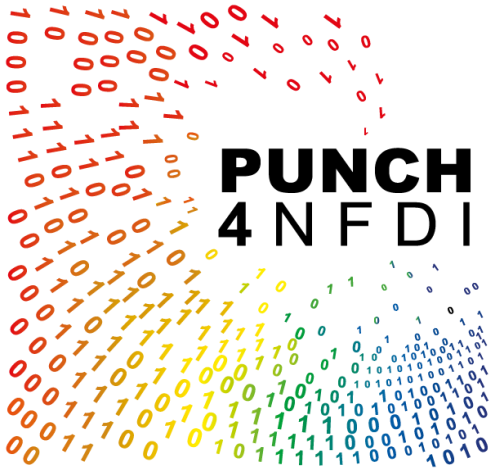


Neural Networks (or AI) on FPGAs

Arno Straessner

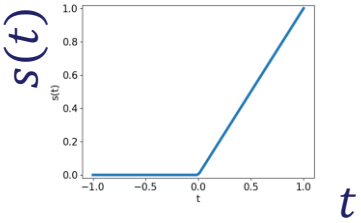
TECHNISCHE UNIVERSITÄT DRESDEN

TA5 WP2 working meeting

20.06.2022

Neural Networks on FPGAs

- Advantages of FPGAs:
  - High I/O bandwidth
  - Large number of DSPs for multiplier-adder functions:
    - fits to the classic neural network operations: linear combination of inputs, combined with activation function



$$s\left(\sum_{i=0}^{N} a_i \times x_i + b\right)$$

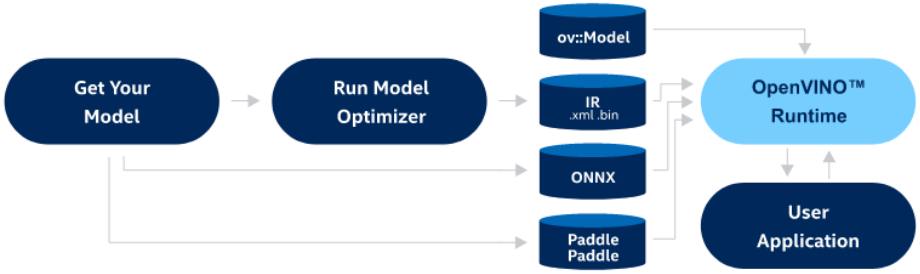
  - Parallel processing (multiplexing)
  - Short latency calculations

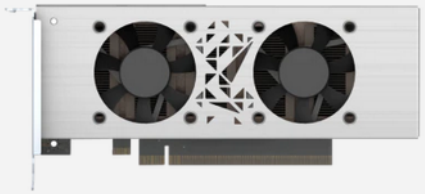
- Two directions of applications:
  - many input streams, small number of output streams: "classic" neural network structure
    - example: trigger processor
  - many input streams, many output streams: "small" neural networks
    - example: detector signal processing

- Limitations:
  - FPGA resources
  - power dissipation (100 W per FPGA) and cooling

Neural Networks on FPGAs

- Specilized AI hardware and frameworks by INTEL and Xilinx:

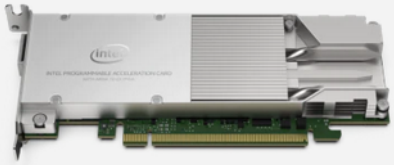




Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA

The IEI Mustang-F100-A10 is an Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA GX in a small form factor, PCI Express* (PCIe*) acceleration card that is supported natively by the OpenVINO™ toolkit to deliver low-latency video inference for edge and, cloud deployment. It has been designed to add acceleration capabilities to PCIe* host platforms and has been validated on the IEI TANK-870AI compact IPC for those with space and power constraints.

Learn more →



Intel® Programmable Acceleration Card with Intel® Arria® 10 GX FPGA (Intel® PAC with Intel® Arria® 10 GX FPGA)

Intel® FPGA-based acceleration platforms include PCIe*-based programmable acceleration cards, socket-based server platforms with integrated FPGAs, and others that are supported by the Acceleration Stack for Intel® Xeon® CPU with FPGAs. Intel platforms are qualified and validated for several leading original equipment manufacturers (OEM) server providers to support large scale FPGA deployment.

Learn more →

Intel® Stratix® 10 NX FPGA

Intel's first AI-optimized FPGA for high-bandwidth, low-latency AI acceleration for applications such as natural language processing and fraud detection.

