Except for the historical information contained herein, certain matters in this presentation are forward-looking statements. These forward-looking statements and any other forward-looking statements that go beyond historical facts that are made in this presentation are subject to risks and uncertainties that may cause actual results to differ materially. Important factors that could cause actual results to differ materially include: global economic conditions; NVIDIA’s reliance on third parties to manufacture, assemble, package and test NVIDIA’s products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA’s existing product and technologies; market acceptance of NVIDIA’s products or NVIDIA’s partners' products; design, manufacturing or software defects; changes in consumer preferences and demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA’s products or technologies when integrated into systems and other factors.

NVIDIA has based these forward-looking statements largely on its current expectations and projections about future events and trends that it believes may affect its financial condition, results of operations, business strategy, short-term and long-term business operations and objectives, and financial needs. These forward-looking statements are subject to a number of risks and uncertainties, and you should not rely upon the forward-looking statements as predictions of future events. The future events and trends discussed in this presentation may not occur and actual results could differ materially and adversely from those anticipated or implied in the forward-looking statements. Although NVIDIA believes that the expectations reflected in the forward-looking statements are reasonable, the company cannot guarantee that future results, levels of activity, performance, achievements or events and circumstances reflected in the forward-looking statements will occur. Except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances. For a complete discussion of factors that could materially affect NVIDIA’s financial results and operations, please refer to the reports NVIDIA files from time to time with the SEC, including NVIDIA’s most recent Annual Report on Form 10-K, Quarterly Reports on Form 10-Q, and Current Reports on Form 8-K. Copies of reports NVIDIA files with the SEC are posted on NVIDIA’s website and are available from NVIDIA without charge.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements within are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

NVIDIA

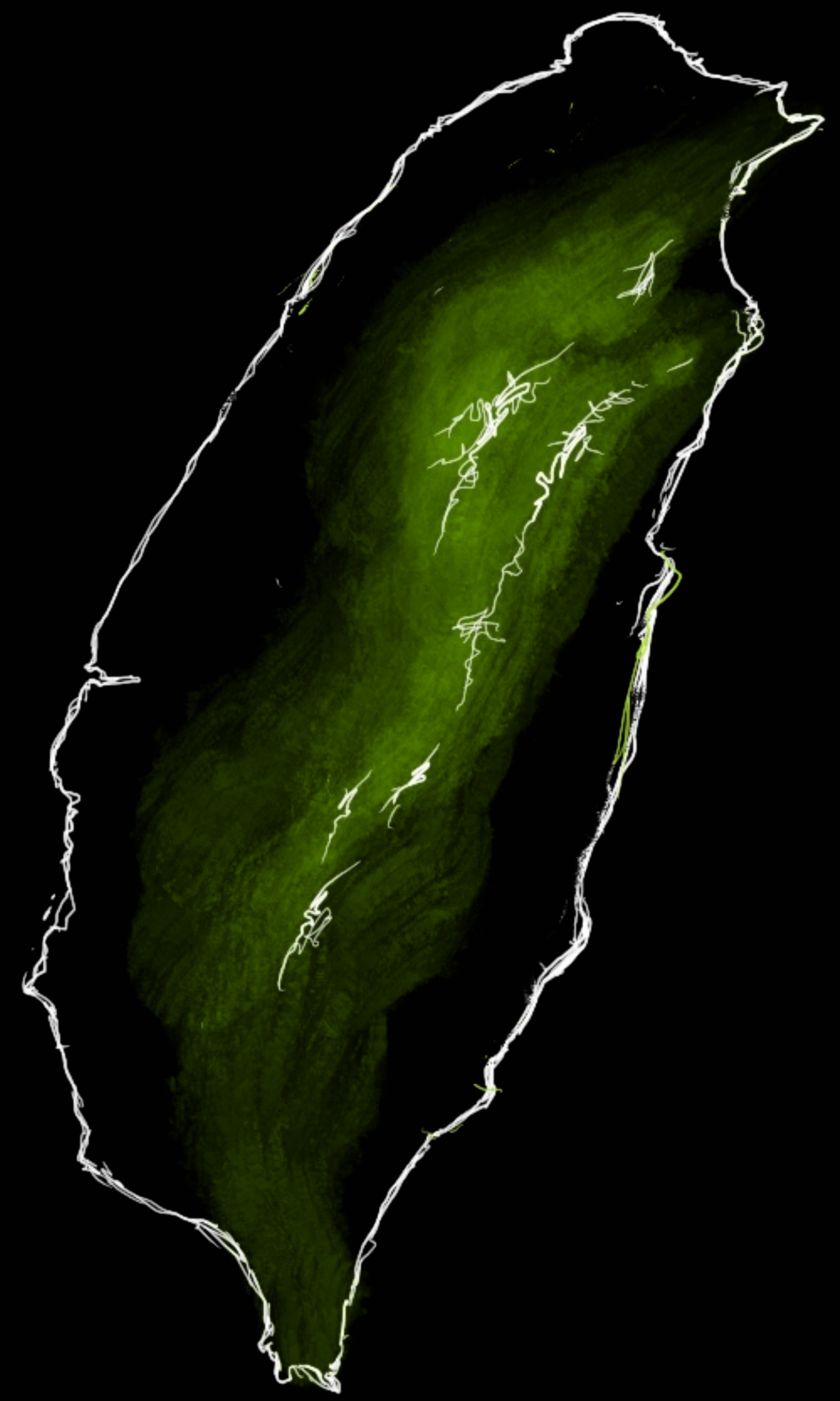
AAEON an ASUS assoc. co.	acer	ADLINK	ADVANTECH	aetina	AIMobile Industrial Mobile Systems	aira	安宏生醫 AnHorn Medicines Innovation · New Therapeutics	APMIC	AppWorks
ASE	ASRock Rack	ASUS	AVALANCHE COMPUTING	AVerMedia	AXIOMTEK	B!gGo	CHENBRO	ChoozMo 集仕多股份有限公司	中原大學 Chung Yuan Christian University
COLORFUL	COMPAL	COOLER MASTER	Coretronic	DeepMentor	DeepRad.AI	DELTA	DMKTZ	EDOM TECHNOLOGY CO., LTD.	EverFocus Your Safety, Our Focus

Footprint-AI Making machine learning for everyone

FORTUNE AI TECHNOLOGIES

Foxconn HON HAI TECHNOLOGY GROUP

Garage+



GIANT GROUP

GIGABYTE

GLIACLOUD

GMI

HOMEE.AI

義守大學 I-SHOU UNIVERSITY

INFINITIES 數位無限軟體

ingrasys

INNO3D

Inventec

InWin

KENMEC 廣運 AUTOMATION TECHNOLOGY

Lanner

LEADTEK

LEDA TECHNOLOGY

律果科技 Legal Tech Inc.

Linker Vision

LITEON 光寶科技 光電事業部

manli

MEDIATEK

MetAI

MiTAC

msi

國立成功大學 National Cheng Kung University 1931

國立臺灣科技大學 NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

國立陽明交通大學 NATIONAL YANG MING CHIAO TUNG UNIVERSITY

NCHC 財團法人國家實驗研究院 國家高速網路與計算中心 National Center for High-performance Computing

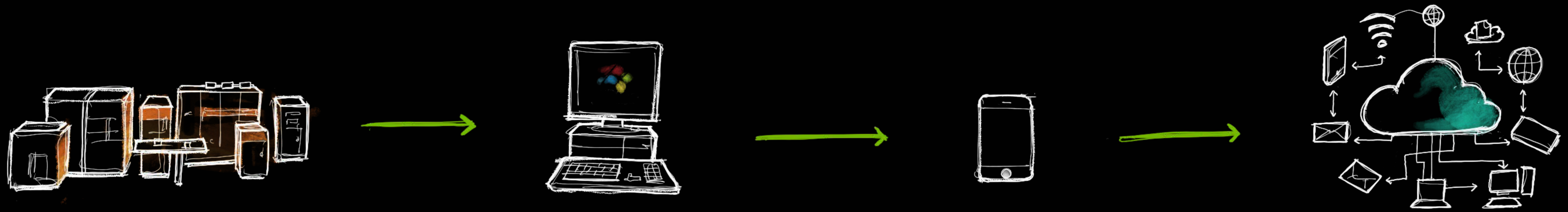
NEOUSYS TECHNOLOGY

NEXCOM

NHRI

onyx Smart Healthcare an ASUS assoc. co.

PALIT	PEGATRON	PROFET AI	QCT	Quanta Computer	SHIH CHIEN UNIVERSITY 實踐	SOLOMON	南臺科技大學	SPIL 矽品精密 Siliconware	
SPINGENCE	StarFab accelerator	Teck Stream	SUPERMICRO	TAMKANG UNIVERSITY	TM ROBOT	TRI innovation	thermaltake	TREND MICRO	tsmc
東海大學 TUNGHAI UNIVERSITY	TWS TAIWAN WEB SERVICE	UMC	Unimicron 欣興電子	Vecow	wistron	wiwynn	YUAN	YUAN ZE UNIVERSITY 1989	ZOTAC



1964
IBM S/360

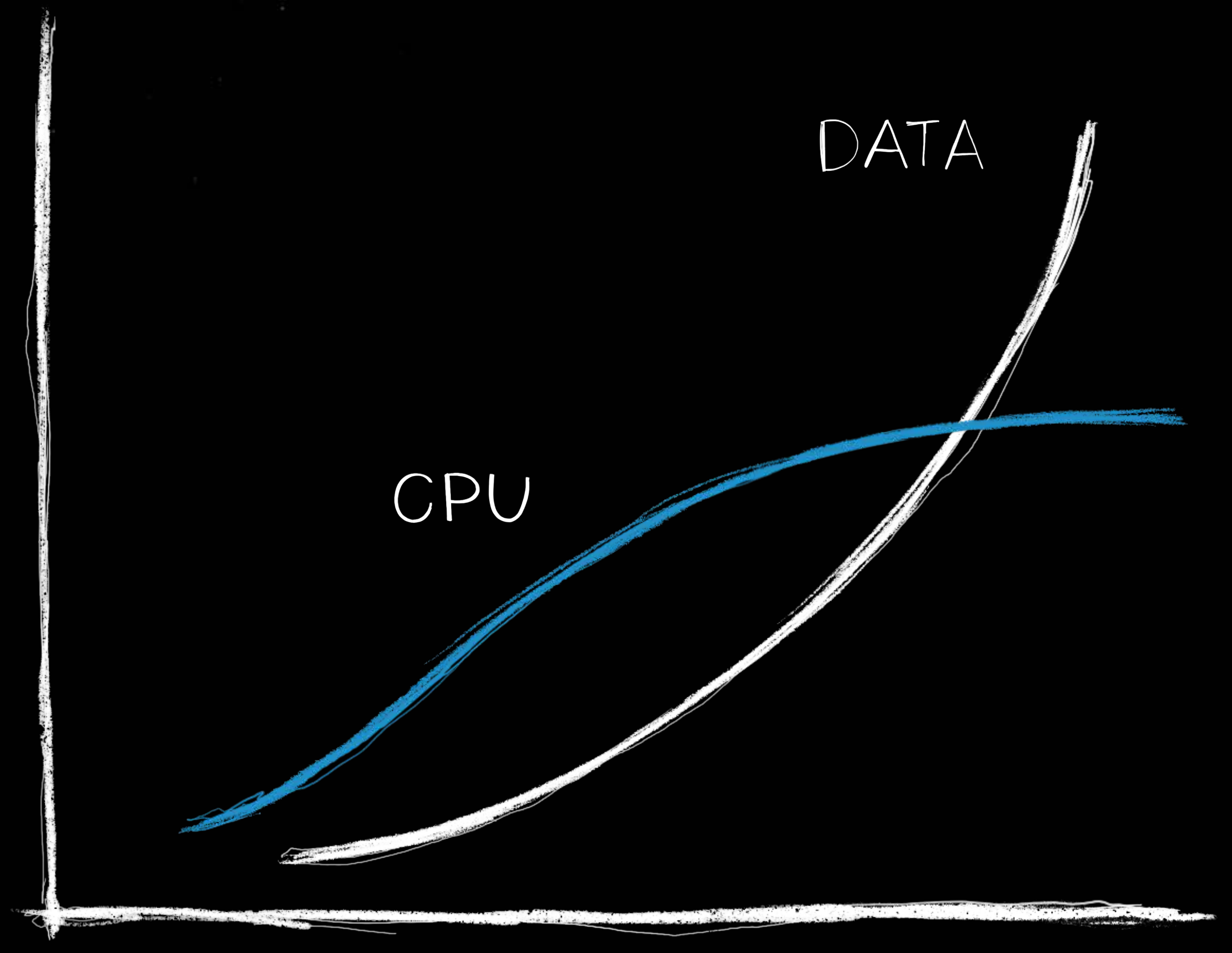
1995
WINDOWS 95
PENTIUM

2007
IPHONE

2007
INTERNET
MOBILE CLOUD

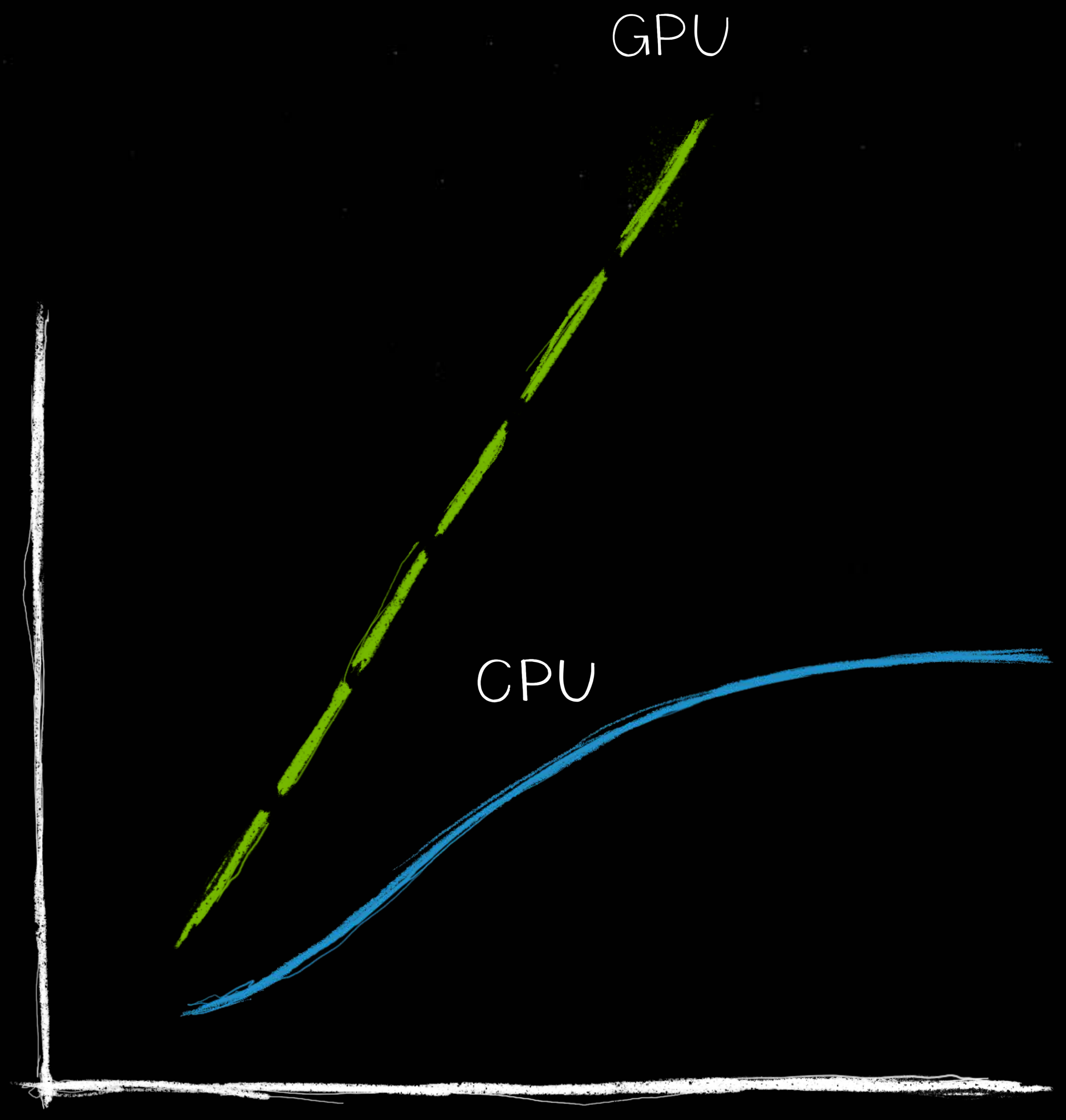
新的運算時代正在開始

A NEW COMPUTING AGE IS STARTING



CPU SCALING SLOWS

. . . AND
COMPUTE DEMAND GROWS
EXPONENTIALLY



GPU – ACCELERATED COMPUTING

2006
CUDA

加速每個應用程式

ACCELERATE EVERY APPLICATION

CPU
1t
100t
CPU
GPU
+
1t

= ~100X SPEED-UP

~3X POWER

~1.5X COST

60X PERF / $ OR 98% SAVINGS

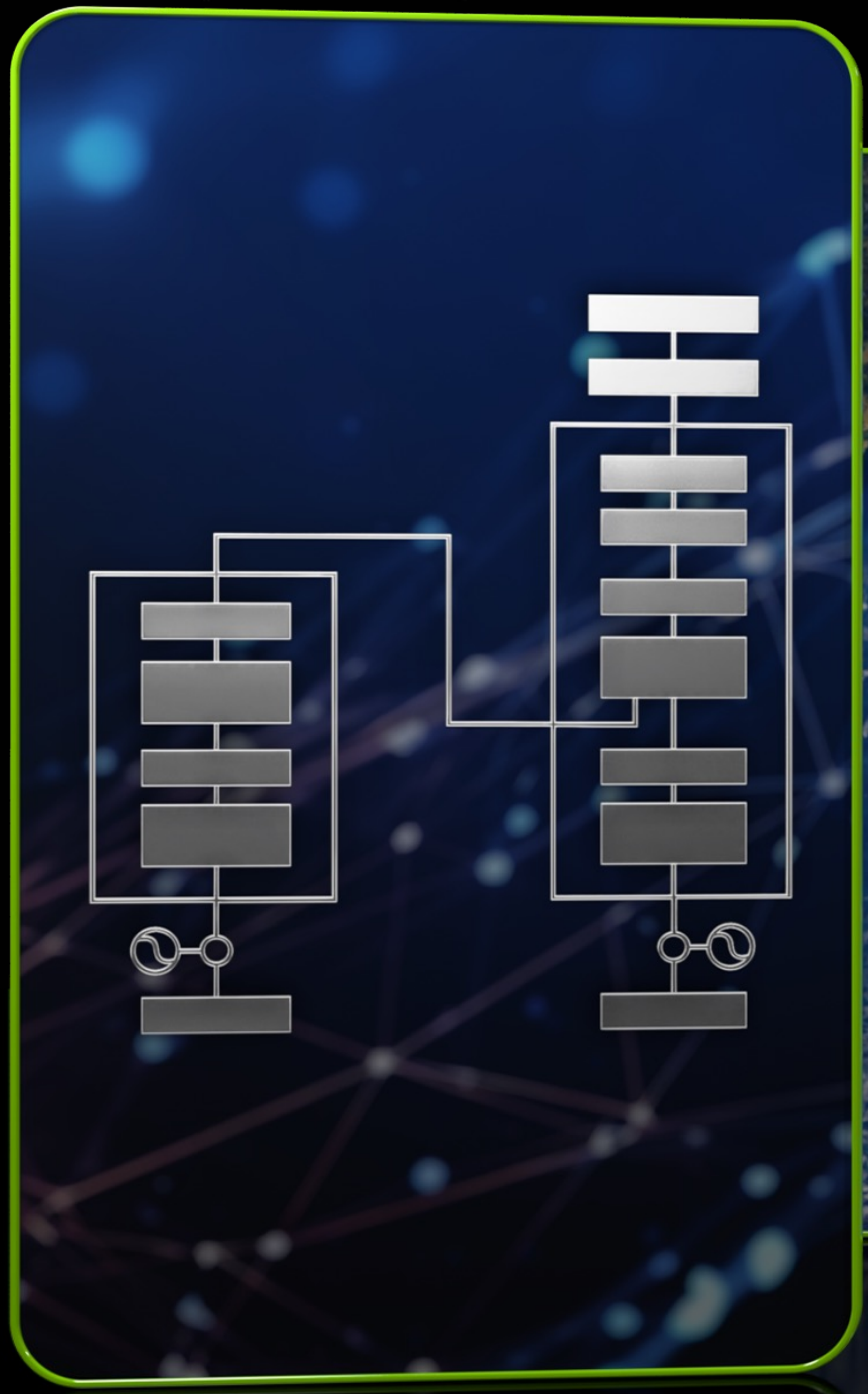
30X PERF / W OR 97% SAVINGS

買越多 … 省越多

"THE MORE YOU BUY… THE MORE YOU SAVE"

