NVIDIA GTC

2026 Highlights


SAP center


NVIDIA
GTC


GTC26


NVIDIA
GTC
It All
Starts Here
Proud Sponsor
ORACLE


NVIDIA
Serve Robotics


nvidia.com/gtc
NVIDIA
GTC
Brilliant Minds,
Breakthrough
Discoveries
San Jose Convention Center
March 16-19, 2026


NVIDIA
GTC
South Registration


Santa Clara St
NVIDIA
GTC


cover every single layer of the five layer pick of artificial intelligence
from land power and shell the infrastructure.


NVIDIA
GTC


Keynote
Entrance


NVIDIA
GTC
It All
Starts Here
nvidia.com/gtc
SAP center


NVIDIA
GTC
LIVE


NVIDIA
GTC

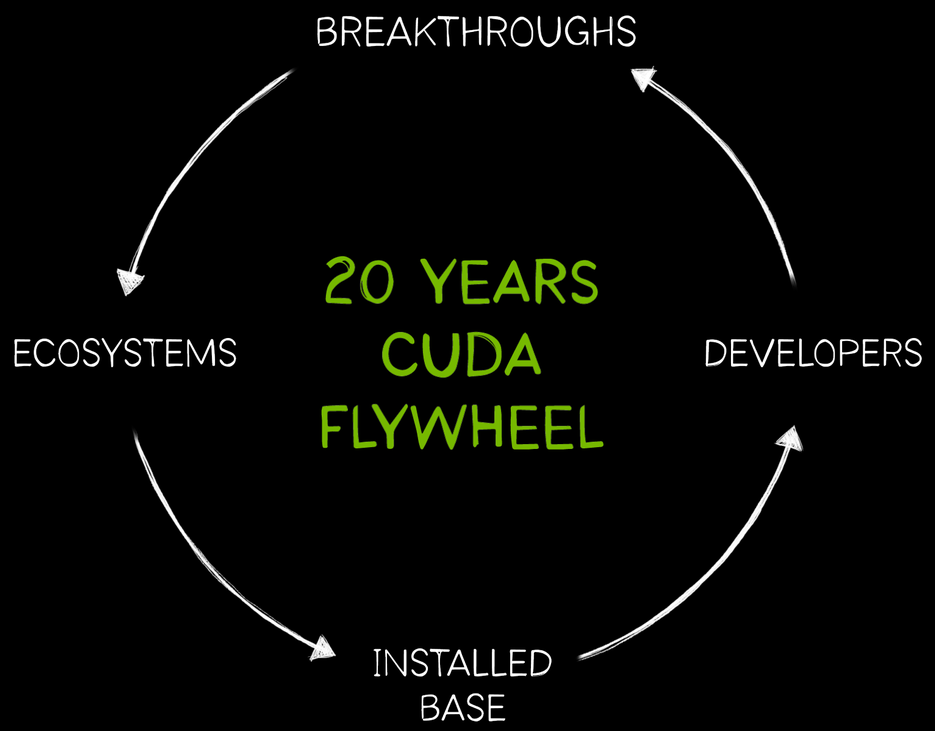

SOUTH TICKET OF


NVIDIA
GTC
LIVE


KO / K5
Invited Guests
Reserved

“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.

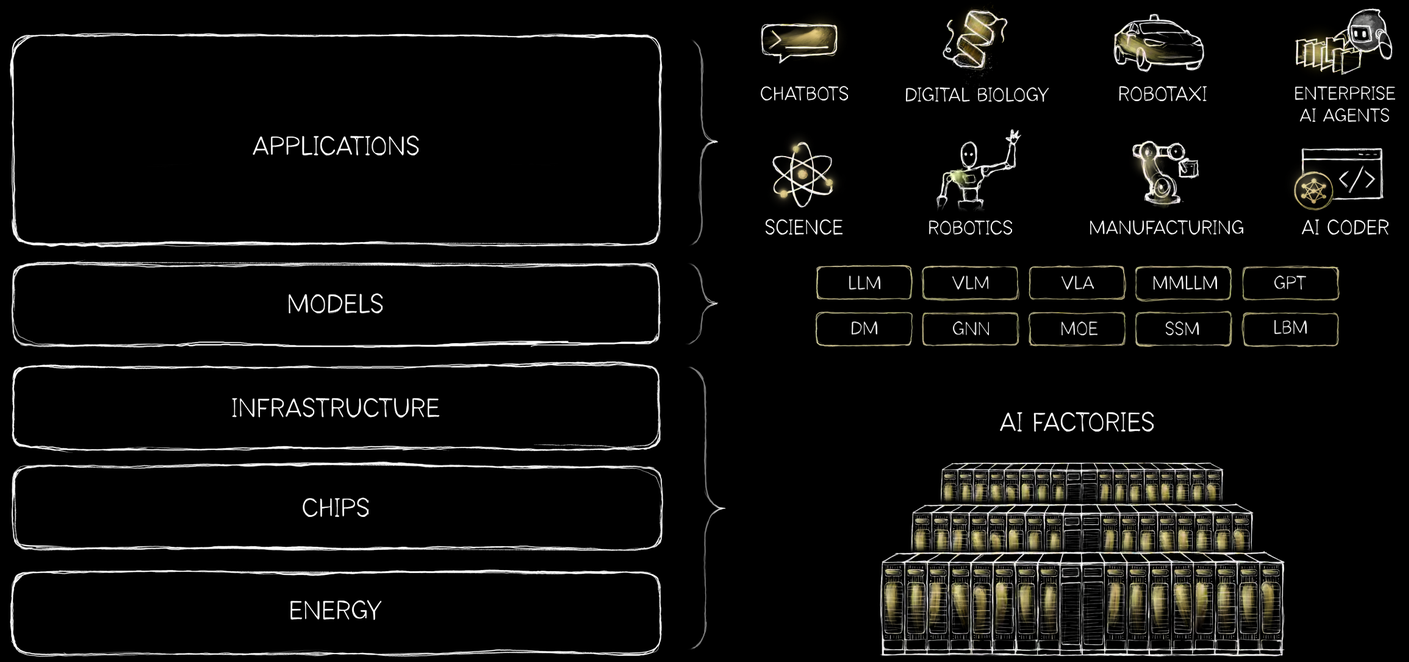
CUDA-X turns algorithms into infrastructure across AI, data, and science, compounding performance at scale.

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

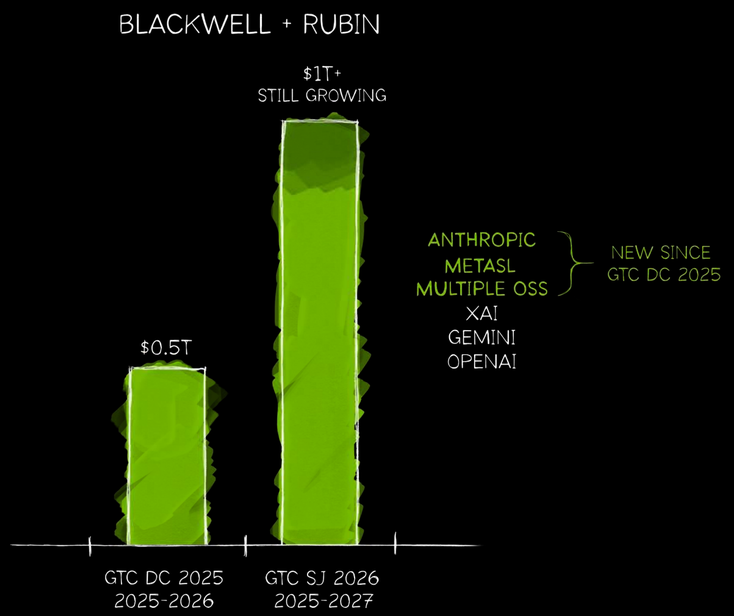
AI does not scale in isolation. It scales as a system. Energy powers chips, chips power infrastructure, infrastructure runs models, and models enable applications. Each layer compounds the next. If one stalls, the system stalls. AI factories demand all five layers advance together.

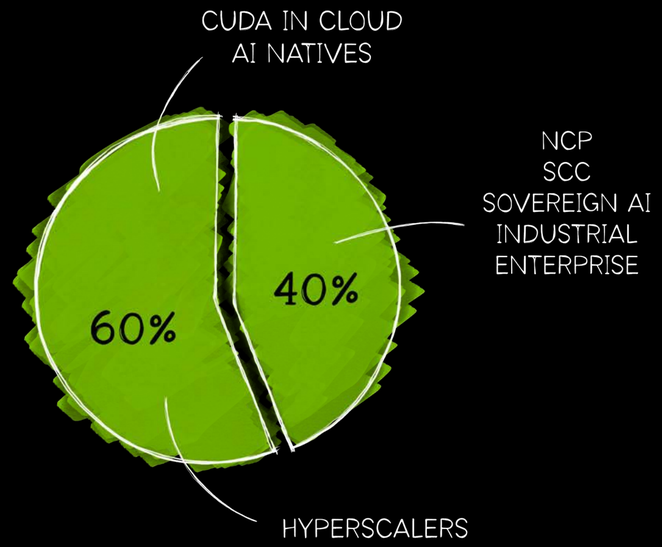


"Every company in the world today needs to have an agentic system strategy."

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.





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

The Wall Street Journal

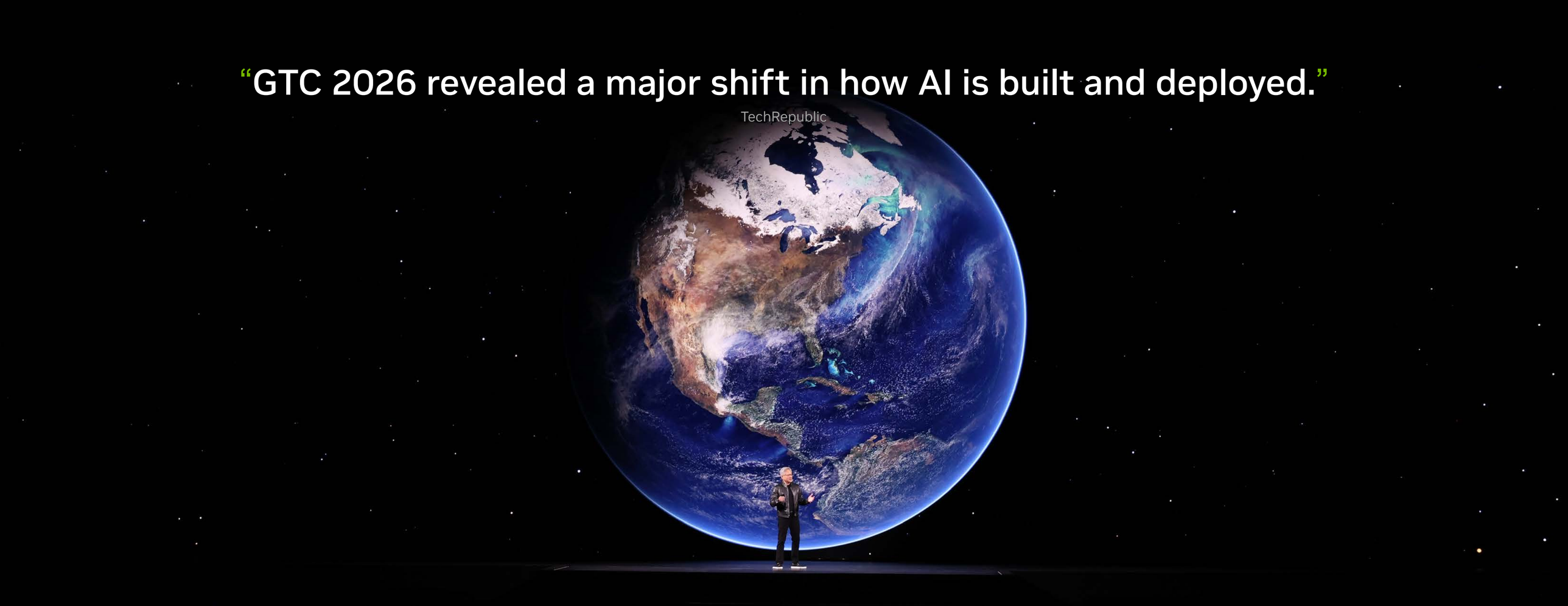
One year ago, we saw roughly $500 billion of demand. Today, that visibility is at least $1 trillion through 2027.

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

"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.

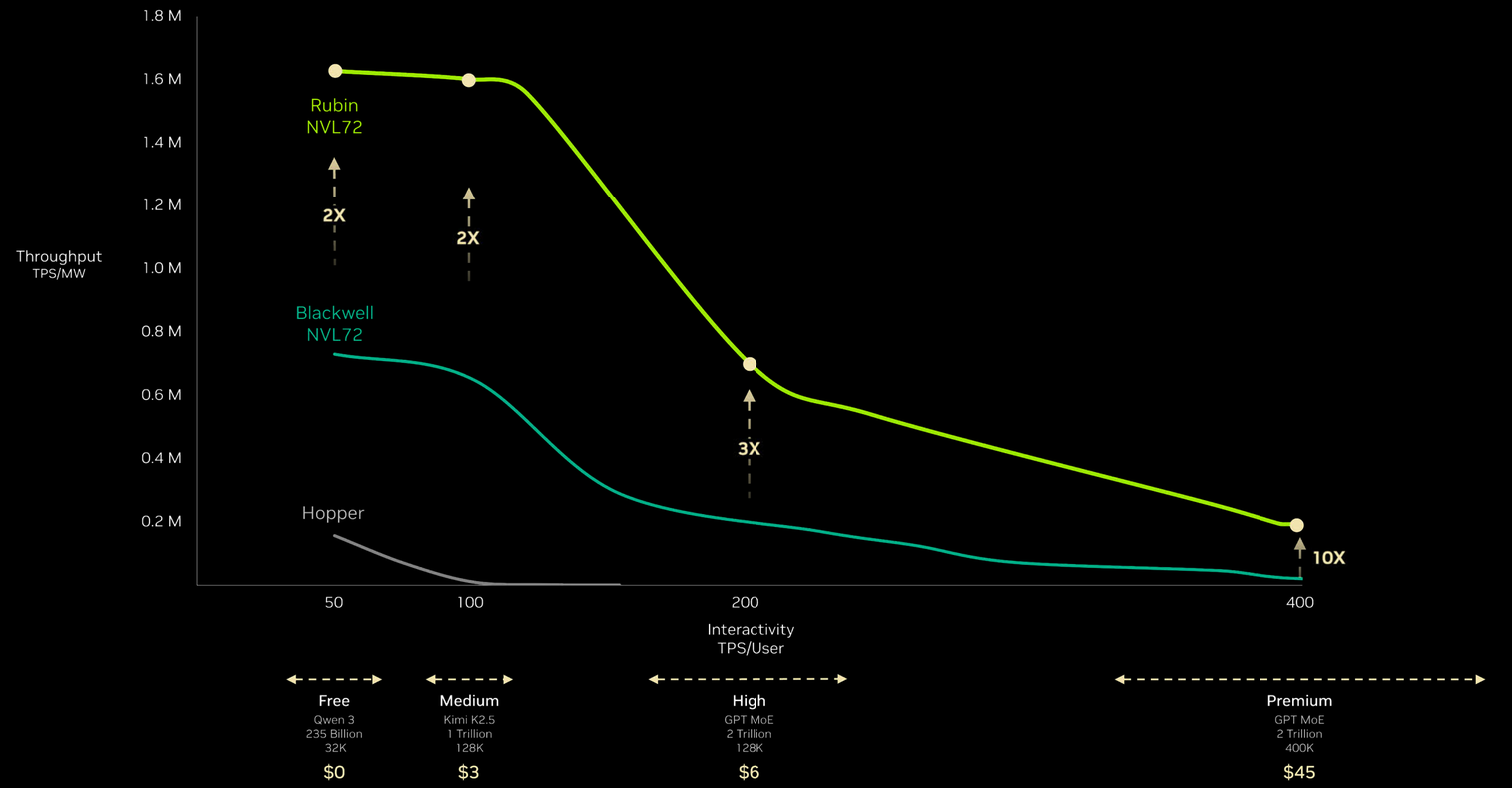
“NVIDIA clearly has the lowest cost per token by a country mile.”

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.



"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.

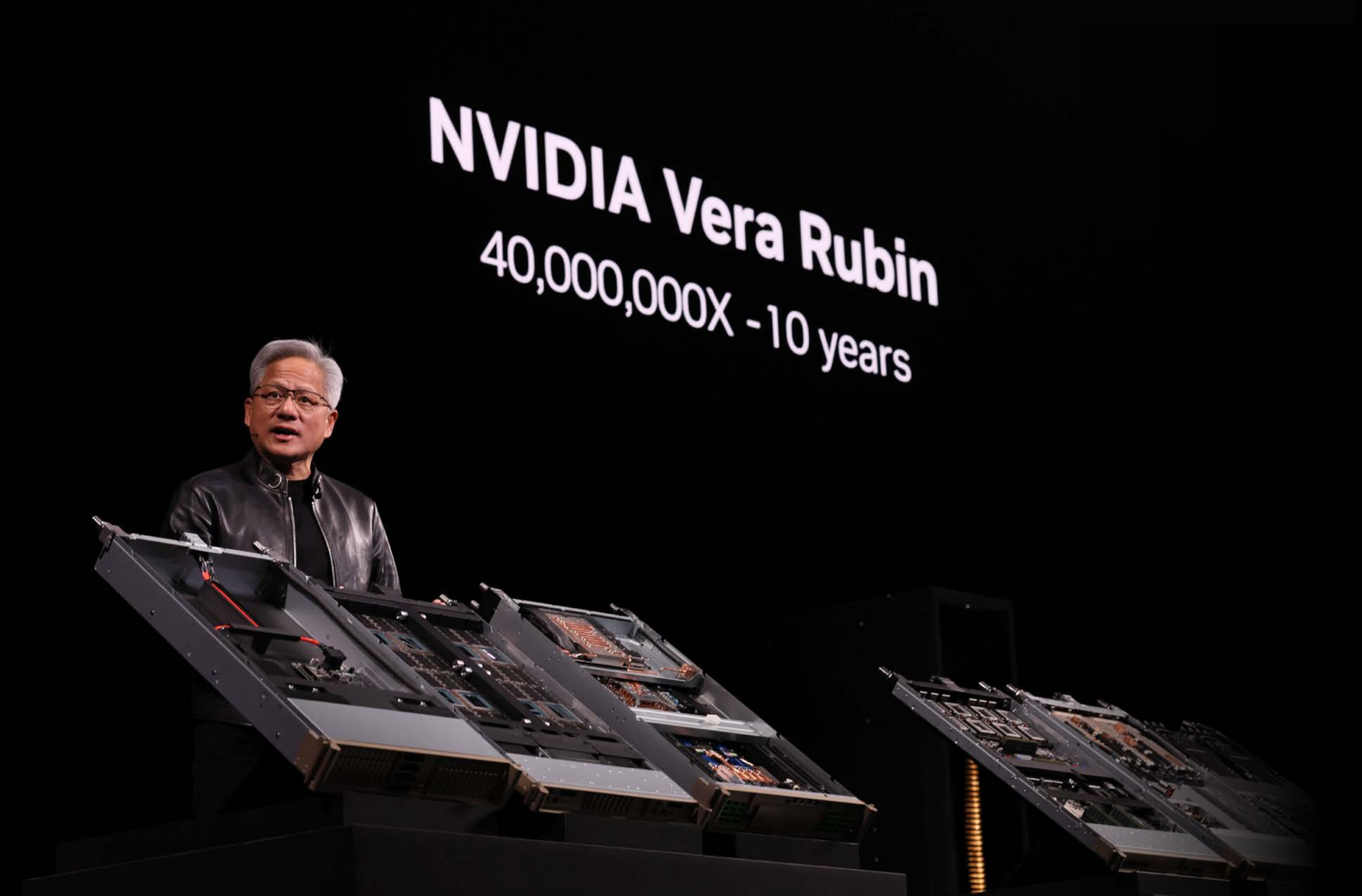
Higher throughput within a fixed power envelope lowers cost per token and expands revenue.

