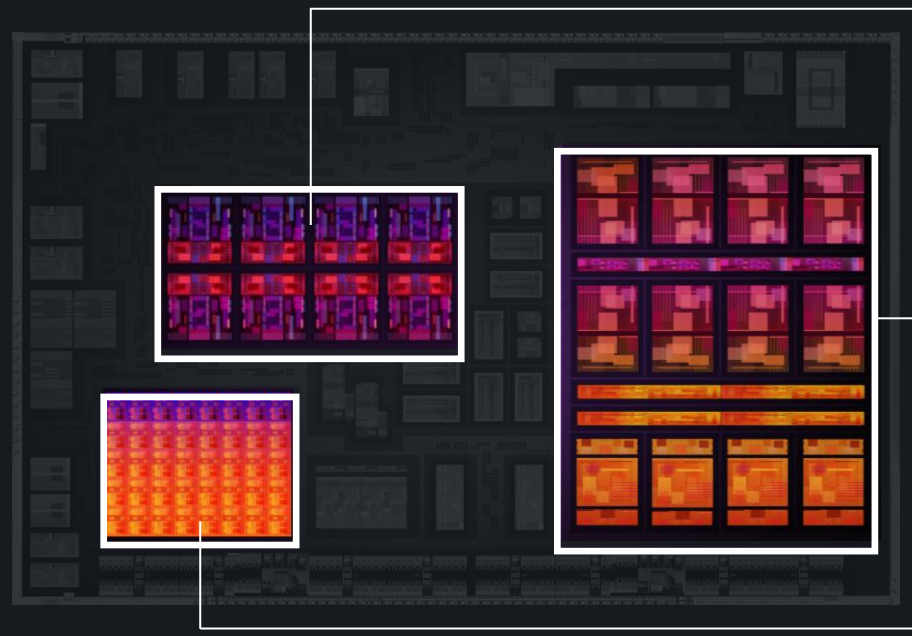
AI PC drive the future of AI in personal computing

AMD
RYZEN AI

Providing the Next Level of NPU, CPU, and GPU Architectures for Next-Gen AI PC Experiences

3rd Generation

AMD Ryzen™ AI



AMD RDNA 3.5

Next-Gen GPU
Up to 16 Compute Units

ZEN 5

Next-Gen CPU
Up to 12 Cores, 24 Threads

AMD XDNA 2

Next-Gen NPU
Industry-leading 50+ NPU TOPS

See endnote STX-04 and GD243

NPU Spatial Architecture for Concurrency

Architecture

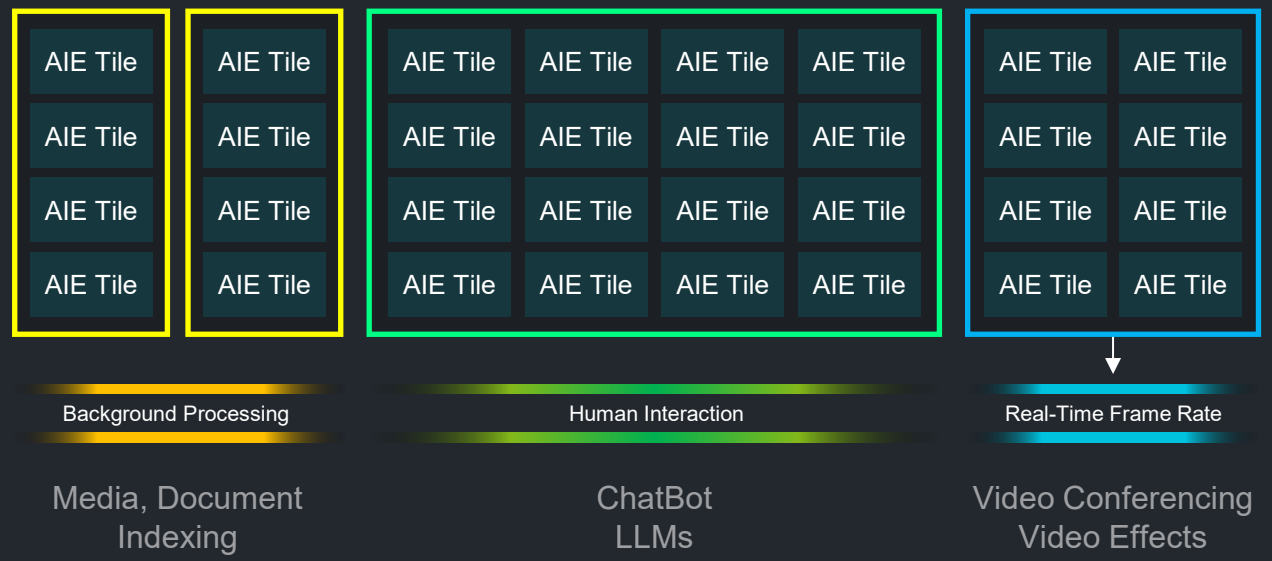
AMD Ryzen™ AI NPUs have spatial dataflow architecture ideal for AI workloads

