Taking Intelligent Video Analytics to the Next Level

Avi Baum, CTO
Hailo

Outline

- What is **Intelligent** video analytics?
- What is the **"current"** level?
- What is the **"next"** level? … And the need for **more AI**
- How to get there?
  - Addressing the need for **more AI**
  - Hardware integration
  - Software integration
- Example case: **Advanced Analytics in ITS**
- Closing remarks

What's the "current" level?

- Most analytics is based on **basic object detection** with additional heuristics applied to conclude more advanced insights
- Processing limits impose **down selection** among video sources
- Processing limits impose video **sub-sampling** both in **resolution** and **time**
- Reliance on manpower to address **high miss rates**
- Relying on **manual configuration**

What's the "next" level?

Using **advanced algorithms** to apply scene understanding and extract **context-aware insights**

Enhancing **existing** application

Improving **accuracy**

Lowering **latency**

Enabling **new** applications

Identify **intent**

Offering **closed-loop** control

Minimizing operator interaction

Need for more AI

- When it comes to vision, AI has become the mainstream for perception
- Two factors come into play as the industry matures:
  - **On one hand…**
    - Progress in neural networks models are more **accurate** while **less demanding**
    - Applying domain-understanding further **relaxes** the required **processing capacity**
  - **On the other hand…**
    - Product grade **processing pipelines** more **complex** and therefore more demanding
    - **New capabilities** are added to extract more information, requiring **more compute**

How much AI capacity is needed? (Basic)

$$\boldsymbol{TOPS} = N_{cameras} \cdot \left(\frac{frames}{sec}\right) \cdot \left(\frac{pixels}{frame}\right) \cdot \left(\frac{TOPS}{pixel}\right)$$

- $N_{cameras}$ ➔ cameras per device ➔ determines **total cost of ownership**
- Frames/sec ➔ frame rate ➔ determines **temporal performance**
- Pixels/frame ➔ resolution ➔ determines **spatial performance**
- TOPS/pixel ➔ model capacity ➔ determines **accuracy**

How much AI is needed? (*slightly more* Advanced)

$$\boldsymbol{TOPS} = N_{cameras} \cdot \left(\frac{frames}{sec}\right) \cdot \left[\left(\frac{pixels}{frame}\right) \cdot \left(\frac{TOPS}{pixel}\right) + \left(\frac{\boldsymbol{region}}{\boldsymbol{frame}}\right) \cdot \left(\frac{\boldsymbol{pixels}}{\boldsymbol{region}}\right) \cdot \left(\frac{\boldsymbol{TOPS}}{\boldsymbol{pixel}}\right)\right]$$

- Real deployment usually requires **multi-stage pipeline** processing
- This requires **more compute capacity** to address real deployment scenarios
- **Peak-to-average** ratio reflecting worst and typical cases, needs to be considered for proper design
- Each "region" (ROI) is further processed; the amount of processing is determined by
  - Number of ROIs
  - Size of ROIs
  - Processing intensity per ROI

Why do we care?

$$\boldsymbol{TOPS} = N_{cameras} \cdot \left(\frac{frames}{sec}\right) \cdot \left[\left(\frac{pixels}{frame}\right) \cdot \left(\frac{TOPS}{pixel}\right) + \left(\frac{\boldsymbol{region}}{\boldsymbol{frame}}\right) \cdot \left(\frac{\boldsymbol{pixels}}{\boldsymbol{region}}\right) \cdot \left(\frac{\boldsymbol{TOPS}}{\boldsymbol{pixel}}\right)\right]$$

- 1st stage is governed by the **input BW**
  - Proper design need to guarantee it isn't becoming a system bottleneck
- 2nd stage is governed by the **quality** (accuracy) of the 1st stage
  - High false positives will result in excess compute demand
  - High false negative will result in poor results

