GreenNLP-NVAITC NLP session @ University of Helsinki

Niki Loppi, PhD, Sr. AI/HPC Solutions Architect, NVIDIA Helsinki

Agenda

- Introduction – NVAITC (Niki Loppi)
- **10.05-10.30:** NVIDIA inference microservices and blueprints for NLP (Andrea Pilzer)
- **10.30-11.00:** NVIDIA NeMo overview (Giuseppe Fiameni)
- **11.00-12.00:** Live demo: Synthetic Data Generation & Fine Tuning with NeMo (Giuseppe Fiameni)
- **12.00-13:00:** open discussion/networking

Introduction + NVAITC Finland

NVIDIA AI Technology Center (NVAITC) Finland

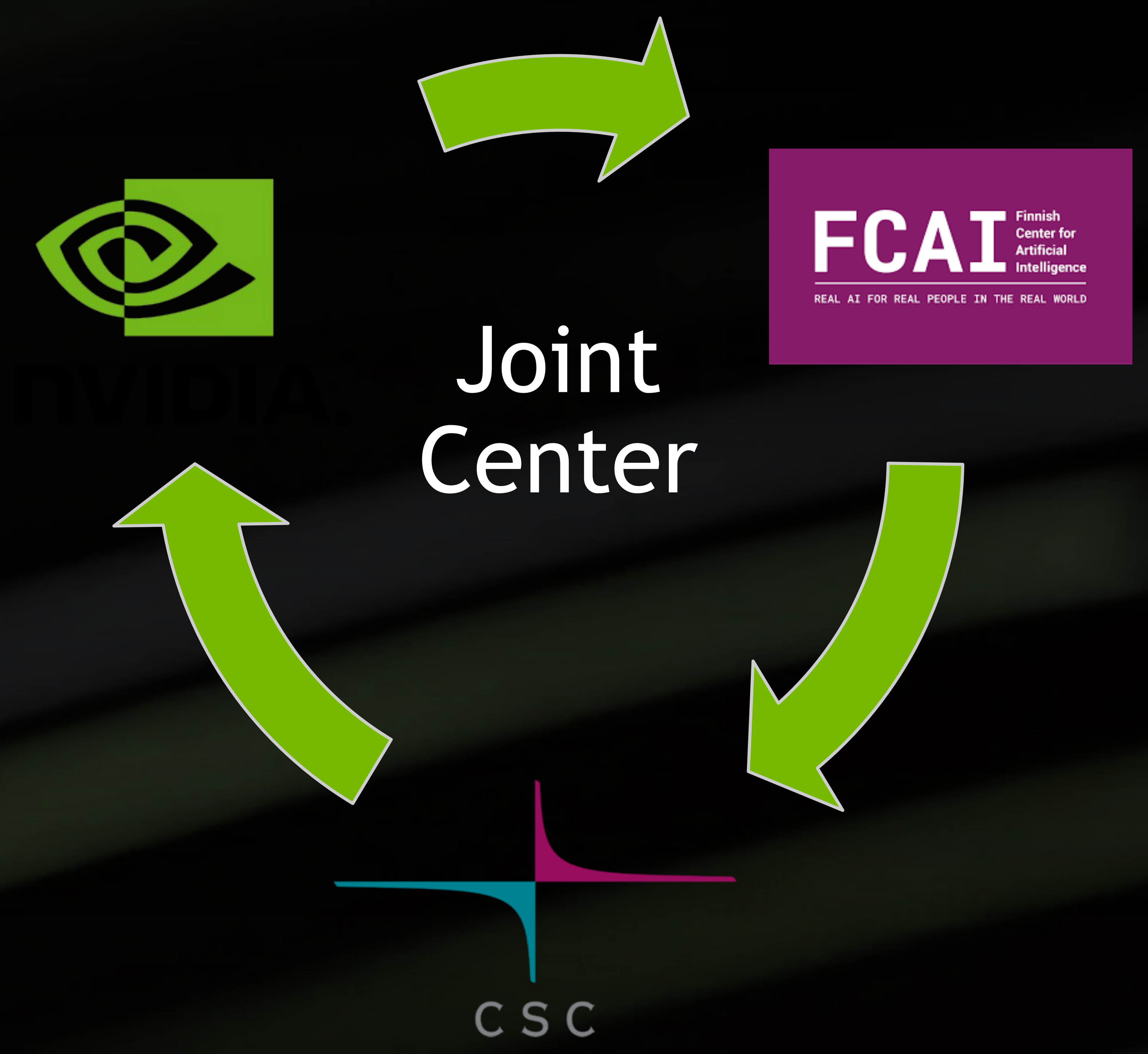
Enabling academics at all levels to do their research more efficiently

- A joint research center between FCAI/Finnish Universities, CSC and NVIDIA
- 26 projects, 25 publications, 450+ people trained

- **What we can offer:**
    - NVAITC scientist/engineer to collaborate on your problems related to GPU-accelerate AI or HPC research
    - Help to harness the accelerated compute capability at CSC
    - A point of contact to access global NVIDIA expertise/advise
    - NDA is in place and IP stays with the researchers

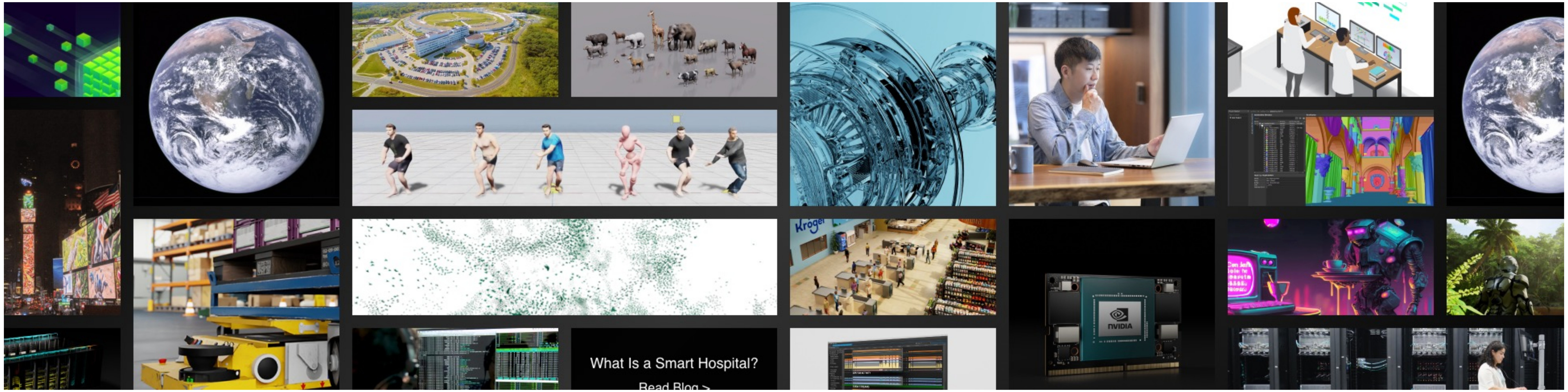
- **What we expect in return:**
    - Depending on the level of our contribution, an acknowledgement or co-authorship in the next publication

