goto;

GOTO
Copenhagen 2021

#GOTOcph

BEST PRACTICES FOR REAL-TIME INTELLIGENT VIDEO ANALYTICS

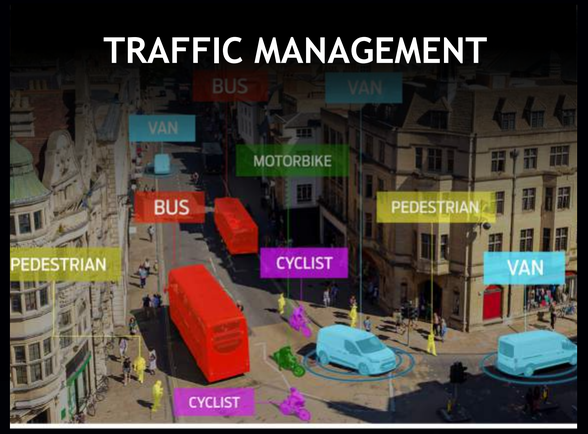
DR. EKATERINA SIRAZITDINOVA, NOVEMBER 2021

WHY INTELLIGENT VIDEO ANALYTICS?

WHY INTELLIGENT VIDEO ANALYTICS?

Every industry is going through rapid innovation to bring intelligent insights

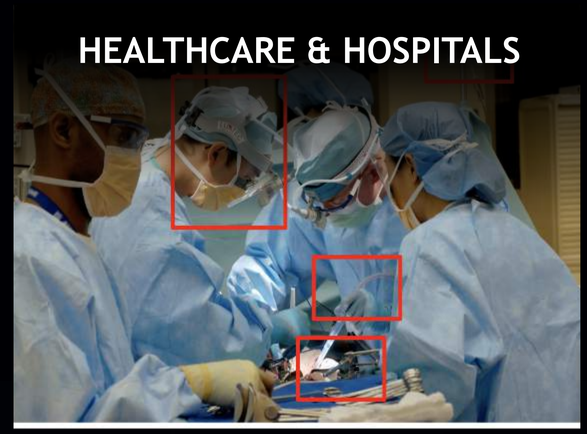
TRAFFIC MANAGEMENT



MANUFACTURING

RETAIL

HEALTHCARE & HOSPITALS



FACTORIES

PUBLIC SPACES

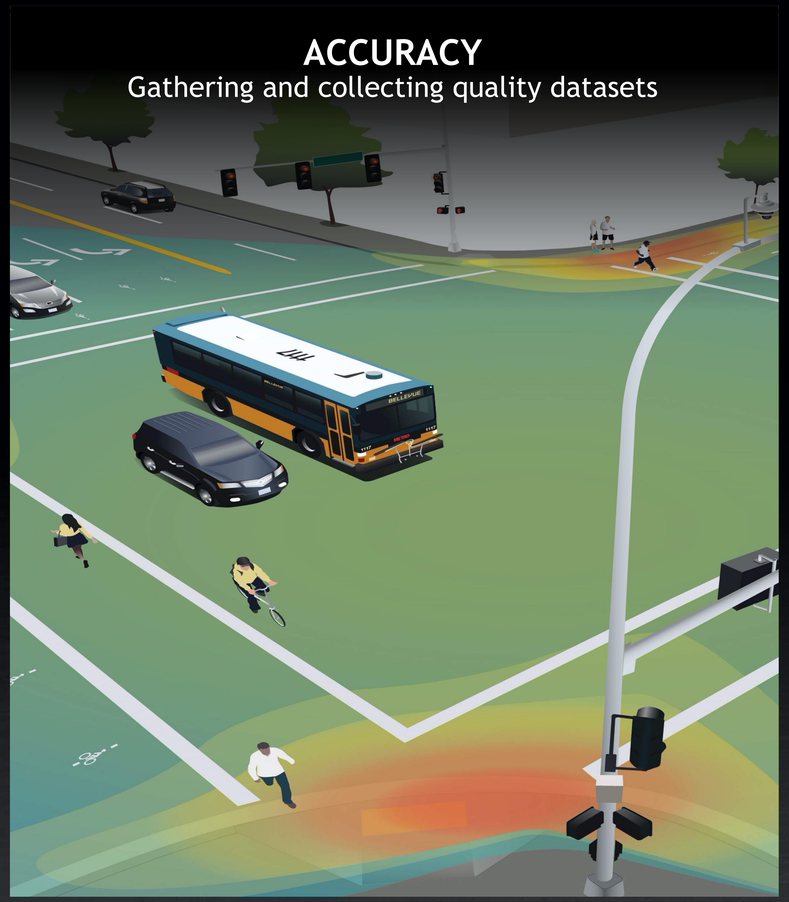
WAREHOUSE & LOGISTICS

TRANSPORTATION HUBS

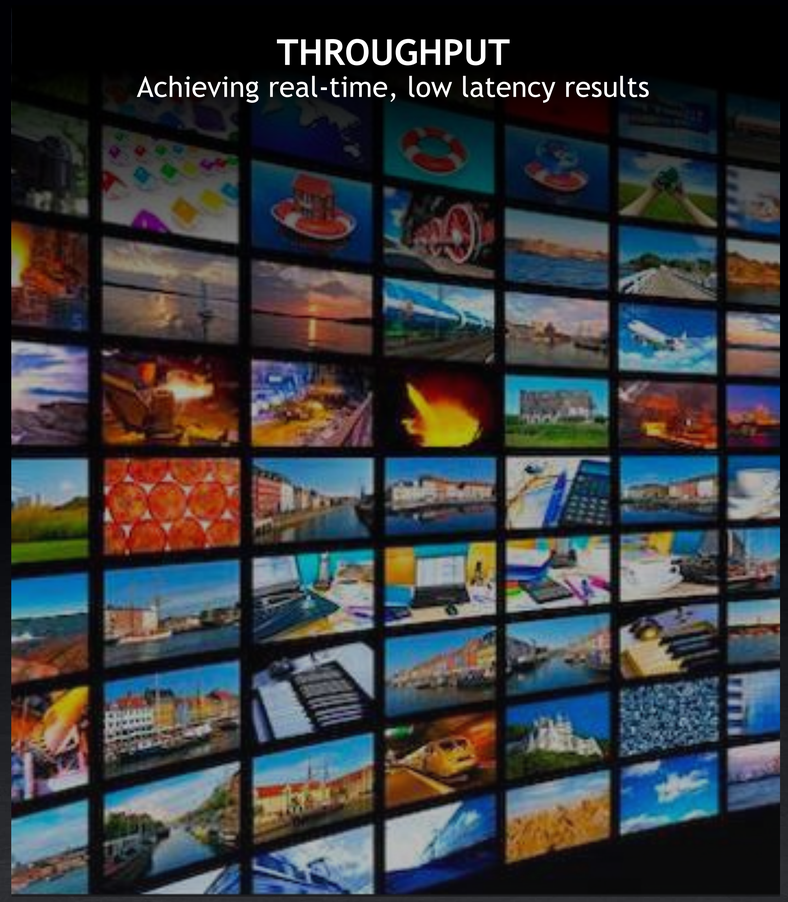
NVIDIA

CHALLENGES WITH INTELLIGENT VIDEO ANALYTICS

ACCURACY
Gathering and collecting quality datasets



THROUGHPUT
Achieving real-time, low latency results



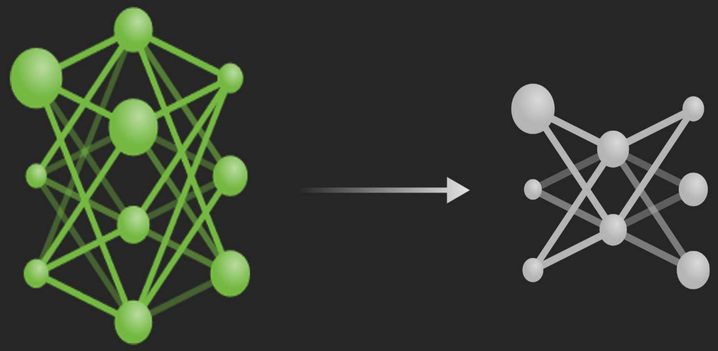
DEPLOYMENT AT SCALE
Real-time, large scale deployment across the world

