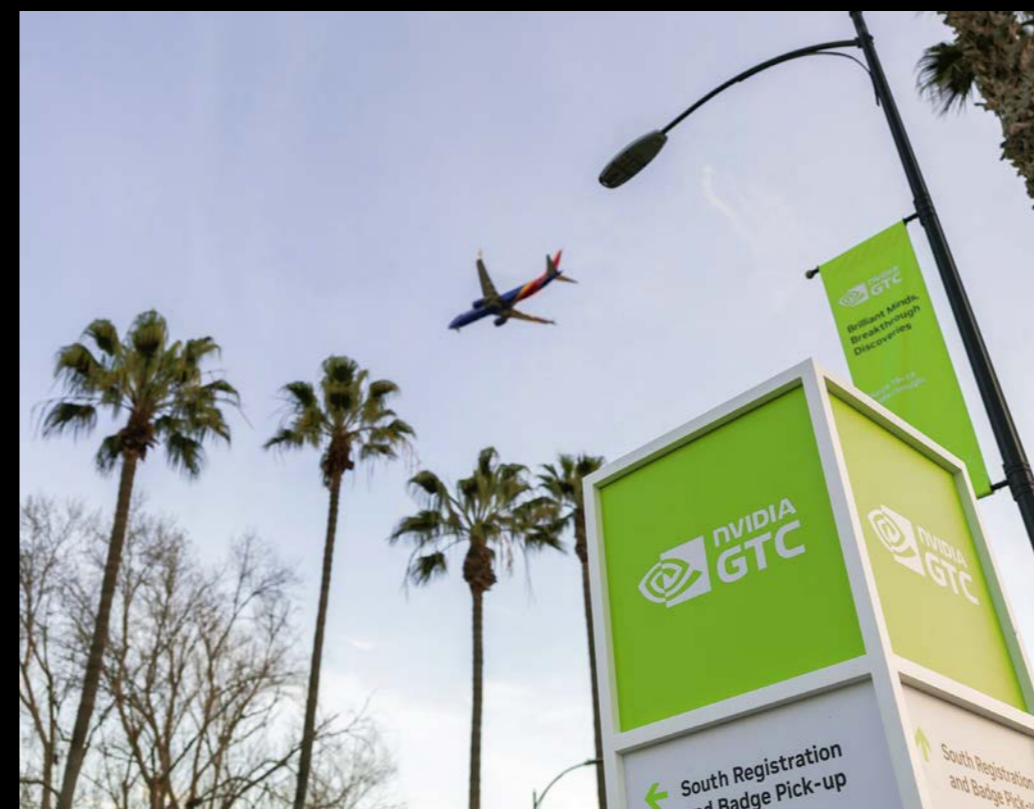
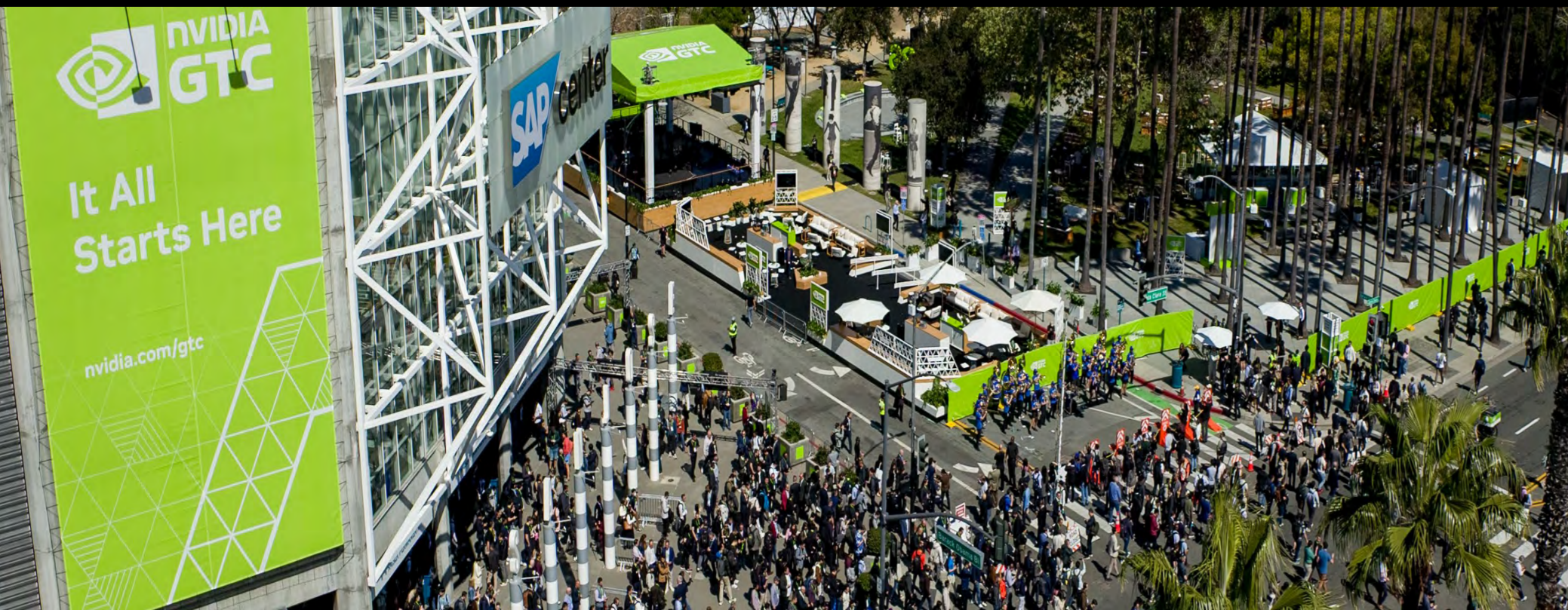




# 2026 Highlights







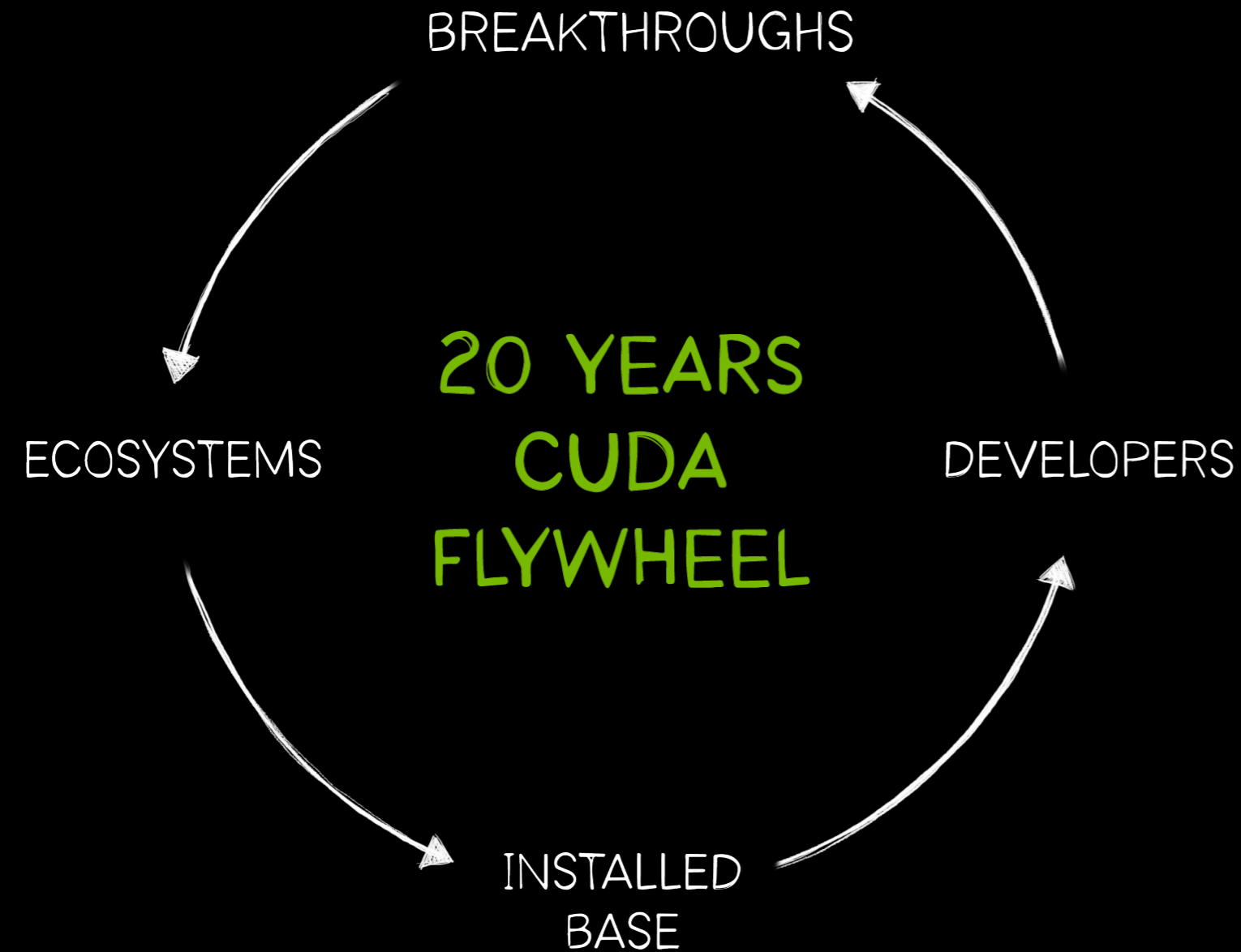






**“This is GTC, the AI festival of everything that drives the tech world.”**

Jim Cramer, CNBC



**“CUDA has over 4 million developers, 3,000+ optimized applications, and deep integration into every major AI framework.”**

Forbes

For 20 years, CUDA has built the developer flywheel powering accelerated computing.

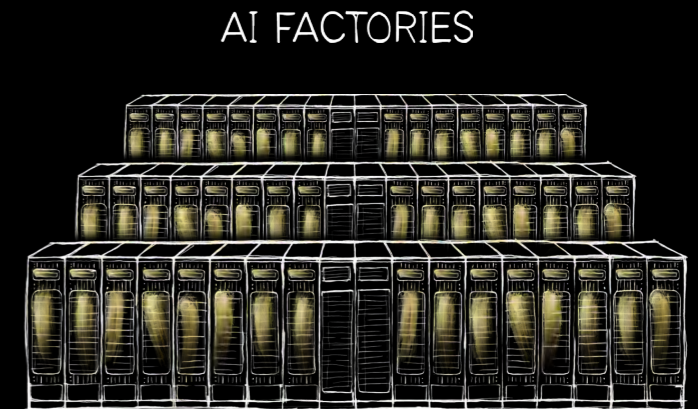
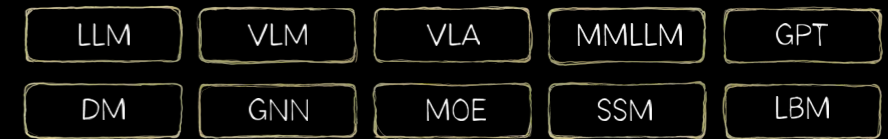
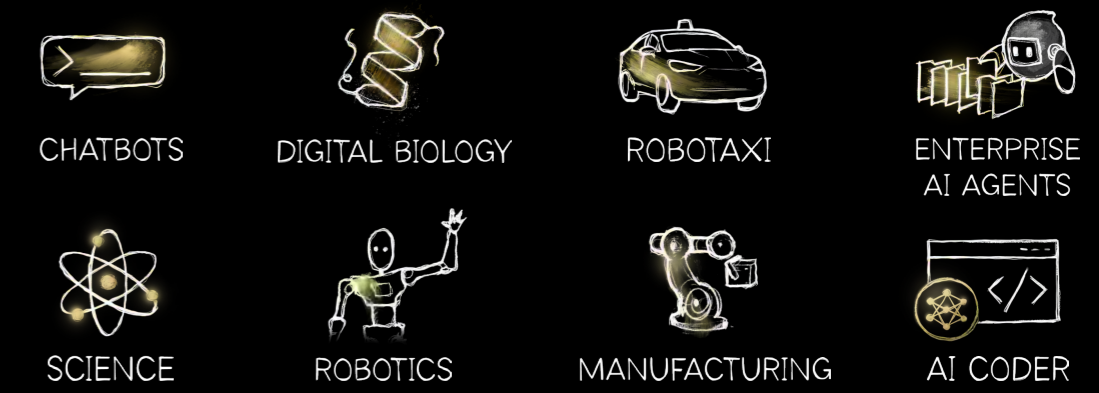
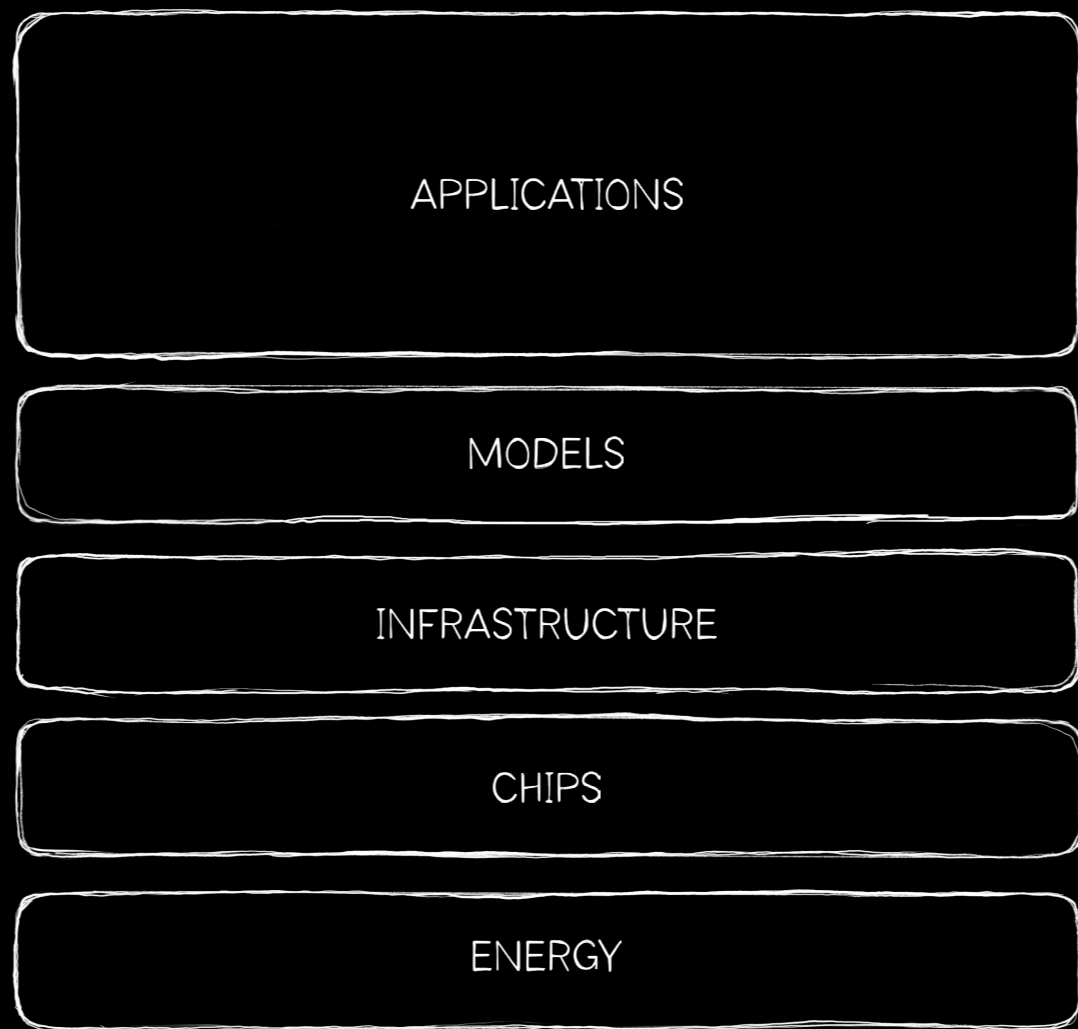
CUDA-X turns algorithms into infrastructure across AI, data, and science, compounding performance at scale.

As compute grows, software drives efficiency—defining throughput, cost per token, and how much intelligence the world can produce.

“AI depends on five layers—energy, chips, infrastructure, models, and applications—and that all five need to scale together.”

Fortune

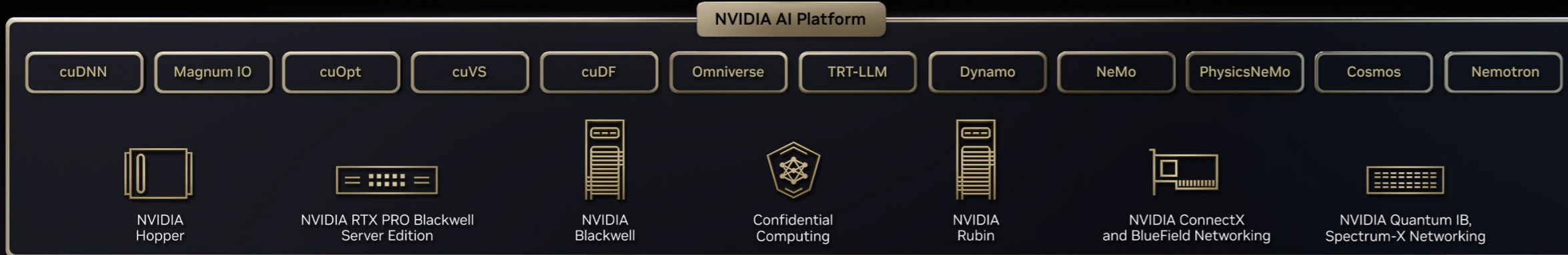
AI does not scale in isolation. It scales as a system. Energy powers chips, chips power infrastructure, infrastructure runs models, and models enable applications. Each layer compounds the next. If one stalls, the system stalls. AI factories demand all five layers advance together.



# NVIDIA ❤️ AI Natives

<p>AI for Auto</p>	<p>AI for Customer Support</p>	<p>AI for Engineering</p>	<p>AI for Healthcare</p>	<p>AI for Robotics</p>	<p>AI for Software Development</p>
--------------------	--------------------------------	---------------------------	--------------------------	------------------------	------------------------------------

<p>DL Frameworks</p>	<p>Agent Frameworks/Protocols</p>	<p>Frontier Model Builders</p>	<p>Model to Production</p>
----------------------	-----------------------------------	--------------------------------	----------------------------

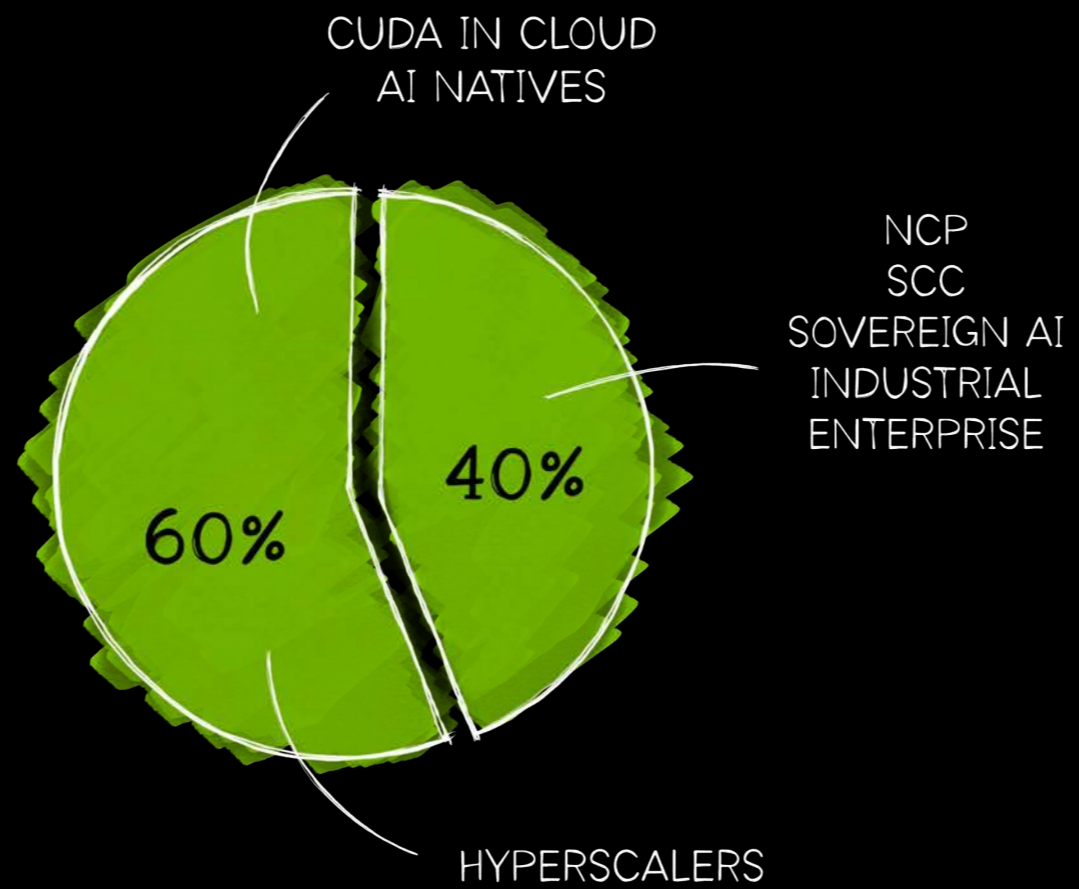
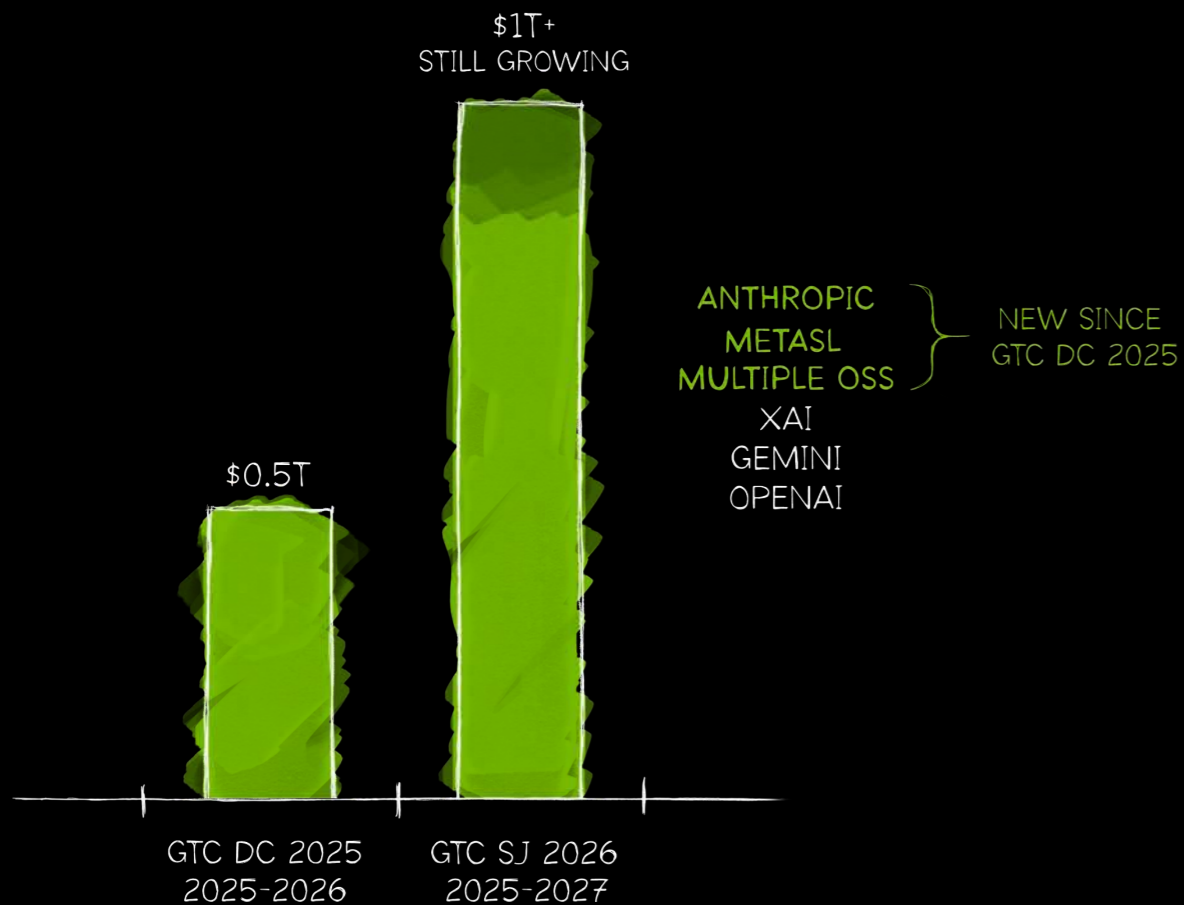


“Every company in the world today needs to have an agentic system strategy.”

Business Insider

AI has expanded beyond research into every sector—AI natives, enterprises, industrial systems, and sovereign infrastructure. Startups and incumbents are building on accelerated computing, and every industry now depends on inference at scale. This is no longer adoption. It is the emergence of a universal computing platform.

