Efficiency Unleashed: The Next-Gen NXP i.MX 95 Applications Processor for Embedded Vision

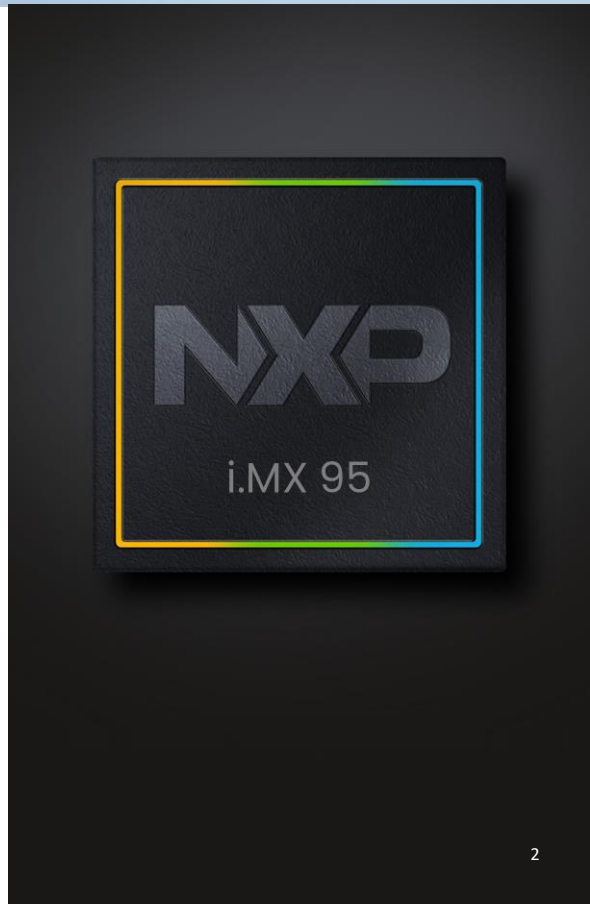
James Prior

Senior Product Manager

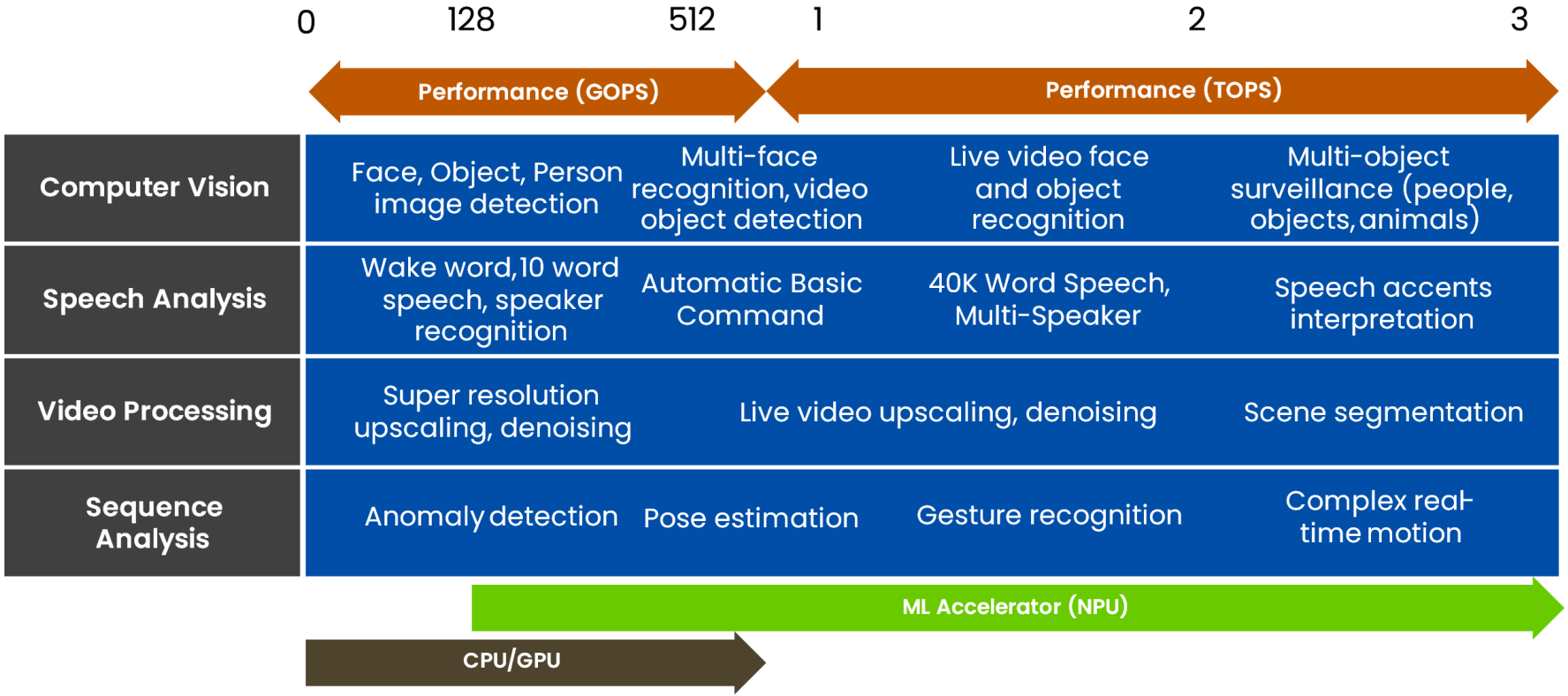
BL SCE, NXP Semiconductors

Edge AI Needs More Efficient ML/AI

- Enable real-time analytics and actuation
  Not hampered by network latency
- On-board machine learning
  Precise and fast detection, classification, adaptation