TIPS AND TRICKS FOR EFFICIENT AI VIDEO ANALYTICS

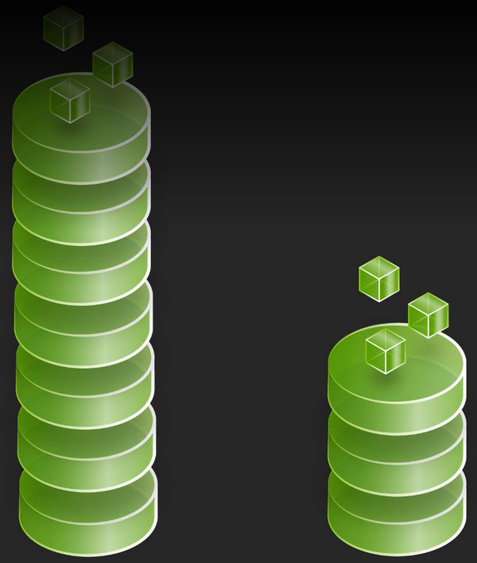
TRANSFER LEARNING

Transferring learned features from one model to another

FEATURE TRANSFER



KEY BENEFITS

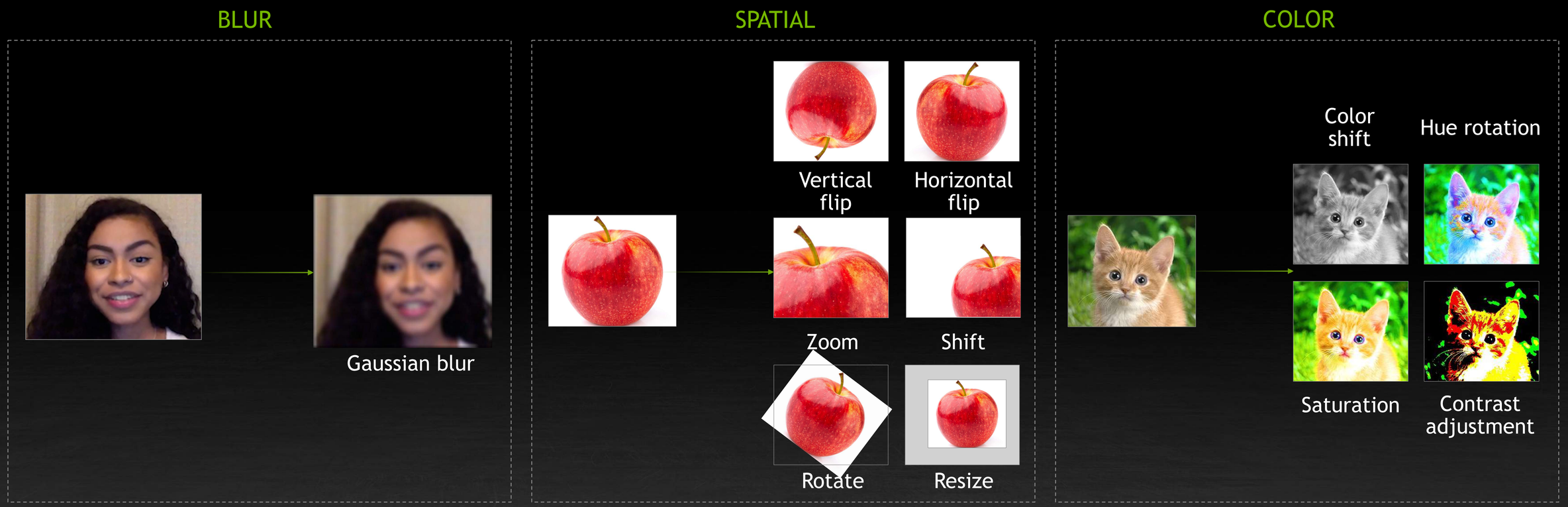


Less data required to train accurately

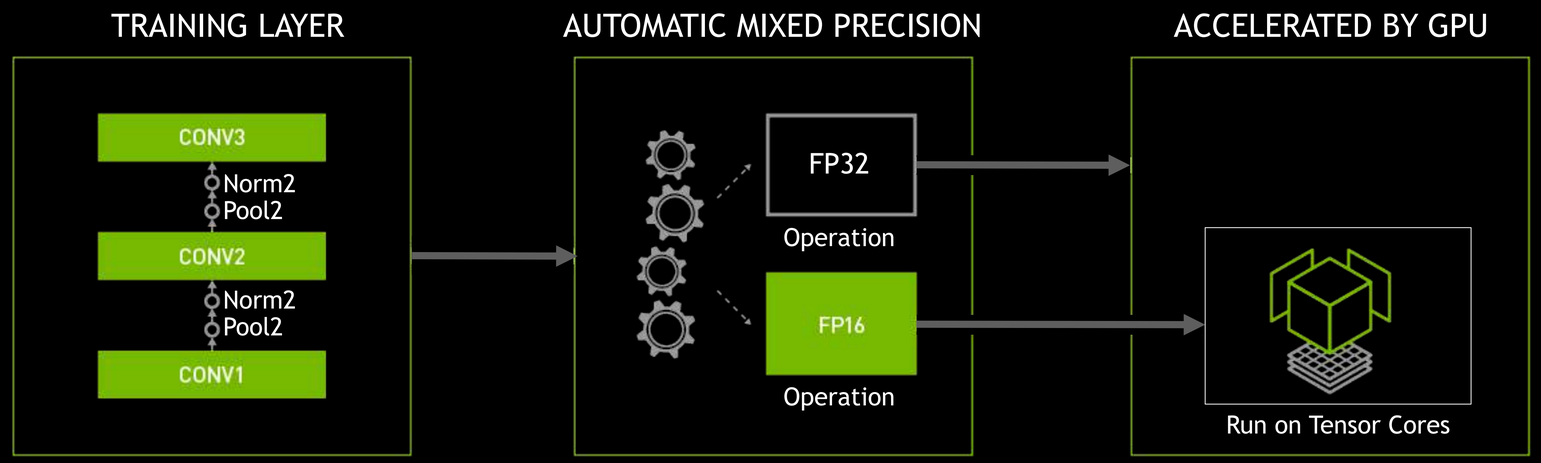
Reduce training time and cost

DATA AUGMENTATION

Extending the dataset by applying simple transformations

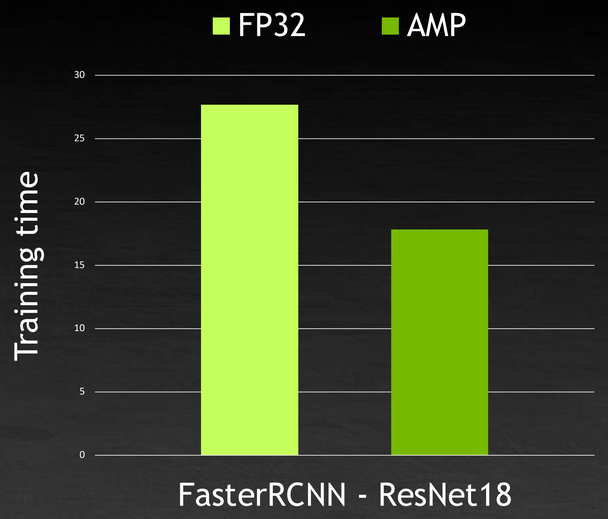


AUTOMATIC MIXED PRECISION (AMP)



Train with half-precision while maintaining network accuracy as with single precision in order to:

- Speed-up math intensive operations by using Tensor Cores
- Speed-up memory intensive operations by using half the bytes
- Train larger models or larger batches
- Reduce memory requirements



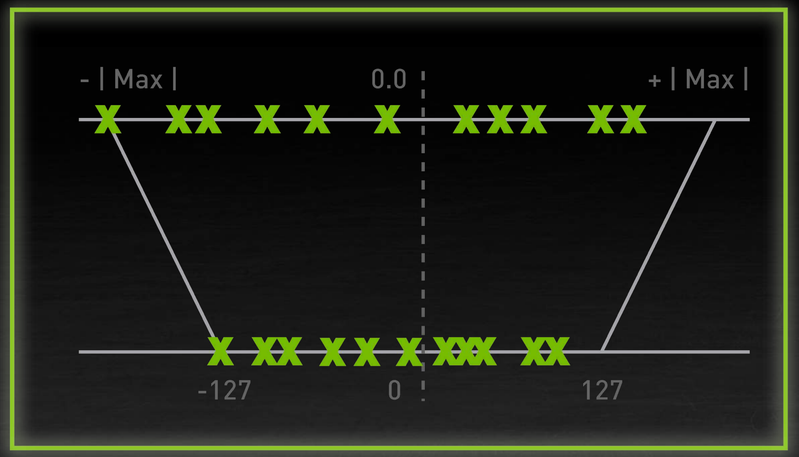
1.5x speedup

QUANTIZATION

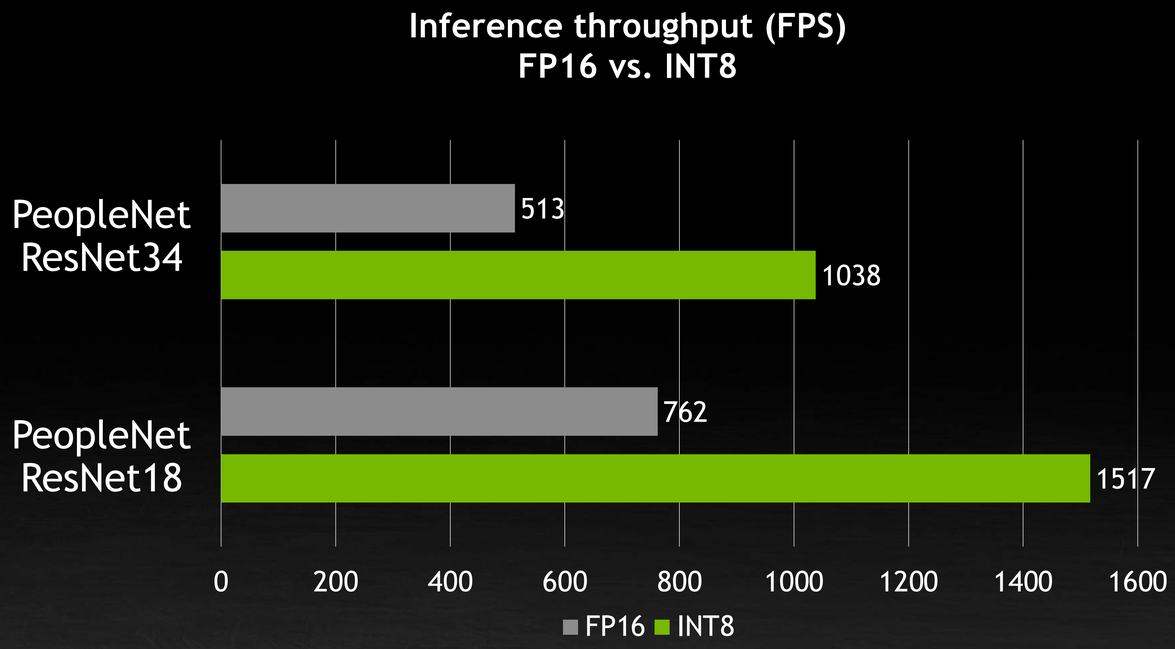
Reducing bits per weight

Post Training Quantization (PTQ) for quantization after training is done

Quantization Aware Training (QAT) for modelling quantization error from weights and tensors during training



