Hastlayer - Implementing on Xilinx Alveo Accelerator Cards

Zoltán Lehóczky @ Lombiq
Ernő Dávid @ Wigner

GPU Day 2020

20.10.2020

lombiq

HASTLAYER

be the hardware

Let’s talk about you!

You’re doing some number crunching.

Number crunching like in…

- Artificial intelligence, machine learning
- Image and video processing, computer vision
- Algorithmic trading
- Data compression
- Scientific computations and physics problems

To make faster you can…

- Profile and optimize it ✓
- Parallelize it ✓
- Use faster and/or more hardware ✓

To make faster you can…

- Profile and optimize it ✓
- Parallelize it ✓
- Use faster and/or more hardware ✓

To make faster you can…

- Profile and optimize it ✓
- Parallelize it ✓
- Use faster and/or more hardware ✓

To make faster you can…

- Profile and optimize it ✓
- Parallelize it ✓
- Use faster and/or more hardware ✓

To make faster you can…

- Profile and optimize it ✓
- Parallelize it ✓
- Use faster and/or more hardware ✓
- …

To make faster you can…

- Profile and optimize it ✓
- Parallelize it ✓
- Use faster and/or more hardware ✓
- Use heterogeneous computing: GPUs, FPGAs… !

Let’s explore the last part a bit.

FPGAs?

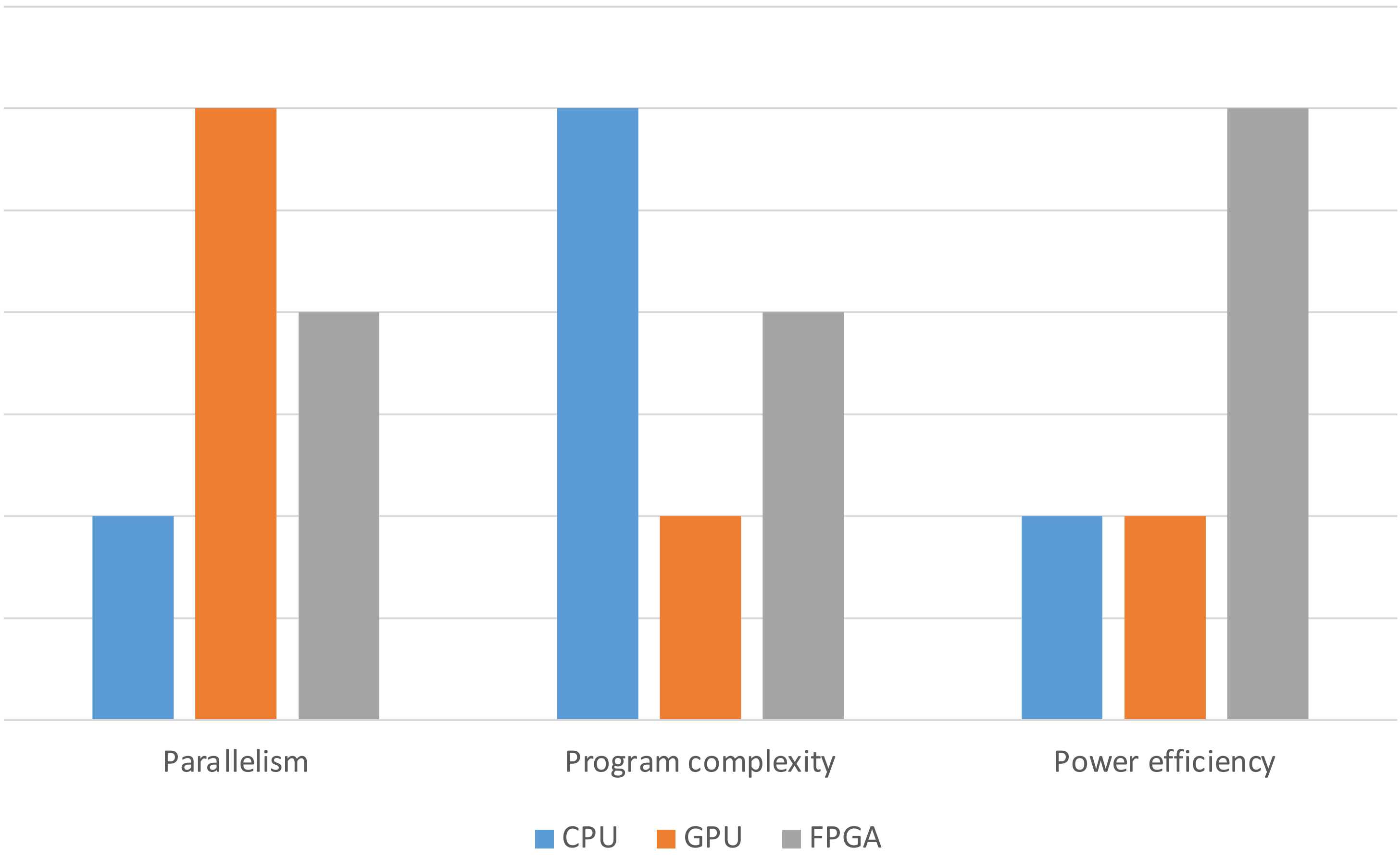
- Field-Programmable Gate Array
- Can behave like any other chip (with limitations)
- Can dynamically be „re-wired”

