USE DEEPSTREAM AND THE TLT TO DEPLOY STREAMING ANALYTICS AT SCALE

AGENDA

1. Real-time Streaming for Video Analytics
2. Framework for Analyzing Video
3. Understand the Basics: DeepStream SDK 3.0
4. Build with DeepStream: Example Applications
5. NVIDIA Transfer Learning Toolkit

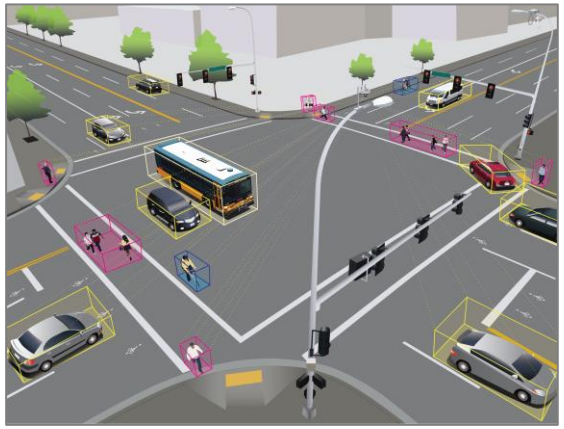

REALTIME STREAMING
VIDEO ANALYTICS

REALTIME STREAMING VIDEO ANALYTICS

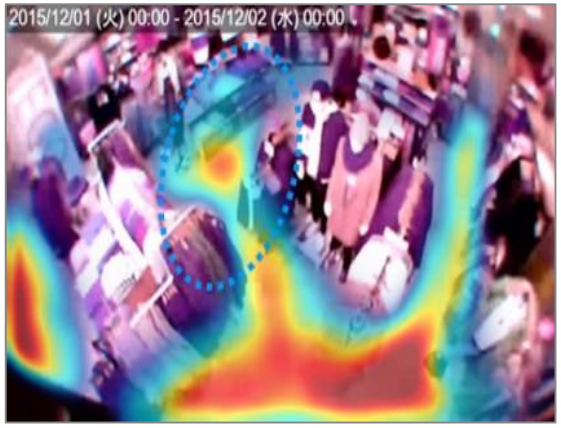
Access Control

Managing operations

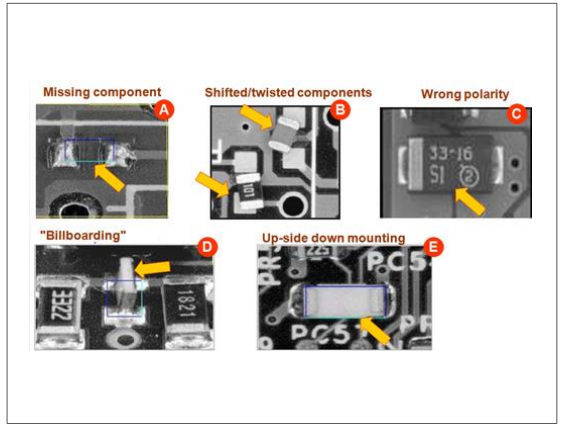
Parking Management



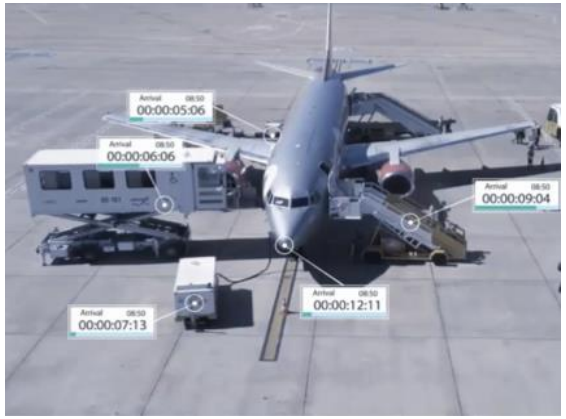
Traffic Engineering



Retail Analytics



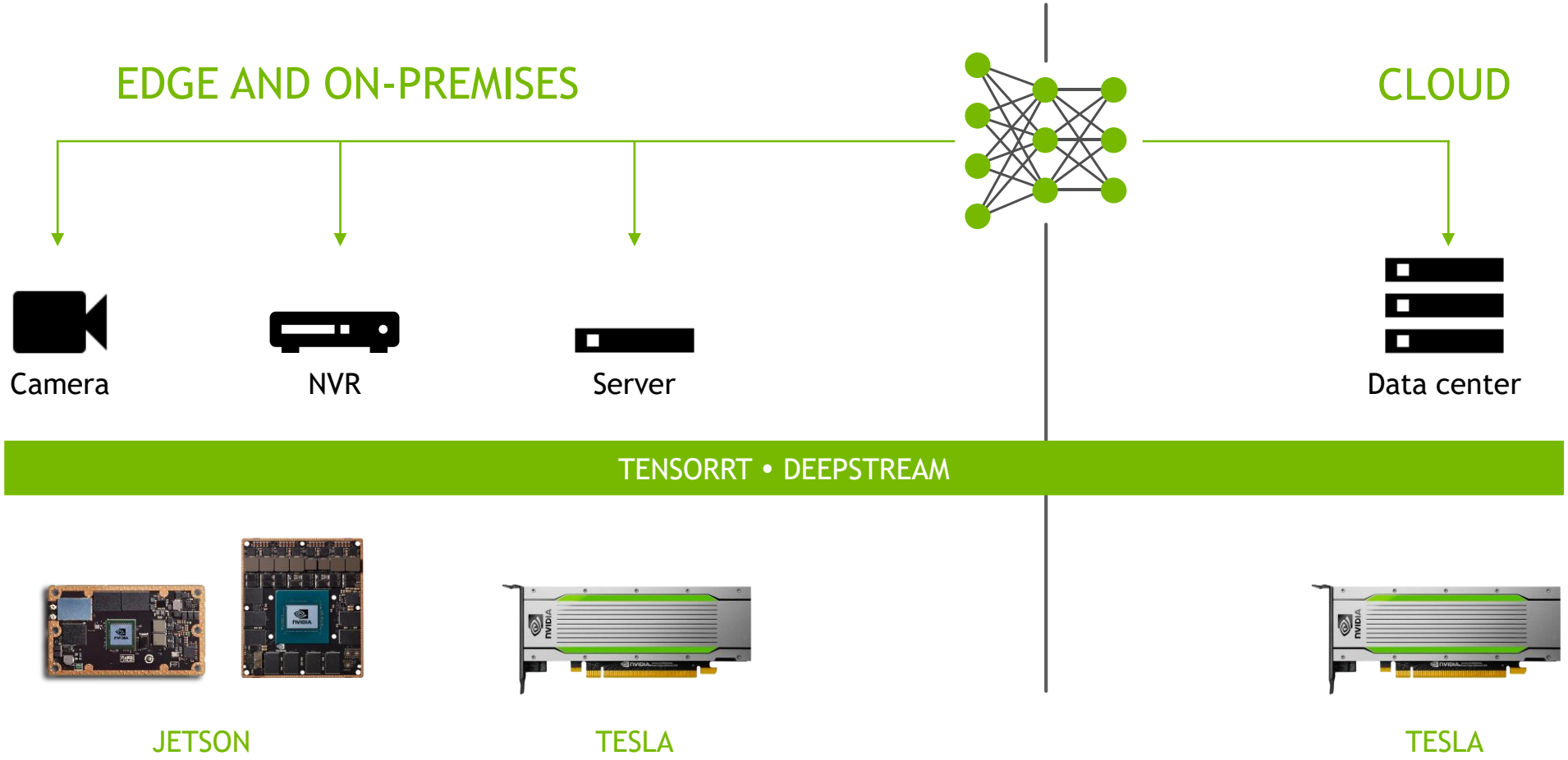
Optical Inspection



Managing Logistics

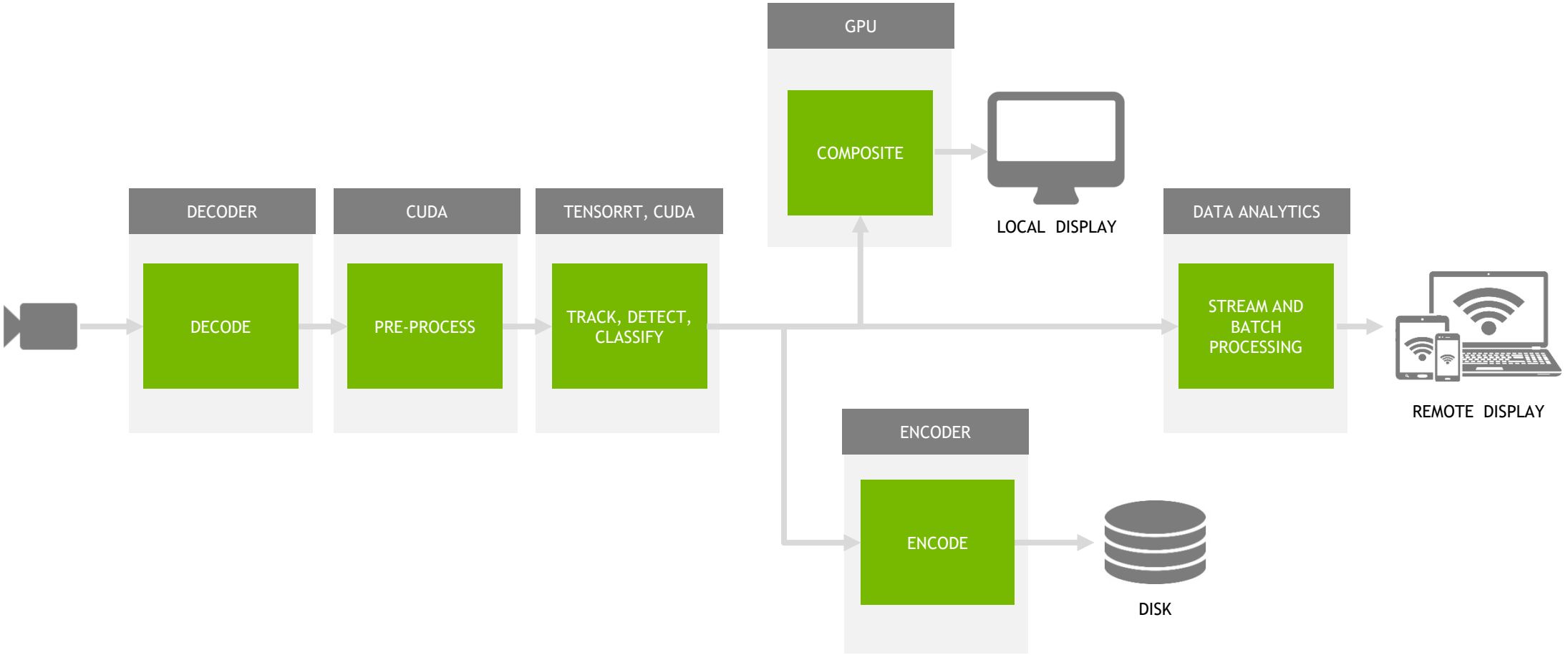
Content Filtering

DEPLOY FROM EDGE TO CLOUD




FRAMEWORK FOR
ANALYZING VIDEO

REALTIME STREAMING VIDEO ANALYTICS




UNDERSTAND THE BASICS:
DEEPSTREAM SDK 3.0

DEEPSTREAM SDK 3.0

Plugins (build with open source, 3rd party, NV)	Analytics - multi-camera, multi-sensor framework	Development Tools
DNN inference/TensorRT plugins	DeepStream in containers, Multi-GPU orchestration	End to end reference applications
Communications plugins	Tracking & analytics across large scale/ multi-camera	App building/configuration tools
Video/image capture and processing plugins	Streaming and Batch Analytics	End-end orchestration recipes & adaptation guides
3rd party library plugins … …	Event fabric	Plugin templates, custom IP integration

