2021 embedded VISION summit

VIRTUAL | MAY 25-28

Productizing Edge AI Across Applications and Verticals: NEC Case Study and Insights

Orr Danon

CEO, Hailo Technologies

Tsvi Lev

Managing Director of NEC Research Center Israel and Corporate Vice President, NEC

HAILO NEC

NEC – Hailo Collaboration

Background

- Started 1 year after company inception
- Initial focus on Public Safety
- Expanded to multiple projects and teams

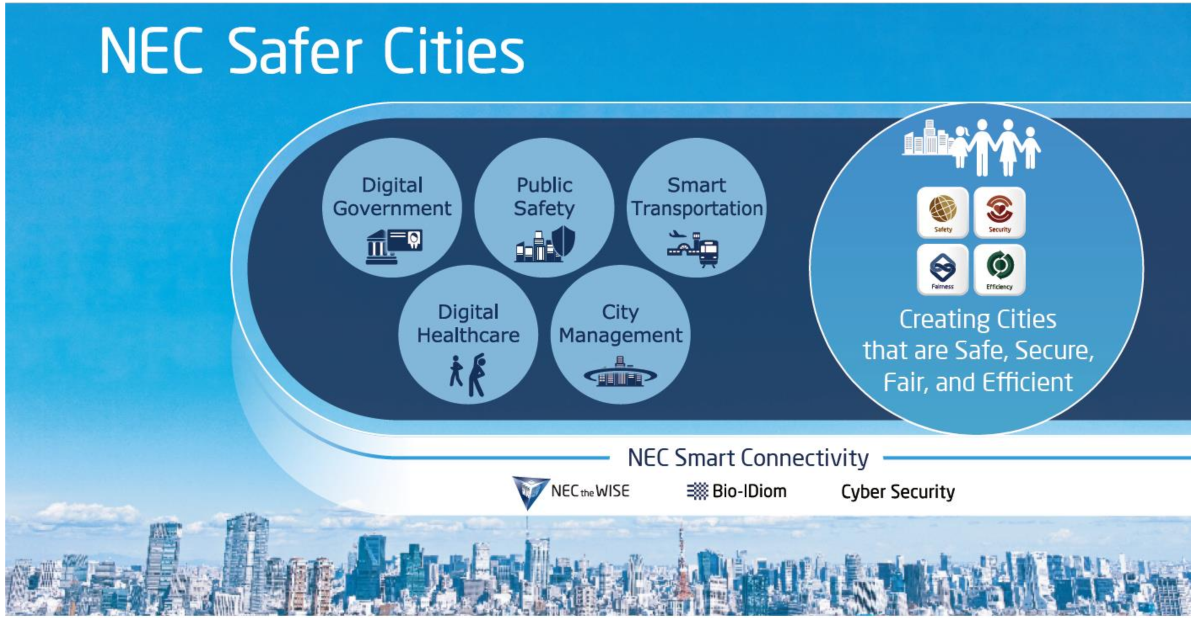
Key Factors for Success

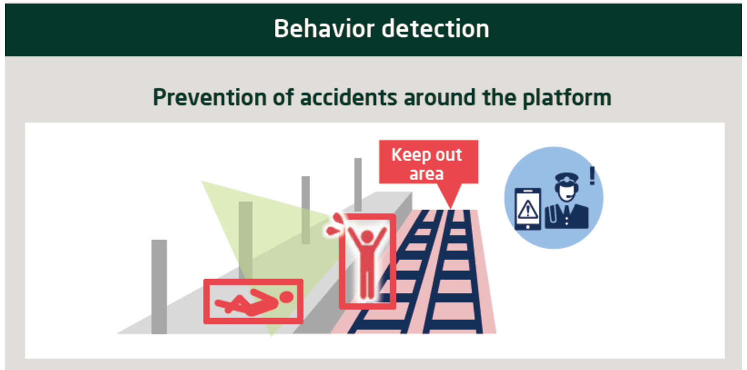
- Transparency
- Strong and open-minded technical teams
- Clear value to customer

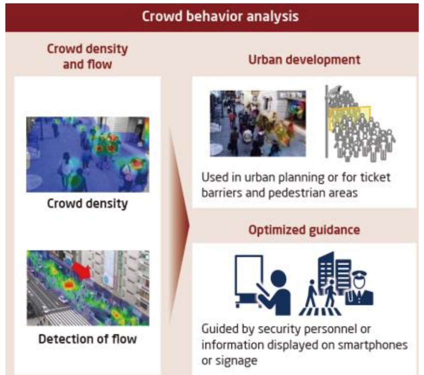
NEC Markets and Positioning

- NEC operates on 5 continents, providing physical safety, failsafe communications and operations solutions
- Biometrics and video analytics – a major product line
- Video-based traffic management – a growth area
- Cost per pixel (camera) has fallen >100X, transmission and storage costs remain high

