



Full Stack, 3 Chips, Data Center Scale

30 Million CUDA Downloads

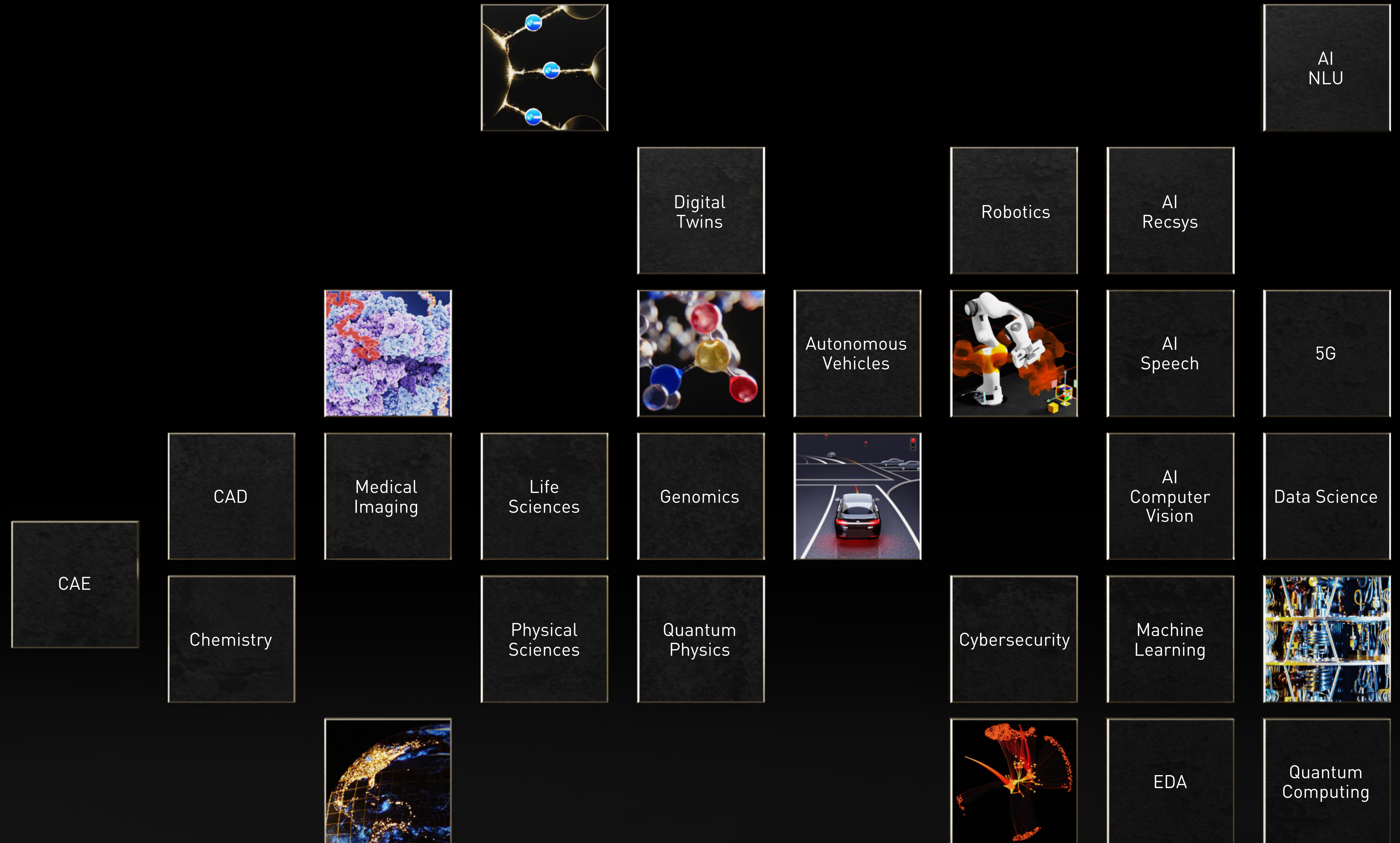
3 Million Developers

2,700 Applications

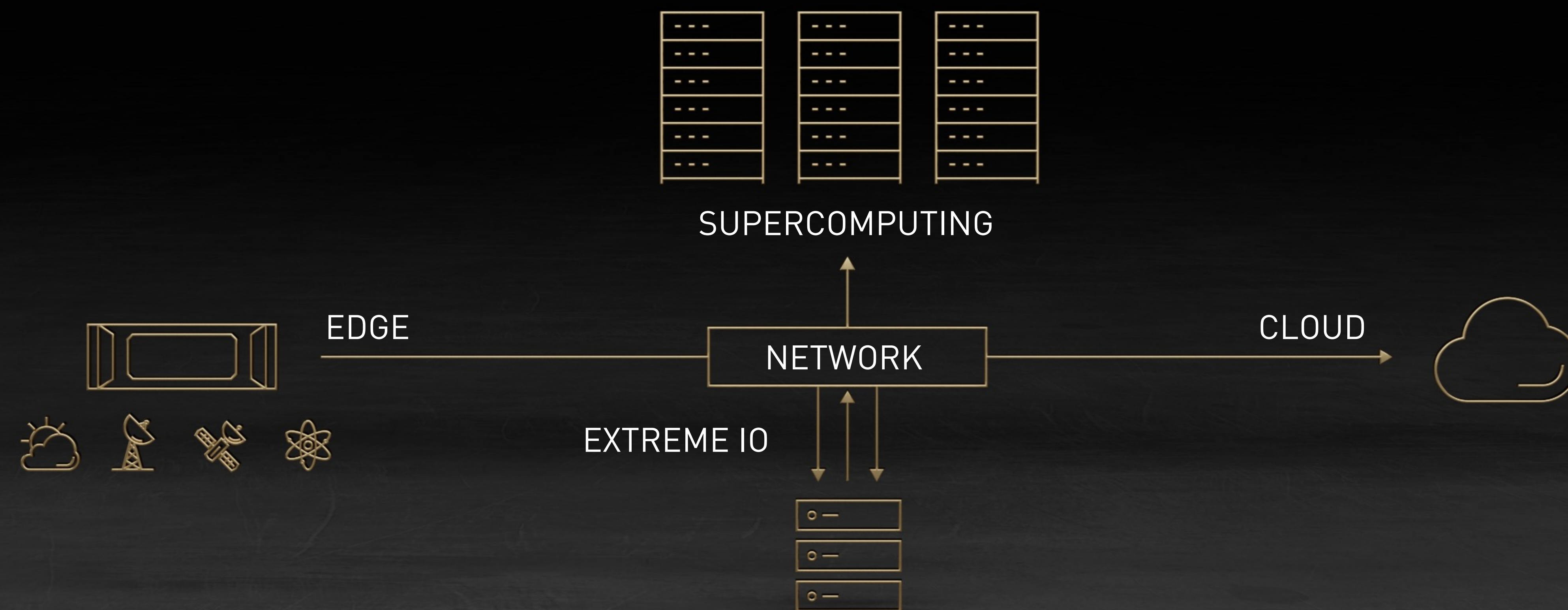
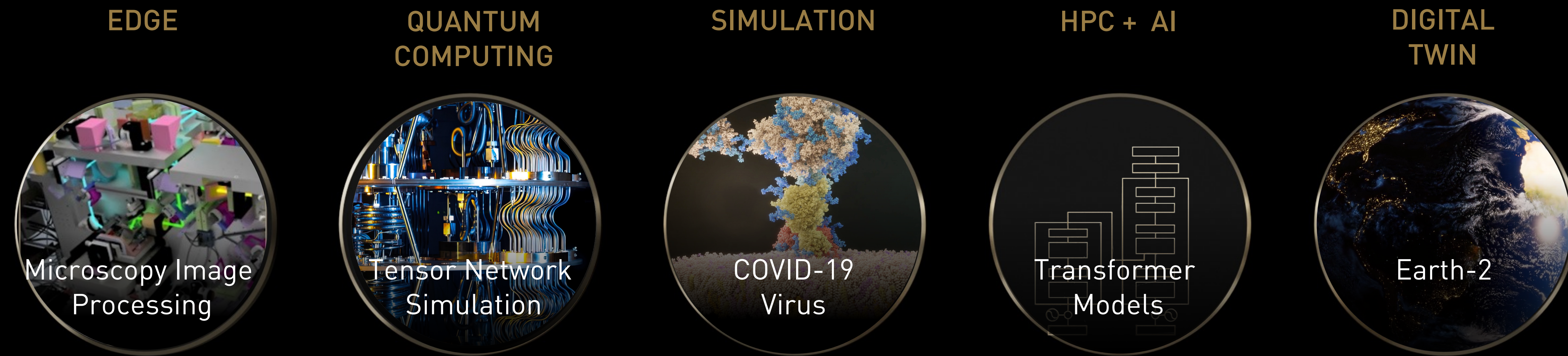
450 SDKs



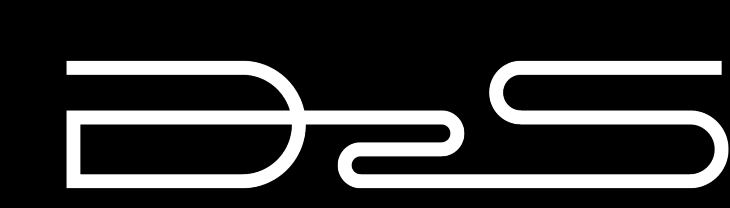
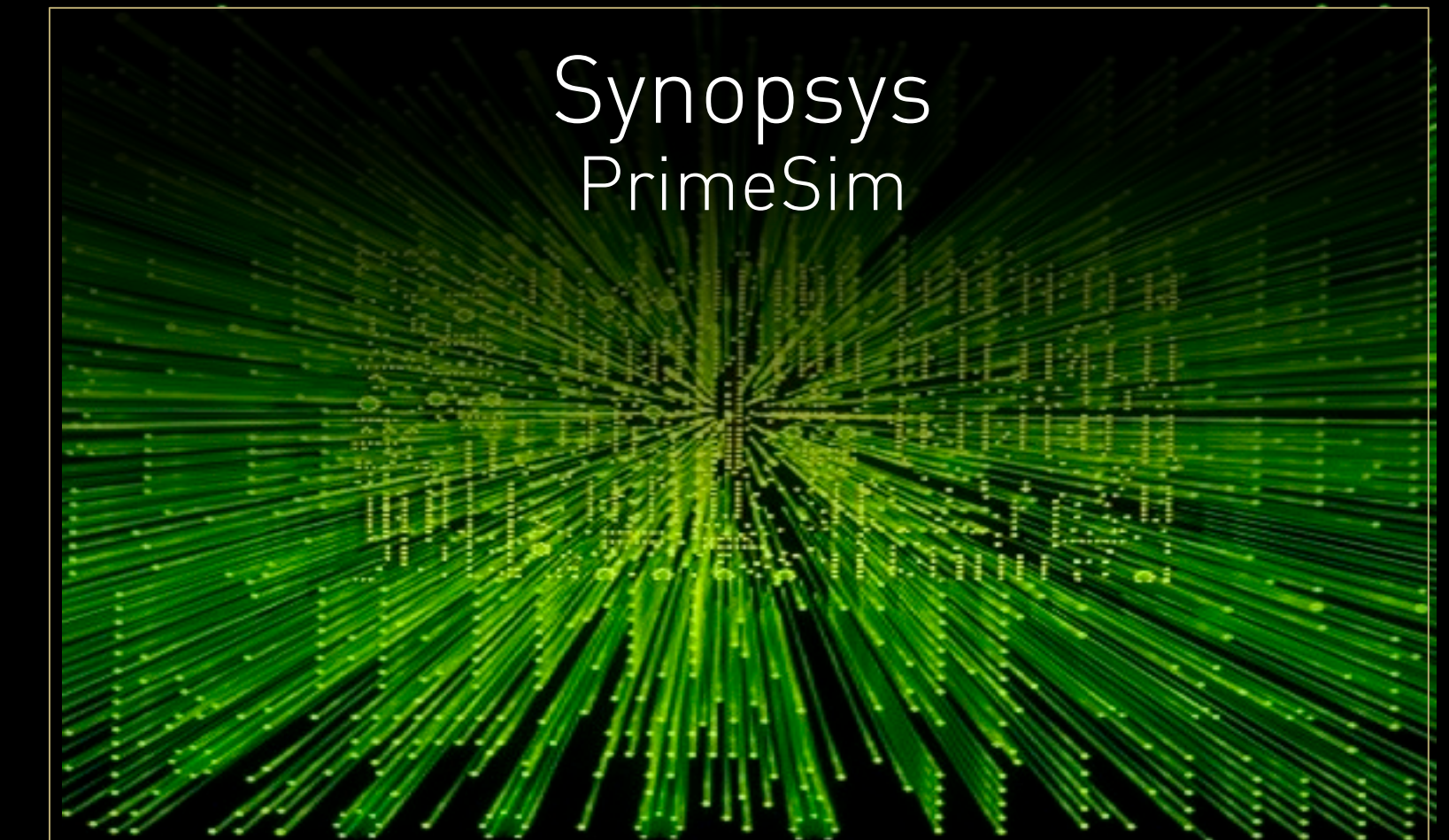
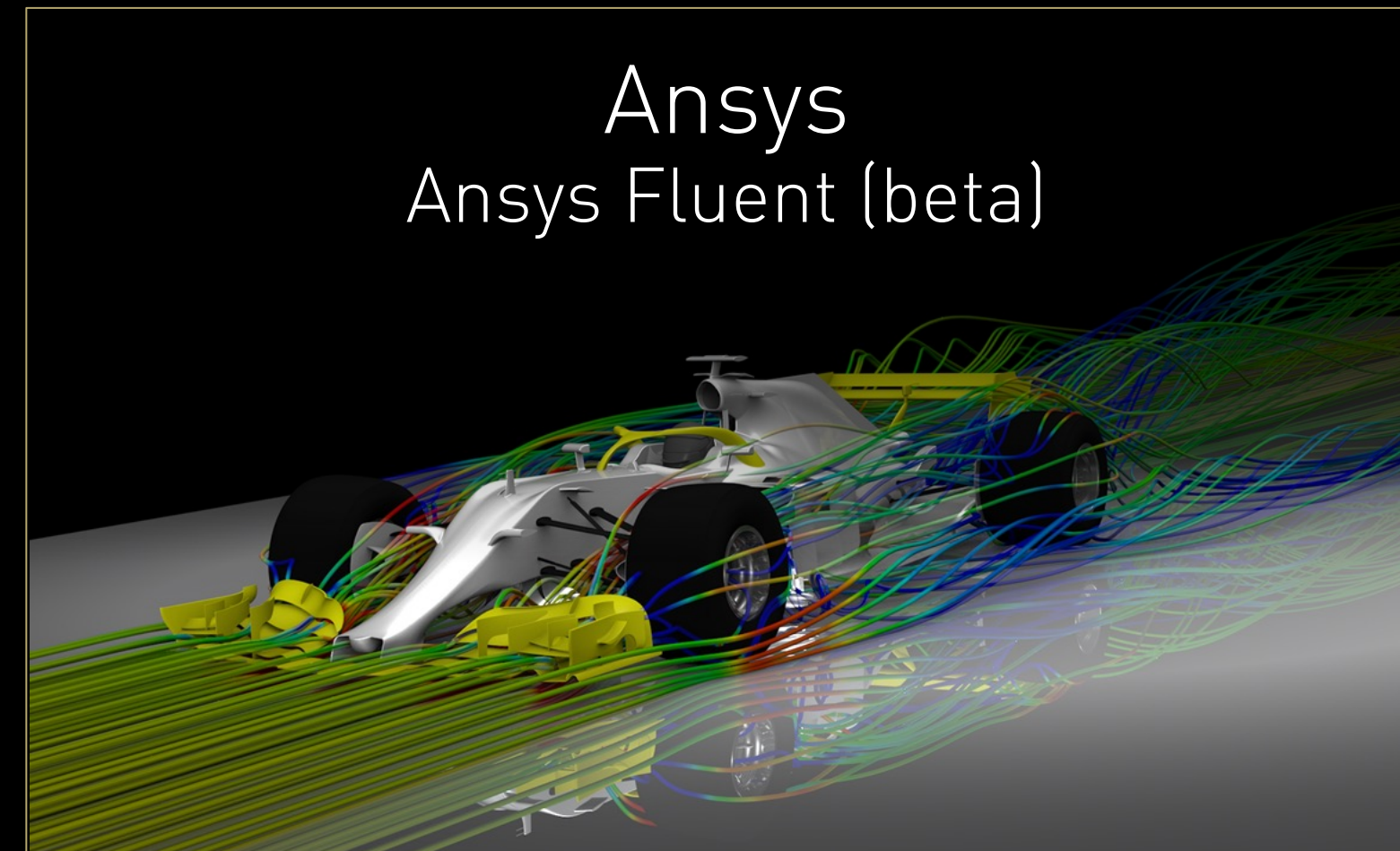
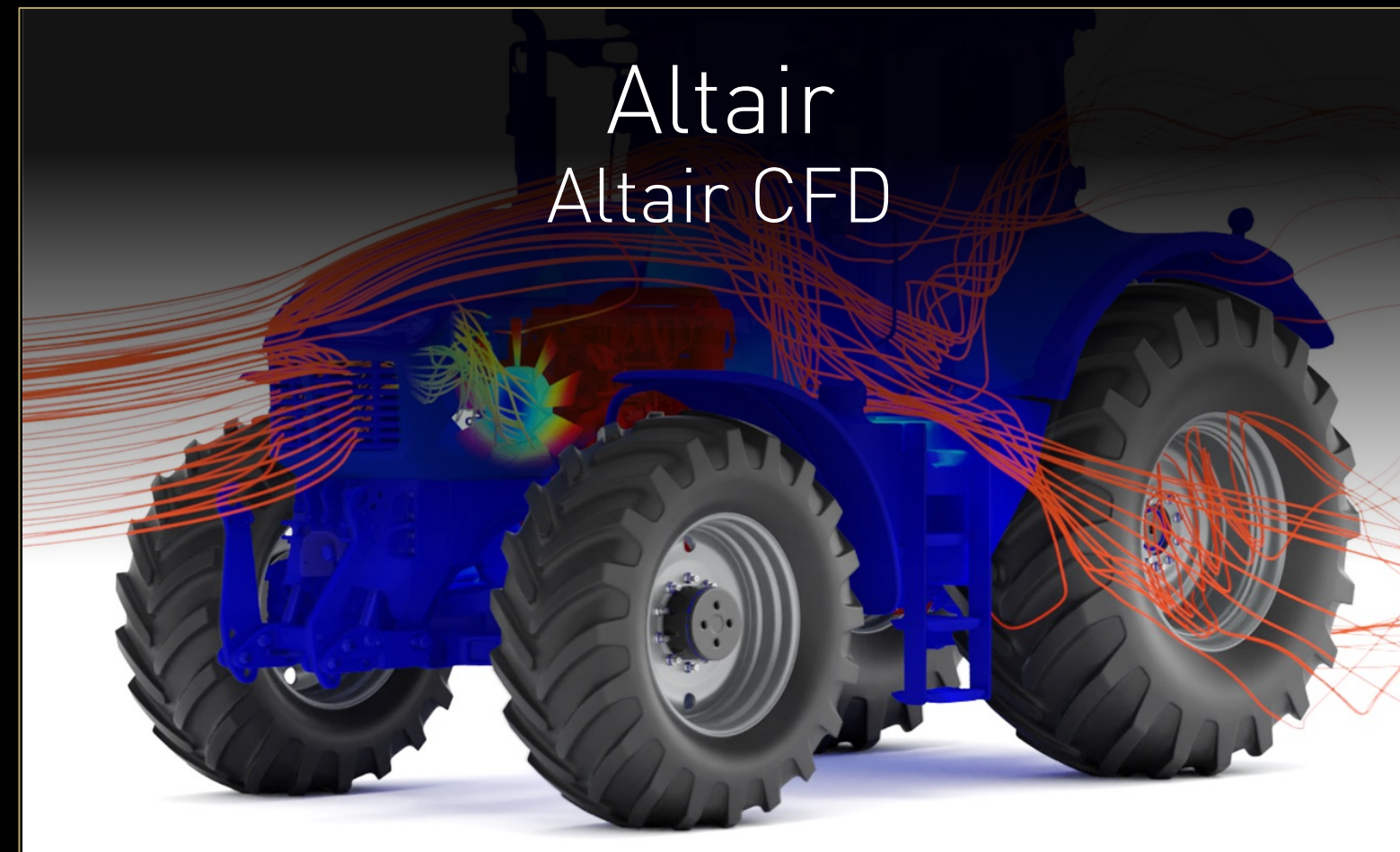
NVIDIA
Accelerated Computing



