
# CUDA PLATFORM OVERVIEW

# CUDA PLATFORM

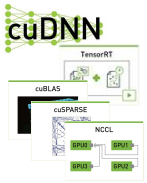
**People**

*GPU Users* | *Domain Specialists* | *Problem Specialists* | *New Developers and Optimization Experts*

**Domain Descriptors**


**Applications**

**Frameworks**




**Libraries**


**Directives and Standard Languages**

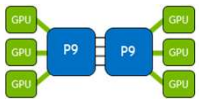

**Extended Standard Languages**

*Ease of use* ⟶ *Specialized Performance*

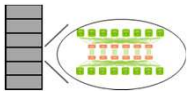
**Systems**



GeForce
Quadro
Tesla PCI




Supercomputing



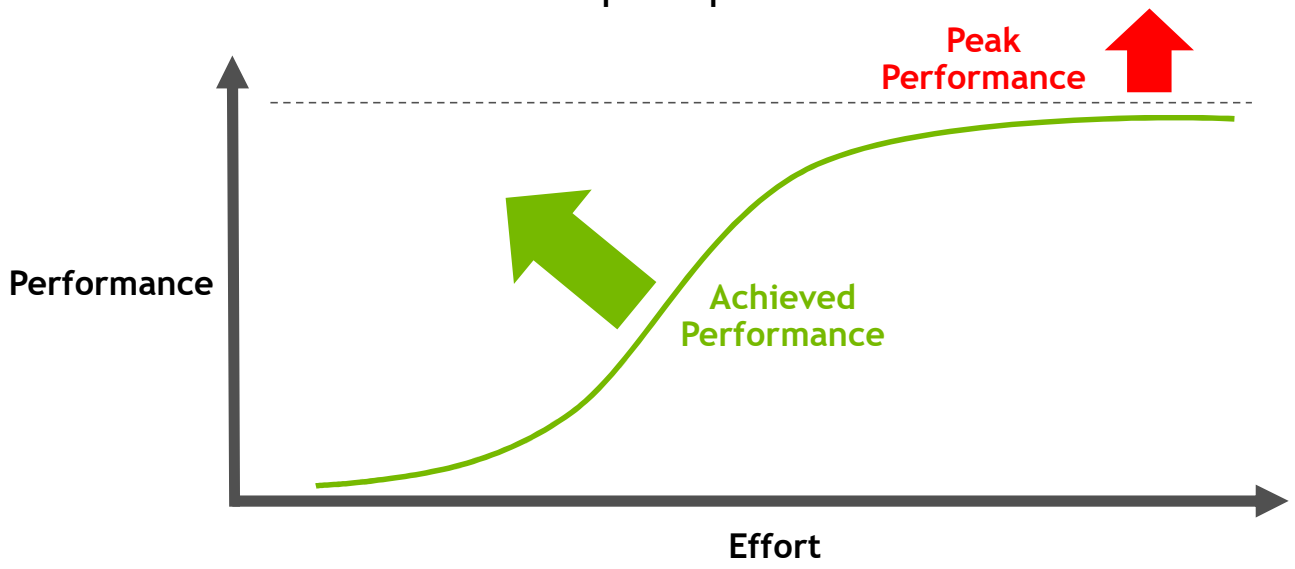
Cloud + HGX/DGX

**Foundation**

Programming Model, GPU Architecture, System Architecture, Tools

# CUDA PLATFORM GOALS

- Every layer of the CUDA Platform has two fundamental goals:
  - Maximize achievable peak performance
  - Minimize effort to achieve peak performance




## CUDA Platform Layers

Applications

Frameworks & DSLs

Libraries

Directives and
Standard Languages

Extended
Standard Languages

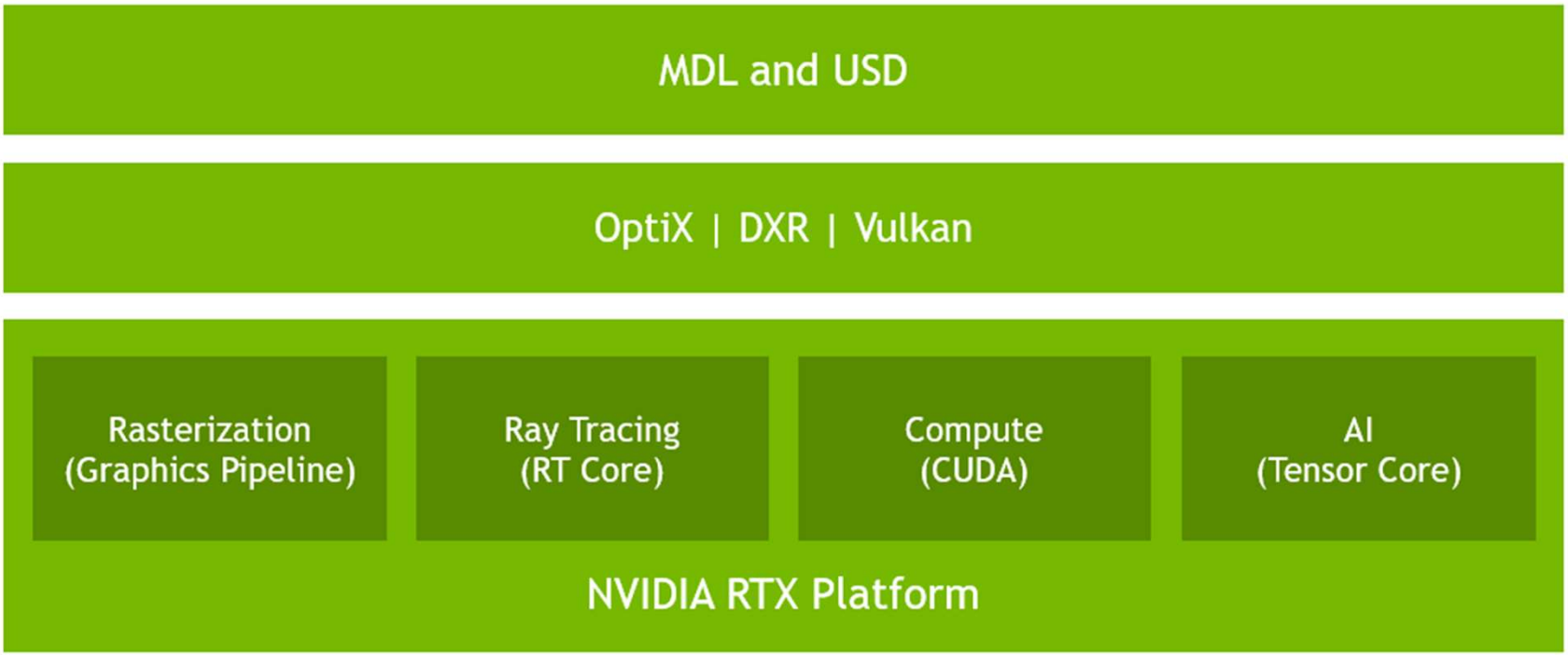
System Architecture

GPU Architecture

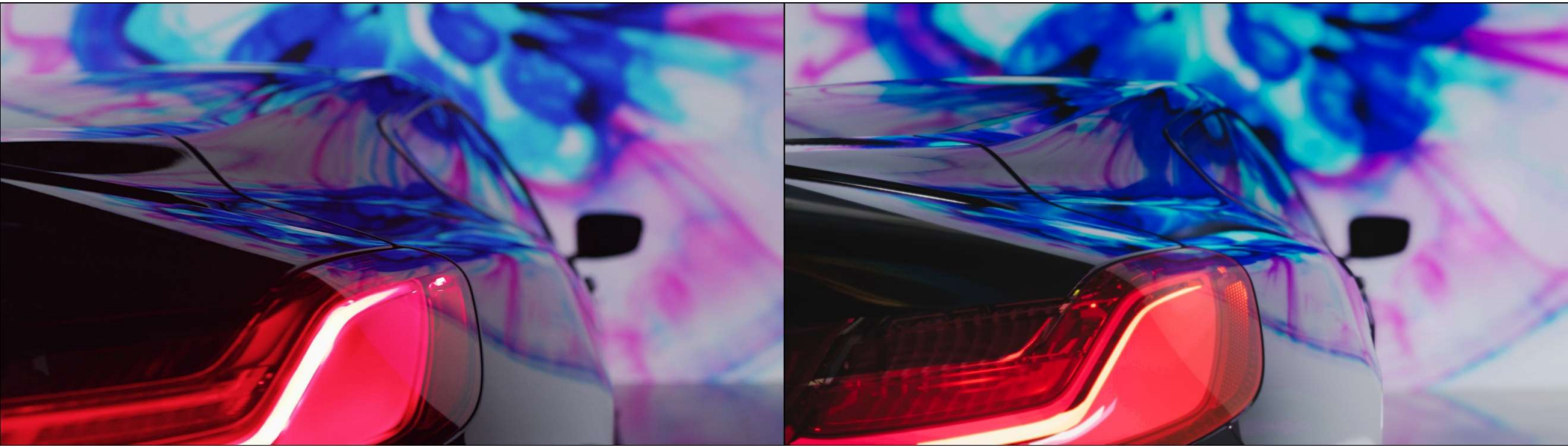
# NVIDIA CUDA-X
# GPU-ACCELERATED COMPUTING PLATFORMS

