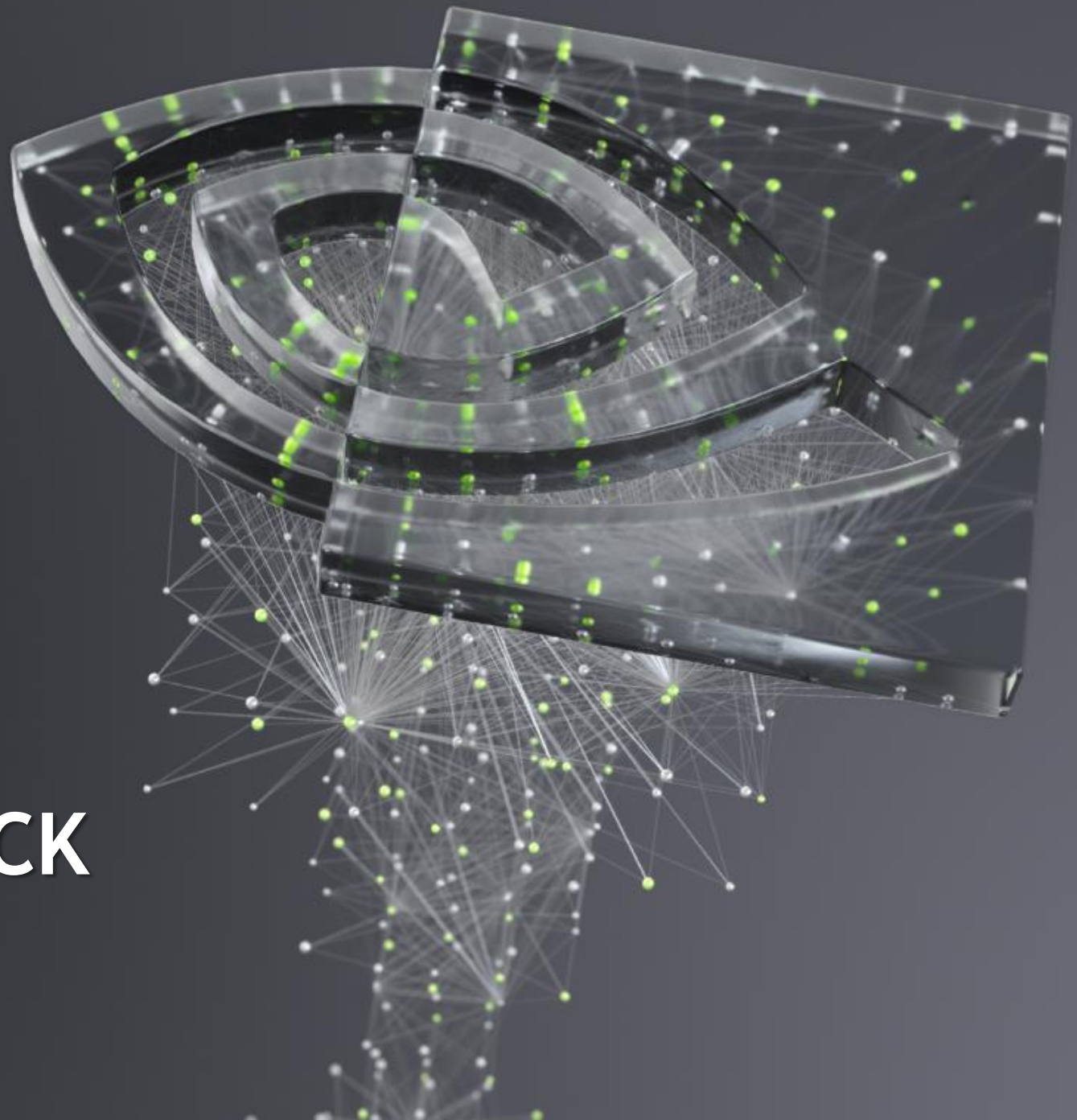


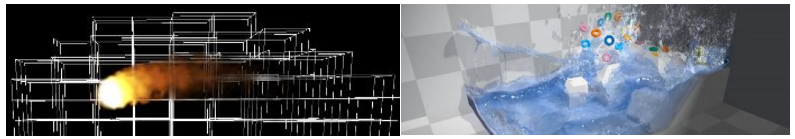
# NVIDIA FULL AI STACK

Amit Kumar

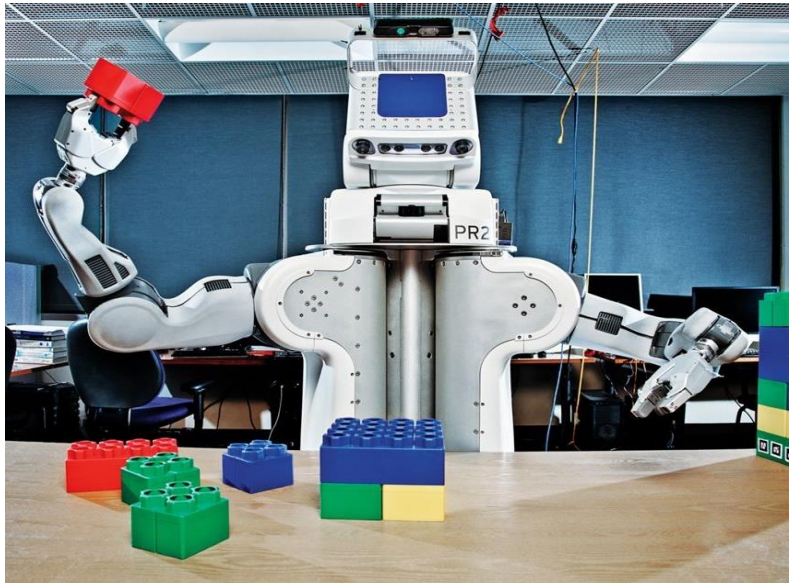
# NVIDIA
# “THE AI COMPUTING COMPANY”

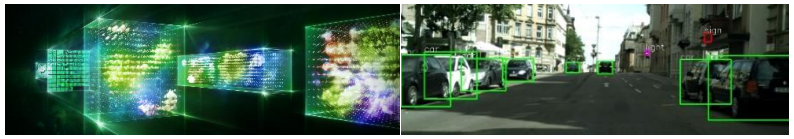


GPU Computing



Computer Graphics



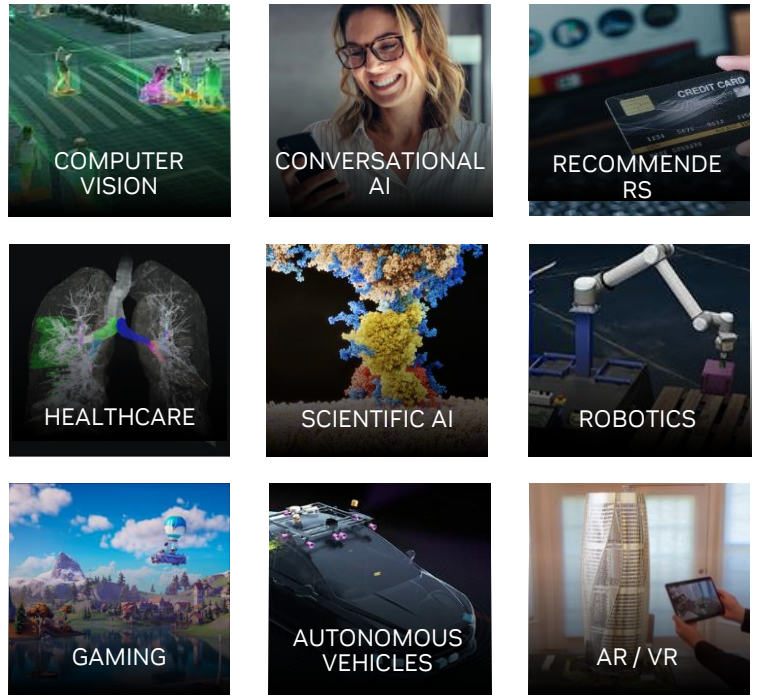
Artificial Intelligence

# NVIDIA Is the Leader in Enterprise AI

## Providing the Frameworks and Platforms for Development and Deployment

- 1B CUDA GPUs
- 250 ExaFLOPS in the Cloud
- 2,000 GPU Apps
- 10M+ CUDA Downloads
- 450+ SDKs + AI Models
- 12K AI Startups
- 3.5M Developers

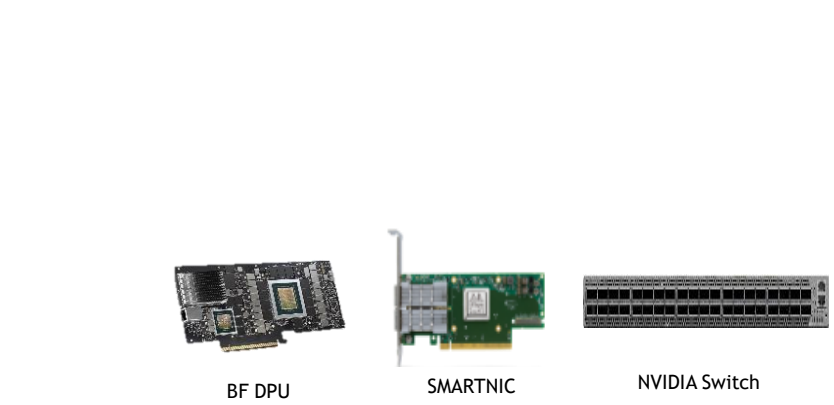
TRANSFORMING INDUSTRIES

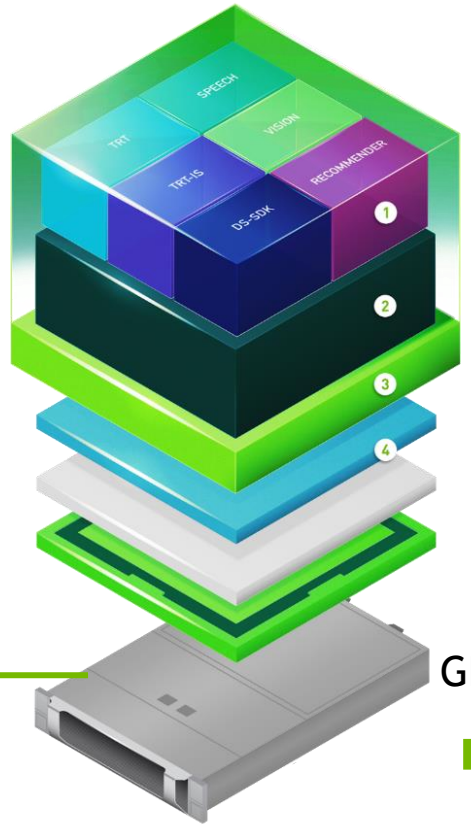



*Breakthroughs in deep learning around 2012 brought AI into focus, but only NVIDIA had the strategy, vision, and roadmap to invest in supporting these now mainstream AI workloads,"*
***Forrester Wave, AI Infrastructure, Q4 2021***

# NVIDIA COMPLETE END-TO-END PLATFORM

## Fastest AI Solution – Easily Deployable Into Production




AI models (NGC)

Libraries (CUDA-X)

Enterprise OS

GPU Accelerated OEM Servers

SCALABLE

Vetted reference design to
server, rack and data center

AI READY DATA CENTER

NVIDIA

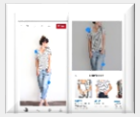
# NVIDIA DATA CENTER PLATFORM

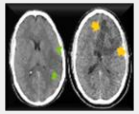
## Single Platform Drives Utilization and Productivity

**CUSTOMER USE CASES**


Speech

Translate

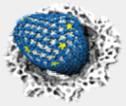

Recommender

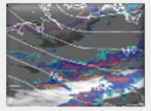

Healthcare

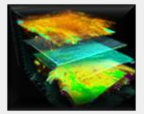

Manufacturing


Finance

CONSUMER INTERNET & INDUSTRY APPLICATIONS


Molecular Simulations


Weather Forecasting


Seismic Mapping

SCIENTIFIC APPLICATIONS


Creative & Technical


Knowledge Workers

VIRTUAL GRAPHICS

**APPS & FRAMEWORKS**

python™ TensorFlow mxnet Chainer ONNX RAPIDS PYTORCH

Amber NAMD +600 Applications

CATIA Ps AUTODESK 3DS MAX Windows 10

**CUDA-X & NVIDIA SDKs**

MACHINE LEARNING: cuDF | cuML | cuGRAPH

DEEP LEARNING: cuDNN | CUTLASS | TensorRT

HPC: OpenACC | cuFFT

VIRTUAL GPU: vDWS | vPC | vAPPS

CUDA & CORE LIBRARIES - cuBLAS | NCCL

**TESLA GPUs & SYSTEMS**


TESLA GPU

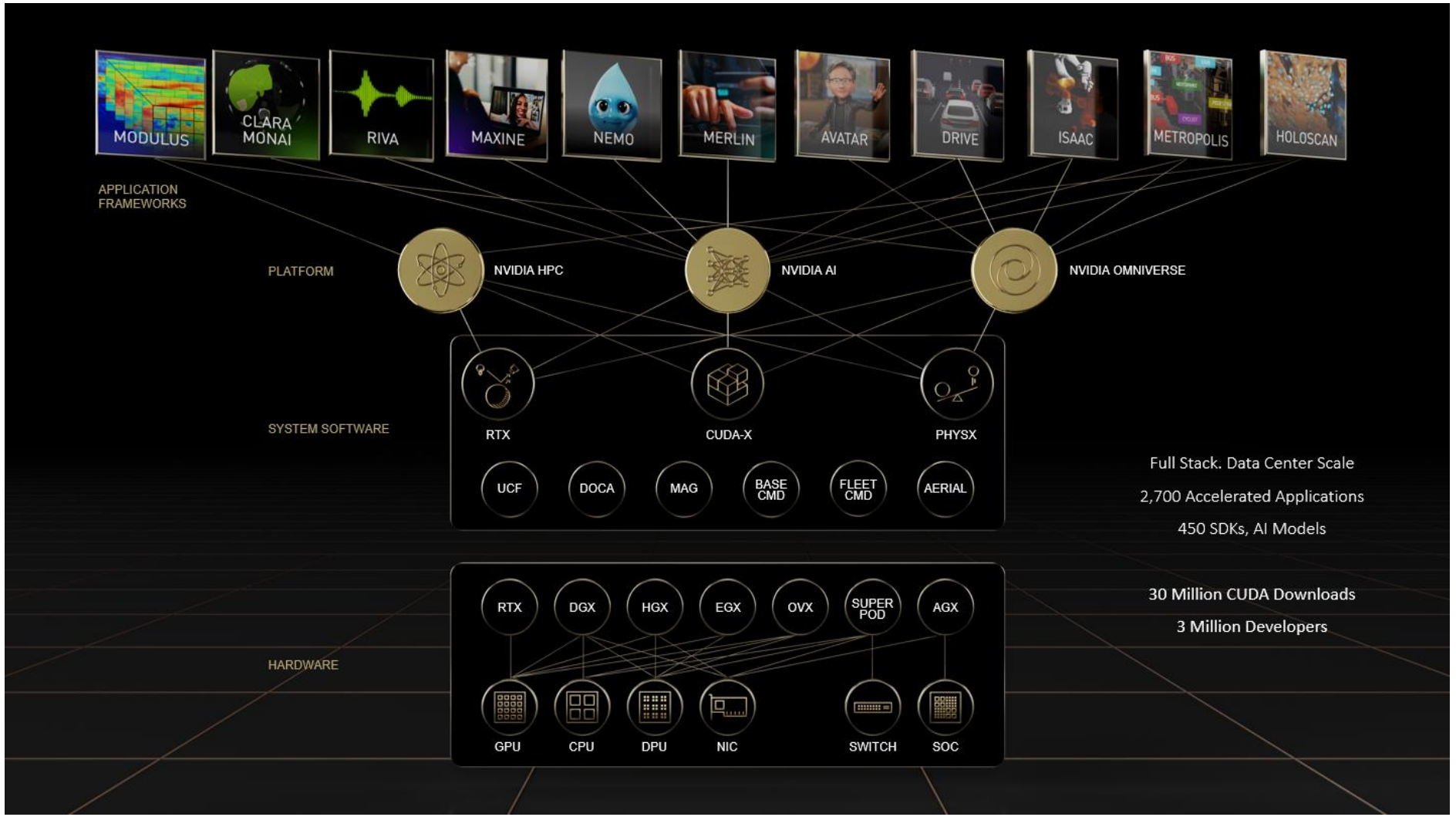

NVIDIA DGX A100


NVIDIA HGX


EVERY OEM


EVERY MAJOR CLOUD

# NVIDIA DATA CENTER PLATFORM




# NVIDIA SDK

# NVIDIA END-TO-END SOFTWARE STACK

Deep Learning Streamlined From Conception to Production at Scale

## NVIDIA Optimized Models Across Multiple Frameworks

Computer Vision

Speech

Recommender

GANs

mxnet PyTorch TensorFlow

...

## Train

Train at-scale

Re-Train with your data on with Tao

## Inference


TensorRT

## Scale

Multi-Instance GPU (MIG) for diverse users

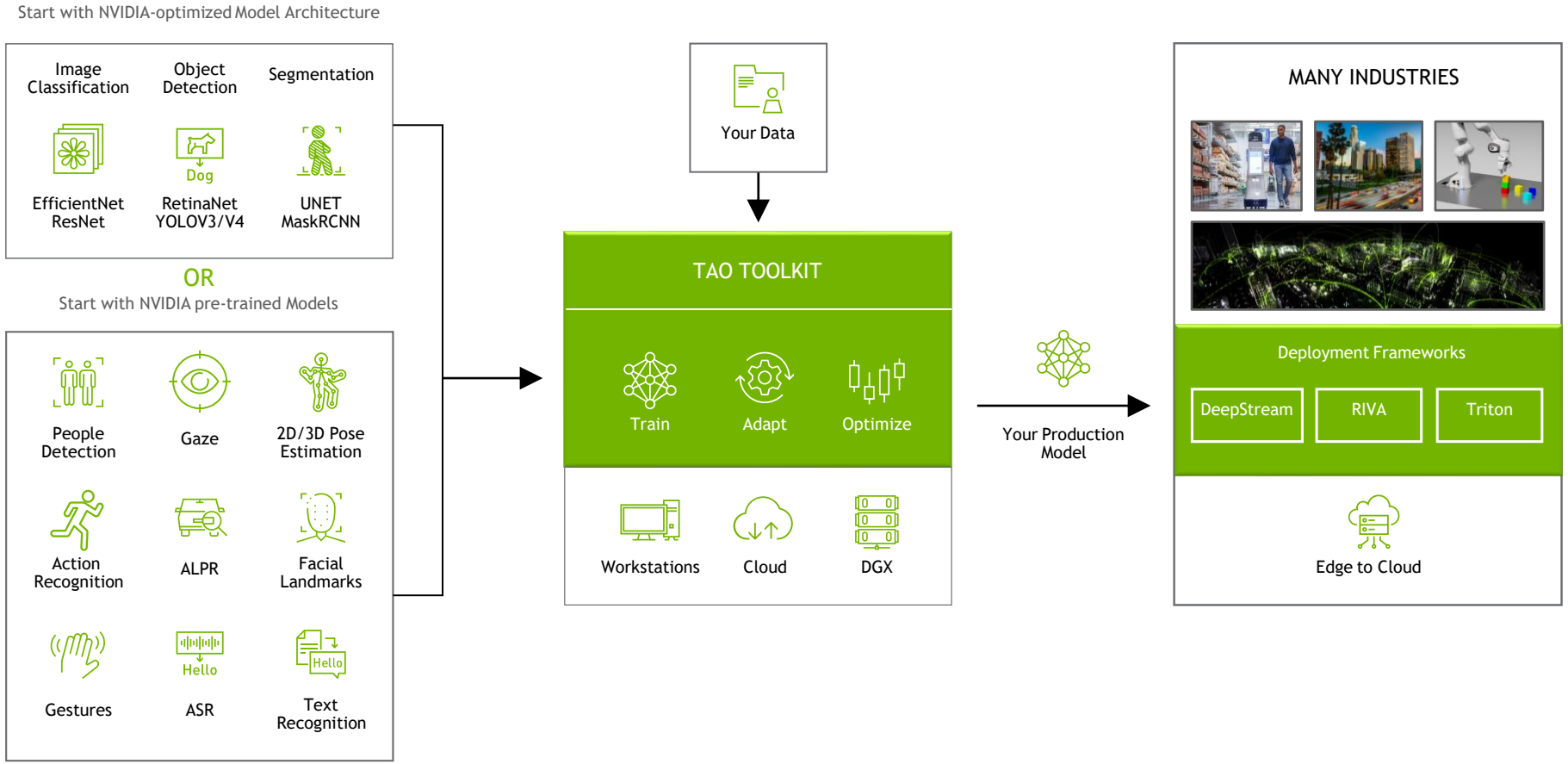

Triton Inference Server (Instant Deployment)

# NVIDIA SUPERCHARGING AI WORKFLOWS

1 Choose from NVIDIA's Library of Pre-trained Models OR Model Architectures

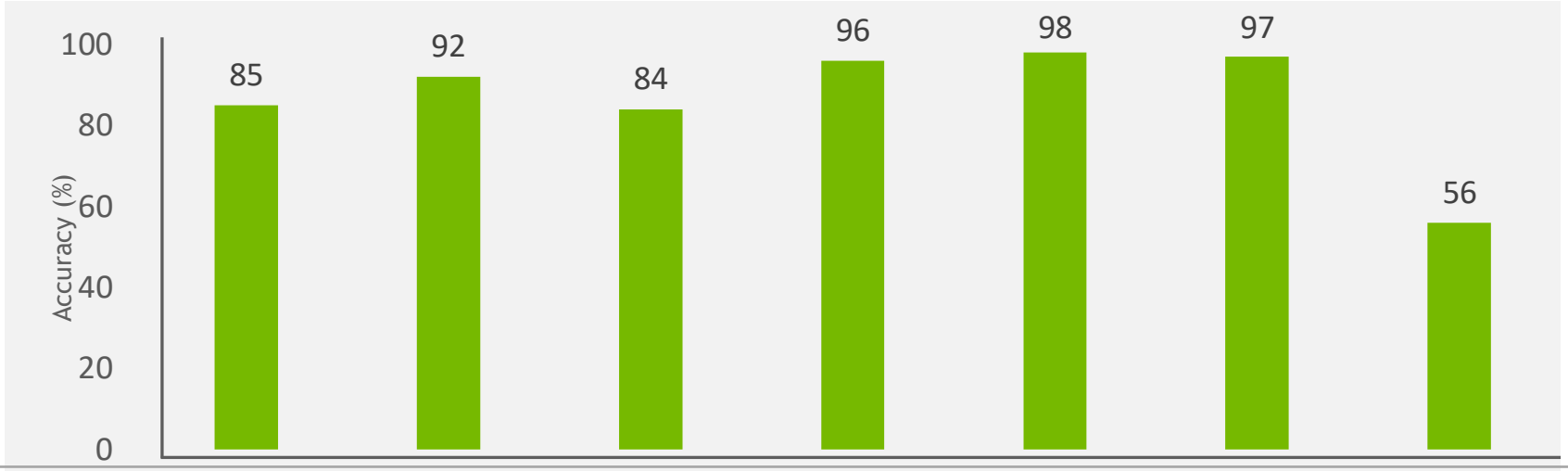
2 Quickly train, adapt, and optimize models to your unique application