# BLACKWELL + RUBIN



**“\$1 trillion in sales as new computing era begins.”**

The Wall Street Journal

One year ago, we saw roughly \$500 billion of demand. Today, that visibility is at least \$1 trillion through 2027.

This is not speculative. It is infrastructure being built—across hyperscalers, model builders, sovereign clouds, and enterprises. And this only reflects Blackwell and Rubin. The AI factory buildout is just beginning.

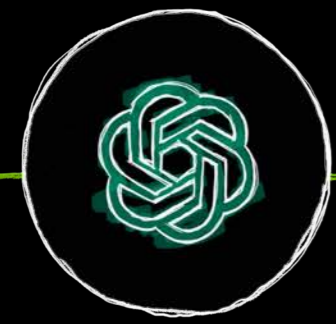
2027 is a step-function in AI factory scale.

**“GTC 2026 revealed a major shift in how AI is built and deployed.”**

TechRepublic



## Inference Inflection Arrives 10,000X ChatGPT Compute



2023

MODELS & CONTEXT 10X  
TOKENS 10X



2024

MODELS & CONTEXT 100X  
TOKENS 100X



2025

**“NVIDIA made it clear at GTC that it is the provider of entire AI computing systems powering the new ‘inference’ phase of AI.”**

Fortune

AI was defined by training. Now it is defined by inference. Models reason, agents use tools, and every interaction generates tokens continuously at scale. Inference is no longer episodic—it is persistent, compounding, and driving exponential compute demand. This is the inflection point.

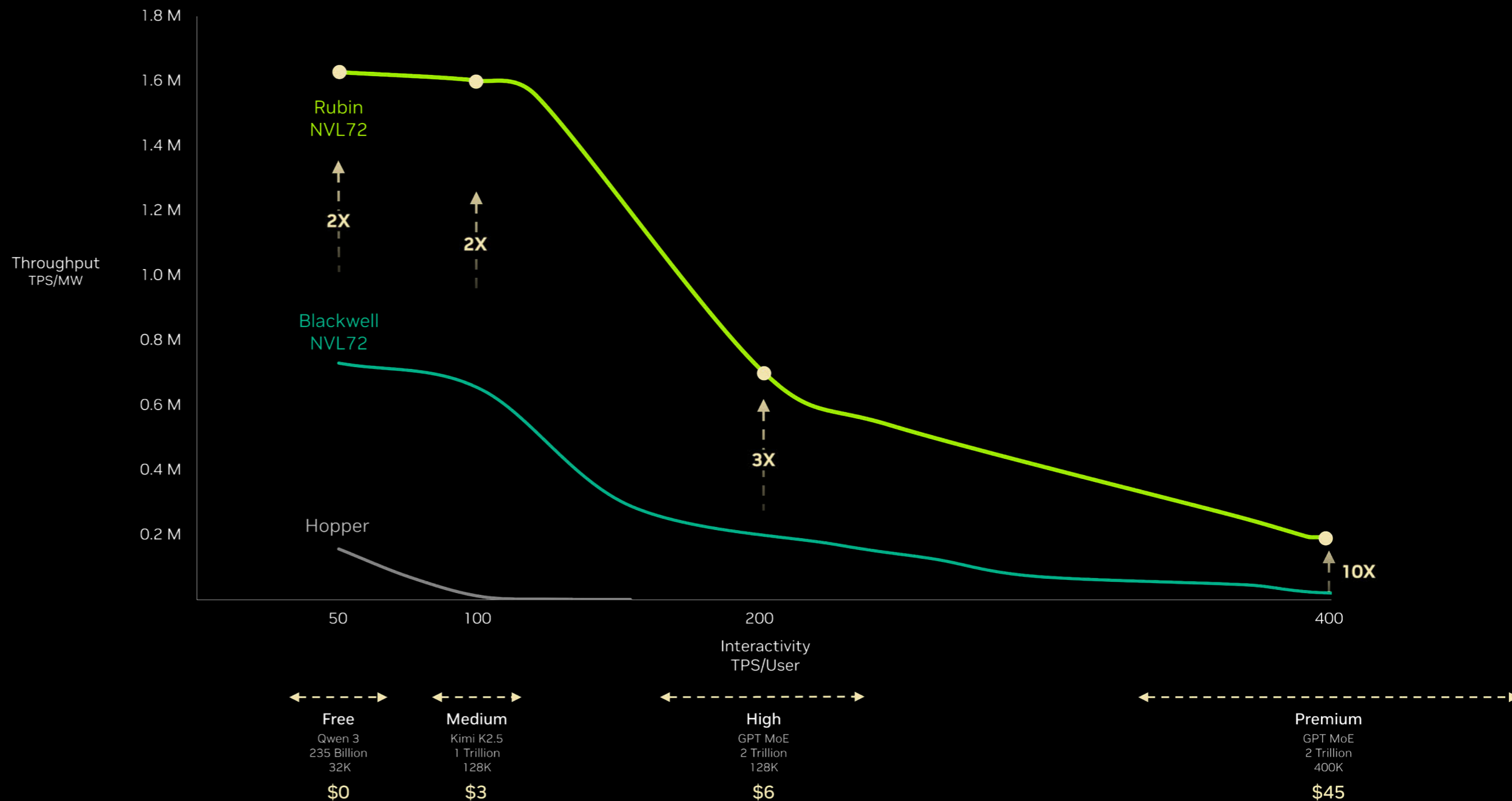
“NVIDIA clearly has the lowest cost per token by a country mile.”

CNBC

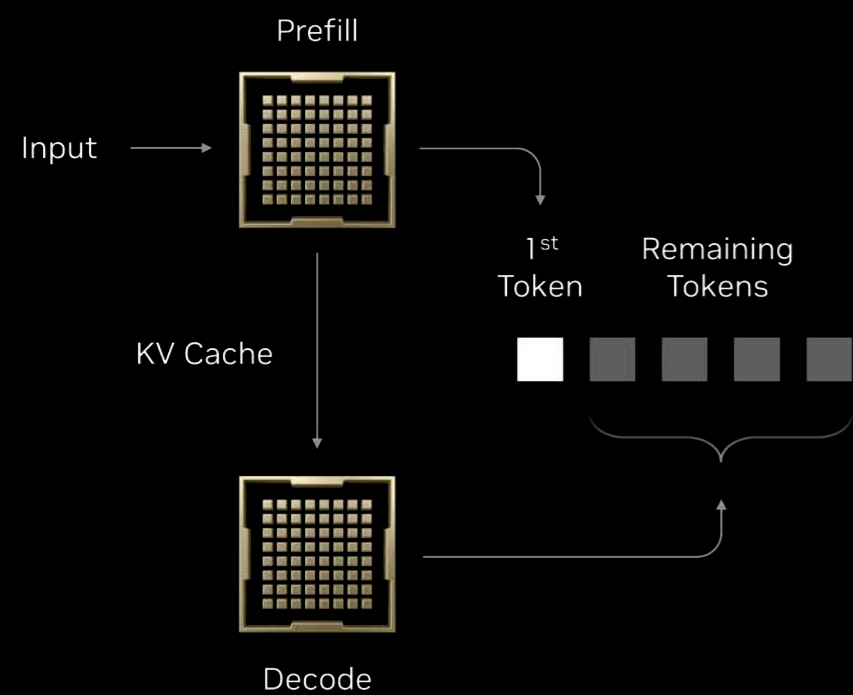
AI data centers are not storage. They are token factories. They convert energy into tokens that directly drive corporate revenue.

Through extreme co-design across silicon, systems, networking, and software, NVIDIA delivers the lowest cost per token.

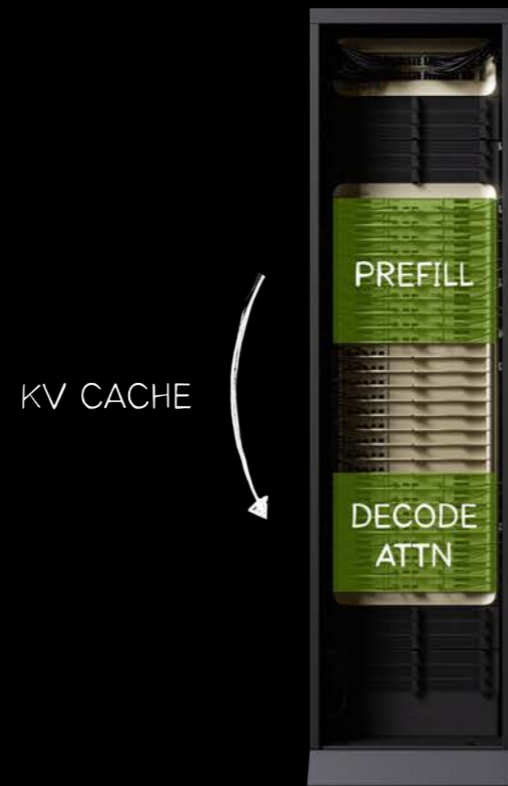
In the AI factory era, tokens per second per watt translate directly into revenue. NVIDIA-powered AI factories generate the highest revenue.



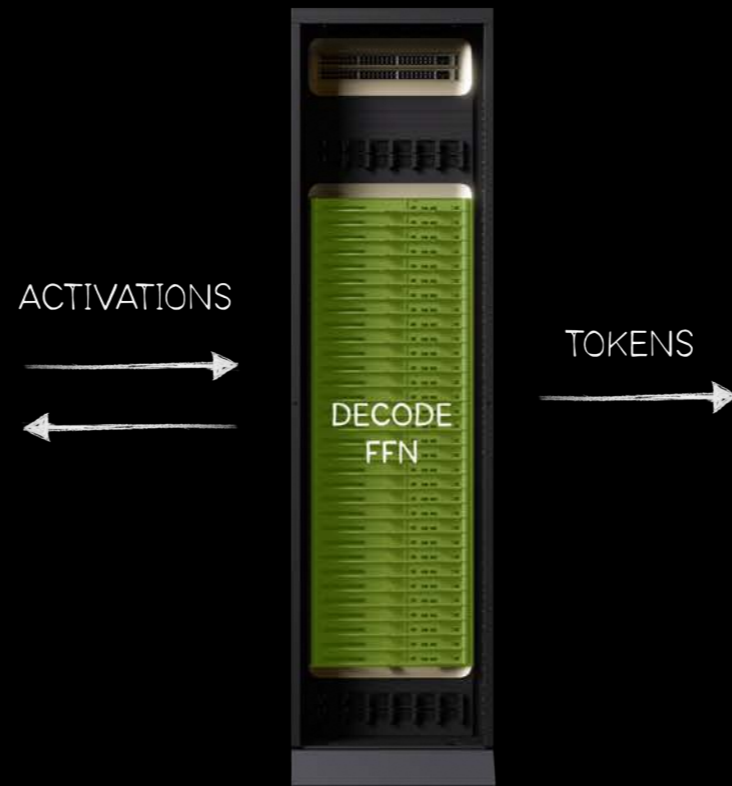
### NVIDIA Dynamo



### Vera Rubin NVL72



### Groq 3 LPX



**“Dynamo boosts inference performance by up to 7x, lowering token cost and increasing revenue.”**

Stock Titan

NVIDIA Dynamo is the operating system of the AI factory. It orchestrates GPUs and memory across rack-scale systems to run generative and agentic inference at scale. With Blackwell and GB300 NVL72, it drives up to 50x performance gains. Integrated into an open ecosystem, it ensures maximum throughput and efficiency.

# “The Inference King has been crowned.”

SemiAnalysis

GB300 NVL72 redefines the economics of inference. SemiAnalysis’ InferenceX testing shows performance per watt improving by nearly 50x in real-world workloads. At gigawatt scale, that is decisive.

Higher throughput within a fixed power envelope lowers cost per token and expands revenue.

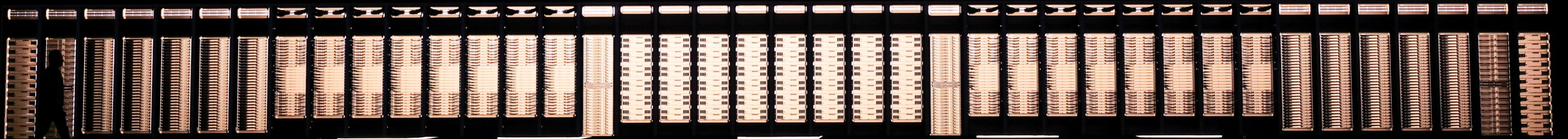
Cost per token defines leadership.



**“Vera Rubin is 10x more efficient than its predecessor.”**

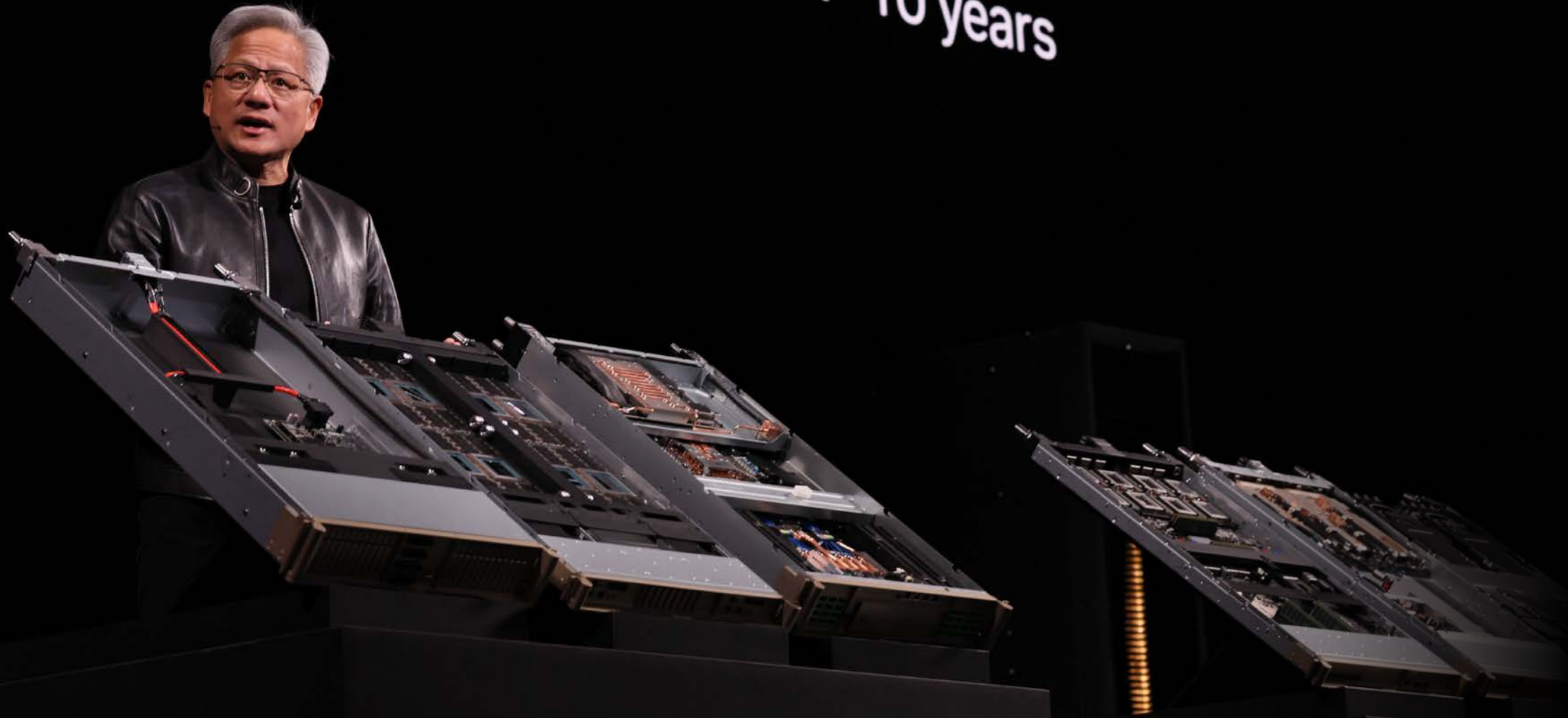
CNBC

Vera Rubin is not just an NVL72 rack. It is a full AI factory pod. Blackwell systems scaled at the rack level. Rubin scales at the pod level—integrating compute, memory, networking, and software into a single production unit. This shift increases throughput, reduces cost per token, and redefines how AI factories are built.



# NVIDIA Vera Rubin

40,000,000X - 10 years



**“Vera Rubin will deliver 10x higher inference throughput per watt and roughly one-tenth the cost per token.”**

Melius

Vera Rubin is a fully integrated AI computing system. Hardware and software are co-designed end to end, operating as one architecture for reasoning and agentic AI at scale.

Seven chips work together as one system: CPU, GPU, LPU, NVLink, SuperNIC, DPU, and Spectrum-X. This level of integration defines performance, efficiency, and AI factory economics. NVIDIA’s platform spans CUDA-X, DSX infrastructure, and Vera Rubin systems.

# “NVIDIA reinvents the CPU for the age of agentic AI.”

SiliconANGLE

Vera is our first fully custom data center CPU. It is built for the AI factory.

Agentic systems need a control plane—workflow orchestration, memory management, CPU-native execution alongside accelerated compute. Vera integrates tightly with Rubin systems to deliver predictable latency and coherence at scale. This completes the architecture.



## Announcing NVIDIA Vera CPU Launch Partners

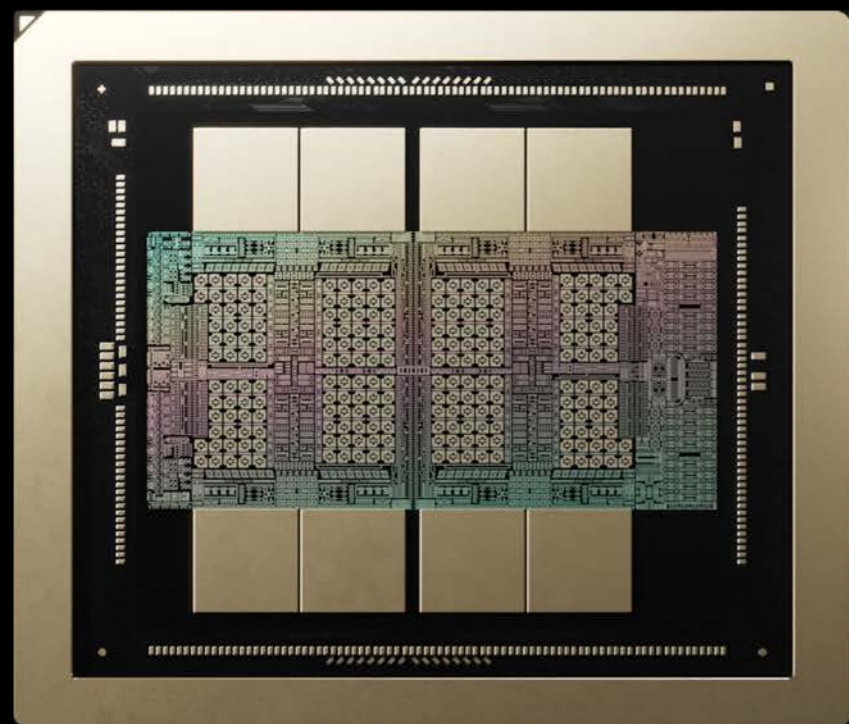
### Cloud



### OEMs and ODMs



Rubin GPU



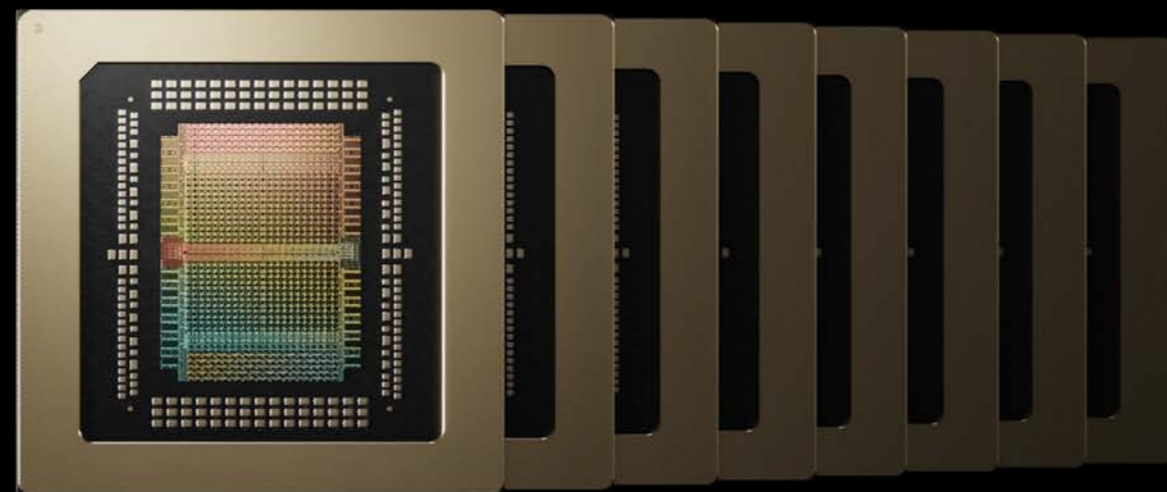
288 GB HBM4

22 TB/s

50 PFLOPs (NVFP4)

336B Transistors  
+ 2.5T (HBM4)

Groq 3 LPU



4 GB SRAM

1,200 TB/s SRAM Bandwidth

9.6 PFLOPs (FP8)

784B Transistors

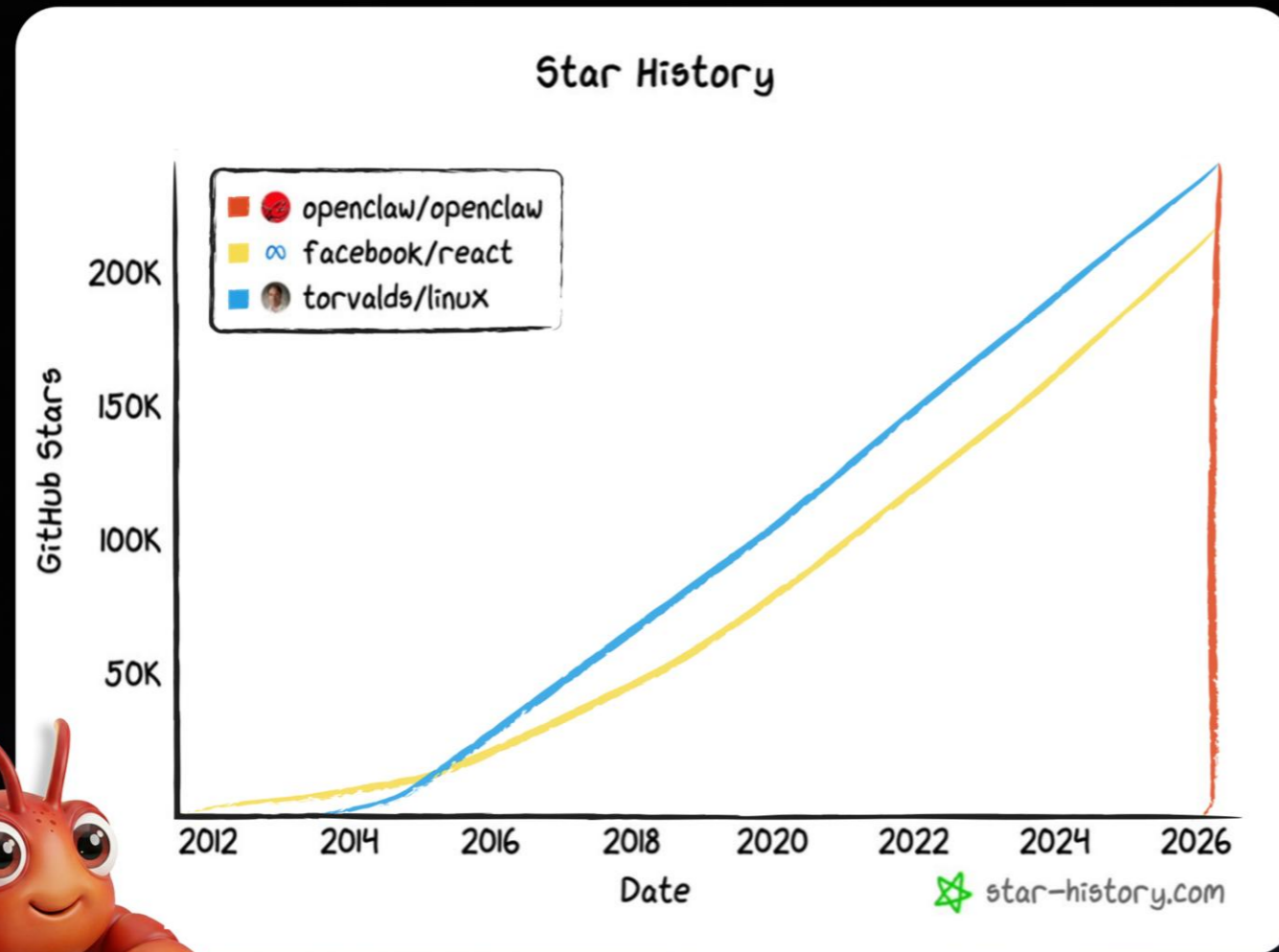
**55X**

“Groq will play a role similar to Mellanox, extending the architecture as an ‘accelerator’ for low-latency decode.”