# RTX PLATFORM

## Accelerating the creative process



# RTX REAL TIME RAY TRACING

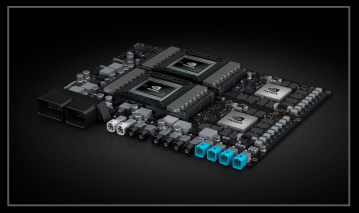


# THE DRIVE INITIATIVE

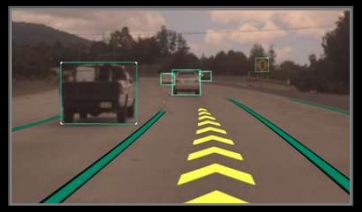
DGX Saturn V

Constellation

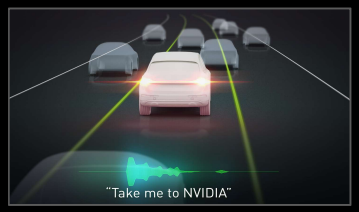
Xavier



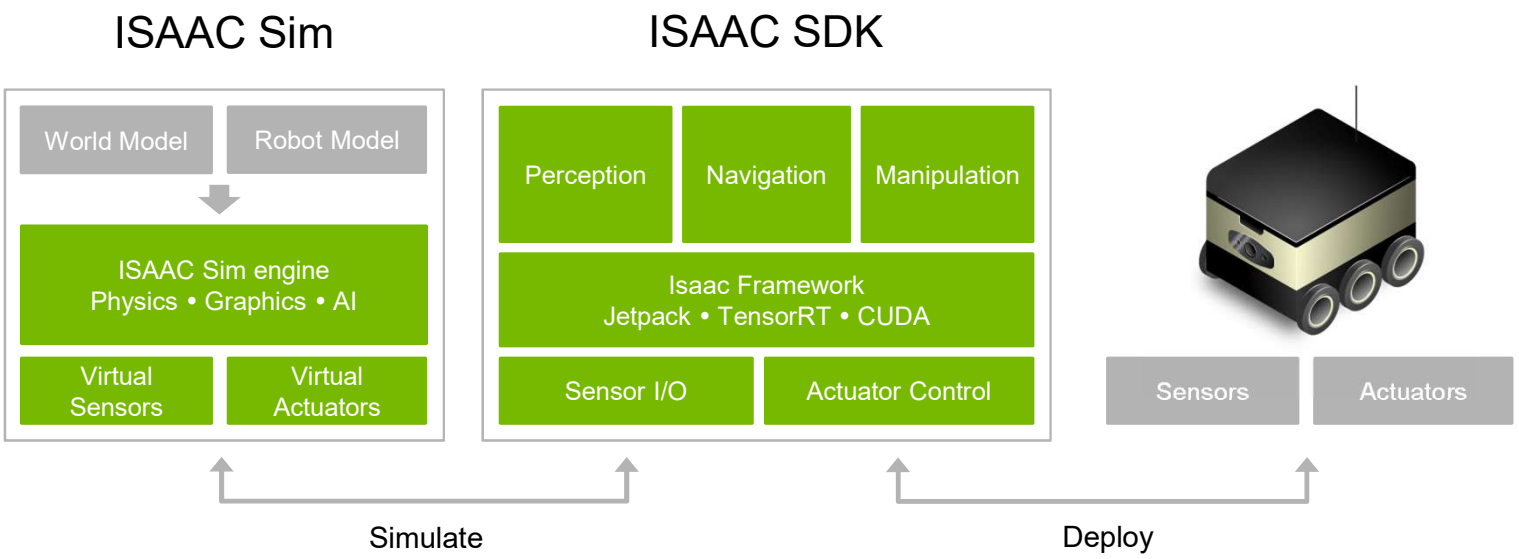
DRIVE AV

