INTELIGENCIA ARTIFICIAL DESCOMPLICADA

Marcel Saraiva
msaraiva@nvidia.com

João Paulo Navarro
jpnavarro@nvidia.com

NVIDIA —
A LEARNING MACHINE

NVIDIA tem continuamente se reinventado nas ultimas duas décadas.

Com as nossa invenção da GPU em 1999 acelreramos o Mercado de game em PCs, redefinimos a computação gráfica moderna e revolucionamos a computação paralela. Mais recentemente a computção em GPU habilitou as era da Intelligencia Artificial.

NVIDIA é uma “maquina de aprendizado” que constante envolviennto se adapta a novas oportunidades que são dificies de resolver, que somente nós conseguimos enfrentar e que importa para o mundo.

GRAPHICS

GPU COMPUTING

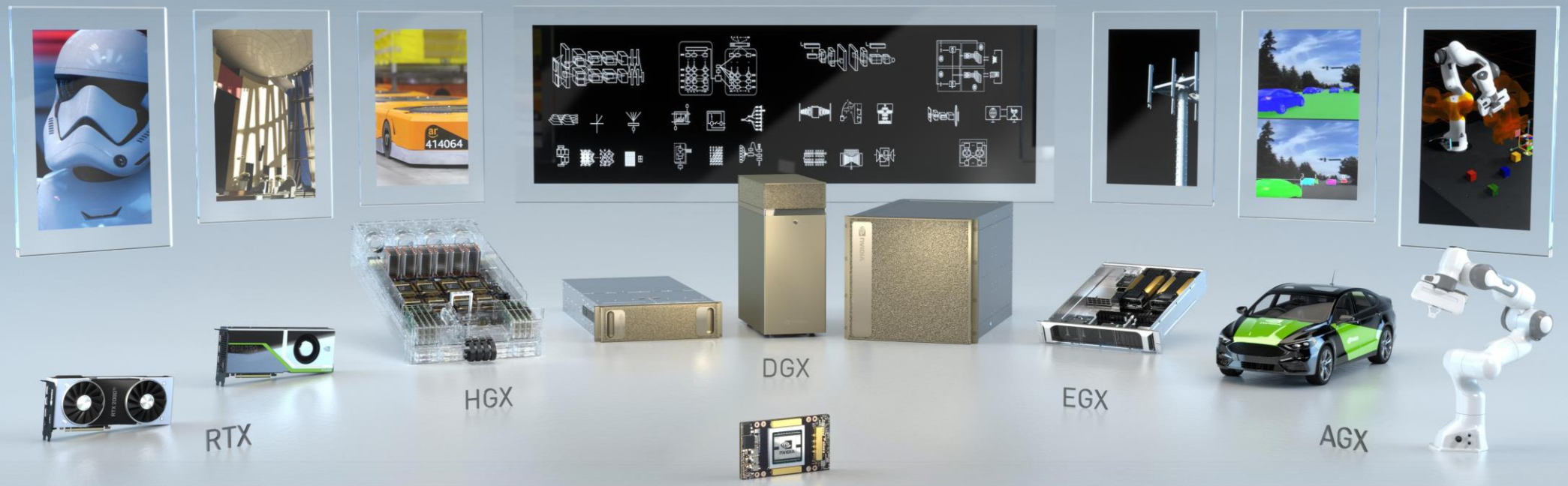
AI

“It has been a long road from inventing the GPU to accelerate gaming to reinventing the GPU to be the most diverse and powerful coprocessor we have ever seen.”

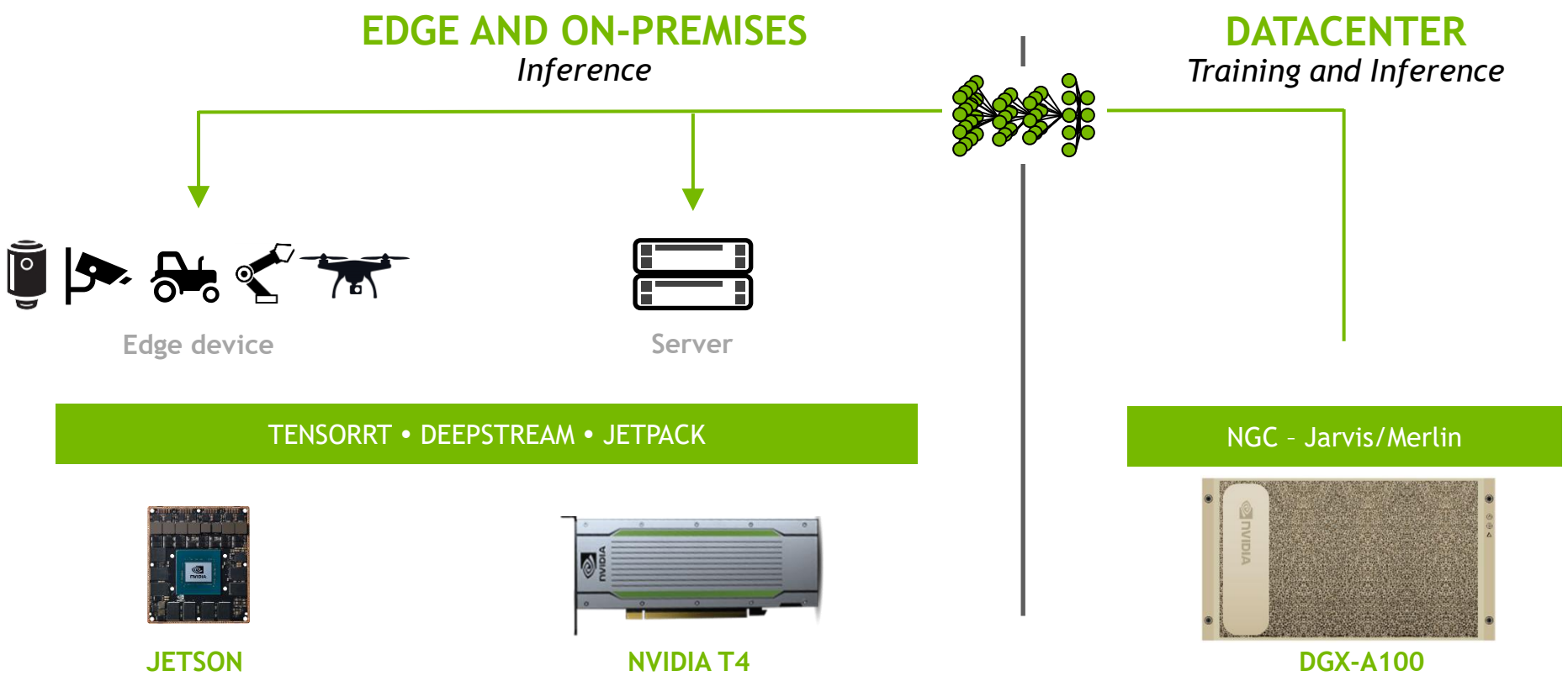
THE NEXT PLATFORM

NVIDIA pioneered accelerated computing to tackle challenges ordinary computers cannot. We make computers for the da Vincis and Einsteins of our time so that they can see and create the future.

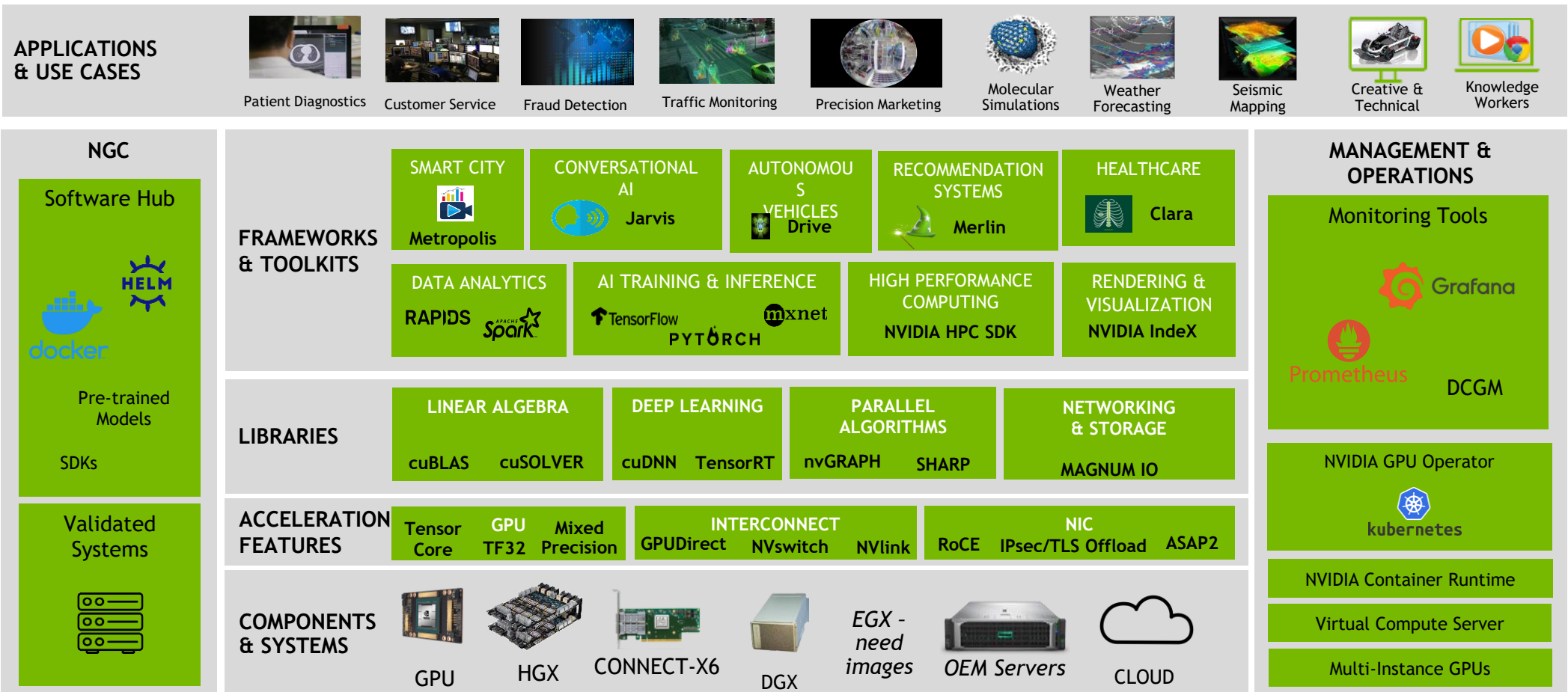
Accelerated computing requires more than just a powerful chip. We achieve incredible speedups through full-stack invention—from the chip and systems to the algorithms and applications they run.



END TO END AI


EDGE AND ON-PREMISES
Inference
DATACENTER
Training and Inference
Edge device
Server
TENSORRT • DEEPSTREAM • JETPACK
NGC – Jarvis/Merlin
JETSON
NVIDIA T4
DGX-A100

NVIDIA ACCELERATED COMPUTING PLATFORM



MULTIPLE ON-RAMPS TO ACCELERATED AI

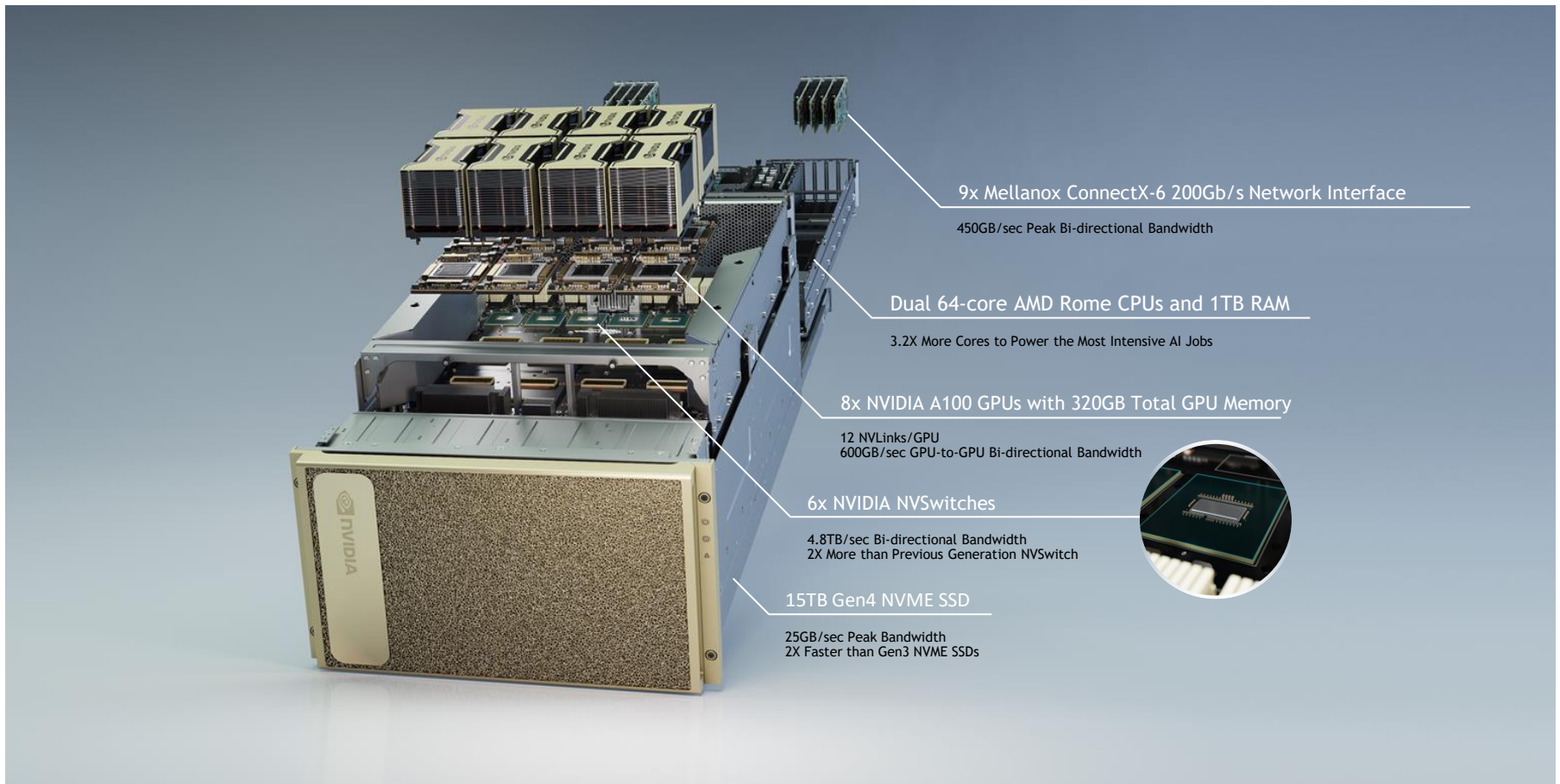
DGX Servers and
Super-POD Clusters

GPU-Accelerated Servers from
Your OEM of choice

aws
Alibaba Cloud
Azure
BaiduCloud
Google Cloud
ORACLE CLOUD
Tencent Cloud

Cloud-based AI Acceleration

GAME-CHANGING PERFORMANCE FOR INNOVATORS



MARKET FORCES

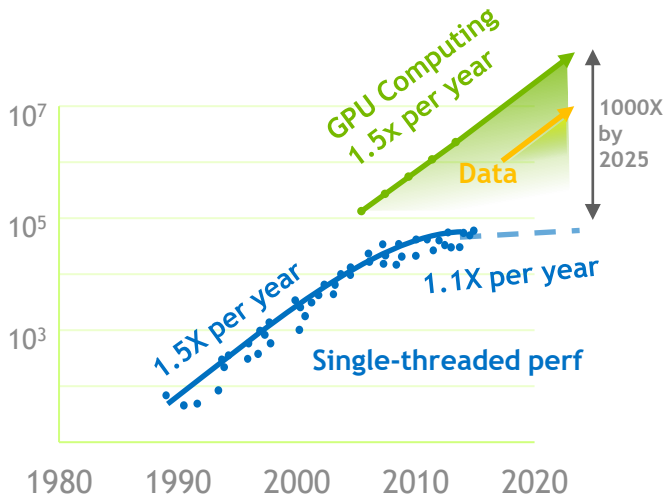
Applications

AI

DATA SCIENCE

HPC

Performance



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

Data Growth Source: *Mapping the Future of Silicon for AI* - September 2017

Old Economics

300 DUAL CPU SERVERS

$3,000,000

180KW

MARKET FORCES

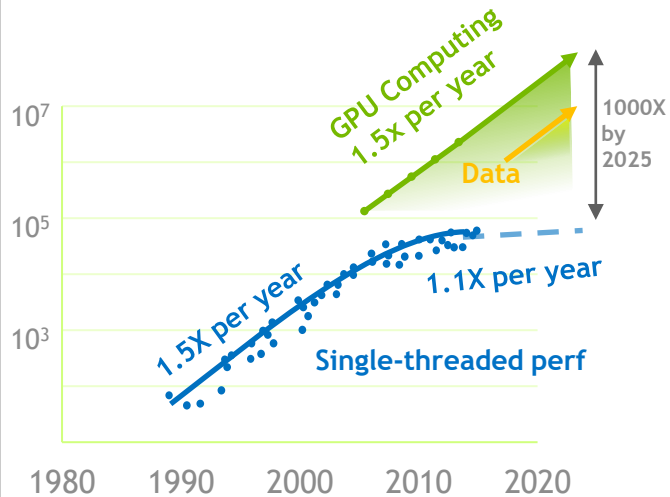
Applications

AI

DATA SCIENCE

HPC

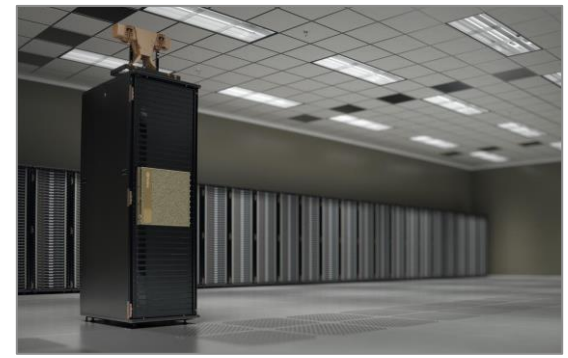
Performance



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

Data Growth Source: *Mapping the Future of Silicon for AI* - September 2017

New Economics



DGX-A100
$199,000
6.5KW

1/15 THE COST, 1/27 THE POWER, 1/60 THE SPACE

JETSON MOMENTUM

700,000+
Developers

3,000+
Customers

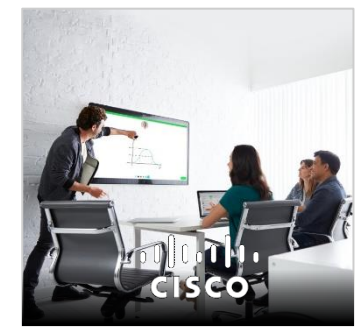
Traffic

Healthcare

Agriculture

Autonomous Drone

Delivery Robot



Collaboration

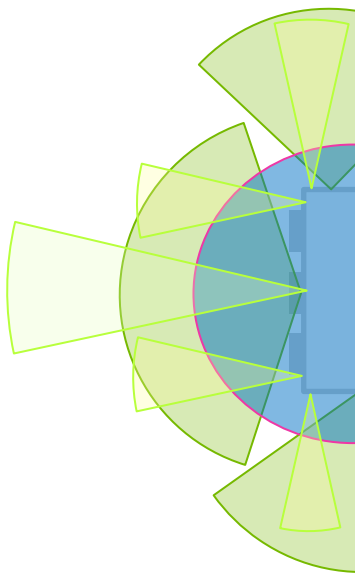
Warehouse

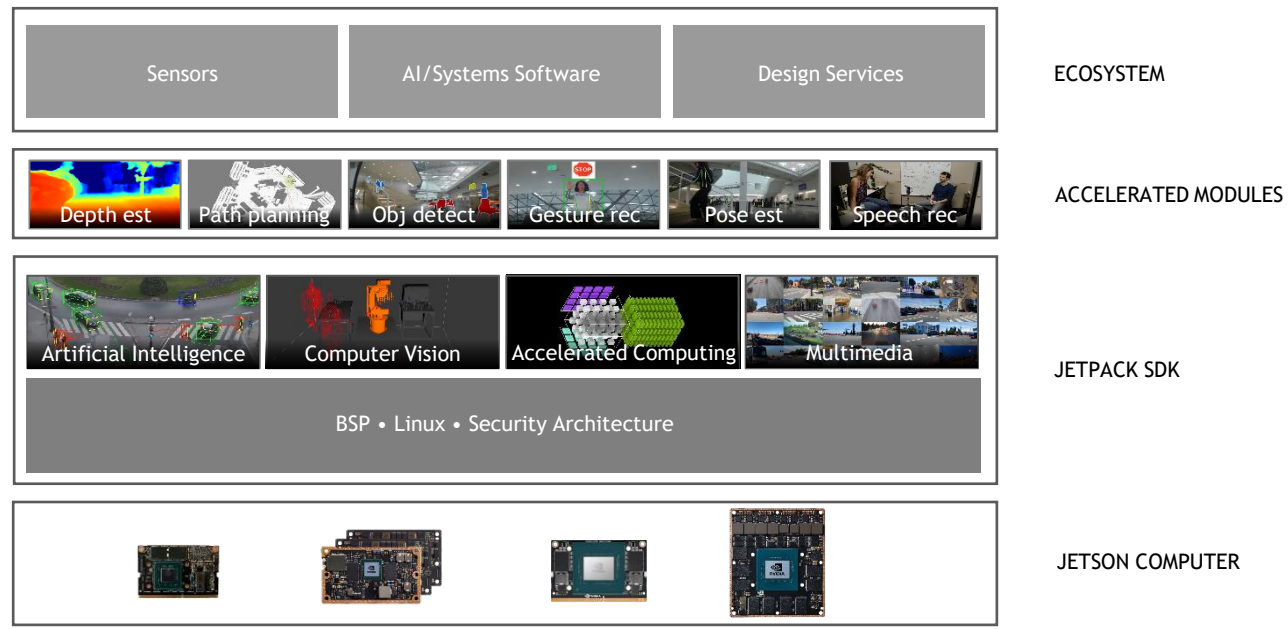
Optical Inspection

Digital Education

NVIDIA JETSON
SOFTWARE-DEFINED AUTONOMOUS MACHINES

Powerful and efficient AI, CV, HPC | Rich Software Development Platform
Open Platform





