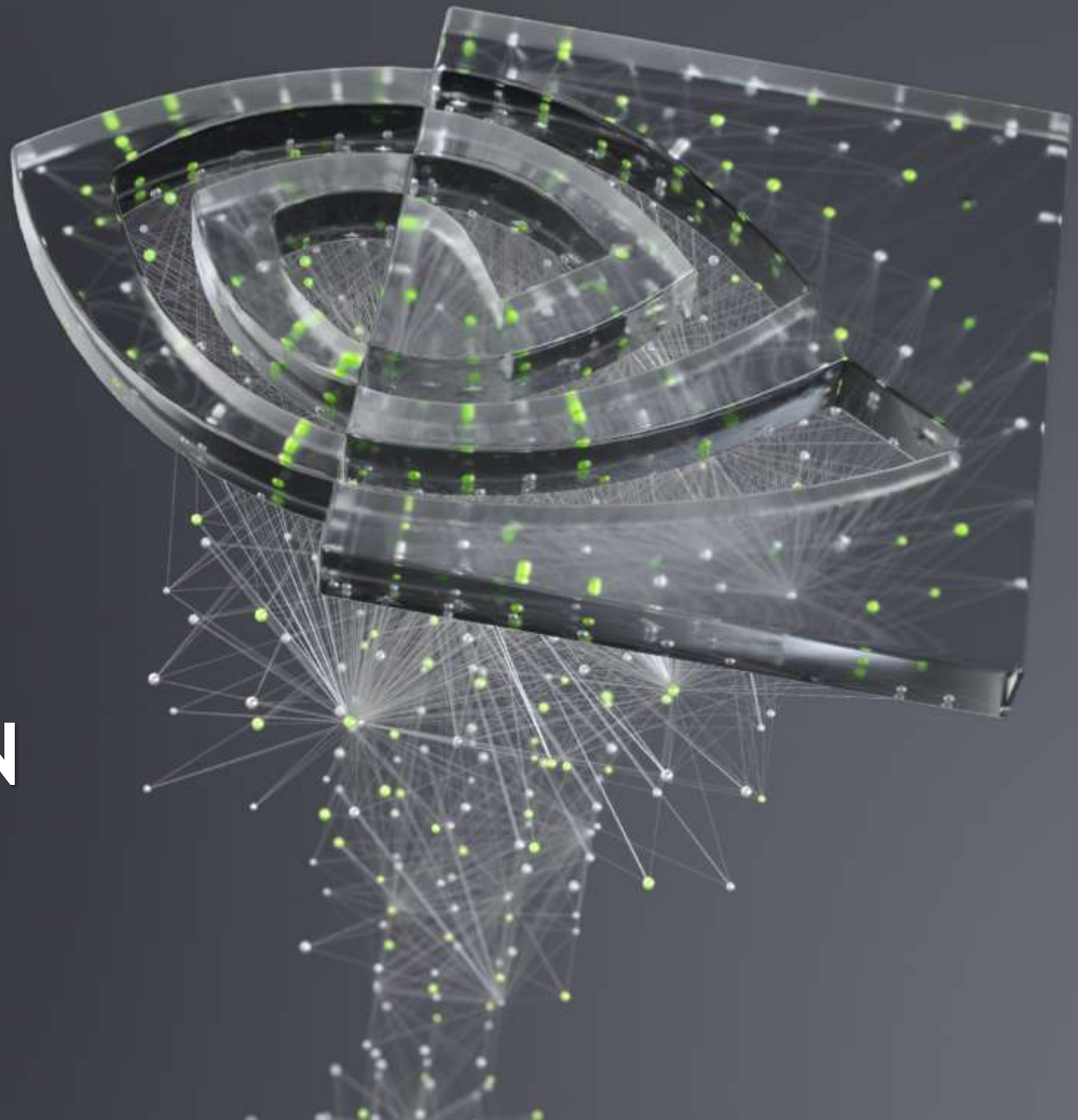
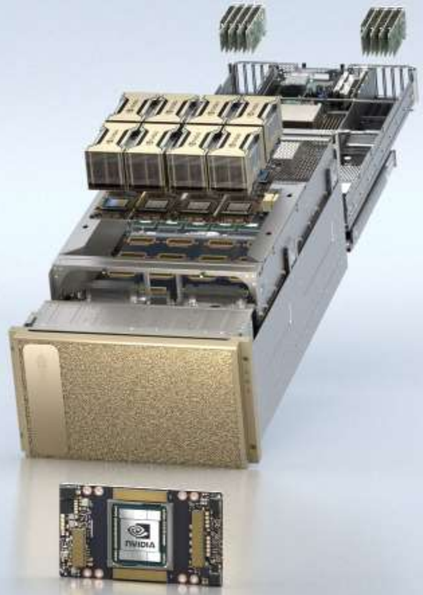


# NVIDIA APPLICATION FRAMEWORKS

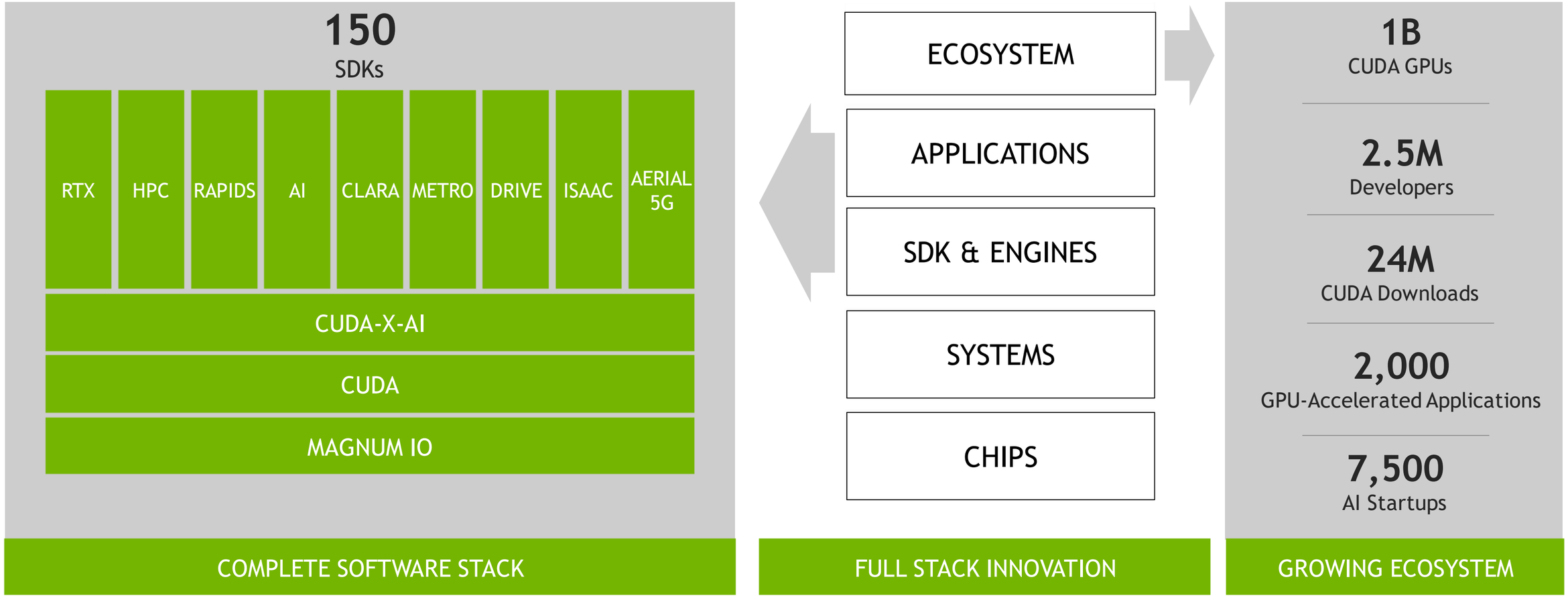
Pedro Mário Cruz e Silva
Solutions Architect Manager
Enterprise Latin America
NVAITC-Latam (NVIDIA AI Technology Center)

# SOLVING PROBLEMS ORDINARY COMPUTERS CANNOT SOLVE

# NVIDIA IS A FULL STACK COMPUTING PLATFORM

## Amazing Innovation and Expansion of NVIDIA Ecosystem



# NVIDIA DATACENTER PLATFORM

**BUSINESS APPLICATIONS**

Customer Engagement | Patient Diagnostics | Fraud Detection | Quality Assurance | Industrial Automation | Precision Marketing | Molecular Simulations | ++

**NGC**

**SOFTWARE HUB**

HELM

*Certified Containers*

*Pre-trained Models*

*SDKs*

**APPLICATION FRAMEWORKS**

| SMART CITY | CONVERSATIONAL AI | AUTONOMOUS VEHICLES | RECOMMENDATION SYSTEMS | HEALTHCARE | ++ |
|---|---|---|---|---|---|
| *Metropolis* | *Jarvis* | *Drive* | *Merlin* | *Clara* | ... |

**DEVELOPER TOOLKITS**

| ML & DATA ANALYTICS | AI TRAINING & INFERENCE | HIGH PERFORMANCE COMPUTING | RENDERING & VISUALIZATION |
|---|---|---|---|
| RAPIDS, dmlc XGBoost | TensorFlow, TensorRT, PYTORCH, mxnet | *NVIDIA HPC SDK* | *IndeX, OptiX, CloudXR, MDL* |

**ACCELERATION LIBRARIES**

| COMPUTE | NETWORKING, STORAGE & SECURITY |
|---|---|
| *CUDA-X* | *DOCA, MAGNUM IO* |

**OPERATIONS**

VIRTUAL GPU SW

TRITON INFERENCE SERVER

FLEET COMMAND

NVIDIA GPU Operator

Red Hat

vmware

**NVIDIA CERTIFIED**

**VALIDATED SOLUTIONS**

**SERVERS & CLOUD**

DGX HGX

Purpose Built

EGX

Mainstream & Edge

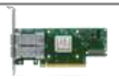
CSP Instances

**HARDWARE TECHNOLOGIES**

GPU

NVSwitch

BF DPU



SMART NIC

NVIDIA Switch

**MONITORING**

DCGM

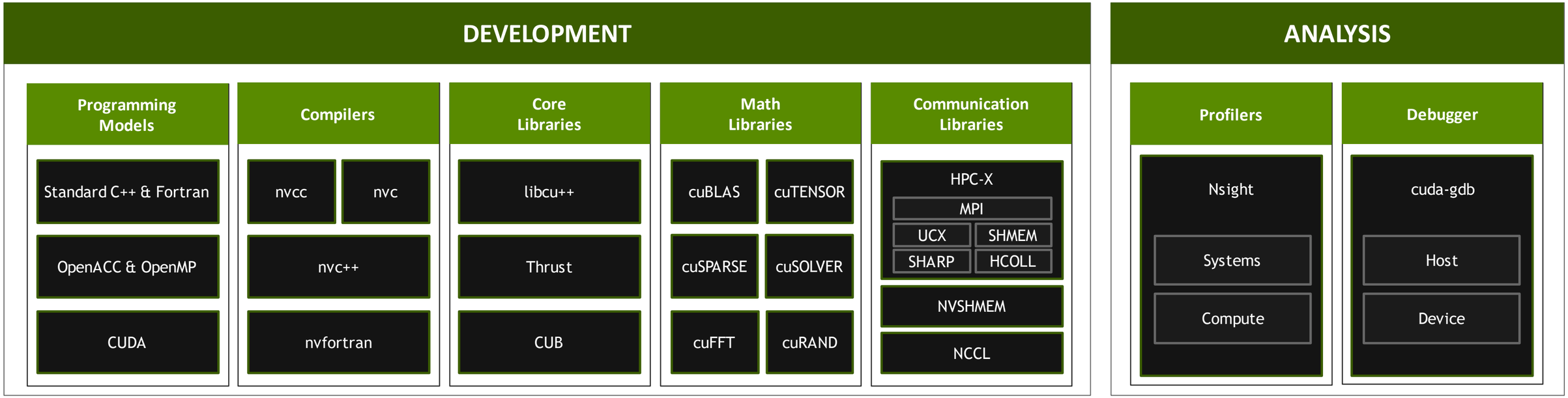
Grafana

Prometheus

UFM

# NVIDIA HPC SDK

## Fully Integrated Developer Environment for the NVIDIA Platform: GPU, CPU and Interconnect



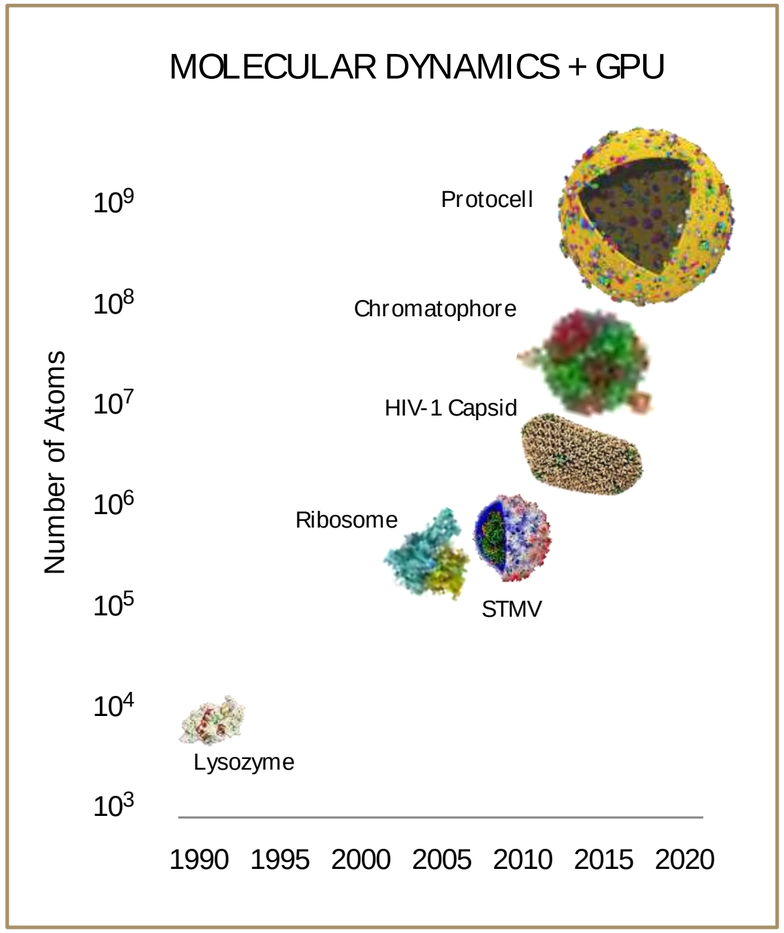

Available at developer.nvidia.com/hpc-sdk, on NGC, via Spack, and in the Cloud

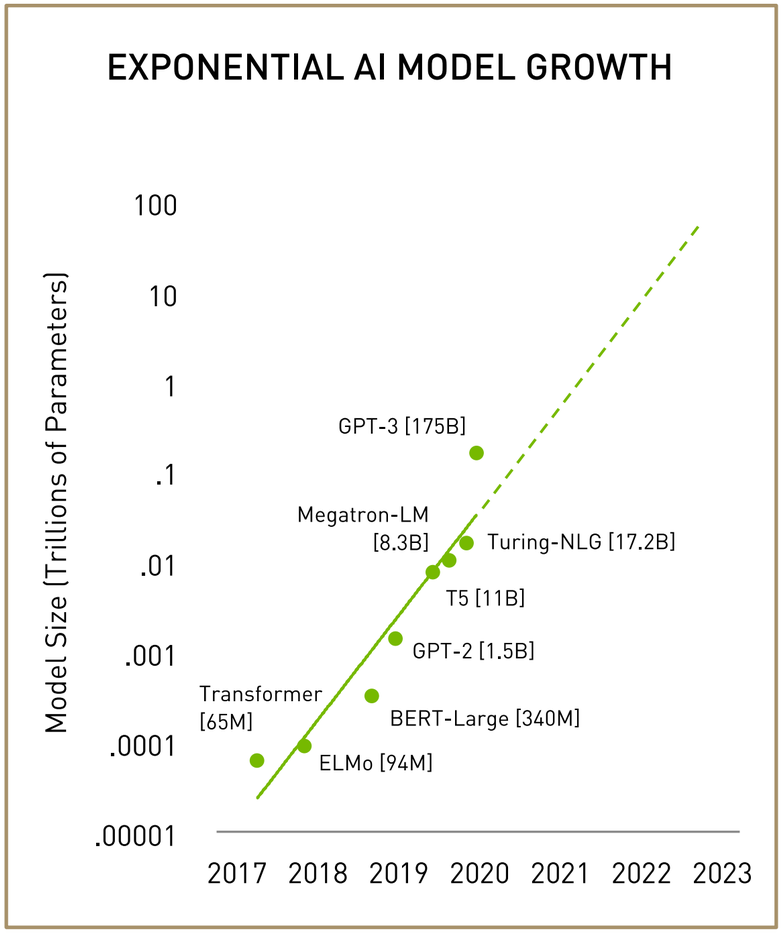
Arm | x86 | Power

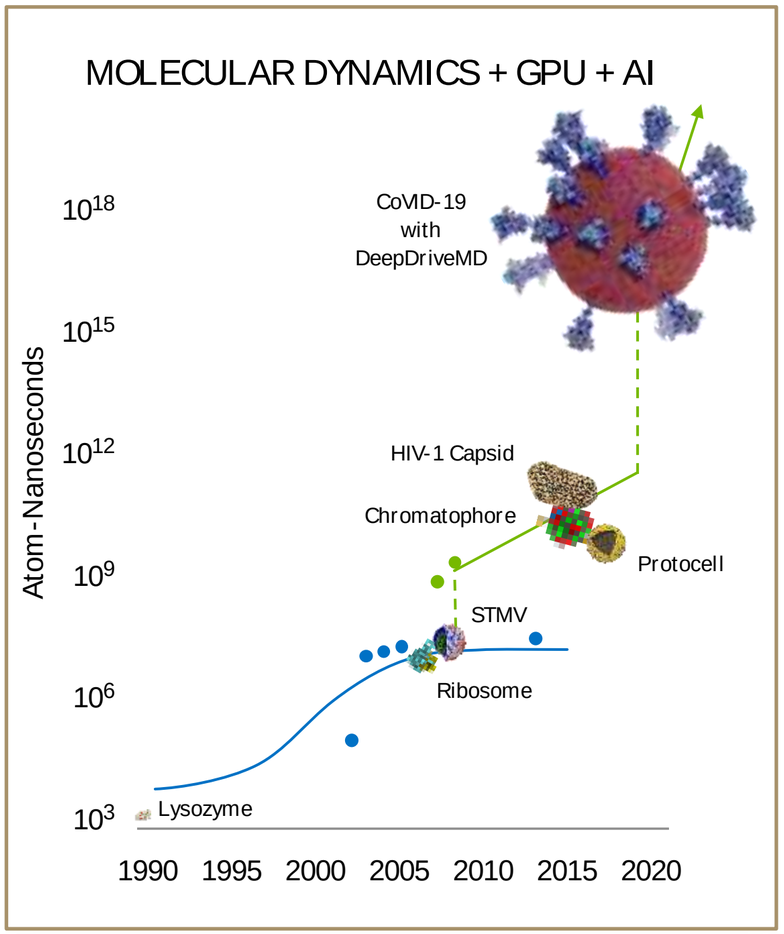
Libraries | Accelerated C++ and Fortran | Directives | CUDA

7-8 Releases Per Year | Freely Available

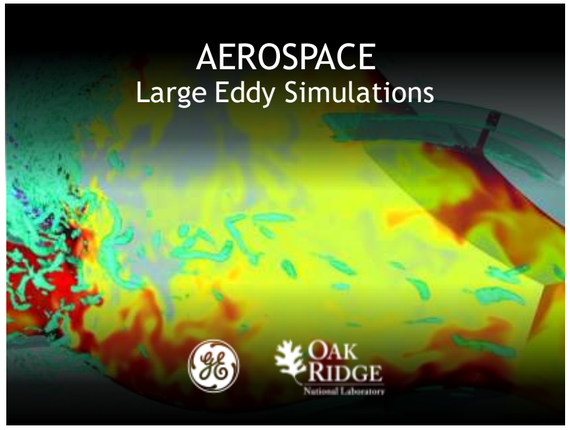
# INDUSTRIAL HPC AT THE TIPPING POINT

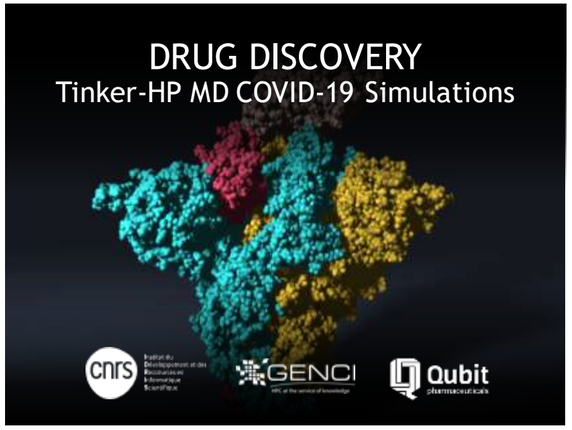


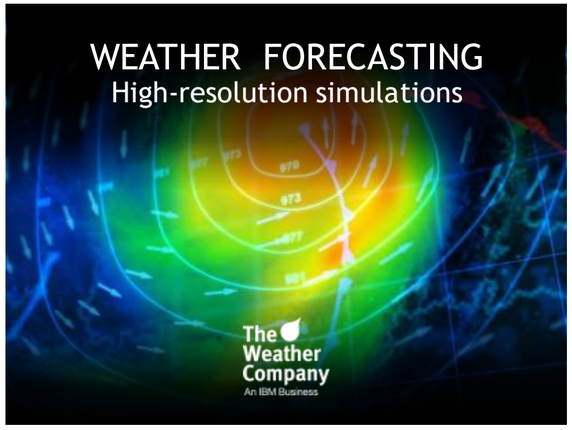
# INDUSTRIAL HPC DELIVERS TRANFORMATIONAL BREAKTHROUGHS







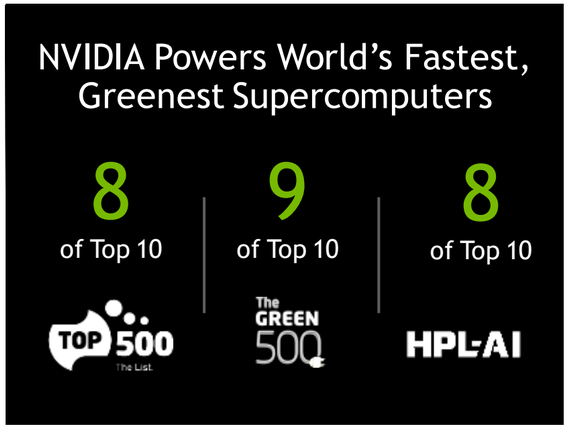




# POWERING NEXT WAVE OF ACCELERATED AI SUPERCOMPUTING







# NVIDIA POWER WORLD'S FASTEST AND MOST EFFICIENT, AI SUPERCOMPUTERS

TOP 500 The List.

**8**
of Top 10

**68%**
Overall

**20%**
More InfiniBand Systems v. ISC20

The GREEN 500

**9**
of Top 10

**3.5X**
Higher Energy Efficiency v. non-GPU Green500 Systems

**29.5GF/W**
Greenest NVIDIA System

HPL-AI

**8**
of Top 10

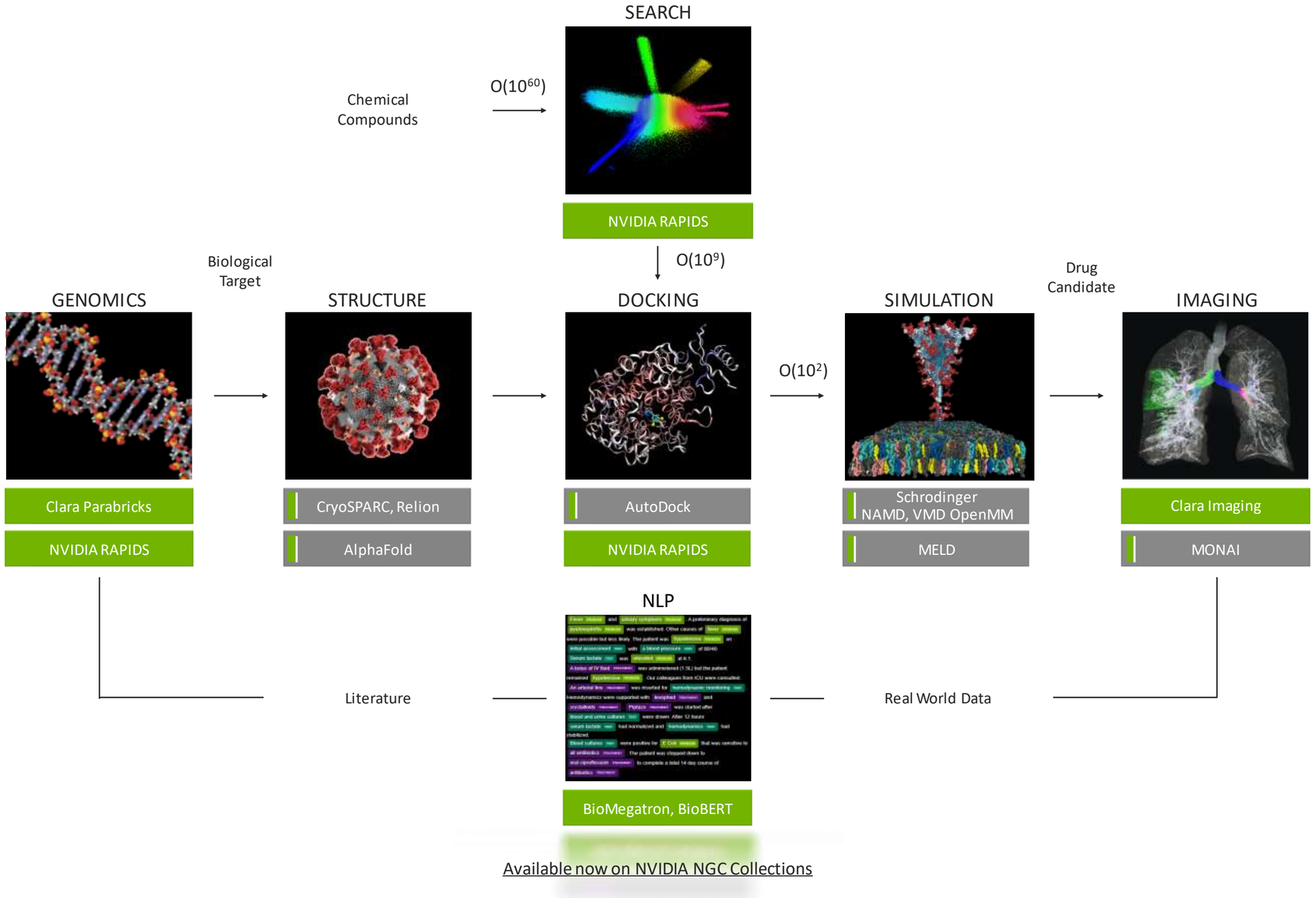
**1.1EF**
Mixed Precision AI Performance on SUMMIT

