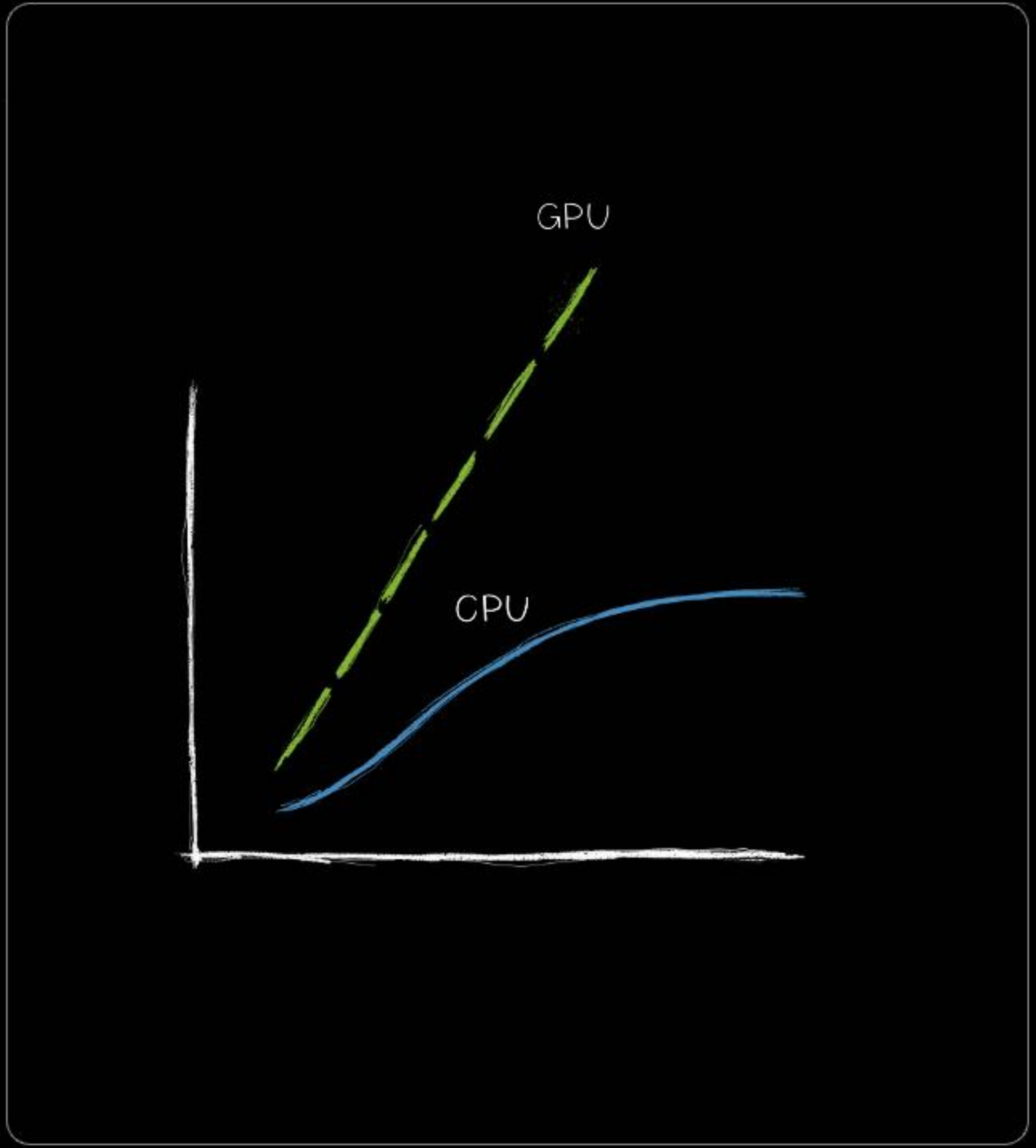
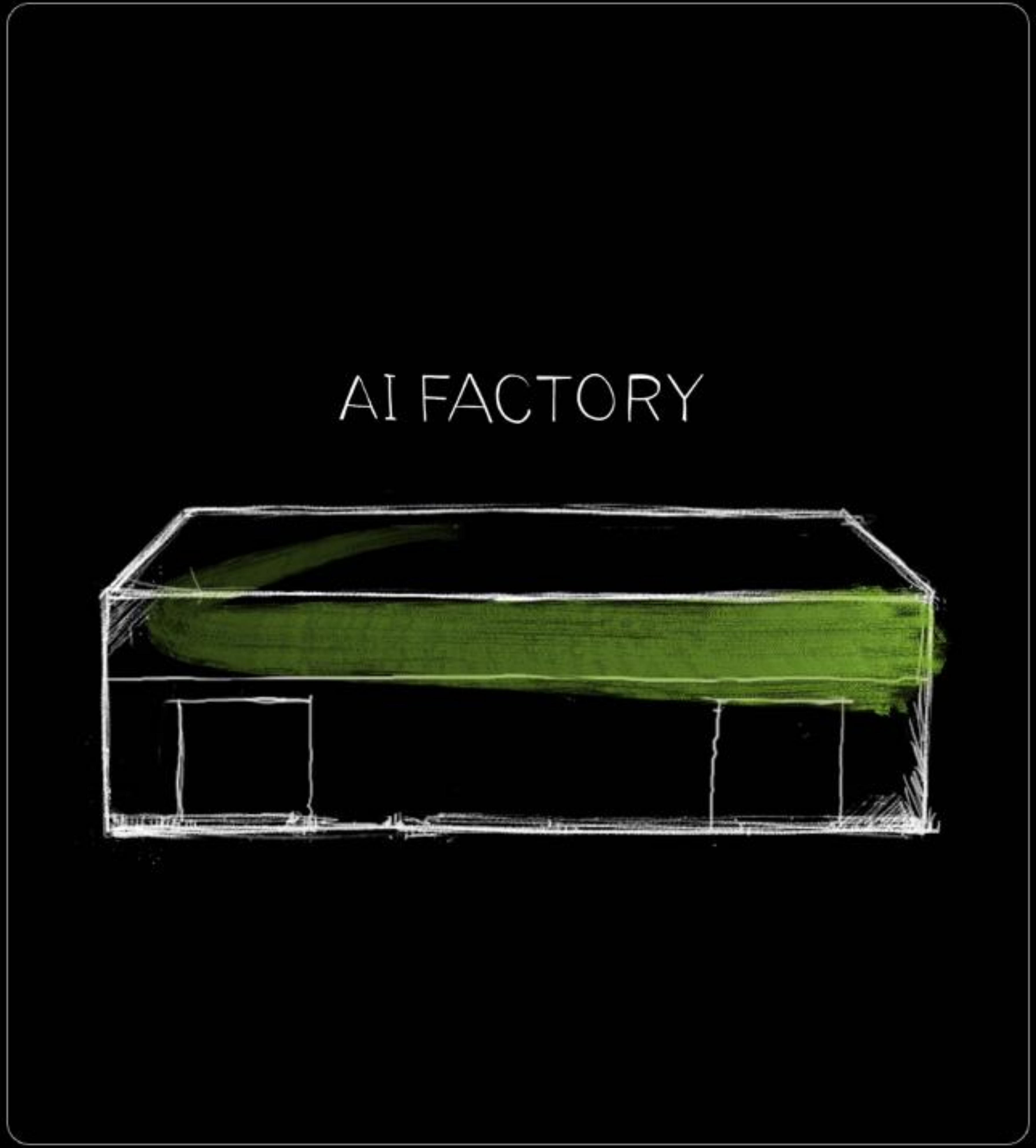
NVIDA AI Factory