→ **Edge video processing is required**






Crowd behavior analysis
Crowd density and flow
Crowd density
Detection of flow
Urban development
Used in urban planning or for ticket barriers and pedestrian areas
Optimized guidance
Guided by security personnel or information displayed on smartphones or signage

Edge AI Market Realities – the NEC View

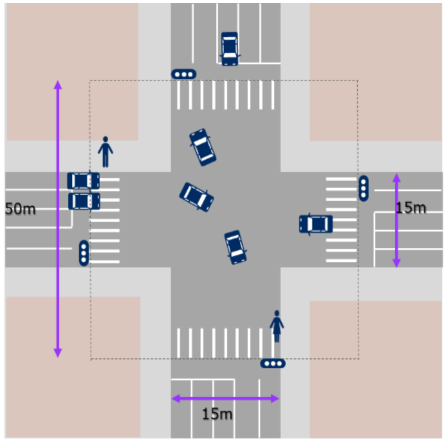
- AI models keep evolving, customers expect “human equivalent” performance
- Making SOTA models run on the edge is not trivial, and most edge AI chips support only a some of the layers, architectures and fixed-point ranges
- Key considerations Include:
  - Can it support the DNN model?
  - Fast CPU interface (bus and drivers)
  - Tight, real-time support for missing features/bugs

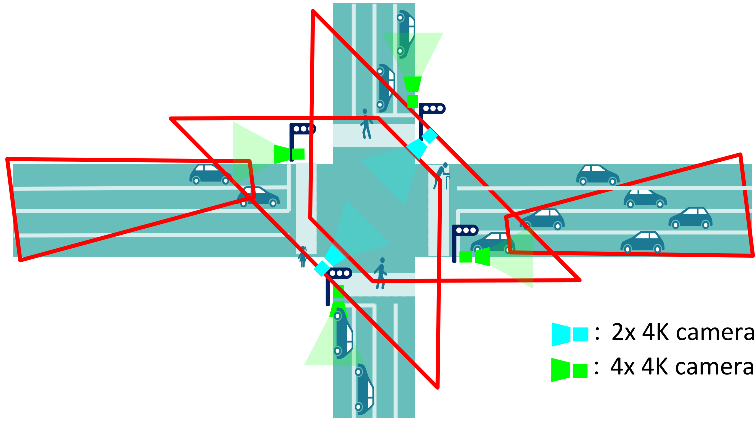
Takeaway from Implemented Projects

- Biometrics – the switch to fixed point requires careful QA and precise layer implementation – customers are intolerant to ‘new’ mistakes
- Video Analytics (traffic) – TOPS/W is nice but modern models need a lot of memory for interim layers – sometimes you need to split and context switch
- Video Analytics (safety) – pixels outpace compute – with more compute, even an existing model can work on higher resolution inputs and yield better results. Raw ‘muscle power’ can deliver the better overall performance!

Project Example

- Traffic video analytics: reliable real time recognition of vehicles, pedestrians
- 4 to 6 cameras per junction, in the future: FHD up to 4K
- SOTA YOLO model for object recognition with additional tracking and pedestrian analytics
- A single TensorPC with 2 Hailo-8 cards can reach 4 FHD streams at 30 FPS each
- 4K stresses CPU, not the Hailo-8 cards – this can be addressed with a stronger CPU





🟦 : 2x 4K cameras inbound

🟩 : 4x 4K cameras outbound

Key Takeaways

- More pixels win – AI works better in higher resolutions. More AI power → better performance
- Fast CPU drivers and data bus are critical
- Conversion of models to the edge is hard, but changing your model is harder – use edge AI that can run your existing models
- Go big or go home – pick a solution that seems “slightly oversize” in compute power and features

Edge AI Platforms

- **Two system designs to support a range of projects**

	Mid-Range	High End
Edge Box	Compulab Fitlet2	Compulab TensorPC
Dimensions X*Y*Z [cm]	11.2 * 8.4 * 3.4	20 * 20 * 3.5
Video Interfaces	1/2	4/8/16
AI Performance	26 TOPS (1 module)	26-104 TOPS (1-4 modules)
Power Consumption (typical)	5W-15W	20W-50W