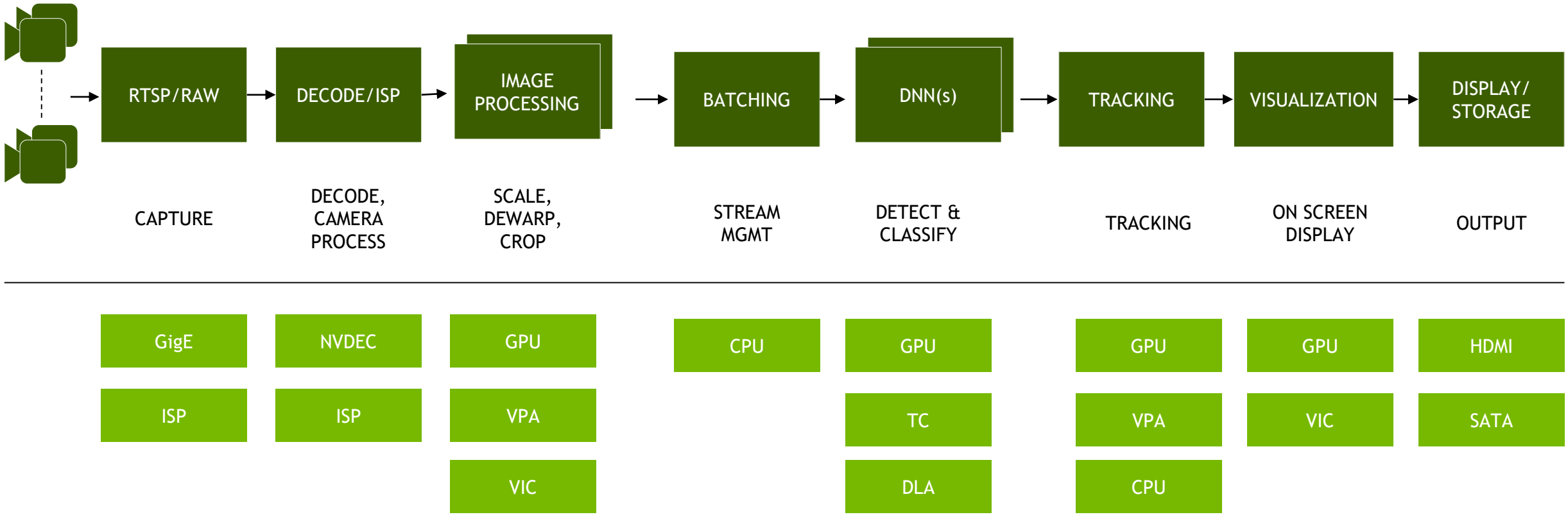
DeepStream SDK

TensorRT | Multimedia APIs/ Video Codec SDK | Imaging & Dewarping library | Metadata & messaging | NV containers | Message bus clients | Multi-camera tracking lib

Linux, CUDA

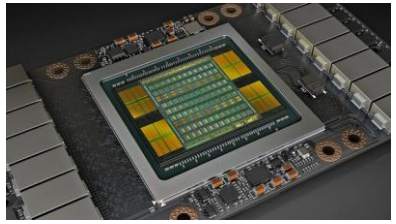
Perception infra - Jetson, Tesla server (Edge and cloud) | Analytics infra - Edge server, NGC, AWS, Azure

DEEPSTREAM STREAMING ARCHITECTURE



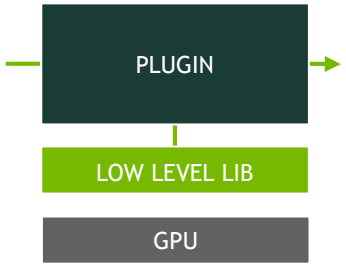
WHAT'S NEW IN DEEPSTREAM 3.0

LATEST PLATFORMS - TESLA T4, Jetson XAVIER



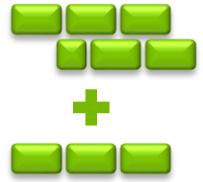
TensorRT 5, CUDA 10

NEW PLUGINS



Increased capability and throughput

DYNAMIC STREAM MANAGEMENT



Add, remove, modify streams on the fly

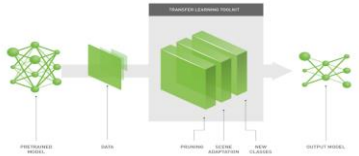
CONNECT EDGE TO CLOUD

Stream and Batch Analytics on Metadata

EASY TO SCALE AND MANAGE

Deploy in Docker Containers

HIGH EFFICIENCY AND THROUGHPUT WITH Transfer Learning Toolkit

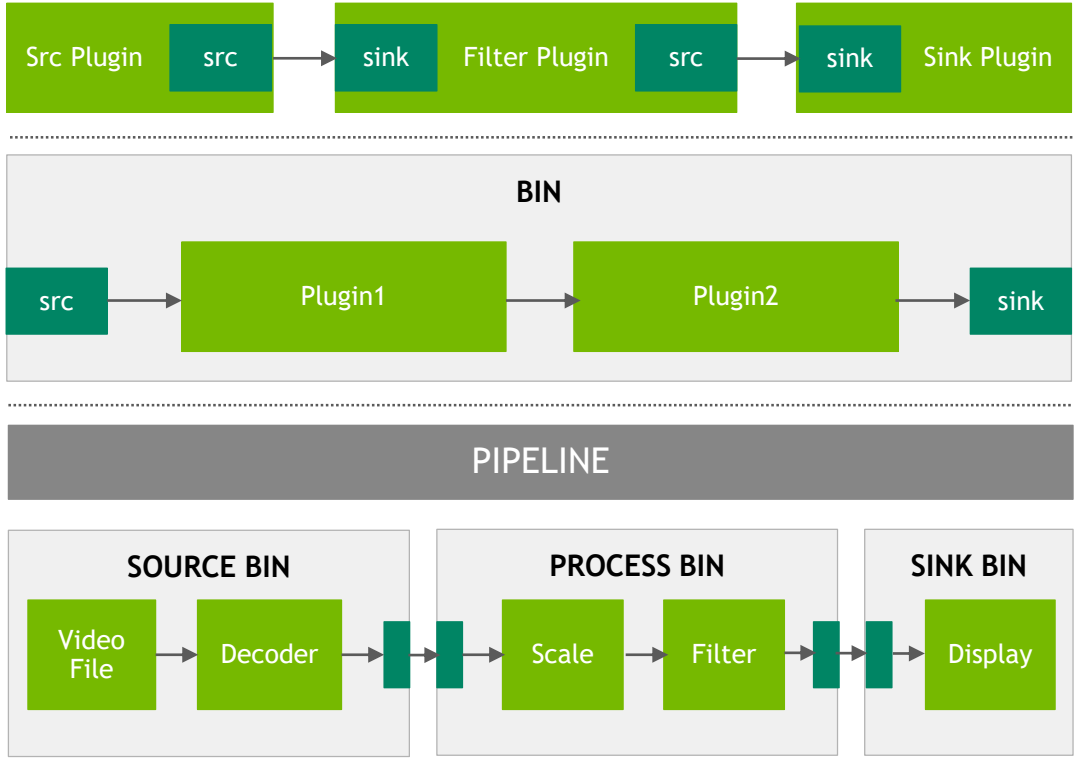


Transfer Learning Toolkit model files are plug-n-play

Dynamic stream management and containers available on Tesla only. Jetson support coming soon

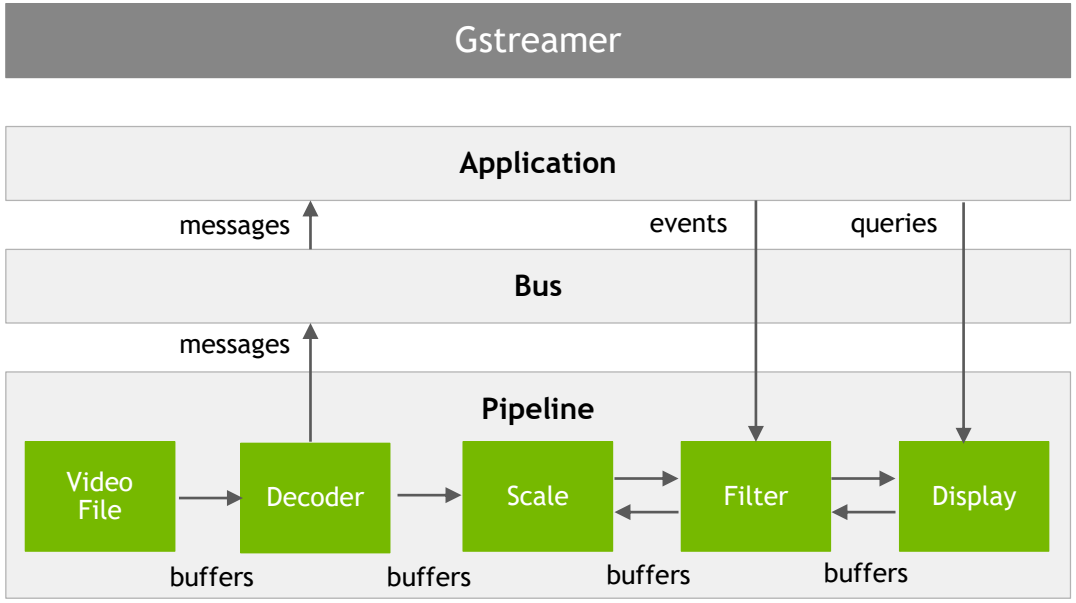
GSTREAMER

Level	Component	Function
1	PLUGINS	Basic building block connected through PADs
2	BINS	A container for a collection of plugins
3	PIPELINE	Top level bin providing a bus and managing the synchronization



