CEVA

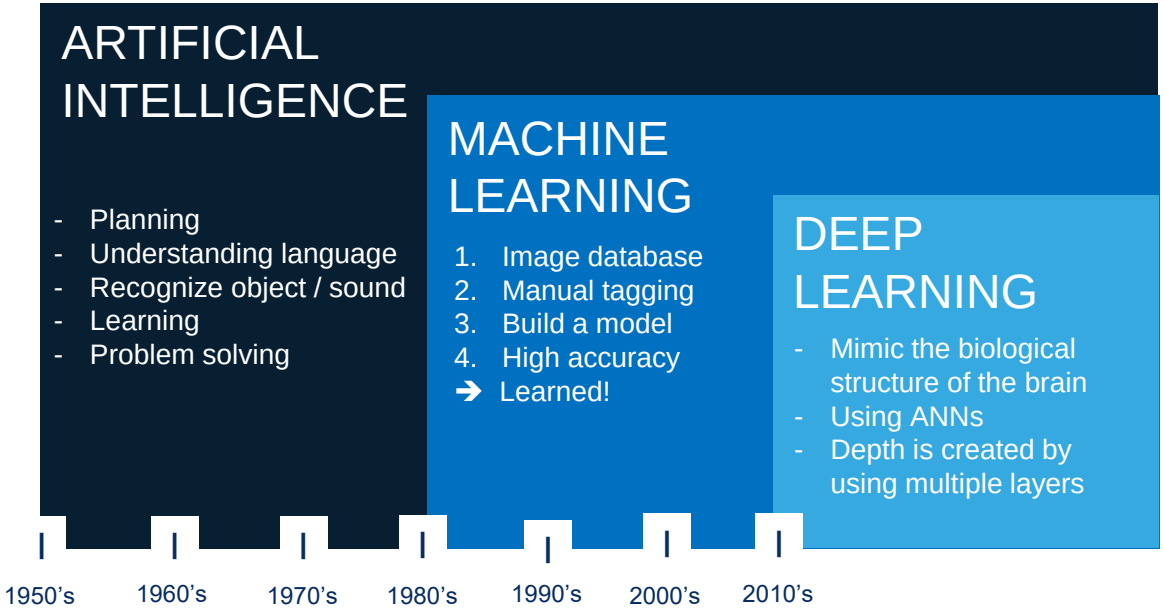
CEVA TECHNOLOGY
SYMPOSIUM SERIES

12th YEAR

Specialized AI Processor for Deep-Learning Inference at the Edge

Liran Bar, Director of Product Marketing, CEVA

What's the difference between AI, ML, and DL?

