Machine Learning on the Edge

Tashia Mehdi

Field Applications Engineer, MPUs – EMEA

Edge-AI Seminar

June 2024

Agenda

- Introduction
- ML on the Edge
- Machine Learning Solution
- Advancing State of the Art in ML
- Demo Room

A position of strength to better serve our 26,000+ customers

We accelerate breakthroughs that advance the world through our semiconductor technology leadership

EMPLOYEES IN

30+ COUNTRIES

Headquartered in Eindhoven, Netherlands

~34,000

TEAM MEMBERS

9,500+

Patent Families

$13.28B

Annual Revenue [1]

60+

Year History

~12,000

R&D team members

[1] Posted revenue for 2023 – Please refer to the Financial Information page of the Investor Relations section of our website at www.nxp.com/investor for additional information

SECURE CONNECTIONS FOR A SMARTER WORLD…HAS EVOLVED

AUTOMOTIVE

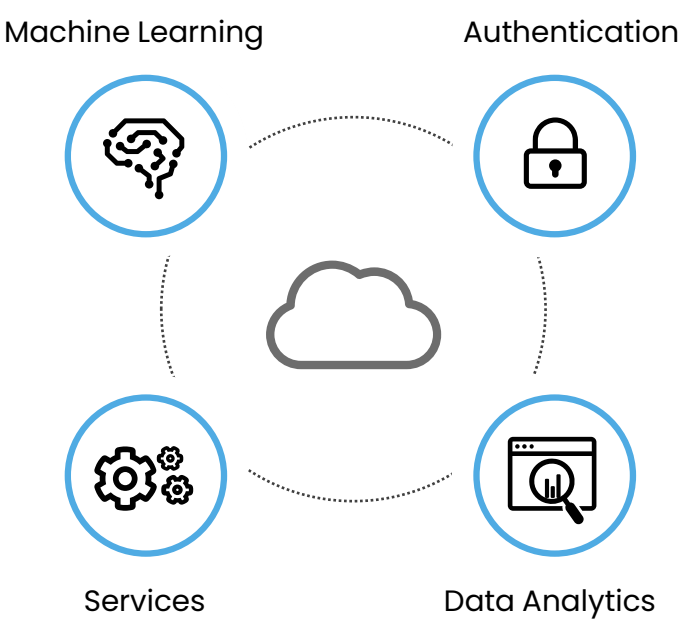
INDUSTRIAL

MOBILE

COMMUNICATION INFRASTRUCTURE

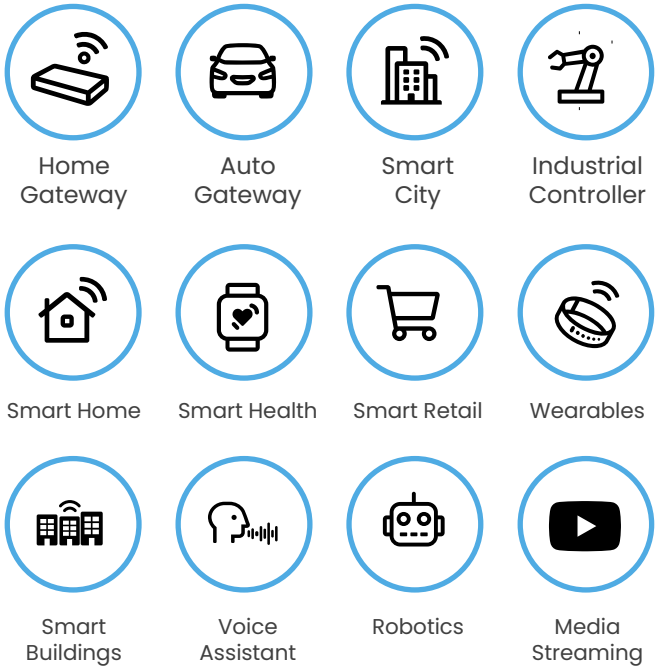
Focus Verticals

CLOUD INFRASTRUCTURE



ENABLING TECHNOLGIES

EDGE TO EDGE



ML on the edge

i.MX 9 SERIES OF APPLICATIONS PROCESSORS
ADDING TO OUR PORTFOLIO

Software Compatible

Pin-to-pin Compatible

- i.MX 6QuadPlus
- i.MX 6Quad — A9
- i.MX 6DualPlus
- i.MX 6Dual
- i.MX 6DualLite
- i.MX 6Solo
- i.MX 6SoloX — A9 M4
- i.MX 6SoloLite
- i.MX 6SLL
- i.MX 6UltraLite
- i.MX 6ULL — A7
- i.MX 6ULZ

Arm® v7-A

i.MX 8 Family — A72 A53 M4
Advanced graphics, vision & performance

i.MX 8M Family — A53 M4/M7 NPU
General purpose Edge processing

i.MX 8ULP Family — A35 M33
Ultra low power with graphics

i.MX 8X Family — A35 M4
Safety certifiable & efficient performance
Arm ® v8-A (32-bit/64-bit)

i.MX 7 Family — A7 M4
Flexible efficient connectivity

i.MX 7ULP Family — A7 M4
Ultra Low Power with Graphics
Arm® v7-A (32-bit)

i.MX 95 Family
ML Vision, Safety Enabled Platforms, HMI, Powerful Android/Linux Platforms
FuSa A55 M33 M7 NPU ISP GPU

i.MX 93 Family
ML Vision, Automotive Edge, Industrial HMI, Smart Appliances, Industrial & IoT Gateway
A55 M33 NPU

i.MX 91 Family
Secure Linux Controller Platforms, IoT & Industrial RTOS, Connected IoT & Industrial
A55

...and Future i.MX 9

Scalable Software

Scalable Platform
Arm® v8.2-A (32-bit/64-bit)

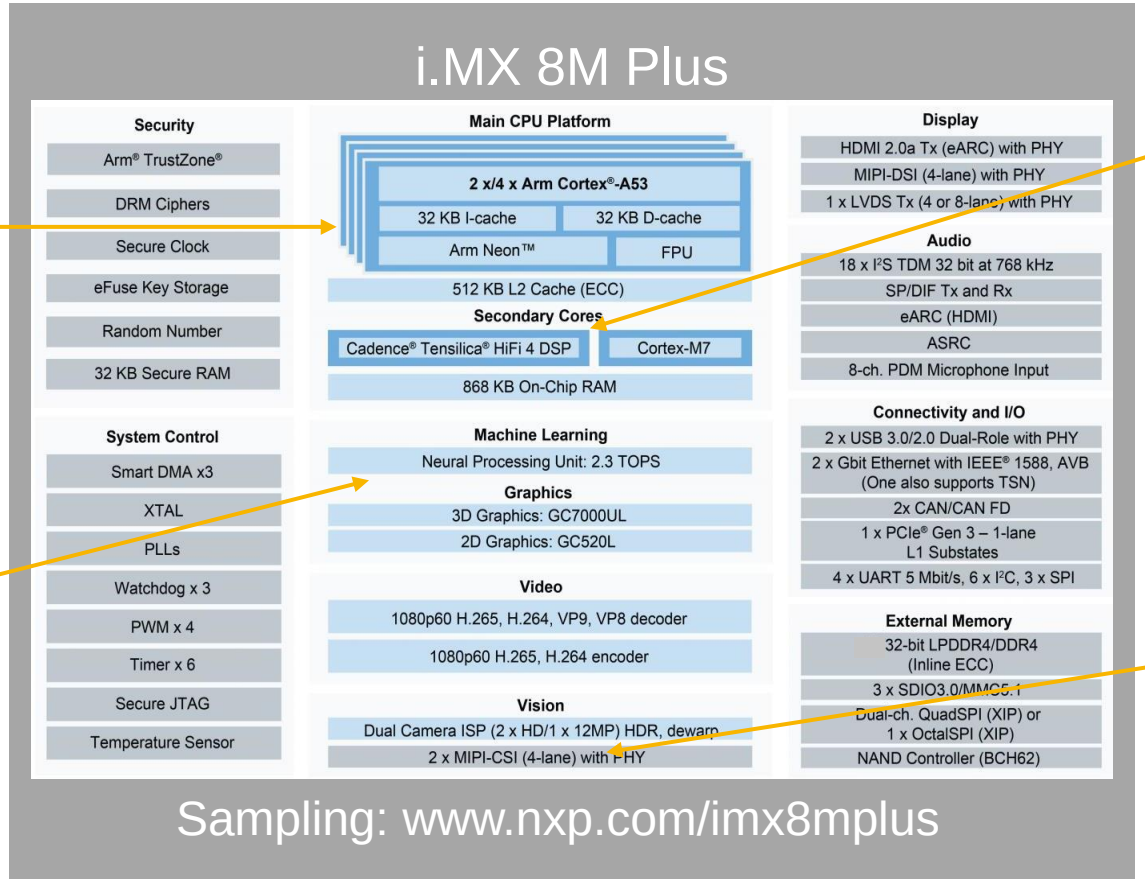
i.MX 8M Plus machine learning compute engines