NVIDIA CUDA 函式庫開拓新市場

NVIDIA CUDA LIBRARIES OPEN NEW MARKETS

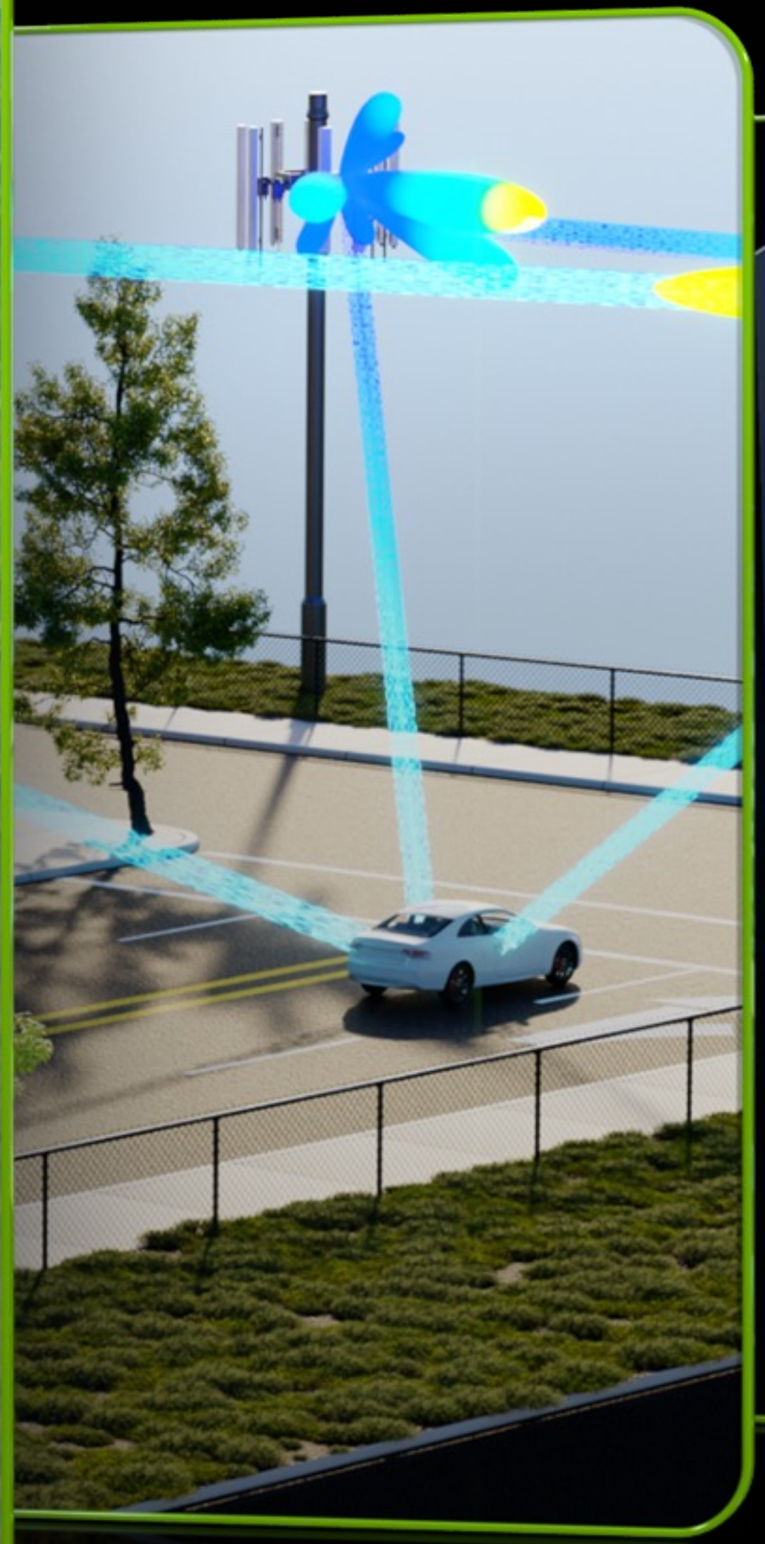


cuDNN

Deep Learning

Modulus

AI Physics

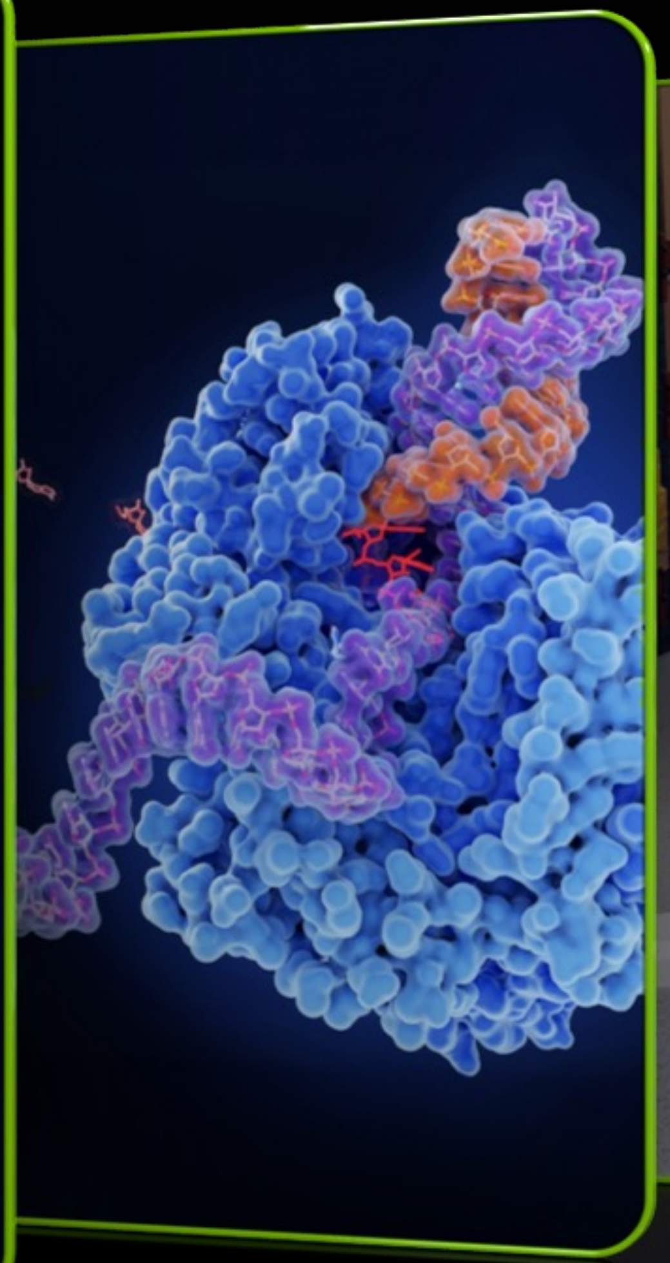


Aerial RAN

AI Radio

cuLITHO

Computational Lithography

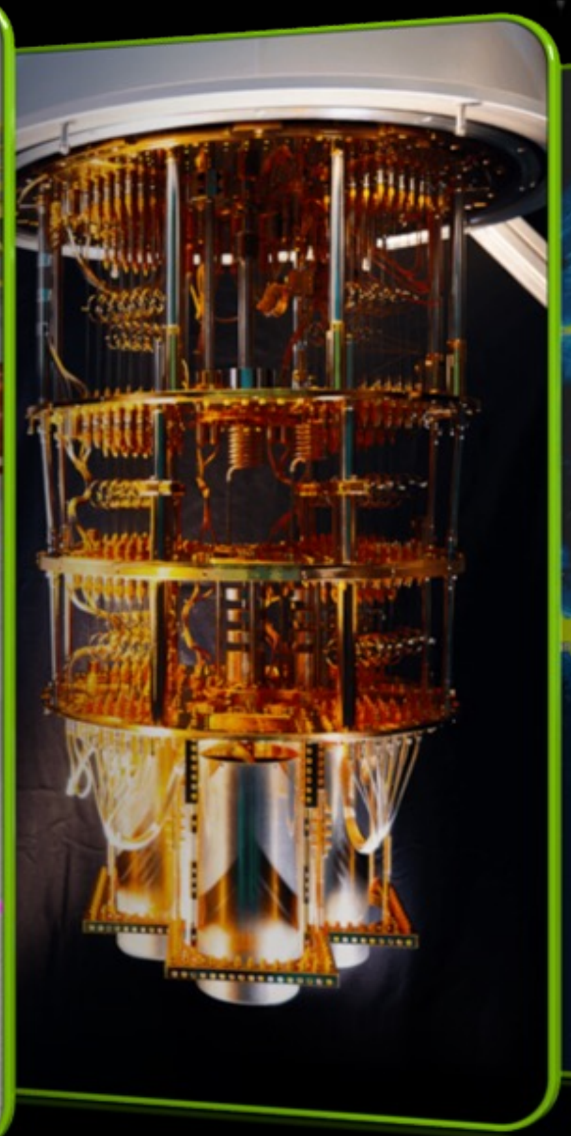


Parabricks

Gene Sequencing

cuOPT

Combinatorial Optimization

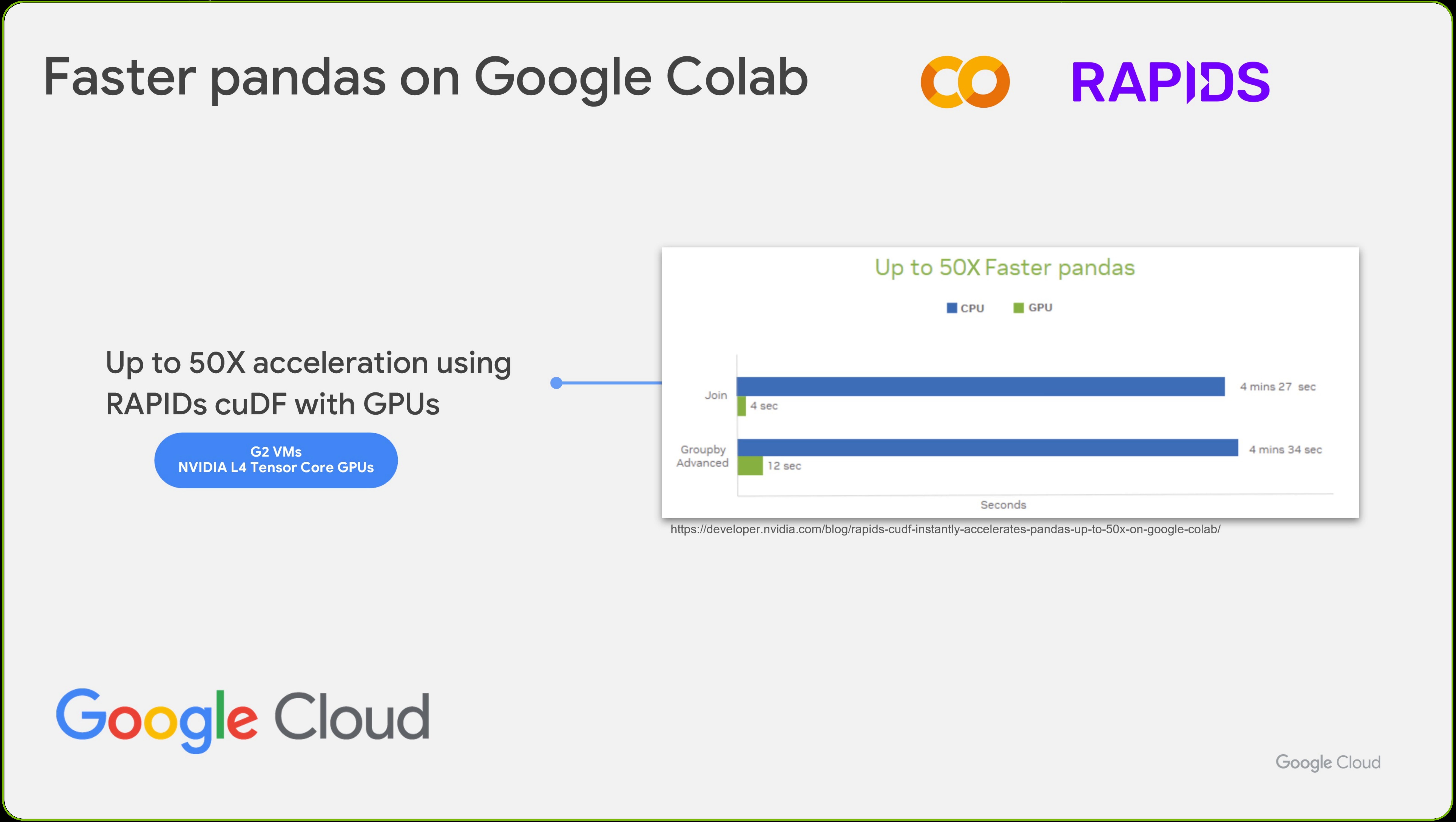


cuQUANTUM

QC Simulation

cuDF

Data Processing



Faster pandas on Google Colab

Up to 50X acceleration using RAPIDs cuDF with GPUs

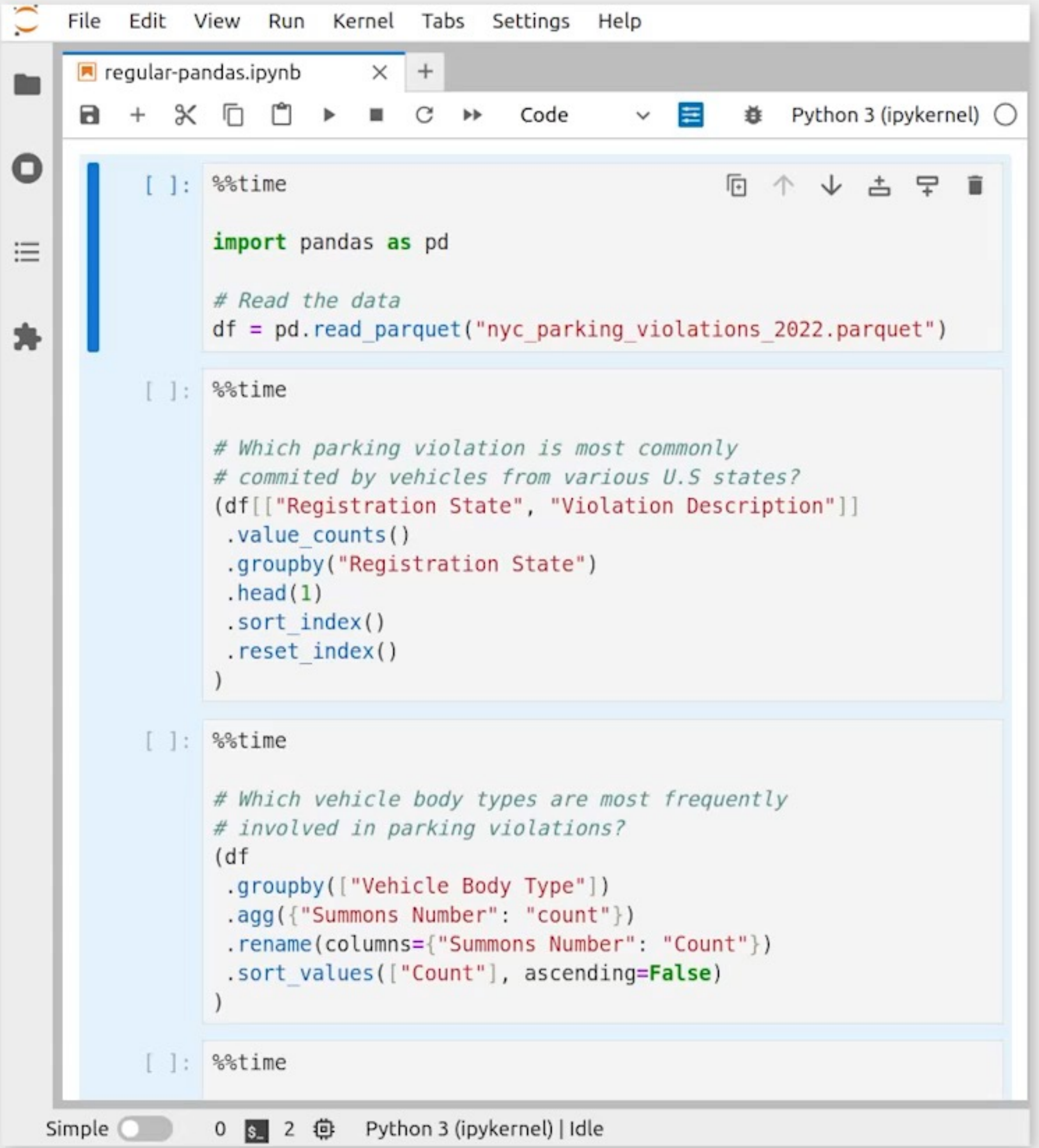
G2 VMs
NVIDIA L4 Tensor Core GPUs

https://developer.nvidia.com/blog/rapids-cudf-instantly-accelerates-pandas-up-to-50x-on-google-colab/

Google Cloud

NVIDIA 宣佈推出 GOOGLE COLAB 中的 PANDAS-CUDF

NVIDIA ANNOUNCES PANDAS-CUDF IN GOOGLE COLAB

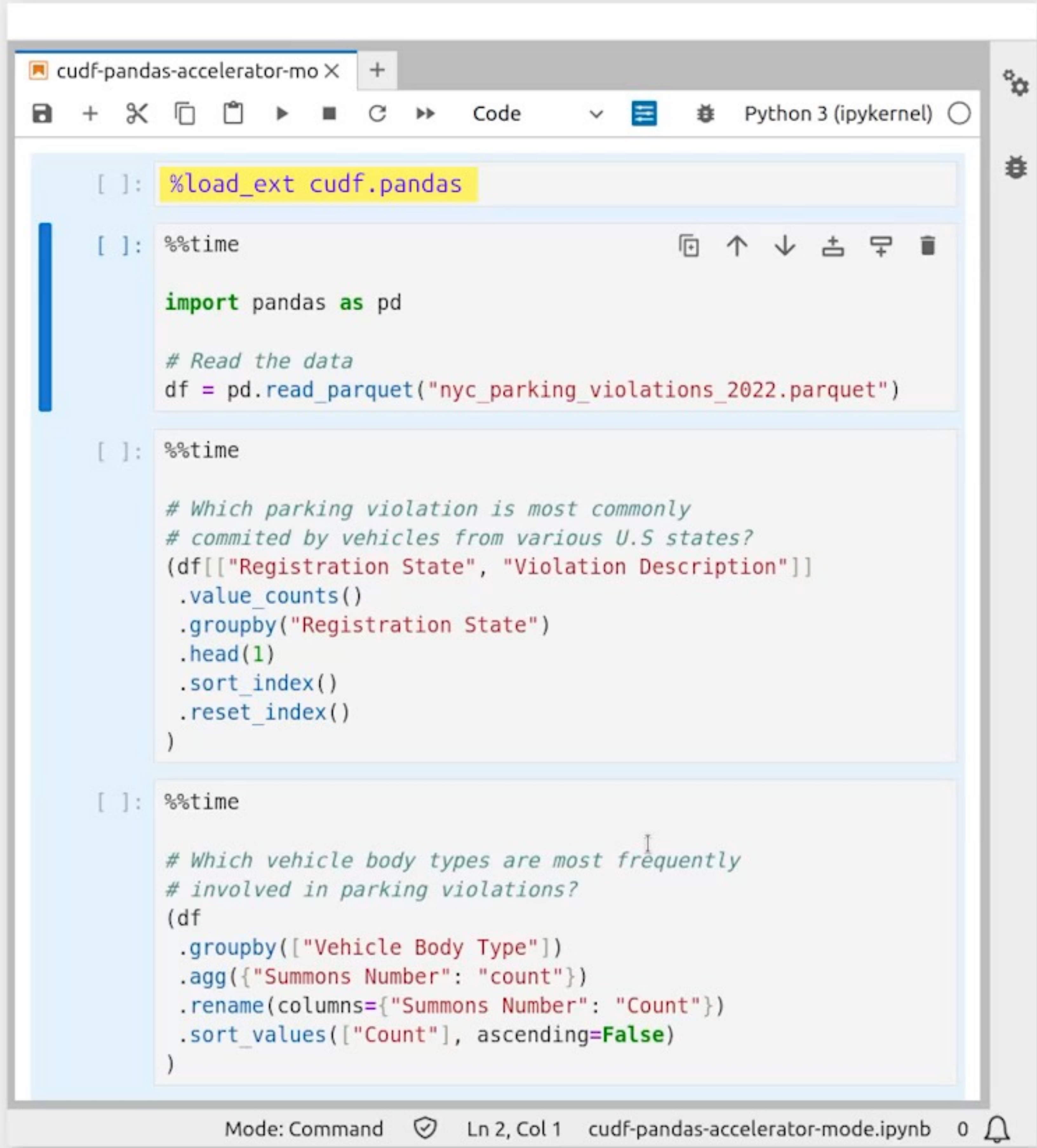


```
[ ]: %%time

     import pandas as pd

     # Read the data
     df = pd.read_parquet("nyc_parking_violations_2022.parquet")
```

```
[ ]: %%time

     # Which parking violation is most commonly
     # commited by vehicles from various U.S states?
     (df[["Registration State", "Violation Description"]]
      .value_counts()
      .groupby("Registration State")
      .head(1)
      .sort_index()
      .reset_index()
     )
```

```
[ ]: %%time

     # Which vehicle body types are most frequently
     # involved in parking violations?
     (df
      .groupby(["Vehicle Body Type"])
      .agg({"Summons Number": "count"})
      .rename(columns={"Summons Number": "Count"})
      .sort_values(["Count"], ascending=False)
     )
```

```
[ ]: %%time
```