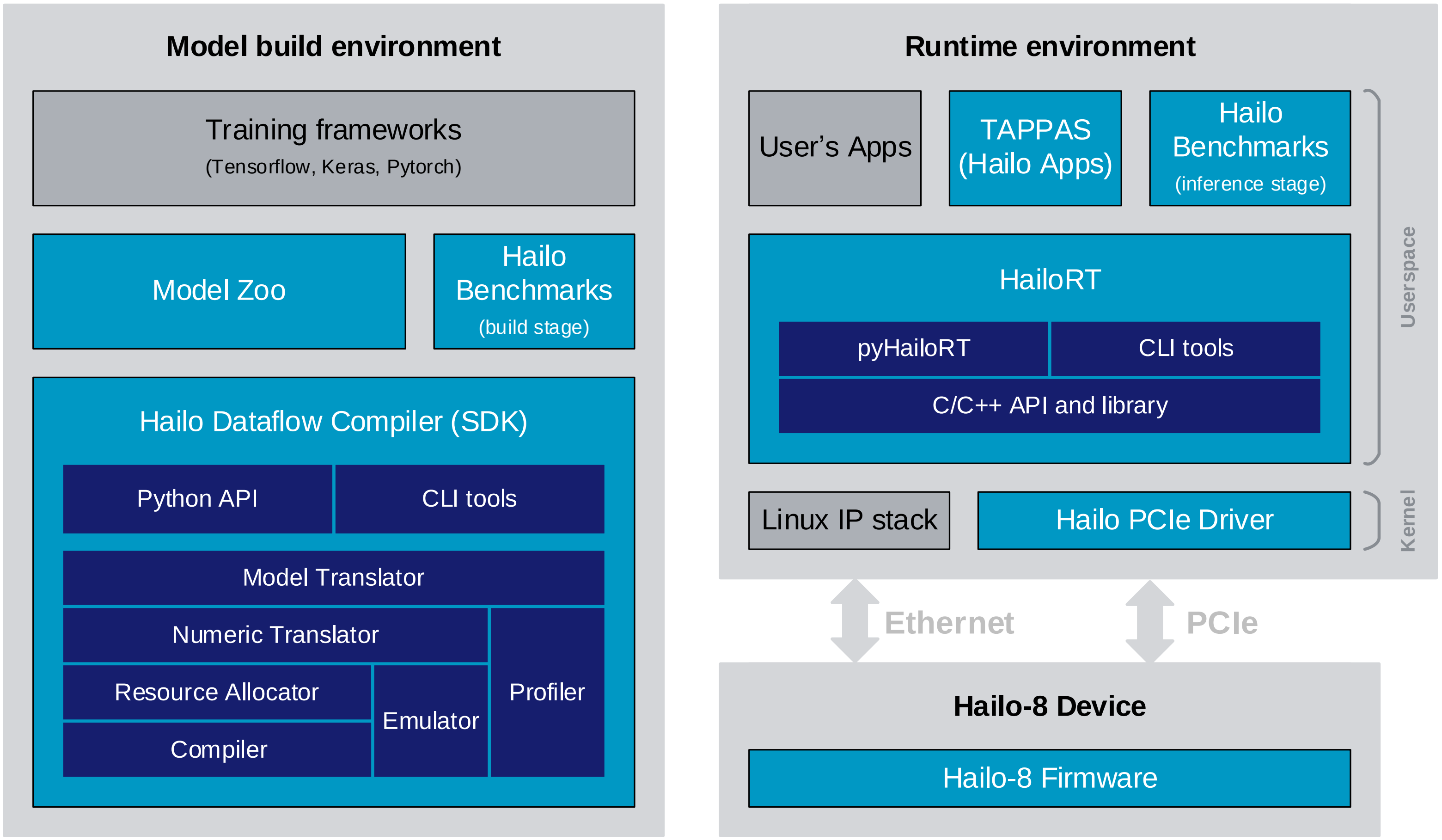
HAILO NEC

AI Acceleration Modules

- PCIe interface
- ARM + x86 support
- Power consumption
  - Ex: ResNet-50, 1200 FPS @ 3.8W
  - Near-linear
  - Low power modes support
- Form factors
  - M.2 (A+E, B+M, M)
  - mPCIe (full size)

Software Stack

Build Flow



Model build environment

Training frameworks (Tensorflow, Keras, Pytorch)

Model Zoo

Hailo Benchmarks (build stage)

Hailo Dataflow Compiler (SDK)

Python API

CLI tools

Model Translator

Numeric Translator

Resource Allocator

Compiler

Emulator

Profiler

Runtime environment

User's Apps

TAPPAS (Hailo Apps)

Hailo Benchmarks (inference stage)

