Harnessing the Power of High Performance Computing for AI / ML

Vineet Tyagi, Principal FAE, Supermicro
Gabriele Paciucci, Solutions Architect, Nvidia

Confidential

- Computing landscape
- HW Design Choices
- Scaling HPC / AI

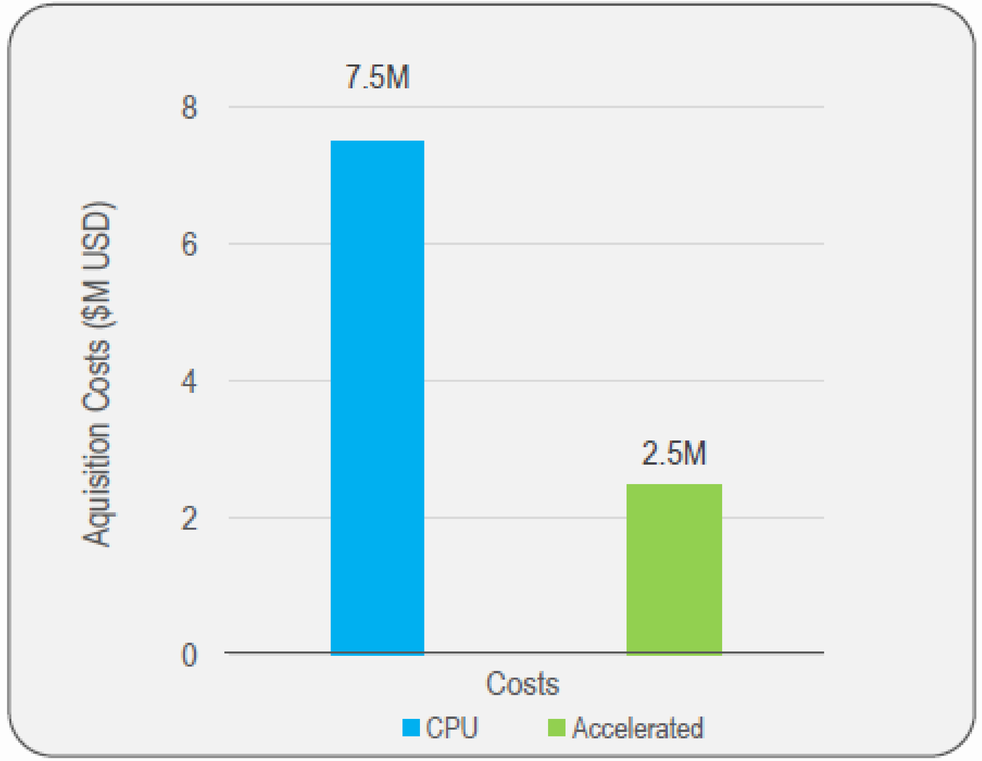
Accelerated Computing is Sustainable Computing

Confidential

Siemens Simcenter StarCCM+ Simulations for Mercedes EQE consume 3X Less Cost and 4X Less Energy

