Scaling i.MX Applications Processors with Discrete AI Accelerators

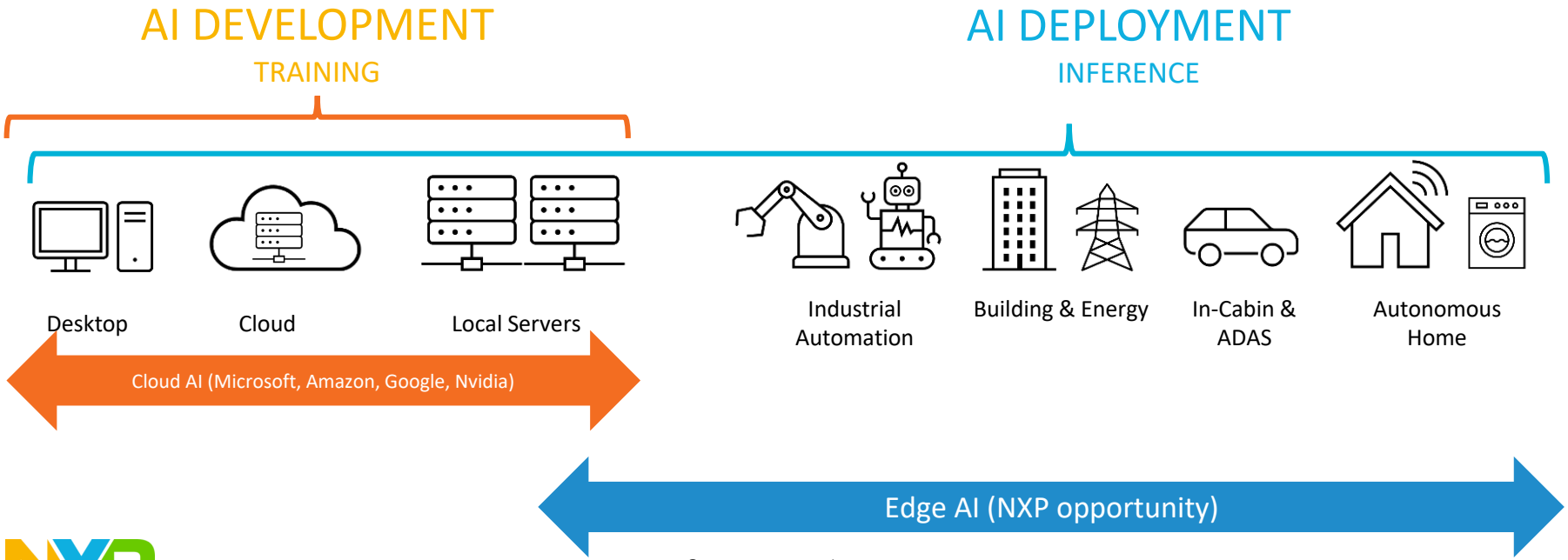
Ali O. Ors

Global Director, AI Strategy and Technologies

NXP Semiconductors

AI Spans from Training in the Cloud to Inference at the Edge

NXP's strength in processing gives us a unique opportunity to shape the deployment of AI at the edge.



Intelligent edge systems enabled by NXP

Expansive processor portfolio

MCX MCUs

i.MX RT Crossover MCUs

i.MX Apps Processors and beyond

AI co-processor

NPU

Differentiated HW and SW enablement

eIQ® Neutron NPU

Highly scalable and optimized integrated dedicated AI acceleration

eIQ® Toolkit

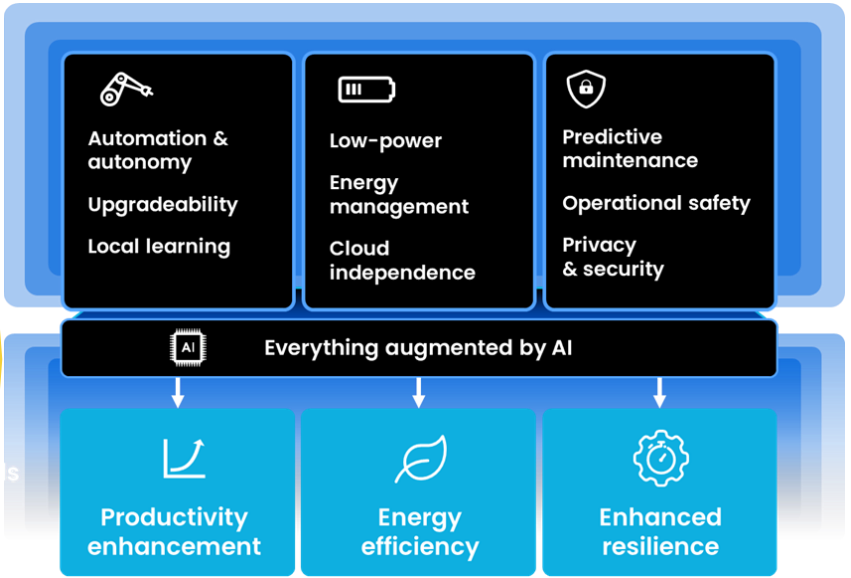
AI/ML software toolkit for model Creation, optimization and porting