THE CHALLENGES OF AI TRANSFORMATION

Enterprises Need Infrastructure That Supports the Lifecycle of AI Innovation

From Inspiration

AI practitioners need the right tools for exploration:

- Iterating to the best model, with less effort expended
- Fastest time-to-solution for every training run
- Insulation from the bleeding edge of AI open source

To Production

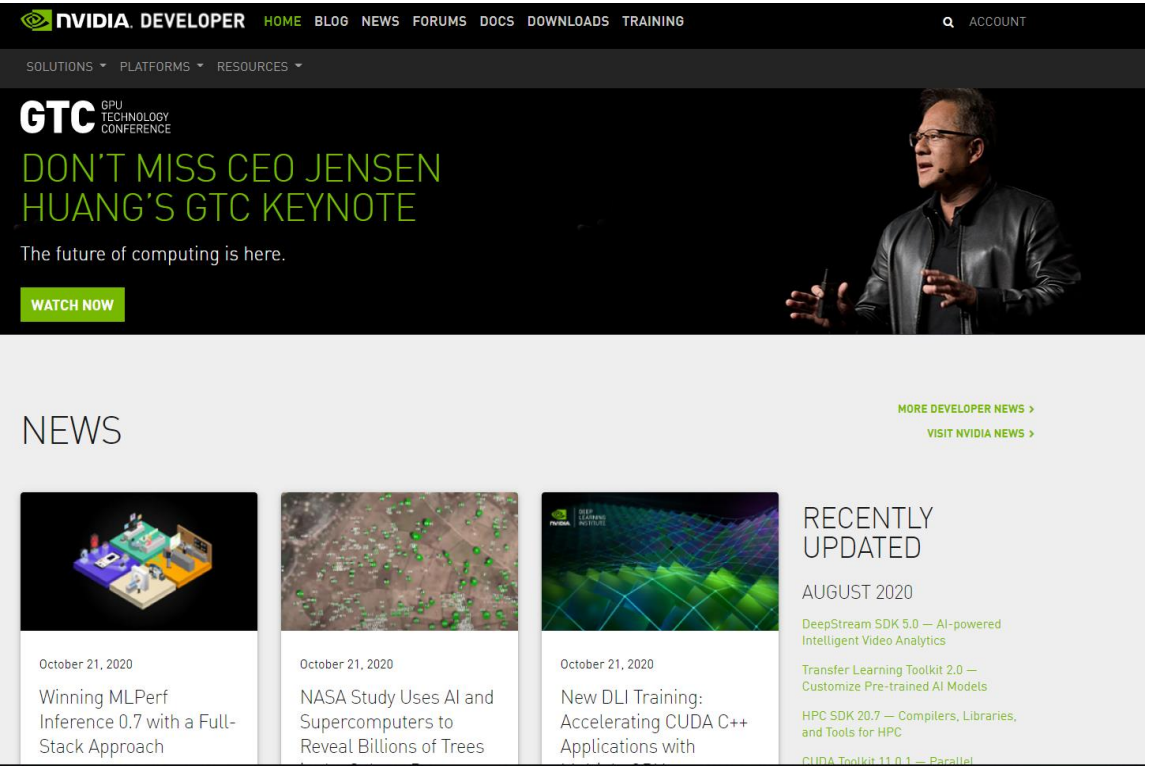
IT needs a standardized approach for AI infrastructure:

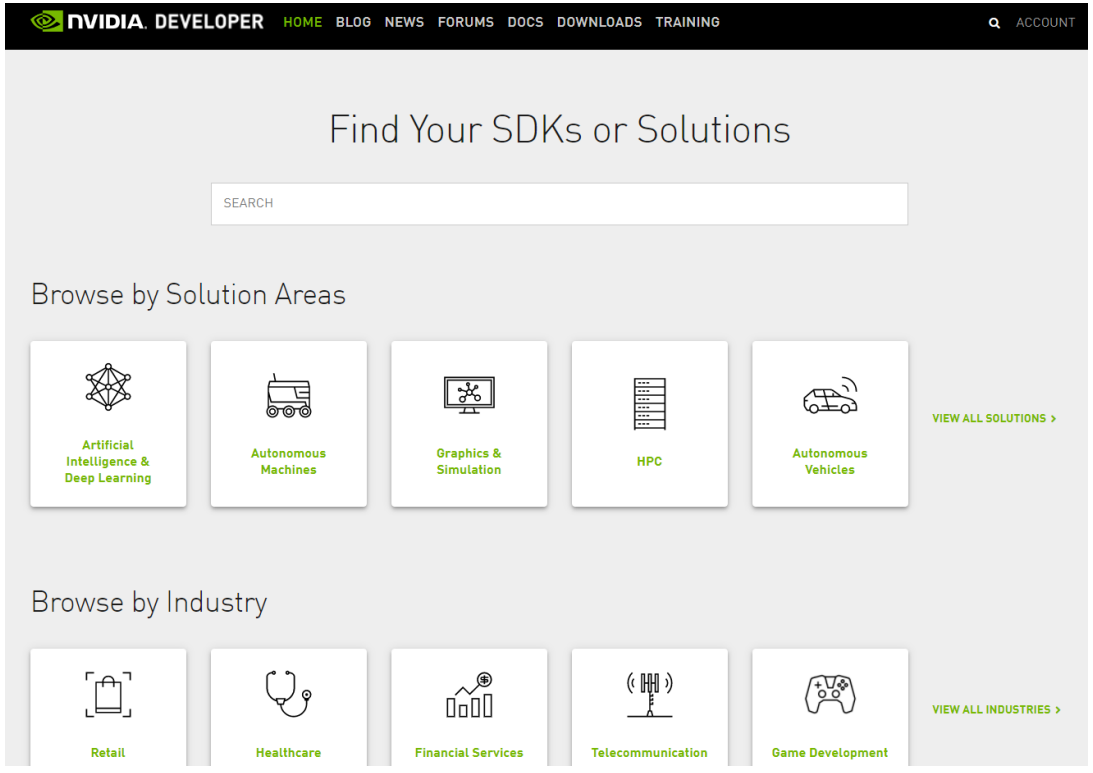
- Simplified infrastructure planning, heterogenous workloads & users
- Security at every layer, operations peace-of-mind
- Linearly predictable performance with scale

NVIDIA TOOLS

NVIDIA DEVELOPER SITE

Developer.nvidia.com





NGC - GPU-OPTIMIZED SOFTWARE

Build AI Faster, Deploy Anywhere

NGC

CONTAINERS

150+

TRAINED MODELS

100+

HELM CHARTS

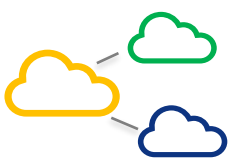
ML, Inference

AI TOOLKITS & SDKs

Healthcare | Smart Cities | Conversational AI | Robotics | more

ENCRYPTED

x86 | ARM | POWER



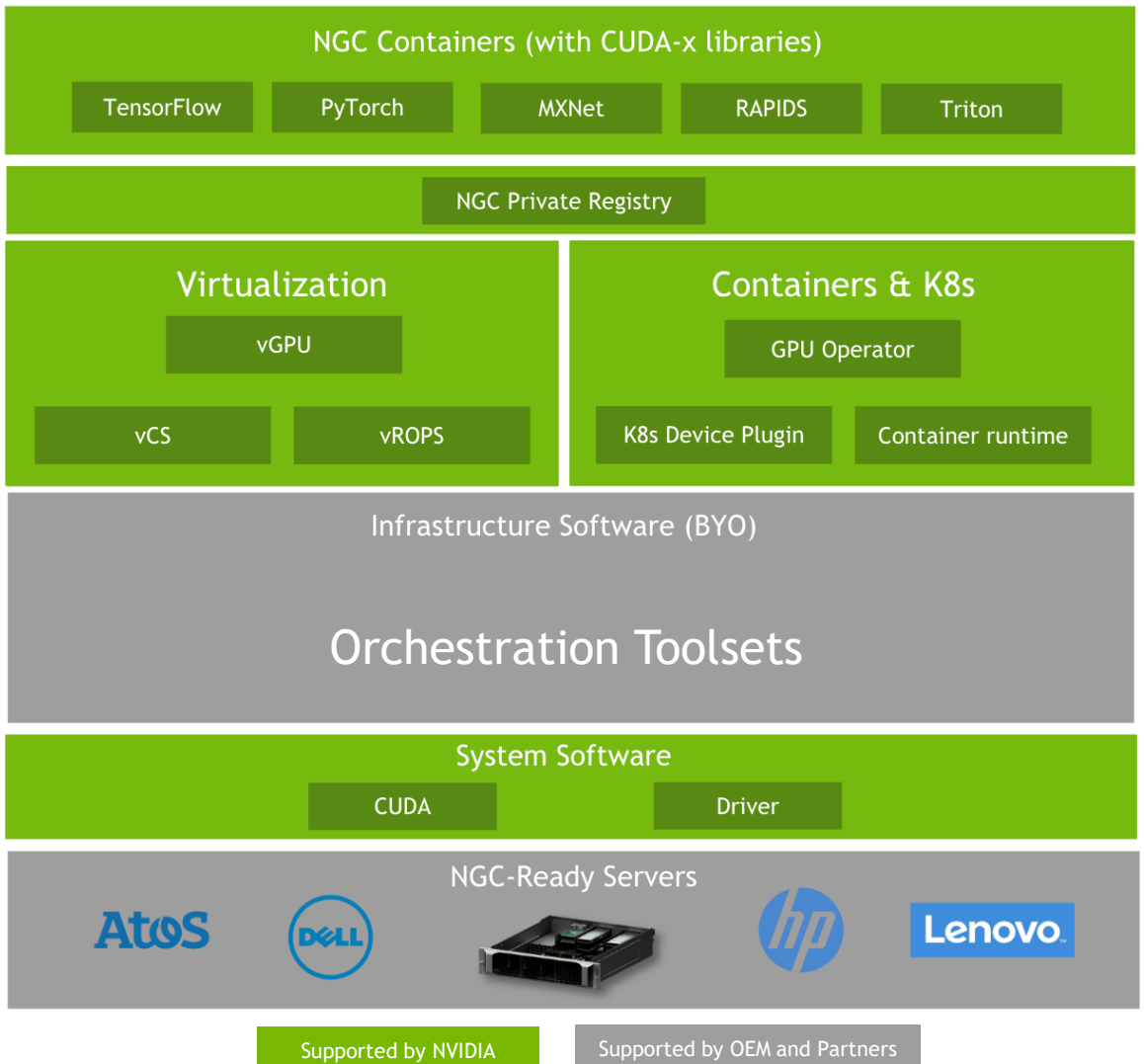
MULTI-CLOUD

ON-PREM

HYBRID CLOUD

EDGE

NGC SUPPORT SERVICES



Minimize system downtime

Direct access to enterprise-grade support from NVIDIA's AI experts to help troubleshoot issues

Build optimized AI solutions

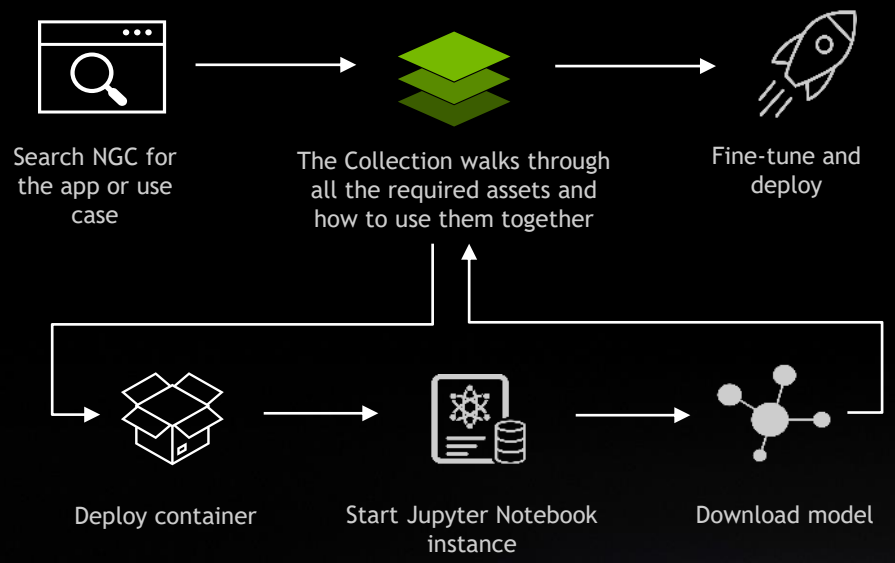
Scale-up and scale-out with help from NVIDIA experts on how best to maximize GPU performance