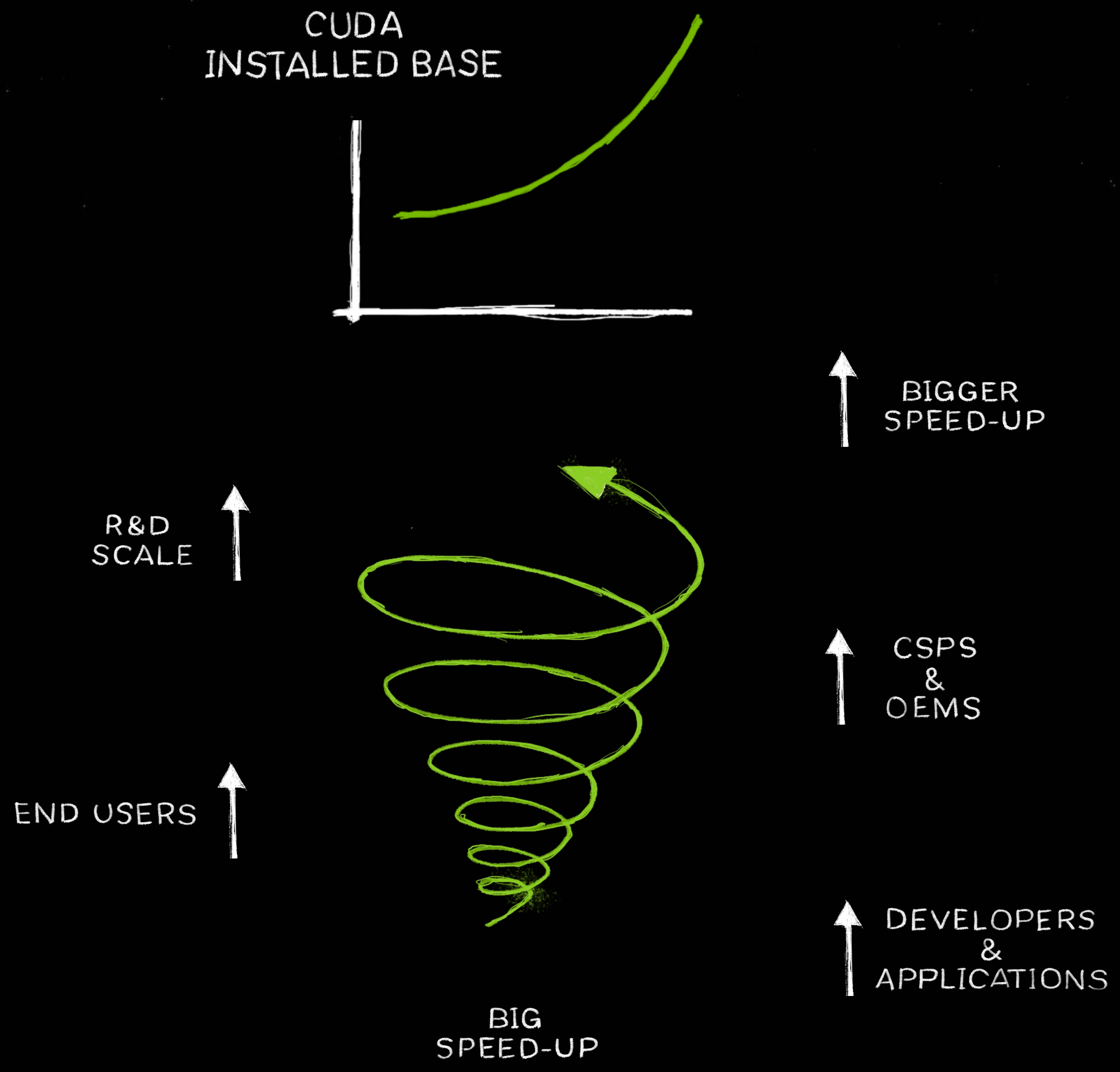

cudf-pandas-accelerator-mo
Code
Python 3 (ipykernel)
Mode: Command
Ln 2, Col 1
cudf-pandas-accelerator-mode.ipynb
0

```
[ ]: %load_ext cudf.pandas
```

```
[ ]: %%time

     import pandas as pd

     # Read the data
     df = pd.read_parquet("nyc_parking_violations_2022.parquet")
```

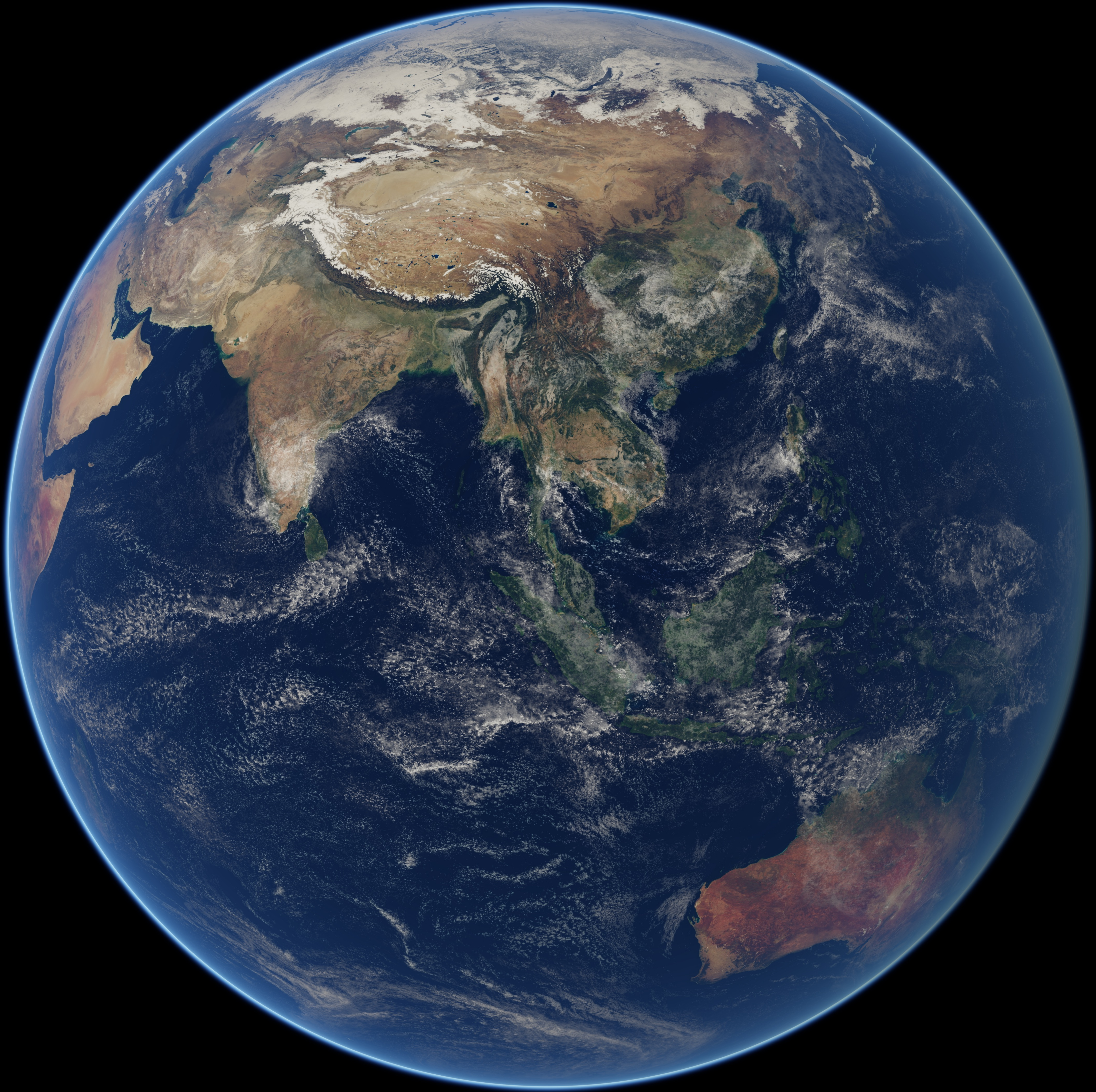
```
[ ]: %%time

     # Which parking violation is most commonly
     # commited by vehicles from various U.S states?
     (df[["Registration State", "Violation Description"]]
      .value_counts()
      .groupby("Registration State")
      .head(1)
      .sort_index()
      .reset_index()
     )
```

```
[ ]: %%time

     # Which vehicle body types are most frequently
     # involved in parking violations?
     (df
      .groupby(["Vehicle Body Type"])
      .agg({"Summons Number": "count"})
      .rename(columns={"Summons Number": "Count"})
      .sort_values(["Count"], ascending=False)
     )
```

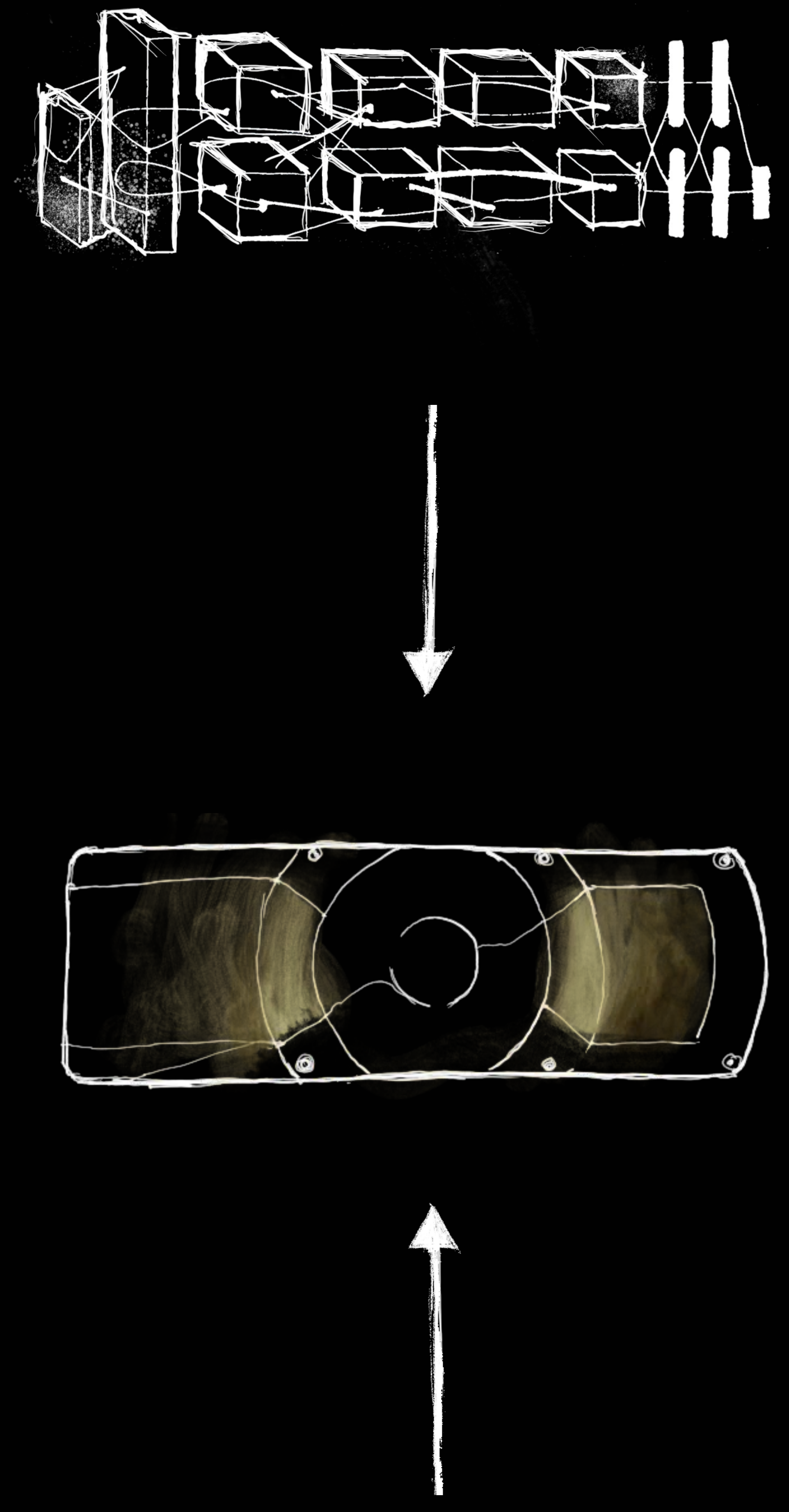


CUDA 實現良性循環

CUDA ACHIEVES VIRTUOUS CYCLE

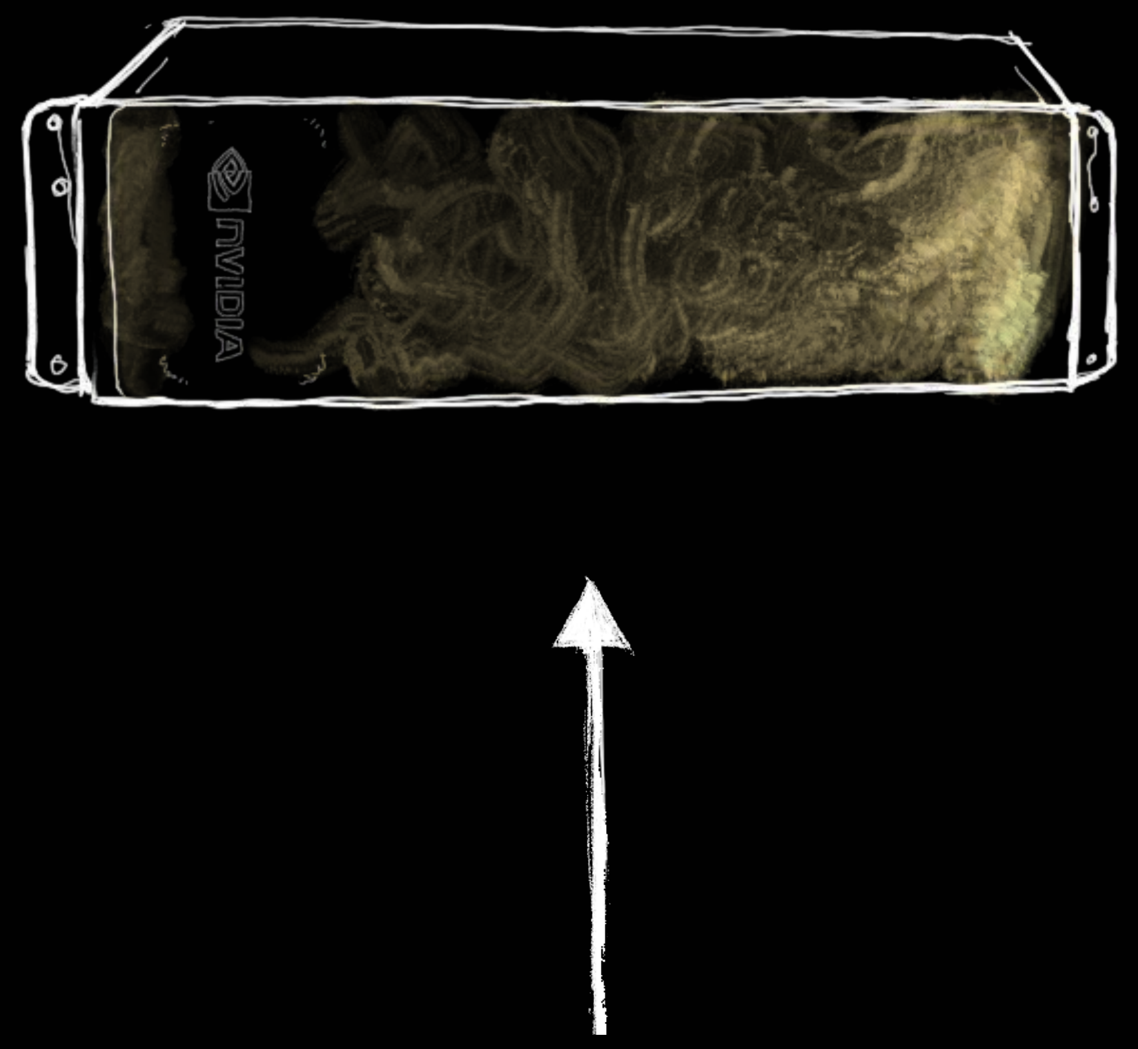


2012
ALEXNET
"FIRST CONTACT"

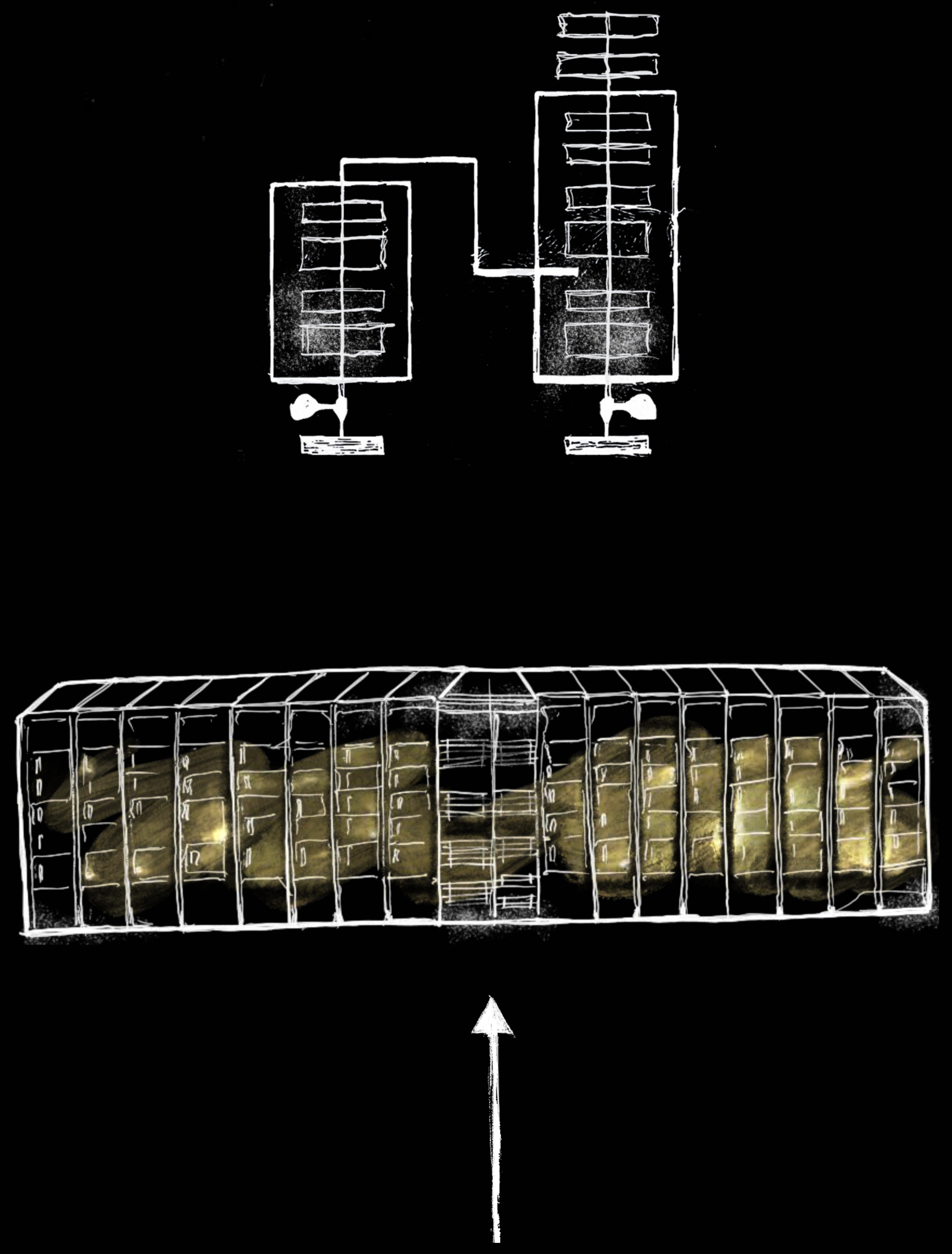


DATA

2016
FIRST DGX DELIVERED



DATA

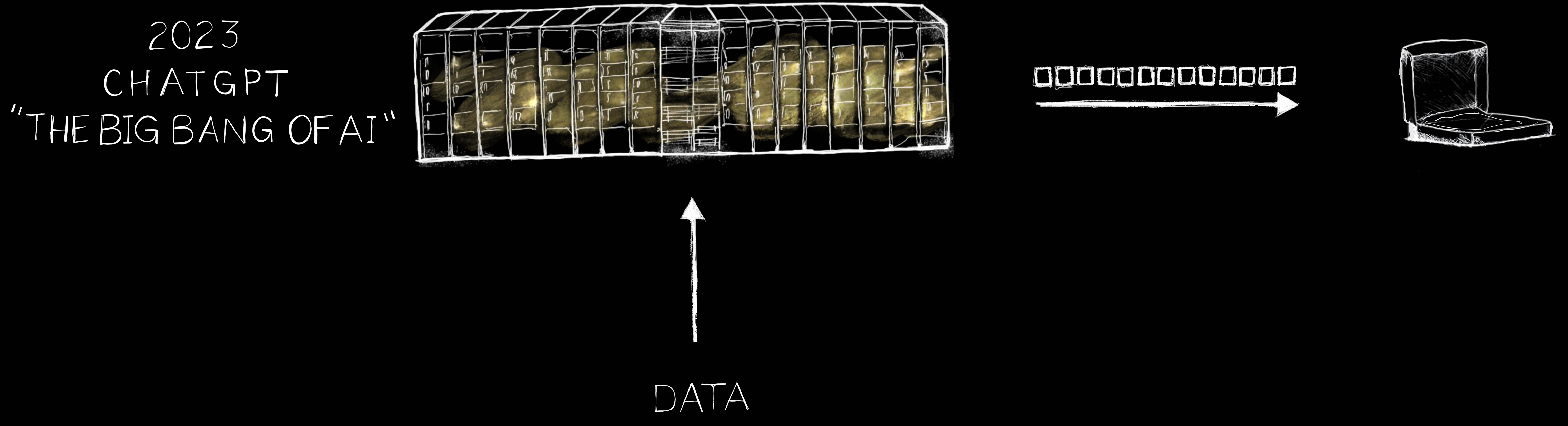


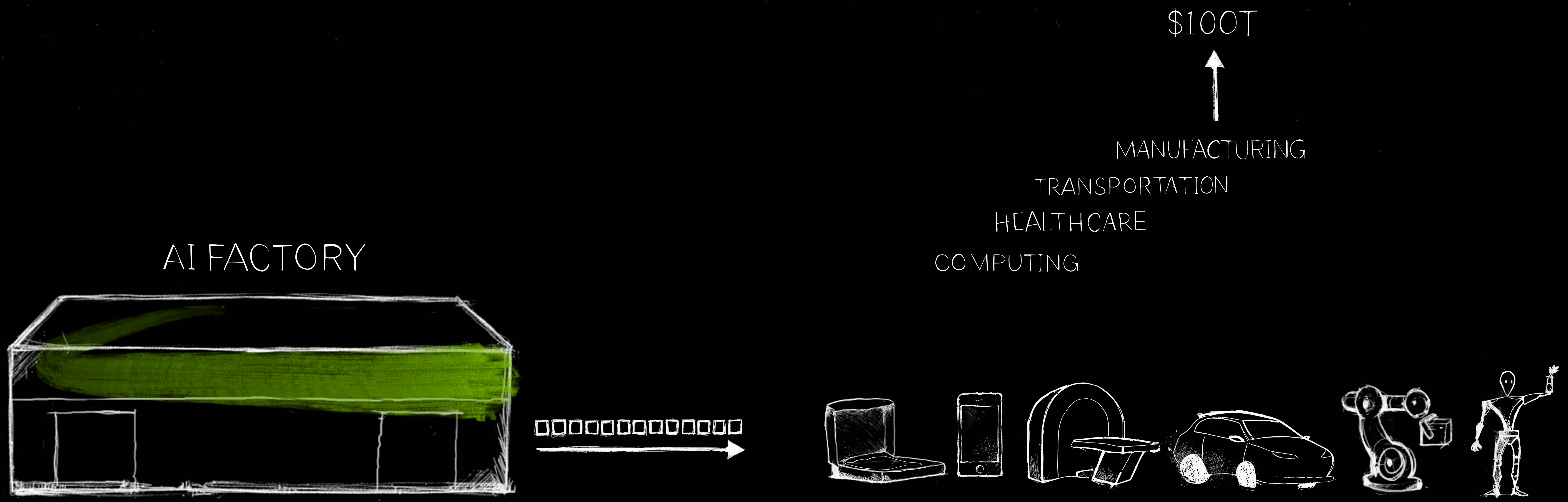
2017
TRANSFORMER

DATA

ChatGPT can make mistakes. Check important info.