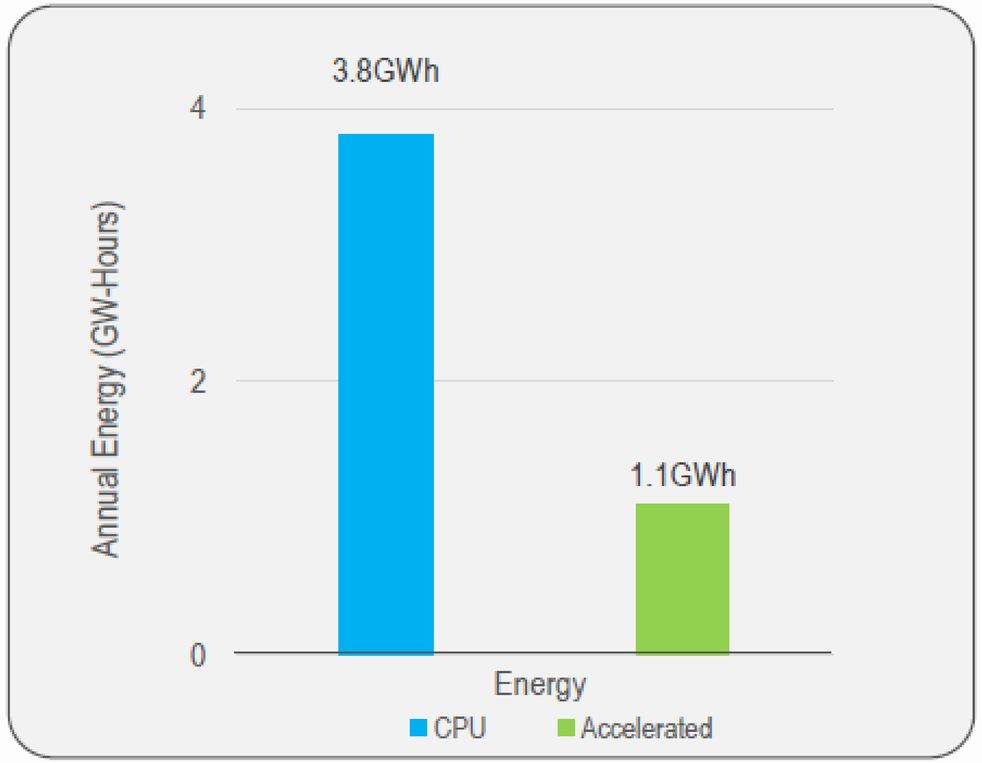
3X LESS COST

Same Throughput



4X LESS ENERGY

Same Throughput

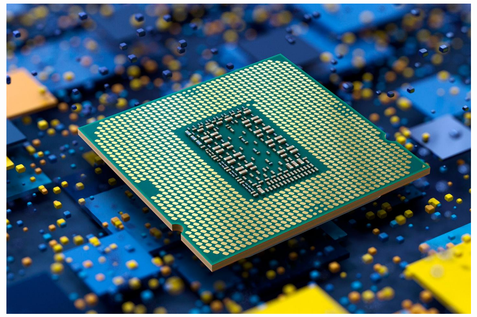


Current Computing Landscape

Confidential

CPUs

The traditional HPC Platform



GPUs

The faster parallel processing of GPUs leads to a paradigm shift for HPC

XPUs

Breaking the system boundaries

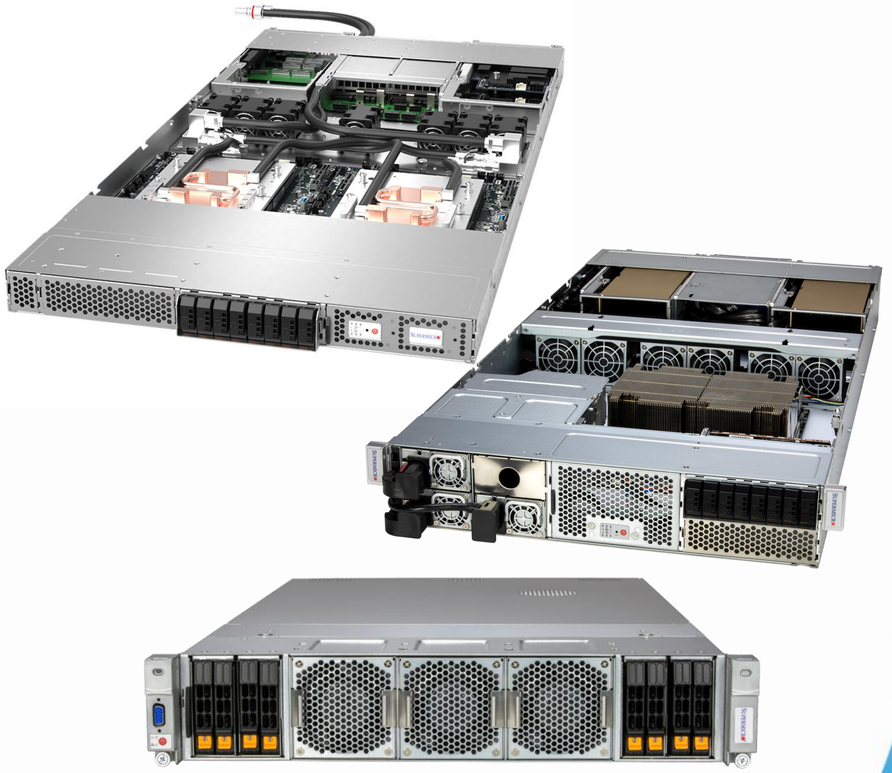
Platform Architectures

Confidential

PCIe GPU Platforms

SXM / OAM Platforms

XPU Platforms



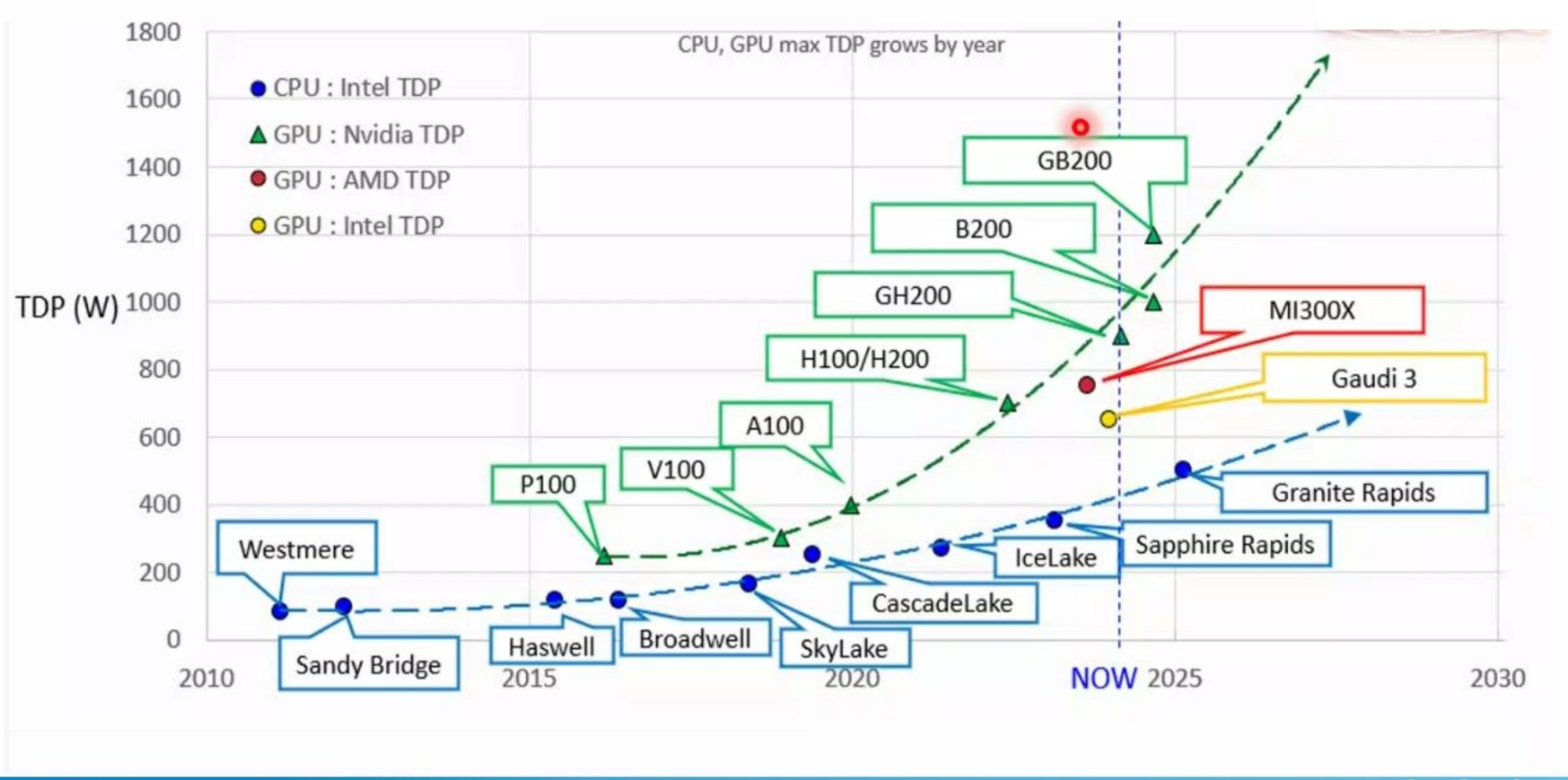
XPUs and their imperatives

Confidential

- Scale out architecture
- Shared memory
- Power
- Liquid Cooling
- Network architecture optimisations

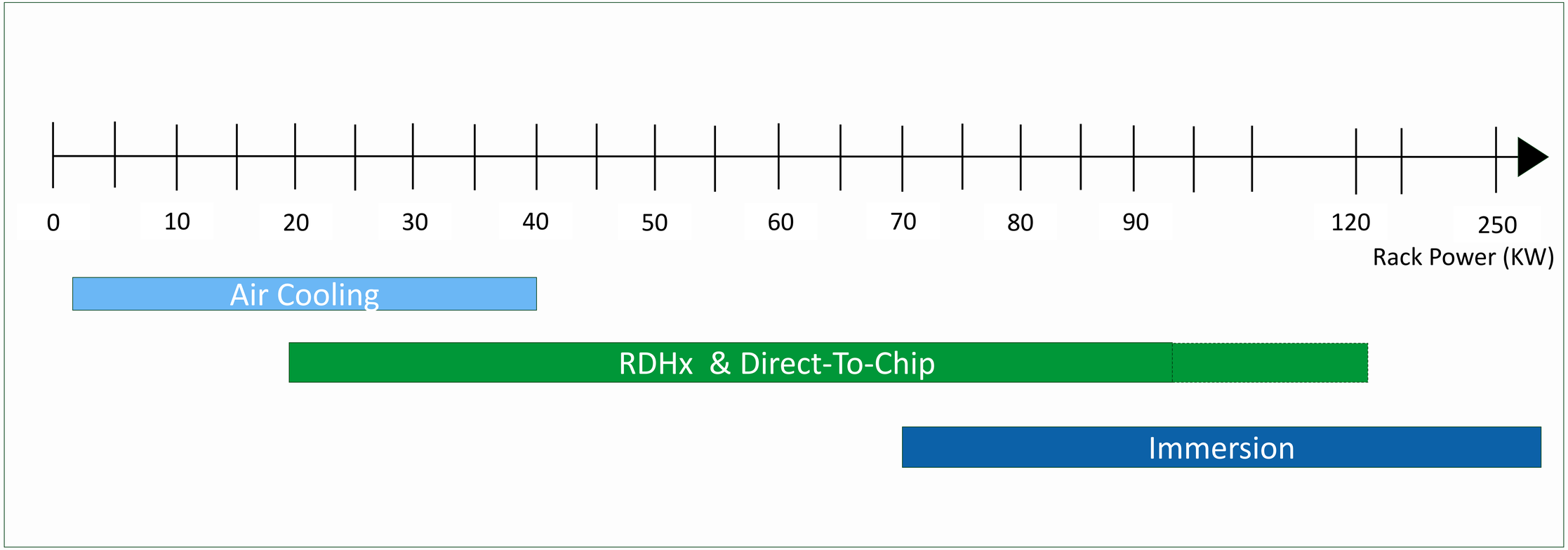
Thermal Design Power Growth

Confidential



Liquid Cooling Inflection Point

Confidential



Factors

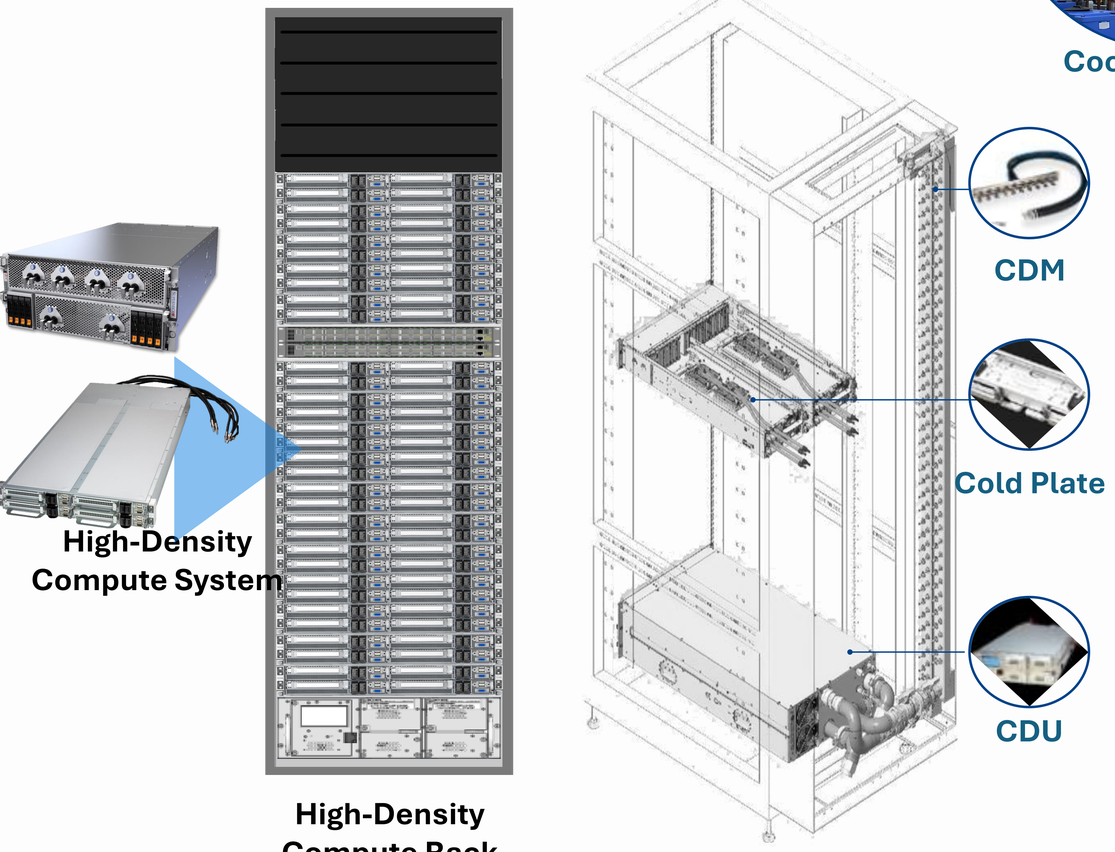
- Rack density
- Power per rack
- Platform Characteristics – GPUs / High TDP CPUs etc
- Operational Considerations – Power costs, additional infrastructure, change required in DC operations etc