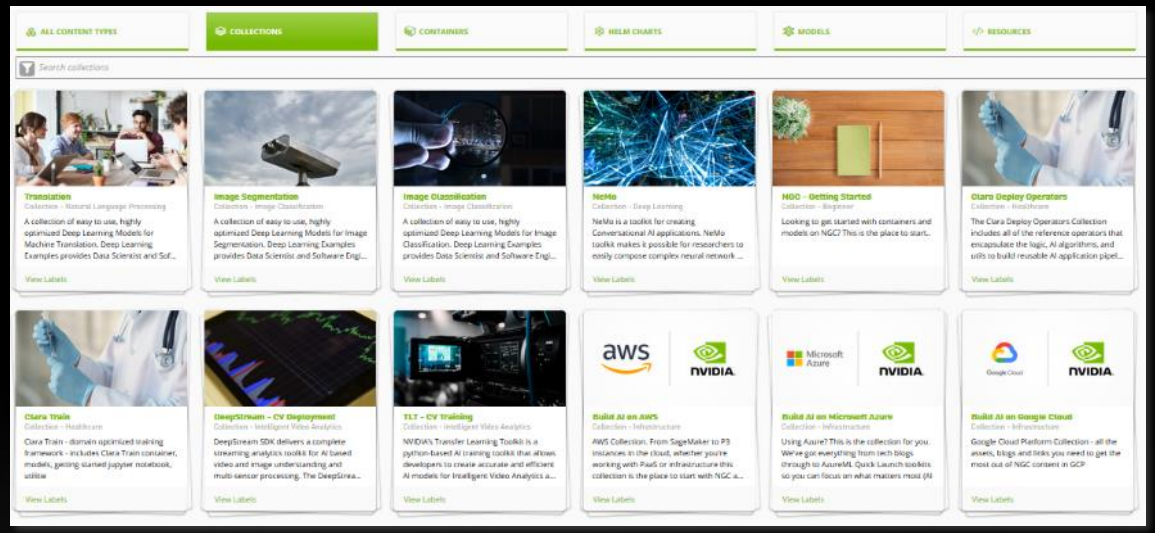
Accelerate time to solution

Support for the entire AI software stack including containers, CUDA and drivers

NGC COLLECTIONS

Everything You Need to Build Your AI Application



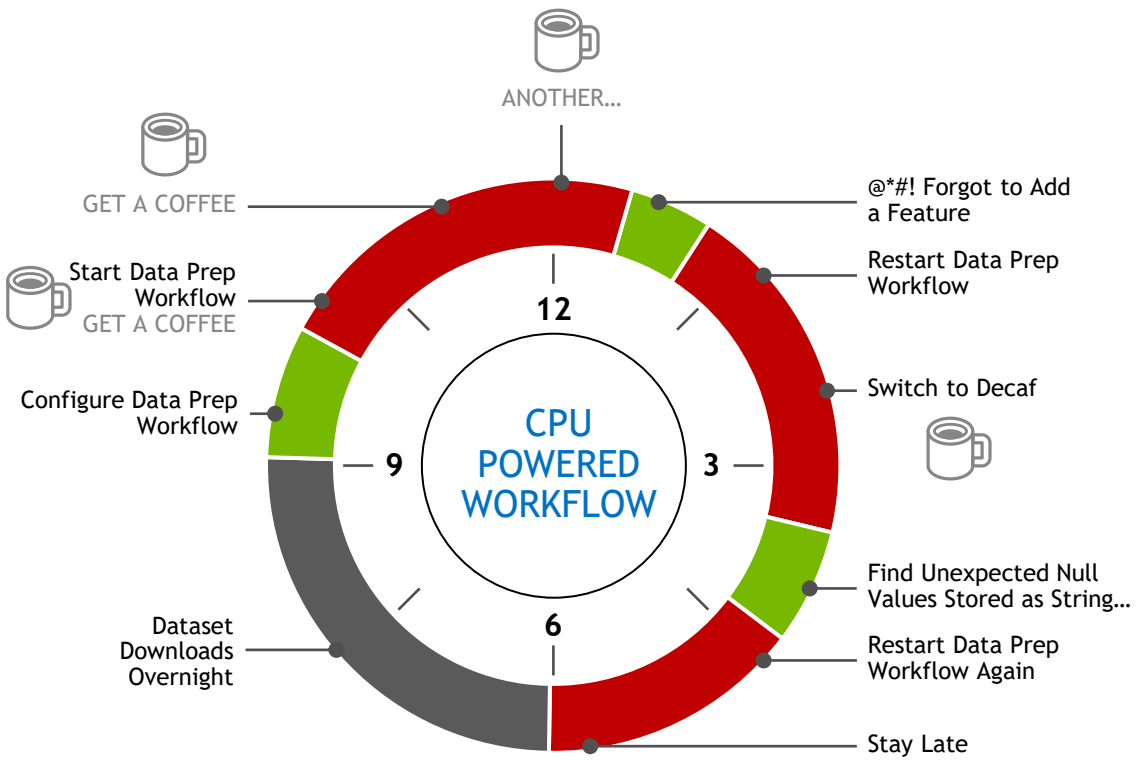


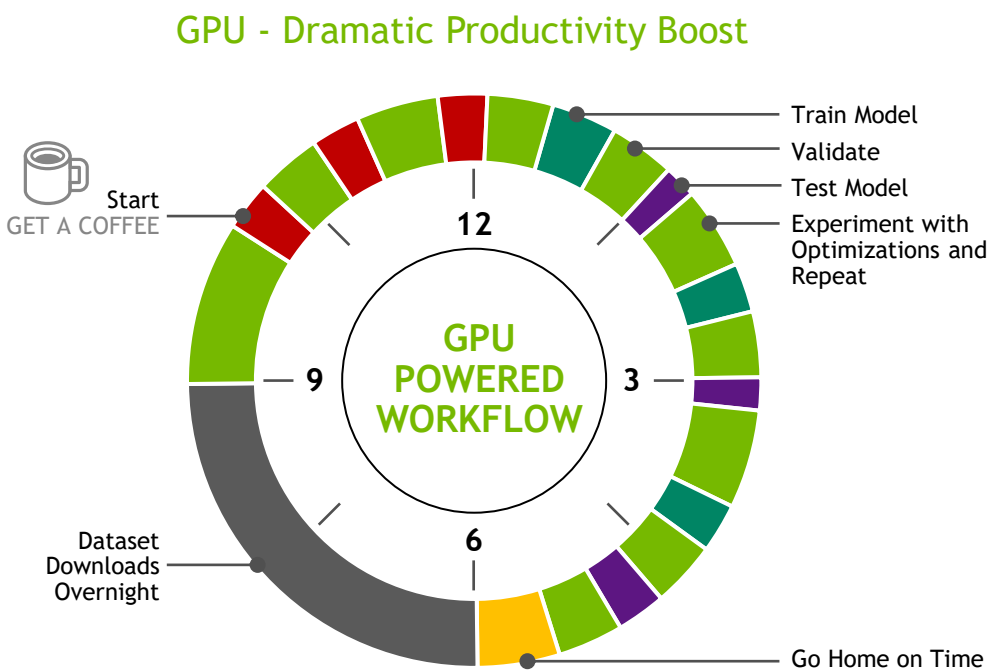
Ready To Use Collections

Conversational AI | Computer Vision | NVIDIA AI App Frameworks

MACHINE LEARNIG - RAPIDS

DAY IN THE LIFE OF A DATA SCIENTIST

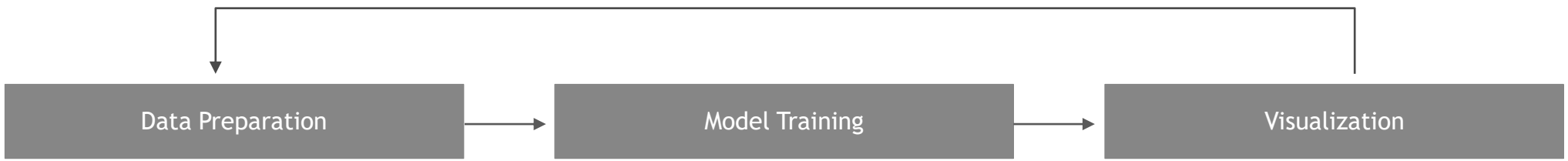


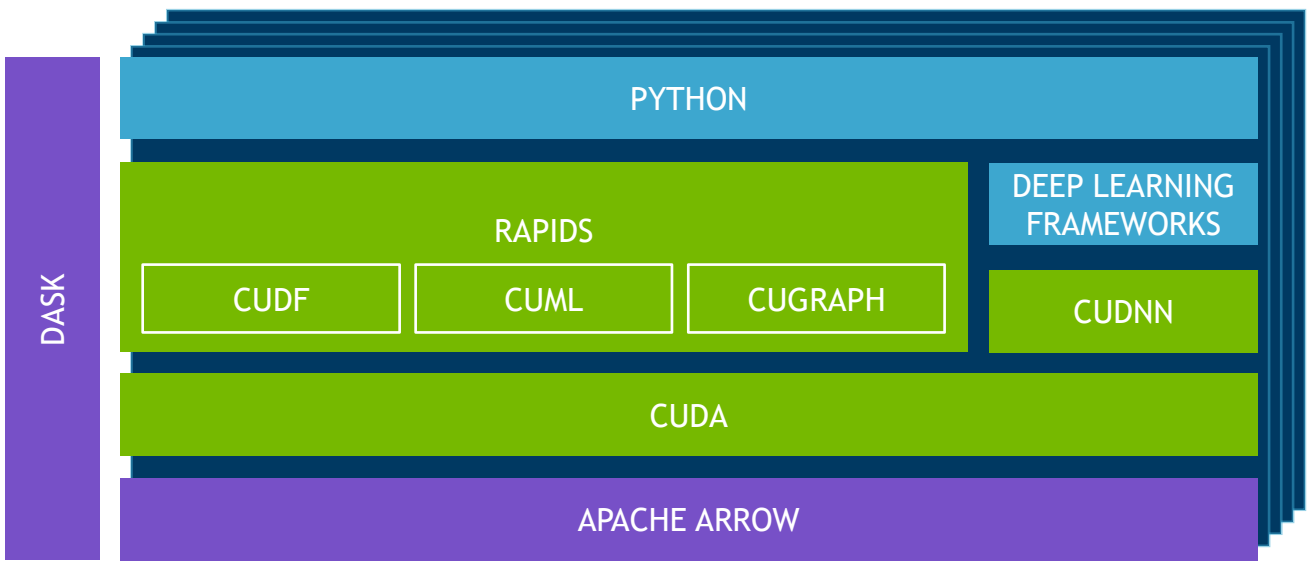


Dataset Collection | Analysis | Data Prep | Train | Inference

RAPIDS — OPEN GPU DATA SCIENCE

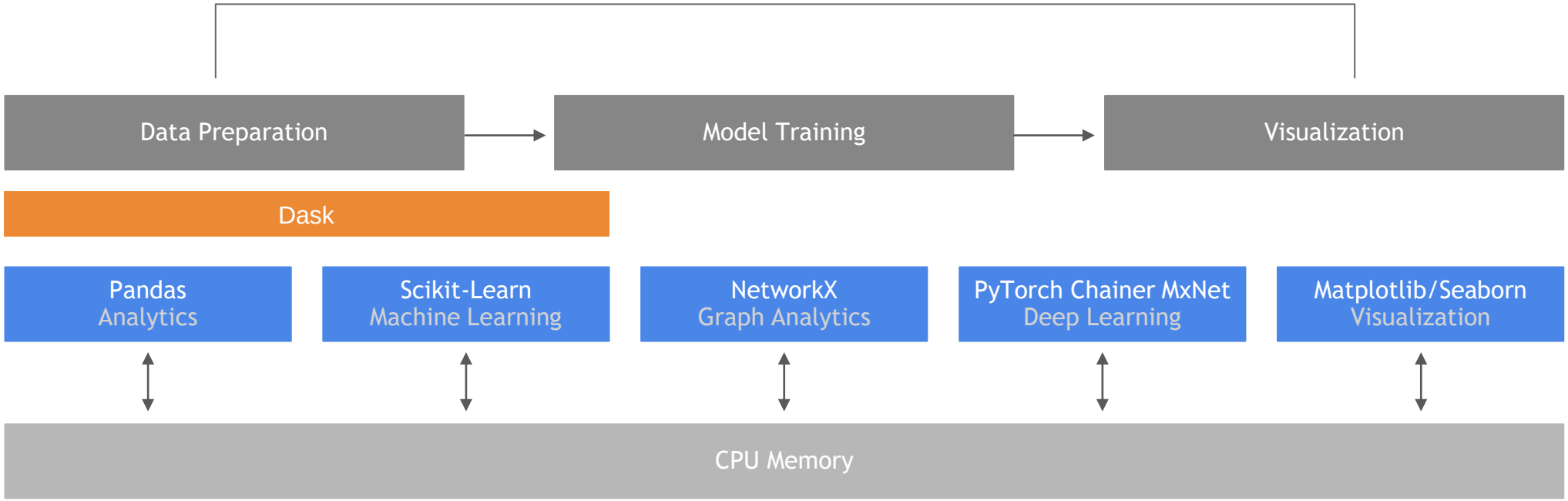
Software Stack





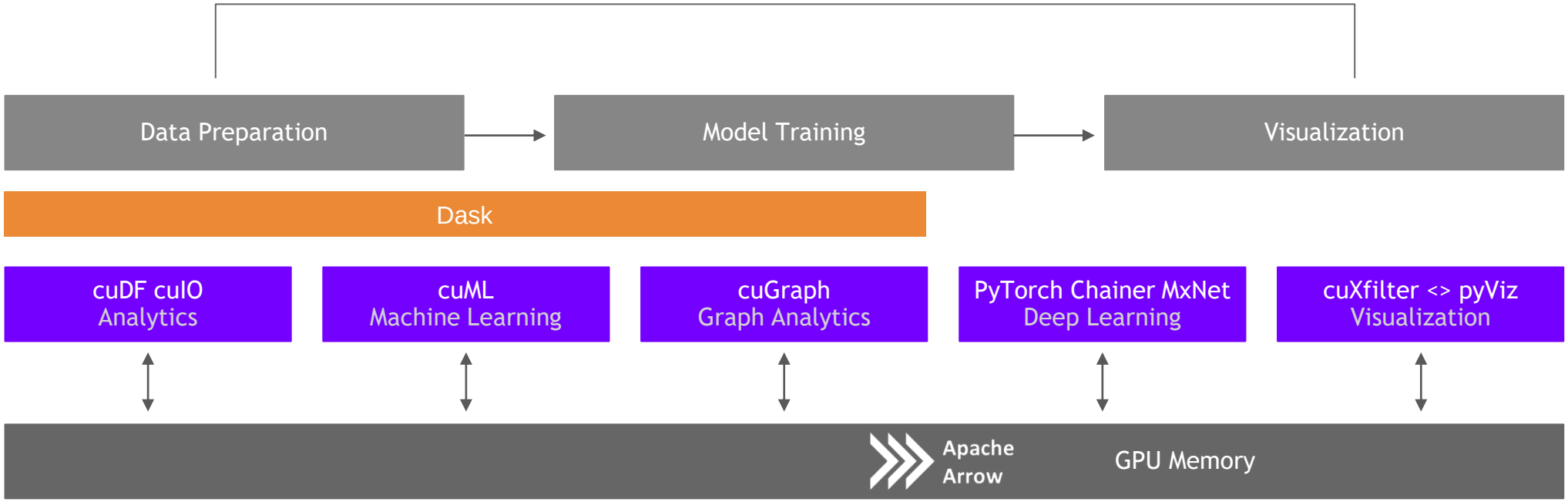
Open Source Data Science Ecosystem

Familiar Python APIs

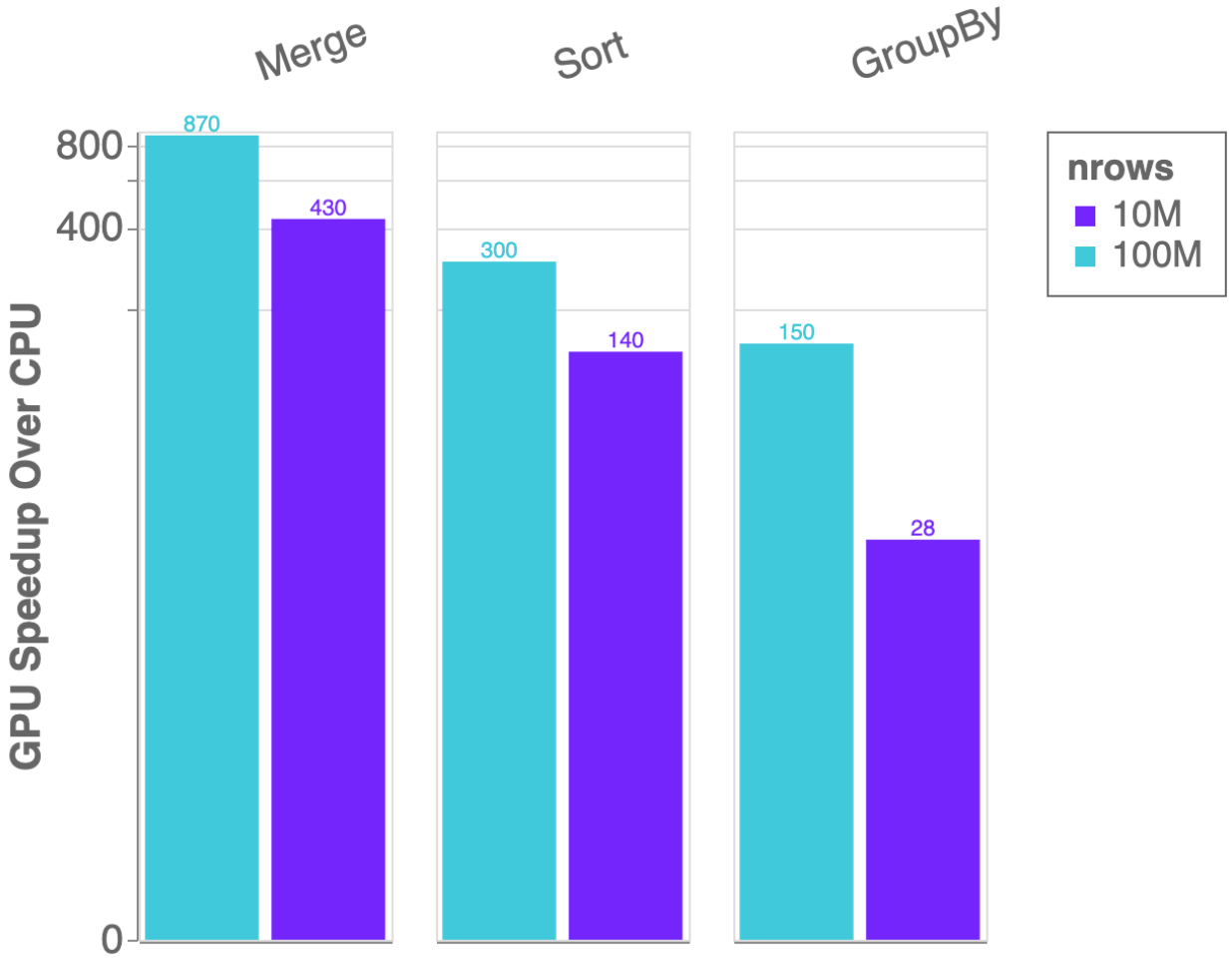


RAPIDS

End-to-End Accelerated GPU Data Science



Benchmarks: single-GPU Speedup vs. Pandas



cuDF v0.9, Pandas 0.24.2

Running on NVIDIA DGX-1:

GPU: NVIDIA Tesla V100 32GB
CPU: Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz

Benchmark Setup:

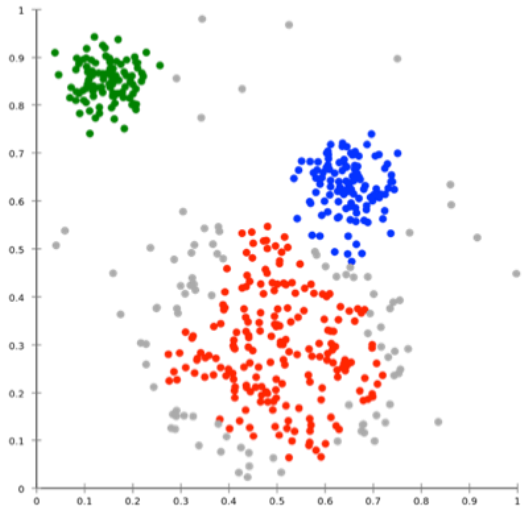
DataFrames: 2x int32 columns key columns, 3x int32 value columns

Merge: inner

GroupBy: count, sum, min, max calculated for each value column

Algorithms

GPU-accelerated Scikit-Learn



Cross Validation

Hyper-parameter Tuning

More to come!

Classification / Regression	Decision Trees / Random Forests Linear Regression Logistic Regression K-Nearest Neighbors
Inference	**Random forest / GBDT inference**
Clustering	K-Means DBSCAN Spectral Clustering
Decomposition & Dimensionality Reduction	Principal Components Singular Value Decomposition UMAP Spectral Embedding
Time Series	**Holt-Winters** Kalman Filtering

Key:

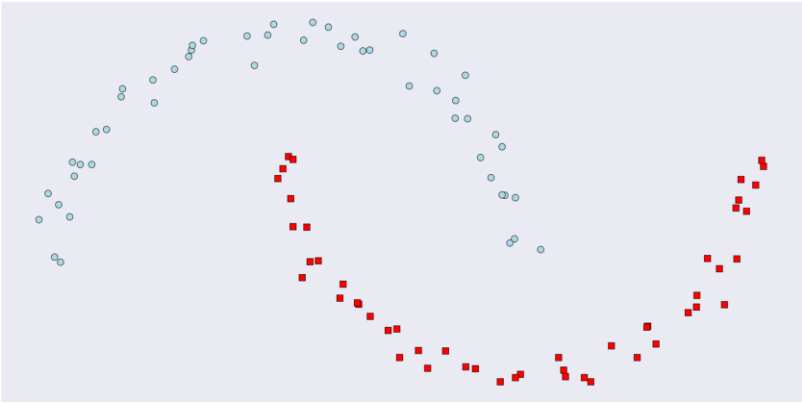
- Preexisting
- **NEW for 0.9**

RAPIDS matches common Python APIs

CPU-Based Clustering

```
from sklearn.datasets import make_moons
import pandas

X, y = make_moons(n_samples=int(1e2),
                  noise=0.05, random_state=0)

X = pandas.DataFrame({'fea%d'%i: X[:, i]
                      for i in range(X.shape[1])})
```

```
from sklearn.cluster import DBSCAN
dbscan = DBSCAN(eps = 0.3, min_samples = 5)

dbscan.fit(X)

y_hat = dbscan.predict(X)
```

