恩智浦i.MX 8M PLUS 应用处理器
边缘计算与机器学习

PUBLIC

弋 方
恩智浦半导体 微处理器市场经理

2021年6月29日

SECURE CONNECTIONS
FOR A SMARTER WORLD

边缘计算实现机器学习的优势

可靠性和可靠性

- 可离线工作
- 避免因网络延迟影响性能

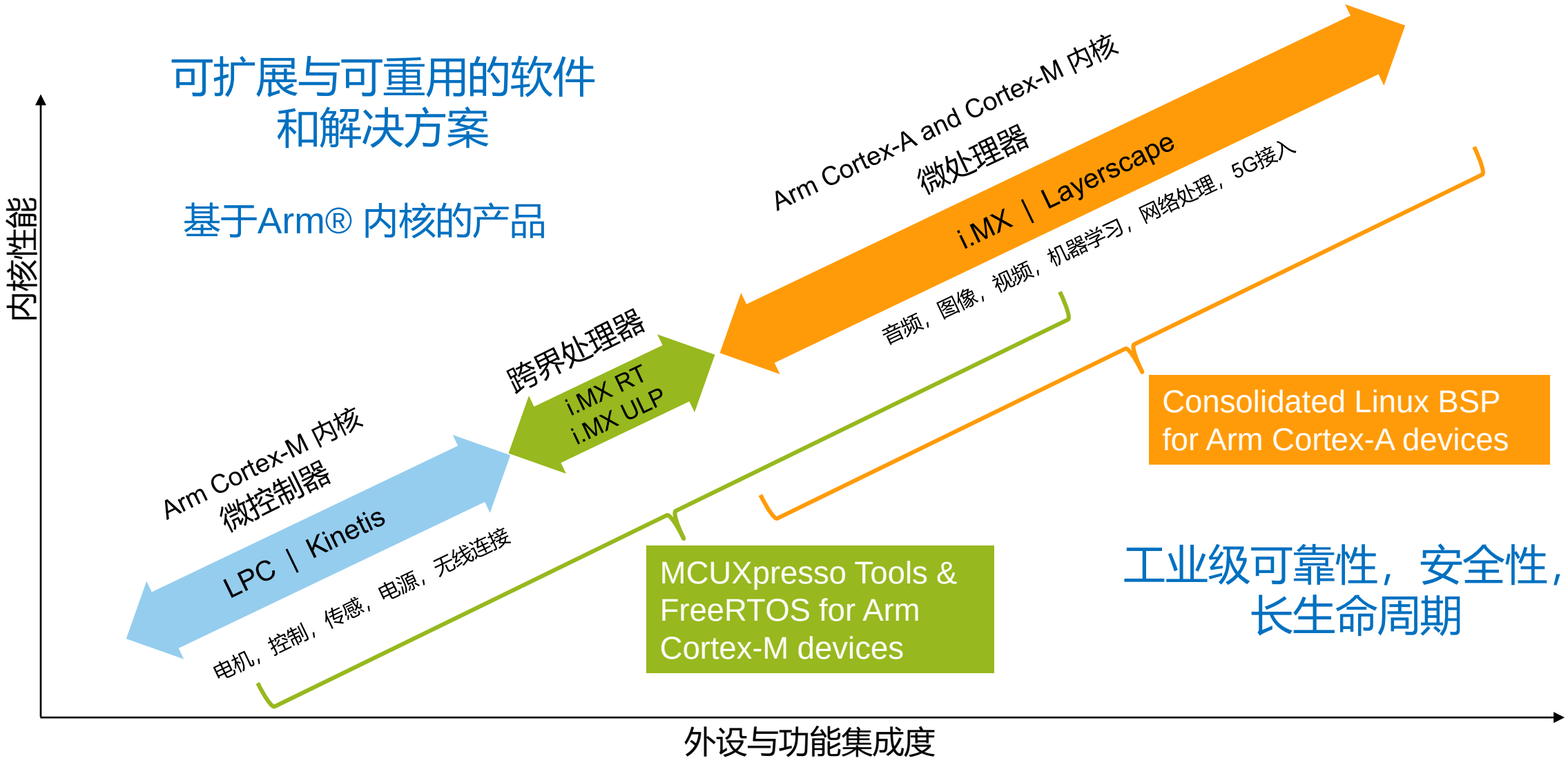
保护数据和隐私

- 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

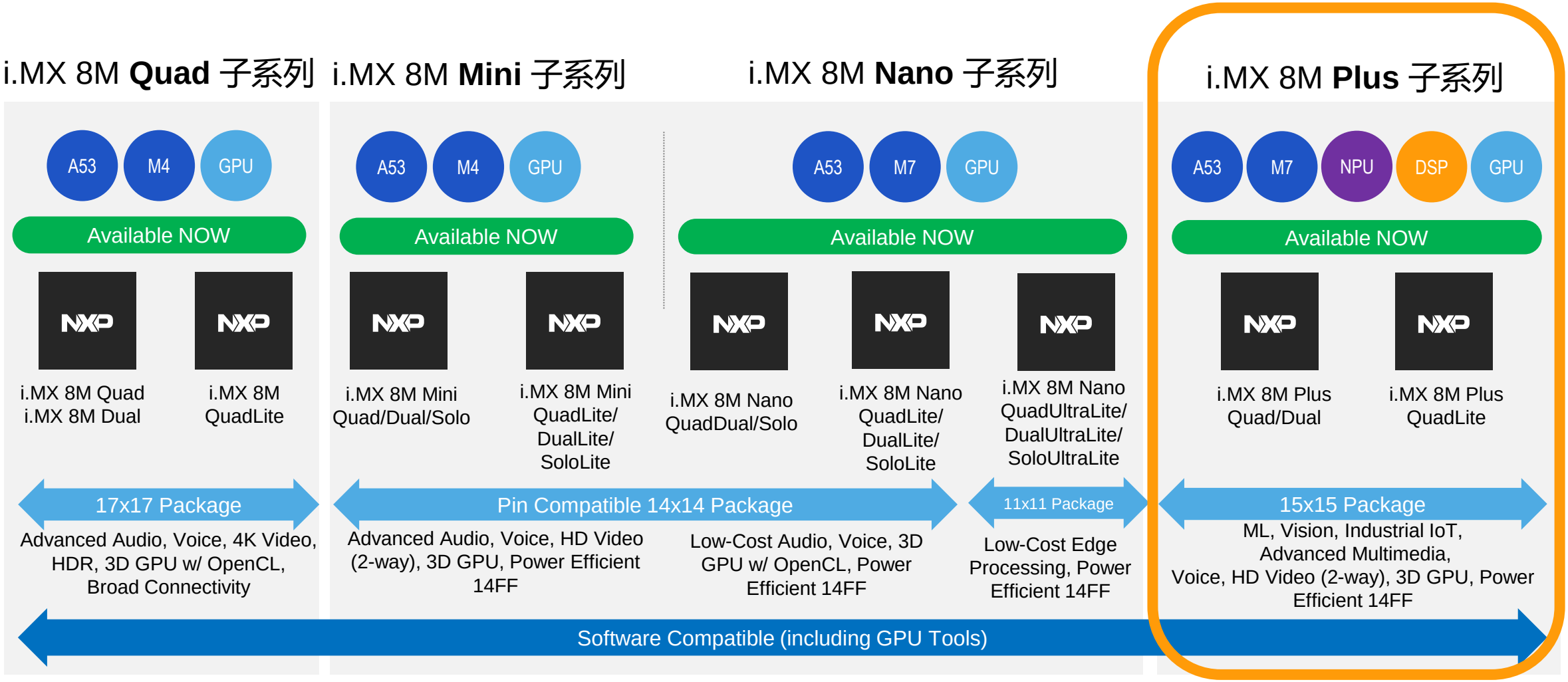
Android	Common Linux Platform	Zephyr	FreeRTOS	Azure RTOS

