Agentic AI design on AMD Ryzen AI

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

www.linkedin.com/in/JensStapelfeldt

AMD
together we advance_

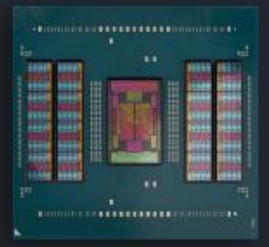
Agentic AI design on AMD Ryzen AI

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

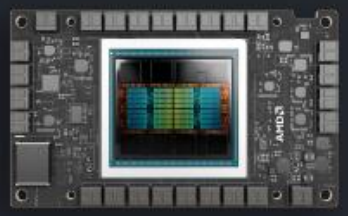
Best End-to-End AI Compute Portfolio in the Industry

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

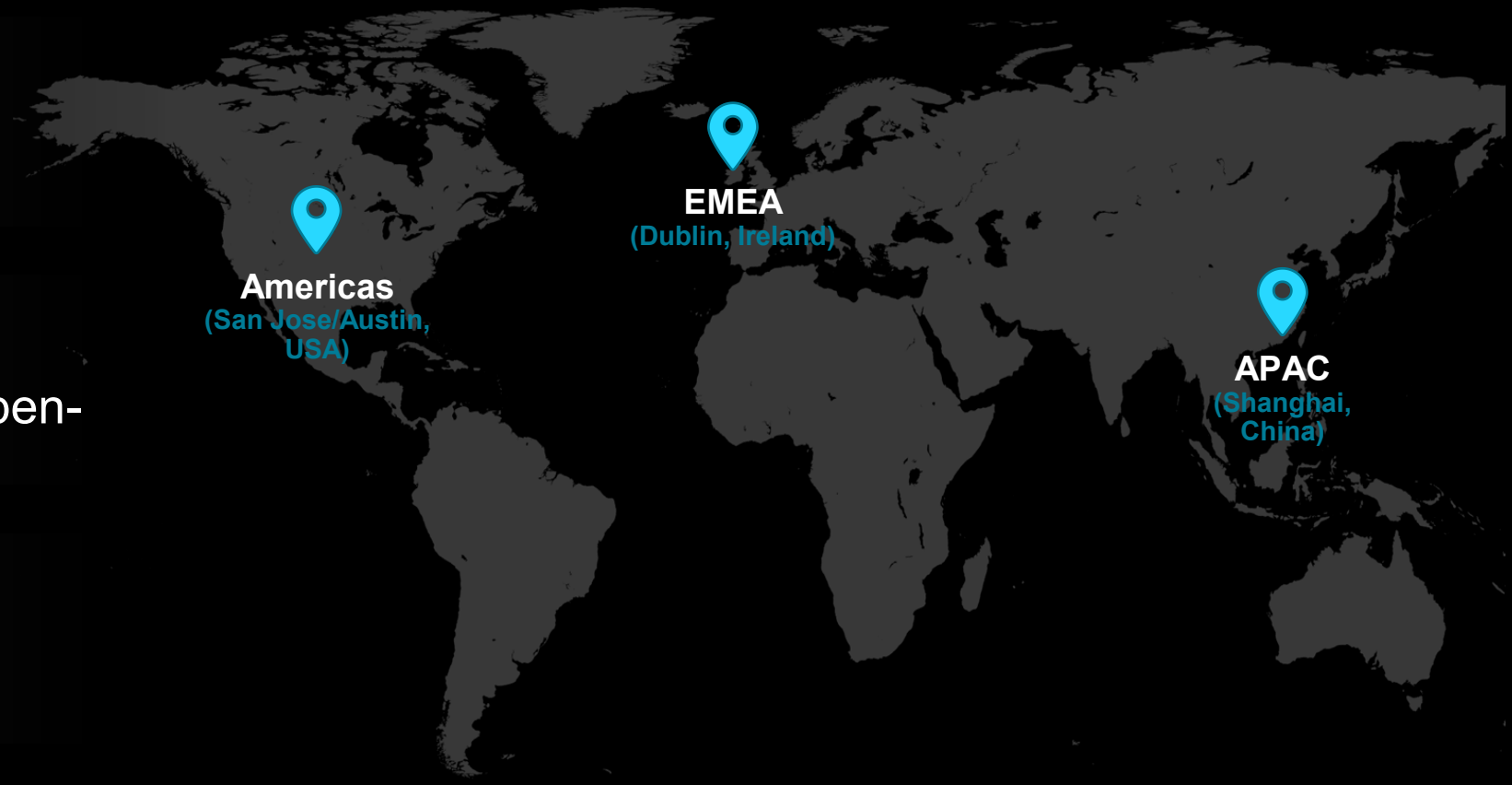
Leadership AI Processing at the edge

See endnote: SHO-000

AI Everywhere, for Everyone

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

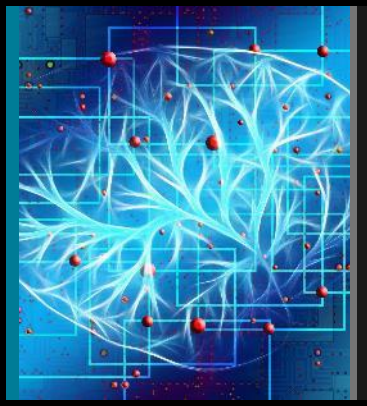
Low-cost Academic Hardware

Donation Program

Teaching Resources

Training

Support



Research Programs

Reasoning & Agents Fuel Compute Surge

Enterprise

GPU

CPU

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)

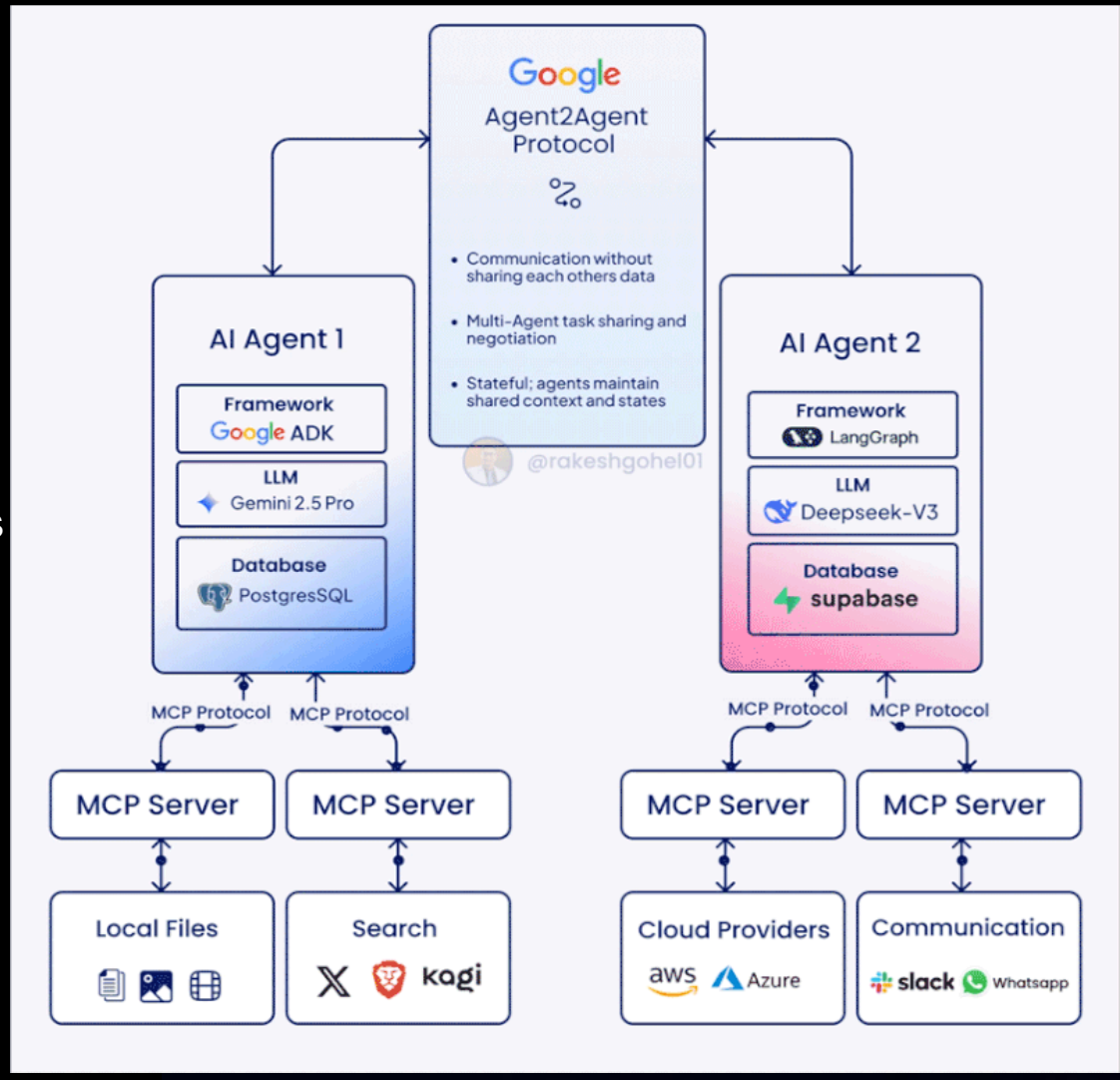
Language models learn from texts. They recognize patterns and generate responses (the likely next word!)

