Fact Sheet

Ara-2

Accelerate Generative AI and Large Language Models at the edge

High-performance, energy-efficient Discrete Neural Processing Units (DPNUs) are programmable to run a wide range of neural networks, including transformers for multi-modal Generative AI and Large Language Models at the edge.

Target applications

- AI assistance, Copilot
- Factory automation
- Smart retail
- Home entertainment

Designed for real-time vision and AI workloads

The Ara-2 DNPU enables real-time Generative AI and Large Language Models execution on AI-enabled compute and embedded systems, delivering low latency, lower operational costs and enhanced data privacy. Its innovative architecture combines balanced compute, large on-chip memory and high off-chip bandwidth to efficiently execute large models.

Features

- Up to 40 eTOPS*
- Access up to 16 GB LPDDR4(X) memory
- Supports AI model frameworks: TensorFlow, PyTorch, ONNX
- Secure boot and root-of-trust processor

*eTOPS = equivalent TOPS

The Ara-2 DNPU delivers 5–8× performance improvement over the Ara-1, with up to 40 eTOPS* performance and integrates 16 GB LPDDR4 memory.

Key benefits

- Enable real-time AI computing and decision-making
- Exceptional performance/watt inference
- Meet high performance needs of computing and embedded systems and laptops
- Process multiple models without incurring switch-time performance penalties
- Ara-2 M.2 (M-Key) and USB modules for compact, plug and play AI acceleration

Maximize edge AI performance

Ara-2 DNPU is designed to maximize edge AI performance while providing the flexibility to adapt as AI models evolve. The architectural flexibility of Ara-2 allows seamless support for current and future workloads — from CNNs to Generative AI and emerging agentic AI approaches — ensuring long-term platform longevity. Delivering up to 40 eTOPs*, Ara-2 can be easily integrated into new or existing embedded systems, making it ideal for upgrading in-field devices and accelerating time to-market for next-gen AI applications.

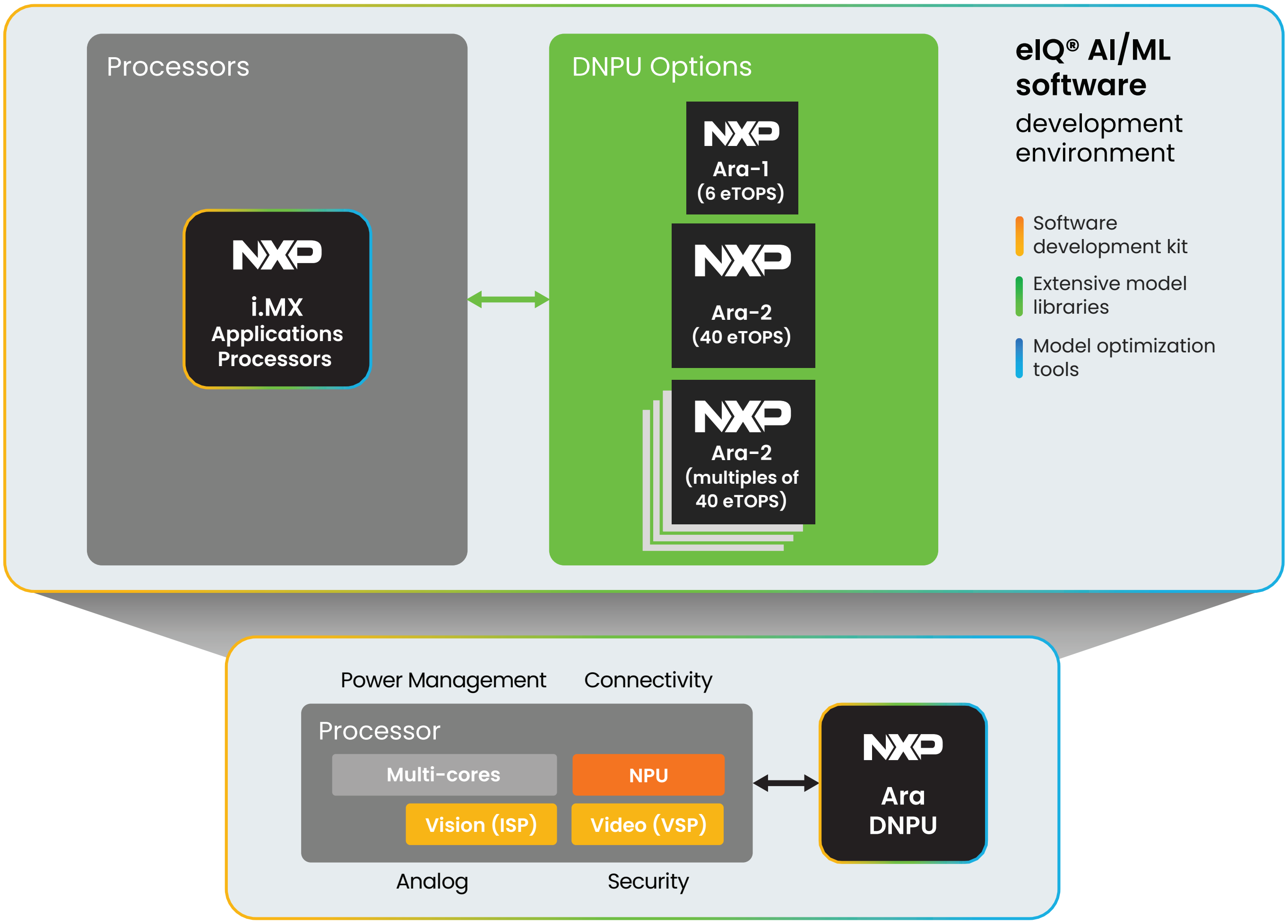
Software development kit

Ara AI Software Development Kit (SDK) streamlines deployment of AI models onto Ara-2 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	LLaMa-7B: 14 output tokens/sec MobileNetV1 SSD: 974 IPS (1.03 ms latency) Up to 40 eTOPS*
Security	Secure boot, root-of-trust processor
Memory interface	Up to 16 GB LPDDR4 (X)
Operating system support (Runtime)	Linux, Windows
Host interface	4-lane PCIe Gen 4, USB3 Gen 2
Chip package	17 mm x 17 mm FCBGA
Power consumption (Typical)	<2 Watts

*eTOPS = equivalent TOPS

NXP intelligent edge AI platform



nxp.com/Ara2