



Agentic AI design on AMD Ryzen AI

Jens Stapelfeldt
Global AI Lead – AUP (AMD University Program)

www.linkedin.com/in/JensStapelfeldt

Agentic AI design on AMD Ryzen AI

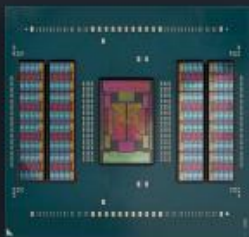
Overview

1. AMD University Program (AUP) – who we are
2. Agentic AI on Ryzen AI
3. Demo – Lemonade Server running local LLM
4. AMD open-source strategy
5. Q & A



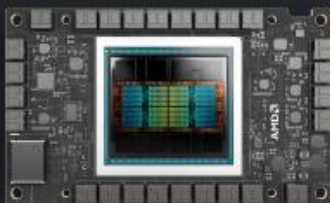
Best End-to-End AI Compute Portfolio in the Industry

AMD EPYC™ Processors



Leading server CPU

AMD Instinct™ Accelerators



World's best GPU accelerator

AMD Pensando™ Networking



Premier programmable
DPUs & AI NICs

AMD Ryzen™ AI AMD Radeon™ AI Processors



Most powerful client
AI processors

AMD Versal™ Adaptive SOCs




Leadership AI
Processing at the edge

AI Everywhere, for Everyone

A vertical panel showing a blurred background of green and red financial data, likely a stock market ticker, with numbers like 4093.21, 4093.09, 4093.08, and 4093.54 visible.

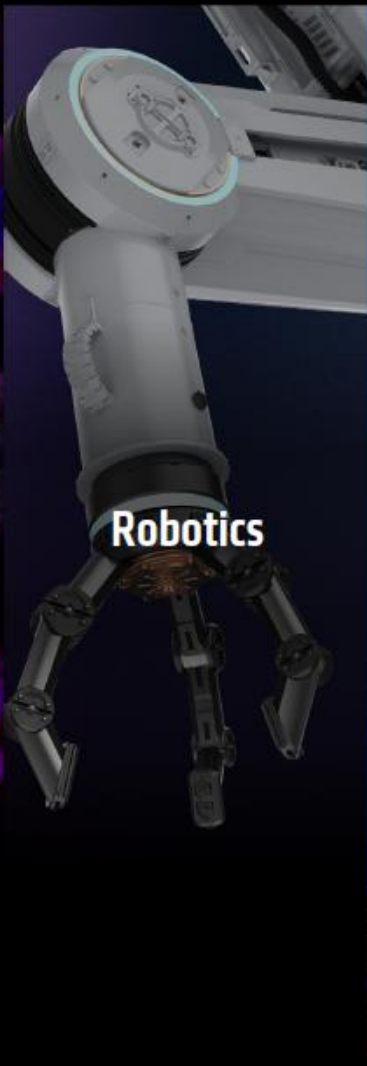
**Financial
Services**

A vertical panel showing a close-up of industrial machinery with blue and white components, possibly a robotic assembly line.


Manufacturing

A vertical panel showing a microscopic view of cells, with blue and purple hues, representing healthcare and life sciences.

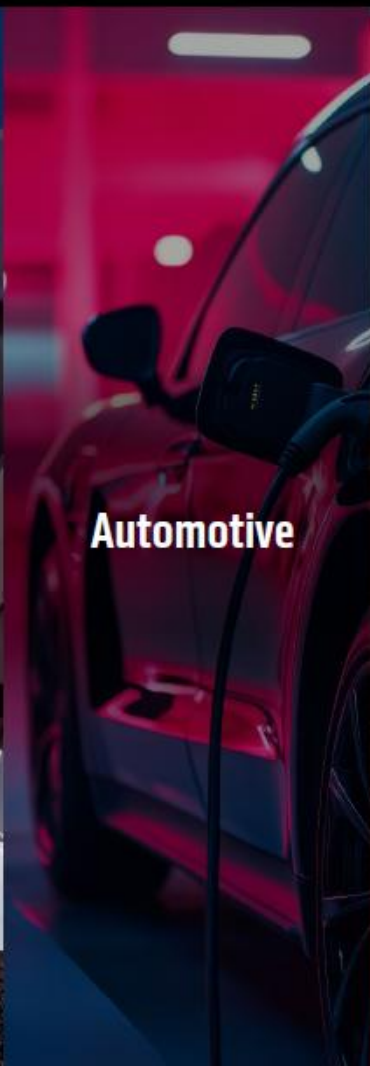
**Healthcare and
Life Sciences**

A vertical panel showing a white robotic arm with a blue light ring, set against a dark background, representing robotics.

Robotics

A vertical panel showing a large satellite dish antenna pointing towards the sky, representing communications.

Communications

A vertical panel showing a red car with a black charging cable plugged into its side, representing the automotive industry.

Automotive

A vertical panel showing a woman with curly hair and glasses, smiling while working on a laptop, representing productivity.

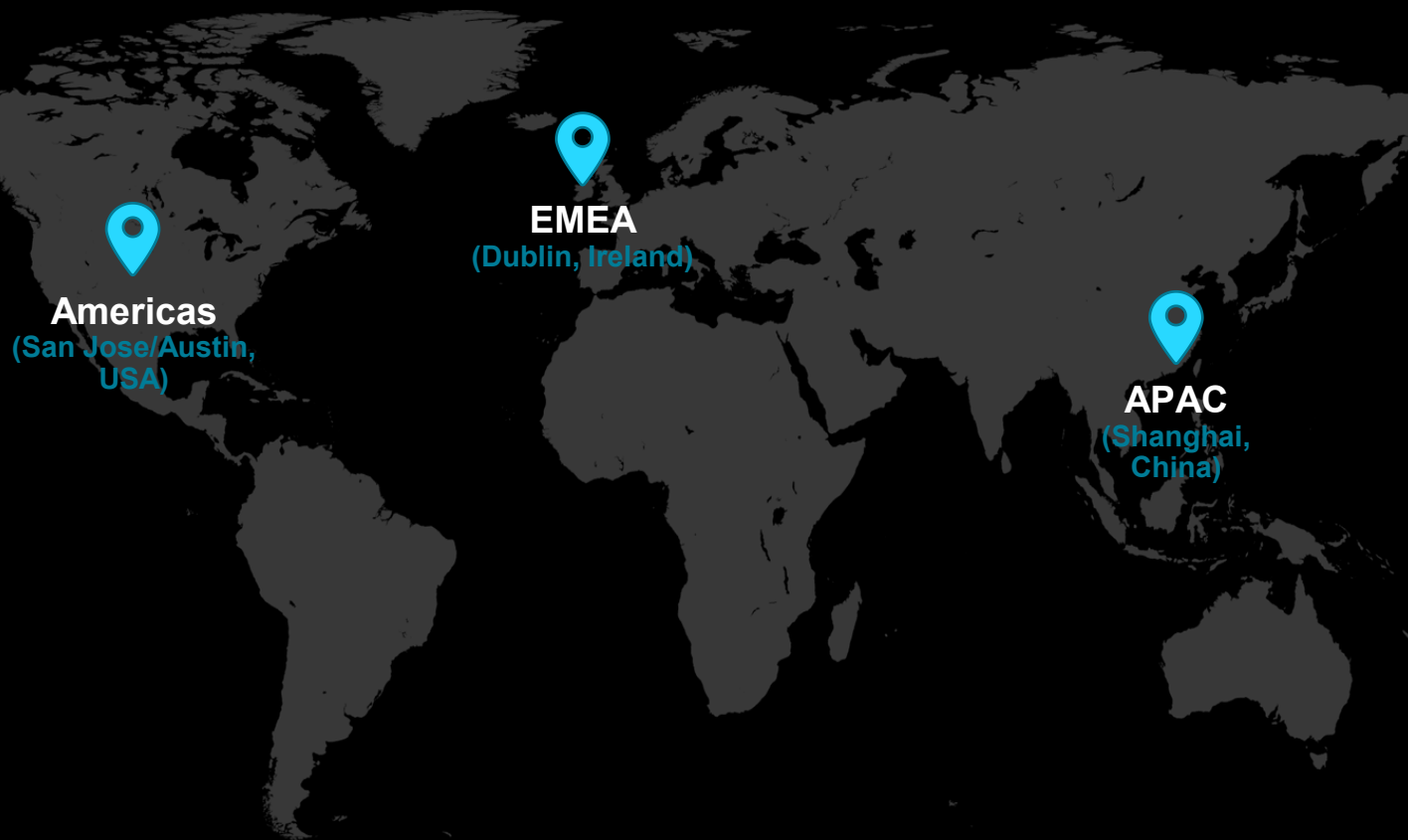
Productivity

AMD University Program AUP, part of CTO, RAD

Dedicated world-wide
technical team

Supporting High Performance
and Adaptive Compute in a open-
source Eco-System

25+ years experience
working with academia



AMD University Program (AUP) - What We do



**Low-cost
Academic
Hardware**



**Donation
Program**



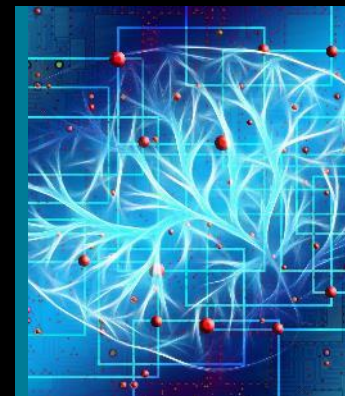
**Teaching
Resources**



Training



Support



**Research
Programs**

Reasoning & Agents Fuel Compute Surge



Leadership GPU
Lowers TCO



Leadership CPU
Powers apps



Openness
Accelerates Innovation

AI Beyond the Data Center

Inference Scaling Across Cloud to Edge to Client



Cloud



Edge



Client

Driven by domain-specific compute engines & open software stack

From ChatGPT to Agentic AI

ChatGPT (LLMs)

