RaiderChip

Accelerating Artificial Intelligence

GenAI NPU IP core

Product Brief

RaiderChip's GenAI NPU IP core is the first **Generative AI accelerator** licensable for use **in your FPGA based products**.

What is Generative AI?

The most advanced **Artificial Intelligence** capable of creating original content, such as text, images, ... thinking, reasoning and interacting like a human being.

Examples: **ChatGPT, Meta Llama**, or **Microsoft Phi** Large Language Models.

Why a Hardware IP core?

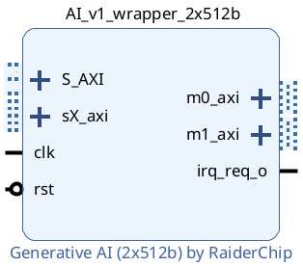
Generative AI models demand a **huge computing and memory bandwidth,** making it challenging for **CPUs** to run Artificial Intelligence efficiently. This results in **very low speeds** and the full utilization of processor and memory.

RaiderChip GenAI NPU Hardware IP core is designed to maximize memory bandwidth and computing performance, ensuring **real-time AI capabilities** while **keeping your CPU free** for other software tasks.

Which AI models can it run?

Any LLM model of your choice. Like Chat, Vision, Text-to-Speech, Automatic Speech Recognition, etc.

Fully accelerated models are **Meta Llama-2, 3, 3.1, 3.2, Microsoft Phi-2 and Phi-3**, **DeepSeek-R1** distills, **Qwen 2.5** LLMs, plus more being added every month.



Simple GenAI IP block for Vivado

Why Generative AI in my product?

Enhance your solutions with stand-alone, offline AI capabilities using our IP core, requiring no internet connection or subscriptions and ensuring full privacy.

- **Next level assistant**: LLMs understand and adapt to all user levels, from children to seniors
- **Multilingual**, **commercial-friendly** LLM AI models
- No User Manuals: **the device is the User Guide**
- Ideal for **privacy-sensitive** applications with offline functionality
- **Autonomous** operation anywhere, anytime
- **Subscription-free**, avoiding usage fees
- Supports custom, **fine-tuned models**

Generative AI is the Universal User Interface, providing text or optional voice interaction with your product's users.

Bring intelligence to your products today.

RaiderChip

Accelerating Artificial Intelligence

GenAI NPU IP core

Variants, Performance and Size

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: **2 tokens/sec**	Llama 3.2-1B: **30 tokens/sec** Llama 3.1-8B: **5 tokens/sec**	Llama 3.2-1B: **52 tokens/sec** Llama 3.1-8B: **10 tokens/sec**	Llama 3.2-1B: **80 tokens/sec** Llama 3.1-8B: **17 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	 89K 143K 386 10/64	 141K 211K 650 10/64	 244K 347K 1178 10/64	 451K 610K 2234 10/64
SMALLEST SUPPORTED FPGA	AMD Versal VE2202 / VM1102	AMD Versal VE1752 / VM1302	AMD Versal VE1752 / VM1402 / VM1502	AMD Versal VP1202 / VM2502

GenAI NPU is supported on all devices of the **AMD Versal and AMD UltraScale+ FPGA** families. Contact us for more information on support for other FPGA devices.

Check the latest product information on https://raiderchip.ai

Or get in touch at info@raiderchip.ai

Versal FPGA SoM fitting GenAI NPU 1x (actual size)

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

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