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.



“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.

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

Announcing NVIDIA Vera CPU Launch Partners

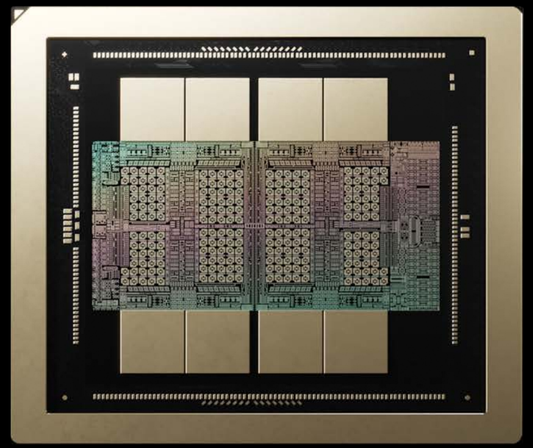
Cloud

Meta · Oracle Cloud Infrastructure · CoreWeave · Crusoe · Lambda · Nebius · Nscale · together.ai

OEMs and ODMs

Cisco · Dell Technologies · HPE · Lenovo · Supermicro · Aivres · ASRock Rack · ASUS · Compal · Foxconn · Gigabyte · hyve · Inventec · Mitac · msi · Pegatron · QCT · Wistron · Wiwynn

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 **55X**

9.6 PFLOPs (FP8)

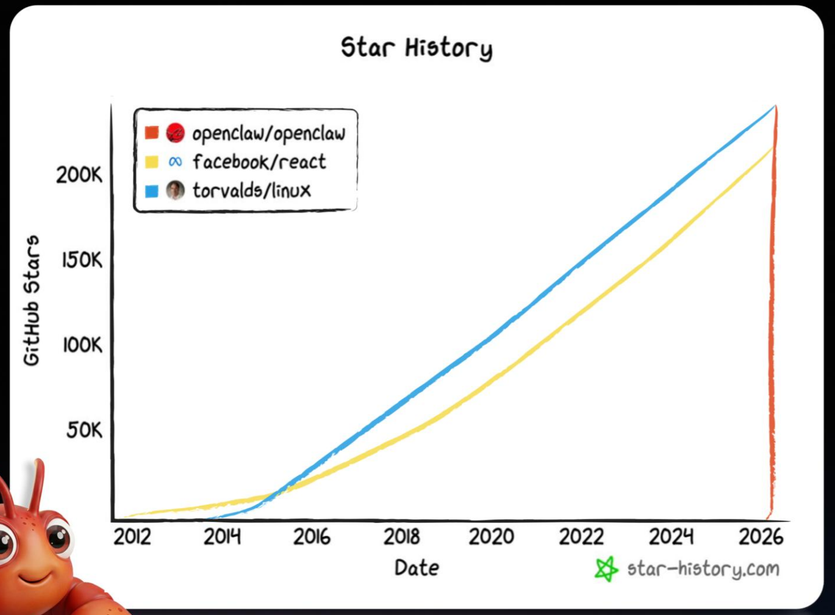
784B Transistors

“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.

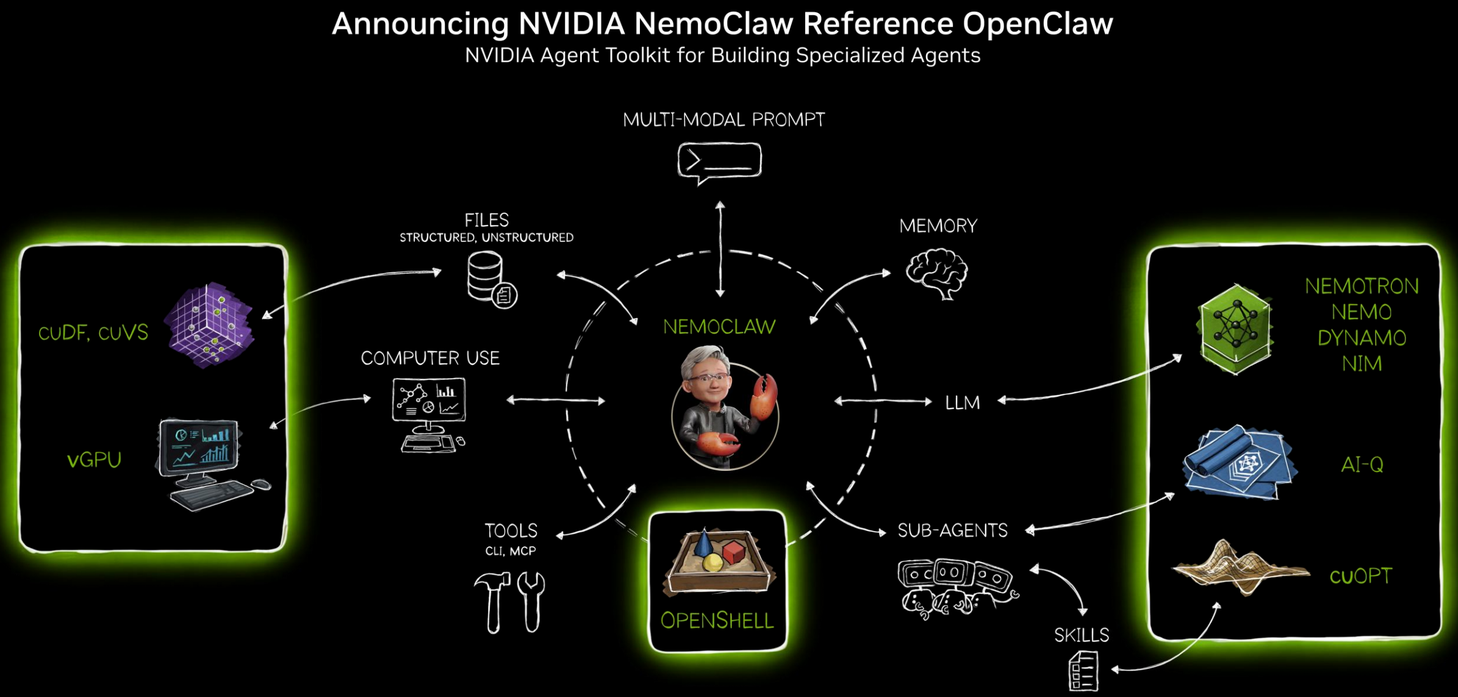
“NemoClaw brings security, scale to the agent platform taking over AI.”

VentureBeat

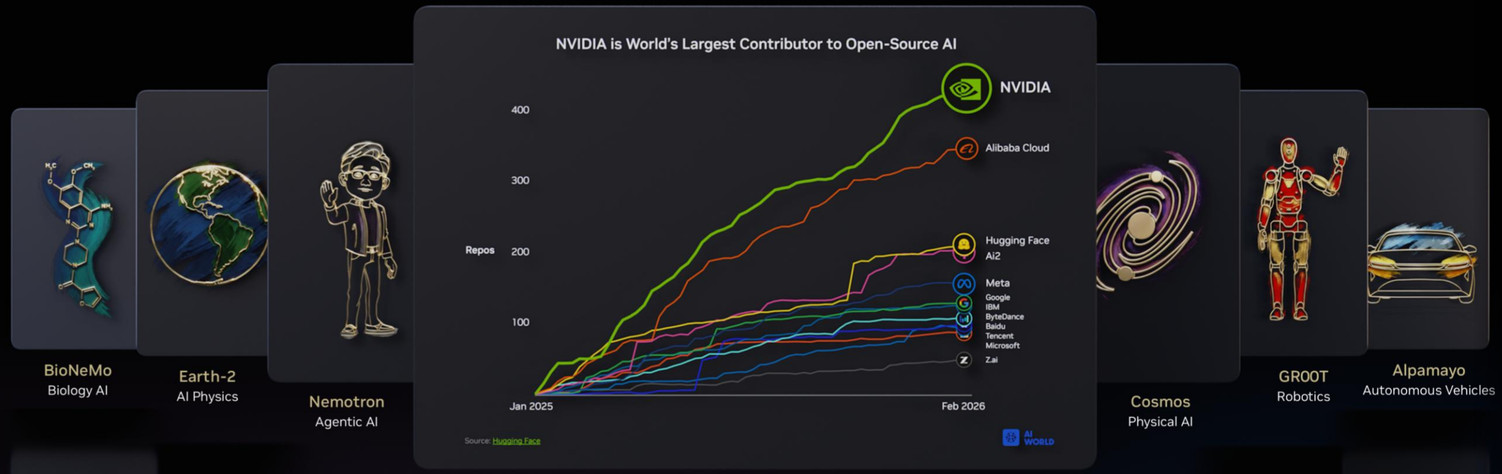
OpenClaw makes agentic systems possible. NemoClaw makes them deployable.

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

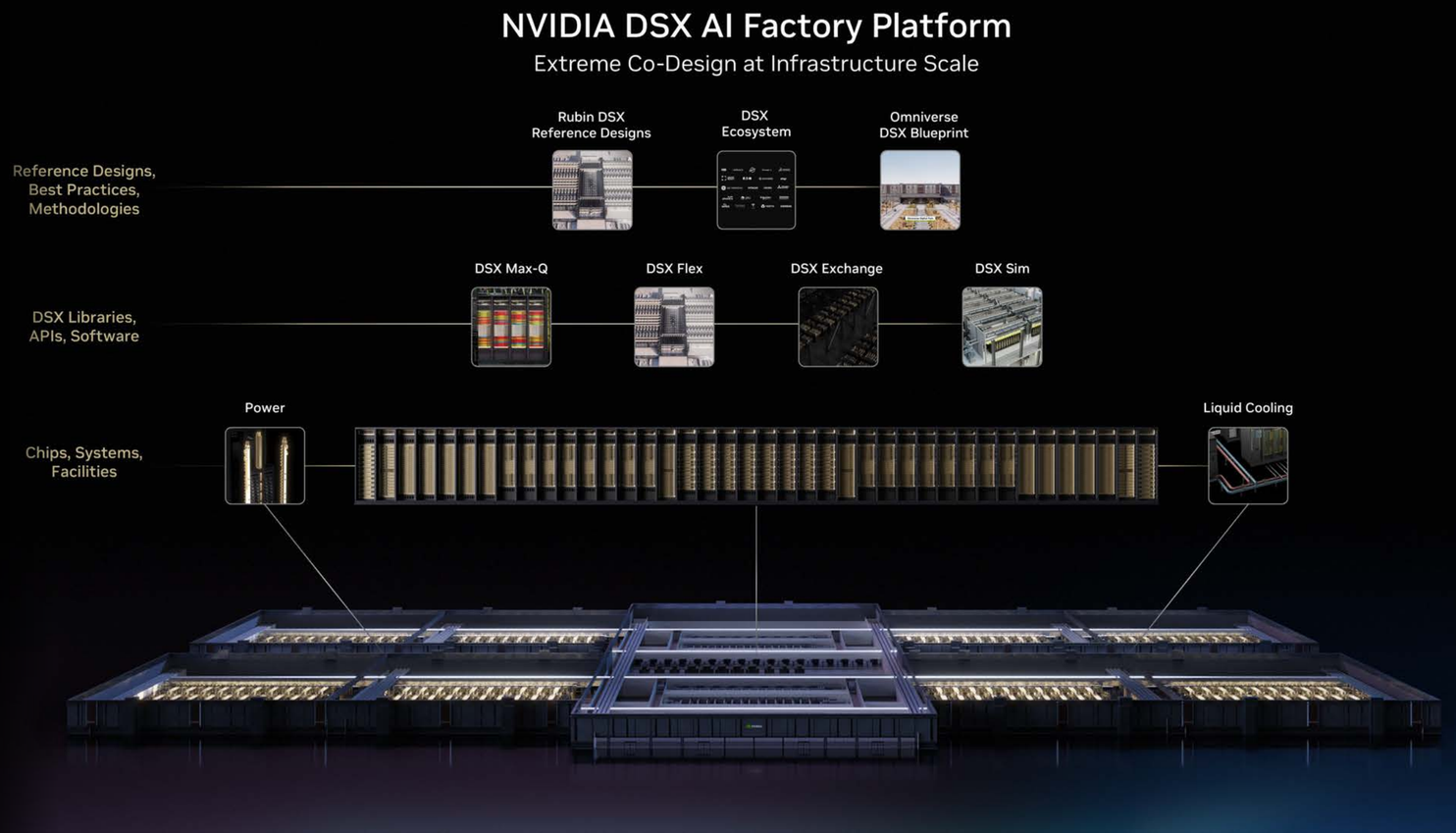
AI is expanding beyond centralized labs and proprietary platforms. AI natives, enterprises, industrial systems, sovereign initiatives, and regional cloud providers are building infrastructure tailored to their industries and data. Open frontier models enable this shift, allowing AI to be built and deployed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

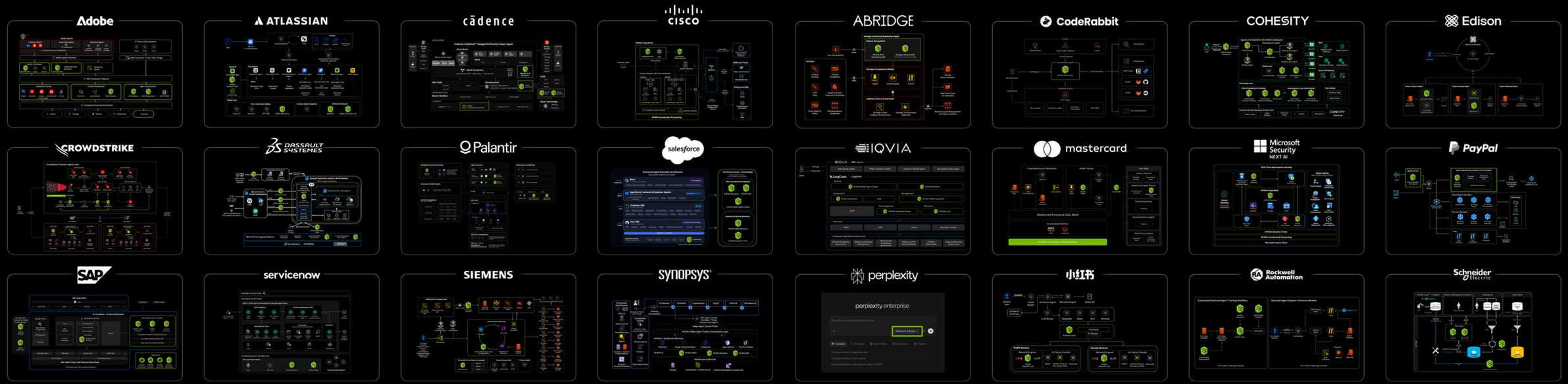
Punchbowl News

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.



Alibaba · Automation Anywhere · aws · baseten · ByteDance · CLOUDIAN · crewai · 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 Cloud Infrastructure · OpenEnv · openCode · OpenRouter · OpenHands · PRIME Intellect · PyTorch · SoftBank · Tencent Cloud · together.ai · unsloth · VAST · VULTR · WEKA

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."

WCCFTech

"$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."

