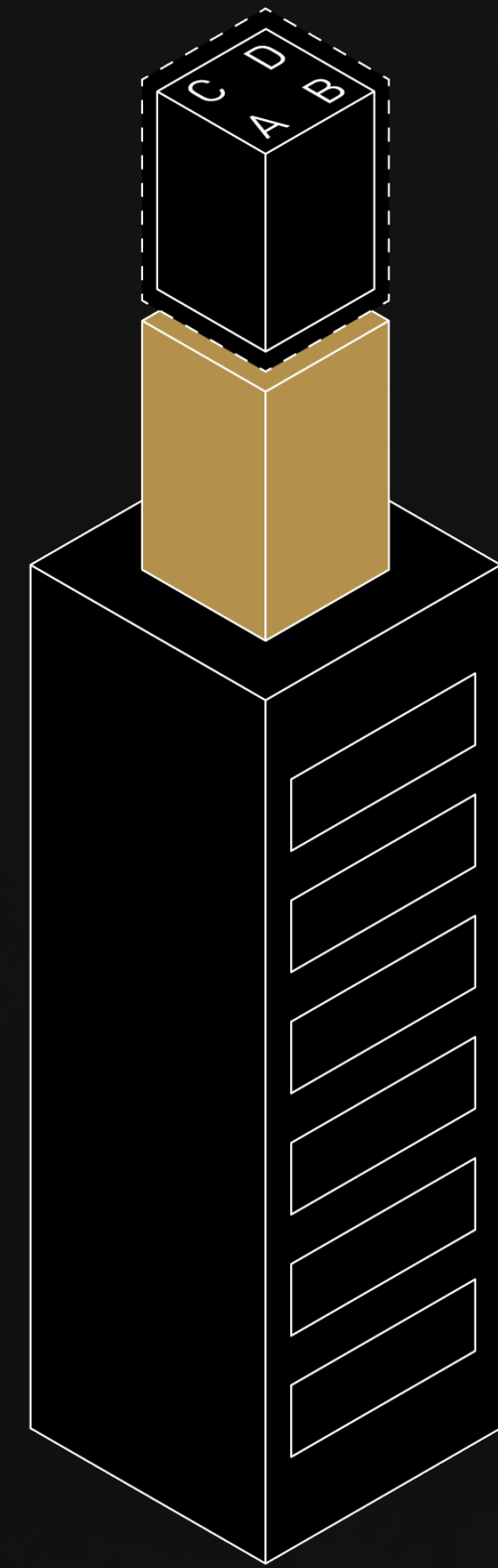


Disaggregated,
Micro-Services, Scaled Out

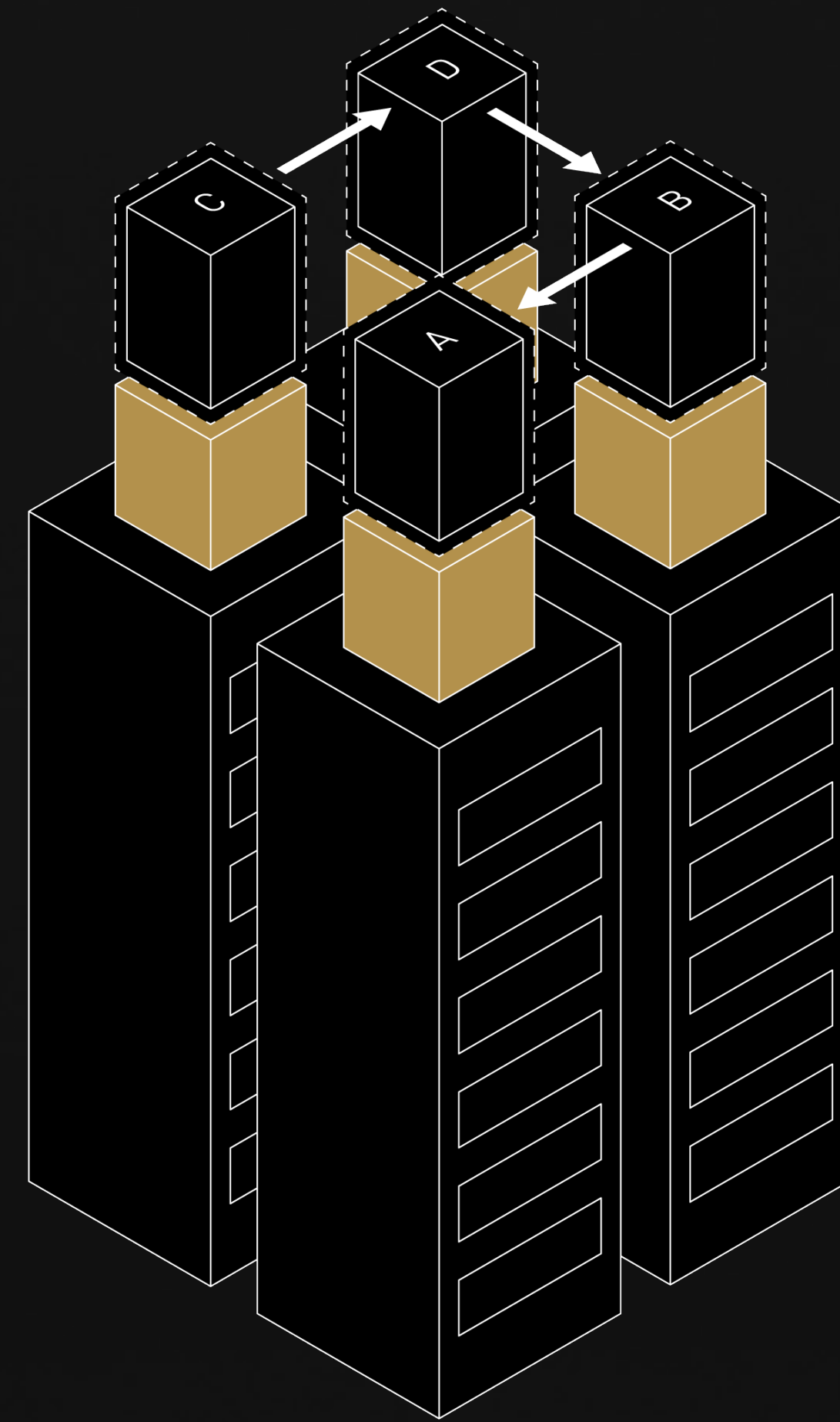
DATA CENTER IS THE NEW UNIT OF COMPUTING



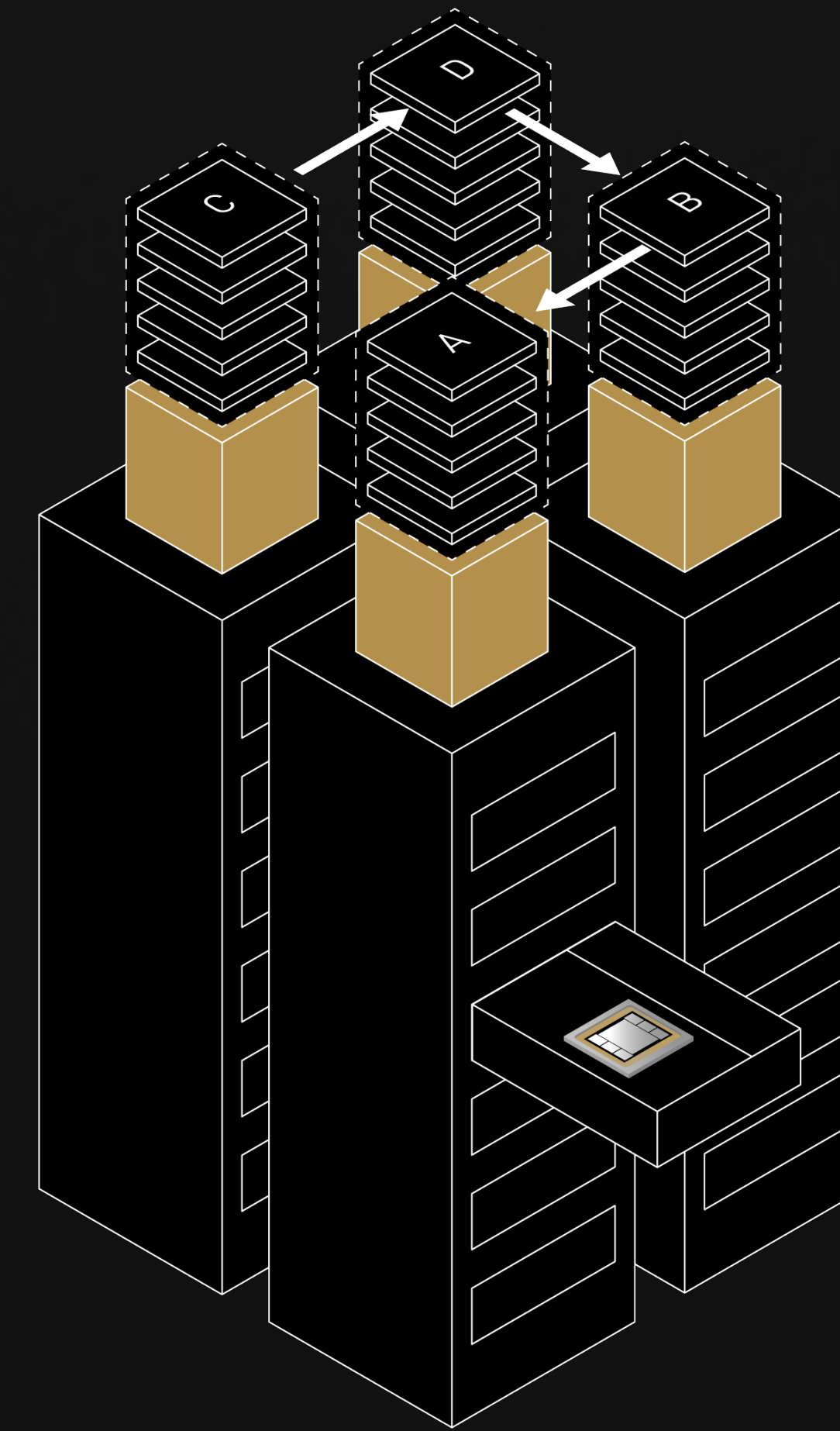
Monolithic



Software-Defined
Data Center

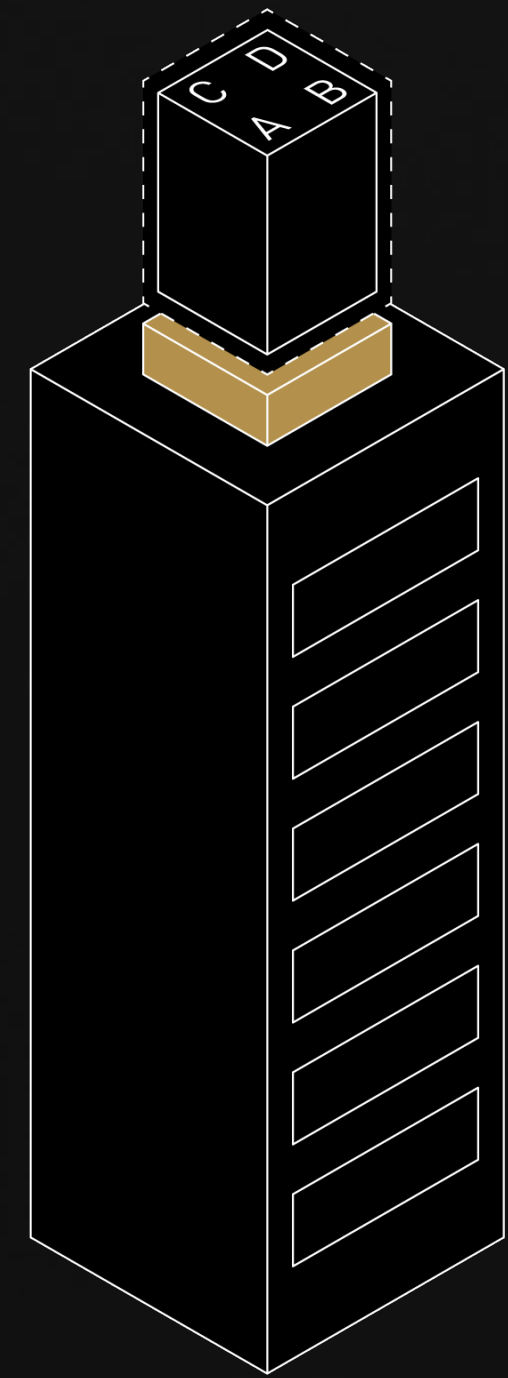


Disaggregated,
Micro-Services, Scaled Out

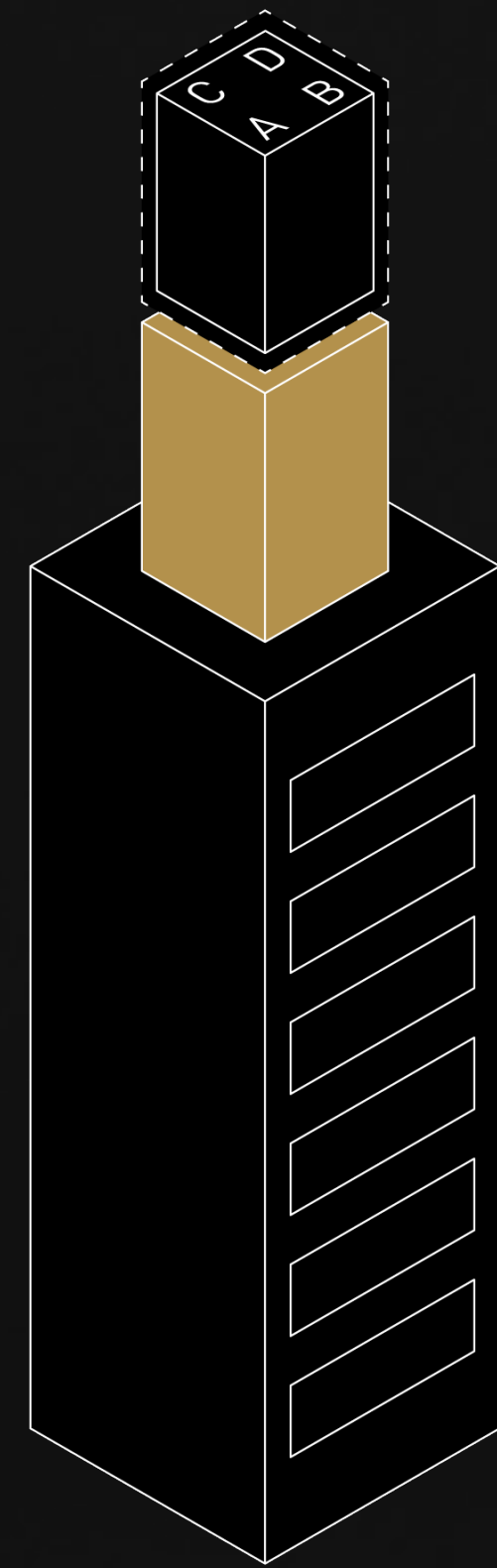


GPU-Accelerated
Computing

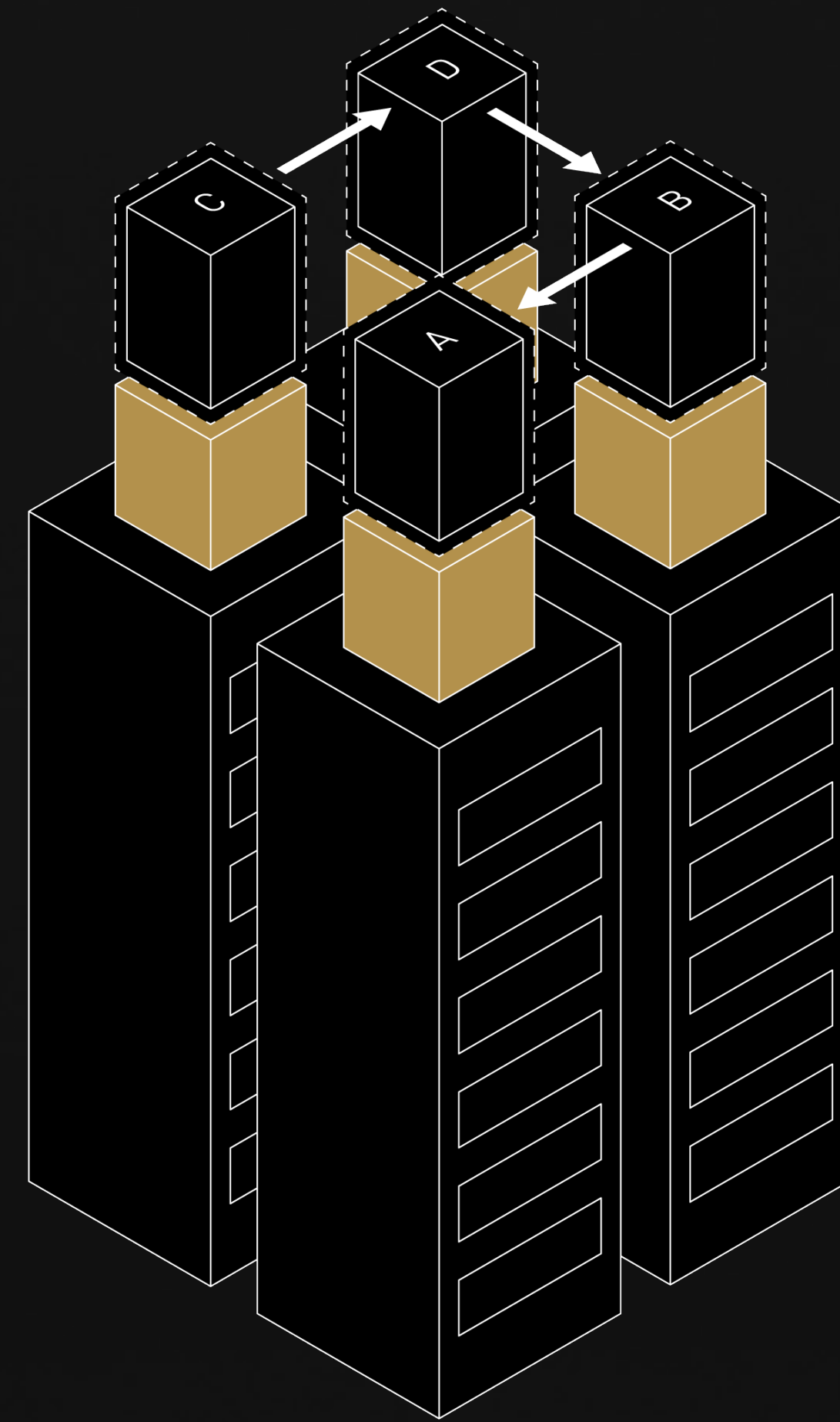
DATA CENTER IS THE NEW UNIT OF COMPUTING



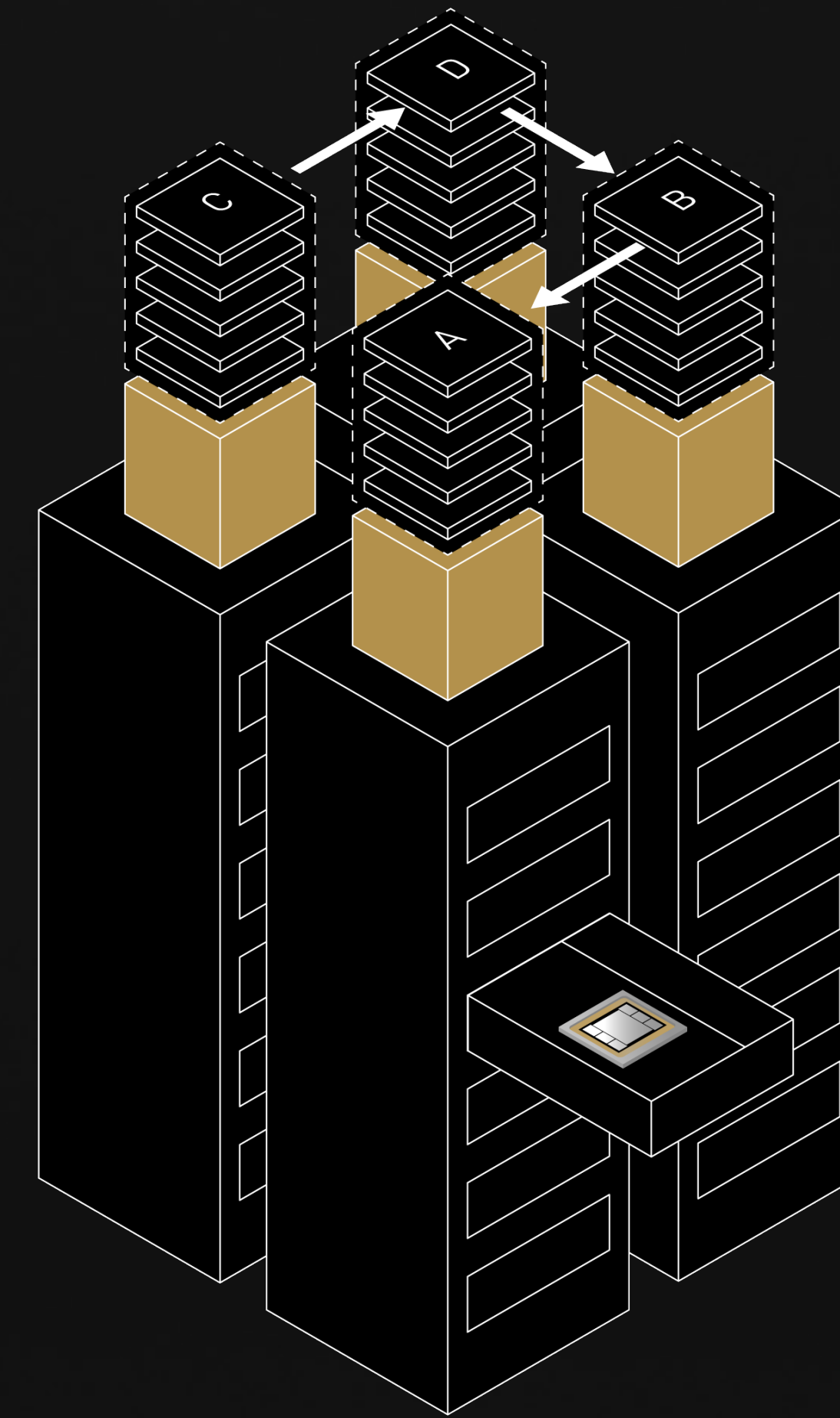
Monolithic



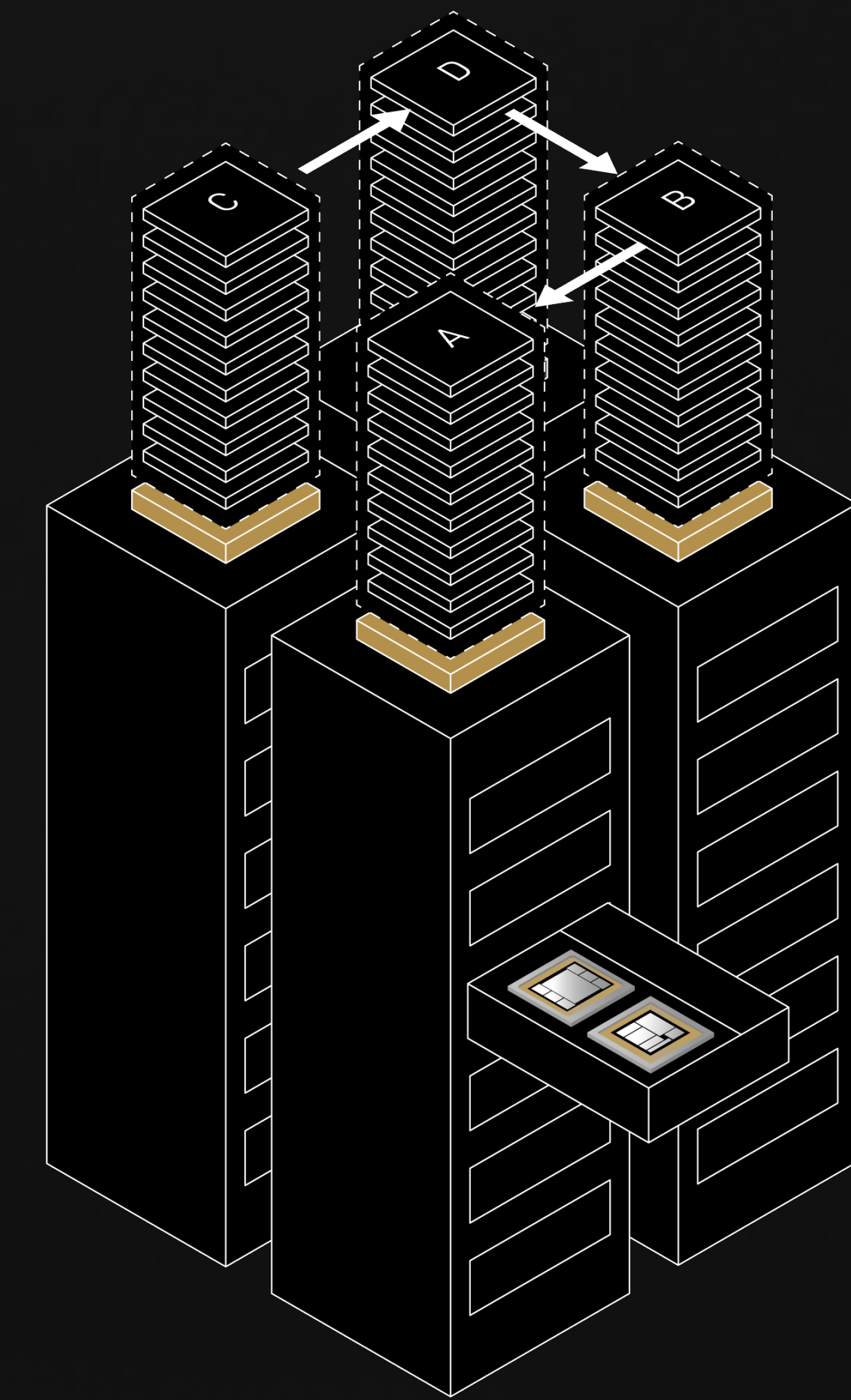
Software-Defined
Data Center



Disaggregated,
Micro-Services, Scaled Out



GPU-Accelerated
Computing

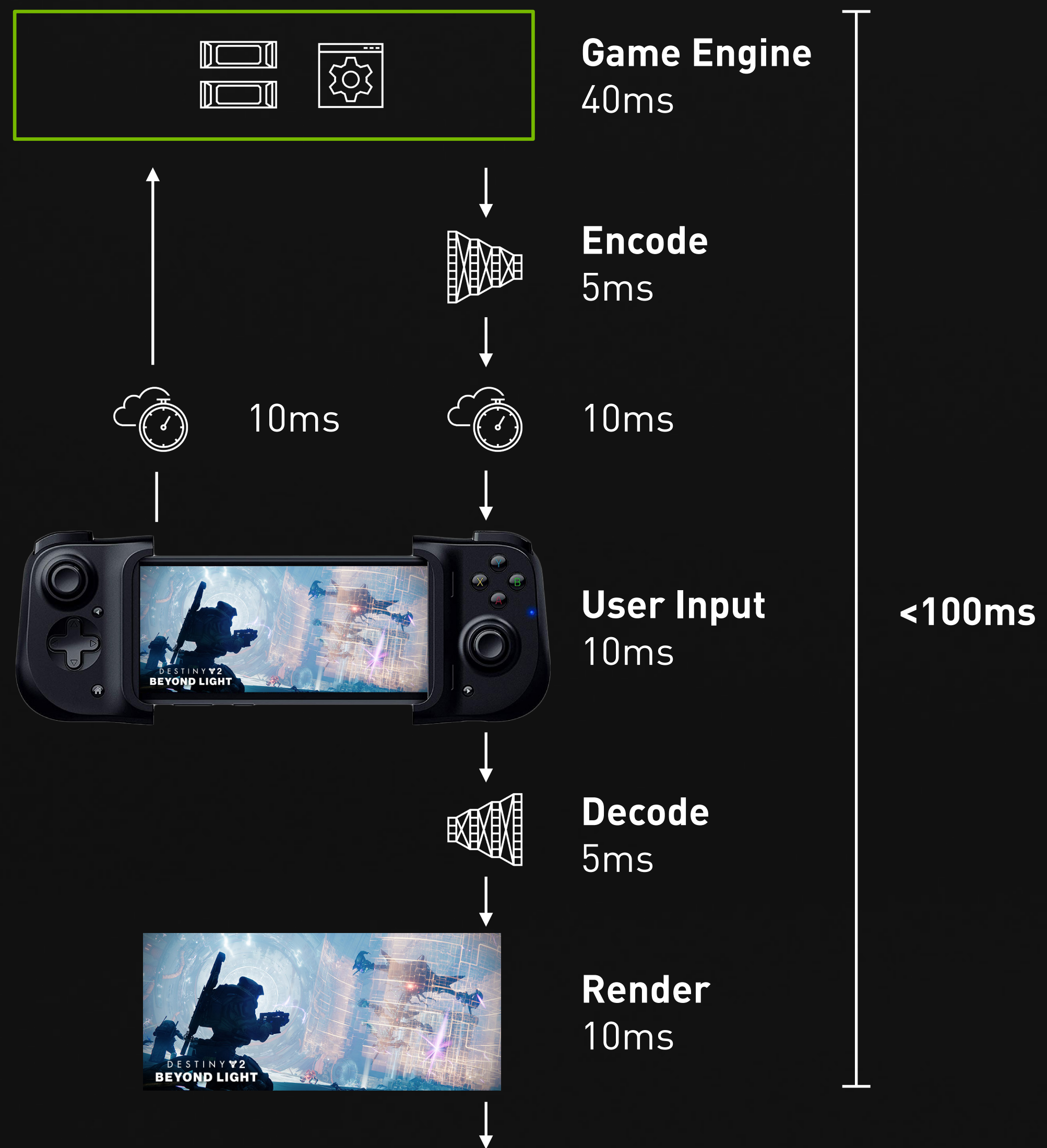


DPU-Accelerated
Data Center Infrastructure

BLUEFIELD SECURES AND ACCELERATES GEFORCE NOW CLOUD GAMING

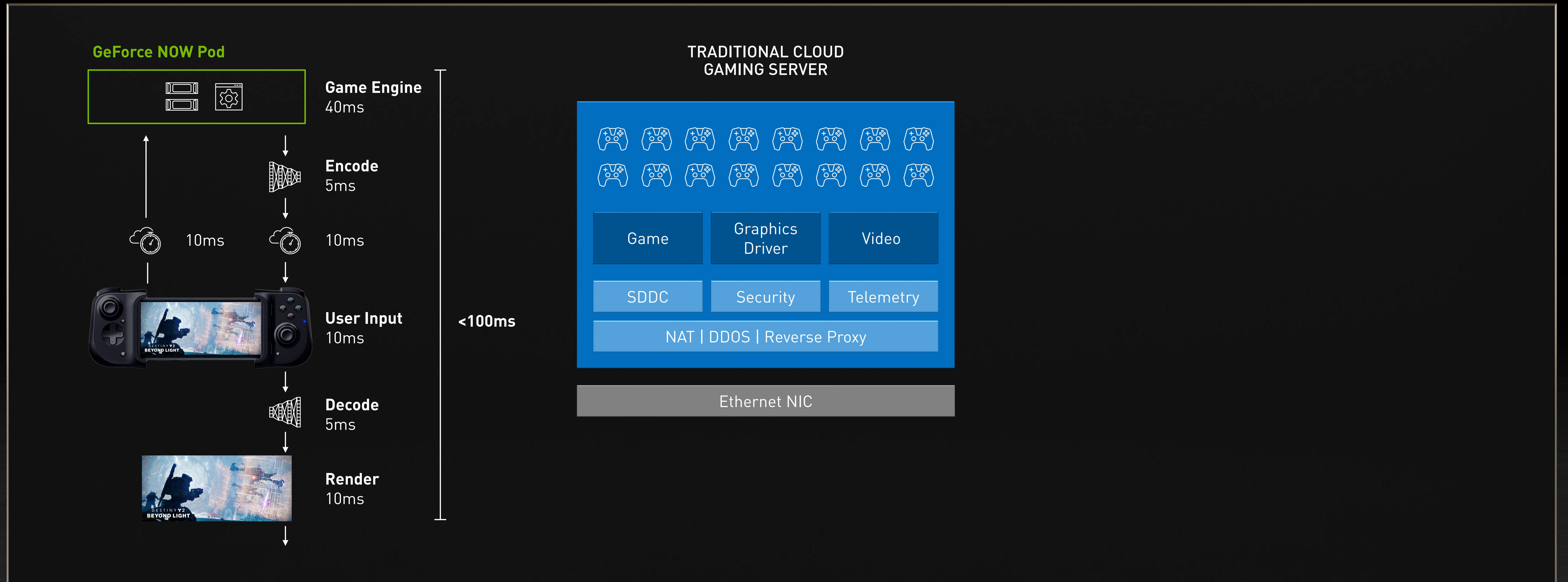
Isolate and Secure Infrastructure | High Quality-of-Service | More Concurrent Users

GeForce NOW Pod



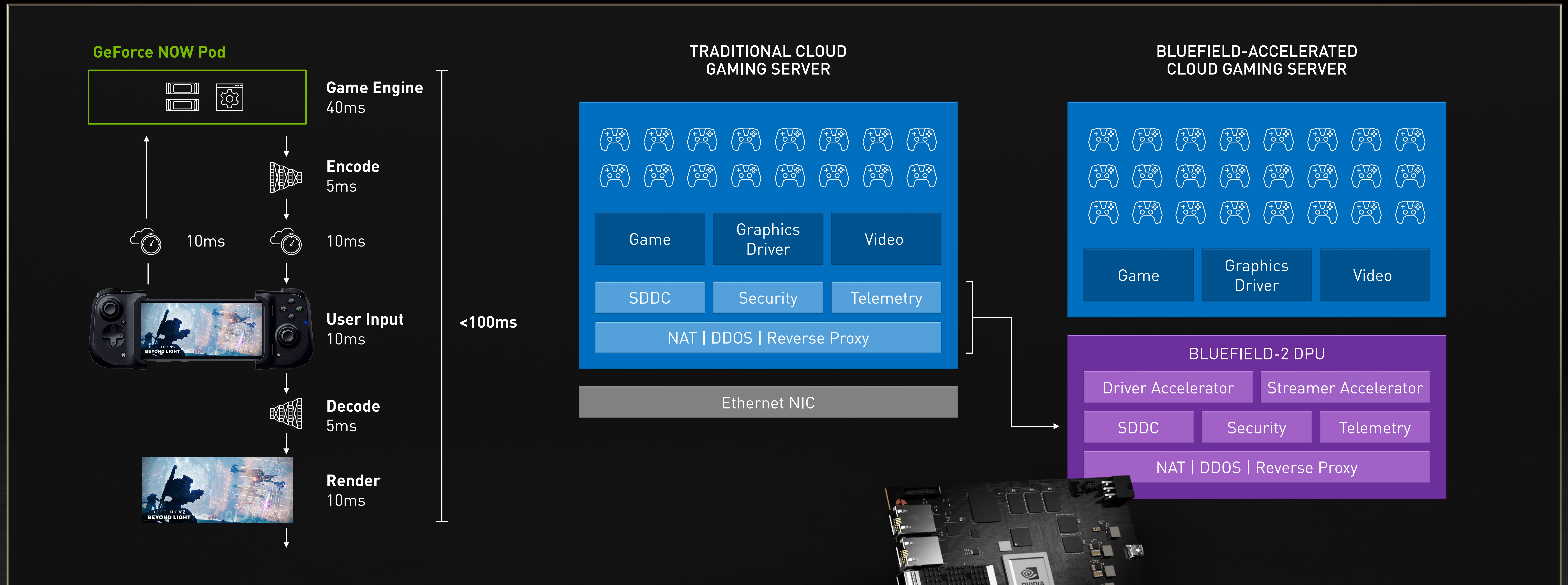
BLUEFIELD SECURES AND ACCELERATES GEFORCE NOW CLOUD GAMING

Isolate and Secure Infrastructure | High Quality-of-Service | More Concurrent Users



BLUEFIELD SECURES AND ACCELERATES GEFORCE NOW CLOUD GAMING

Isolate and Secure Infrastructure | High Quality-of-Service | More Concurrent Users



ANNOUNCING DOCA 1.0 AND BLUEFIELD PARTNER ECOSYSTEM

Data Center on a Chip Architecture

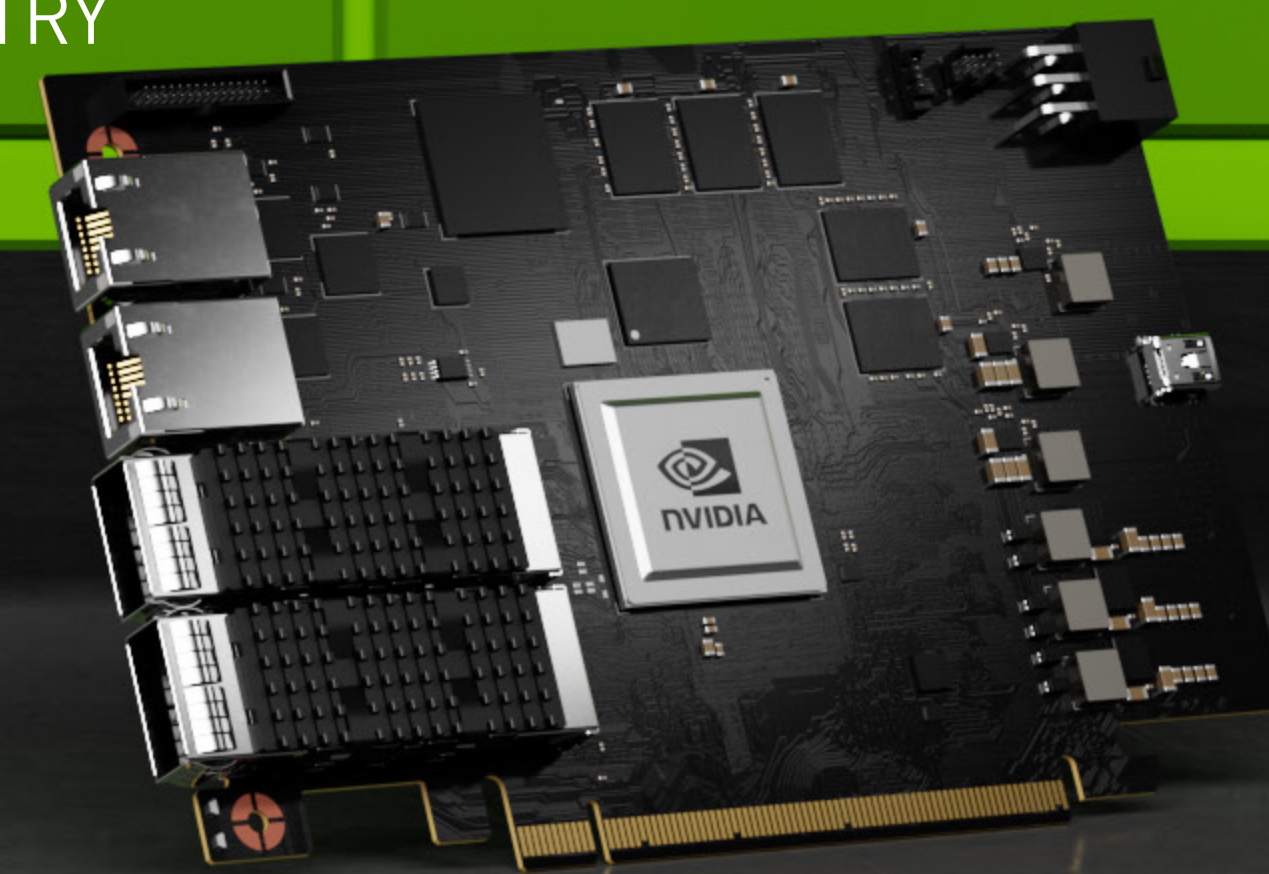
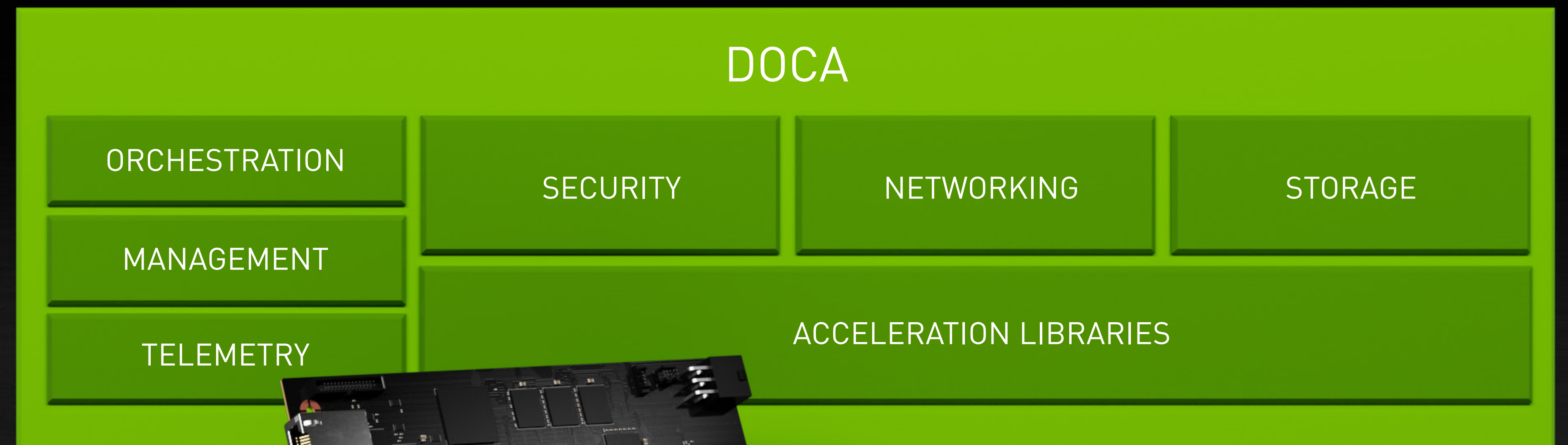
Software Development Framework for BlueField DPUs