3 Integrate your customized models into your application and deploy

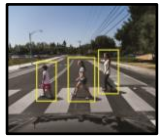



* Choose from over 100+ model combinations on NGC

# HIGH PERFORMANCE PRE-TRAINED VISION AI MODELS




| Models | Accuracy |
|---|---|
| Facial Landmark | 6.1-pixel error |
| Gaze Estimation | 6.5 RMSE |


PeopleNet


PeopleSem SegNet

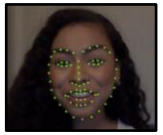

TrafficCamNet


FaceDetectIR


LPD

LPR


2D Pose Estimate


Facial Landmark


Gaze Estimation

| | PeopleNet | PeopleSem SegNet | TrafficCamNet | FaceDetectIR | LPD | LPR | 2D Pose Estimate | Facial Landmark | Gaze Estimation |
|---|---|---|---|---|---|---|---|---|---|
| Nano | 11 | 1.4 | 18 | 104 | 66 | 94 | 5 | 125 | 98 |
| Xavier NX | 296 | 17 | 340 | 2000 | 1158 | 564 | 48 | 747 | 923 |
| AGX Xavier | 462 | 28 | 656 | 3915 | 1880 | 1045 | 84 | 1451 | 1627 |
| A30 | 4163 | 330 | 4991 | 26635 | 12207 | 15960 | 1515 | 10078 | 15172 |
| A100 | 6001 | 519 | 9520 | 50541 | 21931 | 26600 | 2686 | 23117 | 26534 |

15+ Pre-trained models - download for free from NGC

NVIDIA

# ENABLING BEYOND PRE-TRAINED AI MODELS

## 100+ Combination of Model Architectures and Backbones

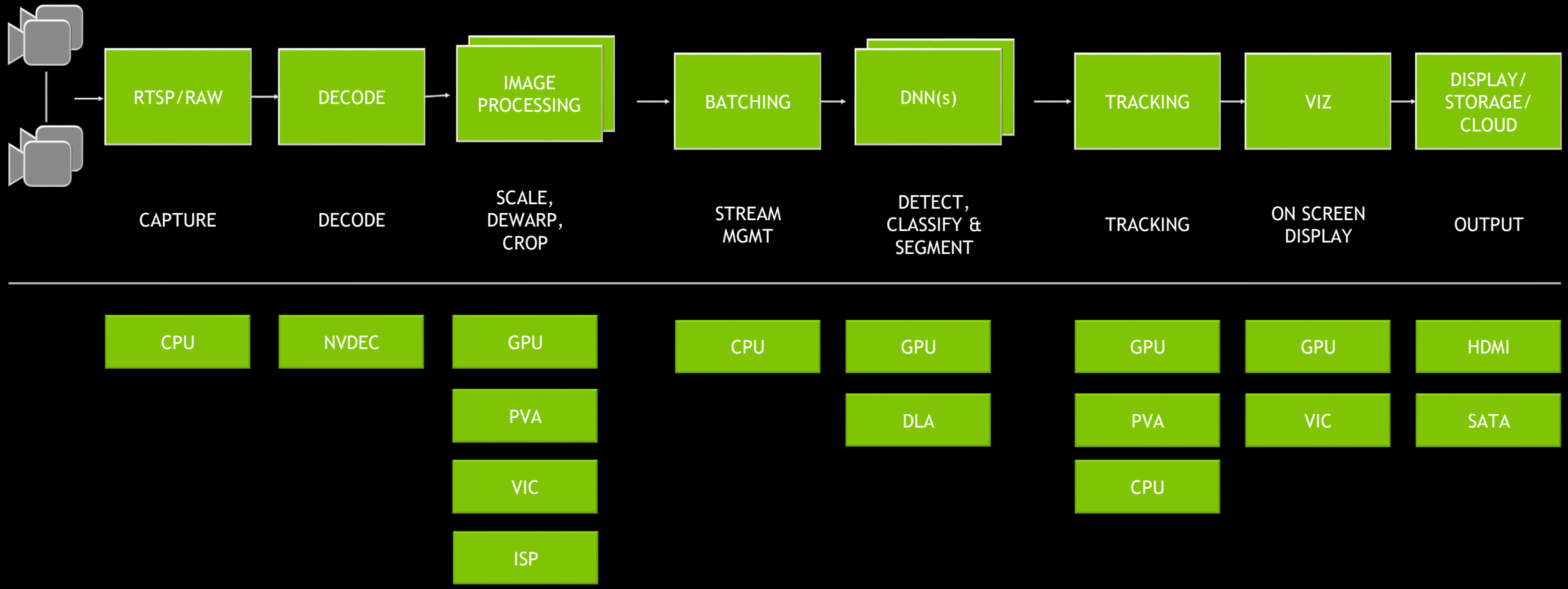
| | Image Classification | Object Detection | | | | | | | | | Segmentation | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DetectNet_V2 | FasterRCNN | SSD | YOLOV3 | YOLOV4 | RetinaNet | DSSD | EfficientDet | YOLOV4 Tiny | MaskRCNN | UNET |
| ResNet10/18 /34/50/101 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ |
| VGG16/19 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ |
| GoogLeNet | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| MobileNet V1/V2 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| SqueezeNet | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| DarkNet 19/53 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| CSPDarkNet 19/53 | ✓ | | | | | ✓ | | | | ✓ | | |
| EfficientNet B0-B5 | ✓ | | ✓ | ✓ | | | ✓ | ✓ | ✓ | | | |

Pre-trained weights trained on OpenImage dataset

New in TAO Toolkit 21-11 release

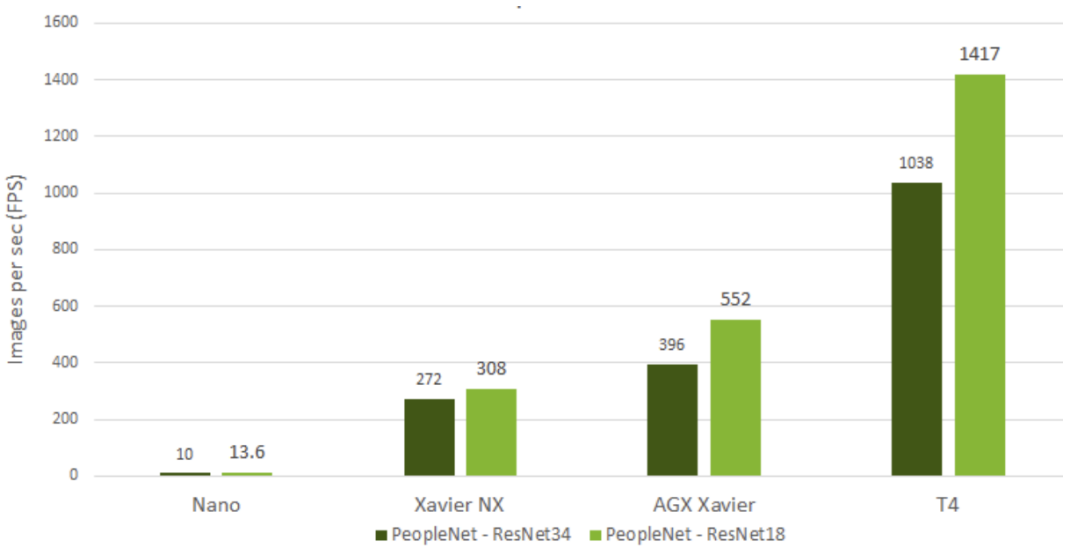
# (IVA)DEEP LEARNING COMPUTER VISION PIPELINES

# DEEPSTREAM GRAPH ARCHITECTURE

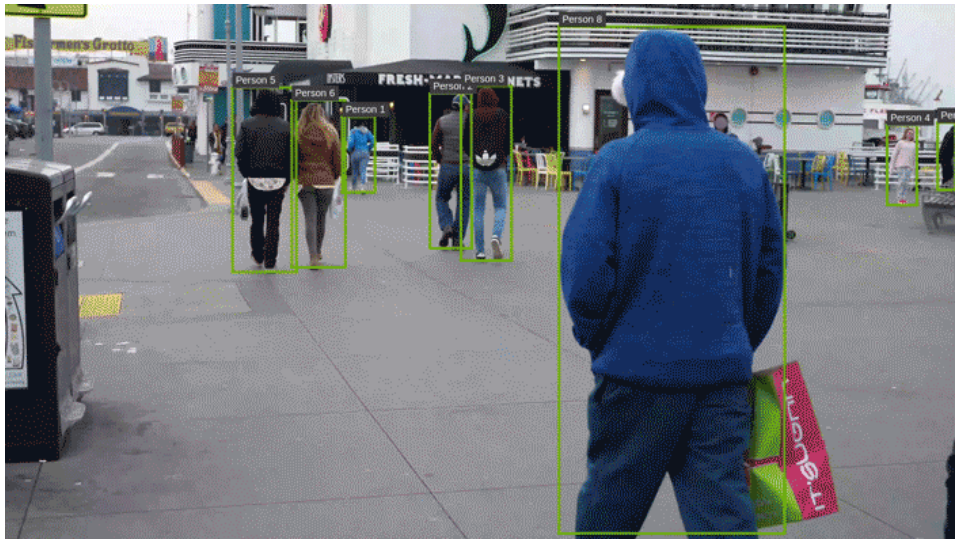


# PEOPLENET: REAL-TIME INFERENCE PERFORMANCE

## Detect persons, bags and faces




VIDEO DEMO

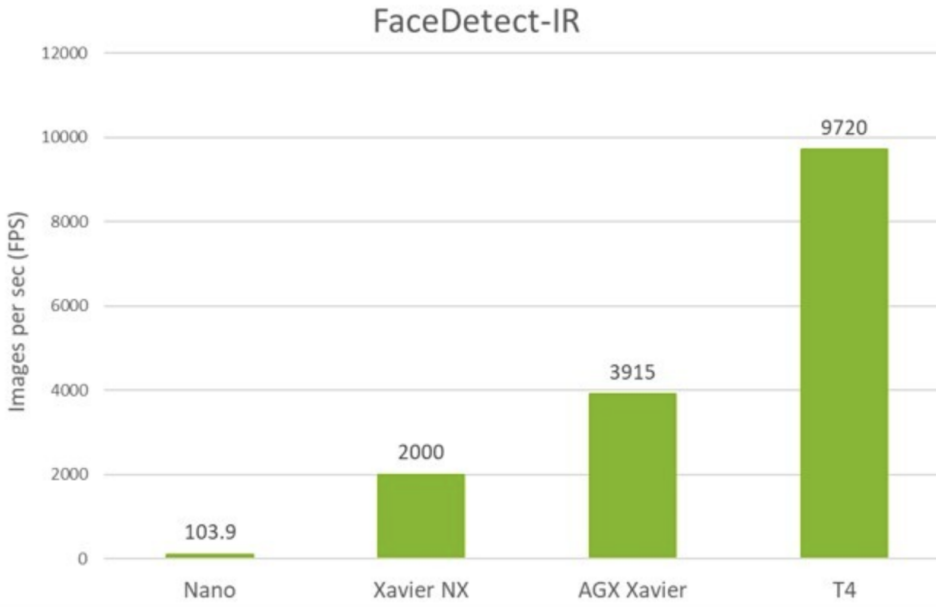



Number of classes: 3
Dataset: 750k frames

| Accuracy |
| --- |
| 84% |

# FACEDETECT-IR: REAL-TIME INFERENCE PERFORMANCE

Detect one or more faces in each image / video





Number of classes: 1
Dataset: 600k images

| Accuracy |
| --- |
| 96.21% |

# STATE OF THE ART NEURAL NETWORK ARCHITECTURES

- Optimized for NVIDIA GPUs
- Support provided for SOTA AI
- Pretrained weights freely available on NGC
- Flexibility to retrain your data with TAO Toolkit to customize your models

| | Image Classification | Object Detection | | | | | | Instance Segmentation |
|---|---|---|---|---|---|---|---|---|
| | | DetectNet_V2 | FasterRCNN | SSD | YOLOV3 | RetinaNet | DSSD | MaskRCNN |
| ResNet 10/18/34/50/101 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| VGG16/19 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| GoogLeNet | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| MobileNet V1/V2 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| DarkNet 19/53 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| SqueezeNet | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | |

Models trained on google open images public dataset
Available to download on ngc.nvidia.com

Object Detection

Image Classification

Instance Segmentation

# END-TO-END AI WITH NVIDIA DEEPSTREAM (DS)

## Reduce development time and increase overall throughput

| | | | | Jetson Nano | Jetson Xavier NX | | | Jetson AGX Xavier | | | T4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model Architecture | Inference Resolution | Precision | Model Accuracy | GPU (FPS)* | GPU (FPS) | DLA1 (FPS) | DLA2 (FPS) | GPU (FPS) | DLA1 (FPS) | DLA2 (FPS) | GPU (FPS) |
| PeopleNet-ResNet18 | 960x544 | INT8 | 80% | 14 | 218 | 72 | 72 | 384 | 94 | 94 | 1105 |
| PeopleNet-ResNet34 | 960x544 | INT8 | 84% | 10 | 157 | 51 | 51 | 272 | 67 | 67 | 807 |
| TrafficCamNet-ResNet18 | 960x544 | INT8 | 84% | 19 | 261 | 105 | 105 | 464 | 140 | 140 | 1300 |
| DashCamNet-ResNet18 | 960x544 | INT8 | 80% | 18 | 252 | 102 | 102 | 442 | 133 | 133 | 1280 |
| FaceDetect-IR-ResNet18 | 384x240 | INT8 | 96% | 95 | 1188 | 570 | 570 | 2006 | 750 | 750 | 2520 |
| VehicleTypeNet - ResNet18 ᴛ | 224x224 | INT8 | 96% | 120 | 1333 | 678 | 678 | 3047 | 906 | 906 | 11918 |
| VehicleMakeNet - ResNet18 ᴛ | 224x224 | INT8 | 91% | 173 | 1871 | 700 | 700 | 3855 | 945 | 945 | 15743 |

**Greater end-to-end throughput using Transfer Learning Toolkit and DeepStream SDK**

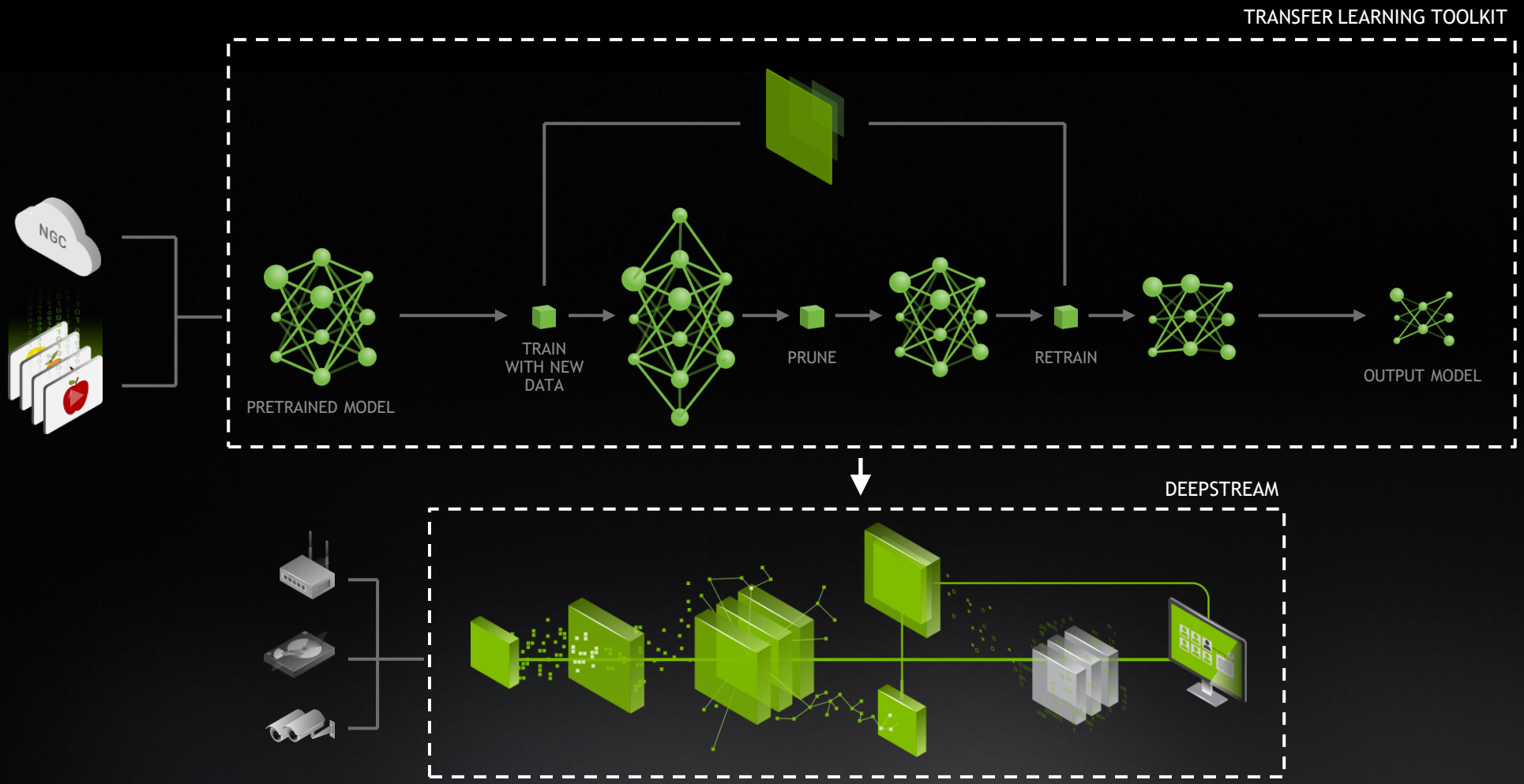
* FP16 inference on Jetson Nano

ᴛ Throughput measured using trtexec and does not reflect end-to-end performance

Easily retrain purpose-built pretrained models with TAO Toolkit and deploy at the edge or the cloud using DS

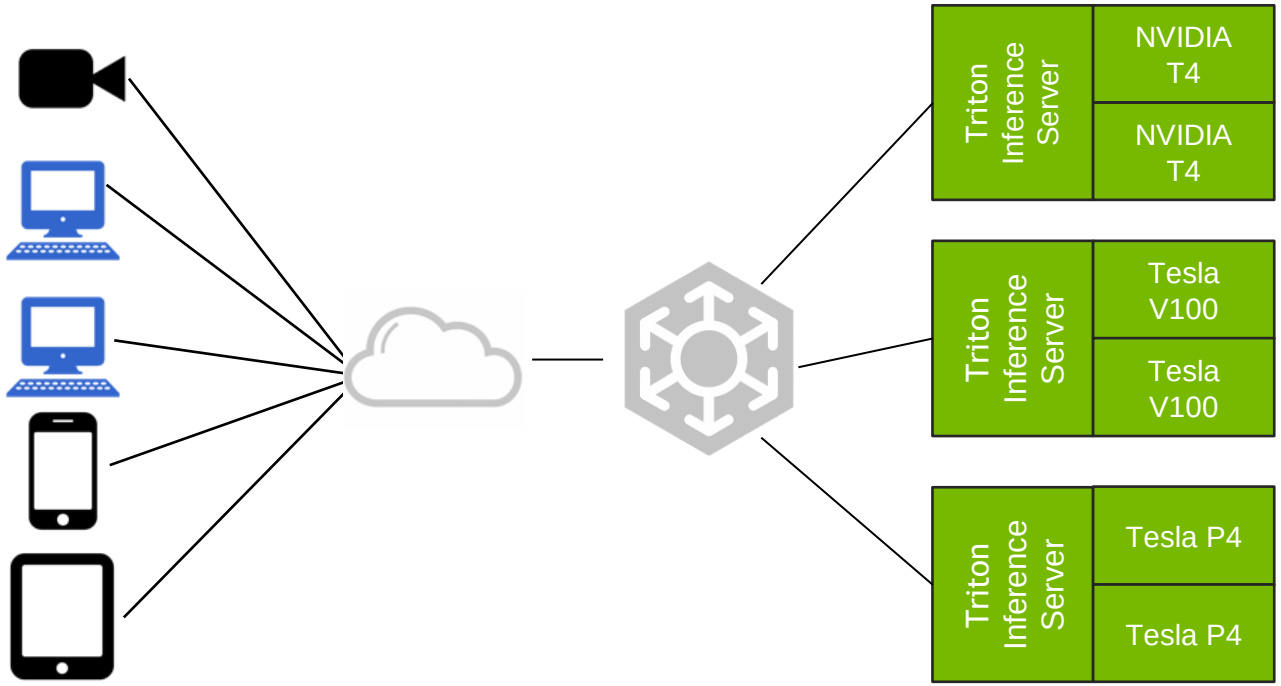
# END-TO-END DEEP LEARNING WORKFLOW

## Accelerate Time to Market and Save on Compute Resources!



# NVIDIA TRITON INFERENCE SERVER

## Production Data Center Inference Server




Maximize real-time inference performance of GPUs

Quickly deploy and manage multiple models per GPU per node

Easily scale to heterogeneous GPUs and multi GPU nodes

Integrates with orchestration systems and auto scalers via latency and health metrics