Liquid Cooling

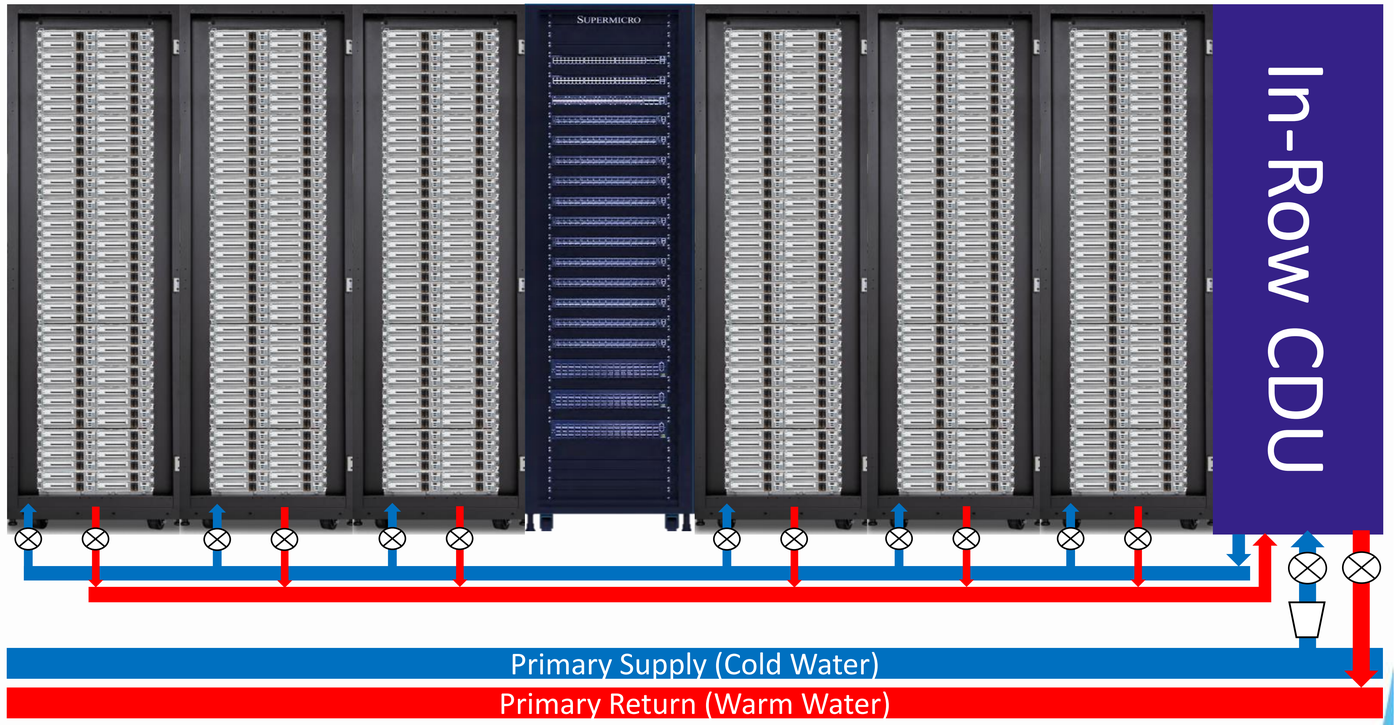
Confidential

In-Rack CDU Configuration

Cooling Tower



In-Row CDU Configuration

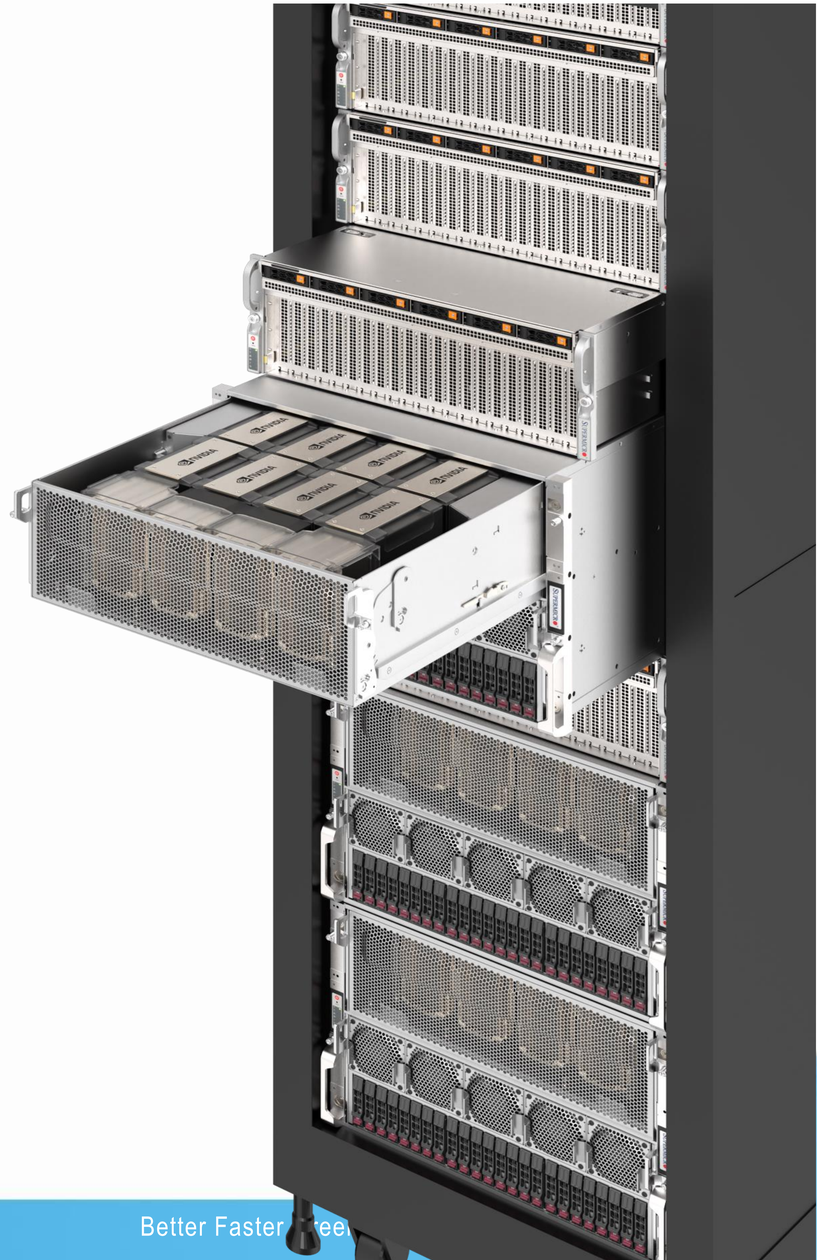


HW Design Choices - HPC & AI

Confidential

- Accelerator / Platform selection
- Workload – multiple vs one (mostly)
- Memory size & coherence
- Scalability
- Available Power
- Fast storage (Lustre / BeeGFS … vs Weka / Vast)
- Network architecture – Any-to-Any vs Rail Optimised
- Cluster level resilience / checkpointing

- Data Pipeline – Cleaning, tagging, pre-processing, ETL, Optimisation ..
- SW Stack



AI Clusters

Confidential

AI Factories

- ☑ Single or few workloads
- ☑ Extremely large AI models

AI Cloud

- ☑ Multi-tenant
- ☑ Variety of workloads

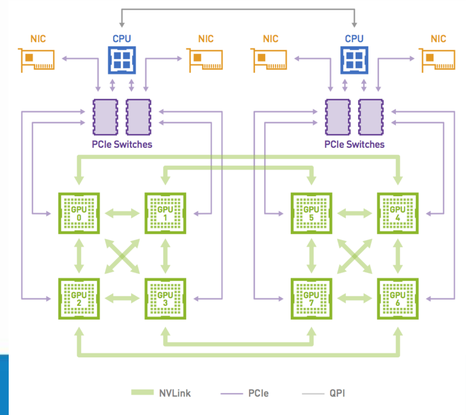
Scaling with HGX / similar Architecture

Confidential

High Density Liquid Cooled

Liquid Cooled Scalable Unit

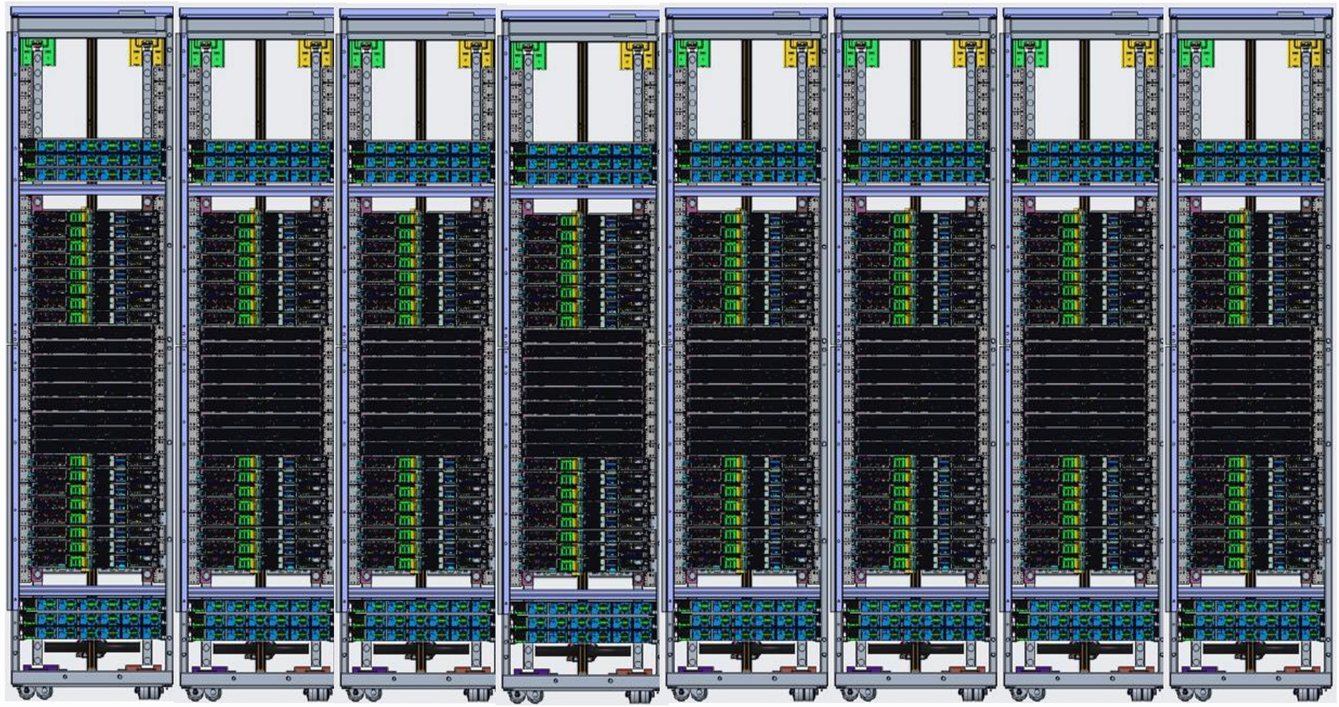
Air Cooled Scalable Unit



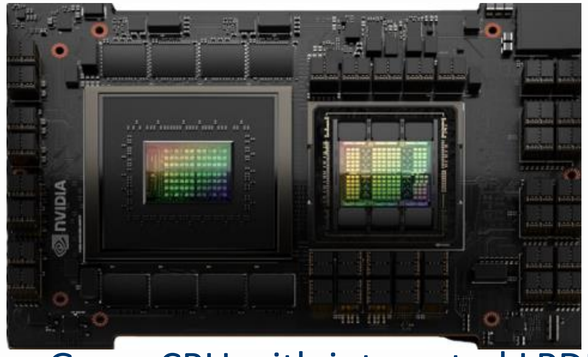
- Liquid cooling can enable higher densities / power efficiencies
- RDMA for faster GPU-to-GPU communications
- Rail Optimised NW architecture