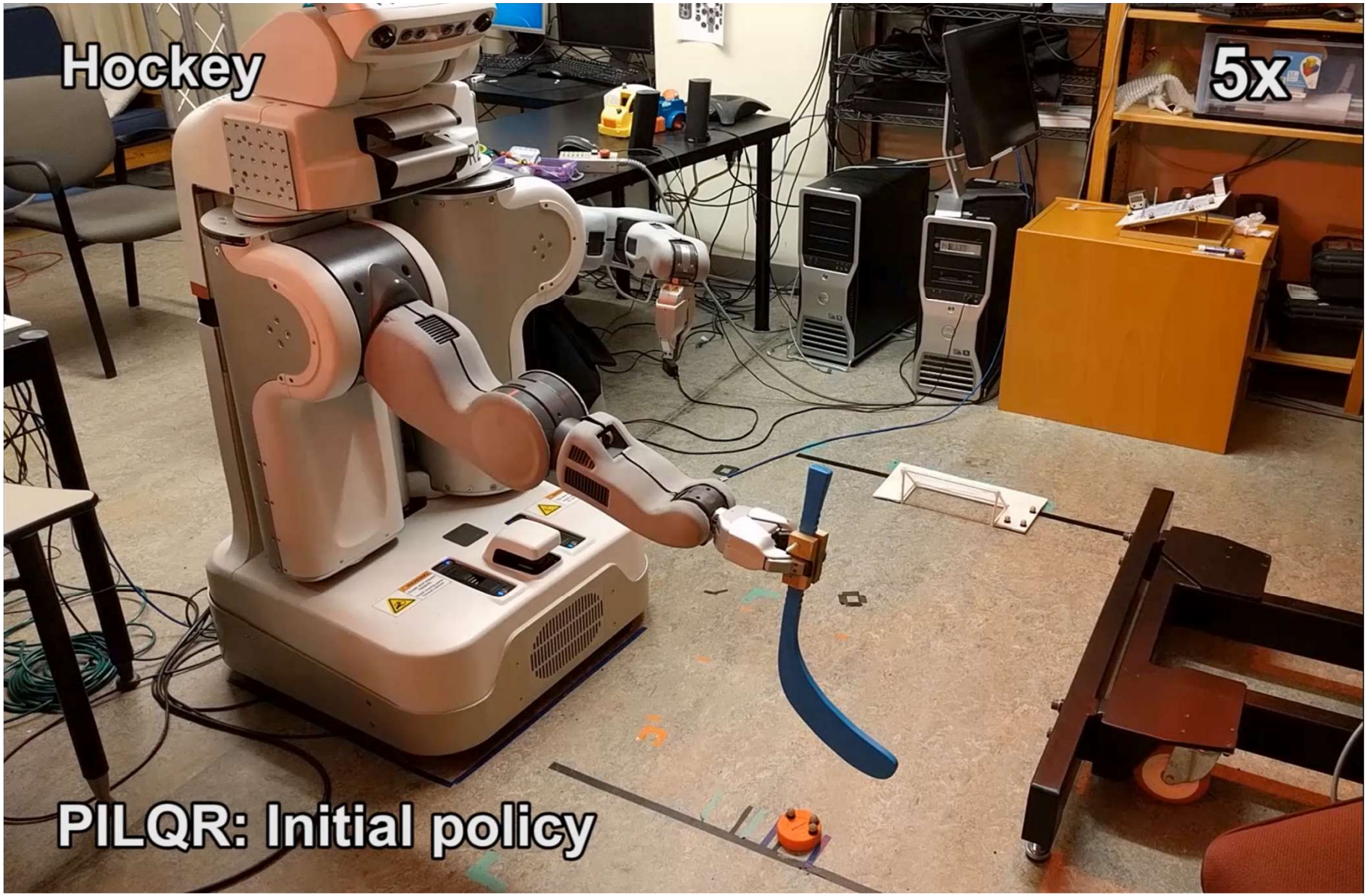


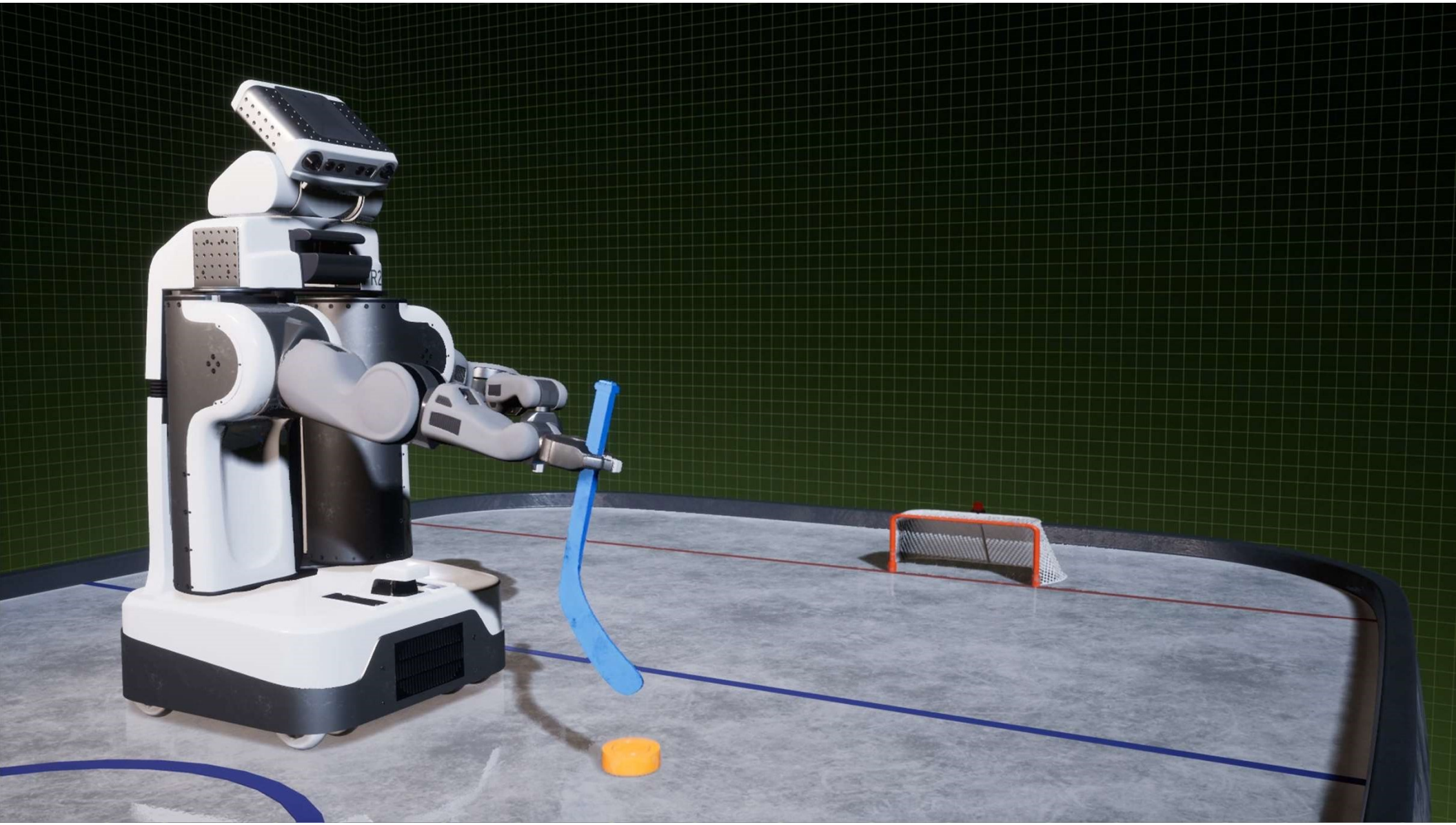
DRIVE IX



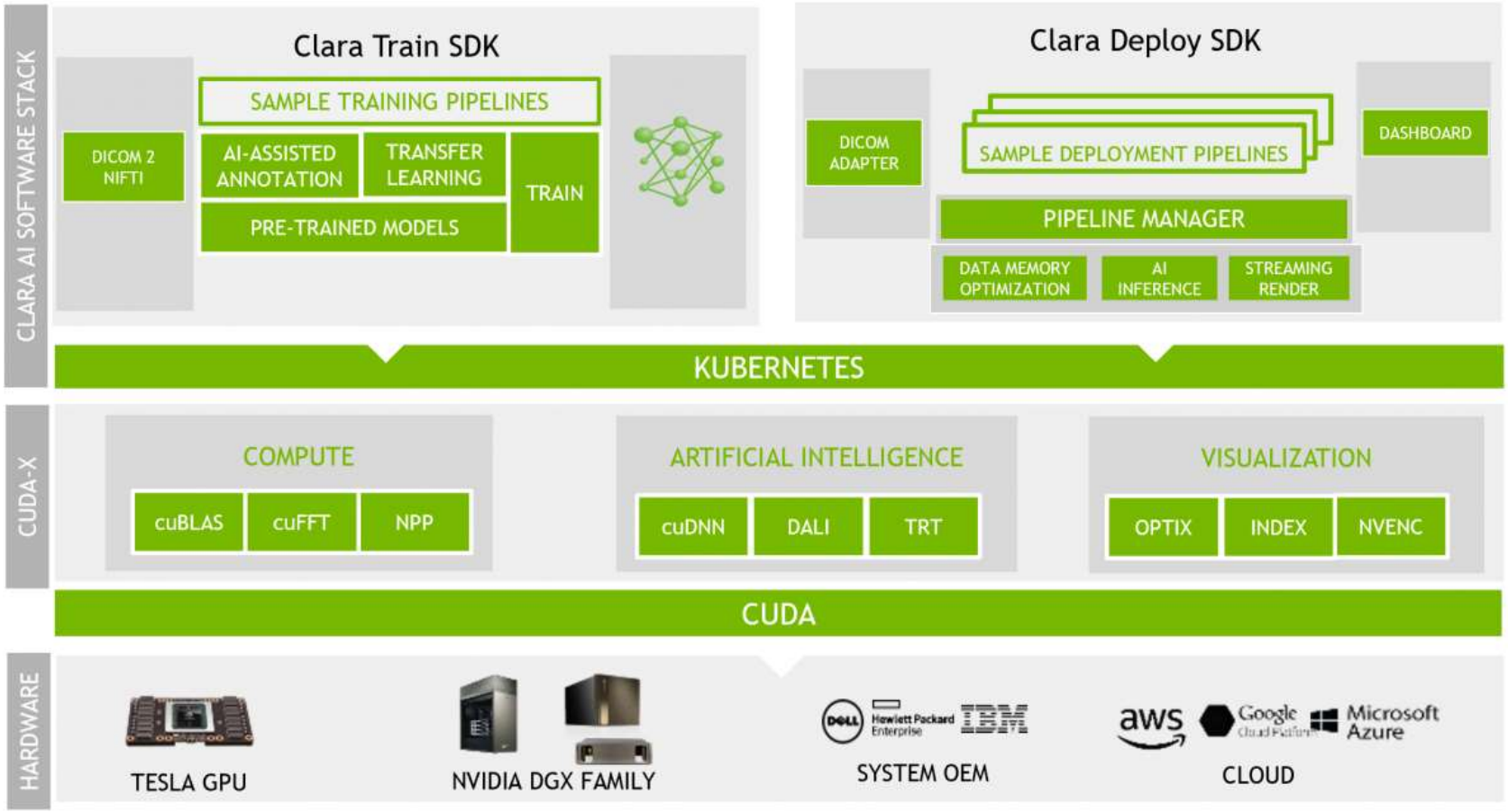

# ISAAC FOR ROBOTICS



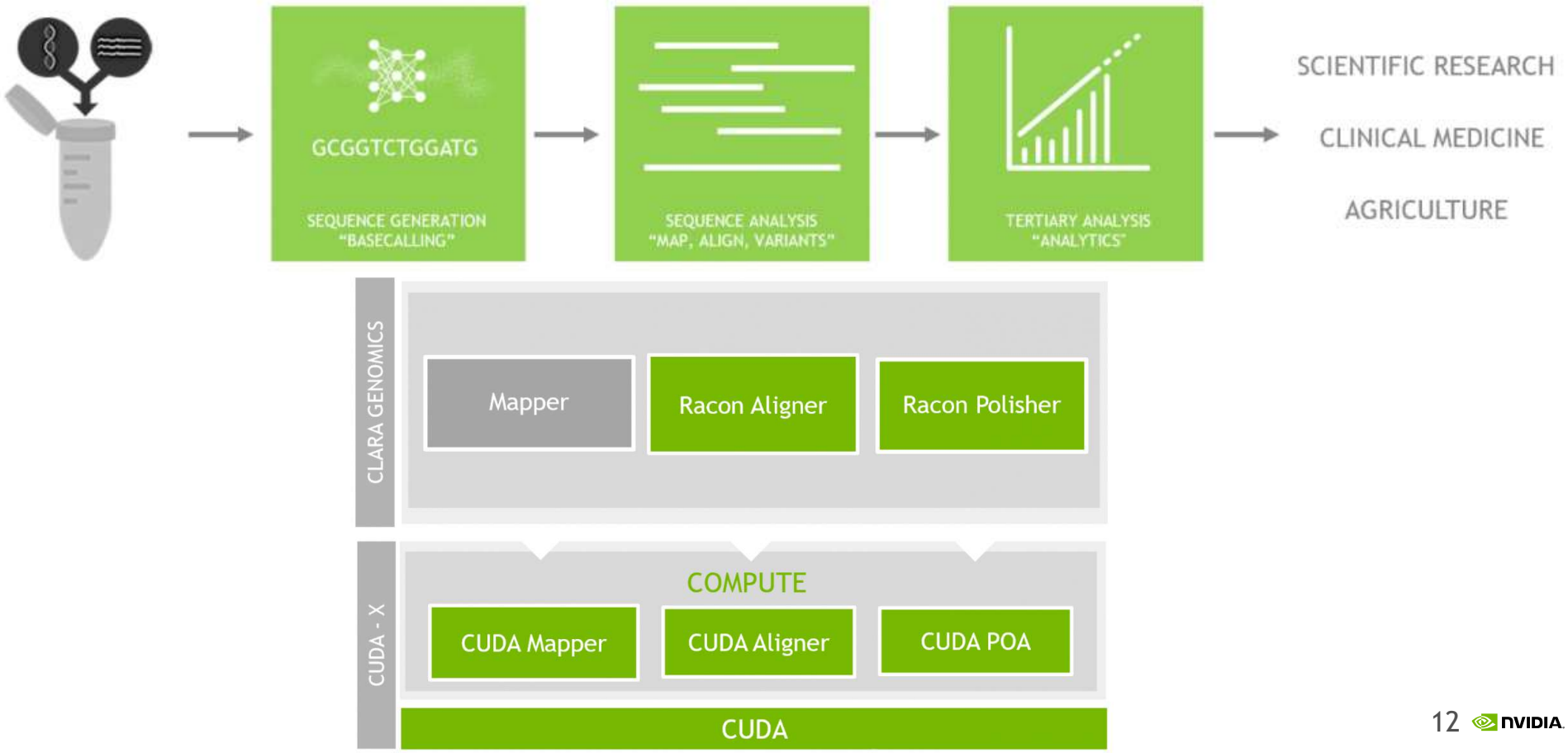

Hockey
5x
PILQR: Initial policy



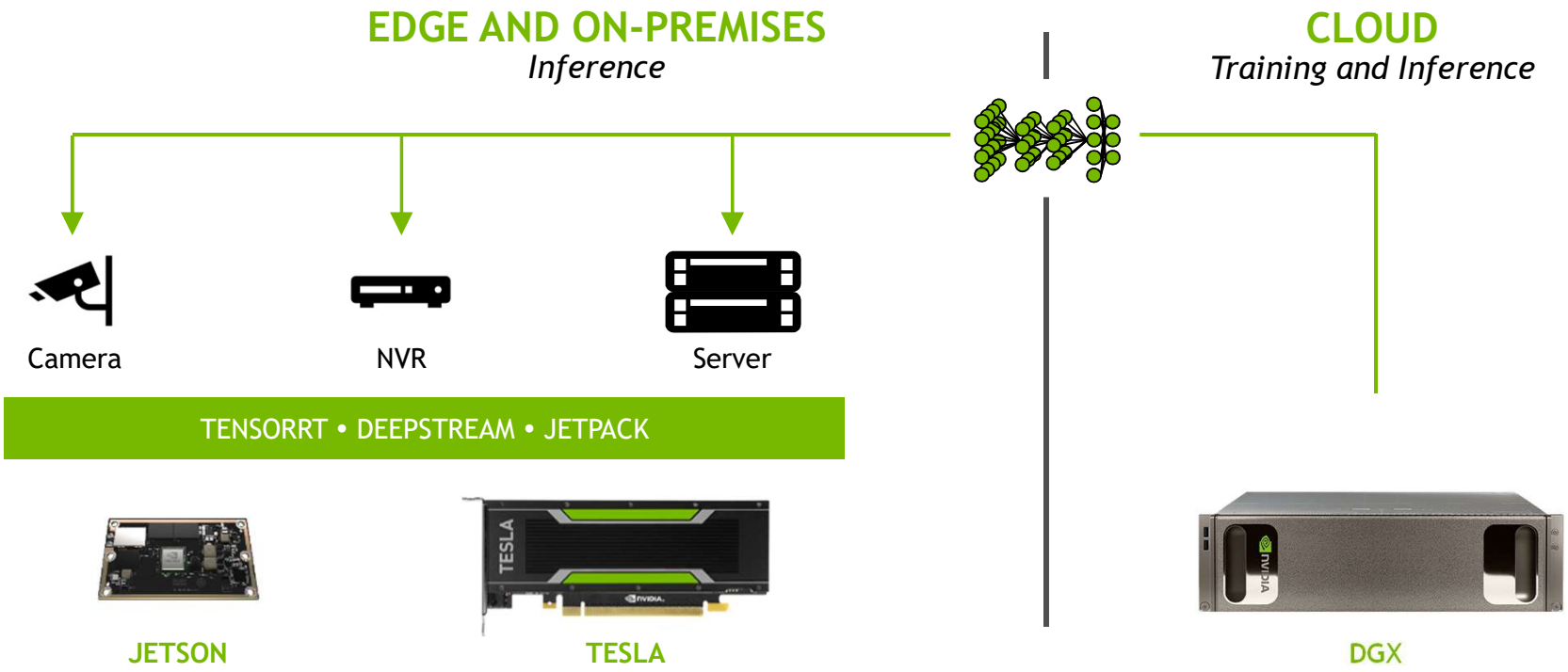
# CLARA AI: HEALTHCARE AND LIFE SCIENCES

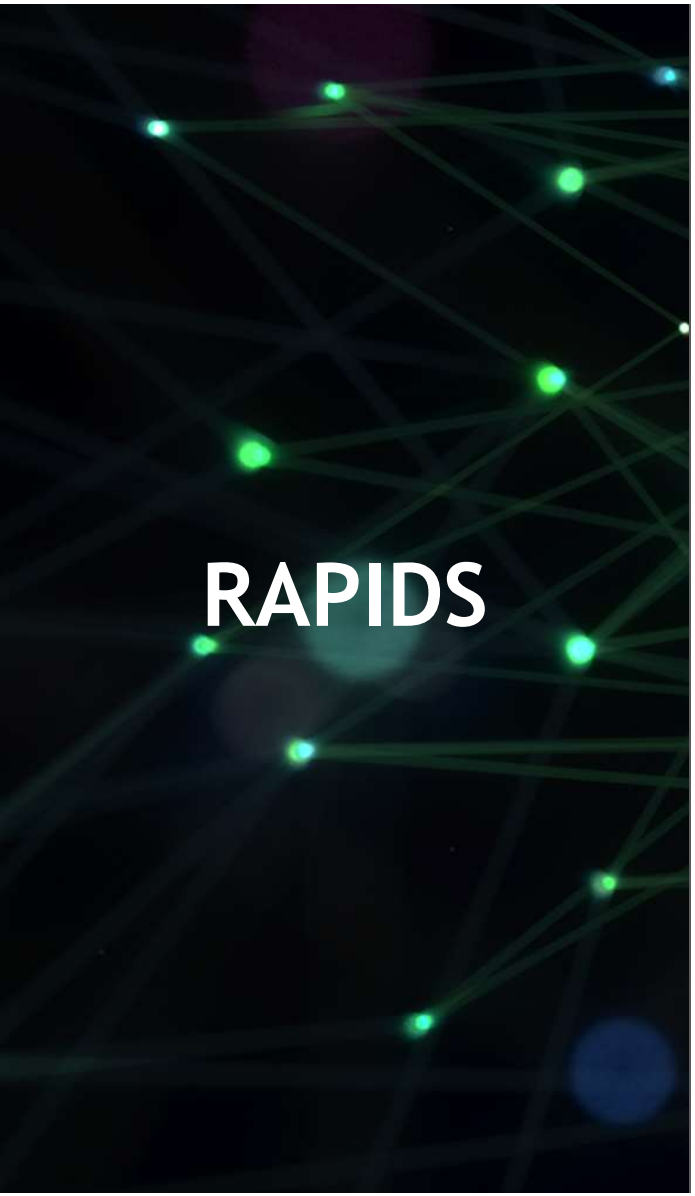


