NVIDIA’s Vision to Transform Energy

Robert Searles | Senior Solutions Architect, NVIDIA

April 22, 2025

NVIDIA

Pioneering Accelerated Computing and AI

Full stack – hardware, networking, software

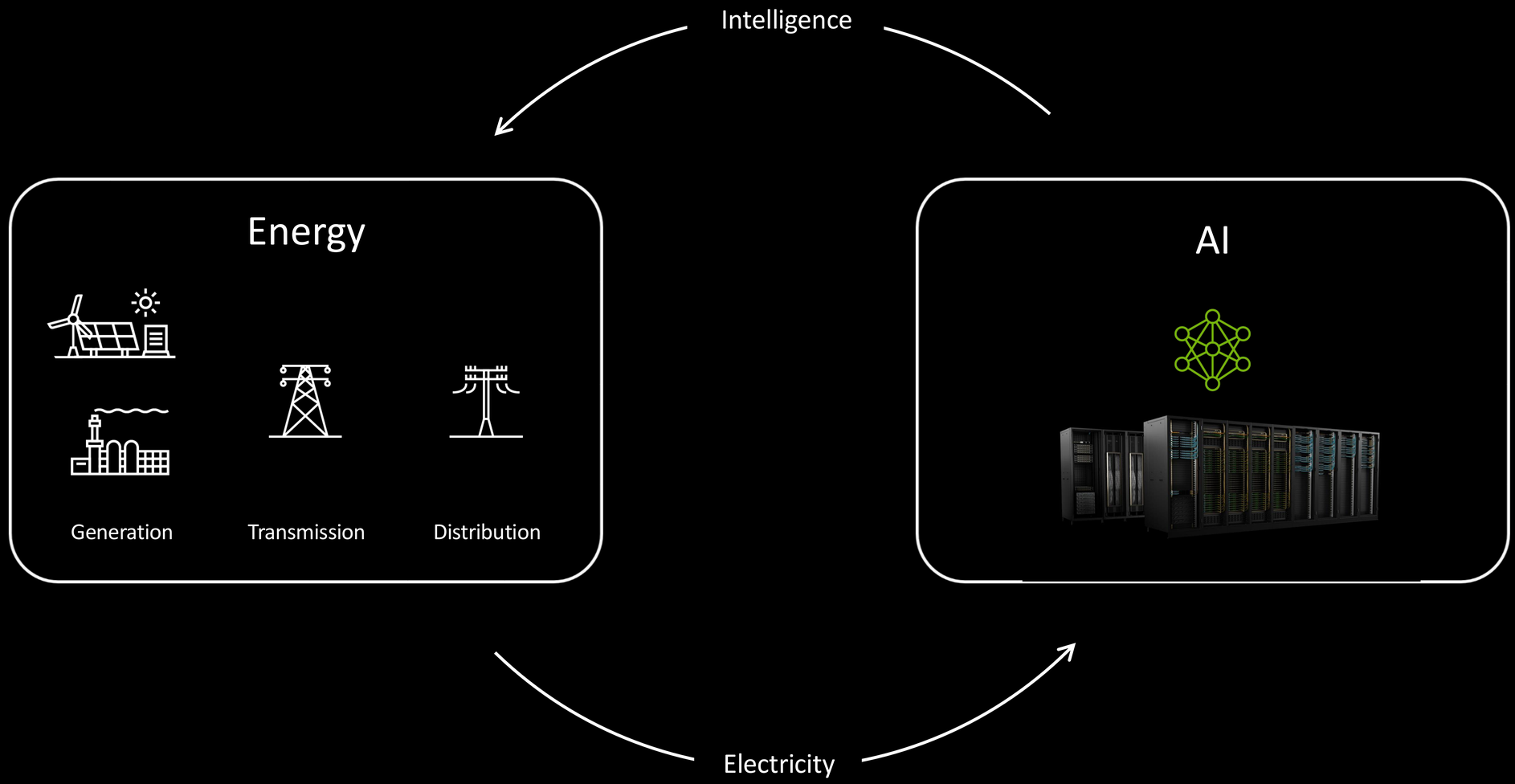
From cloud to data center to edge

40,000+ organizations run NVIDIA AI Platforms

6M developers, 27K startups in ecosystem

Powering the Future – For AI and With AI

The massive opportunity of our industries together



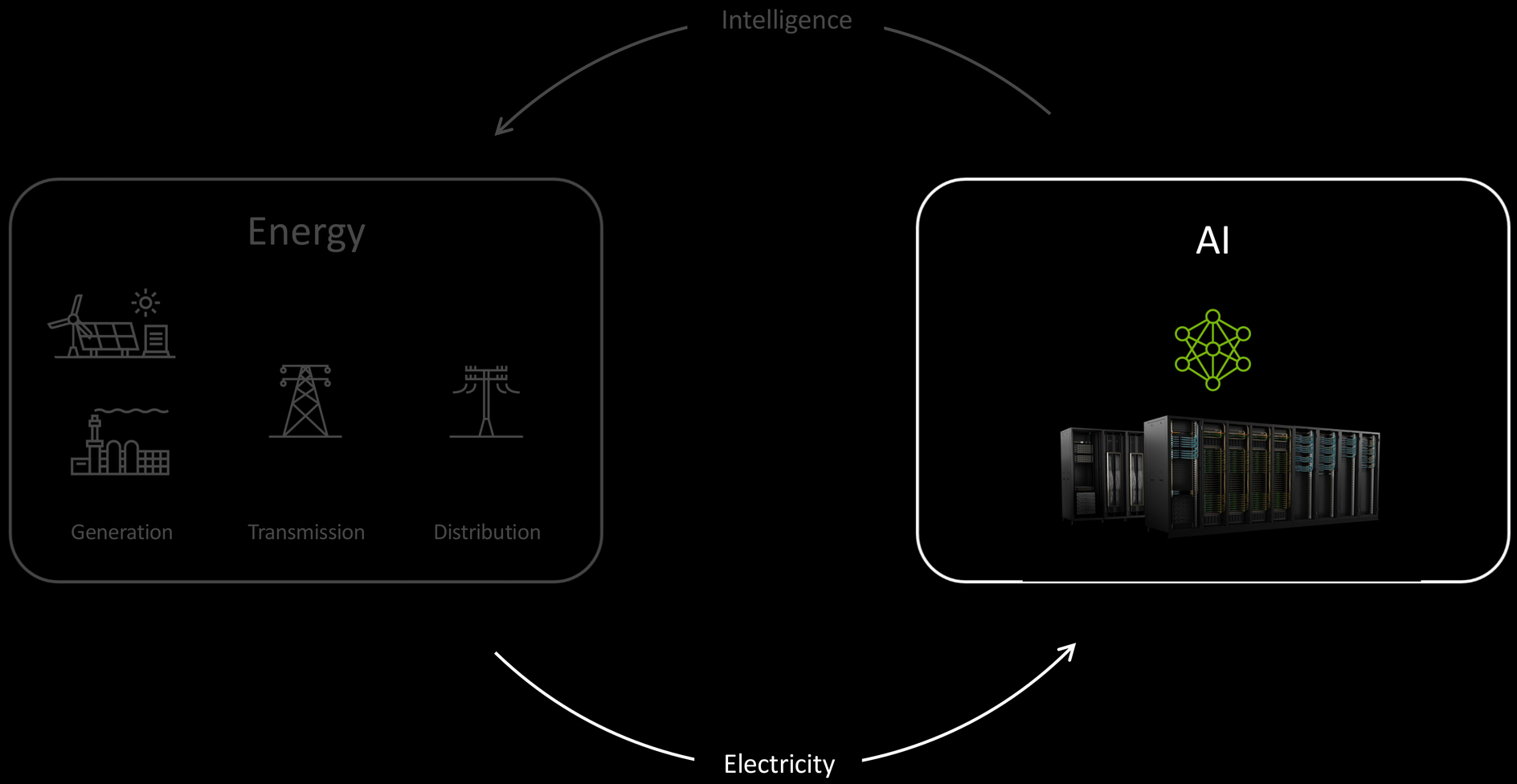

THE WALL STREET JOURNAL.
‘Three New York Cities’ Worth of Power: AI Is Stressing the Grid
Across the nation, utilities are worried about expanding the overburdened power grid, citing high costs and concerns about commitment from data center projects


AP
CLIMATE
Los Angeles fires have scorched largest urban area in California in at least 40 years


The New York Times
A.I.’s Insatiable Appetite for Energy
The soaring electricity demands of data centers and A.I. are straining the grid in some areas, pushing up emissions and slowing the energy transition.


NBC NEWS
WEATHER
Millions in Texas without power as deadly storm brings snow, freezing weather