Addressing the need: Powerful Baseline

Hailo-8™ AI Processor

- 26 TOPS
- 17 x 17 FCBGA

Hailo-8™ M.2 AI Acceleration Module

- PCIe Interface
- M.2 form factor
  - M key (2242/2260/2280)
  - B+M key (2242/2260/2280)
  - A+E key (2230)
- Extended temperature -40° up to 85°

M key 4 lanes | B+M key 2 lanes | A+E key 2 lanes

Hailo-8™ Mini PCIe AI Acceleration Module

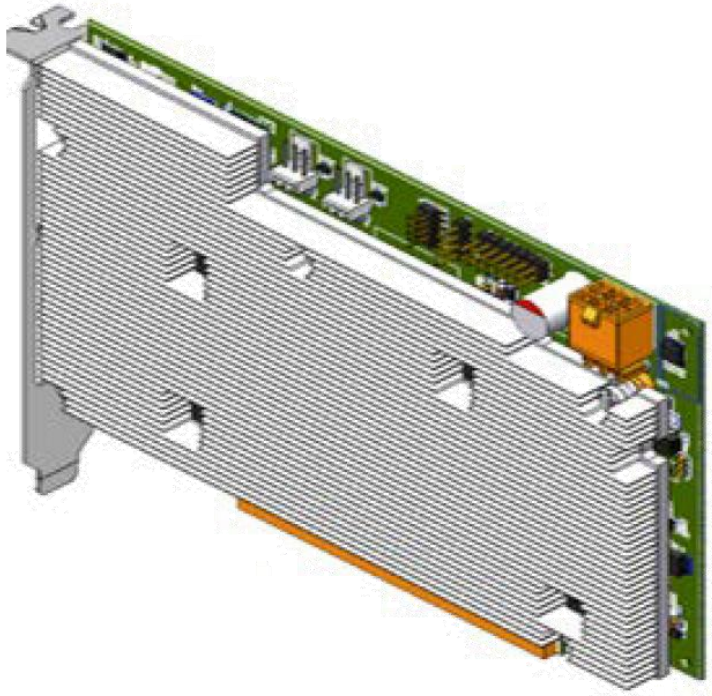
- PCIe Interface
- mPCIe form factor 3050
- Extended temperature support: -40° up to 85°