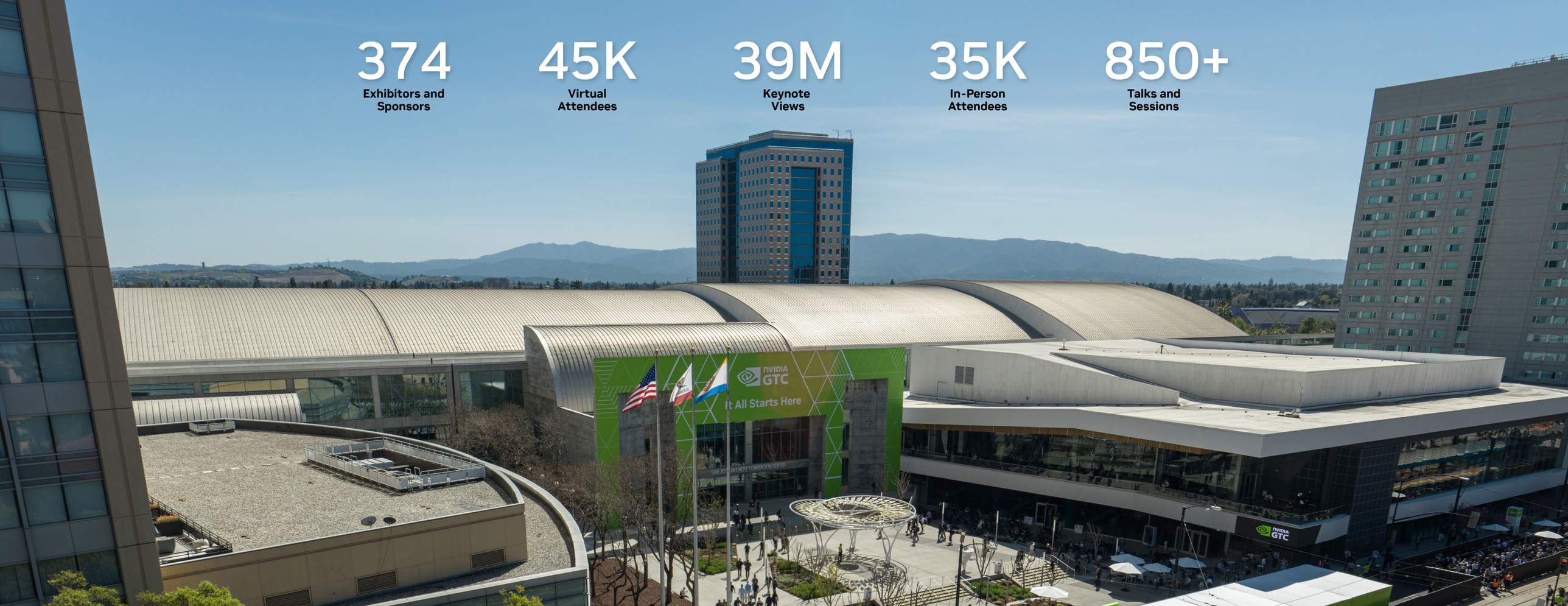
All In Podcast

"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


374
Exhibitors and Sponsors
45K
Virtual Attendees
39M
Keynote Views
35K
In-Person Attendees
850+
Talks and Sessions
NVIDIA GTC
t All Starts Here
NVIDIA GTC


ALL-IN


ALL-IN

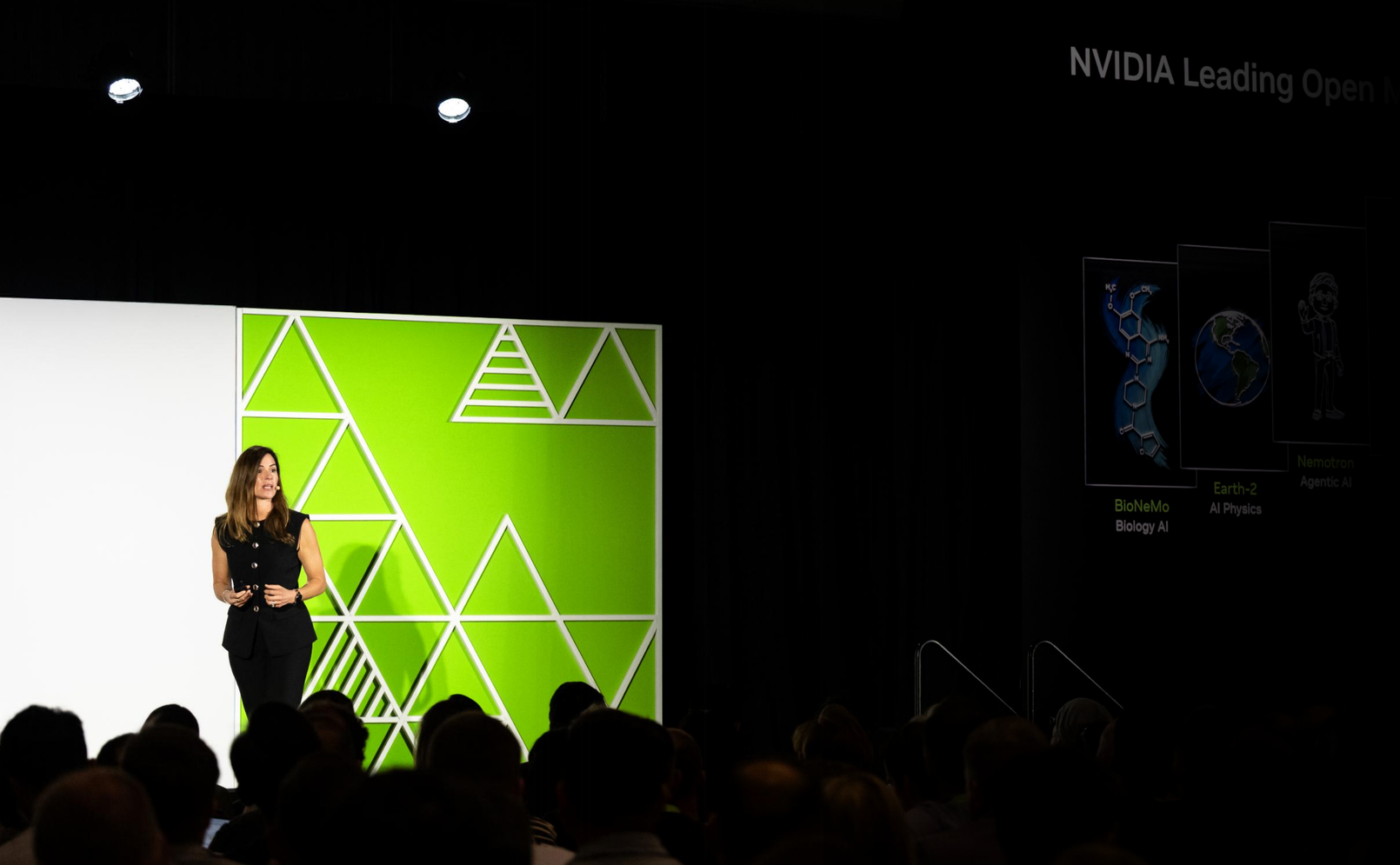

amaran

“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 Lebaredian, NVIDIA



“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

“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.”

Xinzhou Wu, NVIDIA


NVIDIA
GTC
SPX Switches
SPX
SuperNICs
GPU
GPU
GPU
GPU


NVIDIA
GTC


A Gear Store


Staff

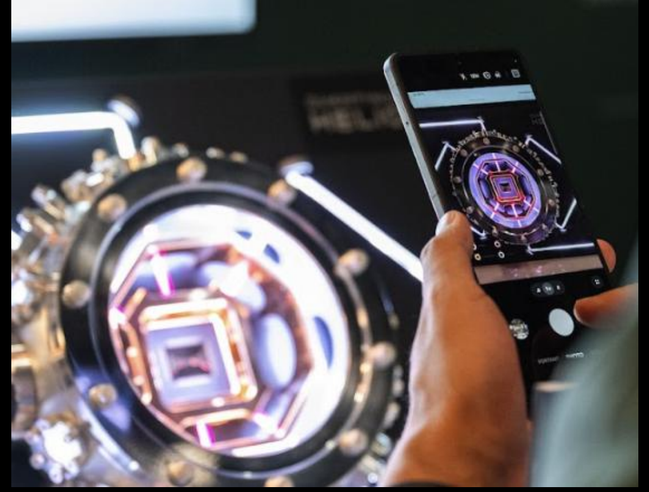

NVIDIA
NVIDIA


Staff


NVIDIA


Secure and simple
Run AI
in the Cloud
VERTIV
FOXCONN


NVIDIA




NVQLink With QubiC
NVQLink With Zurich Instrum


Information


Simplify your
AI journey

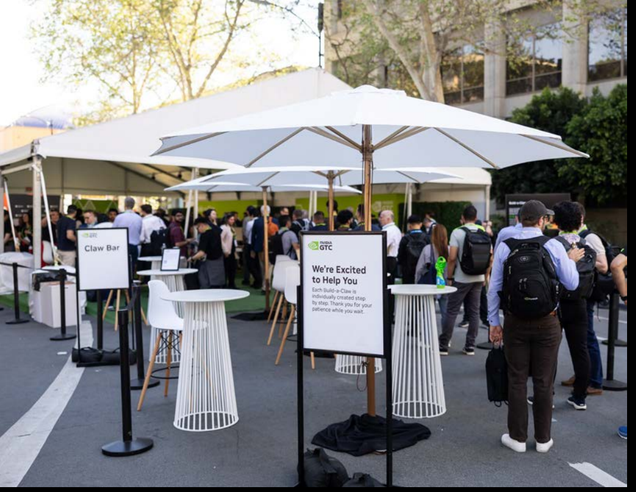

micron


World Wide
Technology

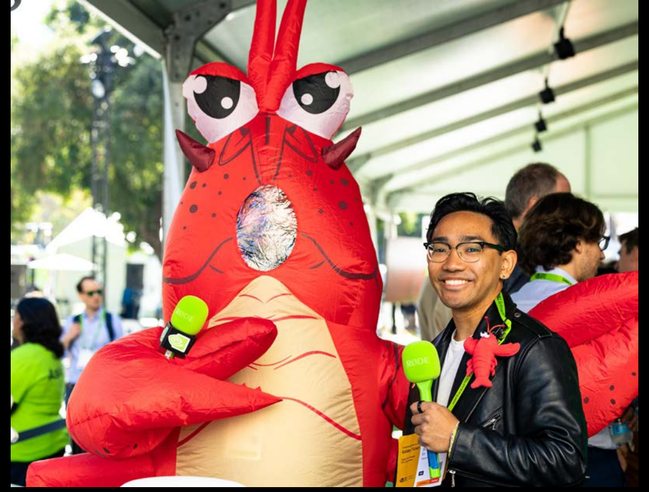

to Productio


Foxconn Omniverse Digital Twins (FODT)
Embodied Intelligence: AI Industrial Humanoid

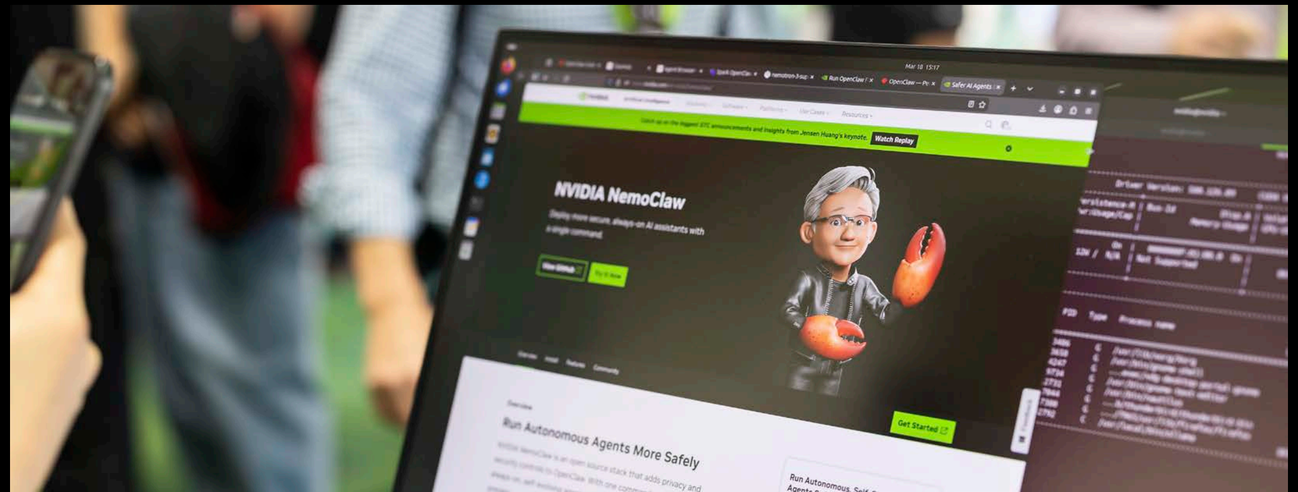

Design - Simula
Factories from Grid

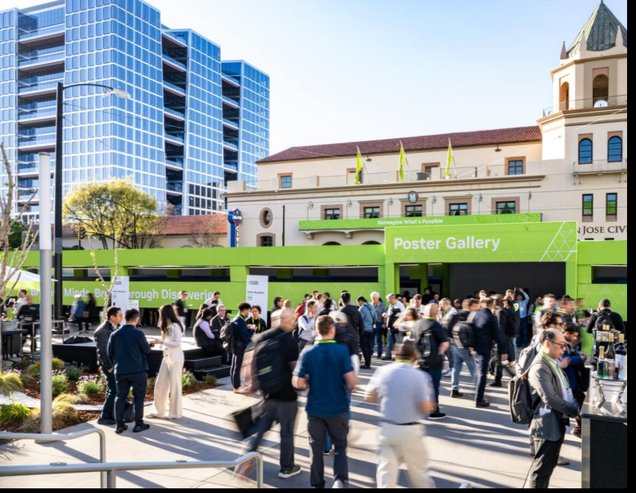

Claw Bar
We're Excited to Help You
Each Build-a-Claw is individually created step by step. Thank you for your patience while you wait.


Build-a-Claw

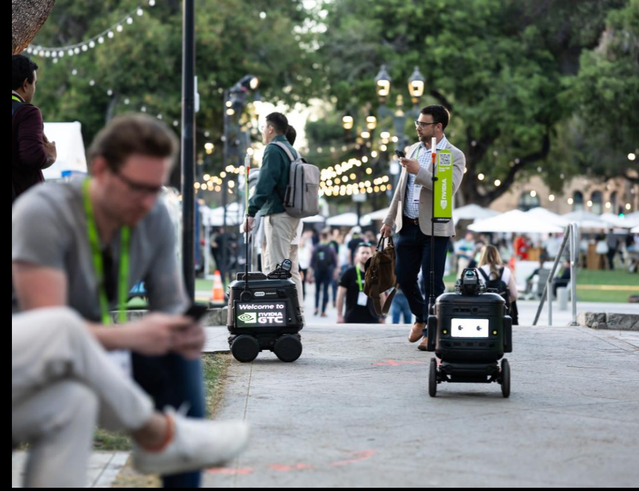



Build-a-Claw


NVIDIA NemoClaw
Watch Replay
Get Started
Run Autonomous Agents More Safely


Poster Gallery


NVIDIA
GTC
GTC Park




Red Hat


NVIDIA


ORDER HERE
PICK UP HERE


GTC
It All Starts Here
Hilton

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.


NVIDIA
GTC
2026


SARAH GUO
3RD PREGAME APPEARANCE
2X HOST
ALFRED LIN
2ND PREGAME APPEARANCE
2026 HOST DEBUT
GAVIN BAKER
2ND PREGAME APPEARANCE
2026 HOST DEBUT
HARRISON CHASE
JACOMO CORBO
JOE CREED
MICHAEL DELL
ANIRUDH DEVGAN
MARK ERELSTONE
AIDAN GOMEZ
MICHAEL INTRATOR
AKI JAIN
TIFFANY JANZEN
1ST 2026
ARTHUR MENSCH
DANIEL NADLER
DINESH NIRMAL
DEEPAK PATHAK
LIN QIAO
SAM RODRIQUES
ROBIN ROMBACH
ARAVIND SRINIVAS
PETER STEINBERGER
RAQUEL URTSAUN
VINCENT WEISSER

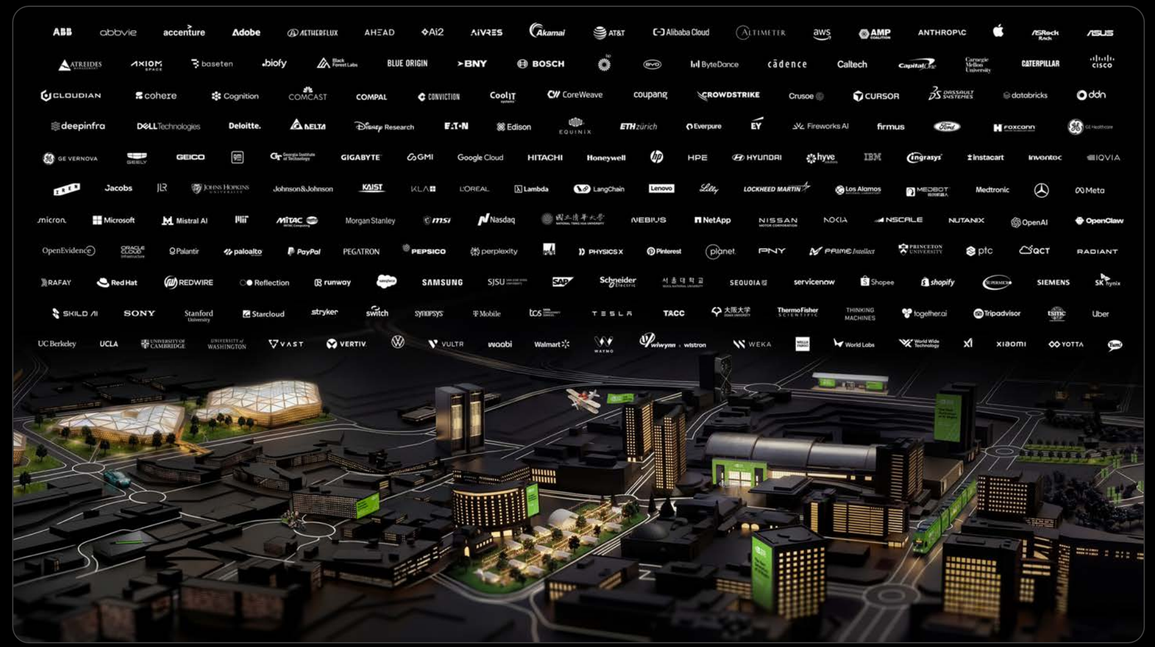
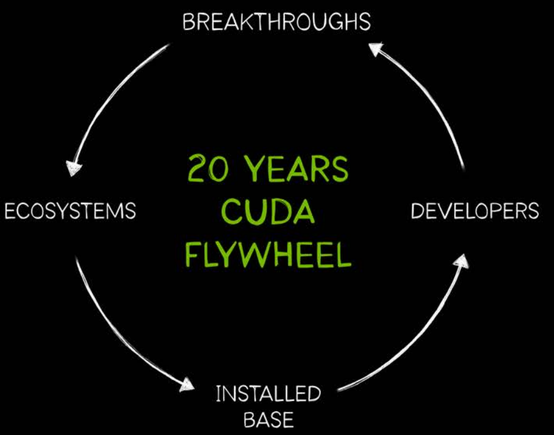
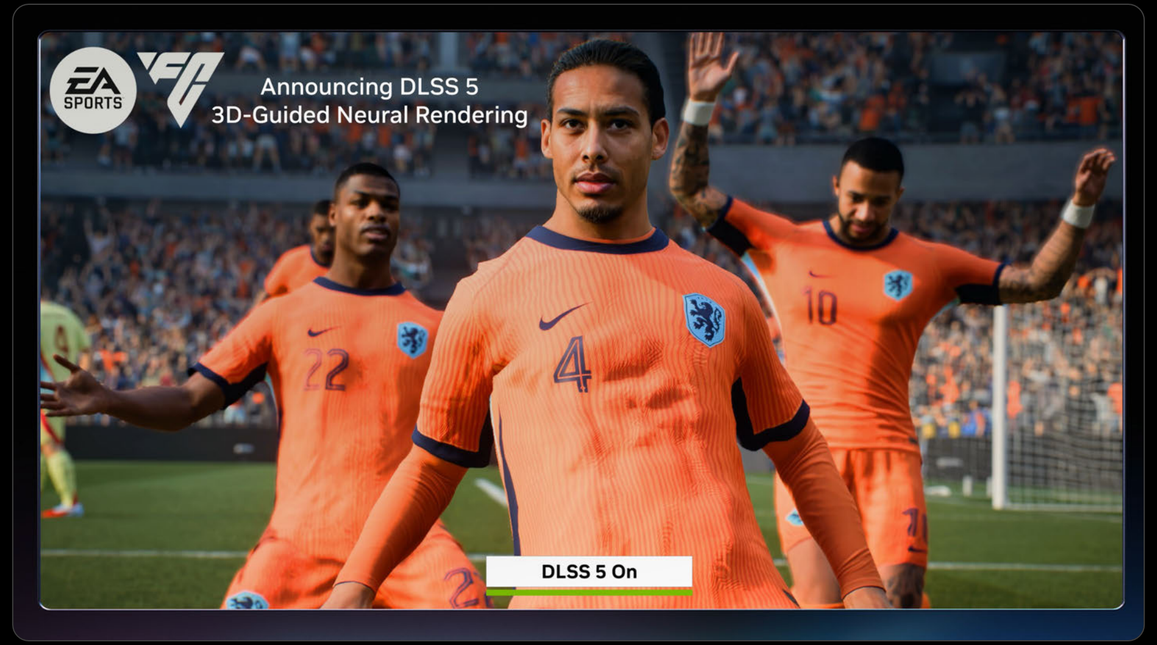
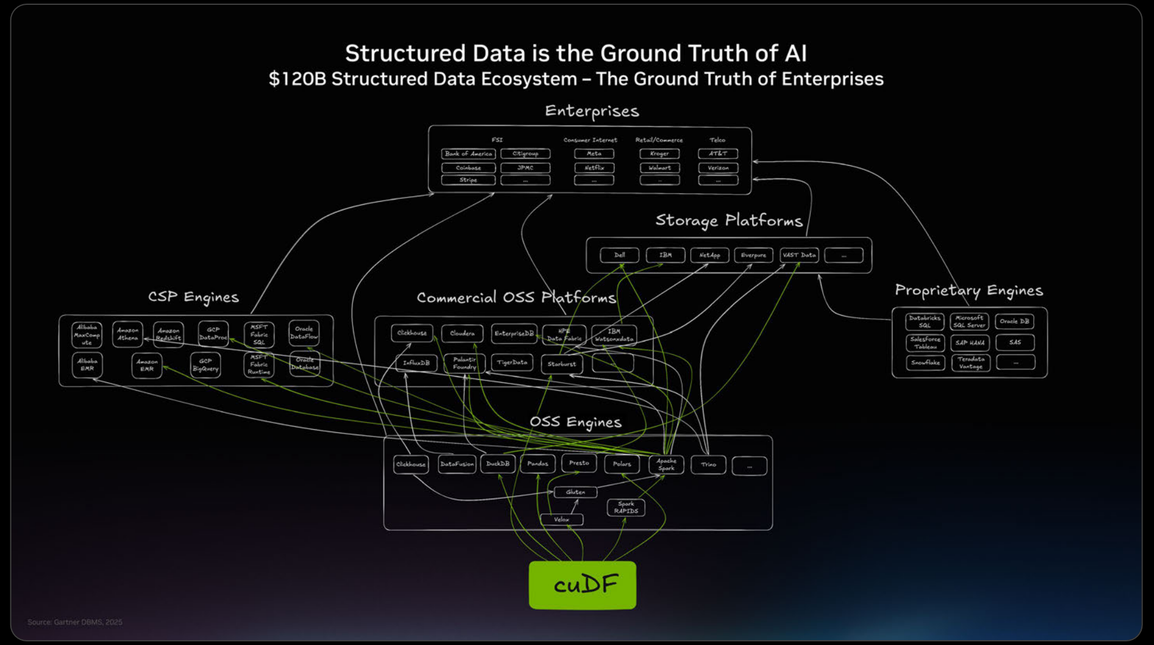