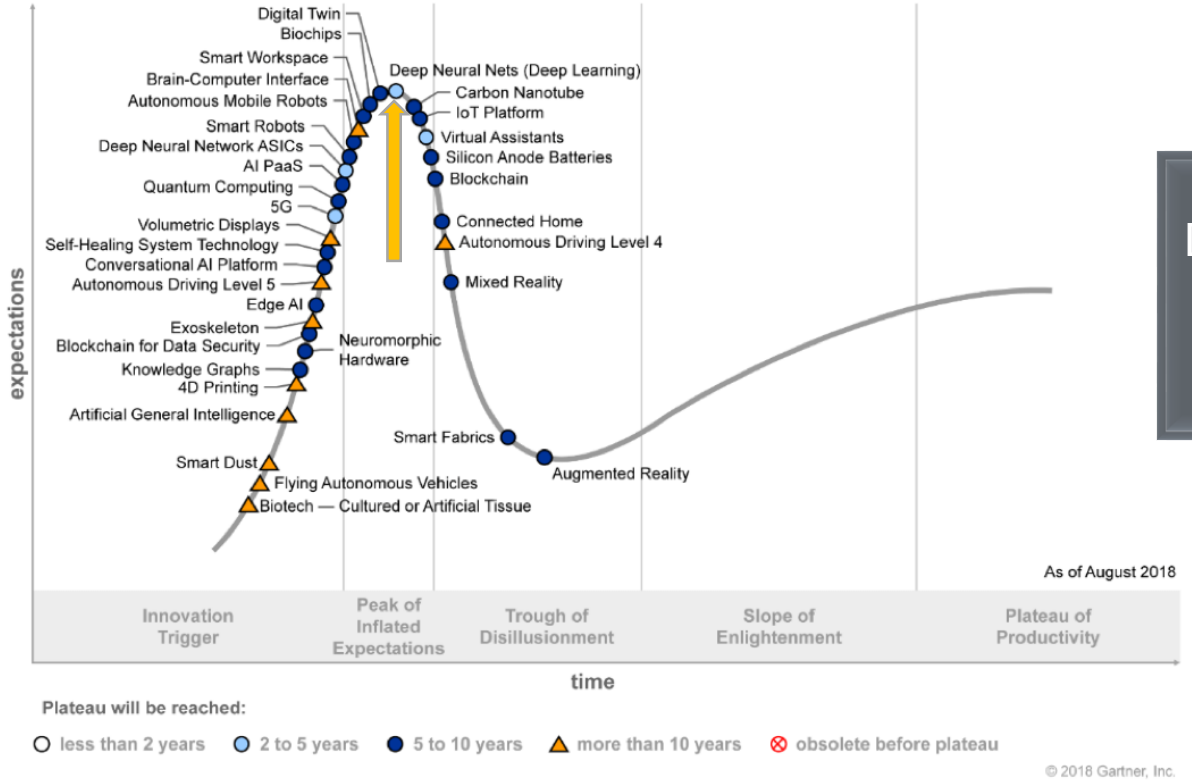


AI: Intelligence demonstrated by machines rather than humans or animals

ML: Giving computers the skills to learn without explicit programming

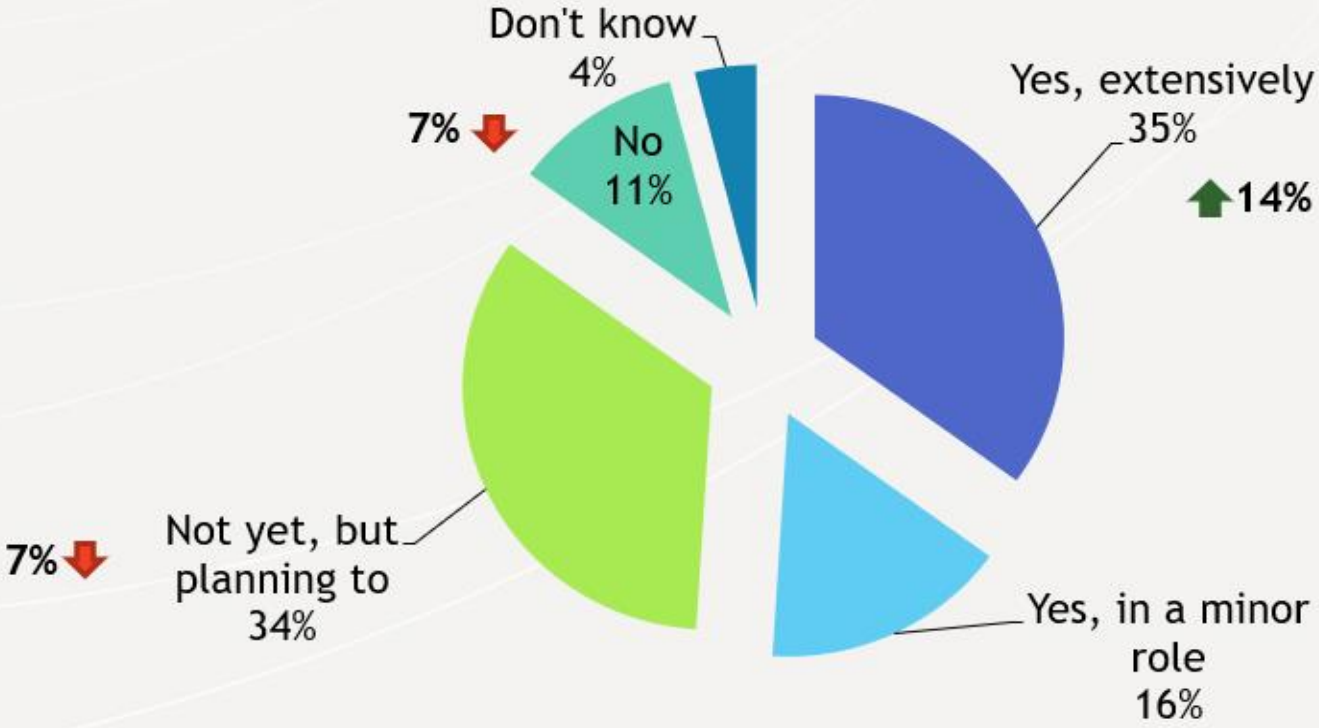
DL: Is an ML subset, examining algorithms that learn and improve on their own