- Reduce data center traffic
  Only process and store relevant data
- Reduce network cost
  Shield cloud from large part of raw data
- Safeguard privacy
  Transmit semantic rather than raw data
- Increase security
  Resilient to offline conditions



AI/ML Enabled Applications with eIQ® Neutron NPU



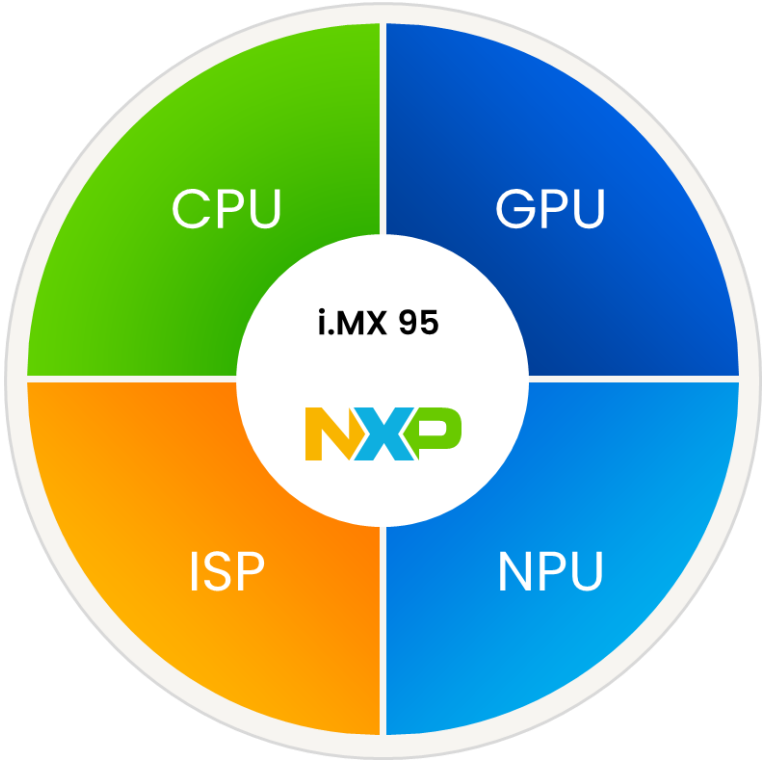
First i.MX Application Processor with eIQ® Neutron NPU