HailoRT

pyHailoRT

CLI tools

C/C++ API and library

Userspace

Linux IP stack

Hailo PCIe Driver

Kernel

Ethernet

PCIe

Hailo-8 Device

Hailo-8 Firmware

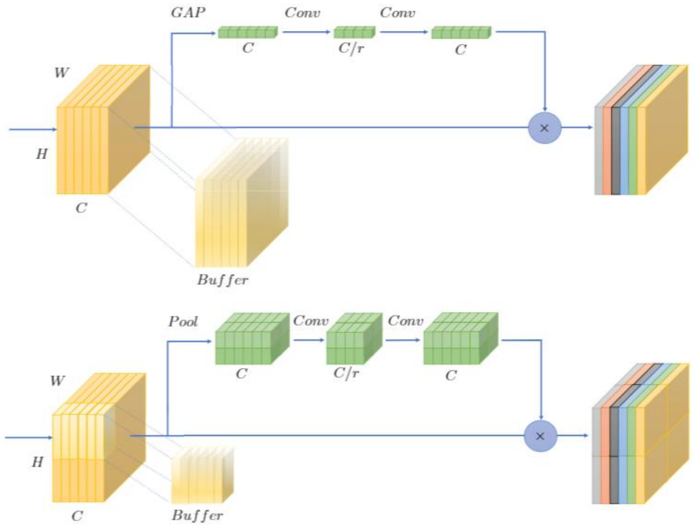
Runtime Flow

🟦 Hailo software component

⬜ Other software component

Collaboration Benefits

- **Early engagement with knowledgeable customer tunes and prioritizes product**
  - → Mixed precision and error metrics
  - → Squeeze-and-Excite (leading to tiled-SE research)
  - → Moving from demos to reference (Hailo TAPPAS)



Collaboration Benefits

- **Understanding customer development flow**

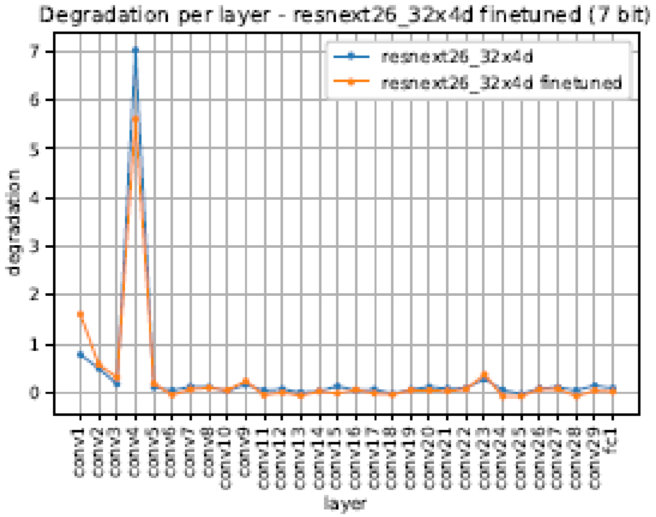
  → Provide tools that bring our ML expertise into customer hands (example – LAT)

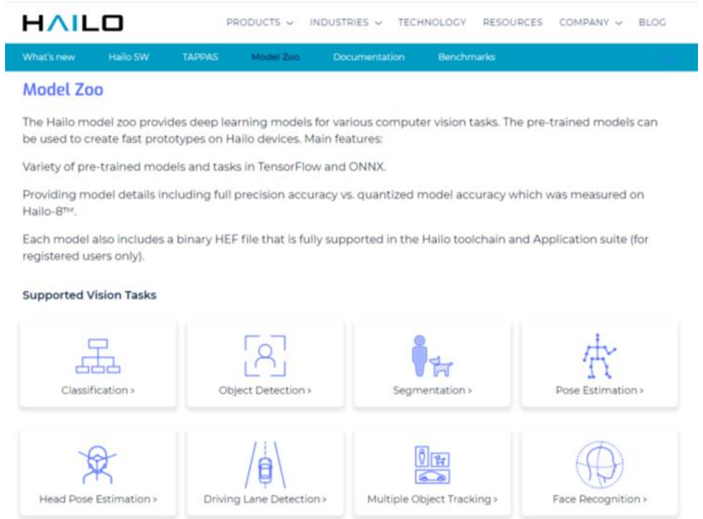
- **Customer needs prioritize highly-optimized models**

  → Inputs to Hailo's model zoo roadmap

- **Roadmap refinement (device N+2 effect)**

- **Compute requirements only increase…**





Model Zoo

The Hailo model zoo provides deep learning models for various computer vision tasks. The pre-trained models can be used to create fast prototypes on Hailo devices. Main features:

Variety of pre-trained models and tasks in TensorFlow and ONNX.

Providing model details including full precision accuracy vs. quantized model accuracy which was measured on Hailo-8™.

Each model also includes a binary HEF file that is fully supported in the Hailo toolchain and Application suite (for registered users only).

Supported Vision Tasks