HAILO

Empowering Intelligence

Hailo-8™

AI Inference Processor

For Edge Devices

Presentation for Haier

September 2022

About Hailo

A leading AI chipmaker for edge devices, founded in 2017

1st generation in MP

Headquartered in Israel with offices in USA, Germany, Japan, China, Korea, Taiwan

Patented structure-defined dataflow architecture

190+ employees with extensive experience from leading tech companies

Total $224M funding including Strategic Investors

NEC & ABB

A growing worldwide partner ecosystem

CES 2020 Innovation Awards Honoree

EU Horizon 2020 Recipient

AI Semi Cool Vendor by Gartner

Best Edge AI Processor of 2021

ISO 9001 Certified

ISO 14001

Hailo-8™ Highlights

The World's Most Powerful and Efficient Edge AI Processor

High Performance
26 TOPS
Efficient AI architecture

Comprehensive SW Tools
Mature dataflow compiler
Efficient RT library

Industrial & Automotive Grades
Industrial: -40° up to 85°
Automotive: -40° up to 105°

Power Efficiency
Typical Power
Consumption: 2.5W

Single Chip Solution
No External DRAM
required

Scalable & Flexible
Multi-streams
Multi-model
Multi-chip

Intelligence Becomes a Necessity

Hailo's **powerful** and **scalable** AI technology enables new capabilities in various markets

Automotive

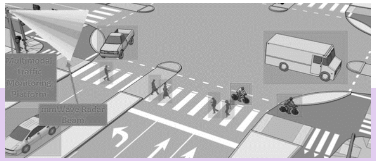
Autonomous
Vehicles, ADAS

Smart City

Public safety & security

Smart Home

Security,
Assisted Living



ITS (Intelligent Transportation System)

Traffic control, Tolling,
Law enforcement

Smart Retail

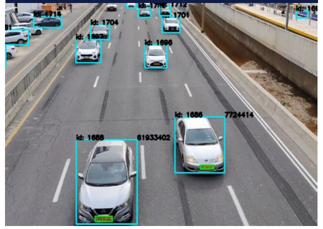
Cashierless Store, Inventory Management

Industry 4.0

Factory Automation

Deep Learning at the Edge with Hailo-8: Use Cases & Target Platforms

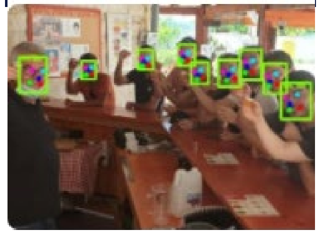
Traffic Management & Tolling



Traffic Monitoring

Intersection Safety

Public Health Monitoring



Autonomous Delivery

Quality Inspection

Factory Safety

Retail Automated Checkout

Smart Building

Advanced Driver Assistance (ADAS)

Front Facing Perception

Access Control

Intelligent Cameras

Intelligent NVR

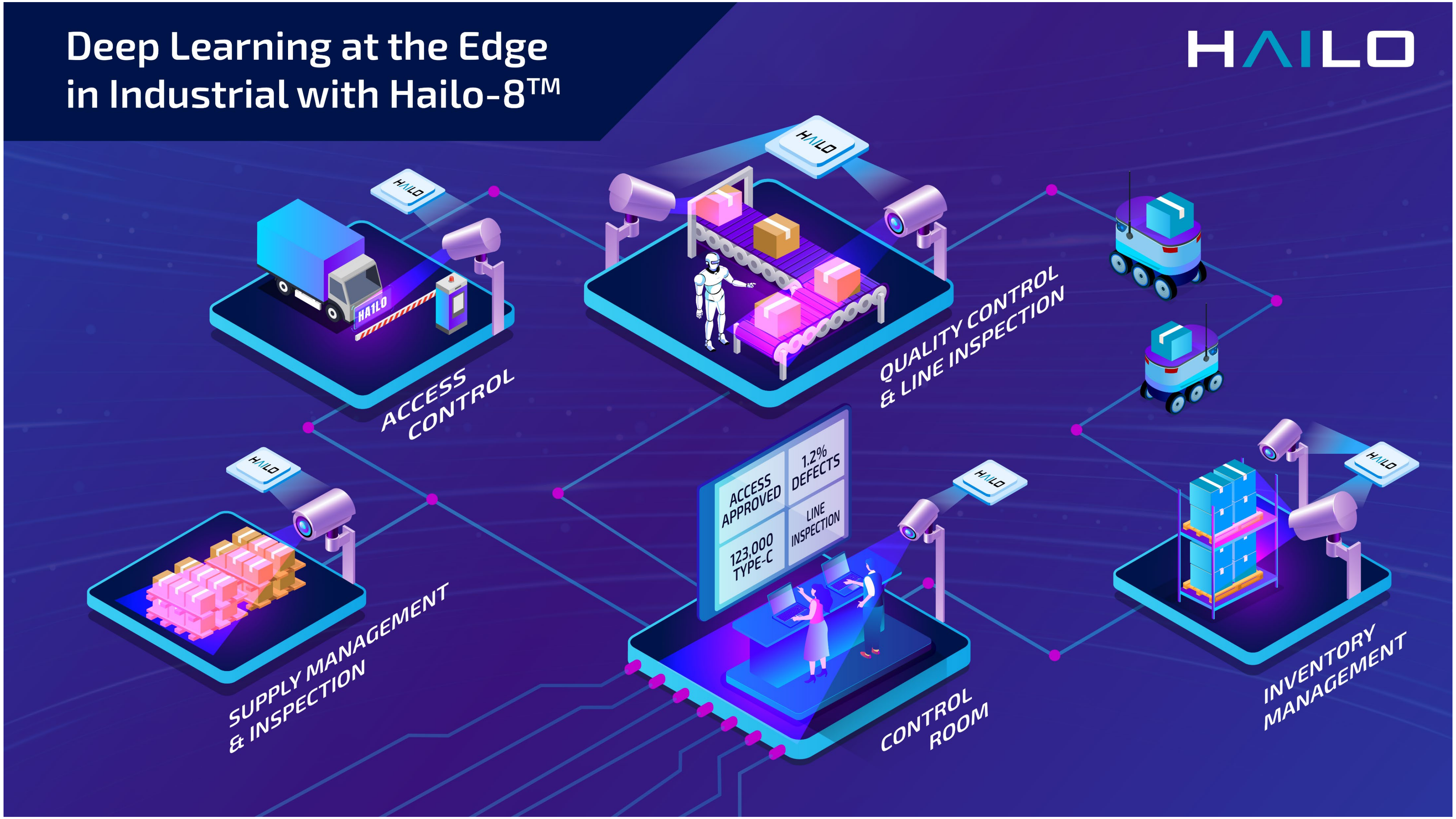
Industrial Gateways & PCs

In-Vehicle Computer

ADAS ECU

Autonomous Mobile Robot

Examples of Deep Learning at the Edge in Smart Factories


Deep Learning at the Edge
in Industrial with Hailo-8™
HAILO
ACCESS CONTROL
QUALITY CONTROL & LINE INSPECTION
SUPPLY MANAGEMENT & INSPECTION
ACCESS APPROVED
1.2% DEFECTS
123,000 TYPE-C
LINE INSPECTION
CONTROL ROOM
INVENTORY MANAGEMENT

