- NVIDIA Platforms
- GH200 (Grace Hopper)

Jay Chen, Data Scientist | Aug 2023.

Platforms for Discovery

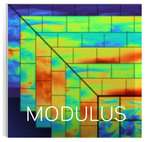
AI APPLICATION FRAMEWORK

Edge Processing

Speech AI

Customer Service

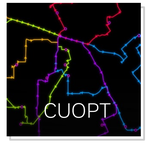
Recommenders



Physics ML

Video

Video Analytics



Logistics

Conversational AI

Robotics

Autonomous Vehicles

Cyber-security

Healthcare

Customer Service

PLATFORMS


NVIDIA HPC


NVIDIA AI


NVIDIA Omniverse

ACCELERATION LIBRARIES

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

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

CUDA Quantum | DOCA | Magnum IO | Aerial | cuFFTMp

4M+ Developers

450+ SDKs/Models

CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS

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

3 CHIPS


GPU

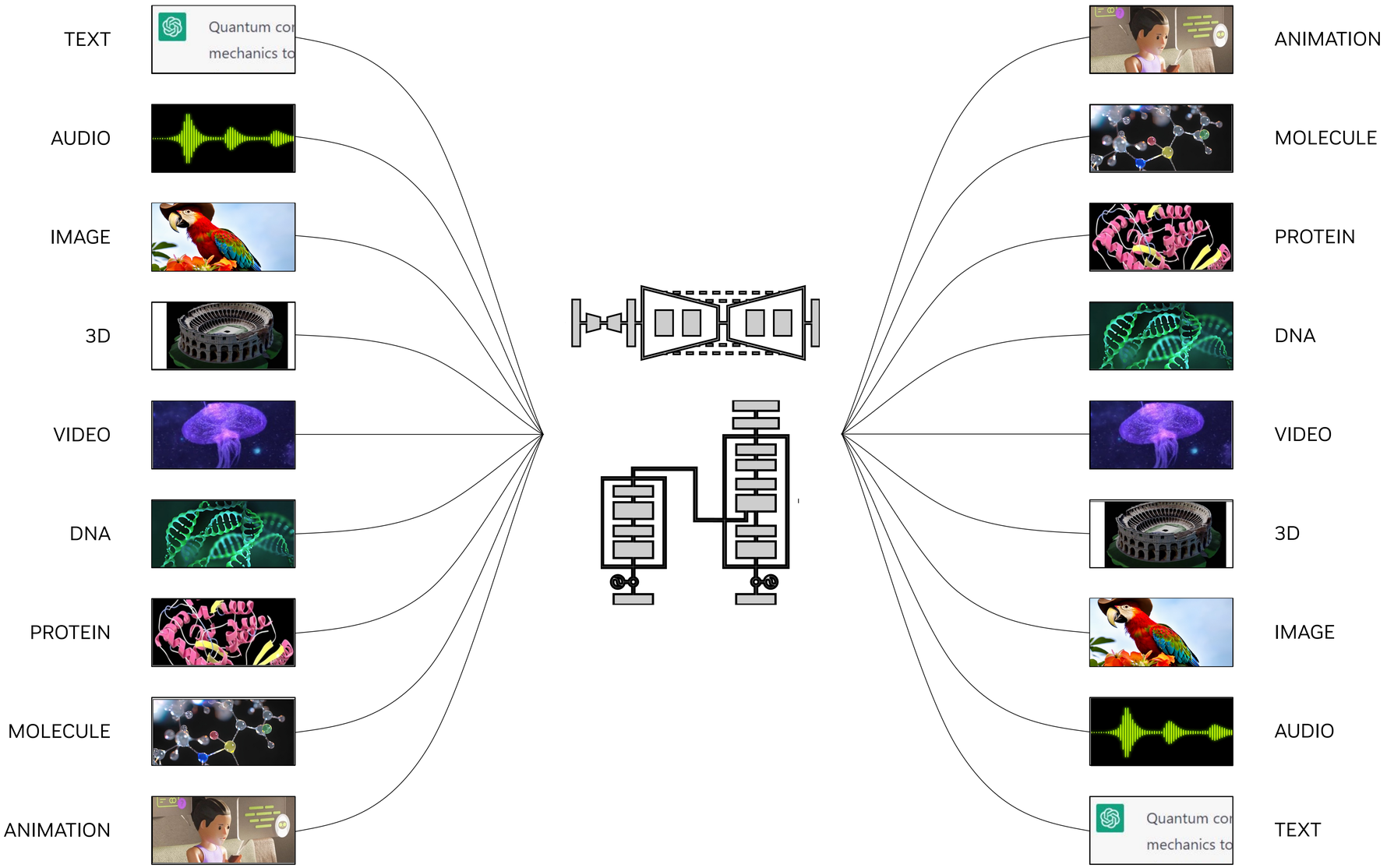

CPU


DPU

NVIDIA

NVIDIA AI Platform - Gen AI

What is Generative AI?