ABB
abbvie
accenture
Adobe
AETHERFLUX
AHEAD
Ai2
AIVRES
Akamai
AT&T
Alibaba Cloud
ALTIMETER
aws
AMP
ANTHROPIC
ASRock Rack
ASUS
ATREIDES
AXIOM SPACE
baseten
biofy
Black Forest Labs
BLUE ORIGIN
BNY
BOSCH
ByteDance
cadence
Caltech
Capital Tech
Carnegie Mellon University
CATERPILLAR
cisco
CLOUDIAN
cohere
Cognition
COMCAST
COMPAL
CONVICTION
CoolIT
CoreWeave
coupang
CROWDSTRIKE
Crusoe
CURSOR
DASSAULT SYSTÈMES
databricks
ddn
deepinfra
DELL Technologies
Deloitte.
DELTA
Disney Research
E.T.N
Edison
EQUINIX
ETH zürich
Everpure
EY
Fireworks AI
firmus
FORD
FOXCONN
GE Healthcare
GE VERNOVA
GEICO
Georgia Institute of Technology
GIGABYTE
GMI
Google Cloud
HITACHI
Honeywell
HPE
HYUNDAI
hyve
IBM
Imagry
instacart
inventec
IQVIA
Jacobs
Johns Hopkins University
Johnson&Johnson
KAIST
KLA
LORÉAL
Lambda
LangChain
Lenovo
Lilly
LOCKHEED MARTIN
Los Alamos
MEDBOT
Medtronic
Meta
micron
Microsoft
Mistral AI
MIT
MITAC
Morgan Stanley
mtsi
Nasdaq
NEBIUS
NetApp
NISSAN
NOKIA
NSCALE
NUTANIX
OpenAI
OpenClaw
OpenEvidence
ORACLE
Palantir
paloalto
PayPal
PEGATRON
PEPSICO
perplexity
PHYSICS X
Pinterest
planet
PNY
Prime Intellect
PRINCETON UNIVERSITY
ptc
QCT
RADIANT
RAFAY
Red Hat
REDWIRE
Reflection
runway
SAMSUNG
SJSU
SAP
Schneider Electric
SEQUOIA
servicenow
Shopee
shopify
SIEMENS
SK hynix
SKILD AI
SONY
Stanford University
Starcloud
stryker
switch
SYNOPSYS
T-Mobile
TCS
TESLA
TACC
ThermoFisher Scientific
THINKING MACHINES
togetherai
Tripadvisor
UBTECH
Uber
UC Berkeley
UCLA
University of Cambridge
University of Washington
VAST
VERTIV
VOLKSWAGEN
VULTR
woobi
Walmart
WAYMO
wayve
wistron
WEKA
World Labs
World Wide Technology
xAI
XIAOMI
YOTTA

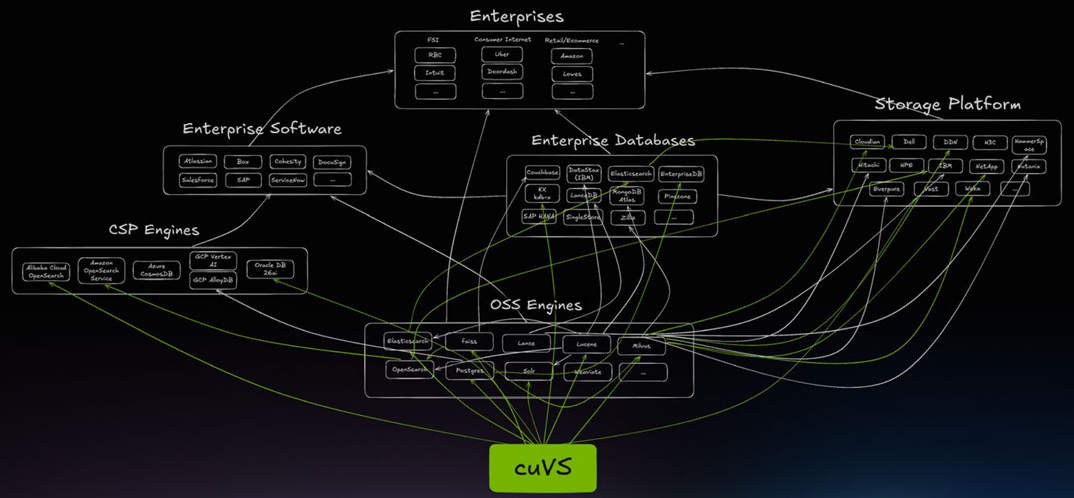

BREAKTHROUGHS
ECOSYSTEMS
20 YEARS
CUDA
FLYWHEEL
DEVELOPERS
INSTALLED
BASE


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
Enterprises
FSI
Consumer Internet
Retail/Commerce
Telco
Storage Platforms
Dell
IBM
NetApp
Everpure
VAST Data
CSP Engines
Commercial OSS Platforms
Proprietary Engines
OSS Engines
cuDF
Source: Gartner DBMS, 2025

Unstructured Data is the Context of AI

100's of Zettabytes Per Year of Unstructured Data – Growing Exponentially

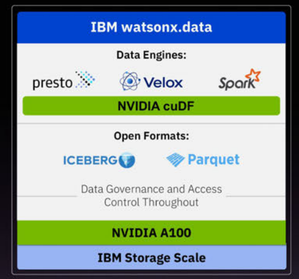


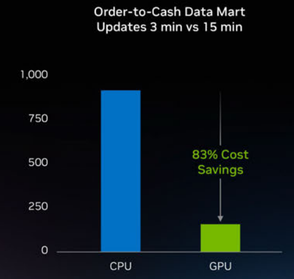
Source: IDC Global DataSphere, 2025

NVIDIA ❤ IBM ❤ Nestlé

*"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

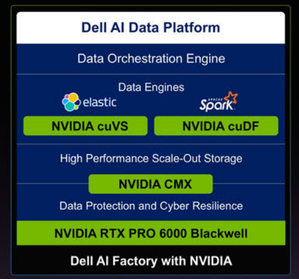


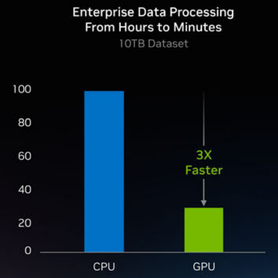


NVIDIA ❤ DELL Technologies ❤ NTT DATA

"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

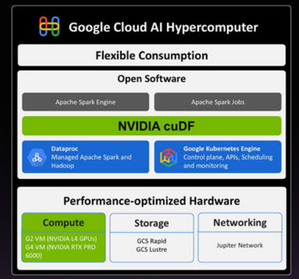


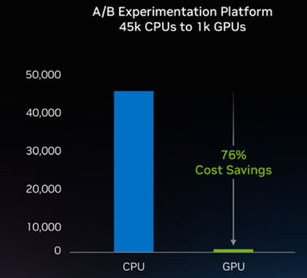


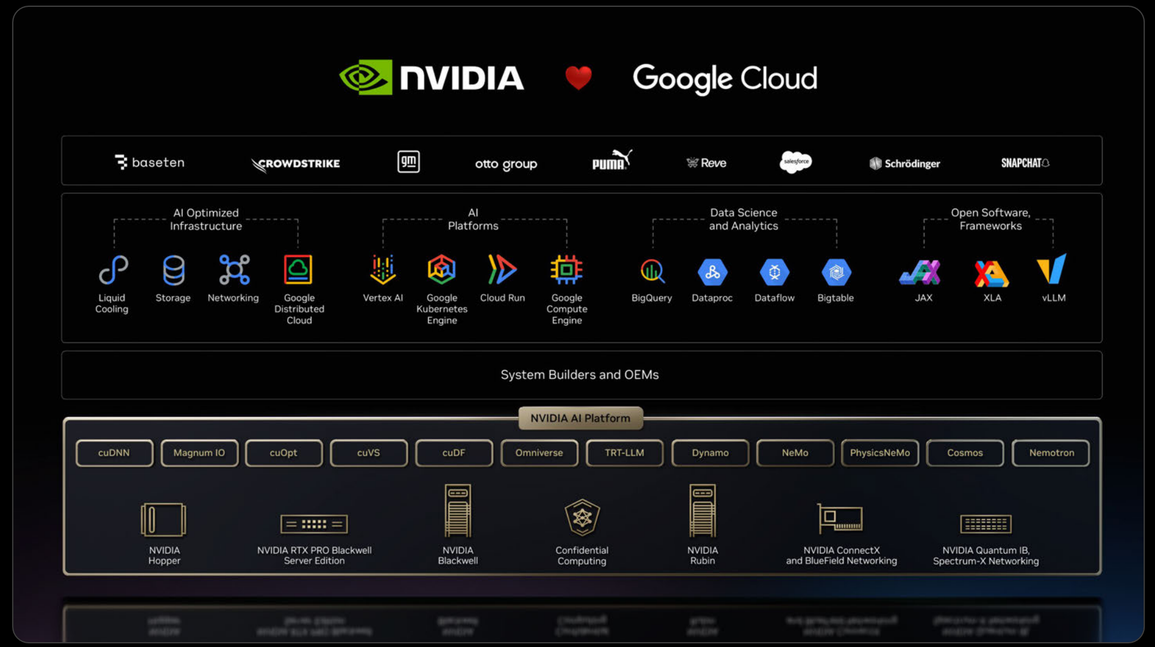
NVIDIA ❤ Google Cloud ❤

"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

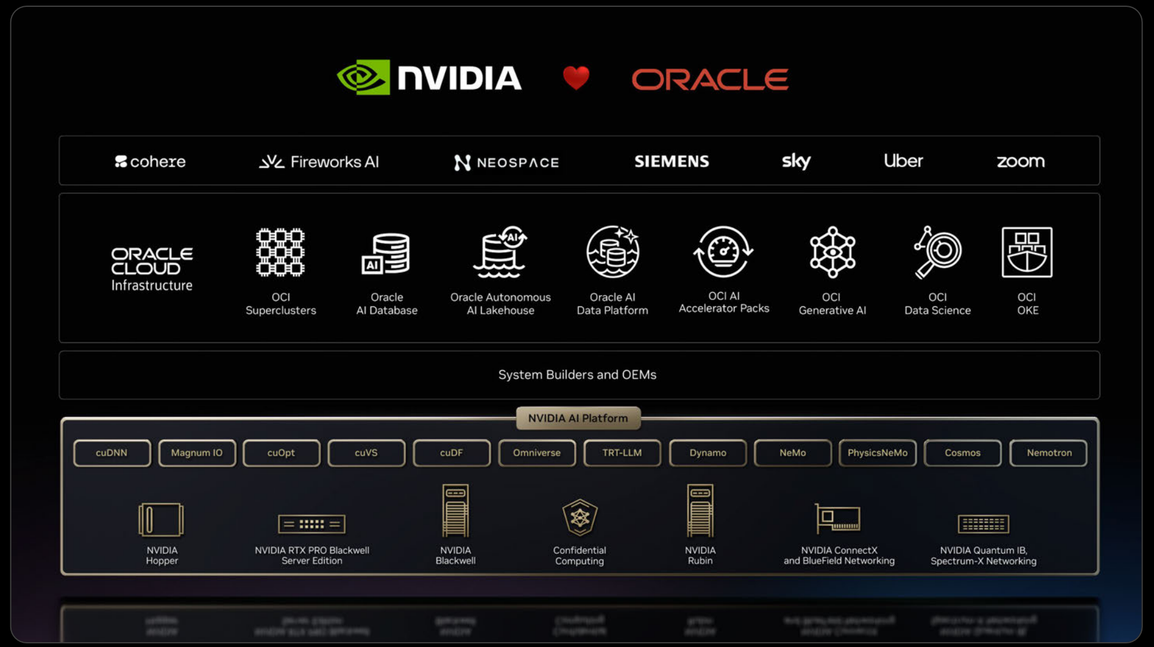


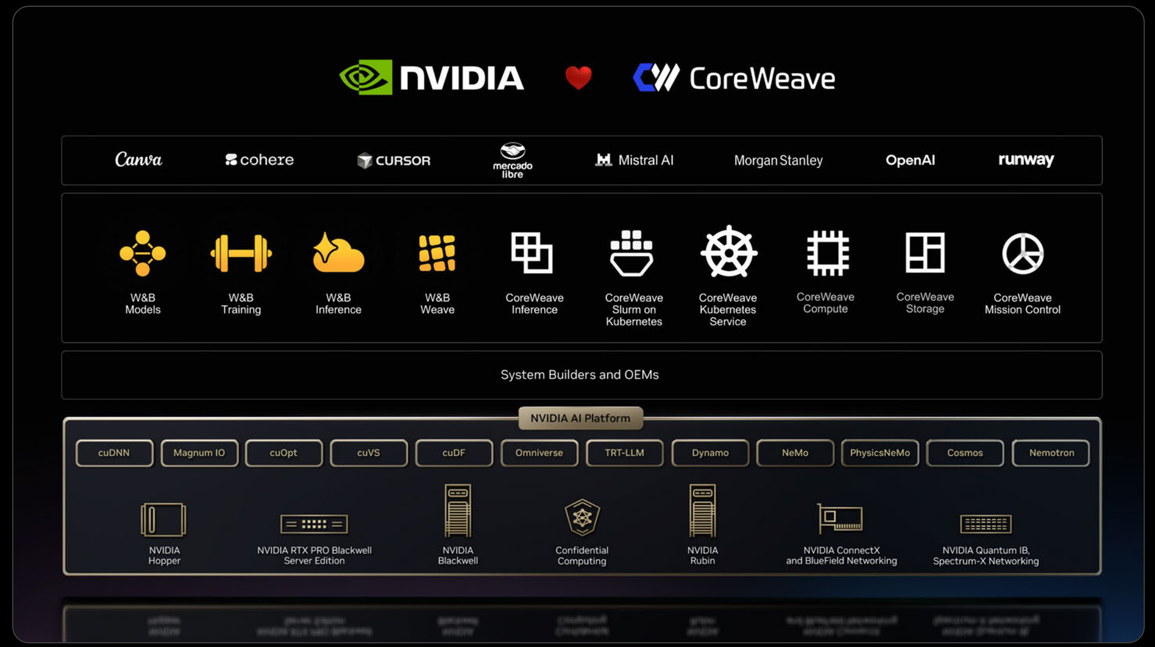



NVIDIA
Google Cloud
baseten
CROWDSTRIKE
otto group
PUMA
Reve
Schrödinger
SNAPCHAT
AI Optimized Infrastructure
Liquid Cooling
Storage
Networking
Google Distributed Cloud
AI Platforms
Vertex AI
Google Kubernetes Engine
Cloud Run
Google Compute Engine
Data Science and Analytics
BigQuery
Dataproc
Dataflow
Bigtable
Open Software, Frameworks
JAX
XLA
vLLM
System Builders and OEMs
NVIDIA AI Platform
cuDNN
Magnum IO
cuOpt
cuVS
cuDF
Omniverse
TRT-LLM
Dynamo
NeMo
PhysicsNeMo
Cosmos
Nemotron
NVIDIA Hopper
NVIDIA RTX PRO Blackwell Server Edition
NVIDIA Blackwell
Confidential Computing
NVIDIA Rubin
NVIDIA ConnectX and BlueFlield Networking
NVIDIA Quantum IB, Spectrum-X Networking