Smart Factory Key Use-cases Supported by Hailo Solutions

Manufacturing Management

- High accuracy quality control and line inspection
  - Real-time counting, defect detection and product analysis
- Robot control
  - Material handling, assembly and processing
- Predictive maintenance
  - Anomaly detection in real-time
- Efficiency and bottlenecks analytics

Building Management System (BMS)

- Logistics management and automated warehouses
  - Inventory tracking and monitoring

Autonomous Mobile Robot (AMR)

- Autonomous navigation
  - Object detection and classification, route planning, anomaly detection, complex pattern recognition

Safety

- Detects hazards, obstacles and dangerous machine movements
- People detection , identification, counting, tracking, physical conflict, face mask detection, safety rules violations
- Proof of evidence in case of incidents or accidents
- Enables prompt incidents handling

Security

- Access Control
  - Automatic people access control by face recognition
  - Automatic vehicles access control by fast real-time detection
- Perimeter protection
  - Detects person/vehicle entrance to restricted areas
  - Detects person/vehicle crosses a predefined line
- Generate extensive metadata, enables:
  - Analytics search by event type or classification
  - Appearance search for location of a person or object

Hailo-8™ Key Values for Smart Factory

Comprehensive Solution

- Supports multiple use-cases (quality inspection, Robot control, Security, …) simultaneously, in real-time
- Scalable solution up to 312 TOPS

High Accuracy Detection

- Real-time AI processing with high FPS for detection product anomalies, hazards, and people / objects in highest accuracy
- Best performance by utilizing state-of-the art Deep Learning models

High Reliability and Low Maintenance Cost

- Low power consumption ~2.5W
- Extended temperature range support of -40°C to 85°C
- Fanless device ➔ No need for active colling solution

Cost Effective Solution

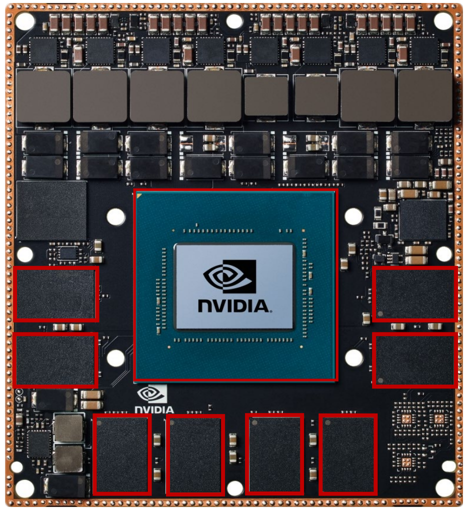
- Cost-effective modules and cards
- A single Hailo-8™ device can process multiple video streams in real-time at the edge
- Enable small footprint and fanless design
- Lower dependency on network bandwidth and cloud services usage for AI analytics
- Extend product lifetime with introduction of new features

Low Development Efforts and short TTM

- Simpler, efficient and fast integration
- Wide availability of production ready solutions w/Hailo-8™

Unprecedented AI Performance

NVIDIA AGX Xavier



General Purpose GPU
+ External Memory

Hailo-8™

Dedicated AI Chip
No External Memory

ResNet-50 Benchmark

Device	Total Power [Watt]	Total Power Efficiency [TOPS/W]
Hailo-8™	1.7	2.8
Nvidia Xavier AGX	32	0.14

Conditions:

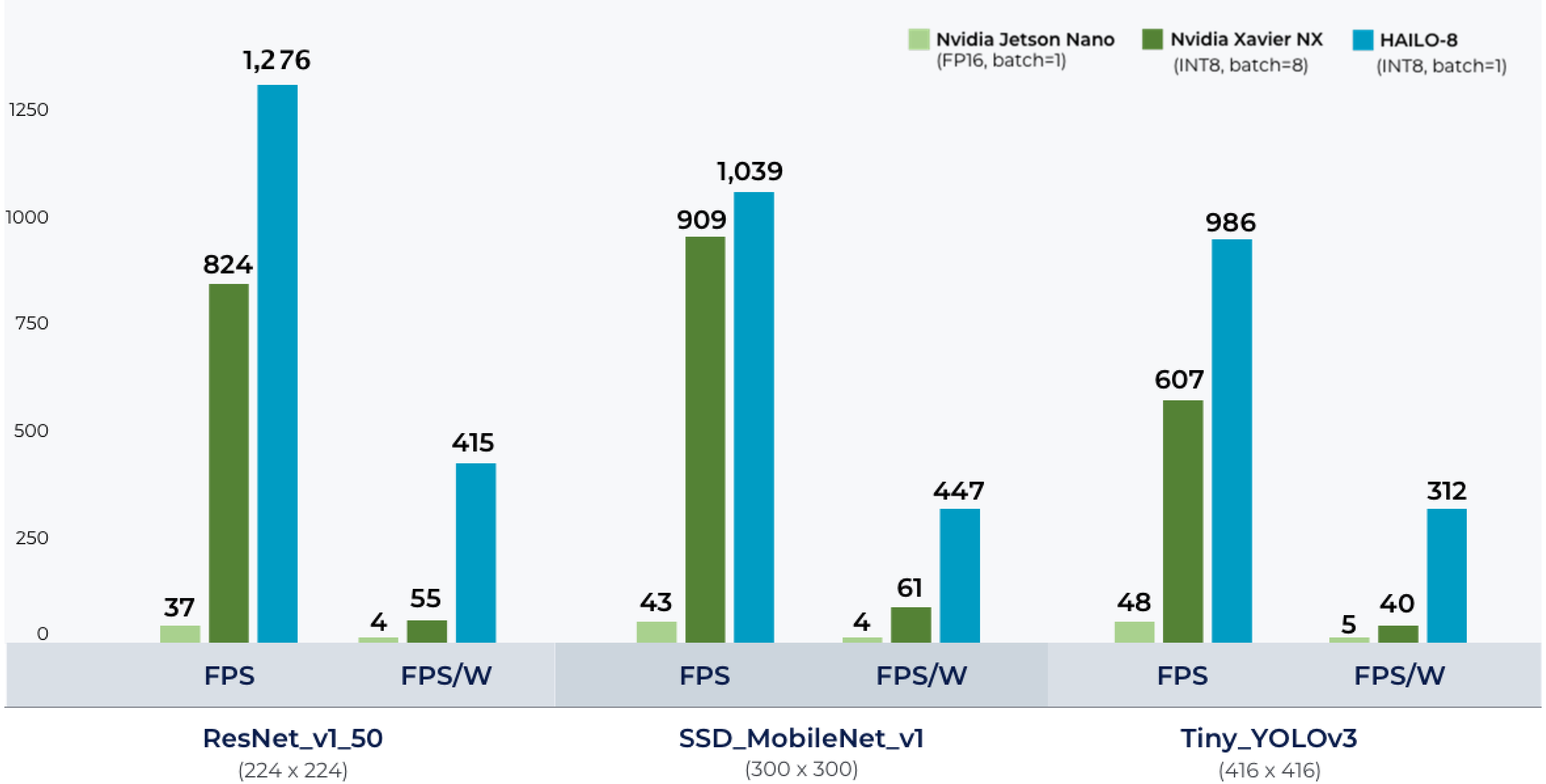
- TOPS (8-bit): Xavier 32 TOPS, Hailo-8 26 TOPS
- 224x224 image resolution feed @ 656 FPS
- 8-bit precision
- Batch size = 1