Oberon: Largest Memory Supercomputer for Exascale Computing

Confidential

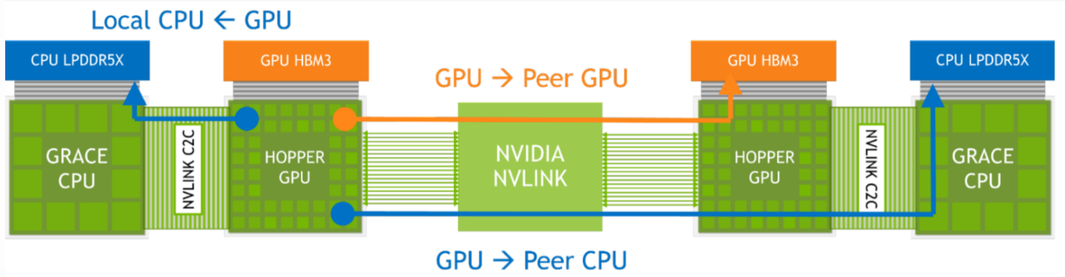


NVIDIA GH200 Grace Hopper Superchip



- One Grace CPU with integrated LPDDR5X and one H100 Tensor Core GPU (Hopper) on mezzanine module
- Fast NVLink-C2C interface between CPU and GPU
- Consumes 5X less power and provides 7X bandwidth compared to the latest PCIe technology
- Integrated with ***Direct to Chip cooling***

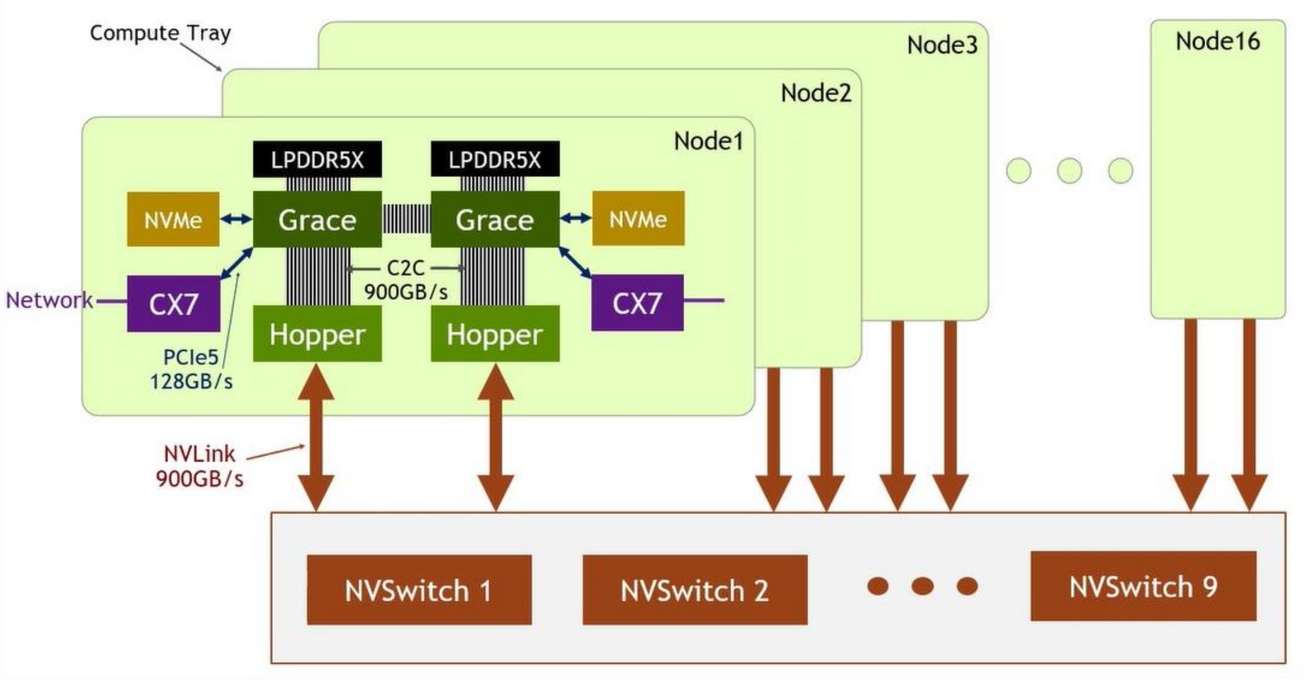
- NVLink Switches combine **256 GH200 superchips**, allowing them to perform **as a single GPU.**
- 1 exaflop of performance
- 144 terabytes of shared memory
- 48x more NVLink bandwidth than the previous generation



Scaling with XPU Architecture

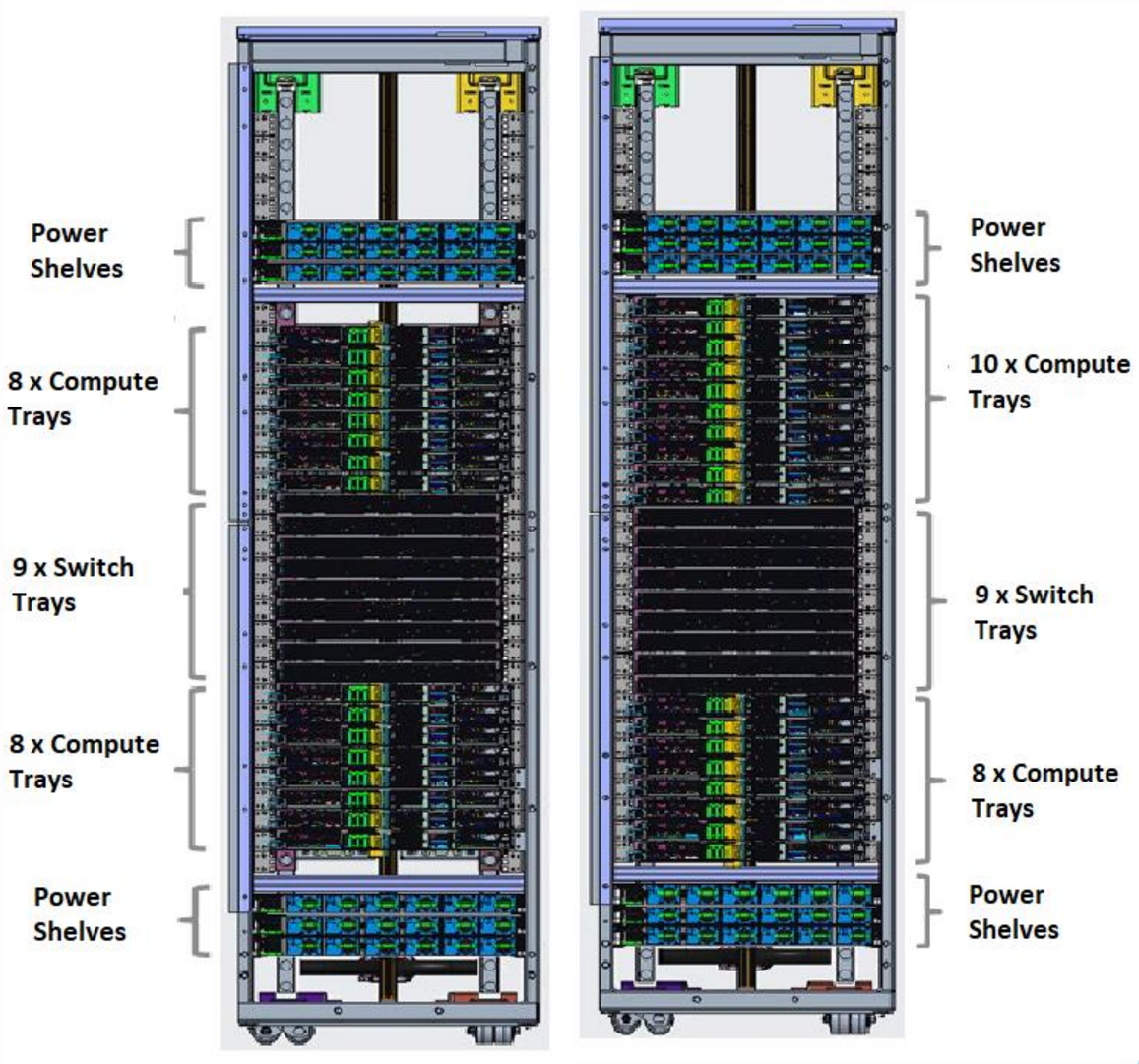
Confidential

Oberon



- Rack Scale Solution for GPUs connected via NVLink
- Oberon Building Blocks
  - Compute Trays (2 x GH200 modules per tray)
  - NVLink Switch Trays
- ORv3 Racks
- Hybrid Cooling Trays: GH200 modules & NVLink switches are LC, rest air cooled

Oberon GH200: Oberon GB200:

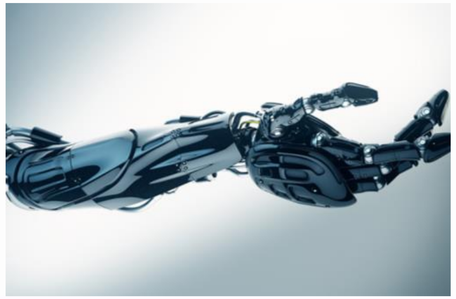


Direct to Chip Liquid cooled, ORV3 racks

Looking Ahead

Confidential

- Power& Liquid Cooling
- Ultra Ethernet vs Infiniband
- CXL for memory pooling / sharing
- Lower precision data types for AI Inference
- CISC vs RISC in HPC/AI



SUPERMICRO®

www.supermicro.com

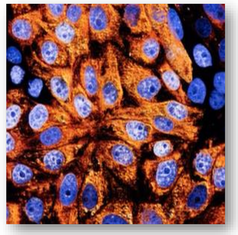
Powering Generative AI for Scientific Computing