LLM
Large Language Model





AI FACTORY

$100T

MANUFACTURING

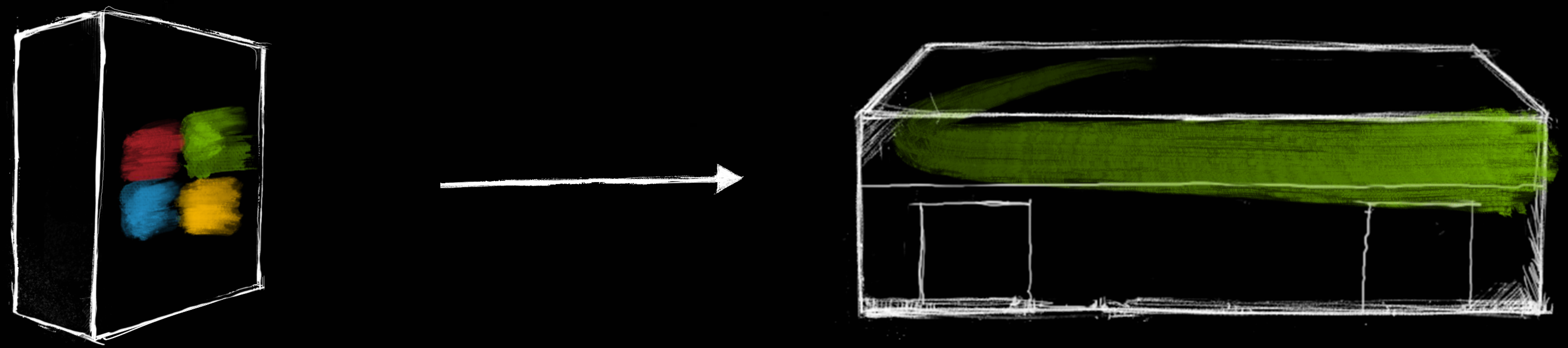
TRANSPORTATION

HEALTHCARE

COMPUTING

新產業革命

"A NEW INDUSTRIAL REVOLUTION"



SOFTWARE FACTORY	AI FACTORY
TOOLS	SKILLS
RETRIEVAL	GENERATIVE
INSTRUCTIONS	LLMS
CPU	**GPU**

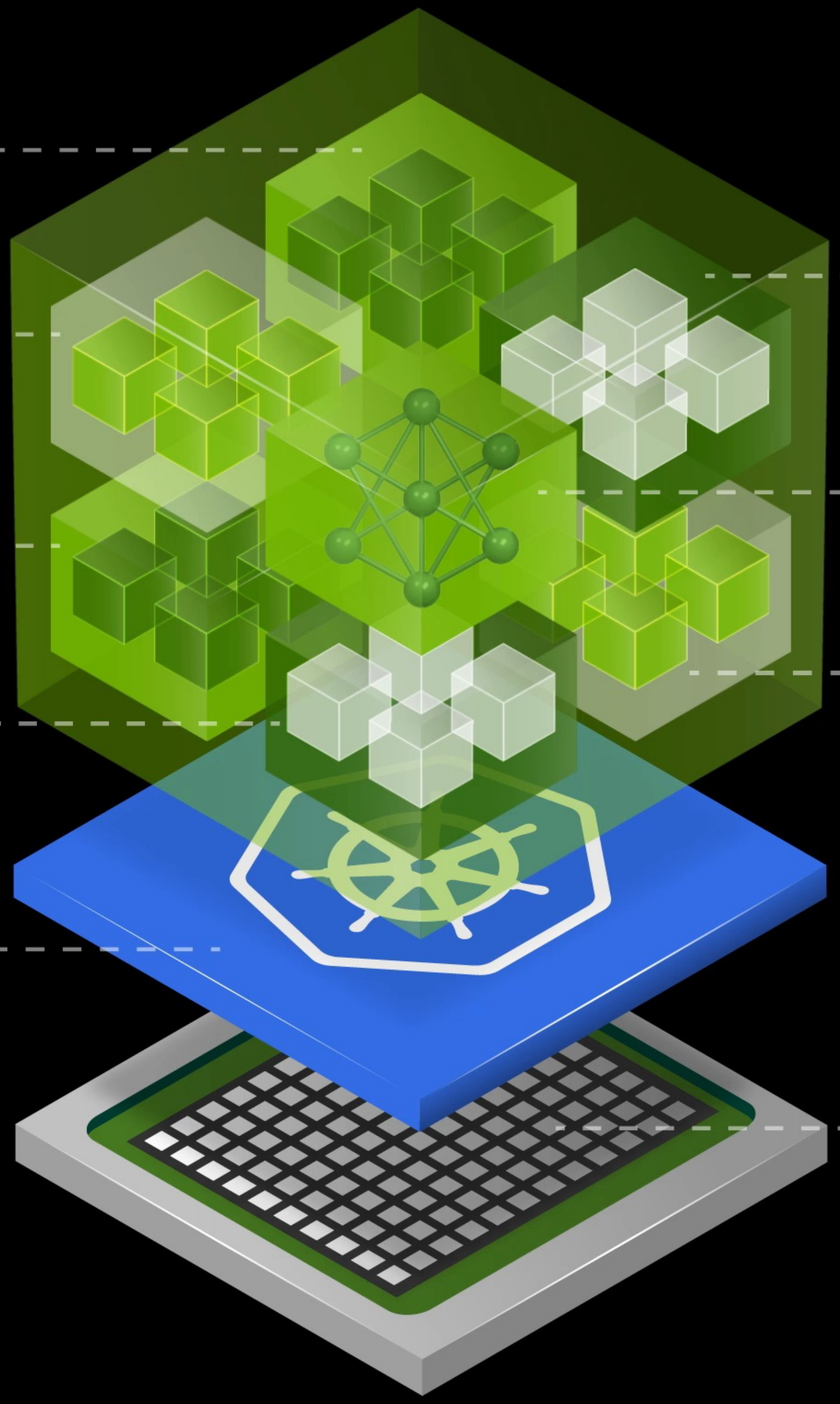
生成式人工智慧推動全棧重塑

GENERATIVE AI DRIVES FULL STACK REINVENTION

NVIDIA INFERENCE MICROSERVICE

Pre-Trained AI Models

Packaged and Optimized to Run Across CUDA Installed Base



Industry-Standard APIs
Text, Speech, Image, Video Biology

Triton Inference Server
cuDF, CV-CUDA, DALI, NCCL,
Post-Processing Decoder

Cloud-Native Stack
GPU Operator, Network Operator

Enterprise Management
Health Check, Identity, Metrics,
Monitoring, Secrets Management

Kubernetes

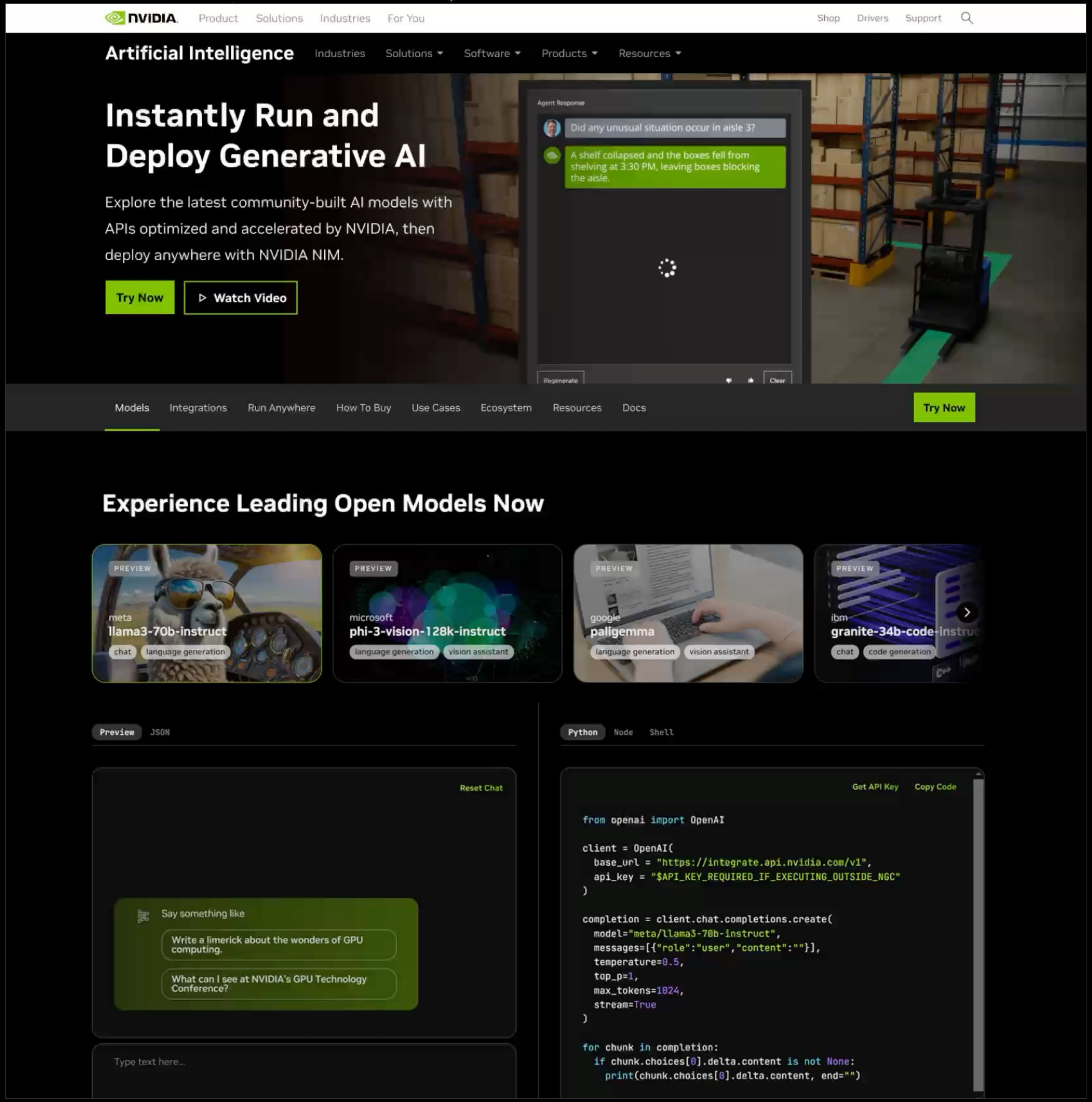
TensorRT-LLM and Triton
cuBLAS, cuDNN, In-Flight Batching,
Memory Optimization, FP8 Quantization

Optimized AI Model
Single GPU, Multi-GPU, Multi-Node

Customization Cache
P-Tuning, LoRA, Model Weights

NVIDIA CUDA

Installed Base of 100s of Millions of CUDA GPUs



Did any unusual situation occur in aisle 3?

A shelf collapsed and the boxes fell from shelving at 3:30 PM, leaving boxes blocking the aisle.

Instantly Run and Deploy Generative AI

Explore the latest community-built AI models with APIs optimized and accelerated by NVIDIA, then deploy anywhere with NVIDIA NIM.

Experience Leading Open Models Now

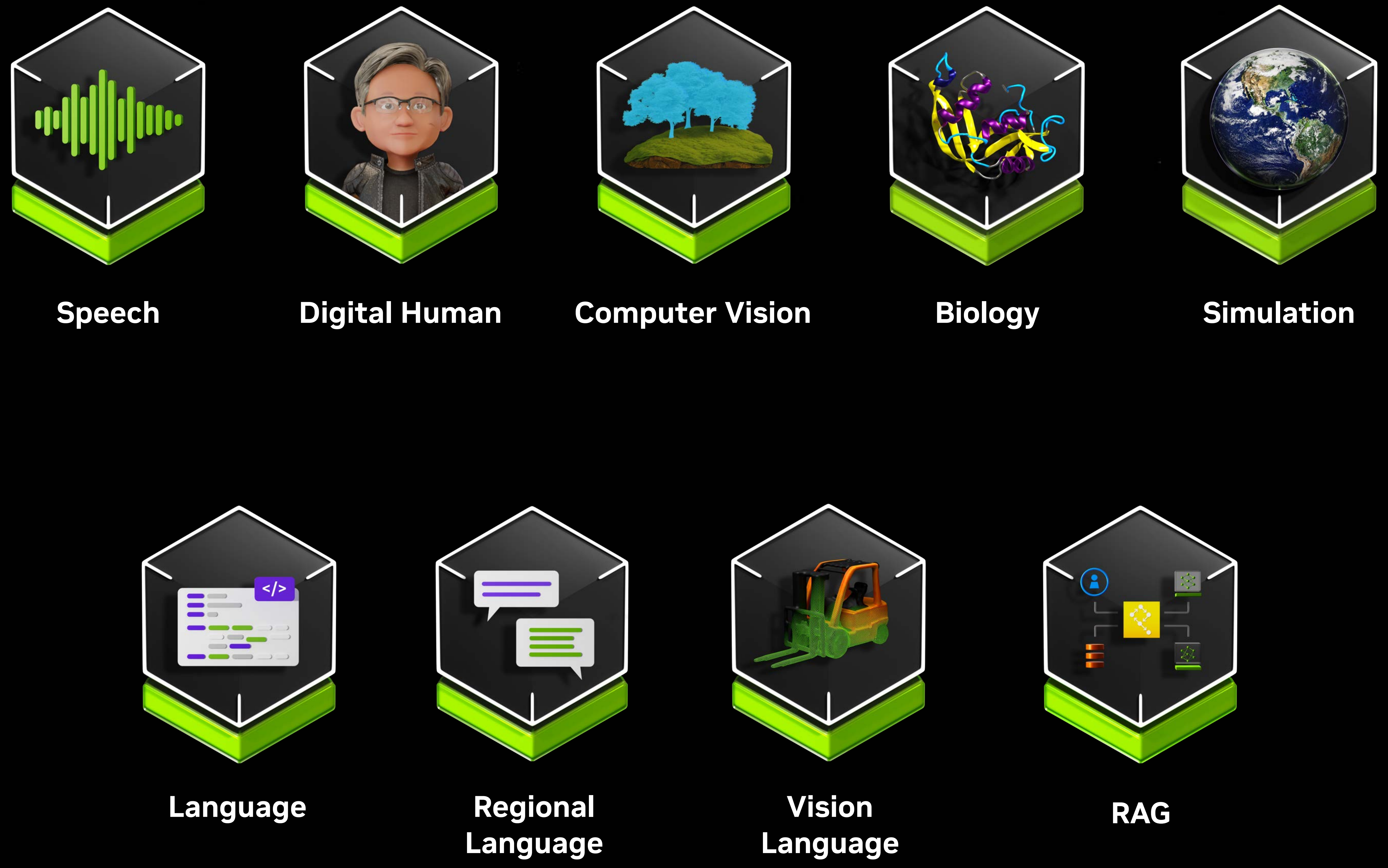
Say something like

Write a limerick about the wonders of GPU computing.

What can I see at NVIDIA's GPU Technology Conference?

```python
from openai import OpenAI

client = OpenAI(
  base_url = "https://integrate.api.nvidia.com/v1",
  api_key = "$API_KEY_REQUIRED_IF_EXECUTING_OUTSIDE_NGC"
)

completion = client.chat.completions.create(
  model="meta/llama3-70b-instruct",
  messages=[{"role":"user","content":""}],
  temperature=0.5,
  top_p=1,
  max_tokens=1024,
  stream=True
)

for chunk in completion:
  if chunk.choices[0].delta.content is not None:
    print(chunk.choices[0].delta.content, end="")
```

AI.NVIDIA.COM



Speech

Digital Human

Computer Vision

Biology

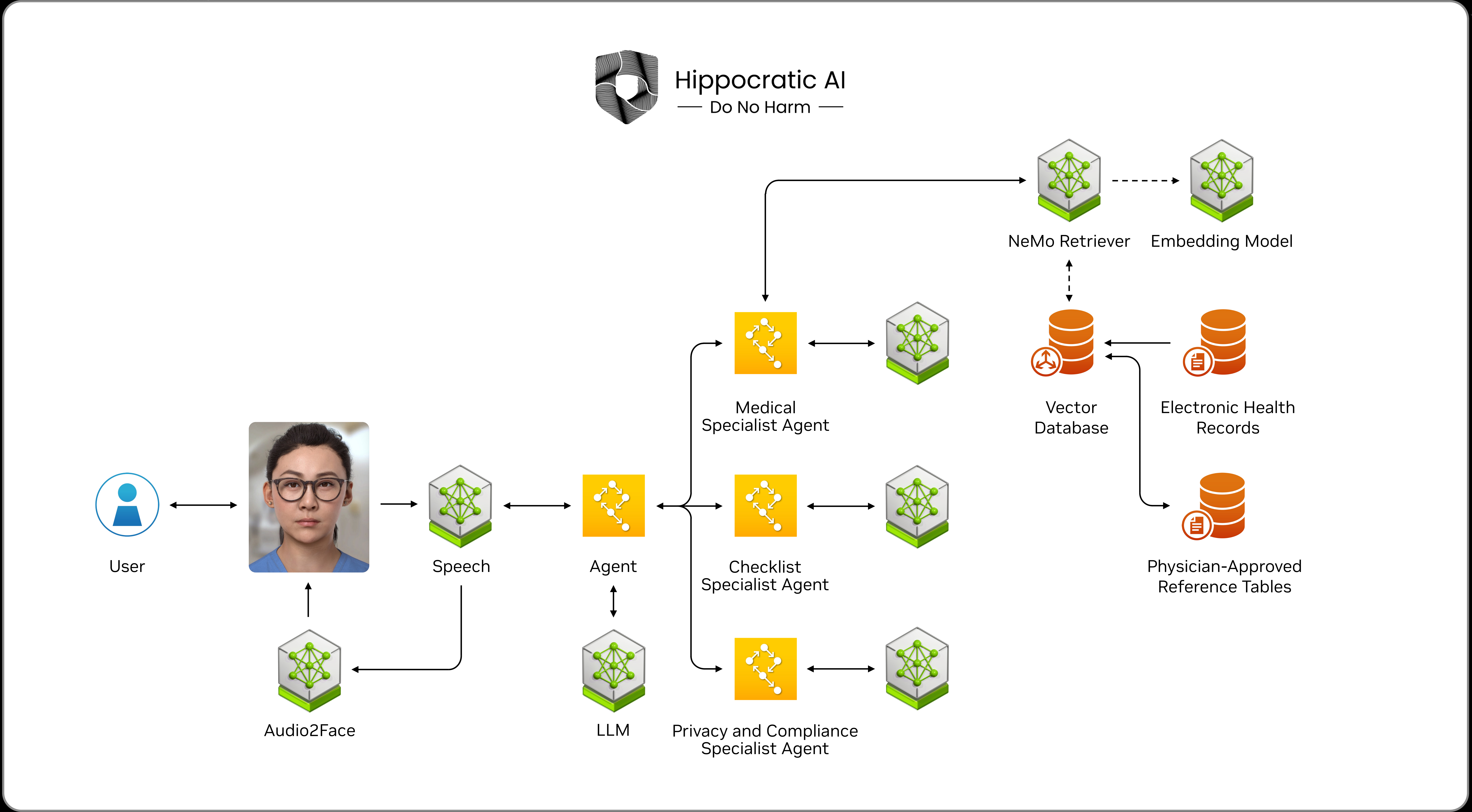
Simulation

Language

Regional Language

Vision Language

RAG

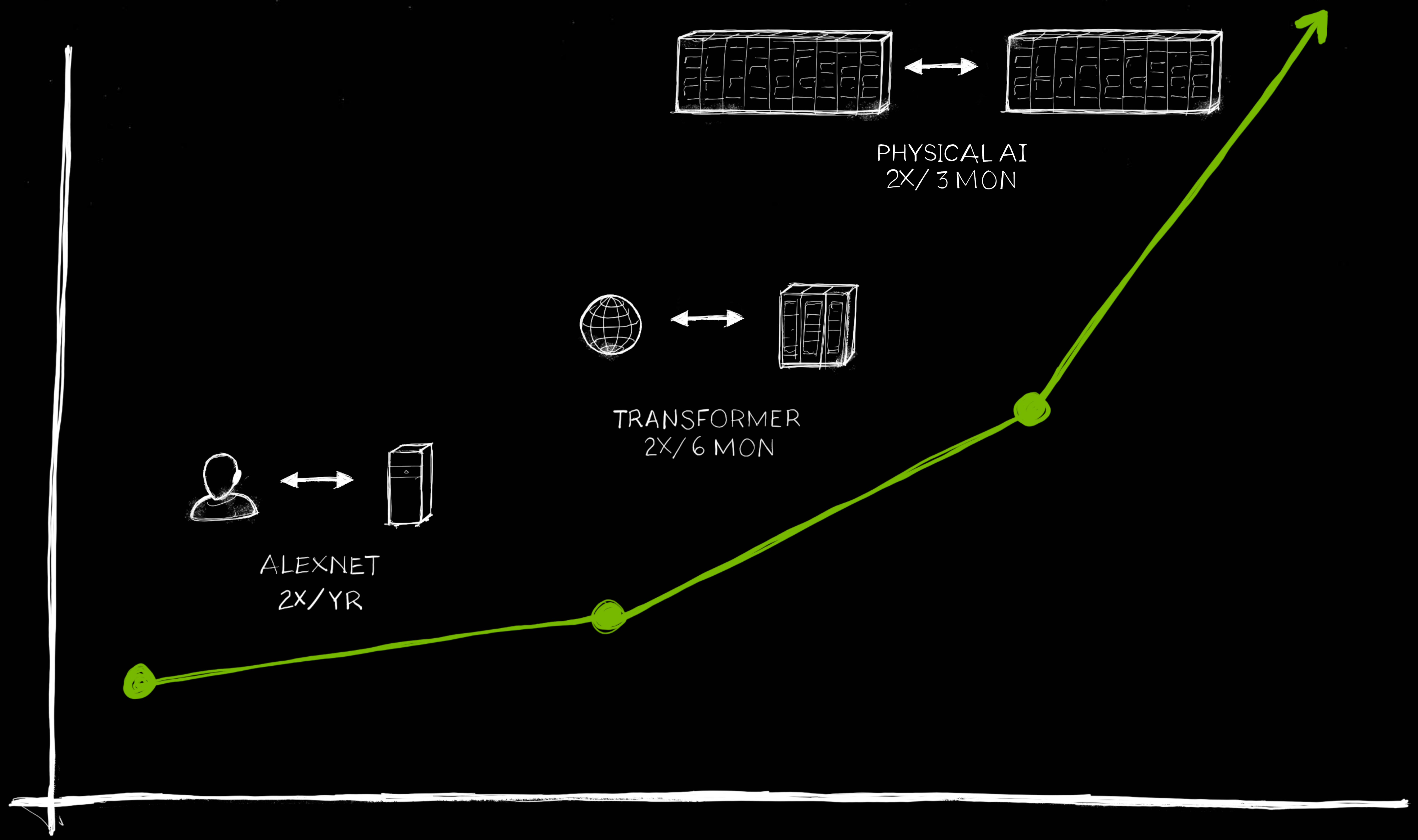

Hippocratic AI
Do No Harm
NeMo Retriever
Embedding Model
Medical Specialist Agent
Vector Database
Electronic Health Records
User
Speech
Agent
Checklist Specialist Agent
Physician-Approved Reference Tables
Audio2Face
LLM
Privacy and Compliance Specialist Agent