Hype Cycle for Emerging Technologies



Deep Learning is still at the peak

Use of Neural Networks to Perform Computer Vision Functions

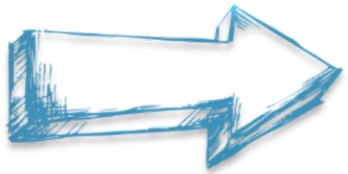


AI Market Trends

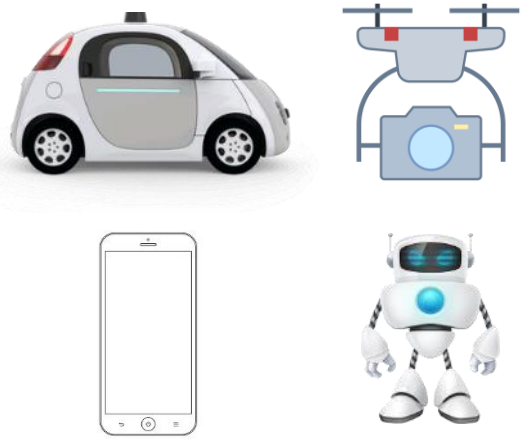
Intelligence (keep) Moving to the Edge

Cloud AI

Intelligence & Analytics



Edge Devices

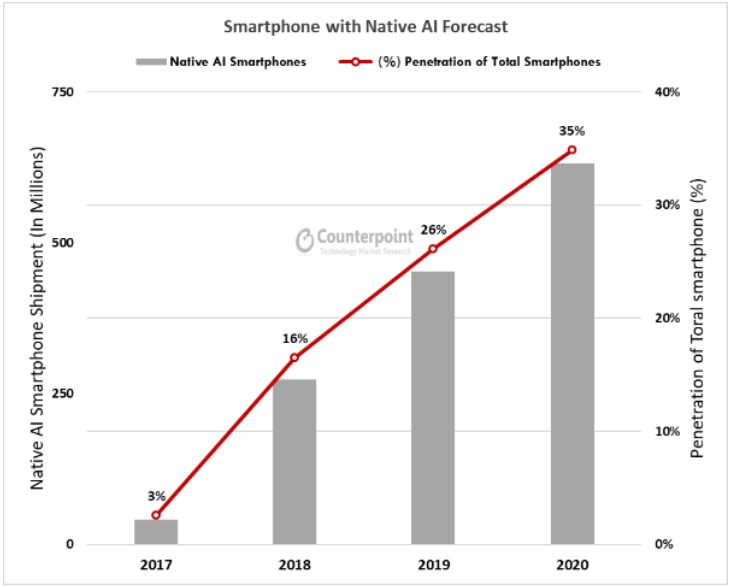


Key Benefits

Low Latency | Low Power | Low Cost | High Privacy | High Reliability

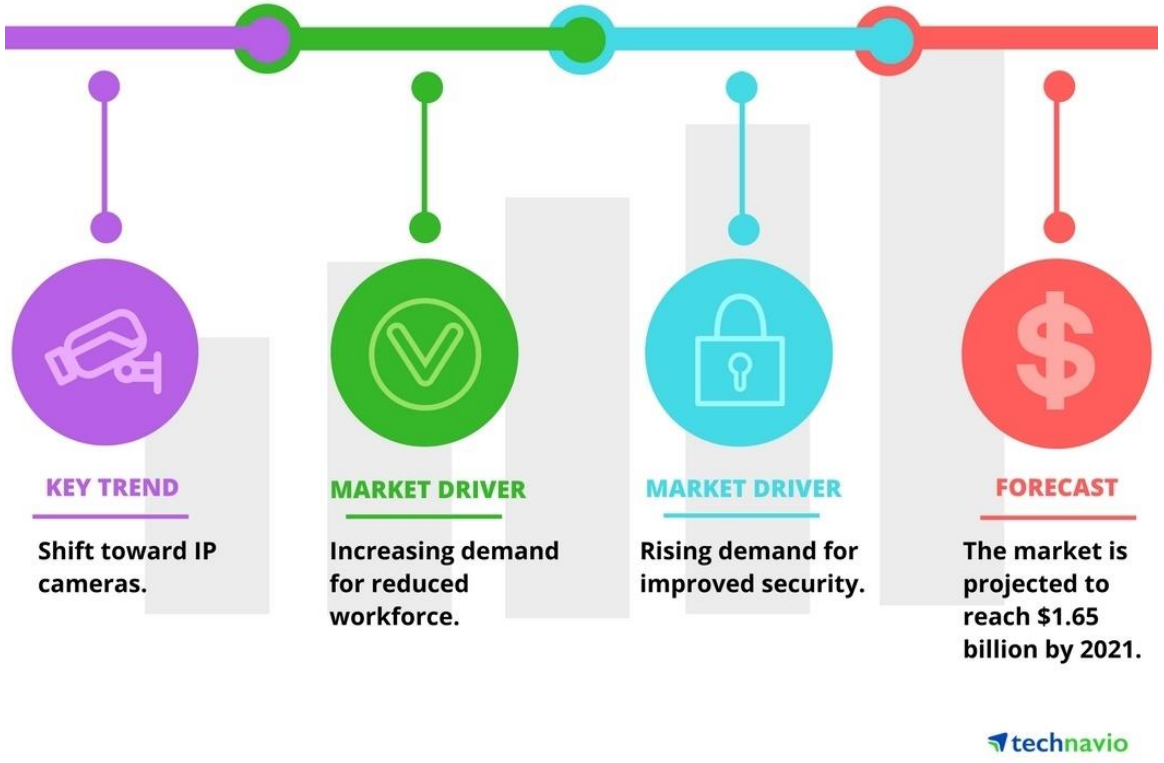
AI Smartphones Market

	A11 Bionic • Neural Engine x 2 • 10nm • Sep 2017	**A12 Bionic** • Neural Engine x8 • 7nm • Sep 2018
HUAWEI	**Kirin 970** • NPU / Image DSP • 10nm • Sep 2017	**Kirin 980** • NPU x2 • 7nm • Sep 2018
MEDIATEK	**Helio P60** • (APU) Image DSP • 12nm • Feb 2017	**Helio P70** • (APU) Image DSP x2 • 7nm • Oct 2018
QUALCOMM	**SDM 845** • Hexagon 685 • 10nm • Dec 2017	**SDM 855** • ??? • 7nm • Dec 2018



1 in 3 smartphones to be shipped in 2020 will natively embed ML & AI capabilities at the chipset level

Intelligent Video Analytics Market



Intelligent Video Analytics Market

One of the **KEY TRENDS** for this market will be the rising trend of **SMART CITIES**

Will grow by more than US$64 billion by 2022 as it particularly picks AI-based video surveillance as a key trend for the future

Facial recognition

Skynet program in China

IoT powered by AI

WHY NEURAL NETWORKS?

- Because there are more image sensors than people on the planet!

Intelligent Video Analytics Roadmap

viisights
Intelligence by vision

Future
Prediction
Recommend on actions to predicted situation of interest

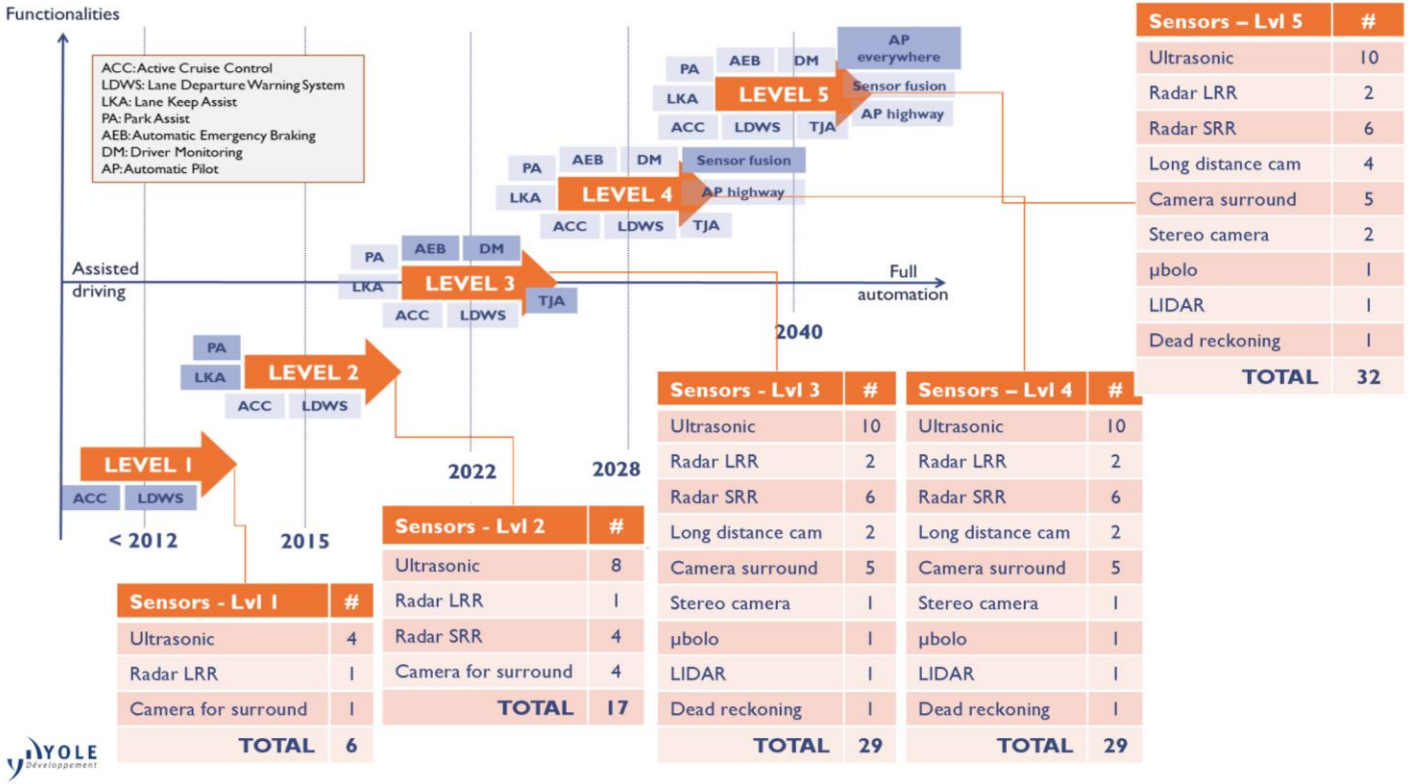
Roadmap
Fusion
Integration of video understanding data with multiple data sources