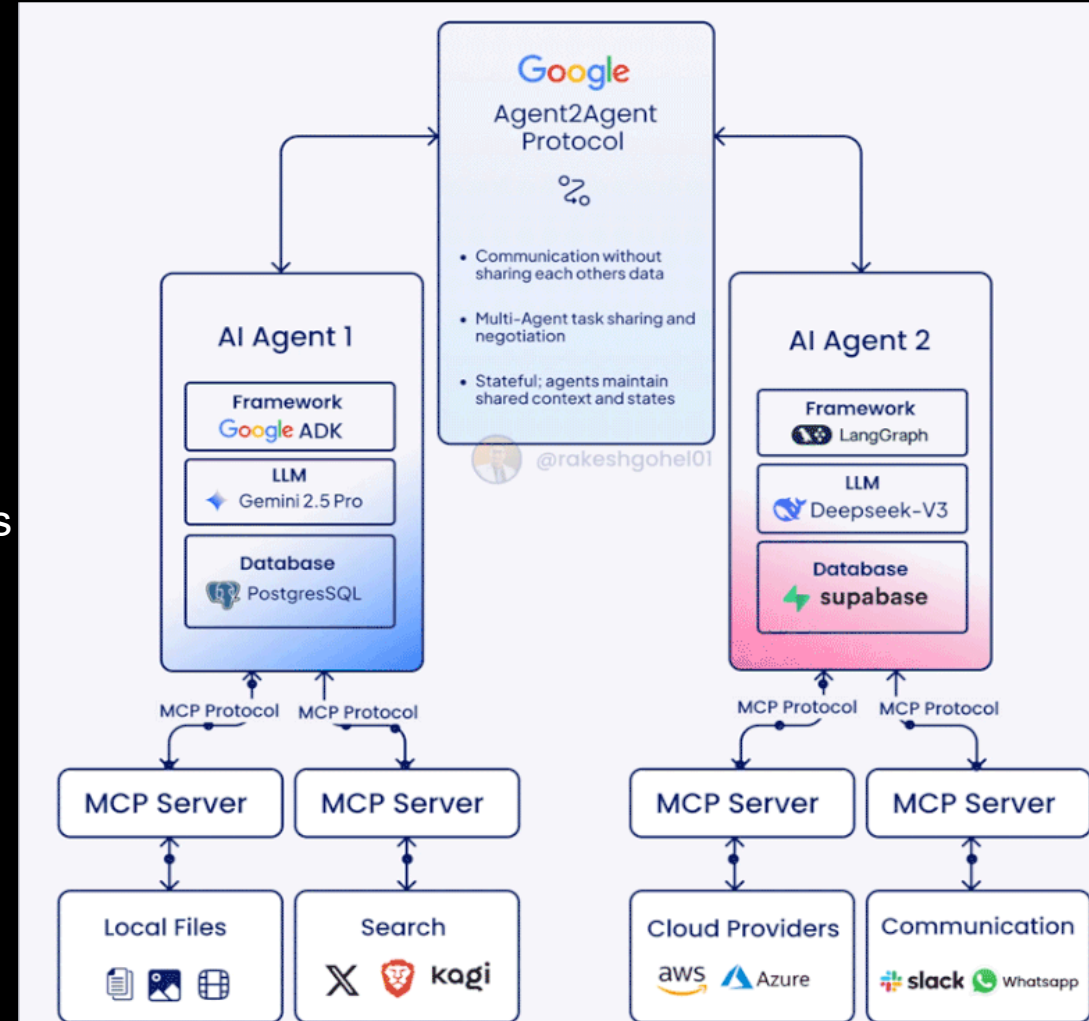
Language models learn from texts. They recognize patterns and generate responses (the likely next word!)

RAG-Systems

Retrieval Augmented Generation (RAG). Accessing one's own data

Agentic AI and AI Agents

AI systems act independently. They solve complex tasks without intervention





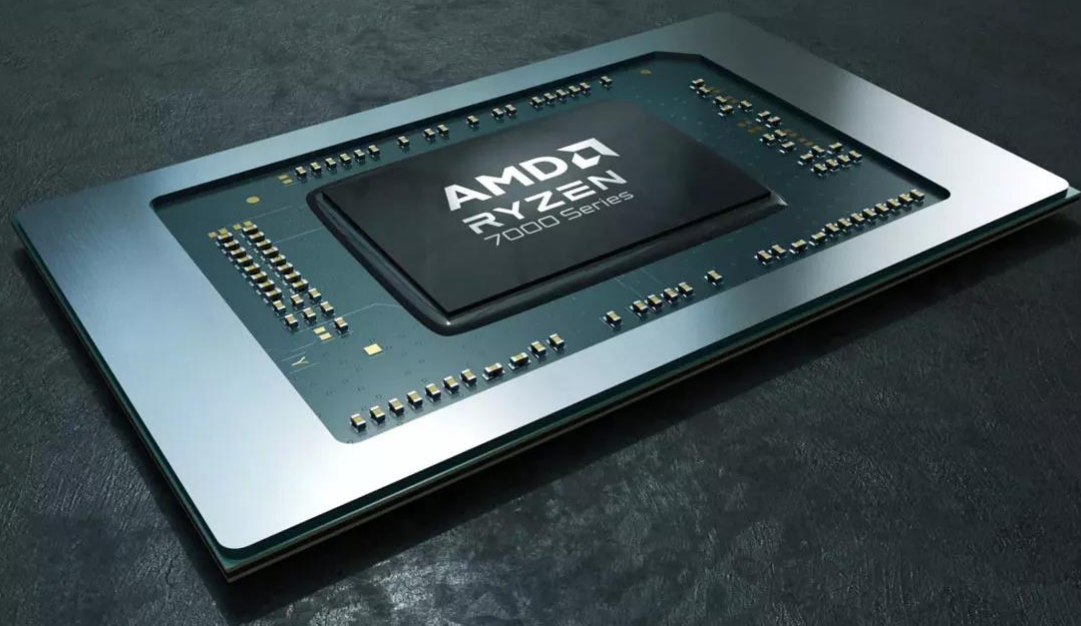
AI PC
**Drive the future of AI in
personal computing**

AMD 
RYZEN AI

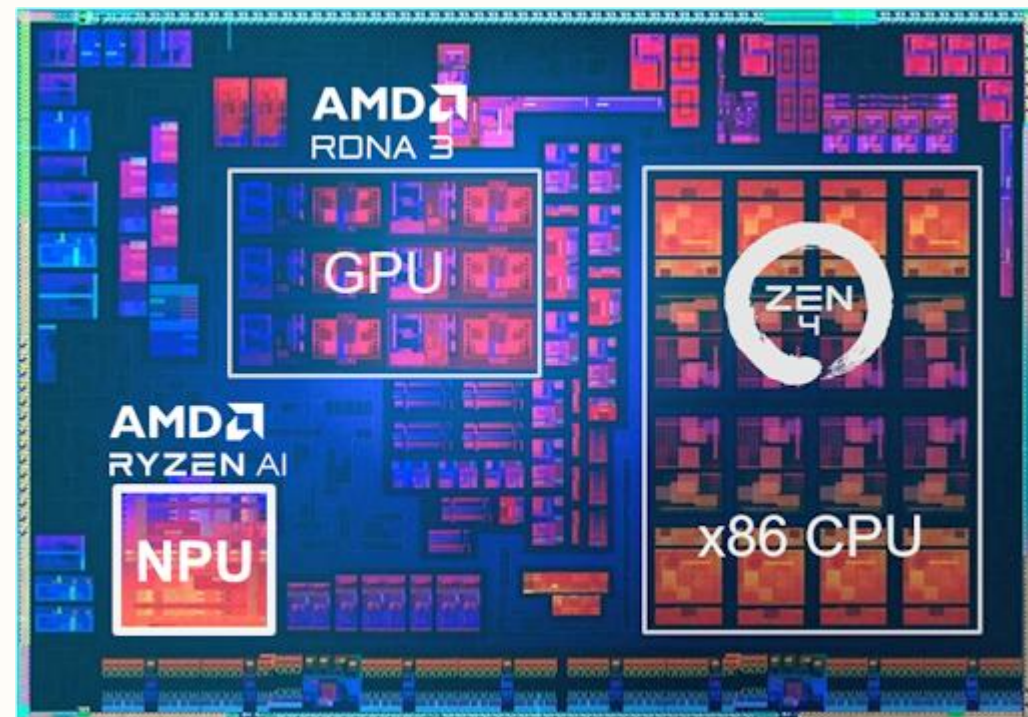
AMD RYZEN AI



Ryzen 7040 'Phoenix'



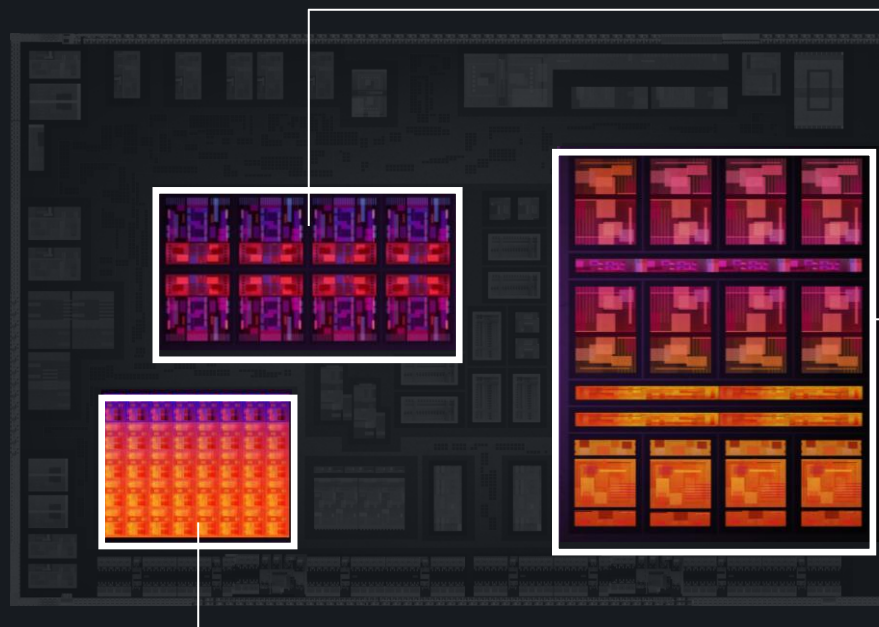
First x86 chips to integrate
CPU, GPU and *IPU*



Phoenix floorplan

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

3rd Generation
AMD Ryzen™ AI
Best in class AI platform



AMD
RDNA 3.5

Next-Gen GPU
Up to 16 Compute Units



Next-Gen CPU
Up to 12 Cores, 24 Threads

AMD
XDNA 2

Next-Gen NPU
Industry-leading 50+ NPU TOPS

AMD Ryzen™ 7040 Series - Introduction

"ZEN 4" Core

- High performance and efficient x86 cores
- Up to 13%* higher IPC

RDNA™ 3 Graphics

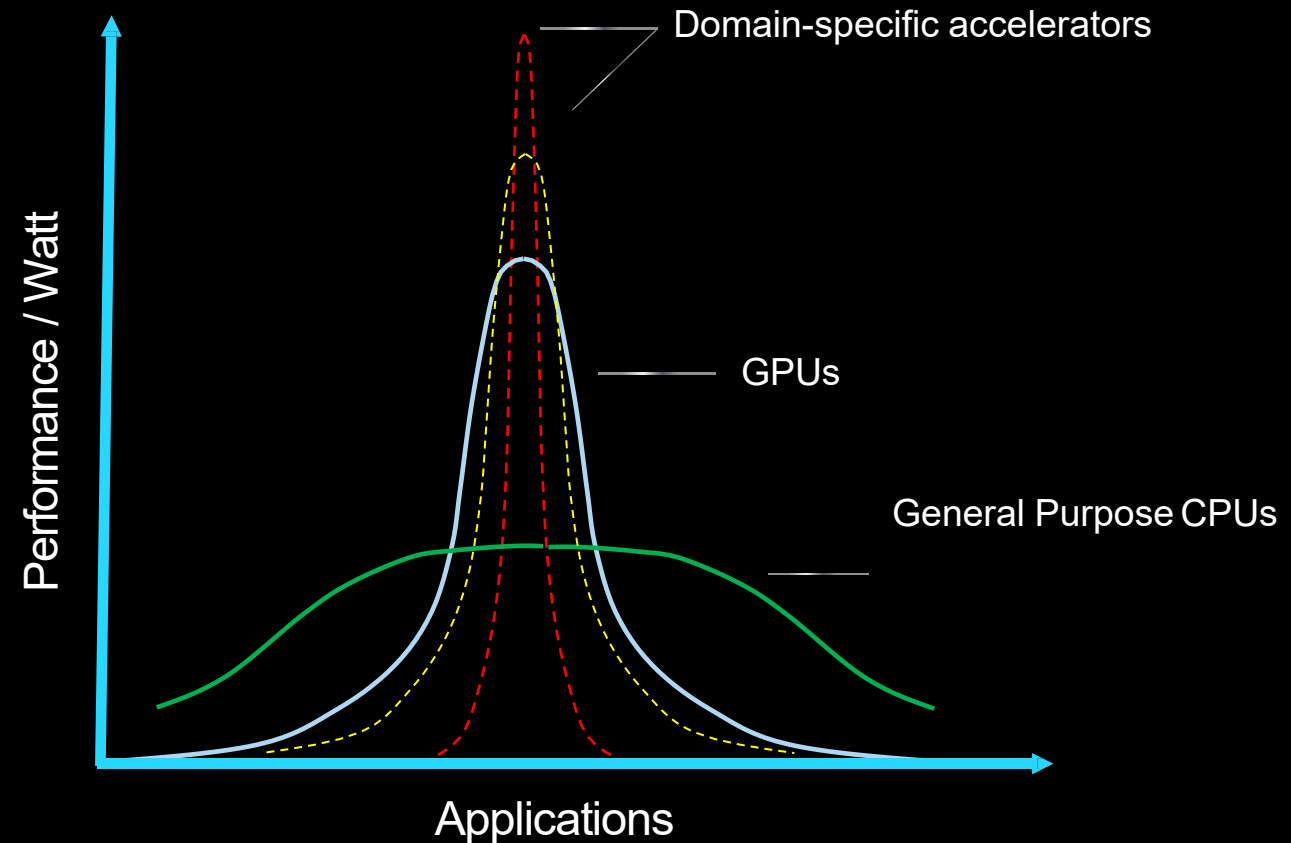
- Improved perf/W per compute unit

XDNA AI Engine

- First integrated AI engine on an x86 processor, powering AMD Ryzen™ AI

Technology

- Higher transistor density with TSMC N4P
- Accelerators include video, display, audio