Image by SparkFun Electronics, Boulder, USA

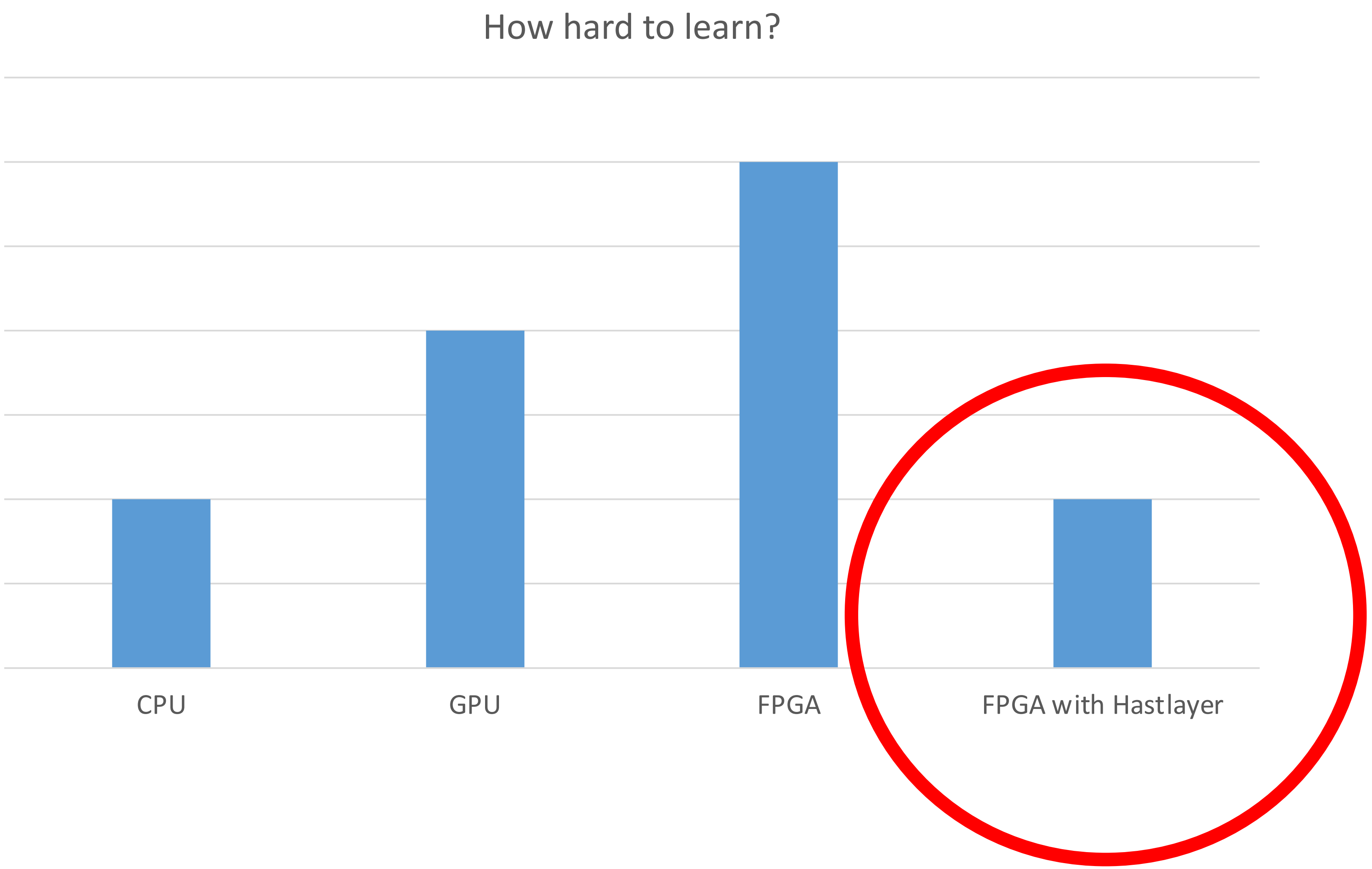
FPGAs!