eIQ® Time Series Studio

Automated ML model creation from sensor signals

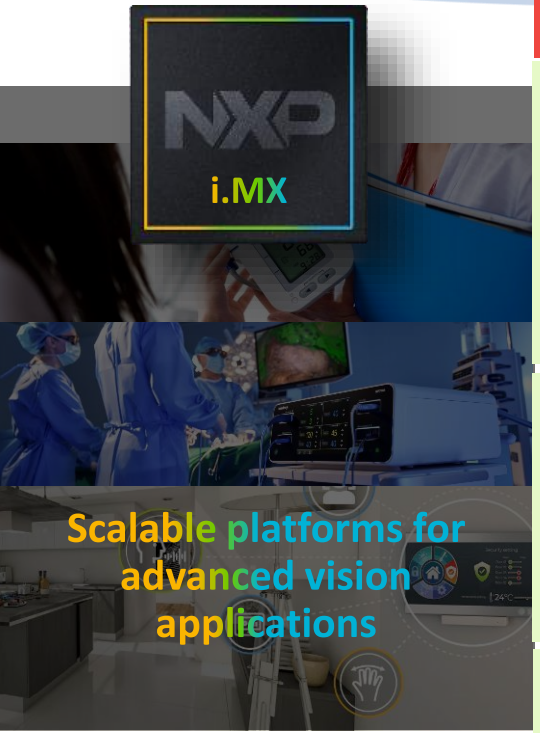
eIQ® GenAI Flow

Context aware generative AI application development



Engaging with our customers to develop system solutions and solve challenges together

EMBEDDED VISION SUMMIT



i.MX 8M Mini
Essential HMI & Vision Platform

Compute
- Linux & Android OS
- up to 4 Arm Cortex-A53 cores at 1.8GHz
- Embedded real-time M4 CPU
- 32-bit LPDDR4/DDR4/DDDR3 L Memory

Visualization
- 8 GFLOPs 3D GPU
- OpenGL ES 2.0, Vulkan, OpenCL 1.2
- 2D GPU
- MIPI-DSI 1080P60 Display

Intelligence
- CPU-based Neural Net enabled via NXP eIQ Toolkit

Vision
- Up to 4 cameras with MIPI-CSI virtual lanes
- 1080P video encode/decode
- Pixel Compositor

Platform
- Commercial & Industrial Temperature Qualification
- 10-year longevity
- Secure Boot

Connectivity
- PCIe Gen. 2 x1 Lane
- 1G Ethernet
- 2x USB 2.0
- 3x SD/eMMC

i.MX 8M Plus
Powerful HMI & Vision Platform with Edge AI & Industrial Connectivity

Compute
- Linux & Android OS
- up to 4 Arm Cortex-A53 cores at 1.8GHz
- Embedded real-time M7 CPU
- 32-bit LPDDR4/DDR4/DDDR3 L Memory

Visualization
- 16 GFLOPs 3D GPU
- OpenGL ES 3.1, Vulkan, OpenCL 1.2
- 1.3 Gpixel/s 2D GPU
- MIPI-DSI/ LVDS/ HDMI 1080P60 Display

Intelligence
- Neural Net Acceleration by embedded VSI VIP8000 NPU, GPU, and CPU
- Enabled via NXP eIQ Toolkit

Vision
- Up to 4 cameras with MIPI-CSI virtual lanes
- 375 Mpixel/s Image Signal Processor
- 12MP @ 30fps / 8MP @ 45fps
- 1080P video encode/decode

Platform
- Commercial & Industrial Temperature Qualification
- 10-year longevity
- Secure Boot plus Cryptographic Accelerator

Connectivity
- PCIe Gen. 3 x1 Lane
- 2x 1G Ethernet (1 w/TSN)
- USB 3.0 + 2.0
- 3x SD/eMMC
- 2x CAN-FD

i.MX 95 Family
Advanced HMI & Vision Platform with Safety, Security, and Next-Gen Edge AI

Compute
- Linux & Android OS
- up to 6 Arm Cortex-A55 cores at 1.8GHz
- Embedded real-time M7 CPU
- NXP SafeAssure Safety Domain
- 32-bit LPDDR5 /LPDDR4X Memory

Visualization
- 64 GFLOPs Arm Mali 3D GPU
- OpenGL ES 3.2, Vulkan 1.2, OpenCL 3.0
- 3 Gpixel/s 2D GPU
- 4K30P MIPI-DSI + 2x 1080P30 LVDS/ triple Display

Intelligence
- Embedded NXP eIQ® Neutron 1024S NPU
- up to 3x more AI acceleration than 8M Plus
- eIQ support for NPU, GPU, & CPU
- LLM and VLM support

Vision
- Up to 8 cameras with MIPI-CSI virtual lanes
- 500 Mpixel/s ISP with RGB-IR
- 12MP @ 45fps / 8MP @ 60fps
- 4K60P vid. codec
- Safe 2D display pipeline

