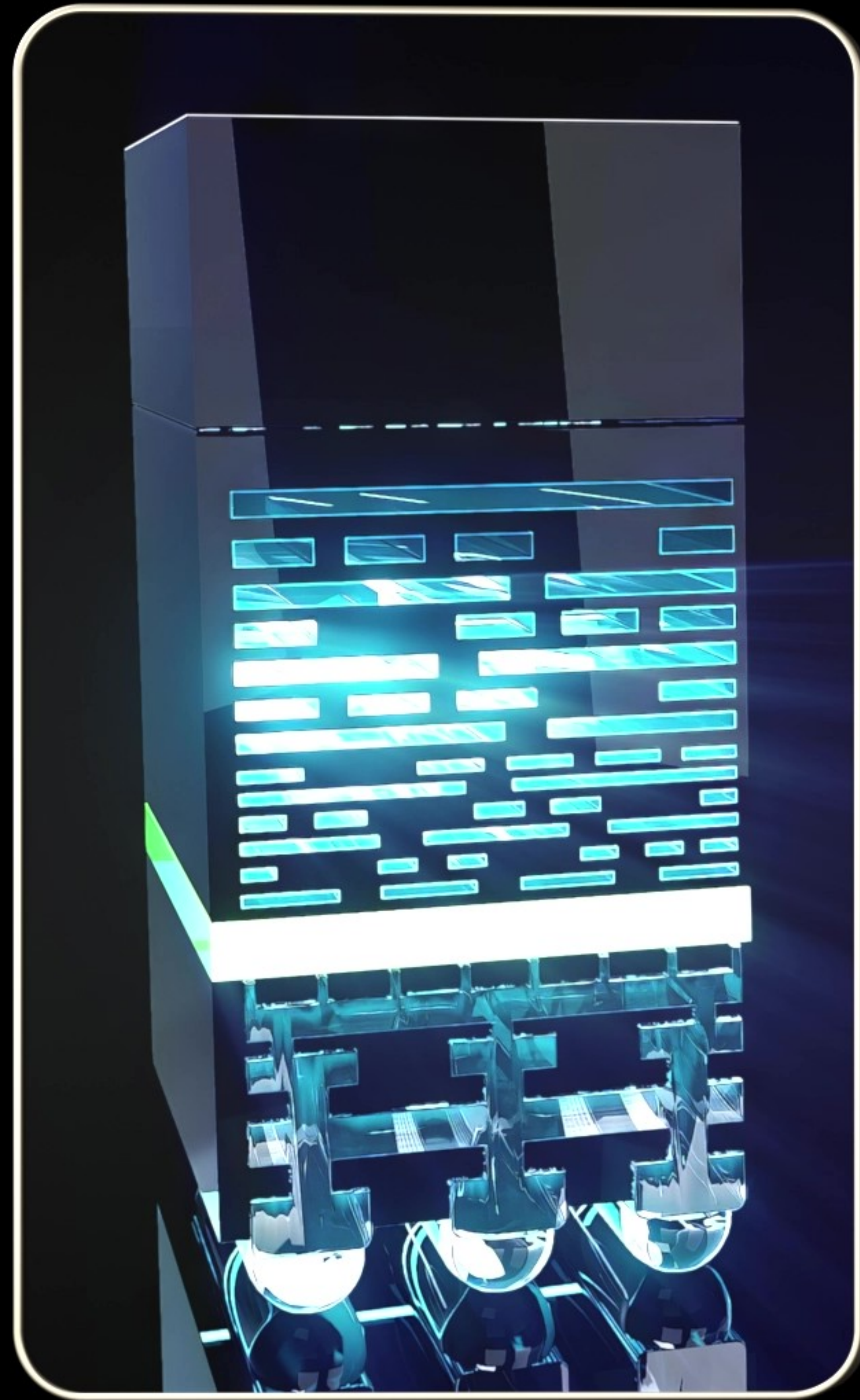
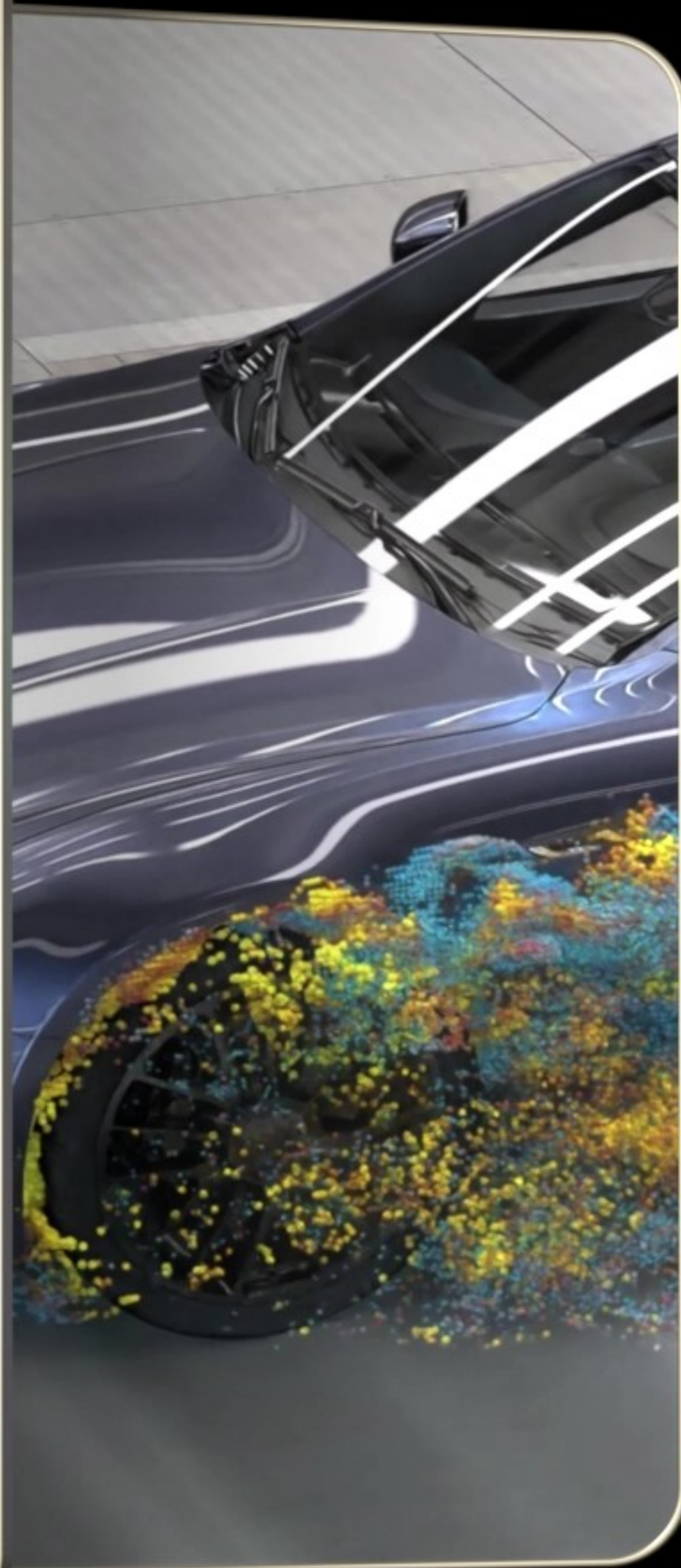








cuLitho



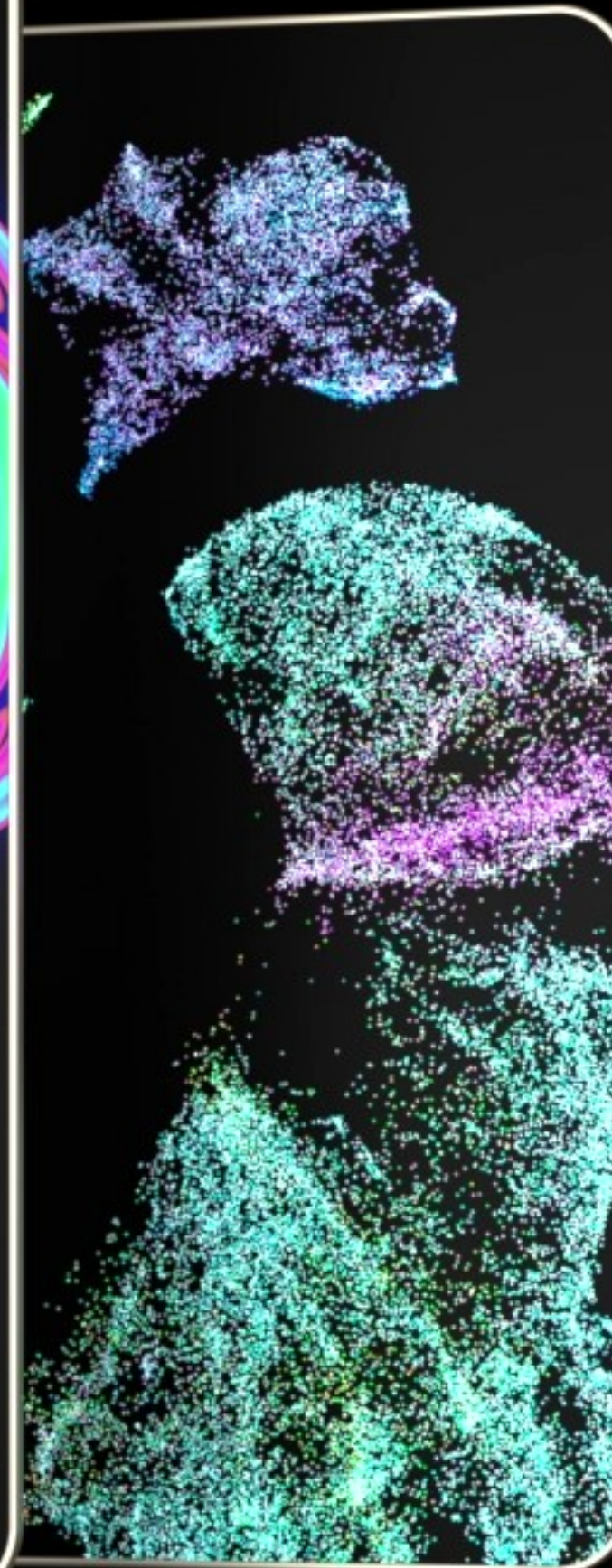
cuDSS
AmgX



cuOpt



Warp



cuDF
cuML



Dynamo
cuDNN



cuEquivariance
cuTensor



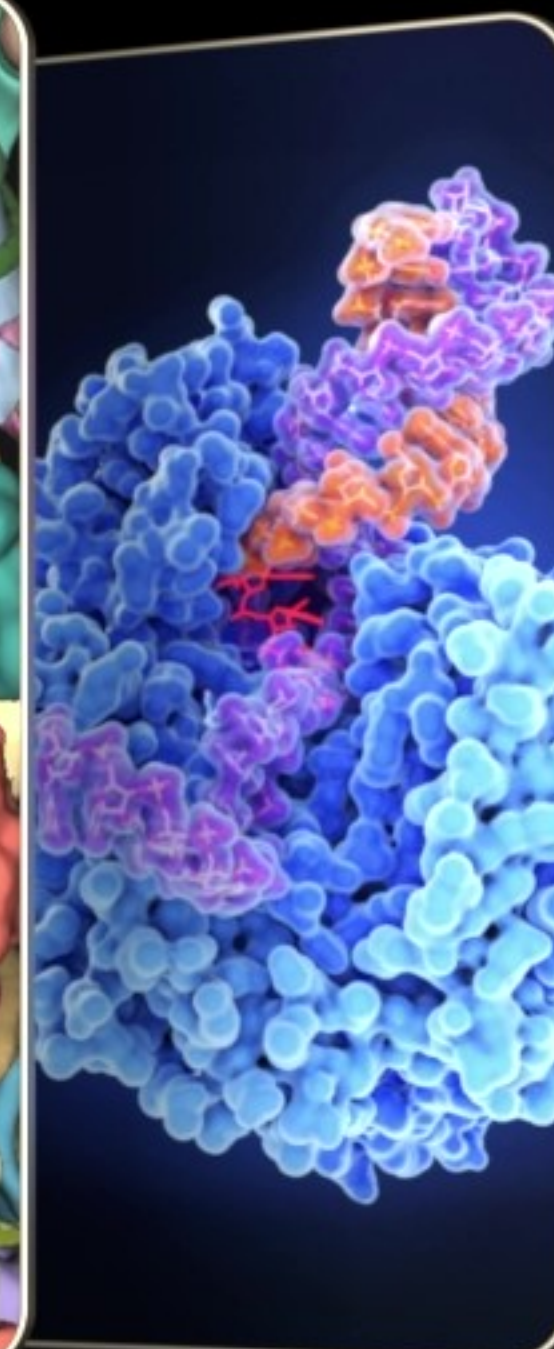
Aerial
Sionna



Earth-2



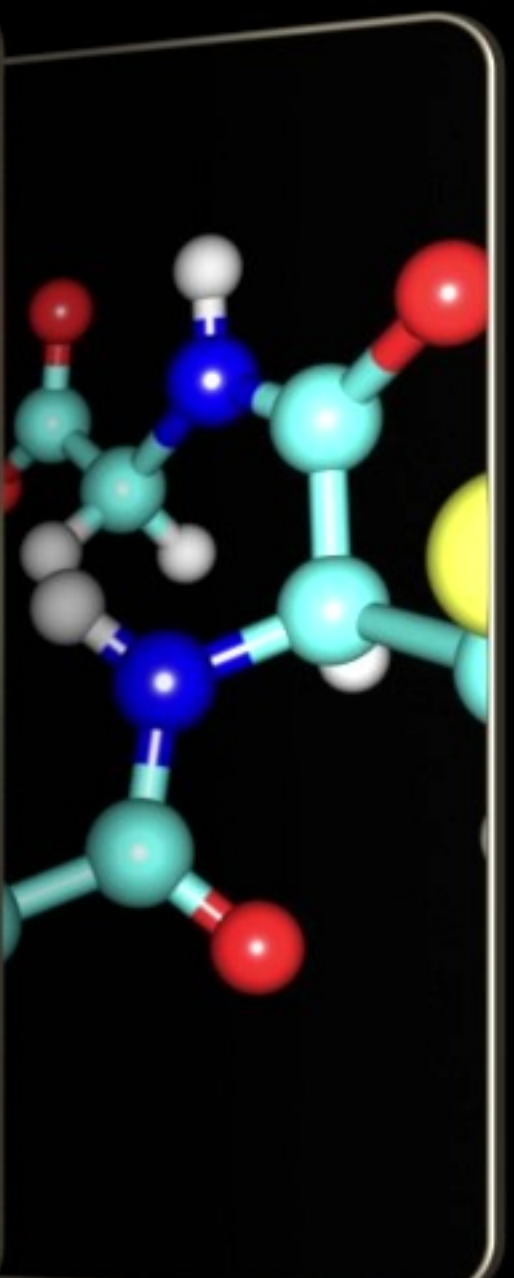
MONAI



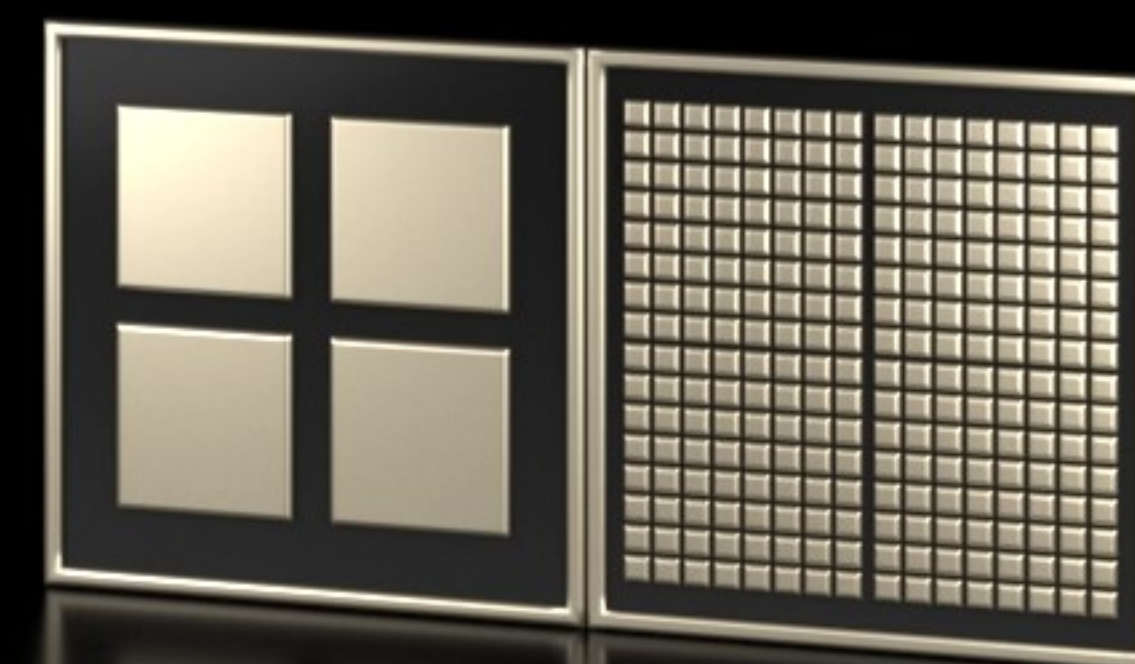
Parabricks



cuQuantum
CUDA-Q

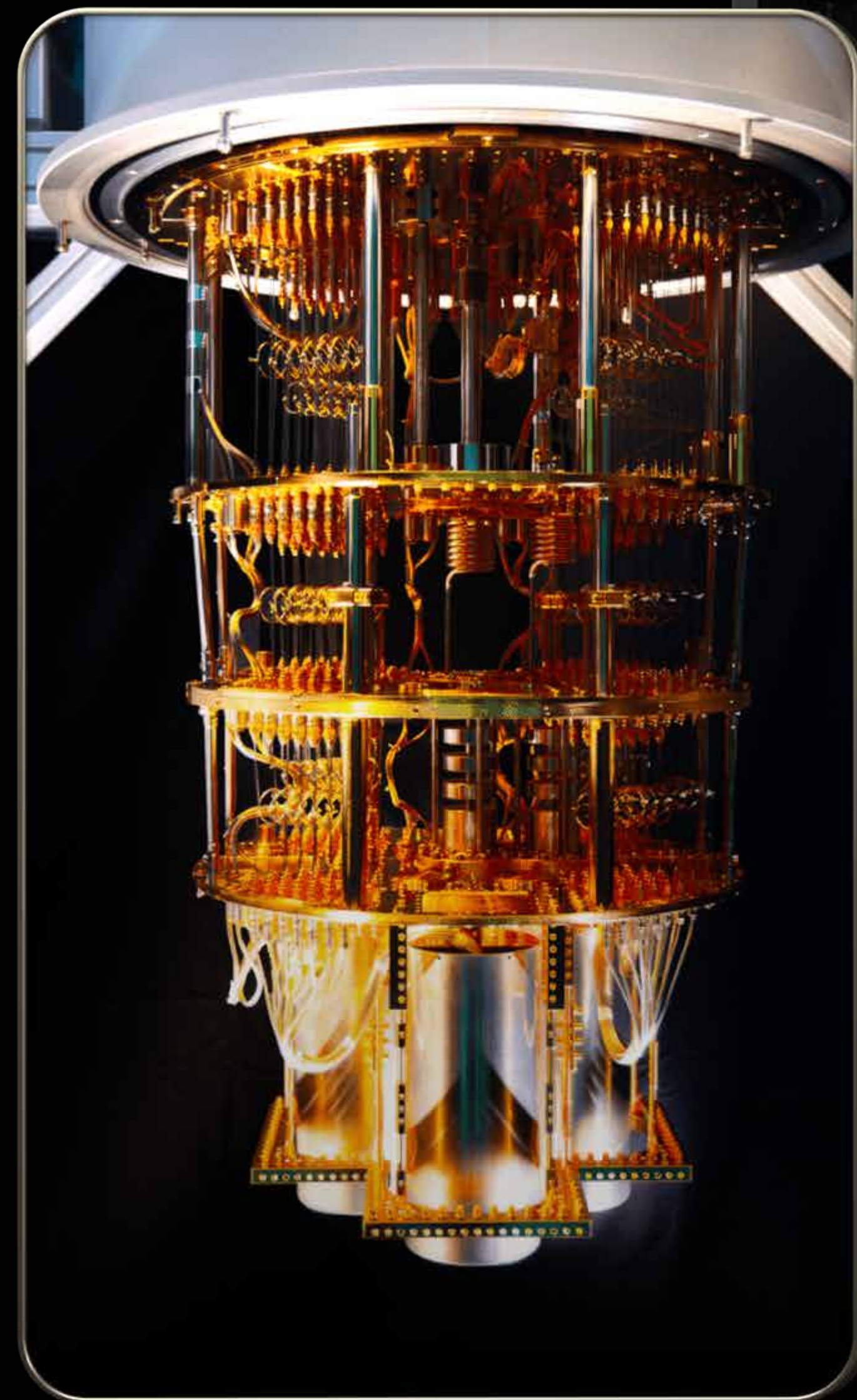


cuPyNumeric



NVIDIA CUDA-Q

Quantum-GPU Classical Computing Across Europe



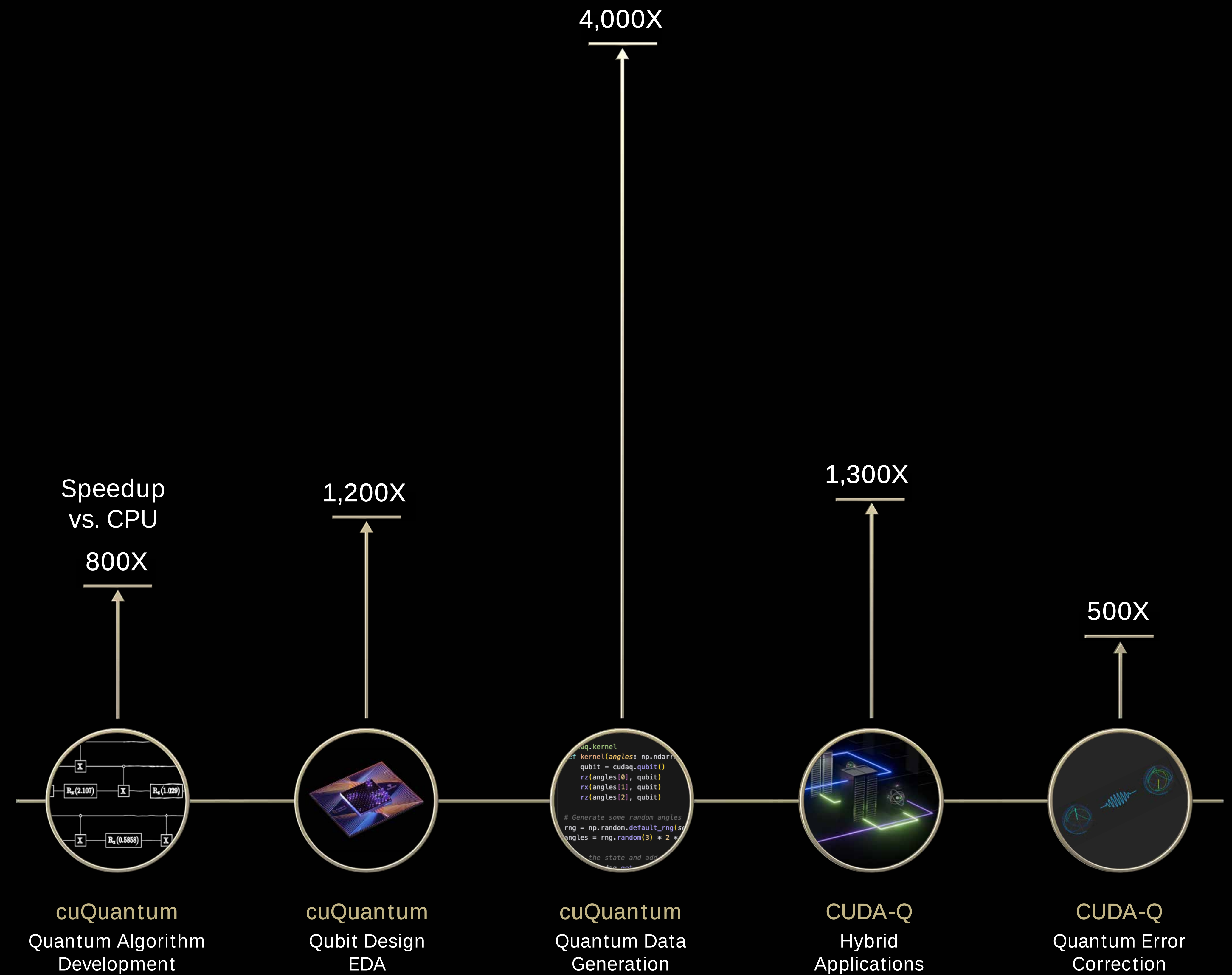
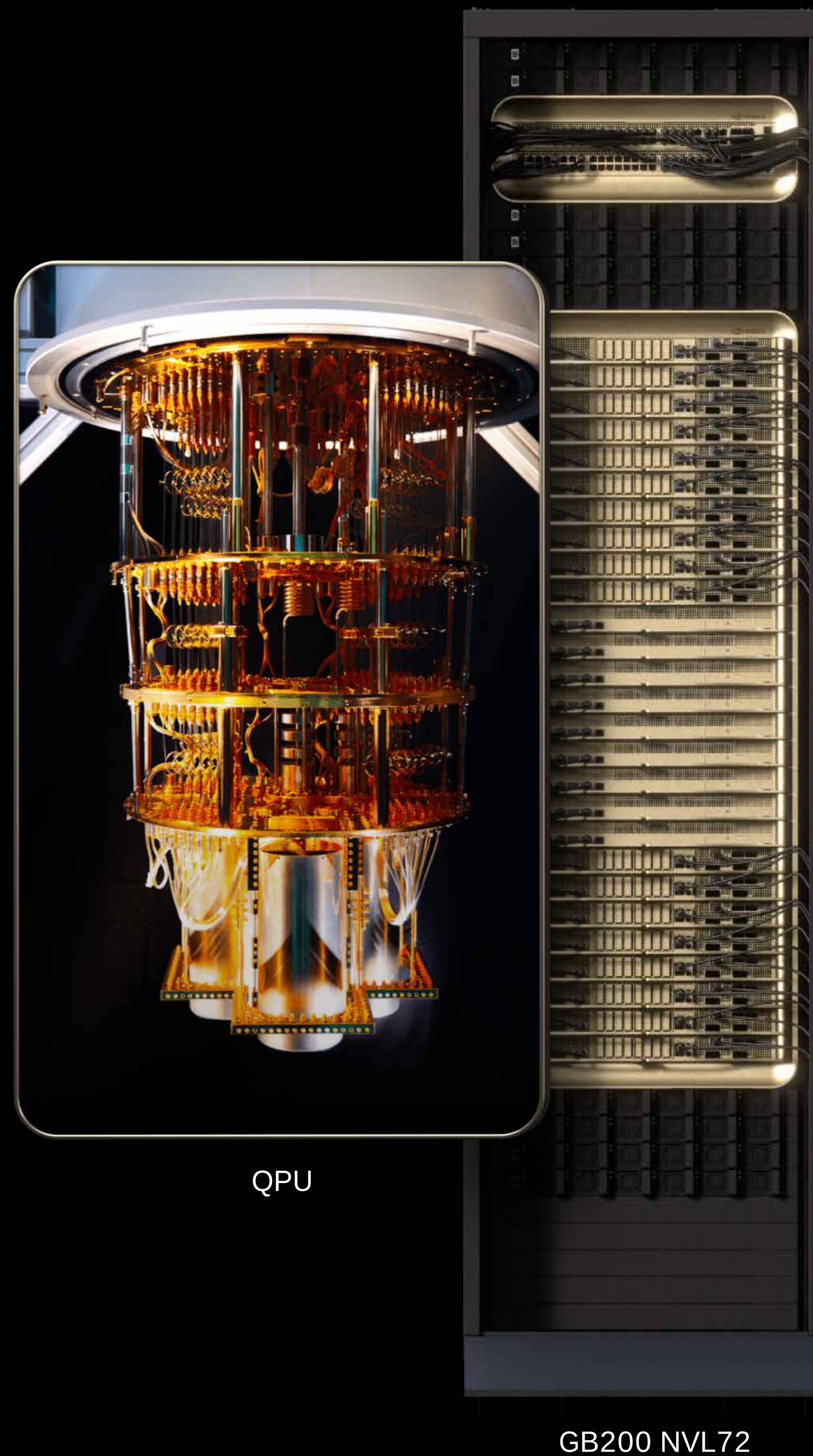
QPU