X15 Better
Area Efficiency

X20 Better
Power Efficiency

Unprecedented Performance at the Edge

Hailo-8 offers higher performance and as much as x8 the power efficiency of Nvidia's best edge device



Remarks

- SDK version 3.9.0 (June 2021), measured at room temp on a single Hailo-8 device through PCIe interface on a Hailo EVB. System host: Intel® Core™ i5-9400 CPU @ 2.90GHz)
- **Xavier NX results are using batch=8** (while Hailo-8 and Jetson Nano are using batch=1) and that **Jetson Nano is limited to FP16** (while Hailo-8 and Xavier NX are INT8). Nvidia results for batch=1 and INT8, respectively, are expected to be lower.
- FPS & power figures for Nvidia Jetson Nano and Xavier NX are sourced from the Nvidia website and Github repo, retrieved 12/07/21

Hailo-8™ Unprecedented AI Performance and Power Efficiency

Intel Myriad X

87 FPS

35 FPS/W

Google Edge TPU

385 FPS

275 FPS/W

Hailo-8™

2,613 FPS

1,267 FPS/W

The Hailo-8™ M.2 AI Acceleration module is the highest performing AI module on the market

Hailo-8™ delivers better throughput: x30 better than Myriad X and x6 than Edge TPU

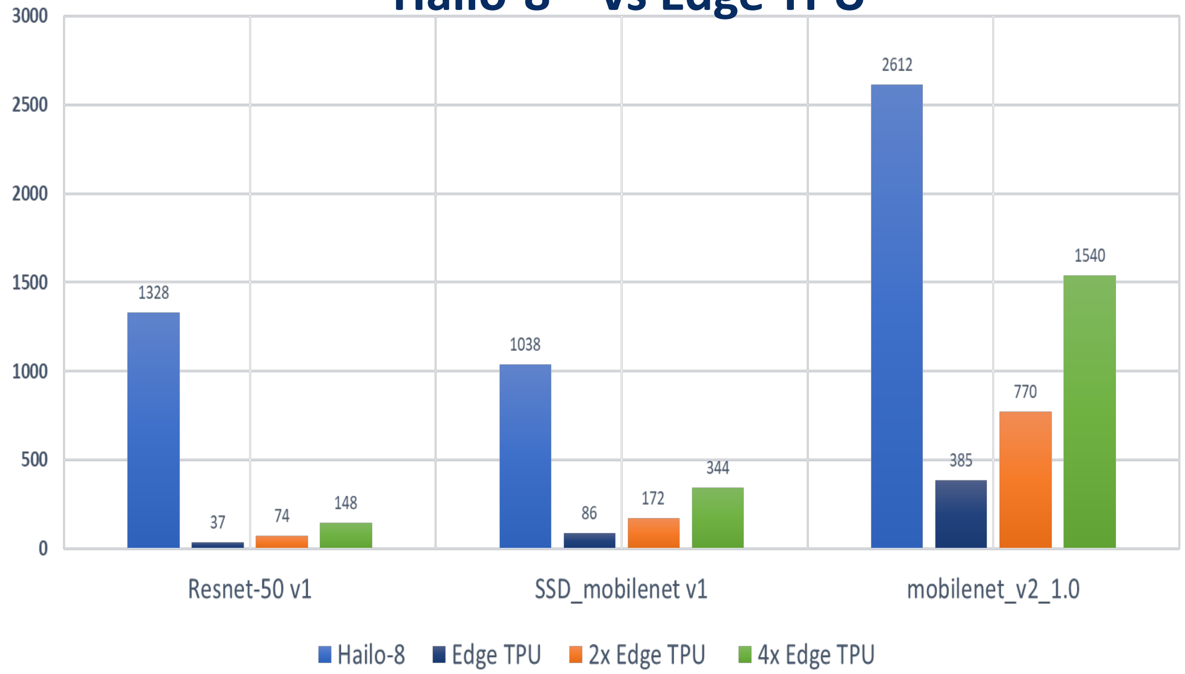
Power Efficiency is x30 better than Myriad X and x4 than Edge TPU

Provides the scalability to run advanced video analytics NN models in high-resolution & high-frame rate

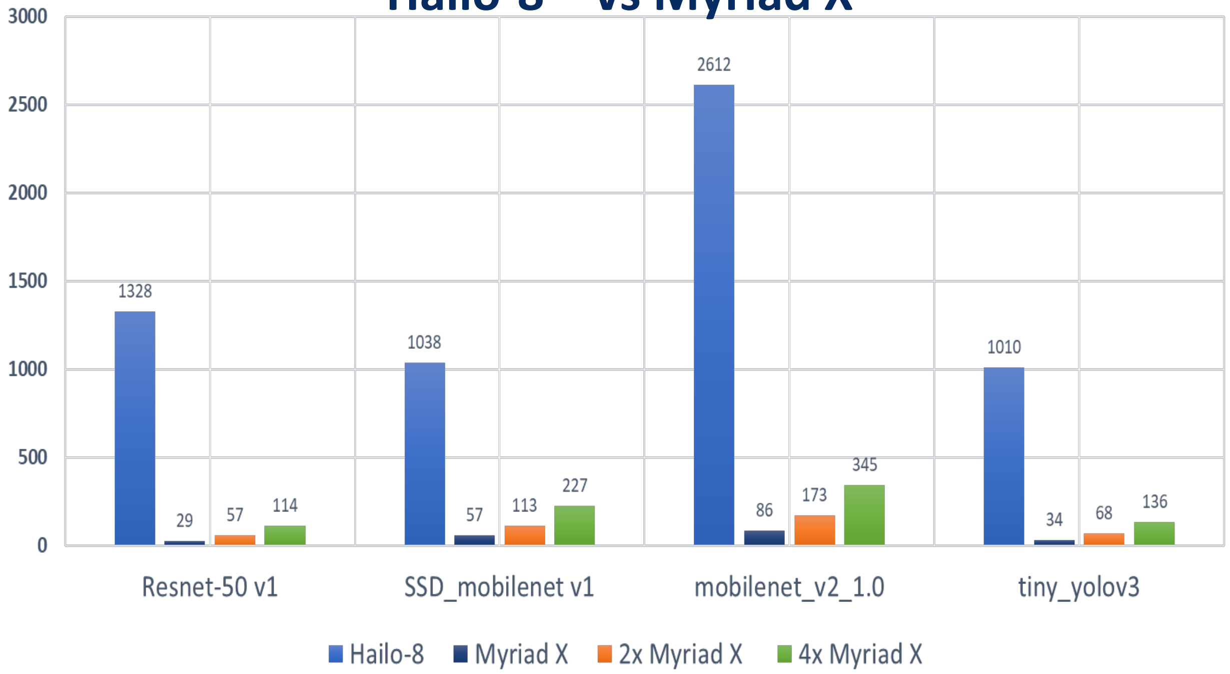
** Based on published performance (intel, Google) on the MobileNet_v2_1.0 BM, input resolution: 224x224*

Hailo-8™ Unprecedented Performance at the Edge

Hailo-8™ vs Edge TPU



Hailo-8™ vs Myriad X



Hailo-8™ **outperforms Edge TPU** by as much as **x10**, and by **x2** vs 4 Edge TPU devices

