Ministry of Electronics & Information Technology

सत्यमेव जयते

सुस्वागतम्
நல்வரவு
ସୁସ୍ୱାଗତମ
సుస్వాగతం
સુસ્વાગતમ
সুস্বাগতম
ಸುಸ್ವಾಗತ
സുസ്വാഗതം
সুস্বাগতম
ਸੁਆਗਤਮ
सुस्वागतम्
خوش آمدید

NPSF User Workshop 2023

24 July – 24 Aug

Presentation On

NVIDIA Accelerated Computing

Arihenth Vijayan – Quantiphi

25th July

National PARAM Supercomputing Facility

CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING

प्रगत संगणन विकास केंद्र

सी डैक CDAC

॥ ज्ञानादेव तु कैवल्यम् ॥

NVIDIA

Agenda

- Accelerated Computing Overview
- NVIDIA DGX A100
- NVIDIA DGX SuperPOD
- NVIDIA Deep Learning Tools and Frameworks
  - Frameworks for LLM/Speech
  - Modulus(FourCastNet)
- Top 10 HPC Apps
- Case Study of DGX SuperPOD for Large Scale Workload

Evolving Nature of Workloads Driving Accelerated Computing

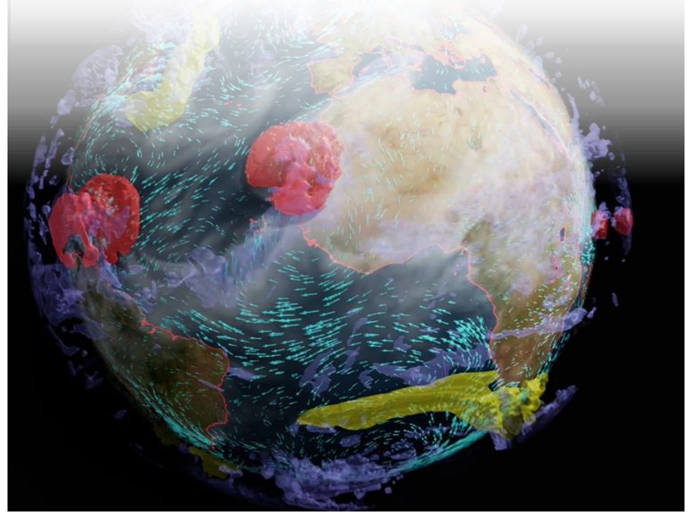
Innovation Powered by Fusion of AI and Scientific Computing (HPC) Across Every Industry

AI	Scientific Computing (HPC)

PRODUCT RECOMMENDATIONS

TEXT GENERATION

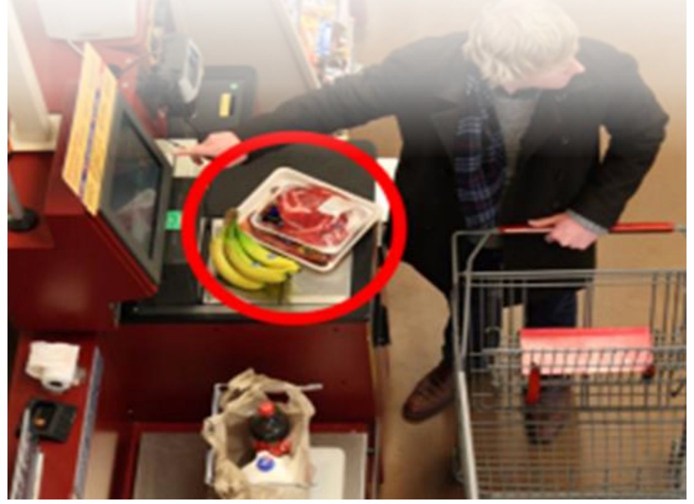
CLIMATE CHANGE



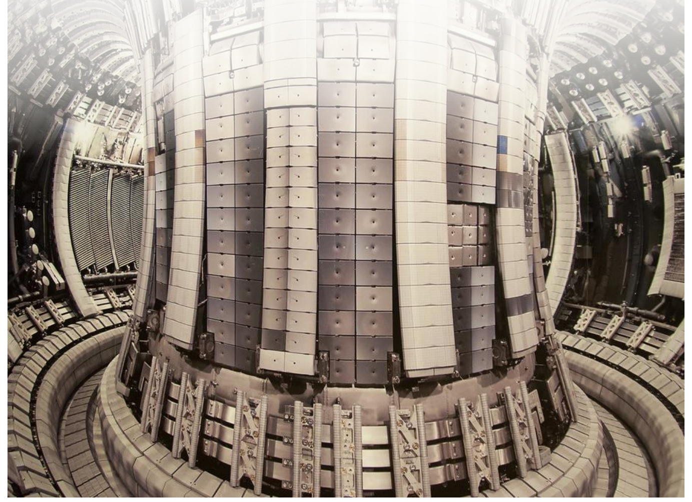
DRUG DISCOVERY

MEDICAL DIAGNOSTICS

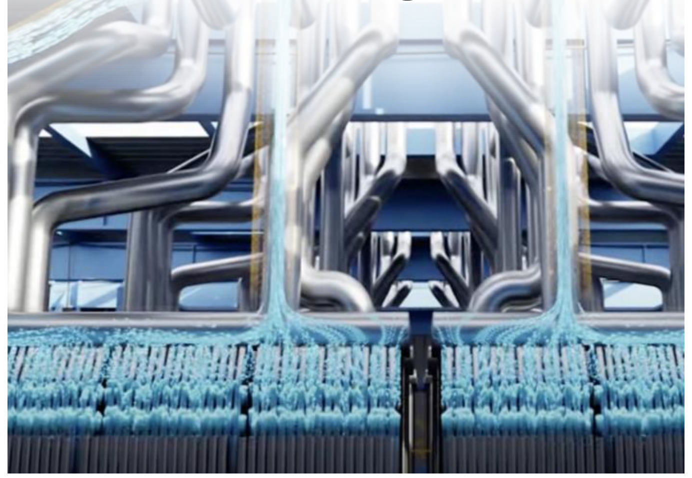
ASSET PROTECTION



RENEWABLE ENERGY

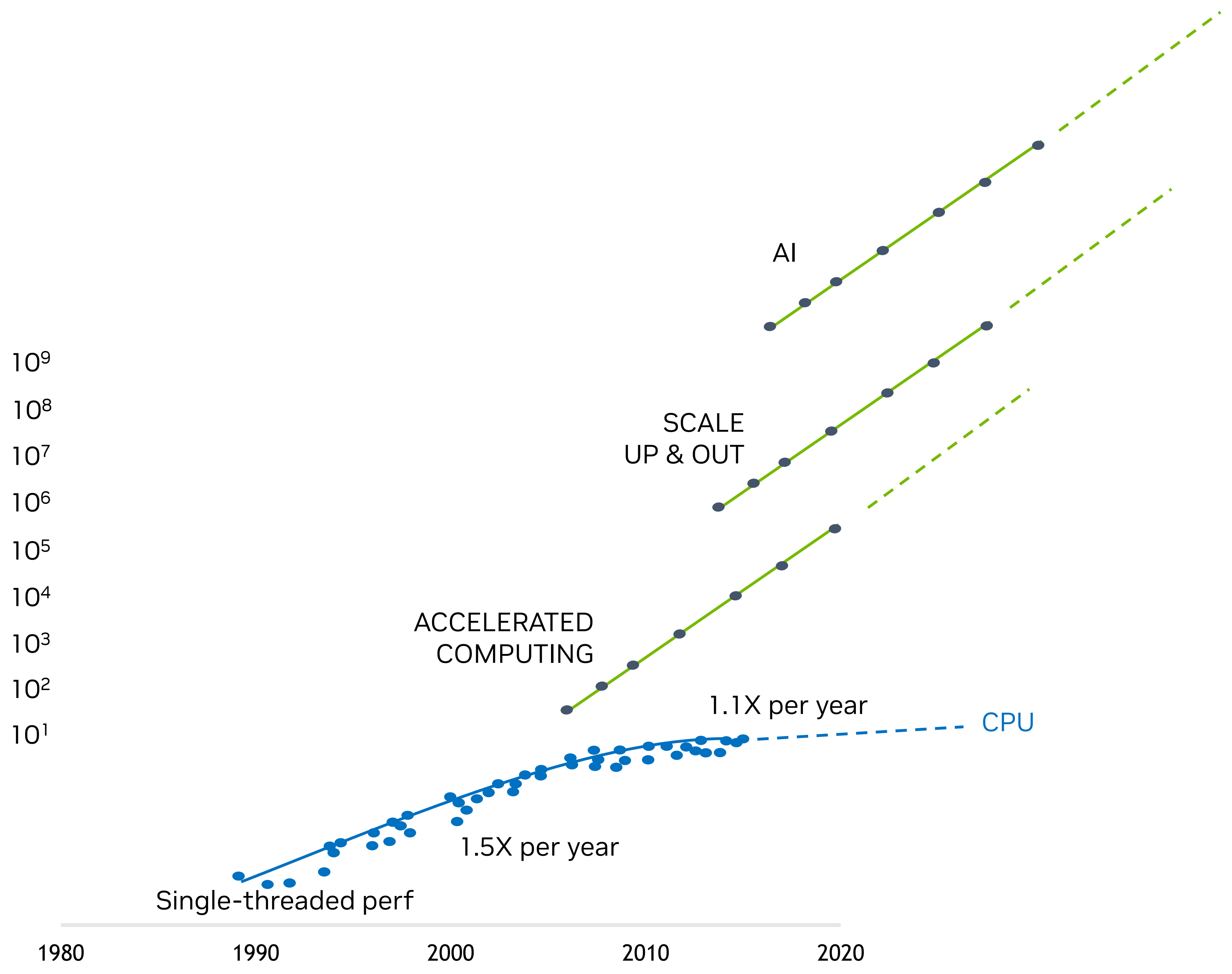


INDUSTRIAL DIGITAL TWINS



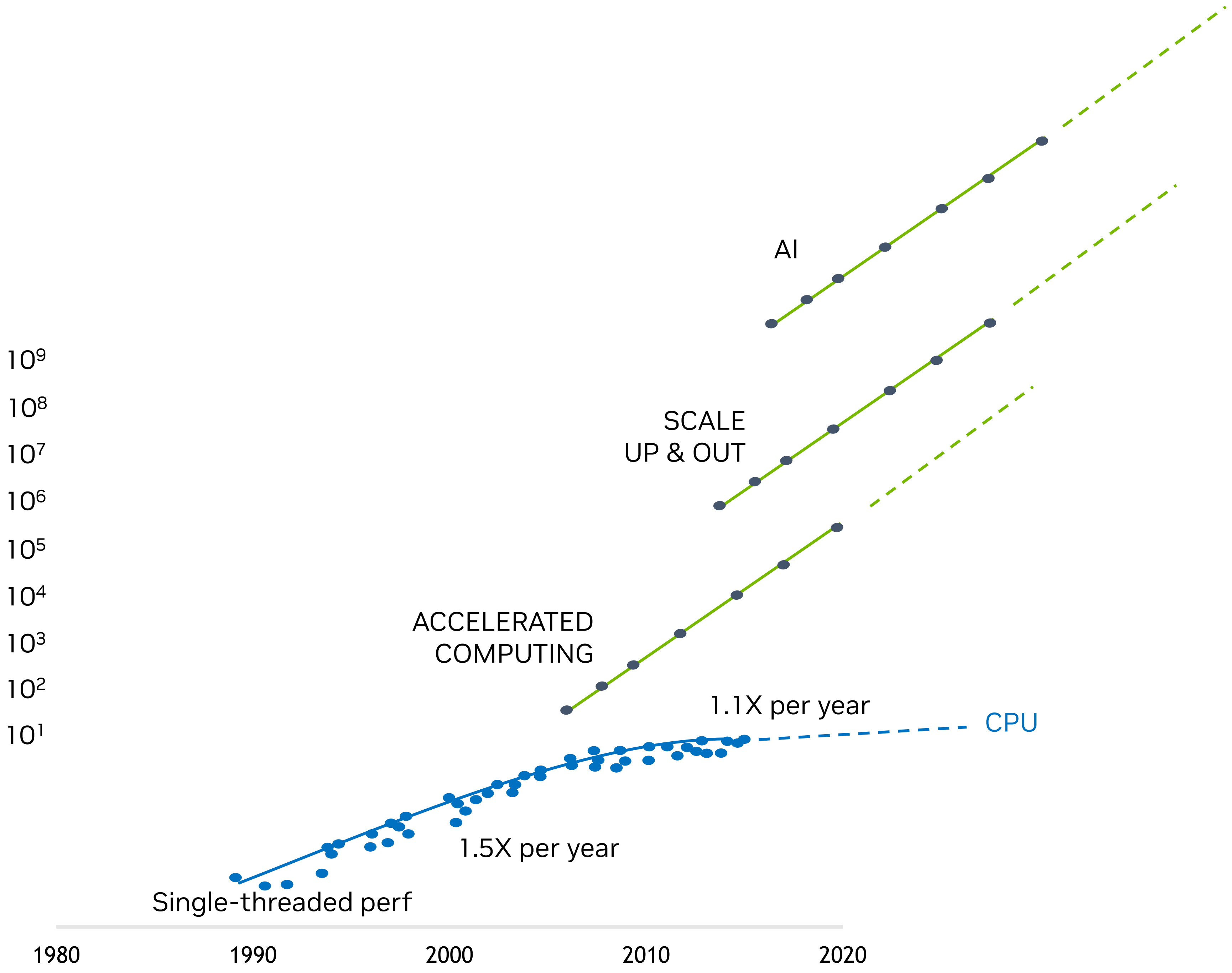
Getting Million-X Speedups to Power AI and Scientific Computing

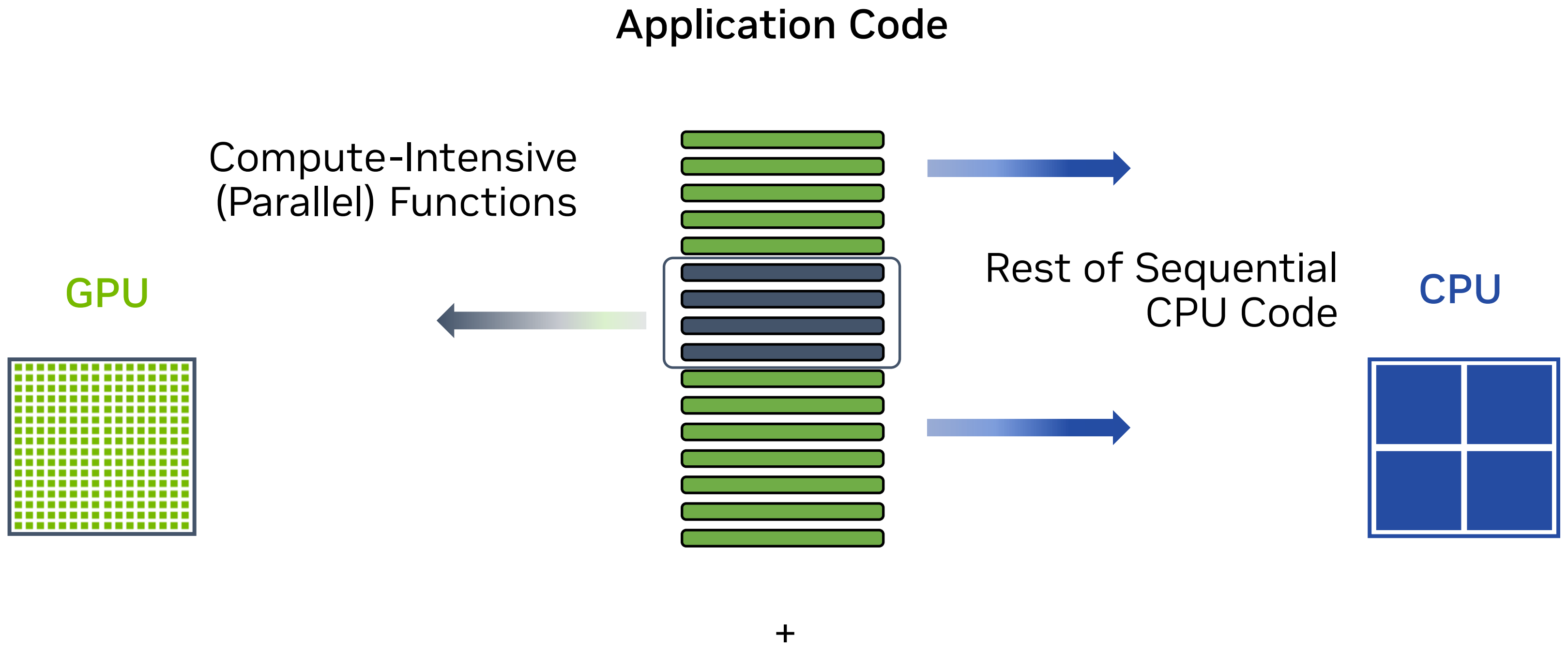
Accelerated Computing + AI Provides the Compute Required