Offload, Accelerate, and Isolate Infrastructure Processing

Support for Hyperscale, Enterprise, Supercomputing and Hyperconverged Infrastructure

Transparent Offload and Acceleration of VMware ESX

Software Compatibility for Generations of BlueField DPUs



ANNOUNCING NVIDIA BLUEFIELD-3

400 Gbps Data Center Infra Processor

Offloads and Accelerates Data Center Infrastructure

Isolates Application from Control and Management Plane

Powerful CPU – 16x Arm A78 Cores

Process Networking, Storage, and Security at 400 Gbps

22 Billion Transistors

DATA PATH ACCELERATOR

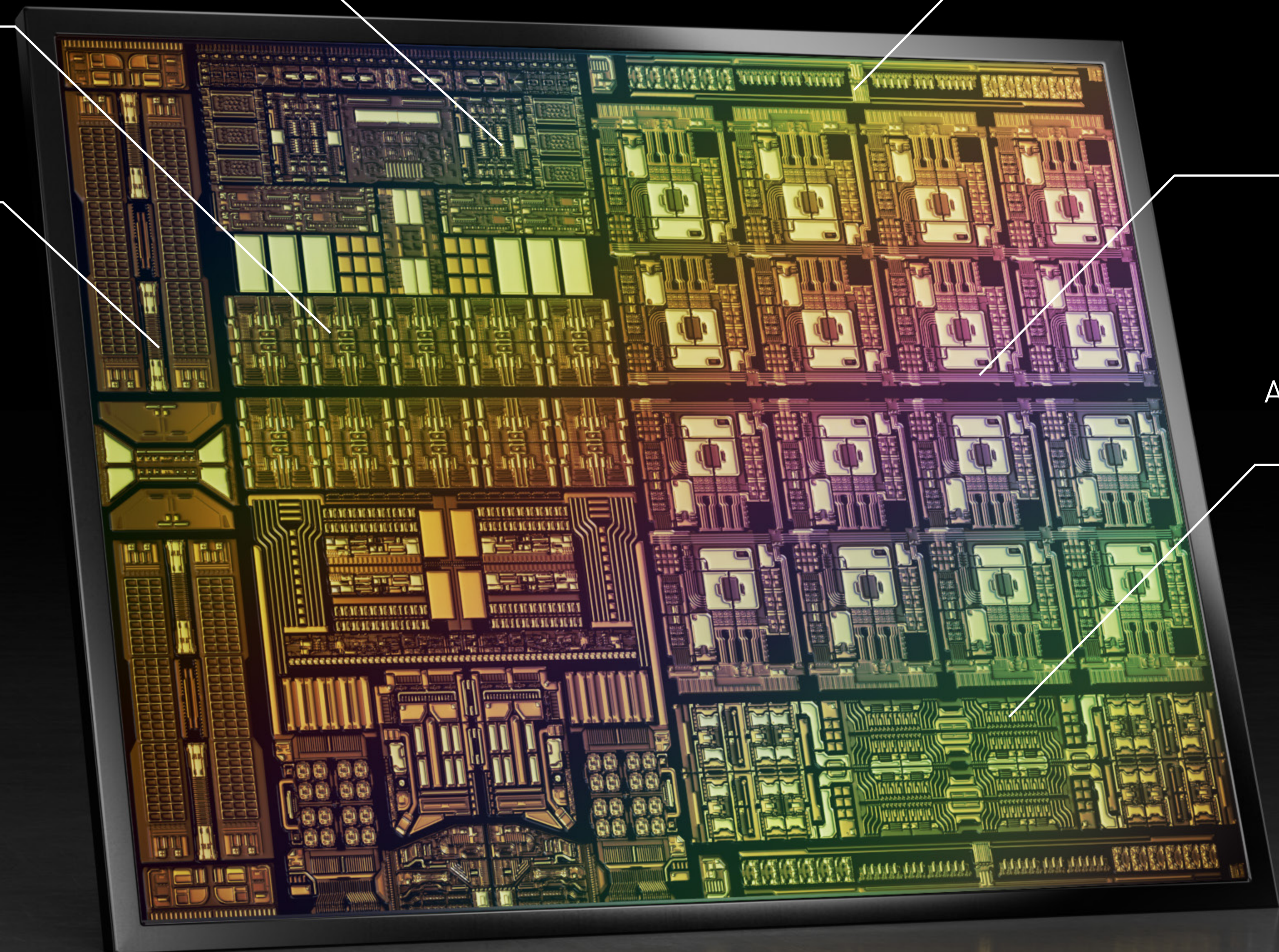
DDR5 MEMORY INTERFACE

CONNECTX-7

ARM CORES

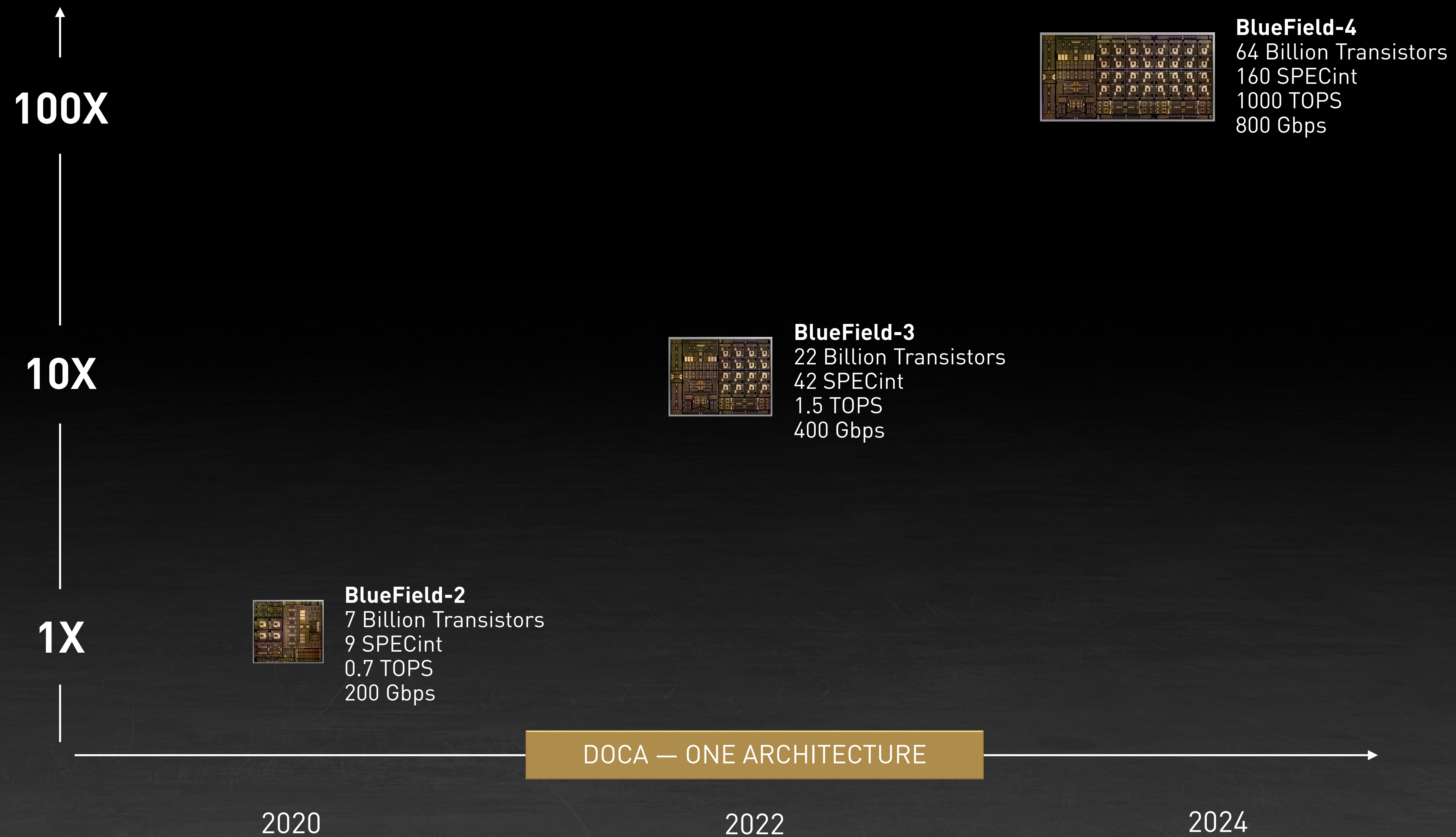
PCIe GEN 5.0

ACCELERATION
ENGINES



EXPONENTIAL GROWTH IN DATA CENTER INFRASTRUCTURE PROCESSING

Cloud-Native | Disaggregation | Micro-Services | AI | Zero-Trust Security



NVIDIA DGX

The Computer of AI Researchers

9 of 10

Top Universities

8 of 10

Top Global Telcos

7 of 10

Top U.S. Hospitals

7 of 10

Top Consumer Companies

6 of 10

Top U.S. Banks

7 of 10

Top Car Makers

10 of 10

Top Aerospace Companies



DGX SUPERPOD
AI Data Center As-a-Product



DGX A100
AI Data Center Building Block

DGX STATION A100
AI Data Center-in-a-Box

ANNOUNCING NVIDIA DGX STATION 320G

Workgroup AI Supercomputer-in-a-Box

Plug-into-the-Wall Instant AI Infrastructure

2.5 petaFLOPS

320 GB at 8 TB/sec

7.68 TB NVMe

28 MIGs

1500W and < 37db

\$149,000 or \$9,000/Month Subscription



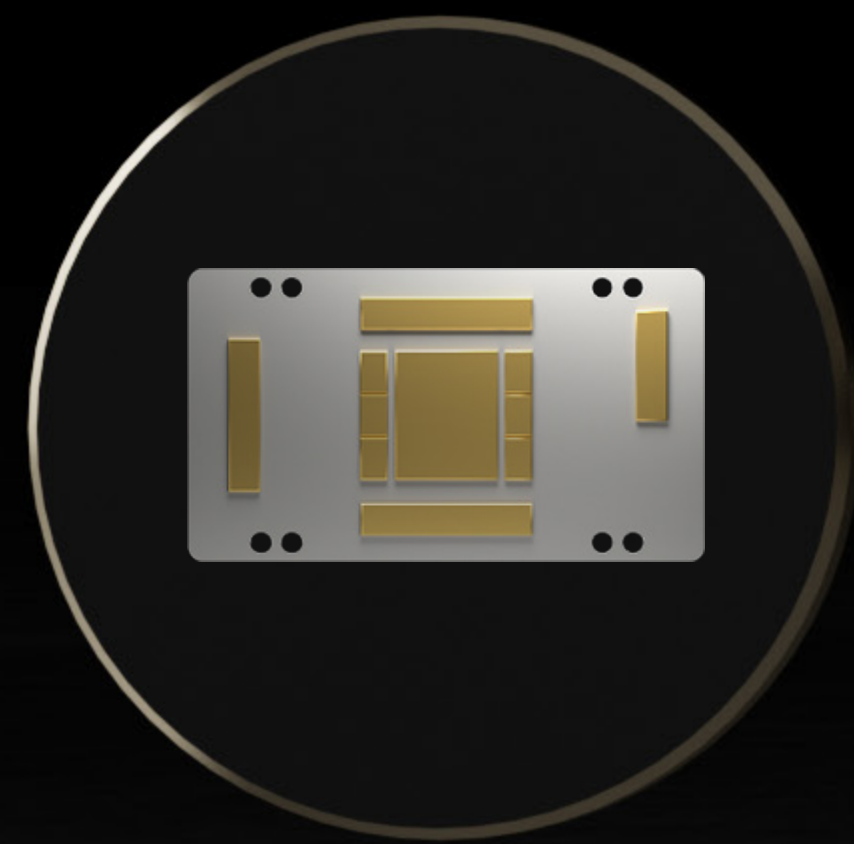


DGX STATION

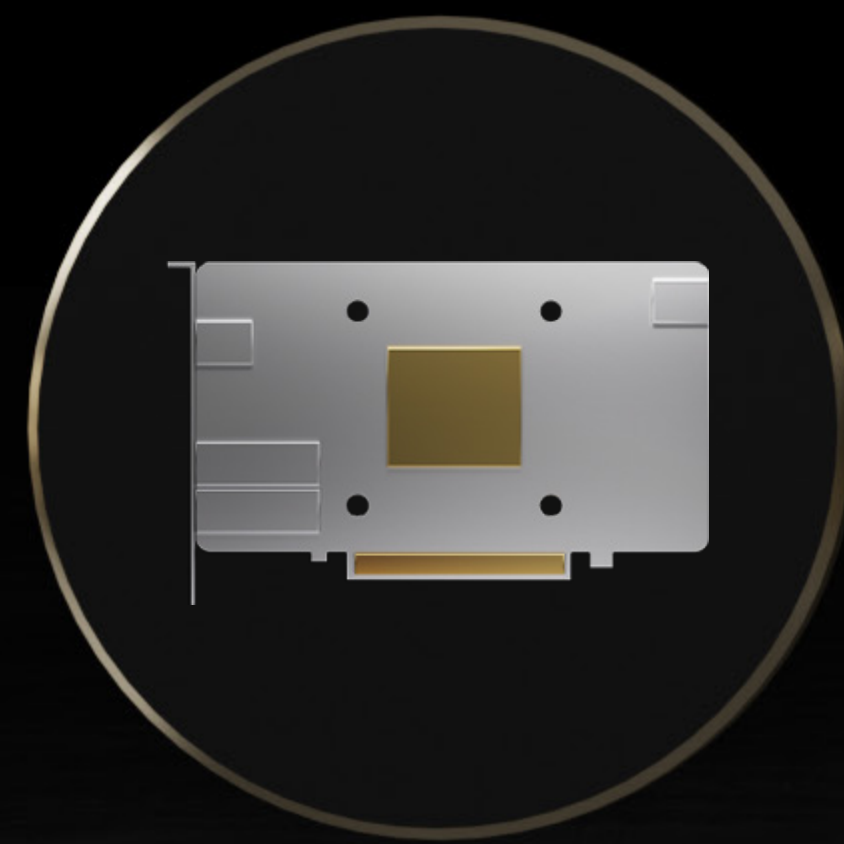
 NVIDIA

ANNOUNCING THE NEW DGX SUPERPOD

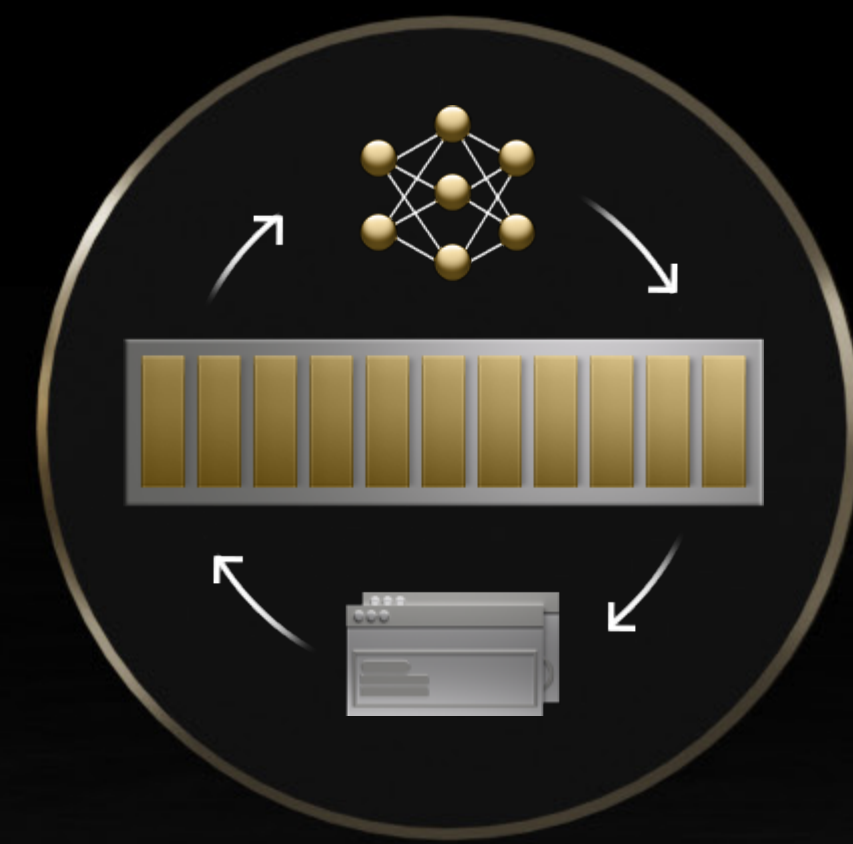
World's First Cloud-Native Supercomputer | Secured by NVIDIA BlueField | Multi-Tenant Bare-Metal Performance



A100 80GB



BLUEFIELD-2

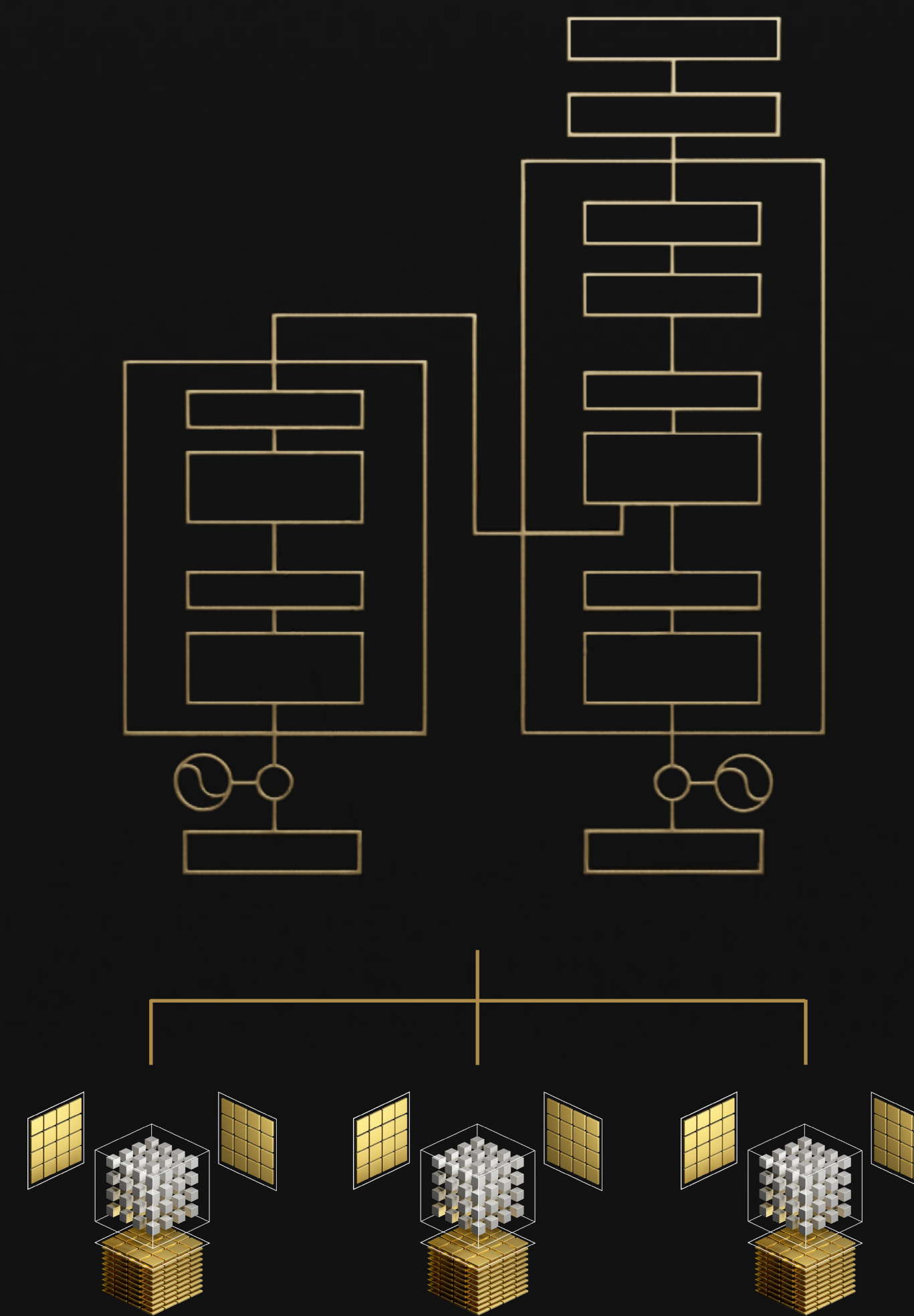
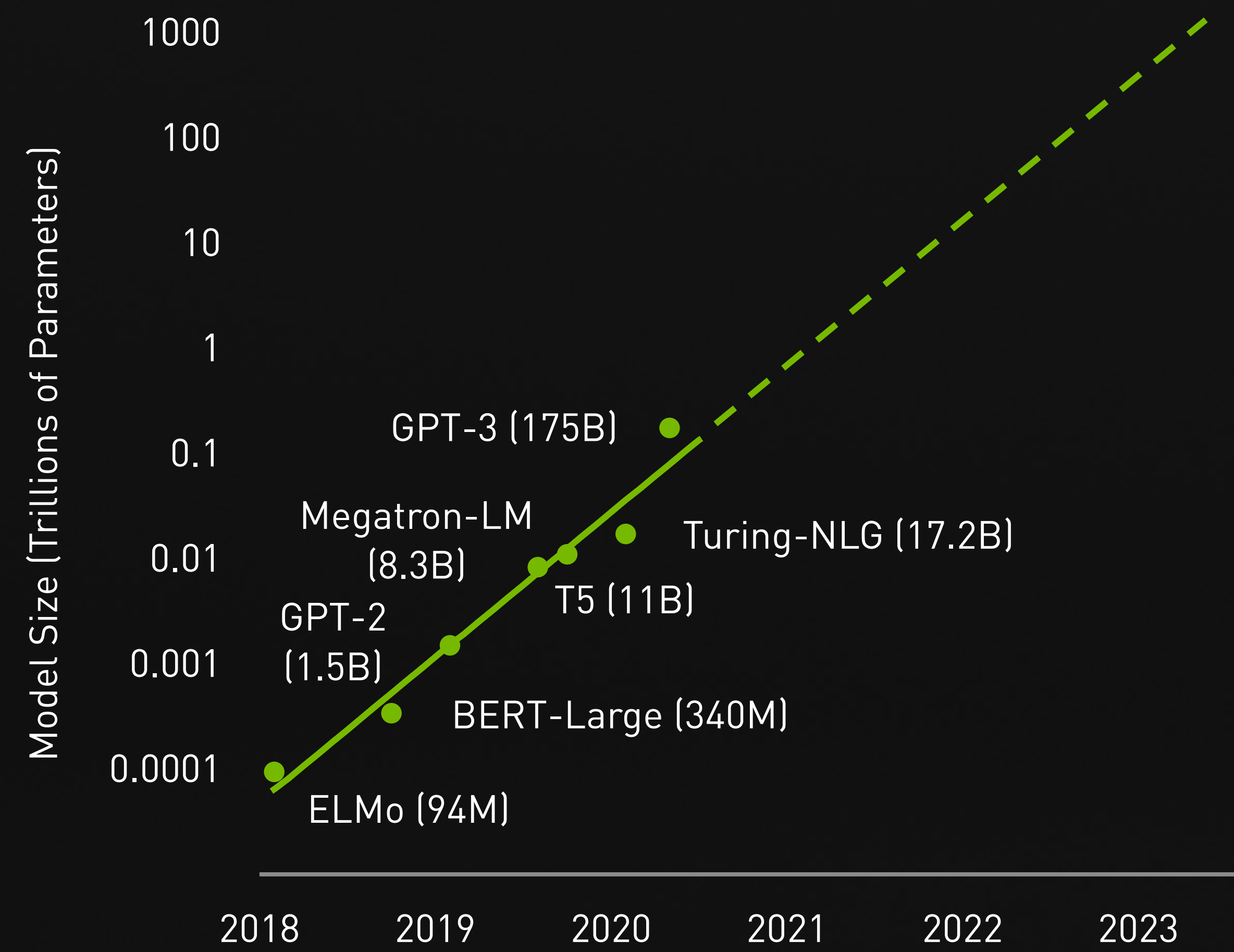


BASE COMMAND



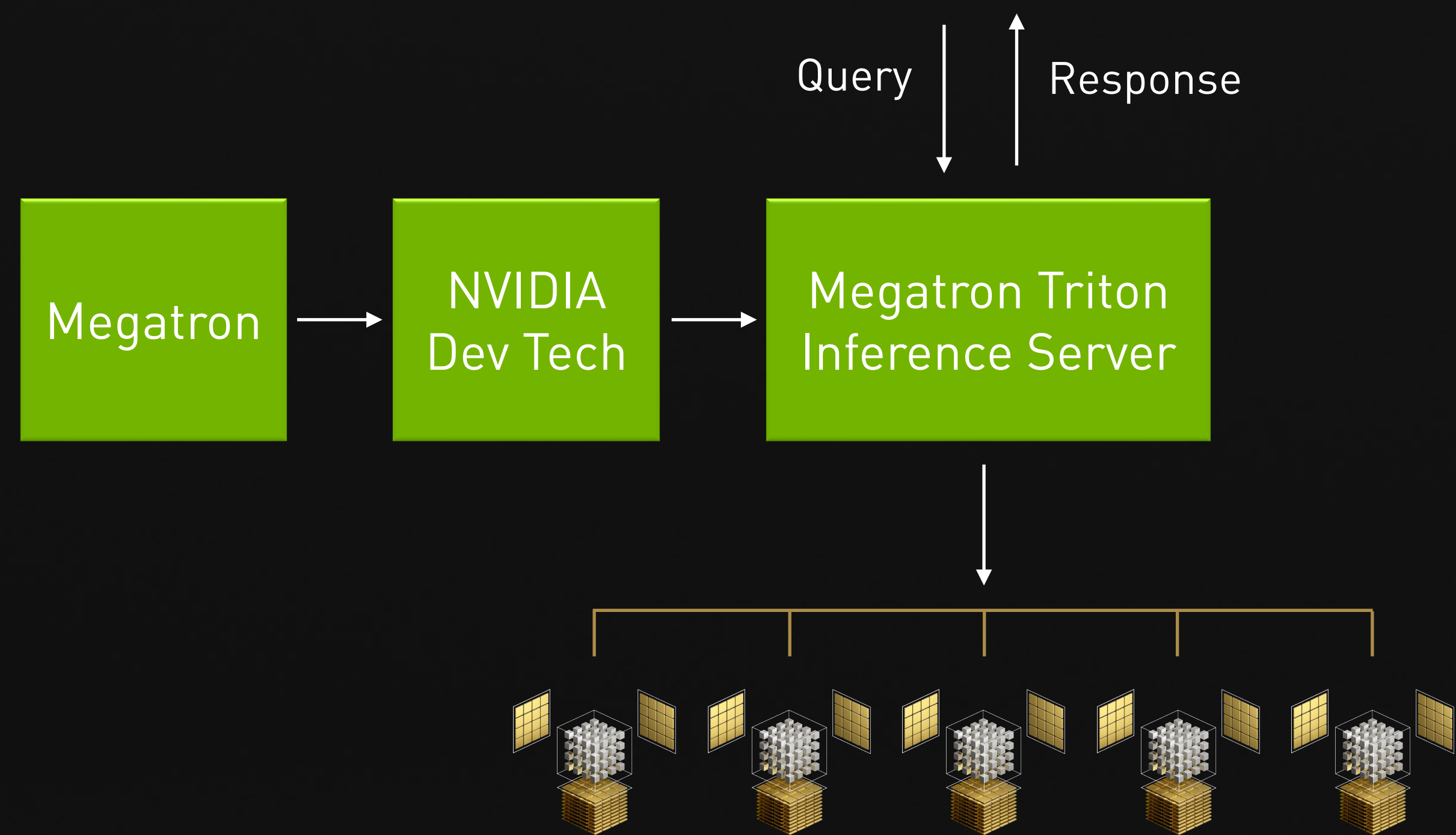
NVIDIA MEGATRON TRAINS TRANSFORMERS

100 TRILLION PARAMETERS MODELS BY 2023



NVIDIA MEGATRON TRITON

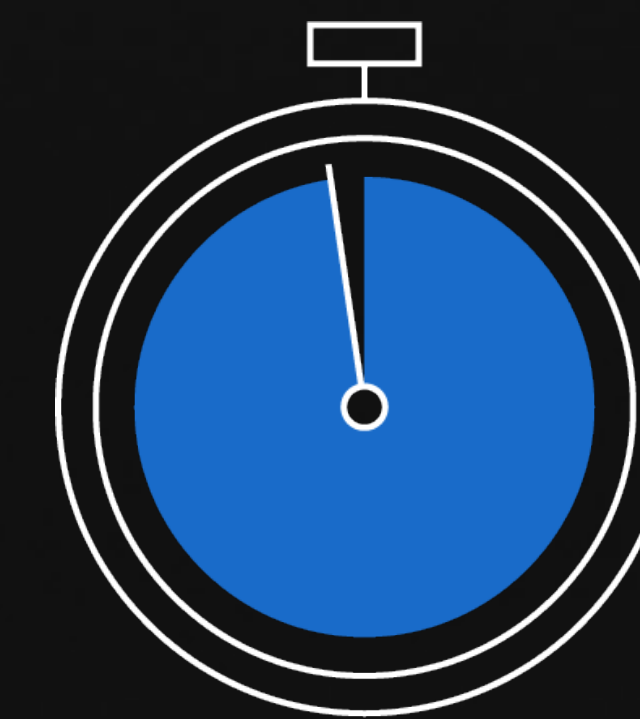
RUN GPT-3 BASED CHATBOTS



16 QUERIES
1 Second
DGX A100



1 QUERY
>1 Minute
Dual Socket CPU Server



NAVER

ANNOUNCING NAVER ADOPTS DGX SUPERPOD TO CREATE LANGUAGE UNDERSTANDING AI SERVICES

#1 Internet Tech Platform in Korea

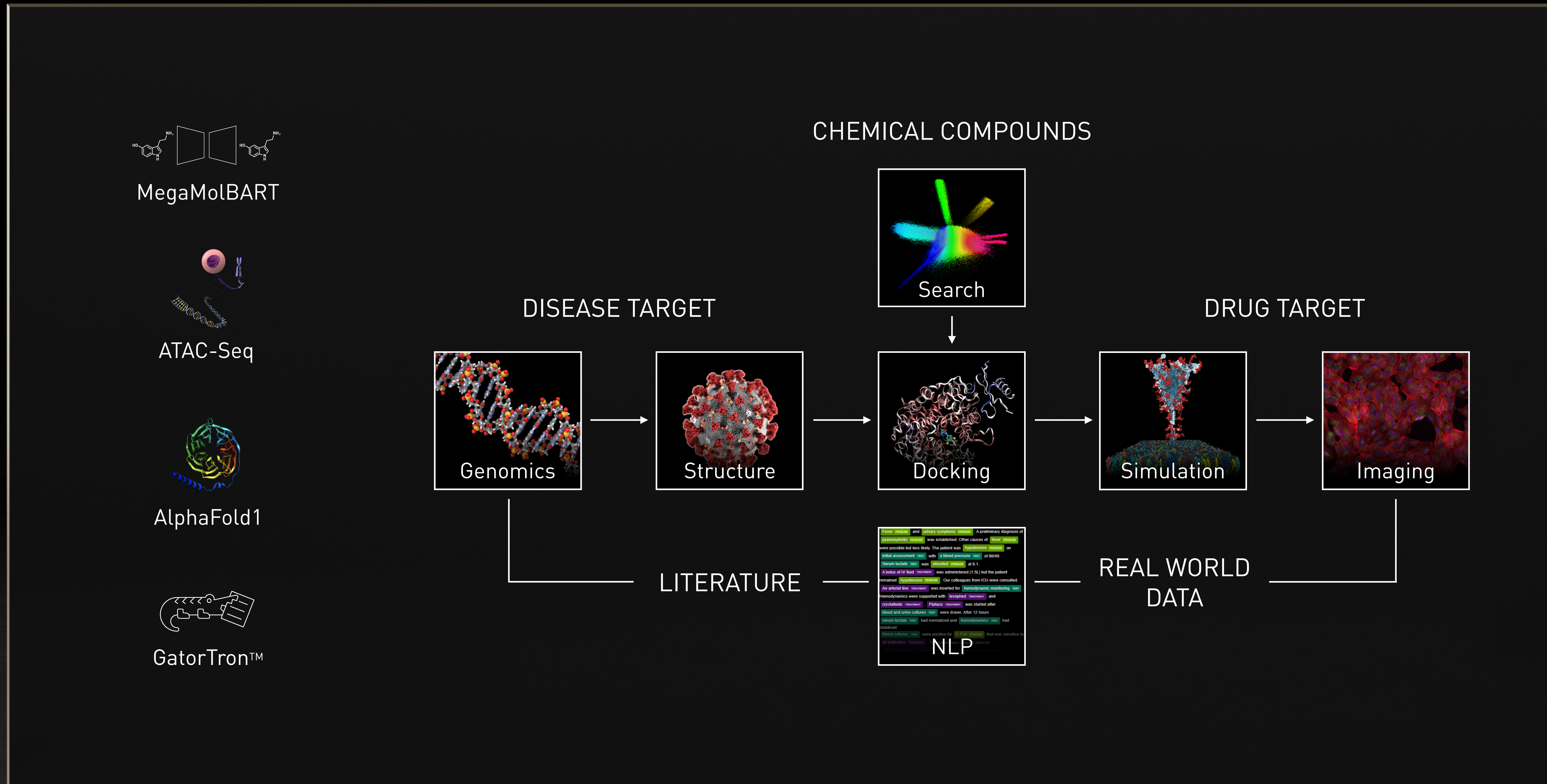
Construct AI R&D Belt Across Asia and Europe

Develop Giant AI Language Models

DGX SuperPOD to Build Global-Scale
AI Supercomputing Infrastructure



NVIDIA CLARA DISCOVERY



ANNOUNCING OXFORD NANOPORE TECHNOLOGIES ADOPTS NVIDIA DGX



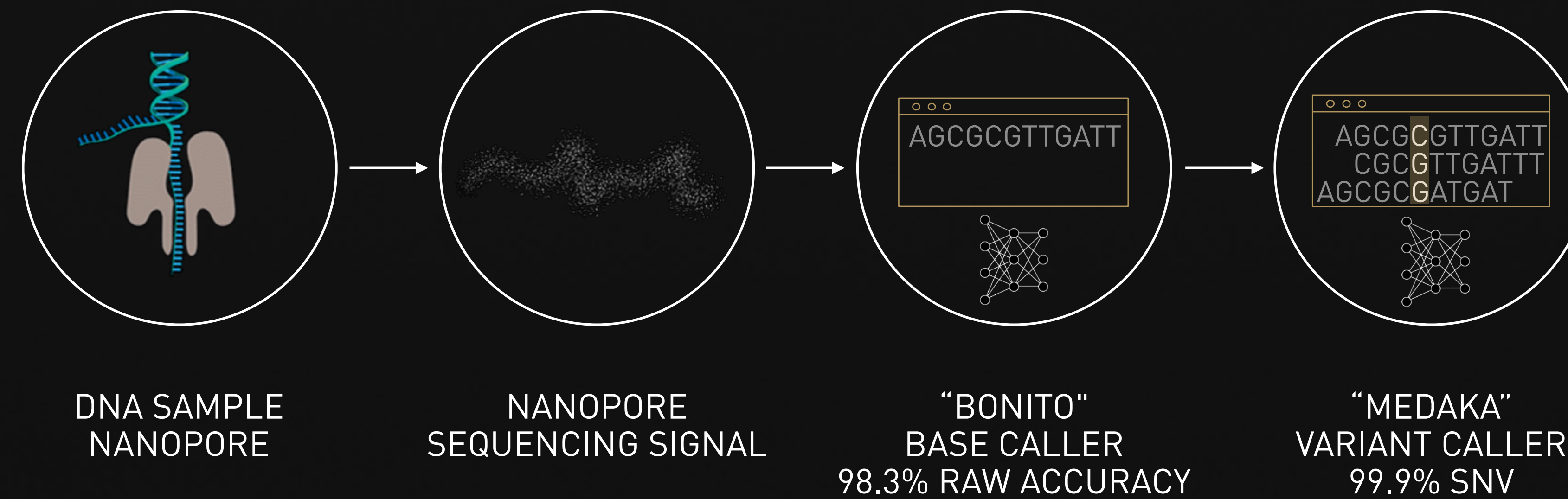
GridION with V100



MinION Mk1C with AGX

NANOPORE SEQUENCE PROCESS

Model Trained on DGX SuperPOD | Inference on DGX or V100 or Jetson



SCHRÖDINGER®

ANNOUNCING
SCHRÖDINGER AND NVIDIA
ACCELERATE DRUG DISCOVERY

\$1.25 Trillion Dollar Industry | 10+ Years
Development | 3,000 Pharma Companies

Accelerating Schrodinger's Physics and Machine Learning-
Based Computational Drug Discovery Tools with NVIDIA
Clara Discovery Acceleration Libraries on NVIDIA DGX A100