Hailo-8™ **outperforms Myriad X** by as much as **x26**, and by **x6** vs 4 Myriad X devices

- **Hailo-8 figures are based on SDK Q1 2022 version, measured at room temperature on Hailo-8 device through PCIe interface on a Hailo-8 evaluation board (system host: Intel Core i5-9400 CPU @ 2.90GHz)**
- **Edge TPU figures are for batch=1 and INT8, while Myriad X is batch=1 and FP16**
- **Intel Myriad X figures sourced from: https://docs.openvinotoolkit.org/latest/openvino_docs_performance_benchmarks_openvino.html , retrieved April 2022**
- **Google Edge TPU figures sourced from here and here retrieved April 2022; FPS is converted from latency in ms (1 divided by ms/1000)**

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

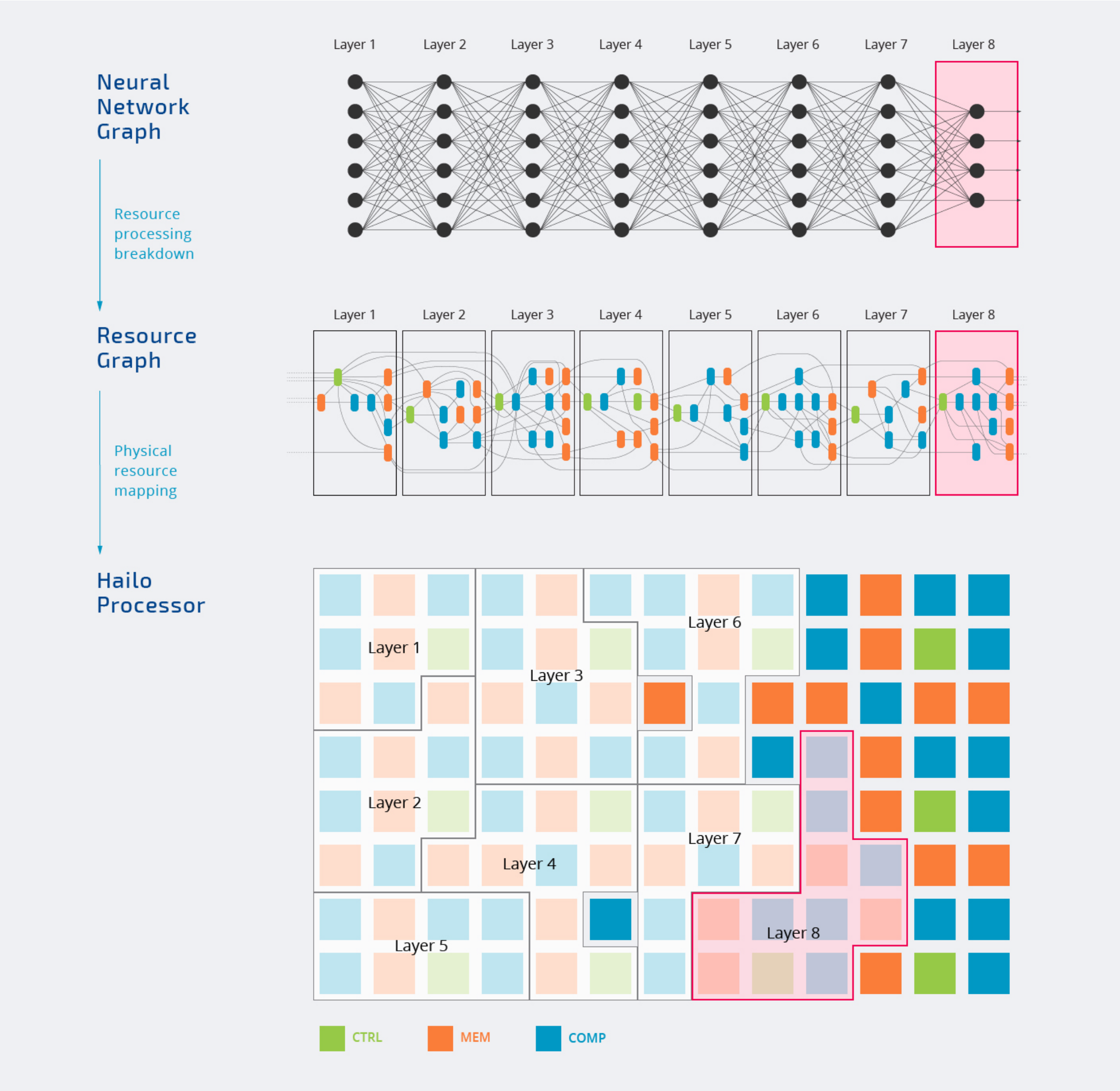
Hailo-8™ Measured Benchmarks

Model	Type	Input Resolution	FPS	Total Power [W]	FPS/W
ResNet-v1-50	Classification	224x224	1,328	3.1	428
MobileNet-v2-1.0	Classification	224x224	2,613	2.1	1,267
MobileNet_v3[4]	Classification	224x224	3,468	1.8	1,878
RegNetx_800mf	Classification	224x224	2,447	2.0	1,232
EfficientNet-M	Classification	240x240	889	3.2	278
SSD-MobileNet-v1	Object Detection	300x300	1,038	2.3	452
Tiny-YOLOv3	Object Detection	416x416	1,010	3.2	315
YOLOv3[5]	Object Detection	608x608	60	4.2	14
YOLOv4[5]	Object Detection	512x512	72	3.1	23
YOLOv5m	Object Detection	640x640	218	4.3	50

Notes:

1. **Based on Dataflow compiler version 3.14.0 (Q1 2022)**
2. **Measurements are done in room temperature through PCIe interface on Hailo-8 evaluation board**
3. **System host: Intel(R) Core(TM) i5-9400 CPU @ 2.90GHz**
4. **MobileNet-V3 - the benchmarked model flavor is Mobilenet V3 Large Minimalistic**
5. **Performance figures are gives for processing 8 simultaneous streams**