操作系统，BSP

i.MX	MPUs	Crossover	MCUs

芯片

I.MX 8M 系列产品概览



Scalable series of **FOUR** Arm V8 64-bit (/32-bit) based SoC Families

I.MX 8M PLUS 目标应用

机器学习和工业自动化

- Machine Vision and Robot Controller
- Industrial Computer, Gateways, HMI
- Printers and Scanners
- Machine Visual Inspection
- Factory Automation

智慧家庭，智慧楼宇，智慧城市

- Safety, Security and Surveillance
- Fleet Analytics
- Traffic Monitor and Flow Optimization
- Vision Payment Systems
- Targeted Advertisement
- Service Drones
- Alarm and AI Server Hubs
- Home Patient and Elderly Monitor

专业级音视频设备

- Surround sound and sound bars
- Audio/video receiver
- Immersive Audio Products
- Wireless or networked smart speakers
- Personal Assistant
- Voice-assisted products

I.MX 8M 系列参数对比表

Product	i.MX 8M Quad / Quad Lite	i.MX 8M Mini / Mini Lite	i.MX 8M Nano / Nano Lite	i.MX 8M Nano UltraLite	i.MX 8M Plus
Production Status	Mass Production	Mass Production	Mass Production	Mass Production	Mass Production
Main CPU	2-4xA53 1.5GHz, 1MB L2	**1-4xA53 1.8GHz, 512KB L2**	1-4xA53 1.5GHz, 512KB L2	1-4xA53 1.4GHz, 512KB L2	**2-4xA53 1.8GHz, 512KB L2**
MCU/DSP	M4 266MHz	M4 400MHz	M7 up to 750MHz	M7 up to 750MHz	**M7 800MHz, HiF i4 800 MHz**
DDR	x16 or x32 LPDDR4/DDR4/DDR3L	x16 or x32 LPDDR4/DDR4/DDR3L	x16 LPDDR4/DDR4/DDR3L	x16 LPDDR4/DDR4/DDR3L	**x32 LPDDR4/DDR4 Inline ECC**
GPU	3D – GC7000L (OpenGL® ES 2.1/3.0/3.1, OpenCL™ 1.2, Vulkan®)	2D – GC320 3D – GC NanoULTRA (OpenGL ES 2.1)	GC7000UL (OpenGL ES 2.1/3.0/3.1, OpenCL 1.2, Vulkan)	-	2D - GC520L 3D – GC7000UL (OpenGL ES 2.1/3.0/3.1, OpenCL 1.2, Vulkan)
Security	CAAM, RDC, Arm® TrustZone®	CAAM, RDC, Arm TrustZone	CAAM, RDC, Arm TrustZone	CAAM, RDC, Arm TrustZone	CAAM, RDC, Arm TrustZone
AI/ML	OpenCL CPU: 32 GOPS	A53	OpenCL CPU, GPU: 32 GOPS	A53	**ML Accel 2+ TOPS**
SRAM	128KiB + 32KiB	256KiB + 32KiB	512KiB + 32KiB	512KiB + 32KiB	768KiB + 32KiB
Camera	2x MIPI CSI (4-lane)	1x MIPI CSI (4-lane)	1x MIPI CSI (4-lane)	1x MIPI CSI (4-lane)	**2x MIPI CSI (4-lane), 2 ISPs (3-exposure HDR)**
Display	**HDMI 2.0a Tx, MIPI DSI (4-lane), eDP; HDR10, HLG, Dolby Vision**	1x MIPI DSI (4-lane)	1x MIPI DSI (4-lane)	-	**HDMI 2.0a Tx (eARC), MIPI DSI (4-lane), 1x LVDS (8-lane)**
OSD Overlay	**4Kp60**	1080p60	1080p60	-	**4Kp30**
Video Decode	**4Kp60 HEVC, VP9, 4Kp30 H.264, legacy codecs**	**1080p60 HEVC, H.264, VP9, VP8**	None	None	**1080p60 H.265, H.264, VP9, VP8**
Video Encode	None	1080p60 H.264, VP8	None	None	1080p60 H.265, H.264
Connectivity	**PCIe, SDIO, USB**	**PCIe, SDIO, USB**	SDIO, USB	SDIO, USB	**PCIe, SDIO, USB**
Audio	20x I2S TDM (32b @384KHz), S/PDIF Tx+Rx	20x I2S TDM (32b @384KHz), 8ch PDM DMIC input, S/PDIF Tx+Rx	**12x I2S TDM (32b @384KHz), ASRC, 8ch PDM DMIC input, S/PDIF Tx+Rx**	**12x I2S TDM (32b @384KHz), ASRC, 8ch PDM DMIC input, S/PDIF Tx+Rx**	**18x I2S TDM (32b @384KHz), ASRC, 8ch PDM DMIC input), S/PDIF Tx+Rx**
Expansion I/O	**2x USB3.0, 2x PCIe Gen 2**	**2x USB2.0, 1xPCIe Gen 2**	1x USB2.0	1x USB2.0	**2x USB 3.0, 1x PCIe Gen 3**
Network, Storage	1x GbE, 2x SD/eMMC, MLC/SLC NAND	1x GbE, 3xSD/eMMC, MLC/SLC NAND	1x GbE, 3xSD/eMMC, MLC/SLC NAND	1x GbE, 3x SD/eMMC, MLC/SLC NAND	**2x GbE (1x TSN), 2x CAN-FD, 3x SD/eMMC, MLC/SLC NAND**
Process	28nm	**14nm FinFET**	**14nm FinFET**	**14nm FinFET**	**14nm FinFET**
Package	17x17mm, 0.65mm	**14x14mm, 0.5mm de-pop**	**14x14mm, 0.5mm de-pop**	11x11mm, 0.5mm de-pop	15x15mm, 0.5mm de-pop