Hailo-8™ Century Evaluation Platform

- PCI Interface
- Multi-chip configuration
- 104 TOPS
- Typical power: 25 W

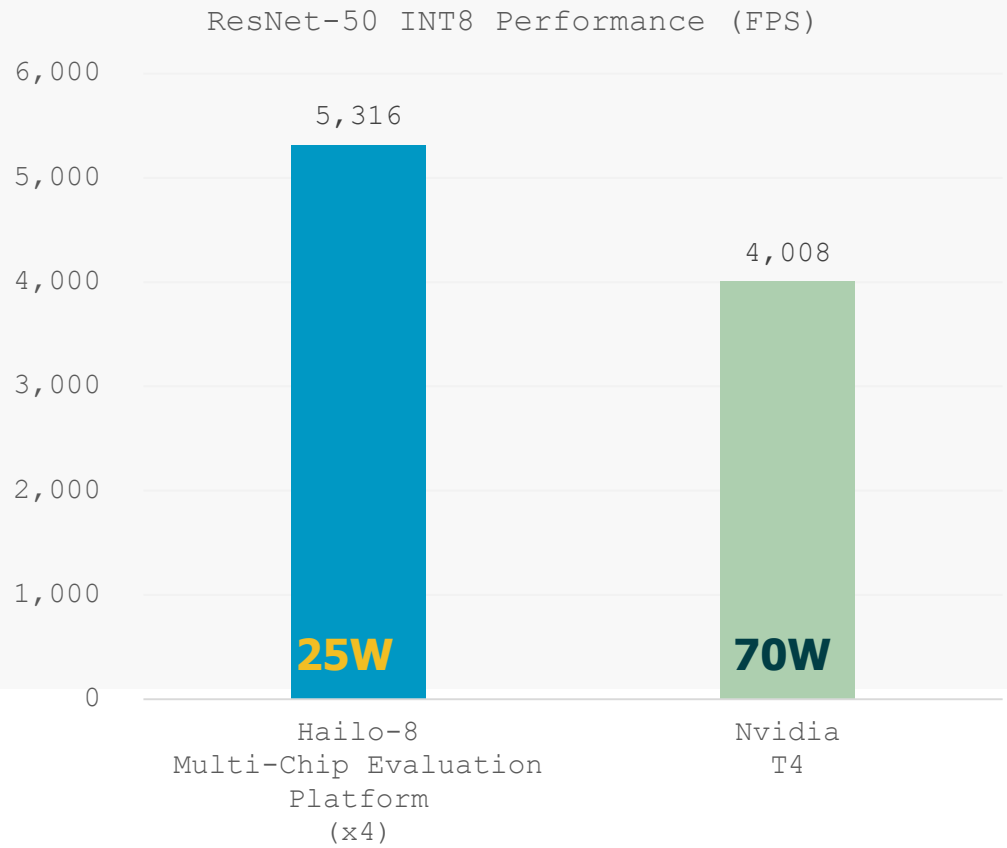
Addressing the need: Capacity Scaling

- High-performance video analytics, **single-slot PCIe accelerator** card
- **Efficient**, high-performance design
  - Delivering up to **156 TOPS** for vision processing at up to **35 W**
  - Higher **cost-efficiency** (TOPS/$) compared with existing solutions
  - Guaranteed product **longevity** and **extended temperature** range
- **Mature** software toolchain
  - Supports state-of-the-art NN models and application pipelines
  - Low transition barrier from existing solutions
  - Resource assignment **granularity** enabling elastic workload assignment
  - **Physically separate** assignment for data protection