Platform
- Extended Industrial Temp. Qual.
- 15-year longevity
- EdgeLock Secure Enclave + V2X Cryptographic Accelerator

Connectivity
- 2x PCIe Gen. 3 x1 Lanes
- 1x 10G Eth. (w/TSN)
- 2x 1G Eth. (w/TSN)
- USB 3.0 + 2.0
- 3x SD/eMMC
- 5x CAN-FD

NXP i.MX 95 Family for Automotive Edge, Industrial, & IoT

EMBEDDED VISION SUMMIT

Safety

Ditch the hypervisor and simplify building safety capable platforms with the first-generation on-die i.MX functional safety framework. Featuring NXP Safety Manager, Safety Documentation, & NXP Professional support to enable ISO26262 (ASIL-B) / IEC61508 (SIL-2) computing platforms, including 2D display pipeline.

Intuitive Decisions

Deliver increased accessibility and augment complex interfaces with Generative AI-enhanced voice command & control with the first i.MX applications processor to integrate the new, efficient NXP eIQ® Neutron neural processing unit.

Connect & Secure

Build secure, private applications with peace of mind based on the combined capabilities of integrated security and authentication acceleration, including post-quantum cryptographic capabilities, and lifecycle management.

Visualize & Act

Responsive HMI for IoT, Industrial, and Automotive applications are easily created with NXPs partner ecosystem, unlocked by a powerful modern 3D graphics processor combined with strong, efficient hexacore application processor performance.