Structure Defined Dataflow Architecture



Neural Network Graph

Resource processing breakdown

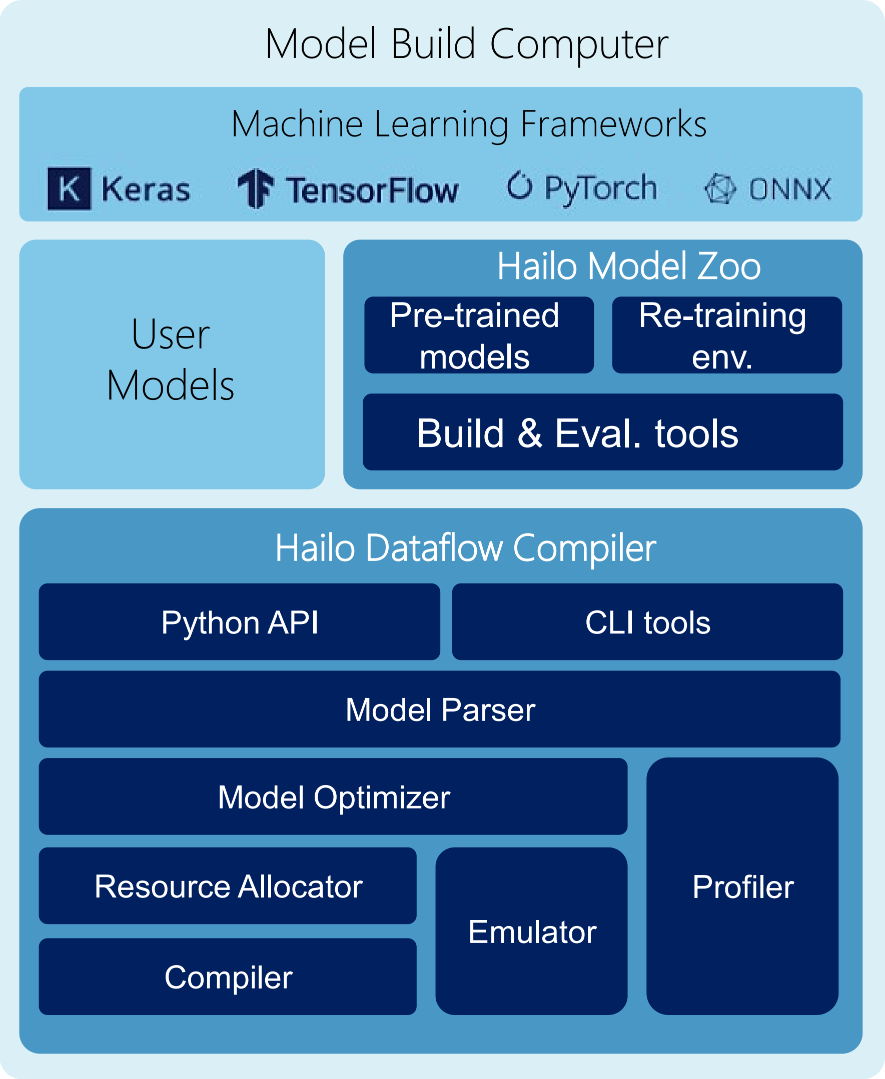
Resource Graph

Physical resource mapping

Hailo Processor

Hailo Software Toolchain and Developer Tools

Model Build Environment



Model Build Computer

Machine Learning Frameworks

Keras TensorFlow PyTorch ONNX

User Models

Hailo Model Zoo

Pre-trained models

Re-training env.

Build & Eval. tools

Hailo Dataflow Compiler

Python API

CLI tools

Model Parser

Model Optimizer

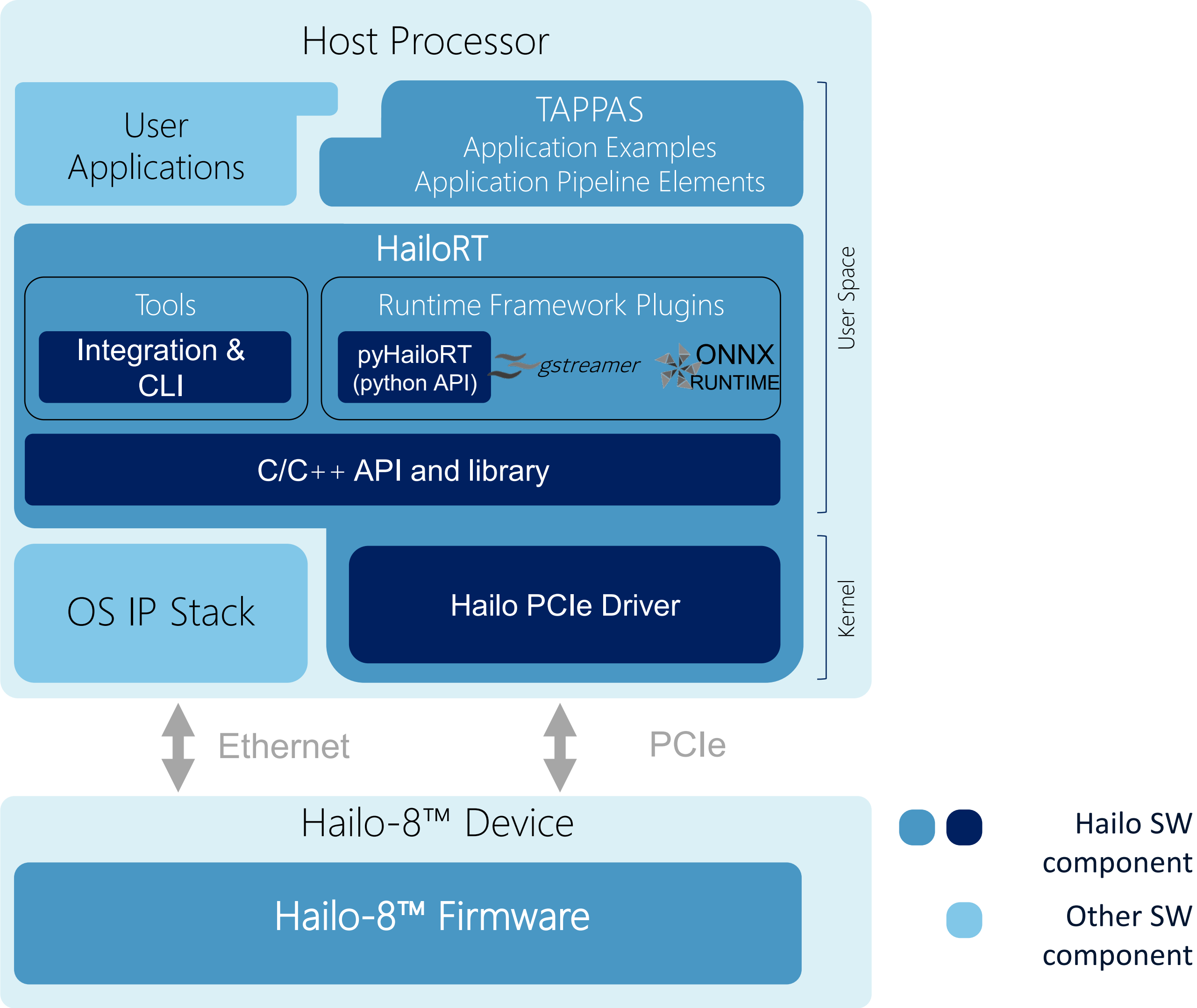
Resource Allocator

Compiler

Emulator

Profiler

Runtime Environment



Host Processor

User Applications

TAPPAS

Application Examples

Application Pipeline Elements

HailoRT

Tools

Integration & CLI

Runtime Framework Plugins

pyHailoRT (python API)

