恩智浦处理器
机器学习算法工具链

弋 方
Email: fang.yi@nxp.com
恩智浦半导体 微处理器市场经理

2021年8月26日

SECURE CONNECTIONS
FOR A SMARTER WORLD

边缘计算实现机器学习的优势

可靠性和实时性

- 可离线工作
- 避免因网络延迟影响性能

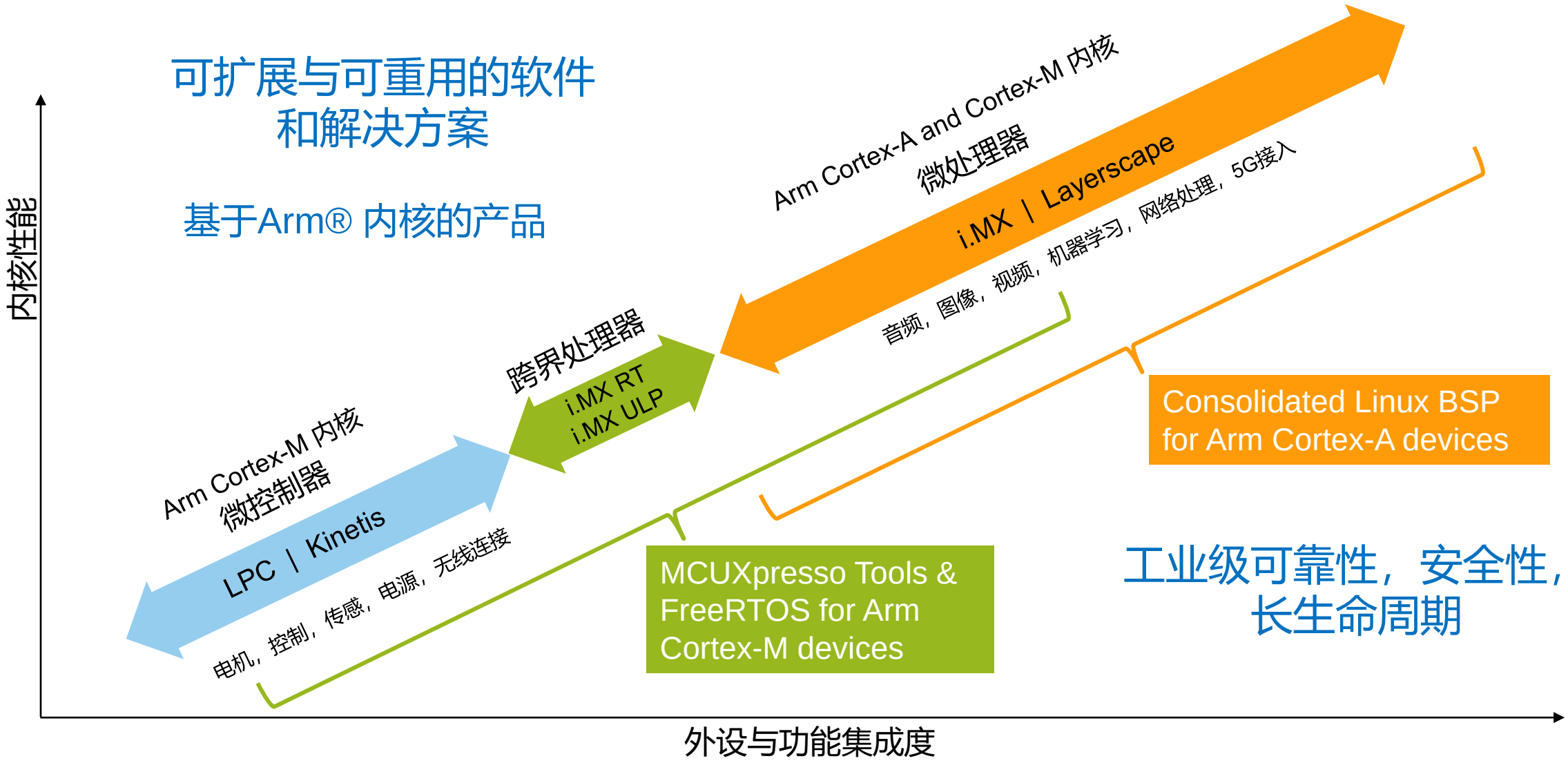
保护数据和隐私

- 避免将完整的原始数据上传云端
- 仅传输识别和理解之后的内容

降低数据通讯成本和云端负载

- 降低云端计算量
- 降低云端数据存储量
- 降低对通信带宽的要求

恩智浦提供广泛的微控制器和微处理器产品用于边缘计算



恩智浦完整的软硬件产品和生态

EDGEVerse

Industrial Edge	IoT Edge	Network Edge

垂直应用领域

CLOUD PLATFORM EXAMPLES

aws · Microsoft Azure · Alibaba Cloud · Google Cloud · Baidu 百度 · salesforce · PREDIX · IBM Cloud

云服务

NXP eIQ™ Machine Learning	Audio Solutions	Key injection EdgeLock-2-Go	Turnkey solutions & SW services	Partners
Edge Intelligence		Security & Provisioning		Device Management