Getting Million-X Speedups to Power AI and Scientific Computing

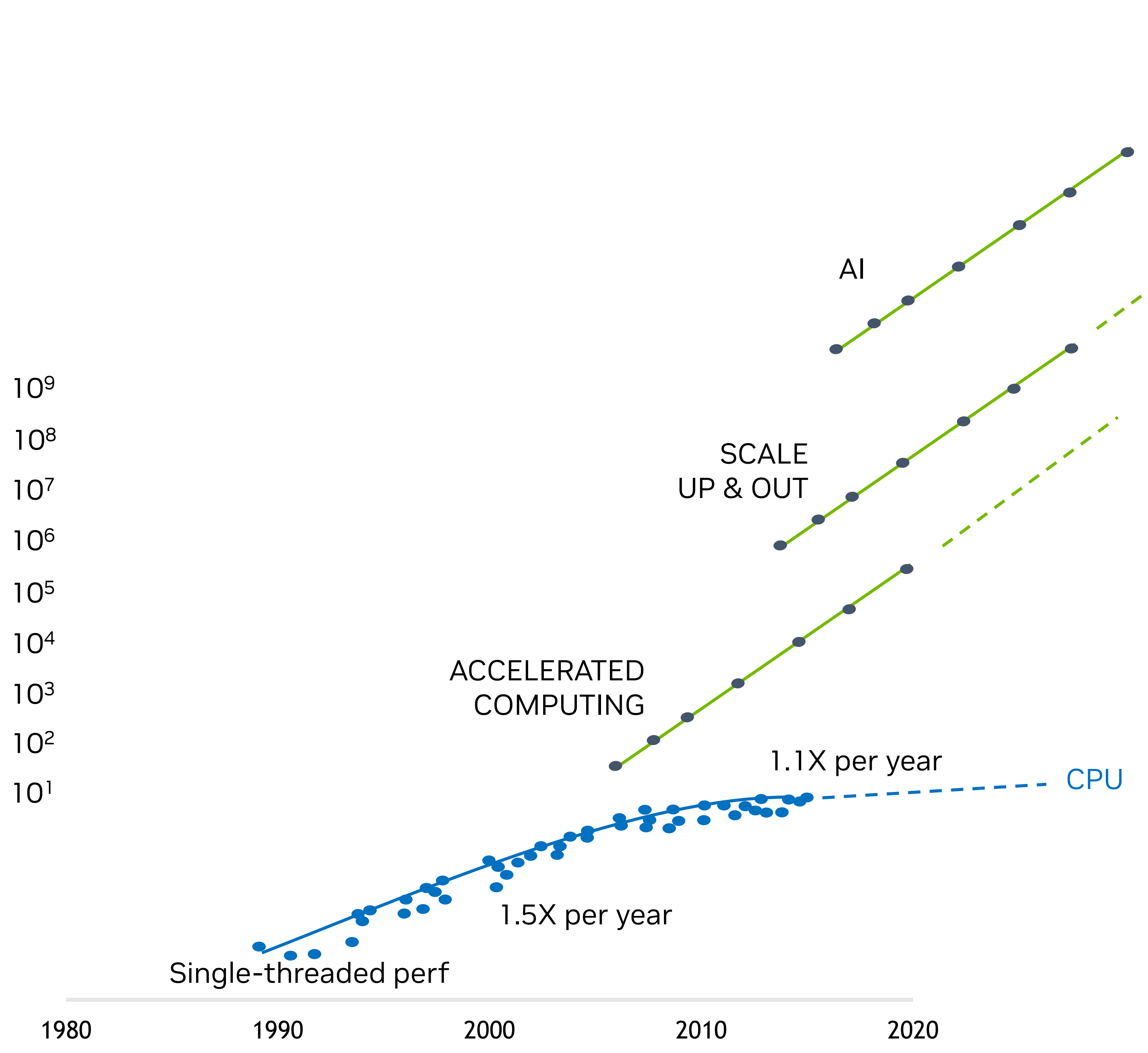
Accelerated Computing + AI Provides the Compute Required

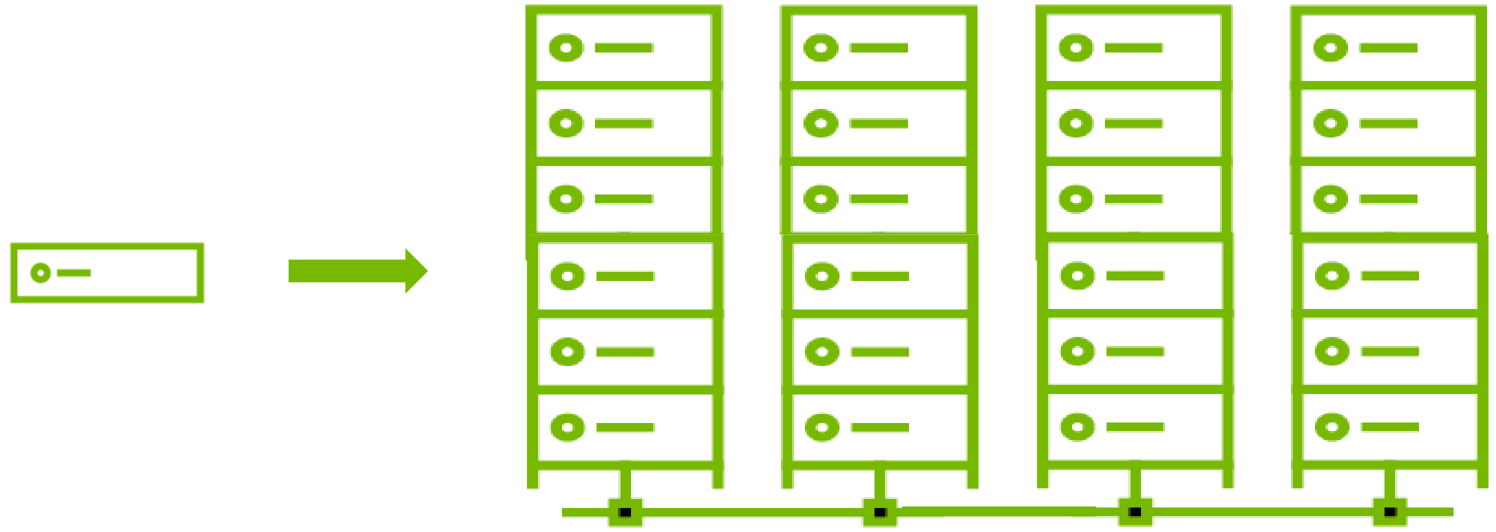




Getting Million-X Speedups to Power AI and Scientific Computing

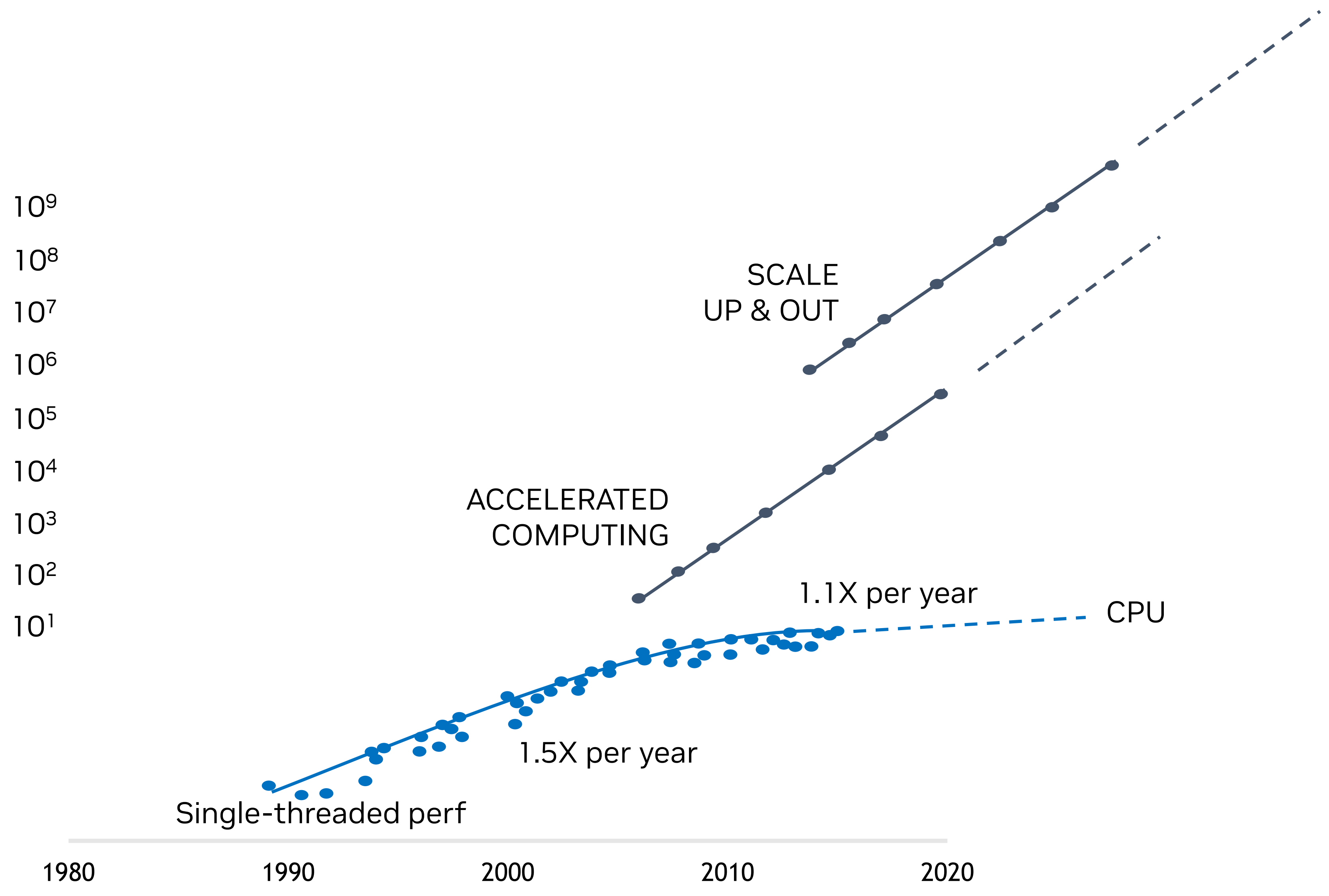
Accelerated Computing + AI Provides the Compute Required