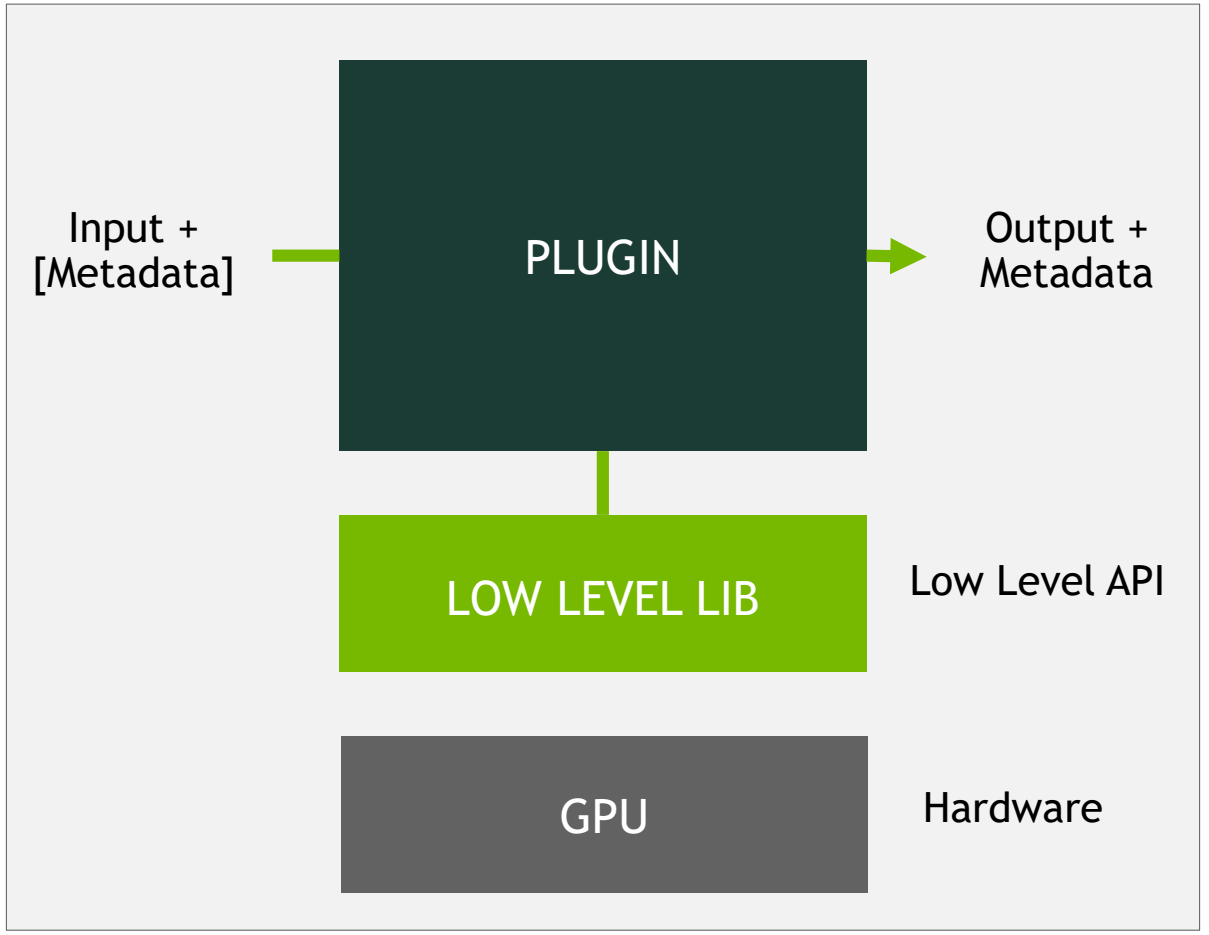
GSTREAMER

Object	Function
Buffers	Pass streaming data between plugins in the pipeline
Events	Send info between plugins or from the application to plugins
Messages	Post info on message bus for collection by the application
Queries	Allow applications to request information from the pipeline



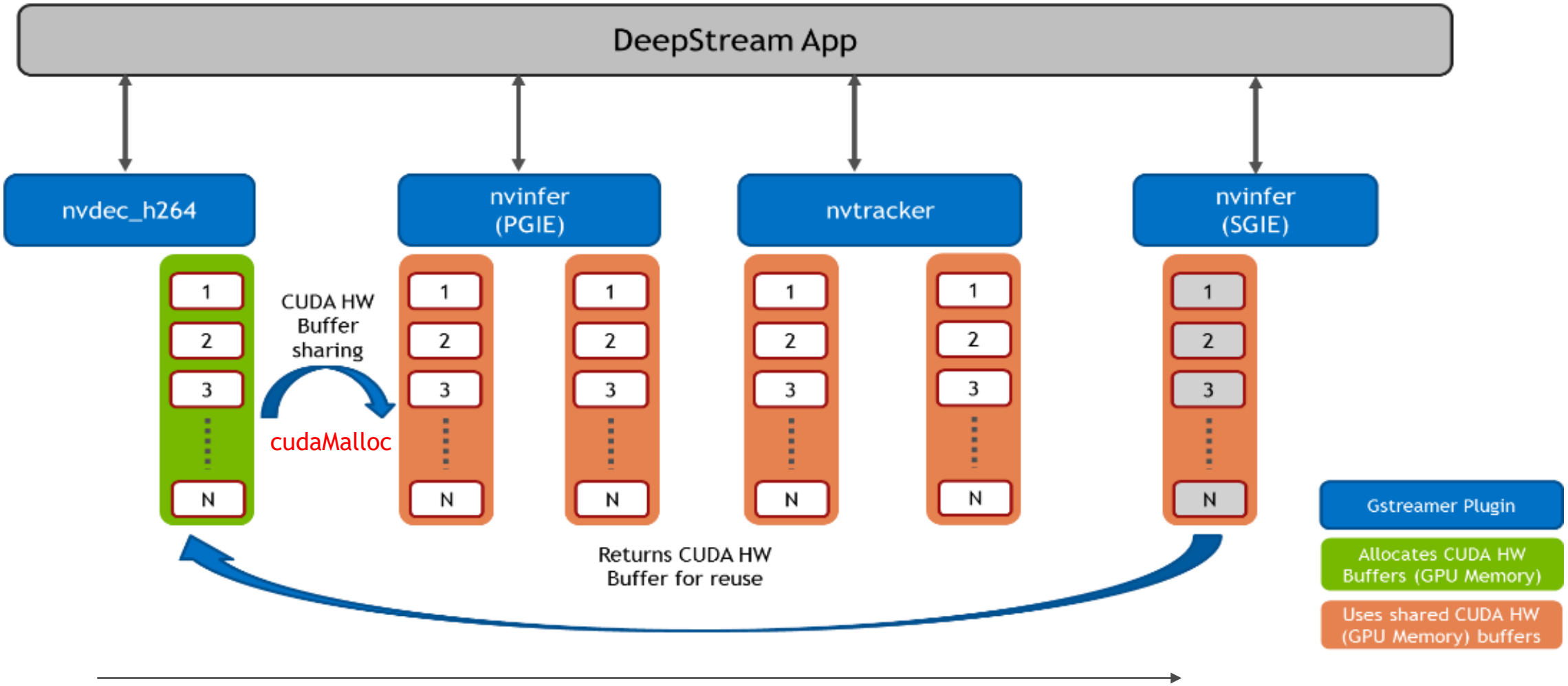
DEEPSTREAM BUILDING BLOCK

- A plugin model based pipeline architecture
- Graph-based pipeline interface to allow high-level component interconnect
- Heterogenous processing on GPU and CPU
- Hides parallelization and synchronization under the hood
- Inherently multi-threaded



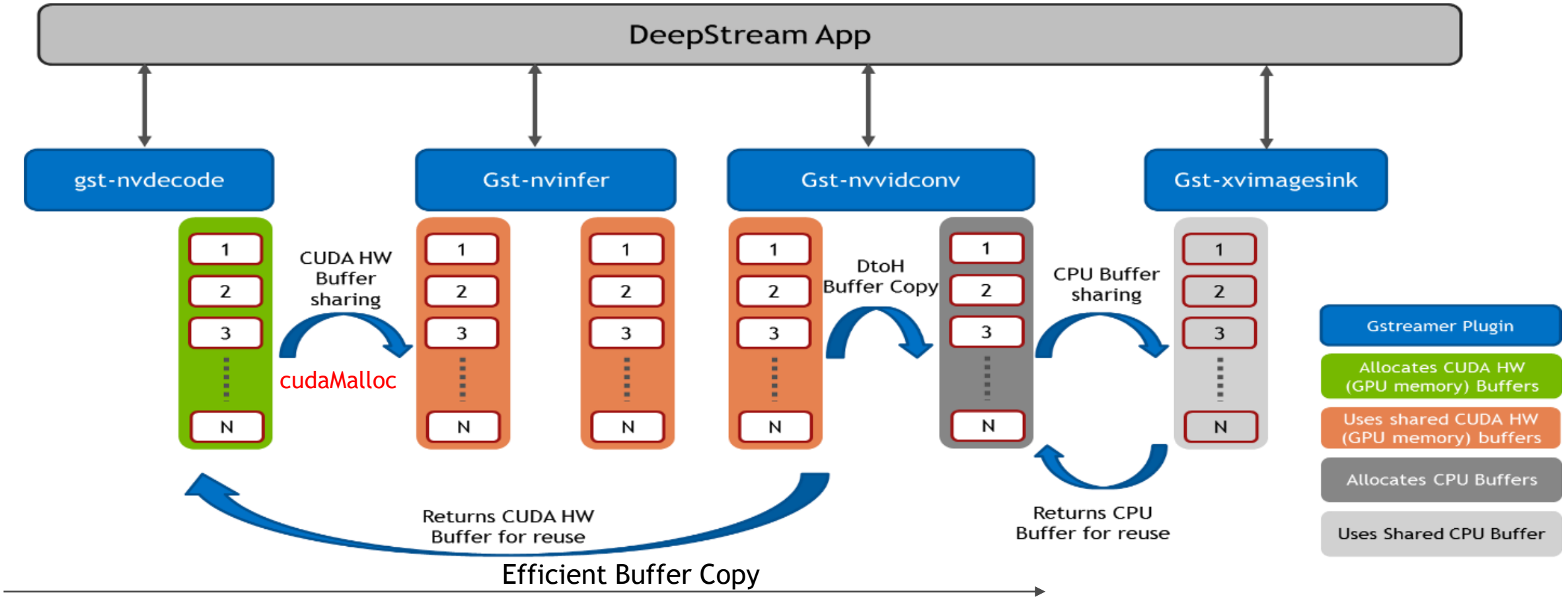
MEMORY MANAGEMENT

Efficient Memory Management



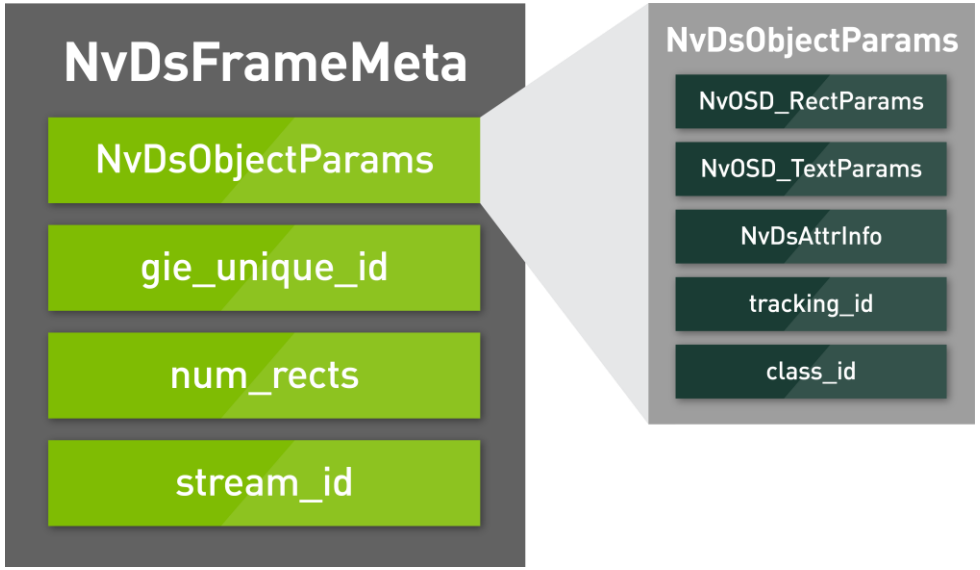
MEMORY MANAGEMENT

GPU to CPU Copy



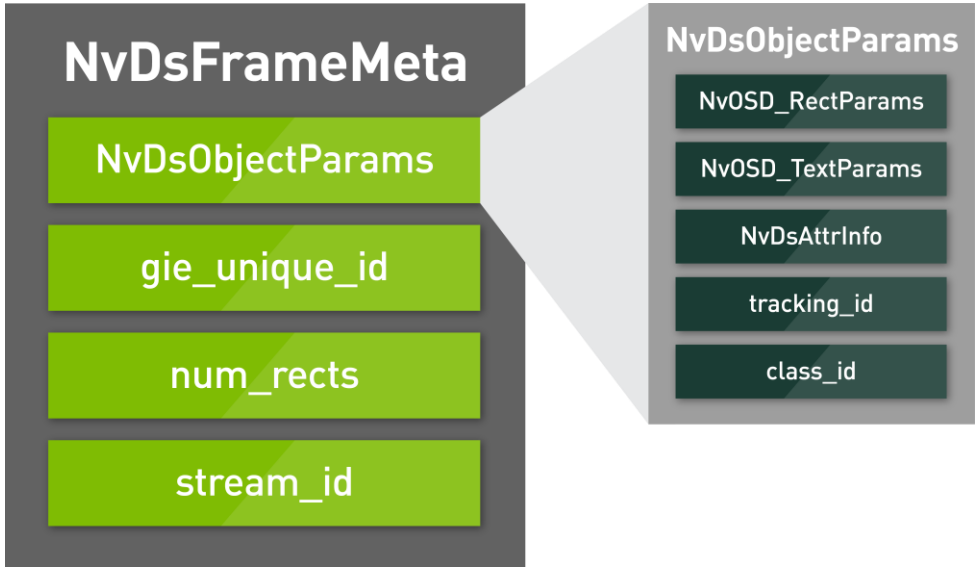
METADATA STRUCTURE

- **NvDSObjectParams** – Contains a subset of metadata information for an object detected in the frame.
- **GIE_Unique_ID** - Multiple neural networks get assigned a unique ID.
- **Num_rects** – Number of objects detected in the frame.
- **Stream_Id** – In case of multi-stream, to identify we need a stream id to associate which stream the data belongs to.

