Software for AI-powered Platforms

Qualcomm and NVIDIA software overview

Łukasz Grzymkowski, Engineering Manager
Arrow Electronics

What it feels like to be a data scientist since 2022

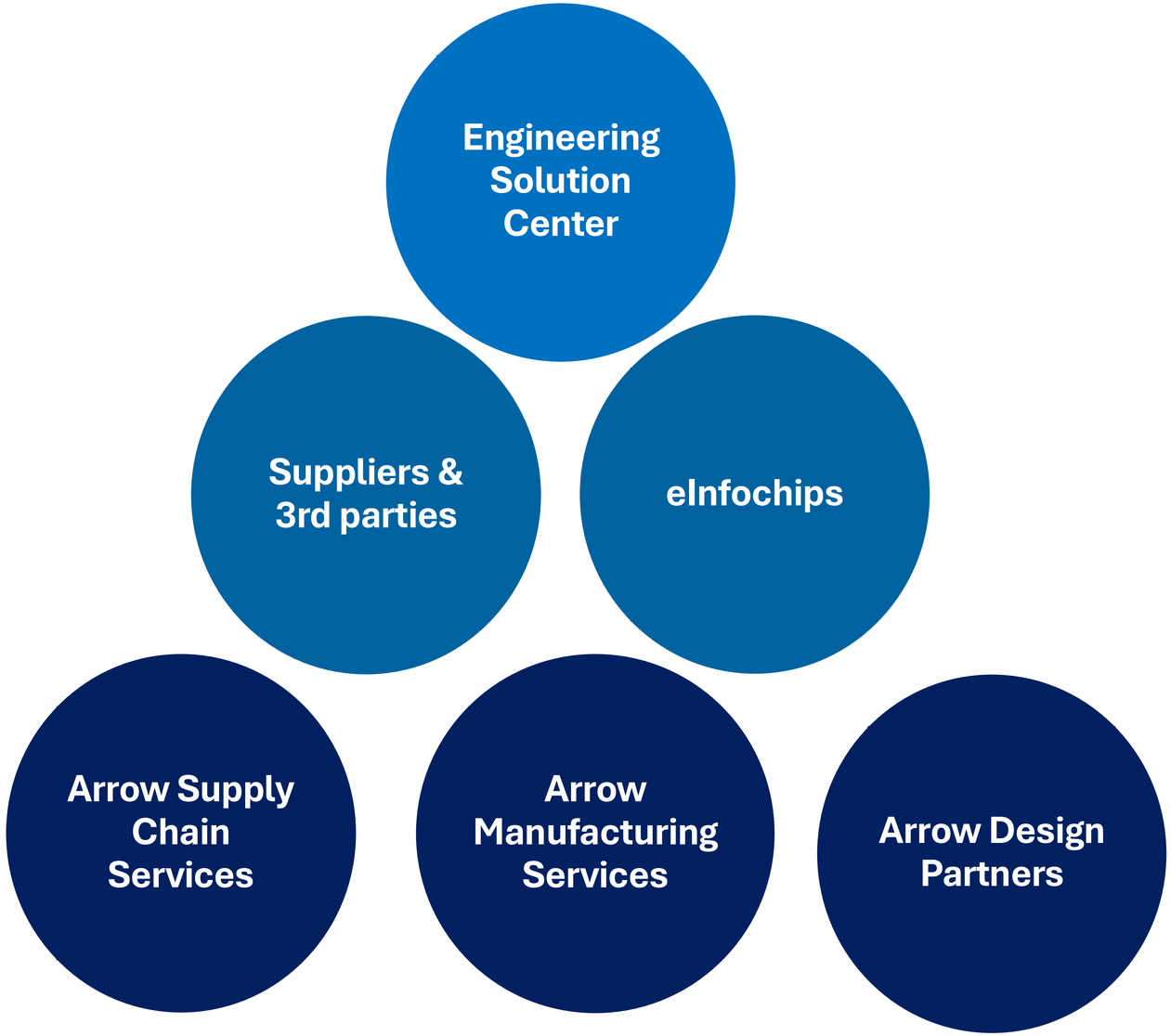
Engineering Solution Center

Design house supporting customer projects and offering outsourcing services and consultancy

Complete product development capabilities:
- Hardware and electronic engineers
- Software, Firmware engineers
- Data Scientists, ML engineers

Close connection with Arrow suppliers and partners.

Support model:
Consultancy (time & material)
Fixed bid (statement-of-work)



Case studies: AI and Machine Learning

Coffee machines

TinyML using AC measurements only to understand utilization of grinder and coffee machine

Tasks include detecting type of coffee, volume, grinding settings, percolation time, detecting anomalies.

Experiment and dataset collection, through model selection and deployment on MCUs.

Conference Equipment

Extract and recognize speaker voice from mixture of voices.

Recognize through face detection, detect lip movement for active speaker.

Conference tools with microphone array, cameras to be equipped with AI to improve user experience.

Multiple models running on microprocessor with acceleration.

Multi-modal LLM fine-tuning

Image and text-modality to label video files, recognize objects and people from digital archives.

Fine-tuning of Pixtral model using LORA technique.

Use case is to label movies, scenes, sport events, digitalized photo albums, but also human-interface for vision monitoring systems (retail, warehouse logistics).

Firmware and Machine Learning Services

Machine Learning and Software Services

We offer **end-to-end support for machine learning, artificial intelligence and firmware development** to meet your needs and objectives.

Embedded Systems

Hands-on expertise in developing all crucial parts of the system including BSP, driver development, kernel patches integration, application development.

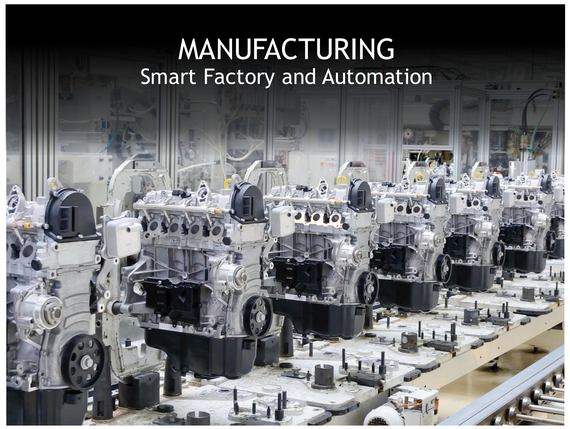
Artificial Intelligence and Machine Learning

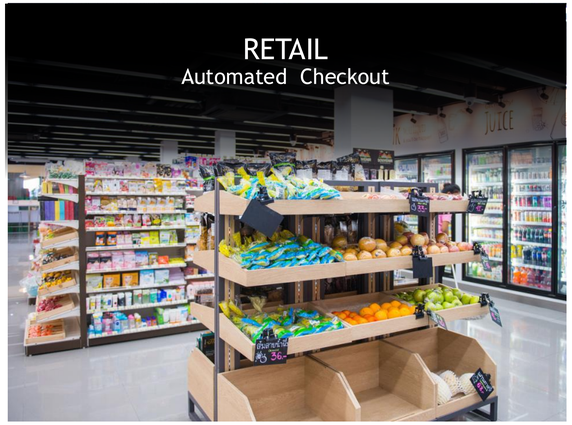
Deep Learning and Machine Learning Services. Highly-customized embedded solutions running on advanced machine learning algorithms.

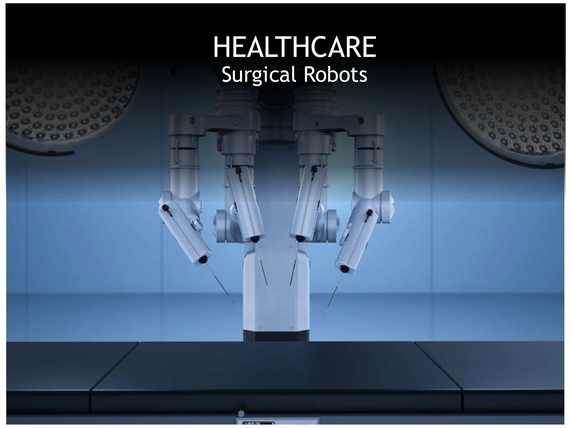
THE C PROGRAMMING LANGUAGE
C++
python™
yocto PROJECT
Zephyr™
freeRTOS
Azure RTOS ThreadX
TensorFlow
Hugging Face
PyTorch
Bluetooth™
zigbee
LoRa
UWB
THREAD
Wi Fi™
NB-IoT
Microsoft Azure
amazon web services

EDGE AI TRANSFORMS NEARLY ALL INDUSTRIES

AI could potentially deliver an additional economic output of around US $13 trillion by 2030







Qualcomm

Qualcomm Dragonwing

Powered by leading-edge on-device AI, computing, and connectivity, **Qualcomm Dragonwing™** products and solutions unlock potential, so today's industries and businesses can scale to new heights.

Purpose-built

A portfolio of solutions purpose-built to address the connectivity, computing and Intelligence needs to elevate potential across industries.

Edge intelligence

Next-gen intelligence, custom-designed for business and industry. On-device hardware and software AI solutions for multiple use cases.

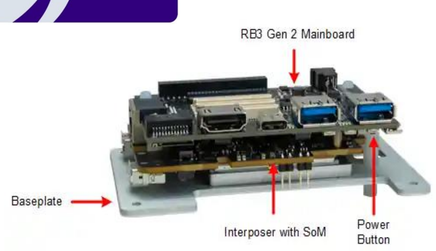
Established and trusted

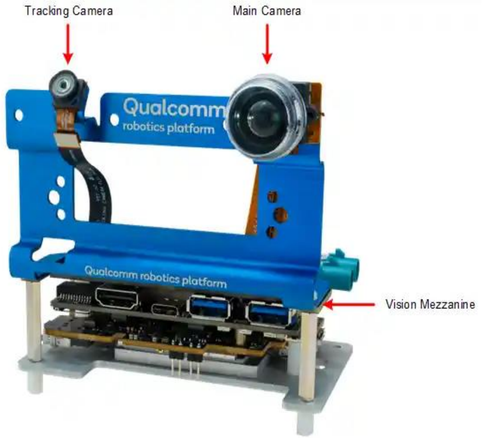
Dragonwing is built on decades of trusted innovation and expertise, bringing cutting edge technologies across industries.