软件赋能与服务

SECURITY | CONNECTIVITY | GRAPHICS | AUDIO / VIDEO | NETWORKING

中间件与参考代码

ubuntu · yocto PROJECT · CentOS · Open Industrial Linux

MCUXpresso

Android | Common Linux Platform | Zephyr | FreeRTOS | Azure RTOS

操作系统，BSP

i.MX | MPUs | Crossover | MCUs

芯片

EIQ™机器学习软件开发环境

嵌入式开发人员

数据分析师

机器学习专家

eIQ 工具集

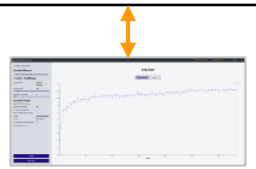
eIQ 用户接口

模型优化工具

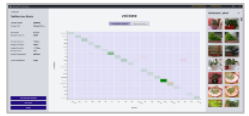
提供用户自己的数据处理流程

数据整理

模型选择



模型训练、量化、优化



模型验证、部署

提供用户自己的模型工作流程

PyTorch Caffe2 ONNX

PyTorch Caffe2 ONNX

eIQ inference

Applications running on EdgeVerse processors

异常检测 | 声音识别 | 人脸识别 | 多目标分类 | 复杂情况诊断

eIQ inference with…

TFLite | TFMicro | Glow | DeepView™

Arm® Cortex®-M, DSP

TFLite | Arm NN | ONNX Runtime | DeepView™

Arm Cortex-A, GPU, DSP, NPU, microNPU

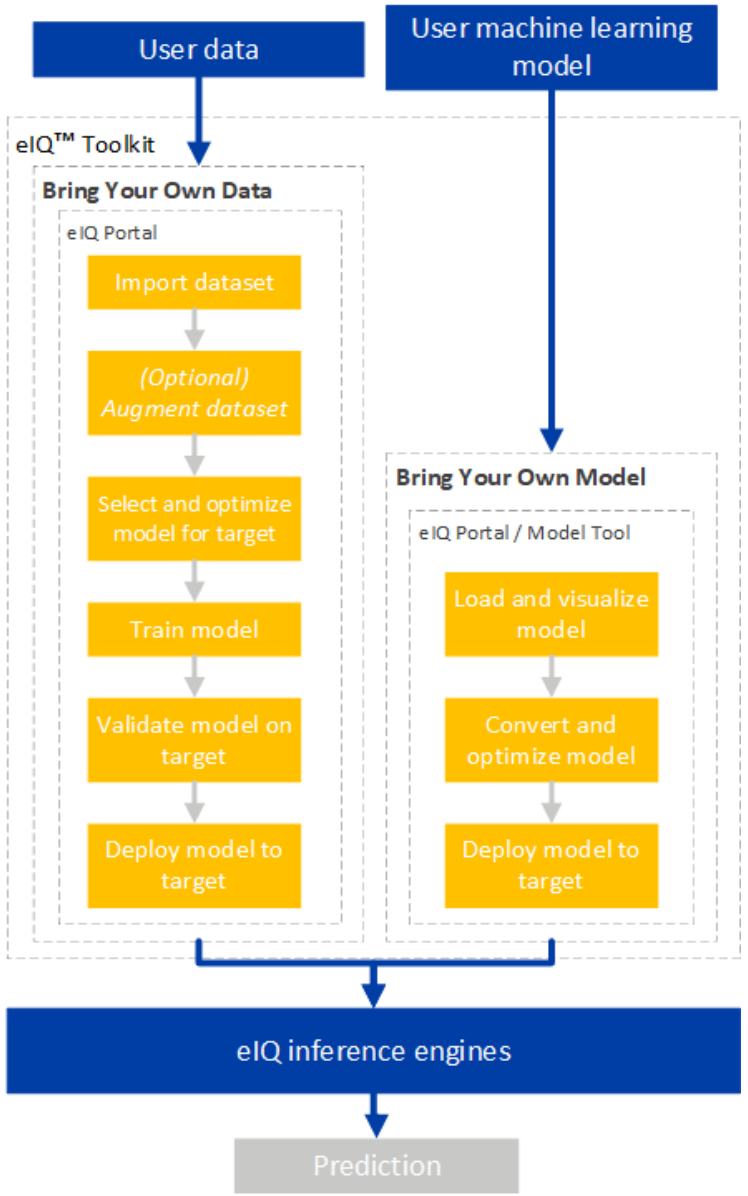
eIQ 工具箱

SECURE CONNECTIONS
FOR A SMARTER WORLD

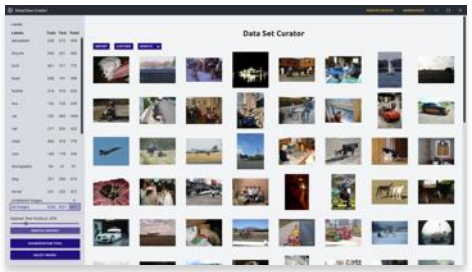
eIQ 工具箱

- eIQ工具箱，是恩智浦的机器学习软件开发环境，适用于开发机器学习算法，并将其部署于恩智浦微控制器，微处理器等不同的硬件平台上

- eIQ工具箱，包含两种典型使用方法，分别是
  - *Bring Your Own Data (BYOD) 基于原始数据*
  - *Bring Your Own Model (BYOM) 基于已有模型*

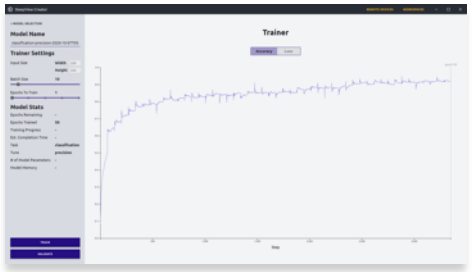


基于原始数据： DeepView™ Bring Your Own Data Workflow (BYOD)

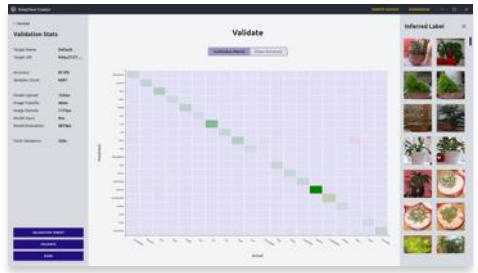


数据导入与标定

模型选择，硬件平台选择



模型训练，优化，量化



模型验证与测试

基于已有模型: DeepView™ Bring Your Own Model Workflow (BYOM)

PyTorch

ONNX

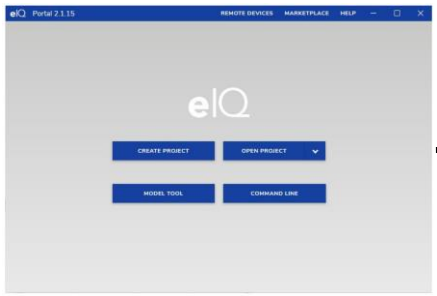
Caffe2

公有或私有模型

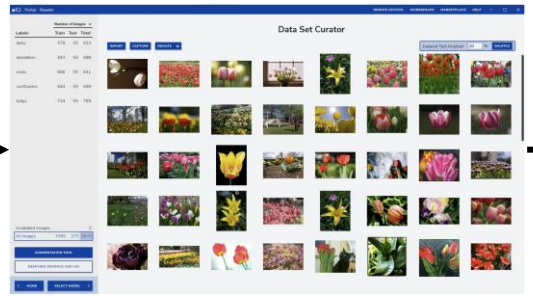
模型格式转换，优化，量化

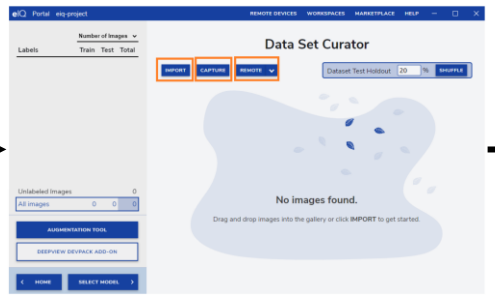
嵌入式板级推理实现

基于原始数据： DeepView™ Bring Your Own Data Workflow (BYOD)

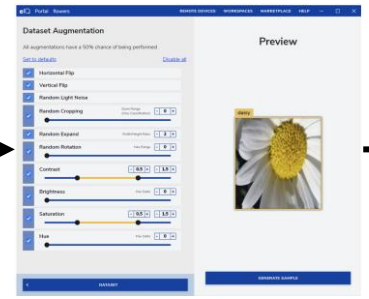


创建工程

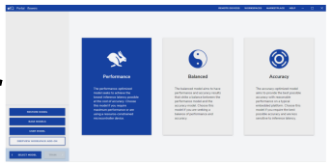




数据导入与标定

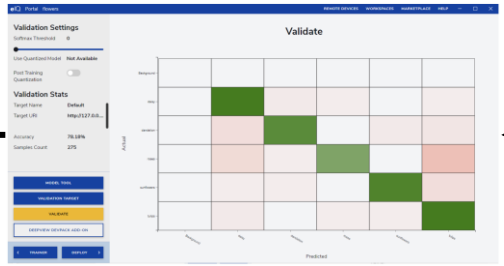


数据增强



模型选择

模型训练



模型验证与测试

模型分析

目标设备实现推理

基于已有模型: DeepView™ Bring Your Own Model Workflow (BYOM)

Source/ Destination	RTM	TFLite (MLIR/TOCO)	ONNX	RTM q	TF:OTE q	ONNX q
1.x Pb	Y	Y	Y	N	Y	Y
Saved Model (Folder/TAR)	Y	Y	Y	Y	Y	Y
Keras (h5)	Y	Y	Y	Y	Y	Y
RTM		N	N		N	N
TFLite (MLIR/TOCO)	Y		Y	N	N	Y
ONNX	Y	Y		N	Y	Y
TFLite q (MLIR/TOCO)	Y		Y	N		
ONNX q	Y	Y		N		