- If you use Bing or Azure, you’ve used them!
- Found in routers, X-ray machines, self-driving cars…
- You need to be a hardware engineer to utilize them

CPU vs GPU vs FPGA



But!



What’s Hastlayer?

computer program → FPGA logic

computer program → computer chip

logic expressed as software → logic expressed as hardware

.NET (C#, VB, C++, F#, Python, PHP, JavaScript…) → FPGA logic

The benefits of FPGAs for us all

- Performance increase for parallel compute-bound algorithms
- Higher power efficiency
- Still only software development

Benchmarks

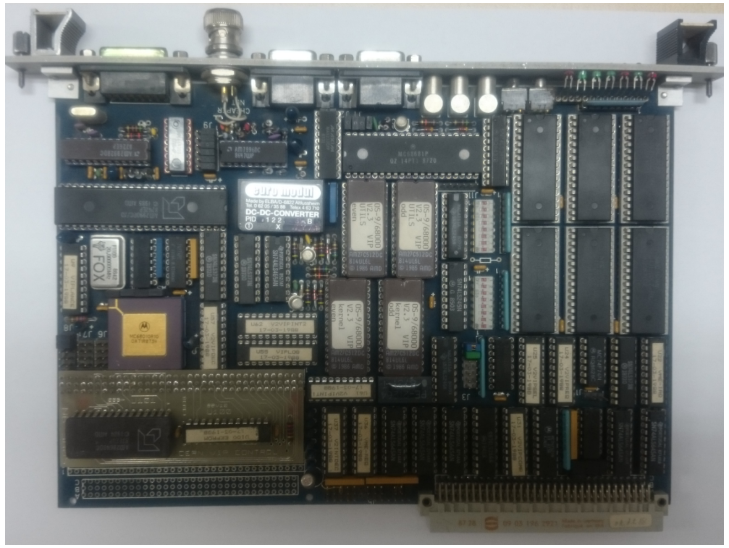
Algorithm	Nexys speed advantage	Nexys power advantage	Alveo speed advantage	Alveo power advantage
ImageContrastModifier	1% (net)	4700%	1486%	569%
MonteCarloPiEstimator	15%	5233%	189%	22%
ParallelAlgorithm	391%	23600%	421%	120%

Demo: Hands-on Hastlayer

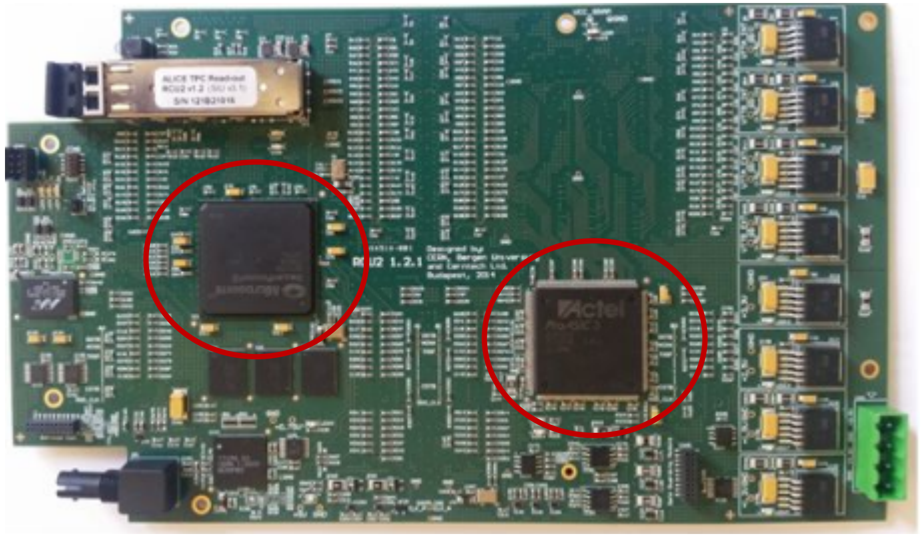
What’s under the hood?