GB200 NVL72

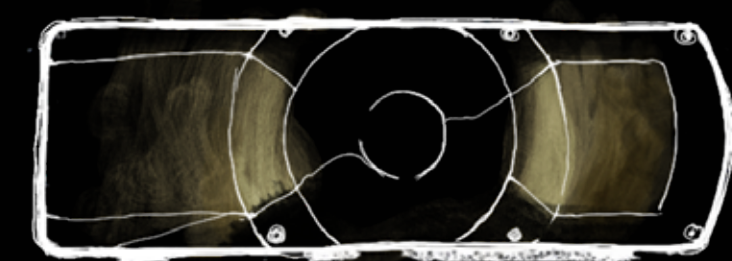


NVIDIA CUDA-Q

Quantum-GPU Classical Computing



"FIRST CONTACT"



2012 ALEXNET

PERCEPTION AI

SPEECH RECOGNITION
DEEP RECSYS
MEDICAL IMAGING

GENERATIVE AI

DIGITAL MARKETING
CONTENT CREATION



AGENTIC AI

CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE



PHYSICAL AI

AUTONOMOUS VEHICLES
GENERAL ROBOTICS

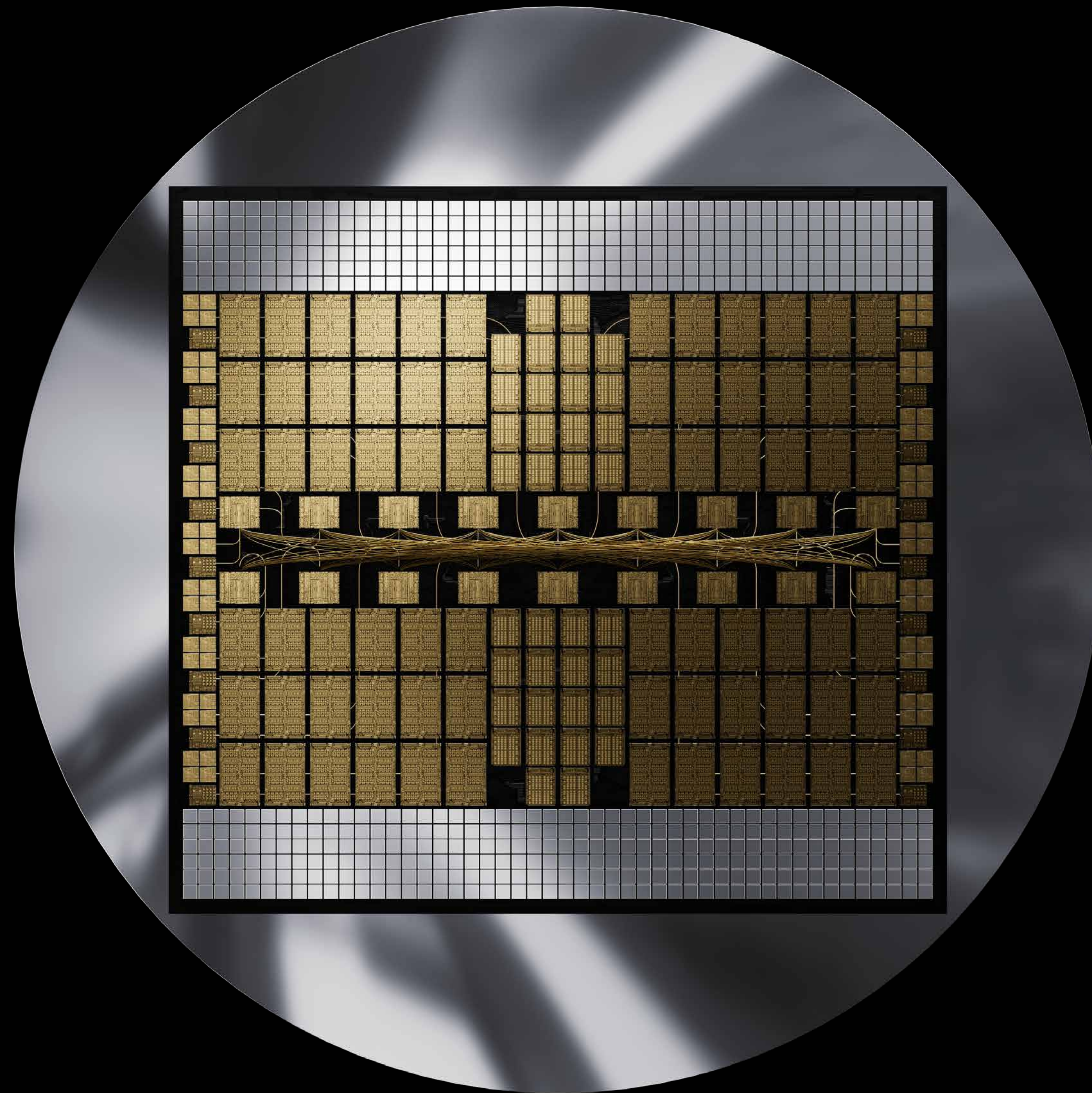


GB300 NVL72

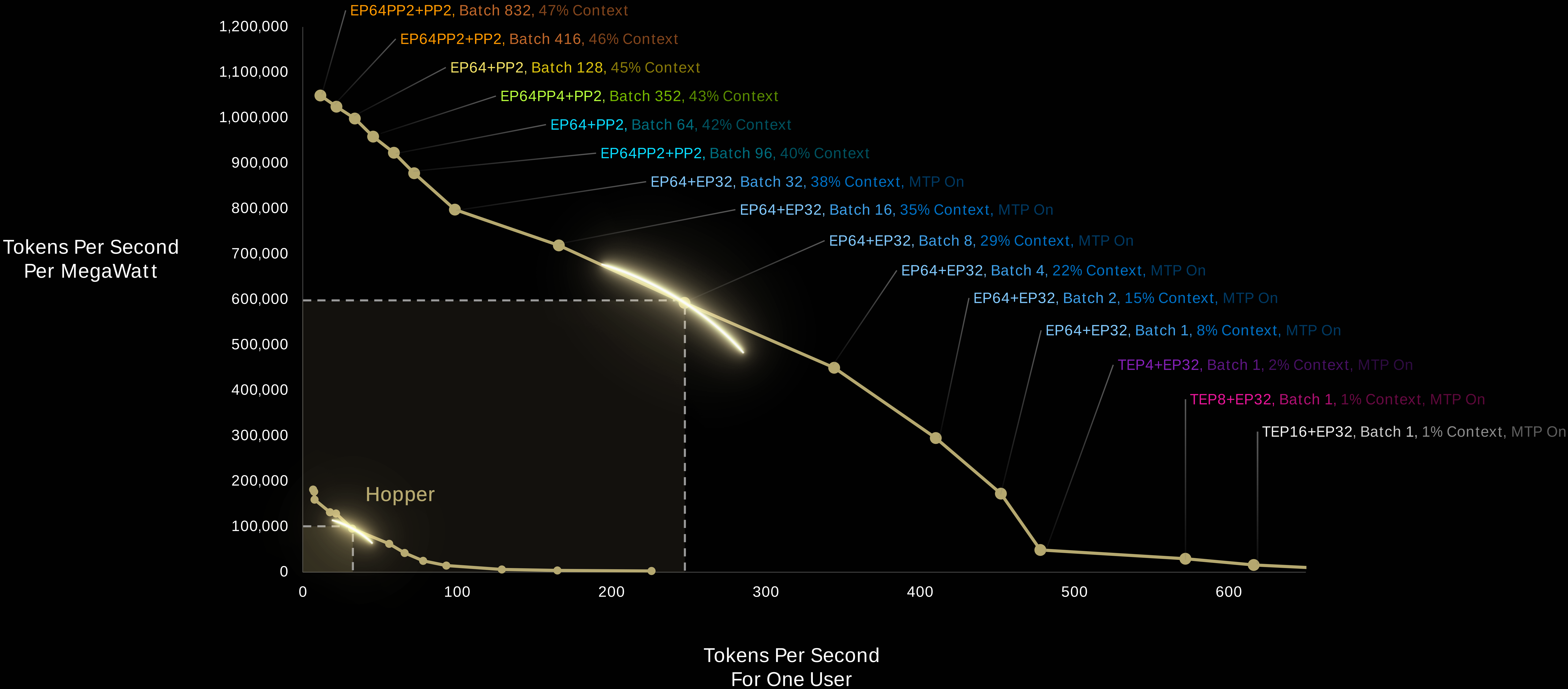
Grace Blackwell NVL72
“A Thinking Machine”

Grace Blackwell NVL72

“One Giant GPU”

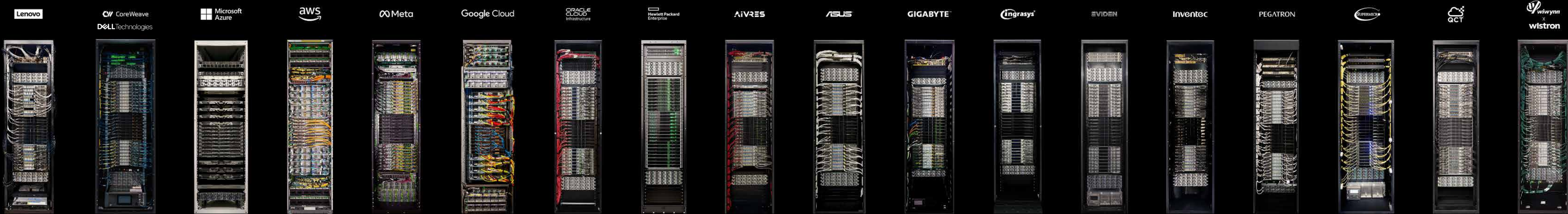


NVIDIA Blackwell Massive Leap in Reasoning Inference Performance





Grace Blackwell in Full Production



One Architecture – From Cloud AI, Enterprise AI, Personal AI, to Edge AI



One Architecture – From Cloud AI, Enterprise AI, Personal AI, to Edge AI



DGX Spark

NEW



DGX Station

NEW



GB200 / GB300 NVL72



RTX PRO Workstation

NEW



RTX PRO Laptop

NEW



RTX PRO Server

NEW



DGX B200 / B300

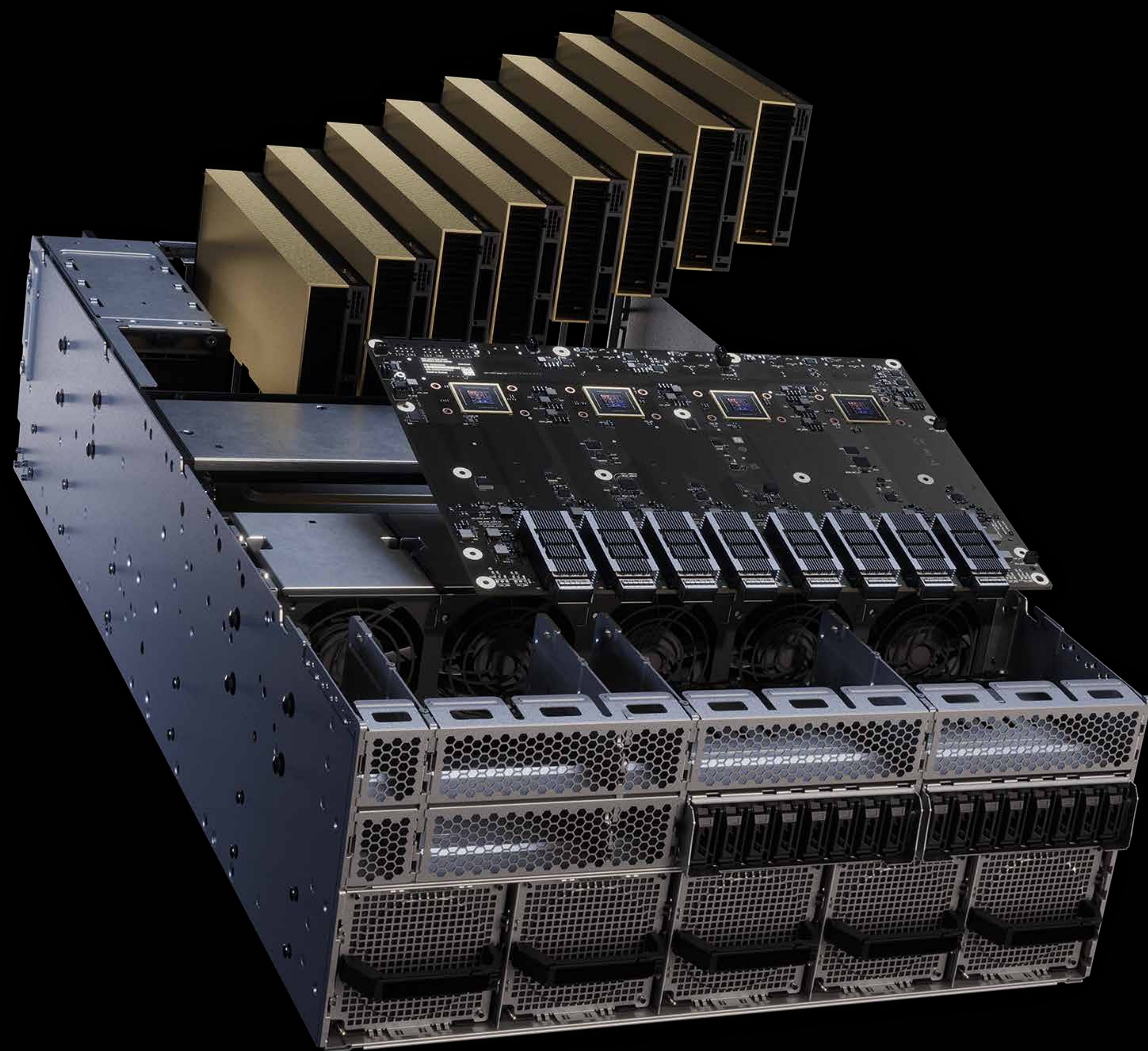


RTX Storage Servers

NEW



Spectrum-X Switch



NVIDIA RTX PRO Server

Enterprise AI and Industrial AI Computing Platform

Available Now



A NEW INDUSTRIAL REVOLUTION

AI FACTORY

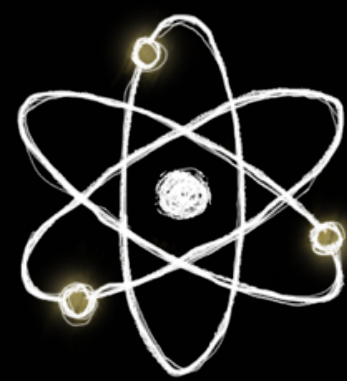
\$100T



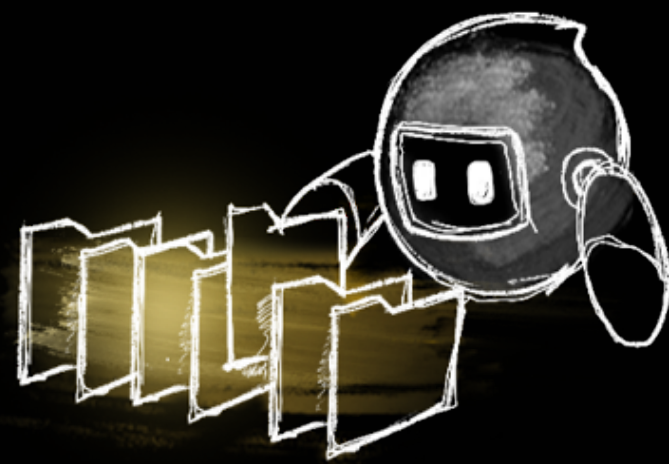
TOKENS
"UNITS OF INTELLIGENCE"

□□□□□□□□□□□□□□

→



SCIENCE



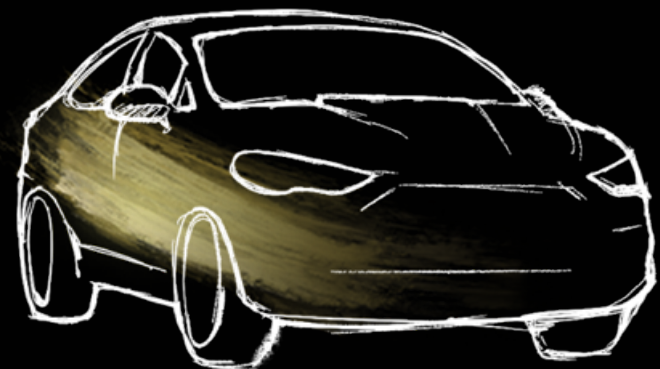
ENTERPRISE



HEALTHCARE



MANUFACTURING



Stargate
200 MW, Scaling to 1.2 GW
Abilene, Texas



Courtesy of Crusoe

European Telcos Building AI Infrastructure With NVIDIA

FASTWEB



European Cloud Service Providers Building AI Infrastructure With NVIDIA

beyond.pl

CUDO
COMPUTE

DataCrunch

OUTSCALE

Domyn

firebird

Fluidstack

G CORE

Mistral AI

NEBIUS

nebul

net zero compute

NEXUS
CORE SYSTEMS

NORTHERN
DATA GROUP

NSCALE

OVHcloud

Polarise.

Scaleway

SESTERCE



European Supercomputing Centers Building AI Infrastructure With NVIDIA



European Union Building 20 AI Factories

Barcelona Supercomputing Center (Spain)

CINECA (Italy)

CSC (Finland)

GENCI/CEA (France)

GRNET (Greece)

Jülich Supercomputing Center (Germany)

Linköping University (Sweden)

LuxProvide (Luxembourg)

Poznan Supercomputing Center (Poland)

Sofia Tech Park (Bulgaria)

University of Maribor (Slovenia)

University of Stuttgart (Germany)

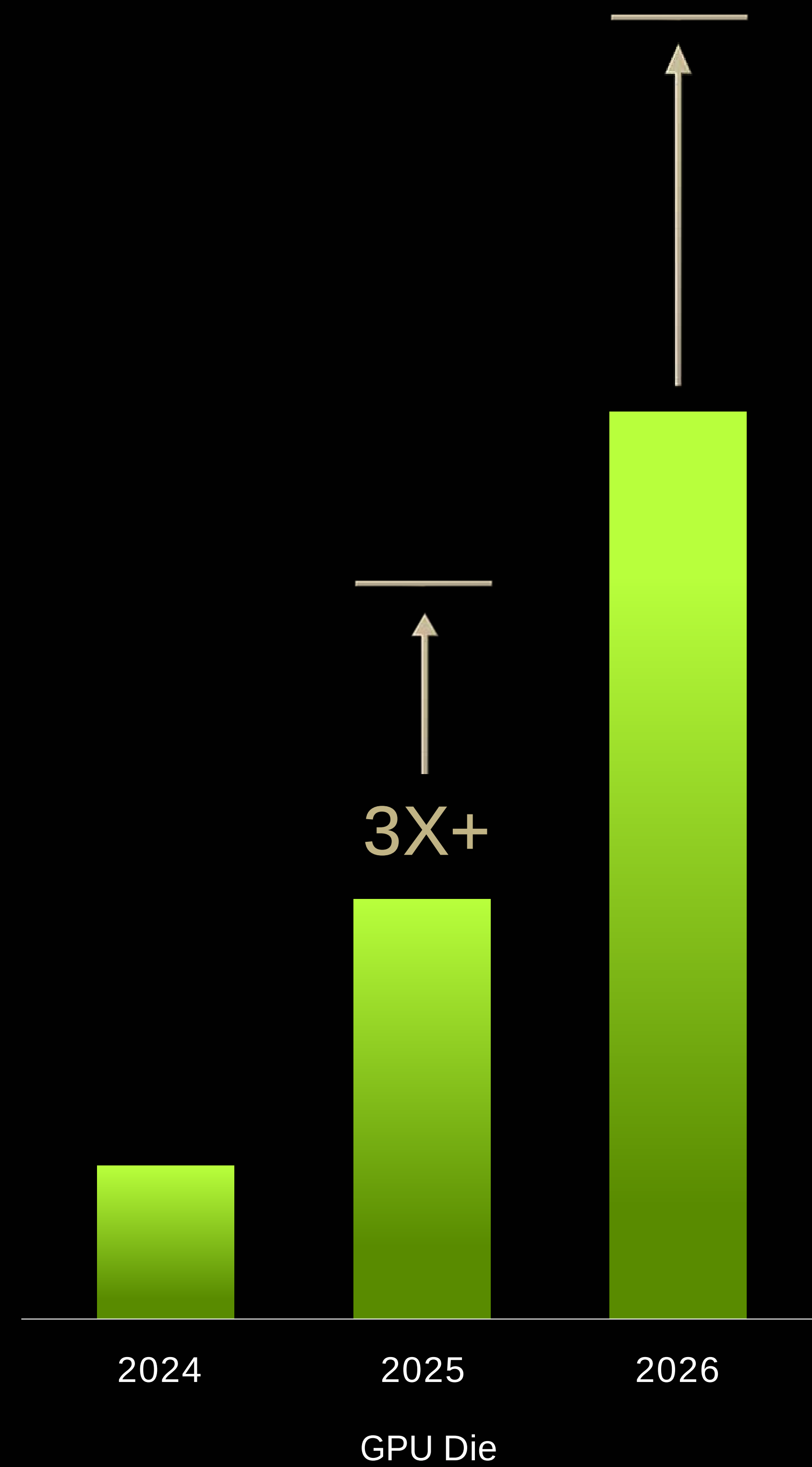
Vienna Scientific Cluster (Austria)

2 AI Factories (Location TBD)

5 Gigafactories (Location TBD)

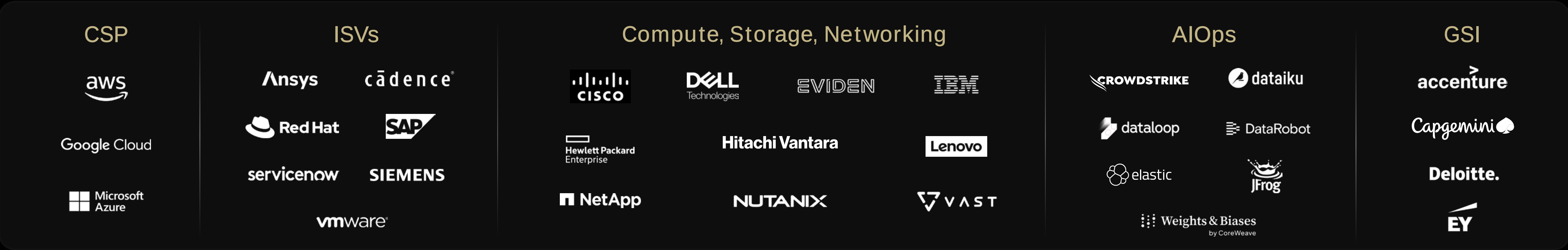
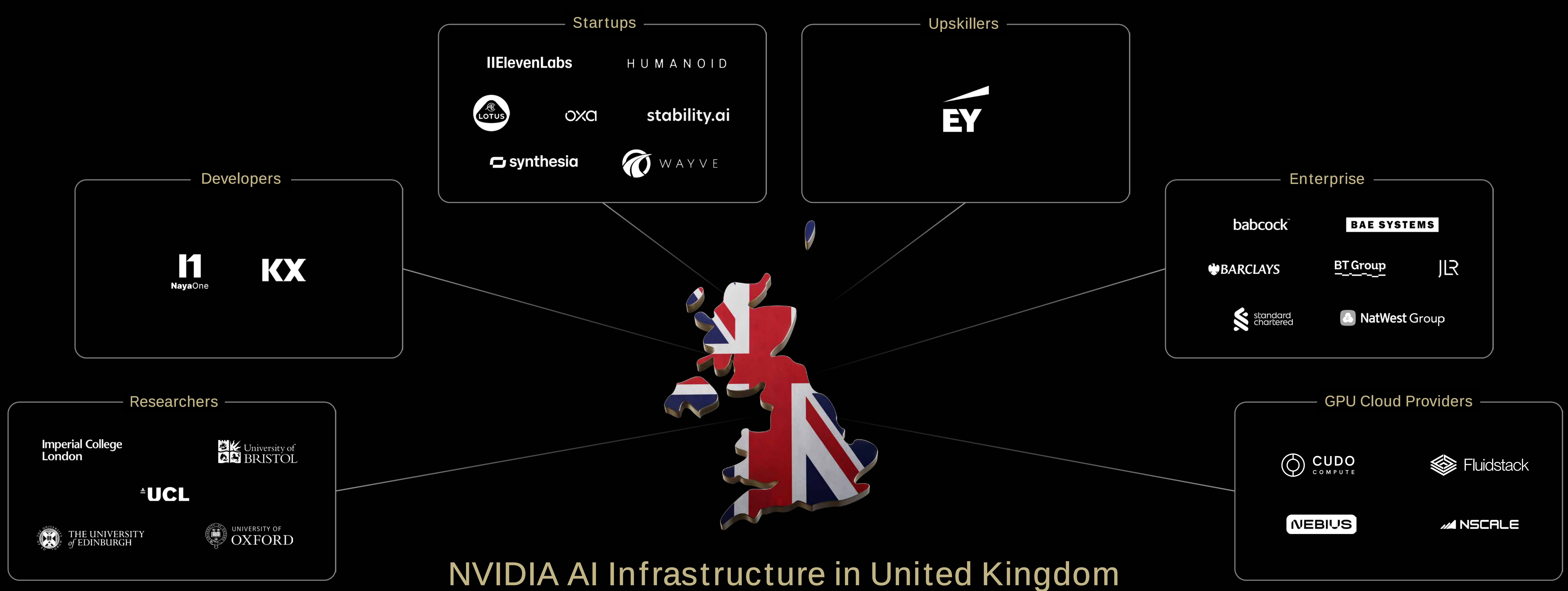