Quantize(x, r) = round(s * clip(x, -r, r))
where r = |Max| and s = 127 / r



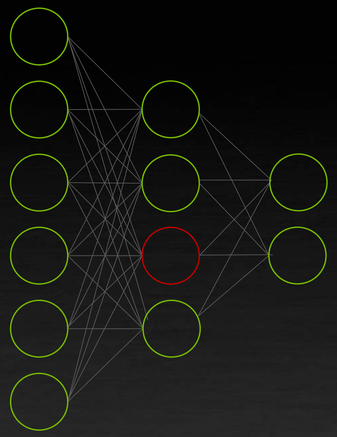
2x throughput Increase

NETWORK PRUNING

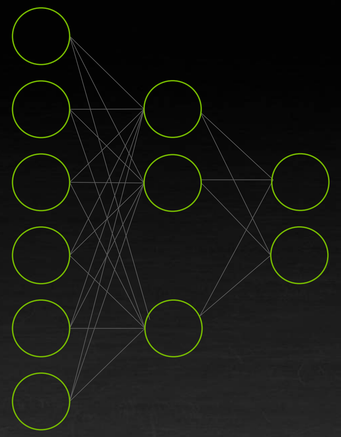
Reducing the number of weights

2-step process

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

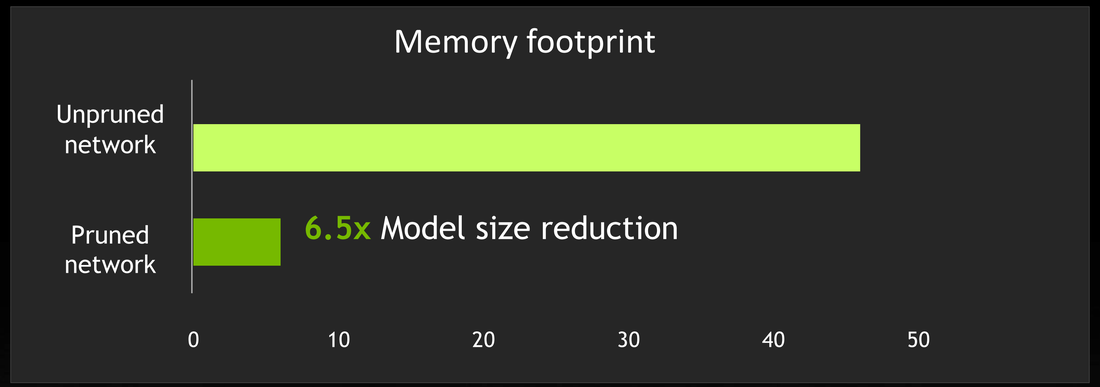


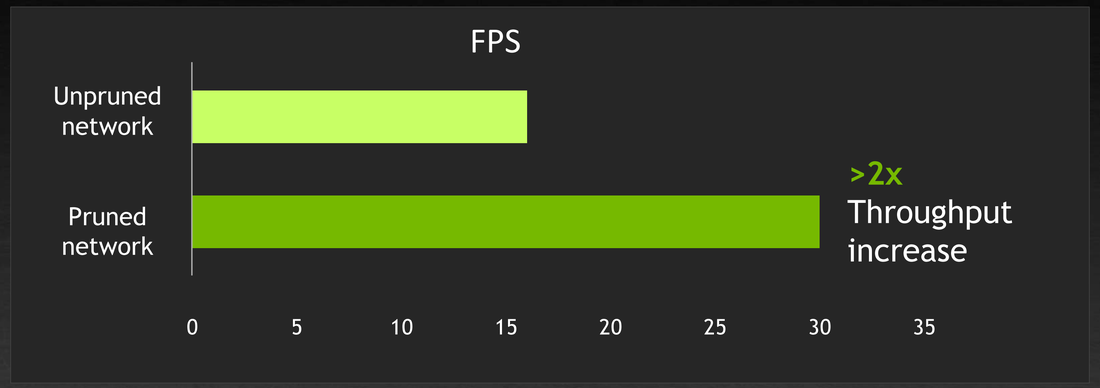
6 inputs, 6 neurons,
32 connections



6 inputs, 5 neurons,
24 connections

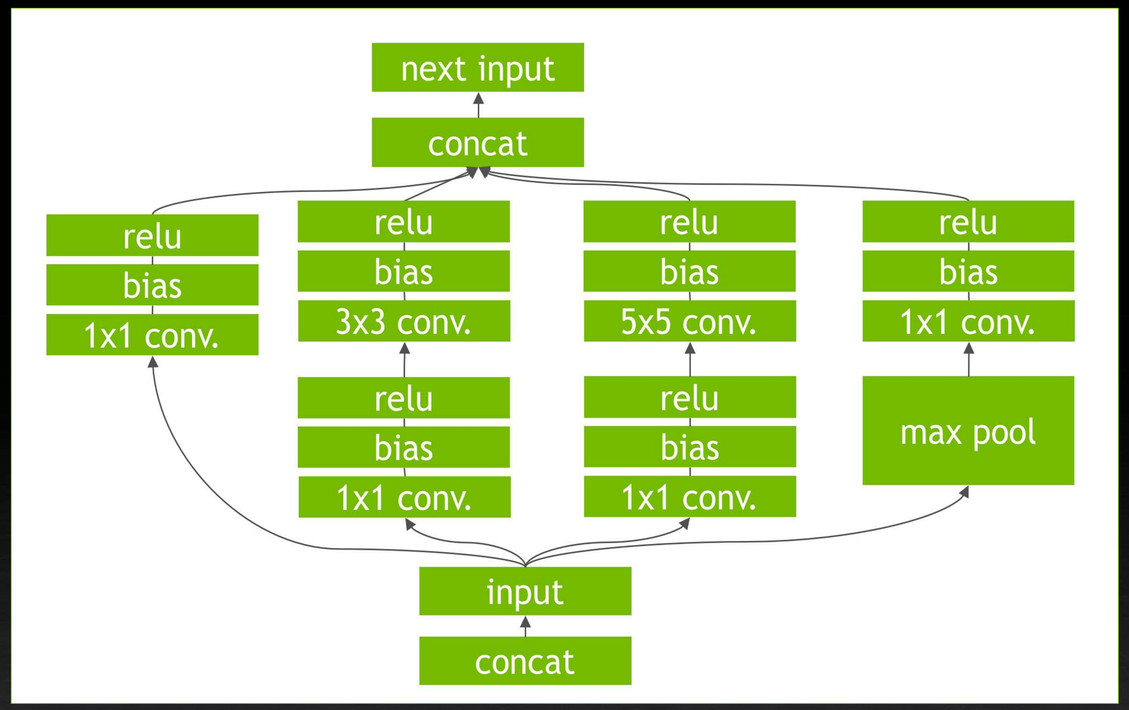
Network - ResNet18 4-class



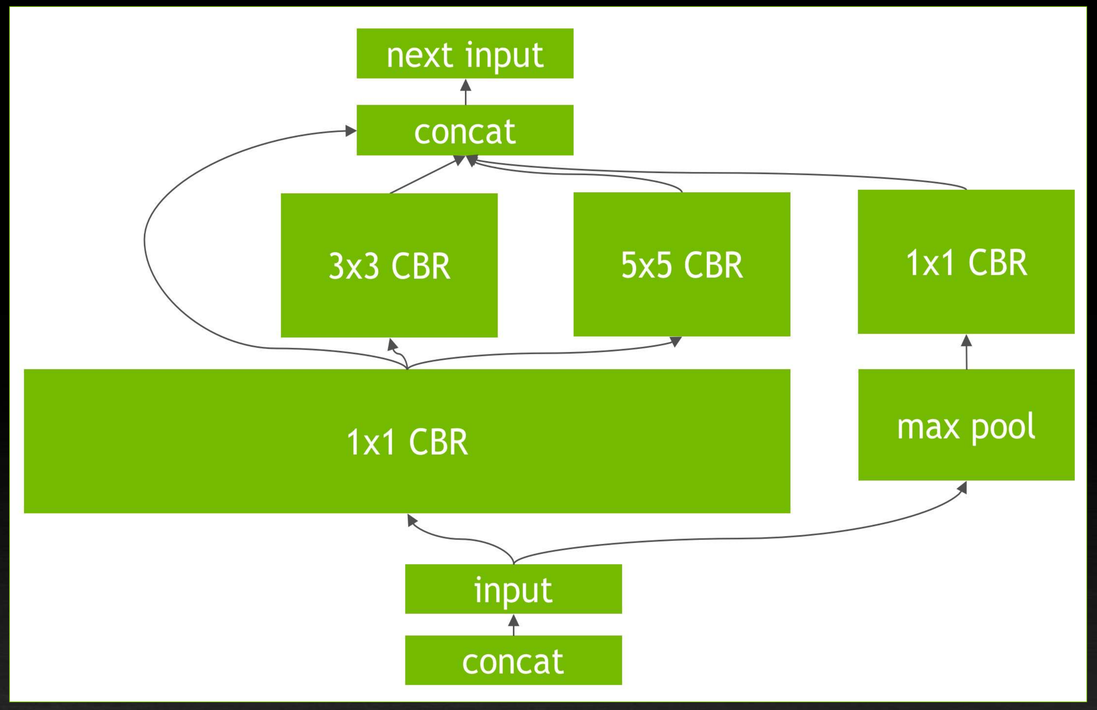


NETWORK GRAPH OPTIMIZATIONS

Layer & tensor fusion



Before layer fusion

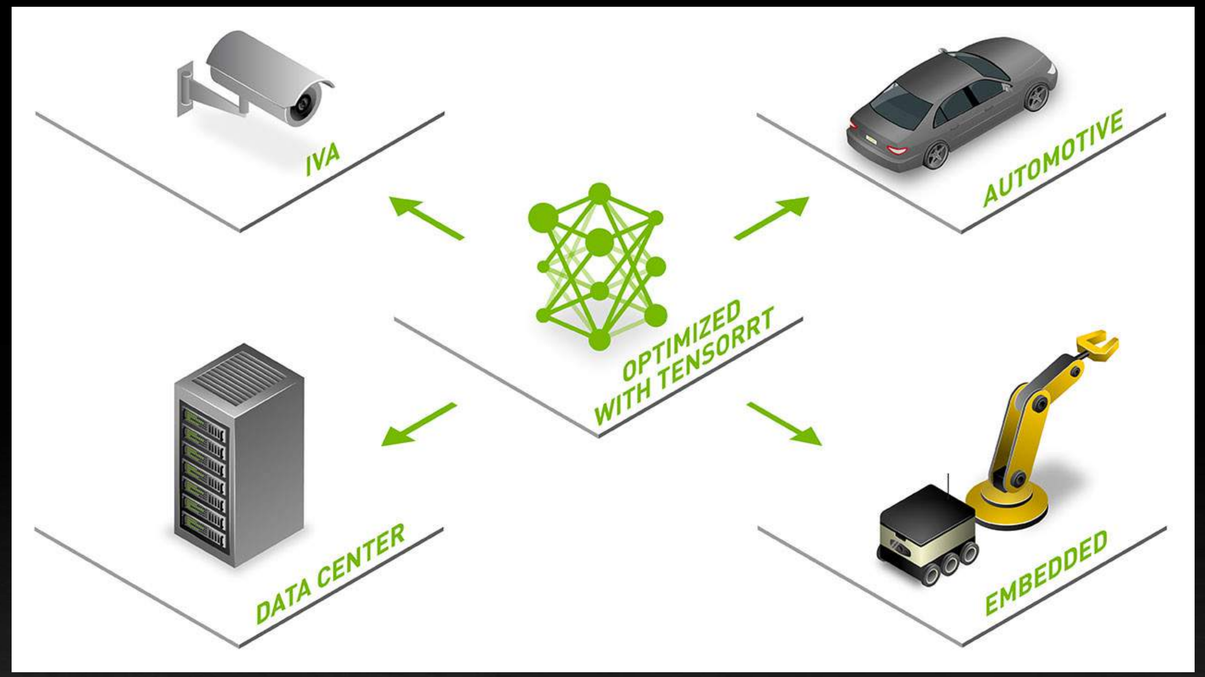


After horizontal and vertical layer fusion

KERNEL AUTO-TUNING

Picking right algorithms depending on your deployment hardware

- Target hardware platform
- Batch size
- Input dimensions
- Filter dimensions
- Tensor layout
- Specific algorithms implementation
- ***100s of specialized kernels otimized for every GPU platform***

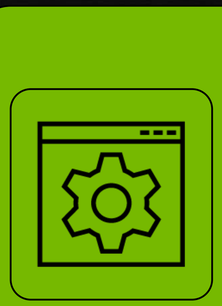


Nano TX2 Xavier NX Xavier AGX x86 GPUs

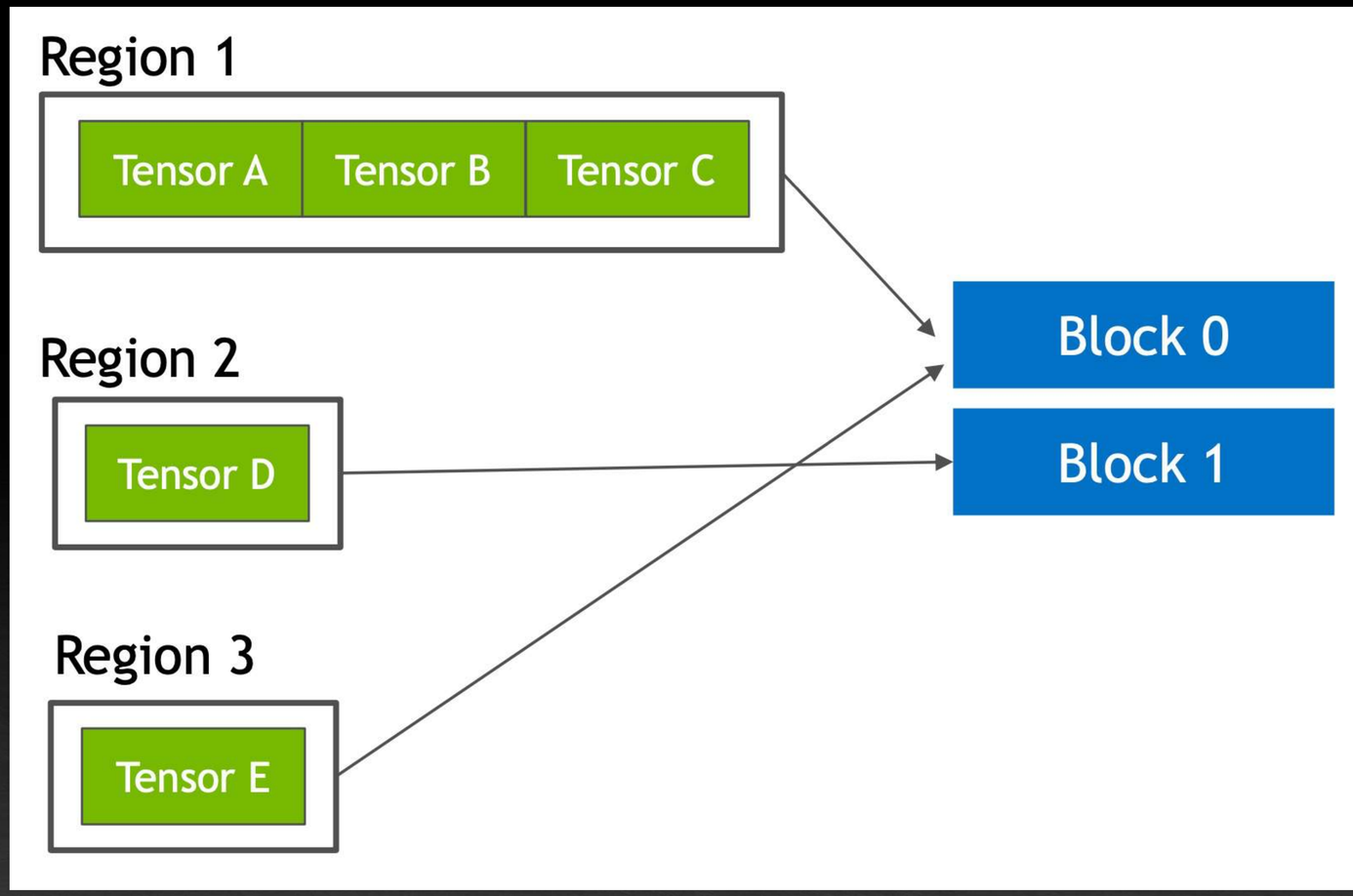
DYNAMIC TENSOR MEMORY UPON INFERENCE

Reduced network memory footprint & improved memory re-use

- Combining tensors into regions
  - Region lifetime is a section of network execution time
- Assigning regions to blocks
  - Regions assigned to a block have disjoint lifetimes

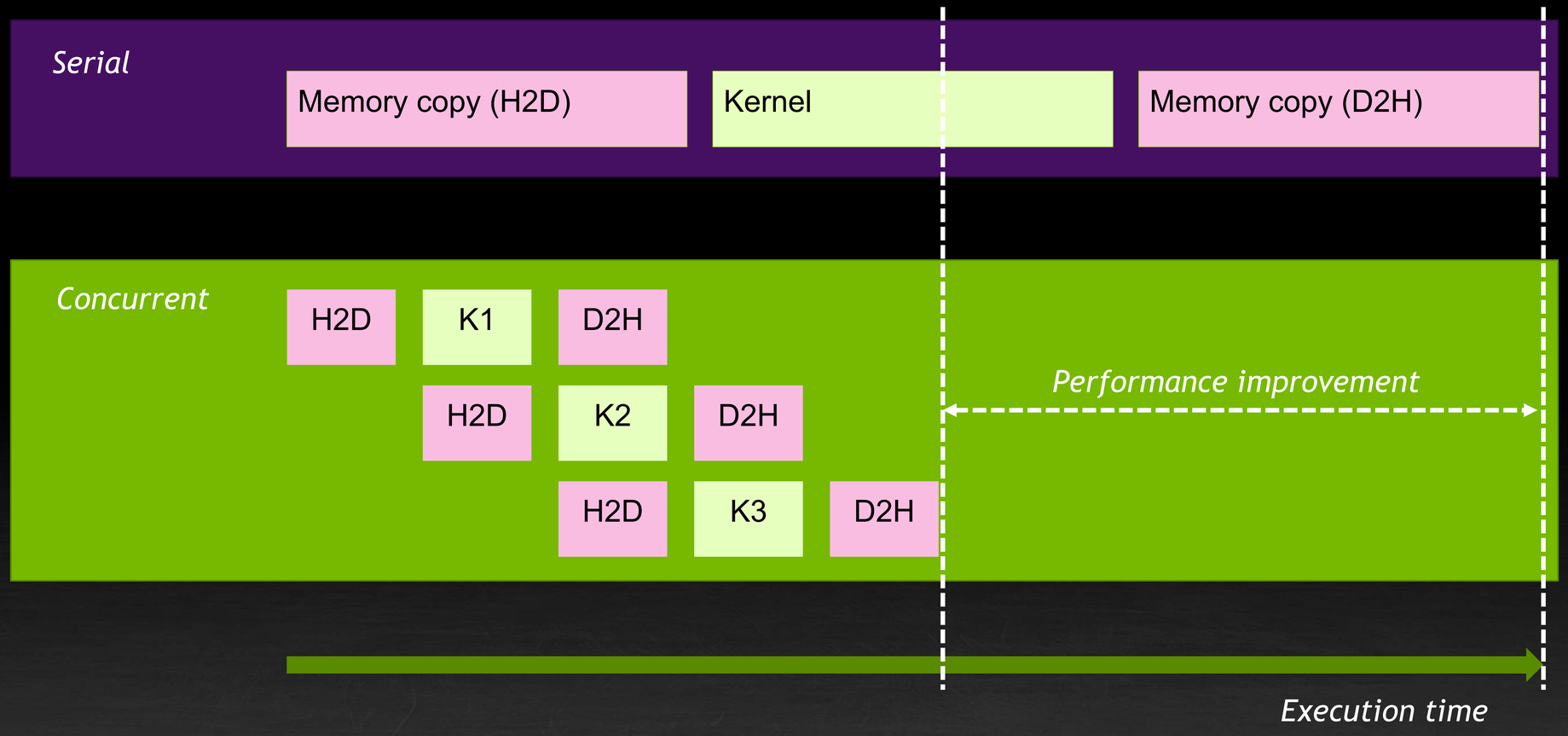


Similar to ***register allocation in compiling***, the process of assigning a large number of target program variables onto a small number of registers



