# これからの計算工学に
# NVIDIA GPUがもたらすものとは

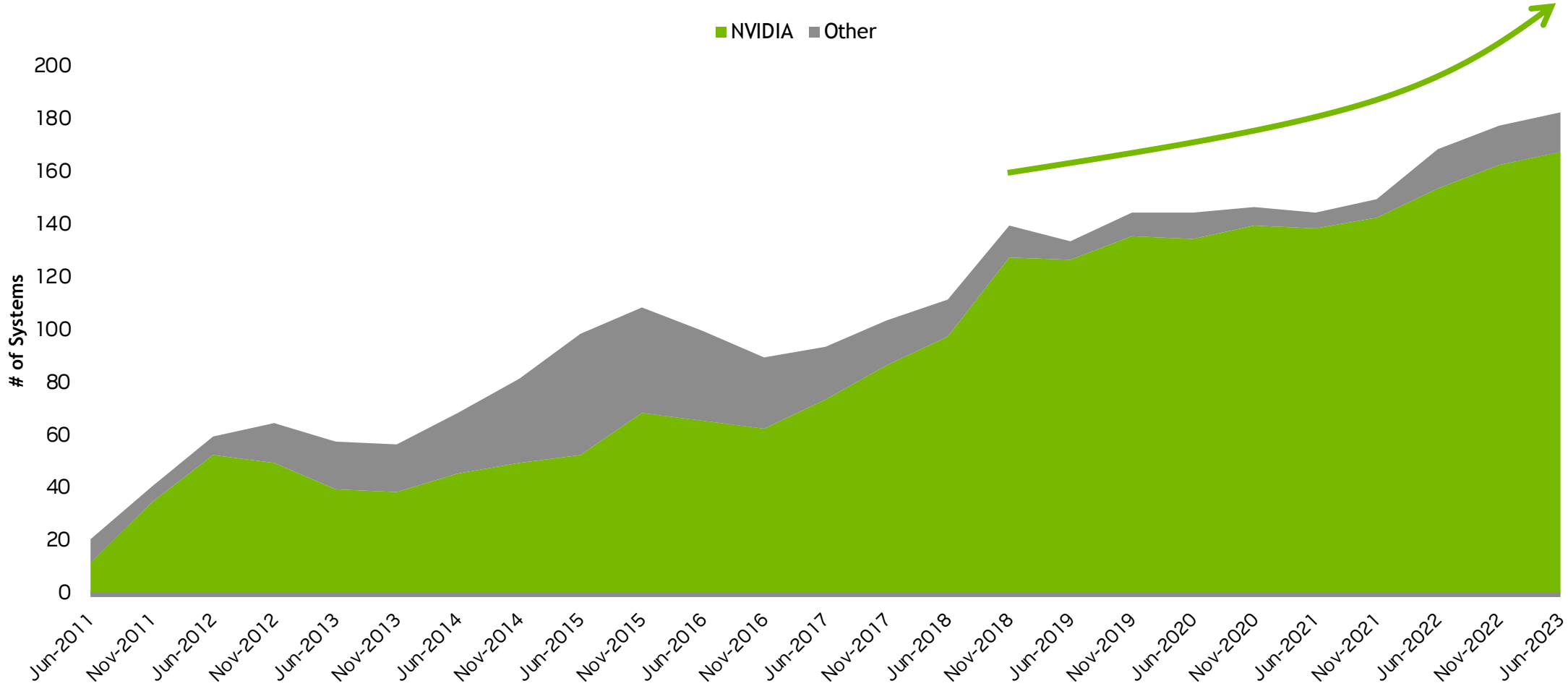
プロメテックグループ
（プロメテック・ソフトウェア/GDEPソリューションズ）
**NVIDIA**