Arrow

Qualcomm RB3 Gen 2

Development kits



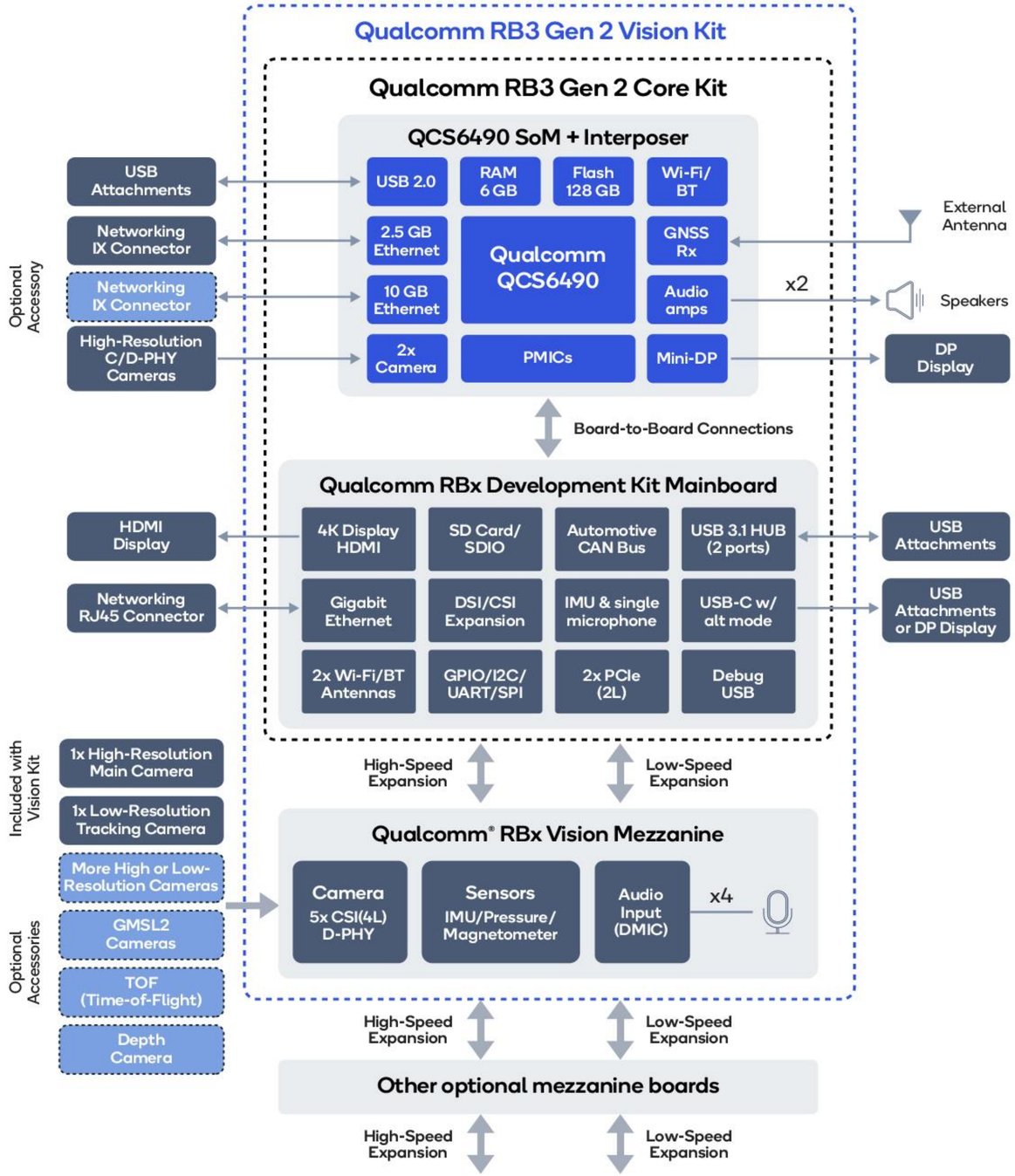


Core Kit

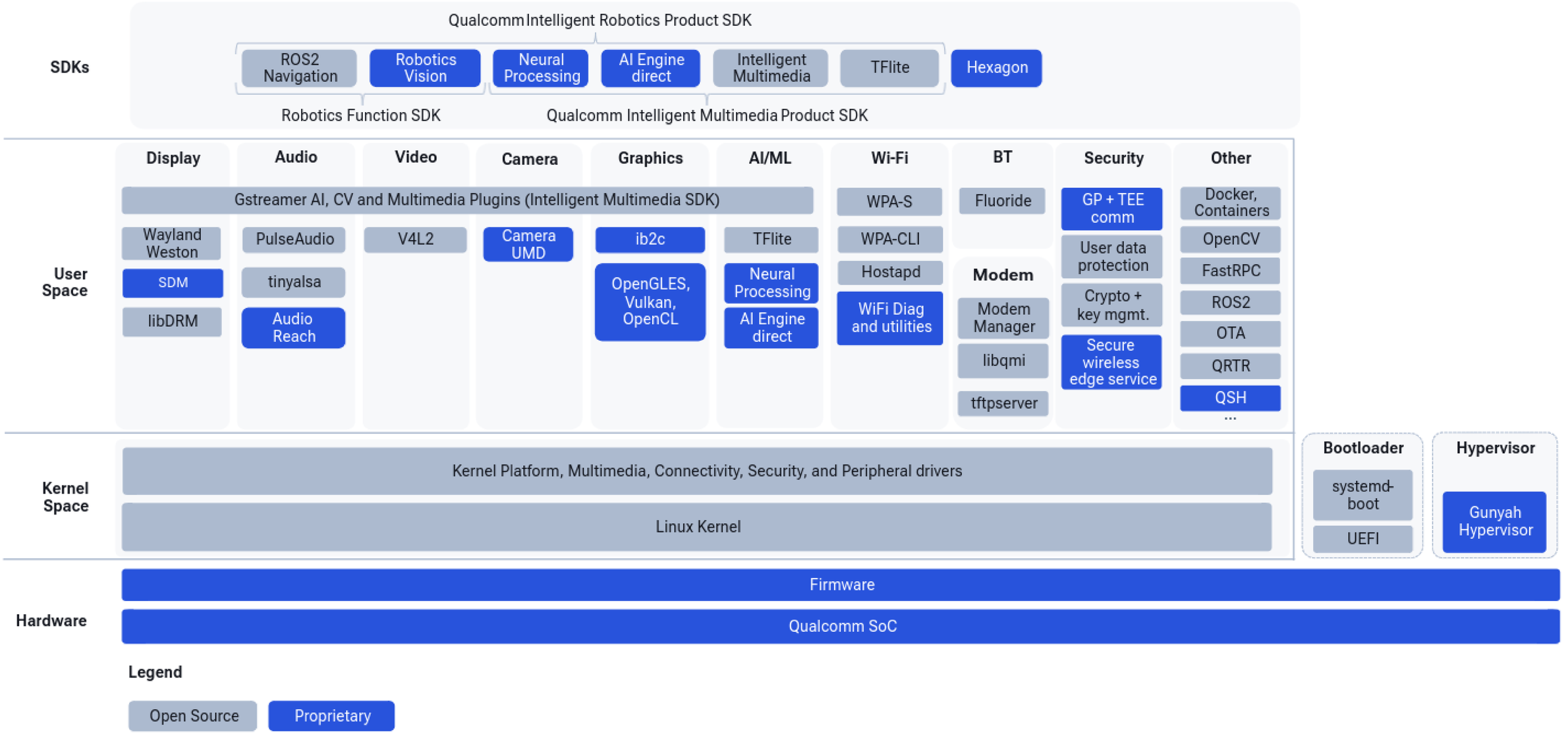
- Qualcomm QCS6490 processor RB3 Gen 2 mainboard
- System on module (SoM) + interposer
- Power supply (12 V)
- USB-A to USB-C cable
- 2 speakers
- Pick tool to help access DIP switches

Vision Kit

Qualcomm RB3 Core kit +
Qualcomm Vision mezzanine board
IMX577 camera
OV9282 tracking camera



Application development environments

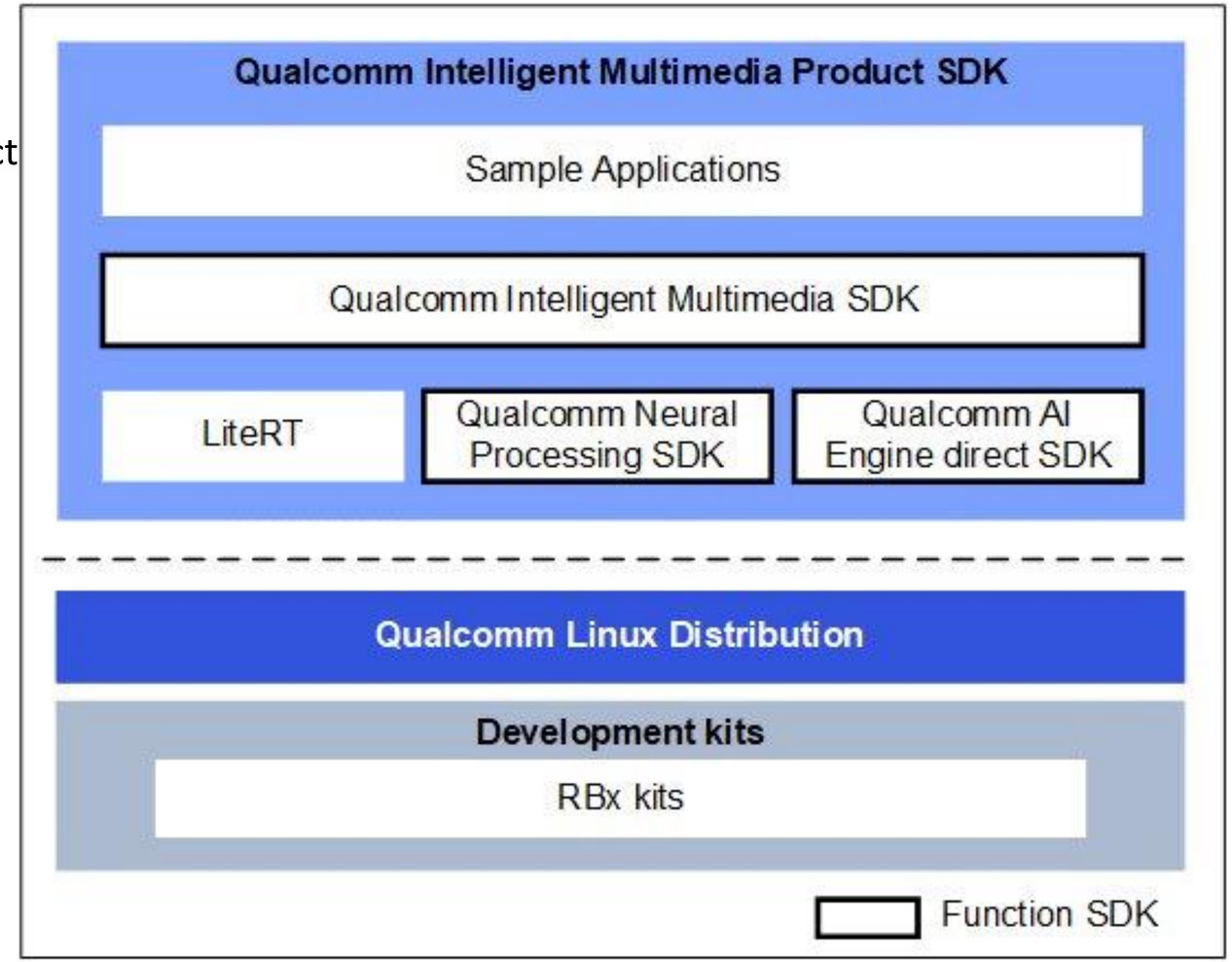


Application development environments

QIMP

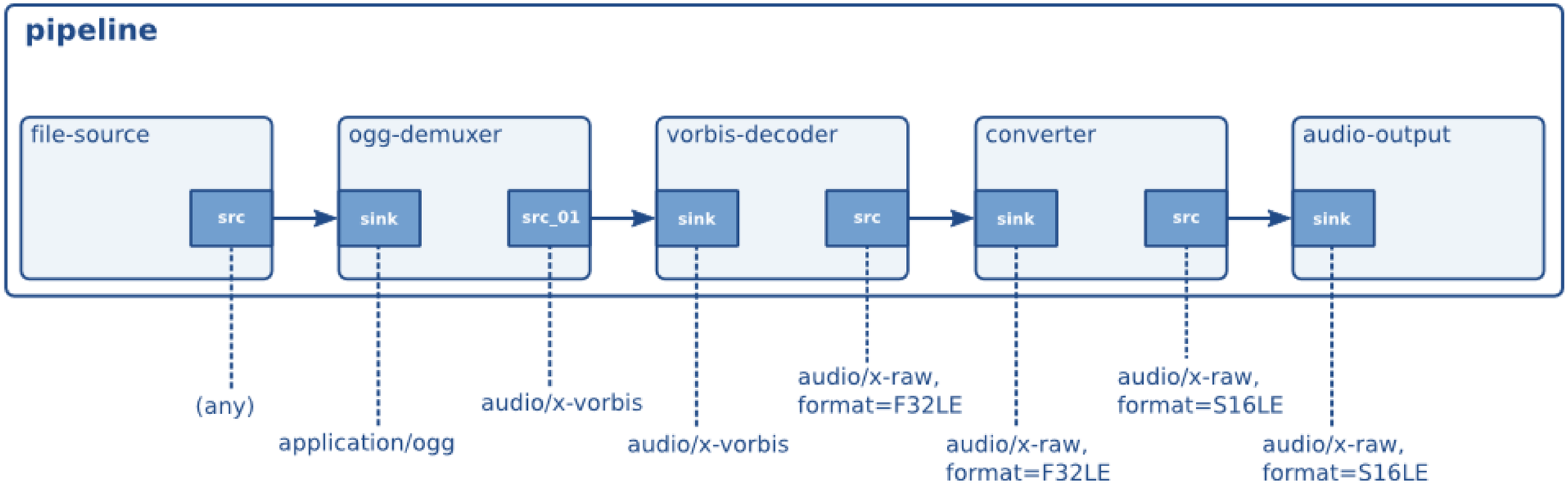
The **Qualcomm Intelligent Multimedia Product** (QIMP) SDK is a collect
of four standalone function SDKs:

- Qualcomm® **Intelligent Multimedia** SDK (IM SDK)
- Qualcomm® **Neural Processing** SDK
- Qualcomm® **AI Engine direct** SDK
- TensorFlow Lite (currently known as **Lite Runtime** or **LiteRT**).



QIMP SDK software stack

Architecture



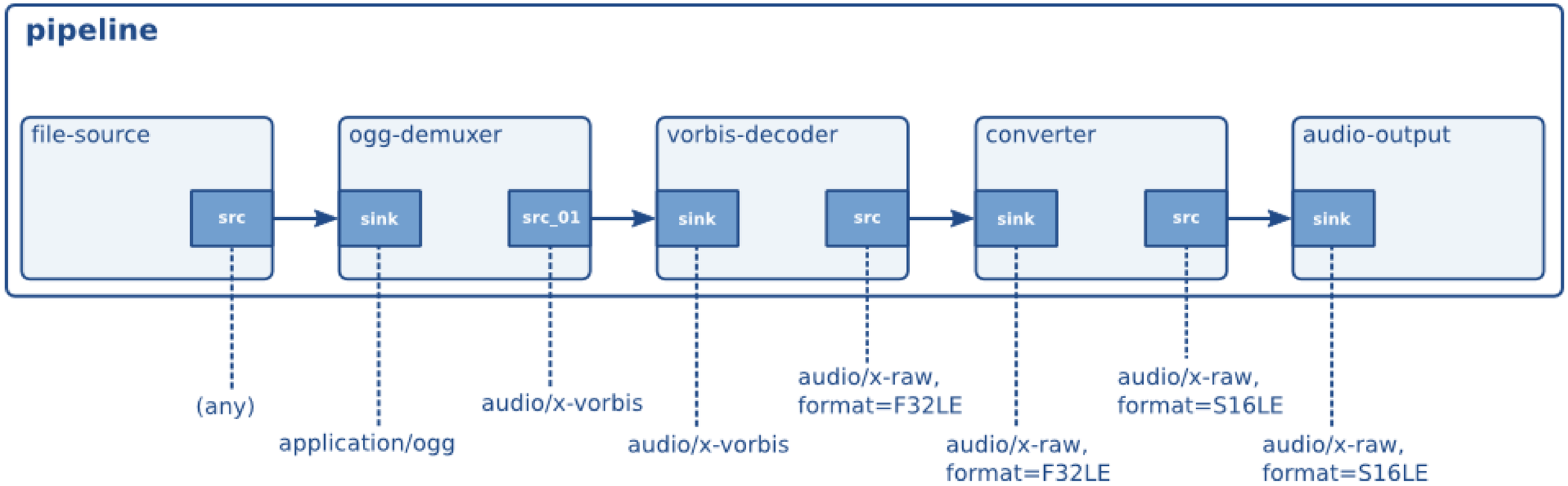
Modular pipeline system consists of interconnected bins called Gstreamer **plugins.**

Each plugin is responsible for specific tasks (e.g., **decoding, encoding, filtering**).

Opensource Multimedia Framework

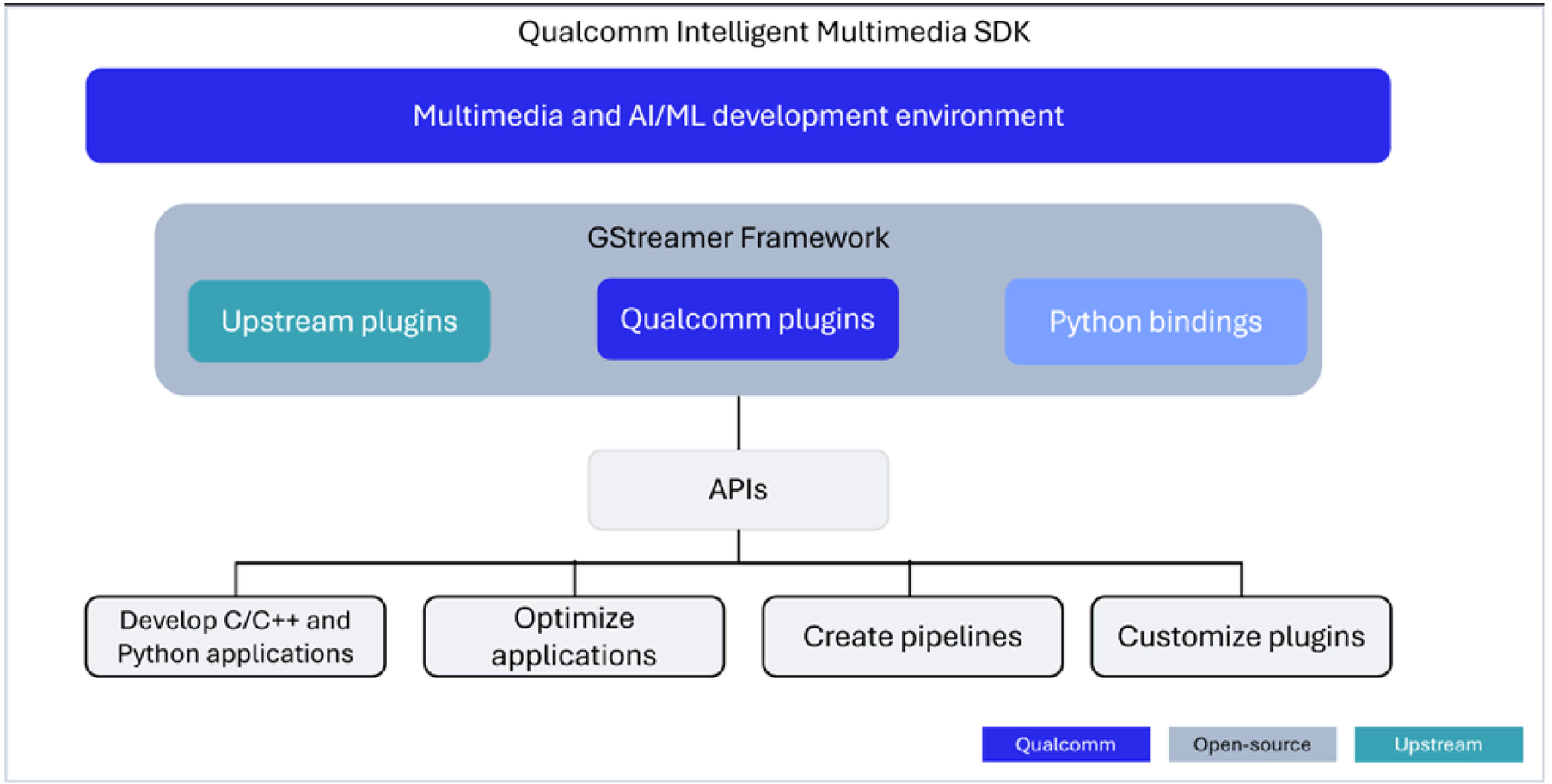
GStreamer is a framework designed to handle **audio, video, and streaming** operations. It provides a **flexible pipeline-based architecture**, making it ideal for applications such as:

- **media players**
- **video processing**
- **real-time streaming**
- **embedded multimedia solutions**



gstreamer

Qualcomm and Gstreamer plugins



Qualcomm IM SDK overview

gstreamer

Display, camera, encode and decode plugins

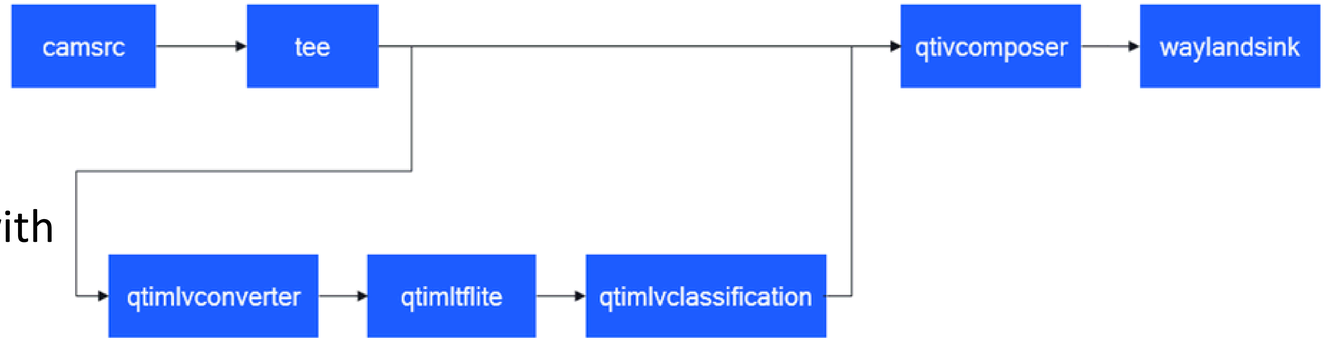
ARROW

gstreamer

Machine Learning plugins

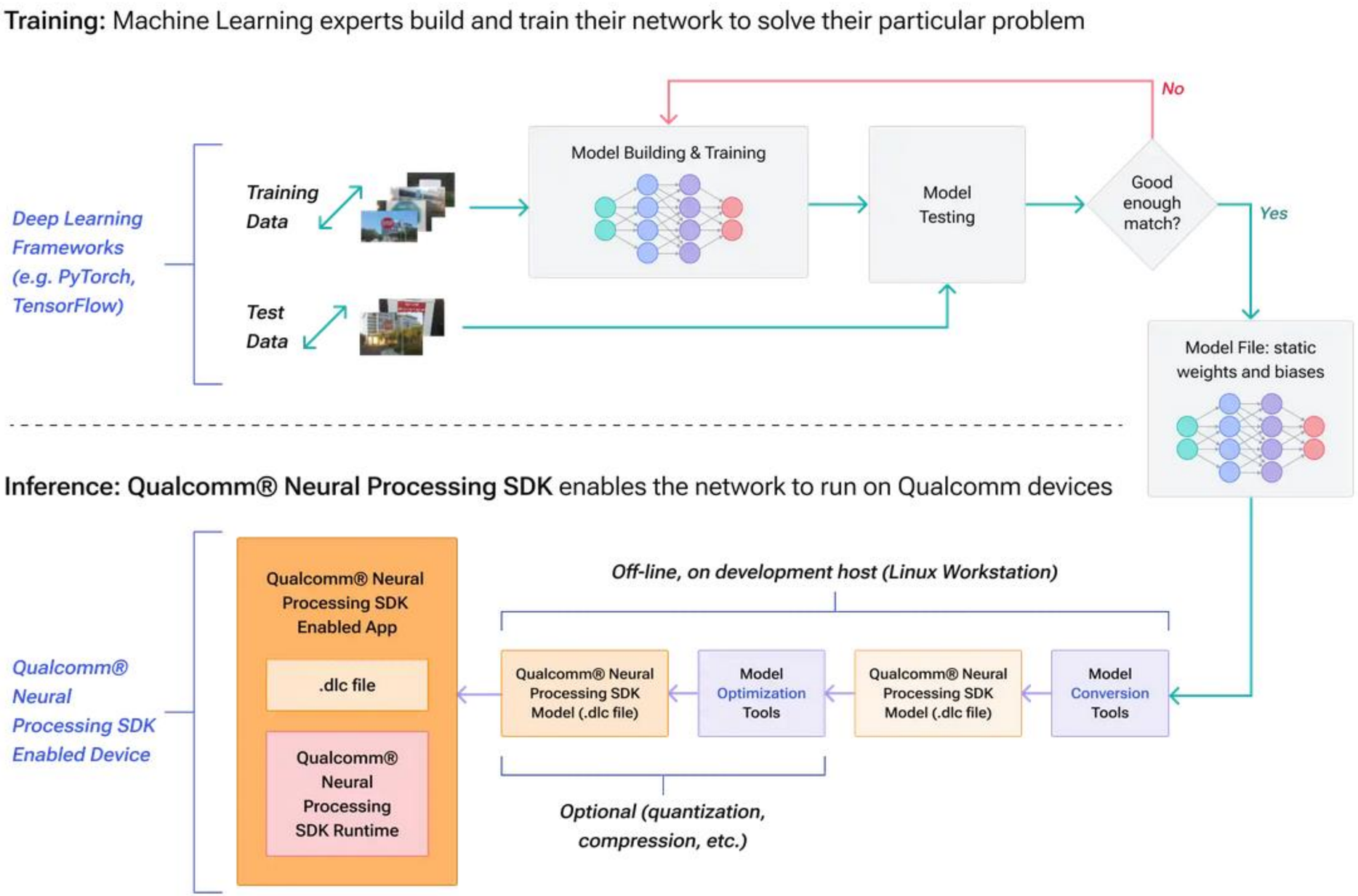
Gstreamer plugins for AI

Example of image classification and display with **LiteRT**



Process	Plugin	Overview
Preprocessing	qtimlvconverter	Converts video to tensor stream with **color conversion, scaling, and normalization**.
Inferencing	qtimltflite	Loads and runs the model, modifies the graph, and outputs inference results.
Postprocessing	qtimlvclassification	Converts inference results into video/text, applies thresholds, and prepares output.
Compositing	qtivcomposer	Merges the original video stream with classification results into a single output.

Qualcomm Neural Processing SDK Workflow



Qualcomm AI Hub

FILTER BY Clear All

Search models

Domain/Use Case

- Audio
- Computer Vision
- Generative AI
- Multimodal

Device

Search devices

Chipset

- ✓ Qualcomm QCS6490 (Proxy)
- Qualcomm QCS8250 (Proxy)
- Qualcomm QCS8275 (Proxy)

39 models

DeepLabV3-Plus-MobileNet-...

Quantized Deep Convolutional Neural Network model for semantic segmentation.

Semantic Segmentation

DenseNet-121-Quantized

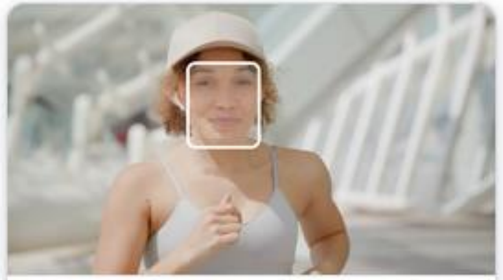
Imagenet classifier and general purpose backbone.

Image Classification

Facial-Attribute-Detection-Q...

Comprehensive facial analysis by extracting face features.

Object Detection



Lightweight-Face-Detection...

face_det_lite_quantized is a face detection model.

Object Detection

Facial-Landmark-Detection-...

Facial landmark predictor with 3DMM.

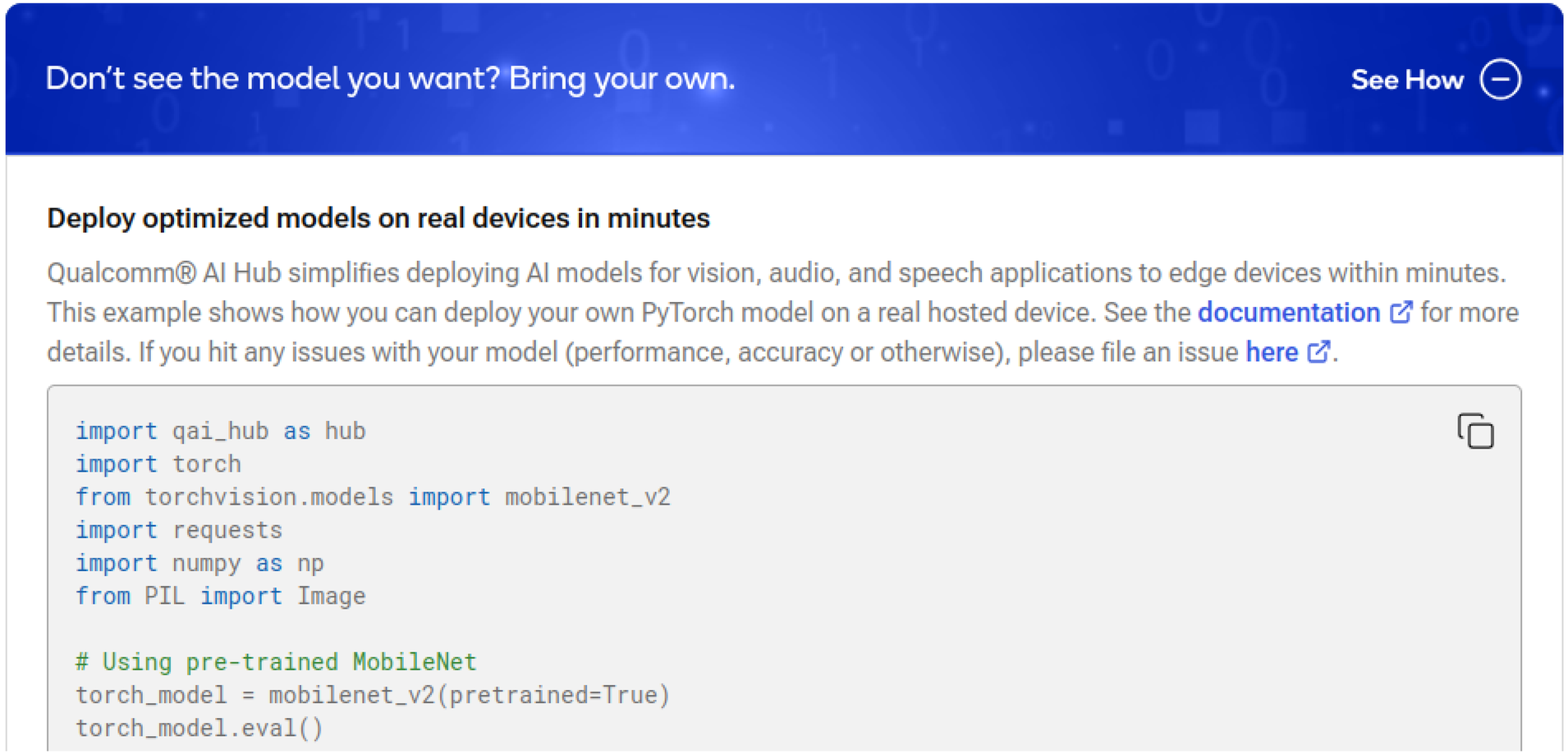
Pose Estimation

FCN-ResNet50-Quantized

Quantized fully-convolutional network model for image segmentation.

Semantic Segmentation

Qualcomm AI Hub – Bring your own model



Using Qualcomm AI Python package, you can easily deploy your own model.

EDGE IMPULSE

The Edge AI Platform

Gartner
COOL VENDOR 2022

What is Edge Impulse?

The industry leading Edge AI development platform helping your teams take **any data**, develop **any model** and deploy to **any target.**