**2X**
Submission Over Previous List

NVIDIA

# CLARA HEALTHCARE

# NVIDIA CLARA DISCOVERY




Available now on NVIDIA NGC Collections

# BUILDING A MODEL FOR THE INDUSTRY


Partnering on World's First AI Pharma Lab

Unlock the power of Biomedical Data
Digital Pathology, Imaging, Genomics


Building UK's Fastest Supercomputer for AI Healthcare Research

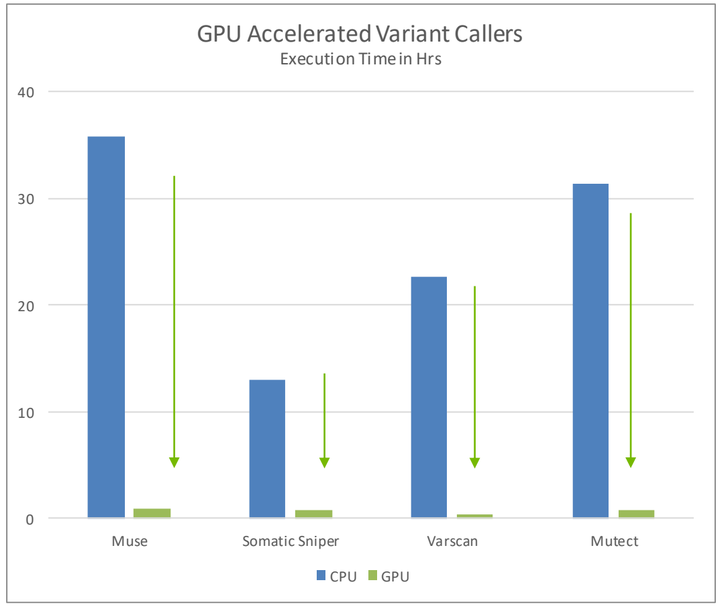
Top 30 on Top 500 | Top 3 on Green 500
80 DGX A100 | NGC Optimized Software

# NVIDIA CLARA PARABRICKS

## Improving Accuracy, Speed and Scale of Computational Genomics

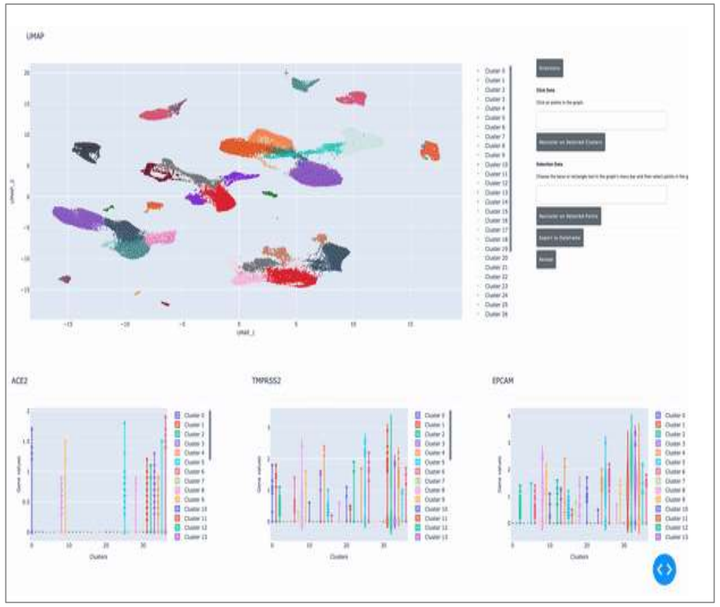

### Accelerated Best Practices

Industry Standard GATK Pipelines
Germline, Somatic, RNA-Seq
30x Faster WGS, 12x Faster WES




### Detecting Diversity of Variants

DeepVariant
Multiple Somatic Callers, Days -> 1Hr




### Analytics at Scale

GWAS & Single Cell
NVIDIA RAPIDS GPU Data Science

*Clara Parabricks NGC Collection*

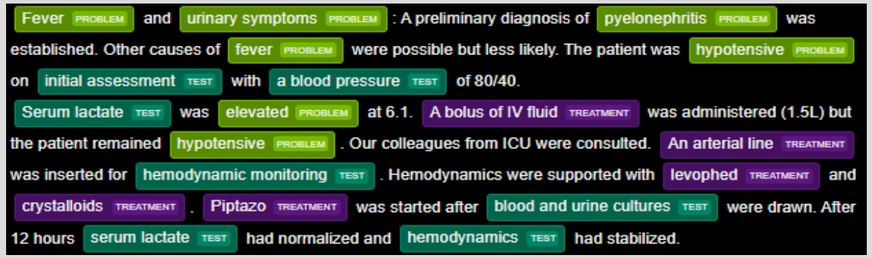


NVIDIA

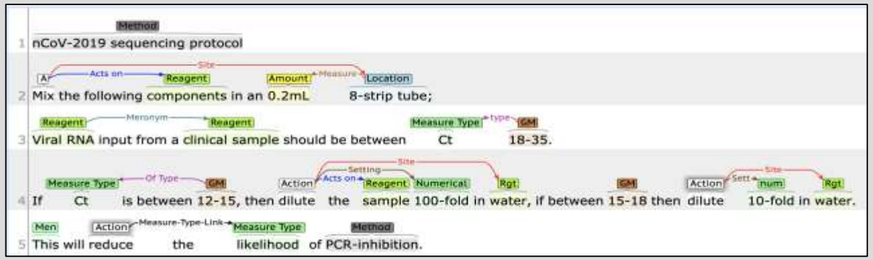
# NEW BREAKTHROUGH IN BIOMEDICAL NLP

## NVIDIA BioMegatron Achieves State-of-the-Art

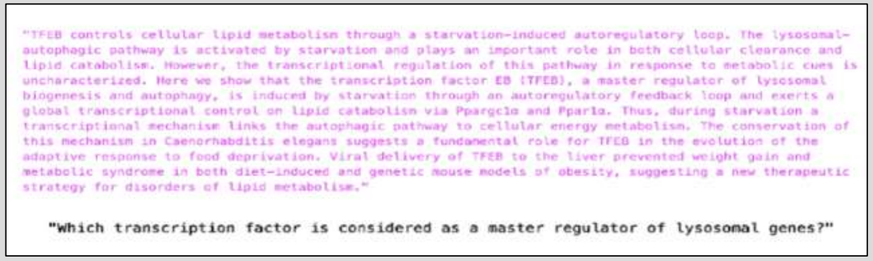
Named Entity Recognition




Relation Extraction




Question Answering




BioMegatron: Larger Biomedical Domain Language Model

| | Benchmark | Model | #Parameters | Vocabulary | Prec | Rec | F1 |
|---|---|---|---|---|---|---|---|
| NER | BC5CDR-chem | BioBERT | 110m | BERT-cased | 90.0 | 93.4 | 91.7 |
| | | PubMedBERT | 110m | PubMedBERT-vocab (30k) | 92.1 | 93.2 | 92.6 |
| | | **BioMegatron** | **345m** | **Bio-vocab-30k** | 92.1 | **93.6** | **92.9** |
| | | BioMegatron | 345m | Bio-vocab-50k | **92.9** | 92.0 | 92.5 |
| | | BioMegatron | 800m | BERT-cased | 91.3 | 92.9 | 92.1 |
| | | BioMegatron | 1.2b | BERT-uncased | 92.0 | 90.5 | 91.3 |
| | BC5CDR-disease | BioBERT | 110m | BERT-cased | 85.0 | 89.4 | 87.2 |
| | | PubMedBERT | 110m | PubMedBERT-uncased (30k) | **86.2** | 88.4 | 87.3 |
| | | BioMegatron | 345m | Bio-vocab-30k | 85.2 | 88.8 | 87.0 |
| | | **BioMegatron** | **345m** | **Bio-vocab-50k** | 86.1 | **91.0** | **88.5** |
| | | BioMegatron | 800m | BERT-cased | 85.8 | 90.1 | 87.9 |
| | | BioMegatron | 1.2b | BERT-uncased | 83.8 | 89.2 | 86.4 |
| | NCBI-disease | BioBERT | 110m | BERT-cased | 85.0 | 90.0 | 87.5 |
| | | PubMedBERT | 110m | PubMedBERT-uncased (30k) | 85.9 | 87.7 | 86.8 |
| | | BioMegatron | 345m | Bio-vocab-30k | 85.6 | 88.6 | 87.1 |
| | | BioMegatron | 345m | Bio-vocab-50k | 83.7 | **90.4** | 87.0 |
| | | **BioMegatron** | **800m** | **BERT-cased** | **87.0** | 88.8 | **87.8** |
| | | BioMegatron | 1.2b | BERT-uncased | 83.5 | 90.1 | 86.7 |

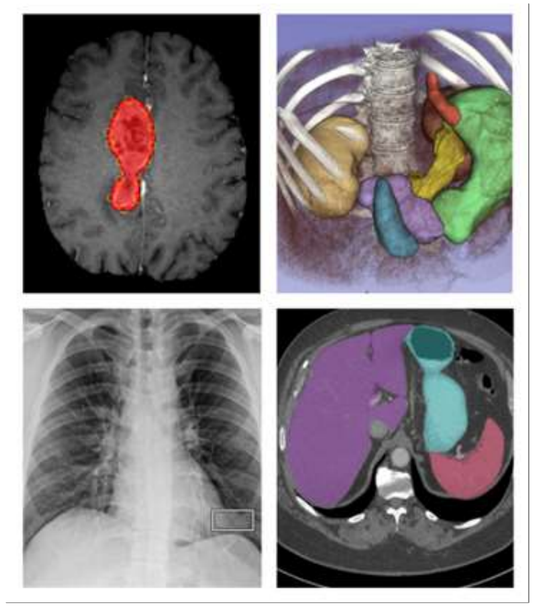
| | Benchmark | Model | #Parameters | Vocabulary | Prec | Rec | F1 |
|---|---|---|---|---|---|---|---|
| RE | ChemProt | BioBERT | 110m | BERT-cased | 76.5 | 73.3 | 74.8 |
| | | PubMedBERT | 110m | PubMedBERT-uncased (30k) | 73.6 | 77.7 | 75.6 |
| | | BioMegatron | 345m | Bio-vocab-30k | 77.8 | 72.5 | 75.1 |
| | | **BioMegatron** | **345m** | **Bio-vocab-50k** | 74.5 | **79.7** | **77.0** |
| | | BioMegatron | 800m | BERT-cased | 80.4 | 68.9 | 74.3 |
| | | BioMegatron | 1.2b | BERT-uncased | **82.0** | 65.6 | 72.9 |

| | Benchmark | Model | #Parameters | Vocabulary | SAcc | LAcc | MRR |
|---|---|---|---|---|---|---|---|
| QA | BioASQ-7b-factoid | BioBERT-Base | 110m | BERT-cased | 30.8 | 64.1 | 41.1 |
| | | BioBERT-Large | 345m | BERT-cased | 42.8 | 62.8 | 50.1 |
| | | **BioMegatron** | **345m** | **BERT-uncased** | 46.2 | **62.6** | **52.5** |
| | | BioMegatron | 800m | BERT-uncased | 45.2 | 58.6 | 50.4 |
| | | BioMegatron | 1.2b | BERT-uncased | **47.4** | 60.9 | 52.4 |

*Source: https://arxiv.org/abs/2010.06060*

# NVIDIA CLARA IMAGING AI FRAMEWORK

## Discovering Biomarkers in Imaging Data

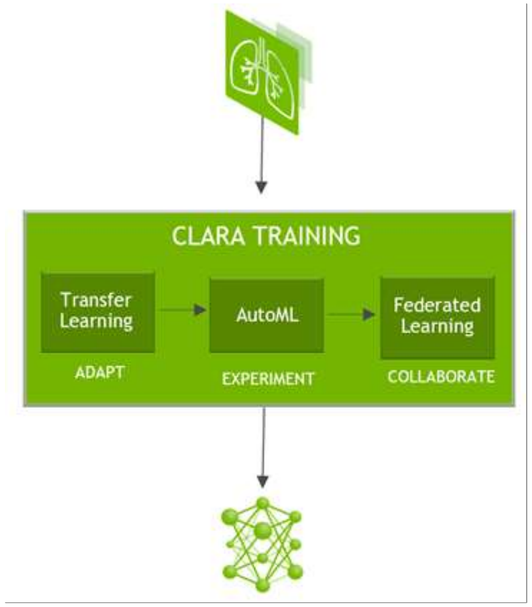


### Pre-Trained Models

Multi-Organ, Multi-Modal
2D & 3D Classification
3D Segmentation

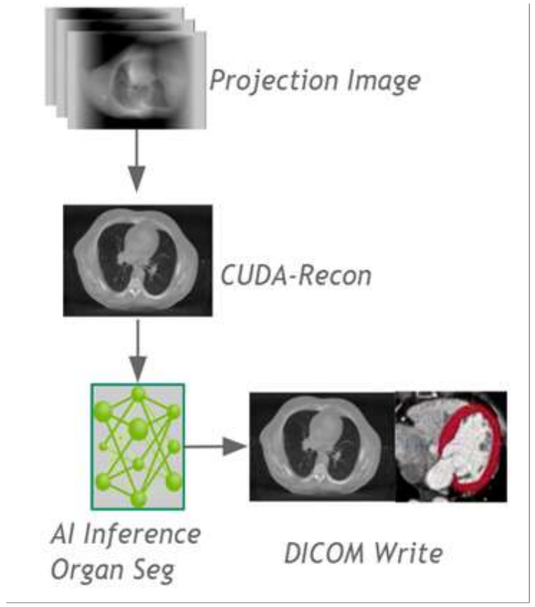
### AI Assisted Annotation

3D Slicer, OHIF, MITK, Fovia
Re-Train Model w/ User Input
10x Faster Annotation




### Optimized Training Frameworks

Multi-GPU | Multi-Node
Mixed Precision | Fast I/O
55x Faster vs Native OSS




### Scalable Deployment Framework

Multi-Stage Compute Pipelines
Multi-Stage AI Pipelines
Health IT Integrations

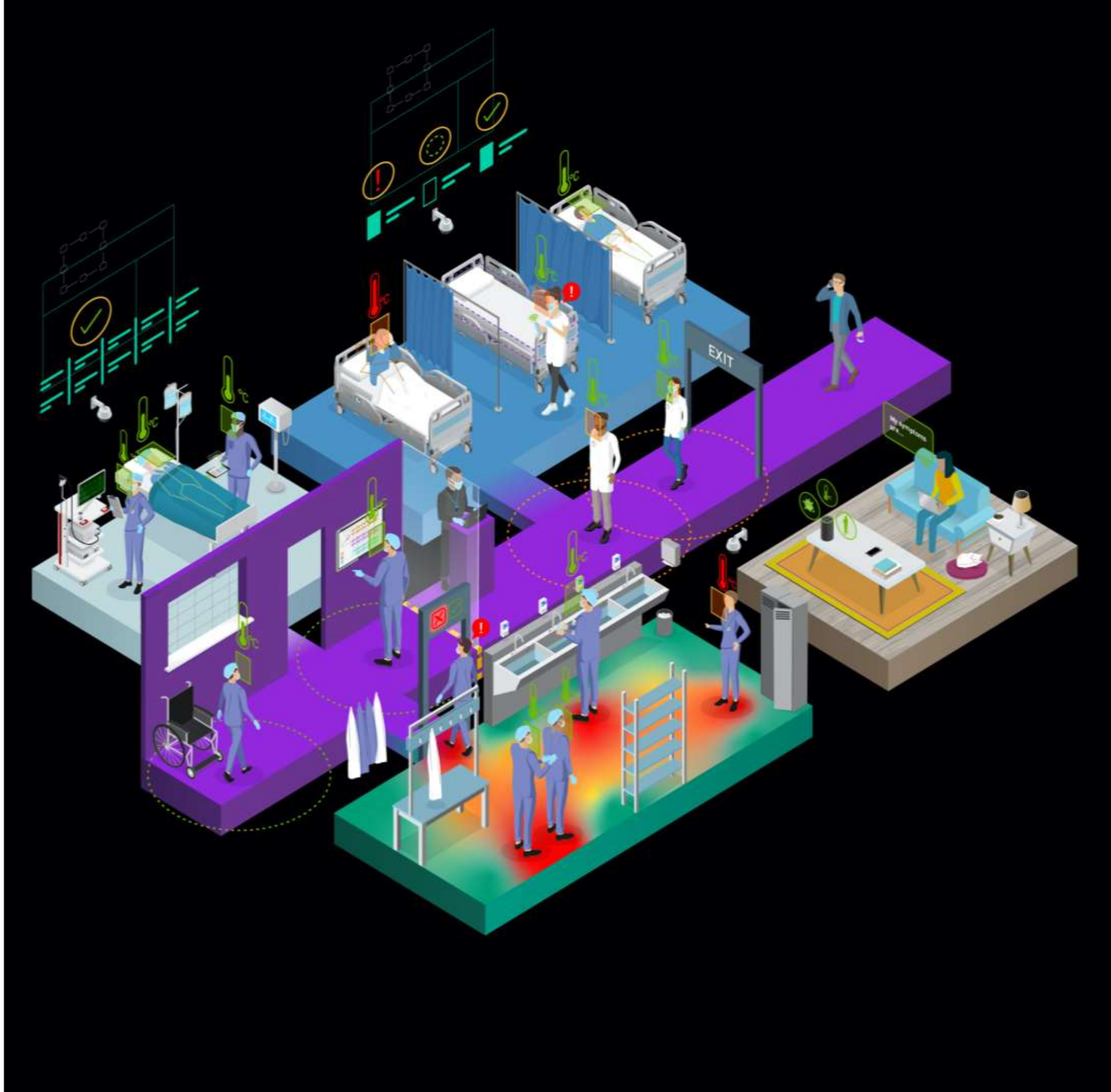
# SPEEDING THE WAY TO SMART HOSPITALS

The delivery of healthcare is increasingly more challenging with having to do more with less resources.

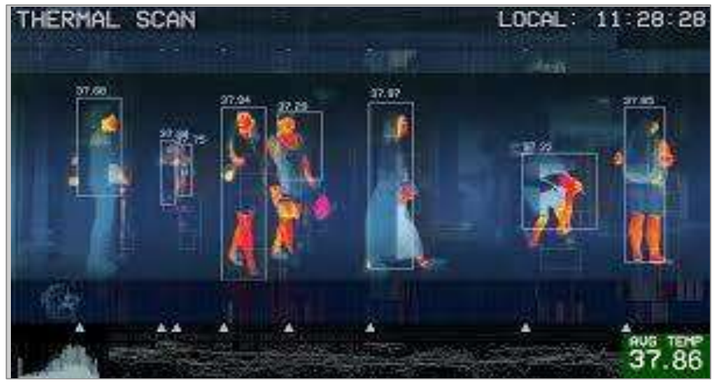
Smart sensors such as cameras and speakers can act as eyes and ears to ensure safety and operational excellence of healthcare facilities as well as advance care for patients.

It's now possible with smart sensors to measure for fever or absence of protective gear, monitor crowds and safe social distancing, interact with high risk patients keeping both patients and staff safe and informed.

NVIDIA Clara Guardian is an application framework and partner ecosystem accelerating the development and deployment of smart sensors and sensor fusion anywhere in the hospital or health system.



# SMART HOSPITAL USE CASES




Body Temperature Screening

Patient Monitoring

Surgery Analytics

Safe Social Distancing

Fall Prevention

Contactless Controls

# OMNIVERSE

# REVOLUTIONIZE COLLABORATION

## For Widespread Teams

### UNITES TEAMS, TOOLS, AND ASSETS

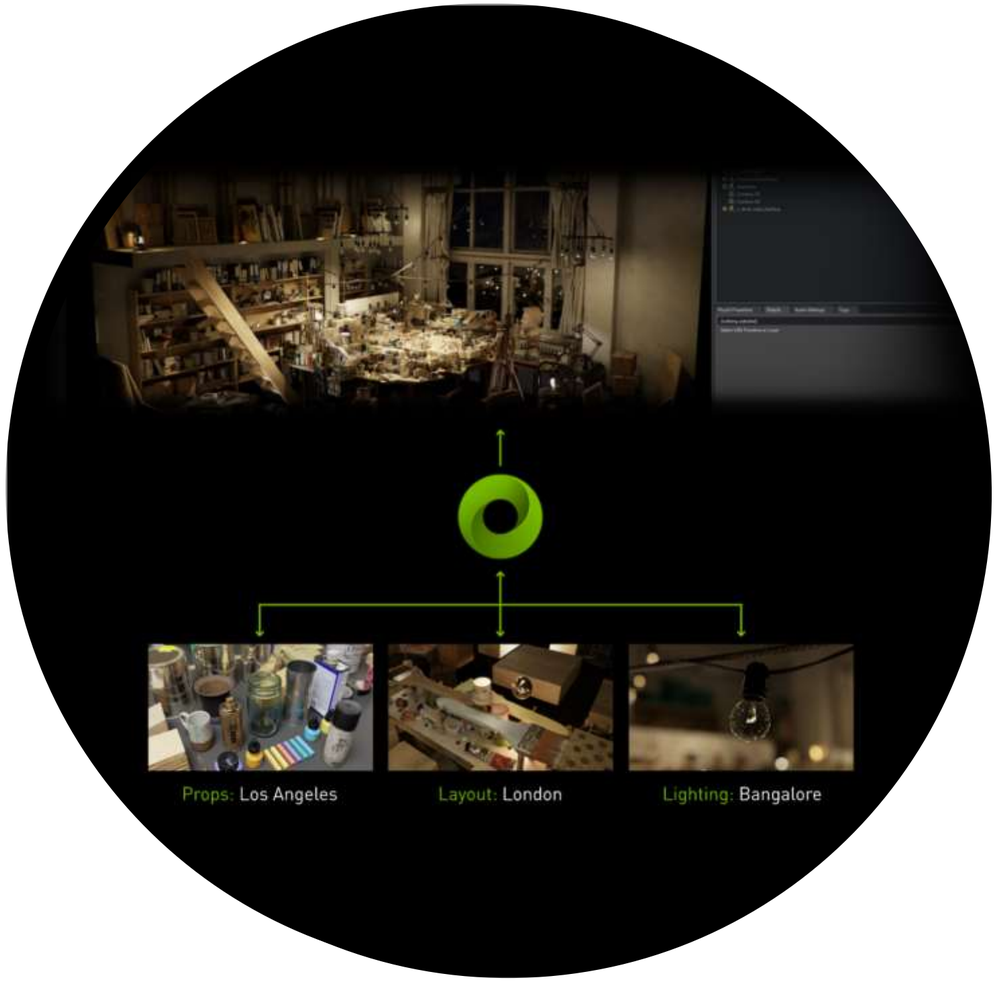
Workflows are simplified as updates, iterations, and edits are instantaneous with no need for data preparation.

### SIMPLIFIES IT CHALLENGES

Easily scale to any organizational level, integrate with any IT infrastructure, and support the building of custom applications, extensions, and experiences.

### DEMOCRATIZES PHOTOREAL RENDERING

Beautiful, high-fidelity models can be instantly shared to any device with one click to ray-traced or path-traced rendering. Omniverse delivers scalable, final-frame-quality rendering in real time.



# ADVANCED TOOLS AND TECHNOLOGIES

## Foundational Platform Components

NUCLEUS

CONNECT

KIT

SIMULATION

RTX RENDERER

NVIDIA OMNIVERSE

# CUTTING EDGE APPLICATIONS

## Core Omniverse Apps

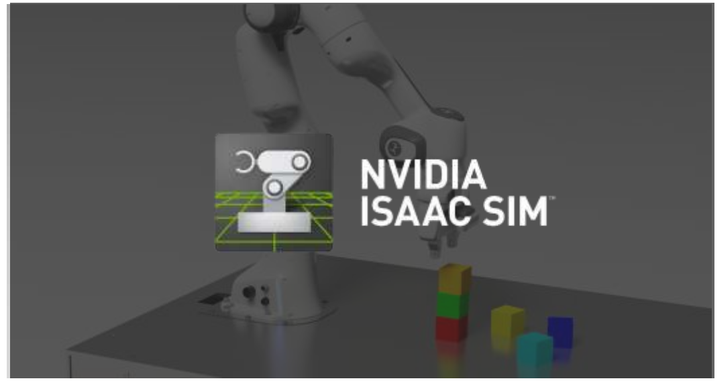

FOR GAME DEVELOPERS, ANIMATORS


FOR DESIGNERS, CREATORS, ENGINEERS


FOR 3D DEEP LEARNING RESEARCHERS


FOR GEFORCE RTX GAMERS


FOR ARCHITECTS, DESIGNERS, ENGINEERS




FOR ROBOTICISTS, SIMULATION SPECIALISTS

# POWER LARGE SCALE SIMULATION

## Physically accurate, complex virtual worlds

**Anything that is Built will be Visualized**

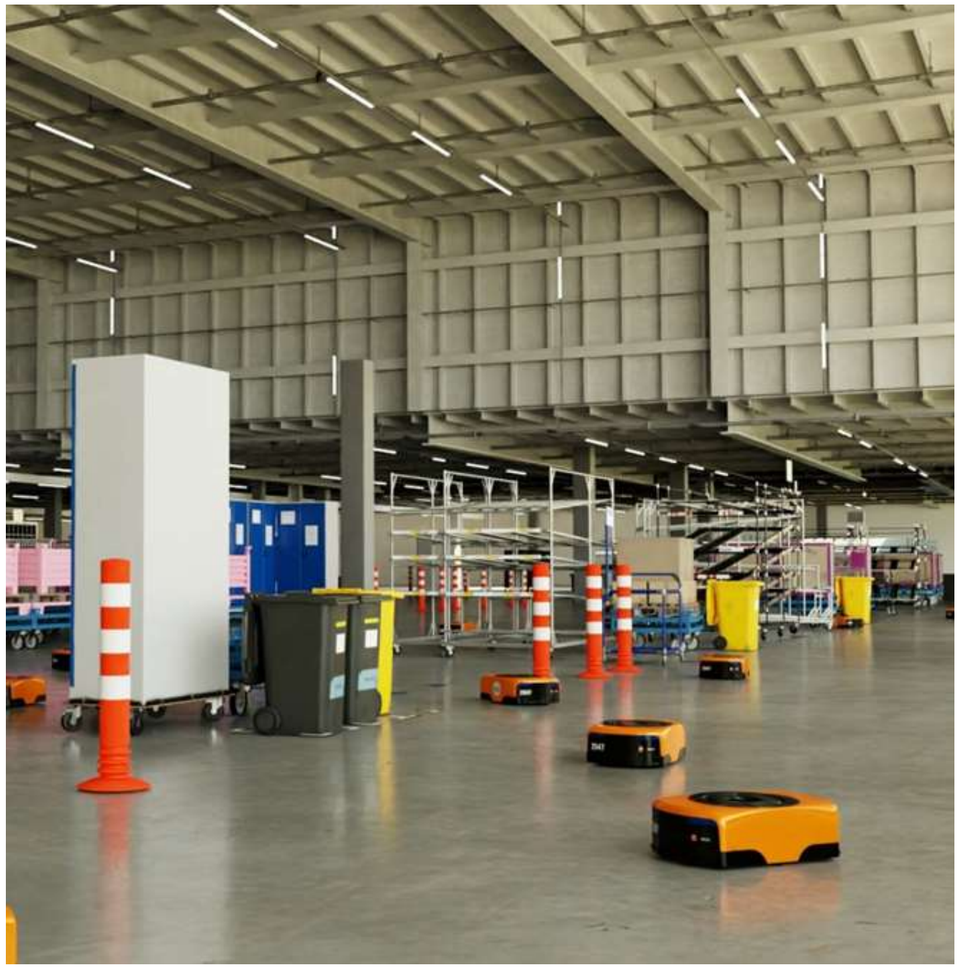
True-to-reality simulation achieves faster time-to-production with higher build quality than purely physically-based prototypes

**Anything that Moves will be Autonomous**

Autonomous machines excel in maximum efficiency and accuracy from product assembling, to warehouses management.

**Anything Autonomous will be Simulated**

AI agents can only achieve intelligence by training in scalable, photorealistic environments that obey the laws of physics.



# BUILDING THE FACTORY OF THE FUTURE

## BMW Group

BMW Group implemented an end-to-end system based on NVIDIA technologies, with logistics robots developed using one software architecture, running on NVIDIA's open Isaac robotics software platform. Navigation robots transport material autonomously, while manipulation robots select and organize parts. NVIDIA technology was used to design, train, develop, simulate, and deploy the system. Robots are virtually trained and tested using NVIDIA Isaac Sim operating in NVIDIA's Omniverse virtual environment. Multiple BMW Group personnel in different geographies can all work in one simulated environment.

BMW GROUP

# TRANSFORMING VFX WORKFLOWS

## Industrial Light & Magic

“NVIDIA continues to advance state-of-the-art graphics hardware, and NVIDIA Omniverse showcases what is possible with real-time ray tracing. The potential to improve the creative process through all stages of VFX and animation pipelines will be transformative.”