Learn more →

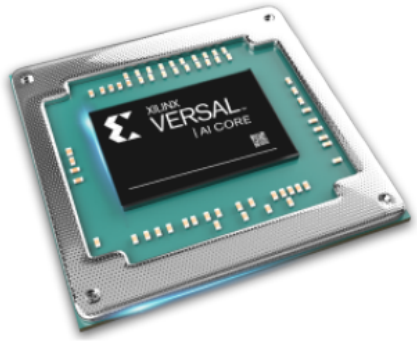
Neural Networks on FPGAs

- Specilized AI hardware and frameworks by INTEL and Xilinx:

Available in Versal Portfolio

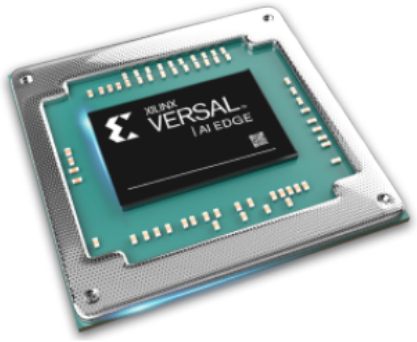
AI Engine and AI Engine-ML architectures are available in both Versal AI Core and Versal AI Edge devices.

Versal AI Core Series



Versal AI Core series delivers breakthrough AI inference and wireless acceleration with AI Engines that deliver over 100X greater compute performance than today's server-class CPUs. Featuring the highest compute in the Versal portfolio, applications for Versal AI Core ACAPs include data center compute, wireless beamforming, video and image processing, and wireless test equipment.