# CLARA: GENOMICS



# METROPOLIS: AI CITY EDGE TO CLOUD



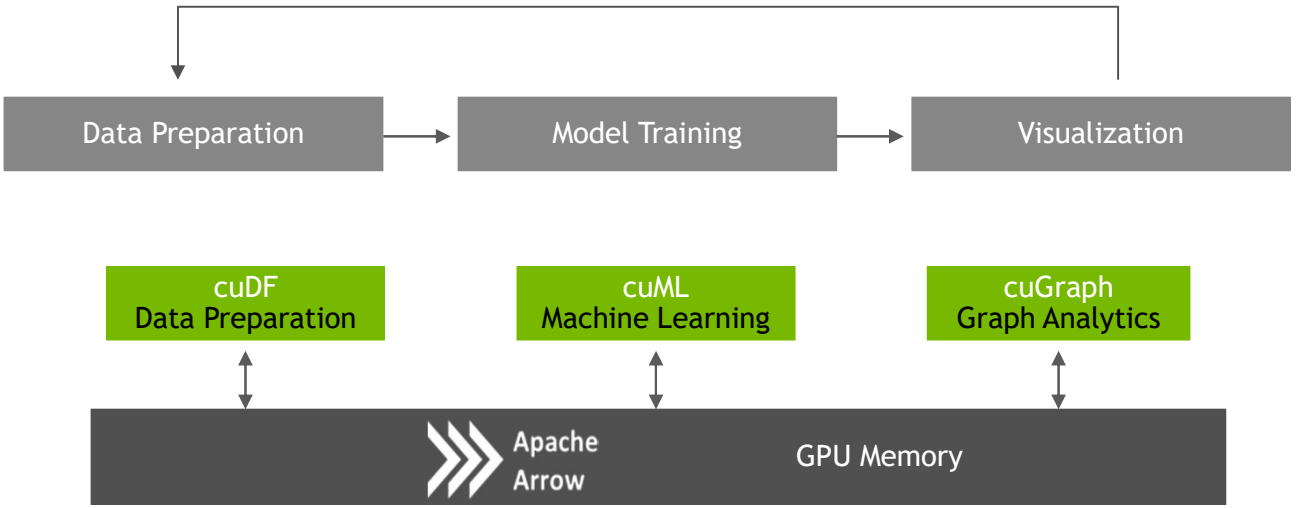



# AI: RAPIDS FOR ML

## GPU Accelerated End-to-End Data Science

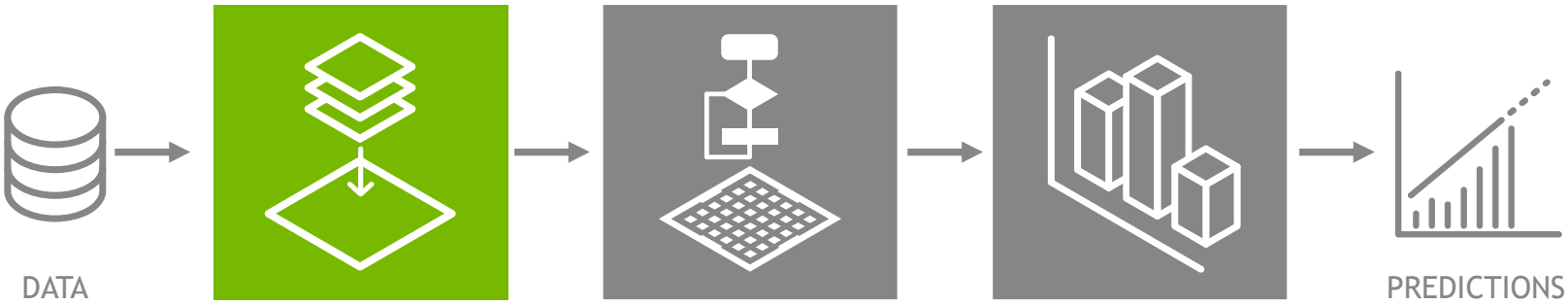
RAPIDS is a set of open source libraries for GPU accelerating **data preparation** and **machine learning.**