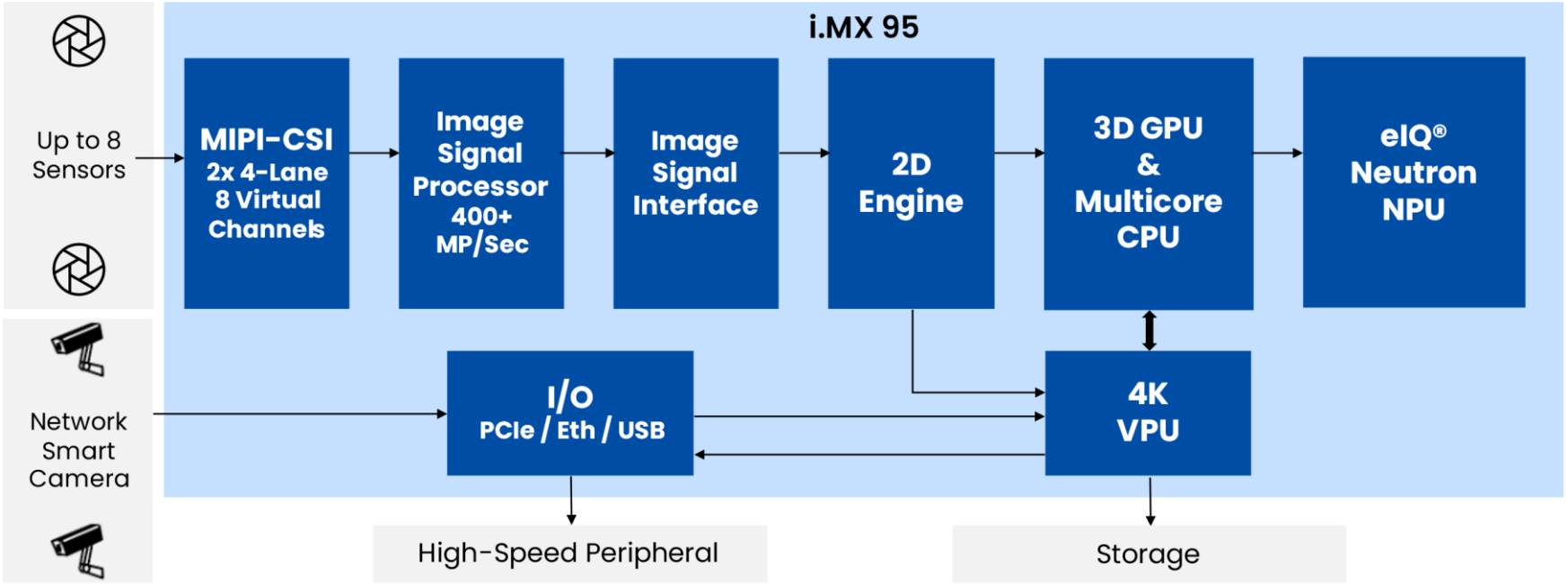
NXP.com/iMX95

Connectivity Leadership:
UWB, Wi-Fi, NFC, RFID, & BT

Co-Developed Platforms:
PMIC, Wi-Fi, Sensors, & More

Deep Application Insights:
26,000 Customers & Growing

i.MX 95 Vision Processing Pipeline



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

i.MX 95 and i.MX 8M Plus ISP

Specification/Feature	i.MX8M Plus	i.MX 95
Pixel Throughput	375 Megapixels/Sec	**500 Megapixels/Sec**
Image Resolution	12MP @ 30fps 8MP @ 45fps	**12MP @ 45 fps** **8MP @ 30 fps**
Streaming Mode Support	Yes	**Yes**
Memory-to-Memory Support	No	**Yes**
RGB-IR Support	No	**Yes (4x4 Array Pattern)**
High Dynamic Range (HDR) Support	12-bit	**20-bit**
Chromatic Aberration Support	Yes	**No**
Statistics	Block	**Advanced Auto White Balance (AWB)**
Output Formats	YUV 420 YUV 422	**YUV 420, YUV 422** **YUV 444, RGB 888**
S/W Enablement	3rd Party	**NXP Provided Toolchain**
OS Support	Linux oriented	**OS Agnostic S/W Stack**
S/W Stack	V4L Layer provided on top of a native S/W Stack	**Direct Integration into V4L** **LibCamera support (Default)**

Compute Bound vs Memory Bound

Compute-bound and memory-bound are terms used to describe the limitations of a computational task based on different factors:

Compute-Bound

A task is considered compute-bound when its performance is limited by processing power and the number of computations that need to be performed. Convolution Neural networks, **CNNs are typically compute bound** in embedded systems.

Memory-Bound

A task is considered memory-bound when its performance is limited by the speed and bandwidth of the memory system. **Generative AI workloads with large multi-billion parameter models are typically memory-bound** in embedded systems.

So the size and bandwidth of DDR memory available determines the time to first token (TTFT) and token per second (TPS) performance.

Generative AI and Transformer Models

Transformers and Generative AI dominating new AI development

What is Generative AI?

- Generative AI refers to deep-learning models that can take raw data and “learn” to generate probable outputs when prompted.
- Generative AI focuses on creating new content and data, while Traditional AI solves specific tasks with predefined rules.
- **Generative AI Models are based on the “Transformer” architecture**

How are Convolutional Neural Networks (CNNs) and Transformer Models different?

- Transformers require substantially more compute and have lower data or parameter parallelism.
- Transformers require a higher dynamic range of data which makes them less edge friendly.
- Transformers need more training data and training GPU performance to surpass CNN results.
- Transformer models are much larger than typical CNN models.

Transformer acceleration needs substantially more resources than more traditional convolutional AI models!

NXP to Acquire Kinara

California-based technology leader in offering **flexible, energy-efficient discrete NPUs** for Industrial and IoT Edge applications.

Discrete NPUs

Two generations capable of a variety of neural networks, incl. advanced generative AI

>500k NPUs shipped to date

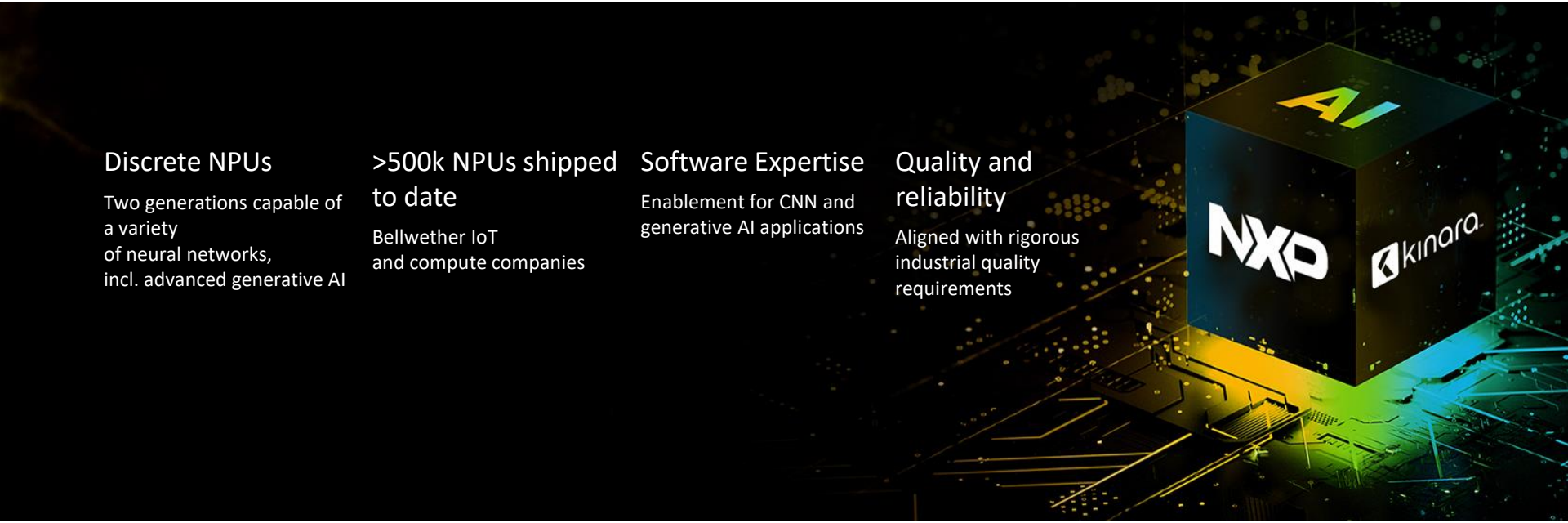
Bellwether IoT and compute companies

Software Expertise

Enablement for CNN and generative AI applications

Quality and reliability

Aligned with rigorous industrial quality requirements



Two Generations of AI Accelerators Optimized for Traditional & Generative AI Workloads

Ara-1

Latency Optimized for Edge Applications

10x Capex/TCO improvement over GPUs

Generative AI Capable

6 eTOPs. Up to 2GB LPDDR4

Ara-2: Vision, Multi-modal LLMs

Computer Vision, Generative AI optimized!

