加速在边缘端部署生成式AI和大语言模型

高性能、高能效的独立神经处理单元（DNPU）具备可编程能力，可运行包括Transformer在内的多种神经网络，支持在边缘端部署多模态生成式AI和大语言模型。

目标应用

- 工业自动化
- 楼宇与能源管理
- 座舱与高级驾驶辅助系统（ADAS）
- 自动化家居

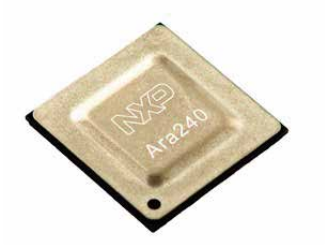
专为实时可见和AI工作负载而设计

Ara240 DNPU可在支持AI的计算平台和嵌入式系统上实时运行生成式AI及大语言模型（LLM），实现低延迟和更低运营成本，同时进一步增强数据隐私保护。Ara240采用创新架构，将均衡的算力、大容量片上存储与高带宽片外互联相结合，能够高效运行大型模型。

特性

- 高达40 eTOPS的算力*
- 可支持高达16GB的LPDDR4内存
- 支持TensorFlow、PyTorch和ONNX等AI模型框架
- 安全启动，内置信任根处理器

*eTOPS = 等效每秒万亿次运算



与Ara-1相比，Ara240 DNPU可实现5–8倍性能提升，算力可达40 eTOPS*，集成16GB LPDDR4内存。

核心优势

- 实现实时AI计算与决策
- 卓越的每瓦推理性能
- 满足计算系统、嵌入式系统及笔记本电脑对高性能的需求
- 同时处理多个模型，避免因模型切换造成性能损耗
- 提供Ara240 M.2（M-Key）和USB模块，实现简约、即插即用的AI加速方案

大幅提升边缘AI的性能

Ara240 DNPU旨在大幅提升边缘AI的性能，同时具备灵活的适应能力，可随AI模型的不断演进进行扩展。Ara240处理器采用灵活的架构设计，能够无缝支持当前及未来的各类工作负载，涵盖卷积神经网络（CNN）、生成式AI模型以及新兴的智能体AI应用，从而延长平台的使用寿命。Ara240的算力可达40 eTOPS*，能够轻松集成到全新或现有的嵌入式系统，非常适合用于现场设备升级，并能加速新一代AI应用的上市进程。

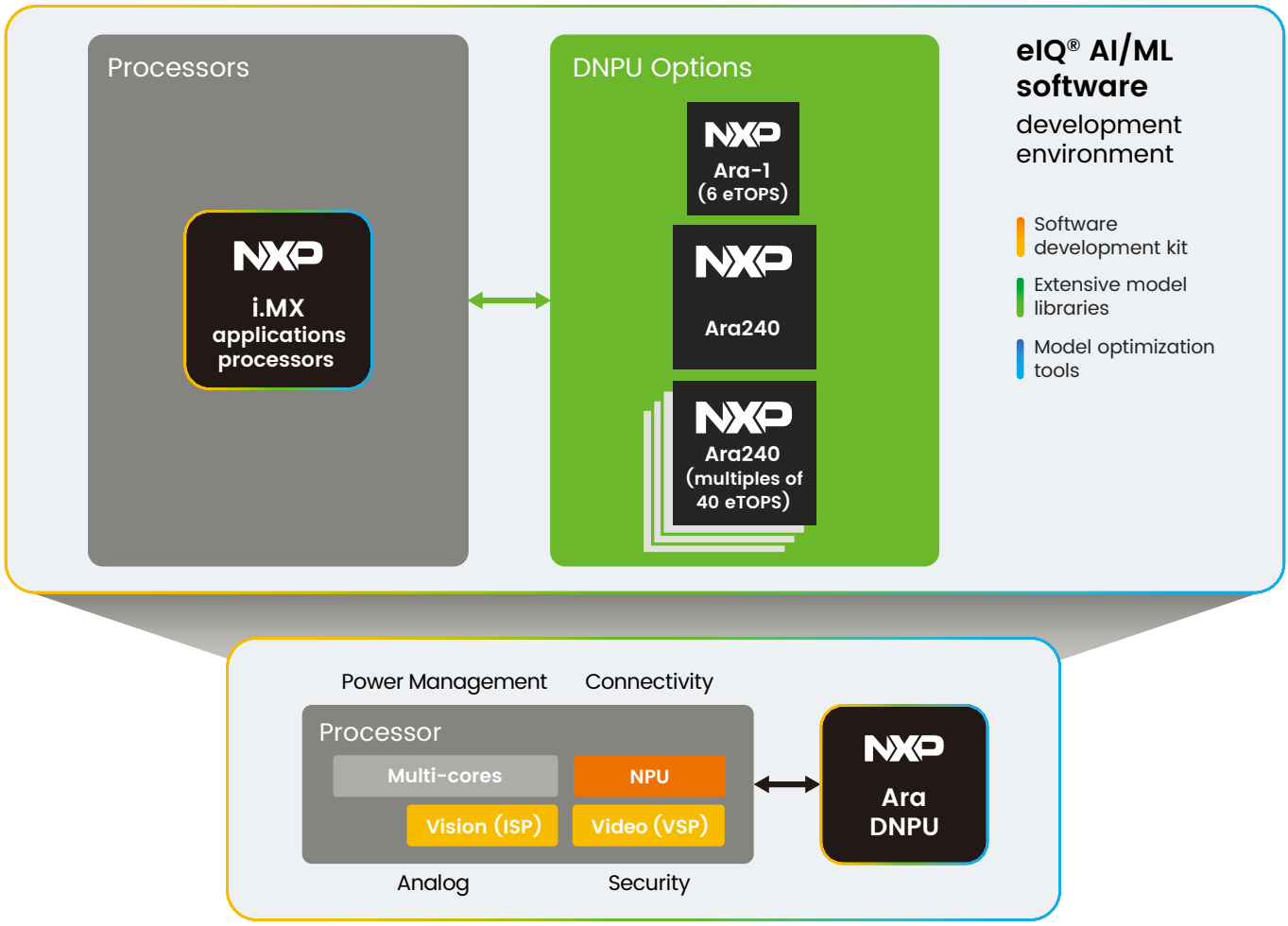
软件开发套件

Ara软件开发套件（SDK）可简化AI模型在Ara240 DNPU及相关模块上的部署。该SDK配备可扩展的编译器，支持卷积神经网络（CNN）和大语言模型（LLM）等多种模型架构。同时，它还提供高效的数据流优化、灵活的量化方法以及对多种数据类型的支持，能够为任意AI计算图确定高效的数据流与计算流方案。

规格	
支持的AI模型框架	TensorFlow、PyTorch、ONNX
性能	Llama2-7B: 每秒输出14个词元 ResNet34: 660 IPS（每秒推理次数） YOLOv8n: 313 IPS（每秒推理次数） 高达40 eTOPS*
信息安全	安全启动、信任根处理器
内存接口	高达16GB LPDDR4
支持的操作系统（运行时）	Linux
主机接口	4通道PCIe Gen 4, USB3.2 Gen 1
芯片封装	17 mm x 17 mm FCBGA
典型功耗	6.5瓦

*eTOPS = 等效每秒万亿次运算

恩智浦智能边缘AI平台



nxp.com/Ara240

文档编号：KINARAARA2FS REV 2