Build real-world datasets at scale

Develop custom ML solutions fast

Deploy intelligent edge products

Powering the Largest Edge AI Ecosystem

160,000+
Developers

400,000+
Developer & Enterprise Projects

TRUSTED BY LEADING ENTERPRISES

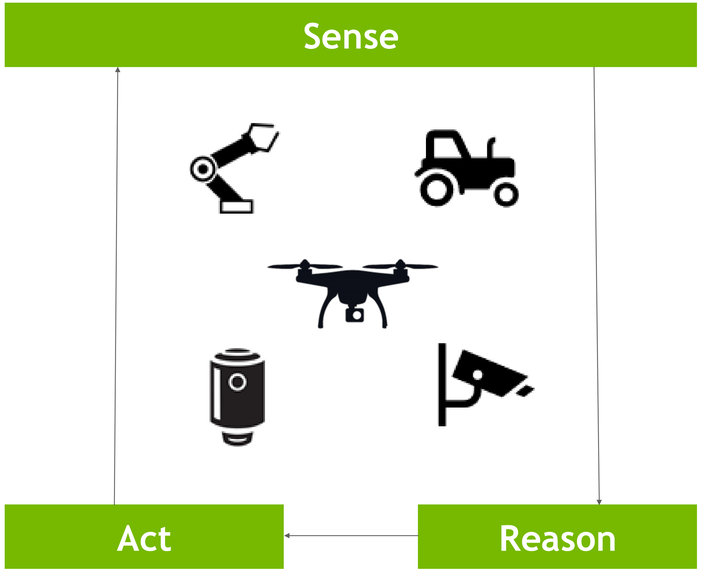
NVIDIA ECOSYSTEM AND SOFTWARE OVERVIEW FOR EMBEDDED AND ROBOTICS

NVIDIA JETSON

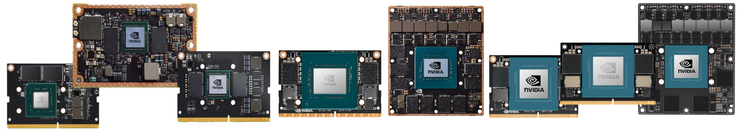
Software-Defined AI Platform

AI at the Edge

Sensor Fusion & Compute Performance



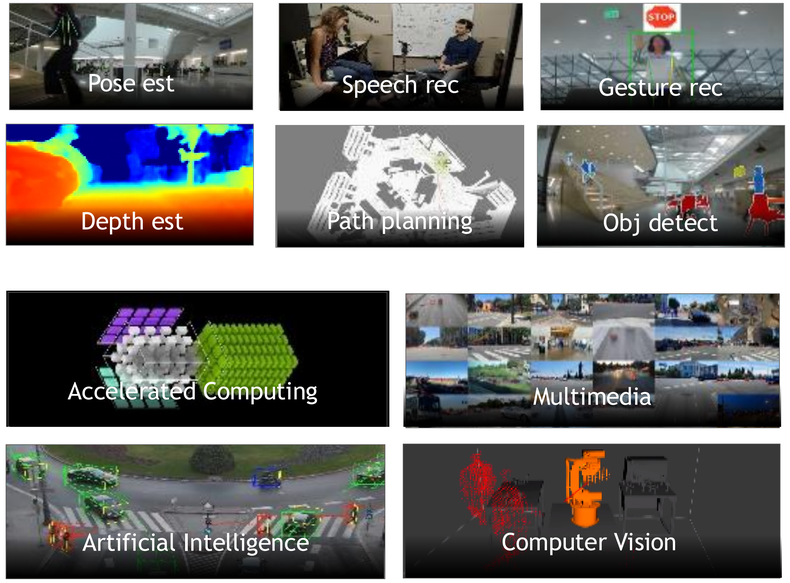
JETSON COMPUTER



Autonomous Machines: The Future of AI | NVIDIA

SOFTWARE DEFINED

SDK, Design Tools, Libs, GEMs



Jetpack SDK · CUDA · TensorRT · Triton · ONNX · ROS

Jetson Software | NVIDIA Developer

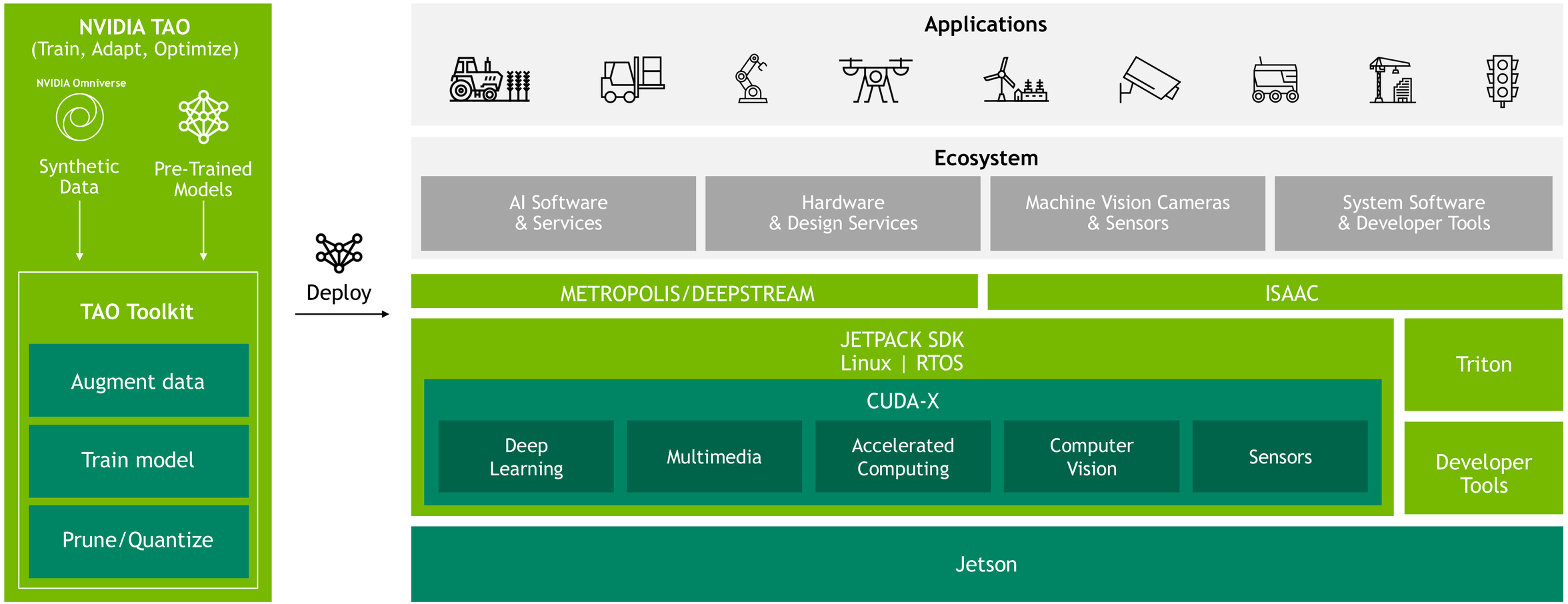
ECOSYSTEM

Expertise, Time to Market

Jetson Ecosystem | NVIDIA Developer

SAME POWERFUL NVIDIA SOFTWARE - FROM THE CLOUD TO THE EDGE

Jetson Software provides end-to-end acceleration for AI applications and accelerate your time to market



JETSON ORIN ACCELERATES AI APPLICATIONS FROM SPECIAL SILICON DESIGN

OFFLOAD the GPU with powerful accelerators

Deep Learning Accelerator

A fix function accelerator for deep learning inference workloads

DLA

Programmable Vision Accelerator

Dual Vector Processing Units(VPUs) that support various computer vision kernels

PVA

Video Image Compositor

Enables support for various image processing features

VIC

Optical Flow Accelerator

A hardware accelerator for computing optical flow and stereo disparity between frames

OFA

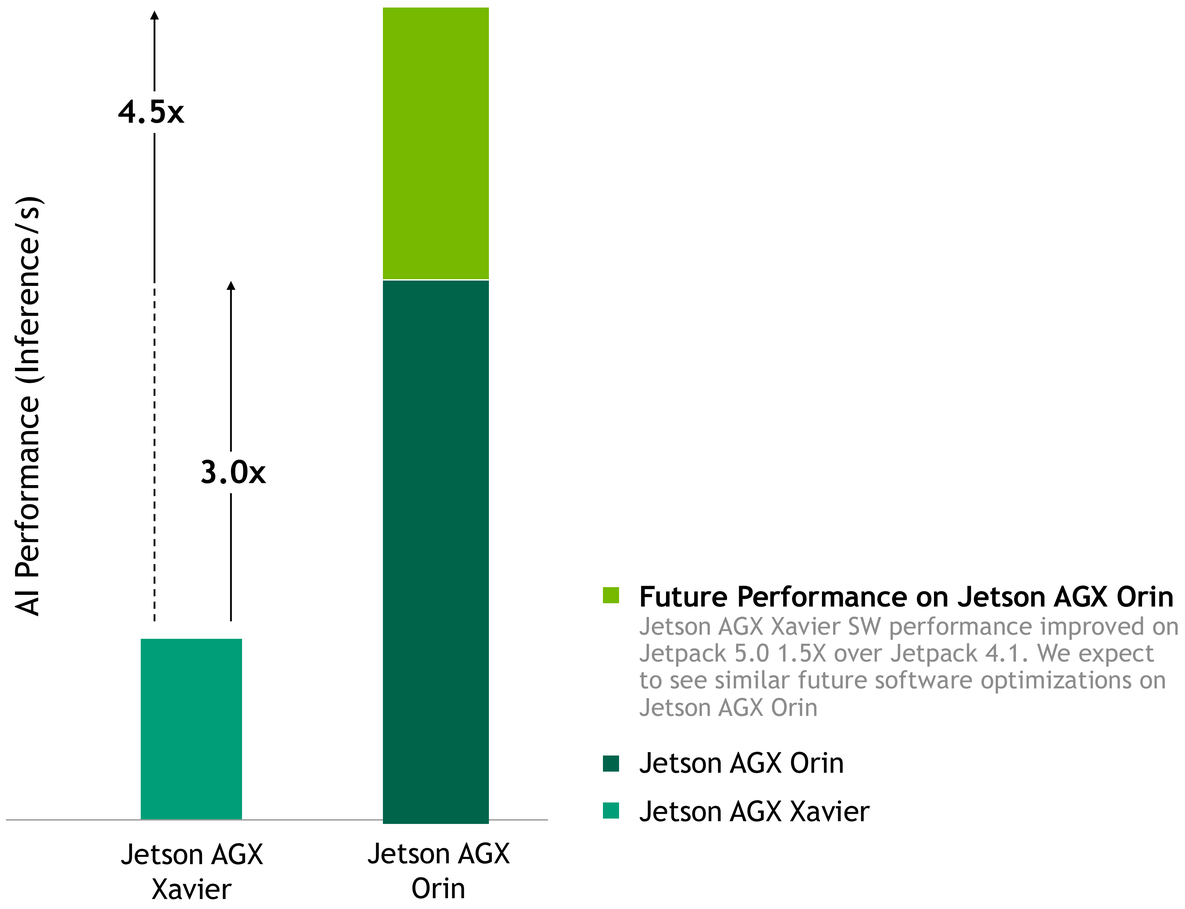
	AGX Orin Performance	GPU Performance	2x DLA Performance
PeopleNet (V2.3)	1294 fps	890 fps	404 fps
Dashcam Net	1895 fps	1308 fps	587 fps

	Jetson AGX Xavier	Jetson AGX Orin	Jetson AGX Orin
Engine	GPU	GPU	PVA + VIC + OFA
Stereo Disparity* (SGM)(540P)	80fps	166fps	102.7 fps

GIANT LEAP IN PERFORMANCE FOR NEXT GEN AI

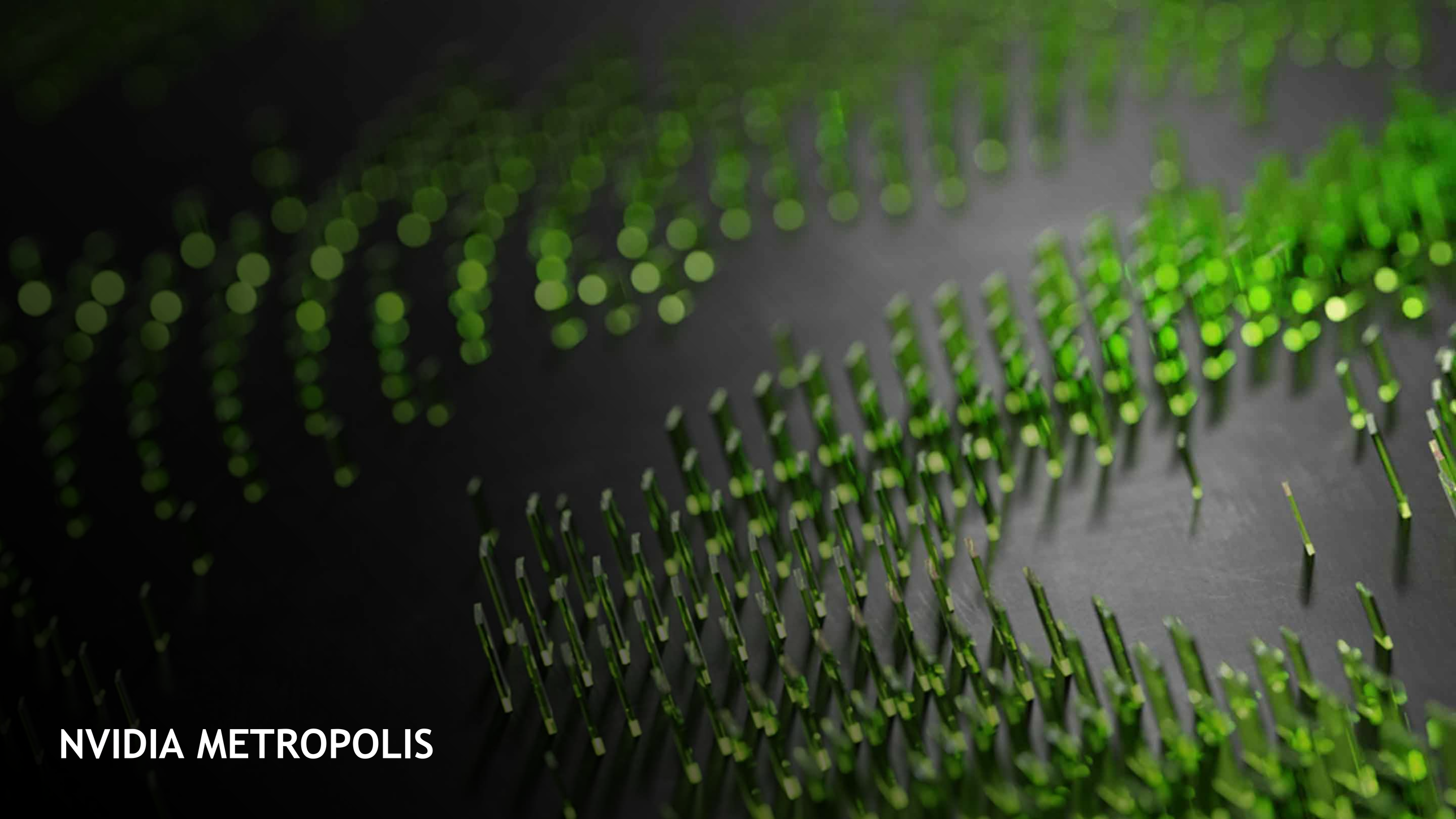
VISION AI AND CONVERSATIONAL AI PRETRAINED MODELS

	Jetson AGX Xavier	Jetson AGX Orin
PeopleNet (V2.3)	418	1294
Action Recognition 2D	471	1577
Action Recognition 3D	32	105
LPR Net	1190	4118
Dashcam Net	670	1895
BodyPose Net	172	559
ASR: Citrinet 1024	19	44
NLP: BERT-base	271	780
TTS: Fastpitch-HifiGAN	36	95



** These are Dense Models | * PeopleNet used here is v2.3 with pruned performance. Previous results used are v2.5 unpruned data.*

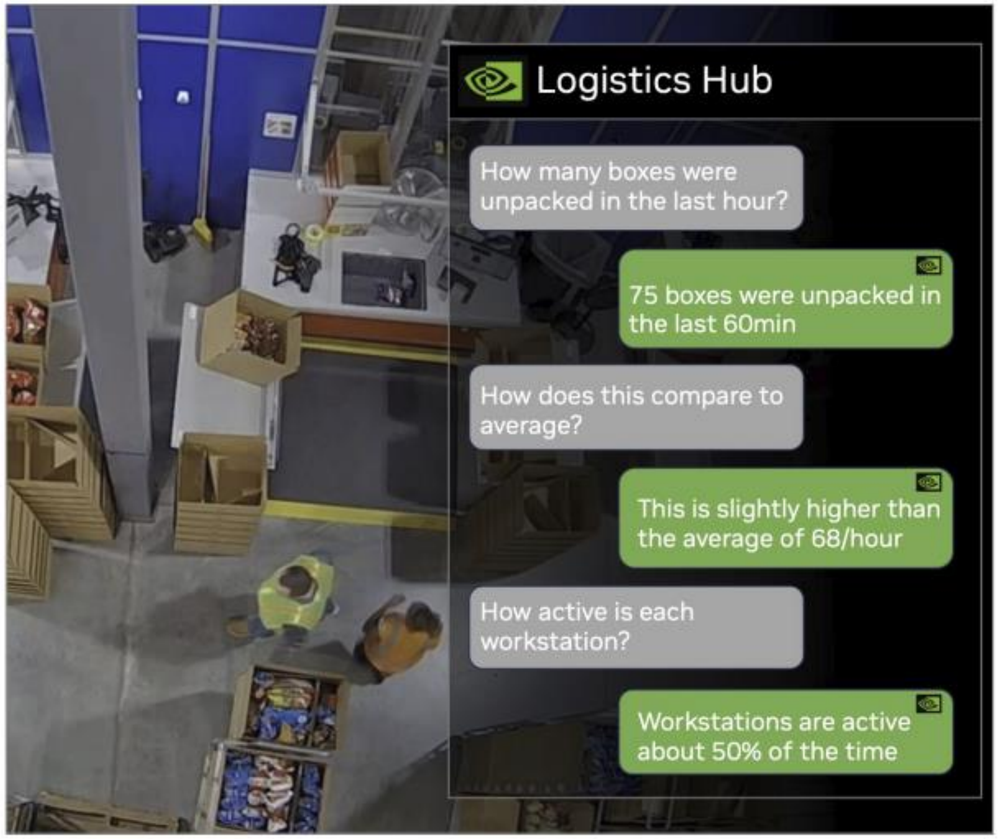
NVIDIA


NVIDIA METROPOLIS

GENERATIVE AI IS TRANSFORMATIVE

Legacy CNNs

Rigid/Rule-Based | Requires tons of Labeled Data
Slow Development Cycle



Generative AI

Zero-Shot Learning | Generalizable
Faster Development Cycle | Natural Language Prompts