WORKLOADS OF THE MODERN DATA CENTER



SIMULATION



+300

other ISV apps

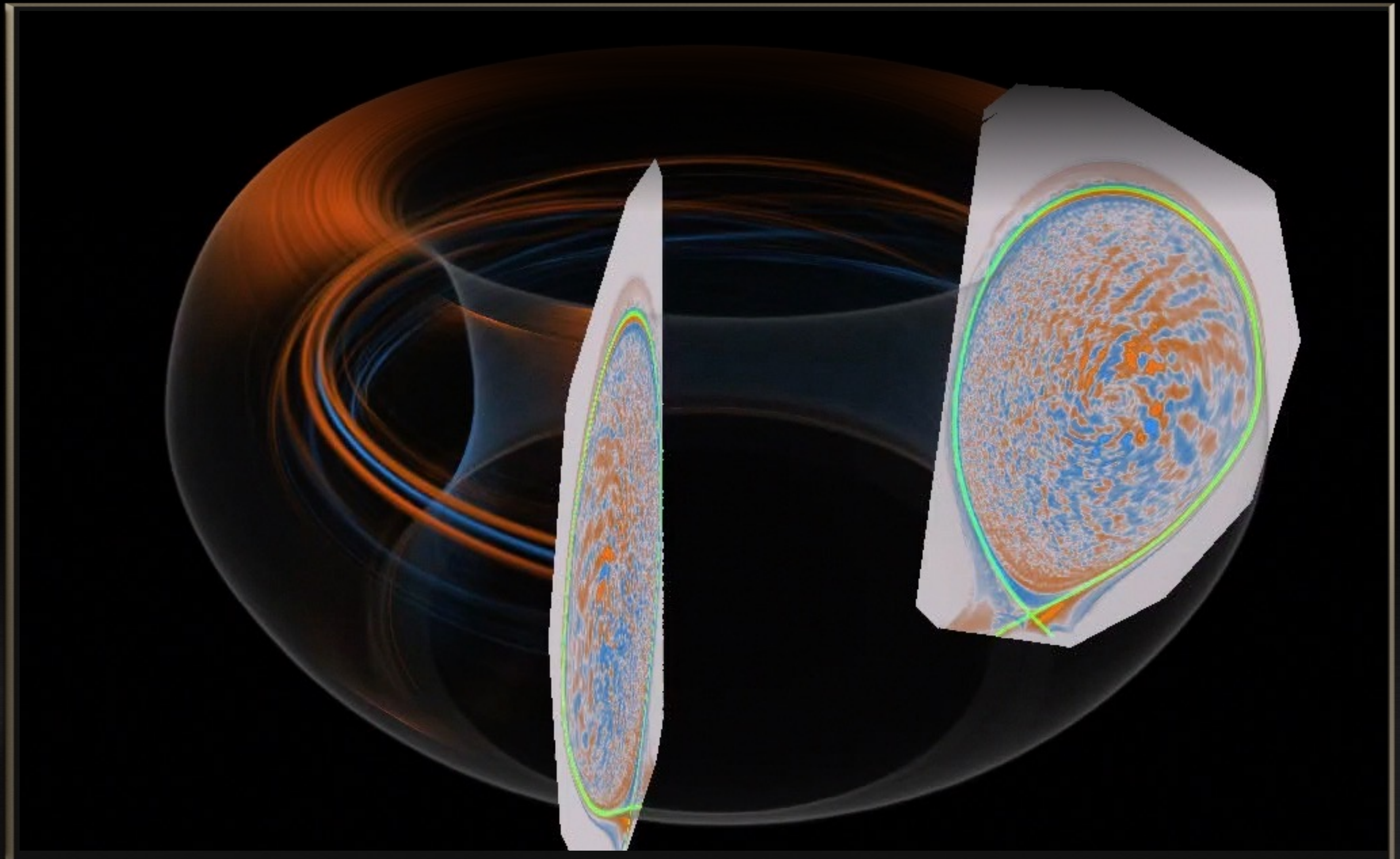
NEW WAVE OF INDUSTRIAL HPC ACCELERATION

SIMULATION

Advancing Research For A Sustainable Future

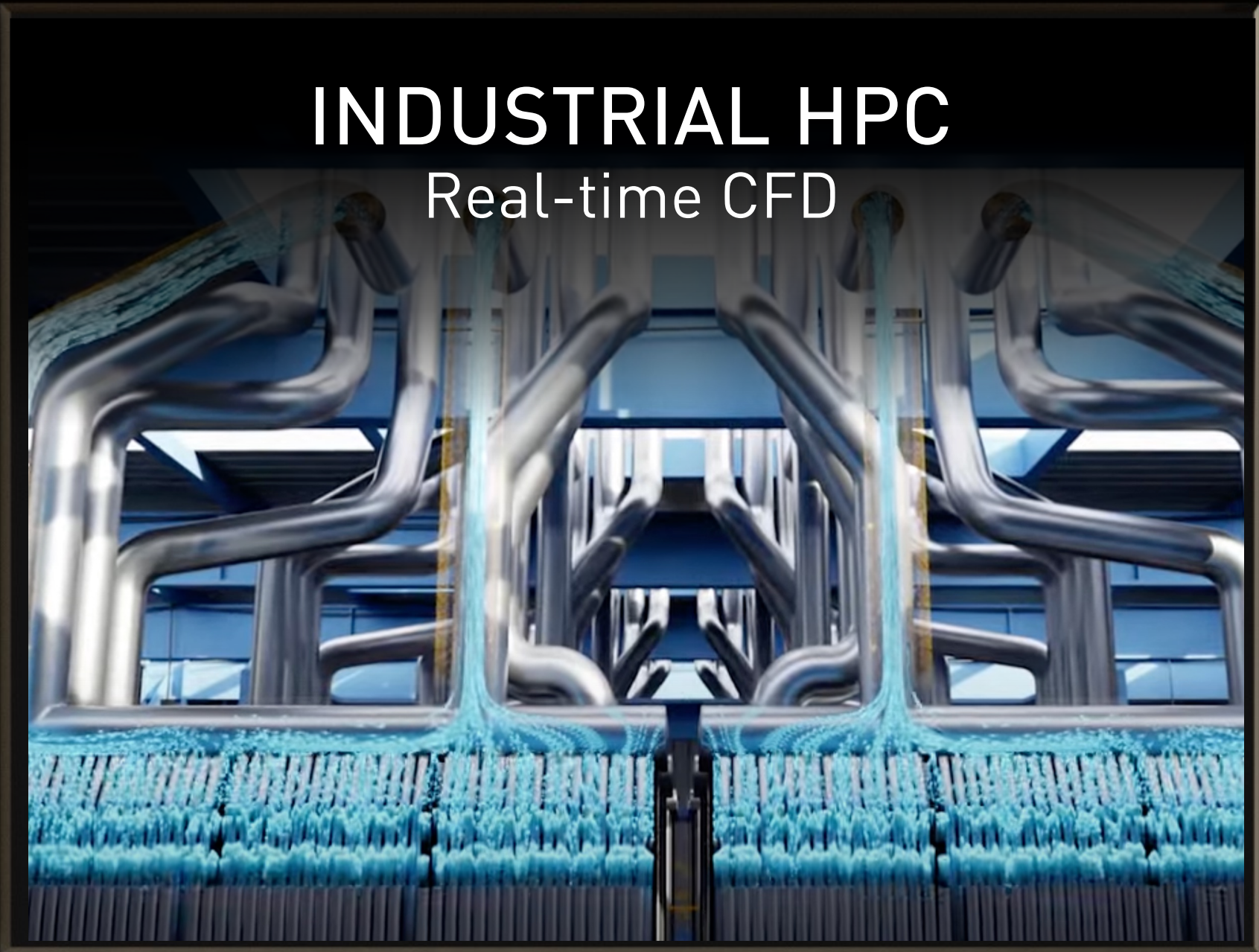
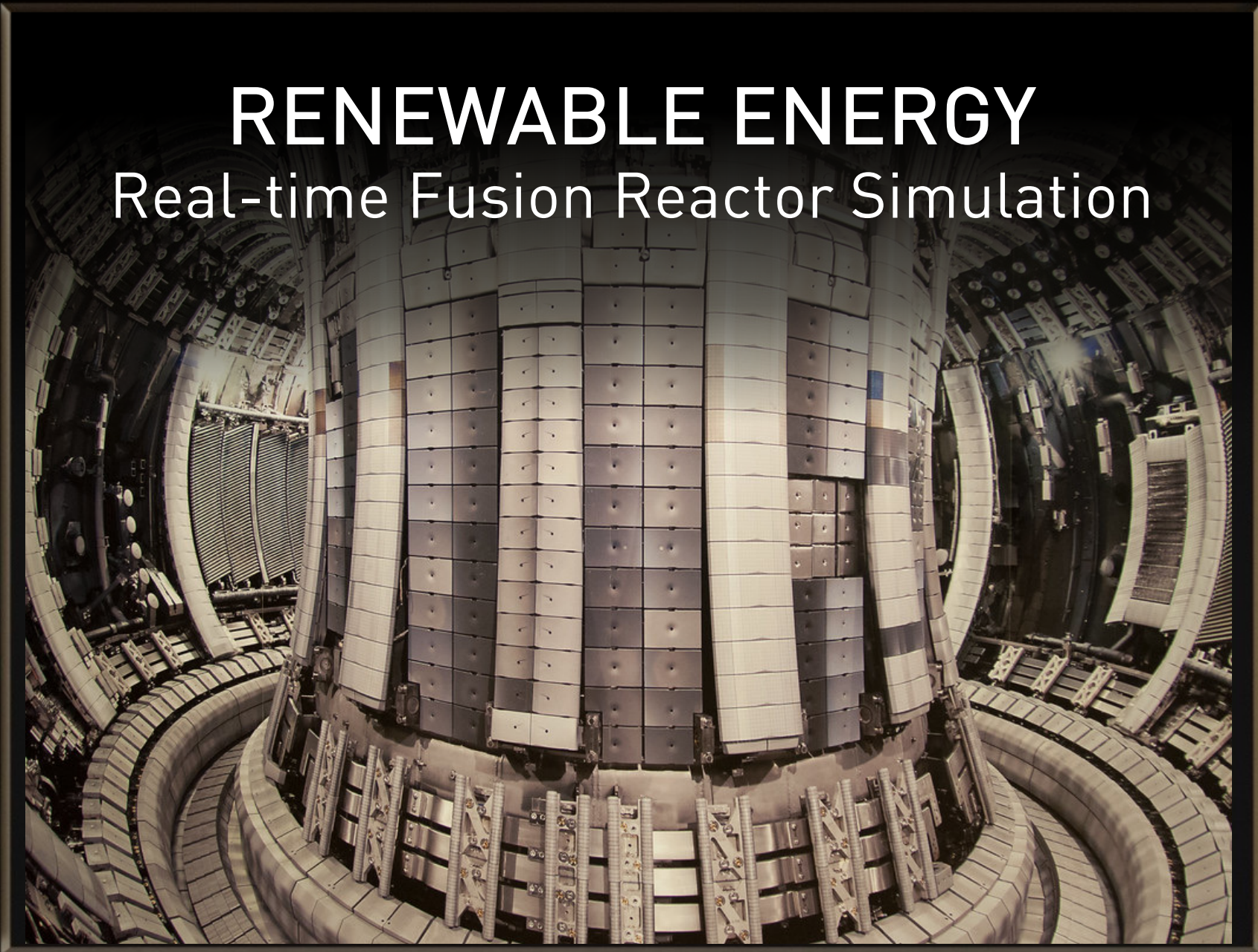
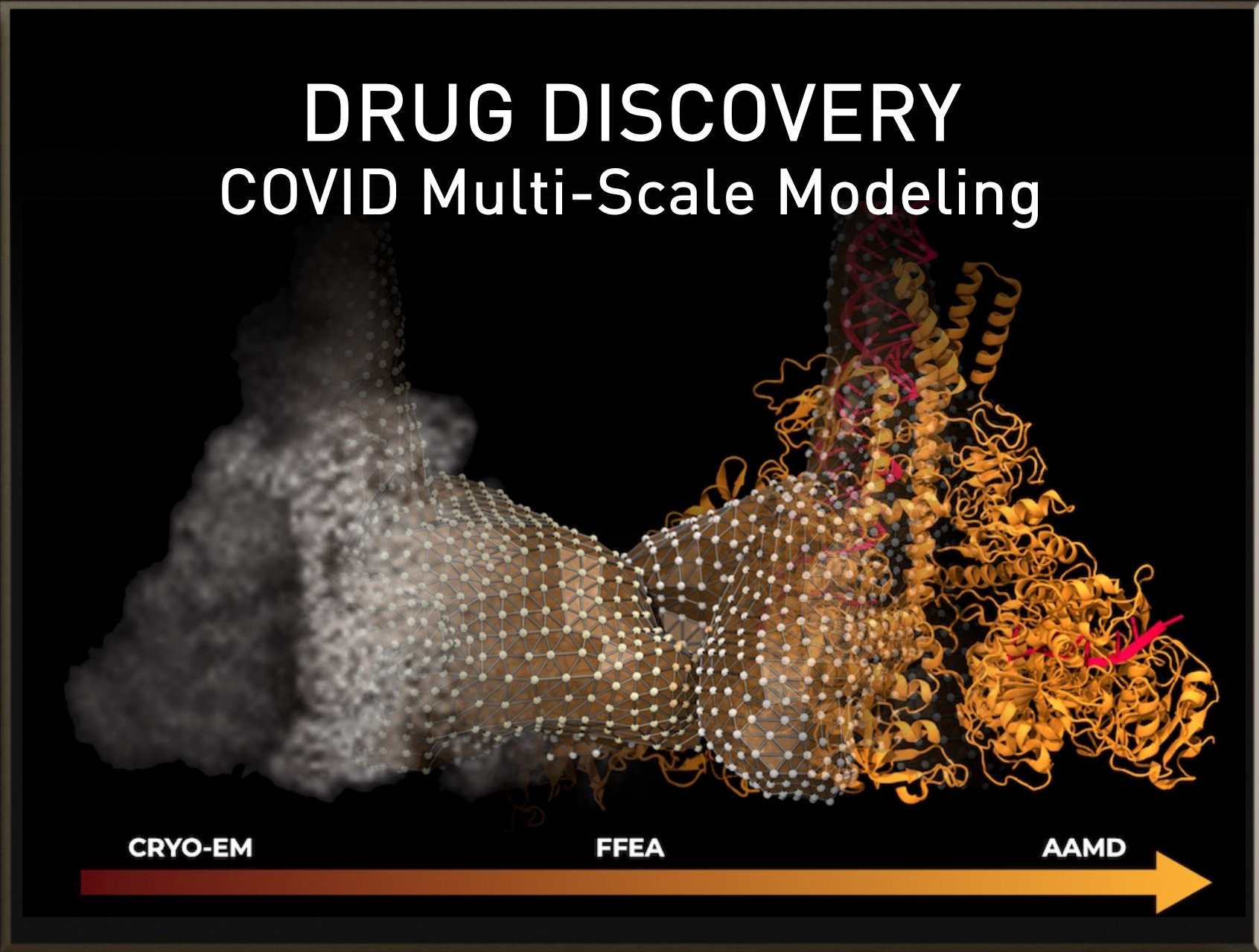
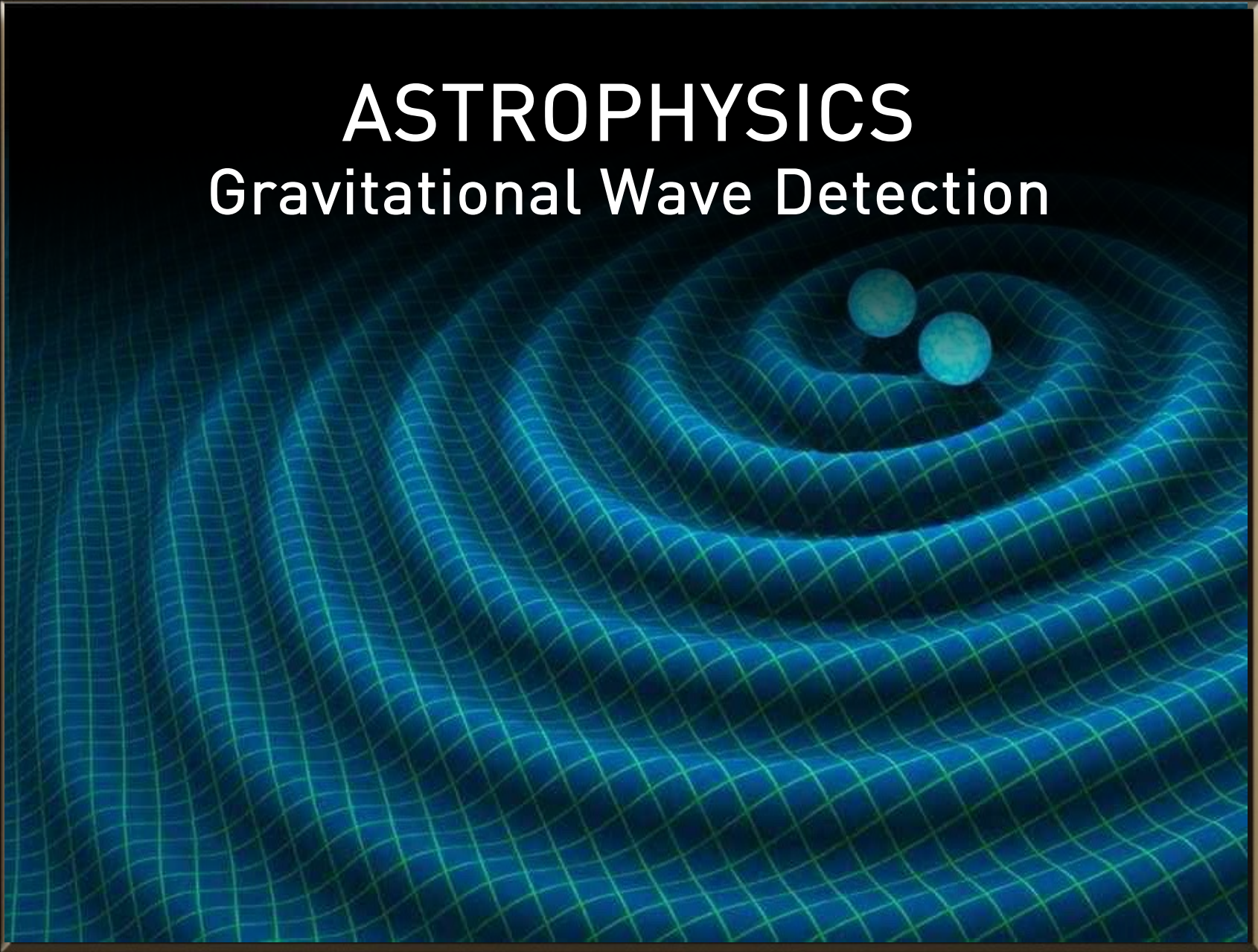
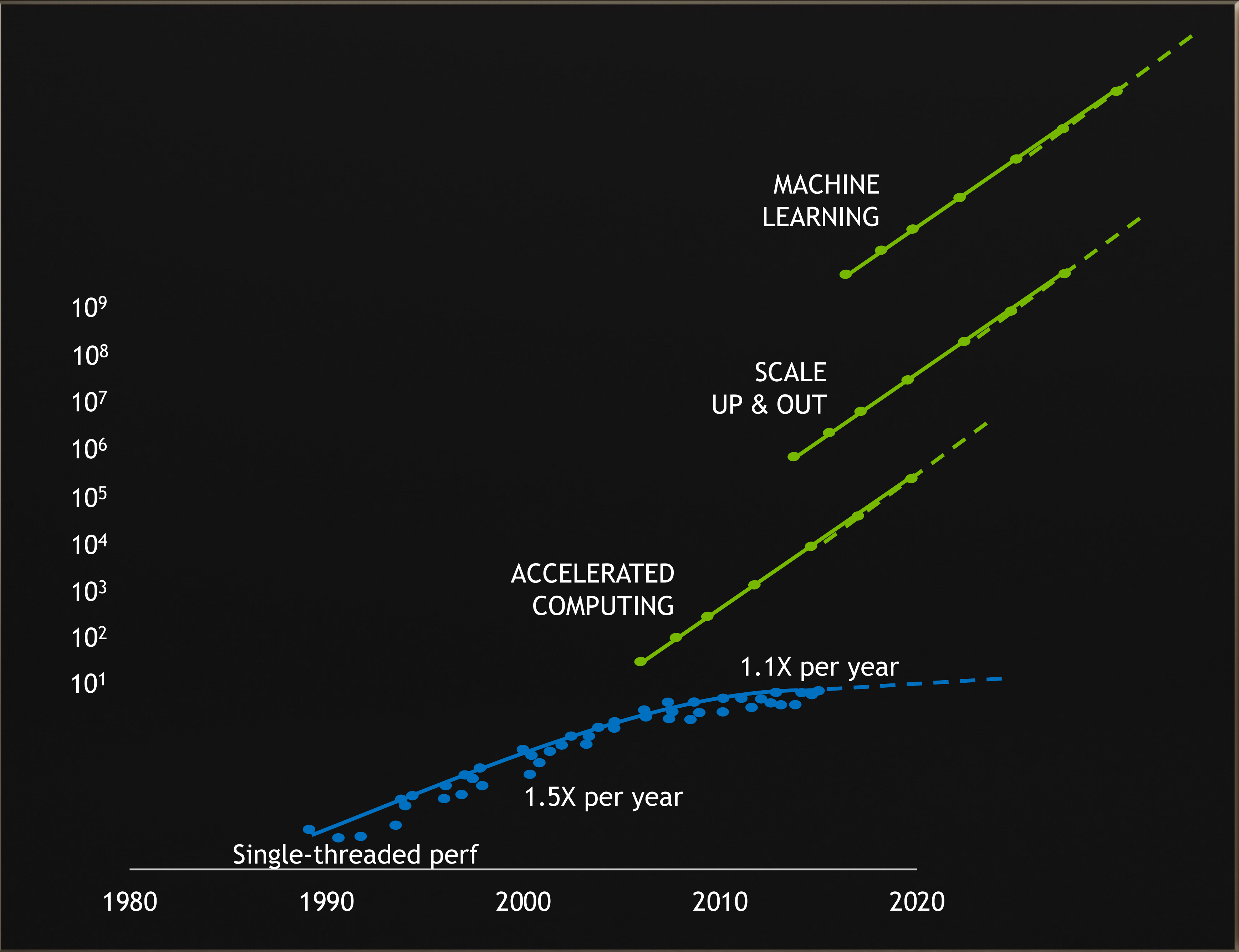
FIRST FUSION REACTOR SIMULATION WITH ELECTROMAGNETIC EFFECTS AT SCALE

- Models how electromagnetic micro-turbulence impacts electron leakage
- Results could impact ITER's design
- 256 nodes | 1,024 A100 GPUs on Perlmutter
- Running XGC written in C++ with Kokkos



MILLION-X SPEEDUP FOR INNOVATION AND DISCOVERY

Combination of Accelerated Computing, Data Center Scale and AI



HPC AT THE EDGE

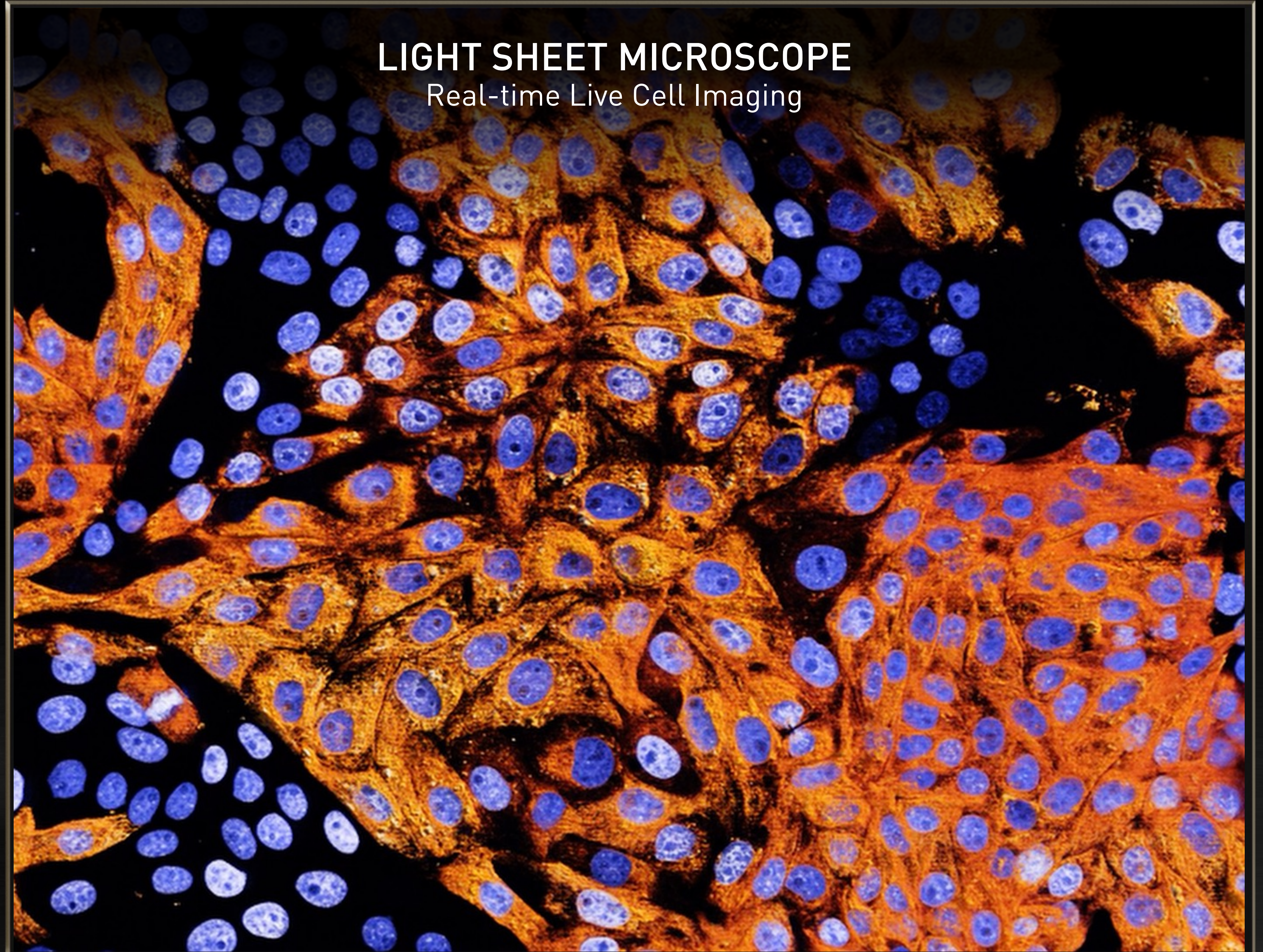
Light Sheet Microscopy with NVIDIA Clara Holoscan