Pin Compatible

Highlighted Feature

I.MX 8M PLUS 特性与优势

	Capability	Feature	Benefits
Machine Learning	**Machine Learning**	• Neural Processing Unit (NPU) up to **2.3 TOPS**	• Edge Inference and intelligence • No cloud dependency, privacy, better user experience
	Vision System	• Up to 2 cameras (**2x MIPI-CSI**), 1080p60; • Image Signal Processor (**ISP**) up to 375MPix/s, HDR, de-warp	• 4K Vision / 1080p Stereo Vision. • HDR extracts the maximum image details in high contrast scenes • De-Warp: Fisheye lens correction reduce overall optical system cost
	Voice	• Low Power **Voice Accelerator**, • **8ch-PDM** digital microphone input.	• Voice systems processed at the edge. • Privacy, better user experience, no cloud dependency • Low noise and cost-effective microphone system
Advanced Multimedia	**HD Video**	• 1080p60 **video decoding**, **(H.265)** • 1080p60 **video encoding, (H.265)**	• High resolution video compression for cloud upload or local storage • H.265 compression is twice more effective than H.264
	3D/2D Graphics	• **2D and 3D GPU** - 1Gpix/s	• 3D and 2D graphics for rich HMI and user experience
		• OVG 1.1, OGLES 3.1, Vulkan, OCL 1.2 FP	• Vulkan requirement for Android 10 onwards
	Advanced Audio	• 18x I2S TDM, DSD512, SP/DIF Tx + Rx, 8-ch PDM	• Latest audio technology including Immersive audio support
		• HDMI 2.0b Tx + **eARC, ASRC**	• eARC for high-quality audio for cost-optimized soundbars
Industrial Network & Reliability	**Industrial Network**	• **2x Gigabit Ethernet** (1x w/ **TSN**) **2x CAN-FD**	• Low latency industrial network for precise control • robust industrial control network • Gateway application with dual Ethernet.
	Memory Reliability	• **ECC** in L2, **In-line ECC** on DDR bus • 14FF process	• High Industrial system reliability for Safety Industrial Level SIL3. • Low SER rates
	Real time Processing	• **Cortex-M7 @800MHz**	• High real-time processing power, Cortex-A complex off loading • Eliminating an external MCU
Performance & Connectivity	**High Performance Low power**	• **4x CortexA53 @1.8GHz,** • 14nm FinFET, Low Power, high performance	• Applications running <2.0W. Deep Sleep Mode < 20mW
	Display Interfaces	• **MIPI-DSI**, **HDMI** 2.0b, **LVDS** 4/8-lane, supports 2x1080p60 or 1x4kp30	• Multiple display interface options working simultaneously.
	High-Speed Interfaces	• **2xUSB 3.0** (5Gb/s each) **PCIe 3.0** • 3x SDIO 3.0 800Mb/s	• Fast connections to Wi-Fi, FPGAs, co-processors. • Fast data transfer, e.g., media files

I.MX 8M PLUS 处理器产品框图

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)

Voice Acceleration co-processor:

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

Machine Learning (ML) accelerator: 2.3 TOPS performance

Package:

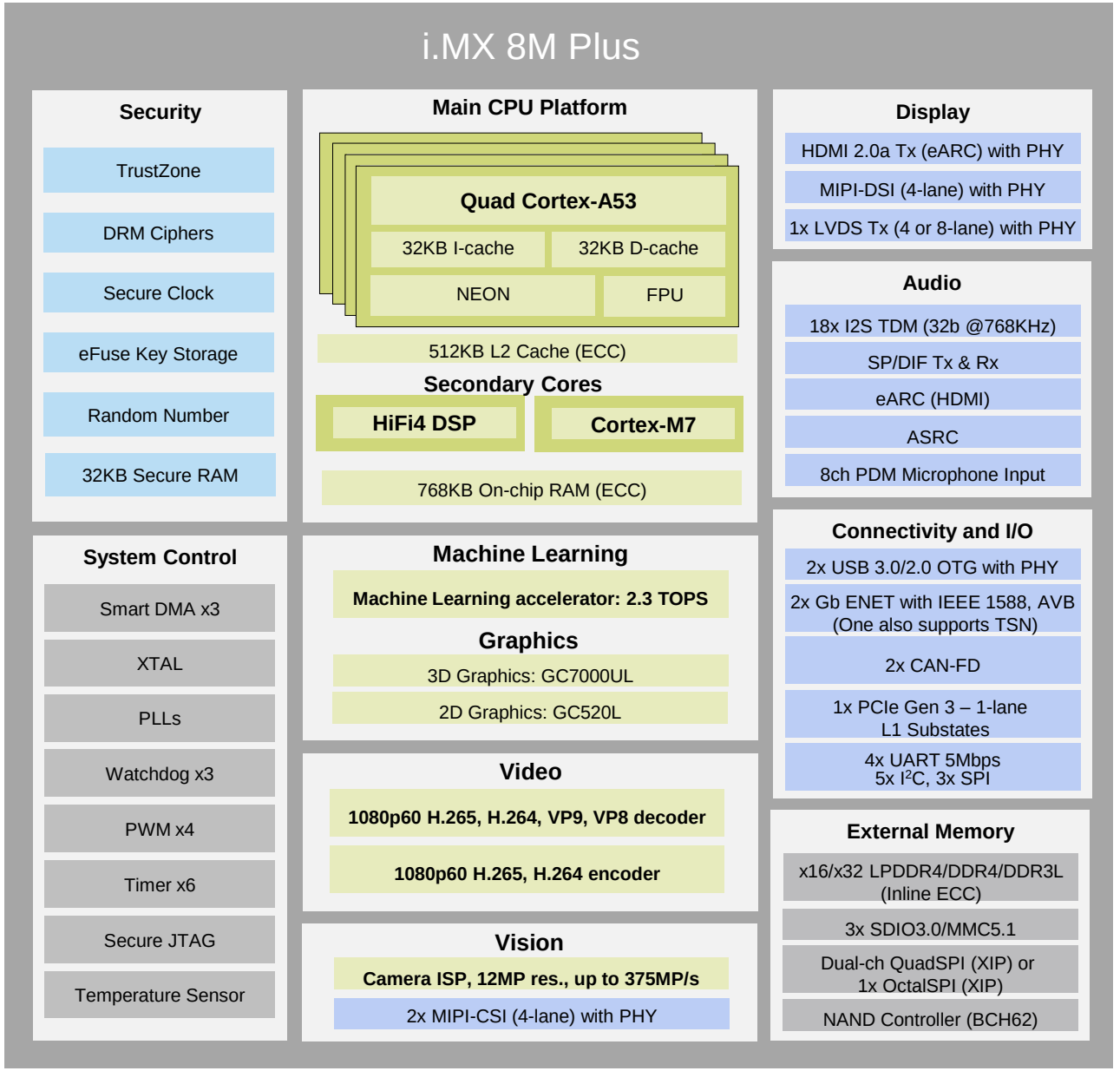
#1 FCBGA 15x15mm, 0.5mm pitch, depop (consumer and industrial)