gstreamer

ONNX RUNTIME

C/C++ API and library

OS IP Stack

Hailo PCIe Driver

User Space

Kernel

Ethernet

PCIe

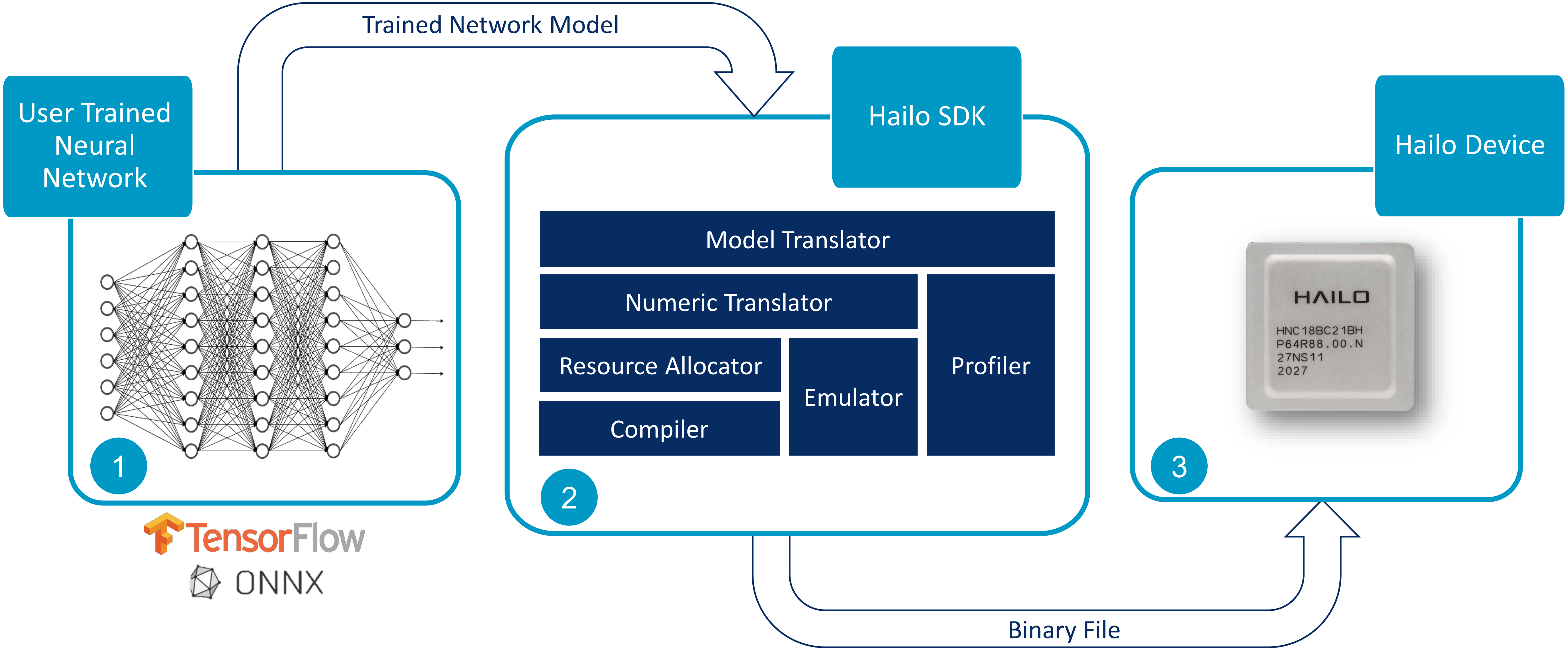
Hailo-8™ Device

Hailo-8™ Firmware

Hailo SW component

Other SW component

Hailo Dataflow Compiler



Trained Network Model

User Trained Neural Network

1

TensorFlow

ONNX

Hailo SDK

Model Translator

Numeric Translator

Resource Allocator

Compiler

Emulator

Profiler

2

Hailo Device

3

Binary File

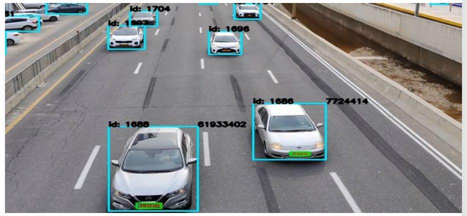
Hailo AI Template APPlications And Solutions (TAPPAS)

Suite of high-performance, pre-trained template AI tasks and applications elements with production-grade pipeline

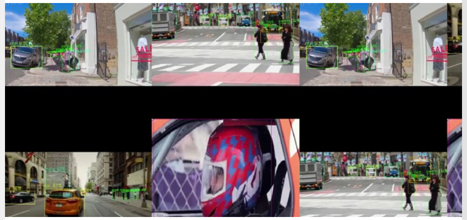
- Suitable for variety of categories and industries
- Useful for demos and can be used as reference designs
  - Accelerate time to market by reducing development and deployment effort
  - Model(s) can be easily replaced

https://hailo.ai/developer-zone/tappas-apps-toolkit/

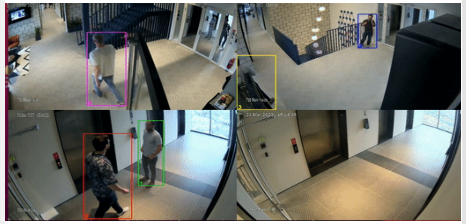
License Plate Recognition



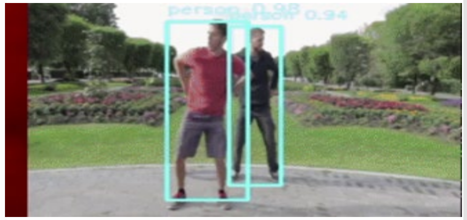
Multi Streams Multi Device Object Detection



Multi Person Multi Camera Tracking



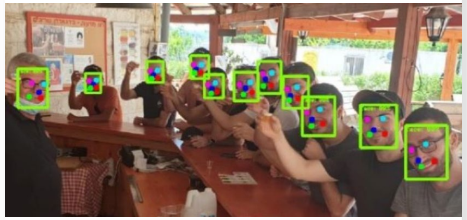
Object Detection and Depth Estimation



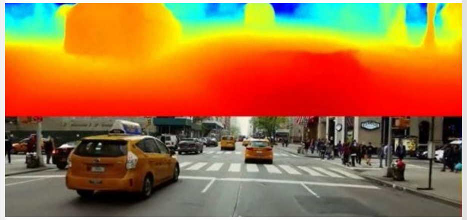
Semantic Segmentation

Pose Estimation

Facial Detection & Recognition

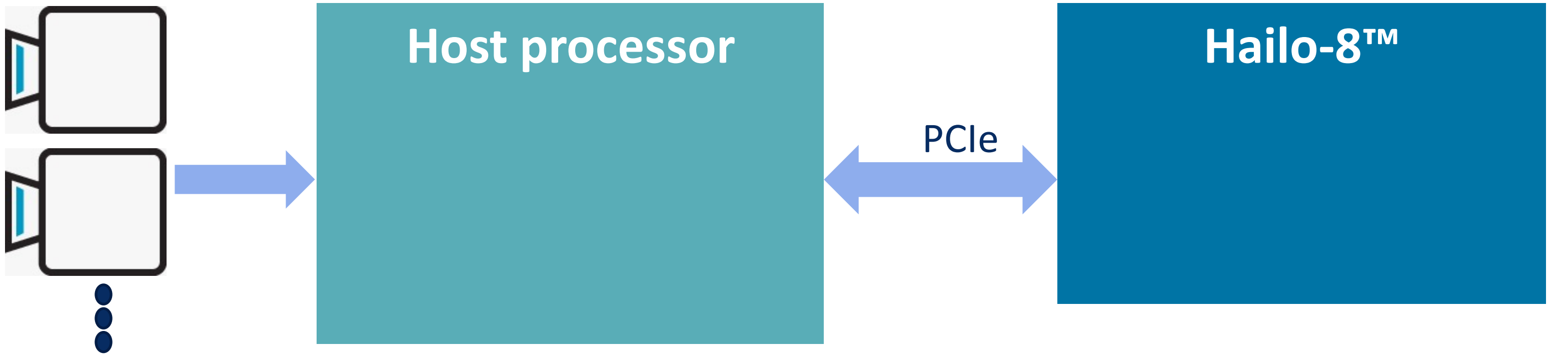


Depth Estimation



Instance Segmentation

Hailo-8™ System Usage



Host processors support

- Intel X86 - Celeron, i3, i5, i7, Atom, Xeon, …
- AMD X86
- ARM based
  - i.MX8
  - Layerscape (LX2160)
  - S32G
  - Raspberry Pi
  - FPGA SoC – Xilinx Zynq
  - Renesas R-CAR V3H/V4H
  - SocioNext SC2A11
  - …