OSS website: rapids.ai



# GPU-ACCELERATED DATA SCIENCE WORKFLOW

## NVIDIA Accelerated Data Science Solution, Built on CUDA-X AI




### DATA PREPARATION

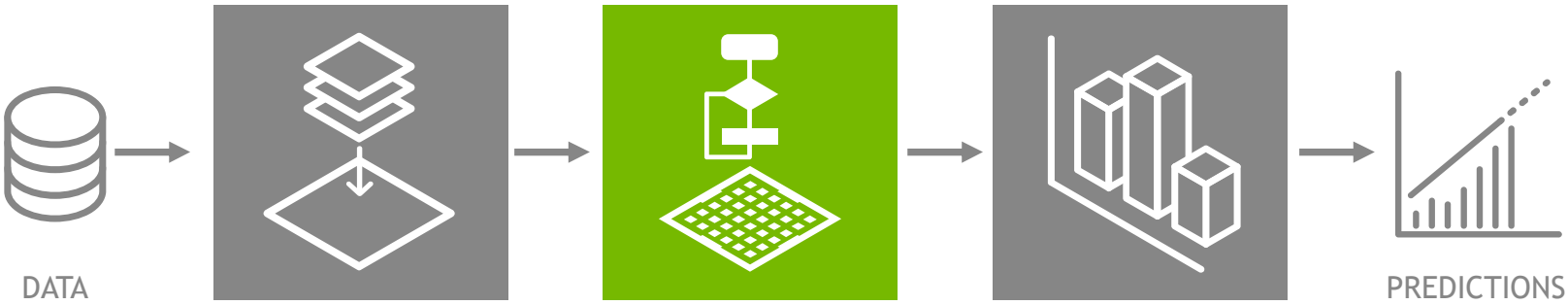
GPUs accelerated compute for in-memory data preparation

Simplified implementation using familiar data science tools

Python drop-in Pandas replacement built on CUDA C++. GPU-accelerated Spark (in development)

# GPU-ACCELERATED DATA SCIENCE WORKFLOW

## NVIDIA Accelerated Data Science Solution, Built on CUDA-X AI



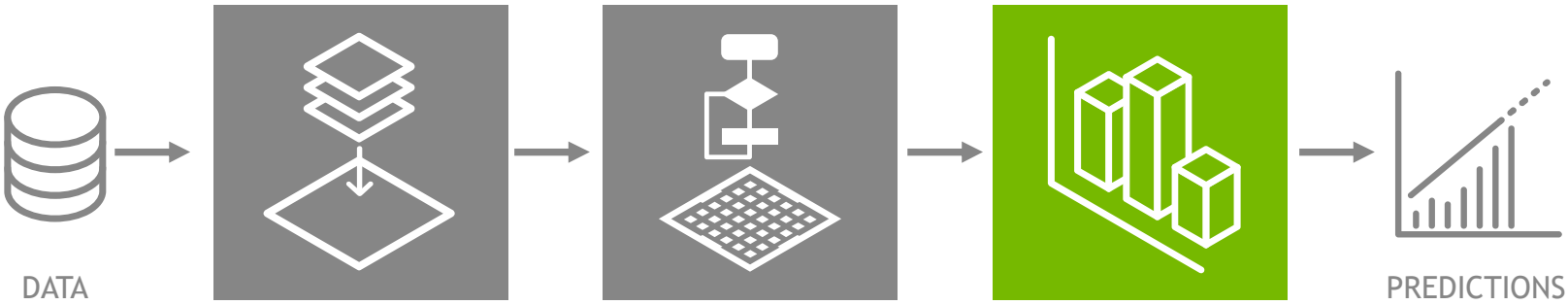

### MODEL TRAINING

GPU-acceleration of today's most popular ML algorithms

XGBoost, PCA, K-means, k-NN, DBScan, tSVD …

# GPU-ACCELERATED DATA SCIENCE WORKFLOW

## NVIDIA Accelerated Data Science Solution, Built on CUDA-X AI



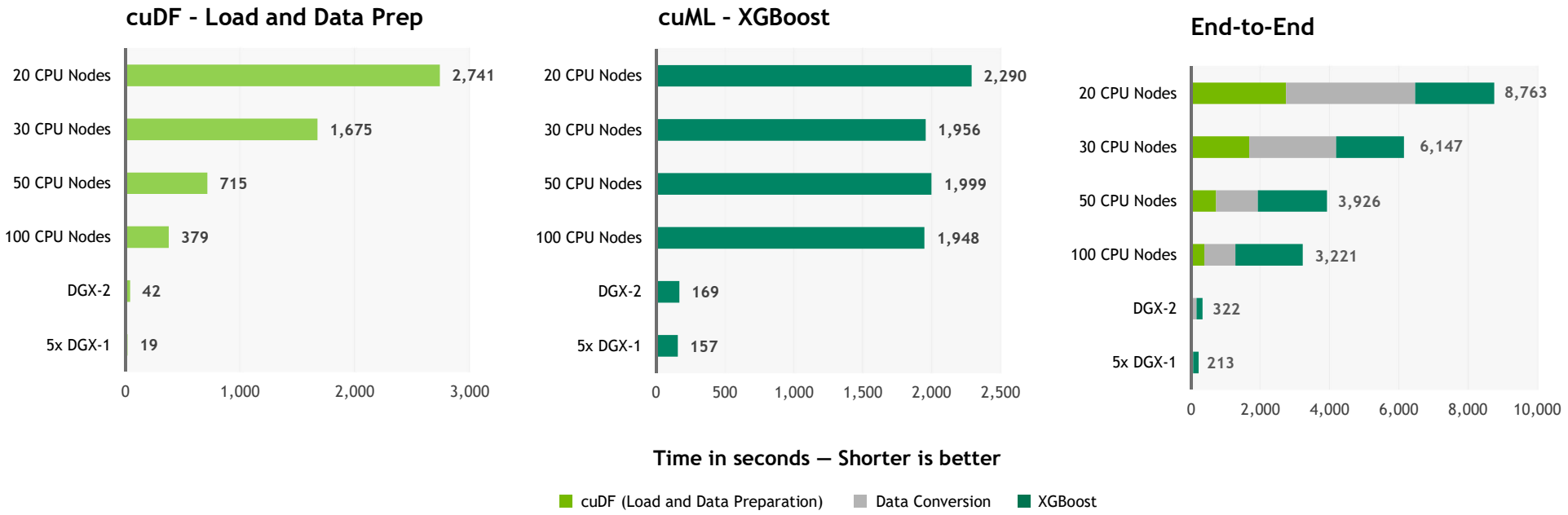

### VISUALIZATION

Effortless exploration of datasets, billions of records in milliseconds

Dynamic interaction with data = faster ML model development

Data visualization ecosystem (Graphistry & OmniSci), integrated with RAPIDS

# BENCHMARKS




### Benchmark

200GB CSV dataset; Data preparation includes joins, variable transformations.

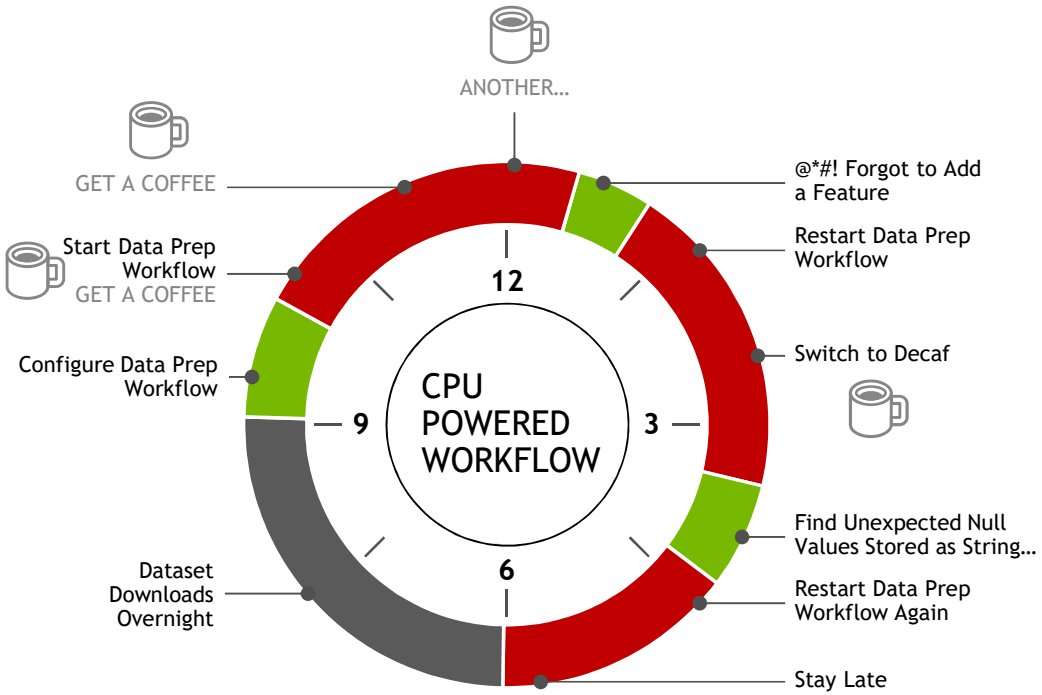
### CPU Cluster Configuration

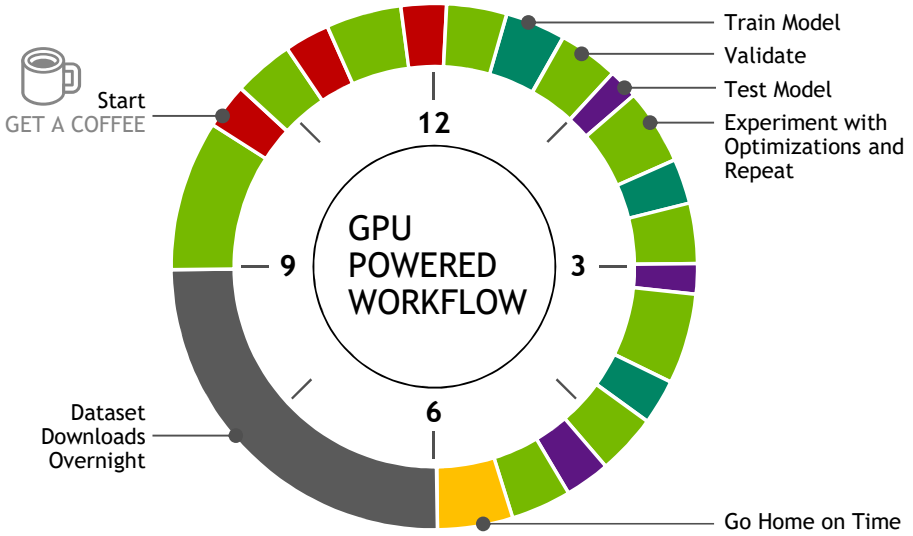
CPU nodes (61 GiB of memory, 8 vCPUs, 64-bit platform), Apache Spark