MULTISTREAM CONCURENT EXECUTION

For better GPU utilization and higher throughput



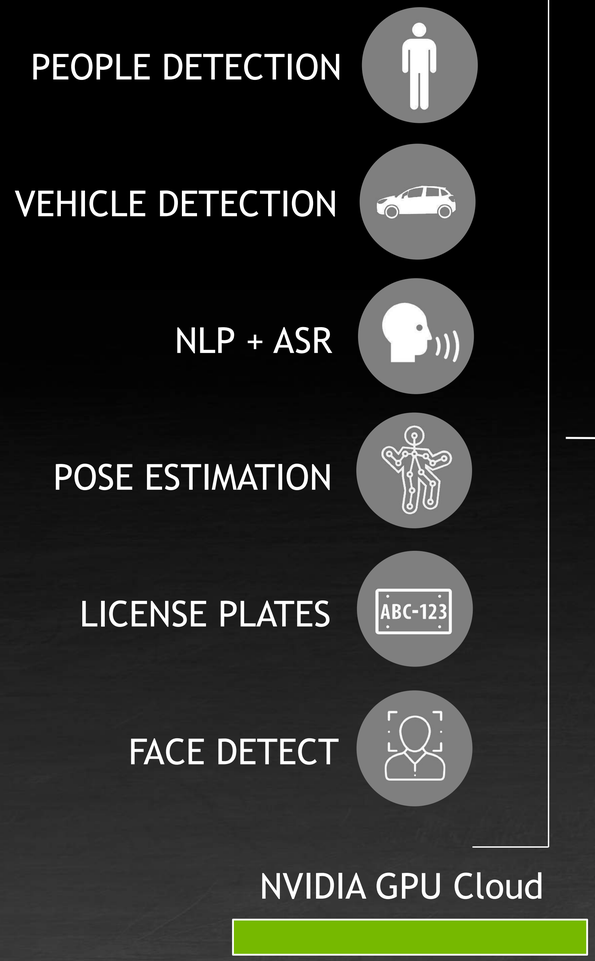

FREE NVIDIA PRODUCTS DESIGNED TO
MAKE YOUR AI APPLICATIONS EFFICIENT

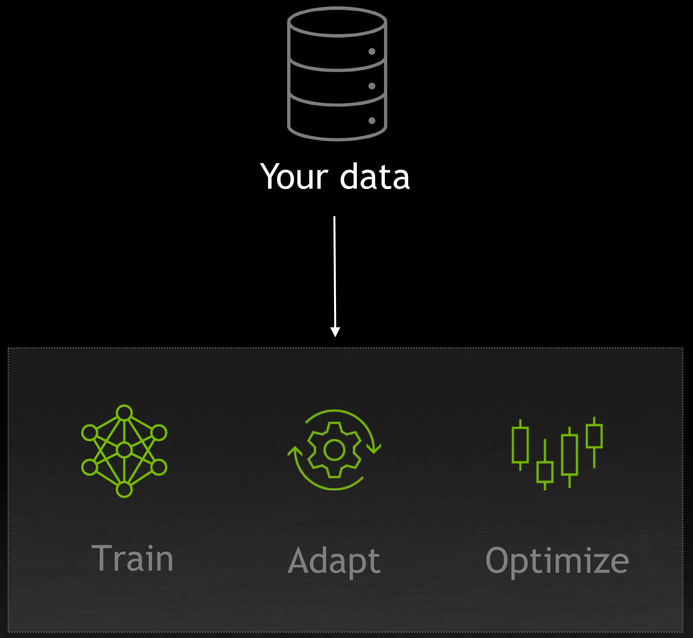

“NVIDIA is not a GPU company. It’s a platform company.”
Jensen Huang, CEO of NVIDIA

NVIDIA'S END-TO-END AI WORKFLOW

Develop & deploy production ready solutions

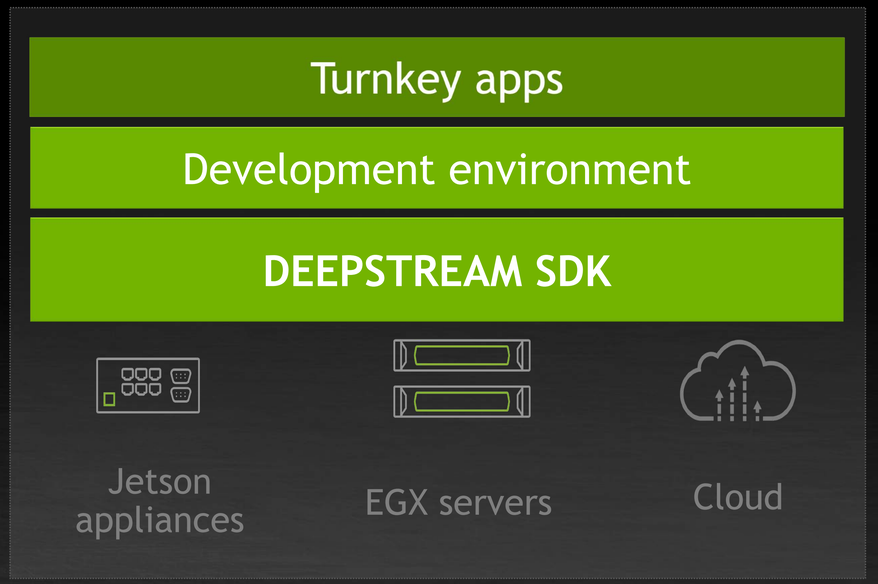
PRE-TRAINED MODEL LIBRARY



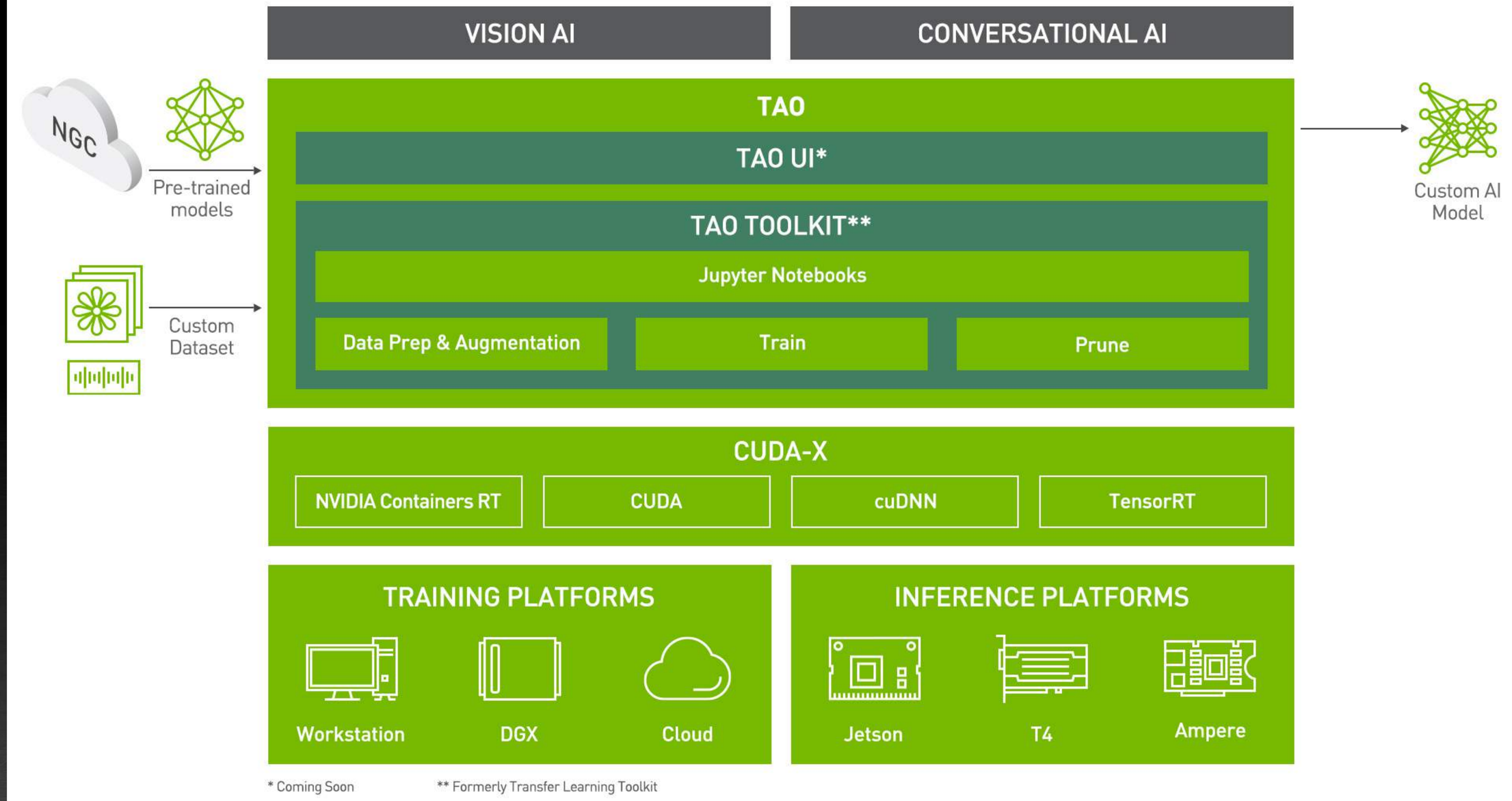


Your production model

DEVELOPMENT AND DEPLOYMENT



TAO TOOLKIT


VISION AI
CONVERSATIONAL AI
NGC
Pre-trained models
Custom Dataset
TAO
TAO UI*
TAO TOOLKIT**
Jupyter Notebooks
Data Prep & Augmentation
Train
Prune
Custom AI Model
CUDA-X
NVIDIA Containers RT
CUDA
cuDNN
TensorRT
TRAINING PLATFORMS
Workstation
DGX
Cloud
INFERENCE PLATFORMS
Jetson
T4
Ampere
* Coming Soon
** Formerly Transfer Learning Toolkit

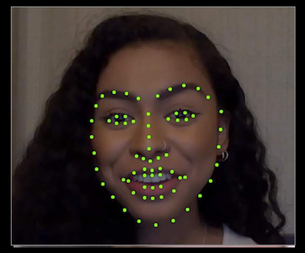
HIGH PERFORMANCE PRE-TRAINED VISION AI MODELS

Download for free from https://ngc.nvidia.com/

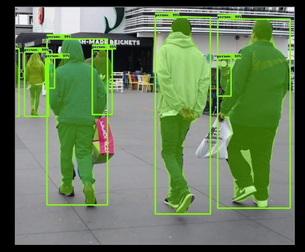
- Optimized for high throughput
- Trained for >80% accuracy
- Production-ready
- Adaptable with NVIDIA TAO