2D tiled array of compute tiles with a flexible interconnect enabling data locality and runtime partitioning

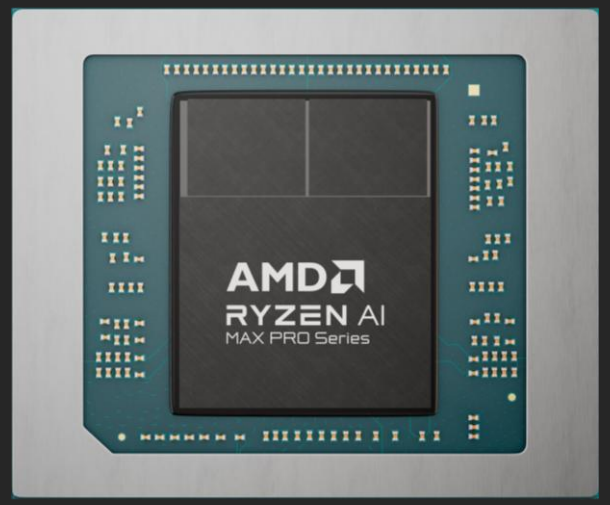
Benefit of Concurrency and Spatial Processing

- Spatial Configurability
- Efficient Multitasking
- Performance Segregation

Architecture



Introducing



AMD RYZEN™

AI Max PRO

Series Processors

Designed to power a new generation of compact Copilot+ PC workstations

Cutting Edge CPU and Memory

- Desktop-class “Zen 5” CPU cores
- Up to 128GB unified memory

Powerful Certified Graphics

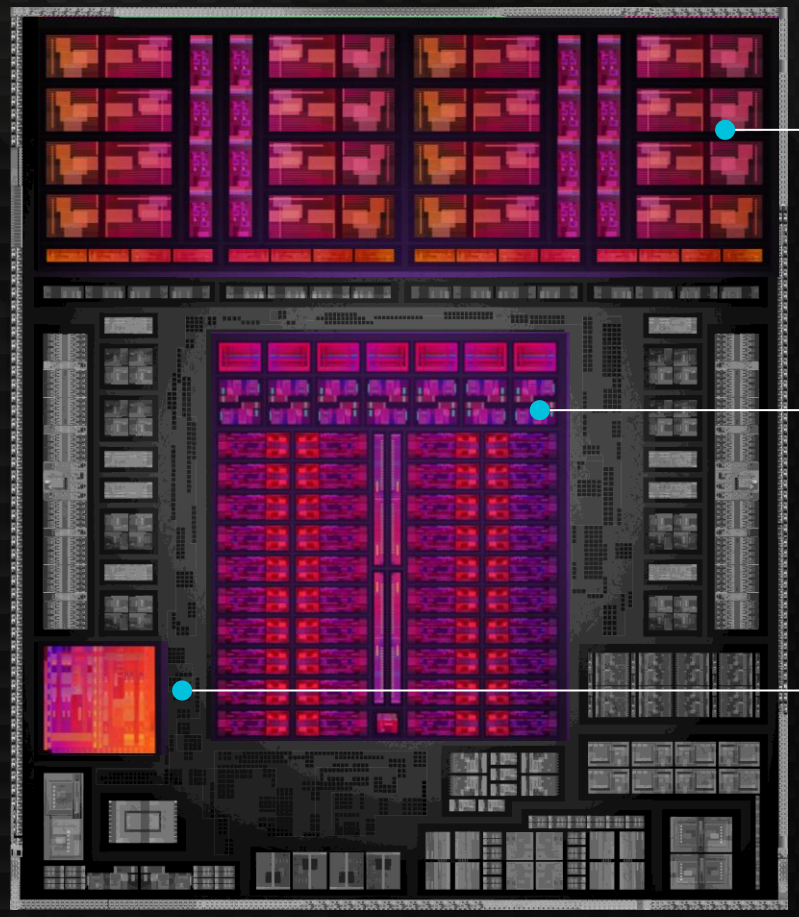
- Integrated GPU with discrete-level performance
- Up to 96GB flexibly configurable VRAM