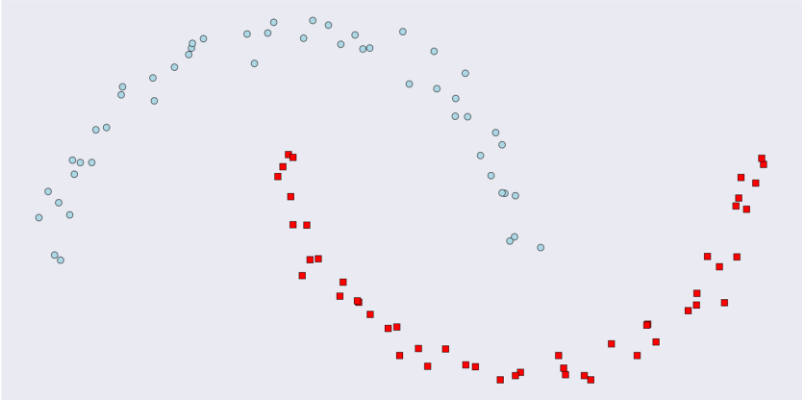


RAPIDS matches common Python APIs

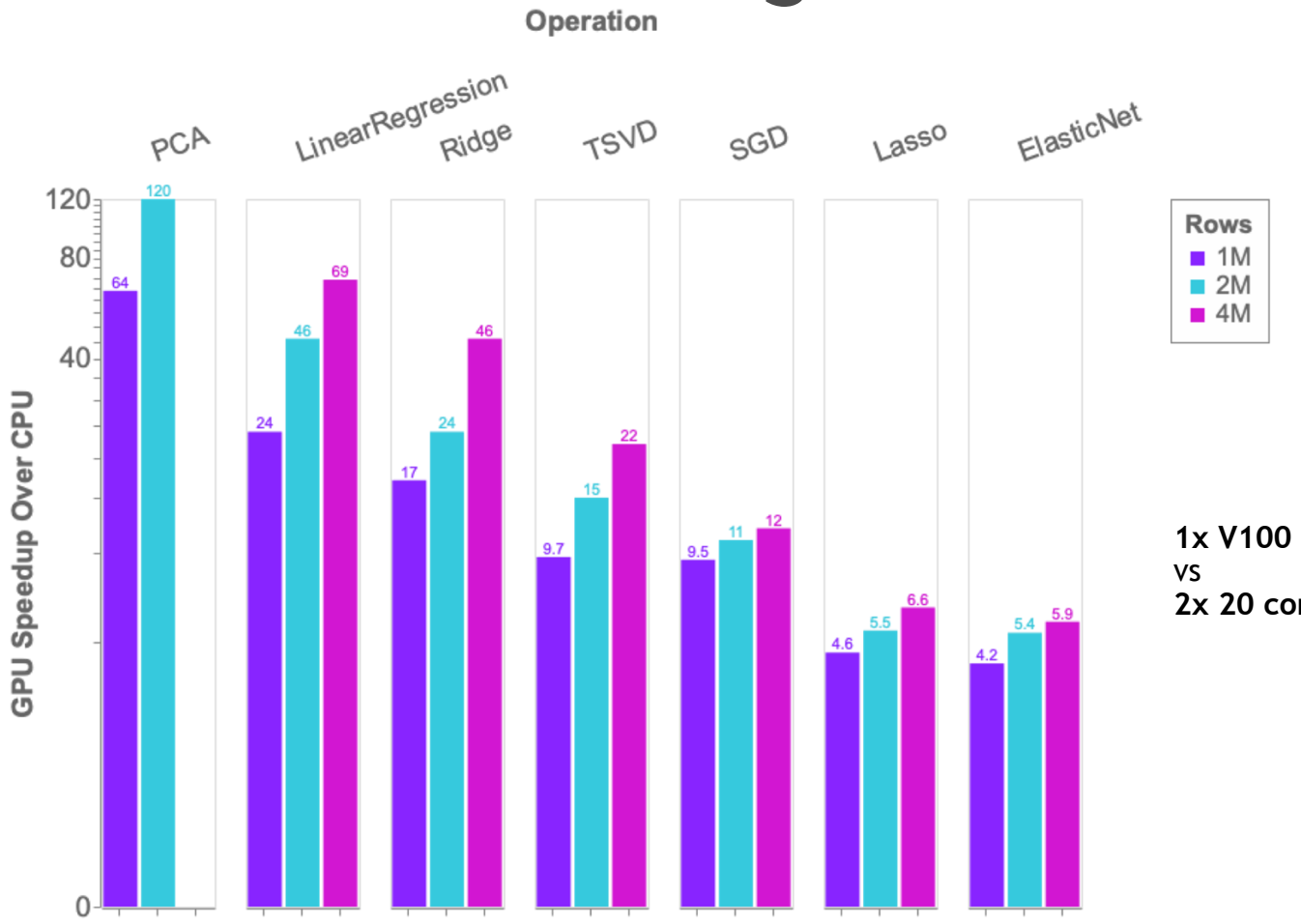
GPU-Accelerated Clustering

```
from sklearn.datasets import make_moons
import cudf

X, y = make_moons(n_samples=int(1e2),
                  noise=0.05, random_state=0)

X =   cudf.DataFrame({'fea%d'%i: X[:, i]
                      for i in range(X.shape[1])})
```

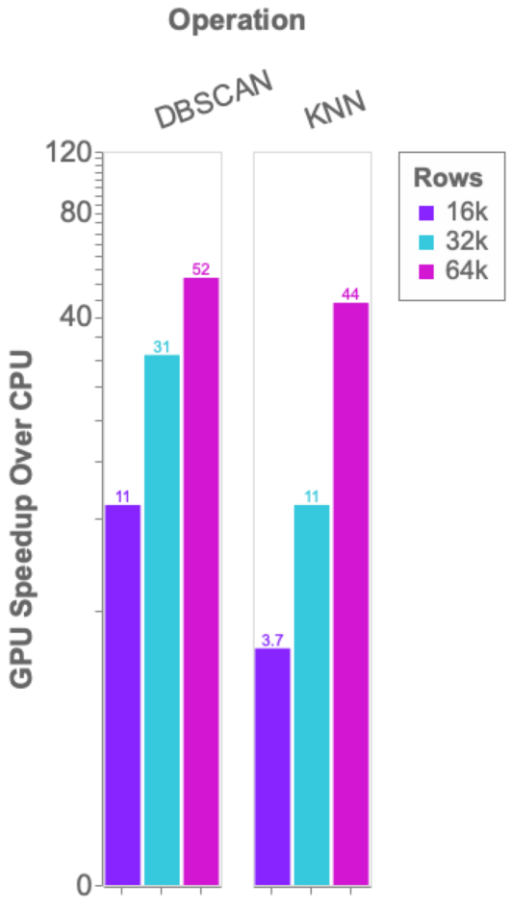
```
from cuml          import DBSCAN
dbscan = DBSCAN(eps = 0.3, min_samples = 5)

dbscan.fit(X)

y_hat = dbscan.predict(X)
```



Benchmarks: single-GPU cuML vs scikit-learn

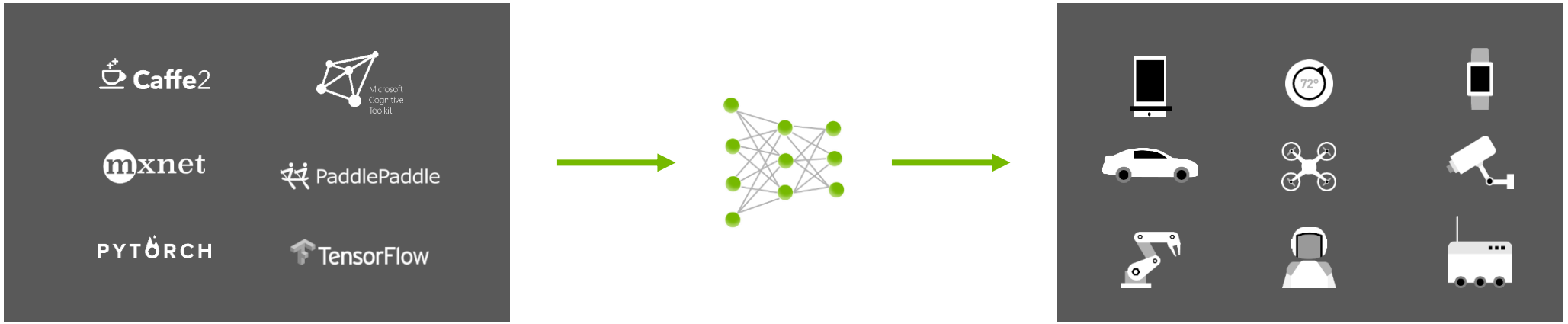


1x V100
vs
2x 20 core CPU



TENSORRT

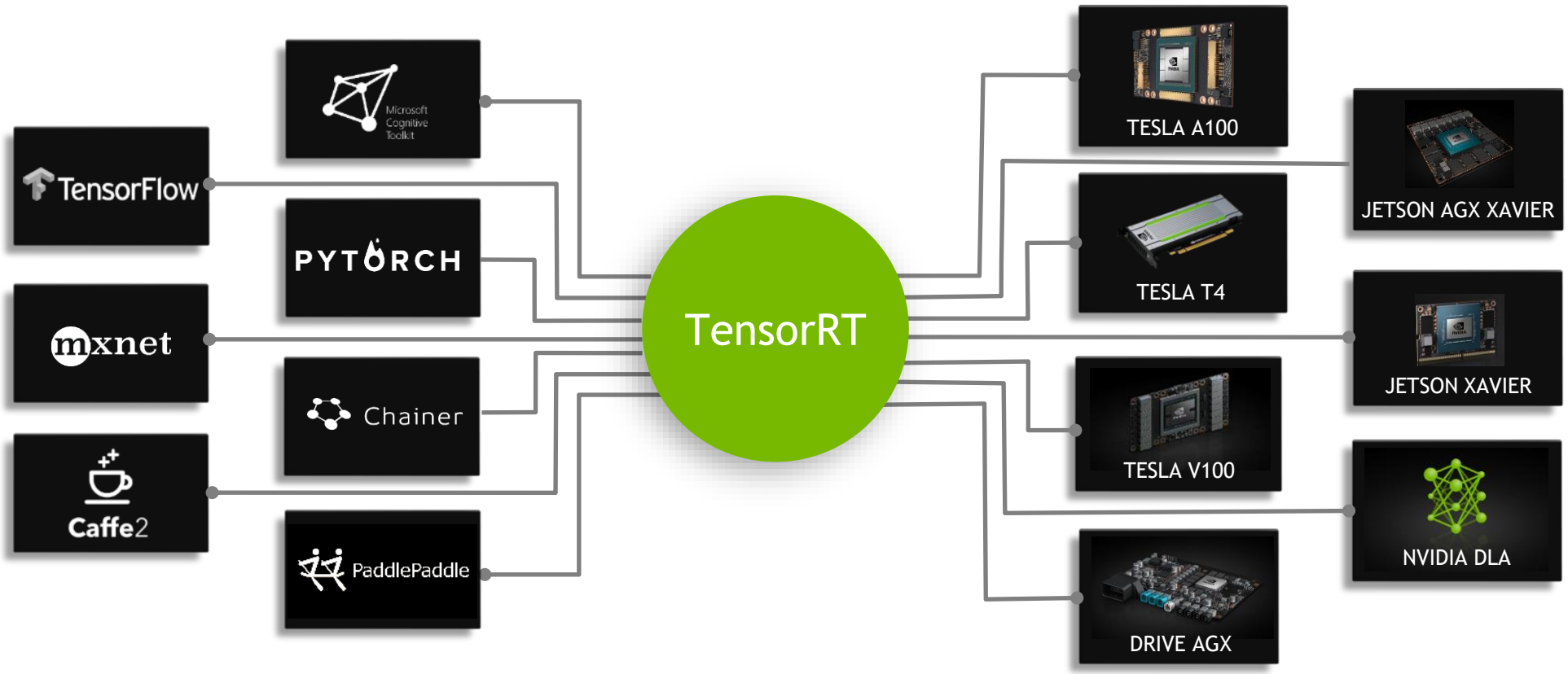
AI INFERENCE NEEDS TO RUN EVERYWHERE



Training DNN Model Inference

NVIDIA TensorRT

From Every Framework, Optimized For Each Target Platform



NVIDIA TensorRT

SDK for High-Performance Deep Learning Inference

Optimize and Deploy neural networks in production

Maximize throughput for latency-critical apps with compiler & runtime

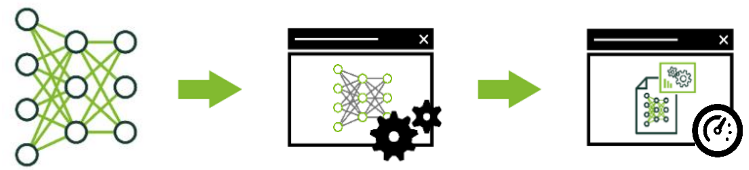
Deploy responsive and memory efficient apps

FP32, TF32, BFLOAT16, FP16 & INT8

Optimize every network including CNNs, RNNs and Transformers

Accelerate every framework - ONNX support, TensorFlow integration

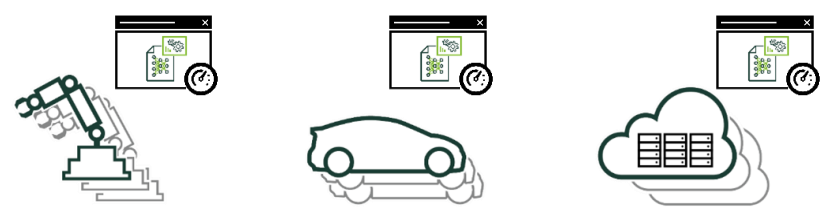
Run multiple models on a node with containerized inference server



Trained DNN

TensorRT Optimizer

TensorRT Runtime Engine



Embedded

Automotive

Data center

Jetson

Drive

Tesl
a

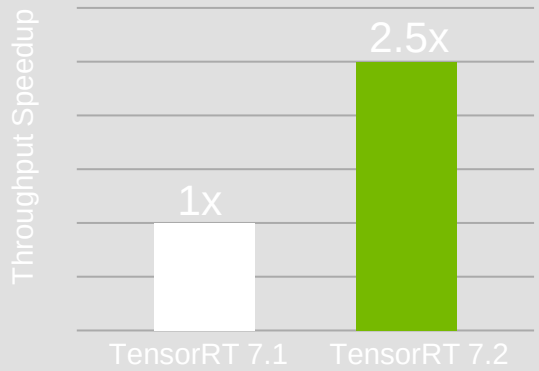
developer.nvidia.com/tensorrt

ANNOUNCING TensorRT 7.2

2000+ Kernel Optimizations | CNNs, RNNs, MLPs, Transformers

New Optimizations for AI-based Audio & Video Workloads

2.5x Faster Recommenders (NCF) Vs TensorRT 7.1



Improved performance with Tensor Core in INT8 mode

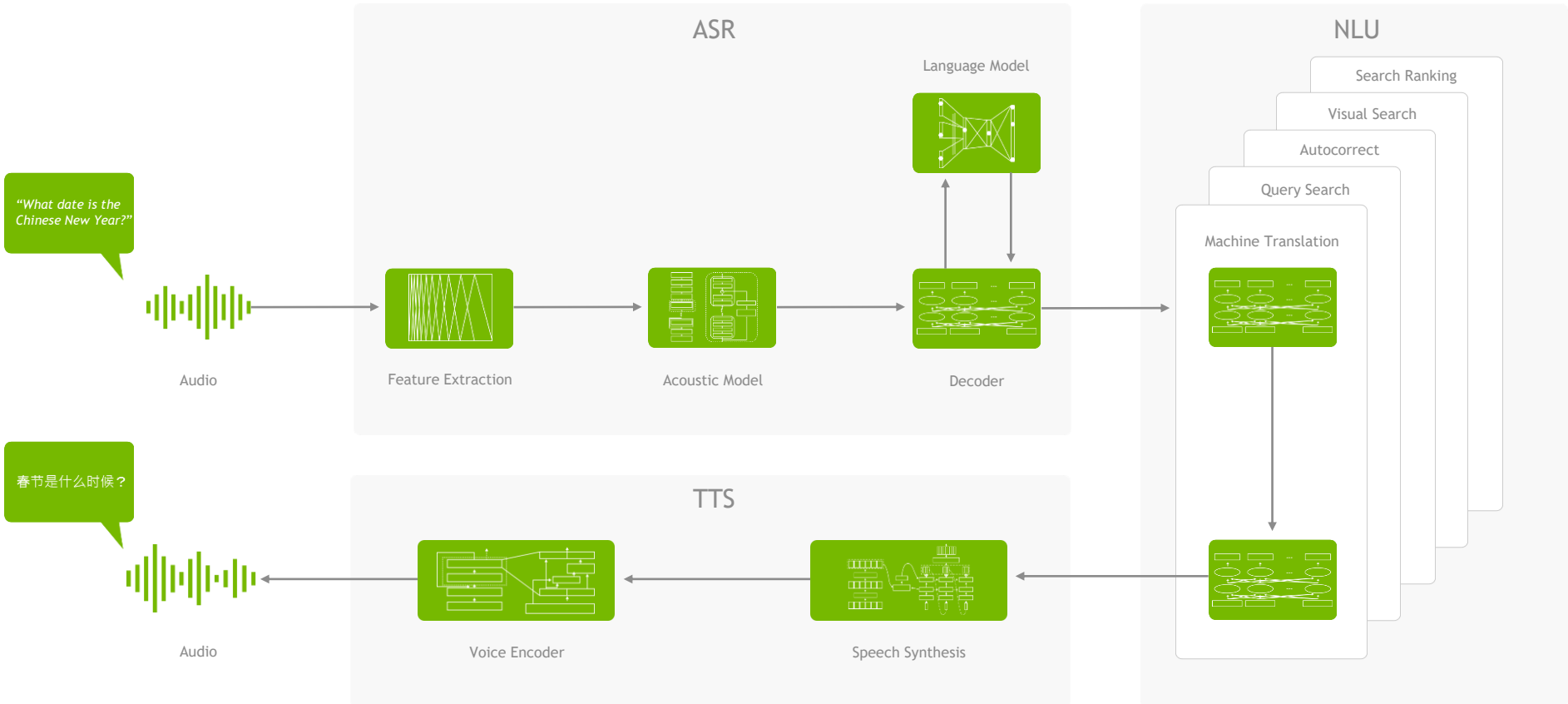
2x Faster RNNs Vs TensorRT 7

New RNNs optimizations for apps such as Anomaly & Fraud Detection

Available in Q4, 2020 to members of the NVIDIA Developer Program from
developer.nvidia.com/tensorrt

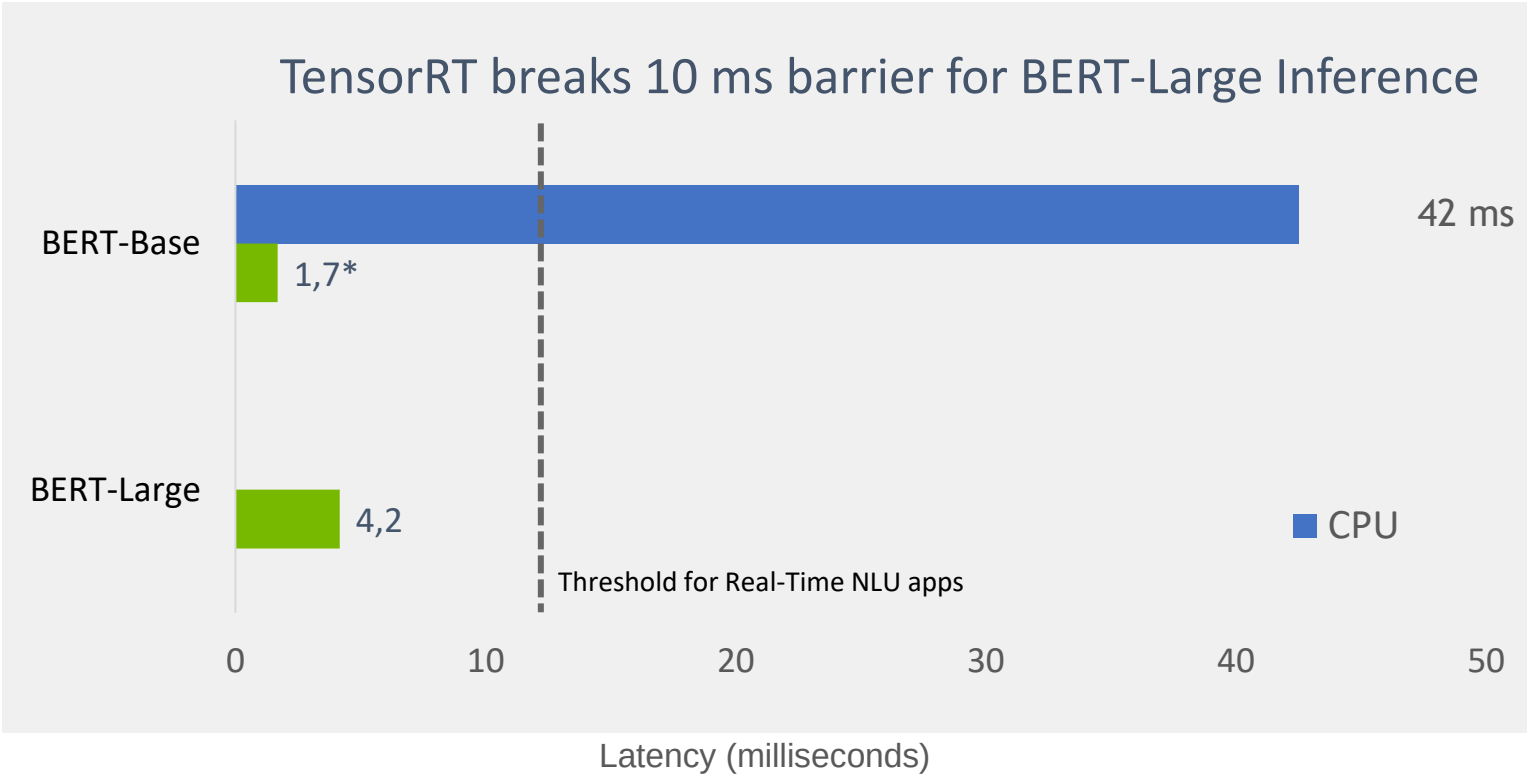
TensorRT ENABLES INTERACTIVE CONVERSATIONAL AI