5-8X Performance improvement over Ara-1

Up to 40 eTOPs

Up to 16GB LPDDR4

* eTOPS: equivalent TOPs , performance comparison used to derive value as the ARA architecture is not a traditional MAC Array

Ara-2 High Level Features

- Up to 40 eTOPS*. 6.5 W typ. power. 17 mm x 17 mm EHS-FCBGA
- Host interface (x86 or ARM) PCIe or USB
  - PCIe: Up to 4-lane Gen 3/4 Endpoint. x1, x2 and x4 modes. 16 Gbps per lane
  - USB: 3.2 Gen1/2. 10 Gbps. Supports USB Type-C connector. Gen2 also supported
- External DDR memory options: Up to 16 GB density
  - 1-2GB for most vision use cases and 4/8/16 GB for Gen AI
  - LPDDR4 or 4X
  - Single 64-bit or two 32-bit memory devices
- Industrial grade qualified (-40 to 85C ambient)

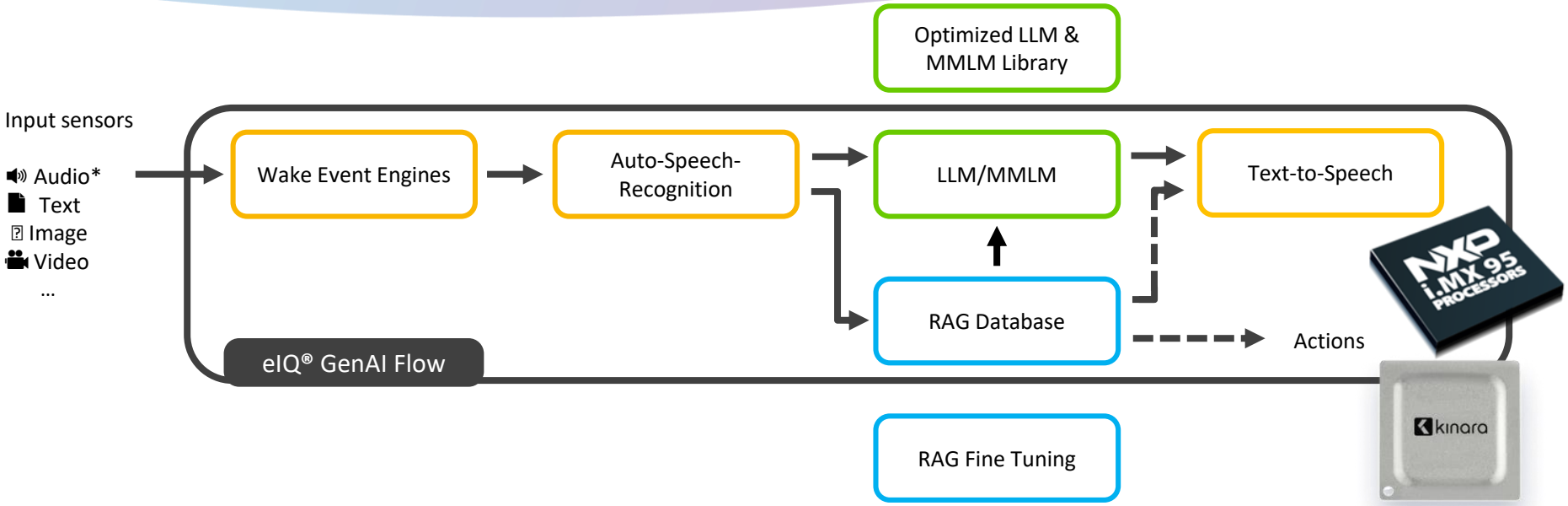
* eTOPS: equivalent TOPs , performance comparison used to derive value as the ARA architecture is not a traditional MAC Array

Why ARA Discrete AI Accelerators

System level features for selecting a discrete AI accelerator:

- **Performance and Efficiency:** Ara devices reduce the time and energy required for wide array of AI tasks like deep learning, large language models, multi-modal generative AI models.
- **Parallel Processing:** Ara devices can handle multiple data streams and multiple concurrent model executions.
- **Scalability:** Ara accelerators can be scaled to handle larger workloads or expanded AI applications. This scalability ensures that AI systems can grow and adapt to increasing demands without significant overhauls.
- **Memory bandwidth:** Ara devices support high transfer rate DDR, which is needed to run multi billion parameter generative AI models.
- **Connectivity:** Ara devices support up to 4 lanes of PCIe gen3/4 for handling high bandwidth connections when pairing with host controllers to provide inference on more data inputs. Ara devices also support USB and Ethernet connection options to provide flexibility in system design
- **Flexibility:** Ara devices have programmability and flexibility allowing newer models and operators to be supported without any hardware changes.
- **SW Enablement:** Ara devices are supported by an intelligent AI compiler that automatically determines the most efficient data and compute flow for any AI graph.

eIQ GenAI Flow: Bringing Generative AI to the Edge



Library of Functional Blocks
Necessary building blocks needed to create real Edge GenAI applications.