Wccftech

Inference is no longer monolithic. Reasoning and token generation run on different architectures, each optimized for its task. Dynamo orchestrates these systems, coordinating compute, memory, and networking for maximum throughput. Groq extends NVIDIA’s infrastructure stack, much like Mellanox did for networking, adding a new pillar for low-latency decode. Together, output rises and cost per token falls within the same power envelope.

## Inference Inflection Arrives



**“NVIDIA goes all-in on agents at GTC with toolkits, OpenClaw and models.”**

SDxCentral

AI is moving beyond chatbots. It now reasons, uses tools, and executes multi-step tasks. It doesn't just respond—it works continuously. This is the inflection point for agentic AI.

OpenClaw is the operating system for this new class of systems. It standardizes how agents are built, orchestrated, and deployed at scale—marking the shift from prompt-driven AI to autonomous software.

# Announcing NVIDIA NemoClaw Reference OpenClaw

NVIDIA Agent Toolkit for Building Specialized Agents

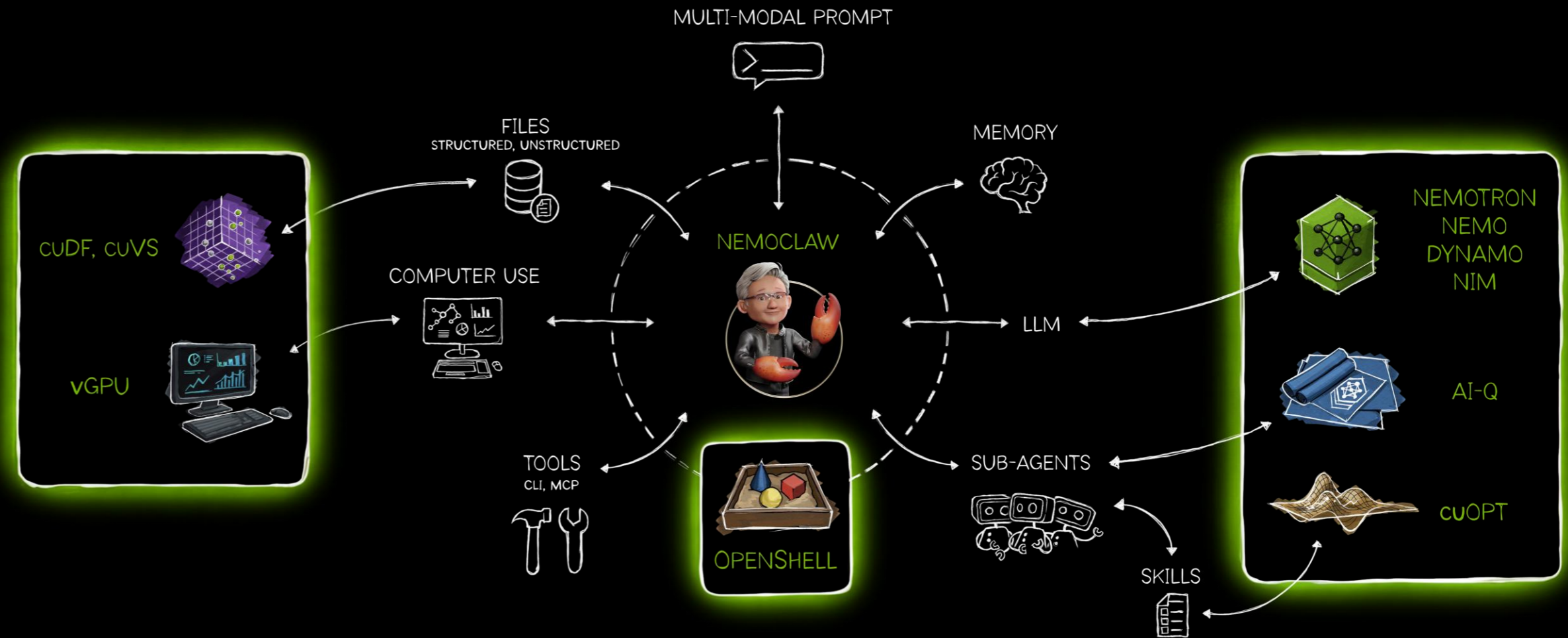
“NemoClaw brings security, scale to the agent platform taking over AI.”

VentureBeat

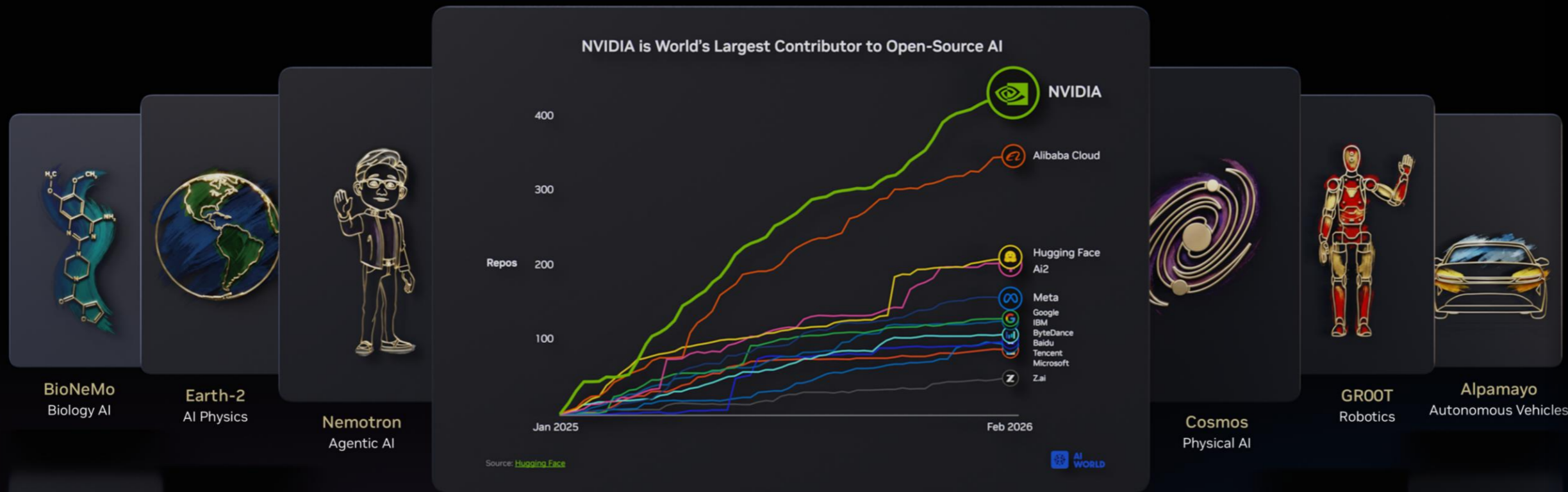
OpenClaw makes agentic systems possible. NemoClaw makes them deployable.

Agents can access sensitive data, execute code, and communicate externally. That requires enterprise-grade guardrails.

NemoClaw integrates policy enforcement, privacy controls, and network protections directly into the runtime—enabling safe, scalable adoption across cloud and on-prem environments.



## NVIDIA Leaderboard Topping Open Models



“Open ecosystems drive hardware adoption. What’s new is they now extend beyond compute into models, agents, and orchestration.”

Bain & Company

Open frontier models and OpenClaw have moved AI from generation into execution. Models reason, agents take action, and connect to tools, memory, and workflows at scale. NVIDIA’s open model ecosystem enables developers and enterprises to build and deploy agentic systems across industries.

## The World Building Regional AI With NVIDIA Nemotron



“The imperative is clear: sovereign AI is essential for economic growth.”

Civo

AI is expanding beyond centralized labs and proprietary platforms. AI natives, enterprises, industrial systems, sovereign initiatives, and regional cloud providers are building infrastructure tailored to their industries and data. Open frontier models enable this shift, allowing AI to be built and deployed across languages, domains, and regulatory environments.

Through the Nemotron Coalition, NVIDIA is helping developers and enterprises build, customize, and scale AI across regions and industries.

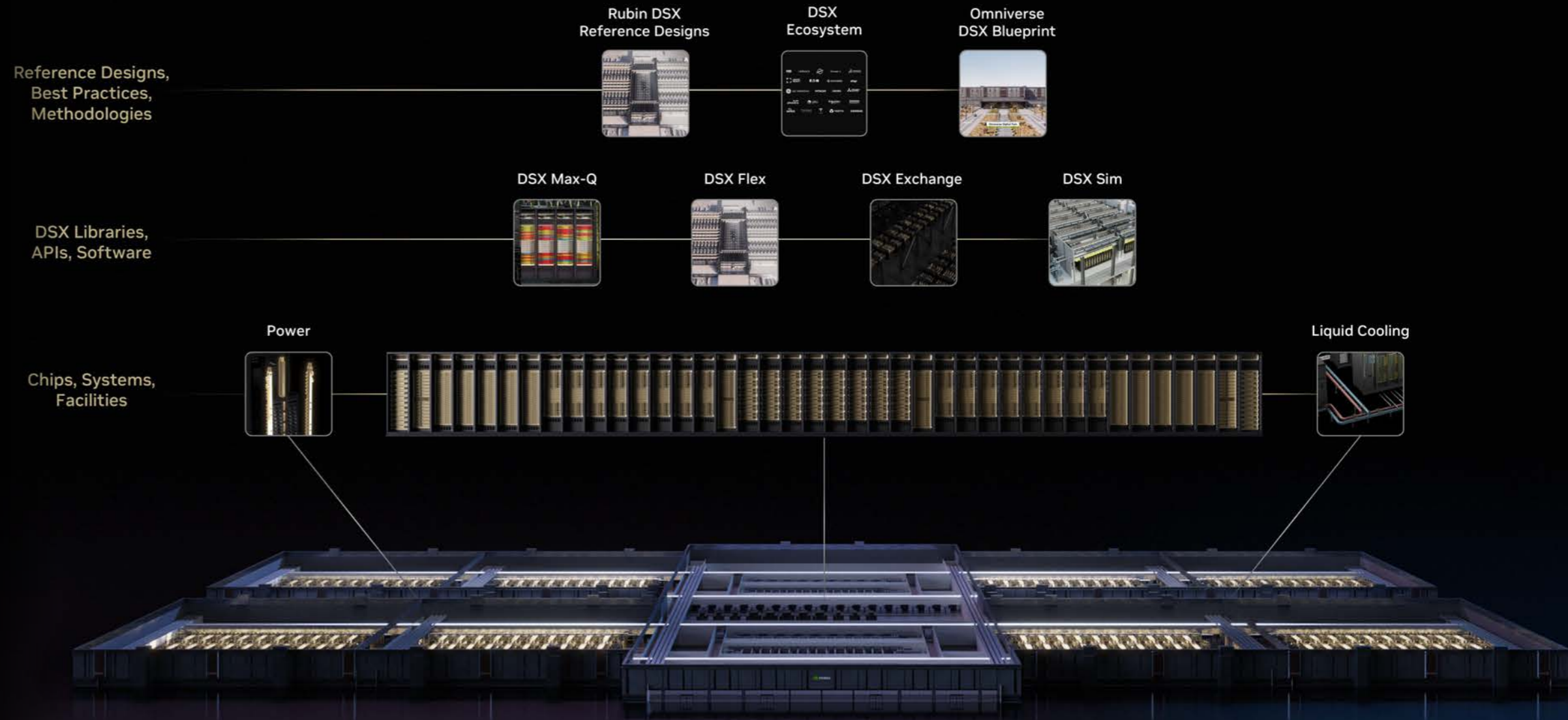
“A turnkey solution to help plan, build and maintain AI factories.”

Forbes

DSX defines how AI factories are built as a system. It integrates compute, networking, storage, power, and cooling into a single architecture—from rack to pod to gigawatt scale. With simulation and co-design, factories are optimized before deployment, maximizing tokens per watt and ensuring the highest throughput and lowest cost per token.

## NVIDIA DSX AI Factory Platform

Extreme Co-Design at Infrastructure Scale



# “NVIDIA expands robotics ecosystem at GTC as physical AI moves toward large-scale deployment.”

TrendForce

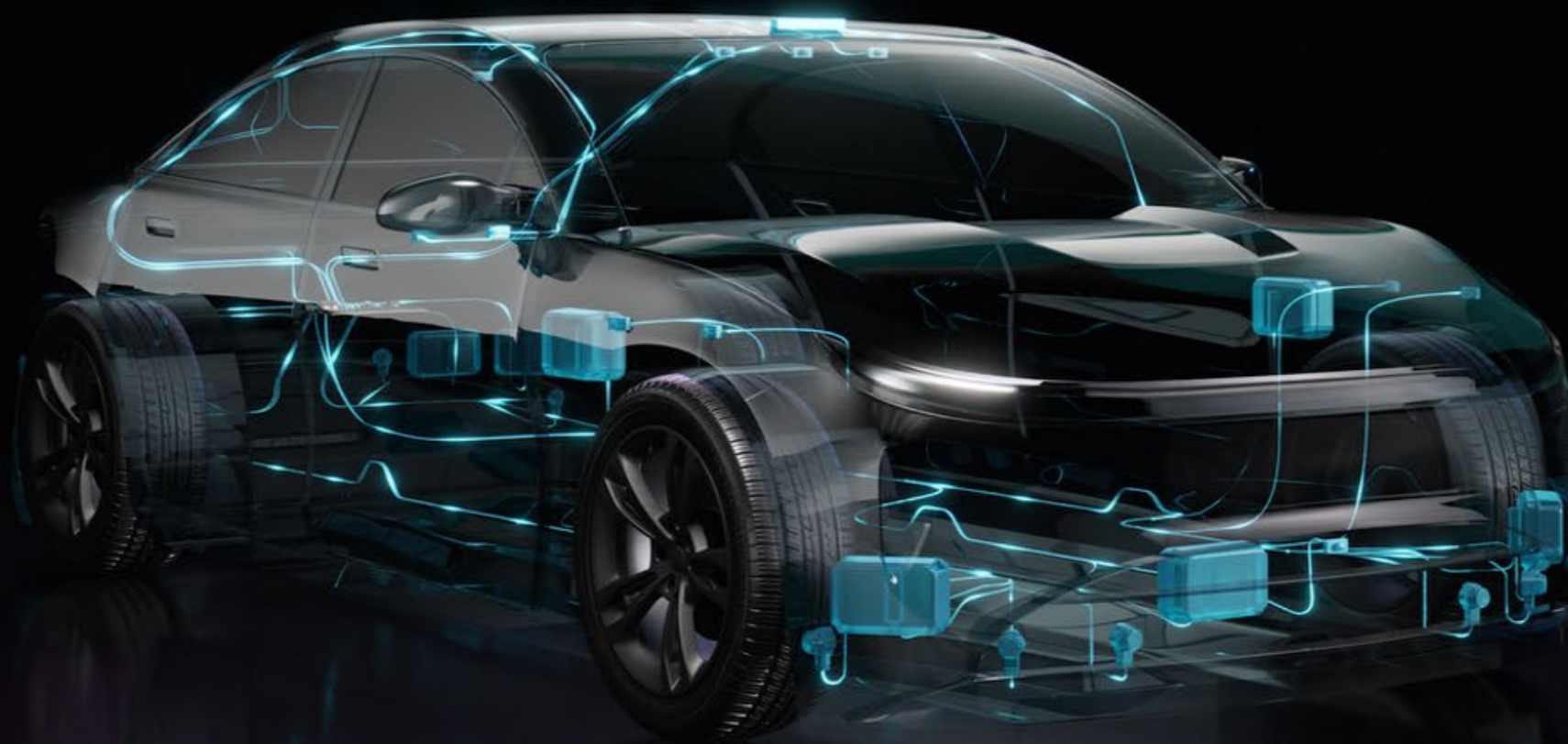
Physical AI is the next wave of AI. It is moving from simulation into the real world. Robotics, industrial automation, and autonomous systems are scaling across manufacturing, logistics, and mobility. NVIDIA works with virtually every major robotics developer. Its full-stack platform spans training, simulation, and deployment, enabling a global ecosystem where robots become a new class of workforce.



# “Uber, NVIDIA plan robotaxi rollout in 28 cities starting next year.”

Reuters

Uber and NVIDIA are scaling autonomous mobility worldwide. Powered by NVIDIA DRIVE, robotaxi fleets are rolling out across 28 cities, bringing autonomous transportation into everyday use. This marks the transition from pilot programs to global deployment, as autonomous systems move from demonstration into commercial infrastructure at scale.



## “NVIDIA betting big on ‘physical AI’.”

Punchbowl News

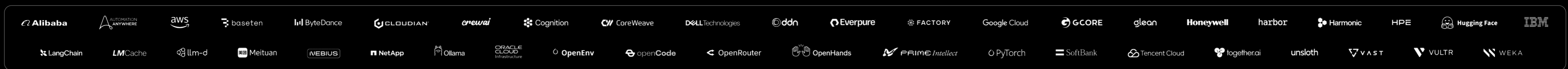
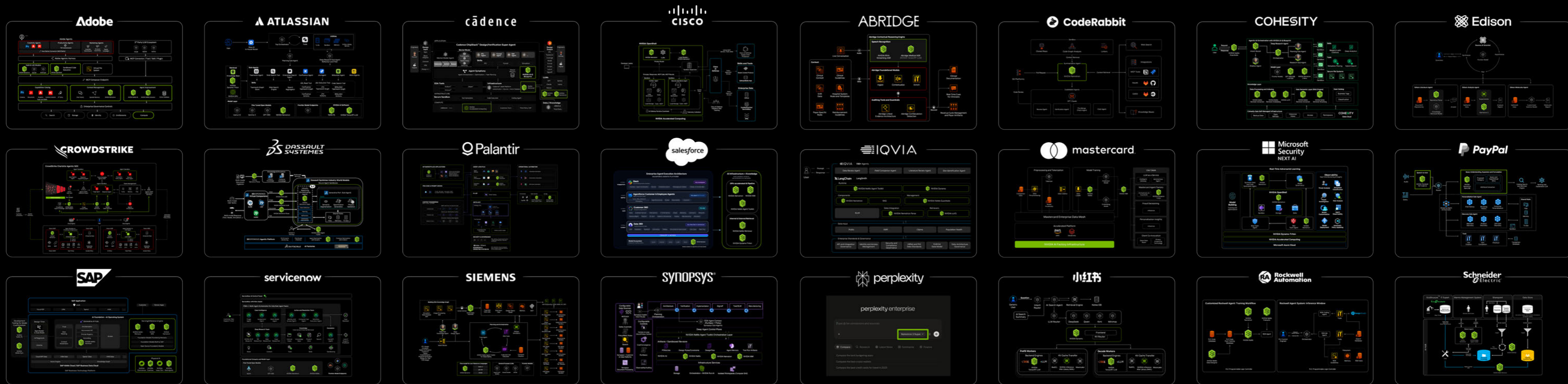
Physical AI is moving from simulation into the real world. In partnership with Disney, NVIDIA brought Olaf to life using physics-based simulation, foundation models, and real-time robotics. At the same time, automakers and mobility leaders are standardizing on NVIDIA’s AV platform to scale autonomous systems globally. Physical AI is no longer experimental. It is entering production across industries.



# “A vast ecosystem of hardware, software, and industrial partnerships.”

DIGITIMES Asia

NVIDIA's ecosystem spans AI natives, enterprises, frontier model builders, and global cloud providers. Developers build on CUDA and NVIDIA AI software, bringing new workloads to AI factories worldwide. As token demand grows, compute scales with it—connecting customers, infrastructure, and cloud into a single, global production system. This ecosystem is what enables NVIDIA to scale AI globally.



## What the analysts said:

“NVIDIA-based inference has clear cost per token leadership that gets better with Rubin.”

Morgan Stanley

“Continued leadership backed by broadening full-stack end-to-end pipeline, extreme co-design with customers, and supply assurance.”

BofA

“Strong growth outlook for 2027, consistent with our estimates and well above the Street.”

Goldman Sachs

“The \$1T+ in Blackwell and Vera Rubin purchase orders/demand through CY27 is explicitly a floor.”

JPMorgan

“We came out of the keynote reassured in NVIDIA’s roadmap and continued ability to innovate faster than the competition.”

Citi

“Super Bowl atmosphere, impromptu concert, and technical breakthroughs ... GTC at its finest.”

Cantor

“Frankly we increasingly wonder how anyone else can compete with this.”

Bernstein

“NVIDIA is the clear end-to-end infrastructure partner with the enabling software stack.”

Melius

## What the press said:

“NVIDIA is sustaining its leadership in the AI chip market.”

Reuters

“NVIDIA is the engine powering the AI revolution.”

Lex Fridman Podcast

“Vera Rubin offers a stunning 40,000,000x compute growth within a decade.”

Wccfttech

“\$1 trillion won't be enough to meet AI demand.”

Fortune

“NVIDIA remains at the forefront of the AI demand curve for 2026 and beyond.”

Handelsblatt

“Every industry is here. Every tech company is here. Every AI company is here.”

All-In Podcast

“NVIDIA is the default platform where businesses build serious AI systems.”

The Street

“No other event has brought together as many industries, companies, AI influencers, VCs, and startups in one place.”