NVIDIA CLARA HOLOSCAN PLATFORM



LIGHT SHEET MICROSCOPE

Real-time Live Cell Imaging

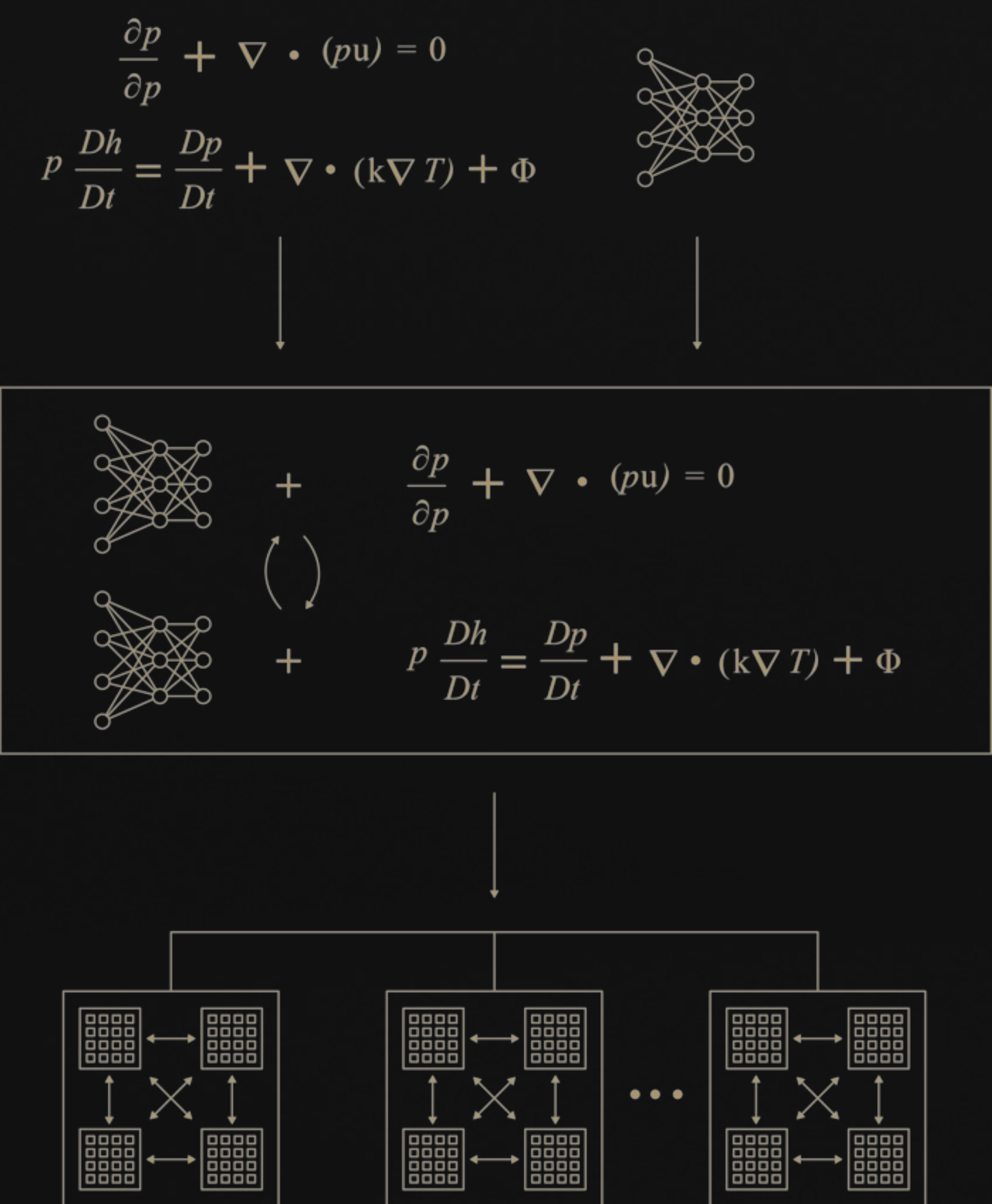


NVIDIA MODULUS FOR SCIENTIFIC DIGITAL TWINS

Battling Climate Change With Physics-ML Accelerated Digital Twins

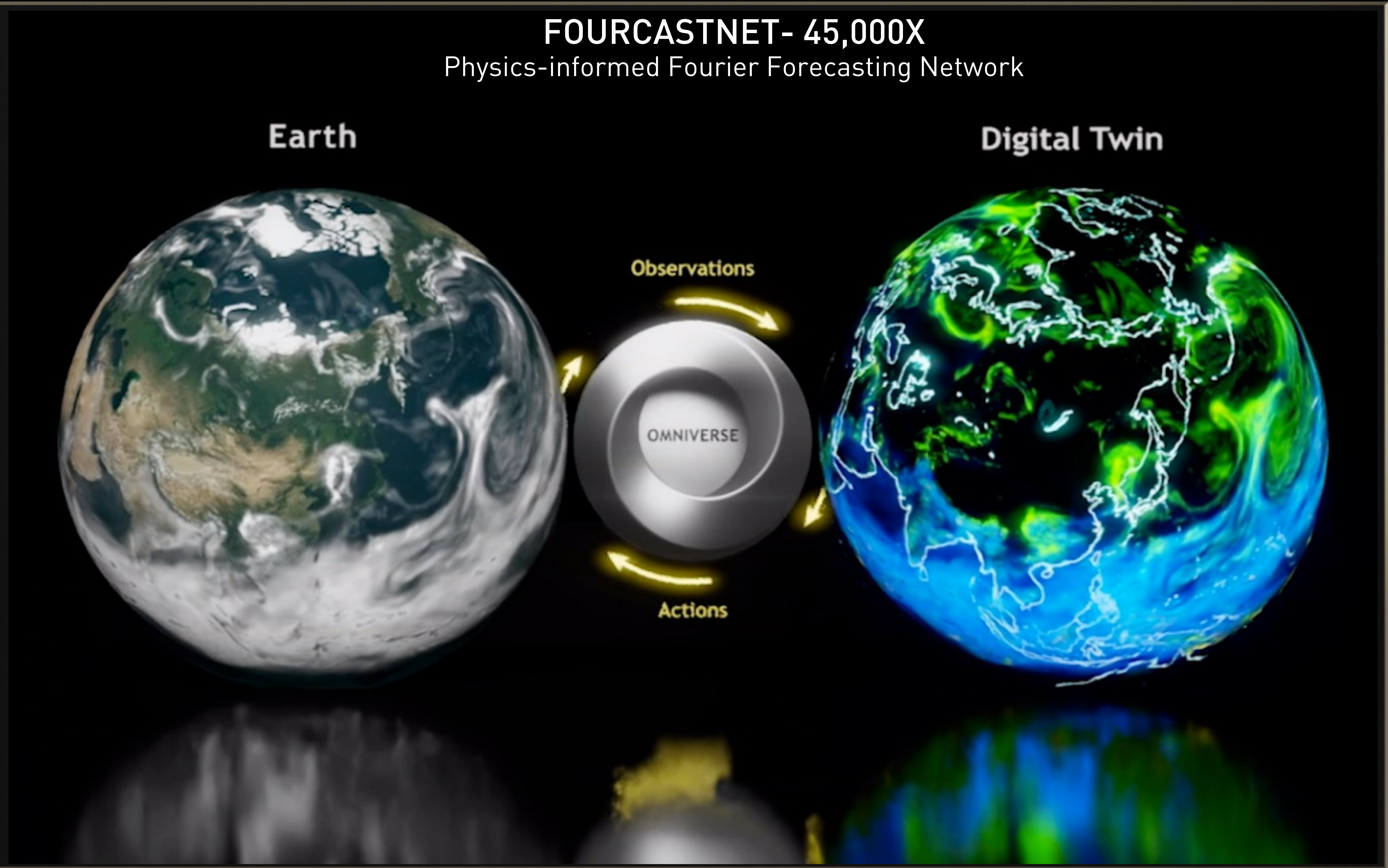
MODULUS

New AFNO , Python, and Omniverse Integration



FOURCASTNET- 45,000X

Physics-informed Fourier Forecasting Network

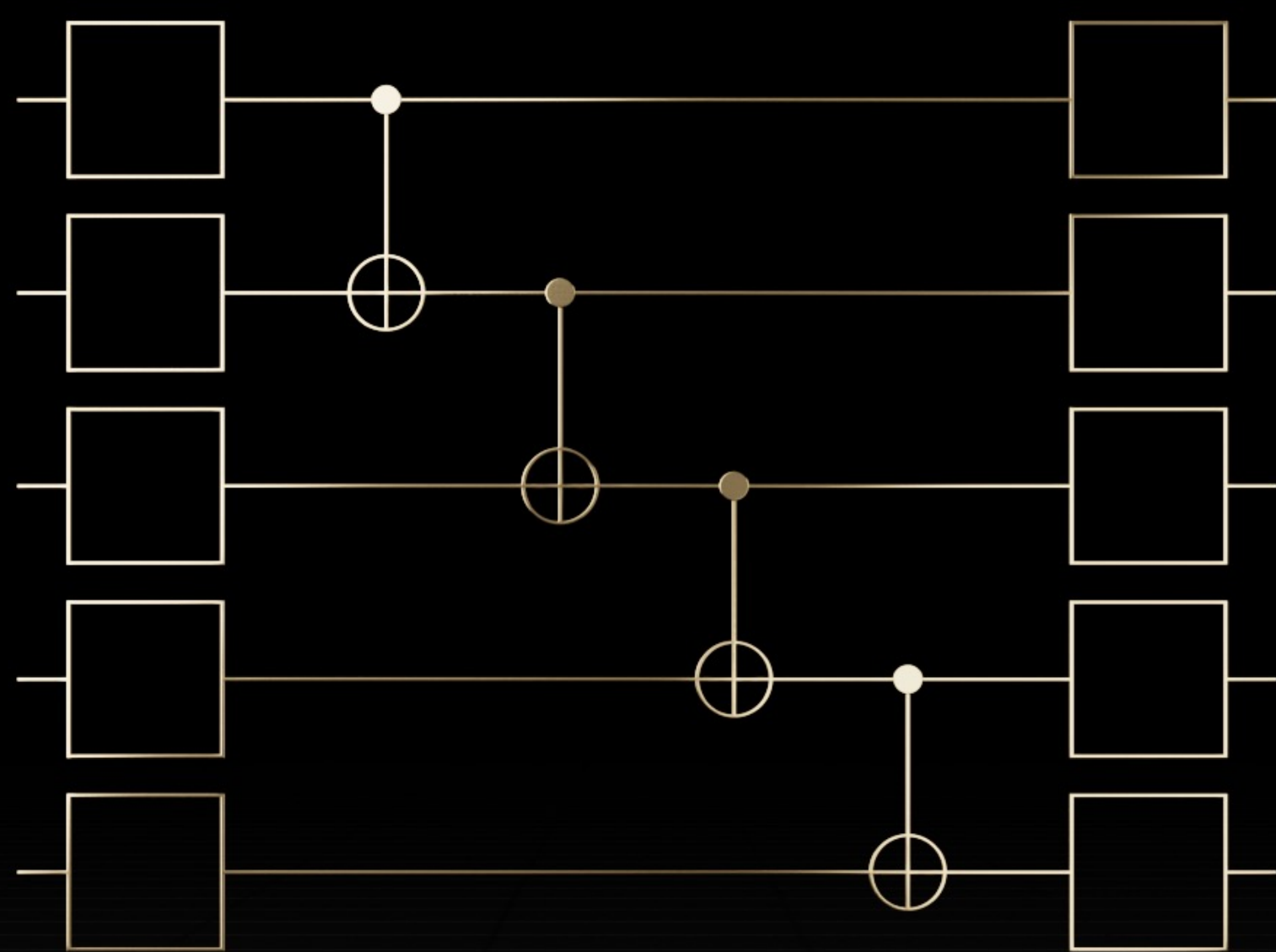


QUANTUM COMPUTING

Research the Computer of Tomorrow with the Most Powerful Computer Today

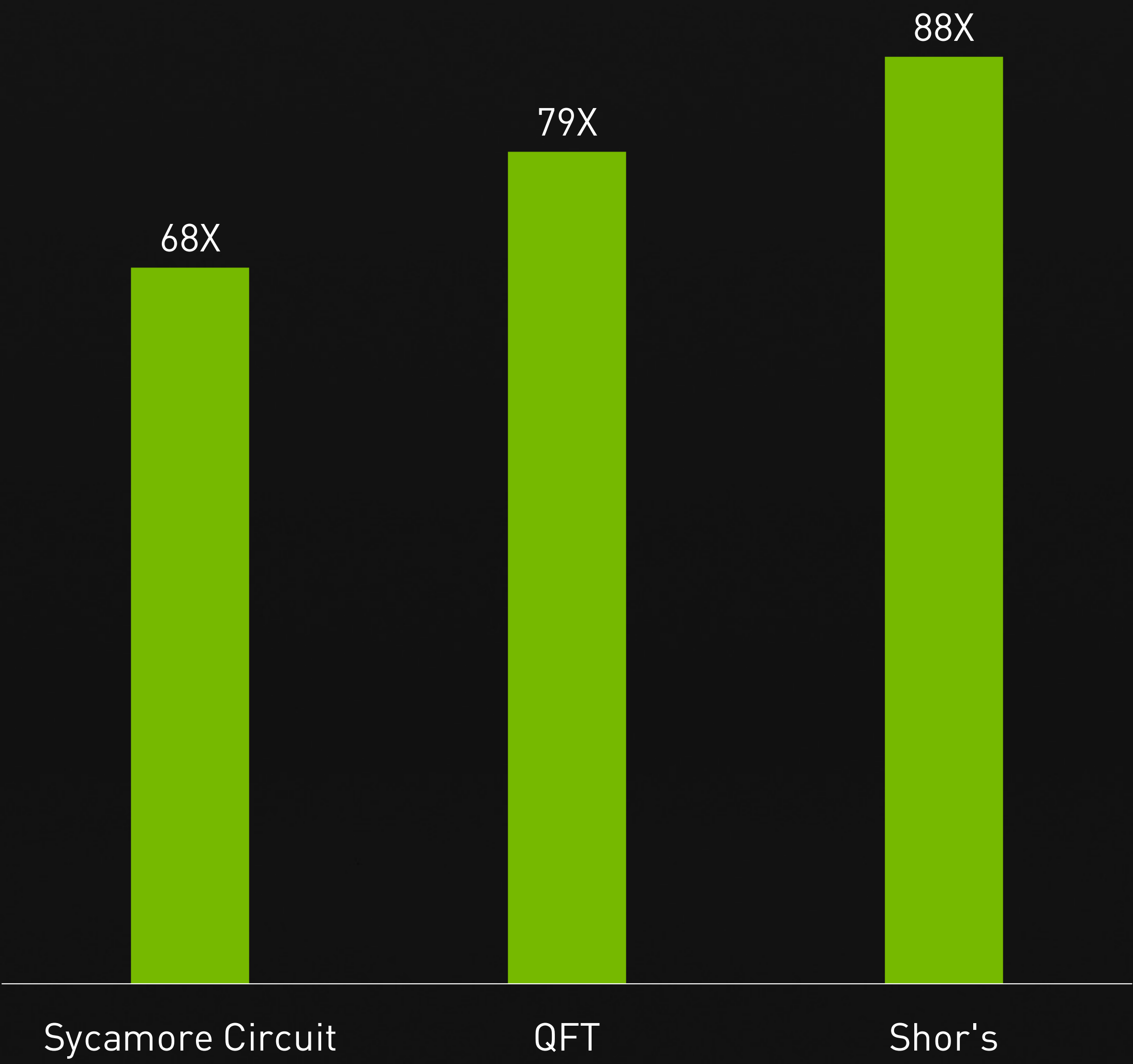
EXPANDING ECOSYSTEM

New releases and partnerships



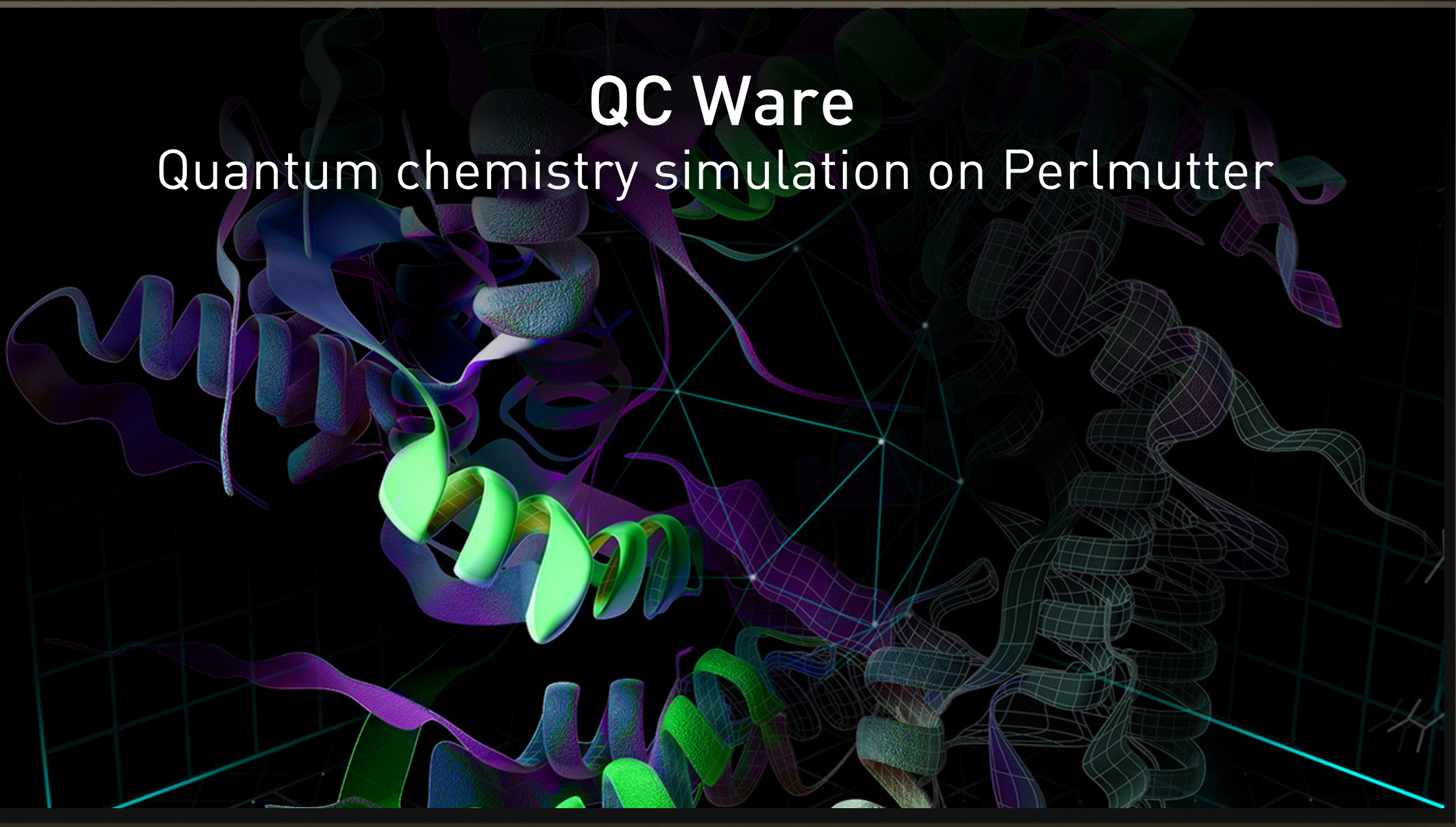
CUQUANTUM DGX APPLIANCE

88X Speedup on Quantum Algorithms



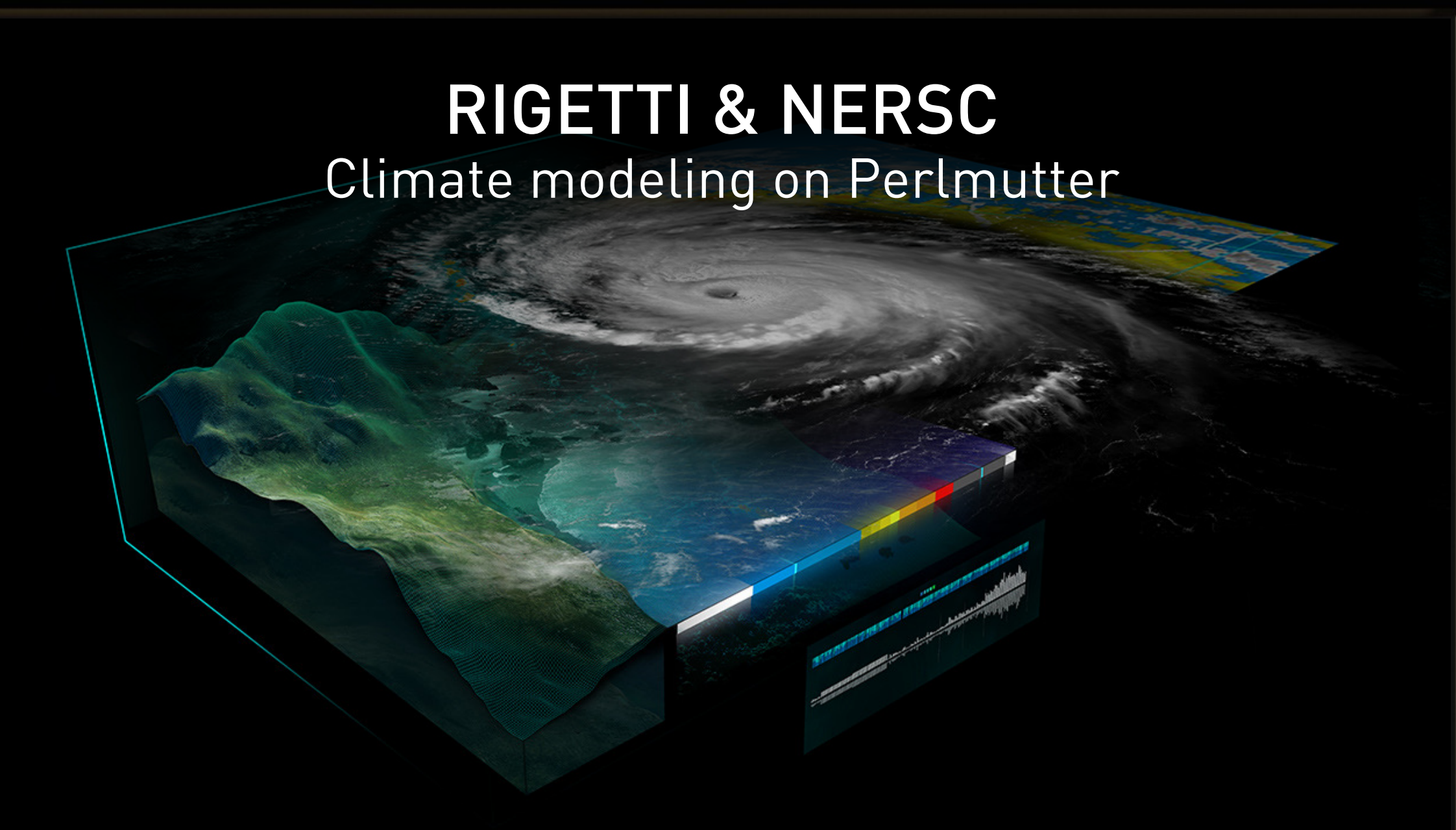
QC Ware

Quantum chemistry simulation on Perlmutter