FPGAs in the Wigner

- Wigner RC - DAQ (Data Acquisition) laboratory
- HW and FW for physical experiments (e.g. CERN ALICE)

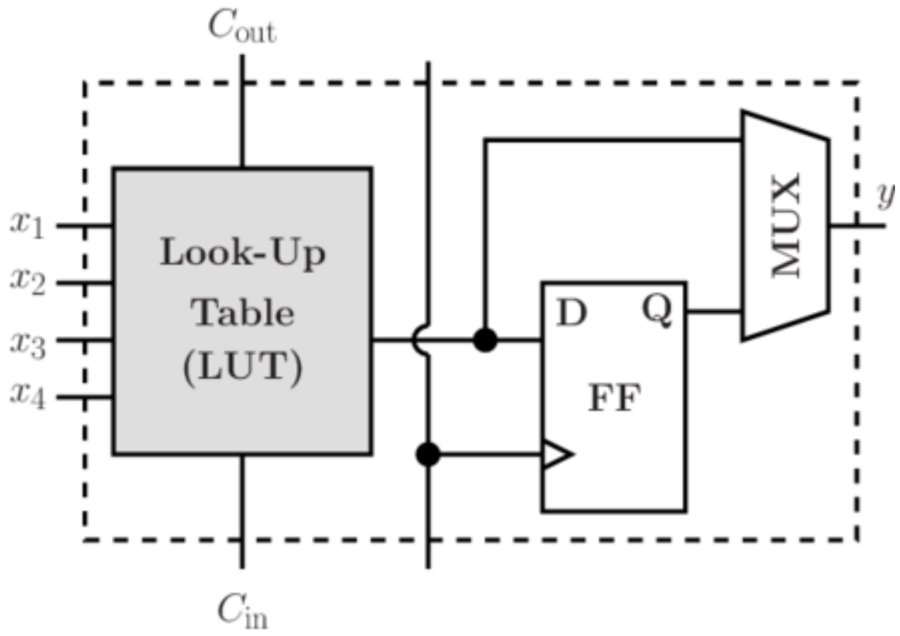


CERN 198x

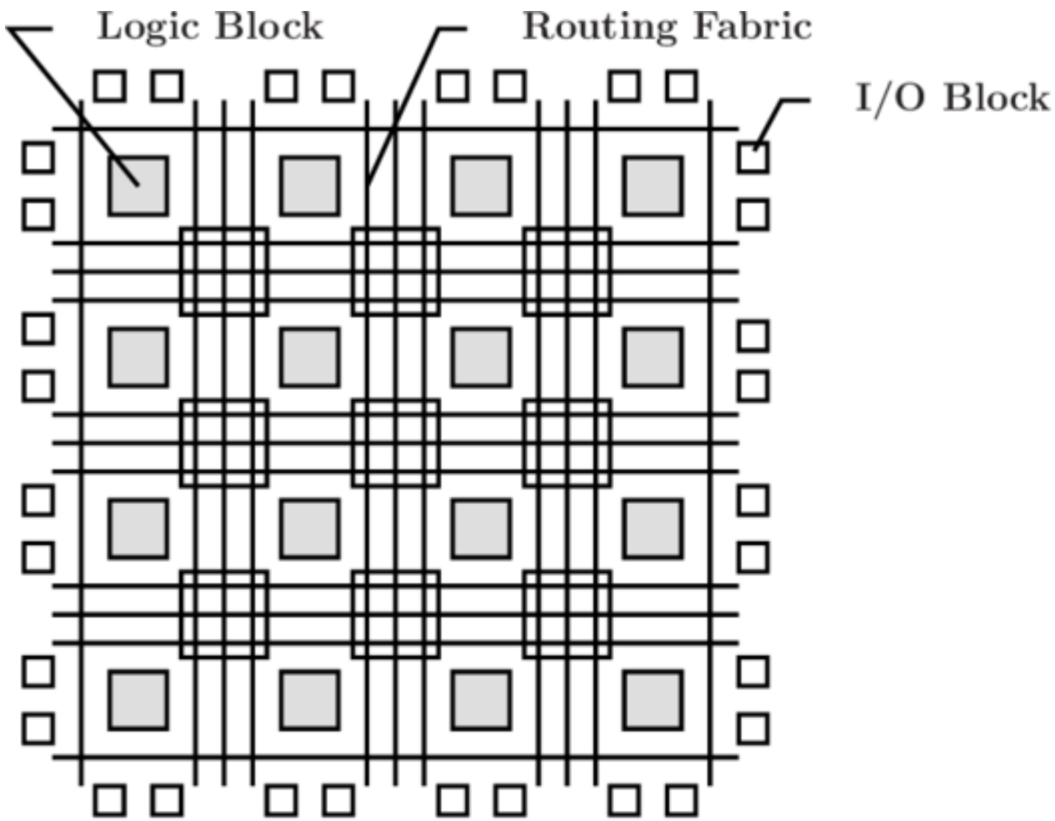


CERN 2016-2019

Inside the FPGA



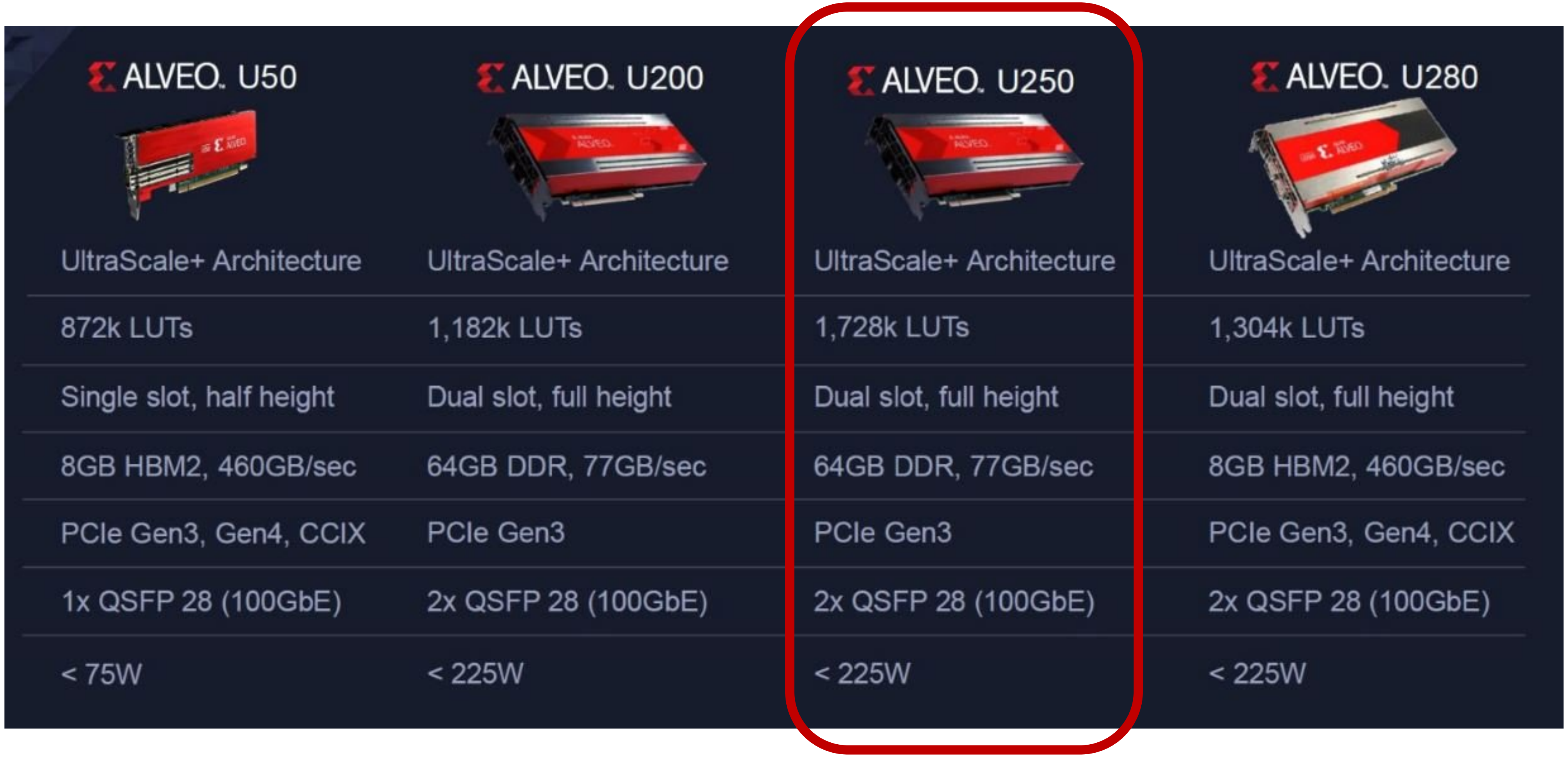
Logic Block



FPGA Fabric

Source: ResearchGate

Xilinx Alveo Product Lineup