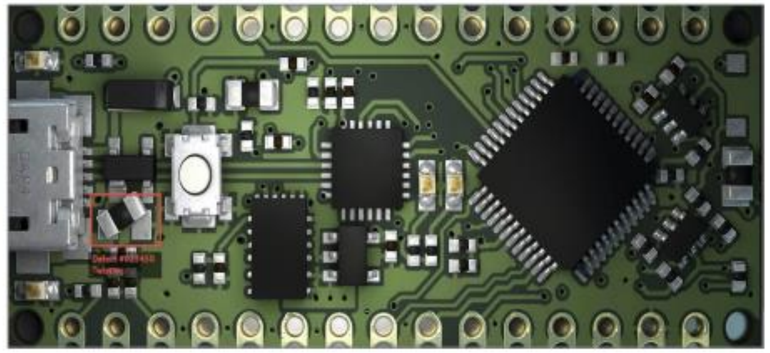
GENERATIVE AI ACROSS INDUSTRIES

Operational Efficiency in Warehouses

Real-Time Asset Tracking

Autonomous Planning & Navigation

Robot Programming

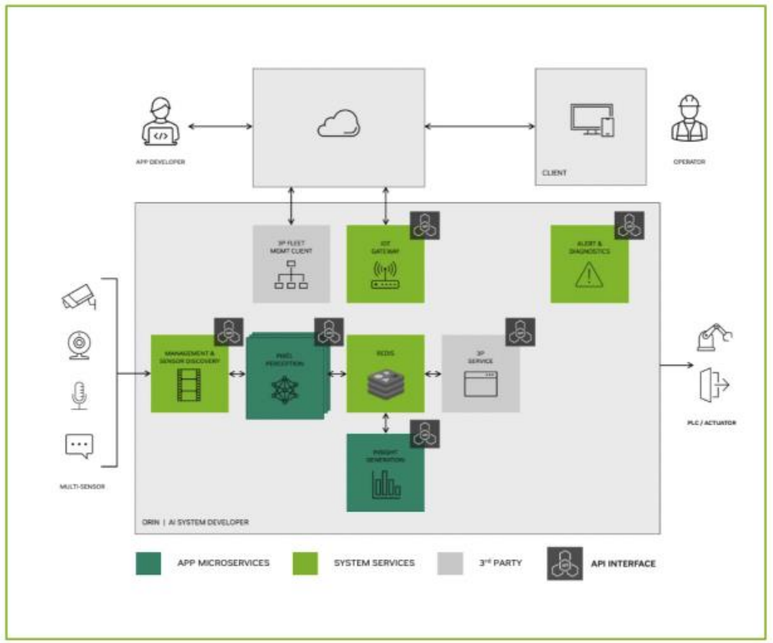


Defect Inspection in Factories

Human-Robot Interaction

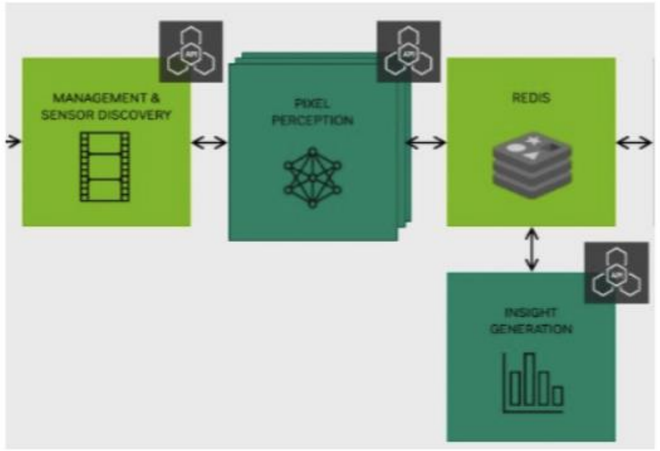
METROPOLIS APIS & MICROSERVICES FOR JETSON

Architecture



- Containerized
- Modular
- OTA
- Extensible

App Framework

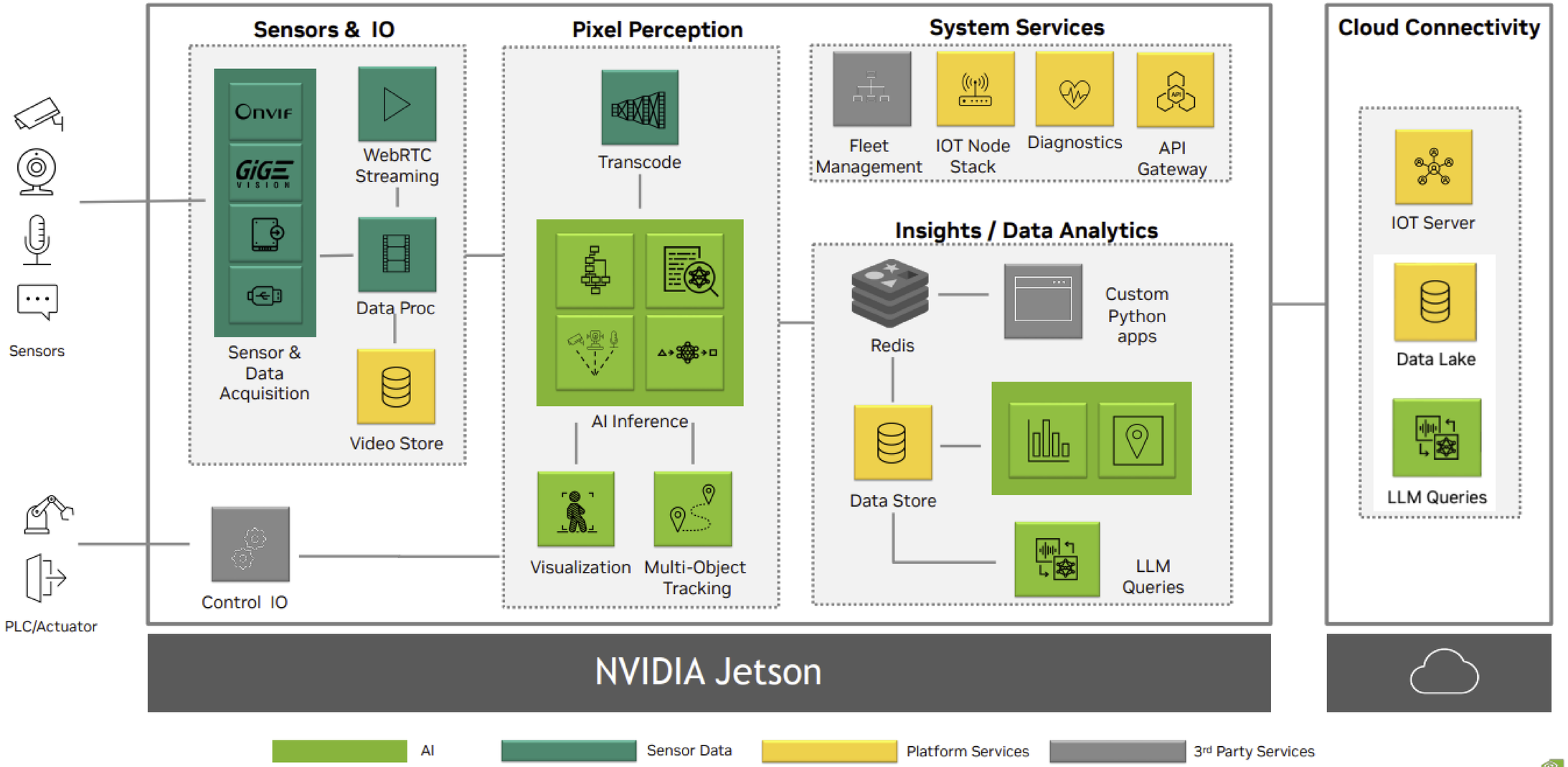


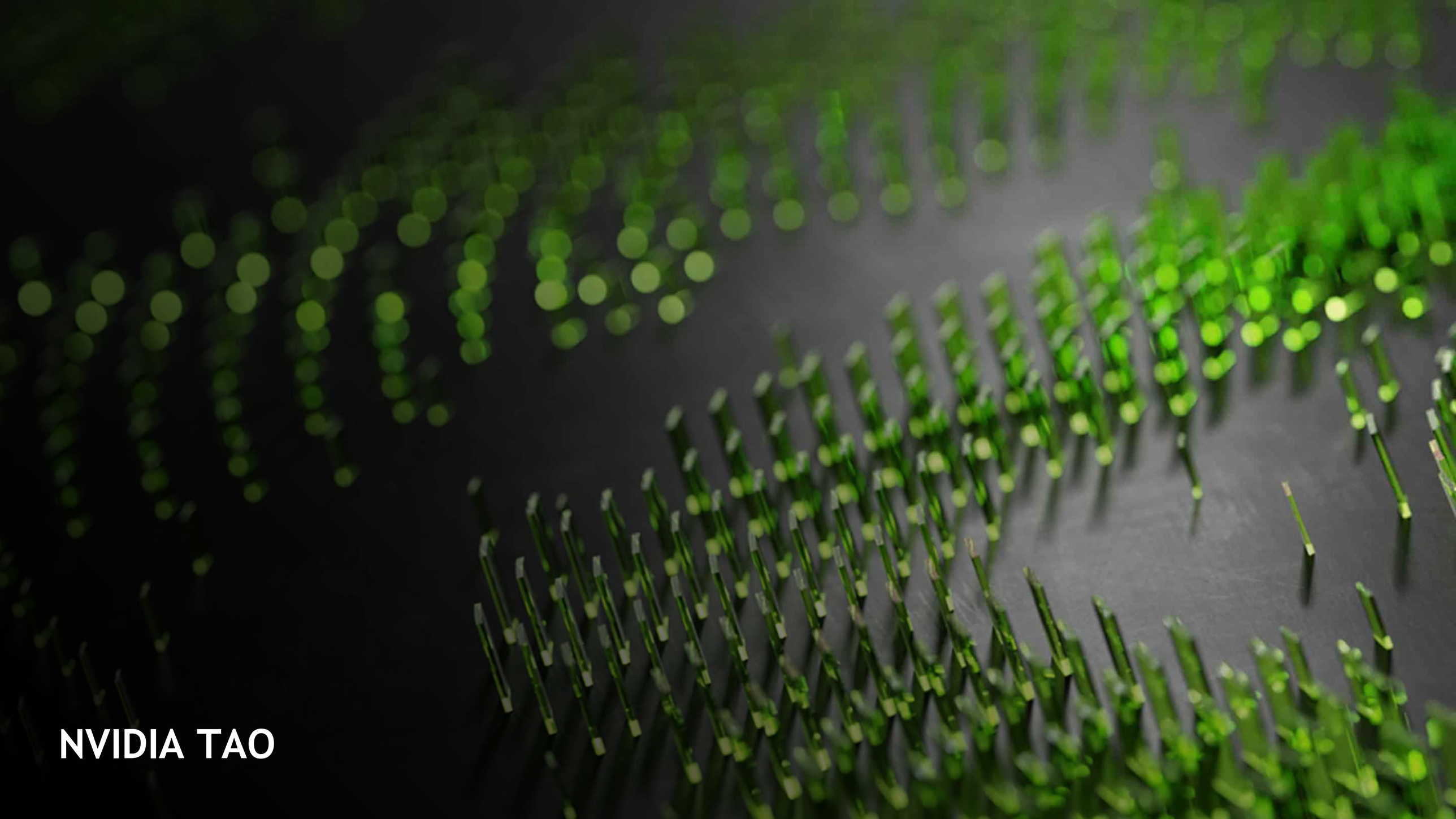
- Video Storage Toolkit
- Perception Pipeline
- Analytics
- Overlay Visualization

Gen AI

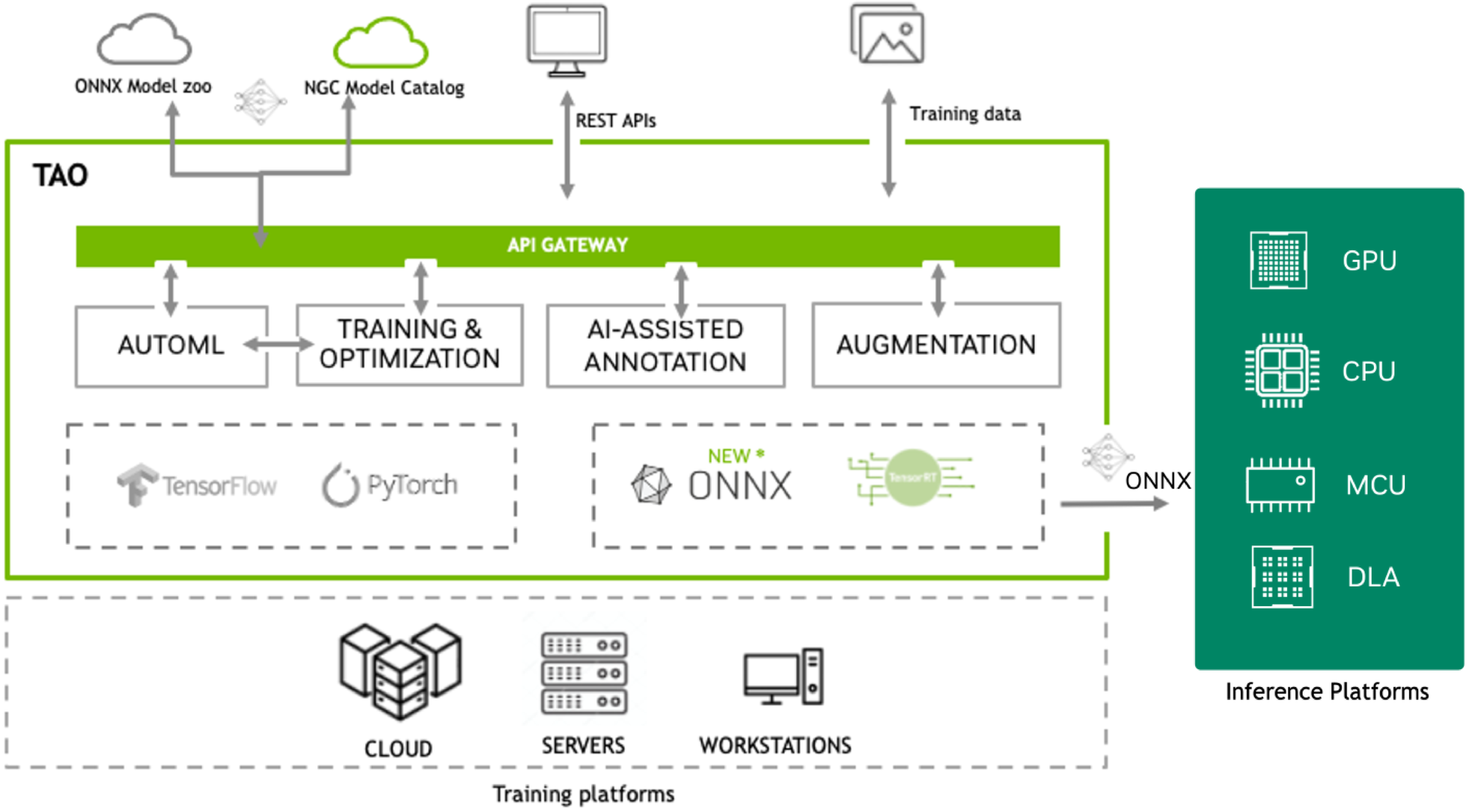
- LLM, LMM Queries
- Natural Language Prompts
- Highly Interactive
- Zero-shot Learning