People detection

Gaze estimation

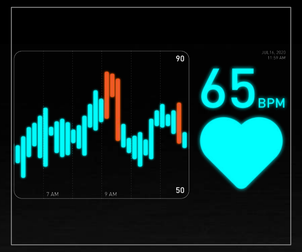


Facial landmark



People segmentation

Gesture recognition



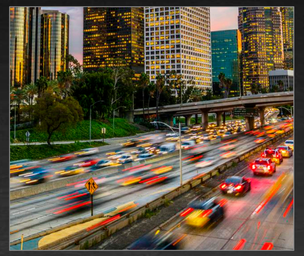
Heart rate estimation

Emotion recognition

Pose estimation

Face detect IR

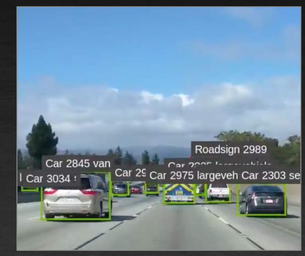
Dash camera vehicle detection



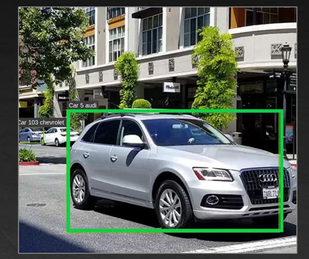
Vehicle & pedestrian detection

License plate detection

License plate recognition



Vehicle type net



Vehicle make net

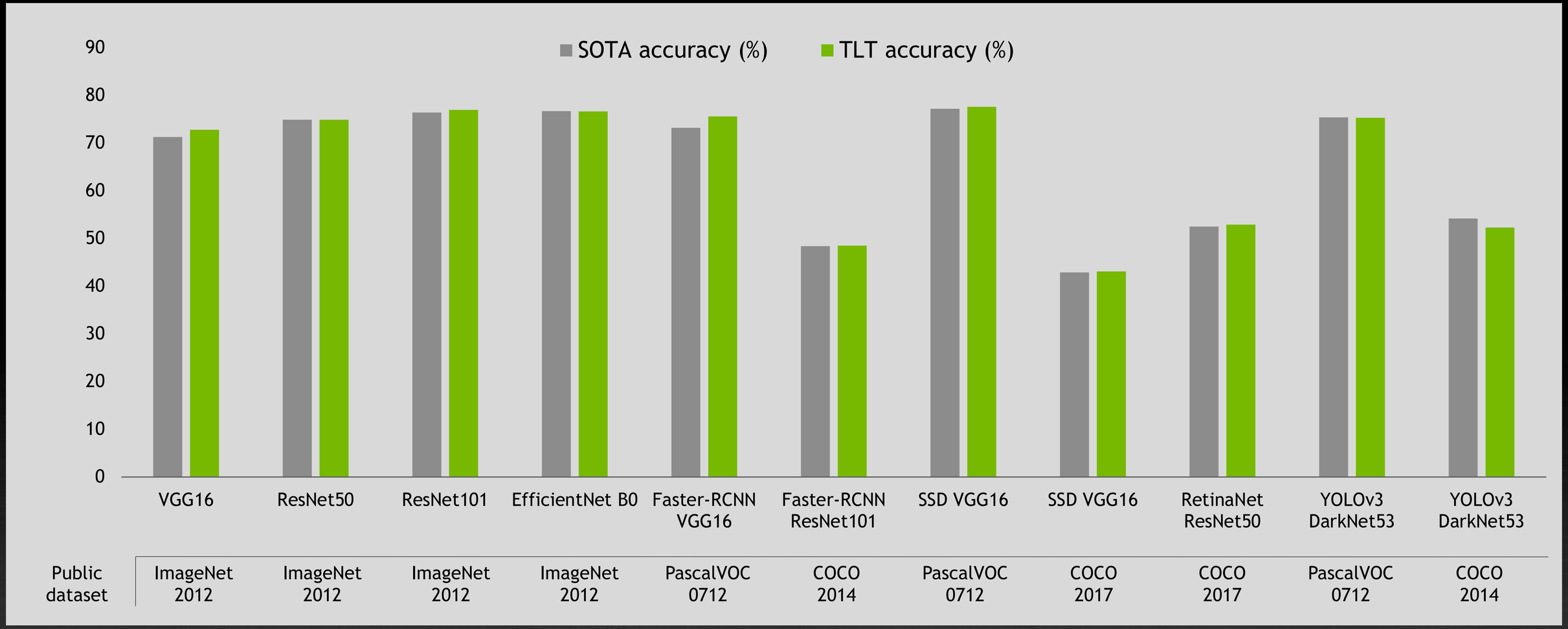
ENABLING BEYOND PRE-TRAINED AI MODELS

100+ combinations of model architectures and backbones

	Image Classification	Object Detection							Segmentation	
		DetectNet_V2	FasterRCNN	SSD	YOLOV3	YOLOV4	RetinaNet	DSSD	MaskRCNN	UNET
ResNet10/18 /34/50/101	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
VGG16/19	✓	✓	✓	✓	✓	✓	✓	✓		✓
GoogLeNet	✓	✓	✓	✓	✓	✓	✓	✓		
MobileNet V1/V2	✓	✓	✓	✓	✓	✓	✓	✓		
SqueezeNet	✓	✓		✓	✓	✓	✓	✓		
DarkNet 19/53	✓	✓	✓	✓	✓	✓	✓	✓		
CSPDarkNet 19/53	✓					✓				
Efficient Net B0/B1	✓		✓	✓			✓	✓		

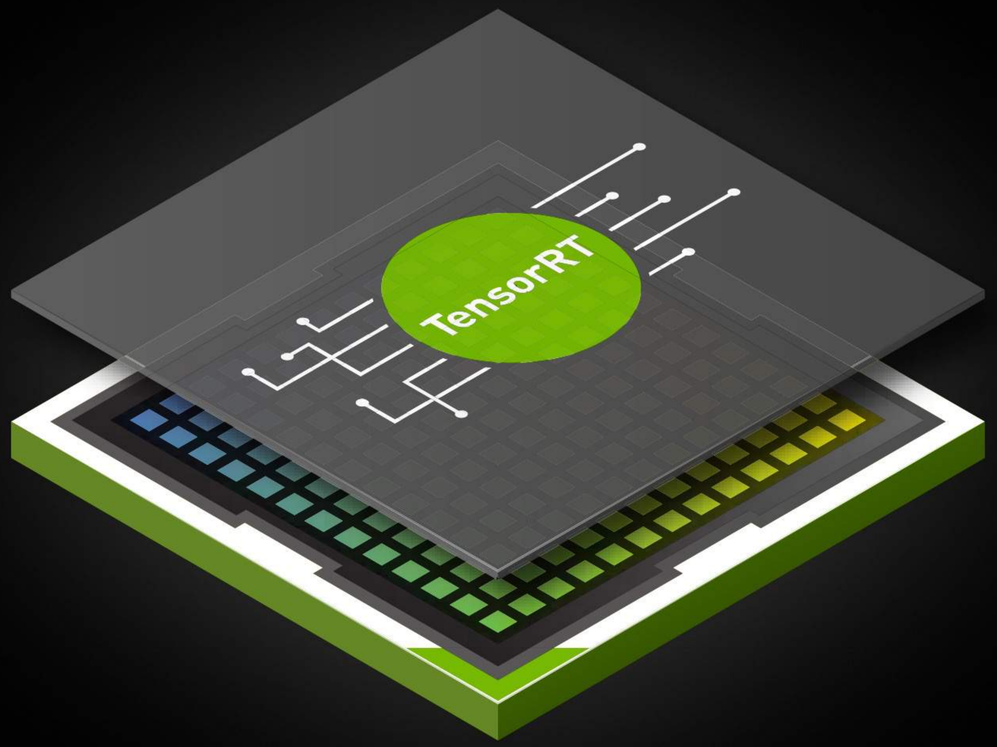
Pre-trained weights trained on OpenImage dataset

ACHIEVING STATE OF THE ART ACCURACY FOR PUBLIC DATASETS



NVIDIA TENSORRT

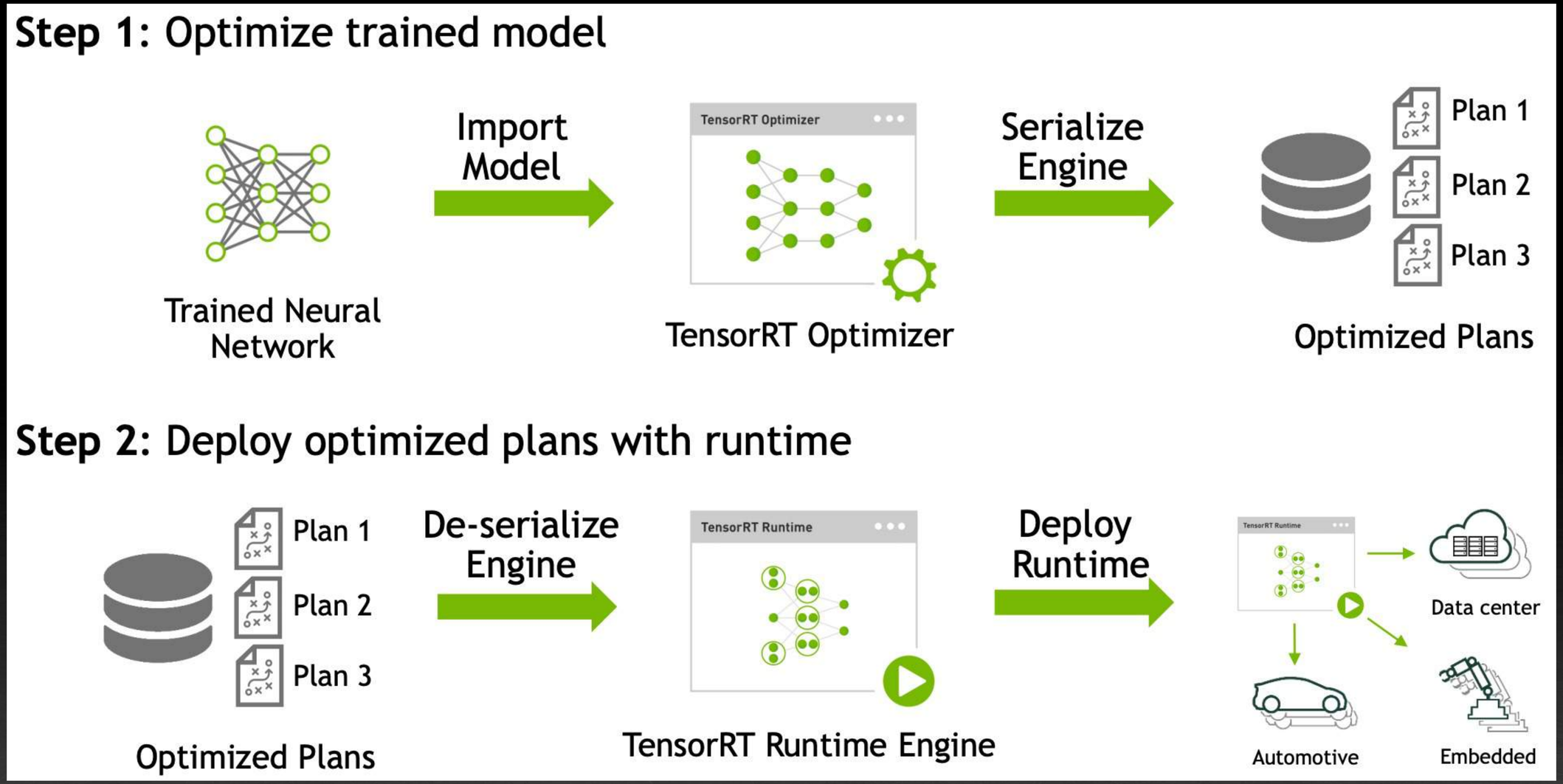
TensorRT Optimizer & TensorRT Runtime



- Layer and Tensor Fusion
- Weight and Activation Precision Calbration
- Kernel Auto-Tuning
- Dynamic Tensor Memory
- Multi-Stream Execution