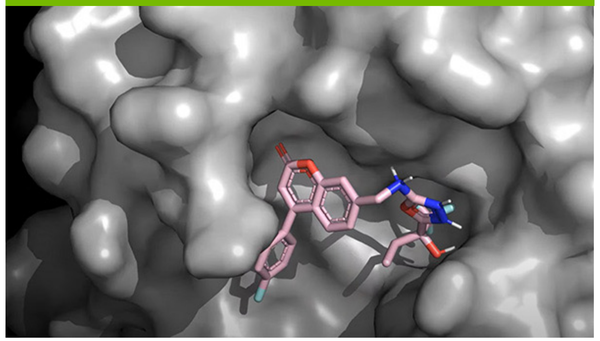
NVIDIA AI Foundations

Foundations to Build and Run Your Generative AI

NVIDIA AI FOUNDATIONS

NeMo

BioNeMo



Picasso

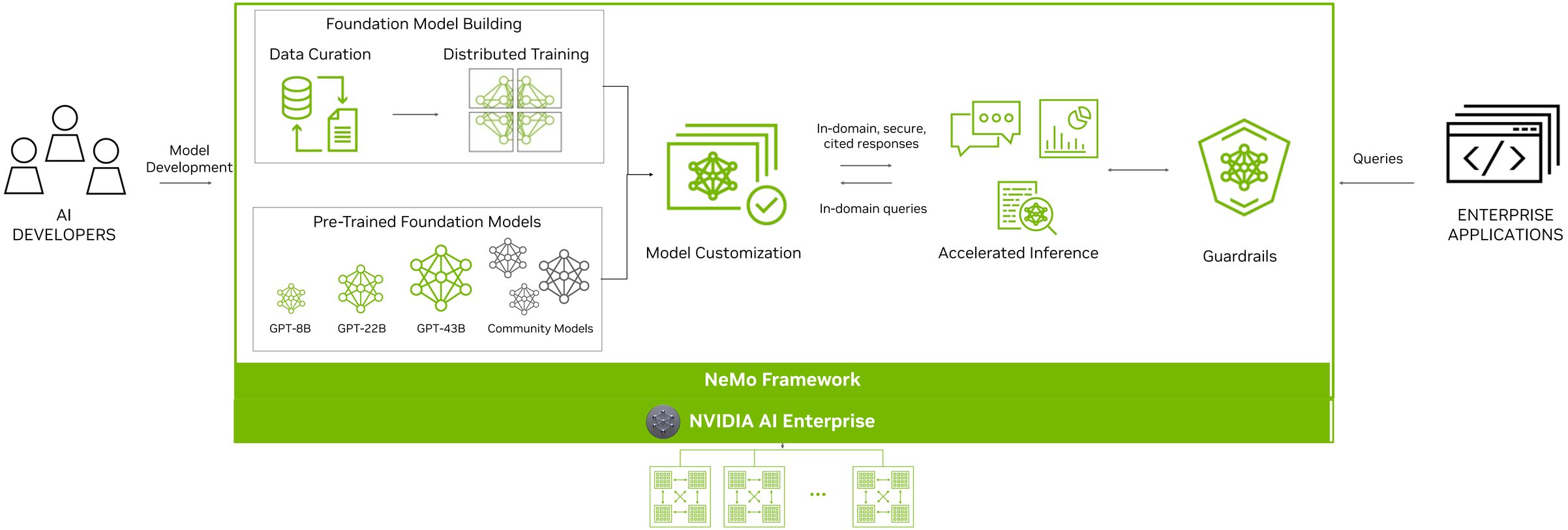
A photo of a golden retriever puppy wearing a green shirt. The shirt has text that says "NVIDIA rocks". Background office. 4k dslr.

NVIDIA AI ENTERPRISE

NVIDIA Accelerated Compute Infrastructure

NVIDIA NeMo for Custom LLMs

End-to-end, open and interoperable platform to build, customize and deploy generative AI models anywhere



Multi-modality support

Build language, image, generative AI models

Data Curation @ Scale

Extract, deduplicate, filter info from large unstructured data @ scale

Optimized Training

Accelerate training and throughput by parallelizing the model and the training data across 1,000s of nodes.

Model Customization

Easily customize with P-tuning, SFT, Adapters, RLHF, AliBi

Deploy at-scale Anywhere

Run optimized inference at-scale anywhere

Guardrails

Keep applications aligned with safety and security requirements using NeMo Guardrails

Support

NVIDIA experts by your side to keep projects on track

Model Architectures Supported on NeMo

Nemo Framework Now Supports Training And Deployment of Popular Model Architectures

Language Models

GPT
T5, mT5, T5-MoE
BERT

Text-to-Image Models

Prompt: A 'sks' dog mecha robot.

Stable Diffusion v1.5
Imagen
Vision Transformers
CLIP

Image-to-Image Models

Instruction: Make it on a beach

Dreambooth
InstructPix2Pix

Building Generative AI Foundation Models

Efficiently and quickly training models using NeMo

Requirements for Building Foundation Models

Mountains of Training Data

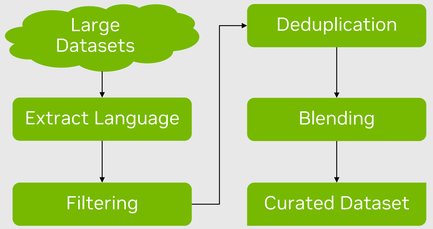
Large-scale compute infrastructure for training & inferencing, costing $10 M+

Deep technical expertise

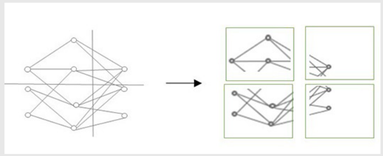
Complex algorithms to build on large-scale infrastructure