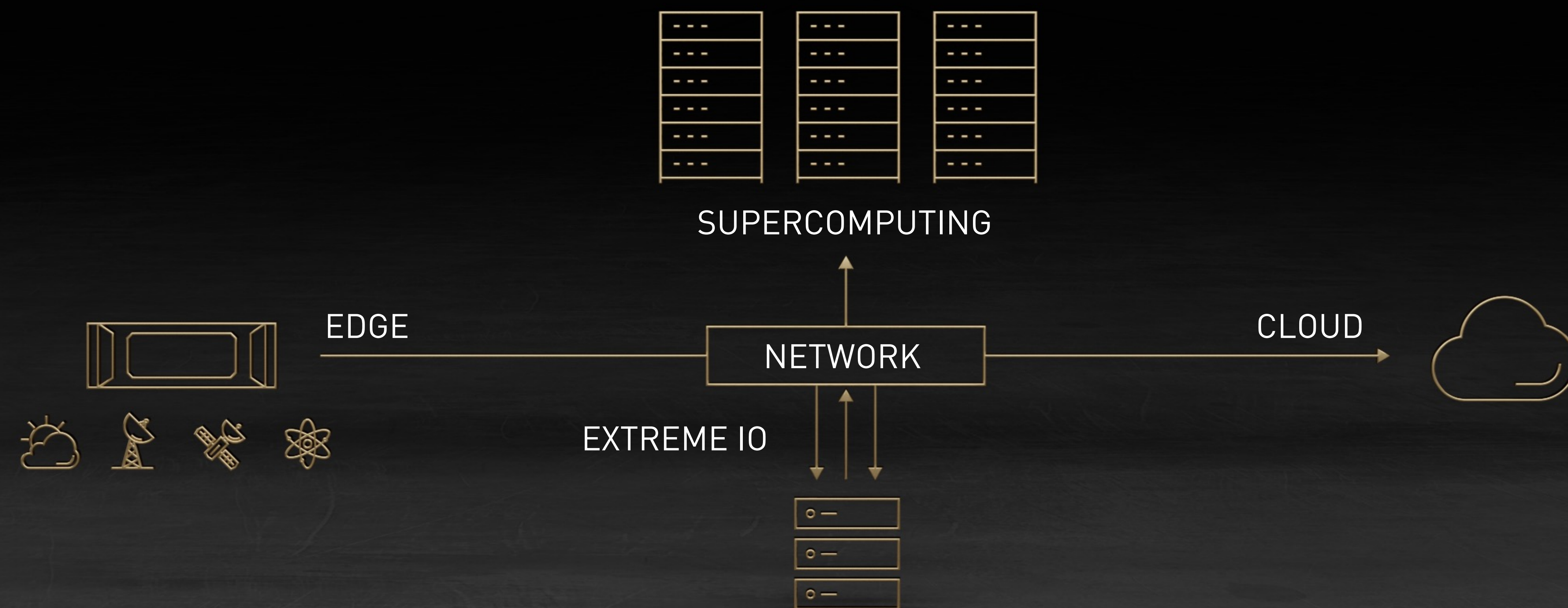
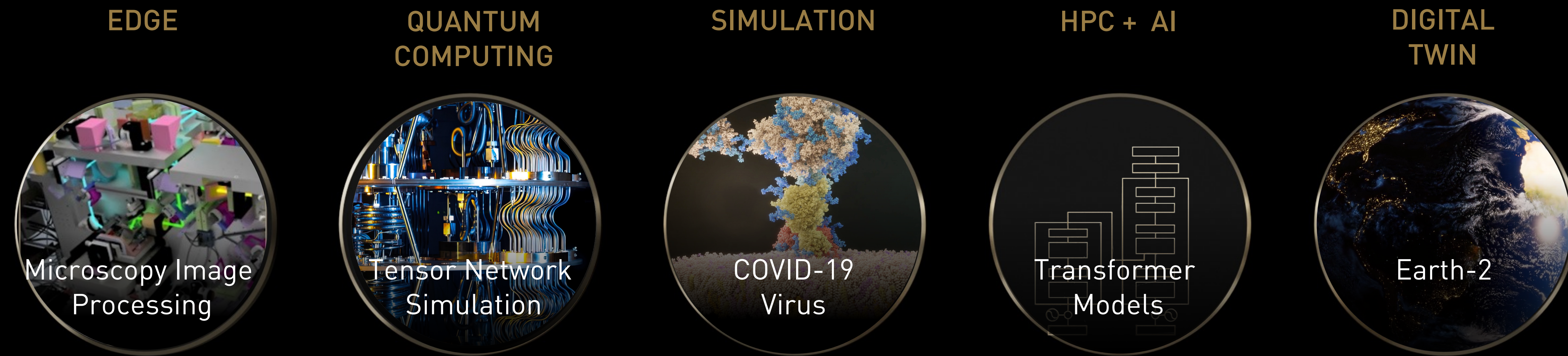


RIGETTI & NERSC

Climate modeling on Perlmutter



WORKLOADS OF THE MODERN DATA CENTER



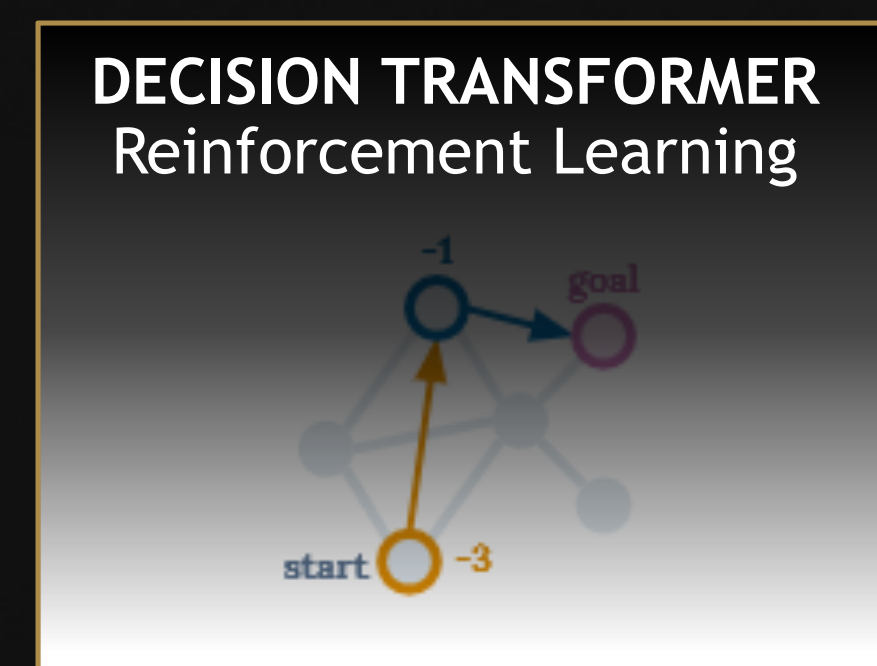
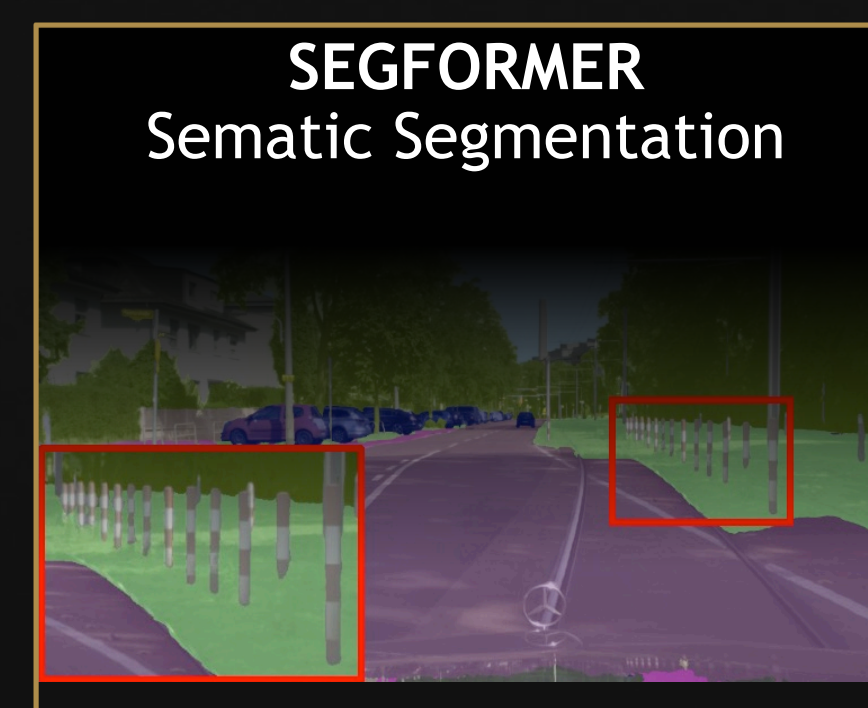


NEXT WAVE OF AI + HPC REQUIRES PERFORMANCE AND SCALABILITY

TRANSFORMERS TRANSFORMING AI + HPC

70%

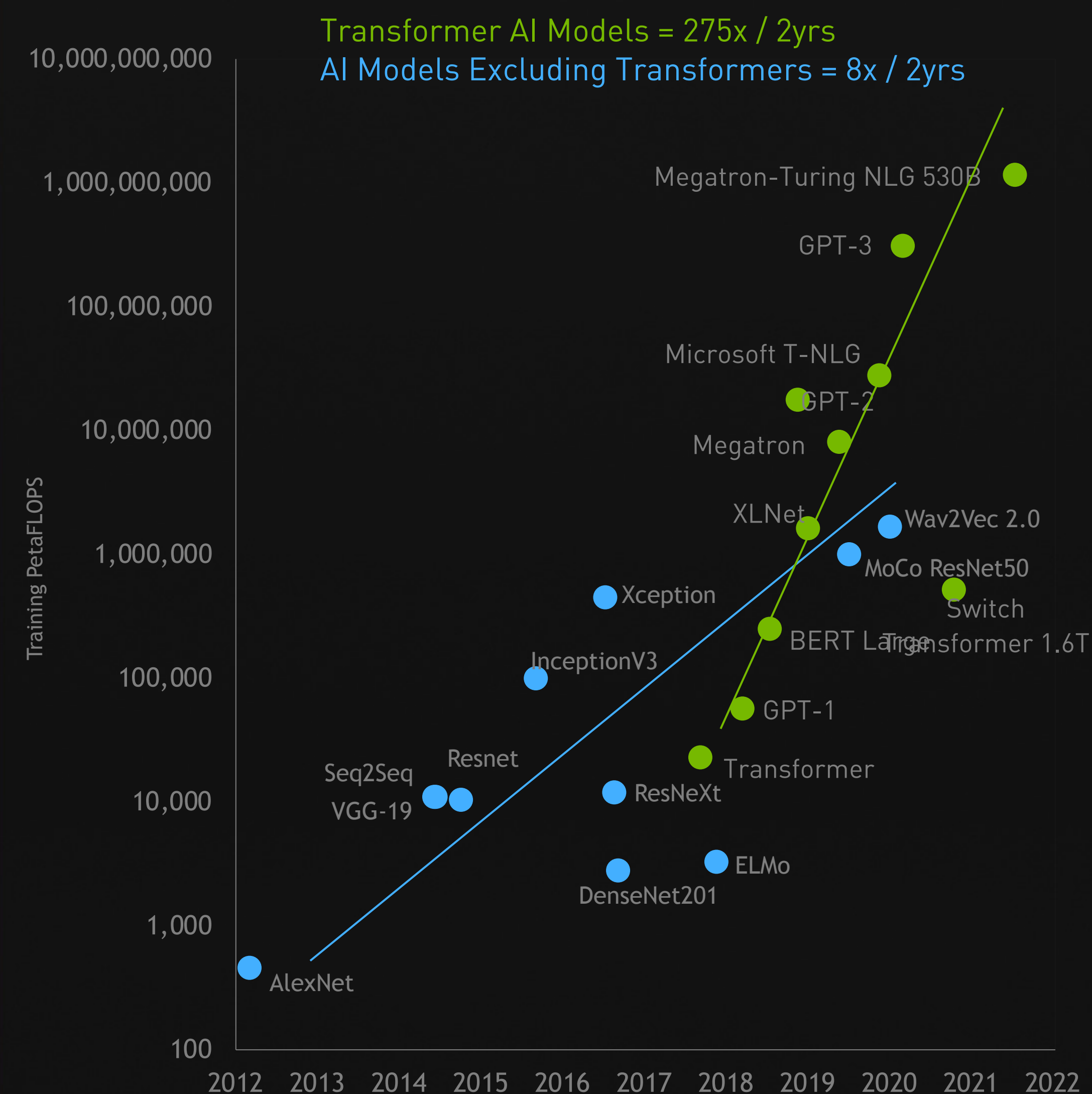
AI Papers In Last 2 Years
Discuss Transformer Models



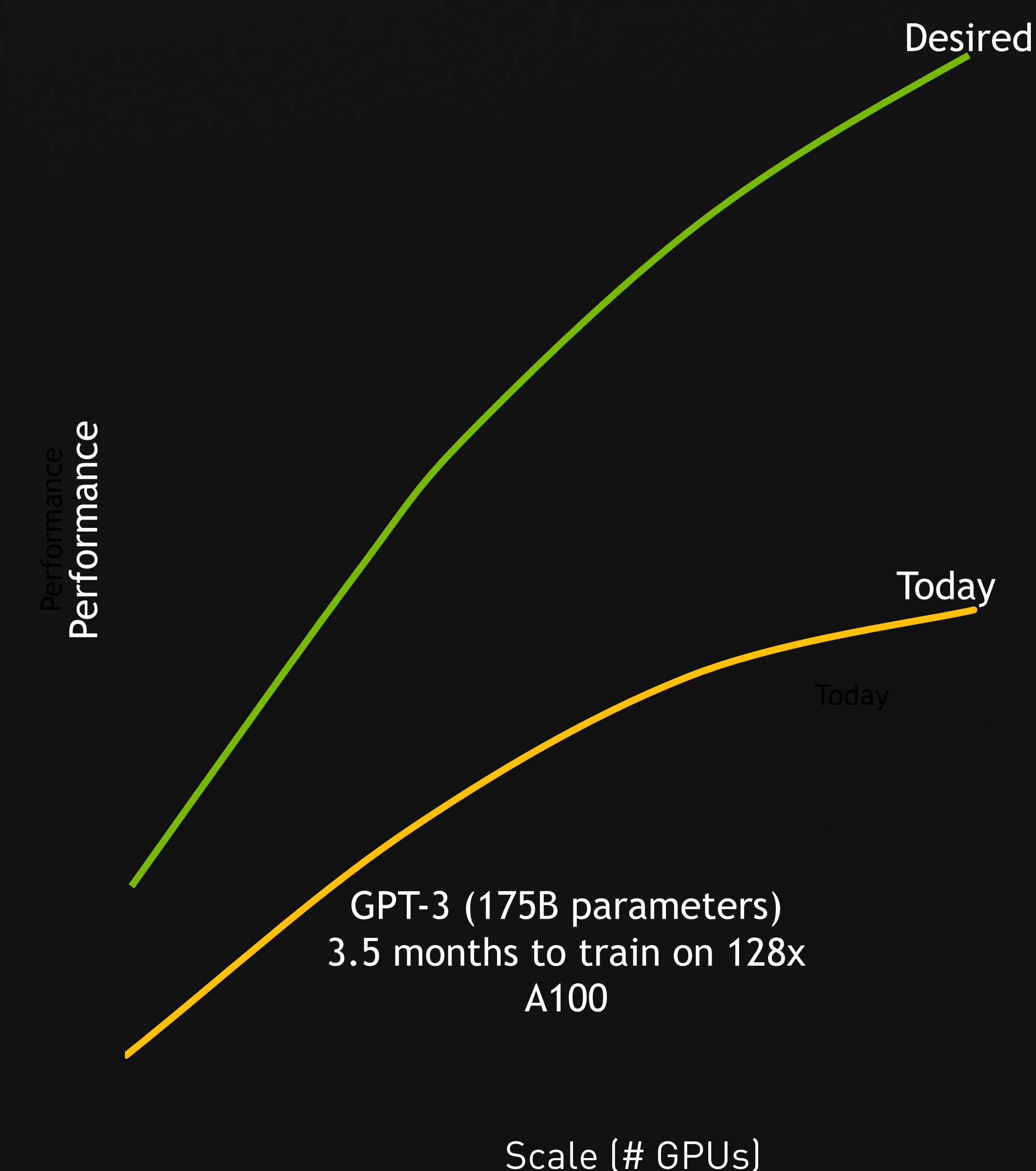
SUPERGLUE LEADERBOARD
Difficult NLU Tasks

Rank	Name	Model	Score
1	Liam Fedus	ST-MoE-32B	91.2
2	Microsoft Alexander v-team	Turing NLR v5	90.9
3	ERNIE Team - Baidu	ERNIE 3.0	90.6

EXPLODING COMPUTE REQUIREMENTS

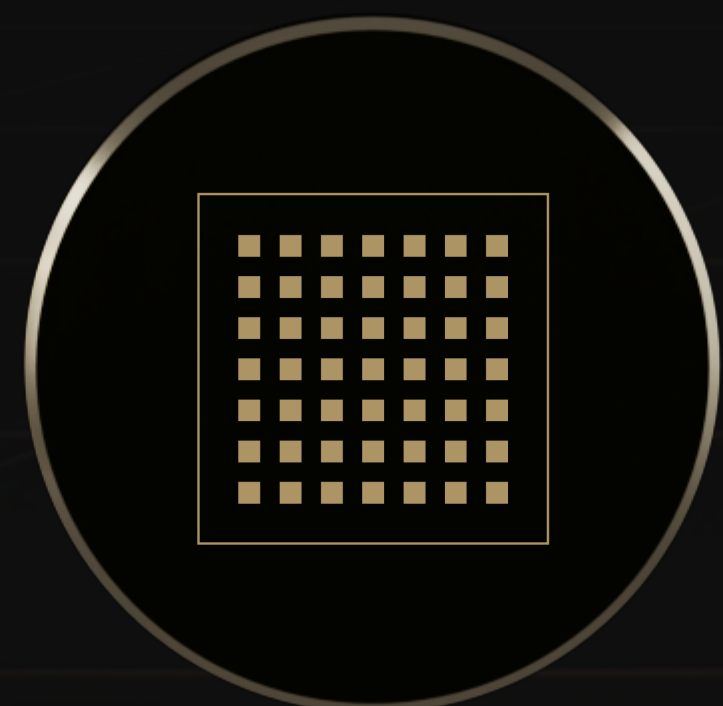
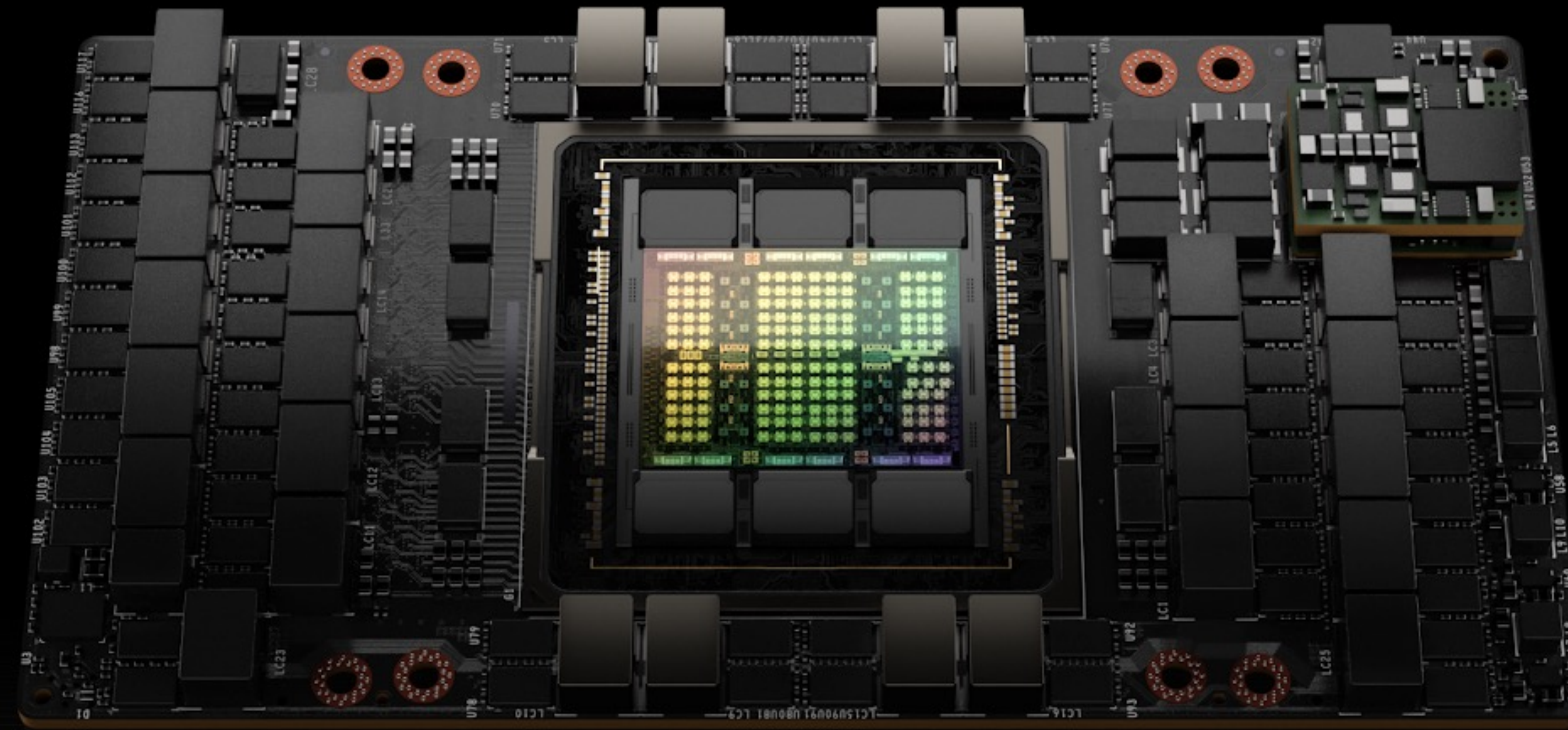