Quad Arm® Cortex® -A53 @ 1.8 GHz

- Speech command recognition
- Object detection classification
- Gesture recognition

Neural Processing Unit (NPU) @ 1 GHz

- Multi-camera classification and detection



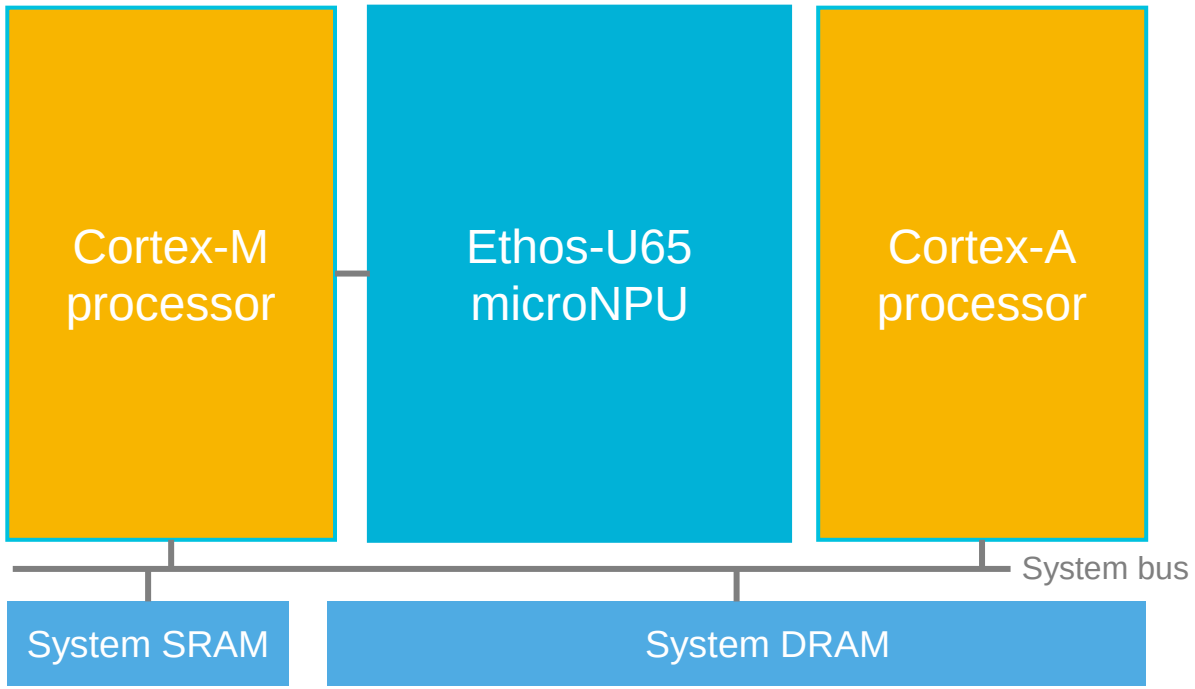
Cortex-M7 @ 800 MHz and HiFi4 DSP @ 800 MHz

- Keyword detection
- Sensor fusion
- Anomaly detection

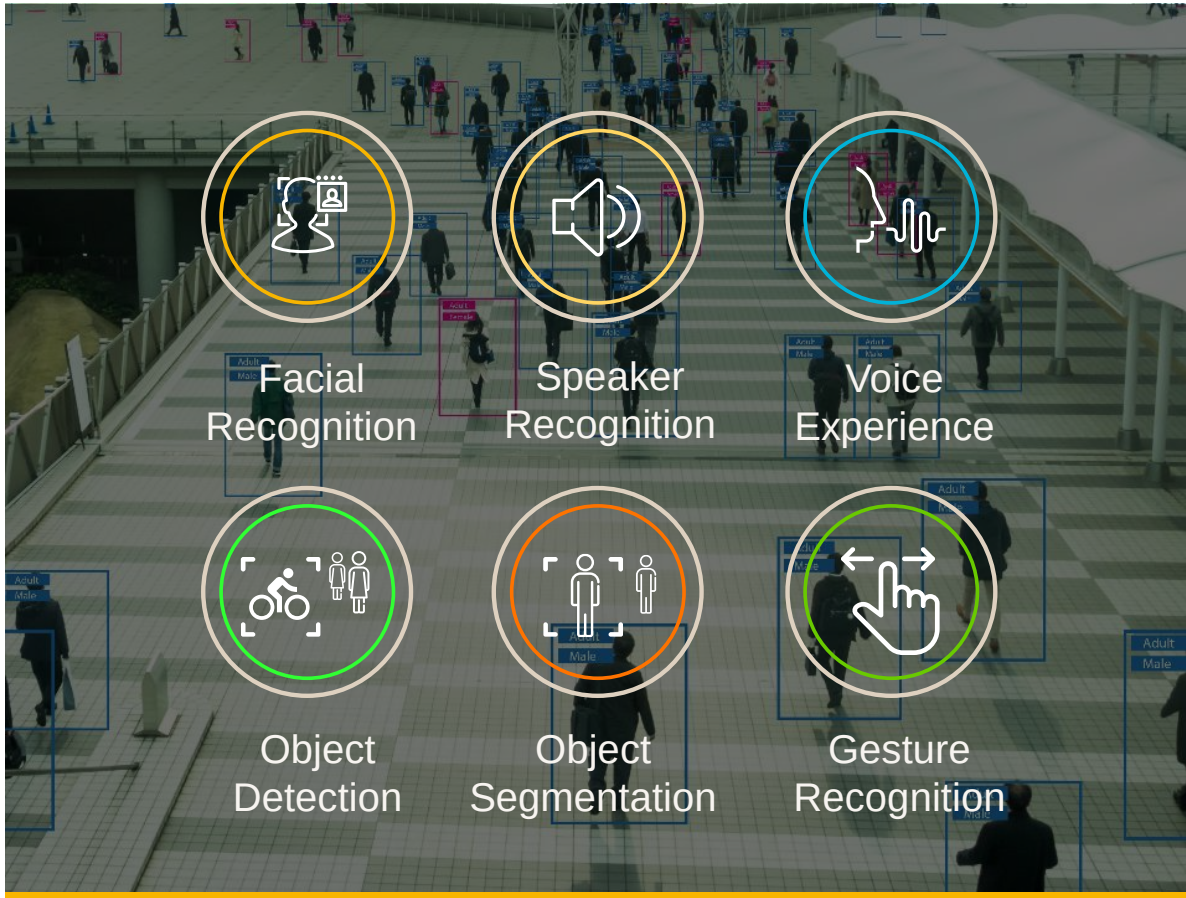
Two-channel Image Signal Processor (ISP) for dewarping and image enhancement

ISI for scaling and other image enhancements

i.MX 93 : Expanding Edge ML with Arm® Ethos™-U65



- High efficiency and small memory footprint
- HW acceleration for high compute NN + Cortex-M for other operations with 0.5 TOPS
- Model compression and on-the-fly weight decompression
- Optimization strategies for DRAM and SRAM
- Comprehensive software and tools with NXP's eIQ® ML Software Development Environment