### DGX Cluster Configuration

5x DGX-1 on InfiniBand network

# DAY IN THE LIFE OF A DATA SCIENTIST

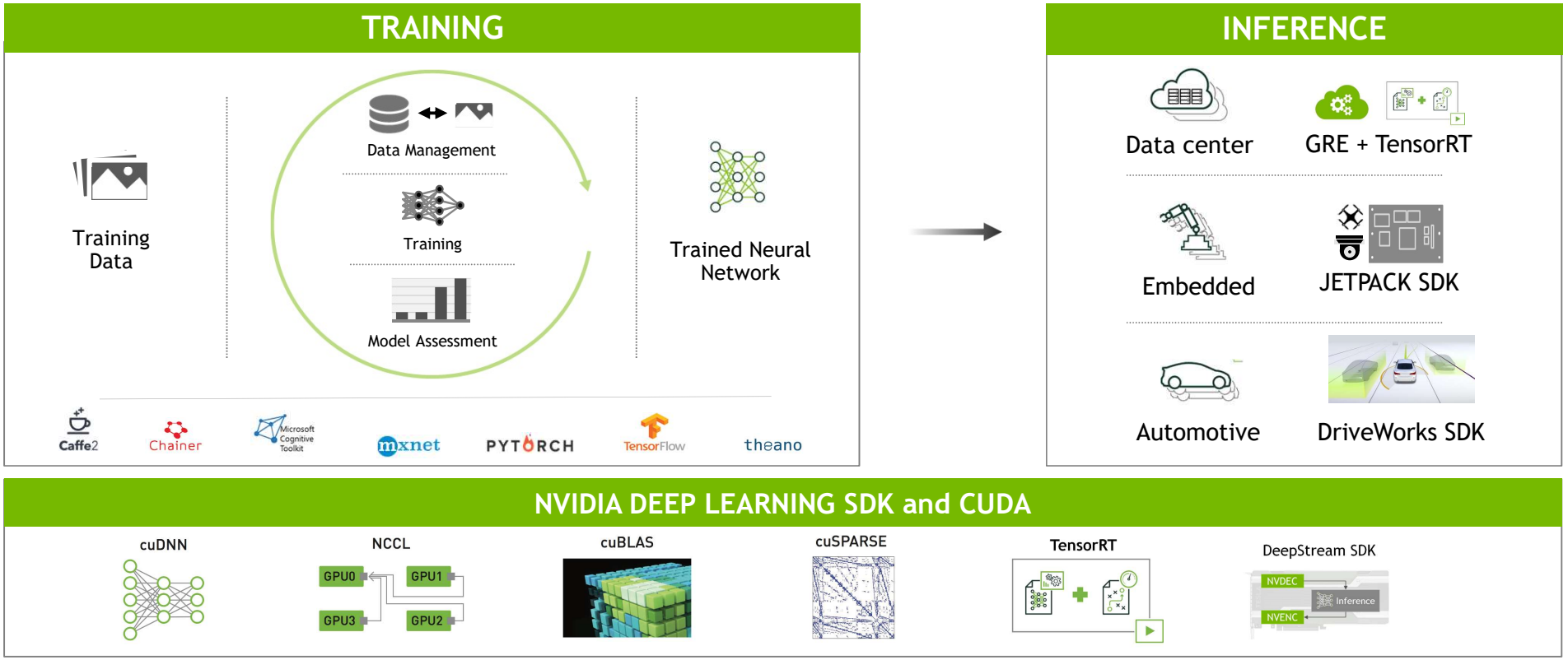



Dataset Collection | Analysis | Data Prep | Train | Inference

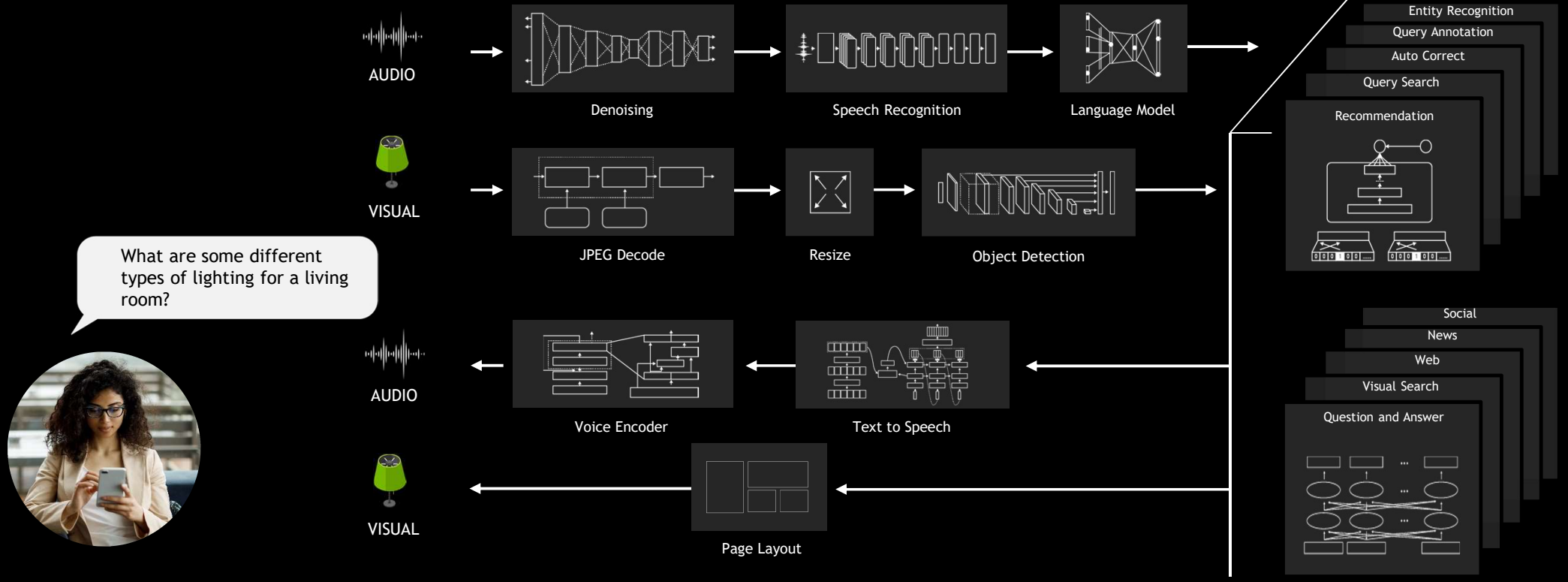

THE #1 DATA SCIENTIST EXCUSE FOR LEGITIMATELY SLACKING OFF:
"MY MODEL'S TRAINING."
HEY! GET BACK TO WORK!
TRAINING!
OH. CARRY ON.

# NVIDIA DEEP LEARNING SOFTWARE PLATFORM




developer.nvidia.com/deep-learning-software

# AI INFERENCE: CONVERSATIONAL SEARCH

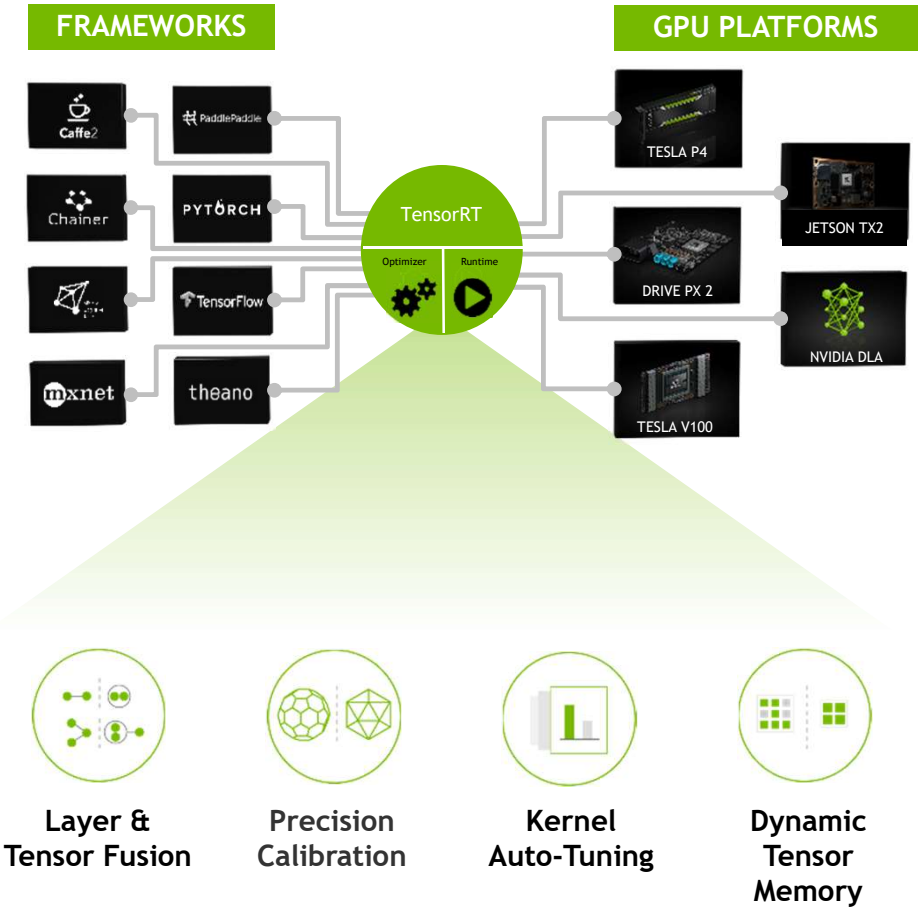



20-30 containers end-to-end | RNN, CNN, MLP in INT8, FP16, FP32 | Latency <300ms

# NVIDIA TensorRT 5

## Fastest Deep Learning Inference Platform

- Data Center, Embedded & Automotive
- In-framework support for TensorFlow
- Support for all other frameworks and ONNX
- Containerized Inference Serving Engine
- Docker and Kubernetes integration
- New Layers and APIs
- New OS Support for Windows and CentOS