RAG
Secure method of fine-tuning: customers' private knowledge sources aren't passed to the LLM training data.

Transformers
Require specific types of optimization to be small and fast enough for edge devices.

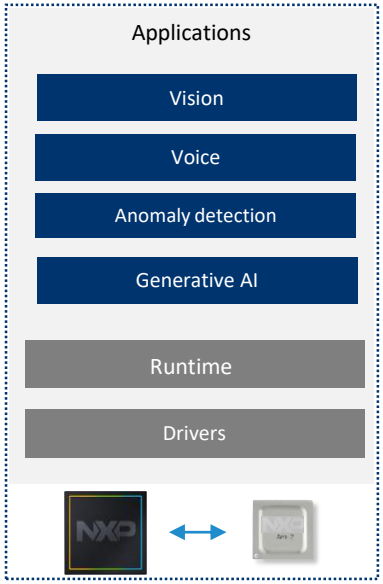
Dashed arrow shows possible pathway using pre-defined intents (no LLM)

Simple i.MX and Ara Decision Tree

Currently using:	Application is:	Wants to:	Recommended path:	Additional expansion:
i.MX 8M Plus	Vision based classification and detection use case	Extend existing product with more AI capabilities and performance	**Add ARA device** on PCIe to reuse existing applications with more AI performance	
i.MX 8M Plus	Vision based classification and detection use case	Design new product with more AI performance and possibly higher resolution camera, or more camera sensors	**Select i.MX 95** to get higher AI inference performance and higher camera pixel throughput	**Add ARA device** to system to extend AI applications with more AI performance
New design	Vision based classification and detection use case	Design state of the art vision AI system	**Select i.MX 95** as applications processor	**Add ARA-2 device** to system to extend AI applications with more AI performance as needed
New design	Gen AI for conversational HMI and system health monitoring	Build a solution with generative AI for better system monitoring and operator user experience	**Select i.MX 95** as applications processor. **Support <4B parameter LLMs**	**Add ARA-2 device** to system to extend for support >4B parameter Gen AI models
New design	Gen AI multi-modal video event and scene understanding	Use gen Ai models to build applications with vision, audio and sensor signals	**Select i.MX 95 and ARA-2 device**	**Add additional ARA-2 device** to system to extend

For guidance only, it is possible to select other i.MX 8M * and i.MX 9* portfolio of applications processors to pair with Ara DNPU devices.

Ara AI SW Enablement



Ara AI SDK combined with i.MX BSP and eIQ AI SW suite comprehensive solution for immediate engagement

Creates optimal execution plan

AI Compiler automatically determines the most efficient data and compute flow for any AI graph

Readily support new operators

Fully programmable compute engines with a neural-optimized instruction set

Efficient dataflow for any network architecture type

Software defined Tensor partitioning and routing optimized for dataflow

Extensible compiler

Converts and schedules models ranging from CNNs to complex vision transformers and Generative AI

Support for multiple datatypes INT8, INT4 and MSFP16

Utilizes flexible quantization methods

Choose between Kinara integrated quantizer or TensorFlow Lite and PyTorch pre-quantized networks

GenAI on the Edge: Cloud Experience on the EDGE

EMBEDDED VISION SUMMIT

Multimodal Gen-AI @ Edge with Voice UI

(Model: LLaVA - Large Language and Vision Assistant)

Available soon in GoPoint

kinara Ara-2

NXP

The image shows a close-up view of an industrial setting, focusing on two pieces of equipment that appear to be part of a larger machine or system. On the left side, there is what looks like an electrical control panel with various components and possibly some wiring or connections visible. The right side features what appears to be a large metal container with latches for securing lids.

There are no people present in this image; it's purely focused on these mechanical parts without any human context. The floor has some water stains and marks suggesting recent cleaning efforts or possible spills from the machinery above.