Energy for AI



The AI Revolution is Here and Accelerating

Perception AI

Speech Recognition
Deep RECSYS
Medical Imaging

Generative AI

Digital Marketing
Content Creation

Agentic AI

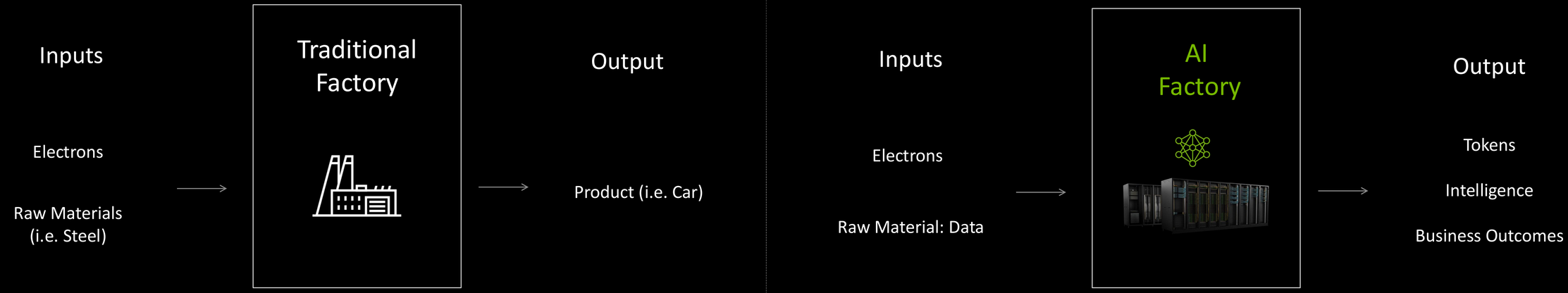
Coding Assistant
Customer Service

Physical AI

Self-Driving Cars
General Robotics