Tailored compute for every client use-case

Dedicated AI Hardware Enables a New Era for PCs

Client



Productivity & experiences
Tremendous pressure on
battery life

Smart PCs



Explosion of Generative AI

**Bing > 100M users since AI
launch**

**Co-Pilot launching to
>400M users**



From Cloud to Hybrid

**Cloud inferencing is costly
for Gen-AI**

**ISVs need Client AI for Gen-
AI features**



Enhanced Experiences

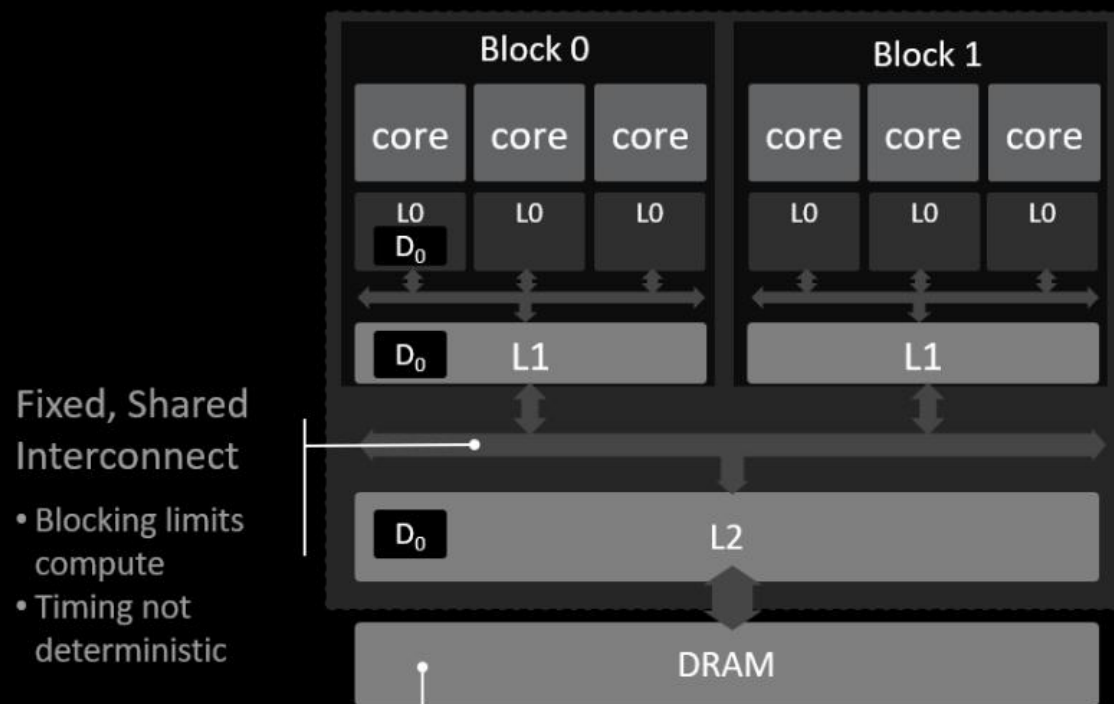
**Client = Personalized and
Private**

Reliable and Fast

TRADITIONAL MULTI-CORE VS. AMD XDNA™

Traditional Multi-Core Processor

(cache-based architecture)



Fixed, Shared Interconnect

- Blocking limits compute
- Timing not deterministic

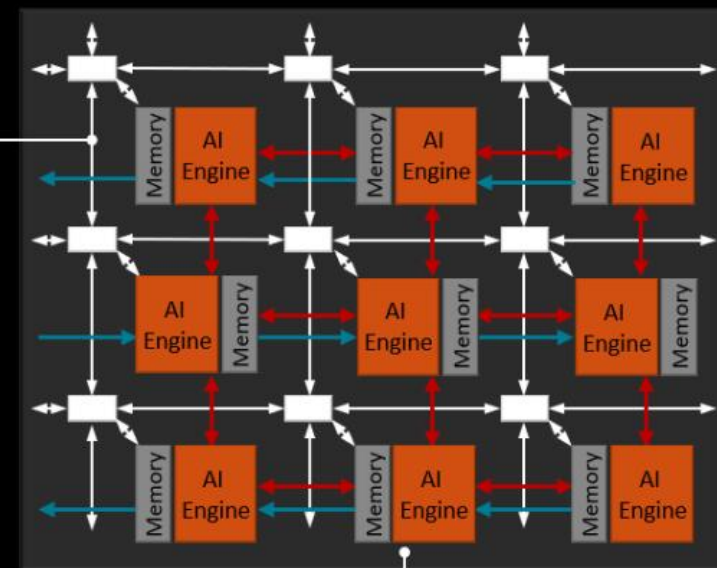
Data Replicated

- Robs bandwidth
- Reduces capacity

AMD XDNA™ AI Engine

(intelligent engine)

- Dedicated Interconnect
- Non-blocking
 - Deterministic

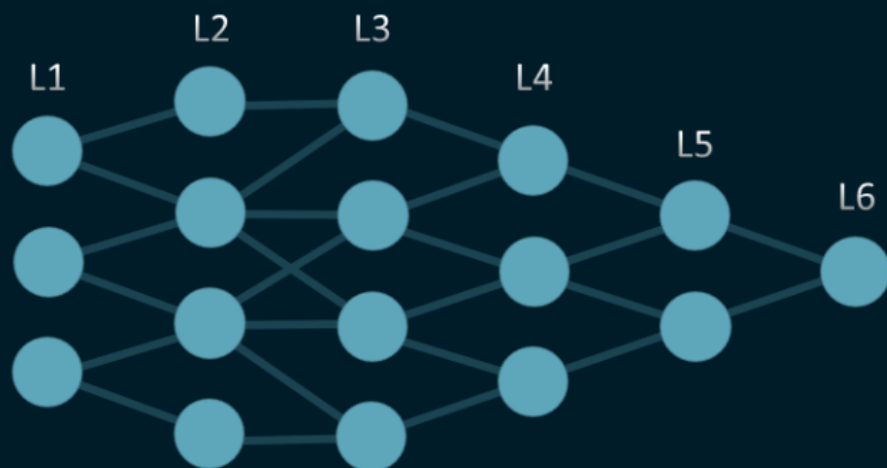


Local, Distributed Memory

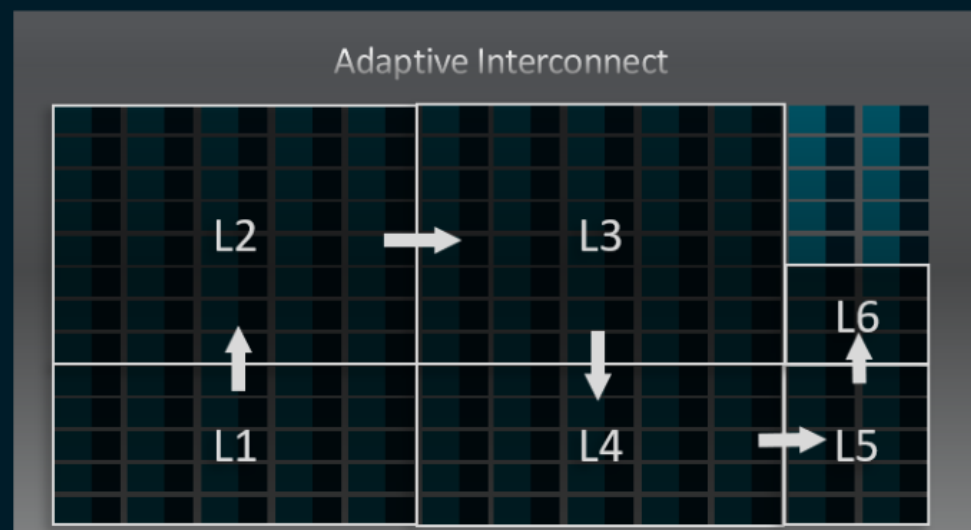
- No cache misses
- Higher bandwidth
- Less capacity required

HOW AMD RYZEN™ AI PROCESSES INFERENCE MODELS

A Neural Network
activates 'Neurons' from Layer to Layer



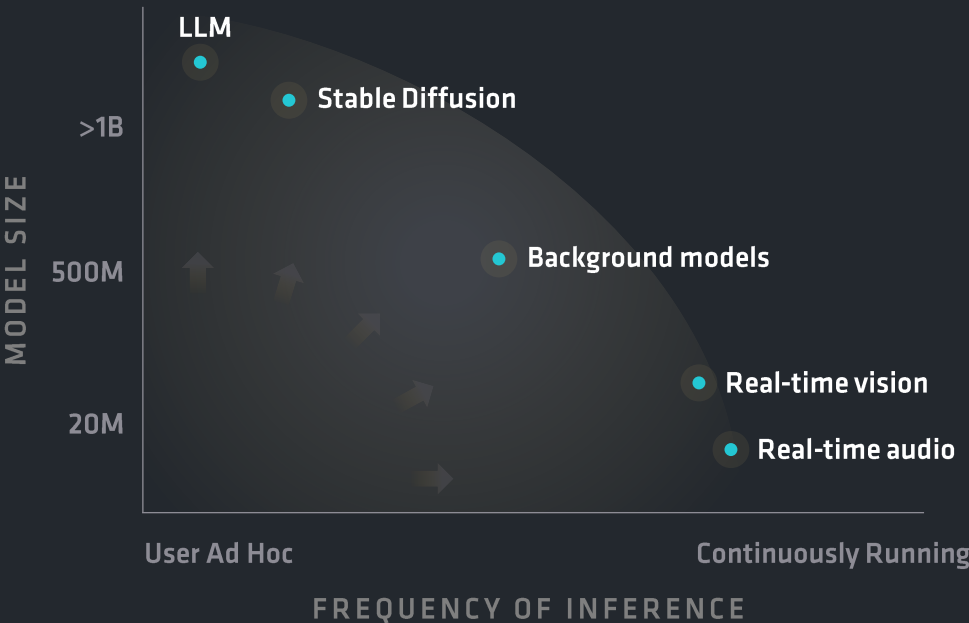
AMD XDNA™:
Adaptive Dataflow Architecture