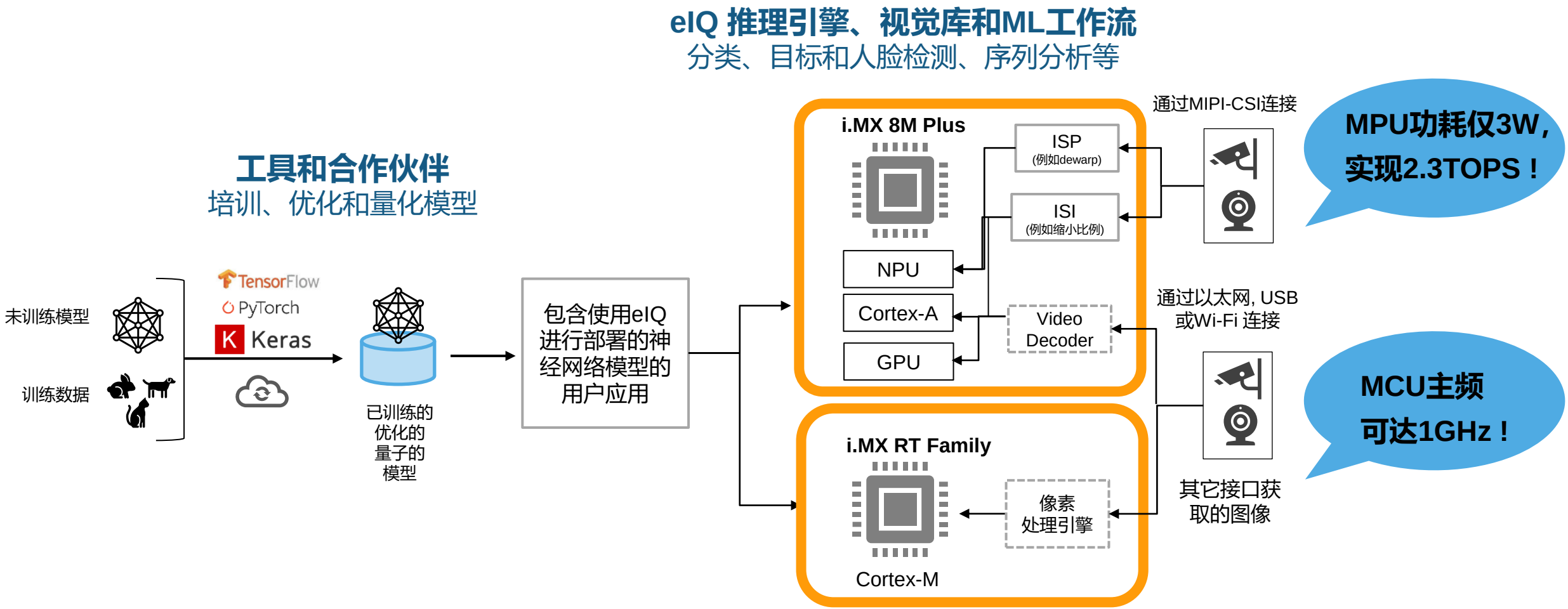
*Notes: (1) **RTM** is for DeepViewRT*
*(2) **q** is for quantized*

获取 eIQ 工具包

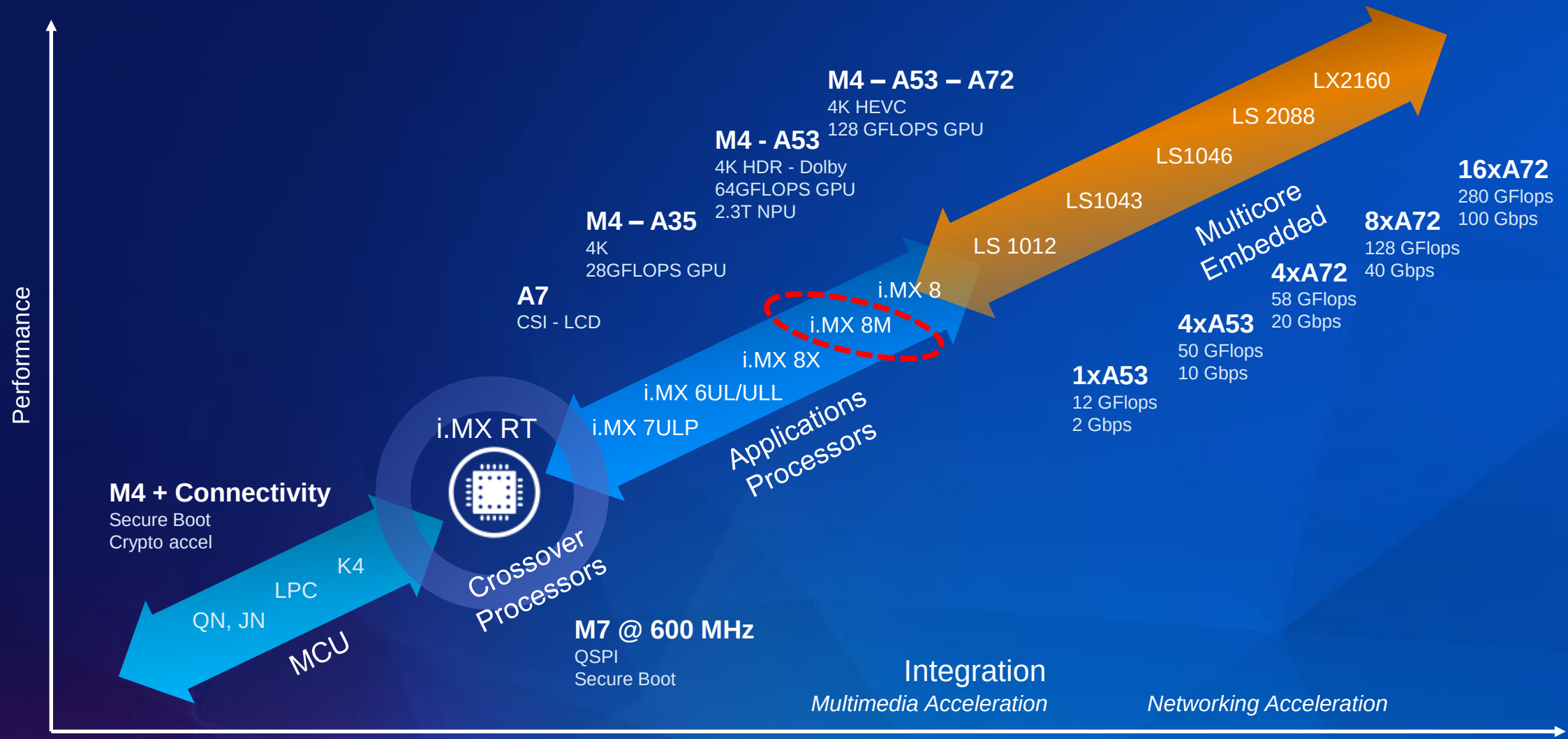
eIQ Toolkit can be found at www.nxp.com/eiq/toolkit

- eIQ Toolkit package includes the eIQ Portal GUI as well as command line tools for model conversion and creation
- The eIQ Toolkit consists of three key components:
  - eIQ Portal
  - eIQ Model Tool
  - eIQ Command-line Tools

恩智浦eIQ工具箱支持多种硬件平台：从MCU到MPU



恩智浦完整的嵌入式处理器平台



MPU平台：i.MX 8M PLUS

Quad Arm® Cortex-A53 up to 1.8 GHz (up to 20,988 DMIPS)
- ARMv8 fully 64-bit capable, 512KB L2 cache (ECC)

Arm Cortex-M7 up to 800MHz with 512KB RAM (ECC)