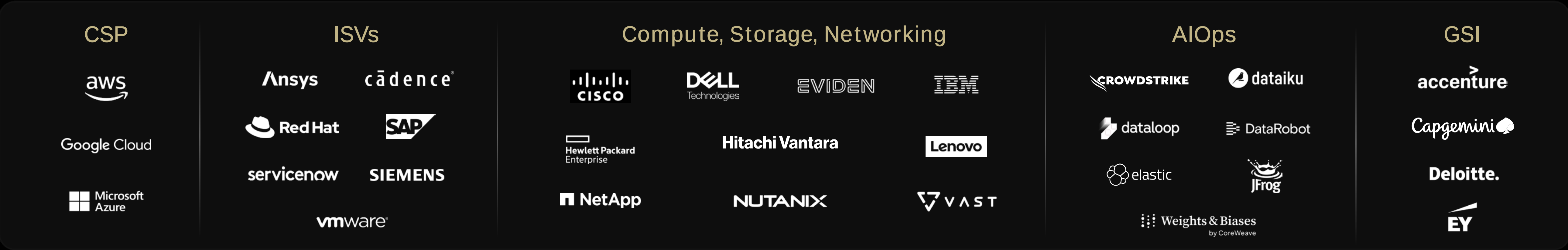
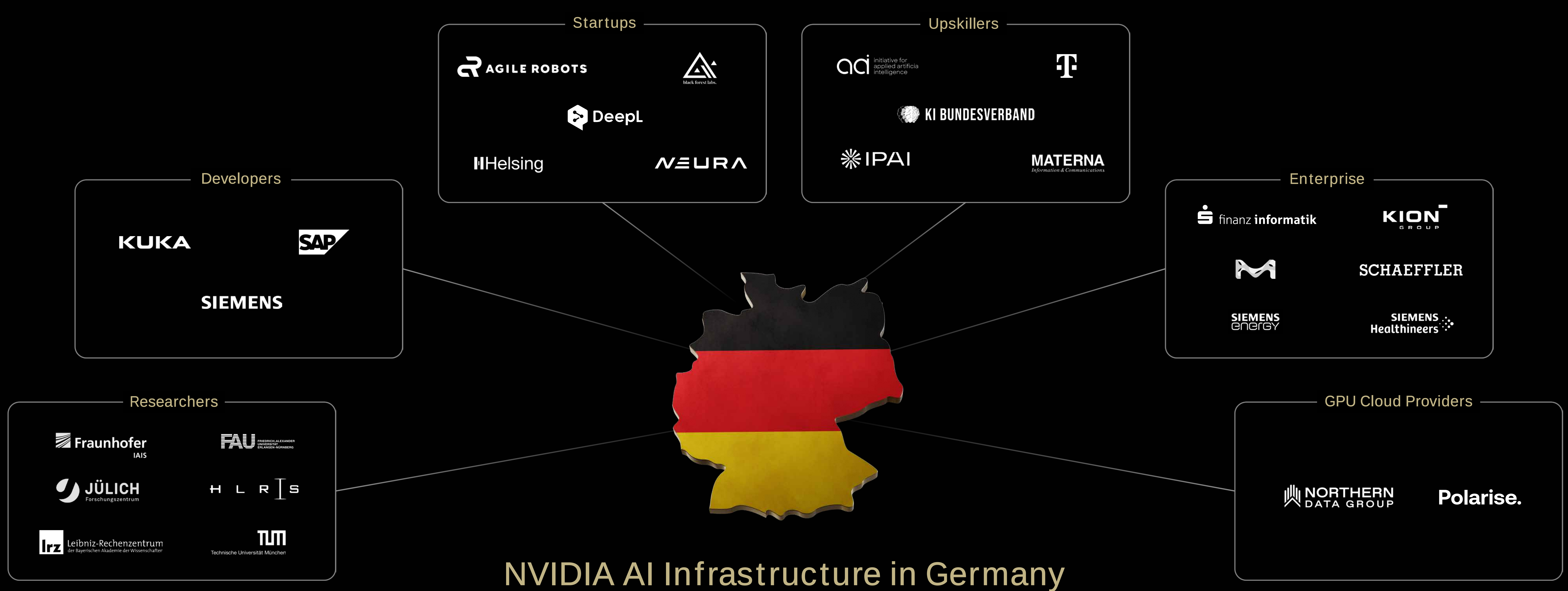
10X Growth Projection Pan-European AI Infrastructure



**NVIDIA Establishes
European AI Technology Centers
to Grow Regional Research and
Startup Ecosystem**

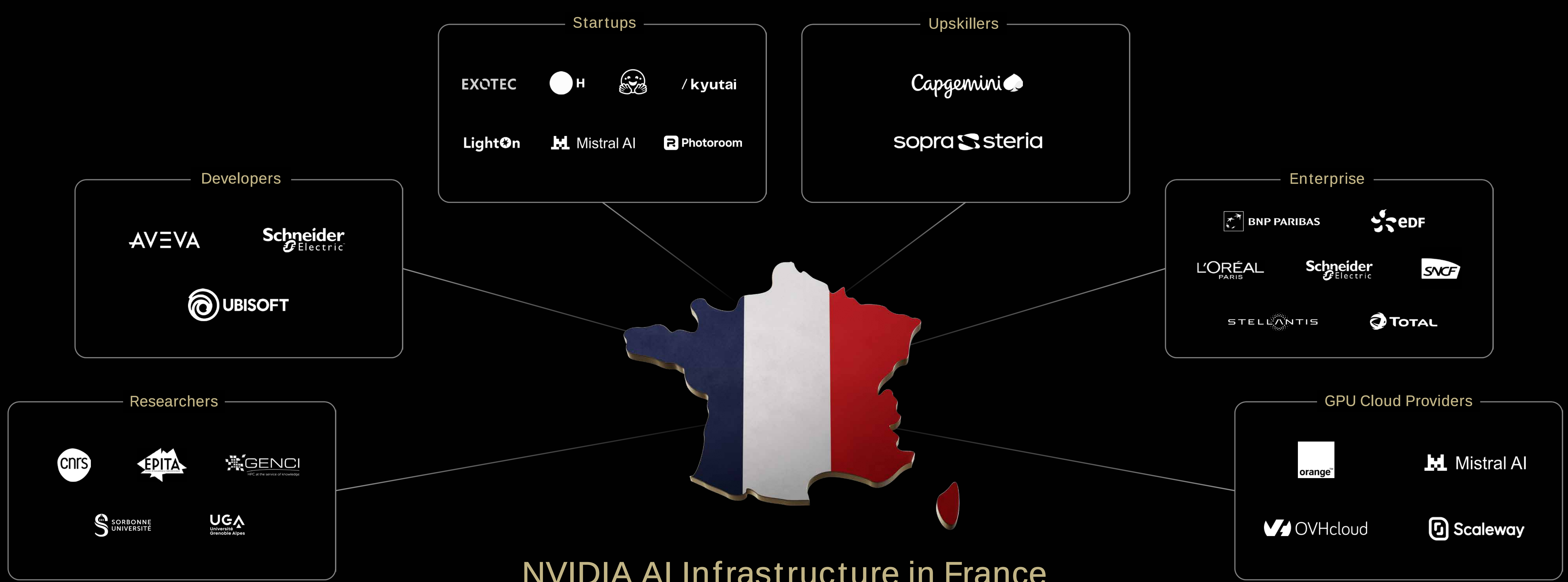




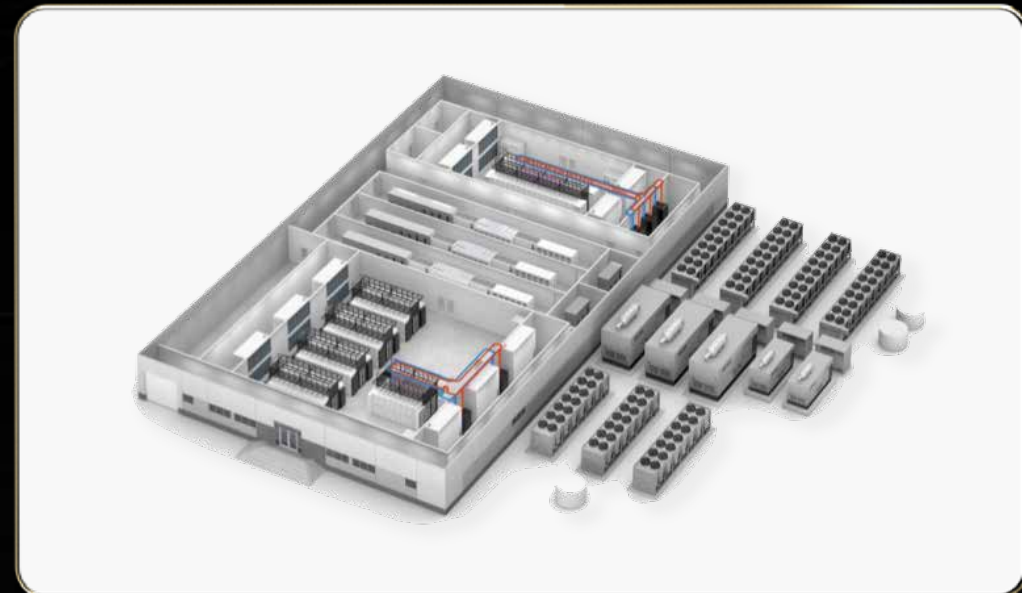




CSP	ISVs	Compute, Storage, Networking	AIOps	GSI
aws	Ansys	cisco	CROWDSTRIKE	accenture
Google Cloud	cadence	DELL Technologies	dataiku	Capgemini
Microsoft Azure	Red Hat	EVIDEN	dataloop	Deloitte.
	SAP	IBM	DataRobot	EY
	servicenow	Hitachi Vantara	elastic	
	SIEMENS	Lenovo	Jfrog	
	vmware	NetApp	Weights & Biases by CoreWeave	
		NUTANIX		
		VAST		



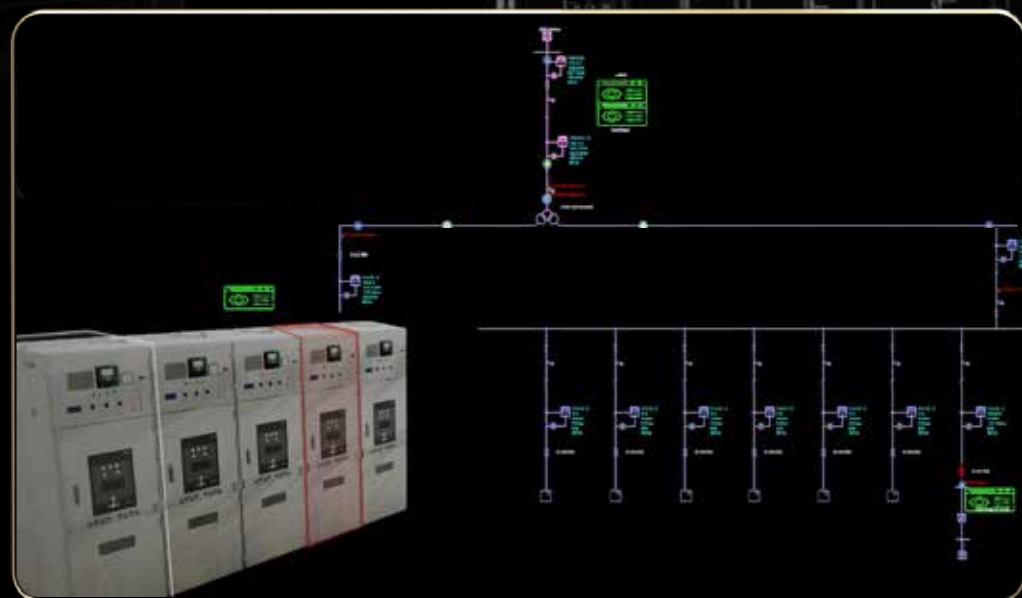
CSP	ISVs	Compute, Storage, Networking			AIOps	GSI
aws	Ansys	CISCO	DELL Technologies	EVIDEN	CROWDSTRIKE	accenture
Google Cloud	Red Hat	Hewlett Packard Enterprise	Hitachi Vantara	Lenovo	dataiku	Capgemini
Microsoft Azure	SAP	NetApp	NUTANIX	VAST	dataloop	Deloitte.
	servicenow				DataRobot	EY
	SIEMENS				elastic	
	vmware				Jfrog	
					Weights & Biases by CoreWeave	



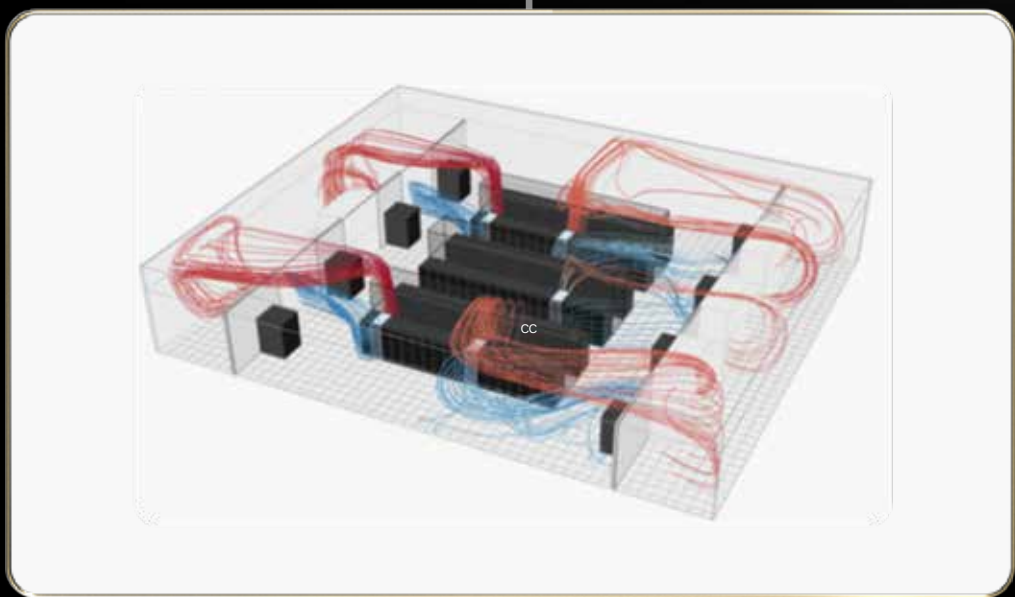
Planning | Concept
Reference Designs



1st Fully Integrated
AI Factory



Electrical Digital
Twin Simulation



EcoStruxure IT Design CFD



Build

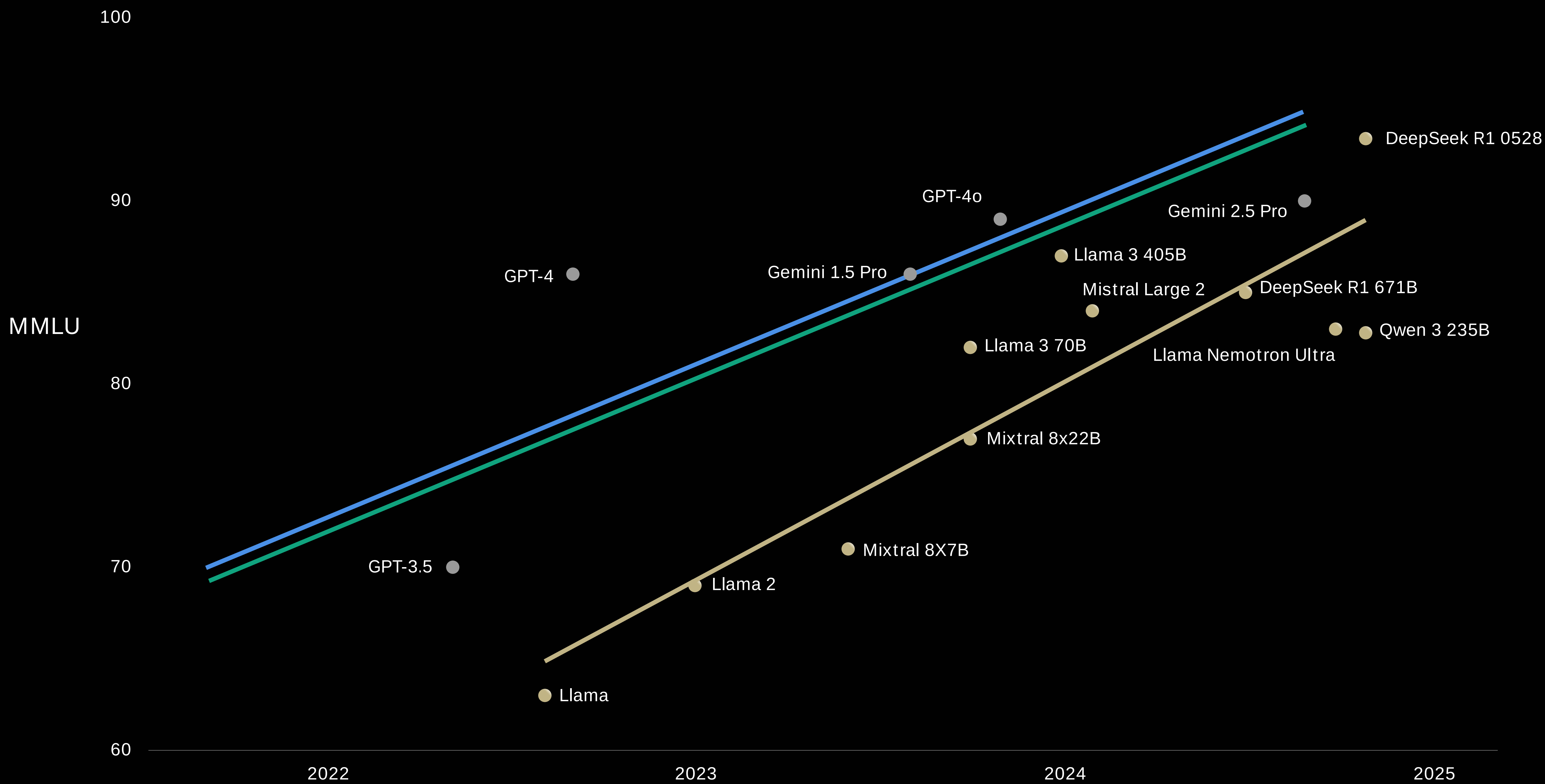


Next Gen
DC Architectures

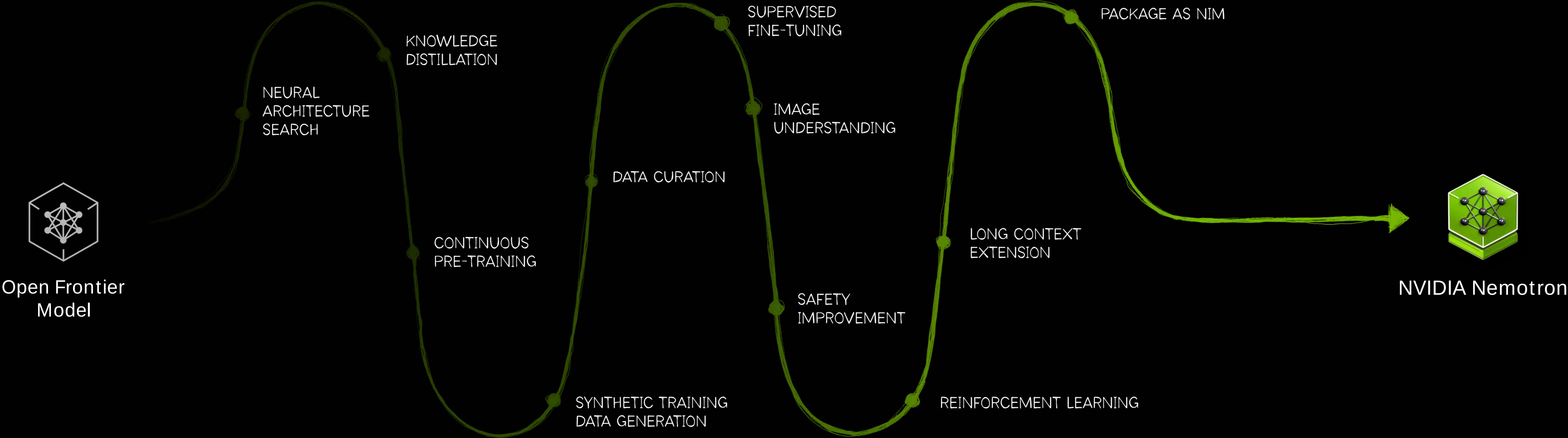


Announcing Mistral and NVIDIA Build AI Cloud

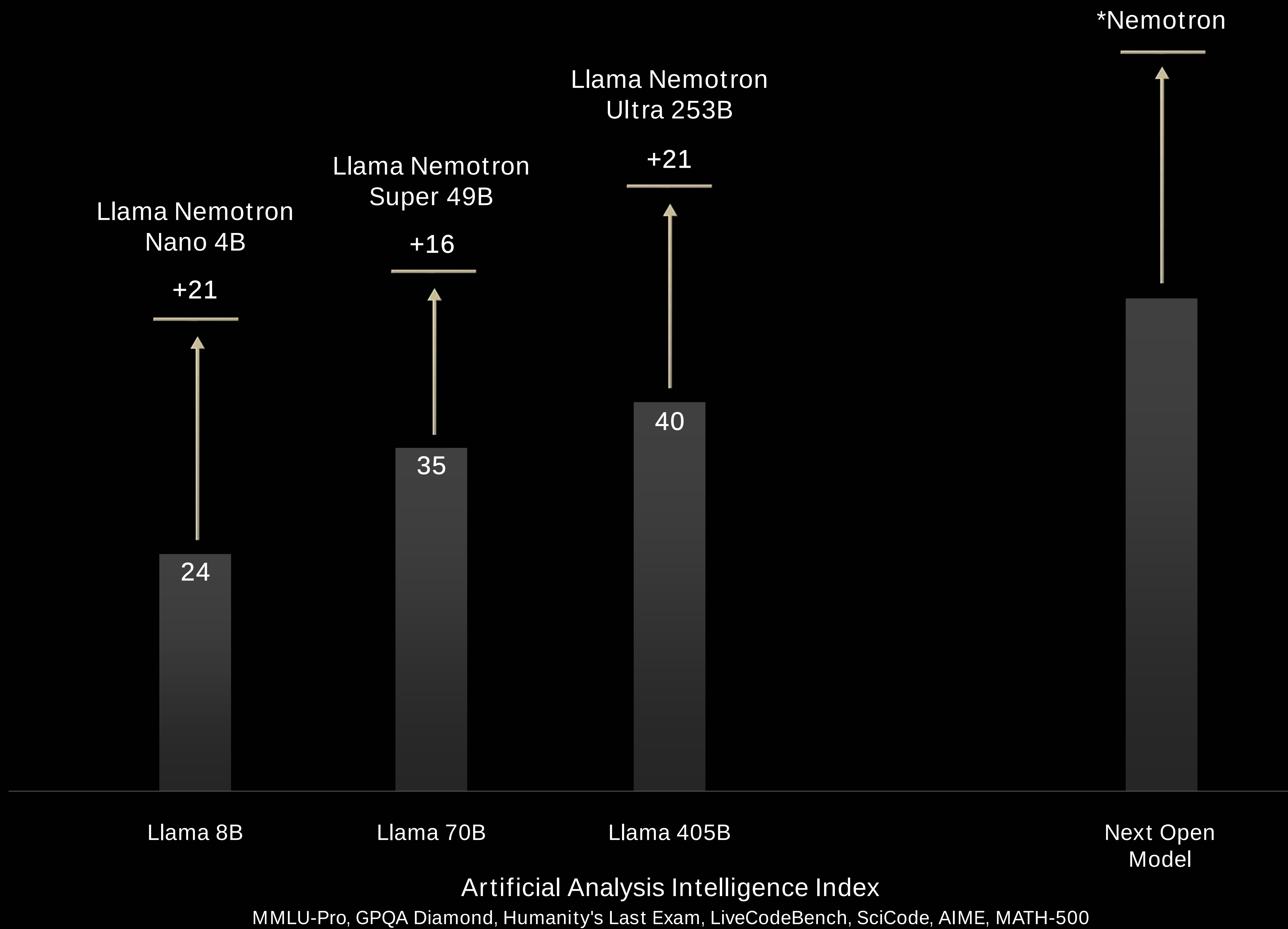
Open-Source AI Models Are Unlocking Global AI Innovation



NVIDIA Nemotron Further Advances Leading Open Models



NVIDIA Nemotron Increases Open Model Efficiency and Accuracy



NVIDIA Nemotron Achieves World-Class Benchmarks



Datasets

Llama Nemotron Post-Training
Trending Reasoning Dataset

OpenCode
Trending Coding Dataset

Nemotron Personas
Represents Real-World Demographics



Models

Llama Nemotron Ultra
Leading GPQA-Diamond Accuracy

Llama Nemotron Nano
Math 500

Llama Nemotron Vision
OCRBench V2

AceReasoning Nemotron
Math, Coding SML

Nemotron-H
Fastest Inference LLM With
Leading Reasoning Accuracy



Agents

Nemotron-CORTEXA
Leading SWE-Bench Accuracy

Model Builders Across Europe

Creating Sovereign LLMs with NVIDIA Nemotron



Perplexity Makes European Nemotron Models Available in Region



Create anything...