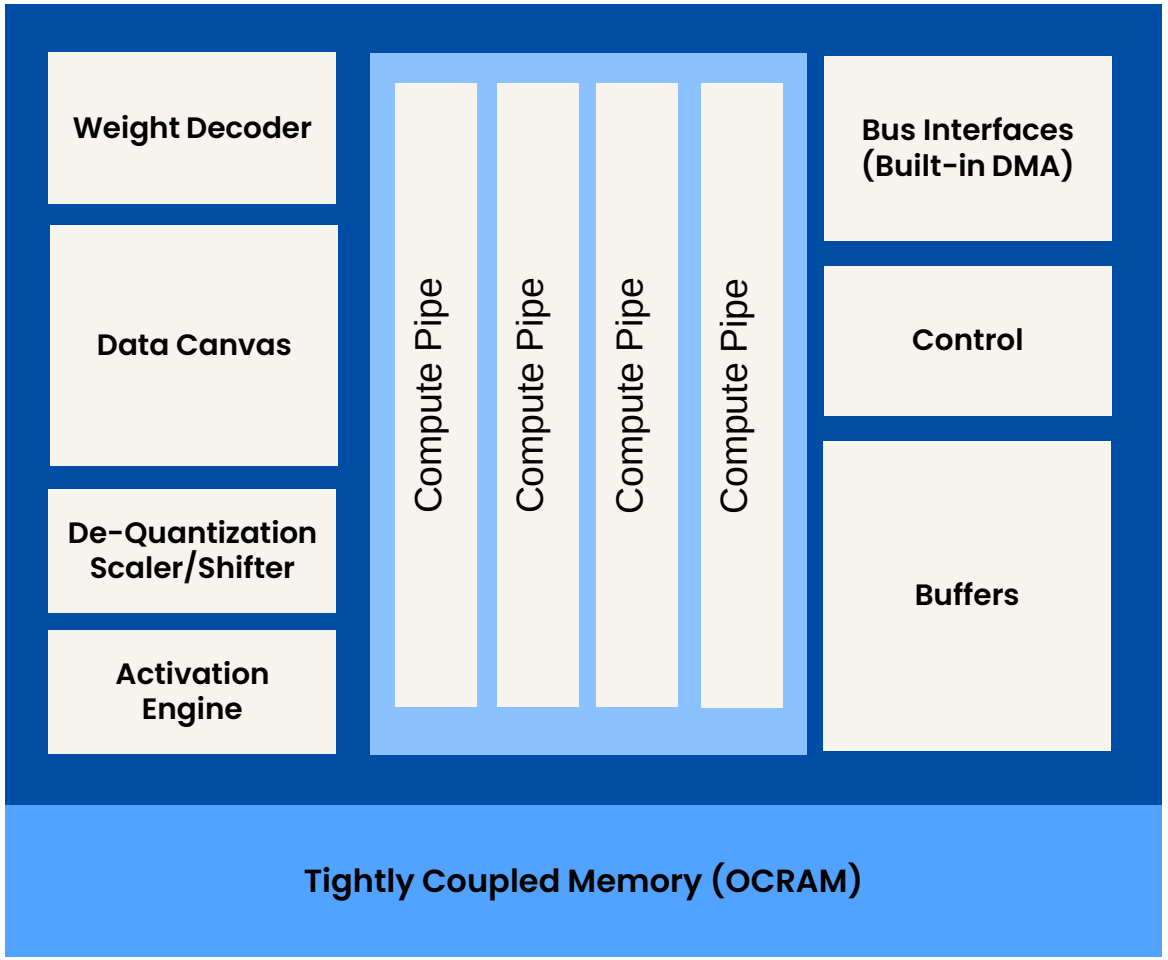


BRINGING MCU-CLASS ML EFFICIENCY TO THE CORTEX-A WORLD

i.MX 95: NXP eIQ® Neutron
Neural Processing Unit (NPU)

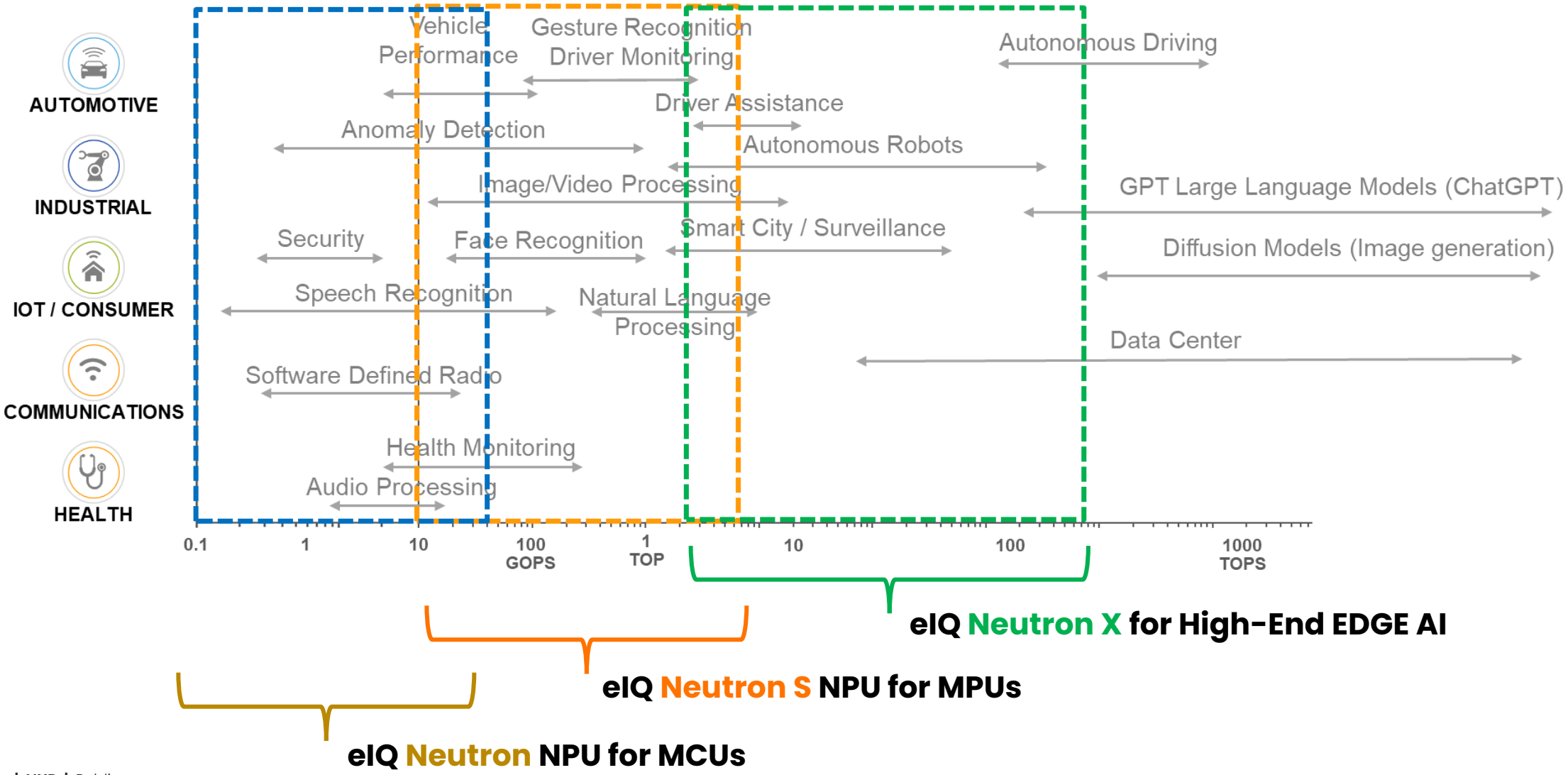
- Single Architecture With Great Scalability
- Optimized for Performance and Power Efficiency
- ML solution development support with eIQ® ML SW Development Environment
  - Supports major NN structures (CNN, MLP, RNN, LSTM, TCN, and more)
  - LLM support exploration underway: LLAMA v2 & Blenderbot
- Internal development provides flexibility to tune solution to better meet our customer needs and the ability to provide ongoing support and generational improvements for changing applications and operator support needs
- Hardware scales from performance efficient 32 Ops/cycle to 2k Ops/cycle and beyond for portfolio coverage with a single architecture, and potential to provide future expansion
- Software support is unified over multiple generations and device portfolio, creating consistent enablement and support solutions for our customers