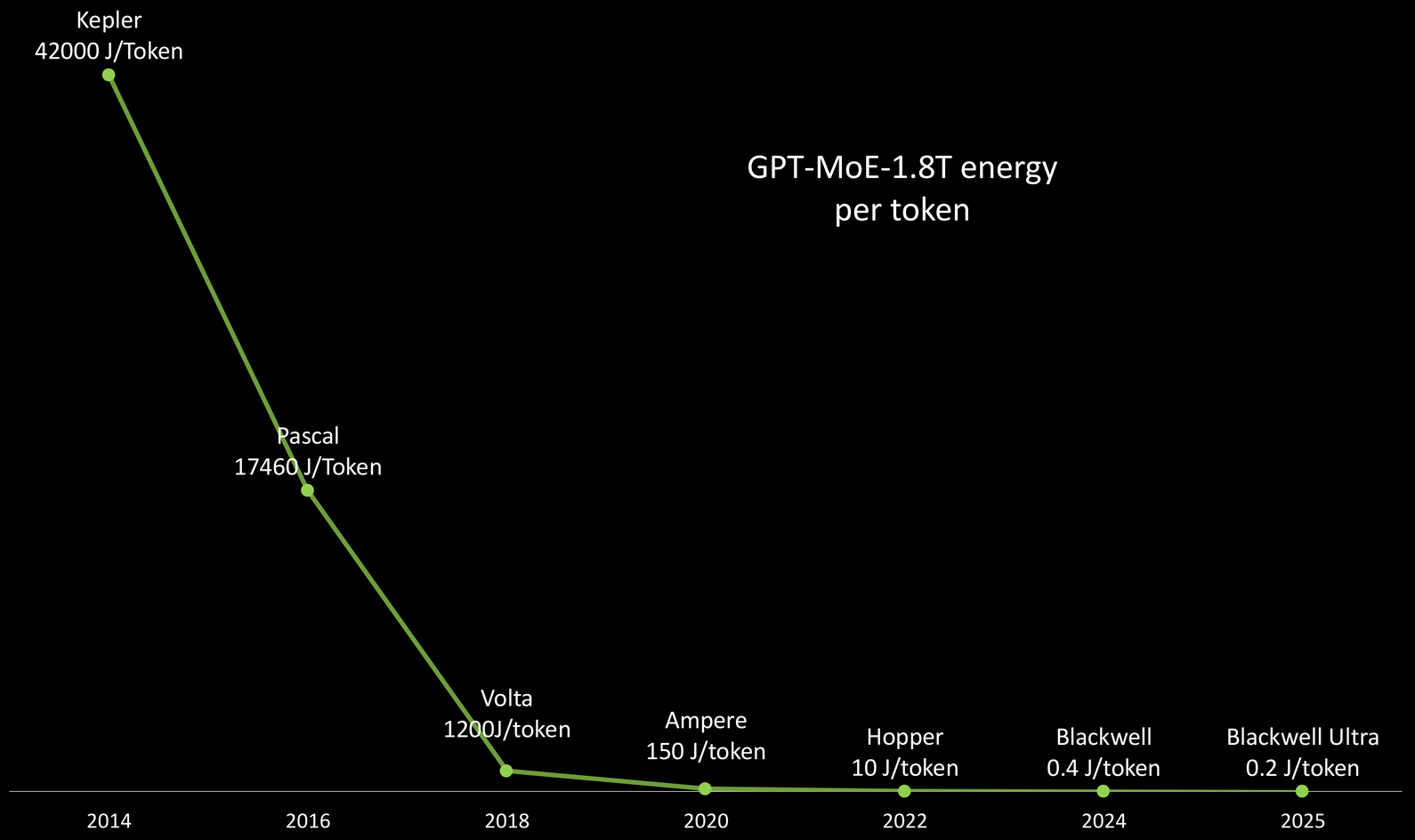
AI Factories Manufacture Intelligence at Scale

A new type of factory producing digital tokens instead of physical products



LLM Inference Continues to Get More Energy Efficient

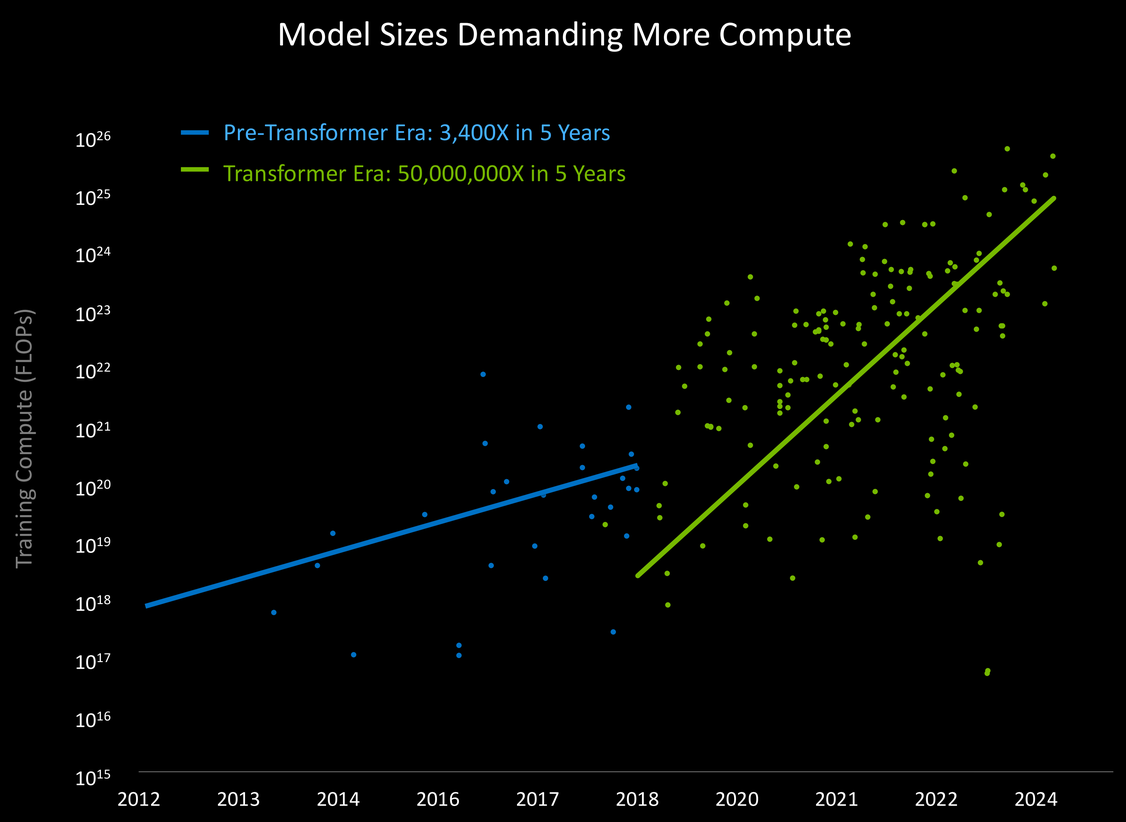
Energy required for tokens drops 200,000X in just over ten years