# これからの計算工学に NVIDIA GPU がもたらすものとは – NVIDIA HPC SDK と NVIDIA Modulus の紹介–

Shinnosuke Furuya, Ph.D., HPC Developer Relations

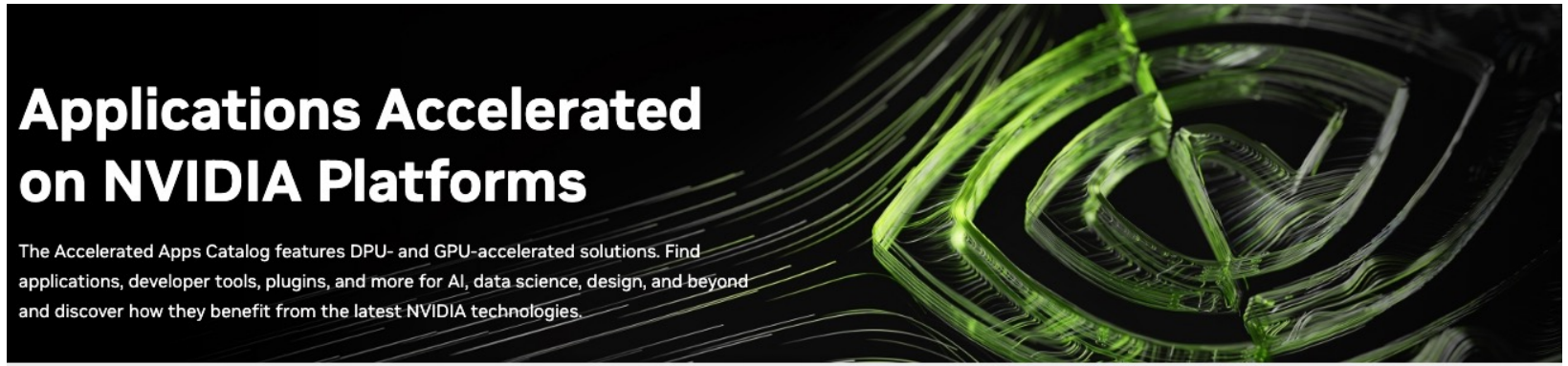
# TOP500 List

## Top 500 Supercomputer with Accelerator



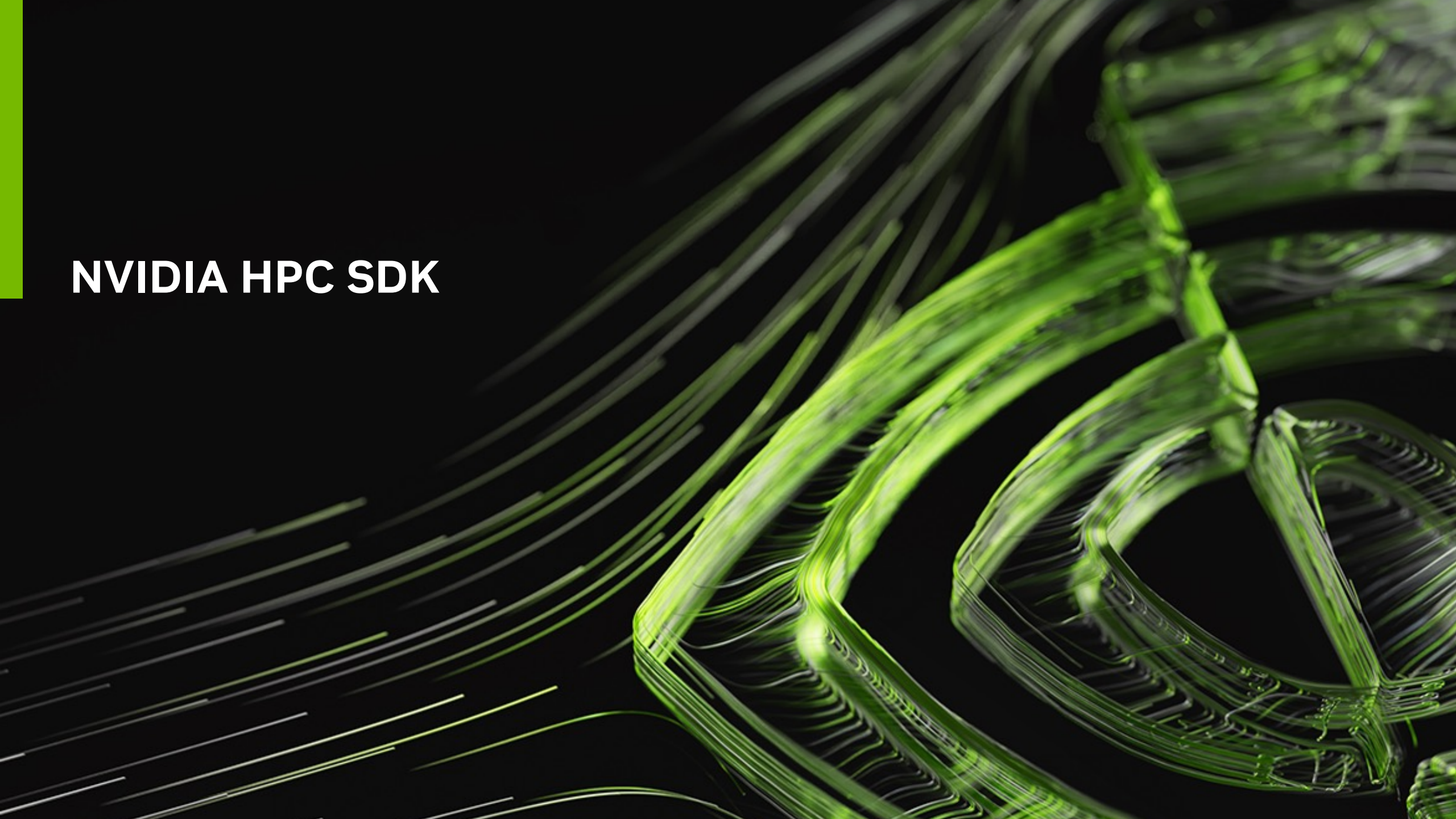

https://top500.org/

# Applications Accelerarted on NVIDIA Platforms




https://www.nvidia.com/en-us/gpu-accelerated-applications/

# NVIDIA HPC SDK

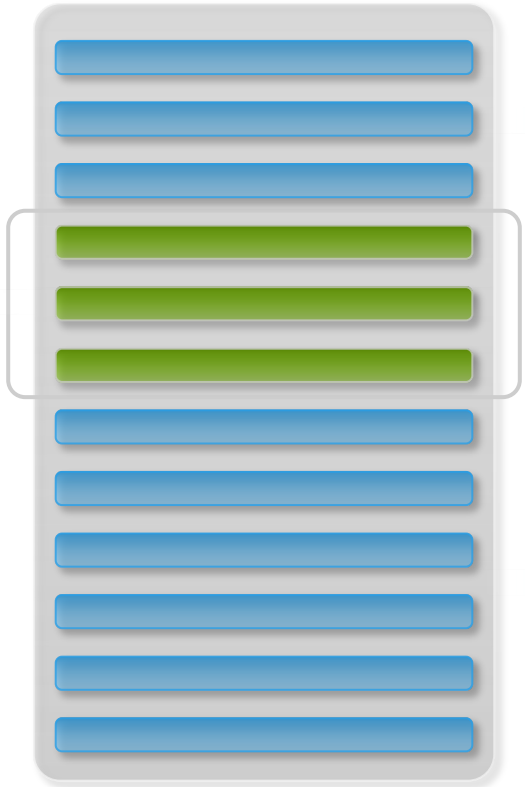


# GPU Computing in a Nutshell

## All GPU Programming Models Follow This Pattern

Program flow and resource allocation is managed by the CPU

Data & GPU “kernels” offload to the GPU

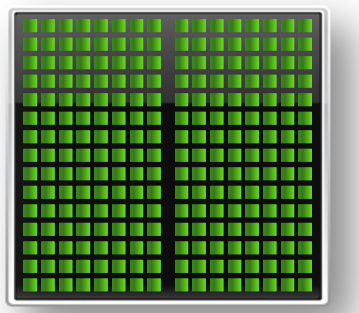


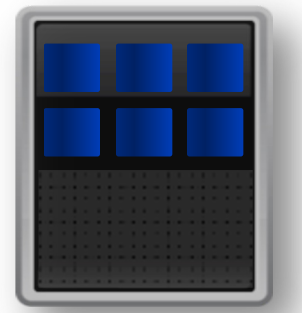
Rest of Sequential CPU Code

**GPU**

Parallelize using CUDA Programming Model

**CPU**





\+ CPU & GPU work together

# 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

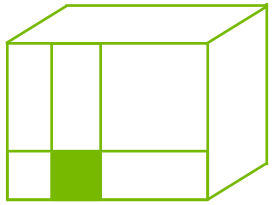
Core | Math | Communication | Data Analytics | AI | Quantum

# NVIDIA Math Libraries

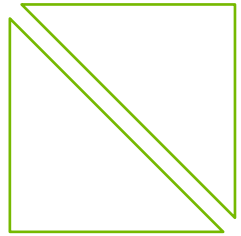
Linear Algebra, FFT, RNG and Basic Math

cuBLAS

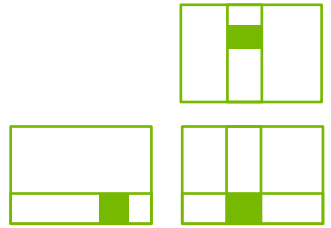
cuSPARSE



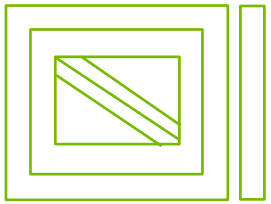
cuTENSOR



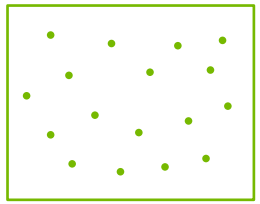
cuSOLVER



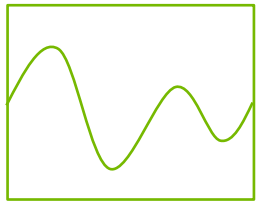
CUTLASS



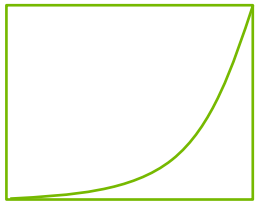
AMGX



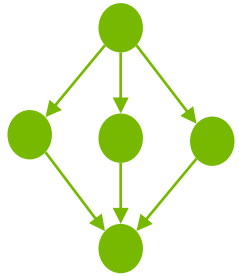
cuRAND



cuFFT



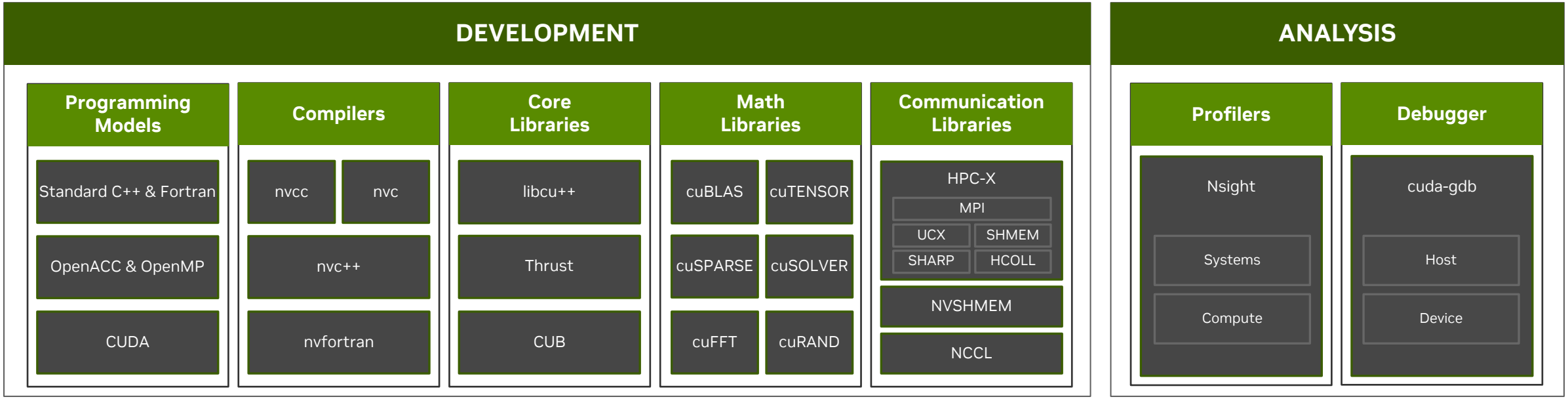
CUDA Math API



Legate

# NVIDIA HPC SDK

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




Develop for the NVIDIA Platform: GPU, CPU and Interconnect
Libraries | Accelerated C++ and Fortran | Directives | CUDA
x86_64 | AArch64 | OpenPOWER
7-8 Releases Per Year | Freely Available

# Choose a Programming Model

They can be ~~only~~ **more than** one

| Libraries | Standard Languages | Compiler Directives | CUDA Languages |
|---|---|---|---|
| • Accelerate common operations with little/no code changes<br>• Expert-tuned performance<br>• Forward support guarantees | • Strong cross-platform support<br>• Single source code for multiple platforms<br>• Reduced learning curve | • High cross-platform support<br>• Single source code for multiple platforms<br>• Reduced learning curve<br>• Additional programmer control | • Exposes full GPU capabilities<br>• Trades portability for performance<br>• Distinct GPU/CPU code paths<br>• Full programmer control |
| Programmer Productivity | | | Programmer Control |