RAG-Systems

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

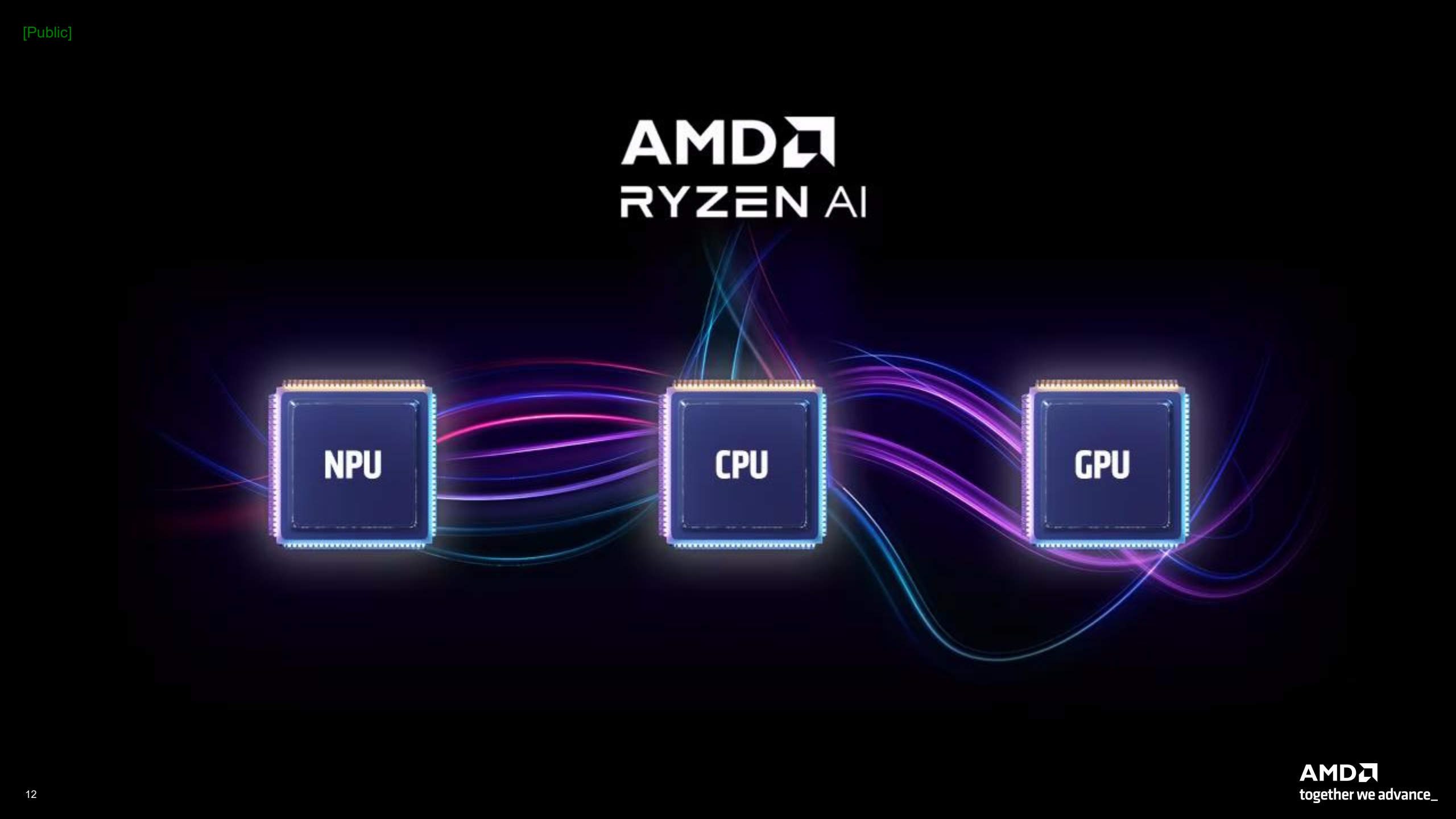
Agentic AI and AI Agents

AI systems act independently. They solve complex tasks without intervention

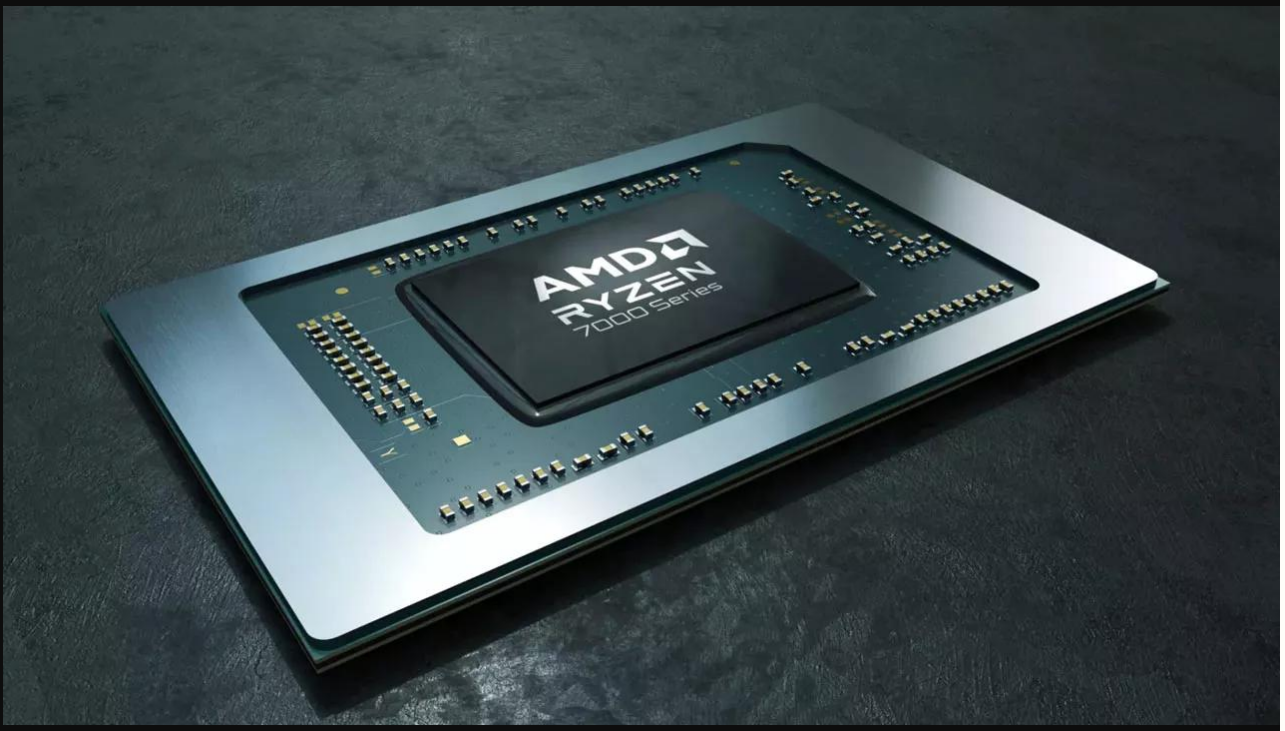


AI PC
Drive the future of AI in personal computing

AMD
RYZEN AI

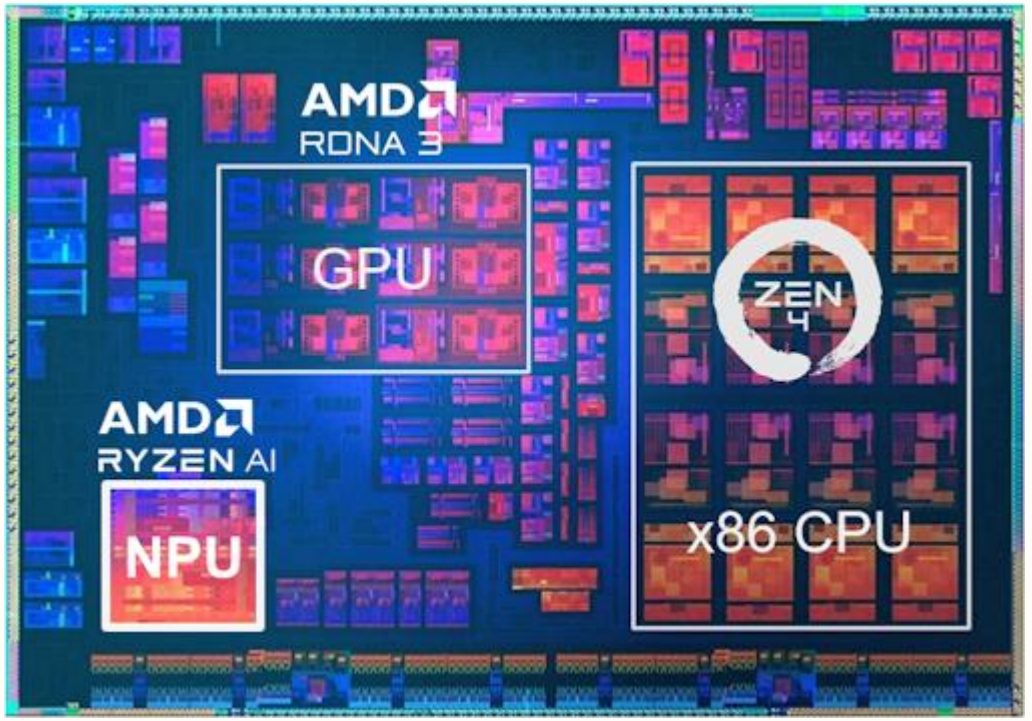


Ryzen 7040 'Phoenix'



