Running AI Directly on an FPGA

Turning a Machine-Learning Model into Specialized Hardware

Can Internet Traffic Reveal the Website?

Data is transmitted across the internet in small chunks called packets. While most internet traffic is encrypted, certain information can still be analyzed such as:

- Packet Size
- Packet Order
- Packet Arrival Time
- Packet Frequency

- Even when traffic is encrypted, the pattern of packet sizes and timing can reveal which website someone is visiting.
- This is a real privacy and cybersecurity problem as it means encryption doesn't guarantee anonymity.

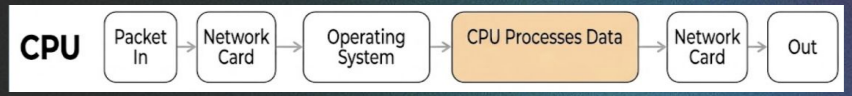
Our Model: WFNet

An AI model can study these patterns to guess which website was visited. The model used here is called WFNet.

A Processor You Can Rebuild

An FPGA (Field Programmable Gate Array) is a computer chip whose internal hardware can be modified after it is manufactured.

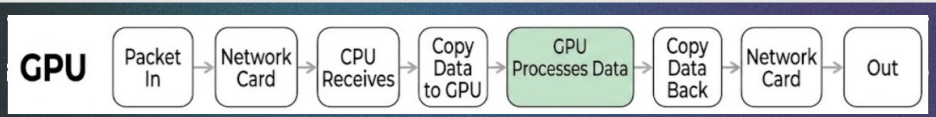
CPU — Motorcycle



- Carries one person at a time
- Moves very quickly
- Repeats the trip for every person

FPGA — Bus

- Carries many people at once
- Each individual trip may be slower
- Moves much more work in parallel



AMD Alveo U200 FPGA

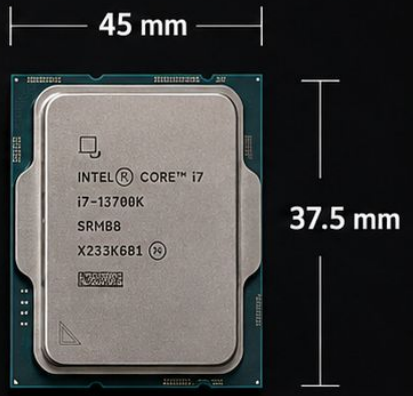
Full Length, Full Height

312 mm x 111 mm (x 1.5” profile)