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

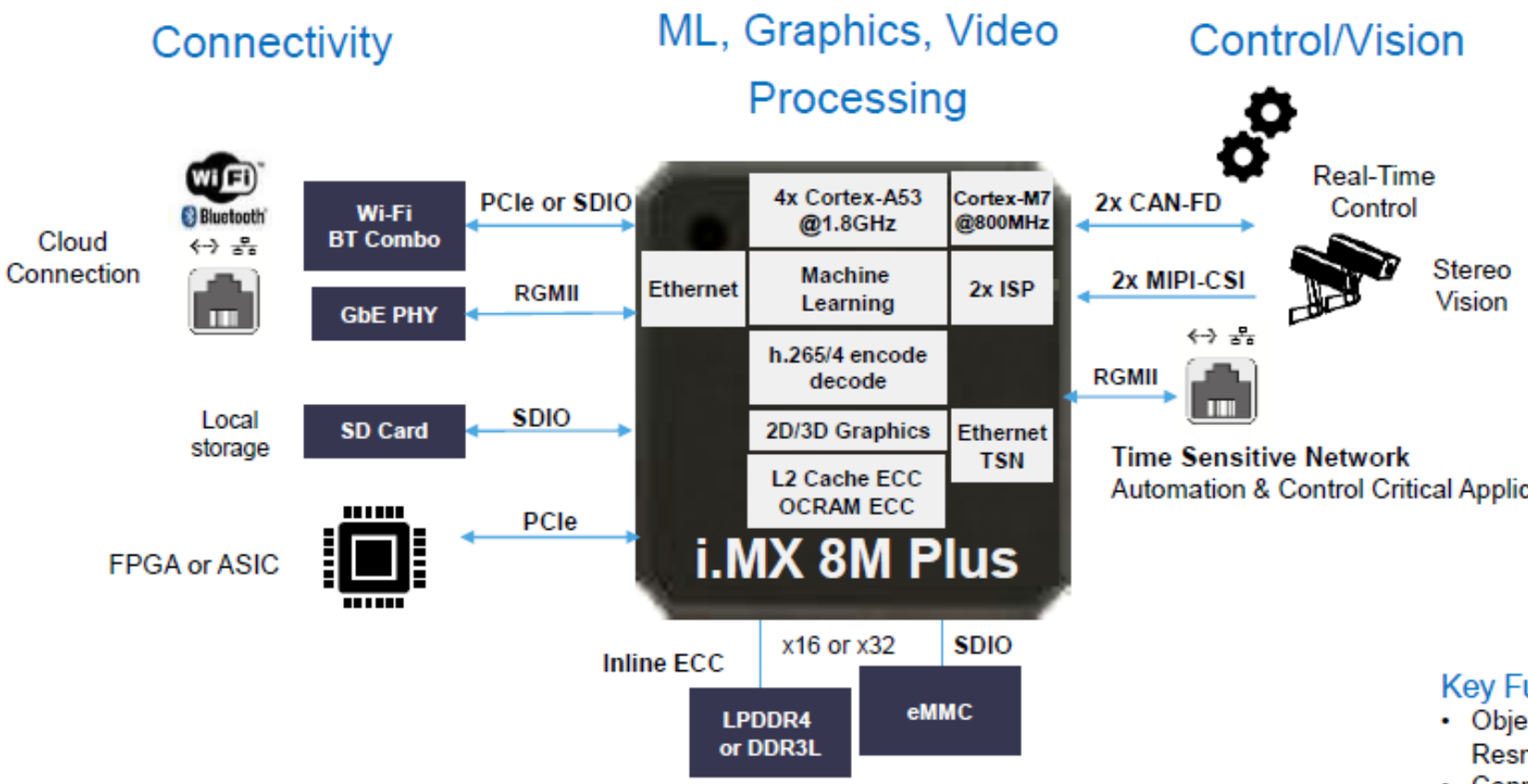
Qualification: Consumer (0C to +95C); Industrial (-40C to +105C)

Feature Highlights:

- **External memory:** x32/x16 LPDDR4/DDR4/DDR3L (Inline ECC)
- 3x SDIO3.0/eMMC5.1
- Dual-channel QuadSPI (XIP) or 1x OctalSPI
- NAND Controller (BCH62)
- SPI NAND
- **Graphics processors:**
- GC7000UL (3D GPU, 2-shaders, OpenGL® ES 3.1, Vulkan®, Open CL™ 1.2 FP)
- GC520L (2D GPU, OpenVG™ 1.1)
- **Video processors:** 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)** with L1 Substates for fast wake from low power mode
- **2x Gigabit Ethernet** with IEEE 1588, EEE and AVB (one with TSN, but no EEE)
- **2x CAN-FD**



I.MX 8M PLUS 针对工业视觉的应用



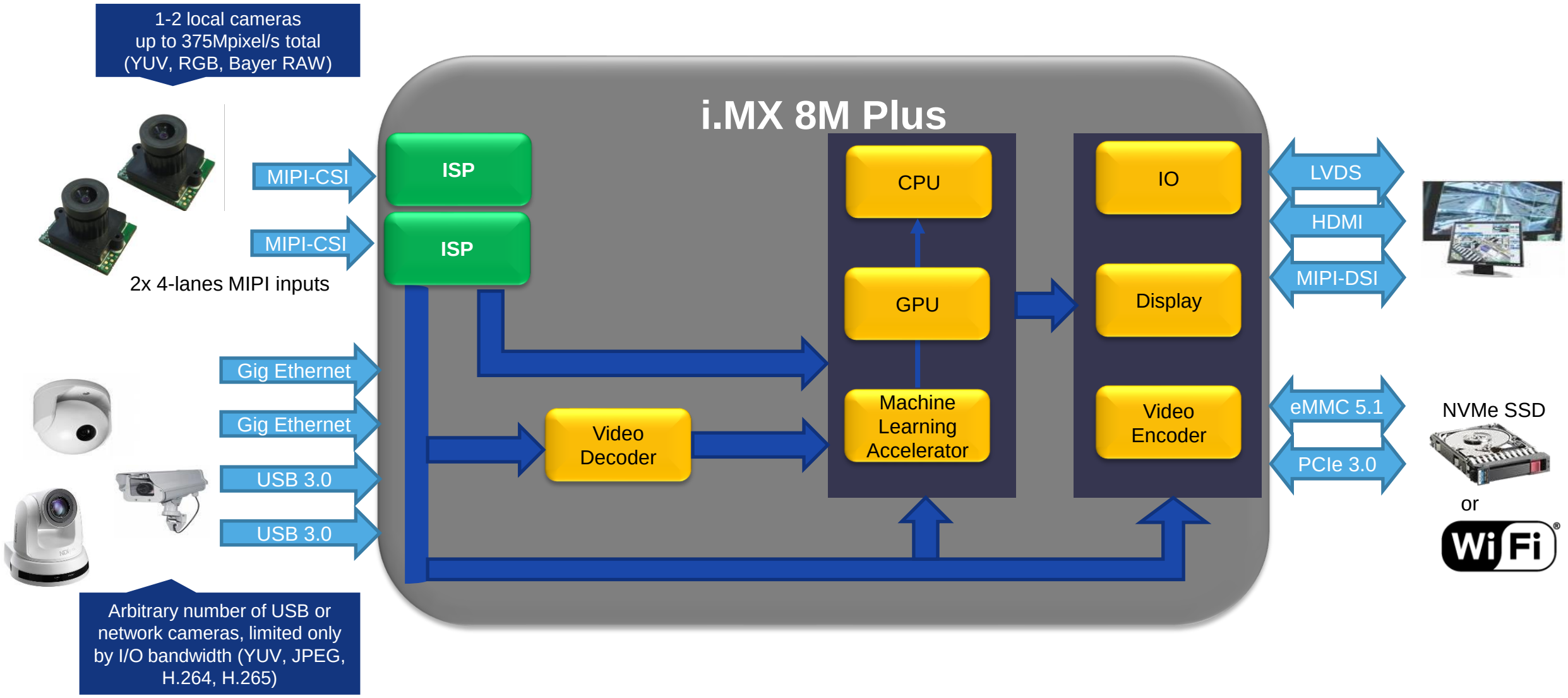
Key Performance Req.
- 1080p video encode
- 2x PCIe and Gb Ethernet
- Wi-Fi
- 2x MIPI-CSI cameras w/ ISP

