III FÓRUM DE INOVAÇÃO

AIOT: AMPLIANDO AS FRONTEIRAS DA INTELIGÊNCIA E CONECTIVIDADE

Realização:

FÓRUM BRASILEIRO DE IOT

MCTI

SiDi 20 anos

Apoio:

nic.br

ABISEMI
ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE SEMICONDUTORES

Qualcomm

III FÓRUM DE INOVAÇÃO

AIOT: AMPLIANDO AS FRONTEIRAS DA INTELIGÊNCIA E CONECTIVIDADE

Plataformas NVIDIA para Desenvolvimento em HPC e IA

Fabio Alves de Oliveira – NVIDIA America Latina

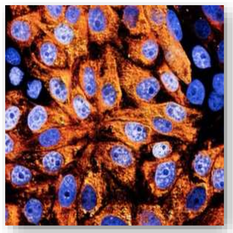
Higher Education and Research

Realização:

Apoio:

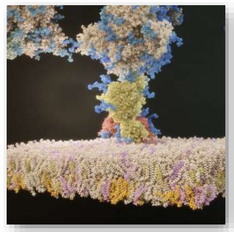
Workloads of the Modern Supercomputer

EDGE

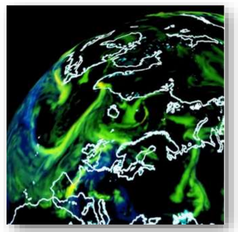


SIM + AI

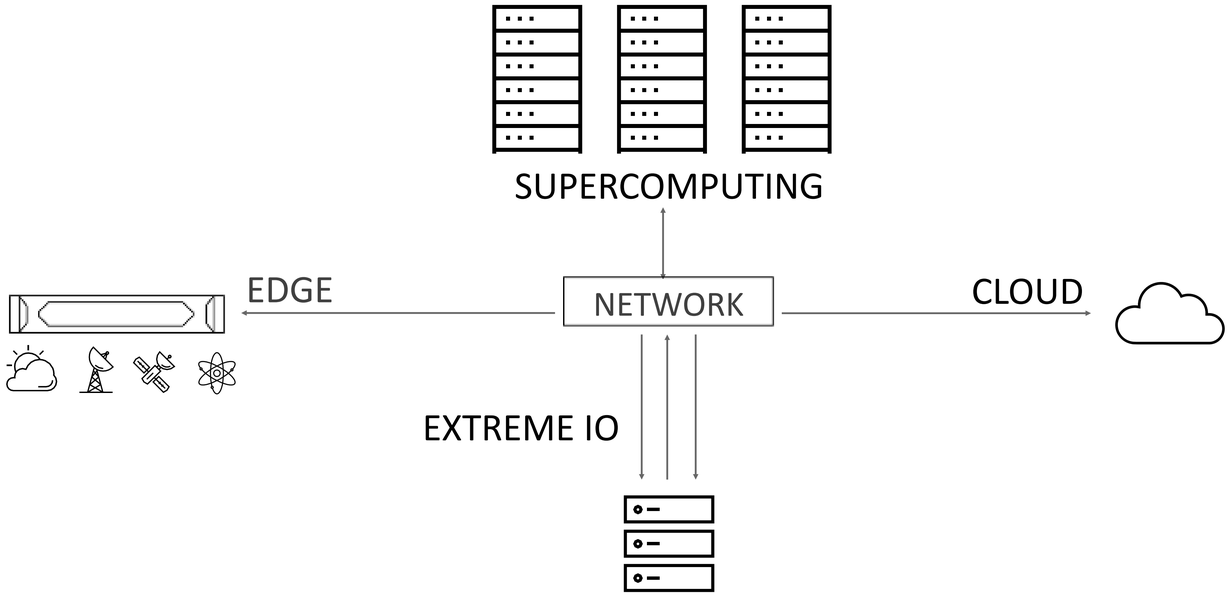
SIMULATION



DIGITAL TWIN



QUANTUM COMPUTING

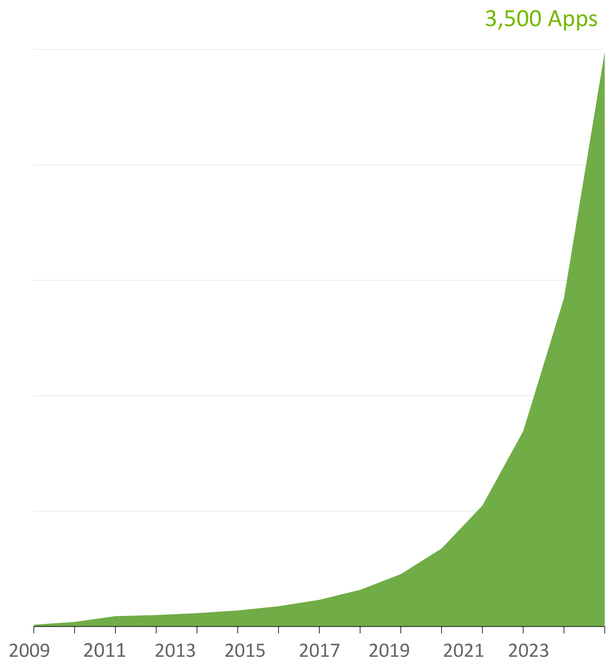


World’s Leading Scientific Computing Platform

Delivering Breakthrough Application Performance

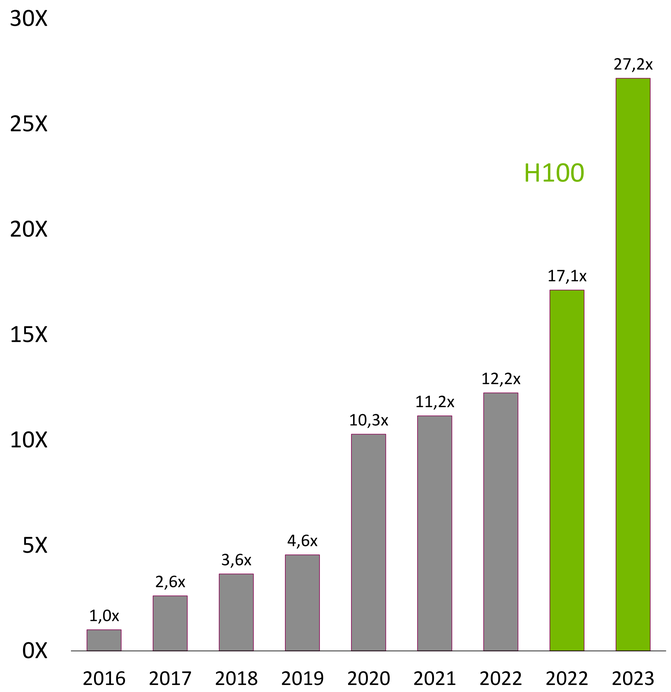
3,500 NVIDIA GPU Accelerated Apps

4M Developers | 17-Year Ecosystem Investment



Simulation Golden Suite Speedup

27X Performance in 8 Years



Powering Worlds Fastest, Greenest Supercomputers

TOP500

The platform of choice

76%
use NVIDIA

1 EF
Added to top10

#1
powered by H100

23
of the top 30

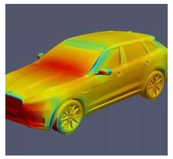
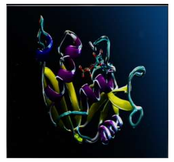
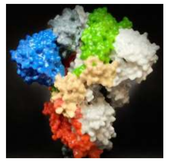
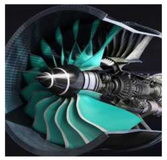
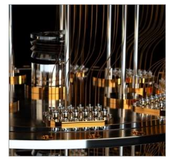
Green500

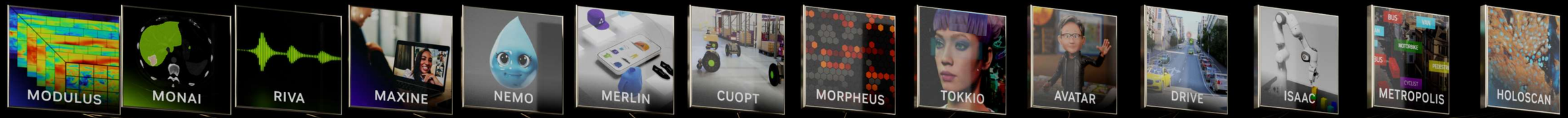
Energy Efficiency

Geometric mean of application speedups relative to 4x P100 GPUs| benchmark applications: Amber [PME-Cellulose_NVE], Chroma [HMC], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT Large Fine Tuner], Quantum Espresso [AUSURF112-jR]

NVIDIA

HPC Reinvented with AI

	Experiments	Simulation	Viz	Edge	HPC + AI	Simulation	Digital Twin	Quantum Computing
								
FEATURE	**PRE-EXASCALE**			**EMERGING POST EXA-SCALE**				
USAGE	BATCH			INTERACTIVE & DISTRIBUTED				
WORKLOAD	SINGLE SIMULATION/ENSEMBLES			SIMULATION/ENSEMBLES, AI TRAINING AND INFERENCE				
EXPERIMENTS	OFFLINE DATA ANALYSIS FOR EXPERIMENTS			MIX OF REAL-TIME ANALYSIS, STEERING AND OFFLINE				
DIGITAL TWINS	IN-SITU VISUALIZATION			INTERACTIVE COMBINATION OF SIMULATION AND OBSERVATIONAL DATA				
QUANTUM COMPUTING	SIMULATION			PREPARING FOR A HYBRID MODEL				
PROGRAMMING MODELS	FORTRAN, C++, MPI, OPENMP			STANDARD PARALLELISM SUPPORT IN FORTRAN, C++, MPI, OPENMP, OPENACC, PYTHON, JULIA, PYTORCH, JAX, TENSORFLOW				
CLOUD	GRID			BURST CAPABILITIES, FASTER REFRESH CYCLE, ACCESS TO LATEST TECHNOLOGY AT SCALE				