Enhanced AI Experiences

- Dedicated Neural Engine (NPU) for Copilot+
- Leadership AI performance to enhance creativity and productivity

AMD Ryzen™ AI Max PRO Series Processors

Redefining Performance For Compact Workstations



ZEN 5

Up to 16 Core CPU
For cutting-edge single- and multi-threaded performance

AMD RDNA 3.5

Up to 40 CU GPU
For powerful integrated, ISV-certified graphics and GPU compute

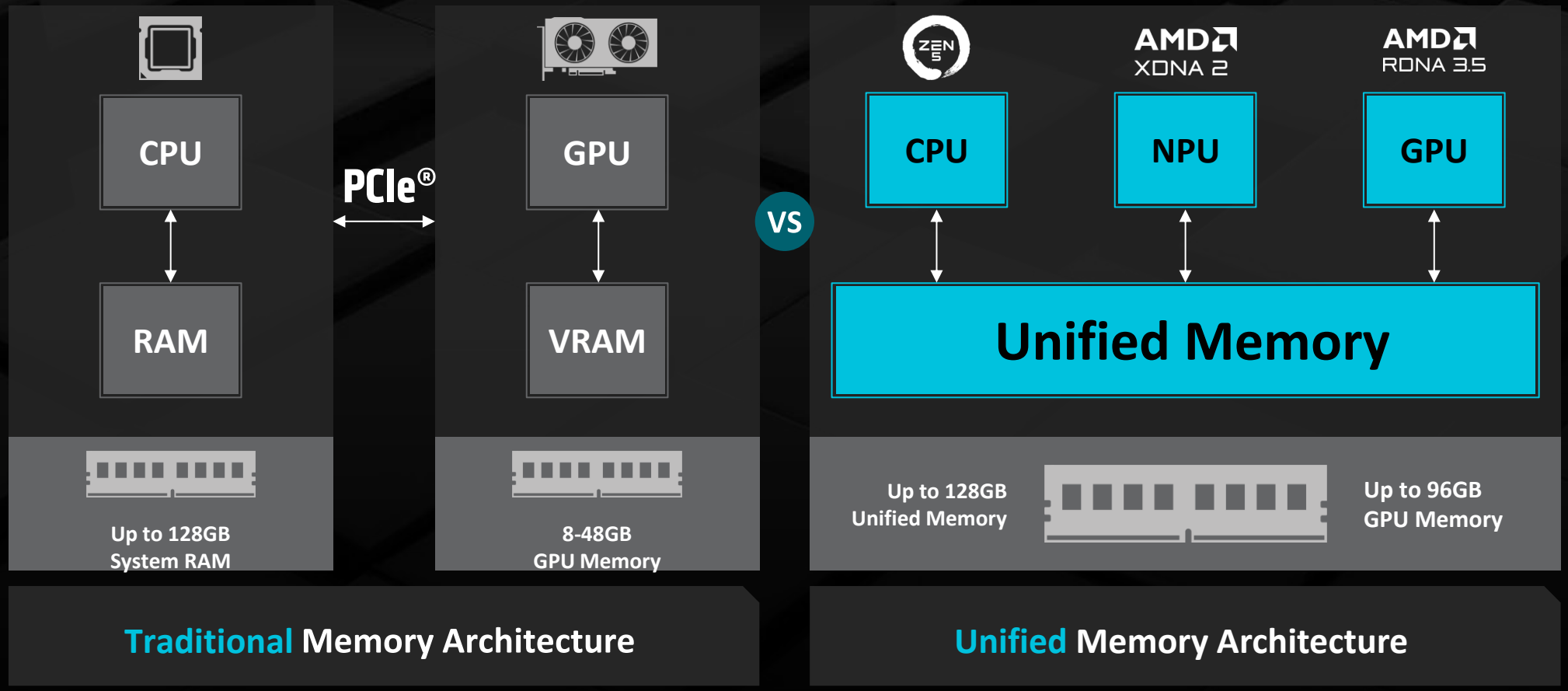
AMD XDNA 2

Up to 50+ TOPS NPU
Energy efficient AI engine for Copilot+PC capabilities*

* See endnote GD-243

AMD Ryzen™ AI Max PRO Series Processors

Unified Memory Architecture enables up to 96GB VRAM



- Work with massively large AI models locally
- Run multiple applications simultaneously
- Handle complex 3D data sets interactively