Now open source for thorough customization and integration


NVIDIA Riva

# SPEECH AI IS EVERYWHERE

## Hundreds of Billions of Minutes Of Speech Generated Daily

Call Center

500M Calls Daily


Virtual Assistants

8B Devices

Online Meetings

200M Daily

Telecom

Finance

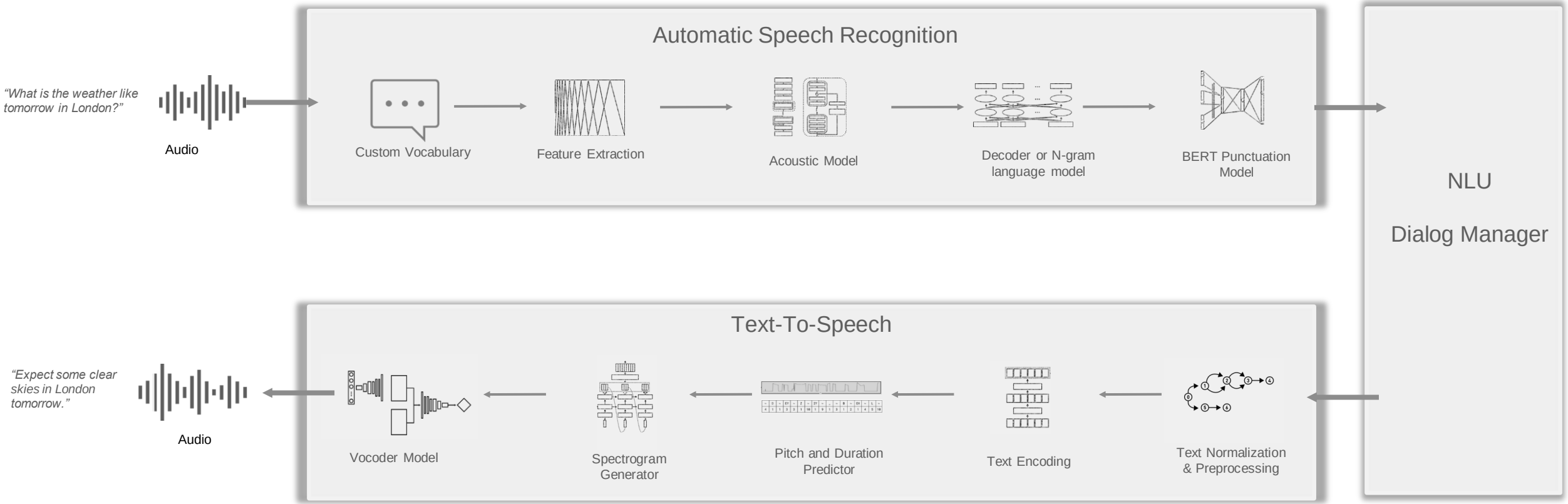
Healthcare

Manufacturing

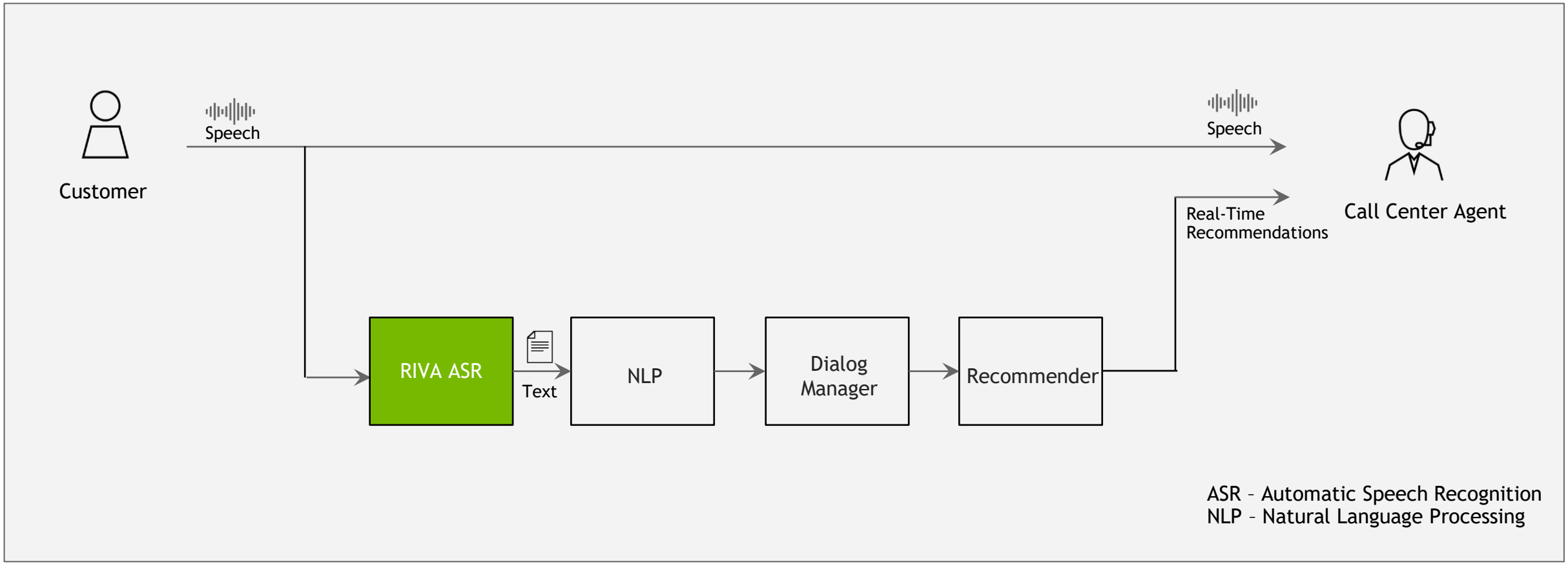
Automotive

# SPEECH AI IS COMPLEX

New Models | Domain-Specific Accuracy | Complex Pipelines | Real-Time Responses



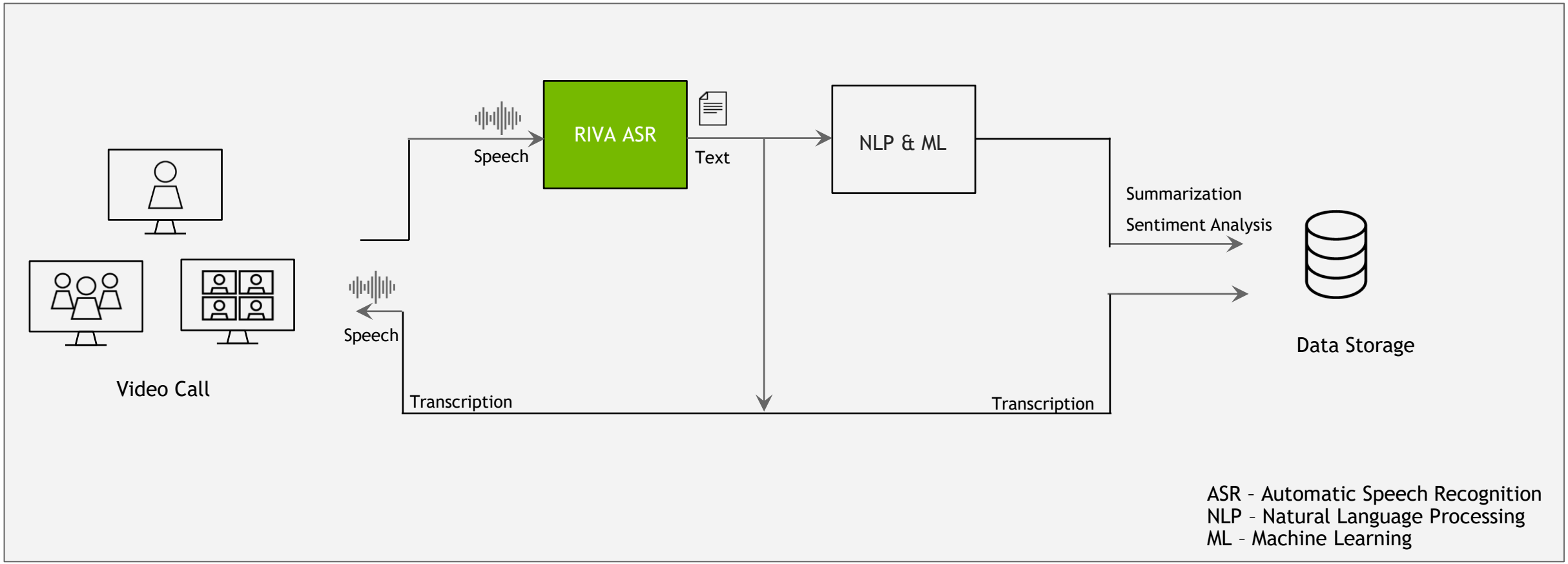
# ASR USE CASE: CALL CENTER AGENT ASSIST



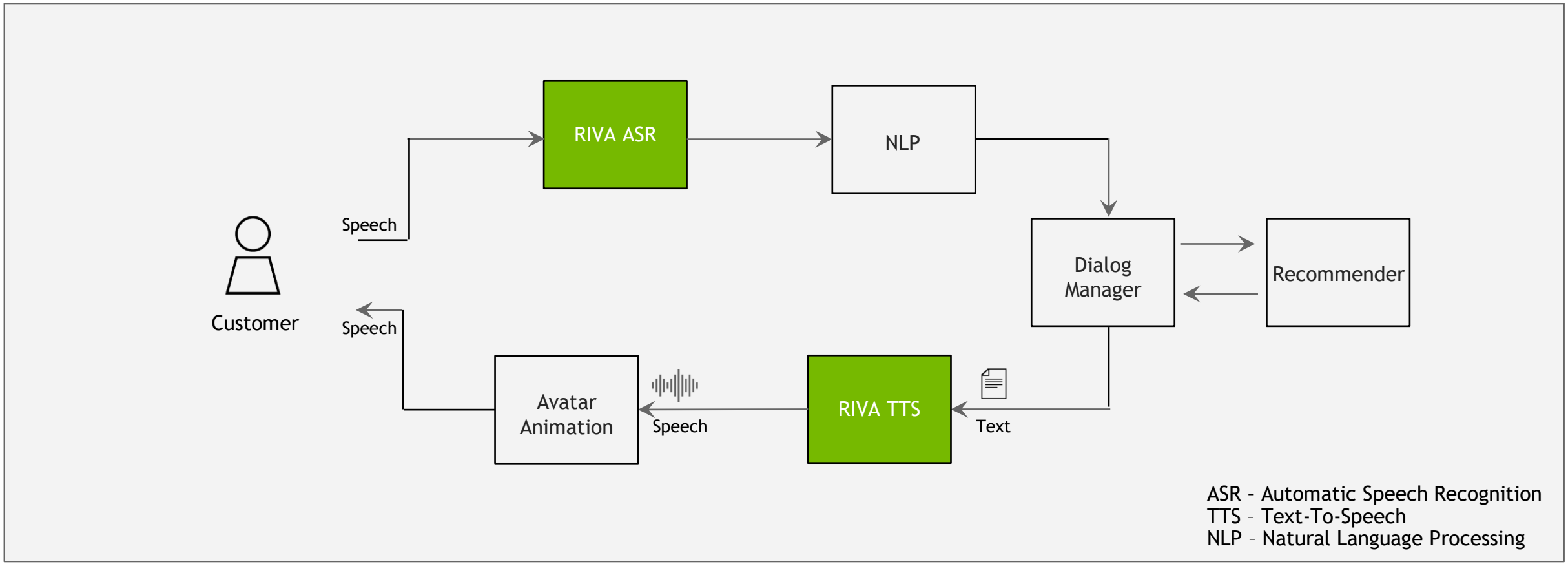

Other Applications in Call Center:

- Transcription
- Digital Assistant

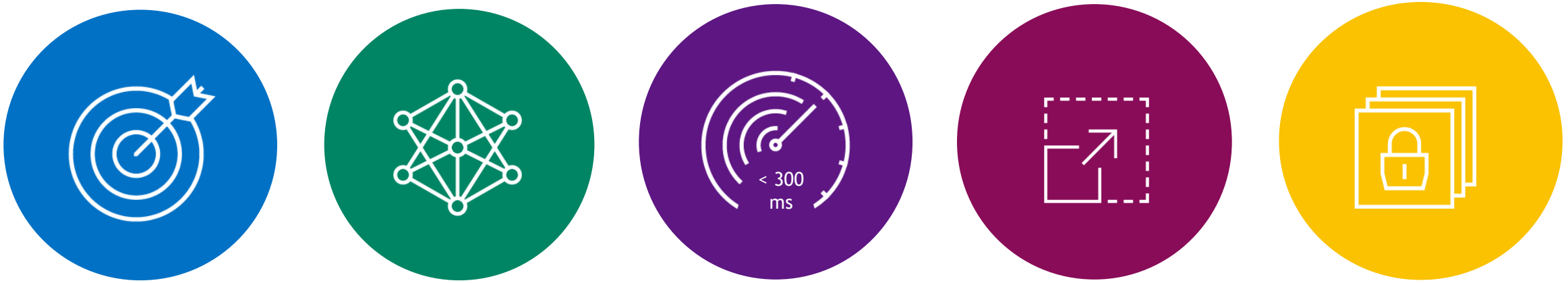
# ASR USE CASE: VIDEO CALL TRANSCRIPTION



# ASR & TTS USE CASE: CONSUMER APPLICATION DIGITAL AVATAR



# RIVA SPEECH AI SOLVES CUSTOMERS PAIN POINTS

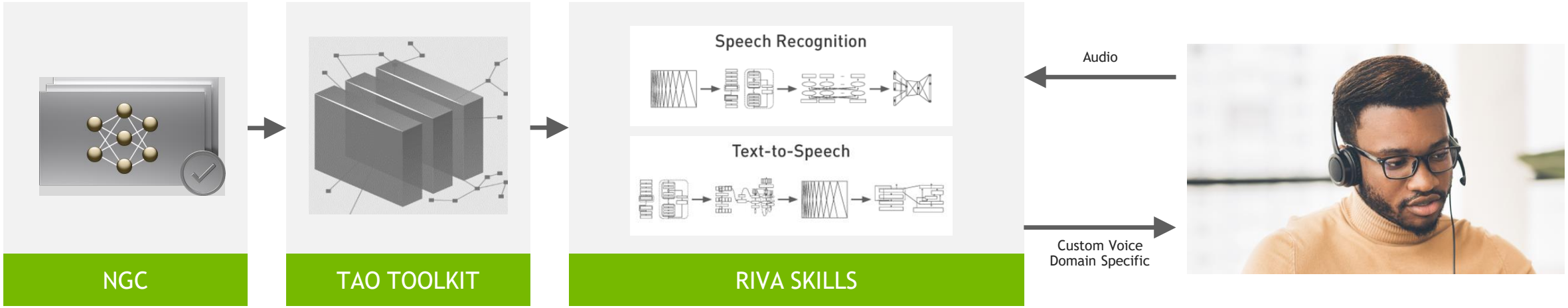



| HIGH ACCURACY | NO ACCESS TO SOTA* SPEECH MODELS | REAL-TIME PERFORMANCE | FLEXIBLE & SCALABLE DEPLOYMENT | DATA OWNERSHIP & PRIVACY |
|---|---|---|---|---|
| Best-in-class accuracy with Speech AI pipeline customization | World-class Deep Learning models & training data | Real-time latency delivery | Large scale deployment on-prem, cloud, and edge | Data storage and processing on customer infrastructure |

*SOTA – State-Of-The-Art

# NVIDIA RIVA

## GPU-Accelerated SDK for Speech AI




- World Class Speech Recognition and Text-to-Speech Skills
- Pre-trained SOTA models trained on 100,000 hours of DGX; Retraining with TAO toolkit (zero coding)
- Flexible customization from data to model to pipeline
- Deploy Services with one Line of code in cloud, on-prem & edge
- Scale to handle hundreds and thousands of real-time streams with <300 ms latency per stream

# PRE-TRAINED SPEECH AI MODELS

## Accurate State-Of-The-Art Models In NGC

Several speech and language pretrained models in NGC to get started

- SOTA models trained over 100,000 hours on NVIDIA DGX™
- Optimized for high-performance training and inference on GPUs
- Customizable with NeMo, fine-tunable with TAO Toolkit, deployable to Riva
- Used across apps such as chatbots, virtual assistants, & transcription services

Automatic Speech Recognition (ASR)

Jasper

Quartznet

Citrinet

Acoustic Model

BERT NER

BERT Punctuation

Post-Processing

Text-To-Speech (TTS)

Fastpitch

Tacotron

Spectrogram Generator

HiFiGAN

WaveGlow

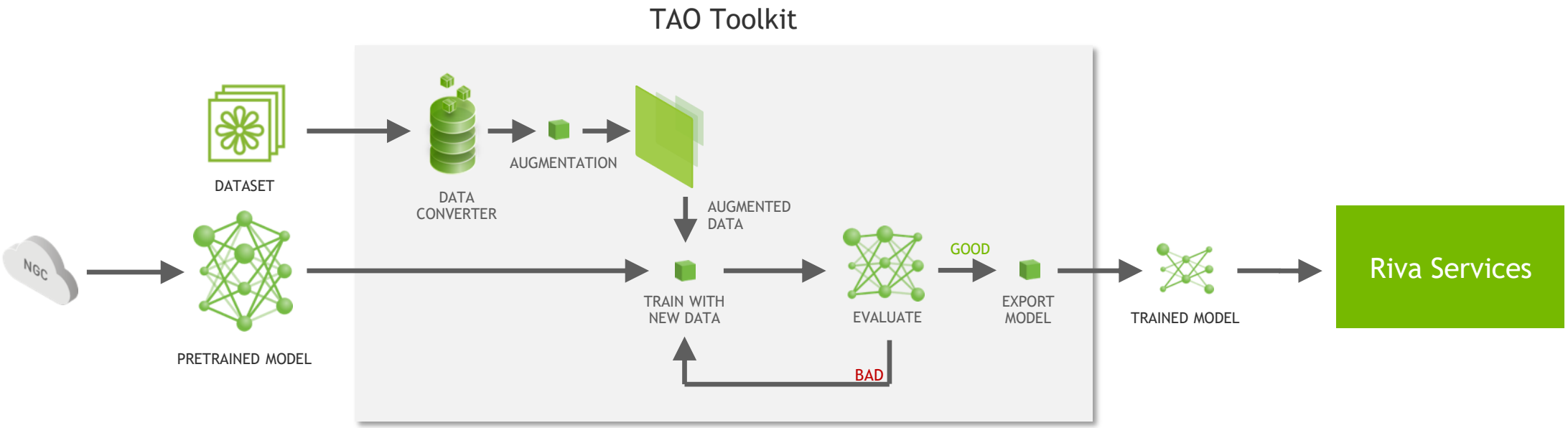
Vocoder

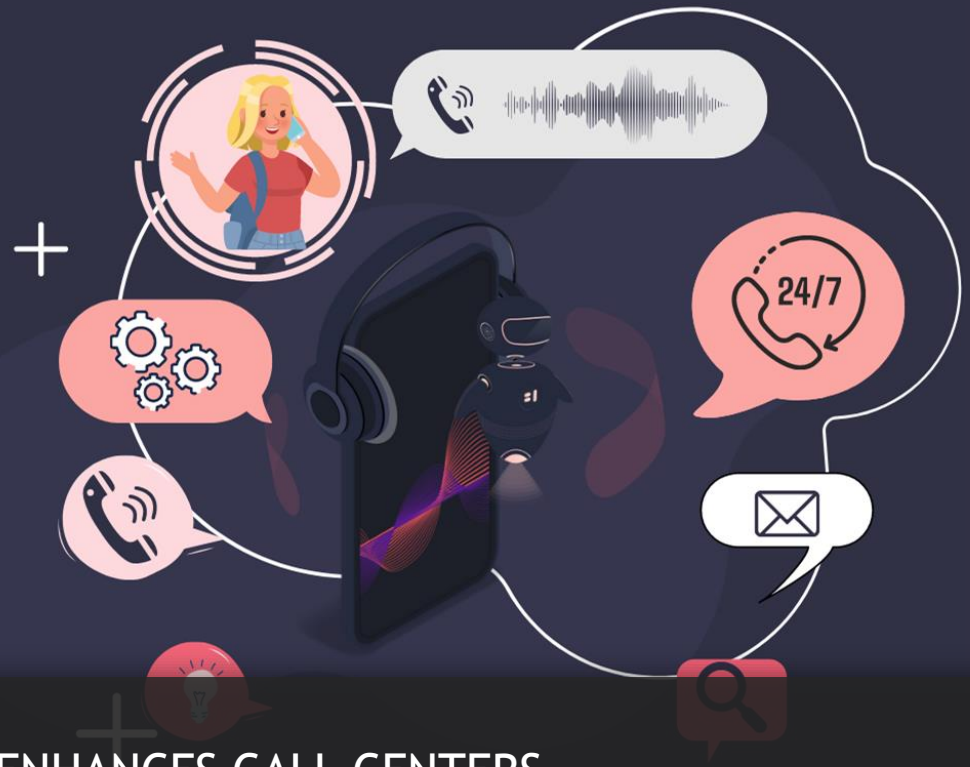
# TAO TOOLKIT WORKFLOW FOR CONVERSATIONAL AI

## Customize Models on Your Domain

Increase accuracy by fine-tuning on proprietary data:

- Zero-coding approach reduces barrier to entry for enterprises
- Use CUDA-X AI libraries, automatic mixed precision and Tensor Cores to achieve highest training performance
- Integrated with Riva to deploy fine-tuned models as real-time services






# ACCURATE, REAL-TIME SPEECH ENHANCES CALL CENTERS

To provide delightful customer experiences, automated call centers must have accurate automatic speech recognition (ASR) and real-time responses.

Floatbot, a leading unified voicebot and chatbot platform, uses the NVIDIA Riva SDK to build customized speech AI applications and deliver real-time performance for its customers in Singapore. The company uses the TAO Toolkit to fine-tune and update its models on data acquired through campaigns.

Using the NVIDIA SDKs, Floatbot reduced response time in Singaporean English by 38% — from 260ms to 162ms — and improved accuracy by 30%.

NVIDIA. Floatbot®

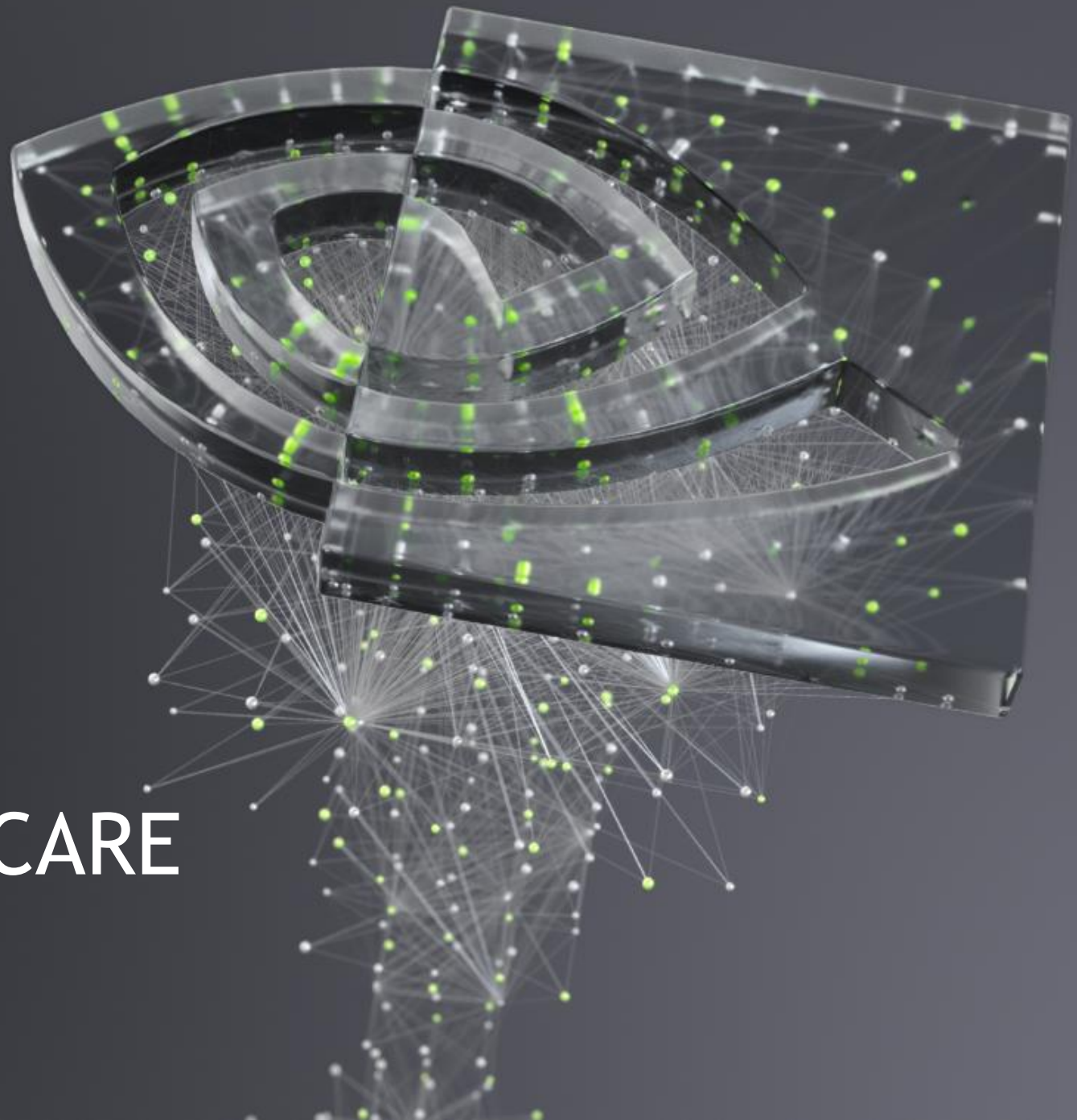
# AI-POWERED READING TUTOR

According to Education Week, 75% of all 4th graders in the U.S. read below their grade level. Early reading assessment and intervention are key to a student's success.

