RaiderChip

Accelerating Artificial Intelligence

GenAI NPU IP core

Product Brief

GenAI Inference on Adaptive SoC

The **GenAI NPU** IP Core brings Generative AI inference directly to AMD Adaptive SoCs.

Deploy state-of-the-art **LLMs and VLMs locally**, right from **HuggingFace**, with **complete control over data, performance and cost**.

No cloud dependency. No third-party APIs. No variable fees.

Artificial Intelligence running on your AMD Versal devices:

- **Fully Offline**
- **Standalone**

AI Models. Freedom to Choose.

Add **vision, language, reasoning** and **voice** to your devices.

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

Alibaba
Qwen 2.5 Coder
Qwen-3 0.6B - 32B

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

Falcon 3 - 1B

Fraunhofer
Teuken - 7B

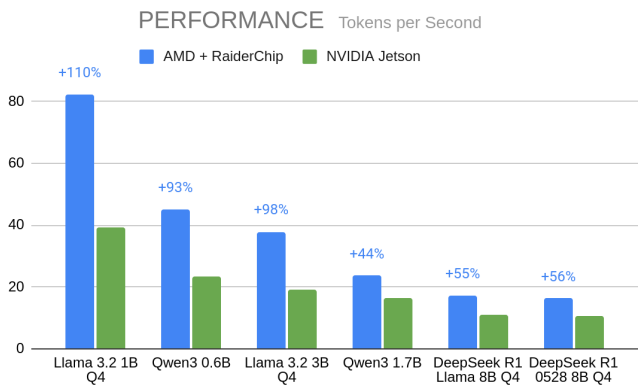
Google Gemma 3 - 1B

OpenAI Whisper ASR

Moondream 2 - VLM

VYVO™ -TTS

Supported AI models



Reference: *"Edge Deployment of Small Language Models, a comprehensive comparison of CPU, GPU and NPU backends"* (arXiv.org/abs/2511.22334)

Edge AI beyond GPUs.

Higher Performance per Watt

The GenAI NPU IP Core running on AMD Versal **surpasses GPU-based platforms, like Nvidia Jetson Orin (Super mode):**

Up to **+110%** higher raw **throughput**

Up to **+63%** higher **energy efficiency**

Up to **+140%** better **Energy Delay Product** (EDP)

RaiderChip NPU redefines energy efficiency.

Next-gen Artificial Intelligence locally

Protect critical data and operations by running fully **on premises** your **AI workloads**. No cloud transfers, no third-party access — **full control of your infrastructure and information**.

Operate independently in **remote environments without** requiring **network** connectivity.

Run cutting-edge open models locally. **No variable subscriptions to third parties, no API fees**.

On premises AI is:

Total Privacy 🔒

True autonomy 🤖

No subscriptions 🚫

Offline ✈️

Precision From Q4 to FP32

FP32 precision to deliver maximum intelligence to **run raw AI models natively**. Or **4-bit and 5-bit Quantization** for a **speed boost.**

Try our Generative AI Accelerator Demo

Everything provided to boot your board and turn it into a standalone **GenAI accelerator on your premises**.

Download AI models from HuggingFace, or run your own, natively on your site.

- Chat with **Vision, Reasoning, Language** models
- Evaluate **latency, throughput**, and **intelligence**

Post-trained and fine-tuned models are **supported natively** without sharing your proprietary weights.

Your model. Your weights. Your control.

Performance & Resource Usage

Select the device that best fits your performance, power, and cost requirements, **we take care of the rest!**

IP CORE VARIANT	1x	2x	4x	8x
MEMORY BANDWIDTH (directly proportional to achievable AI speed)	12 GB/s	24 GB/s	48 GB/s	96 GB/s
AI INFERENCE SPEED with 4-bits Quantization @ 250 MHz	Llama 3.2-1B: **16 tokens/sec** Llama 3.1-8B: **3 tokens/sec**	Llama 3.2-1B: **32 tokens/sec** Llama 3.1-8B: **6 tokens/sec**	Llama 3.2-1B: **58 tokens/sec** Llama 3.1-8B: **11 tokens/sec**	Llama 3.2-1B: **95 tokens/sec** Llama 3.1-8B: **18 tokens/sec**
Supported Data Formats	FP32, FP16, BF16, FP8, Q5_K (5-bits), Q4_K (4-bits)			
IP CORE SIZE LUT REG DSP B/URAM	 74K 141K 384 19/64-256	 110K 187K 648 19/64-256	 181K 277K 1176 19/64-256	 321K 458K 2232 19/64-256
SMALLEST SUPPORTED FPGA	AMD Versal VE2202 / VM1102	AMD Versal VE1752 / VM1302	AMD Versal VE1752 / VM1402 / VM1502	AMD Versal VP1202 / VM2502

RaiderChip
Calvo Sotelo, 4b - 1
39710 Solares - Spain

Phone: (+34) 942 941 060
Email: info@raiderchip.ai
Web: www.raiderchip.ai