Today
Video Understanding
Actions, Events & Scenes

Yesterday
Image Recognition
Objects, Face

CEVA®

ADAS to Autonomy



Autonomous vehicles - Image sensor markets to 2027

Standard saloon

Number of on-board image sensors **0**

2016 top Advanced Driver Assistance System (ADAS) model

Number of on-board image sensors **11**

Fully automated driving (AD) vehicle

Number of on-board image sensors **25+**

World market growth factor 2016-2026

ADAS vehicles **3.45**

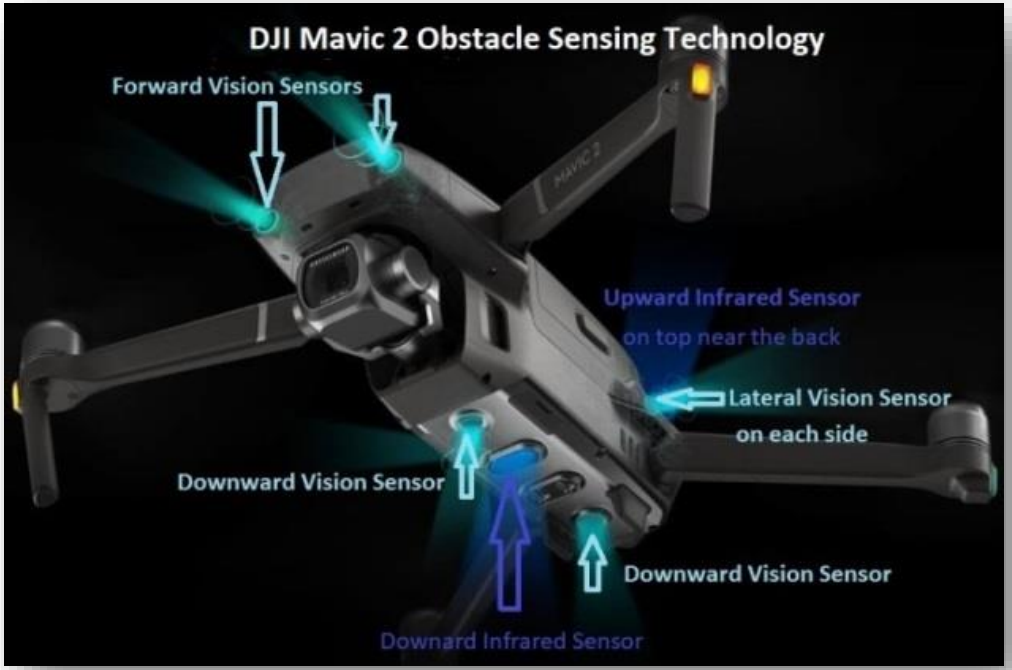
Full AD vehicles **161.4**

Forecast demand AD/ADAS sensor types (million units)

	Cameras	RADAR	LIDAR
2016	52	13	1.6
2021	92	15	3.5
2030	400	40	50

Passenger Awareness

Drones – Obstacle Sense & Avoid

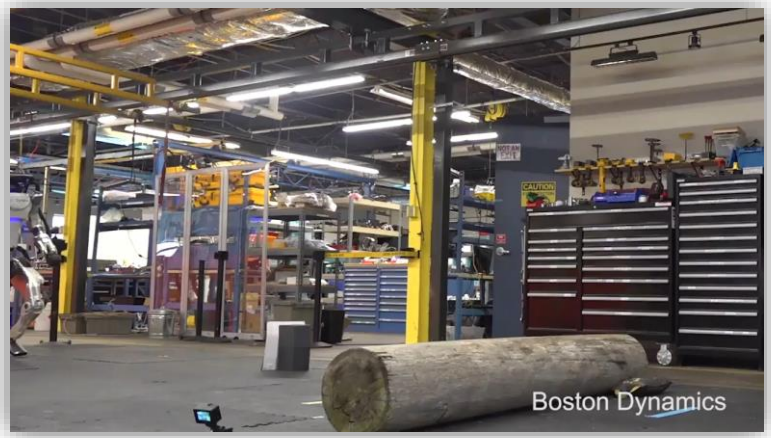


Omnidirectional Obstacle Sensing
Can sense objects in 6 directions

Intelligent Robots

Going from here…

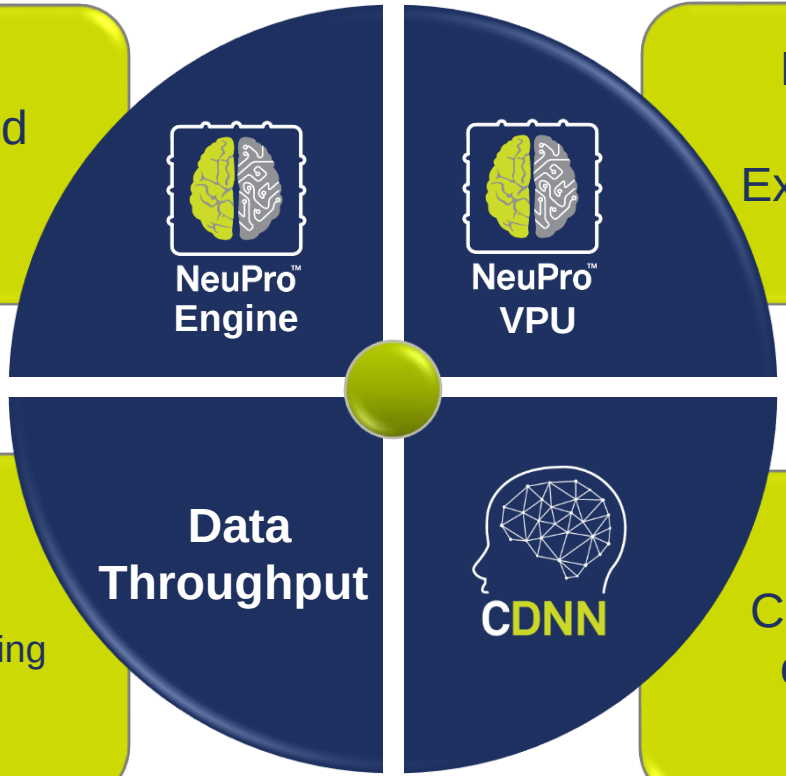
Into the future!