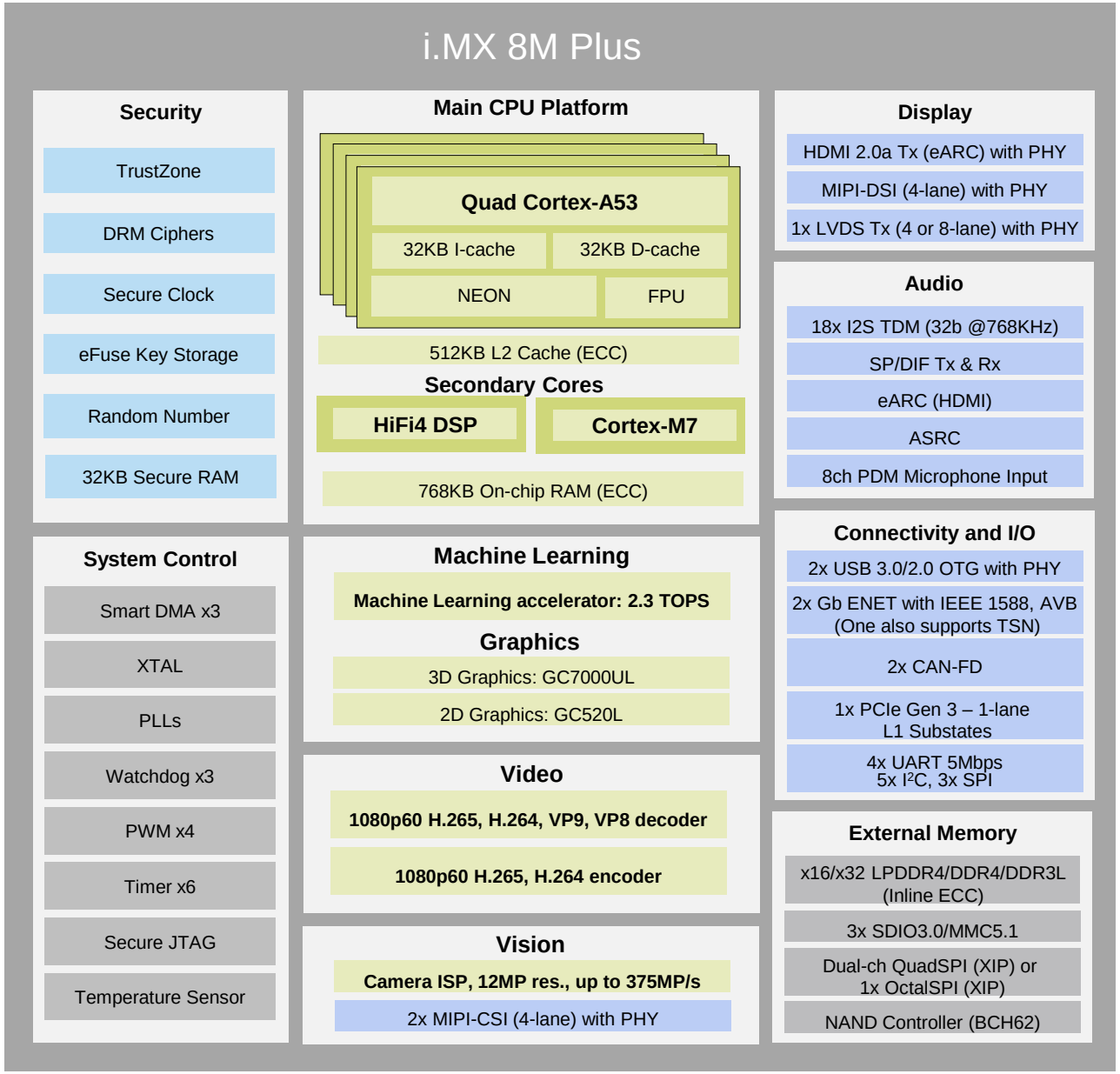
HiFi4 DSP up to 800MHz with 256KB RAM (ECC)

Machine Learning (ML) accelerator: 2.3 TOPS performance

Operating System targets: Linux OS, Android OS, HiFi4 SDK, FreeRTOS

Qualification: Consumer (0C to +95C); Industrial (-40C to +105C)

- x32/x16 LPDDR4/DDR4/DDR3L (Inline ECC)
- 3x SDIO3.0/eMMC5.1
- Dual-channel QuadSPI (XIP) or 1x OctalSPI
- NAND Controller (BCH62)
- SPI NAND
- GC7000UL (3D GPU, 2-shaders, OpenGL® ES 3.1, Vulkan®, Open CL™ 1.2 FP)
- GC520L (2D GPU, OpenVG™ 1.1)
- 1080p60 H.265, H.264, VP9, VP8 decode
- 1080p60 H.265, H.264, VP8 encode
- Display controllers (up to 3 simultaneous displays):
- 1x HDMI 2.0a Tx (eARC) with PHY
- 1x LVDS Tx (4 or 8-lane) with PHY
- 1x MIPI-DSI (4-lane) with PHY
- Camera (up to 2 cameras): 2x MIPI-CSI (4-lane) with PHY
- ISP 12MP resolution, 2x187MP or 1x375MP input rate, HDR, Dewarp
- Audio: 18x I²S TDM (32-bit @ 768KHz), DSD512, SP/DIF Tx + Rx
- 8 channel PDM microphone input
- eARC, ASRC
- 2x USB 3.0/2.0 Type C with PHY
- 1x PCIe Gen 3.0 (1 lane)
- 2x Gigabit Ethernet with IEEE 1588, EEE and AVB
- 2x CAN-FD



MPU硬件平台： i.MX 8M Plus 采用 TensorFlow Lite INT8 模型性能分析

Hardware: i.MX 8M Plus EVK

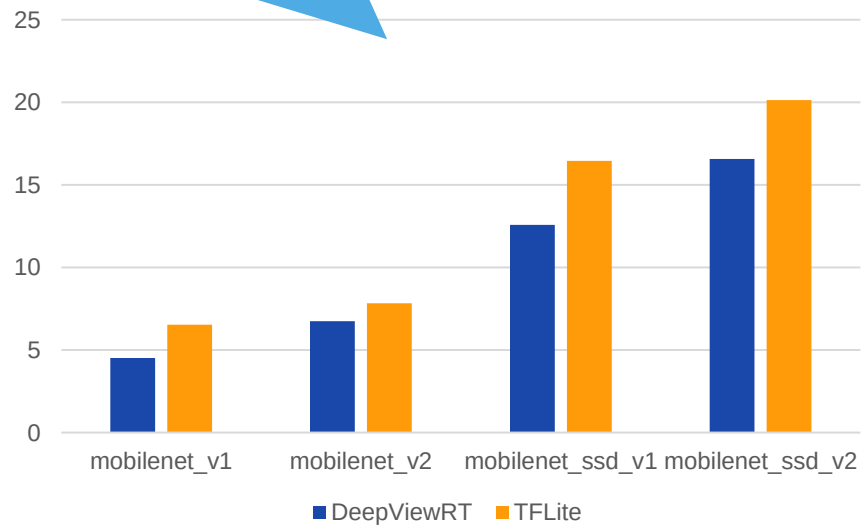
Software: Linux_IMX_5.4.70-2.3.2, TFLite 2.4.0, DeepViewRT 2.4.24

Models INT8 model performance on i.MX 8M Plus CPU and NPU	Time per inference (ms)	
	1x A53 (1.8 GHz) CPU	NPU
inception_v1_INT8	343.48	5.80
inception_v2_INT8	529.75	7.35
inception_v3_INT8	1266.28	17.12
inception_v4_INT8	2658.42	33.17
mobilenet_v1_100_224_INT8	153.49	2.72
mobilenet_v2_100_224_INT8	119.68	3.02
deeplabv3_mnv2_pascal_quant	1089.63	200.08
ssd_mobilenet	210.42	12.76
ssd_mobilenet_v2_coco_quant	273.32	16.29
ssd_mobilenet_v2_face_quant	269.87	7.31