eWEEK

436

Exhibitors and  
Sponsors

45K

Virtual  
Attendees

39M

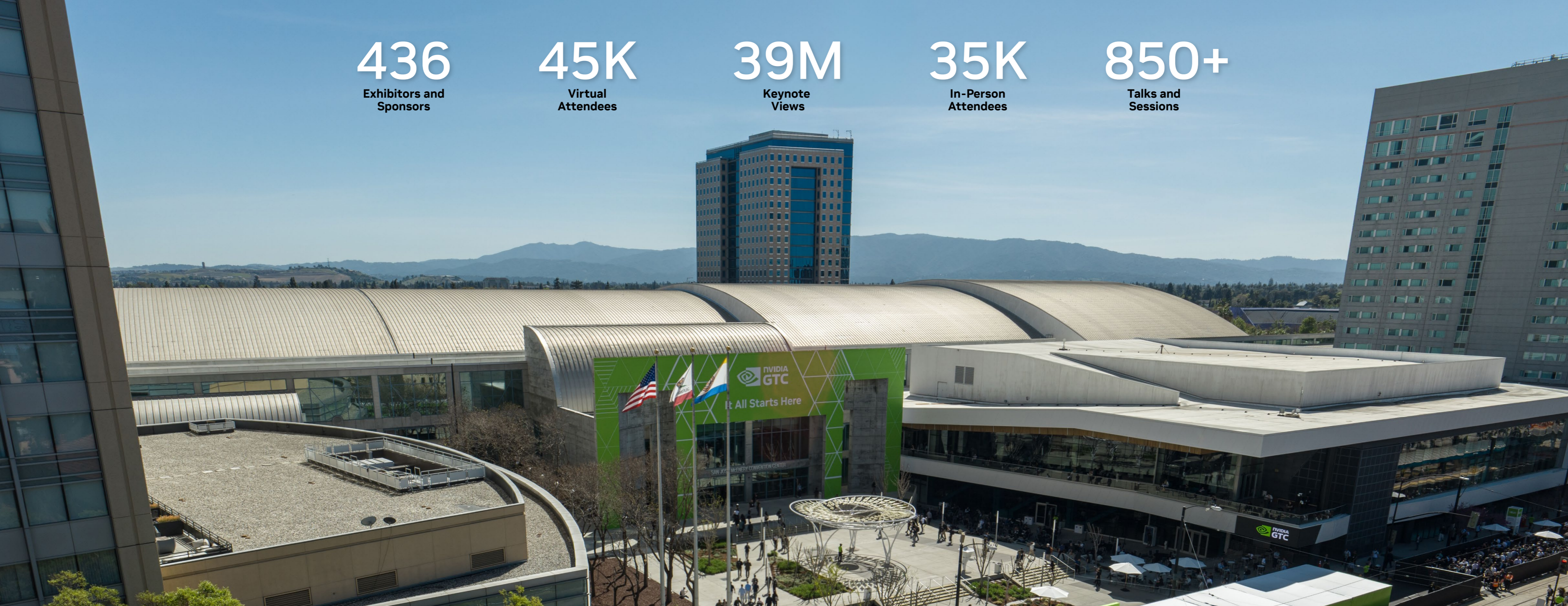
Keynote  
Views

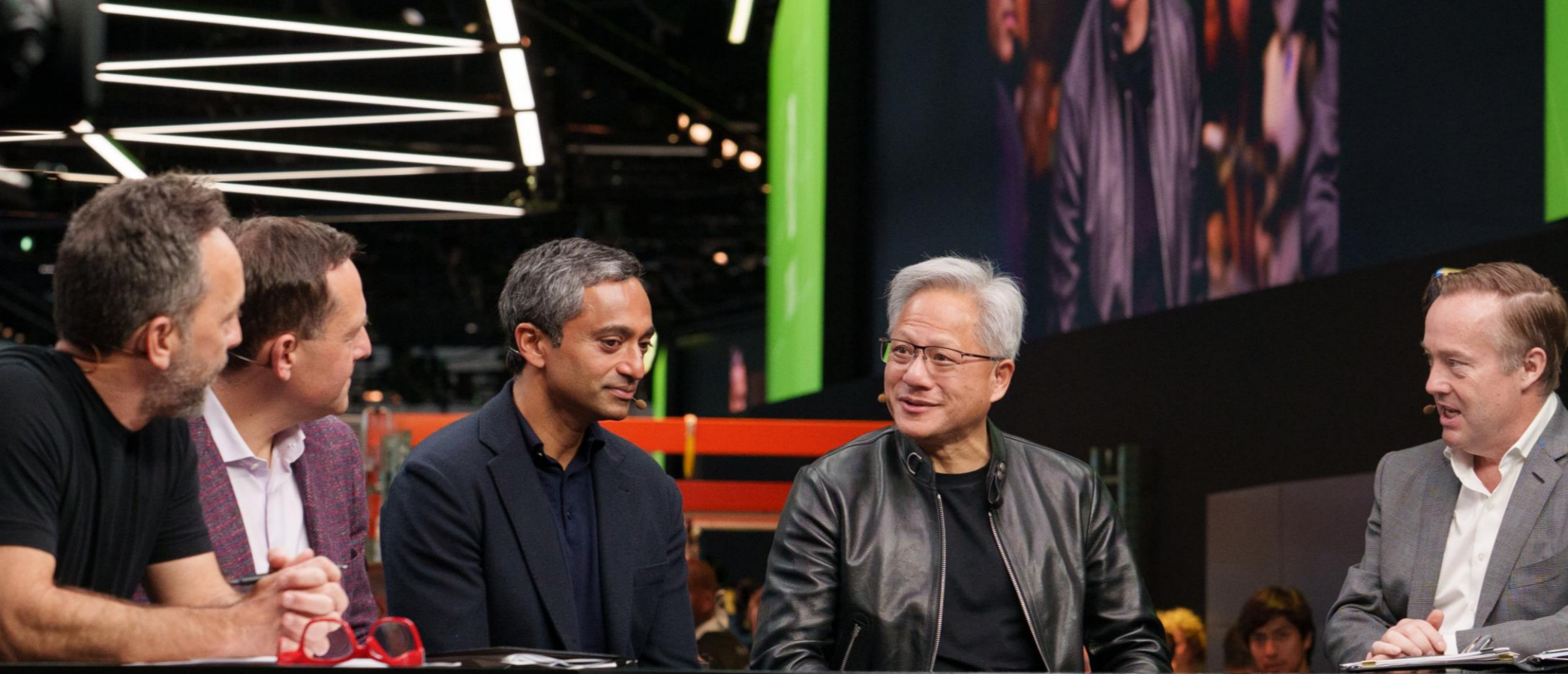
35K

In-Person  
Attendees

850+

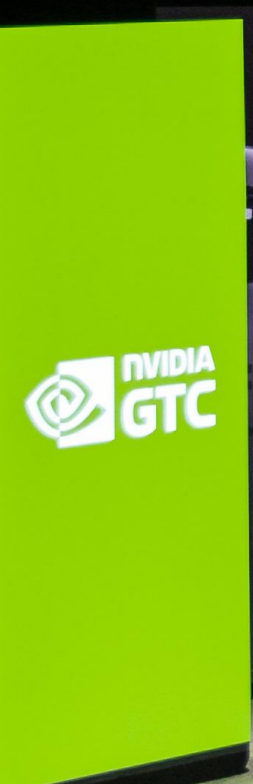
Talks and  
Sessions



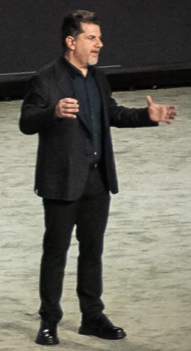
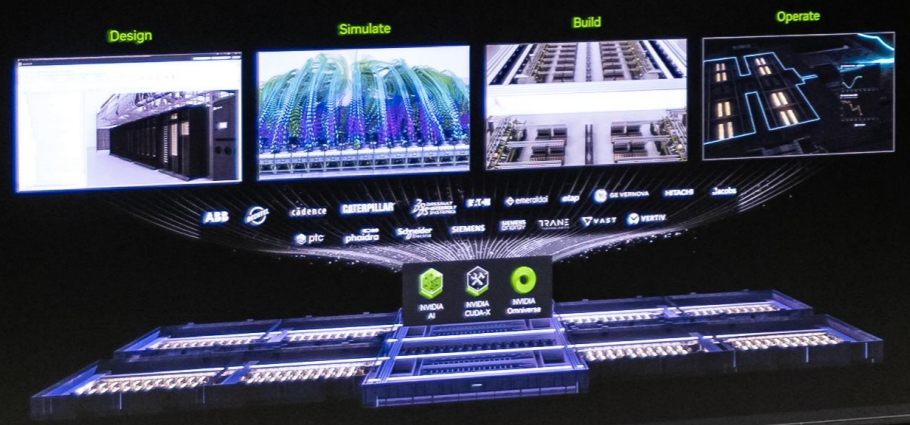


**ALL-IN**





## AI Factories Turn Compute into Revenue



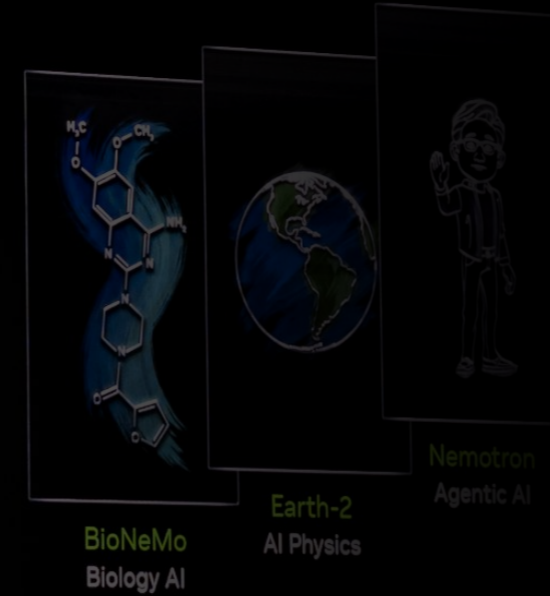
**“Every industrial company will become a robotics company.”**

Manufacturing Dive

“Factories themselves are now robotic systems. If you take the combination of all of those, you can combine them all together to create a system, a factory that’s essentially a robot itself.”

Rev Lebedian, NVIDIA

NVIDIA Leading Open M



## “Agentic AI inflection hits healthcare and life sciences.”

Genetic Engineering & Biotechnology News

“The digital health agent boom is here. It’s creating a new paradigm in science. What is so awesome about this technology is that the way it’s architected is a completely new way of developing software, and it has opened the possibilities for it to be trusted and applied in healthcare from almost day one.”

Kimberly Powell, NVIDIA

with NVIDIA Classical AV Stack



NVIDIA AlpaMax proposes driving behavior while NVIDIA Classical AV Stack intervenes when safety constraints are at risk.

Example:  
A cyclist suddenly moves into the lane, triggering the safety layer to maintain safe distance.



“I took a ride in an NVIDIA-powered autonomous Mercedes at GTC 2026 and it’s convinced me this is the future.”

TechRadar

“AV will be the first mass-produced and mass-deployed physical AI technology and application. And as of today, only 0.006% of the mileage is driven autonomously. We do believe that in the near future, every mile, everything that moves, will become autonomous.”

Xin Zhou Wu, NVIDIA

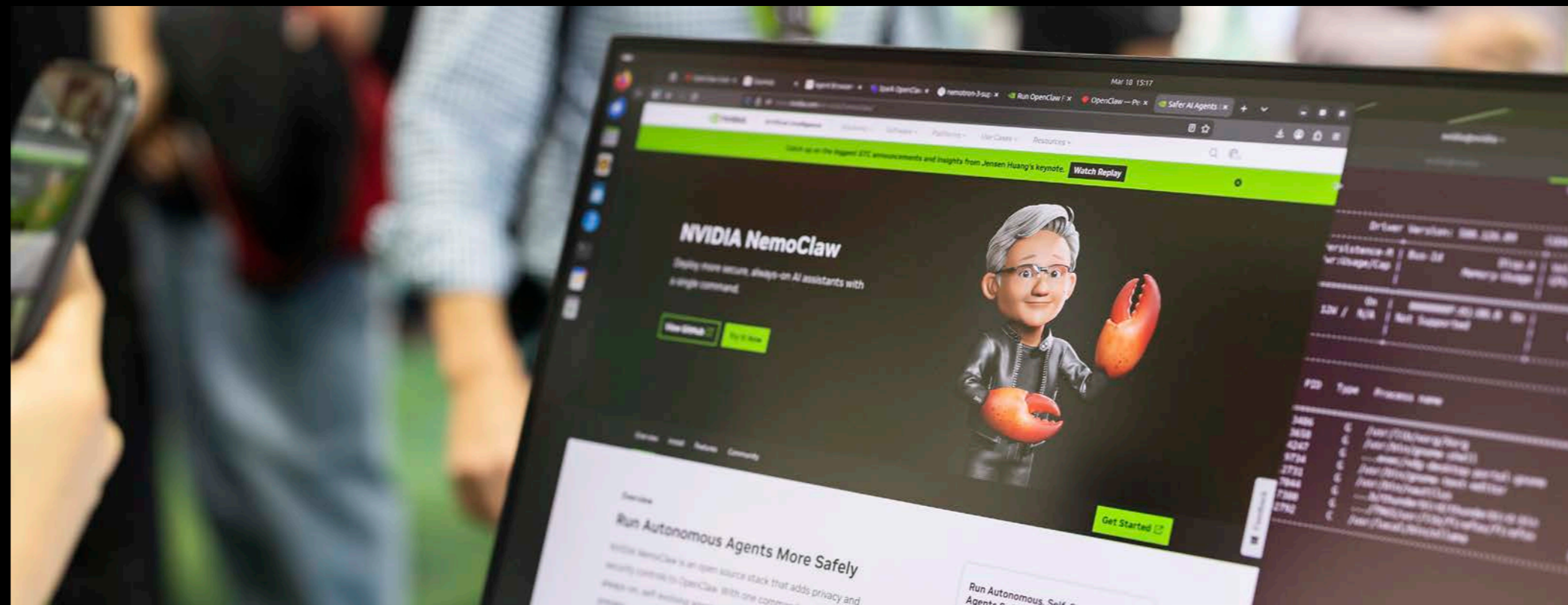
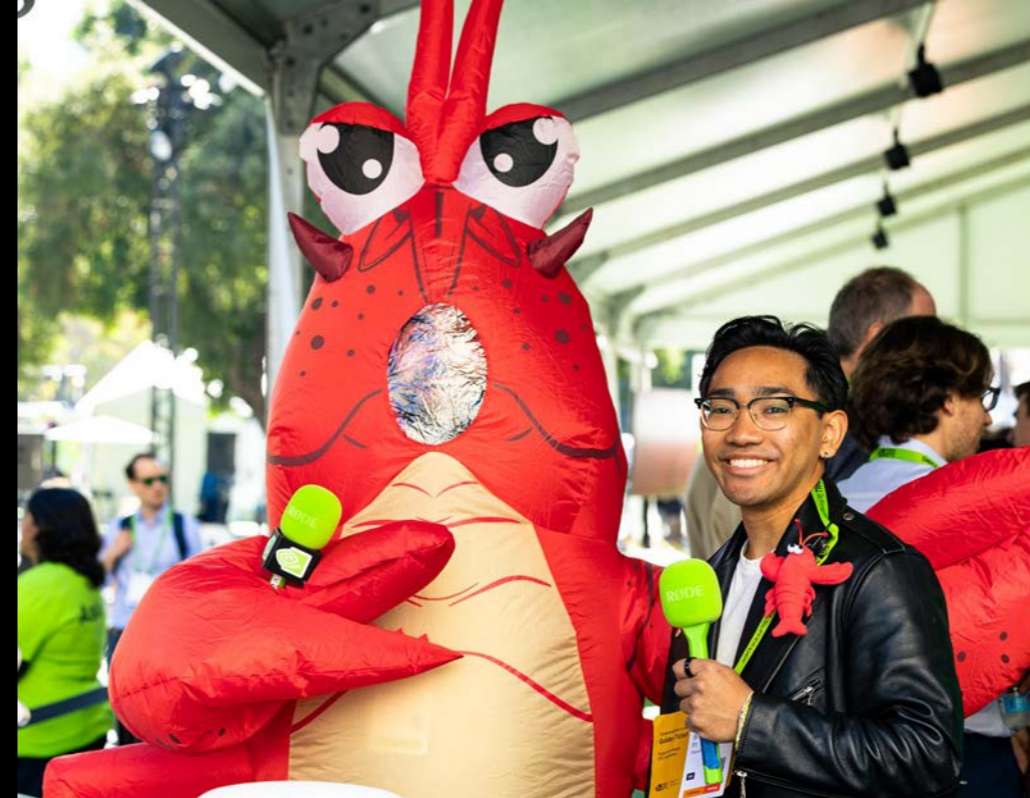
















**This is the next phase of computing.**

AI is infrastructure. A five-layer stack from energy to applications. Inference has become thinking, and every interaction generates tokens. AI factories convert energy into tokens, producing intelligence at scale.

Throughput lowers cost per token and increases revenue. In the AI factory era, tokens per second per watt translate directly into revenue. Through extreme co-design, NVIDIA delivers the lowest cost per token and the highest revenue per watt.

NVIDIA builds the full platform—CUDA-X, DSX infrastructure, and Vera Rubin systems—connected to a global ecosystem.

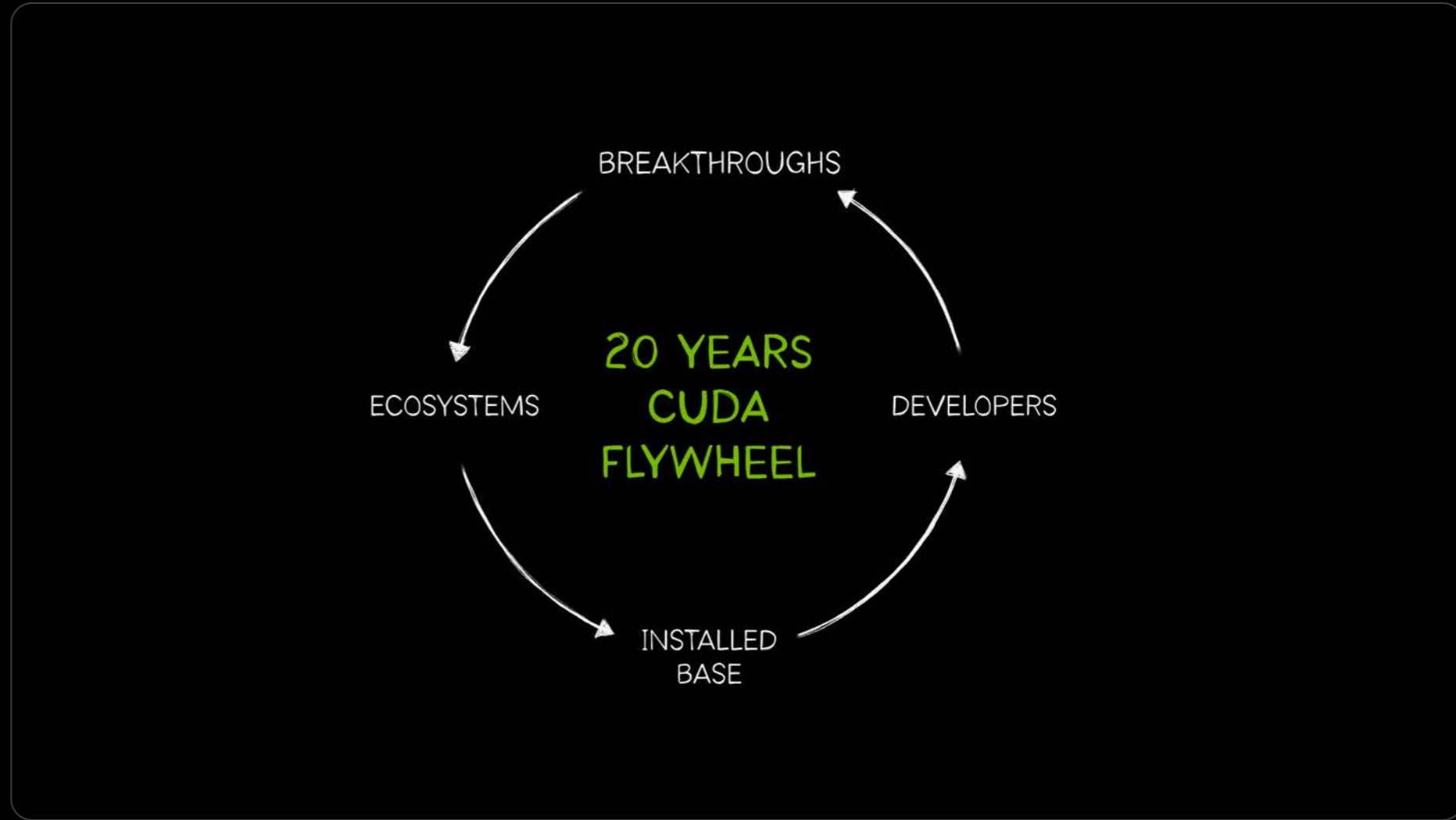
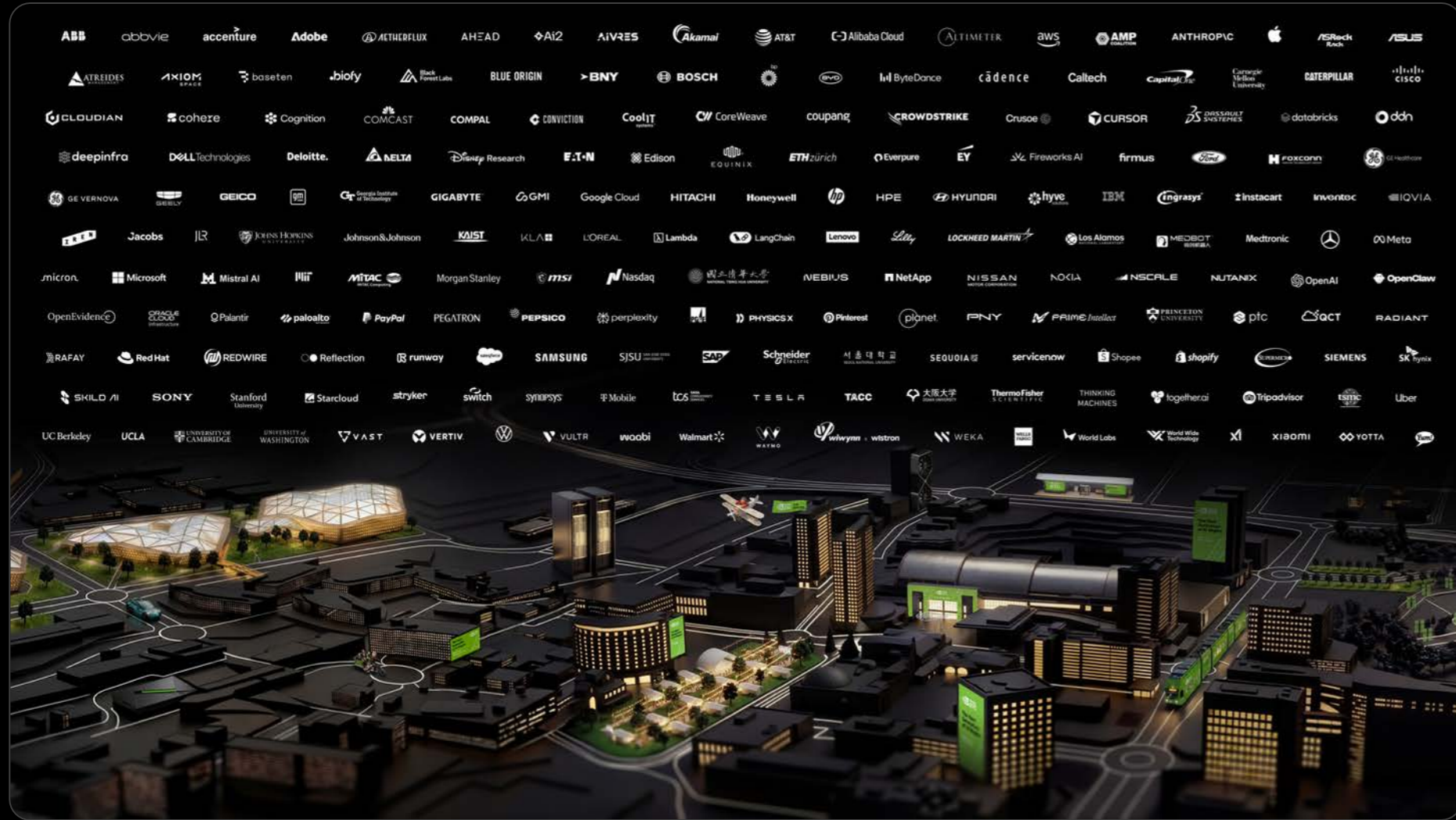
Agentic AI is here. Physical AI is next. We are just at the beginning.

A white, handwritten signature of Jensen Huang, the CEO of NVIDIA, is centered below the text. The signature is fluid and cursive, written in a light color against the dark background.