- **Flexibility & Scalability**
  - **Performance scalability** (1x to 12x Hailo-8 → 26 to 312 TOPS)
  - **Host processor type** (X86 & ARM)
  - **Interface w/Host** (PCIe / Ethernet)

Hailo-8™ Products

Hailo-8™ AI Processor

- 26 TOPS
- Industry-leading power efficiency
- 17 x 17 FCBGA

Hailo-8™ M.2 AI Acceleration Module

- PCIe Interface
- M.2 form factor
  - NGFF M.2 Key M 2242/2260/2280
  - NGFF M.2 Key B+M 2242/2260/2280
  - NGFF M.2 Key A+E 2230
- Extended temperature support: -40° up to 85°

M key
4 lanes

B+M key
2 lanes

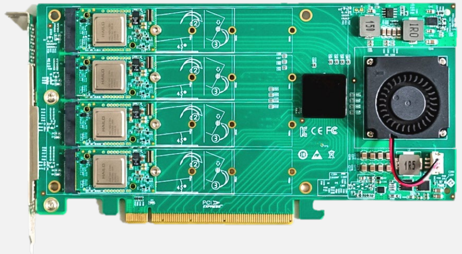
A+E key
2 lanes

Hailo-8™ Mini PCIe AI Acceleration Module

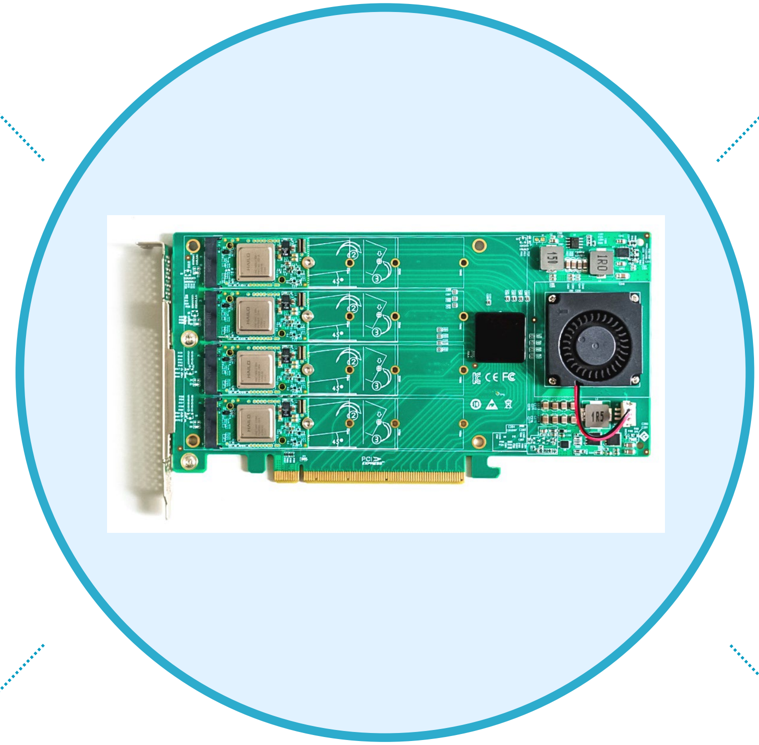
- PCIe Interface
- mPCIe form factor 3050
- Extended temperature support: -40° up to 85°