NXP eIQ Neutron NPU N3-1024S IP



https://www.nxp.com/company/blog/introducing-the-nxp-eiq-neutron-neural-processing-unit-npu:BL-INTRODUCING-THE-NXP-EIQ-NPU

AI/ML COMPUTE WORKLOADS - MAPPING TO NXP EIQ NPU



EDGELOCK™
SECURE ENCLAVE

MULTI-SENSORY EXPERIENCES
STREAMING MEDIA
RICH 2D & 3D GRAPHICS
ADVANCED AUDIO
VOICE PROCESSING
TOUCH SENSING
VISION

ENERGY FLEX ARCHITECTURE
WITH HETEROGENEOUS DOMAIN COMPUTING

INHERENTLY INTELLIGENT
INTEGRATED ML ACCELERATORS

SCALABLE COMPUTE
HIGH PERFORMANCE – FROM SINGLE TO MANY CORE CONFIGURATIONS

ESSENTIAL CONNECTIVITY

BUILT-IN MCU!
REAL TIME RESPONSE FOR THE REAL WORLD

ALWAYS-ON, LOW POWER SENSING

NXP

EDGELOCK™
SECURE ENCLAVE

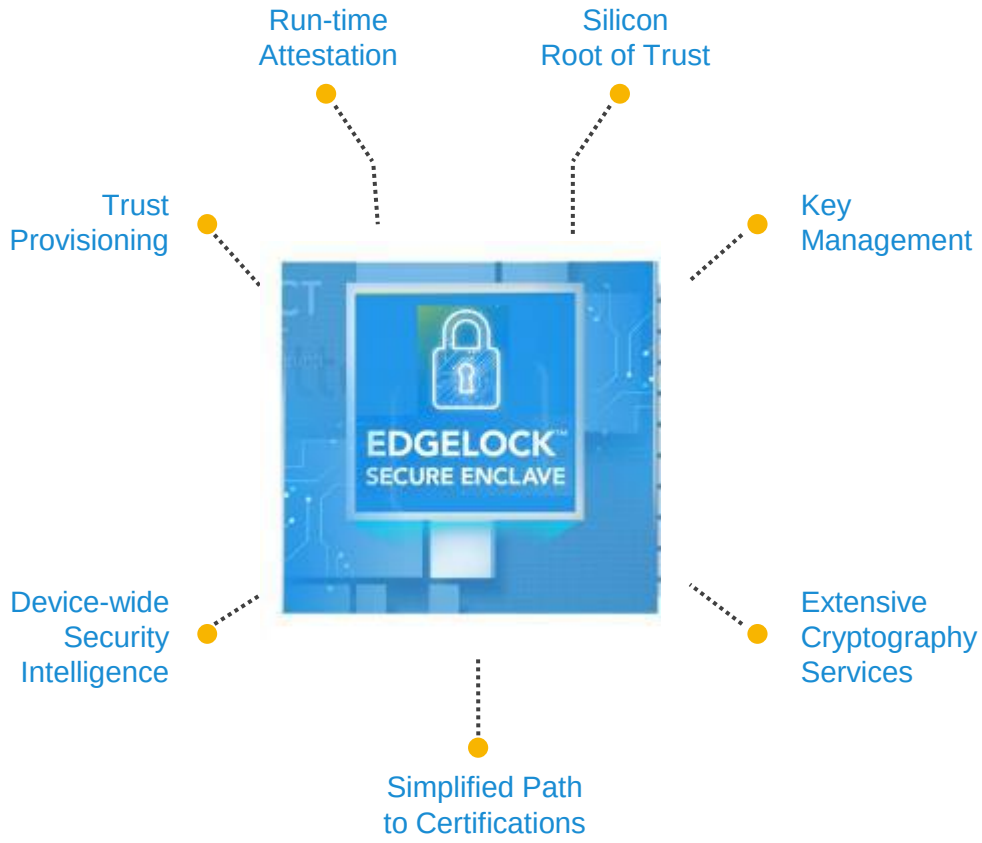
BEYOND CRYPTO
Evolved on-die security with run-time attestation, silicon root of trust, trust provisioning, fine-grain key management augmented by extensive crypto services and simpler path to security certifications

SECURITY “HQ”
Where security is governed – this fortress inside the chip oversees security functions to protect the system against attacks

MANAGED AGENTS
Agents extend security across the chip – distributed outside of the central HQ – to establish and maintain trust of security capabilities

INTELLIGENT
Tracks and manages power transitions to help prevent attack surfaces from emerging on heterogeneous multicore devices

READY TO GO
Preconfigured security policies reduce the complexity and help avoid costly errors for faster time to market



nxp.com/SecureEnclave

eIQ Machine Learning Solution

eIQ® ML SW Development Environment

Embedded Developers

Data Scientists

ML Experts

eIQ® Toolkit

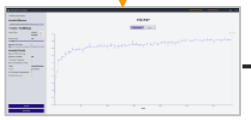
eIQ® Portal

eIQ® Marketplace

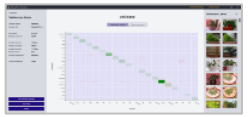
Bring Your Own Data Workflow

Data Curation

Model Selection



Model Training, Optimization, Quantization

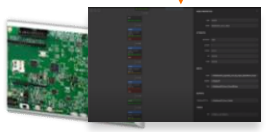


Model Validation

Bring Your Own Model Workflow

PyTorch

ONNX



Applications running on EdgeVerse processors

eIQ® Pipelines

VISION PIPELINE

VOICE/AUDIO PIPELINE

ANOMALY (TIME SERIES) PIPELINE

eIQ® inference

eIQ® inference with…

TFLite-Micro

Arm® Cortex®-M, DSP, NPU

TFLite

Arm Cortex-A, GPU, DSP, NPU

Solutions and Services from NXP and NXP Eco-System Partners

- ML Applications
- Optimized Models
- Optimization Tools and Modules
- Development tools
- Datasets
- Training
- Sensor solutions

Watermarking Neural Networks

Approach: Embed hidden functionality (the watermark) in a machine learning model

✓ No performance penalty

- No drop in accuracy on the primary problem

✓ Robust

- watermark resistant to changes to the network (e.g., pruning weights and nodes)

✓ Provides legal protection against copying/cloning

- Unlike software, cloning/copying a (non-protected) model is typically legally allowed and not prohibited by copyright protection (see [1])
- NXP's watermarking scheme adds copyright protection to a model
- Workflow tool designed around legal workflow for proving copyright infringement

[1] W. Michiels, How do you protect your machine learning investment, EETIMES

✓ Easy deployment

- To embed watermark, training set only needs to be extended with images provided by NXP's watermarking tool
- No changes to training process needed

✓ Easy to use

- detect that a model is a clone by querying the clone and checking the output prediction

True label
monkey

Predicted label
(clone and original)
car

Development of eIQ ecosystem

Enabling NVIDIA’s trained AI models to be deployed on NXP’s edge processing devices

NXP is the **first semiconductor vendor** to integrate NVIDIA TAO Toolkit APIs directly with its AI enablement offering, the eIQ machine learning development environment

NXP Collaborates with NVIDIA to Accelerate AI Deployment with Integration of TAO Toolkit with NXP Edge Devices

March 18, 2024 6:00 PM EDT (UTC-4) by NXP Semiconductors Press Release

SHARE

- NXP is the first semiconductor vendor to integrate NVIDIA TAO Toolkit APIs directly with its AI enablement offering, the eIQ machine learning development environment
- Enables NVIDIA's trained AI models to be deployed on NXP's edge processing devices
- Accelerates AI development by making it easier to deploy trained AI models at the edge