Accelerated Training With NeMo

Data Curation Tool

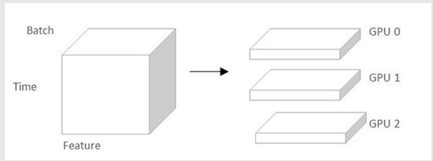


Accelerated Training



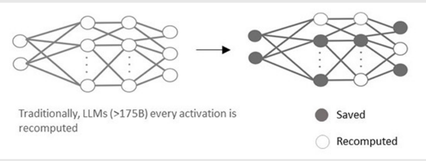
Tensor & Pipeline Parallelism

Accelerated Training



Sequence Parallelism

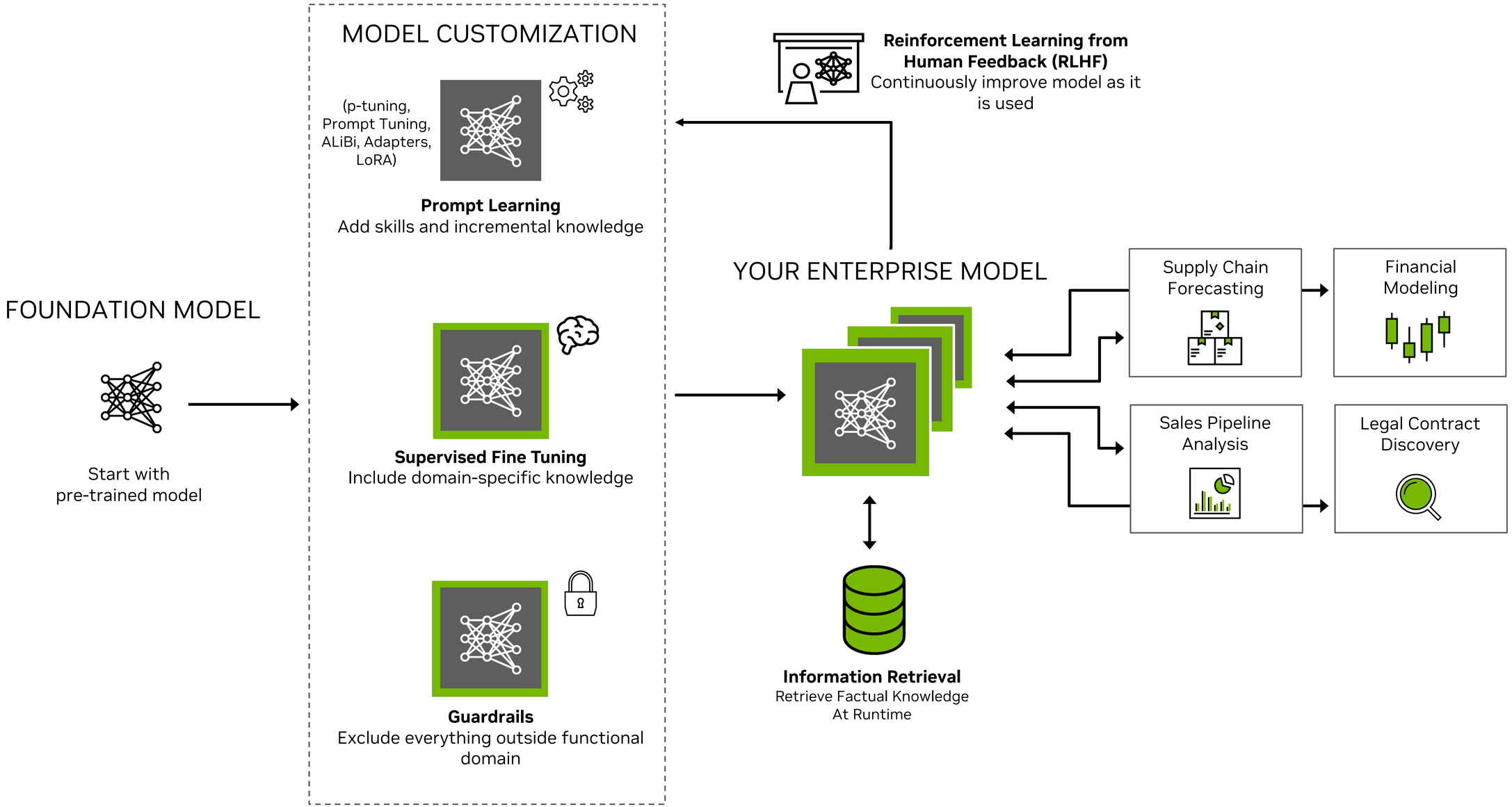
Accelerated Training



Selective Activation Recomputation

Model Customization for Enterprise Ready LLMs

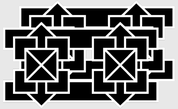
Customization techniques to overcome the challenges of using foundation models



Deploying Large Scale Inference for Generative AI

Production use-cases of Generative AI models at-scale with NeMo

Deploy Models at-Scale using NeMo

	Optimized kernels to accelerate performance for generative AI models
	Tensor/Pipeline parallelism for multi-GPU and multi-node inference
	MPI and NCCL for intra/inter-node communication

NeMo Framework Performance - Training

	Time to train 300B tokens in days (A100) – BF16				Expected speedup with H100 – FP8 & Transformer Engine			
	800 GPUs (100 DGX A100)	480 GPUs (60 DGX A100)	160 GPUs (20 DGX A100)	64 GPUs (8 DGX A100)	800 GPUs (100 DGX H100)	480 GPUs (60 DGX H100)	160 GPUs (20 DGX H100)	64 GPUs (8 DGX H100)
GPT-3: 126M	0.07	0.12	0.37	0.92	1.75x	1.71x	1.68x	1.7x
GPT-3: 5B	0.8	1.3	3.9	9.8	2.67x	2.17x	2.29x	2.39x
GPT-3: 20B	3.6	6	18.1	45.3	2.57x	2.61x	2.66x	2.66x
GPT-3: 40B	6.6	10.9	32.8	82	3x	3.03x	3.04x	3.04x
GPT-3: 175B	28	46.7	140	349.9	3.08x	3.07x	3.07x	3.07x