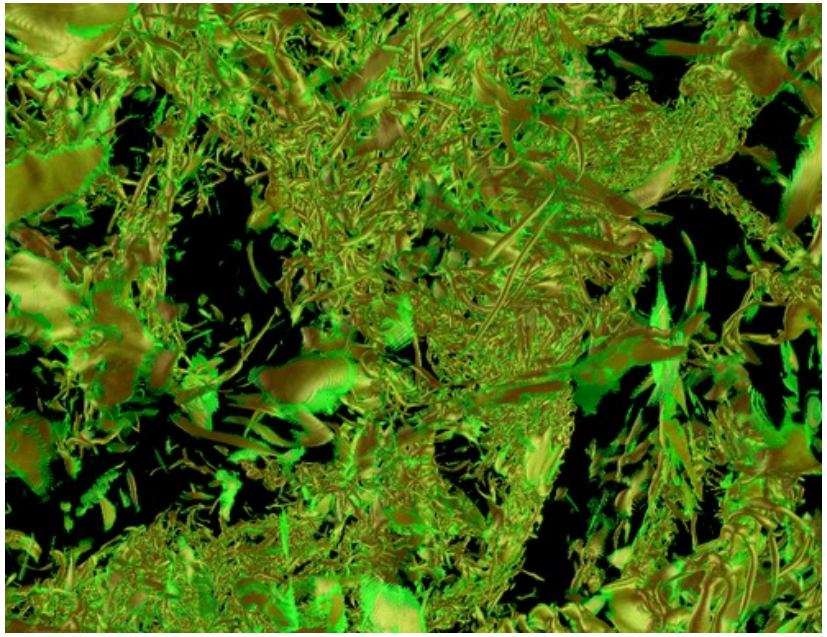
By design these approaches are interoperable,
so you can choose the right balance for your needs

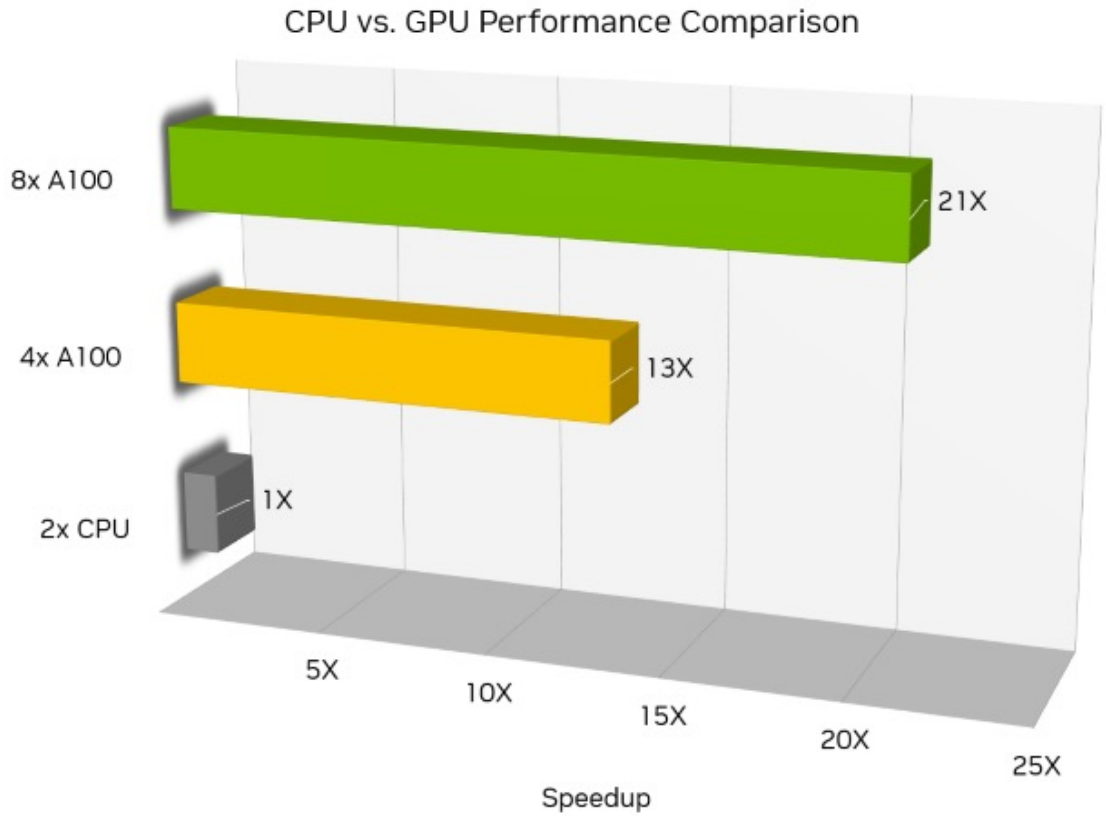
# Magnetohydrodynamics Simulation

## Eliminating Compute Bottleneck with cuFFT

### MHD3D with cuFFT

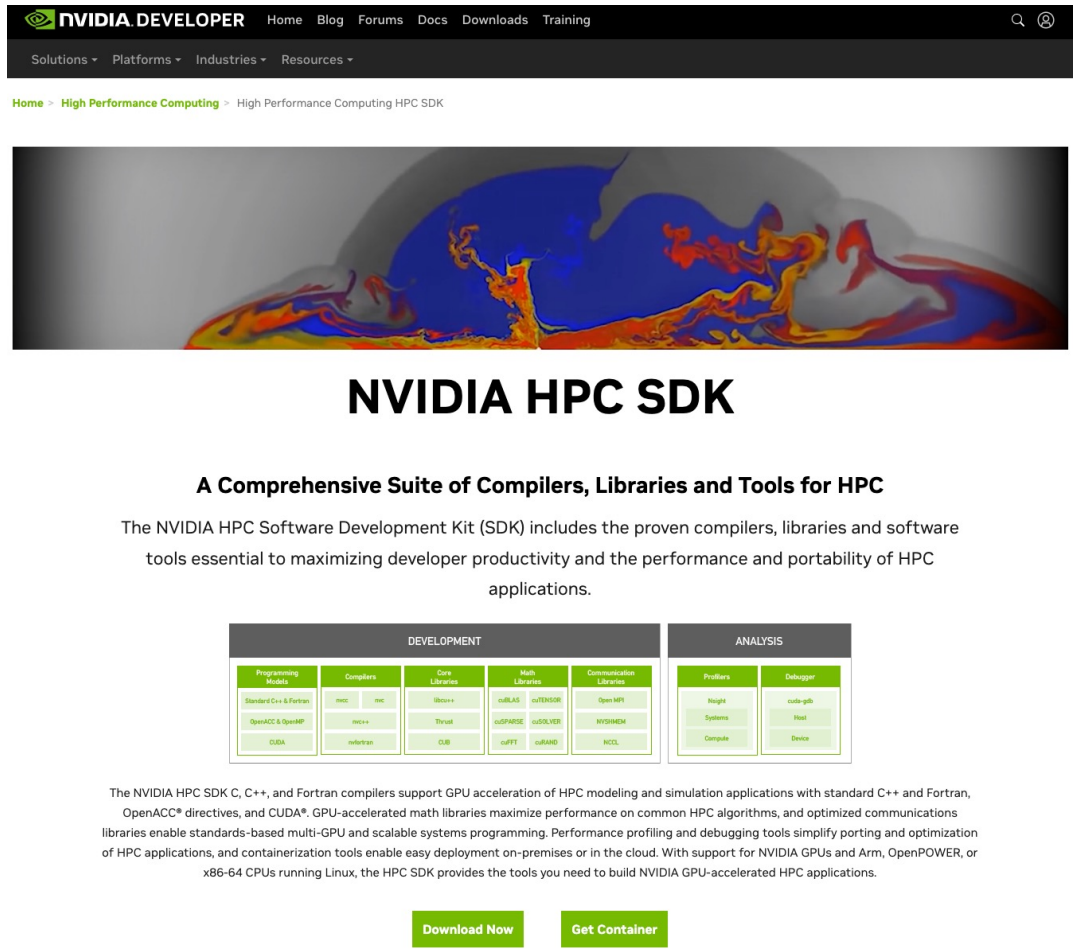
- Incompressible Hall MHD sim bottleneck is 3D FFTs
- MultiGPU offers speedups of 13x and 21x
- 1D FFT of FFTW replaced with cuFFT
- Accelerate transpose operations from Alltoallv to NCCL
- Bottleneck moved to visualization