— Francois Chardavoine, VP of Technology

INDUSTRIAL LIGHT & MAGIC

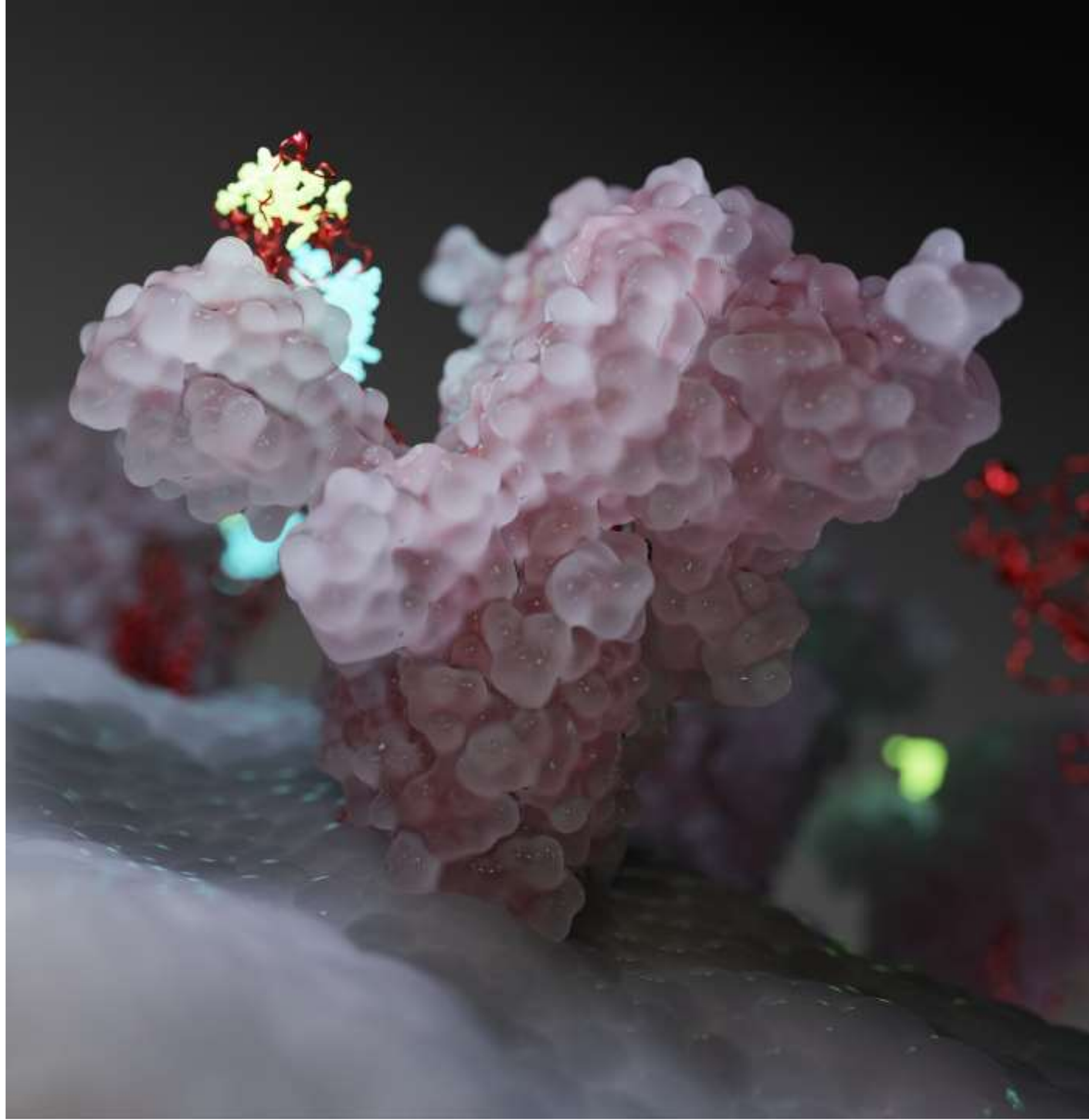
# VISUALIZING A PATH FORWARD

## Folding@home + NVIDIA

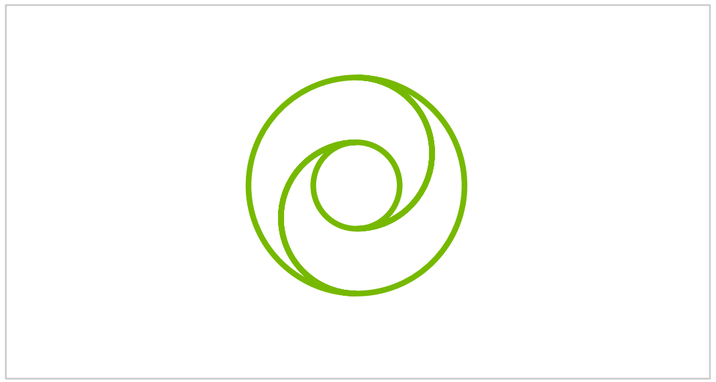
A million citizen scientists donated time on their home systems so the Folding@home consortium could calculate the intricate movements of proteins inside the coronavirus. Then a team of NVIDIA simulation experts combined the best tools from multiple industries to let the researchers see their data in a whole new way.

“Researchers are not confined to scientific visualization tools, they can use the same tools the best artists and movie makers use to deliver a cinematic rendering — we’re bringing these two worlds together.”

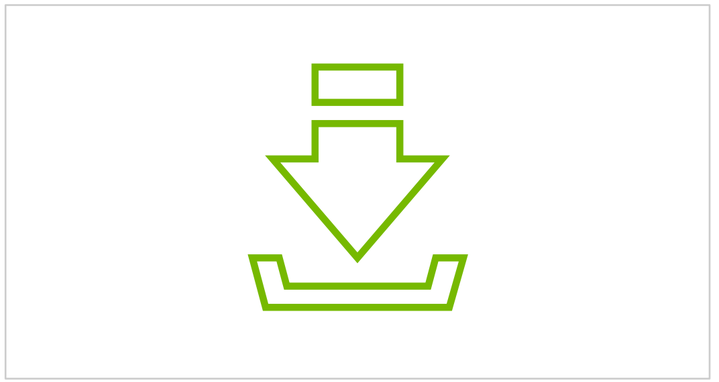
– Peter Messmer, Scientific Visualization, NVIDIA



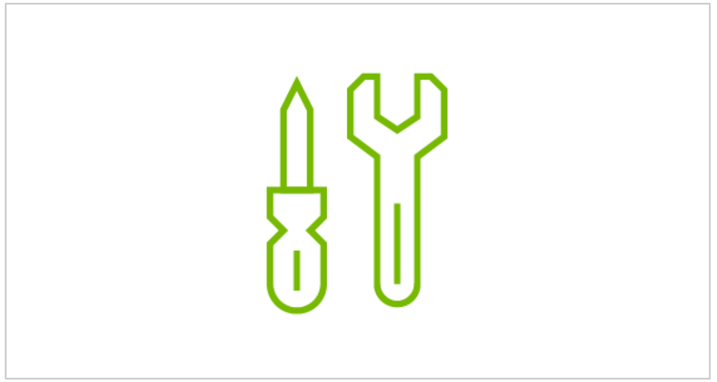
# LEARN MORE ABOUT OMNIVERSE



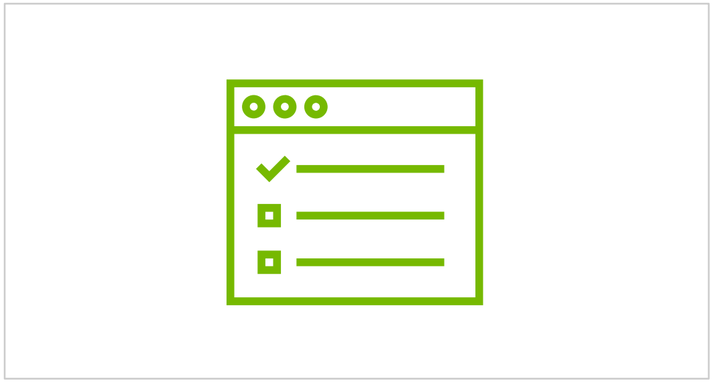
WEBSITE



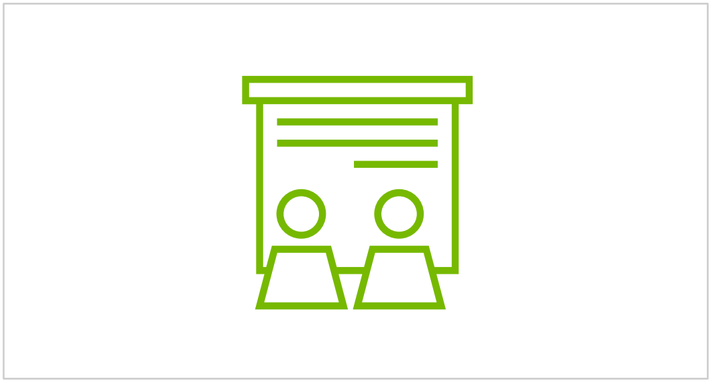
OPEN BETA DOWNLOAD



DEVELOPER TOOLS



EARLY ACCESS CLOSED BETA

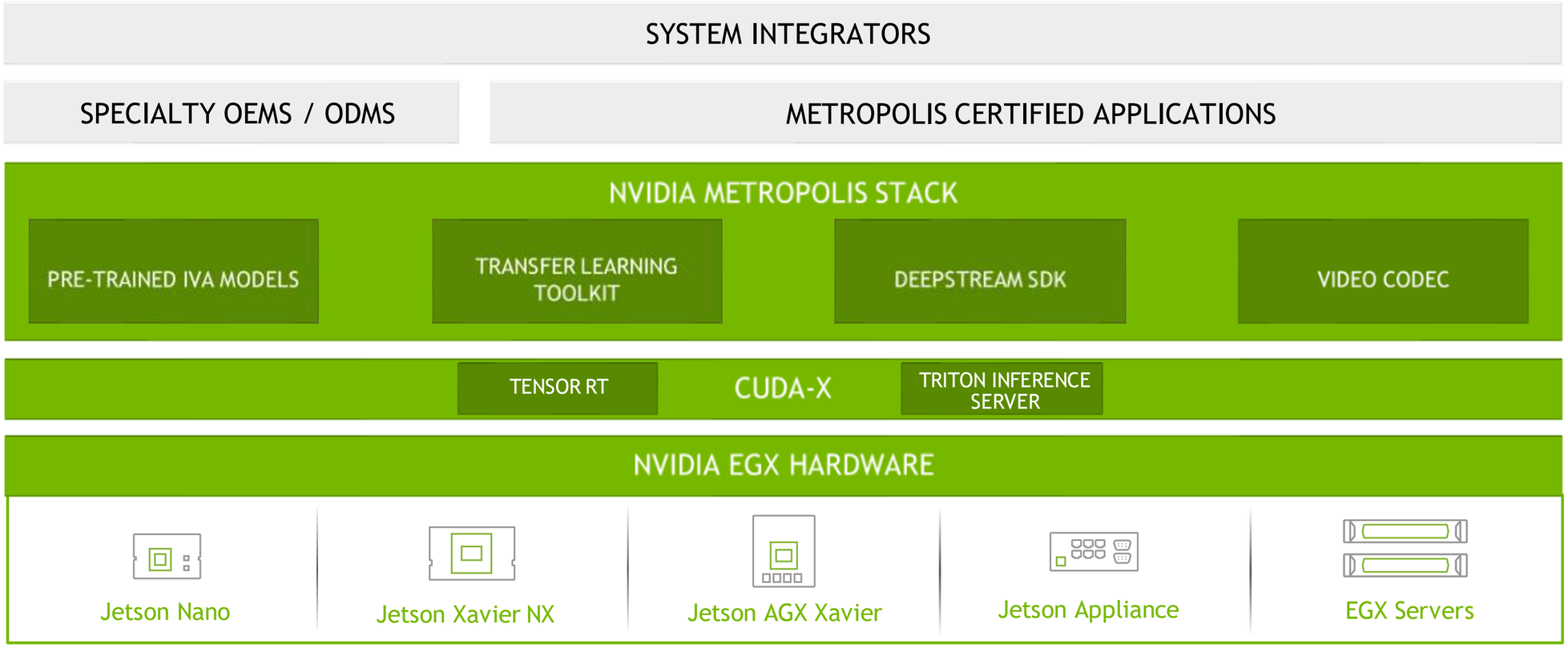


TUTORIAL COLLECTION

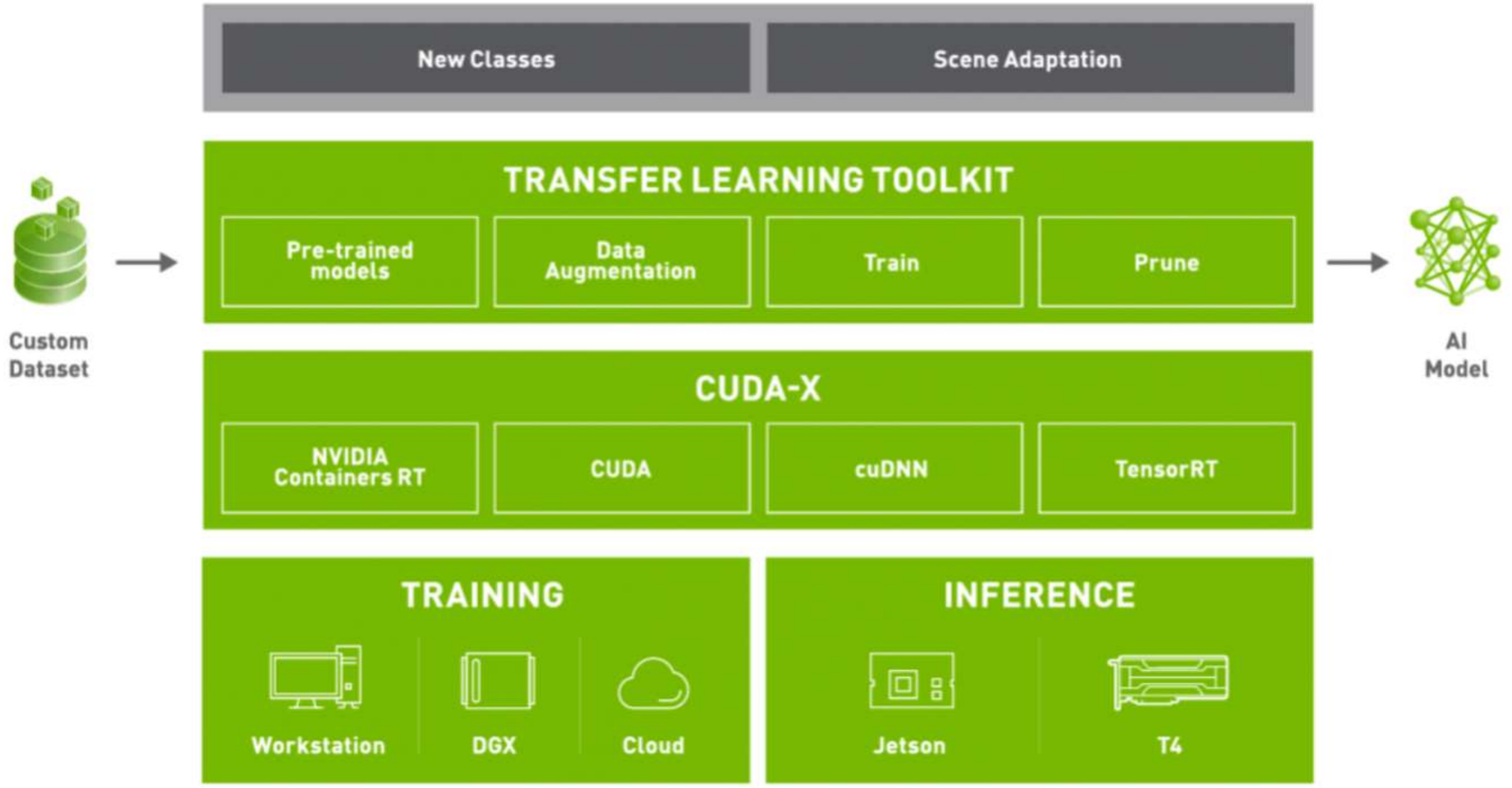
# METROPOLIS
# IVA (INTELLIGENT VIDEO ANALYTICS)

# NVIDIA METROPOLIS

## AI Application Framework for Smart Sensors



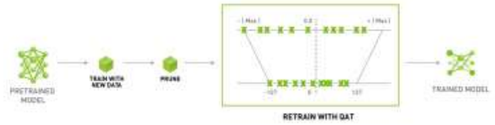
# TRANSFER LEARNING TOOLKIT (TLT)



# TRANSFER LEARNING TOOLKIT 2.0

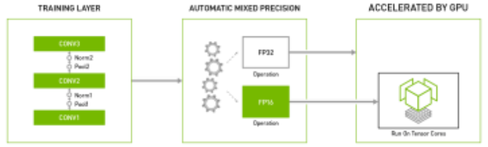
## KEY FEATURES


**Purpose-Built Pretrained Models**



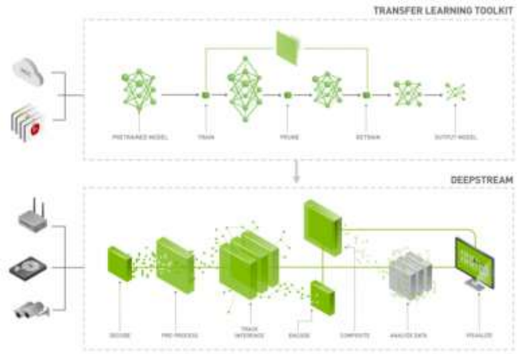

Achieve up to **2X** inference speedup while maintaining comparable accuracy to FP16/FP32

**Quantization Aware Training**




Achieve up to **1.5X** higher performance by training in mixed precision using Tensor Cores

**Automatic Mixed Precision**




Train with TLT and deploy using DeepStream to reduce development time by ~**10X**

**End-to-end vision AI with DeepStream**

# PURPOSE-BUILT PRETRAINED MODELS

## Fastest production path with highly accurate AI

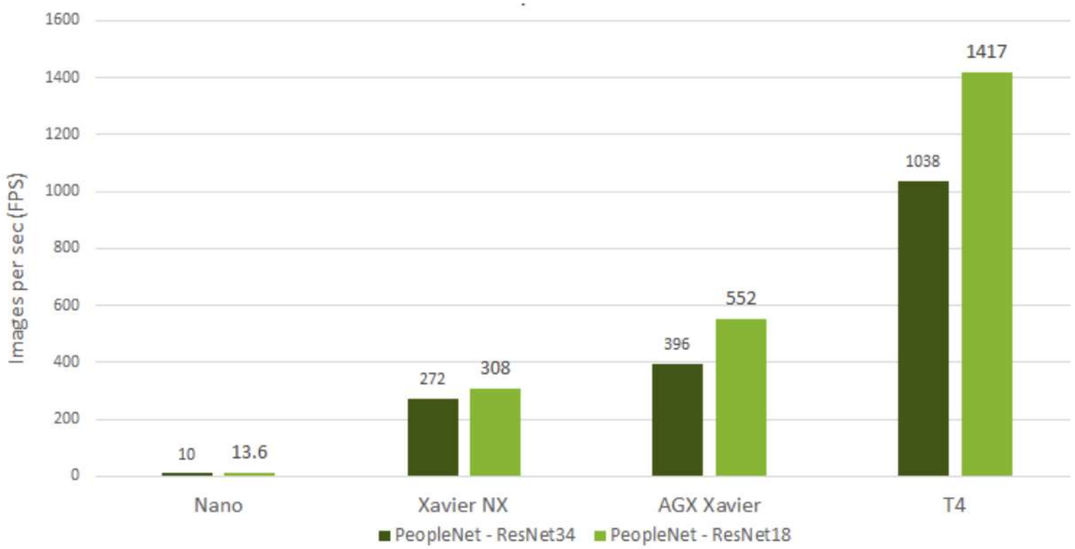
| Model | Network Architecture | Accuracy |
|---|---|---|
| DashCamNet | DetectNet_v2-ResNet18 | 80% |
| FaceDetect-IR | DetectNet_v2-ResNet18 | 96% |
| PeopleNet | DetectNet_v2-ResNet34 | 84% |
| TrafficCamNet | DetectNet_v2-ResNet18 | 83.5% |
| VehicleMakeNet | ResNet18 | 91% |
| VehicleTypeNet | ResNet18 | 96% |

Highly Accurate | Easily Retrainable | DeepStream Deployment Ready

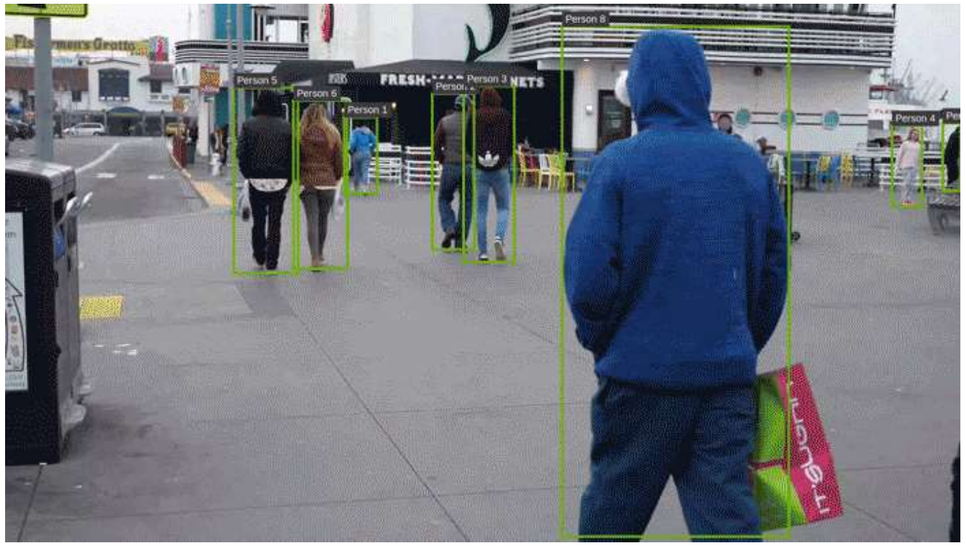
Note: Pruned and unpruned models readily available on NGC for free

# PEOPLENET: REAL-TIME INFERENCE PERFORMANCE

## Detect persons, bags and faces




VIDEO DEMO




Number of classes: 3
Dataset: 750k frames

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

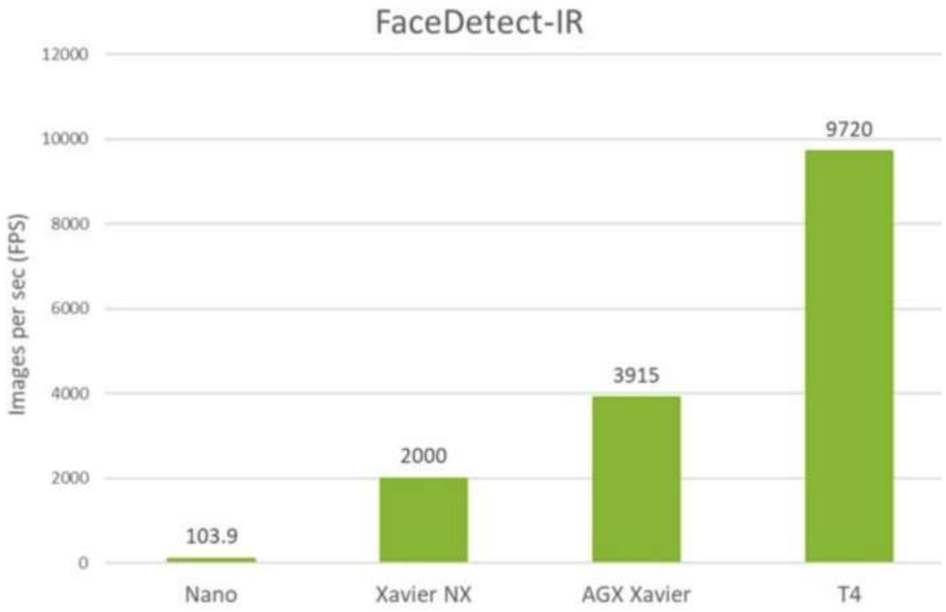
https://ngc.nvidia.com/catalog/models/nvidia:tlt_peoplenet