AI APPLICATION FRAMEWORK

PLATFORMS

NVIDIA HPC

NVIDIA AI

NVIDIA Omniverse

ACCELERATION LIBRARIES

cuNumeric | CV-CUDA | cuQuantum | Parabricks | Sionna | JetPack

RAPIDS | Spark | cuDNN | cuGraph | TensorRT | Triton | Deepstream | Flare

DOCA | Mag IO | Aerial

CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS

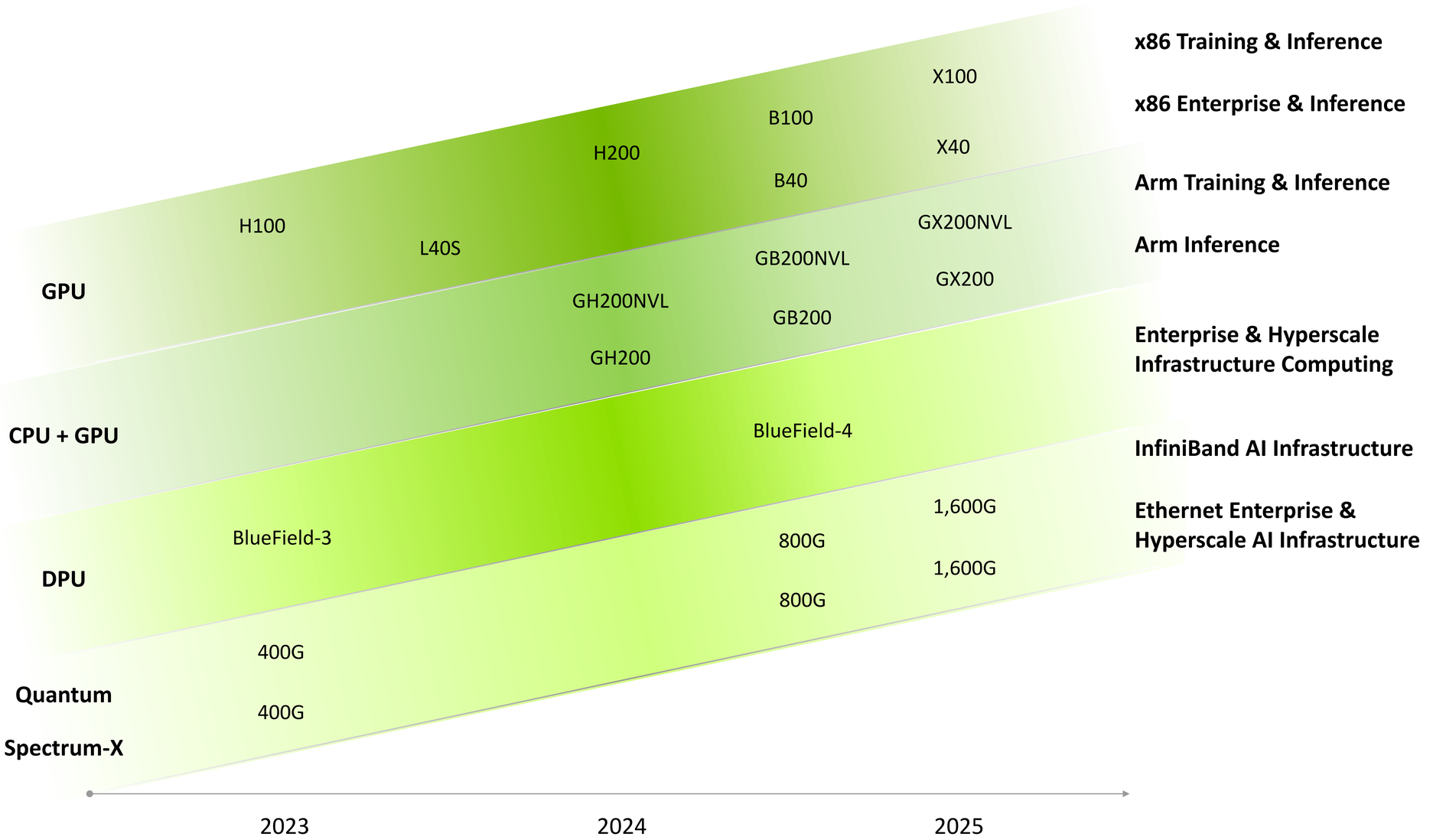
RTX | DGX | HGX | EGX | OVX | Super POD | AGX

3-CHIPS

GPU | CPU | DPU

NVIDIA AI - One Architecture | Train and Deploy Everywhere

One–Year Rhythm



Grace Hopper

NVIDIA GH200 Grace Hopper Superchip

Built for the New Era of AI Supercomputing

CPU to GPU Bandwidth

900GB/s

NVLink-C2C

GPU Memory Bandwidth

5TB/s

HBM3e

Energy Efficiency

50X

MILC Efficiency vs 2S x86 CPUs

QFT Quantum Simulation

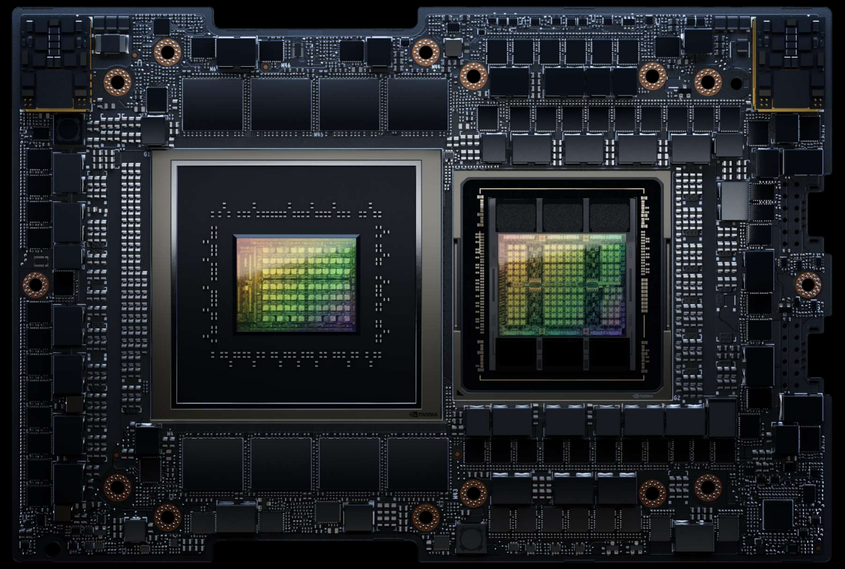
90X

Performance vs 2S x86 CPUs

LLM Inference

200X

Performance vs H100 80GB

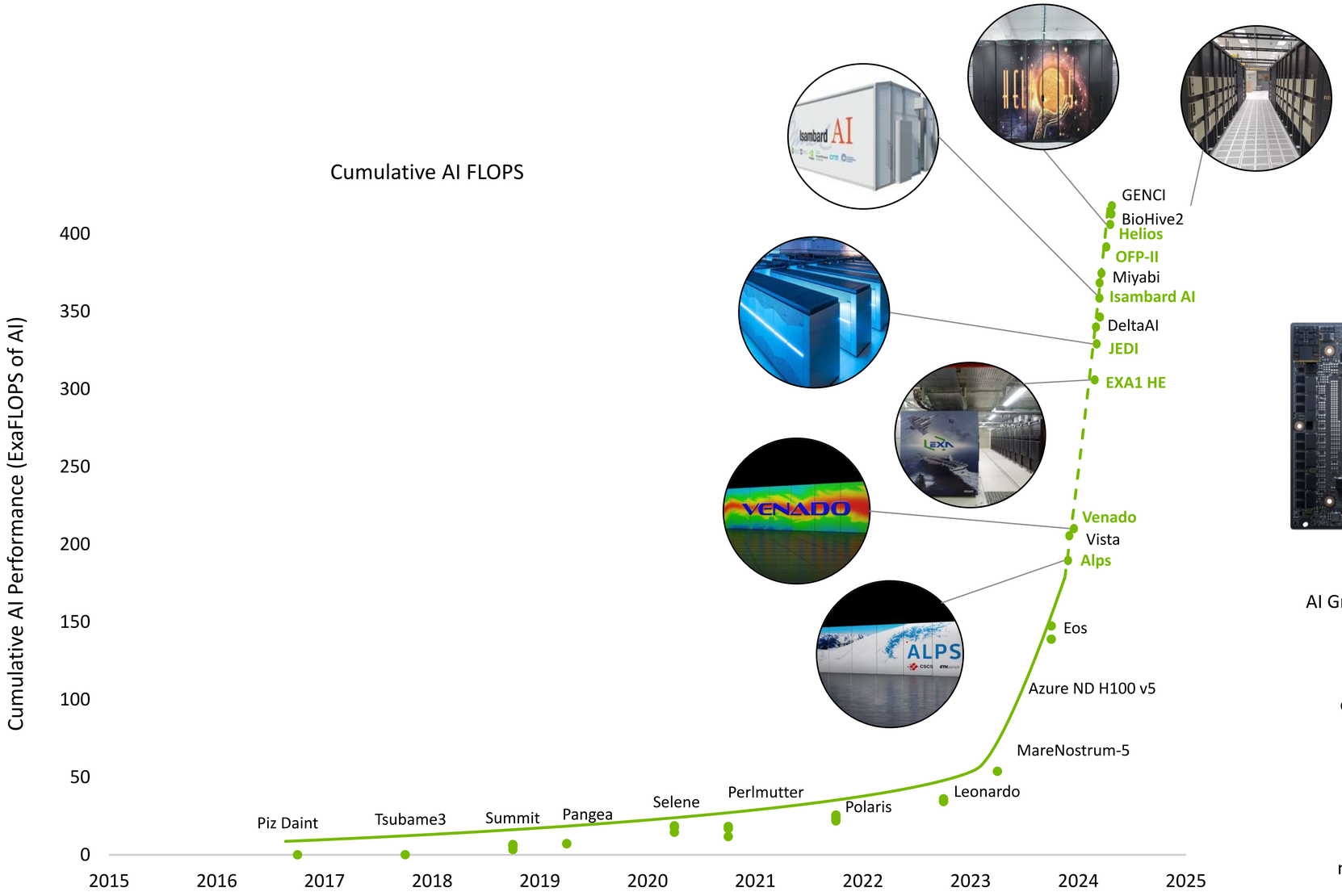


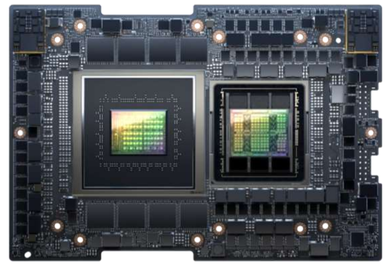
624GB High-Speed Memory | 4 PF AI Perf | 72 Arm Cores

Preliminary measured performance, subject to change
Energy Efficiency: GH200 144GB vs 2S Xeon 8480+ CPU for MILC running dataset: Apex Medium
QFT Quantum Simulation: QFT 2S Xeon 8480+ vs GH200 144GB
LLM Inference: Llama.cpp (2S 8480+) and TensorRT-LLM (GH200, H100) | Max Batch Llama2 70B | Throughput includes time to first token + token generation time

Grace Hopper Powers AI Supercomputing Datacenters

Grace Hopper Will Deliver 200 Exaflops of AI performance for Groundbreaking Research