Advancing State of the Art in ML

75+ billion

Smart connected Devices by 2030

A WORLD THAT ANTICIPATES AND AUTOMATES

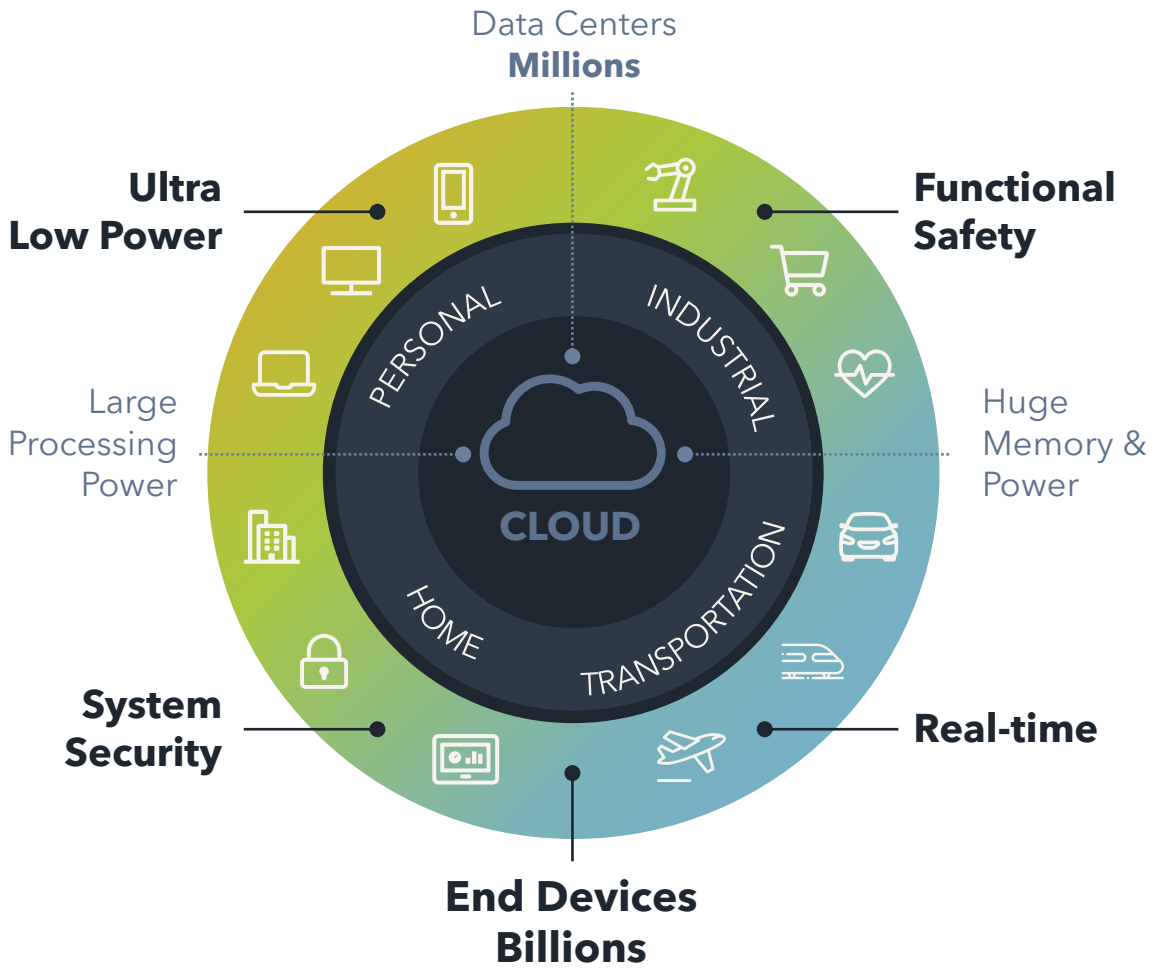
Moving From Cloud to Edge

Enable Real-Time Analytics and Actuation with Local Machine Learning
Not hampered by network latency

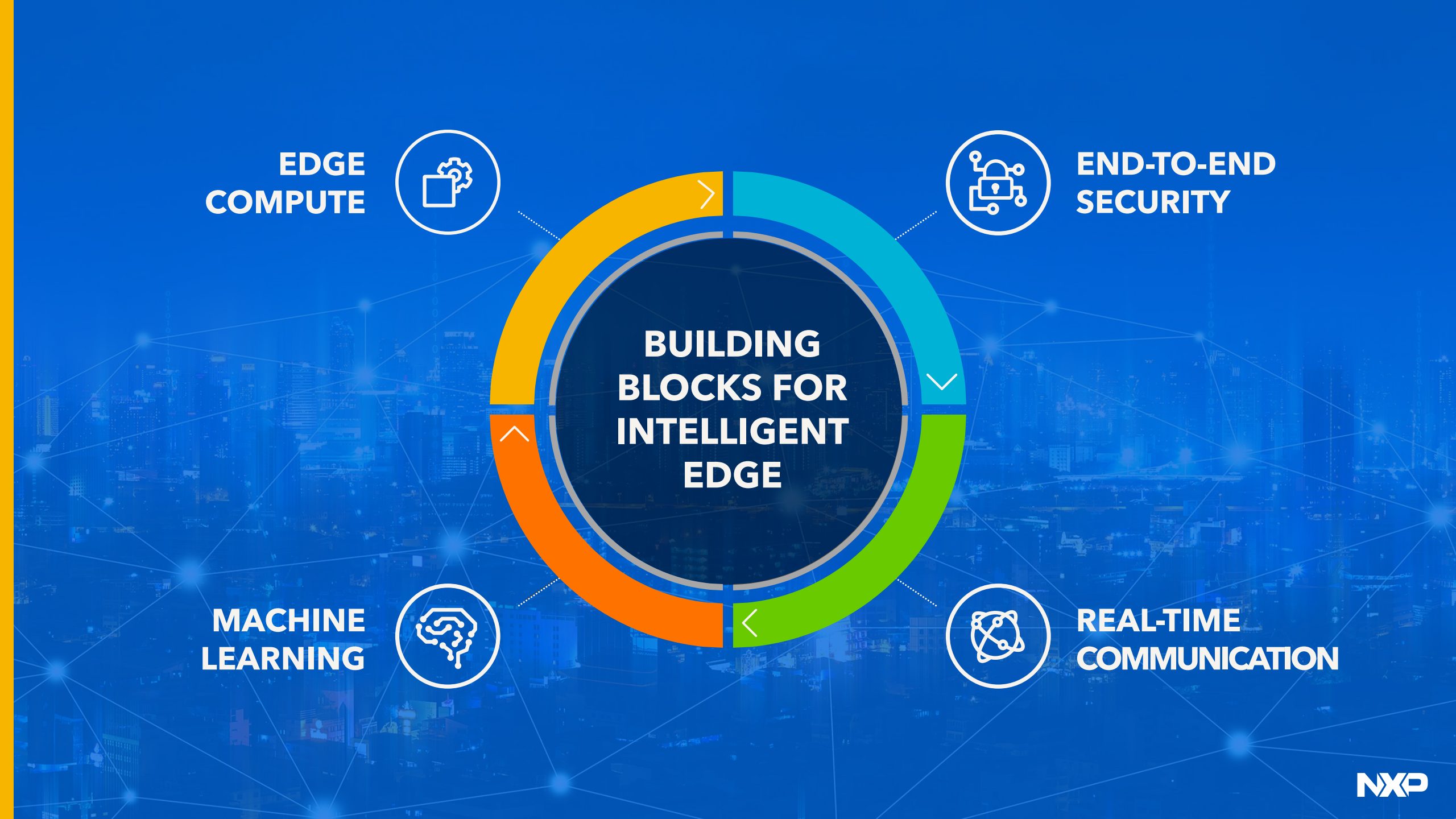
Reduce Data Center and Network Cost
Only transmit, process and store relevant data

Safeguard Privacy
Transmit semantic rather than raw data

Increase Security
Resilient to offline conditions



Data collection, processing and decisions at the edge
Devices securely connected to the cloud


