Hybrid CPU–FPGA Architectures for Efficient AI Inference

Authors: Aleksandar Peulić, Kristina Vasić
stitute of mathematics and informatics, Faculty of Science, University of Kragujevac
Artificial Intelligence Conference, October 2025

Edge AI

Real-Time Decision Making

- Low latency
- Immediate Reaction

- Autonomous vehicles
- Industrial robotics
- Security systems
- Drones

Data Privacy and Security

- Local data processing
- Reduced exposure

- Video surveillance
- Medical data
- Wearables

Energy Efficiency

- Low power consumption
- Cost Reduction

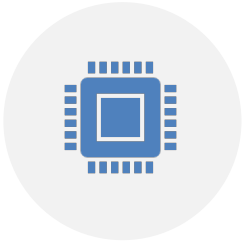
- IOT devices
- Drones

Not network dependant

- Offline Operation

- Rural locations
- Mines
- Sea
- Military systems

Current Trends in Edge AI



Hardware

- NVIDIA Jetson
- Google Coral (Edge TPU)
- Intel Movidius
- Qualcomm Snapdragon with integrated NPUs

Software

- TensorFlow Lite
- ONNX Runtime
- PyTorch Mobile
- Edge Impulse

The Architectural Landscape for Edge AI

CPU Only

General-purpose processing for low-complexity tasks. Flexible but limited in AI throughput and efficiency.

CPU + GPU

Leverages GPU's parallel floating-point capabilities for higher throughput, especially in neural network inference.

CPU + FPGA

Reconfigurable hardware offers custom acceleration, optimal power efficiency, and adaptable data paths for evolving AI models.

CPU + ASIC (NPU/TPU)

Dedicated hardware for specific AI tasks, providing the highest performance and efficiency but with limited flexibility.

CPU + GPU + NPU + DSP

Highly heterogeneous systems combine various accelerators for maximum performance and versatility across diverse AI workloads.

Why and When to Use CPU–FPGA Architectures?

Why:

- Hardware-level **parallelism** and **low latency** processing on FPGA
- **Flexibility** for custom AI layer acceleration
- **Energy efficiency** for embedded / edge environments

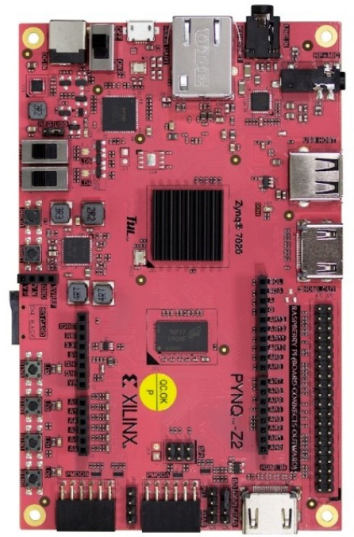
When:

- When **real-time constraints** are critical
- When workloads are **compute-intensive but repetitive** (e.g., CNN layers)
- When **power and size** limitations exclude large GPUs
- When system needs **deterministic timing** and **hardware adaptability**
- When implementing **non-standard or custom AI models —** even experimental architectures in research or industry prototypes