Plabook improves students' reading and comprehension skills through automated reading assessments.

Data Monsters used the NVIDIA Riva SDK to add speech skills that were customized on voice recordings from hundreds of children with varying accents and reading levels to Plabook.

Plabook is a timesaver for teachers. The AI manages students' recordings and highlights errors for the teacher to validate. The process requires minutes, versus hours, of a teacher's time for an entire class.

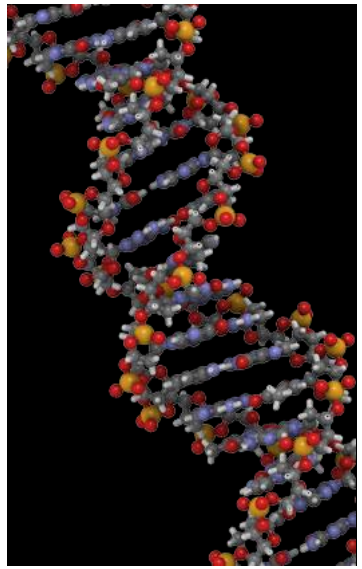


NVIDIA

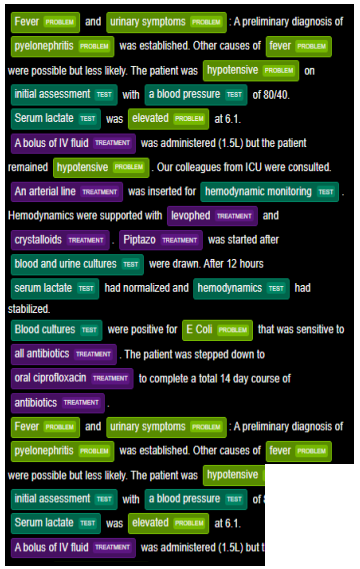
# CLARA FOR HEALTHCARE

# NVIDIA CLARA
# COMPUTATIONAL PLATFORM FOR HEALTHCARE

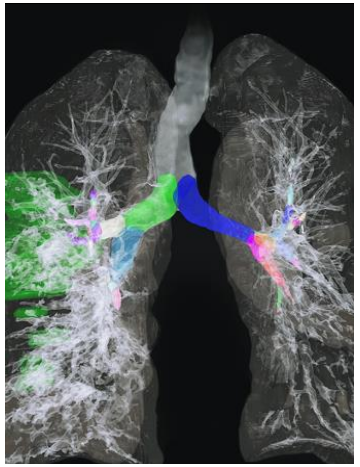
GENOMICS



NLP

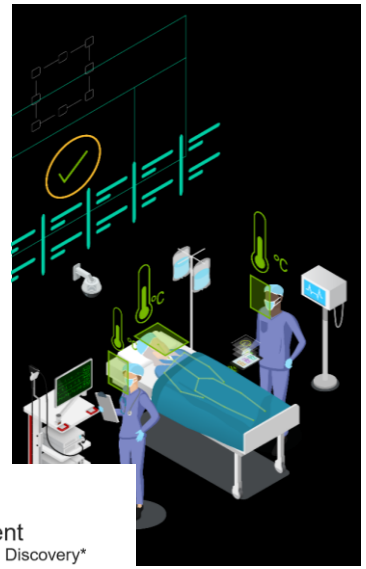



IMAGING

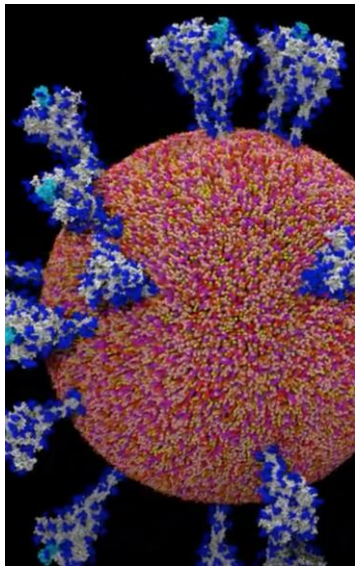


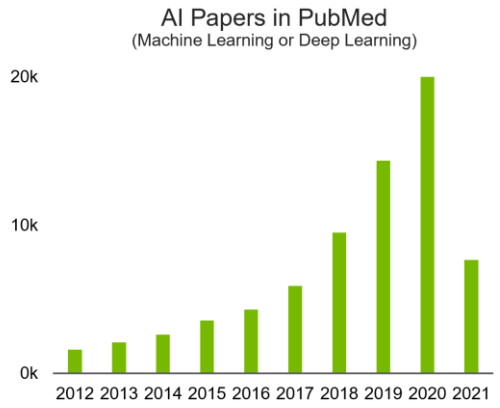
INSTRUMENTS

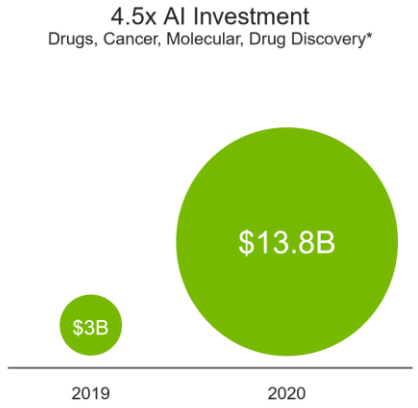

CONVERSATIONAL AI



DRUG DISCOVERY

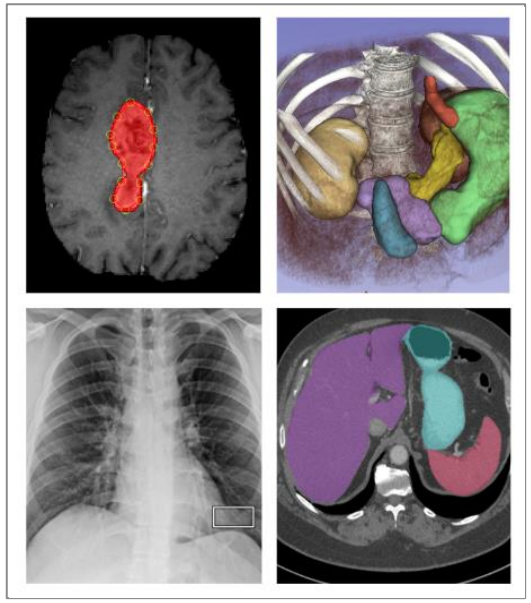




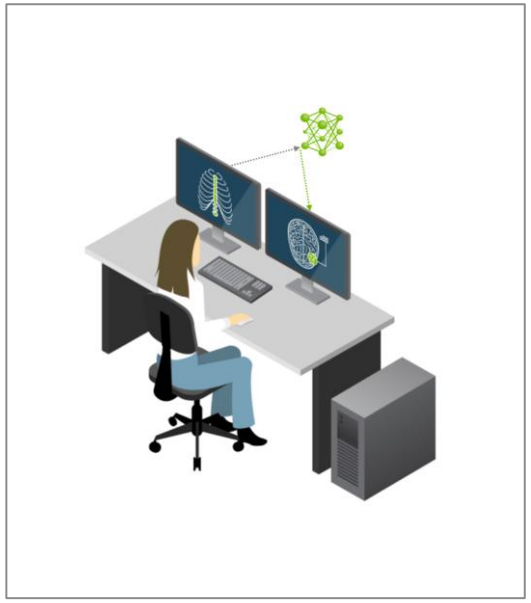



# NVIDIA CLARA IMAGING

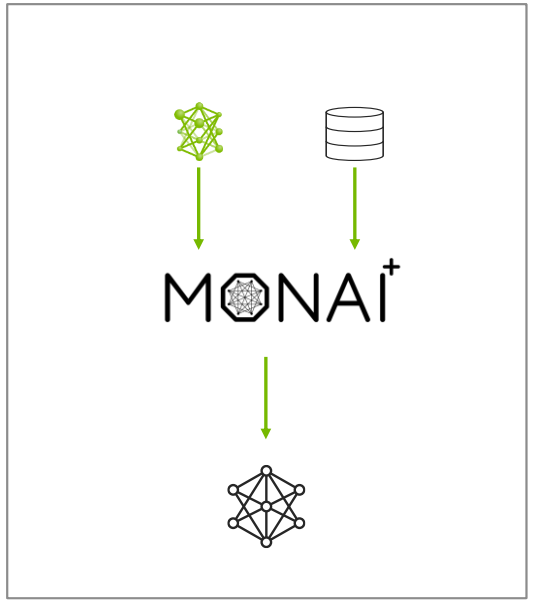
## End-to-End AI Application Framework - Development to Deployment



Pre-Trained Models



Labeling Data



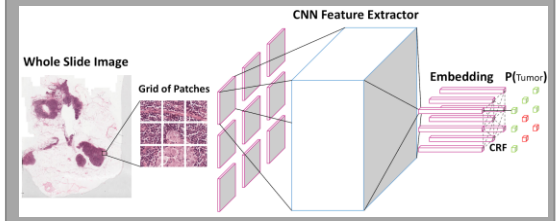

Training

Deployment

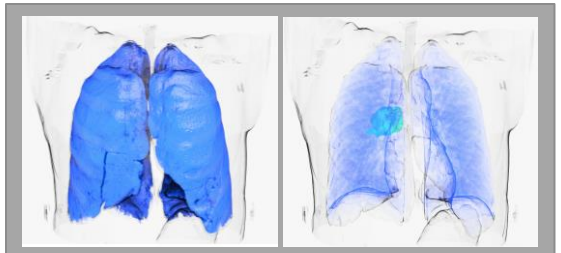
# NVIDIA CLARA PRE-TRAINED MODELS

## Save Thousands of Hours | Millions of Dollars

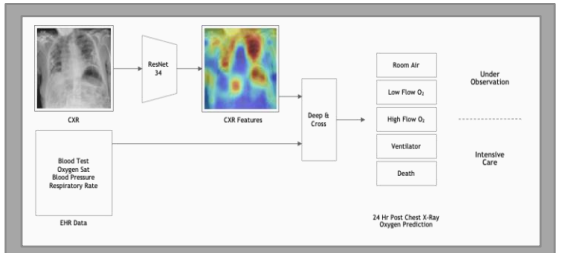
3D SEGMENTATION AND INTERACTIVE ANNOTATION
Spleen | Liver | Brain | Prostate | DeepGrow 2D & 3D



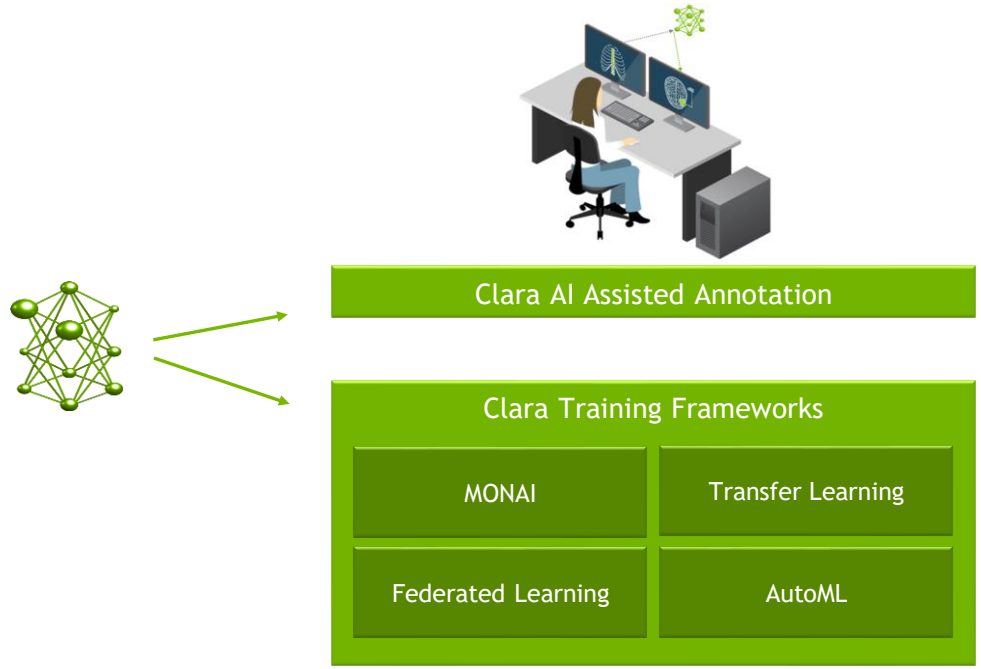

METASTSIS DETECTION - ResNet
Automated Detection of Metastasis WSI of lymph node sections



COVID-19 CT PROGNOSIS
Lung Seg | COVID Classification | Lesion Seg




COVID-19 OXYGEN PREDICTION
Model Developed by Federated Learning | 20 Global Sites




### Clara Imaging Pre-Trained Models in PyTorch

20+ Pre-Trained Models: CT, MRI, X-Ray, Digital Pathology

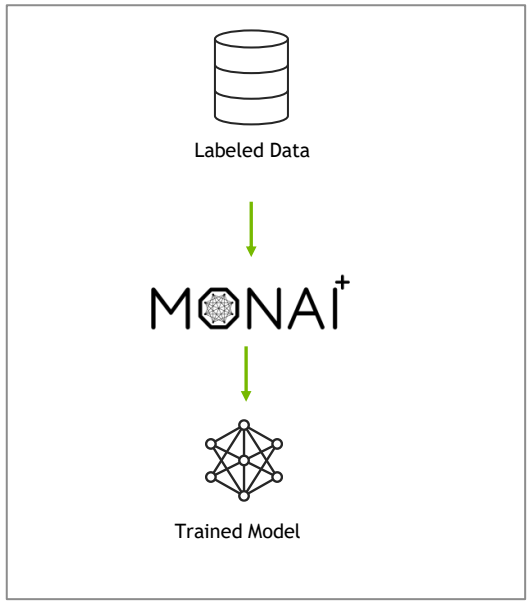
### Data Labeling & Model Training

Jumpstart Complex 3D Data Labeling | Reduce Training Data Needed

# NVIDIA CLARA TRAINING APPLICATIONS

## Open, Extensible & Domain Specialized Training Framework

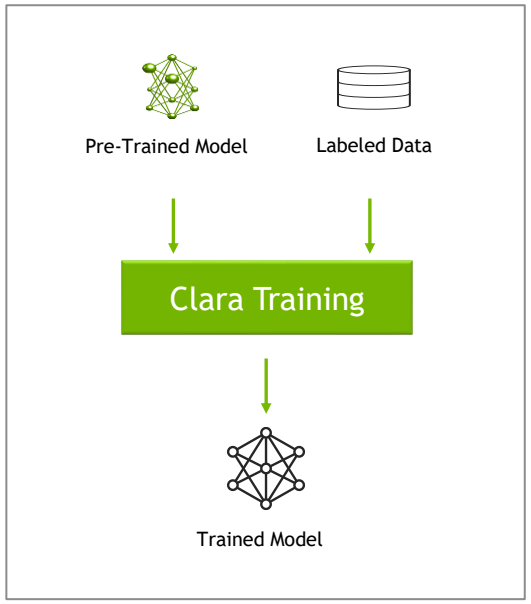
INVENT




### Optimized Training

6x Faster PyTorch Native
New Open-source foundation
Speed of light research collaboration

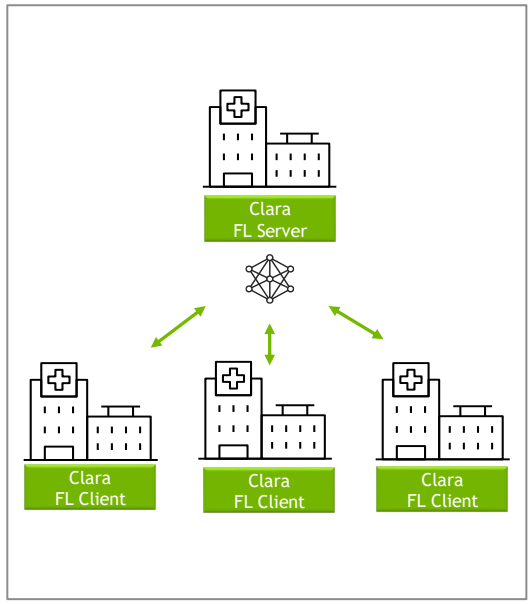
ADAPT




### Transfer Learning

New Pathology Pre-trained model
Reduce Training Time and Cost
Adapt for Local Environment

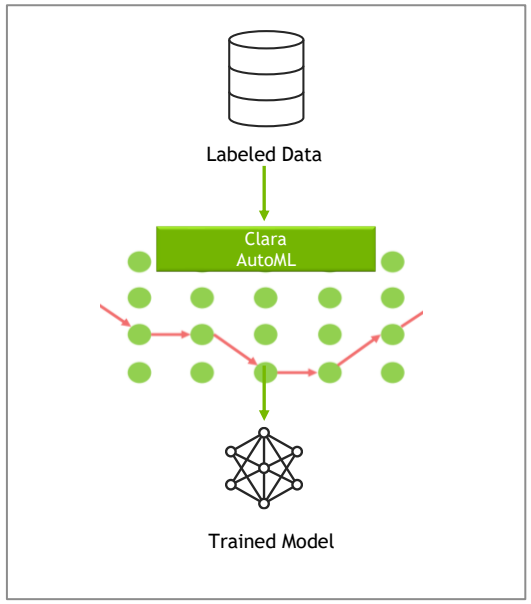
COLLABORATE




### Federated Learning

New Homomorphic Encryption
New Bring your own Trainer
Train w/o Sharing Data

AUTOMATE

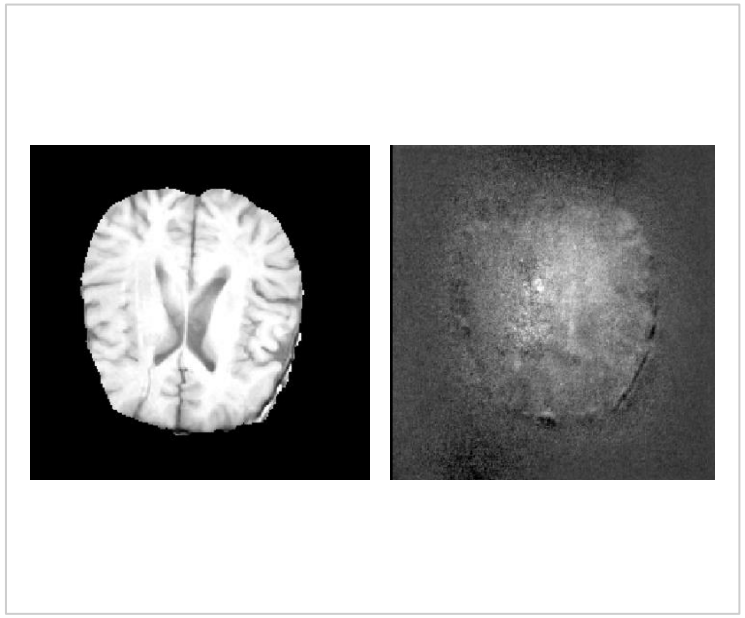


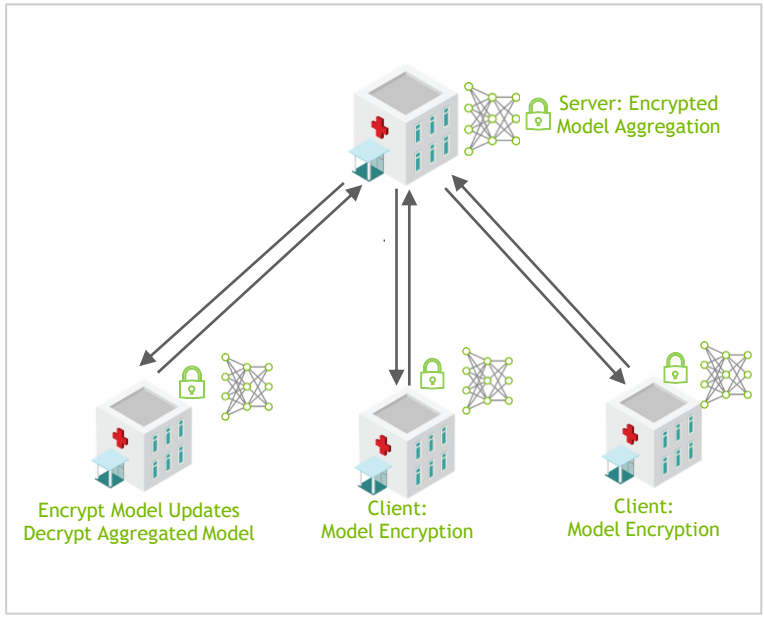

### AutoML

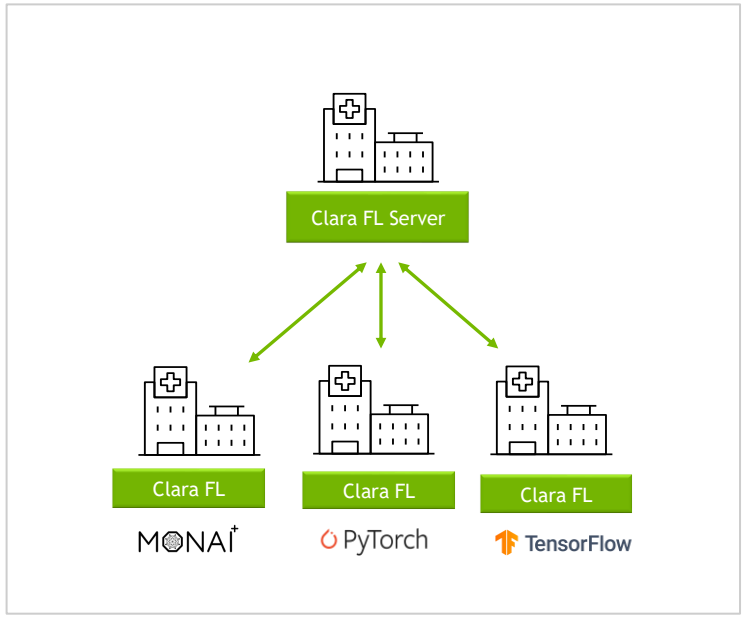
Automate Network Selection
Automate Hyperparameter Search