First x86 chips to integrate

CPU, GPU and *IPU*



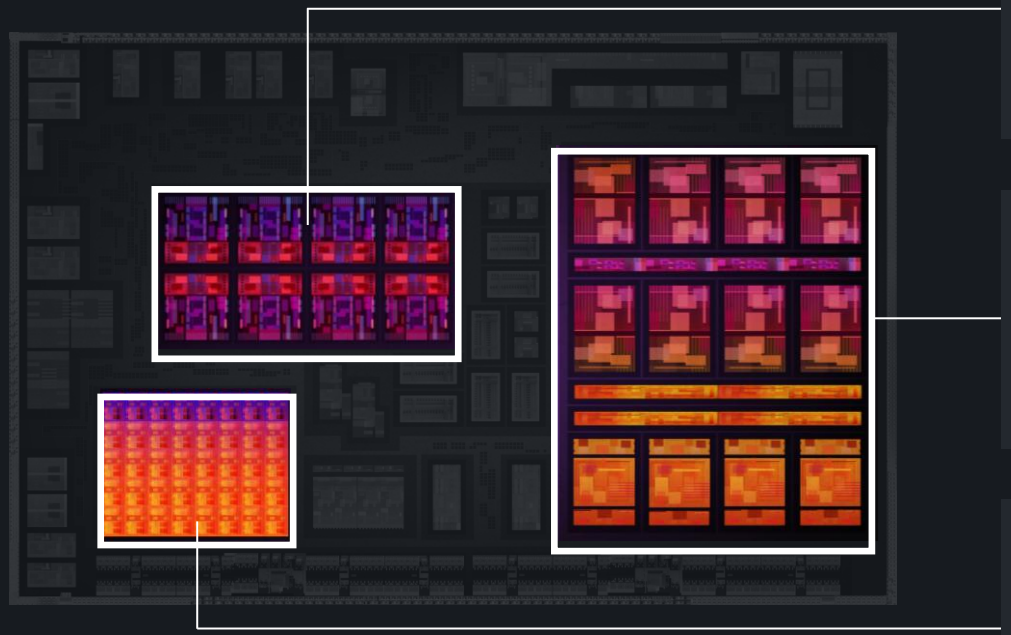
Phoenix floorplan

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

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

AMD Ryzen™ 7040 Series - Introduction

"ZEN 4" Core

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

RDNA™ 3 Graphics

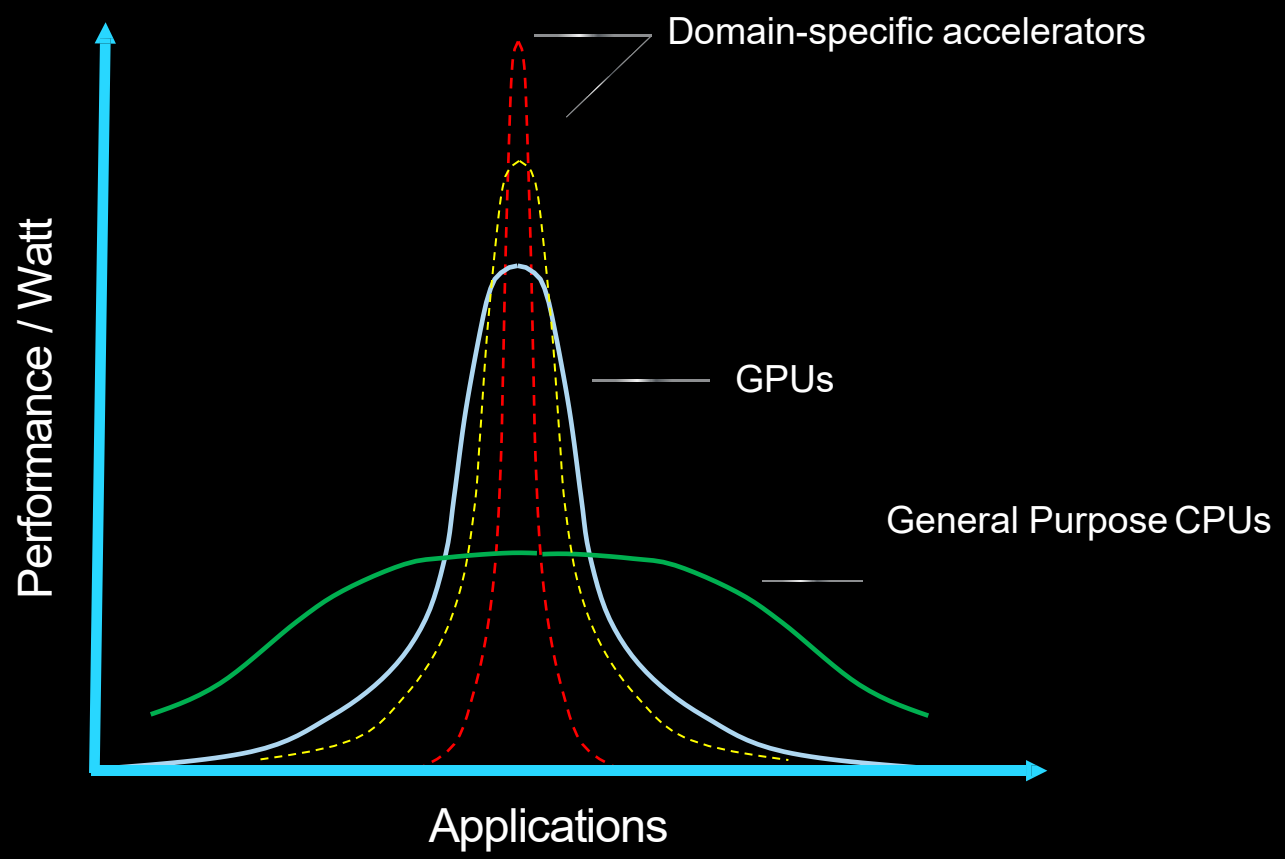
- Improved perf/W per compute unit

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

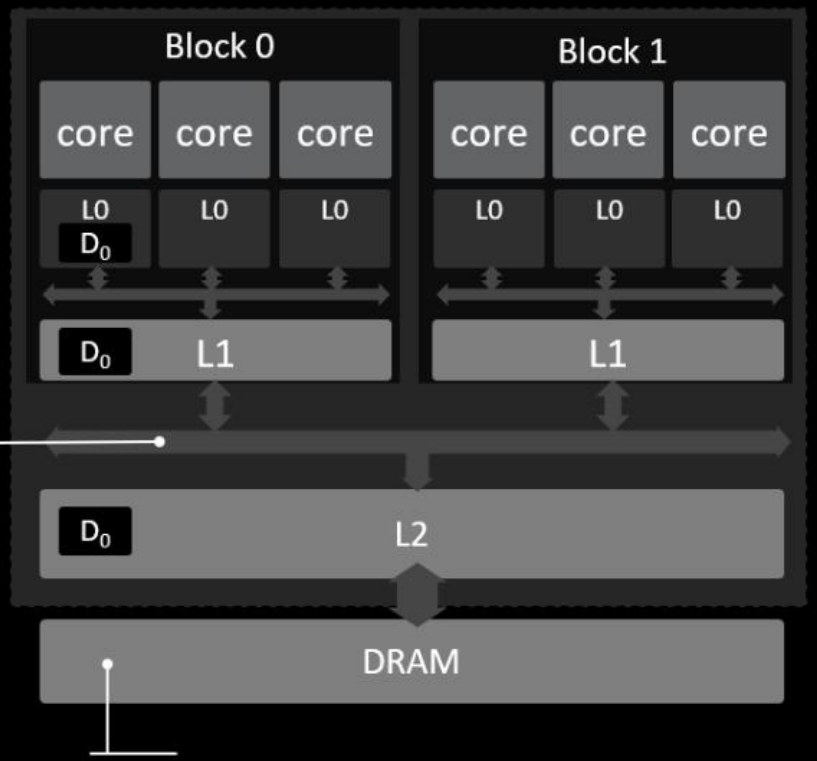
Enhanced Experiences

Client = Personalized and Private

Reliable and Fast

TRADITIONAL MULTI-CORE VS. AMD XDNA™

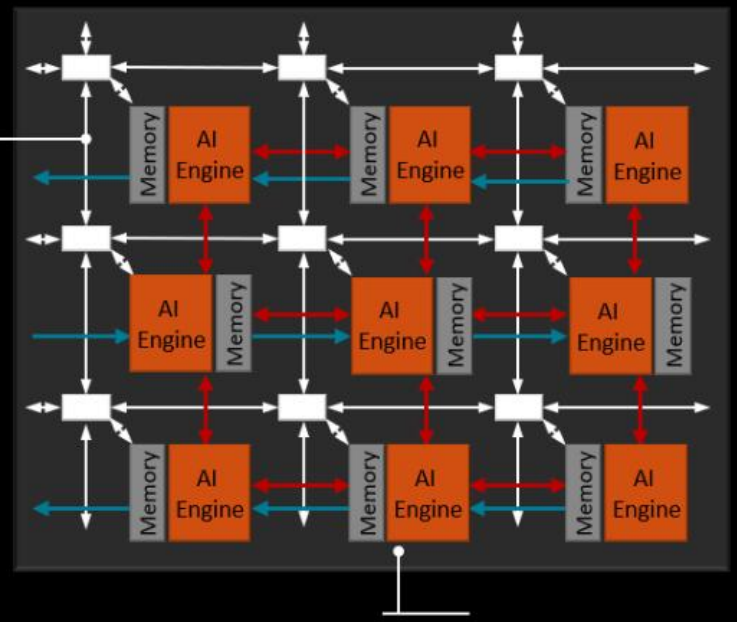
Traditional Multi-Core Processor
(cache-based architecture)



Fixed, Shared Interconnect
- Blocking limits compute
- Timing not deterministic

Data Replicated
- Robs bandwidth
- Reduces capacity

AMD XDNA™ AI Engine
(intelligent engine)

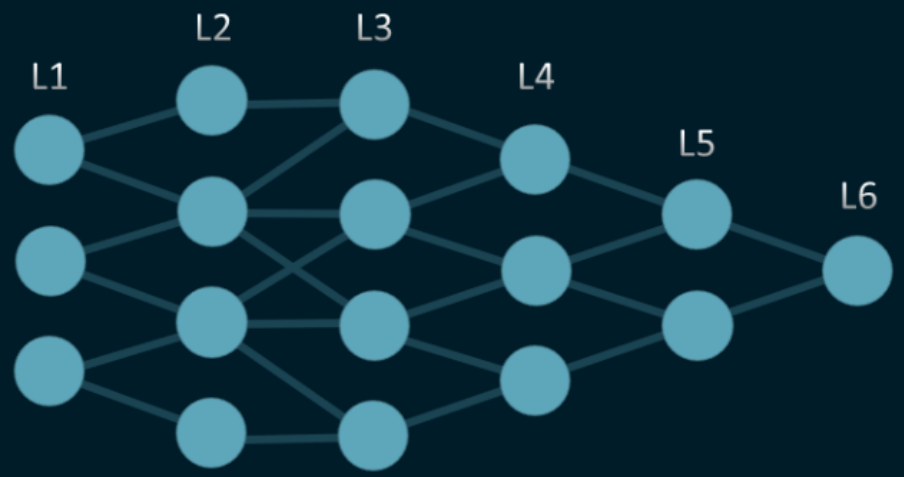


Dedicated Interconnect
- Non-blocking
- Deterministic

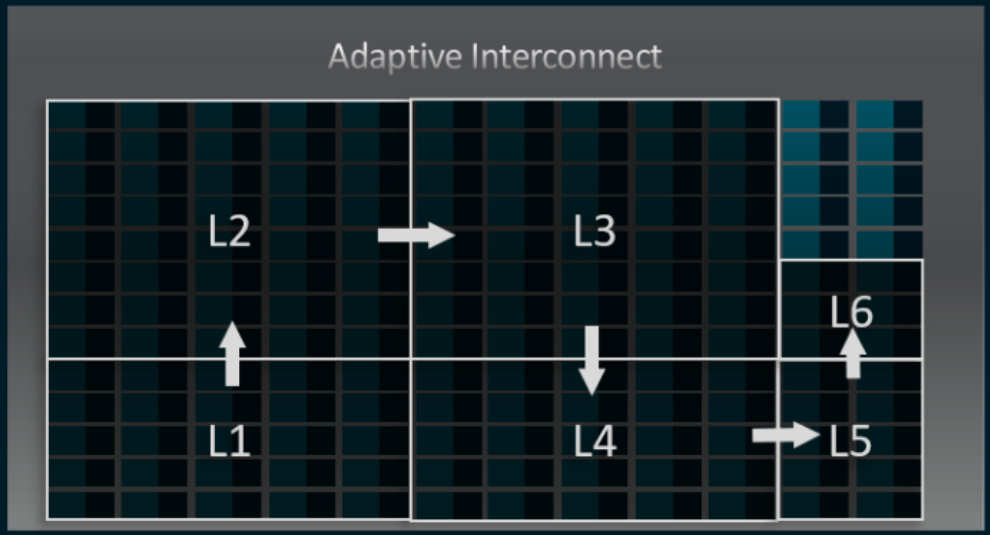
Local, Distributed Memory
- No cache misses
- Higher bandwidth
- Less capacity required

HOW AMD RYZEN™ AI PROCESSES INFERENCE MODELS

A Neural Network activates 'Neurons' from Layer to Layer



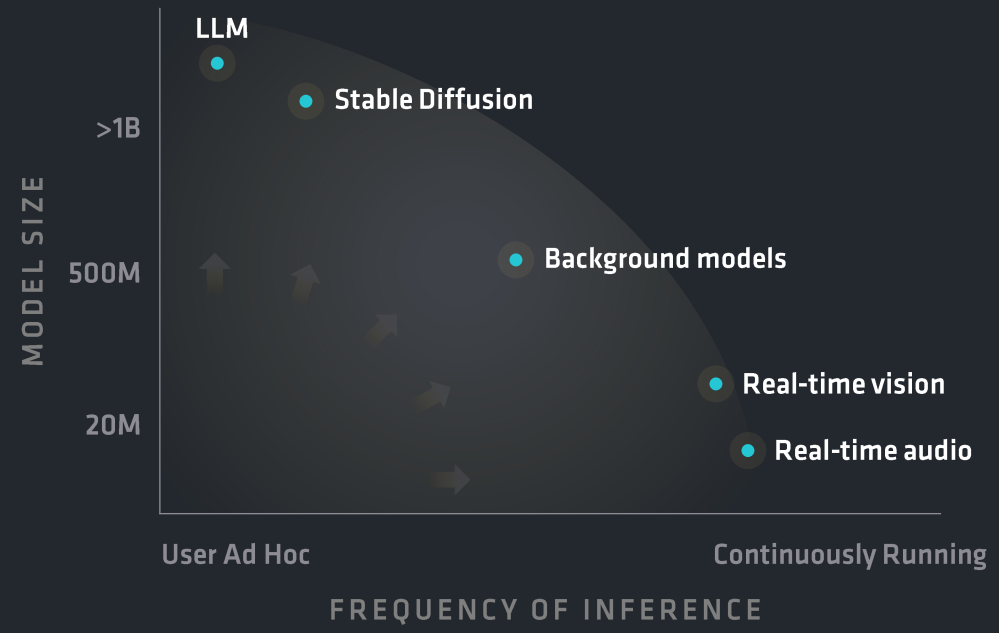
AMD XDNA™: Adaptive Dataflow Architecture



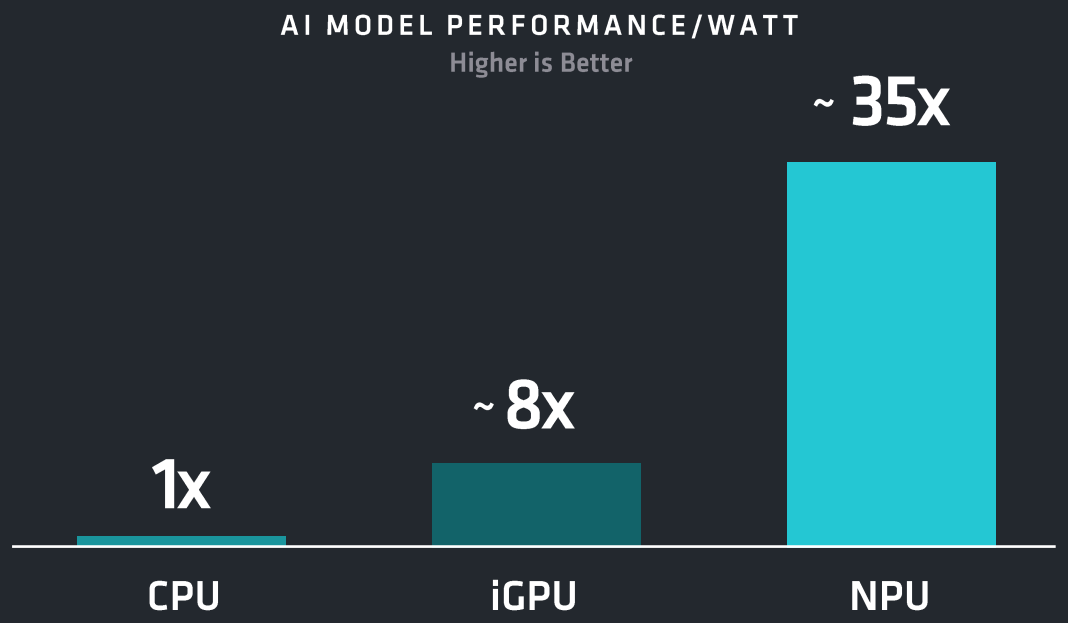
High-Performance, Energy Efficient, and Customizable for AI Workloads

Why NPUs?