EDGE COMPUTE
END-TO-END SECURITY
BUILDING BLOCKS FOR INTELLIGENT EDGE
MACHINE LEARNING
REAL-TIME COMMUNICATION

Developing ML for intelligent edge devices is an investment

$154B

Global Spending on Artificial Intelligence (AI) centric systems in 2023

27%

Compound annual growth rate (CAGR) 2022-2026

Incl. Software

Hardware

Services

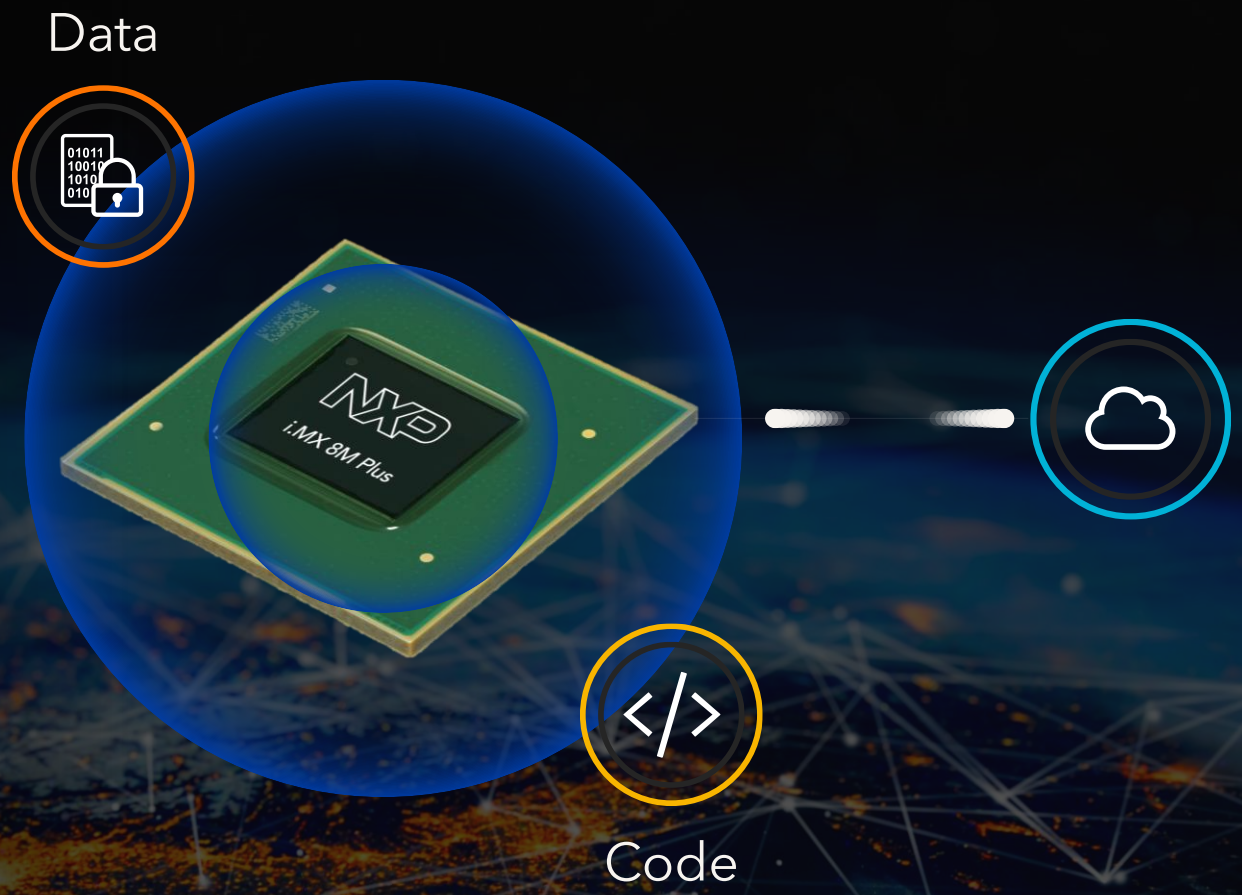
Source: IDC Worldwide Semiannual Artificial Intelligence (2022)

Sophisticated threats **outpacing** security of equipment

Manufacturing was the most attacked sector last year, accounting for **23% of reports** of ransomware

Sources: IBM, 2021

Any security for ML needs to account for the Data, the code and the DEVICE



eIQ® MACHINE LEARNING DEVELOPMENT ENVIRONMENT

Embedded Developers

Data Scientists

ML Experts

eIQ Marketplace

eIQ Toolkit

- Data labeling, curation
- Model conversion training, optimization, quantization
- Model validation, profiling

eIQ Pipelines

AUDIO / VOICE

VISION

TIME SERIES

App SW Packs

STATE MONITOR

IDENTITY RECOGNITION

PERSON DETECTION

and more ….

From MCUs

To Apps Processors

SCALABLE COMPUTE

NXP

NEW THREATS

GENERATIVE AI

AI can be used to self-identify vulnerabilities
Could be trained to defeat its own defenses

Quantum computing

Quantum computers accelerate decryption
Solving the algorithms behind encryption keys that protect our data and infrastructure

CONSTANT INNOVATION

Secure Edge Compute
Continued innovation in embedded security capabilities for attack resistance complemented by trust provisioning and secure on-boarding

AI ML Enablement
Continued innovation to protect the data ,and the IP being generated

Post-Quantum Cryptography
Continued innovation for high-assurance implementations, resistance against side-channel and fault attacks and dedicated PQC hardware

TECHNOLOGY SHOWROOM

JOURNEYS BY DESIRED ENGAGEMENT

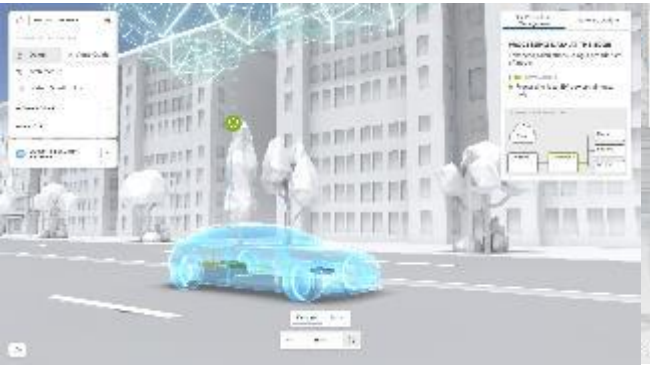
Self-guided tour
Live-streaming at set times
Guided tours