# CLARA FEDERATED LEARNING

## Privacy Preserving & Extensible Collaborative Learning






**DIFFERENTIAL PRIVACY**
Prevent data leakage

**HOMOMORPHIC ENCRYPTION**
Aggregation on Encrypted Models

**PRIVACY PRESERVING**
Collaborate without compromising privacy

**EXTENSIBLE**
Use Cases Beyond Imaging
Use Preferred Training Framework
Standalone Python Package for Easy Integration

# CLARA PARTNERS & INTEGRATORS

## From Academic Medical Centers to Enterprise Imaging























# NVIDIA CLARA NGC COLLECTIONS

## Containers, Models, Samples

Certified Containers

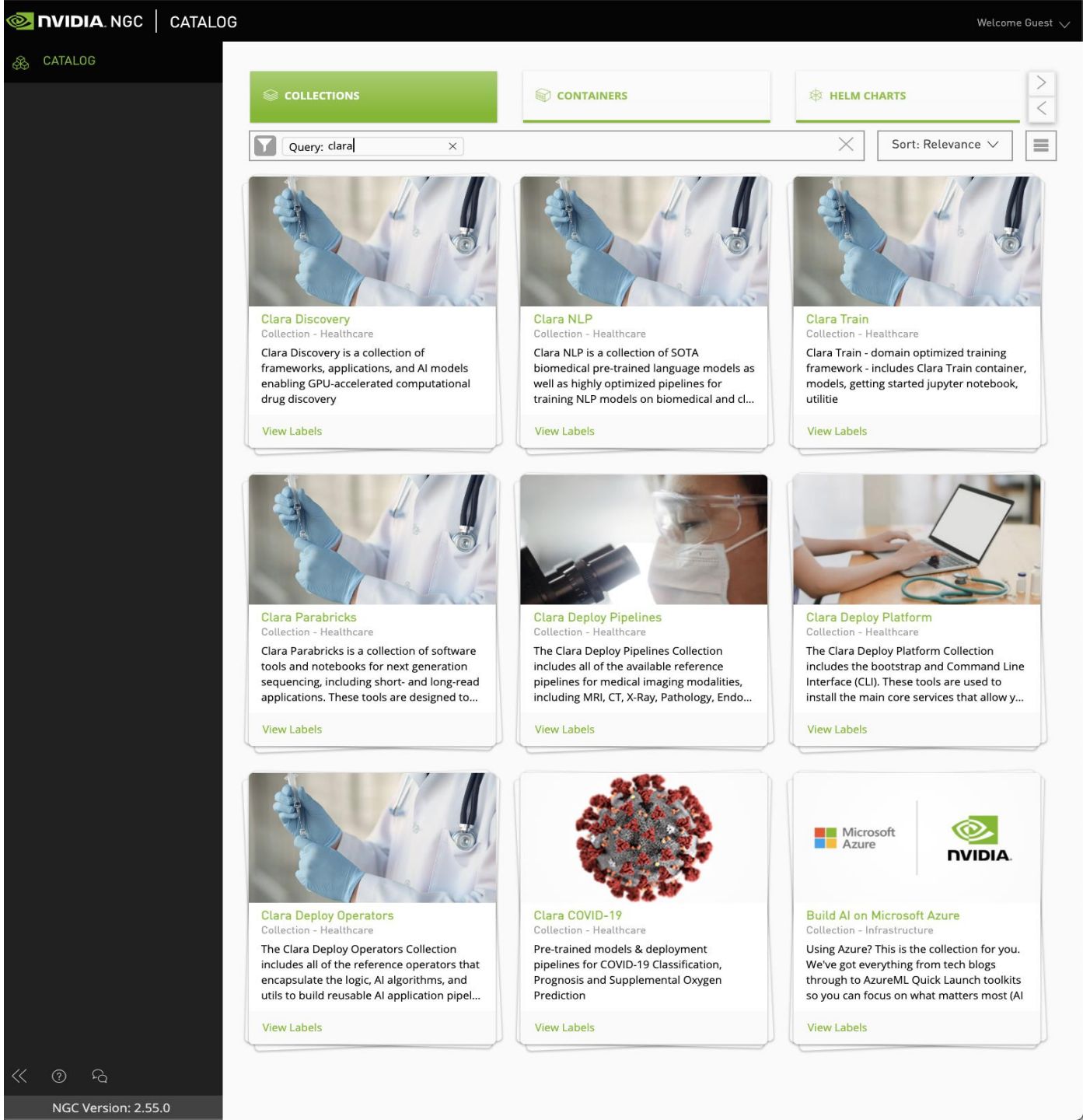
Pre-Trained Models

- COVID-19
- CT, MRI, X-Ray, Digital Pathology, NLP

Portable Workload Deploy On Prem and Cloud

NGC Catalog Now Available in AWS Marketplace

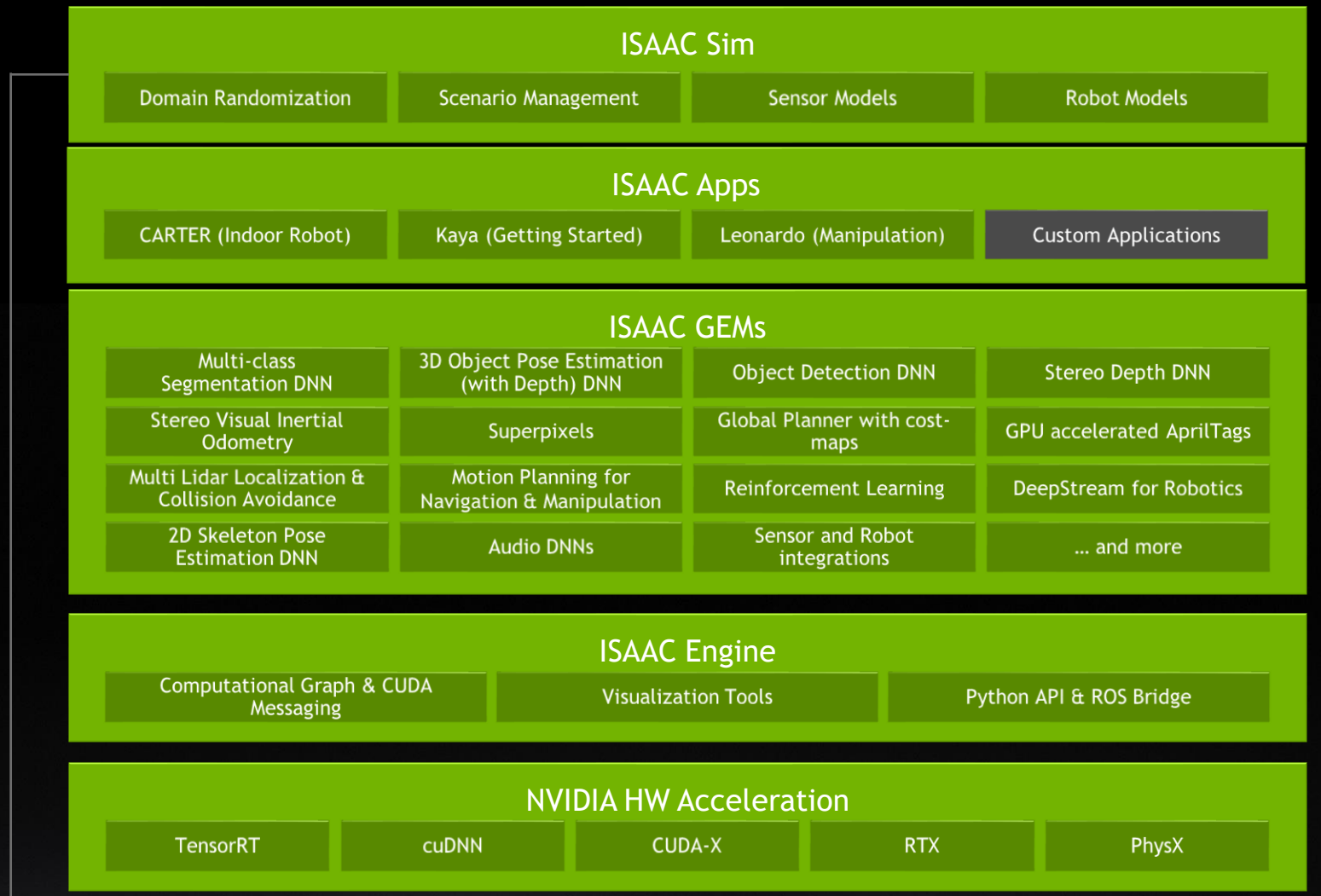
https://ngc.nvidia.com



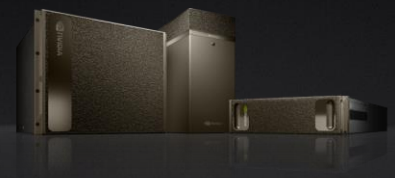
# INDUSTRIAL ROBOTICS

# NVIDIA ISAAC

- Isaac Engine
- Isaac GEMS
- Reference Designs
- Isaac Sim



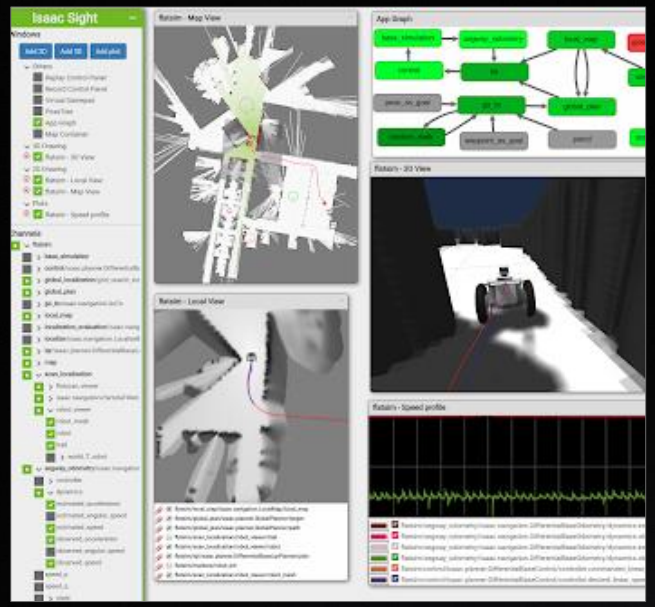

NVIDIA AGX



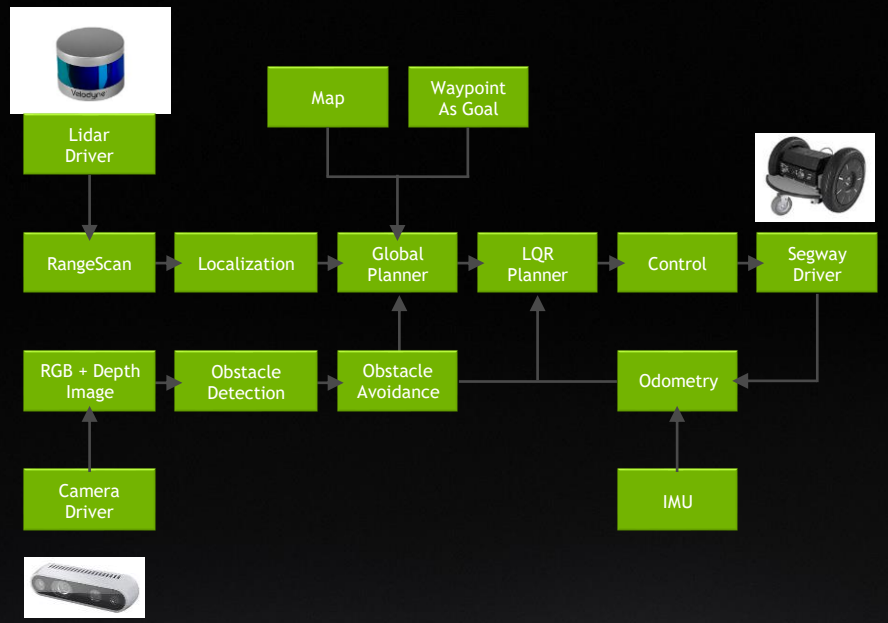
NVIDIA DGX

# ISAAC SOFTWARE

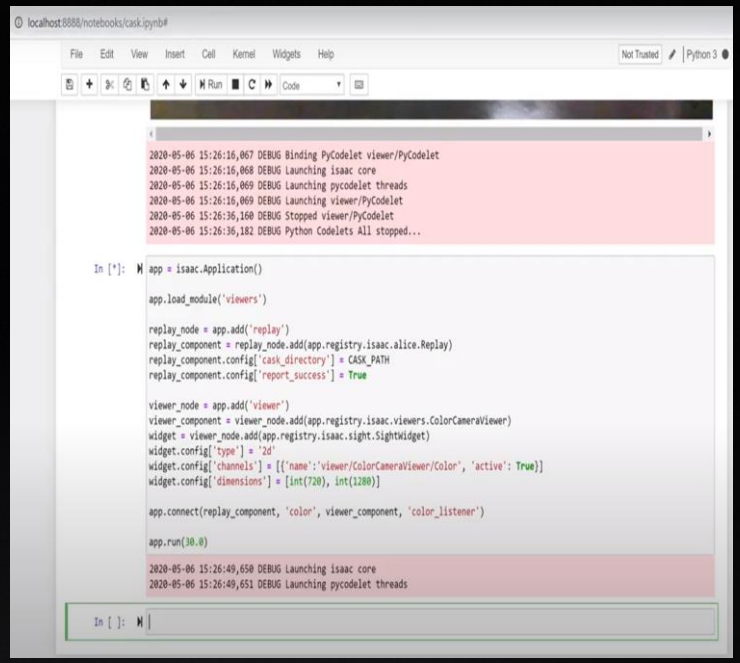
## Isaac Engine




Visualization Tool

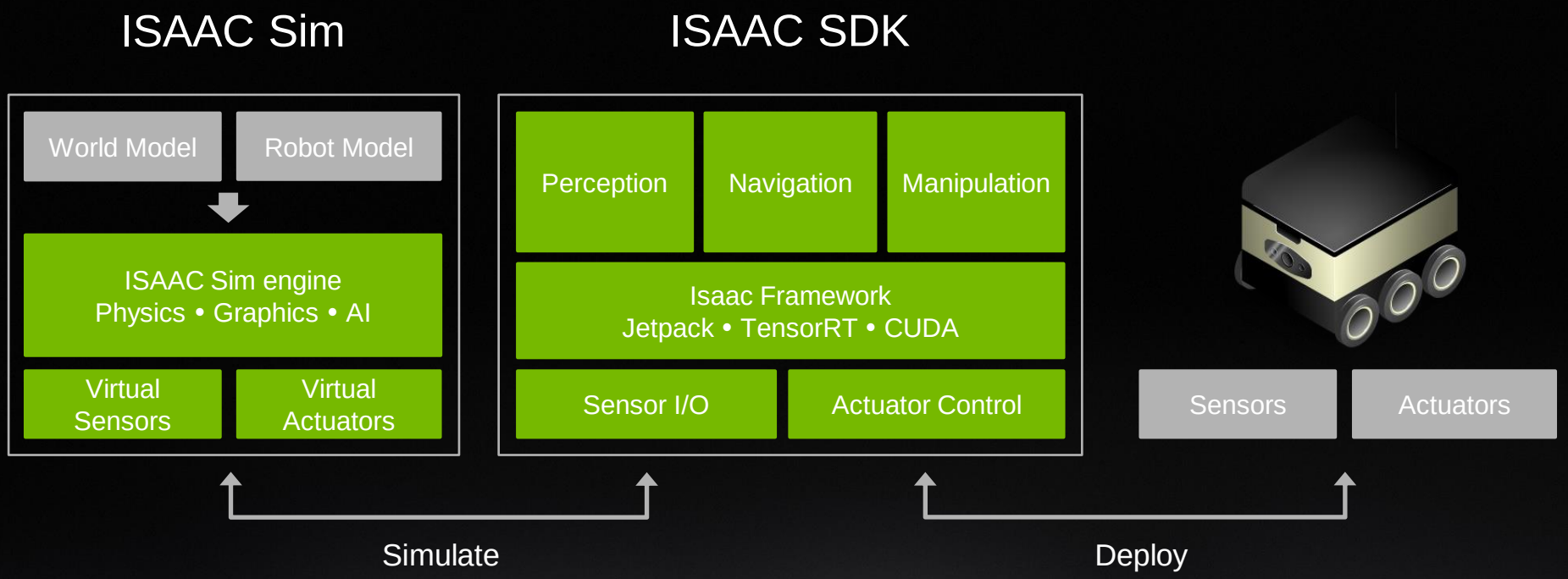


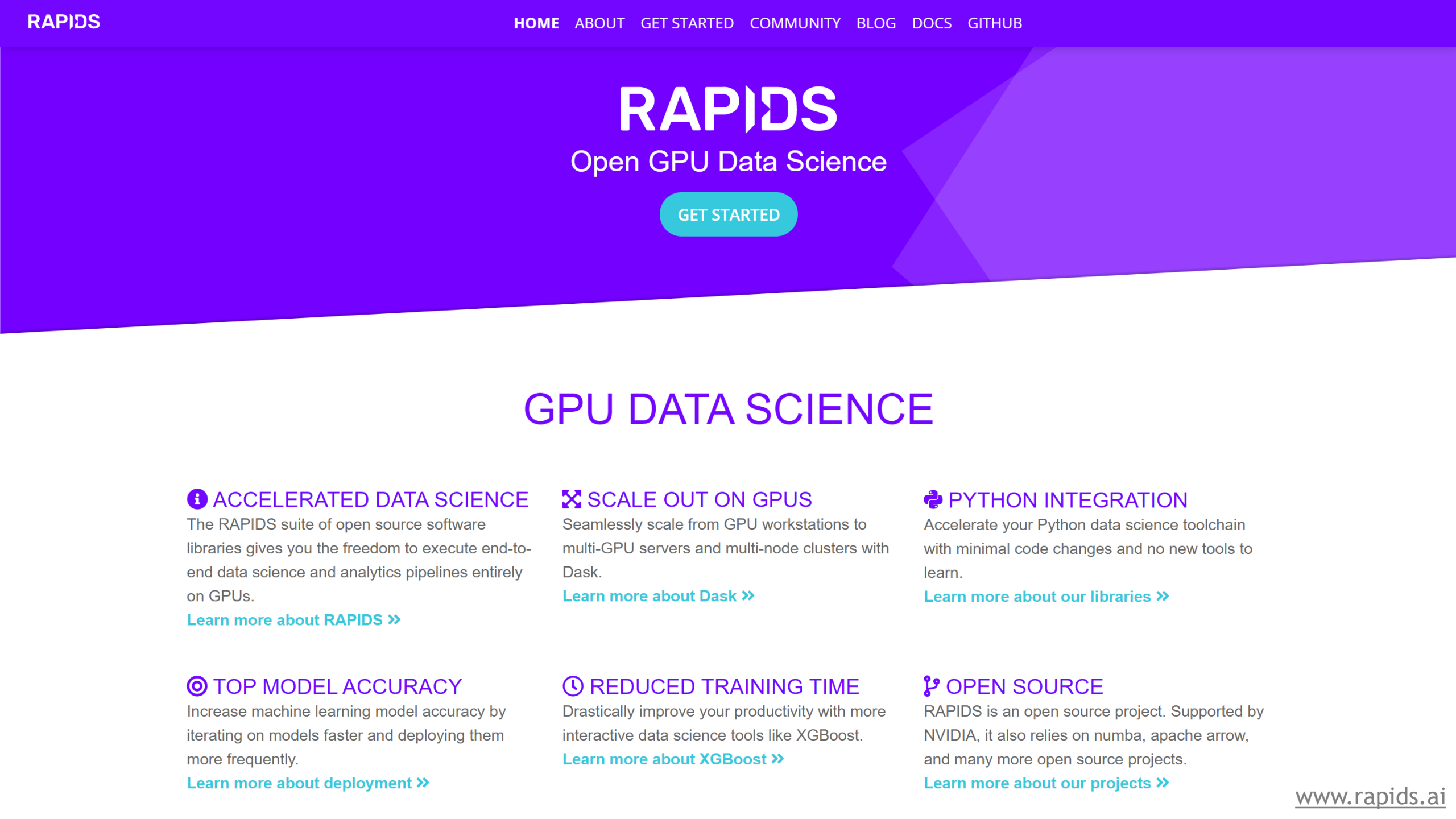

Computational Graph & CUDA Messaging



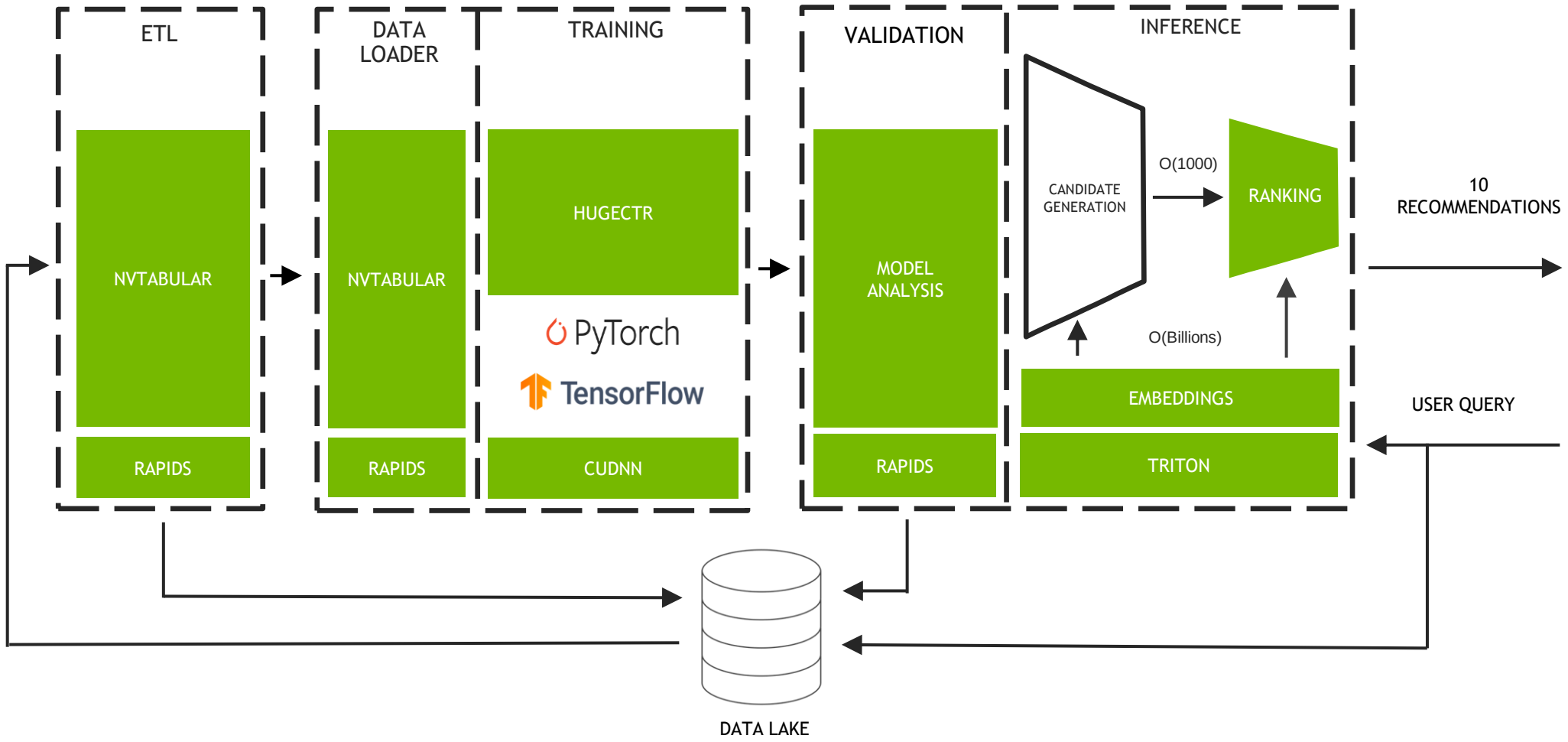

Python API

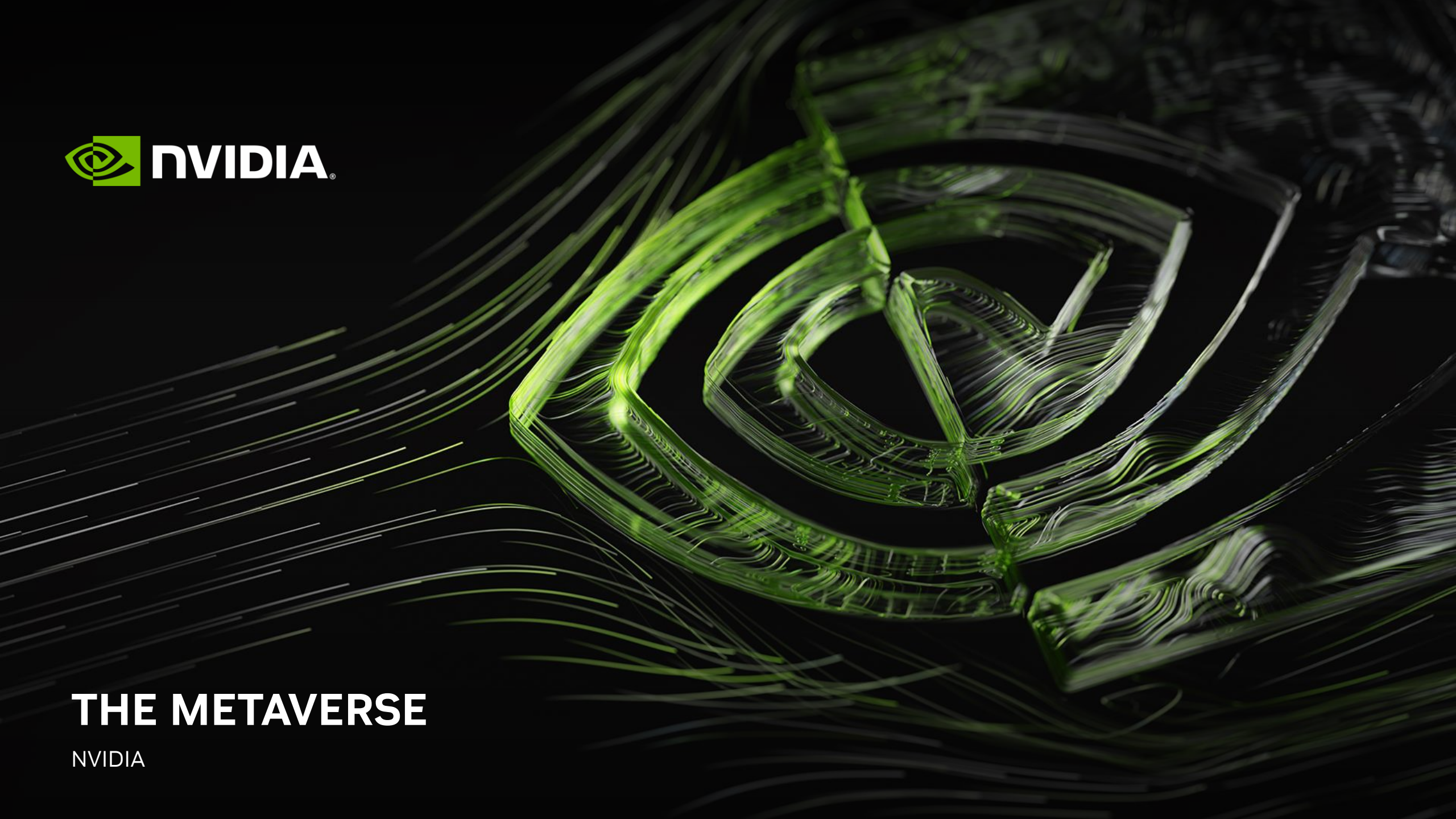
# ISAAC WORKFLOW




RAPIDS
HOME ABOUT GET STARTED COMMUNITY BLOG DOCS GITHUB
RAPIDS
Open GPU Data Science
GET STARTED
GPU DATA SCIENCE
ACCELERATED DATA SCIENCE
The RAPIDS suite of open source software libraries gives you the freedom to execute end-to-end data science and analytics pipelines entirely on GPUs.
Learn more about RAPIDS »
SCALE OUT ON GPUS
Seamlessly scale from GPU workstations to multi-GPU servers and multi-node clusters with Dask.