Key Functions:
- Object detection (MobileNet-SSD, Resnet-SSD)
- Connected to industrial network
- Cloud connection for remote monitoring
- Ensure system reliability
- Recognition

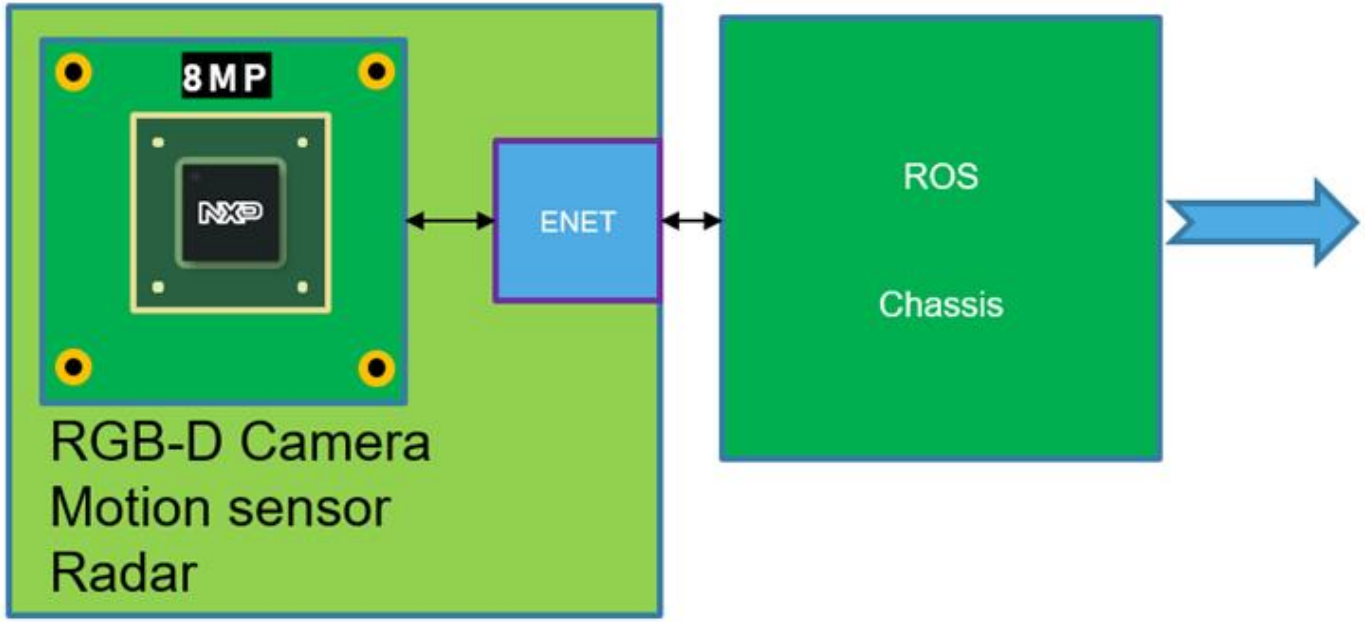
Overview:

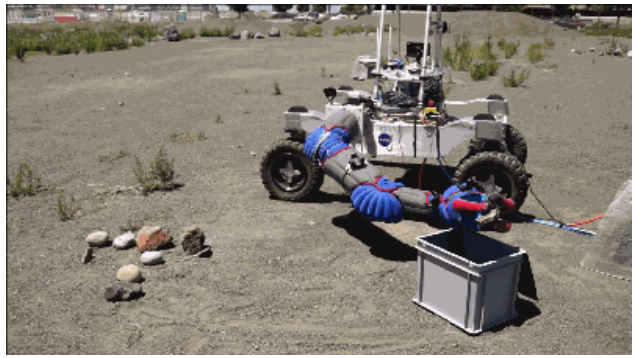
Stereo vision camera feeds, connected to industrial network, connection to the cloud, object detection with Machine Learning at the edge

