

Local, Open, and Energy-Efficient: Harnessing x86 NPUs for Everyday AI Tasks

Ivan Burmaka

Chernihiv Polytechnic National University
Information Technology and Software Engineering Department

October 2, 2026

Outline

- 1 Introduction
- 2 NPU Architecture
- 3 Software Ecosystem
- 4 Performance & Efficiency
- 5 Conclusions

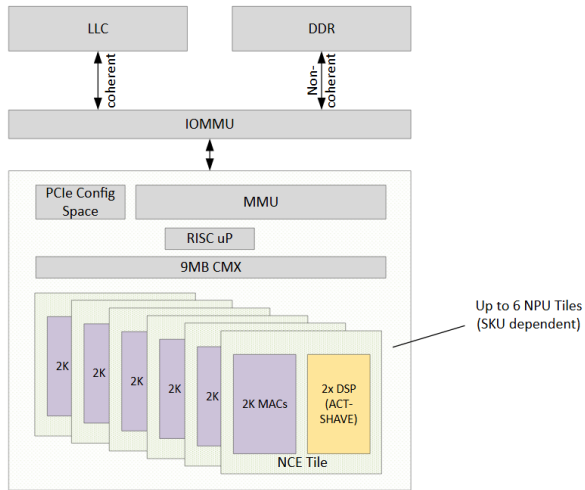
What are NPUs?

Neural Processing Unit (NPU)

- Specialized AI accelerator in laptop CPUs
- Dedicated hardware for AI inference
- Optimized for power efficiency
- Enables on-device AI processing

Why Now?

- Privacy concerns with cloud AI
- Battery life requirements
- Latency-sensitive applications
- Generative AI everywhere



Key Questions

- How do x86 NPU architectures compare?
- What open-source tools enable NPU development?
- What are real-world performance gains?
- Can NPUs handle practical AI workloads efficiently?

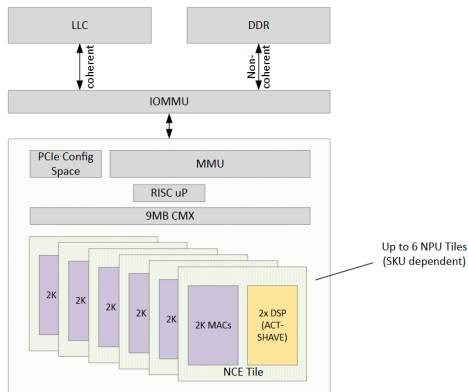
Research Focus

- Intel Core Ultra 200V series
- AMD Ryzen AI (XDNA architecture)
- Open-source software ecosystem
- Practical applications: video segmentation, image generation

Intel NPU Architecture

Core Ultra 200V NPU

- **Host Control:** 64-bit RISC-V microcontroller
- **Deep Learning Accelerator:**
 - 6 Neural Compute Engine (NCE) Tiles
 - 12,288 INT8 MACs total
 - 12 DSP processors
 - 9 MB near-compute memory
- **M2I Engine:** Media preprocessing (crop, scale, color conversion)
- **Precision:** INT4, INT8, FP16



Design Philosophy:

- Hardware-managed scheduling
- Ease of use focus

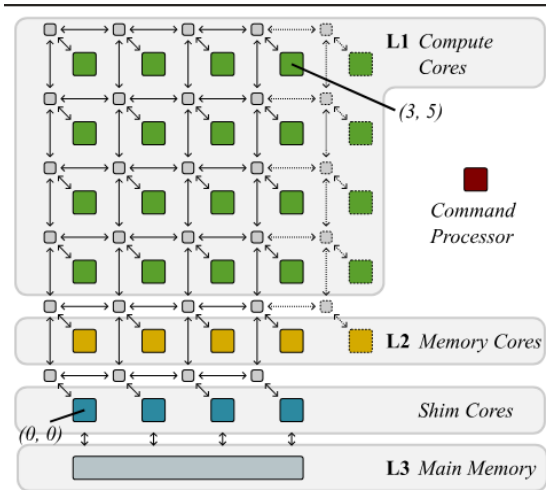
AMD XDNA NPU Architecture

Ryzen AI NPU (XDNA)

- Derived from Versal ACAP
- **AI Engine (AIE) Tiles:**
 - Vector processors for ML ops
 - Local scratchpad memory
 - Dedicated DMA engines
- Packet-switched NoC interconnect
- **Precision:** FP16, INT8, INT4

Design Philosophy:

- Programmable architecture
- Explicit data movement control
- Flexibility for specialized apps



Architecture Comparison

Feature	Intel NPU	AMD XDNA
Approach	Automated	Programmable
Scheduling	Hardware-managed	Software-defined
Target Users	General developers	Optimization experts
Programming	High-level only	High & low-level
Precision	INT4/8, FP16	INT4/8, FP16
Memory	9 MB unified	Distributed scratchpad
Philosophy	Ease of use	Maximum control

Common Ground

Both excel at: matrix multiplication, convolution, transformers, mixed-precision compute, on-chip execution for power efficiency

Open-Source Development Frameworks

Intel Ecosystem

- **OpenVINO:** Primary toolkit
 - Model optimization
 - NNCF quantization
 - Deployment tools
- Fully open-source pipeline
- Linux kernel drivers included

Universal Runtime

- **ONNX Runtime**
- Hardware abstraction
- Execution Providers per vendor
- Unified API

AMD Ecosystem

- **Ryzen AI Software:** PyTorch/TensorFlow support
- **AMD IRON:** Low-level programming (MLIR-AIE based)
 - Python layout scripts
 - C++ compute kernels (AIE API)
 - Problem-specific optimization
- Mixed approach: open drivers, proprietary compiler