Koen van den Berg
AI Factory EMEA
4 November, 2025

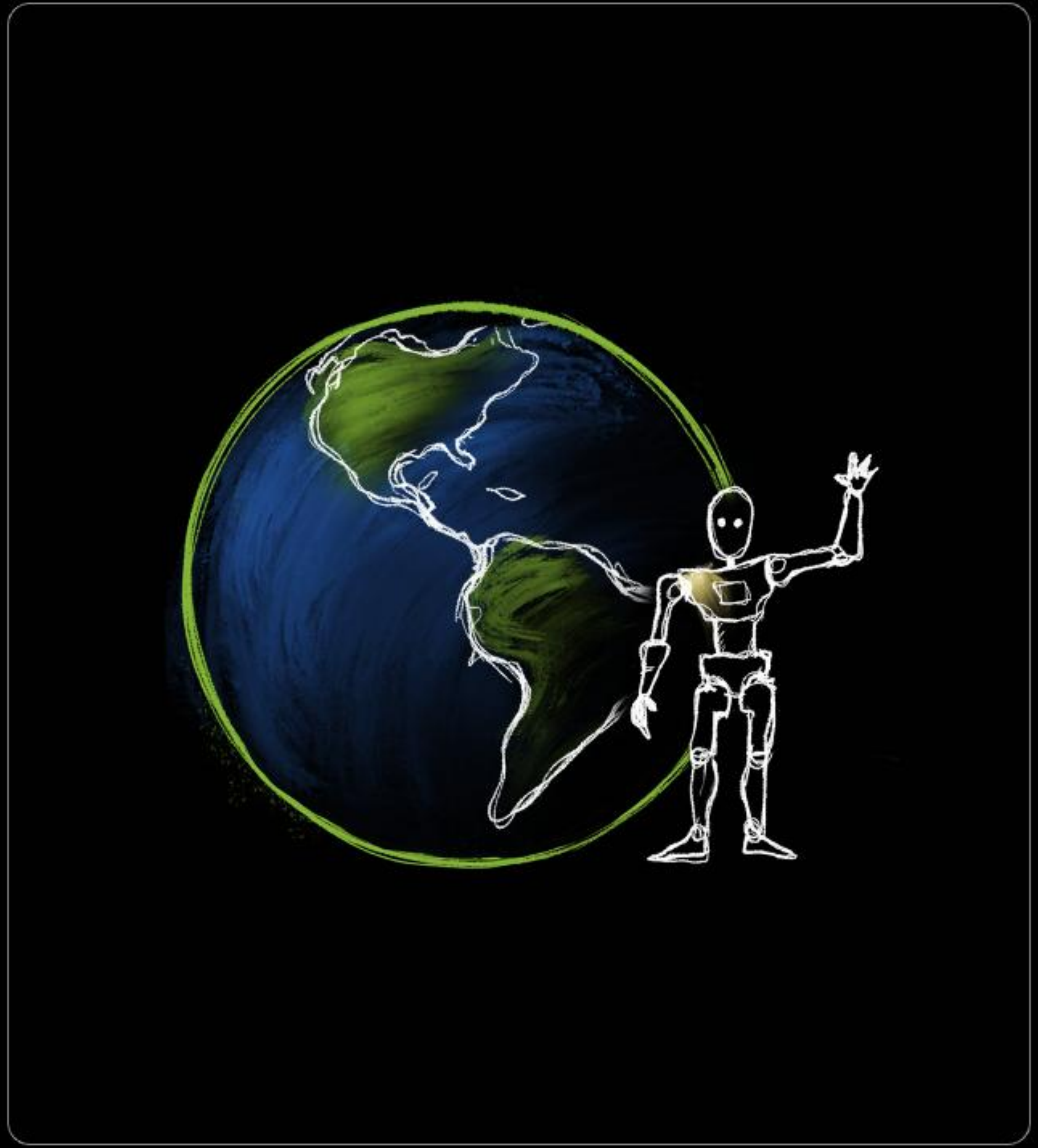
THREE MAJOR TRENDS IN COMPUTING



Traditional -> Accelerated



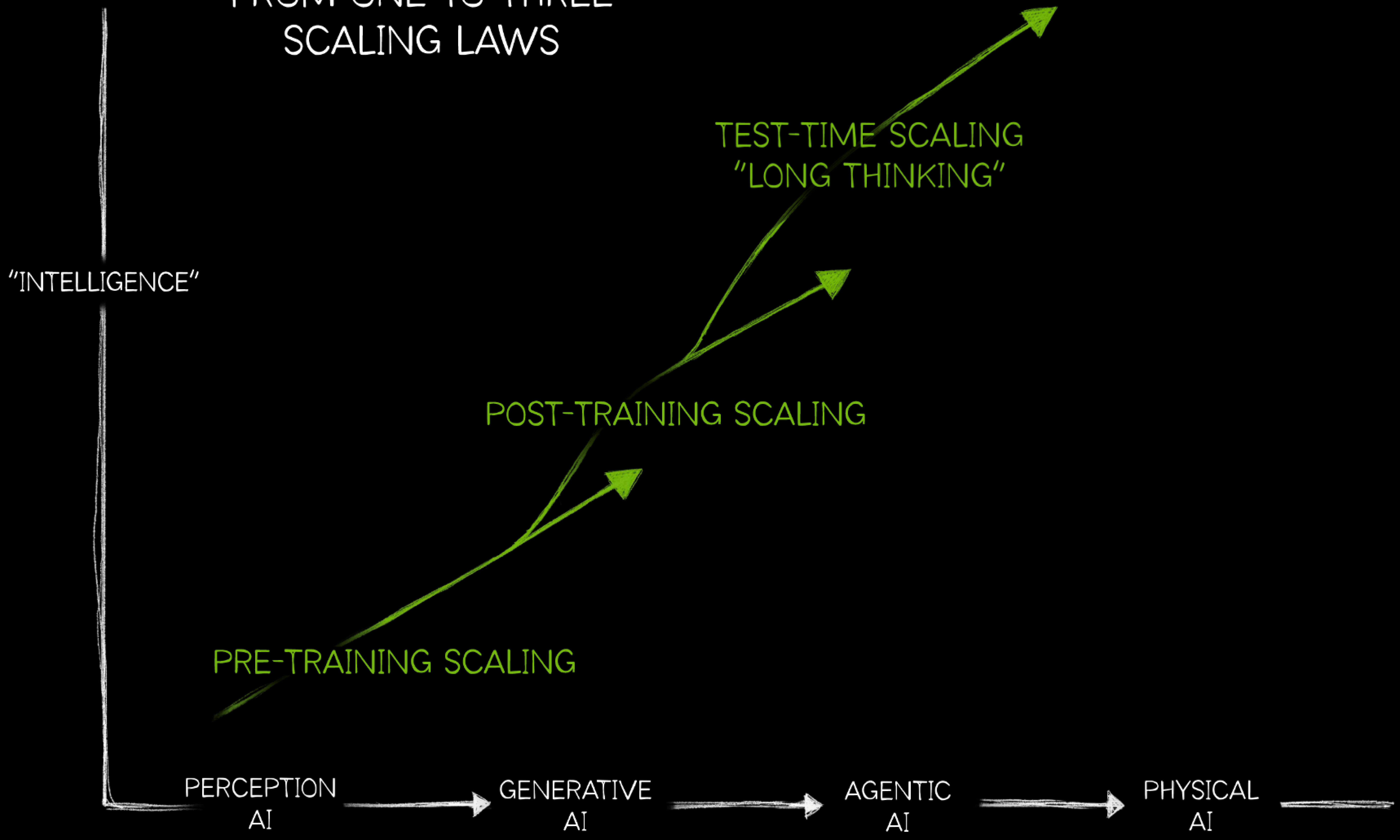
Data Centers -> AI Factories



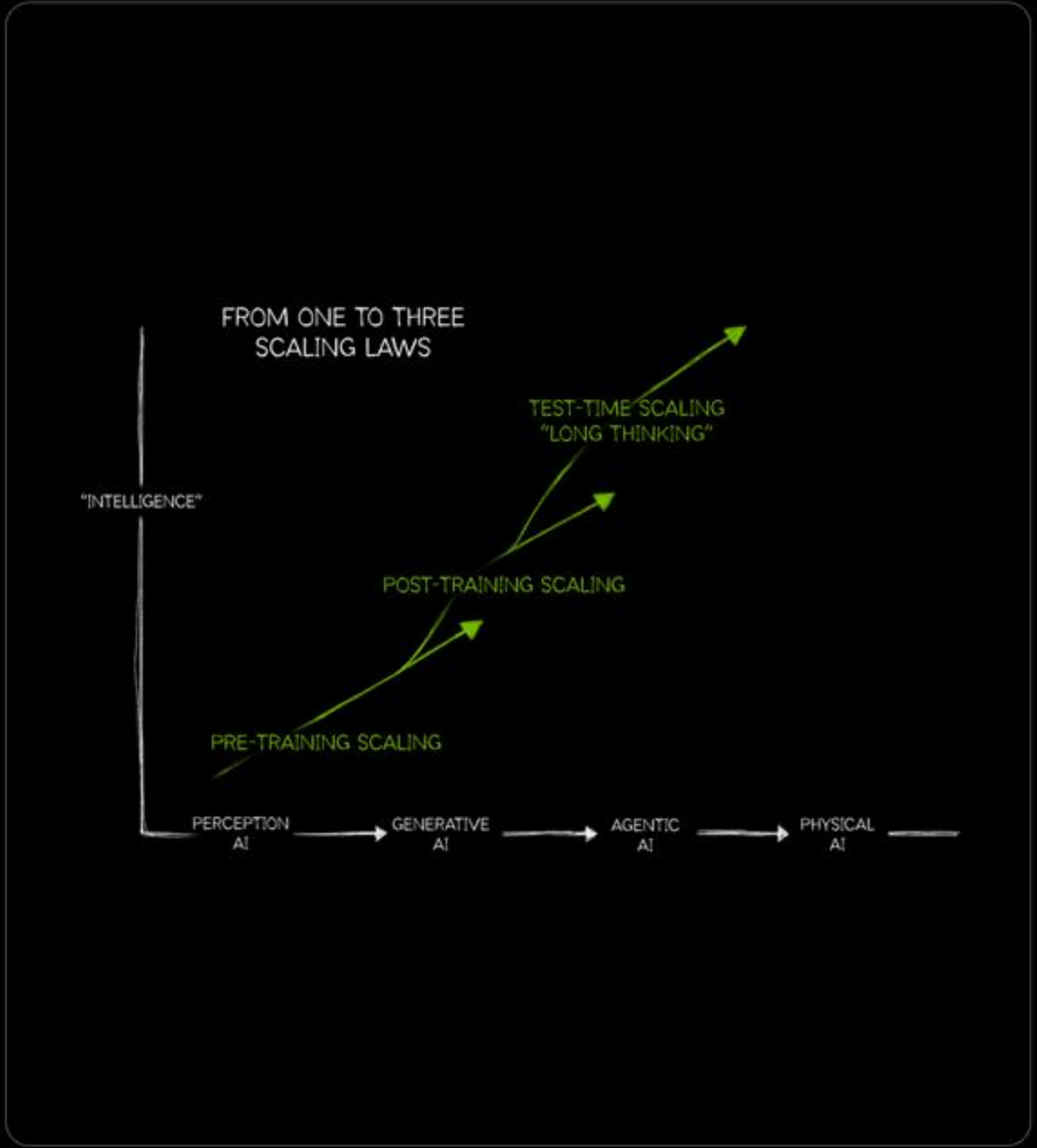
Generative AI -> Agentic AI -> Physical AI


"FIRST CONTACT"
2012 ALEXNET
PERCEPTION AI
SPEECH RECOGNITION
DEEP RECSYS
MEDICAL IMAGING
GENERATIVE AI
DIGITAL MARKETING
CONTENT CREATION
AGENTIC AI
CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE
PHYSICAL AI
AUTONOMOUS VEHICLES
GENERAL ROBOTICS