METROPOLIS APIS & MICROSERVICES FOR JETSON




NVIDIA TAO

NVIDIA TAO TOOLKIT



NVIDIA TAO - WORKFLOW

We will use NVIDIA TAO to optimize and convert the Tiny-YOLOv4

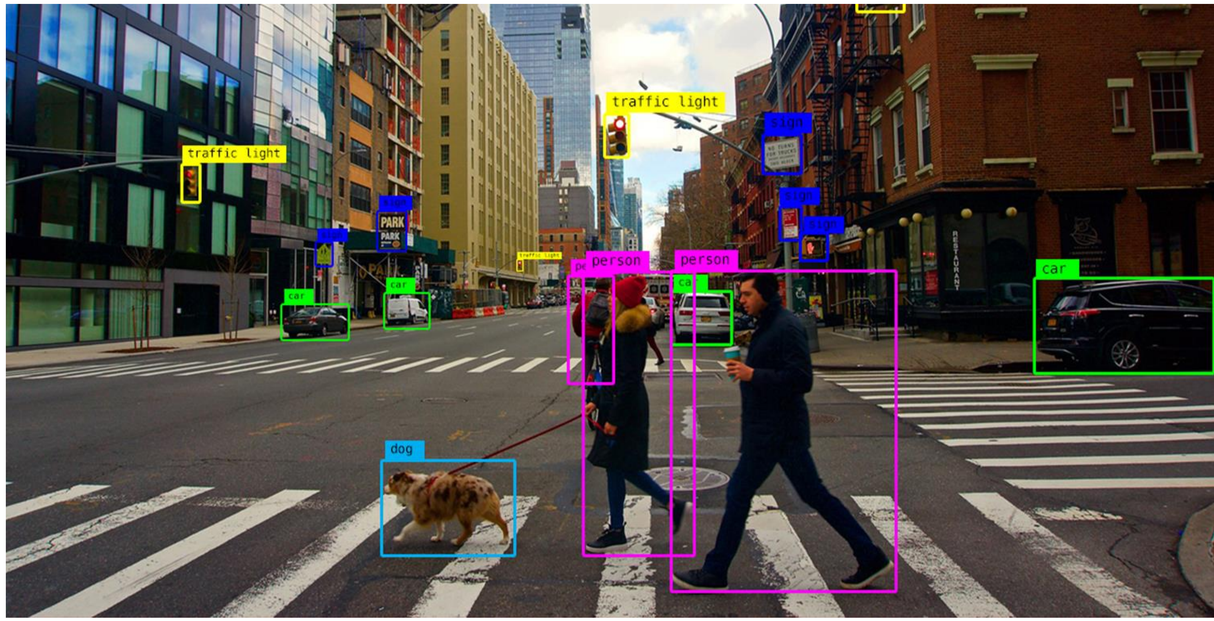
- Setting up the environment and NVIDIA TAO
- Downloading model, dataset
- Training
- Quantization
- Pruning
- Evaluation of the model

References:

We follow https://github.com/NVIDIA-AI-IOT/nvidia-tao

- You can do the exercise (or play with NVIDIA TAO) yourself using free Google Colab instance

YOLOv4 official repo: https://github.com/AlexeyAB/darknet

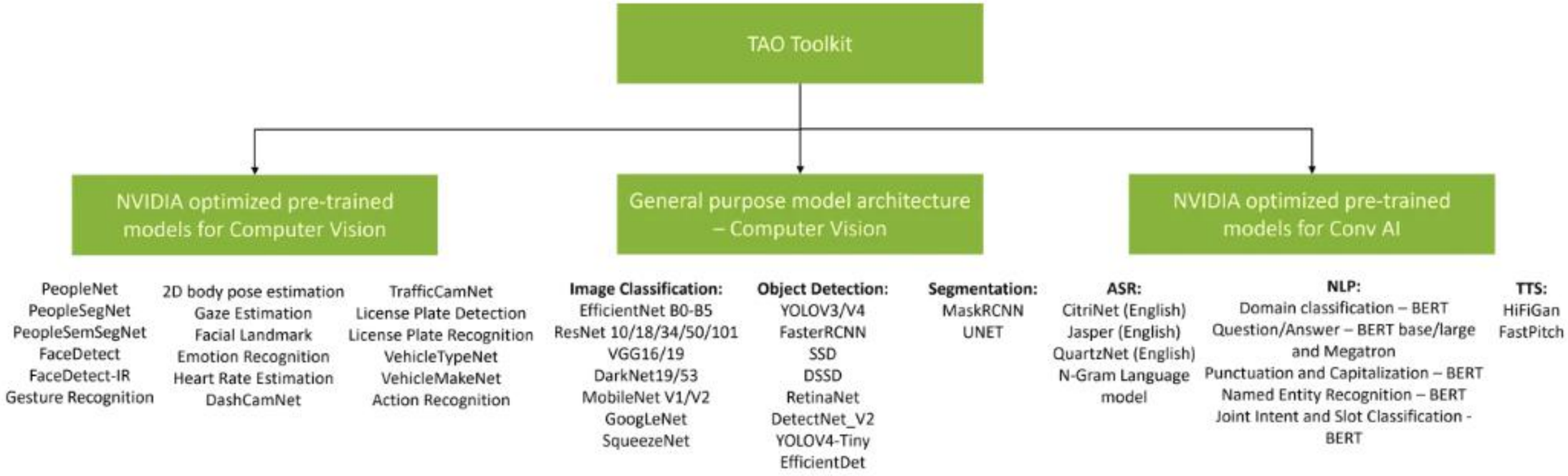


NVIDIA MODEL ZOO

NVIDIA provides extensive number of pretrained models, that can be easily used with Tao and used in commercial applications.

General-purpose vision models – starting point to build more complex models with transfer learning.

Purpose-built models – highly accurate models for specific application, domain and can be used directly or fine-tuned.

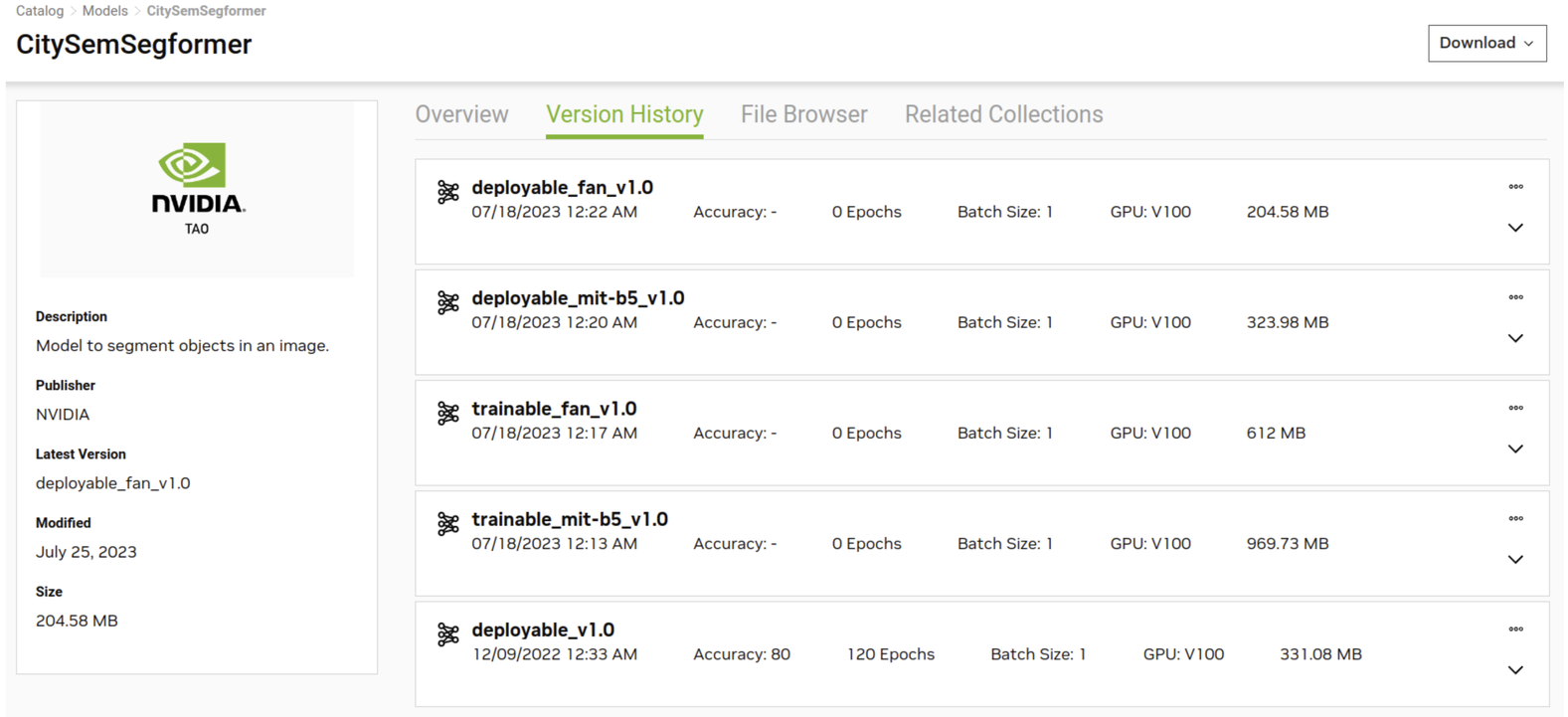


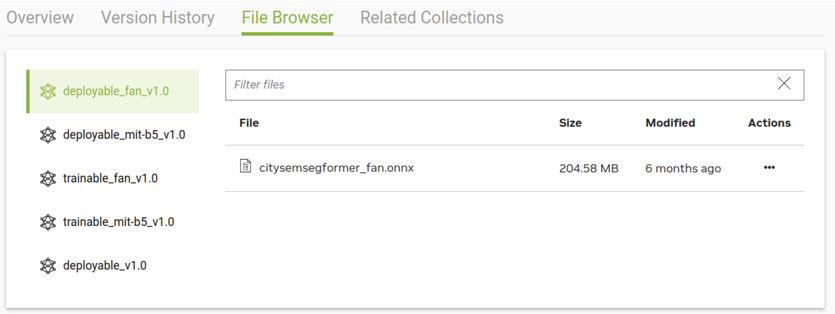
NVIDIA MODEL ZOO

Models are available on NVIDIA catalog (NGC), along with containers, software and resources

Provides model overview, information on dataset, evaluation, performance, usage (optimization, training, deployment).

SegFormer (transformer-based): https://catalog.ngc.nvidia.com/orgs/nvidia/teams/tao/models/citysemsegformer

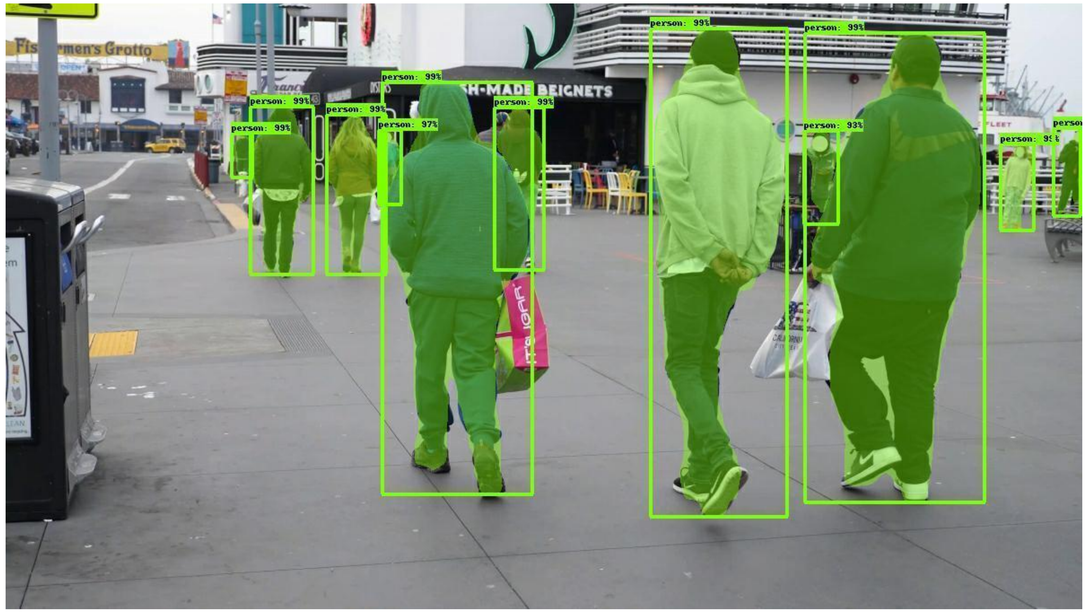




TAO-Converter Commands

FP16

```
./tao-converter -k tlt_encode -p input,1x3x1024x1820,1x3x1024x1820,1x3x1024x1820 -t fp16 -e ./bs1_fp16.engine ./citySemSegFormer.etlt
```



