Paderborn Center for Parallel Computing

The Need for Speed: A High-Performance Driver for FINN

Author: Linus Jungemann
E-mail: linus.jungemann@uni-paderborn.de

eki
ENERGY EFFICIENT AI BY DNN APPROXIMATION FOR FPGAs
https://eki-project.tech

MOTIVATION FOR FINN IN THE DATA CENTER

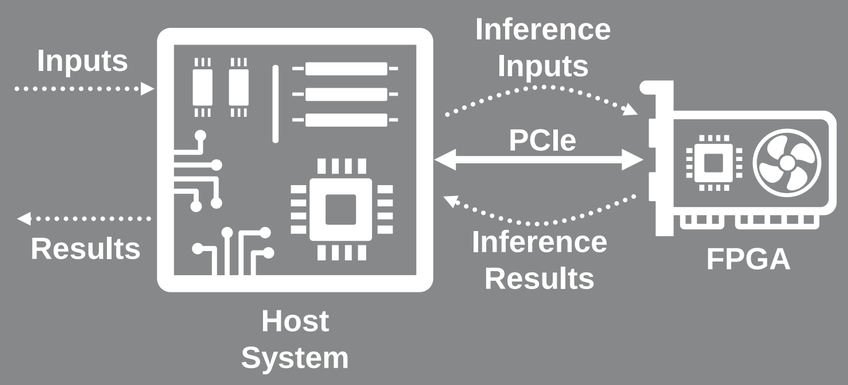
- Increasing percentage of software utilizes AI
- Data Center High-Performance applications of AI include live processing of sensor data [1], network intrusion detection, management & simulation of live systems e.g.: the power grid [2]

- Increasing importance of energy consumption and CO_2 emissions for data center operation
- Growing power consumption of neural network becomes significant factor of energy efficient data center operation

- Using FPGAs for neural network inference (using e.g.: FINN) increases energy efficiency significantly compared to GPUs

- Existing Software for integrating FPGA based neural network accelerators into applications is not suitable
  - High latency
  - Low throughput
  - Redundant and/or inefficient operations

FPGA DEPLOYMENT IN THE DATA CENTER



- FPGA usage for neural network inference in data center follows pattern
  - Data from data source (network, storage) accessed by the host system
  - Preprocessing of data on the host system (e.g.: quantization of input data, first steps of data processing)
  - Offloading of inference to FPGA: Data transferred to FPGA & inference kernel is started by host
  - Output data is transferred back & postprocessed (if necessary)
  - Outputs are sent back to storage/network or data is used for further processing

- Software designed for data center usage must support this pattern efficiently!

PROBLEM: PERFORMANCE OF DEFAULT FINN DRIVER

- Python based
- Not optimized for low latency, high throughput operation
- XRT managed kernels (high overhead kernel starts)
- Redundant stores and loads to/from FPGA memory
- Dynamically grown/extended driver leads to extremely inefficient implementation