AMD Ryzen™ AI Max PRO Series Processors

Enabling New Software Development Experiences

Develop, integrate or use AI

Go beyond the capabilities of discrete GPUs and work locally with large language and diffusion models thanks to a unified memory architecture and up to 96GB VRAM

Get Software Coding Assistance Using LLMs

Improve your DevOps by using a fine-tuned version of Llama 70B Instruct (40GB+) or Code Llama for on-device coding support enabled by a fast GPU with dedicated AI accelerators

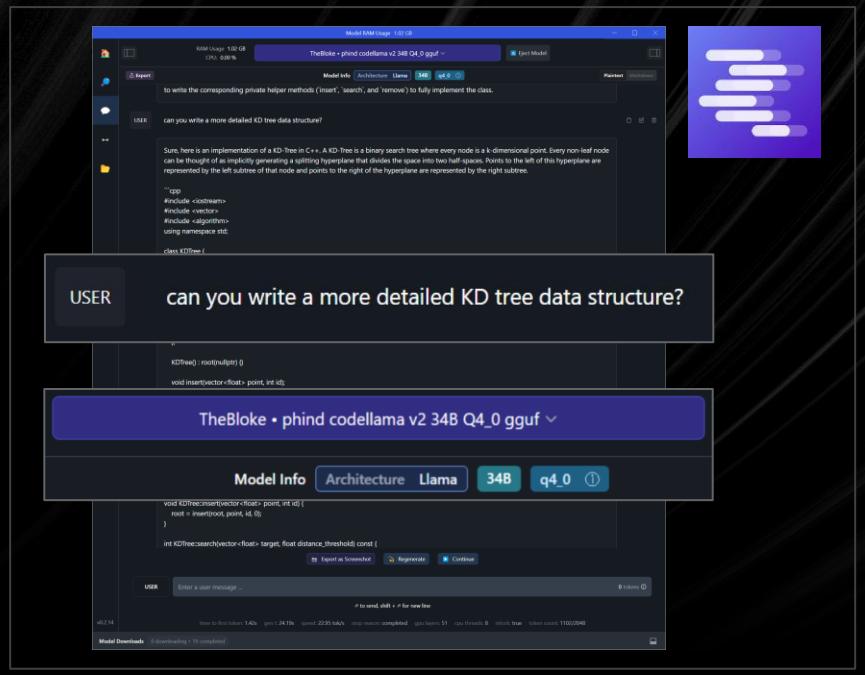
Accelerate Your Development

Up to 16 desktop-class “Zen 5” CPU cores deliver fast code compilation results, at your desk and on the road