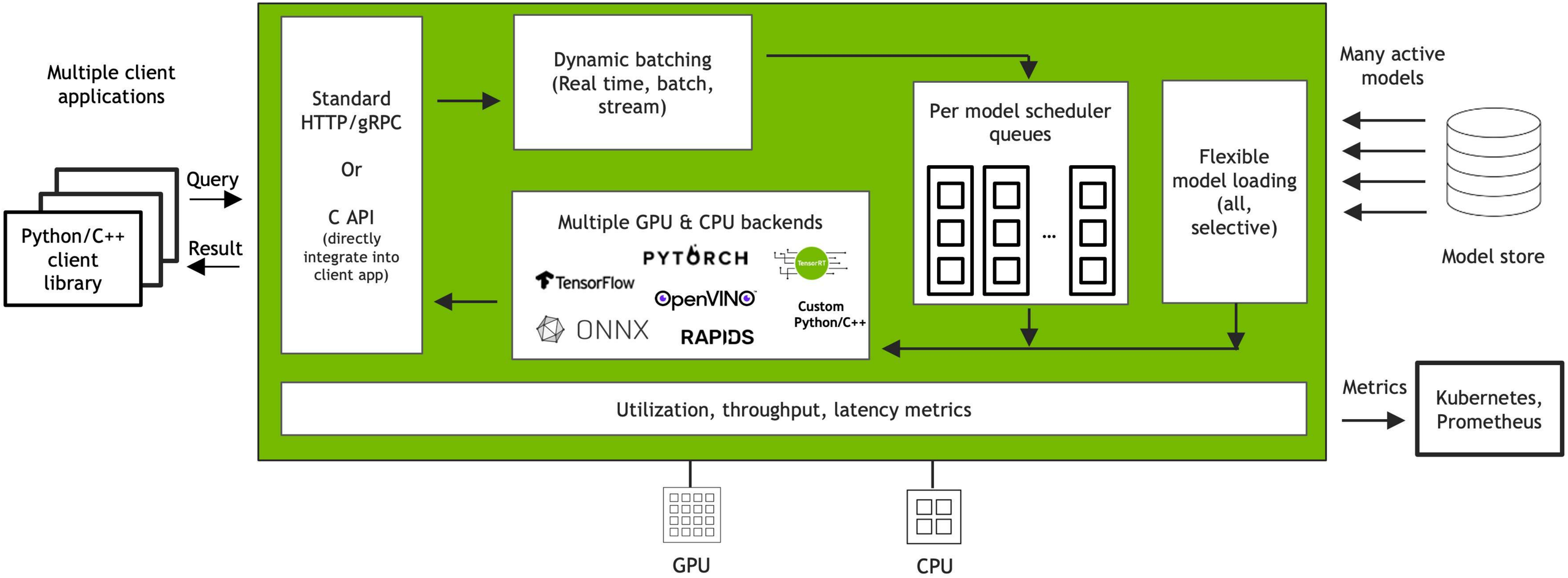
TENSORRT WORKFLOW

Optimize and deploy

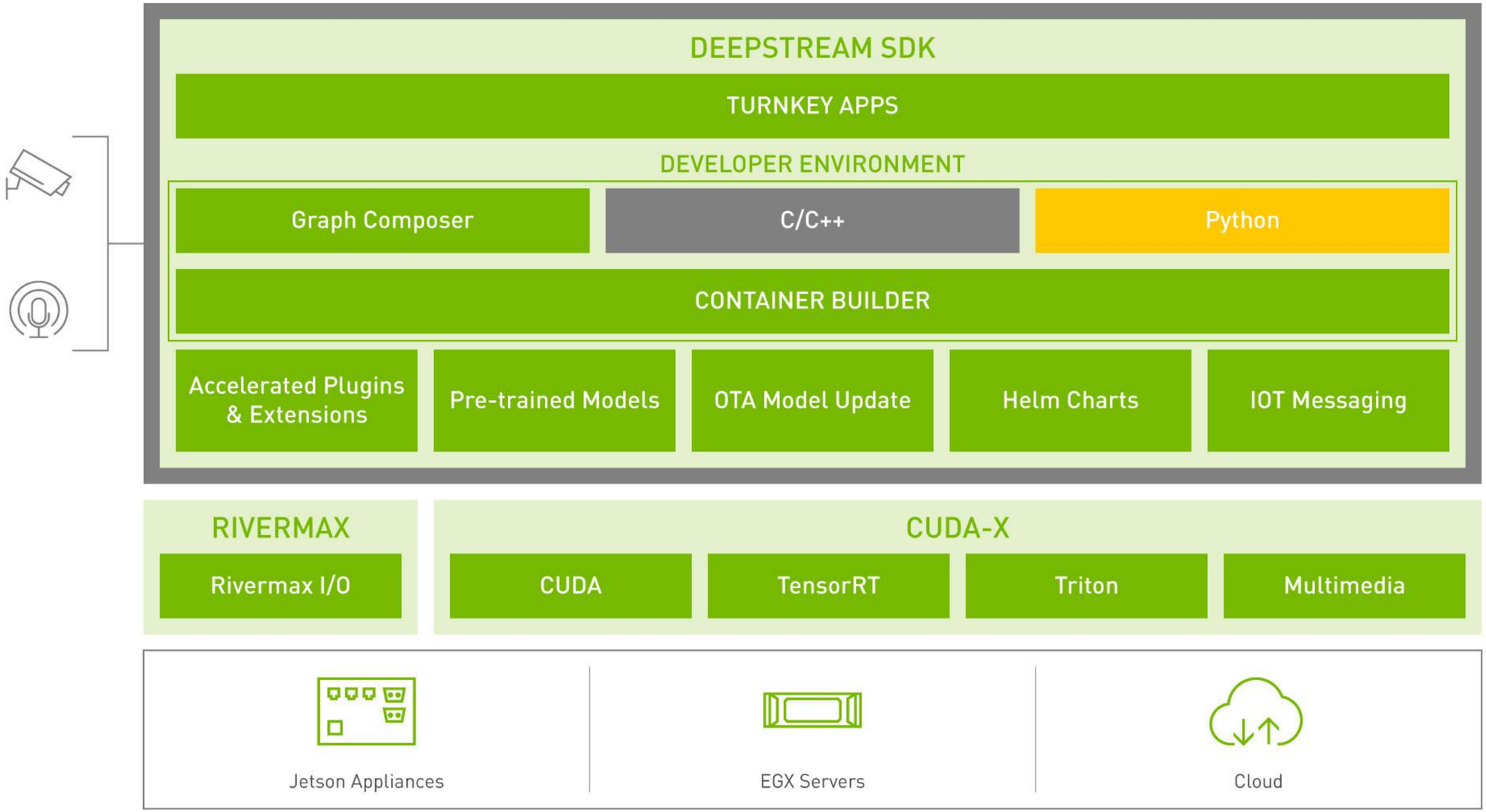


TRITON INFERENCE SERVER

Open-source software for scalable, simplified inference serving

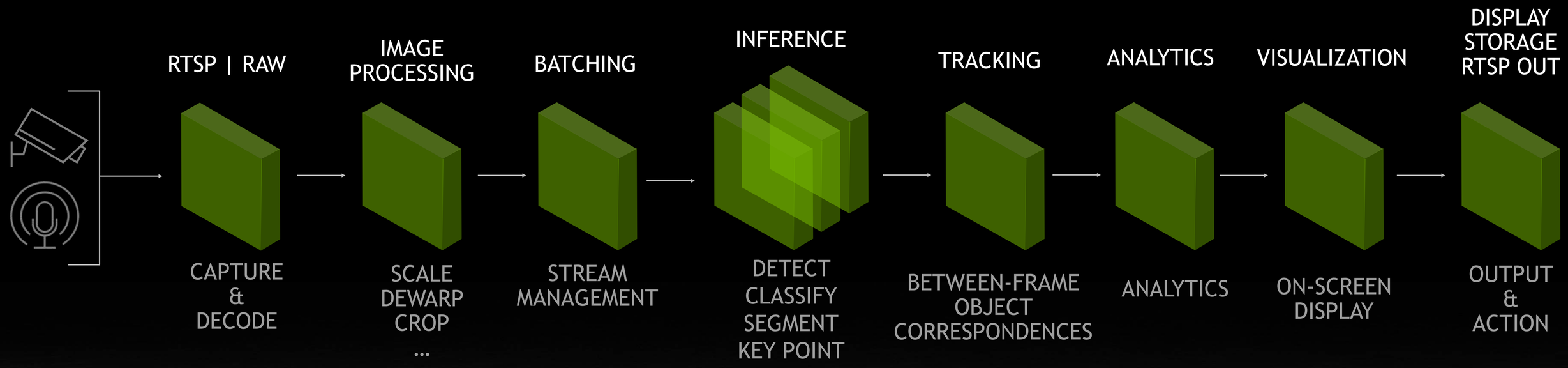


DEEPSTREAM SDK



DEEPSTREAM APPLICATION ARCHITECTURE

End-to-end hardware accelerated pipeline



HARDWARE ACCELERATION							
	ISP	GPU	CPU	GPU	GPU	GPU	GPU	CPU
	GigE	VIC		DLA	CPU	CPU	VIC	HDMI
	NVDEC	CPU						SATA

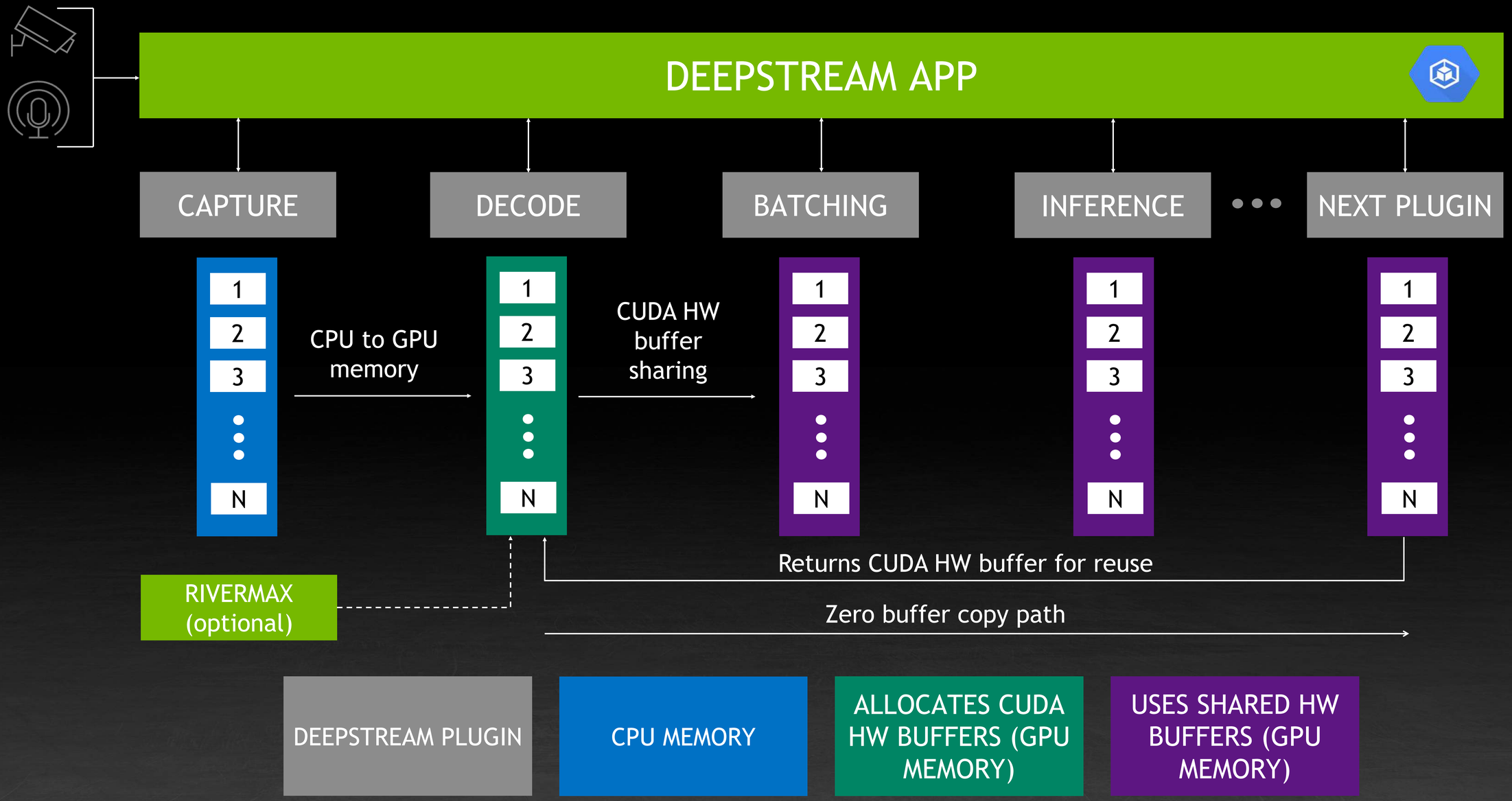
EDGE TO CLOUD DEPLOYMENTS

JETSON APPLIENCES

EGX SERVERS

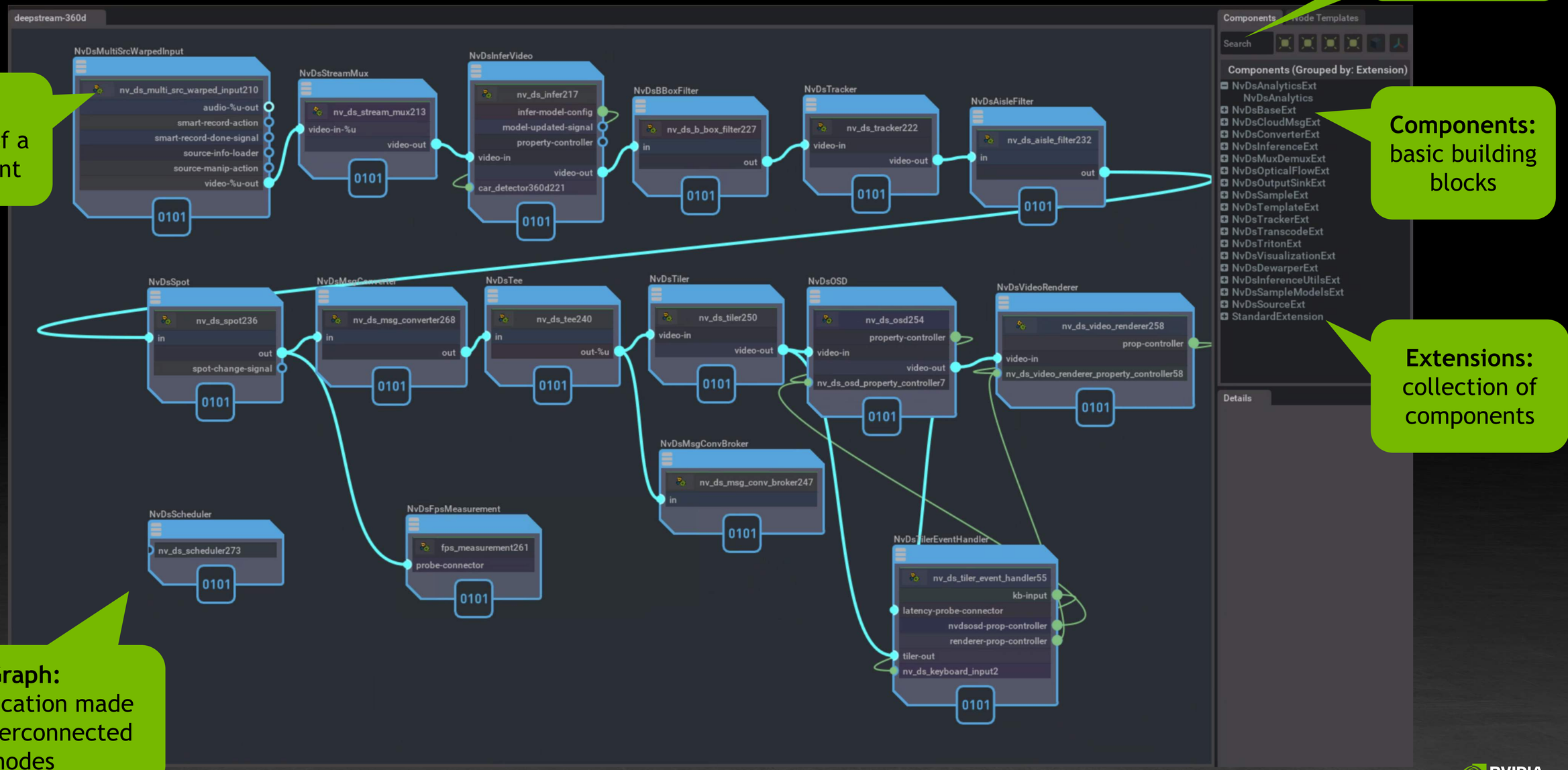
CLOUD

PIPELINE EFFICIENCY WITH ZERO MEMORY COPIES