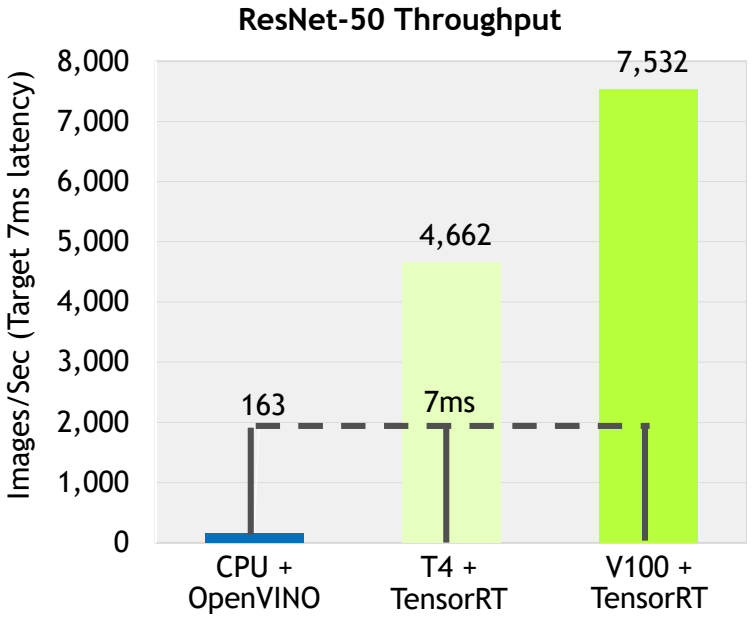



*New in TRT5

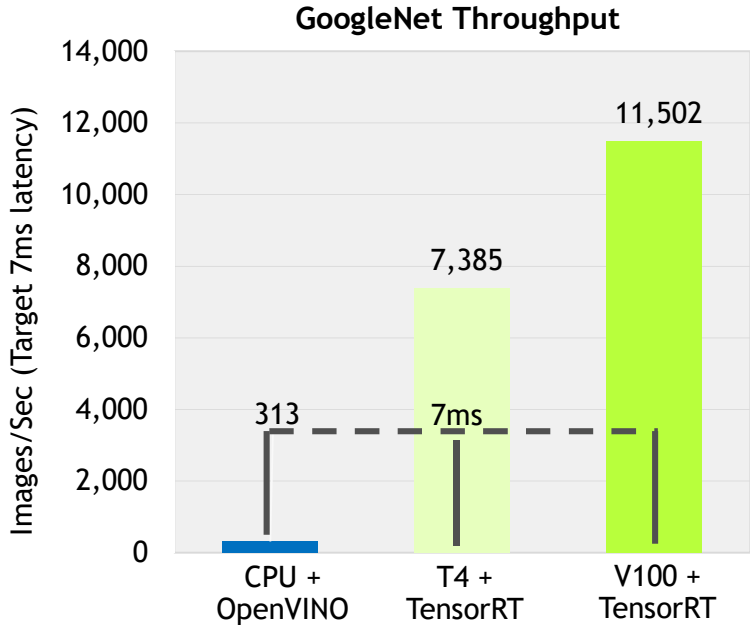
developer.nvidia.com/tensorrt

# NVIDIA TENSORRT 5

## Massive Throughput at Low Latency




CNN - IMAGES




CNN - IMAGES

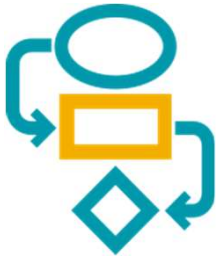
*CPU throughput based on measured inference throughput performance on Skylake-based Xeon Scalable Processor Gold 6140 CPU.*
*GPU Server config: Dual-socket Xeon Gold 6140@2.30GHz, and a single T4 or V100; GPU running TensorRT 5.1 GA vs. Intel OpenVINO Toolkit*
*CUDA 10.1.105; NCCL 2.4.3, cuDNN 7.5.0.56, cuBLAS 10.1.105; NVIDIA Driver 410.104; Batch sizes: T4=ResNet-50 32, Googlenet 52| V100=ResNet-50 52, Googlenet 80*


NVIDIA
NGC

# CHALLENGES UTILIZING AI & HPC SOFTWARE

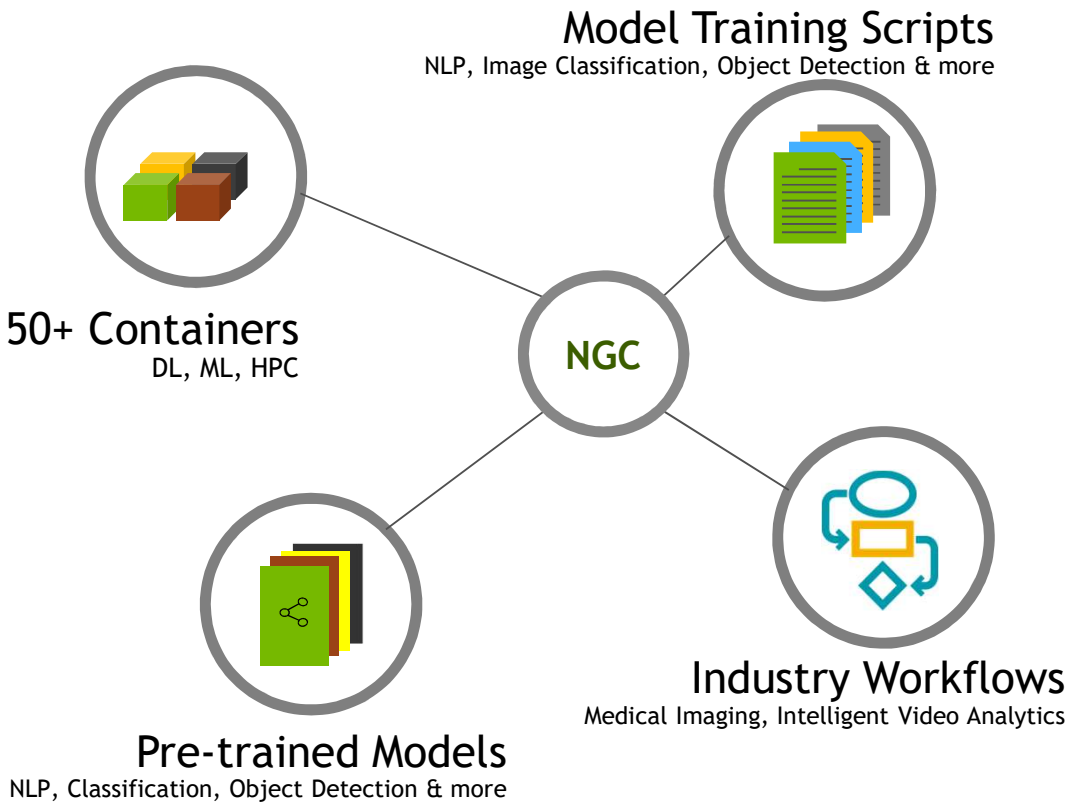
| EXPERTISE | INSTALLATION | OPTIMIZATION | PRODUCTIVITY | MAINTAINENCE |
|---|---|---|---|---|
|  |  |  |  |  |
| Building AI-centric solutions requires expertise | Complex, time consuming, and error-prone | Requires expertise to optimize framework performance | Users limited to older features and lower performance | IT can't keep up with frequent software upgrades |

# NGC – SIMPLIFYING AI & HPC WORKFLOWS

| EMBEDDING EXPERTISE | FASTER DEPLOYMENTS | OPTIMIZED SOFTWARE | HIGHER PRODUCTIVITY | ZERO MAINTENANCE |
|---|---|---|---|---|
|  |  |  |  |  |
| Deliver greater value, faster | Eliminates installations. Simply Pull & Run the app | Key DL frameworks updated monthly for perf optimization | Better Insights and faster time-to-solution | Empowers users to deploy the latest versions without IT support |