Display & Multimedia Capability

Advanced Security

Enhanced Reliability

Functional Safety



Machine Vision Capability

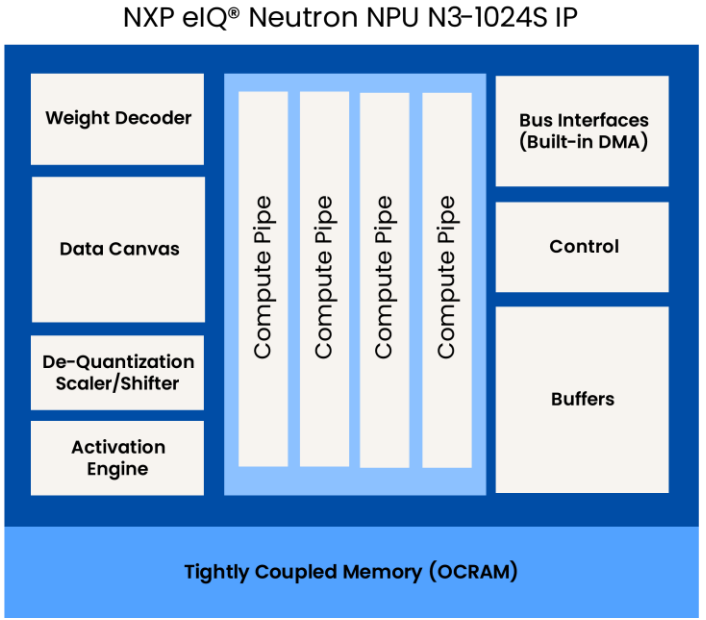
High-Performance Compute

Real-Time Compute

Rich High Speed I/O

NXP eIQ® Neutron NPU
Acceleration for Edge AI & Industry 4.0 Workloads

- Acting on market feedback to deliver scalability & consistency
  - Optimized for Edge AI inference workload performance & power efficiency
  - Supports major NN structures (CNN, MLP, RNN, LSTM, TCN, and more)
  - LLM support enablement: LLAMA v2 & Blenderbot
- Flexibility to tune solution to meet customer needs
  - Hardware scales from 32 Ops/cycle to 2k Ops/cycle
  - Software support is unified and consistent