NVIDIA

# FACEDETECT-IR: REAL-TIME INFERENCE PERFORMANCE

## Detect one or more faces in each image / video



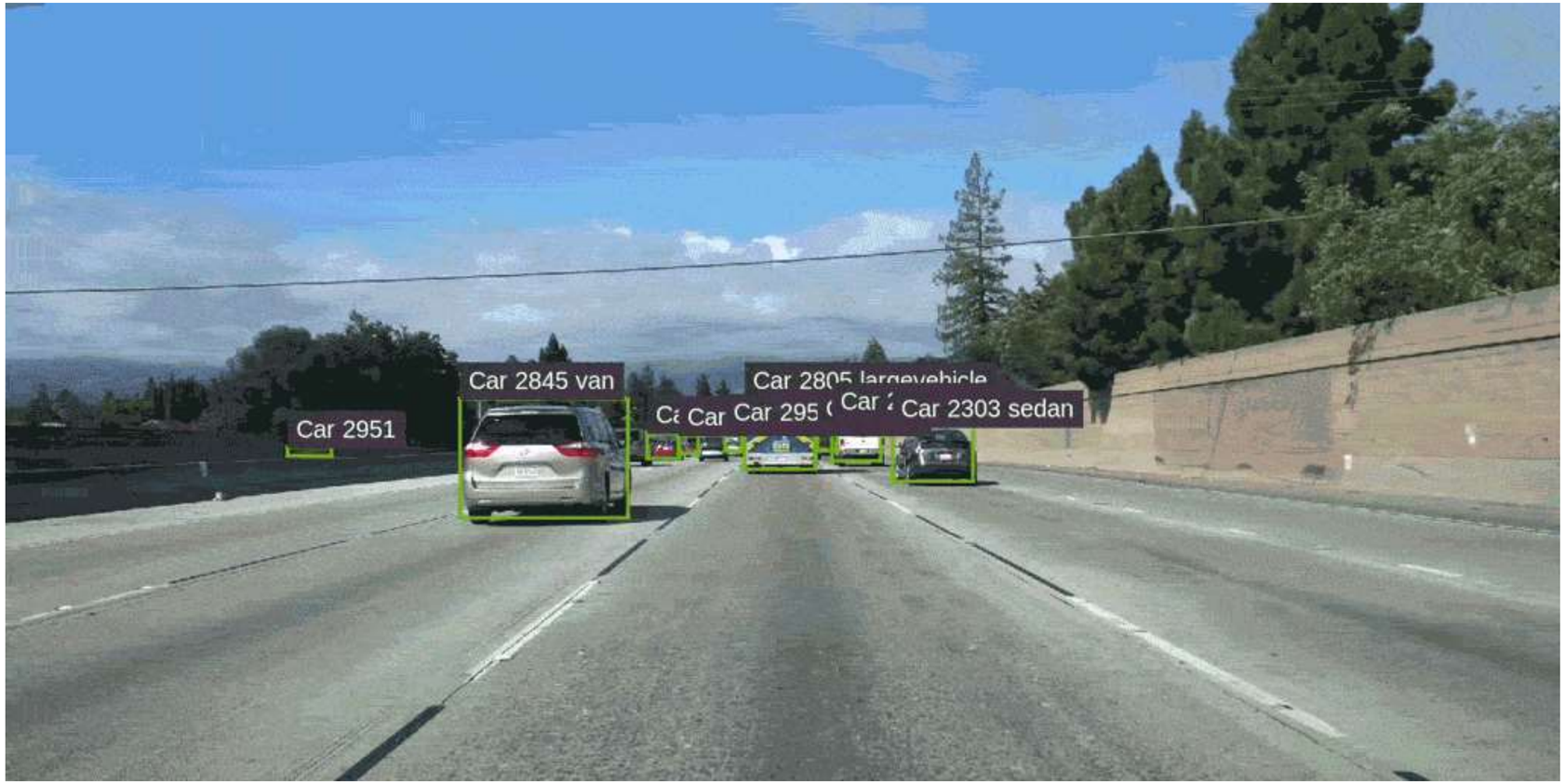


Number of classes: 1
Dataset: 600k images

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

# VIDEO DEMO

## DashCamNet and VehicleTypeNet In-Action



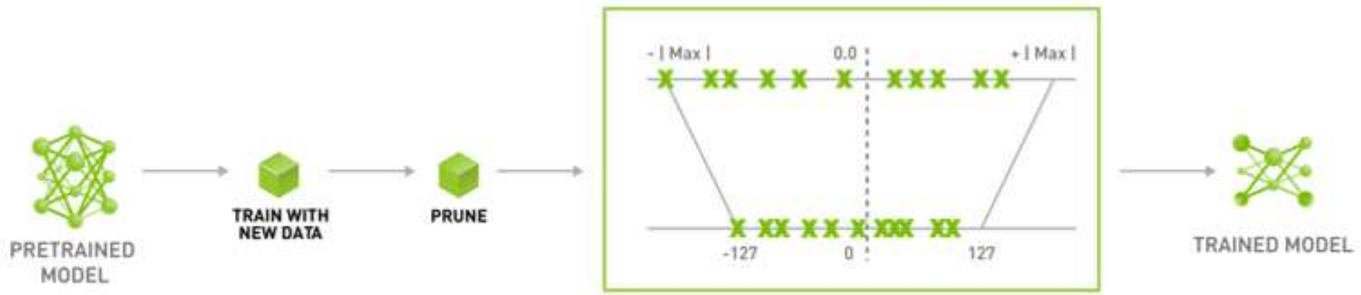
# VIDEO DEMO

## TrafficCamNet and VehicleMakeNet In-Action

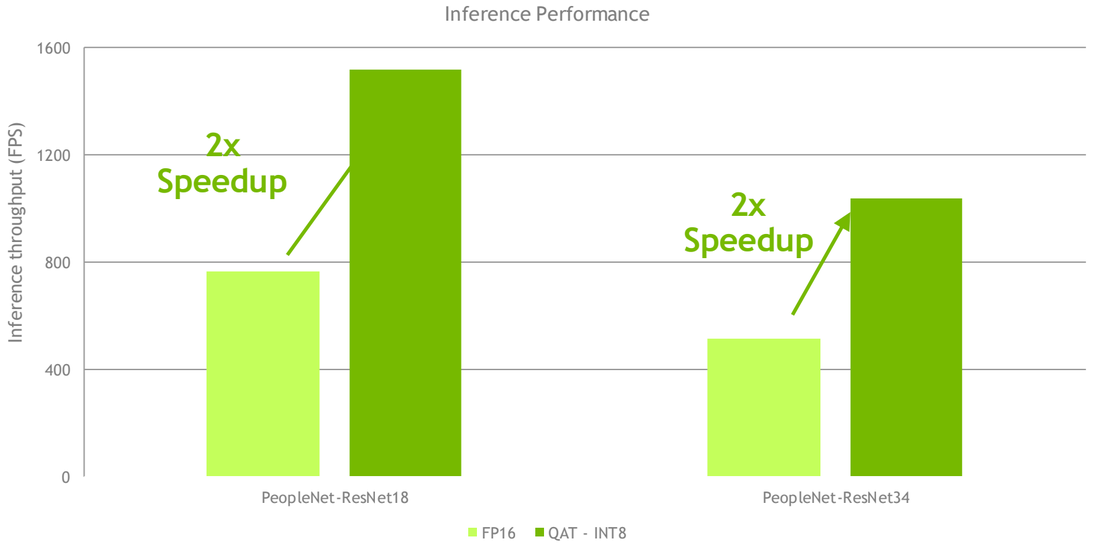
# QUANTIZATION AWARE TRAINING

## Maintain Comparable Performance & Speedup Inference using INT8 Precision

QAT workflow: Train -> Prune -> Re-train with QAT -> Export pruned, INT8 model




QAT supported on all object detection models




Accuracy

| | Baseline FP32 mAP | INT8 mAP with PTQ | INT8 mAP with QAT |
|---|---|---|---|
| PeopleNet-ResNet34 | 78.37 | 59.06 | 78.06 |
| PeopleNet-ResNet18 | 80.2 | 62 | 79.57 |

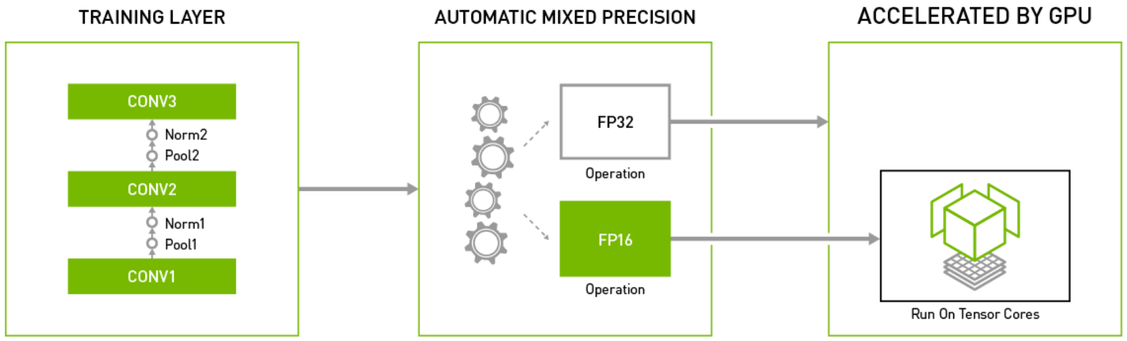
PTQ - Post-training quantization
QAT - Quantization-aware training

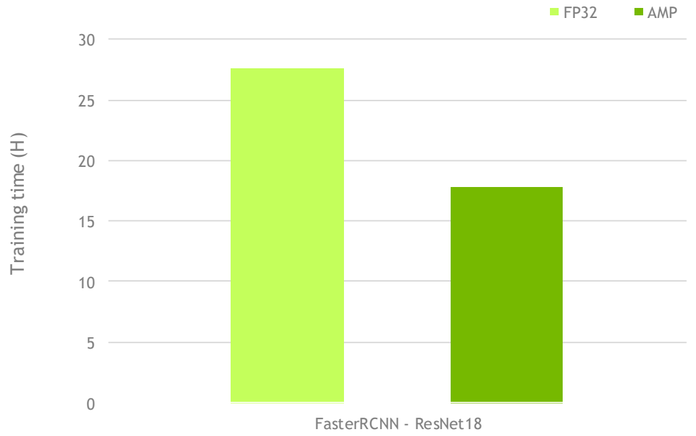
Developer Tutorial on using QAT with TLT

# AUTOMATIC MIXED PRECISION TRAINING USING TLT

## Natively supported with TLT to take advantage of Tensor Cores

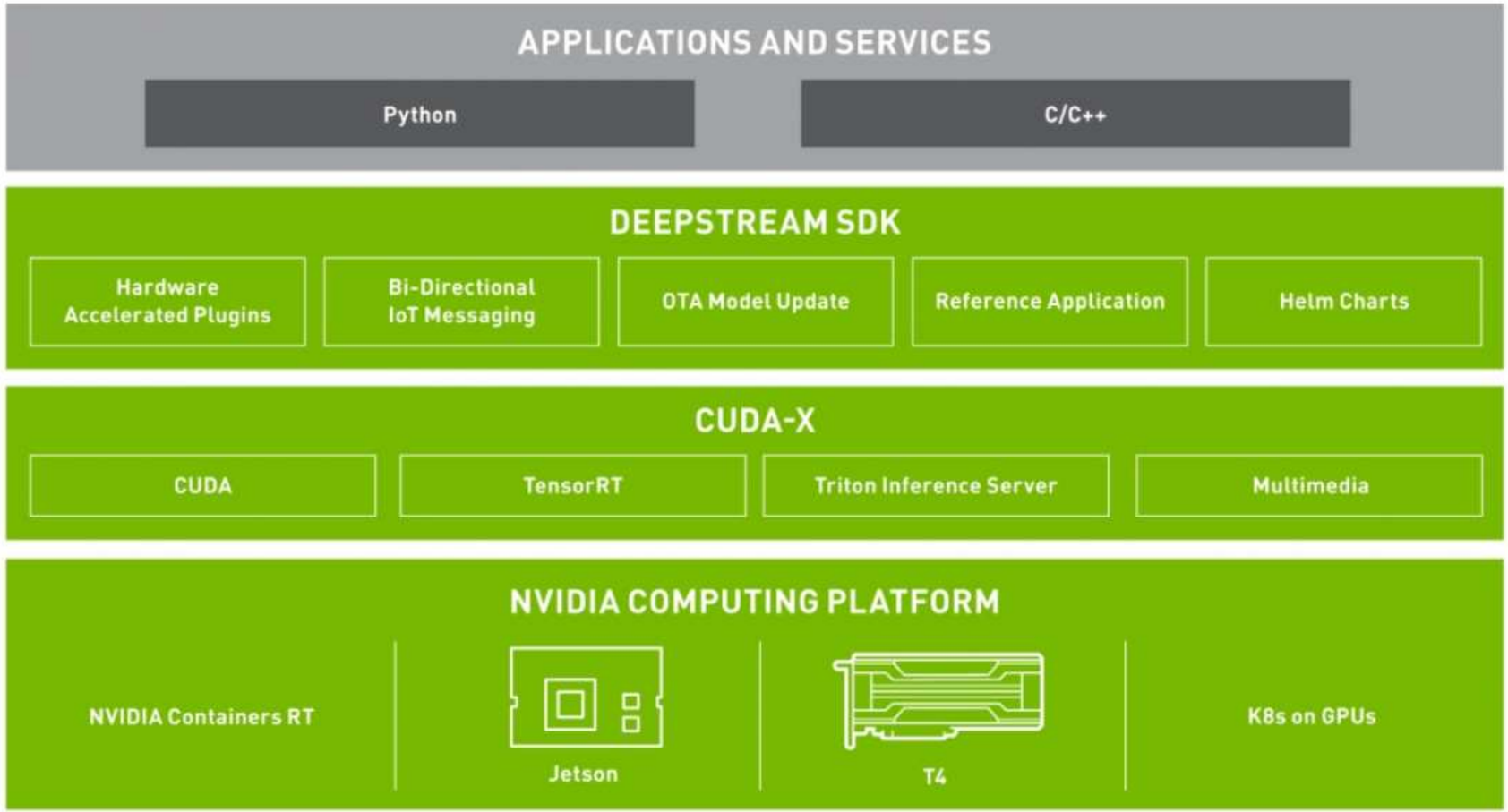
- Speed-up math intensive operations by using Tensor Cores
- Speed-up memory intensive operations by using half the bytes
- Reduce memory requirements




**1.5x Speedup**

# DEEPSTREAM SDK



# DEEPSTREAM 5.0 KEY FEATURES

## Develop in Python

Choose from C/C++ or Python development and achieve comparable performance

## Bi-directional Messaging


Send metadata, monitor health, multimedia. Record events, OTA update, orchestrate apps

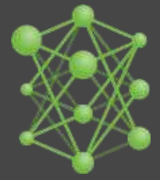
## Over-the-Air Model Update


On-the-fly model update with **zero downtime**

## Smart Record


Selective record saves valuable disk space and provides faster searchability

## Direct Integration to Triton Inference Server



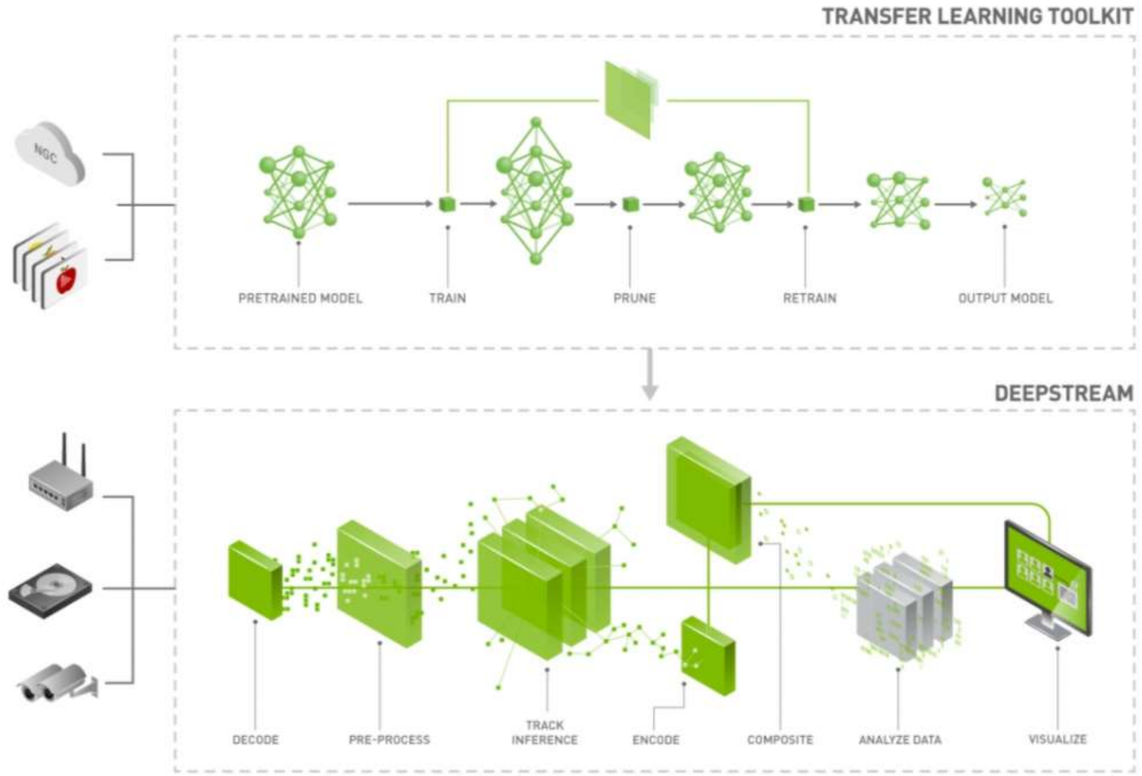
Deploy a model natively in TF, TF-TRT, PyTorch, or ONNX in DeepStream pipeline

## RHEL8

Red Hat Enterprise Linux

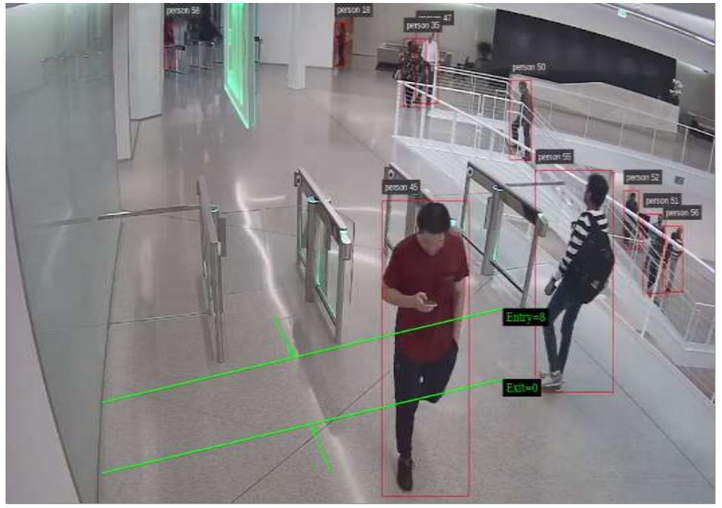
Build and deploy DeepStream apps natively using RHEL

# END-TO-END AI USING TLT AND DEEPSTREAM



# END-TO-END REFERENCE APPS

## Get started today




People Analytics for Crowded Spaces
Using TLT and DeepStream

Social Distancing App
Using DeepStream


Face Mask/No Mask Detection
Using TLT and DeepStream

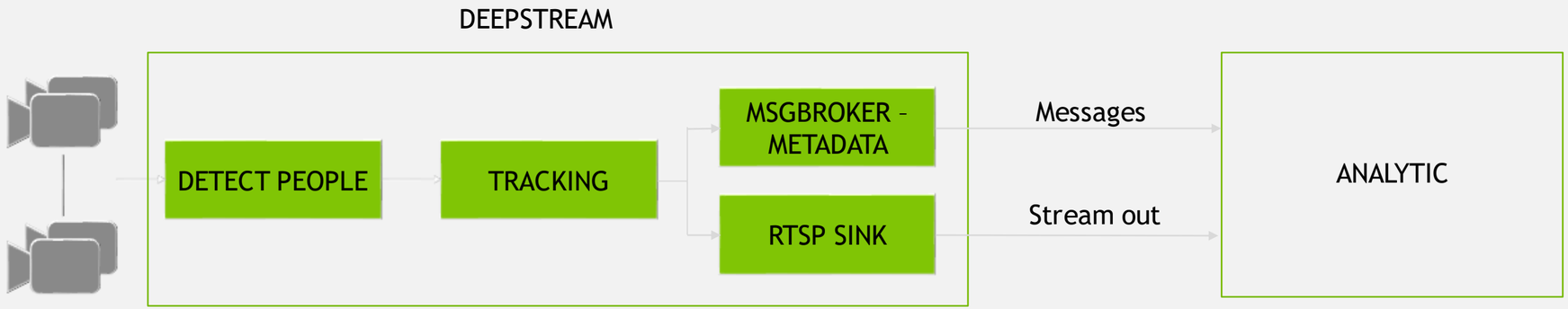
# PEOPLE ANALYTICS FOR CROWDED SPACES

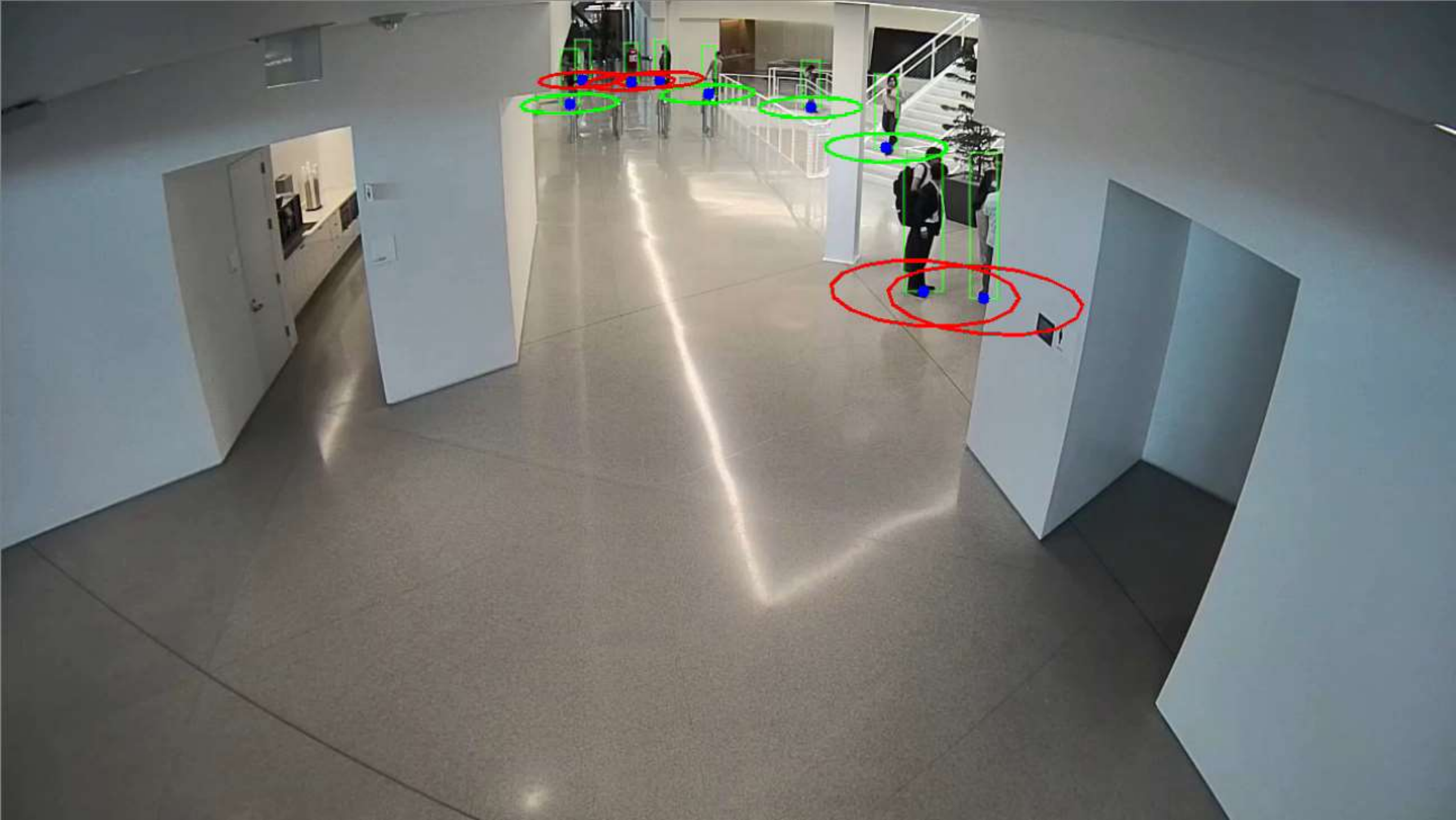
## Reference architecture & demo


Entry=0
Exit=0

# SOCIAL DISTANCING APP

## Reference architecture & demo





# FACE MASK/NO MASK DETECTION

## Reference architecture & demo


### What this project includes

- Transfer Learning Toolkit (TLT) scripts:
  - Dataset processing script to convert it in KITTI format
  - Specification files for configuring tlt-train, tlt-prune, tlt-evalute
- DeepStream (DS) scripts:
  - deepstream-app config files (For demo on single stream camera and detection on stored video file)

### What this project does not provide

- Trained model for face-mask detection; we will go through step by step to produce detetctnet_v2 (with ResNet18 backbone) model for face-mask detection.
- NVIDIA specific dataset for faces with and without mask; we suggest following dataset based on our experiments.

GitHub: https://github.com/NVIDIA-AI-IOT/face-mask-detection

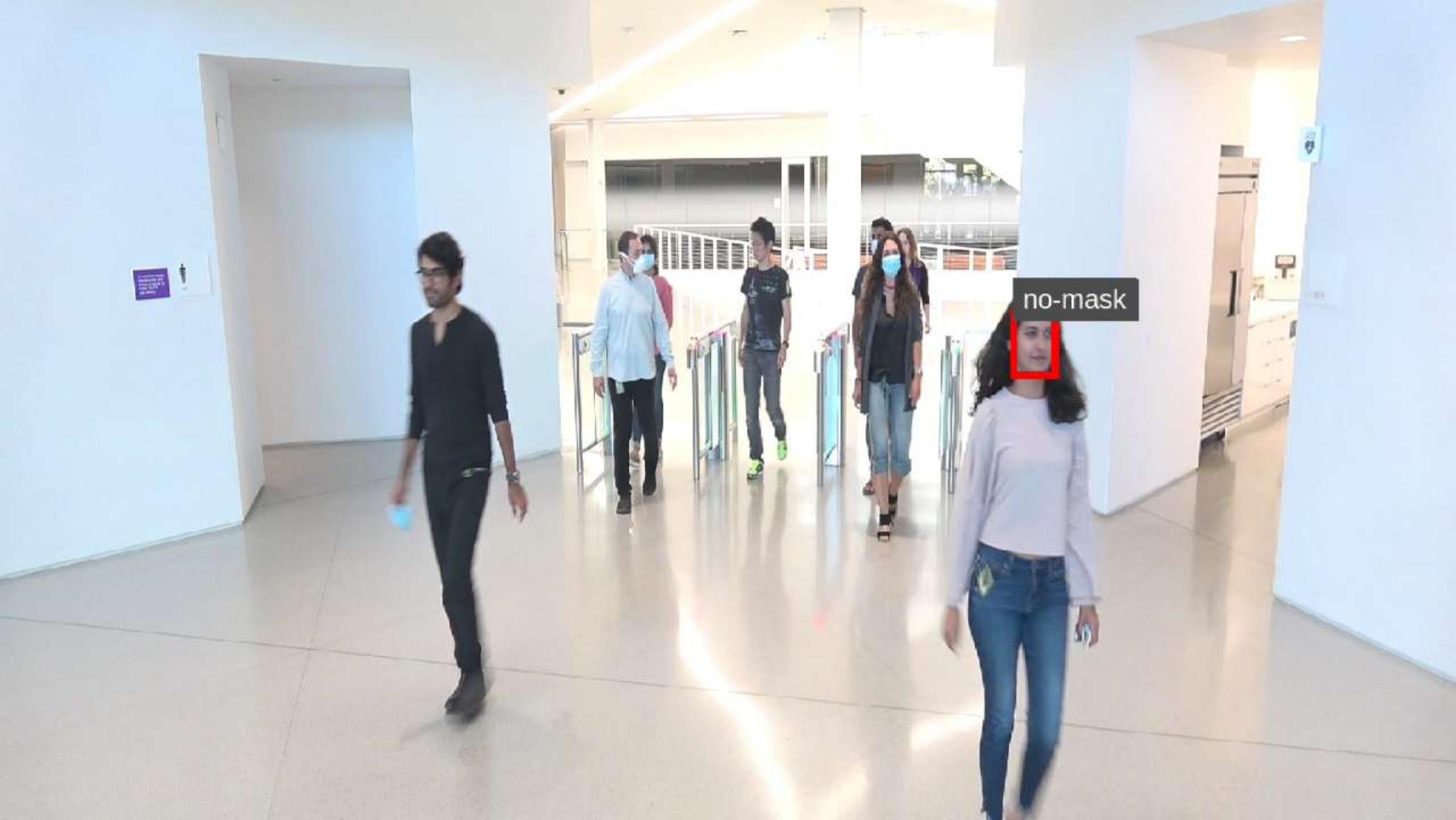
| Dataset Size | mAP (Mask/No-Mask) (%) |
|---|---|
| 27K | 78.98 (91.77, 66.19) |
| 4K | 86.12 (87.59, 84.65) |

| Pruned | mAP (Mask/No-Mask) (%) | Inference Evaluations on Nano | Inference Evaluations on Xavier NX | | Inference Evaluations on Xavier | |
|---|---|---|---|---|---|---|
| | | GPU (FPS) | GPU (FPS) | DLA (FPS) | GPU (FPS) | DLA (FPS) |
| No | 86.12 | 6.5 | 125.36 | 30.31 | 269.04 | 61.96 |
| Yes | 85.50 | 21.25 | 279 | 116.2 | 508.32 | 155.5 |

**TLT pruning increased FPS**

trained on ~4000 training images. **Pruning ratio (pruned model / original model)*100

https://developer.nvidia.com/blog/implementing-a-real-time-ai-based-face-mask-detector-application-for-covid-19/

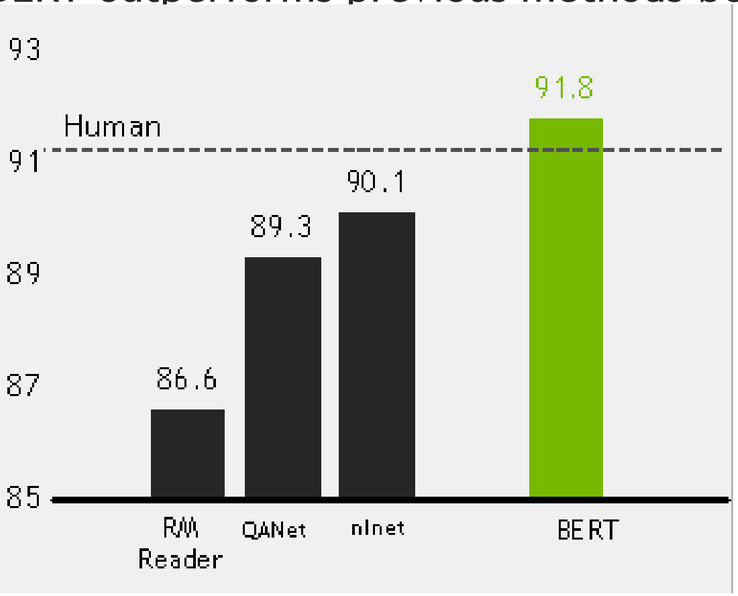

no-mask


AI & DATA SCIENCE

# BERT: FLEXIBILITY + ACCURACY FOR NLP TASKS

“BERT is a method of pre-training language representations, meaning that we train a general-purpose "language understanding" model on a large text corpus (like Wikipedia), and then use that model for downstream NLP tasks that we care about (like question answering).