Best

Selects the best model for each query

BSC

Spanish and Catalan Nemotron

Dicta

Hebrew Nemotron

Domyn

Italian Nemotron

Fraunhofer OpenGPT-X

Pan European Nemotron

UTTER EuroLLM

Pan European Nemotron

LightOn

French Nemotron

Slovak Republic

Slovak Nemotron

SpeakLeash

Polish Nemotron

TII

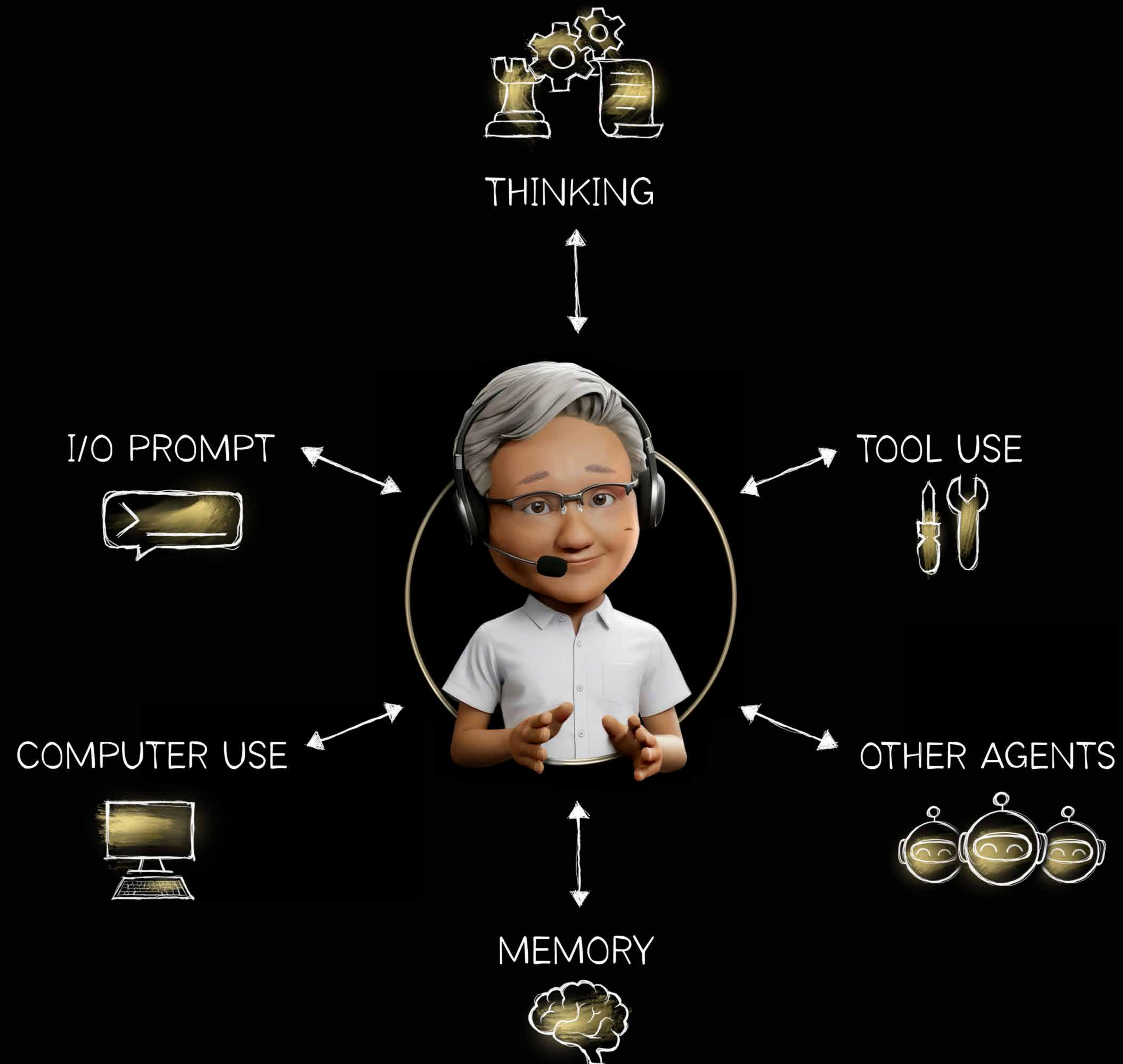
Arabic Nemotron

University College London

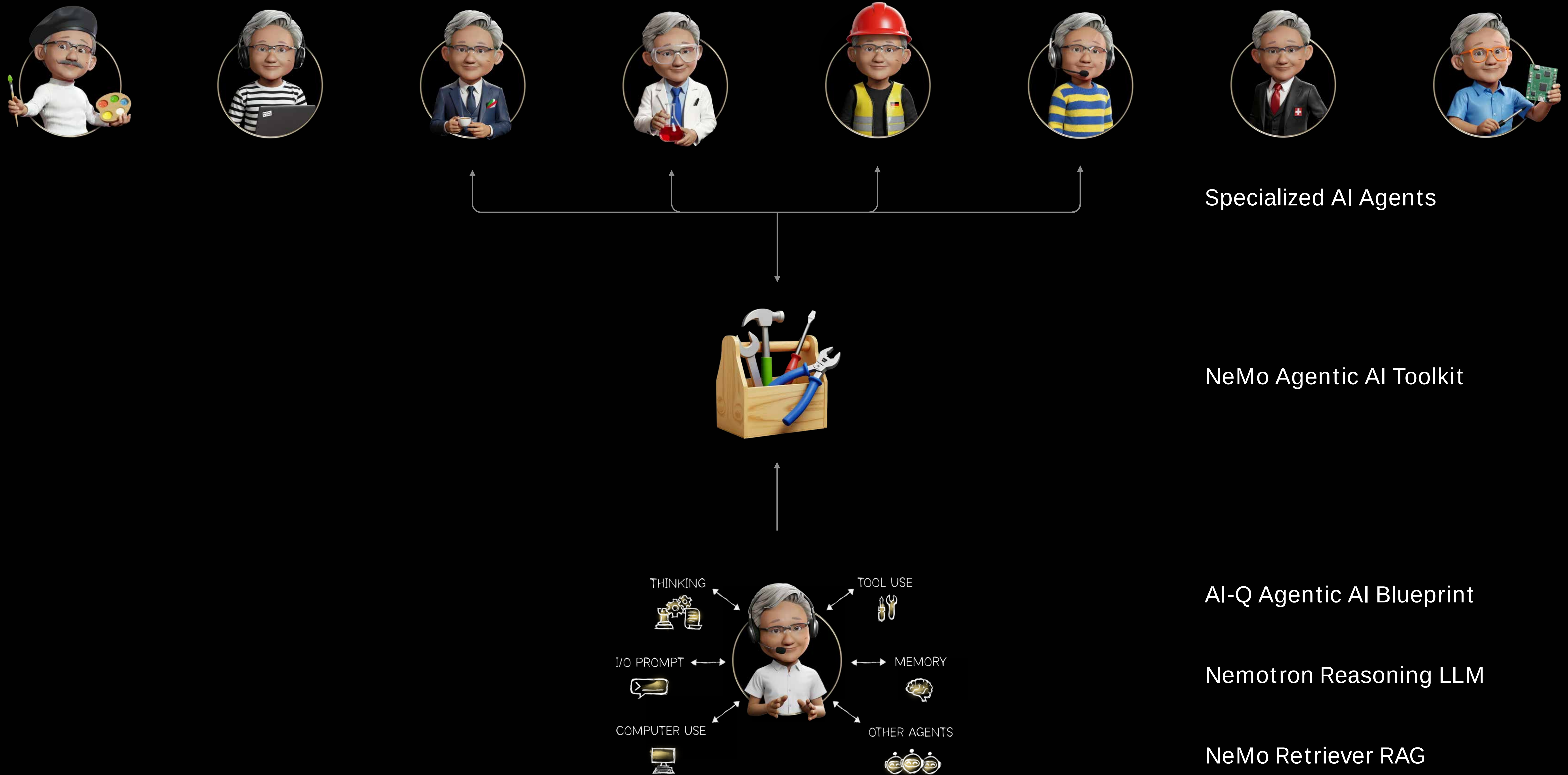
BritLLM Nemotron

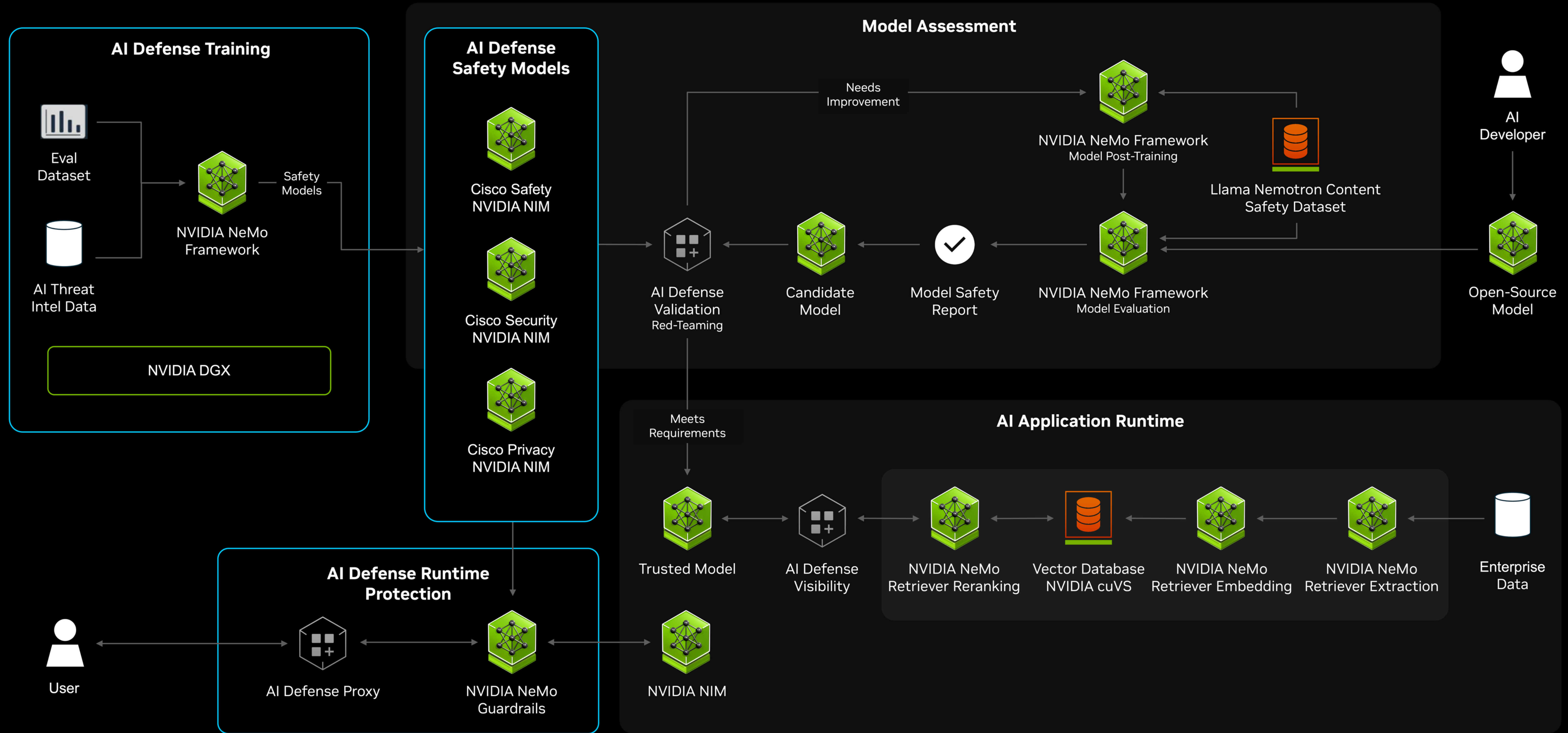
University of Ljubljana

Slovene Nemotron



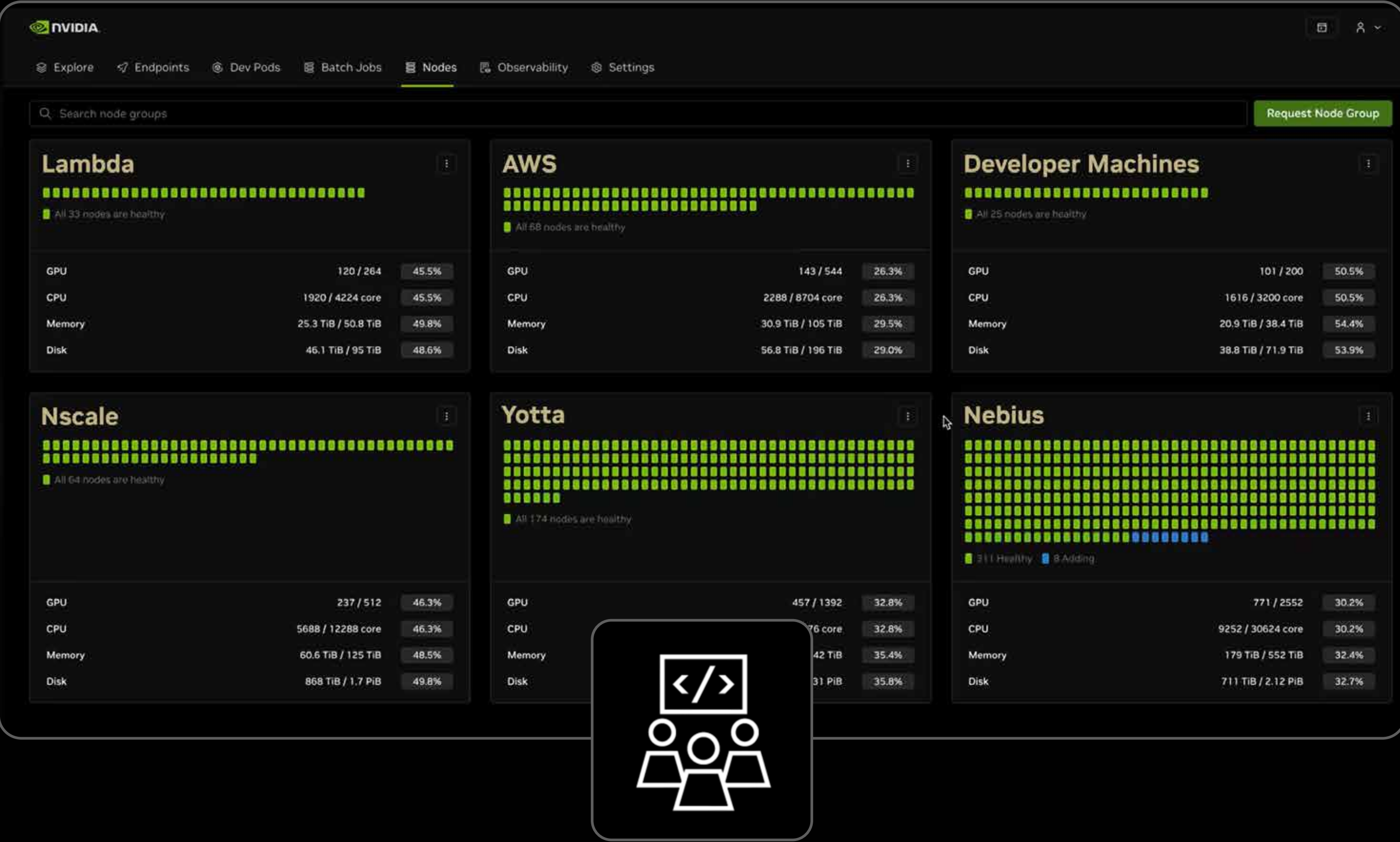
NVIDIA Enterprise AI Agent Platform





Connecting Developers to Global AI Compute

Multi-Cloud | Multi-Tenant | Scalable Training | Global Inference



DGX Cloud Lepton

Global Cloud Providers

aws Microsoft Azure

GPU Cloud Providers

CoreWeave Crusoe firebird Lambda Mistral AI

NEBIUS NSCALE Scaleway together.ai YOTTA

GPU Marketplaces

Andromeda Fluidstack FOUNDRY

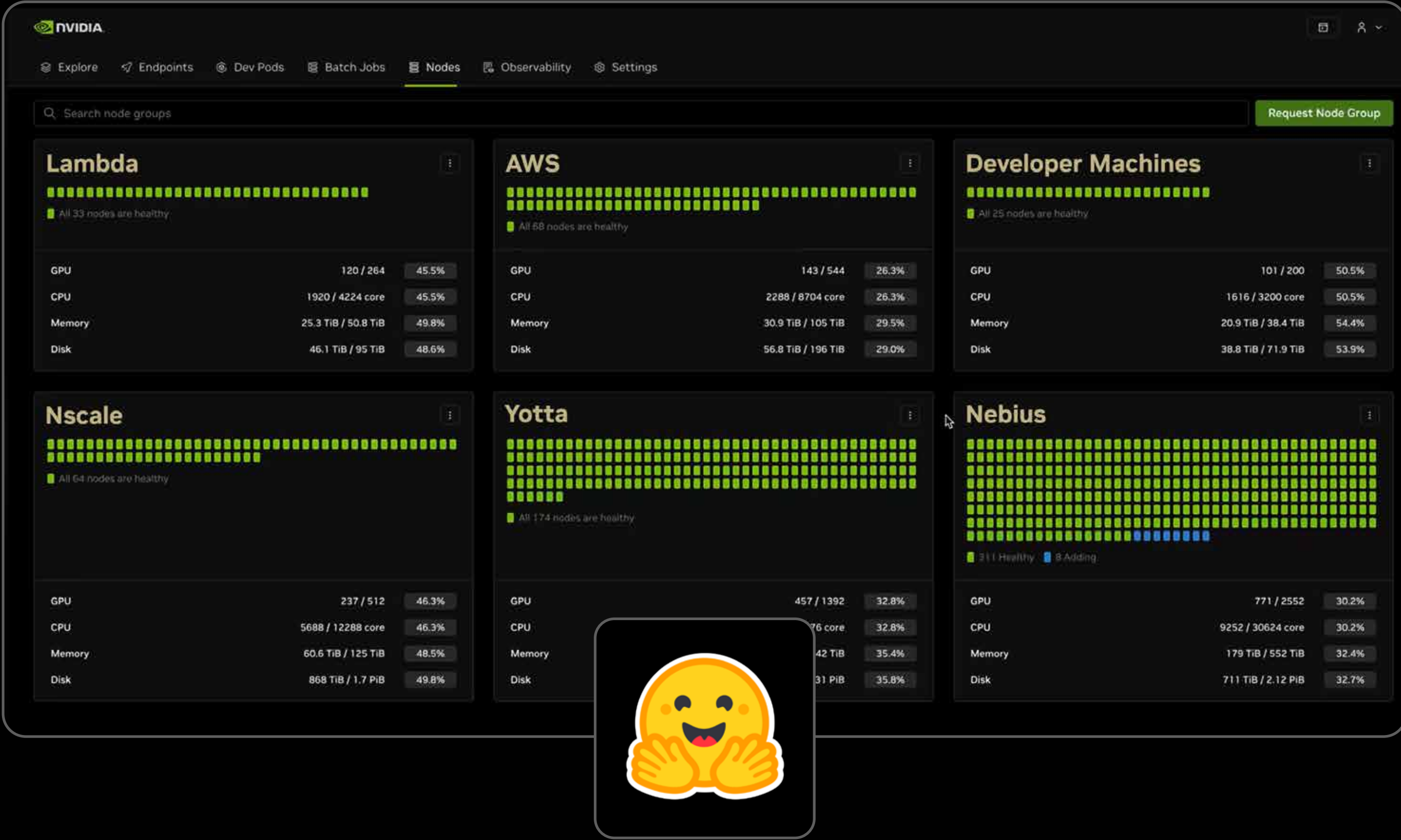
HYDRA The San Francisco Compute Company

Developer Local Environment

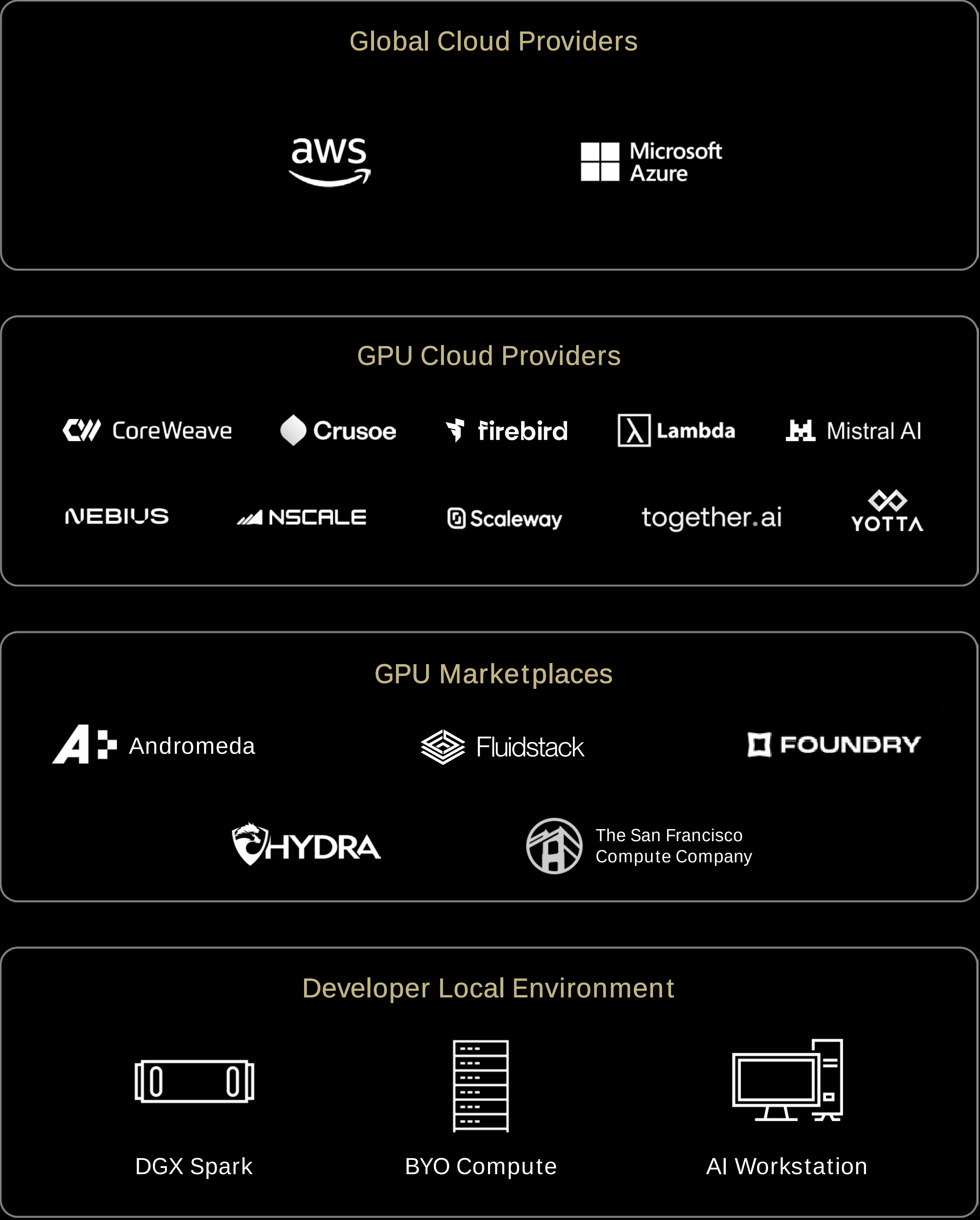
DGX Spark BYO Compute AI Workstation

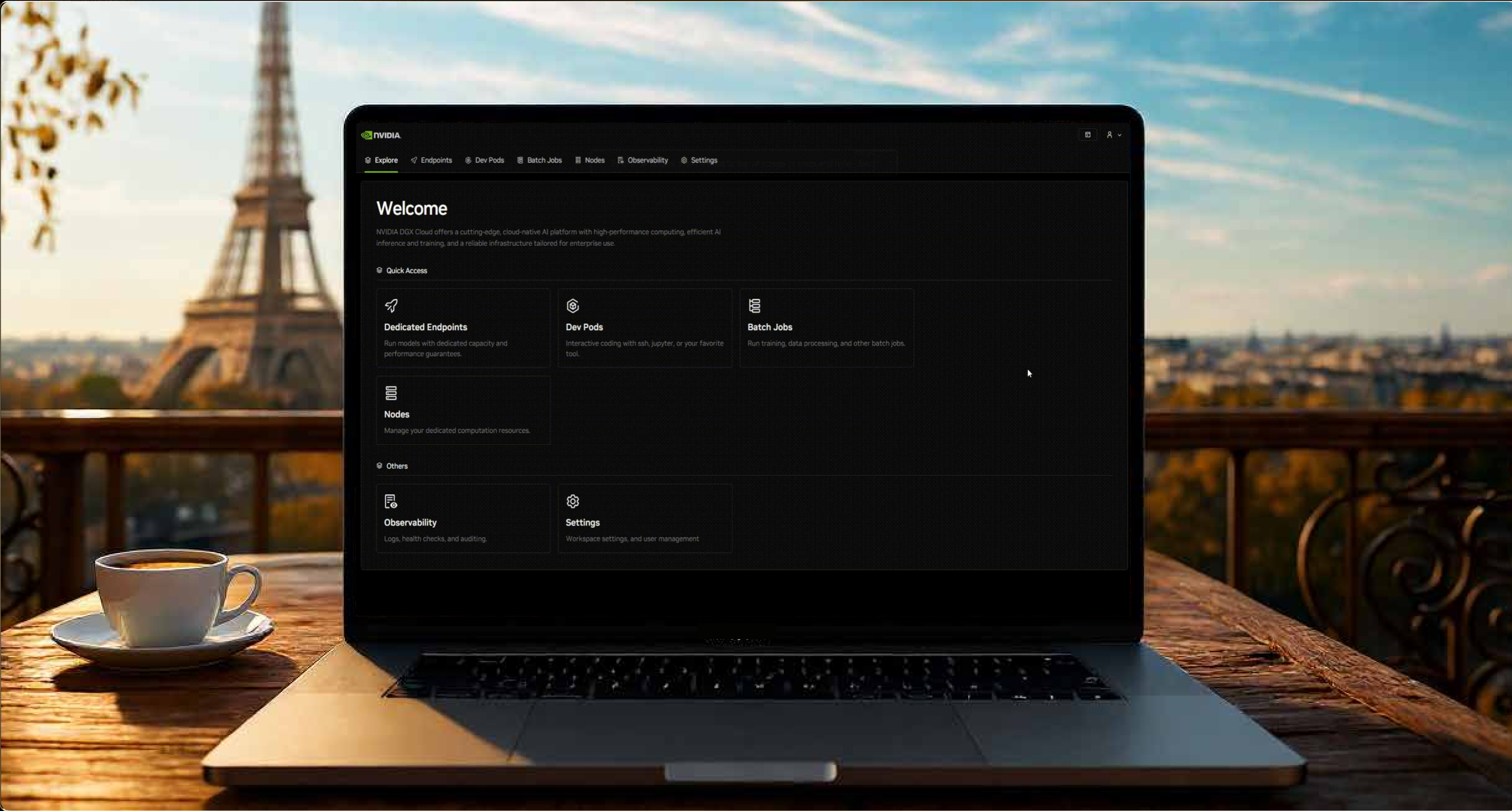
Connecting Developers to Global AI Compute

Multi-Cloud | Multi-Tenant | Scalable Training | Global Inference



DGX Cloud Lepton





 NVIDIA

 Explore

 Endpoints

 Dev Pods

 Batch Jobs

 Nodes

 Observability

 Settings

General settings for your workspace | [Get started](#)

Welcome

NVIDIA DGX Cloud offers a cutting-edge, cloud-native AI platform with high-performance computing, efficient AI inference and training, and a reliable infrastructure tailored for enterprise use.