MONTHS TO TRAIN MODELS



ANNOUNCING NVIDIA HOPPER H100

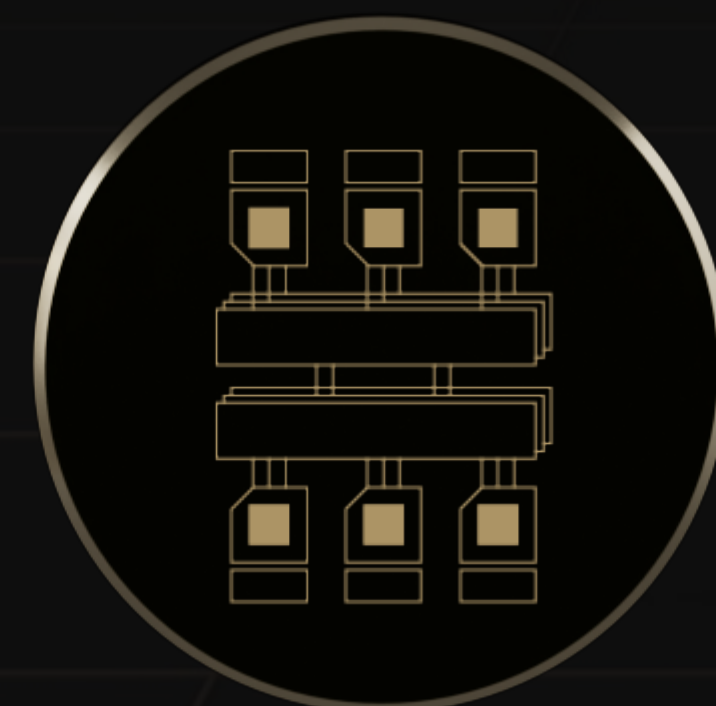
The New Engine for the World's AI Infrastructure



**WORLD'S MOST
ADVANCED CHIP**
80B Transistors



TRANSFORMER ENGINE
6X Transformer Performance



4TH GEN NVLINK
7X PCIe Gen 5



**CONFIDENTIAL
COMPUTING**
Secure Data and AI Models in Use



**2ND GEN
MULTI-INSTANCE GPU**
7X Secure Tenants



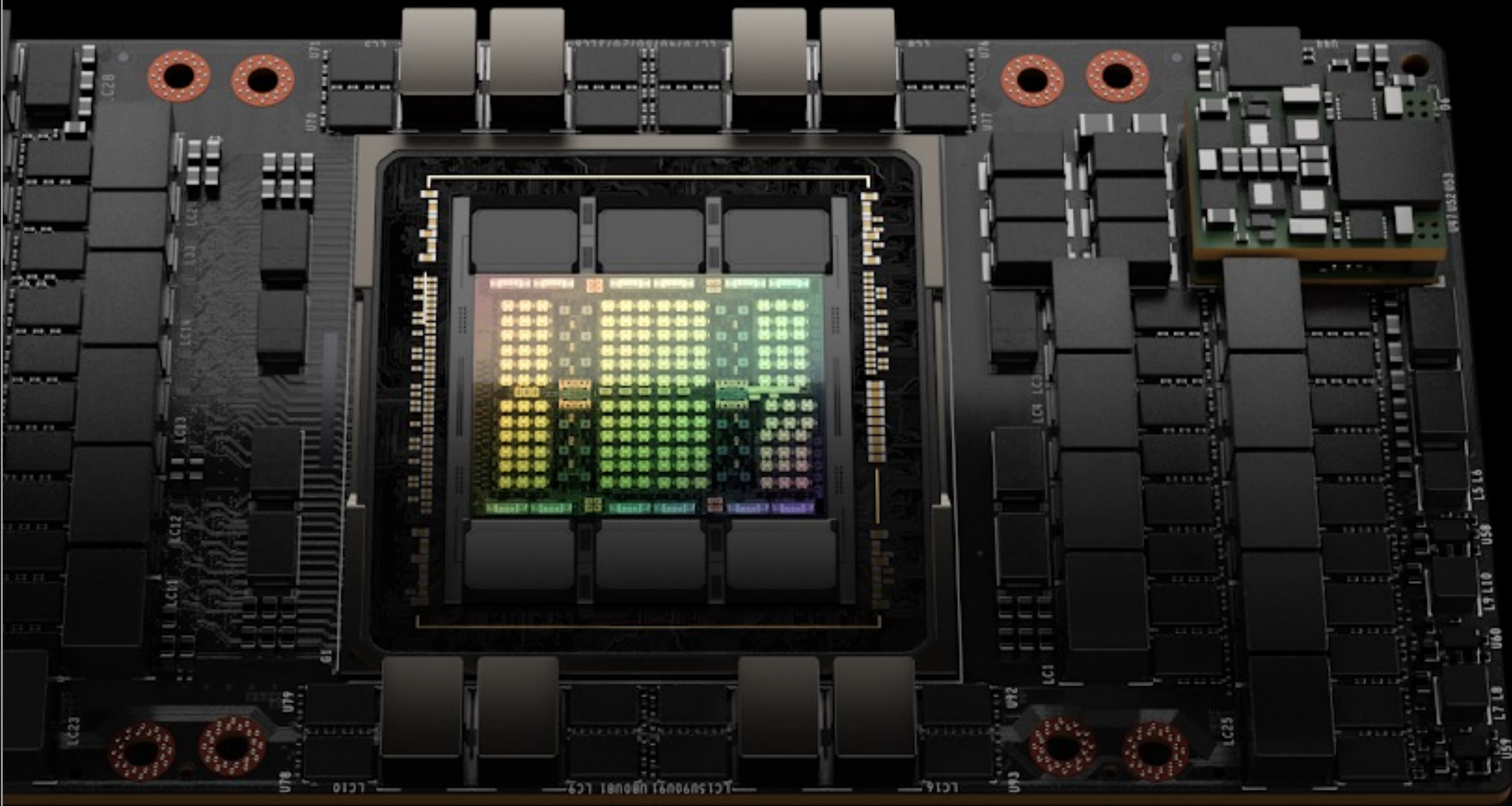
DPX INSTRUCTIONS
7X Dynamic
Prog Performance

ANNOUNCING NVIDIA HOPPER H100

Unprecedented Performance, Scalability, And Security For Every Data Center

NVIDIA H100

FP8	4,000 TFLOPS	6X
FP16	2,000 TFLOPS	3X
TF32	1,000 TFLOPS	3X
FP64	60 TFLOPS	3X
HBM3	3 TB/s	1.5X
PCI Gen5	128 GB/s	2X
4 th Gen NVLink	900 GB/s	1.5X
TDP	700W (SXM)	



Custom TSMC 4N Process | 4.9 TB/s Total External B/W

ANNOUNCING NVLINK SWITCH FOR MULTI-NODE NVLINK CONNECTIVITY

Purpose Built Network, Scales-up to 256 GPU

NVLINK SWITCH SYSTEM

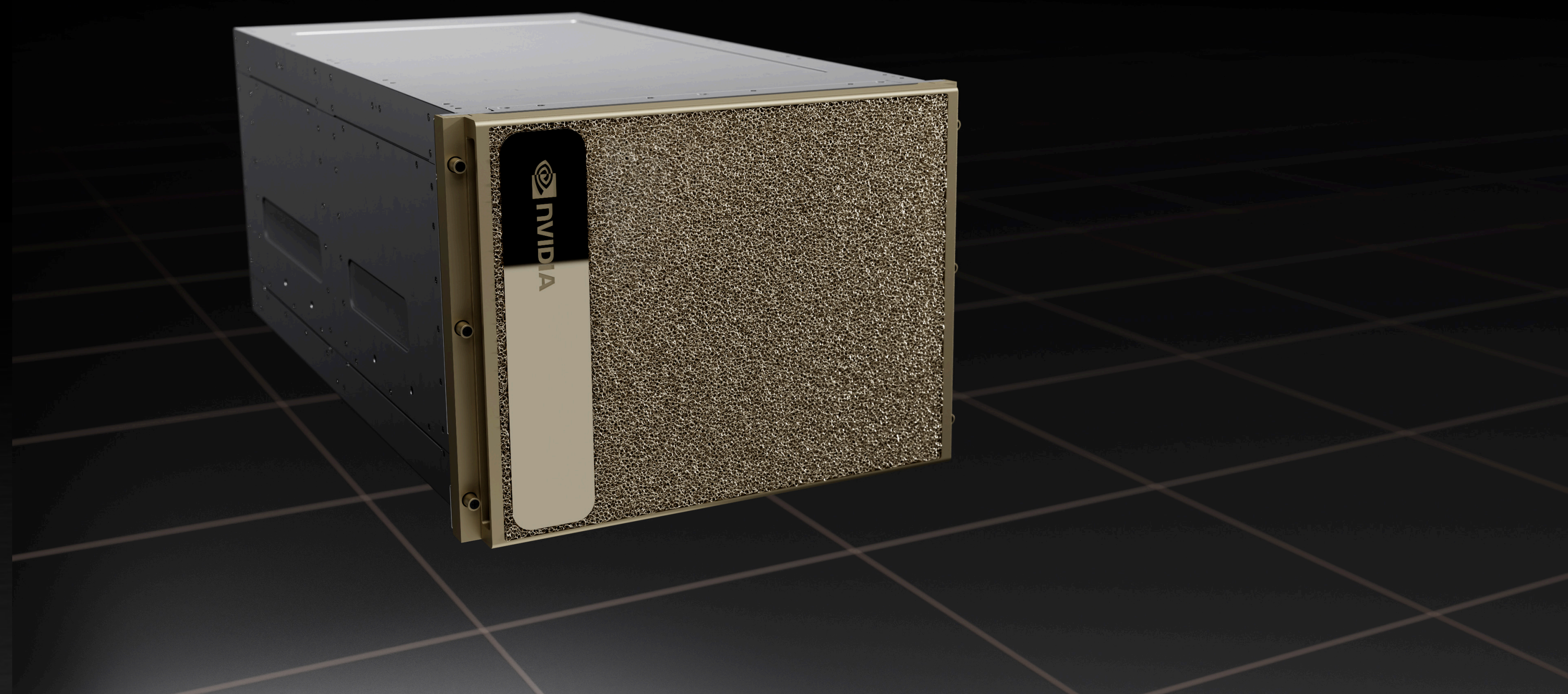
In-Network Compute	192 TFLOPS	15X
Bisection Bandwidth	70.4 TB/s	11X
NVLINK Domain	256 GPUs	32X
4 th Gen NVLink	900 GB/s	



ANNOUNCING NVIDIA DGX H100 SYSTEMS

Best Of Breed Infrastructure For AI Development Built On NVIDIA DGX

NVIDIA DGX H100



8xNVIDIA H100 | 32 PFLOPS AI (6X)
0.5 PFLOPS FP64 (3X) | 640 GB HBM3 | 3.6 TB/s BISECTION B/W

4th Generation of the World's Most Successful
Platform Purpose-Built for Enterprise AI

DGX SuperPOD WITH DGX H100



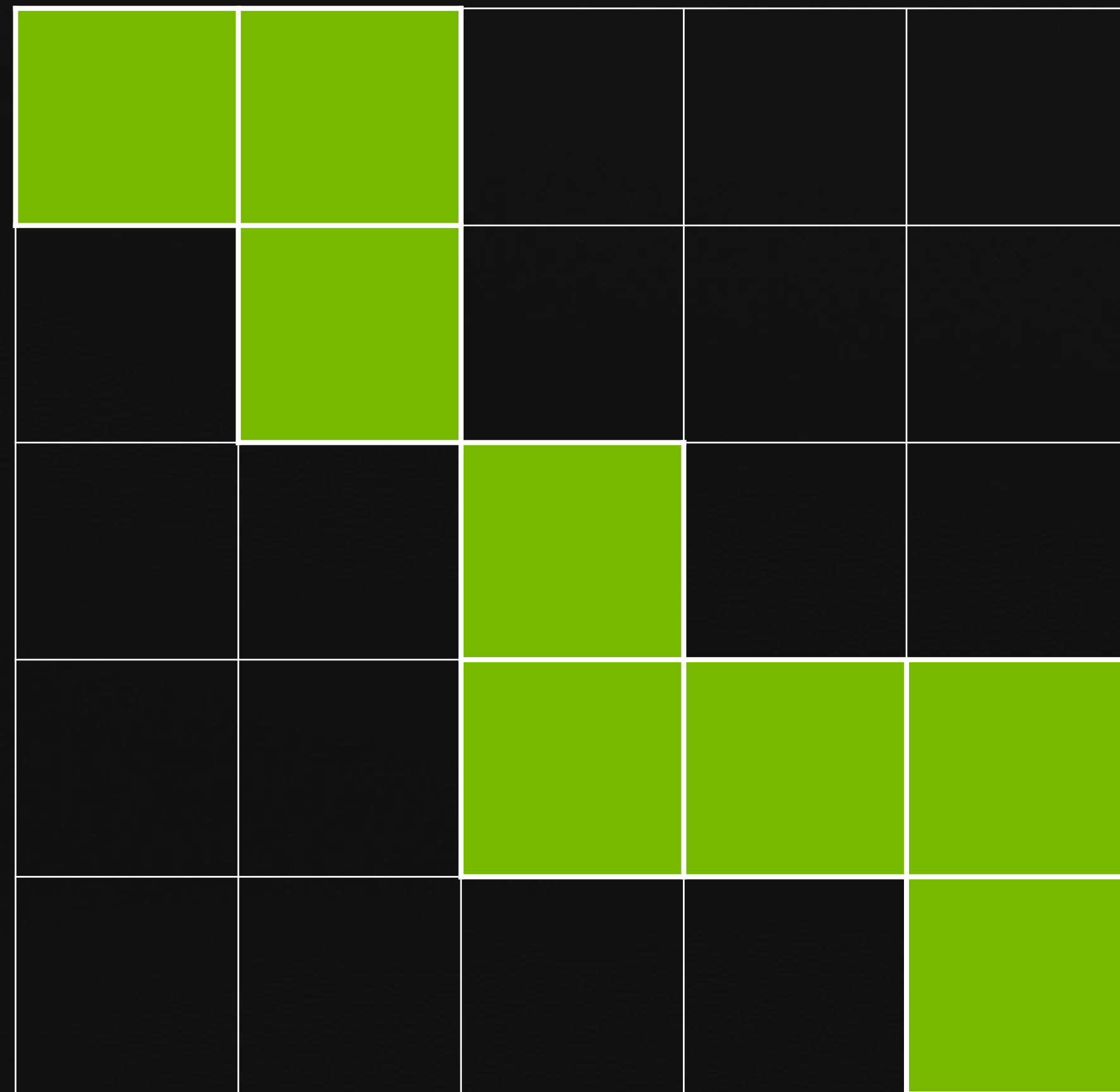
32 DGX H100 | 256 H100 | NVLINK SWITCH SYSTEM
QUANTUM-2 IB | 20TB HBM3 | 70.4 TB/s BISECTION B/W