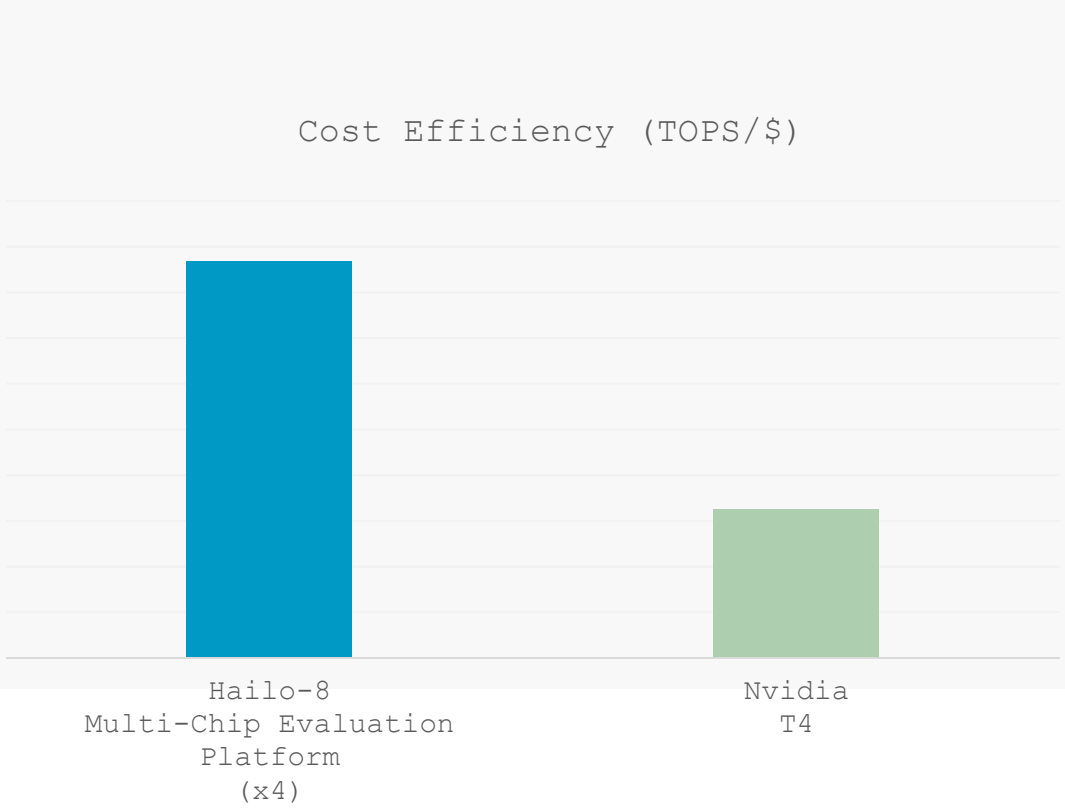


Addressing the need: Capacity Scaling

More performance at **1/3 of the power**



x3 more performance per $



- *Nvidia T4 figures source: https://developer.nvidia.com/deep-learning-performance-training-inference*
- *Based on maximum performance claims and market pricing*