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

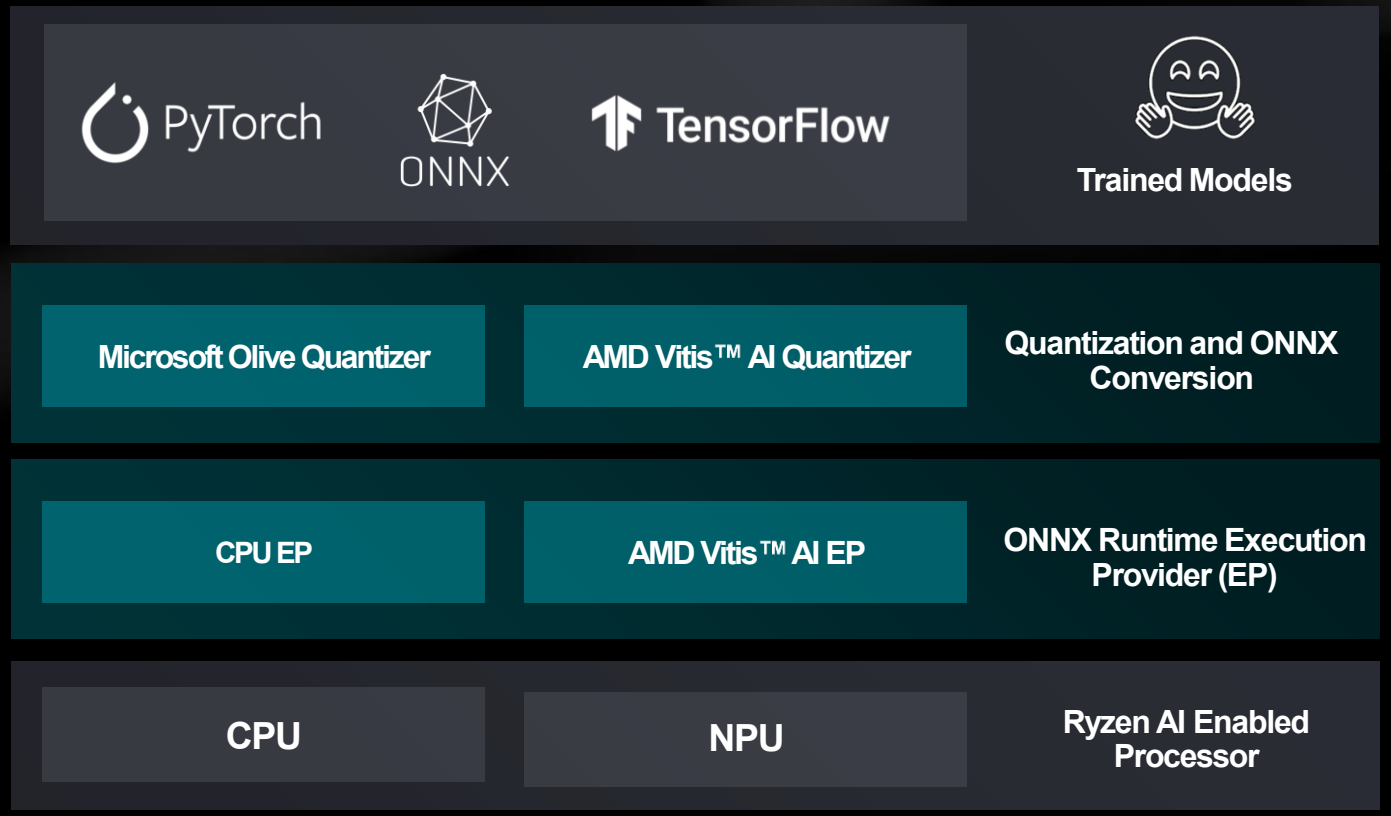


Enhanced AI efficiency matters more now than ever before



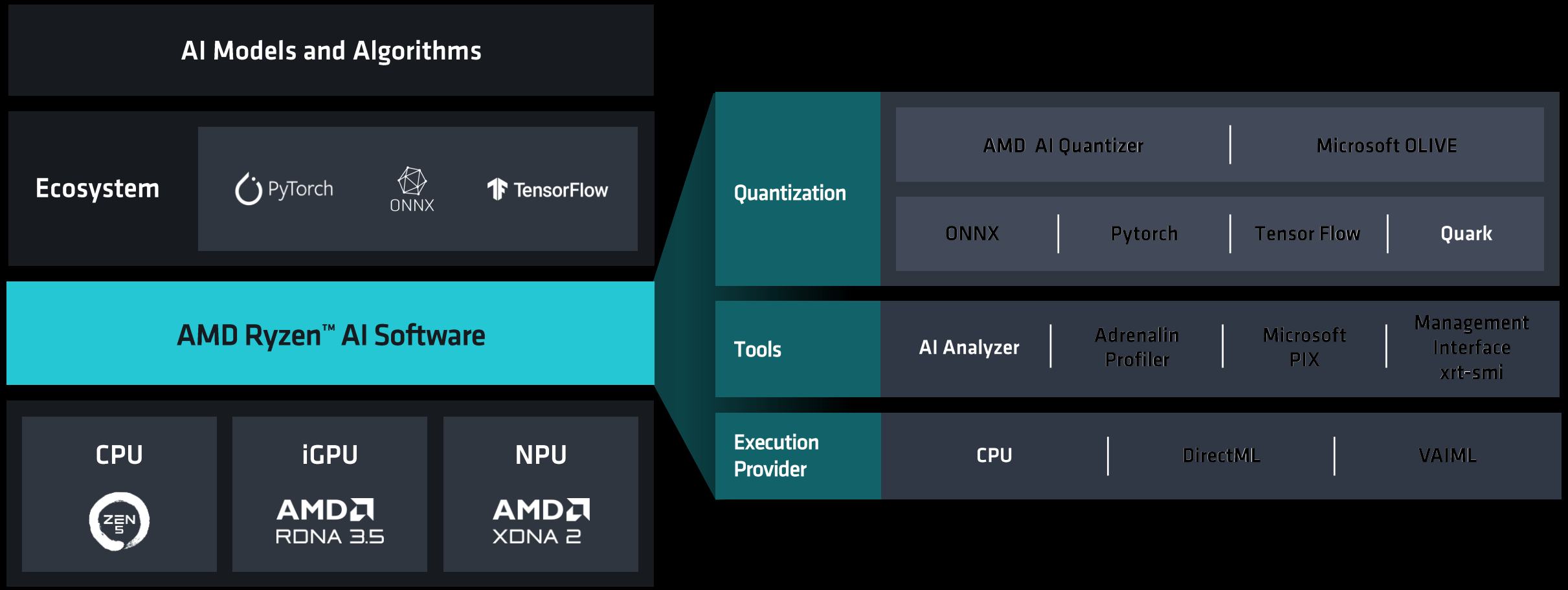
See endnote STX-41

Ryzen™ AI Software Solution

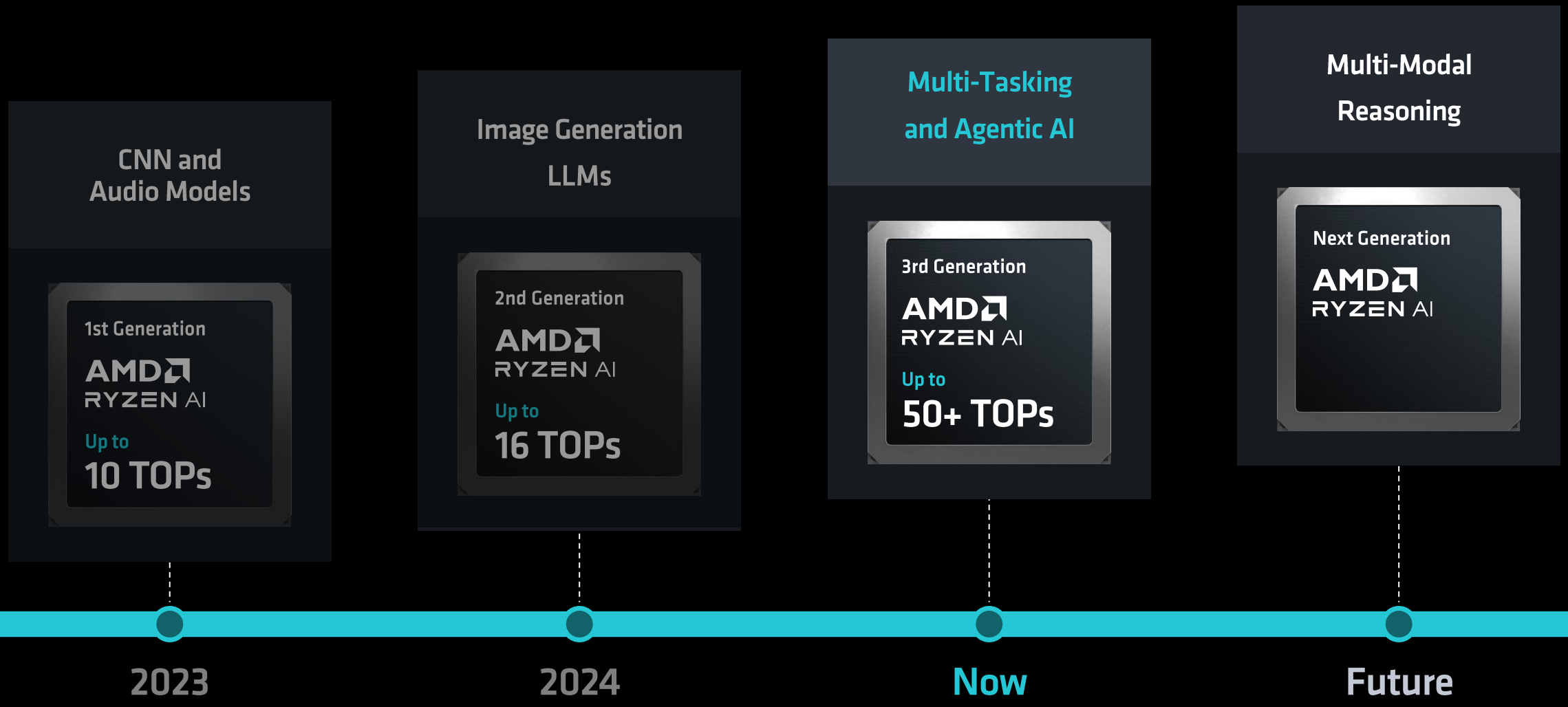


- Enable ONNX Runtime Applications on Ryzen AI Technology
- Build and deploy AI Apps locally
- Run private on-device LLM AI Workloads
- Quick Deployment with optimized pre-trained models
- Single-click installation in minutes

AMD Ryzen™ AI Software



Ryzen AI Roadmap

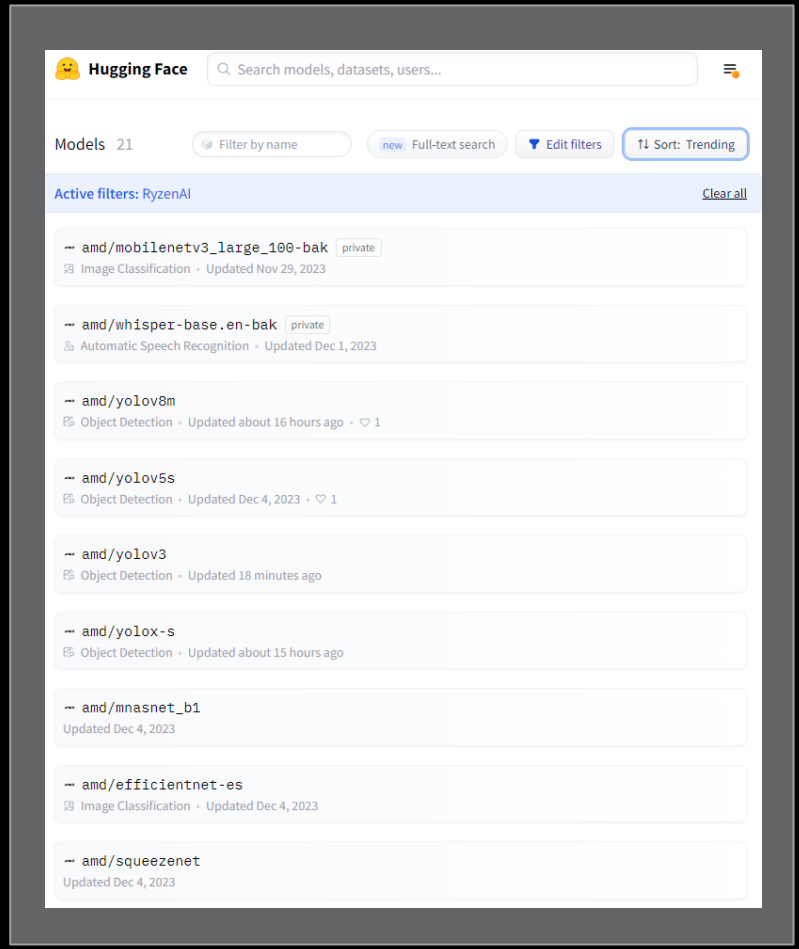


See endnote GD-243

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

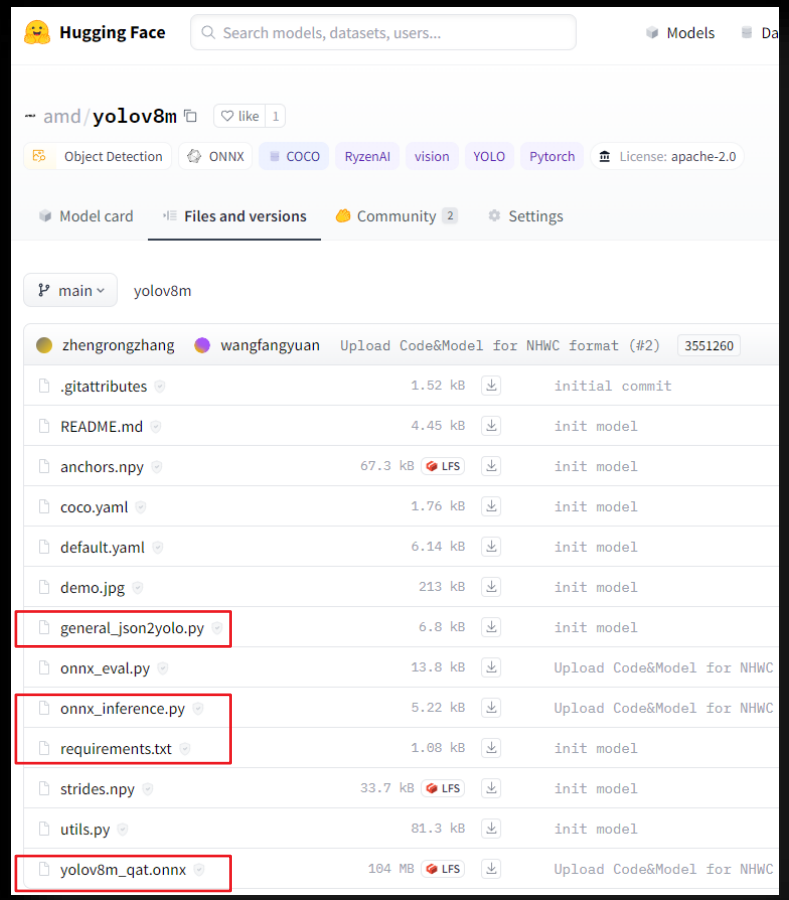
- Get Models
  - GitHub repository
  - Model Zoo on Hugging Face
    - https://huggingface.co/amd