Gabriele Paciucci

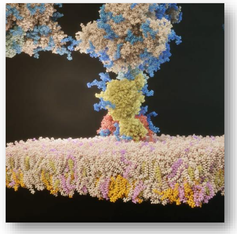
Principal Solution Architect

Pre-Exascale Supercomputing

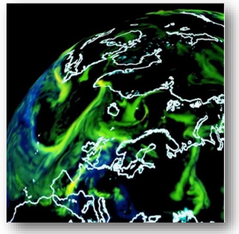
EDGE

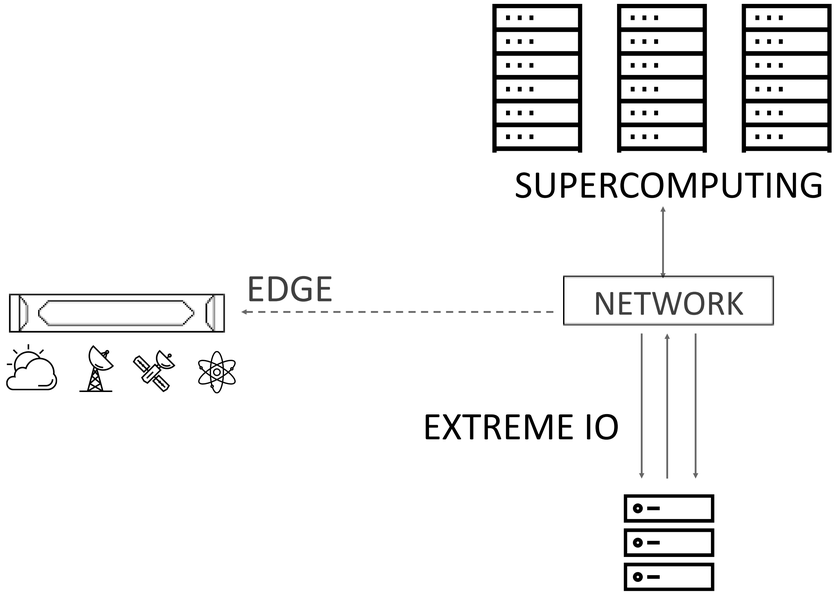


SIMULATION



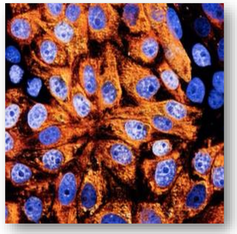
Viz





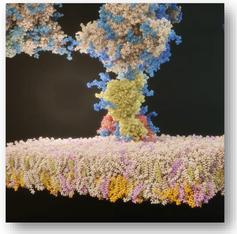
Post-Exascale Supercomputing

EDGE

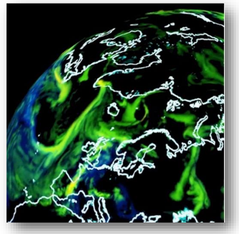


SIM + AI

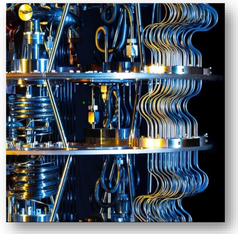
SIMULATION

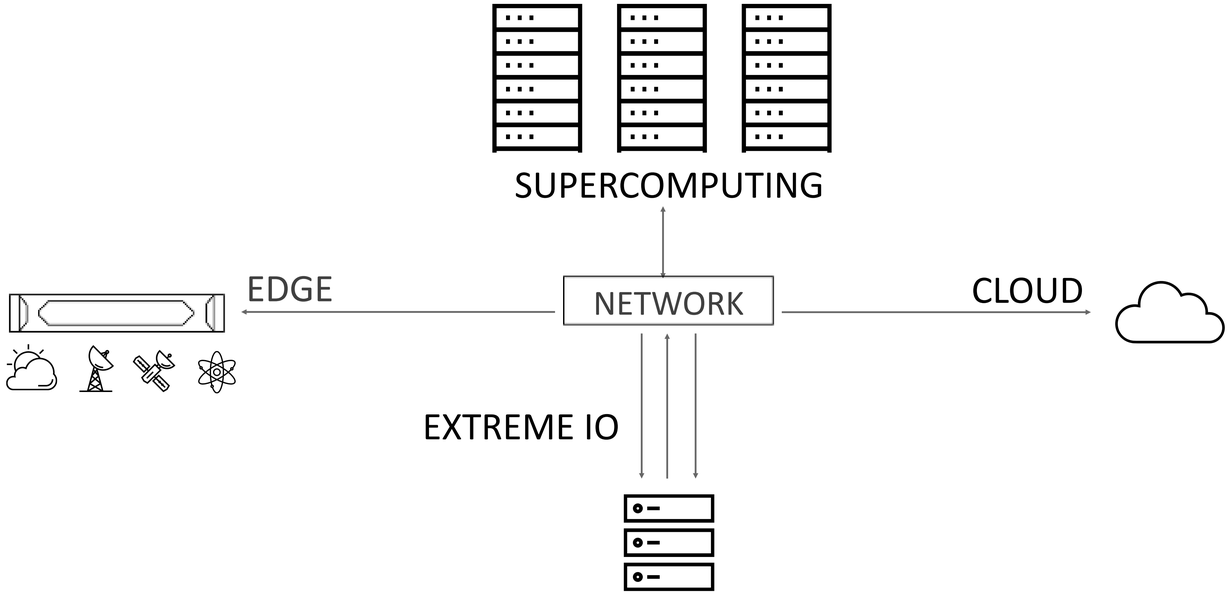


DIGITAL TWIN



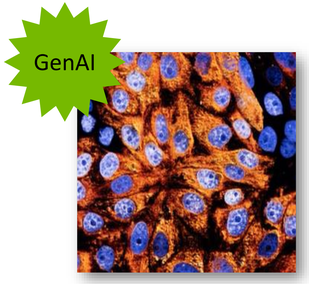
QUANTUM COMPUTING





Workloads of the Modern Supercomputer / AI Factory

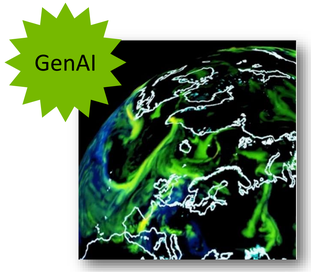
EDGE



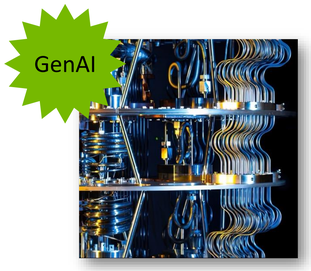
SIM + AI

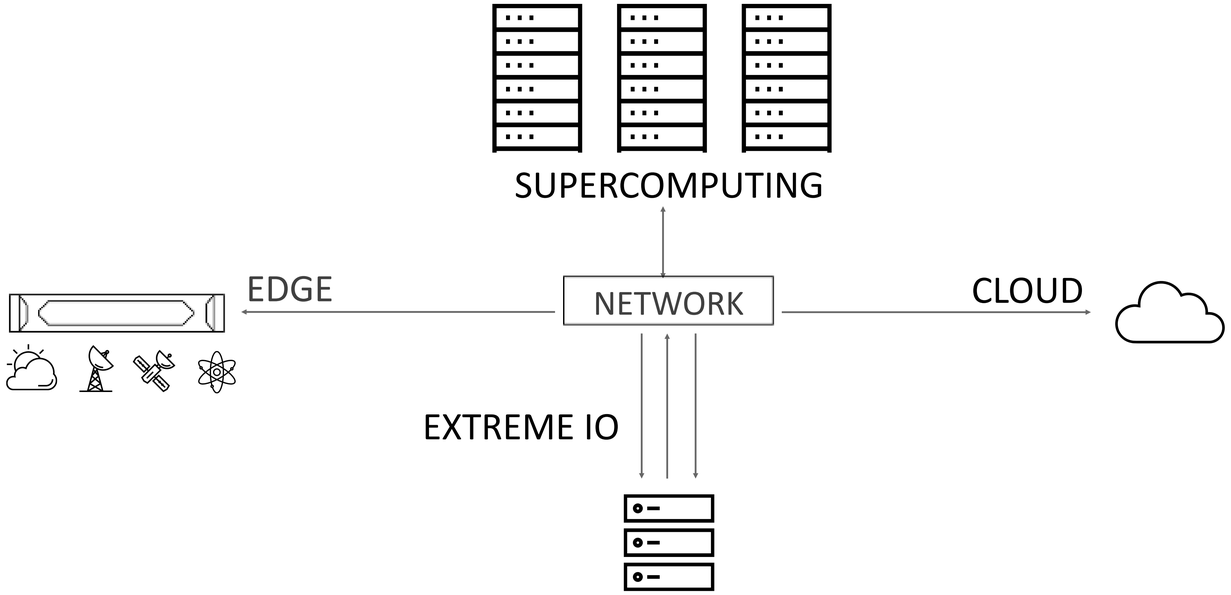
SIMULATION

DIGITAL TWIN

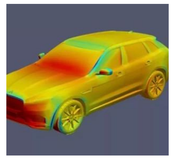
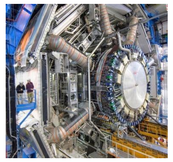
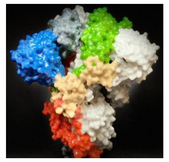
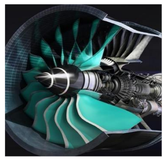
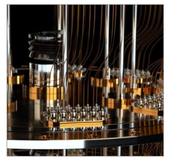


QUANTUM COMPUTING



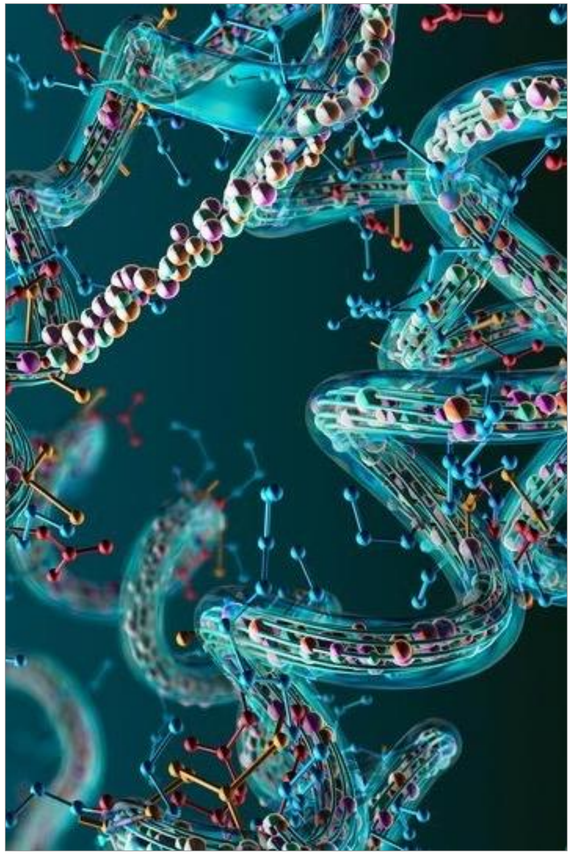


HPC Reinvented with AI (Interactive, Ease of use)