JOURNEYS BY DESIRED FOCUS

Edge & AI/ML
Safety & Security
Connectivity
Analog

40+ VIRTUAL DEMOS

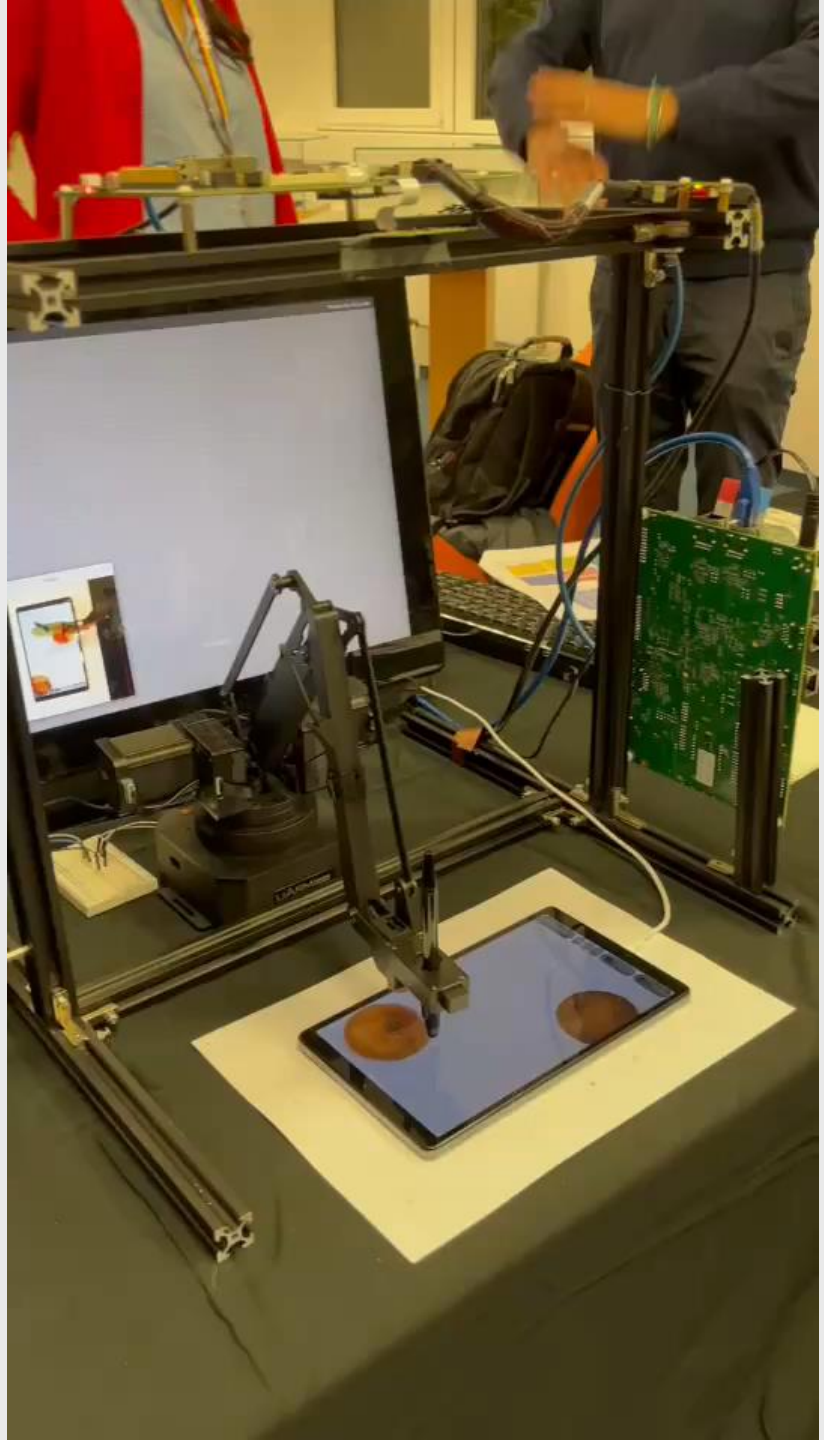
Focus on system solutions
Set up along NXP verticals



SHOWROOM.NXP.COM

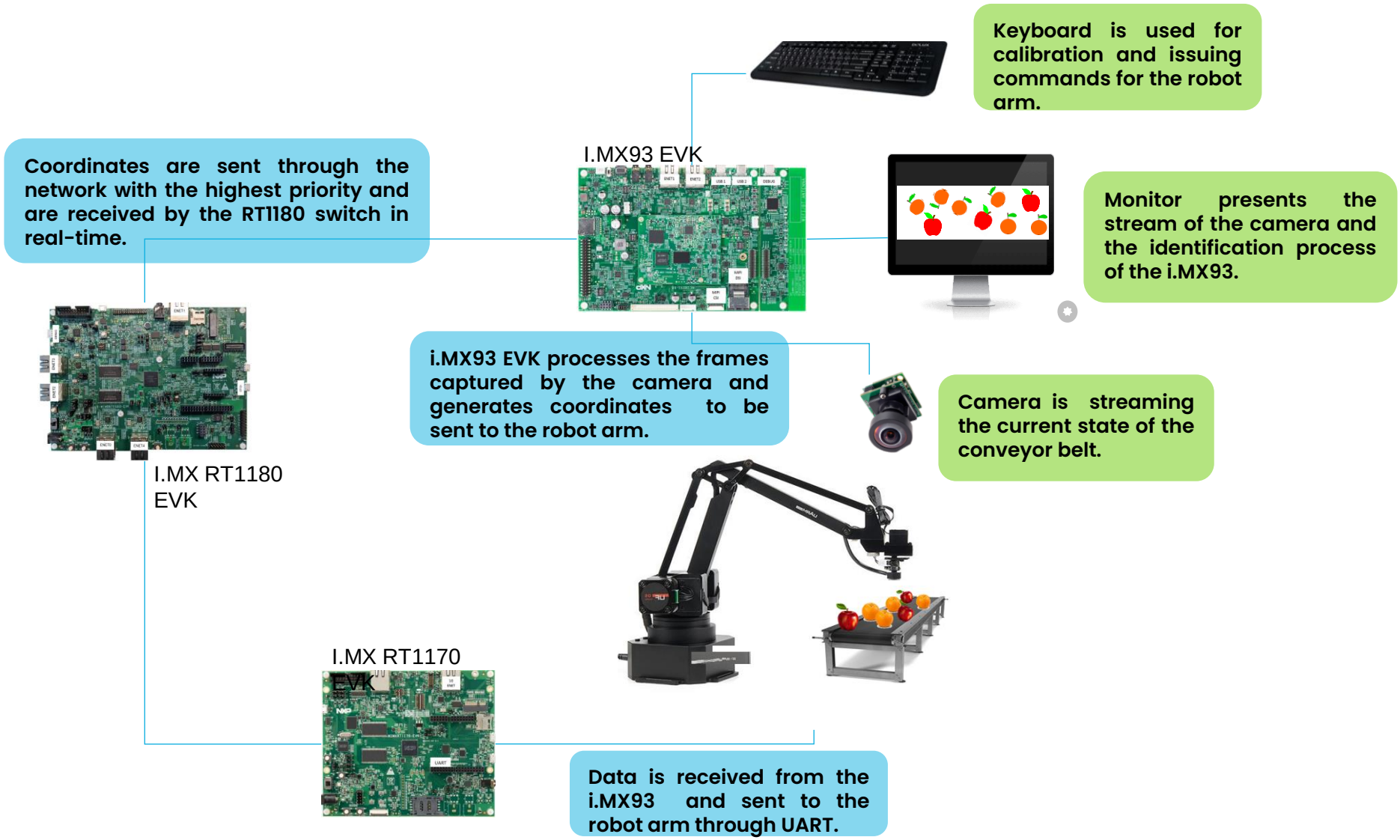
nxp.com

Demos



Fruit Picker demo

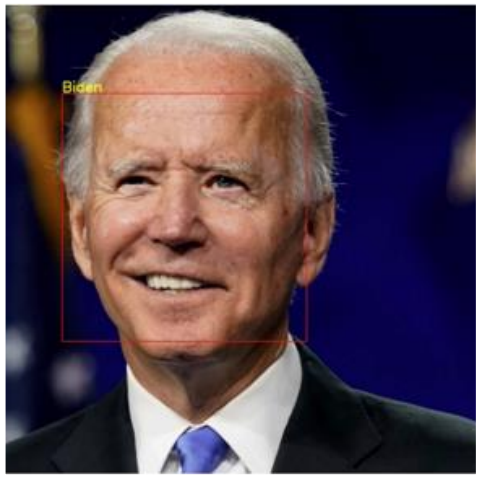
Machine Inspection Demo



eIQ examples + Model Zoo

[GitHub - NXP/eiq-model-zoo: A collection of machine learning models for vision optimized for NXP products]

Face recognition demo



SSD object detection demo

Pose Estimation

[eiq-model-zoo/tasks/vision/pose-estimation at main · NXP/eiq-model-zoo · GitHub]

Hand gesture detection demo

[i.MX Machine Learning User's Guide (nxp.com)]

NXP GoPoint : available at www.nxp.com/imxlinux

The out-of-box experience customers have with NXP MPU innovation

Get running in seconds	Beginner friendly	Poor graphics? No problem.	Like a demo? Reuse it.	Ships on supported reference hardware
Run 20+ demo use cases with a couple of clicks	Demos use easy to understand user interfaces	Launch demos through an altenative text-based interface	Demo source code is open source to jumpstart development	Customers run this application first when they receive hardware

Included Demos

Machine Learning
Object Classification
Object Detection
Pose Detection
Brand Detection
ML Gateway
Face Recognition
DMS Demo
Mask Detection
ML Benchmark

GPU
15 GLES2 Demos
OpenVG 2D Demo

Multimedia
Video Test Demo
Camera using VPU
Multi-Cam Preview
ISP Control Demo
Video Dump Demo
Audio Record
Audio Play
i.MX Voice Control