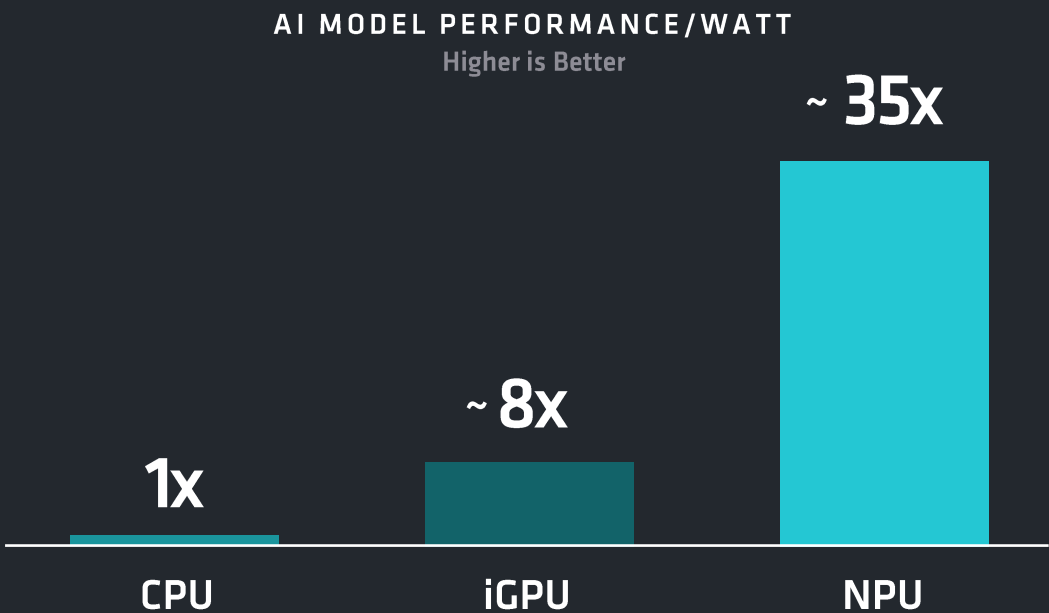
High-Performance, Energy Efficient, and Customizable for AI Workloads

Why NPUs?

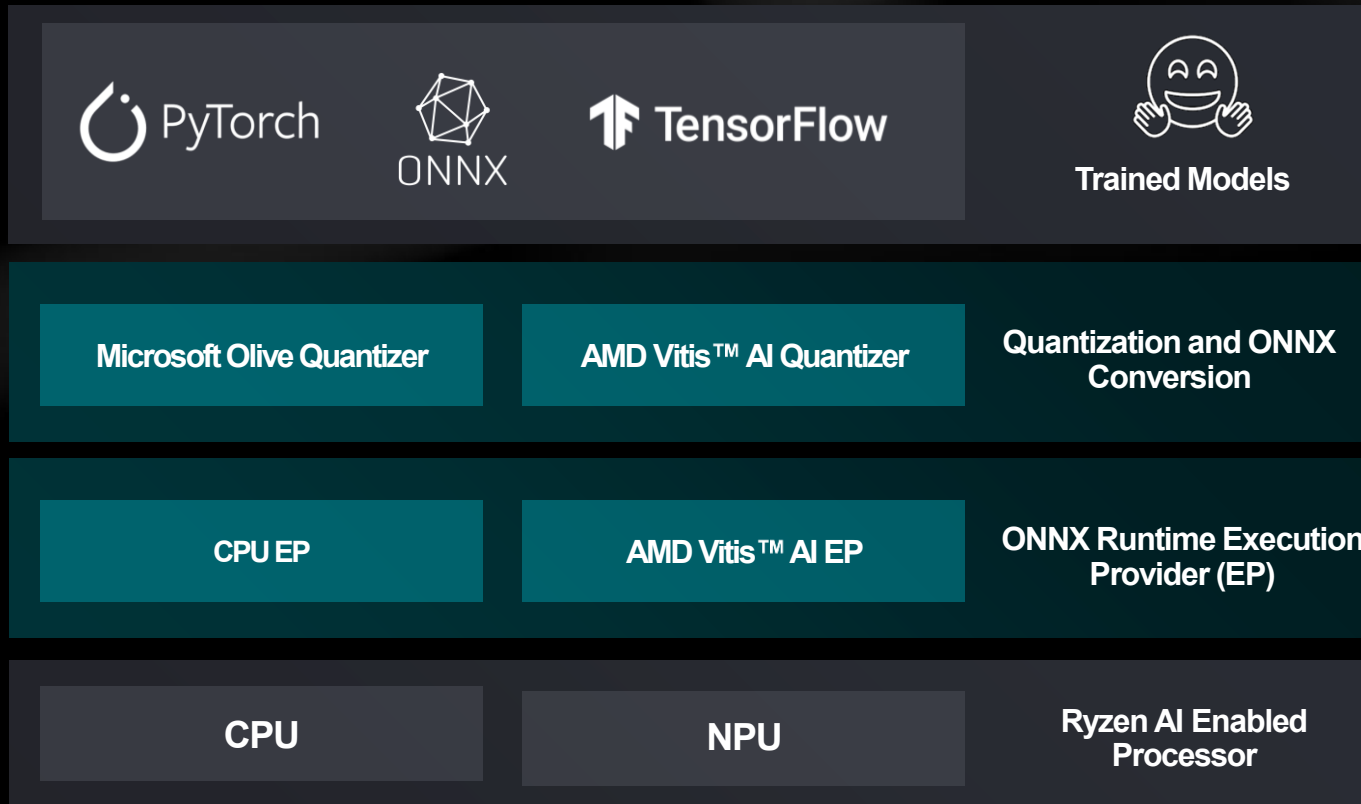
Model size and diversity growing and becoming increasingly integral to the OS



Enhanced AI efficiency matters more now than ever before

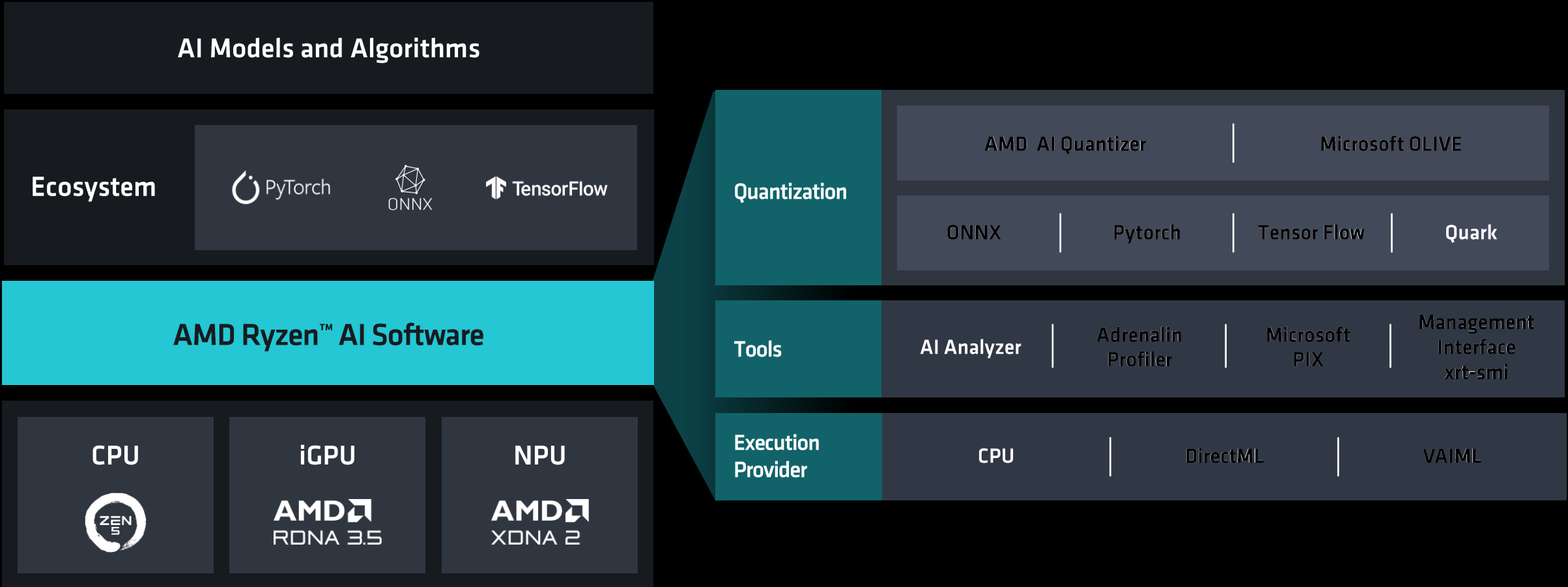


Ryzen™ AI Software Solution

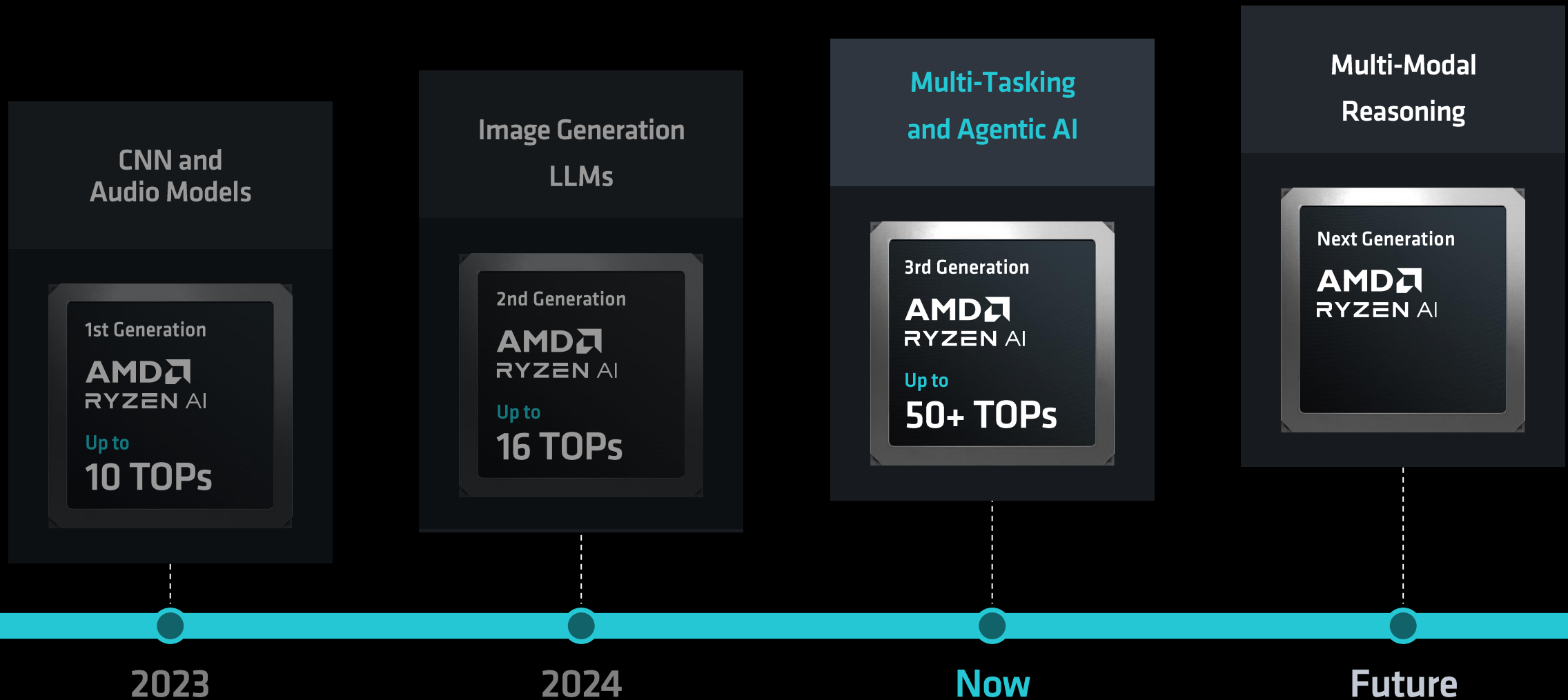


- ❑ Enable ONNX Runtime Applications on Ryzen AI Technology
- ❑ Build and deploy AI Apps locally
- ❑ Run private on-device LLM AI Workloads
- ❑ Quick Deployment with optimized pre-trained models
- ❑ Single-click installation in minutes