The style of the photograph seems candid rather than staged, capturing elements in their natural state within an industrial environment without artistic manipulation beyond framing the subject matter as seen through its lens

CANCEL

CLIP inf. time: 428 ms | Time to first token: 9.363000 s | Tokens/Second: 6.503015

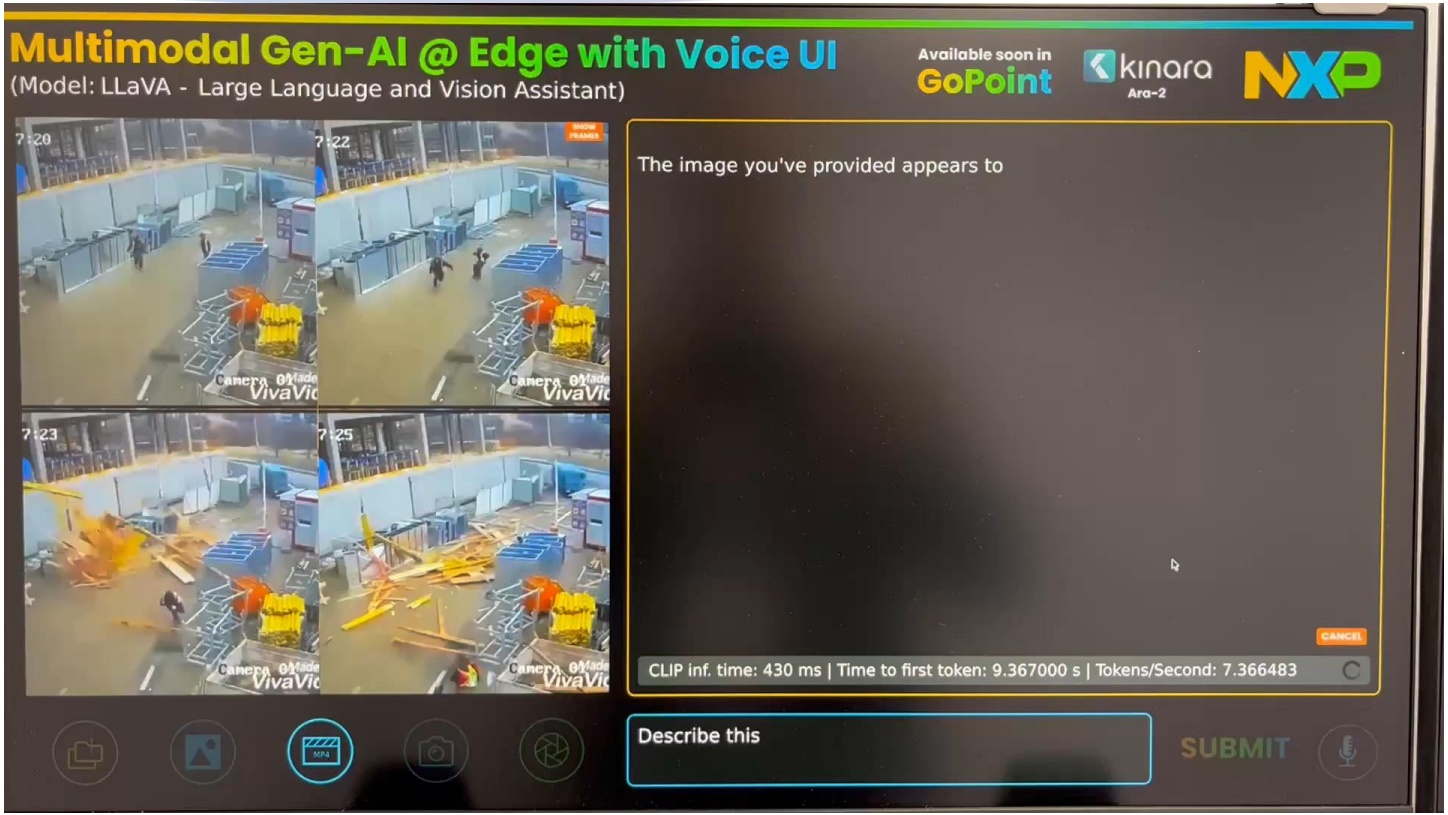
Describe this image

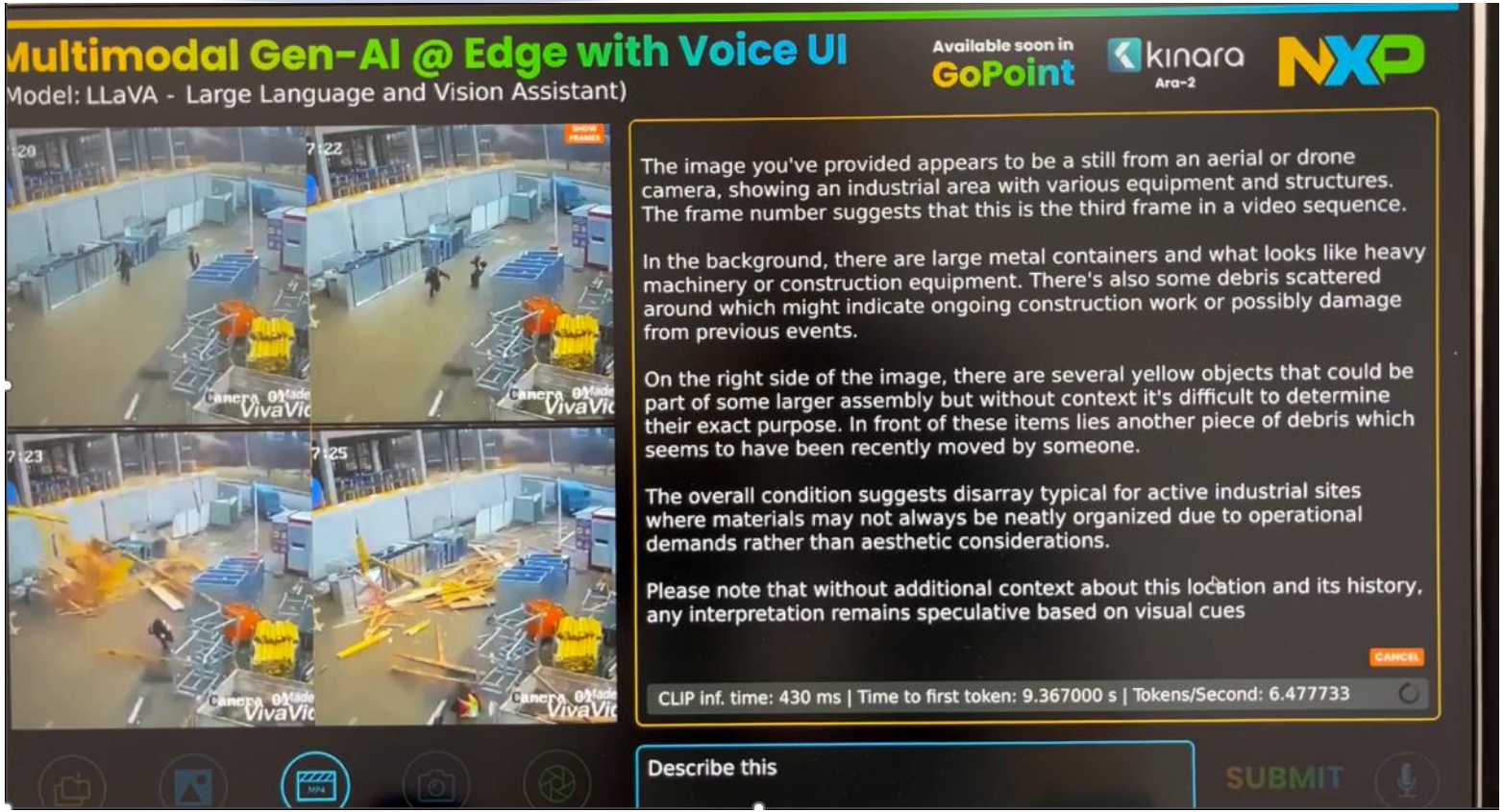
SUBMIT

NXP

GenAI on the Edge: Cloud experience on the EDGE
Occupational Health and Safety GenAI Example

Occupational Health and Safety




Multimodal Gen-AI @ Edge with Voice UI
Model: LLaVA - Large Language and Vision Assistant)
Available soon in
GoPoint
kinara
Ara-2
NXP
The image you've provided appears to be a still from an aerial or drone camera, showing an industrial area with various equipment and structures. The frame number suggests that this is the third frame in a video sequence.
In the background, there are large metal containers and what looks like heavy machinery or construction equipment. There's also some debris scattered around which might indicate ongoing construction work or possibly damage from previous events.
On the right side of the image, there are several yellow objects that could be part of some larger assembly but without context it's difficult to determine their exact purpose. In front of these items lies another piece of debris which seems to have been recently moved by someone.
The overall condition suggests disarray typical for active industrial sites where materials may not always be neatly organized due to operational demands rather than aesthetic considerations.