200 ExaFLOPS
AI Grace Hopper coming online 2024

80%
of Hopper are Grace Hopper

2X
More energy efficient

7
new GH200 supercomputers

Grace CPU

NVIDIA Grace CPU Superchip

Breakthrough Performance and Efficiency for the Modern Data Center

CPU + Memory Power
500W
Grace CPU Superchip TDP

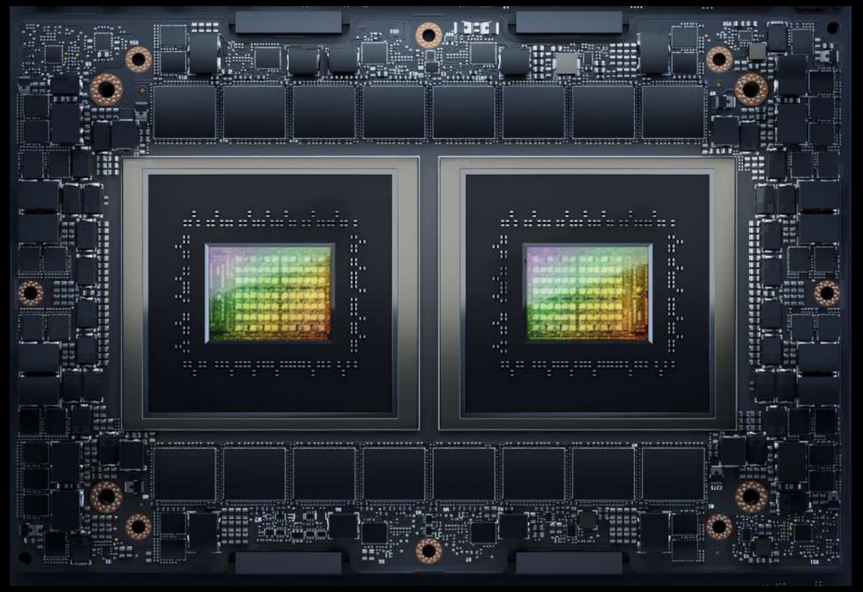
Memory Bandwidth
1 TB/s
LPDDR5X

Green 500
7 Grace systems
Performance, Energy Efficiency with GH200

Energy Efficiency
2X
Performance vs x86 CPU

Weather
1.3X
Performance vs x86 CPU

Graph Analytics
2X
Performance vs x86 CPUs

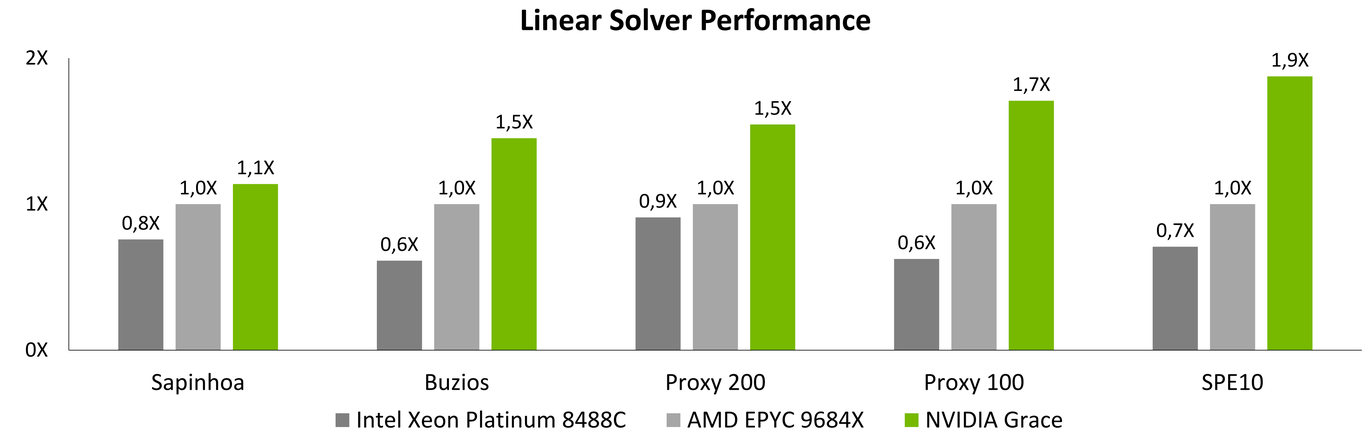


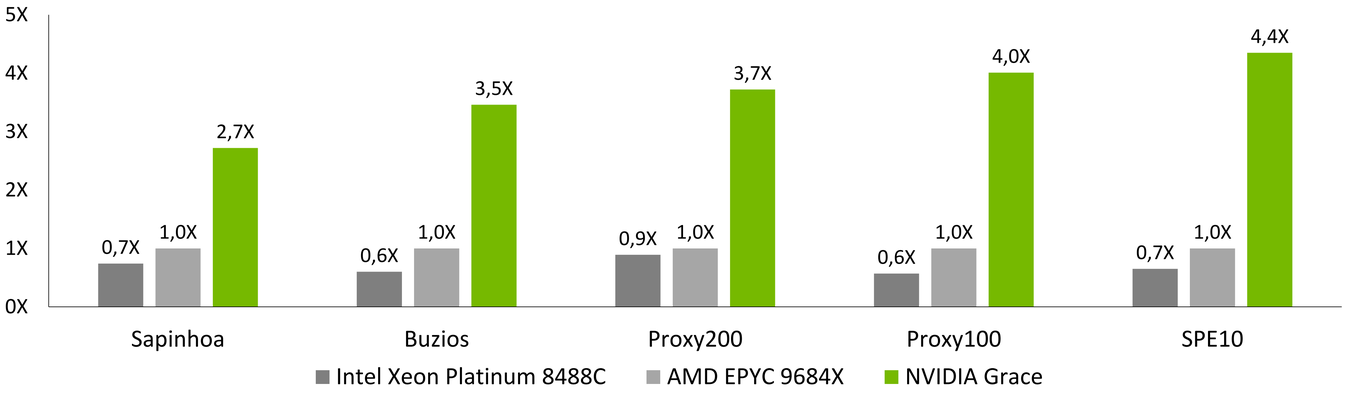
144 Arm Neoverse V2 Cores | 234MB L3 Cache
3.2 TB/s NVIDIA Scalable Coherency Fabric | 960GB LPDDR5X

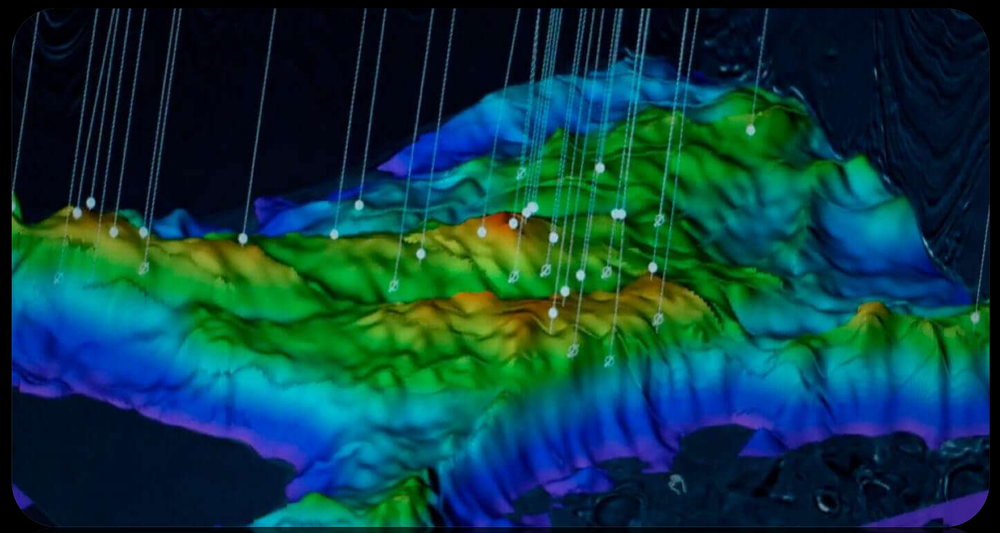
Preliminary measured performance, subject to change
Energy Efficiency: Grace CPU Superchip vs 2S AMD EPYC 9654 and Xeon Platinum 8480+. Geomean of OpenFOAM (Motorbike Large), WFR (CONUS12km), ICON (QUBICC 80 km resolution) specfm3d (four_material_simple_model) and Branson (3D_hohlraum_single_node)
Weather: WRF (CONUS12km) Grace CPU Superchip vs 2S AMD EPYC 9654
Graph Analytics: GAP BS Breadth First Search

Petrobas Reservoir Simulation

NVIDIA Grace CPU delivers over 4X energy efficiency





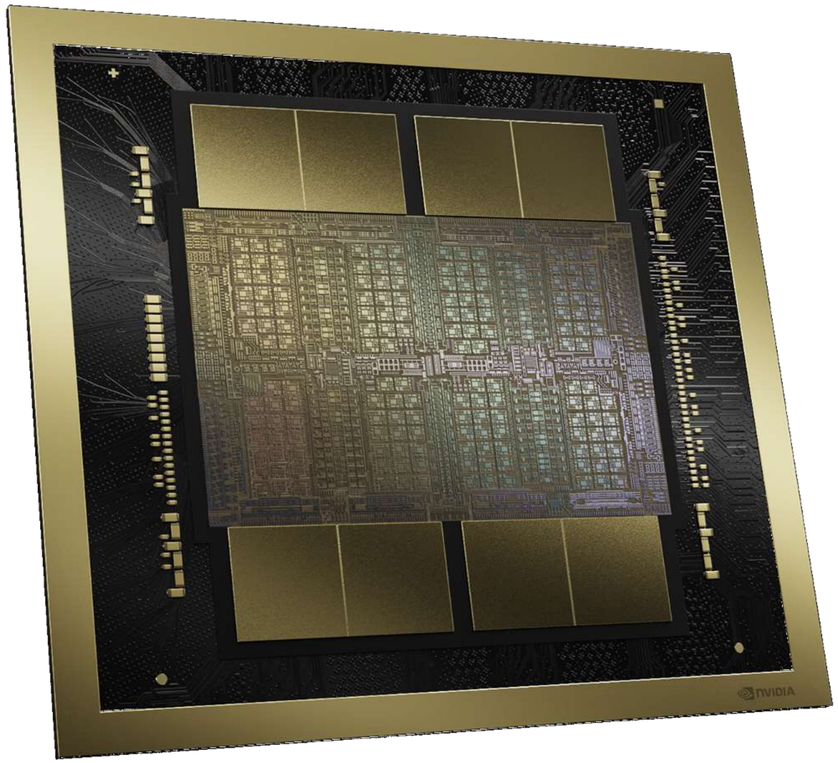


NVIDIA Grace CPU compared with Intel Xeon Platinum 8488C (48 Core) on AWS EC2 R7i and AMD EPYC 9684X (96C) on AWS EC2 R7a
Power based on TDP + memory power
SPE10 (10th SPE Comparative Solution Project) | Búzios (Biggest ultra deep-water oil field of the world) | Sapinhoá (PreSaltreservoir of Santos Basin)
Proxy 100 (Semi Synthetic Model with 100x100m cells; 6,245,051 active cells) | Proxy 200 (Semi Synthetic Model with 200x200m cells; 765,620 active cells