NeMo Framework Performance - Inference

	Data measured on FT in BF16 for A100 [All data for 200/200 in/out sequence length]			Expected performance improvement with H100 – FP8 & Transformer Engine [All data for 200/200 in/out sequence length]		
	Latency (Batch size = 1) (ms)	Throughput at similar latency (token/sec)		Latency (Batch size = 1) (ms)	Throughput at similar latency (token/sec)	
		Throughput (token/sec)	Latency (time per generated token) (ms)		Throughput (token/sec)	Latency (time per generated token) (ms)
GPT-3: 5B (1 GPU)	1514.74	3319.7	9.64	0.66x	4.14x	9.32
GPT-3: 20B (1 GPU)	5387.68	74.24	26.94	0.51x	68.3x	25.24
GPT-3: 20B (2 GPUs)	3275.11	239.83	16.68	0.62x	32.12x	16.62
GPT-3: 175B (8 GPUs)	7936.94	380.87	42.01	0.71x	8.08x	41.6

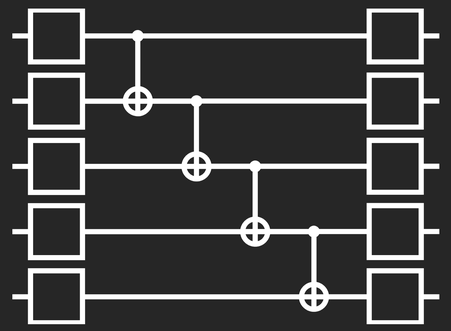
NVIDIA HPC Platform

- Quantum
- HPC+AI

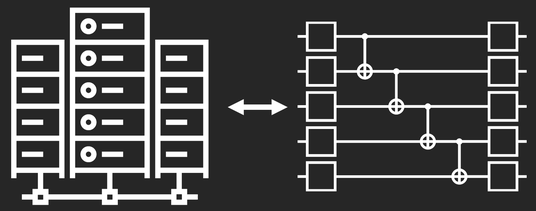
NVIDIA Quantum

Powering Quantum Simulation and Quantum-Integrated Accelerated Computing

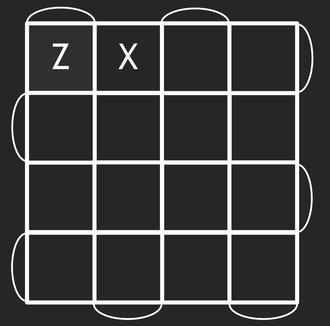
Quantum Algorithms Research



Quantum-Integrated Applications



Error Correction, Calibration, Control



cuQuantum
Accelerated Quantum Simulation

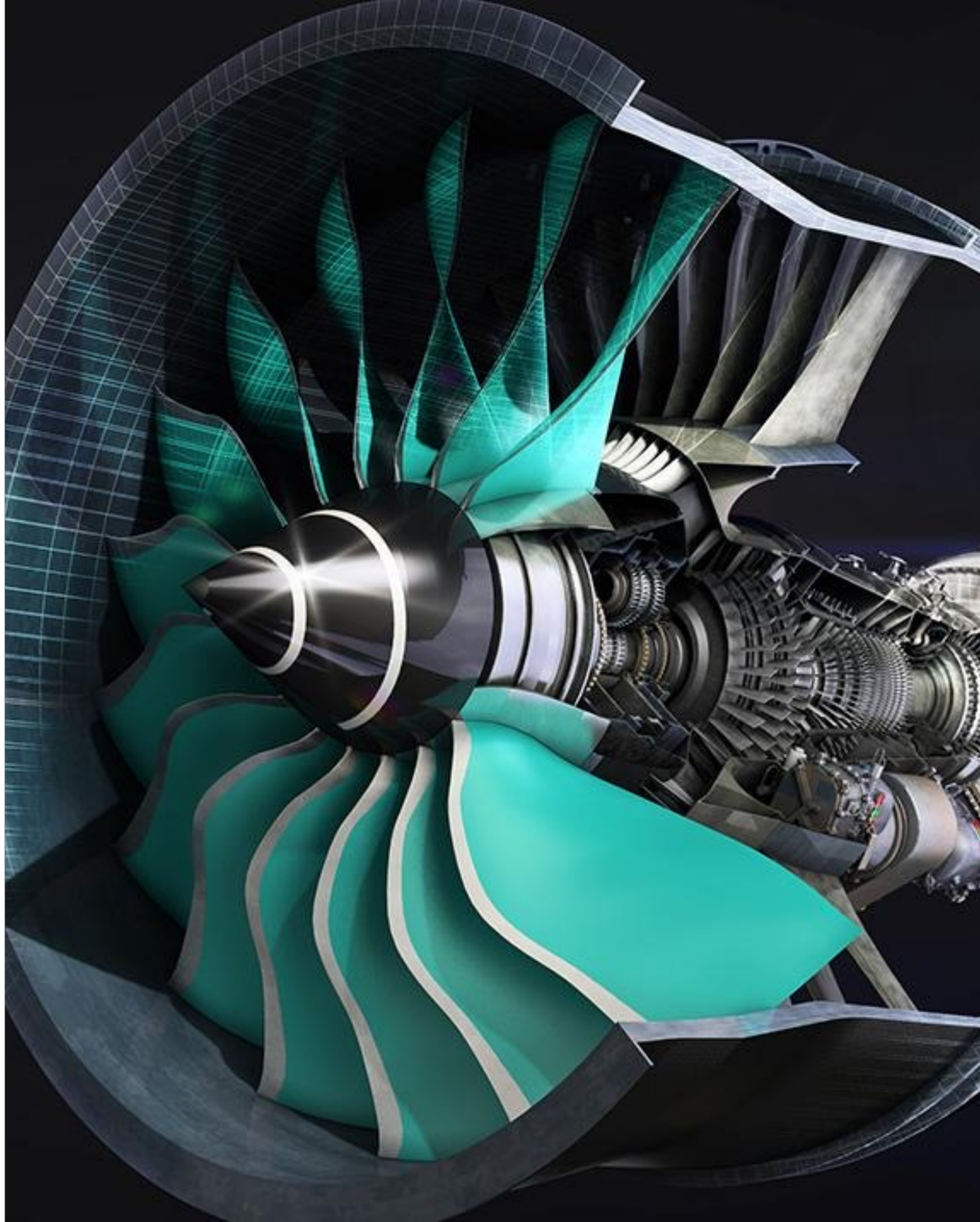
CUDA Quantum
Quantum-Classical Developer Platform