Now Possible To Run ASR, NLU & TTS Within 300 ms



BERT-LARGE INFERENCE IN 4.2ms

Makes Real-Time Natural Language Understanding Possible



* In our tests, OpenVINO Release 2019 R2 did not execute BERT-Large and exited with an error

BERT Sample Code in TensorRT Repo
Jupyter Python Notebook

FEATURE WALKTHROUGH

NEW FULLY-CONNECTED LAYER OPTIMIZATIONS

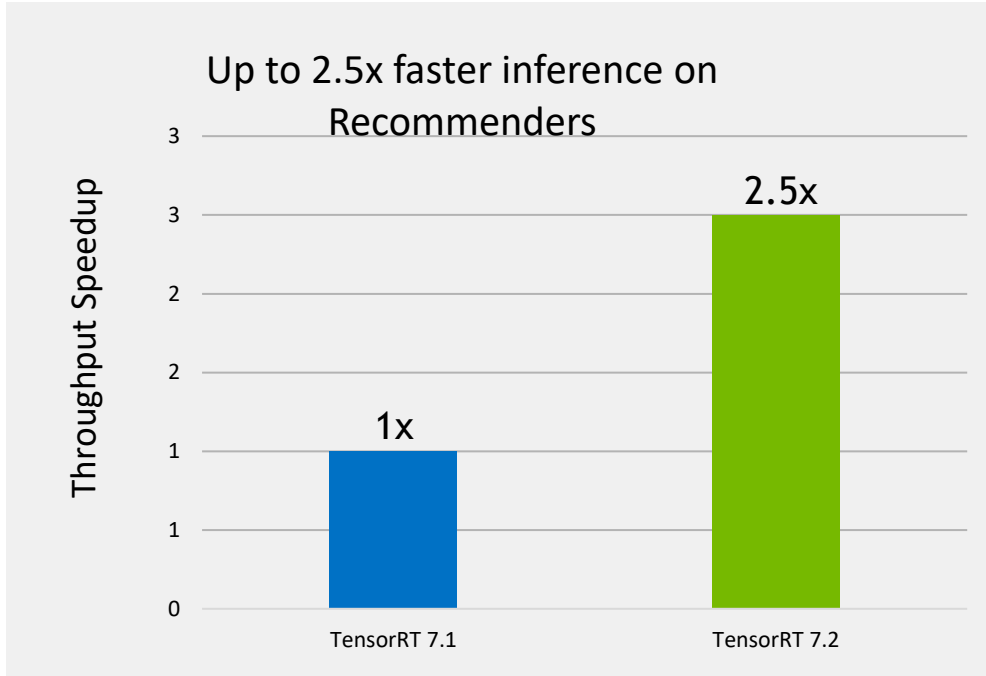
Maximize Recommender, NLU and Object Detection

Accelerate MLP based networks

Replaces FullyConnected layers with 1x1 Convolutions, increasing rate of computations

New optimizations in FullyConnected Layers result in greater performance in networks like MLPs, BERT

Improved performance with Tensor Core in INT8 mode.



GPU: A100; CUDA: 450.51, BS = 65536
Networks: Recommender: NCF, INT8

RECURRENT NEURAL NETWORK OPTIMIZATIONS

High Performance ASR and TTS apps

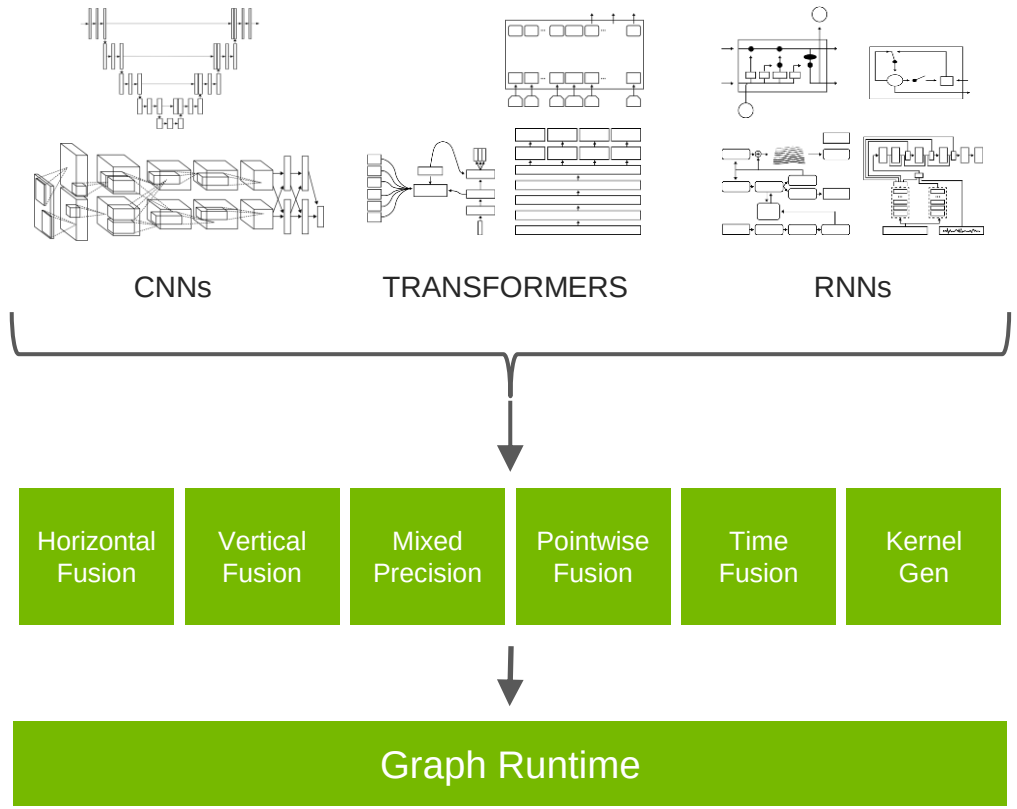
Deploy highly-optimized Conversational AI apps in production environments

New API to define loops found in RNNs

Compiler fuses pointwise ops, generates optimized kernels, and fuses ops across time steps

Run ASR, NLU and TTS within 300 ms, a requirement for real time apps, 10x perf vs CPU

Models Supported: BERT, MT-DNN, RoBERTa, Tacotron 2, WaveRNN, DeepASR, GNMT, LSTM Peephole, LSTM Autoencoder



Blog: Real Time Text-to-Speech using TensorRT

RECURRENT NEURAL NETWORK OPTIMIZATIONS

High Performance ASR and TTS apps

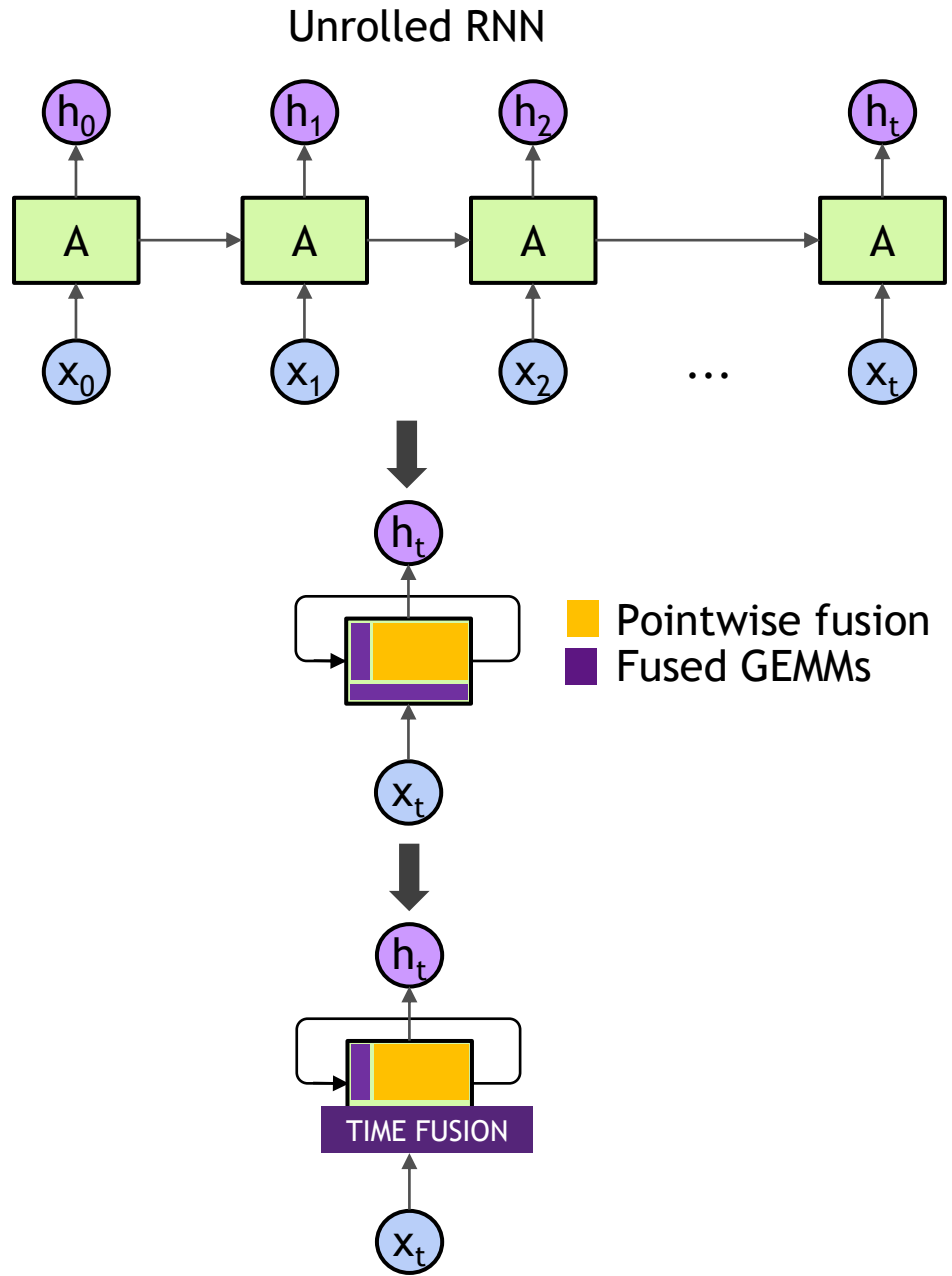
Deploy highly-optimized Conversational AI apps in production environments

New API to define loops found in RNNs

Compiler fuses pointwise ops, generates optimized kernels, and fuses ops across time steps

Models Supported: Tacotron 2, WaveRNN, DeepASR, GNMT, LSTM Peephole, LSTM Autoencoder

Blog: Real Time Text-to-Speech using TensorRT



VARIABLE INPUT SIZE SUPPORT

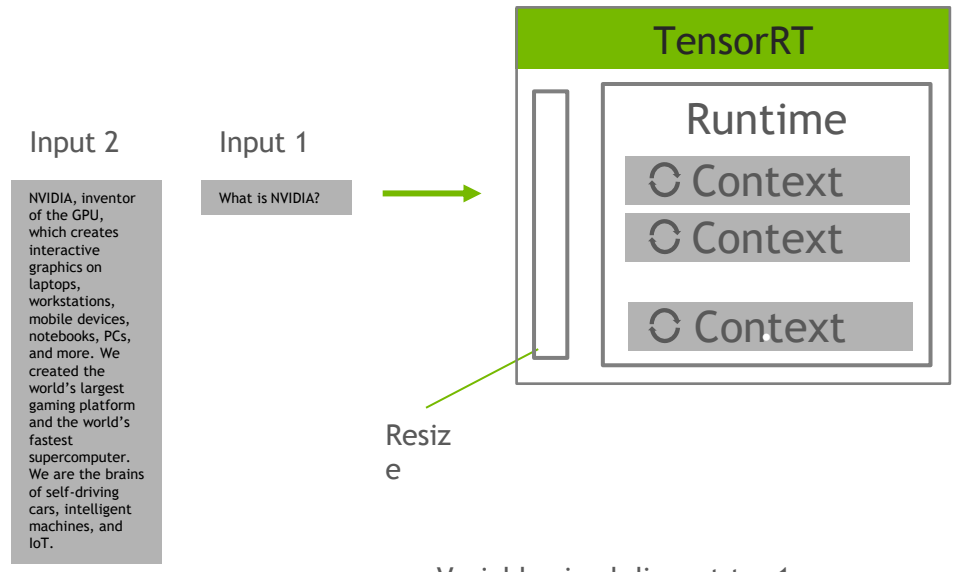
New API to accelerate apps that receive variable sized inputs

Maximize inference performance for apps that receive variable sized inputs

Speed up computer vision, speech and conversational AI apps using easily

High performance across inputs with varying sizes with optimization profiles

Available through new open source ONNX parser



Variable sized dim set to -1

```
auto input = preprocessorNetwork->addInput("input", nvinfer1::DataType::kFLOAT, Dims3{1, -1, -1});
auto profile = builder->createOptimizationProfile(); // create optimization profile
profile->setDimensions(input->getName(), OptProfileSelector::kMIN, Dims3{1, 1, 1}); // min dim
profile->setDimensions(input->getName(), OptProfileSelector::kOPT, Dims3{1, 28, 28});
//optimized dim
profile->setDimensions(input->getName(), OptProfileSelector::kMAX, Dims3{1, 56, 56}); // max dim
```

VARIABLE BATCH SIZE SUPPORT

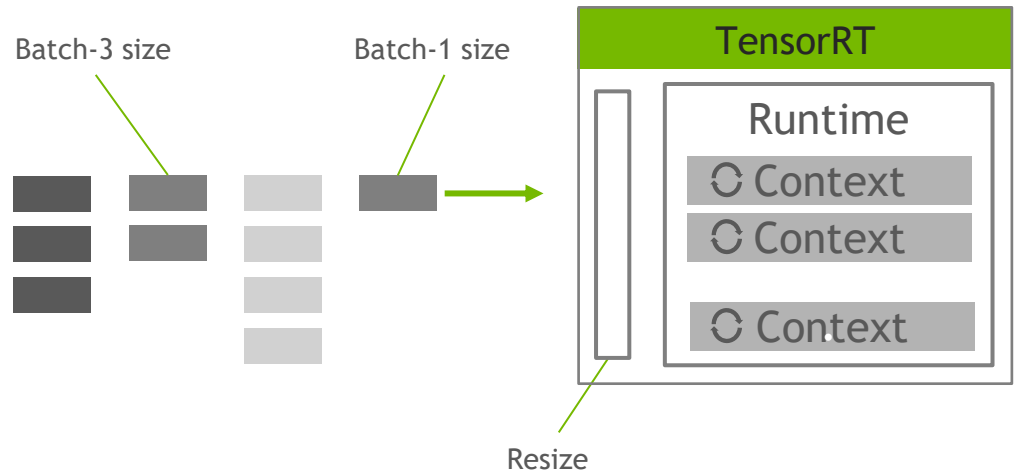
Maximize inference performance for apps with fluctuating loads

New API to efficiently accelerate apps that receive variable batch sizes

Reuse engine across multiple batch sizes efficiently using optimization profiles

Deploy as a service with TensorRT Inference Server

Available through new open source ONNX parser




```
explicit_batch_flag = 1 << int(trt.NetworkDefinitionCreationFlag.EXPLICIT_BATCH)
input_ids = network.add_input(name="input_ids", dtype=trt.int32, shape=(-1, S))
segment_ids = network.add_input(name="segment_ids", dtype=trt.int32, shape=(-1,
S))
input_mask = network.add_input(name="input_mask", dtype=trt.int32, shape=(-1, S))
```

INT8 API & OPTIMIZATIONS

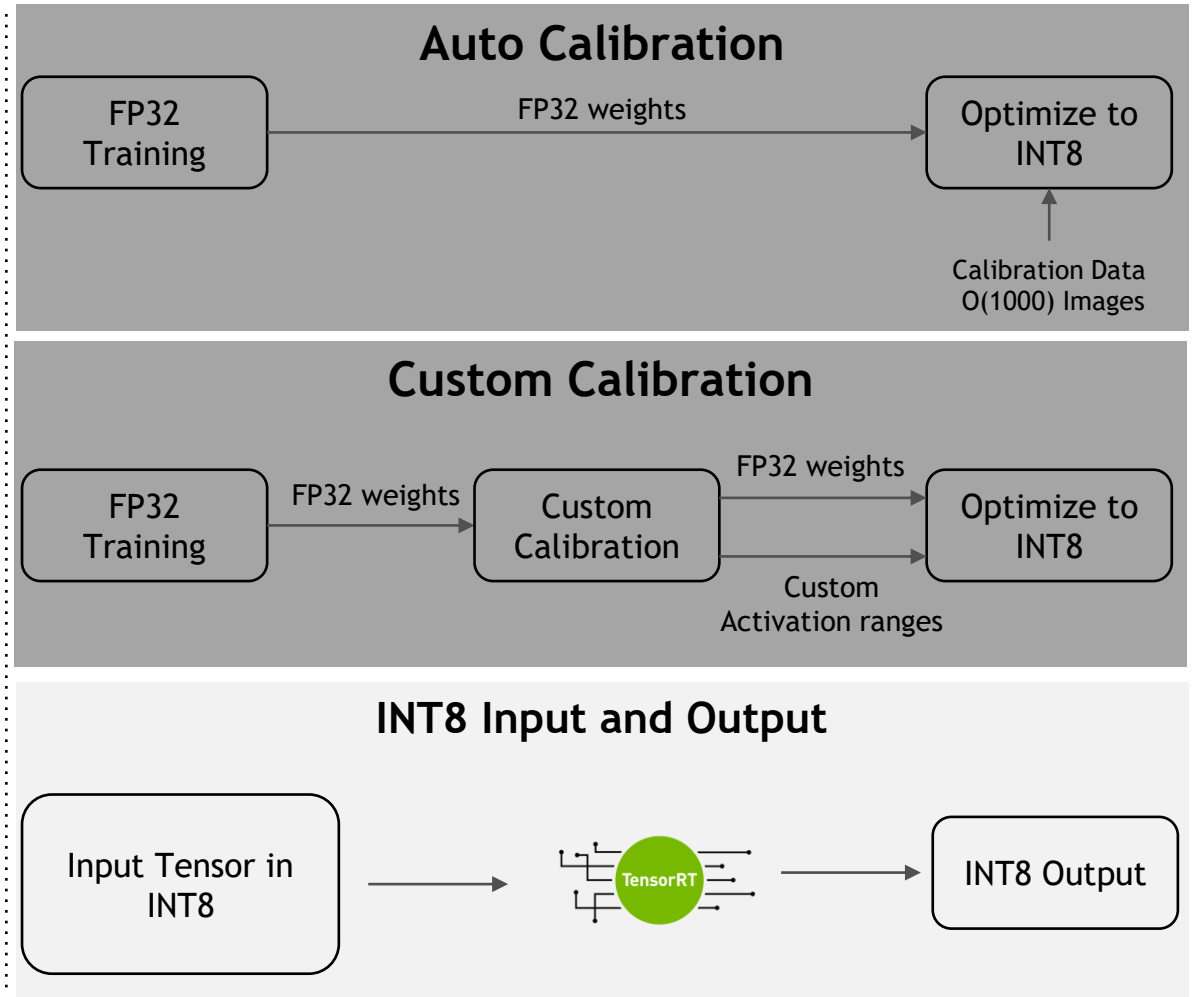
High-performance Optimizations and Flexible APIs For Mixed Precision Inference

Maximize throughput at low latency with mixed precision compute in production

Apply INT8 quantization aware training or custom calibration algorithms with new APIs