Addressing the need: Platform Scalability

1 device
26 TOPS

Up to 2 devices
52 TOPS

Up to 6 devices
156 TOPS

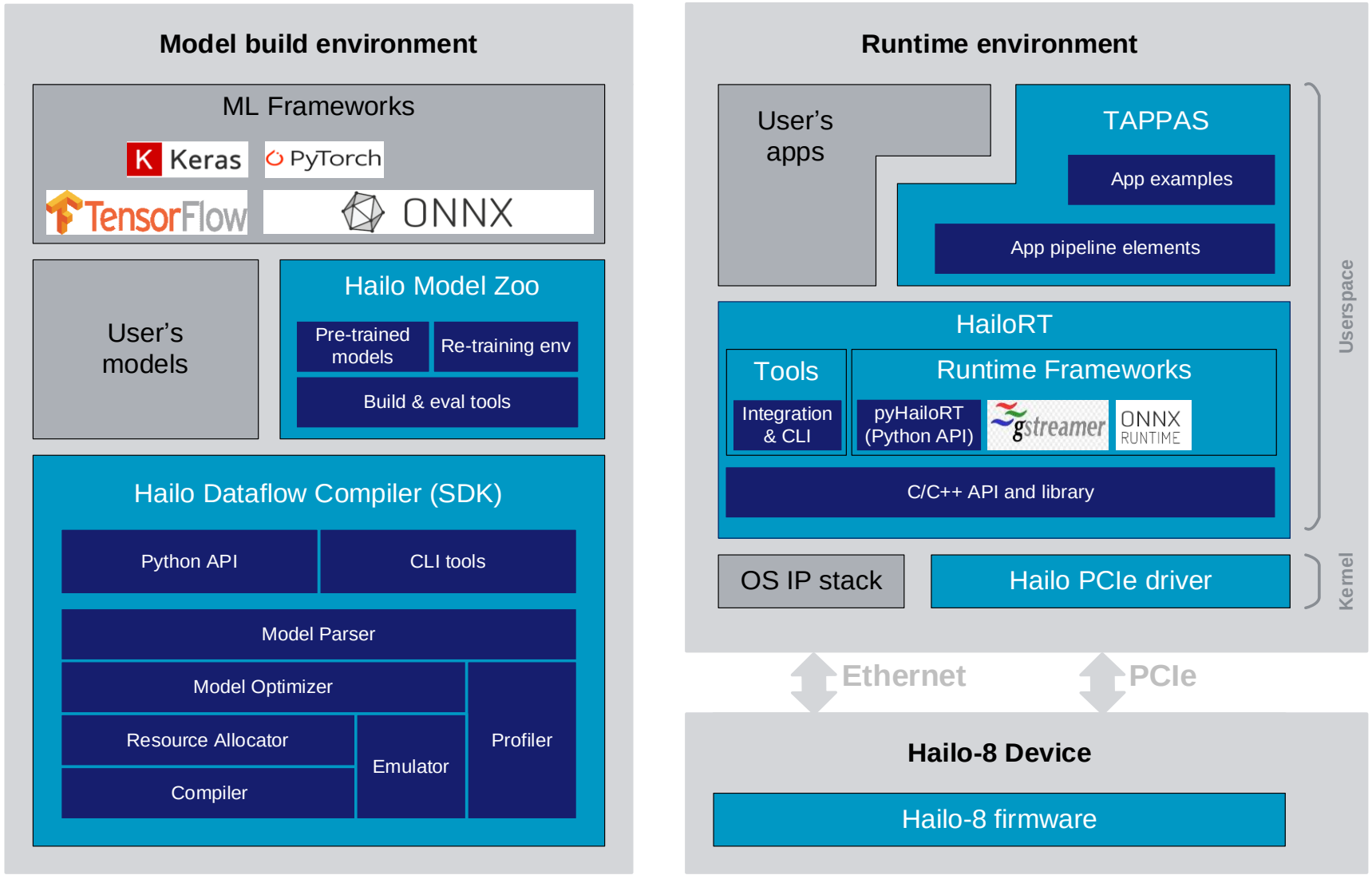
Up to 8 devices
208 TOPS

Up to 12 devices
312 TOPS

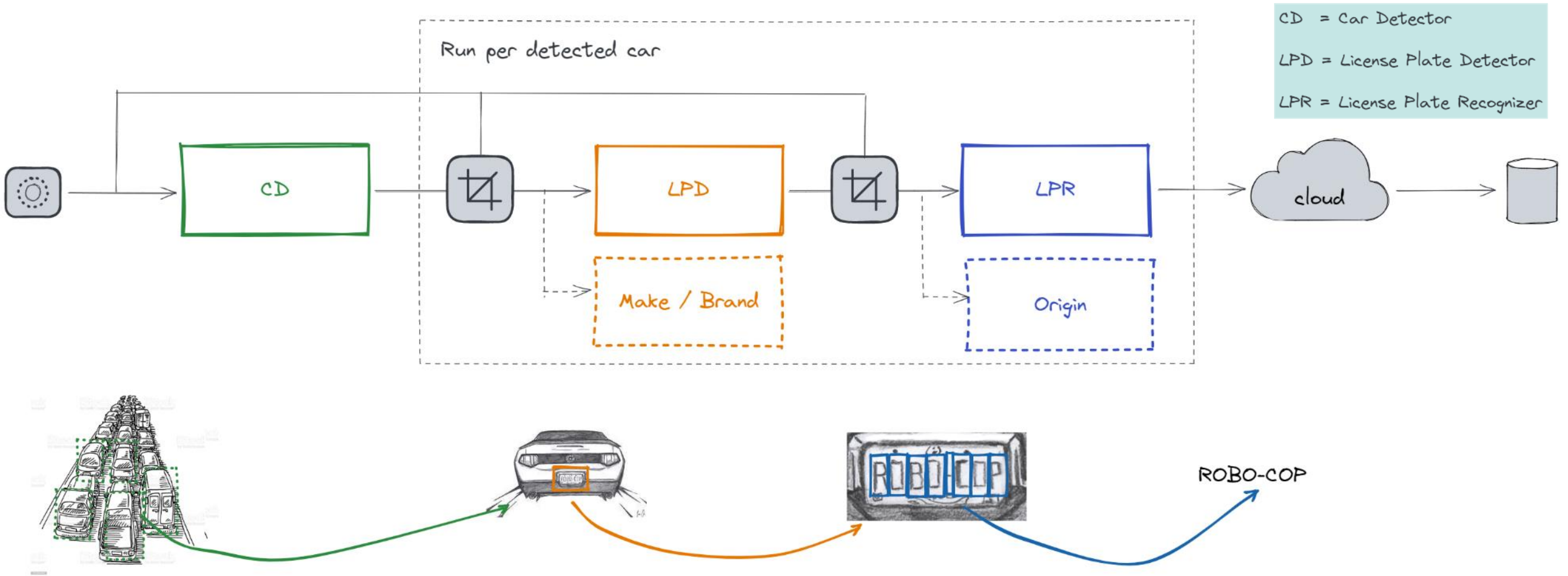
Addressing the need: Rich Ecosystem

AAEON an ASUS assoc. co. · Analog Tech アナログ・テック株式会社 · AXIOMTEK · BASLER
Compulab · DELL · FOXCONN · iBASE
kontron · Lanner · Leopard · MicroSys Creating Embedded Systems
NEXCOM · REGISTERED PARTNER NXP · socionext · Variscite
Vecow

Addressing the need: Rich & Mature Software



Advanced analytics in ITS: ALPR Pipeline



Advanced analytics in ITS: ALPR Pipeline Performance

- Overall system performance is determined by the ability to run the pipeline at the required rate
- The **required rate** is given by the acquisition rate at full resolution and number of detections
- To achieve desirable accuracy, frame rate should allow enough detection for fastest movement

$$Acc_{required} = \left(1 - \left(1 - P_{success_{CD}}\right)^{min\left(\frac{R_{camera}}{V_{car}}, FPS_{camera}\right)}\right)$$

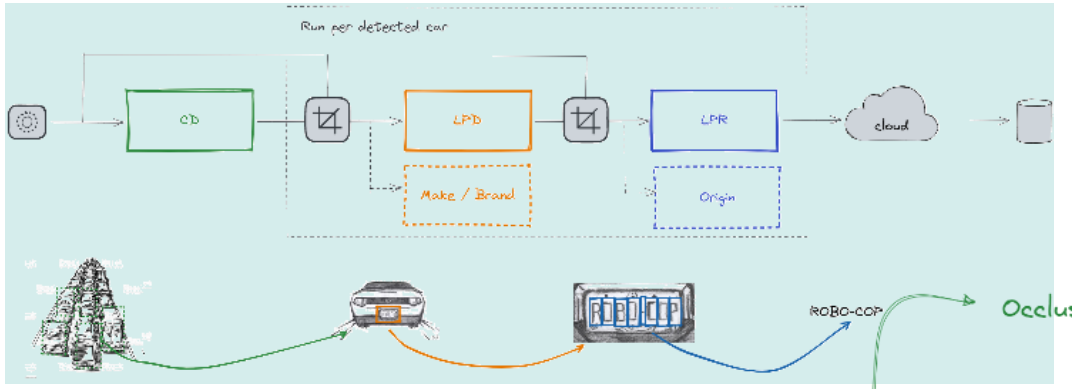
- Detection accuracy determines performance

$$P_{success} = P_{success\_CD} \cdot P_{success\_LPD} \cdot P_{success\_LPR}$$

- High frame rate
- Better detector
- More classes

➔Improved detection accuracy

Advanced analytics in ITS: ALPR Pipeline Performance



Occlusion

Glare

Motion blur

Misleading box

TP TP TP FP FP FP

TN FN

TN FN

TN FN

Undetectable characters
(low accuracy; out-of-odistribution)

Missed plates (low mAP)

Missed objects (low mAP)

Closing Remarks

- Advanced AI in Video Analytics enables
  - Lower total cost of ownership (**TCO**) … By aggregating more cameras
  - Better **accuracy** … By deploying more advanced models
  - **Versatility** … Enabling domain adaptation
  - **Richer experiences** … With more complex application pipelines
  - Solution **scalability** … Develop once run everywhere

- Start design with the rich portfolio of Hailo-8 powered solutions **today!**

Further Information

More on Hailo on out website

Technology

https://hailo.ai/technology/

Demos

https://hailo.ai/resources/#demos

Benchmarks

https://hailo.ai/developer-zone/benchmarks/

contact@hailo.ai

2022 Embedded Vision Summit

You are welcome to visit us,

in our booth #313

to our representatives

and witness our demos