BERT outperforms previous methods because it is the first *unsupervised*, *deeply bidirectional* system for pre-training NLP.”




Super Human Question & Answering

9th October, Google submitted GLUE benchmark

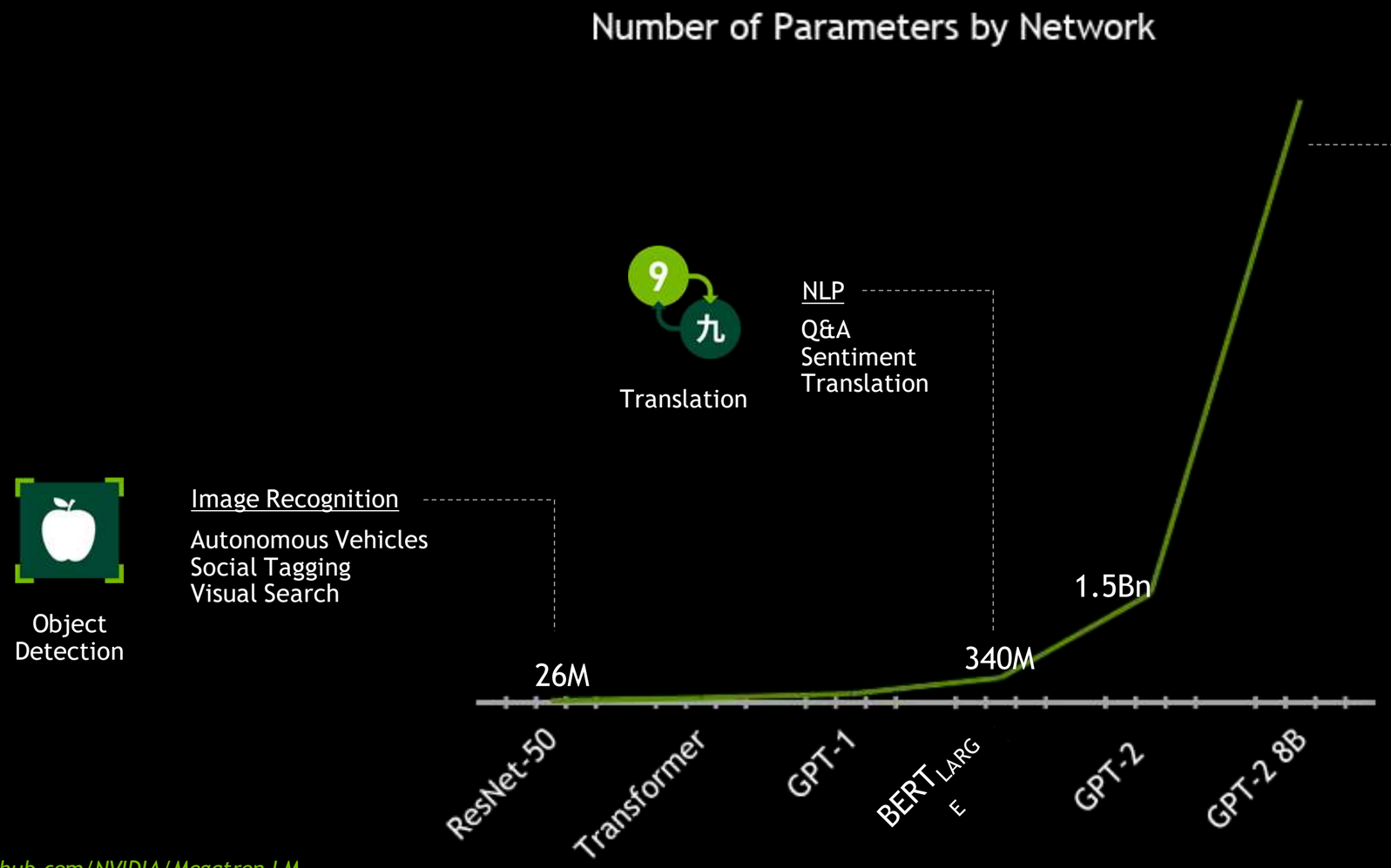
- **Sentence Pair Classification:** MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG
- **Single Sentence Classification:** SST-2, CoLA
- **Question Answering:** SQuAD
- **Single Sentence Tagging:** CoNLL - 2003 NER

| System | MNLI-(m/mm) 392k | QQP 363k | QNLI 108k | SST-2 67k | CoLA 8.5k | STS-B 5.7k | MRPC 3.5k | RTE 2.5k | **Average** - |
|---|---|---|---|---|---|---|---|---|---|
| Pre-OpenAI SOTA | 80.6/80.1 | 66.1 | 82.3 | 93.2 | 35.0 | 81.0 | 86.0 | 61.7 | 74.0 |
| BiLSTM+ELMo+Attn | 76.4/76.1 | 64.8 | 79.8 | 90.4 | 36.0 | 73.3 | 84.9 | 56.8 | 71.0 |
| OpenAI GPT | 82.1/81.4 | 70.3 | 87.4 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.1 |
| $\text{BERT}_{\text{BASE}}$ | 84.6/83.4 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| $\text{BERT}_{\text{LARGE}}$ | **86.7/85.9** | **72.1** | **92.7** | **94.9** | **60.5** | **86.5** | **89.3** | **70.1** | **82.1** |


49

# DEEP LEARNING MODELS INCREASING IN COMPLEXITY

## Next-Level Use-Cases Require Gigantic Models



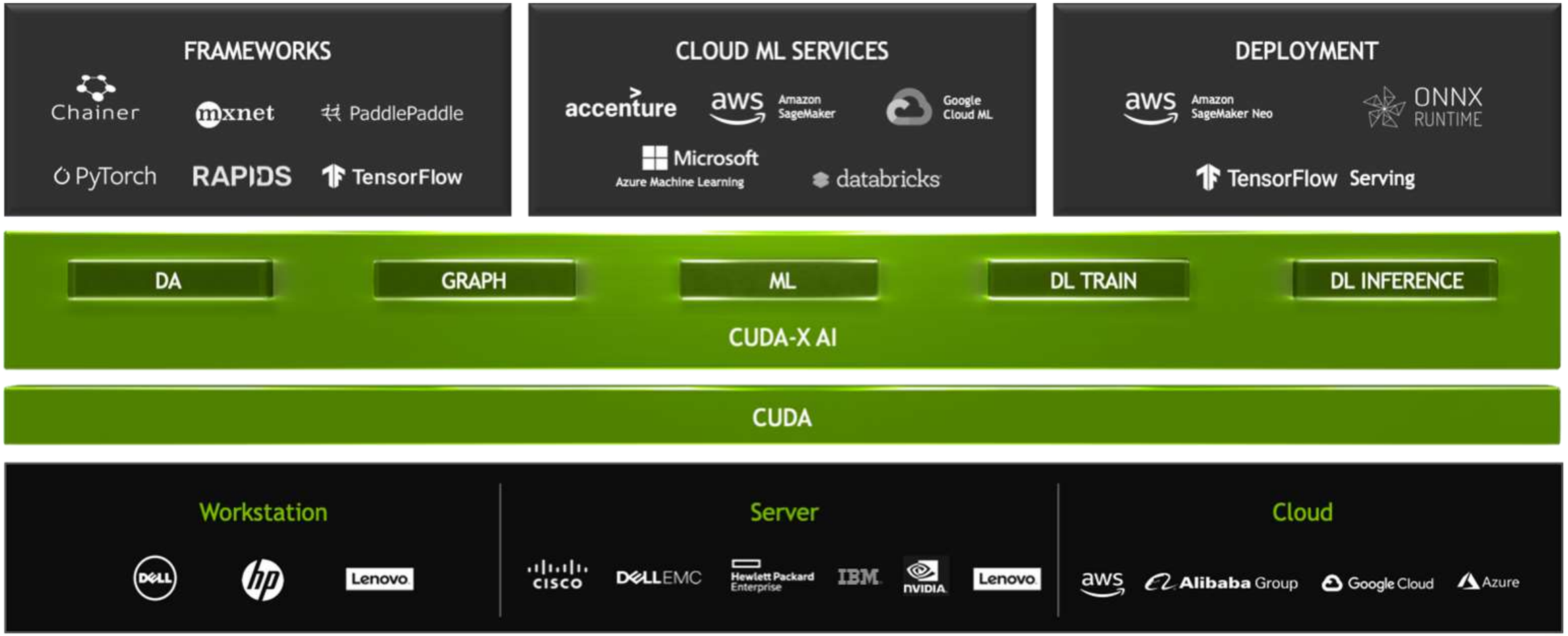

**Project Megatron**

**8.3B parameters**
**8-way Model Parallel**
**64-way Data Parallel**
**24x larger than BERT**

https://github.com/NVIDIA/Megatron-LM

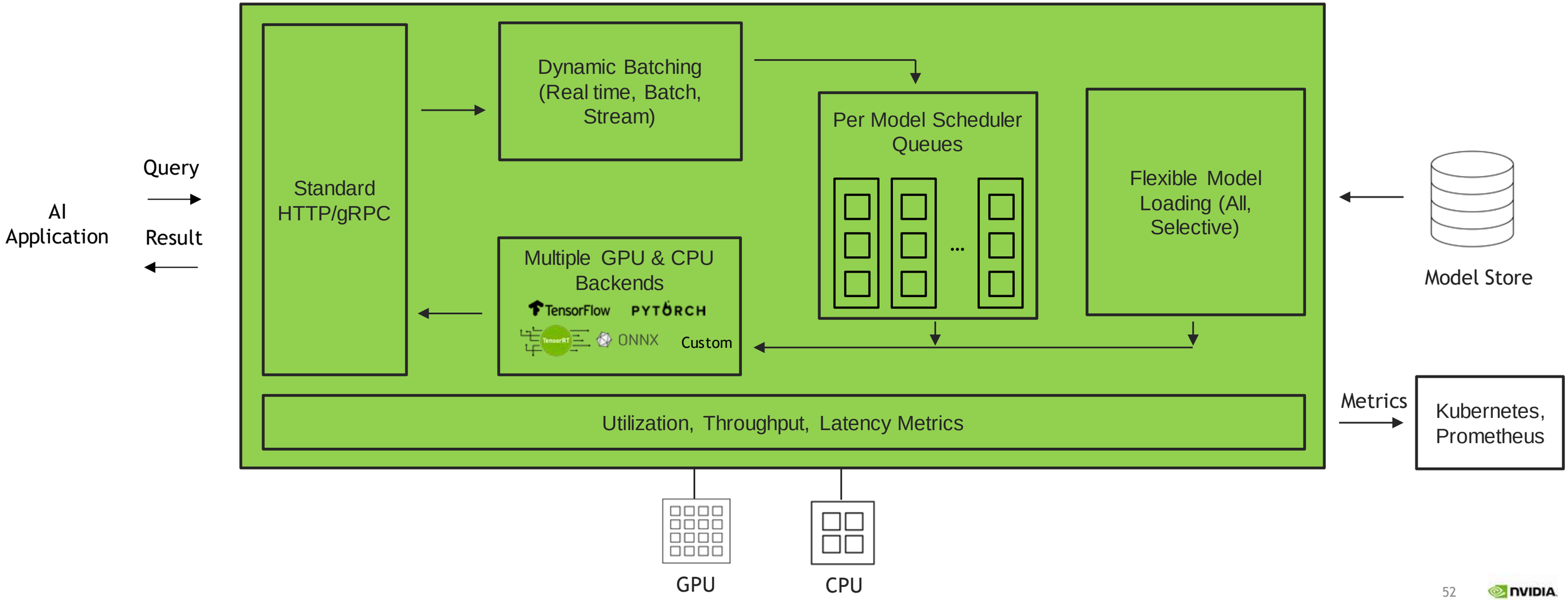
# CUDA-X AI TRANSFORMS DATA SCIENCE

## From Data Science to NVIDIA **Accelerated** Data Science with CUDA-X AI



# TRITON INFERENCE SERVER

## Open-Source Software For Scalable, Simplified Inference Serving



# DEVELOPERS CAN FOCUS ON MODELS AND APPLICATIONS

## Triton Takes Care of Plumbing To Deploy Models for Inference

### Multiple Frameworks

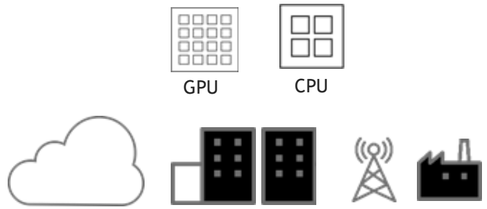

Custom

All Major Framework Backends For Flexibility & Consistency

Concurrent Model Execution For High Throughput & Utilization

Standard HTTP/gRPC Communication

### Inferencing on GPU and CPU




Inference Serving on GPU & CPU Across

Cloud | Data Center | Edge

Bare metal | Virtualization

### Different Types of Queries

Real time Batch Stream Ensemble

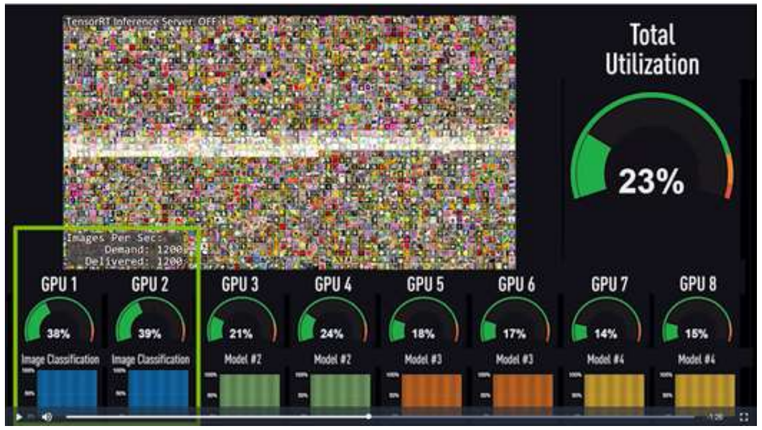
Support For Different Types Of Inference Queries For Different Use Cases

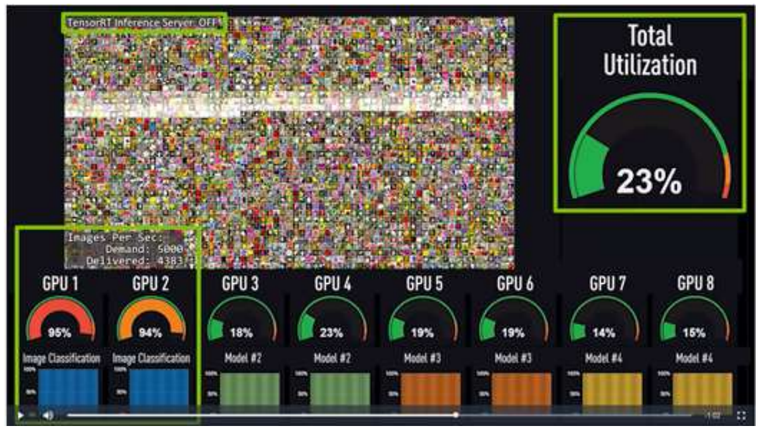
### Dynamic Batching

Dynamic Batching Maximizes Throughput Under Latency Constraint

# DESIGNED TO SCALE & MAXIMIZE GPU UTILIZATION

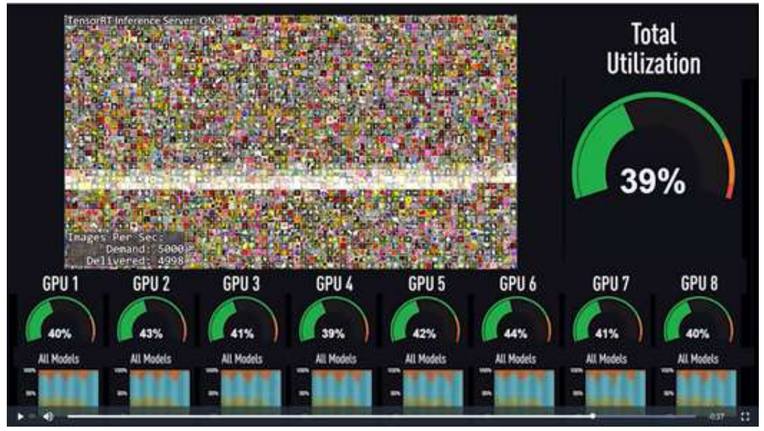
## Illustration of Triton vs Single Model per GPU Implementation

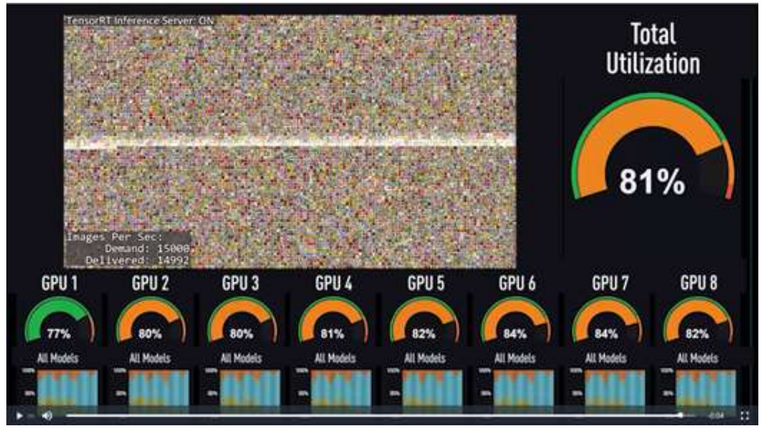



**Single Model per GPU Implementation**
GPU1 & GPU2
struggle with 5000 requests while other GPUs are idle

**Triton Inference Server**
With concurrent model execution, Triton easily handles up to 15000 requests with all GPUs equally utilized

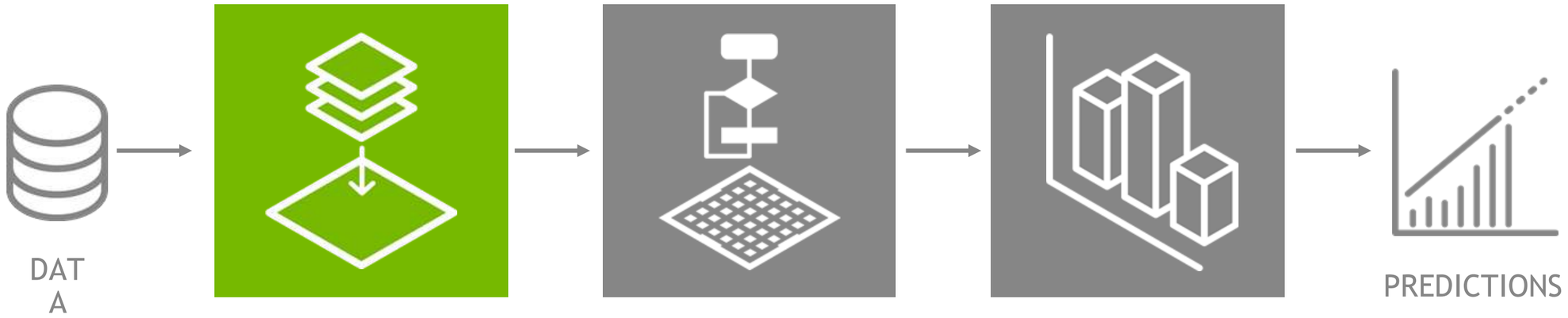



RAPIDS

# GPU-ACCELERATED DATA SCIENCE WORKFLOW WITH RAPIDS

## Built on CUDA-X AI




### DATA PREPARATION

GPUs accelerated compute for in-memory data preparation

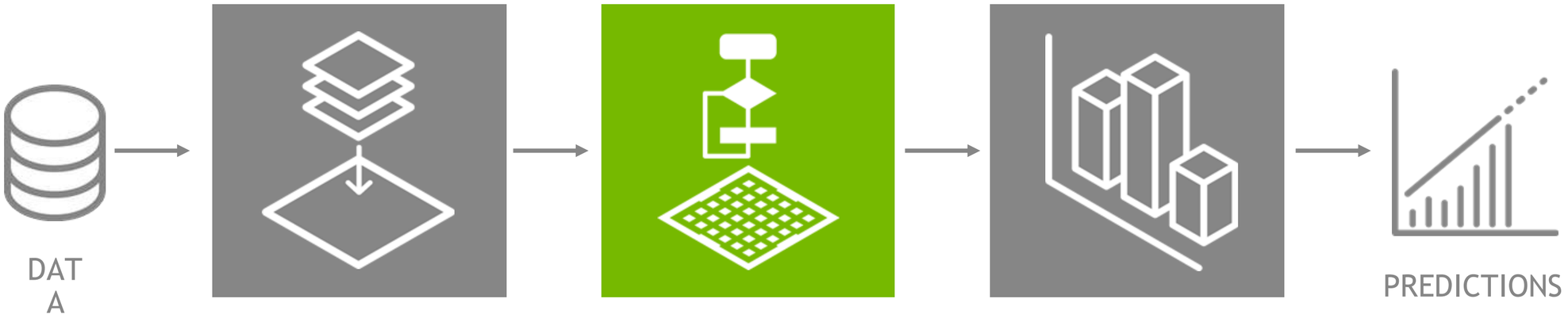
Simplified implementation using familiar data science tools

Python drop-in **pandas** replacement built on CUDA C++.

GPU-accelerated Spark (in development)

# GPU-ACCELERATED DATA SCIENCE WORKFLOW WITH RAPIDS

## Built on CUDA-X AI




### MODEL TRAINING

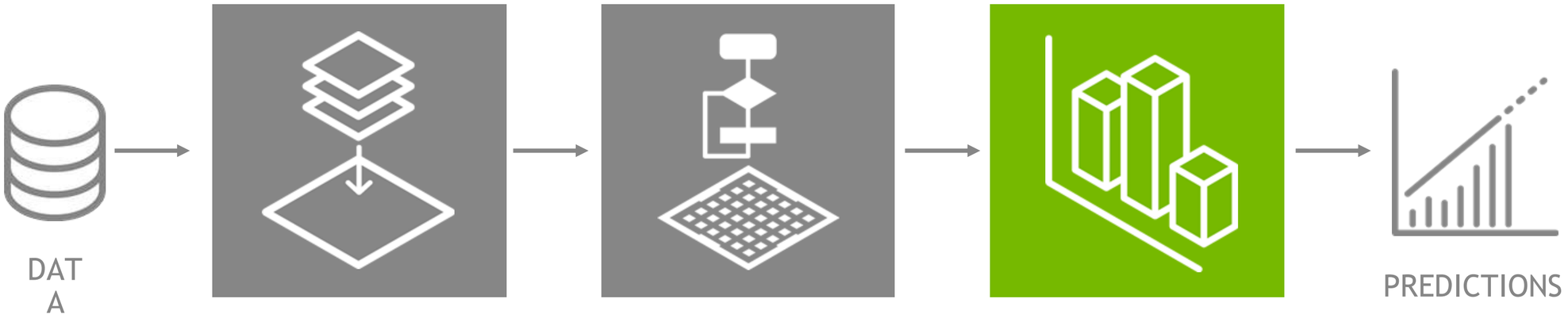
GPU-acceleration of today's most popular ML algorithms such as **XGBoost**

Also available are PCA, K-means, k-NN, DBScan, tSVD, and many more

Easy-to-adopt, scikit-learn like interface

# GPU-ACCELERATED DATA SCIENCE WORKFLOW WITH RAPIDS

## 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

# USE WITH MINIMAL CODE CHANGES

## GPU-Acceleration with the same XGBoost Usage

### BEFORE

```
import xgboost as xgb

params = {'max_depth': 3,
        'learning_rate': 0.1}


dtrain = xgb.DMatrix(X, y)
bst = xgb.train(params, dtrain)
```

### AFTER

```
import xgboost as xgb

params = {‘tree_method’: ‘gpu_hist’,
        'max_depth': 3,
        'learning_rate': 0.1}

dtrain = xgb.DMatrix(X, y)
bst = xgb.train(params, dtrain)
```

# DISTRIBUTED XGBOOST

## GPU-Accelerated XGBoost for Large Scale Workloads

GPU-acceleration for **XGBoost** with **Apache Spark** and **Dask**

Multiple nodes and multiple GPUs per node

Explore and prototype models on a PC, workstation, server, or cloud instance and scale to two or more nodes for production training

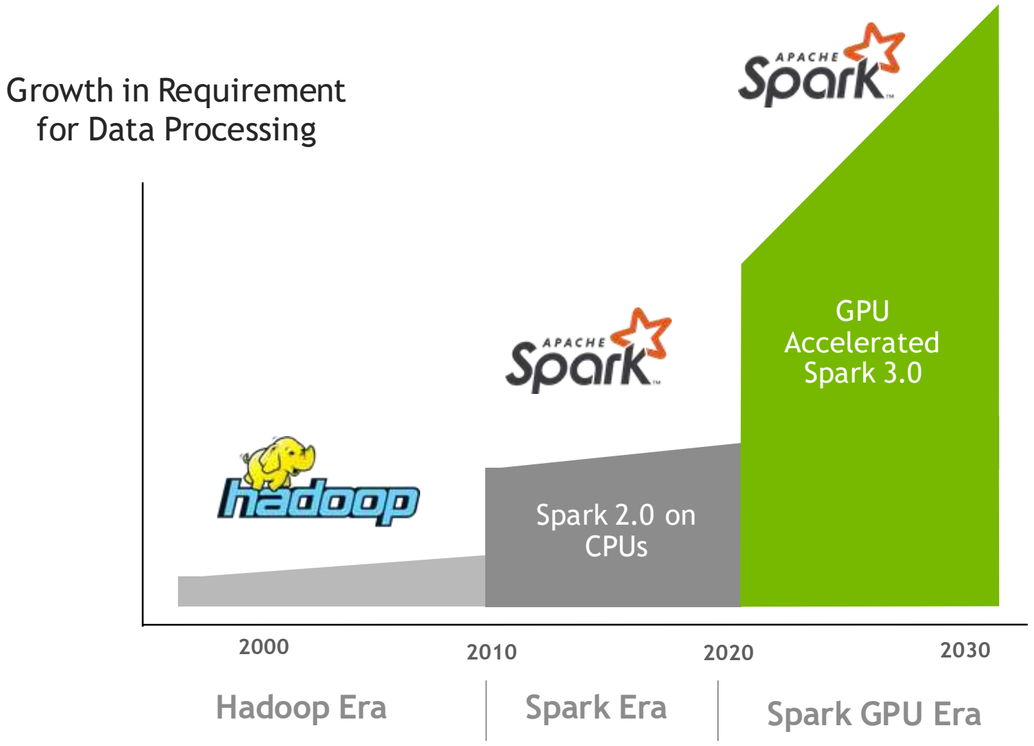
An ideal solution for GPU-accelerated clusters and enterprise scale workloads

Try out Dask support immediately using Google Cloud Dataproc

Download for on-prem and cloud deployments


# THE LEADING PLATFORM FOR SCALE OUT ANALYTICS

## GPU-accelerated Spark 3.0 unifies the pipeline for ETL, Machine & Deep Learning




Originally developed at UC Berkeley (2009)