FROM ONE TO THREE SCALING LAWS



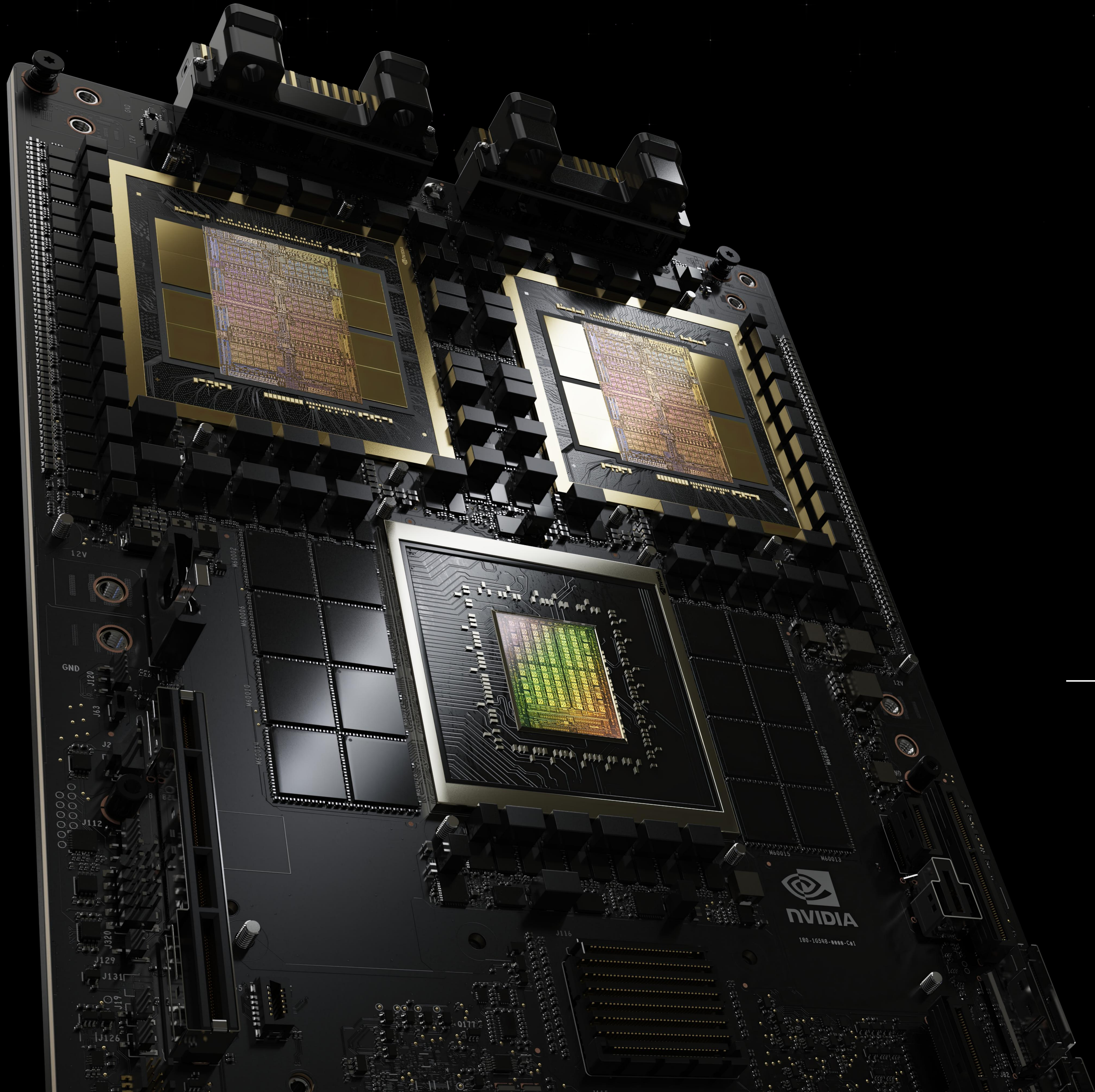
AI Infrastructure Challenges

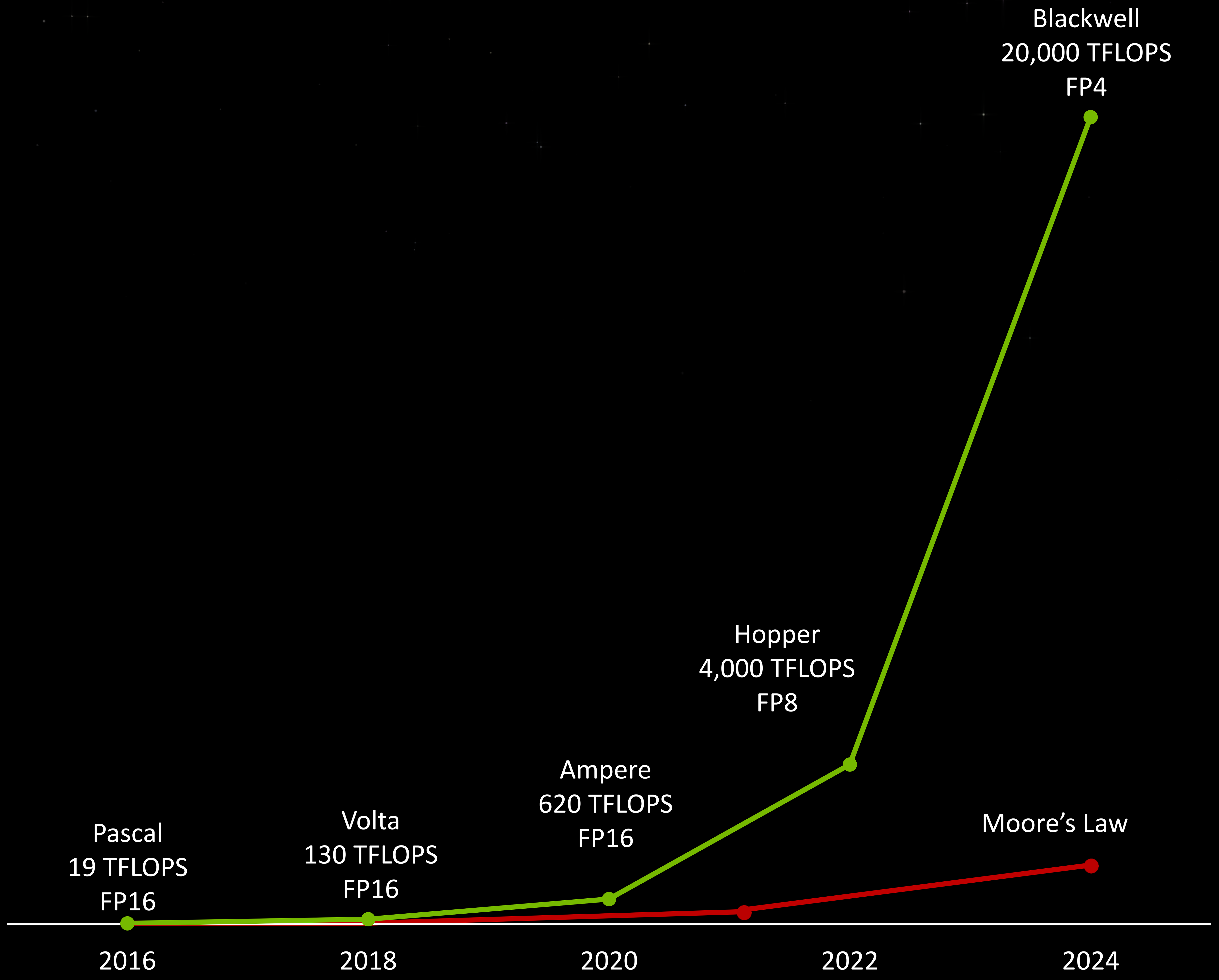


Increasing Compute Demands