Learn more about Dask »
PYTHON INTEGRATION
Accelerate your Python data science toolchain with minimal code changes and no new tools to learn.
Learn more about our libraries »
TOP MODEL ACCURACY
Increase machine learning model accuracy by iterating on models faster and deploying them more frequently.
Learn more about deployment »
REDUCED TRAINING TIME
Drastically improve your productivity with more interactive data science tools like XGBoost.
Learn more about XGBoost »
OPEN SOURCE
RAPIDS is an open source project. Supported by NVIDIA, it also relies on numba, apache arrow, and many more open source projects.
Learn more about our projects »
www.rapids.ai

# NVIDIA MERLIN ACCELERATES EVERY STAGE IN RECOMMENDER PIPELINE

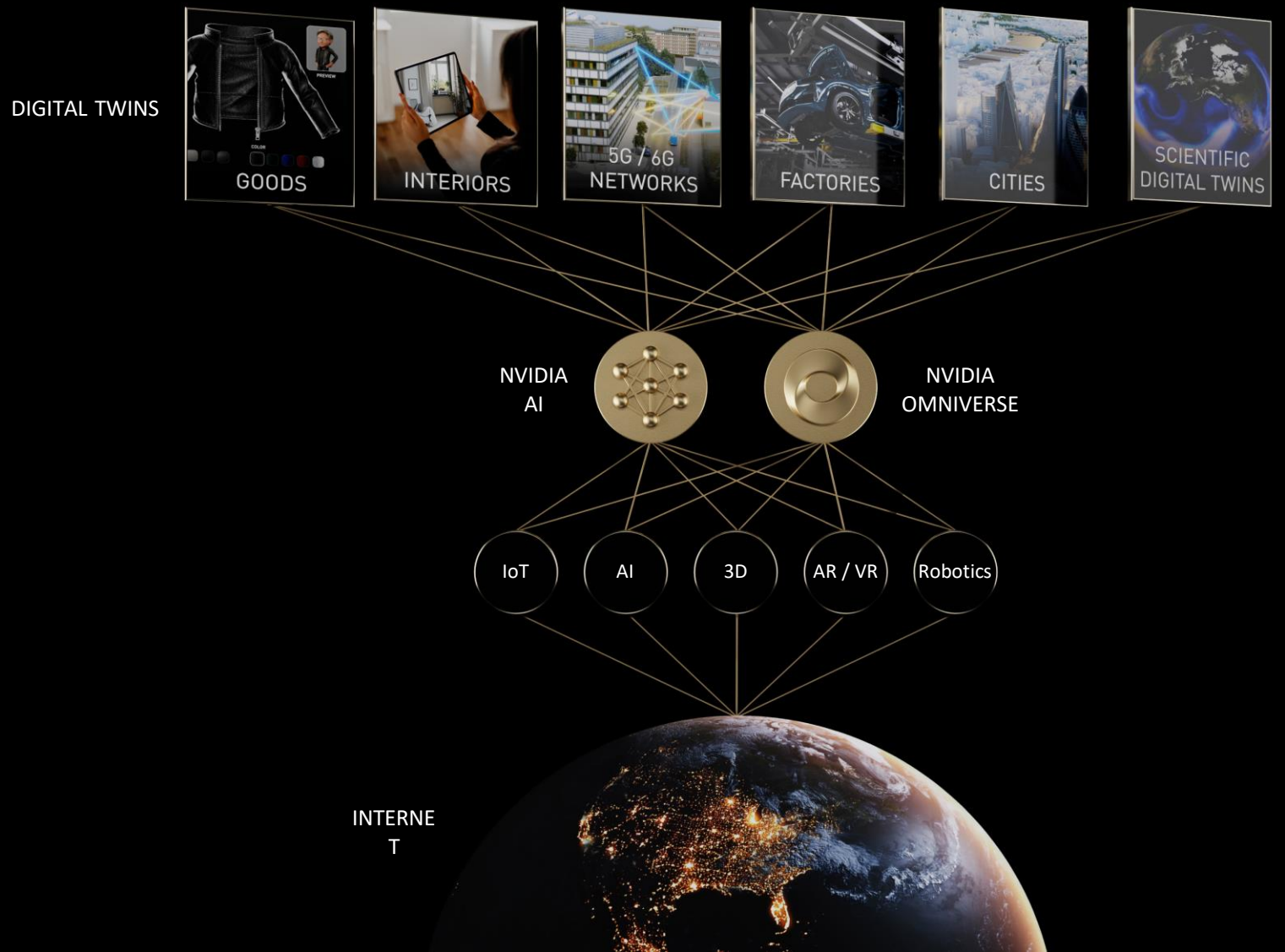




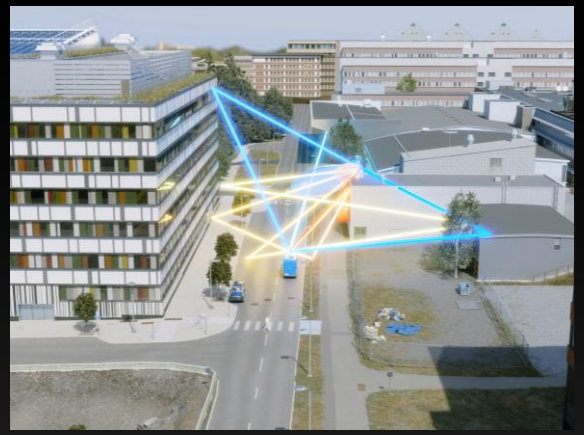

# THE METAVERSE

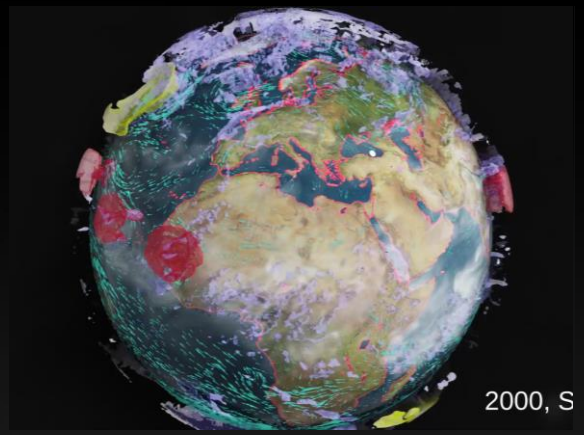
NVIDIA

# THE METAVERSE IS THE 3D EVOLUTION OF THE INTERNET



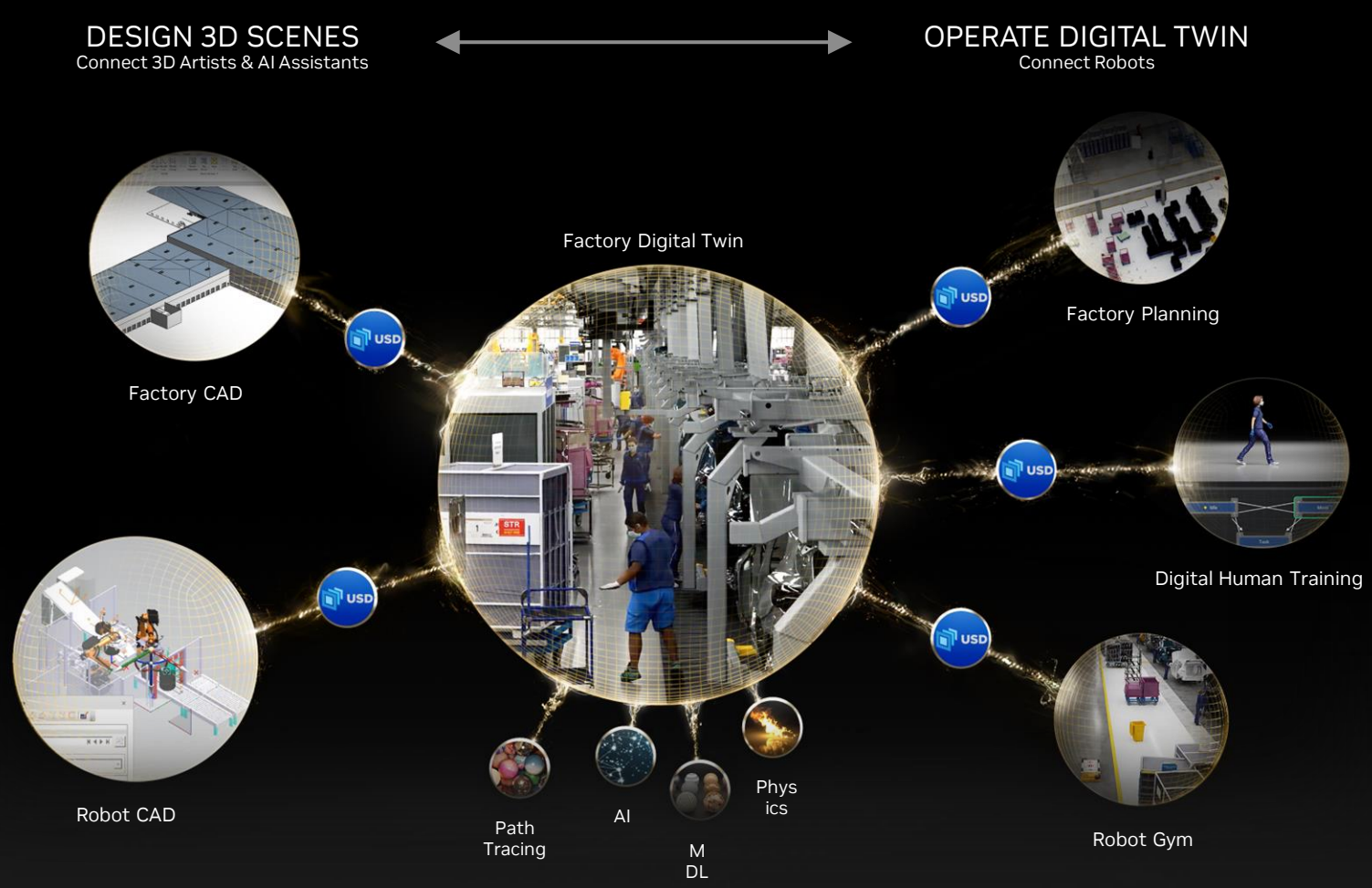
# METAVERSE APPLICATIONS ARE ALREADY HERE TODAY

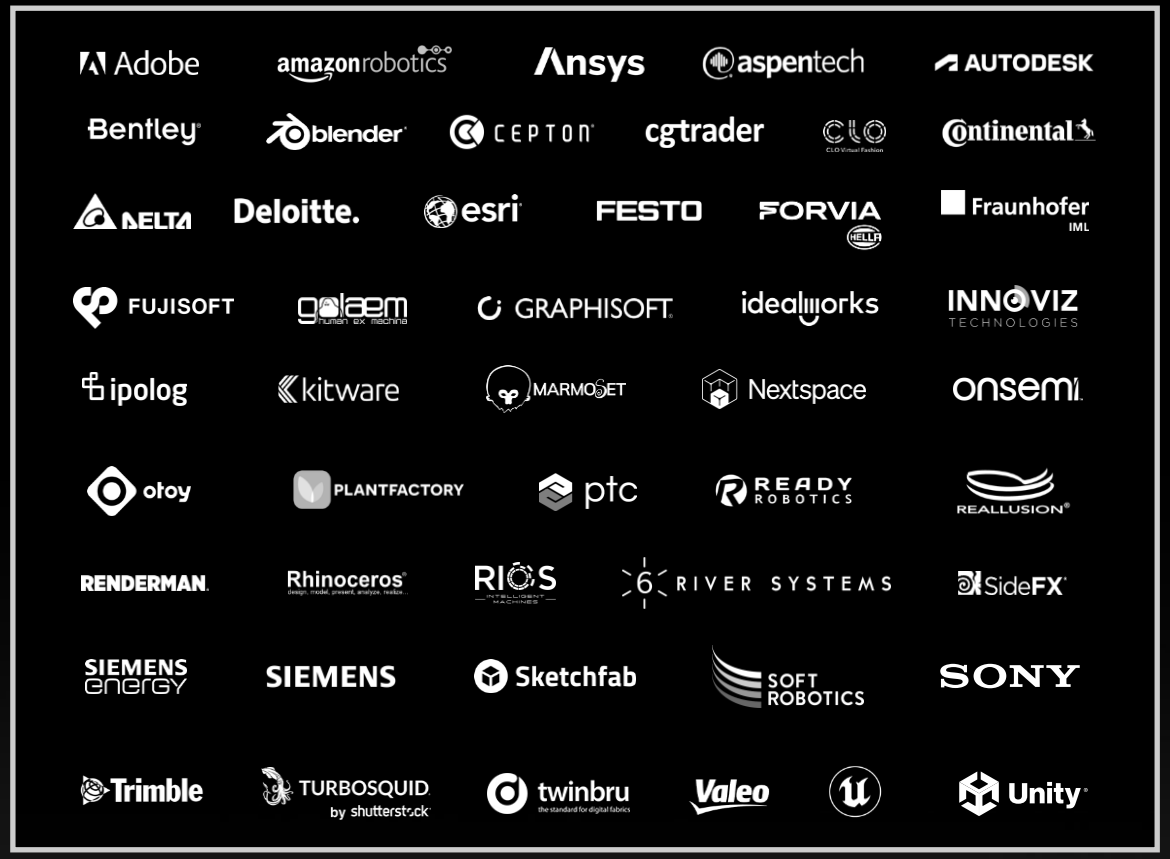
# NVIDIA OMNIVERSE

Computing Platform for Creating and Operating Metaverse Applications



# Omniverse is Already Connecting the World's Industries

## Building the Metaverse Together




Software Partners

Over 150 Universal Scene Description (USD) Connections Across Industry Applications


Adopters

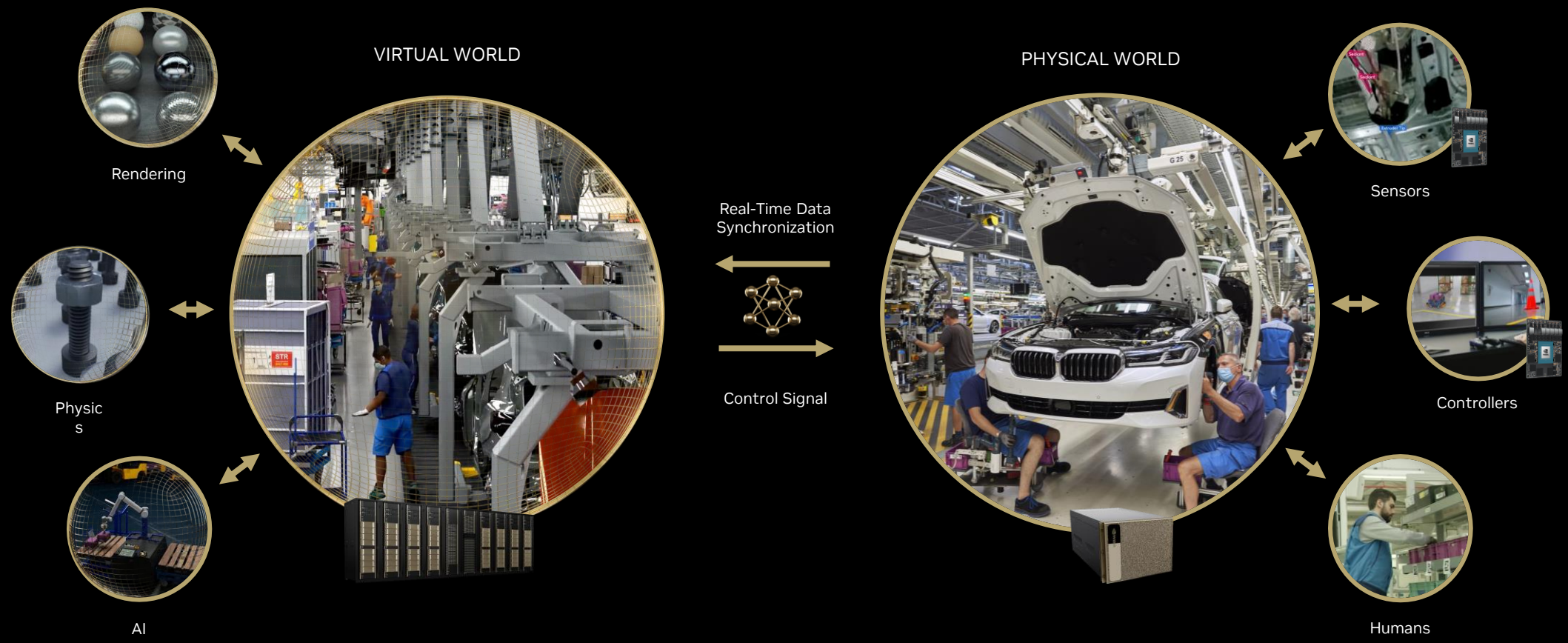
Across Transportation, Retail, Manufacturing, Energy, Telco, and More

# METAVERSE APPLICATIONS ARE ALREADY HERE

| | | | |
|---|---|---|---|
| 3D DESIGN OF GOODS, CONTENT, ENVIRONMENTS | DIGITAL TWINS<br>FOR INDUSTRIAL & SCIENTIFIC USE CASES | TRAINING PERCEPTION AI<br>ROBOTICS, AUTONOMOUS VEHICLES, CV NETWORKS | AVATARS<br>AI OR TELEOPERATED |

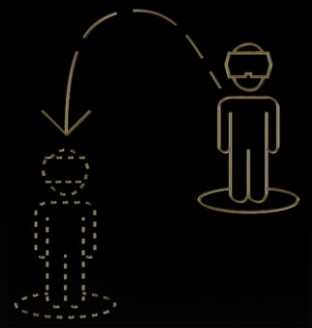
# The Metaverse Unlocks A New Class of Simulation for Enterprises