Control precision per-layer with new APIs

Support for INT8 input and output for TensorRT engine and plugins



3-D CONVOLUTIONS

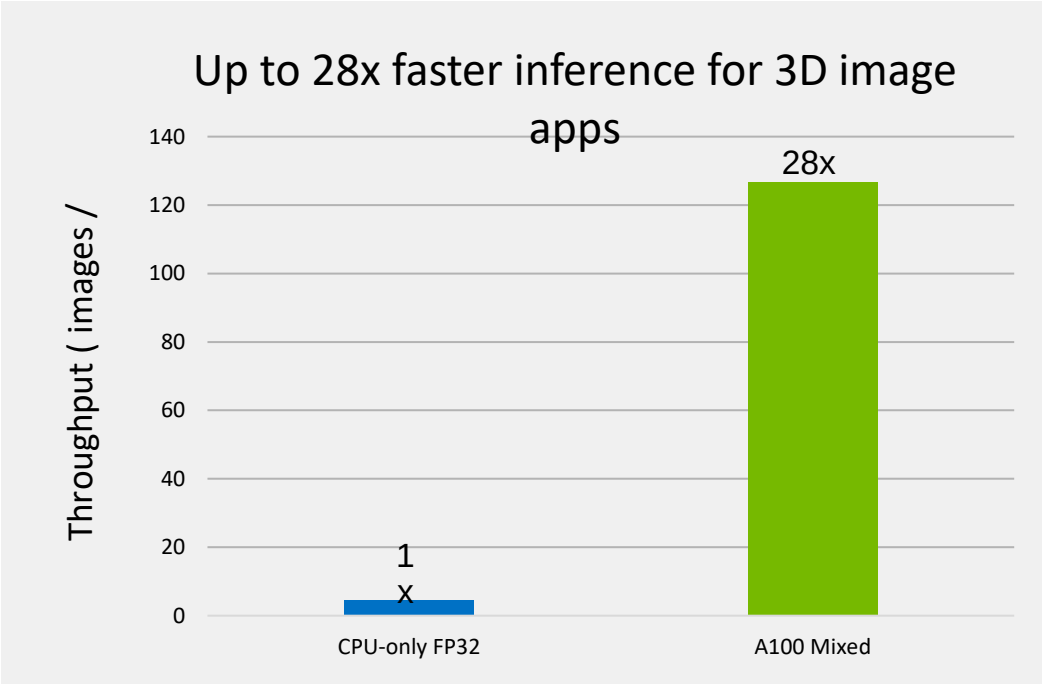
Maximize performance for 3D image based workloads

New layers to accelerate 3D image based apps common in healthcare

Up to 28x faster than CPU-only platforms

New layers for 3D convolution, 3D pooling and 3D deconvolution

Available through ONNX parser



3D UNet with Deconv, Image size = 144x144x144
CPU: Skylake Gold 6140, 2.5GHz, Ubuntu 16.04; 18 CPU threads. OpenVINO 2019 R2, BS=1
Tesla A100; CUDA (455.23); TensorRT 7.2, BS = 1

2-D UNET OPTIMIZATIONS

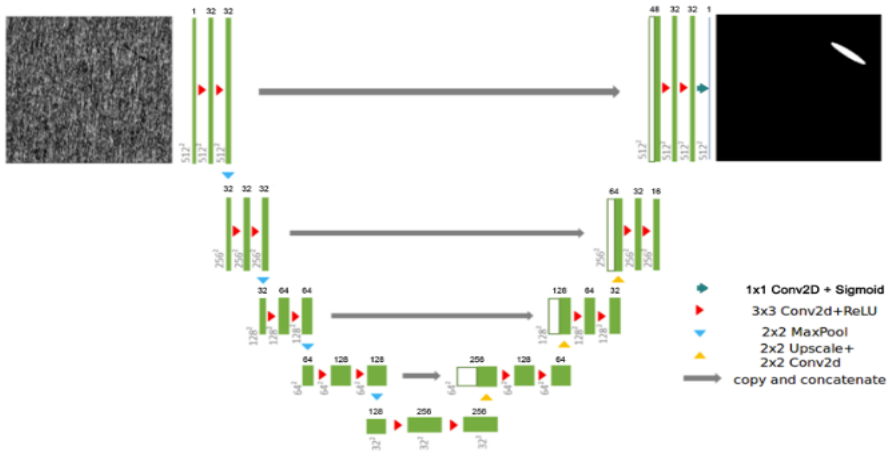
Maximize performance for UNet based workloads

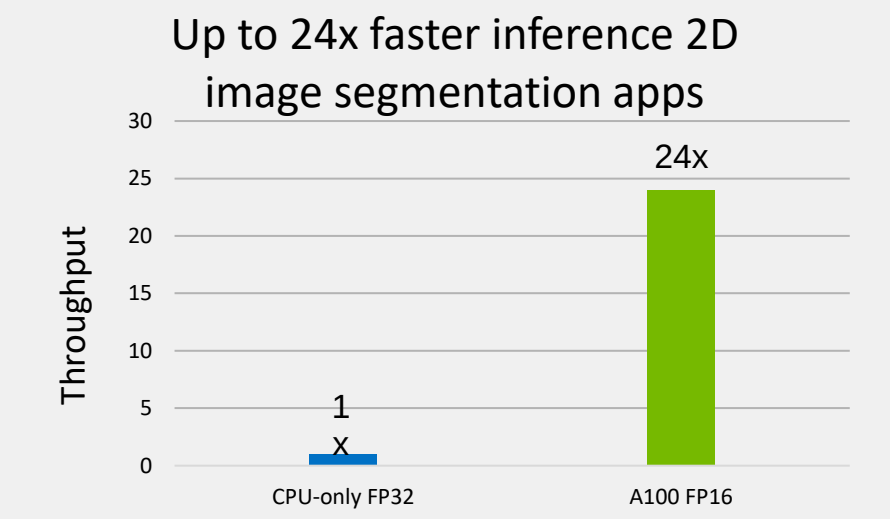
Accelerate 2D UNet based apps common in industrial automation

New resize layer (Upsampling)

Available through new open source ONNX parser

End-to-end workflow sample with pre-trained weights and Jupyter notebook to get started





2D UNet with Industrial, Image size = 512x512
CPU: Skylake Gold 6140, 2.5GHz, Ubuntu 16.04; 18 CPU threads. OpenVINO 2019 R1, BS=1 (error with OpenVINO 2019R2)
Tesla A100; CUDA (450.36); TensorRT 7.1, BS = 1

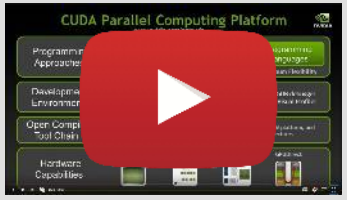
github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/Segmentation/UNet_Industrial

DOWNLOAD TensorRT 7.1 TODAY!

TensorRT DOCUMENTATION

- Installation Guide
- Programming Guide
- API Reference
- Samples

GETTING STARTED RESOURCES



GTC WEBINARS

GitHub

FRAMEWORK INTEGRATIONS

TF
PyTorch
Caffe2
Cognitive Toolkit
Chainer
PaddlePaddle
mxnet
MATLAB

Free download to members of NVIDIA Developer Program soon at
developer.nvidia.com/tensorrt

DEEP STREAM

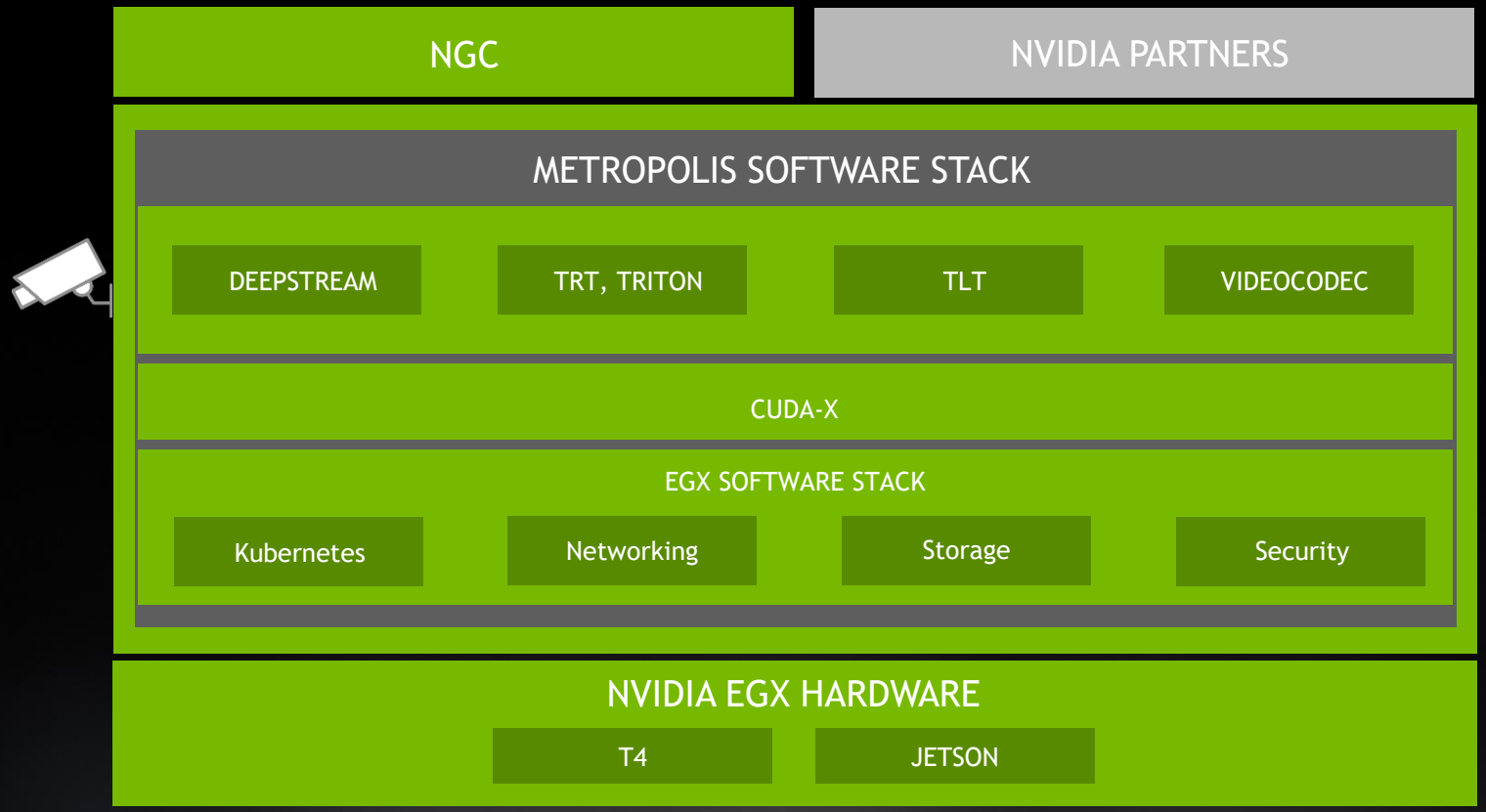
CHALLENGES WITH VIDEO ANALYTICS

Create highly accurate AI

Achieving High Throughput

Deploying at scale

NVIDIA METROPOLIS

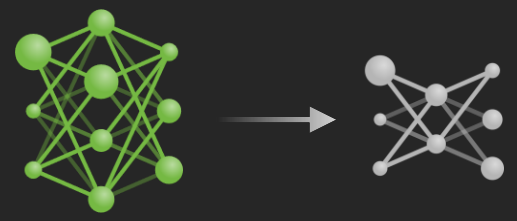


https://www.nvidia.com/en-us/autonomous-machines/intelligent-video-analytics-platform/

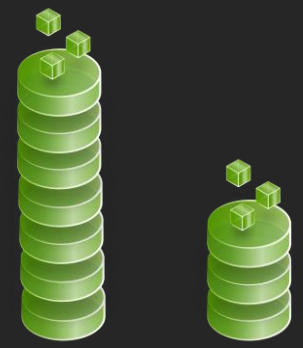
TRAIN WITH TRANSFER LEARNING TOOLKIT

CREATE AI - TRANSFER LEARNING

"Transfer Learning is a process of transferring learned features from one model to another"



Key Benefits

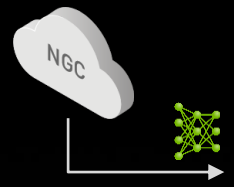


Less Data Required to Train Accurately

Reduce Training Time and Cost

https://blogs.nvidia.com/blog/2019/02/07/what-is-transfer-learning/

NVIDIA TRANSFER LEARNING TOOLKIT (TLT)



New Classes | Model Pruning | Scene Adaptation

TRANSFER LEARNING TOOLKIT

Pre-trained models | Data Preparation & Augmentation | Train | Prune

CUDA-X

NVIDIA Containers RT | CUDA | cuDNN | TensorRT

NVIDIA COMPUTING PLATFORM

TRAINING	INFERENCE
DGX | CLOUD | WORKSTATIONS	JETSON | T4

TRANSFER LEARNING TOOLKIT 2.0

	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	✓	✓	✓	✓	✓	✓	✓	
SqueezeNet	✓	✓		✓	✓	✓	✓	
DarkNet 19/53	✓	✓	✓	✓	✓	✓	✓	

Pre-trained models trained on google open images public dataset
Available to download on ngc.nvidia.com

PURPOSE BUILT PRE-TRAINED NETWORKS

Highly Accurate | Re-Trainable | Out of Box Deployment

PeopleNet

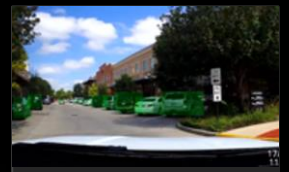
Number of classes: 3
Dataset: 750k frames

84% Accuracy

TrafficCamNet

Number of classes: 4
Dataset: 150k frames

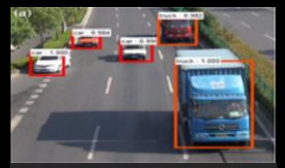
83.5% Accuracy



DashCamNet

Number of Classes: 4
Dataset: 160k frames

80% Accuracy



VehicleTypeNet

Number of classes: 12
Dataset: 56k frames

96% Accuracy

VehicleMakeNet

Number of classes: 20
Dataset: 60k Frames

91% Accuracy

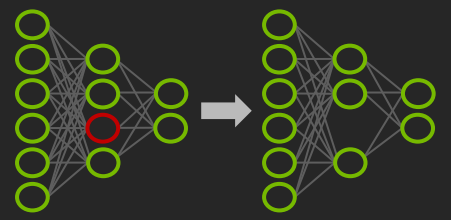
FaceDetect-IR

Number of classes: 1
Dataset: 600k images

96% Accuracy

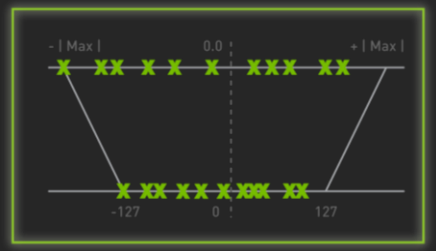
TLT KEY FEATURES

MODEL PRUNING



Reduce memory and increase inference throughput

QUANTIZATION AWARE TRAINING



Improve INT8 accuracy

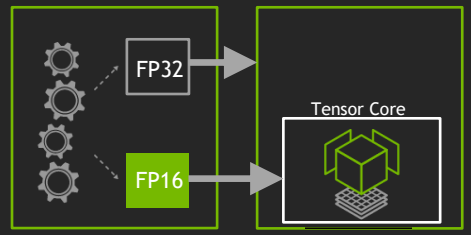
DEPLOYMENT WITH DEEPSTREAM

End-to-end AI application for video analytics

MULTI-GPU TRAINING

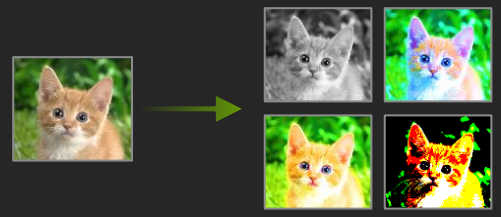
Speedup training time

AUTOMATED MIXED PRECISION (AMP)



Improve training time by using Tensor Cores on GPU

DATA AUGMENTATION



Improve accuracy with color and spatial augmentation

END-TO-END REAL TIME PERFORMANCE

				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	960 x 544	INT8	80%	14	218	72	72	384	94	94	1105
PeopleNet - ResNet34	960 x 544	INT8	84%	10	157	51	51	272	67	67	807
TrafficCamNet - ResNet18	960 x 544	INT8	84%	19	261	105	105	464	140	140	1300
DashCamNet - ResNet18	960 x 544	INT8	80%	18	252	102	102	442	133	133	1280
FaceDetect-IR - ResNet18	384 x 240	INT8	96%	95	1188	570	570	2006	750	750	2520

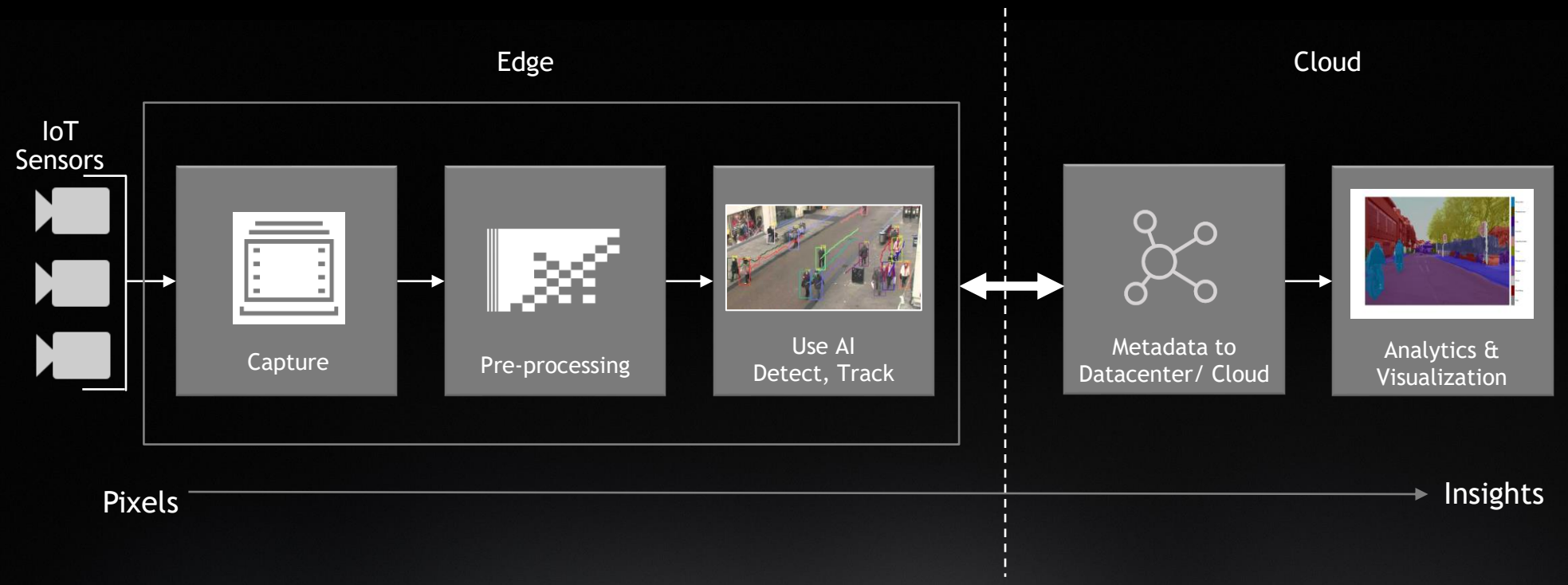
* FP16 inference on Jetson Nano

End-to-end performance using DeepStream SDK

BUILD WITH DEEPSTREAM

DEEPSTREAM - MANY INDUSTRIES, FLEXIBLE DEPLOYMENT

IVA APPLICATION WORKFLOW



DEEPSTREAM SOFTWARE STACK

Applications and Services

- Python
- C/C++

DEEPSTREAM SDK

- Hardware Accelerated Plugins
- Bi-directional IoT Messaging
- OTA model update
- Reference Applications & Helm Charts

CUDA-X

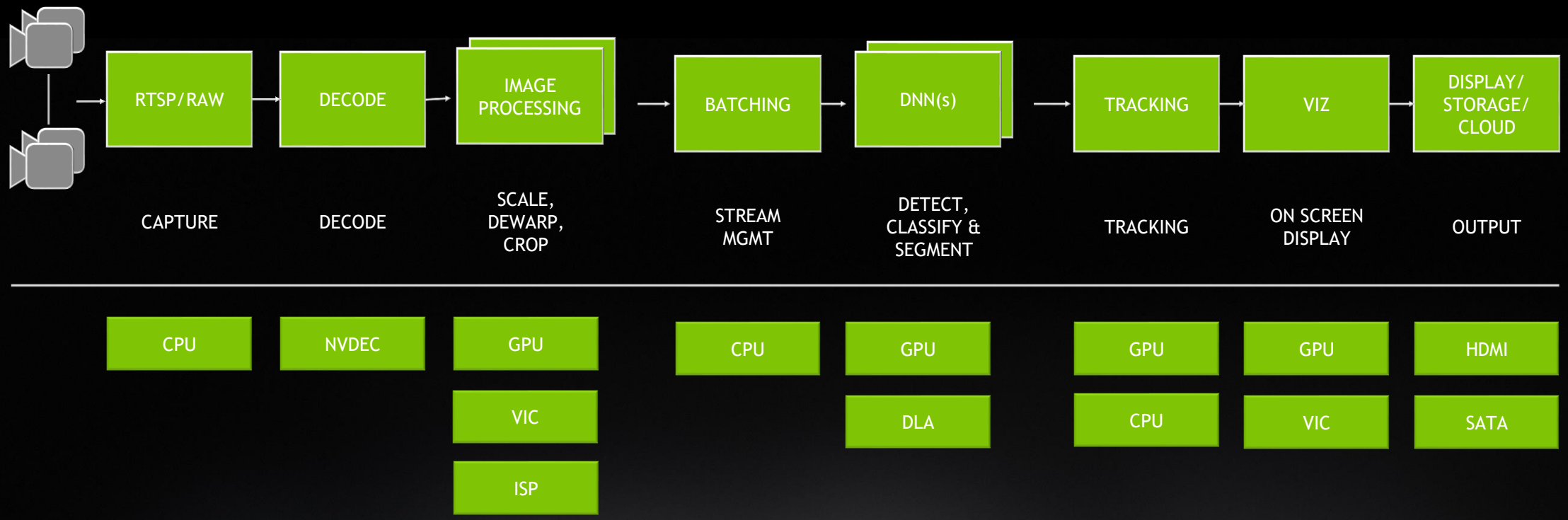
- CUDA
- TensorRT
- Triton Inference Server†
- Multimedia

NVIDIA COMPUTING PLATFORM - EDGE TO CLOUD

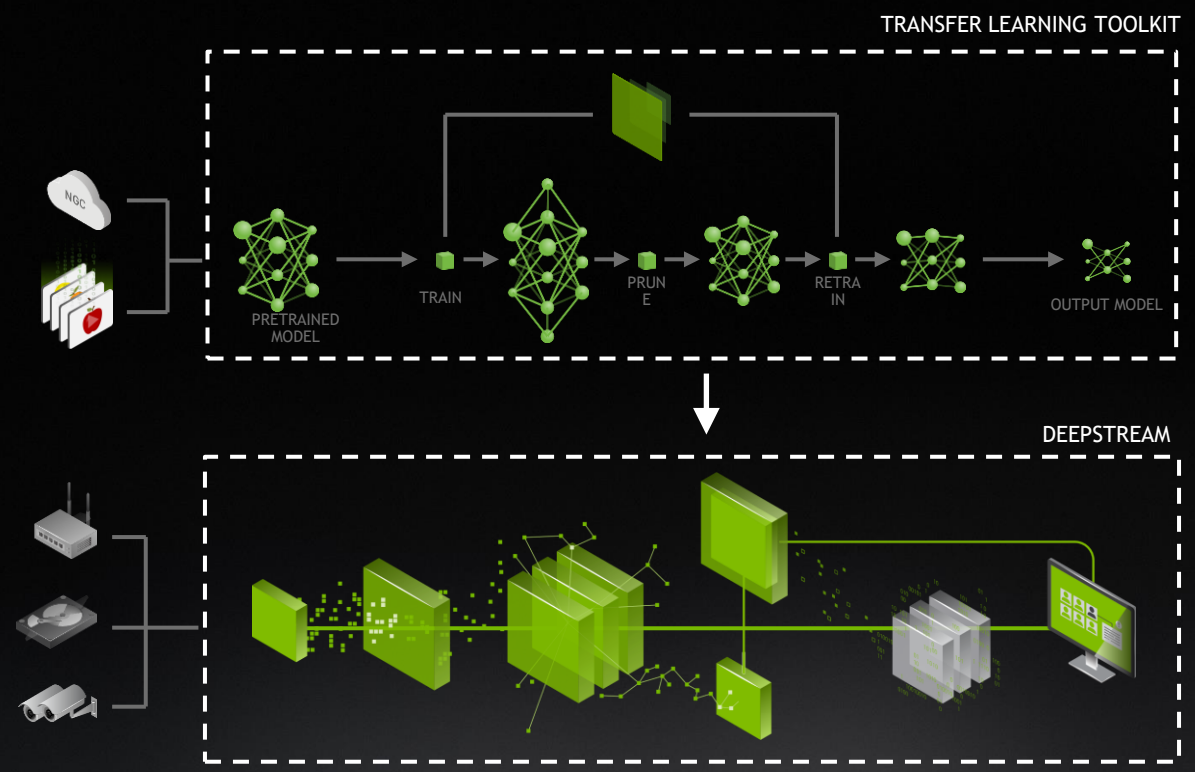
- NVIDIA Containers RT
- JETSON | TESLA | EGX
- Kubernetes ON GPUs

† - Formerly TensorRT Inference Server

DEEPSTREAM GRAPH ARCHITECTURE

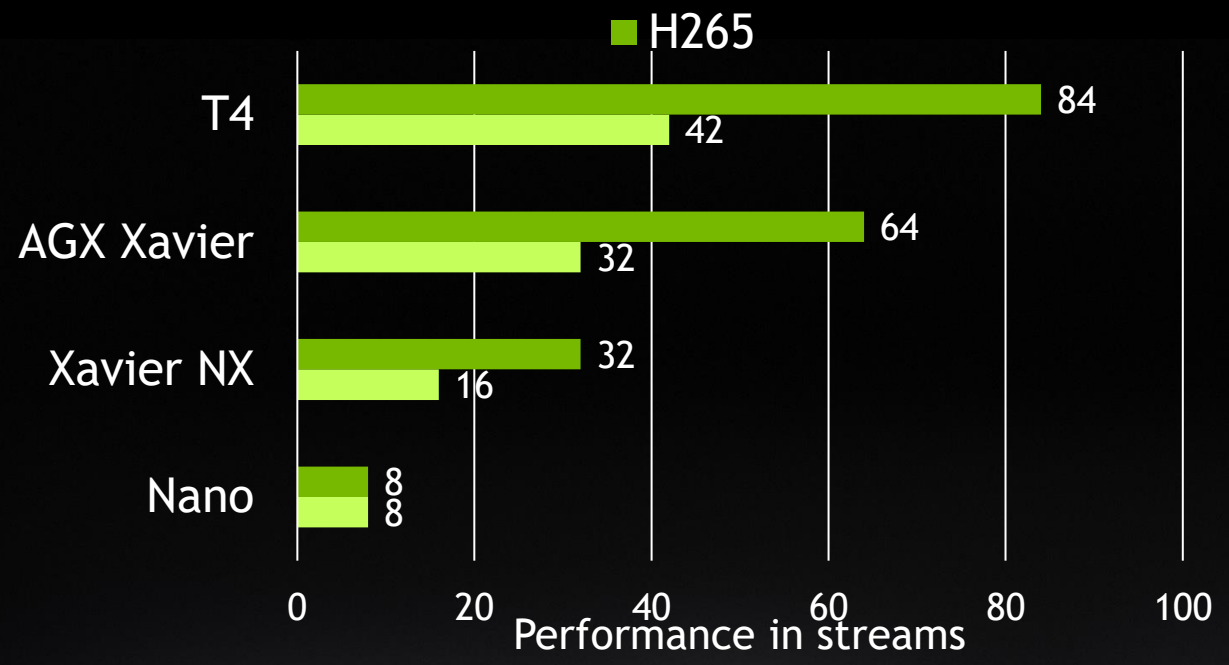


END-TO-END DEEP LEARNING WORKFLOW



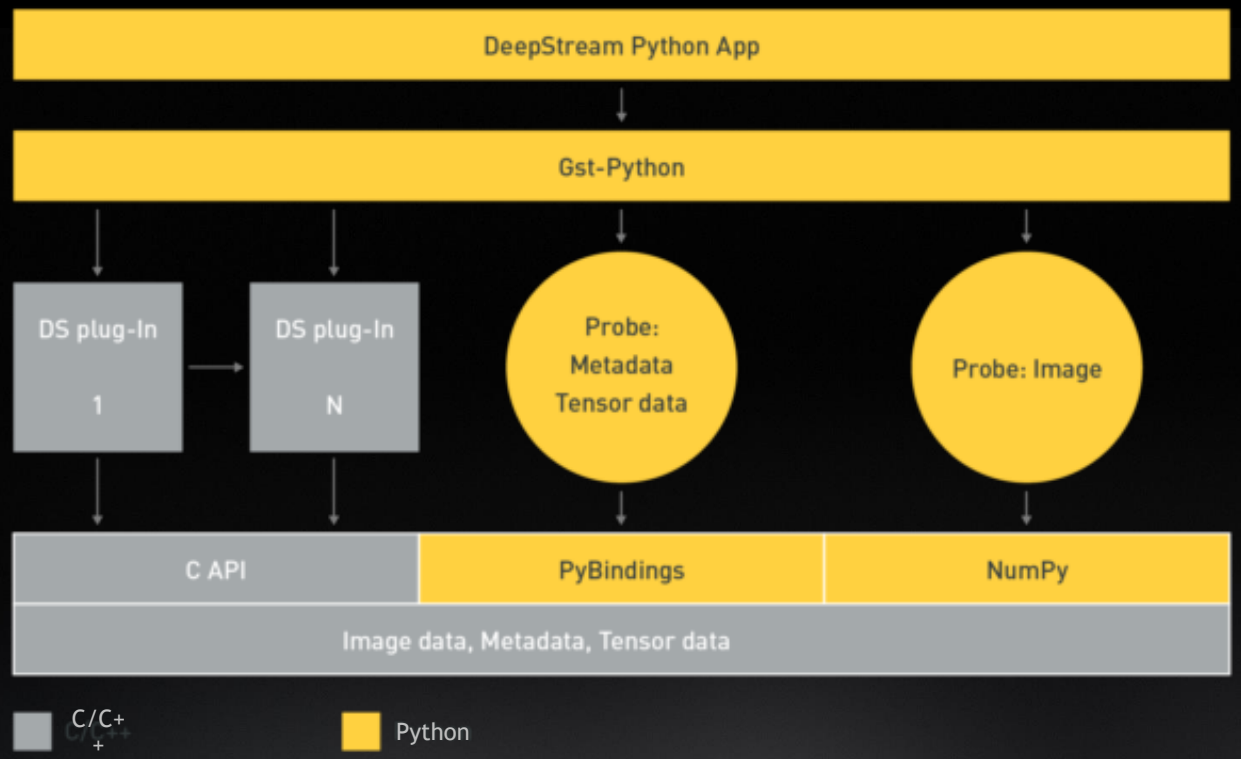
ACHIEVING REAL-TIME PERFORMANCE

Number of 1080p, 30fps streams processed with DeepStream



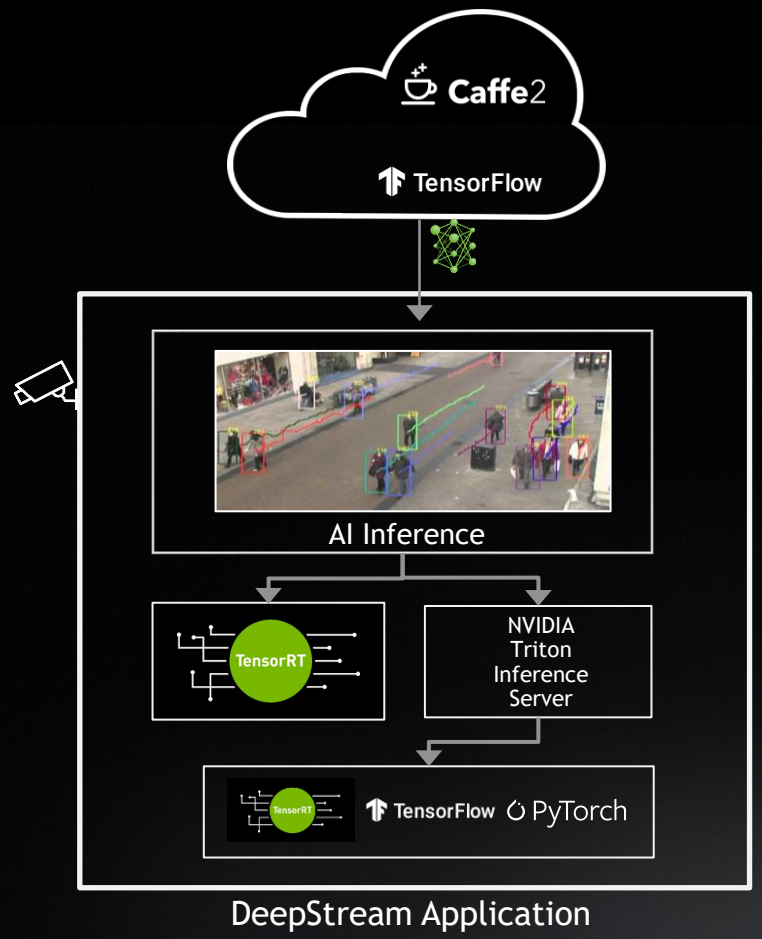
Data generated using DeepStream reference app
Full performance data in DeepStream performance documentation
Watch the performance optimization video tutorial

PYTHON SUPPORT



https://github.com/NVIDIA-AI-IOT/deepstream_python_apps

DEEPSTREAM WITH TRITON INFERENCE SERVER



	TensorRT	Triton Inference Server
Pros	Highest Throughput	Highest flexibility
Cons	Custom layers require writing plugins	Less performant than a TensorRT solution

GETTING STARTED WITH DEEPSTREAM

GETTING STARTED APPLICATIONS

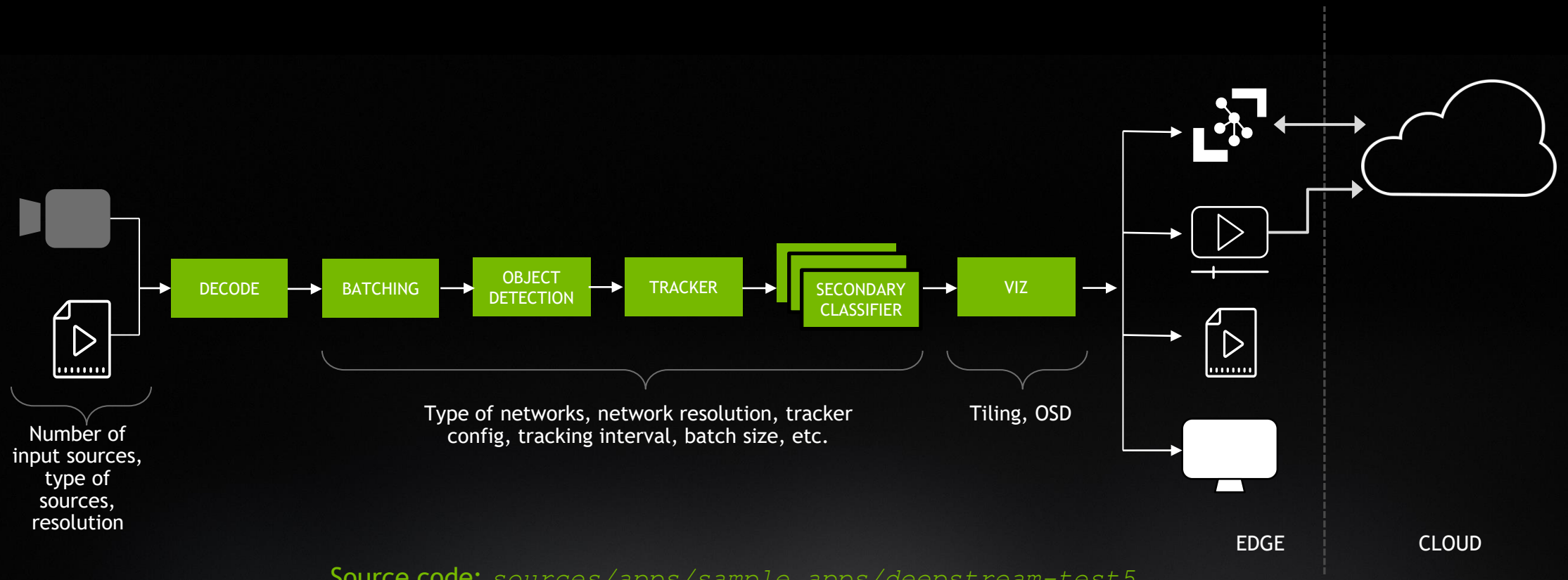
Available in C and Python

Name	Function	
deepstream-test1	DeepStream Hello world. Single video from file to on screen display with bounding box	Decode → Batching → Object detection
deepstream-test2	Builds on test1 and adds secondary object classification on detected objects	••• Object detection → Classific-ation
deepstream-test3	Builds on test1 and adds multiple video inputs	Decode → Batching → Object detection
deepstream-test4	Builds on test1 and adds connections to IoT services thru the nvmsgbroker plugin	••• Object detection → Message Converte r → Message Broker

C/C++ apps | Python apps

END-TO-END DEEPSTREAM APP

DeepStream-test5

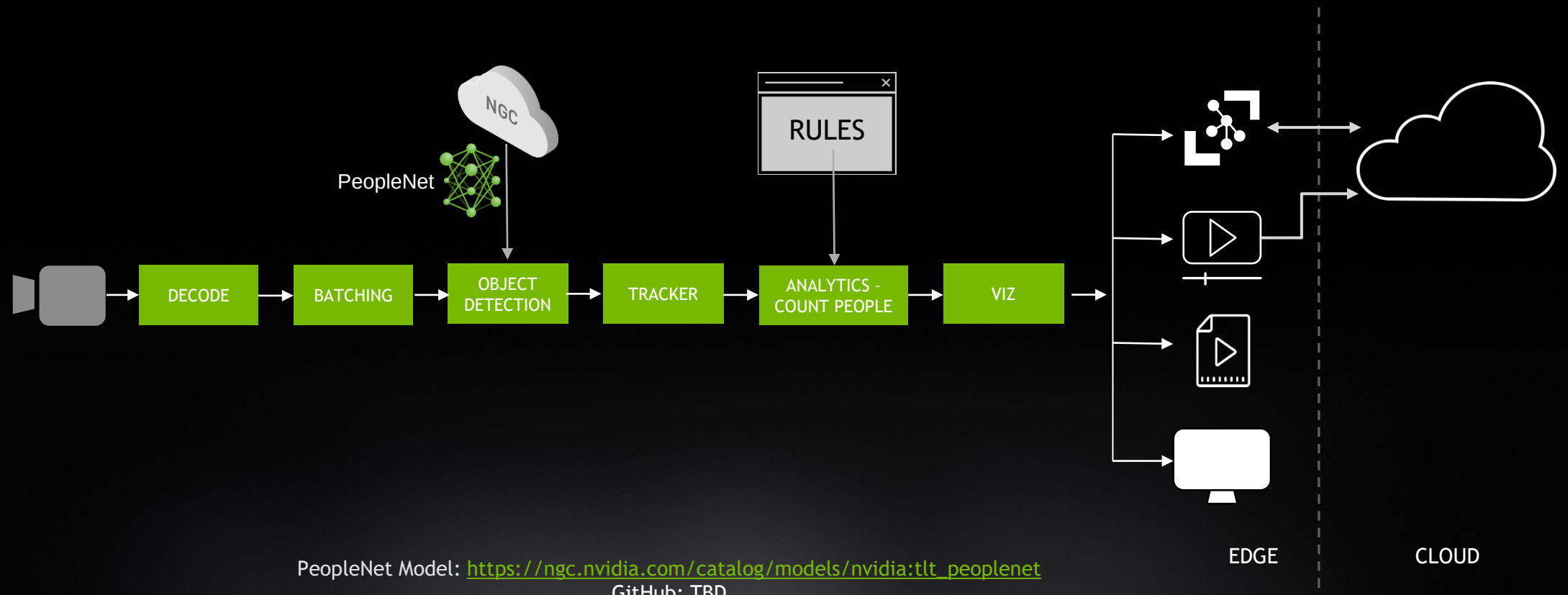


Source code: `sources/apps/sample_apps/deepstream-test5`

Python app coming soon

END-TO-END APPS

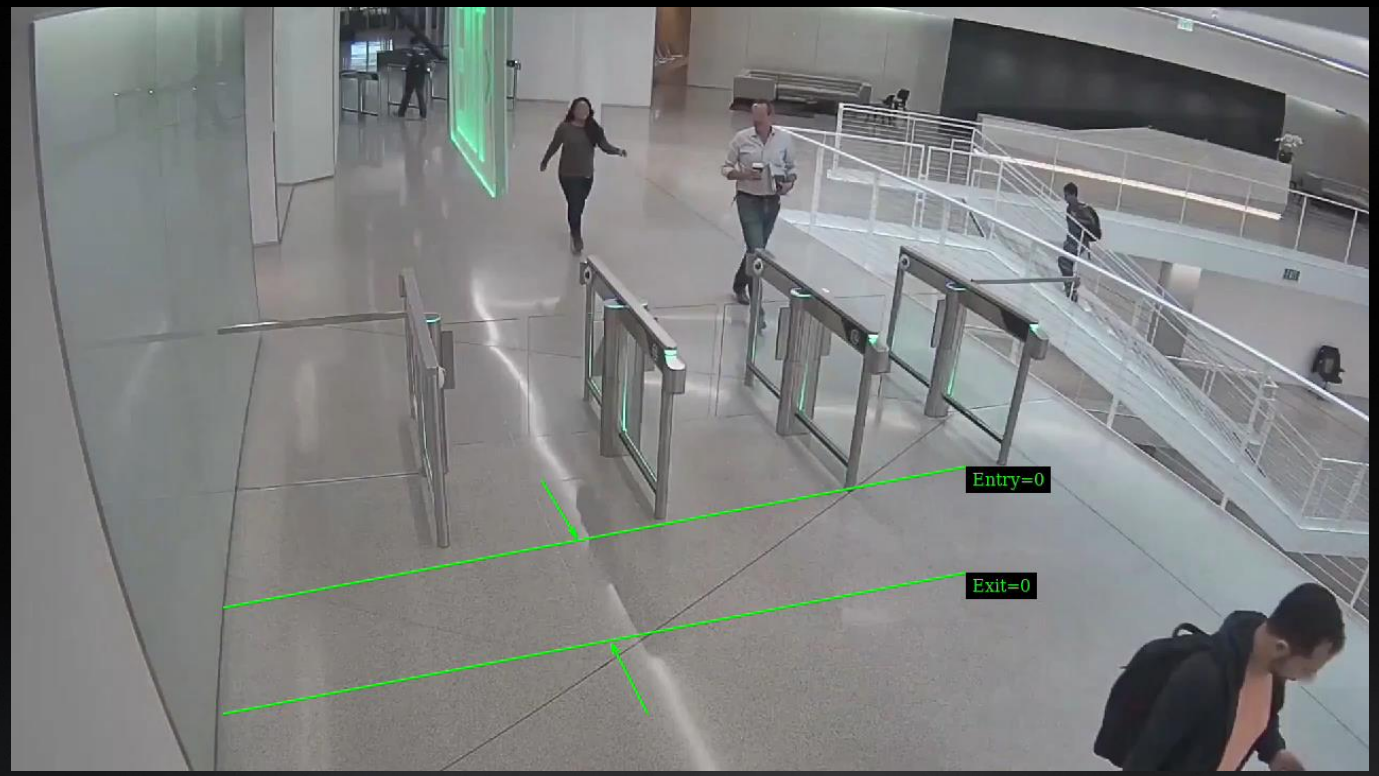
DEEPSTREAM APPLICATION



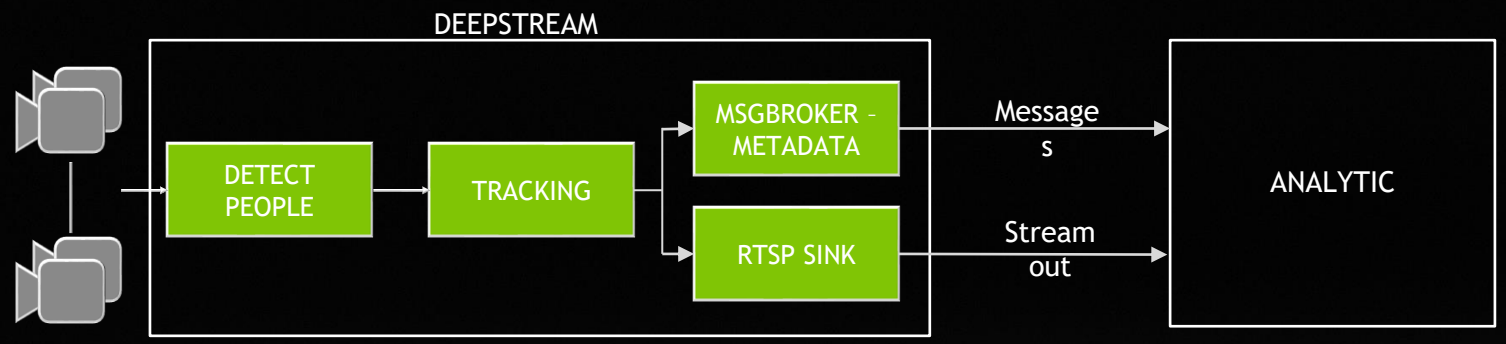
PeopleNet Model: https://ngc.nvidia.com/catalog/models/nvidia:tlt_peoplenet
GitHub: TBD

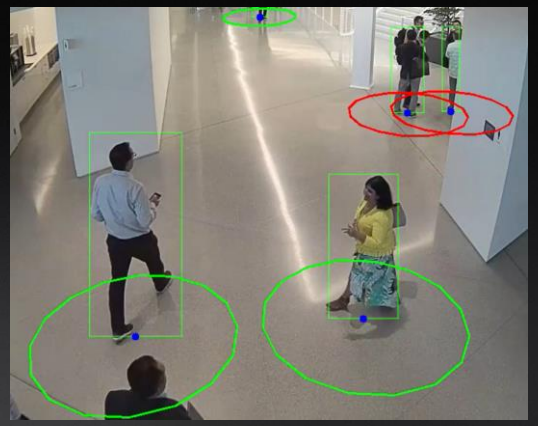
PEOPLE COUNTING

Demo

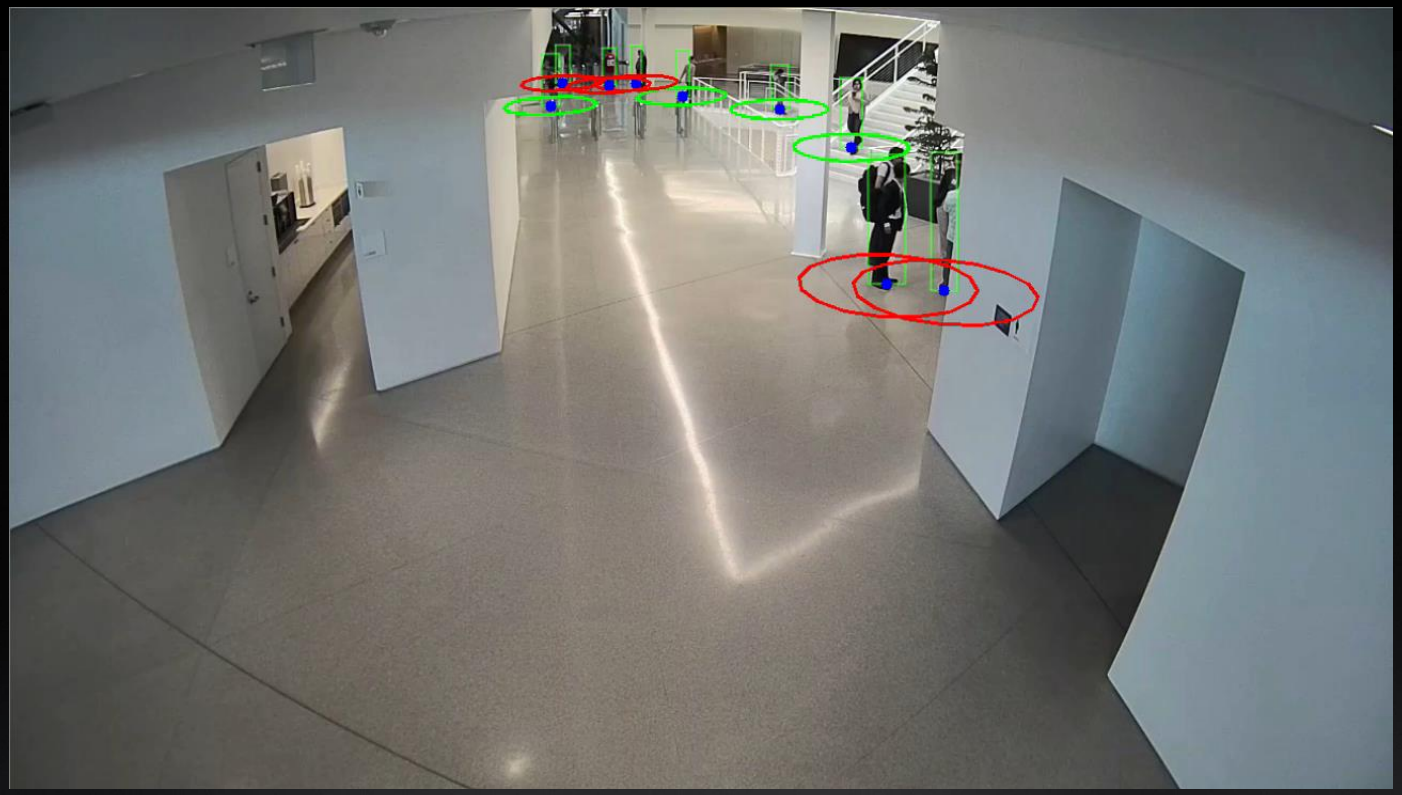


SOCIAL DISTANCING APP



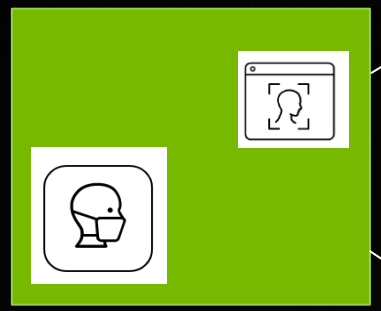


SOCIAL DISTANCING APP



FACE MASK DETECTION

Jupyter notebook, developer recipe to build with an open source dataset



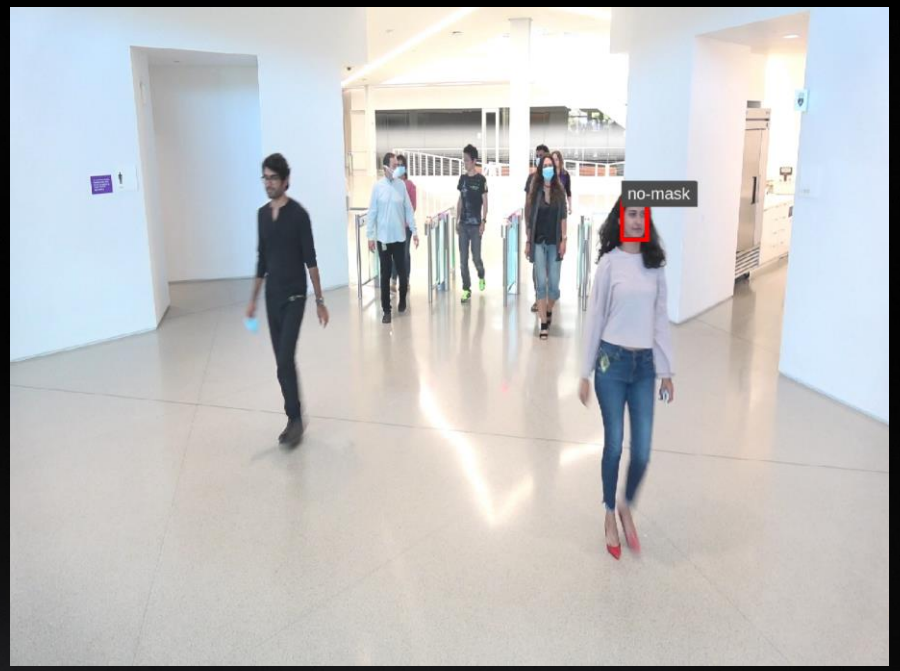
Use Cases

Hospitals
Workspaces
Mass Transit Hubs
Stadiums
Warehouses

Key Components

Transfer Learning Toolkit
DeepStream SDK

GitHub Repo

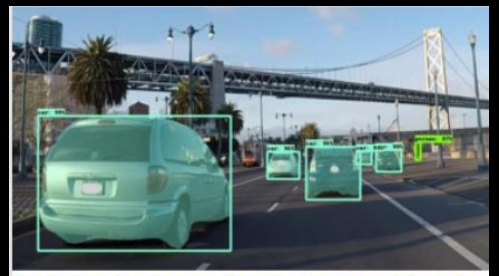


Developer Blog
GitHub Repo

What this project does not provide:

- Trained model for face-mask detection
- NVIDIA specific dataset for faces with and without mask

NEW DEVELOPER CONTENT



Training Instance Segmentation Models Using Mask R-CNN on the NVIDIA Transfer Learning Toolkit

Tutorial

Building Intelligent Video Analytics Apps Using NVIDIA DeepStream 5.0 (Updated for GA)

Developer blog

Deploying Real-time Object Detection Models with the NVIDIA Isaac SDK and NVIDIA Transfer Learning Toolkit

Tutorial

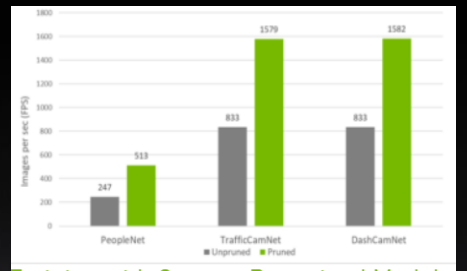
Building a Real-time Redaction App Using NVIDIA DeepStream, Part 1: Training

Tutorial

Improving INT8 Accuracy Using Quantization Aware Training and the NVIDIA Transfer Learning Toolkit

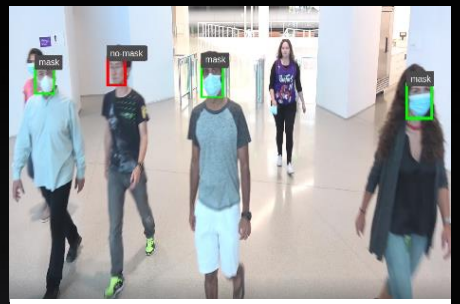
Tutorial

Video Tutorial



Training with Custom Pretrained Models Using the NVIDIA Transfer Learning Toolkit

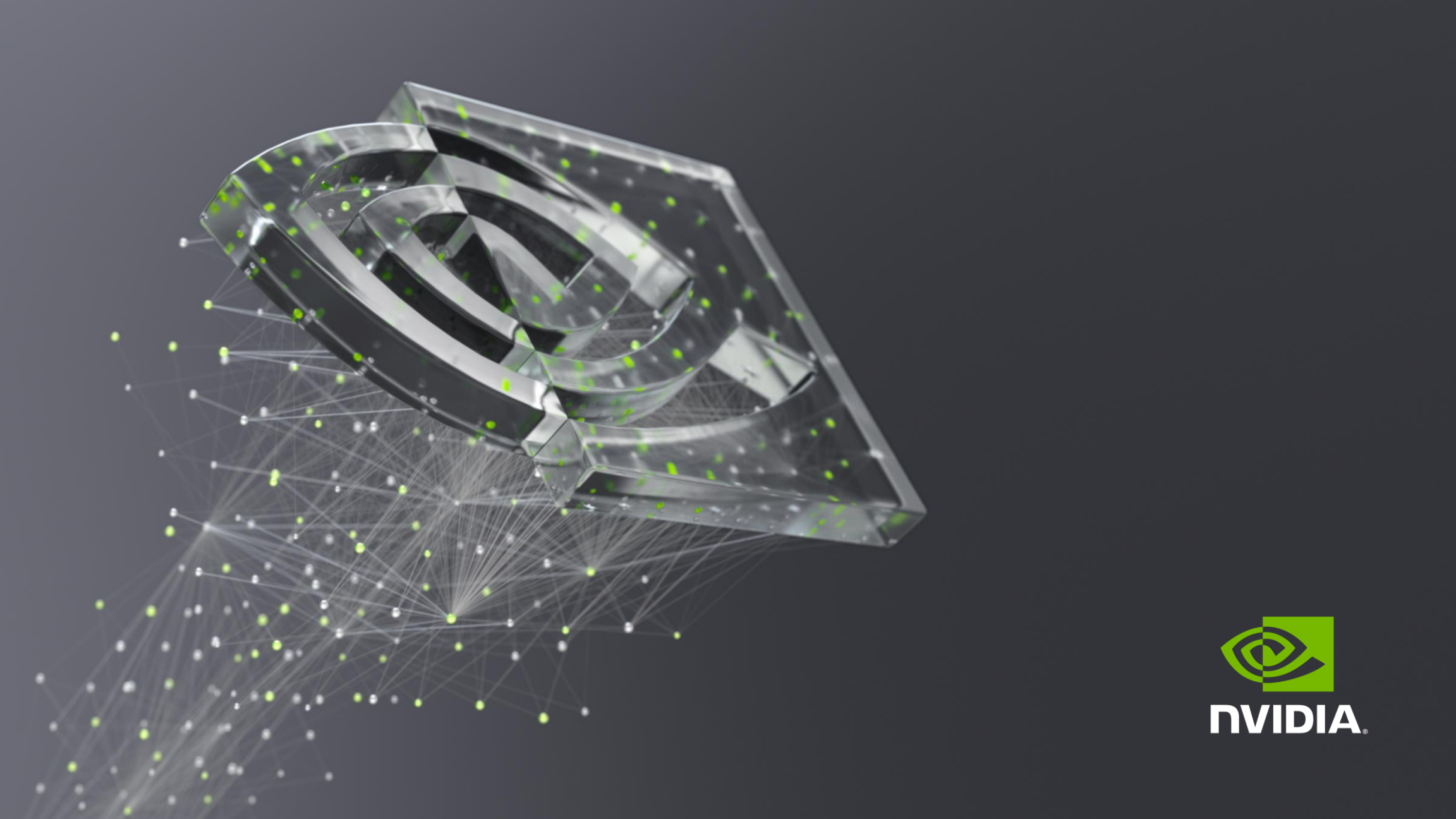
Tutorial



Implementing real-time AI-based face mask detection

Tutorial

Enroll in the NVIDIA Developer Program to get the latest developer updates


NVIDIA