Requires proven scale-up performance

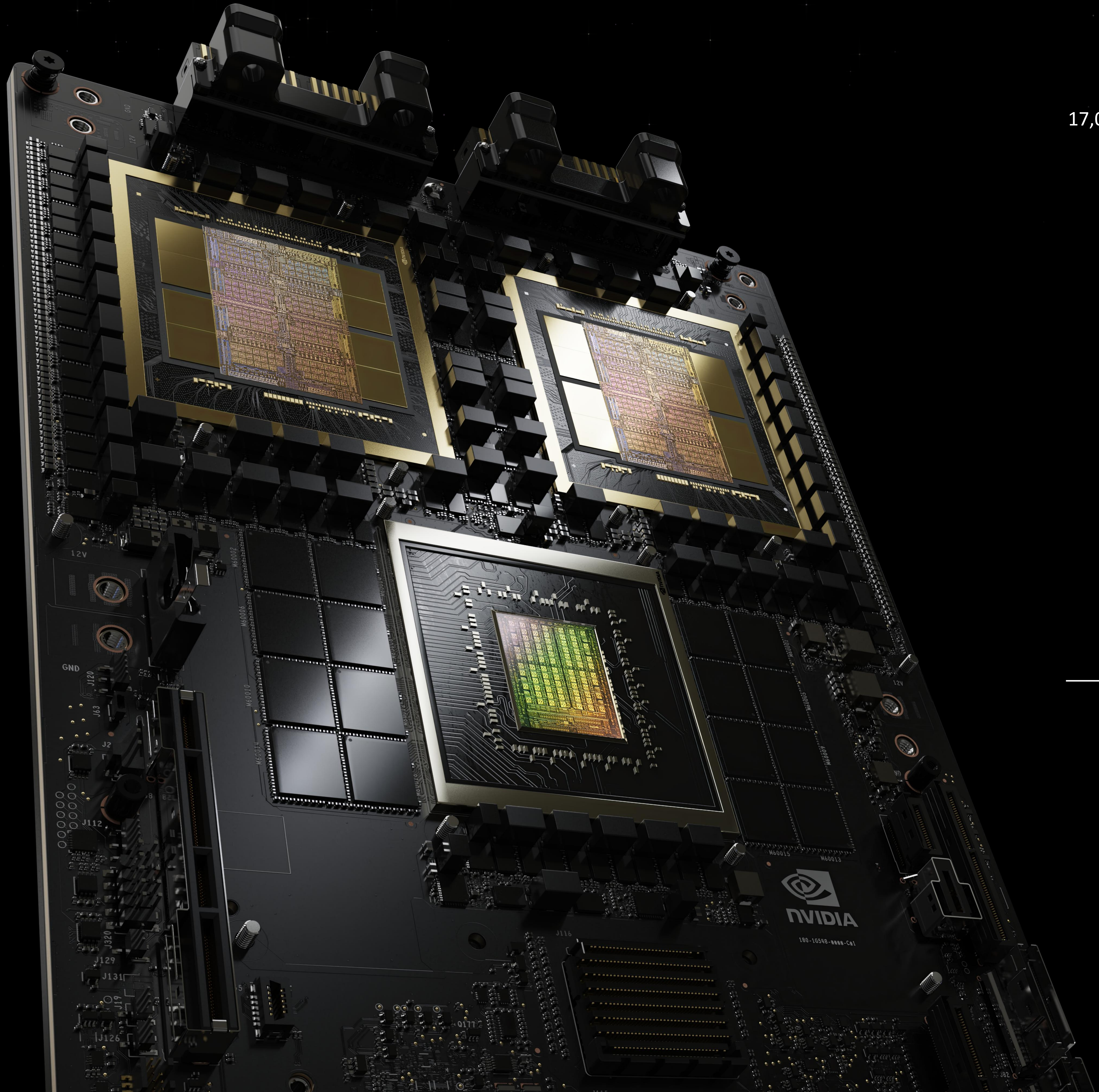
Increasing Scale-up Demands
Requires fast-paced, rack-scale roadmaps