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`

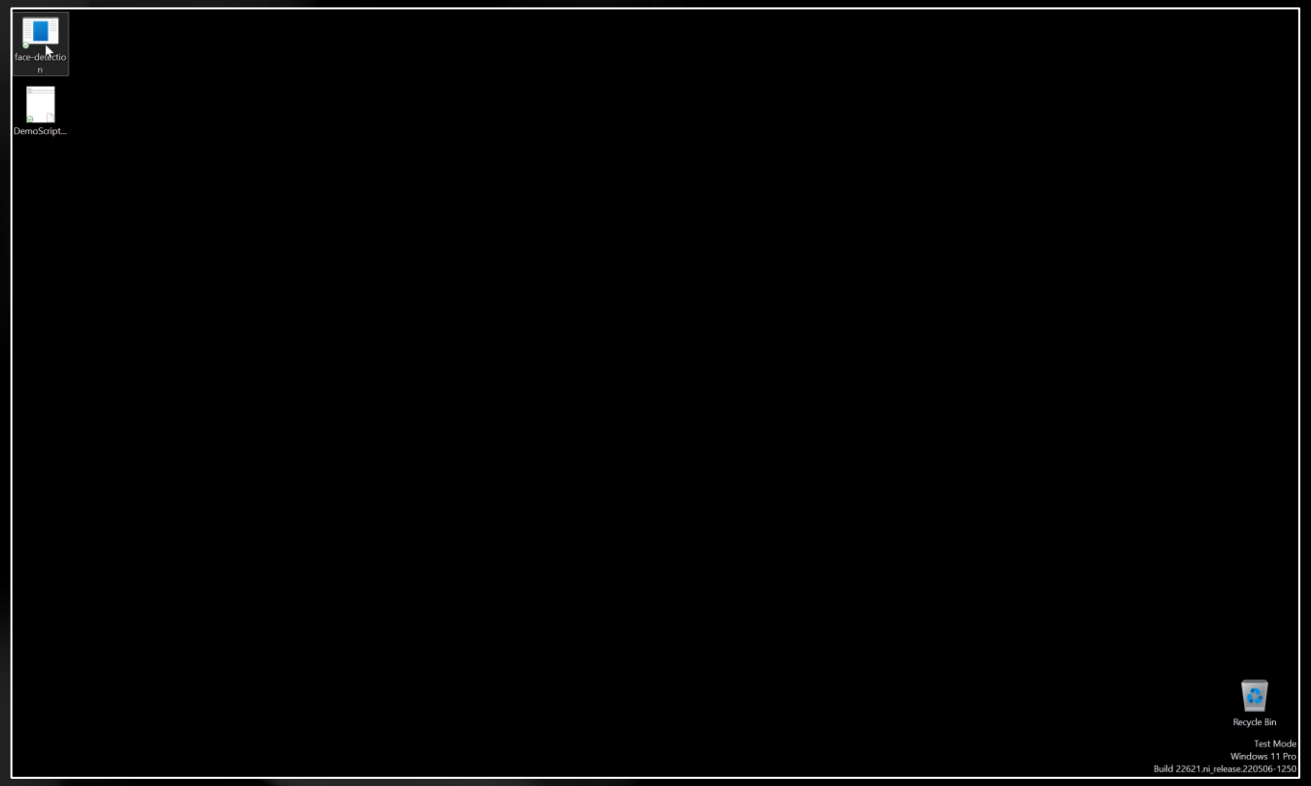
Run inference for a single image

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

Test accuracy of the quantized model

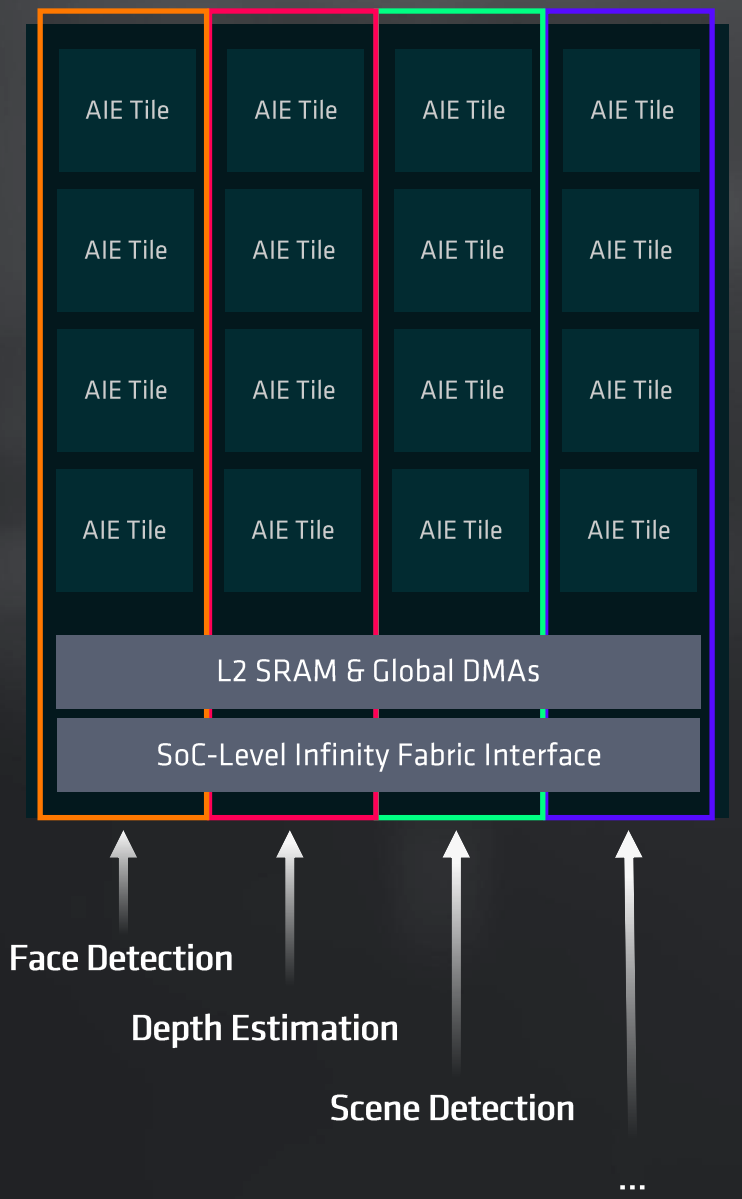
- `python onnx_eval.py --ipu --provider_config vaip_config.json`

Running multiple apps with no performance degradation

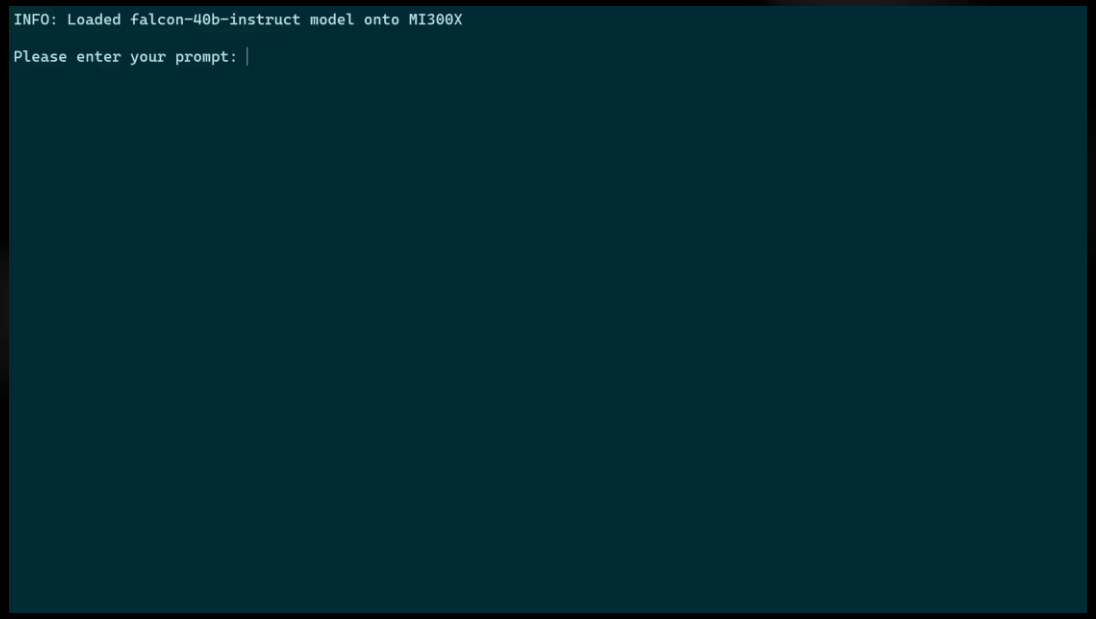


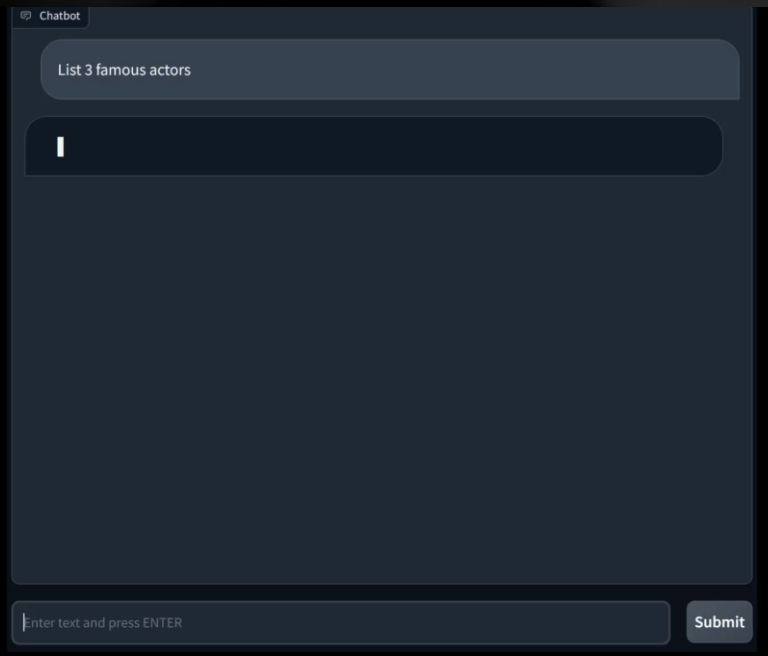
Spatial partitioned AI multi-task acceleration

Ryzen™ AI



Scaling LLMs from cloud to endpoint





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™

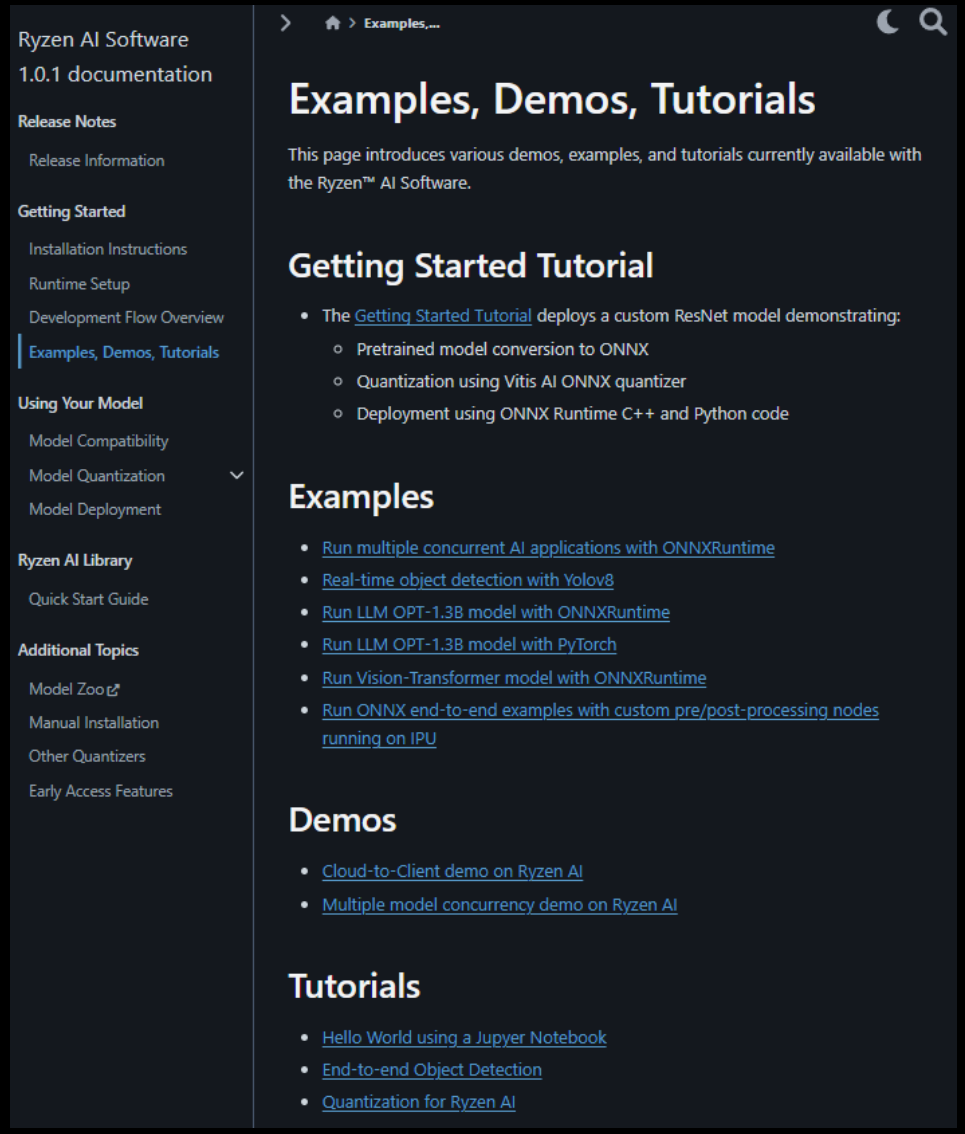
**INT8 models

“Hawk Point” AI PC processors now shipping

AMD Ryzen™ 8040 Series Processors

ZEN 4	AMD RDNA 3	AMD XDNA	Up to 39 TOPS
8 Core | 16 Threads	Graphics	16 TOPS	Total processor performance

Ryzen AI Docs/GitHub



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.