# NGC: GPU-OPTIMIZED SOFTWARE HUB

## Simplifying DL, ML and HPC Workflows




**Simplify Deployments**

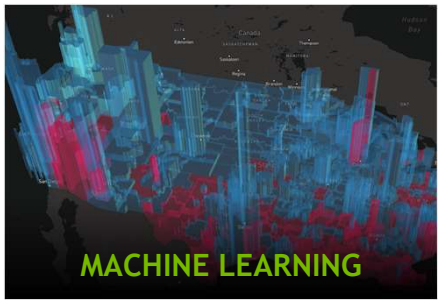
**Innovate Faster**

**Deploy Anywhere**

# GPU-OPTIMIZED SOFTWARE CONTAINERS

## Over 50 Containers on NGC


TensorFlow | PyTorch | more

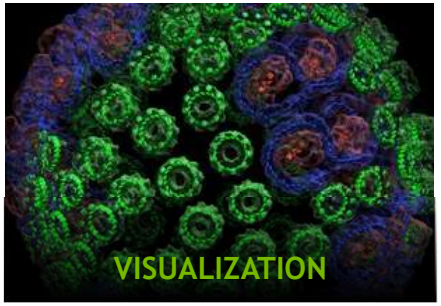



RAPIDS | H2O | more


TensorRT | DeepStream | more


NAMD | GROMACS | more


Parabricks

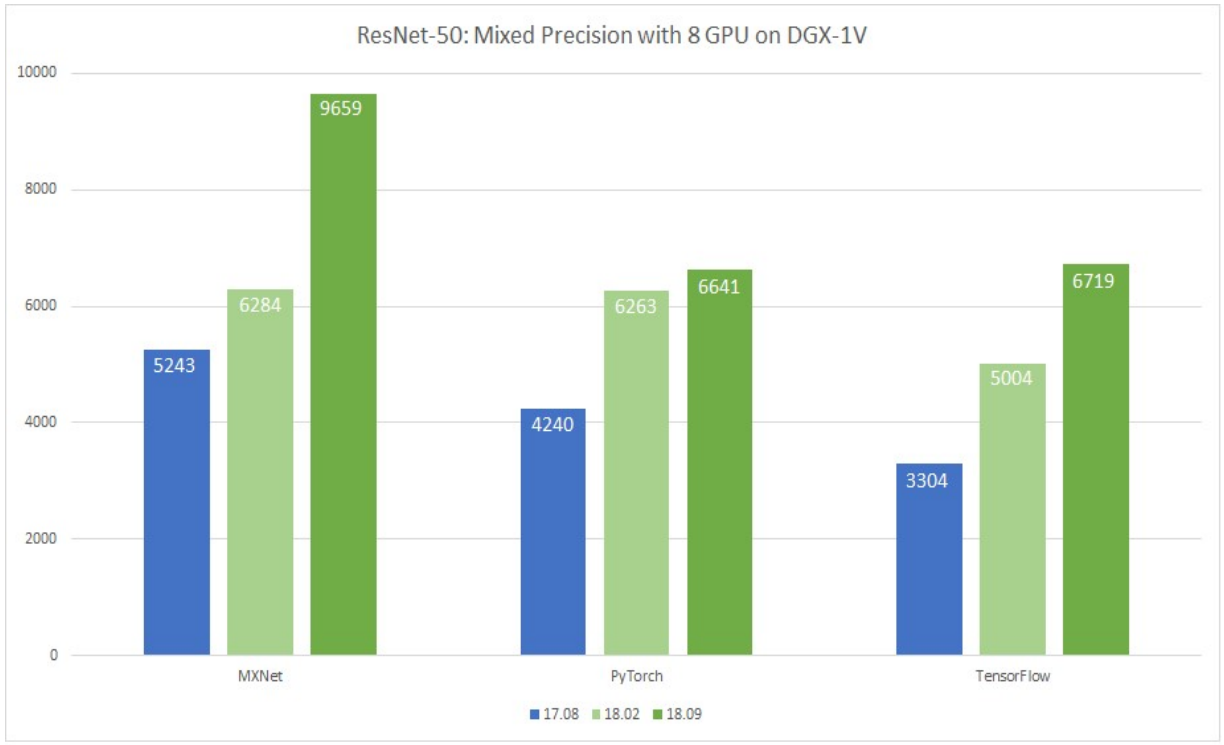


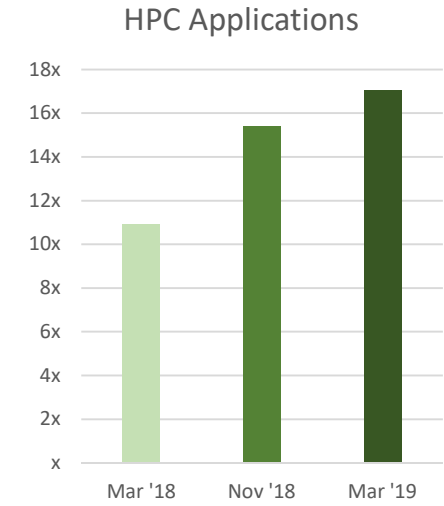

ParaView | IndeX | more

# CONTINUOUS IMPROVEMENT

## NVIDIA Optimizations Delivers Better Performance on the Same Hardware

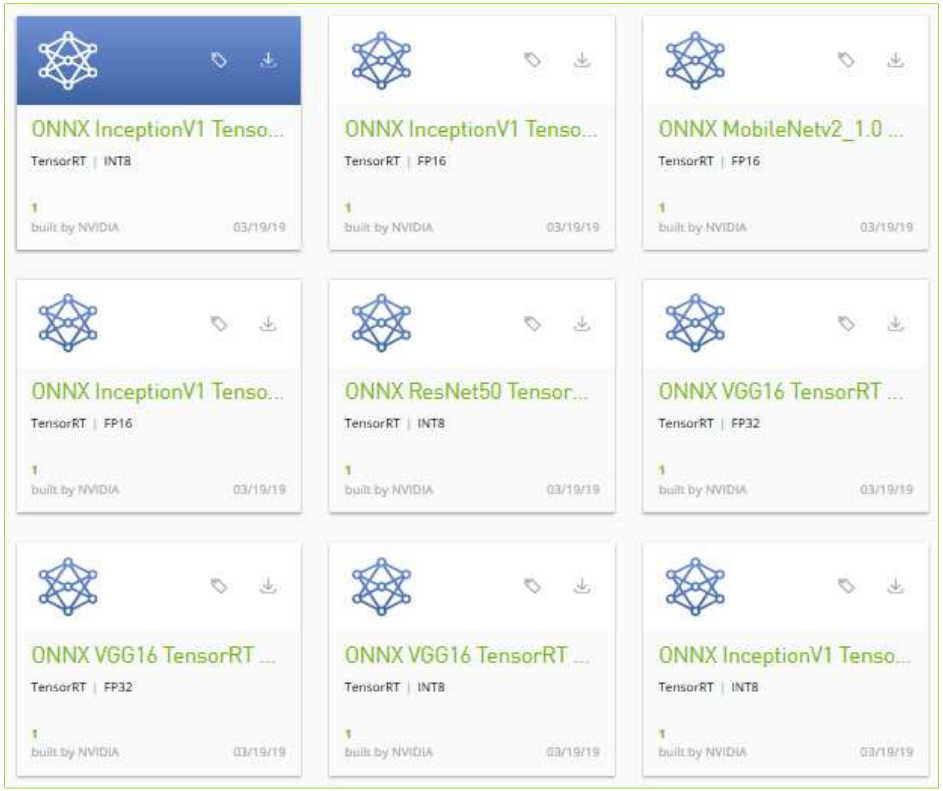
**Over 12 months, up to 1.8X improvement with mixed-precision on ResNet-50**




Speedup across Chroma, GROMACS, LAMMPS, QE, MILC, VASP, SPECFEM3D, NAMD, AMBER, GTC, RTM | 4x V100 v. Dual-Skylake | CUDA 9 for Mar '18 & Nov '18, CUDA 10 for Mar '19

# NGC MODEL REGISTRY




## Repository of Popular AI Models

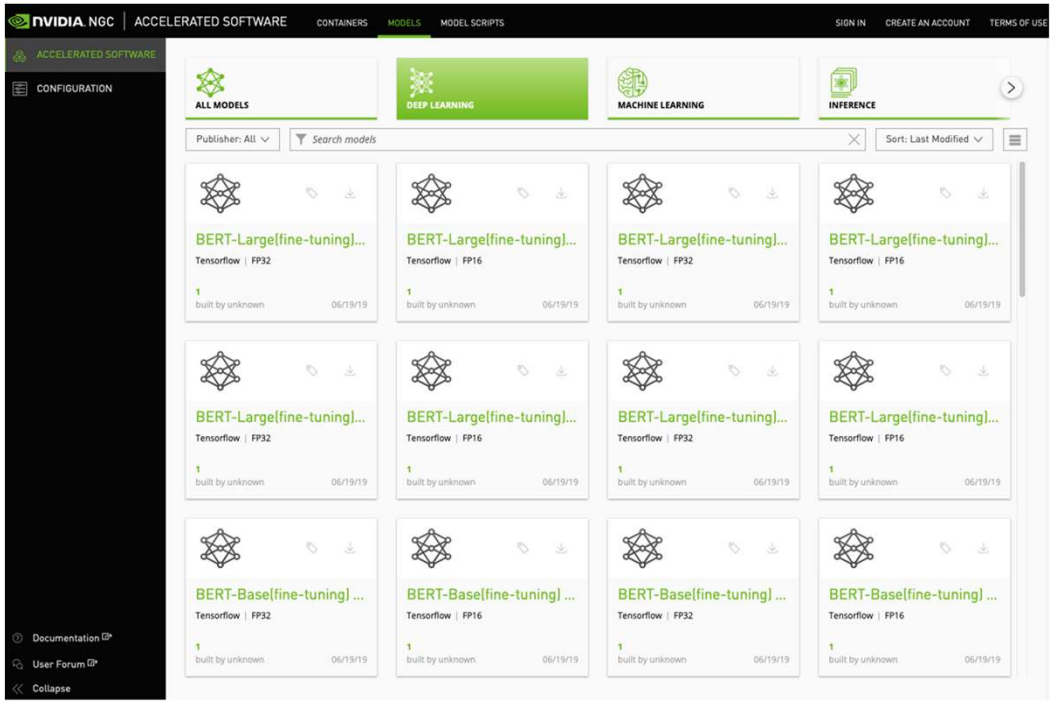
- Starting point to retrain, prototype or benchmark against your own models
- Use As-Is or easily customize
- Ready for inference with Tensor Cores
- Precision: INT8, FP16, FP32
- Optimized for multiple GPU architectures

# LANGUAGE MODELLING: BERT for TensorFlow

https://ngc.nvidia.com/models

Many pre-trained BERT models available:

- FP32 and Mixed Precision
- Base Model: ~110M parameters, 12 transformer layers faster but lower accuracy
- Large Model: ~340M parameters, 24 transformer layers, higher accuracy computational more demanding




List and Download available

- WEB UI: ngc.nvidia.com/models
- CLI: `curl -L https://api.ngc.nvidia.com/v2/models/nvidia/bert_tf_v1_1_base_fp32_384/versions/1/zip > bert.zip`


NVIDIA