ALVEO U50	ALVEO U200	ALVEO U250	ALVEO U280
UltraScale+ Architecture	UltraScale+ Architecture	UltraScale+ Architecture	UltraScale+ Architecture
872k LUTs	1,182k LUTs	1,728k LUTs	1,304k LUTs
Single slot, half height	Dual slot, full height	Dual slot, full height	Dual slot, full height
8GB HBM2, 460GB/sec	64GB DDR, 77GB/sec	64GB DDR, 77GB/sec	8GB HBM2, 460GB/sec
PCIe Gen3, Gen4, CCIX	PCIe Gen3	PCIe Gen3	PCIe Gen3, Gen4, CCIX
1x QSFP 28 (100GbE)	2x QSFP 28 (100GbE)	2x QSFP 28 (100GbE)	2x QSFP 28 (100GbE)
< 75W	< 225W	< 225W	< 225W

Source: Xilinx

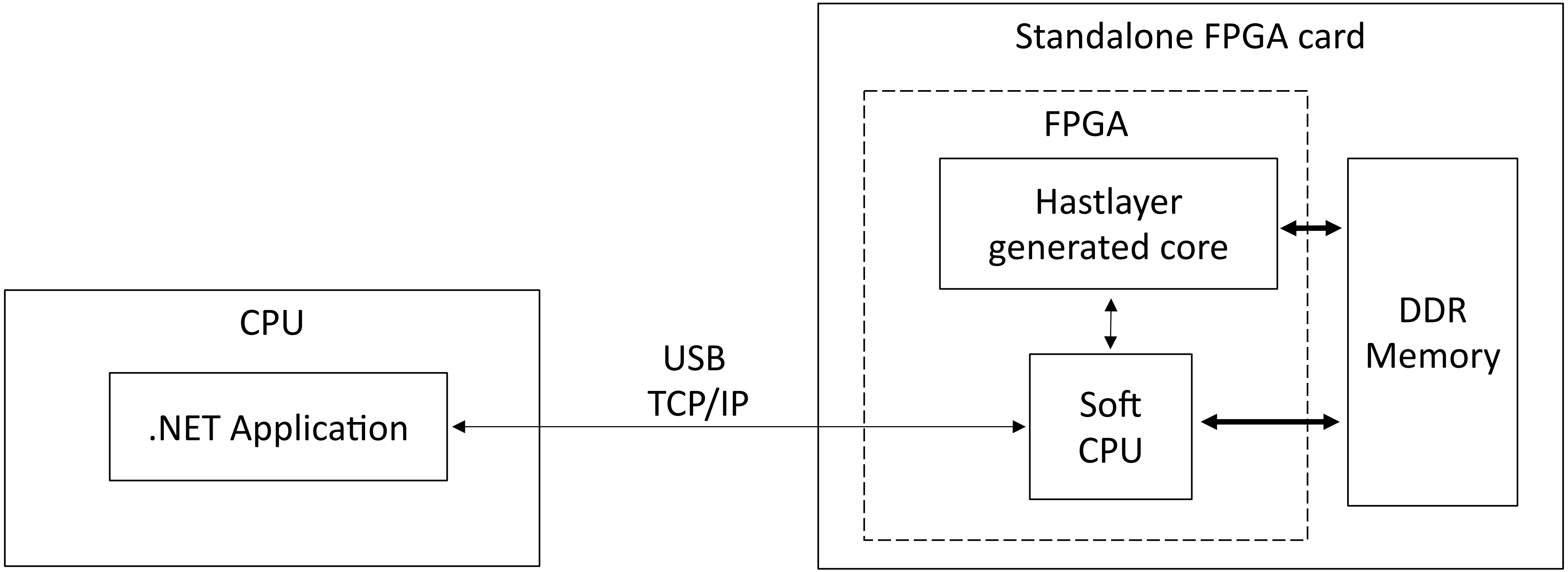
FPGA Tools Evolution

- Early days: by hand, schematic
- Now: VHDL, (System)Verilog
- Future?: High Level Synthesis
  - Xilinx / IntelFPGA OpenCL
  - Hastlayer
  - …