Models INT8 model performance on i.MX 8M Plus NPU	Time per inference (ms)	
	DeepViewRT	TFLite
mobilenet_v1_1.0_224 (Per channel)	4.51	6.53
mobilenet_v1_1.0_224 (Per tensor)	3.38	-
mobilenet_v2_1.0_224 (Per channel)	6.75	7.84
mobilenet_v2_1.0_224 (Per tensor)	4.01	-
mobilenet_ssd_v1_300(Per channel)	12.59	16.46
mobilenet_ssd_v1_300(Per tensor)	8.67	-
mobilenet_ssd_v2_300(Per channel)	16.57	20.15
mobilenet_ssd_v2_300(Per tensor)	10.92	-

专用硬件加速器NPU可提高约50倍的推理性能！

eIQ DeepViewRT 相对于TFlite，可以提高约20%

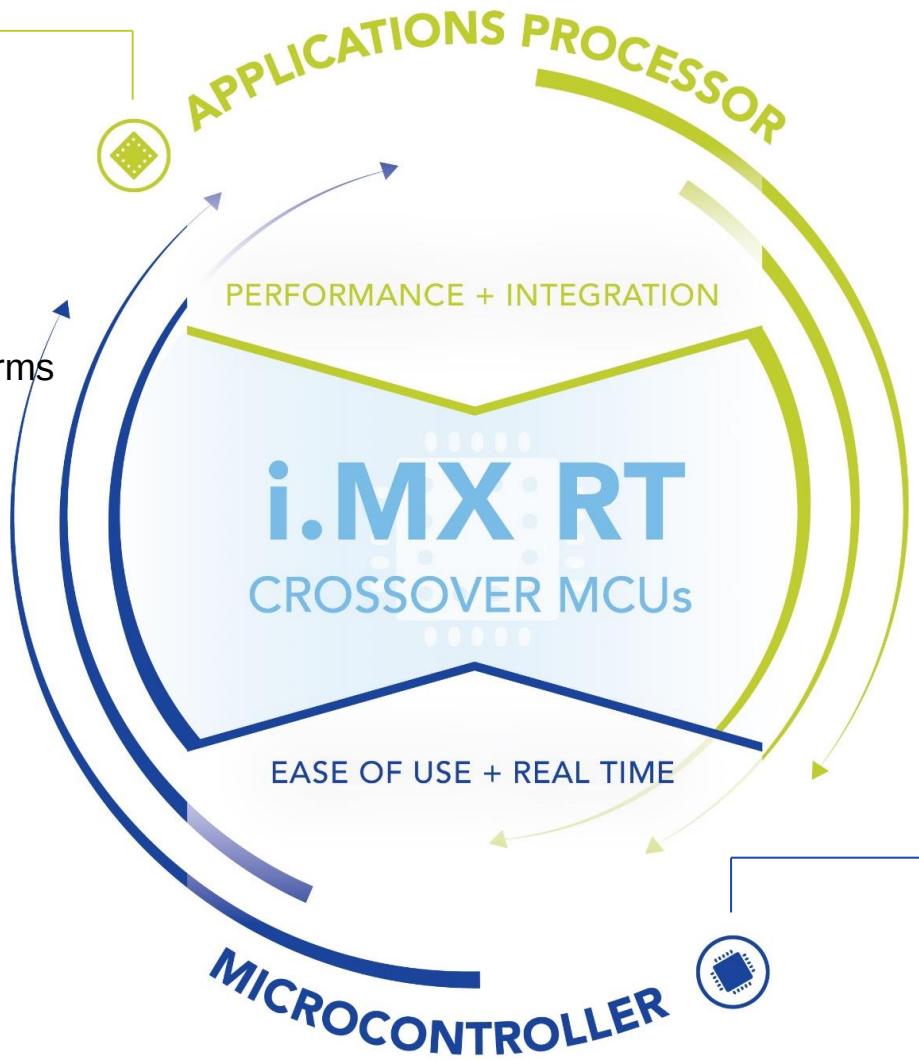


恩智浦跨界处理器i.MX RT系列

i.MX
APPLICATIONS PROCESSORS

- Arm® Cortex®-A class and Cortex-M cores
- 600 MHz to 2 GHz performance
- Rich HMI experience
- Full open-source OS platforms

Linux ANDROID

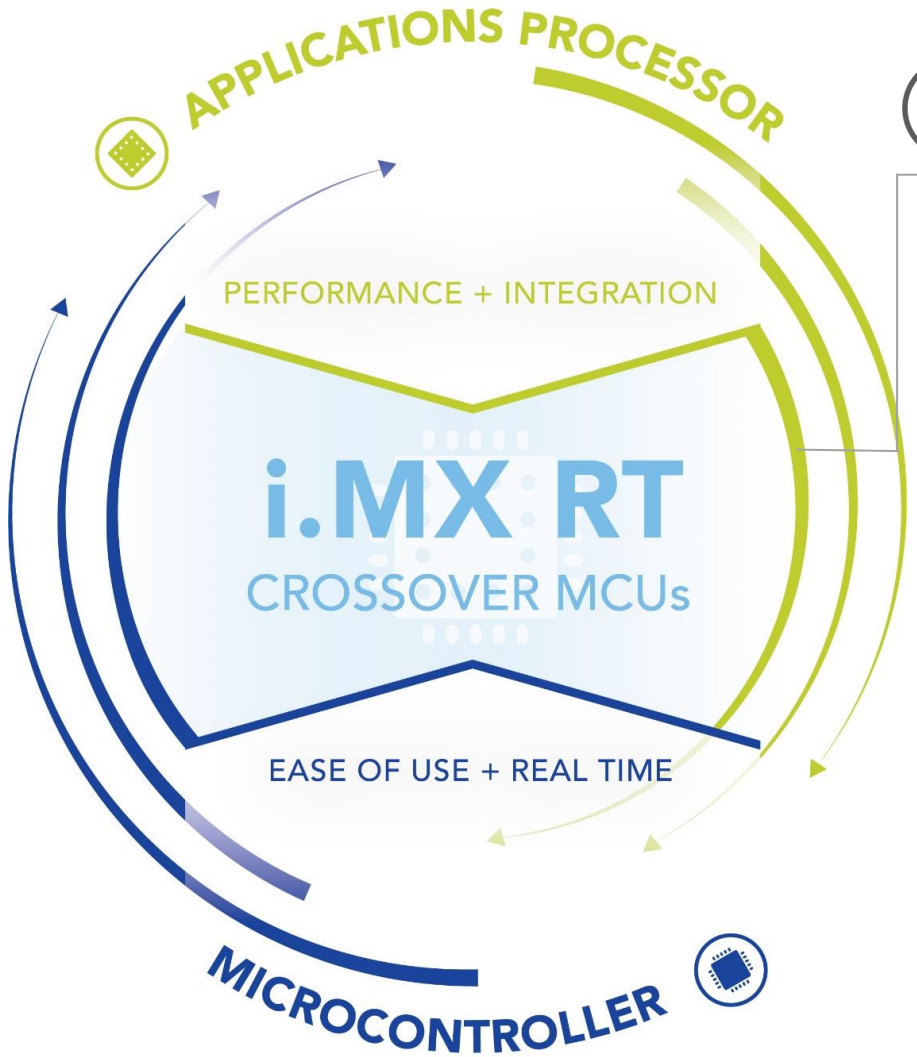


- Arm® Cortex®- M cores
- Performance up to 300 MHz
- Embedded memory
- Comprehensive tool support
- Extensive NXP and partner software