生成式人工智慧實現數位護士、客服、導師等

GENERATIVE AI ENABLES DIGITAL NURSE, CUSTOMER SERVICE AGENTS, TUTORS, ETC

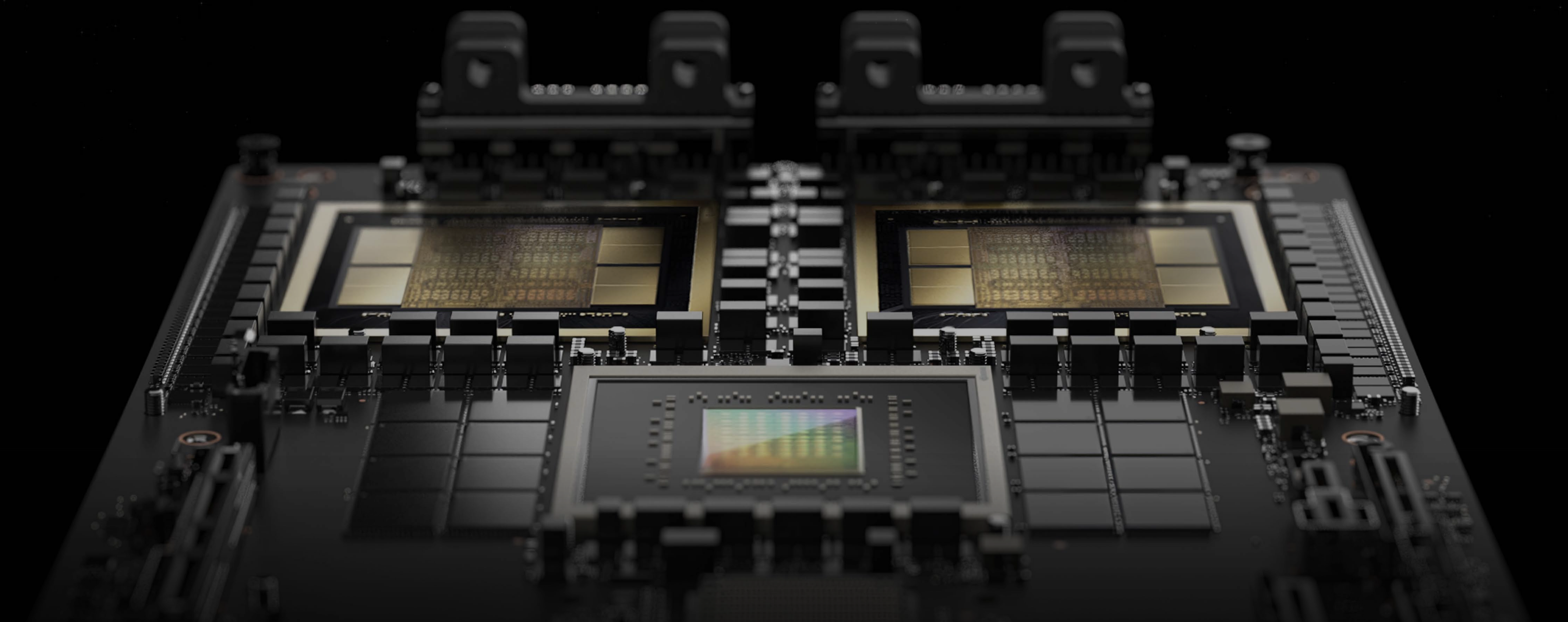
發表新款 RTX AI 電腦
現已超過 200 款 RTX AI 筆記型電腦 | 高達 700 AI TOPS | 7X 生成式人工智慧

ANNOUNCING NEW RTX AI PCs
Now Over 200 RTX AI Laptops | Up to 700 AI TOPS | 7X Generative AI



運算模型規模呈指數級成長

MODEL COMPUTE SCALE GROWS EXPONENTIALLY



AI SUPERCHIP
208B Transistors

TRANSFORMER ENGINE
FP4/FP6 Tensor Core

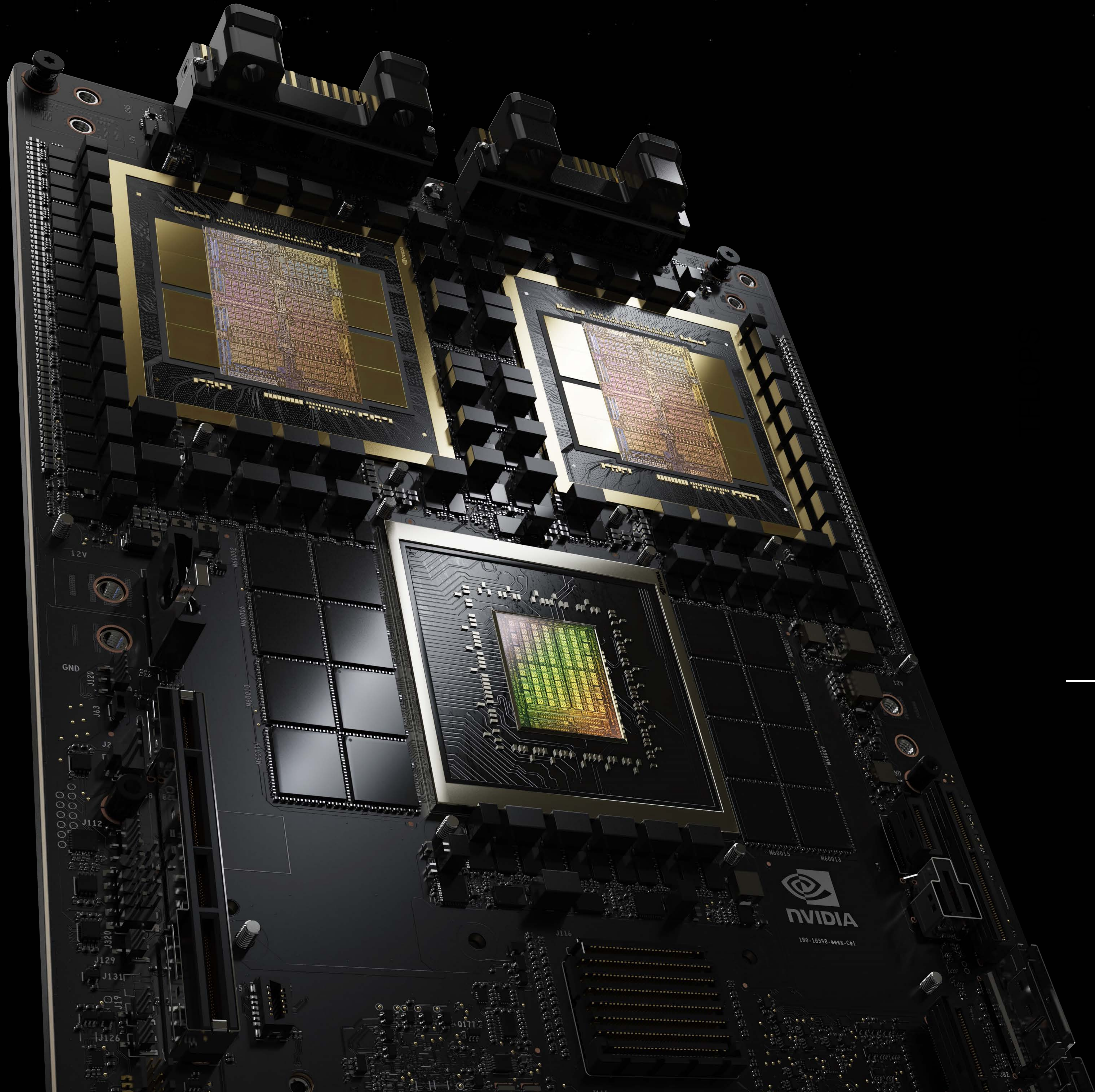
SECURE AI
Full Performance
Encryption and TEE

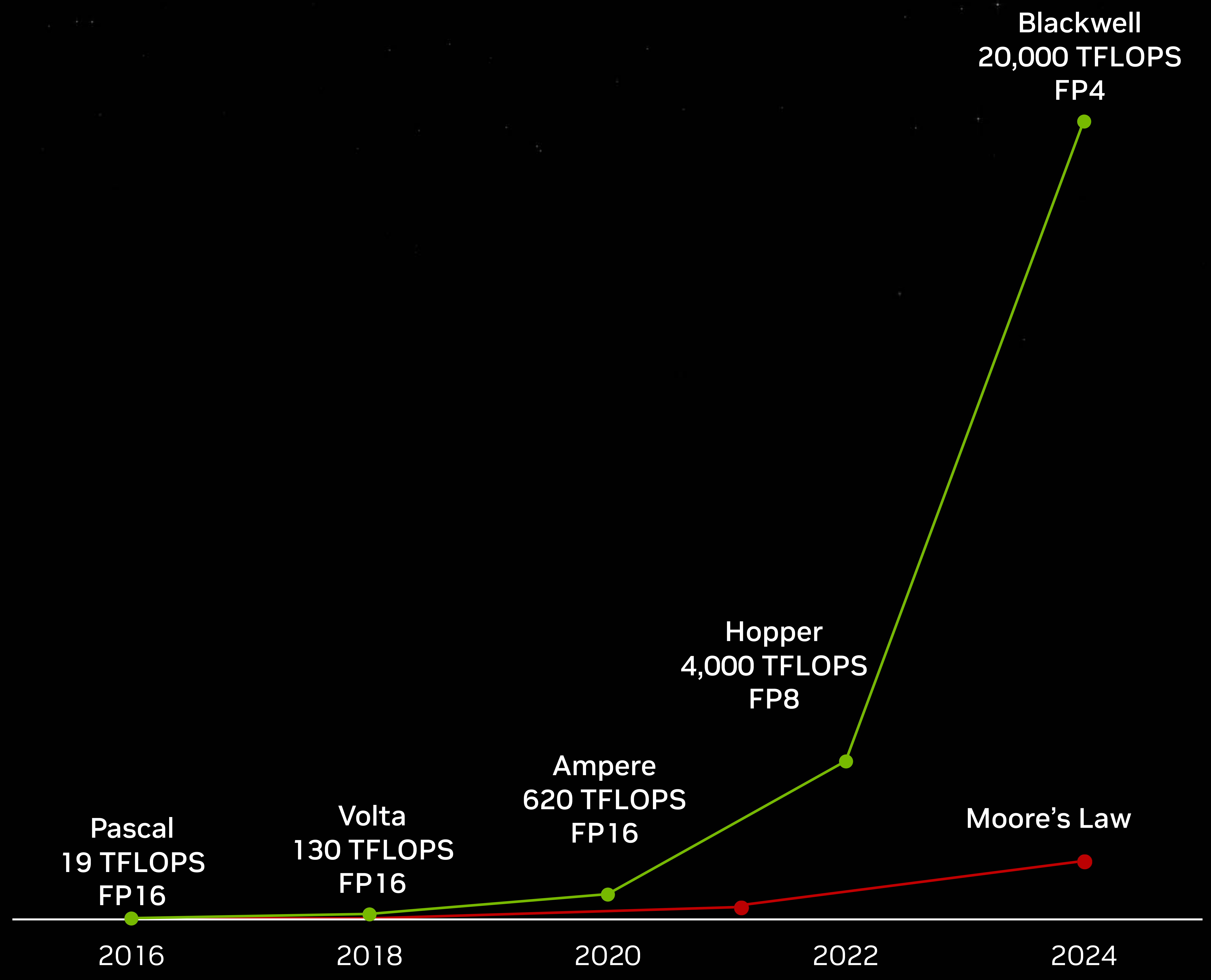
5th GENERATION NVLINK
Scales to 576 GPUs