Blackwell

NVIDIA Blackwell

The Engine of the New Industrial Revolution



Built to Democratize Trillion-Parameter AI

20 PetaFLOPS of AI performance on a single GPU

4X Training | 30X Inference | 25X Energy Efficiency & TCO

Expanding AI Datacenter Scale to beyond100K GPUs

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

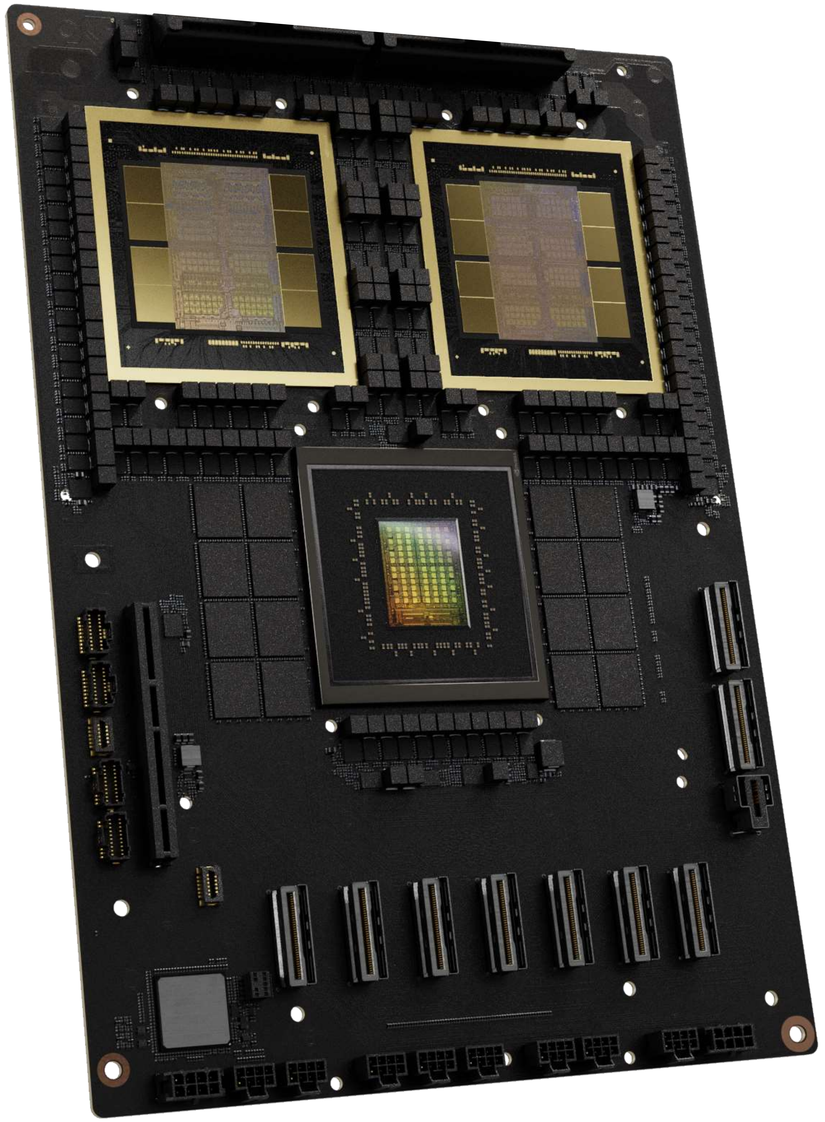
RAS ENGINE
100% In-System
Self-Test

SECURE AI
Full Performance
Encryption & TEE

DECOMPRESSION ENGINE
800 GB/s

GB200 SUPERCHIP

Optimized for Supercomputer-Scale Science



72 Grace CPU Arm cores

40 PetaFLOPS FP4 AI Inference

20 PetaFLOPS FP8 AI Training

16 TB/s of GPU memory bandwidth

864 GB Fast Memory

DGX GB200

Delivers New Unit of Compute

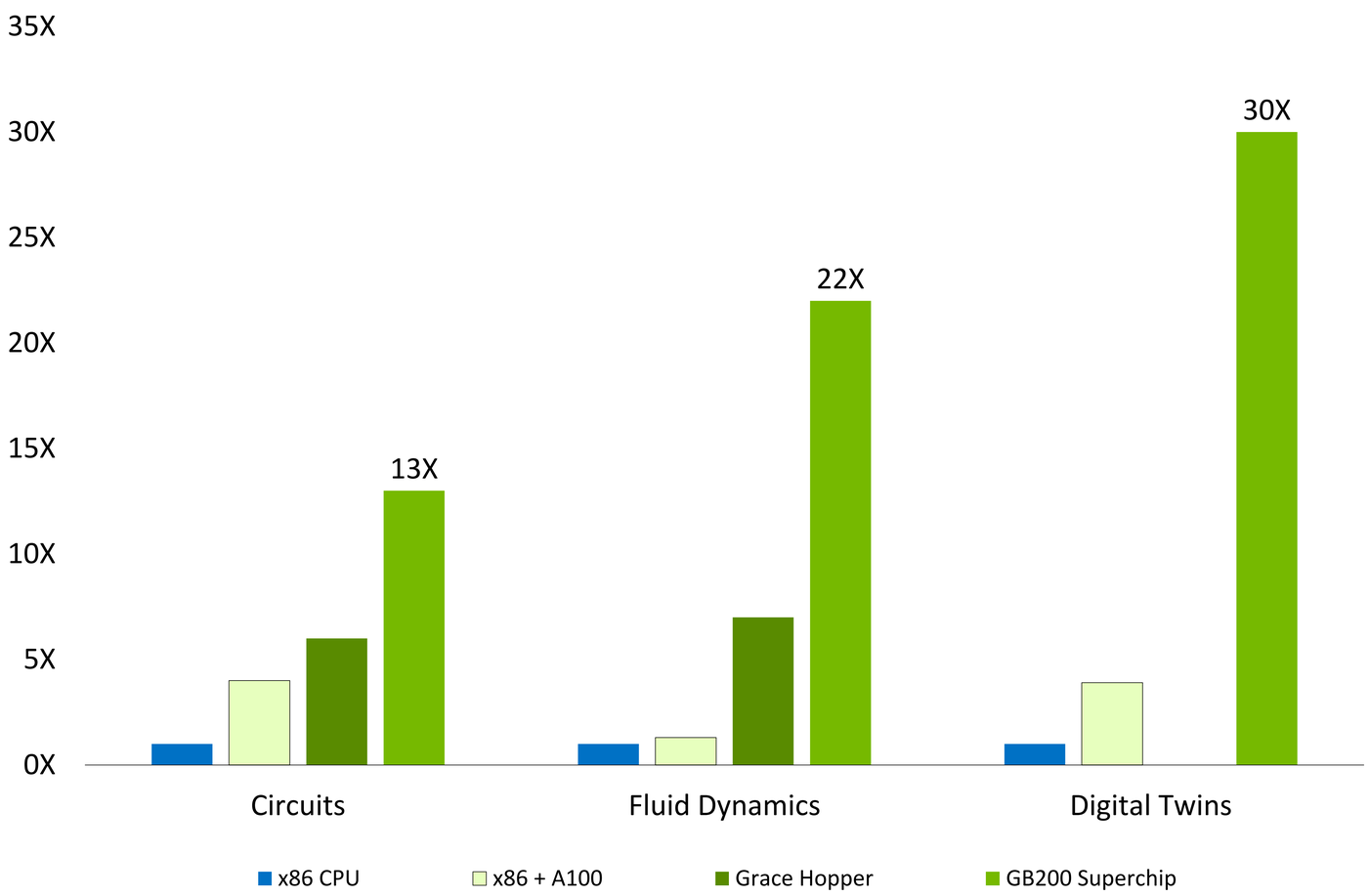
DGX GB200

- 36 GRACE CPUs
- 72 BLACKWELL GPUs
- Fully Connected NVLink Switch Rack

Training	720 PFLOPs
Inference	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node All-to-All	130 TB/s
Multi-Node All-Reduce	260 TB/s

Blackwell Pushes Boundaries of Engineering

Supercharging Double Precision Performance for Product Design Simulations



Projected performance subject to change. CPU, A100, GH200 performance measured. 2x x86, 1x A100 GPU, 1x H100 GPU, and Grace Blackwell GB200 Superchip. SpectreX (Circuits) dataset: Design 2 TSMC N5, Fidelity (Fluid Dynamics) dataset: Gear pump 2M , Reality (Digital Twins) dataset: Case #1 18.6M cells

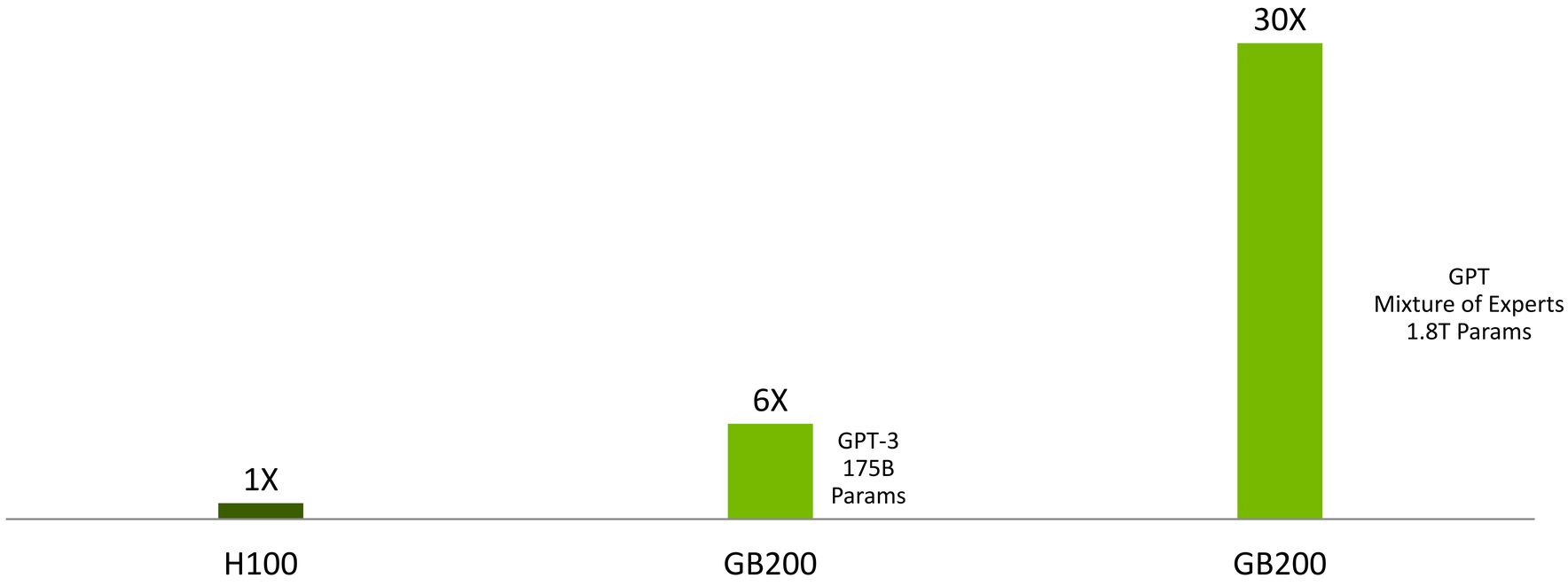
Blackwell — Driving the Era of Generative AI

30X realtime inference, 25X improved energy efficiency

30X
Higher Throughput

25X
Energy Efficiency

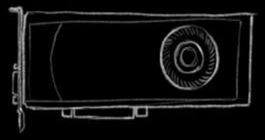
25X
Lower TCO



Projected performance subject to change
Token-to-token latency (TTL) = 50 milliseconds (ms) real time
GPT-3 175B: First token latency (FTL) 2s; input sequence length = 2,048, output sequence length = 128, 4 HGX H100 air-cooled 400GB IB Network vs 2 GB200 Superchips liquid-cooled NVLink; per GPU performance comparison,
GPT-MoE-1.8T: FTL = 5s; input sequence length = 32,768, output sequence length = 1,024 , 8 HGX H100 air-cooled 400GB IB Network vs 18 GB200 Superchips liquid-cooled NVL36; per GPU performance comparison

Sovererign AI

2012
ALEXNET
"FIRST CONTACT"



"CAT"

TEXT

TEXT

IMAGE

VIDEO

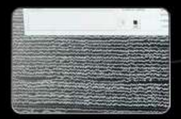
TEXT

SPEECH

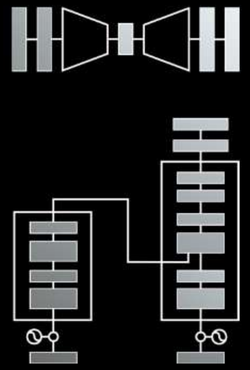
MULTI-MODAL

AMINO ACID

BRAINWAVES



LEARN AND UNDERSTAND EVERYTHING



IMAGE

VIDEO

IMAGE

3D

SOUND

ANIMATION

MANIPULATION

PROTEIN

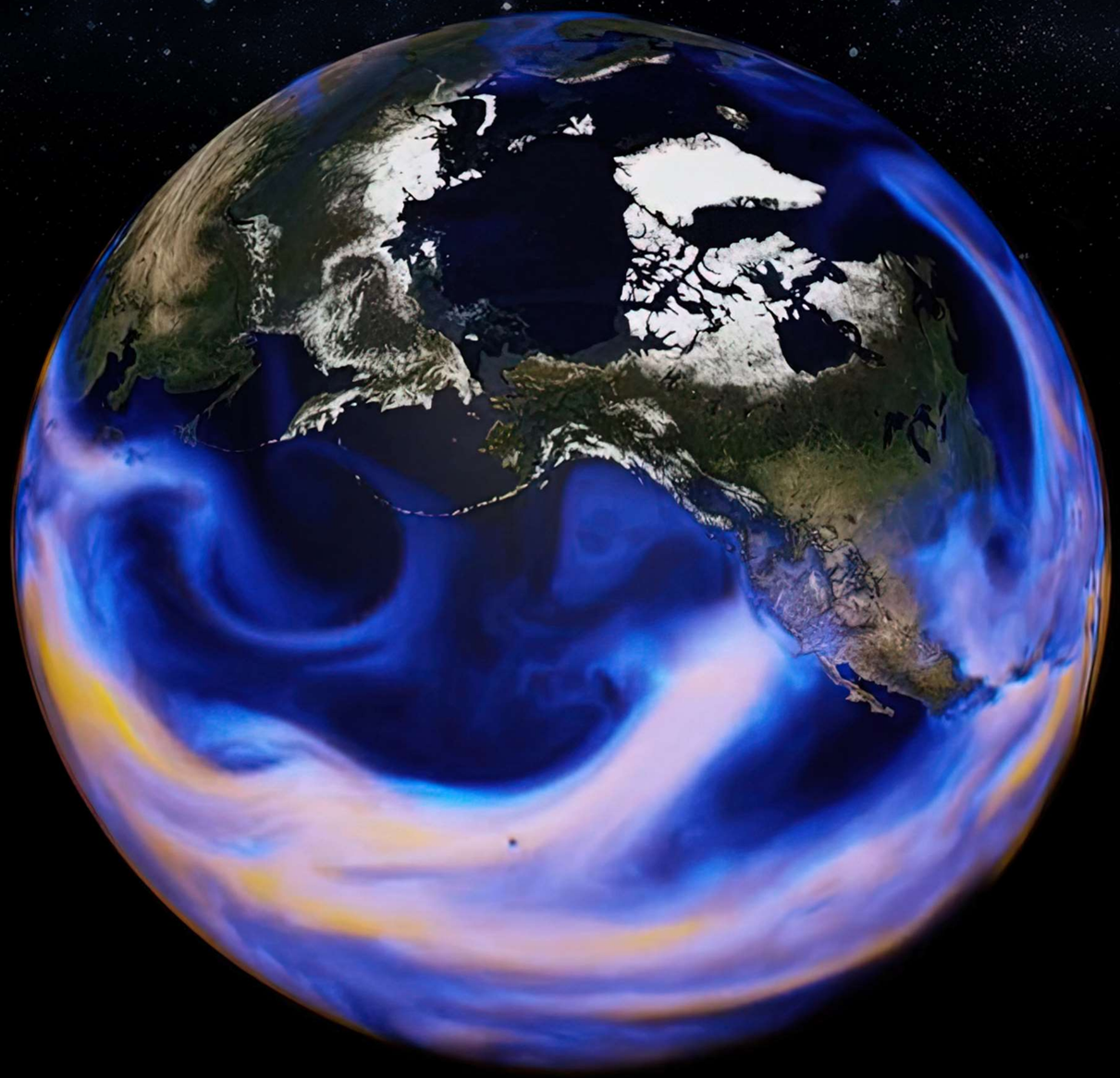
SPEECH

Understand Planet Earth for building climate resilience

Digital Twins for Solving the world’s biggest challenges, including climate change