IAR SYSTEMS KEIL Tools by ARM SEGGER

KINETIS & LPC
MCUs

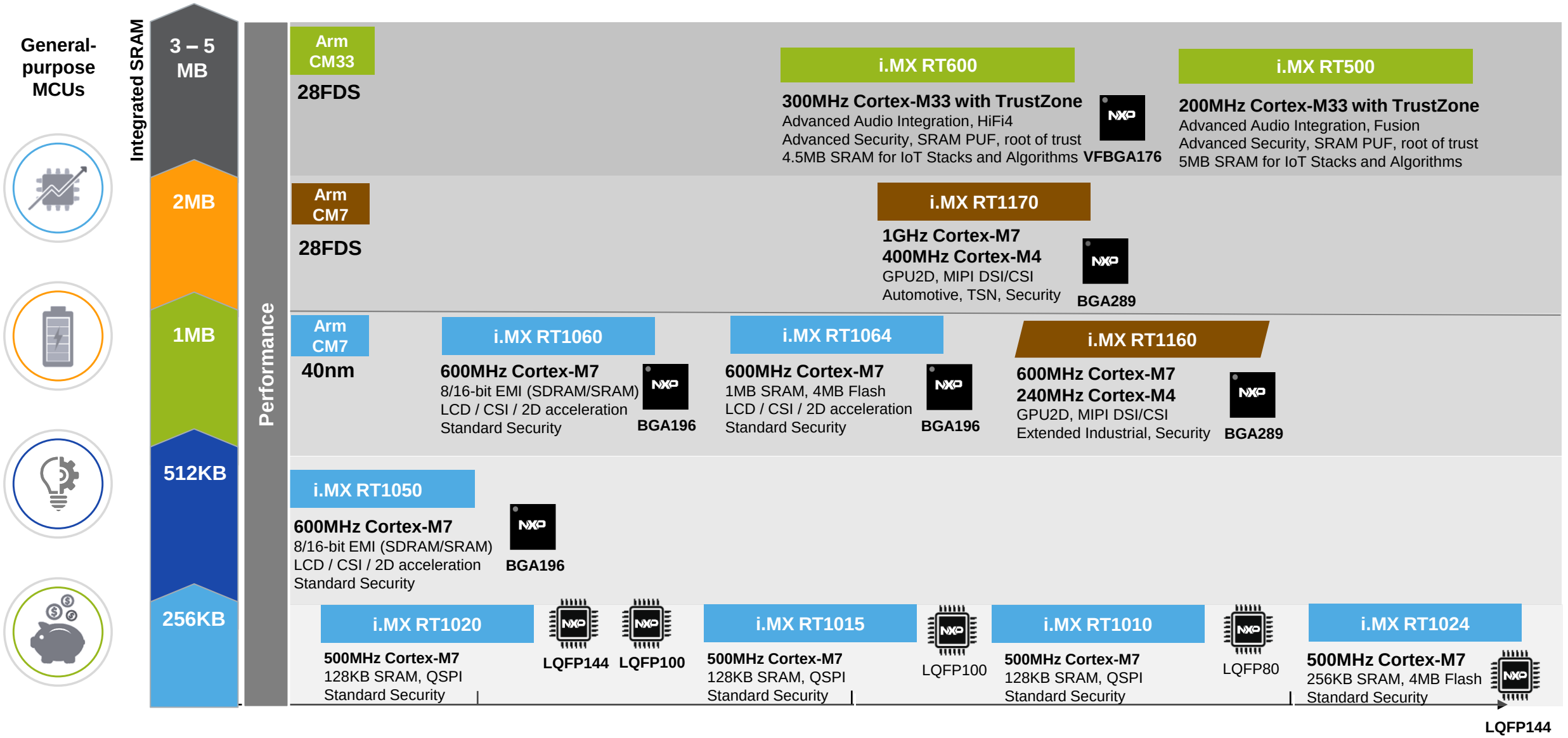
恩智浦跨界处理器i.MX RT系列



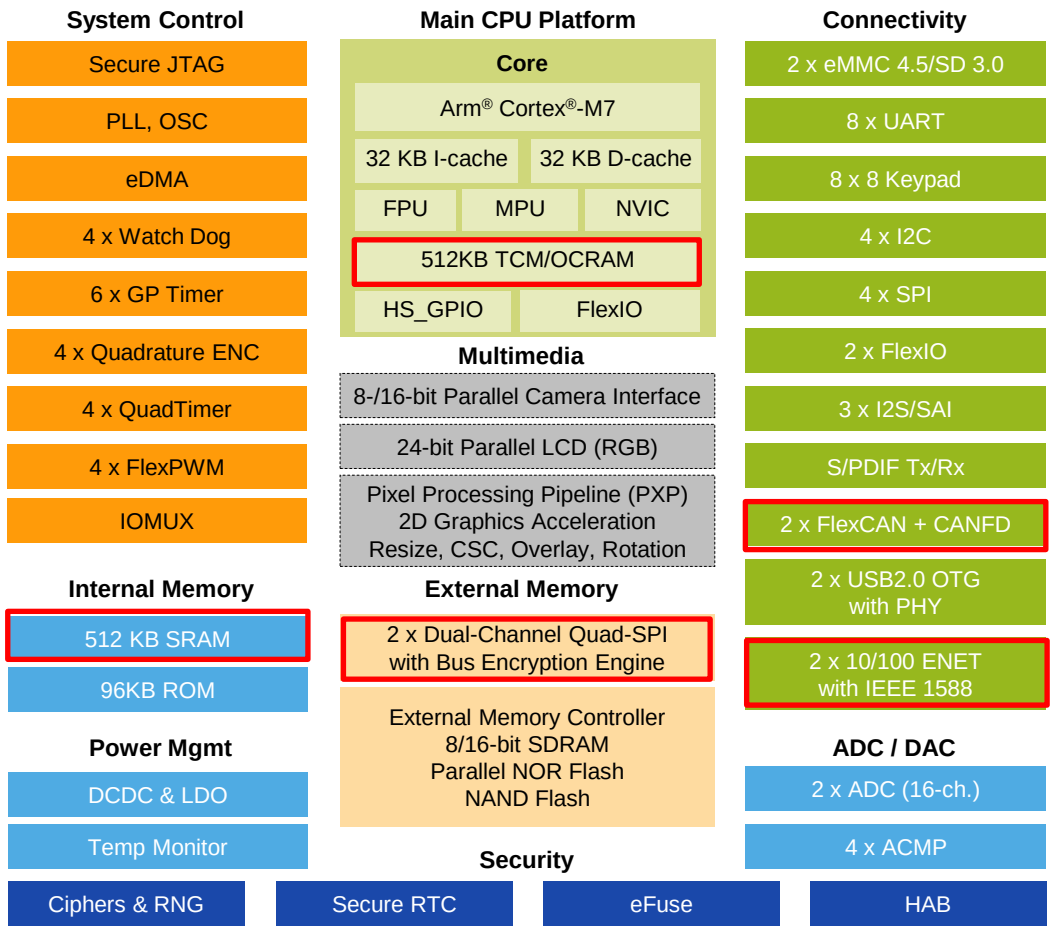
i.MX RT
CROSSOVER MCUs

- Arm® Cortex® -M cores
- 1 GHz performance
- High integration
- Deterministic instructions
- Short latency
- Comprehensive tool support
- Extensive NXP and partner software