1 ExaFLOPS of AI Performance in 32 Nodes
Scale as large as needed in 32 node increments

NEW DPX INSTRUCTIONS

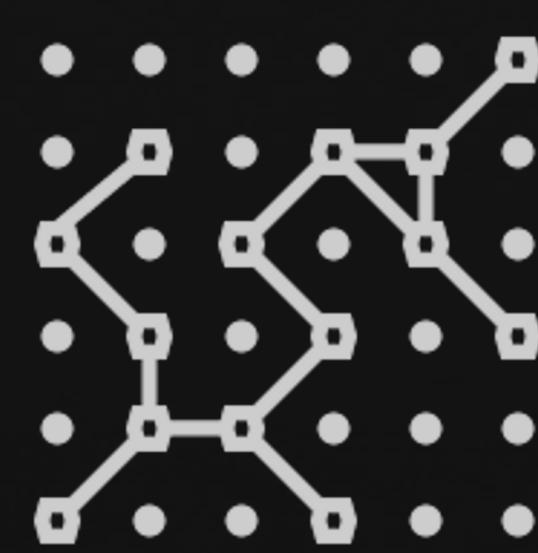
DYNAMIC PROGRAMMING

Exponential to polynomial time problem solving



A BROAD RANGE OF USE CASES

from genomics to routing optimization



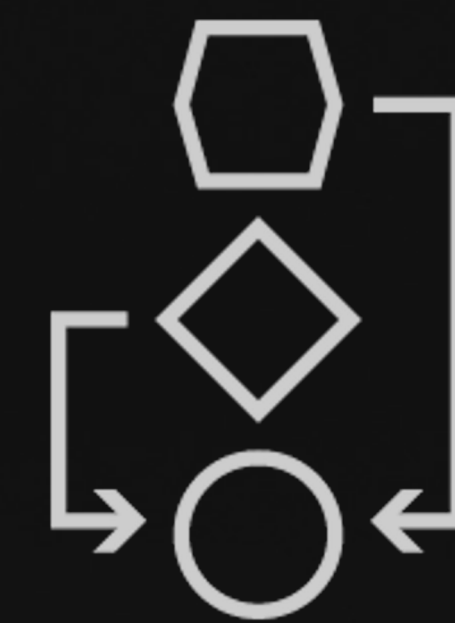
Optimization



Omics



Graph Analytics

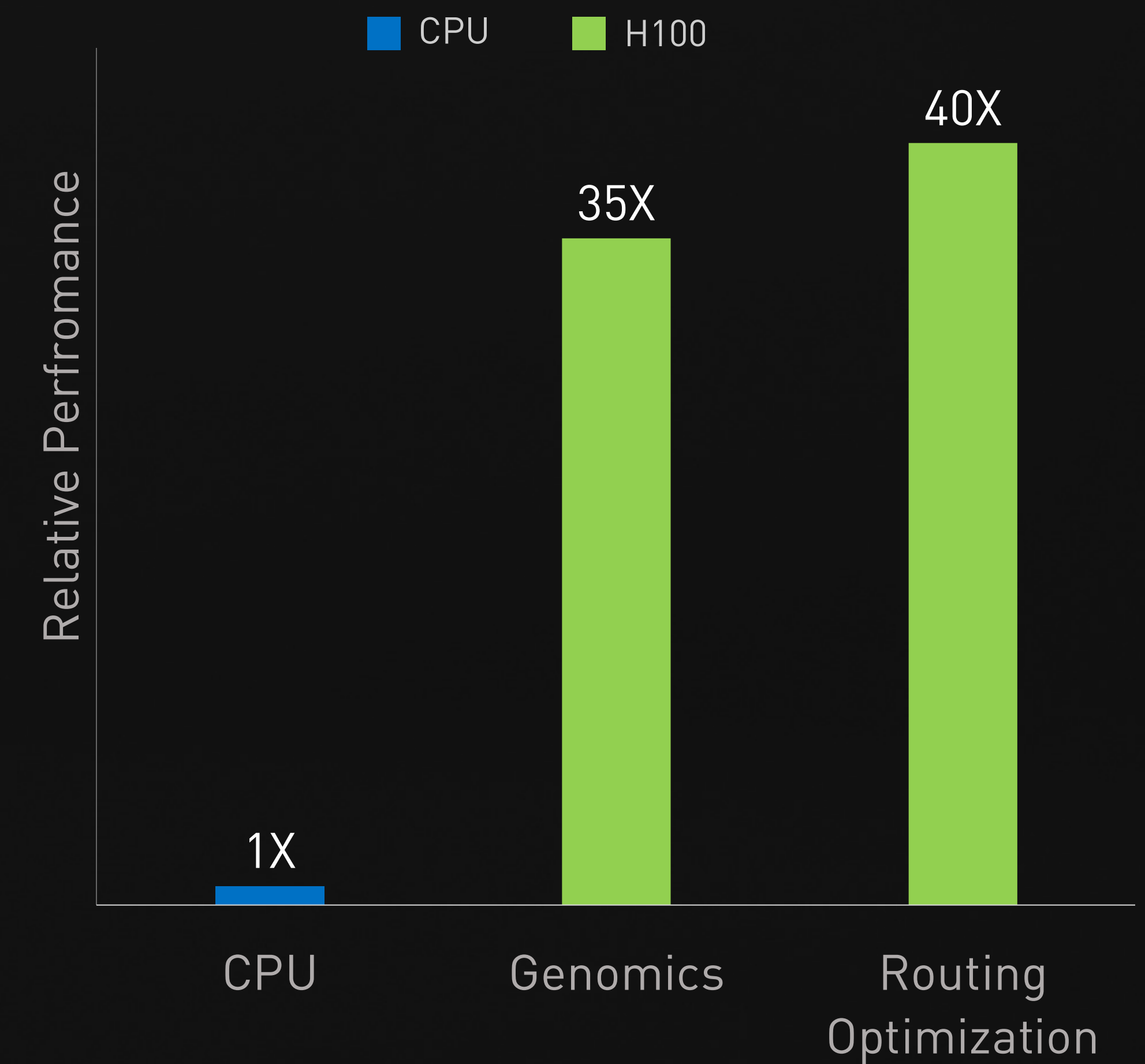


Data Processing

REAL-TIME PERFORMANCE

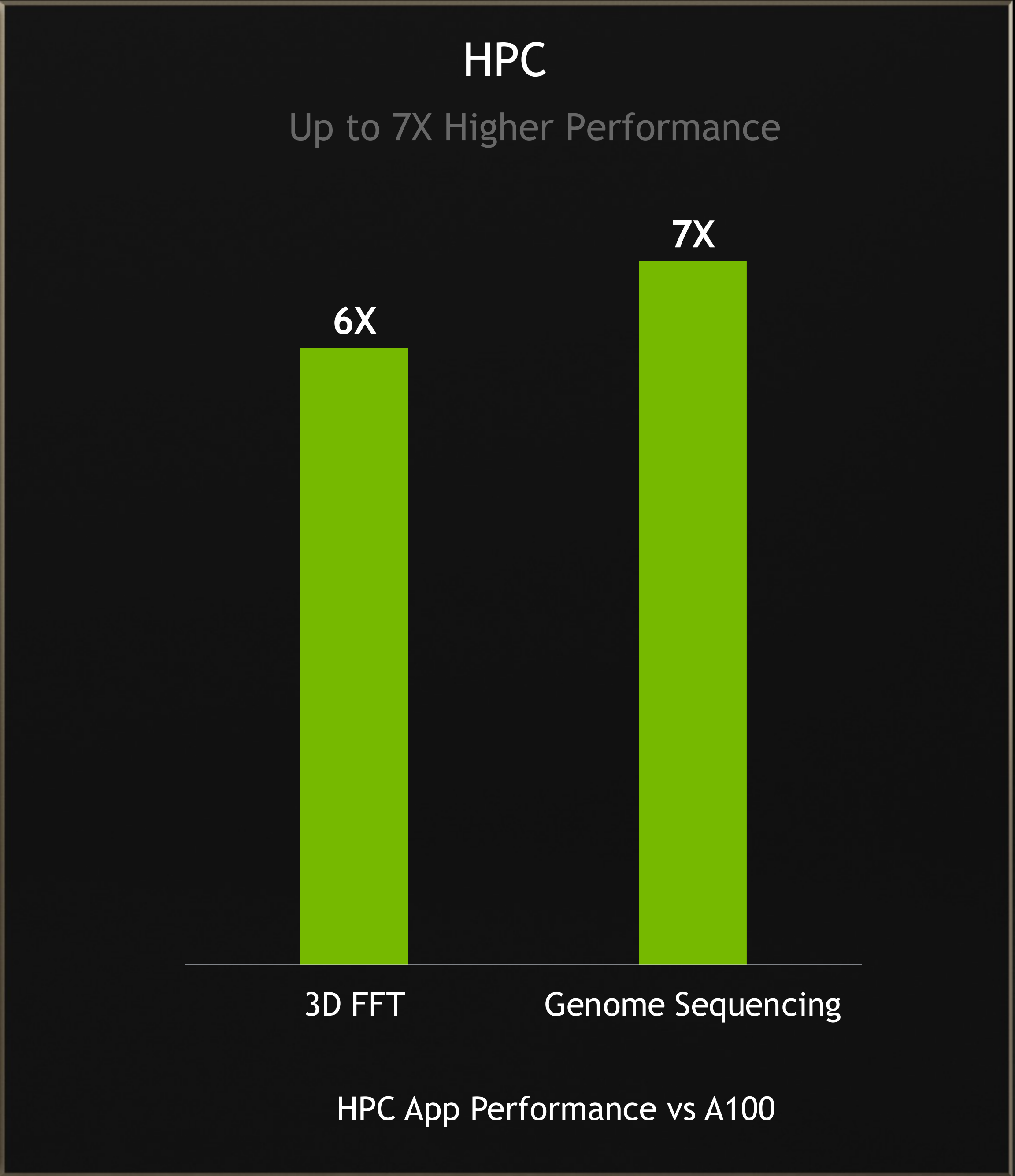
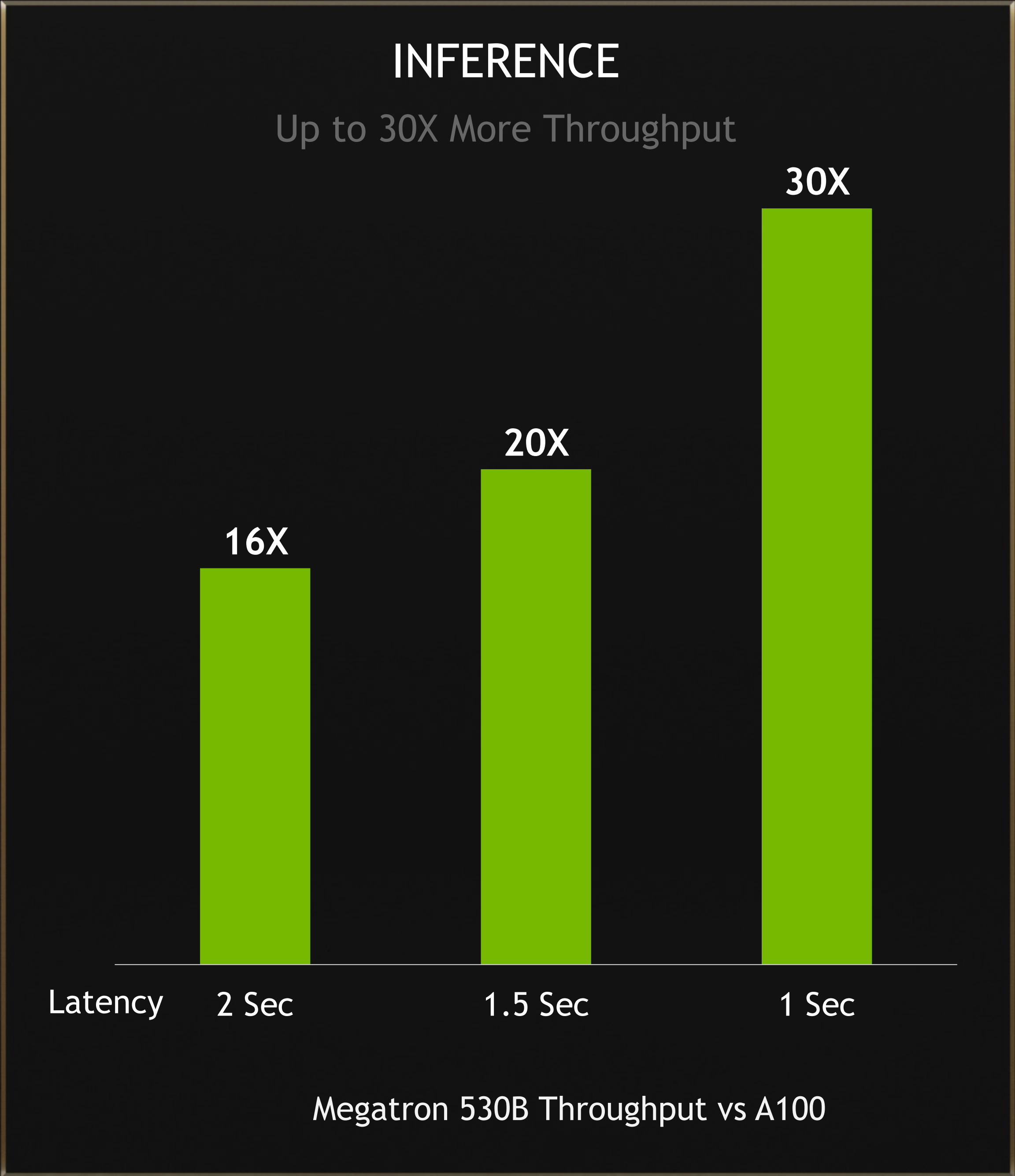
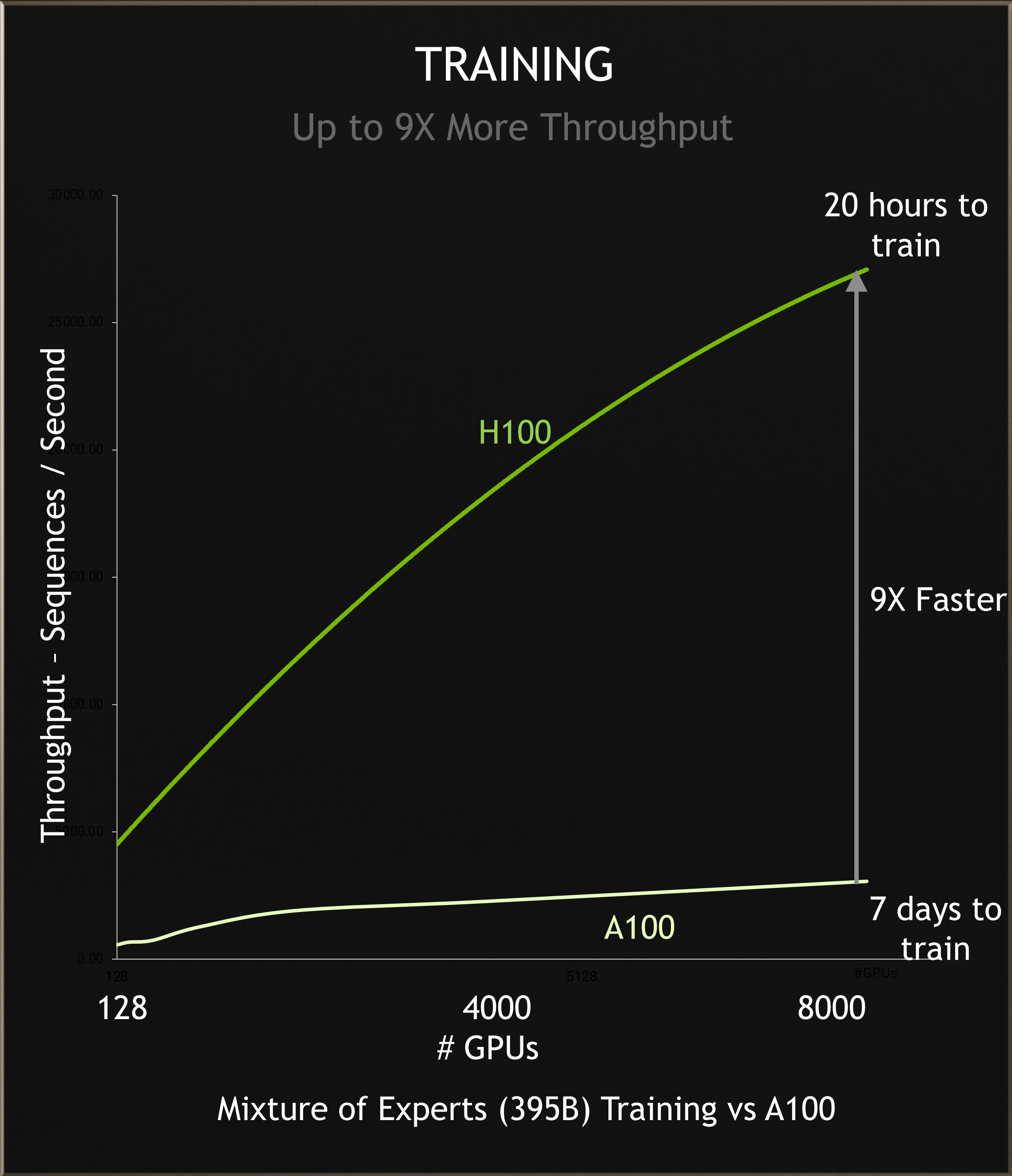
40X speedup

Dynamic Programming Algorithm Speedup



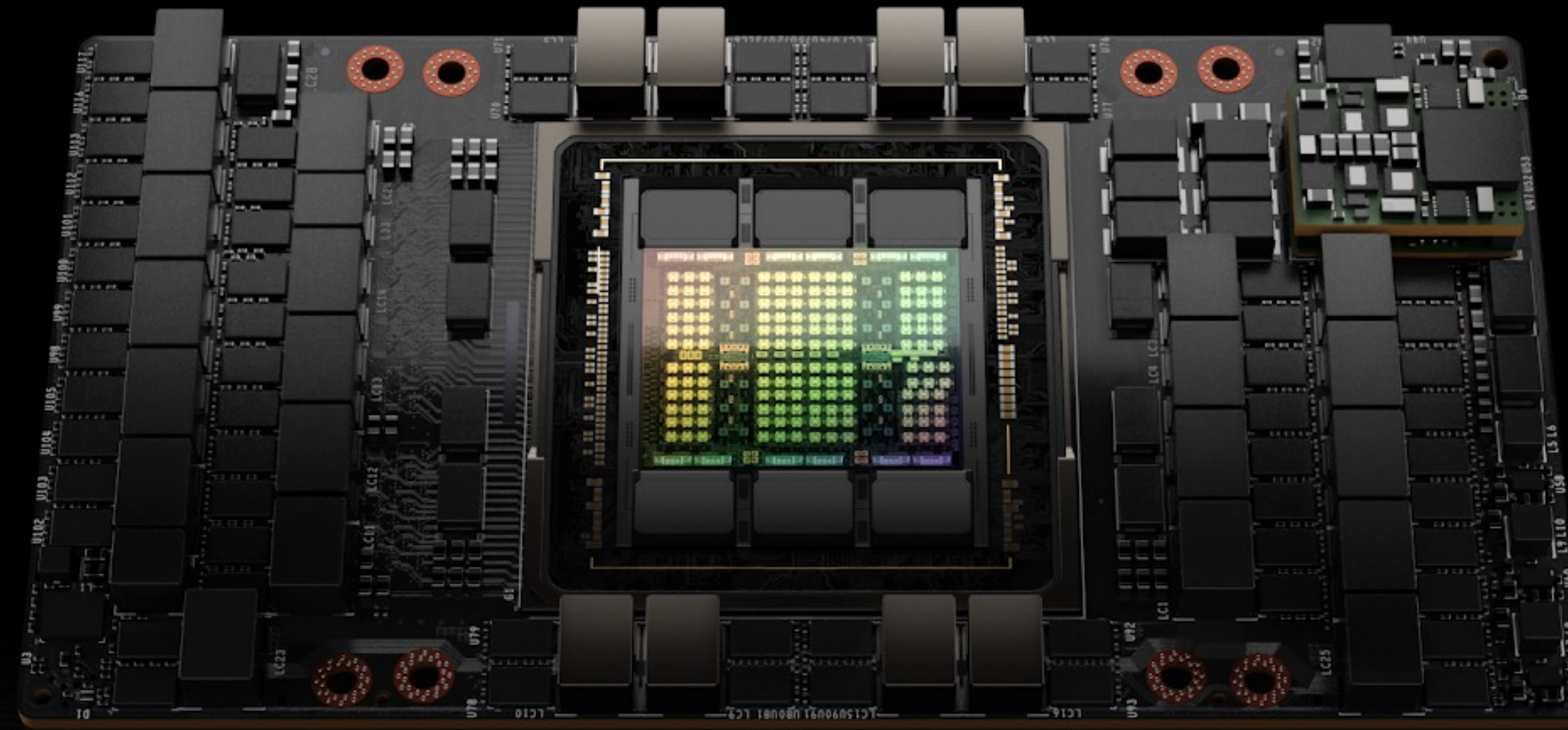
H100 DELIVERS A MASSIVE PERFORMANCE LEAP

Performance And Scalability For Next Generation AI and HPC Breakthroughs



Projected performance subject to change
Training Mixture of Experts (MoE) Transformer Switch-XXL variant with 395B parameters on 1T token dataset | Inference on Megatron 530B parameter model based chatbot for input sequence length=128, output sequence length =20 | HPC 3D FFT (4K^3) throughput | A100 cluster: HDR IB network | H100 cluster: NVLink Network, NDR IB
HPC Genome Sequencing (Smith-Waterman) | 1 A100 | 1 H100

ANNOUNCING H100 COMING TO TACC SUPERCOMPUTING CENTER



ANNOUNCING NVIDIA EOS SUPERCOMPUTER

The World's Most Advanced AI Infrastructure

NVIDIA Eos

DGX SuperPOD powered by 576 DGX H100 systems | 500 Quantum-2
IB Switches | 360 NVLink Switches

FP8	18 EFLOPS	6X
FP16	9 EFLOPS	3X
FP64	275 PFLOPS	3X
In-Network Compute	3.7 PFLOPS	36X
Bisection Bandwidth	230 TB/s	2X
NVLINK Domain	256 GPUs	32X