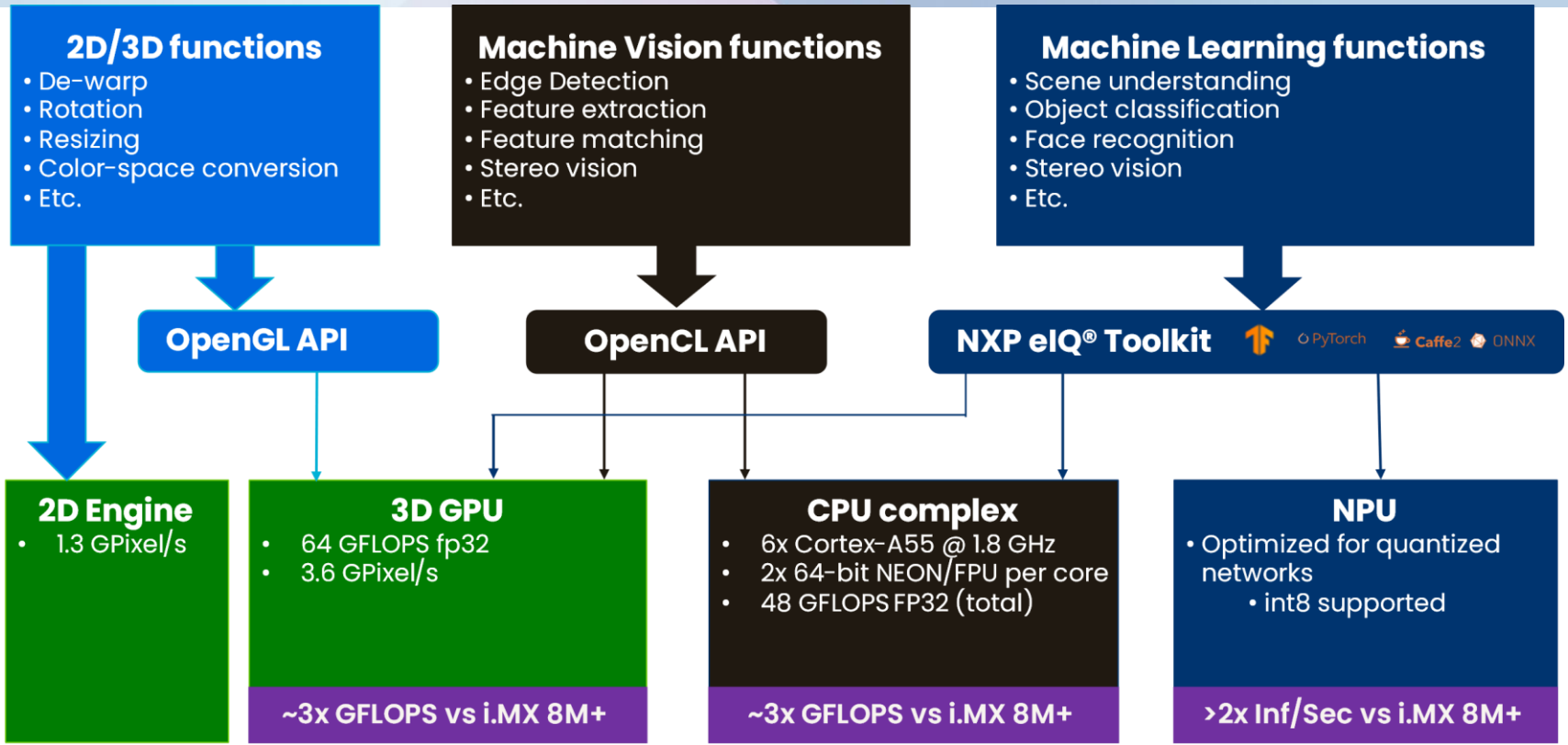


NXP eIQ® Neutron NPU Performance

Benchmark / Performance (Inferences/Sec)	i.MX 8M Plus VeriSilicon VIP8000 2.3 TOPS[1]	NXP i.MX 93 ARM ETHOS U65 0.5 TOPS[2]	NXP i.MX 95 eIQ Neutron N3-1024S 2 TOPS[3]
MobileNet-v1	368	236	**1112**[4]
MobileNet-v2	332	282	**721**
Inception v3	30	30	**101**
ResNet50-v1	60	20	**125**
SSD-MobileNet_v2	137	76	**350**
Performance relative to Neutron N3-1024S	0.48	0.35	**1**
Efficiency relative to Neutron N3-1024S	0.41	0.71	**1**

1: From NXP i.MX8M Plus (measured)
2: From benchmark document by Arm shared with NXP.
3: Projected based on cycle-accurate simulations.
4. Measured on i.MX 95 A1 pre-validation silicon. Preliminary results subject to change.

i.MX 95 Embedded Vision Function Mapping



eIQ® ML Software Development Environment

Embedded Developers

Data Scientists

ML Experts

Bring Your Own Data

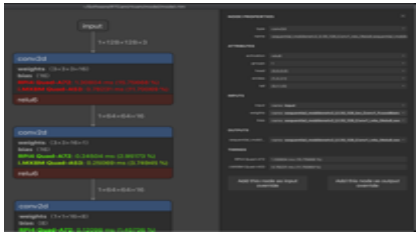
Bring Your Own Model

Extensible **workflow tools** for dataset management, model training, optimization, deployment, and profiling

PyTorch

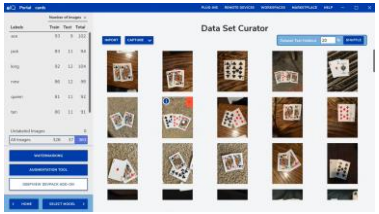
ONNX

eIQ Toolkit



Command Line

or



Graphical UI

Pipelines

AUDIO / VOICE

VISION

TIME SERIES

STATE MONITOR

IDENTITY RECOGNITION

PERSON DETECTION

and more ....