恩智浦i.MX RT 系列跨界处理器



i.MX RT106X 主流性能跨界处理器



High Performance Real Time System

- Cortex-M7 up to 600MHz , 50% faster than any other existing M7 products
- 20ns interrupt latency, a TRUE Real time processor
- 512KB SRAM + 512KB TCM/OCRAM

Rich Peripheral

- Motor Control: Flex PWM X 4, Quad Timer X 4, ENC X 4
- 2x USB, 2x SDIO, 2x CAN + 1x CANFD, 2x ENET with 1588, 8xUART, 4x SPI, 4X I2C
- 8/16-bit CSI interface and 8/16/24-bit LCD interface
- 2x Qual-SPI interface, with Bus Encryption Engine
- Audio interface: 3x SAI/ SPDIF RX & TX/ 1x ESAI

Security

- TRNG&PRNG(NIST SP 800-90 Certified)
- 128-AES cryptography
- Bus Encryption Engine: Protect QSPI Flash Content

Specifications

- Package: MAPBGA196 | 10x10mm^2, 0.65mm pitch (130 GPIOs)
  MAPBGA196 | 12x12mm^2, 0.8mm pitch
- Temp / Qual: -40 to 105°C (Tj) Industrial / 0 to 95°C (Tj) Consumer

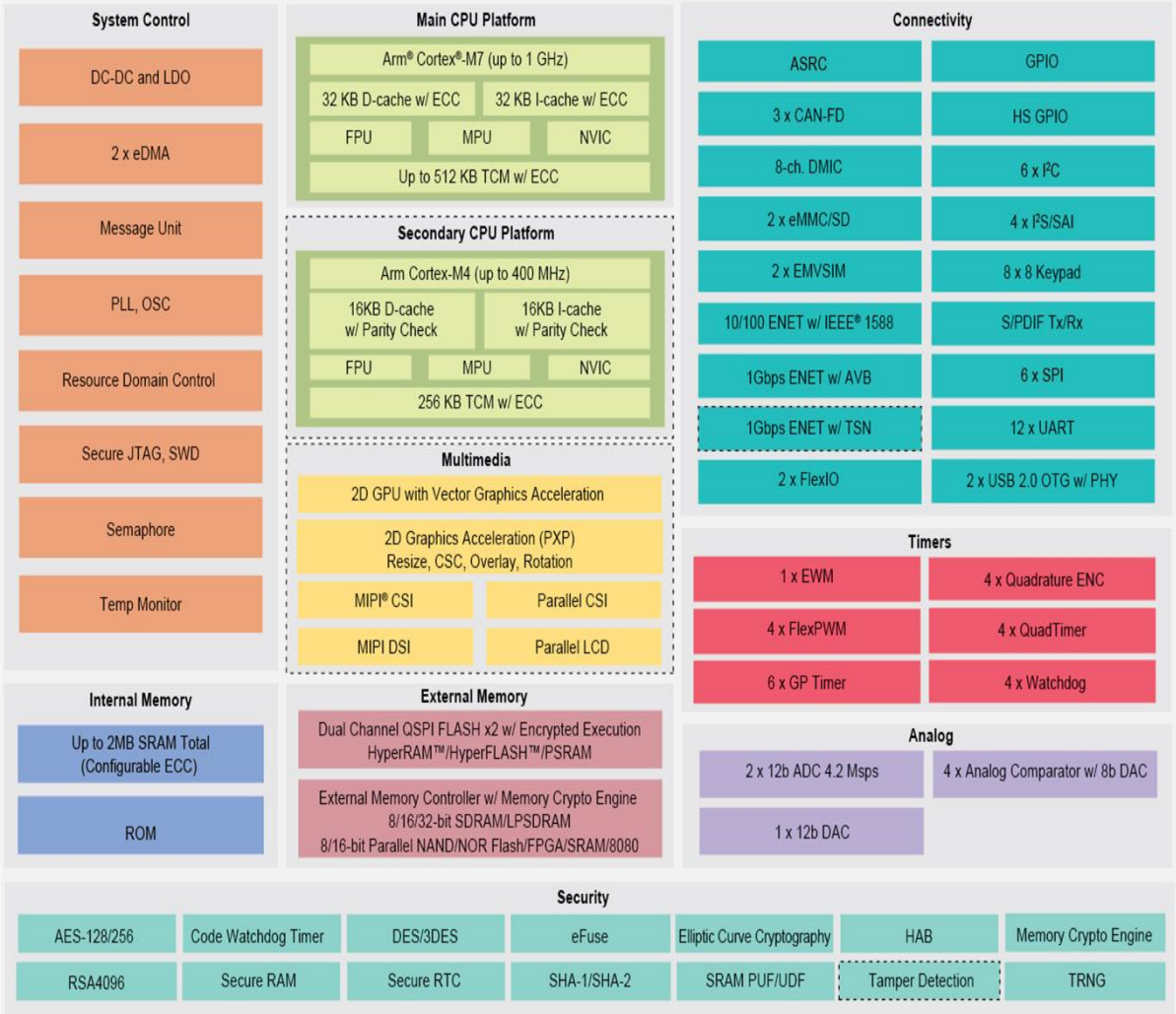
Ease of Use

- MCUXpresso with SDK, IAR, Keil
- FreeRTOS
- Comprehensive ecosystem

Low BOM Cost

- Competitive Price
- Fully integrated PMIC with DC-DC
- QSPi and SDRAM interface