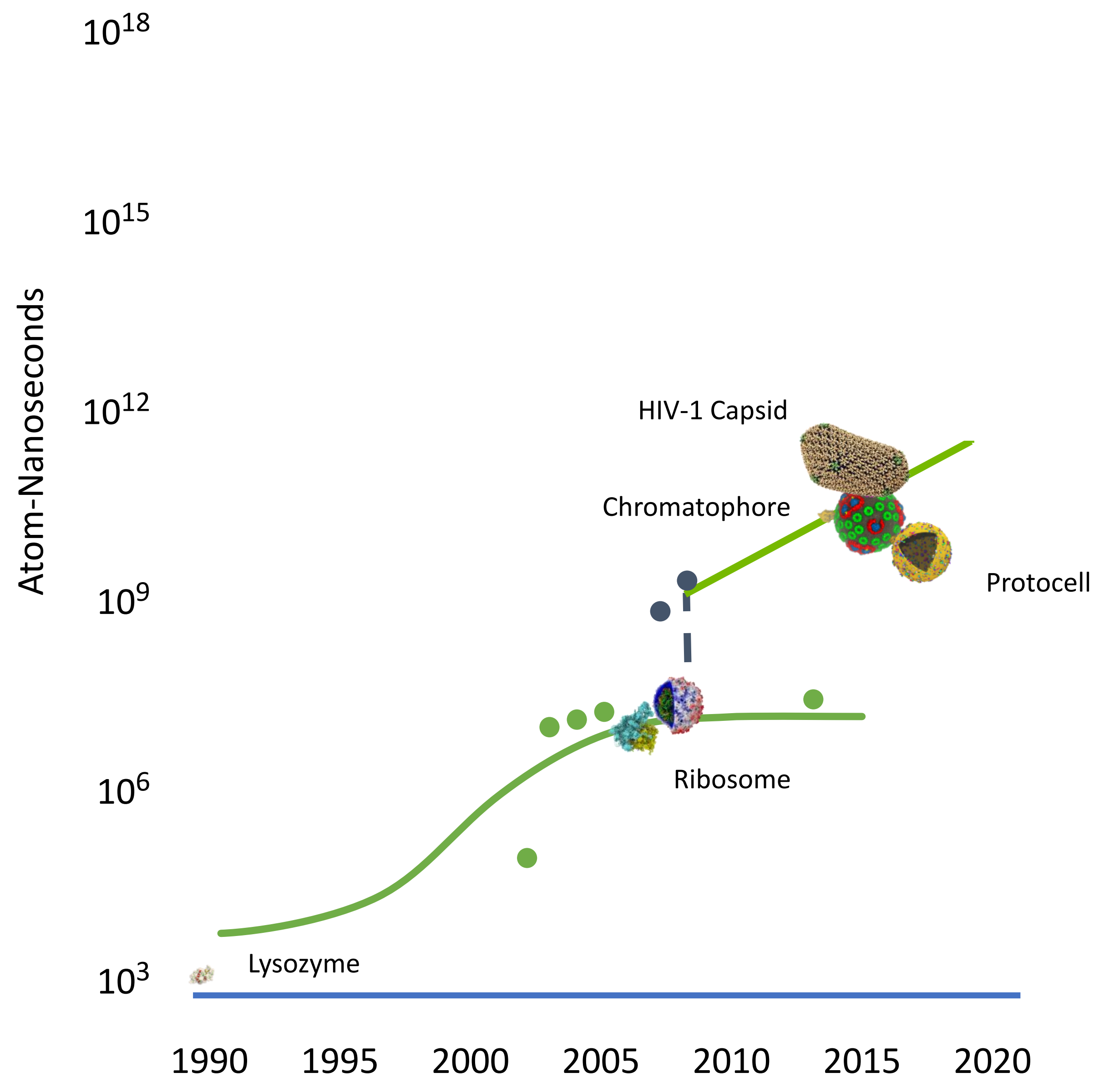


Getting Million-X Speedups to Power AI and Scientific Computing

Accelerated Computing + AI Provides the Compute Required

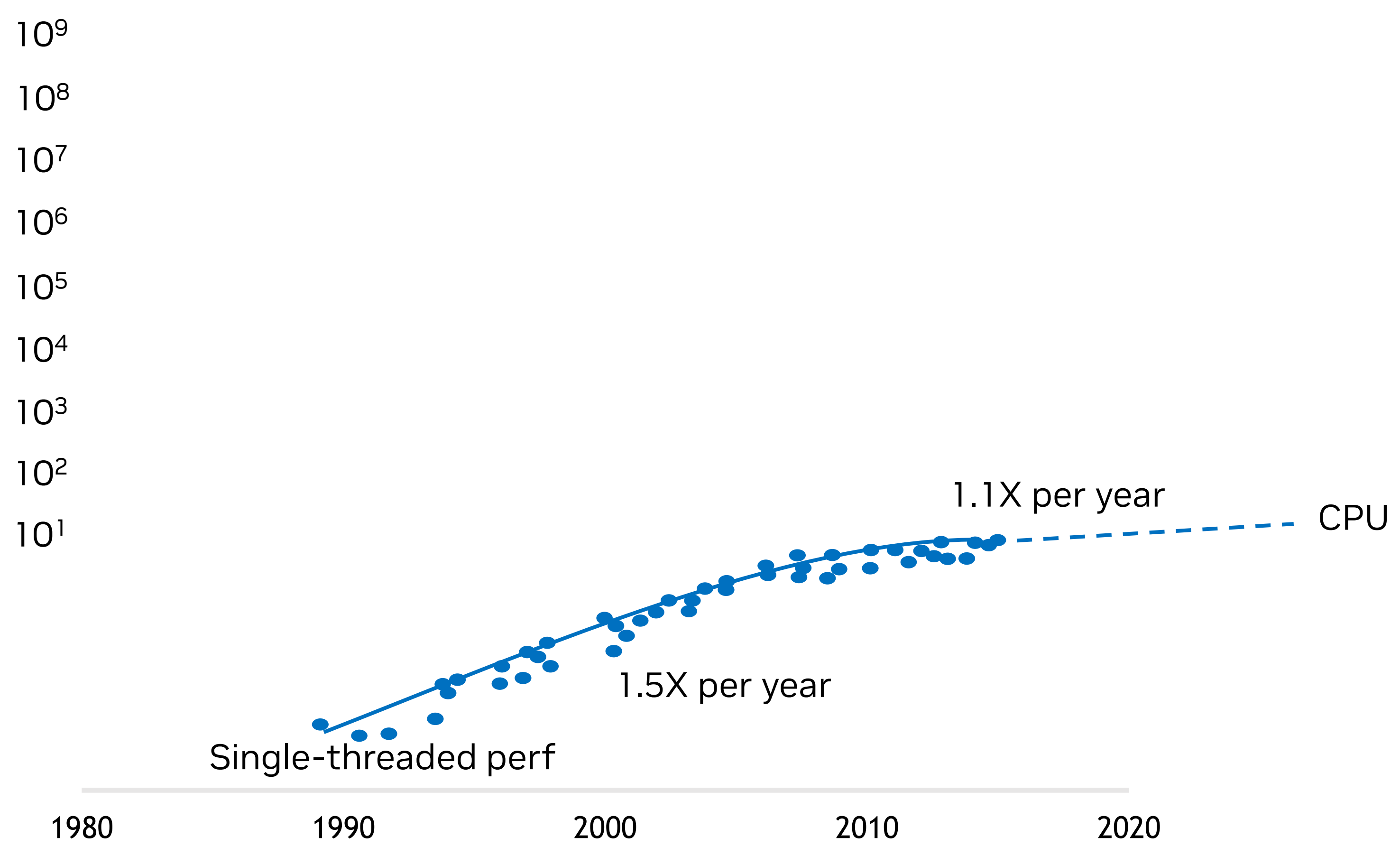


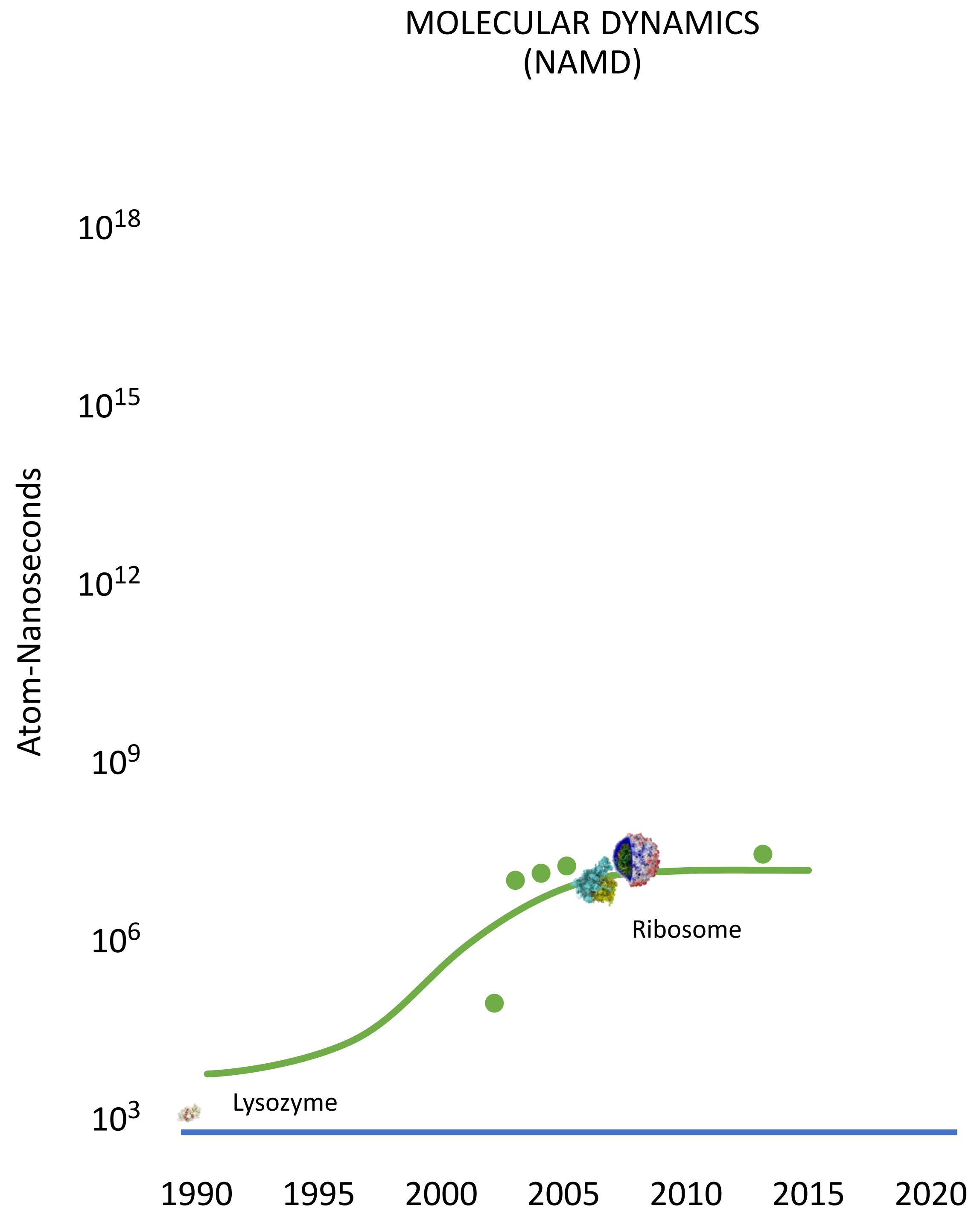
MOLECULAR DYNAMICS + GPU
(NAMD)



Getting Million-X Speedups to Power AI and Scientific Computing

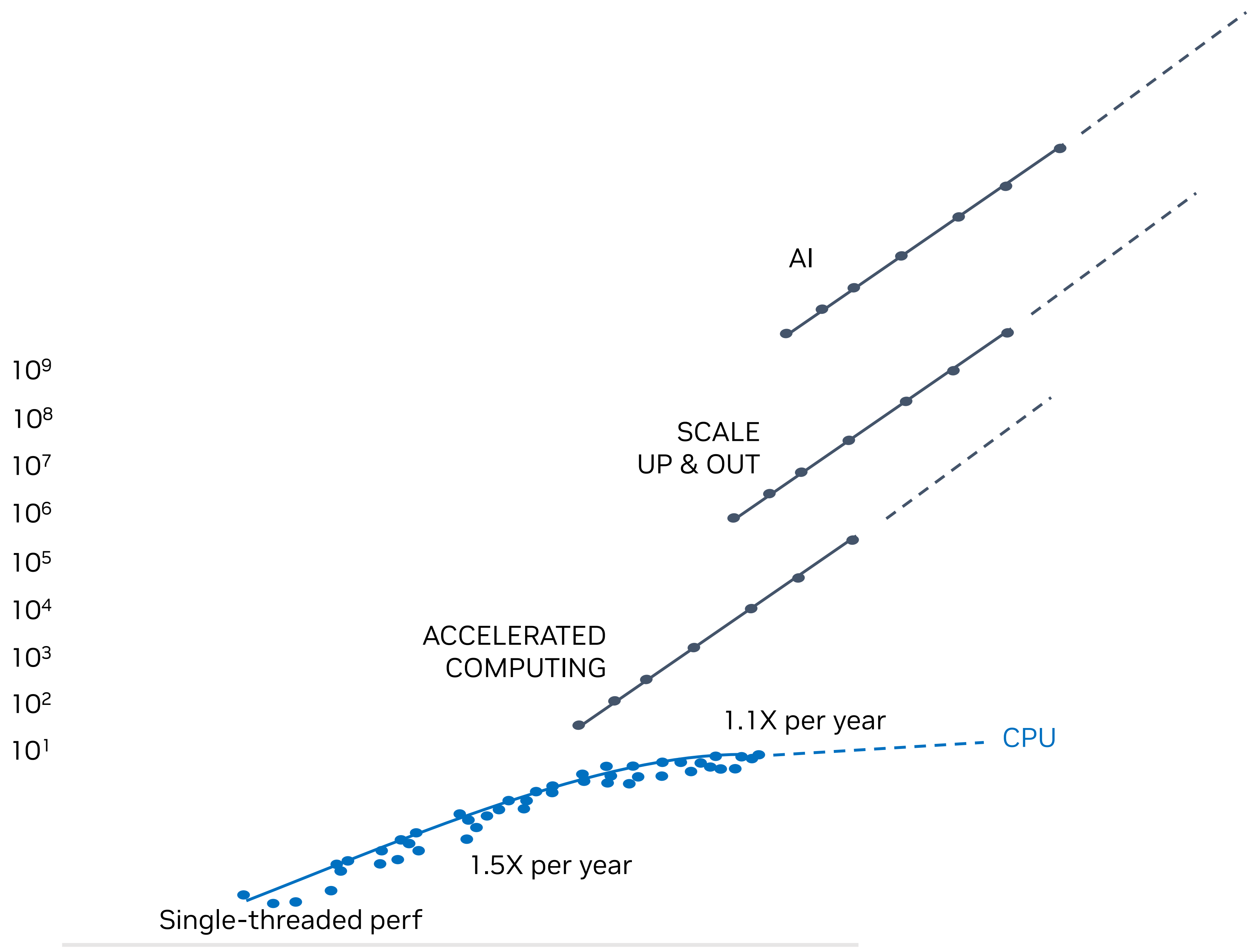
Accelerated Computing + AI Provides the Compute Required

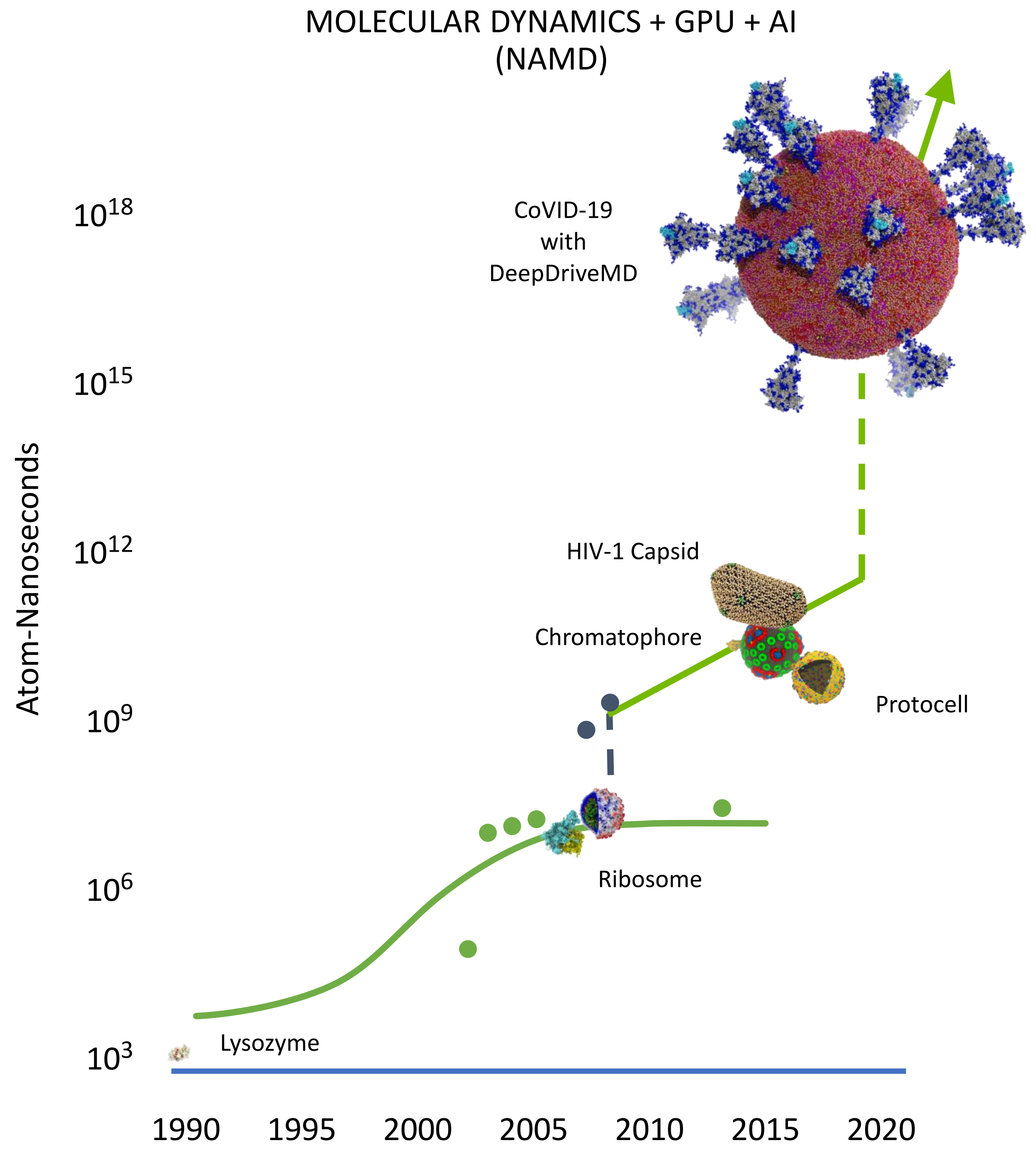




Getting Million-X Speedups to Power AI and Scientific Computing

Accelerated Computing + AI Provides the Compute Required

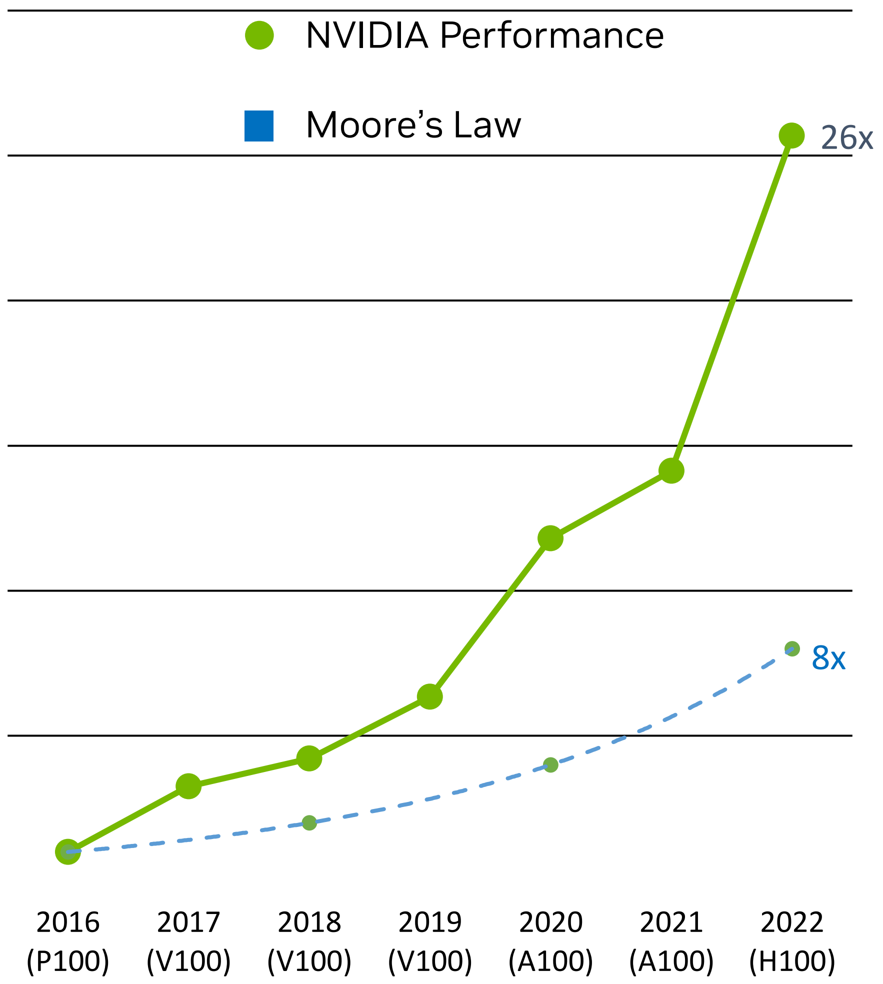




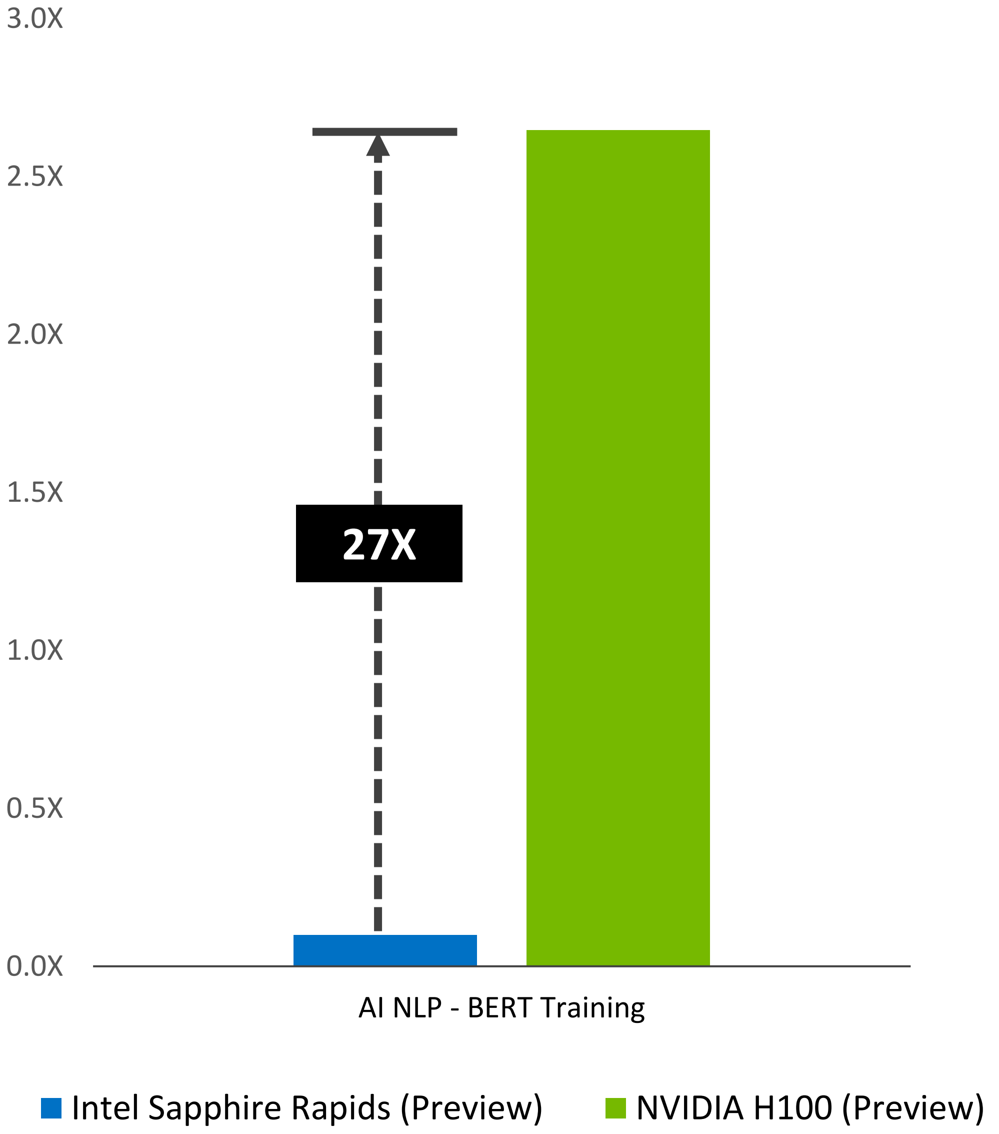
Massive Leaps in Delivered Application Performance