# NVIDIA GTC 2026





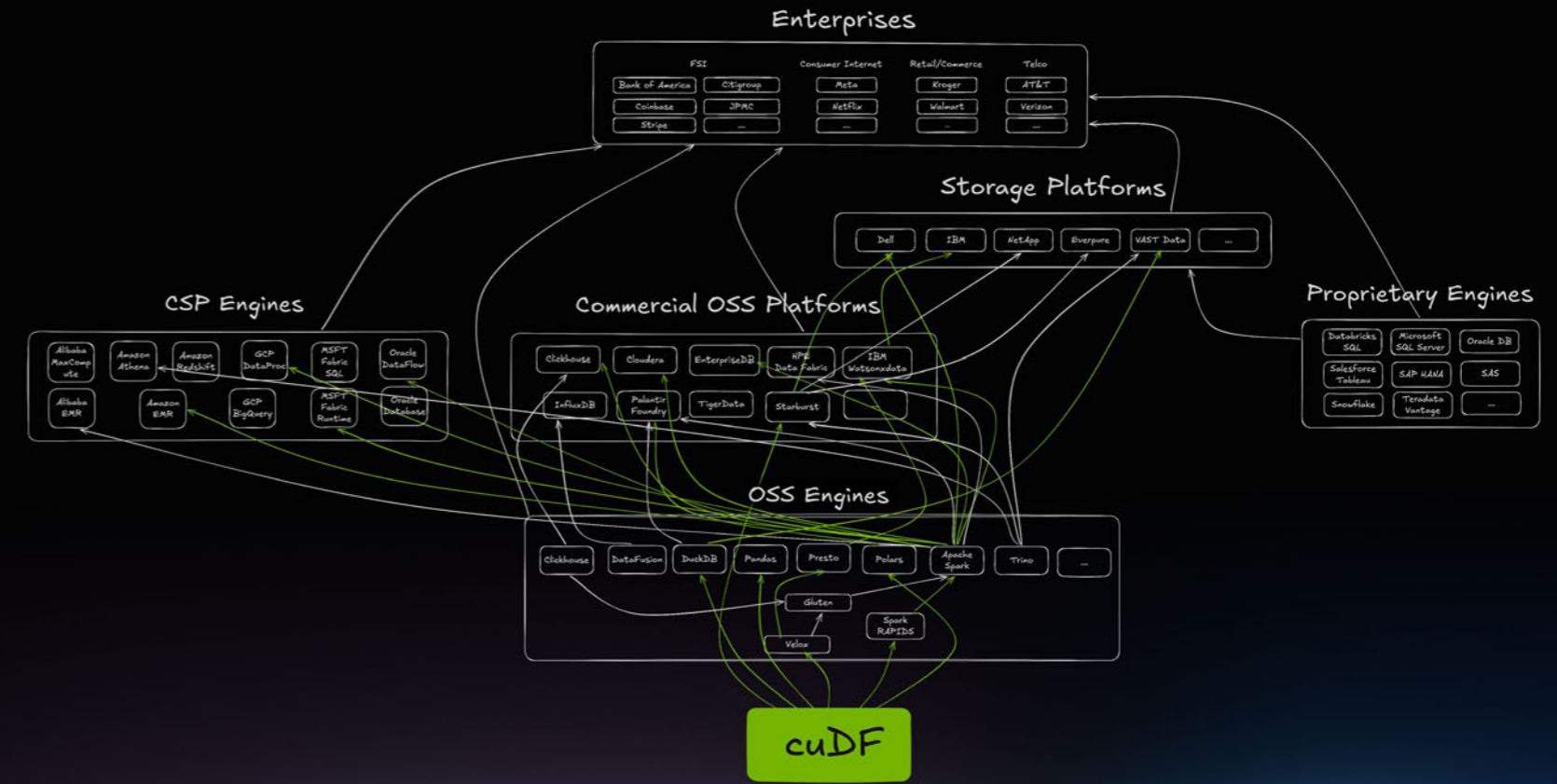


Announcing DLSS 5  
3D-Guided Neural Rendering



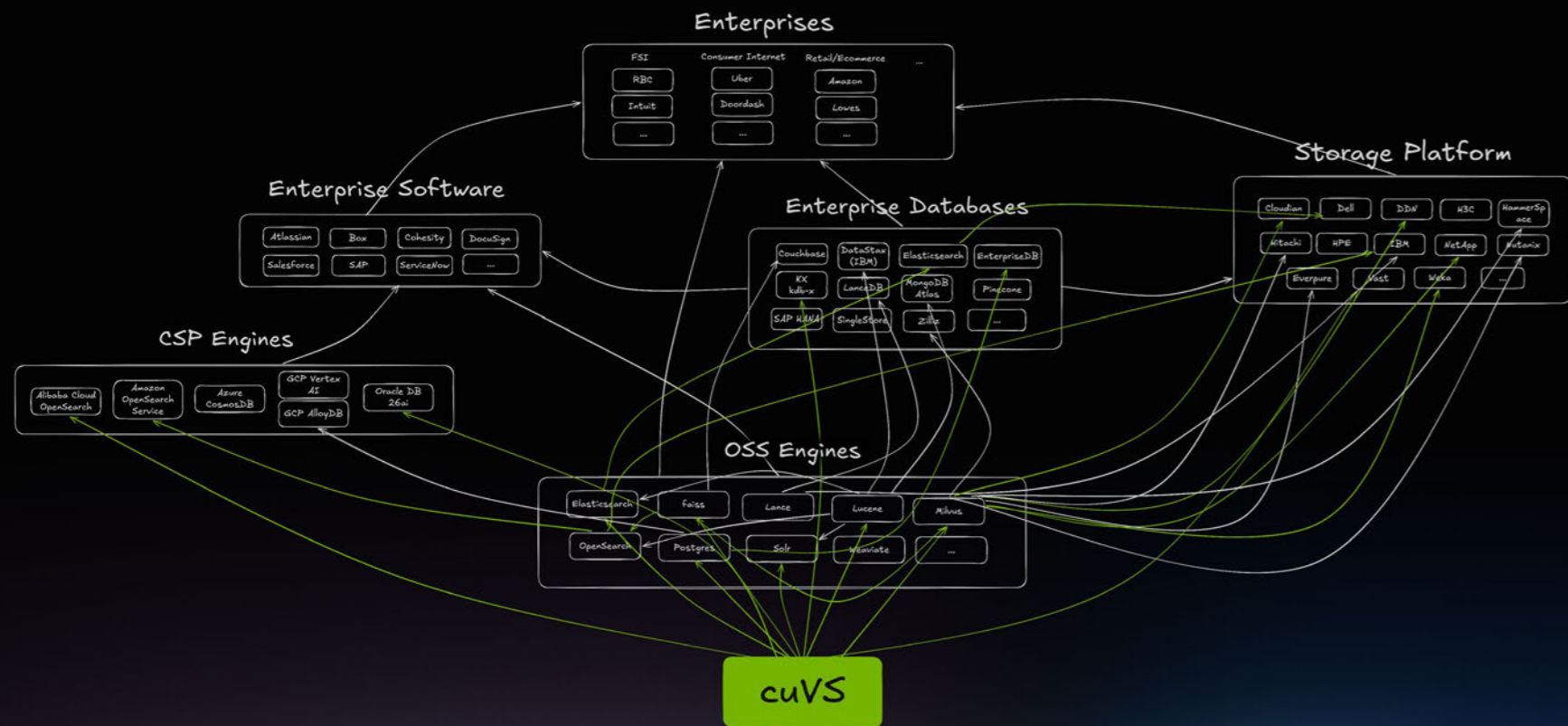
DLSS 5 On

### Structured Data is the Ground Truth of AI \$120B Structured Data Ecosystem – The Ground Truth of Enterprises



Source: Gartner DBMS, 2025

## Unstructured Data is the Context of AI 100's of Zettabytes Per Year of Unstructured Data – Growing Exponentially

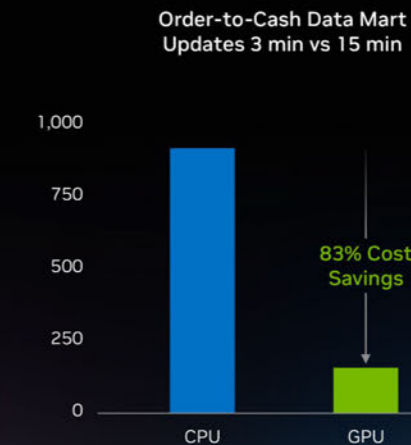
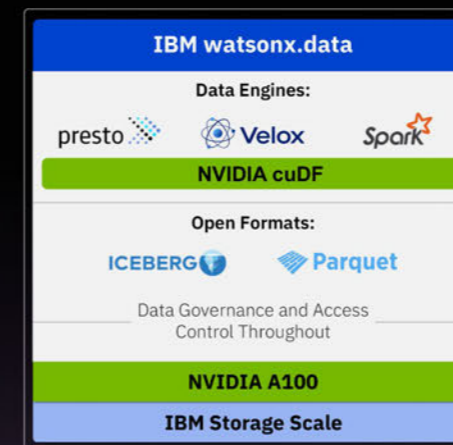


Source: IDC Global DataSphere, 2025



“Working with IBM and NVIDIA, early results show we can **refresh global operations data in minutes** at reduced cost—making a capability that can be turned into **tangible business impact** in areas such as manufacturing or warehousing.”

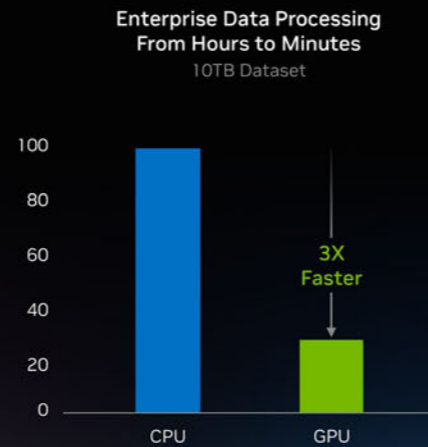
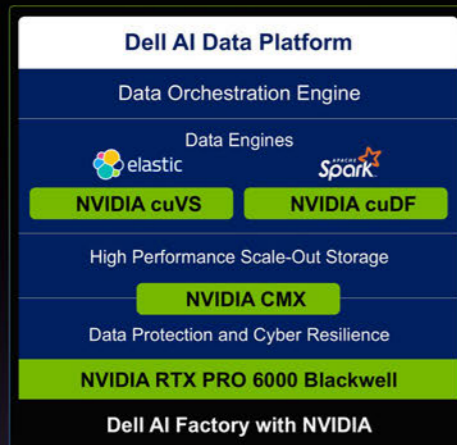
— Chris Wright, Chief Information and Digital Officer of Nestlé





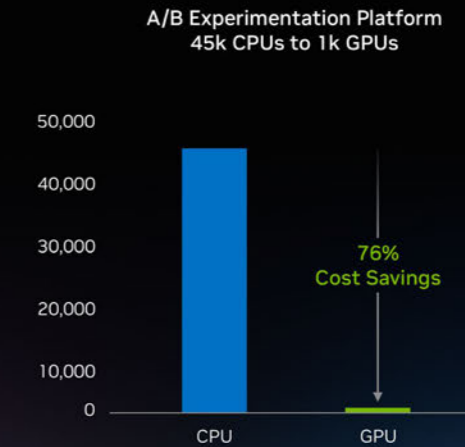
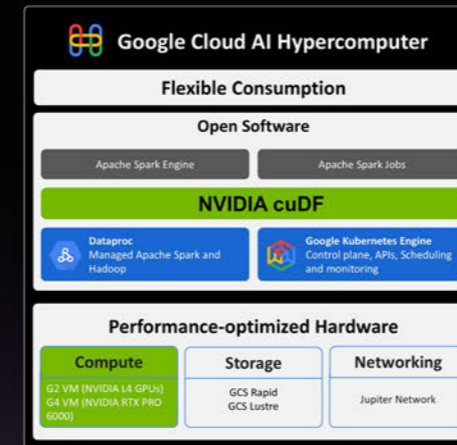
“Dell AI Data Platform with NVIDIA helps us build scalable and repeatable data pipelines that drive automation—processing massive data volumes in *minutes instead of hours* and delivering transformative value for our clients.”

— Abhijit Dubey, CEO of NTT DATA



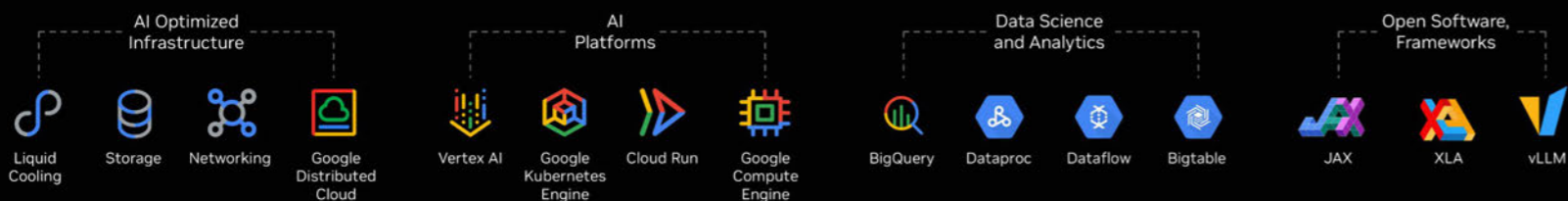
“Our collaboration with NVIDIA and Google Cloud helps us *innovate faster* for more than a billion Snapchatters worldwide. By *lowering costs and scaling experiments* across petabytes of data, we’re delivering AI-powered experiences more quickly and efficiently.”

— Saral Jain, CIO of Snap



**nvidia** ❤️ **Google Cloud**

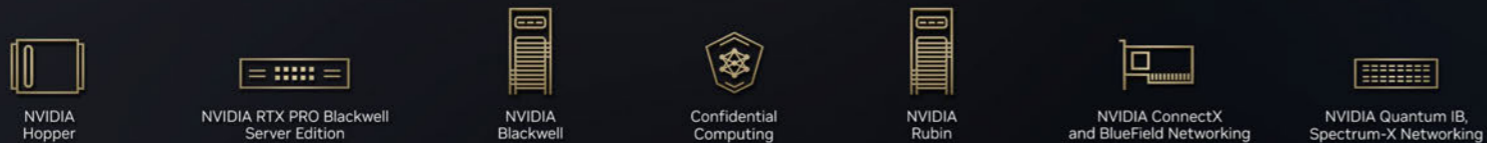
baseten CROWDSTRIKE gtm otto group PUMA Reve salesforce Schrödinger SNAPSHOT



System Builders and OEMs

**NVIDIA AI Platform**

cuDNN Magnum IO cuOpt cuVS cuDF Omniverse TRT-LLM Dynamo NeMo PhysicsNeMo Cosmos Nemotron



**nvidia** ❤️ **aws**

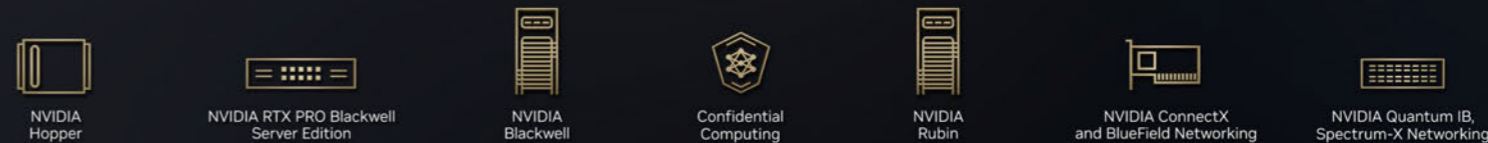
CROWDSTRIKE databricks HUMAIN OpenAI perplexity Palantir qodo salesforce snowflake

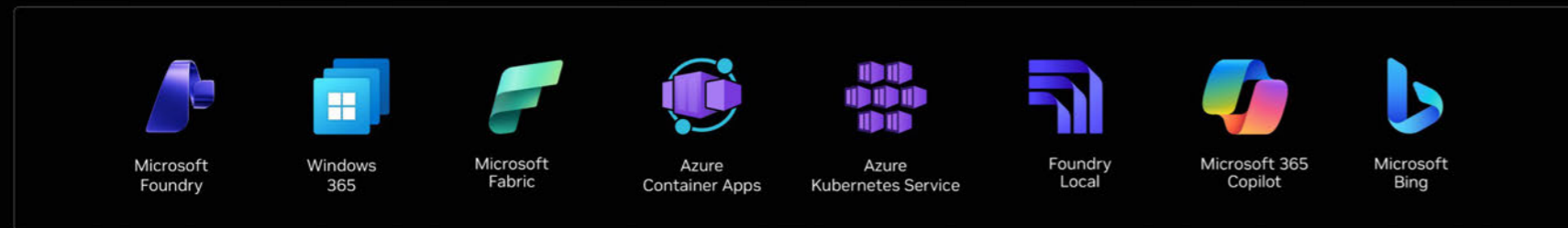
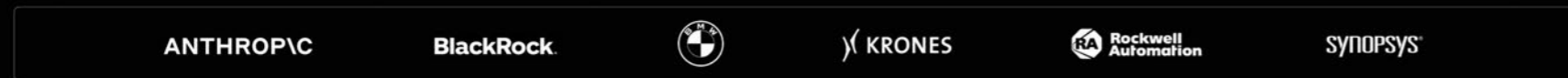


System Builders and OEMs

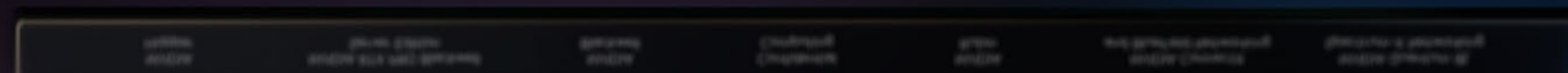
**NVIDIA AI Platform**

cuDNN Magnum IO cuOpt cuVS cuDF Omniverse TRT-LLM Dynamo NeMo PhysicsNeMo Cosmos Nemotron

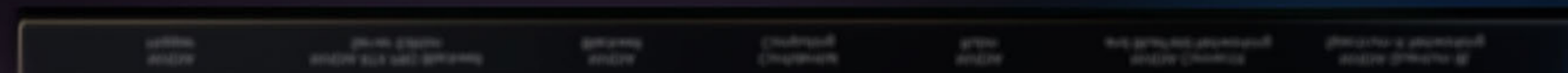
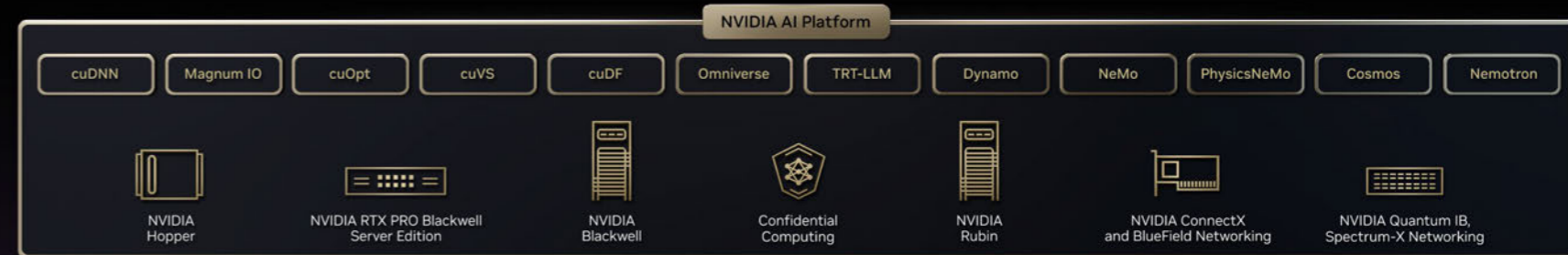




System Builders and OEMs

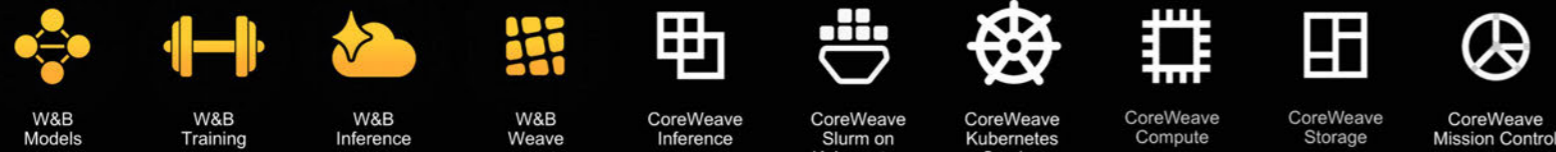


System Builders and OEMs





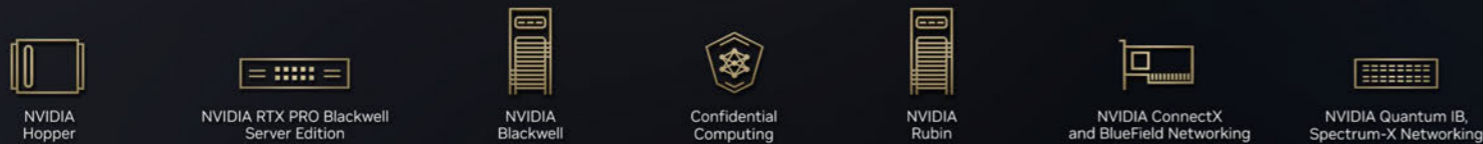
Canva cohere CURSOR mercado libre Mistral AI Morgan Stanley OpenAI runway



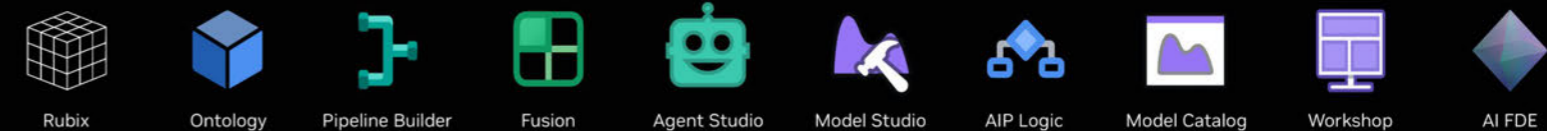
System Builders and OEMs

NVIDIA AI Platform

cuDNN Magnum IO cuOpt cuVS cuDF Omniverse TRT-LLM Dynamo NeMo PhysicsNeMo Cosmos NemoTron



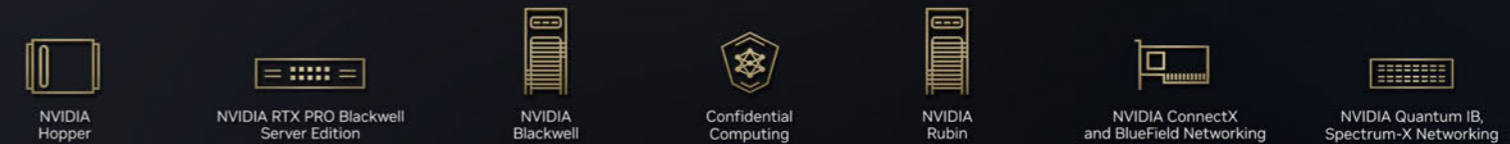
accenture CenterPoint Energy Hertz LOWE'S TETON RIDGE

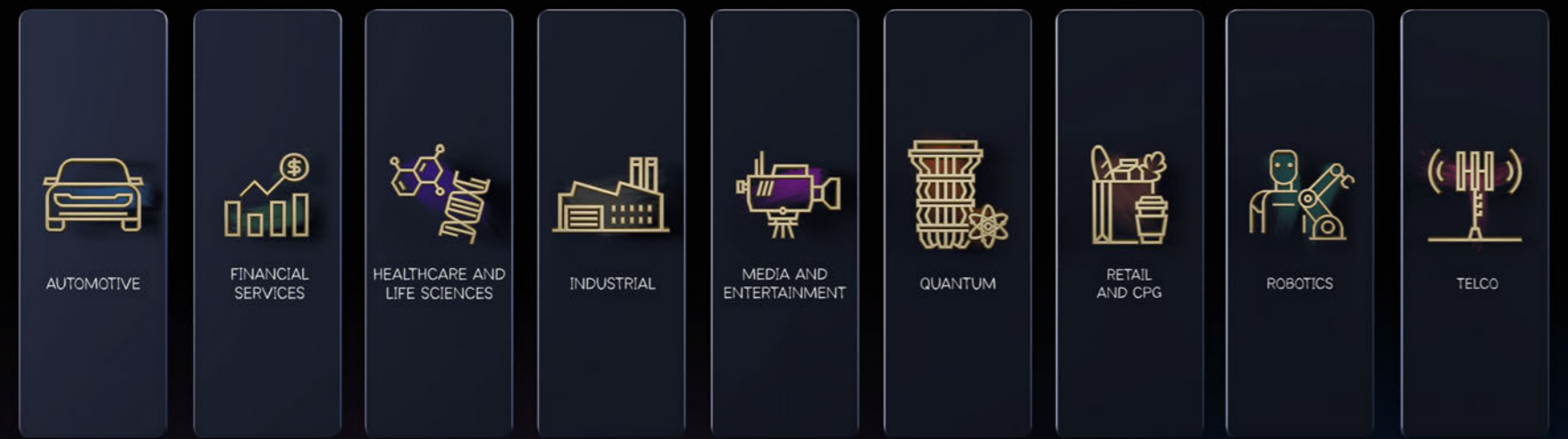


Dell

NVIDIA AI Platform

cuDNN Magnum IO cuOpt cuVS cuDF Omniverse TRT-LLM Dynamo NeMo PhysicsNeMo Cosmos NemoTron





# NVIDIA ❤️ AI Natives

AI for Auto: Auroro, Al momento, nuro, pony, woobl, WAYVE

AI for Customer Support: Decagon, IIElevenLabs, Fin, kore.ai, SIERRA, synthesio

AI for Engineering: BeyondMath, Emmi AI, luminary, PHYSICS X, ACURVE INTELLIGENCE

AI for Healthcare: Arc Institute, biohub, Boltz, Chai Discovery, Edison, LILA, Protein Design

AI for Robotics: AGIBOT, ASILE ROBOTS, FieldAI, FIGURE, Generalist, Physical Intelligence (x), SKILD AI

AI for Search: brave, gleon, perplexity, you.com

AI for Software Development: CodeRabbit, Cognition, cursor, FACTORY, greptile, Lovable, opencode, OpenHands, qodo, tabnine

DL Frameworks: PyTorch

Inference Frameworks: SGL, vLLM

Agent Frameworks/Protocols: A2A, AutoGen, cressai, LangChain, Model Context Protocol, OpenCLAW

Frontier Model Builders: AI2, ANTHROPIC, Back Forest Labs, cohere, Decart, deepsook, Gemma, humans&, Inception, KIMI, Magic, Mistral AI, OpenAI, poolside, Qwen, Reflection, runway, THINKING MACHINES, X

Model to Production: Anyscale, baseten, BigPanda, Braintrust, DataRobot, Distil, Fireworks AI, H2O.ai, LogicMonitor, Modal, NUTANIX, OpenRouter, Red Hat, Replicate, scale, snowflake, togetherai

NVIDIA AI Platform

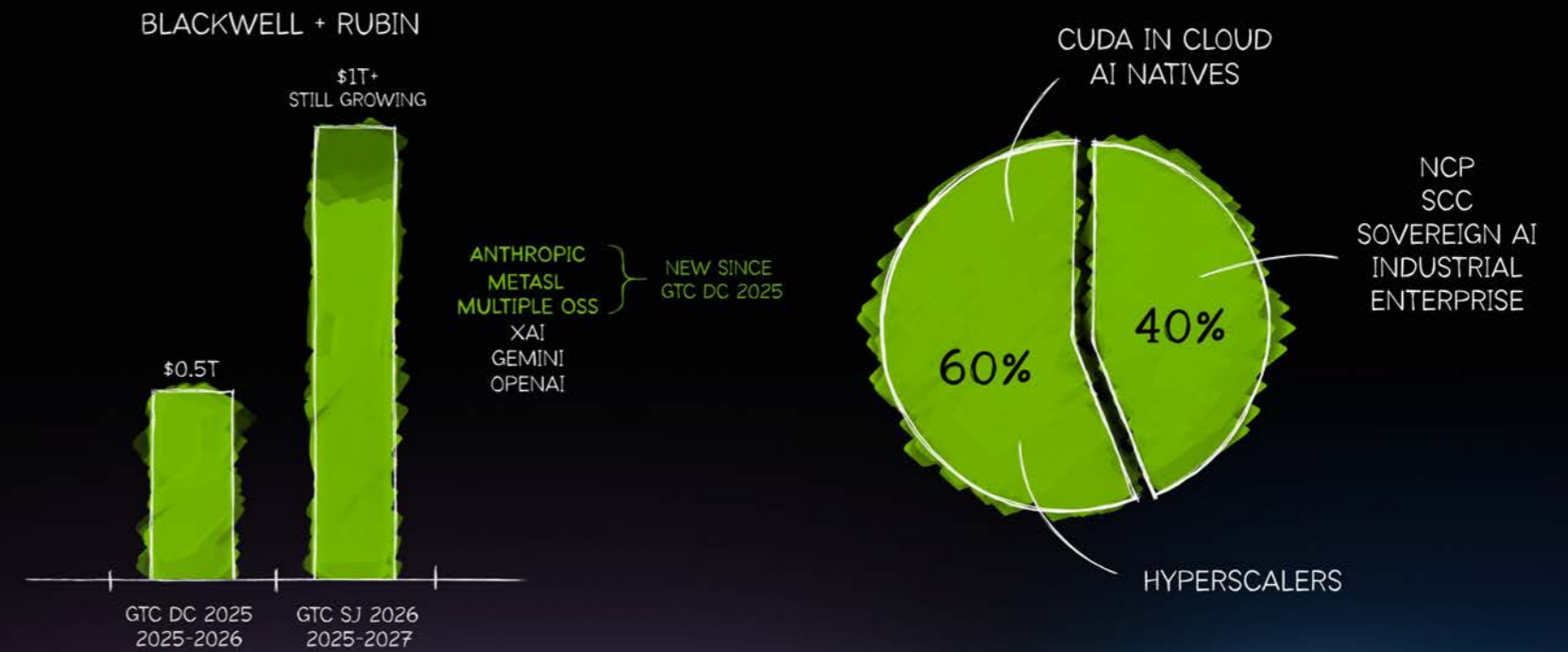
cuDNN, Magnum IO, cuOpt, cuVS, cuDF, Omniverse, TRT-LLM, Dynamo, NeMo, PhysicsNeMo, Cosmos, Nemontron