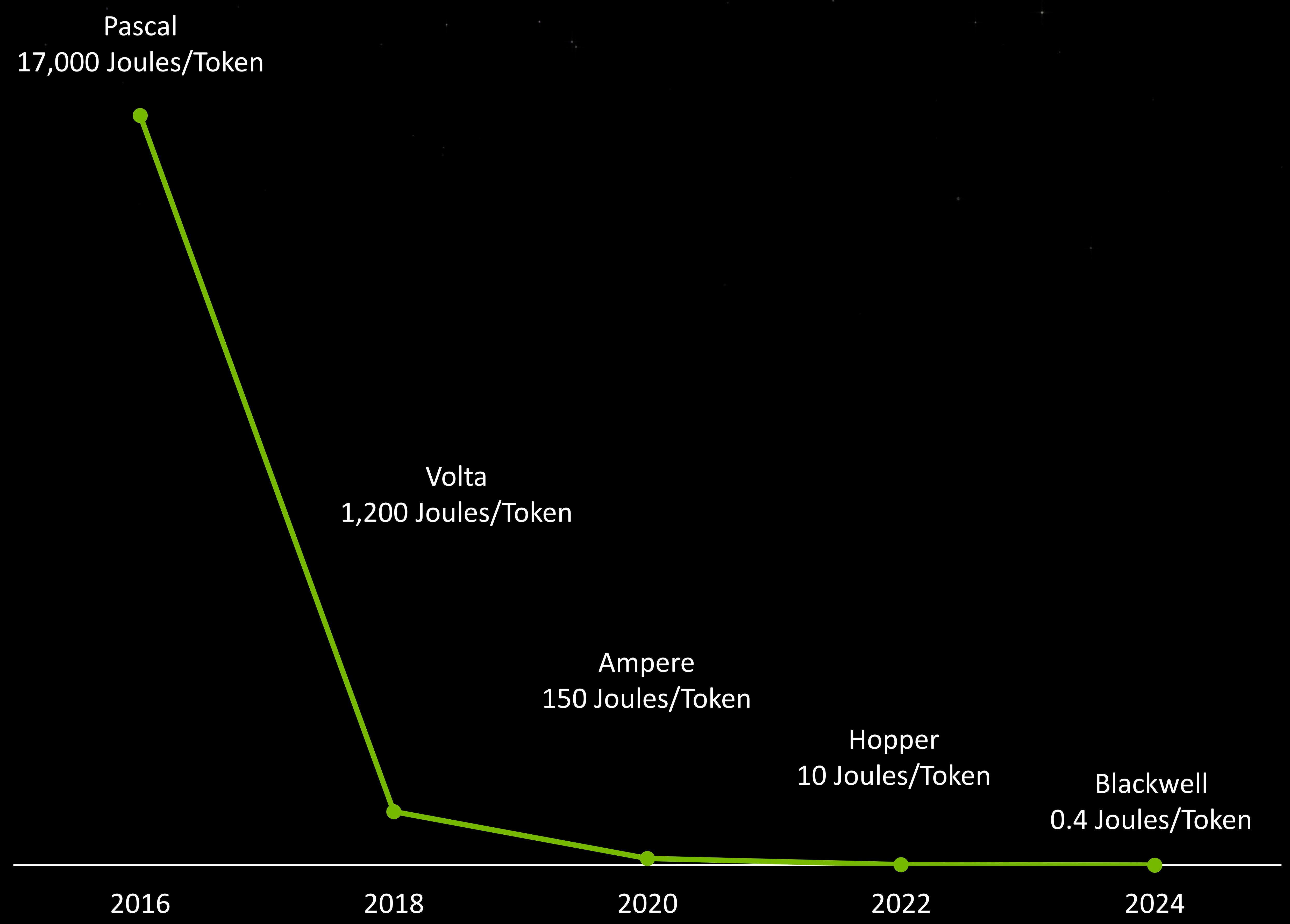
Increasing Infrastructure Complexity
Requires Rich Ecosystem and Common Infrastructure