Quantum Integrated GPU Supercomputing
DGX | HGX | **DGX Quantum**

[ISC23]
Rolls Royce simulates world's largest circuit for computational fluid dynamics

Enabled by cuQuantum multi—node QC simulation

- cuQuantum multi-node simulation on A100 system
- Applied quantum computing to novel CFD application
- 10,000,000 gate depth circuit with 39 qubits simulated on 64 A100 GPUs
- Partnership between Rolls Royce, NVIDIA, Classiq
- Target Net Zero for 2050



[ISC23]
Announcing NVIDIA-Jülich Joint Lab For Quantum-Classical Computing

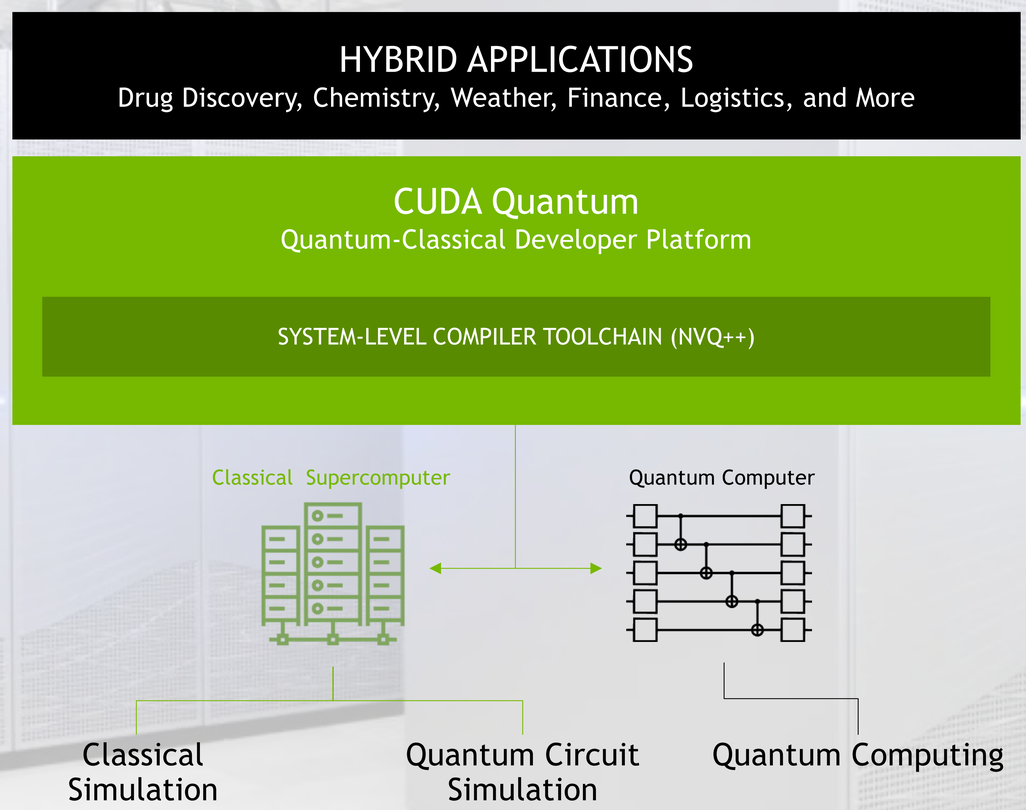
Enabling Breakthroughs in Quantum-Classical Research and Development

- Built on Jülich Unified Infrastructure (JUNIQ)
- Open-Source CUDA Quantum Programming Model integrated within Jülich's modular supercomputing architecture
- Creating a lab of the future that integrates both HPC and quantum computing technologies