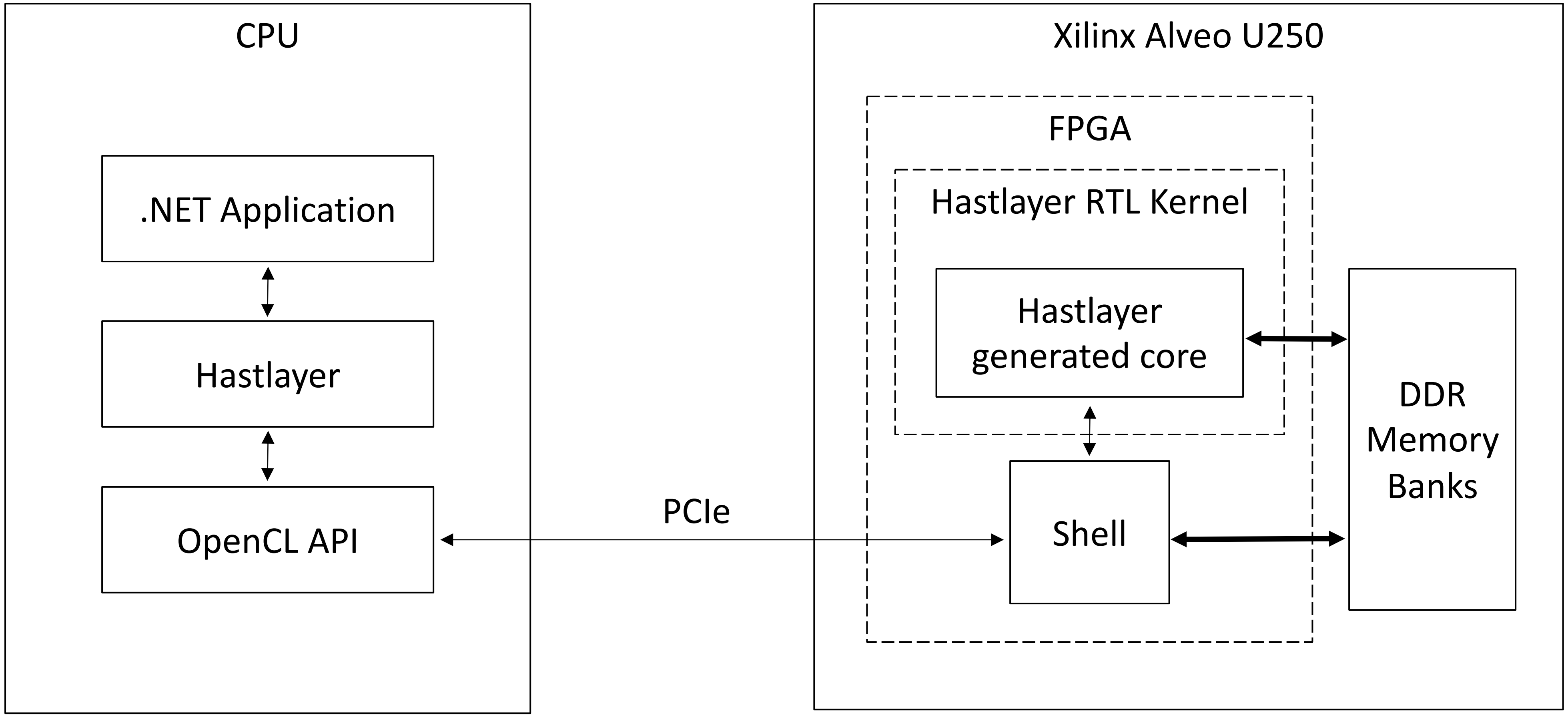
- Wigner GPU Lab and Lombiq collaboration since 2018