Quick Access



Dedicated Endpoints

Run models with dedicated capacity and performance guarantees.



Dev Pods

Interactive coding with ssh, jupyter, or your favorite tool.



Batch Jobs

Run training, data processing, and other batch jobs.



Nodes

Manage your dedicated computation resources.

Others



Observability

Logs, health checks, and auditing.



Settings

Workspace settings, and user management



SAP Applications

Joule

Cloud ERP

CRM

Spend

HXM

...



Customer



Partner
Apps

AI Foundation – AI Operating System

Development Tooling

Generative AI
Hub in SAP AI
Launchpad

Libraries

Generative AI Hub

Trust

Data Masking

Content Moderation

Orchestration

Harmonized API

Prompt Registry

Grounding



NVIDIA NeMo Retriever

Access

Serving/Inference Engine



NVIDIA NeMo
Retriever



NVIDIA NIM



Mistral With
NVIDIA NIM

Foundation Models Provided by Partners

Foundation Models Built by SAP

Open Source Foundation Models

SAP HANA Cloud / SAP Business Data Cloud

Cloud ERP Data

CRM Data

Spend Data

HXM Data

...

Vector Engine

Knowledge Graph



User



Sana Main Agent



Llama Nemotron
Plan, Reflect, Evaluate

Uses



Tools



Documents



Meetings



Emails



Indexing



NVIDIA NeMo Retriever
Extraction



NVIDIA NeMo Retriever
Embedding



NVIDIA NeMo Retriever
Reranking



Vector Search



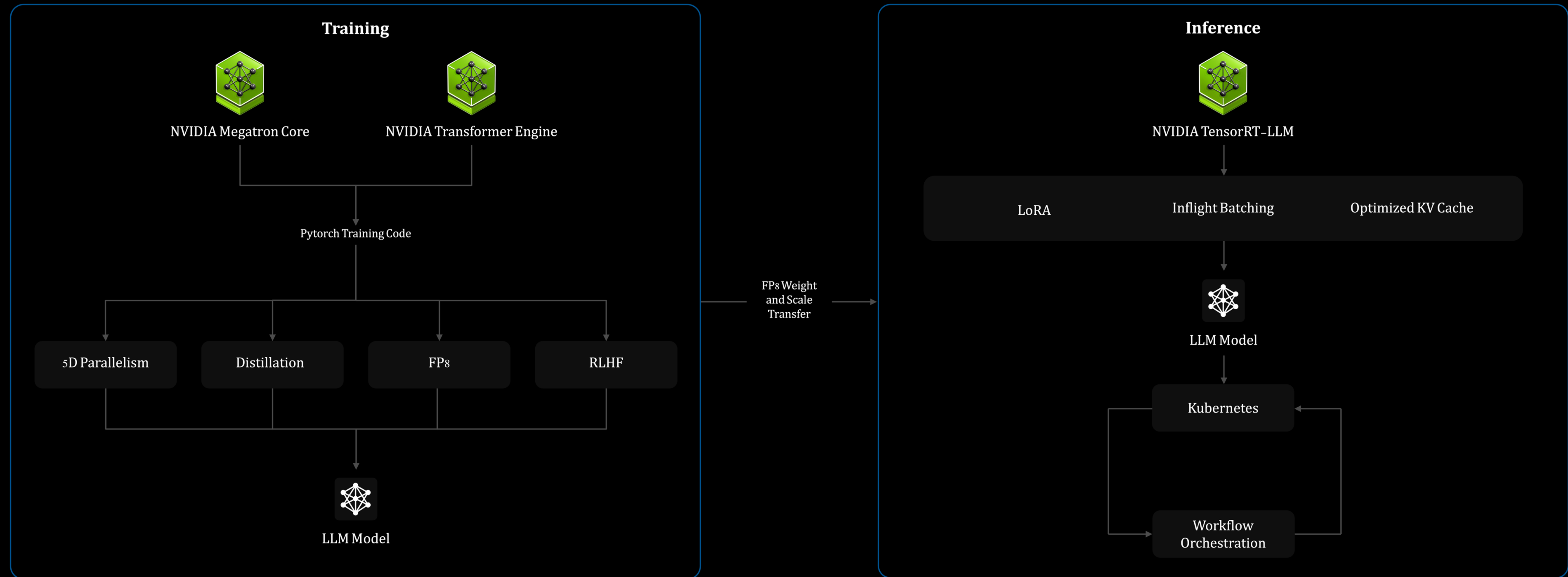
Internet Search



MCP



Third Parties

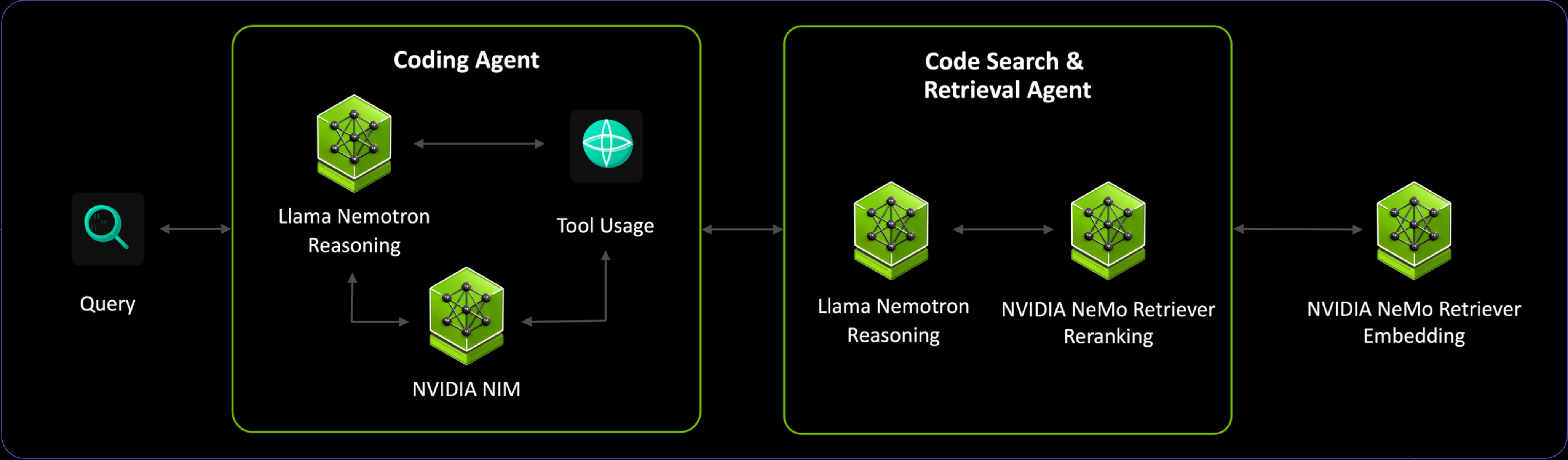






Retrieval & Generation Pipeline

Ingest Pipeline

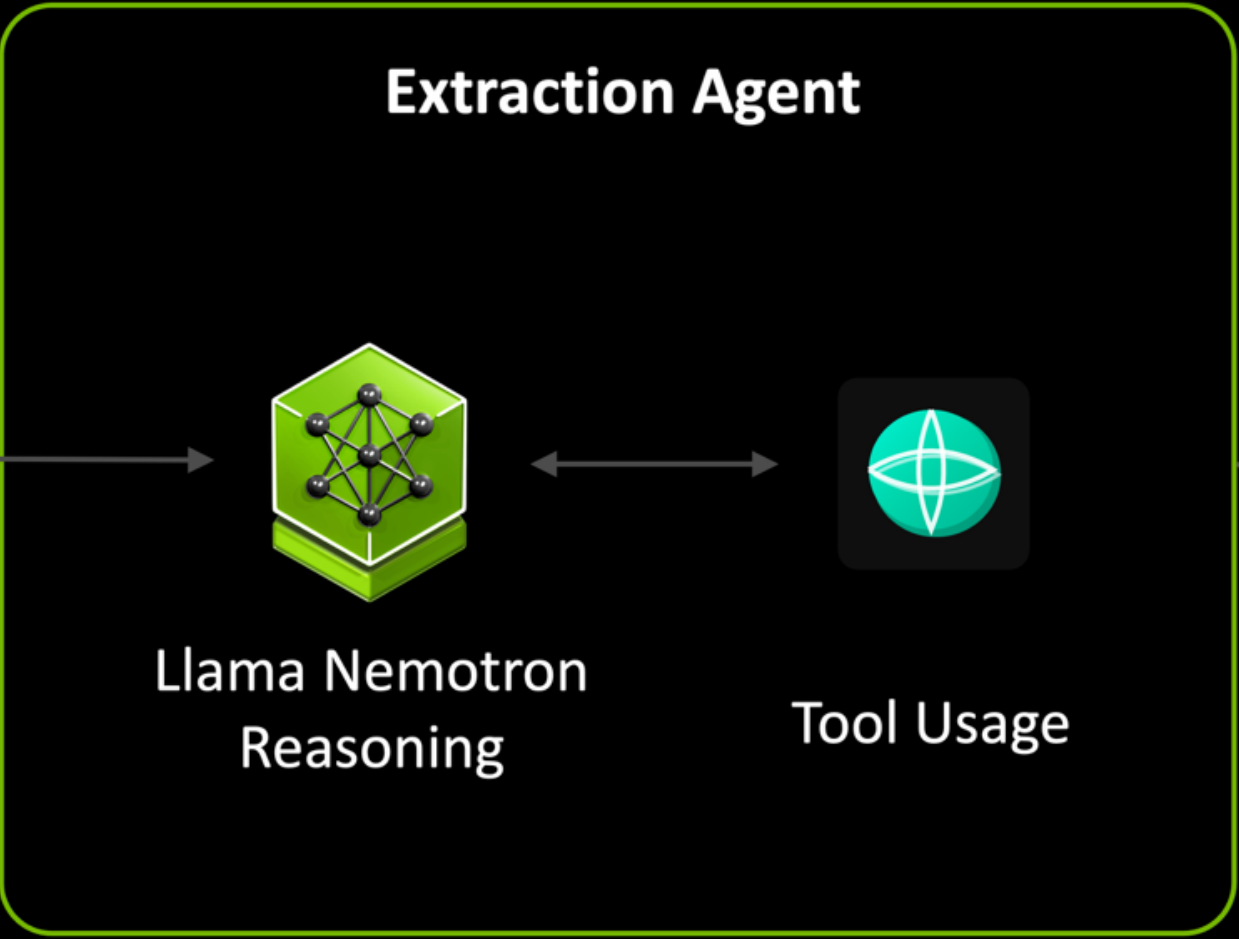


Similarity - "Find similar code to this"
Reference - "Find code using this"
Semantic - "Find code that does X"

Codebase & Software-Related Content

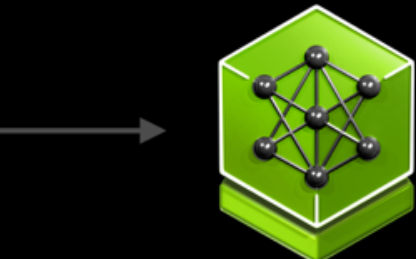


Enterprise Software Files



Extraction Agent

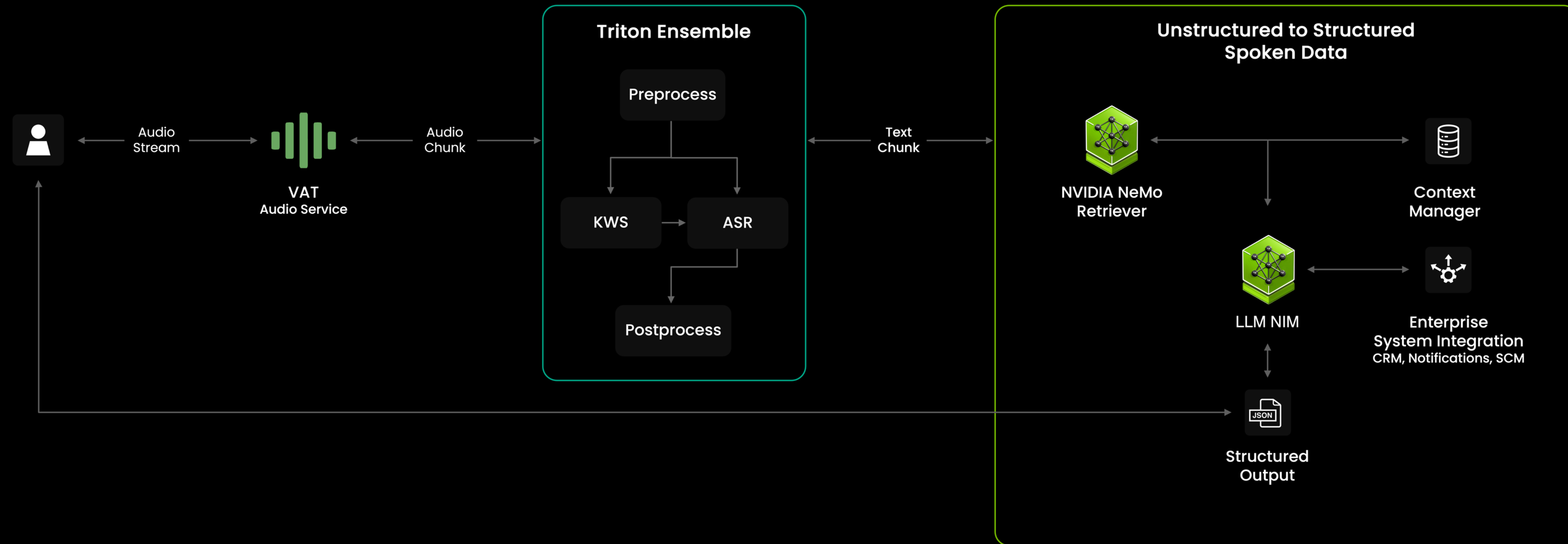
Language-specific Analysis
Class-specific Analysis
Hierarchical Chunking