Achieve Experimentally-Accurate Simulations



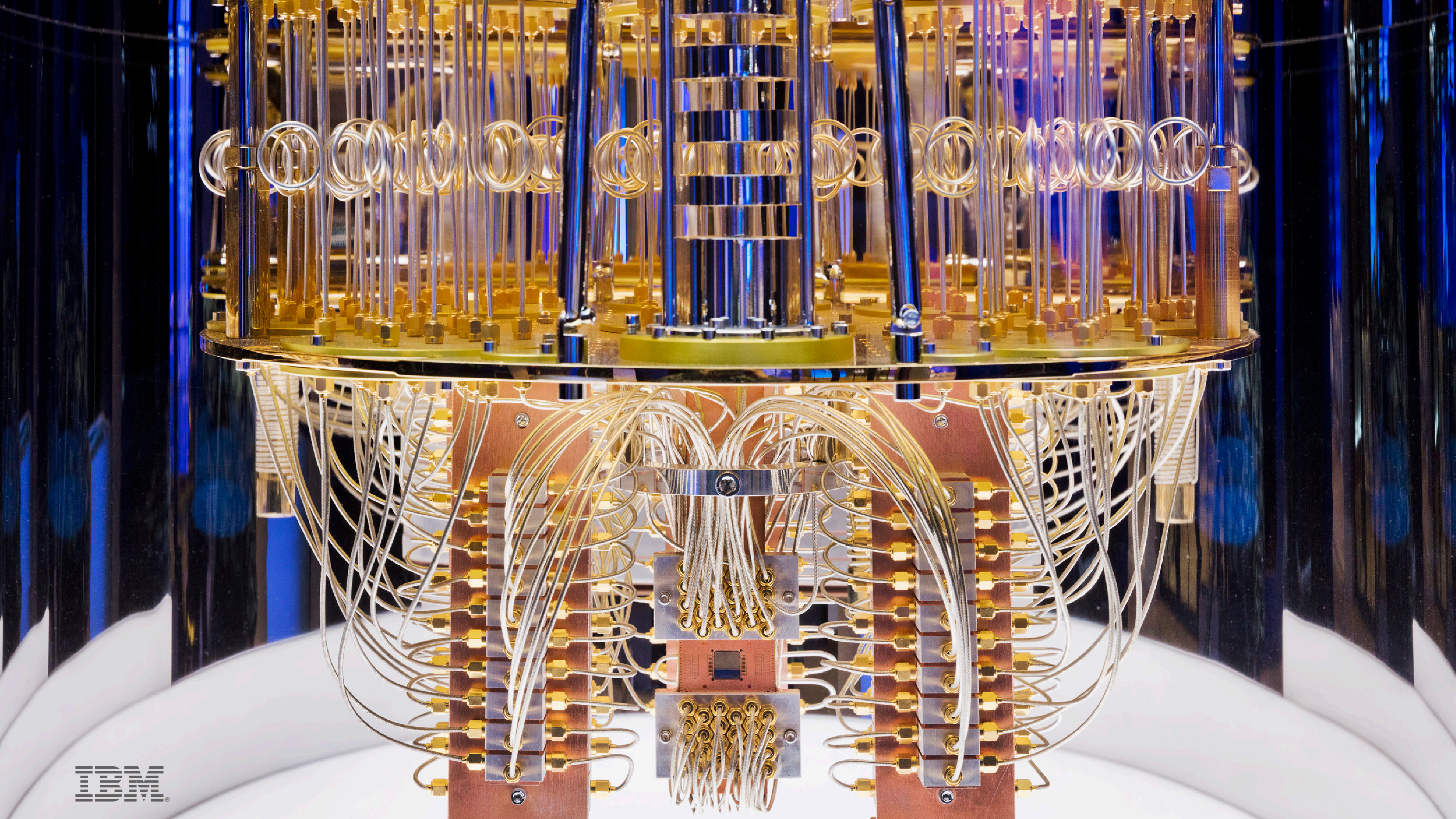
ANNOUNCING
RECURSION BUILDS
PHARMA AI SUPERCOMPUTER
WITH NVIDIA DGX SUPERPOD

BioHive-1 Aims to Decode Biology and Industrialize Drug Discovery

Recursion OS Built on NVIDIA DGX SuperPOD Generates, Analyzes, and Derives Insight from Biological and Chemical Datasets

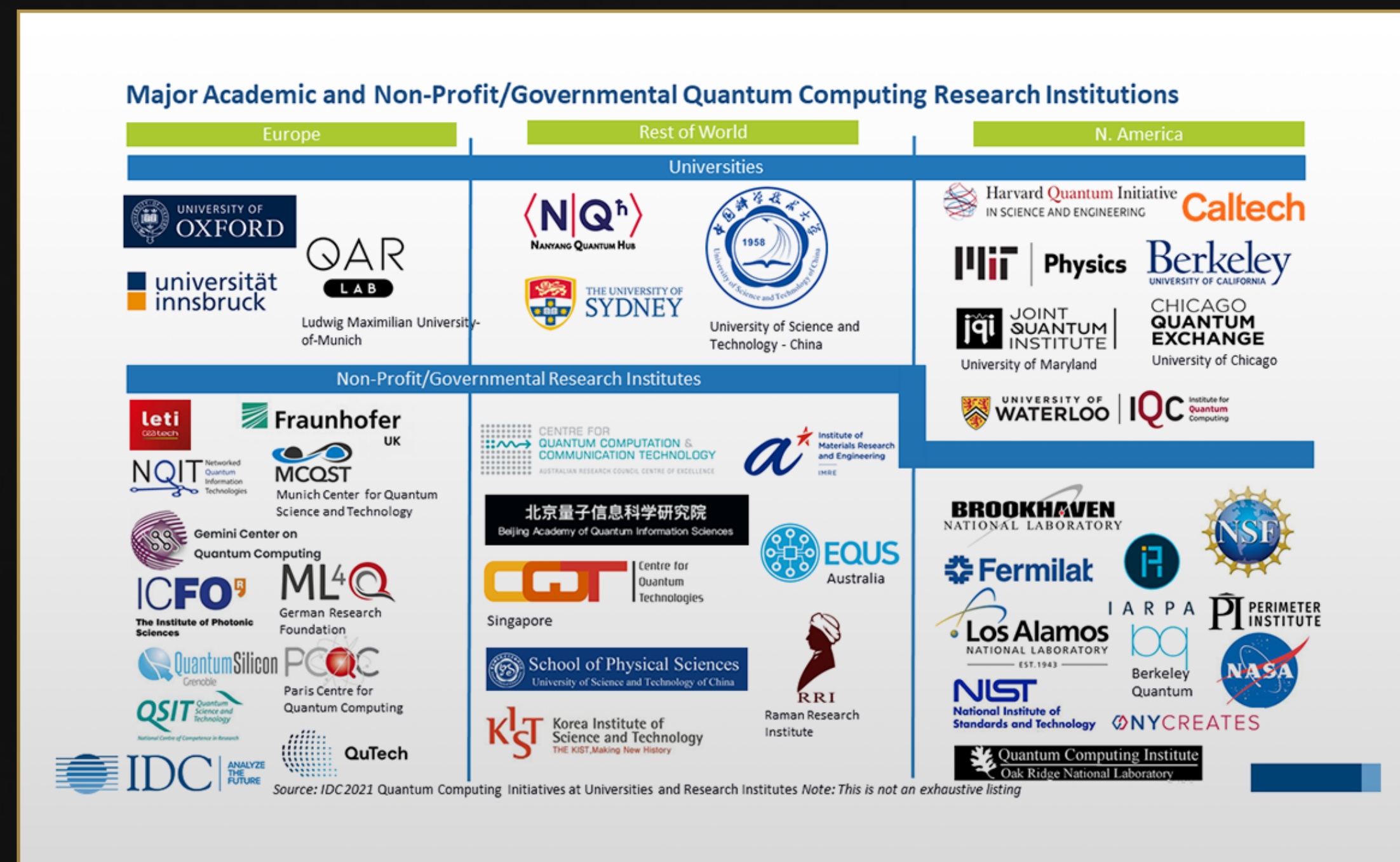
Generate up to 9 Million Images, or Approximately 80 Terabytes of Data, Across up to 1.5 Million Experiments per Week



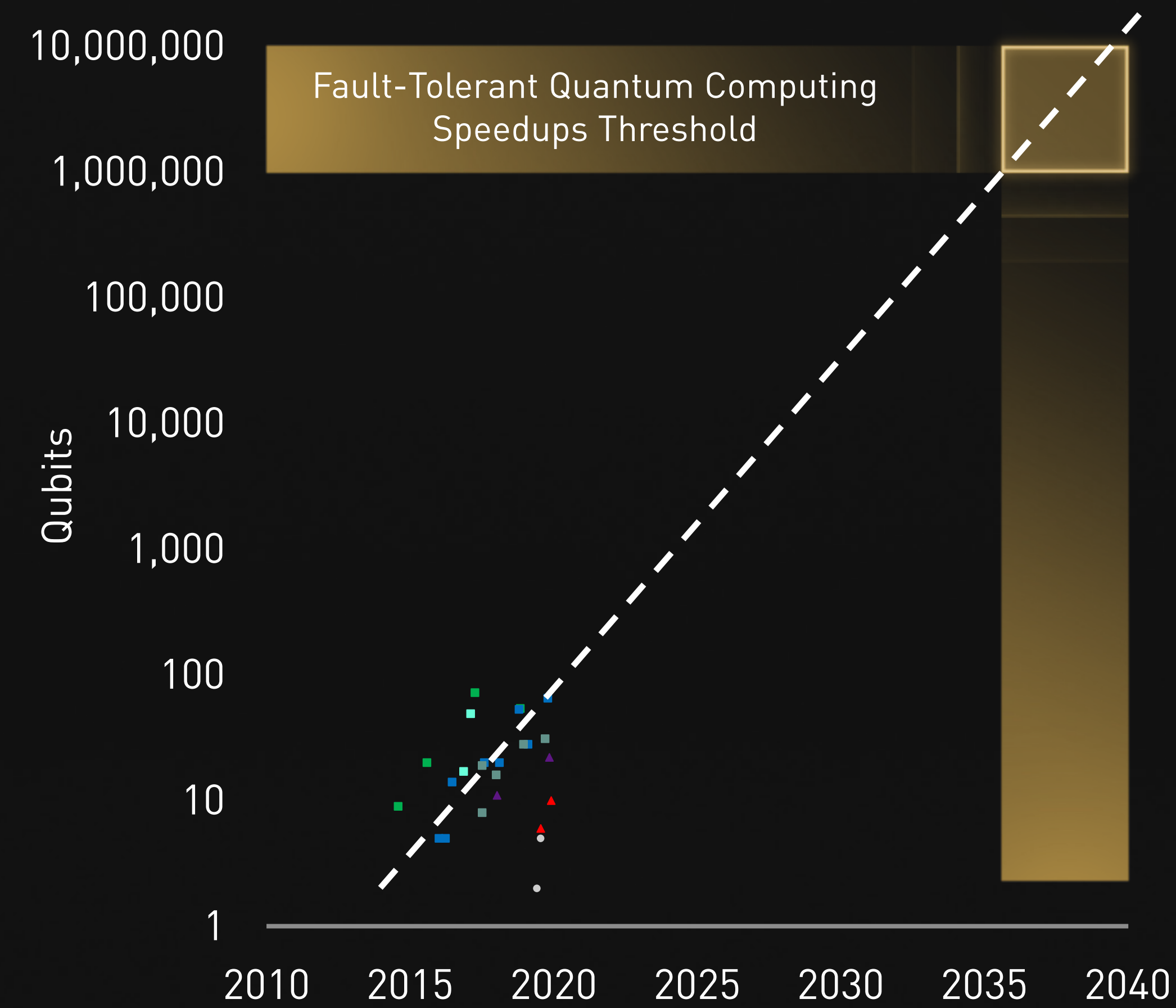


GLOBAL QUANTUM COMPUTING RACE

QUANTUM COMPUTING RESEARCH ECOSYSTEM



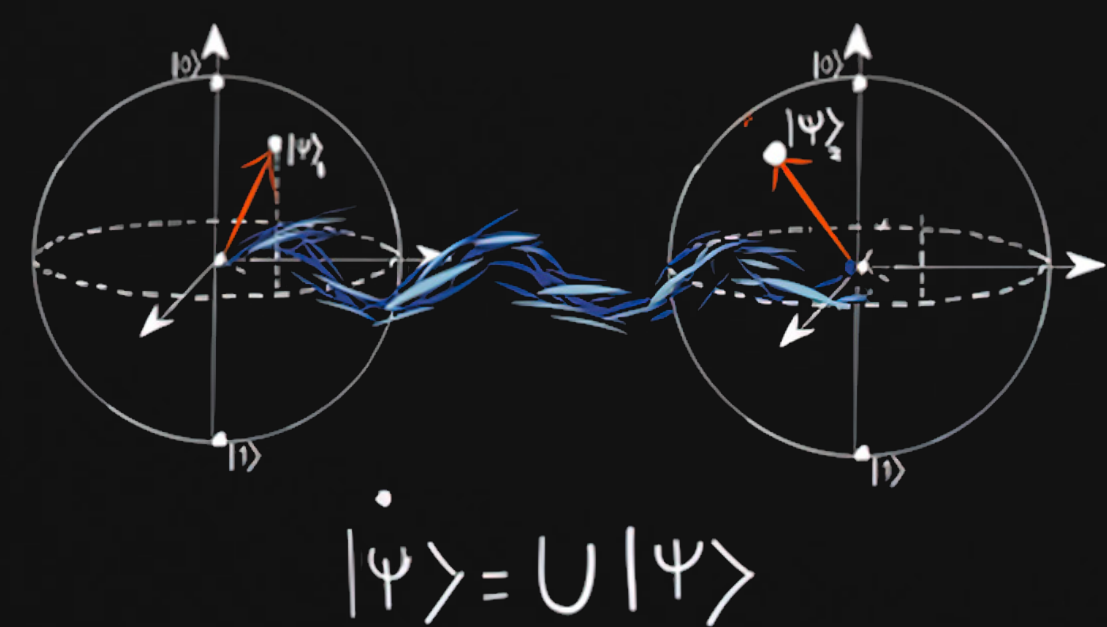
DOUBLING PHYSICAL QUBITS EVERY YEAR



ANNOUNCING NVIDIA CUQUANTUM

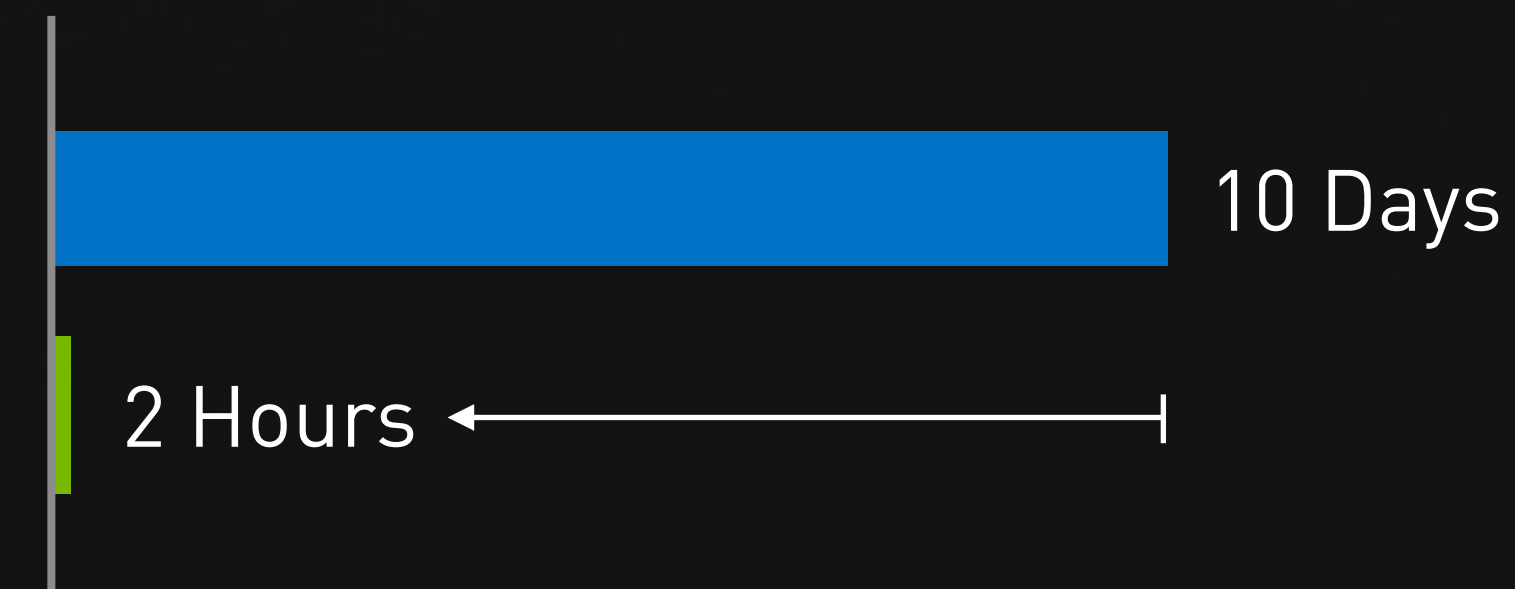
Research the Computer of Tomorrow on the Most Powerful Computer Today

GPU-ACCELERATED QUANTUM SIMULATIONS



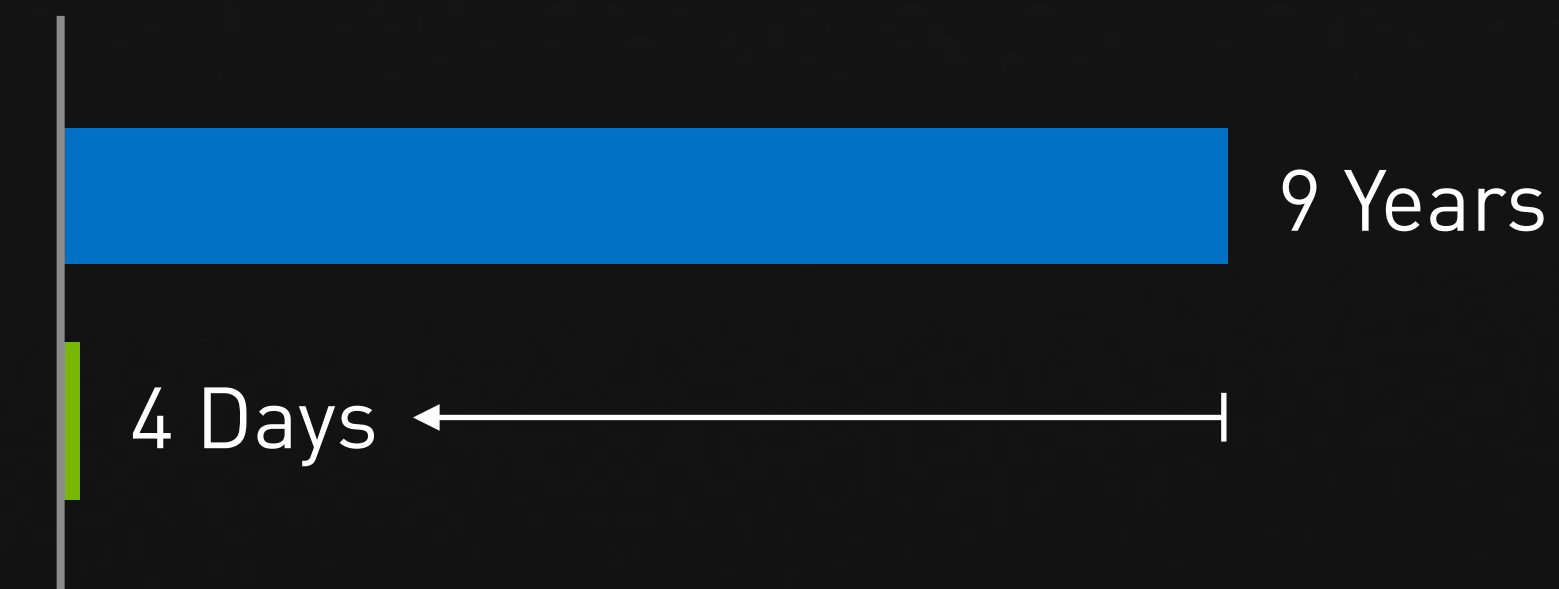
STATE VECTOR SIMULATION

Scales to 10's of Qubits



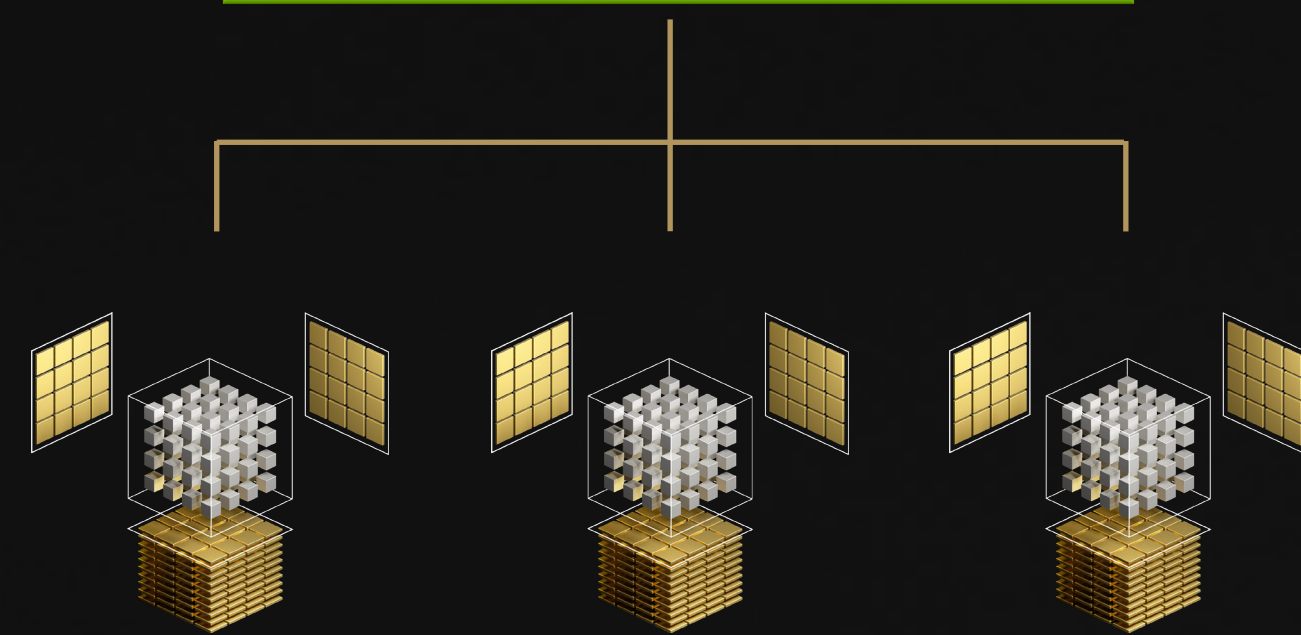
TENSOR NETWORK SIMULATION

Scales to 1000's of Qubits



■ Dual-CPU ■ DGX A100

cuQuantum



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

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



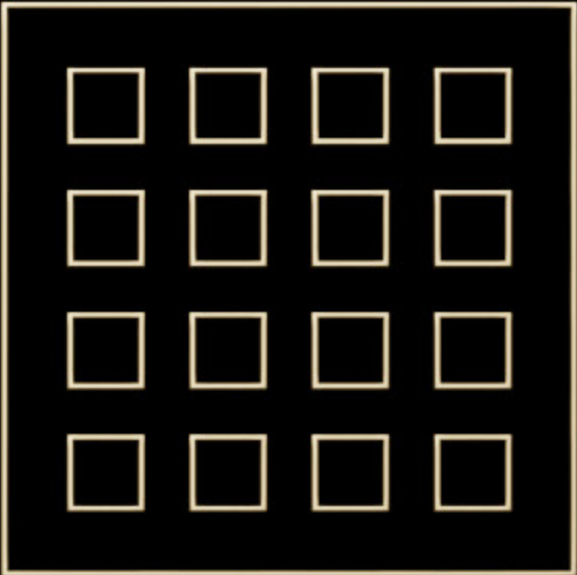
DIVERSE DATA CENTER ARCHITECTURES



ENTERPRISE COMPUTING



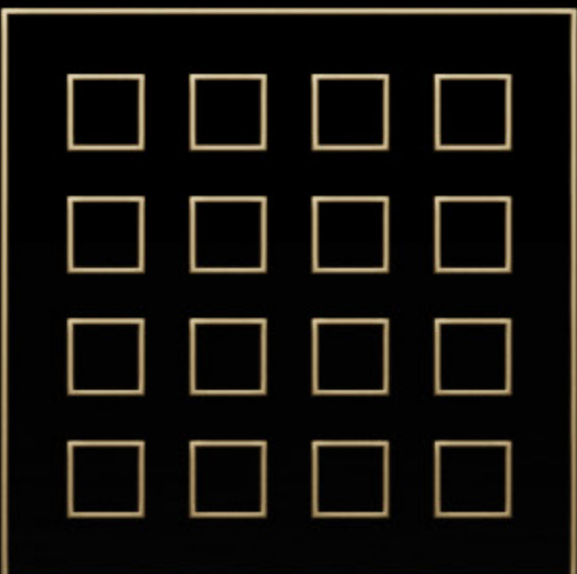
HYPERSCALE



SCIENTIFIC COMPUTING



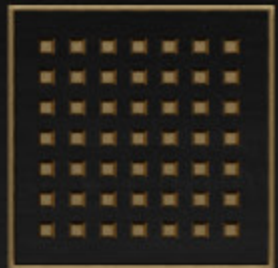
STORAGE SERVERS



ACCELERATED HPC



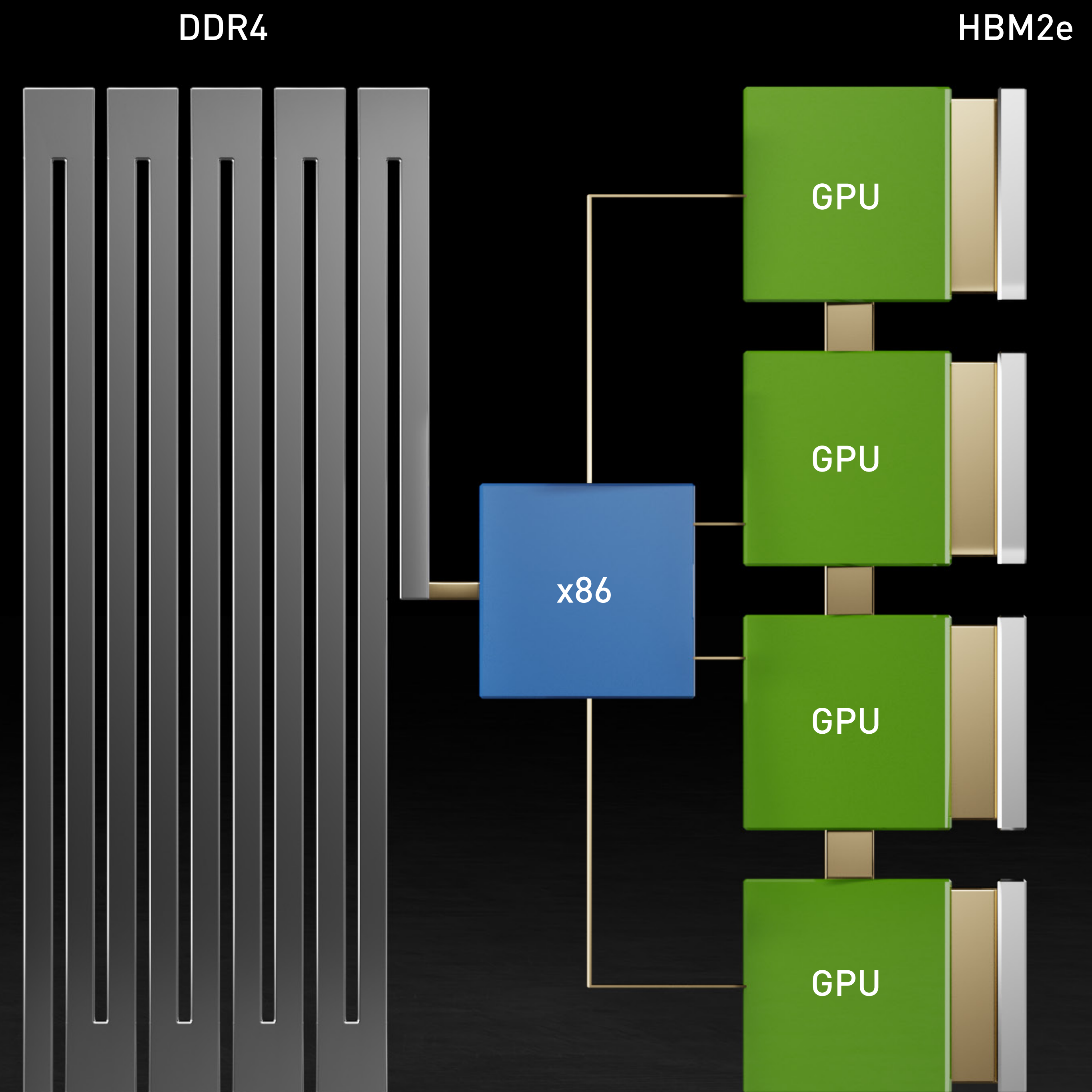
ACCELERATED HYPERSCALE



DATA-COMPUTE DEMAND GROWING FASTER THAN SYSTEM BANDWIDTH

GPU Starved by CPU Memory and PCIe Bandwidth

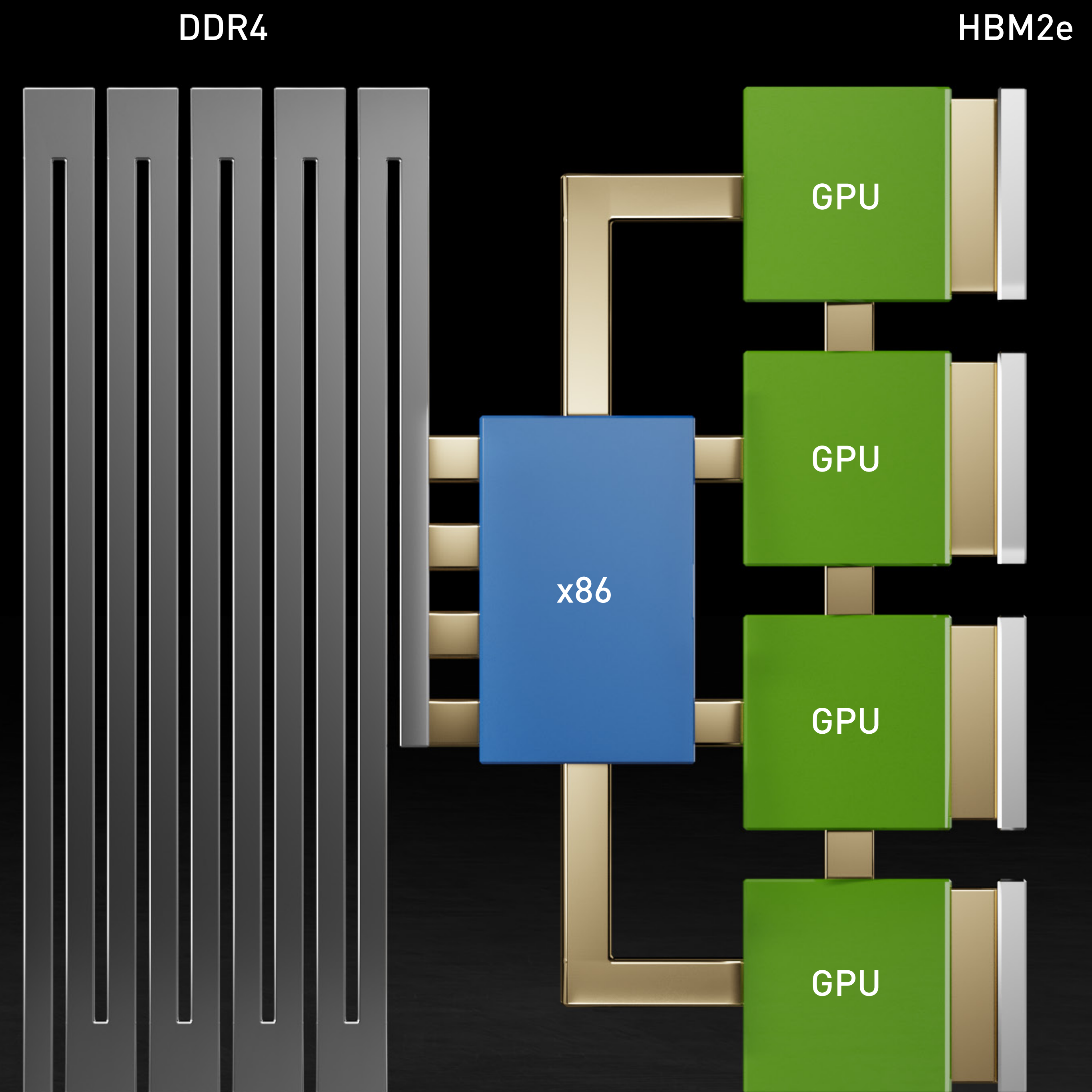
GPU	8,000	GB/sec
CPU	200	GB/sec
PCIe Gen 4	16	GB/sec
Mem-to-GPU	64	GB/sec



DATA-COMPUTE DEMAND GROWING FASTER THAN SYSTEM BANDWIDTH

GPU Starved by CPU Memory and PCIE Bandwidth

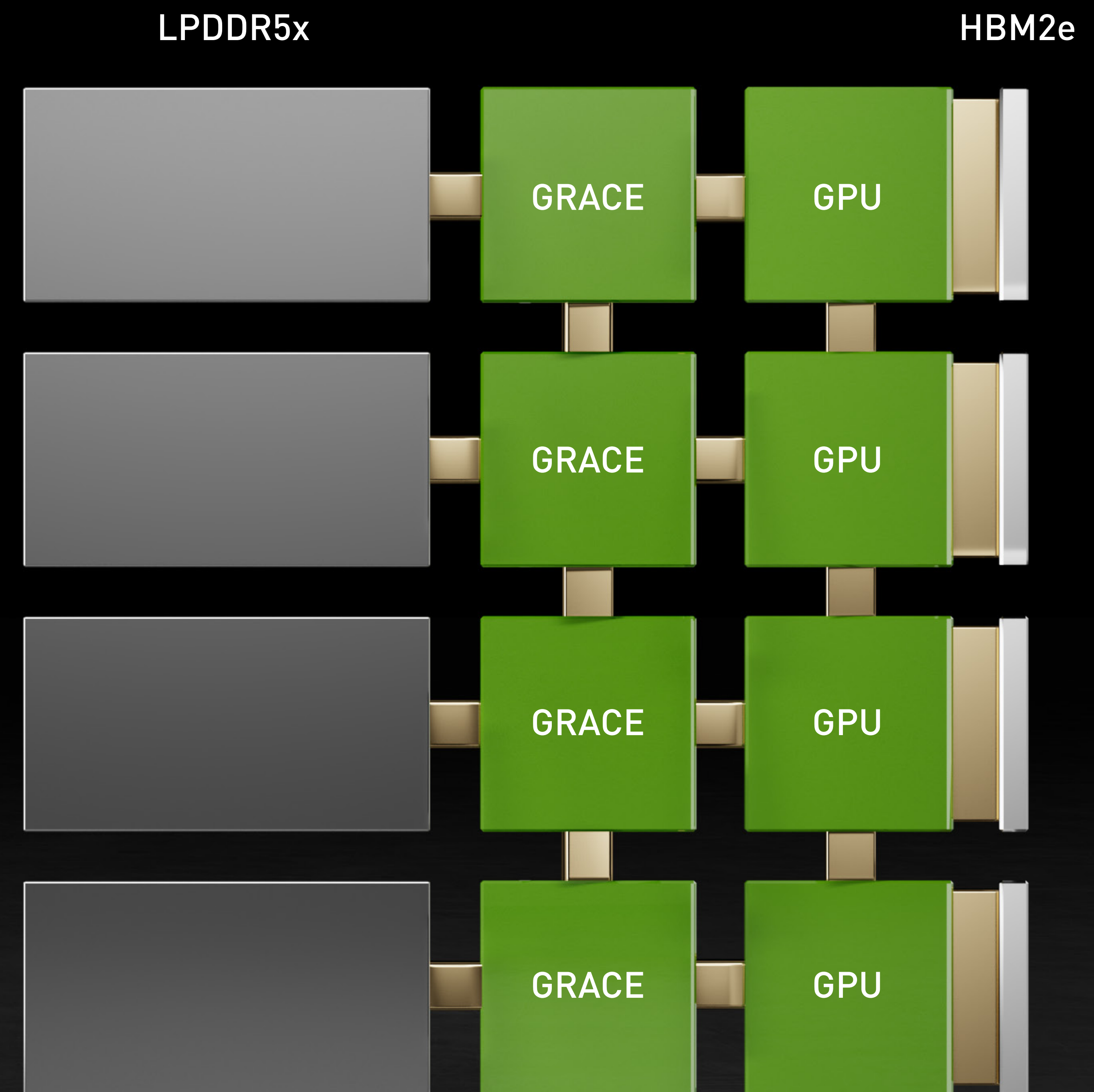
GPU	8,000	GB/sec
CPU	200	GB/sec
PCIE Gen 4	16	GB/sec
Mem-to-GPU	64	GB/sec



A NEW COMPUTING ARCHITECTURE FOR AI AND DATA SCIENCE

30X Increase System Memory to GPU

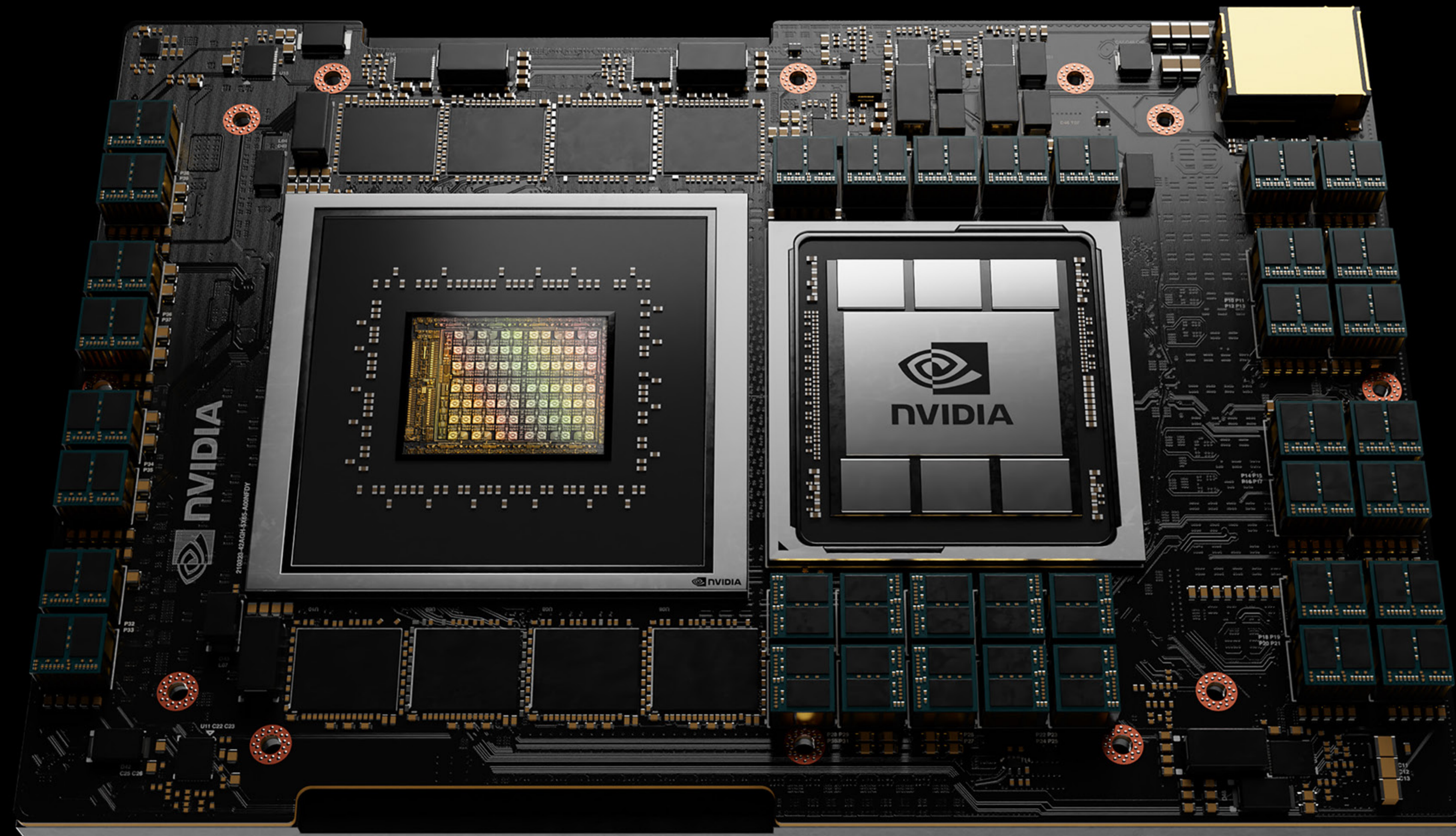
GPU	8,000	GB/sec	
CPU	500	GB/sec	
NVLINK	500	GB/sec	
Mem-to-GPU	2,000	GB/sec	30X





ANNOUNCING NVIDIA GRACE

CPU Designed for Giant-Scale AI and HPC Accelerated Computing



ANNOUNCING THE WORLD'S FASTEST SUPERCOMPUTER FOR AI

20 Exaflops of AI

Powered by NVIDIA Grace CPU and
Next Generation NVIDIA GPU

HPC and AI for Scientific and Commercial Apps

Advance Weather, Climate, and Material Science



CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre



**Hewlett Packard
Enterprise**



NVIDIA



SUPERCOMPUTING COMMUNITY EMBRACES ARM

“Alps will use NVIDIA’s novel Grace CPU to converge AI technologies and classic supercomputing in one single powerful data center infrastructure.”