AMD Ryzen™ AI Software

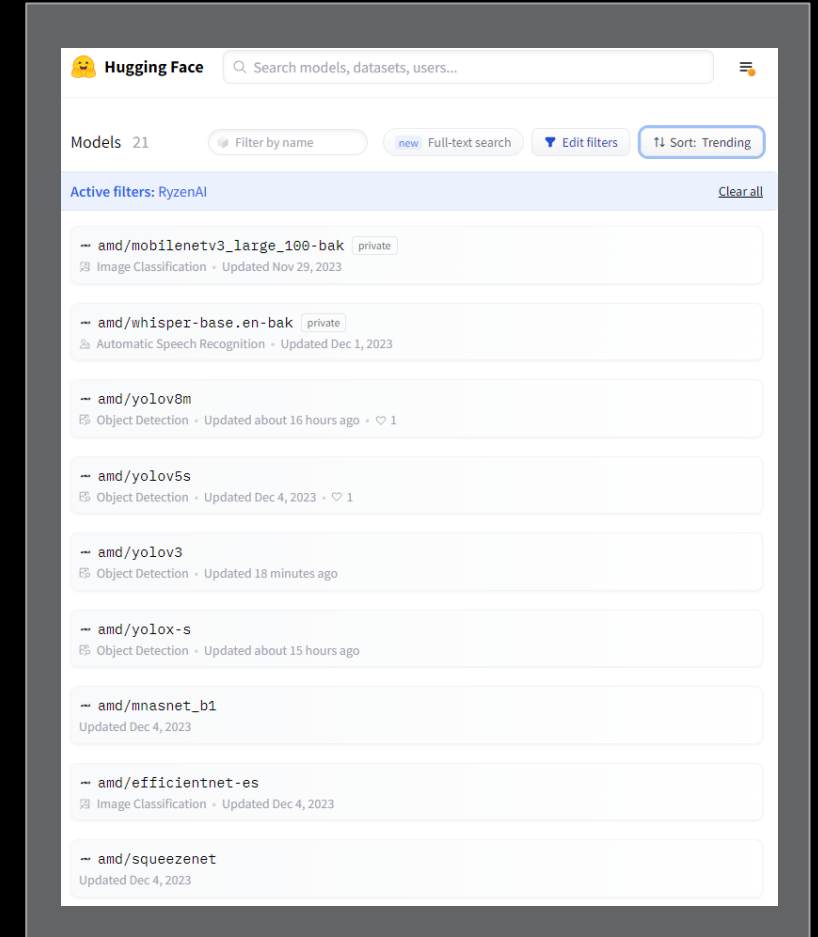


Ryzen AI Roadmap



Installation Steps to Run Models on Client

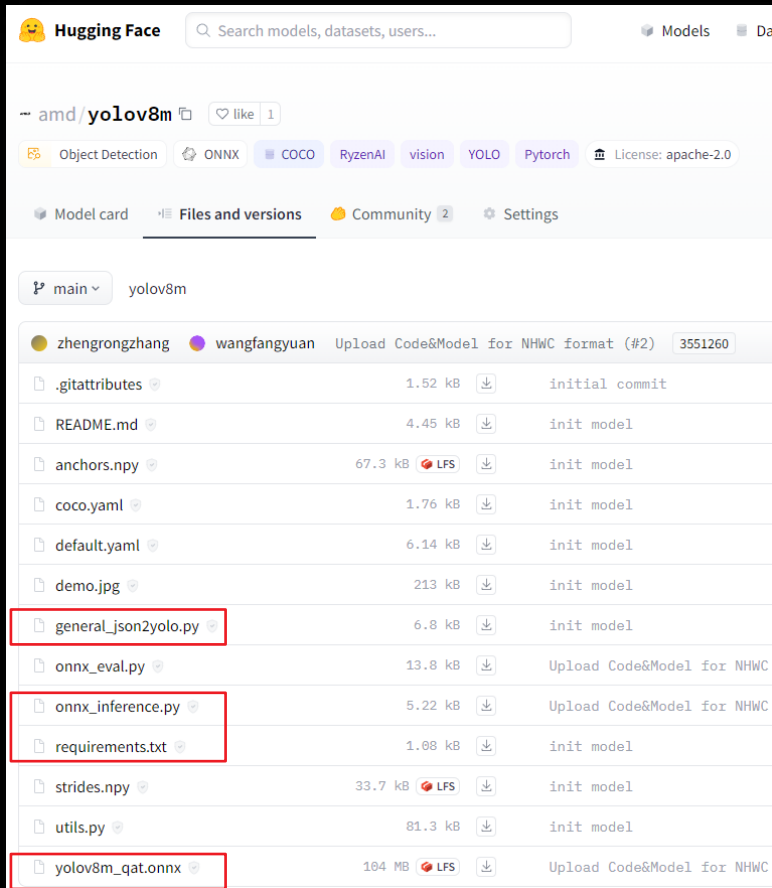
- Install Anaconda
- Install Visual Studio 2019 Community Edition
- Install cmake ≥ 3.26
- Install Python ≥ 3.9
- Install IPU Driver = 10.1109.8.110
- Install Git
- Create conda env with env.yaml
- Install ONNXRUNTIME
- Install VitisAI Engine Provider
- Get Models
 - GitHub repository
 - Model Zoo on Hugging Face
 - <https://huggingface.co/amd>



Pre-trained and Pre-quantized model zoo on Hugging Face

<https://huggingface.co/models?other=RyzenAI>

Get Started with Pre-Quantized Models on Hugging Face



Installing required packages

- ▲ `pip install -r requirements.txt`

Data Preparation (optional: for accuracy evaluation)

- ▲ `python general_json2yolo.py`

Run inference for a single image

- ▲ `python onnx_inference.py -i INPUT_IMG_PATH -o OUTPUT_IMG_PATH`
- ▲ `--ipu --provider_config vaip_config.json`

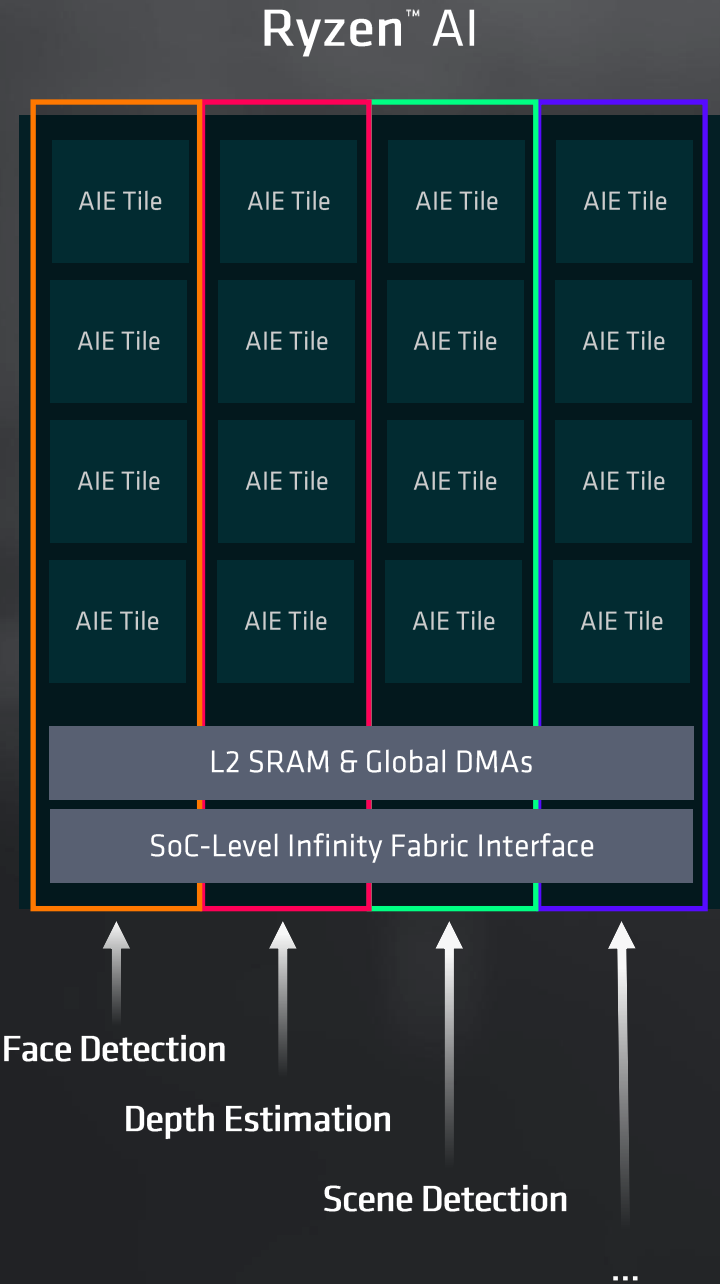
Test accuracy of the quantized model

- ▲ `python onnx_eval.py --ipu --provider_config vaip_config.json`

Running multiple apps with no performance degradation



Spatial partitioned AI multi-task acceleration



Scaling LLMs from cloud to endpoint

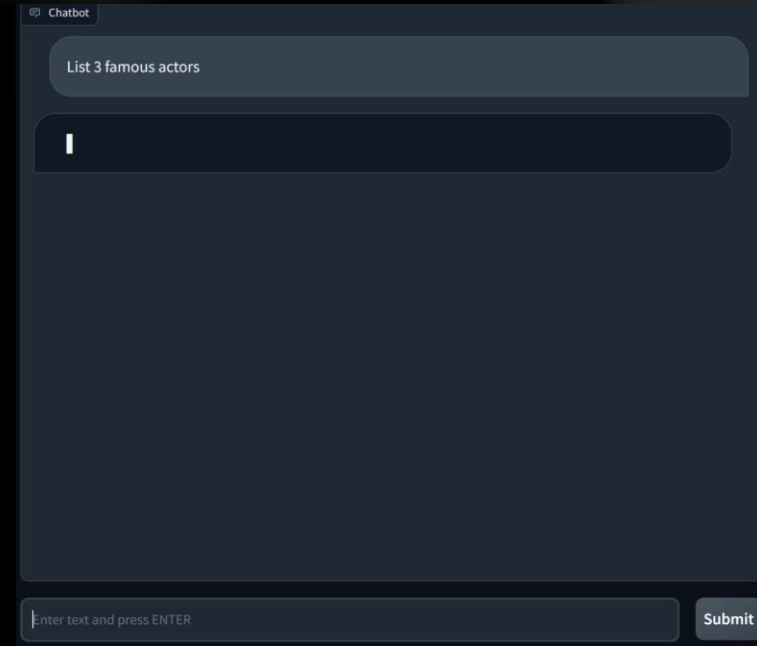
```
INFO: Loaded falcon-40b-instruct model onto MI300X  
Please enter your prompt: |
```