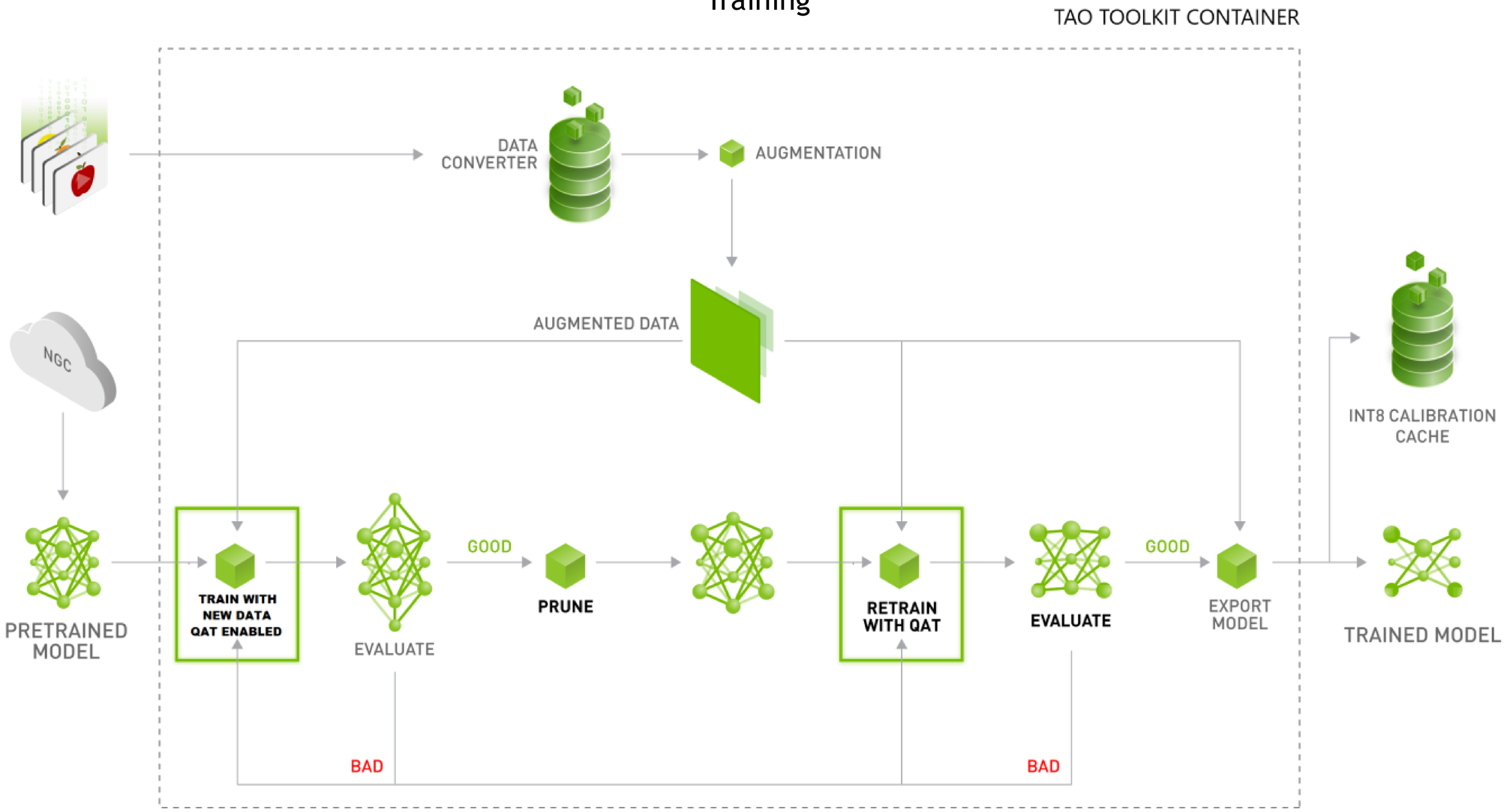
PeopleSegNet

PeopleSemSegNet

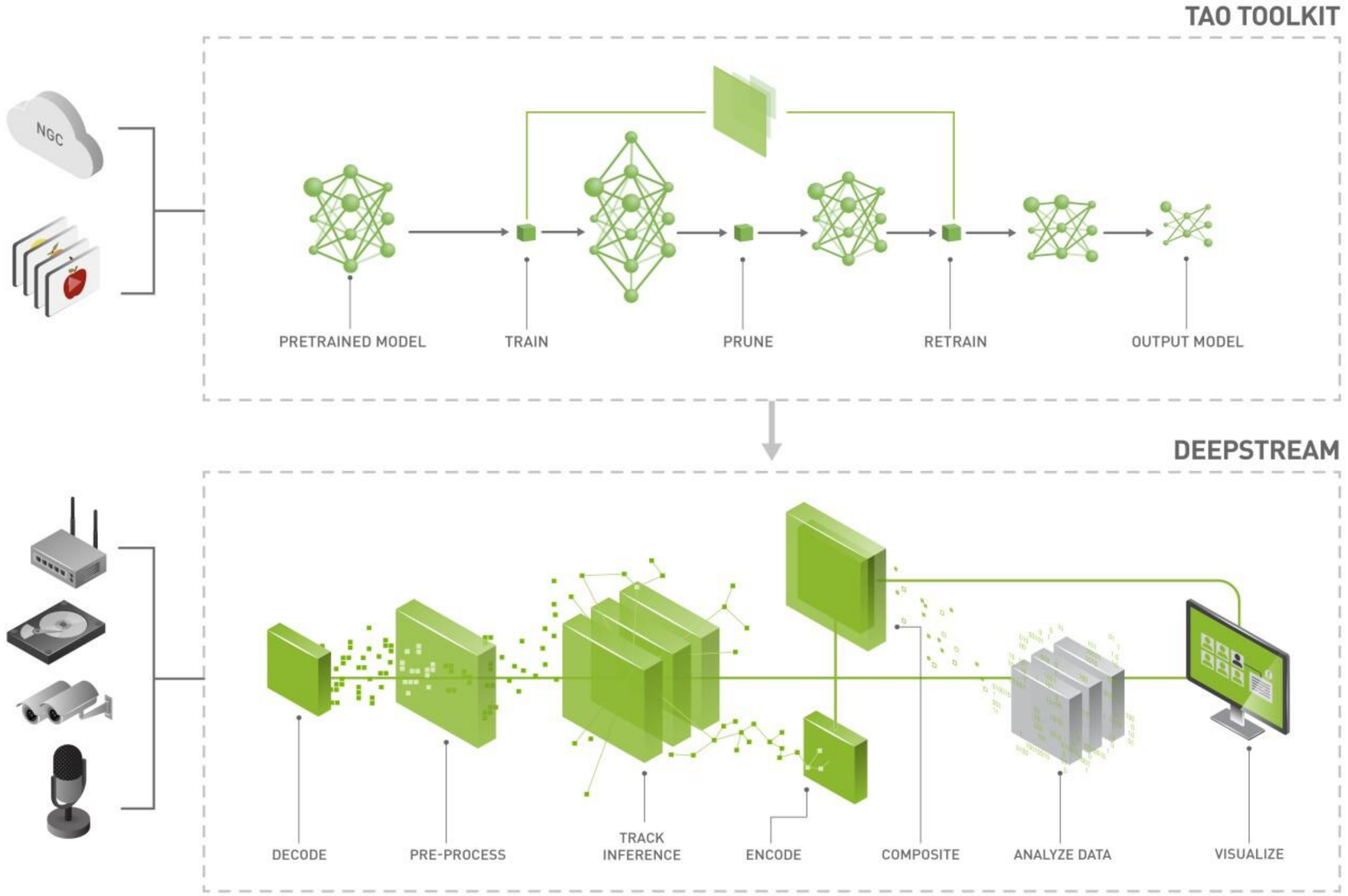
Visual ChangeNet-Segmentation

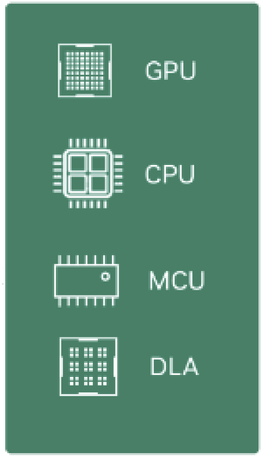
USING TAO TOOLKIT TO TRAIN A MODEL

Training



EXPORTING THE MODEL





Inference Platforms

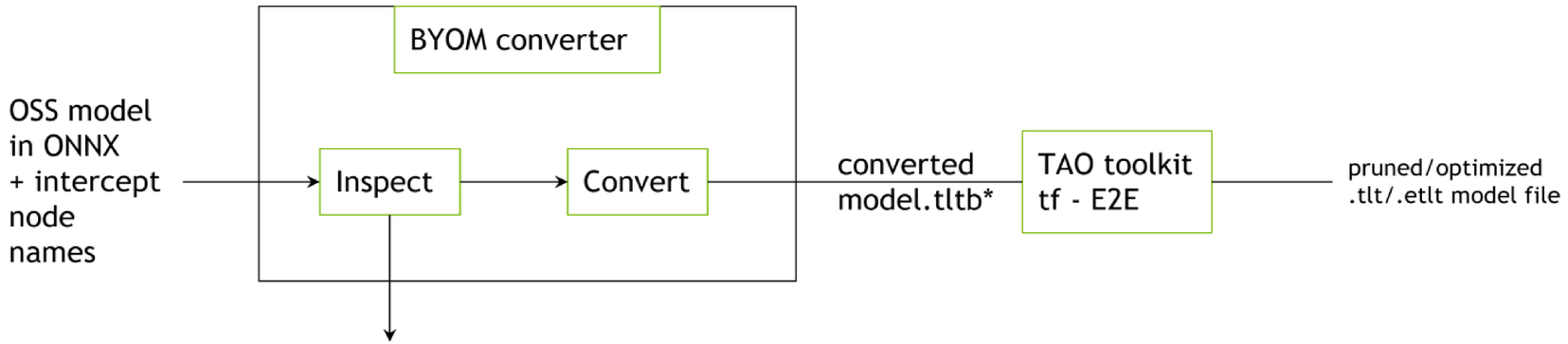
BRING YOUR OWN MODEL

Convert ONNX model for TAO

Bring Your Own Model (BYOM) is a Python-based package that converts any open-source ONNX model to a TAO-comaptible model. The TAO BYOM Converter provides a CLI to import an ONNX model and convert it to Keras. The converted model is stored in .tltb format, which is based on EFF.

Limitations: only classification, UNet supported, only channel_first models are supported

https://docs.nvidia.com/tao/tao-toolkit/text/byom/byom_converter.html#supported-onnx-nodes-in-tao-byom

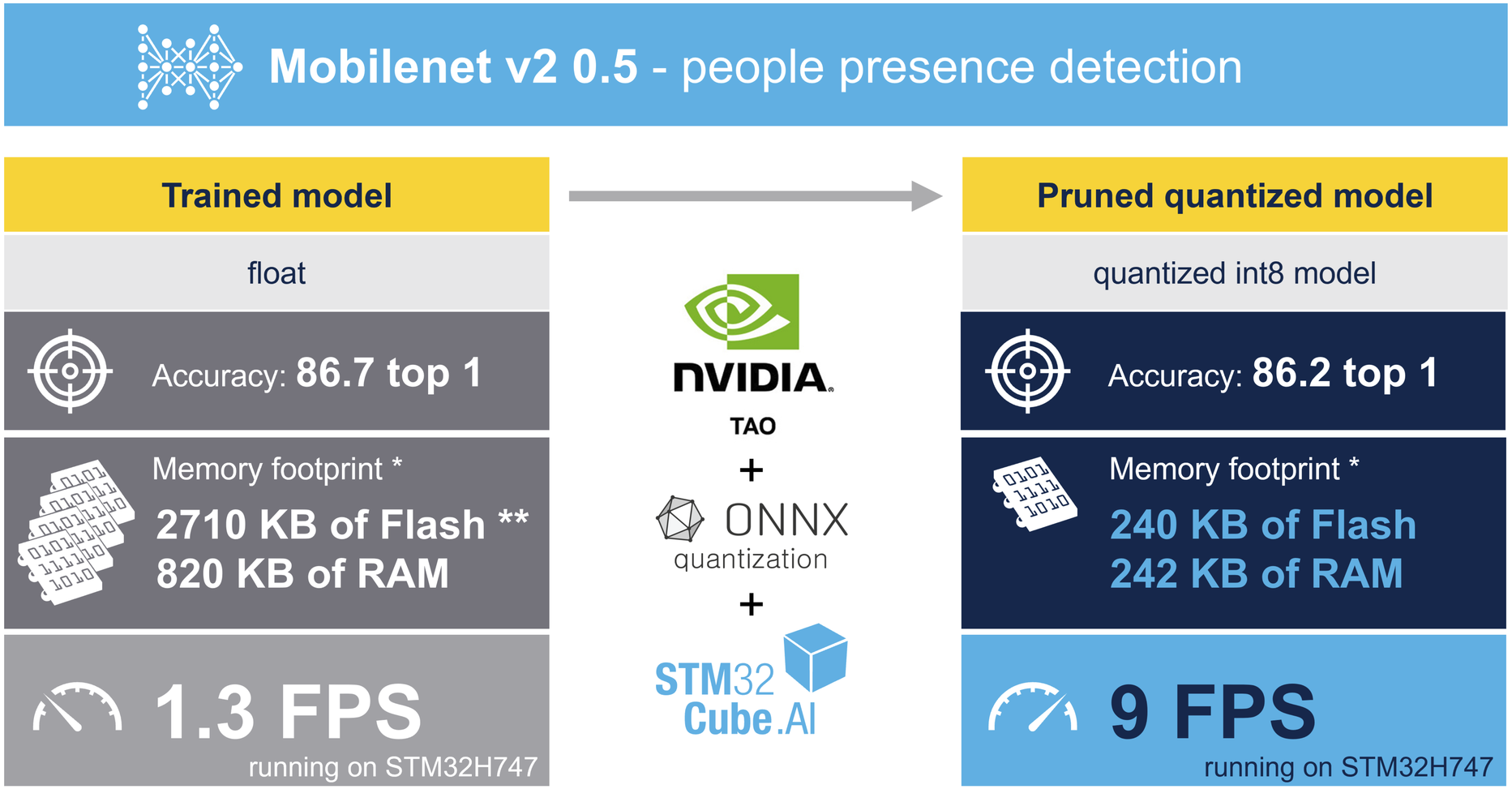


What's included in model.tltb:

1. *custom layer implementation*
2. *model architecture*
3. *model weights*

- TAO Toolkit targeted models
  - Image Classification
  - UNet Semantic Segmentation

NVIDIA TAO ON MCU

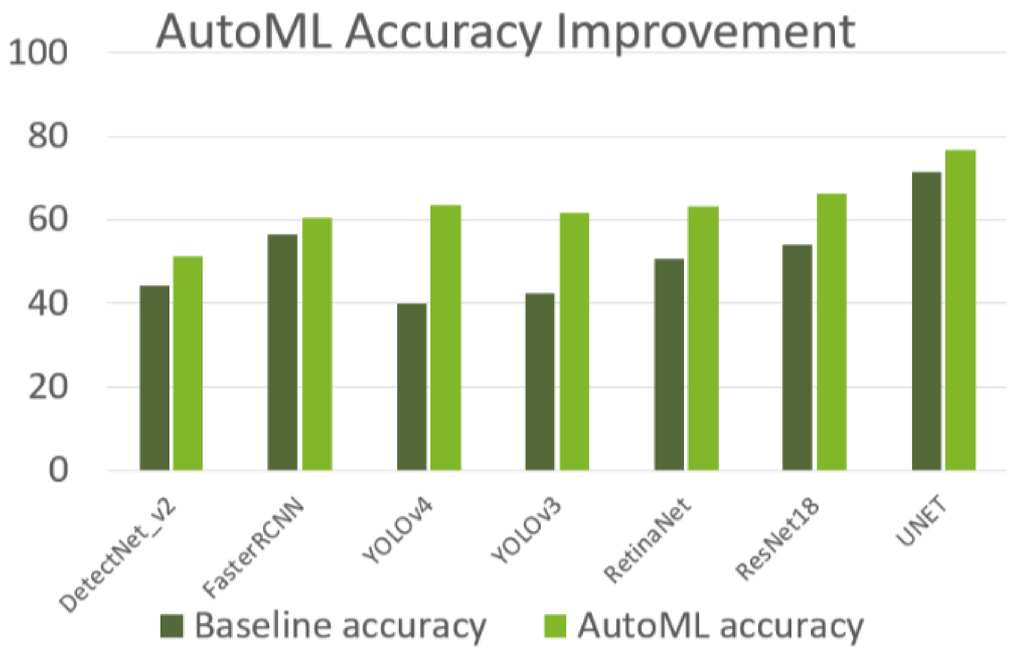


** after optimization with STM32Cube.AI*
*** implies the use of external memories*

up to 12 FPS on STM32H735

AUTOML IN TAO

Object Detection	Classification	Segmentation	Other
DINO NEW	FAN NEW	SegFormer NEW	License Plate
D-DETR NEW	GC-ViT NEW	Mask RCNN	recognition
DetectNet V2	ResNet	UNet	
EfficientDet	EfficientNet		
FasterRCNN	MobileNet		
YoloV3	DarkNet		
YoloV4	Multi-task		
YoloV4_Tiny	classification		
SSD			
RetinaNet			