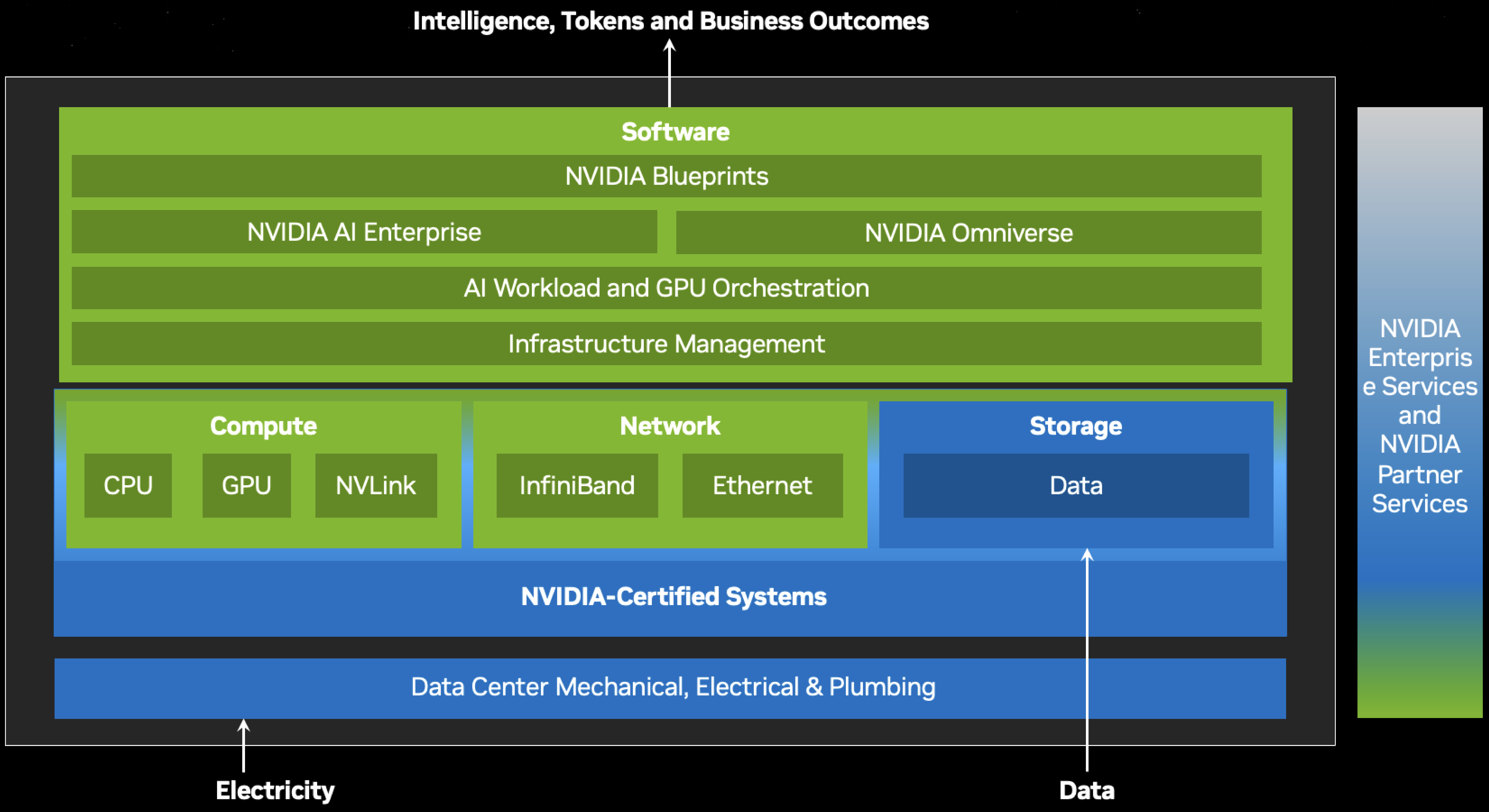
1,000X AI COMPUTE IN 8 YEARS





45,000X ENERGY REDUCTION IN 8 YEARS TO GENERATE GPT4-1.8T TOKENS

NVIDIA Provides a Full Stack for AI Factories



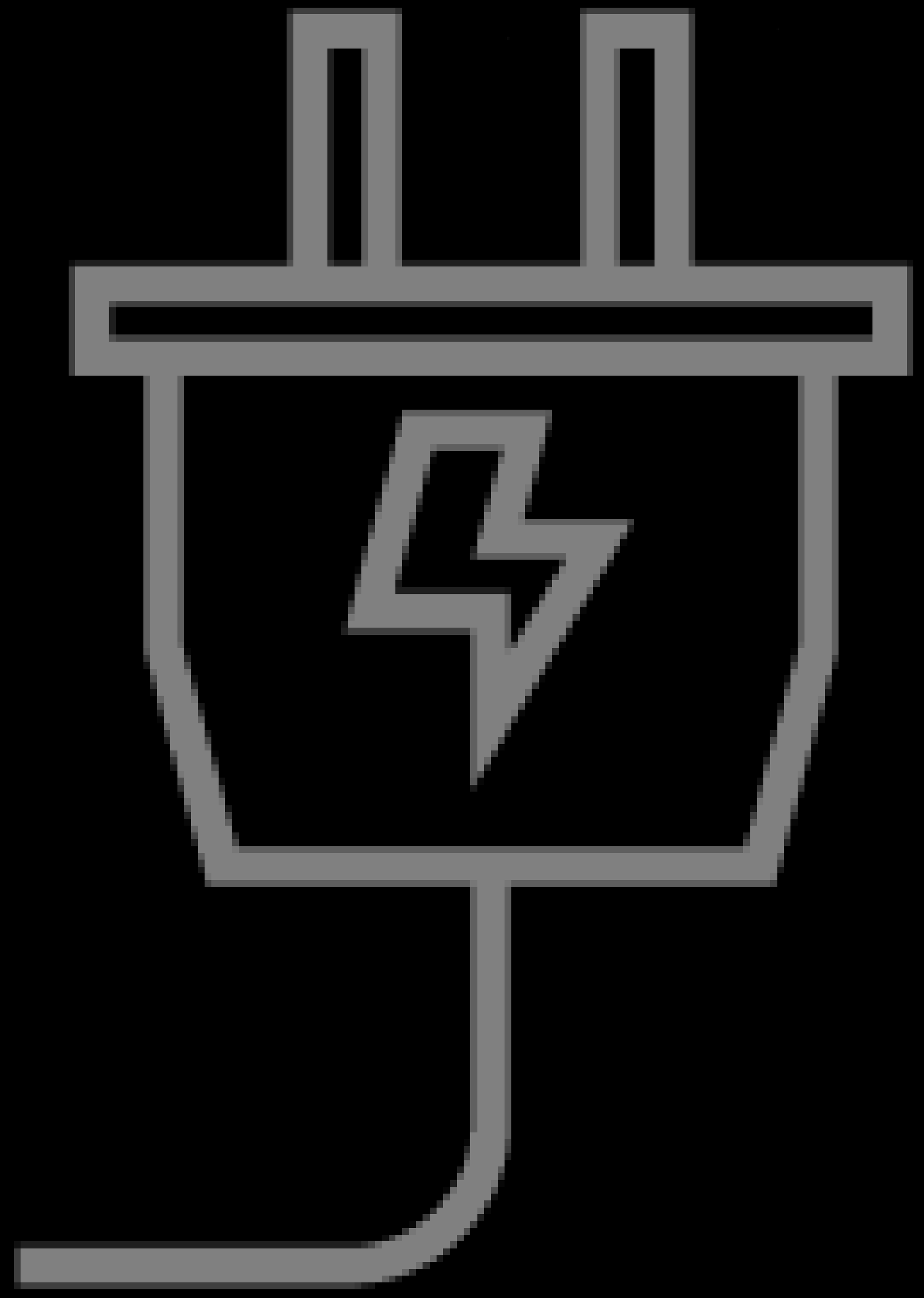
Complexities of Inference Require Full Stack Innovation