Accelerated Computing Significantly Outperforms Moore's Law Based CPU-Architectures

26X AI+HPC Performance in 6 Years
Relentless Full Stack Innovation



MLPerf AI Leadership
Massive Speedups on State-of-the Art AI



Driving Net Zero and Lowest TCO
Accelerated Computing is Highly Efficient

NVIDIA Hopper
Compared to CPUs

26X

more **energy efficient** at MLPerf AI Inference

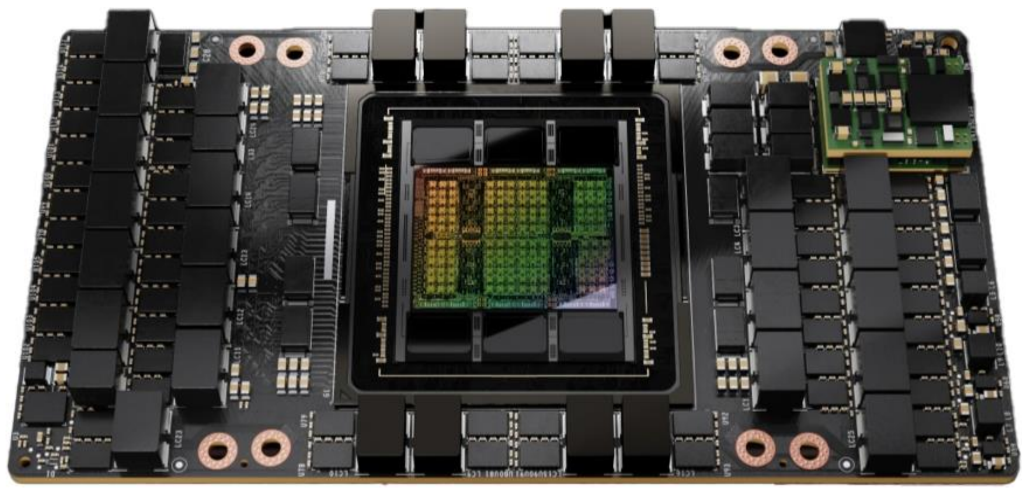
CPU		GPU
920	10X fewer nodes	92
$57M	15X lower TCO	$3.7M

Left Panel: Geometric mean of application speedups vs. P100 | 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]; TensorFlow [ResNet-50], VASP 6 [Si Huge], |GPU node: with dual-socket CPUs with 4x P100, V100, or A100 GPUs. H100 values shown for 2022 projected performance subject to change
Center Panel: Per-chip performance is not a primary metric of MLPerf™ Training. All accelerator based on 8-chip submissions and closest chip count used for Intel Sapphire Rapids results, normalized to A100 | Format: Chip count, submitter, MLPerf ID | BERT: 8x NVIDIA 2.1-2091, 16x Intel 2.1-2089 | MLPerf™ name and logo are trademarks. See www.mlperf.org. for more information.
Right Panel: Energy Efficiency based on re-production of latest commercially available A100 results and latest available CPU (Intel 8380) inference MLPerf (1.1) models. Scaling to H100 results with A100 vs H100 GPU results MLPerf (2.1) inference | Cost/Space comparison example based on latest available NVIDIA A100 GPU and Intel CPU inference results in the commercially available category of the MLPerf (1.1) industry benchmark

NVIDIA GPU DPU and CPU Drive Full Stack Performance

State-of-the-Art Hardware Portfolio and Relentless Software Execution

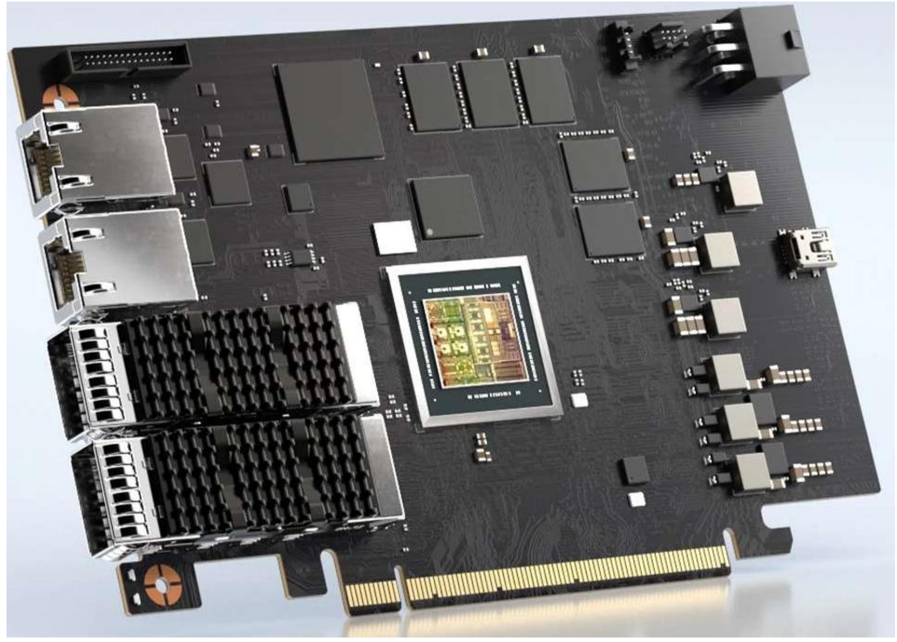
GPU
Accelerate Compute Intensive Functions



Compute Intensive Functions

- AI
- Scientific Computing
- Data Analytics

DPU
Network Infrastructure Acceleration



Infrastructure Offload

- Storage
- Security
- Networking

CPU
Orchestration with Direct Interconnect

Orchestration Management

- Management
- Data I/O

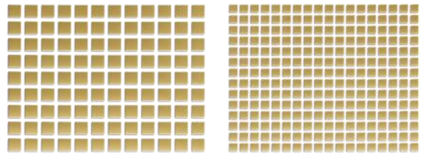

MODULUS
MONAI
RIVA
MAXINE
NEMO
MERLIN
CUOPT
MORPHEUS
TOKKIO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN
AI APPLICATION FRAMEWORK
PLATFORMS
NVIDIA SCIENTIFIC COMPUTING (HPC)
NVIDIA AI
NVIDIA Omniverse
ACCELERATION LIBRARIES
cuNumeric
CV-CUDA
cuQuantum
Parabricks
Sionna
JetPack
RAPIDS
Spark
cuDNN
cuGraph
TensorRT
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

NVIDIA DGX A100

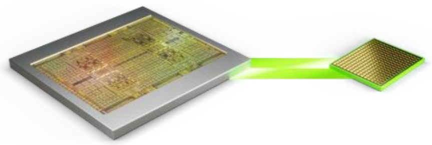
DGX A100

GAME-CHANGING PERFORMANCE FOR INNOVATORS

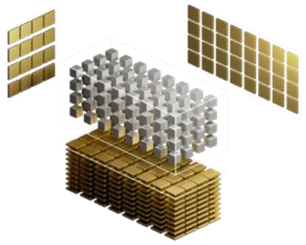
SUPERCHARGING HIGHEST PERFORMING AI SUPERCOMPUTING GPU



80GB HBM2e
For largest
datasets and models

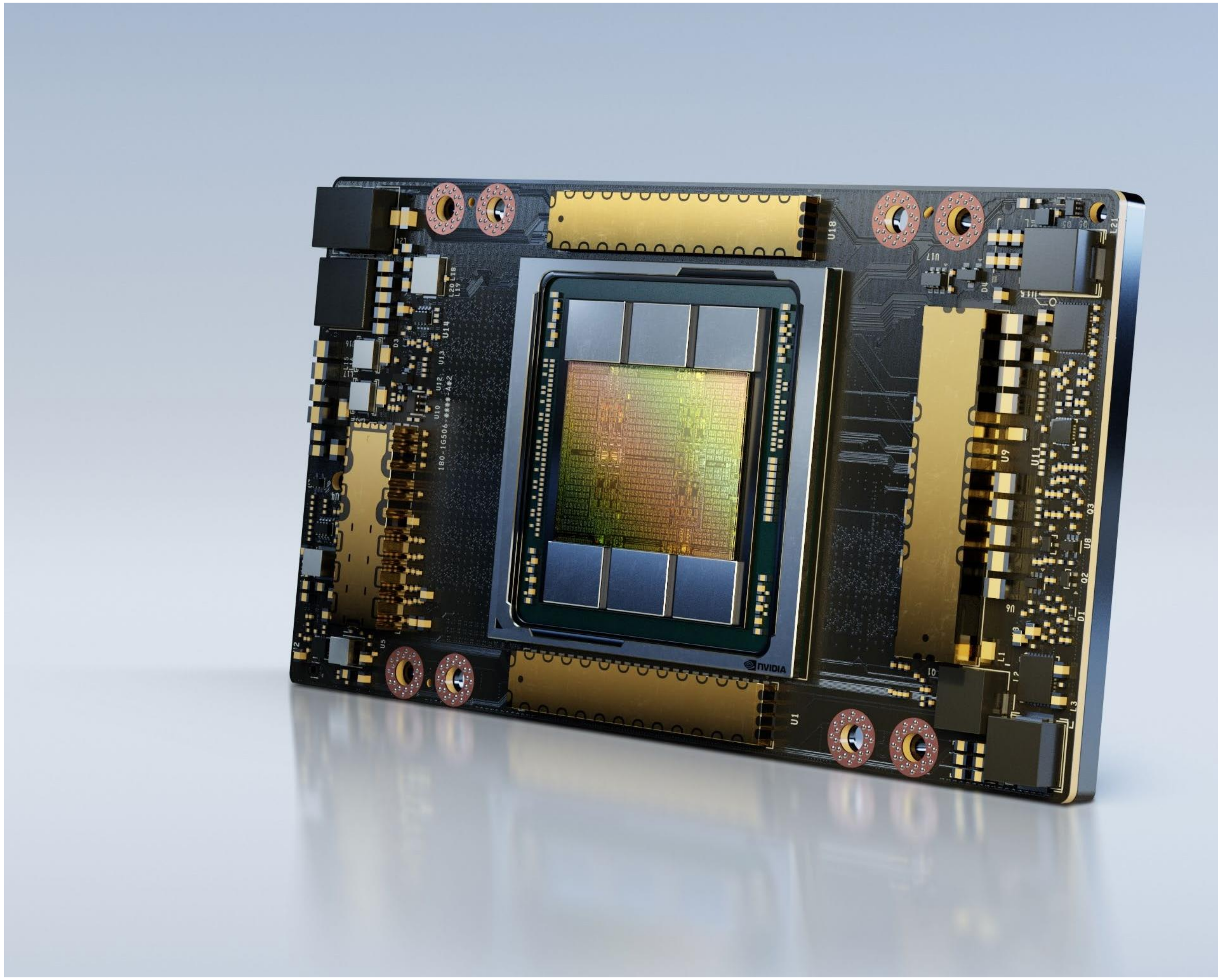


2TB/s +
World's highest memory
bandwidth to feed the world's
fastest GPU



3rd Gen Tensor Core

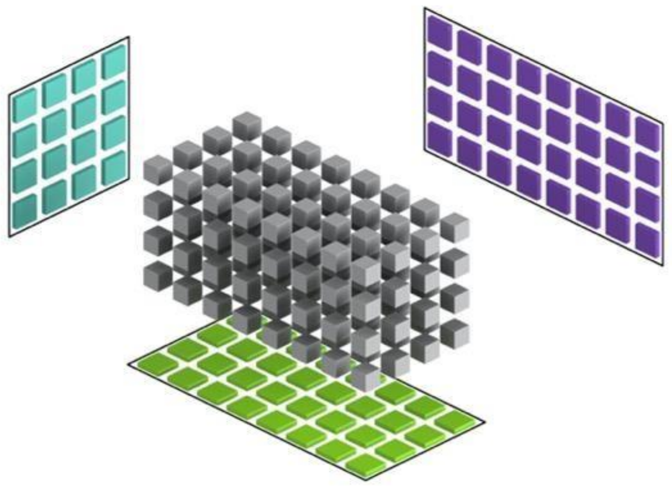
Multi-Instance GPU



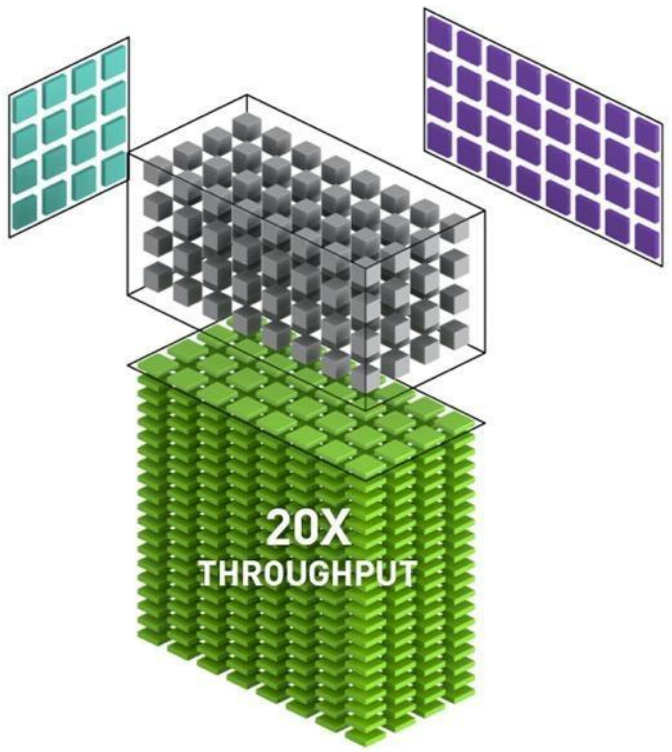
DGX A100

TF32 TENSOR CORES : 20X Higher FLOPS for AI, Zero Code Change

NVIDIA V100 FP32



NVIDIA A100 Tensor Core TF32 with Sparsity



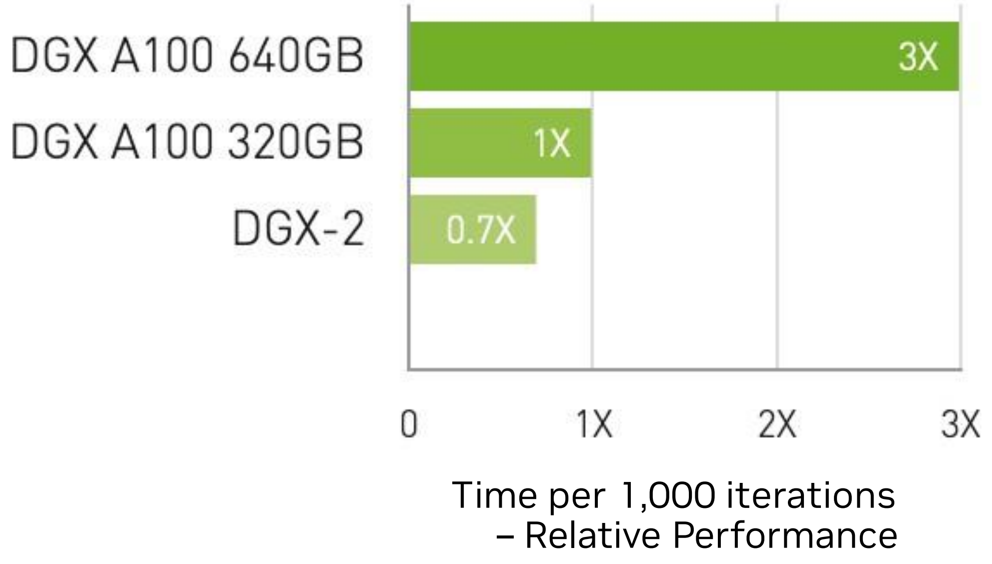
20X Faster than Volta FP32 | Works like FP32 for AI with Range of FP32 and Precision of FP16
No Code Change Required for End Users | Supported on PyTorch, TensorFlow and MXNet Frameworks Containers