GPT-MoE-1.8T energy per token

NVIDIA

Growing Energy Usage in Data Centers

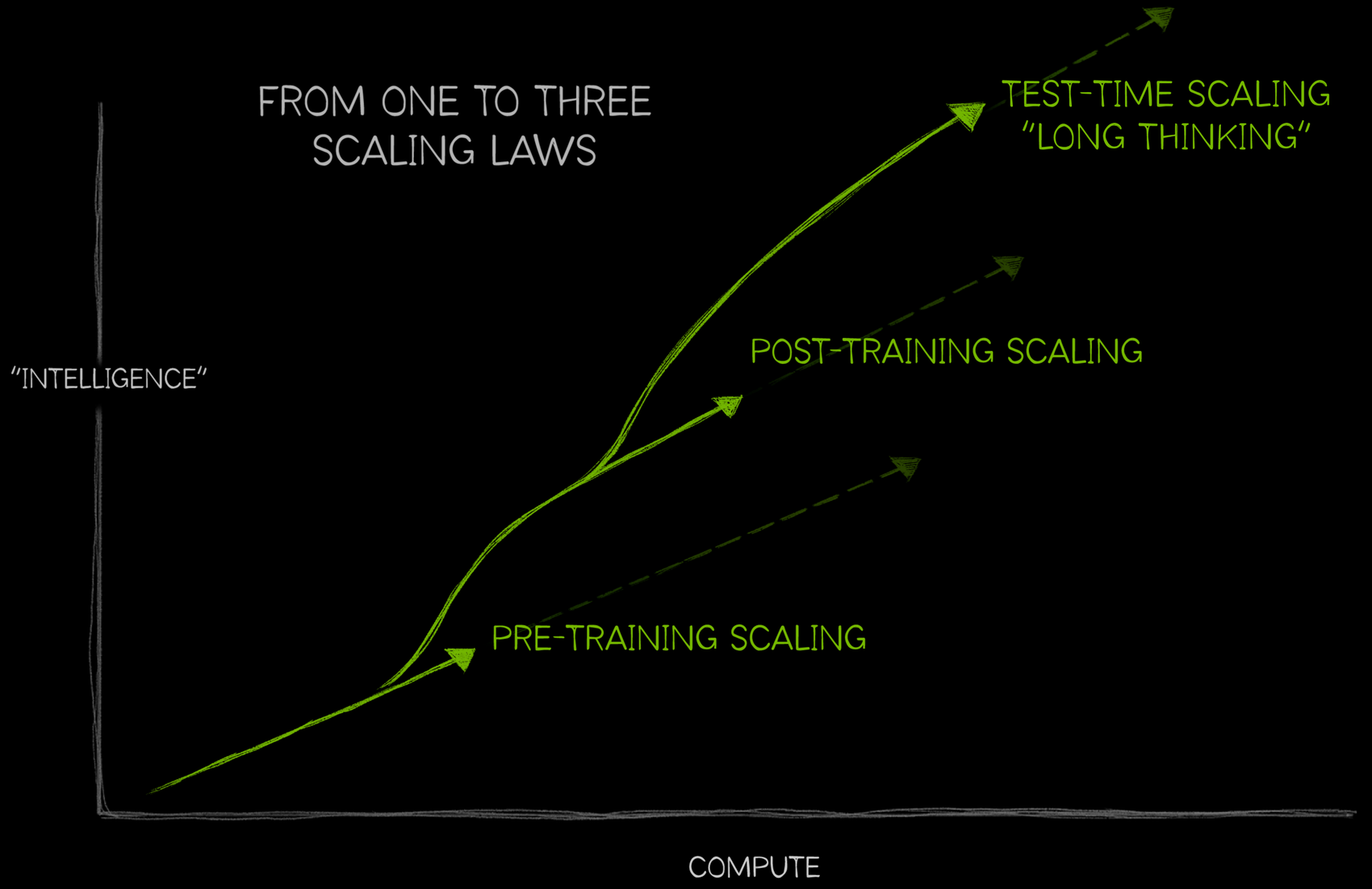


Data Center Electricity Usage	>300 TWh/year
Share of Global Energy Usage	2%
Forecasted Share of Energy Usage	5% by 2030

Source: Model sizes: Github AI Memory Wall, Arxiv papers | Data Center Electricity Usage: IEA Report

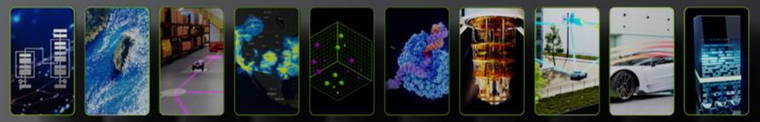
AI Scaling Laws Drive Exponential Demand for Compute

