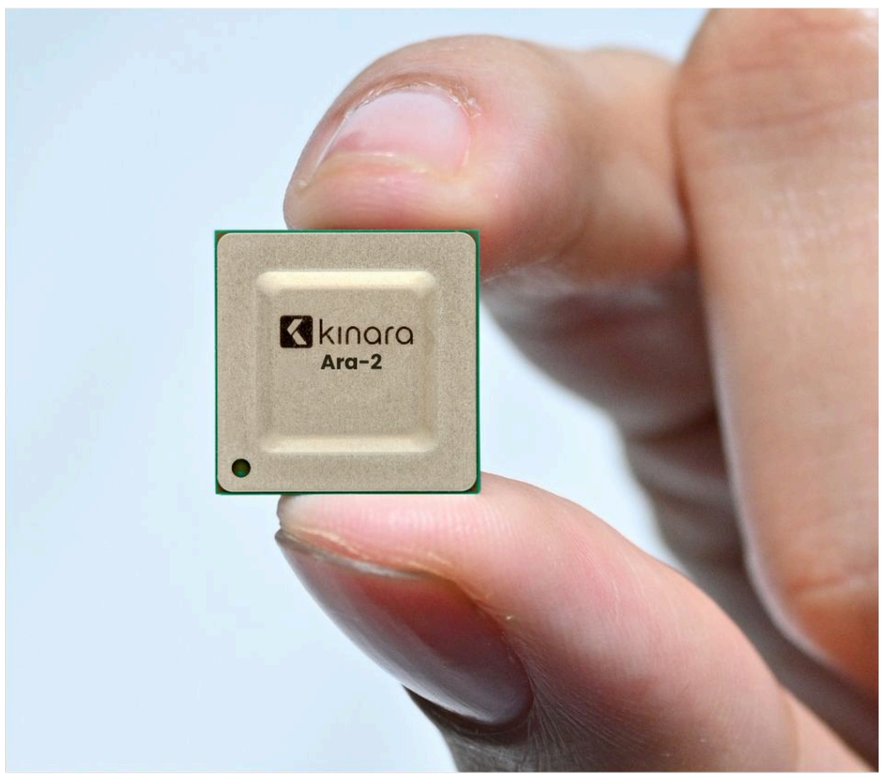
kinara
Leading Edge AI

# PRODUCT BRIEF

## Ara-2 Processor

## Shaping the future of Generative AI at the Edge

Meet the Kinara Ara-2 AI processor, the leader in energy efficient Edge AI processors, capable of tackling the massive compute demands of Generative AI and transformer-based models with unmatched cost-efficiency.




Built on the same flexible and efficient dataflow architecture as the Ara-1, the 40 TOPS Ara-2 boosts performance up to 8x, resulting in tremendous increases in compute efficiency.

### A Powerhouse Accelerator for Edge AI Applications

- Run complex Generative AI models including large language models (LLMs) on AI PCs, 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, all while minimizing cost and maximizing efficiency.
- Utilize Ara-2's real-time optimized design with 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.

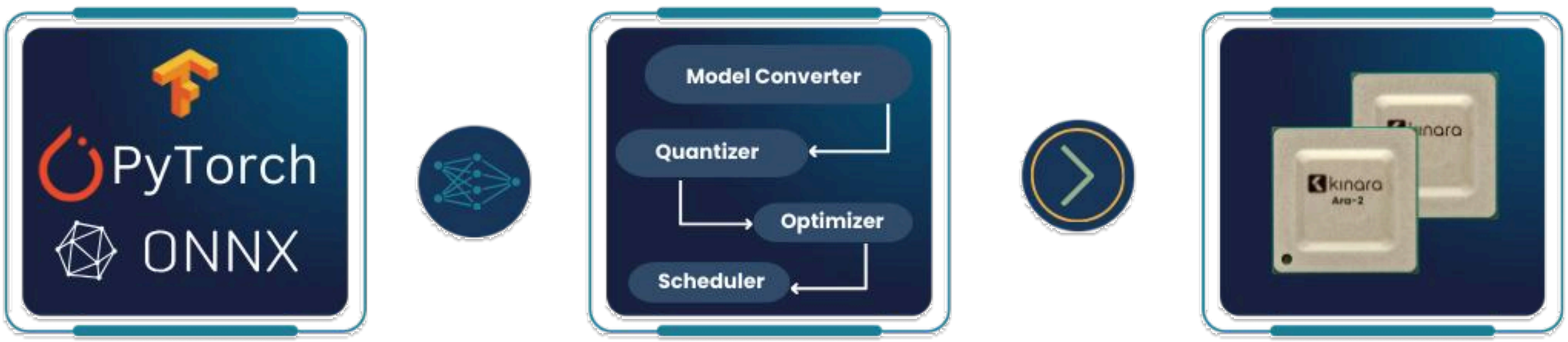
### AI Acceleration for Applications

- AI Assistance, Copilot
- Gaming
- Smart Retail
- Physical Security
- Factory Automation

kinara
Leading Edge AI

# Key Features

| | |
|---|---|
| **AI Model Frameworks Supported** | TensorFlow, PyTorch, ONNX |
| **Performance** | Stable Diffusion 1.4: 7 secs/image<br>LLaMA-7B: 12 output tokens/sec<br>MobileNetV1 SSD: 974 IPS (1.03 ms latency) |
| **Security** | Secure Boot, Root-of-trust processor, encrypted interface |
| **Memory** | Up to 16GB LPDDR4(X) |
| **Operating System Support** | Linux, Windows |
| **Host Interface** | 4-lane PCIe Gen4, USB3 Gen 2 |
| **Chip package** | 17 mm x 17 mm FCBGA |
| **Power Consumption (Typical)** | <2 Watts |




Kinara's end-to-end software seamlessly migrates trained AI models, including **pre-quantized models**. We enable 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**.