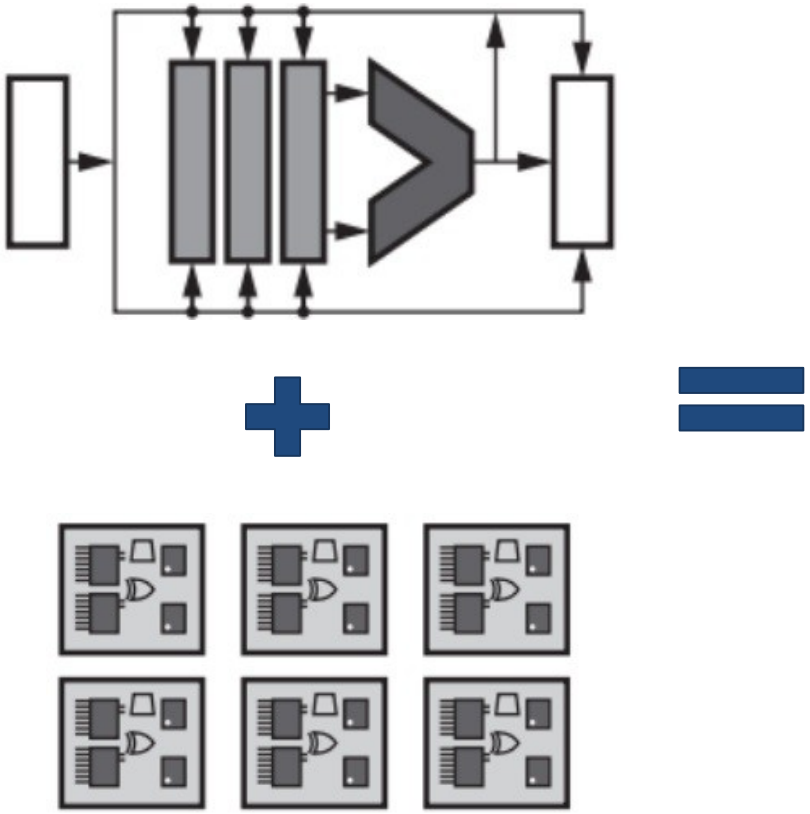
CPU + FPGA Architecture

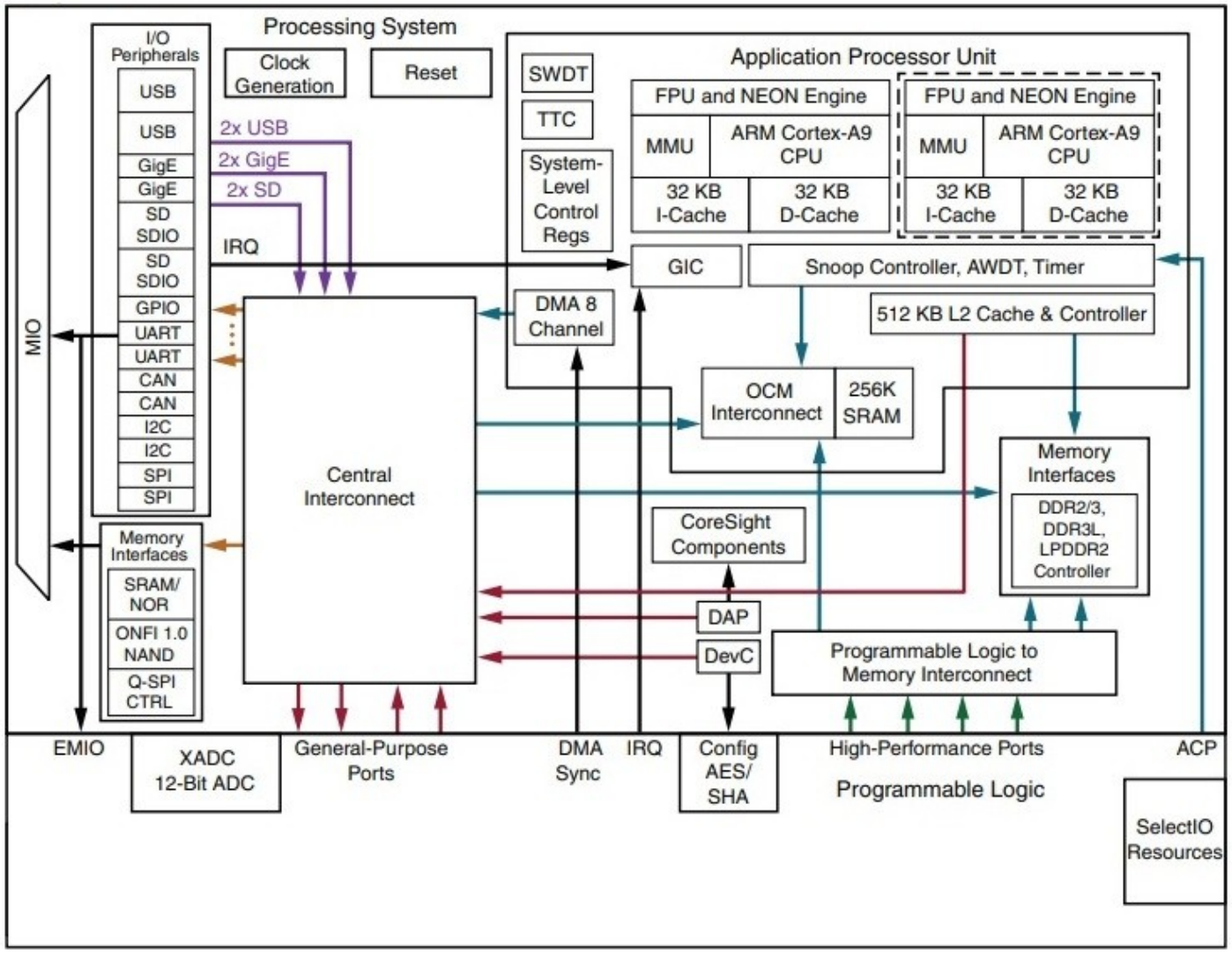
- CPU handles control flow and communication.
- FPGA accelerates heavy AI operations like convolution and matrix multiplication.
- Partial reconfiguration enables real-time hardware updates.