App Code Hub

Inference

TFLite for Microcontrollers

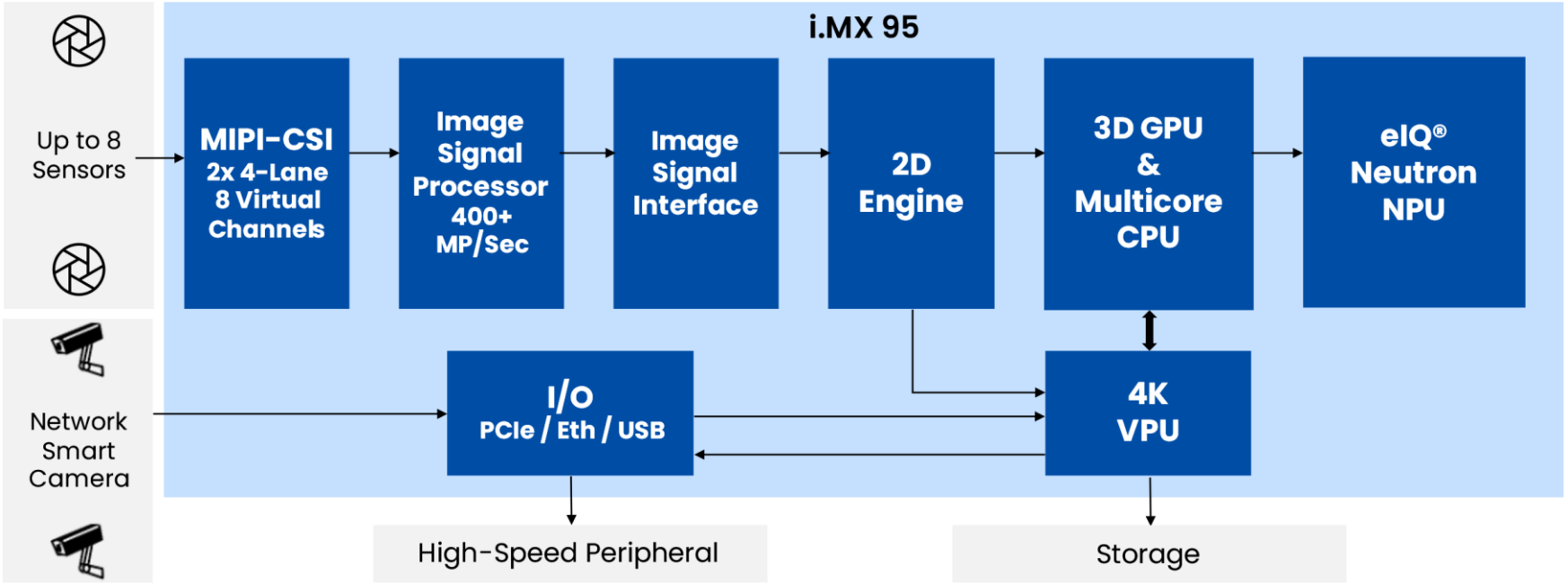
Arm® Cortex®-M, DSP, eIQ Neutron NPU

TFLite

Arm Cortex-A, GPU, DSP, NPU, eIQ Neutron NPU

Inference engine support for compute engines (CPU, NPU, GPU, DSP) available on NXP MPU and MCU

i.MX 95 Vision Processing Pipeline



Up to Single 12MP high resolution camera - 4096x3072p30 / 3820x2160p60

Powerful Immersive Graphics

Arm® Mali™ G310 3D GPU

- Superb code portability & ease of use
  - Strong app ecosystem
  - Android Play store
  - Industry support by HMI engines
- Valhall architecture with tile-based rendering
  - Modern API support including Vulkan, OpenGL
  - Area- & energy-efficiency focus
  - QoS support with finer granularity

- Enhanced, separate 2D GPU enables safety-context streams
  - Supports (de)warp function
  - Higher pixel throughput @ 1300 mega-pixel/sec (vs. NXP i.MX 8M Plus)