New “long thinking” supercharges inference scaling



AI Accelerated Computing Infrastructure Platform

CUDA-X LIBRARIES



ACCELERATED SOFTWARE STACK

CUDA • DOCA • NCCL

Cluster-Scale Software

System Software

Chip Software

NVIDIA AI ENTERPRISE
AGENTIC AI LIBRARIES

NIM
CUDA-ACCELERATED
AI-IN-A-CONTAINER

NVIDIA OMNIVERSE
PHYSICAL AI LIBRARIES

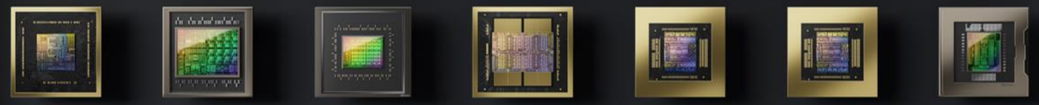
GB200 NVL72 SUPERPOD

NVLINK SPINE

NVLINK SWITCH

BLACKWELL
HGX-NODE

GRACE BLACKWELL
MGX-NODE



CHIPS PURPOSE-BUILT FOR AI SUPERCOMPUTING
GPU | CPU DPU NIC | NVLINK SWITCH | IB SWITCH | ENET SWITCH

QUANTUM & SPECTRUM
SWITCH

NVIDIA

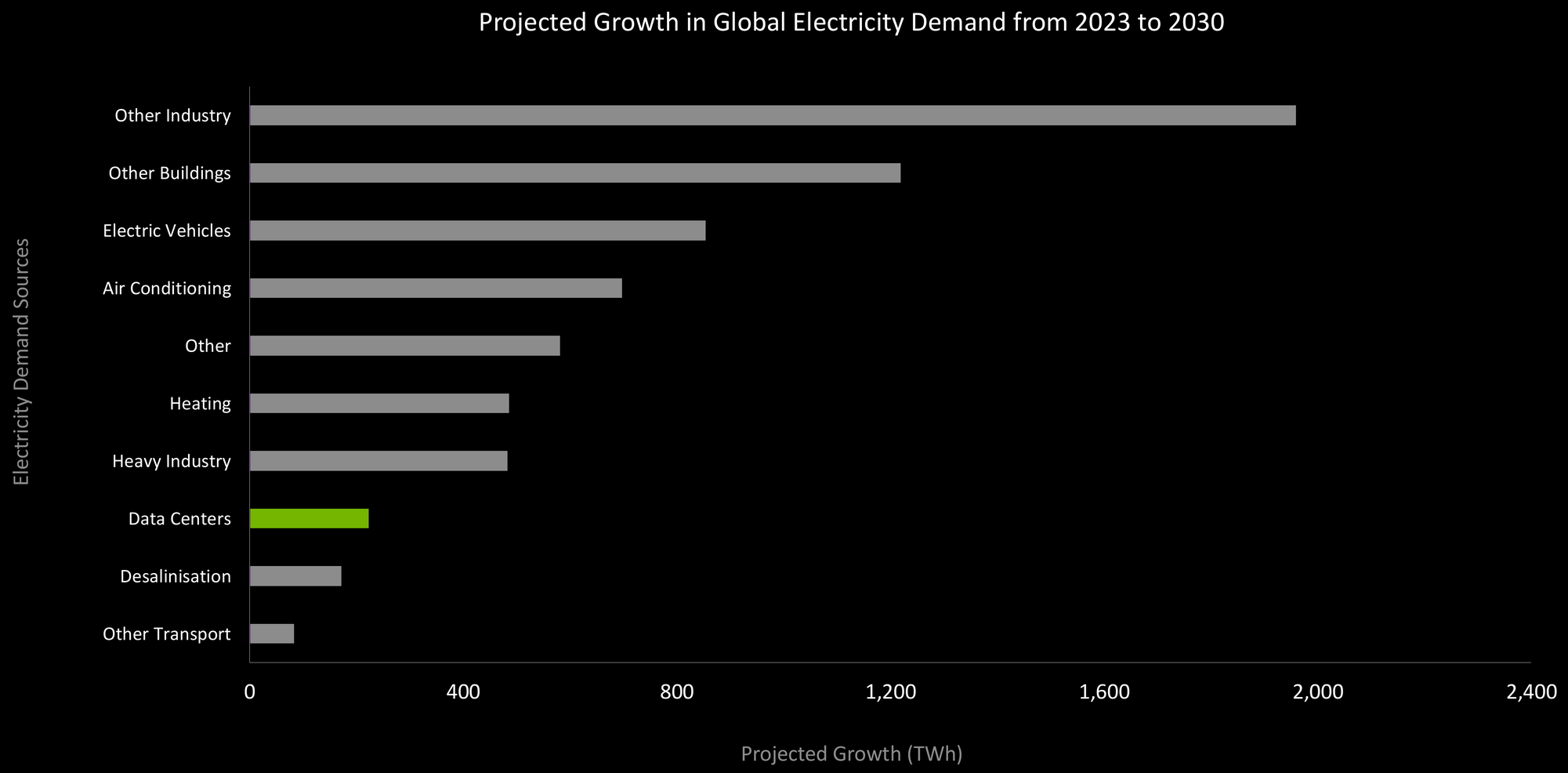
Announcing

New Omniverse Blueprint Advances AI Factory Design and Simulation

NVIDIA


NVIDIA
NVIDIA

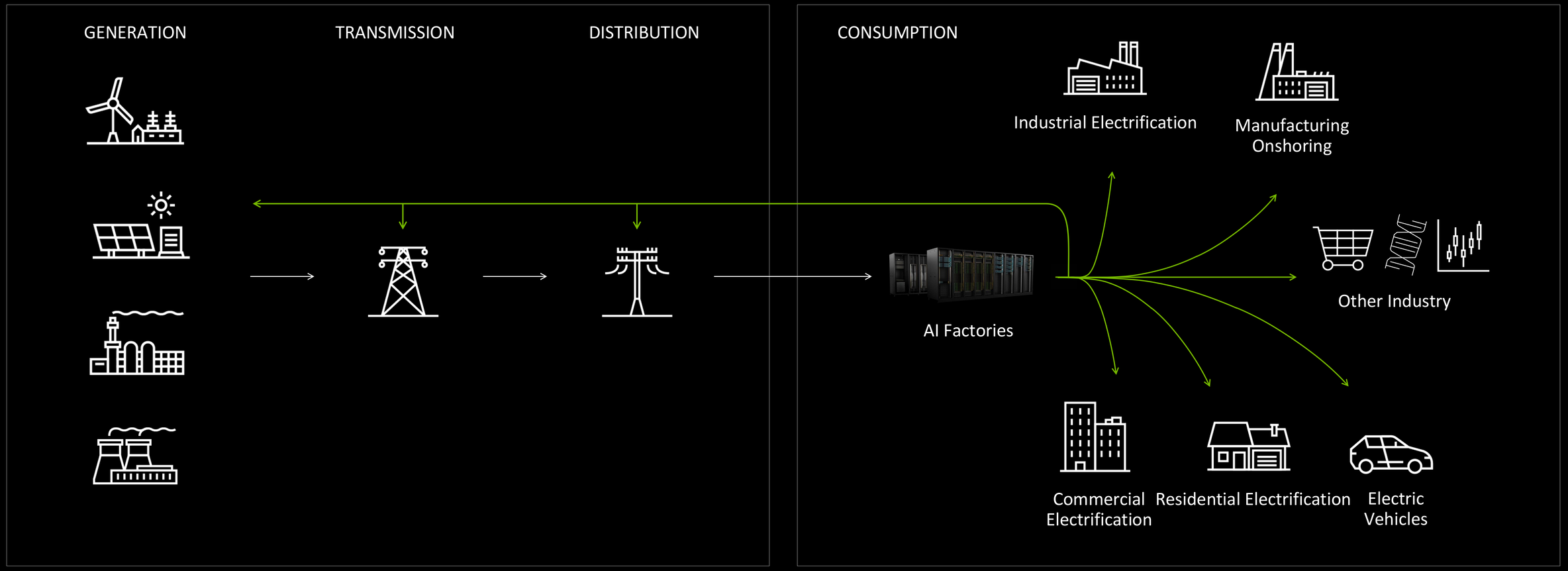
Rising Energy Demand Goes Beyond Data Centers and AI



Source: International Energy Agency (IEA). World Energy Outlook 2024.

AI Can Help Solve Longstanding and Emerging Energy Challenges

Optimize energy production, delivery, and consumption to meet new demand and reduce waste