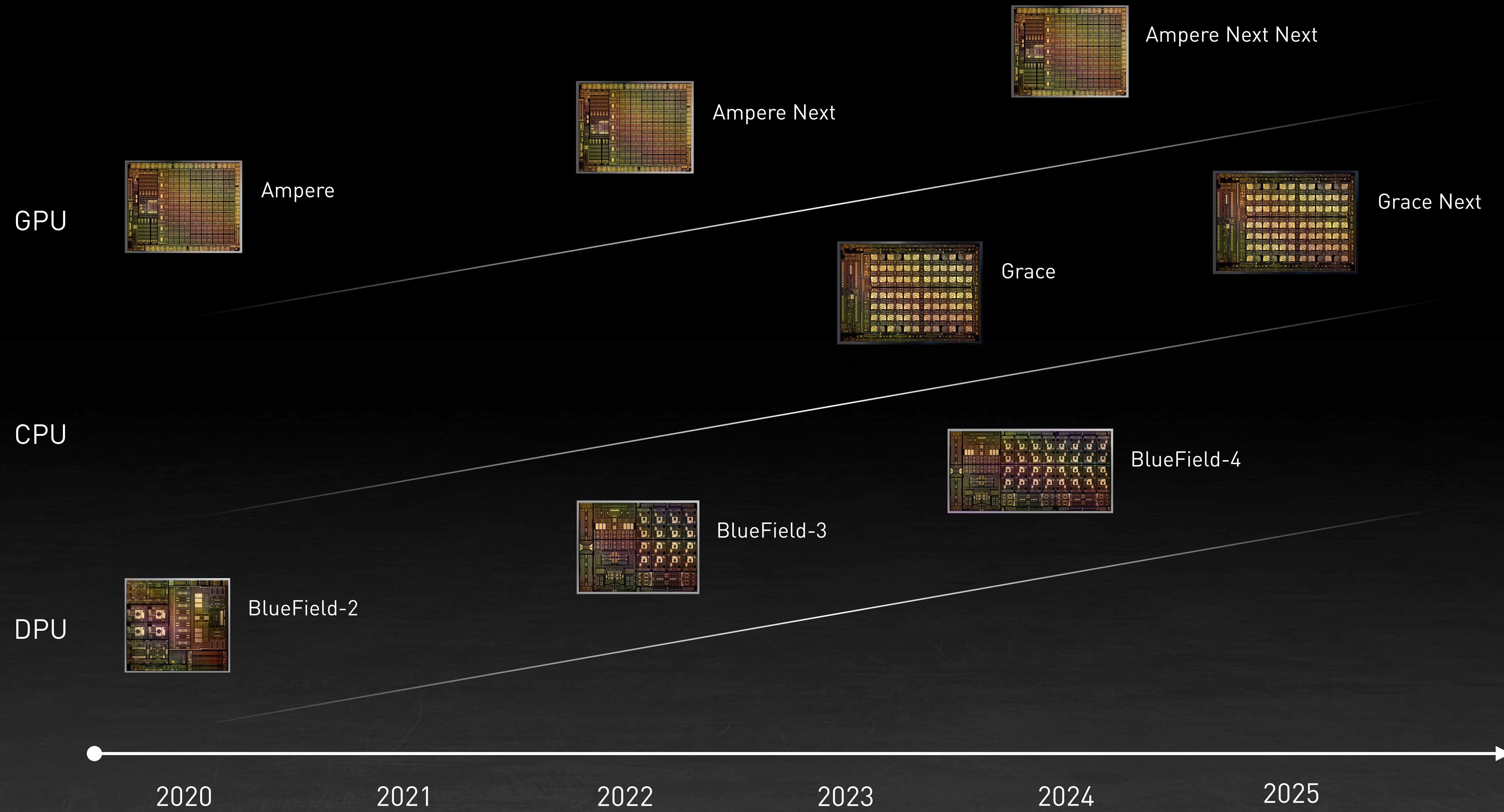
-Thomas Schulthess
Director CSCS

“Thanks to NVIDIA’s new Grace CPU, we’ll be able to deliver advanced scientific research using high-fidelity 3D simulations and analytics with data sets that are larger than previously possible.”

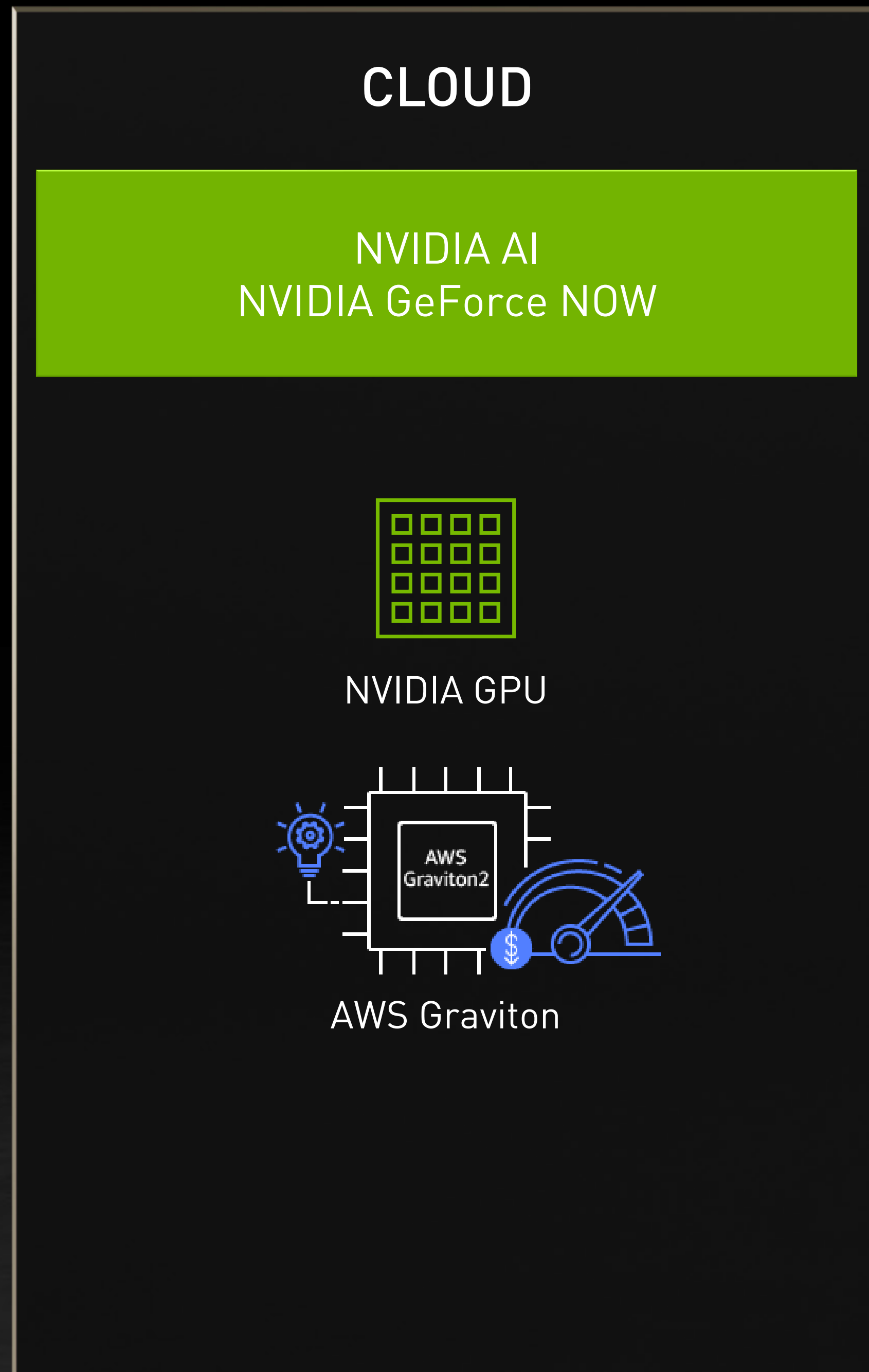


-Thom Mason
Director LANL

3 CHIPS. YEARLY LEAPS. ONE ARCHITECTURE.



EXPANDING ARM IN THE CLOUD



Bringing Together AWS Graviton2 CPU and NVIDIA GPU in 2H21

Android Gaming and AI Inference in the Cloud

AWS Graviton2 Delivers Significantly Better Price Performance

NVIDIA GPUs Deliver Performance to Scale Streaming

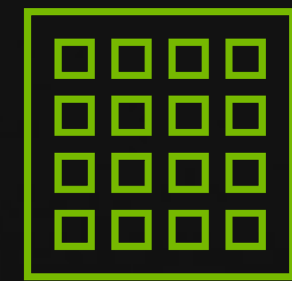
Easiest Way to Move Android Gaming to the Cloud



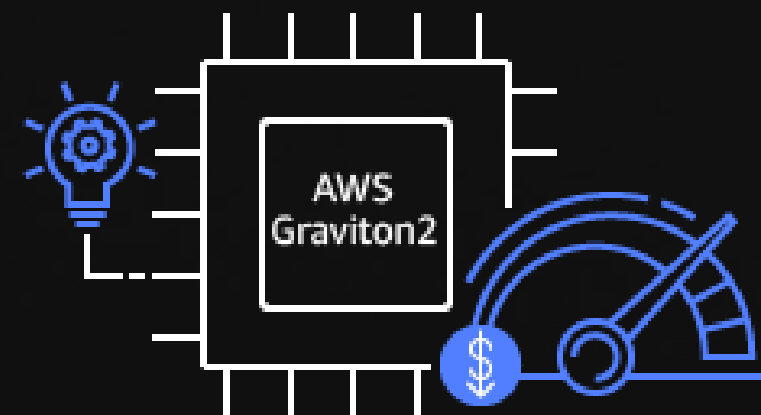
EXPANDING ARM ECOSYSTEM BEYOND MOBILE

CLOUD

NVIDIA AI
NVIDIA GeForce NOW



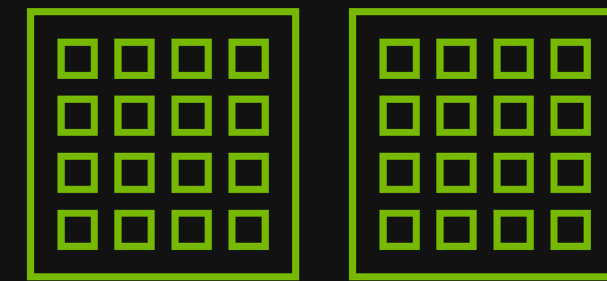
NVIDIA GPU



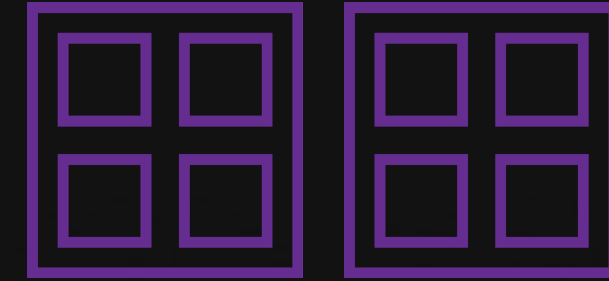
AWS Graviton

SCIENTIFIC COMPUTING

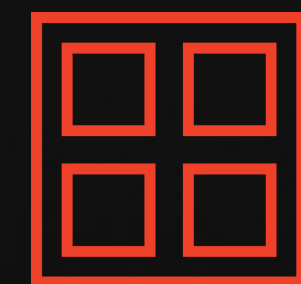
NVIDIA AI
NVIDIA HPC



NVIDIA A100



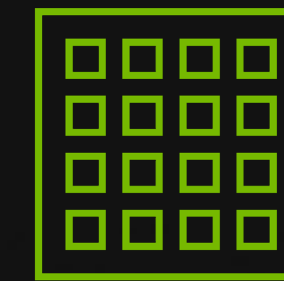
NVIDIA BlueField-2



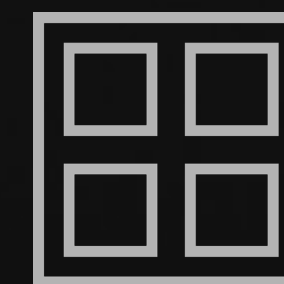
Ampere Altra

EDGE AND ENTERPRISE

NVIDIA AI
NVIDIA DeepStream



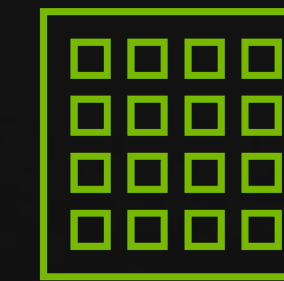
NVIDIA T4



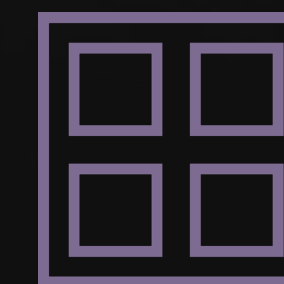
Marvell OCTEON

PC

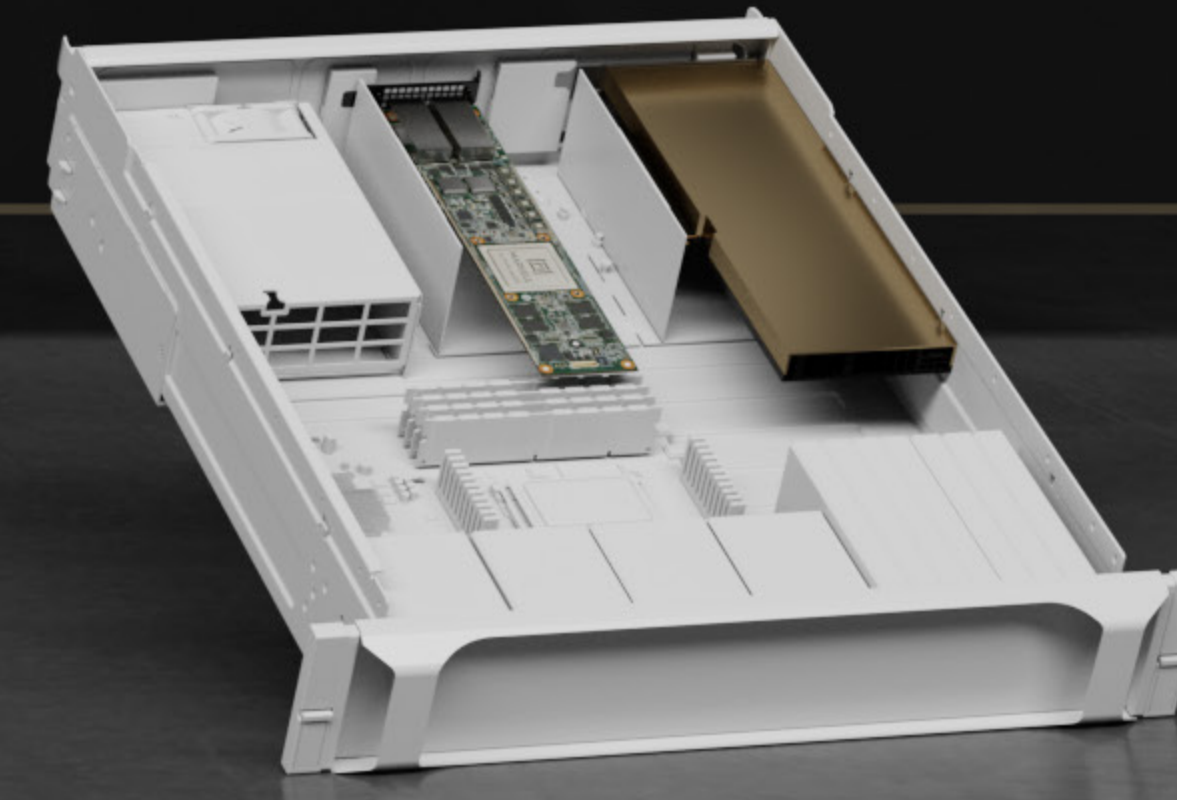
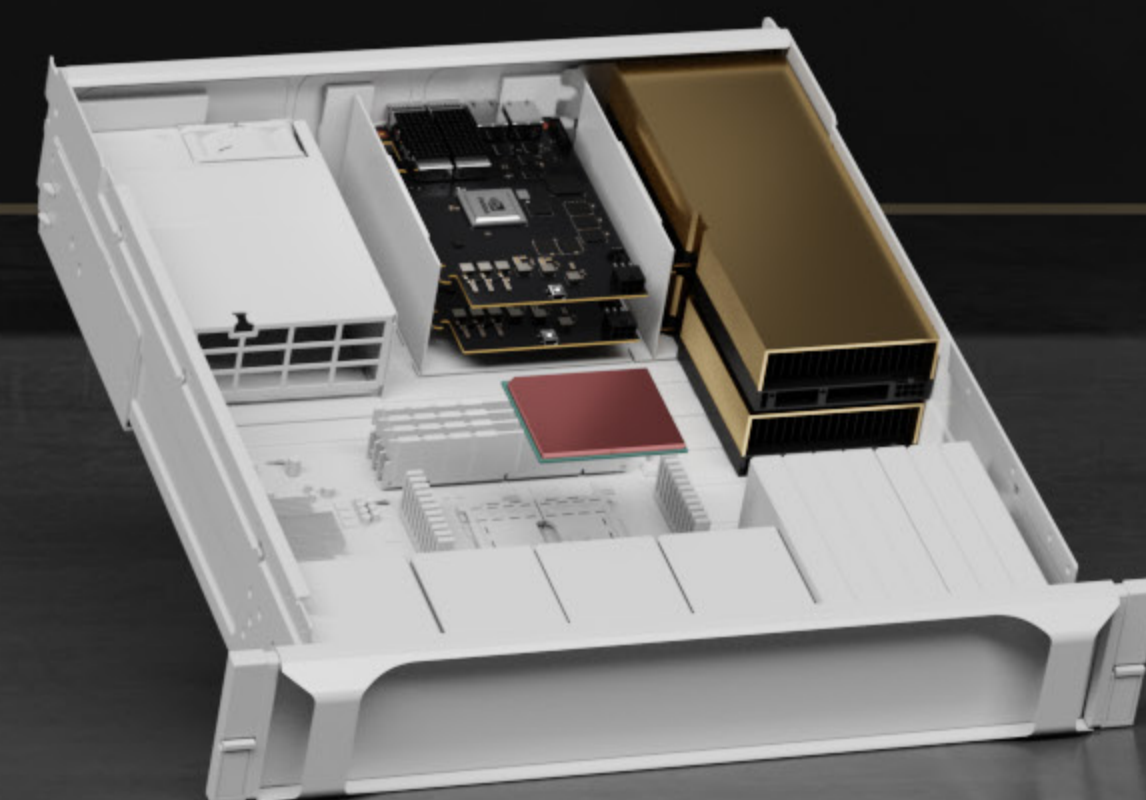
NVIDIA AI
NVIDIA RTX



GeForce RTX 30 Series

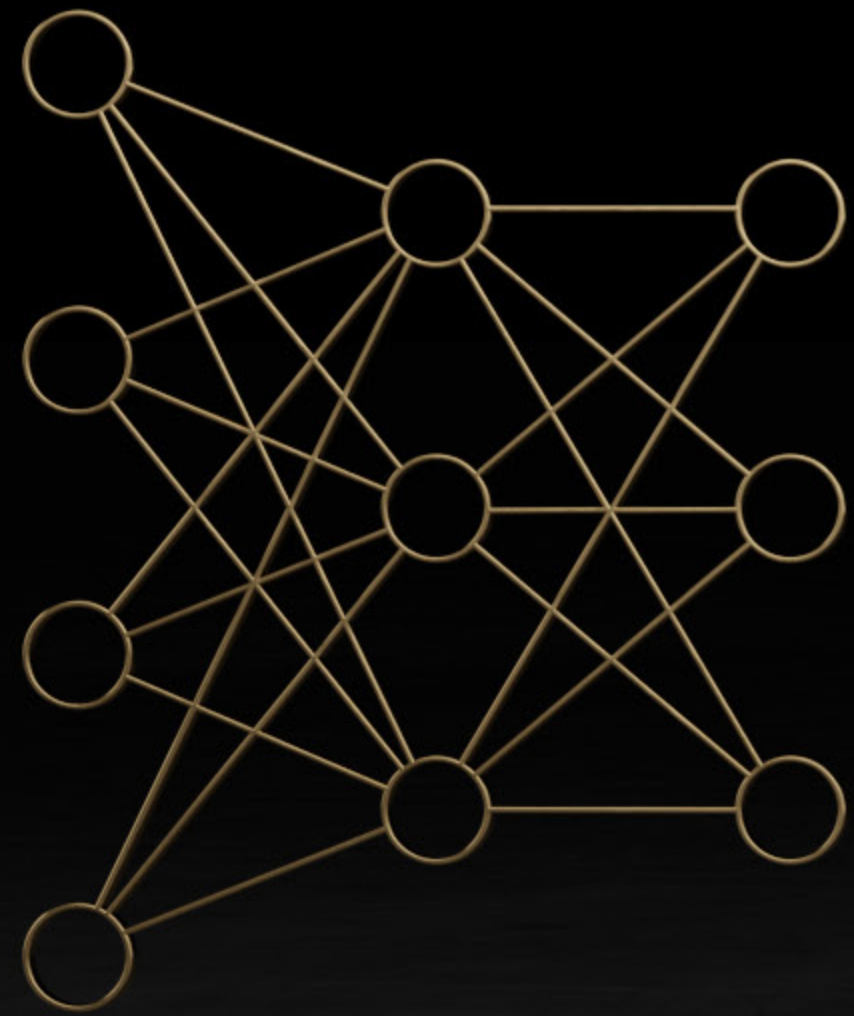


MediaTek MT819x



WAVES OF AI

AI Computing



Cloud



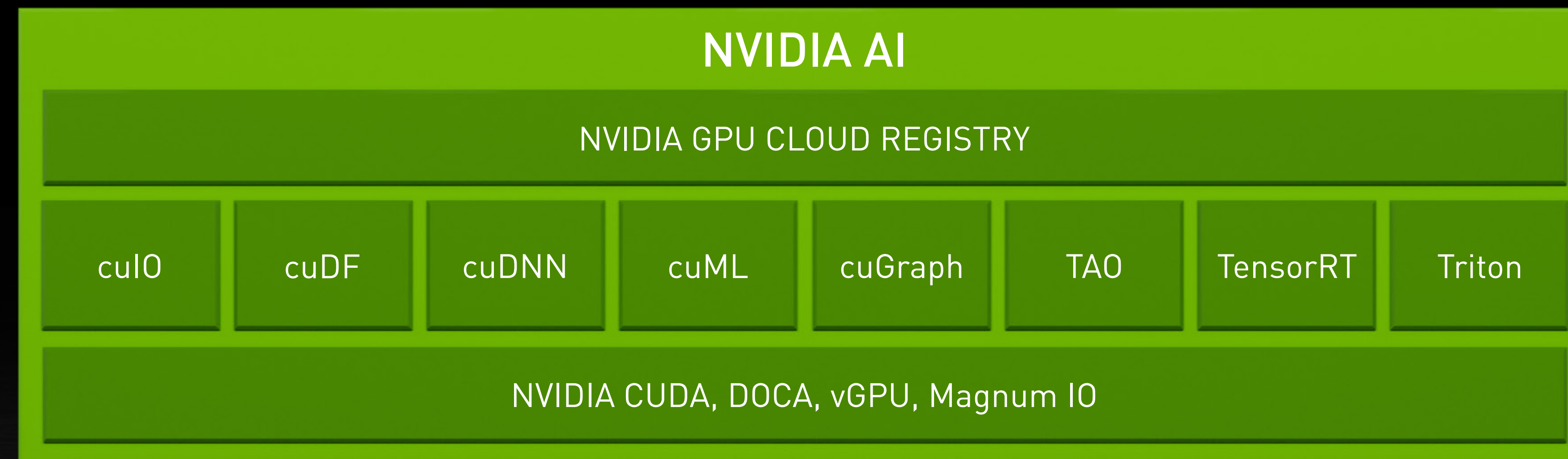
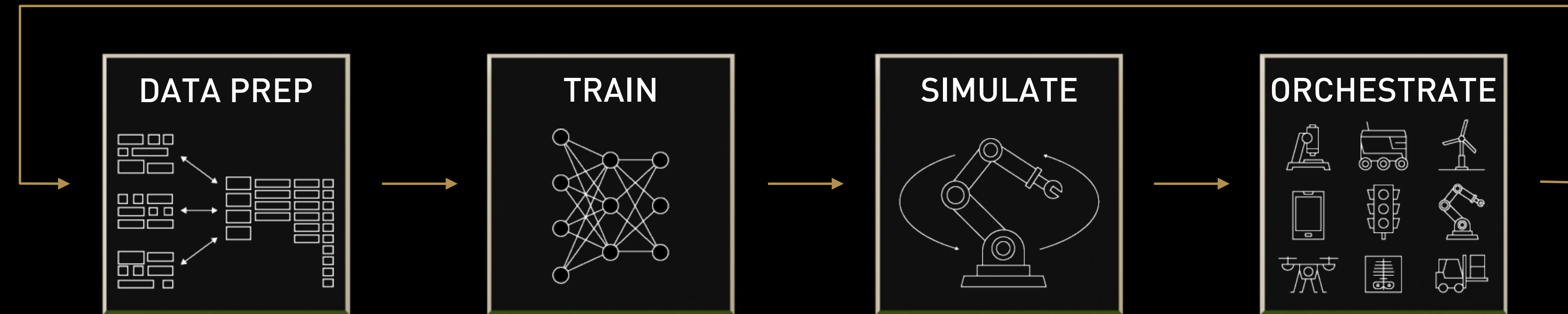
5G Industrial Edge



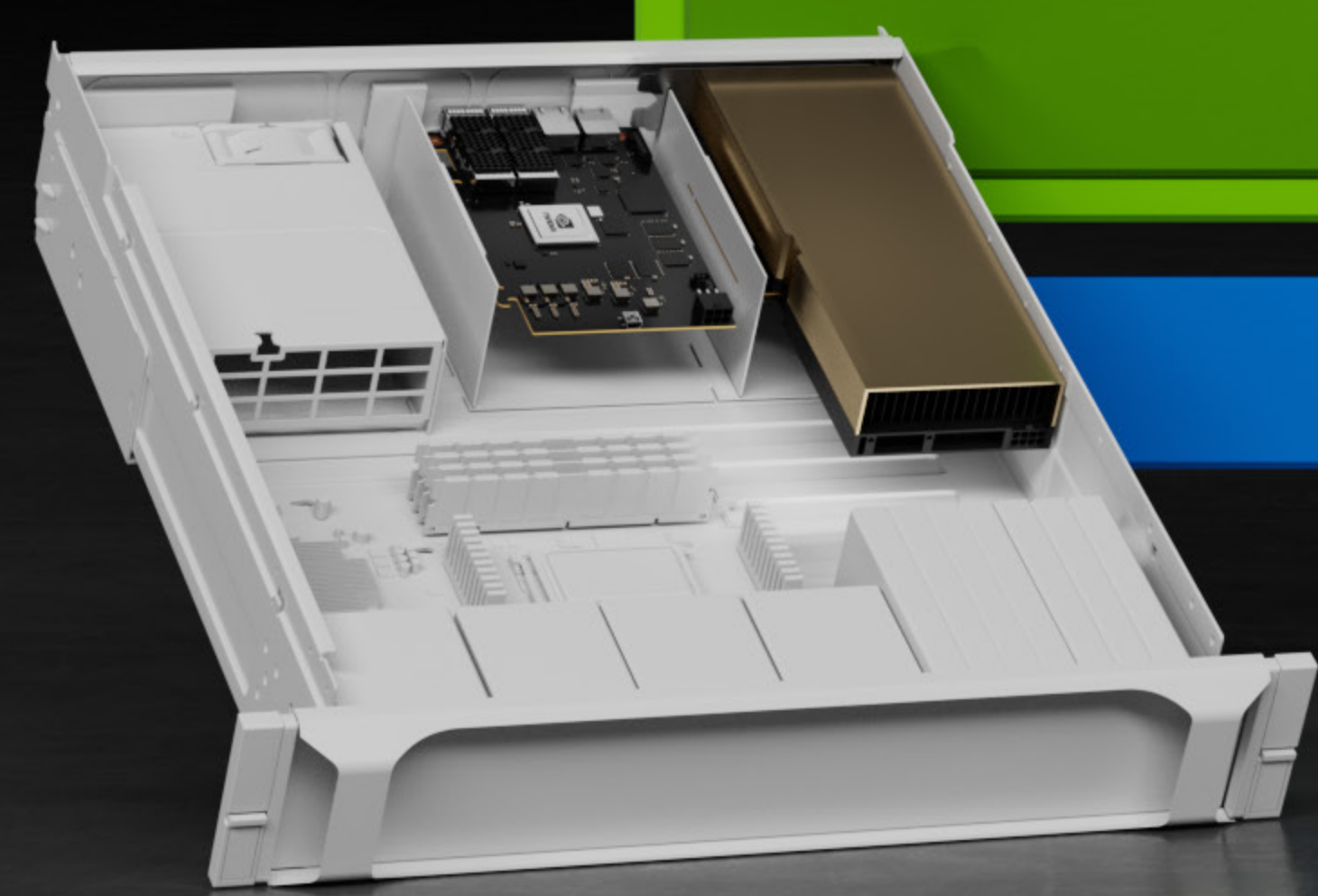
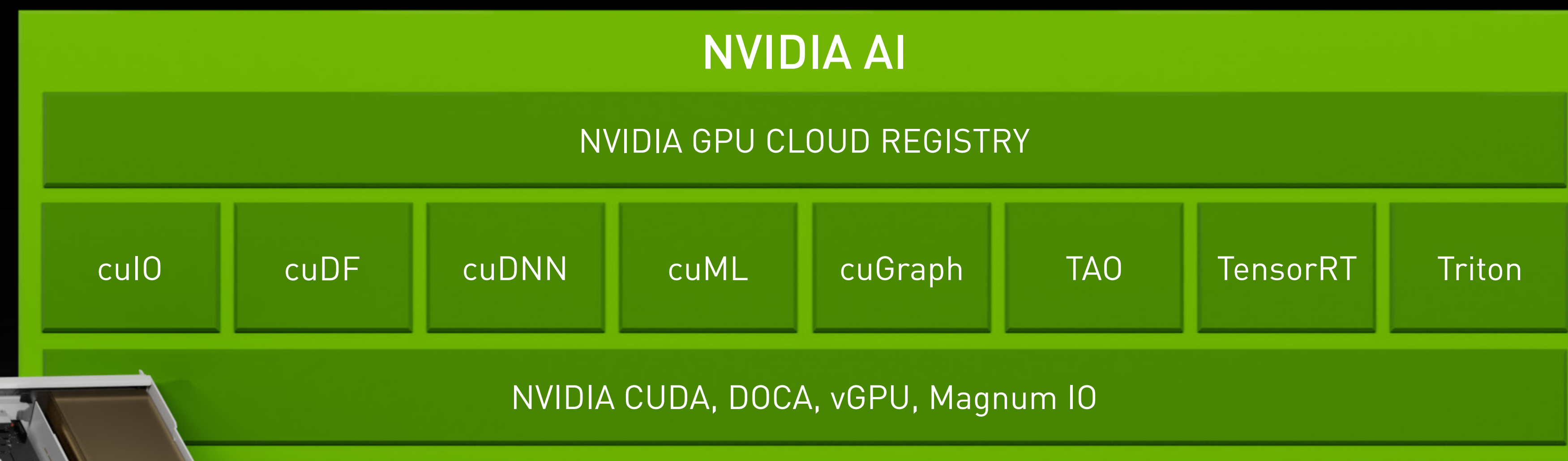
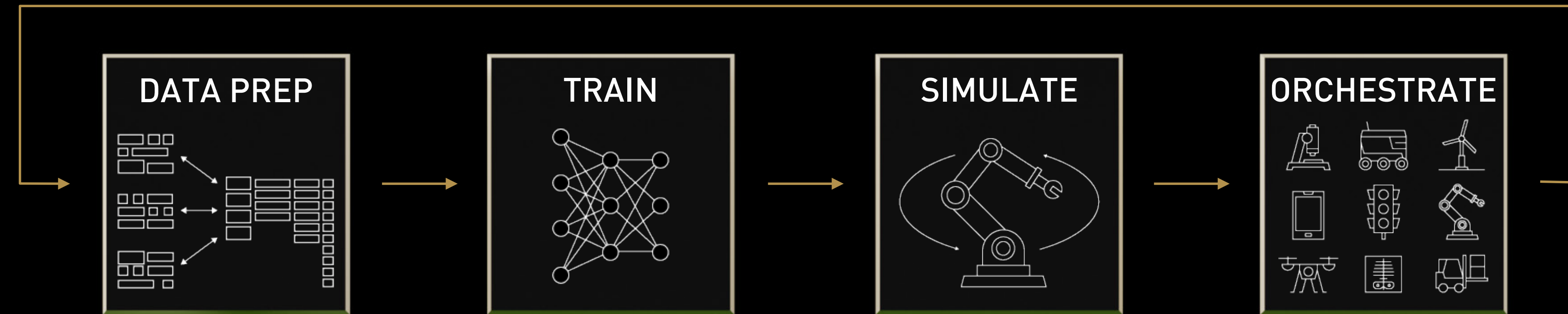
Robotics



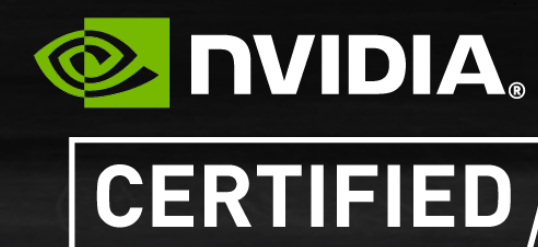
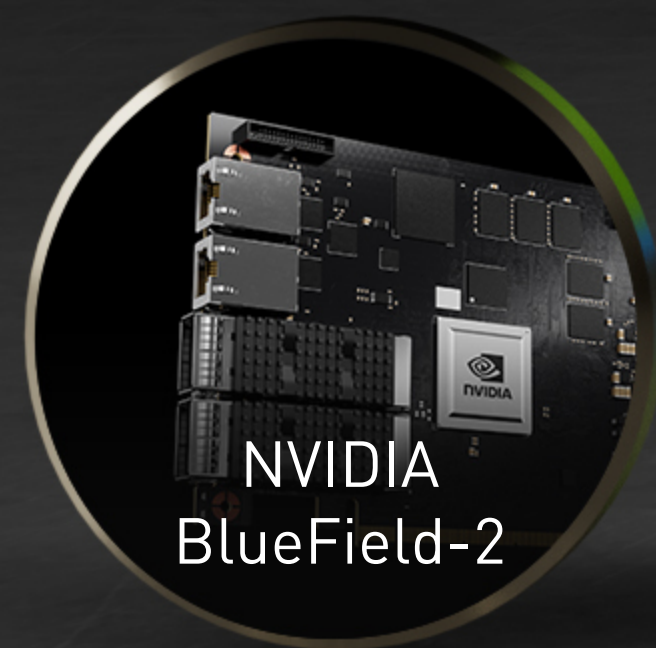
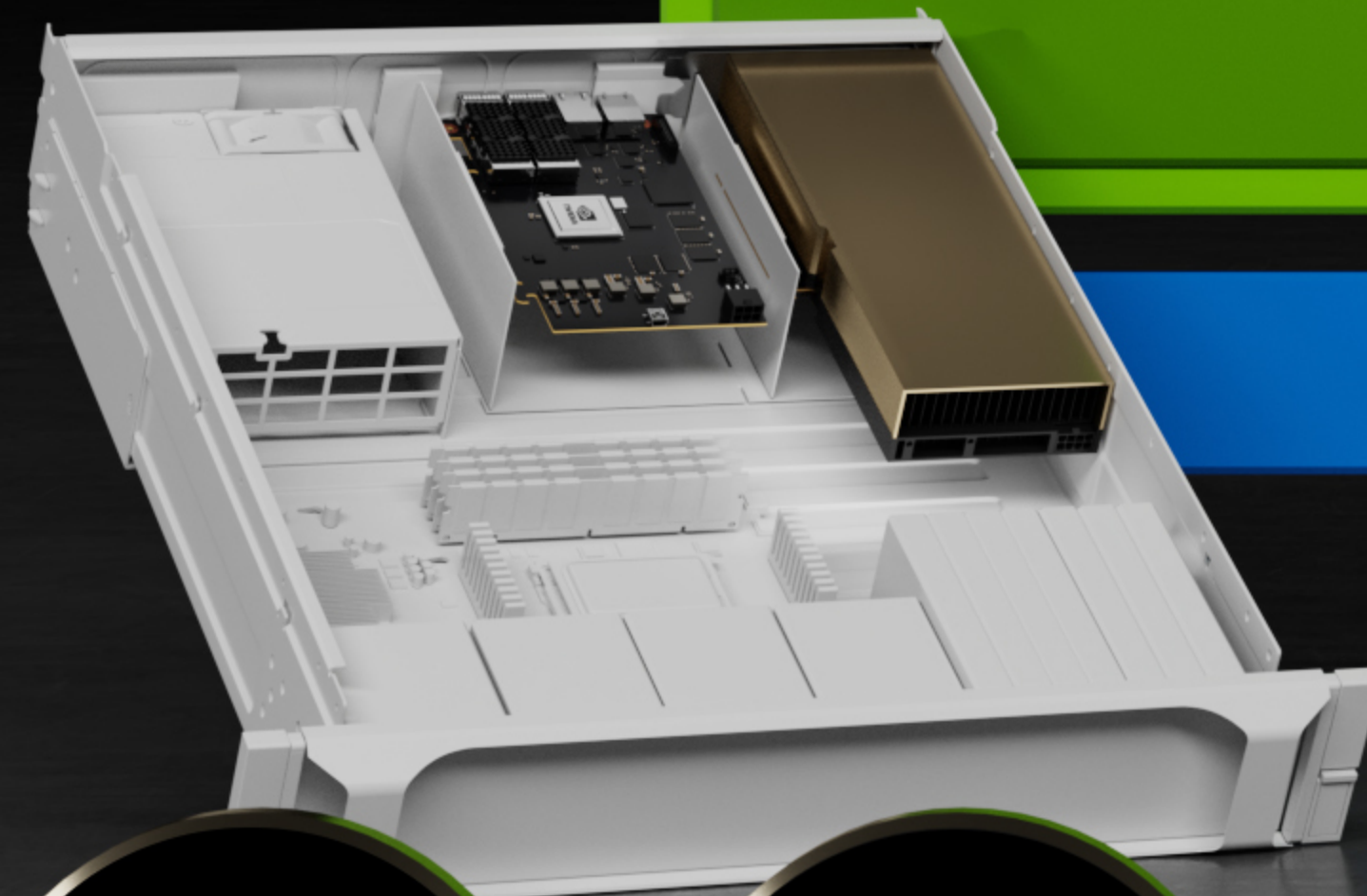
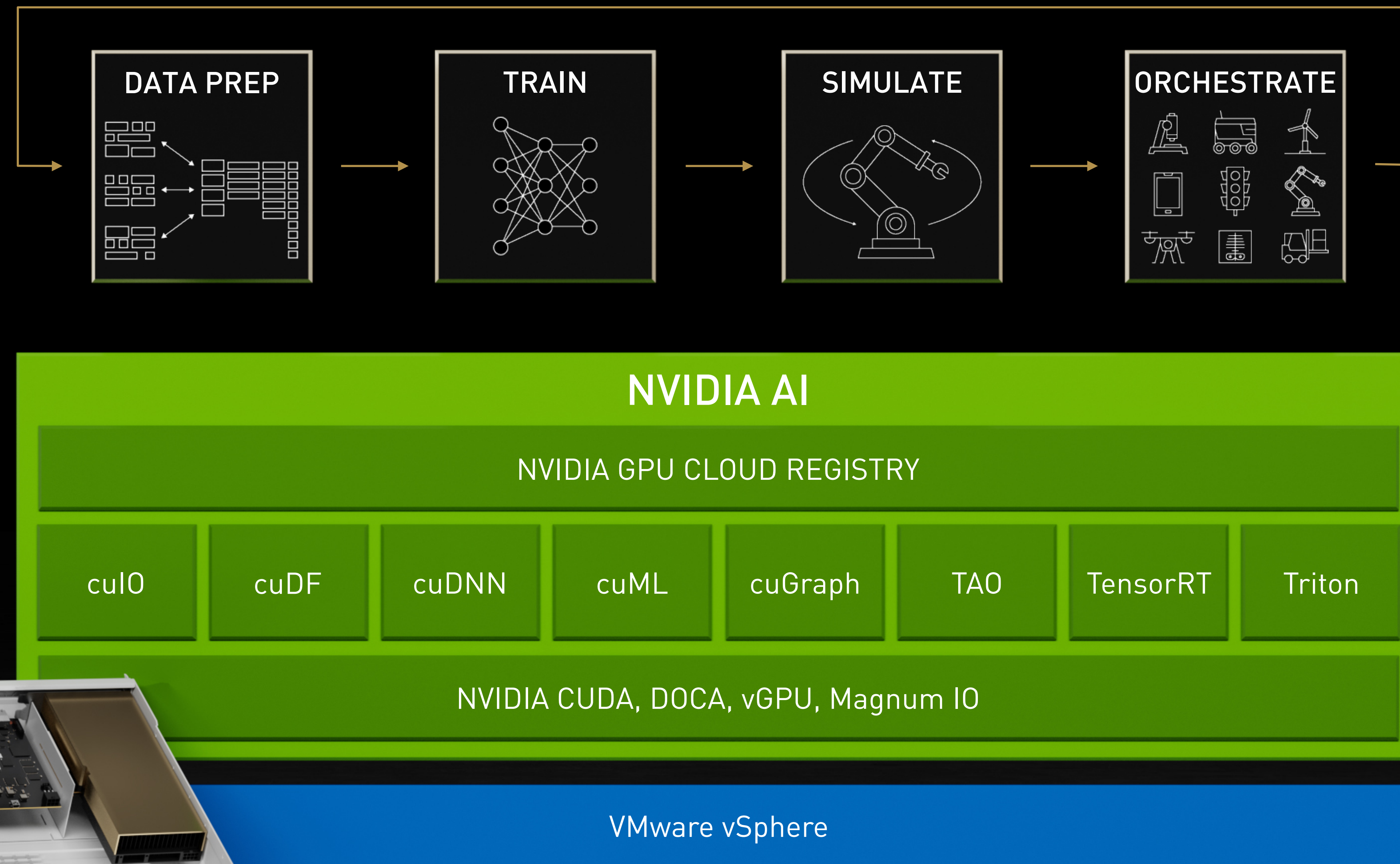
NVIDIA AI