NVIDIA GRAPH COMPOSER

Drag + drop development environment



Node: Instance of a Component

Graph: An application made with interconnected nodes

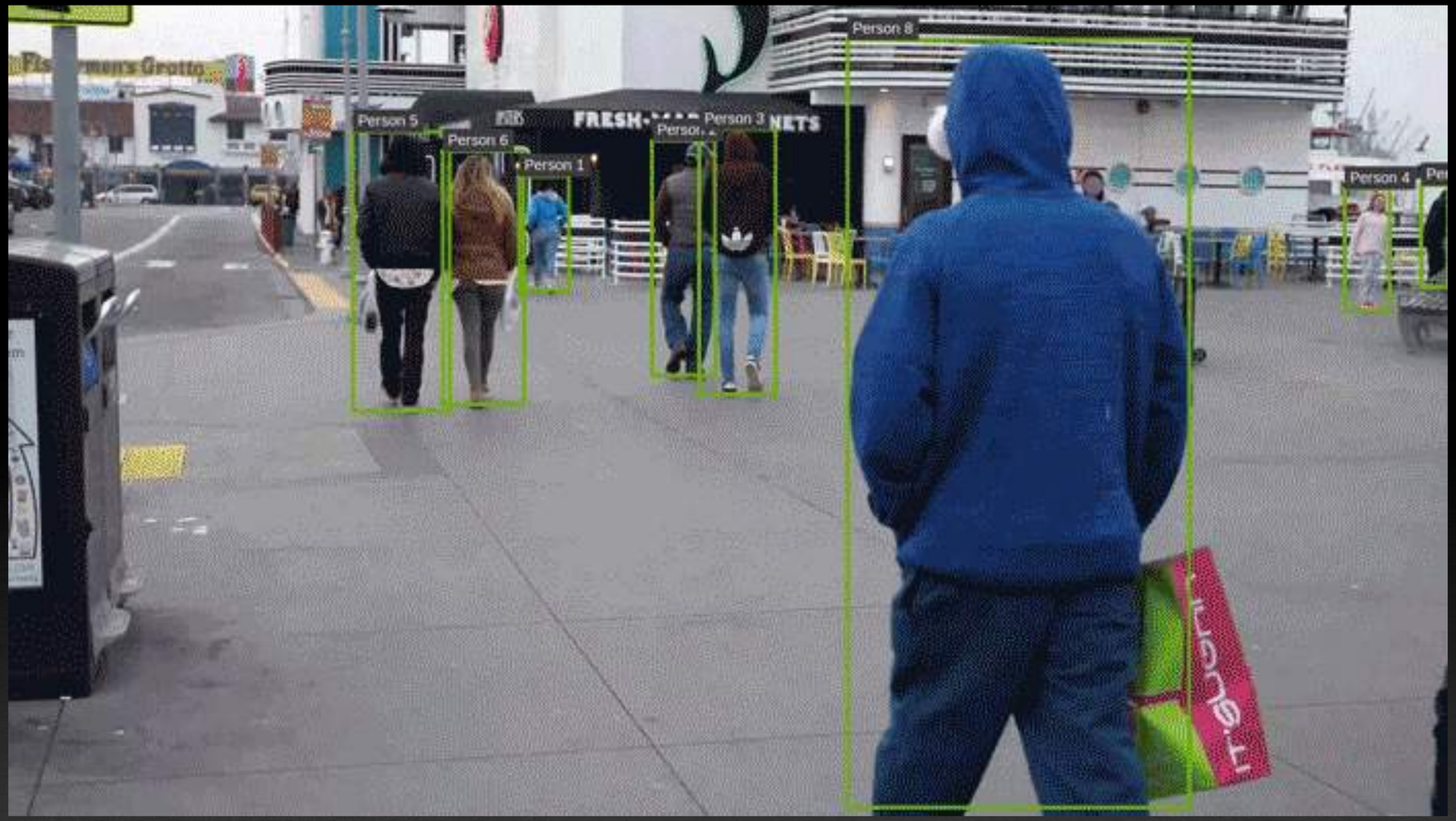
150 building blocks

Components: basic building blocks

Extensions: collection of components

DEEPSTREAM VIDEO DEMO

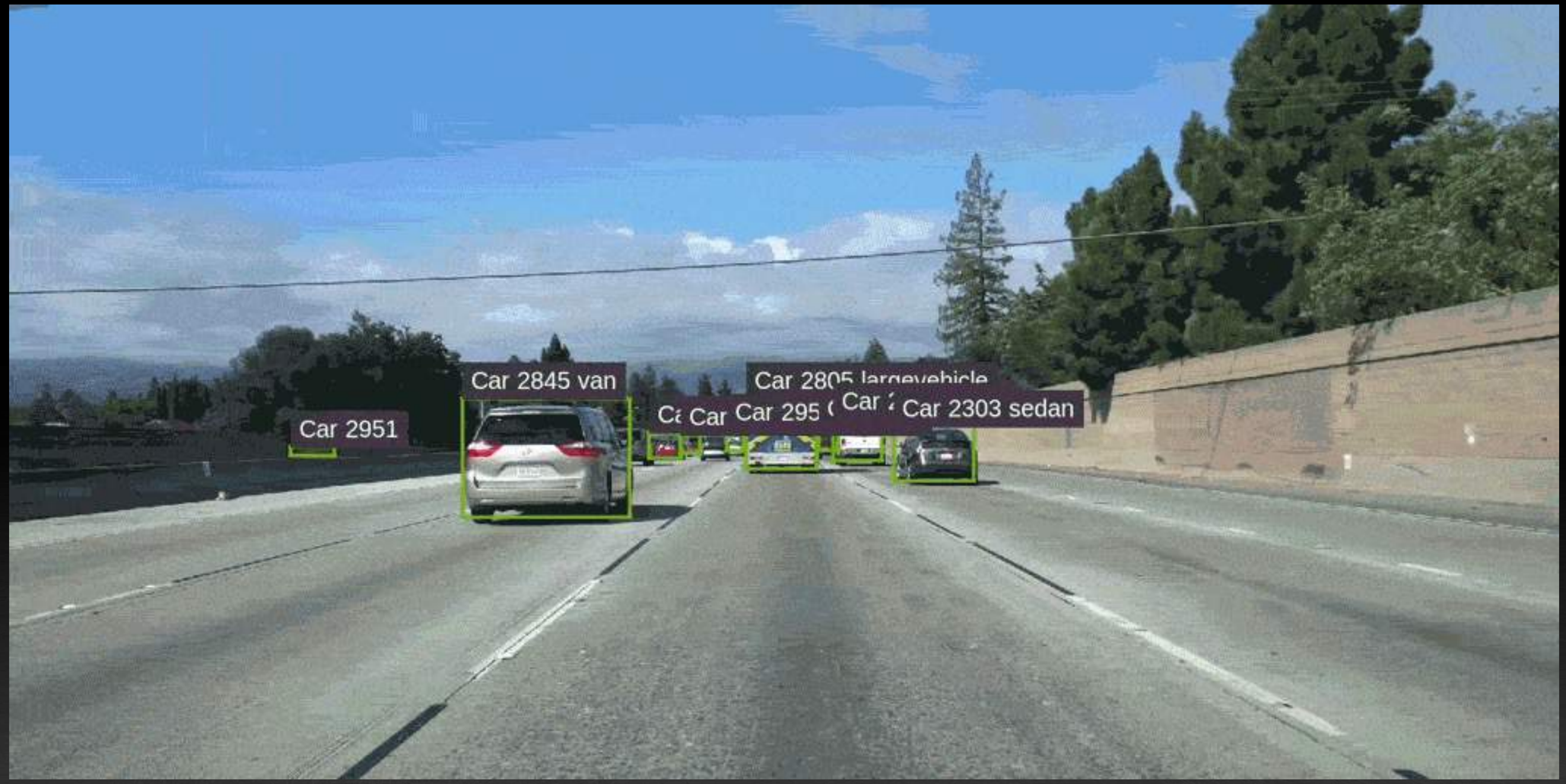
PeopleNet: detecting people



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

DEEPSTREAM VIDEO DEMO

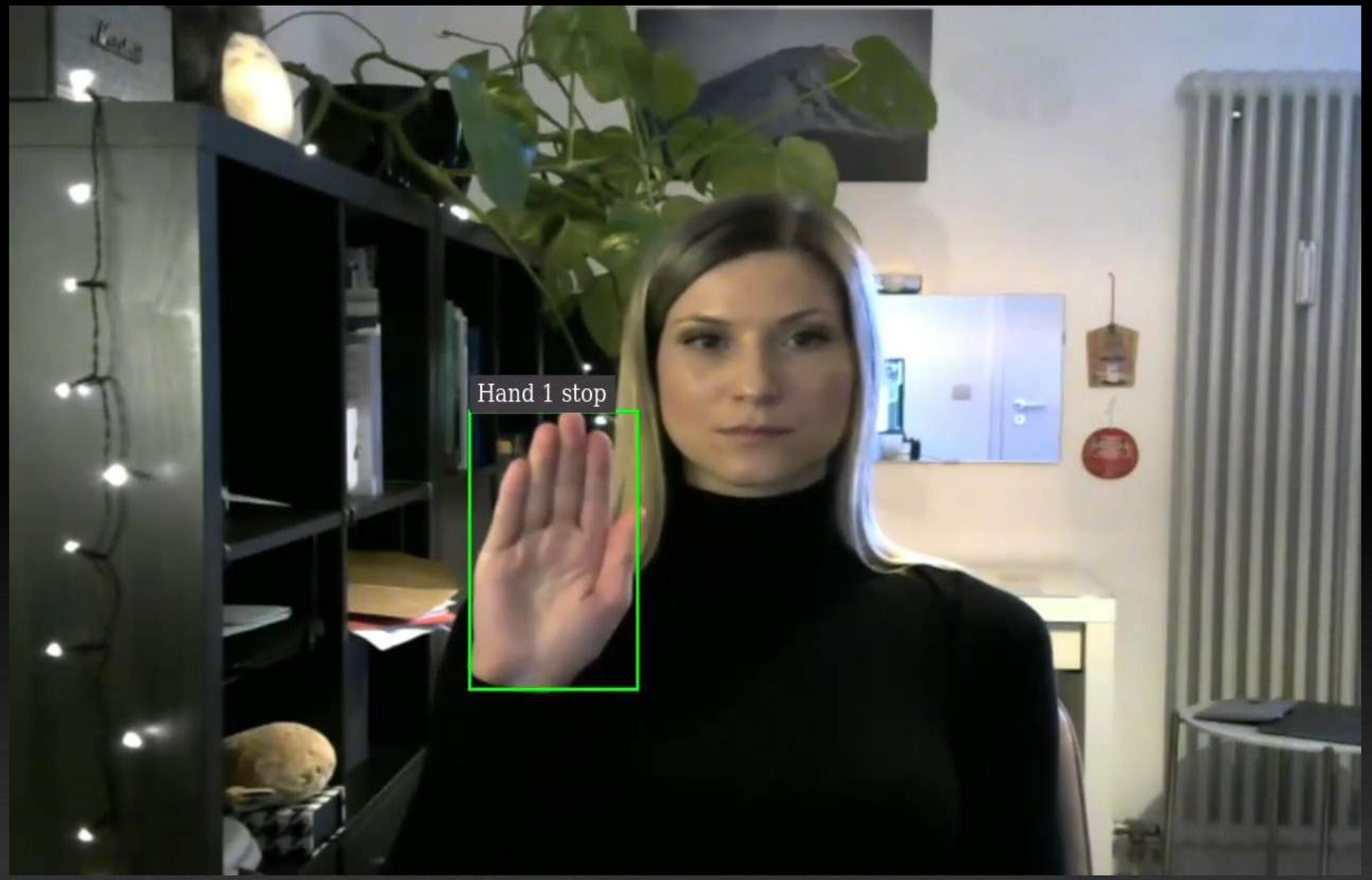
DashCamNet and VehicleTypeNet in action



https://ngc.nvidia.com/catalog/models/nvidia:tlt_dashcamnet
https://ngc.nvidia.com/catalog/models/nvidia:tlt_vehiclemakenet