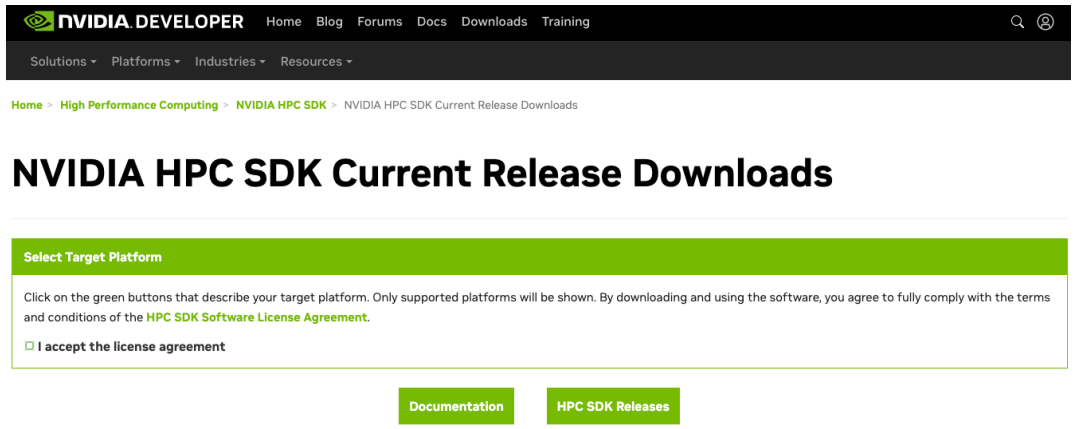



*Wisteria-Aquarius – Xeon Platinum 8360Y | A100 40GB

# Getting Started with NVIDIA HPC SDK




Go to https://developer.nvidia.com/hpc-sdk
and click “Download Now”




Read the “HPC SDK Software License Agreement”,
click “I accept the license agreement”,
and proceed to select your platform, etc.

# Getting Started with NVIDIA HPC SDK

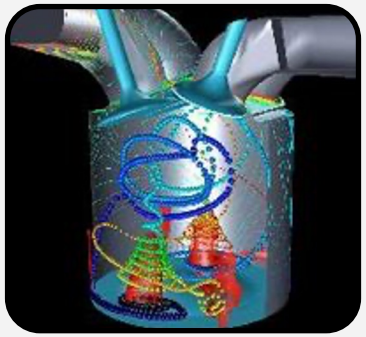
## Resources

- Latest Version: 23.5

- Product Page
  - https://developer.nvidia.com/hpc-sdk
- NGC Container Image
  - https://catalog.ngc.nvidia.com/orgs/nvidia/containers/nvhpc
- Documentation
  - https://docs.nvidia.com/hpc-sdk/index.html

# NVIDIA Modulus

# AI Powered Computational Domains



**Computational Eng.**

Solid & Fluid Mechanics,

Electromagnetics,

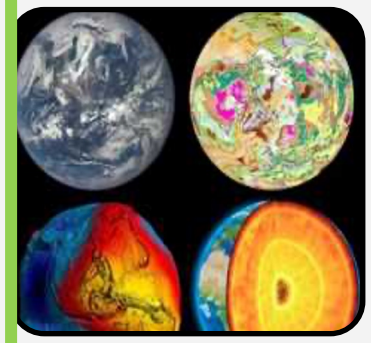
Thermal, Acoustics,

Optics, Electrical,

Multi-body Dynamics,

Design Materials,

Systems



**Earth Sciences**

Climate Modeling,

Weather Modeling,

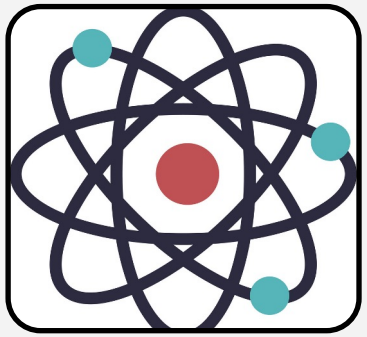
Ocean Modeling,

Seismic Interpretation

**Life Sciences**

Genomics,

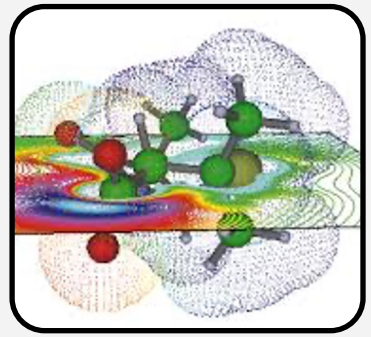
Proteomics



**Computational Physics**

Particle Science,

Astrophysics



**Computational Chemistry**

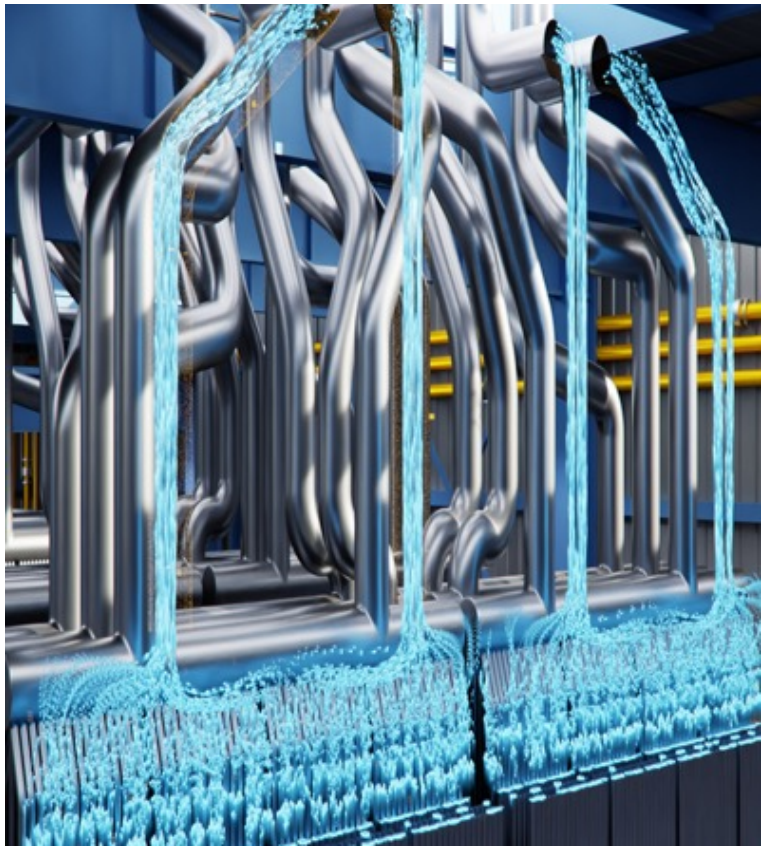
Quantum Chemistry,

Molecular Dynamics

Process/Product Design, Manufacturing, Testing, In-Service

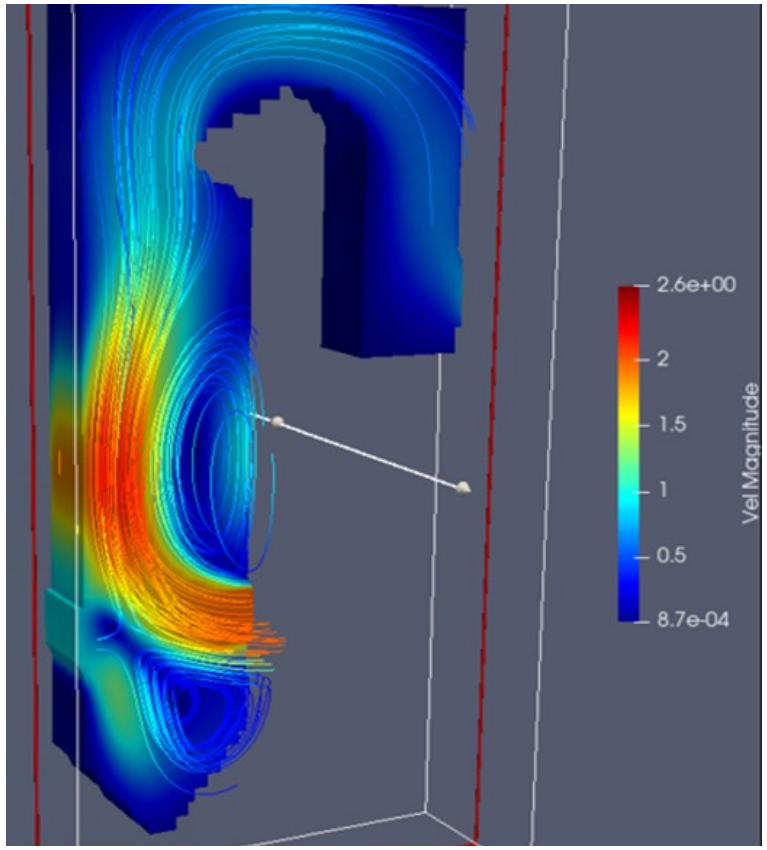
# Developing Digital Twins with Physics-ML