Let’s Build an Accelerator with Hastlayer

- Hastlayer generated core – VHDL module generated from .NET assembly by the Hastlayer SDK

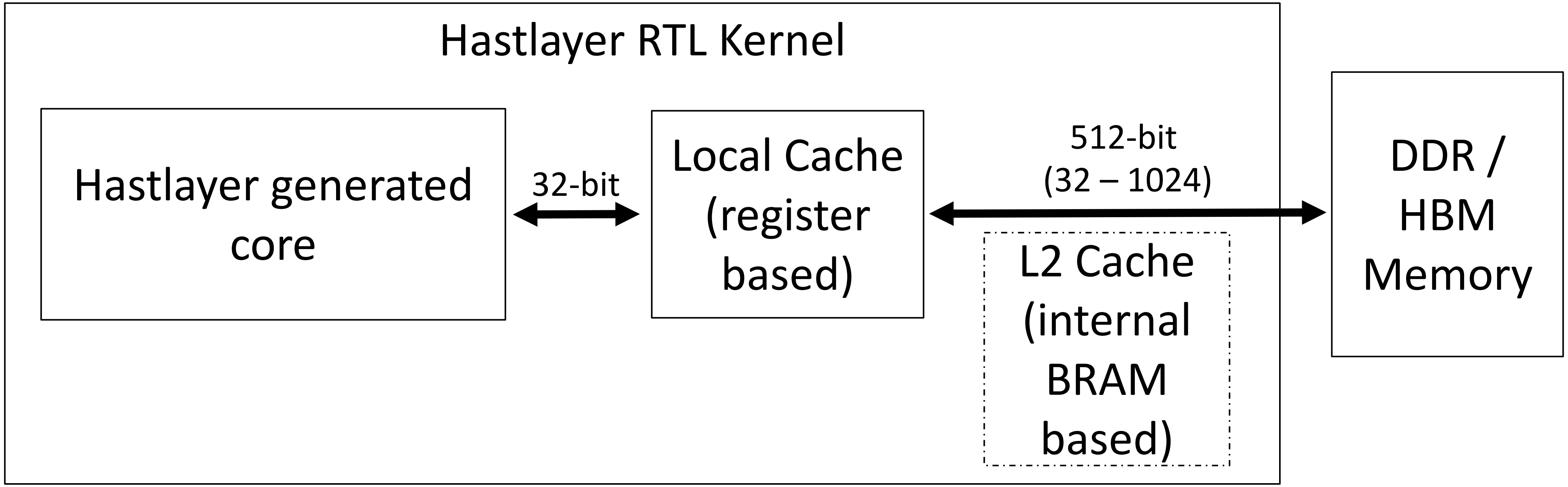


Xilinx Alveo Card Execution Model



Challenging areas – Memory Access

- Hastlayer generated core - 32-bit memory access
- Early stage: 32-bit external memory access
- Now: 512-bit external memory access + local cache



Challenging areas – Memory Access

- ImageProcessingAlgorithms sample runtimes:

AXI and Cache Width [bit]	DDR Memory [ms]	HBM Memory [ms]
32	80,4	71,5
64	44,5	40,2
512	12,6	11,9
1024	10,3	10,0

Challenging Areas – Kernel Frequency

- Higher parallelization → Bigger combinational logic → Longer propagation delay → Slower kernel frequency

- Target frequency: 300 MHz (3.33 ns between two flip-flops)

- Ongoing work: Inject more pipeline registers → Split the combinational logic to smaller parts → Higher kernel frequency

Challenging Areas – Build Time

- Not a Hastlayer issue but an Alveo card development feature
- Simplest design: 3-4 hours to build
- The whole system is rebuilt every time
- (existing FPGA configuration can be loaded immediately)

- Xilix: promises that FPGA partial reconfiguration will be supported in the future (when only the kernel is replaced)

Further plans – Embedded FPGA Support

- Support Xilinx Zynq family (FPGA fabric + ARM CPU)
- Aerospace industry, on board of drones and satellites (image processing)
- Requires AMBA AXI bus support - we already have for the Alveo system

Are you ready to *be* the hardware?

- crew@hastlayer.com
- https://hastlayer.com
- https://github.com/Lombiq/Hastlayer-SDK/
- https://lombiq.com