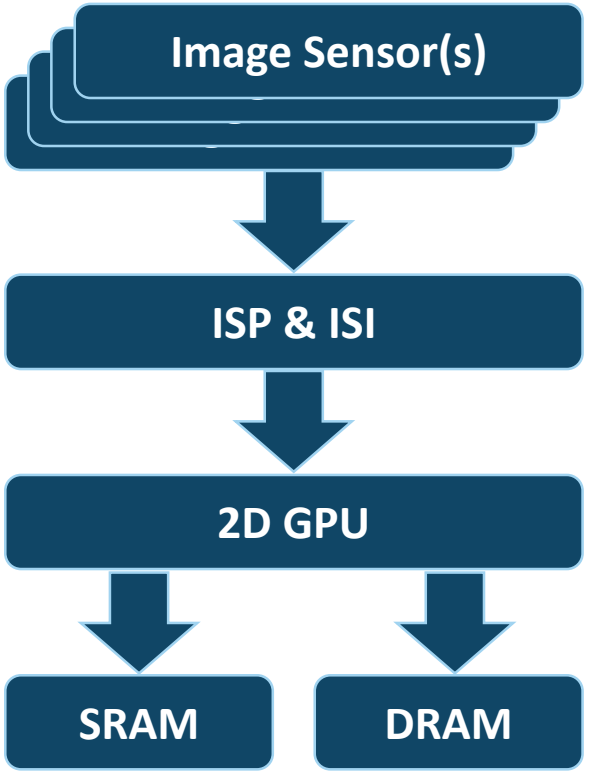
3D GPU Benchmarks

Specification/Benchmark	i.MX8 QXP/DXP GC7000 Lite	i.MX8DM GC7000XSVX	i.MX8QM 2xGC7000XSVX	i.MX 95 Mali-G310 v2
GPU Clock Freq [core/shader; MHz]	700 / 850	800 / 1000	800 / 1000	**1000**
GFLOPS (FP32/FP16)	51.2 / 25.6	64 / 128	128 / 256	**64/128**
Triangle Fill-rate (Mega-triangles/sec)	234	267	267 + 267 (dual) 267 (bridged)	**400**
Pixel Fill rate (Giga-pixels/Sec)	1.4	1.6	3.2	**4**
GFXBench30 Manhattan offscreen1920x1080 (FPS)	5.4	10.7	16.3	**20.3**
GFXBench31 Manhattan31 offscreen (FPS)	3.2	6.9	12	**13**
GLBenchmark27 TREX Onscreen (FPS)	18	26	39	**35**
Antutu v8.4.x (Terracotta-Vulkan) score	TBD	522	534	**984**
Antutu v8.4.x (Refinery-OpenGLES31 + AEP) score	N/A	6849	7603	**18261**

Preliminary results obtained by NXP internal team; subject to change. Platform configuration and BSP/software version may offer different results

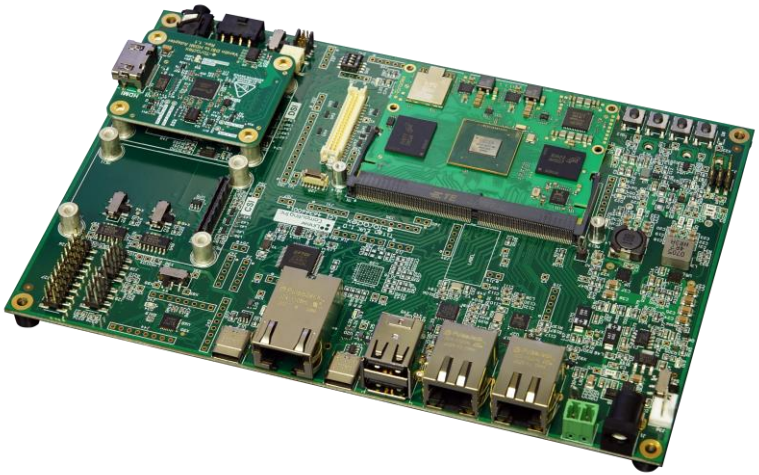
New NXP Image Signal Processor (ISP) Optimized for Machine Vision

- Capture RGB + IR images simultaneously for better low-light image capture, small camera modules, and enhanced facial recognition
- Up to 12MP high-resolution camera
- 20 bit-per-pixel pipeline supports up to 8 sensors w/aggregate 500 MPixel/s throughput
- HDR combining of 2 exposures
- Support color, monochrome & RGB-IR sensors
- Advance de-noising and edge enhancement for low-light conditions
- Single camera memory-to-memory or streaming processing



How To Get NXP i.MX 95 Evaluation Kits?
Apply for Toradex Verdin EVK early access now

Empowering developers to kickstart projects and smoothly transition from proof-of-concept to large-scale production



Toradex
Swiss. Embedded. Computing.