- **Whether you are a beginner or an expert, whether you want to collaborate or just want to ask a quick question. Please email Niki Loppi [nloppi@nvidia.com] or visit our website [fcai.fi/nvaitc]**



NVIDIA Developer & DLI

Join the Dev program and claim 1 free course from the following catalogues



Join the NVIDIA Developer Program

- https://developer.nvidia.com/developer-program
- https://sp-events.courses.nvidia.com/AIDaysEU
- https://sp-events.courses.nvidia.com/AIFactoriesEU

GTC Paris

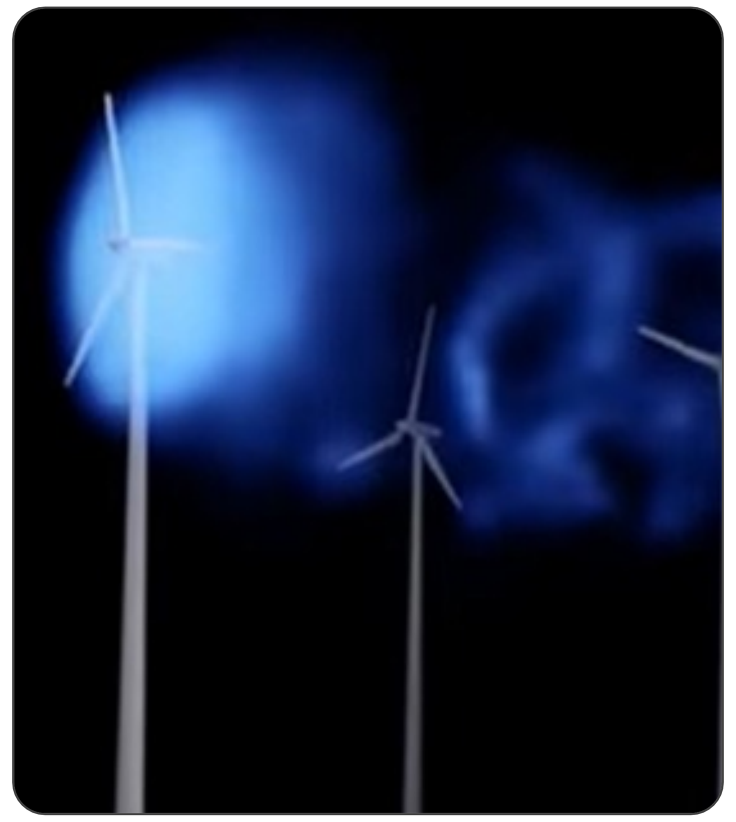
@VivaTech 2025

- In-person
- Presentation and training sessions for all technical levels
- Recorded sessions will be made available at NVIDIA on-demand