JÜLICH Forschungszentrum

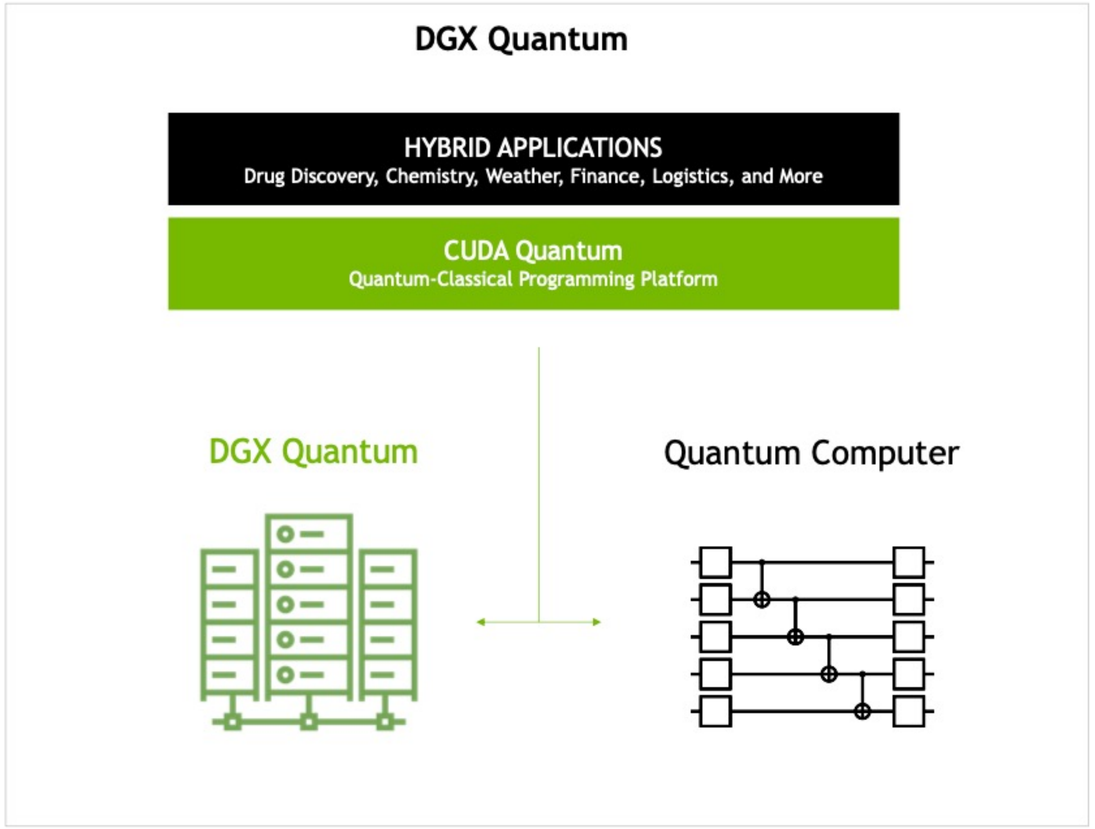
NVIDIA

ParTec Modular Supercomputing



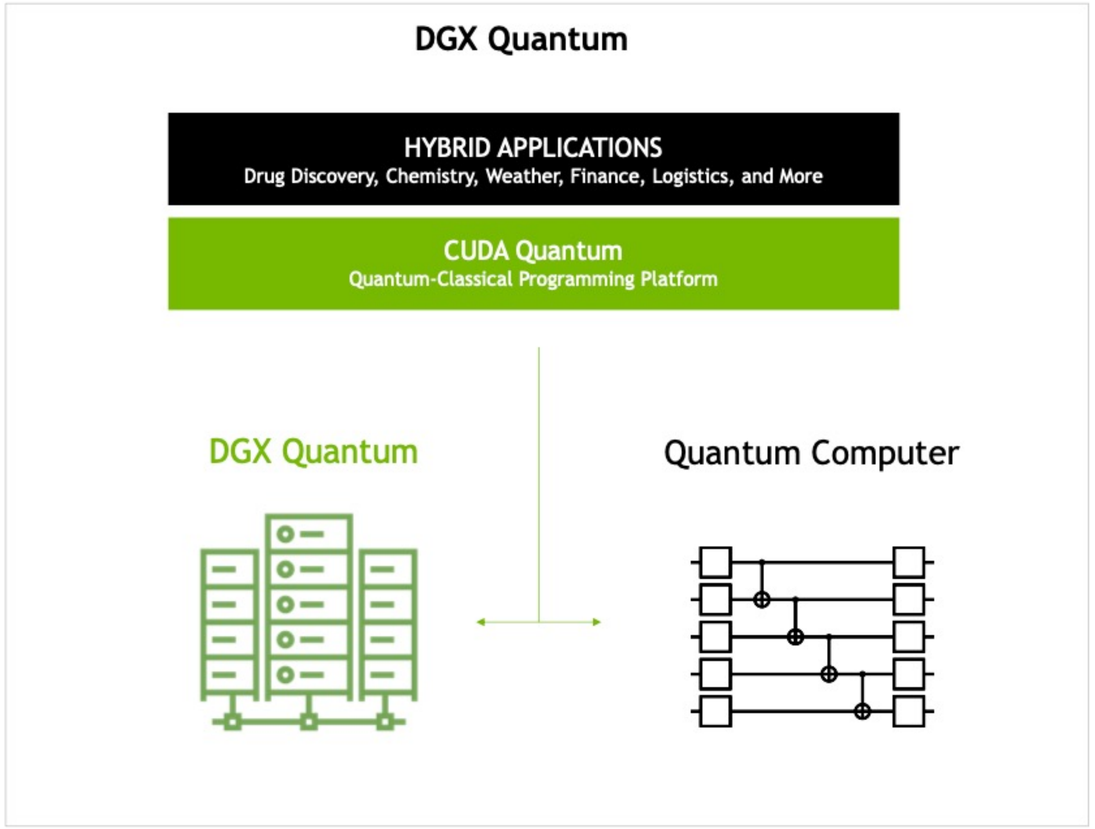
DGX Quantum

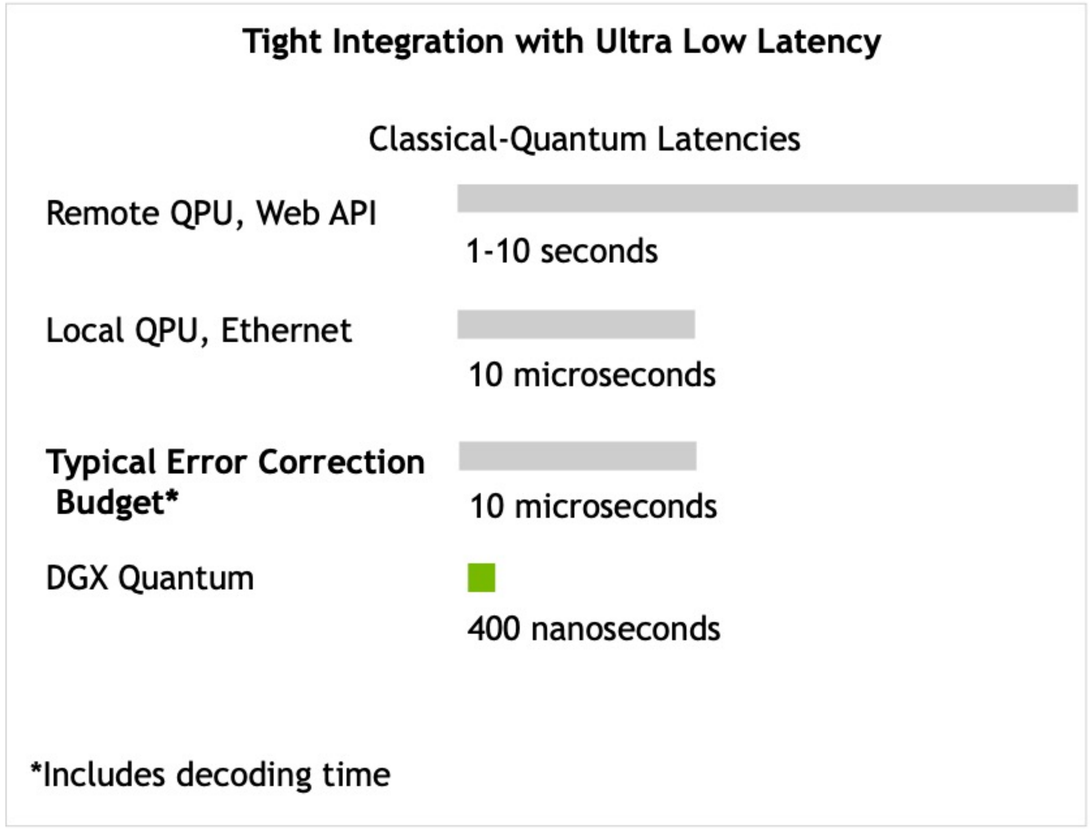
System for Integration of Quantum with GPU Supercomputing



DGX Quantum

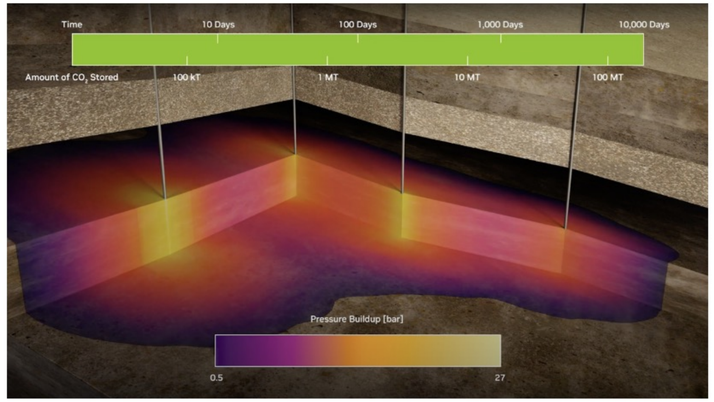
System for Integration of Quantum with GPU Supercomputing





Earth-2

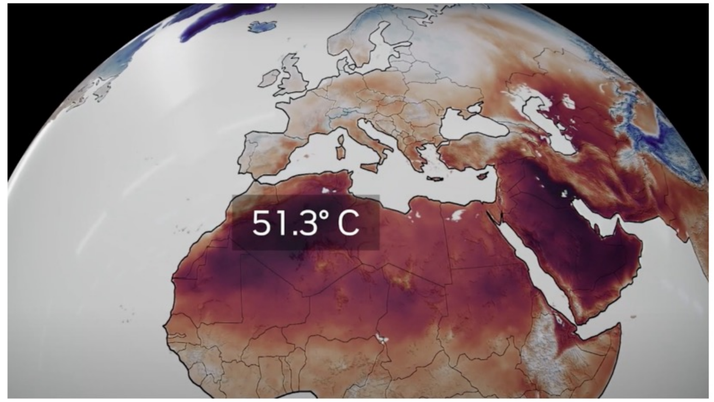
https://www.nvidia.com/en-us/high-performance-computing/earth-2/



Accelerating Carbon Capture and Storage with Fourier Neural Operator and NVIDIA Modulus

By accelerating analysis 700,000X, NVIDIA Omniverse and Modulus can help engineers with the planning and operation of carbon capture and storage, ensuring safe operation and long-term storage and reducing the amount of carbon dioxide released into our atmosphere.