## Virtual World Simulations Live-Linked to the Physical World

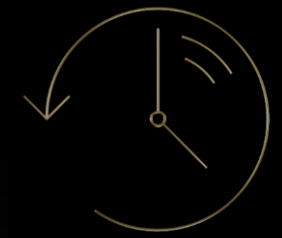


# DIGITAL TWINS GIVE ENTERPRISES SUPERPOWERS

TELEPORTATION

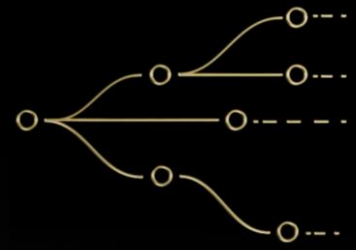


TRAVEL TO THE PAST



TRAVEL TO THE FUTURE

EXPLORE ALTERNATE FUTURES

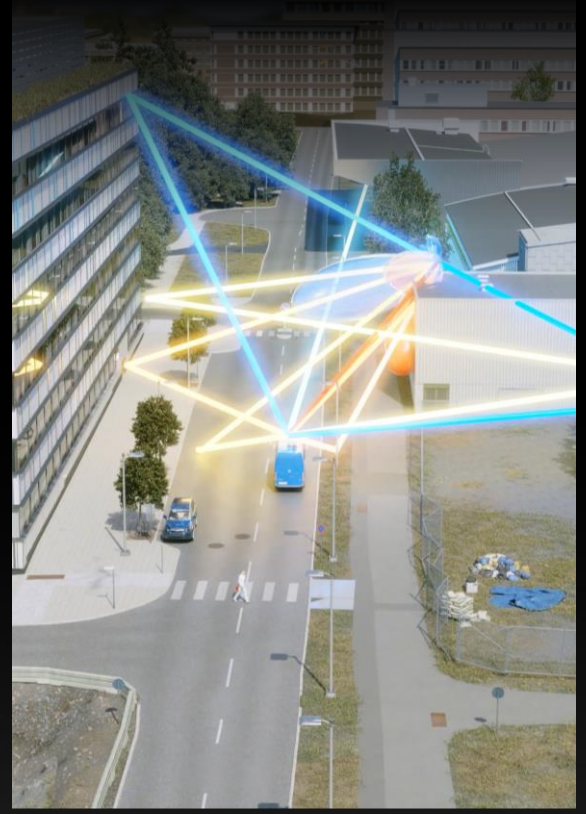


# DIGITAL TWINS WILL ONE DAY EXIST AT EVERY SCALE

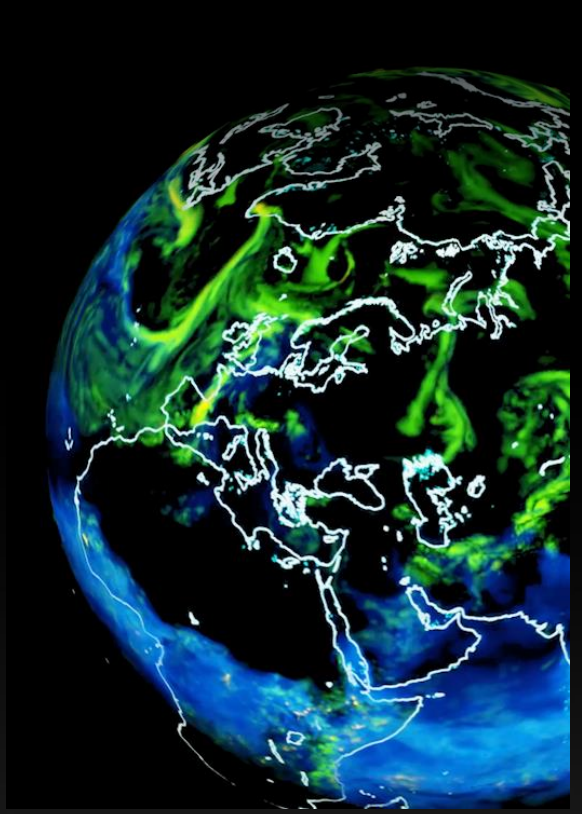
PRODUCT

FACTORY

CITY



PLANETARY



# AVATARS & DIGITAL HUMANS

HOW CAN I HELP?

Vision

Conversation

Speech

Planning & Action

Face Animation

Gesture

Realistic Graphics

Body Animation

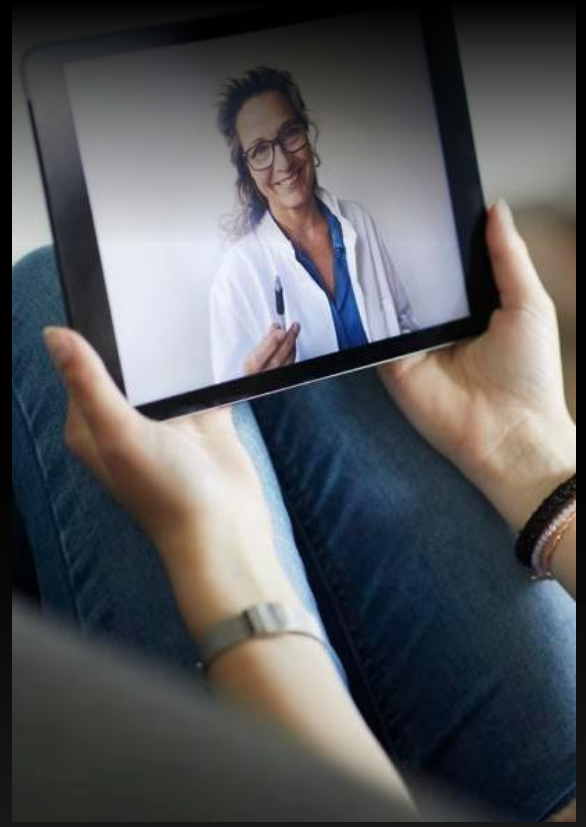
# AVATARS WILL EXIST FOR EVERY APPLICATION

Autonomous or Teleoperated – Realistic, or Fantastical

## GAMING

## CUSTOMER SERVICE


## HEALTHCARE



## WEB CONFERENCING

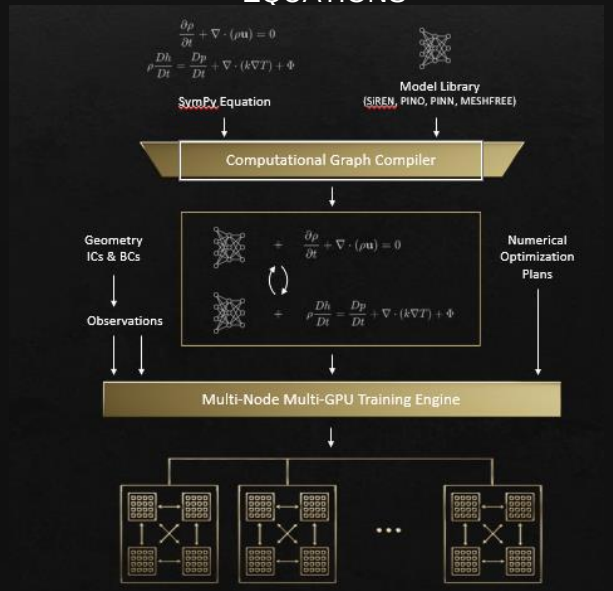

# NVIDIA MODULUS

HTTP://DEVELOPER.NVIDIA.COM/MODULUS

# NVIDIA MODULUS

Framework for developing physics machine learning neural network models

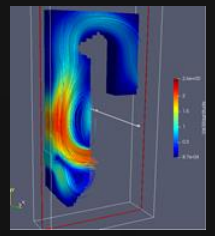
## TRAINING USING BOTH DATA AND THE GOVERNING EQUATIONS

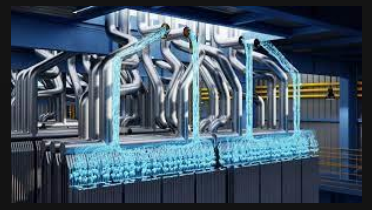



Modulus Latest Release (v. 22.03) Key Highlights:

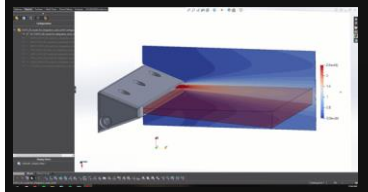
- FNO/AFNO integration to create climate physics-ML models
- Omniverse integration to visualize, infer, and interact in real-time with physics-ML model outputs

## DEVELOP HIGH FIDELITY DIGITAL TWINS

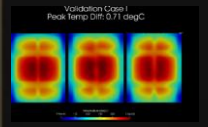


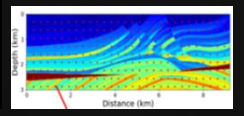


NETL and Siemens Energy use Modulus to build Digital Twins – high fidelity surrogate models – 10,000x faster



Using parameterized models for design optimization





Generalizable methodology for different domains such as fluids, Solid Mechanics, Multiphysics, etc.

## ADOPTION BY LEADING RESEARCH INSTITUTIONS

SIEMENS energy

NATIONAL ENERGY TECHNOLOGY LABORATORY

kinetic vision

## COLLABORATION PARTNERS

BROWN

Caltech

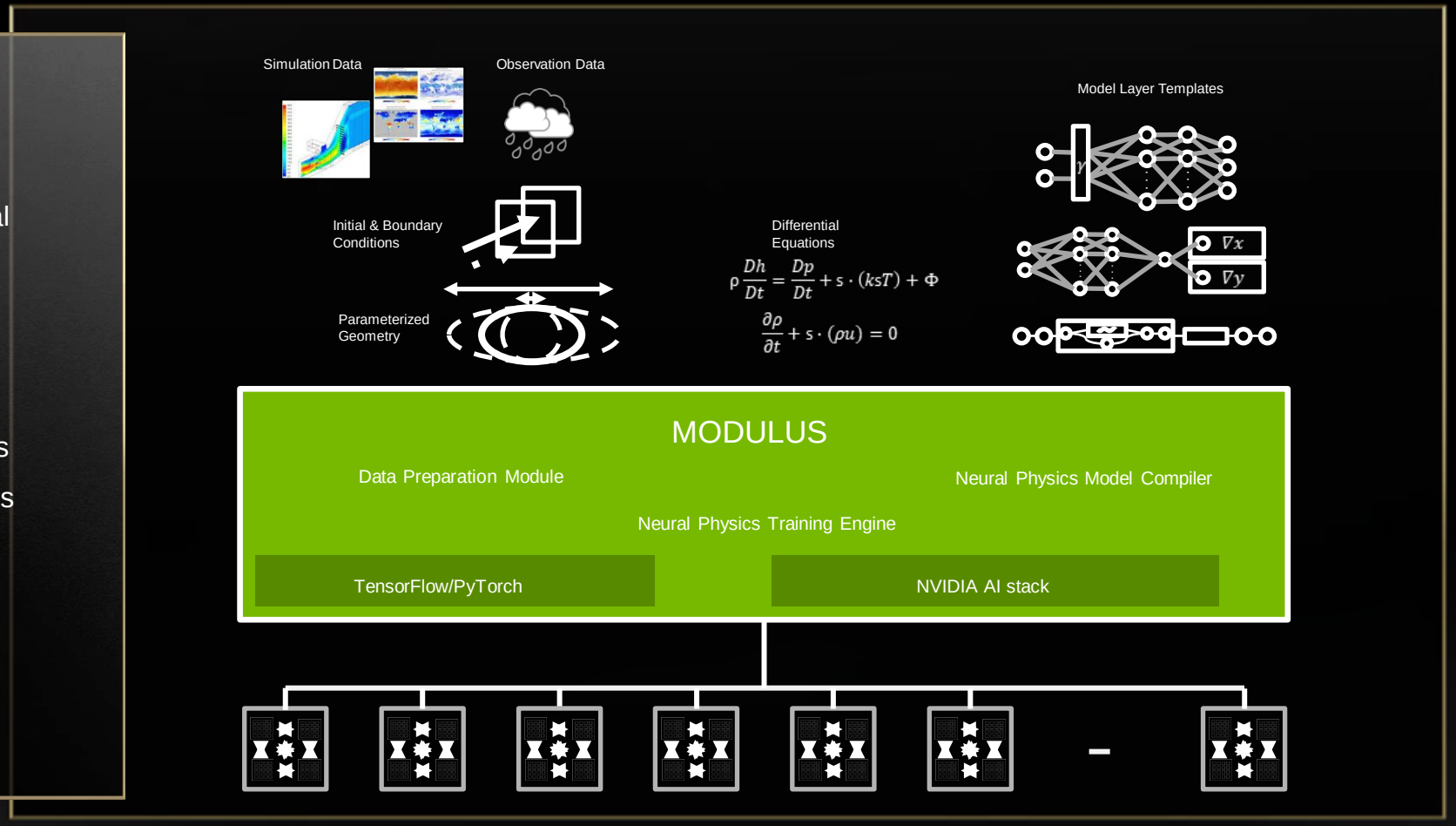
rescale

# NVIDIA MODULUS

Provides a customizable platform to train neural network models using governing equations to predict evolution of multi-scale, multi-physics systems

Generalizes parameterized domain and physics to encapsulate multiple configurations/scenarios in the trained model

Builds a Physics ML model to iterate on the design/operating space



# MODULUS APPLICATIONS

# WIND TURBINE WAKE OPTIMIZATION — SIEMENS GAMESA

## Use Case

- Developing optimal engineering wake models to optimize wind farm layouts
- Simulating the effect that a turbine might have on another when placed in close proximity

## Challenges

- Generating high-fidelity simulation data from Reynolds-averaged Navier-Stokes (RANS) or Large Eddy Simulations (LES) can take over a month to run, even on a 100-CPU cluster.

## Solution

- NVIDIA Omniverse and Modulus enable accurate, high-fidelity simulations of the wake of the turbines, using low-resolution simulations as inputs and applying super resolution using AI.

## NVIDIA Solution Stack

- Hardware: NVIDIA A100, A40, RTX 8000 GPUs
- Software: NVIDIA Omniverse, NVIDIA Modulus

## Outcome

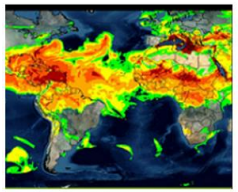
- Approximately 4,000X speedup for high-fidelity simulation
- Optimizing wind farm layouts in real-time increases overall production while reducing loads and operating costs.

Link to Demo

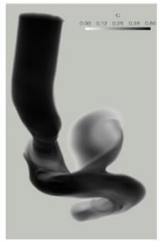
NVIDIA SIEMENS Gamesa RENEWABLE ENERGY

# CATEGORIZATION OF USE CASES

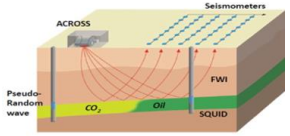
## Inverse & Data Assimilation Problems



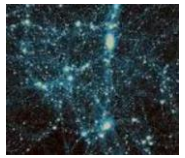
Climate



Medical Imaging

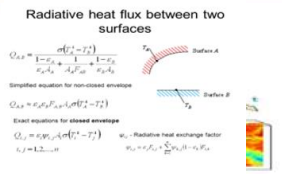



Oil & Gas

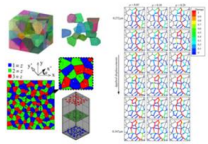


High Energy/
Nuclear Physics

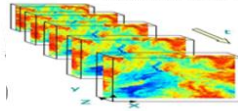
## Improved Physics & Predictions



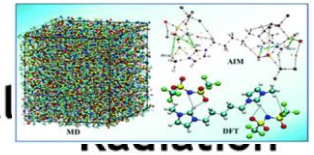

Radiation



Micro-mechanical
Material Model



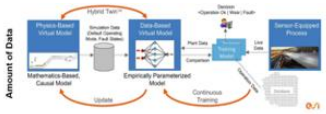
Turbulence



Molecular Dynamics

## Real Time Simulations

Robotics

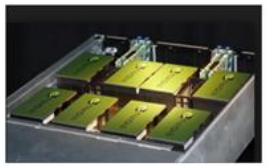


Digital Twin

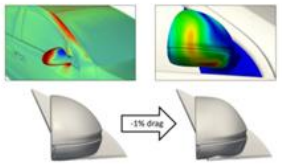
Autonomous
Ride & Handling

Games

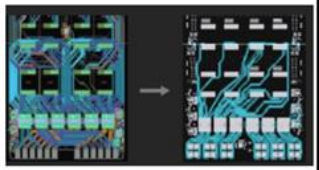
## Digital Design & Manafacturing



Heat Sink




Aerodynamics



Vias on a PCB

*Physics & Data - No Traditional Solver*

*Physics - Traditional Solver (Speed is a limitation)*

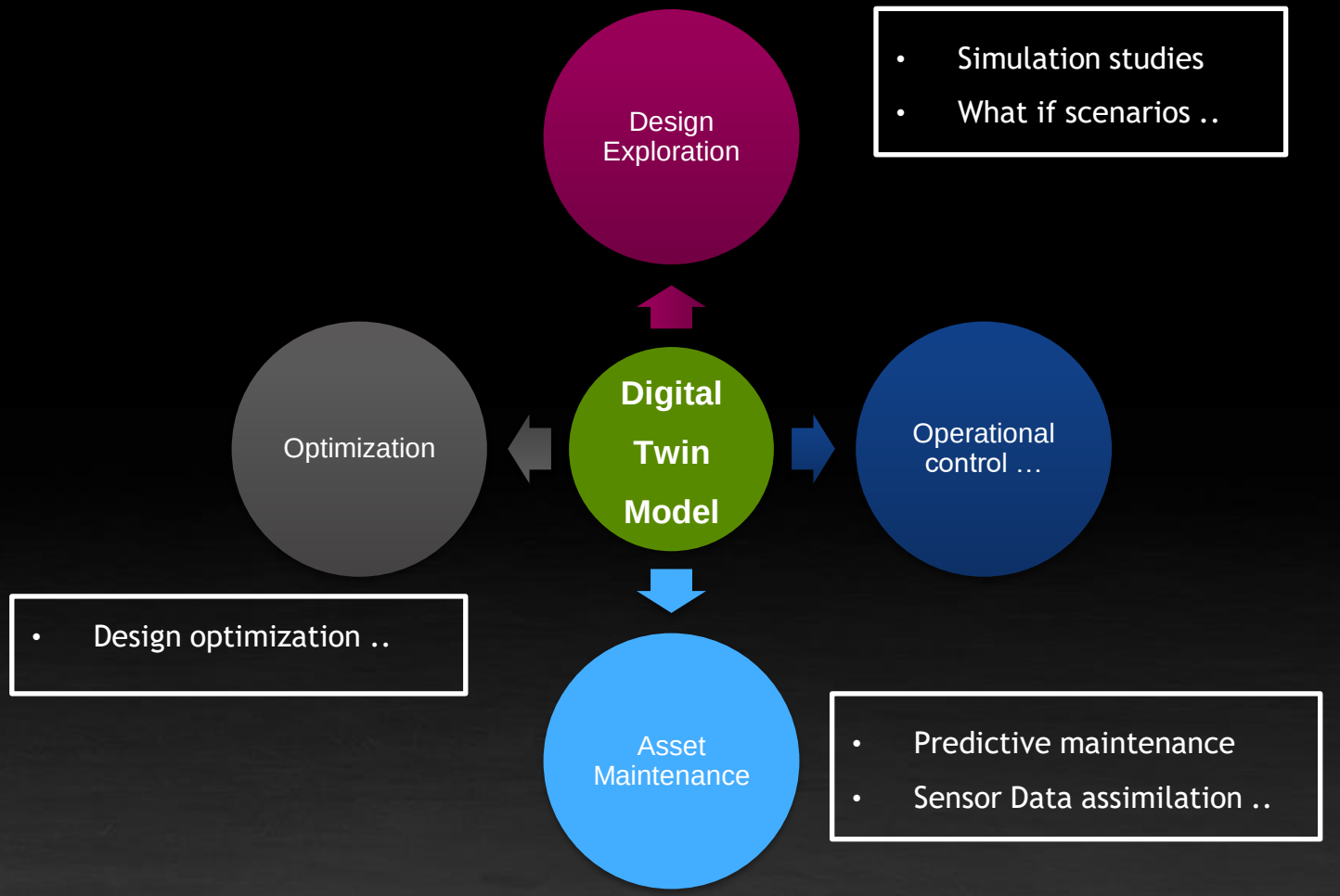
# MODULUS AND DIGITAL TWINS

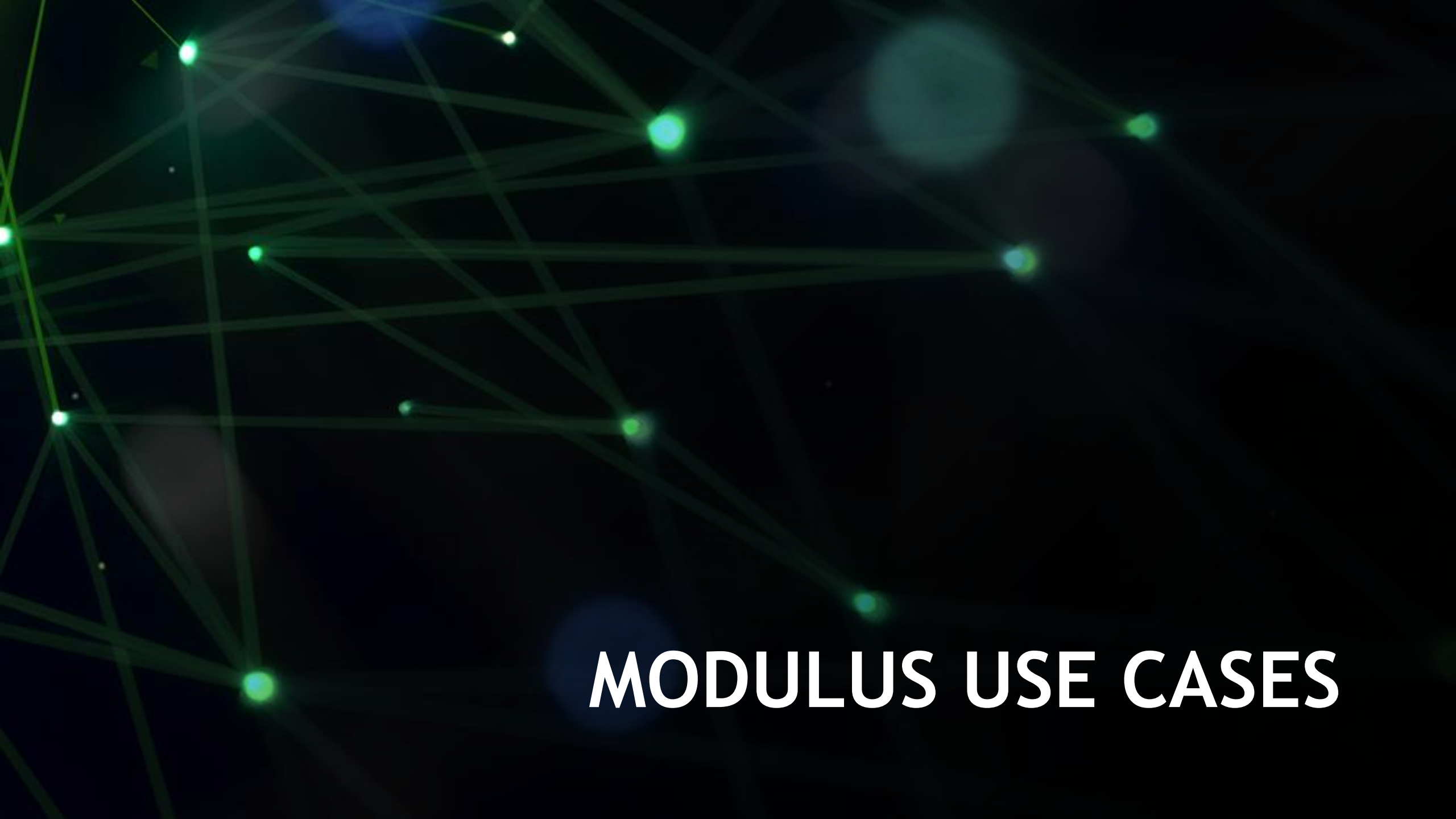
Digital twins are key underpinnings to the digital transformation

- For better product design
- For better maintenance of assets

AI and Physics combined can create robust digital twins to accelerate this digital transformation

Modulus trained physics ML models are 1000 x – 100000x faster while providing accuracy closer to high fidelity simulations.