Energy Production and Delivery

Energy Demand

AI for Energy

Intelligence

Energy

Generation

Transmission

Distribution

AI

Electricity

How AI is Being Leveraged Across Power and Utilities

Becoming a software defined energy provider

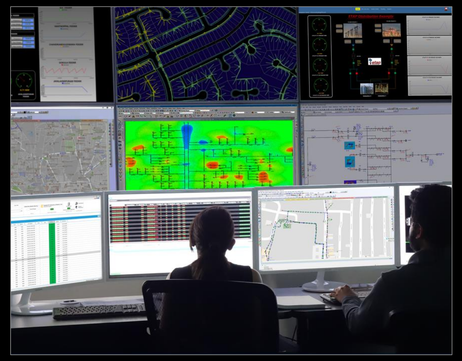
LOWER COSTS

MAXIMIZE RENEWABLE INTEGRATION

IMPROVE RELIABILITY & RESILIENCY

Renewable Generation Prediction

Predictive Maintenance & Asset Inspection



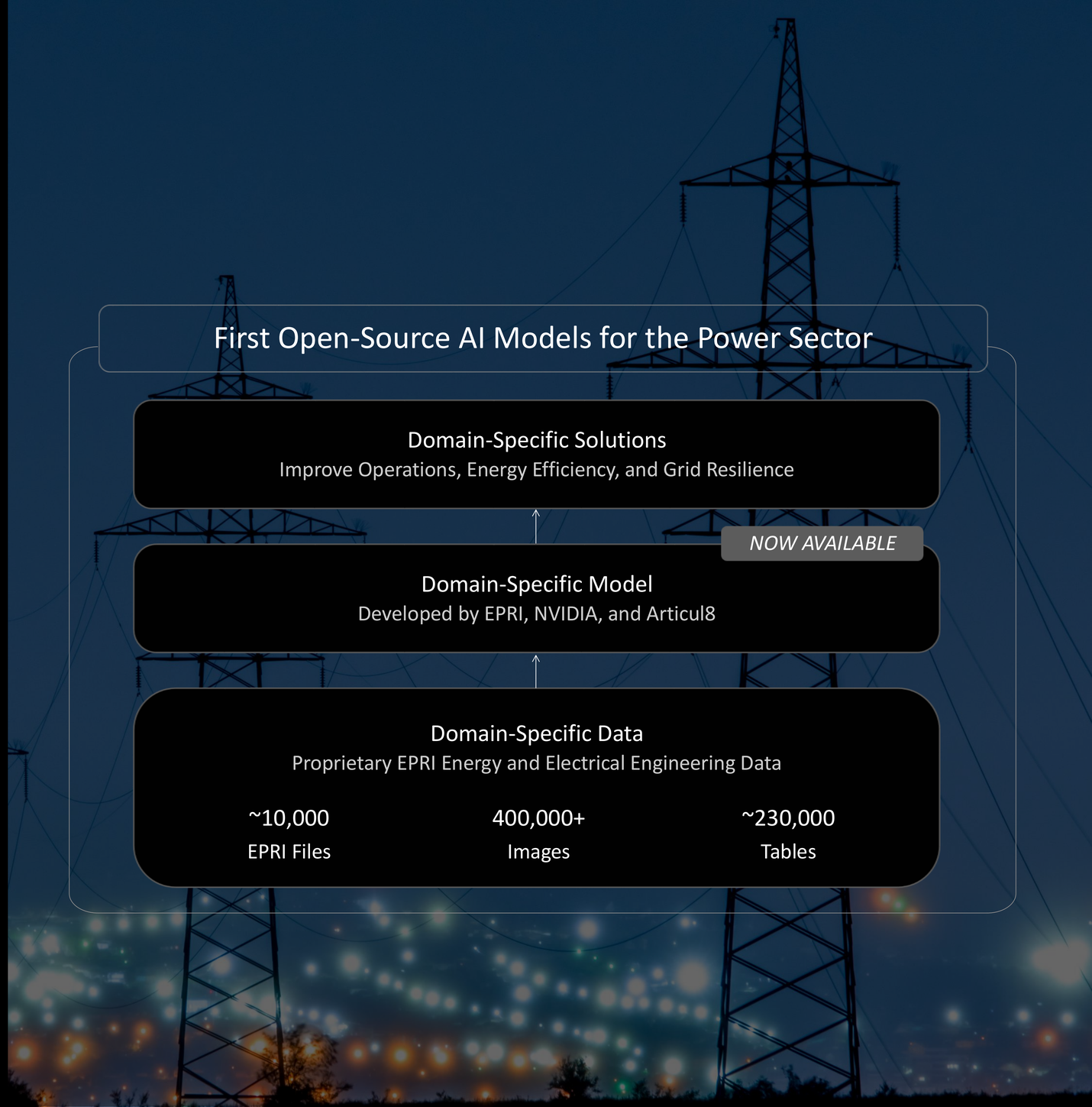
Load & Demand Forecasting

Real-Time Grid Optimization

Cybersecurity & OT Enhancements

Announcing

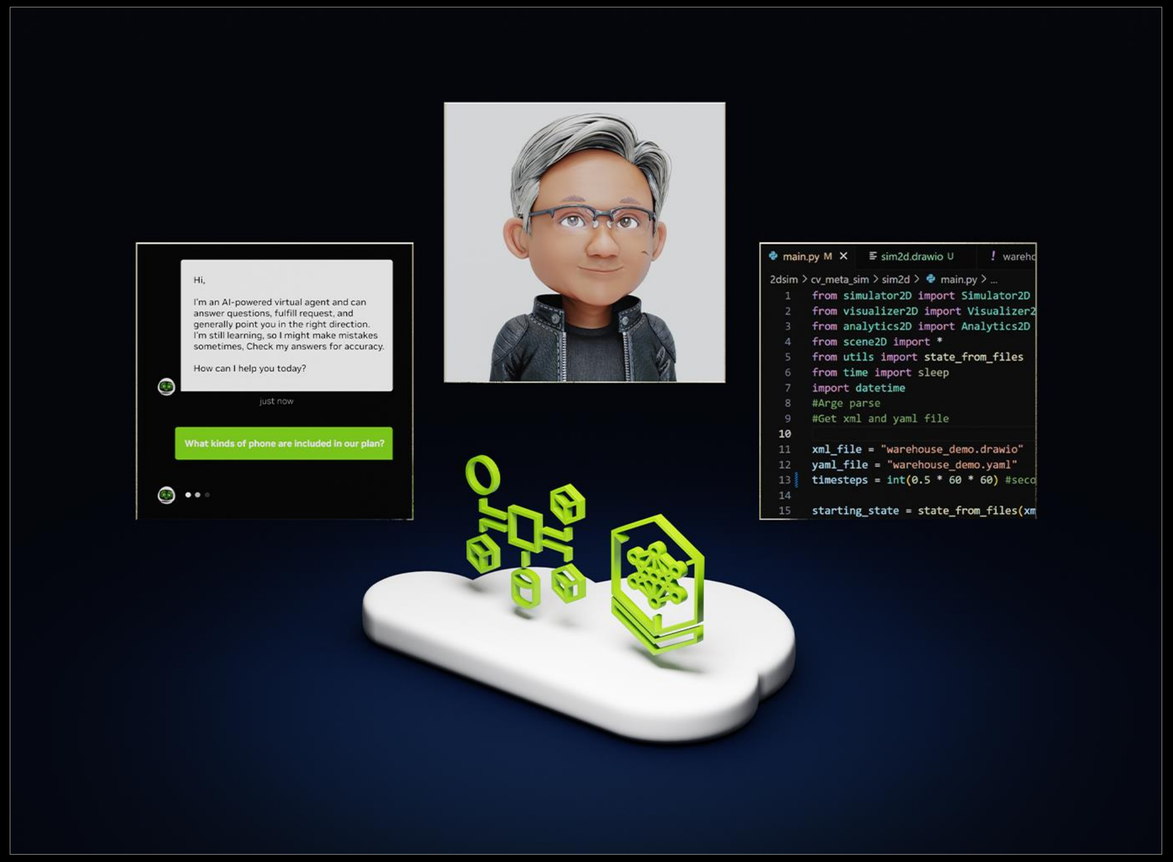
EPRI, NVIDIA, and Collaborators Launch Open Power AI Consortium



