kinara
Leading Edge AI

PRODUCT BRIEF

KU-2, Ara-2 USB Module

Powering Large Language Models on PCs and embedded systems

The KU-2 is a compact USB module that is powered by Ara-2, Kinara's latest AI processor. It provides a plug-and-play option for users to upgrade their PCs and embedded systems to join the new age of Generative AI. Run large language models (LLMs) for increased productivity or stable diffusion models to generate novel images.

Built on the same flexible and efficient dataflow architecture as the Ara-1, the 40 TOPS Ara-2 boosts performance up to eight times, resulting in tremendous increases in compute efficiency.

A Powerhouse Accelerator for Edge AI Applications

· Run large language models at the edge, enabling real-time creativity and productivity

· Fuel powerful transformer-based models, driving breakthroughs in video analytics and natural language processing.

· Revolutionize edge applications from smart cities to smart retail to manufacturing.

· Utilize Ara-2's perfectly balanced compute, on-chip memories, and high off-chip bandwidth to execute very large models with extremely low latency.

· Process multiple models without incurring switch-time performance penalties.

AI Acceleration for Many Applications

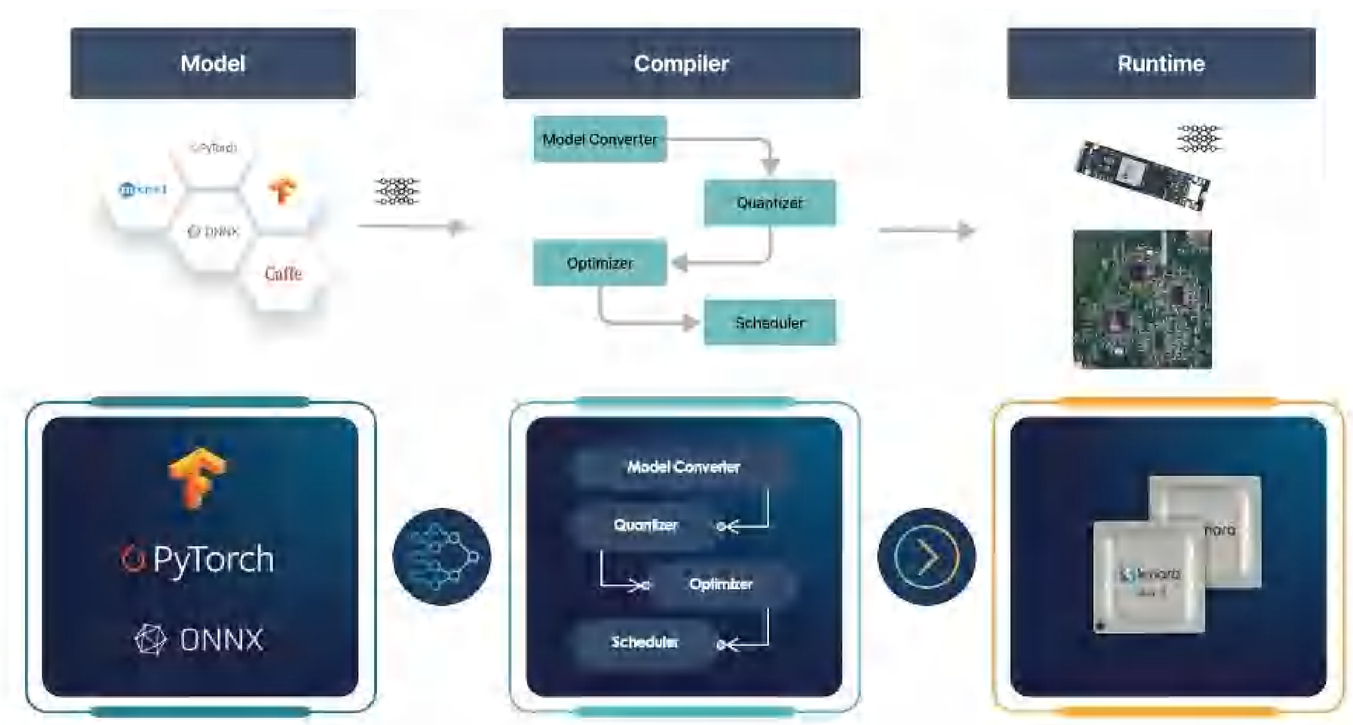
· AI Assistance, Copilot
· Gaming
· Smart retail
· Physical security
· Factory automation

Inquire Now

kinara
Leading Edge AI

Key Features

AI Frameworks	TensorFlow, PyTorch, ONNX
Performance	Stable Diffusion 1.4: 7 sec/image Llama2-7B: 12 output tokens/sec MobileNetV1 SSD: 974 IPS (1.03 ms latency)
Security	Secure Boot, Root-of-trust processor, encrypted interface
Memory (LPDDR4)	8GB / 16GB
Operating System Support	Linux, Windows
Host Interface	USB 3.2 Gen 2 (10 Gbps)
Dimensions (W x L)	30mm x 80mm
Operating Temperature	0°C – 35°C
Power Consumption	3W (typical workload) 8W (full performance)
Thermal Management (Typ.)	Active cooling (heatsink with fan)
Host Processor	x86, Arm
Certification	CE / FCC Class B (pending)
Ordering Information	ARA-U421-C-8 (8GB version) ARA-U421-C-16 (16GB version)



Kinara's end-to-end software seamlessly migrates trained AI models, including pre-quantized models. Our 'software first' approach enables users to run their own custom models without requiring any retraining. In addition, the software can run multiple models on the same stream without any model switching cost - resulting in low latency inferences for edge AI deployments. The software supports all state-of-the-art models including Generative AI and vision transformers.