Please note that without additional context about this location and its history, any interpretation remains speculative based on visual cues
CANCEL
CLIP inf. time: 430 ms | Time to first token: 9.367000 s | Tokens/Second: 6.477733
Describe this
SUBMIT

Why Discrete AI Accelerators

- Leveraging discrete AI accelerators like the Ara-2 offer improvements in several key areas for Edge AI solutions:
  - Performance:
    - They use specialized architectures that are optimized for AI workloads and can provide path to scale beyond the native AI performance for i.MX applications processors.
  - Scalability:
    - These accelerators can be scaled to meet increasing demands, ensuring that systems can grow seamlessly without necessitating changes to the i.MX applications processor. This scalability is crucial for accommodating expanding AI applications and workloads with faster time to market.
  - Flexibility:
    - They can be used to adapt to changing processing needs, like new operators and models like LLMs and emerging paradigms like Agentic AI and Physical AI providing versatility needed to handle diverse and dynamic tasks.

Resources and Links

- AI and Machine Learning at NXP Semiconductors (www.nxp.com/ai)
- eIQ® ML Software Development Environment (www.nxp.com/eiq)
- eIQ GenAI Flow Demonstrator on ACH (https://mcuxpresso.nxp.com/appcodehub?search=dm-eiq-genai-flow-demonstrator)
- eIQ Neutron Neural Processing Unit (NPU) | NXP Semiconductors (www.nxp.com/neutron)
- Kinara AI Accelerators (www.kinara.ai)