Cloud Native | Performance Isolation | Multi-Tenant

Blueprint for OEM and cloud partner offerings



NVIDIA H100 AT EVERY SCALE

To Be Available from World's Leading System Makers
and Cloud Providers

DGX H100 SUPERPOD

Highest Performance Data Centers

H100 HGX and DGX

Highest Performance Servers

H100 CNX and PCIE

Mainstream Servers

CLOUD PROVIDERS

Alibaba Cloud

aws

BAIDU AI CLOUD

Google Cloud

Microsoft
Azure

ORACLE
Cloud Infrastructure

Tencent Cloud

SYSTEM PROVIDERS

Atos

DELL Technologies

FUJITSU

GIGABYTE

H3C

Hewlett Packard
Enterprise

inspur

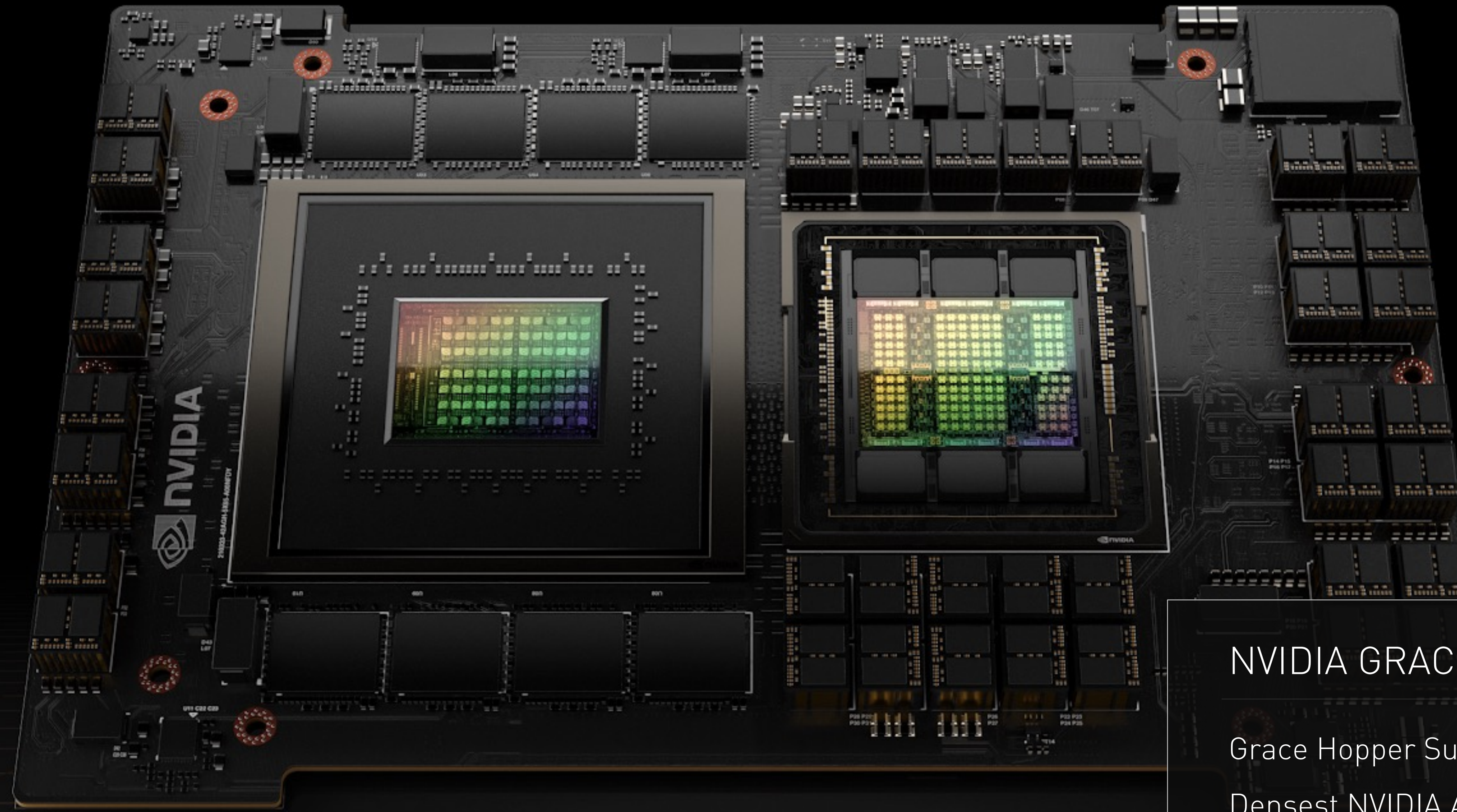
Lenovo

Nettrix 宇畅

SUPERMICRO

ANNOUNCING GRACE HOPPER

CPU+GPU Designed for Giant Scale AI and HPC



NVIDIA GRACE HOPPER

Grace Hopper Superchip Module

Densest NVIDIA Accelerated Computing System

New NVLINK Chip-to-Chip Coherent Inference

900 GB per Second

GRACE HOPPER SUPERCOMPUTER FOR HPC AND AI

20 Exaflops of AI

Powered by NVIDIA Grace CPUs and
H100 GPUs

HPC and AI for Scientific and Commercial Apps

Advance Weather, Climate, and Material Science



CSCS

Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre



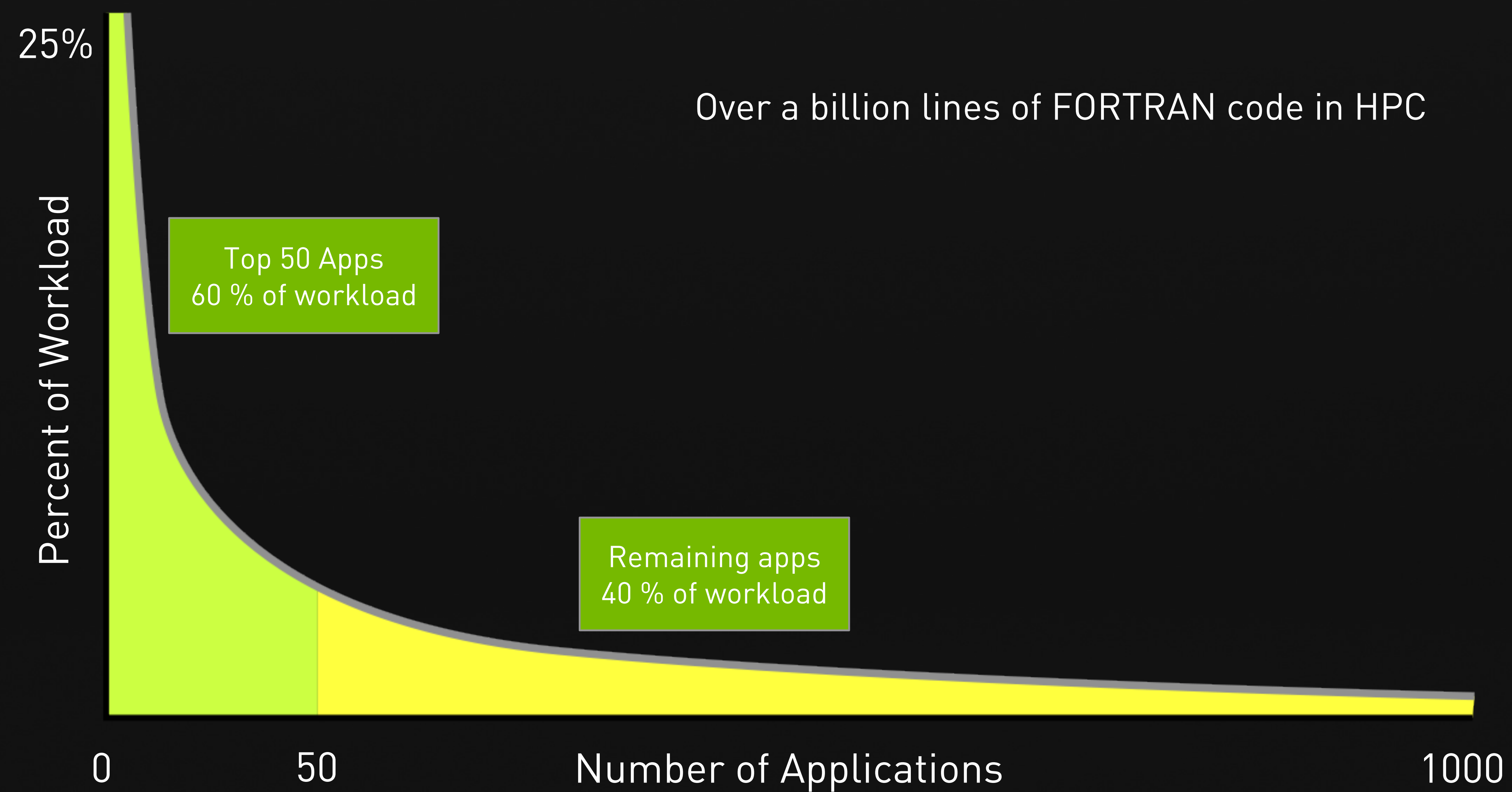
**Hewlett Packard
Enterprise**



nVIDIA

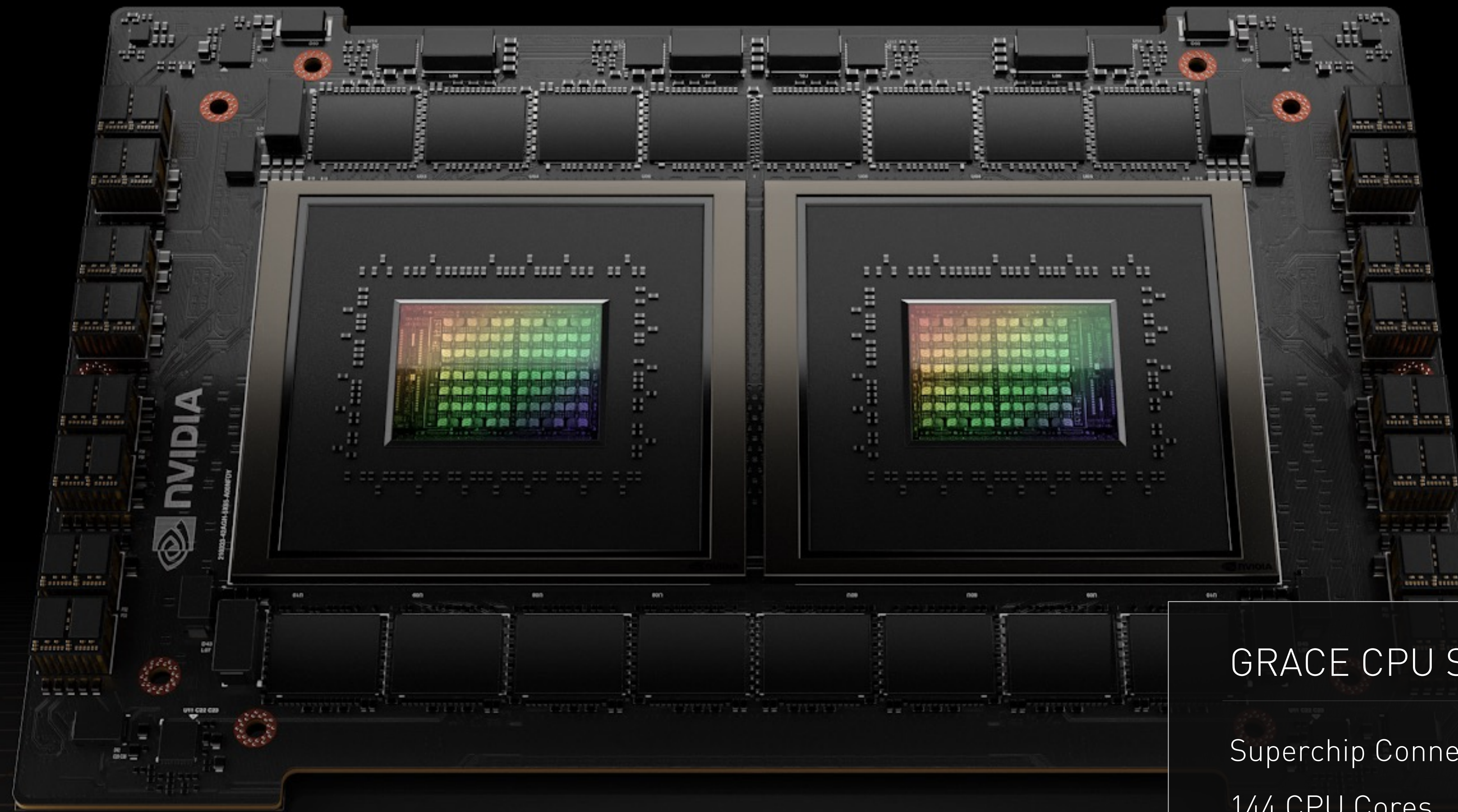


LONG TAIL OF NON-ACCELERATED HPC APPLICATIONS



ANNOUNCING GRACE CPU SUPERCHIP

The Full Power of the Grace



GRACE CPU SUPERCHIP

Superchip Connected by NVLINK-C2C

144 CPU Cores

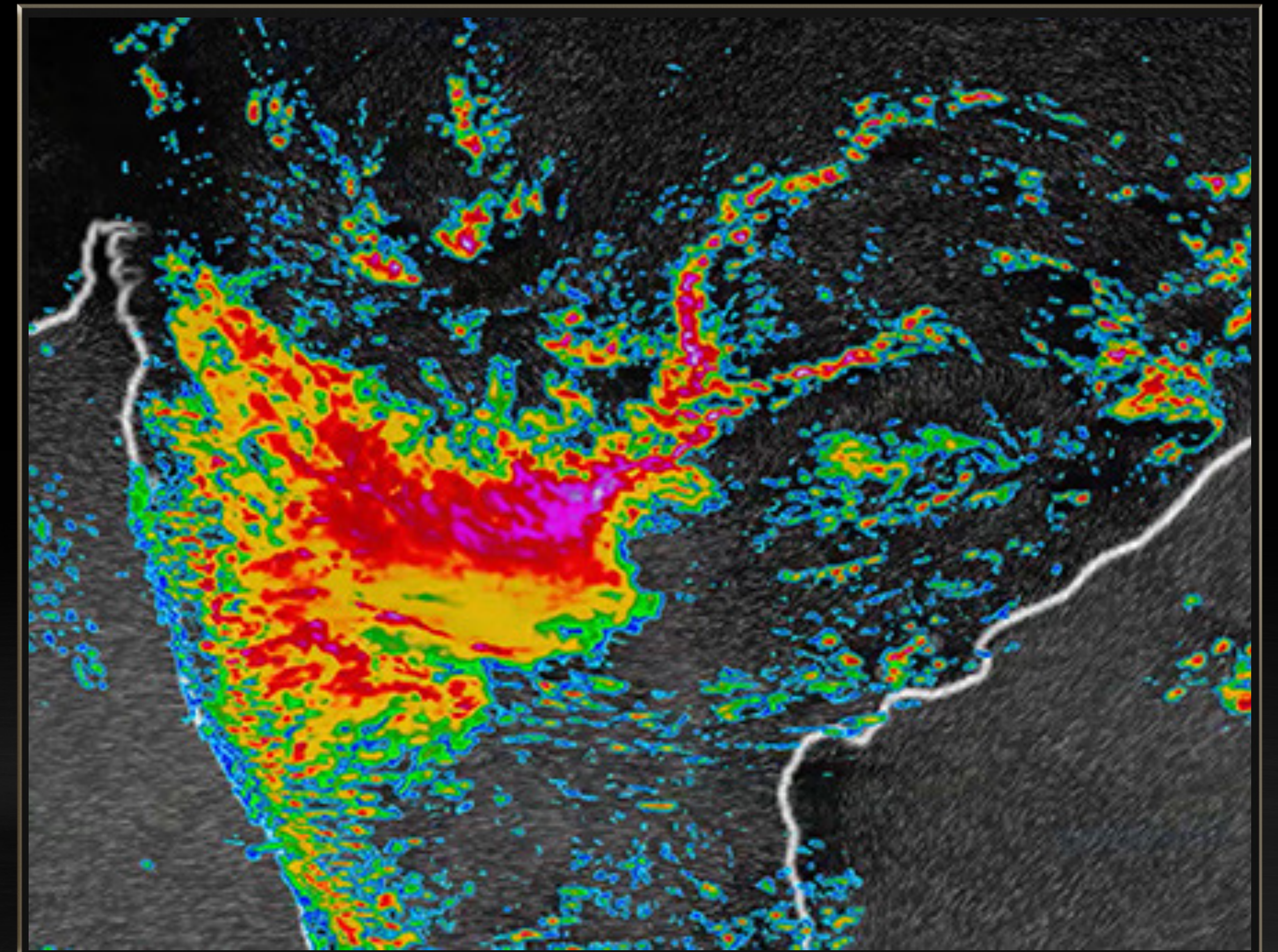
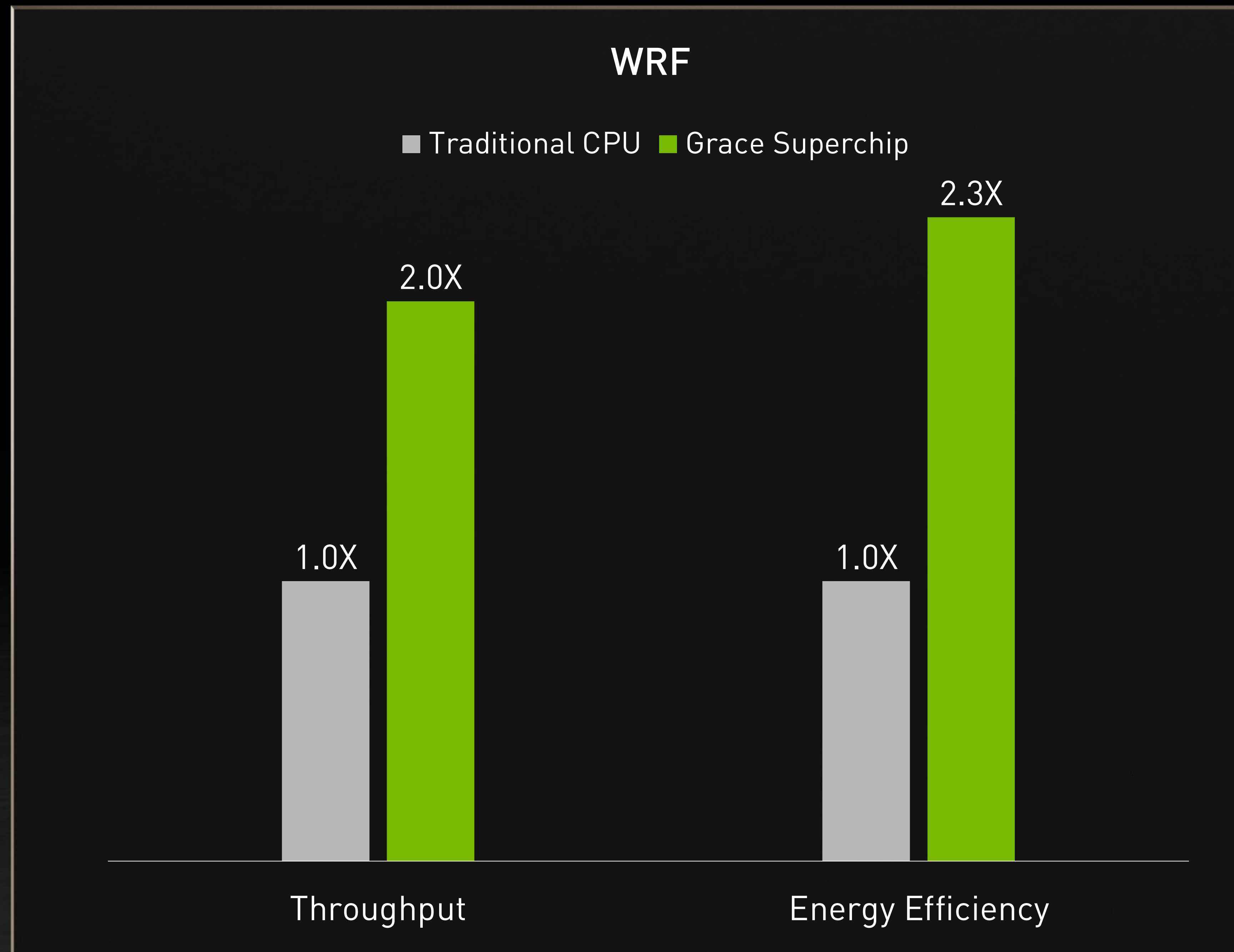
1 Terabytes per Second LPDDR5x with ECC

396MB On-Chip Cache

SPECrate 2017_int_base over 740 est.