# MODULUS USE CASES

# 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

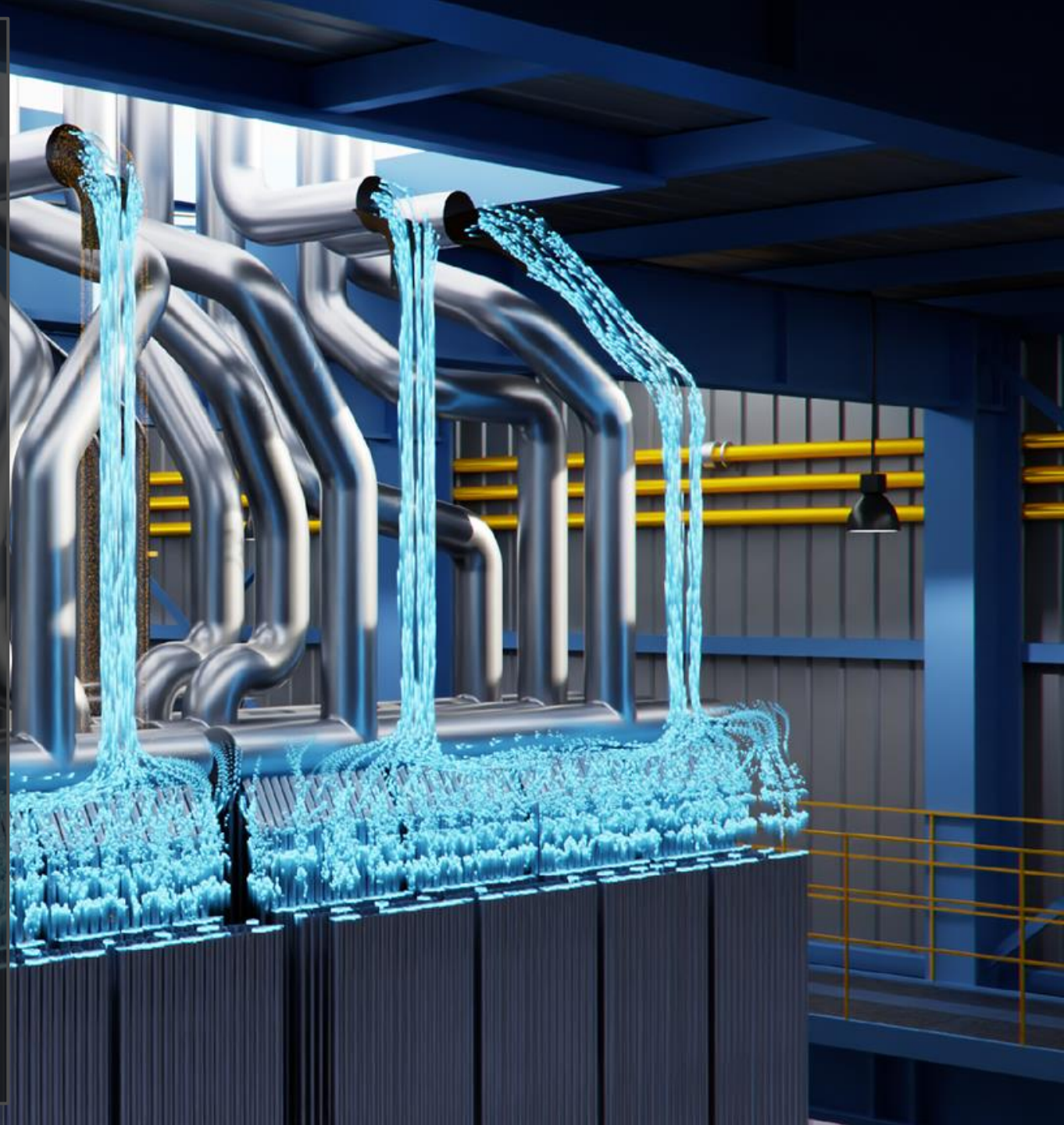
- Hardware: NVIDIA V100 & A100 Tensor Core GPUs
- Software: NVIDIA Modulus, NVIDIA Omniverse

## 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



# DIGITAL TWINS FOR POWER PLANT BOILERS — BATTELLE, NETL

## Use Case

- Using digital twins to accelerate the design and development cycle of a power plant boiler and enable effective carbon capture and storage

## Challenges

- During the power plant development process, many techniques are used to design robust carbon management
  - This requires complex simulations of fluid flow mechanics, heat transfer, and chemical reactions.

## Solution

- Using NVIDIA's Physics Informed Neural Network (PINNs) and Modulus, National Energy Technology Lab developed a digital twin of a boiler capable of modeling turbulent reacting flows.

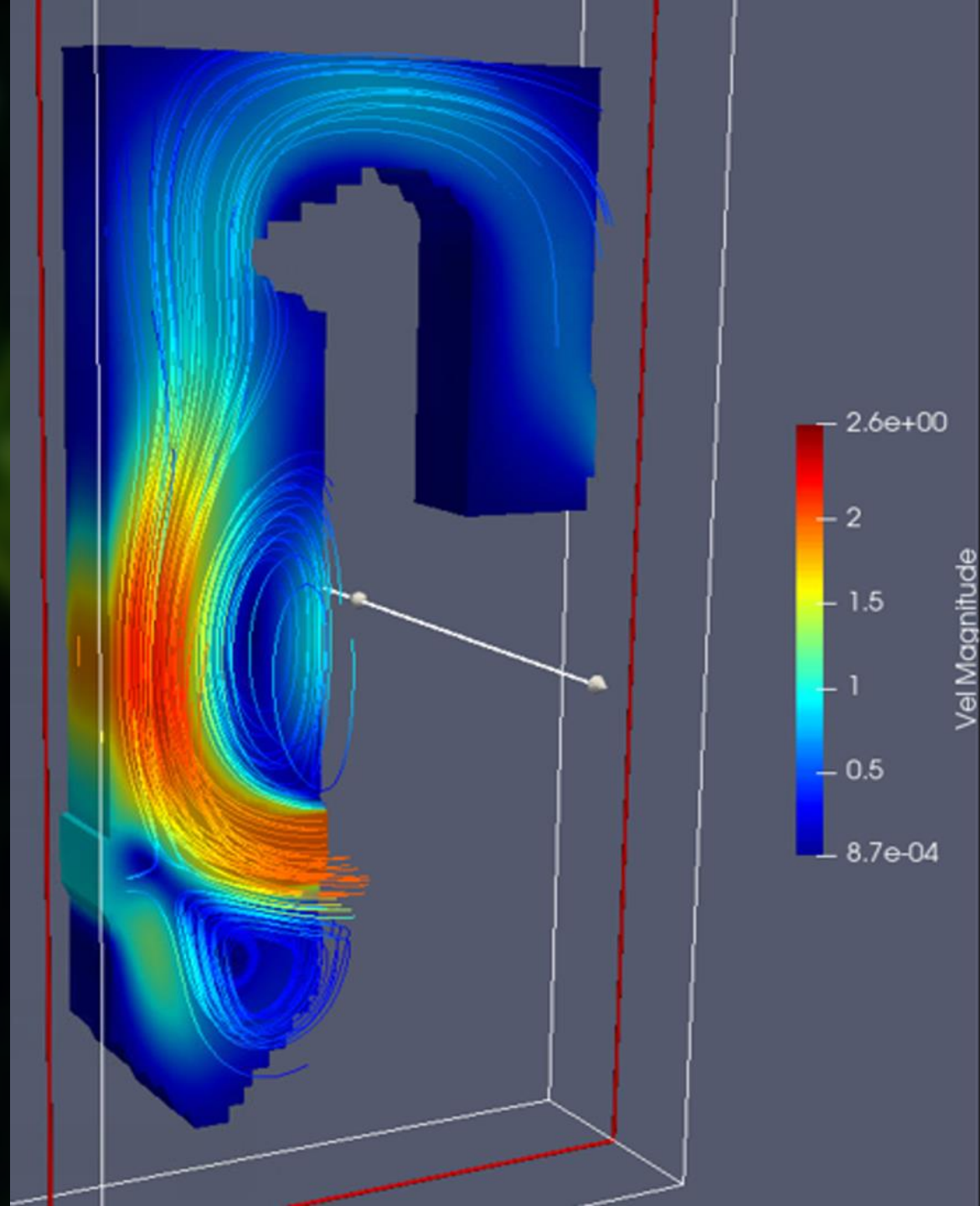
## NVIDIA Solution Stack

- Hardware: NVIDIA V100, A100 GPUs
- Software: CUDA 10.2, NVIDIA Modulus

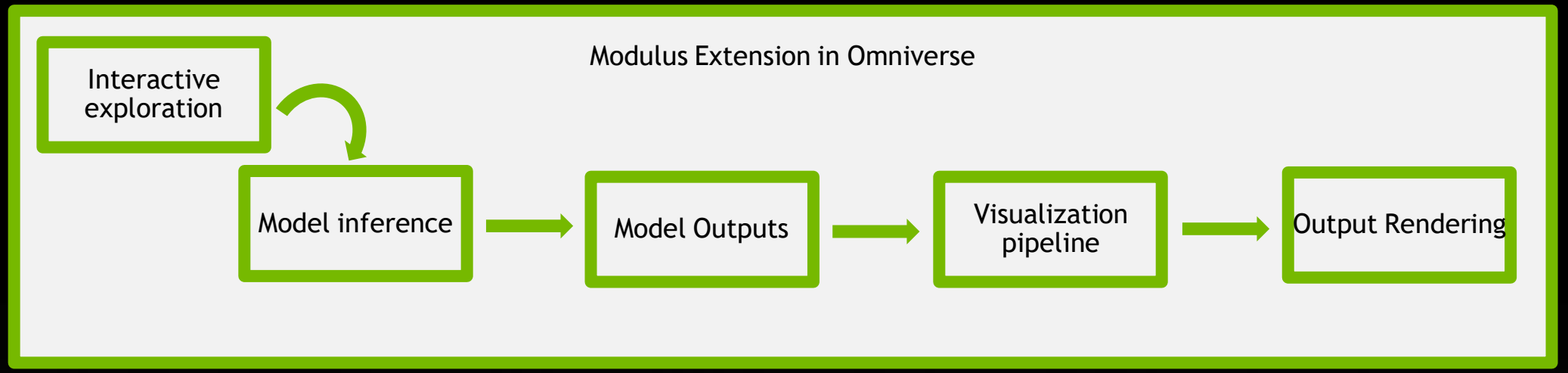
## Outcome

- NETL accelerated the design and development cycle of a powerplant by fast and highly accurate predictions.

NVIDIA BATTELLE NETL NATIONAL ENERGY TECHNOLOGY LABORATORY

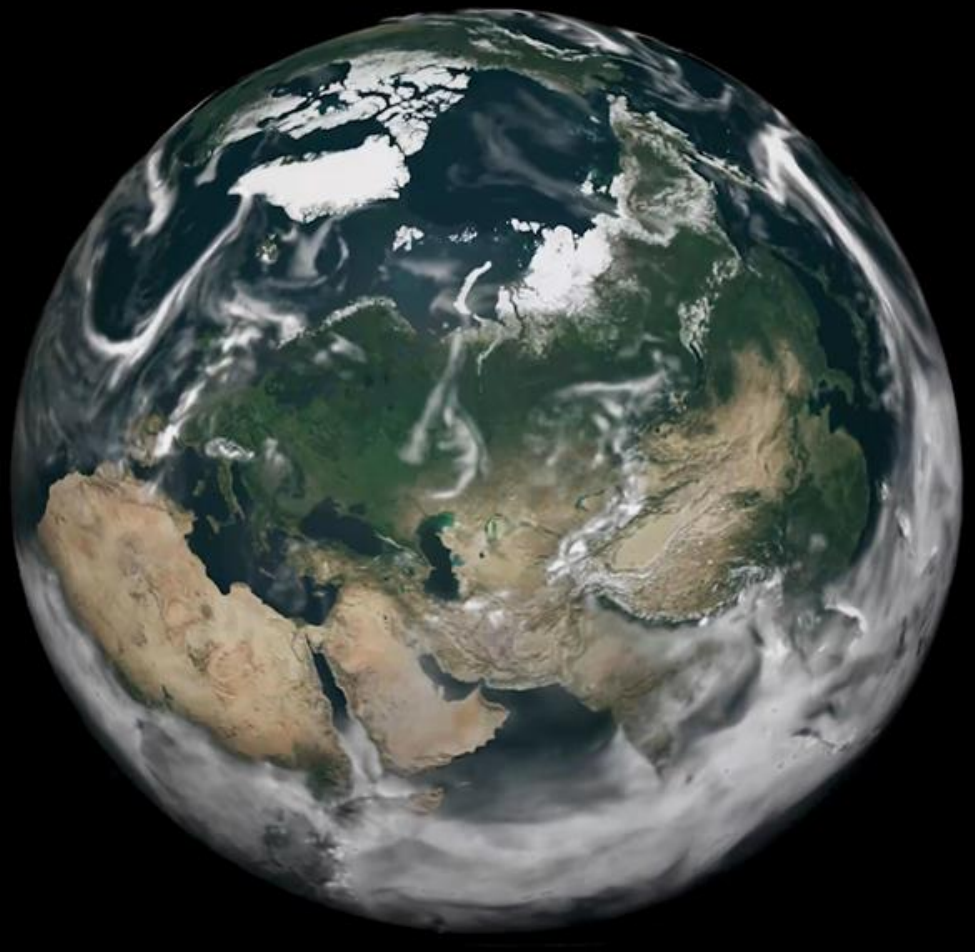


# MODULUS – OMNIVERSE INTEGRATION




- Modulus Omniverse extension:
  - enables importing outputs of Modulus trained model into a visualization pipeline for common output scenarios ex: streamlines, iso-surface
  - provides an interface that enables interactive exploration of design variables/parameters to infer new system behavior


EARTH -2
DIGITAL TWINS



# ACCELERATING EXTREME WEATHER PREDICTION WITH FourCastNet IN NVIDIA MODULUS

## Use Case

- Climate change is making storms both stronger and less predictable, leading to more fires, floods, heatwaves, mudslides, and droughts.
- Predicting global weather patterns and extreme weather events, like atmospheric rivers, is important to quantify any catastrophic event with confidence.

## Challenges

- To develop the best strategies for mitigation and adaptation, we need climate models that can predict the climate in different regions of the globe over decades.

## Solution

- NVIDIA has created a physics-ML model that emulates the dynamics of global weather patterns and predicts extreme weather events, like atmospheric rivers, with unprecedented speed and accuracy.

## NVIDIA Solution Stack

- Hardware: NVIDIA A100
- Software: NVIDIA Omniverse, NVIDIA Modulus

## Outcome

- Powered by the Fourier Neural Operator, this GPU-accelerated AI-enabled digital twin, called FourCastNet, is trained on 10 TB of Earth system data.
- Using this data, together with NVIDIA Modulus and Omniverse, we are able to forecast the precise path of catastrophic atmospheric rivers a full week in advance.

Demo

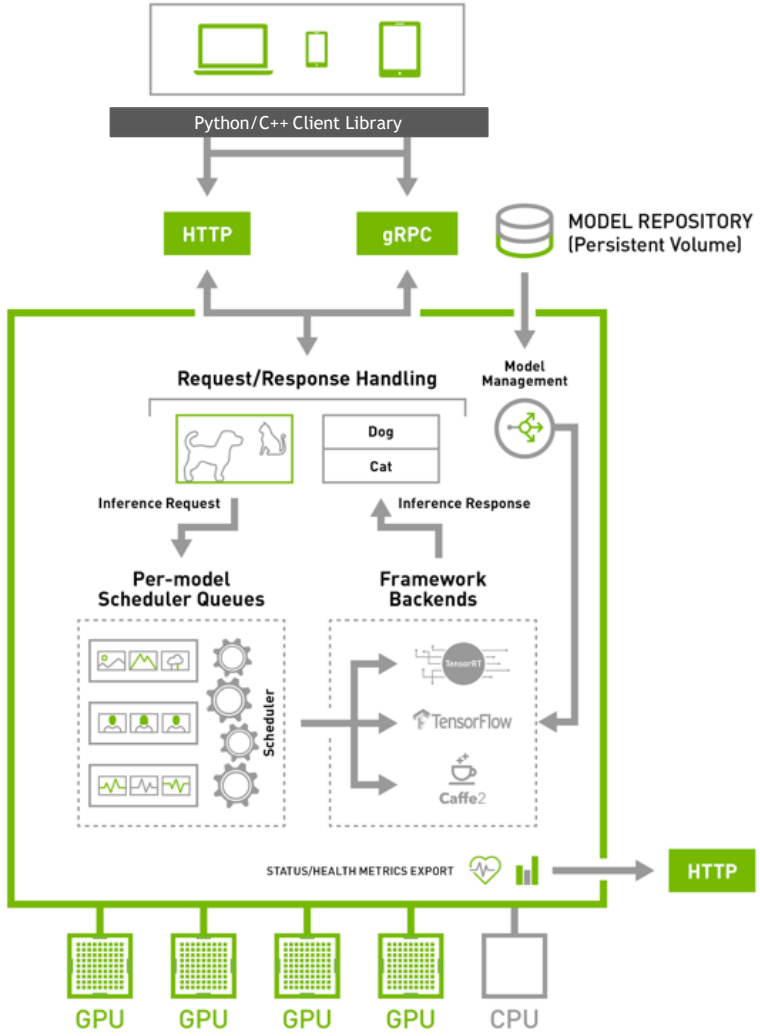
NVIDIA

# Deployment

# TRITON INFERENCE SERVER ARCHITECTURE

## Available with Monthly Updates

NVIDIA Triton Inference Server




Models supported

- TensorFlow GraphDef/SavedModel
- TensorFlow and TensorRT GraphDef
- TensorRT Plans
- Caffe2 NetDef (ONNX import)
- ONNX graph
- PyTorch JIT (.pb)

Multi-GPU support

Concurrent model execution

Server HTTP REST API/gRPC

Python/C++ client libraries

# Jetson @ Edge

# THE JETSON FAMILY

## for AI at the Edge and Autonomous System designs

**JETSON NANO**
0.5 TFLOPS (FP16)

5 - 10W
45mm x 70mm

**JETSON TX2 series**
1.3 TFLOPS (FP16)

7.5 – 15W*
50mm x 87mm

**JETSON Xavier NX**
6 TFLOPS (FP16)
21 TOPS (INT8)

10 - 15W
45mm x 70mm

**JETSON AGX XAVIER series**
11 TFLOPS (FP16)
32 TOPS (INT8)

10 – 30W
100mm x 87mm

AI at the edge

Fully autonomous machines

Same software

* TX2i: 10-20W

# JETSON DEVELOPER KITS

## For Developers, Engineers and Makers

JETSON NANO
5W | 10W
0.5 TFLOPS (FP16)

JETSON TX2
7.5W | 15W
1.3 TFLOPS (FP16)

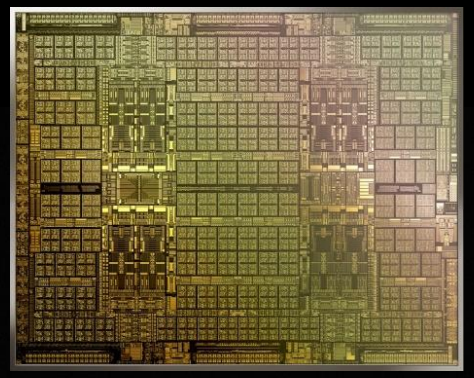
JETSON XAVIER NX
10W | 15W
7 TFLOPS (FP16) | 21 TOPS (INT8)

JETSON AGX XAVIER
10 | 15W | 30W
11 TFLOPS (FP16) | 32 TOPS (INT8)

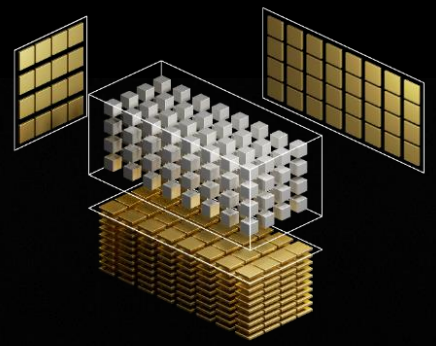
**Multiple developer kits - Same software**

Full specs at developer.nvidia.com/jetson

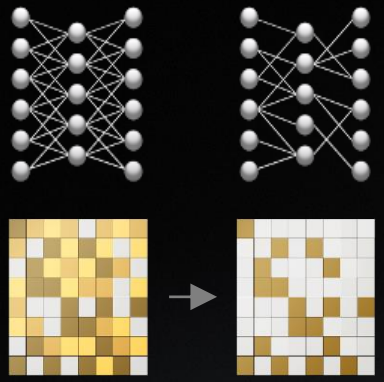
# 5 KEY FEATURES OF A100



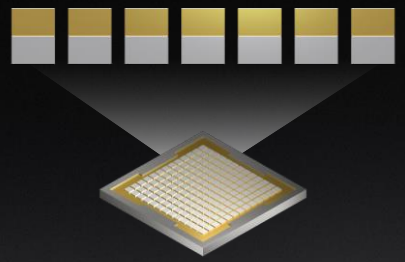
NVIDIA Ampere Architecture
World's Largest 7nm chip
54B XTORS, HBM2



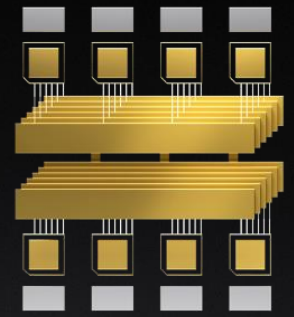
3rd Gen Tensor Cores
Faster, Flexible, Easier to use
20x AI Perf with TF32
2.5x HPC Perf



New Sparsity Acceleration
Harness Sparsity in AI Models
2x AI Performance



New Multi-Instance GPU
Optimal utilization with right sized GPU
7x Simultaneous Instances per GPU



3rd Gen NVLINK and NVSWITCH
Efficient Scaling to Enable Super GPU
2X More Bandwidth


Thank
You
NVIDIA