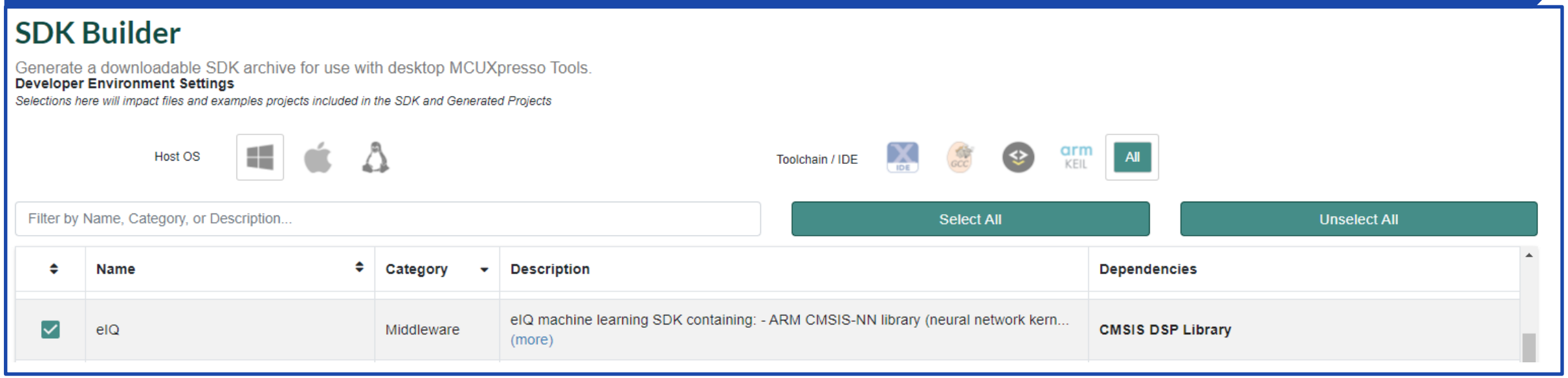
I.MX RT117X 高性能跨界处理器



- Arm Cortex-M7 processor, **1 GHz**, 32KB/32KB L1 Cache, 512KB TCM
- Arm Cortex-M4 processor, **400MHz**, 16KB/16KB L1 Cache, 256KB TCM
- **2MB on-chip SRAM** (including TCM for CPU core)
- Parallel LCD Display up to WXGA (1280x800) @60fps
- 8/16/24-bit Parallel Camera Sensor Interface
- 2-lane MIPI CSI and 2-lane MIPI DSI
- **2D GPU & Graphics Accelerator**
- 8/16/32-bit SDRAM controller up to 200MHz
- 8/16-bit Parallel NOR FLASH / NAND FLASH / SRAM
- 2x Dual-channel FlexSPI Interfaces with On-the-Fly decryption, supporting serial NOR, serial NAND, HyperBUS devices
- 2x eMMC 5.0/SD 3.0/SDIO Port
- 2x USB 2.0 OTG, HS/FS, Device or Host with PHY
- Audio: 4x I²S/SAI, 1x S/PDIF Tx/Rx, ASRC, 8-ch digital microphone input
- **3x ENET: 1Gbps ENET w/ AVB + 10/100 ENET w/ IEEE 1588 + 1Gbps ENET w/ TSN**
- 2x 12-bit ADC, 4.2Msample/s, up to 20 input channels total
- 4x Analog comparator, 1x 12-bit DAC
- Full PMU Integration, DCDC+LDOs

MCU硬件平台： 中间件已集成于MCUXpresso SDK

eIQ is included as part of MCUXpresso SDK. Select the “eIQ” middleware in MCUXpresso SDK Builder



eIQ Inference Engines in MCUXpresso SDK 2.9

	RT595	RT685	RT1050	RT1060	RT1064	RT1160	RT1170
CMSIS-NN			■	■	■		■
TensorFlow Lite			■	■	■		■
TensorFlow Lite for Microcontrollers			■	■	■		■
Glow		■	■	■			■
DeepViewRT					■		*

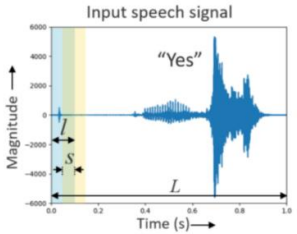
eIQ Inference Engines in MCUXpresso SDK 2.10 (July 2021)

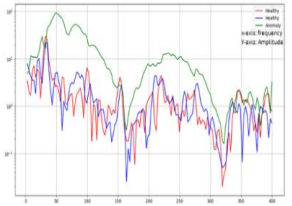
	RT595	RT685	RT1050	RT1060	RT1064	RT1160	RT1170
CMSIS-NN							
TensorFlow Lite							
TensorFlow Lite for Microcontrollers	■	■	■	■	■	■	■
Glow		■	■	■	■	■	■
DeepViewRT			■	■	■	■	■

* Only some DeepViewRT support in standard SDK. Need special download link for DeepViewRT Label Image example

MCU硬件平台： MCUXpresso SDK 2.10 参考实例

CIFAR-10	Keyword Spotting (KWS)	Label Image	Anomaly Detection	Handwritten Digit Detection
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	Use FRDM-STBC-AGM01 sensor board for accelerometer anomaly analysis (Select “agm01” board)	Determine handwritten digit values using LeNET MNIST model



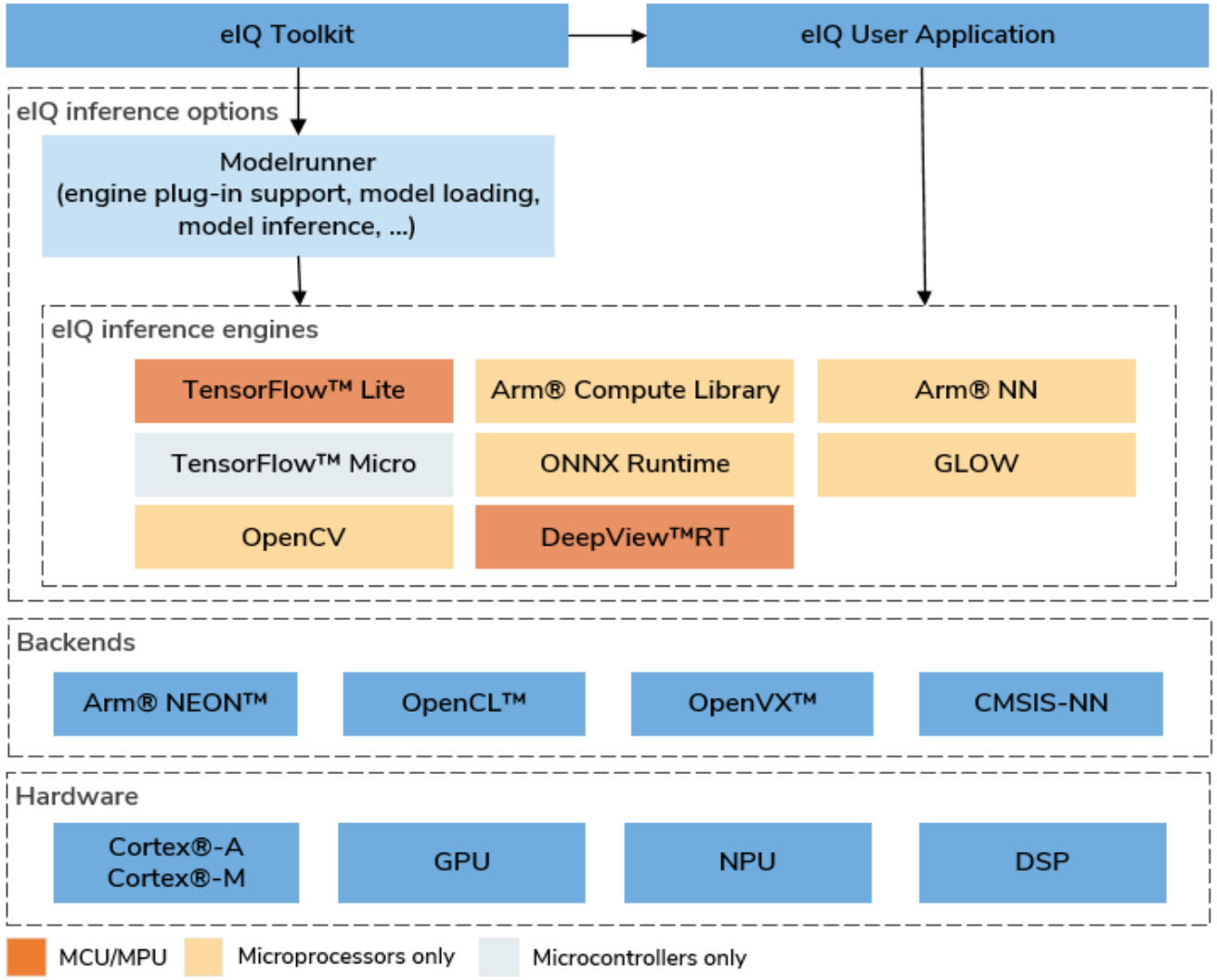


	i.MX RT595	i.MX RT685	i.MX RT1050	i.MX RT1060	i.MX RT1064	i.MX RT1160	i.MX RT1170
TFLM CIFAR10							
TFLM KWS							
TFLM Label Image							
TFLM Anomaly Detection							
Glow MNIST							
Glow CIFAR10							
DeepViewRT Label Image							
DeepViewRT ModelRunner							
DeepViewRT ModelRunner-Glow							

模型运行器 (MODEL RUNNER)

- 模型运行器（ModelRunner） 处理通信并充当中间推理引擎，将计算委托给其他推理引擎。

- 模型运行器（ModelRunner）为了提升执行效率，会根据实际可用的硬件单元，使用更底层的适配技术。



SECURE CONNECTIONS
FOR A SMARTER WORLD