NeuPro Introduction

NeuPro - Holistic Philosophy

Specialized ,Optimized AI Accelerator



Programmable Vector Processor for Expandability & computer Vision Co-processing

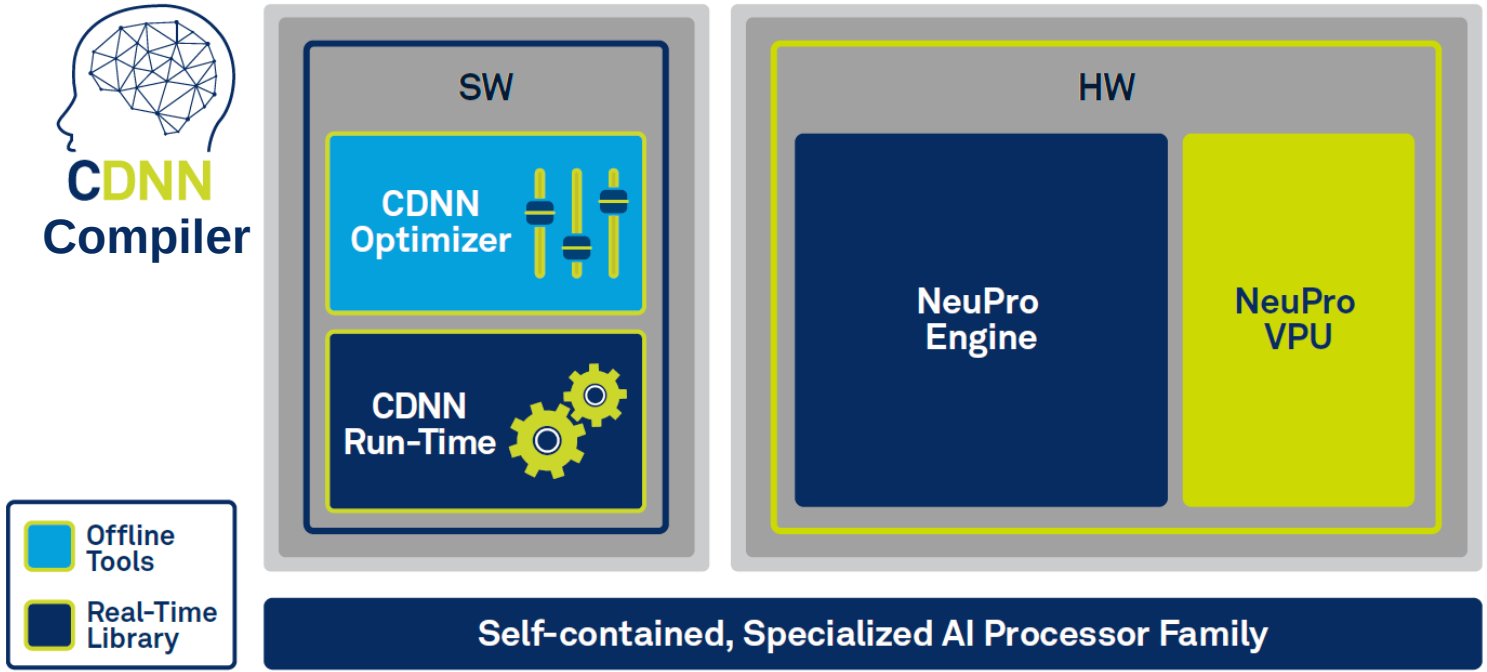
- Advanced DMA features
  - Multi dimension DMAs
  - DMA Queue Manager
  - Automatic padding/cropping
- Internal buffers & paths
- AXI Buss Bridges

Neural Network Compiler includes offline optimizer and runtime software framework

NeuPro - AI Processor Highlights

- Self-contained, four specialized AI processors
- Scalable performance up to 4K 8x8 MACs, up to 12.5 TOPS
- Composed of
  - **NeuPro Engine** - Convolution, Fully Connected, Activation and Pooling layers
  - **NeuPro VPU** - Fully programmable vector processor, simultaneous processing
- Dynamic quantization supported - 8-bit and 16-bit mix
  - Per layer precision vs. performance tradeoff – via CDNN
- All layer types and NN topologies supported
  - Maximize performance via CDNN SW compiler
- Optimized Data Bandwidth
  - Internal busses, DMAs, smart data re-use
- Use models: standalone AI or Computer Vision and AI combo

NeuPro AI Processor



CEVA
NeuPro™
by CEVA

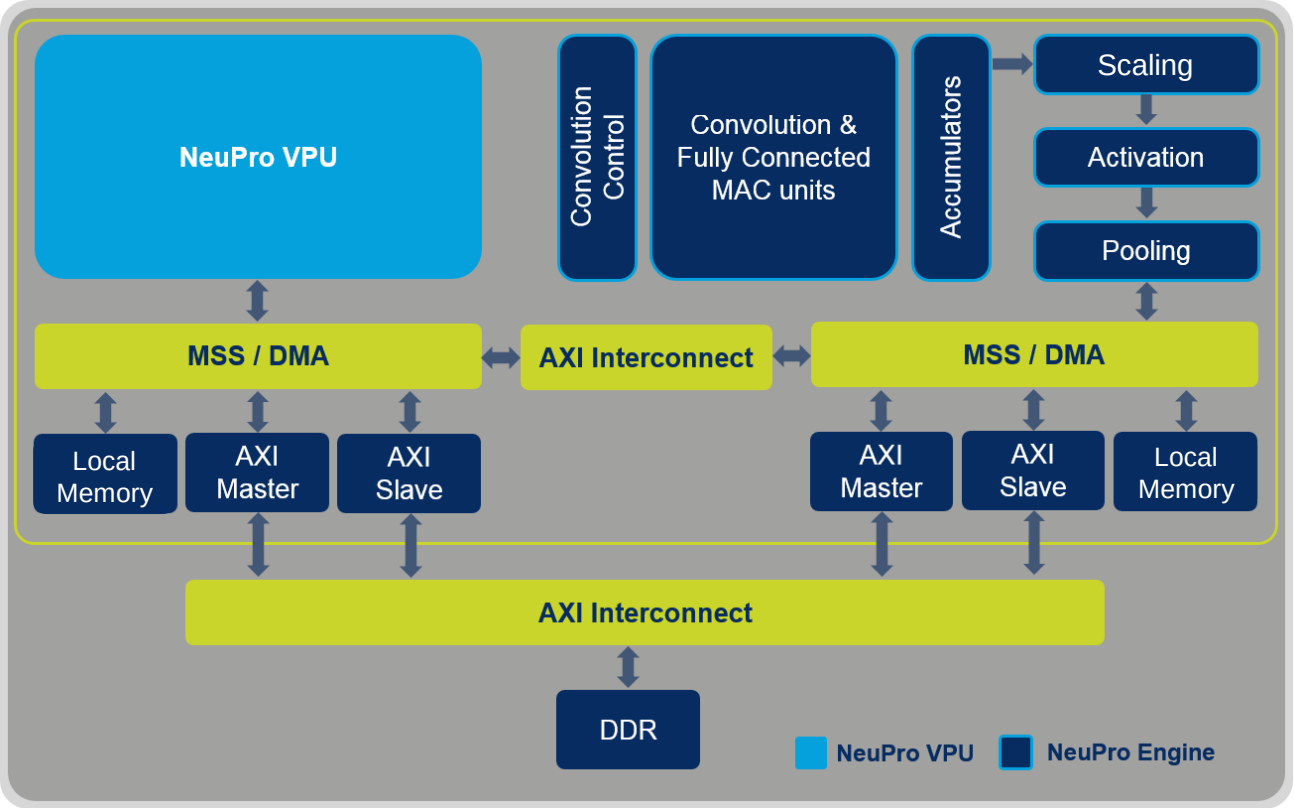
NeuPro AI Processors Family

Product Name	MAC Configuration			Target Market
	8x8	16x8	16x16	
NP4000	4096	2048	1024	High-performance edge processing in enterprise surveillance and autonomous driving
NP2000	2048	1024	512	High-end smartphones, surveillance, robots and drones
NP1000	1024	512	256	Mid-range smartphones, ADAS, industrial applications and AR/VR headsets
NP500	512	256	128	IoT, wearables and cameras

