



FINN+

Towards Hassle-Free Co-Design of FPGA DNN Inference Accelerators

Fast ML Conference 2025, Zürich

Felix Jentzsch
Computer Engineering Group, Paderborn University



About Us

- FINN+ team
 - Four PhD students from Paderborn, Germany
 - Collaboration between Computer Engineering Group and Paderborn Center for Parallel Computing (PC²)
- EKI Research Project¹
 - Funded by the German Ministry for the Environment
 - Goal: Increase energy efficiency of DNN inference via custom-tailored FPGA dataflow accelerators
 - Focus on data center and energy awareness

¹ <https://www.eki-project.tech/>



HOCHSCHULE
HAMM-LIPPSTADT



EKI Project partners



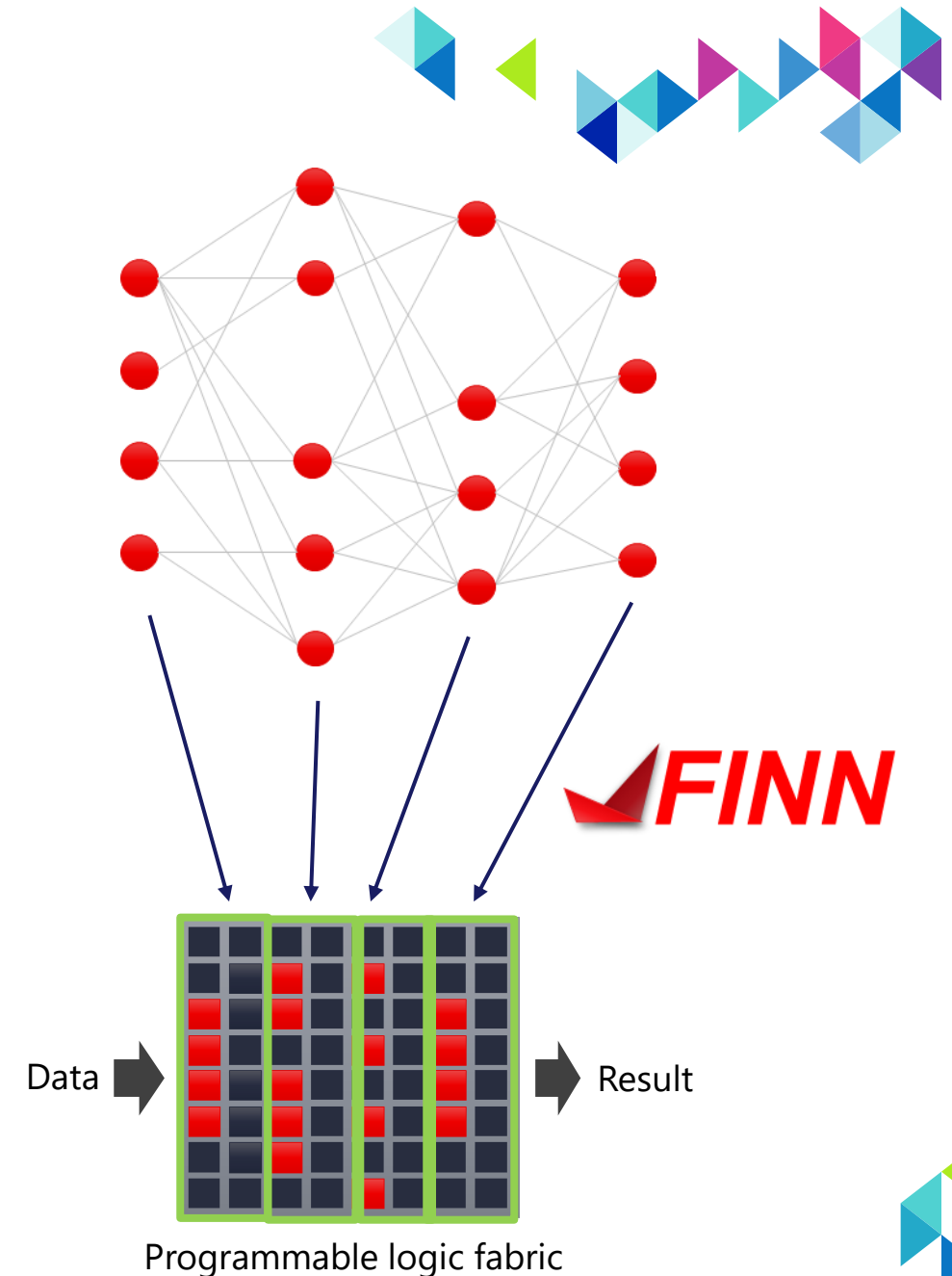
Agenda

- 1. The FINN+ Project**
- 2. DNN to FPGA Co-Design Flow**
- 3. Latest Features**
- 4. Highlight: ML-based QoR Estimation**