DGX A100 PERFORMANCE

Up to 3X Higher Throughput on DGX A100 640GB

DLRM Training

Up to 3X Higher Throughput
for AI Training on Largest Models

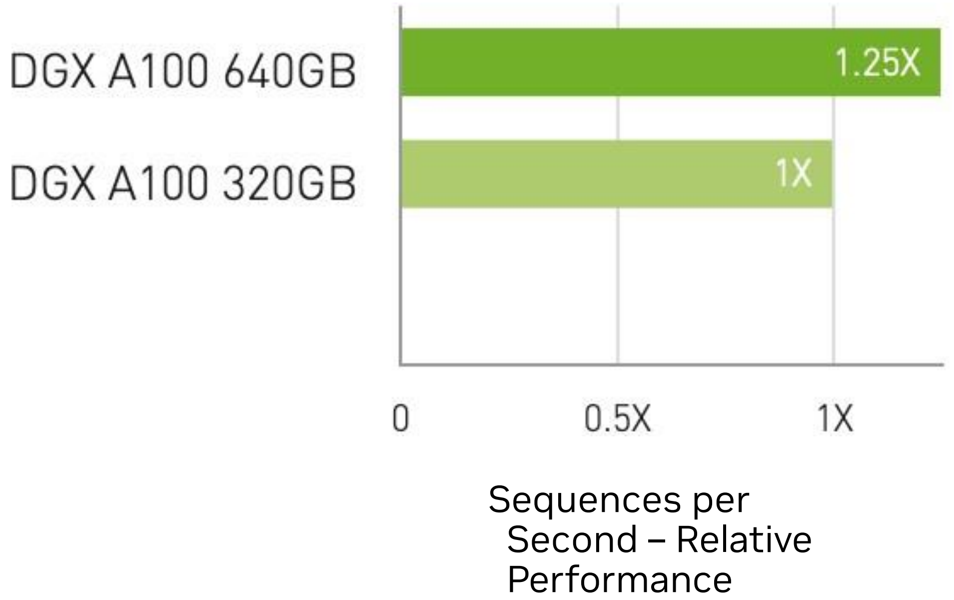


Large Model Training

DLRM (Huge CTR framework), FP16 precision | 1x DGX A100 640GB batch size = 48 | 2x DGX A100 320GB batch size = 32 | 1x DGX-2 (16x V100 32GB) batch size = 32. Speedups normalized to number of GPUs

RNN-T Inference

Up to 1.25X Higher Throughput
for AI Inference

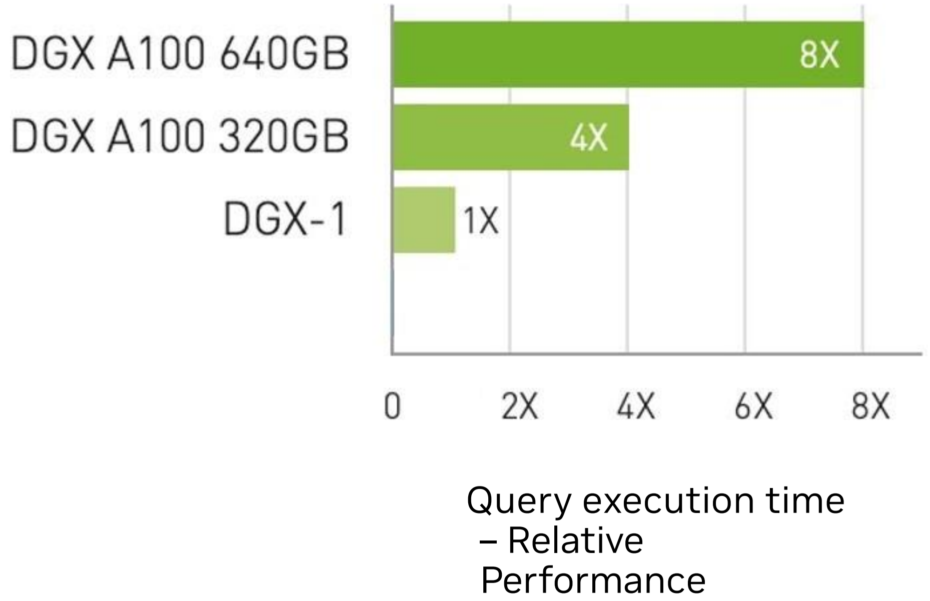


Inference on MIG

MLPerf 0.7 Single stream latency, RNN-T measured with [1/7] MIG slices. Framework: TensorRT 7.2, dataset = LibriSpeech, FP16 precision

Big Data Analytics

2X Faster Query Execution
30 Queries on 1TB dataset



Analyzing Datasets

GPU-BDB | 30 analytical retail queries, ETL, ML, NLP on 1TB dataset 1x DGX-1 V100 256GB | 1x DGX A100 320GB | 1x DGX A100 640GB RAPIDS 0.19, Dask 2021.03.1, UCX 1.9

NVIDIA DGX SuperPOD

NVIDIA DGX SUPERPOD

Next generation AI supercomputing infrastructure

The DGX SuperPod is designed to minimize system bottlenecks and maximize performance for the diverse nature of AI and HPC workloads. It provides:

- A modular architecture constructed from Scalable Units.
- A hardware and software infrastructure built around the DGX SuperPod
- The ability to quickly deploy and update the system.
- Management and monitoring services configured for High Availability (HA).

DGX SUPERPOD REFERENCE ARCHITECTURE

Codesigned by DL scientists, application performance engineers and system architects

Compute
Powerful nodes each with 8 NVIDIA A100 GPUs, a large memory footprint, and NVLink / NVSwitch based fast connections between the GPUs for computing to support the variety of DL models in use.

Storage
A storage hierarchy that can provide maximum performance for the needs of various dataset structures.

Network
A low-latency, high-bandwidth, HDR InfiniBand interconnect designed with the capacity and topology to minimize bottlenecks.

The basic unit of SuperPod is a Scalable Unit (SU) with 20 DGX A100 nodes, InfiniBand networking components and storage

NVIDIA Deep Learning Frameworks and Tools

NVIDIA COMPUTING PLATFORM

Frameworks and Tools

BUSINESS APPLICATIONS
- Customer Engagement
- Patient Diagnostics
- Fraud Detection
- Quality Assurance
- Industrial Automation
- Precision Marketing
- Molecular Simulations
- ++

NGC

SOFTWARE HUB
- HELM
- *Certified Containers*
- *Pre-trained Models*
- *SDKs*

APPLICATION FRAMEWORKS
- SMART CITY – *Metropolis*
- CONVERSATIONAL AI – *Riva*
- AUTONOMOUS VEHICLES – *Drive*
- RECOMMENDATION SYSTEMS – *Merlin*
- HEALTHCARE – *Clara*
- ++ – ...

DEVELOPER TOOLKITS
- ML & DATA ANALYTICS – RAPIDS, dmlc XGBoost
- AI TRAINING & INFERENCE – TensorFlow, TensorRT, PYTORCH, mxnet
- HIGH PERFORMANCE COMPUTING – NVIDIA HPC SDK
- RENDERING & VISUALIZATION – IndeX, OptiX

ACCELERATION LIBRARIES
- COMPUTE – *CUDA-X*
- NETWORKING, STORAGE & SECURITY – *DOCA*, *MAGNUM IO*

OPERATIONS
- TRITON INFERENCE SERVER
- FLEET COMMAND
- NVIDIA GPU Operator
- Red Hat
- vmware

NVIDIA CERTIFIED

VALIDATED SOLUTIONS

SERVERS & CLOUD
- DGX, HGX – Purpose Built
- EGX – Mainstream & Edge
- CSP Instances

HARDWARE TECHNOLOGIES
- GPU
- NVSwitch
- BlueField DPU
- SMART NIC
- Mellanox Switch

MONITORING
- DCGM
- Grafana
- Prometheus
- UFM

COMPONENTS OF A TYPICAL AI PIPELINE

Development

Deployment

Data Processing → **Model Training** → **Model Optimization** → **Private Model Repository** → **Inference Serving** ⇄ **AI Applications**

Useful NVIDIA SDKs

- NVIDIA DALI
- NVIDIA TAO
- NVIDIA NEMO
- NVIDIA NGC
- NVIDIA MEGATRON-LM
- NVIDIA RIVA
- Local/Cloud Storage
- NVIDIA DEEPSTREAM
- NVIDIA RIVA
- NVIDIA TRITON INFERENCE SERVER
- TensorRT

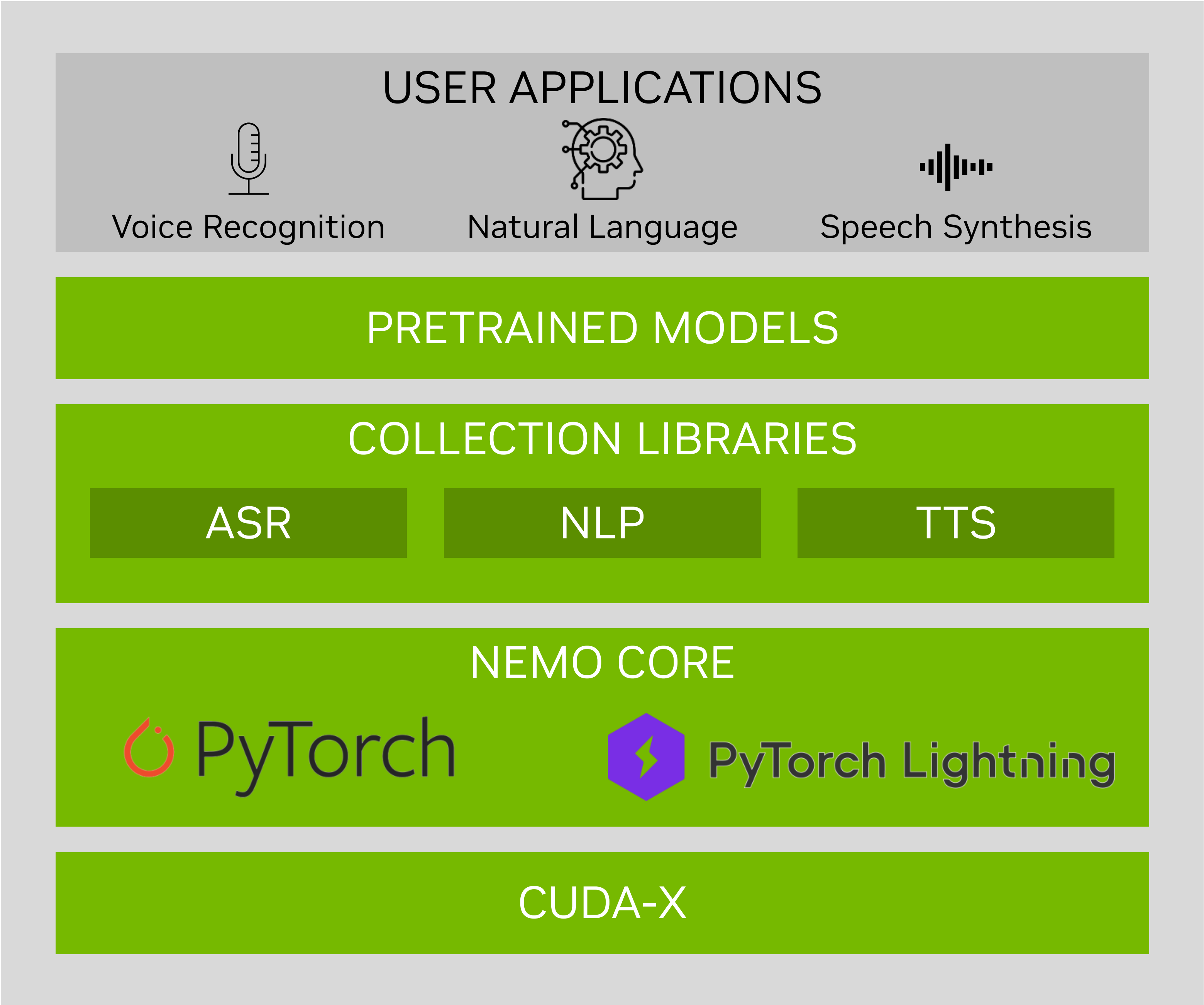
Personas Involved

- Data Engineer
- Data Scientist
- ML Engineer
- MLOps Engineer
- DevOps Engineer
- Application Developer

LLM and Speech Tools/Frameworks

Toolkit for Building SOTA Conversational Models - NVIDIA NeMo Framework / Toolkit

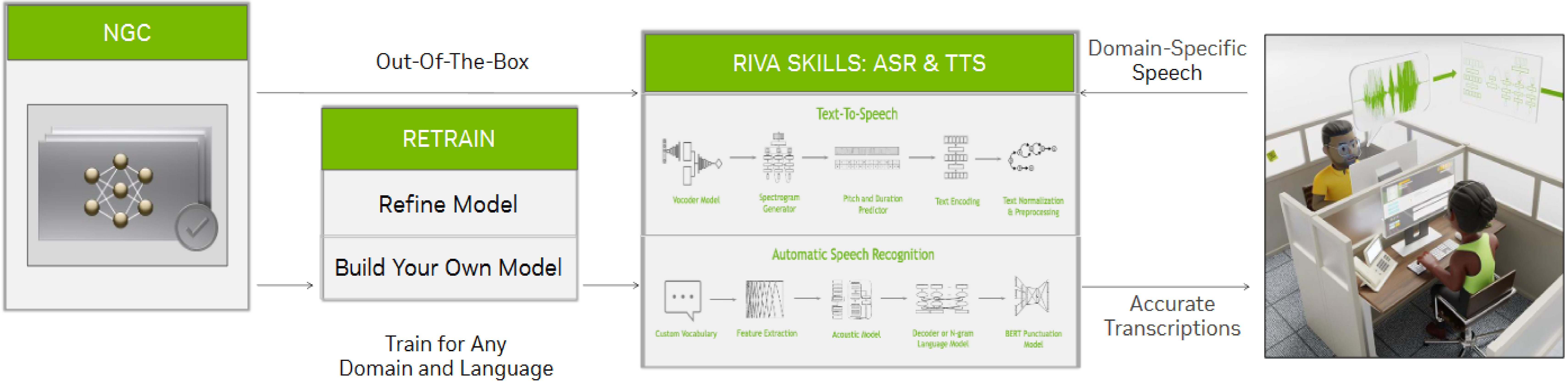
- NVIDIA NeMo™ is an end-to-end cloud-native enterprise framework for developers to build, customize, and deploy generative AI models with billions of parameters.
- Toolkit/Framework for Conversational AI
  - Speech
    - ASR
    - TTS
  - Large Language Models (LLM)
  - Natural Language Processing (NLP)
- Support Expanding Set of Languages:
  - 8 for ASR
  - 5 for NLU



LLM and Speech Tools/Frameworks

Toolkit for Building SOTA Conversational Models

NVIDIA RIVA - Simple end-to-end workflow for making enabled based conversational application

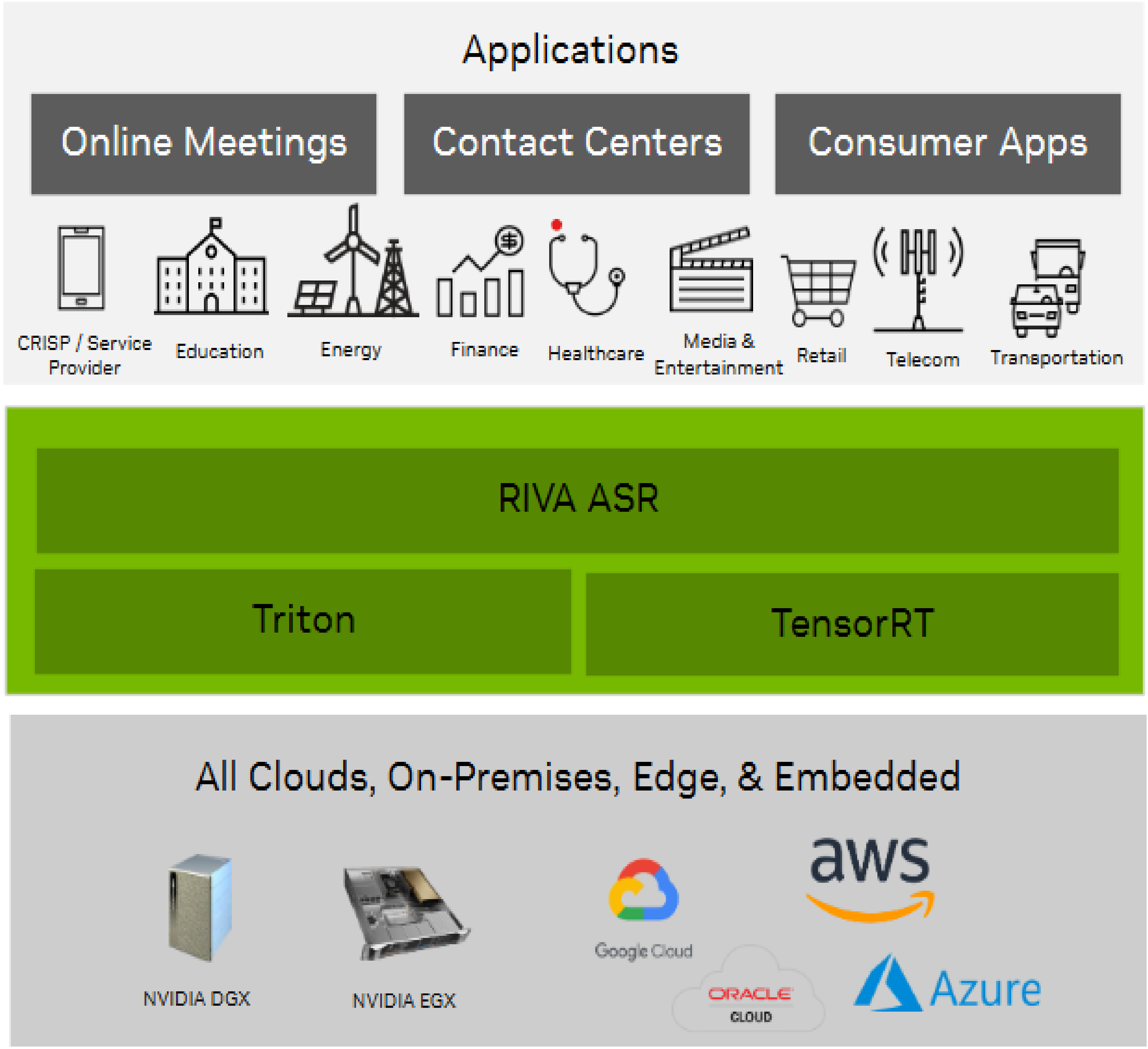


- Highly customizable → highly accurate
- GPU-accelerated → real-time
- Highly scalable: hundreds of thousands of concurrent users
- Deployable everywhere: on-prem, all clouds, edge, embedded