ANNOUNCING NVIDIA EGX ENTERPRISE PLATFORM

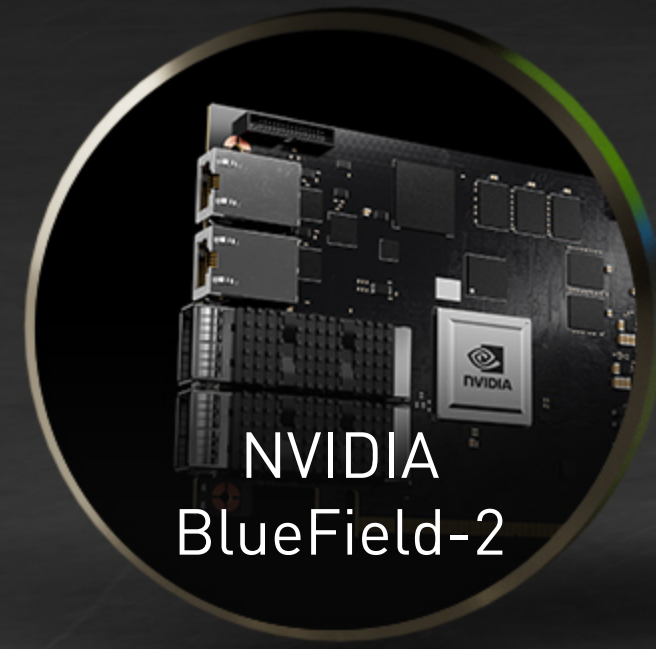
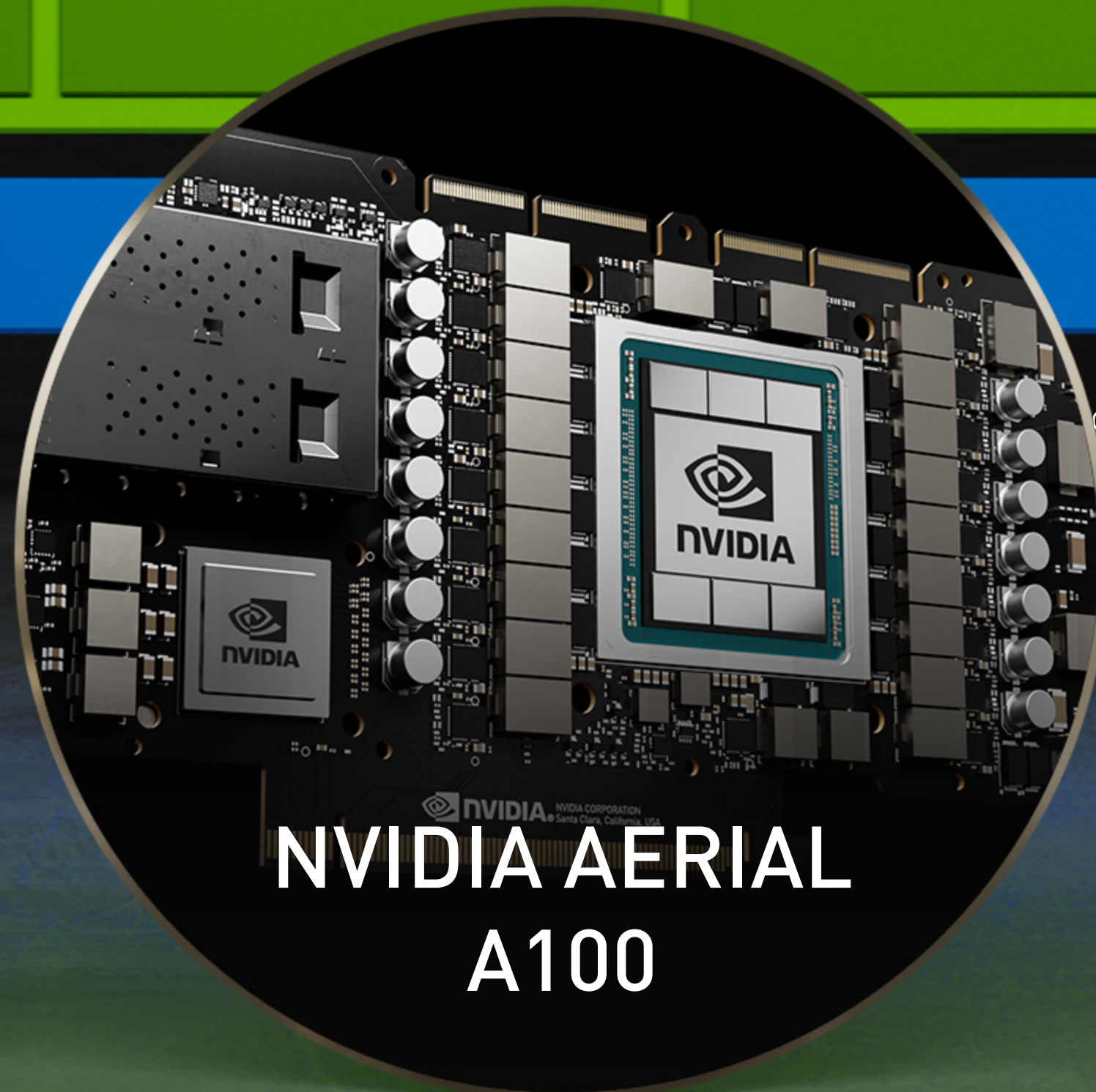
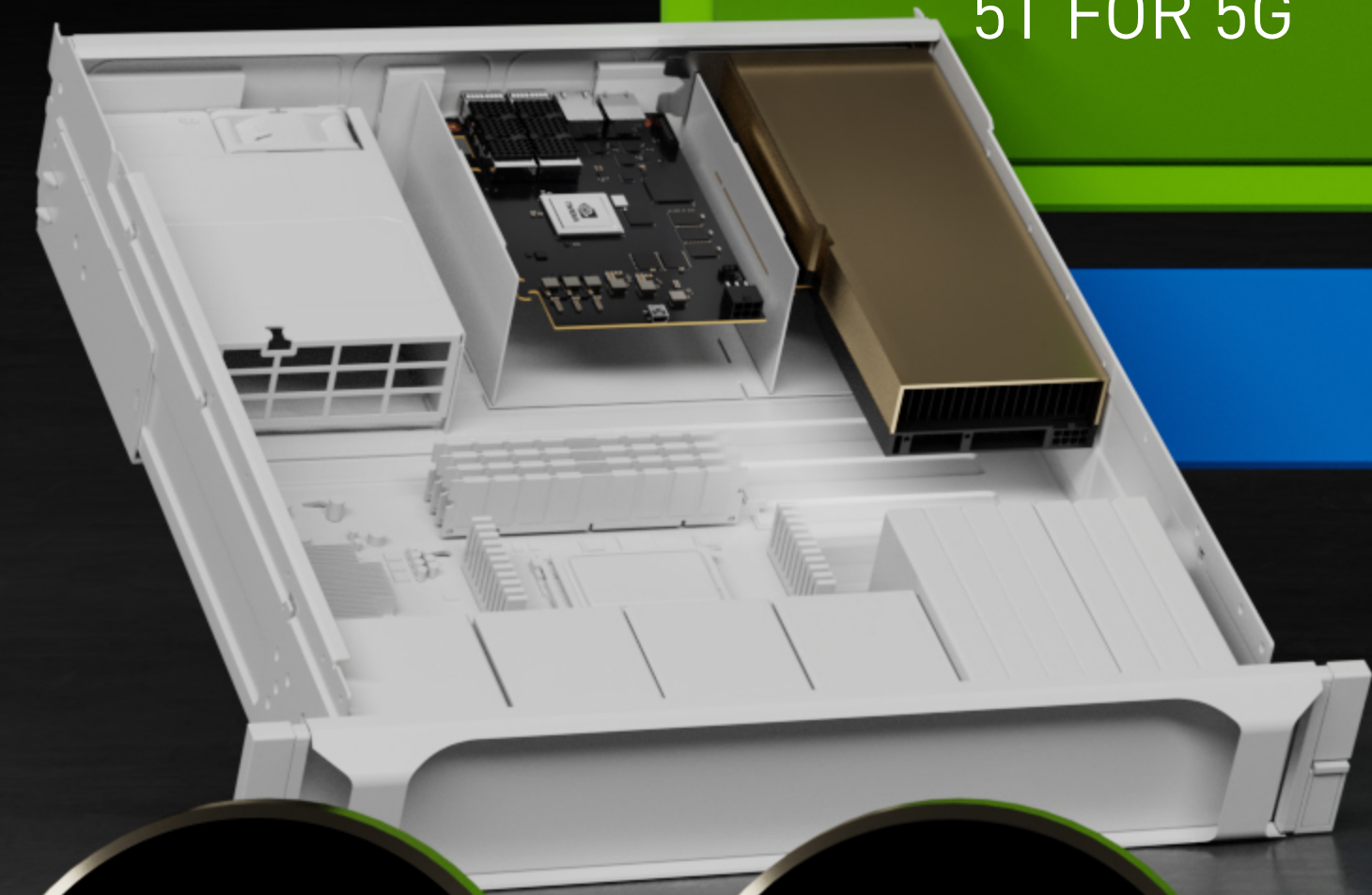
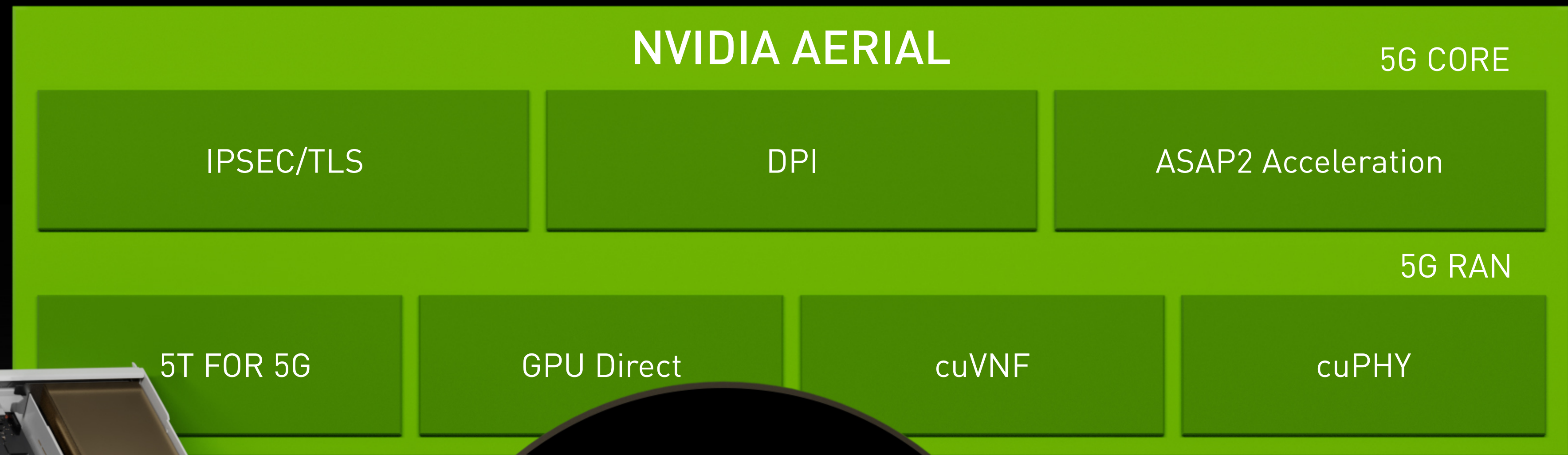


ANNOUNCING NVIDIA EGX ENTERPRISE PLATFORM



ANNOUNCING NVIDIA AERIAL A100

AI-on-5G



ANNOUNCING
GOOGLE CLOUD AND NVIDIA
PARTNER TO DELIVER AI-ON-5G

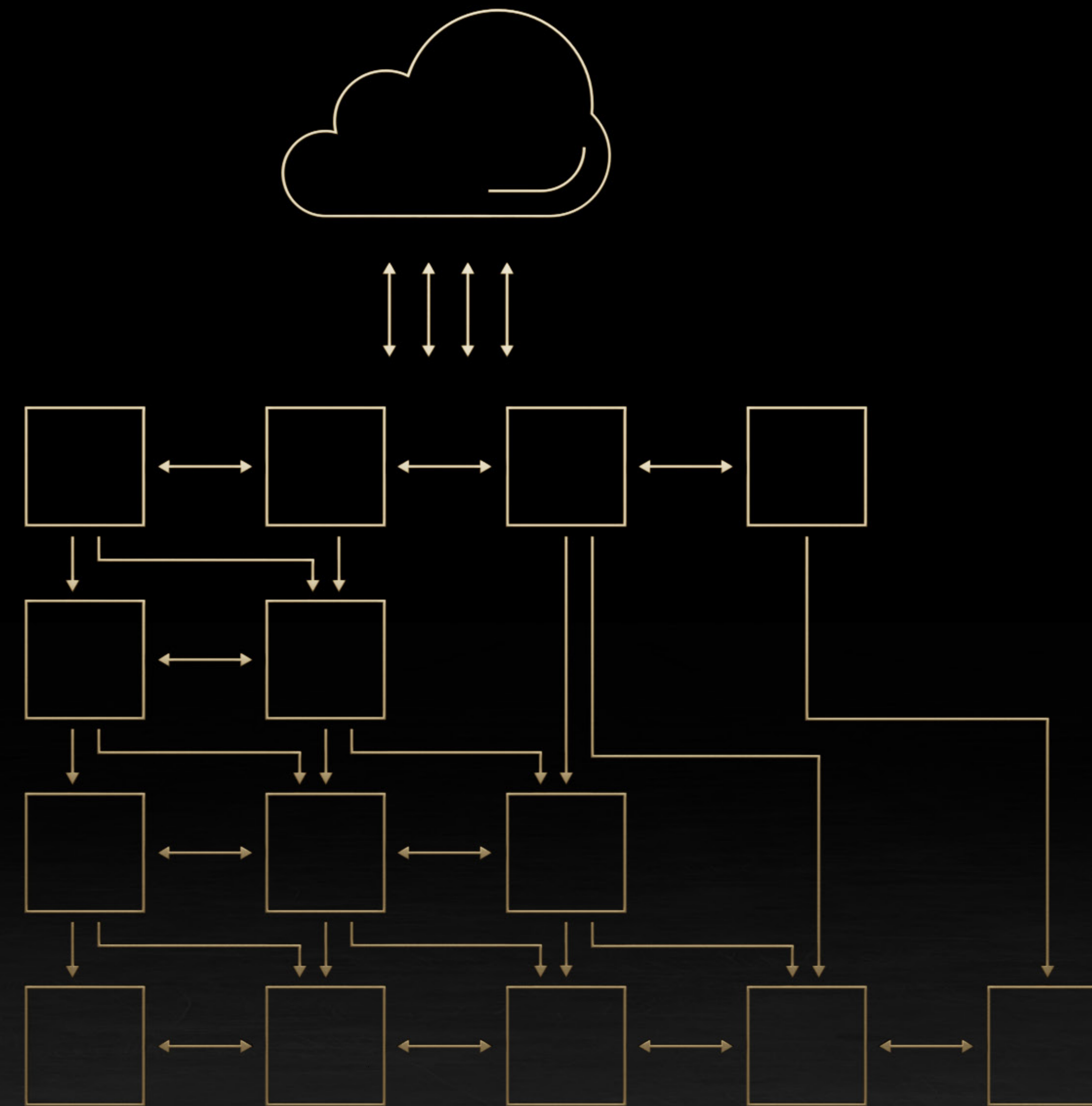
Anthos-Enabled 5G Edge to Deliver Low-Latency,
Secure and Mission-Critical Edge AI Applications

Enables the Rapid Delivery of New Services and
Applications at the 5G Edge

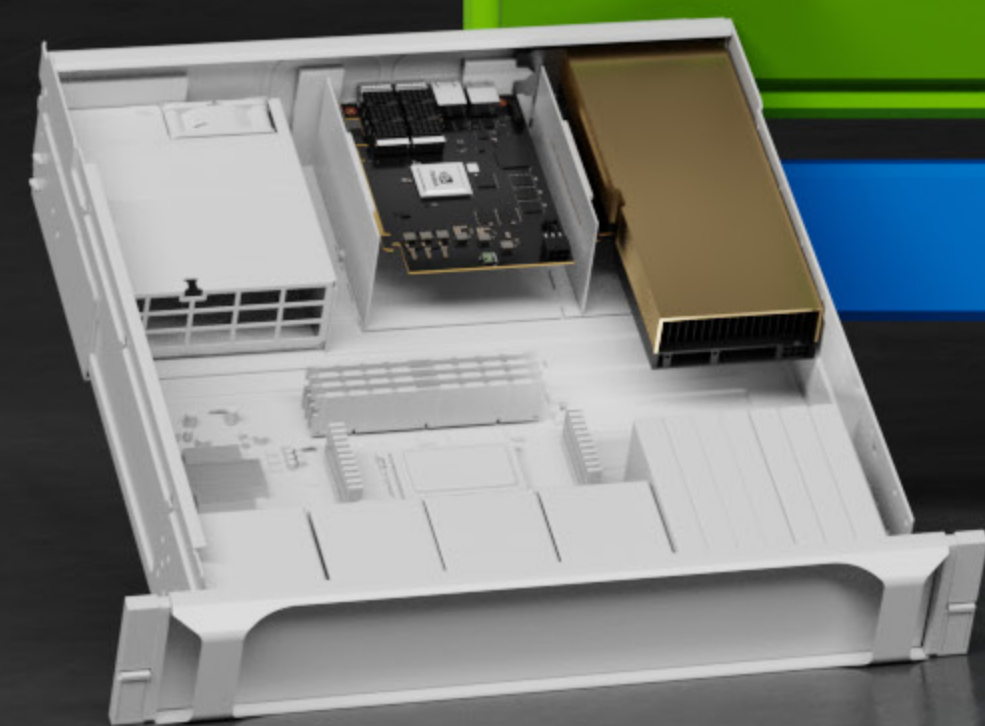
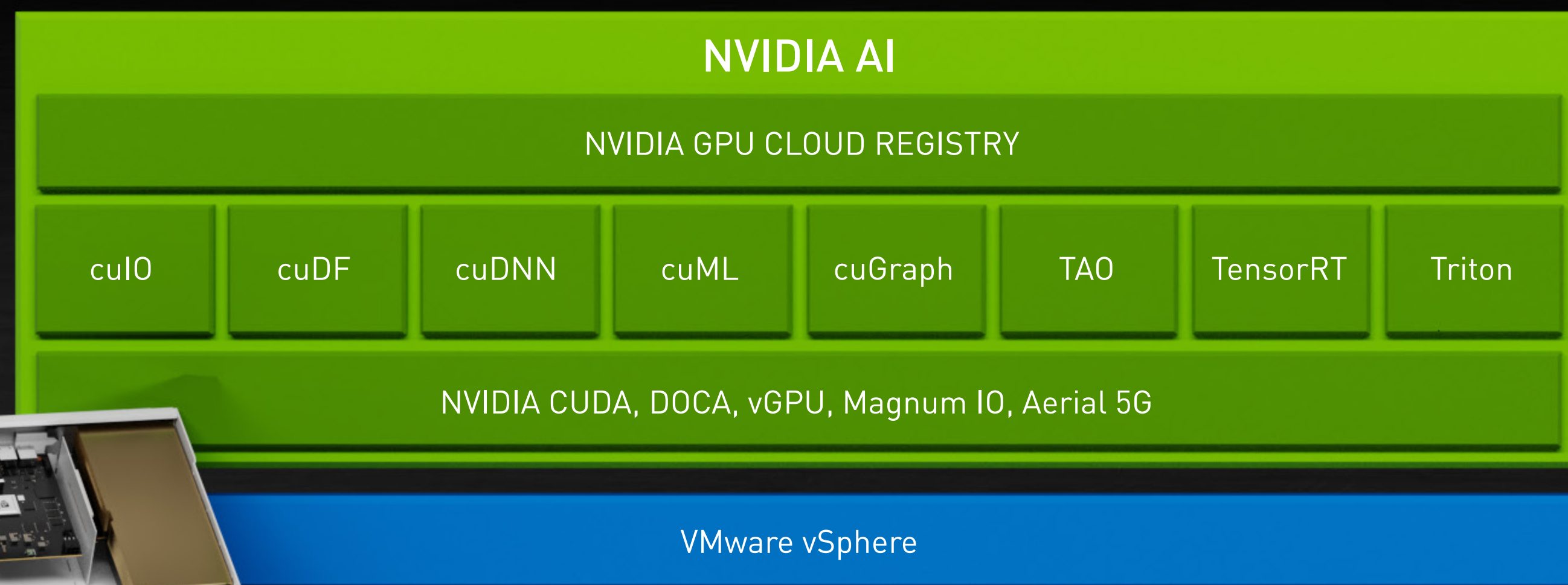
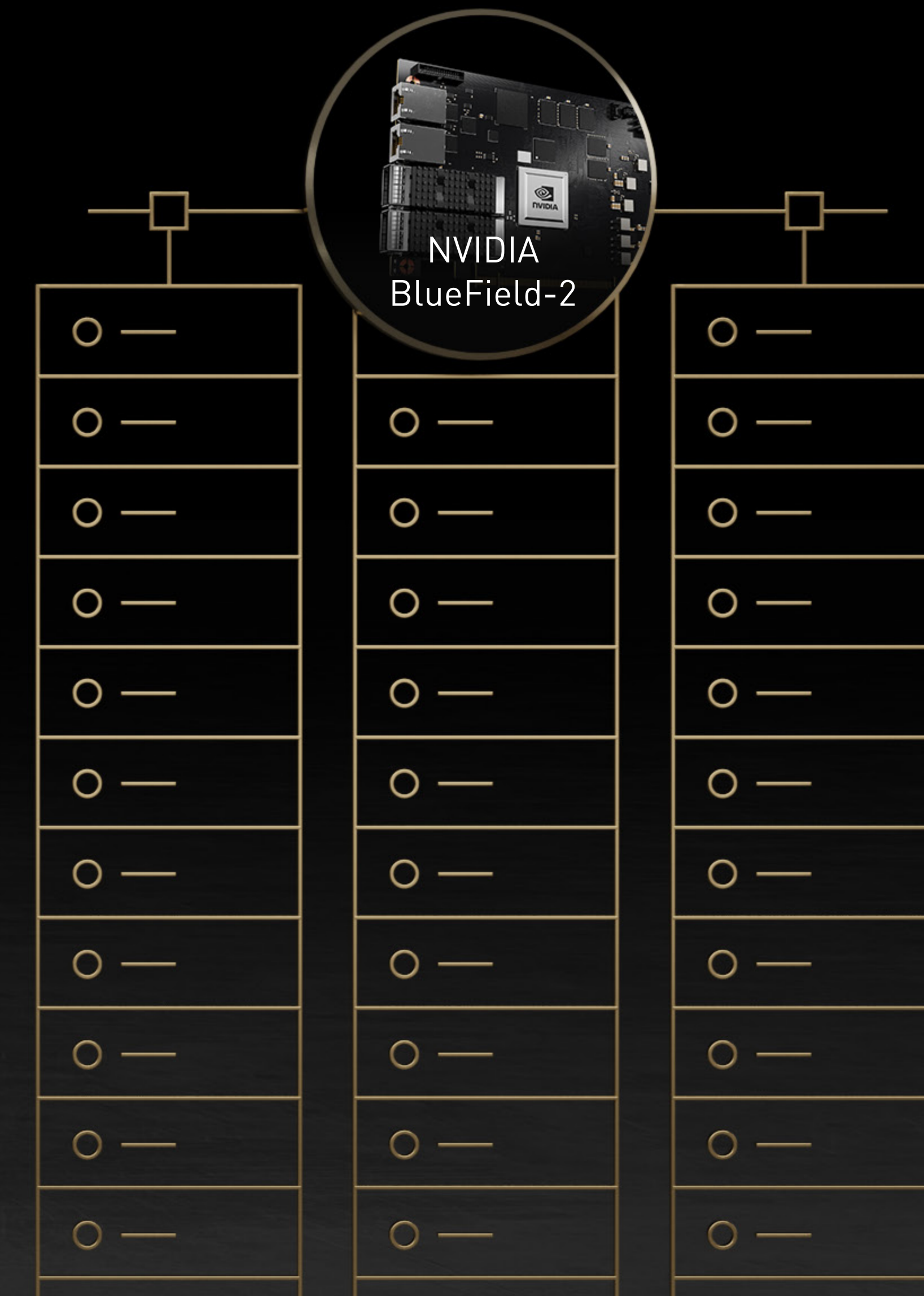
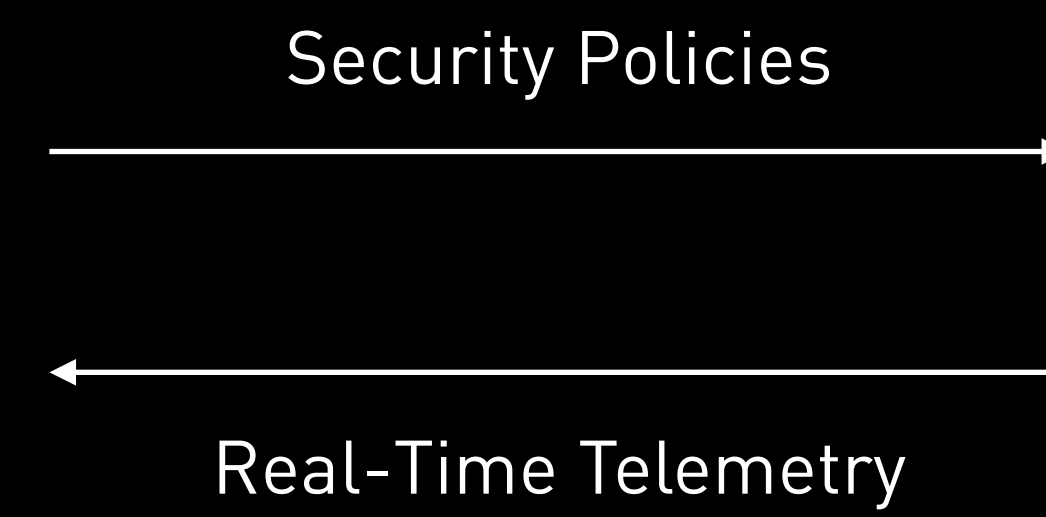
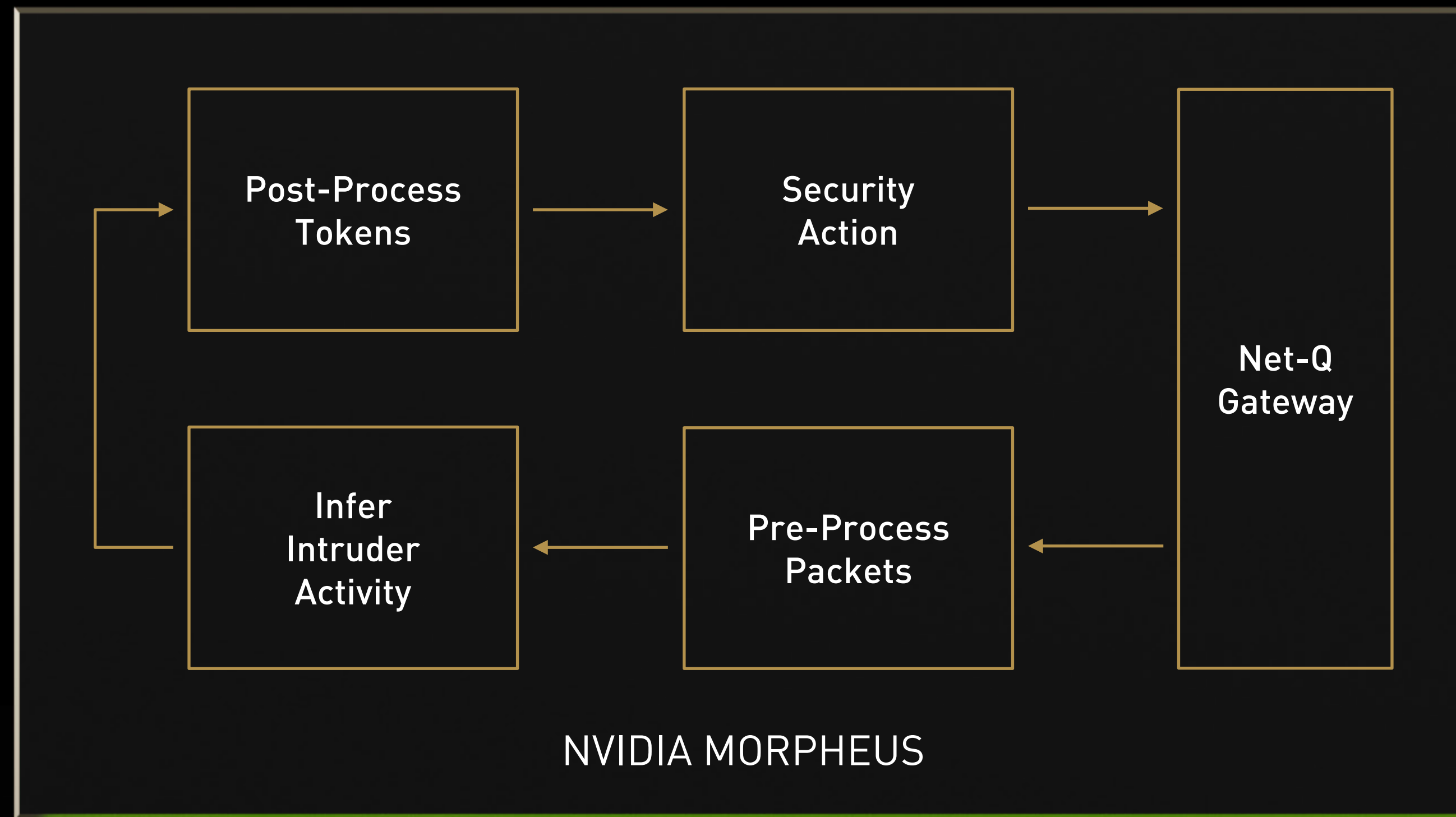
Provides a Consistent Platform for Application
Deployments from Cloud to the Edge

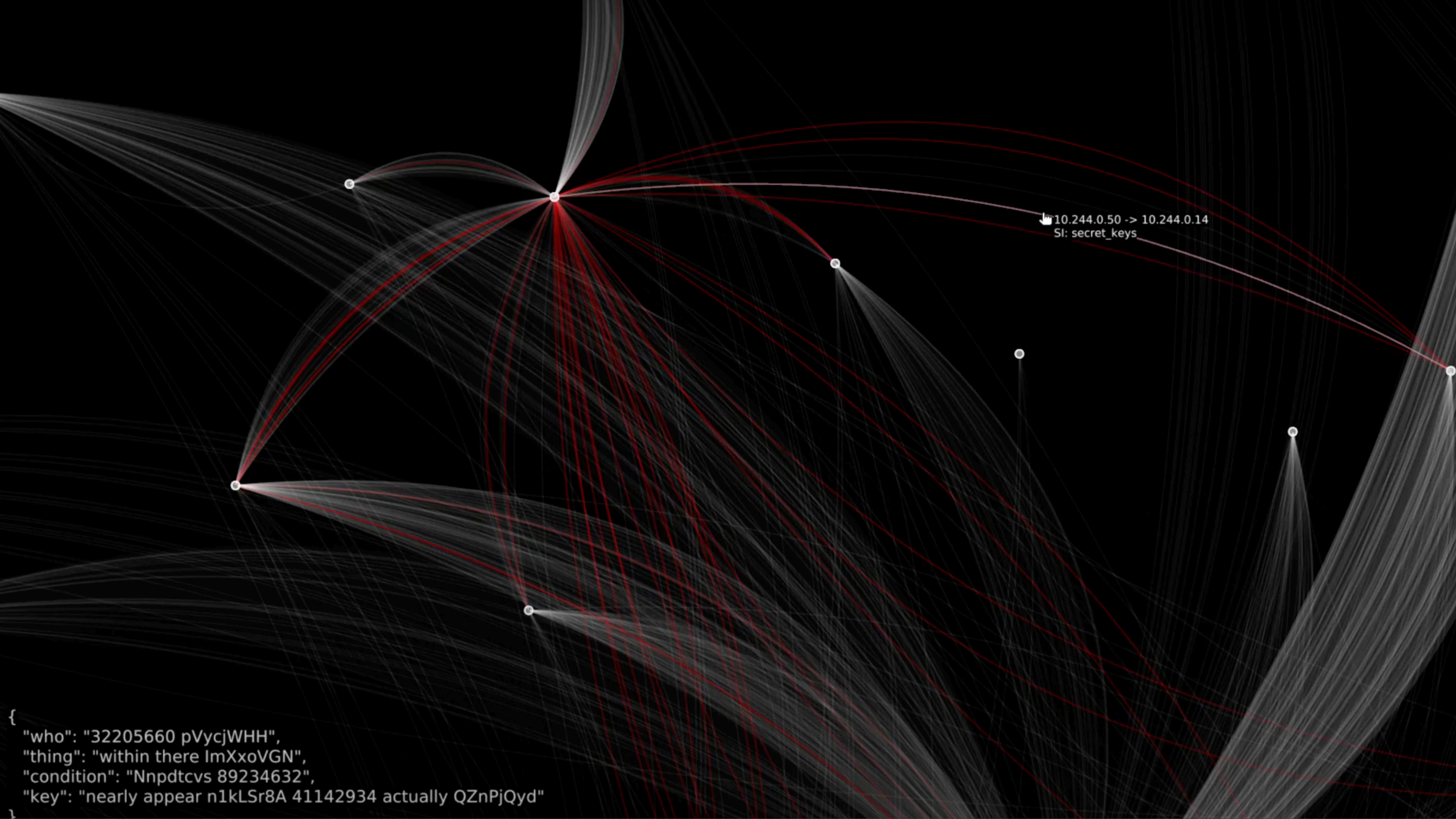
Google Cloud

EVERY USER AND WORKLOAD OF A DATA CENTER IS A SECURITY THREAT



ANNOUNCING NVIDIA MORPHEUS





10.244.0.50 -> 10.244.0.14
SI: secret_keys

```
{  
  "who": "32205660 pVycjWHH",  
  "thing": "within there ImXxoVGN",  
  "condition": "Nnpdtevs 89234632",  
  "key": "nearly appear n1kLSr8A 41142934 actually QZnPjQyd"  
}
```