Watch Carbon Capture Storage Demo ›



Predicting Extreme Weather Events Three Weeks in Advance With FourCastNet

By running FourCastNet in NVIDIA Modulus, we were able to generate 21-day weather trajectories of 1,000 ensemble members in a tenth of the time it previously took to do a single ensemble—and with 1,000X less energy consumption.

Watch Predict Extreme Weather Events with FourCastNet Demo ›

Interactive Visualization of High-Resolution, Global-Scale Climate Data in the Cloud

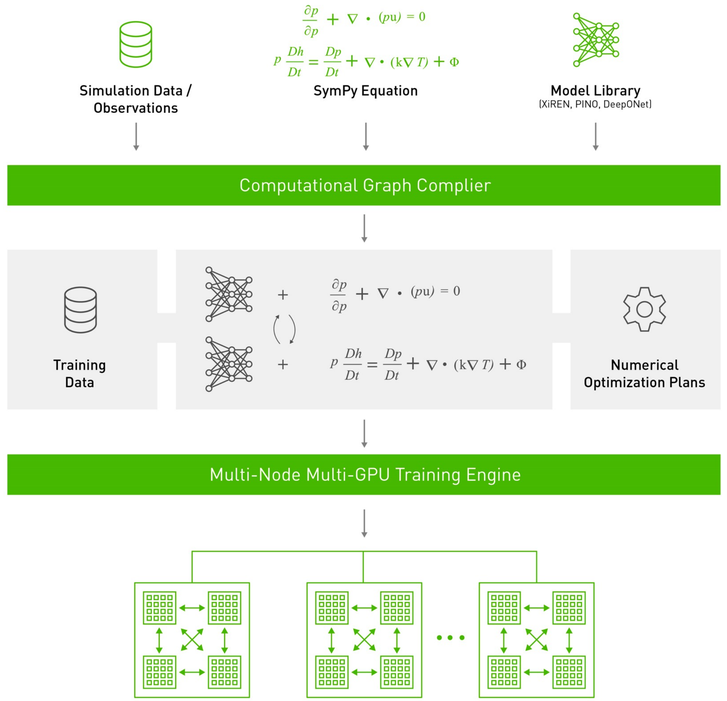
Built on NVIDIA Omniverse and the OpenUSD 3D framework, the Earth-2 platform enables aggregation and visualization of diverse, global-scale climate simulation and geospatial datasets. Made possible with cloud-native technology, visualizations can be explored by anyone around the globe.

Watch Global Climate Visualization Demo ›

NVIDIA Modulus

Open-Source Platform for Developing Physics-Based Machine Learning

TRAINING NEURAL NETWORKS USING BOTH DATA AND THE GOVERNING EQUATIONS

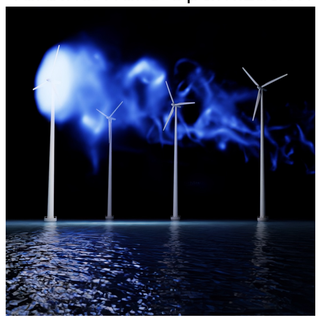


End to end GPU accelerated training pipeline validated across different domains

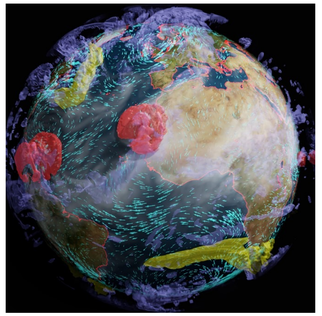
Easy to use Python APIs for domain experts – abstracting the low level details

ADVANCING SCIENTIFIC DISCOVERY WITH MODULUS

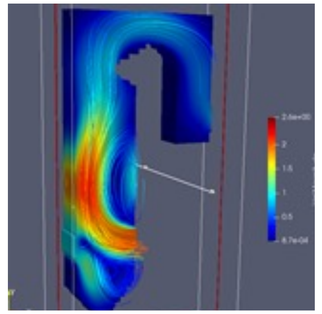
RENEWABLE ENERGY
Siemens Gamesa: Up to 4000X Speedup of Wind Turbine Wake Optimization



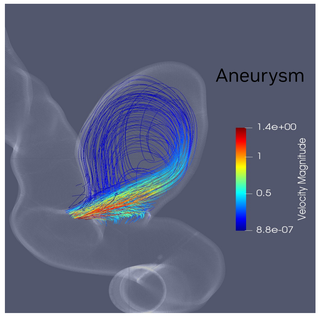
CLIMATE CHANGE
45,000X Speedup of Extreme weather Prediction with FourCastNet



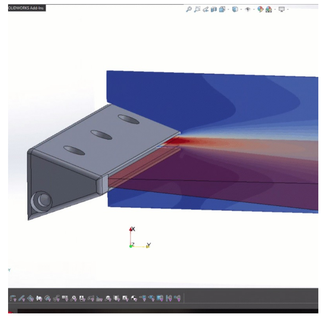
INDUSTRIAL HPC NETL:
10,000X Faster Build Of high-fidelity surrogate models



HEALTHCARE
Achieve high-fidelity results faster for blood flow in inter-cranial aneurysm



DIGITAL TWINS
Kinetic Vision: Design Optimization Using parameterized models



Newly supported, Graph Neural Networks (GNNs) and Recurrent Neural Networks (RNNs) for added ease of use for AI practitioners.