Up to 70B

parameter models* in a single GPU

AMD INSTINCT™ MI GPU

*FP16 models



1B - 7B

parameter models** in a single laptop

AMD RYZEN™

**INT8 models

“Hawk Point” AI PC processors now shipping

AMD Ryzen™ 8040 Series Processors



8 Core | 16 Threads

AMD
RDNA 3

Graphics

AMD
XDNA

16 TOPS

Up to **39 TOPS**

Total processor
performance

Ryzen AI Docs/GitHub

The screenshot shows the Ryzen AI Software 1.0.1 documentation website. The left sidebar contains a navigation menu with the following sections: 'Ryzen AI Software 1.0.1 documentation', 'Release Notes' (with a sub-link 'Release Information'), 'Getting Started' (with sub-links 'Installation Instructions', 'Runtime Setup', and 'Development Flow Overview'), 'Examples, Demos, Tutorials' (highlighted with a blue bar), 'Using Your Model' (with sub-links 'Model Compatibility', 'Model Quantization', and 'Model Deployment'), 'Ryzen AI Library' (with a sub-link 'Quick Start Guide'), and 'Additional Topics' (with sub-links 'Model Zoo', 'Manual Installation', 'Other Quantizers', and 'Early Access Features'). The main content area is titled 'Examples, Demos, Tutorials' and includes a sub-header 'Getting Started Tutorial' with a list of topics: 'The Getting Started Tutorial deploys a custom ResNet model demonstrating: Pretrained model conversion to ONNX', 'Quantization using Vitis AI ONNX quantizer', and 'Deployment using ONNX Runtime C++ and Python code'. Below this is an 'Examples' section with links to 'Run multiple concurrent AI applications with ONNXRuntime', 'Real-time object detection with YOLOv8', 'Run LLM OPT-1.3B model with ONNXRuntime', 'Run LLM OPT-1.3B model with PyTorch', 'Run Vision-Transformer model with ONNXRuntime', and 'Run ONNX end-to-end examples with custom pre/post-processing nodes running on IPU'. The 'Demos' section includes links to 'Cloud-to-Client demo on Ryzen AI' and 'Multiple model concurrency demo on Ryzen AI'. The 'Tutorials' section includes links to 'Hello World using a Jupyter Notebook', 'End-to-end Object Detection', and 'Quantization for Ryzen AI'.

Ryzen AI Software
1.0.1 documentation

Release Notes

Release Information

Getting Started

Installation Instructions

Runtime Setup

Development Flow Overview

Examples, Demos, Tutorials

Using Your Model

Model Compatibility

Model Quantization

Model Deployment

Ryzen AI Library

Quick Start Guide

Additional Topics

Model Zoo

Manual Installation

Other Quantizers

Early Access Features

Examples, Demos, Tutorials

This page introduces various demos, examples, and tutorials currently available with the Ryzen™ AI Software.

Getting Started Tutorial

- The [Getting Started Tutorial](#) deploys a custom ResNet model demonstrating:
 - Pretrained model conversion to ONNX
 - Quantization using Vitis AI ONNX quantizer
 - Deployment using ONNX Runtime C++ and Python code

Examples

- [Run multiple concurrent AI applications with ONNXRuntime](#)
- [Real-time object detection with YOLOv8](#)
- [Run LLM OPT-1.3B model with ONNXRuntime](#)
- [Run LLM OPT-1.3B model with PyTorch](#)
- [Run Vision-Transformer model with ONNXRuntime](#)
- [Run ONNX end-to-end examples with custom pre/post-processing nodes running on IPU](#)

Demos

- [Cloud-to-Client demo on Ryzen AI](#)
- [Multiple model concurrency demo on Ryzen AI](#)

Tutorials

- [Hello World using a Jupyter Notebook](#)
- [End-to-end Object Detection](#)
- [Quantization for Ryzen AI](#)

Visit <https://ryzenai.docs.amd.com/> to install, setup and run examples, demos and tutorials on your own

Introduction to Lemonade Server & Open WebUI



Run LLMs Locally

Use your PC to run large language models without cloud services.



No Coding Required

User-friendly setup with minimal technical knowledge needed.



Two Key Components

Lemonade Server and Open WebUI work together seamlessly.



What is Lemonade Server?

Open-Source SDK

Part of the ONNX project
TurnkeyML with comprehensive tools.

Serve & Benchmark LLMs

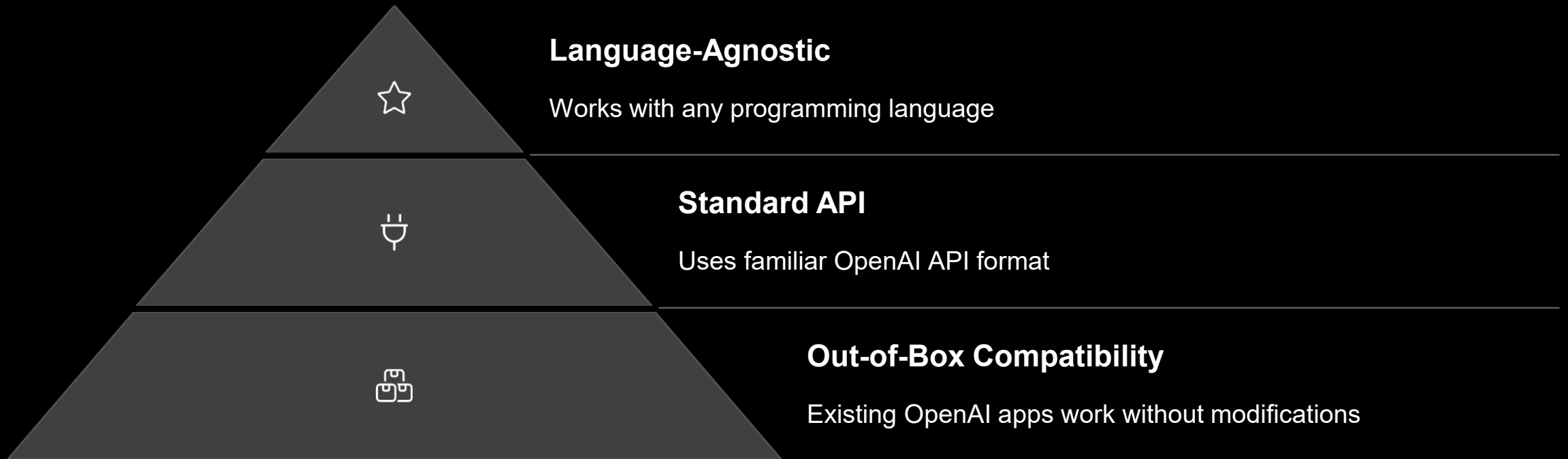
Run and test language models on various hardware.

Hardware Support

Compatible with CPUs, GPUs, and NPUs for flexible deployment.



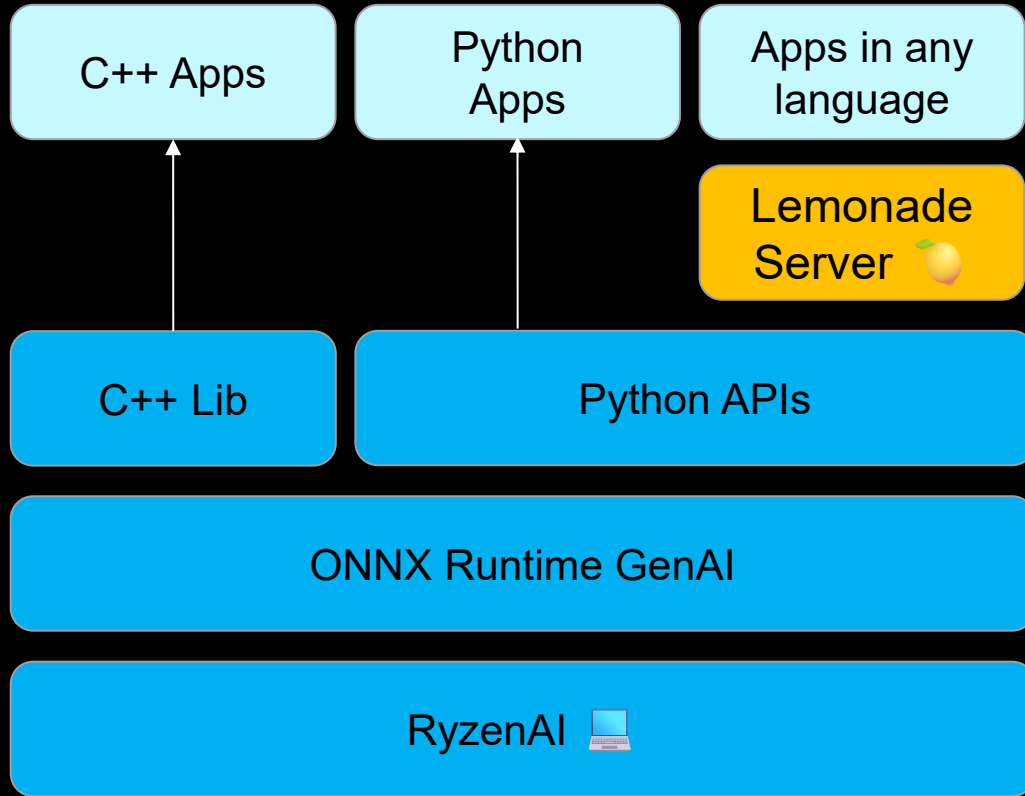
Key Feature: OpenAI API Compatibility



Applications can load LLMs on Ryzen AI and communicate via the standard OpenAI API. No prior knowledge of OGA or Ryzen AI needed.

Lemonade Server Software Stack Overview

SW Stack



Open Source!

github.com \ **lemonade-sdk** \ **lemonade**

OpenAI Compatible!

- Develop using OpenAI Standard

```
from openai import OpenAI

client = OpenAI(base_url=lemonade_url)

completion = client.chat.completions.create(
    model=my_model,
    messages=input_message
)
```

- Connect to dozens of apps without changing a single line of code 🧩



Microsoft
AI Toolkit



Continue.dev



LM Eval
Harness



Open WebUI



CodeGPT

Step 1: Install Lemonade Server



Download Installer

Navigate to TurnkeyML releases page and download the standalone installer.



Run Installer

Click "open file" to run the installer after download completes.



Select Models

Choose which LLM models to download during installation.



Launch Server

Click "Run Lemonade Server" or use the "Lemon desktop icon".



Lemonade Server

Refreshingly fast LLMs on the NPUs of Ryzen AI 300-series PCs.
Integrate with [Open WebUI](#), [AI Toolkit](#), or [your own app](#) in minutes.

Download
for Windows 11



Step 2: Install Open WebUI

Create Conda Environment

Follow the Open WebUI Quick Start Guide to create and activate a conda environment.

Install via Pip

Use pip to install Open WebUI in your activated environment.

Launch Application

Enter the command: open-webui serve in your terminal.

```

+ Portareito preidow

ate conda environment;
ironment ]
da environment (aU);
l data the thwirs fonuxy ;)

stallon
conda environment {
andins(2teraU/linmly,seliur-(coptis insallU)),,
onvtatin.tone adpicr/conda-,insnstalaton)
rep(installin tomxy.it{
topen is puhut ;;
shenu);
}

sil{
conda = if;
def l insalll;i{;
Open WebUI ;
environment's andiv/mewt installatioy,

chasted tank (install:
// the tngialangznazing cenv{lr,
aveiaion l}
};
};

```

Step 3: Configure Open WebUI for Lemonade Server



Create Admin Account

Navigate to Open WebUI URL and set up your admin credentials.