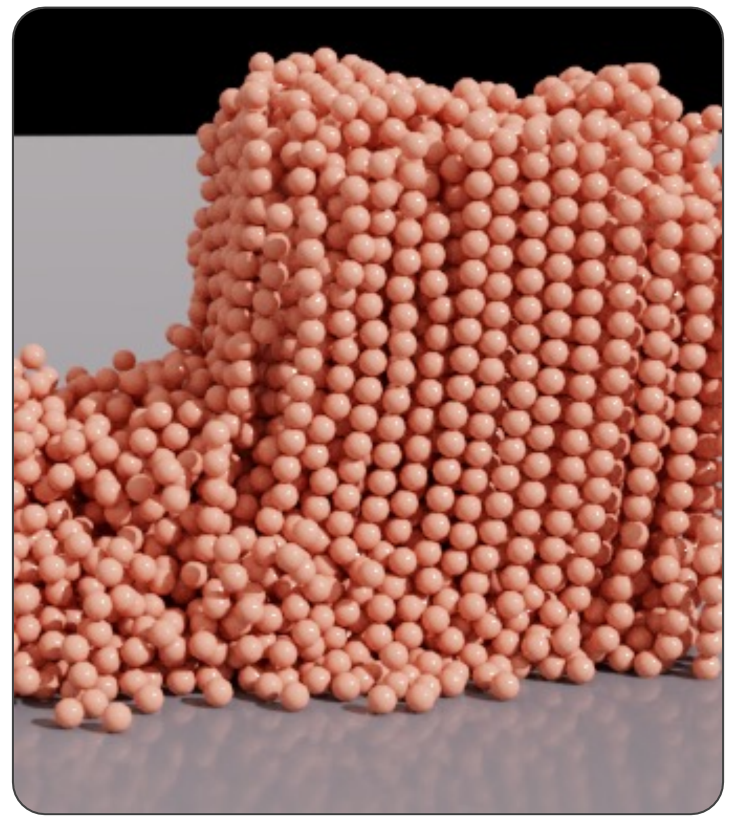
NVIDIA AI & HPC Platform

CUDA-X Accelerates Every Industry

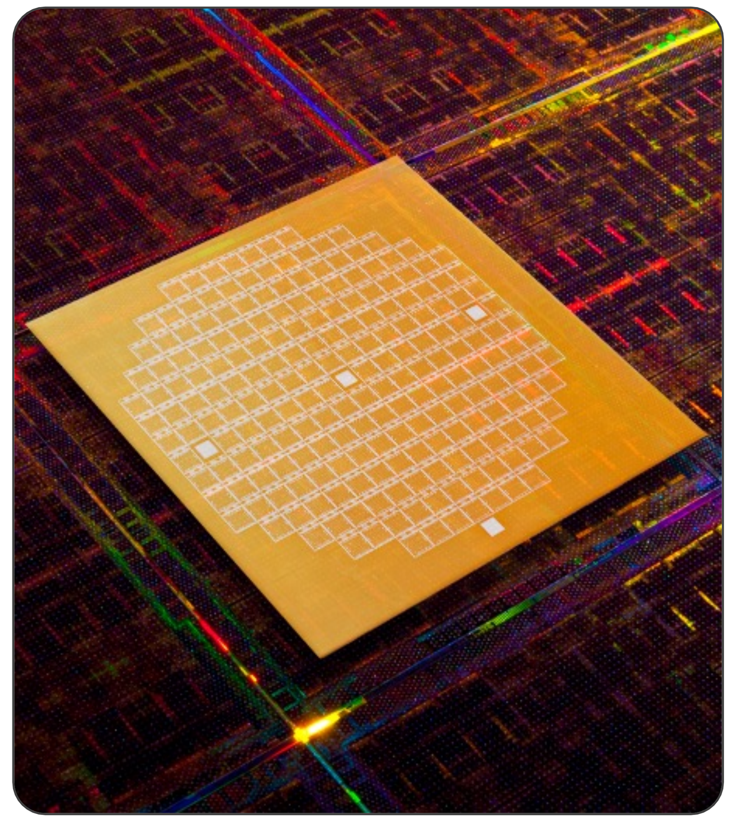
cuDSS
CAE



PhysicsNeMo
AI Physics

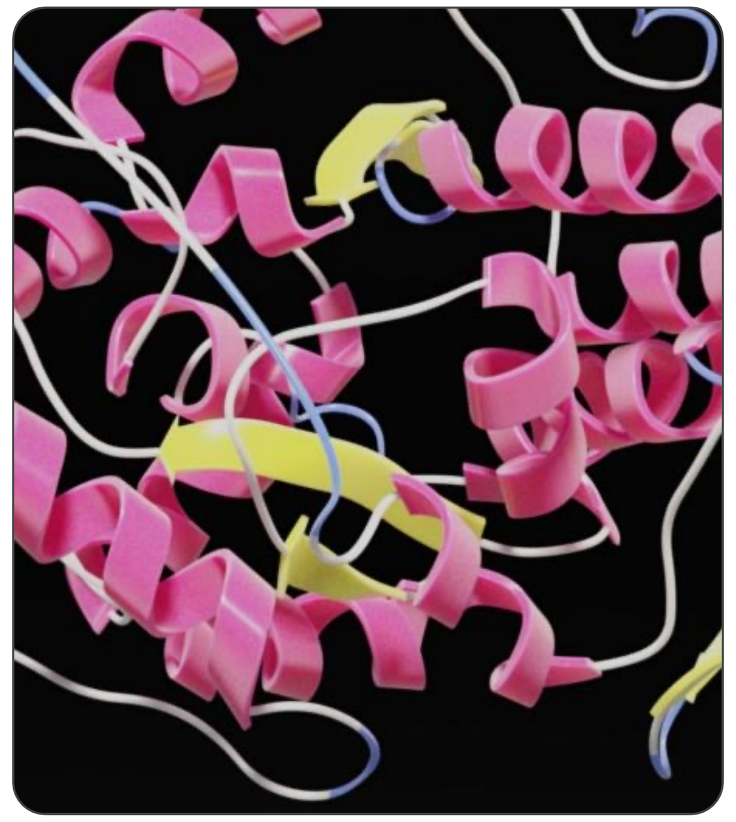


Warp
Physical Simulation