NVIDIA Hopper, NVIDIA RTX PRO Blackwell Server Edition, NVIDIA Blackwell, Confidential Computing, NVIDIA Rubin, NVIDIA ConnectX and BlueField Networking, NVIDIA Quantum IB, Spectrum-X Networking

### Inference Inflection Arrives 10,000X ChatGPT Compute

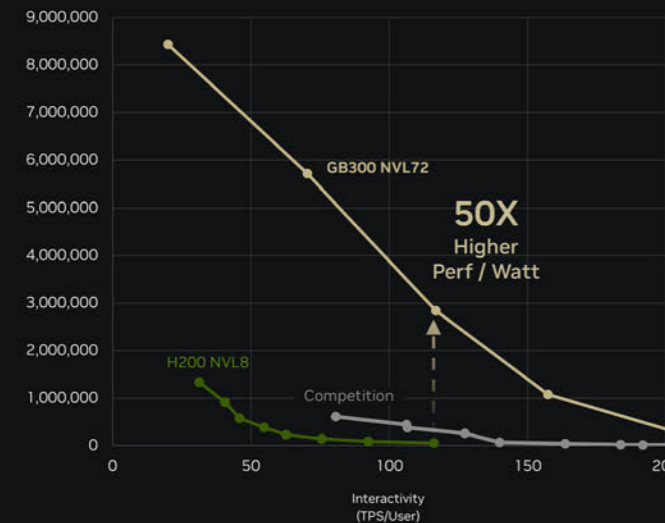


### Inference Inflection Drives Strong Growth NVIDIA Full-Stack Expanding AI to All Regions and Industries

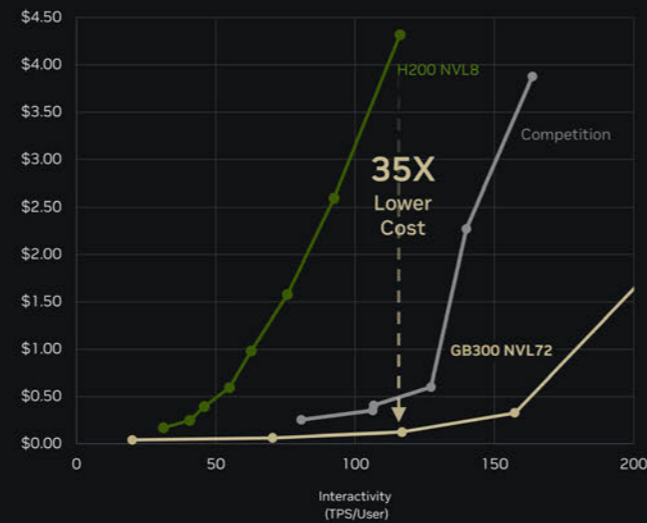


## NVIDIA Extreme Co-Design Revolutionized Token Cost “GB NVL72 Inference King”

### Tokens per Watt Drives Factory Revenue



### Performance Drives Token Cost

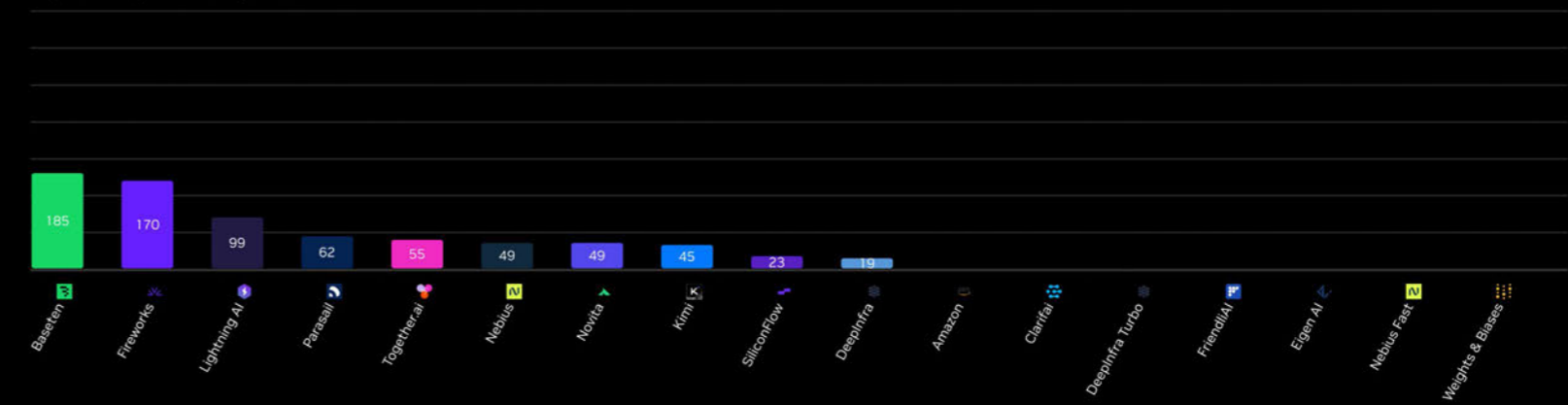


DeepSeek R1 0528 · FP4 · 1K/1K · Source: SemiAnalysis InferenceX

## NVIDIA is the Global Standard for AI Inference at Scale All Leading Inference End-Points Run on NVIDIA

### Output Speed - March 3, 2026

Output Tokens per Second; Higher is better

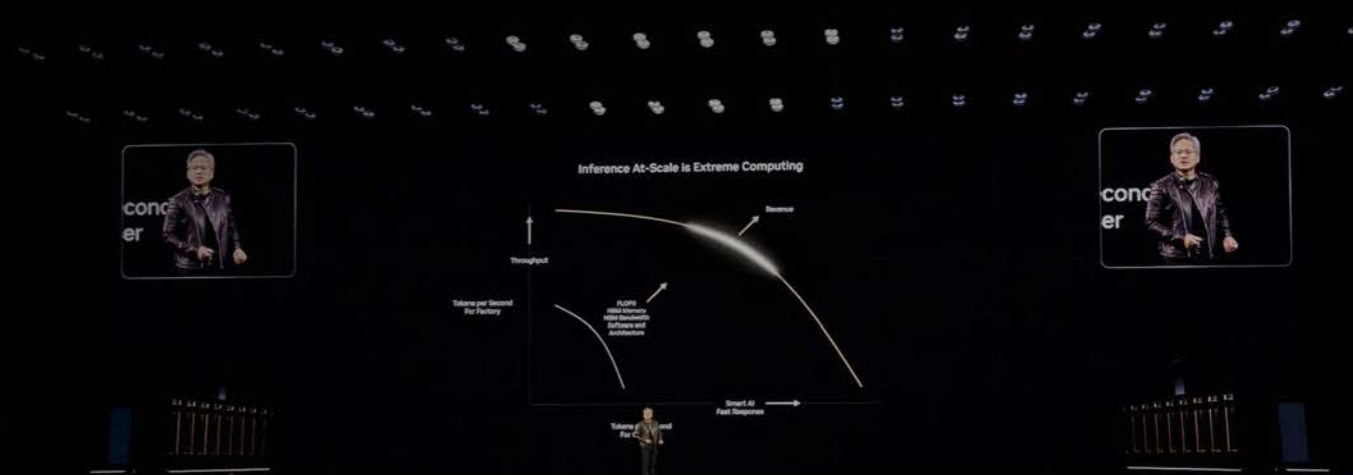


Kimi K2.5 Reasoning

AI Factories are the Industrial Infrastructure of the AI Era  
Inference is the Workload  
Tokens are the New Commodity  
Compute is Revenue



## Inference Performance and Efficiency Drive Company Results



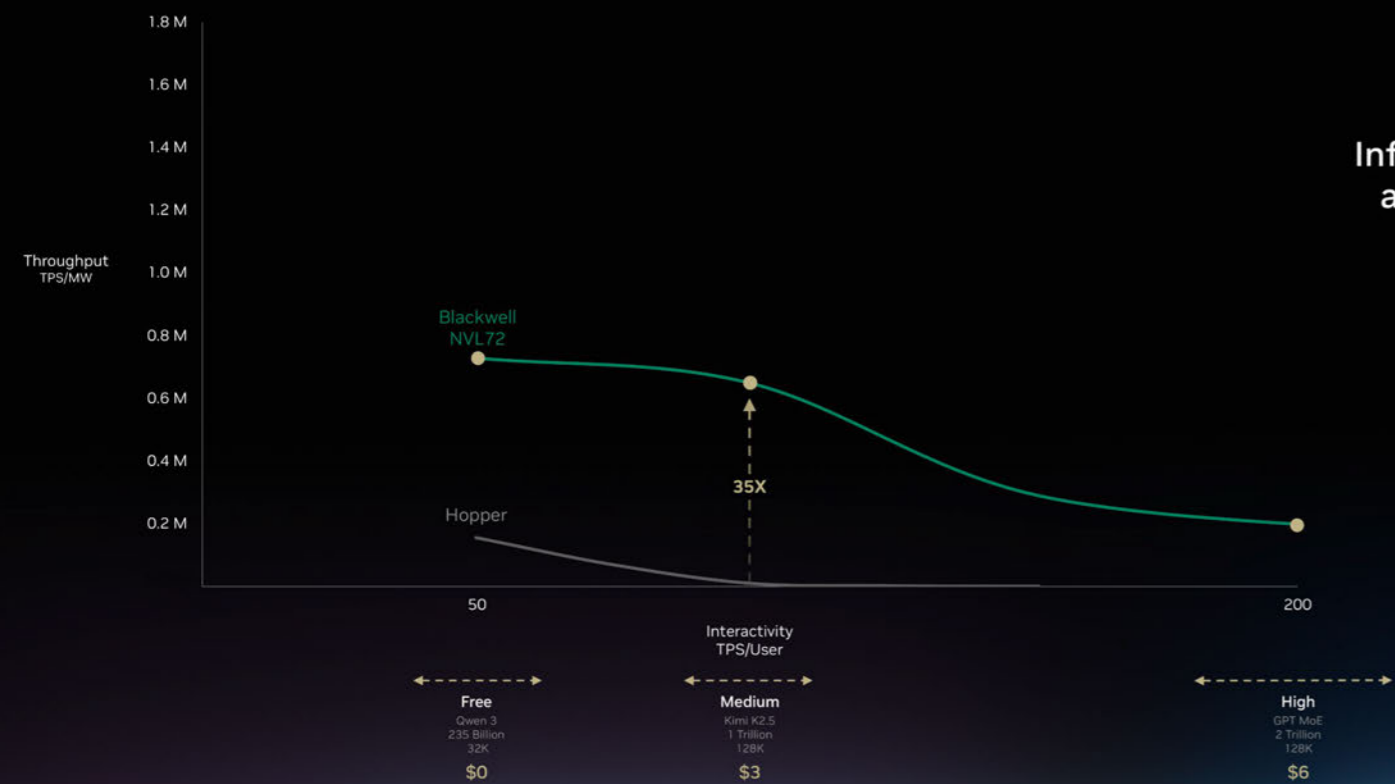
## Inference Performance and Efficiency Drive Company Results



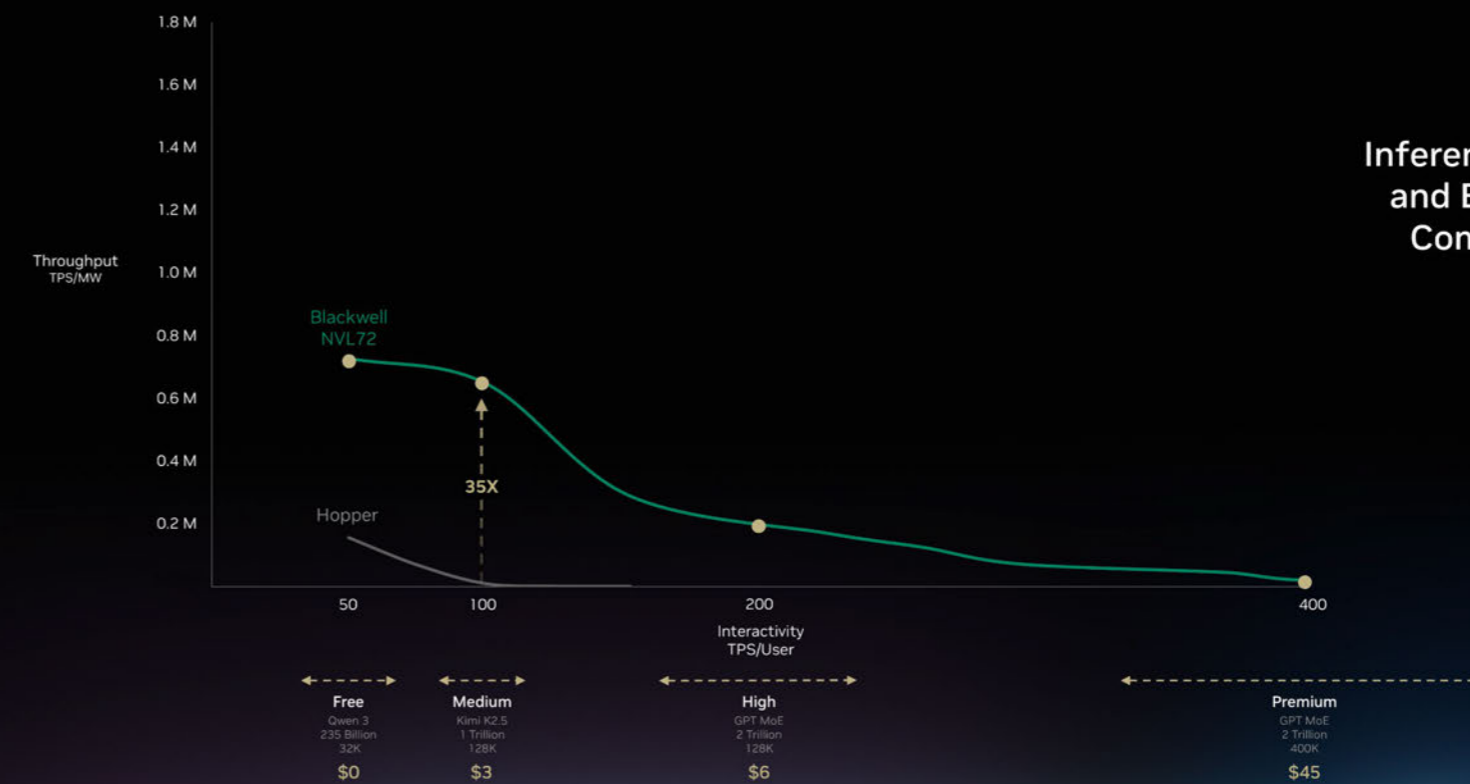
## Inference Performance and Efficiency Drive Company Results

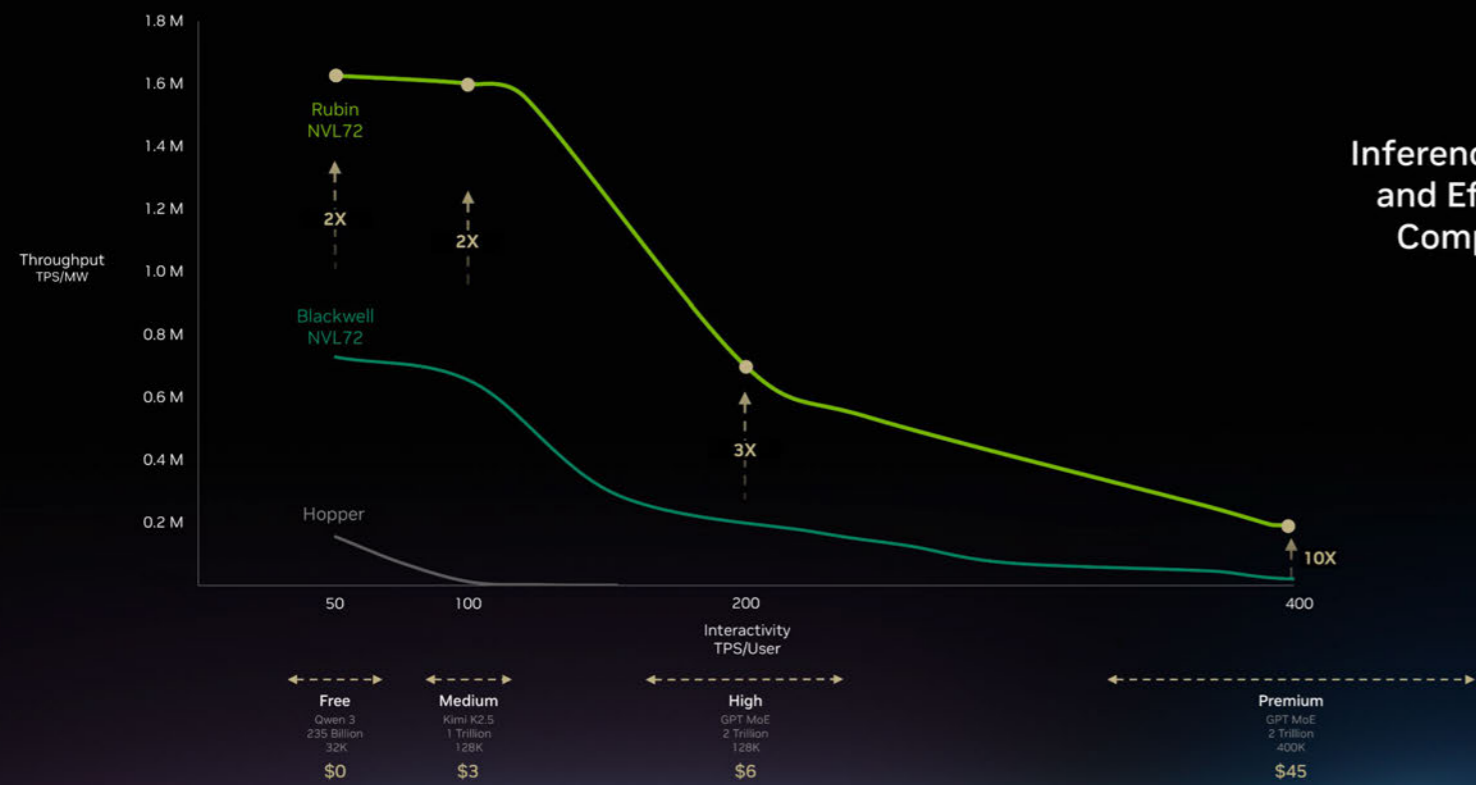


## Inference Performance and Efficiency Drive Company Results



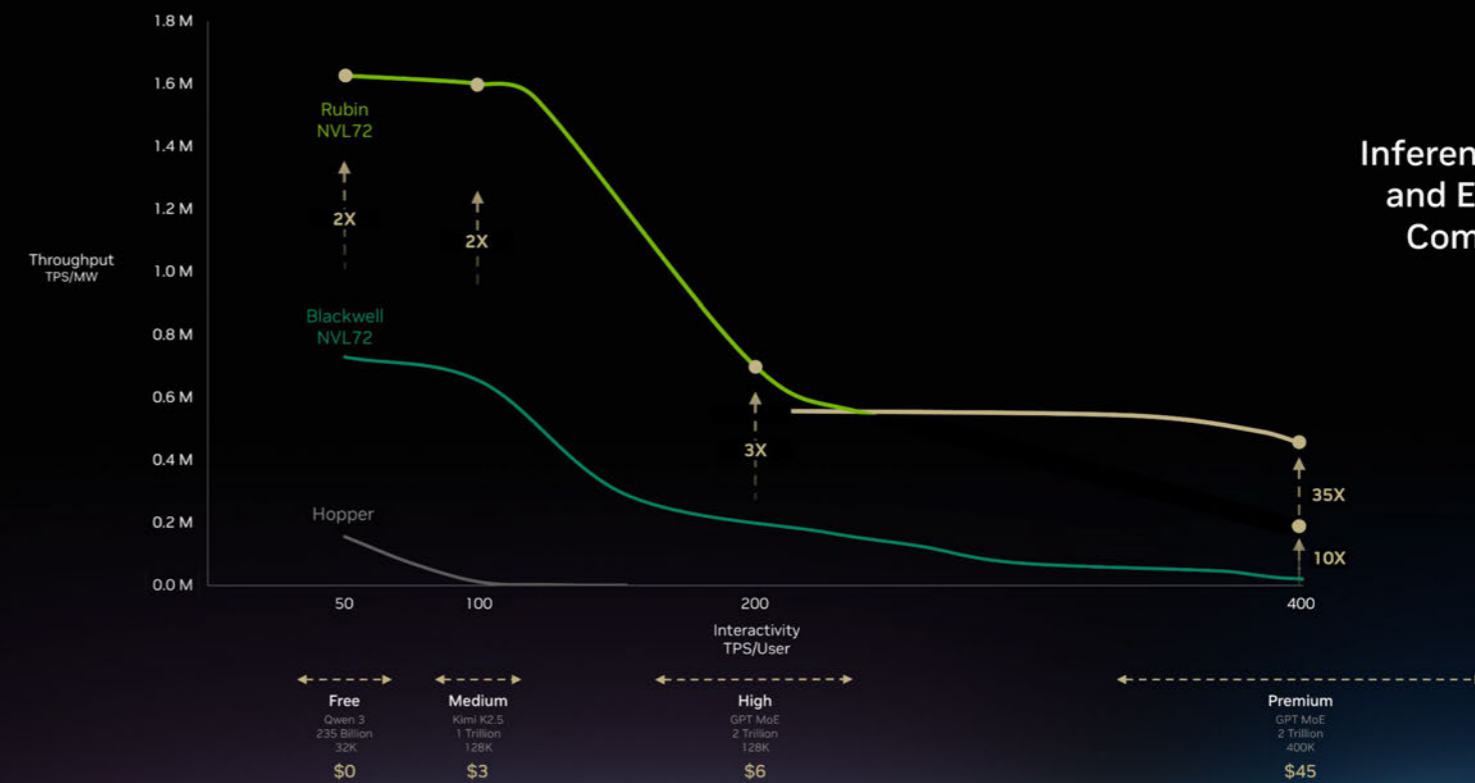
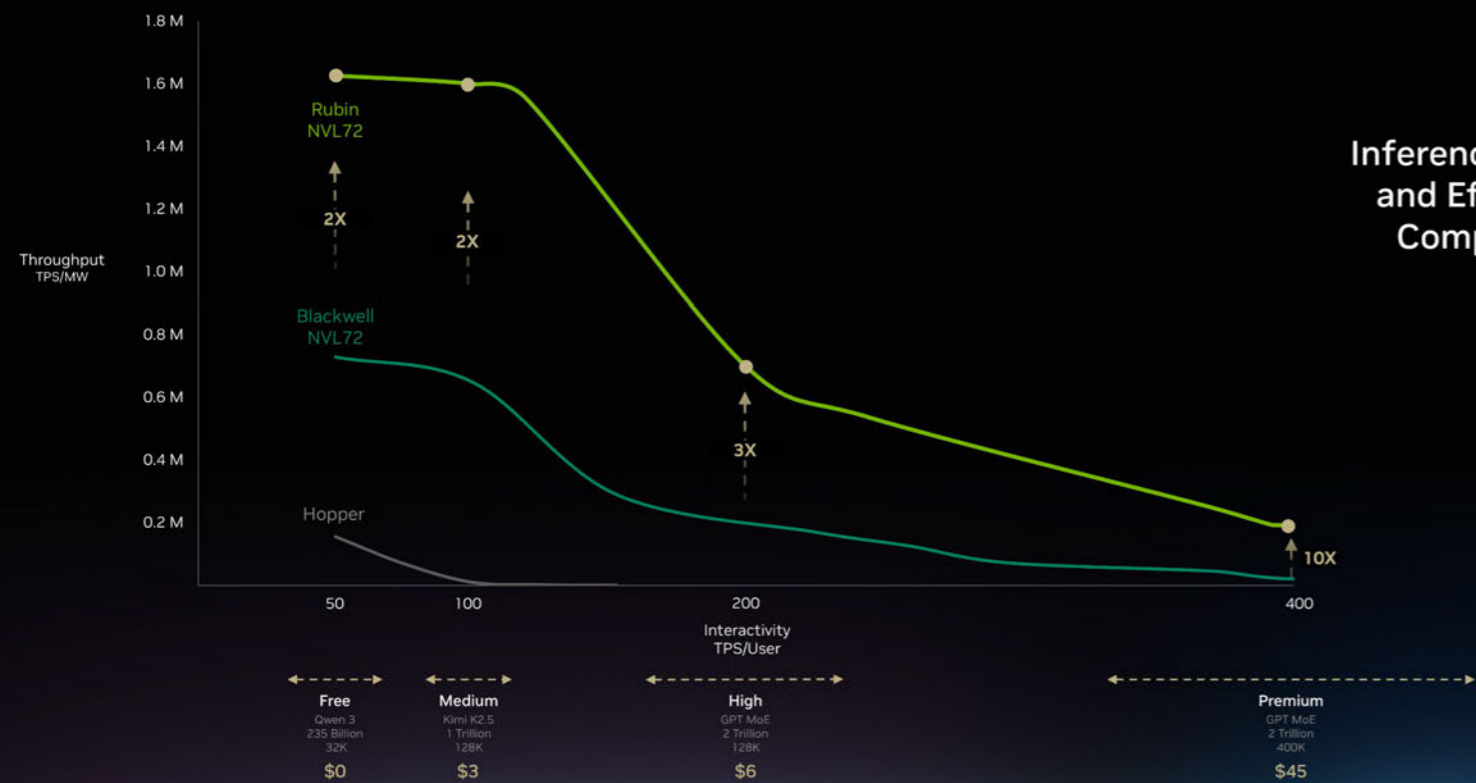
## Inference Performance and Efficiency Drive Company Results