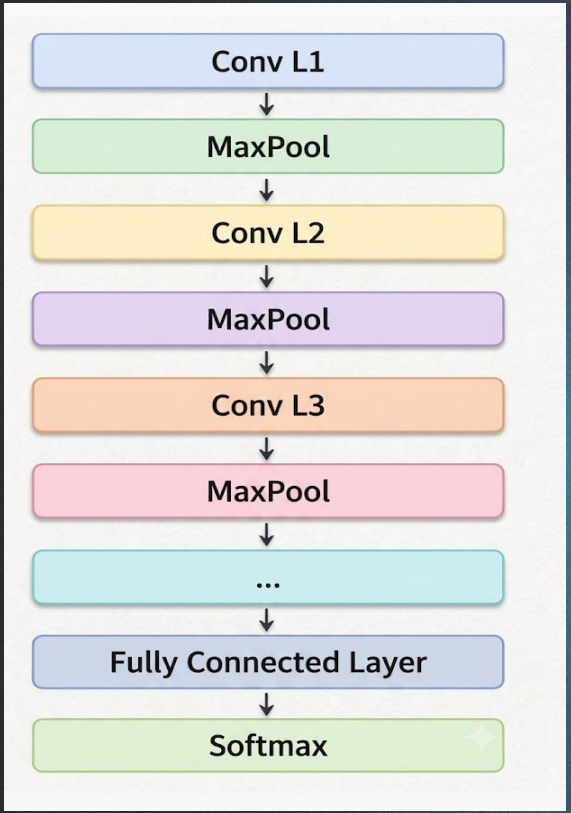
Intel Core i7-13700K CPU

LGA1700 Package

45 mm x 37.5 mm



WFNET Design

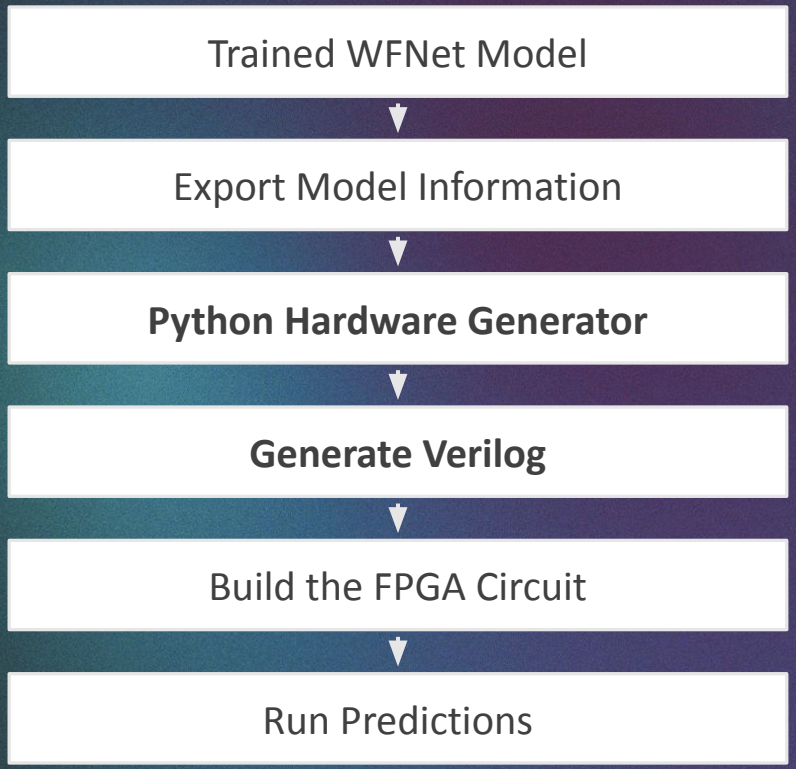


- **Convolution:**
  Multiplies and adds numbers in specific ways to spot patterns in the data

- **Max Pooling**
  Keeps only the strongest signal, drops the rest.

- Repeating these two simple steps thousands of times is what lets the model recognize complex patterns.

From an AI Model to a Circuit



How Many Workers Can the FPGA Handle?

Each hardware worker completes part of the AI calculation. More workers let more calculations happen at the same time.

128 parallel workers logic required 2.6 million lines of code

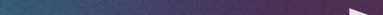
More Workers
Faster processing

The Tradeoff
More hardware, more wiring, and a harder design to build

Our Results:

Before Parallelization

55 Minutes



After Parallelization

20 Seconds

Less than 1% in processing time

Achieved on the same FPGA hardware, by redesigning the model to perform more work in parallel, while preserving prediction accuracy. This is a reduction in required FPGA clock cycles and time in simulation and testing.

The model performed the same task using a fraction of the original computation time.

What We Learned

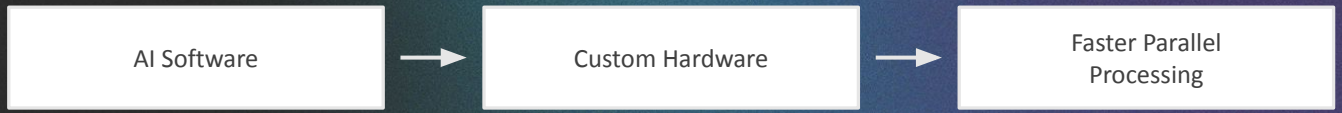
01 AI models can be transformed into hardware.

02 FPGA design is very different from software development. Each design needs significantly more attention to detail to ensure seamless functionality.

03 Optimization matters more than simply generating more code.

04 Small architectural changes can create enormous performance improvements.

This project showed that machine learning can successfully run directly on programmable hardware.



Questions?

Please feel free to visit our booth as well!

What Were We Trying to Do?

AI usually runs as software on a GPU. This project turned an AI model into a physical circuit that runs directly on an FPGA (programmable chip).