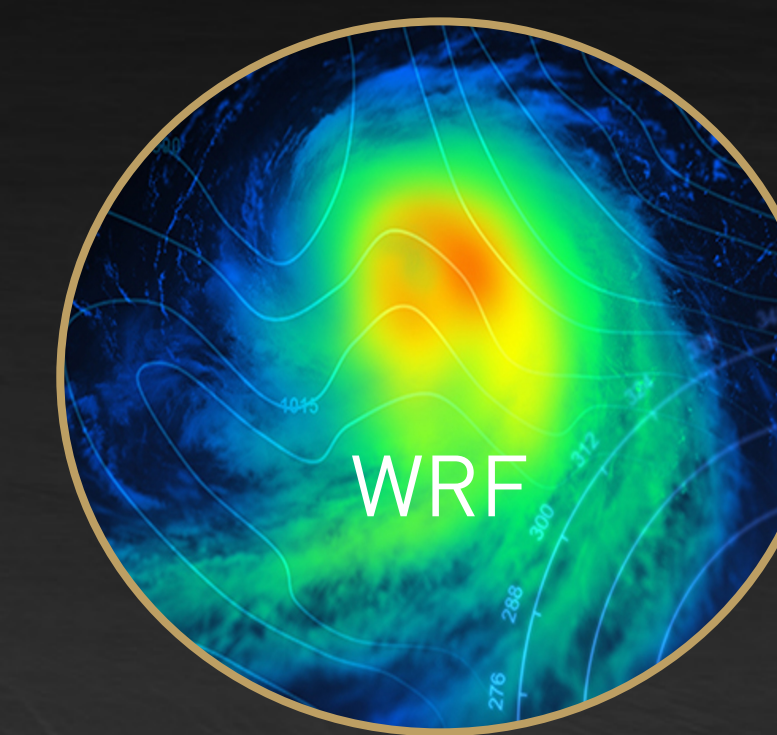
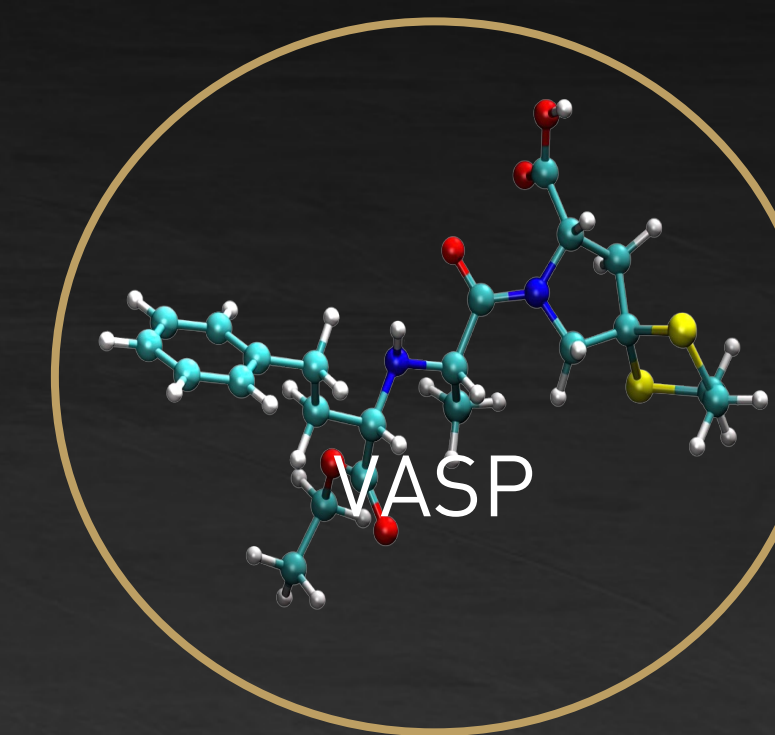
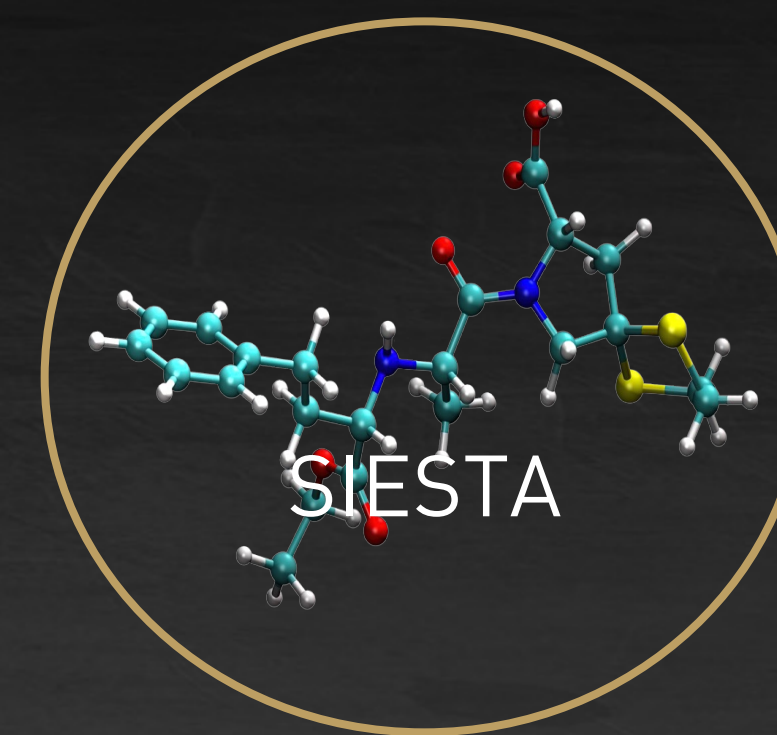
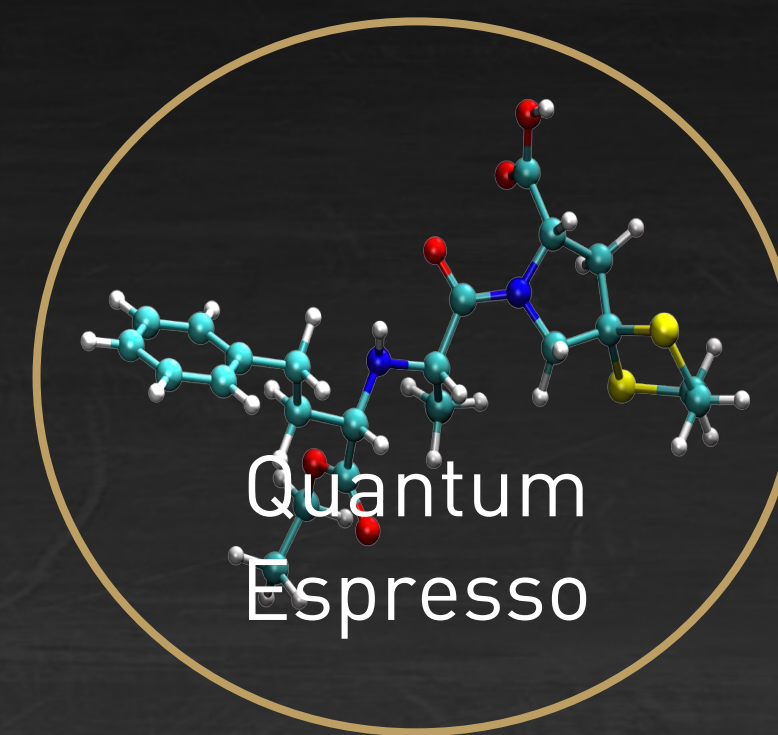
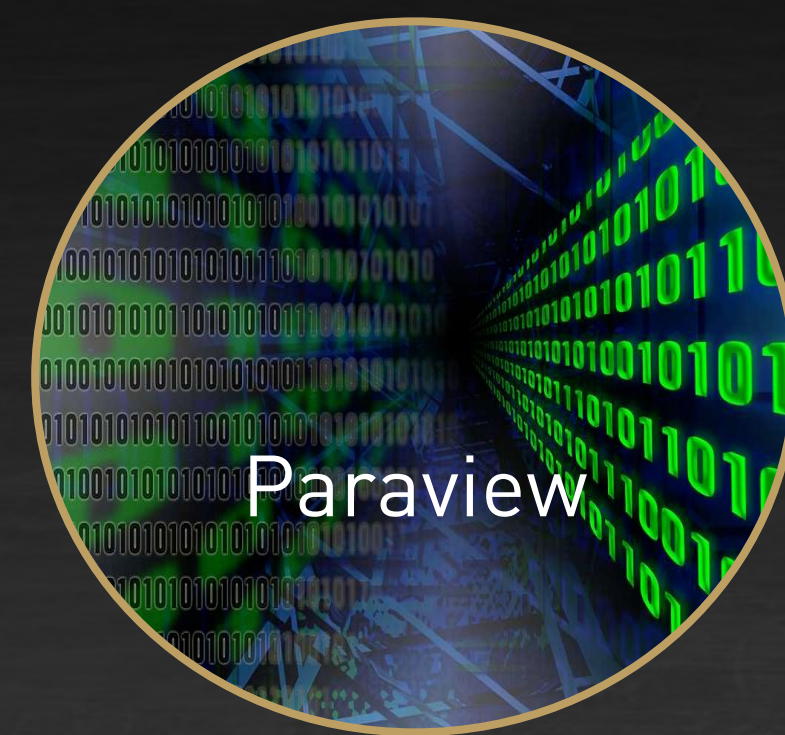
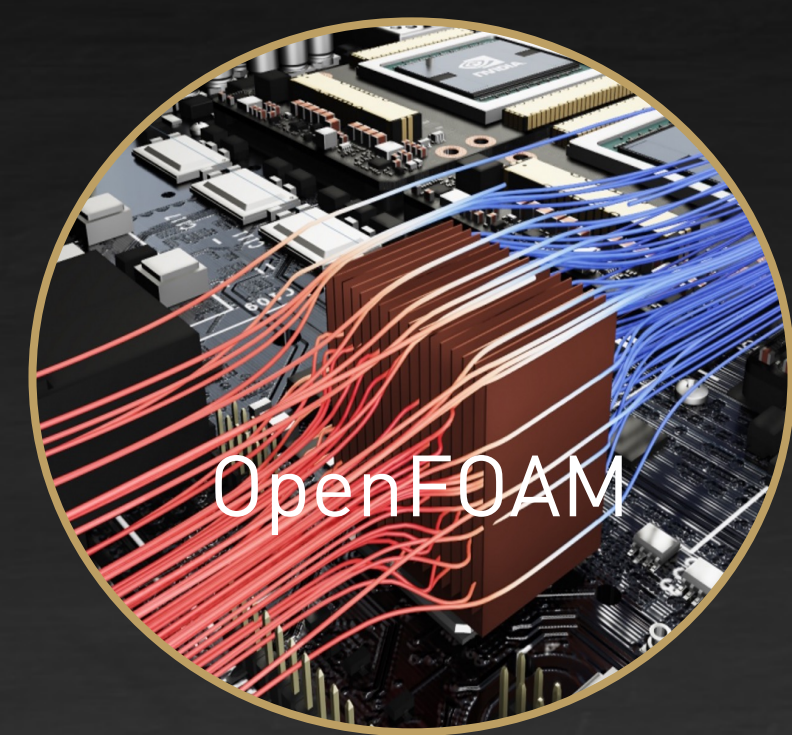
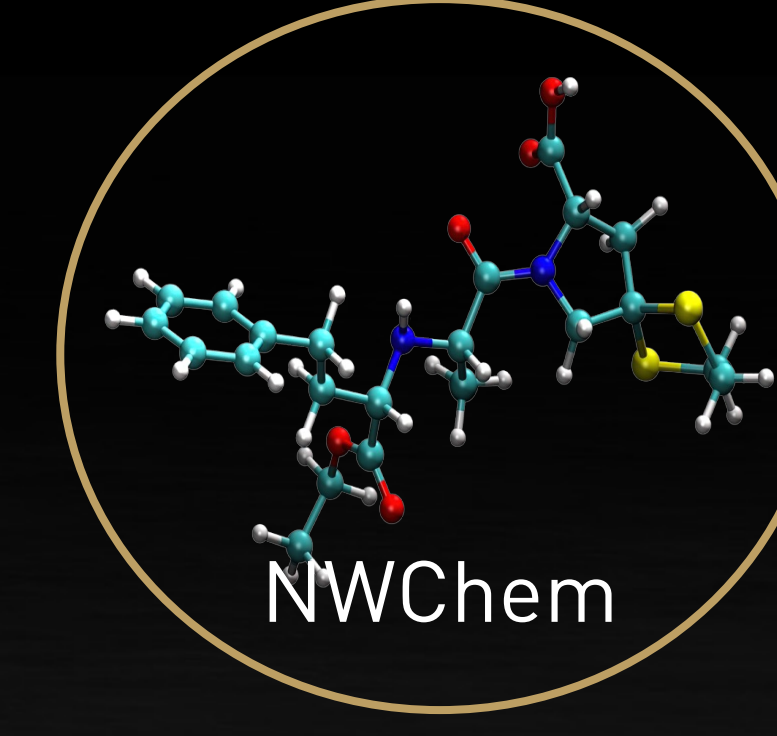
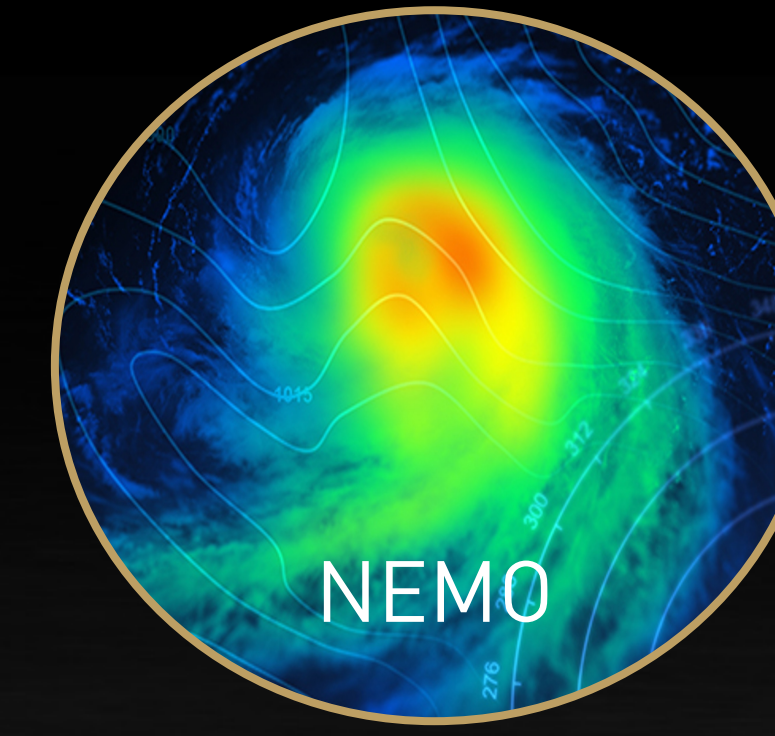
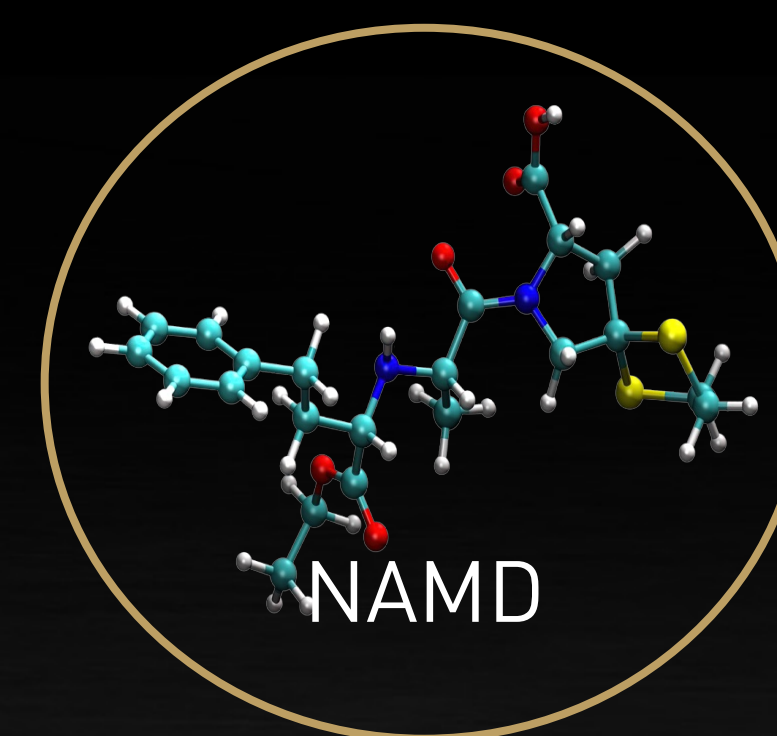
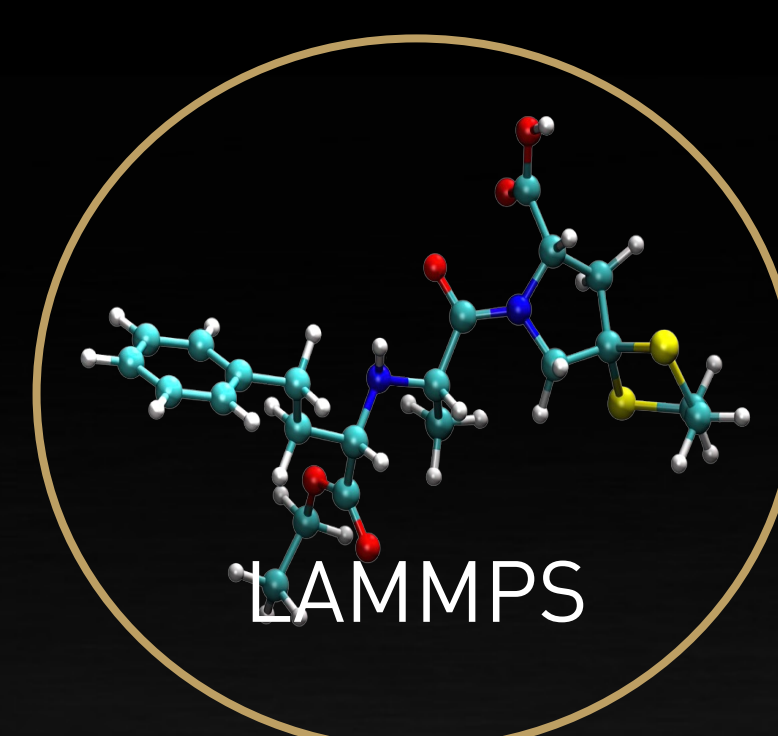
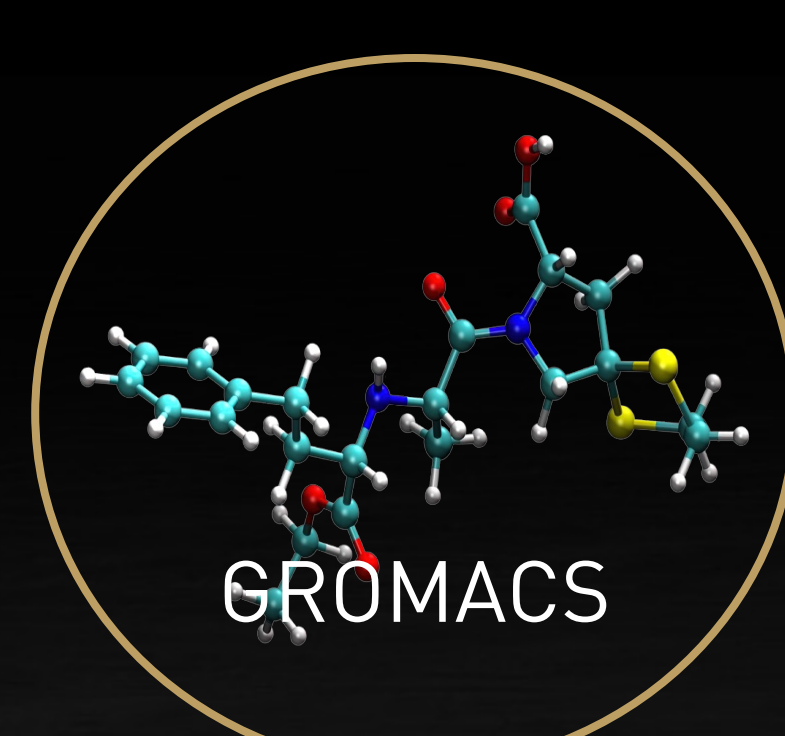
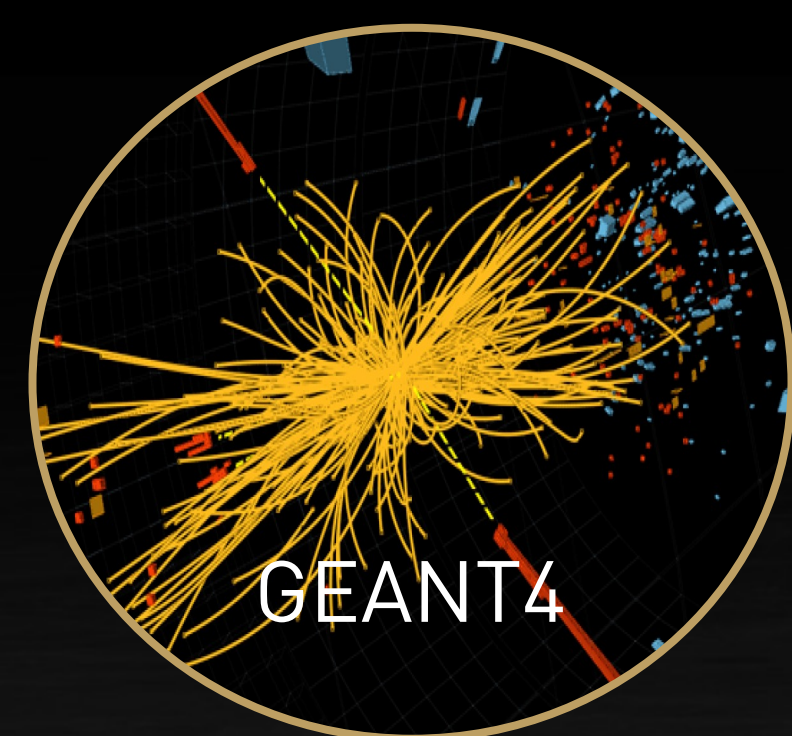
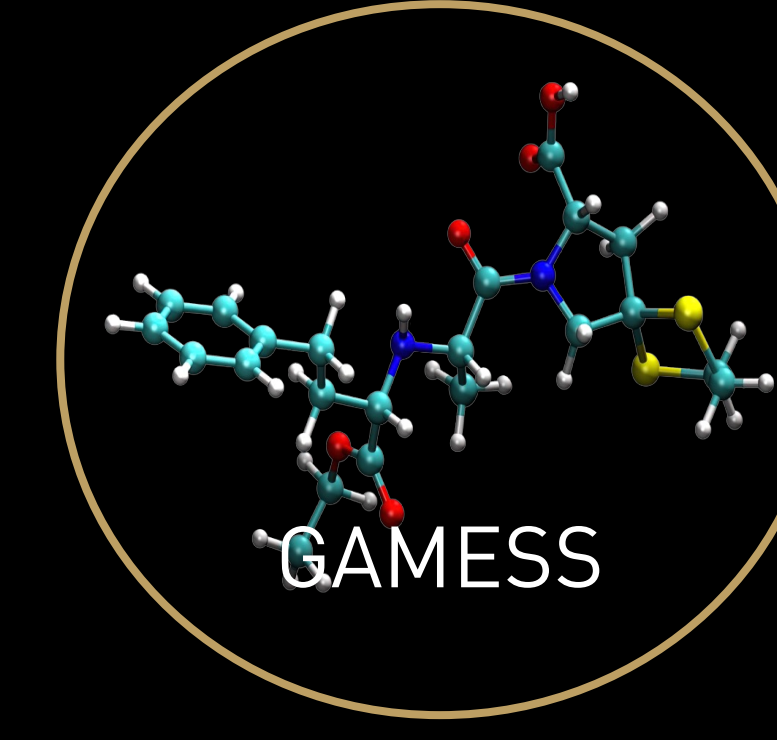
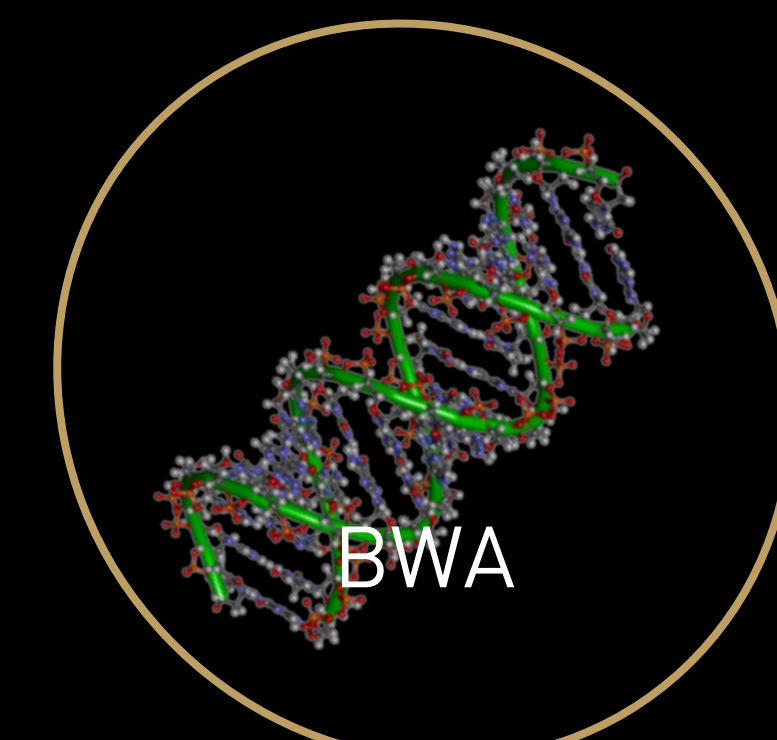
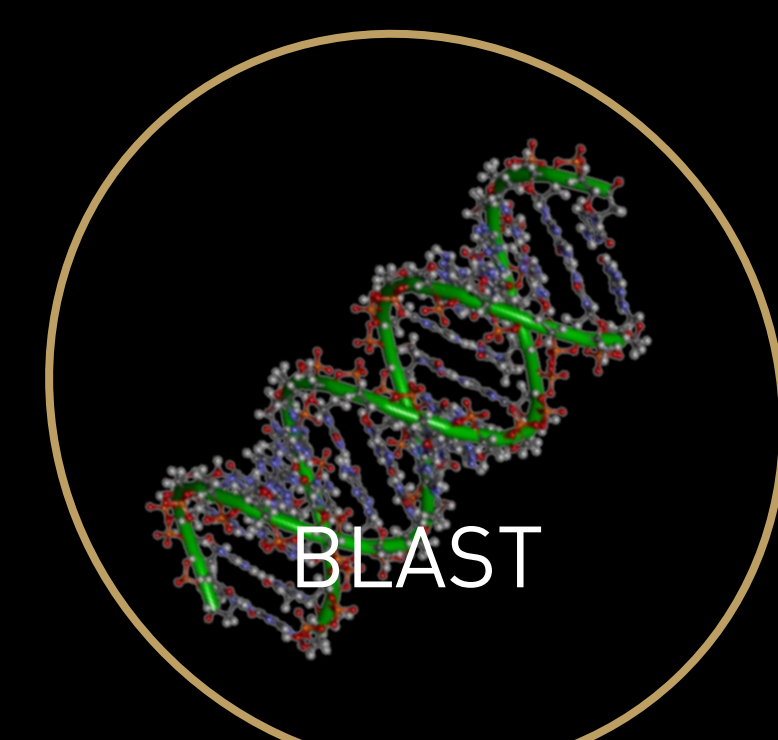
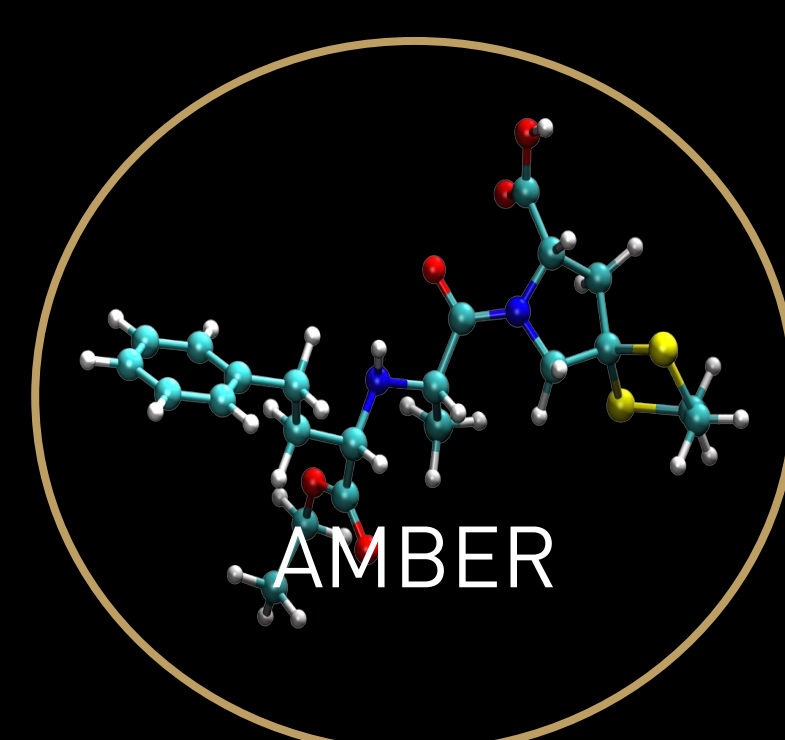
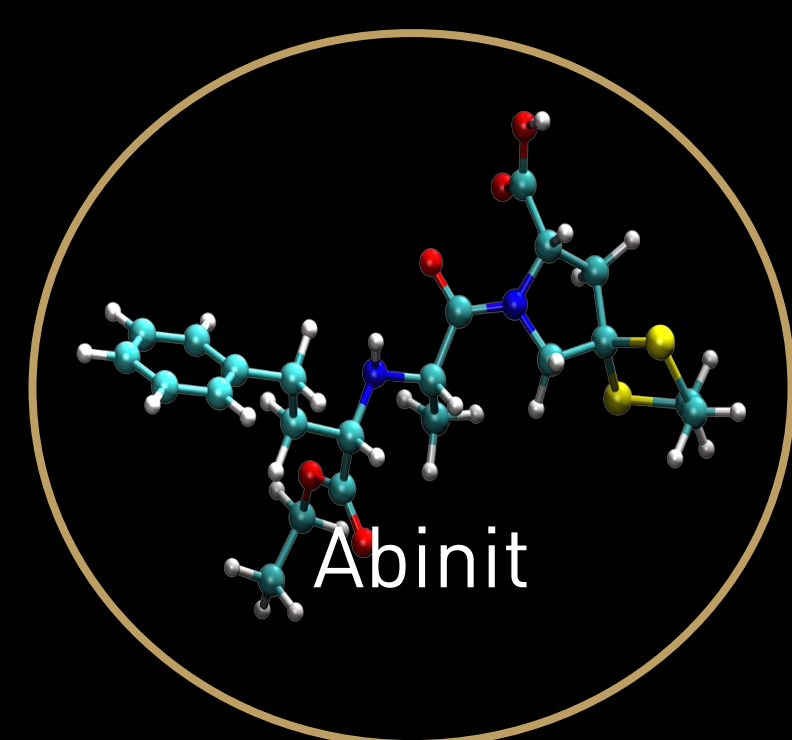
GRACE SUPERCHIP PROVIDES OVER 2X MORE PERF AND EFFICIENCY

HPC applications on ARM-Grace



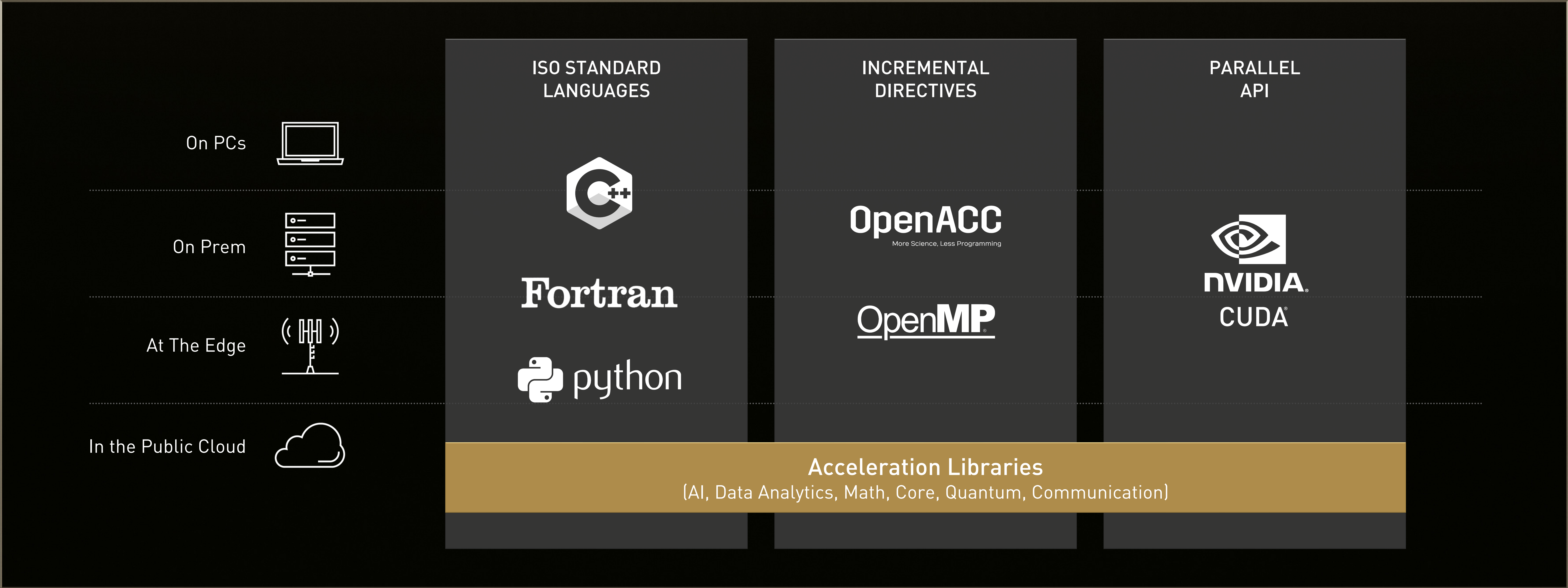
Projected performance subject to change | Application: NCAR WRF standard edition, version 3.9.1.1 | Benchmark: IB4 model (a 4km regional forecast of the Iberian peninsula).
CPU: Dual-socket Intel Ice Lake 8360Y (72 cores) | Grace Superchip (144 cores)

HPC APPLICATIONS RUNNING ON ARM TODAY



PROGRAMMING THE NVIDIA PLATFORM

Develop How You Want, Where You Want





THANK YOU!