DEEPSTREAM VIDEO DEMO

Hand detector adapted with TAO Toolkit and GestureNet is used out the box



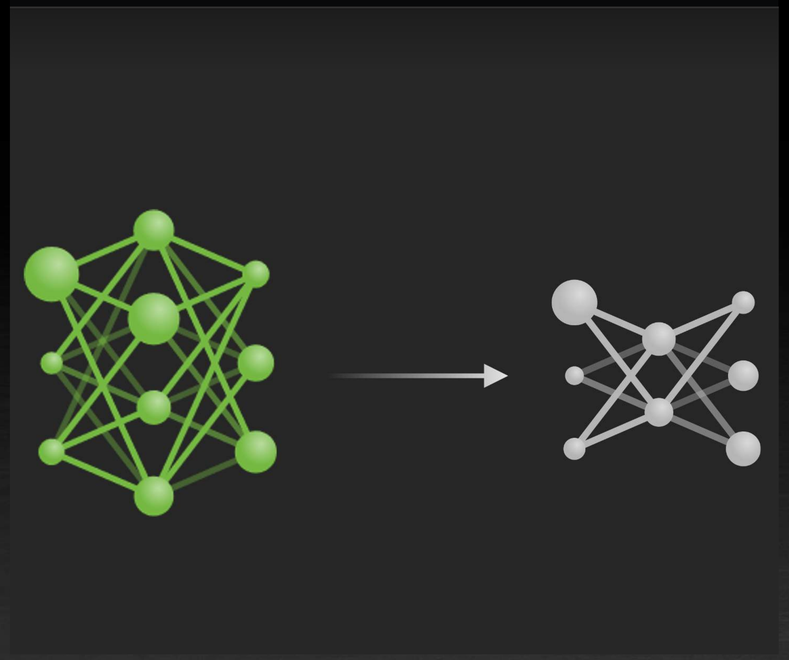
https://github.com/NVIDIA-AI-IOT/gesture_recognition_tlt_deepstream

SUMMARY

For efficient AI video analytics

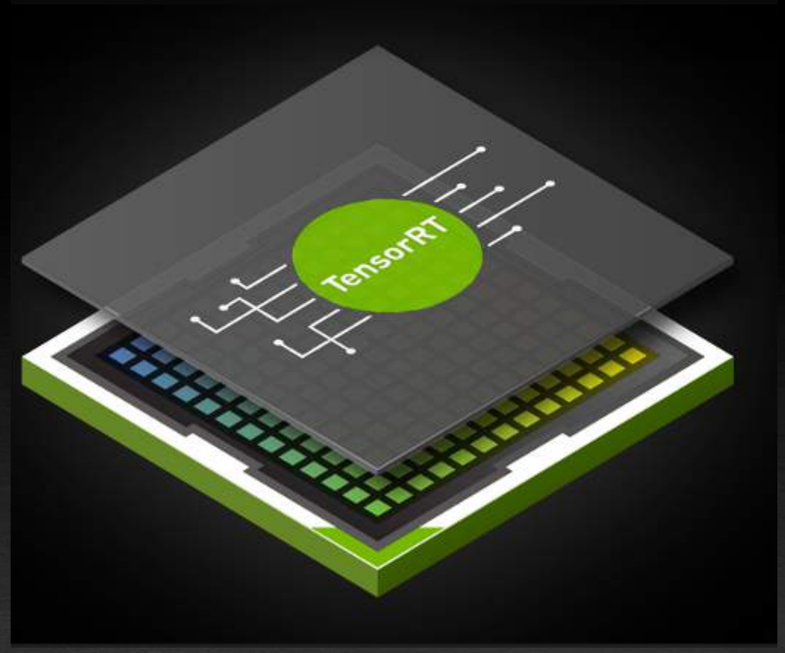
TRAIN

Transfer learning, AMP, multi GPU, data augmentation, pruning and quantization aware training with TAO Toolkit



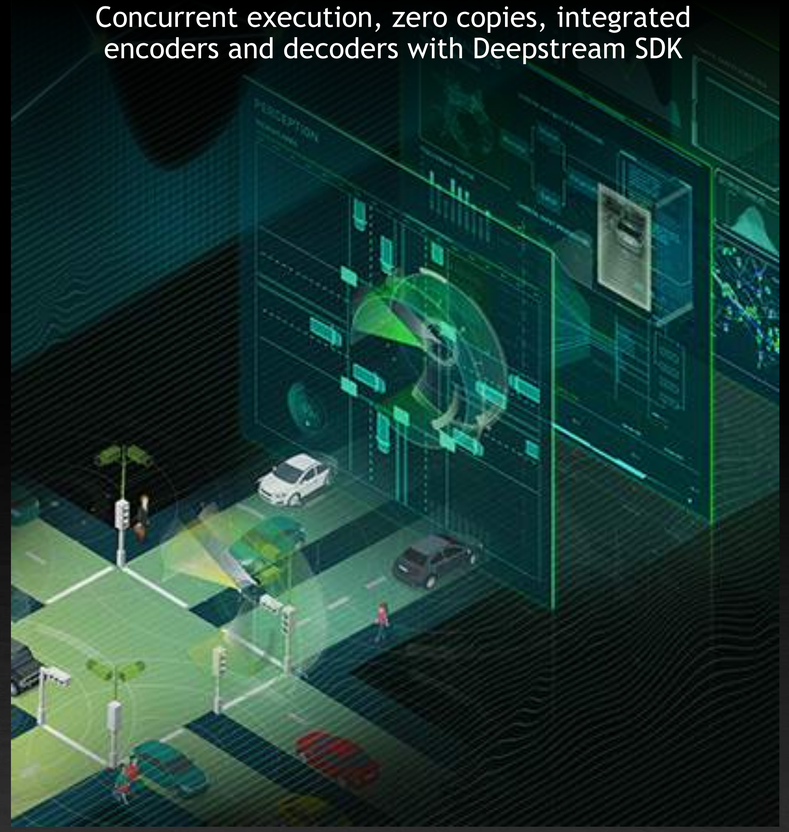
OPTIMIZE

Layer fusion, kernel auto-tuning, dynamic tensor memory, etc. with TensorRT



DEPLOY

Concurrent execution, zero copies, integrated encoders and decoders with Deepstream SDK



DEVELOPER RESOURCES

Powerful end-to-end AI video analytics made easy

TAO Toolkit https://developer.nvidia.com/tao-toolkit

DeepStream SDK https://developer.nvidia.com/deepstream-getting-started

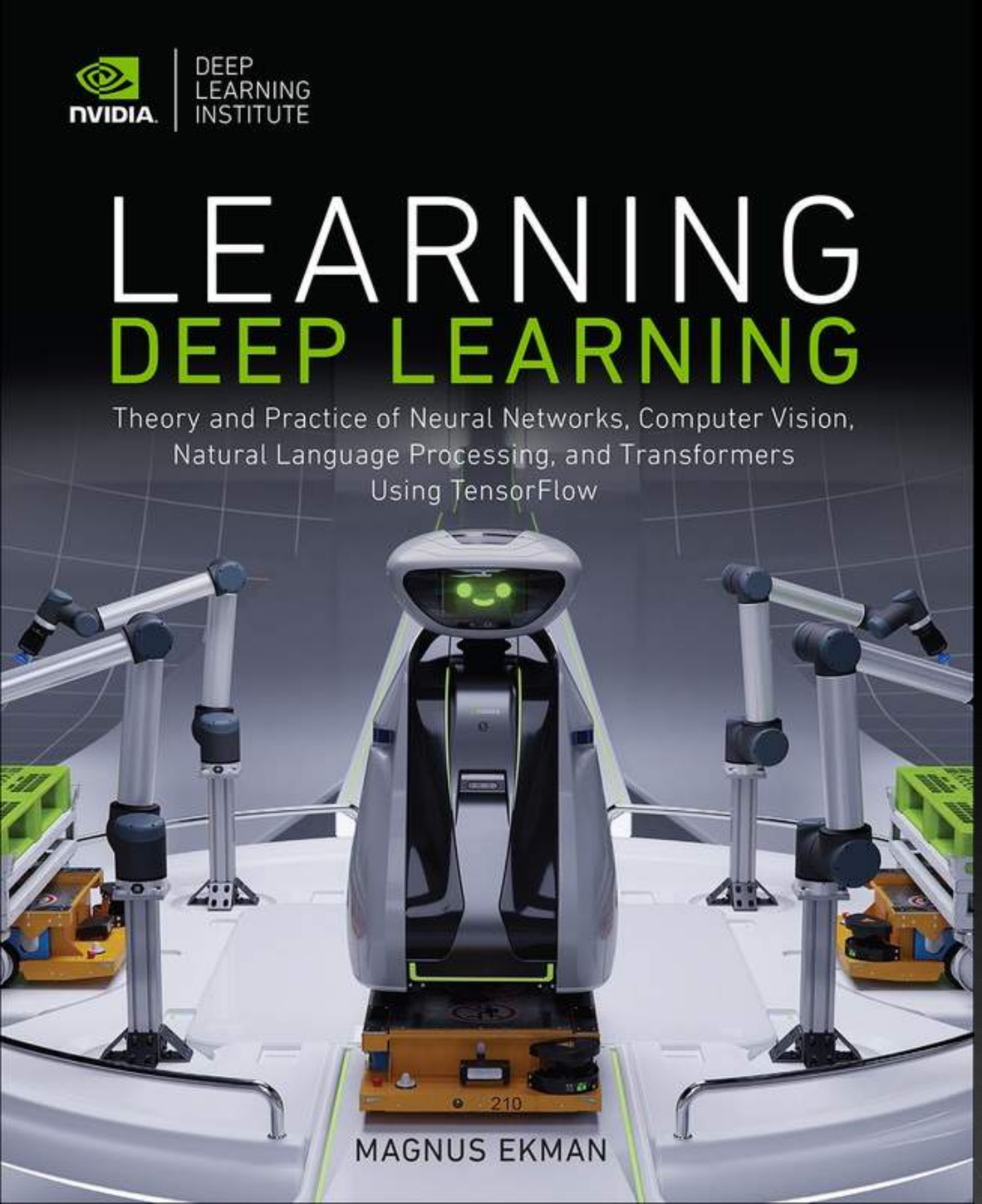
TensorRT https://developer.nvidia.com/tensorrt

Triton Inference Server: https://github.com/triton-inference-server

NVIDIA GPU Cloud (NGC) https://ngc.nvidia.com/

Developer forums https://forums.developer.nvidia.com/

NVIDIA Deep Learning Institute https://www.nvidia.com/en-us/training/



THEORY AND PRACTICE OF NEURAL NETWORKS, COMPUTER VISION, NATURAL LANGUAGE PROCESSING, AND TRANSFORMERS USING TENSORFLOW

- Full-colour guide
- Illuminates both the core concepts and the hands-on programming techniques needed to succeed
- Shows how to build advanced architectures, including the Transformer
- Includes concise, well-annotated code examples using TensorFlow with Keras; corresponding PyTorch examples are provided online

Available at the GOTO Copenhagen Conference Bookstore

NVIDIA