Enhanced Reliability and Functional Safety
FuSa-dedicated **EEPROM** | **OctalSPI** Flash for fast boot

Expansibility and Connectivity
mPCIe and M.2 Key E slots for Cellular Modems, WiFi-6/6E etc.
Micro-SD Card slot | 8x **ADC** Inputs | 4x **I2C** via headers

High-Performance Compute
Featuring a 16GB **LPDDR5** RAM

Energy Flex Architecture
Power Monitoring of SoM and Carrier Board
Control & Optimization of SoC voltage rails

Display and Multimedia
Quad-Lane MIPI **DSI** | Octal-Lane **LVDS**
Audio Codec on-board | **PDM** mics

Machine Vision Capabilities
Quad-Lane MIPI **CSI** (up to 2x w/o DSI)

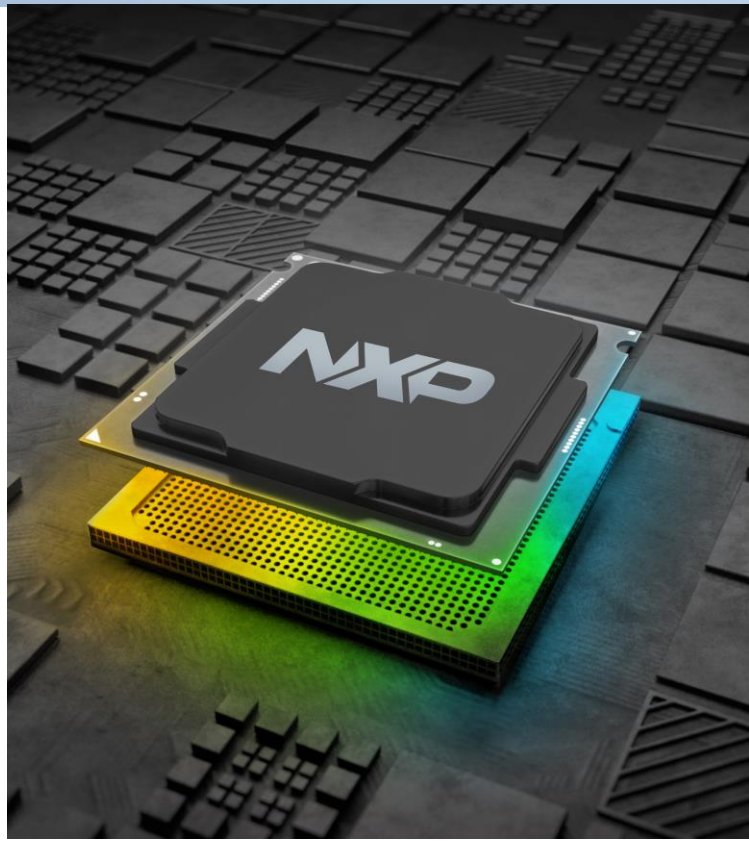
Rich set of Interfaces
10-GbE port | 2x **GbE** ports with **TSN** capability
USB 3.0 Type C | 2x USB 2.0 Type A | 2x **CAN-FD**

SCAN ME

First i.MX Application Processor with eIQ® Neutron NPU

- NXP i.MX 95 Family of Applications Processors
  - For Auto Digital Cluster; Industry 4.0 & HMI ; Commercial IoT Smart Devices & Gateways; & more
  - New NXP eIQ Neutron NPU
  - Premium Arm® Mali™ Graphics
  - 4K Vision Processing
  - Functional Safety
    - ISO 26262 (ASIL-B) / IEC-61508 (SIL 2)
  - Early Access Program Open Now
    - Sampling 2Q24
    - Production 2H25

SCAN ME



Where to find more info on NXP i.MX 95 & AI/ML

NXP i.MX 95 Product Page
https://www.nxp.com/imx95

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

Toradex Verdin i.MX 95 Evaluation Kit
https://www.toradex.com/computer-on-modules/verdin-arm-family/nxp-imx95-evaluation-kit

2024 Embedded Vision Summit

Additional Info:

i.MX 95 In Action For eCockpit with Generative AI

Visit The NXP booth #503