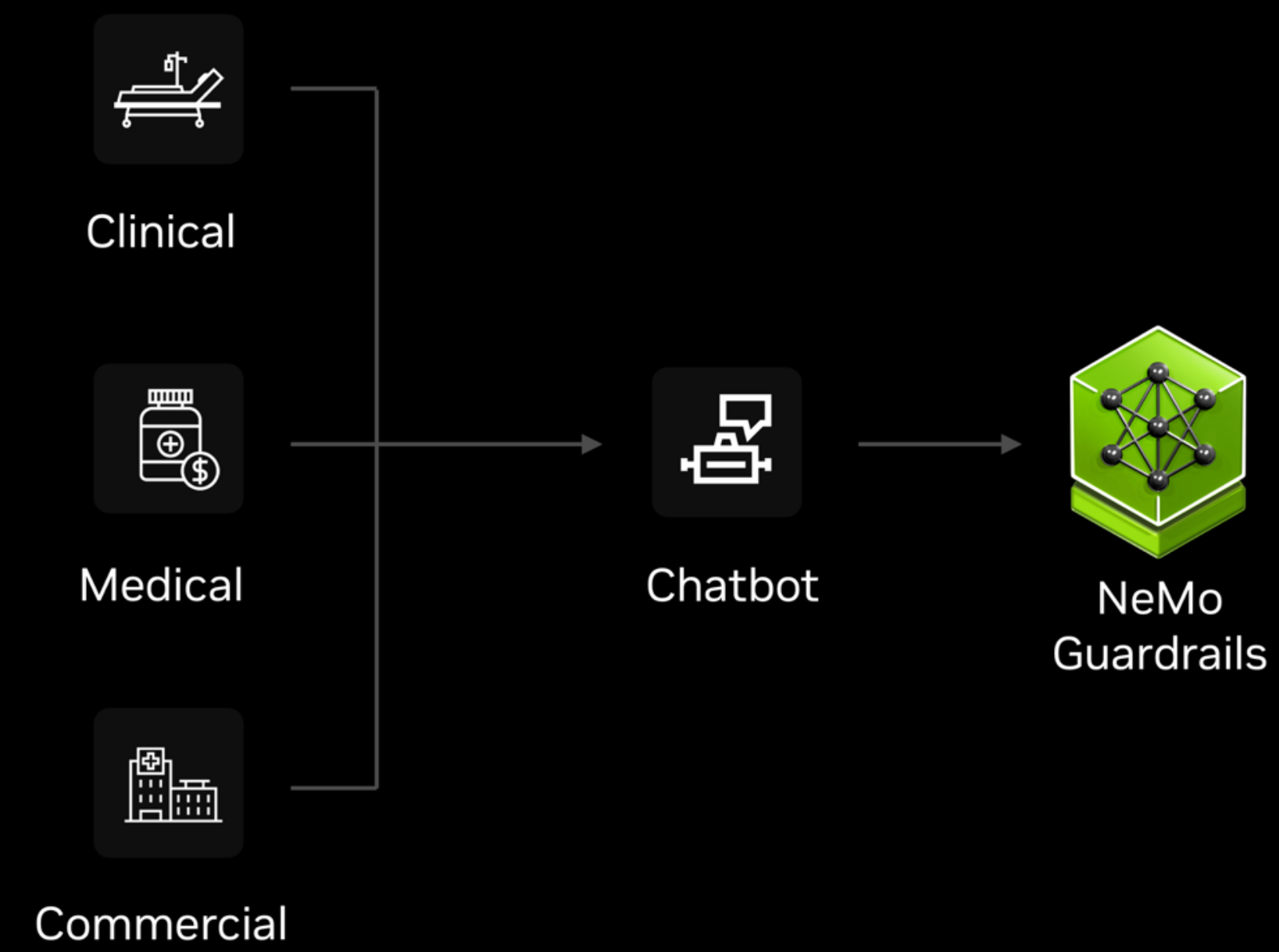


NVIDIA NeMo Retriever Embedding
Qodo-Embed-1

Graph Vector Database

aiOla





Orchestrator Agents

Data Review Agent

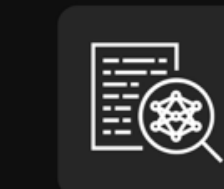
Literature Review Agent

Field Companion Agent

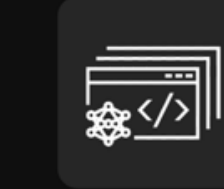
Data Agents



Data Tools



Retriever



Data Integration



RLHF

Data Sources



IQVIA



Public



Client



Llama Nemotron



NVIDIA NeMo Retriever



NeMo Customizer

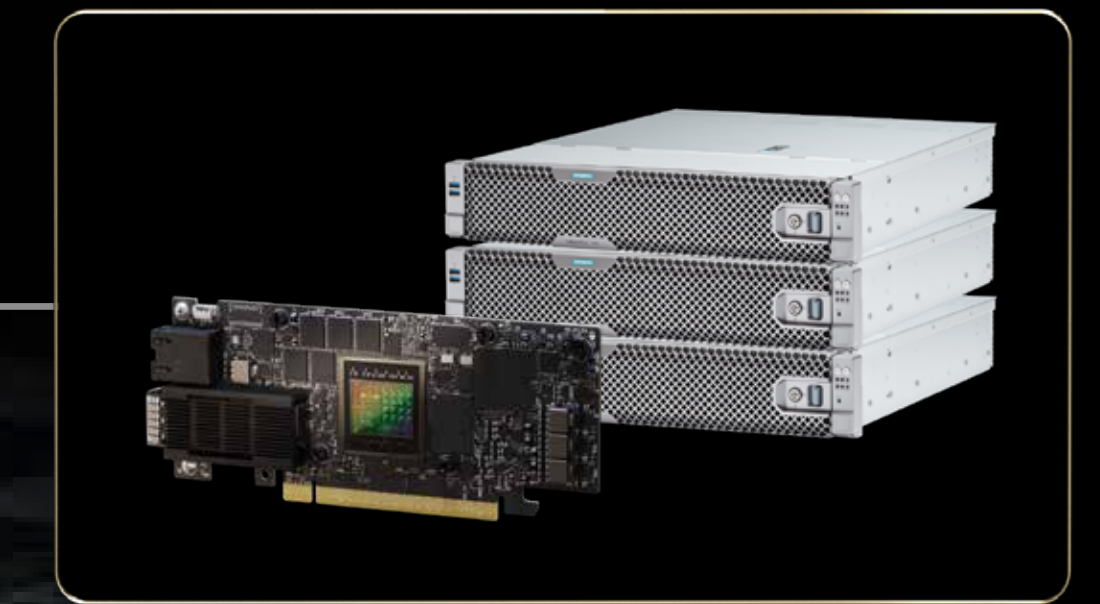
SIEMENS



SIEMENS



Product Design
and Engineering



Secure Edge
Industrial AI



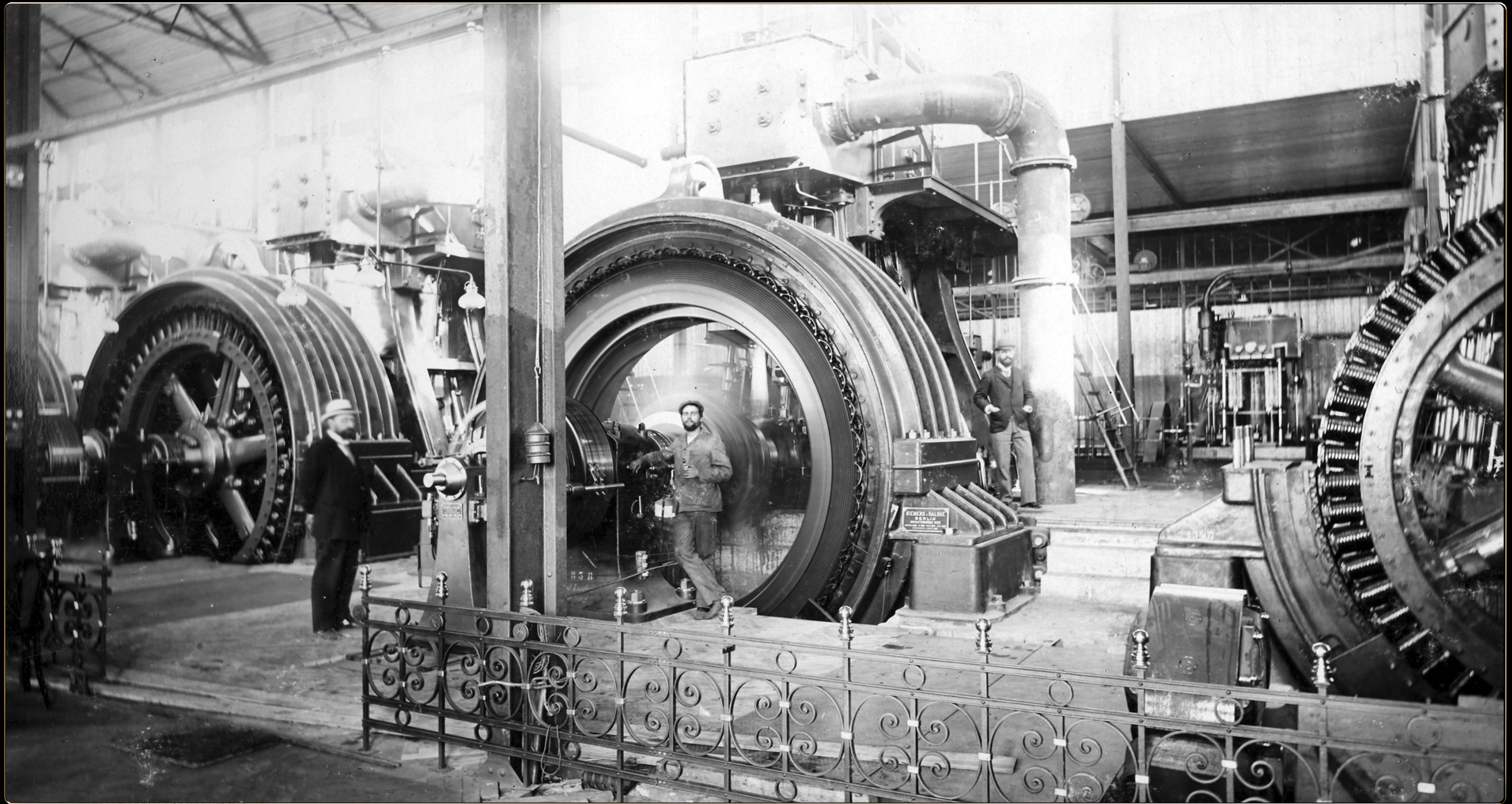
Operational Planning
and Optimization

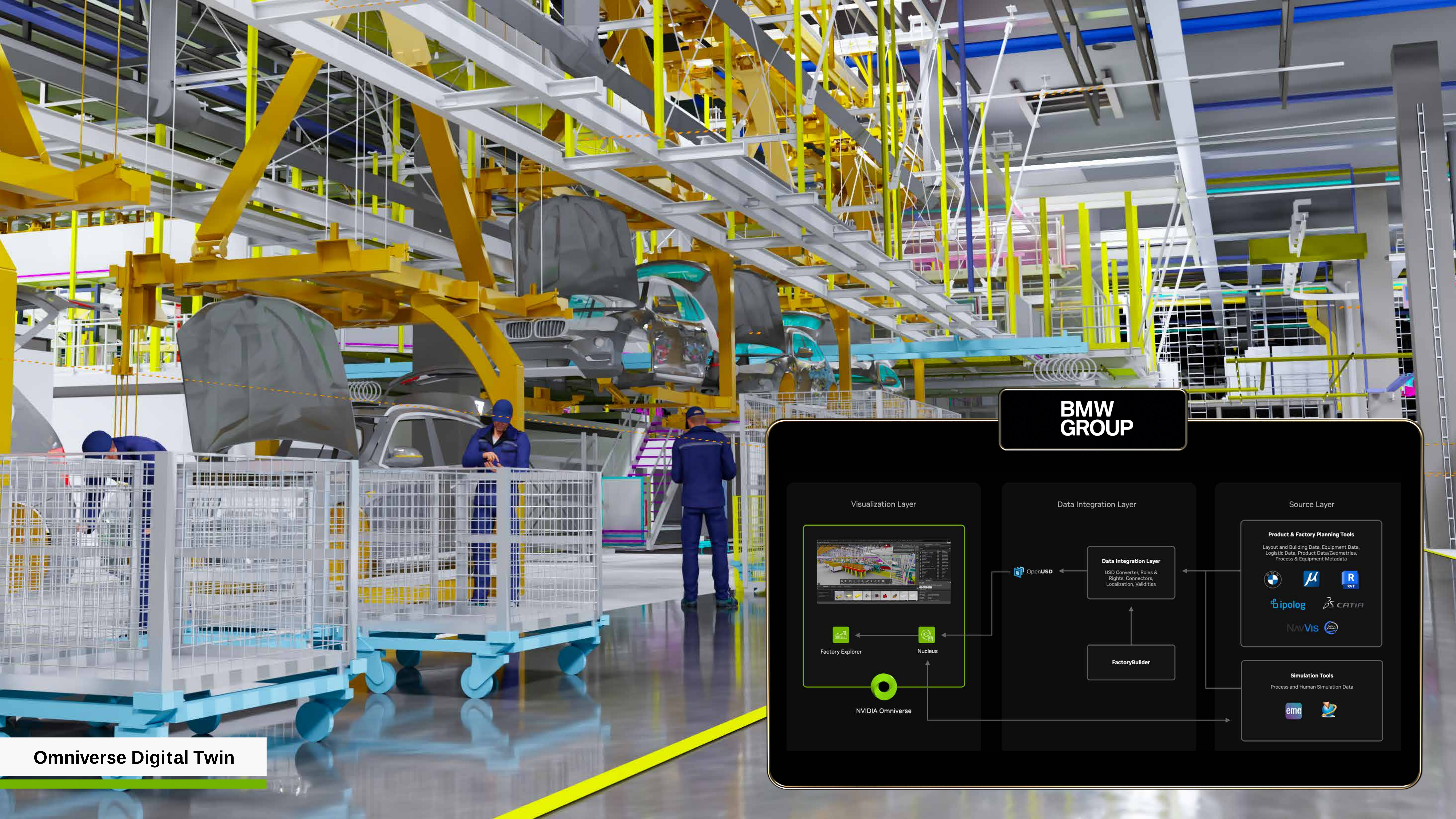


Comprehensive Product
and Factory Digital Twin



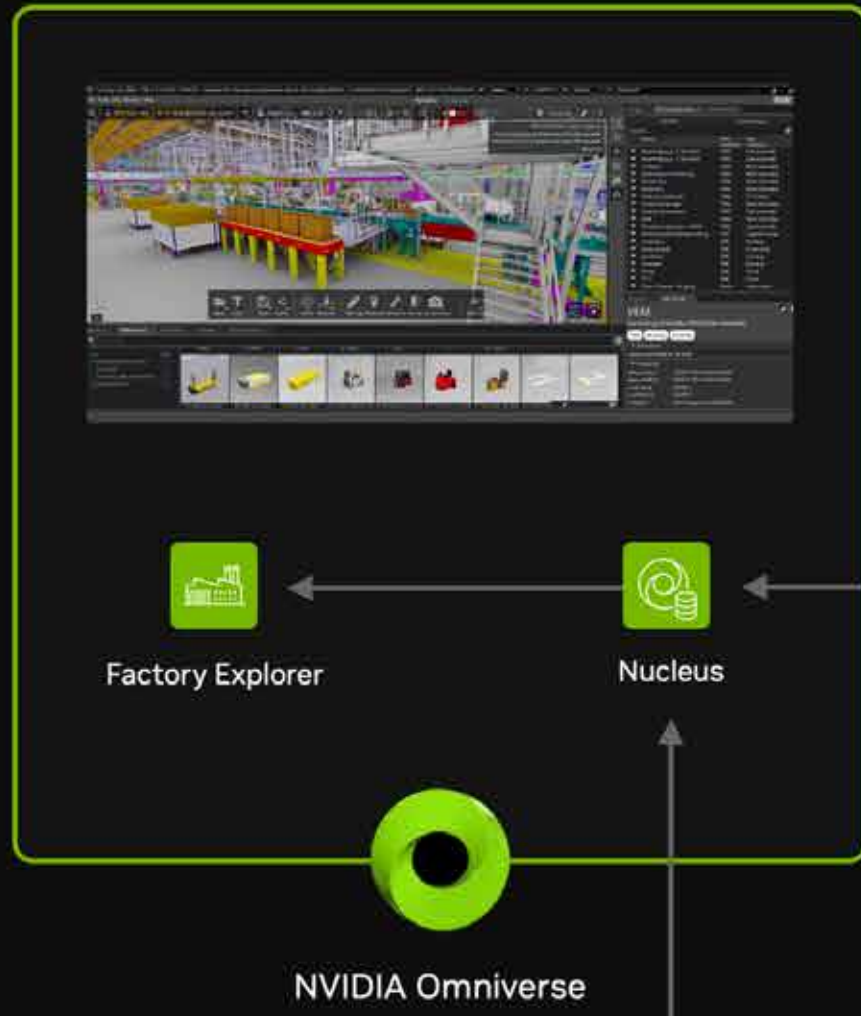
Production
Optimization



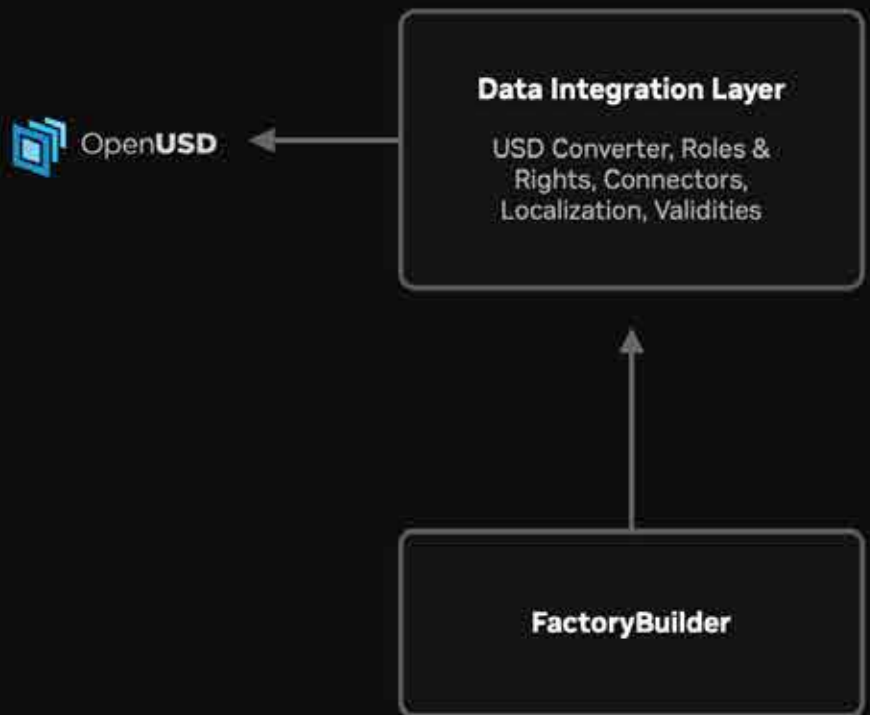


BMW
GROUP

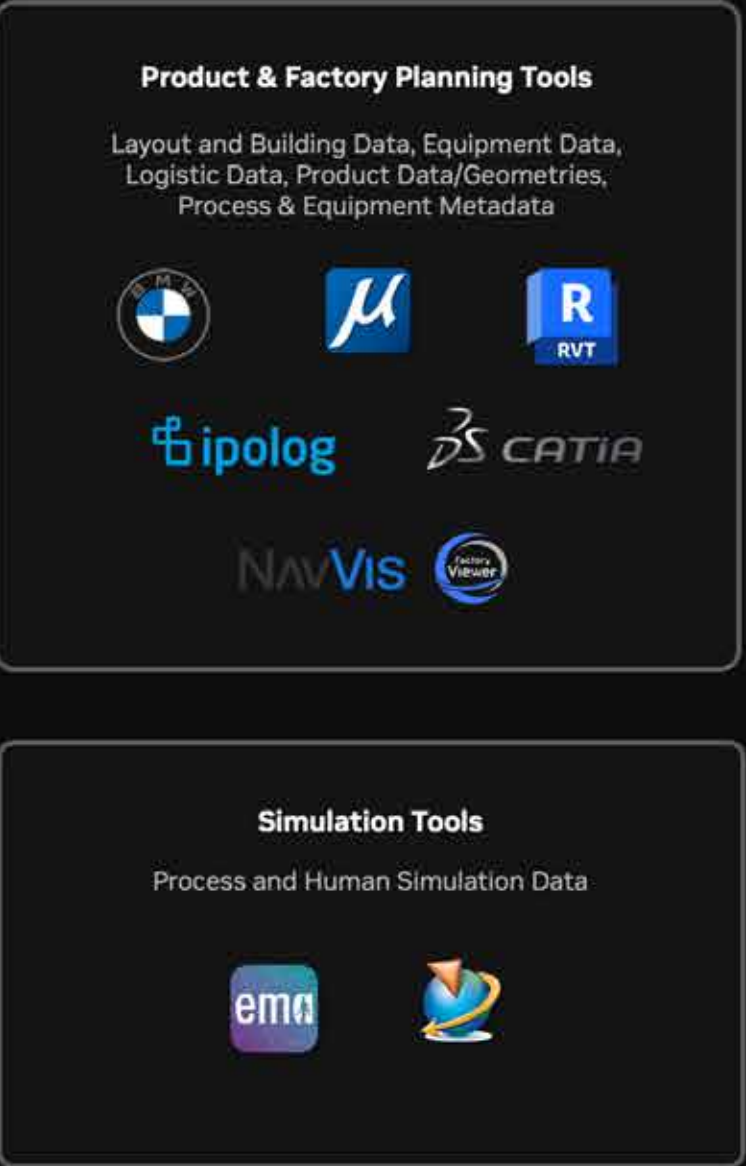
Visualization Layer



Data Integration Layer



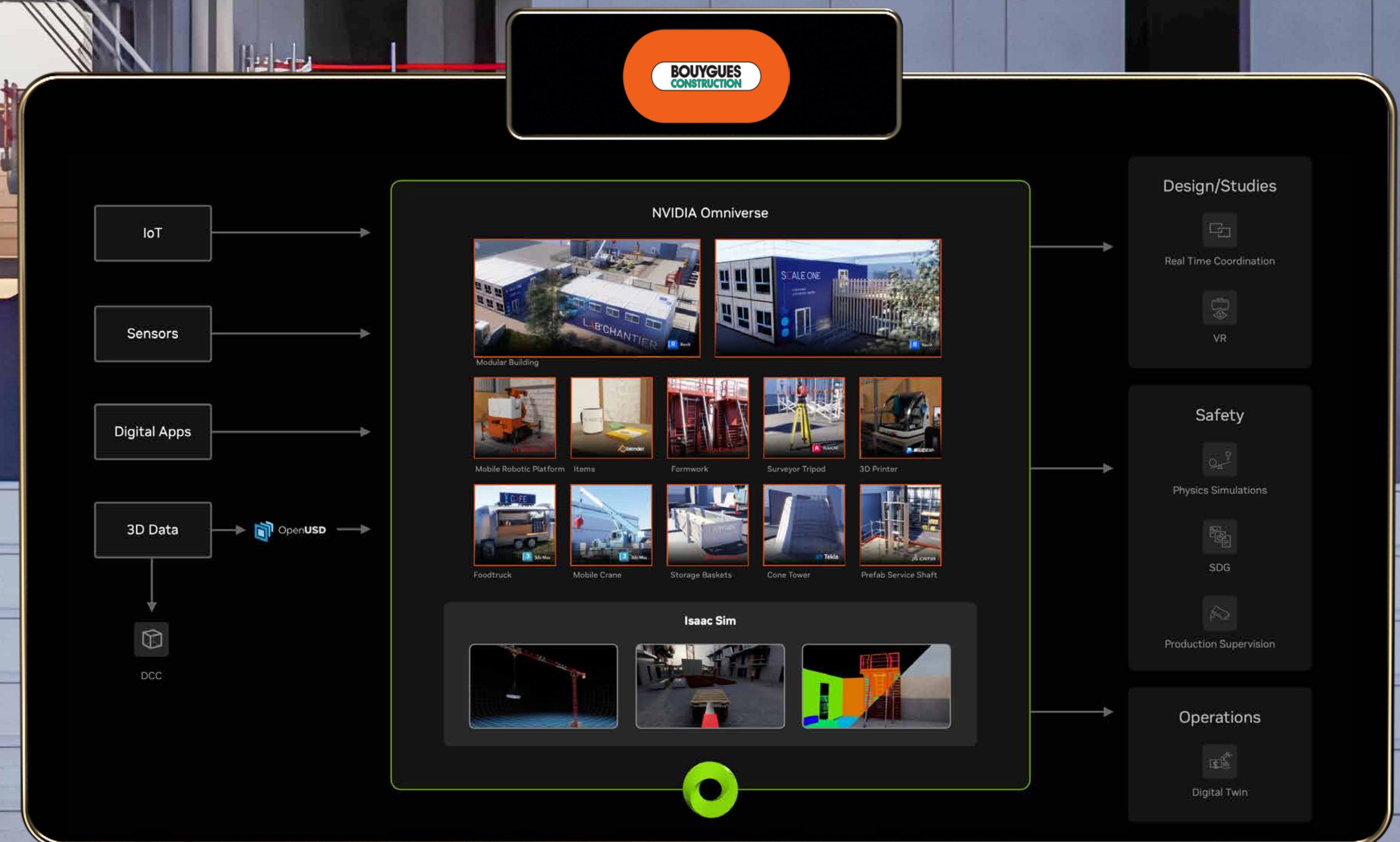
Source Layer



Omniverse Digital Twin

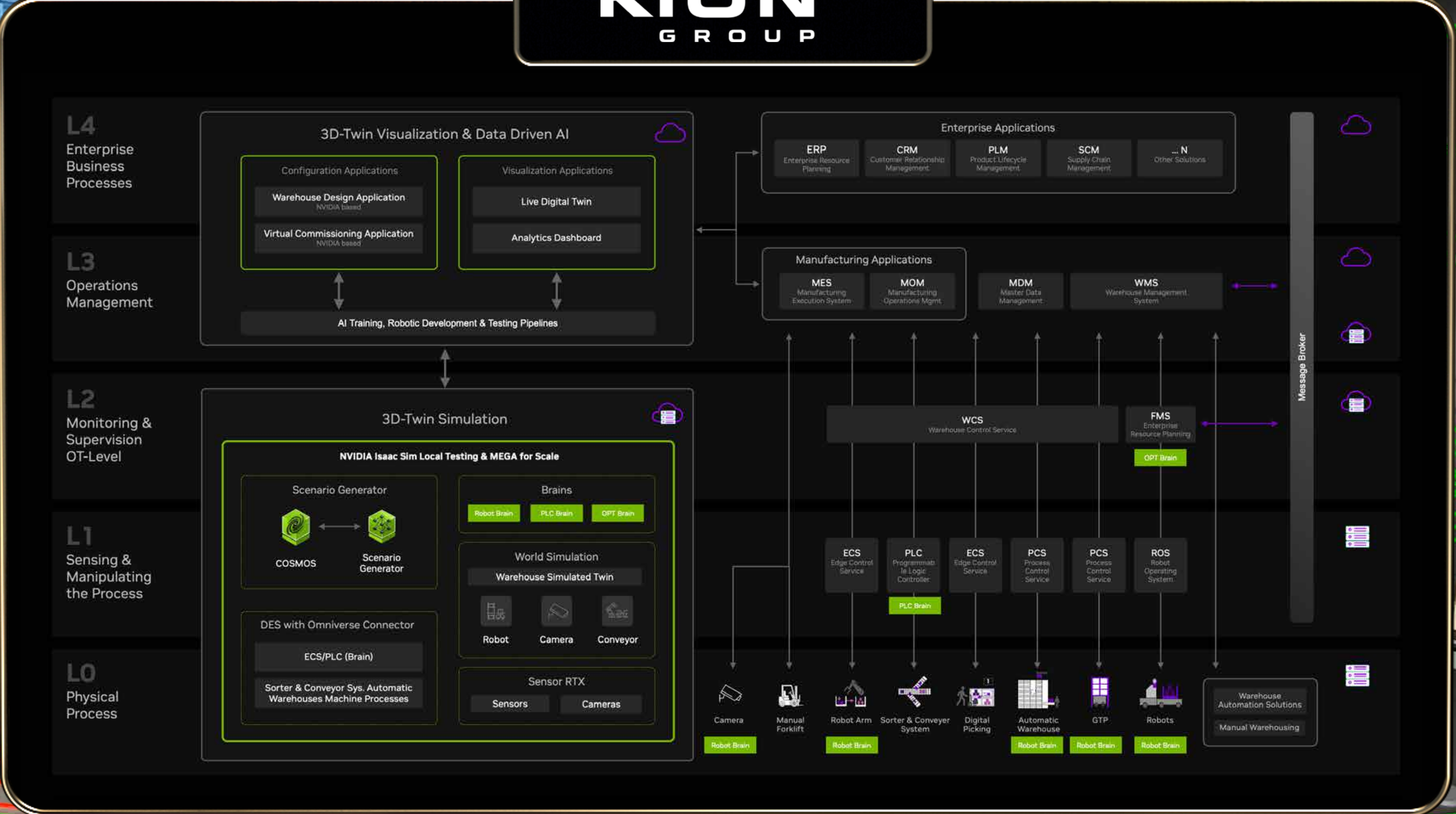


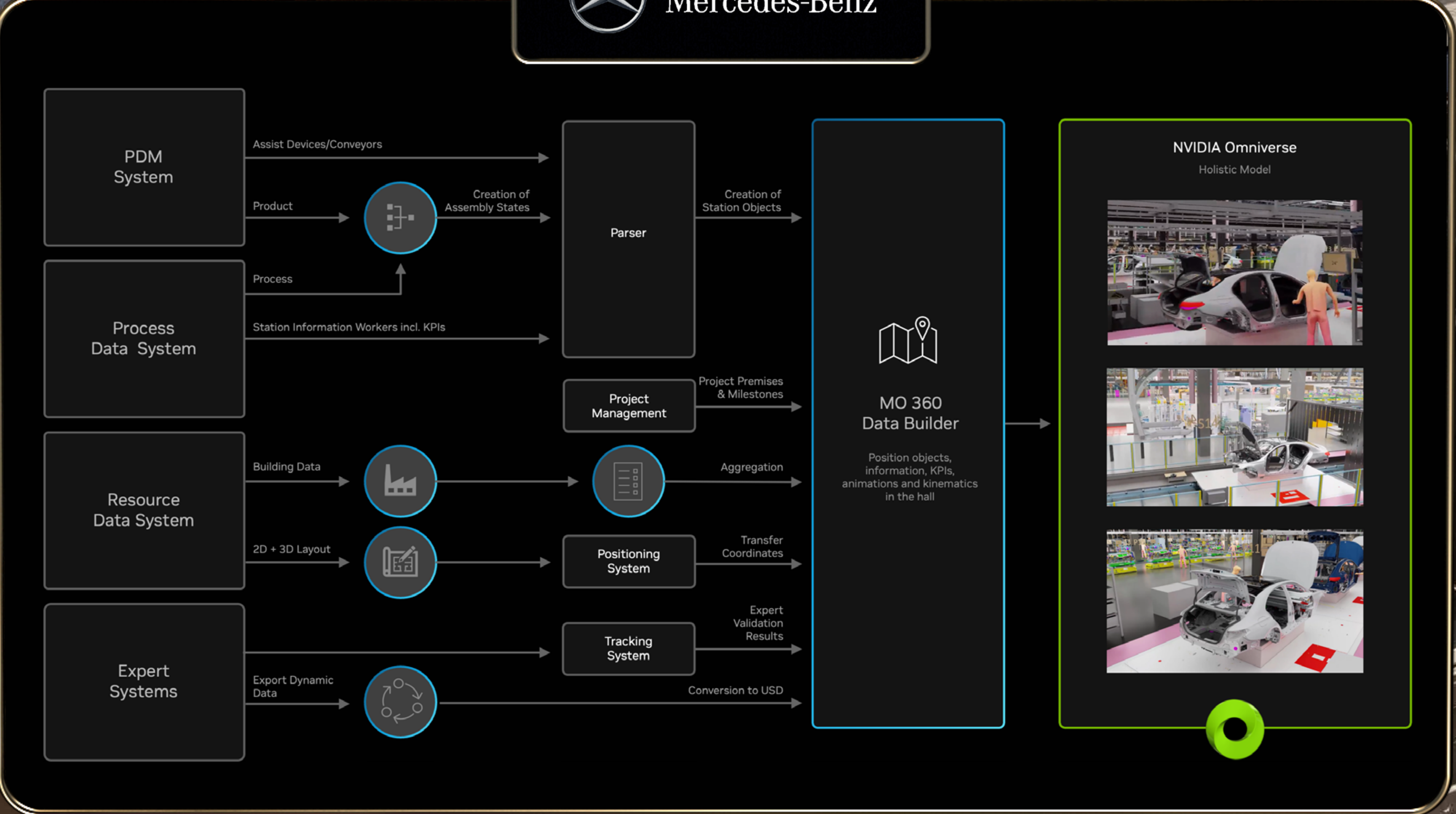
Omniverse Digital Twin



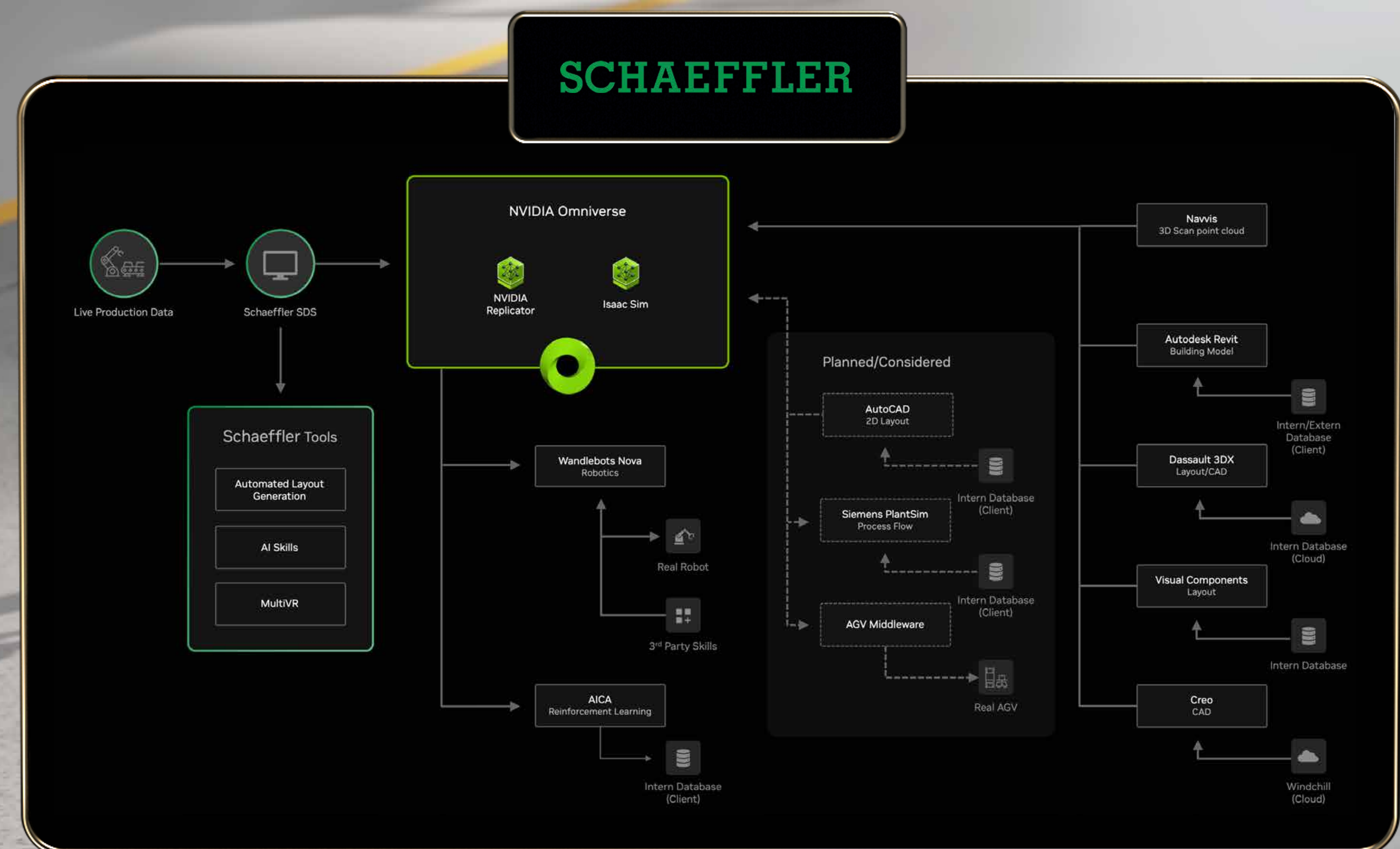


Omniverse Digital Twin



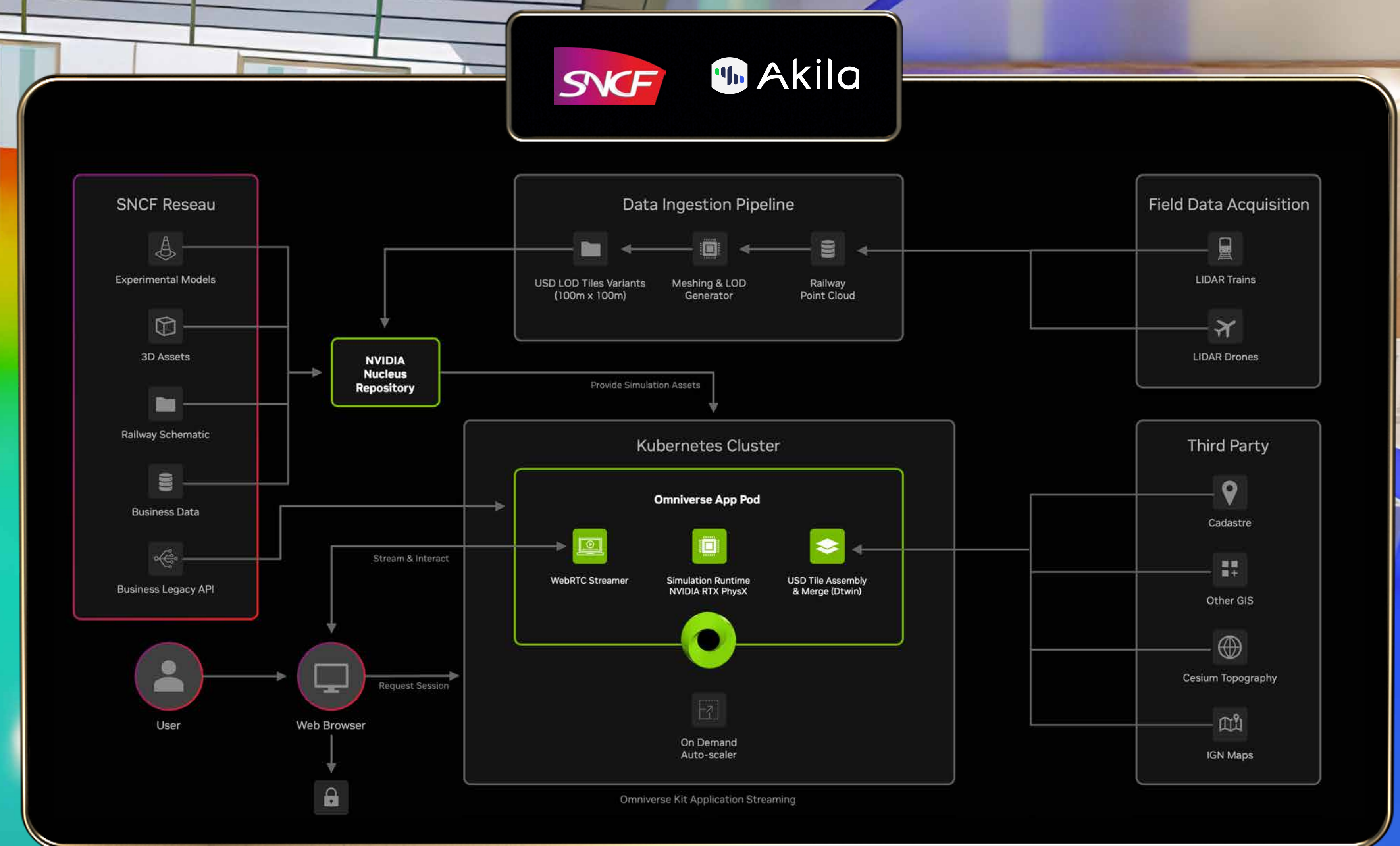


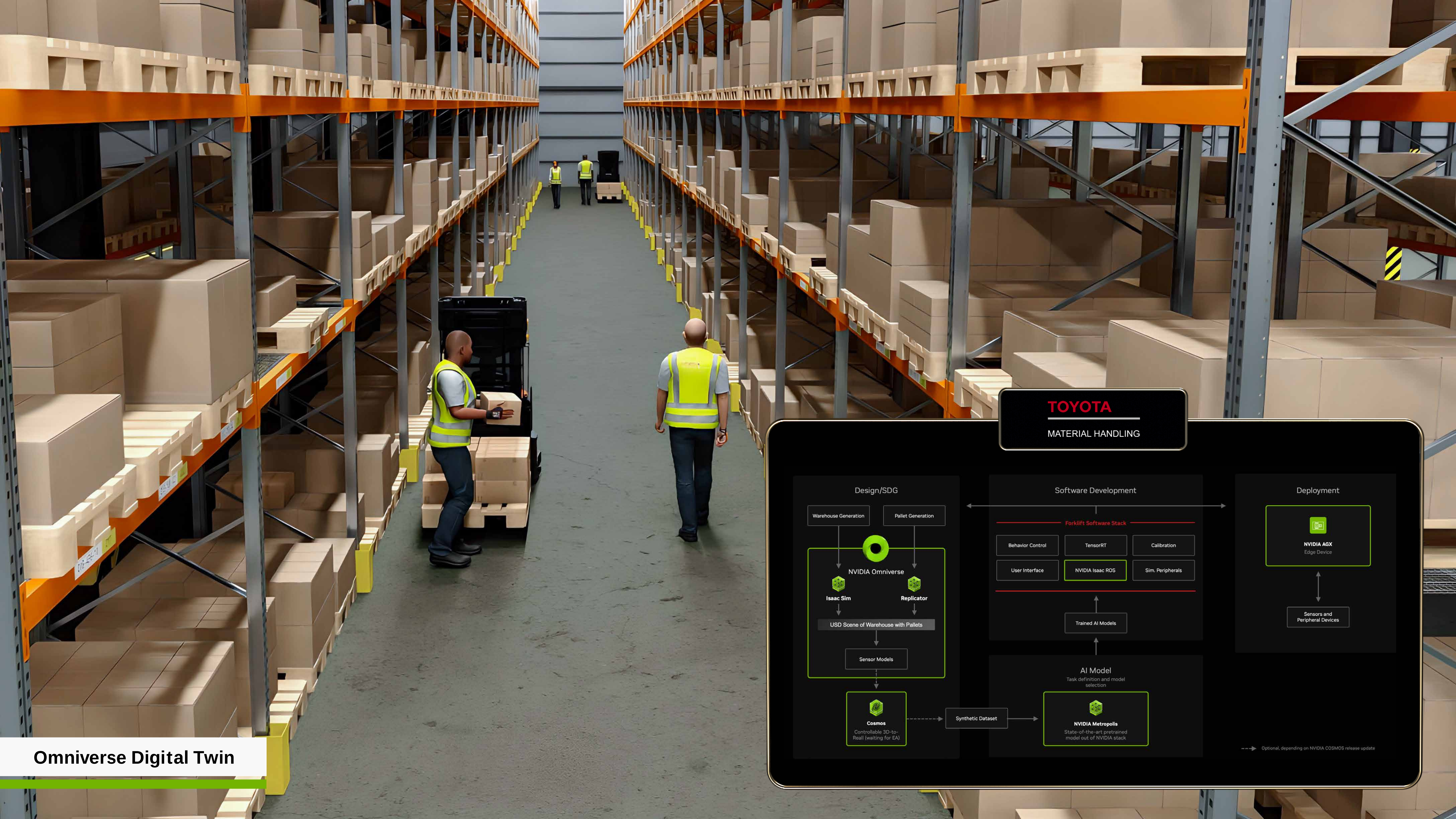
Omniverse Digital Twin





Omniverse Digital Twin

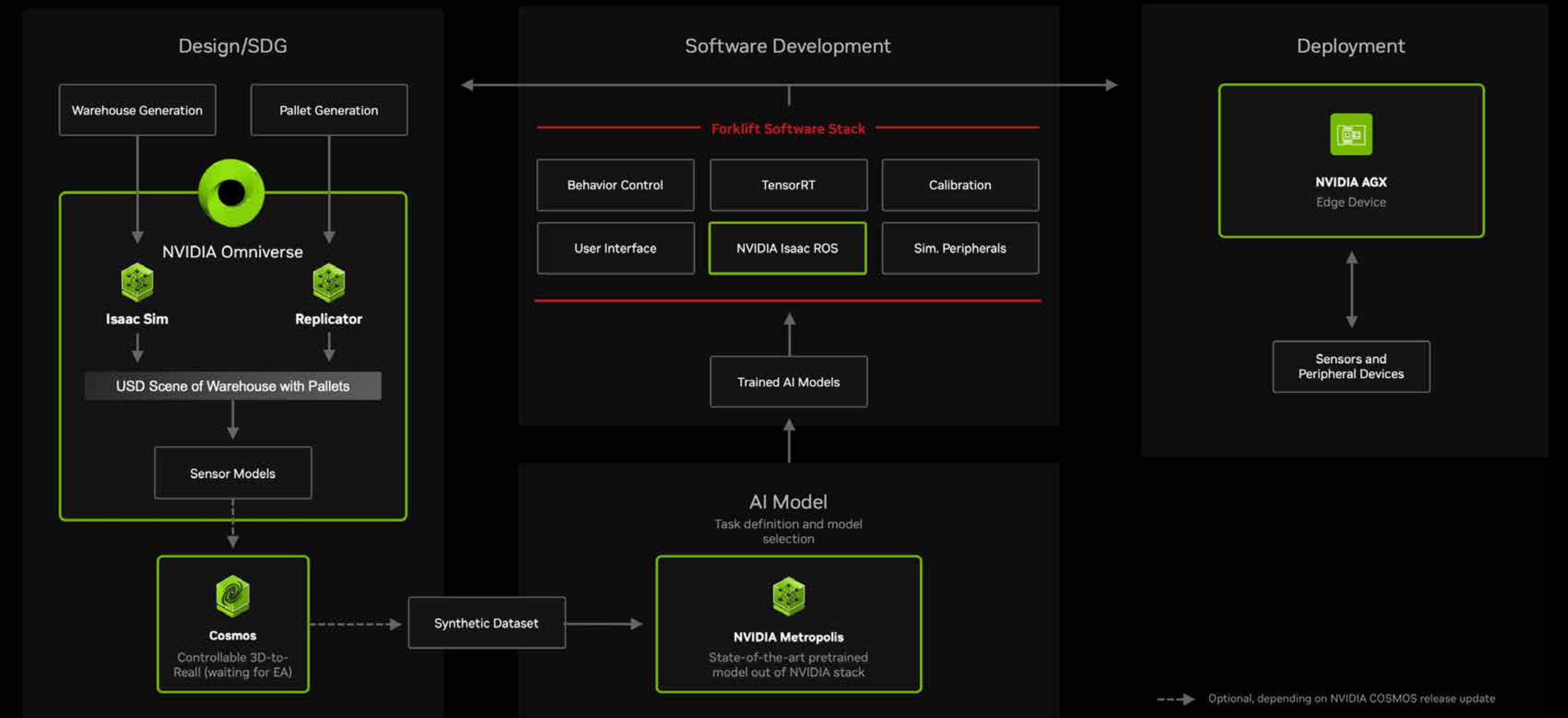


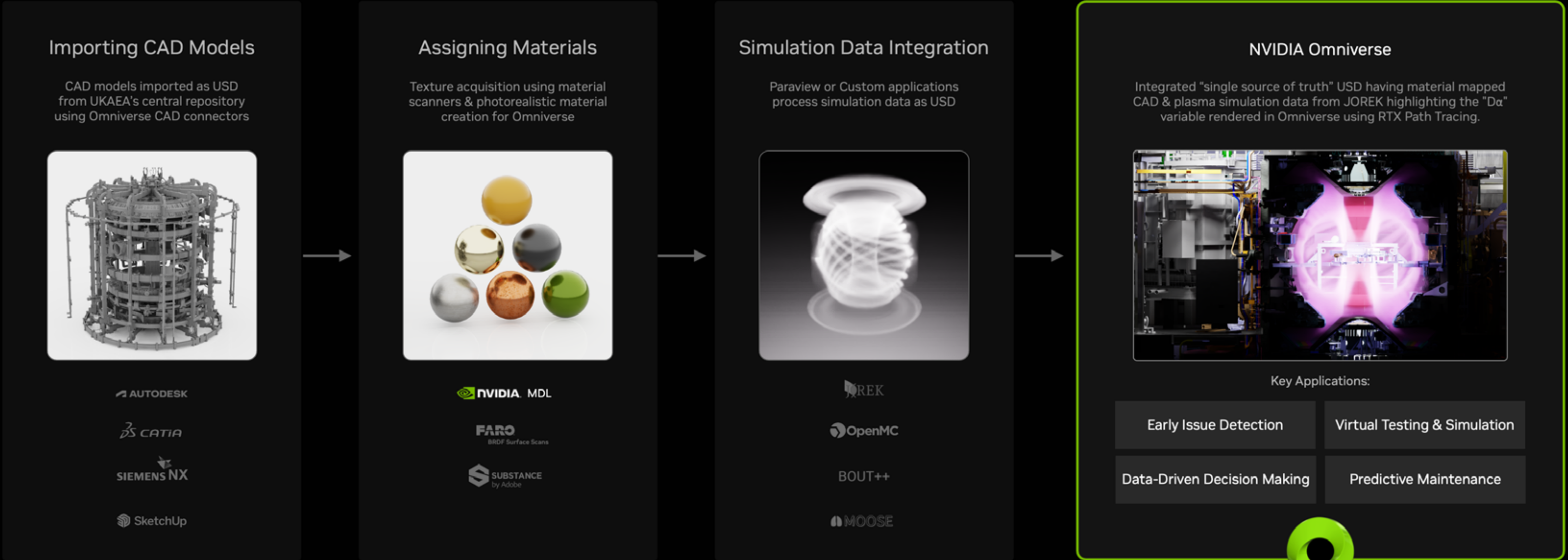


Omniverse Digital Twin

TOYOTA

MATERIAL HANDLING





Accurate Representations



Enhanced Collaboration



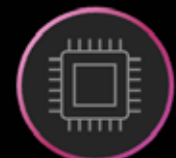
Prototyping & Testing



Data Driven Insights



Immersive Experiences



Customization



Scalability

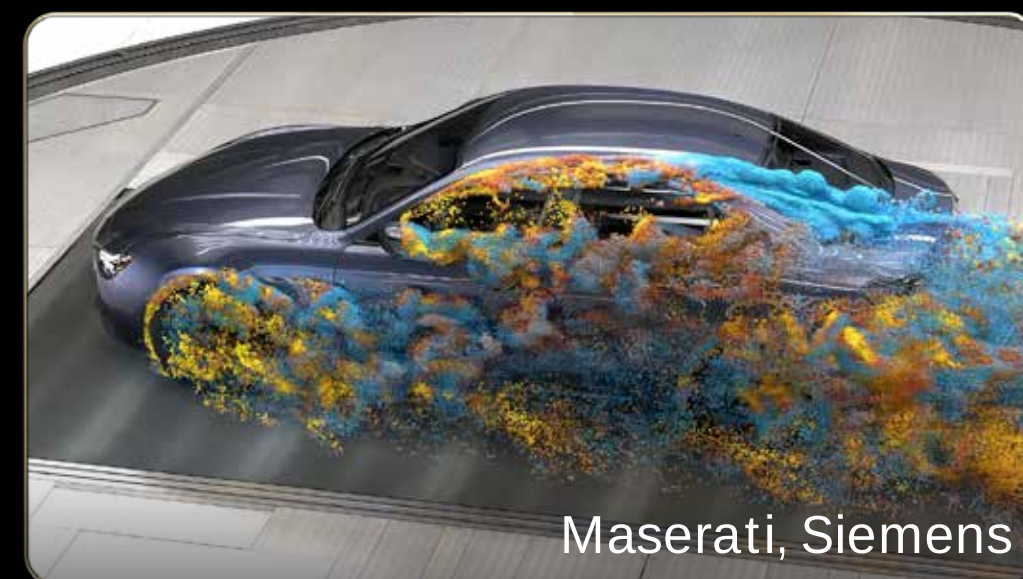
Key Advantages

Announcing NVIDIA Builds World's First Industrial AI Cloud



Siemens

CAD
RTX



Maserati, Siemens

CAE
CUDA-X



BMW

Factory Digital Twin
Omniverse



Schaeffler

Robotics
Isaac and Omniverse

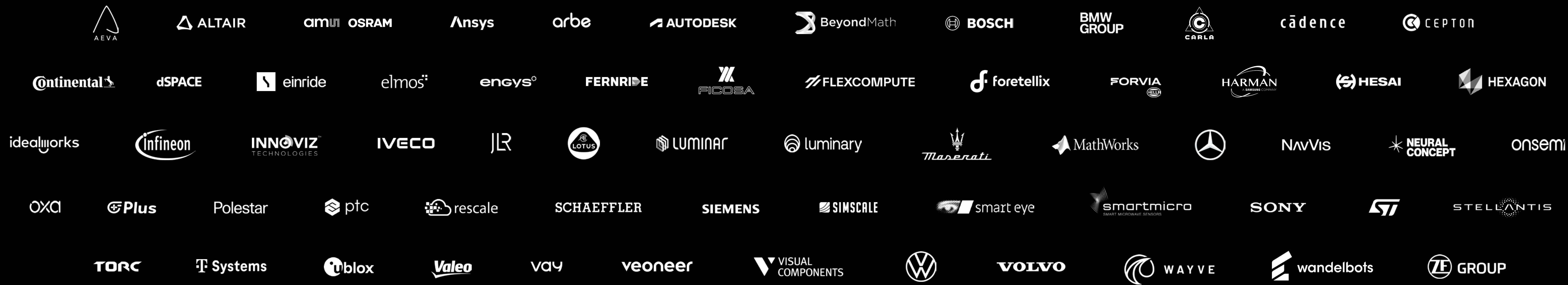


Mercedes-Benz

AV
DRIVE AV and Omniverse

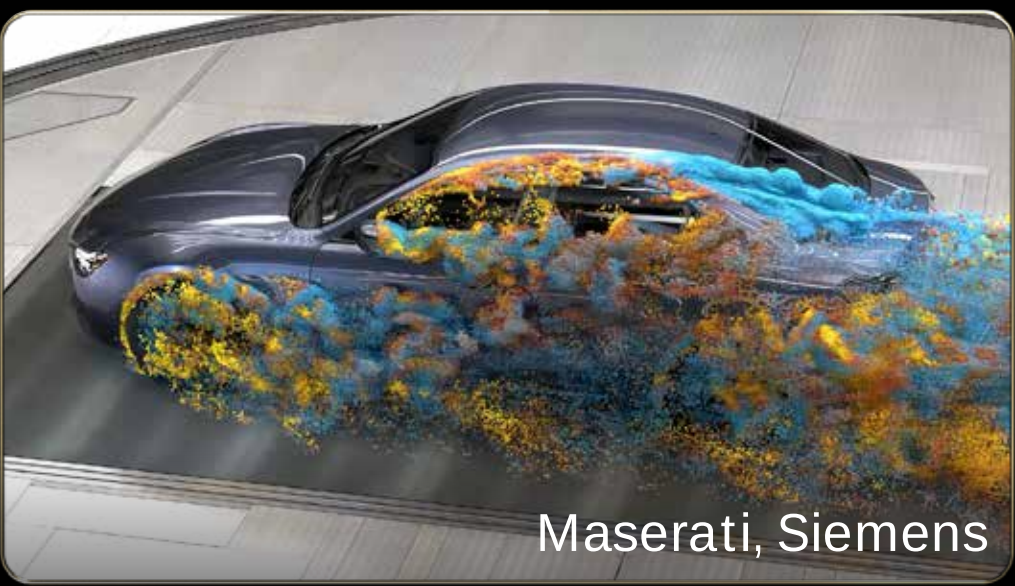


NVIDIA Serving \$2 Trillion Europe Automotive Industry for 15 Years



Siemens

CAD
RTX



Maserati, Siemens