EASY

No AI expertise required to train and optimize hyperparameters

ACCELERATE

Accelerate AI development by automating the tedious, manual process of tuning AI models

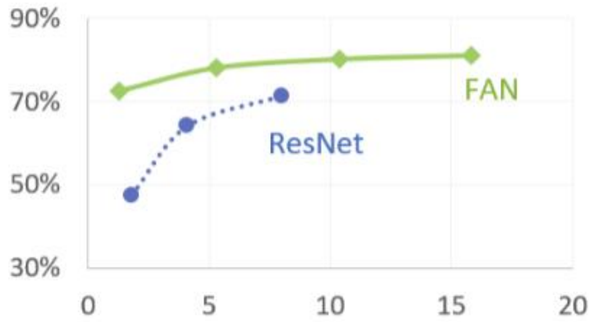
CONFIGURABLE

Sweep from a large selection of hyperparameters

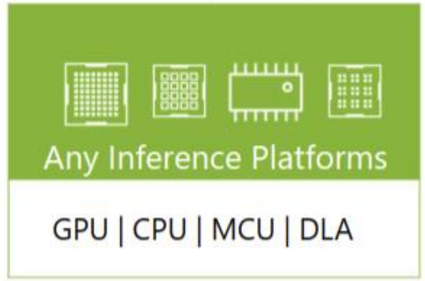
TAO SUMMARY

AI-Assisted Annotation for Mask Segmentation

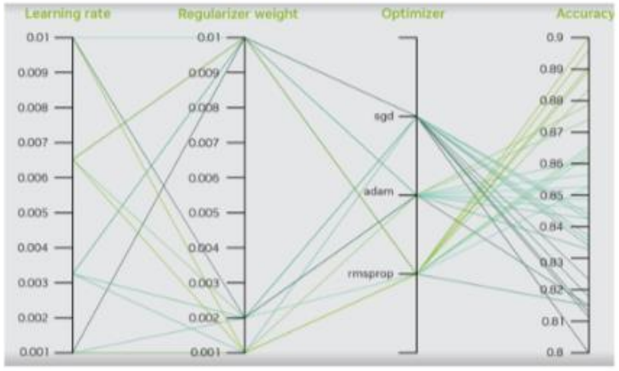
SOTA Vision Transformer Models



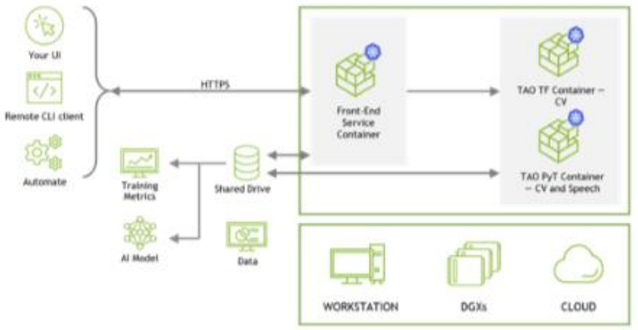
Open Source and deploy on any Platforms



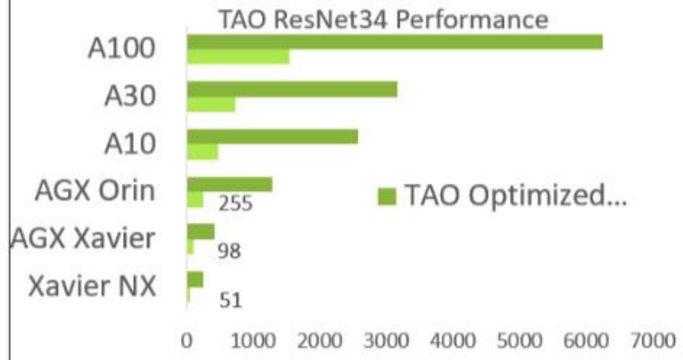
Converge Faster with AutoML



Integrate into Services With Rest APIs



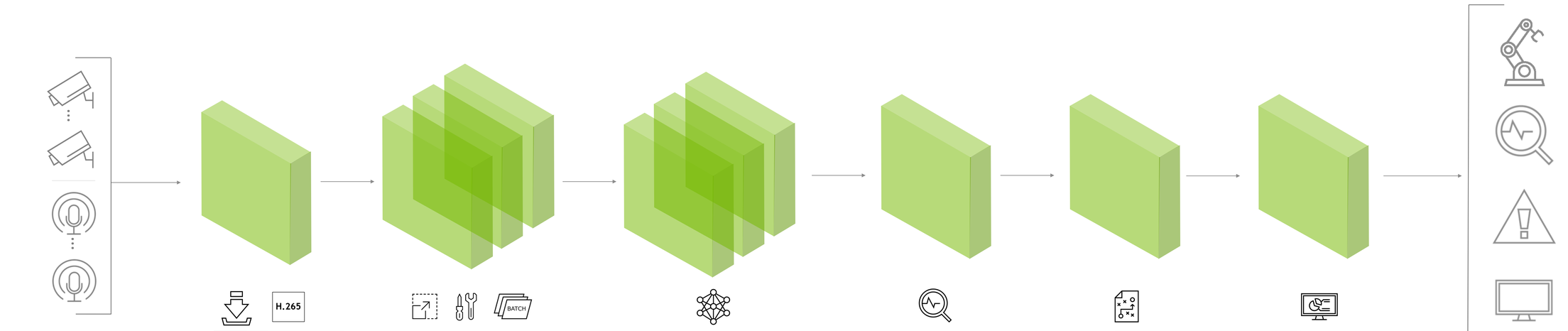
4x Increase In Inference with Pruning and Quantization




NVIDIA DEEPSTREAM

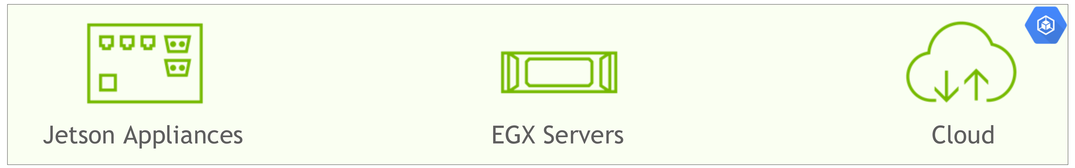
DEEPSTREAM SDK

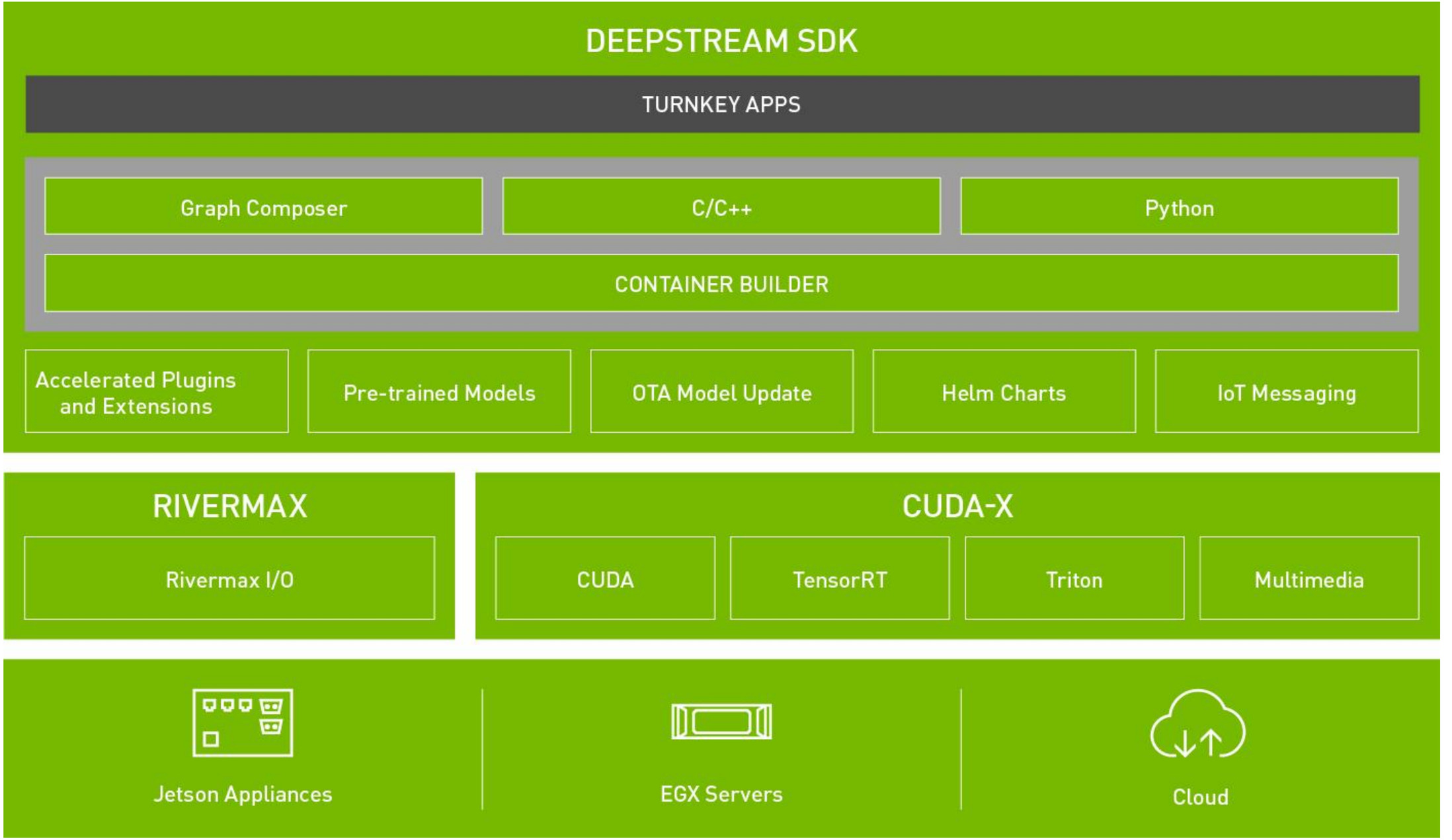
Accelerated and Optimized Applications from Edge to Cloud



Hardware Acceleration	Capture and Decode	Pre-processing and Batching	AI Inference	Tracking	Business Rules and Analytics	Composition
	ISP	GPU	GPU	GPU	GPU	GPU
	GigE		DLA	CPU	CPU	VIC
	NVDEC			VIC		

Edge to Cloud Deployments




DEEPSTREAM SDK
TURNKEY APPS
Graph Composer
C/C++
Python
CONTAINER BUILDER
Accelerated Plugins and Extensions
Pre-trained Models
OTA Model Update
Helm Charts
IoT Messaging
RIVERMAX
Rivermax I/O
CUDA-X
CUDA
TensorRT
Triton
Multimedia
Jetson Appliances
EGX Servers
Cloud

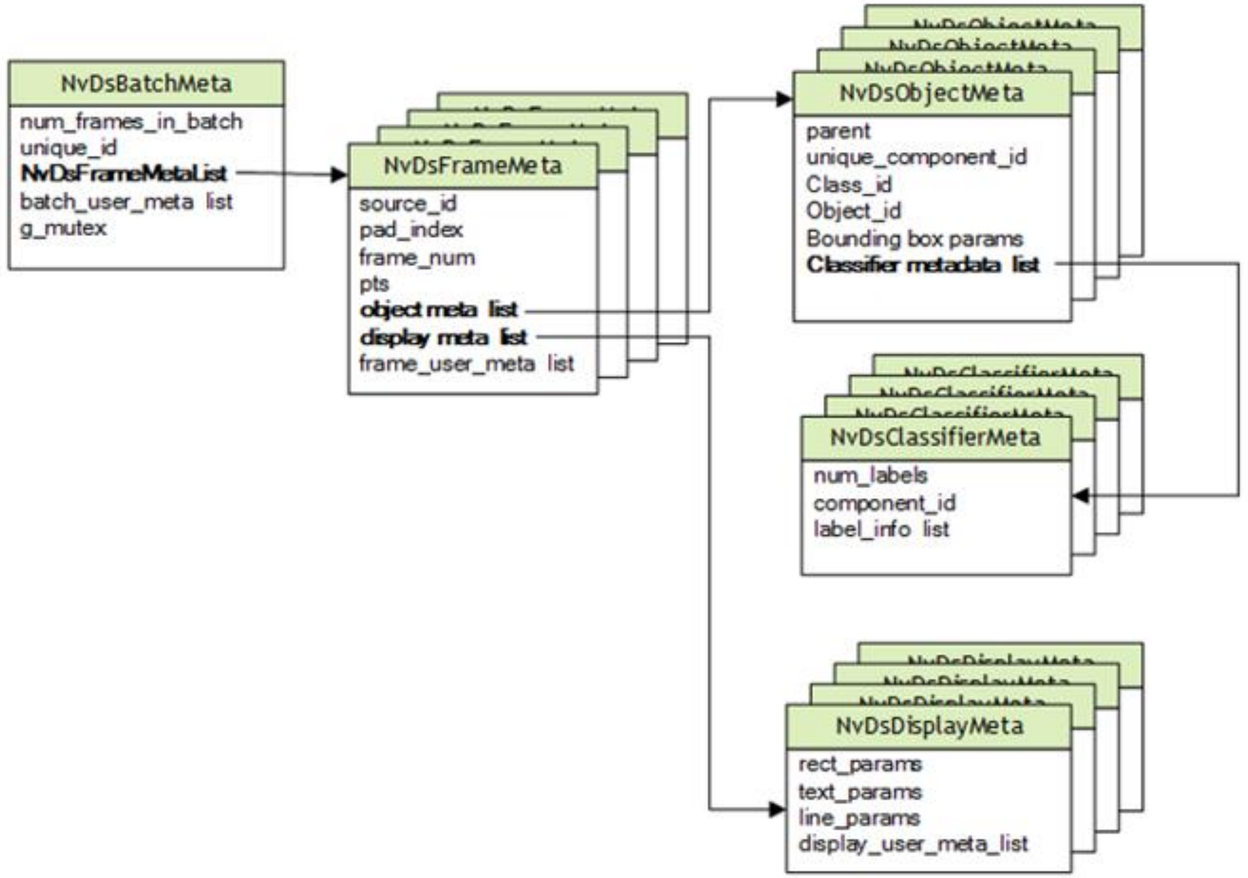
METADATA IN DEEPSTREAM SDK

Gst Buffer is the basic unit of data transfer in GStreamer. Each Gst Buffer has associated metadata. The DeepStream SDK attaches the DeepStream metadata object, NvDsBatchMeta.

User can add custom data, to be propagated through plugins and probes to metadata at the batch, frame, or object level within NvDsBatchMeta. To do that, you must acquire an instance of NvDsUserMeta from the user meta pool by calling nvds_acquire_user_meta_from_pool().

Note: all memory is allocated by C, python code only uses references to memory and has no ownership.

Data includes list of objects, bboxes and tracking bboxes with confidence levels, unique ids of objects, classifier results and pointer to currently processed frame or batch.



Thank You

Five Years Out