WHY KAIRÓS

What if everything ran Locally?

- Full privacy.
- True autonomy.
- No subscriptions.
- Offline by design.
- Everything on device.

Vision | Language | Reasoning | Voice

- Deploy LLM & VLM **right from HuggingFace**
- **No preprocessing** required
- Concurrent **multimodal execution**

MAIN SUPPORTED MODELS

Meta
Llama 2 - 7B
Llama 3.1 - 8B
Llama 3.2 - 1B - 3B

Alibaba
Qwen 2.5 Coder
Qwen-3 0.6B - 32B
Qwen-3 30B-A3B MoE

Microsoft
Phi-2 - 2.7B
Phi-3 mini - 4B
Phi-4 mini - 4B

deepseek
DeepSeek R1 Distill LLaMa - 8B
DeepSeek R1 Distill Qwen 1.5B - 14B
DeepSeek R1 0528 Qwen 3 - 8B

Fraunhofer
Teuken - 7B

OpenAI Whisper ASR

Google Gemma 3 - 1B

Moondream 2 - VLM

vyvo™ -TTS

Mistral Small 24B

TII Technology Innovation Institute — Falcon 3 - 1B

ACCELERATING GENERATIVE AI

The future of AI runs on-device too. No cloud needed.

RAIDERCHIP.AI

LIVE DEMO

Seen it live at our booth? See it in action again.

Try it on your site available under NDA.

RaiderChip
Accelerating Artificial Intelligence

raiderchip.ai
linkedin.com/company/raiderchip

RaiderChip

Accelerating Artificial Intelligence

KAIRÓS-1200

TSMC 7nm FinFET
128-bit LPDDR5X Memory
Embedded RISC-V

Local Generative AI, in silicon

All the intelligence, one chip

INDEPENDENT BENCHMARK

Highest Performance per Watt

KAIRÓS ASIC PROTOTYPE VS NVIDIA JETSON ORIN
102 GB/S VS 102 GB/S MEM BW

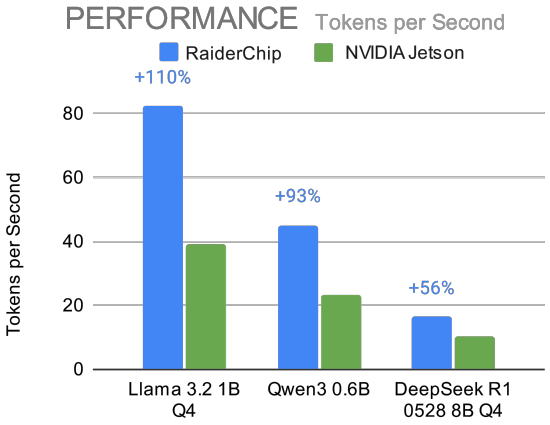
+110% Higher raw throughput

+63% Higher energy efficiency

+140% Better Energy Delay Product (EDP)

Reduced Power Consumption and Silicon cost

Maximum Tokens per Watt with Minimum silicon area · More tokens per dollar from foundry to deployment



RaiderChip NPU Redefines Efficiency

Source: "Edge Deployment of Small Language Models, a comprehensive comparison of CPU, GPU and NPU backends", University of Cantabria https://arxiv.org/abs/2511.22334

NPU ARCHITECTURE

Silicon-level AI. Transformer-optimized.

SPECIFICATIONS

TSMC 7nm FinFET
128-bit LPDDR5X Memory (154 GB/s) · Embedded RISC-V

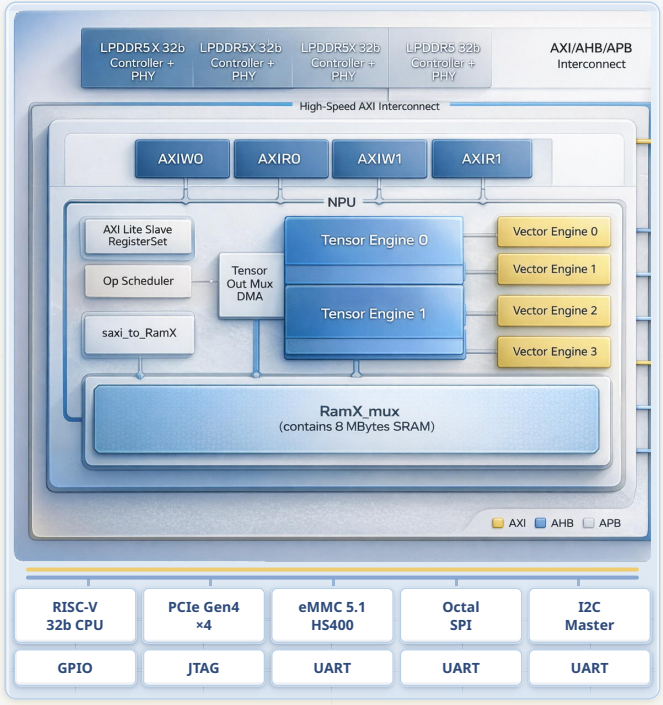
Up to 342 tokens/sec/user
Up to 64K tokens context window

Up to 32B parameters
Transformer-optimized NPU architecture

PRECISION SUPPORT

FP32 · TF32 · FP16 · BF16 · FP8 · Q5 · Q4

NPU BLOCK DIAGRAM



ENABLING REAL AI SYSTEMS

Cloud-level capabilities, delivered locally

Real Conversations

Up to 64K tokens context
Long complex interactions with full memory. No context loss

Deploy Instantly

Preserves original model architecture
Run models right from HuggingFace

Your AI, Your Rules

Full autonomy to run your original, post-trained and custom models. Zero vendor dependency

Full Accuracy

Full floating-point precision, delivering cloud-quality reasoning and outputs with no degradation from compression

Right Intelligence

Balance performance and efficiency
Choose the level of intelligence required for each task
FP32, TF32, BF16, FP16, FP8E5, FP8E4, Q5, Q4

See. Hear. Think. Act.

Foundation for modular, composable AI agent
Runs multiple AI models concurrently
Perception, reasoning, action, audio, vision, language... at the very same time