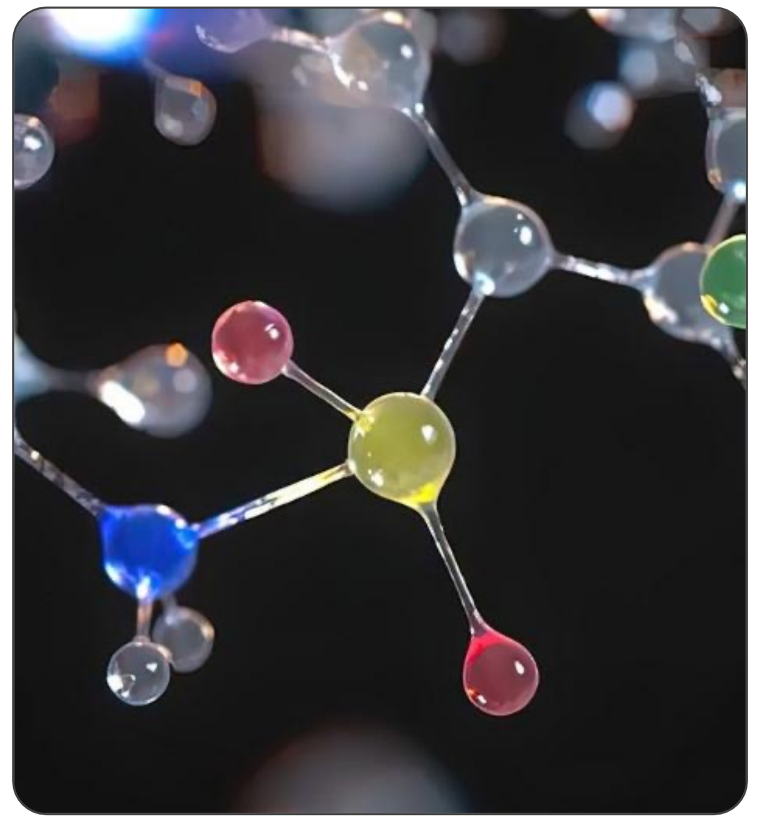


cuLitho
Computational Lithography

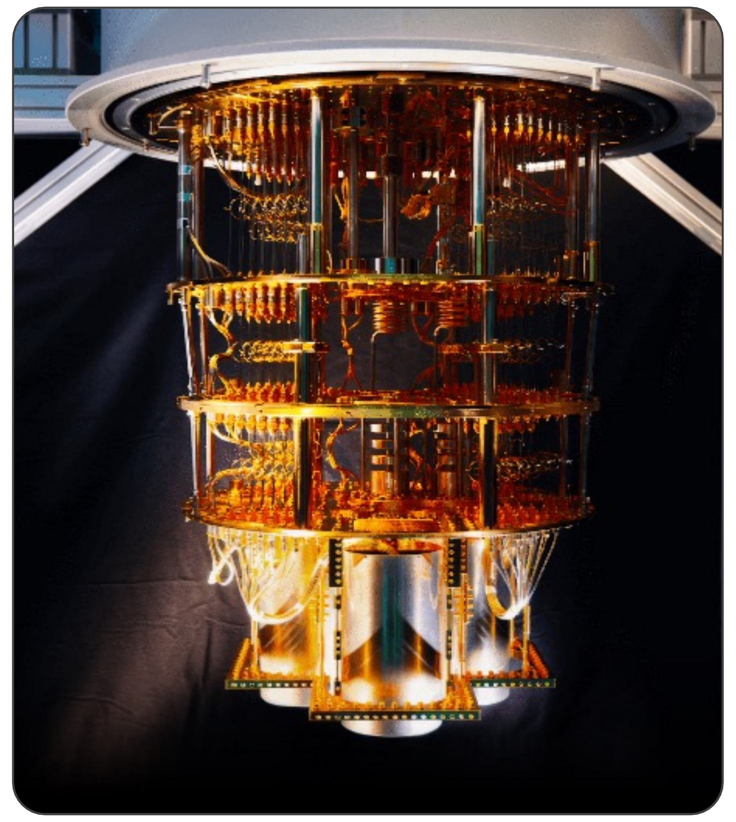
ALCHEMI
AI Materials Science



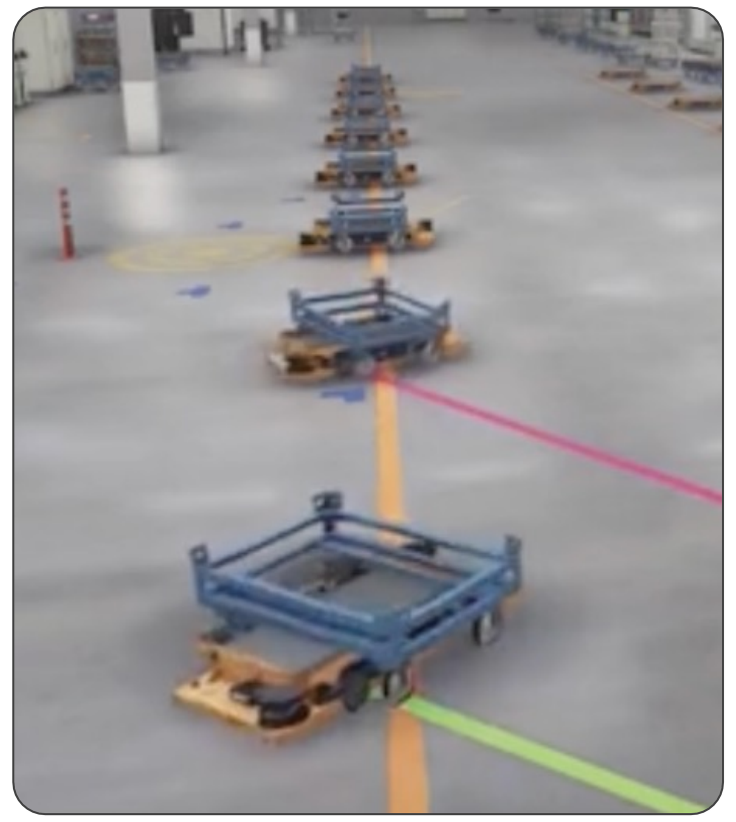
cuEquivariance
Drug & Materials Discovery