LLM and Speech Tools/Frameworks

Toolkit for Building SOTA Conversational Models
NVIDIA RIVA

- SOTA OOTB models trained for 1M+ hrs on 70K hrs of speech
- Support for:
  - 7 languages: English, Spanish, Mandarin, Hindi, Russian, German, & French
  - 5 coming soon: Japanese, Arabic, Korean, Portuguese, & Italian
- 2X accuracy improvement with customizations for:
  - Industry specific jargon
  - Accents & dialects
  - Noisy environments
- Real-time performance far below 300ms for interactive speech apps
- High scale of 100s thousands of concurrent streams
- Runs anywhere: all clouds, on-prem, at the edge, embedded



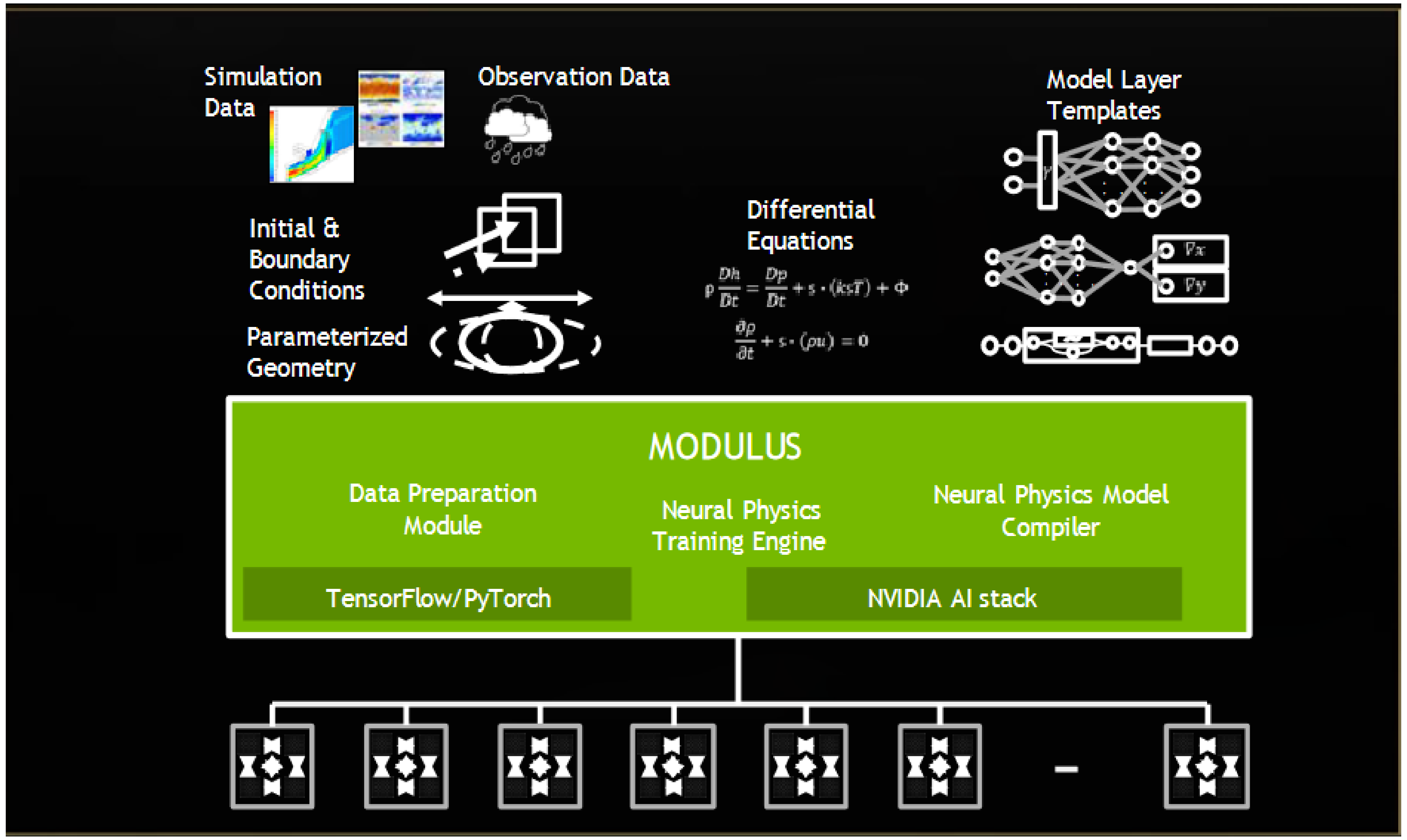
NVIDIA Modulus

What it is:

- Framework for developing physics-ML models – AI framework for science &engineering problems
- Uses simulation and observation data and governing physics equations
  to generate a robust Digital Twin mode

What it is not:

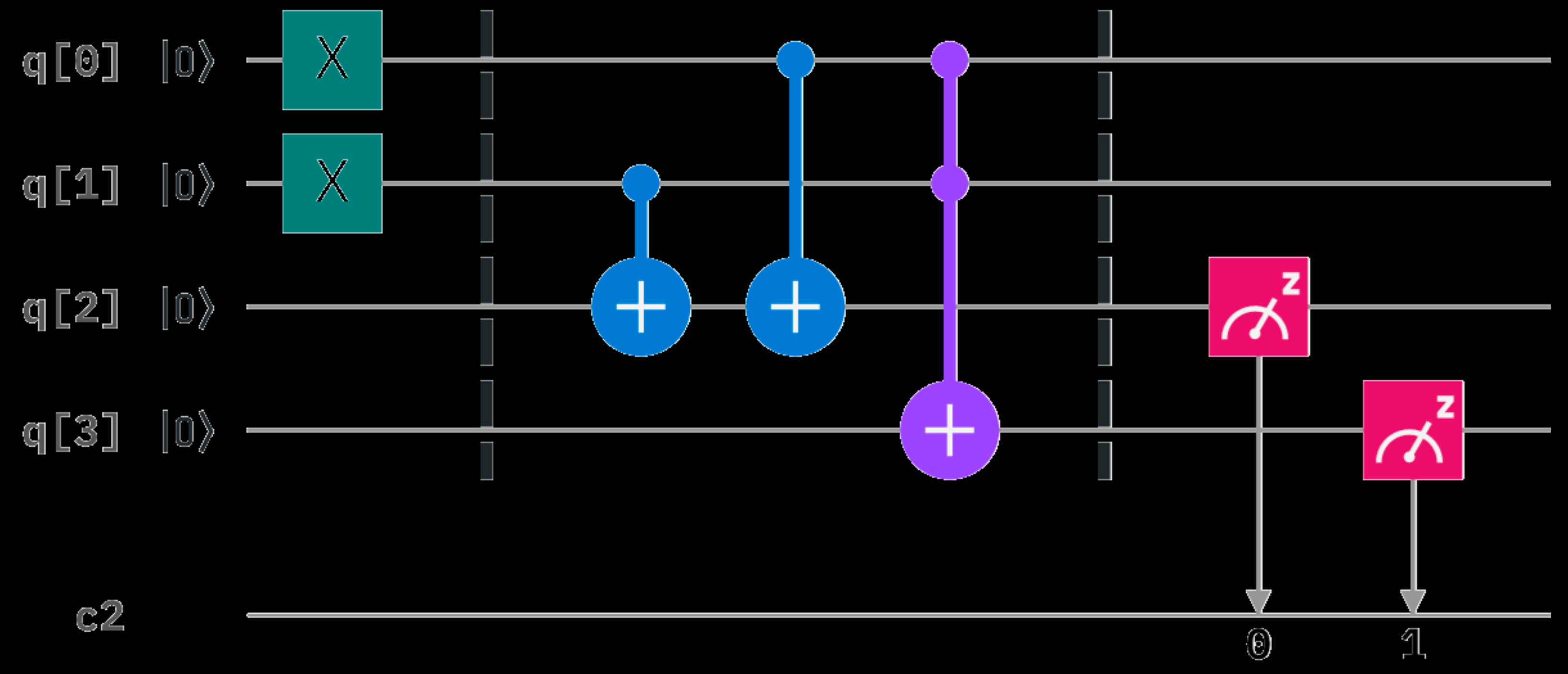
- Not a solver
- Not a simulation platform.



GPU-based supercomputing in the qC ecosystem

Researching the Quantum Computers of Tomorrow with the Supercomputers of Today

Quantum Circuit Simulation
Critical tool for answering today's most pressing questions in Quantum Information Science (QIS):



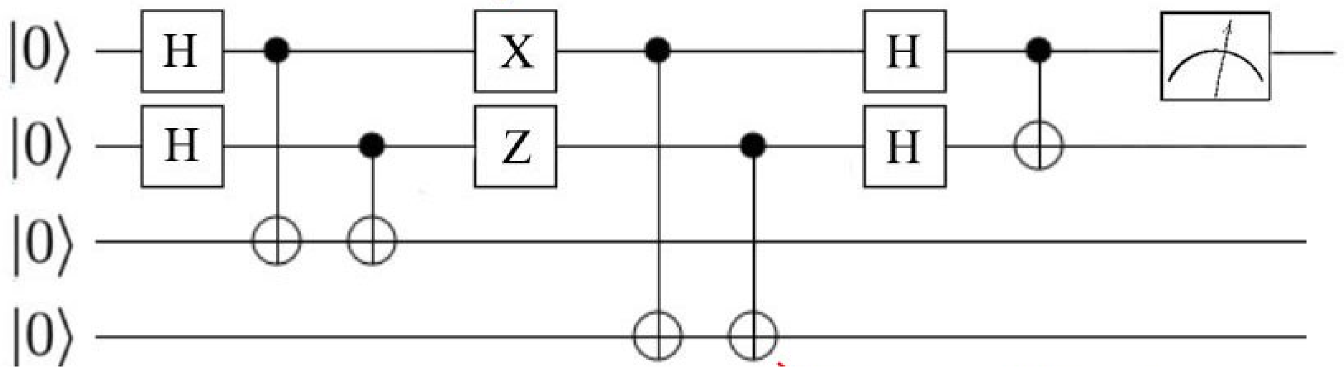
- What quantum algorithms are most promising for near-term or long-term quantum advantage?
- What are the requirements (number of qubits and error rates) to realize quantum advantage?
- What quantum processor architectures are best suited to realize valuable quantum applications?

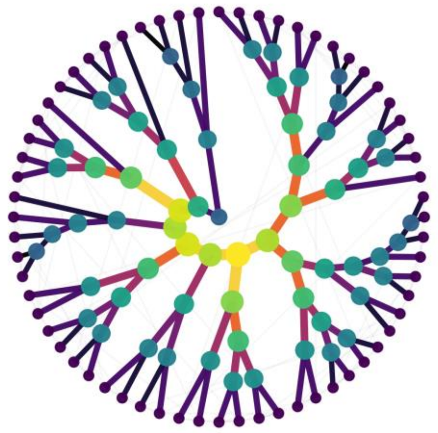
Hybrid Classical/Quantum Applications
Impactful QC applications (e.g. simulating quantum materials and systems) will require classical supercomputers with quantum co-processors


+


- How can we integrate classical HPC systems with quantum computers in an optimal way?
- How can we make use of accelerated classical computing to solve the difficult computational problems needed to use quantum computers effectively?
- How can we enable researchers to easily test quantum algorithms for their applications?

Two Leading Quantum Circuit Simulation Approaches





State vector simulation

“Gate-based emulation of a quantum computer”

- Maintain full 2^n qubit vector state in memory
- Update all states every timestep, probabilistically sample n of the states for measurement

Memory capacity & time grow exponentially w/ # of qubits - practical limit around 50 qubits on a supercomputer

Can model either ideal or noisy qubits

Tensor networks

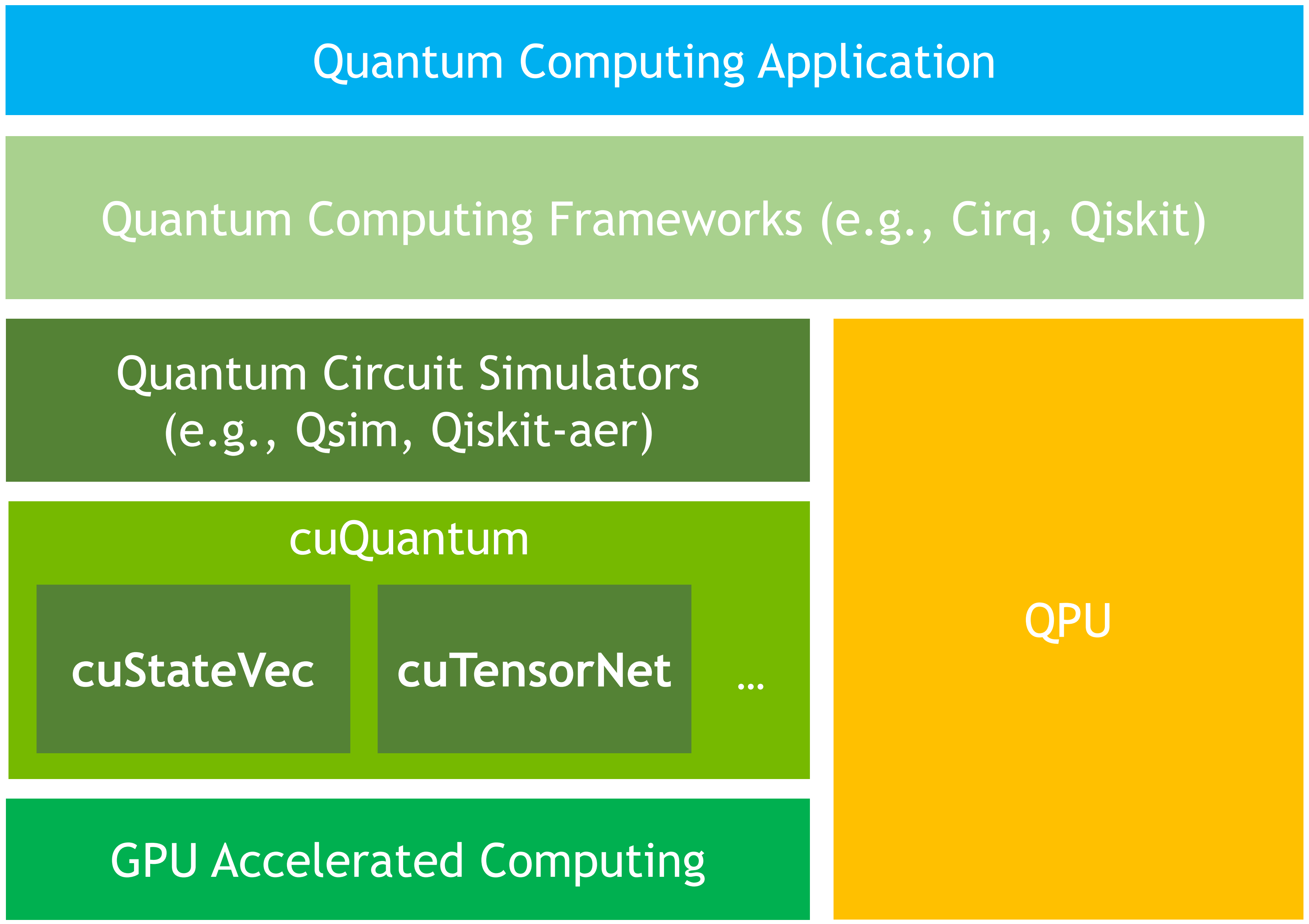
“Only simulate the states you need”

- Uses tensor network contractions to dramatically reduce memory for simulating circuits
- Can simulate 100s or 1000s of qubits for many practical quantum circuits

GPUs are a great fit for either approach

Introducing cuQuantum

- cuQuantum is an SDK of optimized libraries and tools for accelerating quantum computing workflows
- cuQuantum is not a:
  - Quantum Computer
  - Quantum Computing Framework
  - Quantum Circuit Simulator



NVIDIA HPC SDKs

NVIDIA HPC SDK

Available at developer.nvidia.com/hpc-sdk, on NGC, via Spack, and in the Cloud

DEVELOPMENT

Programming Models
- Standard C++ & Fortran
- OpenACC & OpenMP
- CUDA

Compilers
- nvcc
- nvc
- nvc++
- nvfortran

Core Libraries
- libcu++
- Thrust
- CUB

Math Libraries
- cuBLAS
- cuTENSOR
- cuSPARSE
- cuSOLVER
- cuFFT
- cuRAND

Communication Libraries
- HPC-X
  - MPI
  - UCX
  - SHMEM
  - SHARP
  - HCOLL
- NVSHMEM
- NCCL

ANALYSIS

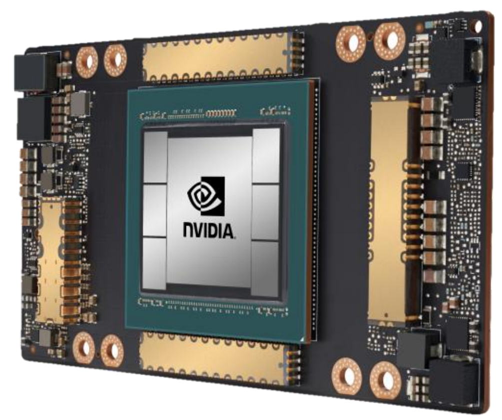
Profilers
- Nsight
  - Systems
  - Compute