	Experiments	Simulation	Viz	Edge	HPC + AI	Simulation	Digital Twin	Quantum Computing
								
FEATURE	**PRE-EXASCALE**			**EMERGING POST EXA-SCALE - AI FACTORY**				
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				

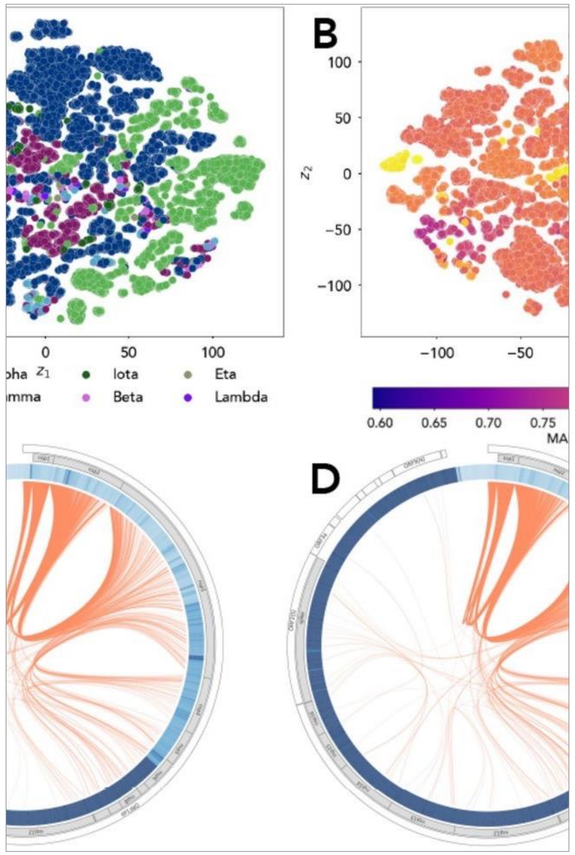
GenAI For Science Research and Discoveries

The Race for Foundation Models for Science is on



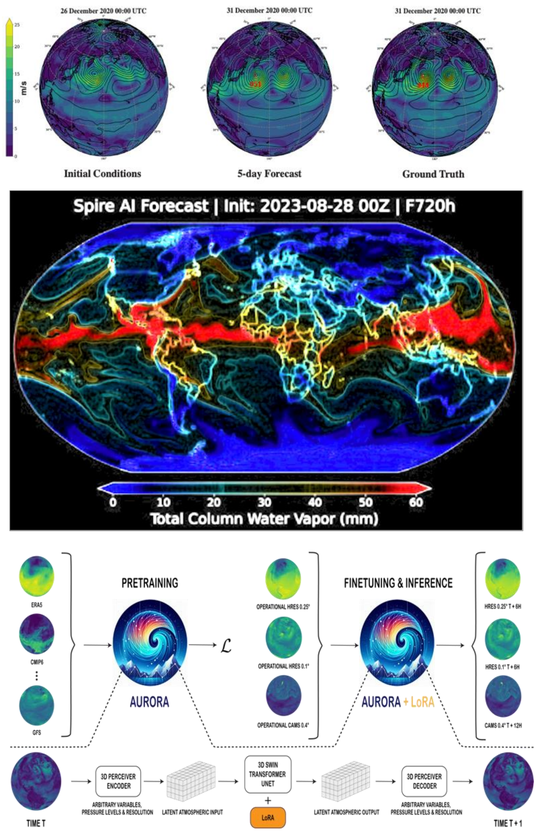
Biology: AlphaFold

2021



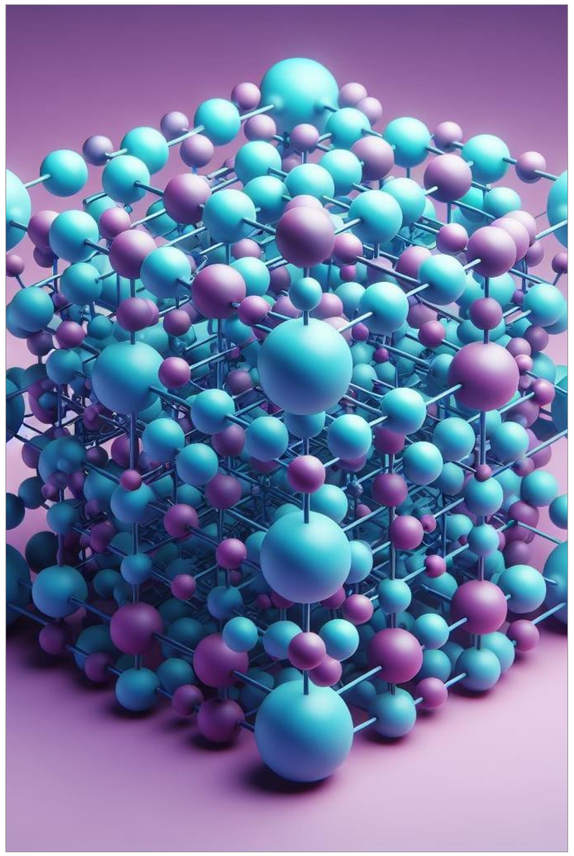
Genomics: GenSLM

Nov 2022



Weather : Stormer, Cordiff, Aurora

2023 to present day

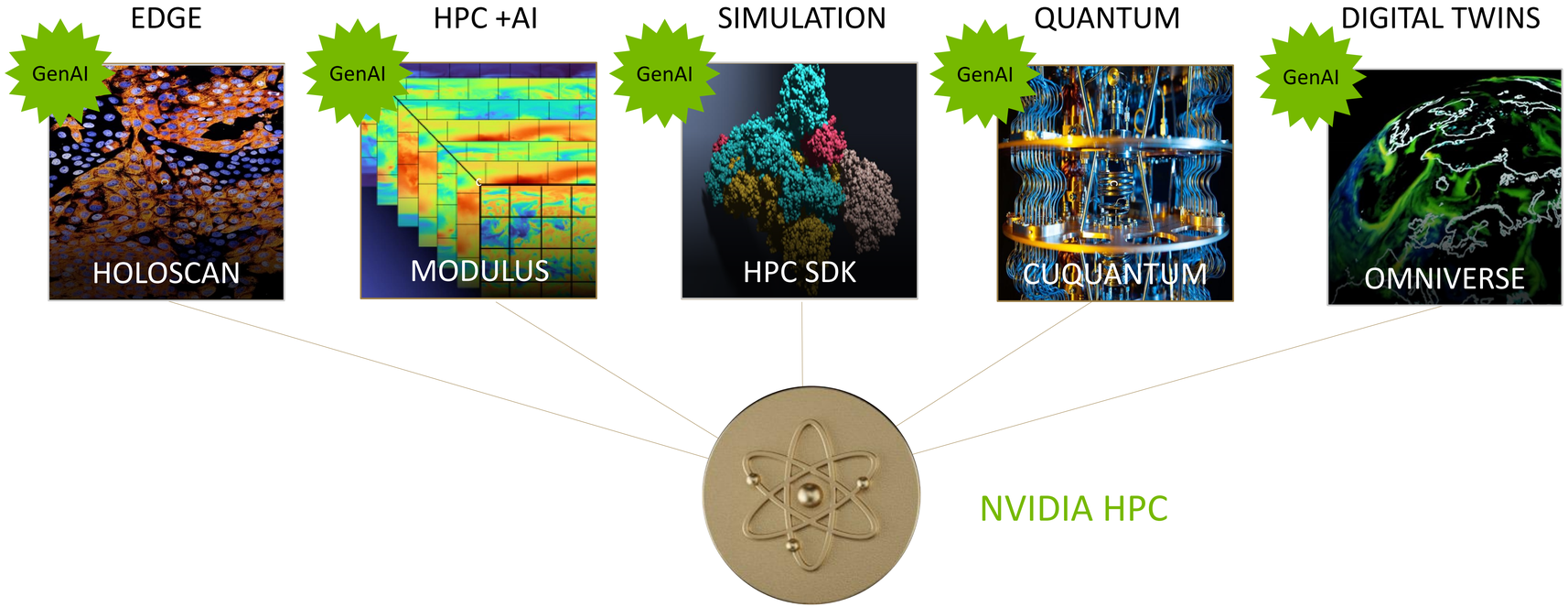


Materials : GNOME, MatterGen, MACE

2023 to present day

NVIDIA Software Platform

Five Workloads of the Modern Supercomputer



DEVELOPMENT

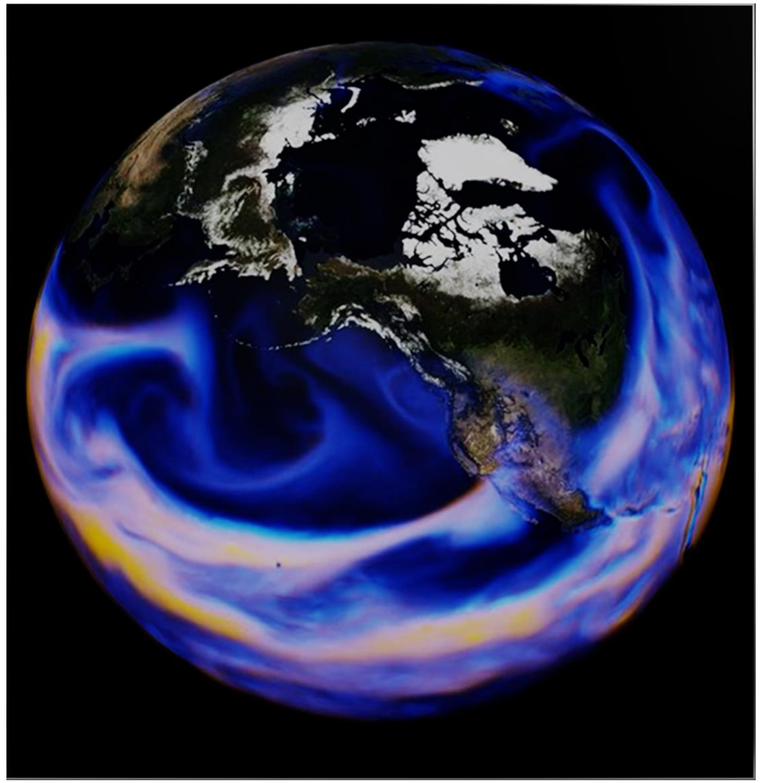
Programming Models	Compilers	Core Libraries	Math Libraries	Communication Libraries
Standard C++ & Fortran	nvcc, nvc	libcu++	cuBLAS, cuTENSOR	HPC-X (MPI, UCX, SHMEM, SHARP, HCOLL)
OpenACC & OpenMP	nvc++	Thrust	cuSPARSE, cuSOLVER	NVSHMEM
CUDA	nvfortran	CUB	cuFFT, cuRAND	NCCL