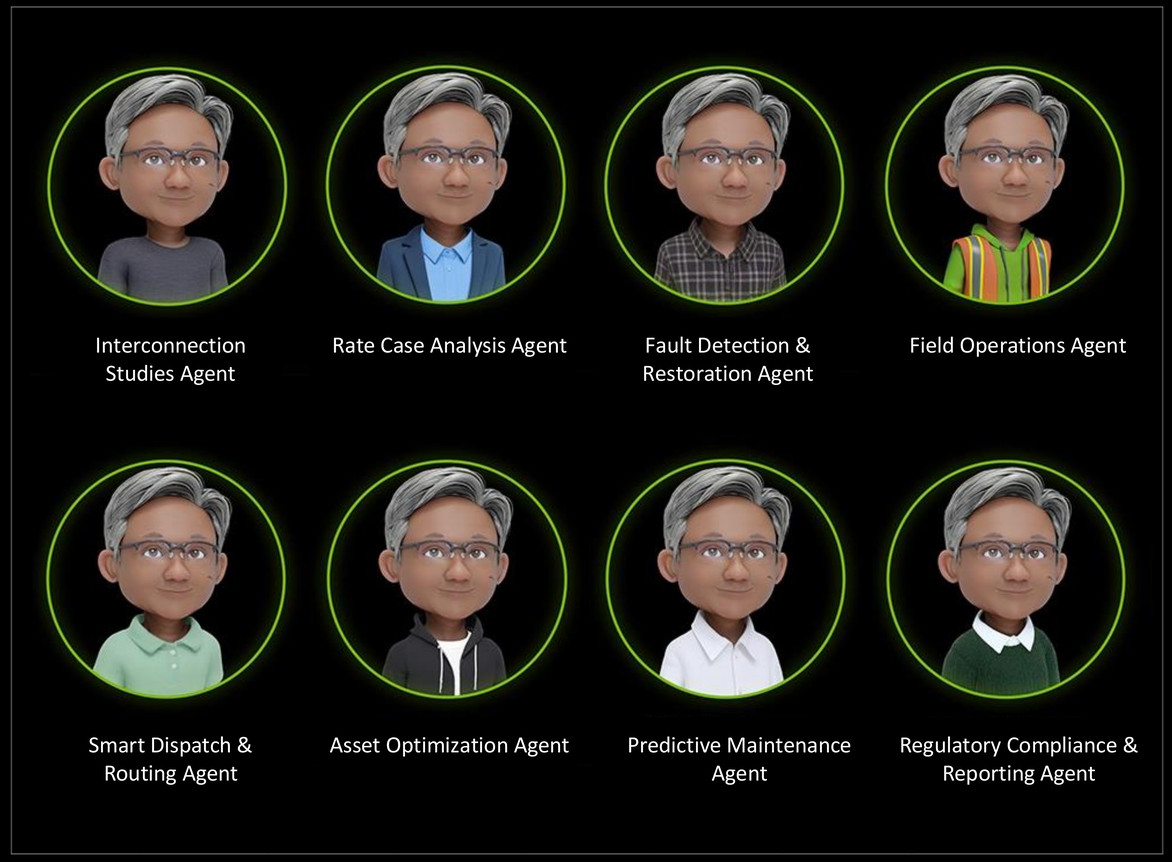
NVIDIA

Agentic AI Enables More Powerful AI Applications

AI agents solve complex problems and use tools to complete work



Agents Augment Humans to Improve Productivity

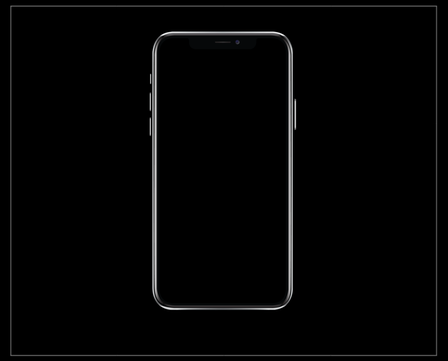


Industry-Specific Agents Will Exist for Every Function

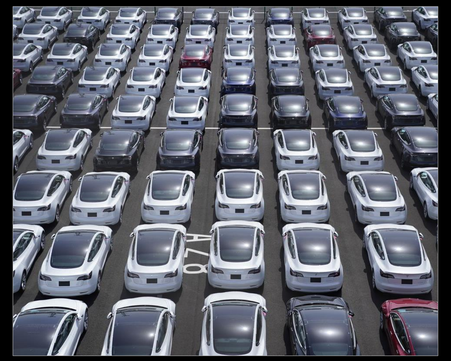
Building a Software-Defined Autonomous Industry

Digital transformation to software-defined industry

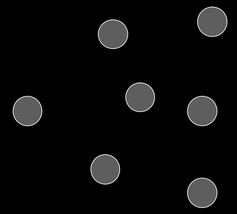
Network of One Vs. Network of Million



Fleet of One Vs. Fleet of Million

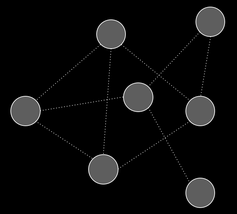


Enabling a Progression to an Autonomous Grid



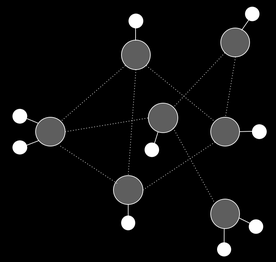
Visibility

Gather real-time data to understand grid conditions



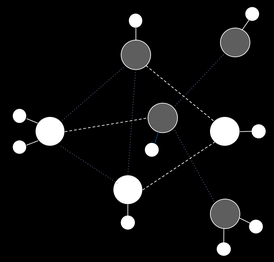
Predictive Analysis

Use the real-time data to forecast and predict grid conditions



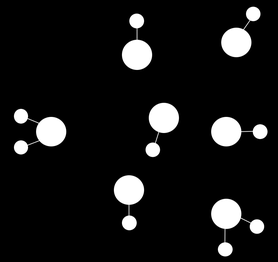
DER Management

Facilitate communications between DERs and the grid so DERs can operate as part of the grid



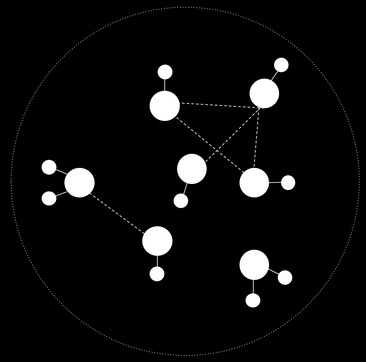
Coordination

Use real-time data and DER management to take preventative action



Dynamic Islanding

Use DERs to power a customer location when the grid goes down



Autonomous Microgrids

Use real-time data and DER management to recover the grid after an emergency

Now — Next — Future

Final Thoughts

- AI enables energy efficient computing
- Complexity of the grid requires AI
- Grid must be software-defined
- Future grid is an autonomous system
- Technology to build an autonomous grid is here
- Open collaboration and partnerships are key
- *Legacy – How did you contribute to a resilient, sustainable energy future?*


NVIDIA