Self-contained, specialized AI processors, scaling in performance for a broad range of end markets

NeuPro HW Block Diagram, Parallelism Flow

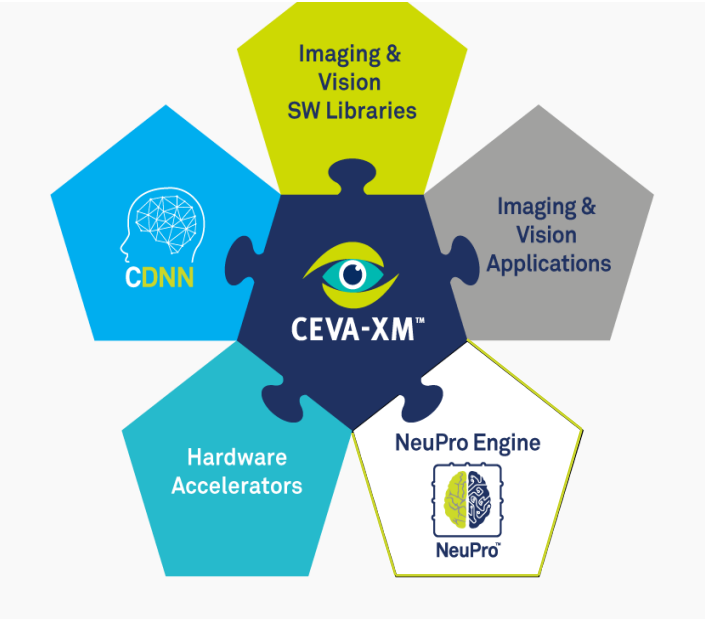
Seamless Handover between **NeuPro Engine** & **NeuPro VPU**



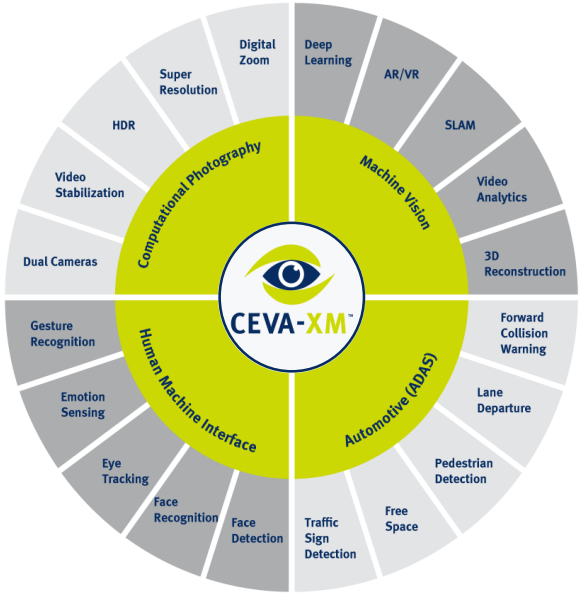
CEVA-XM & NeuPro Engine - Combo Platform