Example: PYNQ-Z2 board used as the experimental platform.



CPU + FPGA Architecture





Typical AI Applications

Real-Time Computer Vision

- Object detection and tracking
- Facial recognition
- Industrial quality inspection

Industrial IoT (IIoT)

- Predictive maintenance
- Process optimization with sensor fusion
- Fault detection in manufacturing lines

Smart Cameras & Surveillance

- Real-time video analytics
- Anomaly detection in security footage
- License plate recognition

Energy Management Systems

- Smart grid monitoring
- Real-time energy consumption optimization
- Anomaly detection in energy distribution

Autonomous Systems

- Drones navigation and obstacle avoidance
- Robotics motion control
- Autonomous vehicles perception pipelines

Medical Imaging & Diagnostics

- Ultrasound and MRI real-time processing
- Portable diagnostic devices
- AI-based anomaly detection in medical data

Edge NLP & Speech Processing

- Voice command recognition
- On-device speech-to-text
- Natural language understanding for embedded systems

Environmental Monitoring

- Precision agriculture analytics
- Weather pattern analysis
- Pollution detection and tracking

Case Study: CNN on CPU+FPGA

- **Objective:** Analyze implementation of Convolutional Neural Networks on a hybrid CPU+FPGA architecture to optimize latency and energy efficiency.
- **Chosen CNN Model:** VGG-16 / ResNet-18
- **Platform / Hardware:** Xilinx Zynq SoC (ARM Cortex-A9 + FPGA fabric)
- **Architecture:** CPU for control and sequential tasks, FPGA for parallel execution of convolution and pooling layers.
- **Key Optimizations:**
  - Pipelined architecture, Fixed-point arithmetic, Memory reuse and minimized data transfers
- **Results & Benefits:**
  - Classification with a peak performance of 169 Gops/s
  - Energy efficiency of 17 Gops/W
  - Frame rate of 5.5 fps on the end-to-end execution of VGG-16, 6.6 fps on ResNet-18.

Reference: NEURAghe (arxiv.org/abs/1712.00994)

Advantages

- ✓ Customizable for specific AI models.
- ✓ Energy-efficient compared to GPUs.
- ✓ Reduced latency for real-time tasks.
- ✓ Upgradable through partial reconfiguration.

vs

Challenges and Limitations

Complex FPGA development process (HDL,Vitis...)

Limited FPGA resources (DSP blocks, BRAM).

Scalability constrained by platform size.

Lack of AI standardization for FPGA workloads.

Comparative Analysis Summary of Edge AI Solutions

Architecture Type	Strengths	Weaknesses	Typical Applications
CPU-only	Easy to program, flexible, low cost	Low parallelism, high latency for AI workloads	Simple IoT devices, control systems
GPU-accelerated CPU	High throughput for parallel workloads, strong support for deep learning frameworks	High power consumption, larger size, higher cost	Autonomous vehicles, robotics, complex vision tasks
CPU + FPGA	Low latency, high energy efficiency, customizable hardware acceleration	Complex design, longer development time	Real-time vision, industrial IoT, autonomous drones
CPU + ASIC (NPU/TPU)	Ultra-high efficiency, optimized for specific AI tasks	Very low flexibility, high upfront cost	Mobile AI, wearable devices, large-scale inference
Standalone FPGA	Highly customizable, can optimize for specific models	Higher complexity, less flexible for diverse workloads	Specialized hardware accelerators, real-time edge computing
Heterogeneous SoC (CPU+GPU+NPU)	Best overall performance, adaptability to diverse workloads	Power-hungry, high cost, complexity in design	Autonomous systems, smart cameras, advanced robotics
TinyML / Microcontroller AI	Ultra-low power, compact, cheap	Limited model size and complexity	Wearables, environmental sensors, ultra-low power IoT devices

Conclusion and Future Work

- Hybrid CPU–FPGA systems offer sustainable AI acceleration.
- Enable dynamic adaptation and efficient computation.
- Future work:
  - Larger FPGA platforms for higher scalability.
  - Integration with RISC-V and AI-optimized cores.
  - Comparative analysis with GPU/ASIC solutions.