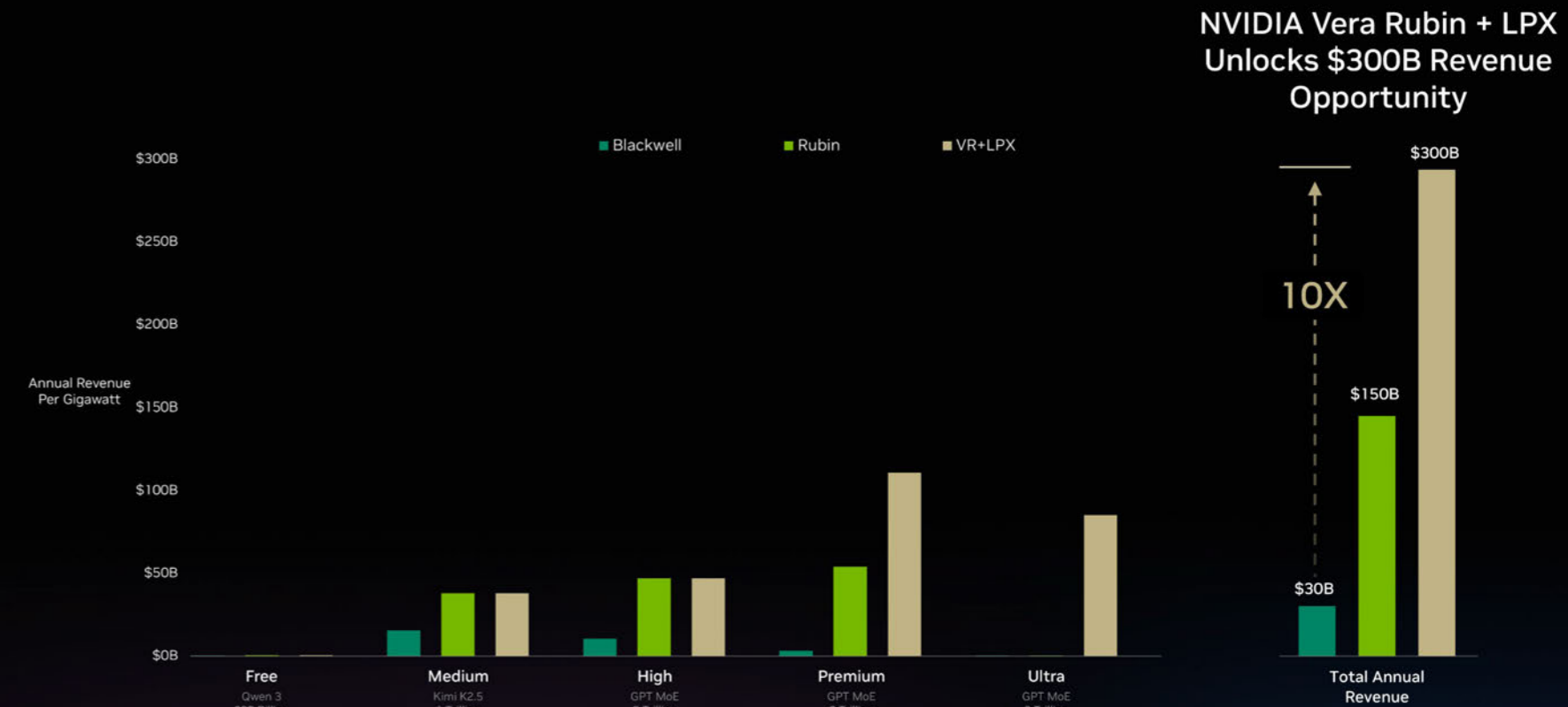
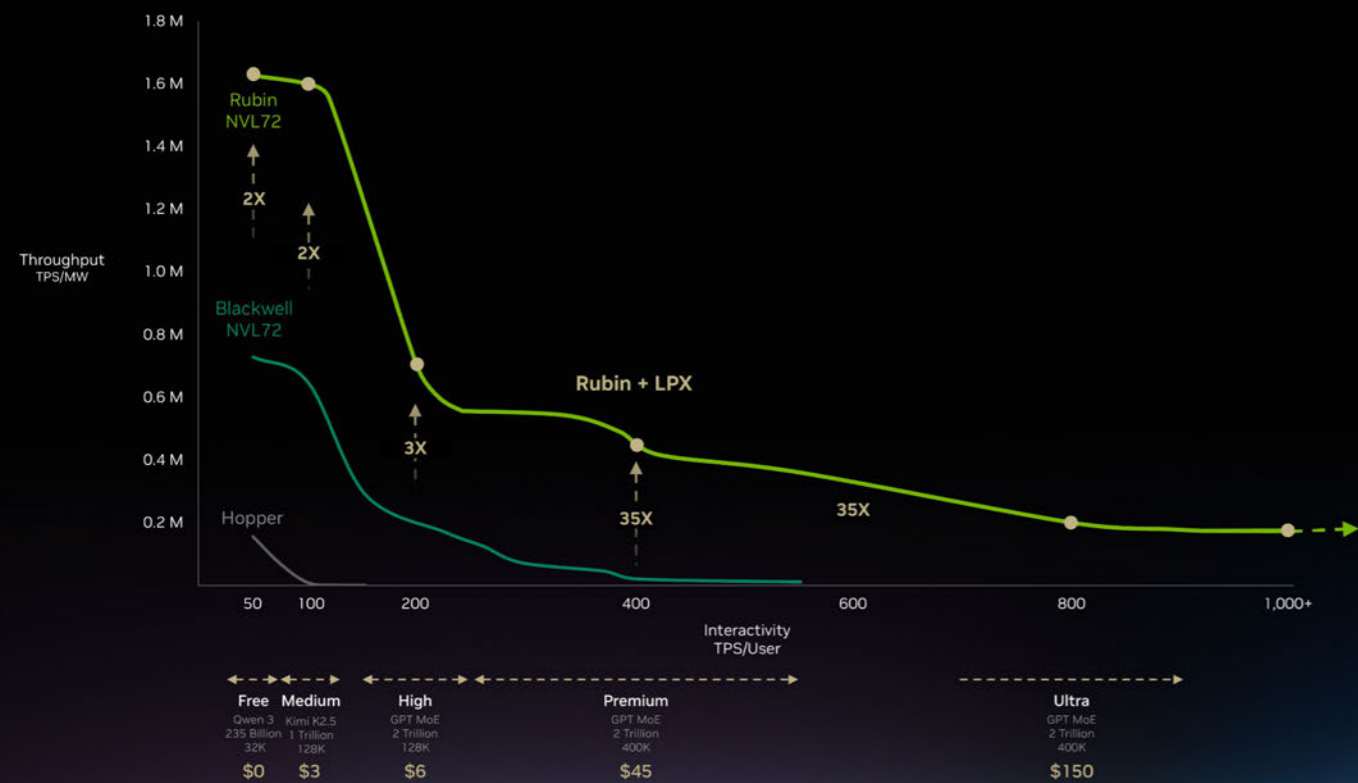




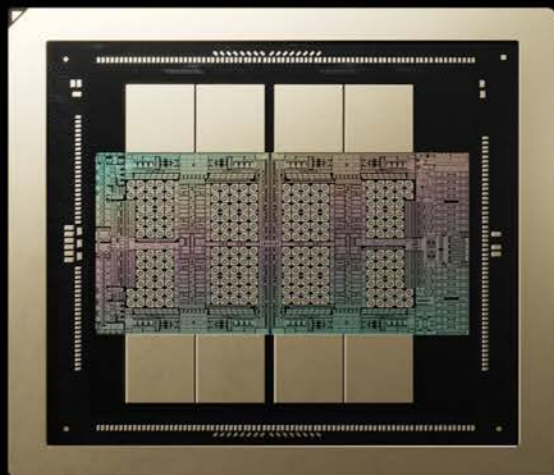
## NVIDIA Vera Rubin Unlocks \$150B Revenue Opportunity



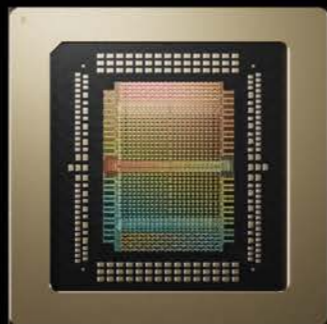




Rubin GPU



Groq 3 LPU

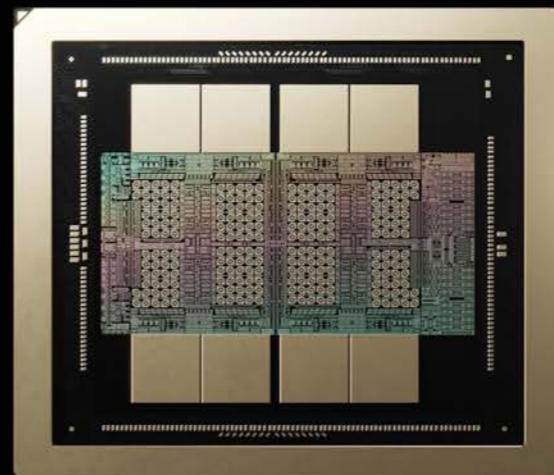


288 GB HBM4  
22 TB/s  
50 PFLOPs (NVFP4)  
336B Transistors  
+ 2.5T (HBM4)

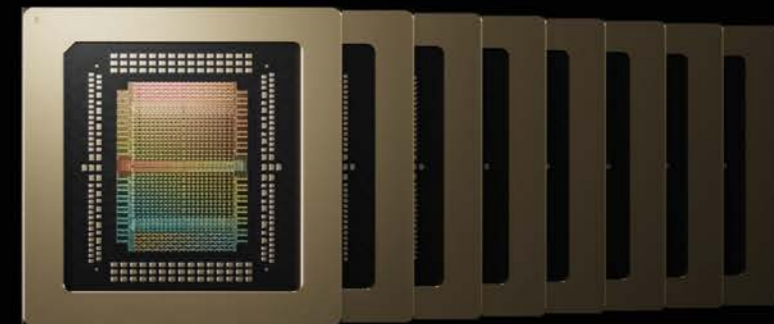
500 MB SRAM  
150 TB/s SRAM Bandwidth  
1.2 PFLOPs (FP8)  
98B Transistors

Uniting Processors of Extreme Performances

Rubin GPU



Groq 3 LPU

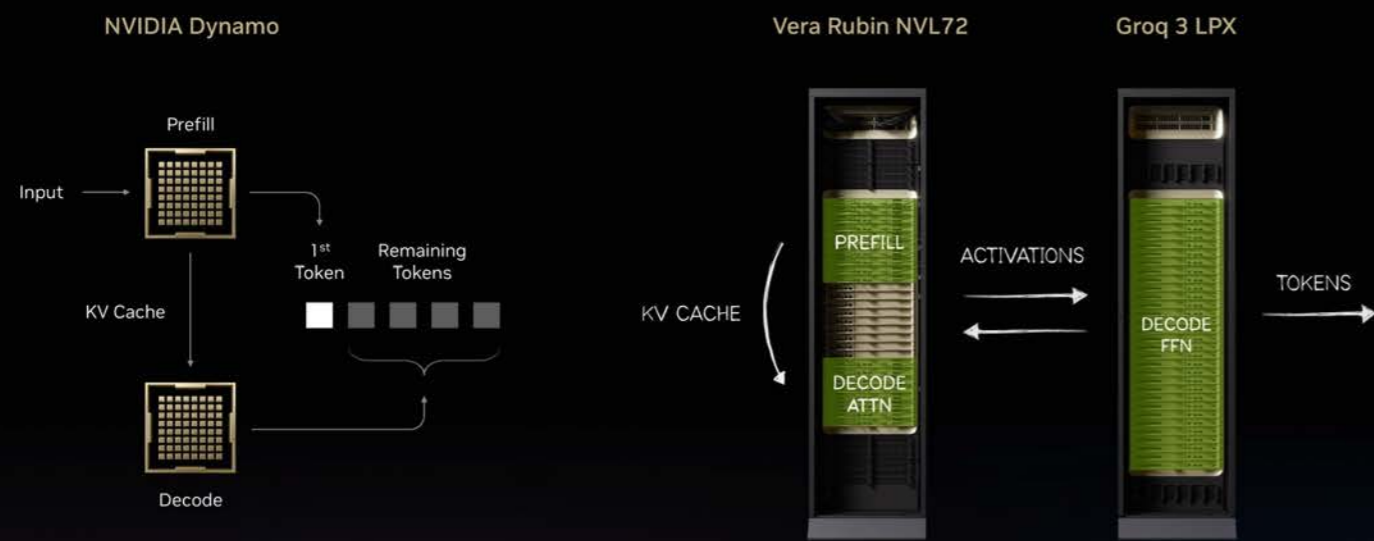


288 GB HBM4  
22 TB/s  
50 PFLOPs (NVFP4)  
336B Transistors  
+ 2.5T (HBM4)

4 GB SRAM  
1,200 TB/s SRAM Bandwidth  
9.6 PFLOPs (FP8)  
784B Transistors

55X

Uniting Processors of Extreme Performances



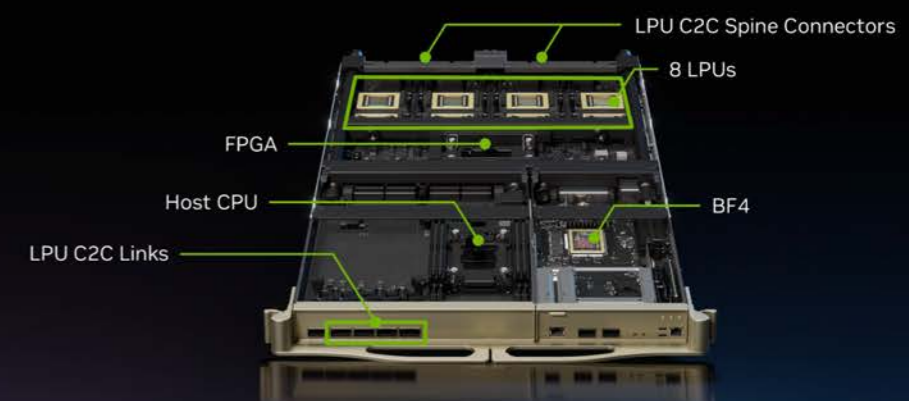
Uniting Processors of Extreme FLOPS and Bandwidth



Available 2H26

### NVIDIA Groq 3 LPX

AI Inference Compute	315 PFLOPS
SRAM Capacity	128 GB
Memory Bandwidth	40 PB/s
Scale-Up Density	256 Chips
Scale-Up Bandwidth	640 TB/s





## Announcing NVIDIA Vera Rubin NVL72 Launch Partners

### AI Labs



### Cloud



### OEMs and ODMs

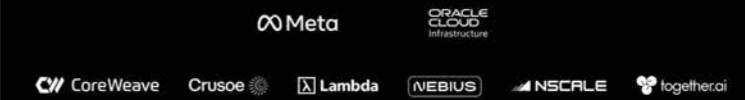


10X Perf/W | 3.6 EF NVFP4 | 1.6 PB/s HBM4 | 260 TB/s NVLink6



## Announcing NVIDIA Vera CPU Launch Partners

### Cloud



### OEMs and ODMs



256 Vera CPUs | 300 TB/s LPDDR5X | ETL Spine | 6.5X Throughput



## Announcing NVIDIA BlueField-4 STX Launch Partners

Cloud

CoreWeave, Crusoe, Lambda, Mistral AI, NEBIUS, ORACLE CLOUD Infrastructure, VULTR

OEMs

AIC, CLOUDIAN, ddn, DELL Technologies, Everpure, HPE, Hitachi Vantara, IBM, MINIO, NetApp, NUTANIX, QCT, SUPRAMEX, VAST, WEKA

5x Tokens/sec | 50 Tb/s Networking BW | 16TB Shared Context / GPU

## NVIDIA Vera Rubin 7 Chips – 5 Rack Systems AI Factory for the Agentic AI Frontier

1 GW AI Factory	X86 + Hopper	Vera Rubin
# of GPUs	600K	300K
AI FLOPS	1.2 ZFLOPS	16 ZFLOPS
All-to-All Scale-up	7.2 TB/s	260 TB/s
Memory BW-per-Domain (GROQ SRAM)	2 EB/s	100 EB/s
Tokens per Second	2M	700M

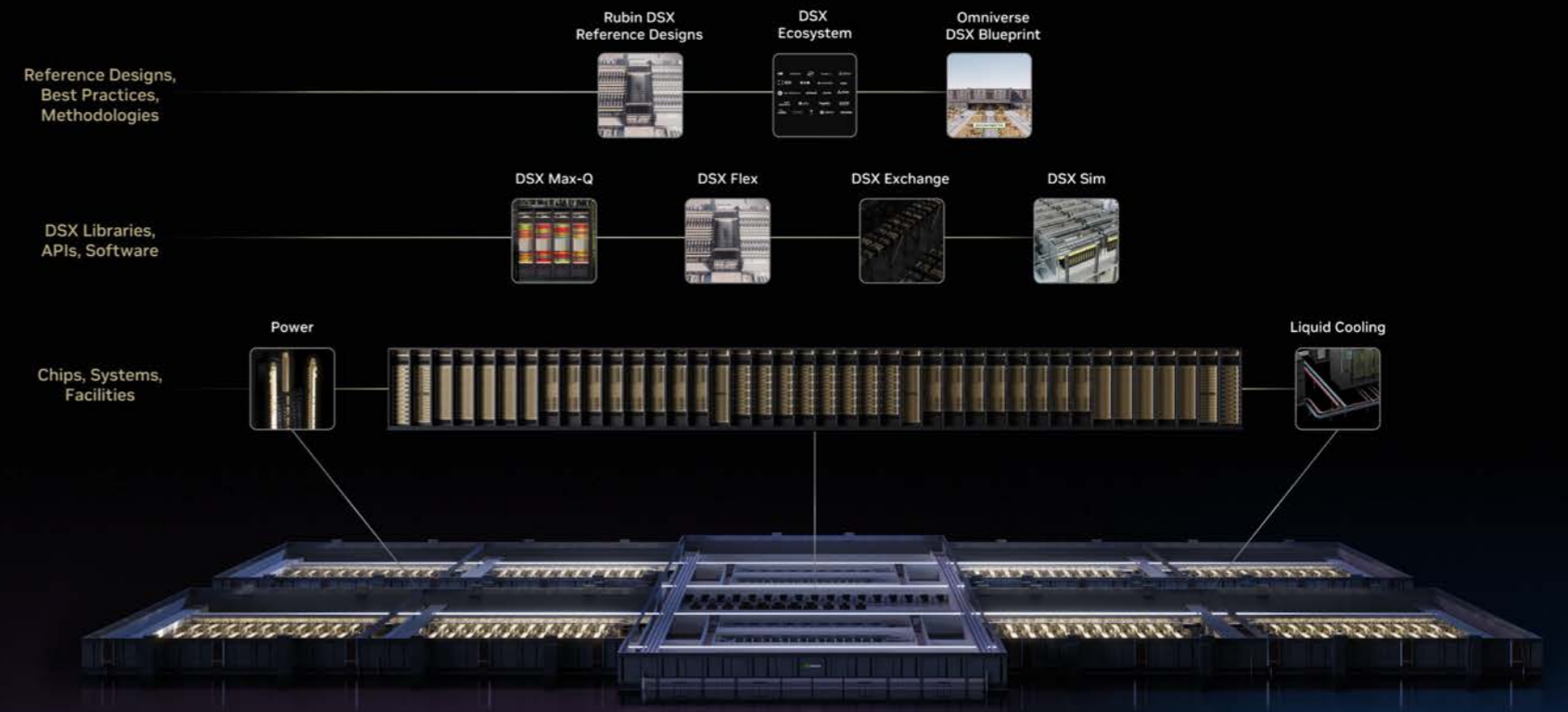


## NVIDIA Extreme Co-Design Delivering X-Factors Every Year From Chips to Racks to AI Factories



## NVIDIA DSX AI Factory Platform

Extreme Co-Design at Infrastructure Scale



# NVIDIA DSX AI Factory Platform

Accelerates Scalable, Energy-Efficient AI Factory Deployment

## DSX Flex

CoreWeave, Crusoe, Digital Realty, emeraldai, NEBIUS, NSCLC, switch

## DSX Max-Q

CoreWeave, Crusoe, Digital Realty, emeraldai, NEBIUS, NSCLC, Red Hat, Fujitsu, phaidra, switch

## DSX Sim

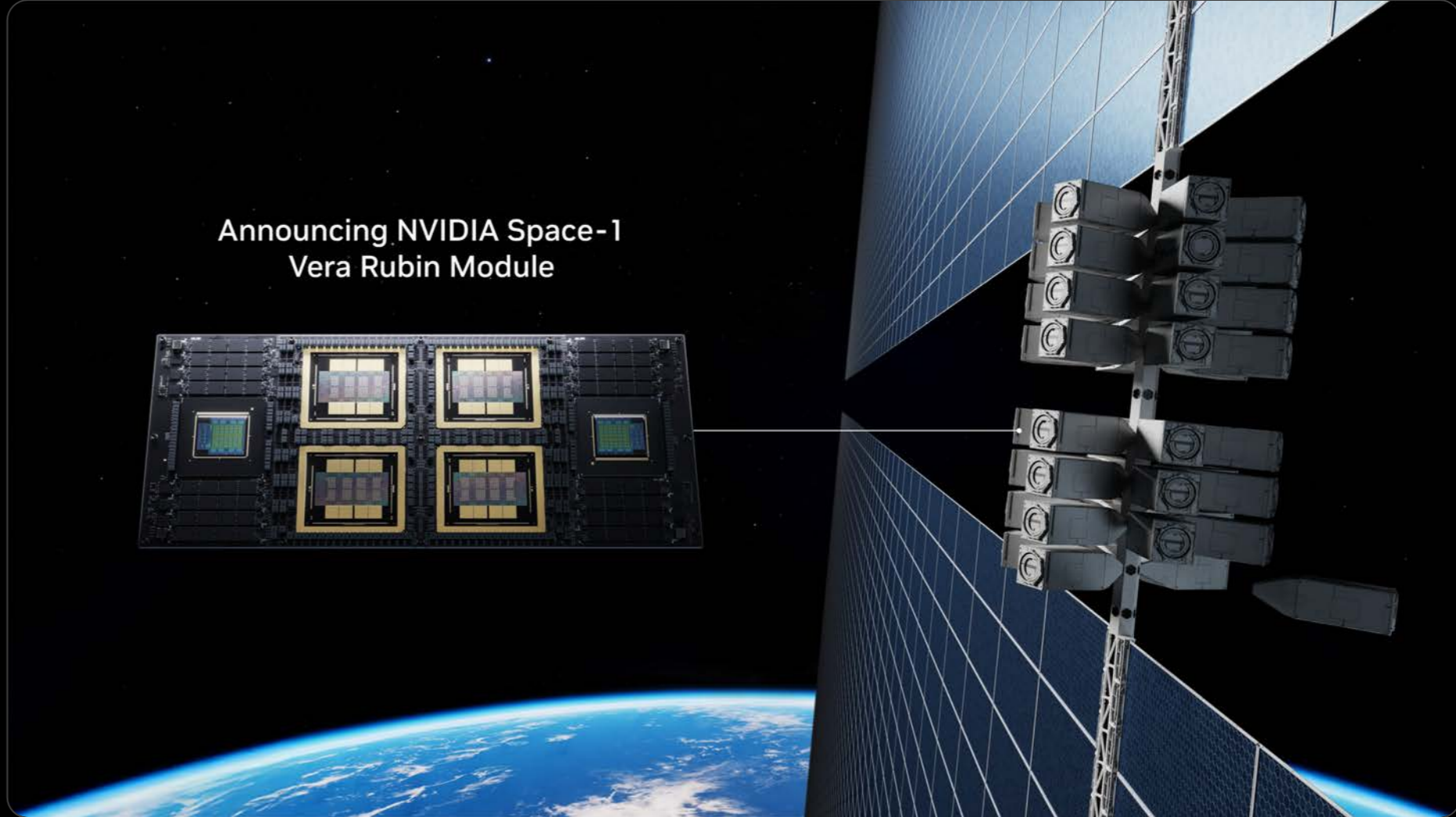
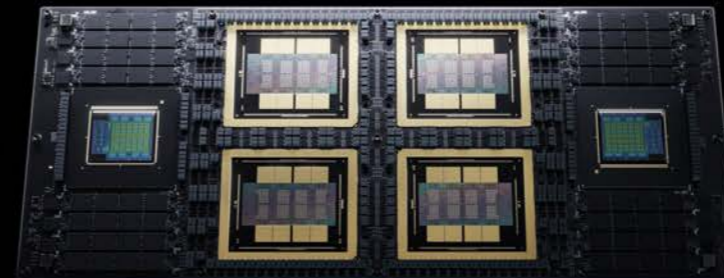
ABB, ARMADA, AVEVA, Cadence, Check Point, Cloudian, Dassault Systèmes, Dell Technologies, F.T.N, etap, GE Vernova, hedgehog, HITACHI, Jacobs, Keysight, metalsoft, Mirantis, Mitsubishi Electric, NetScout Systems, NOKIA, Open Nebula, Palo Alto Networks, PTC, RAFAY, Red Hat, Fujitsu, Schneider Electric, SIEMENS, Siemens Energy, Slurm, Spectro Cloud, TRANE, TrendAI, VAST, vCluster, VERTIV, WEKA