CAE
CUDA-X



BMW

Factory Digital Twin
Omniverse



Schaeffler

Robotics
Isaac and Omniverse



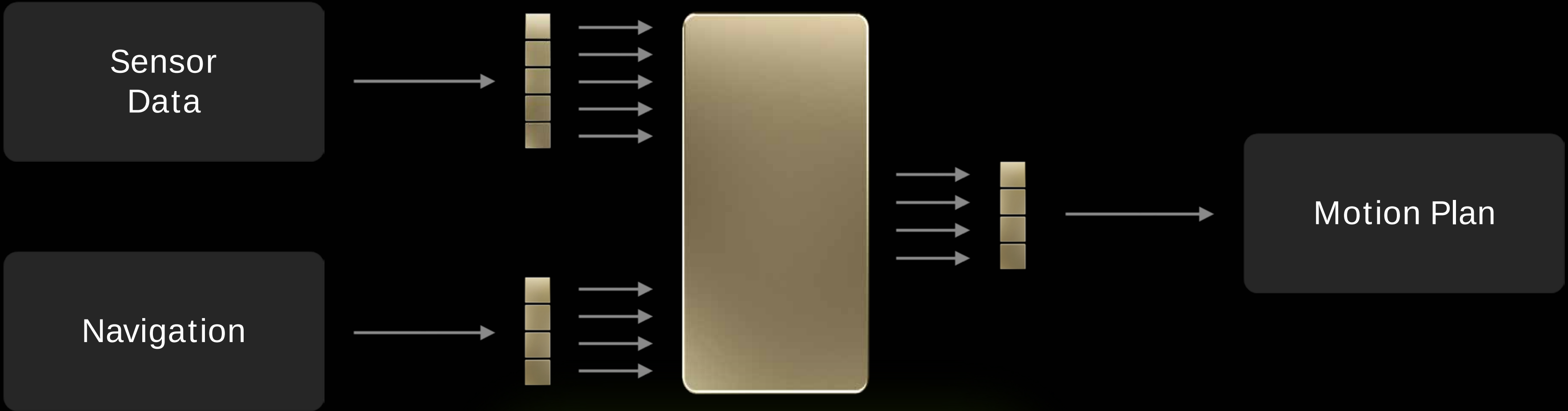
Mercedes-Benz

AV
DRIVE AV and Omniverse

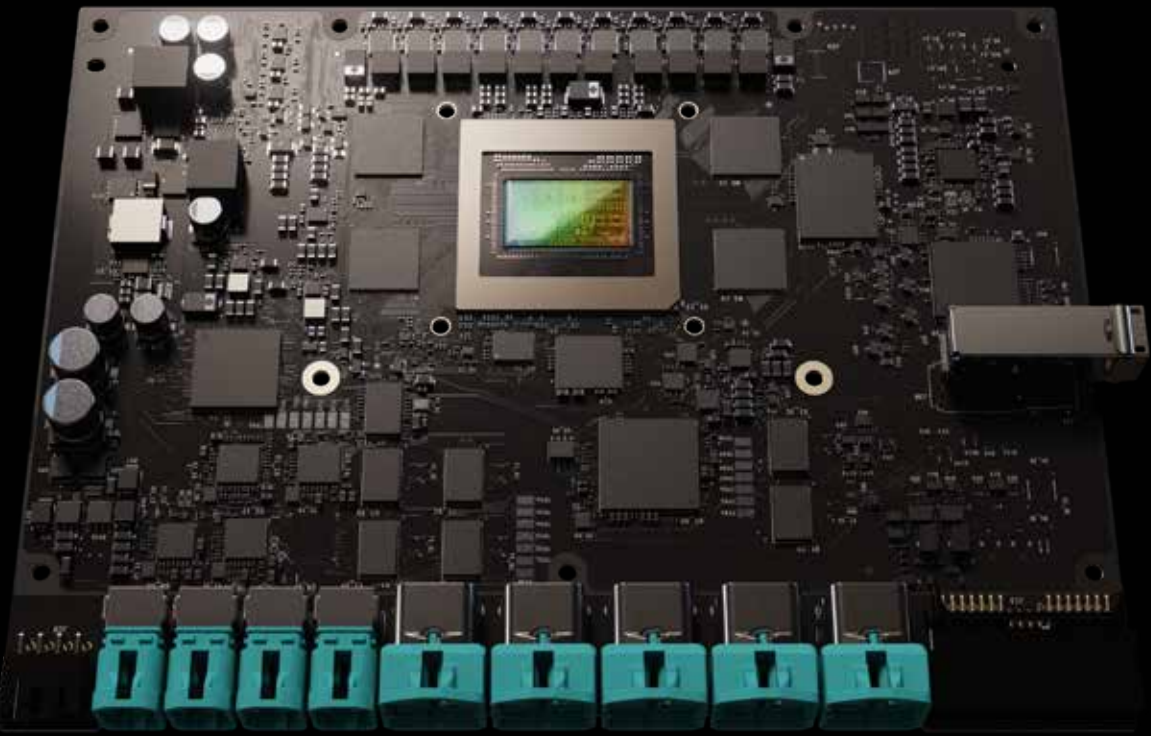


NVIDIA DRIVE

Autonomous Vehicle Platform



NVIDIA DRIVE



DRIVE AGX Thor



Announcing NVIDIA Wins CVPR End-to-End AV Grand Challenge

Generalized Trajectory Scoring for End-to-end Multimodal Planning

Zhenxin Li^{1,2} Wenhao Yao² Zi Wang¹ Xinglong Sun¹ Joshua Chen¹
Nadine Chang¹ Maying Shen¹ Zuxuan Wu² Shiyi Lan¹ Jose M. Alvarez¹
¹NVIDIA ²Fudan University

Abstract

End-to-end multi-modal planning is a promising paradigm in autonomous driving, enabling decision-making with diverse trajectory candidates. A key component is a robust trajectory scorer capable of selecting the optimal trajectory from these candidates. While recent trajectory scorers focus on scoring either large sets of static trajectories or small sets of dynamically generated ones, both approaches face significant limitations in generalization. Static vocabularies provide effective coarse discretization but struggle to make fine-grained adaptation, while dynamic proposals offer detailed precision but fail to capture broader trajectory distributions. To overcome these challenges, we propose GTRS (Generalized Trajectory Scoring), a unified framework for end-to-end multi-modal planning that combines coarse and fine-grained trajectory evaluation. GTRS consists of three complementary innovations: (1) a diffusion-based trajectory generator that produces diverse fine-grained proposals; (2) a vocabulary generalization technique that trains a scorer on super-dense trajectory sets with dropout regularization, enabling its robust inference on smaller subsets; and (3) a sensor augmentation strategy that enhances out-of-domain generalization while incorporating refinement training for critical trajectory discrimination. As the winning solution of the Navsim v2 Challenge, GTRS demonstrates superior performance even with sub-optimal sensor inputs, approaching privileged methods that rely on ground-truth perception. Code will be available at <https://github.com/NVlabs/GTRS>.