ANALYSIS

Profilers	Debugger
Nsight	cuda-gdb
Systems	Host
Compute	Device

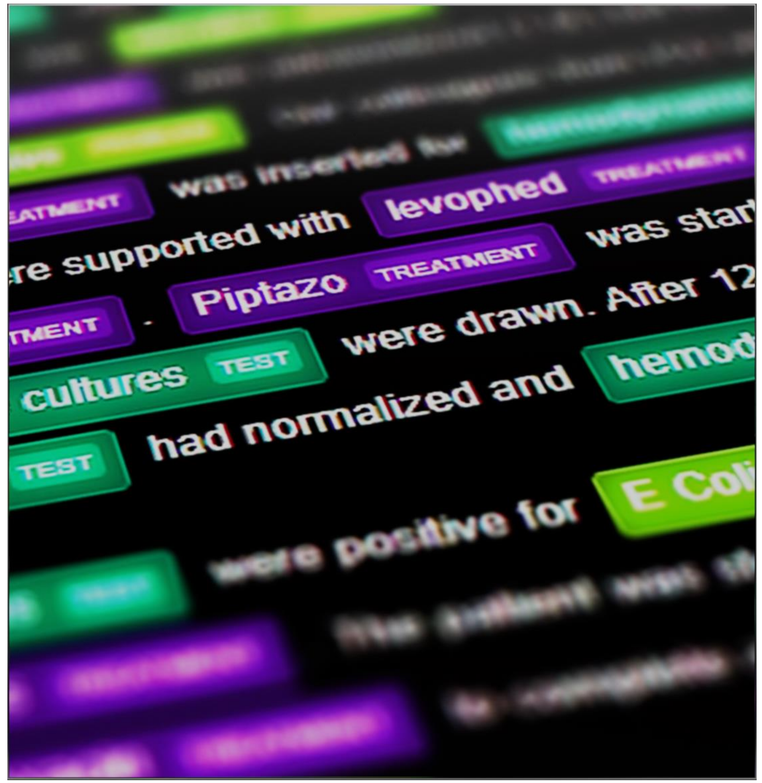
AI Frameworks for Research and Discovery

MODULUS/Earth-2
Physics-ML Model Training and Inference



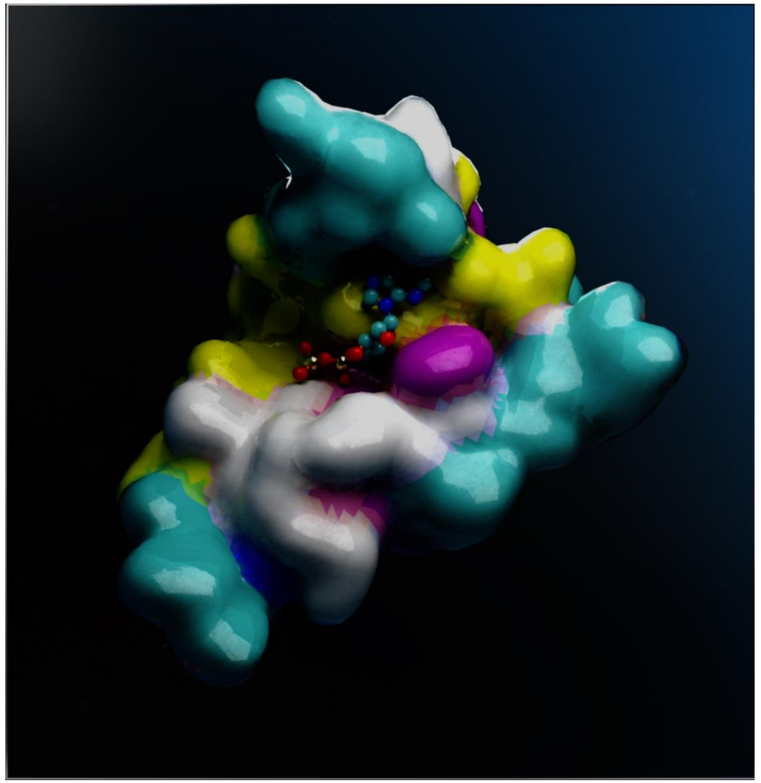
https://developer.nvidia.com/modulus
https://github.com/NVIDIA/modulus