RAS ENGINE
100% In-System Self-Test

DECOMPRESSION ENGINE
800 GB/sec

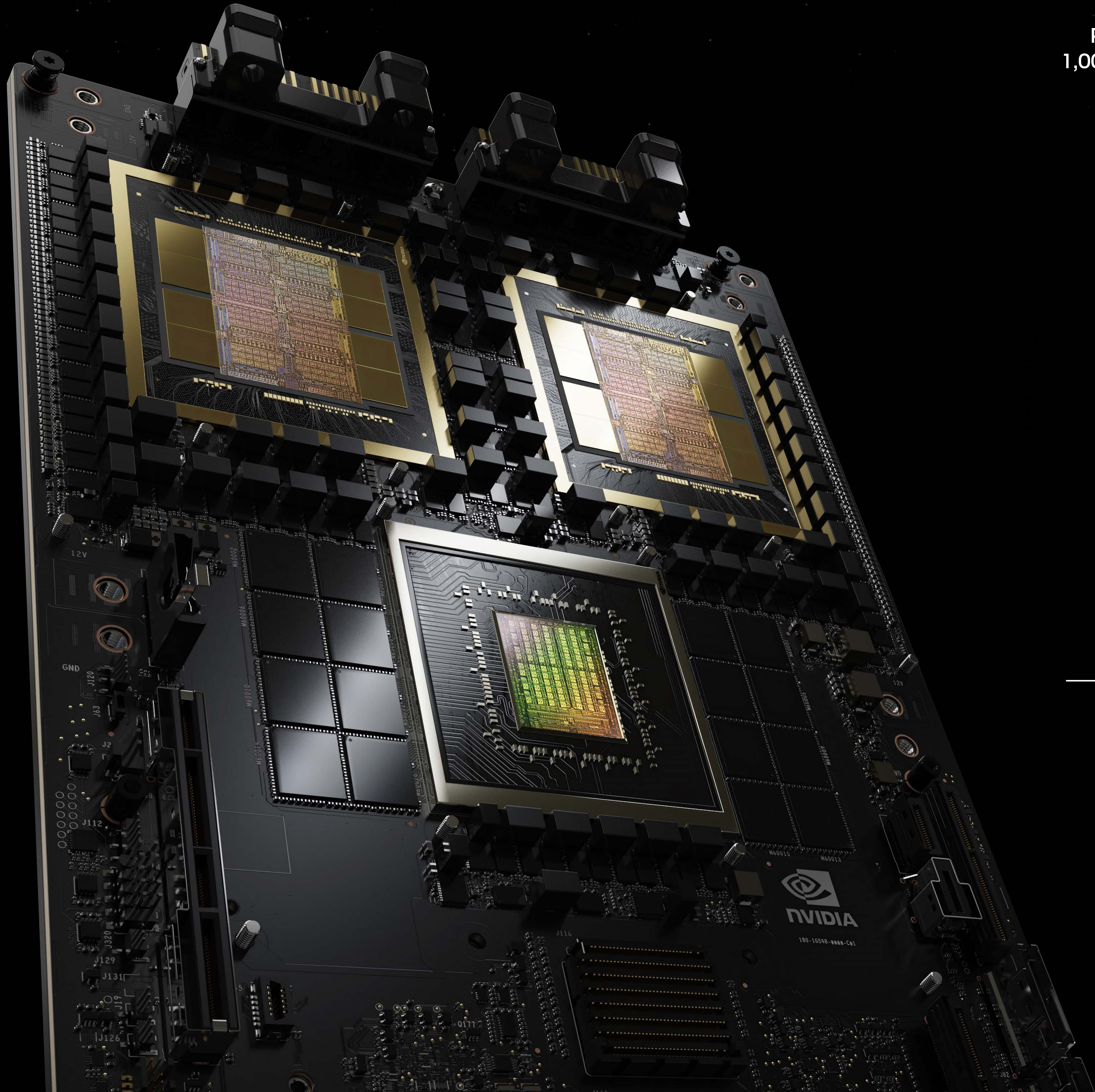
NVIDIA BLACKWELL 平台
兆級參數規模的生成式人工智慧

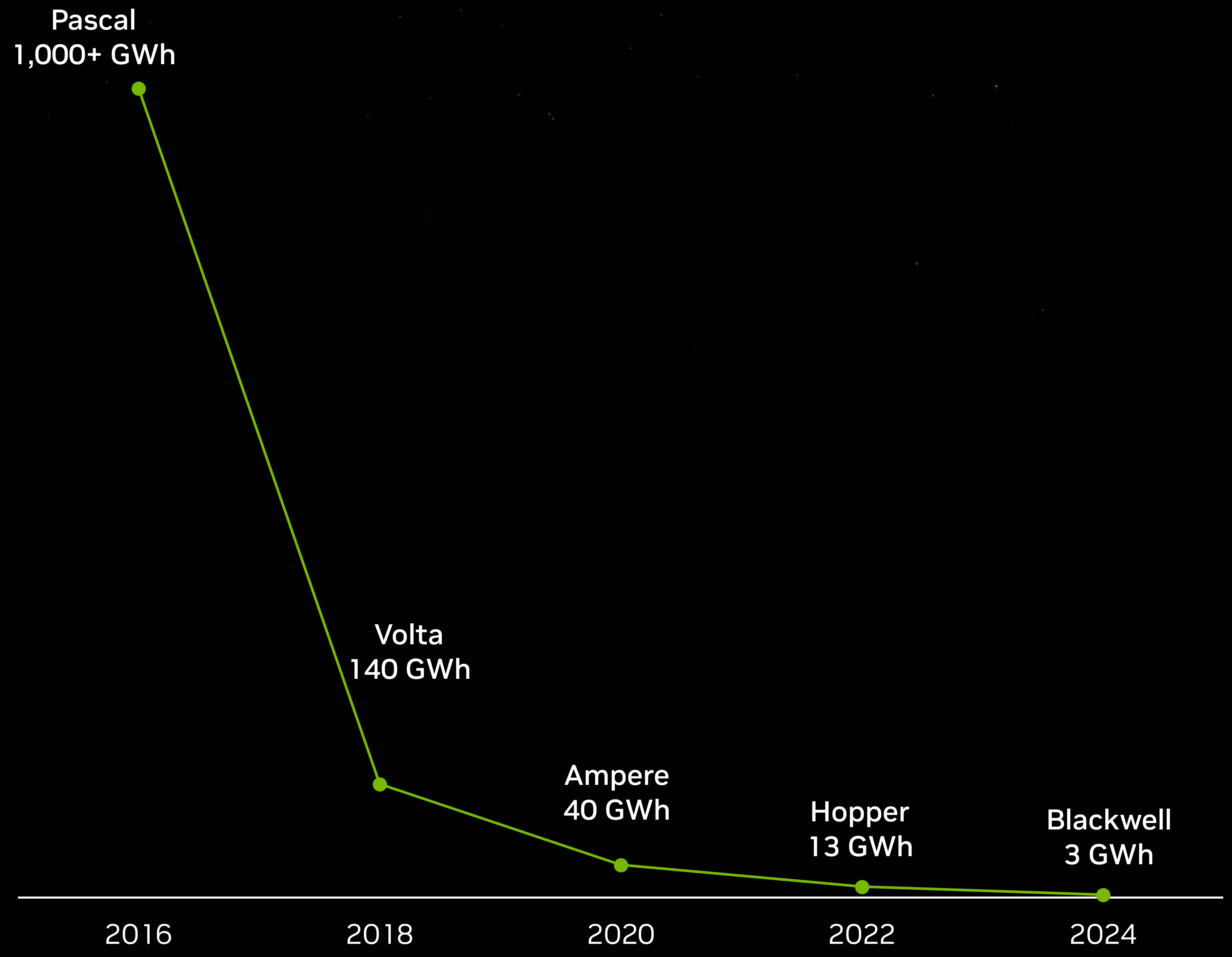




8 年內 1,000X 的人工智慧運算

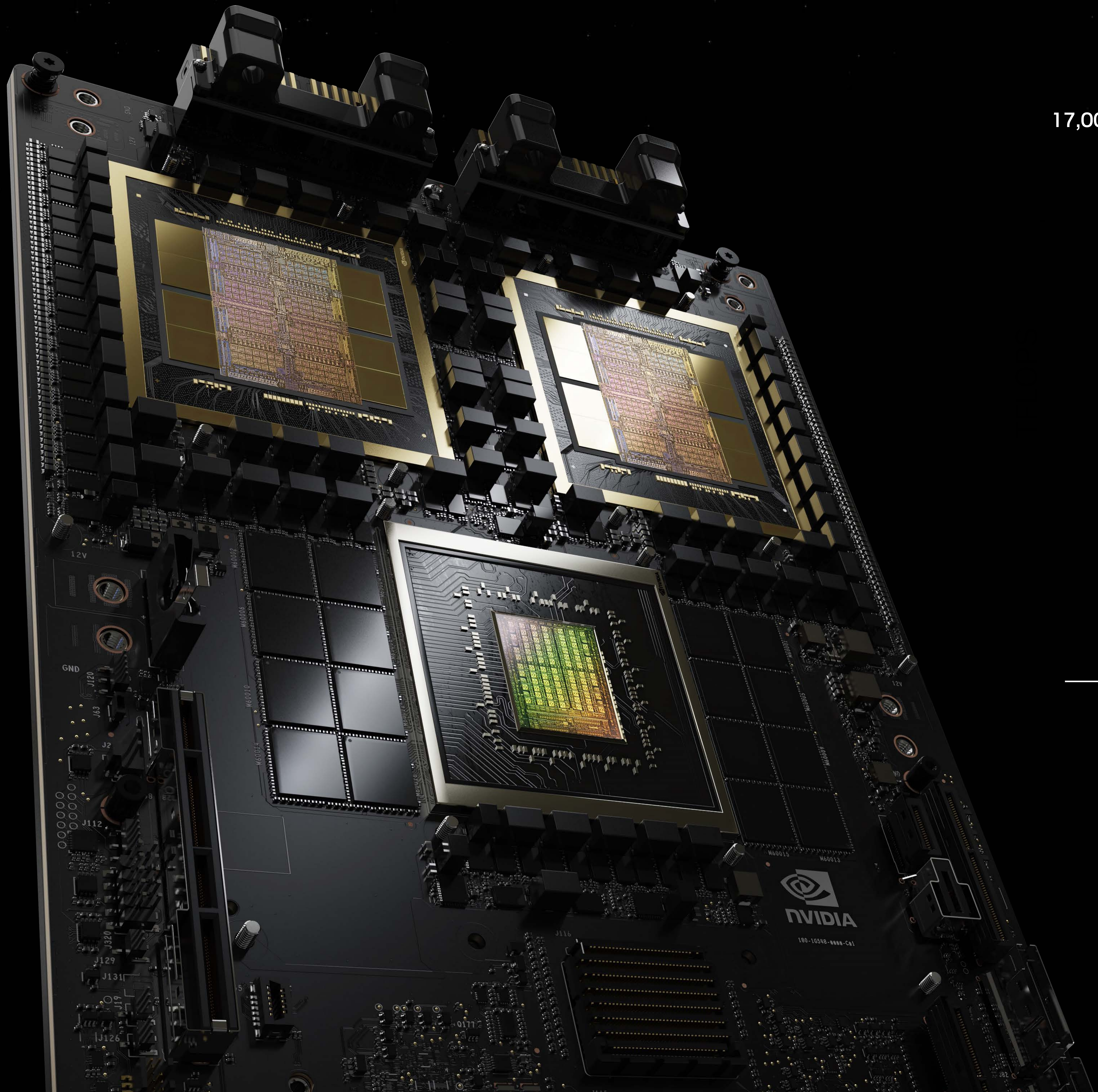
1,000X AI COMPUTE IN 8 YEARS

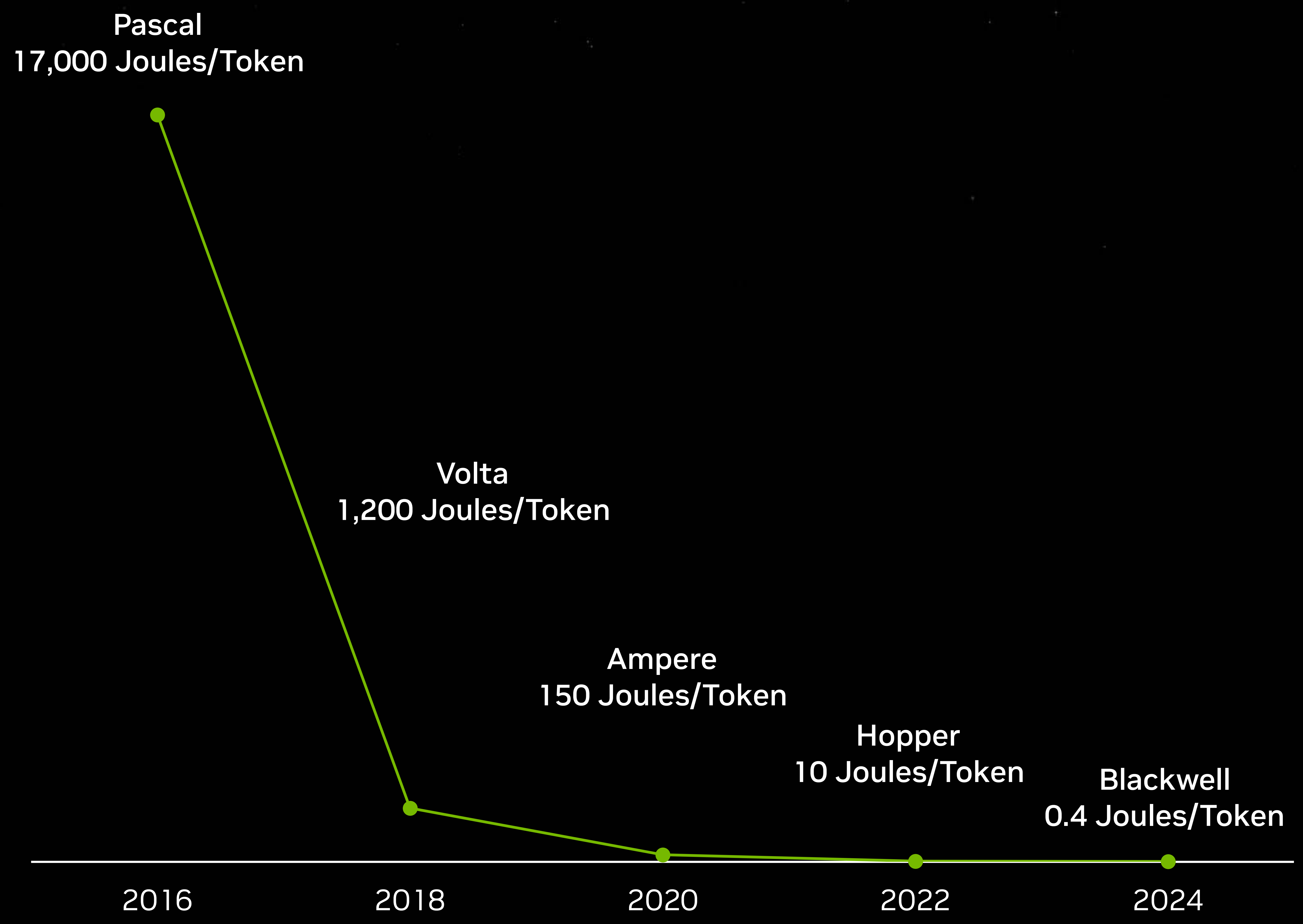




8 年內節能至 1/350
訓練 GPT4-1.8T

350X ENERGY REDUCTION IN 8 YEARS
TO TRAIN GPT4-1.8T

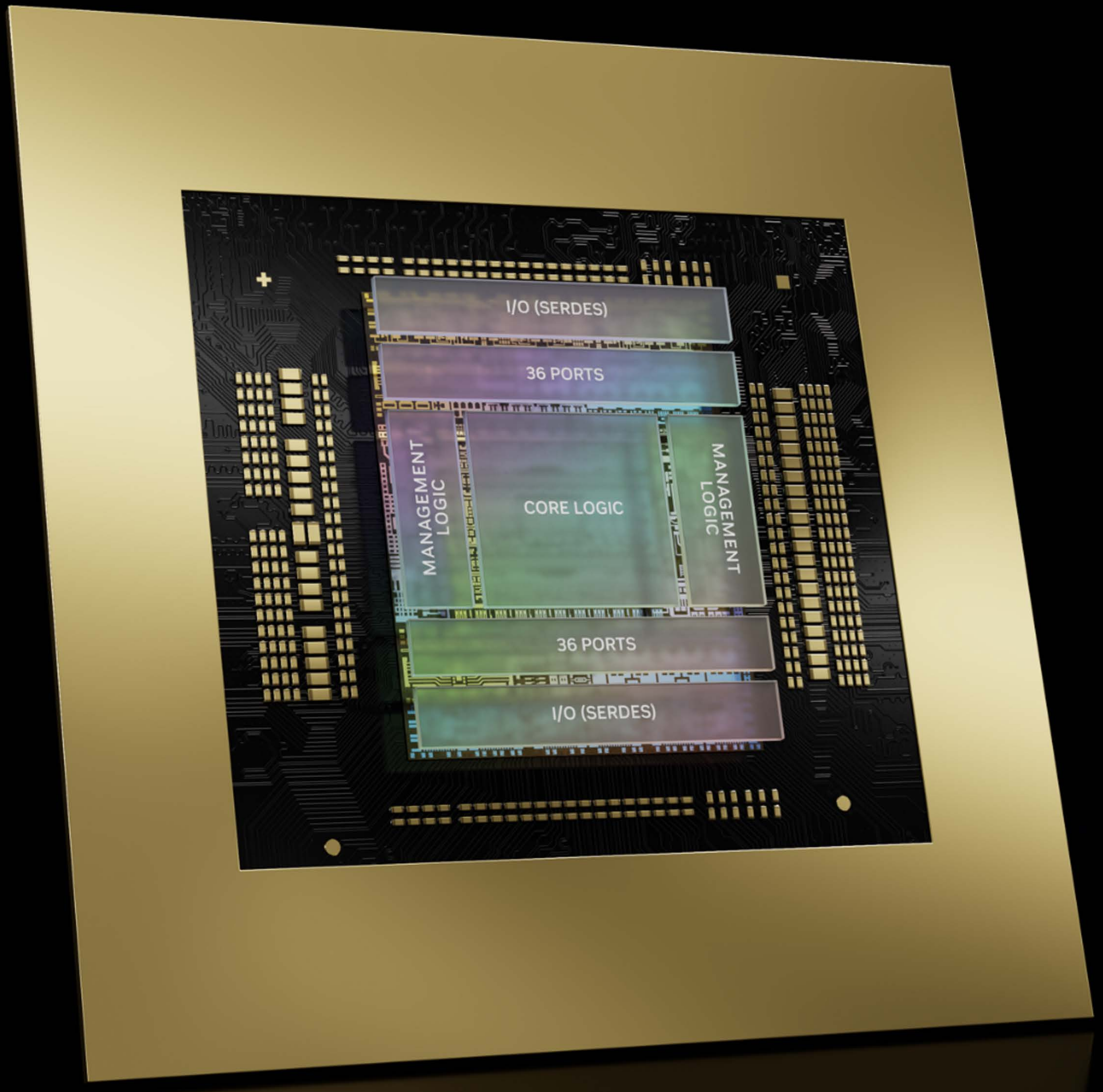




8 年內節能至 1/45,000
生成 GPT4-1.8T 詞元

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

	DGX BLACKWELL	DGX HOPPER	
NVLink Domain	72	8	9X
NVLink BW (TB/s)	130	7	18X
AI FLOPS (PF)	1,440	32	45X
Power (kW)	100	10	10X