AI for Disaster risk monitoring

AI for Modeling & Simulation



Accelerating AI & HPC to Transform the Public Sector

Sovereign Foundation Models

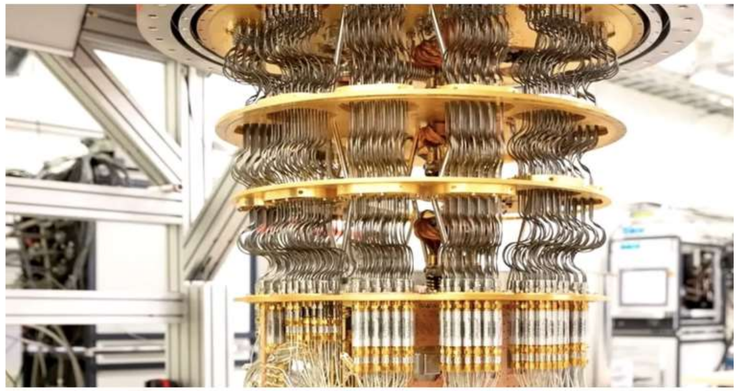
Radar & Signal Processing

Geospatial Intelligence

Digital Twins & Simulation

Data Analytics

Quantum Computing



Generative AI Accelerates Policy Imperatives

Protect National Security

Anomaly Detection | Counter-Disinformation
Training and Simulation | Autonomous Systems and Robotics

Personalized Citizen Services

Virtual Assistance for Social Services | Sentiment Analysis
Language Translation Services | Citizen GPT

Regulatory Compliance and Enforcement

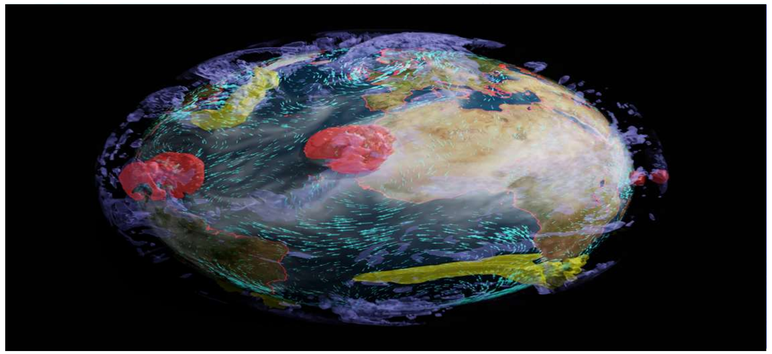
Automated Compliance Monitoring (i.e.., safety, environmental,
Transportation, and vehicle emissions)

Education and Skills Development

Adaptive personalized learning materials | Student Analytics
Virtual Teaching Assistants | Education Policy Analysis

Urban Planning and Smart Cities

Urban Design and Simulation | Public Space Optimization



Climate Action and Policy

Environment GPT – Scenario Generation |
Climate Policy Assessment and Adaptation Planning

Building an AI Nation
Digital Transformation in the Age of AI

Economic Growth | Scientific Discovery | Urban Renewal | National Security | Citizen Services

INFRASTRUCTURE FOR DOMESTIC AI PRODUCTION + CONSUMPTION

AI STRATEGY + POLICY+ ECOSYSTEM + INFRASTRUCTURE

PUBLIC BENEFIT / OUTCOMES

Competitive industries Home-grown innovation Robust startup community	Vaccine development New materials Climate prediction	Efficient public transit Resilient housing Increased livability	Secure borders Early threat detection Energy and food security	Digitised public services Up-skilled workforce Public data exchanges

AI FACTORY IS THE NEW CRITICAL INFRASTRUCTURE

INSTRUMENT FOR SCIENTIFIC DISCOVERY

- Human condition
- Fundamental research
- Industrial innovation

ENGINE FOR ECONOMIC GROWTH

- Start-up ecosystem
- New jobs/sectors
- Retain talent; Workforce productivity

PLATFORM FOR PUBLIC SECTOR INNOVATION

- Improve access to services
- Expand citizen services
- Defend national interests

AI Factories

Manufacturing “intelligence”

Startups – Enterprises - Nations

AI Factory

Accelerated compute + network + storage designed for AI

Access to AI platform, tools and expertise

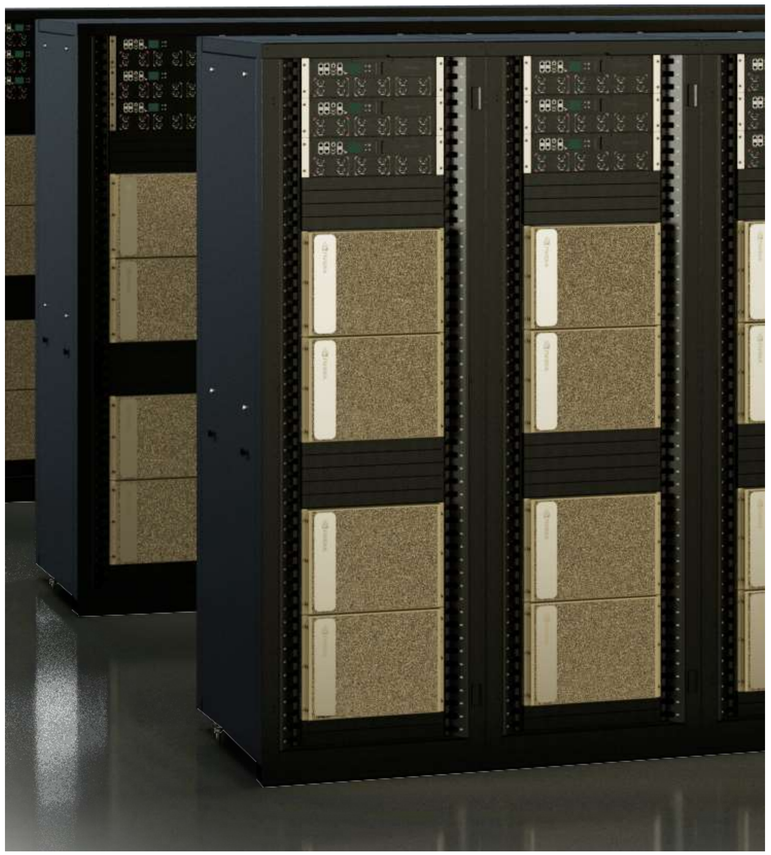
Sovereign AI

New AI-based Supercomputers to Foster Scientific Innovation

Isambard-AI | Isambard-3

United Kingdom

5,280 GH200 GPUs | Grace CPUs with over 55,000 Arm Neoverse V2 Cores



Gefion

Denmark

1,528 H100 GPUs | NVIDIA Quantum-2 InfiniBand | NVIDIA CUDA Quantum

Jean Zay

France

1,456 NVIDIA H100 GPUs | Liquid Cooled | 400Gb/s NVIDIA Quantum-2 InfiniBand

Plano Brasileiro de Inteligência Artificial (PBIA)

IA para o Bem de Todos

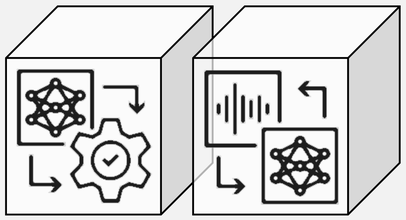
Generative AI for Science

NVIDIA LLM Portfolio

Multiple Products to Help Customers Build Custom LLMs

NeMo Framework

Also available on Git Hub and NGC



GPU Accelerated tools for data preparation, LLM model training, customization and deployment.

NeMo Pretrained Models

Available from NGC and Hugging Face

Community Models

Compatible with NeMo framework

Falcon LLM Llama 2 MPT

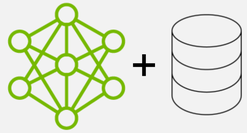
NVIDIA NIM

Microservices for accelerated Inference of AI Models

NeMo Extensions

Early development techniques for researchers/developers

Retriever



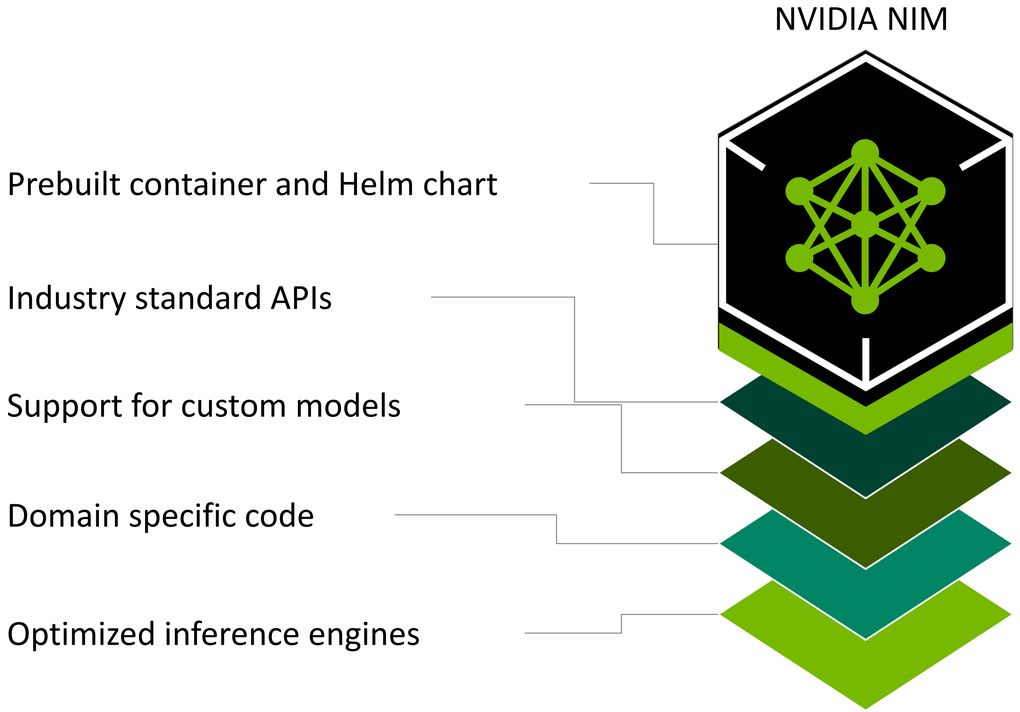
NeMo Guardrails

NeMo SteerLM

https://www.nvidia.com/en-us/ai/

NVIDIA NIM: Inference Microservices for Generative AI

Accelerated runtime for generative AI



Deploy anywhere and maintain control of generative AI applications and data

Simplified development of AI application that can run in enterprise environments

Day 0 support for all generative AI models providing choice across the ecosystem

Improved TCO with best latency and throughput running on accelerated infrastructure