Became top-level Apache Project in 2013

10 years of development /mature & broadly adopted

Optimized for in-memory distributed computing

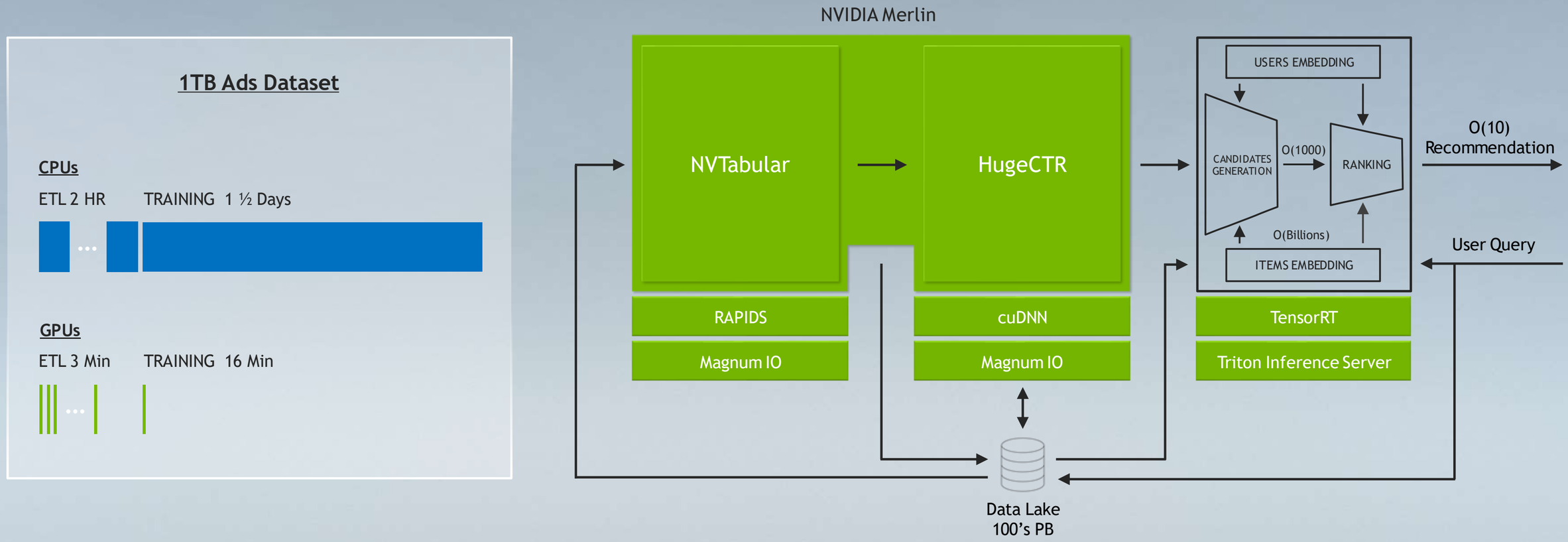
Modern Unified ETL and ML/DL Platform

*"These contributions lead to faster data pipelines, model training and scoring for more breakthroughs and insights with Apache Spark 3.0 and Databricks."*

Matei Zaharia, creator of Apache Spark and chief technologist at Databricks

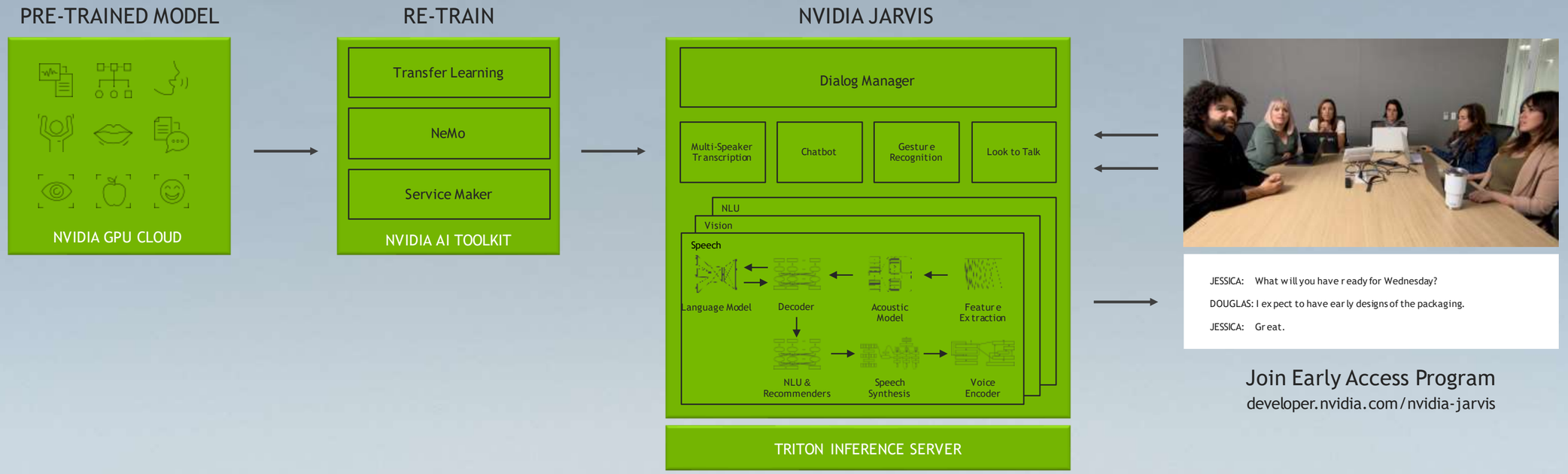
# MERLIN RECOMMENDER SDK

# ANNOUNCING
# NVIDIA MERLIN — DEEP RECOMMENDER APPLICATION FRAMEWORK



# JARVIS + NEMO CONVERSATIONAL AI

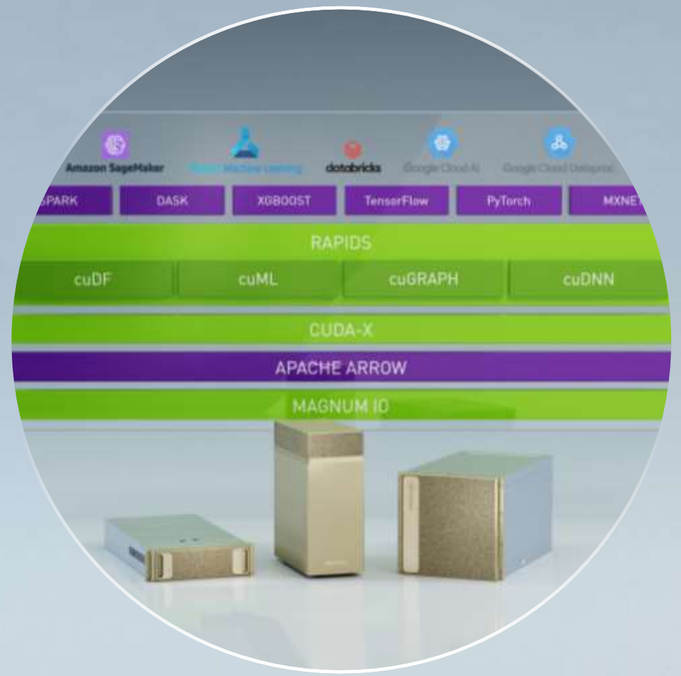
# ANNOUNCING NVIDIA JARVIS
# MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK

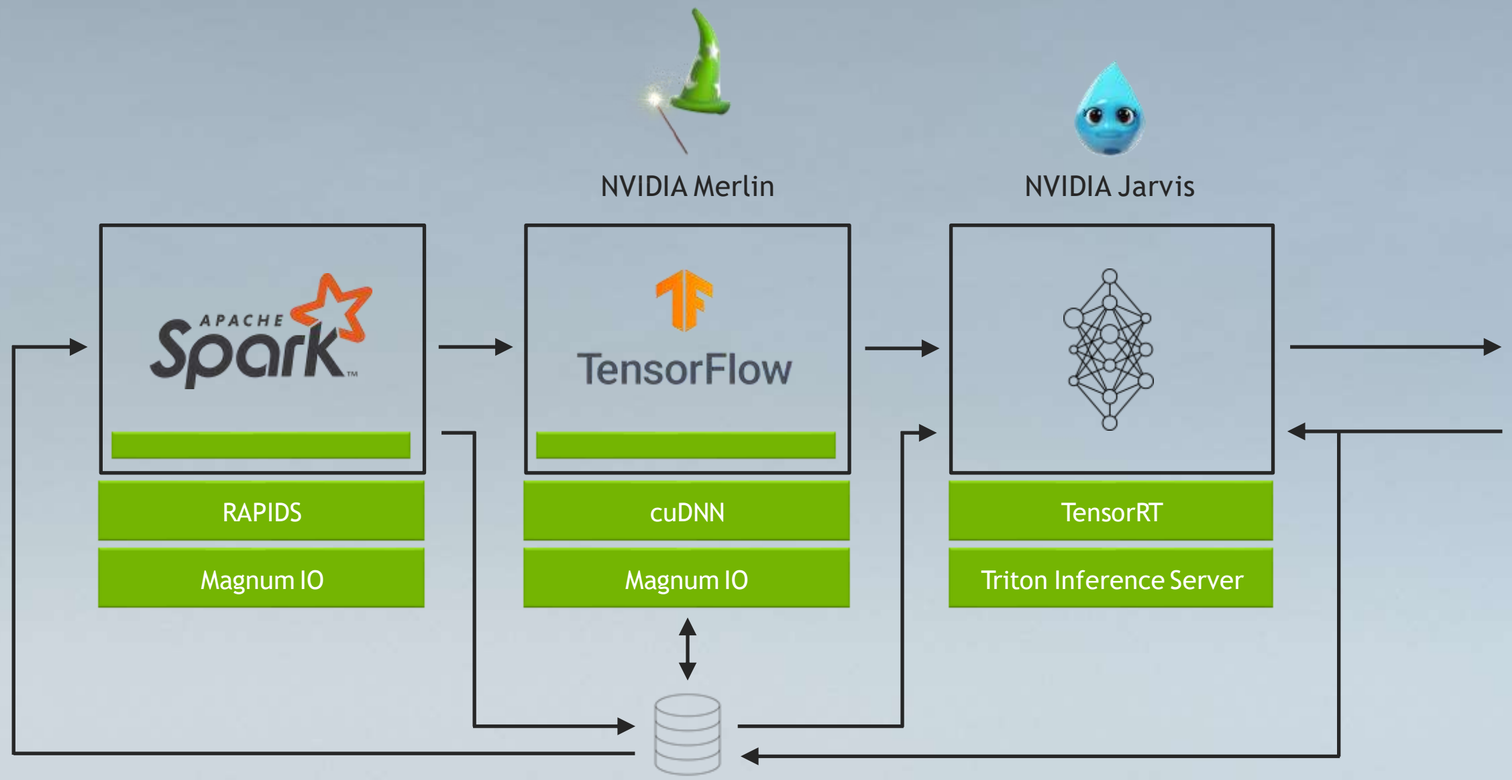



Join Early Access Program

developer.nvidia.com/nvidia-jarvis

# NVIDIA AI


Amazon SageMaker
databricks
DASK
XGBOOST
TensorFlow
PyTorch
RAPIDS
cuDF
cuML
cuGRAPH
cuDNN
CUDA-X
APACHE ARROW
MAGNUM IO


NVIDIA Merlin
NVIDIA Jarvis
APACHE Spark
TensorFlow
RAPIDS
Magnum IO
cuDNN
Magnum IO
TensorRT
Triton Inference Server

# CUQUANTUM

# A NEW COMPUTING MODEL - QUANTUM

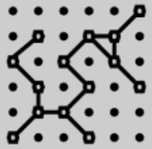
## NEW COMPUTING MODEL

## POTENTIAL USE CASES

Computational Finance

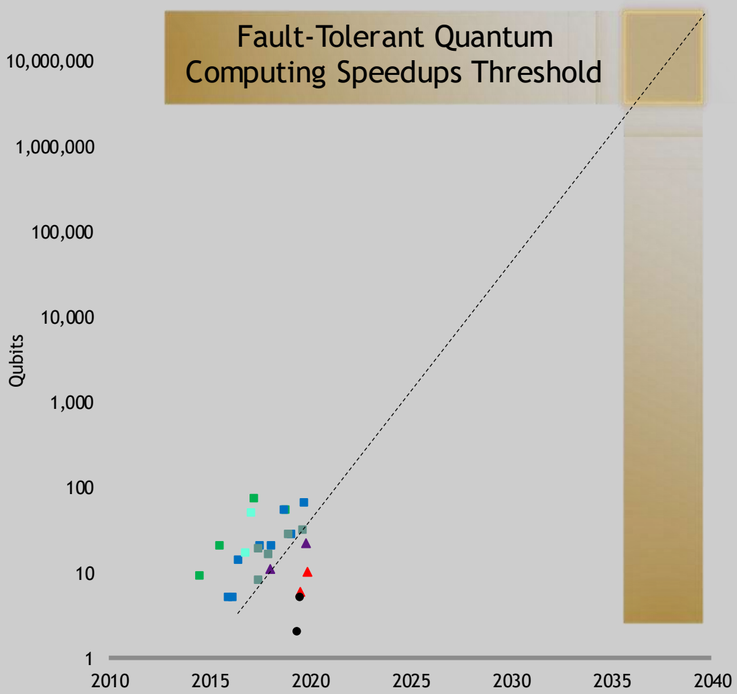
Quantum Chemistry

Cryptography



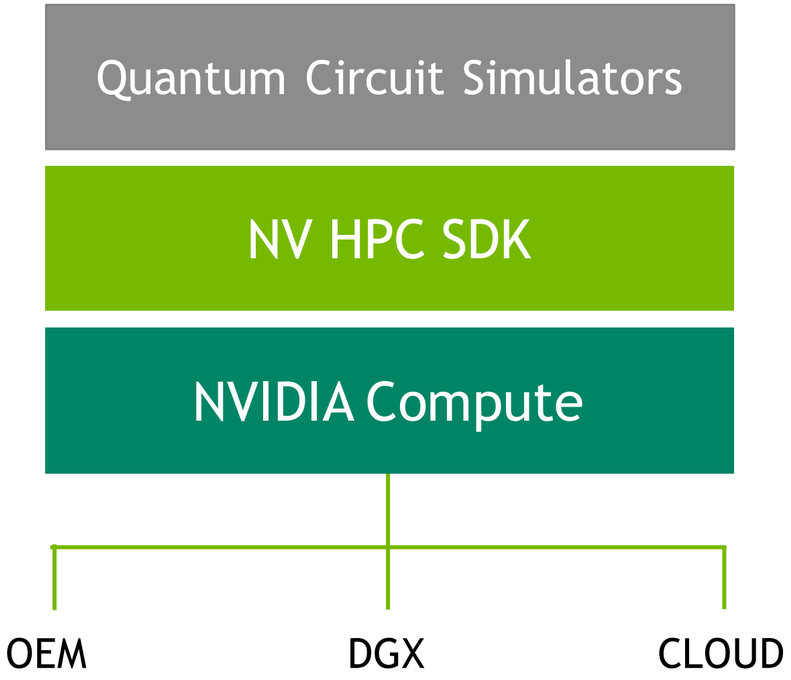
Optimization

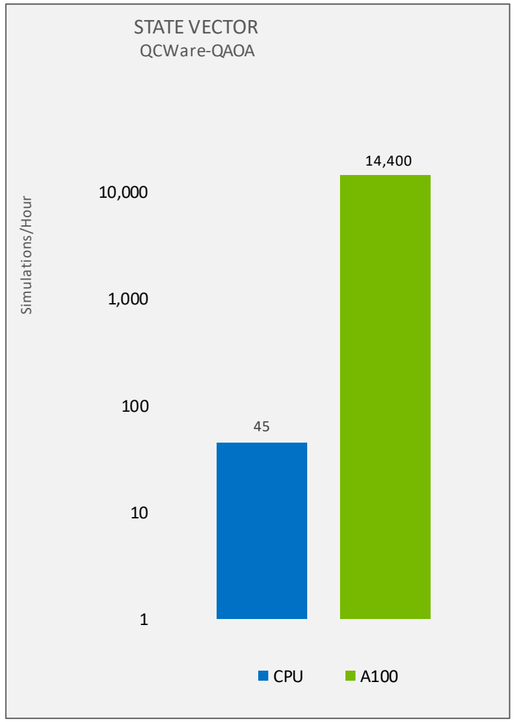
## REQUIRES QUBITS SCALE TO DOUBLE EVERY YEAR

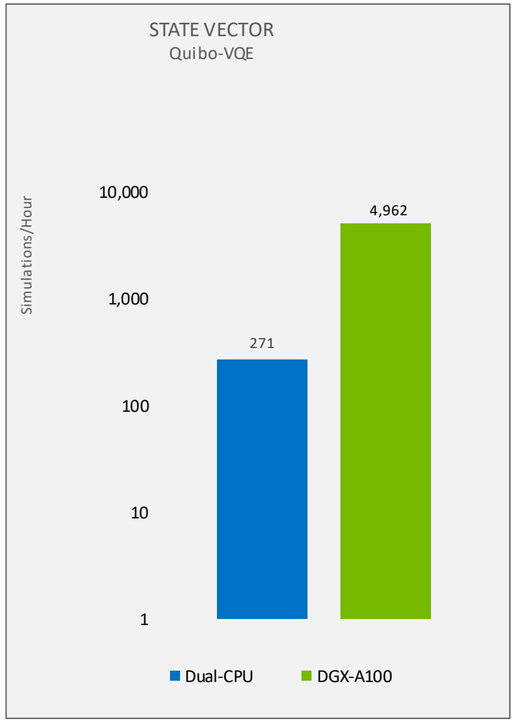


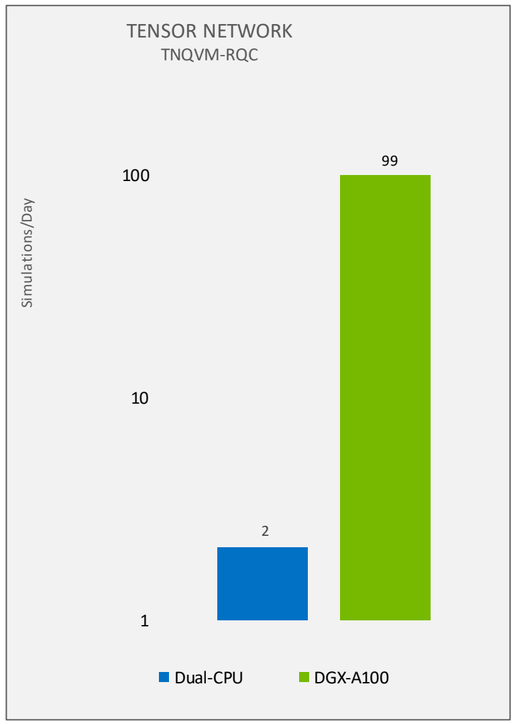
# ACCELERATE QUANTUM CIRCUIT SIMULATIONS TODAY

## With NVIDIA HPC SDK




Footnotes: State Vector, QAOA (complex FP32) - CPU: QC Ware Quasar code on Intel Xeon Platinum 8275CL @ 3.0 GHZ, GPU: QC Ware Vulcan code on 1x NVIDIA A100 SMX4 40GB, based on CUDA kernels | State Vector, VQE (complex FP32) 30 qubits depth m=10, CPU: Quibo on Dual AMD EPYC 7742, GPU: Quibo on DGX-A100 | Tensor Network - 53 qubits, depth m=14, CPU: Estimated TNQVM (ORNL) on Dual AMD EPYC 7742 , GPU: TNQVM (ORNL) on DGX-A100

# NVIDIA CUQUANTUM

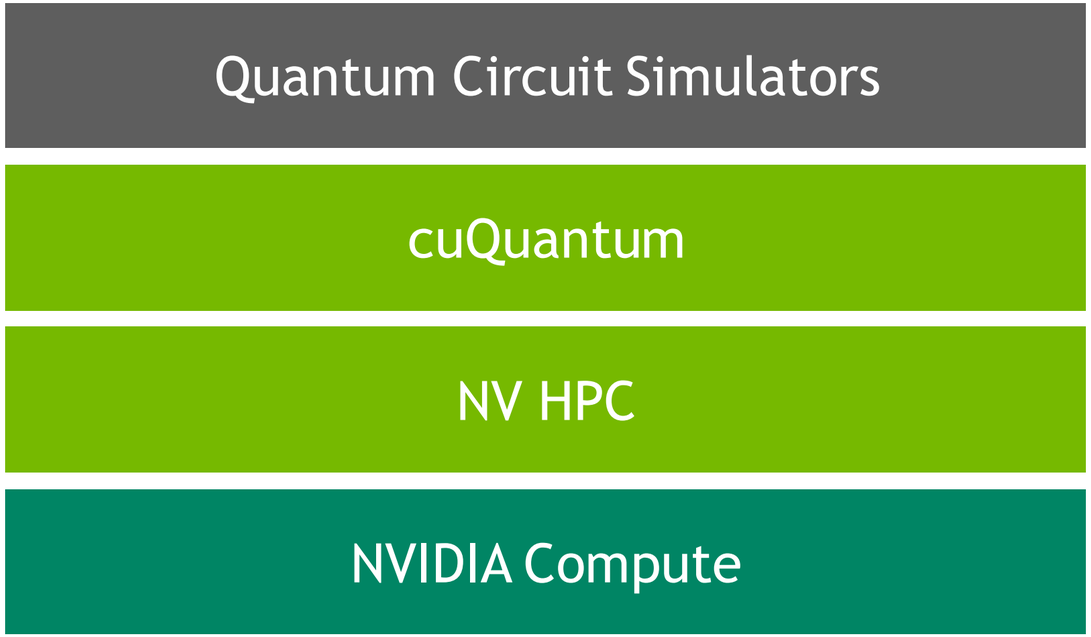
## SDK FOR GPU-ACCELERATED QUANTUM SIMULATIONS

Dramatically Accelerates Quantum Circuit Simulations

Targeting Tensor Network and State Vector Methods

Increased Control and Flexibility with C++ APIs

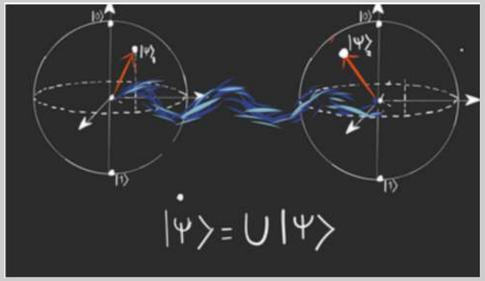
Drop-in Multi-node Execution with Python API



# ANNOUNCING NVIDIA CUQUANTUM

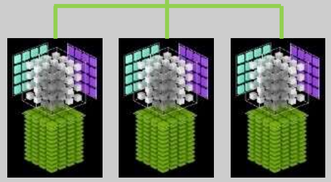
## Researching the Computer of Tomorrow on the Most Powerful Computer Today

### SDK FOR GPU-ACCELERATED QUANTUM SIMULATIONS

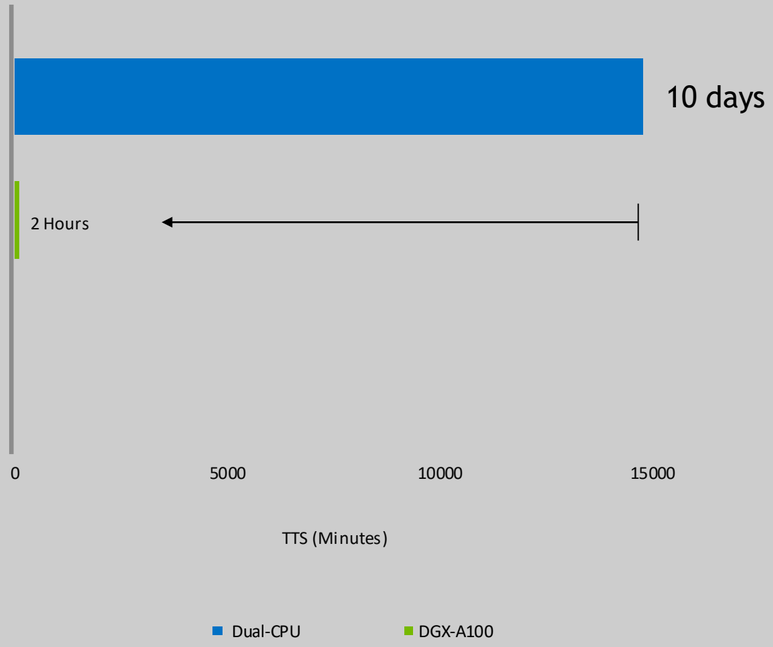



Quantum Circuit Simulation Frameworks

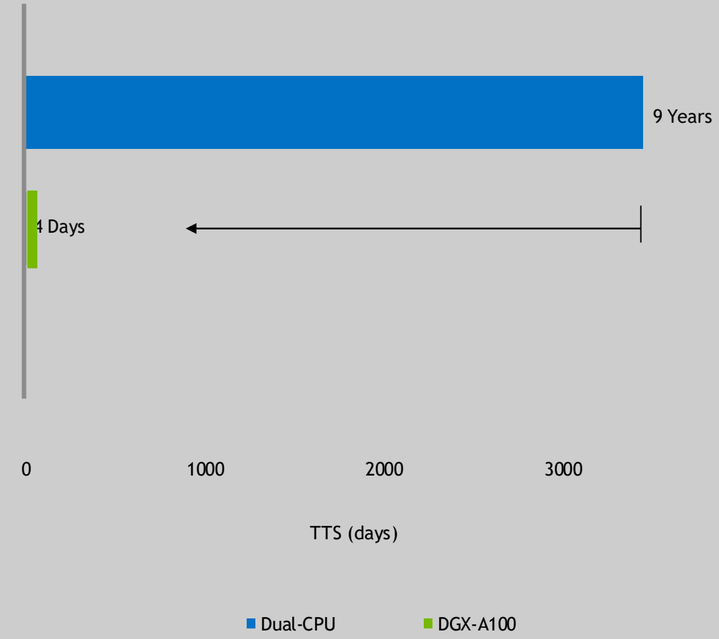
**cuQuantum**



### STATE VECTOR SIMULATION
Scales to 10s of Qubits



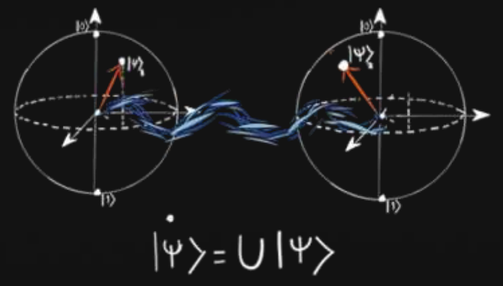

### TENSOR NETWORK SIMULATION
Scales to 1000's of Qubits




Footnotes: State Vector- 1,000 circuits , 36 qubits depth m=10, complex 64 | CPU: Qiskit (IBM) on Dual AMD EPYC 7742 | GPU: Qgate (NVAITC) on DGX-A100
Tensor Network - 53 qubits, depth m=20 | CPU: Estimated Quimb (Caltech) on Dual AMD EPYC 7742 | GPU: Quimb (Caltech) on DGX-A100

# QUANTUM COMPUTING — THE NEXT FRONTIER

Research the Computer of Tomorrow on the Most Powerful Computer Today




cuQuantum

Alibaba.com · amazon · Argonne NATIONAL LABORATORY · Atos · BERKELEY LAB

blueqat · Caltech · CQC · DELL · FLATIRON INSTITUTE

FUJITSU · Google · Honeywell · IBM · IONQ

JÜLICH FORSCHUNGSZENTRUM · Los Alamos NATIONAL LABORATORY EST. 1943 · Microsoft · NASA · NUS National University of Singapore

OAK RIDGE National Laboratory · Pacific Northwest NATIONAL LABORATORY · PsiQuantum · QCWARE · rigetti

STRANGE WORKS · XANADU · ZAPATA


MORPHEUS
CIBERSECURITY

# CYBERSECURITY IS A DATA PROBLEM

Streaming search, multi-environment integration, and quick iterations

**High Velocity Data Streams**

**Heterogeneous Data**

**Integration Cross Environment**

**Quick Iterations**

**Traditional methods and tools are not fast enough and rely too heavily on heuristics**

**197 days**
to identify a breach

**69 days**
to contain a breach

# CURRENT METHODS CANNOT NOT SCALE

Methodologies must change in order to address these issues

## Shortage of Professionals

Currently a worldwide shortage of over 2.93 million cybersecurity professionals, and 53% of companies report a problematic shortage.