1. Introduction

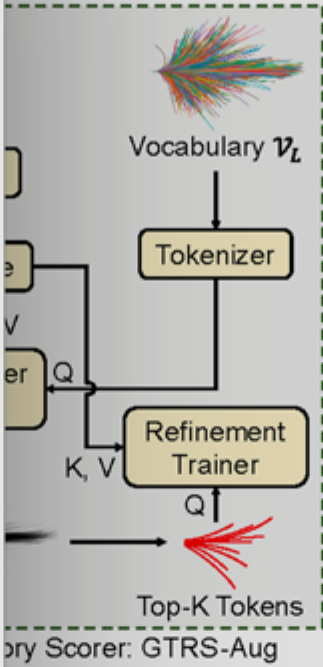
End-to-end multi-modal planning has emerged as a powerful approach in autonomous driving. Unlike traditional uni-modal planners that predict a single trajectory [5, 11, 12], multi-modal approaches [1, 3, 6, 15, 16, 18, 19] generate multiple candidates, enabling greater adaptability during inference. This adaptability supports a wide range of applications, including responding to language instructions [21–23, 28], accommodating different driving styles [15, 25], and navigating complex driving environments [16, 18, 24].

The typical problem of end-to-end multi-modal planning involves evaluating multiple trajectory proposals through scoring [1, 3, 16, 18, 19, 24] given raw sensor data, without access to ground-truth perception. The planner selects the trajectory with the highest likelihood as the decision.

Current trajectory scoring methods generally fall into two categories: (1) scoring a large static vocabulary [3, 16, 18, 24], and (2) scoring a small set of dynamically generated proposals [19, 30]. Both approaches face challenges in generalization. Fixed trajectory vocabularies [3, 16, 18] offer limited flexibility, as they cannot adapt to situations where dynamic proposals are needed. Meanwhile, methods that rely on a small number of dynamic proposals [19, 30] often fail to generalize to unseen trajectories, since the scorer is only exposed to a narrow subset during training. Ideally, a robust scorer should generalize across diverse trajectory distributions—whether static or dynamic—to handle the full complexity of real-world scenarios.

To address these limitations, we propose GTRS (Generalized Trajectory Scoring) for end-to-end multi-modal planning. GTRS is built on a key insight: an effective trajectory scorer must be trained on both coarse and fine-grained trajectory distributions to develop robust generalization capabilities. Our approach contains three complementary techniques, each leading to a dedicated sub-network as shown in Fig. 1:

- Diffusion-based Trajectory Generation (DP):** A diffusion policy (DP) [4] produces diverse trajectory candidates with BEV features as the condition. DP provides the fine-grained details crucial for safety-critical situations that coarse, fixed vocabularies cannot capture. (Sec. 2.1)
- Trajectory Vocabulary Generalization (GTRS-Dense):** We train on a super-dense vocabulary of trajectory samples (16,384 trajectories) covering a wide range of driving scenarios. To maximize generalization of fixed trajectory vocabularies, we propose a trajectory dropout training strategy. The idea is to deliberately create a mismatch between training and inference vocabularies—training the model to effectively generalize to unseen trajectory distributions during inference. (Sec. 2.2)
- Sensor Augmentation with Refinement (GTRS-Aug):** To handle unexpected trajectory events and data distri-



but queries attend to image Encoder [26], and a Diffusion-based trajectory proposals V_{dp} . Transfuser [5] and include a dense supervision on the BEV former, we apply first-order trajectory waypoints to the DDPM scheduling strategy to handle the full complexity of real-world scenarios.

Generalization

Generators provide fine-grained details in their output of possible driving scenarios. To train the model on a wide range of driving scenarios, we introduce a novel technique that trains the model on a wide range of driving scenarios, even

Vocabulary Scorer GTRS-Dense introduces a trajectory tokenizer to process the input, and a model to handle complex interactions between training and inference vocabularies—training the model to effectively generalize to unseen trajectory distributions during inference. (Sec. 2.2)

First, we deliberately train the model on a super-dense vocabulary (16,384 trajectories) covering a wide range of driving scenarios. To maximize generalization of fixed trajectory vocabularies, we propose a trajectory dropout training strategy. The idea is to deliberately create a mismatch between training and inference vocabularies—training the model to effectively generalize to unseen trajectory distributions during inference. (Sec. 2.2)

To handle unexpected trajectory events and data distri-

S_1	EPDMS ₂	EPDMS
40.8	25.6	
48.6	36.7	
50.2	39.7	
52.7	40.8	
57.3	42.0	
59.0	43.4	

ry from V_{dp} during inference. This is from Depth Anything [32].

es, enabling the model to learn similar trajectories. The self-distillation framework average (EMA) copy of the signals:

$$y_t^m - y_t^m, -\delta_m, \delta_m, \quad (1)$$

target score, $\hat{y}_t^m$ is the teacher model's prediction. δ_m ensures the refined target of the ground truth. GTRS-Aug to perform out-of-domain settings without

tion

described above, we combine the trajectory scorers (i.e., inference, as illustrated in Fig. 1) generated by the generator and scored by the scorer. dynamic proposals at inference, is a deliberate design choice of both approaches. By using a vocabulary, the scorer generalizes across a wide range of inference time, the diffusion-based, context-aware trajectory current scene. Meanwhile, the head and potential instability into the training loop, while the model of dynamically generated

and for evaluating end-to-end AV Grand Challenge introduces a new

HC	EC	EPDMS
87.7	36.0	51.3
91.5	25.4	
97.5	79.1	23.1
95.7	75.9	
96.9	40.4	
94.1	49.7	41.7
97.3	46.7	43.4
94.1	41.6	
97.3	40.9	45.3
97.2	54.2	
97.6	54.2	42.1
97.1	56.1	
97.6	49.8	44.7
96.5	52.7	
97.6	50.2	43.4
95.1	48.3	
97.3	53.3	46.6
96.3	47.8	
97.6	51.6	49.4
96.7	56.8	

GTRS-Dense and GTRS-Aug Trajectory Scorer.

izes well to unseen dynamic scenarios, demonstrating strong zero-shot generalization. When combining the generator with random proposals, the model improves by +1.1 EPDMS (43.4), showing that vocabulary generalization significantly

Finally, we evaluate the sensor augmentation strategy. The model achieves the same top-4 as the best GTRS-Dense scorer [18] by a large margin, showing that augmentations and refinement improve trajectory scoring.

variants achieve significant performance gains. By scaling the model to the EVA-VTL [9], our best model on the Navhard Benchmark ensemble of GTRS-Dense achieves 46.6 EPDMS. GTRS-E, an ensemble of all models, approaches the performance of the privileged planner that relies on ground truth, despite our use of challenging scenarios. This demonstrates the exceptional

ized scene representation for autonomous driving. In ICCV, pages 8340–8350, 2023.

Cheolho Park, Benjamin Sapp, et al. Motiondiffuser: Con- ditioned diffusion for prediction using diffusion. In CVPR, pages 181–191, 2024.

Wang, Sangrok Lee, Yuseok Lee, et al. Energy and gpu-computation efficient real-time object detection. In CVPR, pages 181–191, 2024.

Wang, Xin Wen, Pengxiang Li, et al. Xiangpeng: A general- izable trajectory model with human feedback. arXiv preprint arXiv:2503.12820, 2025.

Wang, Shiyi Lan, Zhiding Yu, et al. Jose M Alvarez. End-to-end driving via expert- ly trained diffusion models. arXiv preprint arXiv:2503.12820, 2025.

Wang, Jiahao Li, Jan Kautz, Tong He, et al. ego status all you need for autonomous driving? arXiv preprint arXiv:2503.12820, 2025.

Wang, Shiyi Lan, Zhiding Yu, et al. Jan Kautz, Zuxuan Wu, et al. End-to-end planning with multi-modal planning with multi-modal planning. arXiv preprint arXiv:2406.06978, 2024.

Haoan Yin, Bo Jiang, Cheng Wang, Xiangyu Li, Ying Zhang, et al. Truncated diffusion model for vision. In CVPR, 2025. 1, 2

Vitor Guizilini, Jie Li, and Jan Kautz. Vision language model for autonomous driving with 3d scene representation. In CVPR, 2025. 1, 2

Maria Marcu, Jan Hünemmann, et al. Jamie Shotton, Eliahe Shotton. Vision language model for autonomous driving. arXiv preprint arXiv:2503.12820, 2025.

He Arani, and Oleg Sinavski. End-to-end autonomous driving with vision. arXiv preprint arXiv:2503.09594, 2025.

z, Kashyap Chitta, Li Chen, et al. Jens Beifwenger, Ping Luo, et al. Li. Drivelm: Driving with vision. In ECCV, pages 256–274, 2024.

Chitta, Zhiding Yu, Shiyi Lan, et al. Xiangyu Li, and Jose M Alvarez. End-to-end autonomous driving with test-time adaptation. arXiv:2503.11650, 2025. 1

otie Meng, and Erkang Cheng. End-to-end multi-modal planning with de-



2024



2025



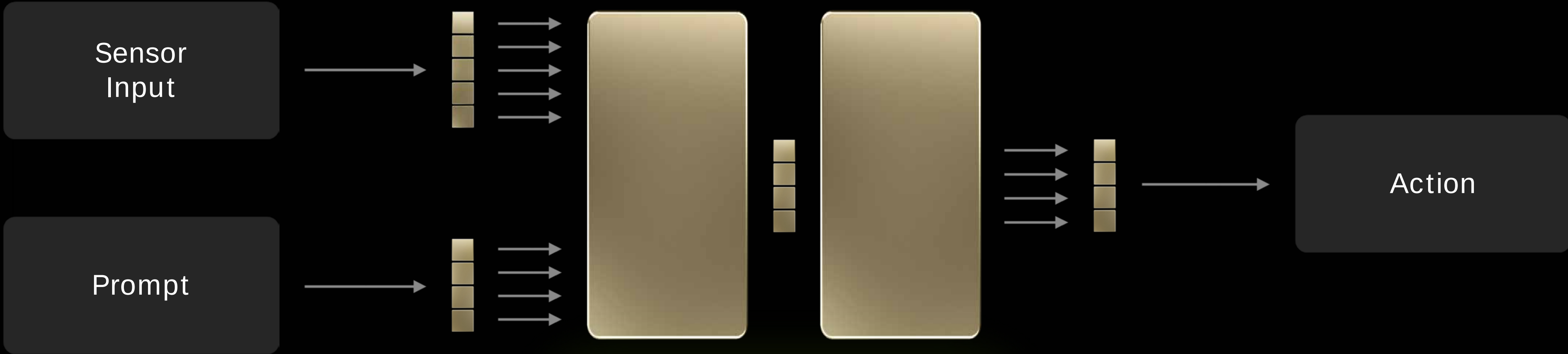
Announcing NVIDIA Halos Recognized by Global Safety Ecosystem

Comprehensive Chip-to-Deployment Safety System



NVIDIA Isaac

Open Robotics Development Platform



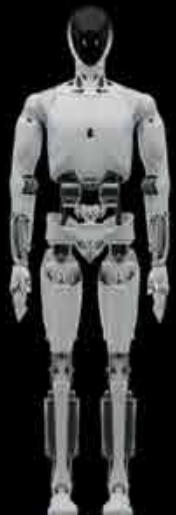
NVIDIA Isaac



Jetson Thor



XPENG
Iron



Apptronik
Apollo



ROBOTERA
Star 1



Agibot
A2



Fourier
GR-2



Boston Dynamics
E-Atlas



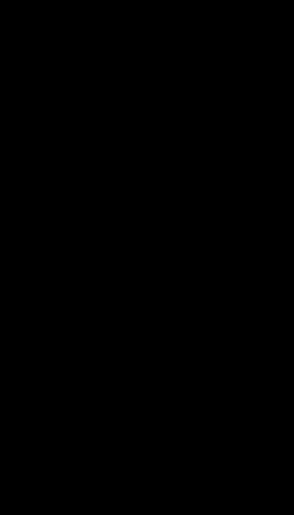
Agility Robotics
Digit



NEURA Robotics
4NE-1



NVIDIA
Jensen Huang



Mentee
MenteeBot



Agile Robots
Dual Arm



Humanoid
HMND 01



1X
NEO



Galbot
G1



Unitree
H1



Figure
Figure 02



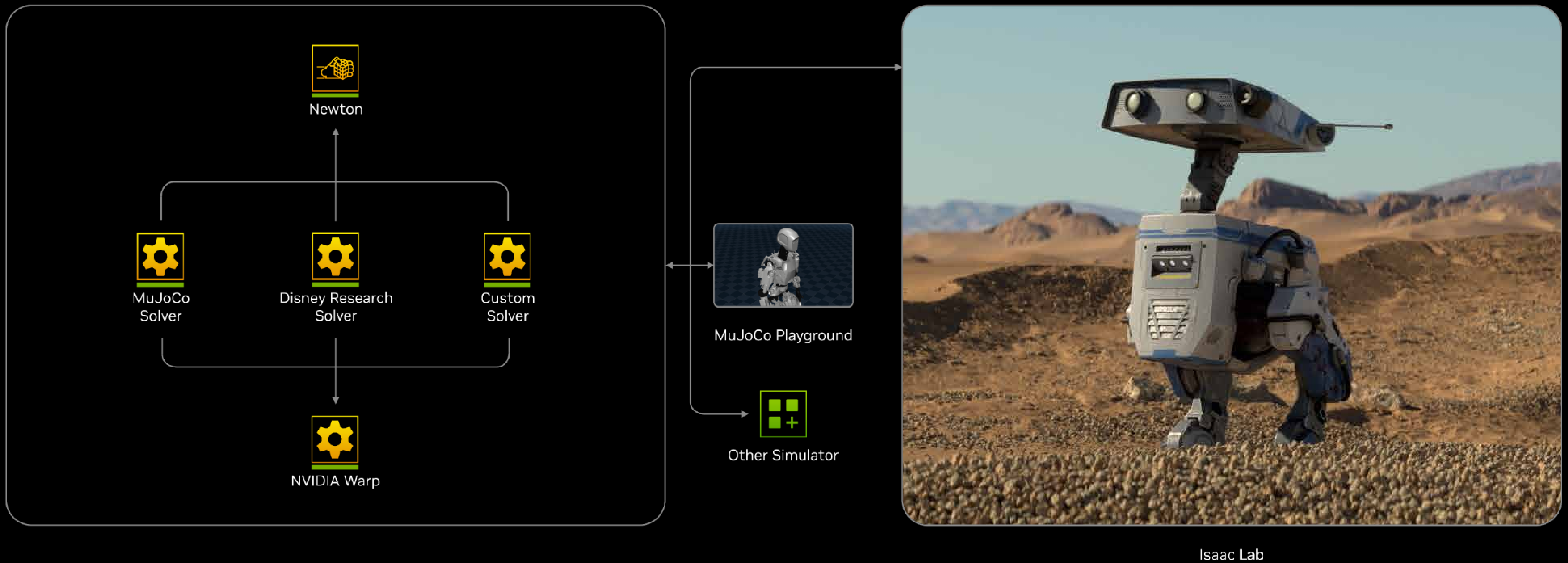
UBTech
Walker S1



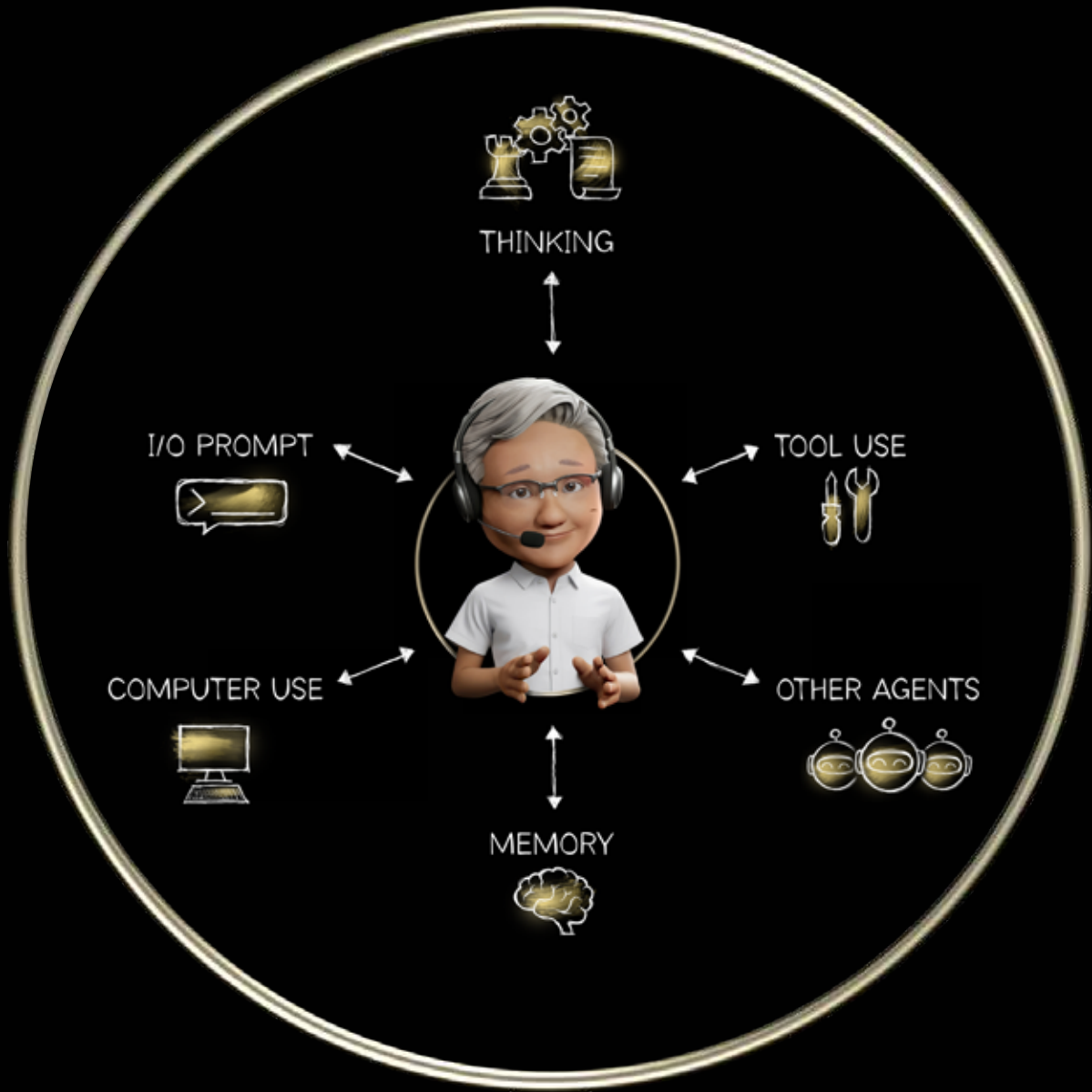
Disney Research

 NVIDIA

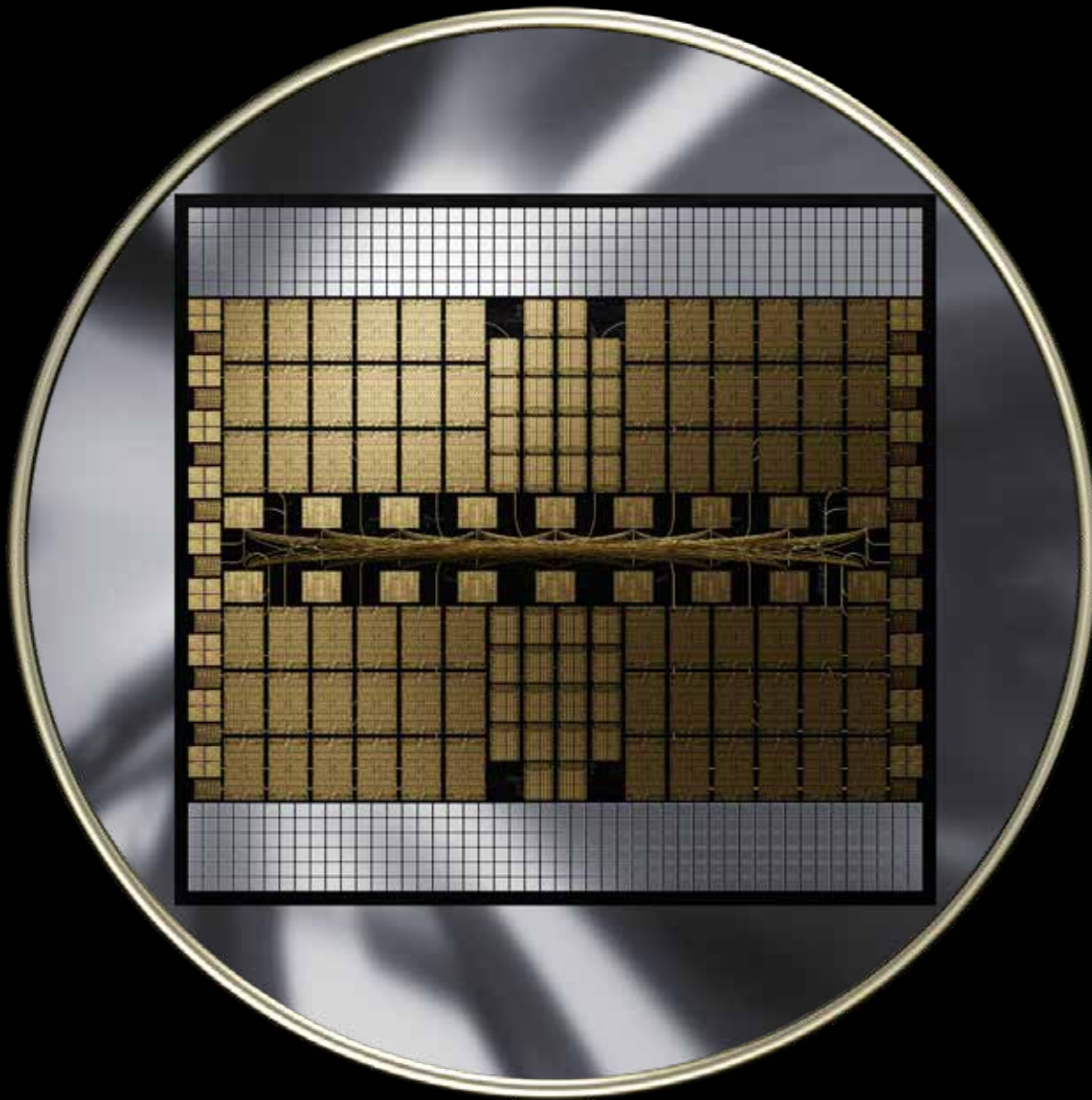
 Google DeepMind



A New Industrial Revolution



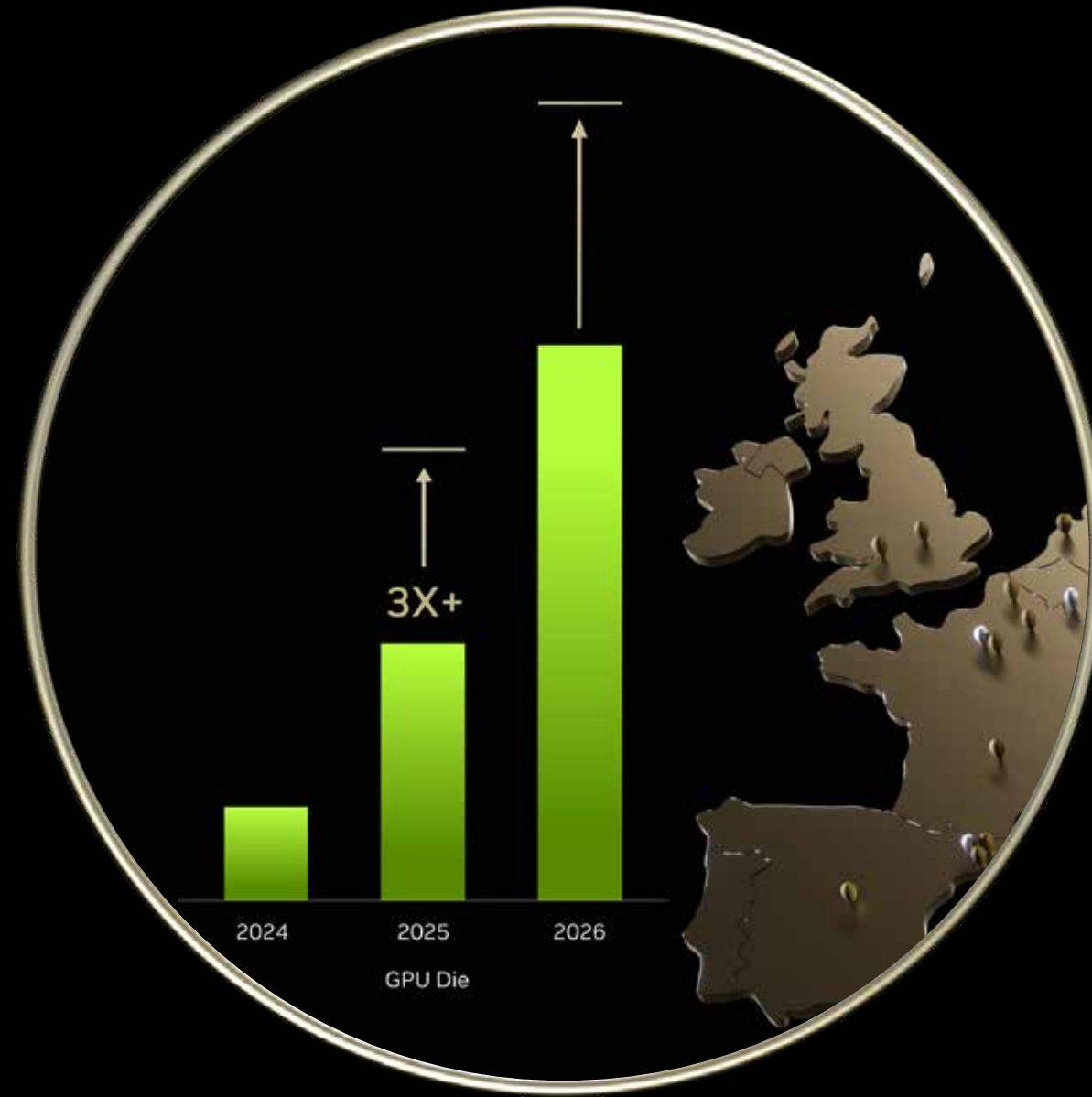
A NEW WAVE OF AI
AGENTS AND ROBOTS



BLACKWELL
“A THINKING MACHINE”



AI FACTORIES –
THE NEW INFRASTRUCTURE



10X EUROPEAN
INFRASTRUCTURE NEXT 2 YEARS