Best accuracy for enterprise by enabling tuning with proprietary data sources

Enterprise software with feature branches, validation and support

NVIDIA NIM for Every Domain

LANGUAGE NIMs


Code Llama 70B


Cohere 35B


Gemma 7B


Jamba


Llama 2 70B


Mistral 7B


Mixtral 8x7B


Nemotron-3 22B Persona


Phi-2

VISUAL / MULTIMODAL NIMs


Adept 110B


Deplot


Edify. Getty


Edify. Shutterstock


FuYu 8B, 55B


Kosmos-2


NeVA


SDXL 1.0


SDXL Turbo

DIGITAL HUMAN NIMs


Audio2Face


Riva ASR

OPTIMIZATION / SIMULATION NIMs


cuOpt


Earth-2

APPLICATION NIMs


Llama Guard


Retrieval Embedding


Retrieval Reranking

DIGITAL BIOLOGY NIMs


DeepVariant


DiffDock


ESMFold


MolMIM

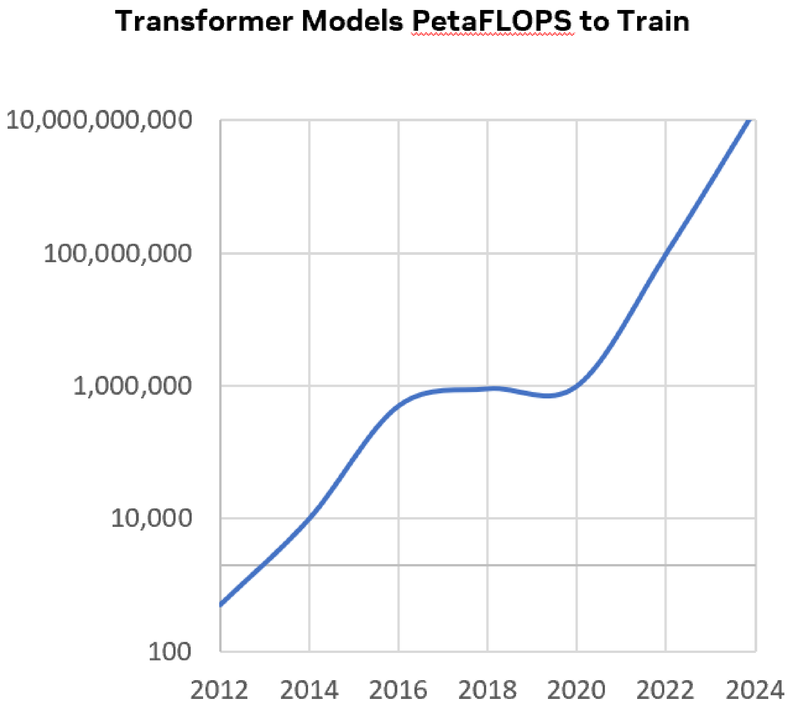

Vista 3D

Accelerated Computing is Sustainable Computing

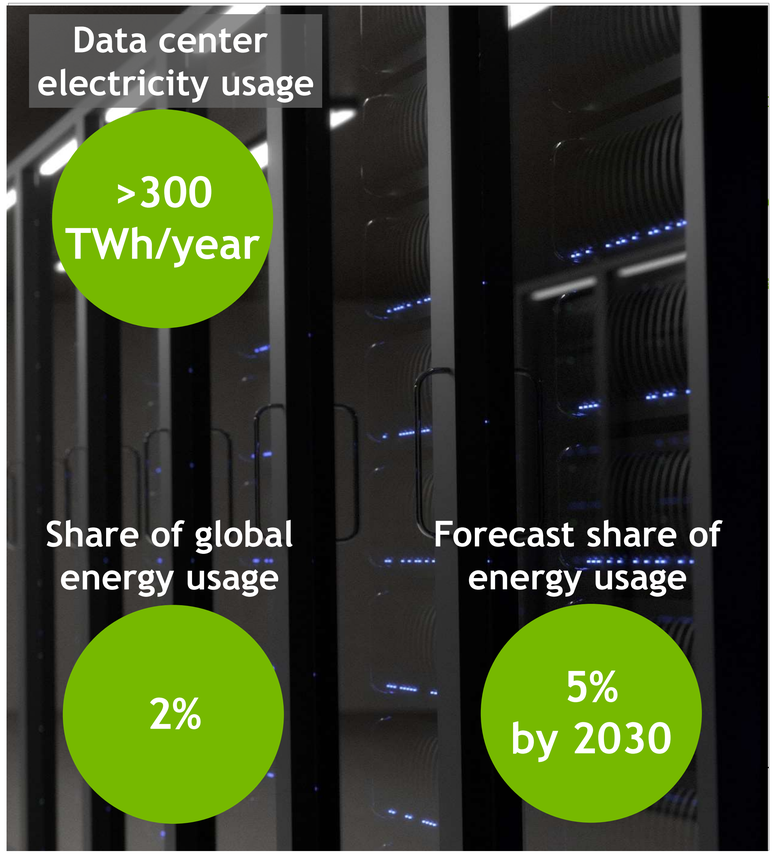
Exploding Energy Usage In Data Centers

Massive AI Models Deliver New Use Cases

Transformer networks power Conversational AI and Metaverse



Model Sizes Demanding More Compute



Data Center Need to Become More Efficient

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

Accelerated Computing Is Sustainable Computing

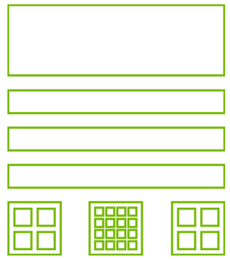
Full Stack, 3 Chips, Data Center Scale

48 Million CUDA Downloads

4.5 Million Developers

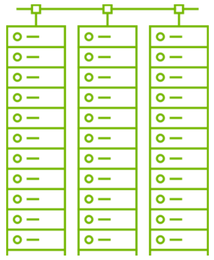
3,000 Applications

600 SDKs & AI Models

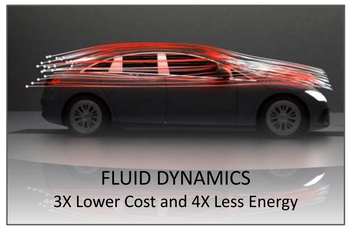


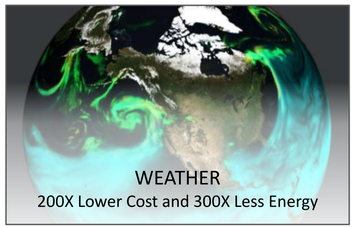
Accelerated Computing

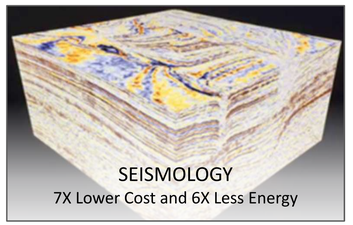
Generative AI

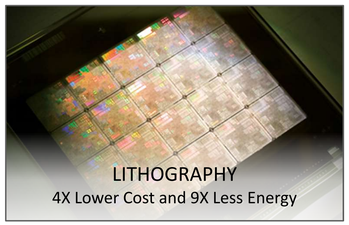


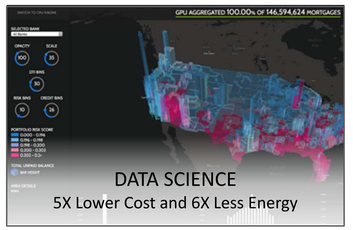
Data Center Scale









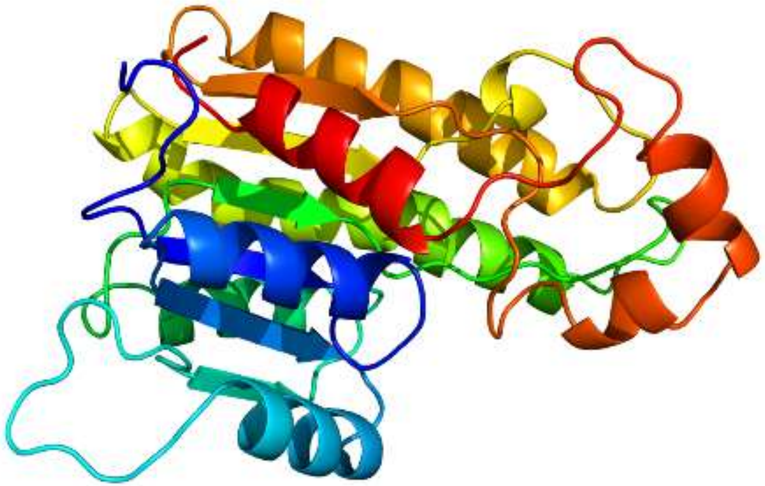


HPC Powered AI Applications Accelerated by NVIDIA Platform

23X Improvement in Energy Efficiency of LLMs Adapted for Scientific Use Cases

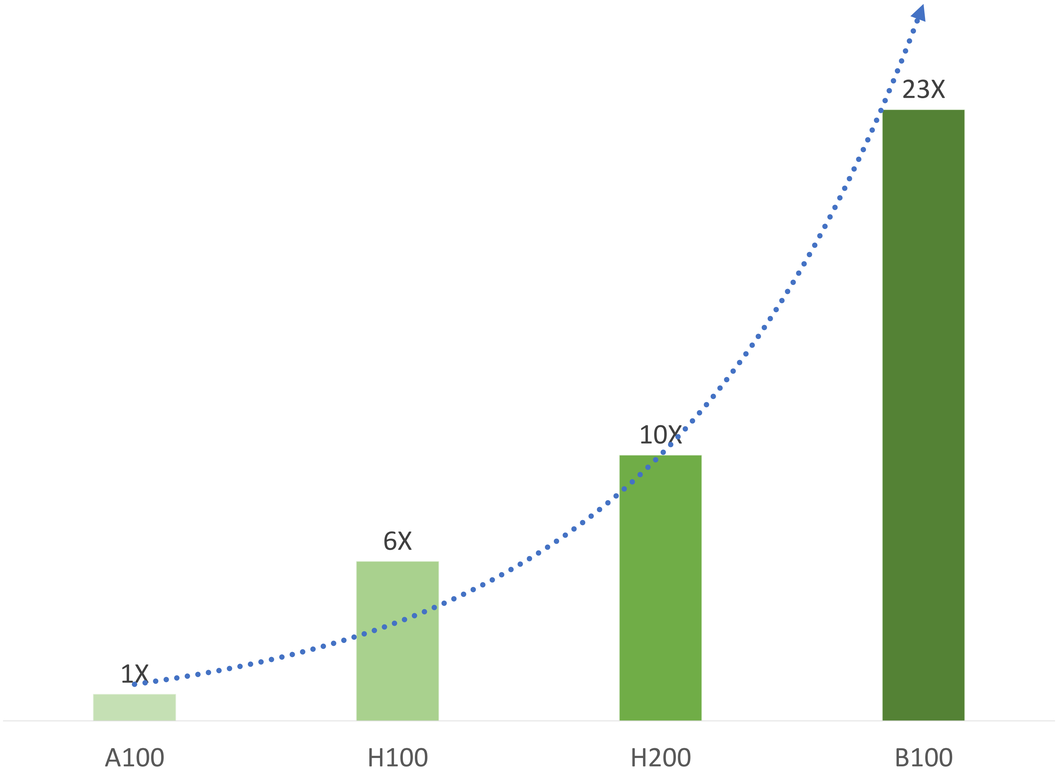
Drug Discovery and Genomics

Transformer based networks like AlphaFold and MegaMolBART enable protein sequencing for drug discovery