Windows ML

- OS-level NPU abstraction
- DirectML backend
- Cross-vendor (Intel, AMD, Qualcomm)

Practical Open-Source Applications

Audacity + OpenVINO Plugins (Intel)

AI-powered audio editing on NPU:

- Noise suppression
- Vocal isolation (stem separation)
- Local music generation

FastFlowLM (AMD Ryzen AI)

- LLM and vision-language models on NPU
- INT4 quantization
- OpenAI-compatible REST API
- Ultra-lightweight footprint
- **Advantage:** Much cooler operation than GPU (no thermal throttling)

AMD Generative AI Workflows

- Stable Diffusion on NPU via ONNX Runtime + Vitis AI EP
- Custom inference pipelines with AMD IRON

Case Study: Real-time Video Segmentation

Developed human-background segmentation for webcam using Windows ML

- Framework: Windows ML (DirectML backend)
- Target: AMD Ryzen AI 7 350
- Models tested: MediaPipe Selfie Segmentation, PP-HumanSeg

Important Consideration: Layer Compatibility

- Not all neural network operations work on NPU
- Incompatible layers offload to CPU → degrades performance & increases power
- Solution: Choose NPU-compatible models
- Result: PP-HumanSeg and MediaPipe models use ~10% of NPU resources

Experimental Setup

Test Platform

- **Processor:** AMD Ryzen AI 7 350
- **Power Mode:** Maximum energy saving
 - Motherboard Eco mode enabled
 - Windows power saving active
 - Minimizes CPU boosting impact

Test Workloads

- 1 **Video Segmentation:** PP-HumanSeg model
 - Real-time background removal
 - Comparison: NPU vs integrated GPU
- 2 **Image Generation:** Stable Diffusion XL
 - Generative AI workload
 - Backends: DirectML (GPU), ROCm (GPU), NPU-optimized

Video Segmentation Results

Metric	NPU	iGPU
Performance	Excellent	Excellent
Power Consumption	Lower	Higher
Power Savings	23%	Baseline
Temperature	Lower	Higher

Key Findings

- Both accelerators provide superior performance for lightweight models
- NPU achieves same throughput with significantly less power
- 23% power reduction translates to extended battery life
- Lower temperatures prevent thermal throttling

Image Generation Results: Stable Diffusion XL

Backend	Performance	Power	Notes
DirectML (GPU)	16× slower	Very High	Performance issues
Vulkan (GPU)	16× slower	Very High	Backend problems
ROCm (GPU)	Moderate	High	Better than DirectML
NPU-optimized	Best	Lowest	Quantized models
Power Savings	75% vs GPU (DirectML/Vulkan)		

Key Insights

- NPU uses optimized model quantization
- 75% power reduction for complex generative tasks
- Surprising finding: NPU competitive even for heavy workloads
- GPU backends on Windows show optimization challenges

Energy Efficiency Summary

Power Reduction

- Video segmentation: **23%**
- Image generation: **75%**

Thermal Benefits

- 20-28°C lower temperatures
- No thermal throttling
- Quieter operation

Battery Life Impact

- **2-4 hours** extended runtime
- Critical for mobile AI workloads
- Enables sustained AI processing

Practical Implications

- All-day AI-powered workflows
- Background AI without battery drain
- Sustainable generative AI usage

Bottom Line

NPU make continuous AI processing practical on battery power

Key Findings

Architecture

- Intel: Automated, hardware-managed, ease-of-use focus
- AMD: Programmable, software-defined, maximum flexibility
- Both: Effective for modern AI workloads (transformers, convolutions, matmul)

Software Ecosystem

- Mature open-source frameworks (OpenVINO, ONNX Runtime, AMD IRON)
- Practical applications already available (Audacity, FastFlowLM)
- Active community development
- Intel: fully open pipeline; AMD: mixed approach

Key Findings

Performance & Efficiency

- **23-75% power reduction** vs integrated GPU
- **2-4 hours extended battery life** for AI workloads
- **20-28°C lower temperatures** (no throttling)
- Competitive performance even for complex generative AI

Broader Implications

Privacy & Independence

- On-device inference eliminates cloud dependency
- Sensitive data stays local
- No network latency
- Works offline

Sustainability

- Lower power = reduced carbon footprint
- Enables always-on AI features
- Extends device battery longevity

Accessibility

- NPUs democratize AI acceleration
- Available in consumer laptops
- Open-source tools lower barriers
- Community-driven innovation

The AI PC Era

- NPUs transition from niche to essential
- Foundation for next-gen computing
- Real-time, private, efficient AI

NPUs establish energy-efficient, privacy-preserving on-device AI as practical reality for battery-powered platforms

Future Directions

Software Maturity

- Better compiler optimization and auto-scheduling
- Expanded model compatibility (more operations on NPU)
- Unified programming models across vendors
- Enhanced tooling and profiling

Applications

- Real-time language translation
- Advanced video conferencing (effects, transcription)
- Local AI assistants with LLMs
- Creative tools (generative art, music, video)
- Privacy-critical applications (healthcare, finance)

References

- 1 Intel Core Ultra 200V Series Processors Datasheet
<https://edc.intel.com/>
- 2 Rösti, A., & Franz, M. (2025). Unlocking the AMD Neural Processing Unit. *arXiv*.
<https://doi.org/10.48550/ARXIV.2504.03083>
- 3 OpenVINO Documentation
<https://docs.openvino.ai/>
- 4 FastFlowLM
<https://fastflowlm.com/>
- 5 AMD Ryzen AI Software Documentation
<https://ryzenai.docs.amd.com/>

Thank You!

Questions?

Ivan Burmaka

`Ivan.Bourmaka@stu.cn.ua`

Chernihiv Polytechnic National University