I.MX 8M PLUS 工业级多路视觉与显示系统

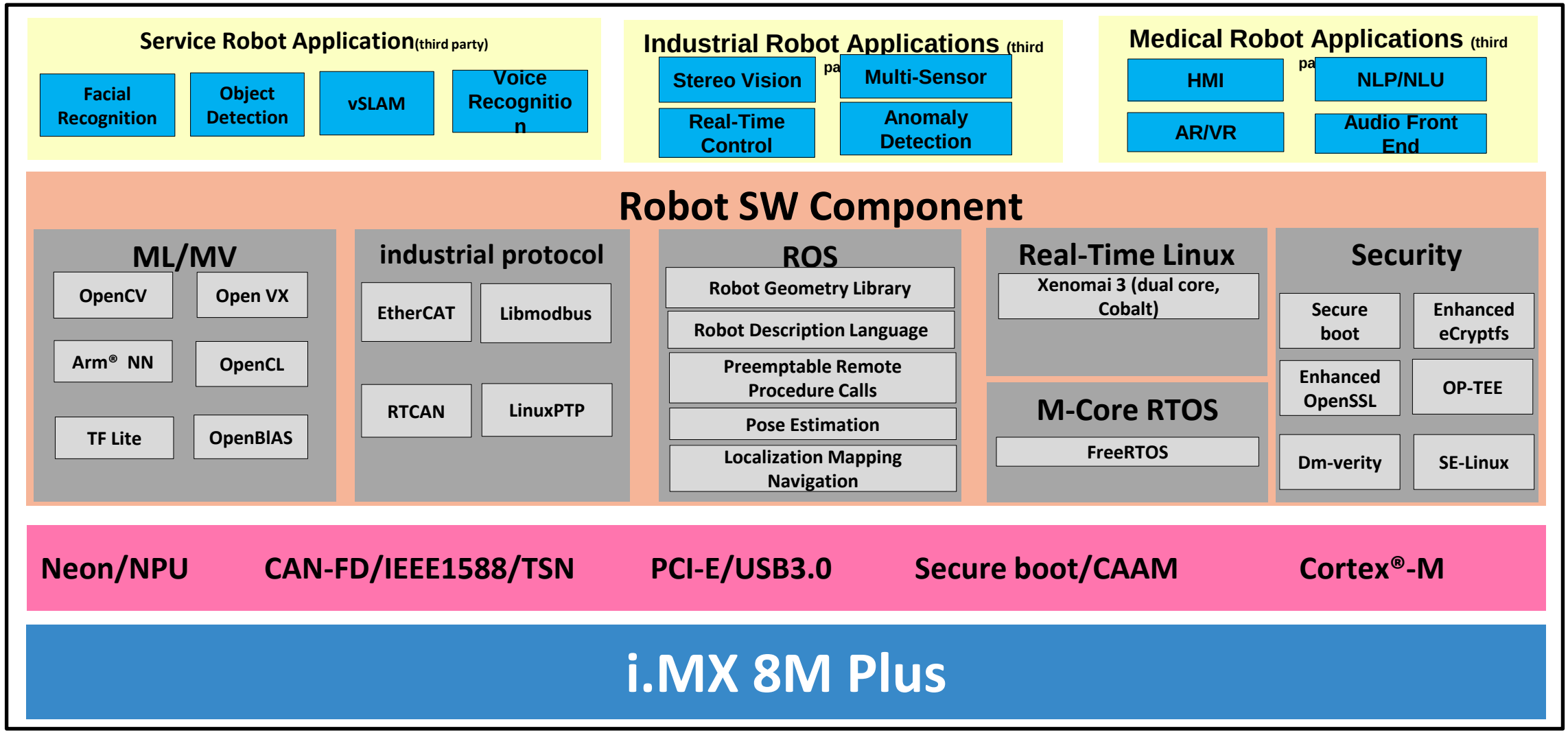


自动导引车应用, AGV (AUTOMATIC GUIDED VEHICLE)





I.MX 8M PLUS AGV及机器人参考平台软件架构



Highlight : Smart, Security, Real-Rime, Reliability

应用于边缘智能的工具箱和推理引擎

机器学习库和开发工具
面向恩智浦的MPU和MCU

部署开源推理引擎

推理引擎的集成与优化:

- i.MX MPU: TensorFlow Lite, Arm NN, ONNX, Tengine
- i.MX RT MCU: Glow, TensorFlow Lite, Arm CMSIS-NN

集成到Yocto Linux BSP 和 MCUXpresso SDK

NXP提供的免费工具，无需单独的SDK或可下载的软件发布:

- i.MX: Yocto里的meta-imx-ml层
- MCU: MCUXpresso SDK的中间件

支持易用性的各种文档资料

端到端的演示示例, 例如摄像头 → 推理引擎

文档: eIQ 用户手册, 软件发行说明, Demo用户手册

基于流行神经网络框架的预训练模型的使用指南, 例如 TensorFlow, PyTorch

技术培训资料, 例如讲座，动手操作，视频

https://community.nxp.com/community/eiq/overview

EDGEVerse | 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

PyTorch Caffe2 ONNX

eIQ 工具集

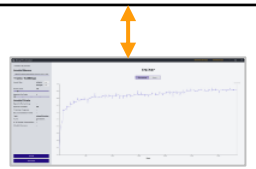
eIQ 用户接口

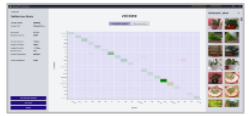
模型优化工具

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

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

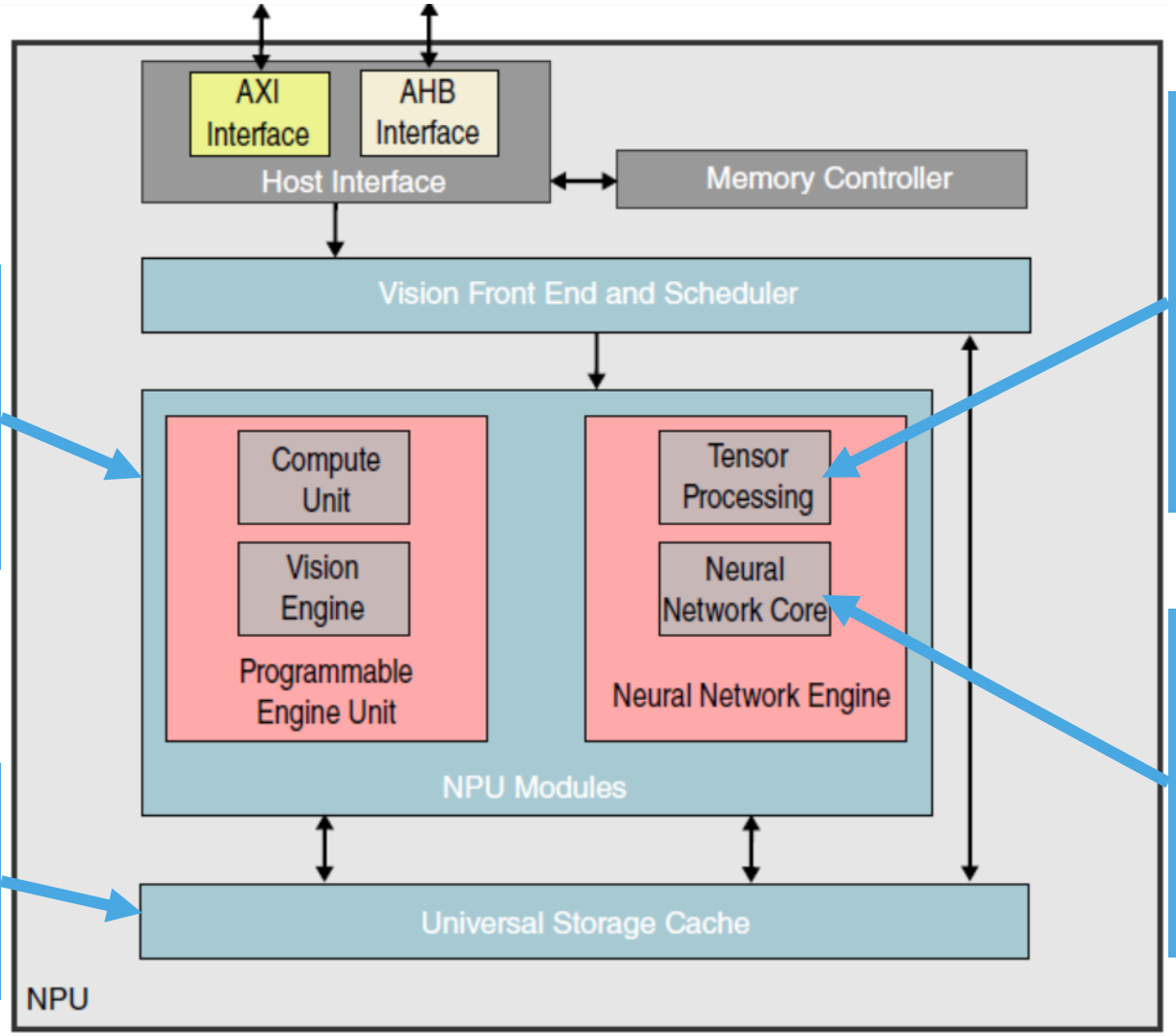
I.MX8M PLUS的机器学习推理引擎NPU

可编程引擎单元
(i.MX8M Plus有1个实例)