## DSX Exchange

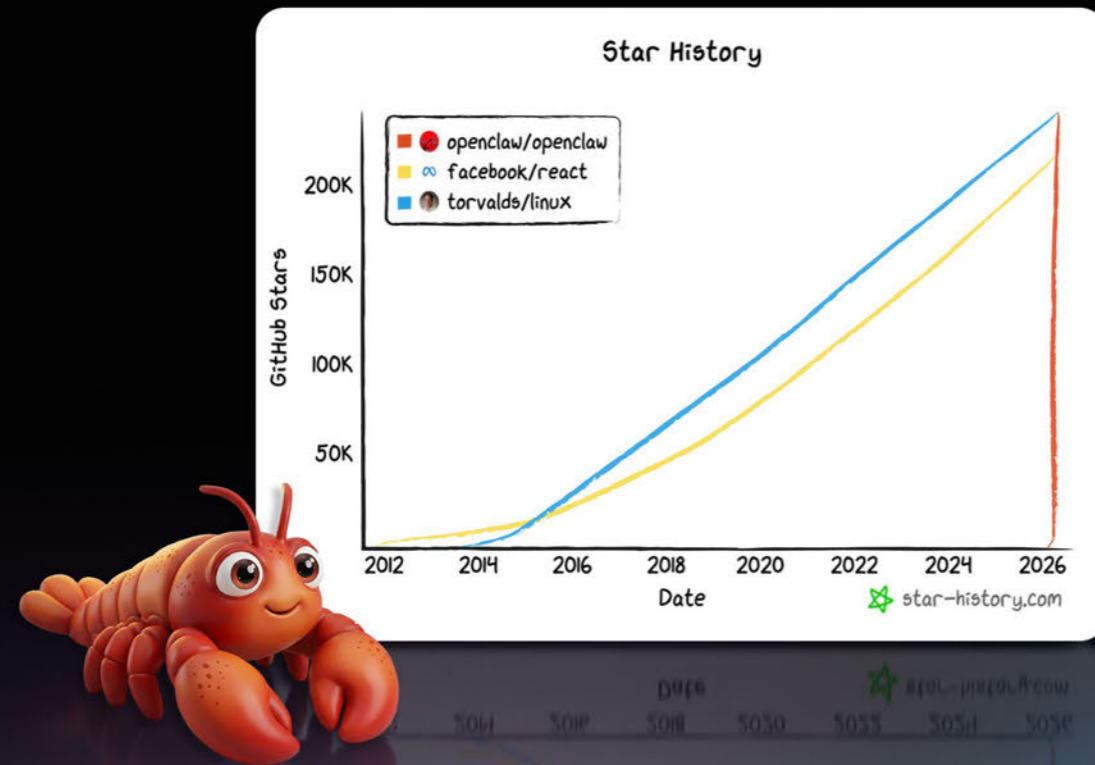
ABB, ARMADA, AVEVA, Cadence, Check Point, Cloudian, CoreWeave, Crusoe, Digital Realty, Dell Technologies, F.T.N, emeraldai, GE Vernova, hedgehog, HITACHI, Keysight, metalsoft, Mirantis, Mitsubishi Electric, NEBIUS, NetScout Systems, NOKIA, NSCLC, Open Nebula, Palo Alto Networks, phaidra, RAFAY, Red Hat, Fujitsu, Schneider Electric, SIEMENS, Siemens Energy, Slurm, Spectro Cloud, switch, TRANE, TrendAI, VAST, vCluster, VERTIV, WEKA



# Announcing NVIDIA Space-1 Vera Rubin Module



## Inference Inflection Arrives

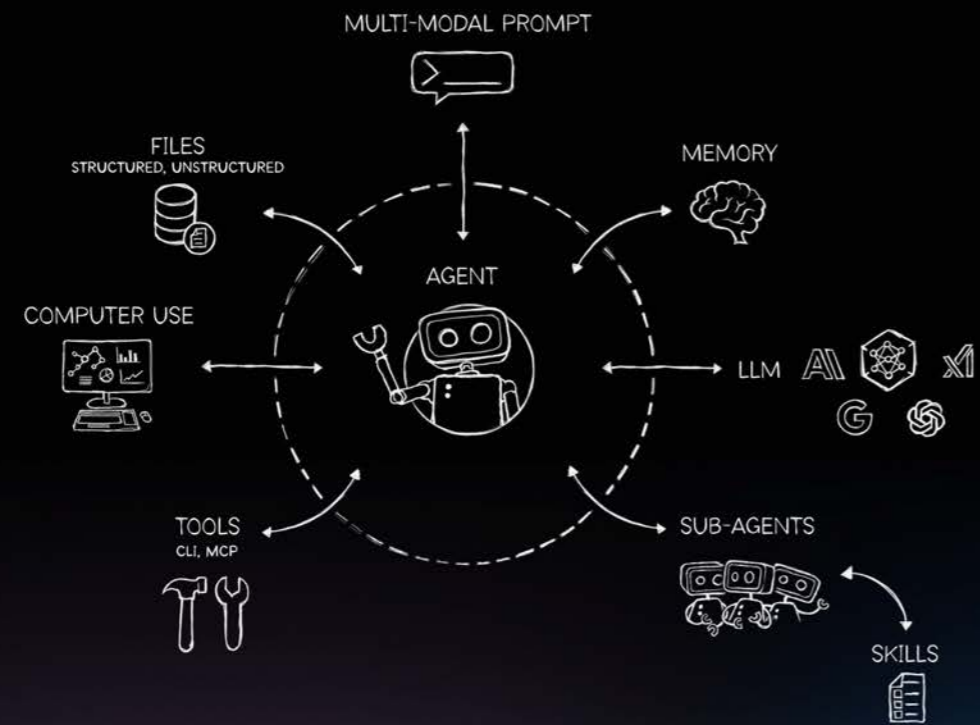


## Announcing NVIDIA NemoClaw for OpenCLaw

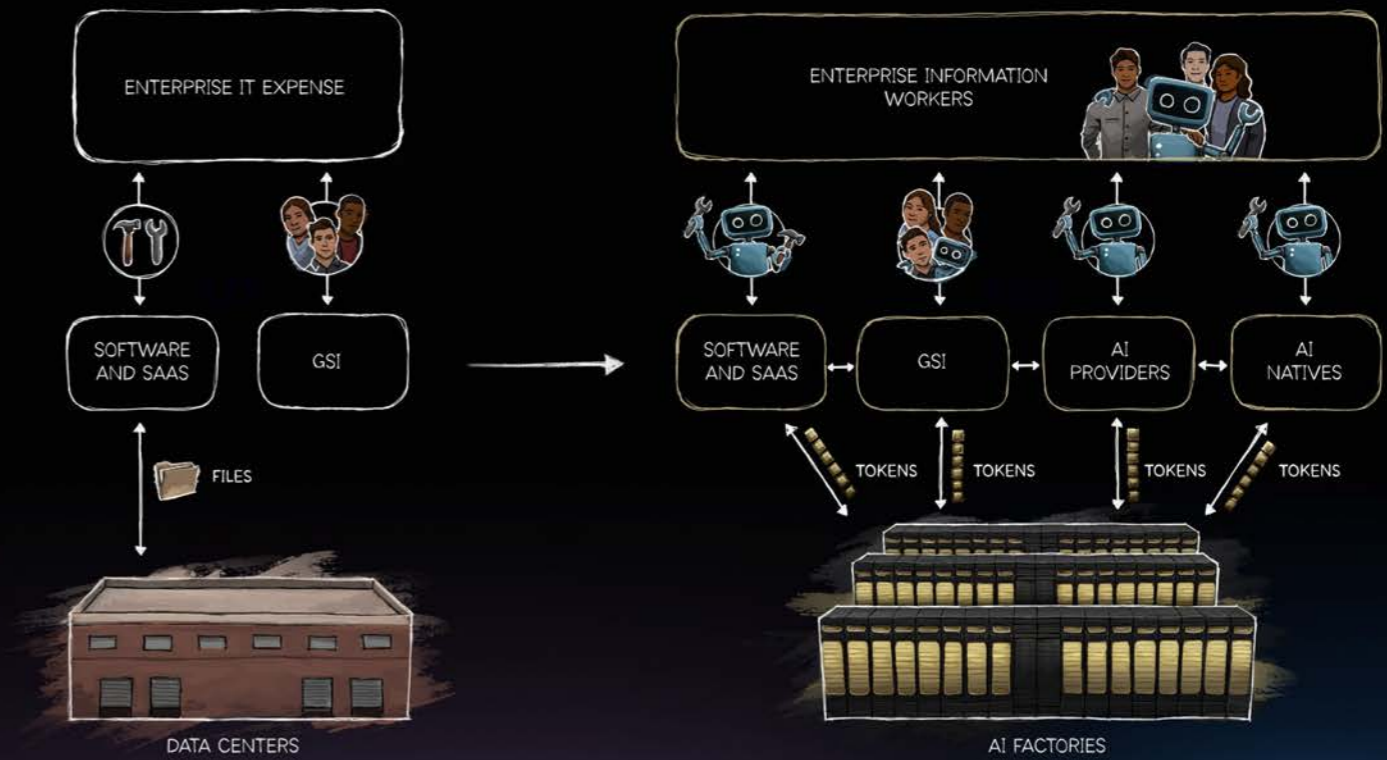
```
ubuntu@brev-ae6ah8vzm: ~  
→ ~ curl -fsSL https://nvidia.com/nemoclaws.sh | bash  
→ ~ nemoclaws onboard
```



## Agents – A New Computing Platform

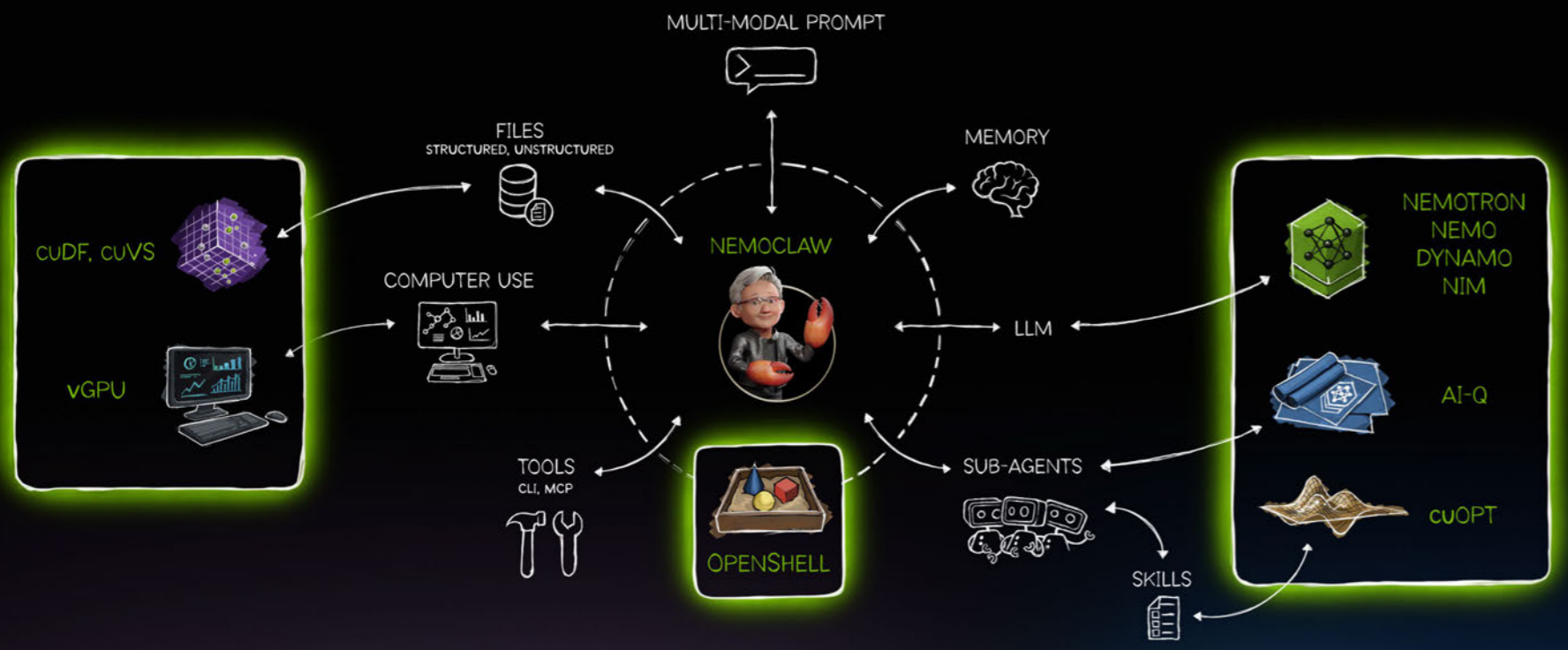


## Enterprise IT Renaissance from SaaS to Agent-as-a-Service

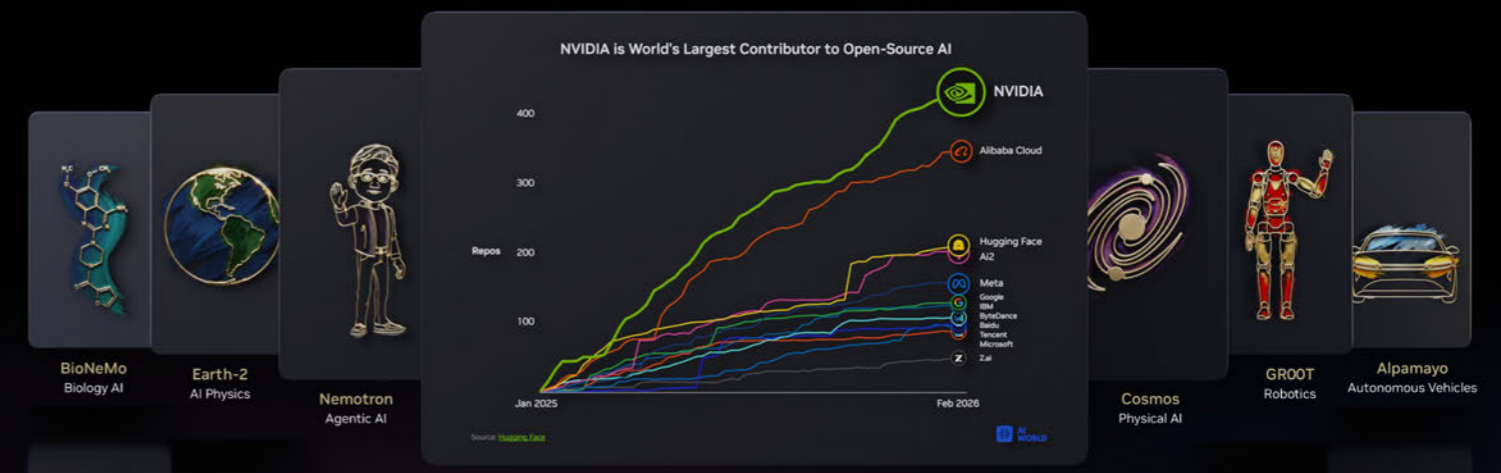


## Announcing NVIDIA NemoClaw Reference OpenClaw

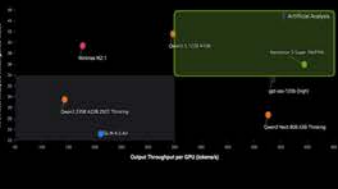
NVIDIA Agent Toolkit for Building Specialized Agents



## NVIDIA Leaderboard Topping Open Models



### AA Intelligence vs Efficiency Nemotron 3 Super



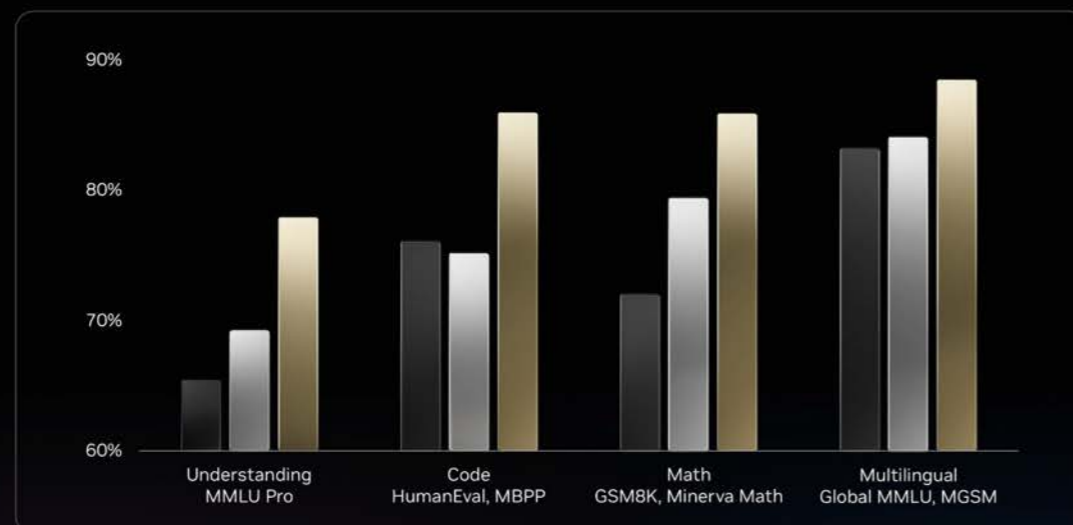
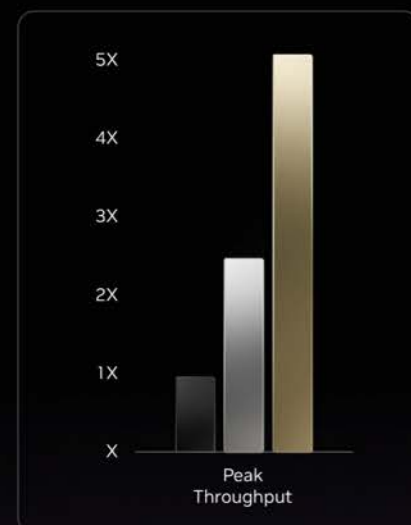
### PinchBench Nemotron 3 Super



## NVIDIA Nemotron 3 Ultra – Best Open Base Model

5X Efficiency and Highest Reasoning Accuracy on NVIDIA GB200 NVL72

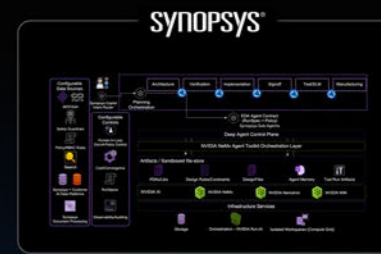
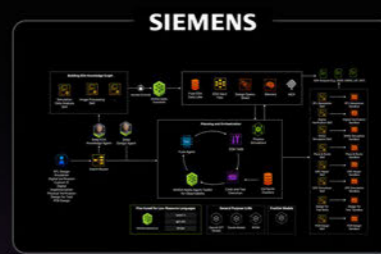
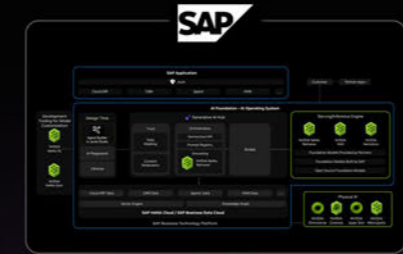
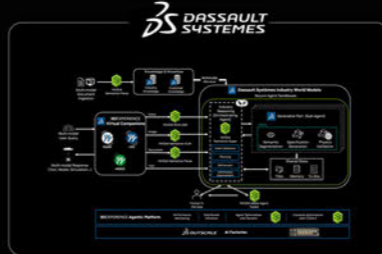
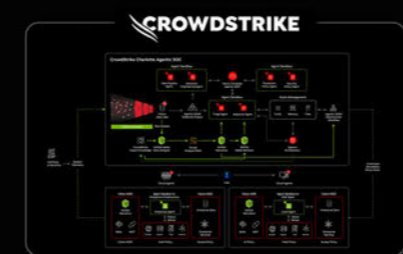
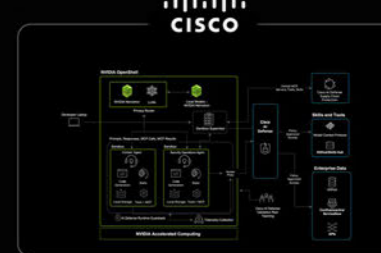
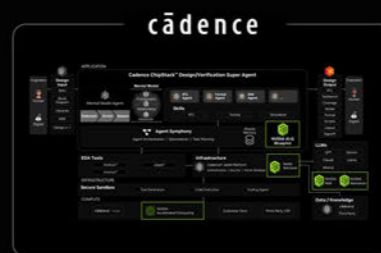
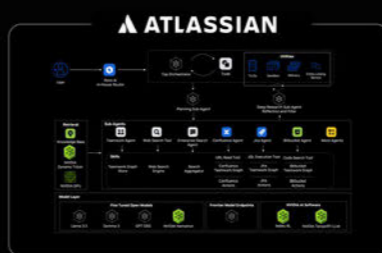
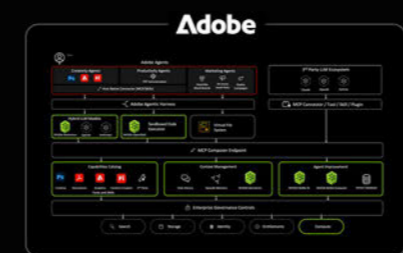
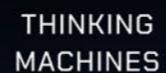
■ GLM   ■ Kimi K2   ■ Nemotron 3 Ultra



## The World Building Regional AI With NVIDIA Nemotron



# Announcing Global AI Leaders Join NVIDIA Nemotron Coalition to Advance Open Frontier Models



- Alibaba, AUTOMATIC ARTISAN, AWS, baseten, ByteDance, CLOUDIAN, crewal, Cognition, CoreWeave, DELL Technologies, ddn, Everpure, FACTORY, Google Cloud
- GCORE, glean, Honeywell, harbor, Harmonic, HPE, Hugging Face, IBM, LangChain, LMCACHE, llm-d, Meituan, NEBIUS, NetApp, Ollama, ORACLE
- OpenEnv, openCode, OpenRouter, OpenHands, PRIME Intellect, PyTorch, SoftBank, Tencent Cloud, togetherai, unsloth, VAST, VULTR, WEKA

**ABRIDGE**

**CodeRabbit**

**COHESITY**

**Edison**

**IQVIA**

**mastercard**

**Microsoft Security**  
NEXT AI

**PayPal**

**perplexity**  
perplexity enterprise

**小红书**

**Rockwell Automation**

**Schneider Electric**

Alibaba, Amazon, AWS, baseten, InByteDance, CLOUDIAN, crewai, Cognition, CoreWeave, DELLTechnologies, ddn, Everpure, FACTORY, Google Cloud, GCORE, glean, Honeywell, harbor, Harmonic, HPE, Hugging Face, IBM, LangChain, LMCache, llm-d, Meituan, NEBUS, NetApp, Olama, OpenAI, OpenAI, OpenRouter, OpenHands, PAIME Intellect, PyTorch, SoftBank, Tencent Cloud, TogetherAI, unsloth, VAST, VULTR, WEKA