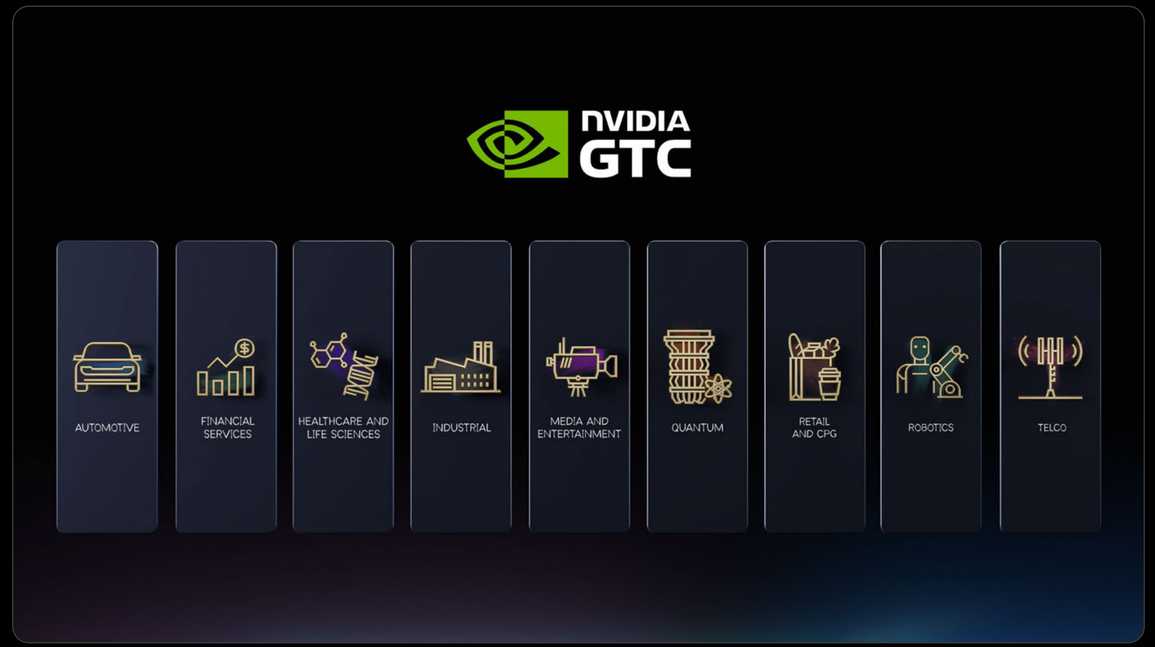

NVIDIA
aws
CROWDSTRIKE
databricks
HUMAIN
OpenAI
perplexity
Palantir
qodo
snowflake
Amazon Braket
Amazon EKS
AWS Nitro Enclaves
AWS Batch
Amazon EMR
Amazon Bedrock
Amazon SageMaker AI
Amazon EC2
Amazon EFA
AWS ParallelCluster
System Builders and OEMs
NVIDIA AI Platform
cuDNN
Magnum IO
cuOpt
cuVS
cuDF
Omniverse
TRT-LLM
Dynamo
NeMo
PhysicsNeMo
Cosmos
Nemotron
NVIDIA Hopper
NVIDIA RTX PRO Blackwell Server Edition
NVIDIA Blackwell
Confidential Computing
NVIDIA Rubin
NVIDIA ConnectX and BlueFlield Networking
NVIDIA Quantum IB, Spectrum-X Networking


NVIDIA
Microsoft Azure
ANTHROPIC
BlackRock
KRONES
Rockwell Automation
SYNOPSYS
Microsoft Foundry
Windows 365
Microsoft Fabric
Azure Container Apps
Azure Kubernetes Service
Foundry Local
Microsoft 365 Copilot
Microsoft Bing
System Builders and OEMs
NVIDIA AI Platform
cuDNN
Magnum IO
cuOpt
cuVS
cuDF
Omniverse
TRT-LLM
Dynamo
NeMo
PhysicsNeMo
Cosmos
Nemotron
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


NVIDIA
ORACLE
cohere
Fireworks AI
NEOSPACE
SIEMENS
sky
Uber
zoom
ORACLE CLOUD Infrastructure
OCI Superclusters
Oracle AI Database
Oracle Autonomous AI Lakehouse
Oracle AI Data Platform
OCI AI Accelerator Packs
OCI Generative AI
OCI Data Science
OCI OKE
System Builders and OEMs
NVIDIA AI Platform
cuDNN
Magnum IO
cuOpt
cuVS
cuDF
Omniverse
TRT-LLM
Dynamo
NeMo
PhysicsNeMo
Cosmos
Nemotron
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


NVIDIA
CoreWeave
Canva
cohere
CURSOR
mercado libre
Mistral AI
Morgan Stanley
OpenAI
runway
W&B Models
W&B Training
W&B Inference
W&B Weave
CoreWeave Inference
CoreWeave Slurm on Kubernetes
CoreWeave Kubernetes Service
CoreWeave Compute
CoreWeave Storage
CoreWeave Mission Control
System Builders and OEMs
NVIDIA AI Platform
cuDNN
Magnum IO
cuOpt
cuVS
cuDF
Omniverse
TRT-LLM
Dynamo
NeMo
PhysicsNeMo
Cosmos
Nemotron
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

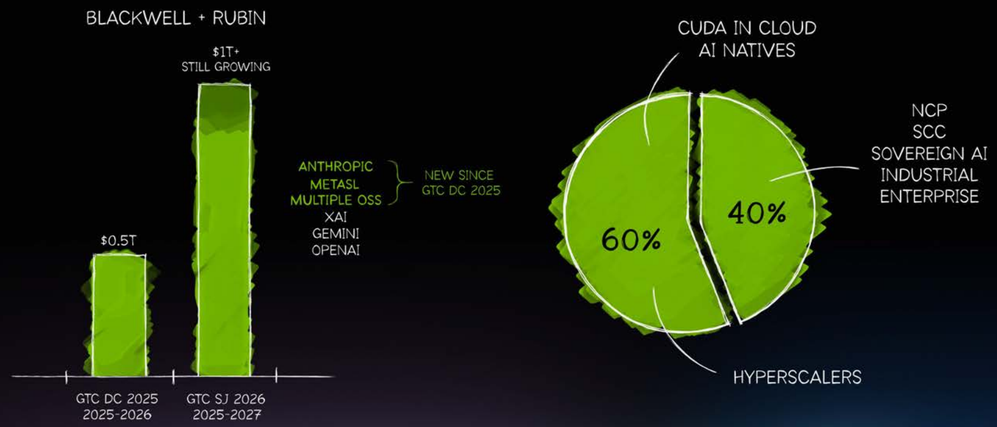

NVIDIA
Palantir
DELL Technologies
accenture
CenterPoint Energy
Hertz
LOWE'S
TETON RIDGE
Rubix
Ontology
Pipeline Builder
Fusion
Agent Studio
Model Studio
AIP Logic
Model Catalog
Workshop
AI FDE
Dell
NVIDIA AI Platform
cuDNN
Magnum IO
cuOpt
cuVS
cuDF
Omniverse
TRT-LLM
Dynamo
NeMo
PhysicsNeMo
Cosmos
Nemotron
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


NVIDIA
GTC
AUTOMOTIVE
FINANCIAL SERVICES
HEALTHCARE AND LIFE SCIENCES
INDUSTRIAL
MEDIA AND ENTERTAINMENT
QUANTUM
RETAIL AND CPG
ROBOTICS
TELCO


NVIDIA
AI Natives
AI for Auto
AI for Customer Support
AI for Engineering
AI for Healthcare
AI for Robotics
AI for Search
AI for Software Development
DL Frameworks
Inference Frameworks
Agent Frameworks/Protocols
Frontier Model Builders
Model to Production
NVIDIA AI Platform
cuDNN
Magnum IO
cuOpt
cuVS
cuDF
Omniverse
TRT-LLM
Dynamo
NeMo
PhysicsNeMo
Cosmos
Nemotron
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

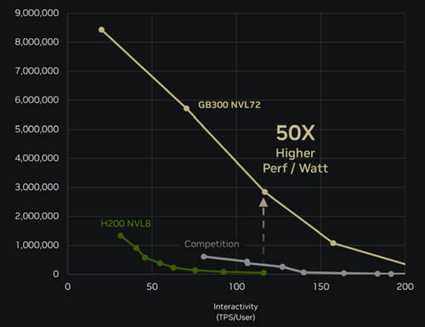
Inference Inflection Drives Strong Growth
NVIDIA Full-Stack Expanding AI to All Regions and Industries



NVIDIA Extreme Co-Design Revolutionized Token Cost

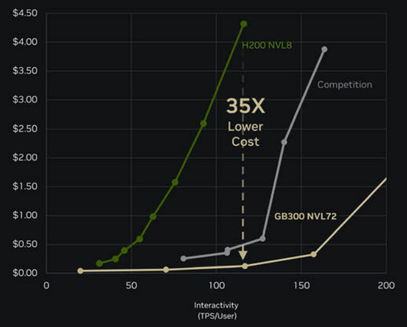
"GB NVL72 Inference King"

Tokens per Watt Drives Factory Revenue



InferenceX by semianalysis

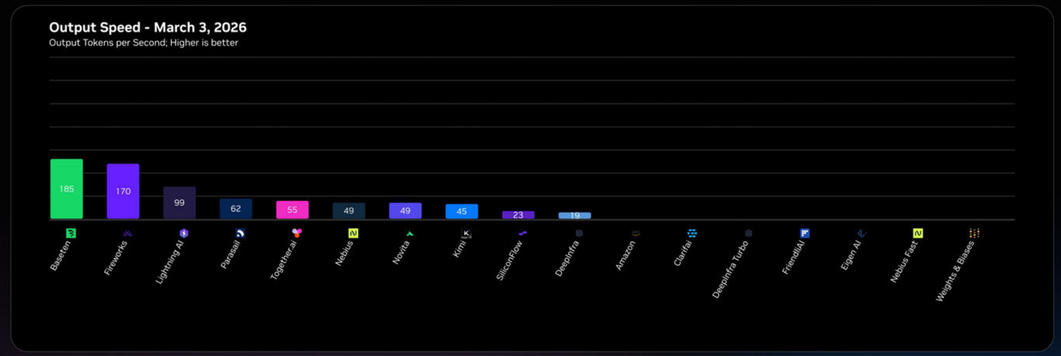
Performance Drives Token Cost



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

NVIDIA is the Global Standard for AI Inference at Scale

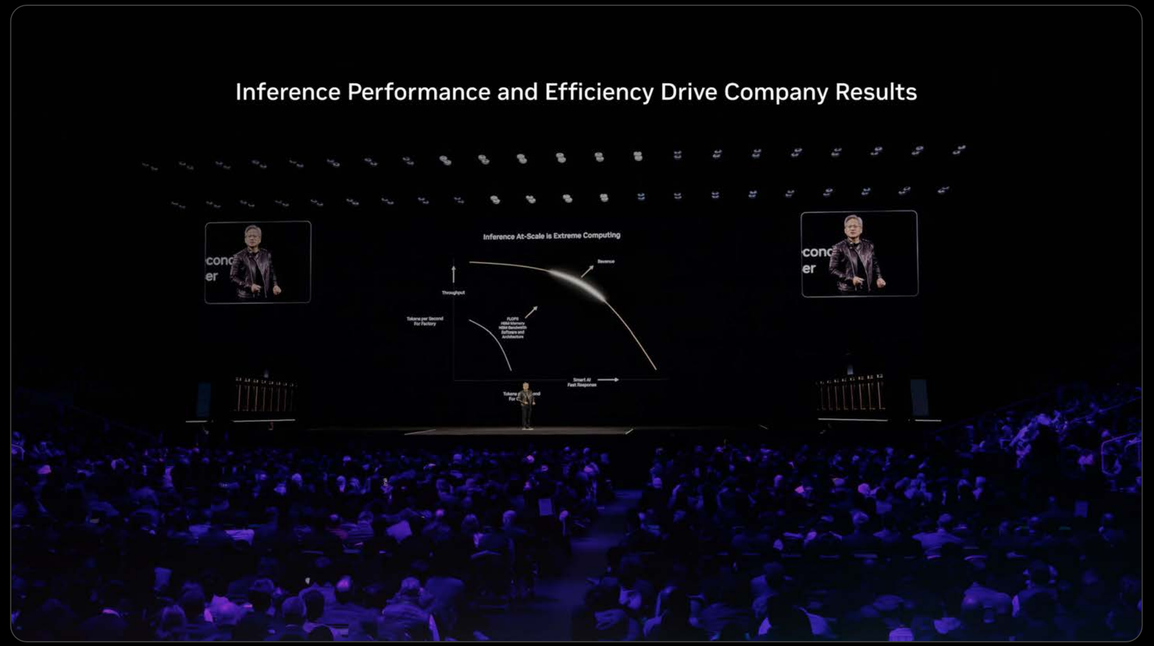
All Leading Inference End-Points Run on NVIDIA

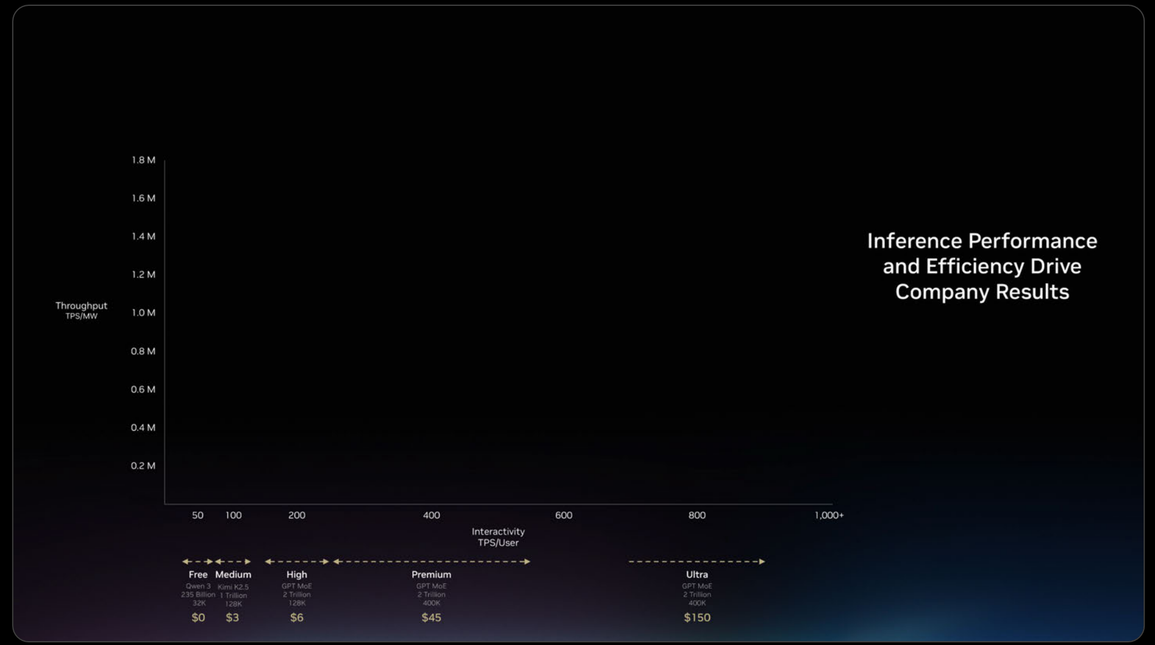


Kimi K2.5 Reasoning

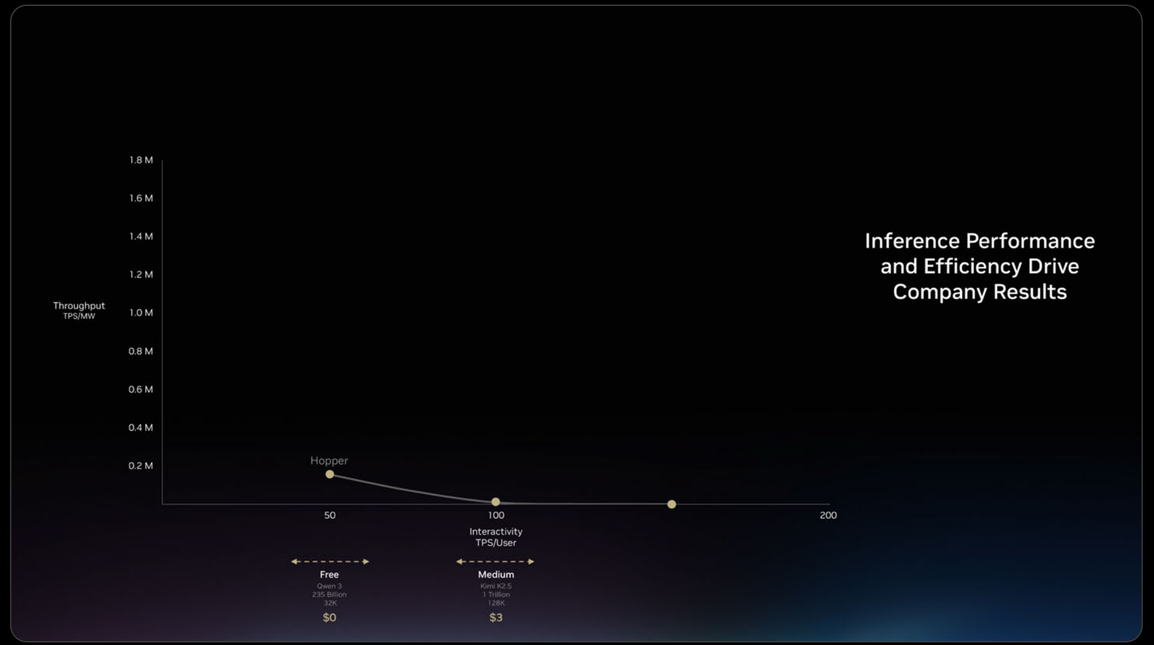
AI Factories are the Industrial Infrastructure of the AI Era
Inference is the Workload
Tokens are the New Commodity
Compute is Revenue