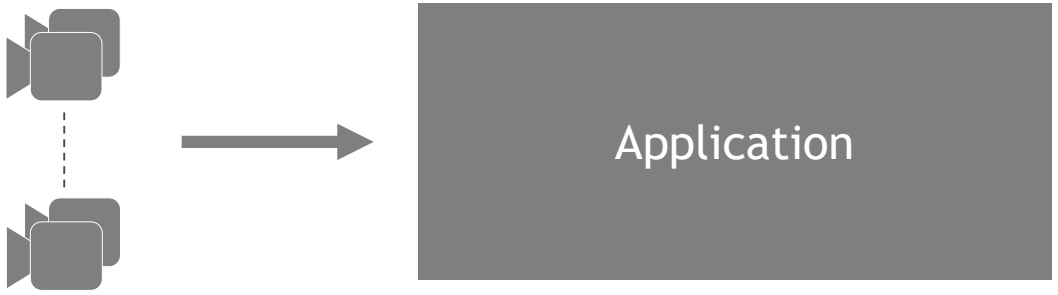


METADATA STRUCTURE

- **NvOSD_RectParams** – Bounding box coordinates
- **NvOSD_TextParams** – Label information required for display (white car, Mercedes, sedan)
- **NvDSAttribinfo** – Attributes of objects (type, color, make)
- **Tracking_ID** – Unique ID of that object from tracker
- **Class_ID** – Type of object (person, vehicle, two-wheeler, road sign)