NVIDIA EGX ENTERPRISE ECOSYSTEM

INFRASTRUCTURE

CANONICAL

Google Cloud

Red Hat

vmware®

INDUSTRIAL EDGE

DATA MONSTERS

DEMATIC

everseen

kineticvision

OSARO

DATA ANALYTICS & MACHINE LEARNING

CLOUDERA

splunk>

ALLUXIO

kinetica

omni·sci

plotly

DIGITAL TWIN

ALTAIR

Ansys

Bentley®
Advancing Infrastructure

CATIA

NETALLIED
SYSTEMS

ptc

SIMULIA

CYBER SECURITY

Check Point®
SOFTWARE TECHNOLOGIES LTD

CLOUDFLARE®

FORTINET®

Guardicore

paloalto®
NETWORKS

5G

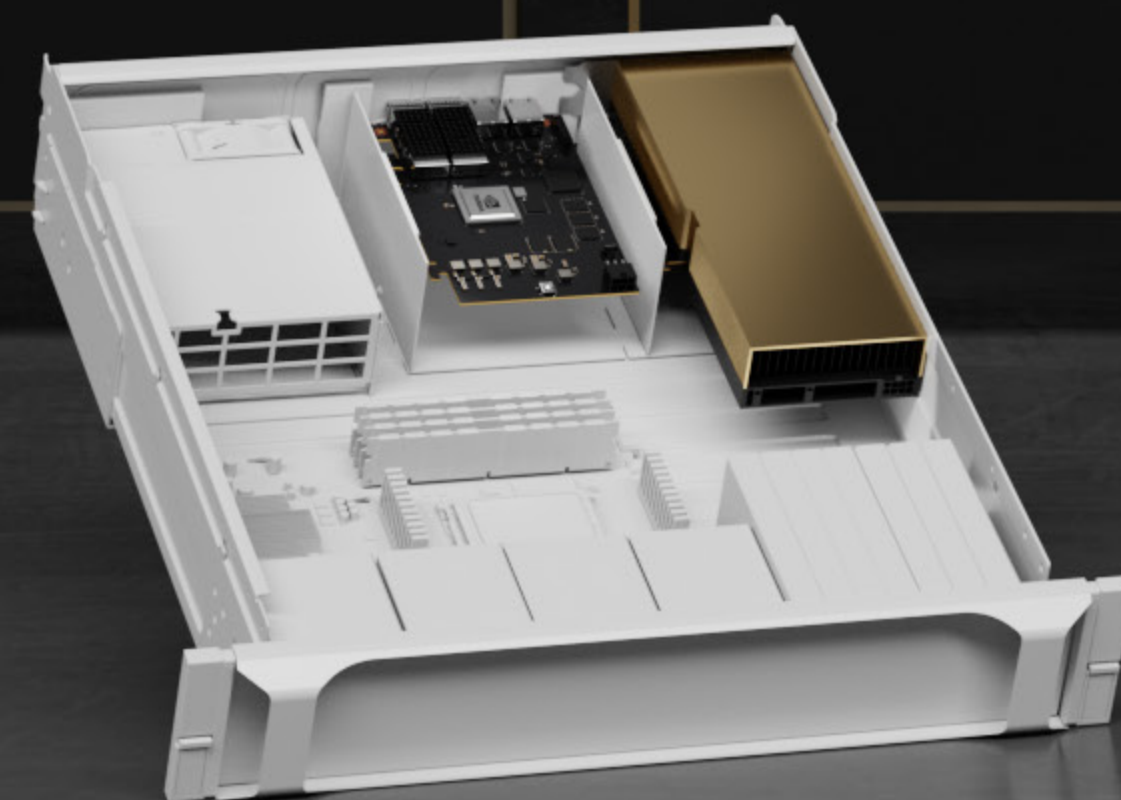
Capgemini

ERICSSON

FUJITSU

MAVENIR

Radisys



NVIDIA
CERTIFIED

AtoS

hp

DELL Technologies

inspur

FUJITSU

Lenovo

GIGABYTE

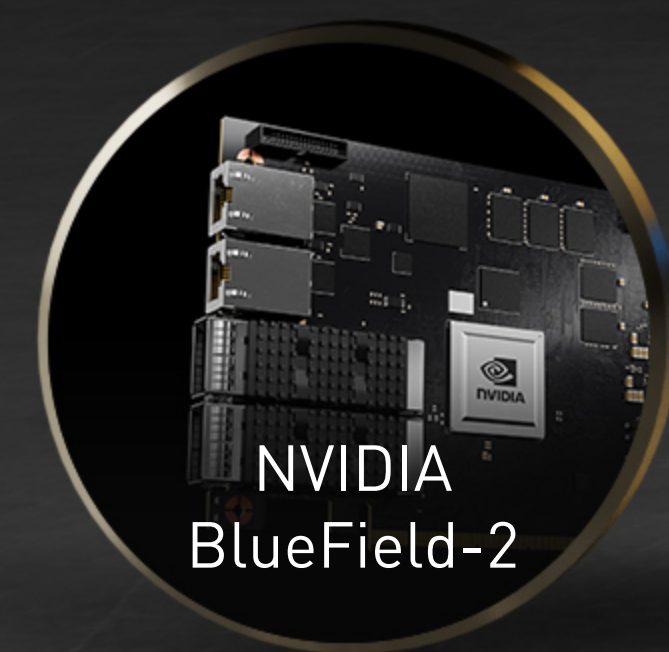
QCT

HBC

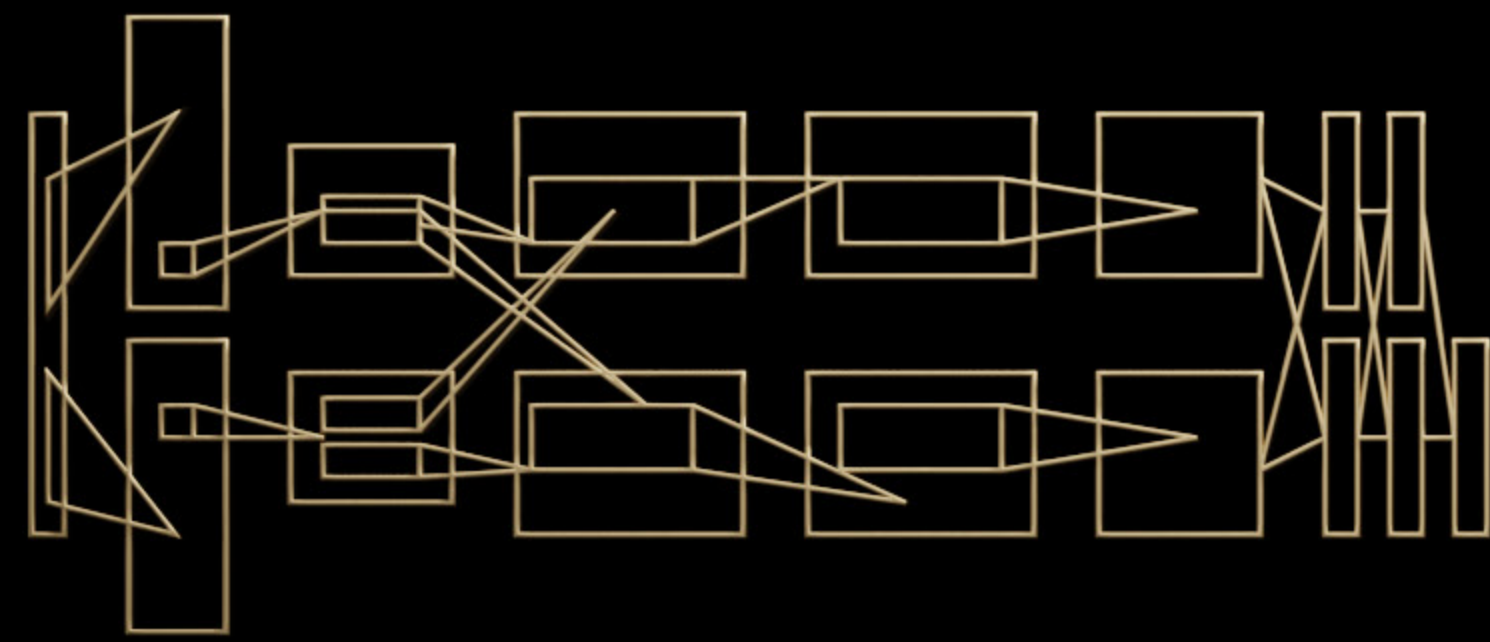
SUPERMICRO

ANNOUNCING NVIDIA EGX ENTERPRISE PLATFORM

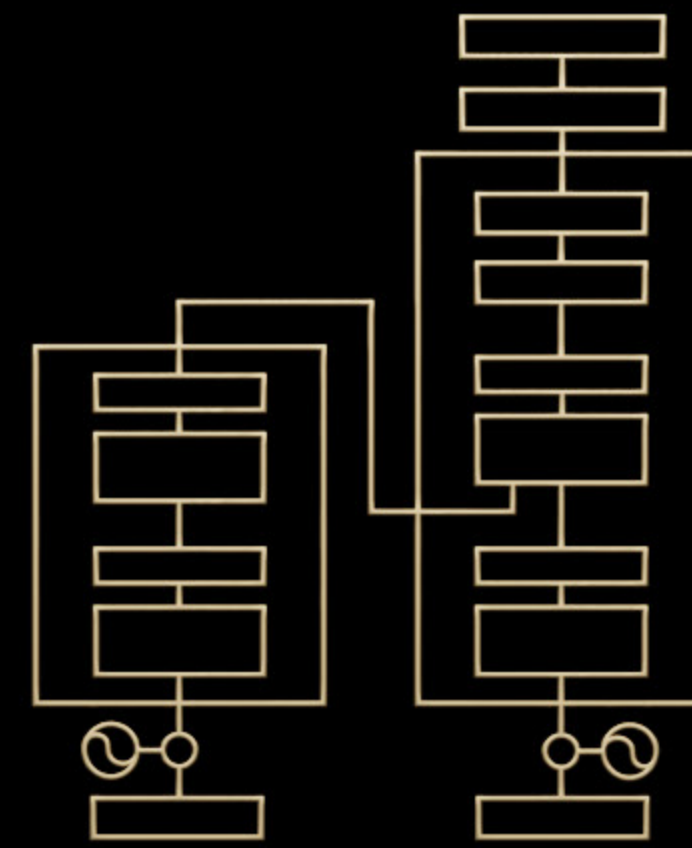
Enterprise-Ready Suite of AI and Data Science Software



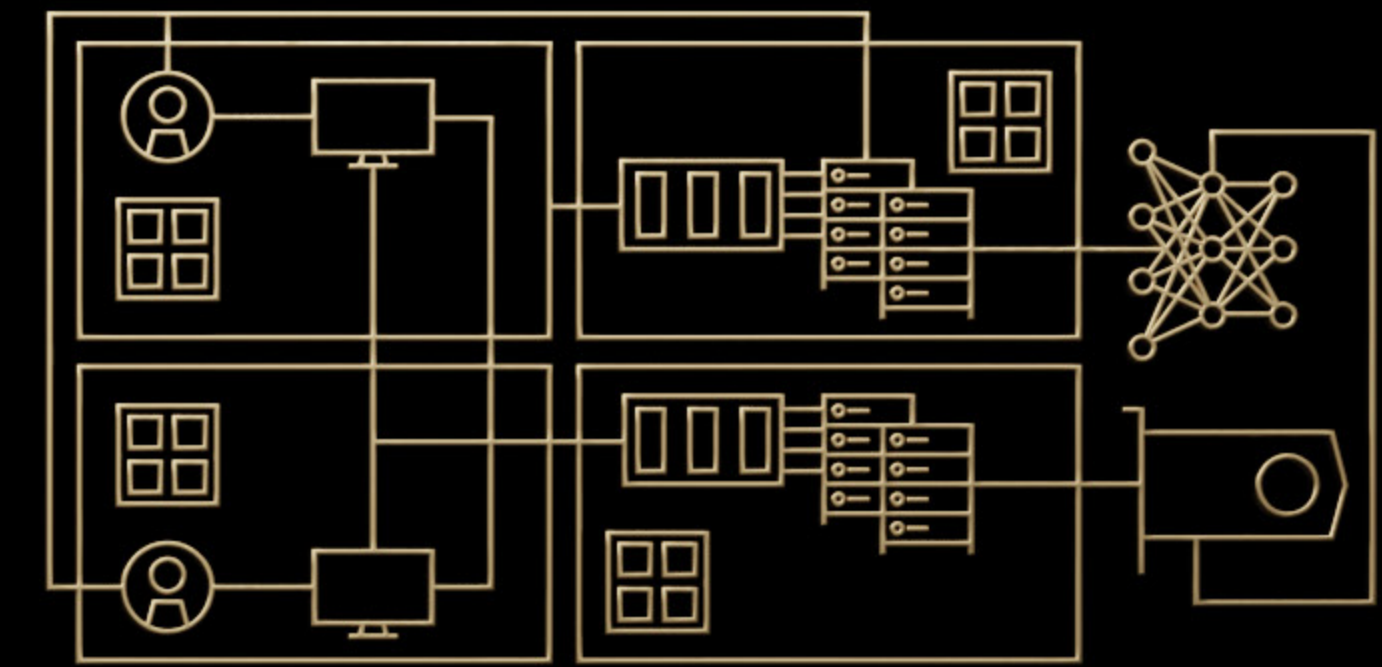
Convolutional Network (CNN)



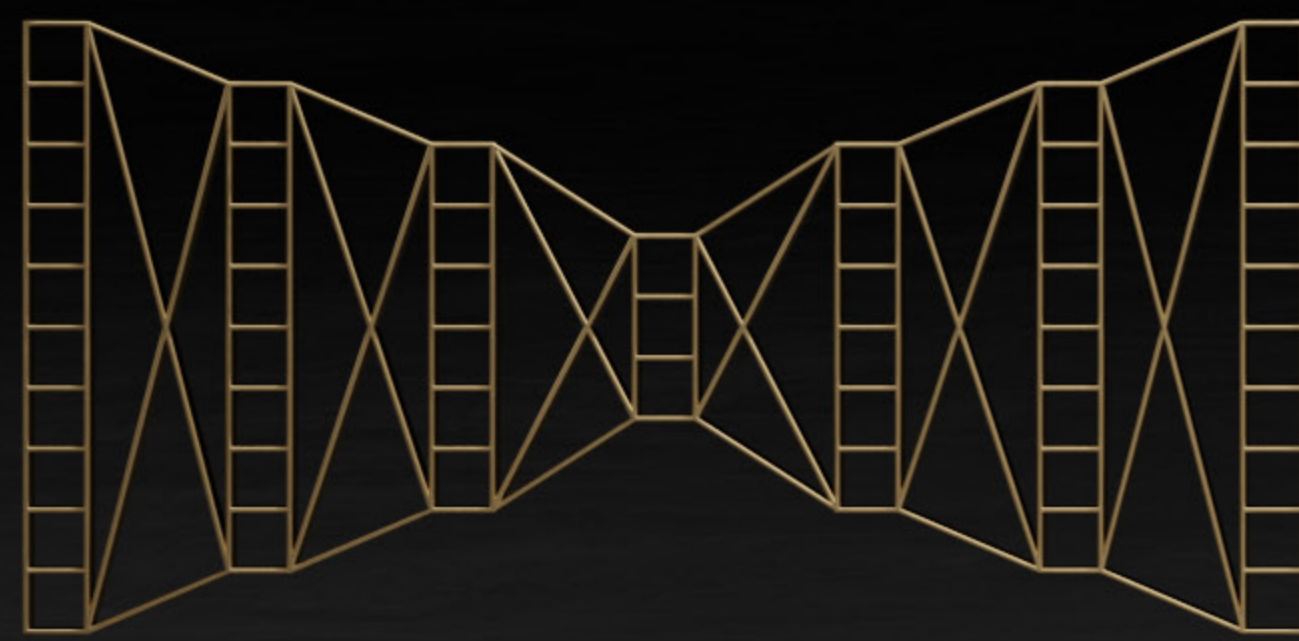
Transformer



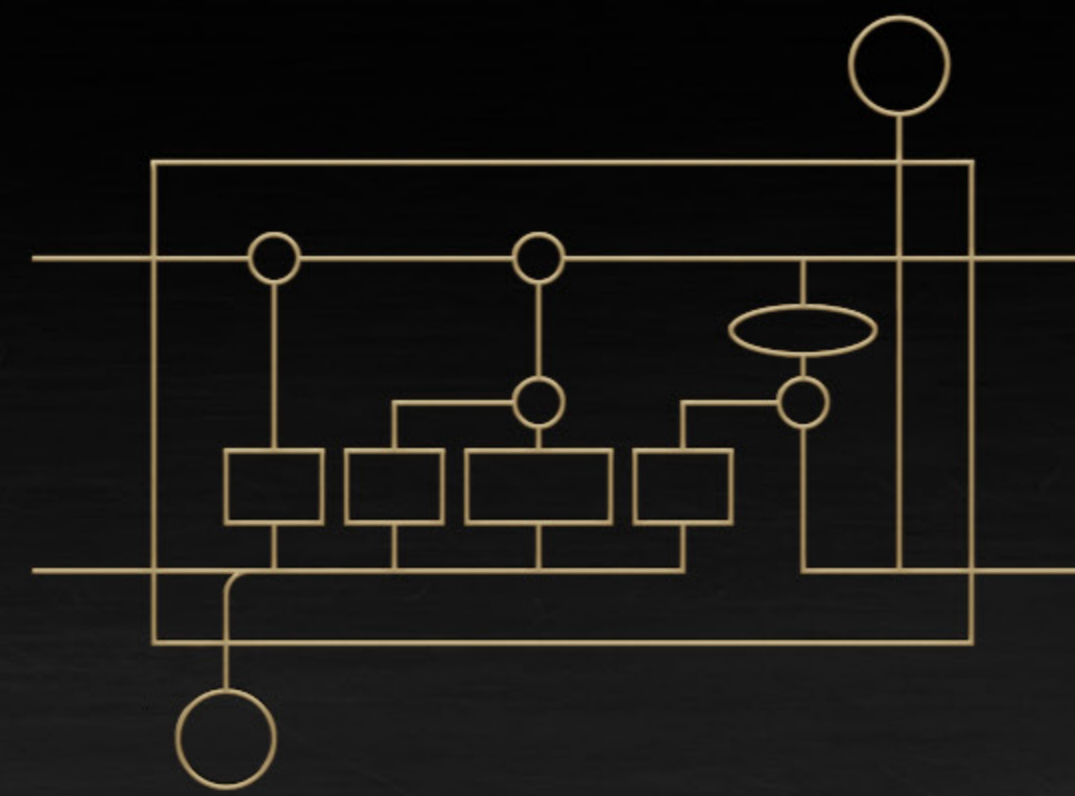
Reinforcement Learning



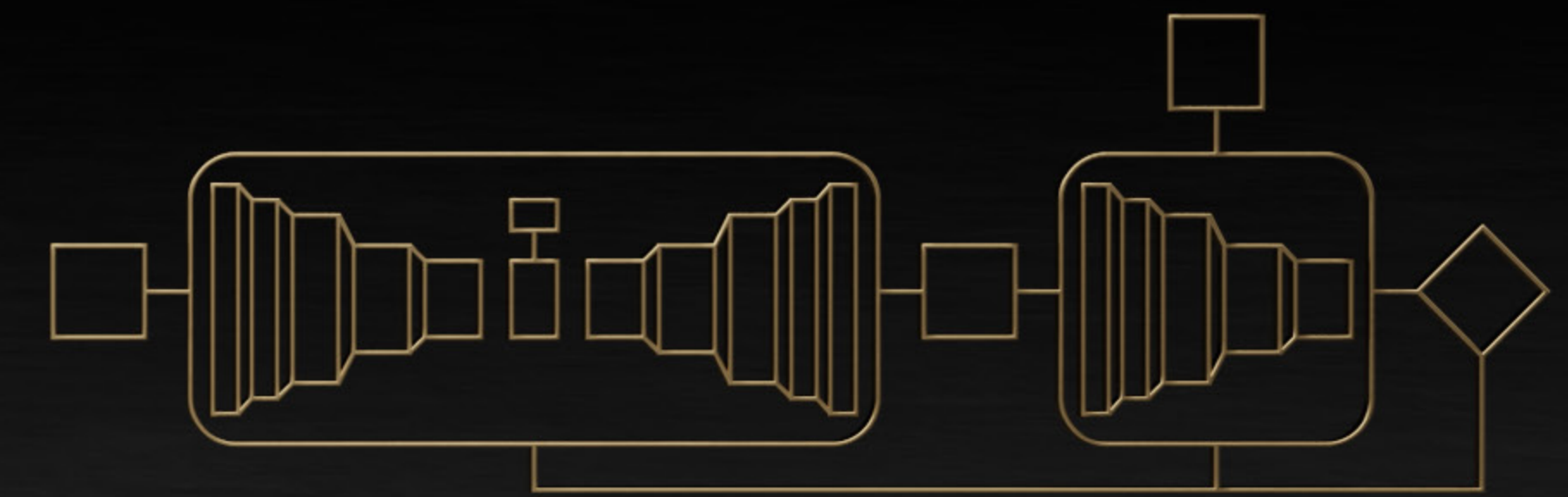
Autoencoders



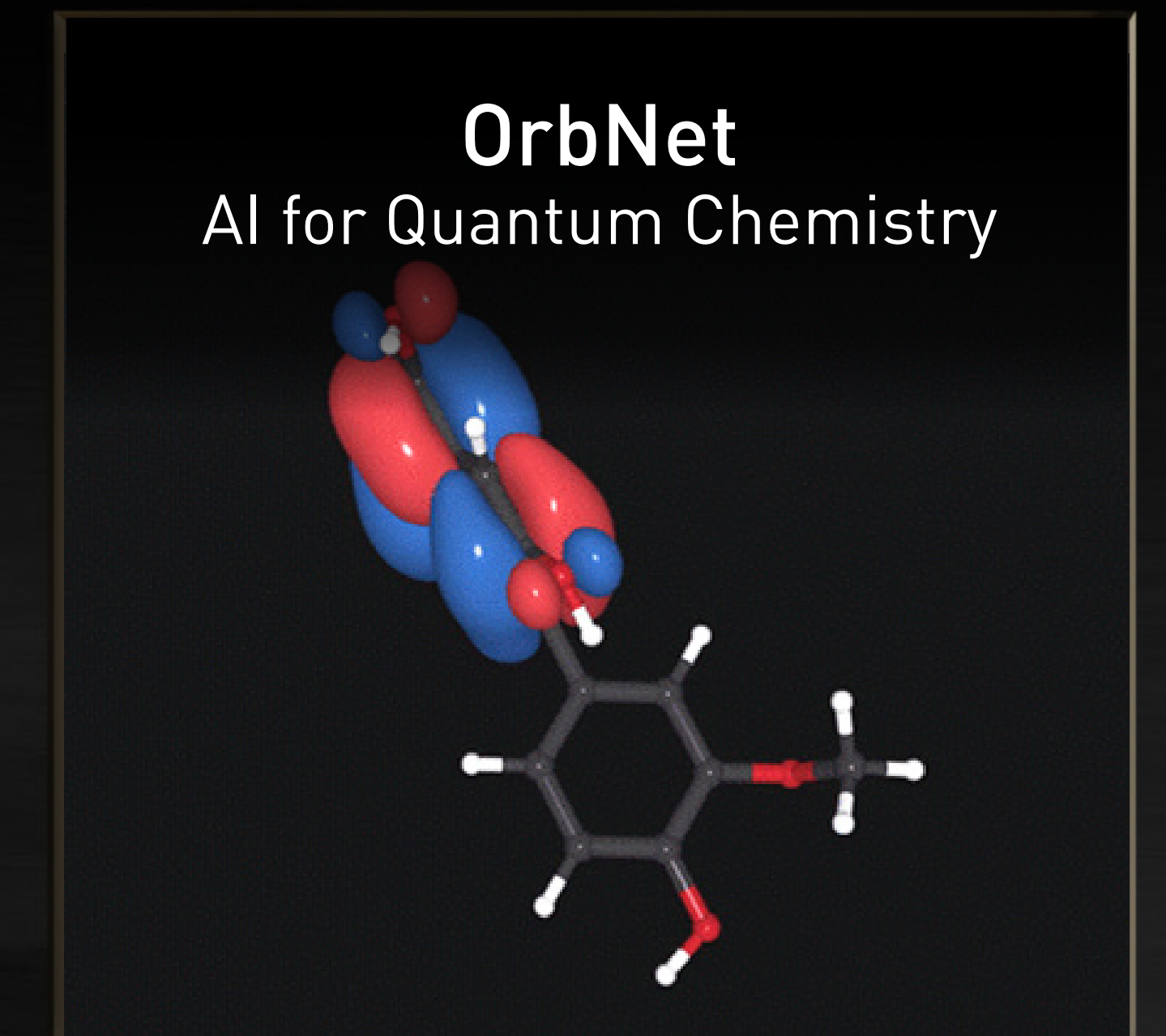
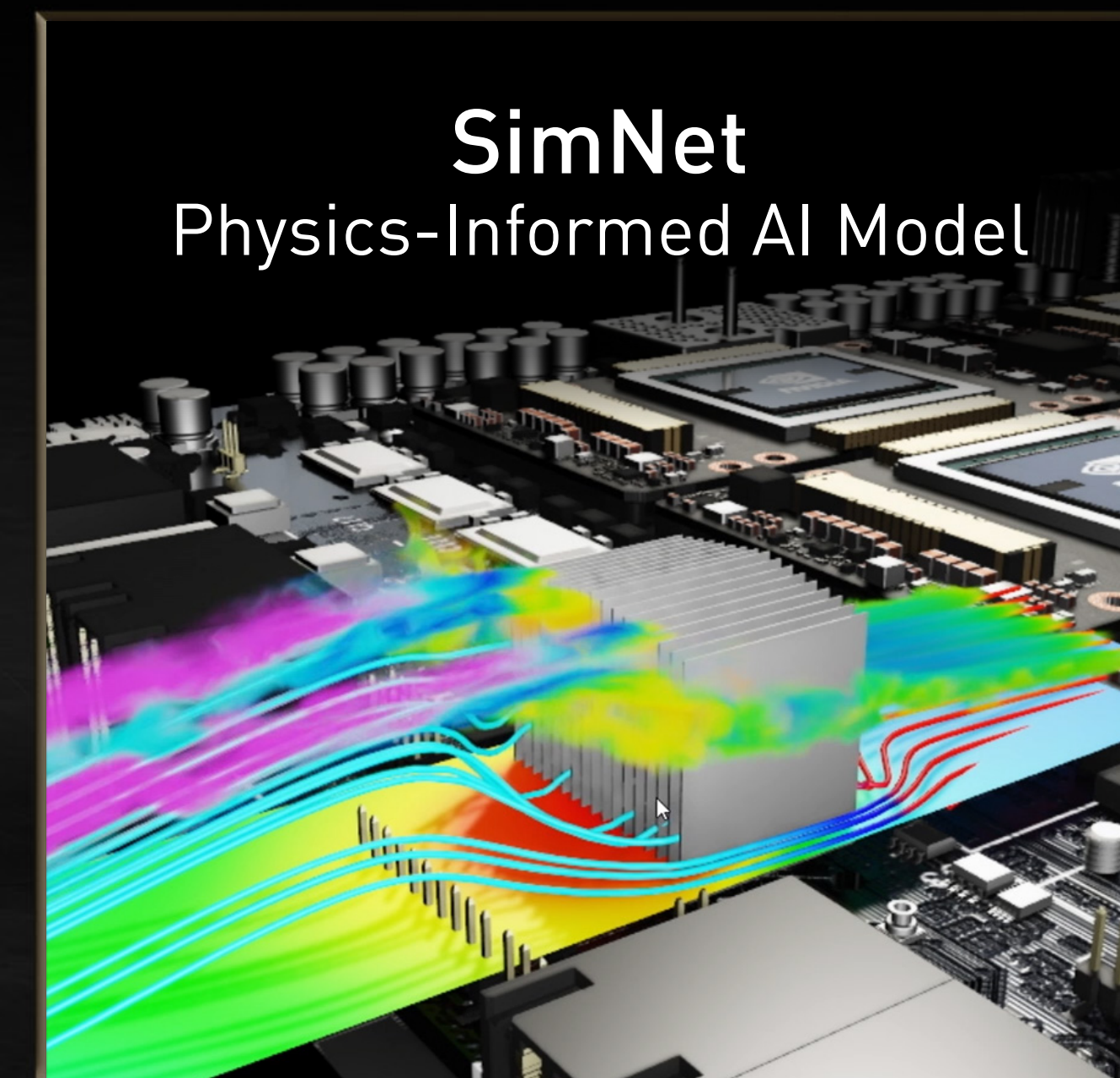
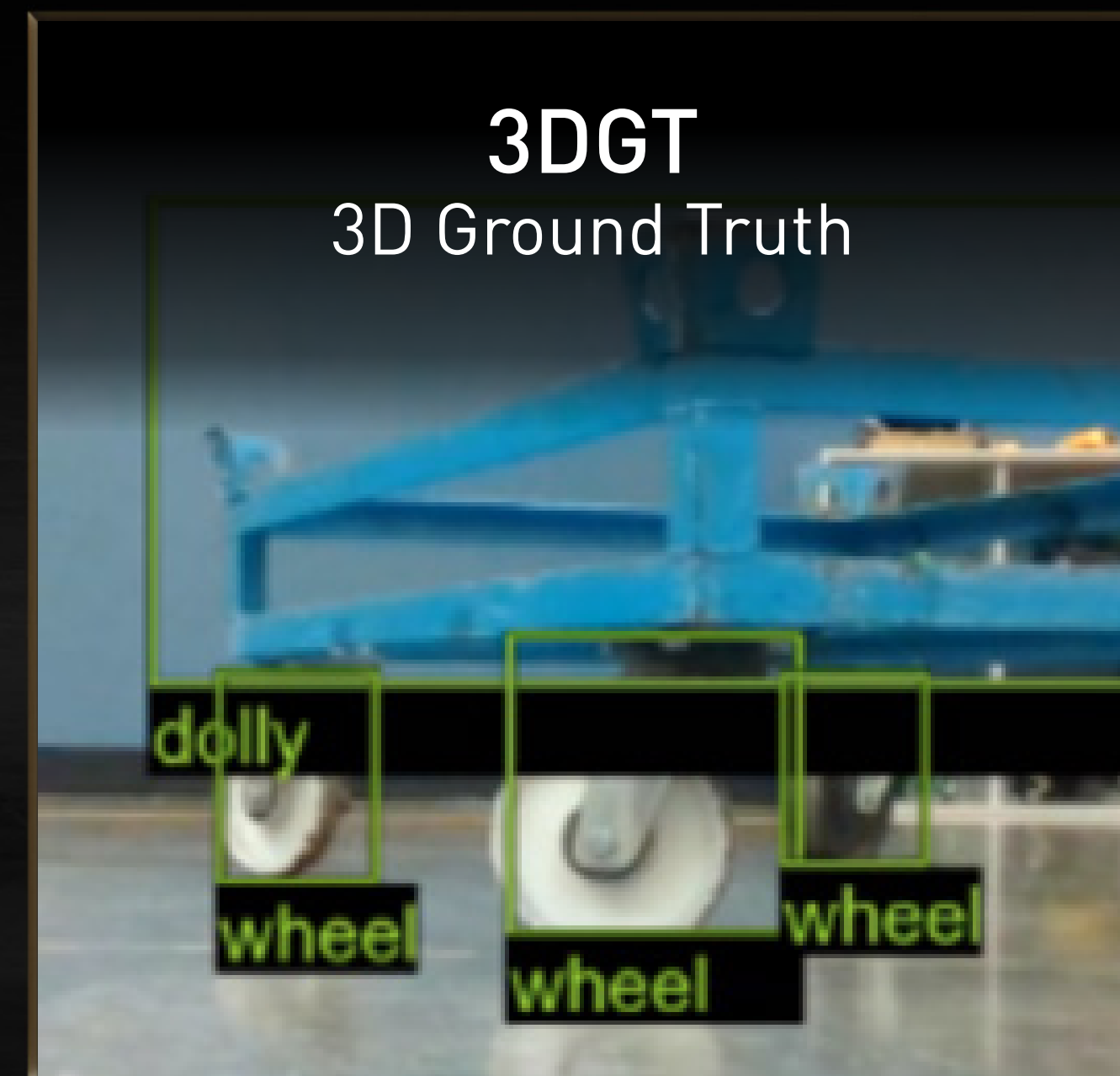
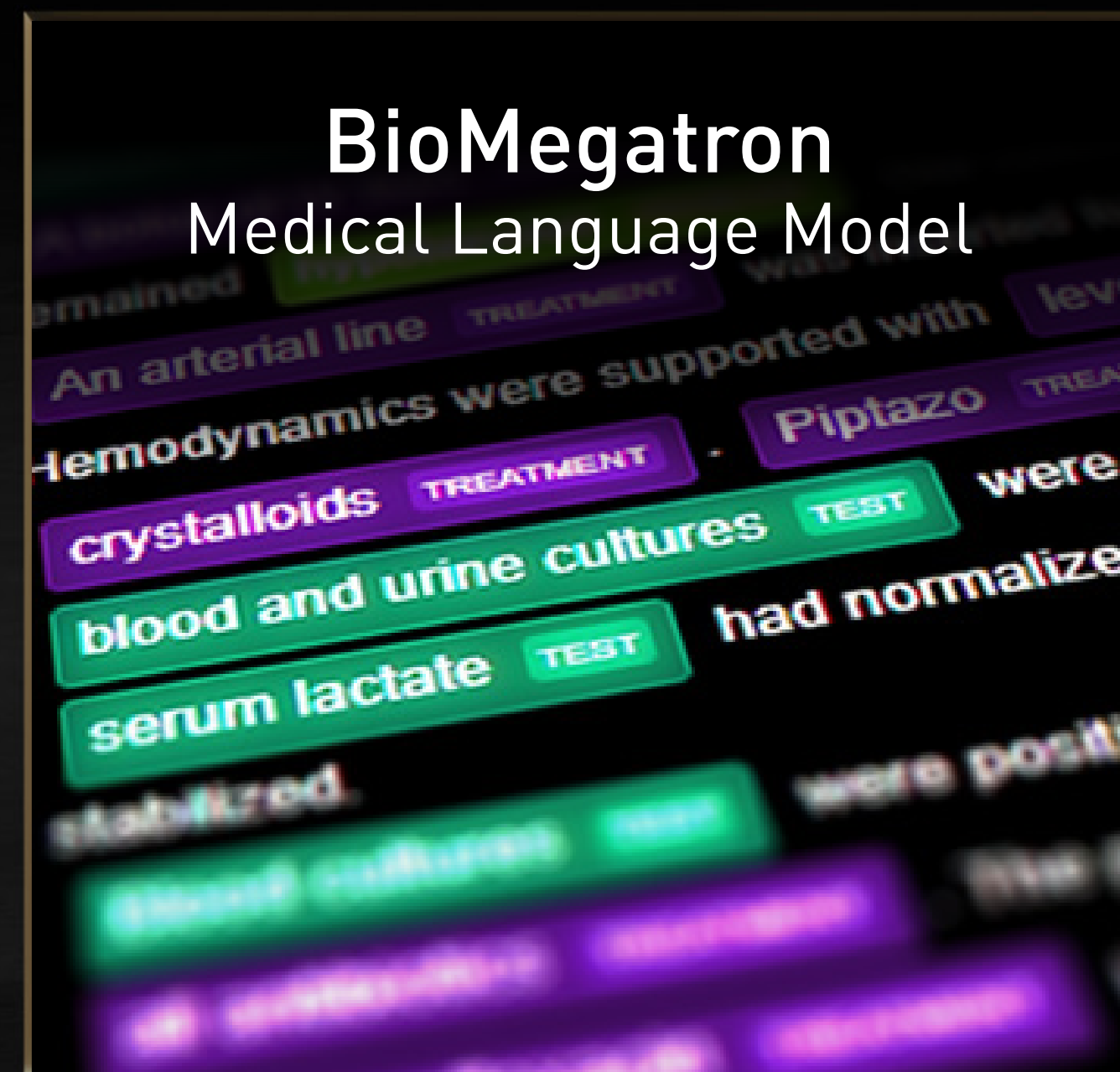
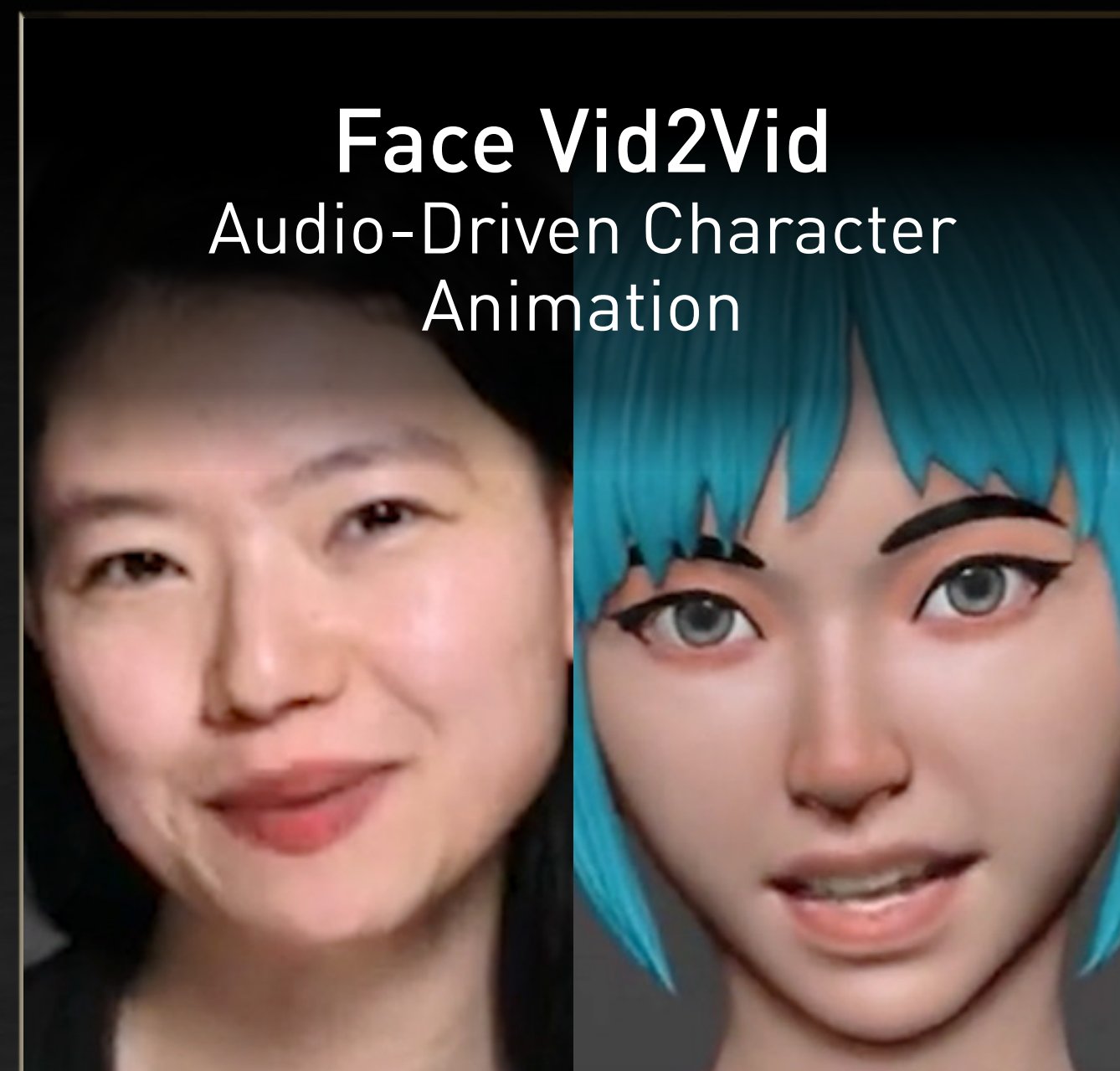
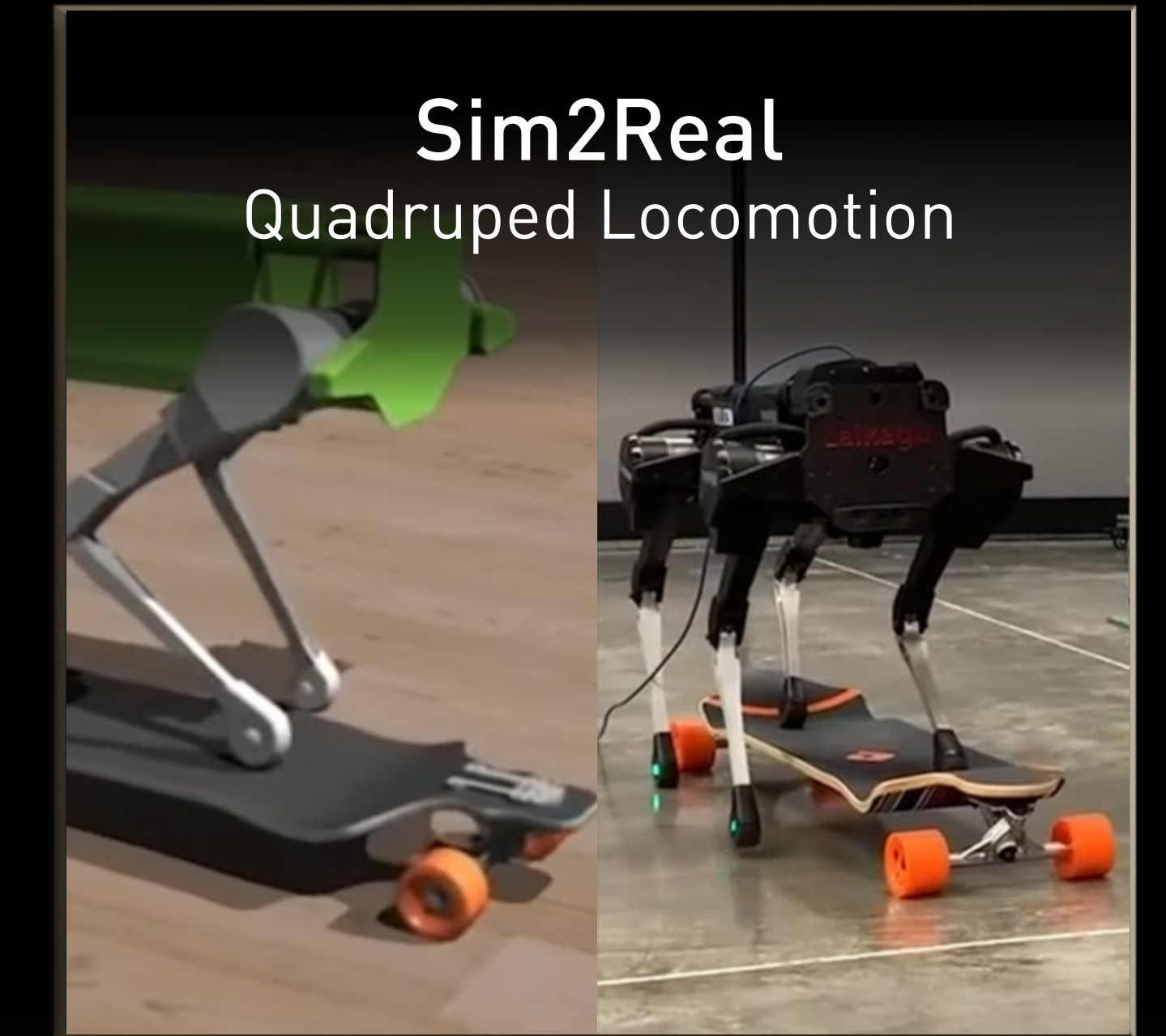
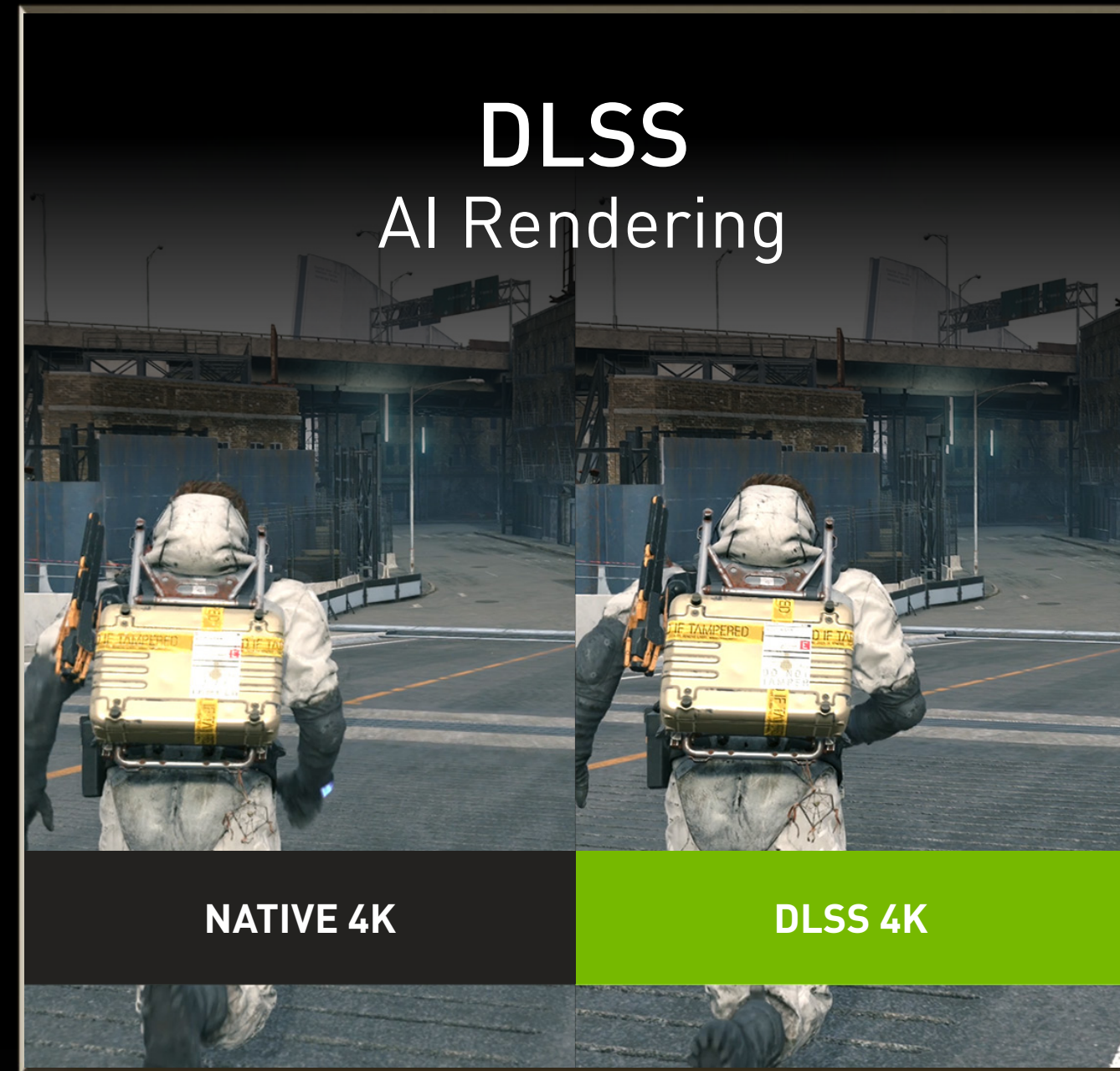
Long Short-Term Memory (LSTM)