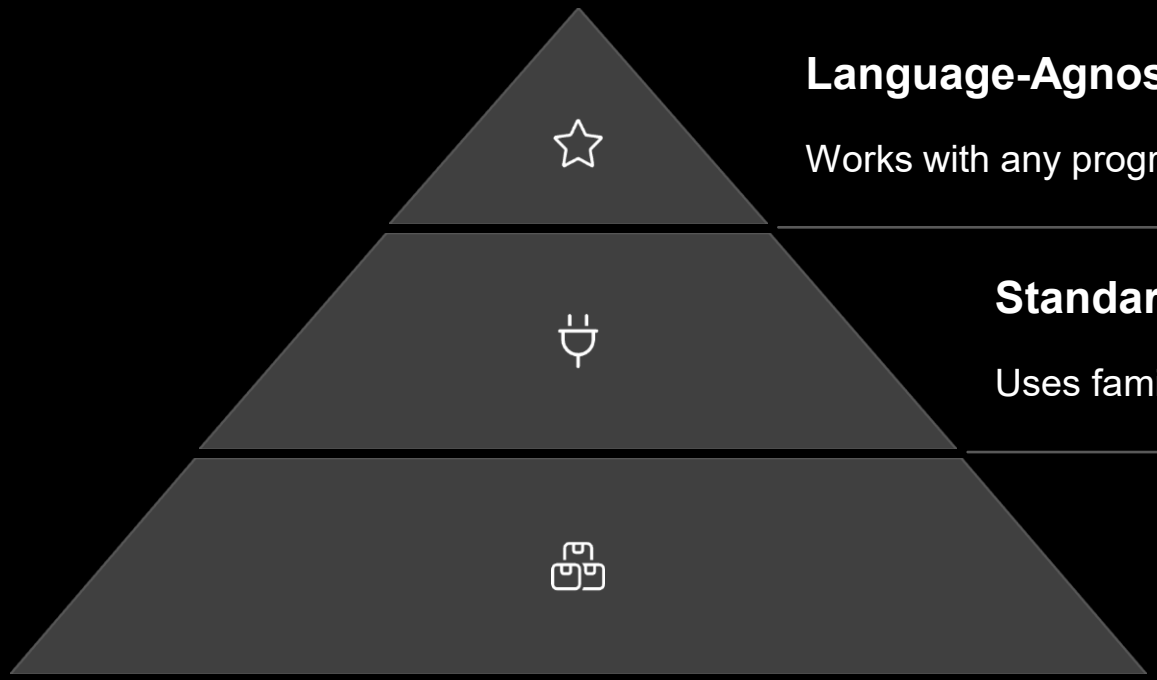
Serve & Benchmark LLMs

Run and test language models on various hardware.

Hardware Support

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

Key Feature: OpenAI API Compatibility



Language-Agnostic

Works with any programming language

Standard API

Uses familiar OpenAI API format

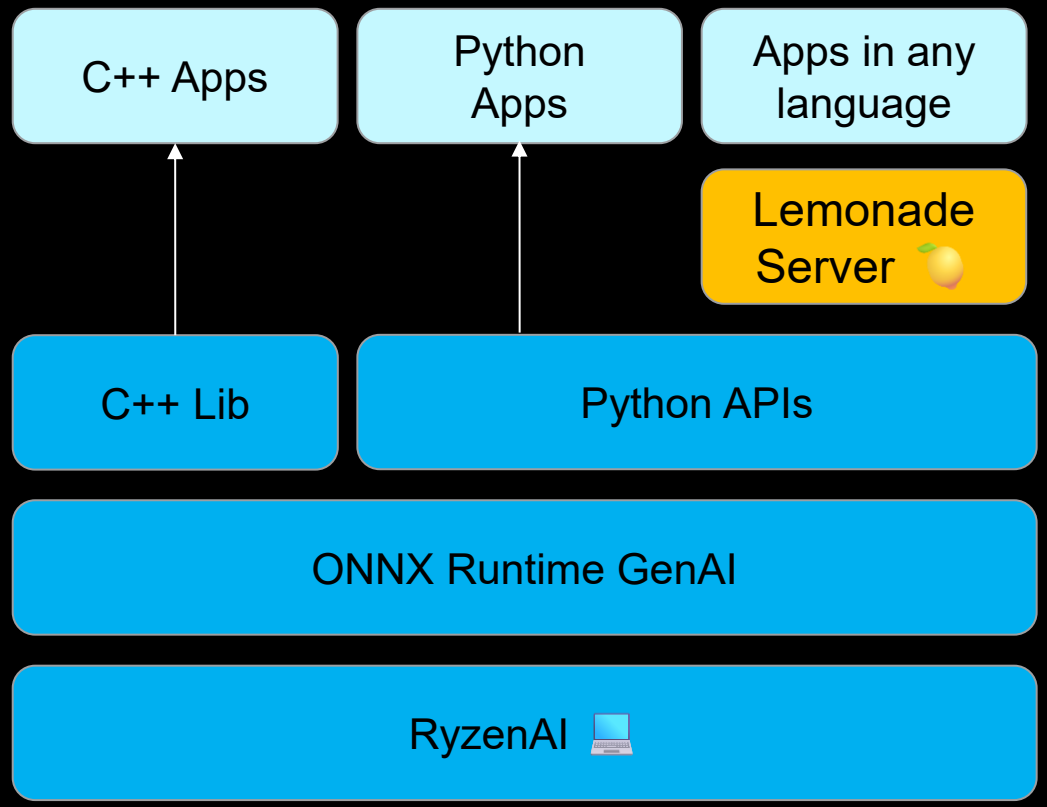
Out-of-Box Compatibility

Existing OpenAI apps work without modifications

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



OpenAI Compatible!