NEMO FRAMEWORK
Developing Scientific Foundational Models at Scale



https://github.com/NVIDIA/NeMo

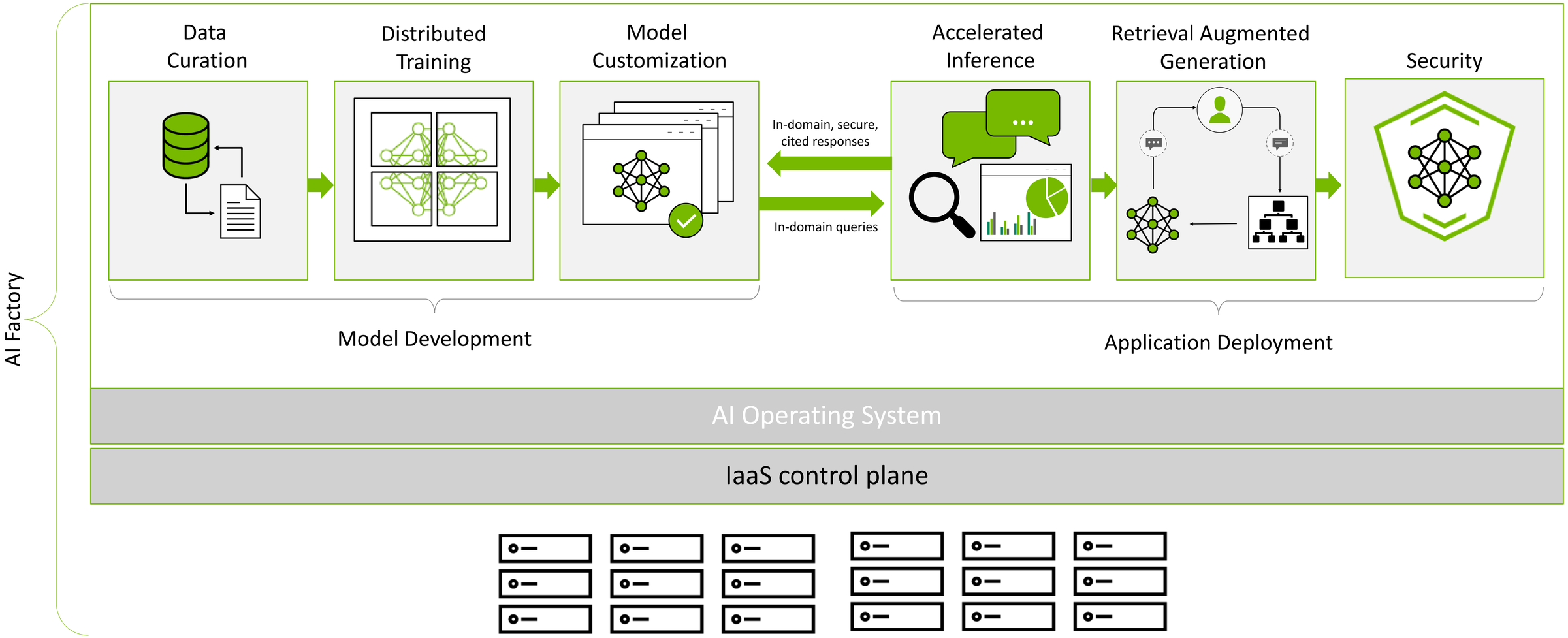
BIONEMO
AI-Driven Drug Discovery



https://www.nvidia.com/en-us/clara/bionemo/

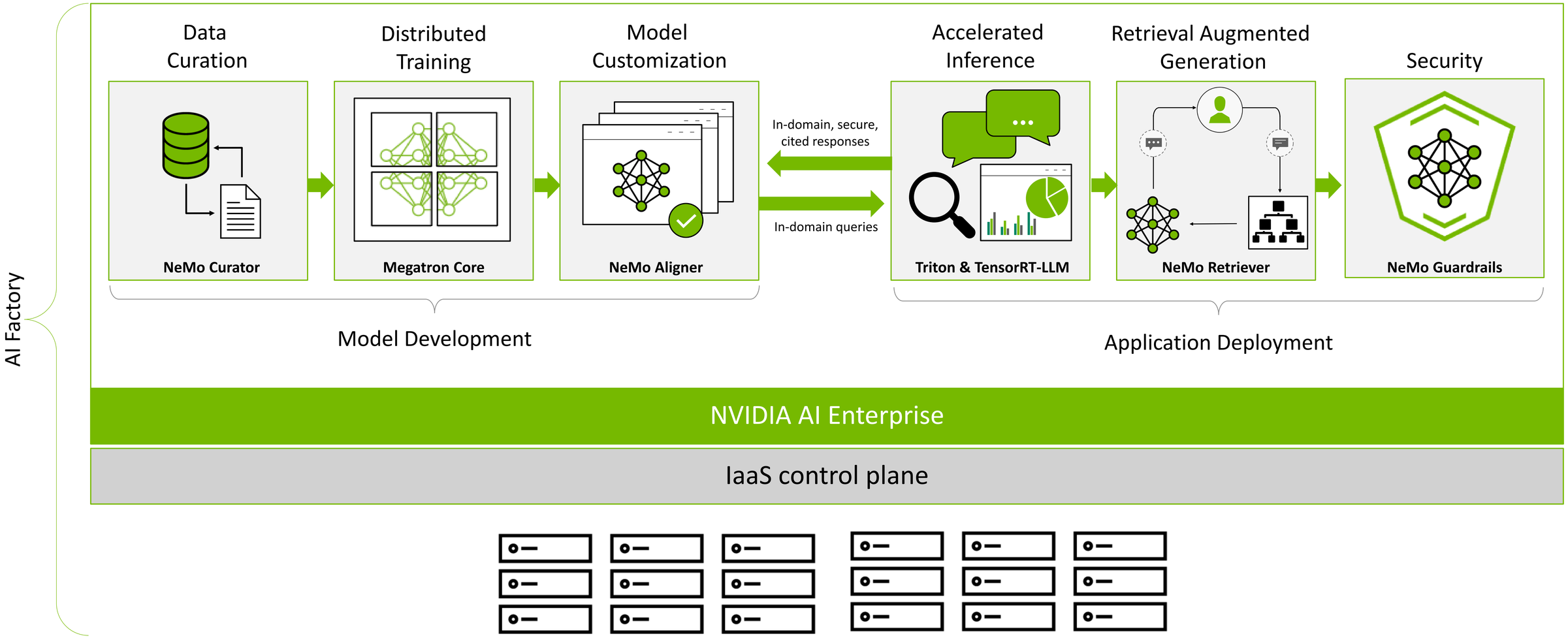
Building Generative AI Applications

Build, customize and deploy generative AI models with NVIDIA NeMo



Building Generative AI Applications

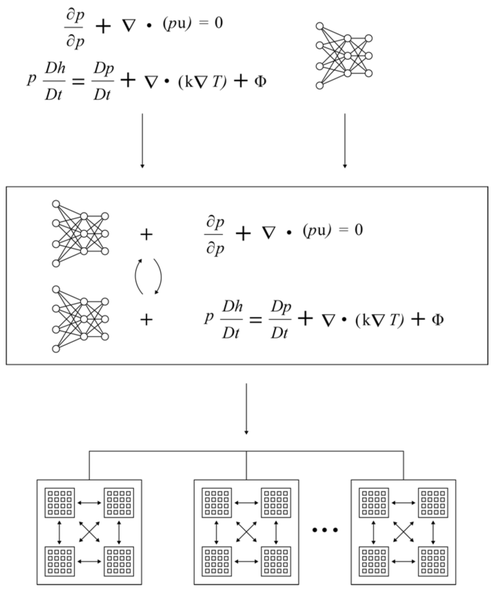
Build, customize and deploy generative AI models with NVIDIA NeMo



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

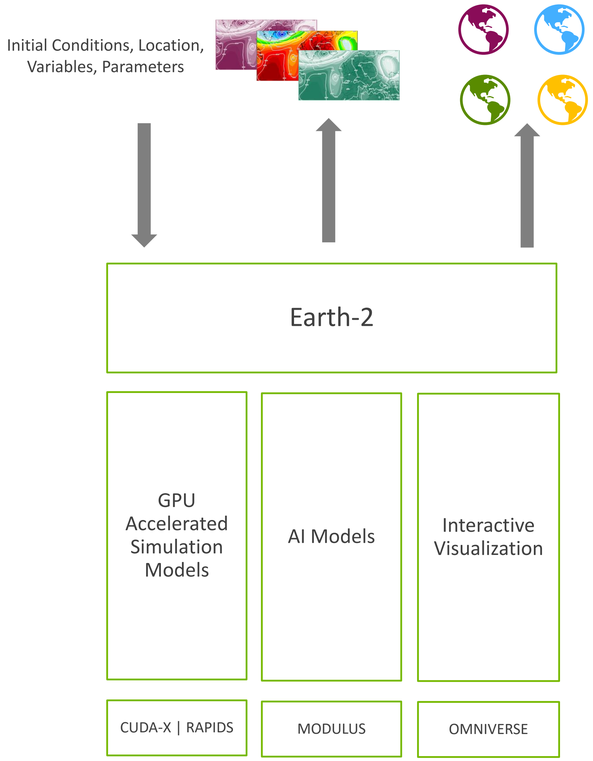
- Build physics-ml models for CFD, Heat Transfer, Structural, Electromagnetics
- Experiment and accelerate training and throughput by parallelizing the model and the training data across multi-node
- Explore physics-ml model architectures – Neural Operators, PINNs, GNNs, Diffusion

Get started today with NVIDIA Modulus

Earth-2: An NVIDIA initiative to boost climate science & climate tech

Using accelerated simulation, visualization, artificial intelligence & informatics.

Cloud Platform for to accelerate weather and climate prediction

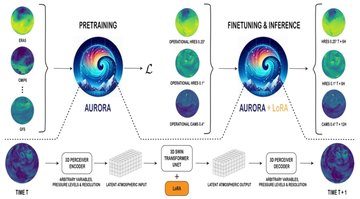


Collaborations with atmospheric modeling groups worldwide

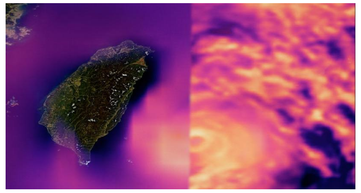
	Model	Organizations	Funding Source
Global:	E3SM, MMF, HOMEXX, SCREAM	US DOE: ORNL, SNL	E3SM, ECP
	MPAS-A	NCAR, UWyo, IBM	WACA II
	FV3 Dycore	NOAA, AI2	SENA, AI2
	NUMA/NEPTUNE	US Naval Res Lab, NPS	ONR
	IFS	ECMWF	ESCAPE, US DOE
	GungHo/LFRic	MetOffice, STFC	PSyclone
	ICON	DWD, MPI-M, CSCS, MCH	PASC ENIAC
	GEOS-5	NASA GMAO	NASA
	CLIMA/NUMA	CLiMA (NASA JPL, MIT, NPS)	Private, US NSF
Regional:	AROME	Meteo France	MF/CNRS
	COSMO	MCH, CSCS, DWD	PASC GridTools
	AceCAST-WRF	TempoQuest	Venture backed

AI for Climate/Weather

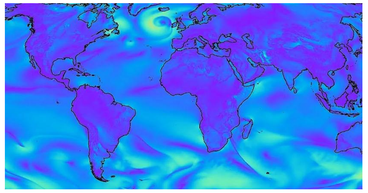
Aurora : A flexible 3D foundation model of the atmosphere
May 2024



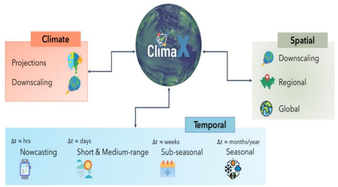
Generative Correction Diffusion Model (CorrDiff) for Km-scale Atmospheric Downscaling
Nov 2023



GraphCast: AI model for faster and more accurate global weather forecasting
Nov 2023



ClimaX, a flexible and generalizable deep learning model for weather and climate science
Jan 2023



BioNeMo Accelerates Drug Discovery with Generative AI

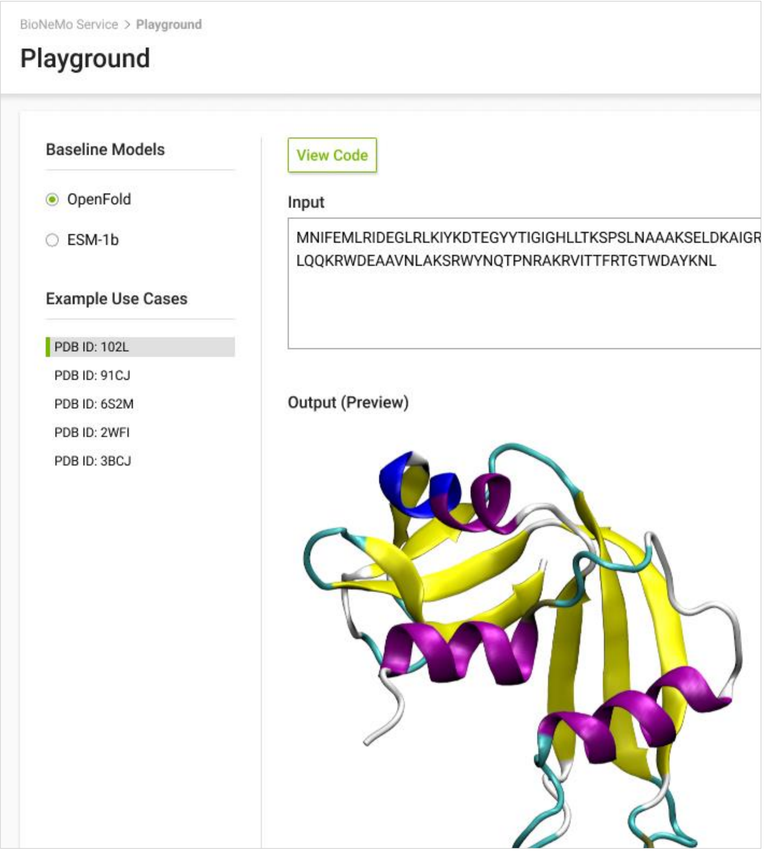
Build Foundation Models | Customize SOTA Models | Run Optimized Models

Customizable SOTA Generative AI

Innovate faster and more competitively using your proprietary datasets to train and fine-tune drug discovery workflows

Easy & Instant Access to Optimized AI

Eliminate the need for building IT infrastructure, managing open source software, optimize for throughput



Seamless & Scalable AI Microservices

Ultimate flexibility in experimenting and building enterprise grade generative AI workflows with GUI & API endpoints

Thank You