Inference Performance and Efficiency Drive Company Results

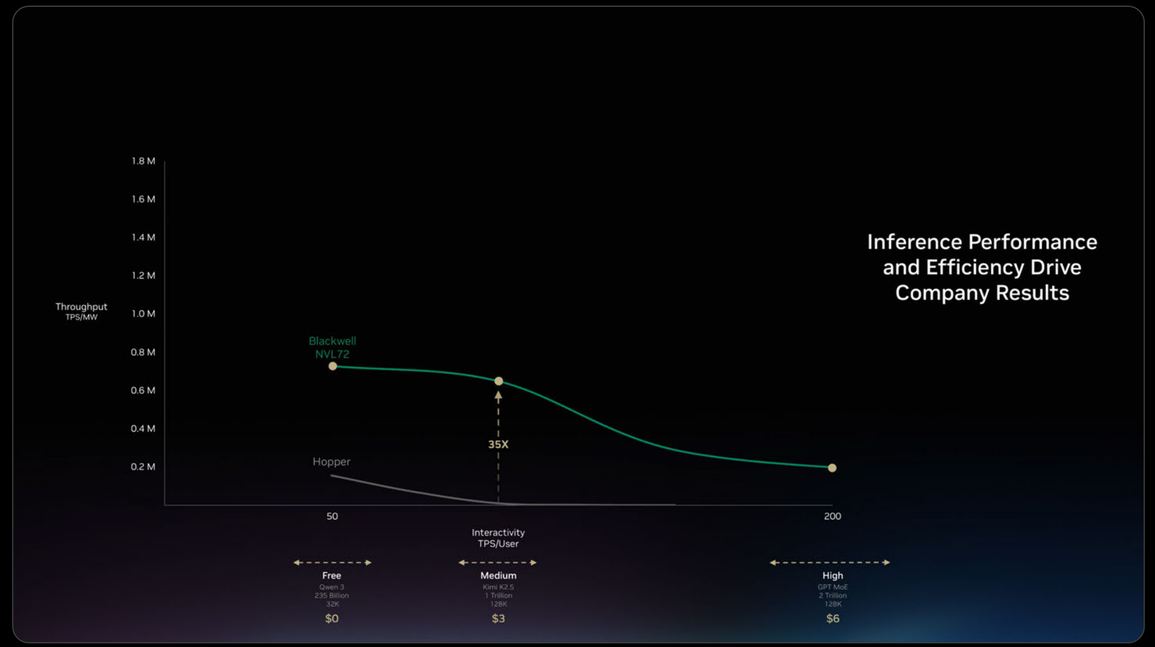


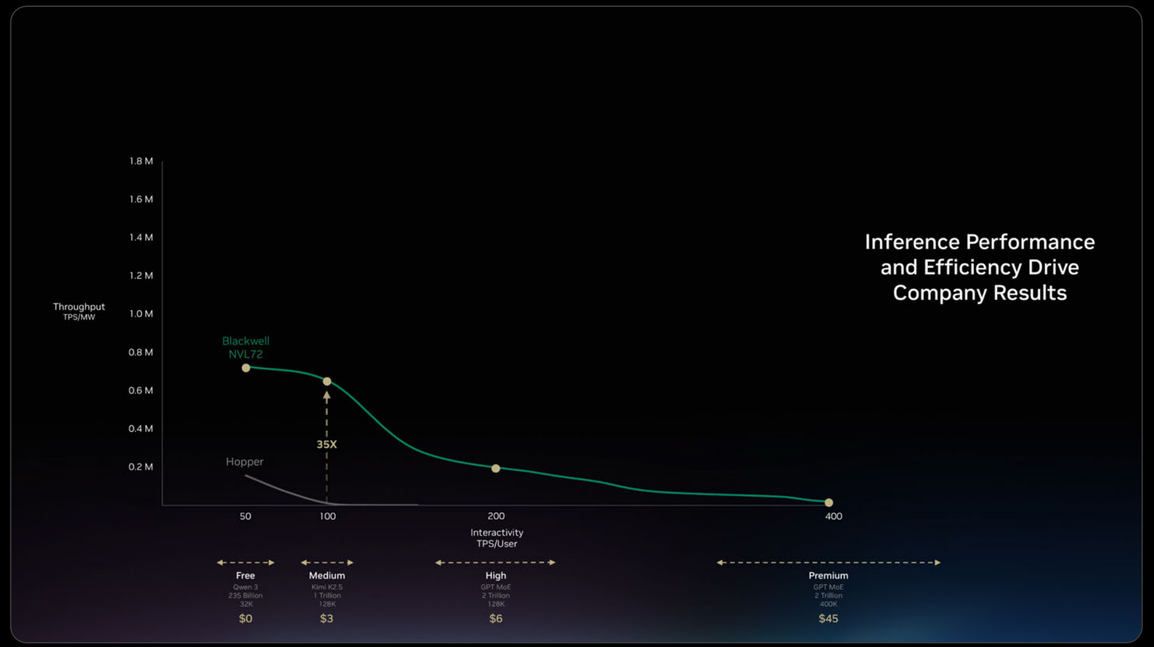


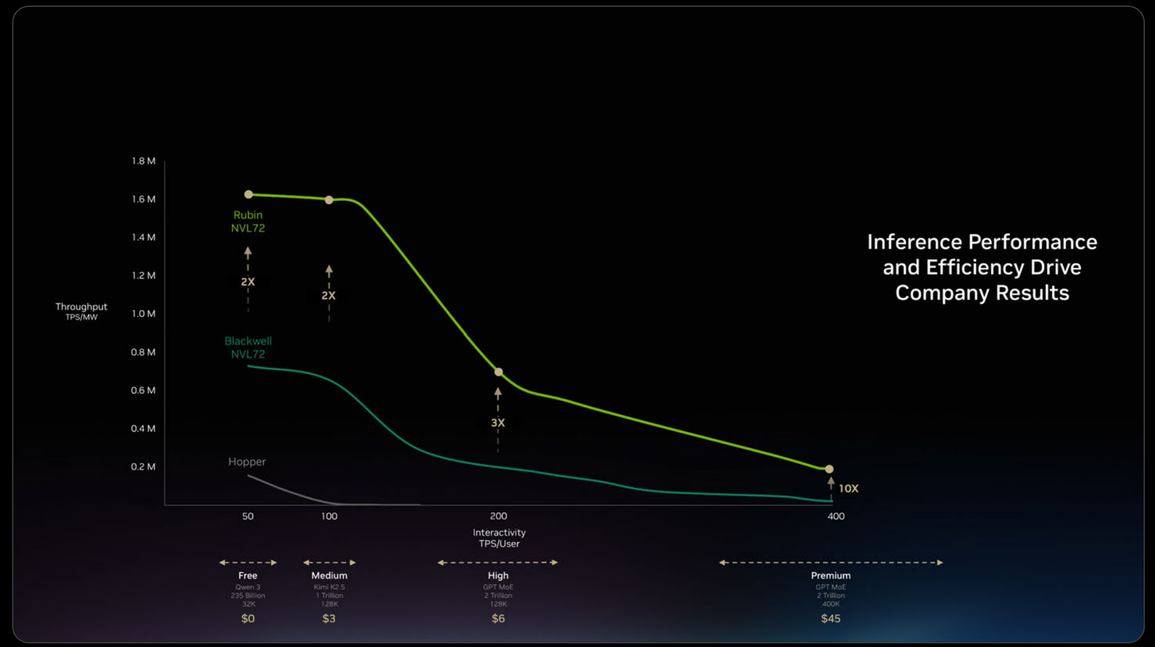
Inference Performance and Efficiency Drive Company Results

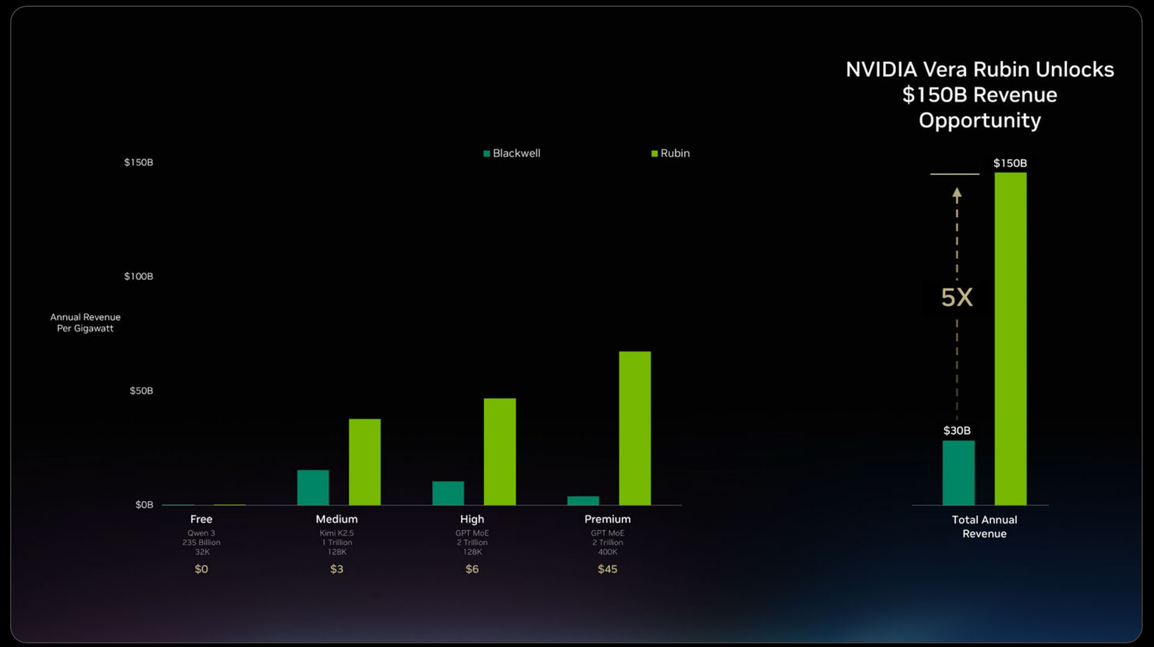


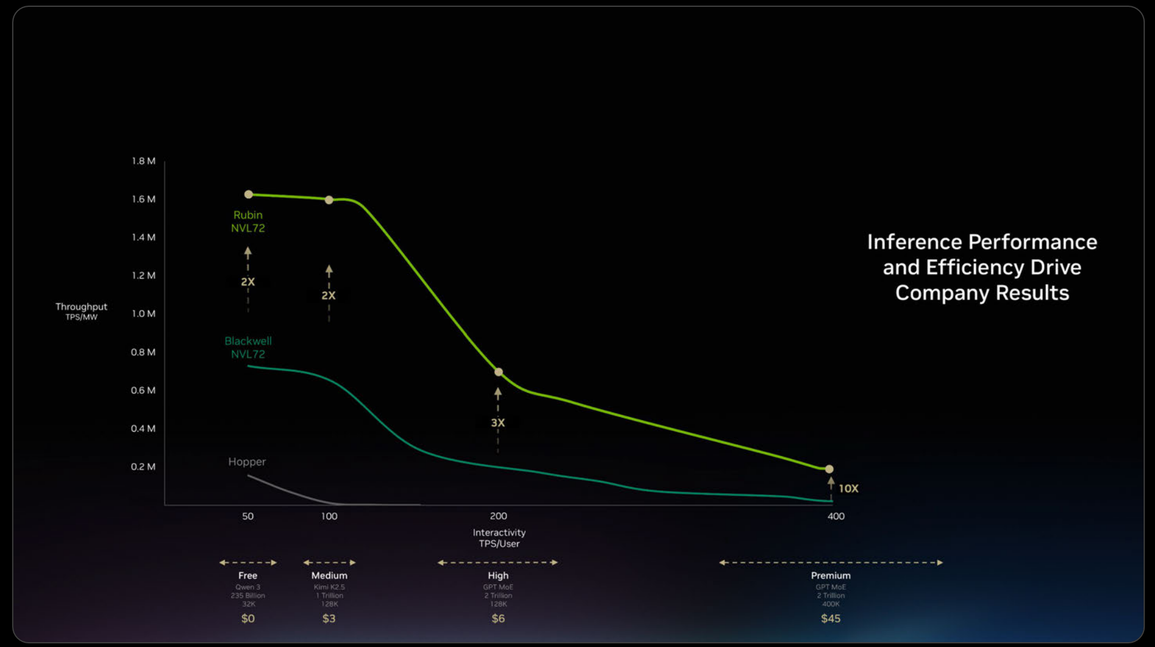
Inference Performance and Efficiency Drive Company Results


Inference Performance and Efficiency Drive Company Results
Throughput TPS/MW
1.8 M
1.6 M
1.4 M
1.2 M
1.0 M
0.8 M
0.6 M
0.4 M
0.2 M
Blackwell NVL72
35X
Hopper
50
200
Interactivity TPS/User
Free
Qwen 3 235 Billion 32K
$0
Medium
Kimi K2.5 1 Trillion 128K
$3
High
GPT MoE 2 Trillion 128K
$6

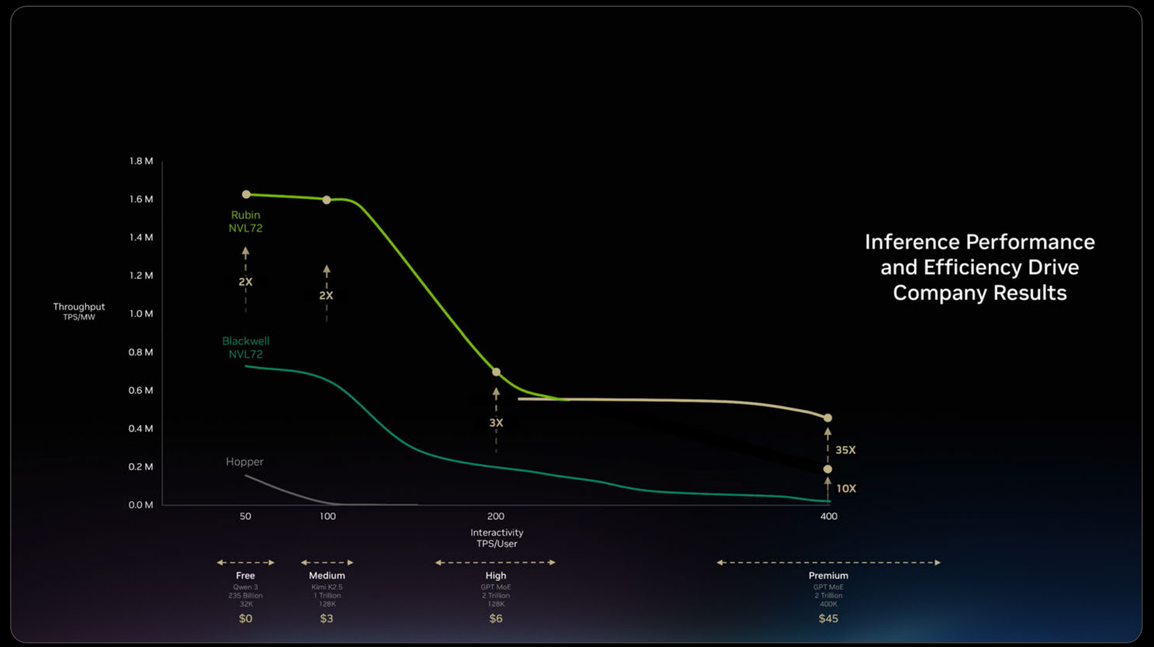

Inference Performance and Efficiency Drive Company Results
Throughput TPS/MW
1.8 M
1.6 M
1.4 M
1.2 M
1.0 M
0.8 M
0.6 M
0.4 M
0.2 M
Blackwell NVL72
35X
Hopper
50
100
200
400
Interactivity TPS/User
Free
Qwen 3 235 Billion 32K
$0
Medium
Kimi K2.5 1 Trillion 128K
$3
High
GPT MoE 2 Trillion 128K
$6
Premium
GPT MoE 2 Trillion 400K
$45


Inference Performance and Efficiency Drive Company Results
Throughput TPS/MW
1.8 M
1.6 M
1.4 M
1.2 M
1.0 M
0.8 M
0.6 M
0.4 M
0.2 M
Rubin NVL72
Blackwell NVL72
Hopper
2X
2X
3X
10X
50
100
200
400
Interactivity TPS/User
Free
Qwen 3 235 Billion 32K
$0
Medium
Kimi K2.5 1 Trillion 128K
$3
High
GPT MoE 2 Trillion 128K
$6
Premium
GPT MoE 2 Trillion 400K
$45

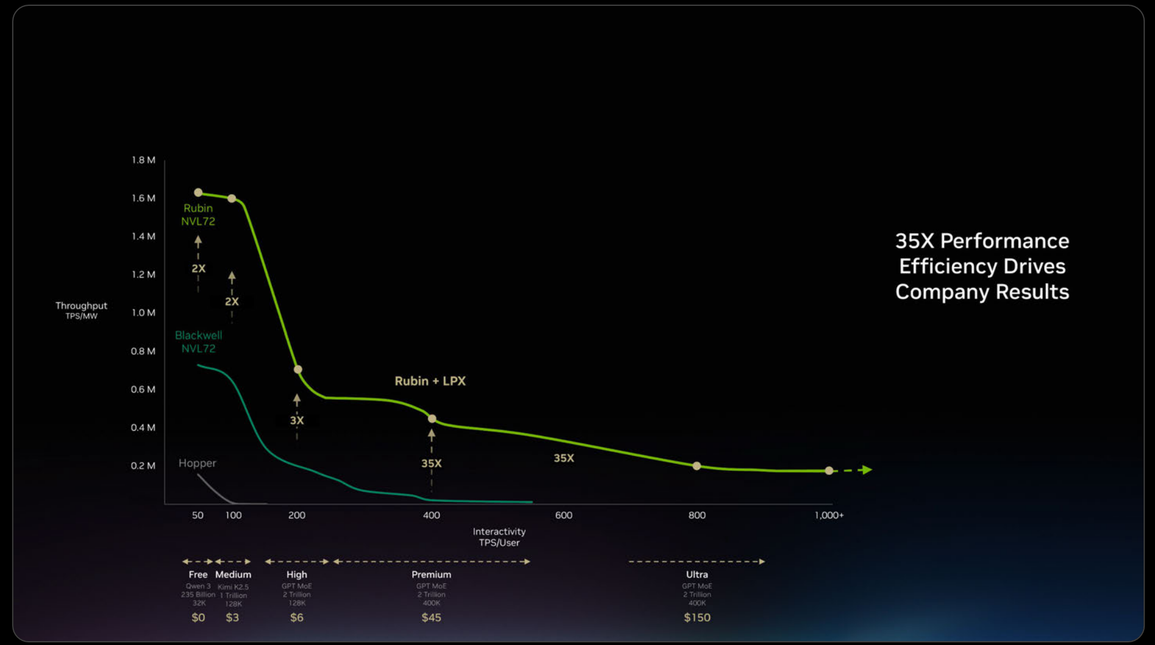

NVIDIA Vera Rubin Unlocks $150B Revenue Opportunity
Blackwell
Rubin
Annual Revenue Per Gigawatt
$150B
$100B
$50B
$0B
Free
Qwen 3 235 Billion 32K
$0
Medium
Kimi K2.5 1 Trillion 128K
$3
High
GPT MoE 2 Trillion 128K
$6
Premium
GPT MoE 2 Trillion 400K
$45
$150B
5X
$30B
Total Annual Revenue