AI outperforms lab discovery techniques at
Cost, speed, permutations

GPT-3 175B LLM Inference Energy Efficiency Performance
Higher is Better



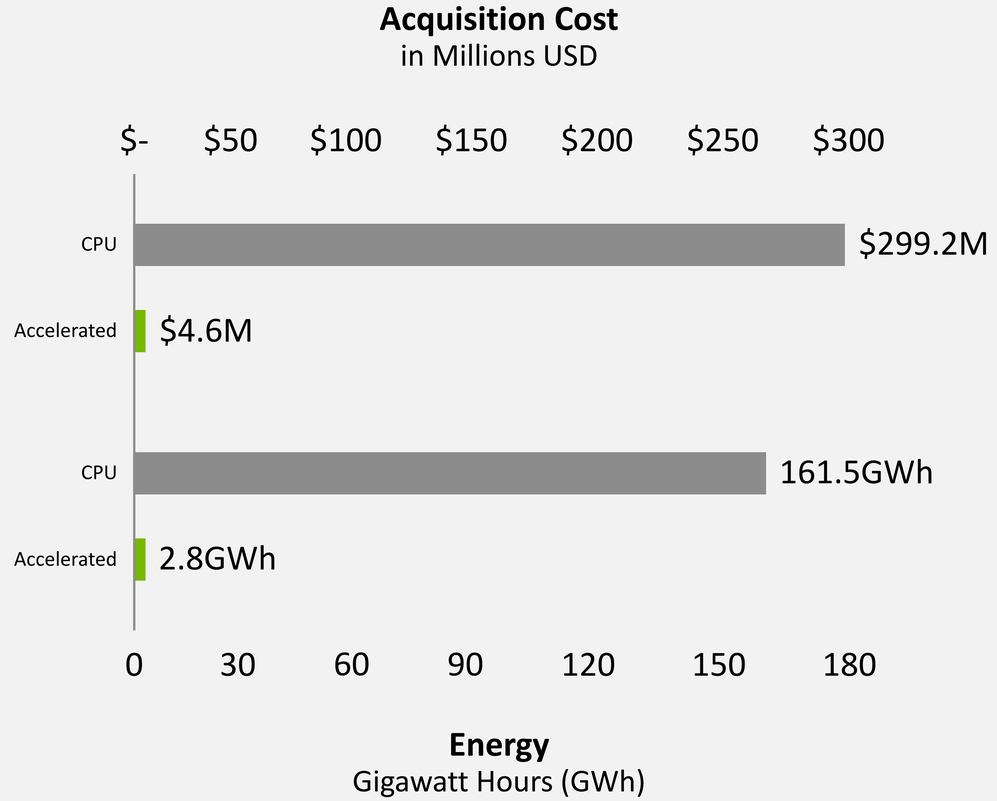
Digital Twins With Physics-ML

AI-accelerated, physics-based, high-fidelity surrogates for interactive digital twins

"By leveraging NVIDIA Modulus, a GPU-based physics-ML framework, we want to accelerate complex, multi-physics simulations with AI-powered surrogates. The reduced computational time enables us to develop energy-efficient digital twins for a sustainable, reliable, and affordable energy ecosystem."

— Georg Rollmann, Head of Advanced Analytics and AI, Siemens Energy

65X Lower Cost and 58X Less Energy

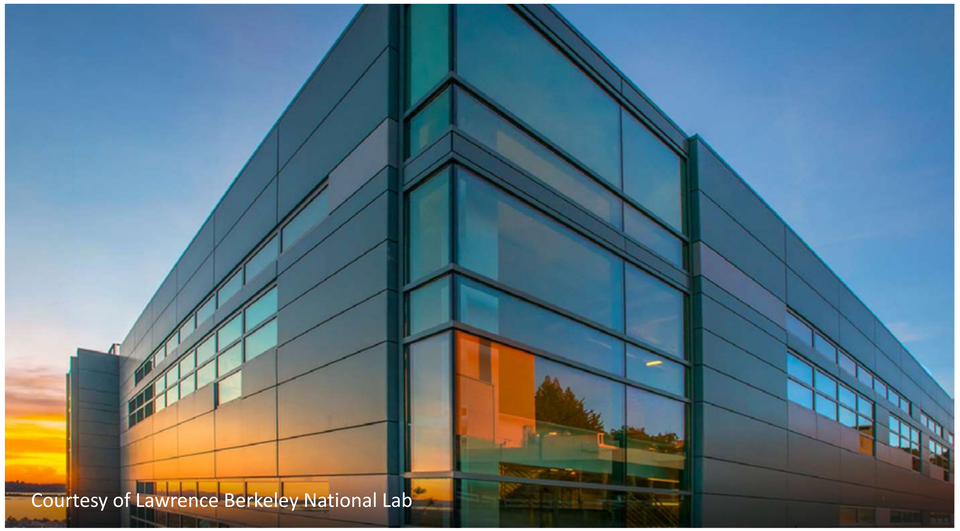


Based on measured performance of Modulus versus open-source CFD solvers like OpenFOAM. CPU: Intel Ice Lake 8380. GPU A100 80GB Tensor Core.

Industrial, Climate, & Life Sciences Domains	200 Billions	200X	$296 Million & 160 GWh
Used in	CPU Hours Consumed Annually	Performance Speedup	Saved Annually

Driving Scientific Discovery

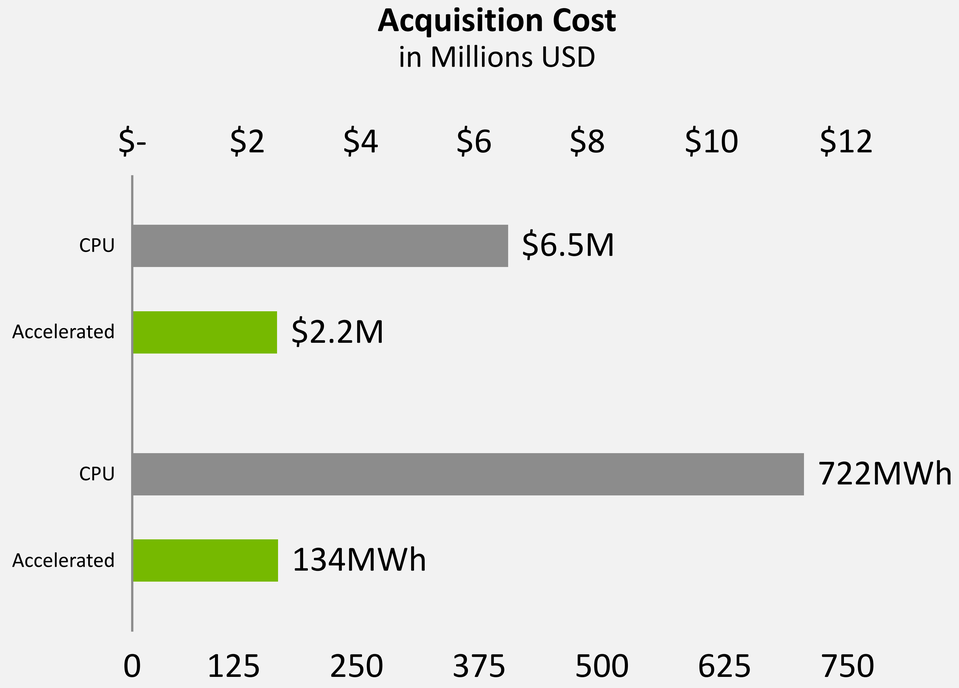
Application efficiency and performance for every supercomputing center



"Researchers need energy-efficient, performant ways for their simulation and AI applications to support their work. At NERSC, we've adopted accelerated computing to support research across a broad variety of science applications."

— Nick Wright, Chief Architect, NERSC

3X Lower Cost and 5X Less Energy



Based on the geometric mean of application speedups and energy consumption vs. dual AMD 7763 / benchmark applications / BerkleyGW / DeepCAM / EXAALT / MILC and monthly prices on Microsoft Azure for instances standard_NC96ADS_A100_v4 and standard_D96ads_v5

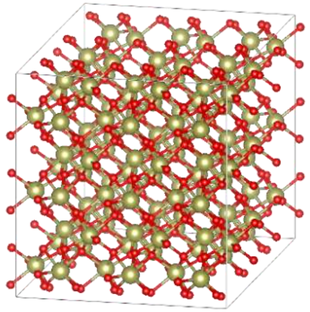
Thousands	16 Billions	12X	$4.3 Million & 588 MWh
Of High-performance Computing and AI Apps	CPU Hours Consumed Annually	Performance Speedup	(MWh) Saved Monthly

Magnum IO Multi-Node Scaling Performance

Magnum IO NVIDIA Common Collectives Library Delivers up to 2.3X Energy Efficiency

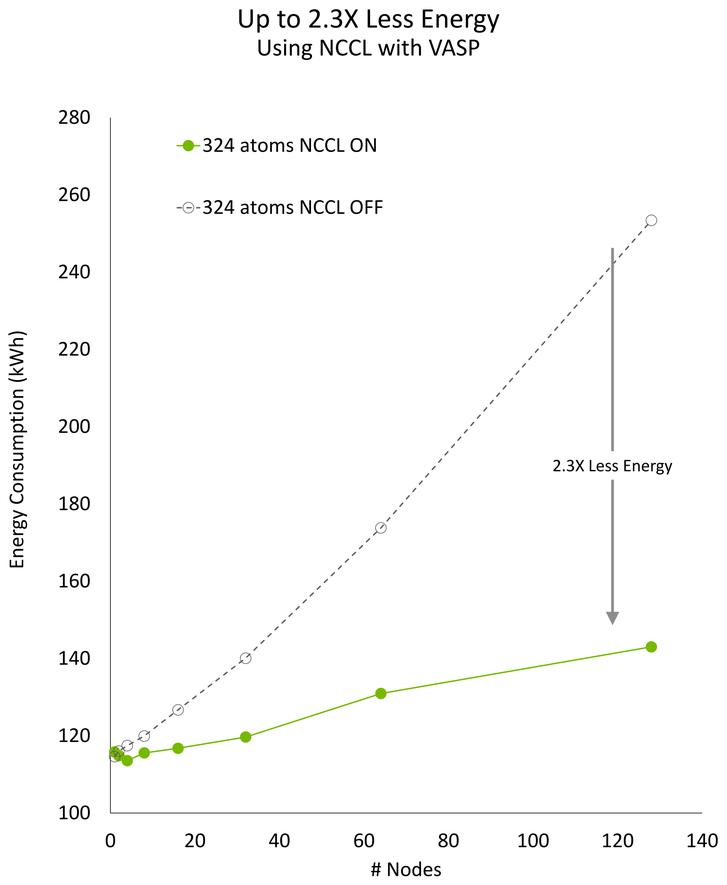
Vienna **A**b initio **S**imulation **P**ackage

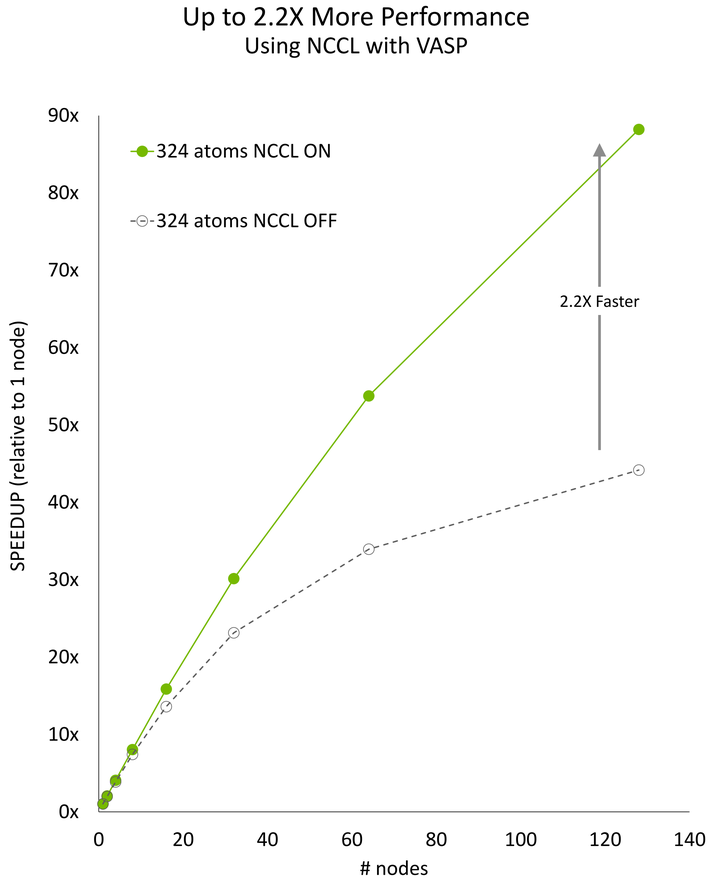
Performs quantum mechanical calculations for systems of atoms to predict properties



Challenge: Computation grows as n^3 with the number of atoms

Solution: Multi-node accelerated computing





Energy consumption and performance measured using VASP 6.4.0 on Selene with 4 A100 80 GB GPUs per node and 4 CX6 Infiniband per node using a Hybrid DFT analysis of Hafnia (HfO2). NCCL OFF runs used openMPI

NVIDIA Programs

Learn. Innovate. Succeed.