Access Settings

Click your profile in the right corner, then settings.



Add Connection

Click "connections" then the plus button to add a connection.



Configure Connection

Enter Lemonade Server URL and use a dash (-) as the API key.

Step 4: Run and Test the LLM

Select Model

Choose one of the hybrid models downloaded during installation.



Enter Prompt

Type a query like "Provide me code for a bubble sort in C".

Verify Response

Confirm the LLM generates appropriate content for your query.



Local Processing

Watch as the LLM runs on your Ryzen AI PC's NPU and iGPU.

References and Further Reading



Official Documentation

Comprehensive guides for [Lemonade Server](#) and [Open WebUI](#) with setup instructions, API reference, and troubleshooting tips.



GitHub Repositories

Explore source code, contribute to development, and track issues at [TurnkeyML/Lemonade](#) and [open-webui/open-webui](#).



Video Tutorials

Step-by-step visual guides for installation, configuration, and advanced use cases on the [AMD Developer YouTube channel](#).

Join the **AMD Developer Community** Discord server to connect with other users, share experiences, and get real-time assistance from the development team.

Demo – Lemonade Server

Introducing Lemonade Server: Local LLM Serving with GPU and NPU Acceleration

Further Resources and Support



Documentation

Detailed instructions in the Lemonade Server Read Me and Examples.



Integration

Try adding Lemonade Server into your own applications.



Support

Contact turnkeyml@amd.com for feedback or questions.

Lemonade Server: AMD's Lightweight Inference Runtime

- **What it is:**

- Lemonade Server is a lightweight inference server designed by AMD to simplify model deployment

- **Features:**

- Open-source and optimized for AMD GPUs via ROCm
- Low-latency, modular, compatible with ONNX and PyTorch models

- **Benefits for Agentic AI:**

- Scalable deployment of LLMs and tool-using agents
- Enables efficient AI inference for logistics, analytics, and real-time systems

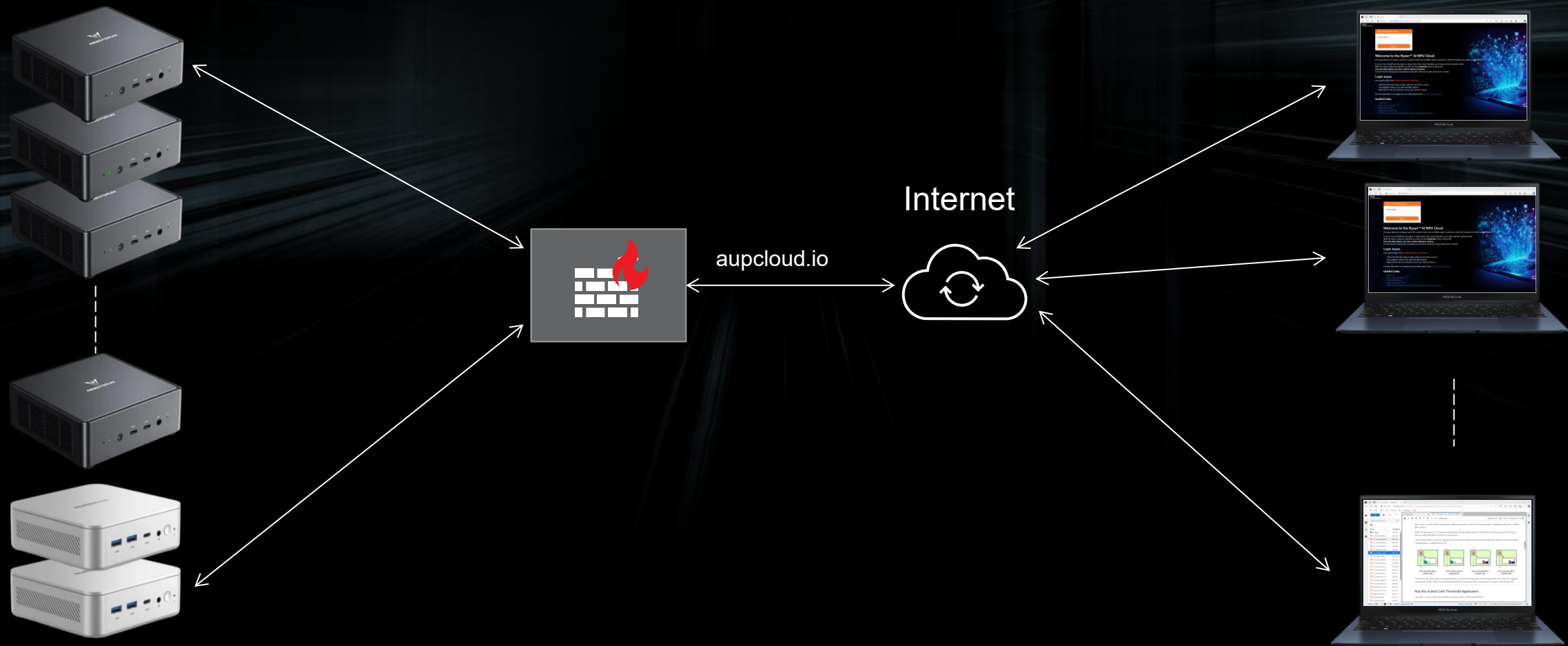
- [Lemonade GitHub](#): Source code and documentation



Run AI Models Locally
on Your PC with
Lemonade Server

Unlock powerful LLM apps without cloud
dependency or coding knowledge.