Generative Adversarial Network (GAN)



NVIDIA'S AI



NGC PRE-TRAINED MODELS

Production-Quality AI Models

Trained by Experts for Enterprise Deployment

Credentials to Find Models You Trust

Continuously Updated to be State-of-the-Art

Adapt with NVIDIA TAO and Orchestrate
with NVIDIA Fleet Command

Reference AI Code Samples to Ease
Application Development

The screenshot shows the NVIDIA NGC Catalog interface for the 'GAZE ESTIMATION' model. The page is divided into a left sidebar and a main content area. The sidebar contains a green header with an eye icon and the text 'GAZE ESTIMATION'. Below this, it lists model details: Application (Gaze detection for a person - point of regard (X, Y, Z) and gaze vector (theta and phi)), Popularity (a bar chart), Domain (Computer Vision), Usage (Unrestricted), License (TLT Licence), Training Dataset (Proprietary dataset with more than 220k images), and Performance (T4 (1698 FPS), Xavier (704 FPS), NX (510 FPS)). At the bottom of the sidebar are tags for 'CV', 'Computer Vision', 'DL', 'Deep Learning', 'Gaze', and 'TLT', along with an 'Expand Credentials' button. The main content area features a 'Download' button in the top right. It includes sections for 'Model Overview', 'Model Architecture', and 'Training Algorithm'. The 'Model Overview' section explains that the model detects a person's eye gaze point of regard (X, Y, Z) and gaze vector (theta and phi). The 'Model Architecture' section describes GazeNet as a multi-input and multi-branch network. The 'Training Algorithm' section states that the training optimizes the network to minimize the root mean square error. Below the text is an image of a woman in a car with gaze estimation lines overlaid on her eyes. Further down, there are sections for 'How to use this model', 'Input', and 'Output'. The 'Input' section specifies that GazeNet is a multi-input network that takes a face crop image, left eye crop image, right crop image, and facegrid. The 'Output' section states that the model outputs a 3D point of regard (X, Y, Z) and gaze vector (theta and phi).

GAZE ESTIMATION

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

Popularity

Domain
Computer Vision

Usage
Unrestricted

License
[TLT Licence](#)

Training Dataset
Proprietary dataset with more than 220k images.

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

CV Computer Vision DL Deep Learning
Gaze TLT


[Expand Credentials](#)

Download

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

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

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



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

Input
GazeNet is a multi-input network, which takes in face crop image, left eye crop image, right crop image, and facegrid.
Face Image which is gray scale. 224 x 224 x 1

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

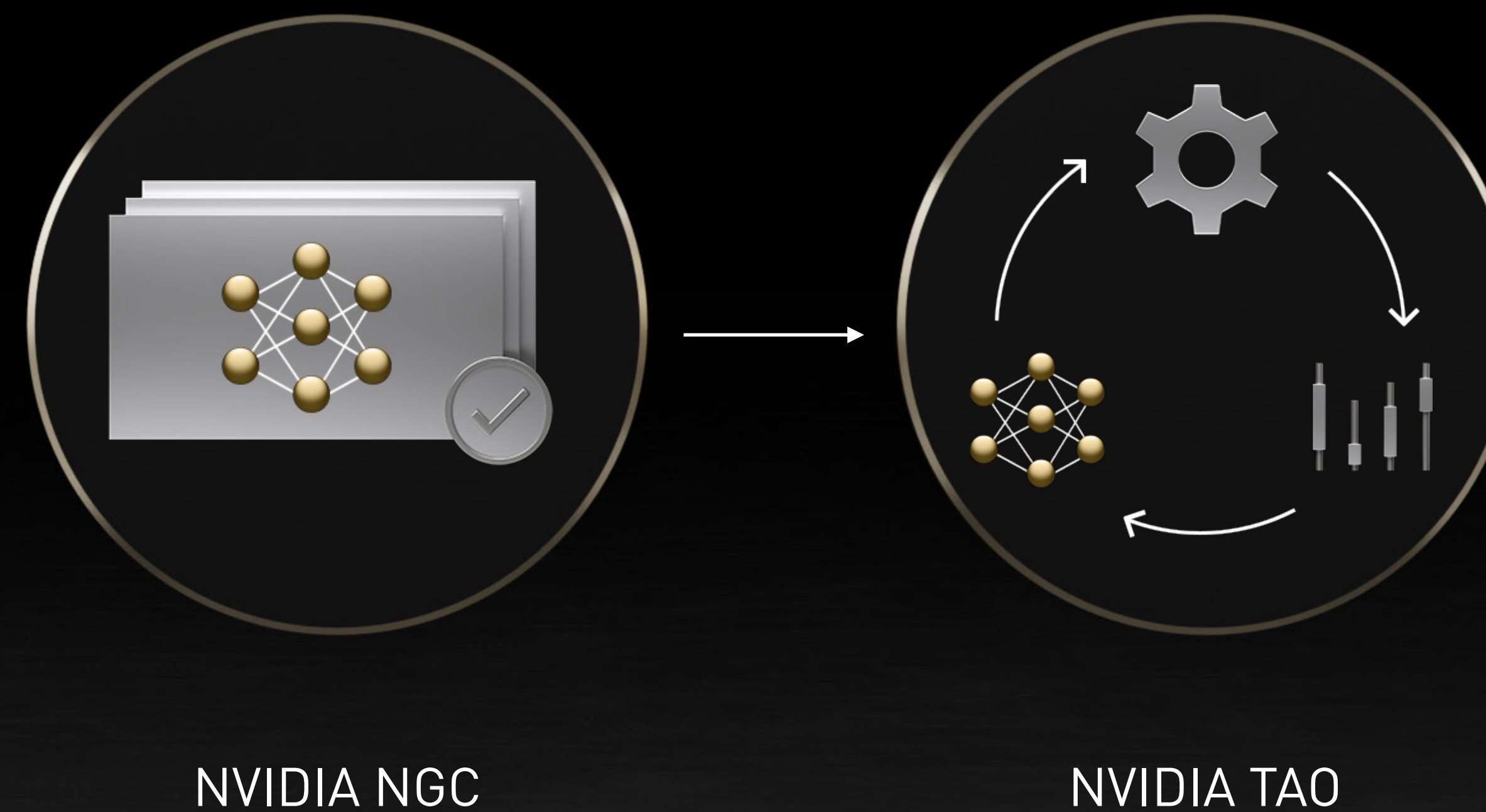
ANNOUNCING NVIDIA TAO FRAMEWORK

Train | Adapt | Optimize

Customize Pre-Trained Models for
Domain-Specific Applications

Federated Learning Enables Model Training
Collaboration while Protecting Data Privacy

Produce State-of-the-Art Models in Hours



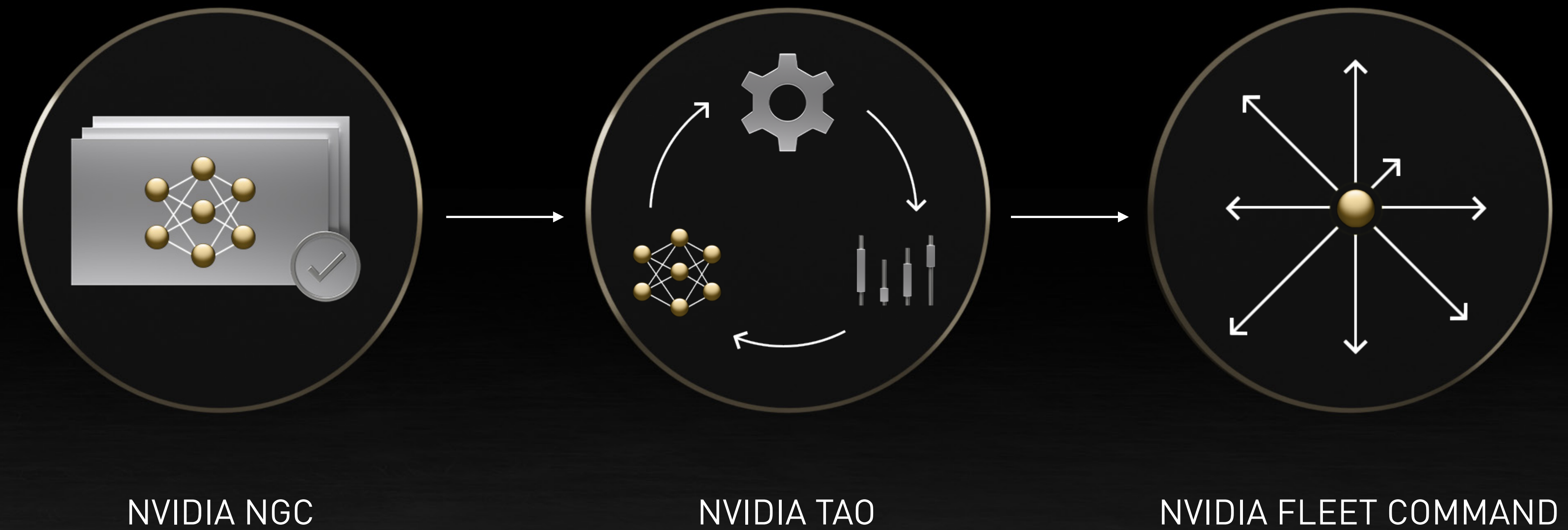
ANNOUNCING NVIDIA FLEET COMMAND

Securely Orchestrate AI Fleet at the
Edge of the Network

Control and Manage Millions of AI-Powered
Devices from Any Cloud

Secure from Boot, Attestation, Uplink and
Downlink, to Confidential AI Enclave

Centrally Monitor Health and Remotely Fix
Edge Systems



NGC PRE-TRAINED MODELS

Production-Quality AI Models

Trained by Experts for Enterprise Deployment

Credentials to Find Models You Trust

Continuously Updated to be State-of-the-Art

Adapt with NVIDIA TAO and Orchestrate
with NVIDIA Fleet Command

Reference AI Code Samples to Ease
Application Development

EMR CXR AI MODEL

Application
Detect likelihood that a COVID-19 patient in the emergency room will need supplemental oxygen.

Popularity
[Bar chart showing popularity]

Domain
Healthcare

Usage
For research purposes only

License
Clara License

Training Dataset
20 different FL client sites with 16,000 cases.

Performance
AUC - 24 hours (94%)
AUC - 72 hours (91%)

Tags: Federated Learning, Deep Learning, Healthcare, Medical imaging, Covid-19, Clara, 20 Federated Learning Sites

[Expand Credentials](#)

Model Overview
The ultimate goal of this model is to predict the likelihood that a person showing up in the emergency room will need supplemental oxygen, which can aid physicians in determining the appropriate level of care for patients, including ICU placement.

Model Architecture
The model uses a pre-trained ResNet34 [1] for image feature extraction together with a deep & cross network [2] to combine it with EMR features.

Training Algorithm
The models were developed within the EXAM consortium using federated learning. The chest x-ray feature extraction branch of the network was pretrained on >200,000 images from CheXpert dataset (on pneumonia vs. rest task) & fine-tuned on ~500 images from Mass Gen Brigham to predict RALE [3] score to evaluate lung edema on CXR.

Diagram:
The diagram illustrates the model's workflow. It starts with two inputs: a CXR (Chest X-ray) image and EHR Data (Blood Test, Oxygen Sat, Blood Pressure, Respiratory Rate). The CXR image is processed by ResNet 34 to generate CXR Features (represented as a heatmap). These CXR Features and the EHR Data are then fed into a 'Deep & Cross' network. The output of this network is a '24 Hr Post Chest X-Ray Oxygen Prediction', which is categorized into four levels: Room Air, Low Flow O₂, High Flow O₂, and Ventilator. The first three levels are grouped under 'Under Observation', while the Ventilator and Death categories are grouped under 'Intensive Care'.

Input
The model takes chest x-ray images and EMR data as input. The EMR data includes Blood Test, Oxygen Sat, Blood Pressure, and Respiratory Rate. The chest x-ray images are converted to PNG format with an associated manifest.csv file.

Output
The model predicts a patient's oxygen needs based on a chest x-ray and patient vitals and lab values (EMR features) by giving a risk score [0...1]. The output is categorized into four levels: Room Air, Low Flow O₂, High Flow O₂, and Ventilator. The first three levels are grouped under 'Under Observation', while the Ventilator and Death categories are grouped under 'Intensive Care'.

NVIDIA JARVIS

State-of-the-Art Conversational AI

GPU-Accelerated ASR, NLU, TTS

Interactive Performance

Customize with NVIDIA TAO

Orchestrate with NVIDIA Fleet Command

Scale-Out with NVIDIA Triton

Run in Every Cloud and at the Edge

NVIDIA NGC | CATALOG Download

NVIDIA Jarvis

Application
Accelerated framework for building multimodal conversational AI skills.

Popularity
[Bar chart showing popularity]

Domain
Conversational AI

License
Jarvis License

1.0.0-b.2 (Latest) Scan Results

Linux / x86 AA

Deep Learning Conversational AI
Inference ASR TTS Text to Speech
NLP Natural Language Processing
Automatic Speech Recognition Transcription
Dialogue Management Call Center

[Expand Credentials](#)

NVIDIA Jarvis

Jarvis is a fully accelerated application framework for building multimodal conversational AI services that use an end-to-end deep learning pipeline. Developers at enterprises can easily fine-tune state-of-art-models on their data to achieve a deeper understanding of their specific context and optimize for inference to offer end-to-end real-time services that run in less than 300 milliseconds (ms) and delivers 7x higher throughput on GPUs compared with CPUs.

The Jarvis framework includes pre-trained conversational AI models, tools in the NVIDIA AI Toolkit, and optimized end-to-end services for speech, vision, and natural language understanding (NLU) tasks.

Fusing vision, audio, and other sensor inputs simultaneously provides capabilities such as multi-user, multi-context conversations in applications

```
graph TD
    subgraph CONVERSATION_AI [CONVERSATION AI]
        ASR[ASR]
        DM[DIALOG MANAGER]
        NLU[NLU]
        TTS[TTS]
        ASR <--> DM
        DM <--> NLU
        NLU <--> TTS
    end
    subgraph RECOMMENDER_SYSTEM [RECOMMENDER SYSTEM]
        Items[(Items)]
        CG[CANDIDATE GENERATION]
        R[RANKING]
        Items --> CG
        CG --> R
    end
    DM <--> CG
```

Getting Started

This release of Jarvis includes Quick Start scripts to help you get started with Jarvis AI Services. These scripts are meant for deploying the services locally for testing and running the example applications.

```
ngc registry resource download-version nvidia/jarvis/jarvis_quickstart:1.0.0-b.3
```

Initialize and start Jarvis. The initialization step downloads and prepares Docker images and models. The start script launches the server.

```
cd jarvis_quickstart_v*
bash jarvis_init.sh
bash jarvis_start.sh
```

Start a container with sample clients for each service.

```
bash jarvis_start_client.sh
```

Next Steps

- [Jarvis Overview](#)
- [Jarvis Documentation](#)
- [Jarvis Developer Forum](#)

ASR Model

None ▾

Translate

En -> Ja ▾

STOP

We're also making tremendous progress in translation and now offer support for five languages.

You can see how fluid the translation is into Japanese.

And it's running in real time with under one hundred milliseconds latency for each sentence.

また、翻訳も飛躍的に進歩しており、現在5カ国語に対応しています。

日本語の翻訳の流動性がわかります。

そして、各文で100ミリ秒未満のレイテンシでリアルタイムで実行されています。

ANNOUNCING NVIDIA MERLIN END-TO-END ACCELERATED RECOMMENDER SYSTEM

GPU-Accelerated ETL, Data Loading,
Training, Inference

Scales Transparently to Target Data Sets
and More Complex Models

Run in Every Cloud and at the Edge

Now Available on NGC

NVIDIA Merlin

NVIDIA Merlin is a framework for accelerating the entire recommender systems pipeline on the GPU: from data ingestion and training to deployment. Merlin empowers data scientists, machine learning engineers, and researchers to build high-performing recommenders at scale. Merlin includes tools that democratize building deep learning recommenders by addressing common ETL, training, and inference challenges. Each stage of the Merlin pipeline is optimized to support hundreds of terabytes of data, all accessible through easy-to-use APIs. With Merlin, better predictions than traditional methods and

Application
Accelerated framework for building large-scale deep learning recommenders

Popularity
[Bar chart showing popularity]

Domain
Recommendation Engine

License
Apache License 2.0

4.0.1-19.11 (Latest) Scan Results

Linux / x86 [AA] >

Recommendation Systems Deep Learning
HugeCTR Machine Learning NVTabular
Triton Merlin

Expand Details

Getting Started

You can pull the training containers with the following command:

```
docker run --runtime=nvidia --rm -it -p 8888:8888 -p 8797:8787 -p 8796:8786 --ipc=host nvcr.io/nvidia/merlin/merlin-training:0.4 /bin/bash
```

The container will open a shell when the run command execution is completed. It should look similar to this:

```
root@02d56ff0738f:/opt/tritonserver#
```

Activate the rapids conda environment by running the following command:

```
root@02d56ff0738f:/opt/tritonserver# source activate rapids
```

Resources

- [Merlin Overview](#)
- [Merlin Documentation](#)
- [Merlin Developer Forum](#)

NVIDIA MAXINE

SOTA AI to Reinvent Virtual Collaboration

SDKs for Video, Audio, and Augmented Reality

AI-Face Codec 10x Lower Bandwidth vs H.264

Jarvis for Conversational AI

Deploy on Client, in Data Center and Every Cloud

Download Today: developer.nvidia.com/maxine

NVIDIA NGC | CATALOG Download

NVIDIA Maxine

NVIDIA Maxine™ is a GPU-accelerated SDK with state-of-the-art AI features for developers to build virtual collaboration and content creation applications such as video conferencing and live streaming.

Maxine's AI SDKs—Video Effects, Audio Effects, and Augmented Reality (AR)—are highly optimized and include modular features that can be chained into your application.

Application
GPU-accelerated SDK with state-of-the-art AI features for developers to build virtual collaboration and content creation applications such as video conferencing and live streaming

Popularity
[Bar chart showing popularity over time]

Domain
Virtual Collaboration

License
Maxine License

1.0-21.04 (Latest) Scan Results

Linux / x86 AA

Deep Learning | Augmented Reality
GAN | Real-time Audio | Real Time Video
AI | Multimedia

[Expand Details](#)

MAXINE

Video & AR
Decode | Eye Contact | Virtual Background
Super Resolution | Encode

Audio
Decode | Denoise | Echo Cancellation | Encode

Jarvis
ASR | Translation

AV Synchron

Video & Audio

Getting Started

You can pull the Maxine container with the following command:

```
docker run -it --name maxine --gpus=all --shm-size=1g --ulimit memlock=-1 --ulimit stack=67108864 -v ${PWD}:/models -p 8000:8000 -p 8001:8001 -p 8002:8002 nvr.io/nvidia/maxine:1.0-21.04
```

The container will open a shell when the run command execution is completed. It should look similar to this:

```
root@02d56ff0738f:/opt/maxine#
```

Activate the Maxine conda environment by running the following command:

```
root@02d56ff0738f:/opt/maxine# source activate maxine
```

Resources

- [Maxine Overview](#)
- [Maxine Documentation](#)
- [Maxine Developer Forum](#)



- AI FACE CODEC
- EYE CONTACT
- MACHINE TRANSLATION
- SPEECH RECOGNITION

that's so much better thanks to NVIDIA Maxine.

No es ideal. Ahora vamos a volver a activar todas las características de Maxine.

ANNOUNCING NVIDIA TRITON INFERENCE SERVER

MODELS

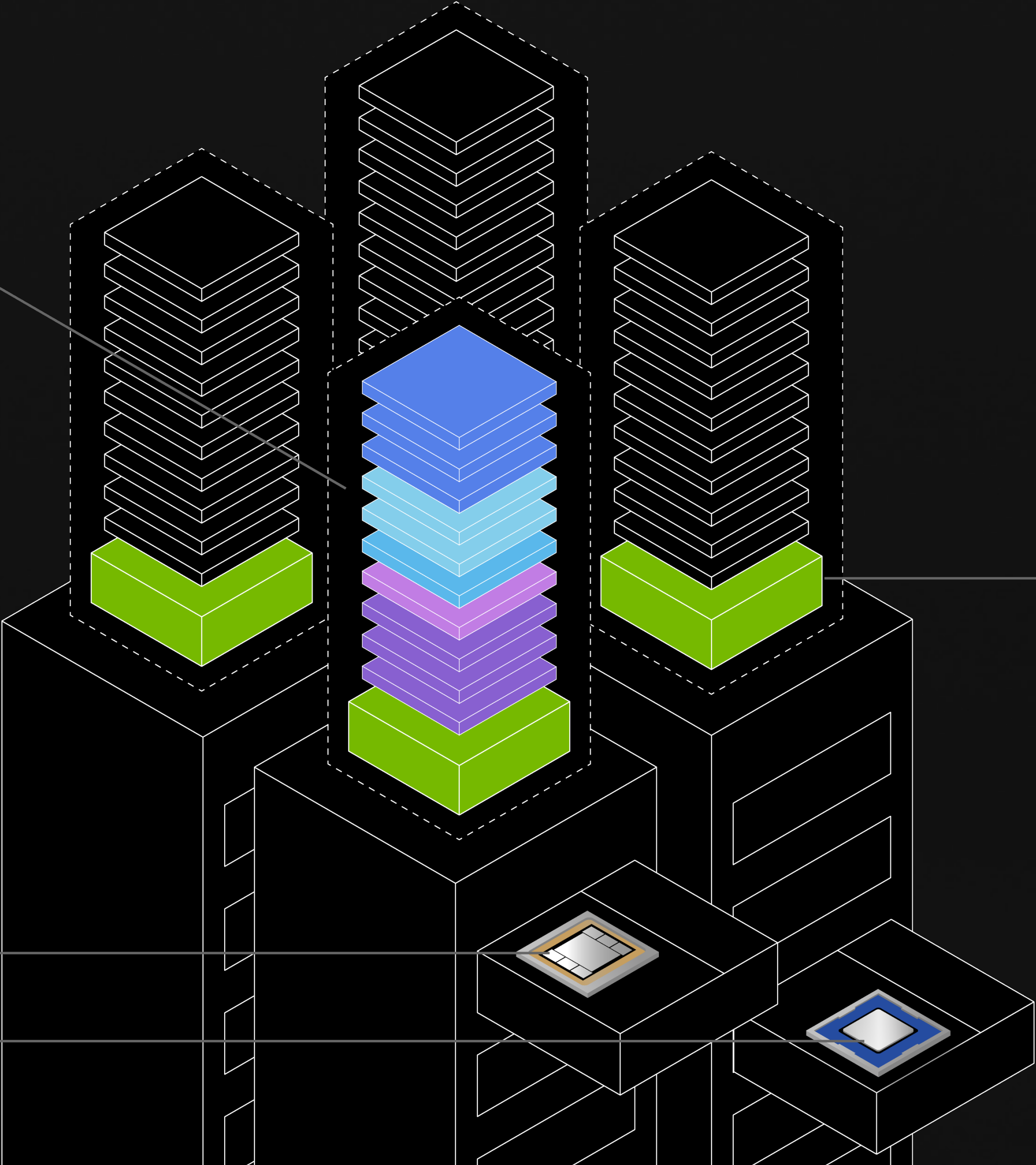
Infinite AI Models
CV, ASR, NLU, TTS, RecSys

Multiple Frameworks
TensorFlow, Pytorch,
ONNX, TensorRT

Varying Services
Batch, Real-Time,
Streaming

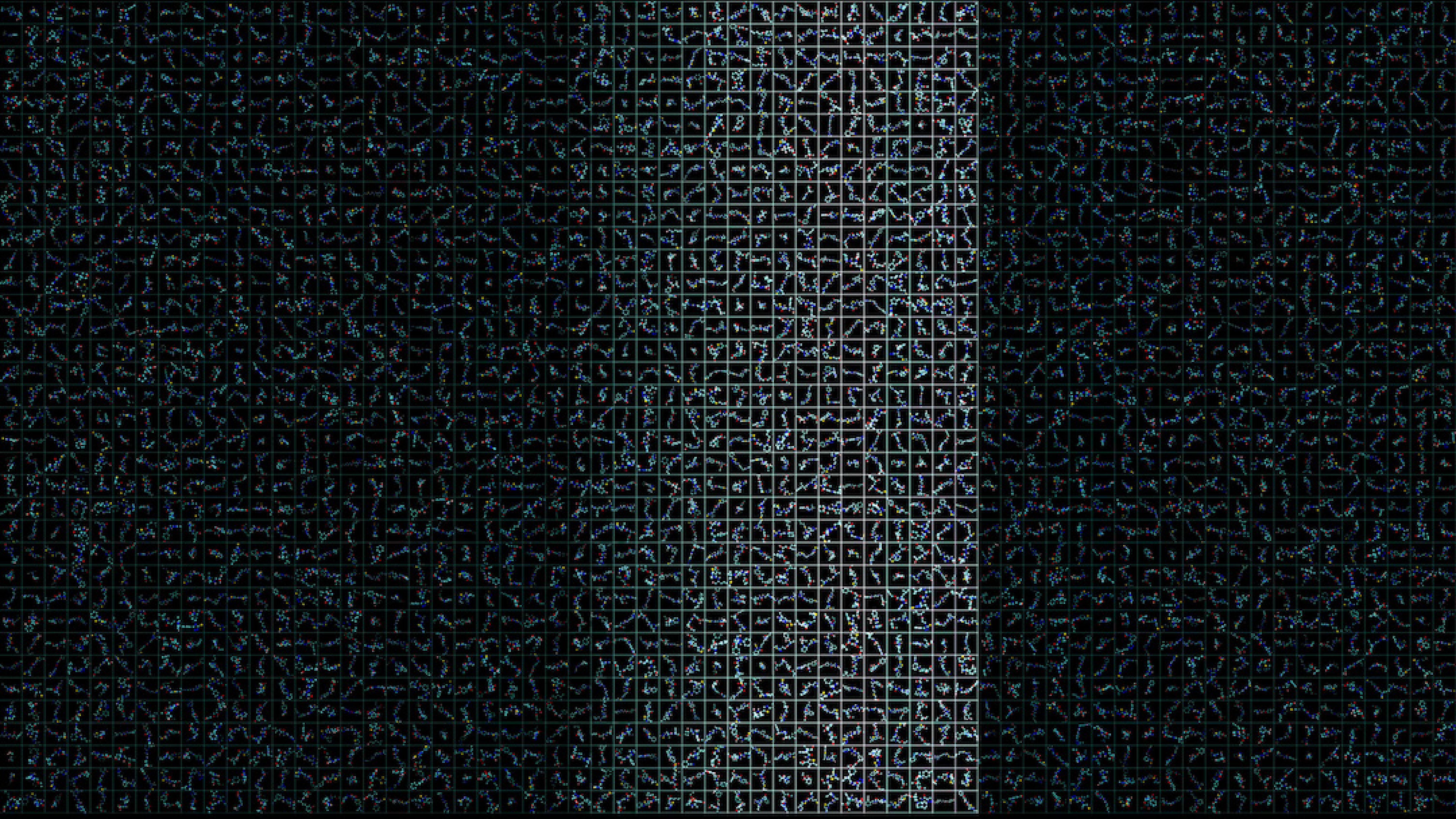
GPU

CPU



TRITON INFERENCE SERVER

Any Model
Any Framework
Any Service Requirement
CPU or GPU



THOUSANDS OF COMPANIES DOING COOL THINGS WITH NVIDIA AI

Analyze 225K
Network Events per Second



Accurately Detect Diseases
in 145M Hearts per Year



Identify Trends in over 300B Pins
for Better Search Results



Tasteful Recommendations
from 600K Restaurants



Postmates

Personalized Playlists
for over 345M Listeners



Award-Winning Customer Care
Using Real-Time ASR

T-Mobile™

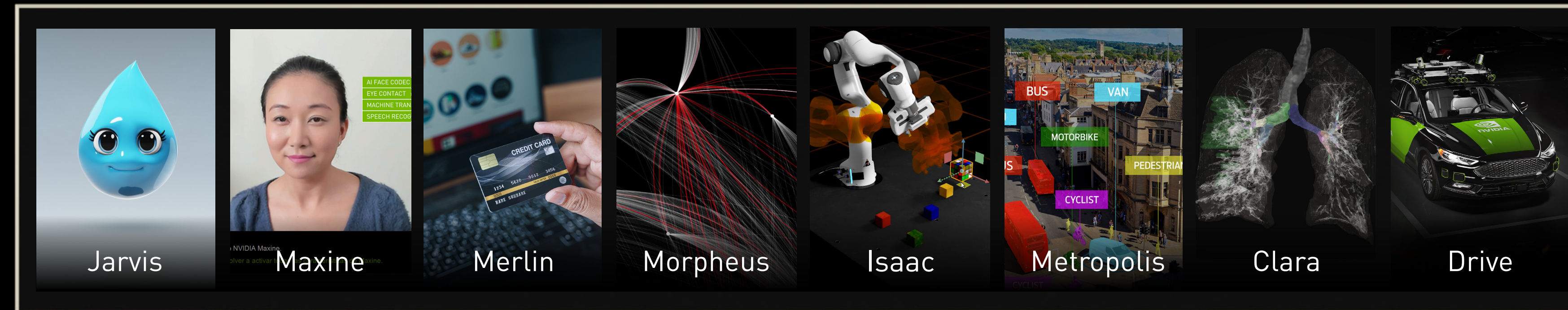
Real-Time Analytics
on 7B Packages per Year