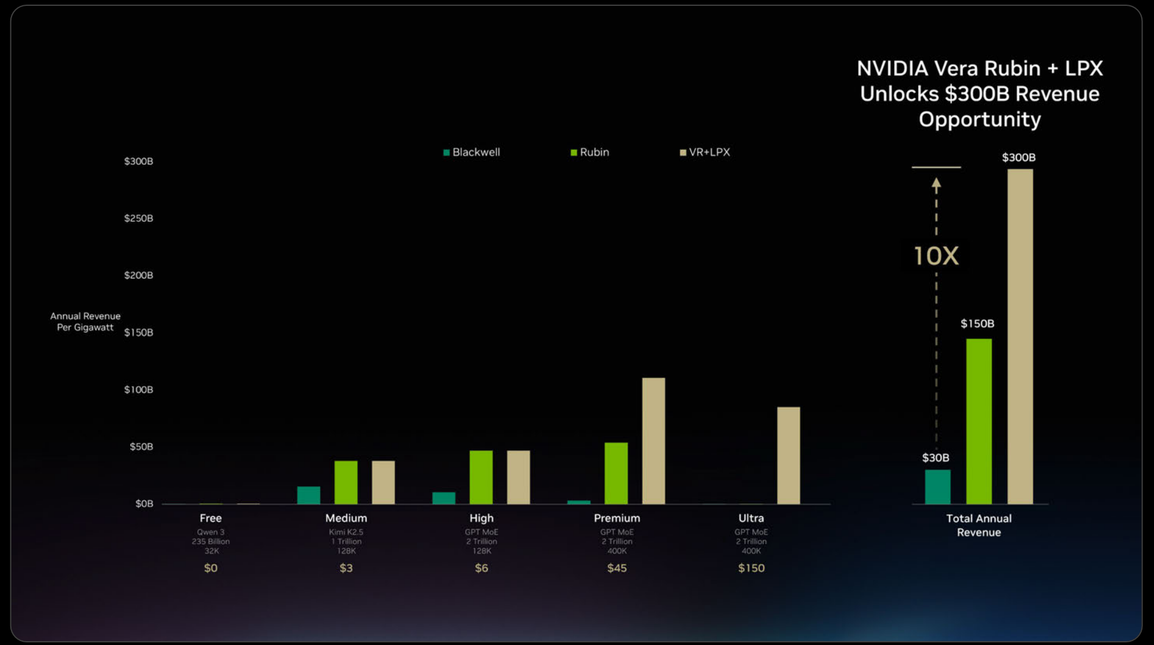


Inference Performance and Efficiency Drive Company Results

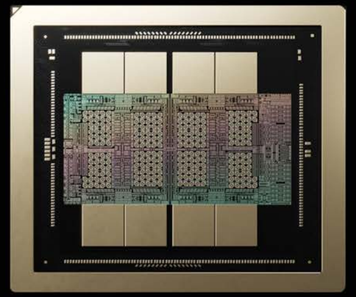


Inference Performance and Efficiency Drive Company Results

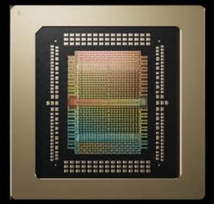

35X Performance Efficiency Drives Company Results
Throughput TPS/MW
1.8 M
1.6 M
1.4 M
1.2 M
1.0 M
0.8 M
0.6 M
0.4 M
0.2 M
Rubin NVL72
2X
2X
Blackwell NVL72
Hopper
Rubin + LPX
3X
35X
35X
50
100
200
400
600
800
1,000+
Interactivity TPS/User
Free
Qwen 3 235 Billion 32K
$0
Medium
Kimi K2.5 1 Trillion 128K
$3
High
GPT MoE 2 Trillion 128K
$6
Premium
GPT MoE 2 Trillion 400K
$45
Ultra
GPT MoE 2 Trillion 400K
$150


NVIDIA Vera Rubin + LPX Unlocks $300B Revenue Opportunity
Blackwell
Rubin
VR+LPX
Annual Revenue Per Gigawatt
$300B
$250B
$200B
$150B
$100B
$50B
$0B
Free
Qwen 3 235 Billion 32K
$0
Medium
Kimi K2.5 1 Trillion 128K
$3
High
GPT MoE 2 Trillion 128K
$6
Premium
GPT MoE 2 Trillion 400K
$45
Ultra
GPT MoE 2 Trillion 400K
$150
10X
$30B
$150B
$300B
Total Annual Revenue

Rubin GPU



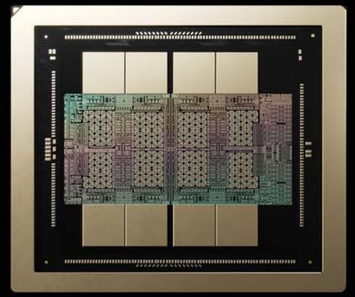
Groq 3 LPU



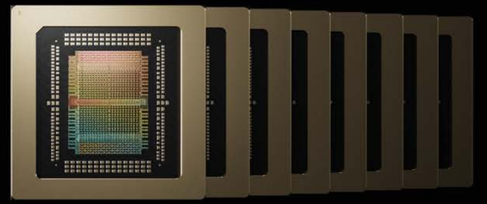
288 GB HBM4	500 MB SRAM
22 TB/s	150 TB/s SRAM Bandwidth
50 PFLOPs (NVFP4)	1.2 PFLOPs (FP8)
336B Transistors + 2.5T (HBM4)	98B Transistors

Uniting Processors of Extreme Performances

Rubin GPU

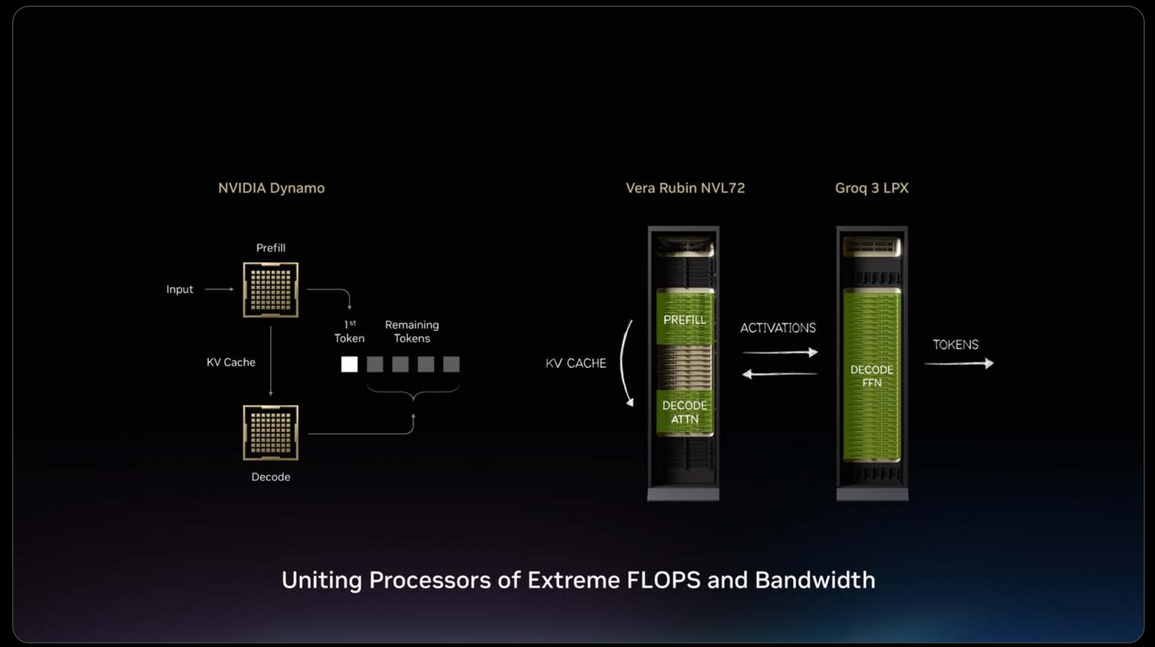


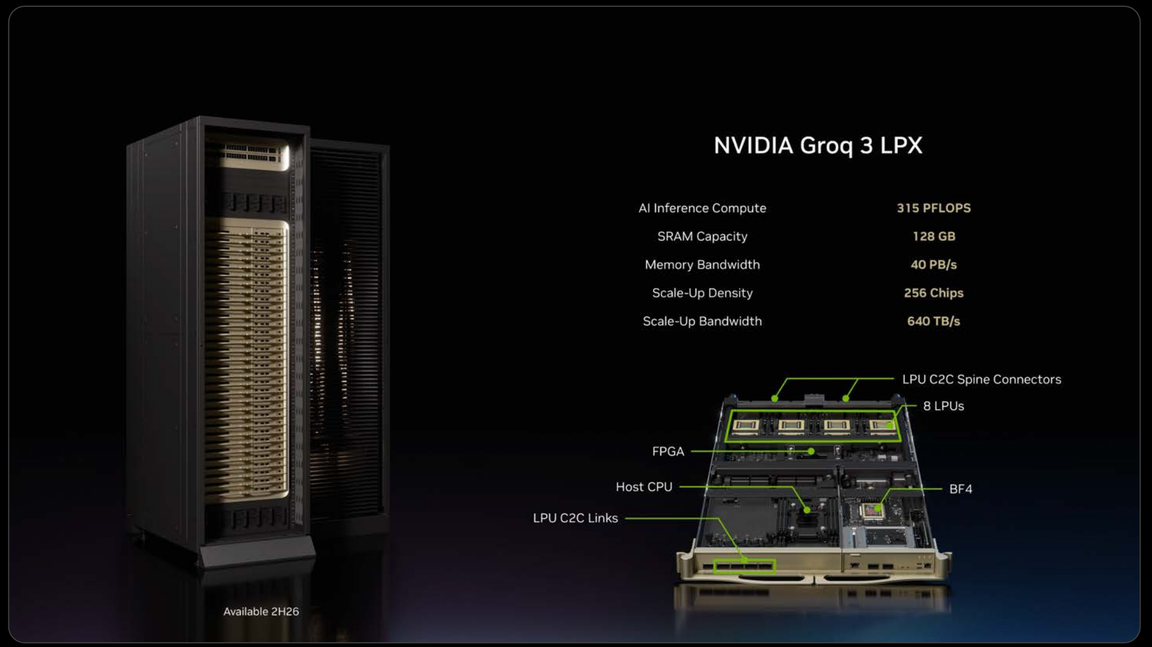
Groq 3 LPU



288 GB HBM4	4 GB SRAM
22 TB/s	1,200 TB/s SRAM Bandwidth **55X**
50 PFLOPs (NVFP4)	9.6 PFLOPs (FP8)
336B Transistors + 2.5T (HBM4)	784B Transistors

Uniting Processors of Extreme Performances


NVIDIA Dynamo
Prefill
Input
1st Token
Remaining Tokens
KV Cache
Decode
Vera Rubin NVL72
Groq 3 LPX
KV CACHE
PREFILL
DECODE ATTN
ACTIVATIONS
DECODE FFN
TOKENS
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
LPU C2C Spine Connectors
8 LPUs
FPGA
Host CPU
BF4
LPU C2C Links

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

Announcing NVIDIA Vera Rubin NVL72 Launch Partners

AI Labs

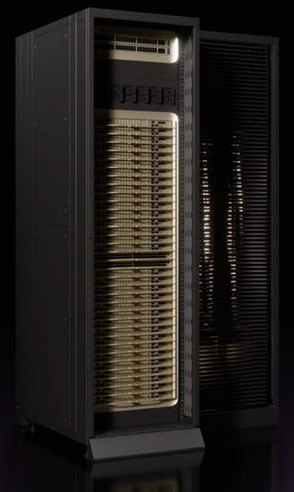
Anthropic · Cursor · Meta · OpenAI · Perplexity · Black Forest Labs · Cohere · Harvey · Mistral AI · OpenEvidence · Runway · Thinking Machines

Cloud

AWS · Google Cloud · Microsoft Azure · Oracle Cloud Infrastructure · CoreWeave · Crusoe · Lambda · Nebius · Nscale

OEMs and ODMs

Dell Technologies · HPE · Lenovo · Supermicro · Aivres · ASUS · Foxconn · Gigabyte · Inventec · Pegatron · QCT · Wistron · Wiwynn



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

Announcing NVIDIA Vera CPU Launch Partners

Cloud

Meta · Oracle Cloud Infrastructure · CoreWeave · Crusoe · Lambda · Nebius · Nscale · together.ai

OEMs and ODMs

Cisco · Dell Technologies · HPE · Lenovo · Supermicro · Aivres · ASRock Rack · ASUS · Compal · Foxconn · Gigabyte · hyve · Inventec · MiTAC · MSI · Pegatron · QCT · Wistron · Wiwynn

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

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 · Supermicro · VAST · WEKA

NVIDIA Vera Rubin

7 Chips – 5 Rack Systems
AI Factory for the Agentic AI Frontier

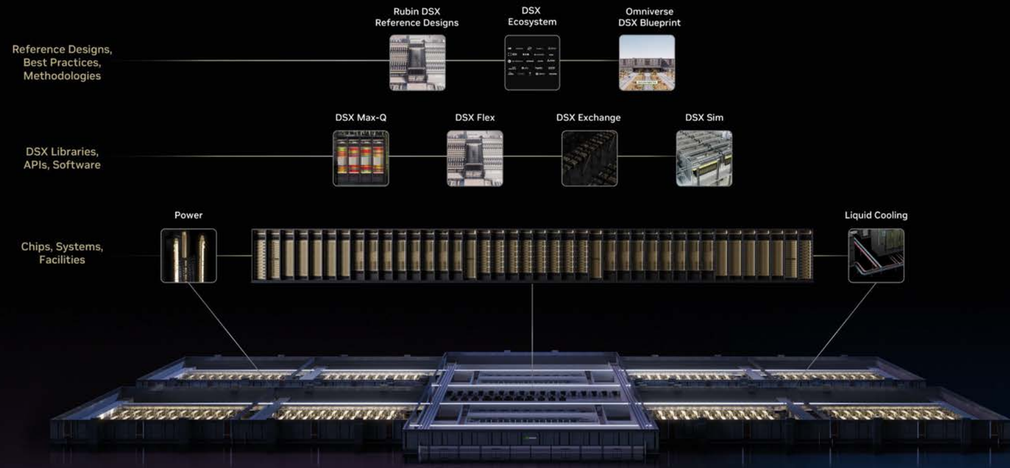
	X86 + Hopper	Vera Rubin
1 GW AI Factory		
# of GPUs	600K	300K
AI FLOPS	1.2 ZFLOPS	16 ZFLOPS
All-to-All Scale-up	7.2 TB/s	260 TB/s
Memory BW-per-Domain (GROQ SRAM)	2 EB/s	100 EB/s
Tokens per Second	2M	700M

Rubin GPU · Vera CPU · CX9 · BF4 · NVLink Switch · Spectrum CPO · Groq 3 LPU

Vera Rubin Compute Tray · NVLink Switch Tray · Vera Compute Tray · BF4 STX Server · Spectrum Switch · Groq 3 Compute Tray

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 · iVEDUS · NSCALE · switch

DSX Max-Q

CoreWeave · Crusoe · Digital Realty · emeraldai · iVEDUS · NSCALE · Red Hat · run:ai · phaidra · switch

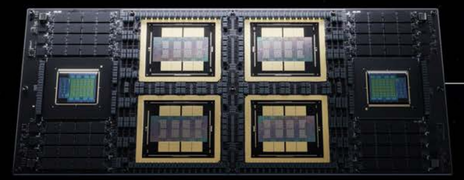
DSX Sim

ABB · ARMADA · AVEVA · GXO · Hexagon · network · cadence · CHECK POINT · CLOUDIAN · DATADOG SYNTHETICS · D&LLTechnologies · E·T·N · etap · GE VERNOVA · hedgehog · HITACHI · Jacobs · KEYSIGHT · metalsoft · MIRANTIS · Mitsubishi Electric · nvtris · NOKIA · Open Nebula · paloalto · ptc · RAFAY · Red Hat · run:ai · Schneider Electric · SIEMENS · SIEMENS Digital Industries Software · slurm · Spectro Cloud · TRANE TECHNOLOGIES · TrendAI · VAST · vCluster · VERTIV · WEKA

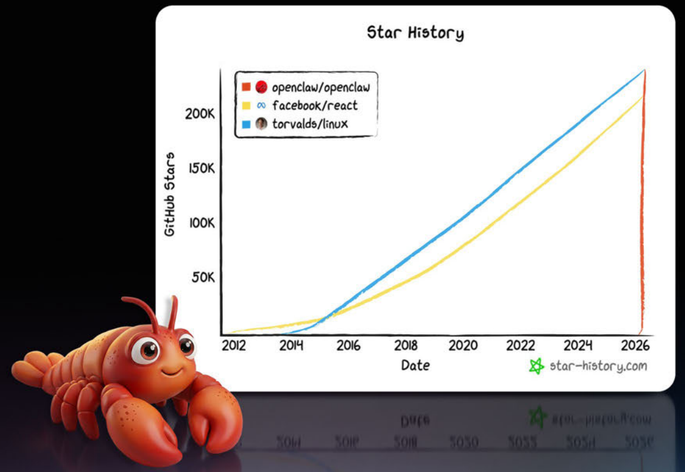
DSX Exchange

ABB · ARMADA · AVEVA · GXO · network · cadence · CHECK POINT · CLOUDIAN · CoreWeave · Crusoe · Digital Realty · D&LLTechnologies · E·T·N · emeraldai · GE VERNOVA · hedgehog · HITACHI · KEYSIGHT · metalsoft · MIRANTIS · Mitsubishi Electric · iVEDUS · nvtris · NOKIA · NSCALE · Open Nebula · paloalto · phaidra · RAFAY · Red Hat · run:ai · Schneider Electric · SIEMENS · SIEMENS Digital Industries Software · slurm · Spectro Cloud · switch · TRANE TECHNOLOGIES · TrendAI · VAST · vCluster · VERTIV · WEKA

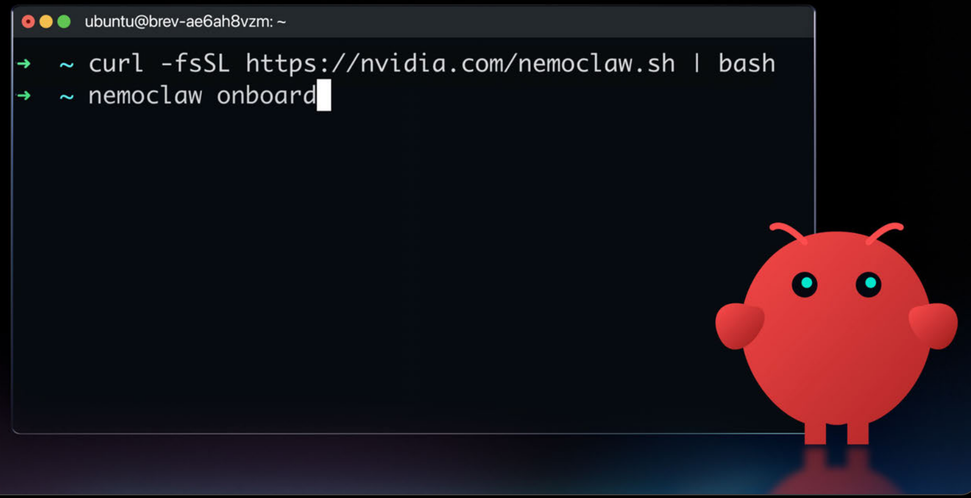
Announcing NVIDIA Space-1 Vera Rubin Module