Inference Performance Drives Revenue and Profit

NVIDIA delivers the highest performance and efficiency for inference

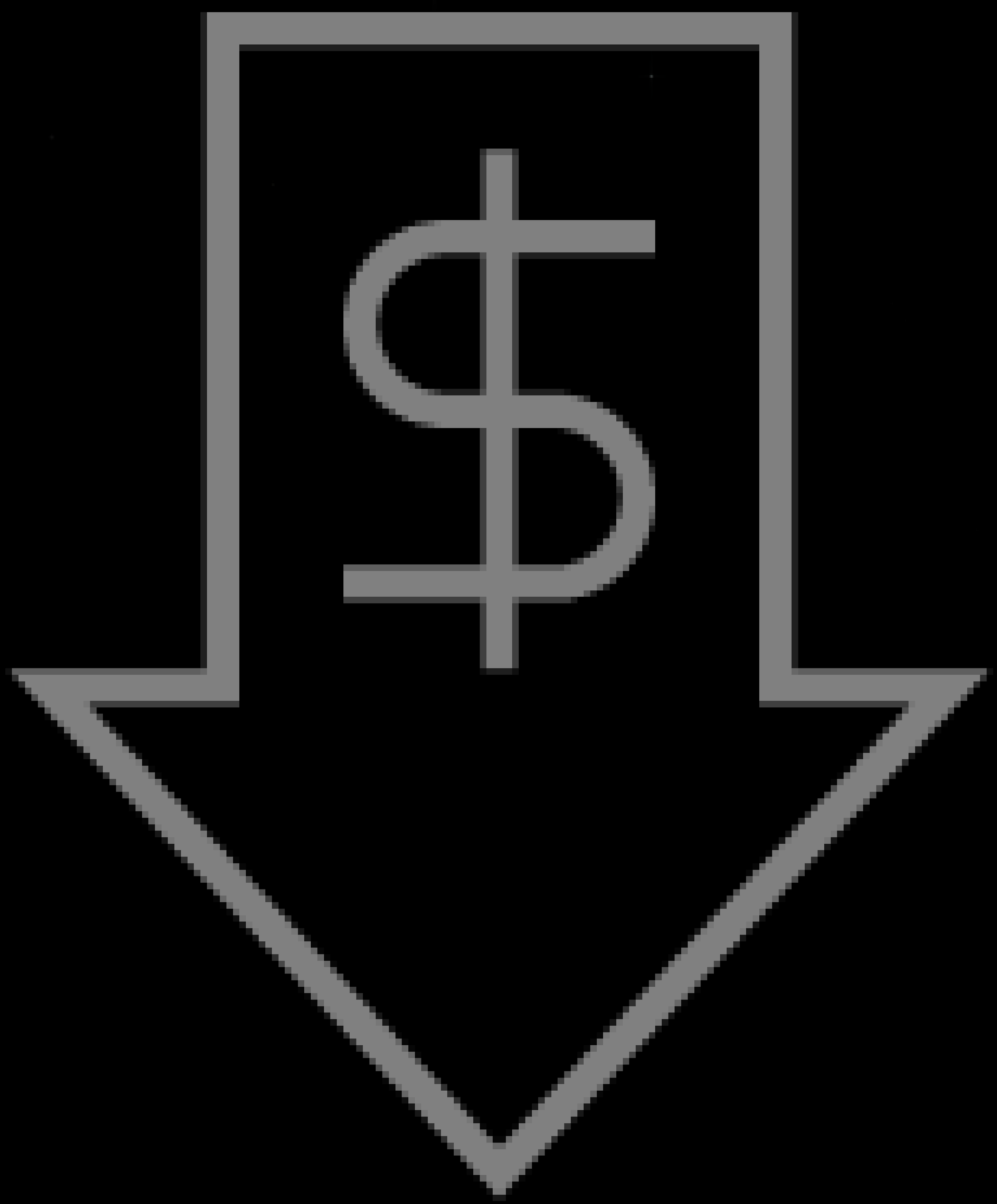
Best Performance

Throughput, interactivity, end-to-end latency



Best Energy Efficiency

Throughput/MW



Lowest TCO

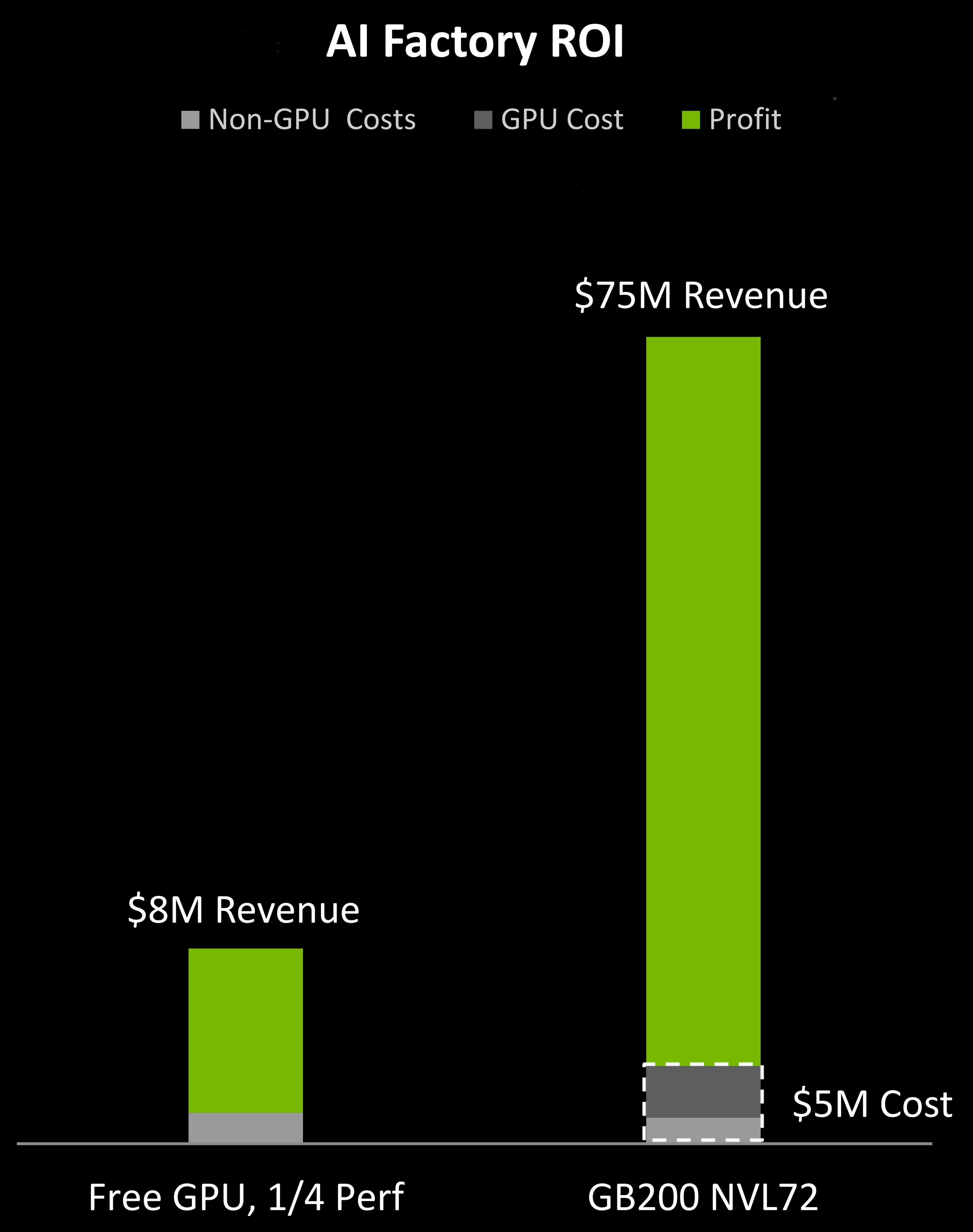
$ per million tokens

Inference Performance Drives Revenue and Profit

Annual Revenue = **Token Throughput** X Price charged per token X Utilization rate

$5M GB200 NVL72 Investment Can Generate $75M Token Revenue

"Free" compute is not cheap enough



Single GB200 NVL72 system performance assuming 3-year operation—DeepSeek R1 based on MLPerf. Assuming $2 per million output token. Actual performance, revenue, ROI may vary. Revenue, profit, and ROI are not primary metrics of MLPerf inference.

AI Factory Compounds Benefits Of NVIDIA Reference Architectures

Software
- NVIDIA AI Enterprise — **5X** Throughput with NIM
- NVIDIA Run:ai — **5X** GPU utilization

Compute
- CPU — **2X** Energy efficiency with Grace Superchip
- GPU — **5X** TFLOPS with Blackwell
- NVLink — **1.5X** Throughput with NVLink

Network
- InfiniBand
- SP-X Ethernet

2X Performance

1.3X Faster Time Deploy

- Time-to-value
- Time-to-first-train
- Time-to-inference
- Scalability

Up to these speeds, may vary based on input/output and model

Thank you!

kvandenberg@nvidia.com