A flexible option for a single unified platform for

Imaging, Computer Vision and Neural networks

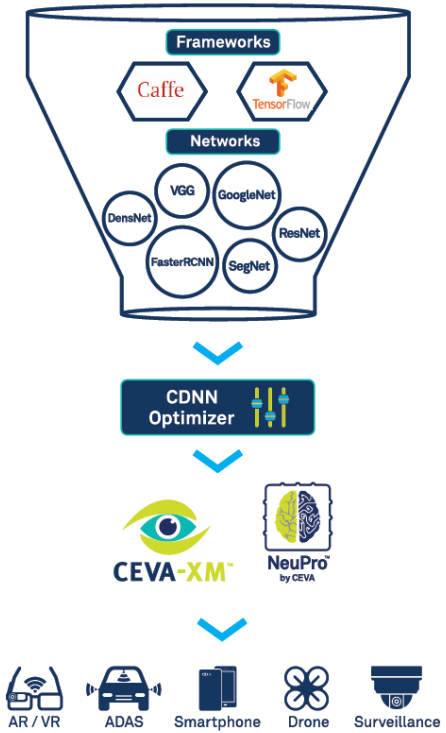


Enabler



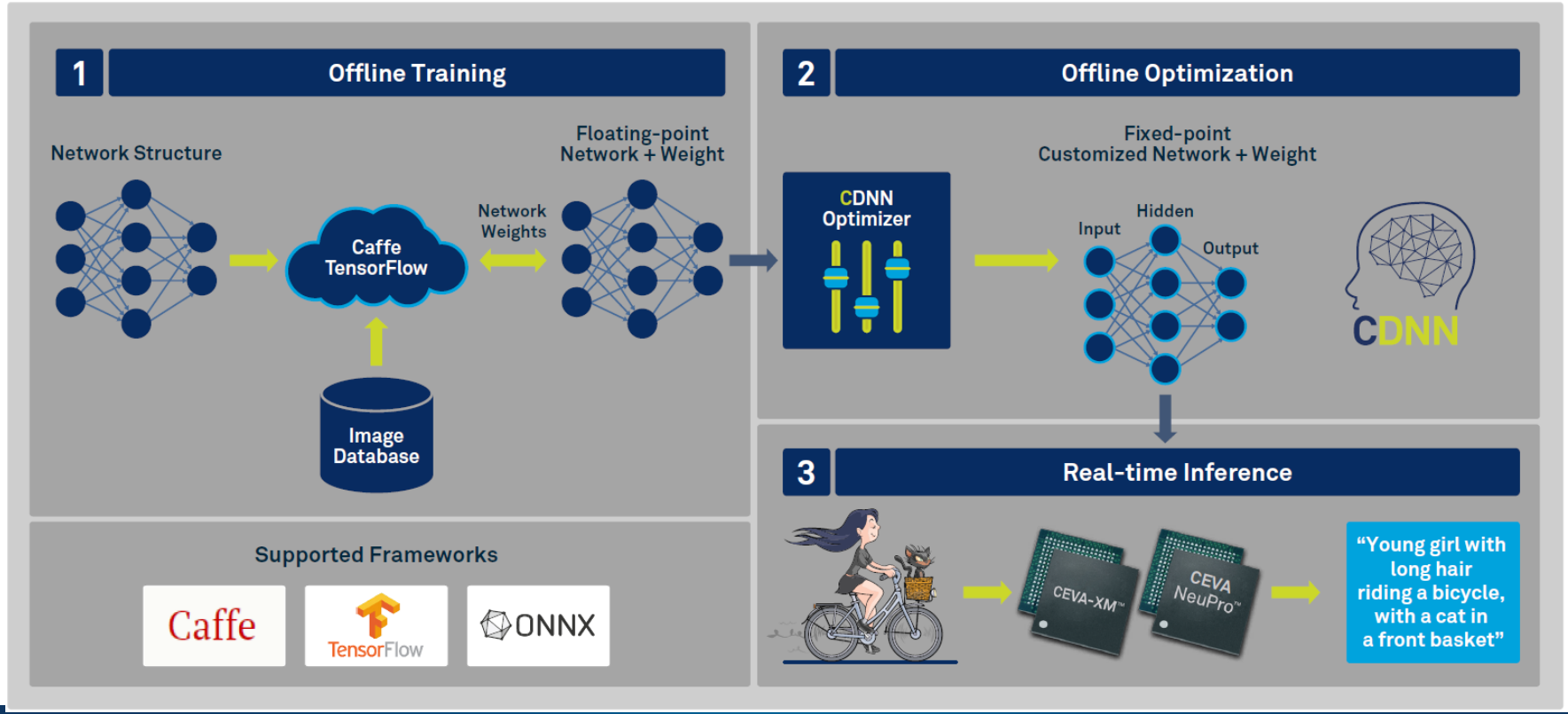
CDNN SW Compiler

CEVA Deep Neural Network (CDNN)

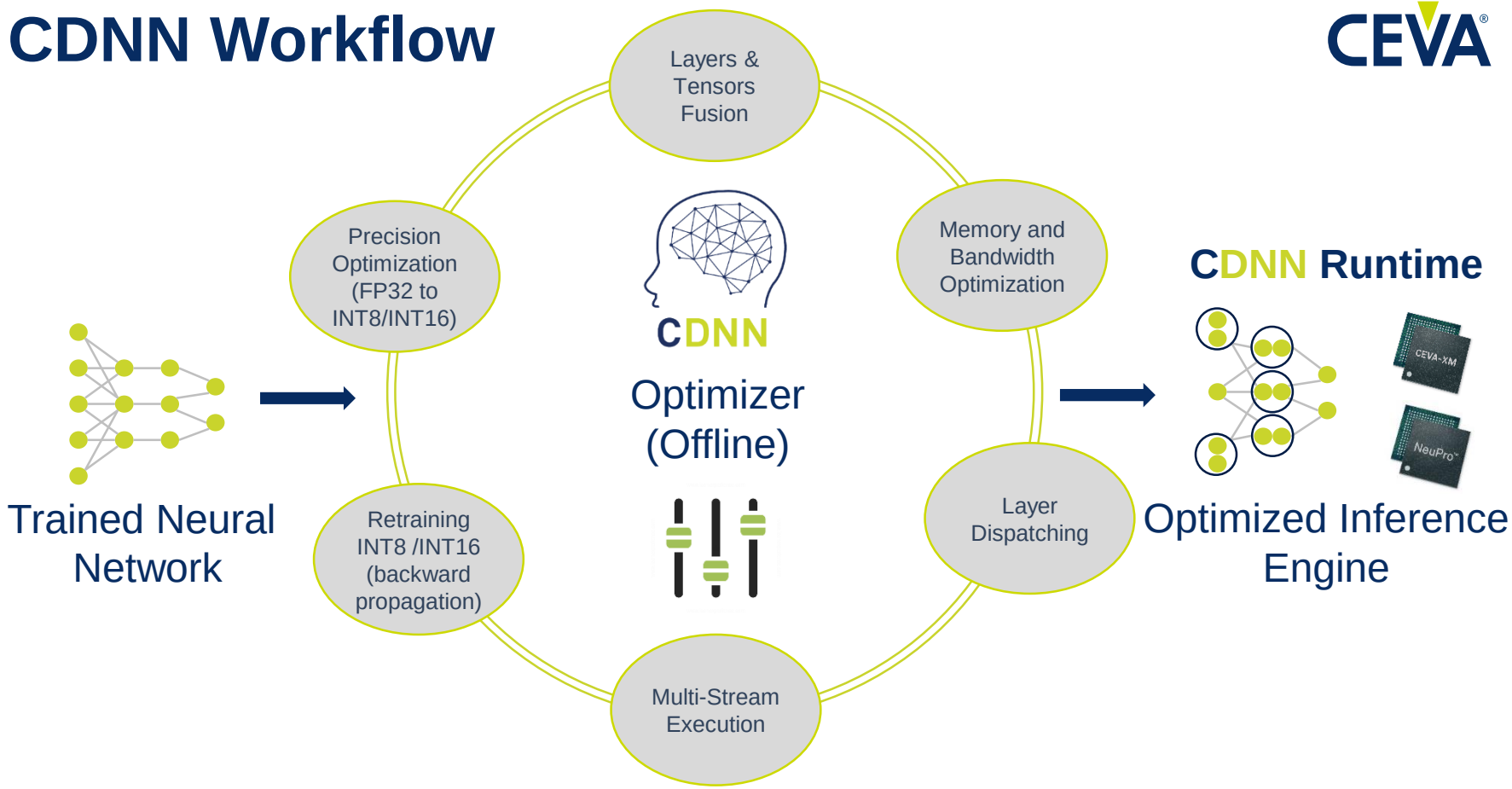


- Neural network SW compiler for Inferencing
- Automatic offline optimization and run-time SW framework
- Mature and robust solution **supporting over 130 NNs**
- Fully optimized for CEVA-XM and NeuPro AI Processors