Parabricks
Gene Sequencing



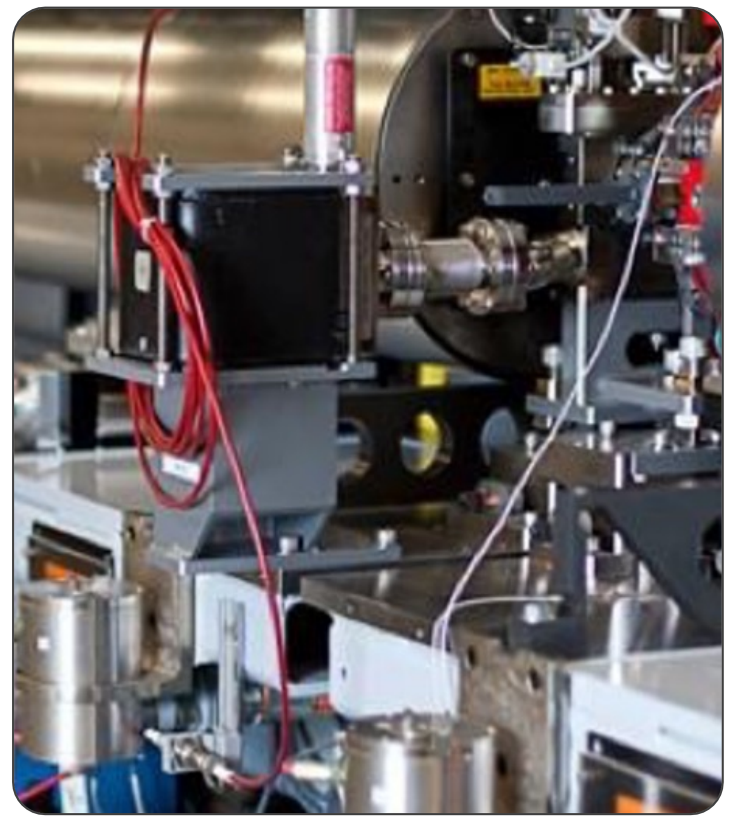
CUDA-Q
Quantum Computing



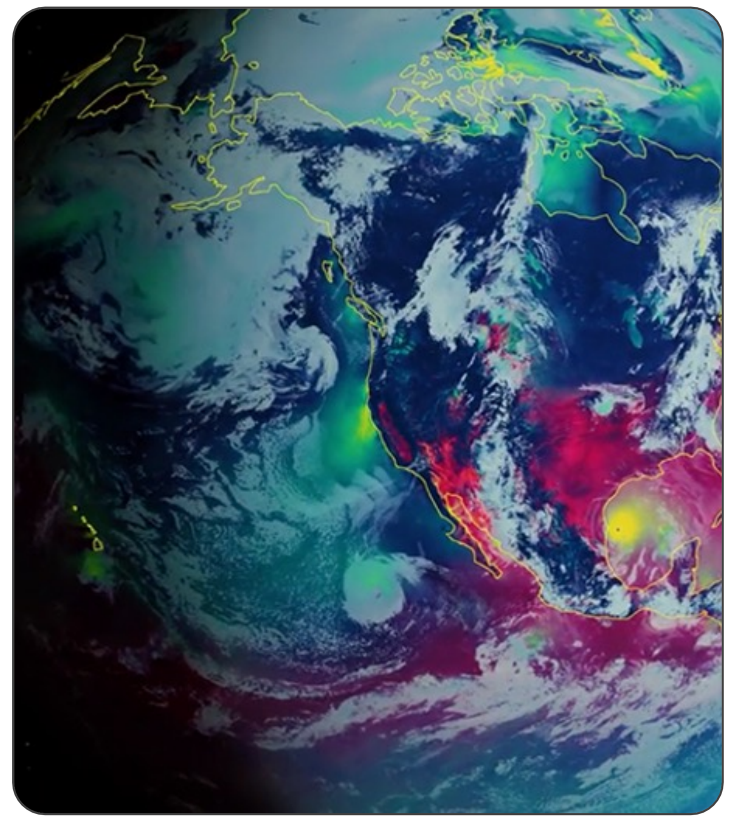
cuOpt
Decision Optimization

cuDF
Data Processing

cuPyNumeric
Numerical Computing

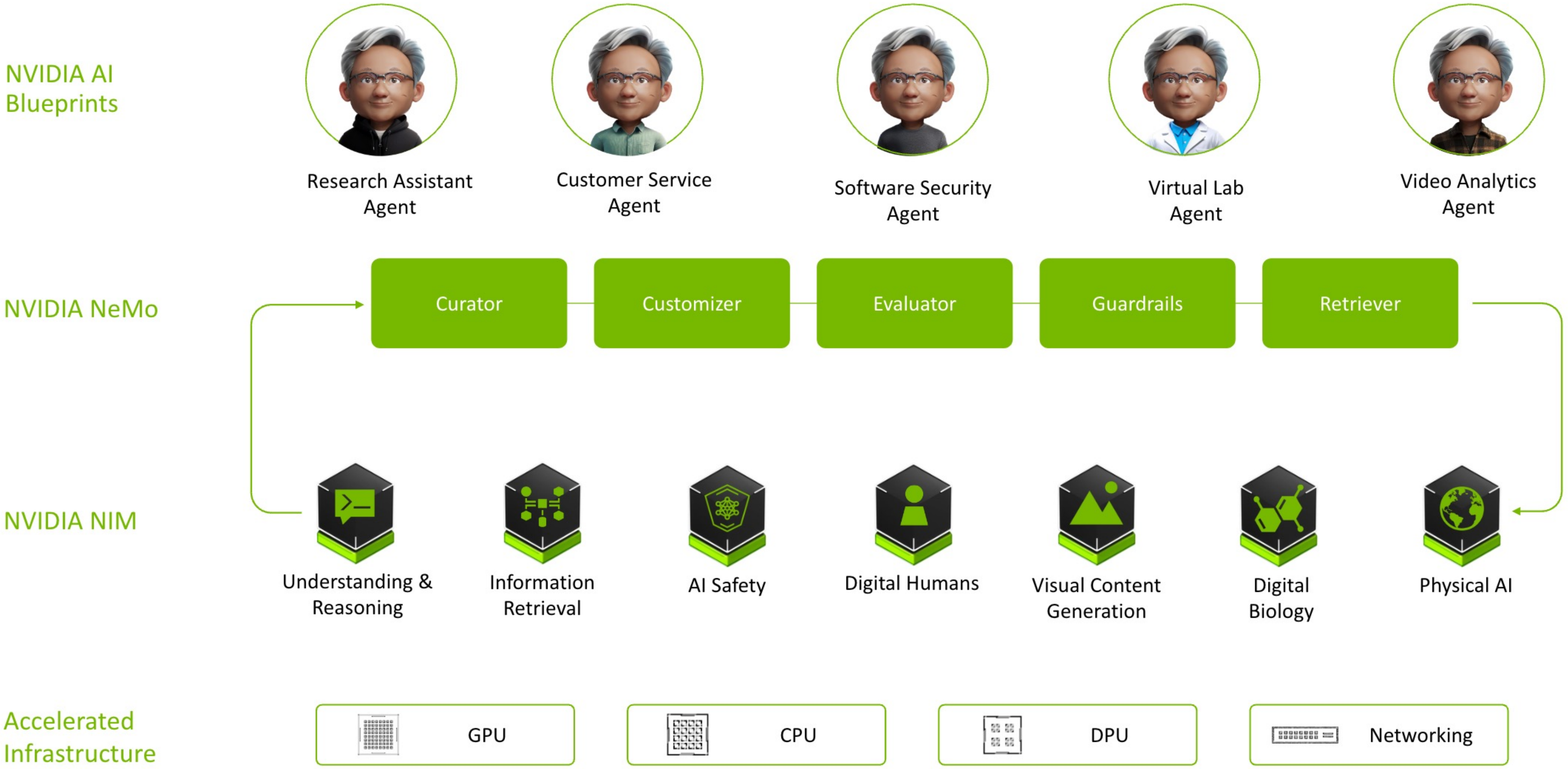


Holoscan
Edge HPC



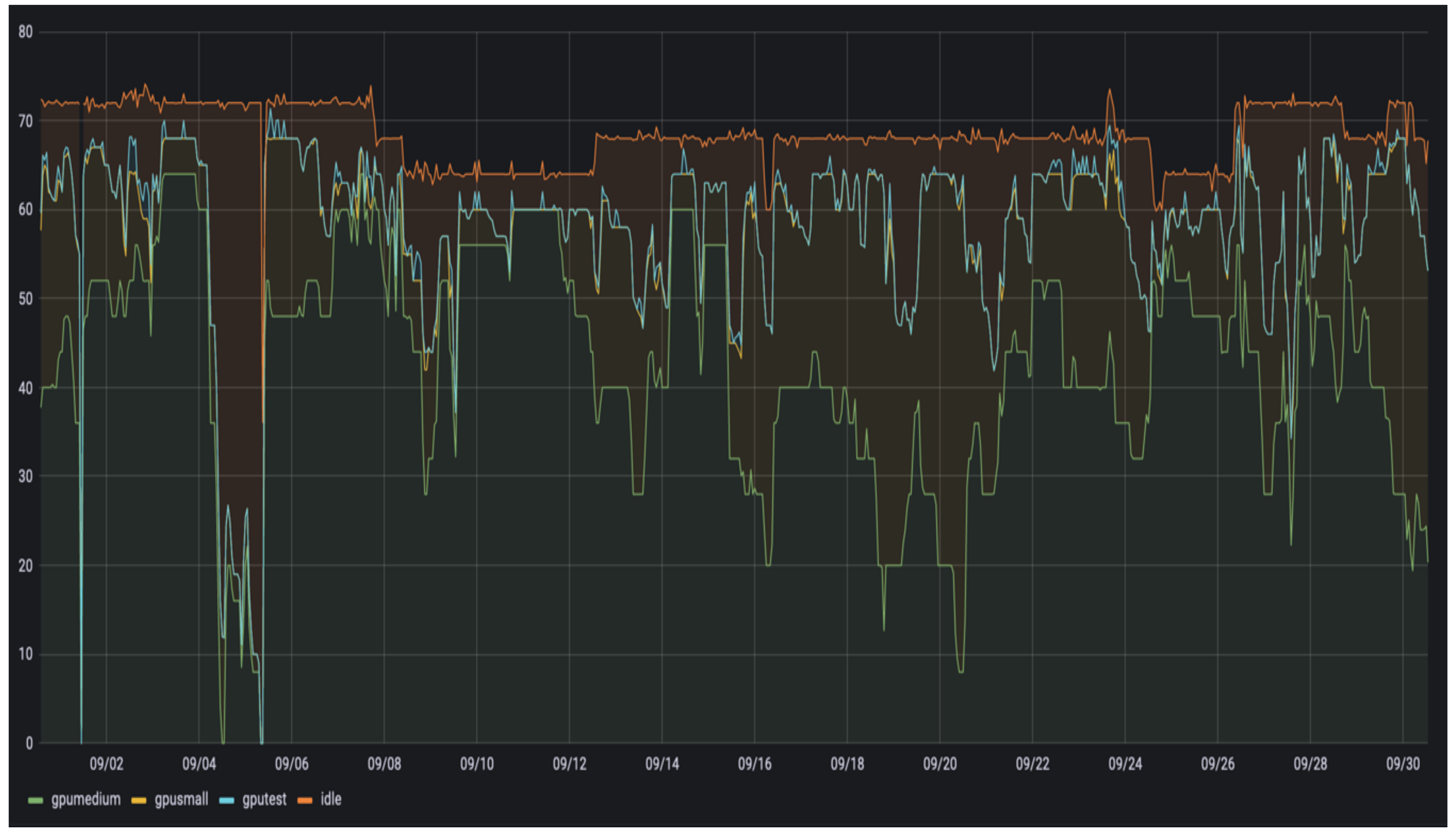
Earth-2
Weather Analytics

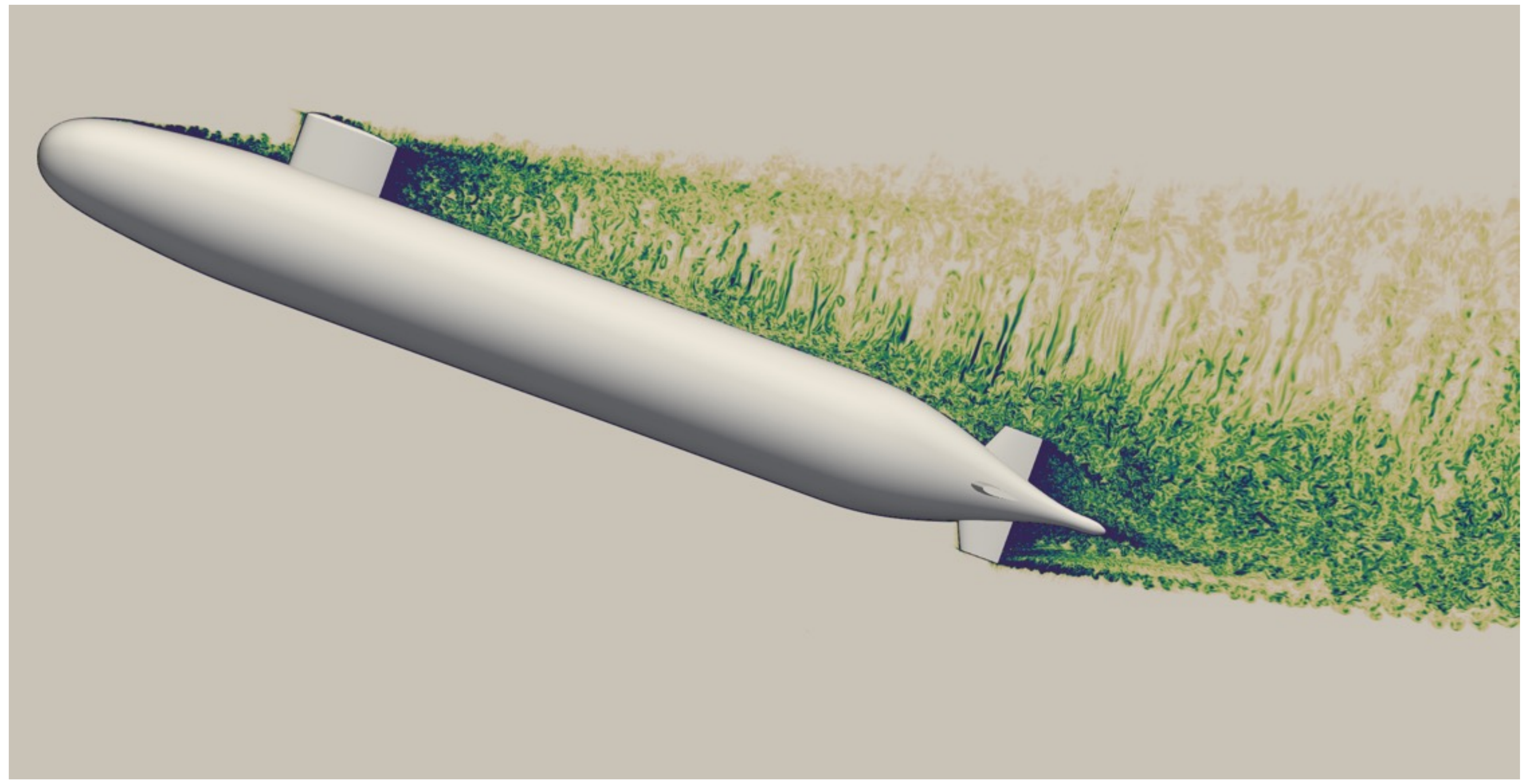
NVIDIA Provides the Building Blocks for Agentic AI



NVAITC Finland by-the-numbers

- Engaged **87+ PIs** from **8 institutions**
- **26** collaboration projects executed across a range domains
  - up to 80 GPU jobs on CSC's Puhti-AI and Mahti-AI
- **25 publications** – NeurIPS, WACV, PoPETs
- More than **450 researchers trained live**
  - Unis of Applied Science program with CSC
  - Fundamentals of DL @ Arcada, CUDA Python @ Oulu
- Finland center spearheading HPC+AI across the NVAITC program
- Roihu with 528 Nvidia GH200 GPUs coming online this year!





Motivation

Addressing common use-cases in research

- The increasing size and complexity of AI models are presenting new technical and engineering challenges, particularly when working with large, multi-format datasets.
- Many of these challenges are recurring across various research groups and often fall outside the typical expertise of AI researchers.
- To address this, we are developing a comprehensive set of playbooks and recipes. These resources are specifically meant to provide our collaborators with a robust foundation for their work to build on.