Source, Source

## Number of Breaches


The current lack of staffing exacerbates the number of breaches occurring.

Source

## Log Collection Outpacing Analysis

The amount and quantity of logs collected on and devices added to networks continues to increase.

Source

## Challenging Vendor Integration

Nearly half of security risks stem from staff being inadequately experienced on products or difficult to use systems.

Source

# MORPHEUS AI CYBERSECURITY FRAMEWORK

PERCEIVE

INFER

ACT

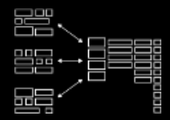
BLUEFIELD 2

AI Apps

NetQ

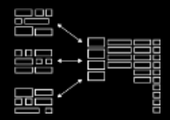
DOCA

DATA PREP



INFERENCE

POST PROCESS



VERIFY

FEEDBACK

NVIDIA MORPHEUS

Pub/Sub

RAPIDS

CLX

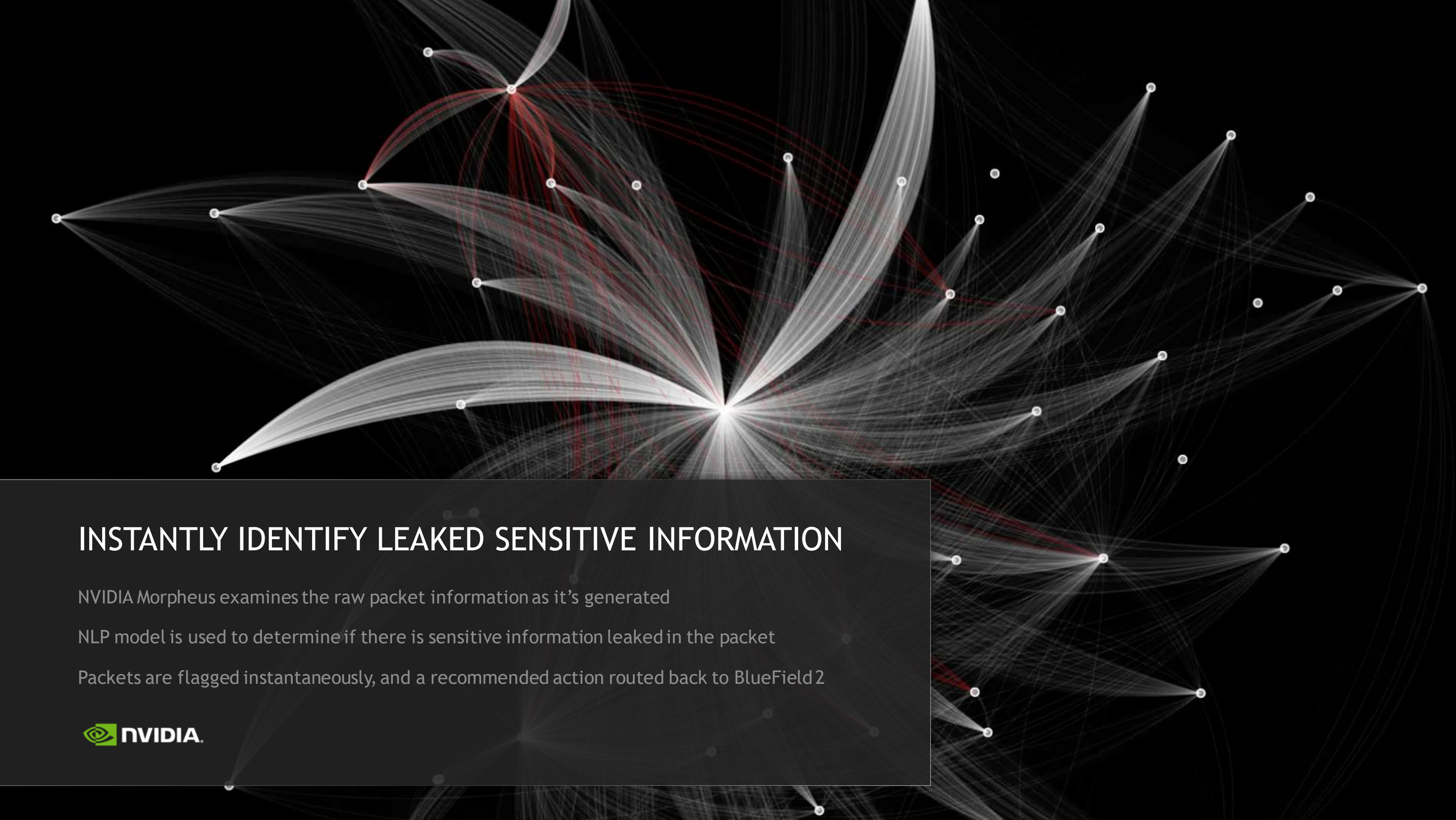
TRITON

TensorRT

cuStreamz

MLFlow

EGX STACK



# INSTANTLY IDENTIFY LEAKED SENSITIVE INFORMATION

NVIDIA Morpheus examines the raw packet information as it’s generated

NLP model is used to determine if there is sensitive information leaked in the packet

Packets are flagged instantaneously, and a recommended action routed back to BlueField 2

NVIDIA


NVIDIA GPU CLOUD
CONTAINERS + MODELS

# NGC – GPU-OPTIMIZED SOFTWARE

## Build AI Faster, Deploy Anywhere

**NGC**

CONTAINERS

150+

TRAINED MODELS

100+

HELM CHARTS

ML, Inference

AI TOOLKITS & SDKs

Healthcare | Smart Cities | Conversational AI | Robotics | more

ENCRYPTED

x86 | ARM | POWER

MULTI-CLOUD

ON-PREM

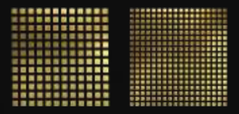
HYBRID CLOUD

EDGE

# GPU (ACCELERATORS)

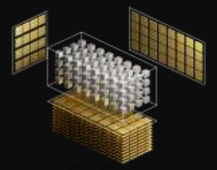
# ANNOUNCING NVIDIA A100 PCIE 80GB

## Supercharging The World's Highest Performing AI Supercomputing GPU

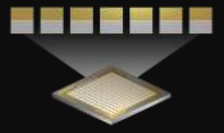


80GB HBM2e
For largest datasets and models

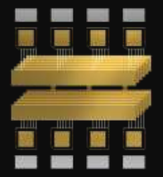
2TB/s
World's highest memory bandwidth
to feed the world's fastest GPU



3rd Gen Tensor Core



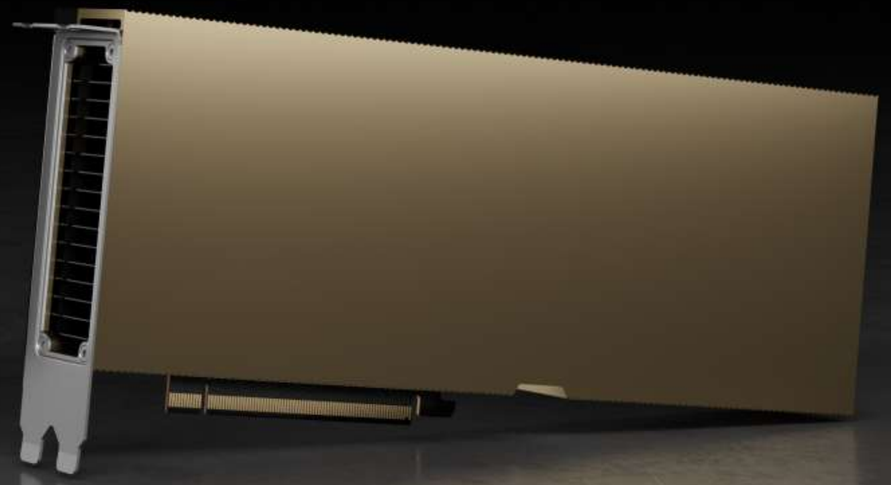
Multi-Instance GPU



3rd Gen NVLink

Atos · cisco · DELL Technologies · FUJITSU · GIGABYTE · H3C · Hewlett Packard Enterprise

inspur · Lenovo · Micron · ORACLE CLOUD · PENGUIN COMPUTING · QCT · SUPERMICRO



# A100 80GB AVAILABLE VIA NVIDIA A100 SXM AND A100 PCIE

## A100 PCIe

For Mainstream Servers

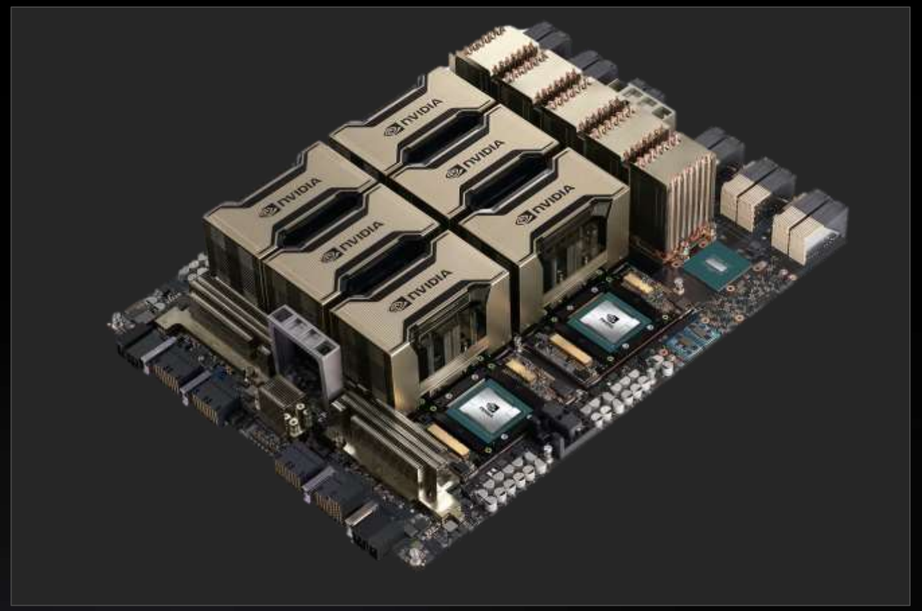
1-8 GPUs per server, optional NVLink Bridge between 2 GPUs

## HGX A100 4-GPU


Scale-Up – Mixed AI & HPC

4 A100s, Fully Connected w/ shared NVLinks

## HGX A100 8-GPU




Scale-up - Fastest Time-to-solution for AI

8 GPUs, Full NVLink B/W between all GPUs with NVSwitch

# ARM ECOSYSTEM

# DIVERSITY OF PROBLEMS REQUIRES ARCHITECTURAL FLEXIBILITY

LARGE SIMULATIONS


TRAINING AI

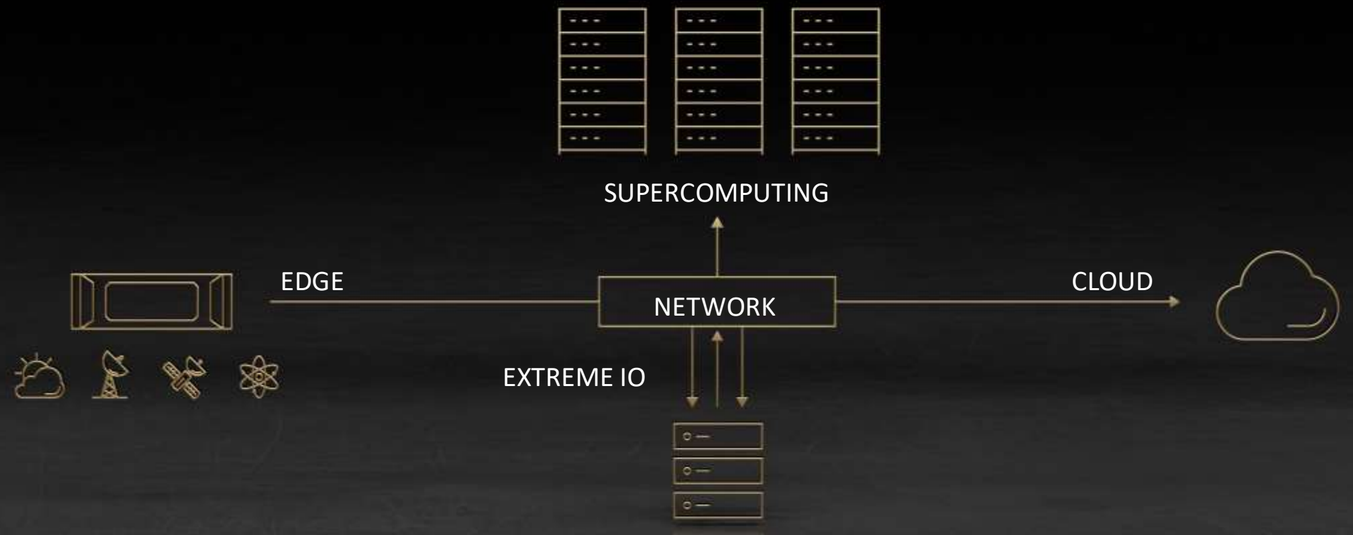

REAL-TIME PROCESSING


3D RECONSTRUCTION


DIGITAL TWIN

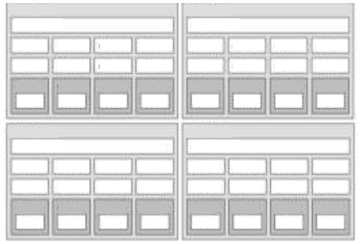

VISUALIZATION


HIGH-THROUGHPUT CLOUD

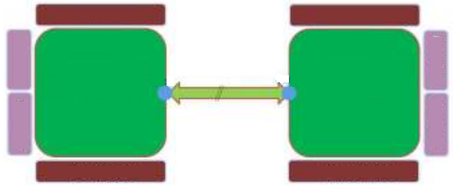



# ARM ENABLES ARCHITECTURAL INNOVATION

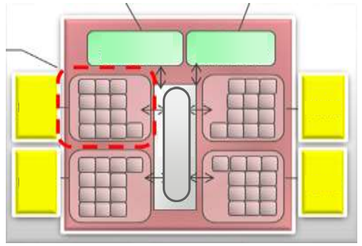
AMAZON GRAVITON
Cloud



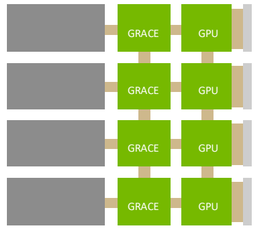
C-DAC
Supercomputing



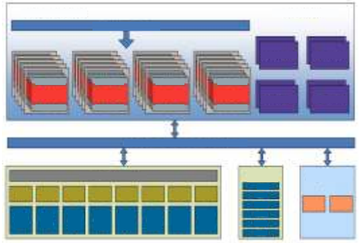
FUJITSU
Supercomputing



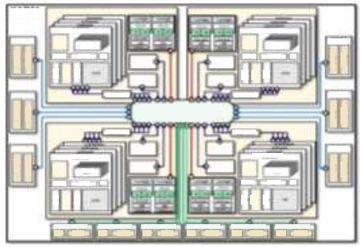
NVIDIA GRACE
Giant AI and HPC



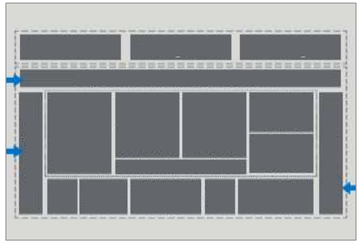

AMPERE COMPUTING
Hyperscale



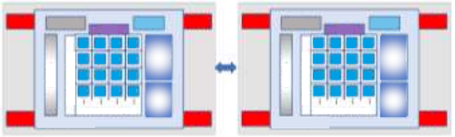
ETRI
Supercomputing



MARVELL
Storage and Edge

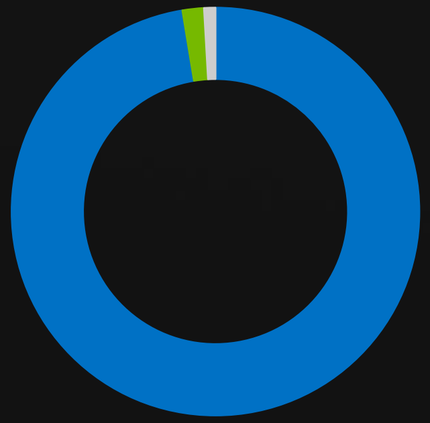


SIPEARL
Supercomputing

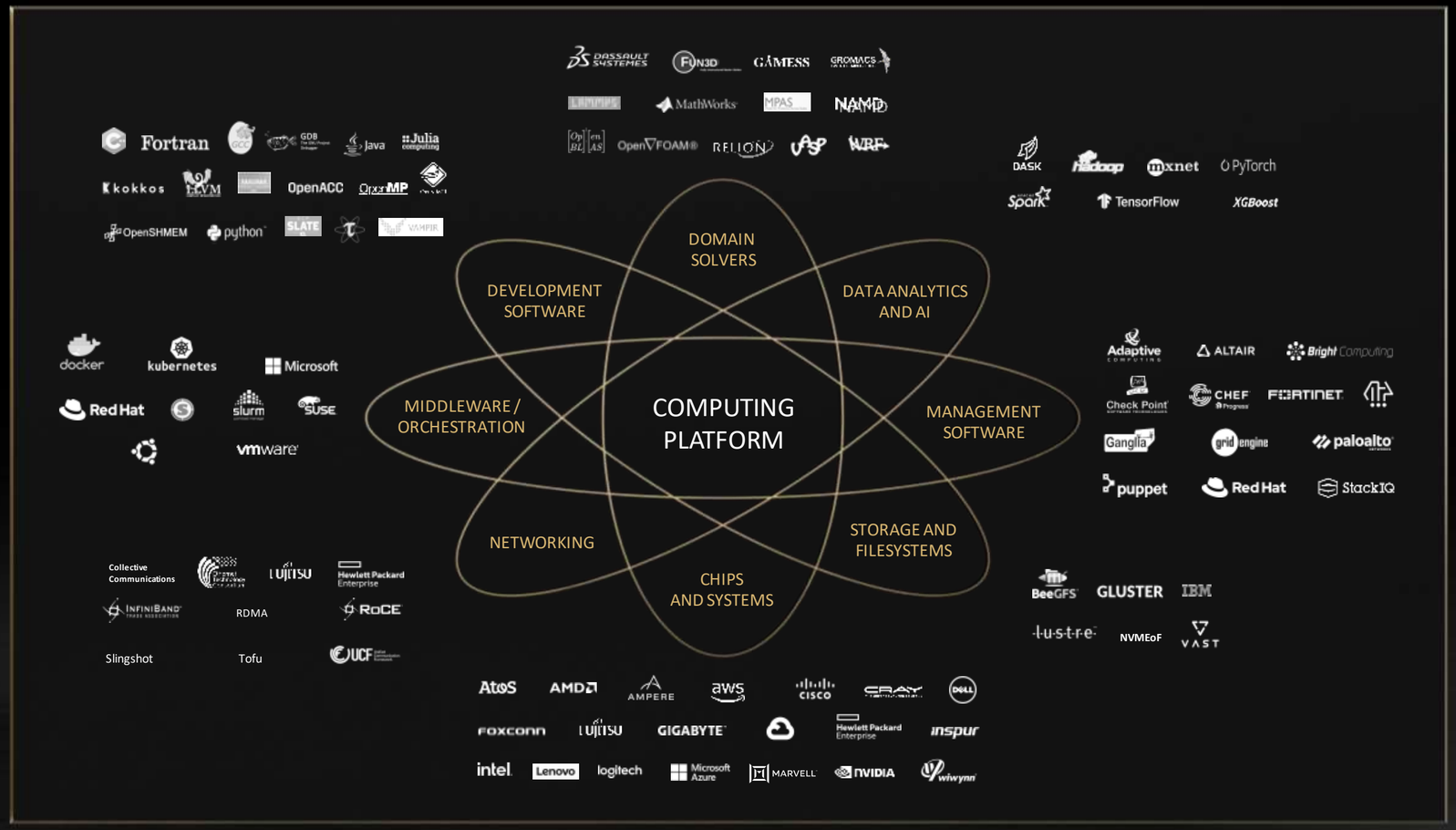


# CREATING A COMPUTING ECOSYSTEM IS HARD

SERVER SHIPMENTS BY CPU ARCHITECTURE
Cloud, Supercomputing, Enterprise Datacenters



| X86 | ARM | OTHER |
|---|---|---|
| 97.4% | 1.7% | 0.9% |




Source: IDC, Worldwide Server Microprocessor Vendor, Architecture, Channel, Socket, and System Category Market Shares, 4Q20, Doc # US47110421, February 2021

# NVIDIA ACCELERATED COMPUTING ECOSYSTEM

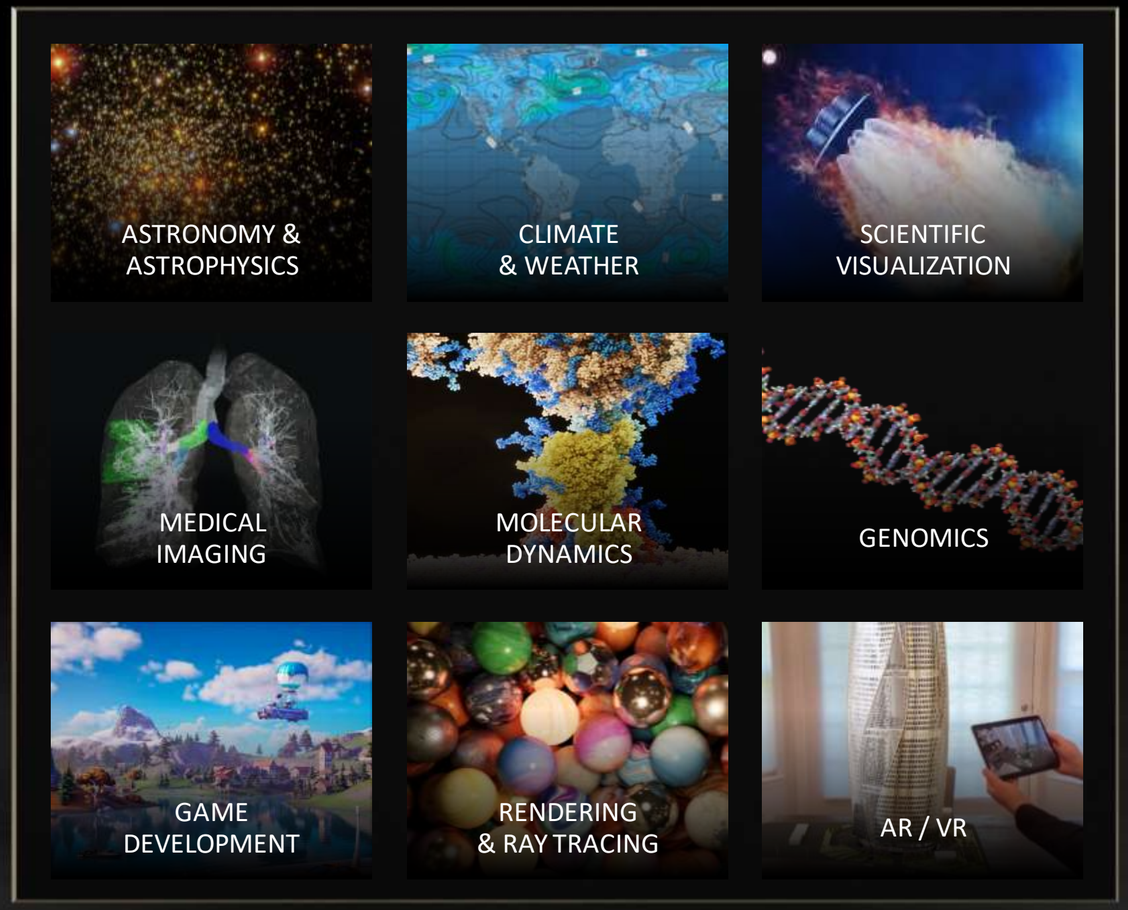
150
SDKs

24M
CUDA Downloads

2.5M
Developers

All Major OEMS

All Major CSPs



# NVIDIA GRACE

CPU Designed for Giant-Scale AI and HPC Accelerated Computing

## 10X More AI Performance
Trillion Parameter NLP Models

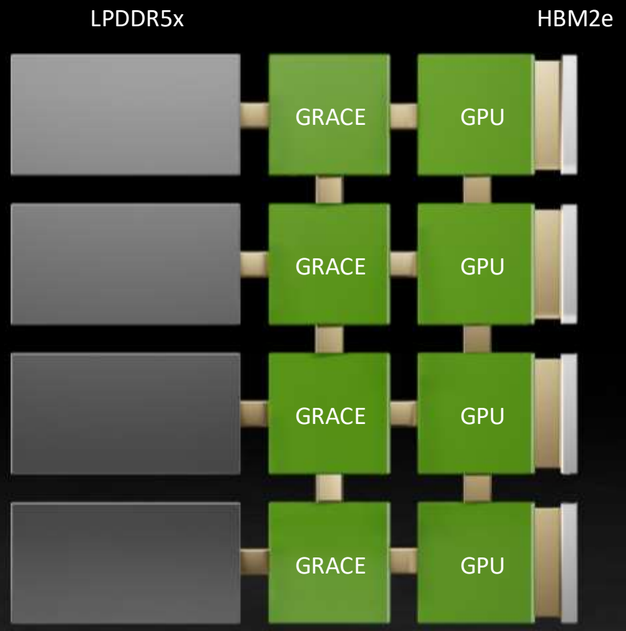
1 MONTH ⟶ 3 DAYS

Fine-Tune Training of 1T Model

REAL-TIME INFERENCE
ON 0.5T MODEL

Interactive Single Node NLP Inference

## NEW COMPUTING ARCHITECTURE
Giant Scale AI and HPC Computing




## 7X More HPC Performance
Giant 3D-FFT

1 WEEK ⟶ 1 DAY

Seismic Data Processing

MAKE NEW DISCOVERIES USING HIGHER RESOLUTION

PROCESS MORE DATA
IN THE SAME TIME

Bandwidth claims rounded to nearest hundred for illustration.
Performance results based on projections on these configurations Grace : 8xGrace and 8xA100 with 4th Gen NVIDIA NVLink Connection between CPU and GPU and x86: DGX A100.
Training: 1 Month of training is Fine-Tuning a 1T parameter model on a large custom data set on 64xGrace+64xA100 compared to 8xDGXA100 (16xX86+64xA100)
Inference: 530B Parameter model on 8xGrace+8xA100 compared to DGXA100.
3D FFT: A 6500 x 6500 x 6500 element 3D FFT on x86 with 8x A100 compared to a Grace server would run 7x faster (thus 1 calendar week to 1 day). This performance estimate is based on standard a profile of seismic data processing applications (e.g. Schlumberger RTM, FSME, Emerson RTM) used in Geo-sciences.

# WORLD’S FASTEST EXASCALE AI SUPERCOMPUTER

20 Exaflops AI | NVIDIA Grace CPU and NVIDIA GPU | HPC and AI

# ACCELERATING ARM ECOSYSTEM ADOPTION

## Application Development Tools and Ready-to-Run Applications

ANALYSIS

Profilers | Debugger

SIMULATION AND VISUALIZATION

HPC Application Containers from NGC

DEVELOPMENT

Compilers | Programming Models | Core Libraries | Math Libraries | Communication Libraries

NVIDIA CUDA, MAGNUM IO, DOCA


AMPERE. Atos सी डैक CDAC GIGABYTE SIPEARL The Silicon Pearl

"The NVIDIA Arm Developer Kits will facilitate the transition of our codes to NVIDIA GPUs and Arm CPUs."