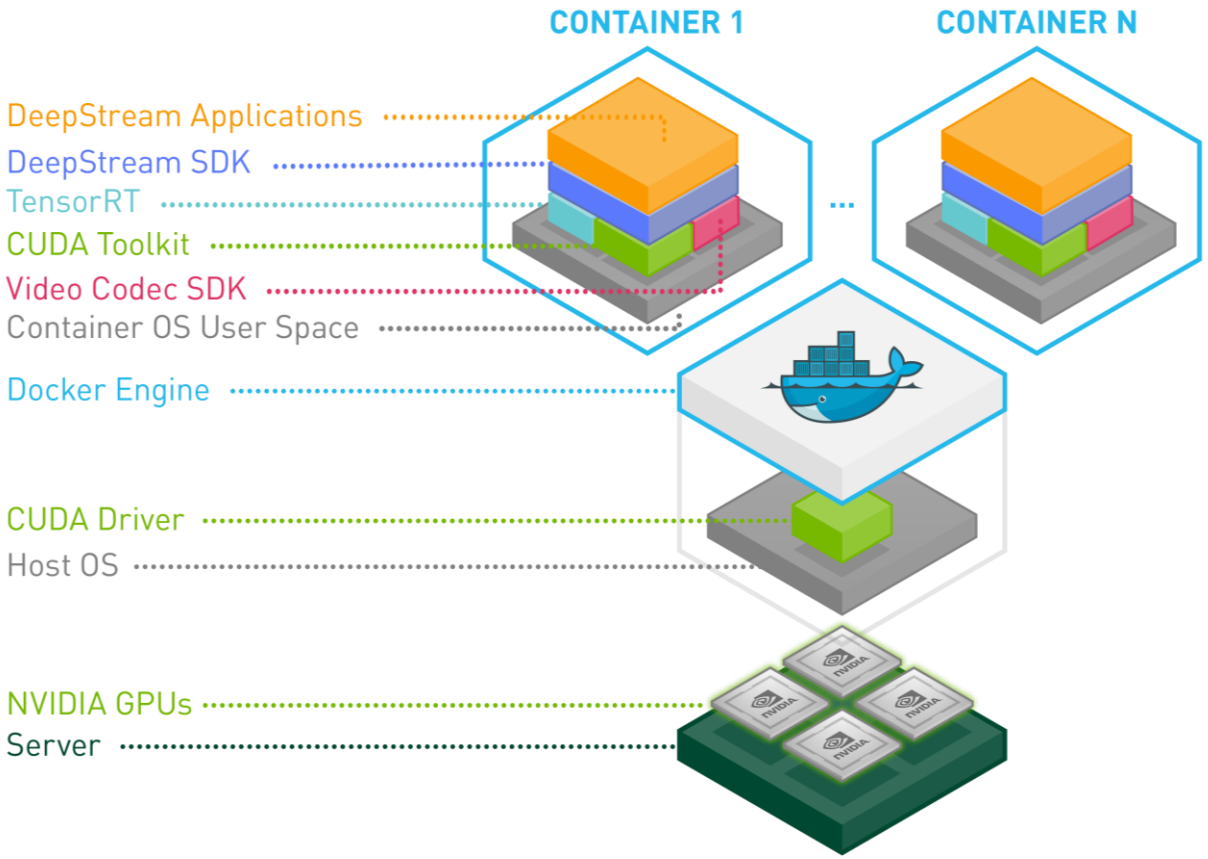


DYNAMIC STREAM MANAGEMENT



1. Add / Remove camera streams
2. Change FPS
3. Change resolutions

SCALE WITH DEEPSTREAM IN DOCKER



NVIDIA GPU CLOUD

Discover GPU-Accelerated Containers

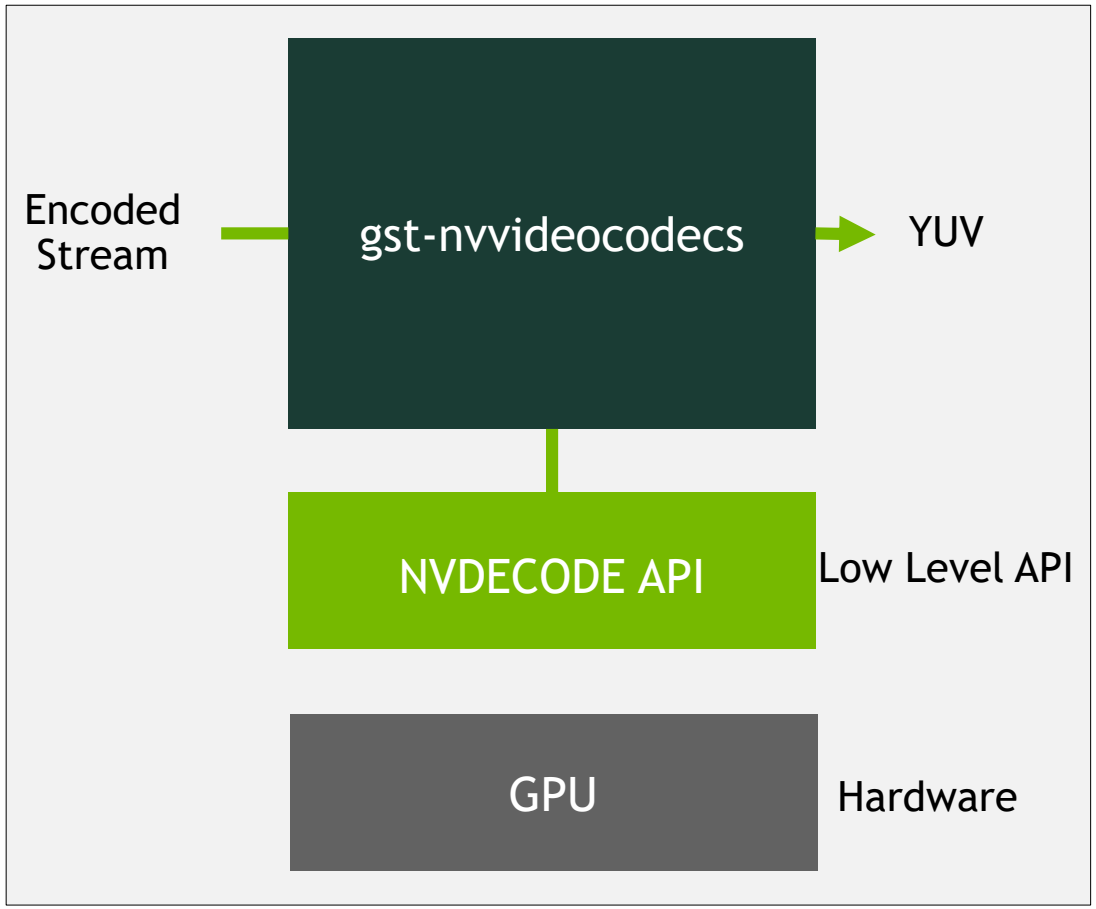
Innovate in Minutes, Not Weeks

Stay Up to Date

https://www.nvidia.com/en-us/gpu-cloud/

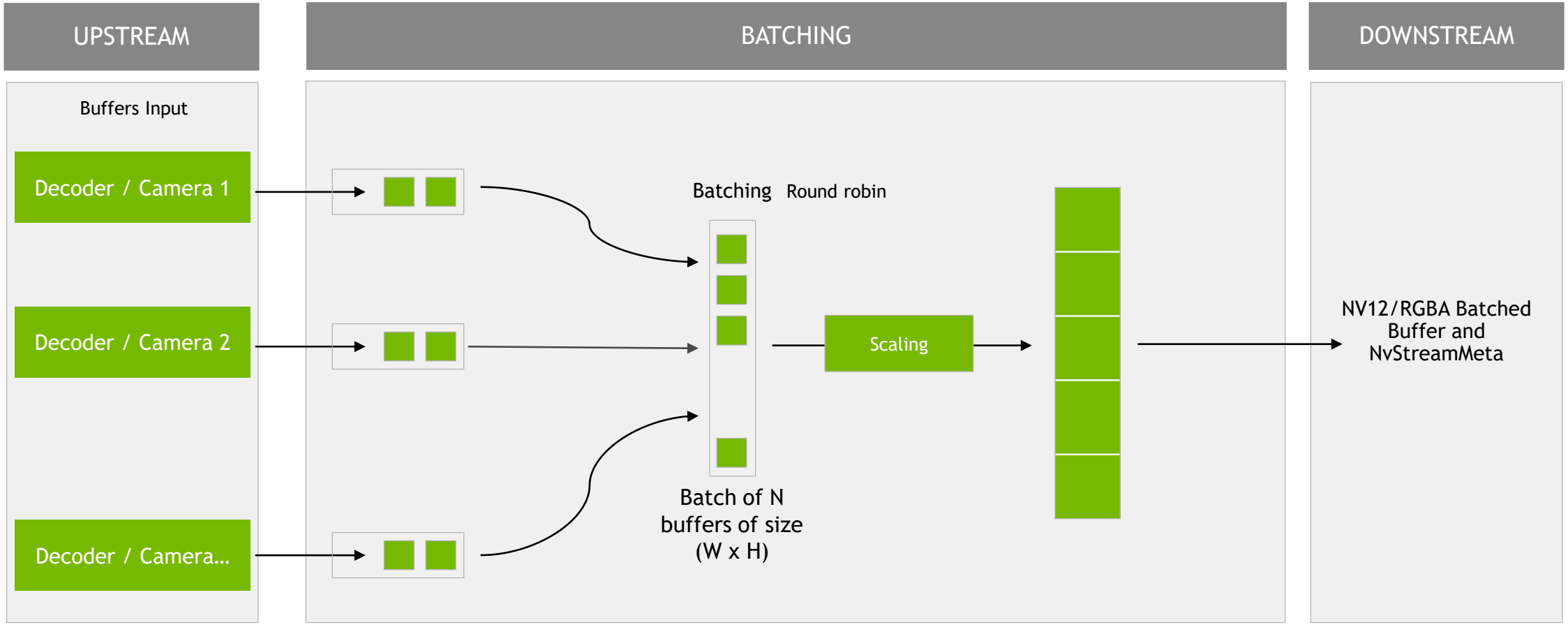
DECODER PLUGIN

gst-nvvideocodecs

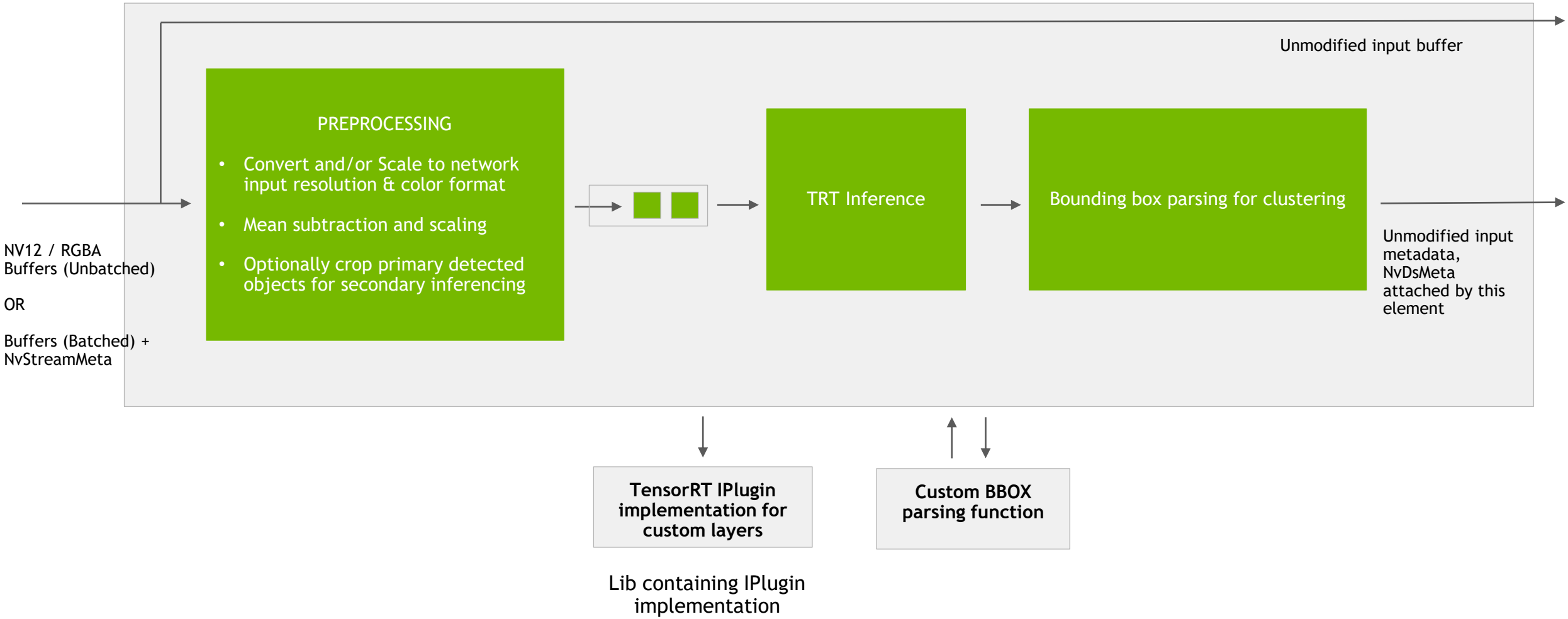


Input	H.264, H.265, VP8, VP9, MPEG2/4
Output	NV12
Parameters	Bit rate control, i-frame decoding

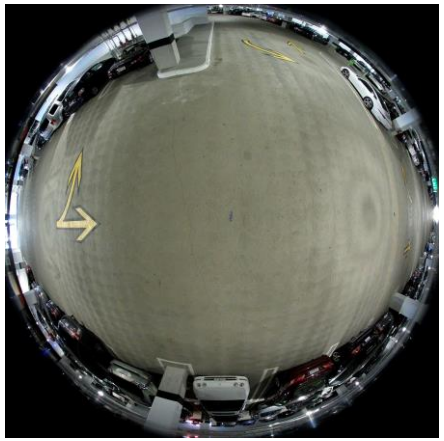
BATCHING - GSTNVSTREAMMUX



GSTNVINFER


Unmodified input buffer
PREPROCESSING
• Convert and/or Scale to network input resolution & color format
• Mean subtraction and scaling
• Optionally crop primary detected objects for secondary inferencing
TRT Inference
Bounding box parsing for clustering
Unmodified input metadata, NvDsMeta attached by this element
NV12 / RGBA Buffers (Unbatched)
OR
Buffers (Batched) + NvStreamMeta
TensorRT IPlugin implementation for custom layers
Custom BBOX parsing function
Lib containing IPlugin implementation

ENABLING 360D CAMERA PROCESSING



NVWARP360
SDK

Equirectangular

Perspective

Cylindrical

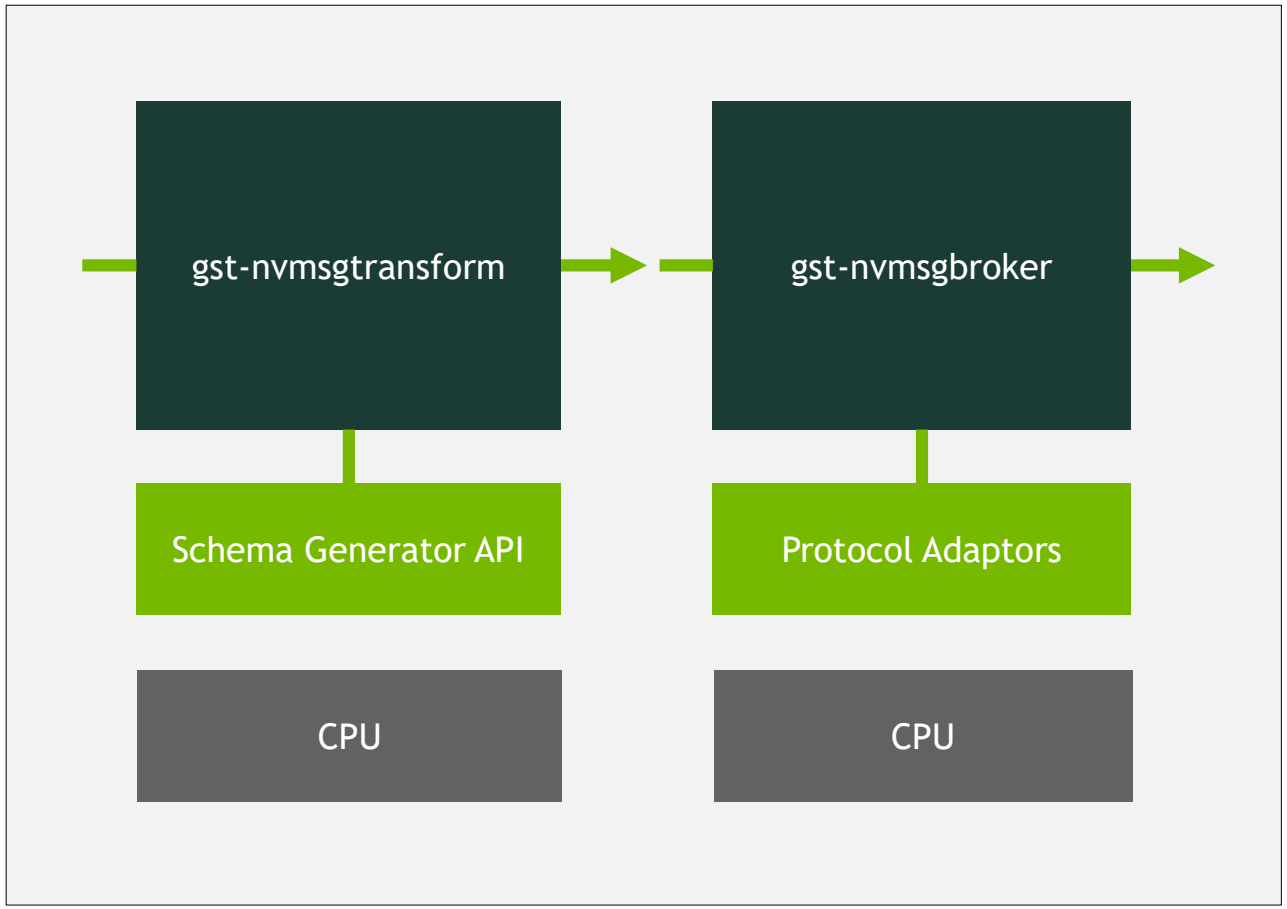
Rotated cylinder

Panini

Pushbroom

METADATA TO MESSAGE BROKERS

gst-nvmsgtransform & gst-nvmsgbroker



Input	Metadata
Output	Message sent over supported protocol
Parameters	Path to shared library implementing message generation from metadata based on schema Protocol, URL, port, topic for message destination Path to shared library implementing adaptor for desired protocol

NVIDIA-ACCELERATED PLUGINS

Plugin Name	Functionality
gst-nvvideocodecs	Accelerated video decoders
gst-nvstreammux	Stream aggregator - muxer and batching
gst-nvinfer	TensorRT based inference for detection & classification
gst-nvtracker	Reference KLT tracker implementation
gst-nvosd	On-Screen Display API to draw boxes and text overlay
gst-tiler	Renders frames from multi-source into 2D grid array
gst-eglglessink	Accelerated X11 / EGL based renderer plugin
gst-nvvidconv	Scaling, format conversion, rotation
Gst-nvdewarp	Dewarping for 360 Degree camera input
Gst-nvmsgconv	Meta data generation
Gst-nvmsgbroker	Messaging to Cloud