## Industrial Digital Twins



**Siemsns Energy HRSG**
PINN | Coupled Flows, Physics

**Siemens Gamesa Windfarm**
PINN/GAN | Super-Resolution

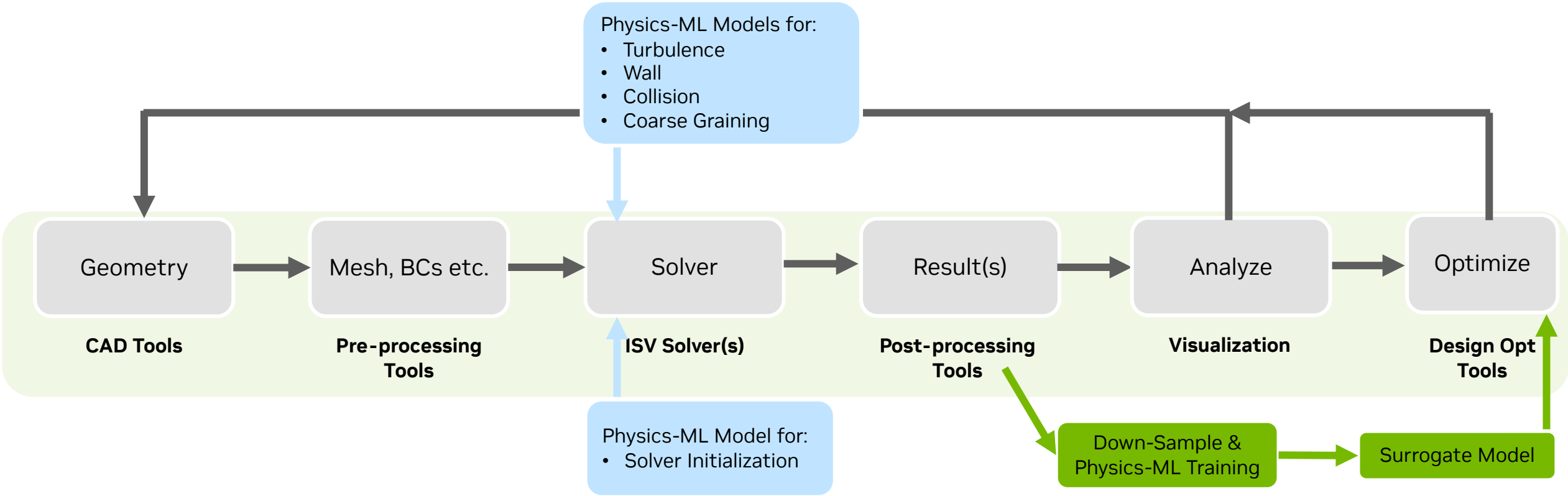



**NETL Power Plant Boiler**
PINN | Multi-Physics, Custom Training

https://resources.nvidia.com/l/en-us-modulus-pathfactory-explore-page

# Simulation Acceleration with Modulus

## Physics-ML Augmented Simulation Workflows

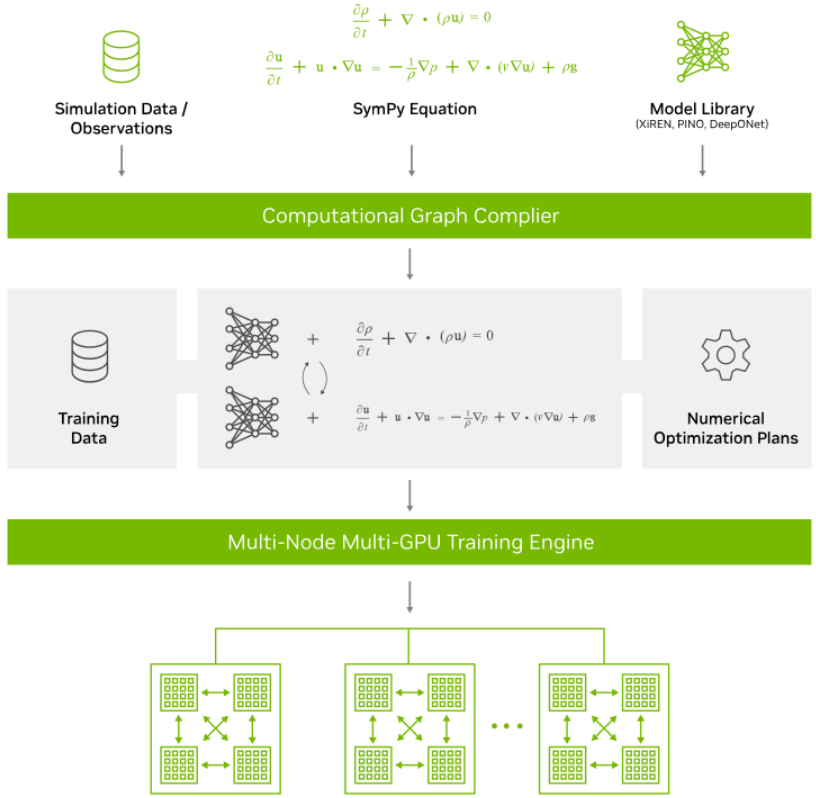


# Open-Source Toolkit for Physics-ML

## NVIDIA Modulus

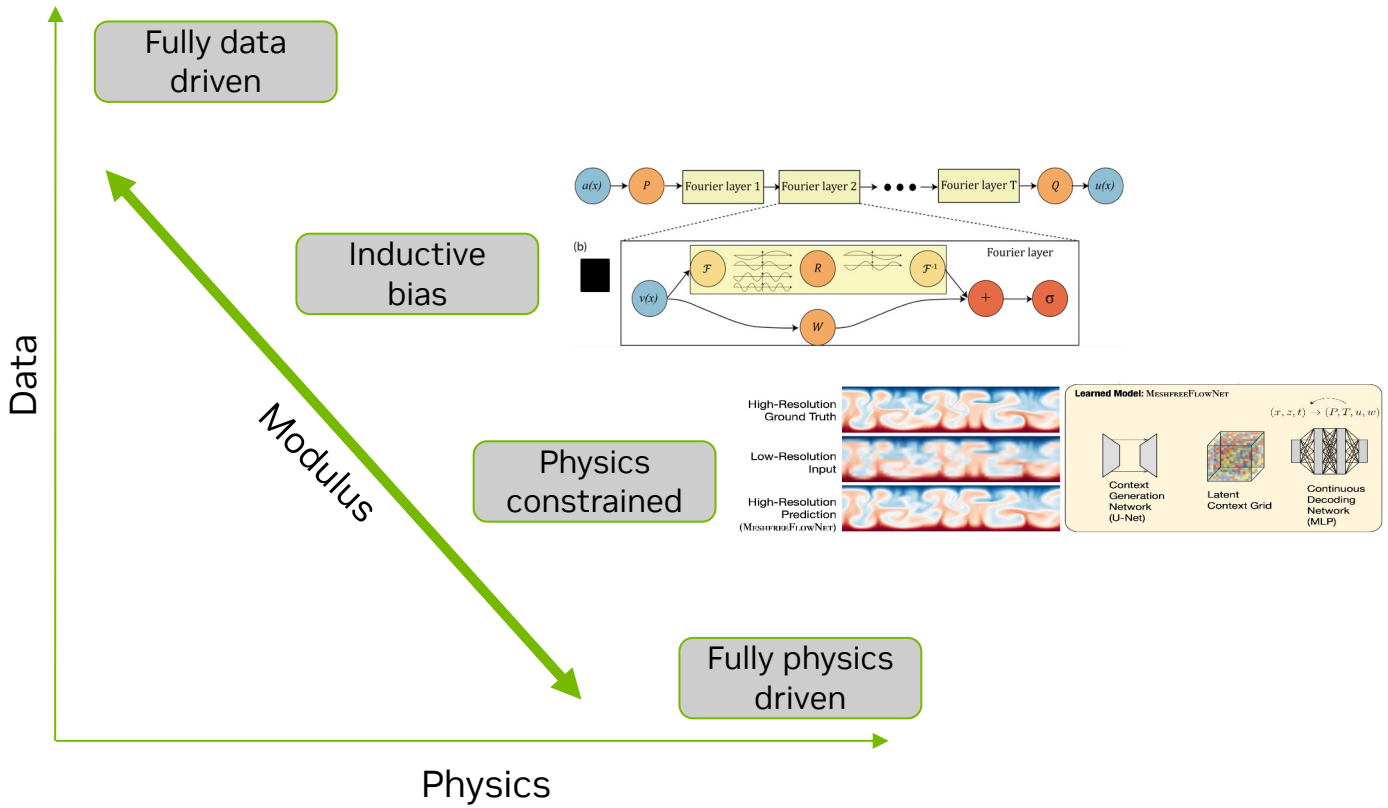
- A customizable platform - training and inference pipeline - using Physics (governing equations) and Data (simulation/observations)
- Python based APIs for ease of use
- Facilitates open collaboration within the Physics-ML scientific community
- Well documented features and functionality for ease of use
- Open-source code – easier to understand and customize
- Import PyTorch models from research for your custom application