Debugger
- cuda-gdb
  - Host
  - Device

Develop for the NVIDIA Platform: GPU, CPU and Interconnect
Libraries | Accelerated C++ and Fortran | Directives | CUDA
7-8 Releases Per Year | Freely Available

HPC Compilers

NVC | NVC++ | NVFORTRAN



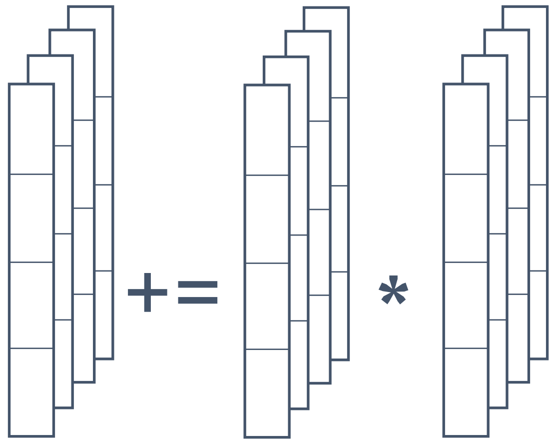
Accelerated

A100
Automatic

Fortran
NVIDIA CUDA
OpenACC
More Science, Less Programming
OpenMP

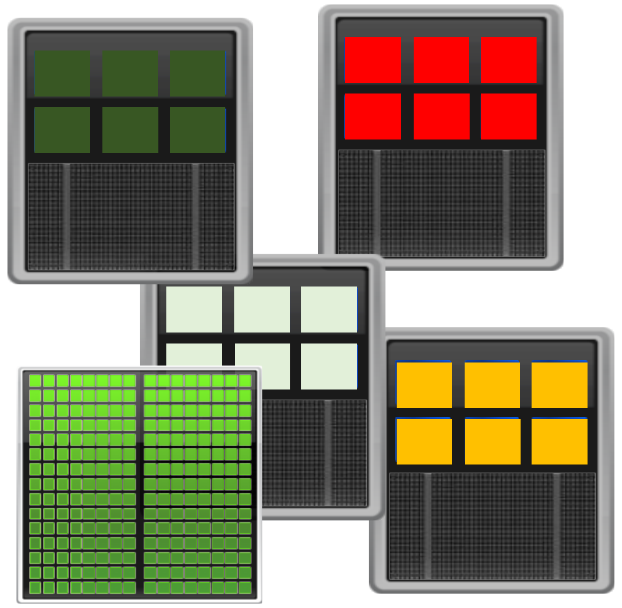
Programmable

Standard Languages
Directives
CUDA



CPU Optimized

Directives
Vectorization



Multi-Platform

x86_64
Arm
OpenPOWER

PROGRAMMING THE NVIDIA PLATFORM

CPU, GPU, and Network

ACCELERATED STANDARD LANGUAGES

ISO C++, ISO Fortran

```
std::transform(par, x, x+n, y, y,
    [=](float x, float y){ return y +
a*x; }
);
```

```
do concurrent (i = 1:n)
   y(i) = y(i) + a*x(i)
enddo
```

```
import cunumeric as np
…
def saxpy(a, x, y):
    y[:] += a*x
```

INCREMENTAL PORTABLE OPTIMIZATION

OpenACC, OpenMP

```
#pragma acc data copy(x,y) {
...
std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});
...
}
```

```
#pragma omp target data map(x,y) {
...
std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});
...
}
```

PLATFORM SPECIALIZATION

CUDA

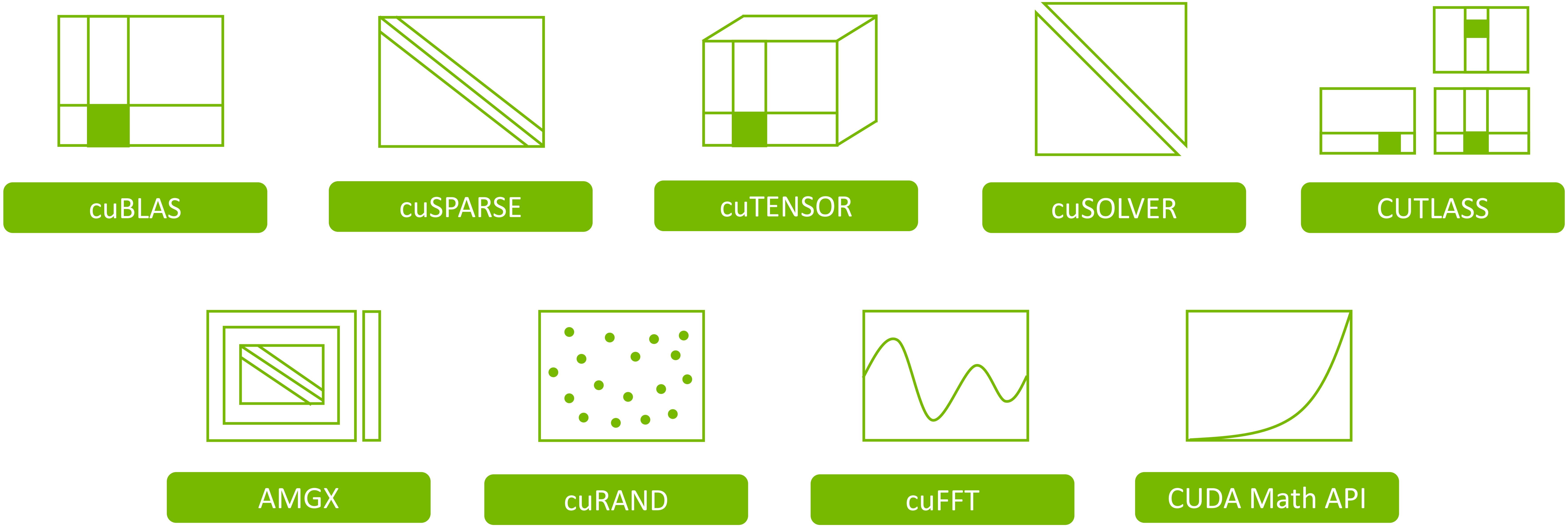
```
__global__
void saxpy(int n, float a,
           float *x, float *y) {
  int i = blockIdx.x*blockDim.x +
          threadIdx.x;
  if (i < n) y[i] += a*x[i];
}

int main(void) {
  ...
  cudaMemcpy(d_x, x, ...);
  cudaMemcpy(d_y, y, ...);

  saxpy<<<(N+255)/256,256>>>(...);

  cudaMemcpy(y, d_y, ...);
```

ACCELERATION LIBRARIES

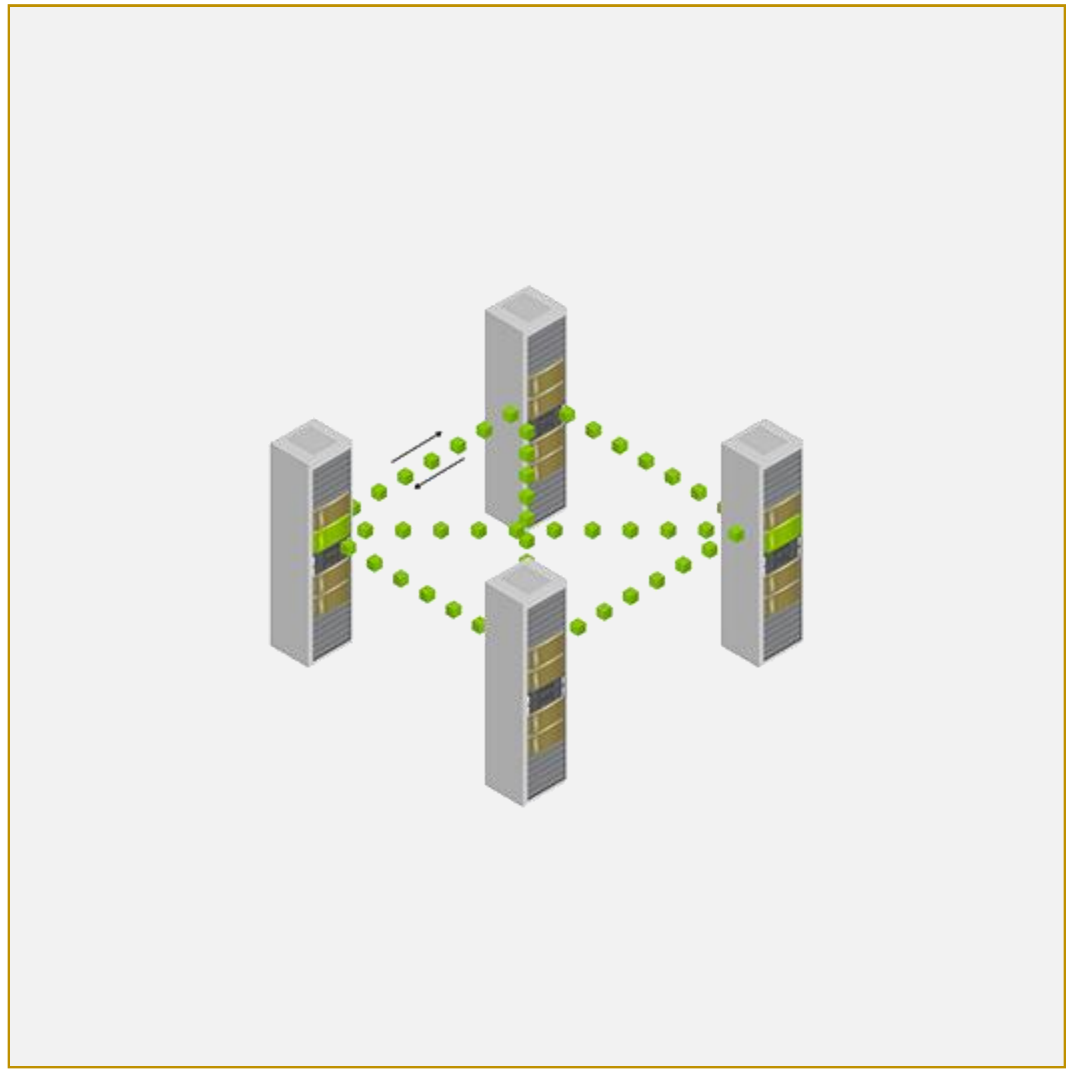
Core	Math	Communication	Data Analytics	AI	Quantum