BUILDING WITH
DEEPSTREAM

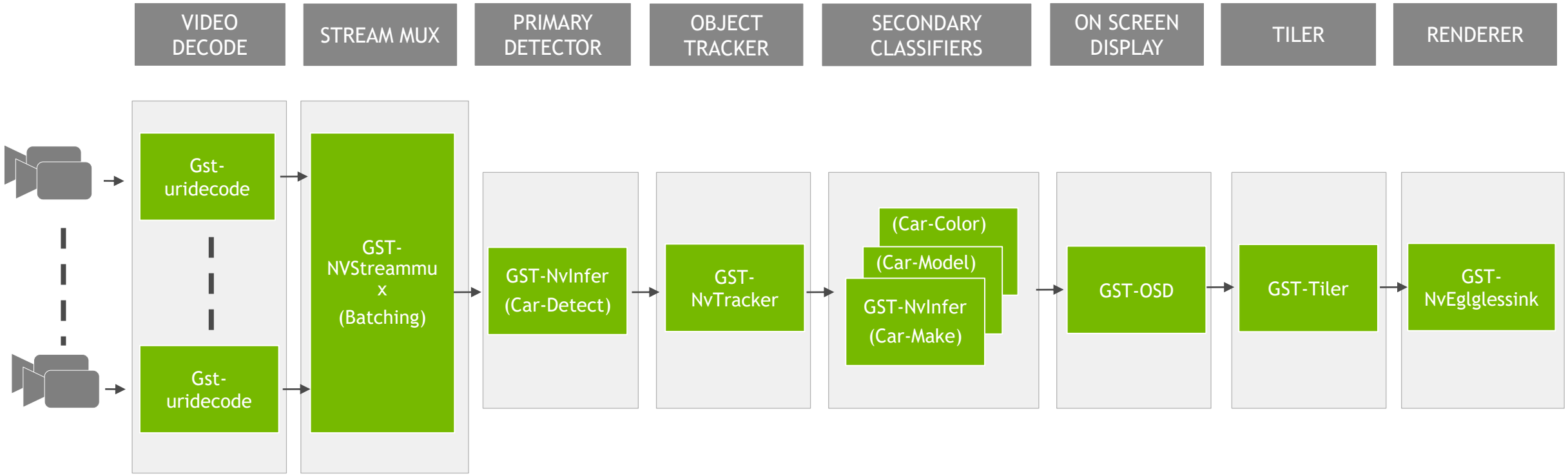
CONFIGURATION FILE

```
enable-perf-measurement=1   //To enable performance measurement
perf-measurement-interval-sec=10  //Sampling interval in seconds for performance metrics
flow-original-resolution=1  //Stream muxer flows original input frames in pipeline
#gie-kitti-output-dir=/home/ubuntu/kitti_data/    // location of KITTI metadata files
```

```
[source0]
enable=1 // Enables source0 input
#Type - 1=CameraV4L2 2=URI 3=MultiURI  //1) Input source can be USB Camera (V4L2)
// 2)URI to the encoded stream. Can be a file,HTTP URI or an RTSP live source
// 3) Select URL from multi-source input
type=3 // Type of input source is selected
uri=file://../../streams/sample_720p.mp4 // Actual path of the encoded source.
num-sources=1 // Number of input sources.
gpu-id=0 // GPU ID on which the pipeline runs within a single system
```

DeepStream reference config file: source4_720p_resnet_dec_infer_tracker_sgie_tiled_display_int8.txt.