- Source code:
  - https://github.com/NVIDIA/Modulus
  - https://github.com/NVIDIA/modulus-launch
  - https://github.com/NVIDIA/modulus-toolchain
  - https://github.com/NVIDIA/modulus-sym



# Open-Source Toolkit for Physics-ML

## Novel NN architectures




- Diverse Physics-ML approaches - Model Zoo:
  - PDE driven Physics-ML recipes
  - Data driven Physics-ML recipes
  - Hybrid (Data + PDE) Physics-ML recipes
- PDE Driven - PINNs:
  - Fourier Feature Network
  - Spatial-temporal Fourier Feature Networks
  - Super Resolution Net …
- Data Driven - Neural Operators:
  - Fourier Neural Operator family (FNO, AFNO, Nested)
  - DeepONet
- GNNs:
  - MeshGraphNet
  - GraphCast
- Hybrid: PINO, ..

# HRSG FLUID ACCELERATED CORROSION SIMULATION — SIEMENS ENERGY

## Use Case

- Detecting and predicting point of corrosion in heat recovery steam generators (HRSGs)

## Challenges

- Using standard simulation to detect corrosion, it took SE at least couple of weeks, and the overall process took 14-16 weeks for every HRSG unit.

## Solution

- Using NVIDIA Modulus Physics-Informed Neural Network, SE simulates the corrosive effects of heat, water and other conditions on metal over time to fine-tune maintenance needs.
- SE can replicate and deploy HRSG plant digital twins worldwide with NVIDIA Omniverse.

## NVIDIA Solution Stack

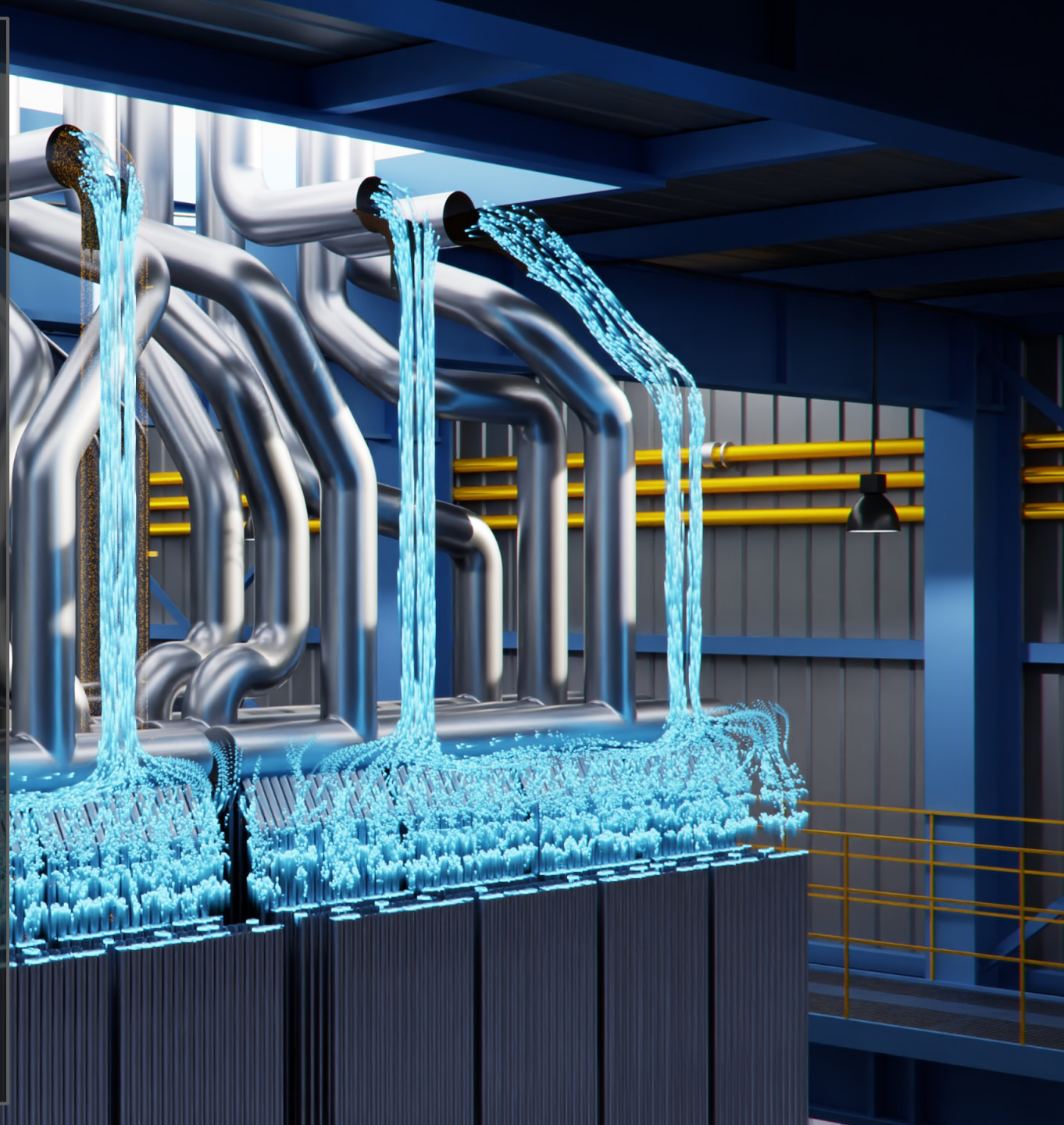
- Hardware: NVIDIA V100 & A100 Tensor Core GPUs
- Software: NVIDIA Modulus, NVIDIA Omniverse

## Outcome

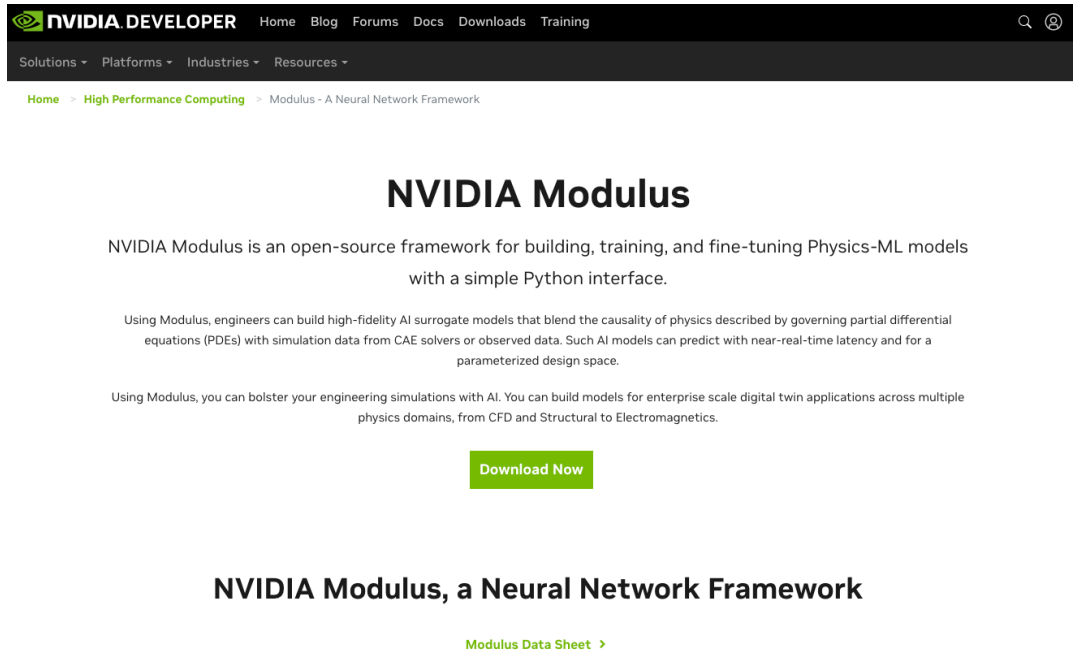
- 10,000X speed-up and inference in seconds can reduce downtime by 70%, saving the industry $1.7 billion annually