Please note that all demos listed will not be available on all boards

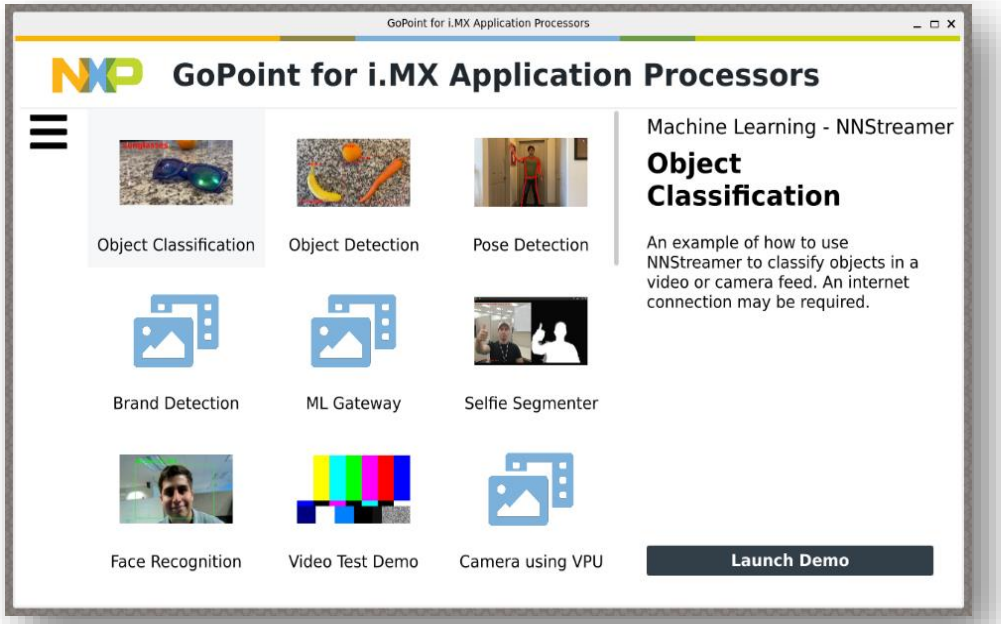
Supported Boards

i.MX 7ULP EVK
All i.MX 8 and 8M EVKs
i.MX 93 EVK
Future i.MX EVKs

Updated every quarter alongside SW team's Linux® release.

Available **today** at **www.nxp.com/imxlinux**

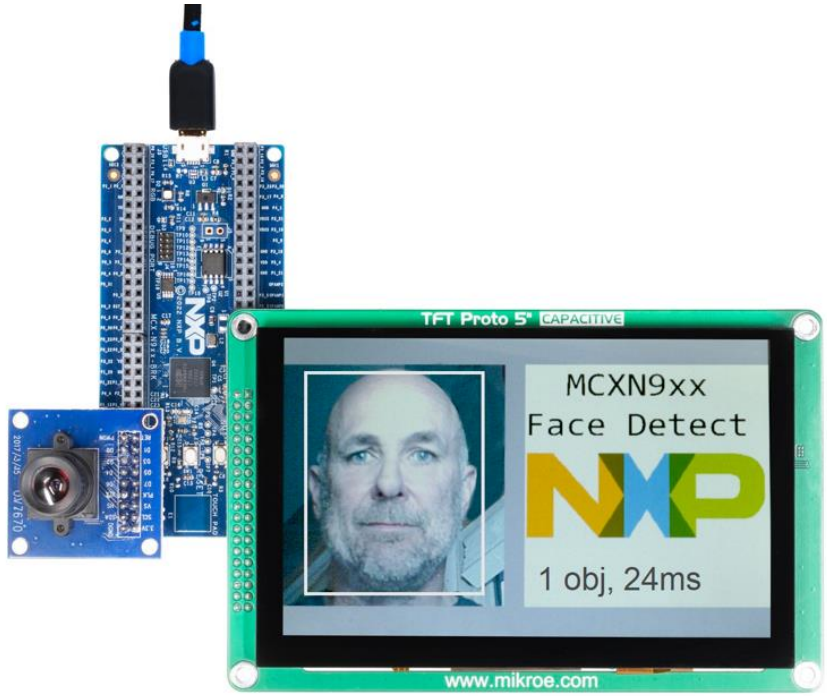
More details at **https://www.nxp.com/docs/en/user-guide/DEXPUG.pdf**



NPU Face Detection Demo

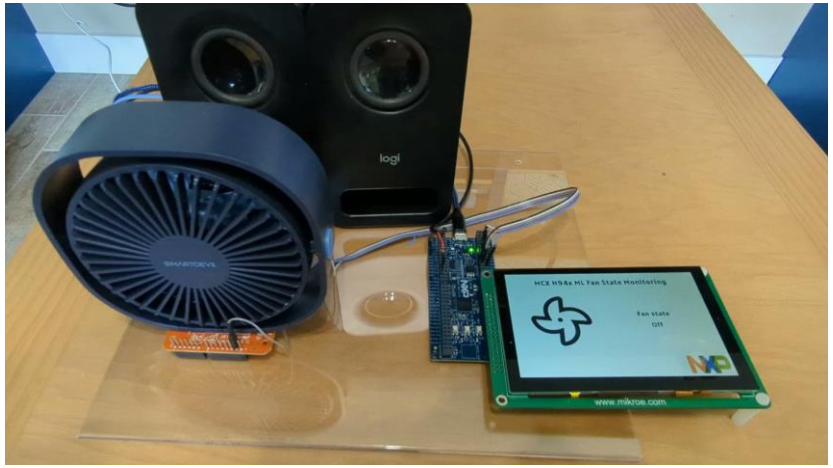
- Face Detection demo on MCX N Series Breakout Board
- Hardware Requirements:
  - **Camera using Smart DMA**
    - OV7670 (w/ optional wide-angle lens replacement)
  - **LCD using FlexIO**
    - Mikroe TFT Proto 5” Capacitive LCD

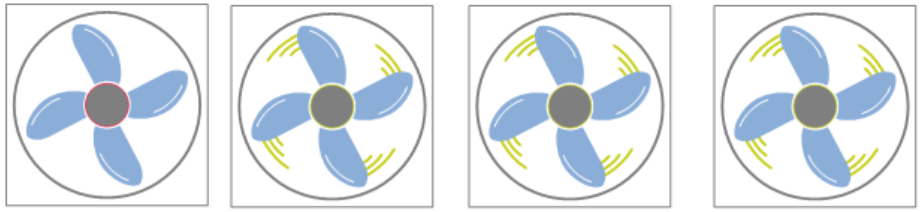
    or
    - NXP Low-Cost LCD



ML State Monitor App SW Pack
Fan State Monitoring And Failure Identification

- Fan State Monitoring and State Identification
- Analyzes vibrations picked up by NXP FXLS8974CF accelerometer on ACCEL-4-CLICK
- Based on ML State Monitor Application Software Pack and adds LCD and Audio features



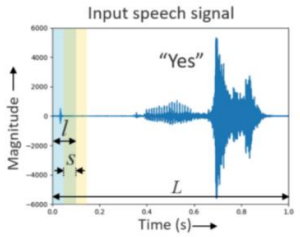


Anomaly Detection with On-Device Training

- Uses classical machine learning to detect anomalies.
- Can train on device the “normal” state and will alert when detects a “non-normal” state

MCUXpresso SDK eIQ Neutron Examples for MCX N

CIFAR-10	Keyword Spotting (KWS)	Label Image	Ultraface
Classifies 32x32 image from camera input into one of 10 categories	Detects specific keywords from microphone input	Classifies 128x128 image from camera input into one of 1000 categories using Mobilenet model. MPP version available as well	Face Detection using Multimedia Processing Pipeline (MPP)



- MPP SDK demos use OV7670 camera and Mikroe TFT Proto 5" Capacitive LCD
- eIQ SDK demos currently use static images/sound pre-loaded into Flash.
- MCXpresso SDK for MCX N9xx 2.14 compatible with Neutron Converter v1.2.0 found in eIQ Toolkit v1.10