NVLink Switch Chip

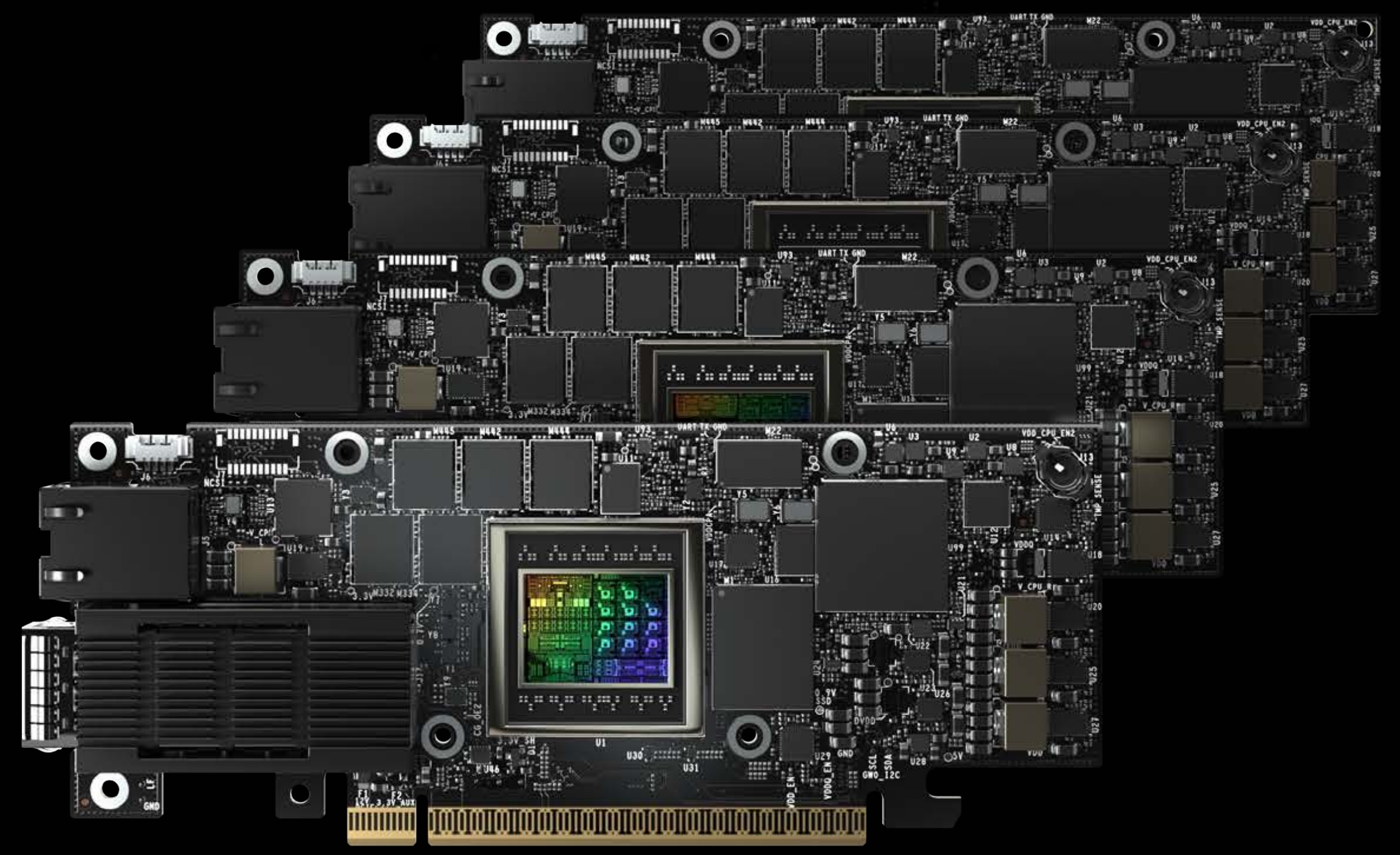
50B Transistors in TSMC 4NP

72-Ports 400G SerDes

4 NVLinks at 1.8TB/sec

7.2TB/sec Full-Duplex Bandwidth

SHARP In-Network Compute – 3.6 TFLOPS FP8

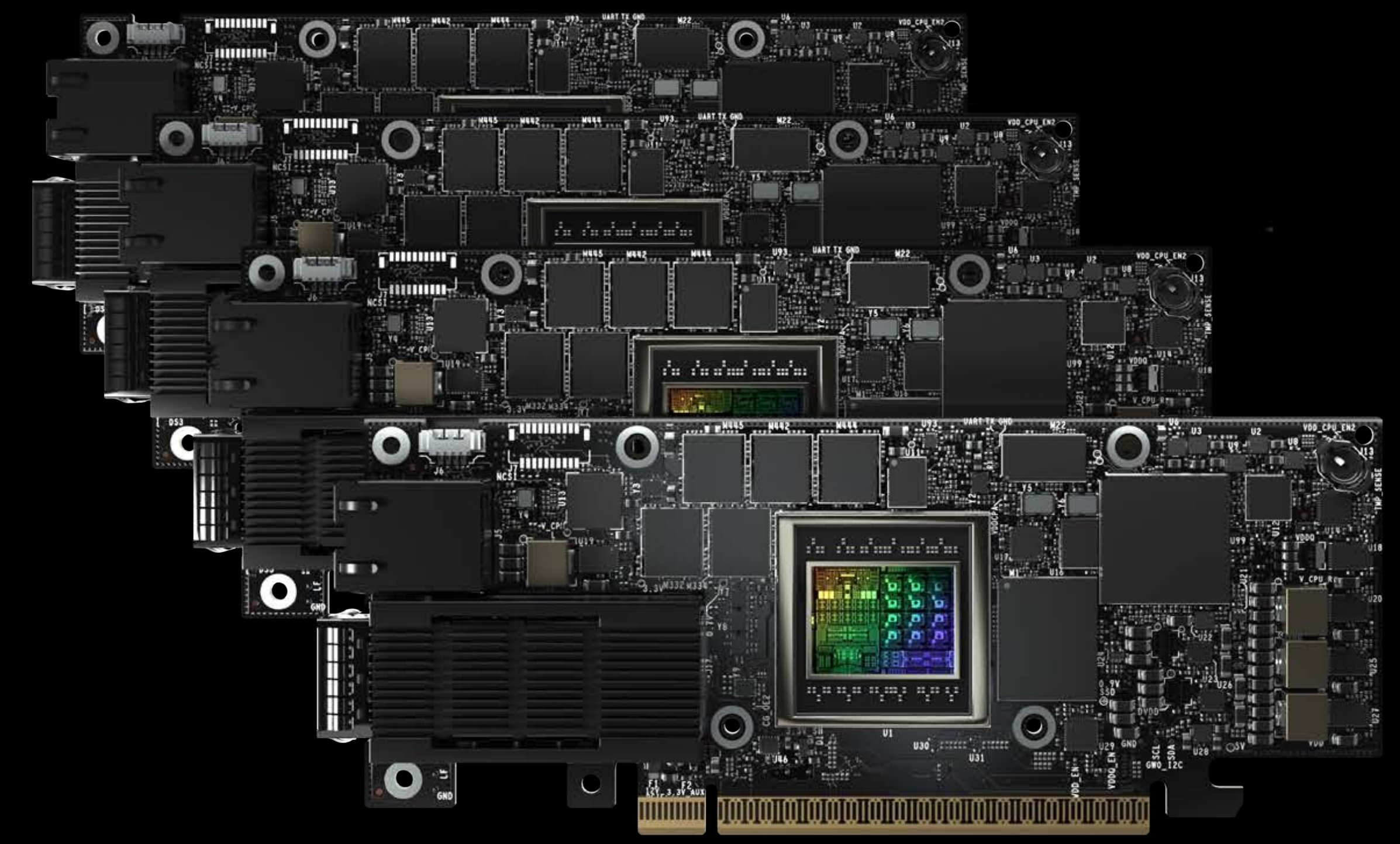


BlueField-3
400G SuperNIC

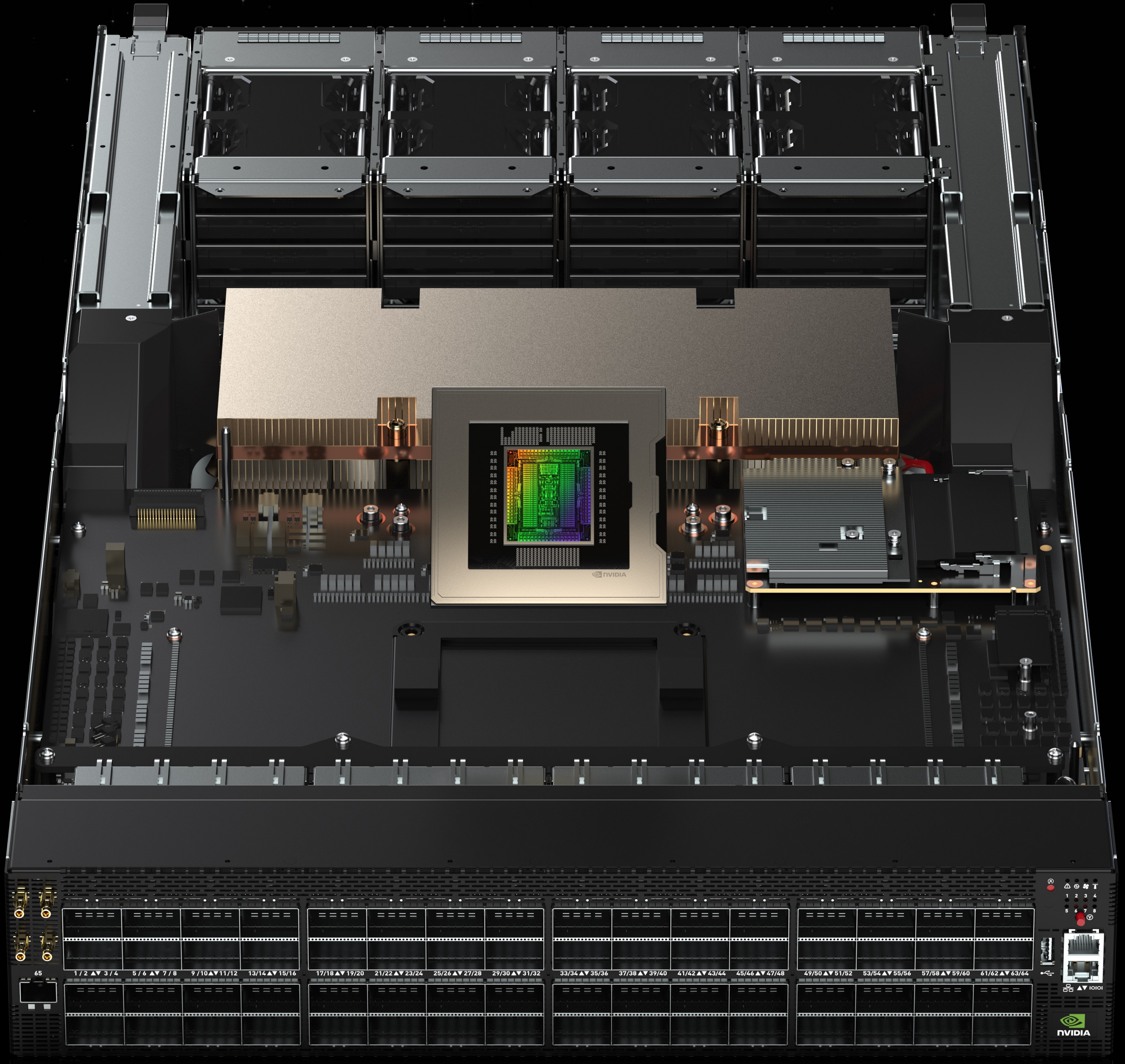
Spectrum-X800
Ethernet Switch

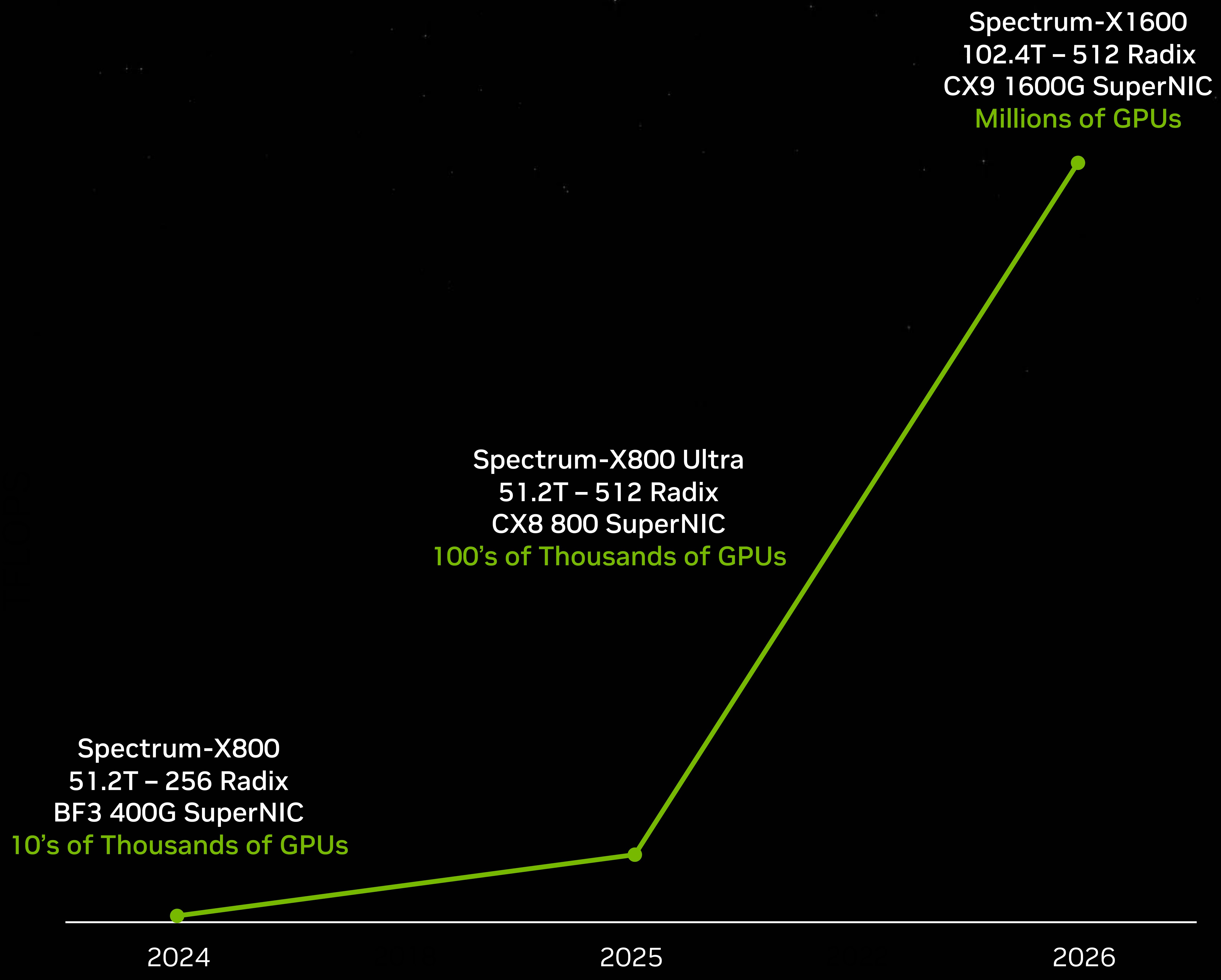
1.6X

Effective
Bandwidth



NVIDIA SPECTRUM-X
為生成式人工智慧推進乙太網路
Supercharging Ethernet for Generative AI





年度 SPECTRUM-X 節奏
擴展到數百萬個 GPU

ANNUAL SPECTRUM-X RHYTHM
SCALING TO MILLIONS OF GPUS

46379.1
46379.1
46379.1

aws
Google Cloud
Microsoft Azure
ORACLE CLOUD Infrastructure

ADEPT
AI21labs
Character.AI
cohere
essential AI
Hugging Face
Inflection

Meta
MISTRAL AI_
OpenAI
perplexity
Recursion
together.ai
X

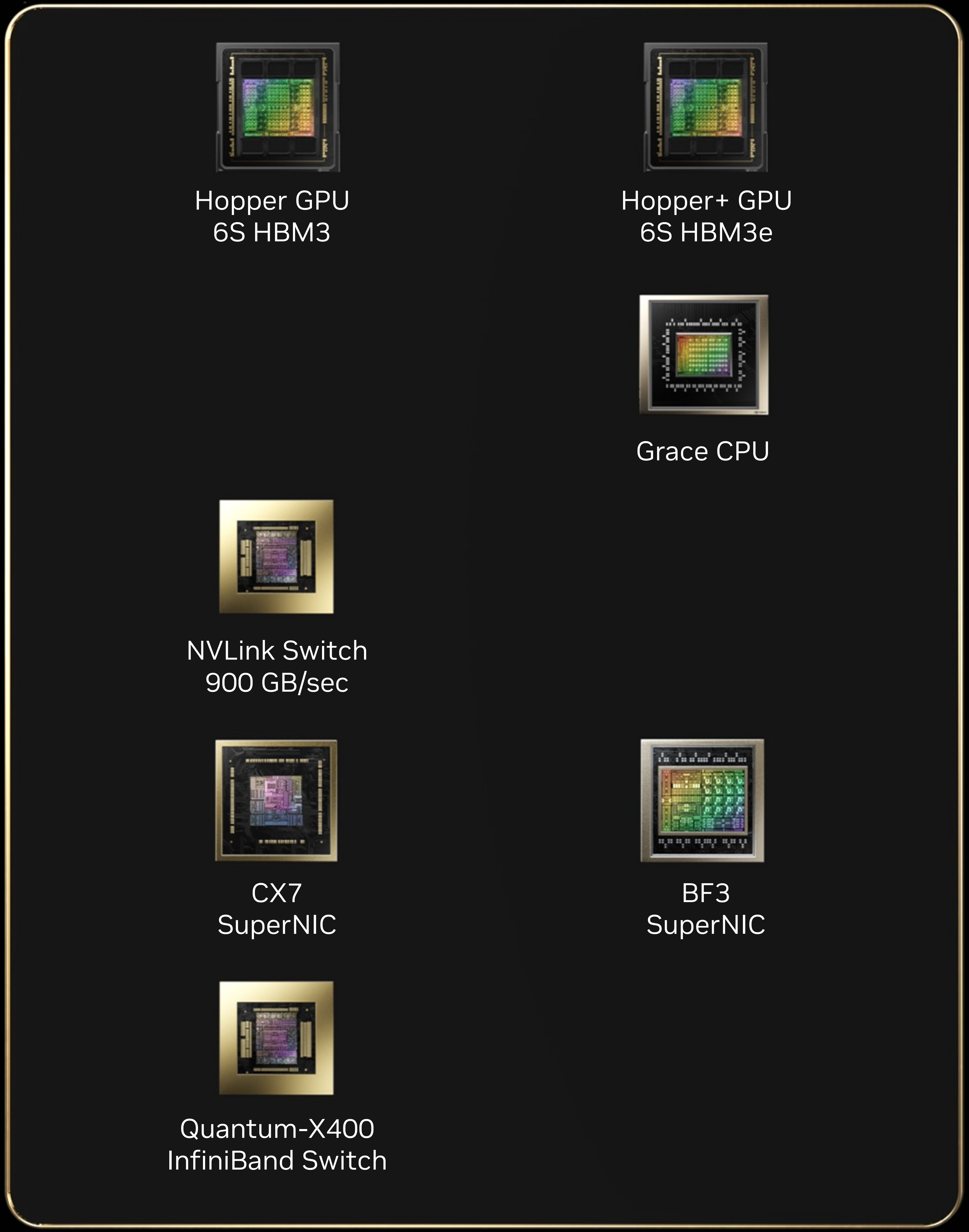
AIVRES
APPLIED DIGITAL
ASRock Rack
ASUS
CISCO
CoreWeave
Crusoe
DELL Technologies

EVIDEN
FOXCONN HON HAI TECHNOLOGY GROUP
FUJITSU
GIGABYTE
Hewlett Packard Enterprise
IBM Cloud
indosat OOREDOO HUTCHISON
Inventec

Lambda
Lenovo
NEXGEN CLOUD
NORTHERN DATA GROUP
PEGATRON
QCT
Scaleway
Singtel

SoftBank
SUPERMICRO
wistron
wiwynn
YOTTA
YTL COMMUNICATIONS
zt Systems

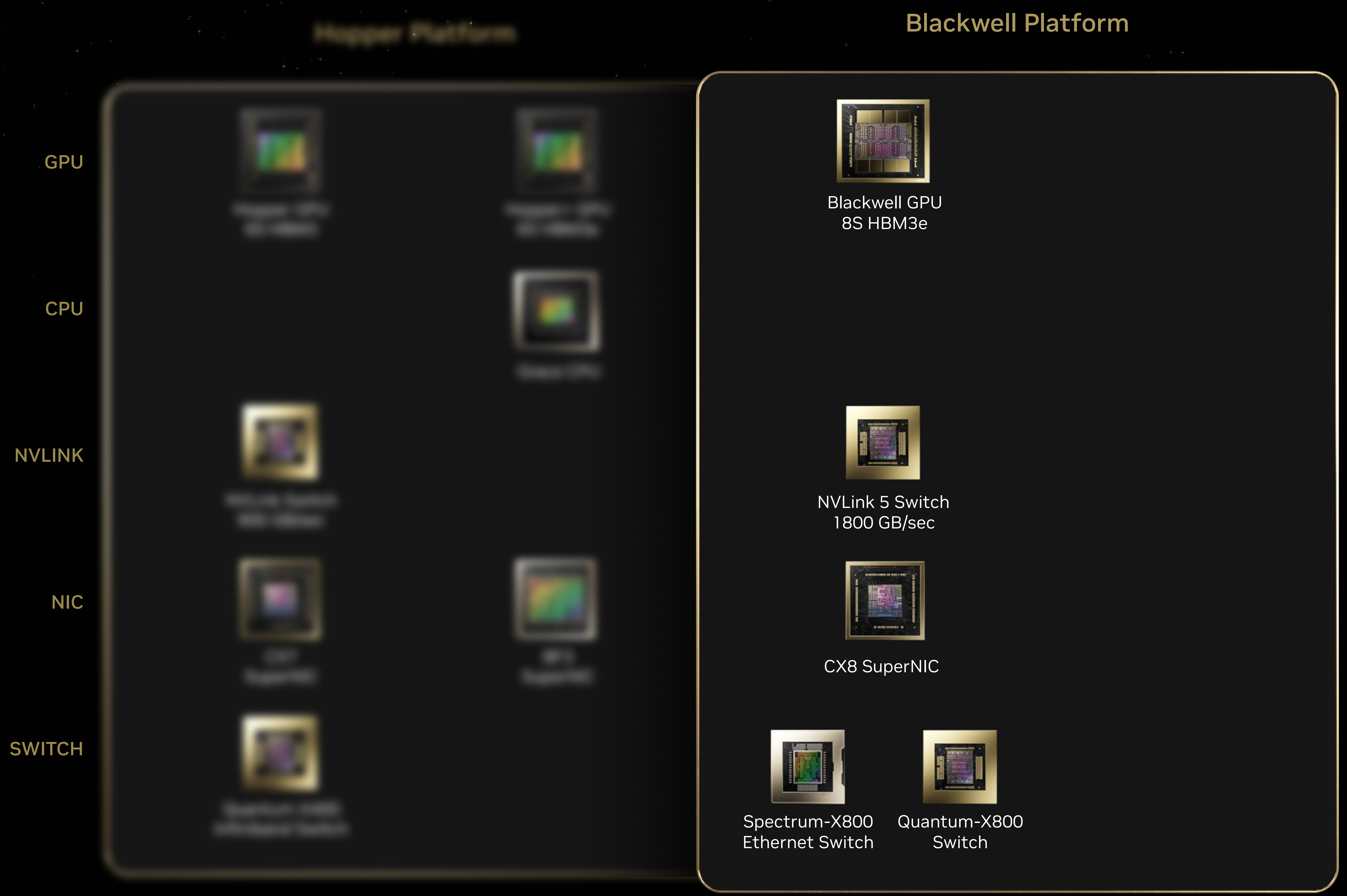
Hopper Platform



GPU	Hopper GPU 6S HBM3	Hopper+ GPU 6S HBM3e
CPU		Grace CPU
NVLINK	NVLink Switch 900 GB/sec	
NIC	CX7 SuperNIC	BF3 SuperNIC
SWITCH	Quantum-X400 InfiniBand Switch	

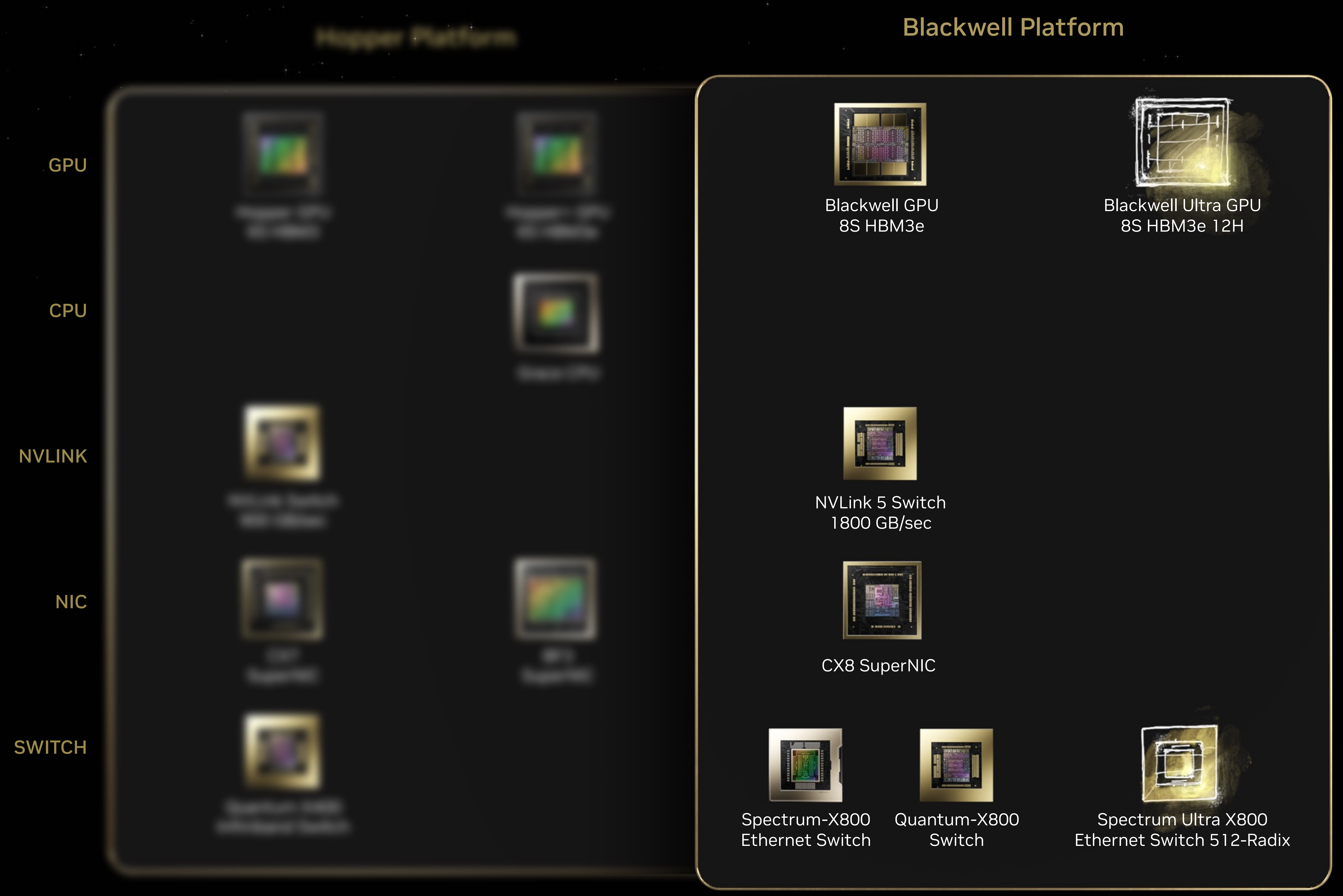
資料中心規模 • 一年節奏 • 技術限制 • 一個架構

DATACENTER SCALE • ONE-YEAR RHYTHM • TECHNOLOGY LIMITS • ONE ARCHITECTURE



資料中心規模 • 一年節奏 • 技術限制 • 一個架構

DATACENTER SCALE • ONE-YEAR RHYTHM • TECHNOLOGY LIMITS • ONE ARCHITECTURE



資料中心規模 • 一年節奏 • 技術限制 • 一個架構

DATACENTER SCALE • ONE-YEAR RHYTHM • TECHNOLOGY LIMITS • ONE ARCHITECTURE

Hopper Platform

Blackwell Platform

Rubin Platform

GPU

CPU

NVLINK

NIC

SWITCH

Rubin GPU
8S HBM4

Rubin Ultra GPU
12S HBM4

Vera CPU

NVLink 6 Switch
3600 GB/sec

CX9 SuperNIC
1600 Gb/sec

X1600
IB/Ethernet Switch

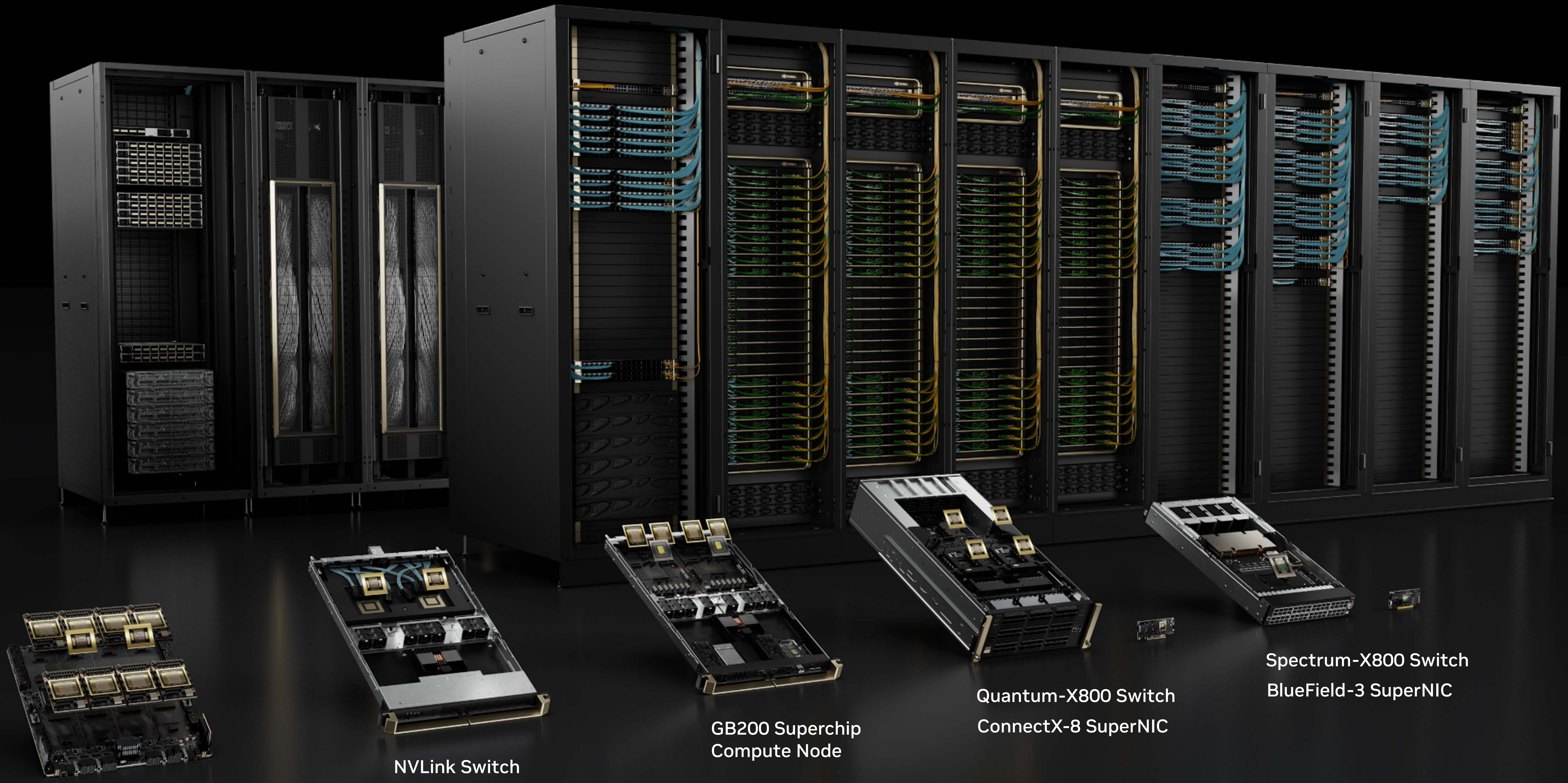
資料中心規模 • 一年節奏 • 技術限制 • 一個架構

DATACENTER SCALE • ONE-YEAR RHYTHM • TECHNOLOGY LIMITS • ONE ARCHITECTURE

	Hopper Platform		Blackwell Platform		Rubin Platform	
GPU	Hopper GPU 6S HBM3	Hopper+ GPU 6S HBM3e	Blackwell GPU 8S HBM3e	Blackwell Ultra GPU 8S HBM3e 12H	Rubin GPU 8S HBM4	Rubin Ultra GPU 12S HBM4
CPU		Grace CPU			Vera CPU	
NVLINK	NVLink Switch 900 GB/sec		NVLink 5 Switch 1800 GB/sec		NVLink 6 Switch 3600 GB/sec	
NIC	CX7 SuperNIC	BF3 SuperNIC	CX8 SuperNIC		CX9 SuperNIC 1600 Gb/sec	
SWITCH	Quantum-X400 Infiniband Switch		Spectrum-X800 Ethernet Switch Quantum-X800 Switch	Spectrum Ultra X800 Ethernet Switch 512-Radix	X1600 IB/Ethernet Switch	
	2022	2023	2024	2025	2026	2027