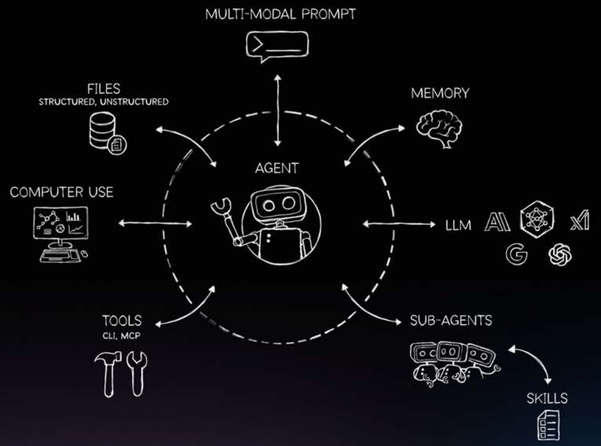
Inference Inflection Arrives



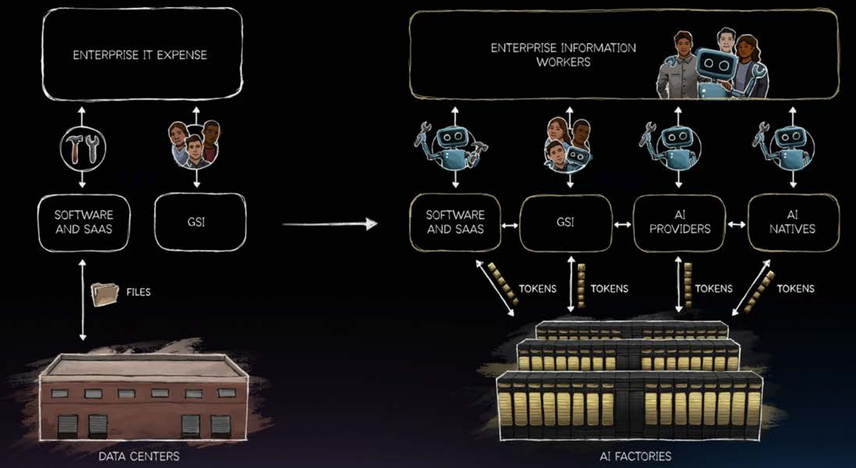
Announcing NVIDIA NemoClaw for OpenClaw

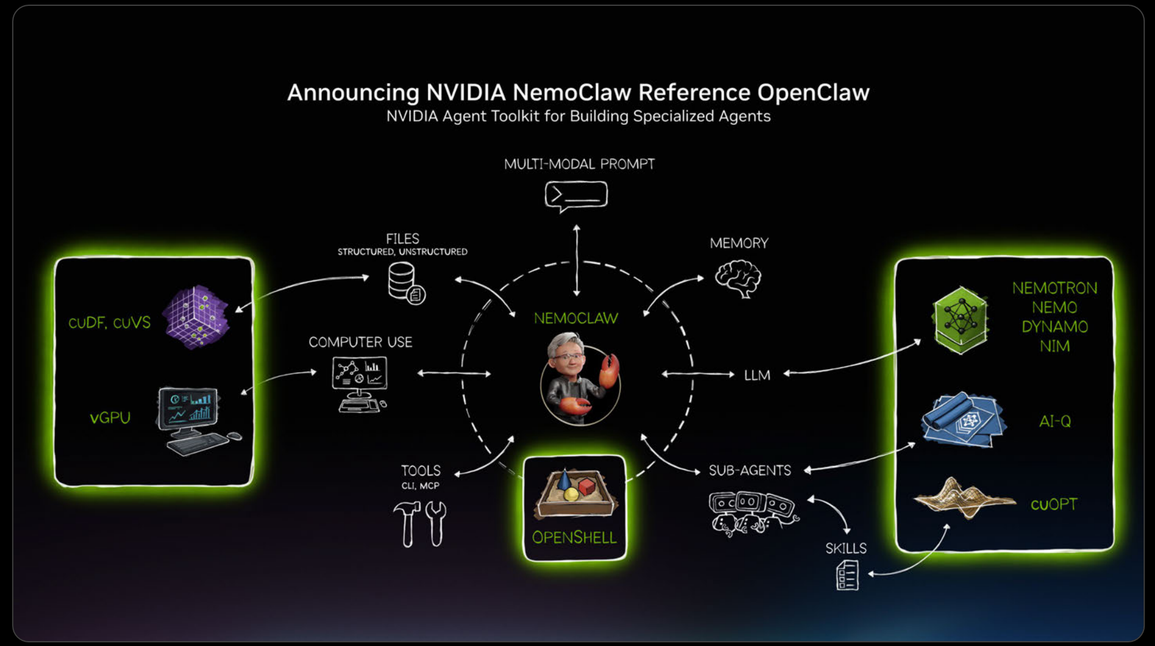


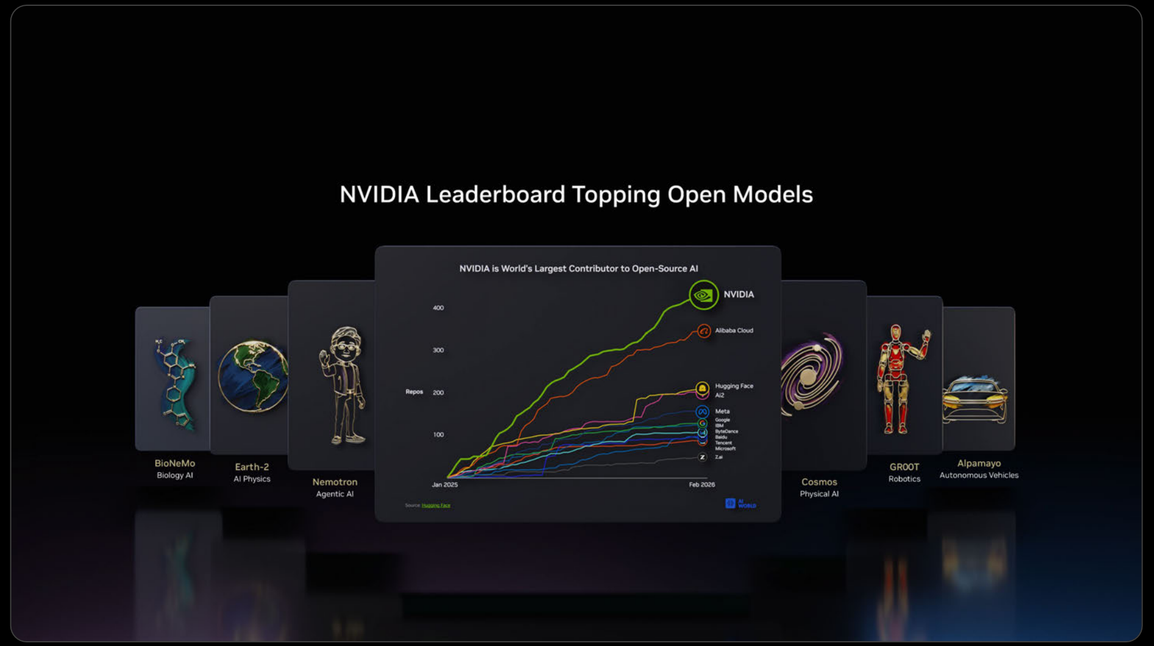
Agents – A New Computing Platform



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




Announcing NVIDIA NemoClaw Reference OpenClaw
NVIDIA Agent Toolkit for Building Specialized Agents
MULTI-MODAL PROMPT
FILES
STRUCTURED, UNSTRUCTURED
MEMORY
CUDF, CUVS
vGPU
COMPUTER USE
NEMOCLAW
LLM
NEMOTRON
NEMO
DYNAMO
NIM
AI-Q
CUOPT
TOOLS
CLI, MCP
OPENSHELL
SUB-AGENTS
SKILLS

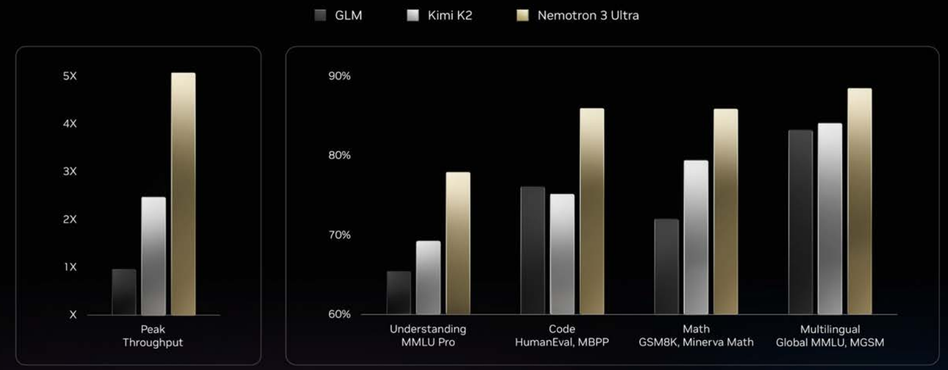

NVIDIA Leaderboard Topping Open Models
NVIDIA is World's Largest Contributor to Open-Source AI
NVIDIA
Alibaba Cloud
Hugging Face
AI2
Meta
Google
IBM
ByteDance
Baidu
Tencent
Microsoft
Z.ai
Repos
400
300
200
100
Jan 2025
Feb 2026
BioNeMo
Biology AI
Earth-2
AI Physics
Nemotron
Agentic AI
Cosmos
Physical AI
GROOT
Robotics
Alpamayo
Autonomous Vehicles

NVIDIA Nemotron 3 Super – Best Open Model for OpenClaw

anthropic/claude-sonnet-4.6	86.9%
anthropic/claude-opus-4.6	86.3%
openai/gpt-5.4	86.0%
nvidia/nemotron-3-super-120b-a12b	85.6%
anthropic/claude-opus-4.5	85.4%
moonshotai/kimi-k2.5	84.8%
qwen/qwen3.5-122b-a10b	84.5%
qwen/qwen3.5-plus-02-15	84.1%
z-ai/glm-5	84.1%
anthropic/claude-sonnet-4.5	83.1%
minimax/minimax-m2.1	82.2%
deepseek/deepseek-v3.2	81.9%
qwen/qwen3.5-397b-a17b	81.7%

NVIDIA Nemotron 3 Ultra – Best Open Base Model

5X Efficiency and Highest Reasoning Accuracy on NVIDIA GB200 NVL72



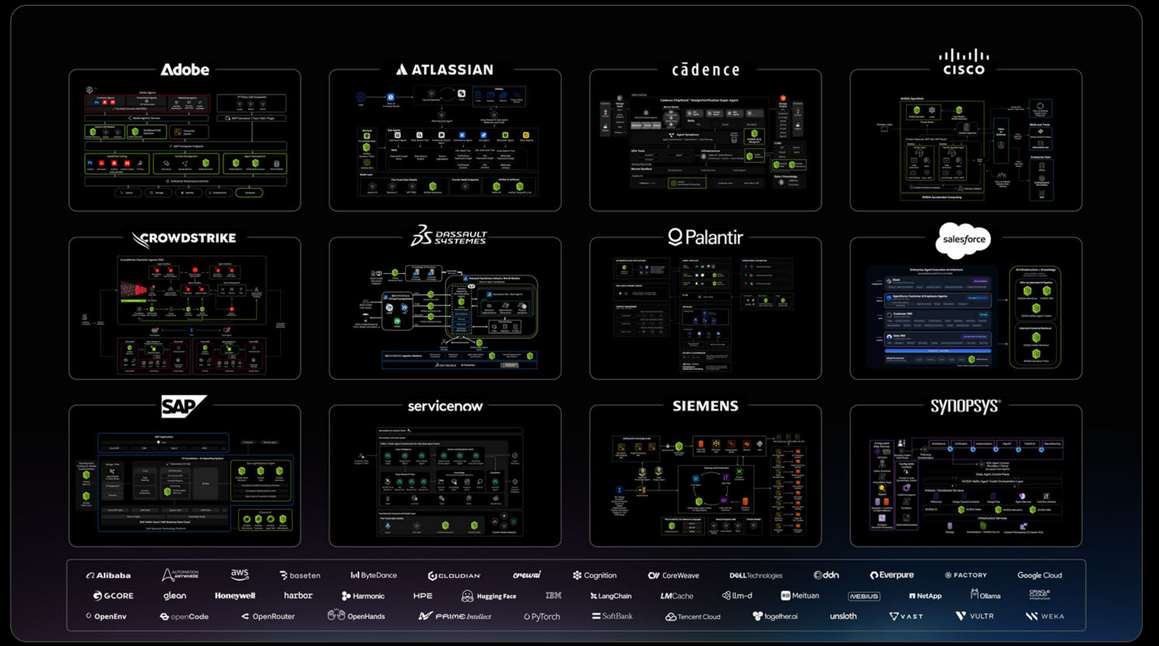
The World Building Regional AI With NVIDIA Nemotron

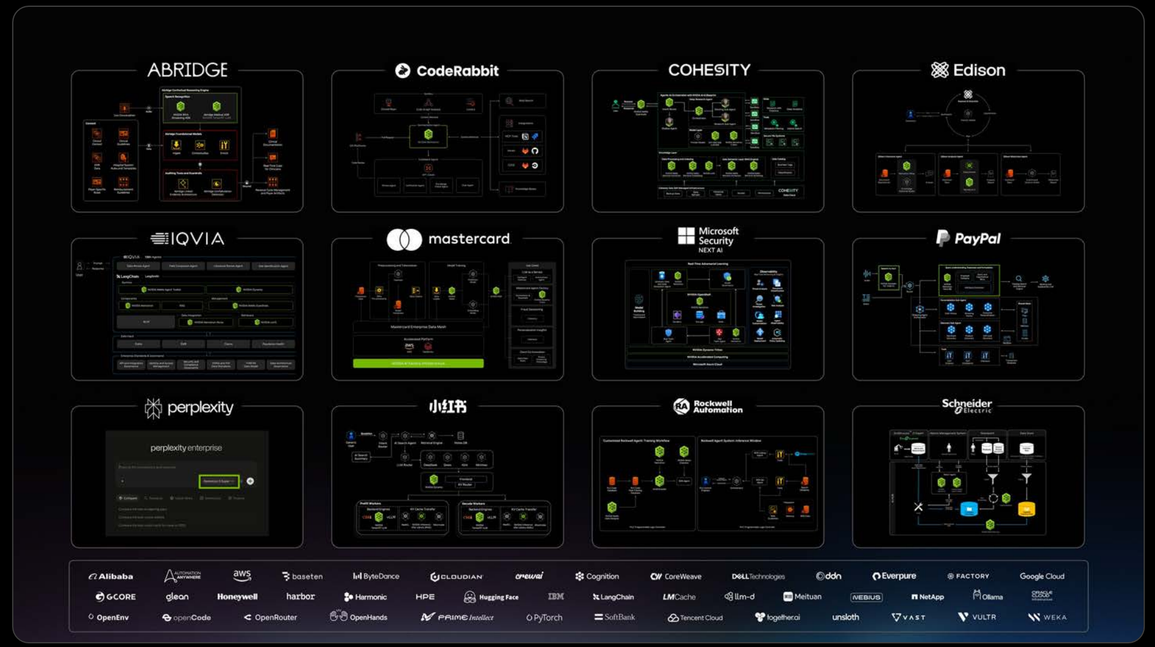
Announcing Global AI Leaders Join NVIDIA Nemotron Coalition
to Advance Open Frontier Models

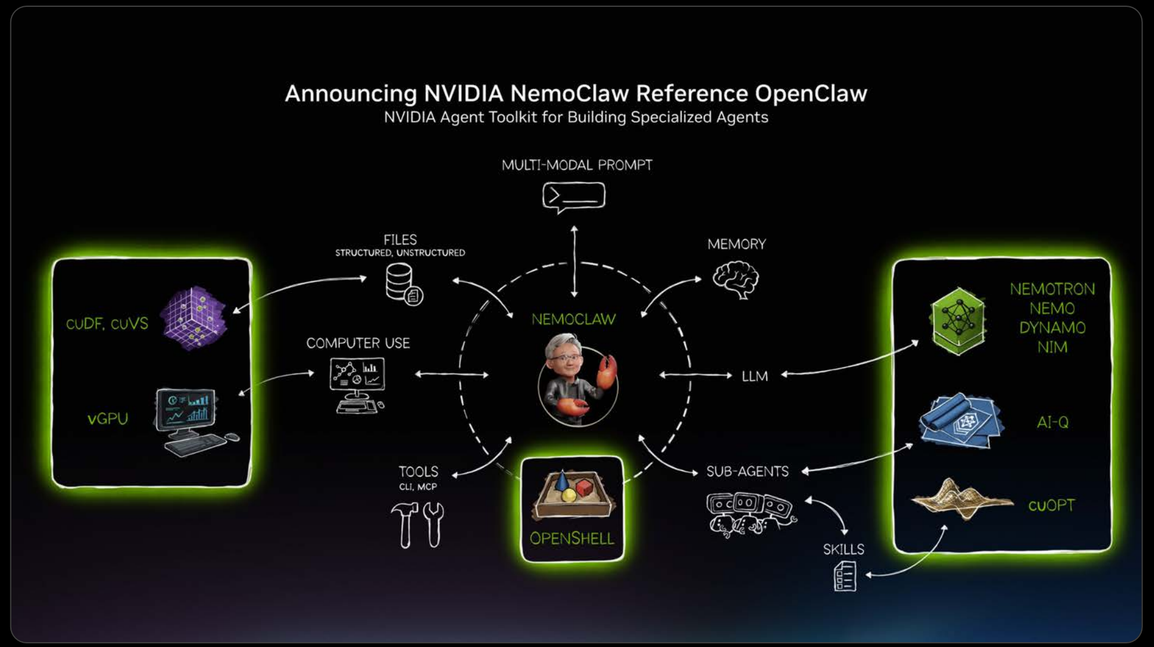
NVIDIA

Black Forest Labs · CURSOR · LangChain · Mistral AI

perplexity · Reflection · sarvam · THINKING MACHINES




ABRIDGE
CodeRabbit
COHESITY
Edison
IQVIA
mastercard
Microsoft Security
NEXT AI
PayPal
perplexity
perplexity enterprise
小红书
Rockwell Automation
Schneider Electric
Alibaba
aws
baseten
ByteDance
CLOUDIAN
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
OpenEnv
openCode
OpenRouter
OpenHands
PRIME Intellect
PyTorch
SoftBank
Tencent Cloud
together.ai
unsloth
VAST
VULTR
WEKA


Announcing NVIDIA NemoClaw Reference OpenClaw
NVIDIA Agent Toolkit for Building Specialized Agents
MULTI-MODAL PROMPT
FILES
STRUCTURED, UNSTRUCTURED
MEMORY
CUDF, CUVS
vGPU
COMPUTER USE
NEMOCLAW
LLM
NEMOTRON
NEMO
DYNAMO
NIM
AI-Q
CUOPT
TOOLS
CLI, MCP
OPENSHELL
SUB-AGENTS
SKILLS


Caterpillar
Mentee
Persona AI
NEURA Robotics
Franka Robotics
Agile Robots
AgiBot
RobotERA
T-Mobile
Zipline
NVIDIA
ABB Robotics
Boston Dynamics
Figure
Foxconn Fii
Agility
1X
LG Electronics
XPENG
Fauna Robotics
Hexagon Robotics
AgiBot
Linde
Universal Robots
Humanoid
Diligent
Serve Robotics
Disney Research
Pollen Robotics
Idealworks
Enchanted Tools
Mercedes-Benz


NVIDIA
GTC