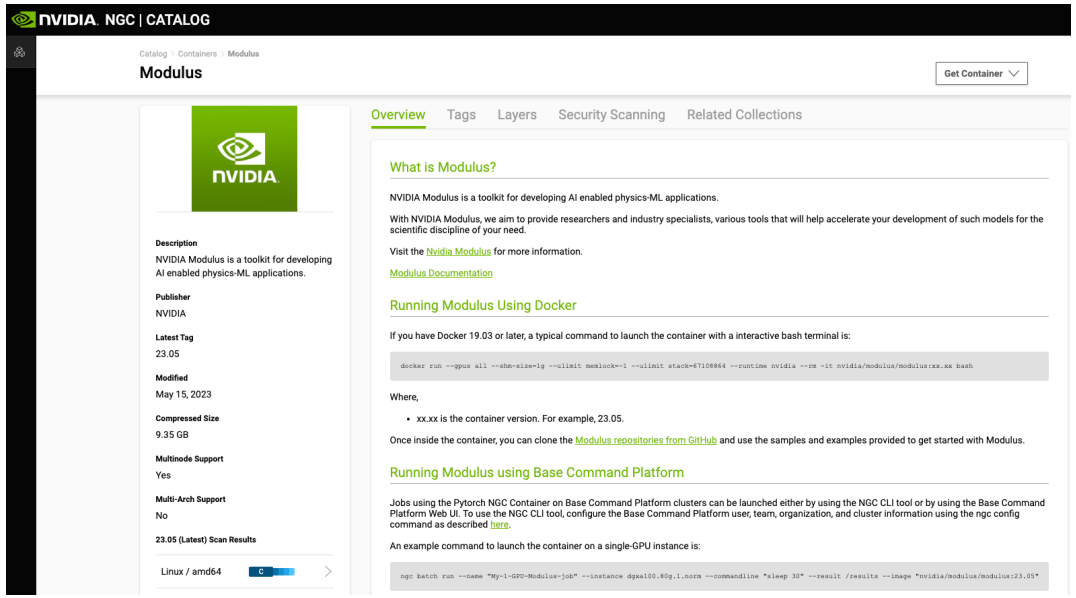
*Link to Demo*

NVIDIA SIEMENS energy



# Getting Started with NVIDIA Modulus




Go to https://developer.nvidia.com/modulus
and click “Download Now”

Follow the instructions described here:

```
$ docker run --gpus all ...
```

# Getting Started with NVIDIA Modulus

## Resources

- Latest Version : 23.05

- Product Page
  - https://developer.nvidia.com/modulus
- NGC Container Image
  - https://catalog.ngc.nvidia.com/orgs/nvidia/teams/modulus/containers/modulus
- Documentation
  - https://docs.nvidia.com/modulus/index.html
- GitHub
  - https://github.com/NVIDIA/modulus
- Japanese Page **<- NEW!**
  - https://developer.nvidia.com/ja-jp/modulus

- Resource Center
  - https://resources.nvidia.com/l/en-us-modulus-pathfactory-explore-page

# 「AI サロゲートモデルでシミュレーションを高速化する方法とは？」

## ソフトウェア ウェビナー シリーズ Vol. 3


現在、機械や建設での設計をはじめ幅広いものづくりにおいて、様々な領域で CAE が用いられています。しかしながら、技術の進歩により、シミュレーションを行うパラメータ数は増え続け、数値解析の結果の出力がただちに得られないケースも増えてきました。

そこでシミュレーションの一部を AI に置き換えるサロゲートモデルの活用が提唱されています。NVIDIA では Physics-ML を開発するためのフレームワーク NVIDIA Modulus をご提供しており、すでに一部の企業がこれを用いて風力発電機やプラントのシミュレーションを行っています。

本ウェビナーでは、Modulus によって何が可能になるのか？得意としている領域とは？ご利用になるための日本語の情報をご紹介いたします。

【日　程】2023 年 7 月 27 日 (木)  14:00 – 15:00 (60 分)

【対　象】大学や企業で CAE 活用を研究の方、CAE に Physics-ML 導入をご検討の方

【主　催】エヌビディア 合同会社

【参加費】無料 / 事前登録制

【配信方法】ON24 Simulive (Q&A はテキストにてライブでご対応いたします)

【お問合わせ】NVIDIA セミナー事務局 (NVIDIAJapanEvent@nvidia.com)

**丹愛彦**

エヌビディア合同会社 ソリューションアーキテクチャ＆エンジニアリング

シニアソリューションアーキテクト

製造業の研究所にて数値流体解析の研究開発に従事したのち、エヌビディア合同会社入社。現在は HPC 分野を中心に技術支援を担当。

**柴田良一 教授**

独立行政法人 国立高等専門学校機構

岐阜高等専門学校 建築学科

建設系機械系を含めた幅広いものづくりを対象に、オープンＣＡＥによる構造解析や流体解析、これらの連成解析を研究分野としている。さらに最近は、数値解析技術と人工知能技術との融合に興味を持ち、PINNsの可能性の検証を進めている。