Hailo-8™ Century Evaluation Platform

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



Hailo-8 Scalability – Hailo-8 Century Evaluation Platform



High Performance
104 TOPS

No. of Devices
4

Passive (fanless) cooling

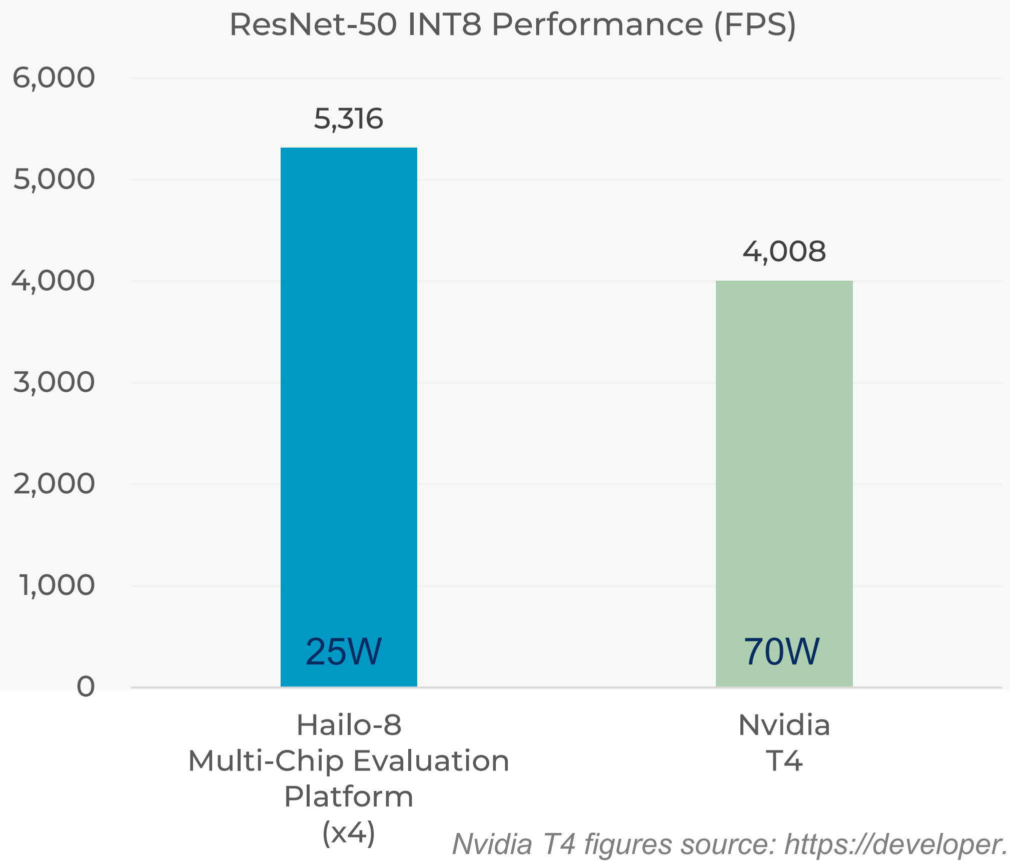
Workload example
Yolo V3 at 400 fps

Low Power
<15 W

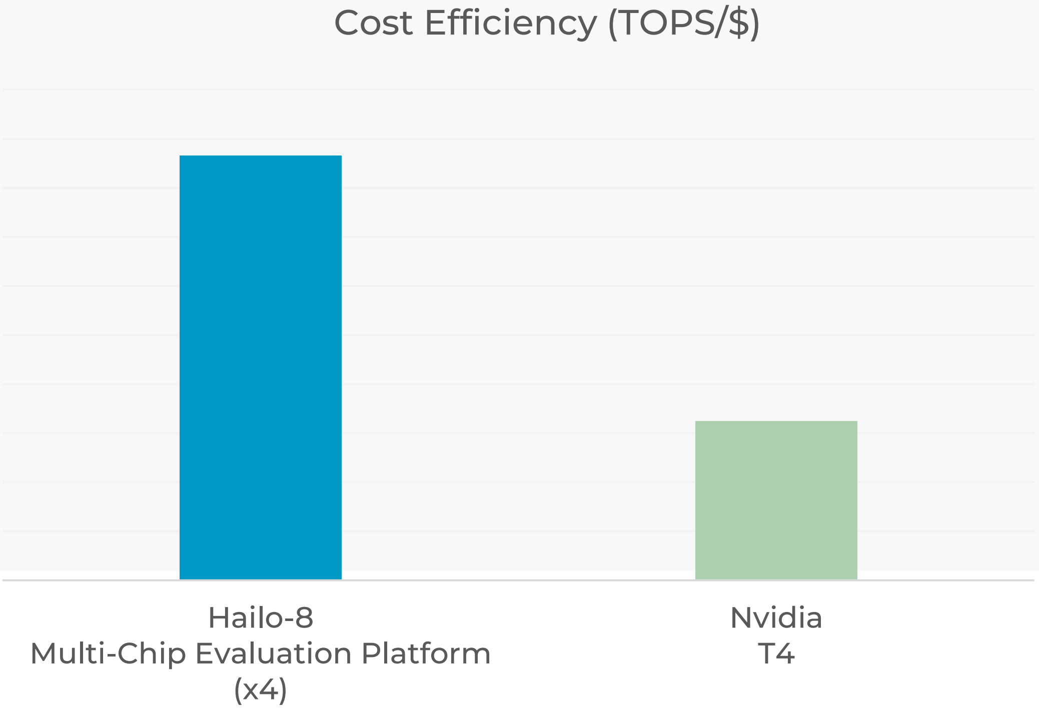
High-efficiency for multi-camera
< 1 W per camera

Hailo-8 Multi-Chip Evaluation Platform – **Cost and Power Efficiency**

Get more performance at 1/3 of the power



Get **X3** more performance for every $ spent



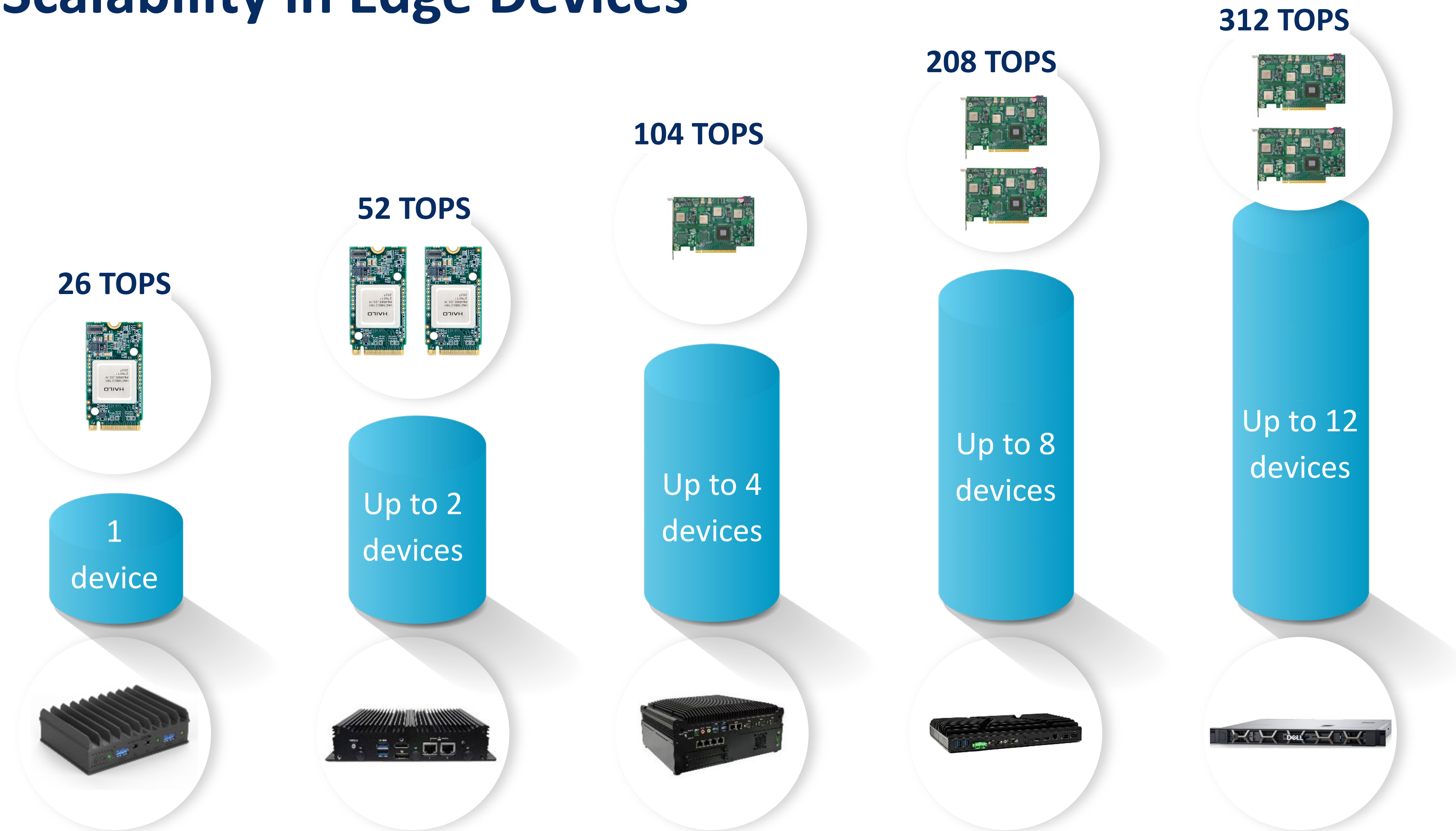
- *Nvidia T4 figures source: https://developer.nvidia.com/deep-learning-performance-training-inference*
- *Based on maximum performance claims and market pricing*

Hailo-8™ Scalability in Edge Devices

x1 to **x12**
devices

26 to **312 TOPS**
of AI processing

Passively
cooled;
Highly **Scalable**;
Multiple
vendors



26 TOPS

1 device

52 TOPS

Up to 2 devices

104 TOPS

Up to 4 devices

208 TOPS

Up to 8 devices

312 TOPS

Up to 12 devices

HAILO

Empowering Intelligence

THANK YOU!

contact@hailo.ai

https://www.hailo.ai