Steve Poole
Chief Architect for Next-Generation Systems
Los Alamos National Laboratory

# THE NVIDIA HPC SDK IS READY FOR ARM

## Auto Vectorization

## Vector Intrinsics

```
float32x2_t vadd_f32 (float32x2_t a, float32x2_t b);
float32x4_t vdupq_n_f32 (float32_t value);
float32x2_t vget_high_f32 (float32x4_t a);
```

## Host Parallelism

ISO Standard Parallelism

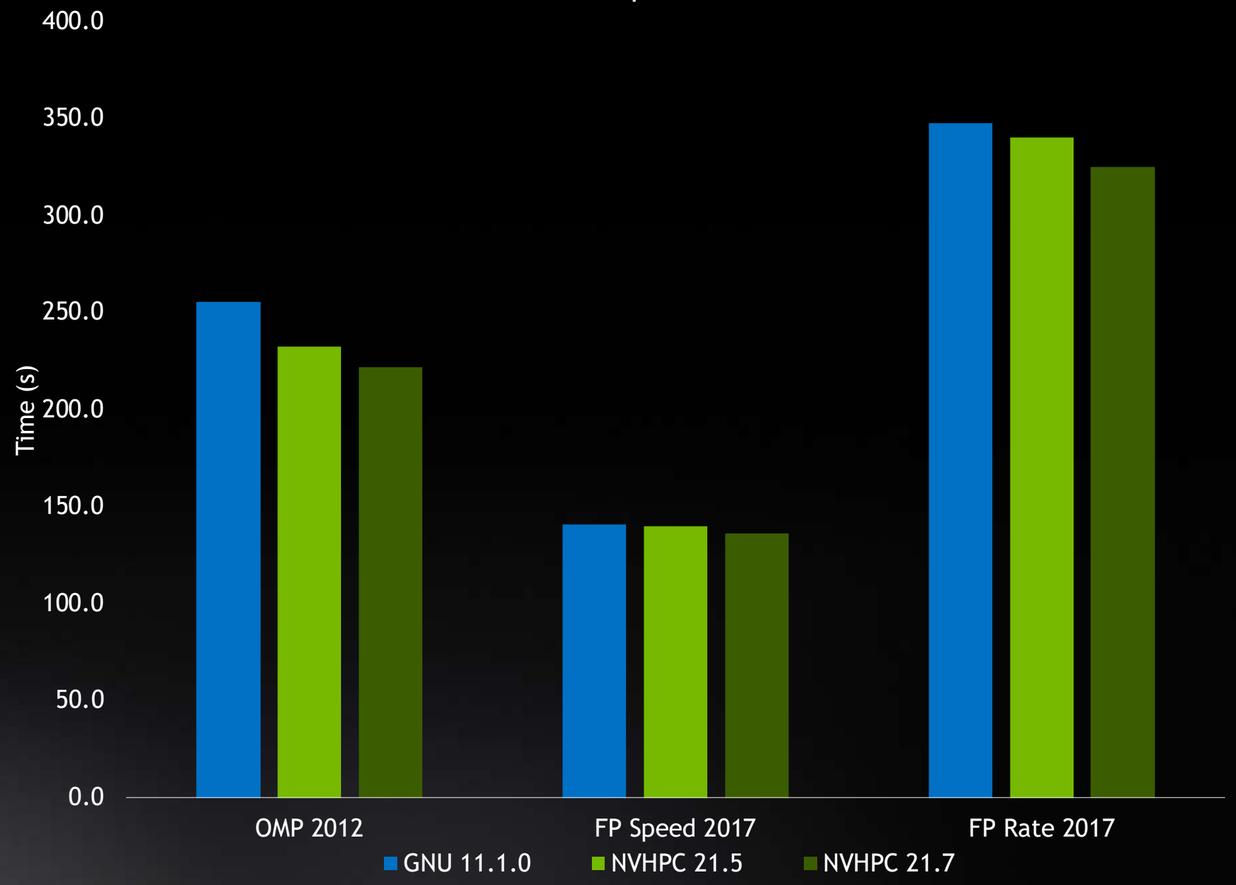
OpenACC

OpenMP

## FP Models

- Precise
- Fast
- Relaxed

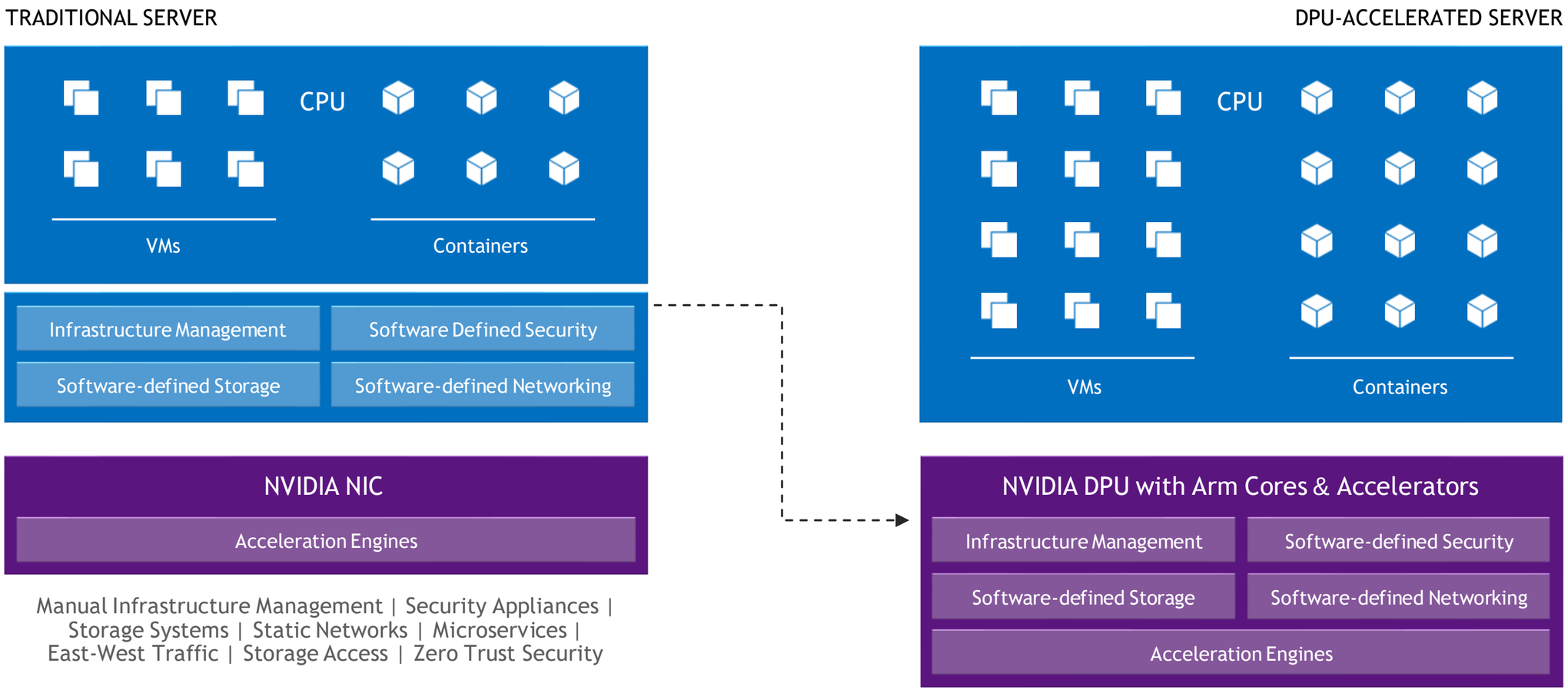
Estimated SPEC OMP® 2012, SPEC CPU® 2017 Floating Point Speed and Rate Geomean on 80 core Ampere Altra – Lower is Better



# DPU
# (DATA PROCESSING UNIT)

# INTRODUCING THE DATA PROCESSING UNIT (DPU)

## Software-defined Data Center Infrastructure-on-a-Chip



# NVIDIA BLUEFIELD-2

## Data Center Infrastructure on a Chip

Up to 200Gb/s Ethernet and InfiniBand, PAM4/NRZ

CX-6 Dx Inside

8 ARM A72 CPUs subsystem – over 2.5GHz

- 8MB L2 cache, 6MB L3 cache in 4 Tiles
- Fully coherent low-latency interconnect

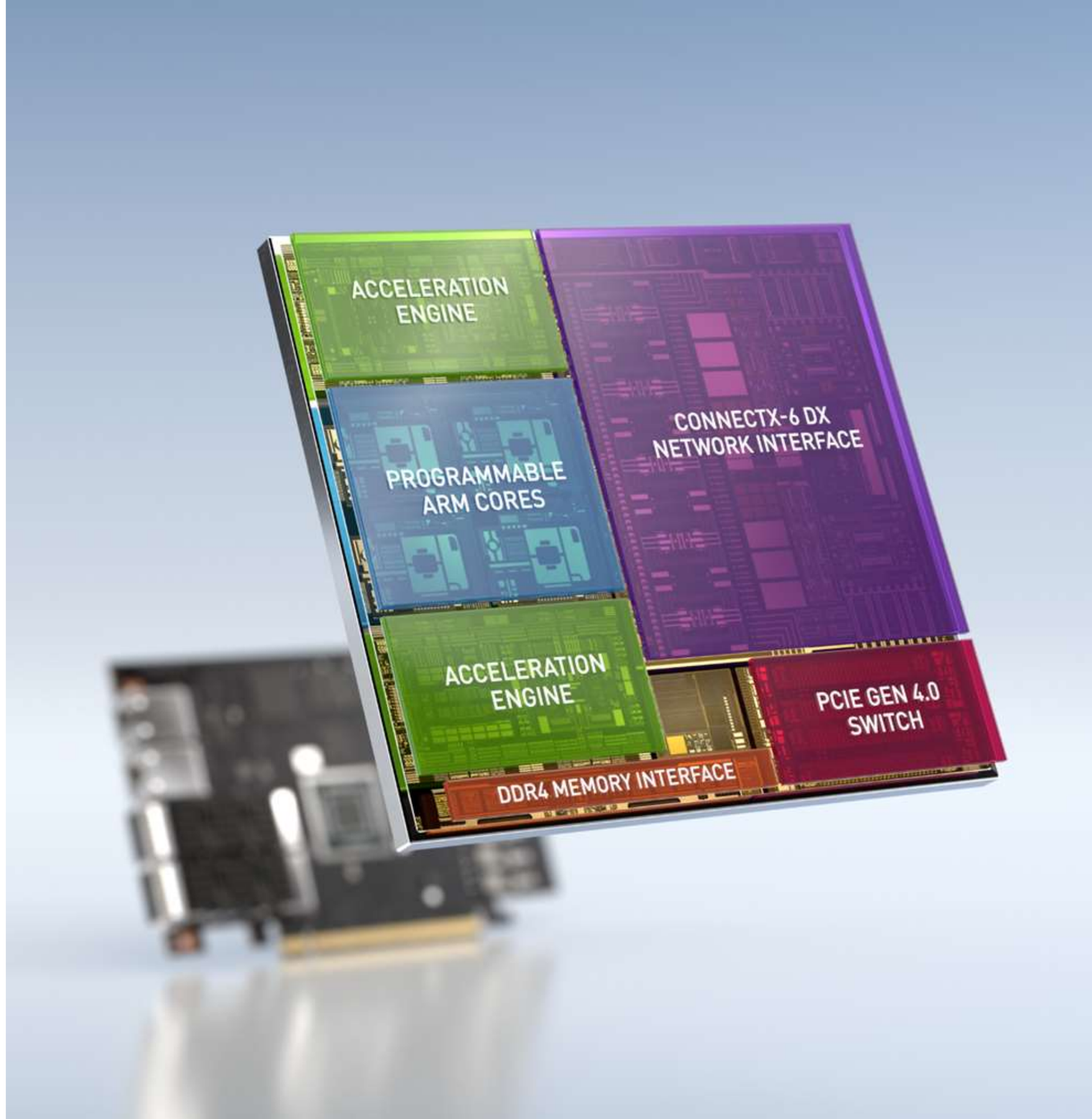
Integrated PCIe switch, 16x Gen4.0

- PCIe Root Complex or End Point modes

Single DDR4 Channel

1GbE Out-of-Band management port

Accelerated Security, Storage, Networking



# NVIDIA DOCA

## Enabling Broad DPU Partner Ecosystem

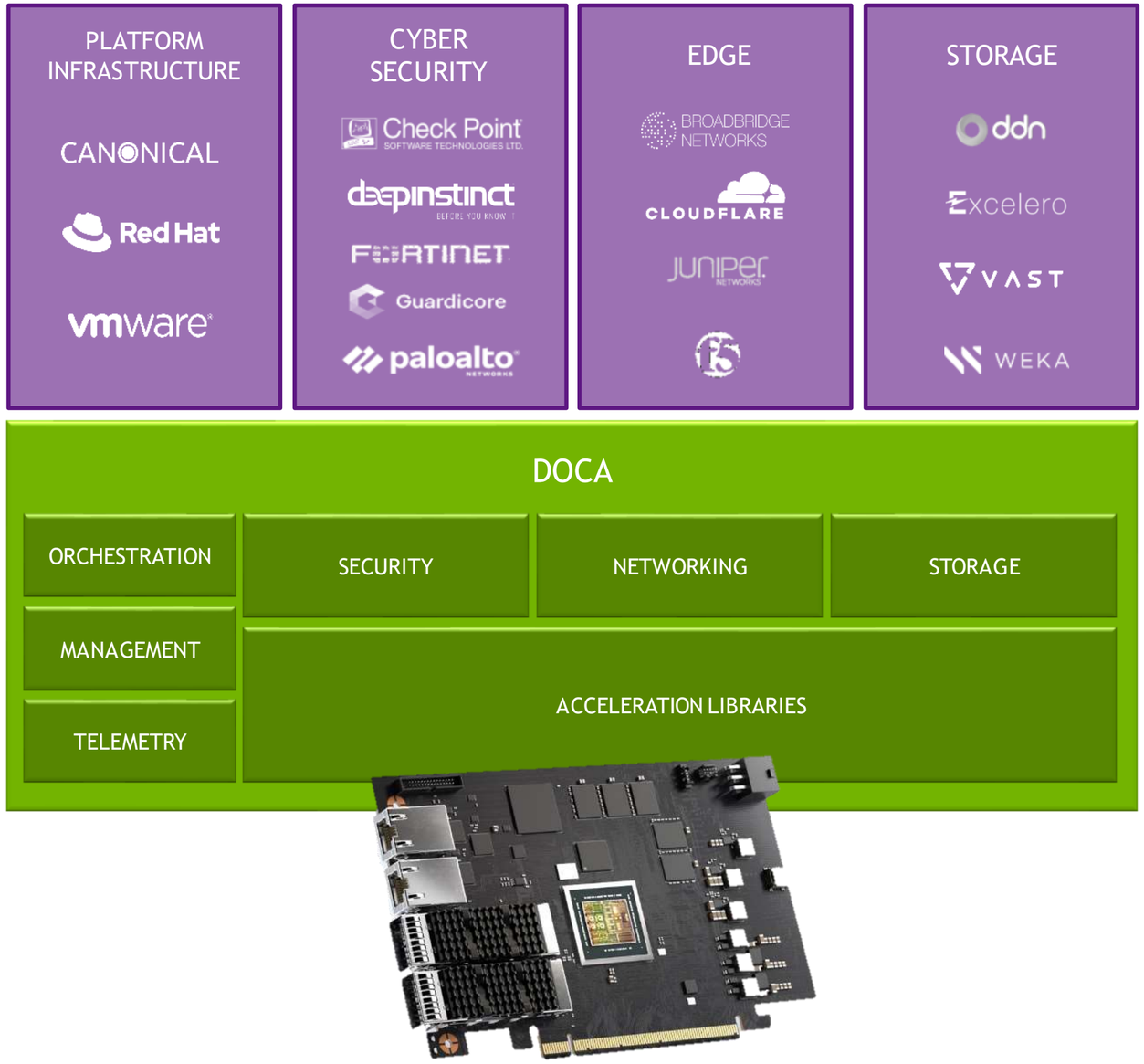
Software application framework for BlueField DPUs

DOCA is for DPUs what CUDA is for GPUs

Protects developer investment for future DPUs

Certified reference applications, APIs & partner solutions

Rich partner ecosystem across industries and workloads



ANNOUNCING

# GPUDIRECT STORAGE

3X Lower CPU Utilization

6X DL Inference

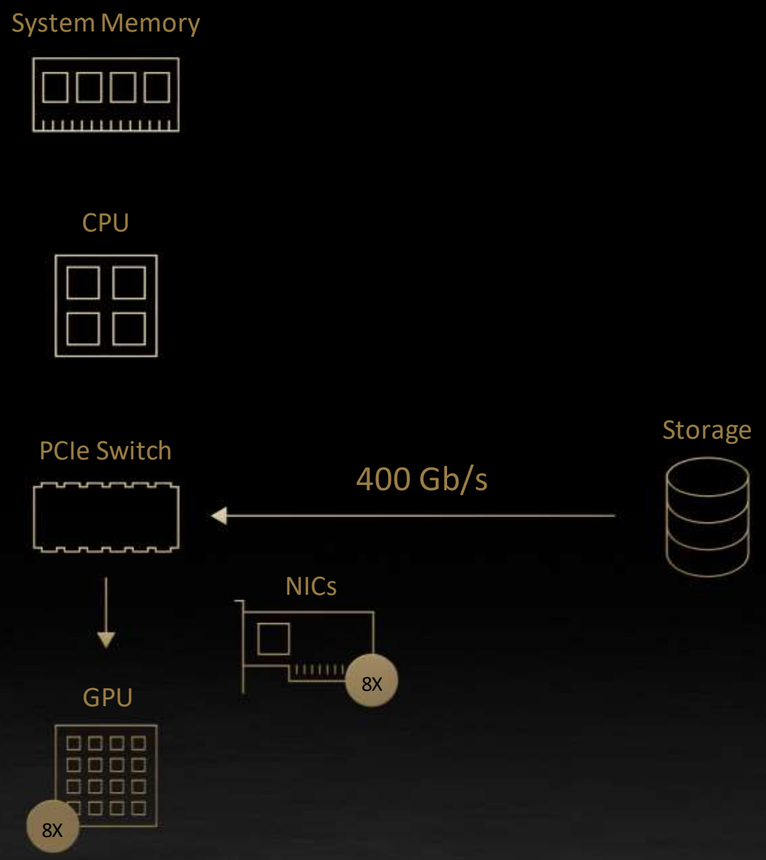
2X Seismic

AVAILABLE NOW

DDN AI · BIG DATA · HPC  DELL EMC  Excelero  IBM

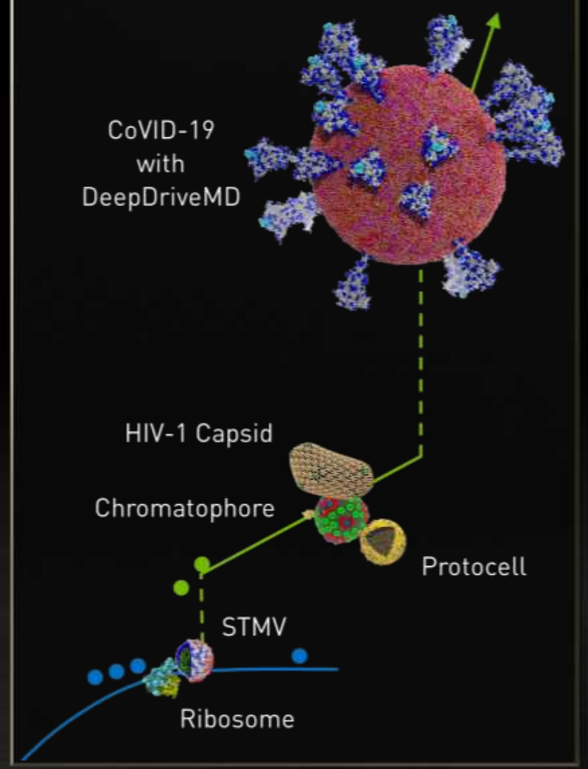
Micron  NetApp  PAVILION

ScaleFlux  VAST  WEKA.IO

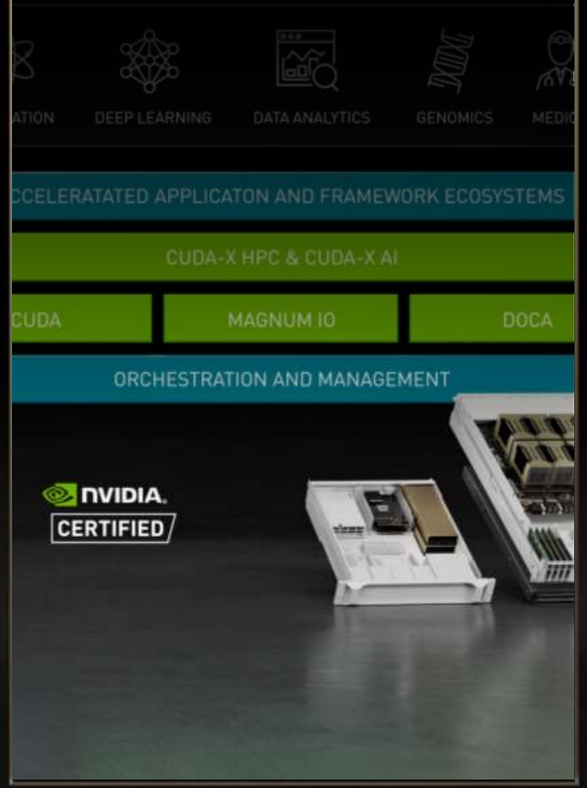


# EXPANDING UNIVERSE OF HPC

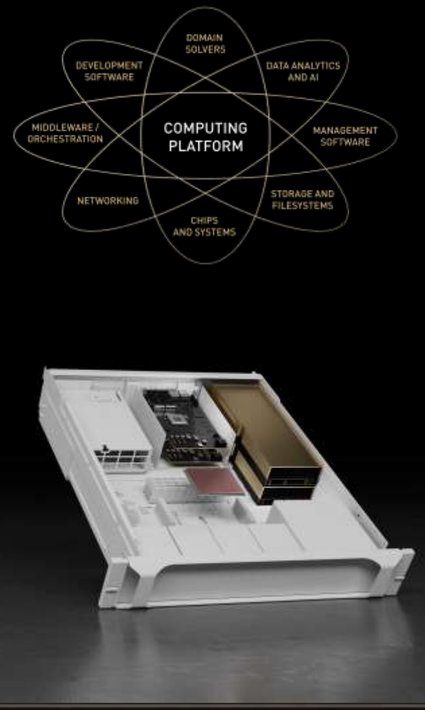
## TIPPING POINT FOR INDUSTRIAL HPC



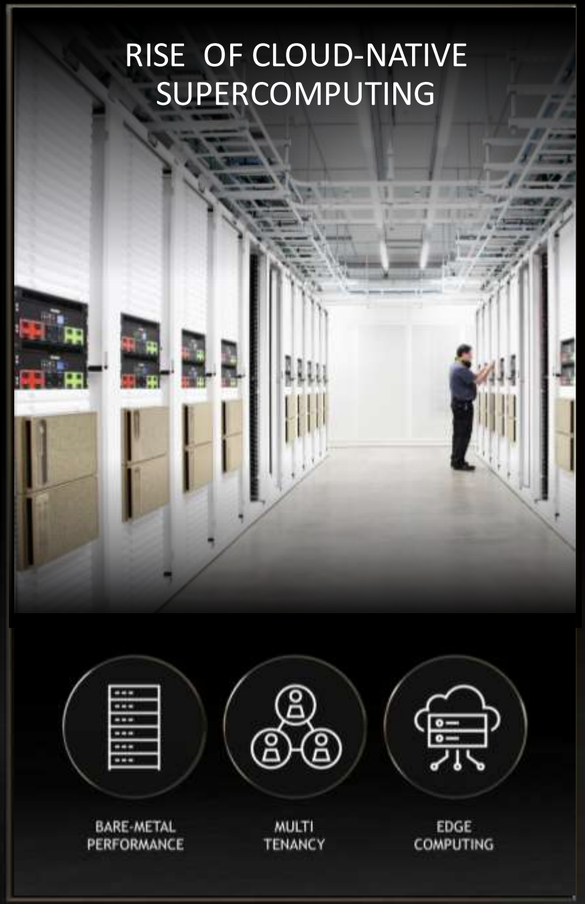

## NVIDIA HGX PLATFORM POWERING HPC & AI

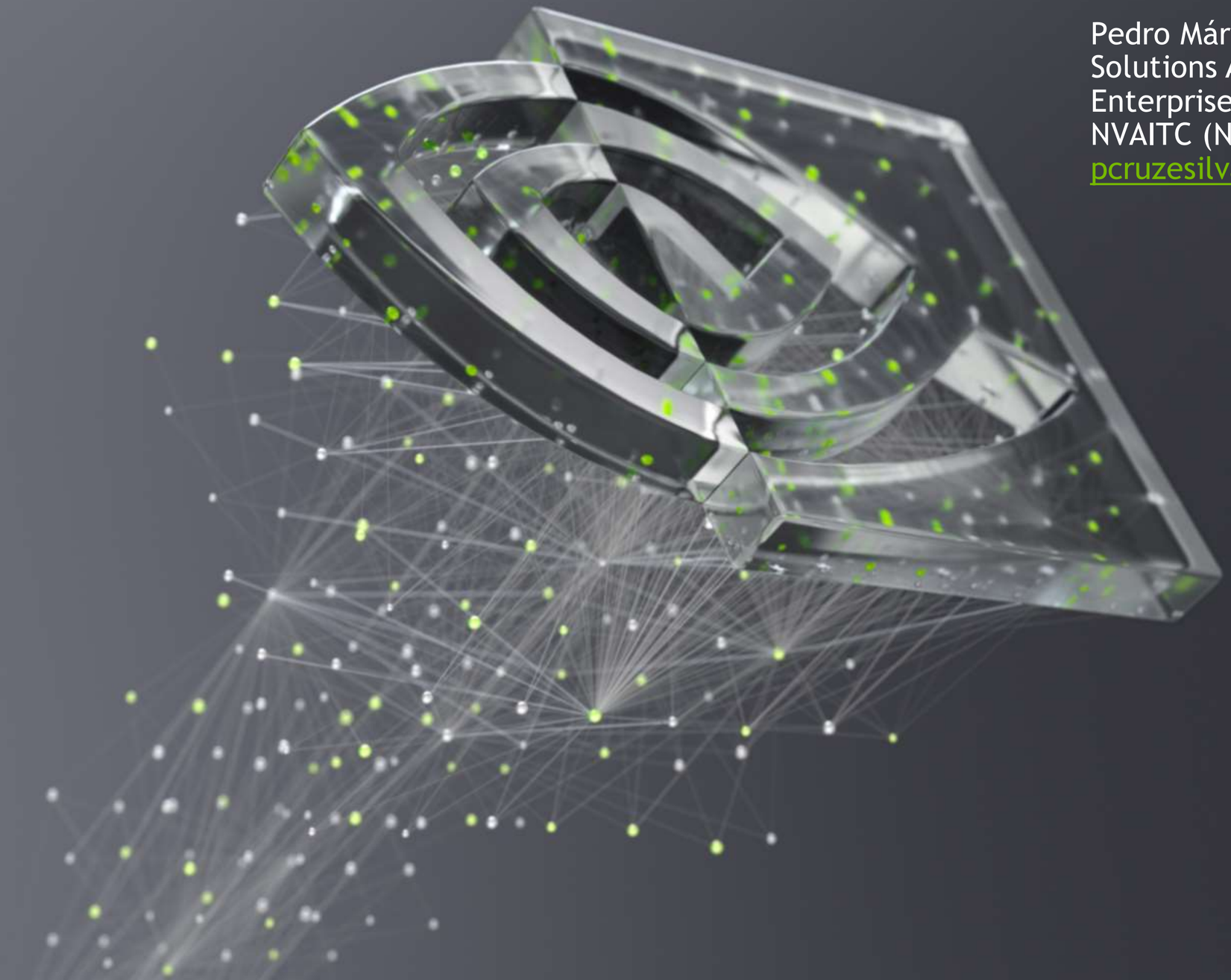



## EXPANDING ARM HPC ECOSYSTEM




## RISE OF CLOUD-NATIVE SUPERCOMPUTING





Pedro Mário Cruz e Silva
Solutions Architect Manager
Enterprise Latin America
NVAITC (NVIDIA AI Technology Center)
pcruzesilva@nvidia.com

# Obrigado
# Gracias
# Thank you

NVIDIA