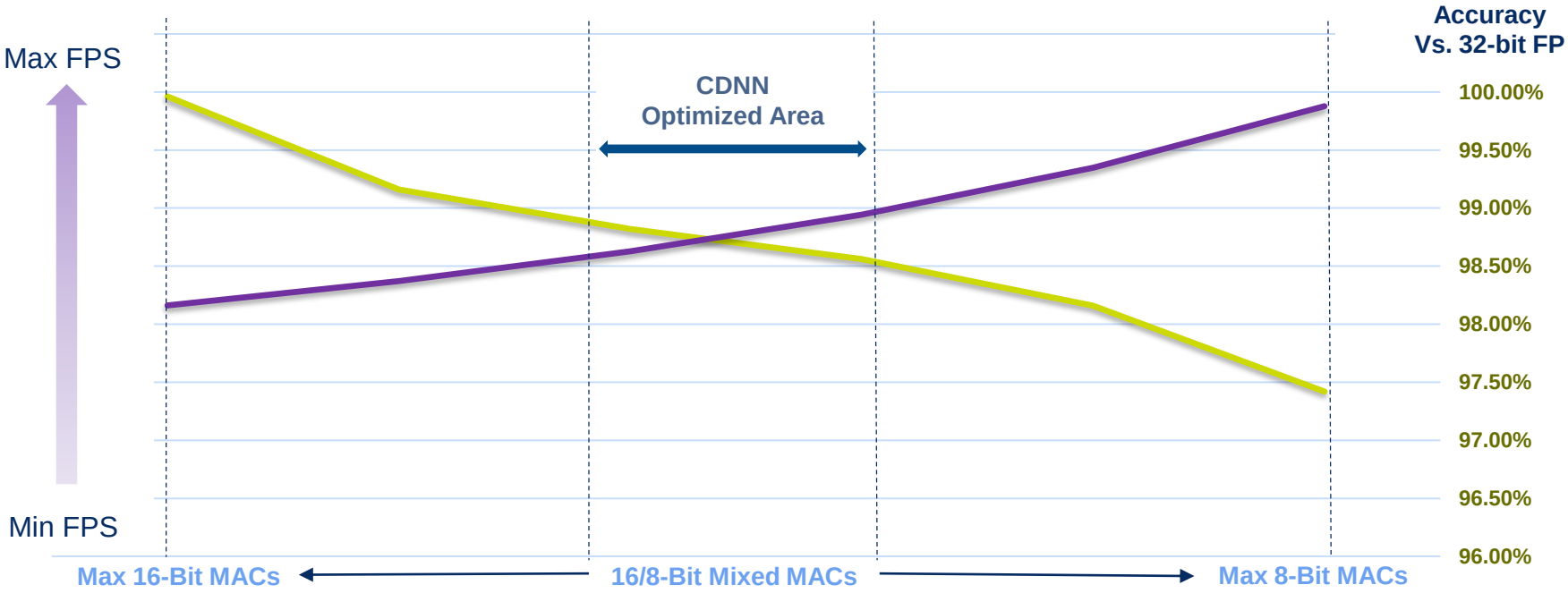
CDNN Compiler Usage Flow



CDNN Workflow



CDNN – Performance Vs. Accuracy

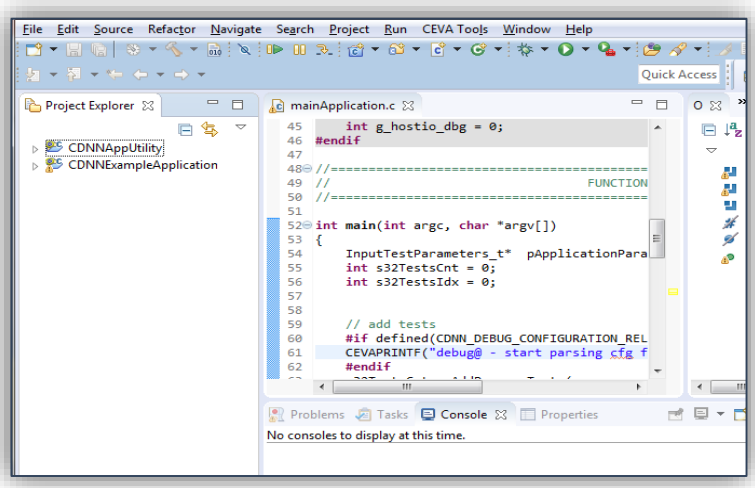


(*) Tested using MobileNet running on 5K images
(**) Percentage of images whose true class matches the network's maximum prediction

CDNN PC Simulation Package

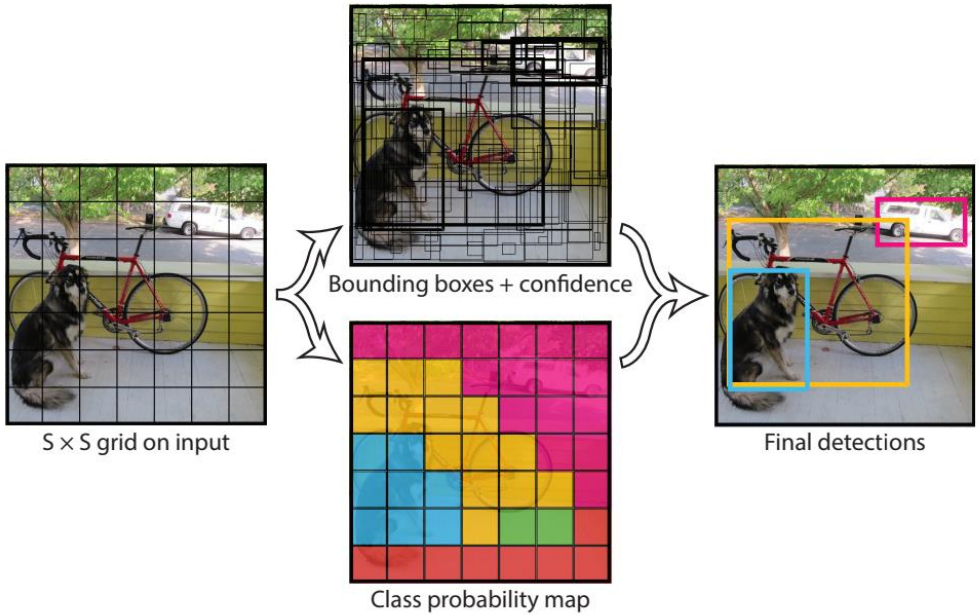
- CDNN Evaluation SW package
- Visual studio environment
- Example reference projects
- Full development flow on PC

  Edit ➔ Build ➔ Execute ➔ Debug
- CEVA-XM and NeuPro simulators are available, covering all configurations



Enables to achieve neural networks cycle count accuracy on a PC before having a dedicated HW

Real-time Tiny Yolo Neural Network Demo



Real-time Tiny Yolo
Neural Network
Running on SoC Powered
by CEVA-XM4

Thank You

Liran Bar, Director of Product Marketing,
CEVA

Email: liran.bar@ceva-dsp.com

www.ceva-dsp.com