資料中心規模 • 一年節奏 • 技術限制 • 一個架構

DATACENTER SCALE • ONE-YEAR RHYTHM • TECHNOLOGY LIMITS • ONE ARCHITECTURE

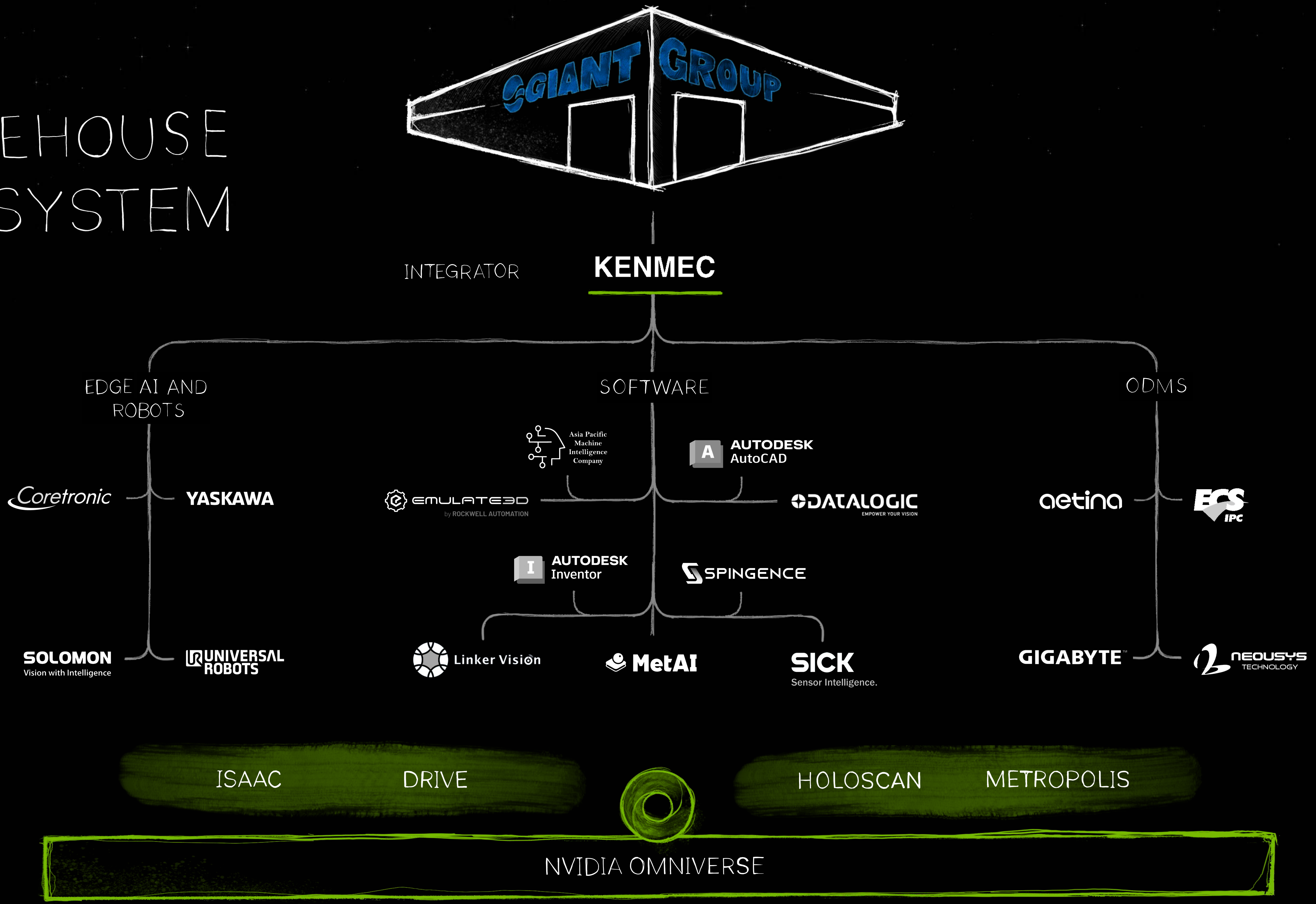


NVIDIA BLACKWELL 平台

NVIDIA BLACKWELL PLATFORM

A 01

WAREHOUSE ECOSYSTEM



INTEGRATOR

KENMEC

EDGE AI AND ROBOTS

Coretronic

YASKAWA

SOLOMON
Vision with Intelligence

UNIVERSAL ROBOTS

SOFTWARE

Asia Pacific Machine Intelligence Company

AUTODESK AutoCAD

EMULATE3D
by ROCKWELL AUTOMATION

DATALOGIC
EMPOWER YOUR VISION

AUTODESK Inventor

SPINGENCE

Linker Vision

MetAI

SICK
Sensor Intelligence.

ODMS

aetina

ECS IPC

GIGABYTE

NEOUSYS
TECHNOLOGY

ISAAC

DRIVE

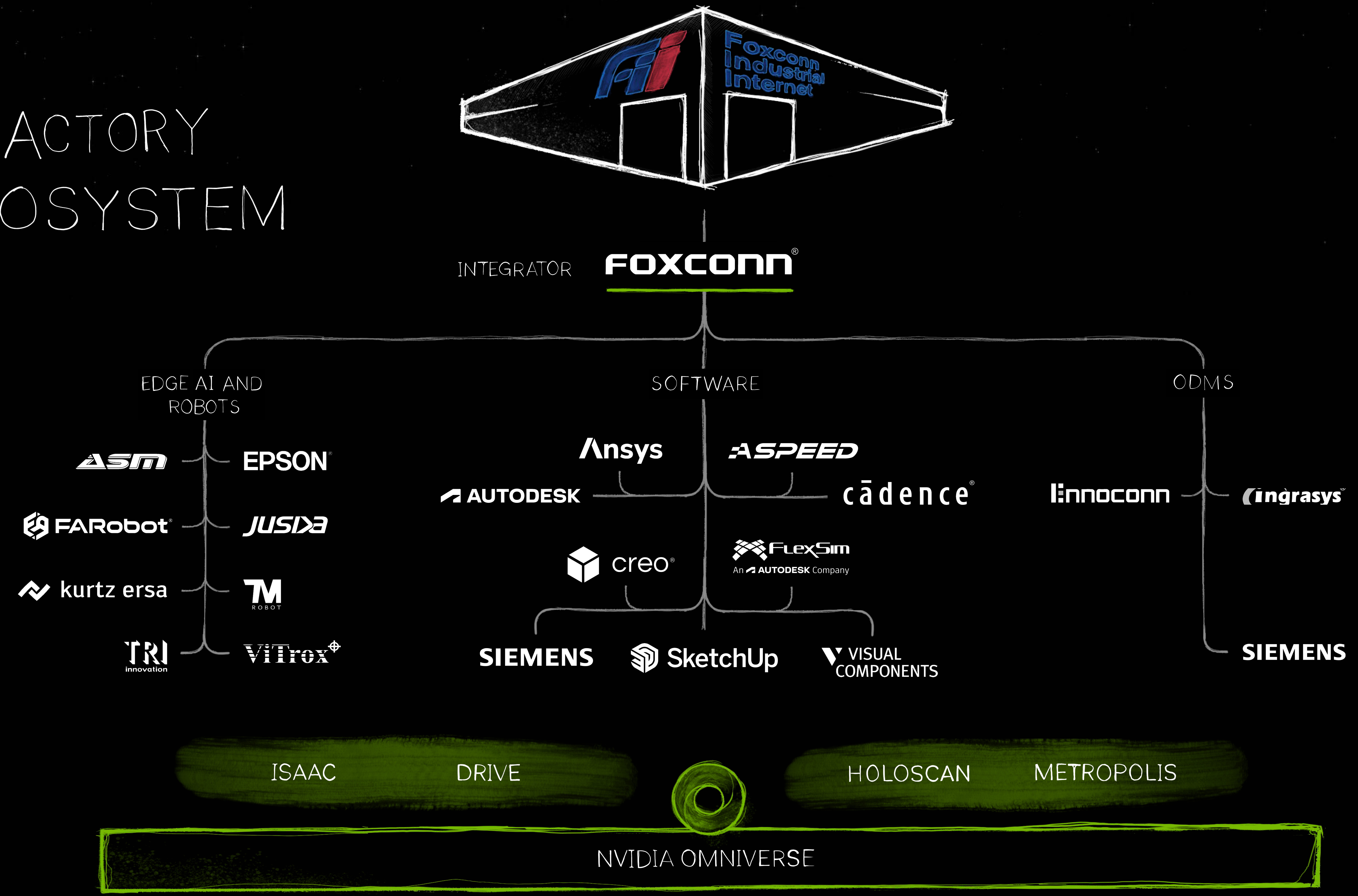
HOLOSCAN

METROPOLIS

NVIDIA OMNIVERSE

I
H
G
F
E
D
C
B
A
A 101
B 201
C 301
D 401
E 501
F 601

FACTORY ECOSYSTEM



INTEGRATOR FOXCONN

EDGE AI AND ROBOTS

- ASM
- EPSON
- FARobot
- JUSDA
- kurtz ersa
- TM ROBOT
- TRI innovation
- ViTrox

SOFTWARE

- Ansys
- ASPEED
- AUTODESK
- cādence
- creo
- FlexSim An AUTODESK Company
- SIEMENS
- SketchUp
- VISUAL COMPONENTS

ODMS

- Innoconn
- ingrasys
- SIEMENS

ISAAC DRIVE HOLOSCAN METROPOLIS

NVIDIA OMNIVERSE

FANUC Robot CRX-20iA/L
FANUC Robot CRX-20iA/L
FANUC Robot CRX-20iA/L
FOXCONN®

DELTA
DELTA
DELTA
DELTA

PEGATRON

wistron

FANUC Robot CRX-20iA/L
FANUC Robot

NVIDIA Omniverse

NVIDIA AI

NVIDIA Isaac Manipulator

Isaac ROS 3.0 Available Now on Github

SIEMENS

SIMATIC Pick AI

NVIDIA Isaac Manipulator

NVIDIA Omniverse

NVIDIA AI

TM Operator Platform
專屬SI、整合專家的新型態二次開發軟體平台解決方案
什麼是 TM Operator Platform?
TMvision
SIEMENS
SMART
SIMPLE
SAFE

ArcBest

NVIDIA Omniverse

NVIDIA AI

ArcBest Vaux

NVIDIA Isaac Perceptor

NVIDIA Omniverse

NVIDIA AI

NVIDIA DRIVE AV

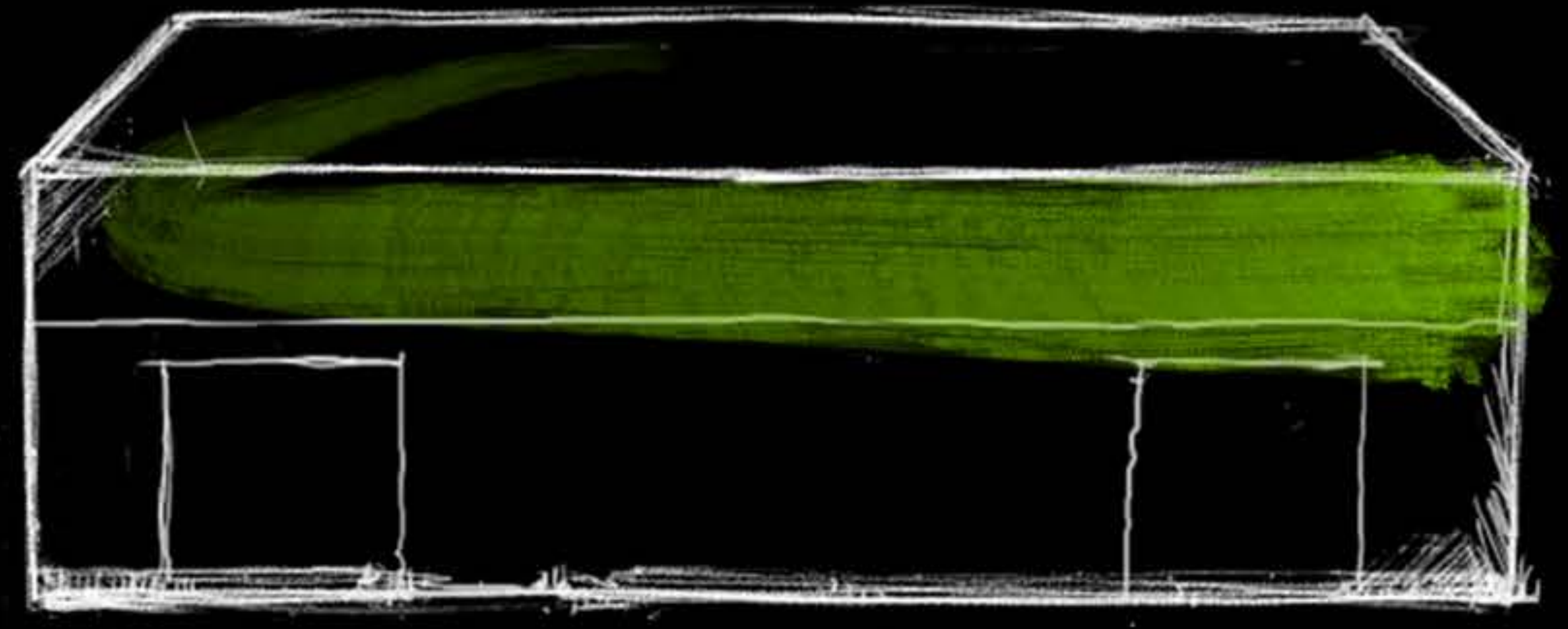
50
40
30
20
10
NO TURNS
SPEED LIMIT 25
NVIDIA

GR00T Stack

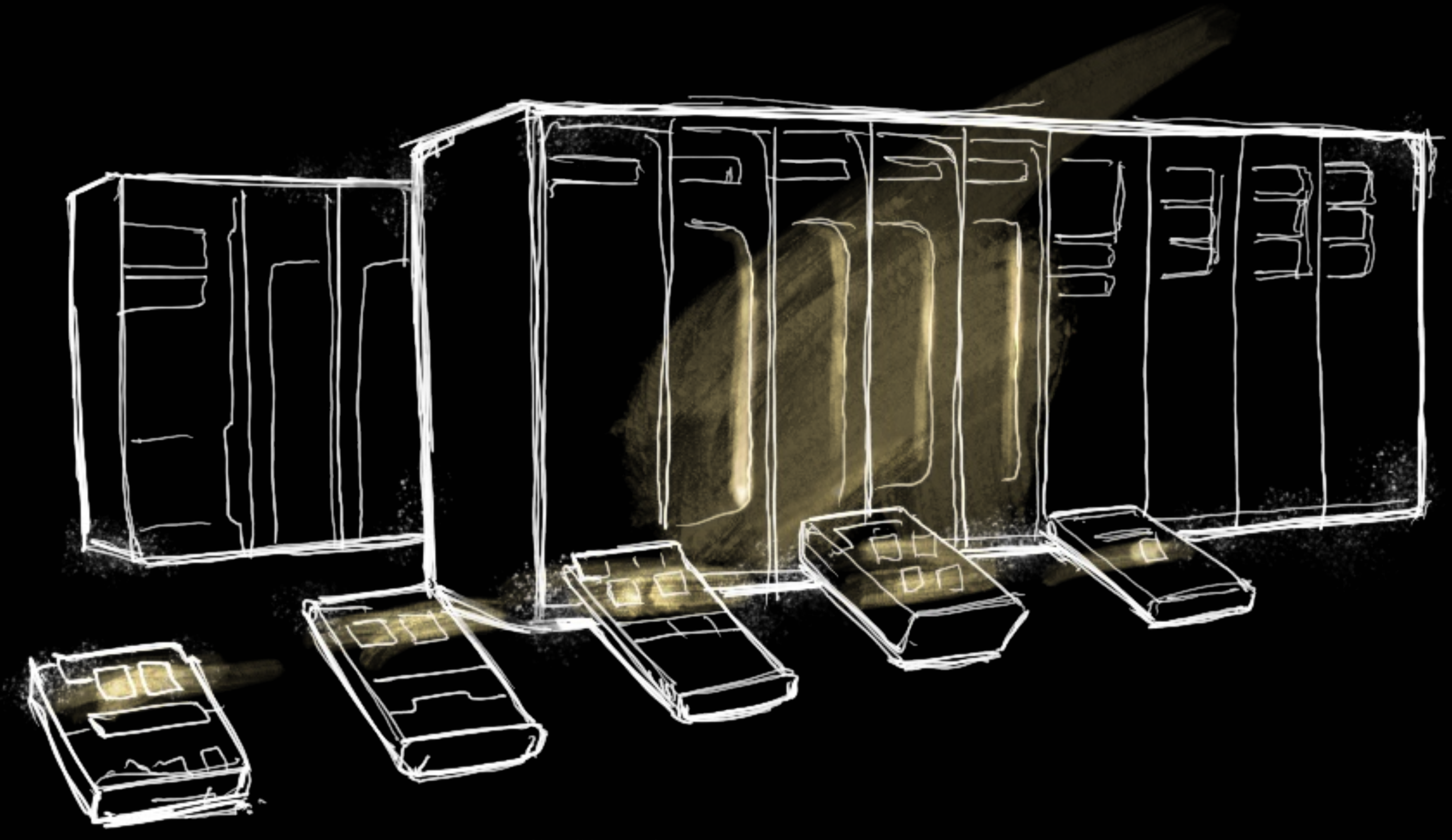
NVIDIA Omniverse

NVIDIA AI

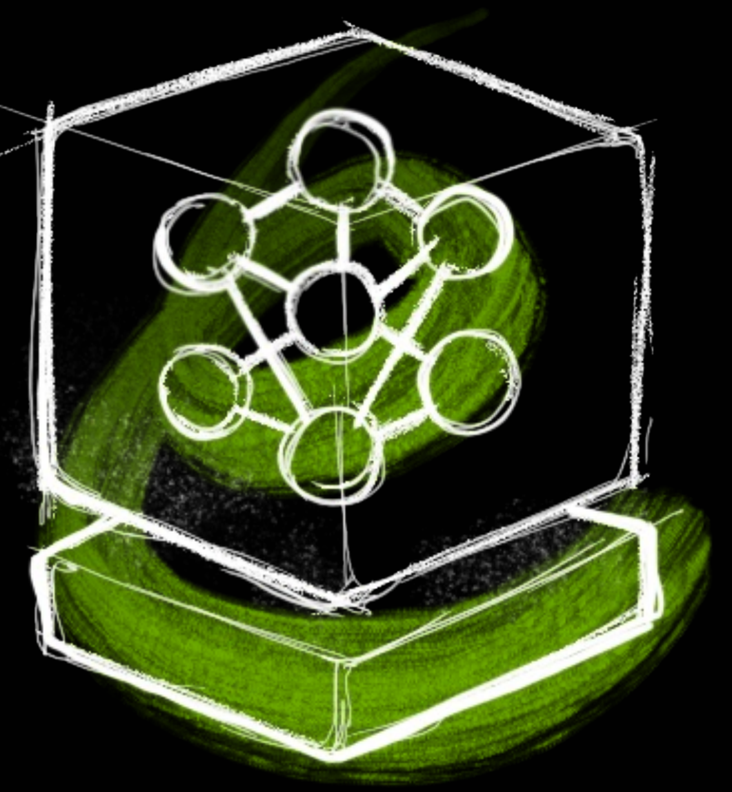
Unitree
A1
BostonDynamics



AI FACTORY



BLACKWELL



NIMS

OMNIVERSE/
ROBOTICS

新産業革命

“A NEW INDUSTRIAL REVOLUTION”

NVIDIA