Fact Sheet

Ara-1

Combine performance, power and value to optimize AI at the edge

Machine Learning acceleration for AI-integrated cameras, edge computer appliances and embedded systems that demand low latency, flexibility and scalability.

Target applications

- Factory automation
- Smart retail
- Home entertainment
- Security and surveillance
- Smart cities

Increase performance by offloading inferencing

Built around an efficient dataflow architecture, Ara-1 Discrete Neural Processing Units (DNPUs) deliver the performance and responsiveness needed for real-time AI computing and decision-making. The Ara-1 enables applications to run multiple AI models with zero latency context-switching. Enables AI-integrated cameras and embedded systems that demand low latency and higher flexibility to accommodate new model operators.

The Ara-1 DNPU delivers up to 10× Capex/TCO efficiency over GPUs, supports generative AI, achieves up to 6 eTOPS* performance and integrates 2 GB LPDDR4 memory.

Features

- Up to 6 eTOPS*
- Supports AI model frameworks: TensorFlow, PyTorch, ONNX
- Patented polymorphic dataflow architecture

*eTOPS = equivalent TOPS

Key benefits

- Enable real-time AI computing and decision-making
- Run multiple AI models with zero latency context-switching
- Ara-1 M.2 (M-Key) and USB modules for compact, plug and play AI acceleration
- Exceptional performance/watt inference
- Achieve low latency results with high accuracy

Software development kit

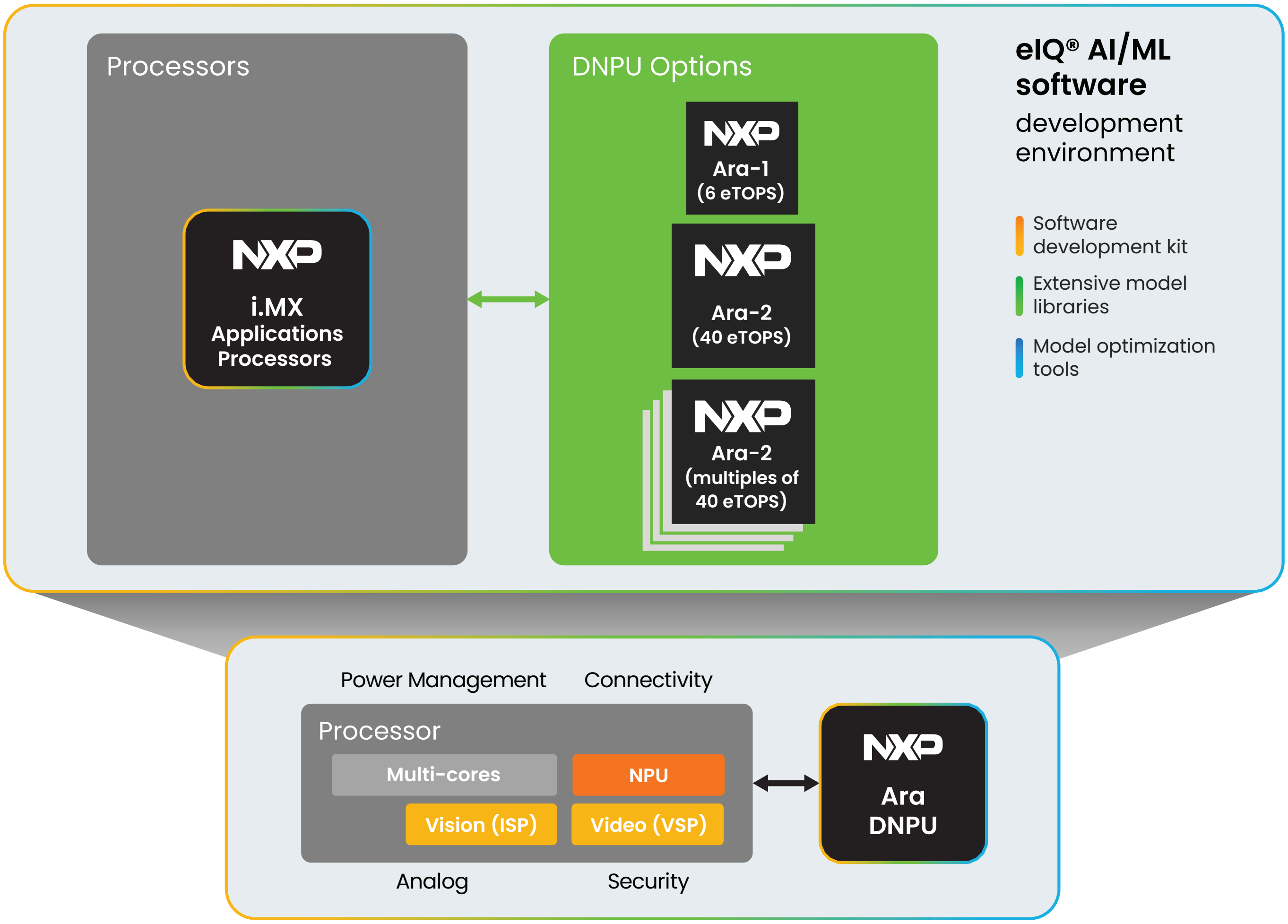
Ara AI Software Development Kit (SDK) streamlines deployment of AI models onto Ara-1 DNPUs and modules. It features an extensible compiler that supports diverse model architectures, including CNNs and Large Language Models (LLMs). The SDK offers efficient dataflow optimization, flexible quantization methods and support for multiple datatypes to determine the most efficient data and compute flow for any AI graph.

Specifications	
AI model frameworks supported	TensorFlow, PyTorch, ONNX
Performance[1]	Resnet50-v1: 100 inferences/sec. MobileNet-v1: 554 inferences/sec. Up to 6 eTOPS*
Operating system support (Runtime)	Linux
Interface	PCIe Gen 3 x 4, USB 3.2 Gen1 x1
Power consumption (Typical)	1.7 W @ 600 MHz
Packaging	15 mm x 15 mm EHS-FCBGA

[1] Maximum performance based on peak computational throughput of Ara-1 (800 MHz) and Host System. Specification subject to change without notice. Performance may vary depending on system configuration.

*eTOPS = equivalent TOPS

NXP intelligent edge AI platform



nxp.com/Ara1