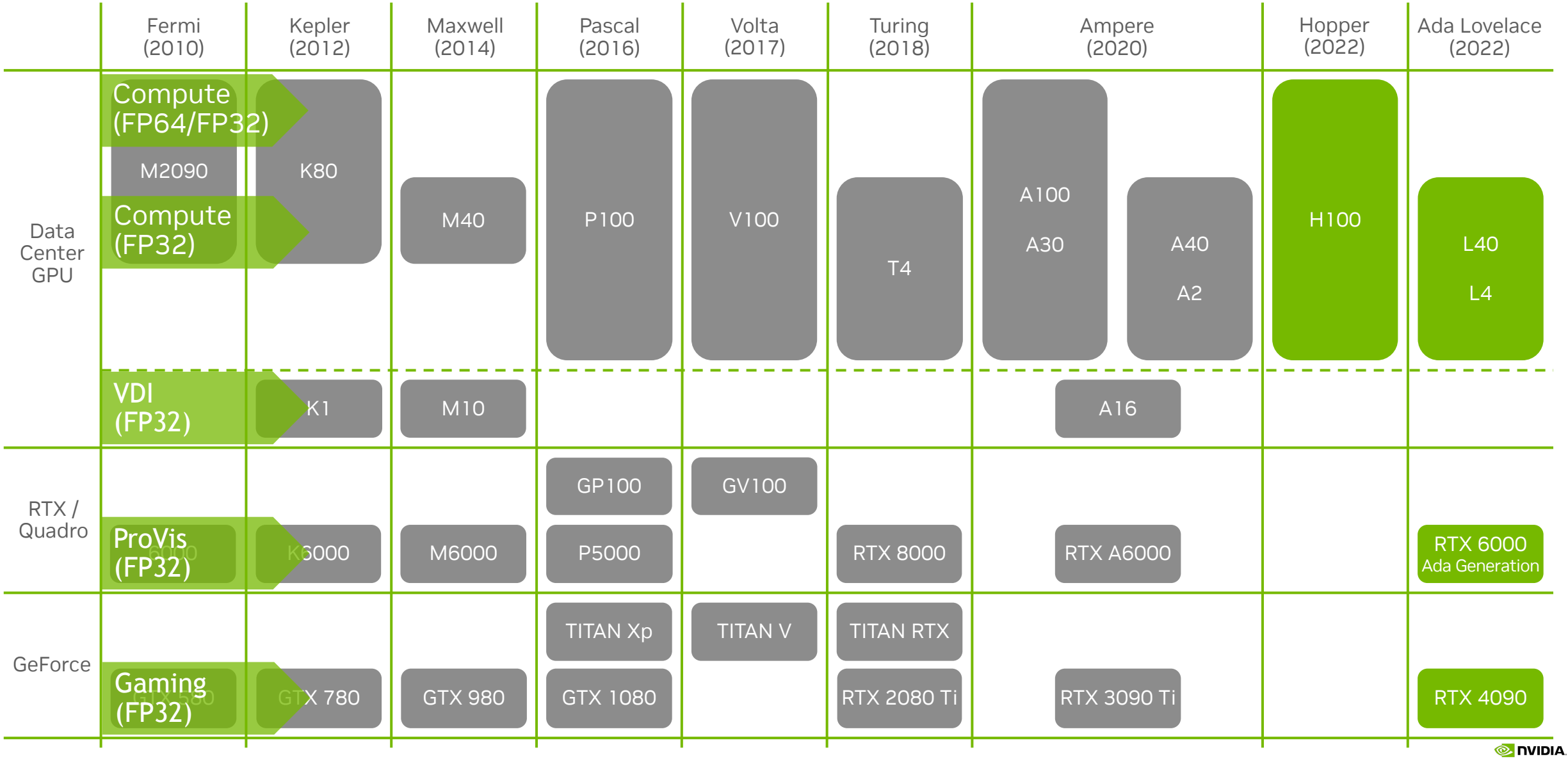
▼詳細はこちら▼

https://event.on24.com/wcc/r/4217685/BBE6855EE53BEA8B0AA3A67FB8FEC3B6/4725196

NVIDIA

# NVIDIA GPUs

# NVIDIA GPUs at a Glance



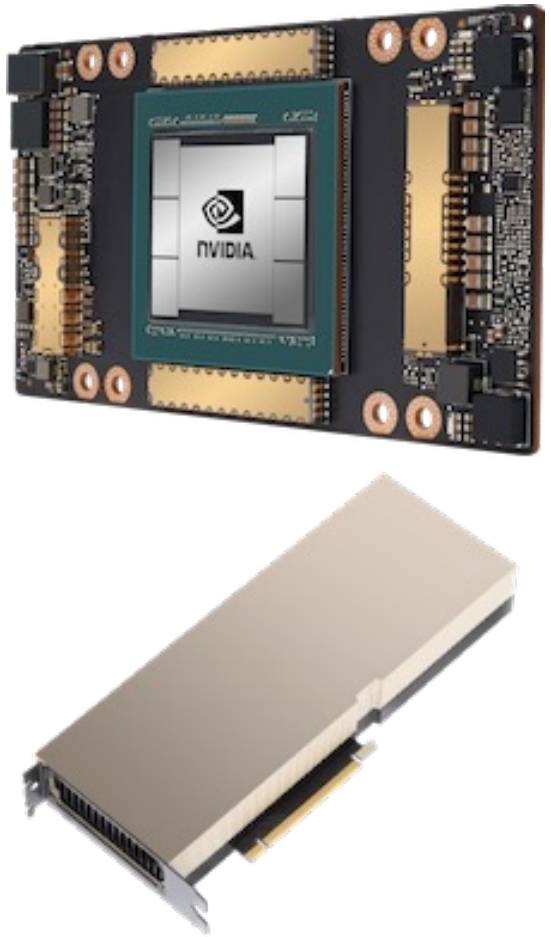
# NVIDIA H100 Tensor Core GPU

- HPC / DL Training / DL Inference / HPDA
  - Exascale HPC / LLM Inference
- Two form factors
  - SXM for HGX / PCIe
- FP64 / FP32
- 4th Generation Tensor Core
  - FP64 / TF32 / BF16 / FP16 / FP8 / INT8
- 4th Generation NVLink
  - 900GB/s (SXM) / 600GB/s up to 2 GPUs via NVLink Bridge (PCIe)
- High-Bandwidth Memory
  - 80GB HBM3 (SXM) / 80GB HBM2e (PCIe) / 188GB HBM3 (NVL; total)
- Transformer Engine
- 2nd Generation Multi-Instance GPU (MIG)

https://www.nvidia.com/en-us/data-center/h100/

# NVIDIA A100 Tensor Core GPU

- HPC / DL Training / DL Inference / HPDA
- Two form factors
  - SXM for HGX / PCIe
- FP64 / FP32
- 3rd Generation Tensor Core
  - FP64 / TF32 / BF16 / FP16 / INT8
- 3rd Generation NVLink
  - 600GB/s (SXM) / 600GB/s up to 2 GPUs via NVLink Bridge (PCIe)
- High-Bandwidth Memory
  - 80GB HBM2e
- Structural Sparsity
- Multi-Instance GPU (MIG)




https://www.nvidia.com/en-us/data-center/a100/

# NVIDIA H100 Gen-to-Gen Comparison

| | NVIDIA H100 | | | NVIDIA A100 | |
|---|---|---|---|---|---|
| GPU Architecture | Hopper | | | Ampere | |
| Form Factor | SXM (SXM5) | PCIe (PCIe Gen5) | NVL (2x PCIe Gen5) | SXM (SXM4) | PCIe (PCIe Gen4) |
| FP64 \| FP32 TFLOPS | 34 \| 67 | 26 \| 51 | 2x34 \| 2x67 | 9.7 \| 19.5 | |
| TF32 TC \| BF16 TC TFLOPS | 494* \| 989* | 378* \| 756* | 2x494* \| 2x989* | 156 \| 312 | |
| FP16 TC \| FP8 TC TFLOPS | 989* \| 1979* | 756* \| 1513* | 2x989* \| 2x1979* | 312 \| NA | |
| Memory | 80GB HBM3 | 80GB HBM2e | 2x94GB HBM3 | 80GB HBM2e | 80GB HBM2e |
| Memory Bandwidth | 3.35TB/s | 2TB/s | 2x3.9TB/s | 2039GB/s | 1935GB/s |
| Max TDP | Up to 700W (configurable) | 300-350W (configurable) | 2x 350-400W (configurable) | 400W | 300W |
| MIG | Up to 7 @10GB | Up to 7 @10GB | Up to 14 @12GB | Up to 7 @10GB | Up to 7 @10GB |
| Interconnect | NVLink: 900GB/s PCIe: 128GB/s | NVLink: 600GB/s PCIe: 128GB/s | NVLink: 600GB/s PCIe: 128GB/s | NVLink: 600GB/s PCIe: 64GB/s | NVLink: 600GB/s PCIe: 64GB/s |

** Double when using sparsity*

# NVIDIA RTX 6000 Ada Generation

- Professional Visualization
- 3rd Generation RT Core
- 4th Generation Tensor Core
- 48GB GDDR6 ECC memory
- PCIe Gen4 x16
- 4x DisplayPort 1.4
  - 4x 4096 x 2160 @ 120Hz
  - 4x 5120 x 2880 @ 60Hz
  - 2x 7680 x 4320 @ 60Hz
- Virtualization-Ready

https://www.nvidia.com/en-us/design-visualization/rtx-6000/

# NVIDIA RTX 6000 Ada Generation Gen-to-Gen Comparison

| | NVIDIA RTX 6000 Ada Generation | NVIDIA RTX A6000 |
|---|---|---|
| GPU Architecture | Ada Lovelace | Ampere |
| CUDA Cores | 18176 | 10752 |
| Tensor Cores | 568 | 336 |
| RT Cores | 142 | 84 |
| Memory Size | 48GB GDDR6 ECC | 48GB GDDR6 ECC |
| Memory Bandwidth | 960GB/s | 768GB/s |
| NVLink | Not supported | 2-way |
| Virtual Workstation | Yes | Yes |
| Media Acceleration | 3 NVENC (+1 AV1 encode)<br>3 NVDEC (+1 AV1 decode) | 1 NVENC<br>2 NVDEC (+1 AV1 decode) |
| Display Connections | 4x DP 1.4 | 4x DP 1.4 |
| Max TDP | 300W | 300W |
| Graphics Bus | PCIe Gen4 x16 | PCIe Gen4 x16 |

# NVIDIA Japan Social Media Directory

## Find Us Online!

- Twitter
  - https://twitter.com/NVIDIAJapan
- Facebook
  - https://www.facebook.com/NVIDIA.JP
- YouTube
  - https://www.youtube.com/user/NVIDIAJapan

- Twitter
  - https://twitter.com/NVIDIAAIJP
- Facebook
  - https://www.facebook.com/NVIDIAAI.JP

- Facebook
  - https://www.facebook.com/NVIDIANetworkingJapan

- Twitter
  - https://twitter.com/NVIDIAGeForceJP
- Facebook
  - https://www.facebook.com/NVIDIAGeForceJP
- Instagram
  - https://instagram.com/nvidiageforcejp
- YouTube
  - https://www.youtube.com/@nvidiageforcejapan44
- Twitch
  - https://www.twitch.tv/nvidiajapan

- Twitter
  - https://twitter.com/NVIDIAStudioJP
- Facebook
  - https://www.facebook.com/NVIDIAStudioJP
- Instagram
  - https://instagram.com/nvidiastudiojp
- YouTube
  - https://www.youtube.com/@nvidiastudiojapan1621


NVIDIA