https://github.com/NVIDIA-AI-Technology-Center

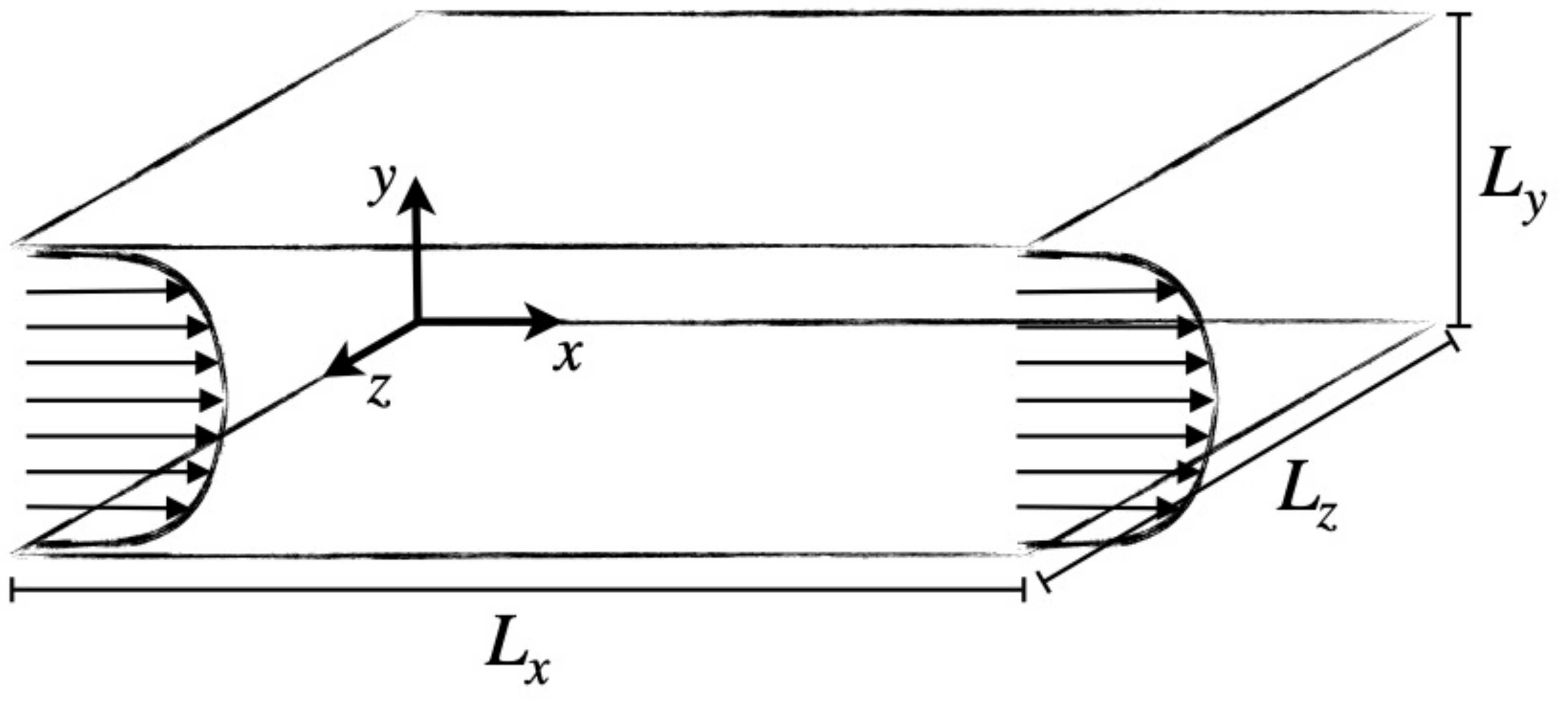
Current list of playbooks

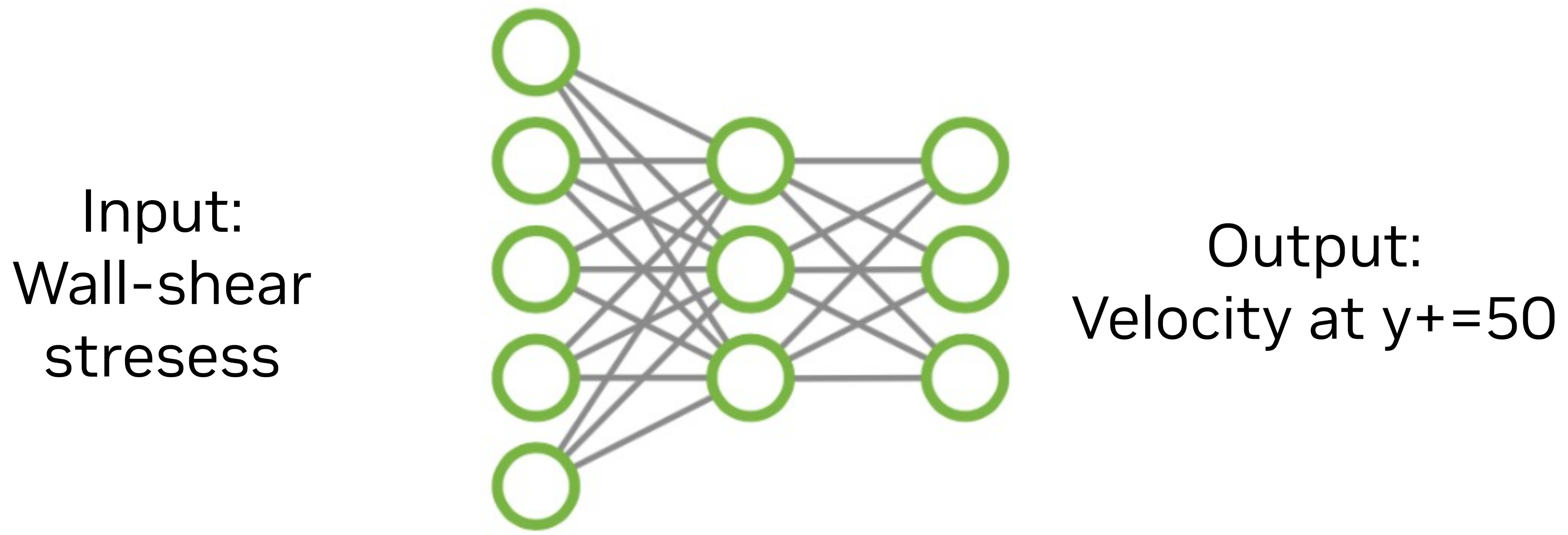
https://github.com/NVIDIA-AI-Technology-Center

1. From Synthetic Data Generation to Model Fine Tuning
2. Multi-scale RAG pipeline
3. Synthetic Data Generation
4. Collaborative data aggregation and 3D visualization of digital twins
5. Online Training of deep learning for HPC Apps
6. NSIGHT Systems profiling on Grace-Hopper
7. Fortran to Python Code Modernization
8. Accelerated Video Processing and Model Training

Online DL training/inference for HPC apps - playboook

(Strongly coupled numerical simulation and AI training/inference)





Test case setup

Now that the TorchFort-enabled CaNS code has been build, the last step is to run the test case. First, let's copy the previously generated CNN model to the case folder.

```
!docker exec <id> cp /files/python_model/cans_fcn.pt /files/reconstruction_case
```

CaNS code will read runtime parameters from an input file 'input.nml' located in the case folder. This example comes with two CaNS input files: 1. input.stats.nml which is used to run the case up to 2000 time units without training to develop the flow and measure the normalisation statistics and 2. input.train.nml to start training from 2000 time units onwards. Let's first run the first part. Please note that we have to add the path to CaNS dependencies to LD_LIBRARY_PATH environment variable as they are not set in the container build.