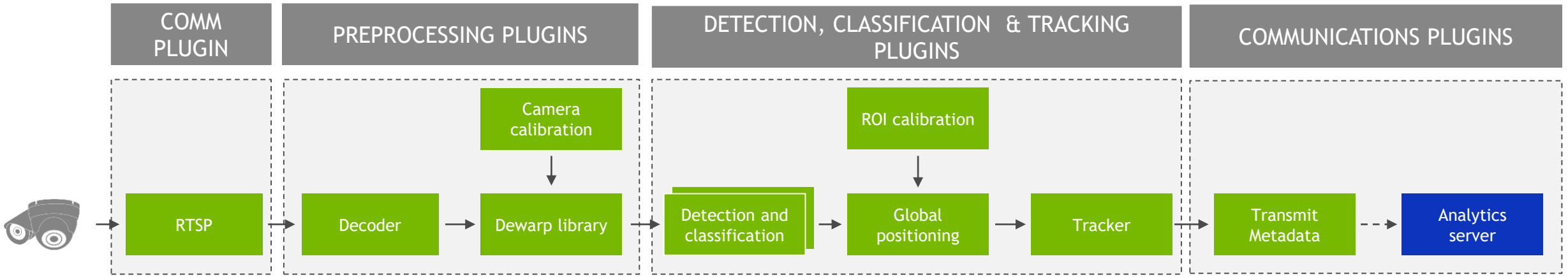
MULTI-STREAM REFERENCE APPLICATION



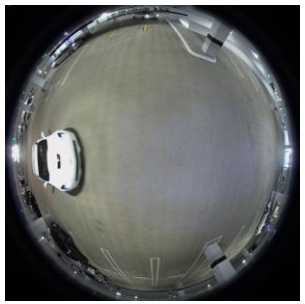

REFERENCE APPLICATION VIDEO


SMART PARKING SOLUTIONS WITH DEEPSTREAM

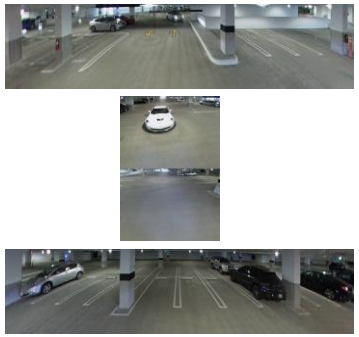
PERCEPTION GRAPH



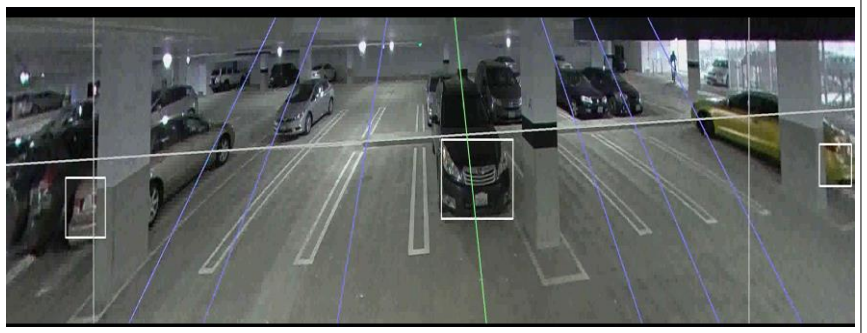
360 Degree Feeds



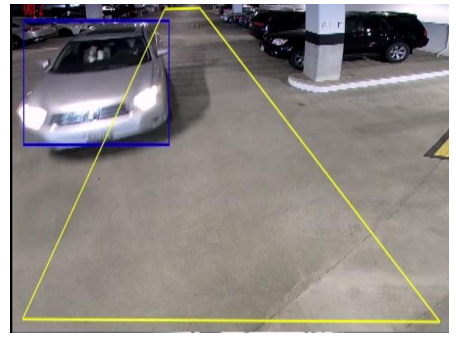
Dewarping



ROI: Lines

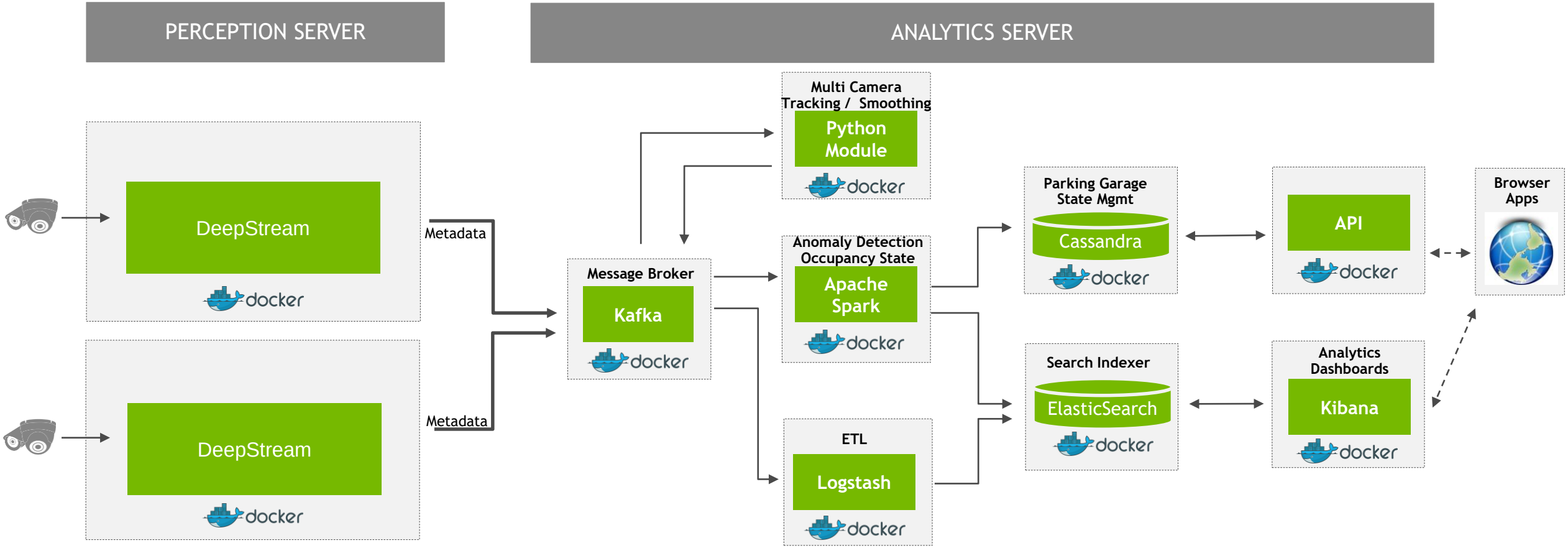


ROI: Polygon



PERCEPTION AND ANALYTICS IN CONTAINERS

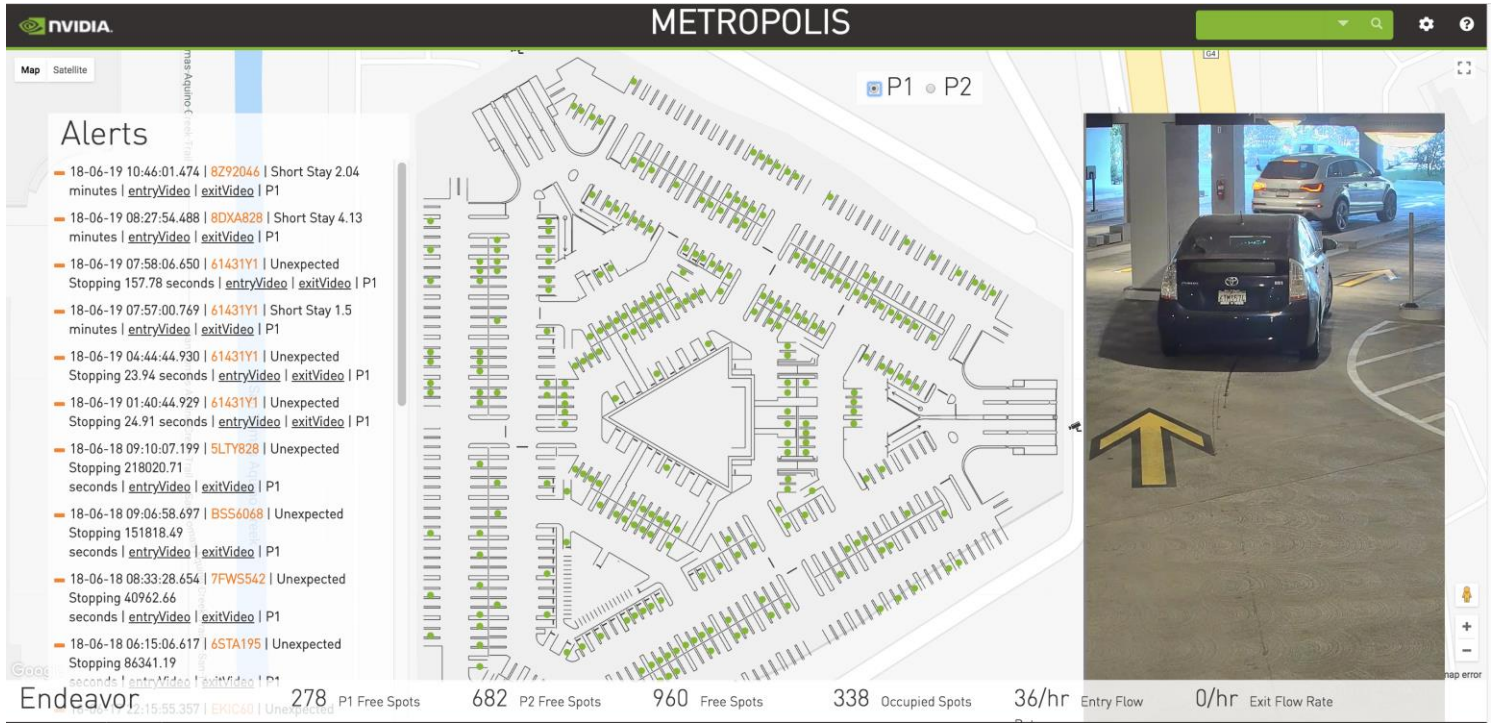
SCALING AND PORTABILITY



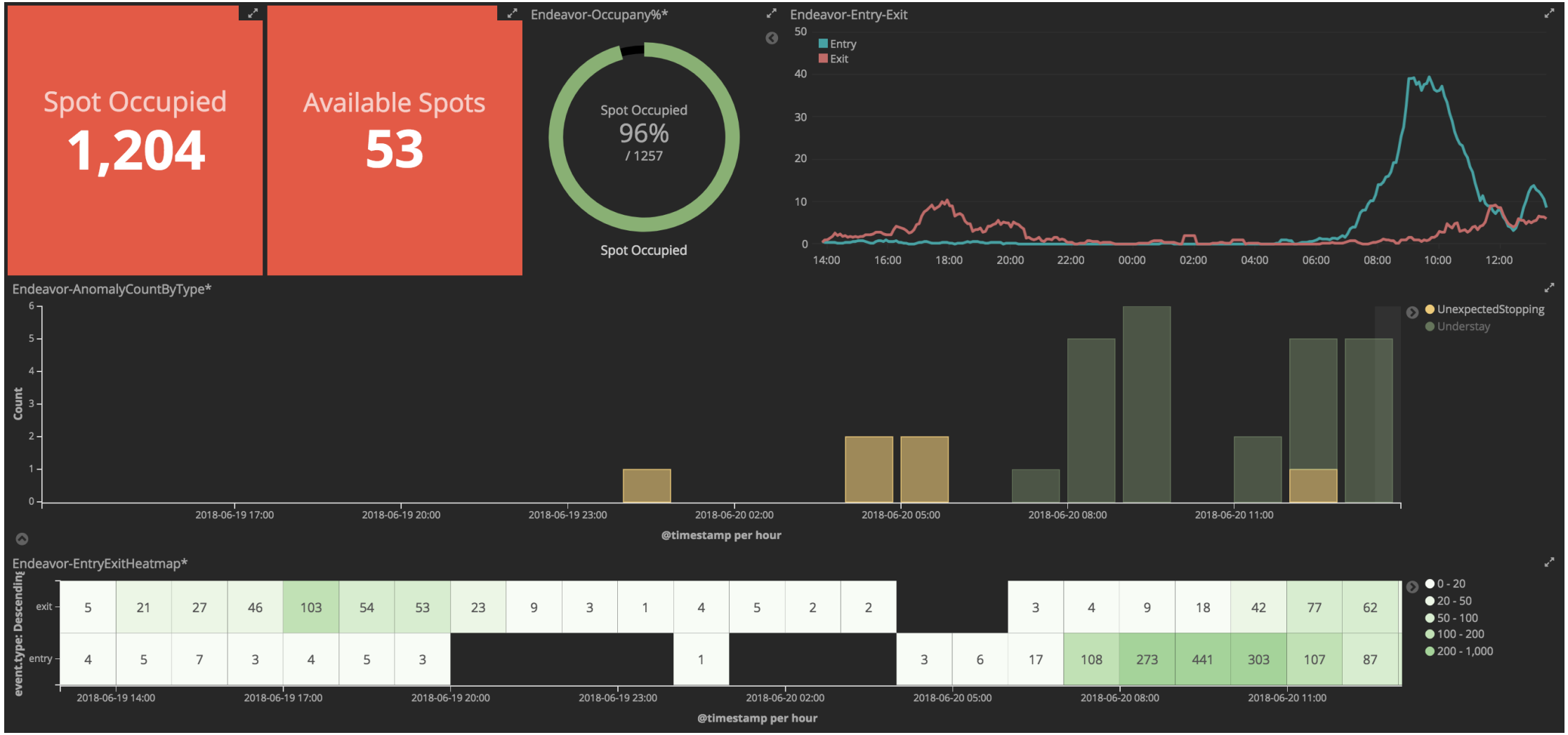
COMMAND CENTER UI

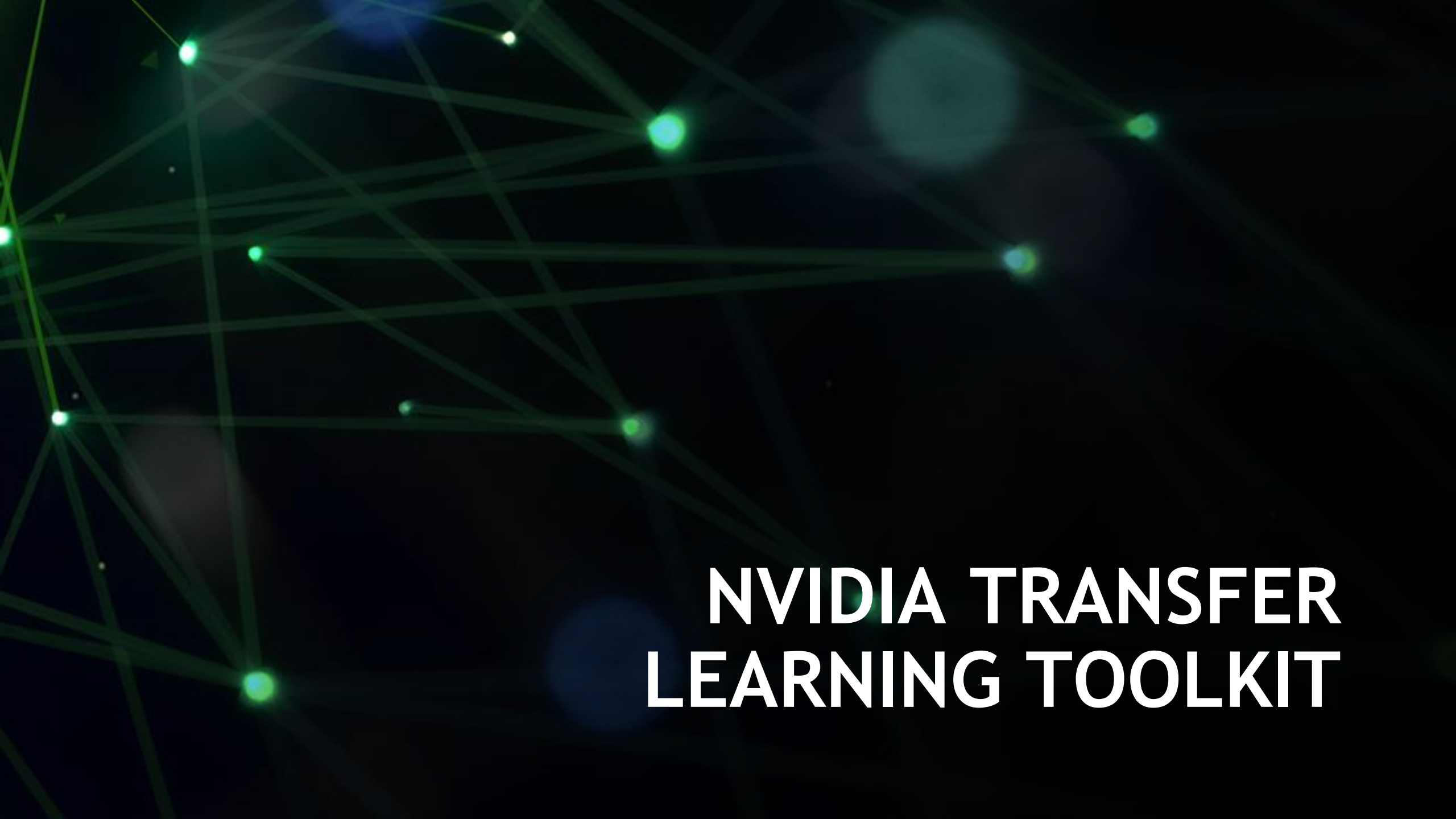
Browser-Based Interface

- Occupancy Map for Multiple Levels
- Sensor Fusion
- Anomalies
- Search Events and Anomalies
- Occupancy Stats and Flow Rates