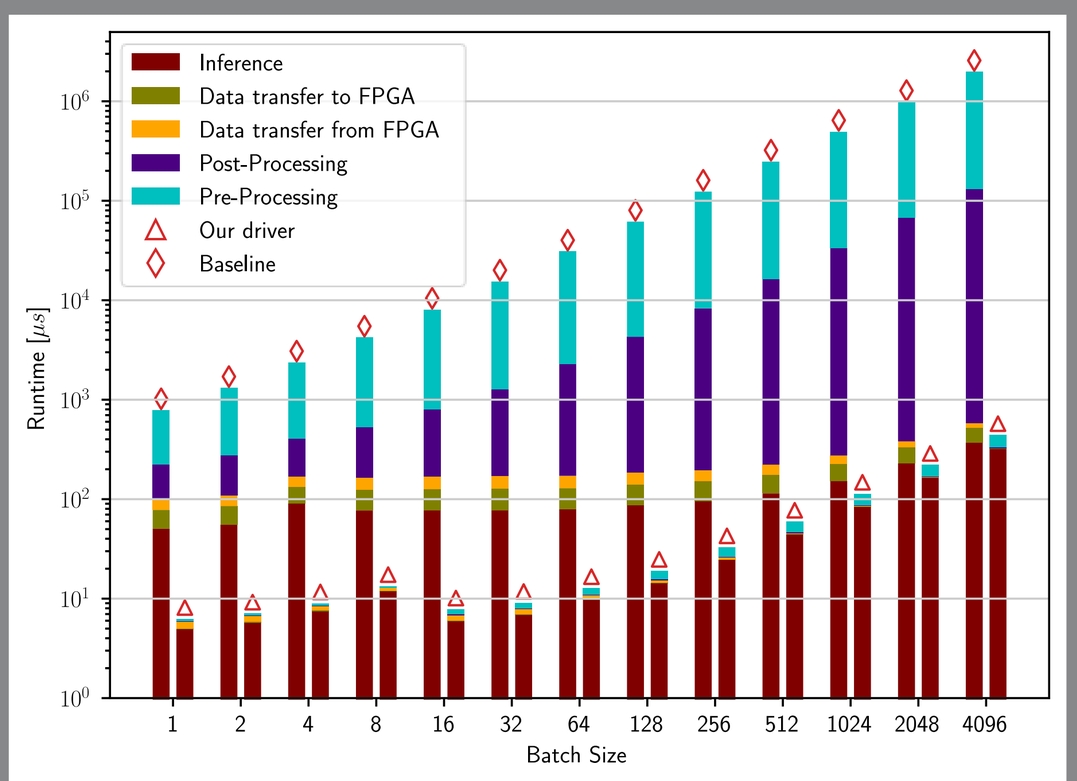
SOLUTION: HIGH-PERFORMANCE DRIVER

- C++ based
- Efficient & Hardware aware
- Multithreading support
- Low overhead user managed XRT kernels
- Host memory access to remove unnecessary FPGA memory stores and loads

BENCHMARKING SETUP

- Benchmarking performed on HACC Nodes of Noctua 2 cluster at PC2
- AMD Alveo U55C FPGA used in combination with 2x AMD Milan 7V13
- FINN v0.10.1
- Network: Particle Jet Classification Network [1], 9 Layers Fully Connected, Feed Forward, 24 8-bit Inputs per Batch, 5 16-bit Outputs per Batch

BENCHMARKING RESULTS



Logarithmic inference runtime in microseconds for the baseline FINN driver and our high-performance driver over different batch sizes. Runtime for both drivers consists of data preprocessing (e.g.: formatting and padding data correctly for FPGA input), data transfer to FPGA, inference, data transfer from FPGA and post-processing (e.g.: removing padding and formatting data into correct output format).

- ~125x to ~4483x improvement in runtime depending on input batch size
- Improvements in runtime also translate into similar improvements in throughput and latency

- Two removed stores and loads to FPGA memory per batch element
- Majority of improvements in runtime are in the Pre-/Post-Processing phase of operation
  - Efficient & multi-threaded data packing and padding contributes majority of speedup
- Inference phase of operation benefits from significantly reduced startup overhead of user managed XRT kernels

OUTLOOK

- Multi-FPGA Support
- Data Transfer Optimization
- Parallel Pre-, Postprocessing and Inference (Asynchronous driver mode)

Github Repository

REFERENCES

- [1] P. Odagiu, Z. Que, J. Duarte, J. Haller, G. Kasieczka, A. Lobanov, V. Loncar, W. Luk, J. Ngadiuba, M. Pierini, P. Rincke, A. Seksaria, S. Summers, A. Sznajder, A. Tapper, and T. K. Aarrestad, “Ultrafast jet classification on FPGAs for the HL-LHC,” vol. 5, no. 3, p. 035017. [Online]. Available: http://arxiv.org/abs/2402.01876
- [2] M. Arpogaus, M. Voss, B. Sick, M. Nigge-Uricher and O. Dürr, "Short-Term Density Forecasting of Low-Voltage Load Using Bernstein-Polynomial Normalizing Flows," in IEEE Transactions on Smart Grid, vol. 14, no. 6, pp. 4902-4911, Nov. 2023, doi: 10.1109/TSG.2023.3254890.

The eki Project is supported by:

Zukunft Umwelt Gesellschaft

Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit

PC² – Shaping the Future of HPC

PADERBORN UNIVERSITY
The University for the Information Society

https://pc2.uni-paderborn.de

Warburger Str. 100 | 33098 Paderborn | Germany