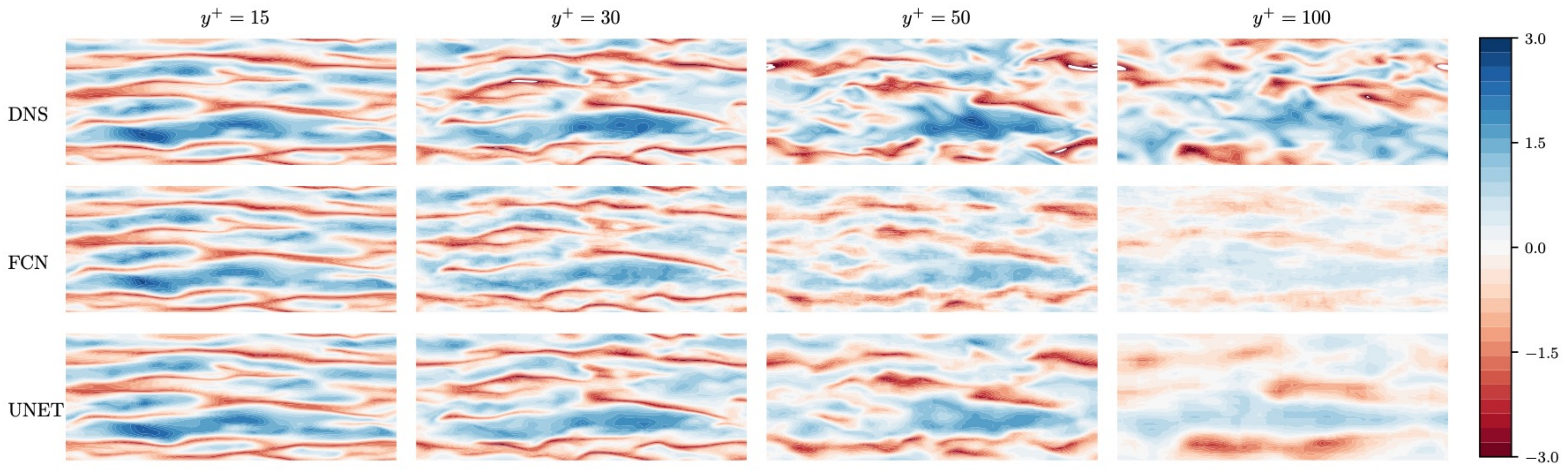
```
!docker exec <id> cp /files/reconstruction_case/input.stats.nml /files/reconstruction_case/input.nml
!docker exec <id> /bin/bash -c 'export LD_LIBRARY_PATH=/opt/CaNS/dependencies/cuDecomp/build/lib:${LD_LIBRARY_PATH} && \
                              cd /files/reconstruction_case && \
                              mpirun -np 1 --allow-run-as-root --bind-to none /opt/CaNS/run/cans'
```

Now that we have a developed checkpoint (to exclude the initial transients) for the simulation and an estimate of the flow statistics for normalisation, we can start the training. We will overwrite the input.nml with it's training counterpart input.train.nml which specifies all necessary runtime parameters for traing e.g.

```
trainbs = 32
nsamples_train = 3200
nsamples_val = 320
```

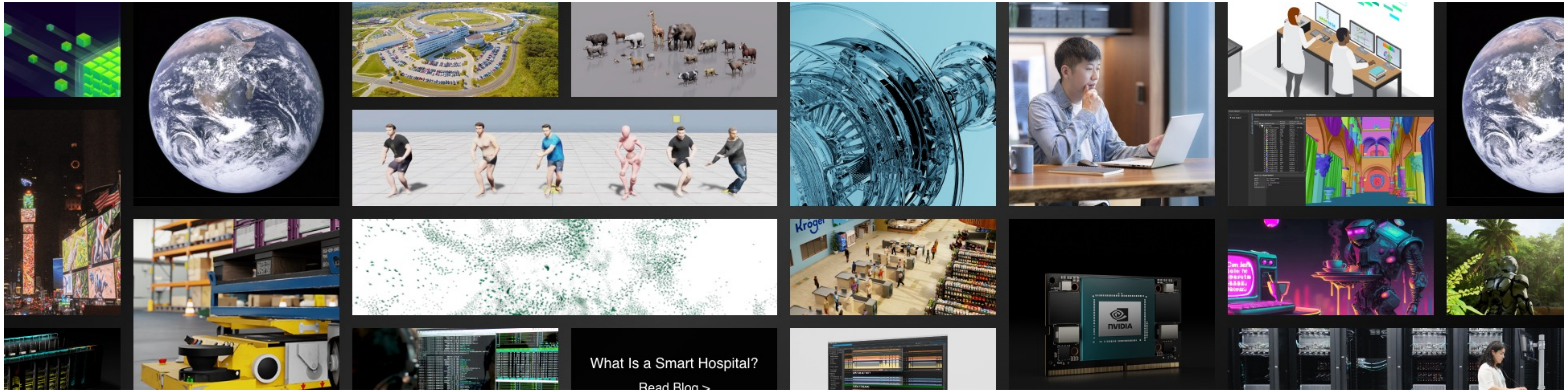
which means that we will undertake nsamples_train/(trainbs * num_gpus) number of training steps, after which we will take nsamples_val/(trainbs * num_gpus) validation steps. Training checkpoints and inference results are save after each validation epoch. For the purposes of this demo the training step will proceed for a total of 500 steps. However, for best results the training should run considerably longer.

```
!docker exec <id> cp /files/reconstruction_case/input.train.nml /files/reconstruction_case/input.nml
!docker exec <id> /bin/bash -c 'export LD_LIBRARY_PATH=/opt/CaNS/dependencies/cuDecomp/build/lib:${LD_LIBRARY_PATH} && \
                              cd /files/reconstruction_case && \
                              mpirun -np 1 --allow-run-as-root --bind-to none /opt/CaNS/run/cans'
```



NVIDIA Developer & DLI

Join the Dev program and claim 1 free course from the following catalogues



Join the NVIDIA Developer Program

- https://developer.nvidia.com/developer-program
- https://sp-events.courses.nvidia.com/AIDaysEU
- https://sp-events.courses.nvidia.com/AIFactoriesEU

GTC Paris

@VivaTech 2025

- In-person
- Presentation and training sessions for all technical levels
- Recorded sessions will be made available at NVIDIA on-demand

Kiitos!