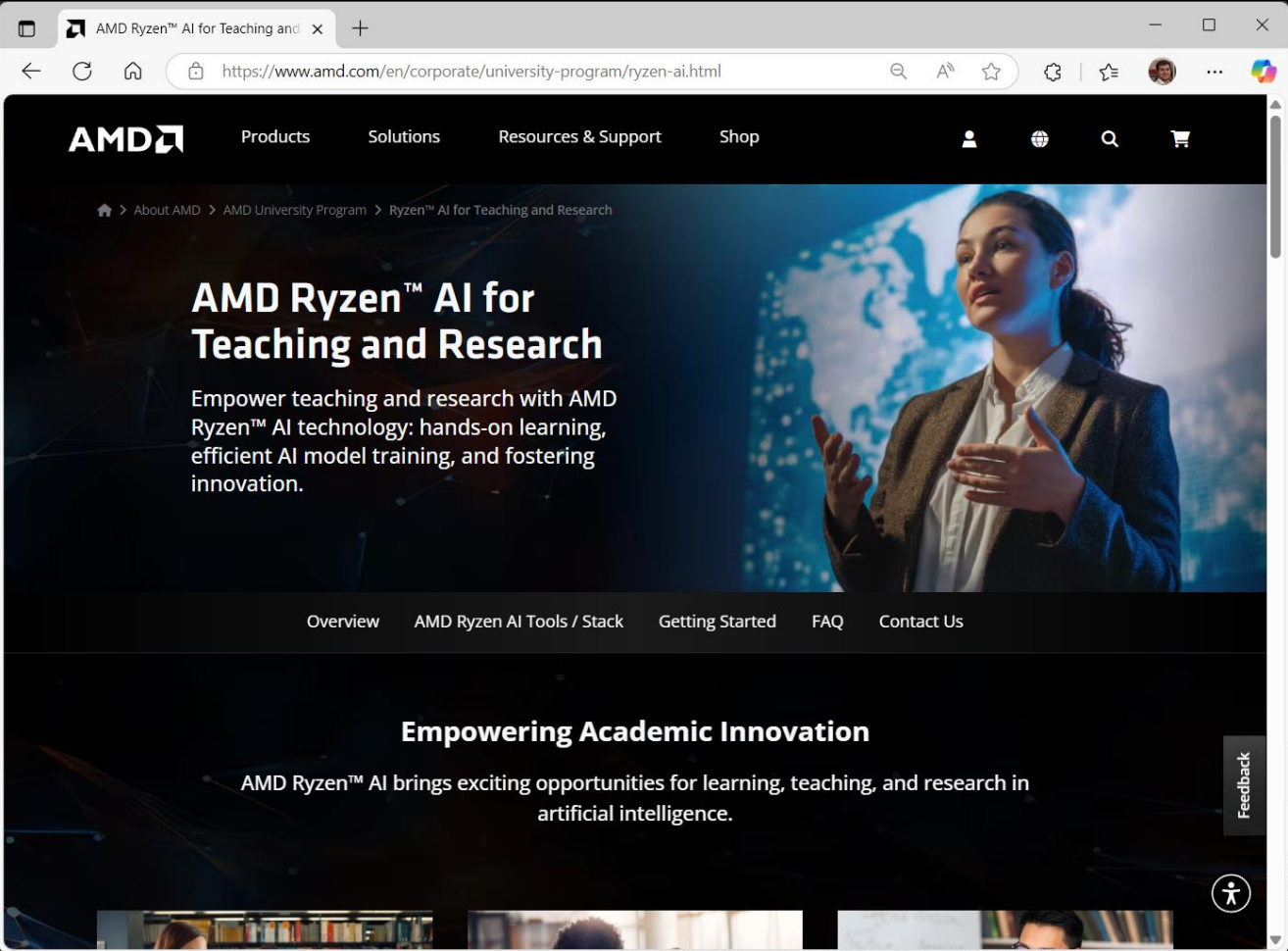
Enabling AI PC Remotely



Call to Action



Get started with Ryzen AI now



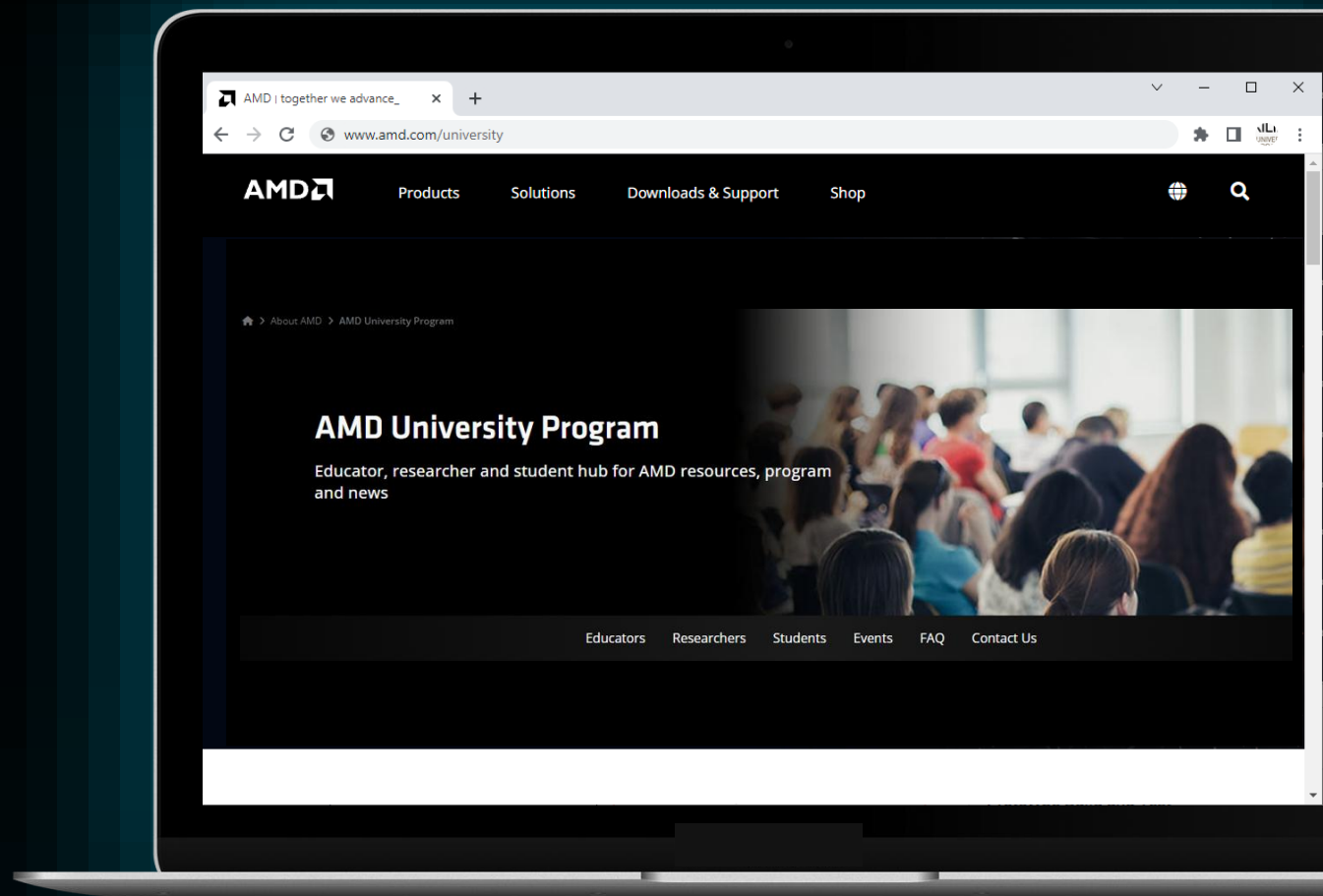
Contact Us

Visit our website to:

- Discover our research programs
- Access educational resources
- Submit a donation request
- Find training & other events

Email us:

aup@amd.com



www.amd.com/AUP

AMD Committed to Open-Source Innovation

Open Development Drives Value & Innovation

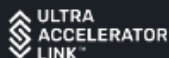
Open Hardware

+

Open Software

+

Open Ecosystem



Ultra Ethernet
Connectivity



Hugging Face



vLLM

SGL

Choice

Flexibility

Rapid Co-Innovation

Portability

Proven

Investing in Full-Stack Solutions

Acquisitions Span Entire AI Value Chain



Over 25 AI Acquisitions & Investments in the Last Year Alone



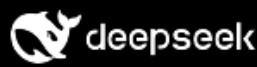
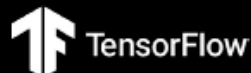
Rapidly Advancing Open Software Capabilities

Day-0 Support
for Leading Models

Accelerated
Pace of Innovation

Broadening
Ecosystem Partnerships

Developer First
Approach to Enablement





Committed to Open-Source Innovation



Hugging Face

1000,000+ models run
out-of-box on AMD
ROCm™ platform



OpenAI Triton

Fully upstreamed AMD ROCm
platform support
Used for key LLM kernel generation



PyTorch

Fully upstreamed AMD ROCm
platform support
Continuous Integration



JAX

vLLM vLLM



Tensor Flow



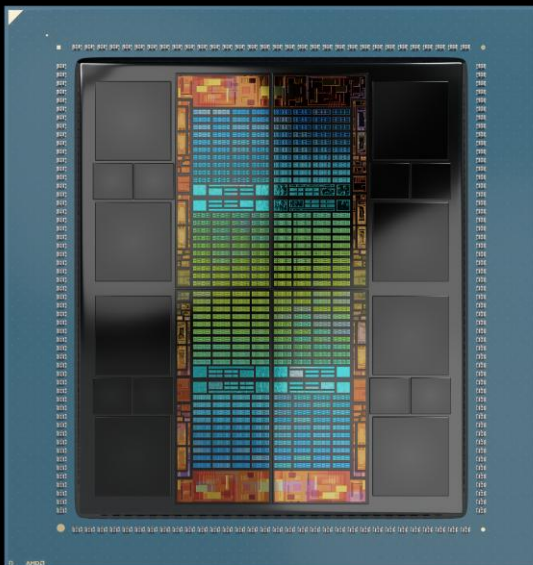
MLIR | IREE



ONXX Runtime



OpenXLA



AMD Instinct™ MI300X Accelerator

Leadership performance

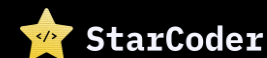
Out-of-box support on popular GenAI models



GPT-4

Llama 2 

Llama 3 



Phi

Falcon LLM



Qwen

MPT



Stable Diffusion

GPT-NeoX

D B R X

AlphaFold 2

OPT



Gemma



Final Thoughts

Agentic AI is not just a technological frontier — it is a collaborative one.
Get in contact with AMD AUP team AUP@amd.com



Students

Provides hands-on experience with AI applications right on their laptops—whether for machine learning projects, data science coursework, or real-time AI-enhanced applications.



Educators

Integrate AMD Ryzen™ AI into their curriculum, enabling interactive lessons on neural networks, automation, and AI-driven creativity.

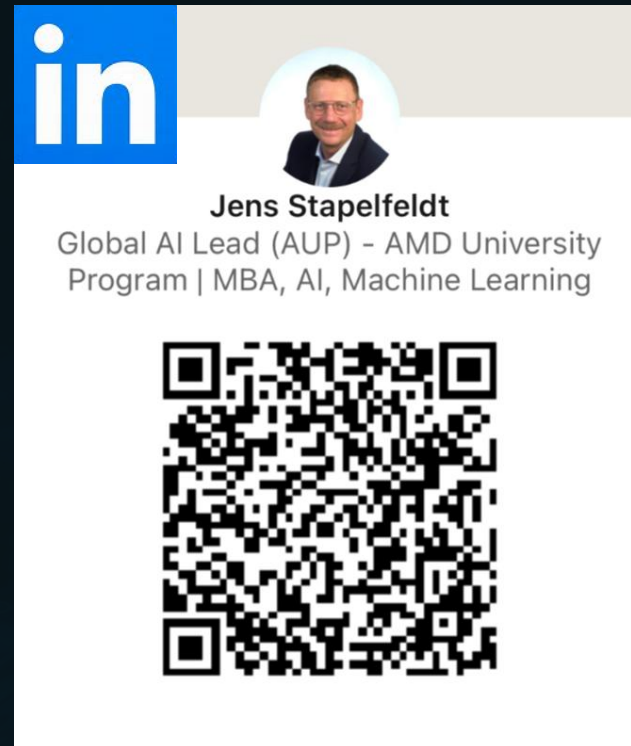


Researchers

Leverage the combined power of the CPU, iGPU, and NPU to experiment, build and optimize local AI models

Q&A

Thank you for your attention! Questions or discussion?





together we advance_

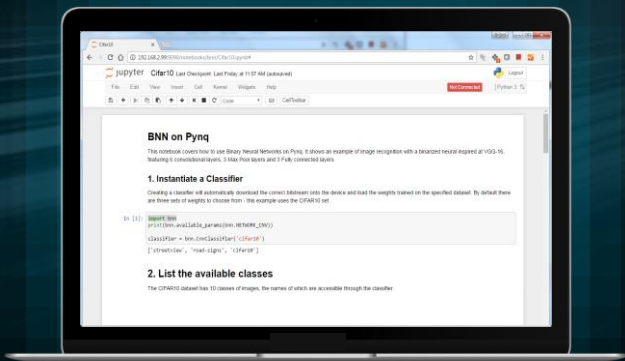
DISCLAIMER

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

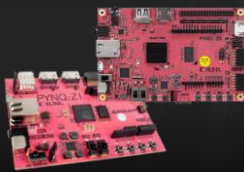
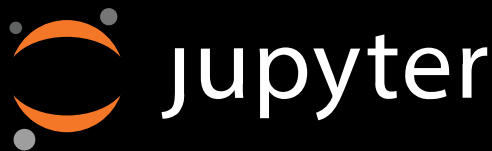
© 2023 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD Instinct, AMD XDNA, Alveo, Artix, EPYC, Kintex, KRIA, Radeon, Ryzen, Spartan, Versal, Virtex, Vitis, Vivado, Zynq, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners.

Enabling Python & Jupyter on Adaptive Compute Platforms

High-level Python
APIs and libraries



AMD
PYNQ



AMD
ZYNQ



AMD
ZYNQ
MPSoC



AMD
ZYNQ
RFSoc

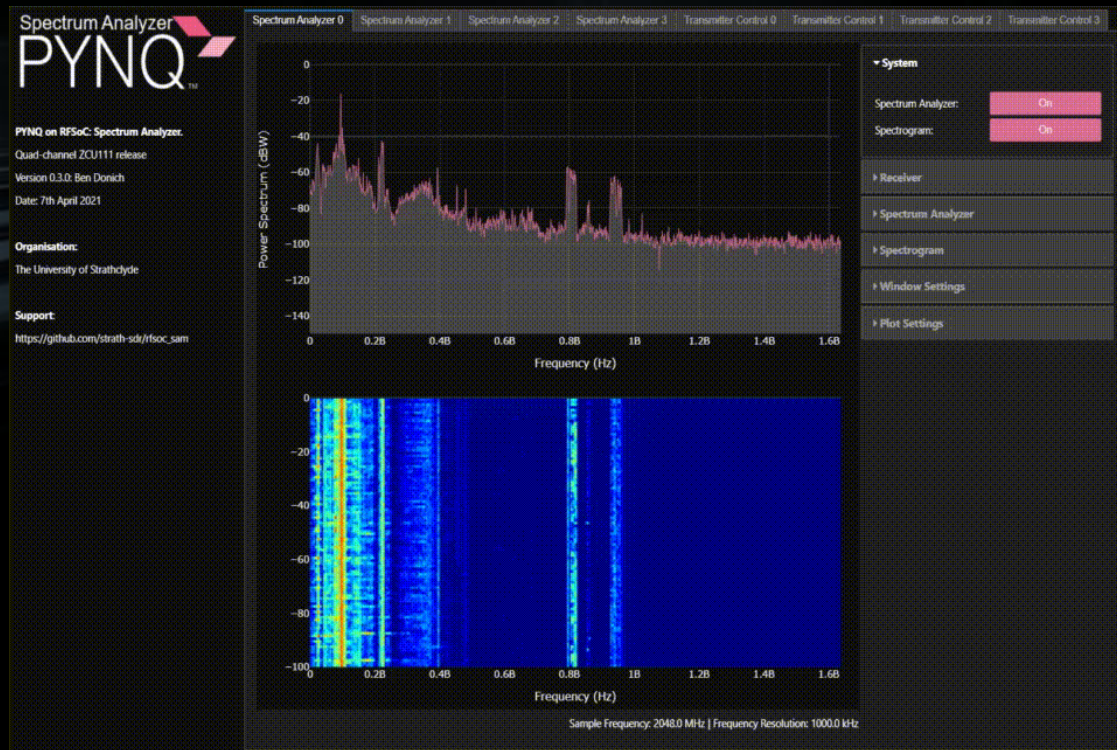


AMD
KRIA



AMD
ALVEO

Open source examples



Open Source PYNQ Spectrum Analyzer
https://github.com/strath-sdr/rfsoc_sam



Open Source RFSoc OFDM Transceiver
https://github.com/strath-sdr/rfsoc_ofdm