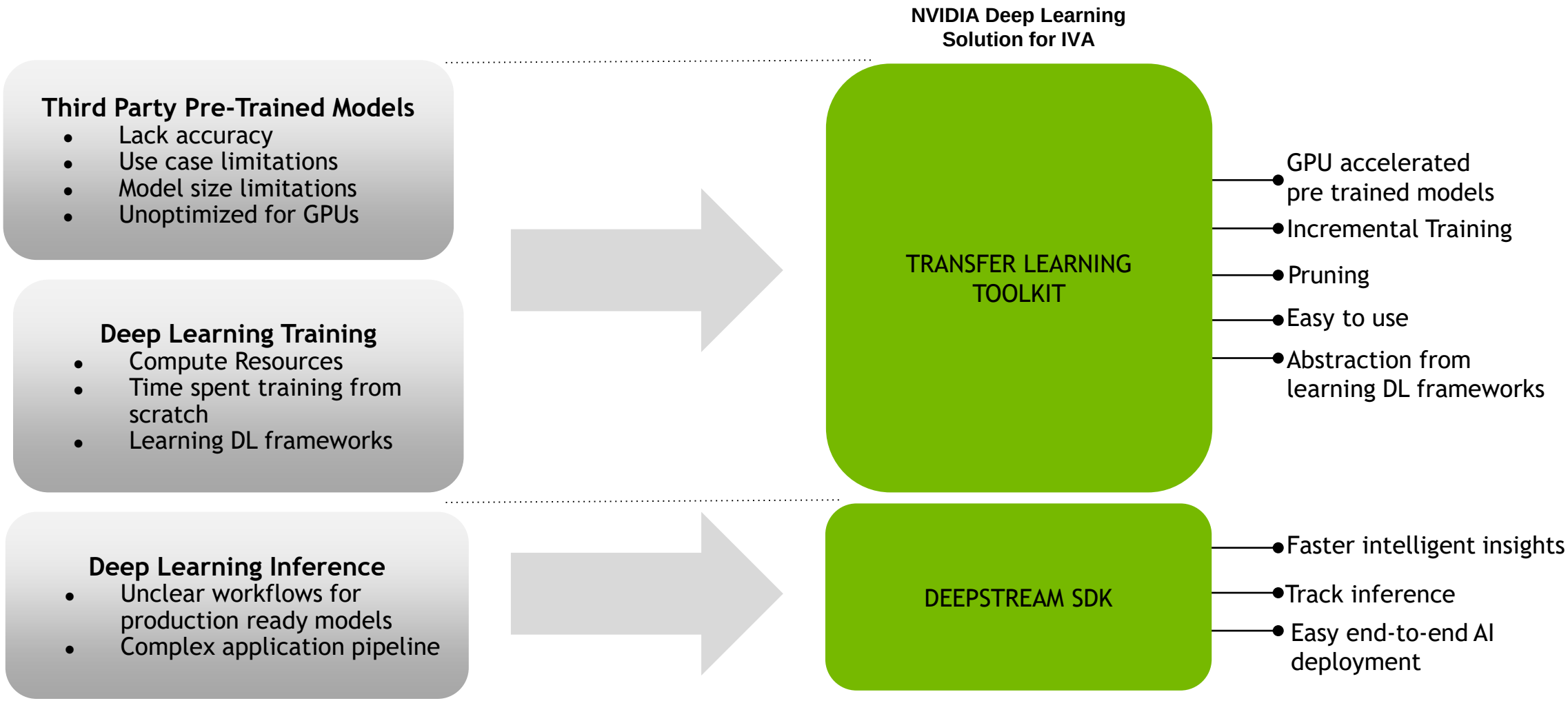


ANALYTICS DASHBOARD




NVIDIA TRANSFER
LEARNING TOOLKIT

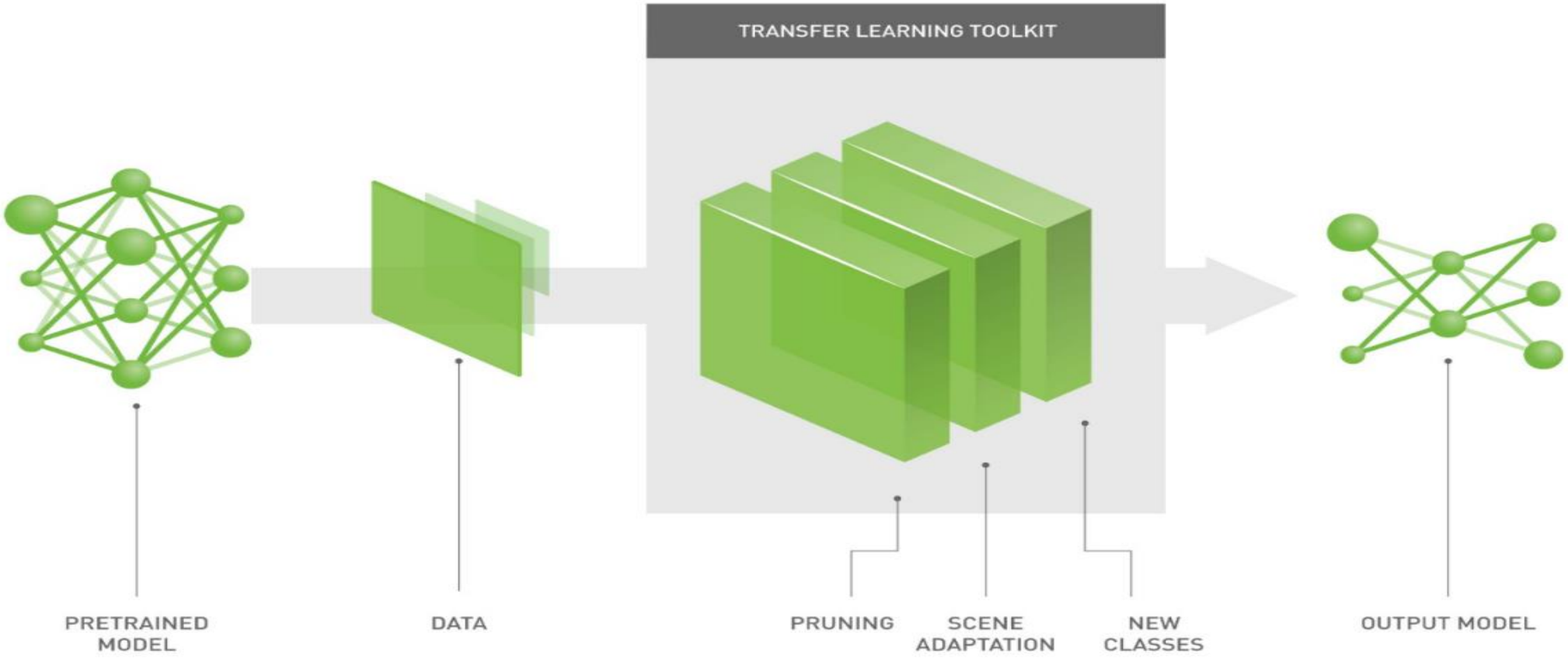
IVA DEEP LEARNING WORKFLOW MANAGEMENT



Accelerate deep learning training with pre-trained models and functions provided by Transfer Learning Toolkit

Transfer Learning Workflow

Fine Tuning * Pruning * Scene Adaptation * New Classes



Output Model ready to be deployed and integrated for us with DeepStream SDK 3.0 applications

Workflow With DeepStream

PRIMARY OBJECT DETECTOR

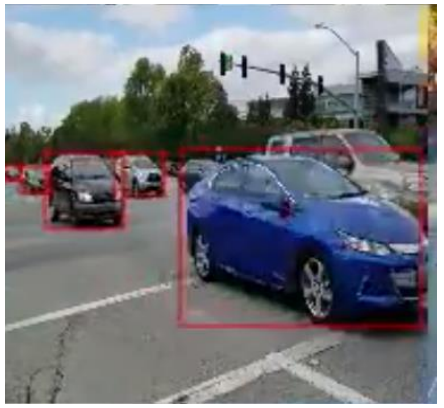
4 class object detector

SECONDARY IMAGE CLASSIFIER

Classifier Car, Truck, Bicycle, Pedestrian

- Train with new data
- Adapt to scene
- Add new classes
- Prune the models
- Export in a DeepStream application

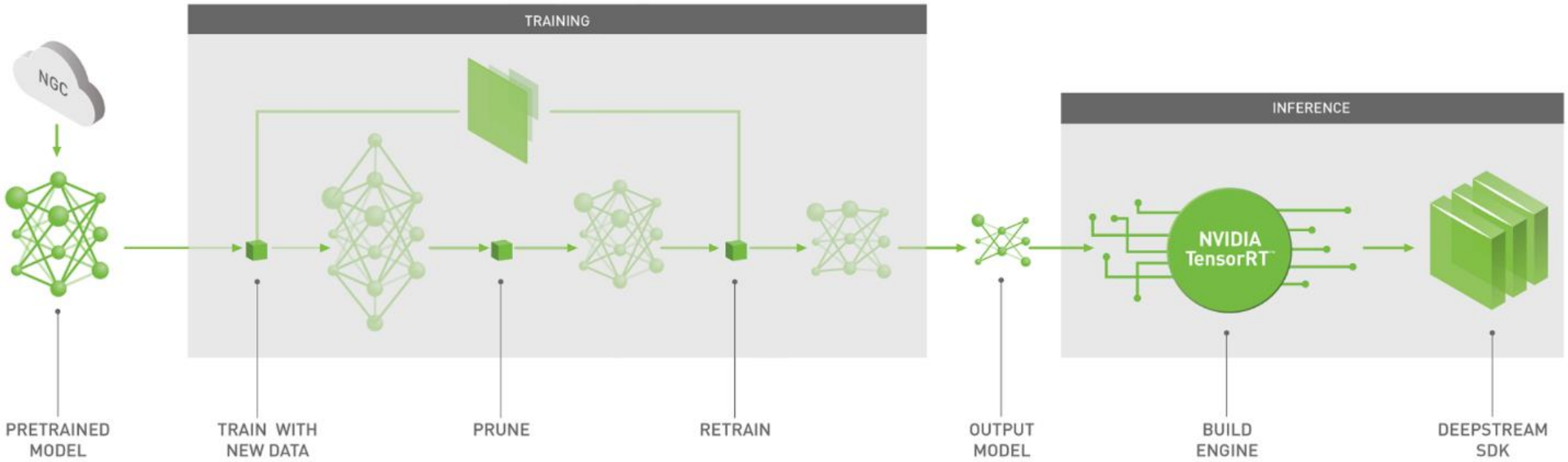
Transfer Learning Toolkit for IVA



DeepStream Application

End to End Deep Learning Workflow

Pre-Trained models * Training & Adaptation pipeline * Applications ready for DeepStream integration



Accelerate time to market and save on compute resources!

FEATURES

Efficient Pre-trained Models

GPU-accelerated models trained on large scale public datasets.

Faster Inference with Model Pruning

Model pruning reduces size of the model resulting in faster inference

Training with Multiple GPUs

Re-training models, adding custom data for multi GPU training using an easy to use tool

Abstraction

Abstraction from having deep knowledge of frameworks, simple intuitive interface to the features

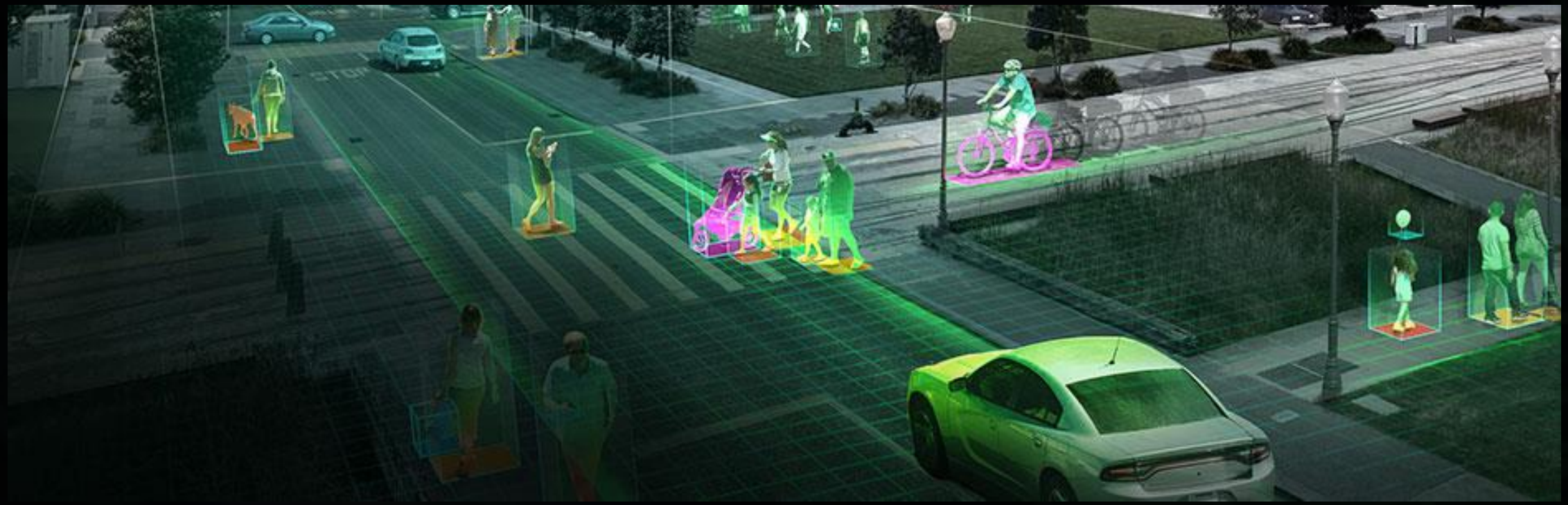
Containerization

Packaged in a container easily accessible from NVIDIA GPU Cloud website. All code dependencies are managed automatically

Integration

Models exported using TLT are easily consumable for inference with **Deep Stream SDK**

START DEVELOPING WITH NVIDIA TRANSFER LEARNING TOOLKIT AND DEEPSTREAM



DEEPSTREAM . TRANSFER LEARNING TOOLKIT . SUPPORT FORUMS