- Develop using OpenAI Standard

```python
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

Open Source!

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

Step 1: Install Lemonade Server

Download Installer

Navigate to TurnkeyML releases page and download the standalone installer.

Run Installer

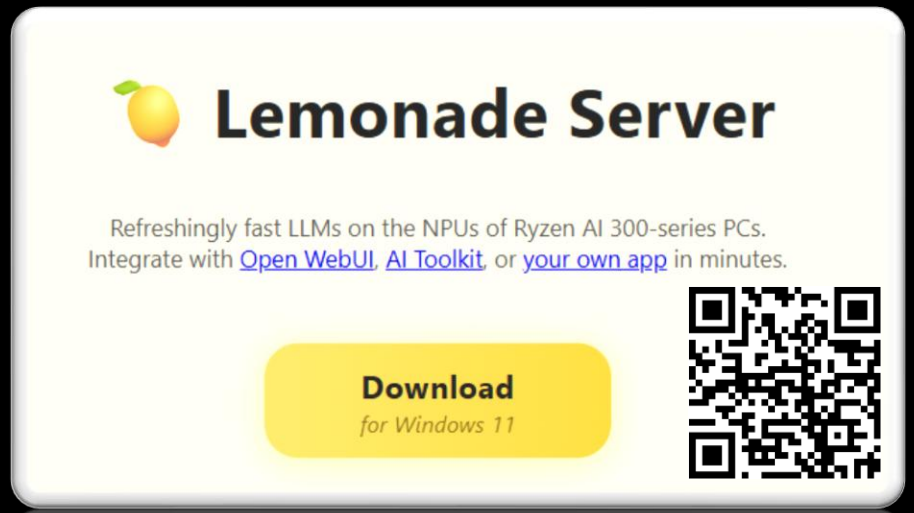
Click "open file" to run the installer after download completes.

Select Models

Choose which LLM models to download during installation.

Launch Server

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



Step 2: Install Open WebUI

Create Conda Environment

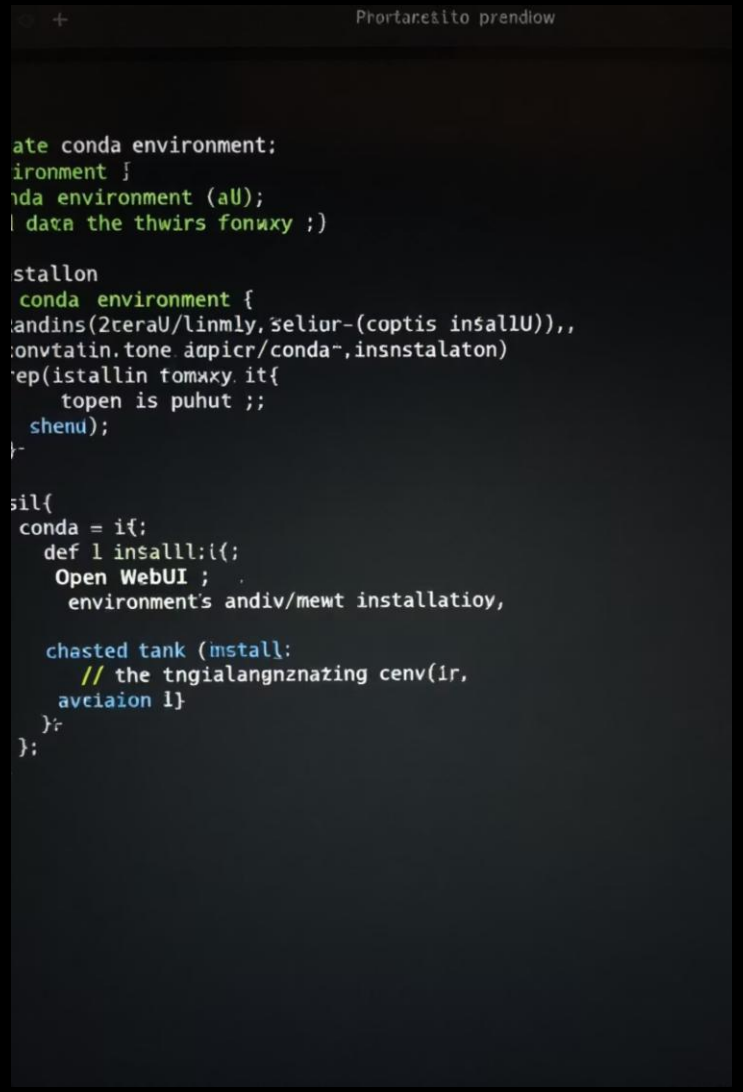
Follow the Open WebUI Quick Start Guide to create and activate a conda environment.

Install via Pip

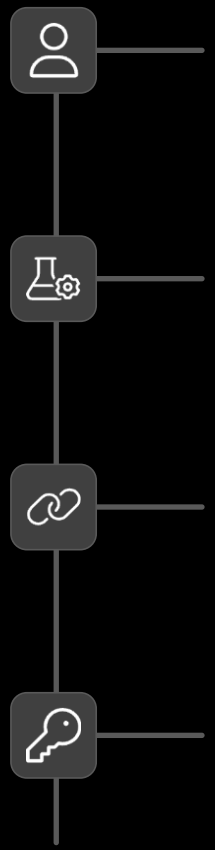
Use pip to install Open WebUI in your activated environment.

Launch Application

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



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.

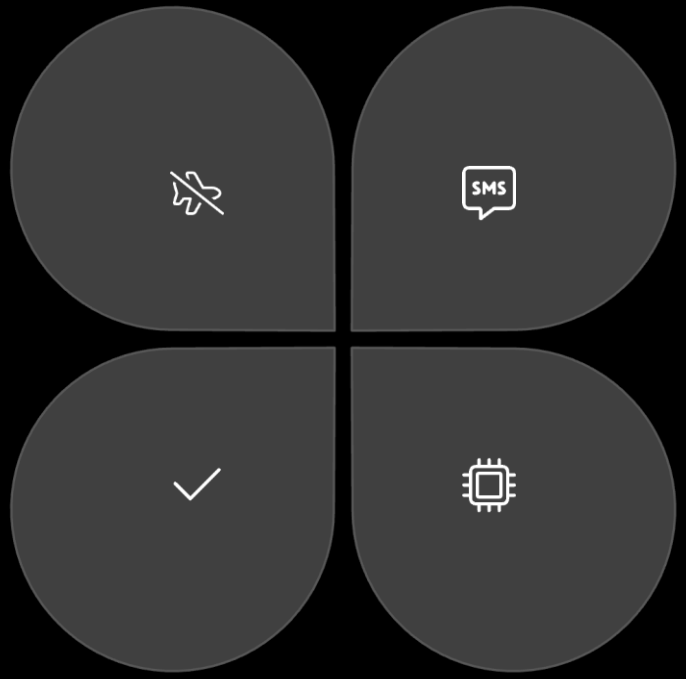
Configure Connection

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.

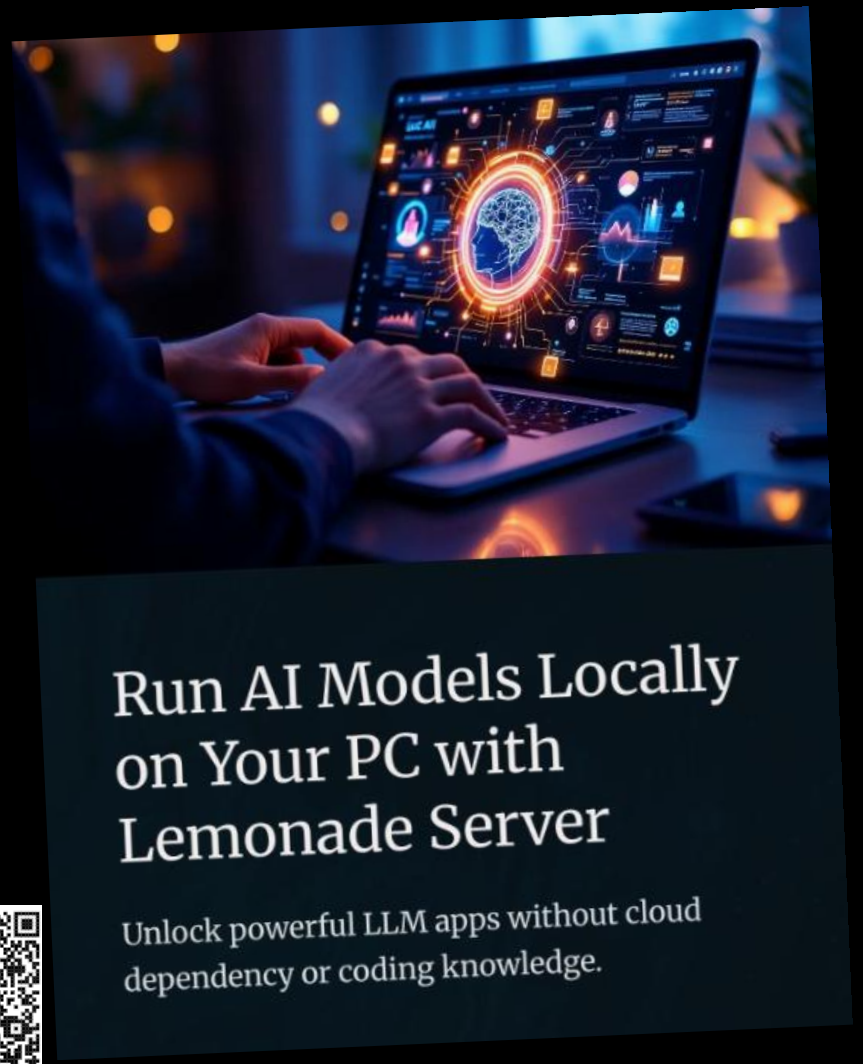
Support

Contact turnkeyml@amd.com for feedback or questions.

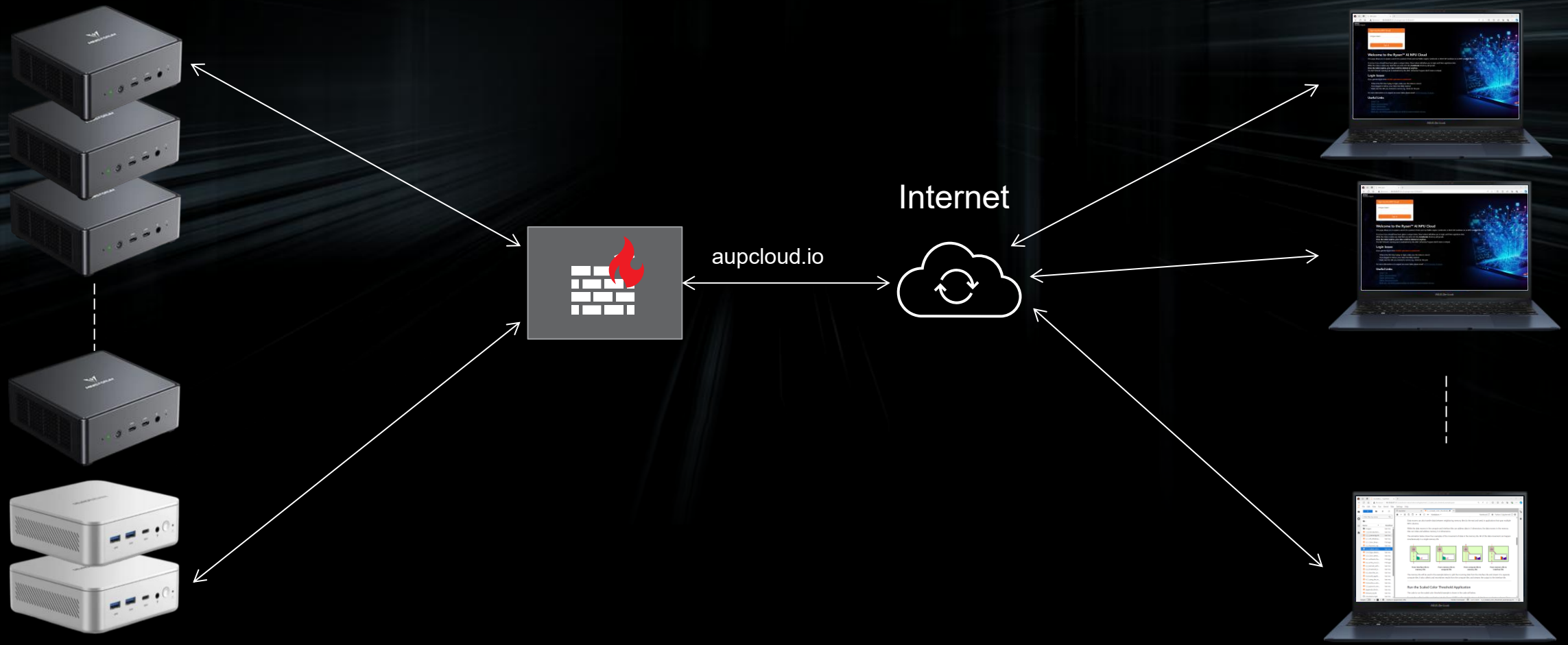
Lemonade Server: AMD’s Lightweight Inference Runtime

- **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

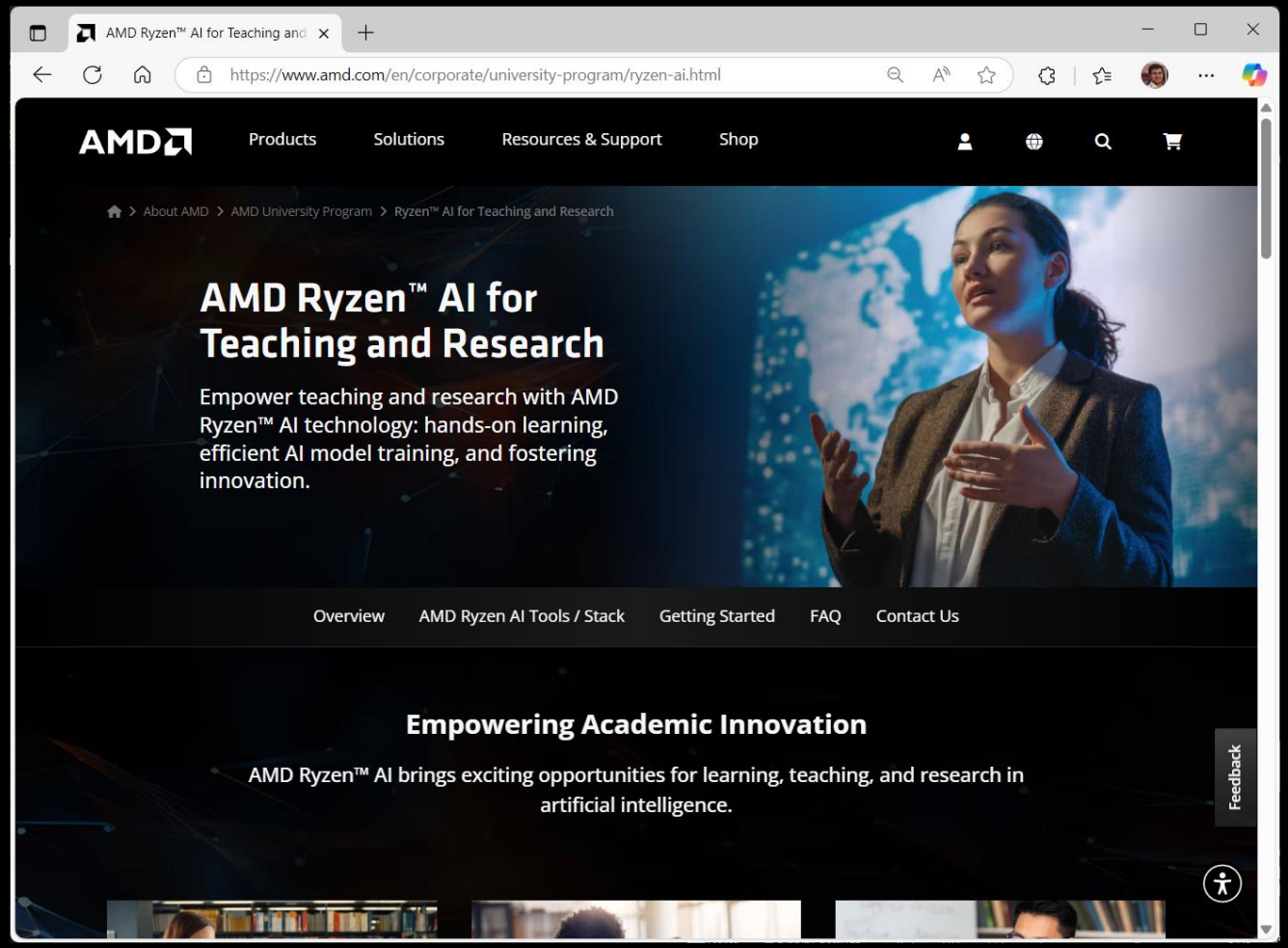


Enabling AI PC Remotely



Call to Action

Get started with Ryzen AI now



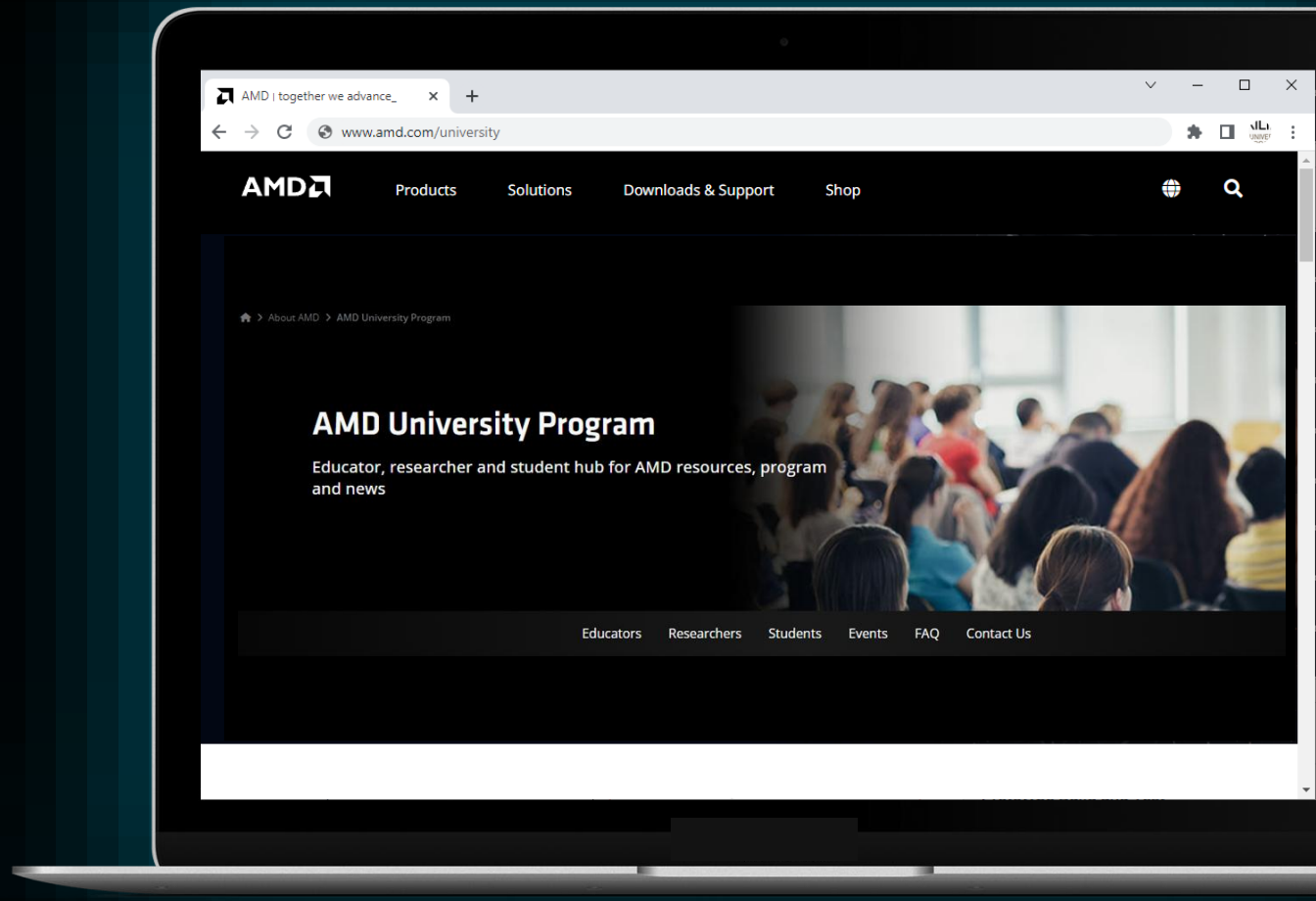
Contact Us

Visit our website to:

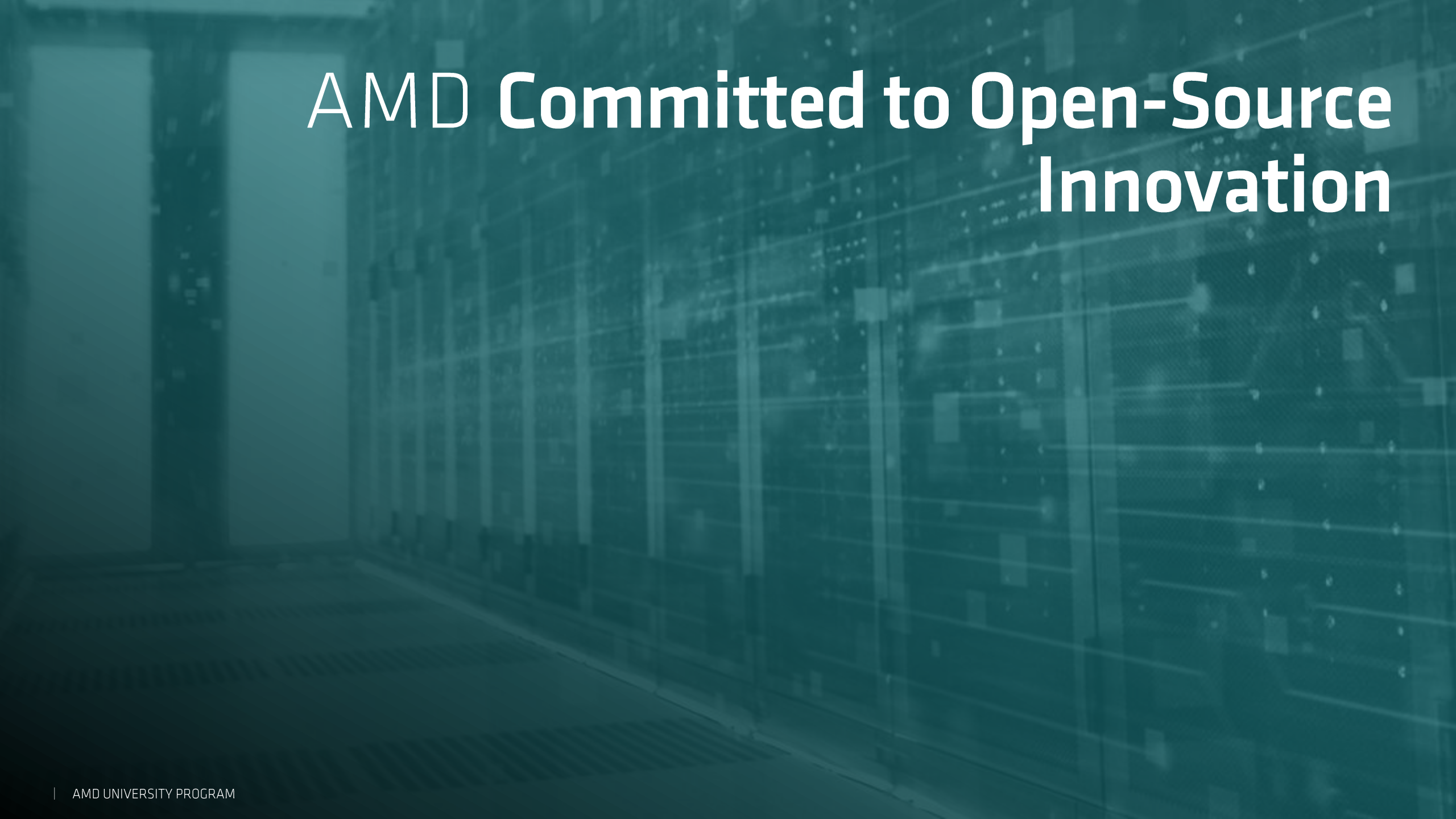
- Discover our research programs