128-bit 向量处理.
整型 8/16/32b, 浮点 16/32b.
灵活的可编程单元
处理张量处理单元和神经网络引擎不能处理的算子/层

通用存储缓存
(i.MX8M Plus有1个实例)

作为本地内存和PPU的L1缓存在NPU模块之间传递数据



张量处理单元
(i.MX8M Plus有3个实例)
整型8/16b, 浮点 16b

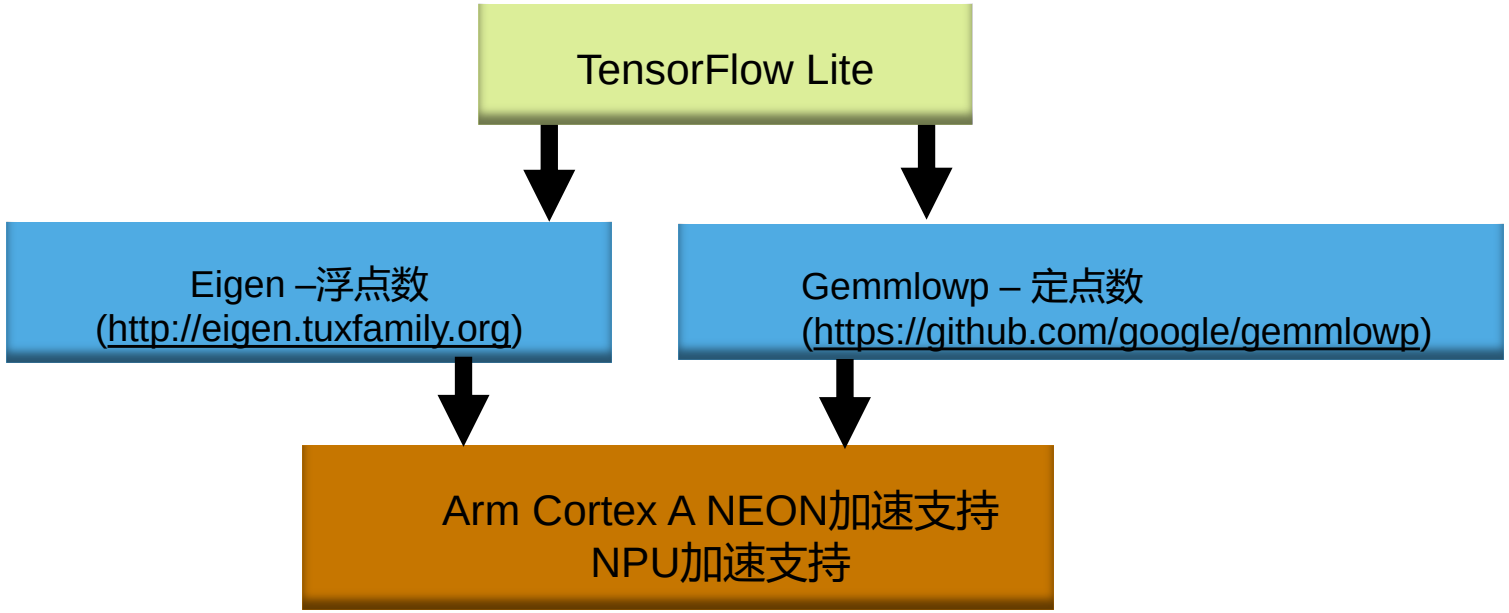
非卷积层. 数据洗牌的多通道处理、规范化，pooling/unpooling, ReLU, LUT
支持网络剪枝、零权重跳过、压缩