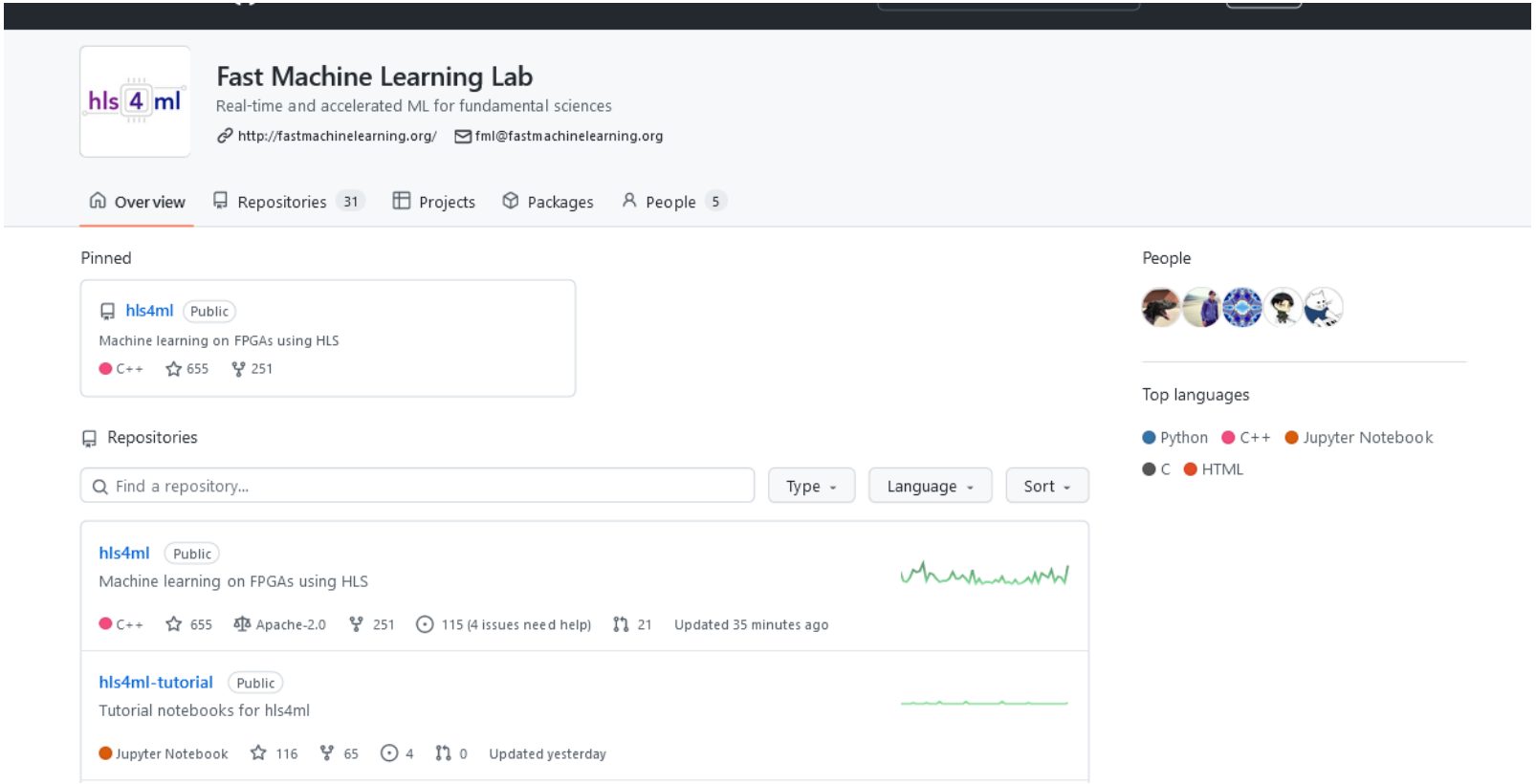
Versal AI Edge Series



Versal AI Edge series delivers 4X AI performance/watt vs. leading GPUs for power and thermally constrained environments at edge nodes. Accelerating the whole application from sensor to AI to real-time control, the Versal AI Edge series offers the world's most scalable portfolio in its class, from intelligent sensor to edge compute, along with hardware adaptability to evolve with AI innovations in real-time systems.

Neural Networks on FPGAs

- Toolkits
  - for example:

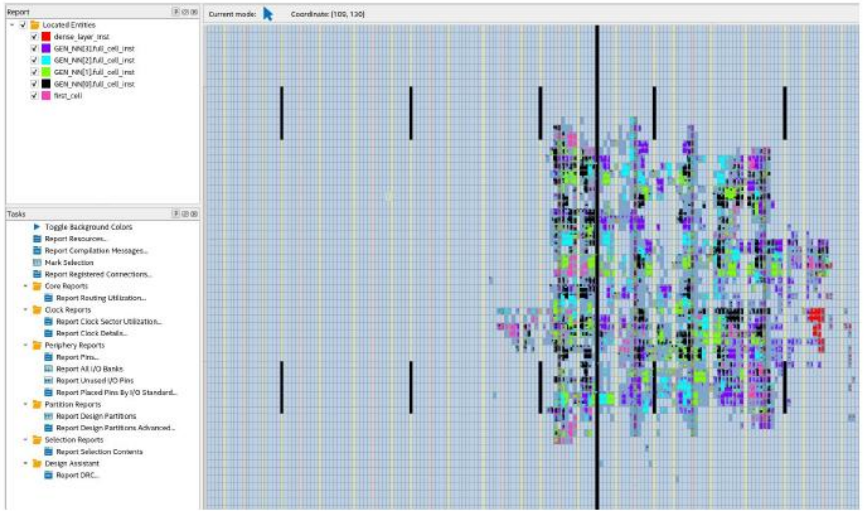


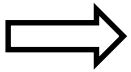
  - missing featurer are added by users, who become developers

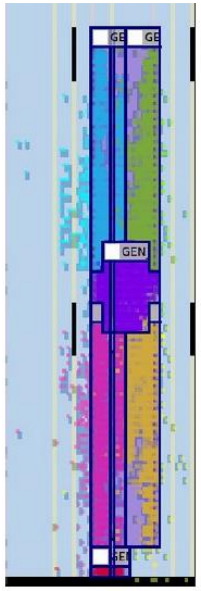
- but often times, "hand-made" conversion of ANNs to firmware
  - goal, for example: high execution frequencies to allow time-multiplexed processing

Neural Networks on FPGAs

- Example of HLS / human "co-training" for the "end-game":
  - placement of ANN cells







structured ANN cell placement reaches higher execution frequencies

  - usage of registers "learned" from HLS, and transferred to VHDL (in this example, HLS was too generous with logic resources)

	ALM	DSP	M20K	Max Frequency
1 NN (Vanilla 5 cells 8 dimensions) in HLS	2,6 %	2,6 %	0,3 %	483 MHz
1 NN (Vanilla 5 cells 8 dimensions) in VHDL	0,7 %	2,4 %	0,6 %	587 MHz

- Full implementation of the NN is better in VHDL than HLS

Summary

- High-end FPGAs allow implementation of ANN/AI

- AI on FPGA can bring more performance for real-time processing
  → need to invest $$-$$$ and power

- Interesting technology which receives much attention
  - example: many conference talks+posters on ANN on FPGAs for ATLAS LAr calorimeters
    → requires training of developers