Grow and validate your skills with training from the NVIDIA Deep Learning Institute.

Developers

http://developer.nvidia.com

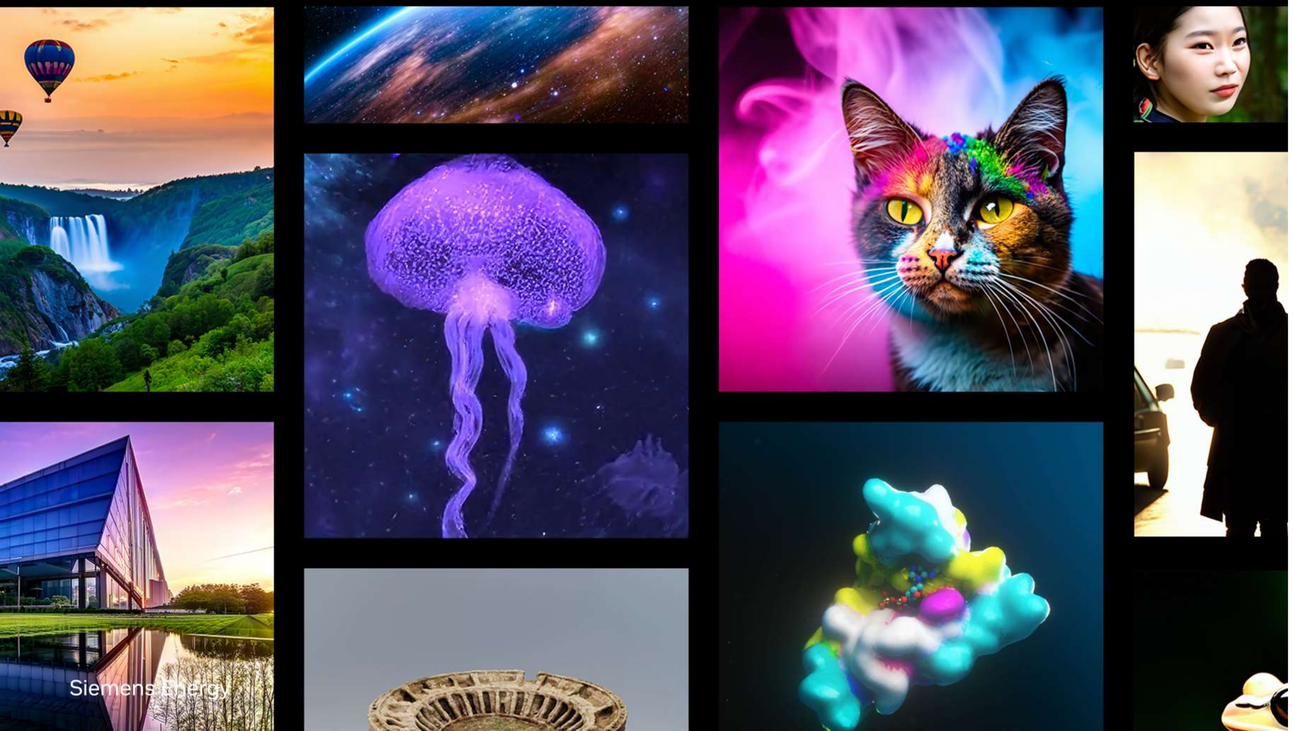
“NVIDIA agora possui um ecossistema robusto e autossustentável de software, universidades, startups e parceiros que permitiram que ela se tornasse referência de um universo recém-criado. ” FORBES

Em menos de dois anos depois dobramos o número de desenvolvedores no nosso programa, isso mostra o grande interesse nas nossas plataformas.

Faça parte deste ecossistema:

Build and Grow Your Generative AI Startup

NVIDIA Inception



NVIDIA Inception

https://www.nvidia.com/pt-br/startups/

O NVIDIA Inception é uma plataforma de aceleração para startups de AI, ciência de dados e HPC que oferece suporte, expertise e tecnologia essenciais para inserção no mercado.

“Participamos de uma série de briefings e eventos da NVIDIA em que fomos apresentados a pessoas que se tornaram nossos clientes. No GTC, você conhece outras empresas que estão desenvolvendo tecnologias complementares e investidores interessados na área que podem ajudar sua empresa.” Adam Bry, Cofundador e CEO, Skydio

NVIDIA | Inception Program

Explore Breakthroughs in AI and Beyond With the Best of GTC

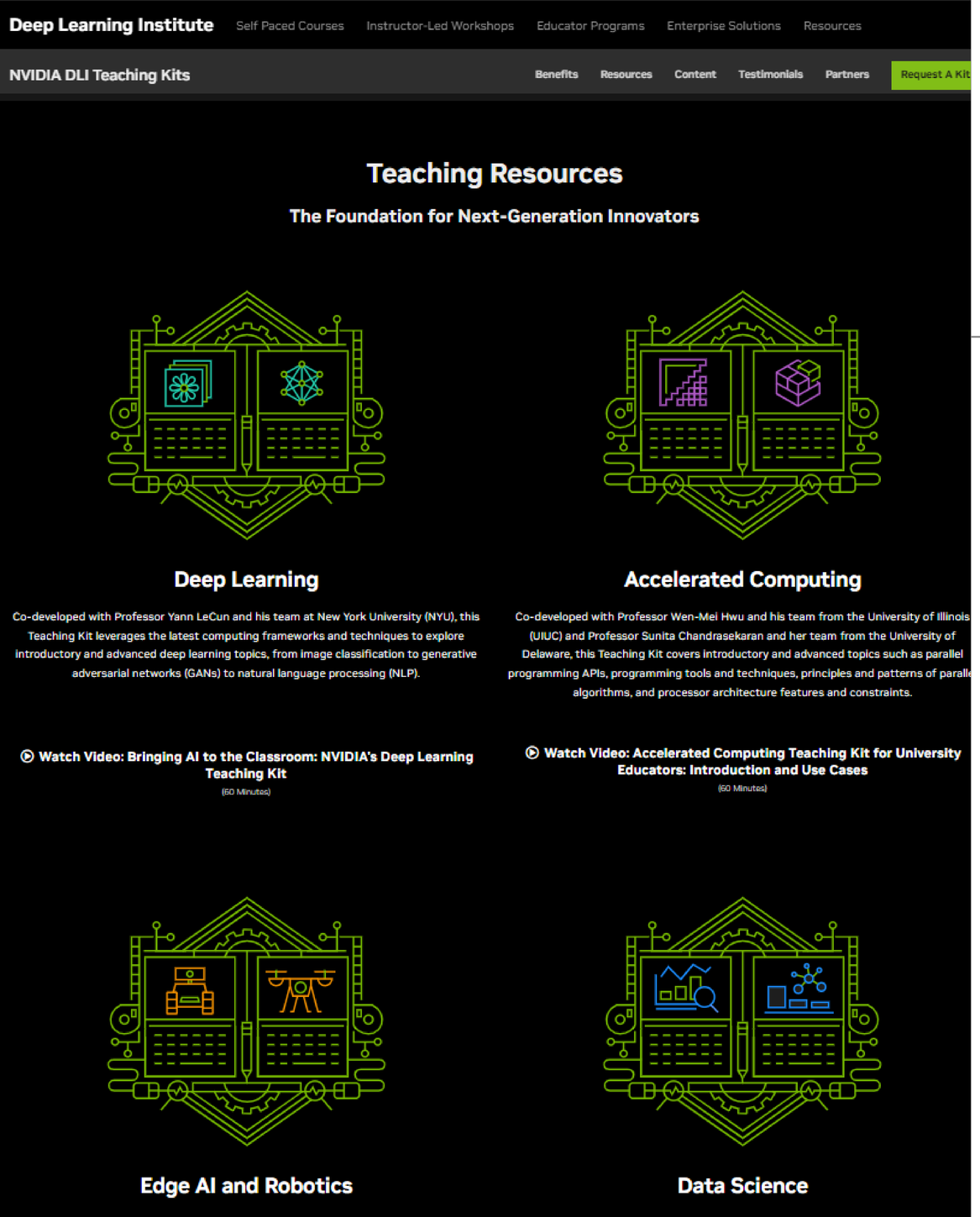
NVIDIA GTC

https://www.nvidia.com/gtc/

Com mais de 650 sessões gratuitas sobre o que há de mais moderno em em IA generativa, metaverso, grandes modelos de linguagem ampla (LLMs), robótica, computação em nuvem e muito mais.

Os participantes podem obter insights e treinamento sobre as tecnologias avançadas que transformam as diversas indústrias.


accenture
ADEPT
Adobe
Alibaba Cloud
amazon
Argonne NATIONAL LABORATORY
Ansys
ANTHROPIC
anyscale
AUTODESK
Baidu 百度
BlackRock
BMW GROUP
BNP PARIBAS
Boston Dynamics
Bristol Myers Squibb
c.ai
cadence
Canonical
Canva
Capital One
CERN
Chan Zuckerberg Initiative
CLEARML
cohere
CoreWeave
Continental
databricks
DATADOG
data iku
dataloop
DataRobot
ddn
Deepgram
DELL
Deloitte.
Disney Research
DOMINO
ebay
IIElevenLabs
EQUINIX
ExxonMobil
Ford
Foster + Partners
FOXCONN
GE Healthcare
gettyimages
Genentech
GIGABYTE
glean
Google
Goldman Sachs
GSK
H2O.ai
HEXAGON
Honeywell
hp
Hewlett Packard Enterprise
Hugging Face
IBM
imbue
JLR
JOHN DEERE
Johnson&Johnson Innovative Medicine
JPMORGAN CHASE & CO.
Kawasaki
KPF
kt
Lambda
LangChain
Lenovo
Lightning AI
LinkedIn
LOCKHEED MARTIN
LlamaIndex
L'ORÉAL
Los Alamos NATIONAL LABORATORY
LOWE'S
MathWorks
MERCK
Meta AI
Medtronic
Microsoft
MISTRAL AI_
MITRE
Modular
moz://a
NASA Langley Research Center
NetApp
NERSC
NETL NATIONAL ENERGY TECHNOLOGY LABORATORY
NETFLIX
NIST National Institute of Standards and Technology
NORTHROP GRUMMAN
NTT Group
OpenAI
ORACLE
orange
PayPal
PEGATRON
perplexity
PETROBRAS
PhotoRoom
PING AN
Pinterest
PIXAR
QCT
Raytheon Technologies
Recursion
Red Hat
rescale
Rockwell Automation
RSS-Hydro
run:ai
runway
SAP
SAMSUNG ELECTRONICS
aramco
scale
slb
servicenow
Shell
shutterstock
SIEMENS
Snapchat
snowflake
SoftBank
STELLANTIS
SUPERMICRO
synthesia
tabnine
together.ai
Tencent 腾讯
Unilever
Uber
ups
U.S. ARMY
NAVY FORGED BY THE SEA
USPTO
verizon
vmware
Weights & Biases
WEKA
WELLS FARGO
wistron
WPP
XL



DLI Teaching Kits

https://www.nvidia.com/en-us/training/teaching-kits/

- Removendo Barreiras para o Ensino de Novas Tecnologias
  - Economiza Tempo de Planejamento de Aulas e Laboratórios
  - Reduz Custos e Necessidades de Infraestrutura
  - Aborda Teoria Acadêmica e Fundamentos
  - Oferta de Curso Singular e Abrangente com Suporte
- Deep Learning
- Accelerated Computing
- Edge AI and Robotics
- Data Science
- Graphics and Omniverse
- Science and Engineering
- Generative AI (NEW)

Obrigado!

Fabio Alves de Oliveira

NVIDIA – Gerente de Negócios

fdeoliveira@nvidia.com

Realização:

Apoio: