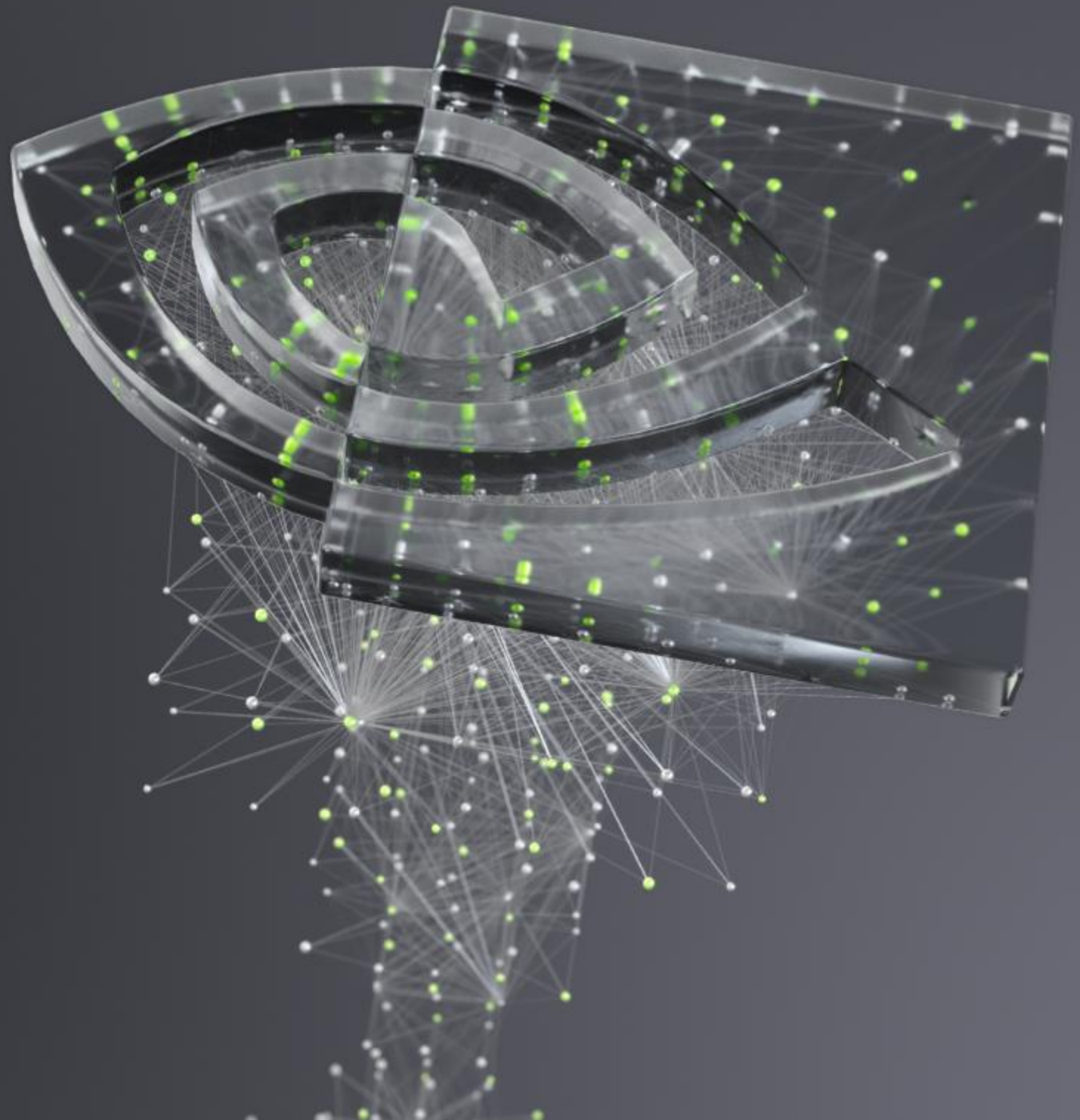


# METROPOLIS SOFTWARE PLATFORM

# AI REVOLUTION

## Measuring Utilization of Space is More Critical Than Ever

Traffic Management

Retail

Factories

Public Safety & Health

Warehouse & Logistics

Stadiums & Casinos

Access Control

Transportation Hubs

**$2T INDUSTRY** — Increase operational efficiency and safety across many industries using AI

# CHALLENGES WITH VIDEO ANALYTICS

Create highly accurate AI

Achieving High Throughput

Deploying at scale

# NVIDIA METROPOLIS

NGC

NVIDIA PARTNERS

METROPOLIS SOFTWARE STACK

DEEPSTREAM | TRT, TRITON | TLT | VIDEOCODEC

CUDA-X

EGX SOFTWARE STACK

Kubernetes | Networking | Storage | Security

NVIDIA EGX HARDWARE

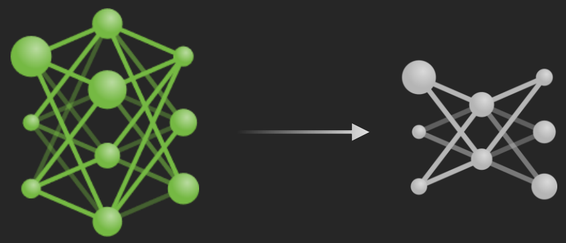
T4 | JETSON

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

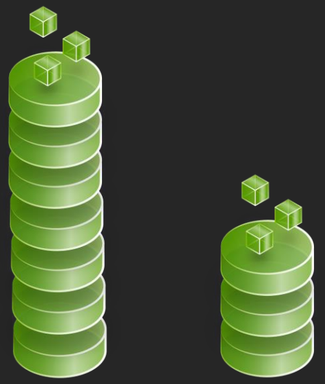
# TRAIN WITH TRANSFER LEARNING TOOLKIT

# CREATE AI - TRANSFER LEARNING

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



## Key Benefits

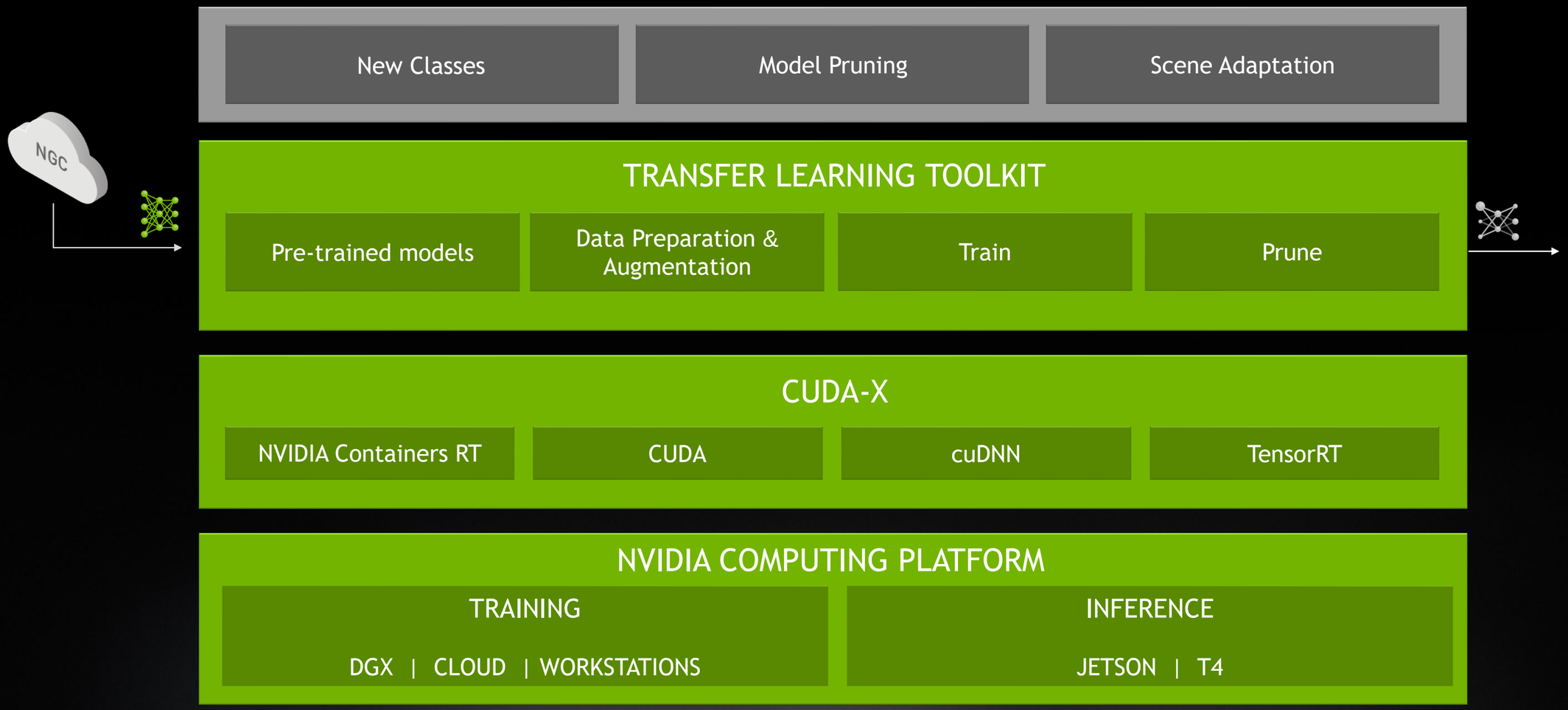


Less Data Required to Train Accurately


Reduce Training Time and Cost

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

# NVIDIA TRANSFER LEARNING TOOLKIT (TLT)



# 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

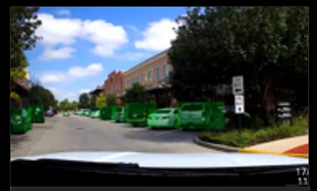
Number of classes: 3
Dataset: 750k frames

**84% Accuracy**

TrafficCamNet

Number of classes: 4
Dataset: 150k frames

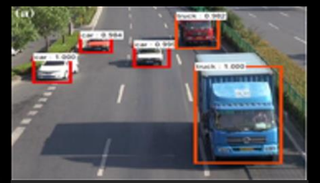
**83.5% Accuracy**



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**

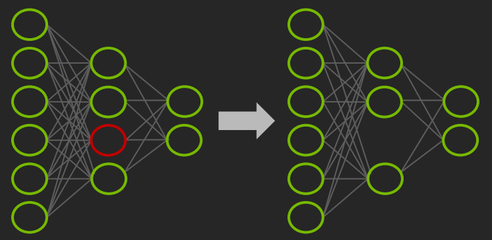
FaceDetect-IR

Number of classes: 1
Dataset: 600k images

**96% Accuracy**

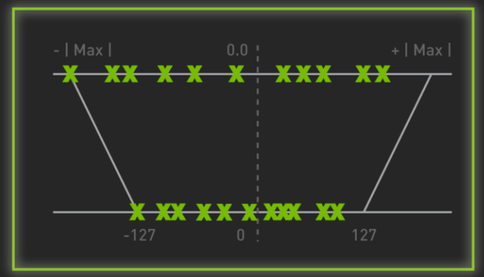
# TLT KEY FEATURES

## MODEL PRUNING



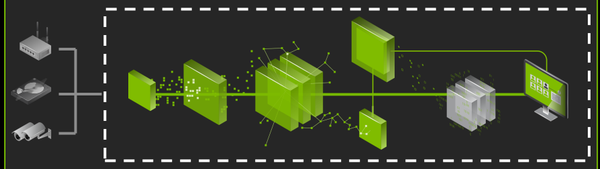
Reduce memory and increase inference throughput

## QUANTIZATION AWARE TRAINING



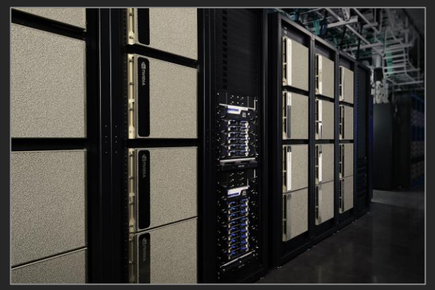

Improve INT8 accuracy

## DEPLOYMENT WITH DEEPSTREAM



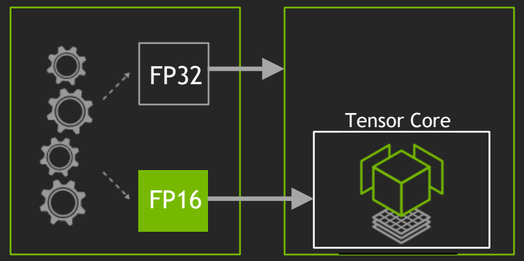
End-to-end AI application for video analytics

## MULTI-GPU TRAINING



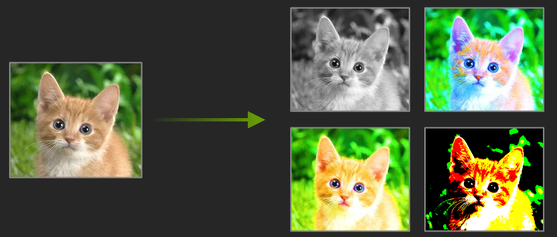
Speedup training time

## AUTOMATED MIXED PRECISION (AMP)




Improve training time by using Tensor Cores on GPU

## DATA AUGMENTATION

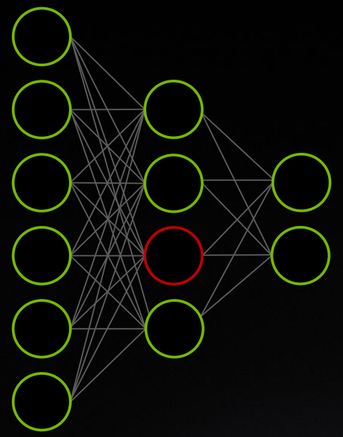


Improve accuracy with color and spatial augmentation

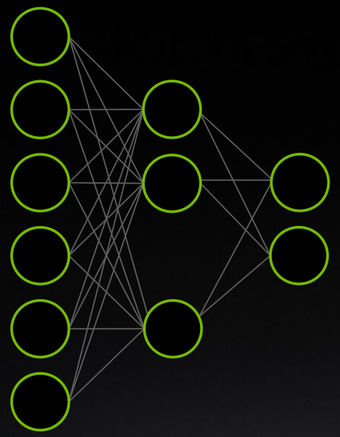
# MODEL PRUNING

## 2 Step Process

1. Reduce model size
2. Incrementally retrain model after pruning to recover accuracy

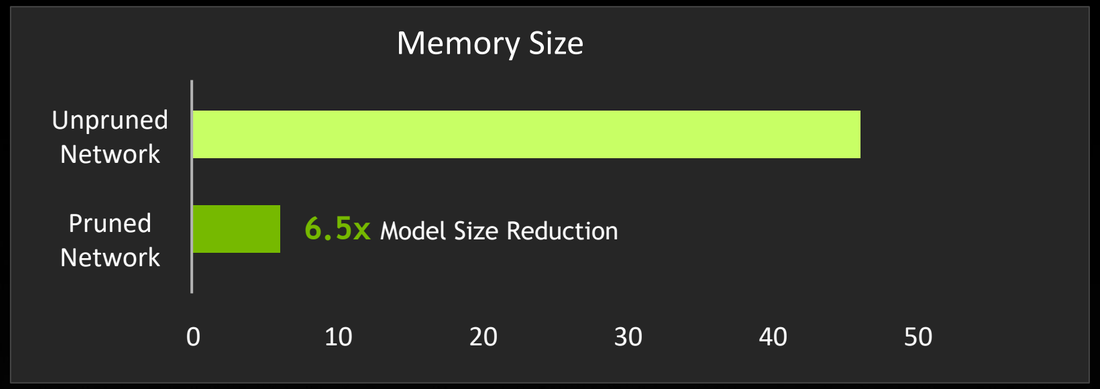


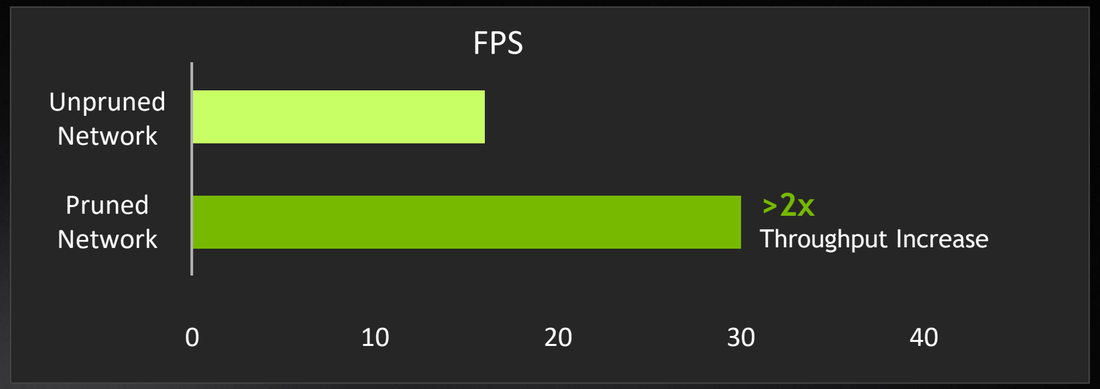
6 inputs, 6 neurons, 32 connections



6 inputs, 5 neurons, 24 connections

## Network - ResNet18 4-class

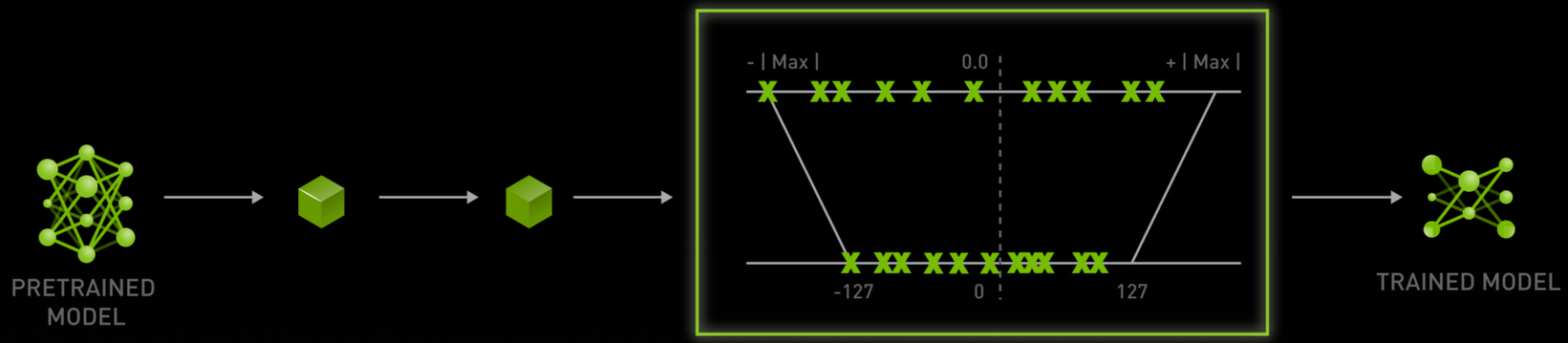





NVIDIA

# QUANTIZATION AWARE TRAINING




QAT workflow:

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

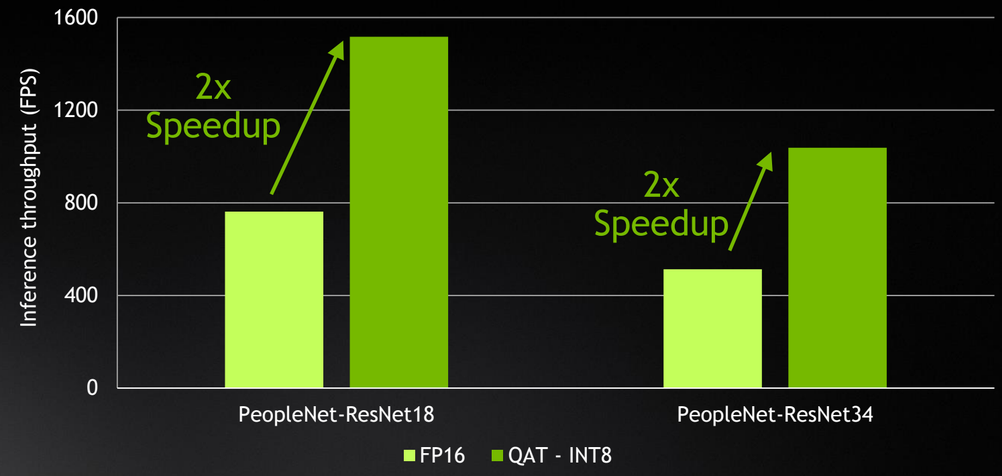
QAT supported on all object detection models

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 |

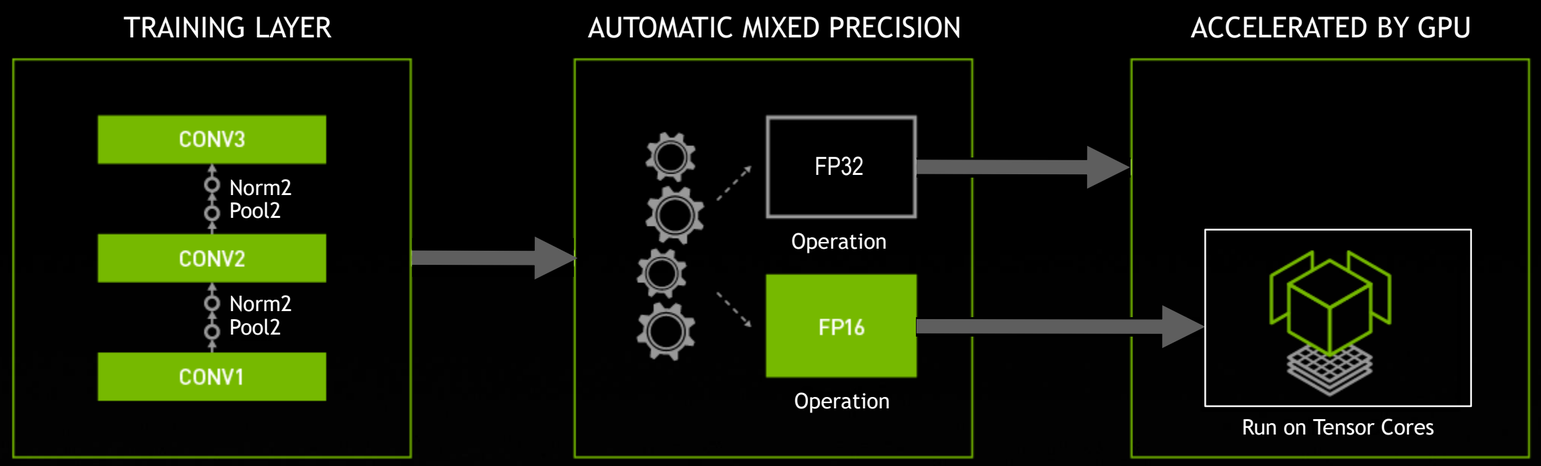
PTQ - Post-training quantization
QAT - Quantization-aware training

Inference Performance




Developer Tutorial on using QAT with TLT

# AUTOMATIC MIXED PRECISION (AMP)

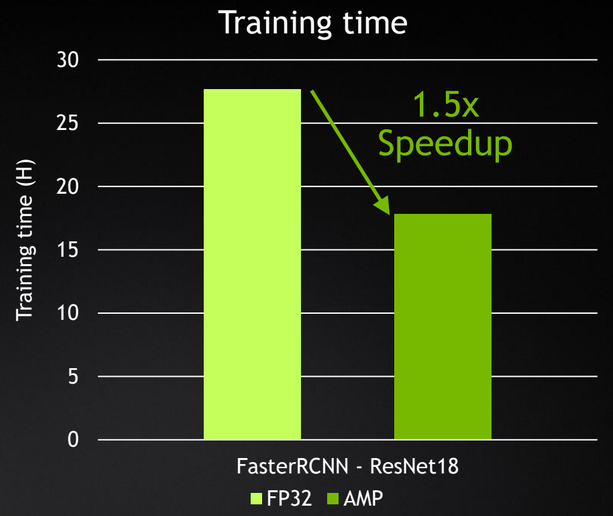



AMP - Train with half-precision while maintaining network accuracy as with single precision

Natively supported with TLT to take advantage of Tensor Cores

Benefits:

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



# INSTANCE SEGMENTATION – MASK R-CNN

Pixel level accuracy vs. an approximate bbox

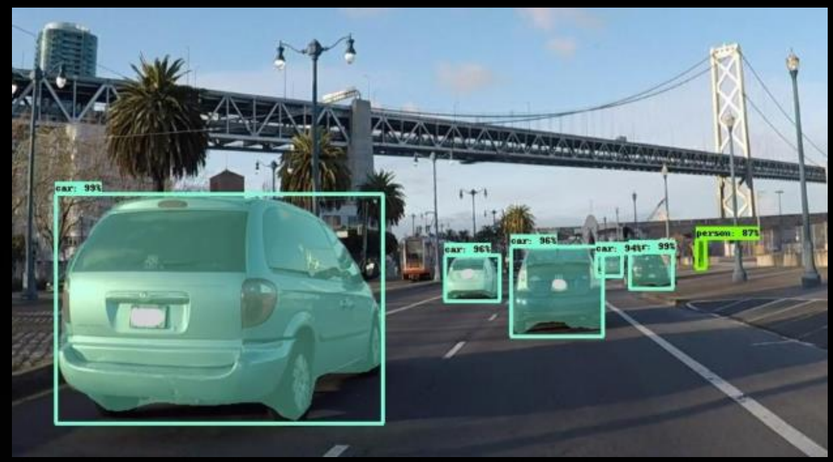
Annotation using COCO format

Model deployment with DeepStream

Comparable accuracy and performance compare to open source

Pre-trained model for ResNet10/18/50/101 on NGC

Developer tutorial on training Mask R-CNN model with TLT



# 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 |

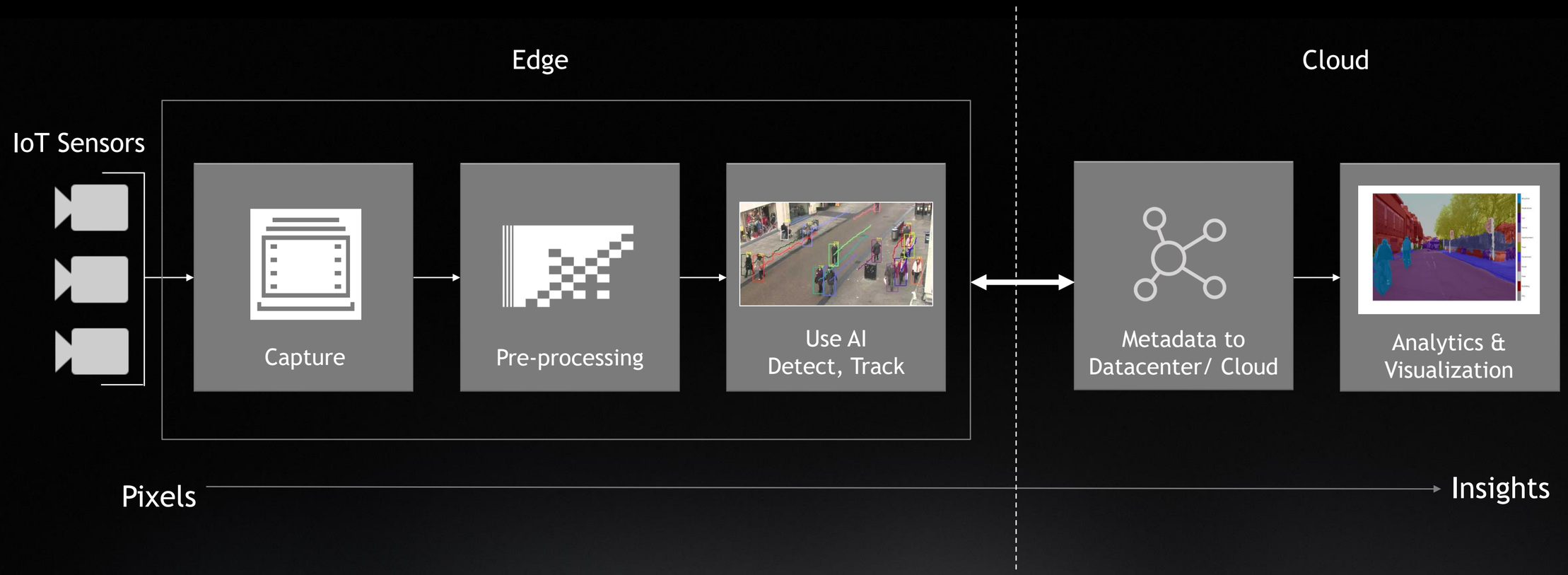
* FP16 inference on Jetson Nano

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

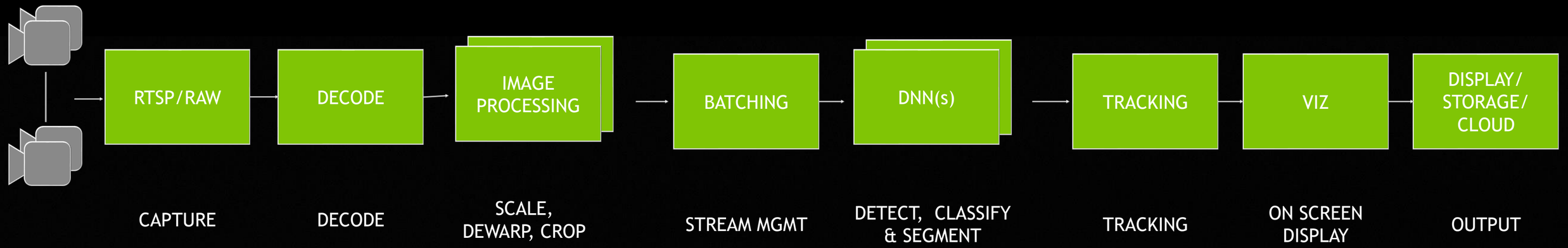
CUDA | TensorRT | Triton Inference Server† | Multimedia

NVIDIA COMPUTING PLATFORM - EDGE TO CLOUD

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

† - Formerly TensorRT Inference Server

# DEEPSTREAM GRAPH ARCHITECTURE



# DEEPSTREAM 5.0 KEY FEATURES

## DEVELOP IN PYTHON

Flexibility of building apps in Python with comparable performance

## INTEGRATION WITH TRITON

Deploy model natively in TF, TF-TensorRT, PyTorch, or ONNX

## RHEL SUPPORT

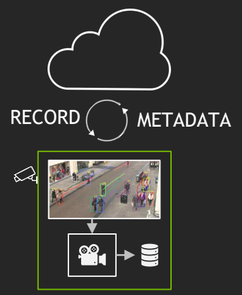
Red Hat Enterprise Linux

Build and deploy DeepStream apps natively using RHEL

## IOT

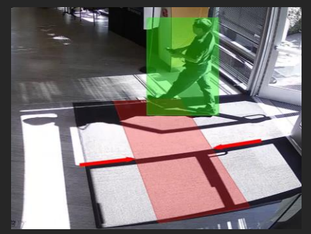
Bi-directional messaging, OTA model update, Security

## SMART RECORD




Save valuable disk space by doing selective record

## PEOPLE ANALYTICS

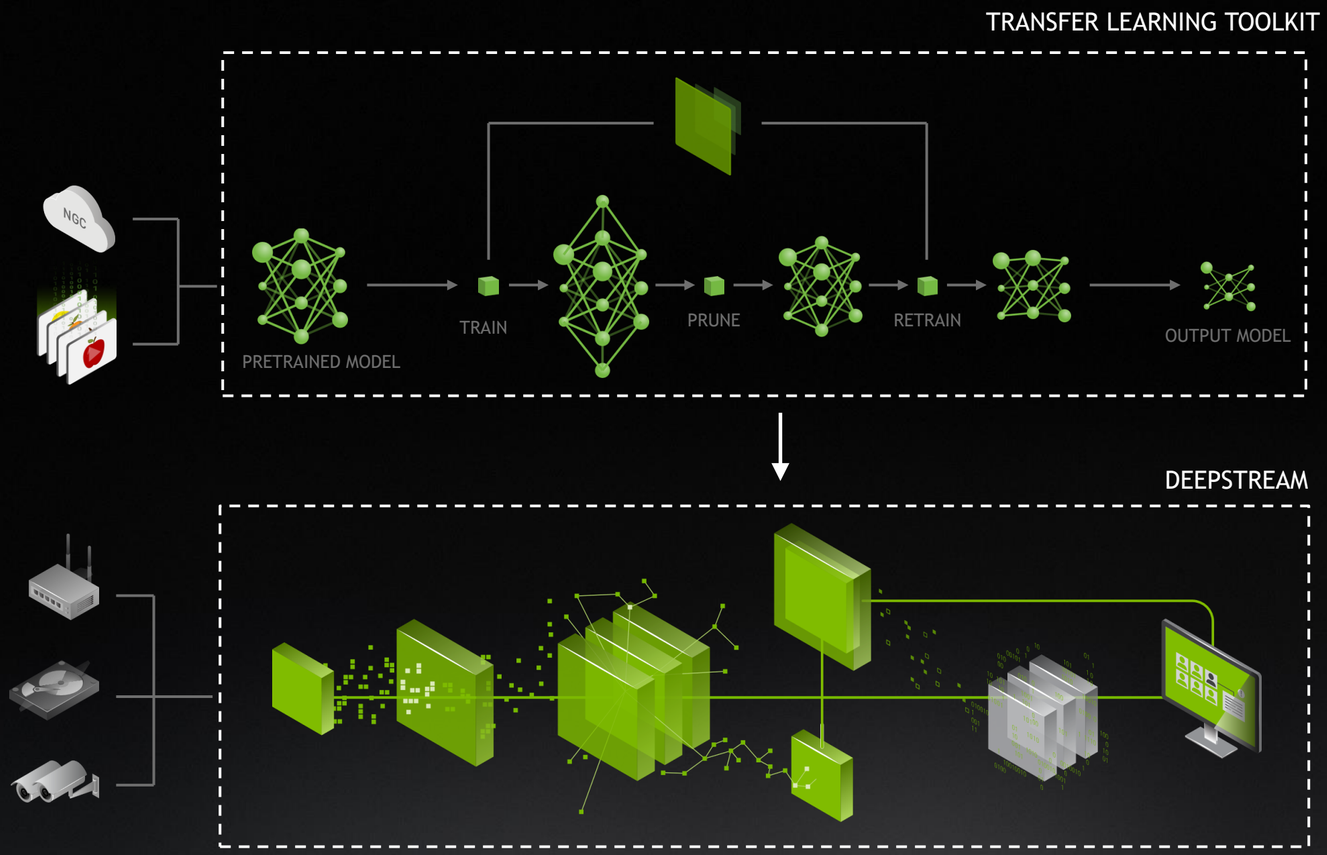


Perform ROI based filtering, line crossing, direction detection

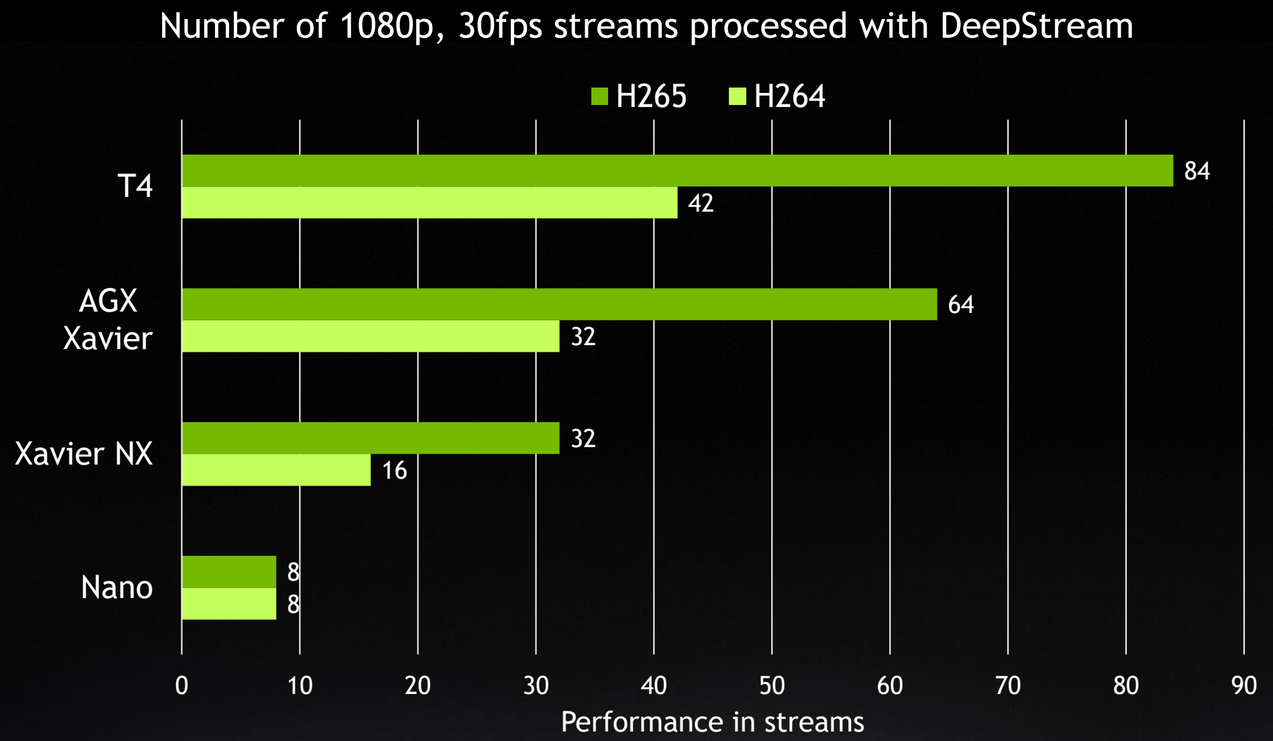


NVIDIA

# END-TO-END DEEP LEARNING WORKFLOW



# ACHIEVING REAL-TIME PERFORMANCE

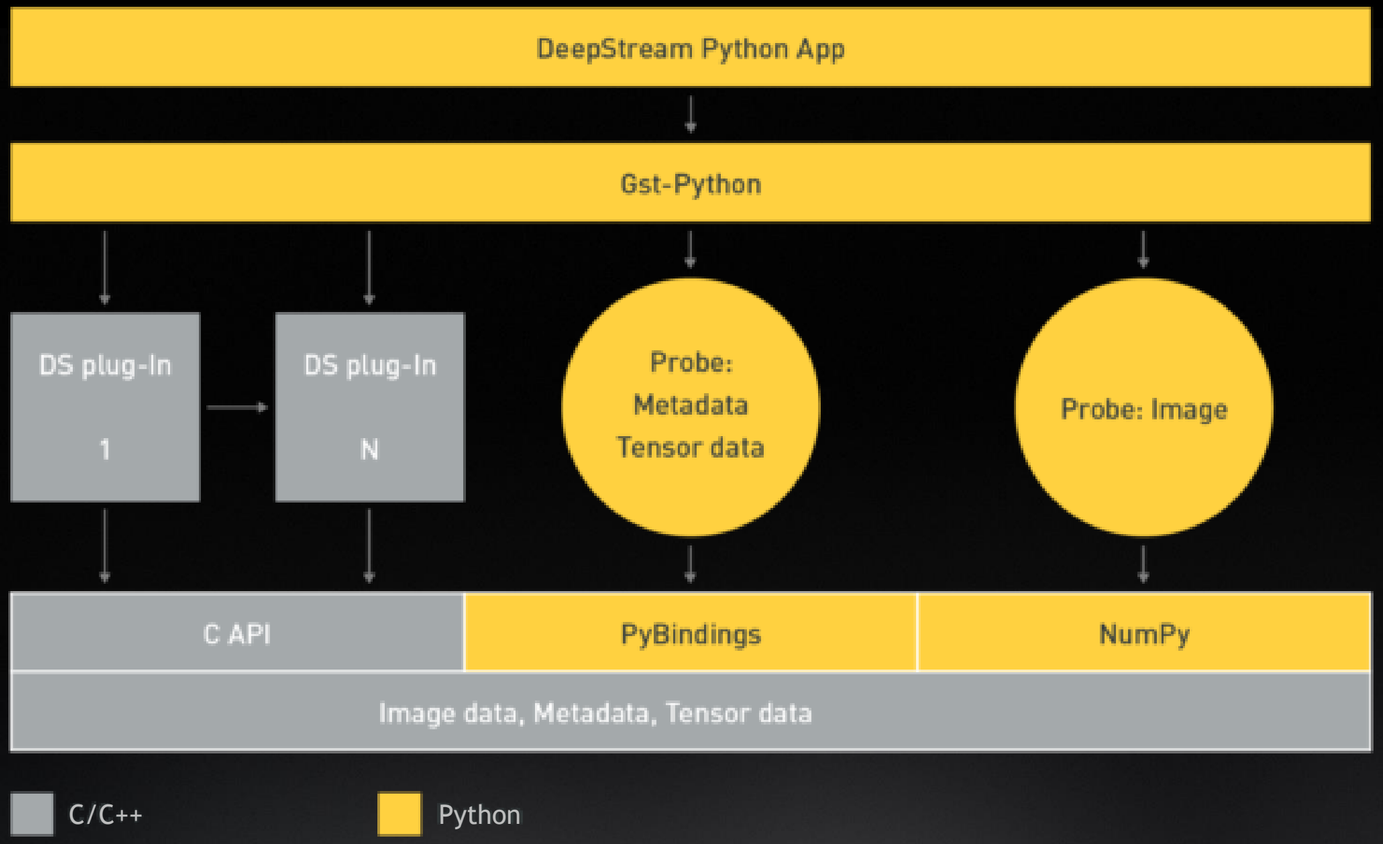



Data generated using DeepStream reference app

Full performance data in DeepStream performance documentation

Watch the performance optimization video tutorial

# PYTHON SUPPORT

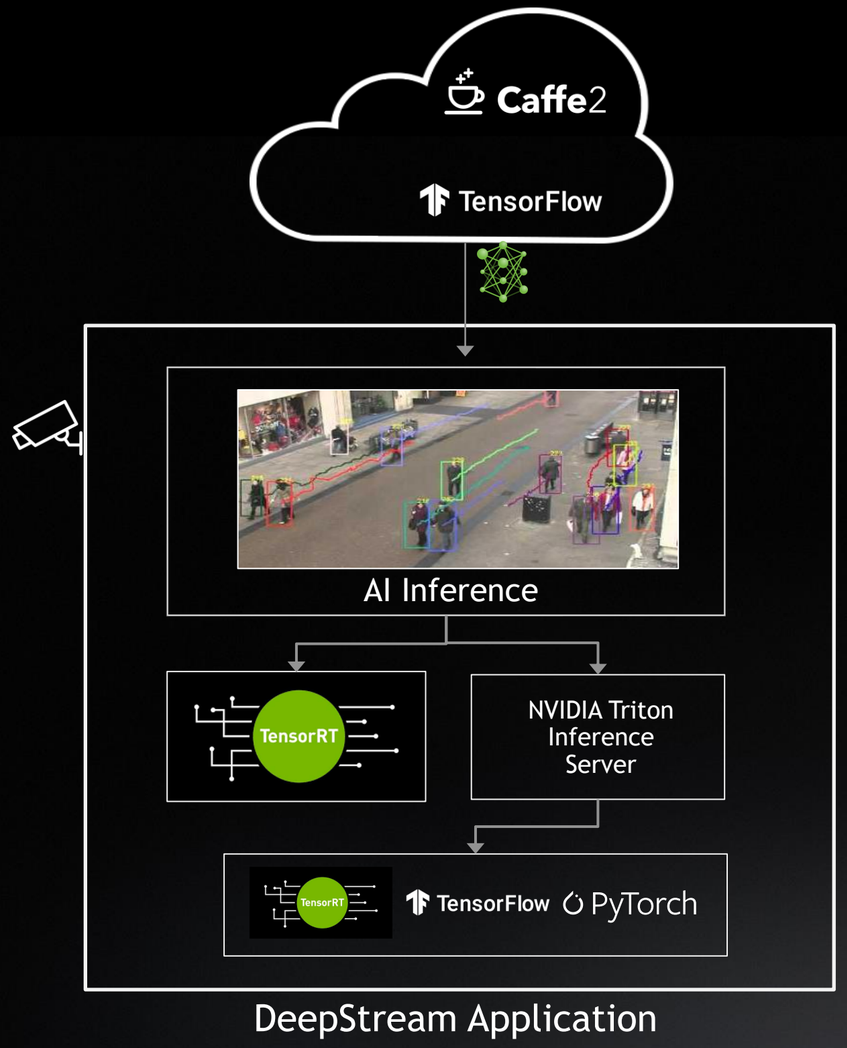



https://github.com/NVIDIA-AI-IOT/deepstream_python_apps

# DEEPSTREAM ACCELERATED PLUGINS

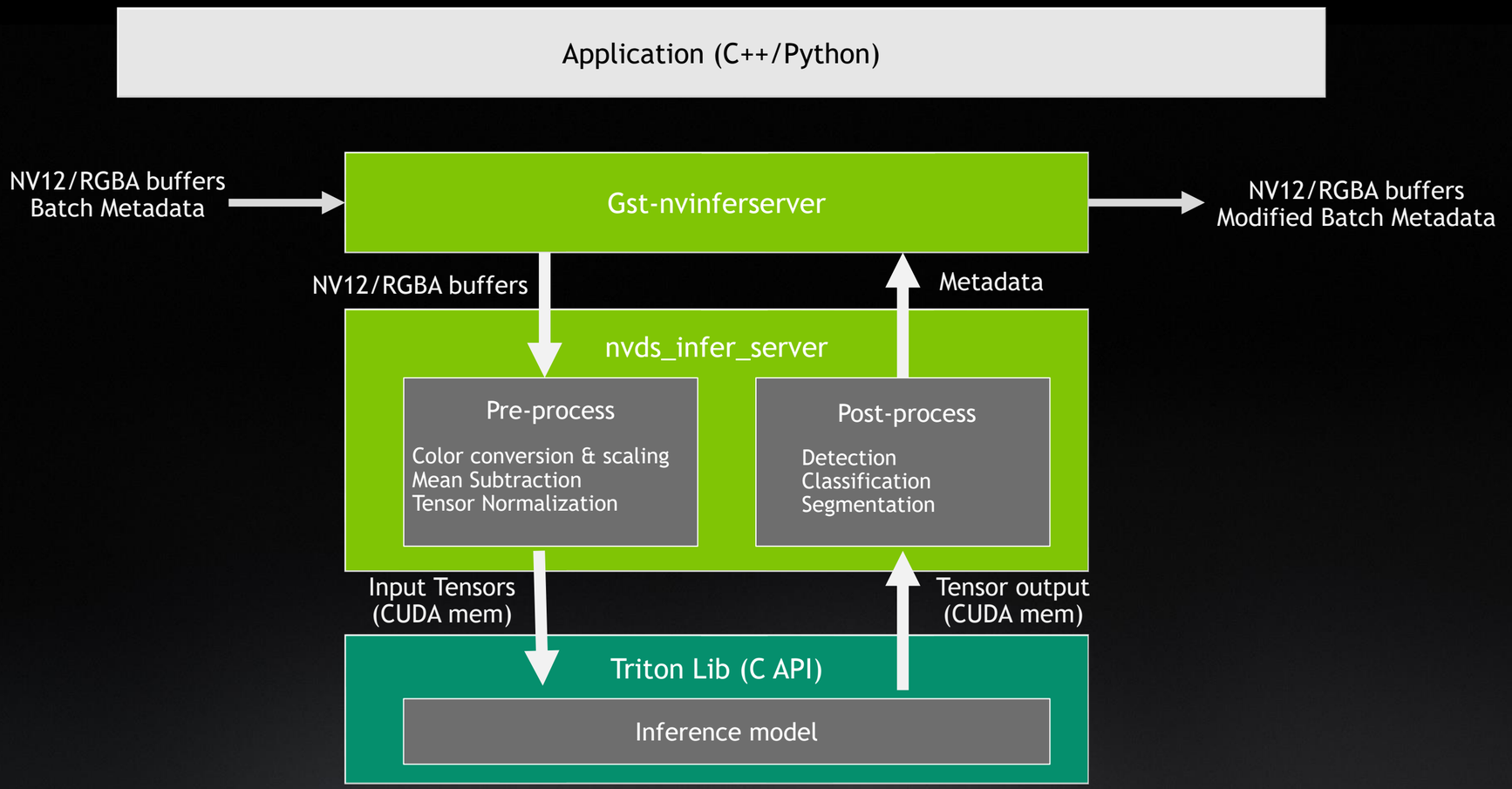
| Plugin Name | Functionality |
|---|---|
| Gst-nvvideo4linux2 | Hardware accelerated decode and encode |
| Gst-nvinfer | DL inference for detection, classification and segmentation |
| Gst-nvinferserver | Plugin to run inference with Triton inference server from DS graph |
| Gst-nvtracker | Reference object trackers; KLT, IOU, NvDCF |
| Gst-nvmsgconv/nvmsgbroker | Metadata generation and messaging to cloud |
| Gst-nvstreammux/nvstreamdemux | Stream aggregation, multiplexing, demuxing, and batching |
| Gst-nvdsosd | Draw boxes and text overlay |
| Gst-nvmultistreamtiler | Renders frames 2D grid array |
| Gst-nveglglessink | Accelerated X11 / EGL rendering |
| Gst-nvvideoconvert | Scaling, format conversion, rotation |
| Gst-nvdewarp | Dewarping for fish-eye degree cameras |
| Gst-nvsegvisual | Visualizes segmentation results |
| Gst-nvof/nvofvisual | Hardware accelerated optical flow and visualization |
| Gst-nvdsanalytics | Perform analytics like ROI, filtering, overcrowding detection |
| Gst-nvjpegdec | Hardware accelerated JPEG decode |

# DEEPSTREAM WITH TRITON INFERENCE SERVER

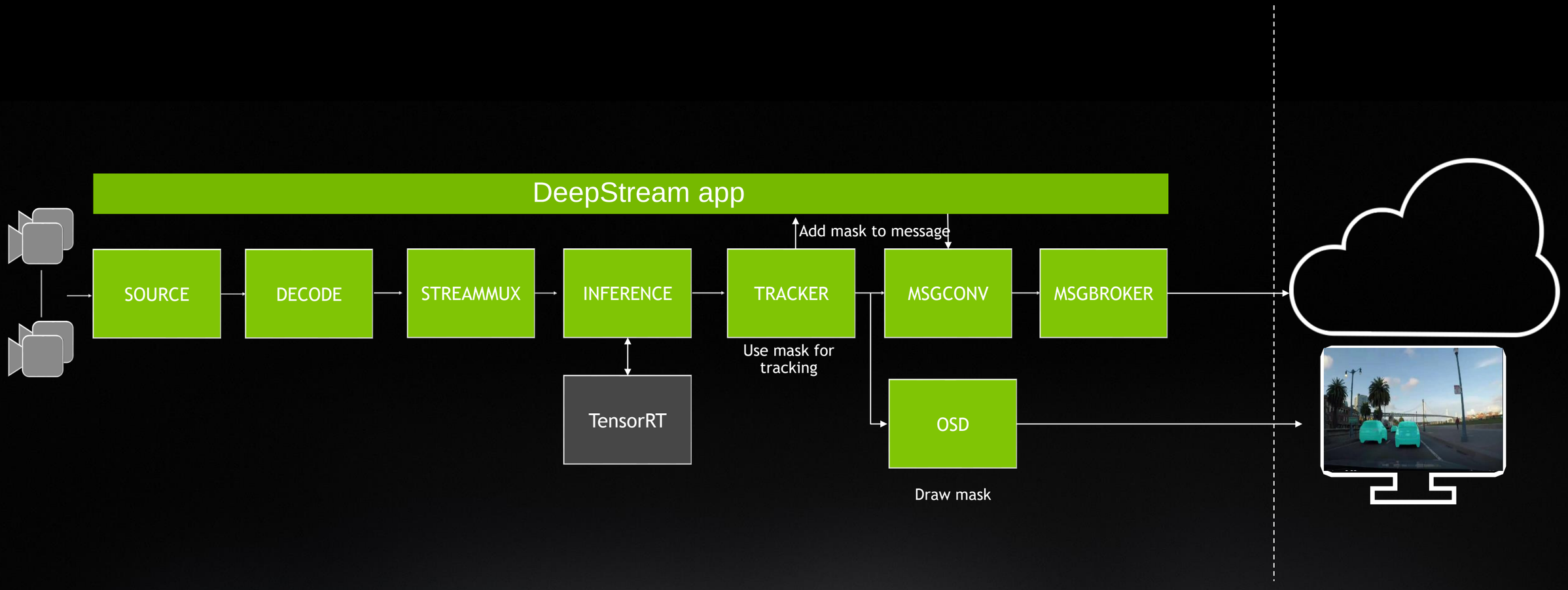



| | TensorRT | Triton Inference Server |
|---|---|---|
| Pros | Highest Throughput | Highest flexibility |
| Cons | Custom layers require writing plugins | Less performant than a TensorRT solution |

# TRITON – DEEPSTREAM ARCHITECTURE



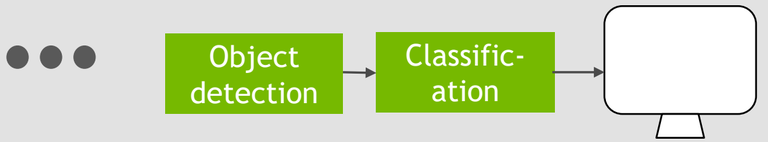
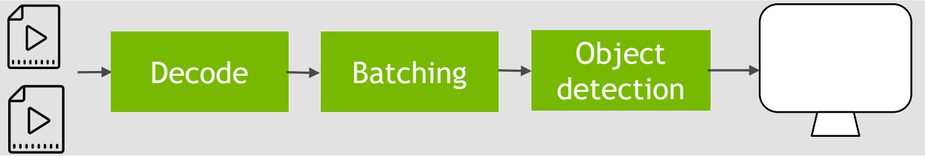
# DEPLOYING MASK R-CNN APPLICATION



# GETTING STARTED WITH DEEPSTREAM

# GETTING STARTED APPLICATIONS

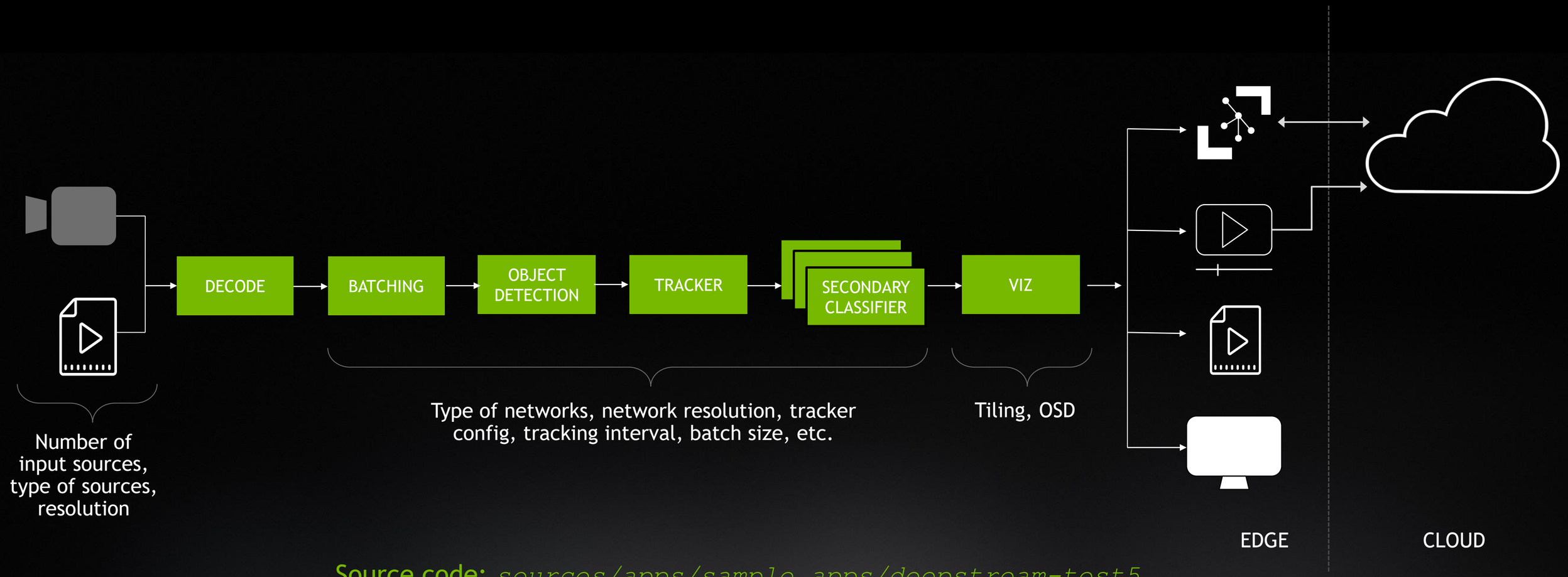
## Available in C and Python

| Name | Function | |
|---|---|---|
| deepstream-test1 | DeepStream Hello world. Single video from file to on screen display with bounding box |   |
| deepstream-test2 | Builds on test1 and adds secondary object classification on detected objects |   |
| deepstream-test3 | Builds on test1 and adds multiple video inputs |   |
| deepstream-test4 | Builds on test1 and adds connections to IoT services thru the nvmsgbroker plugin |   |

C/C++ apps | Python apps

# END-TO-END DEEPSTREAM APP

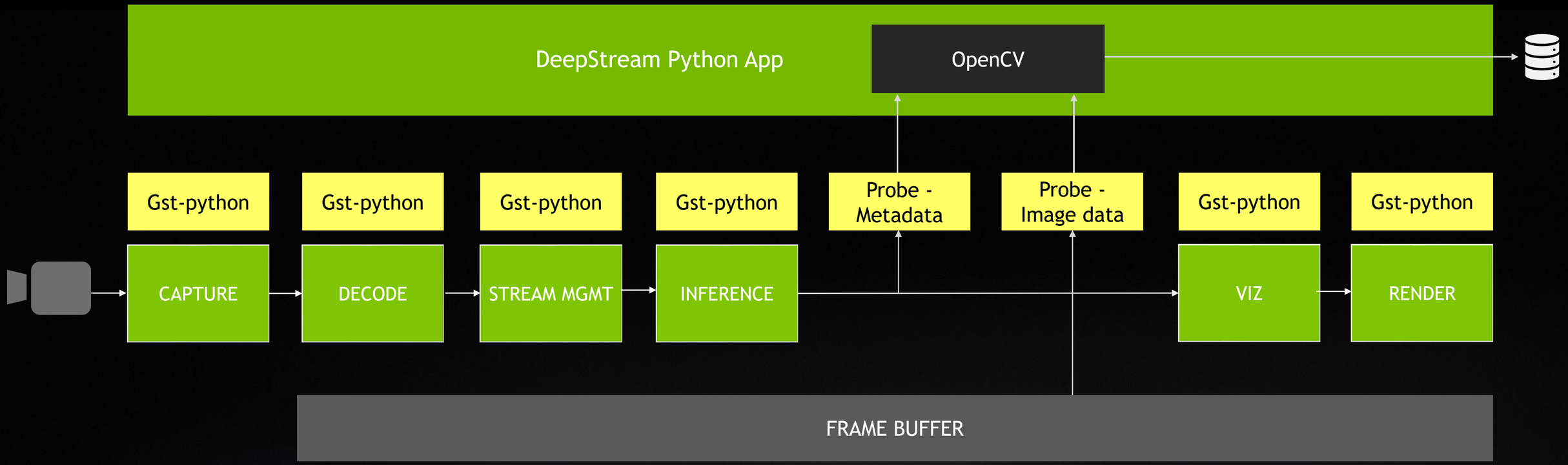
## DeepStream-test5




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

Python app coming soon

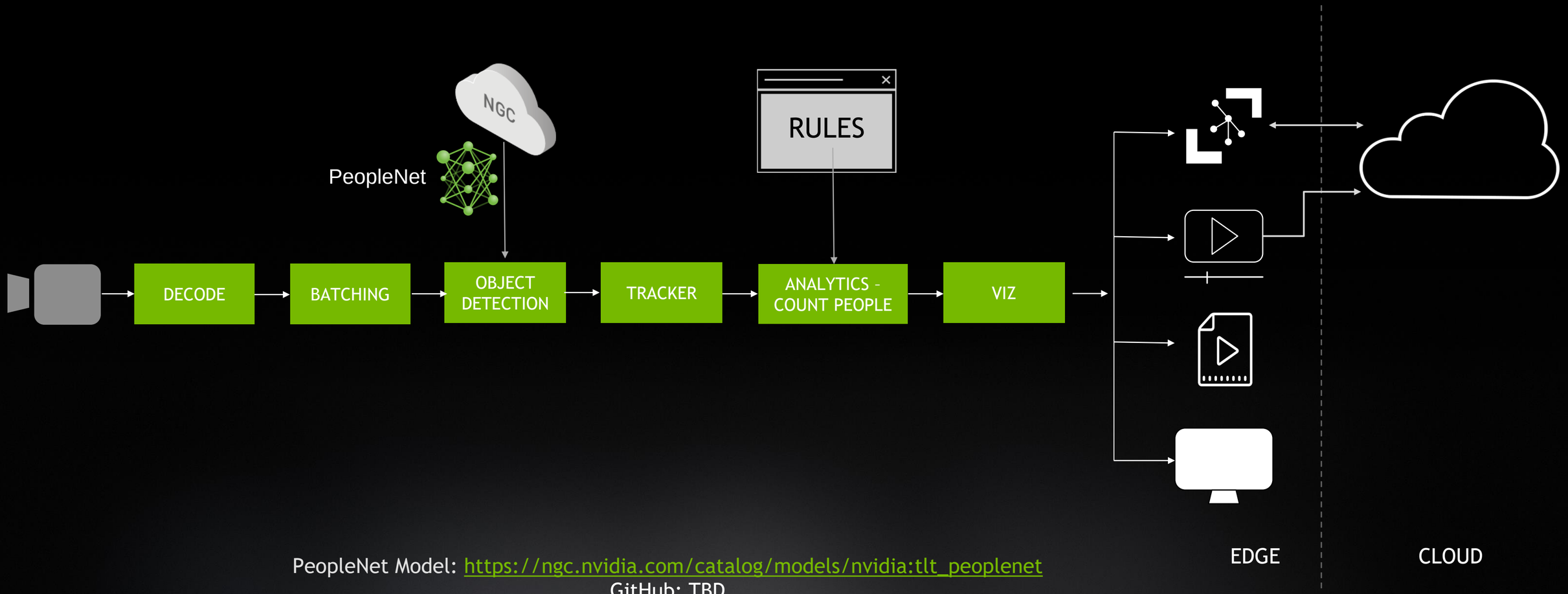
# IMAGE DATA ACCESS IN PYTHON




https://github.com/NVIDIA-AI-IOT/deepstream_python_apps/tree/master/apps/deepstream-imagedata-multistream

# END-TO-END APPS

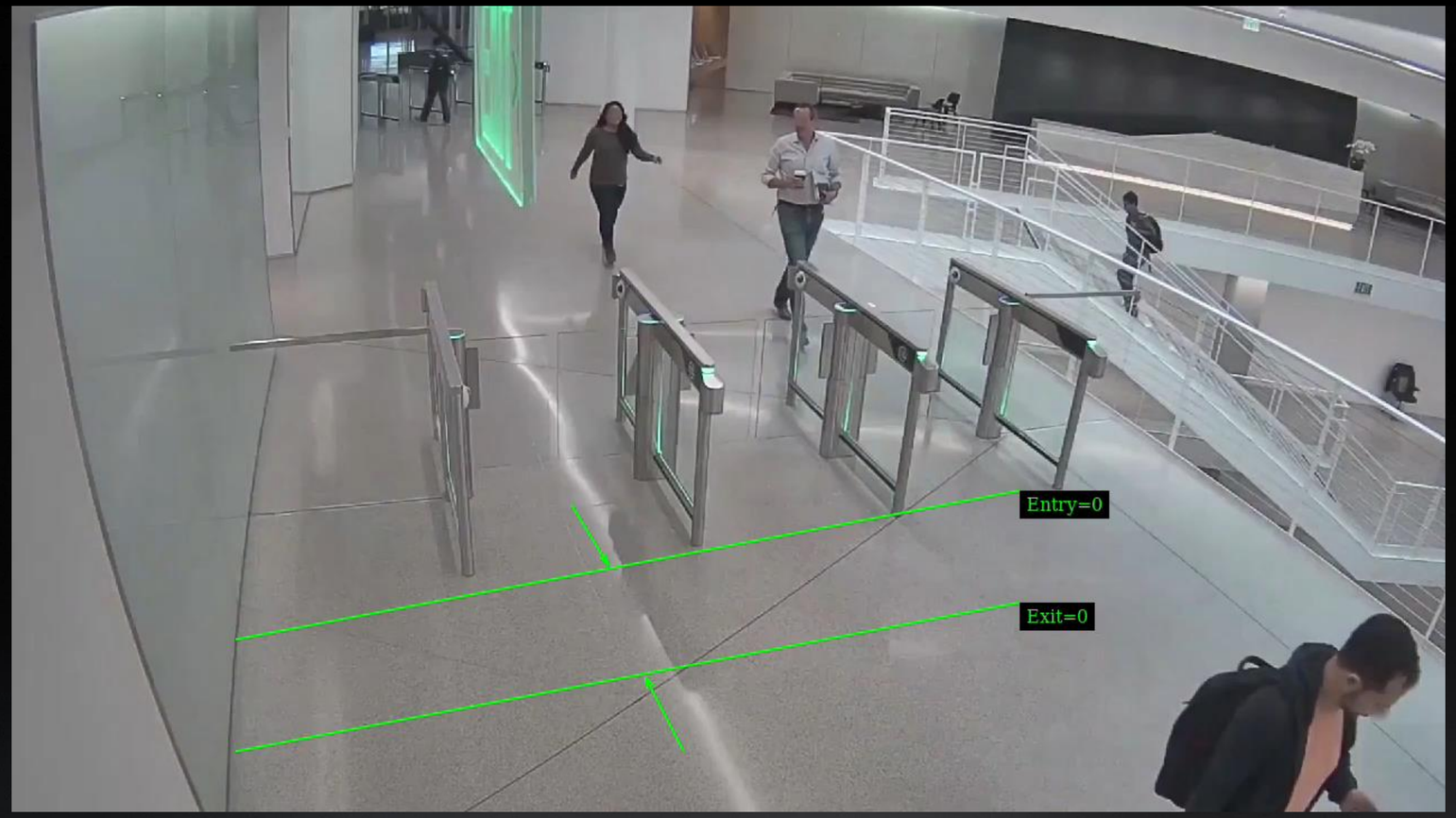
# DEEPSTREAM APPLICATION



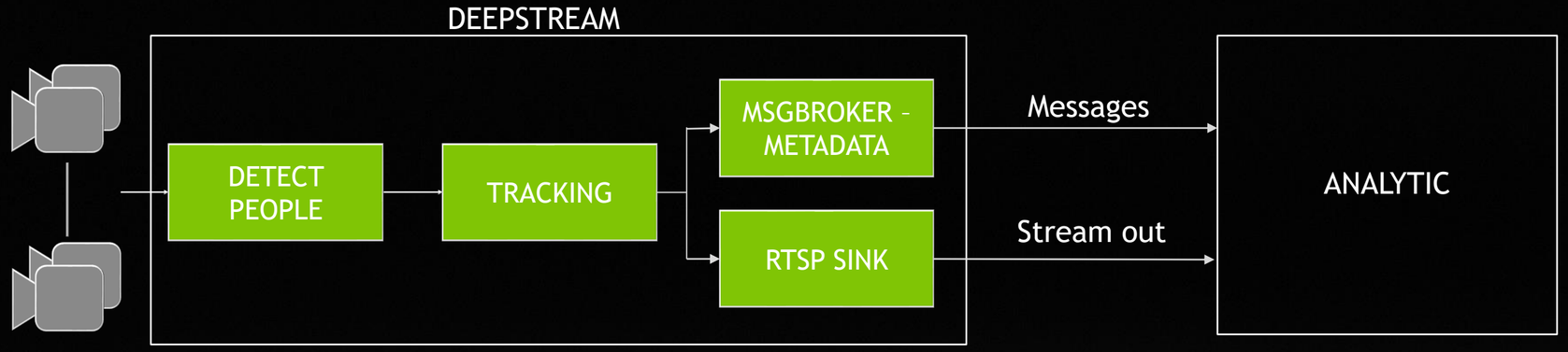

PeopleNet Model: https://ngc.nvidia.com/catalog/models/nvidia:tlt_peoplenet
GitHub: TBD

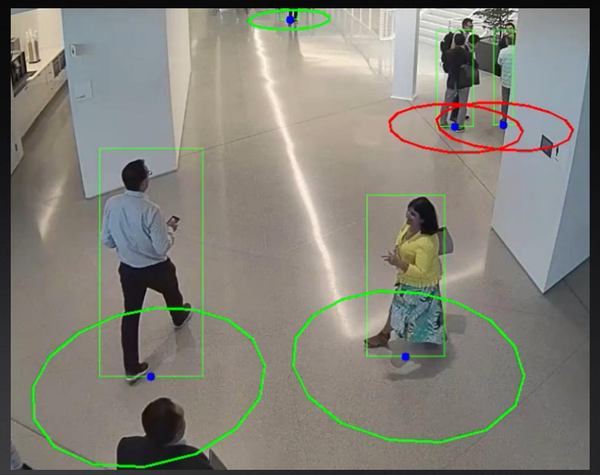
# PEOPLE COUNTING

Demo

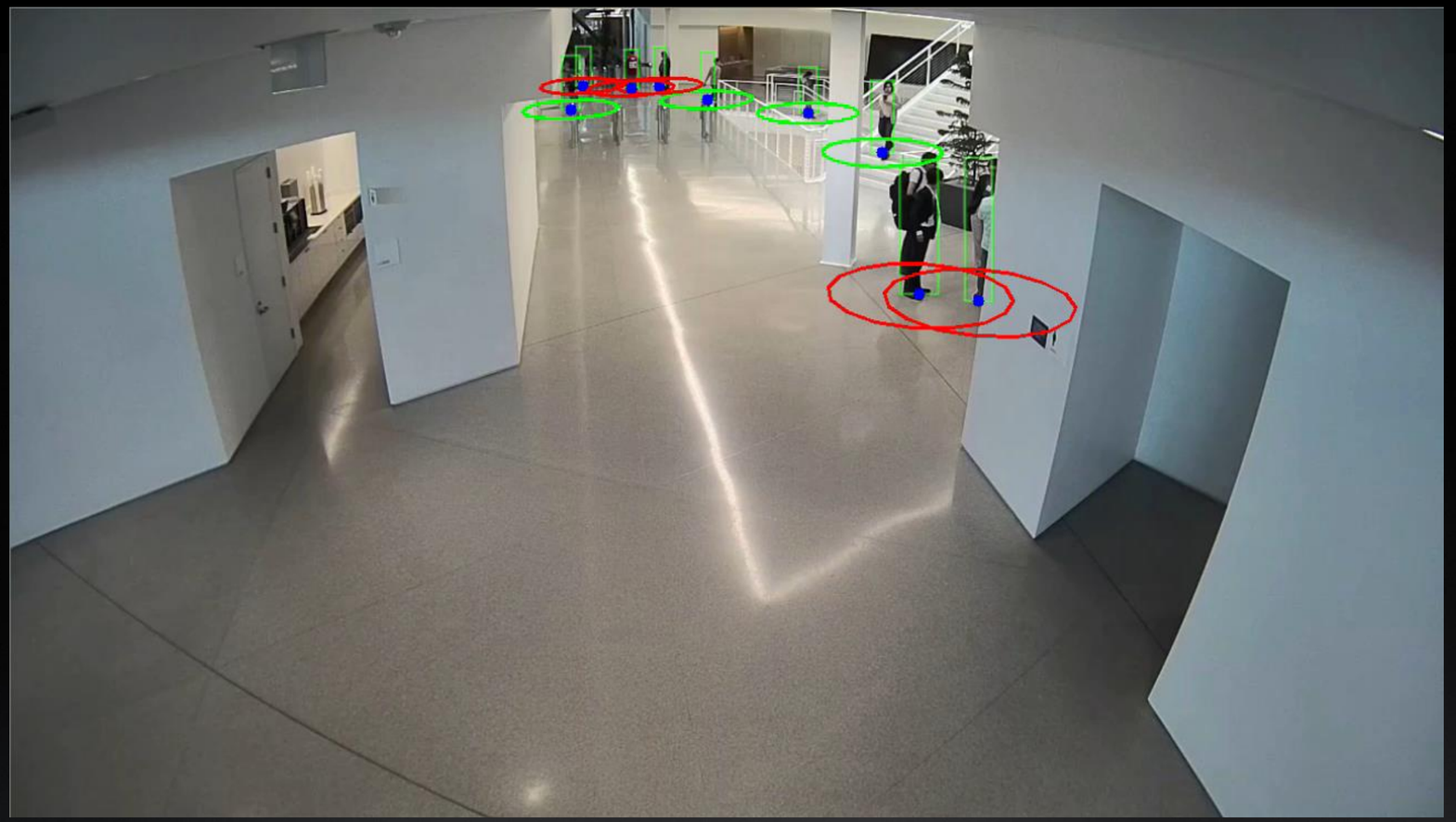


# SOCIAL DISTANCING APP



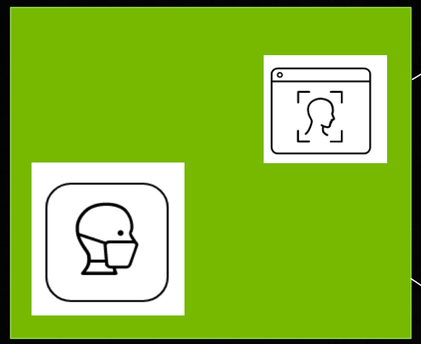



# SOCIAL DISTANCING APP



# FACE MASK DETECTION

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



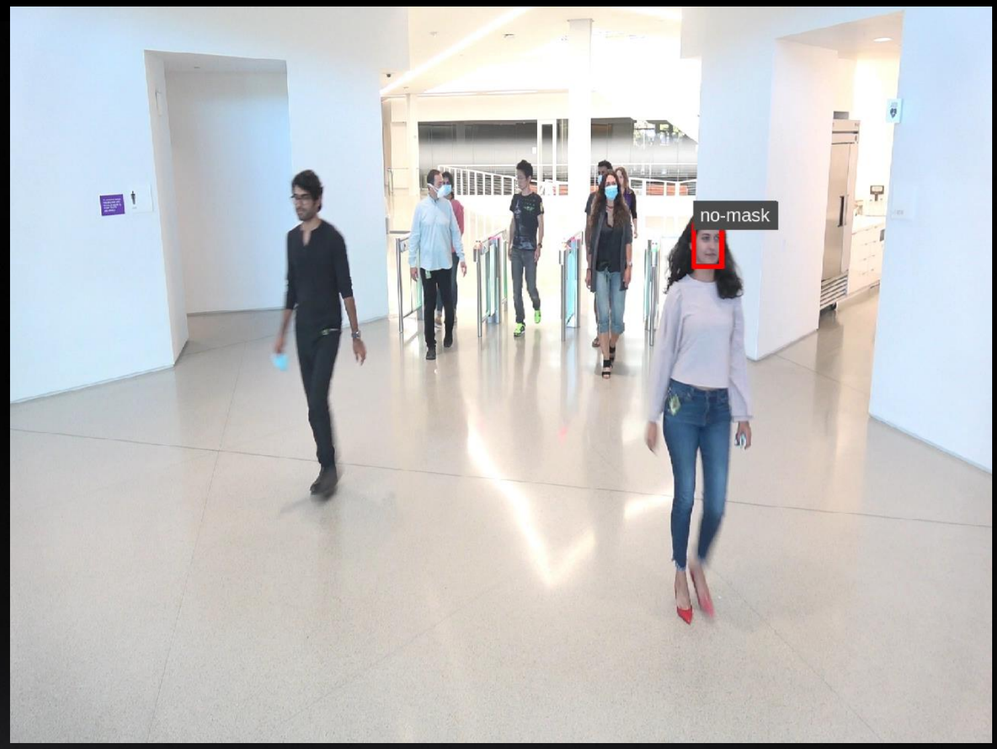
### Use Cases

Hospitals
Workspaces
Mass Transit Hubs
Stadiums
Warehouses

### Key Components

Transfer Learning Toolkit
DeepStream SDK

**GitHub Repo**



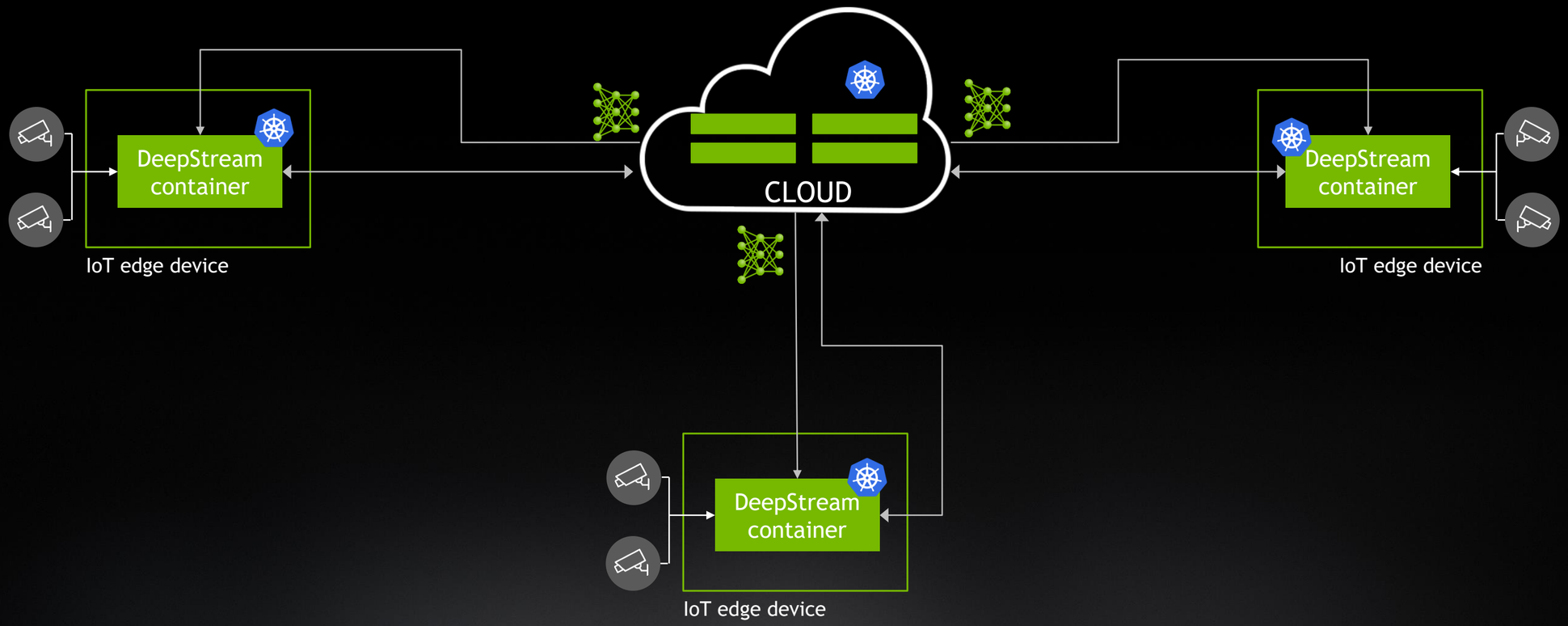

Developer Blog
GitHub Repo

What this project does not provide:

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

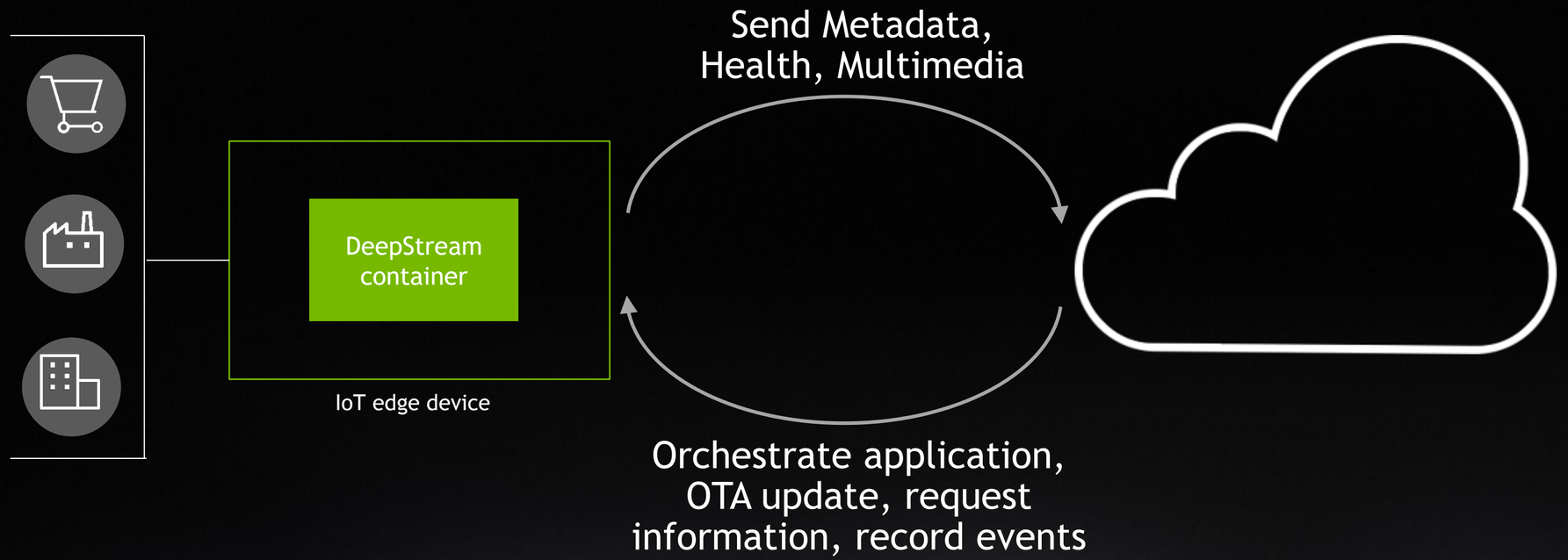
# DEPLOY

# AT-SCALE DEPLOYMENT

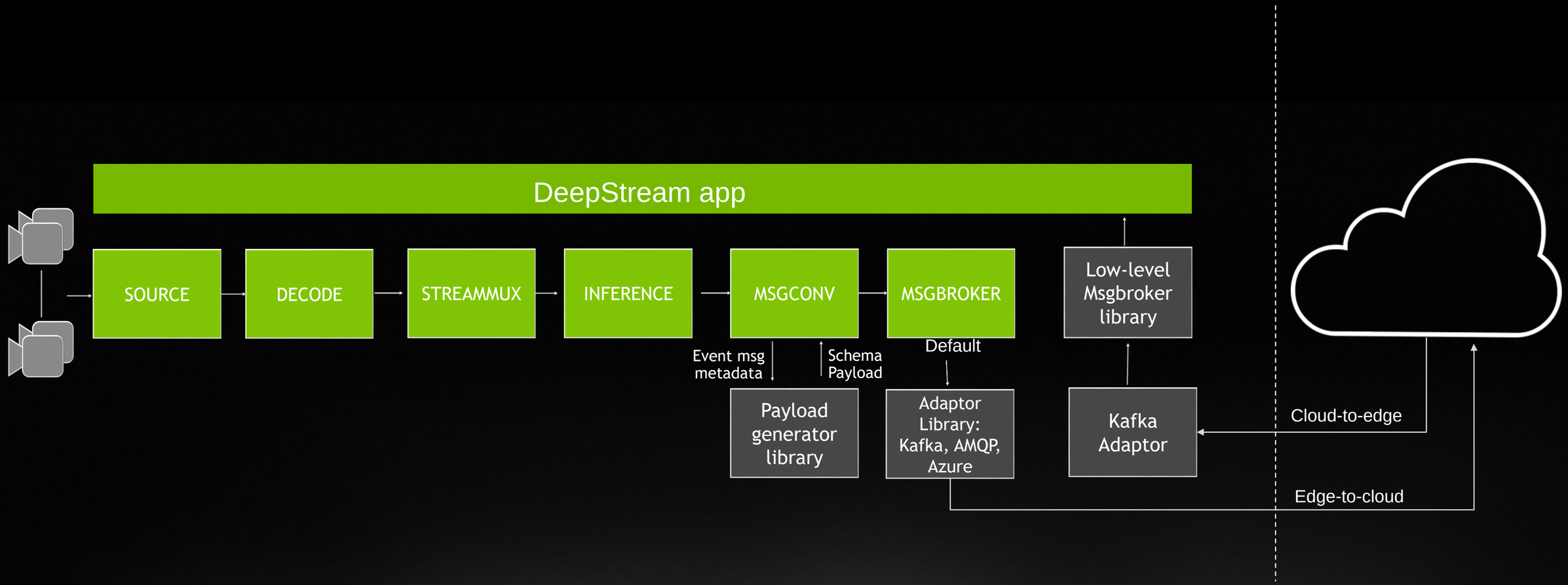



Download the Demo to deploy DeepStream application using Helm charts and Kubernetes on NGC

# BI-DIRECTIONAL IOT COMMUNICATION



# BI-DIRECTIONAL MESSAGING



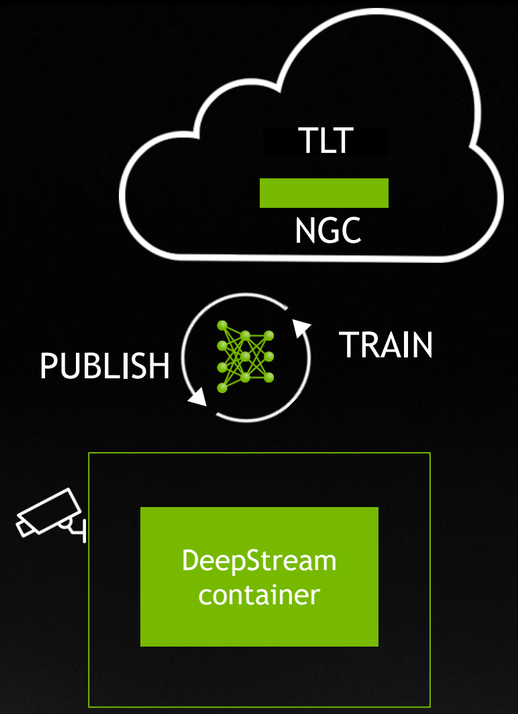

Learn about Bi-directional messaging with DeepStream

# OTA MODEL UPDATE

Use case: Edge deployment that requires frequent model changes

On-the-fly model update with zero Downtime

Example provided in the SDK



# SMART RECORD

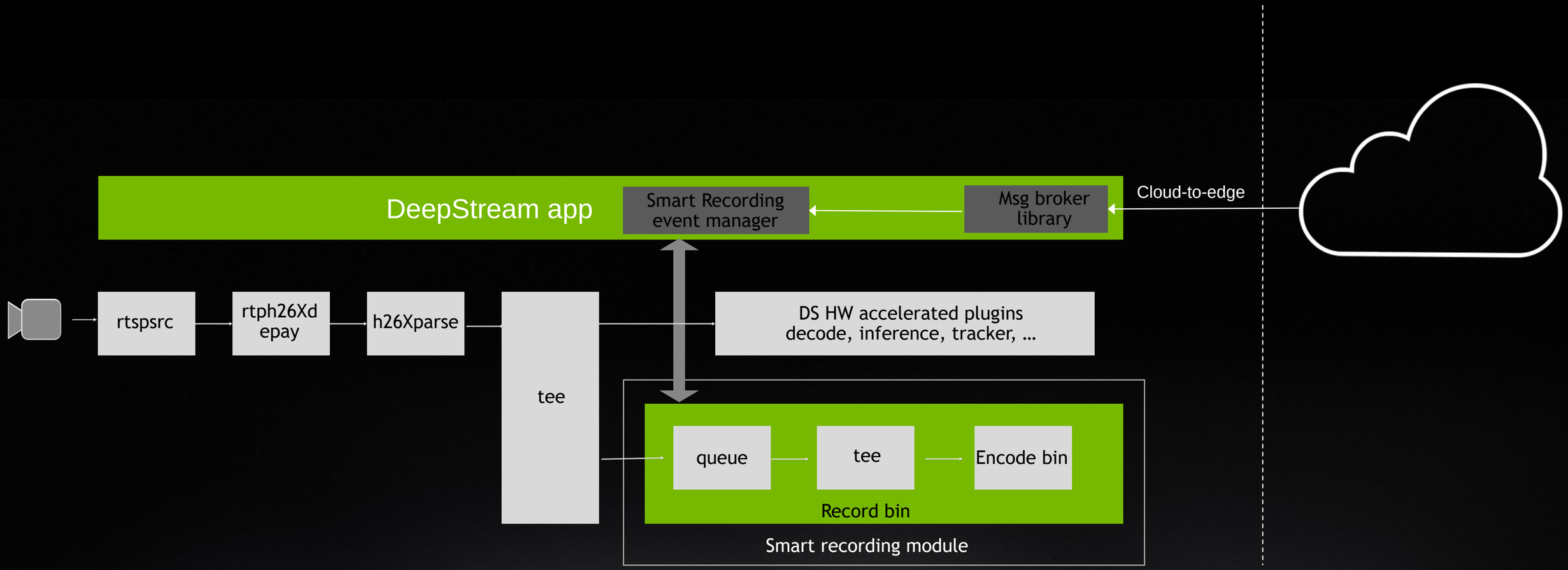

Use case:

- Trigger based event record
- Record anomalies

Benefits:

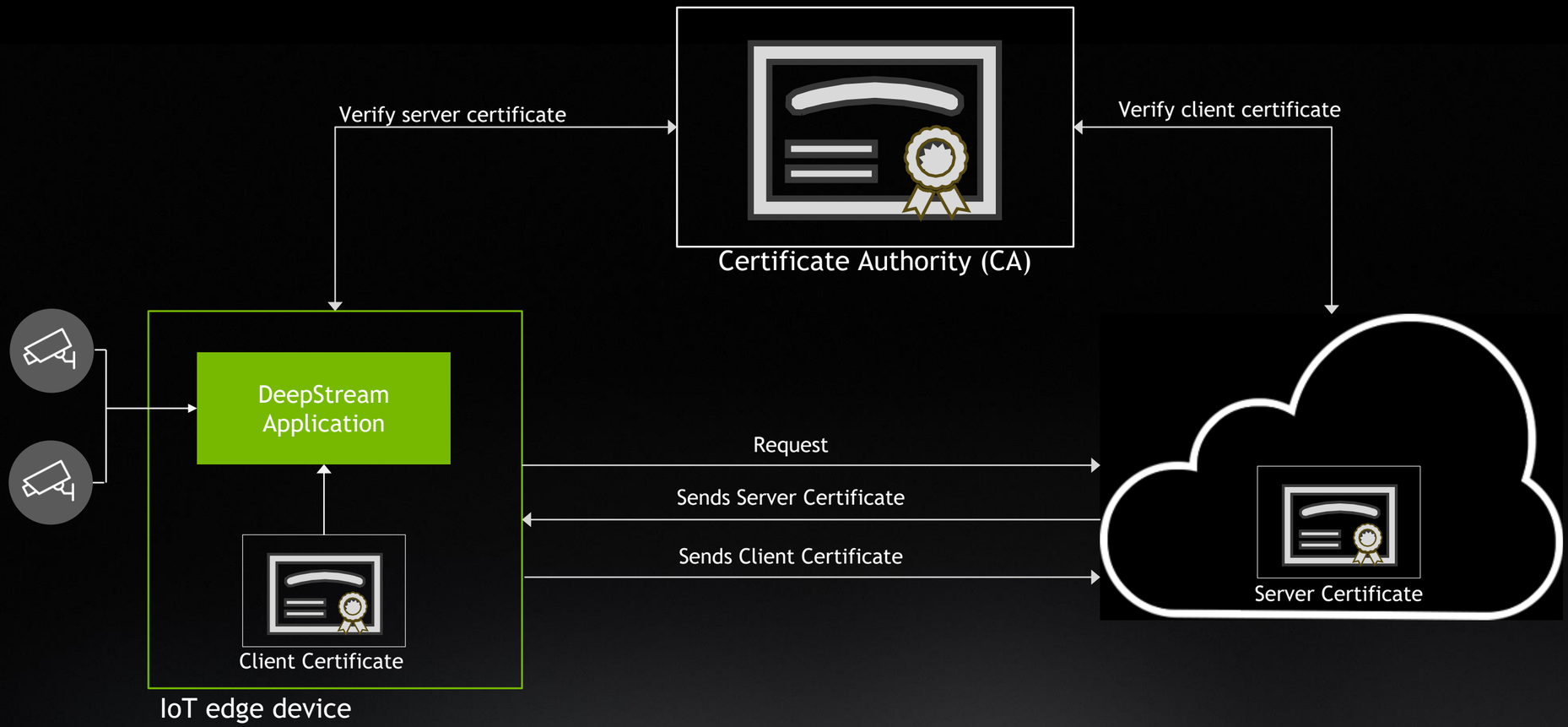
- Selective record saves valuable disk space

# SMART RECORD ARCHITECTURE




Learn more about DeepStream Smart record module

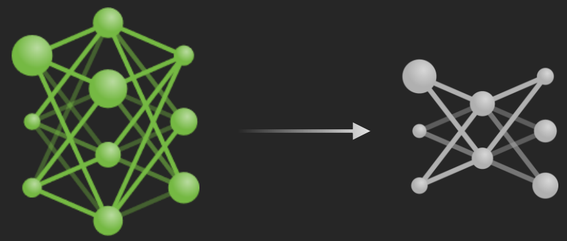
# SECURITY




Secure authentication with SSL certificates and SASL/Plain

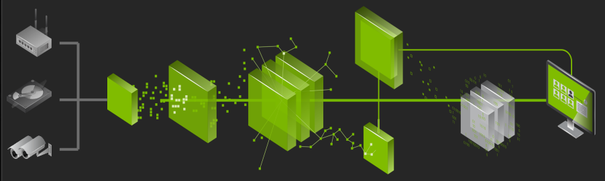
# SUMMARY

## TRAIN



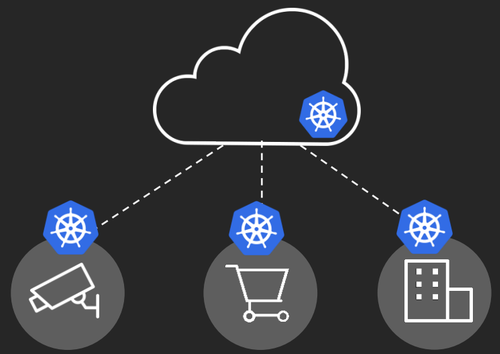
Use Pre-Trained models and TLT to build accurate AI

## BUILD



Build high performance IVA application using DeepStream

## DEPLOY



Use K8s and IoT features in DeepStream to deploy at-scale

# GET STARTED TODAY

## Powerful End to End Intelligent Video Analytics Made Easy

### Pre-trained Models

Pre-trained models on NGC

TLT Object Detection

TLT DetectNet_v2 Object Detection

TLT classification

TLT Instance segmentation

Developer Forums

### Transfer Learning Toolkit

Transfer Learning Toolkit

Documentation

Developer Forums

Getting started resources, sample apps, webinars & tutorials

### DeepStream SDK

DeepStream SDK

Documentation

Developer Forums

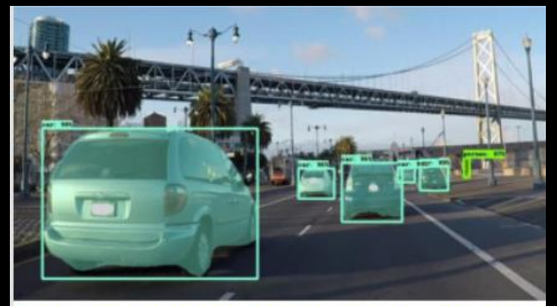
Free Online DLI Course

Getting started resources, sample apps, webinars & tutorials

Sign up for our new webinar on 8/25:

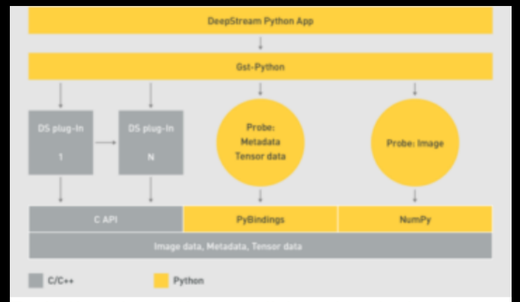
CREATE INTELLIGENT PLACES USING NVIDIA PRE-TRAINED VISION MODELS AND DEEPSTREAM SDK

# 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

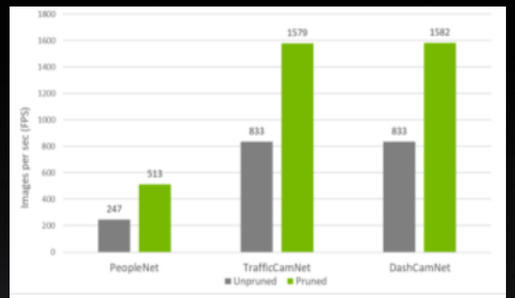
Building a Real-time Redaction App Using NVIDIA DeepStream, Part 1: Training

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

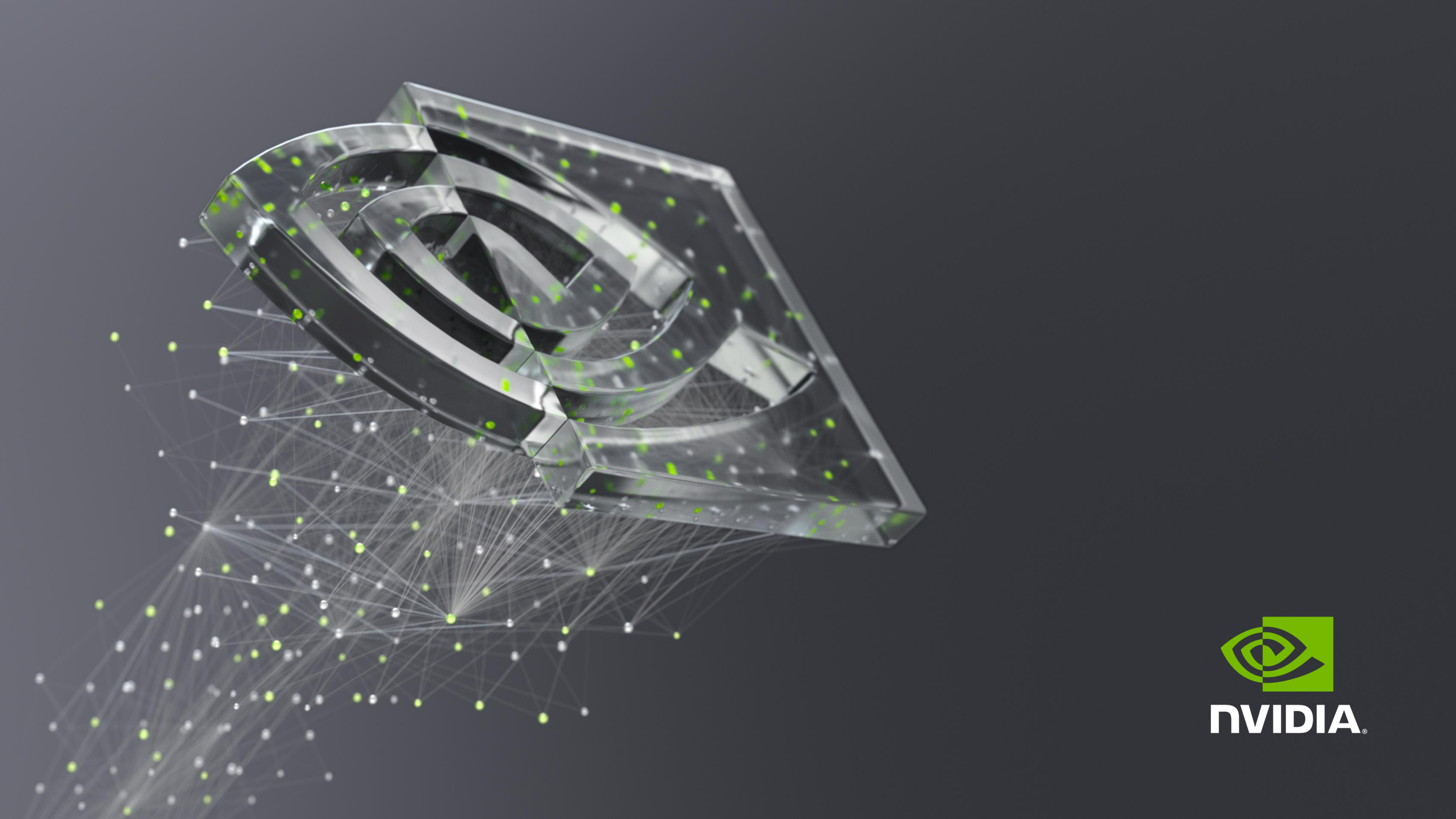

Implementing real-time AI-based face mask detection

Tutorial

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



NVIDIA


NVIDIA