Intelligent Search with SOTA NLU
for 1.2B Users



ANNOUNCING NVIDIA AI ENTERPRISE PLATFORM



NVIDIA AI ENTERPRISE

NVIDIA GPU CLOUD REGISTRY

cuIO

cuDF

cuDNN

cuML

cuGraph

TAO

TensorRT

Triton

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

VMware vSphere



AtoS

DELL Technologies

FUJITSU

GIGABYTE

HBC

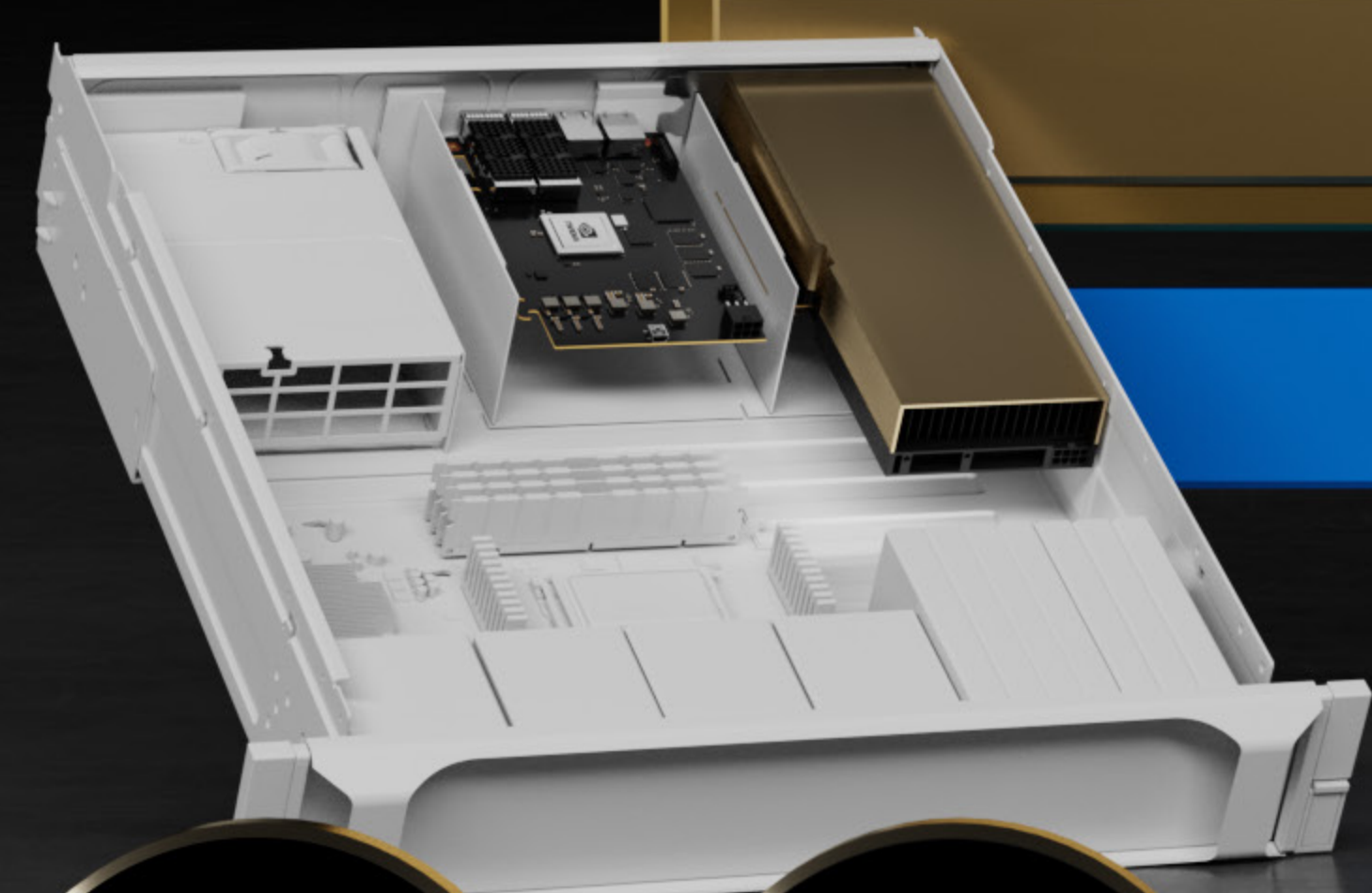
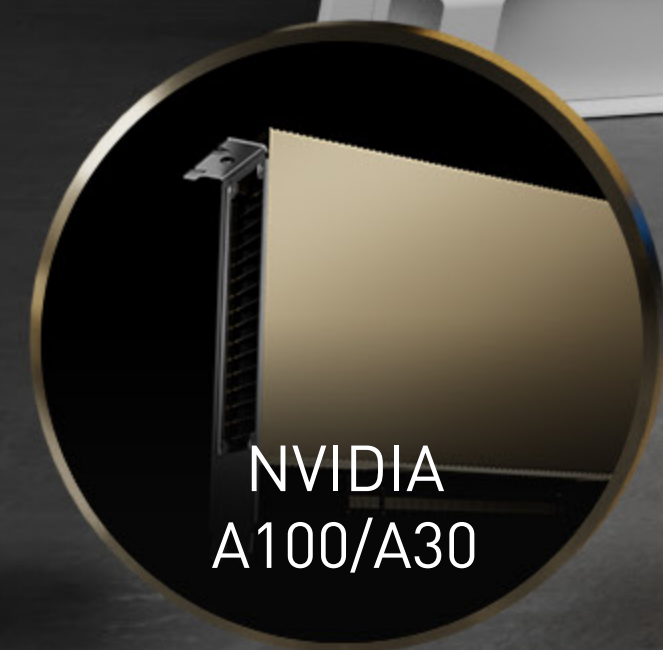
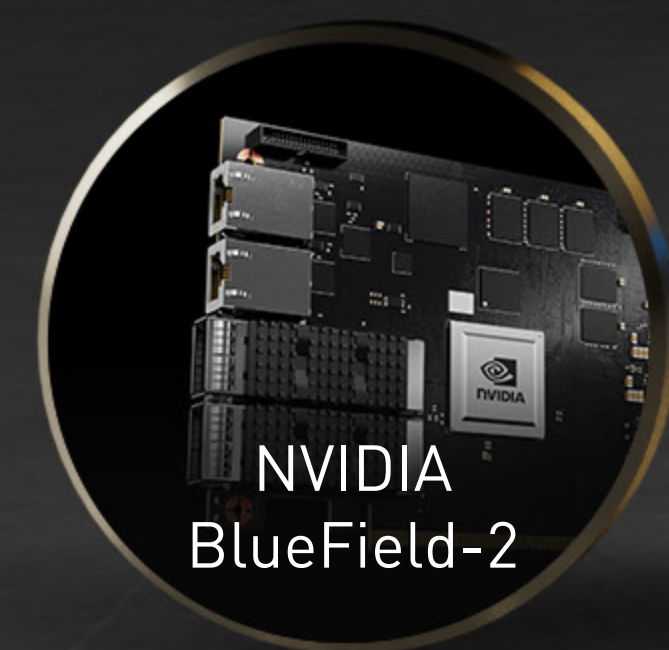


inspur

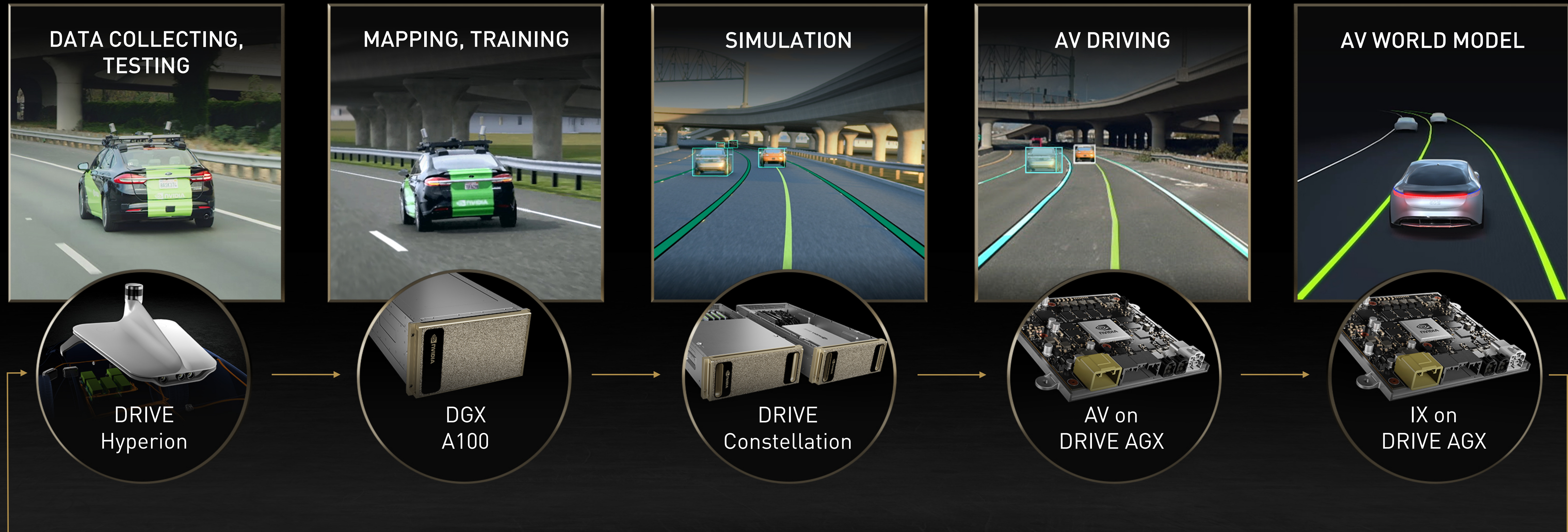
Lenovo

QCT

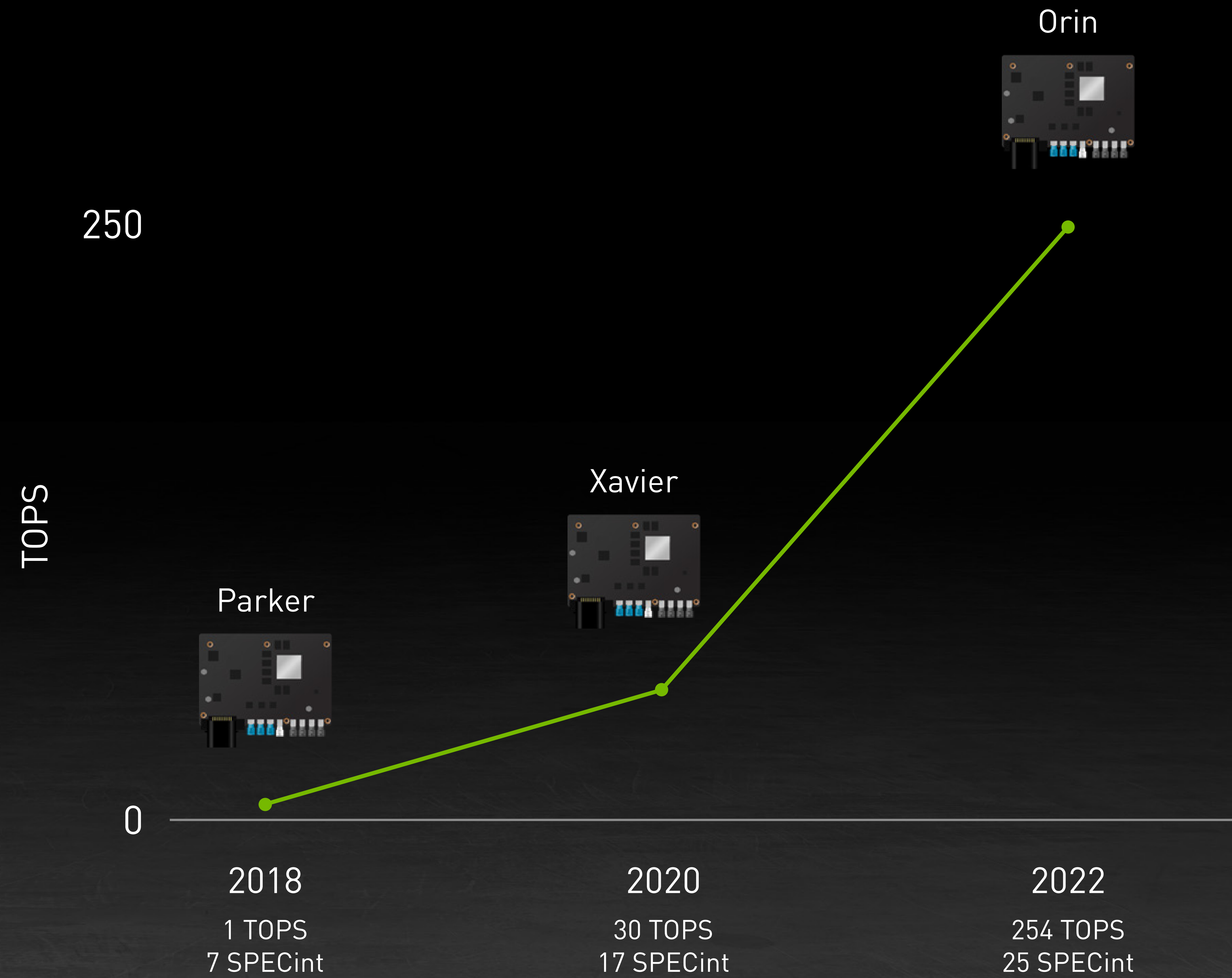
SUPERMICRO



NVIDIA DRIVE AV



TOPS IS THE NEW HORSEPOWER



ANNOUNCING ORIN CENTRAL COMPUTER


1 Chip – 4 Domains

CLUSTER

INFOTAINMENT

PASSENGER
INTERACTION & MONITORING


AV WITH CONFIDENCE VIEW



FUSA Graphics
OS #1



Entertainment Graphics & Multimedia
OS #2

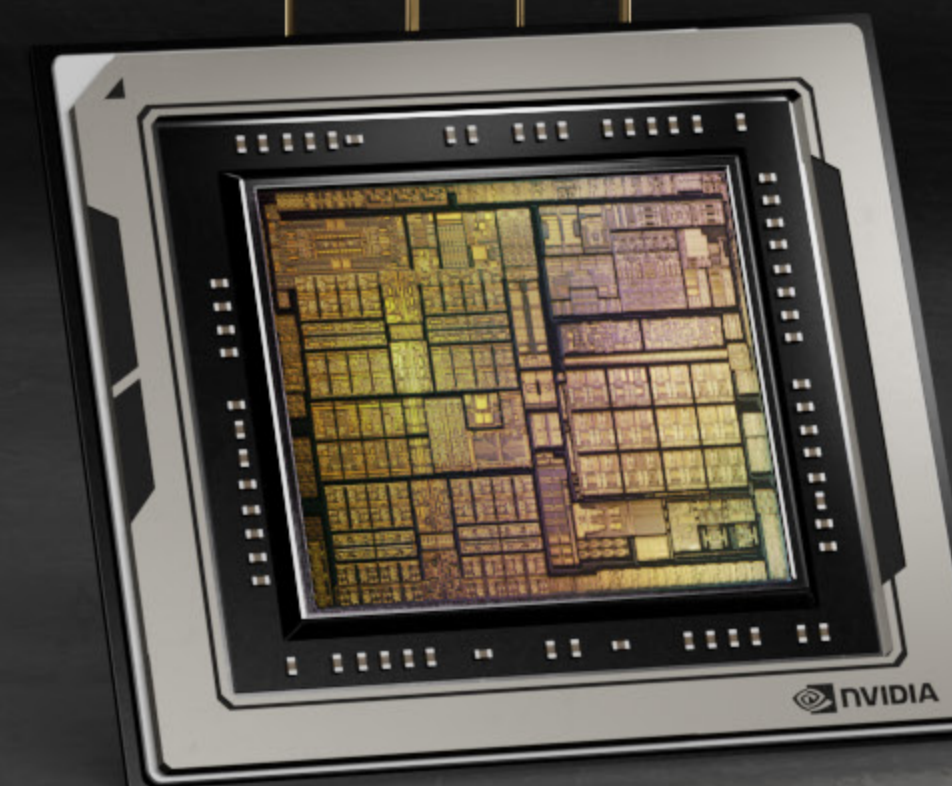


FUSA CV, AR, & Conversational AI
OS #4



FUSA CV & AR
OS #3

ORIN VEHICLE COMPUTER



- Centralized & Software Defined
- Tightly Integrated with AV Applications
- Containerized & Easy to Upgrade
- SOTA Security and Functional Safety

ANNOUNCING
HYPERION 8 AV PLATFORM
State-of-the-Art Advances for
Data Collection, Development and Testing

2x Orin AV Computer

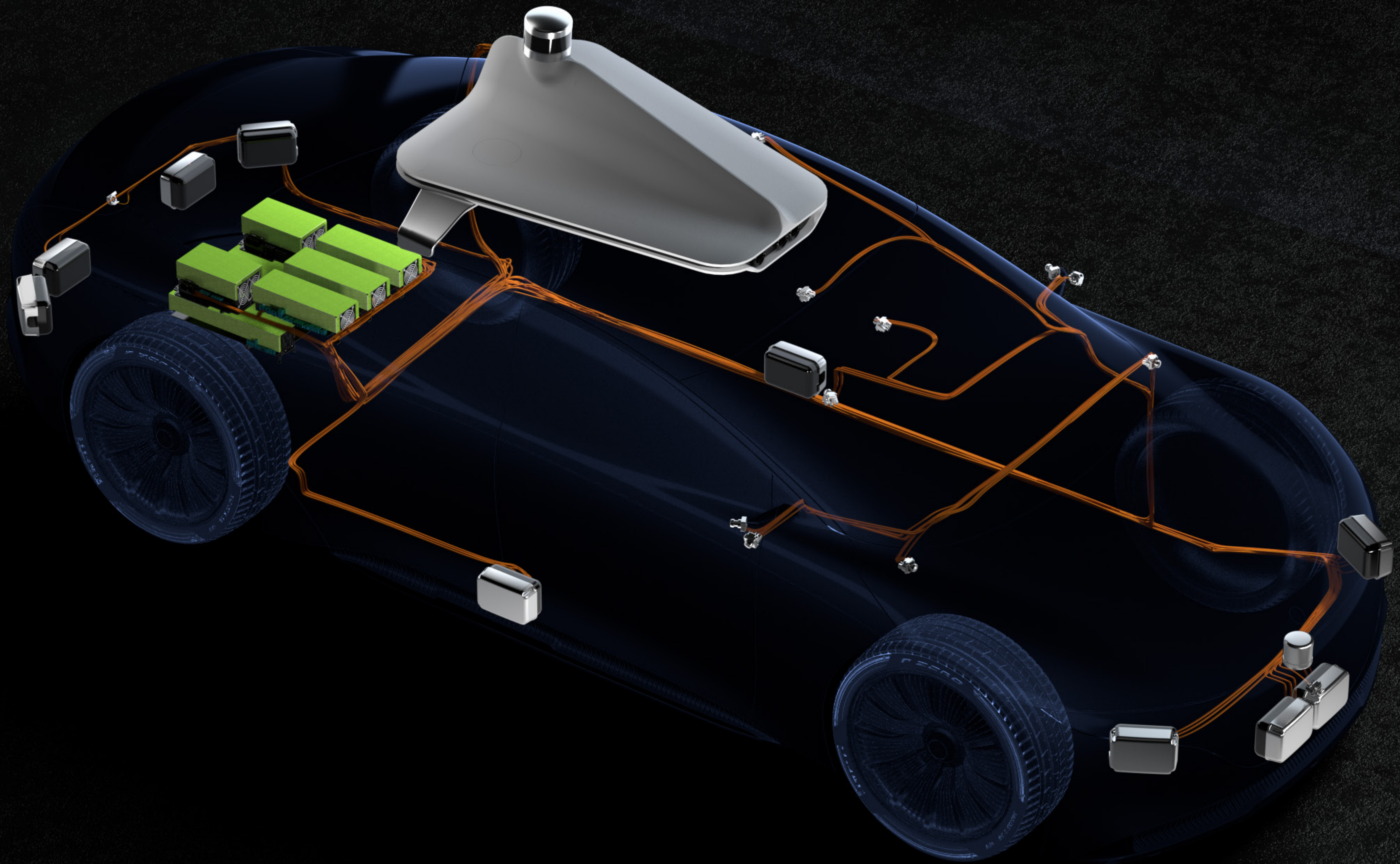
1x Orin IX Computer

4x Orin + 4x MLNX 3D GT Data Recorder

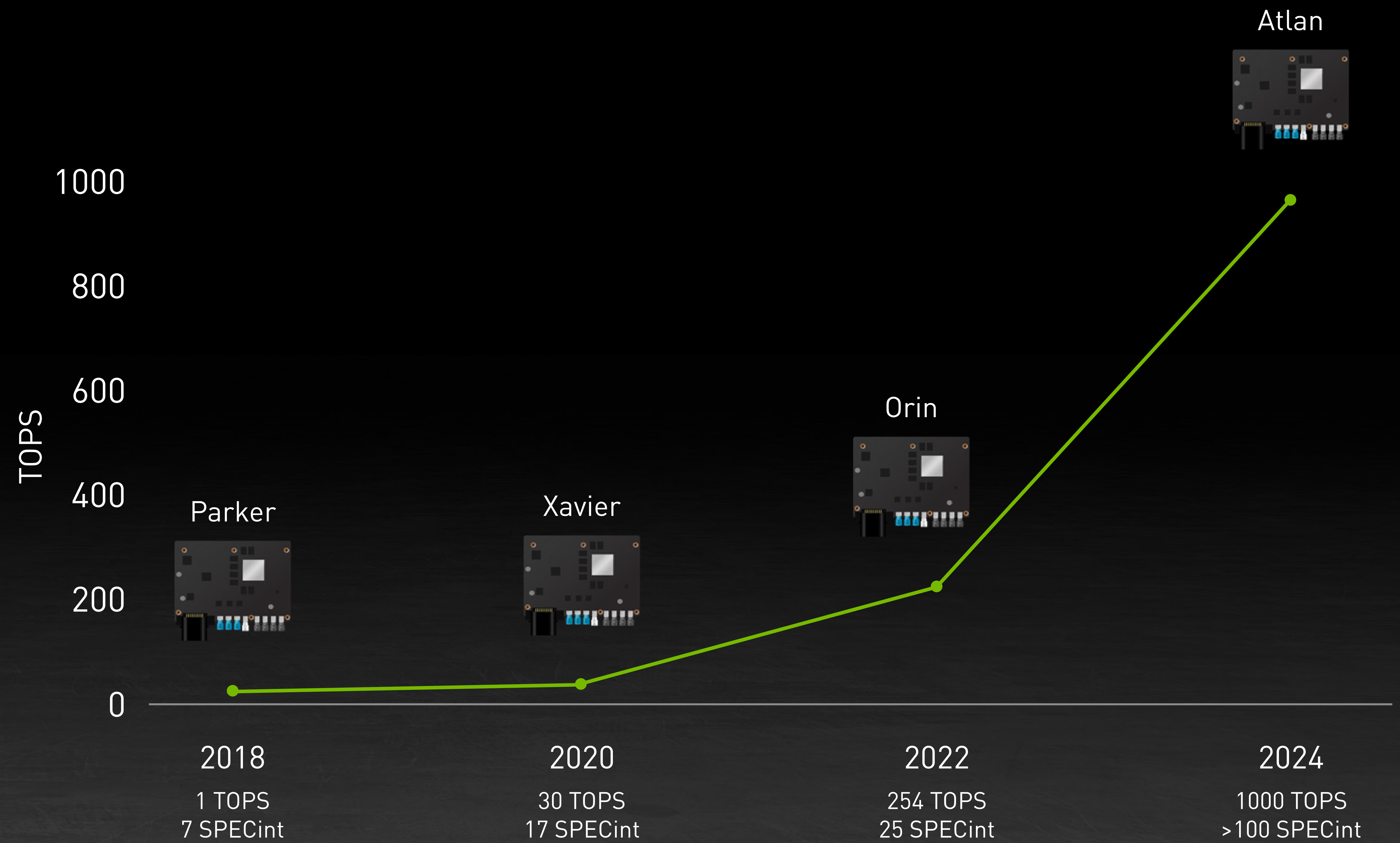
Sensor Suite: 8 Cameras (8MP), 4 Fisheyes (3MP),
3 In-Cabin, 9 Radar, 2 Lidar

Source Access to AV & IX Software Repository

OTA Ready



TOPS IS THE NEW HORSEPOWER



ANNOUNCING NVIDIA DRIVE ATLAN

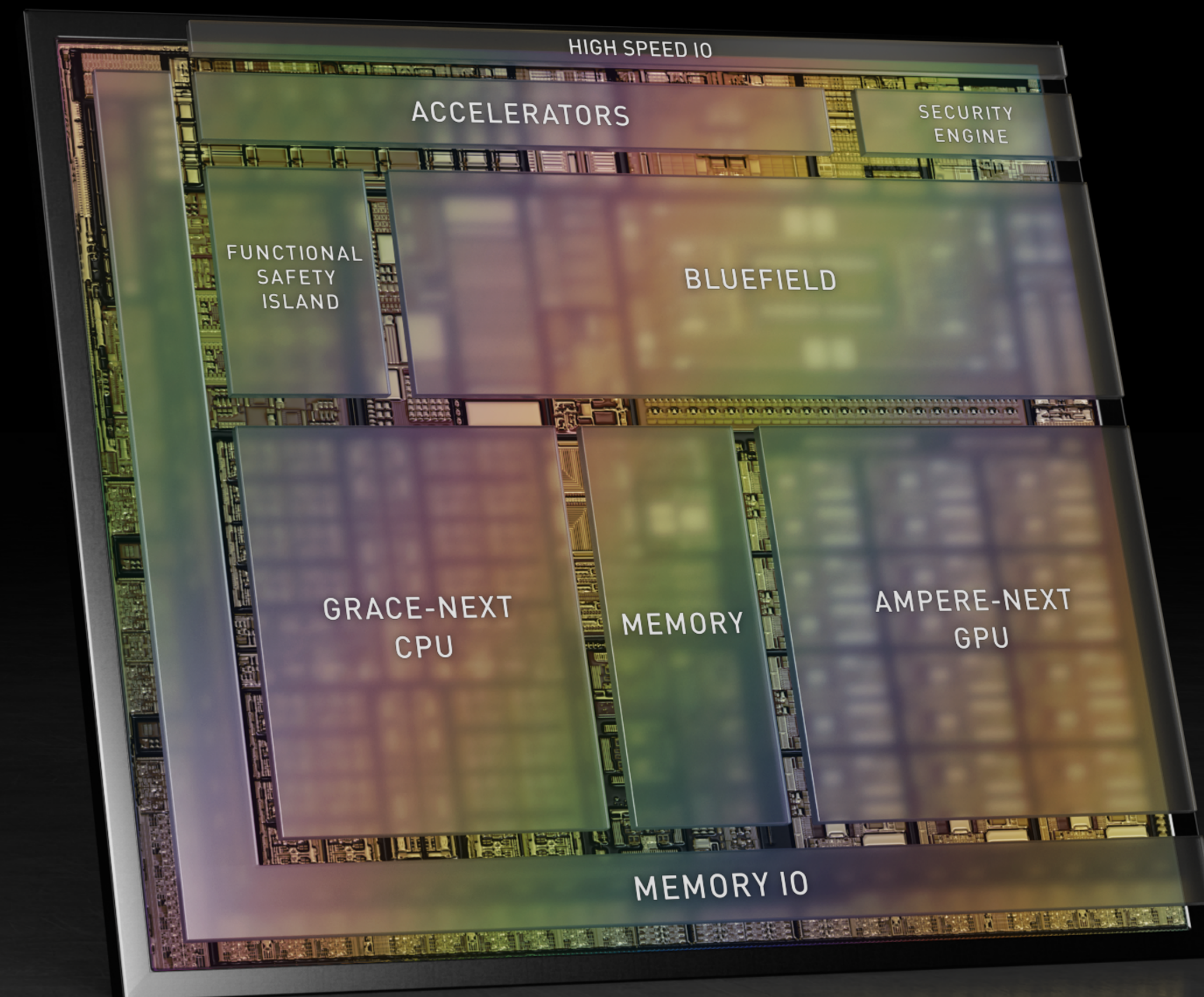
The Next Level – Same Programmable
Architecture

Fusing Next Generation AI and BlueField

Industry's First 1,000 TOPS SoC

400 Gbps Networking with Secure Gateway

ASIL-D Safety Island



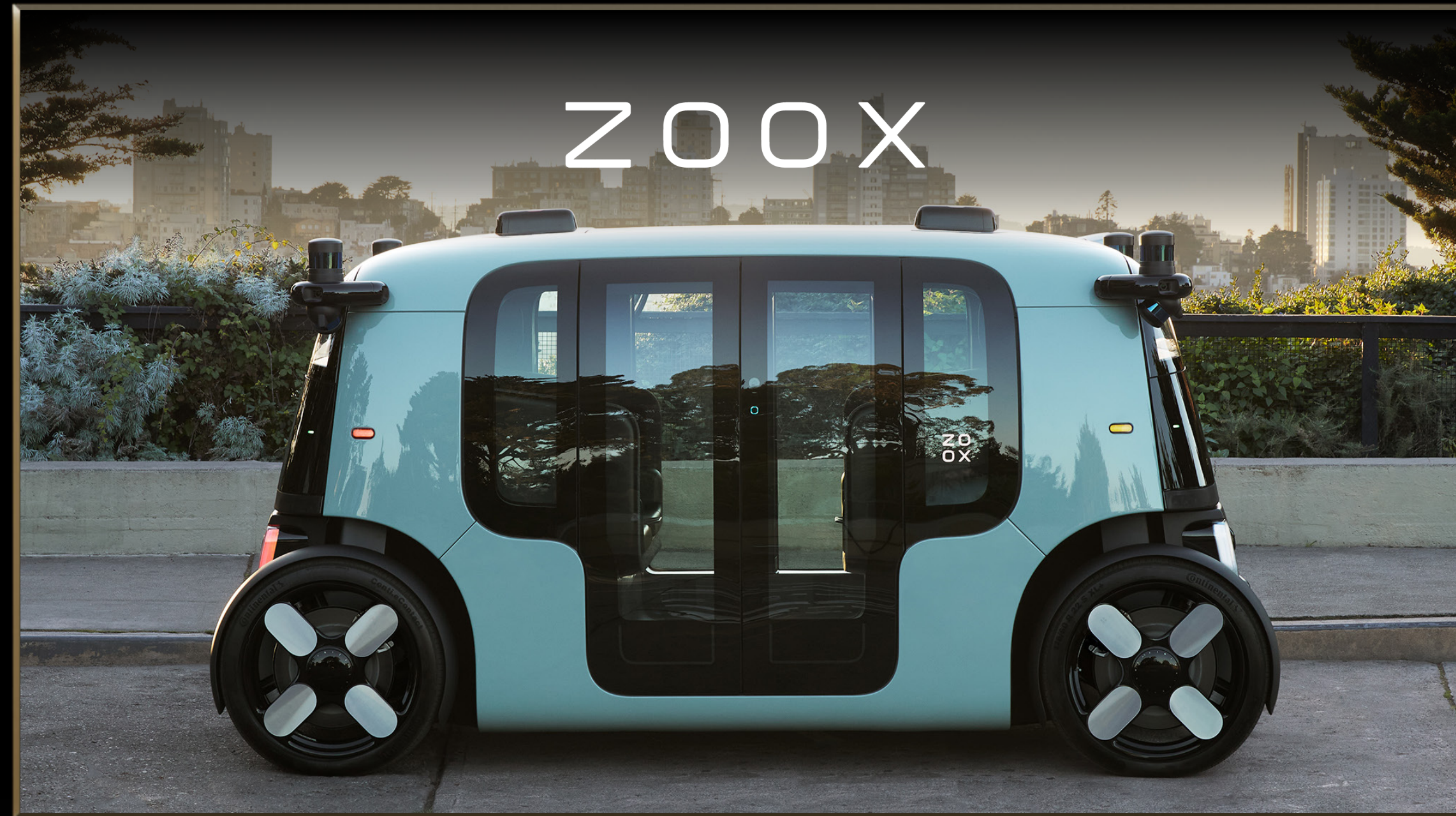
A NEW BREED OF TECHNOLOGY COMPANIES



THE WORLD'S BIG BRANDS HAVE GIANT OPPORTUNITIES



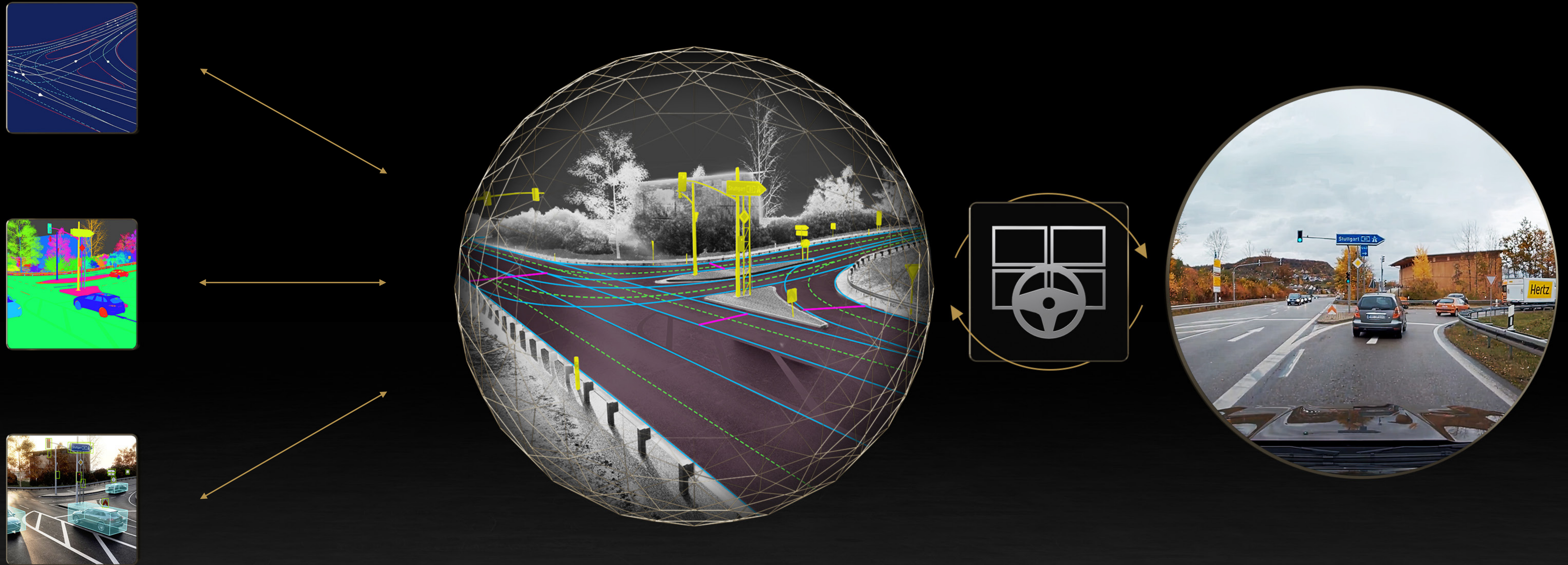
TRANSFORMING 10 TRILLION MILES A YEAR INTO A SERVICE



THE INTERNET OF ATOMS



NVIDIA DRIVE DIGITAL TWIN IN OMNIVERSE



ANNOUNCING NVIDIA DRIVE SIM POWERED BY OMNIVERSE



PHYSICALLY ACCURATE SENSORS

CLOUD-NATIVE WORKFLOW

SCALABLE ARCHITECTURE

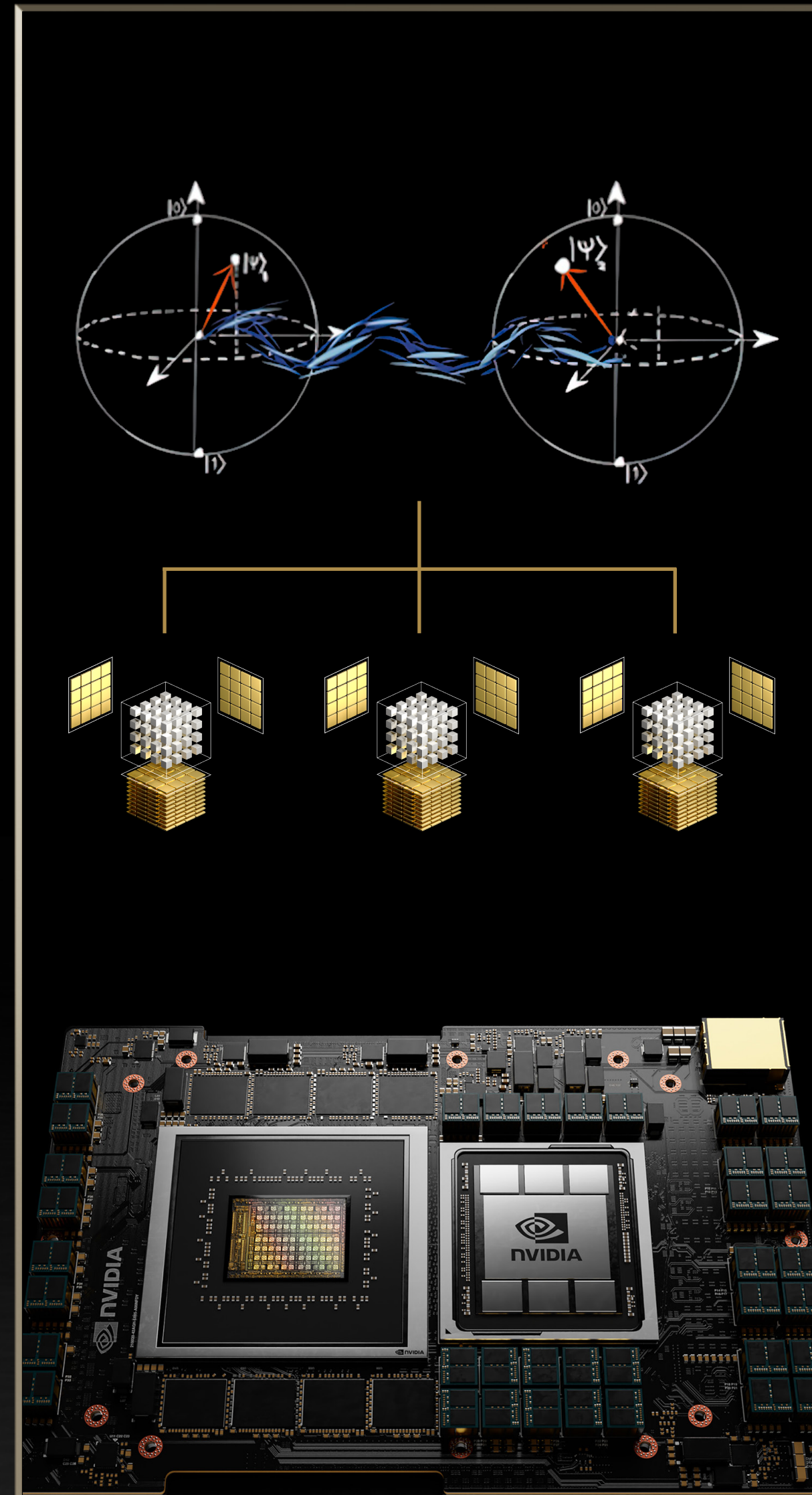
OPEN | MODULAR | EXTENSIBLE



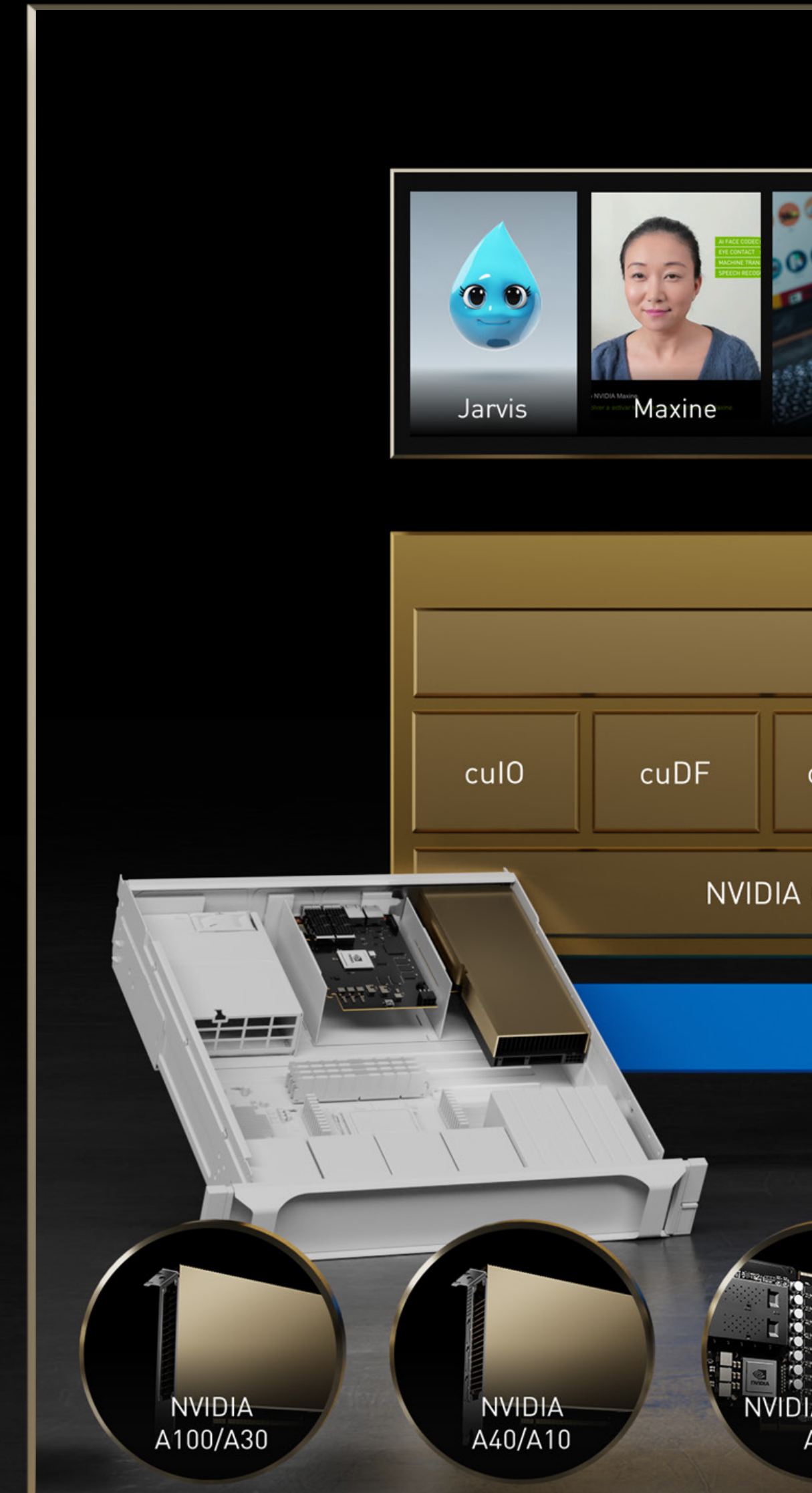
NEW NVIDIA TECHNOLOGIES



Omniverse
Isaac Sim



Megatron | Drug Discovery
Quantum Computing
DGX | Grace
BlueField-3 | DOCA 1.0



Jarvis | Merlin
Maxine | Morpheus
NVIDIA AI
EGX Aerial 5G



DRIVE Sim
Hyperion 8
Atlan
Orin



PRODUCT OF ITALY



TRAIN | ADAPT | OPTIMIZE



TAO
EXTRA VIRGIN
OLIVE OIL

Mado olive oil is luxury you can drizzle. Puncty and fresh, it's made from 100% Italian olives, all picked by hand and pressed the same day. It's an olive oil so tasty it can spirit any dish straight to gastronomic heaven.

Also try our olive oil crushed with fresh lemons
Produced in Italy for Mado Ltd London UK