- Access educational resources
- Submit a donation request
- Find training & other events

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

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

AMD ROCm

Rapidly Advancing Open Software Capabilities

Day-0 Support
for Leading Models

Accelerated
Pace of Innovation

Broadening
Ecosystem Partnerships

Developer First
Approach to Enablement

PyTorch · Triton · Hugging Face · TensorFlow · ONNX · vLLM · SGL

GPT · LLaMA 4 (Meta) · deepseek · Grok · Gemma 3 · Qwen · Command A · MISTRAL AI_

AMD ROCm

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

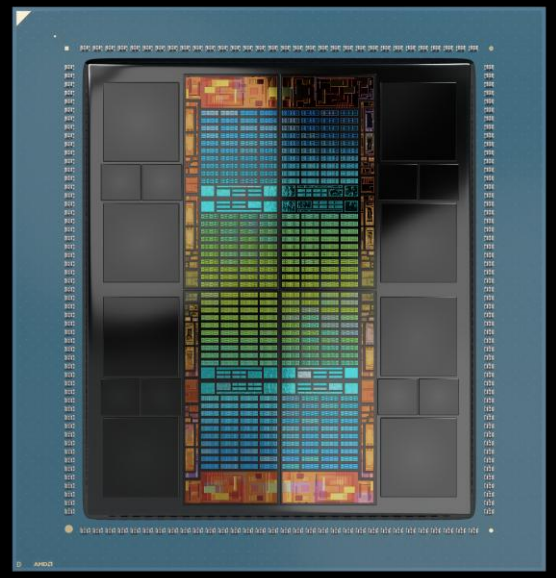
vLLM

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 · Hugging Face · StarCoder · Phi

Falcon LLM · BLOOM · Qwen · MPT · MISTRAL AI_ · Stable Diffusion

GPT-NeoX · D B R X · AlphaFold 2 · OPT · Databricks' Dolly 2.0 · Gemma · VICUNA

deepseek

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

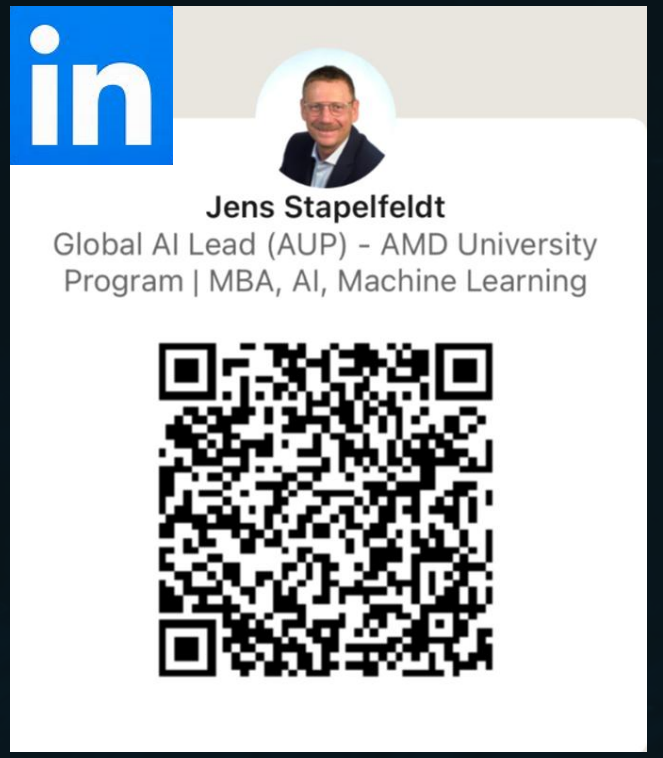
Integrate AMD Ryzen™ AI into their curriculum, enabling interactive lessons on neural networks, automation, and AI-driven creativity.

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?



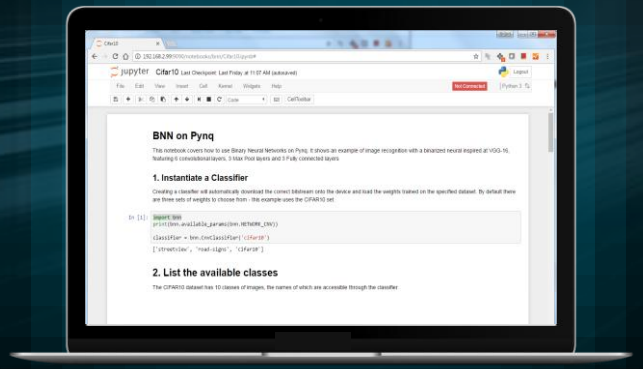

AMD
together we advance_

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

Enabling Python & Jupyter on Adaptive Compute Platforms

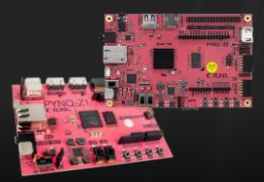
High-level Python APIs and libraries

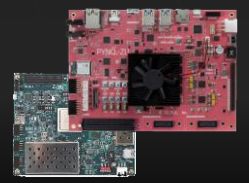


AMD PYNQ

python™

jupyter


AMD ZYNQ

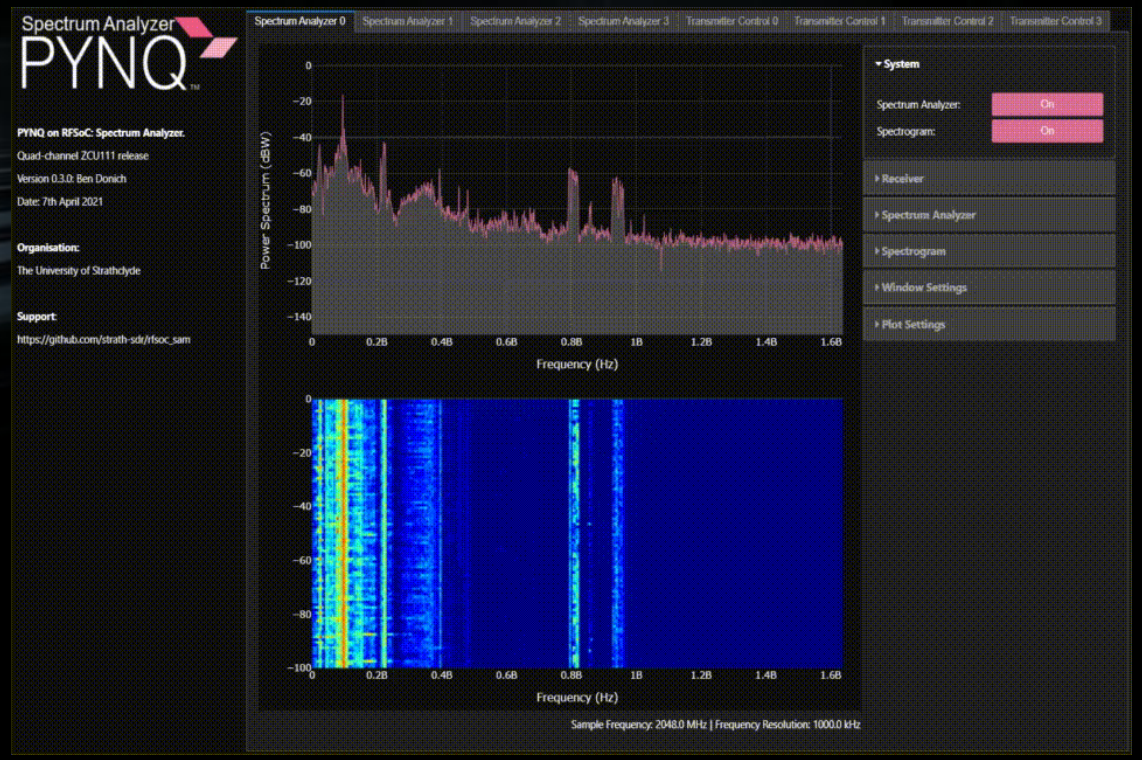

AMD ZYNQ MPSoC


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