NVIDIA MATH LIBRARIES

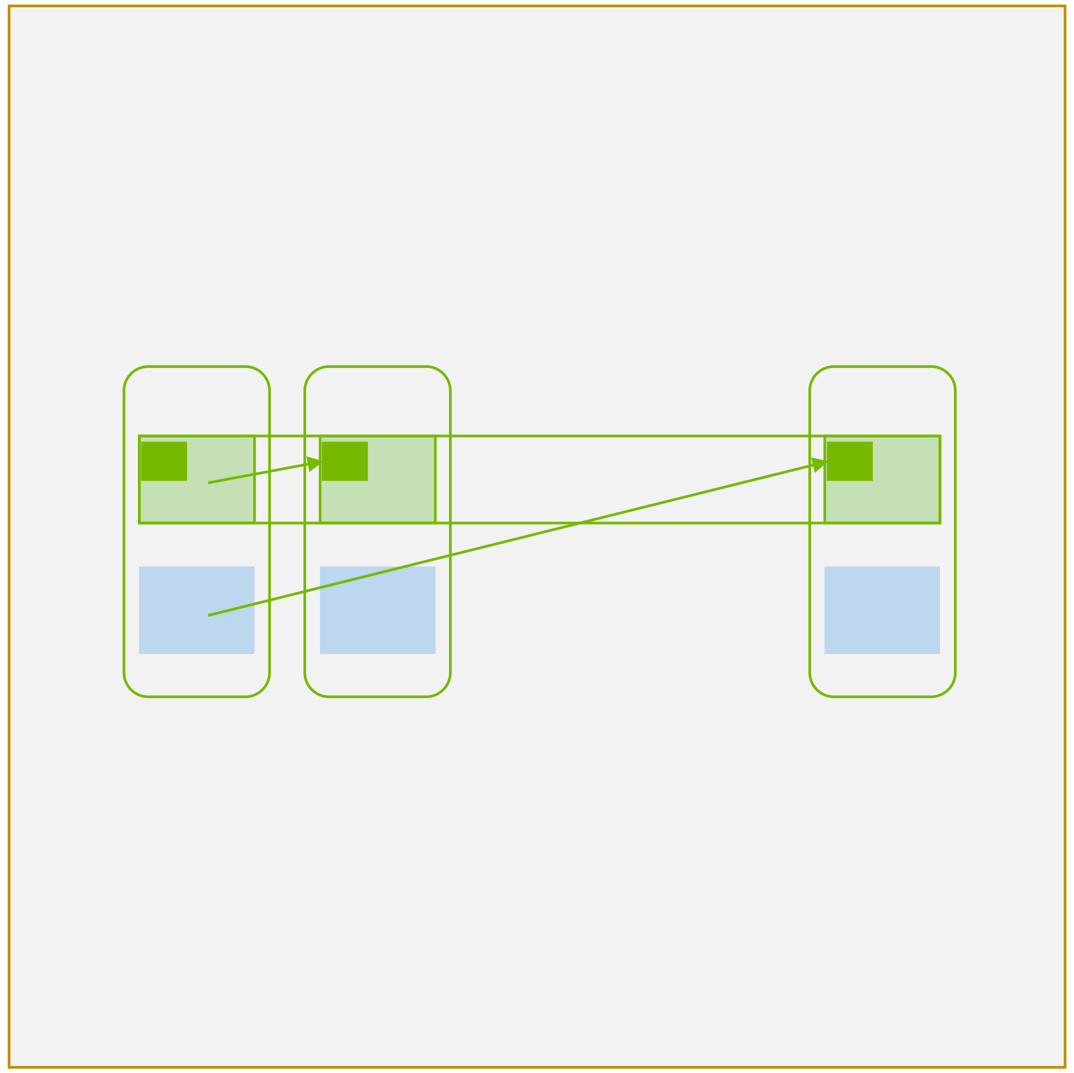
Linear Algebra, FFT, RNG and Basic Math



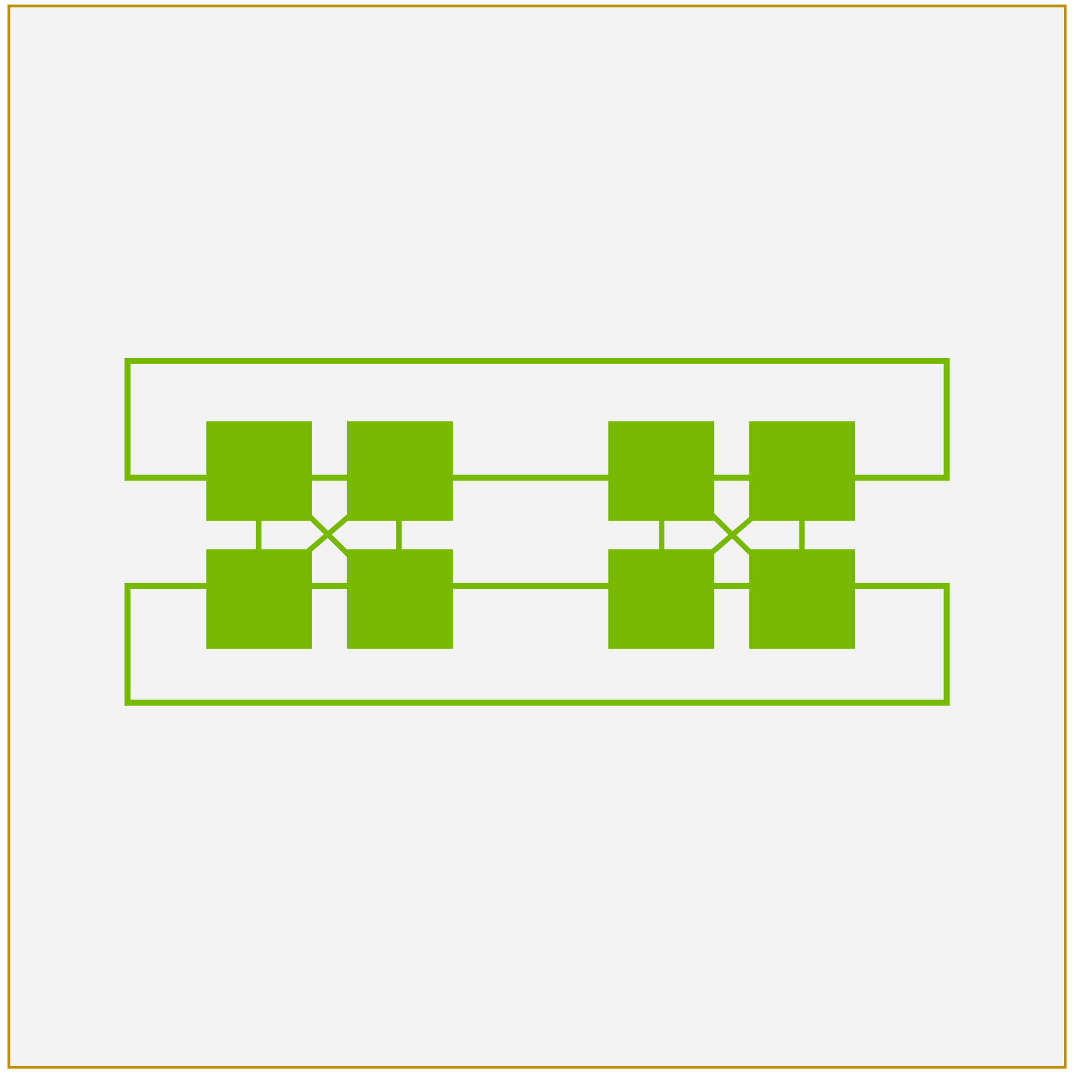
NVIDIA COMMUNICATION LIBRARIES



HPC-X
Optimized whole-system communications



NVSHMEM
Low-latency PGAS programming

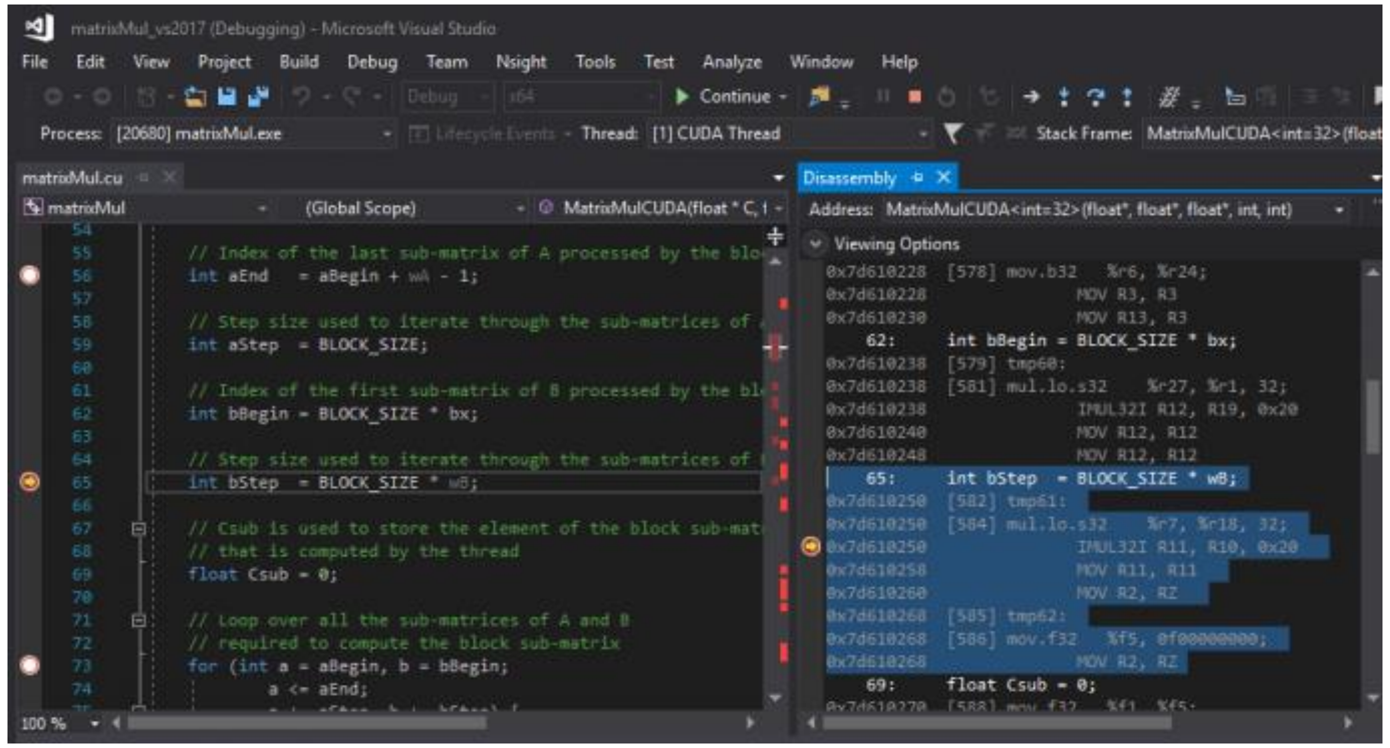


NCCL
Multi-node collectives for accelerators

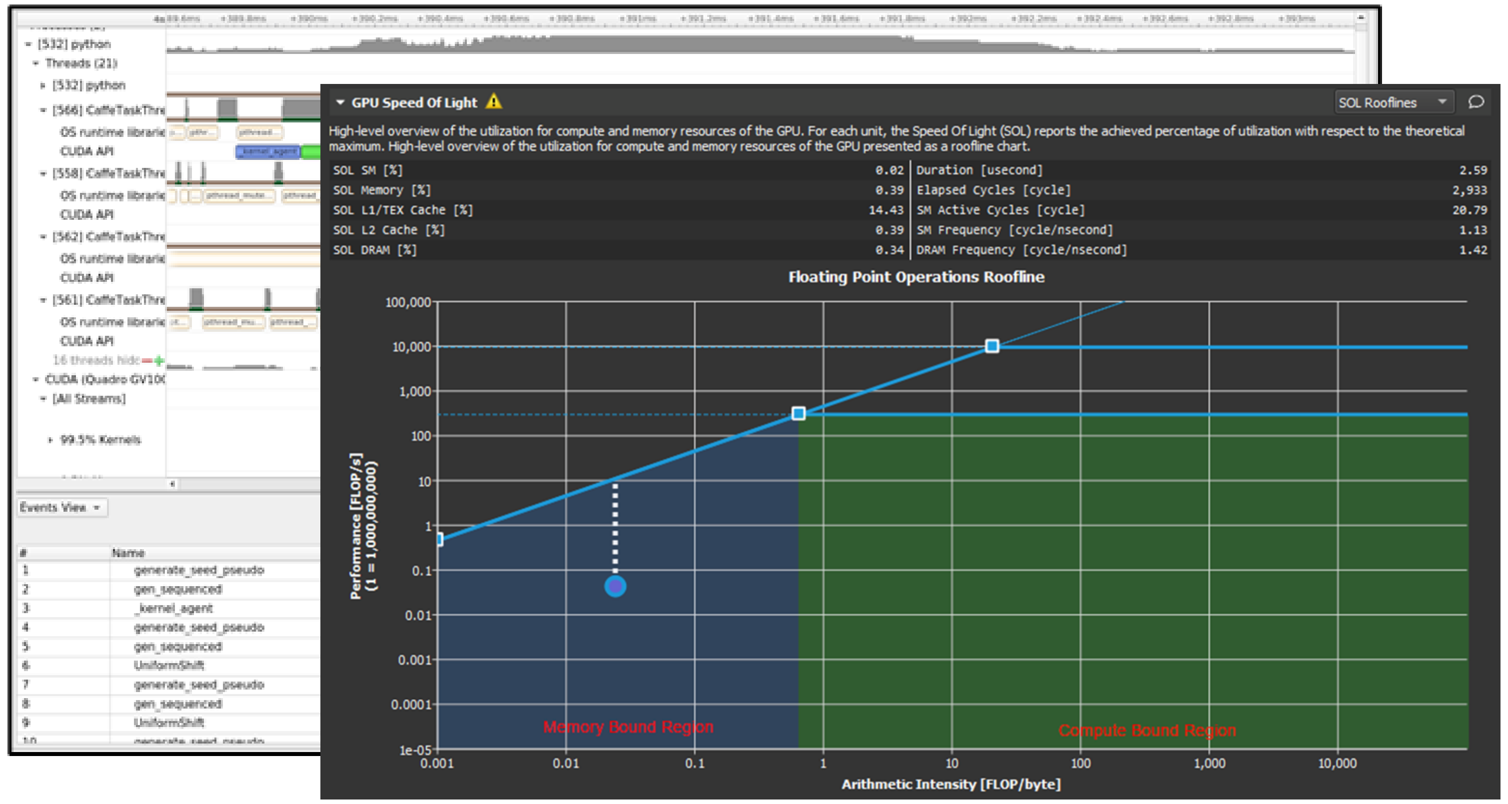
Multi-GPU Programming Models [S31050]

DEVELOPER TOOLS

Debuggers: cuda-gdb, Nsight Visual Studio Edition



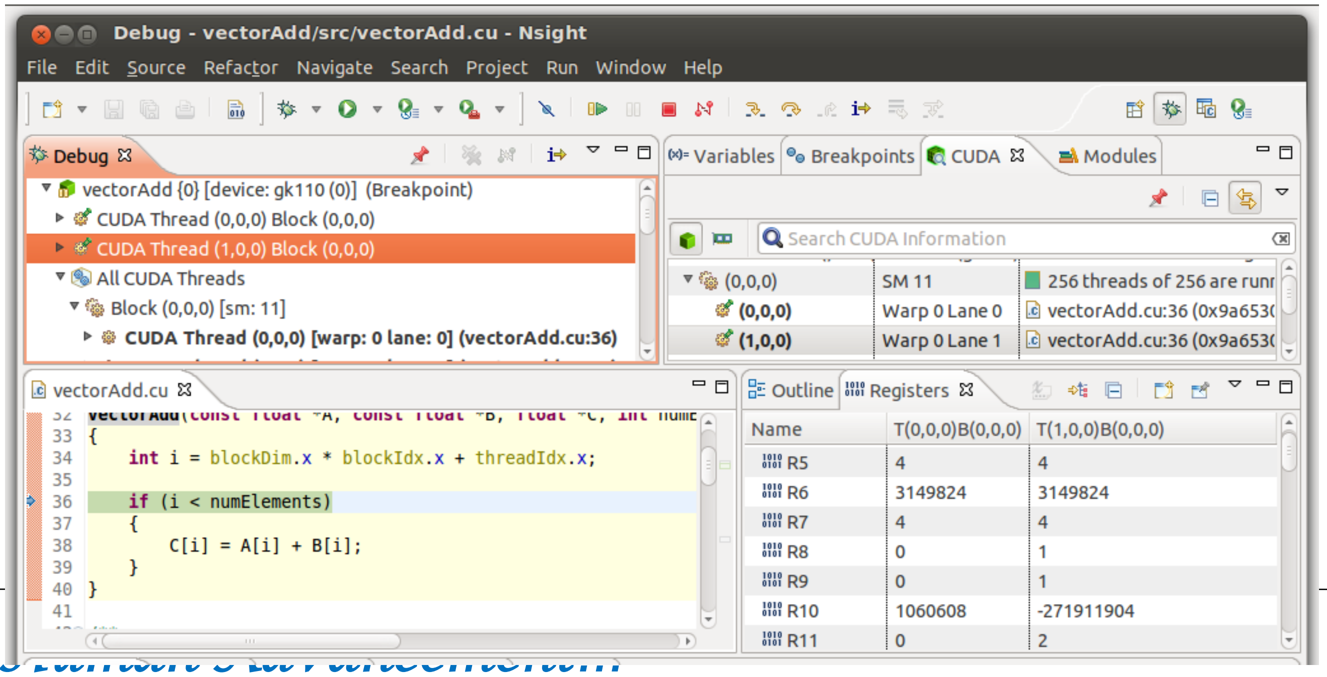
Profilers: Nsight Systems, Nsight Compute, CUPTI, NVIDIA Tools eXtension (NVTX)



Correctness Checker:: Compute Sanitizer

```
$ compute-sanitizer --leak-check full memcheck_demo
========= COMPUTE-SANITIZER
Mallocing memory
Running unaligned_kernel
Ran unaligned_kernel: no error
Sync: no error
Running out_of_bounds_kernel
Ran out_of_bounds_kernel: no error
Sync: no error
========= Invalid __global__ write of size 4 bytes
=========     at 0x60 in memcheck_demo.cu:6:unaligned_kernel(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x400100001 is misaligned
```

IDE integrations: Nsight Eclipse Edition
Nsight Visual Studio Edition
Nsight Visual Studio **Code** Edition



Case Study of DGX SuperPOD for Large Scale Workloads

Digital India
Power To Empower

Choosing DGX? You're in Very Good Company

Thousands of leading companies deploy DGX today

9 OF THE TOP 10 GLOBAL UNIVERSITIES

7 OF THE TOP 10 US HOSPITALS

6 OF THE TOP 10 US BANKS

7 OF THE TOP 10 GLOBAL CAR MANUFACTURERS

8 OF THE TOP 10 GLOBAL TELCOS

10 OF THE TOP 10 US GOVERNMENT INSTITUTIONS

7 OF THE TOP 10 CONSUMER INTERNET COMPANIES

10 OF THE TOP 10 GLOBAL AEROSPACE & DEFENSE COMPANIES

Digital India
Power To Empower

Building Leading Edge AI in Every Industry

Thousands of successful DGX customer deployments to date

HEALTHCARE
- FUJIFILM
- King's College London
- MERCK
- OLYMPUS
- Paige
- Pfizer
- St. Jude Children's Research Hospital
- UCSF ci² center for intelligent imaging
- UFHealth

CONSUMER INTERNET / IT SERVICES
- JD.COM
- Microsoft
- NAVER CLOVA
- RingCentral
- servicenow
- Tencent

MANUFACTURING
- ISHIDA
- MUSASHI
- SONY
- SPIL Siliconware
- wistron

SUPERCOMPUTING
- Argonne National Laboratory
- C-DAC
- NSM
- ILLINOIS NCSA | National Center for Supercomputing Applications
- NERSC
- Pacific Northwest National Laboratory
- Pittsburgh Supercomputing Center
- Sandia National Laboratories
- TACC

AUTOMOTIVE
- BMW GROUP
- Continental
- FORD OTOSAN
- HONDA
- HYUNDAI MOTOR GROUP
- SUBARU
- TOYOTA
- VINFAST
- zenseact

FINANCIAL SERVICES
- CCC Information Services Inc.
- BNY MELLON
- BOREALIS AI
- Deloitte.
- KENSHO
- Riskfuel
- usbank

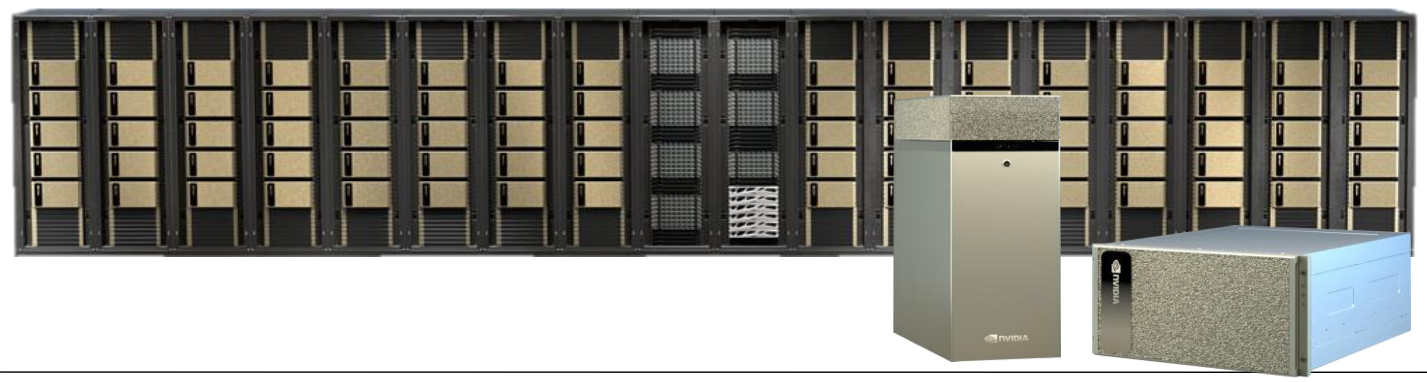
EDUCATION & RESEARCH
- KIT
- MSOE University
- National Taiwan University
- LiU Linköping University
- NYU
- SiDi
- SMU
- UF University of Florida
- UTSA The University of Texas at San Antonio

TELCO
- kt
- docomo
- NTT Resonant
- SK telecom
- TELCONET

GOVERNMENT
- LOCKHEED MARTIN
- NORTHROP GRUMMAN
- UNITED STATES POSTAL SERVICE

RETAIL & CONSUMER GOODS
- BEST BUY
- Domino's
- Kroger
- P&G

ENERGY
- Baker Hughes
- EDF
- Marathon Oil
- Shell
- STONE RIDGE TECHNOLOGY



One Vision. One Goal... Advanced Computing for Human Advancement...

NVIDIA

AI4Barath

Building open-source language AI for Indian languages, including datasets, models, and applications.

To train and evaluate LLM models demands massive distributed computing power, clusters of accelerated-based hardware and memory, reliable and scalable machine learning frameworks, and fault-tolerant systems

Building AI models for Indic languages is challenging tasks specially training a Large Language models.

CDAC-C is Supporting AI4Barath With GPU compute on various research areas below for building Language models, datasets and applications for Indian Languages

- Translation
- Transliteration
- Speech Recognition
- Language Understanding
- Language Generation

Enterprise AI Workflows Developed on DGX Infrastructure

INDUSTRIAL INSPECTION

Baker Hughes

ISHIDA

SPIL 矽品精密 Siliconware

wistron

DRUG DISCOVERY

MERCK

Paige

Pfizer

DIGITAL TWIN

BMW GROUP

LOCKHEED MARTIN

Kroger

FRAUD DETECTION, CLAIMS PROCESSING

CCC INFORMATION SERVICES INC.

BNY MELLON

usbank

LARGE LANGUAGE MODELS / LARGE PRE-TRAINED MODEL

AI SWEDEN

HYUNDAI MOTOR GROUP

JD.COM

kt

AUTONOMOUS SYSTEMS

Continental

VINFAST

zenseact

NVIDIA

सी-डैक
CDAC
प्रगत संगणन विकास केंद्र
CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING

NPSF
National PARAM Supercomputing Facility