神经网络引擎
(i.MX8M Plus有6个实例)
2.3 TOPs 整型8bit

处理卷积层，去卷积，Average Pooling层和全连接层

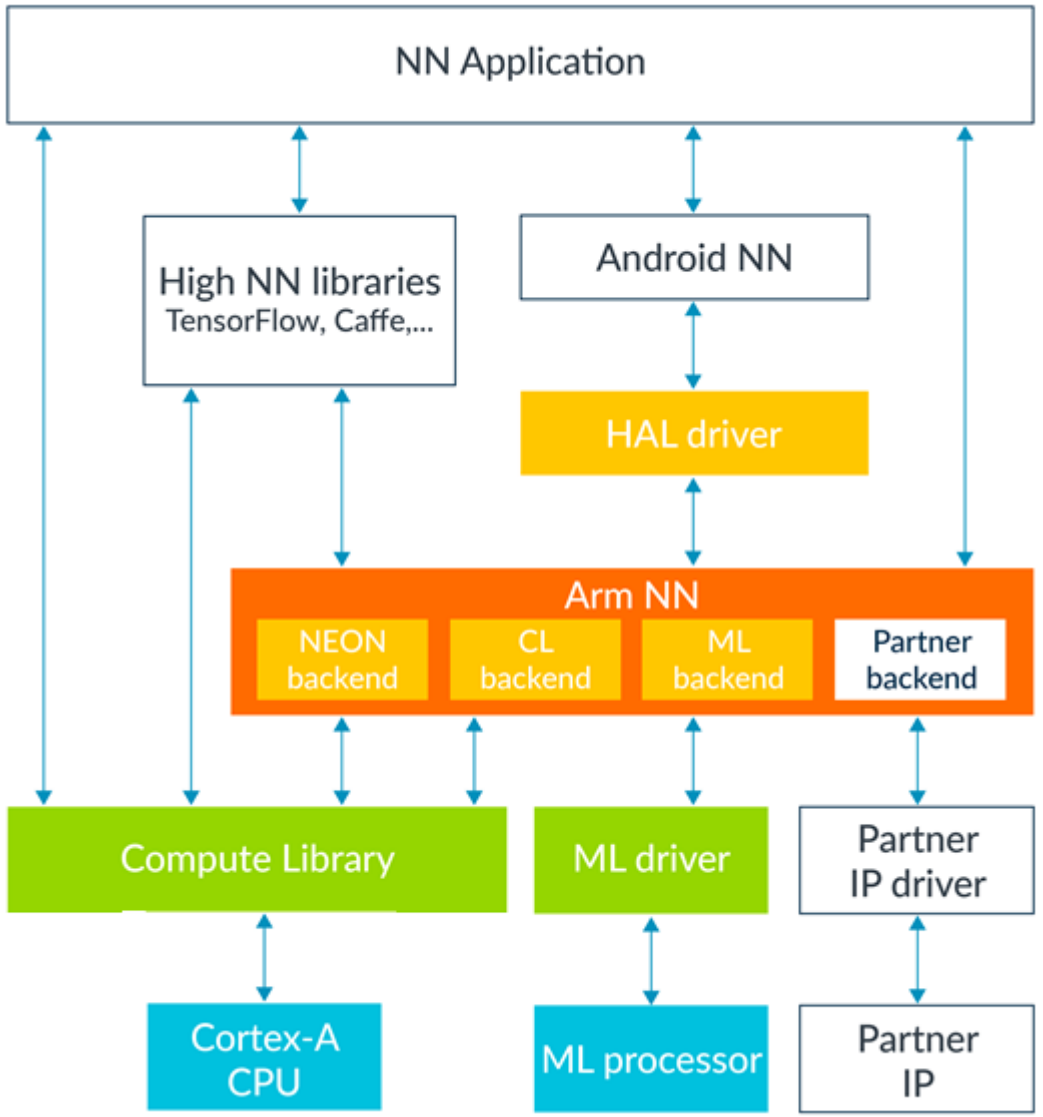
基于TensorFlow Lite的eIQ推理

- 以较低的延迟和较小的二进制大小在边缘侧进行推理
  - TensorFlow模型的简化形式
- 支持TensorFlow操作符的子集；某些模型可能无法转换或需要自定义实现 - https://www.tensorflow.org/lite/guide/ops_compatibility
- 使用OSS库加速矩阵和向量算法(例如 Eigen, GEMMLOWP)



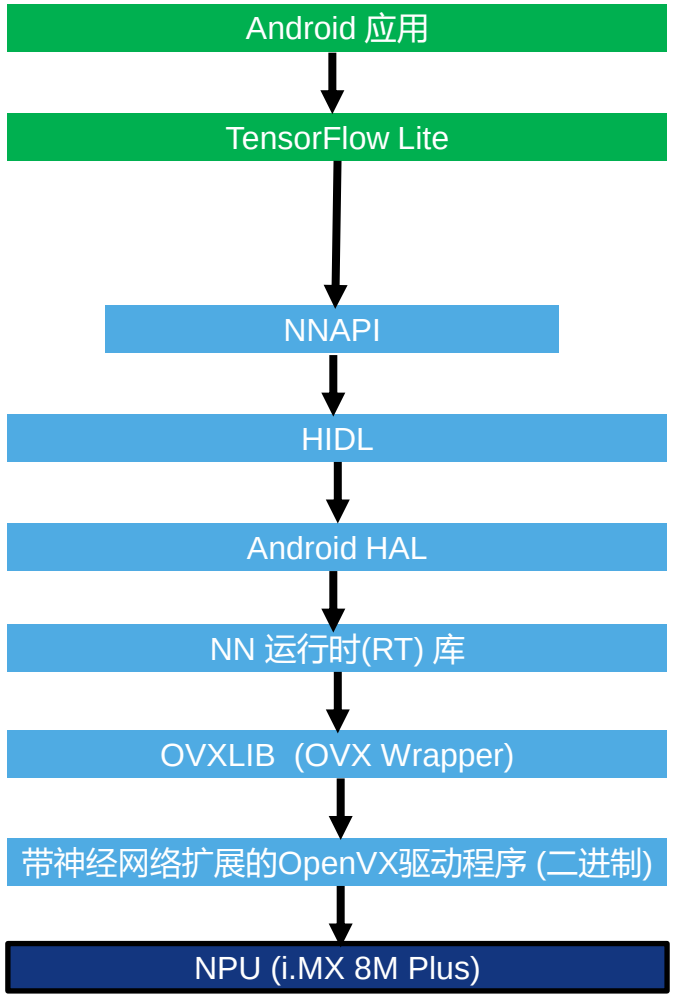
基于ARM NN的eIQ推理

- 中间件推理引擎使用后端API接口将高级库与底层CPU、GPU和其他IP连接起来
- 支持解析的神经网络格式: TensorFlow, TF-Lite, Caffe, ONNX
- 后端将神经网络图的层映射到硬件加速器，即负责执行计算的驱动程序

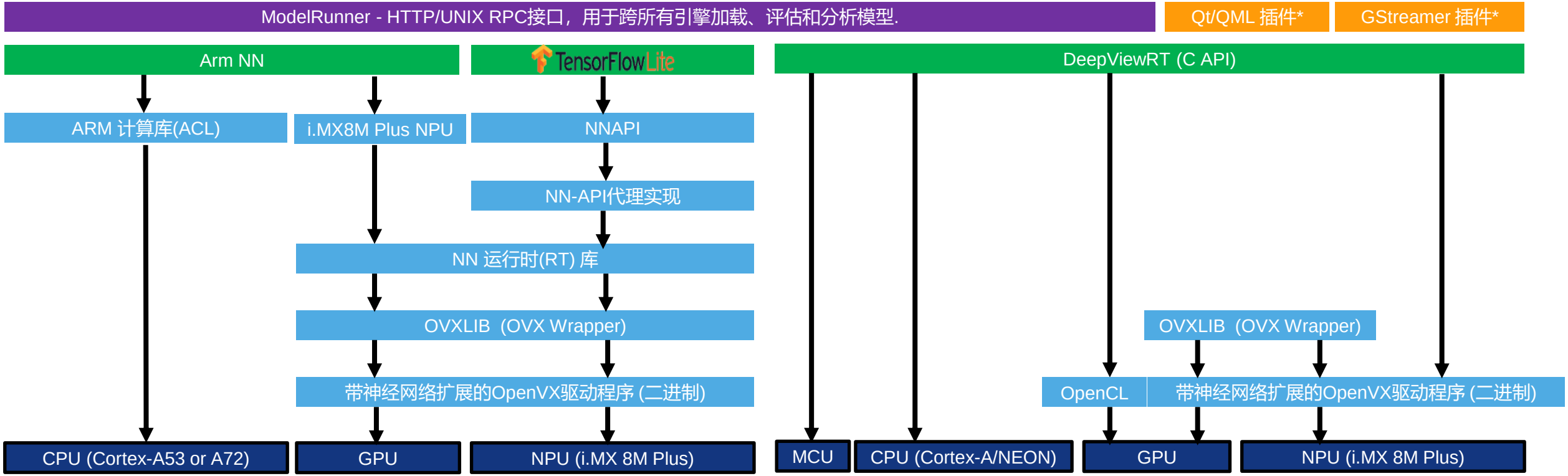


i.MX 8M Plus对Android NN的支持

- 与TensorFlow Lite类似的软件结构
- “自动”作为Android BSP的一部分被启用
- NXP将提供与Linux版本相同的基准测试



DeepView 推理引擎



*Qt/QML and Gstreamer 插件将在2021年迁移到 ModelRunner API以支持所有推理引擎

SECURE CONNECTIONS
FOR A SMARTER WORLD