PyTorch

TensorFlow

ONNX Runtime



AMD Ryzen™ AI Software Ecosystem

Hugging Face | PyTorch | Ollama | LM Studio | AMUSE | LLAMA.CPP

Models

Collaboration
Creative
Security

Open-Source | AMD | Customer

Optimization

Quantization
Pruning
Graph Compilation

Execution

ONNX RUNTIME

Application Integration on AMD Ryzen™ AI laptops

Broad Model Support

1000+ Models
CNN, Transformer

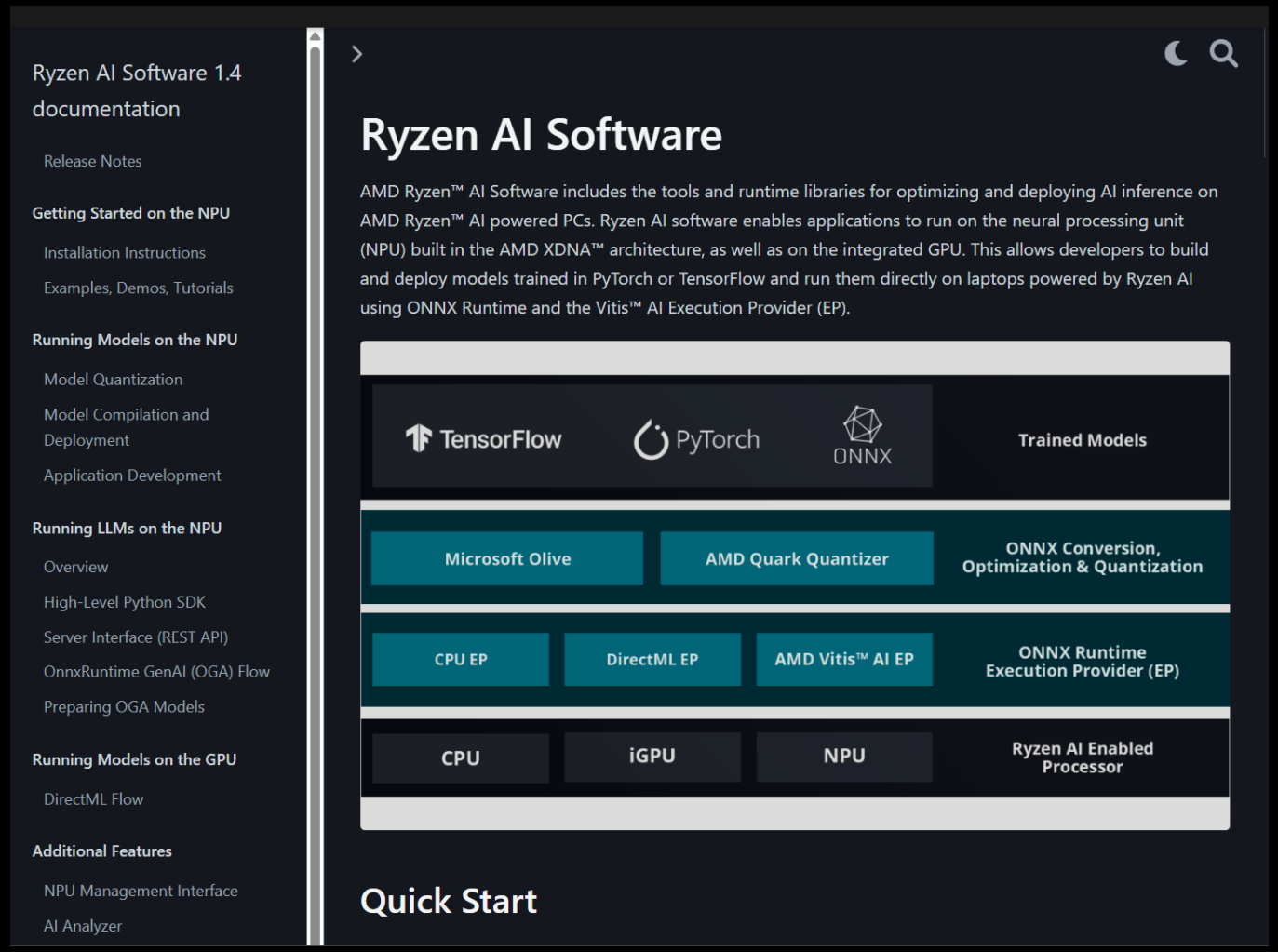
Diverse Data Types
INT4, INT8, BF16, Block FP16

Optimized Halo Models

LLMs
Llama, Mistral, Qwen

Text-to-Image GenAI
Stable Diffusion

Ryzen AI Docs/GitHub



Visit
https://ryzenai.docs.amd.com/

🍋 Lemonade SDK

Lemonade SDK is built on top of OnnxRuntime GenAI (OGA), an ONNX LLM inference engine developed by Microsoft to improve the LLM experience on AI PCs, especially those with accelerator hardware such as Neural Processing Units (NPUs).

The Lemonade SDK is comprised of the following:

- 🌐 **Lemonade Server:** A server interface that uses the standard Open AI API, allowing applications to integrate with local LLMs.
- 🐍 **Lemonade Python API:** Offers High-Level API for easy integration of Lemonade LLMs into Python applications and Low-Level API for custom experiments.
- 🖥️ **Lemonade CLI:** The lemonade CLI lets you mix-and-match LLMs, frameworks (PyTorch, ONNX, GGUF), and measurement tools to run experiments. The available tools are:
  - Prompting an LLM.
  - Measuring the accuracy of an LLM using a variety of tests.
  - Benchmarking an LLM to get the time-to-first-token and tokens per second.
  - Profiling the memory usage of an LLM.

https://github.com/lemonade-sdk/lemonade