FINN+

- Fork¹ of AMD's FINN framework
 - Compiler for FPGA dataflow neural network accelerators (FDNAs)
 - QONNX front-end
 - Vitis/Vivado back-end using optimized HLS & RTL kernels
 - Primarily targets 1 to 4 bit integer quantization (with increasing support for 8+ bits)
- Goals
 - Quickly integrate experimental features
 - Adapt FINN to our specific needs: building and deploying accelerators on our cluster
 - Extend FINN to support our design space exploration research

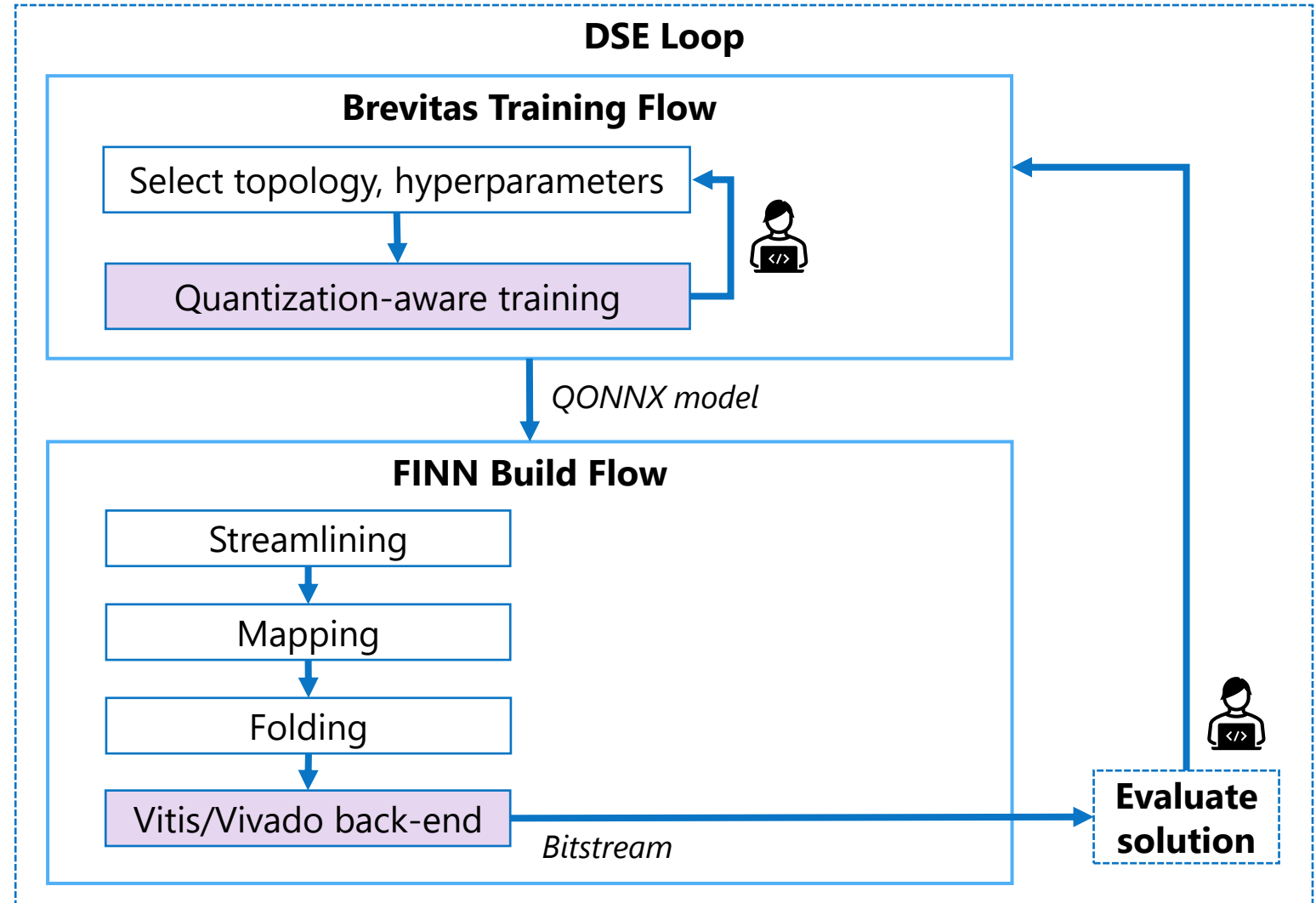
¹ <https://github.com/eki-project/finn-plus>





Co-Design Flow

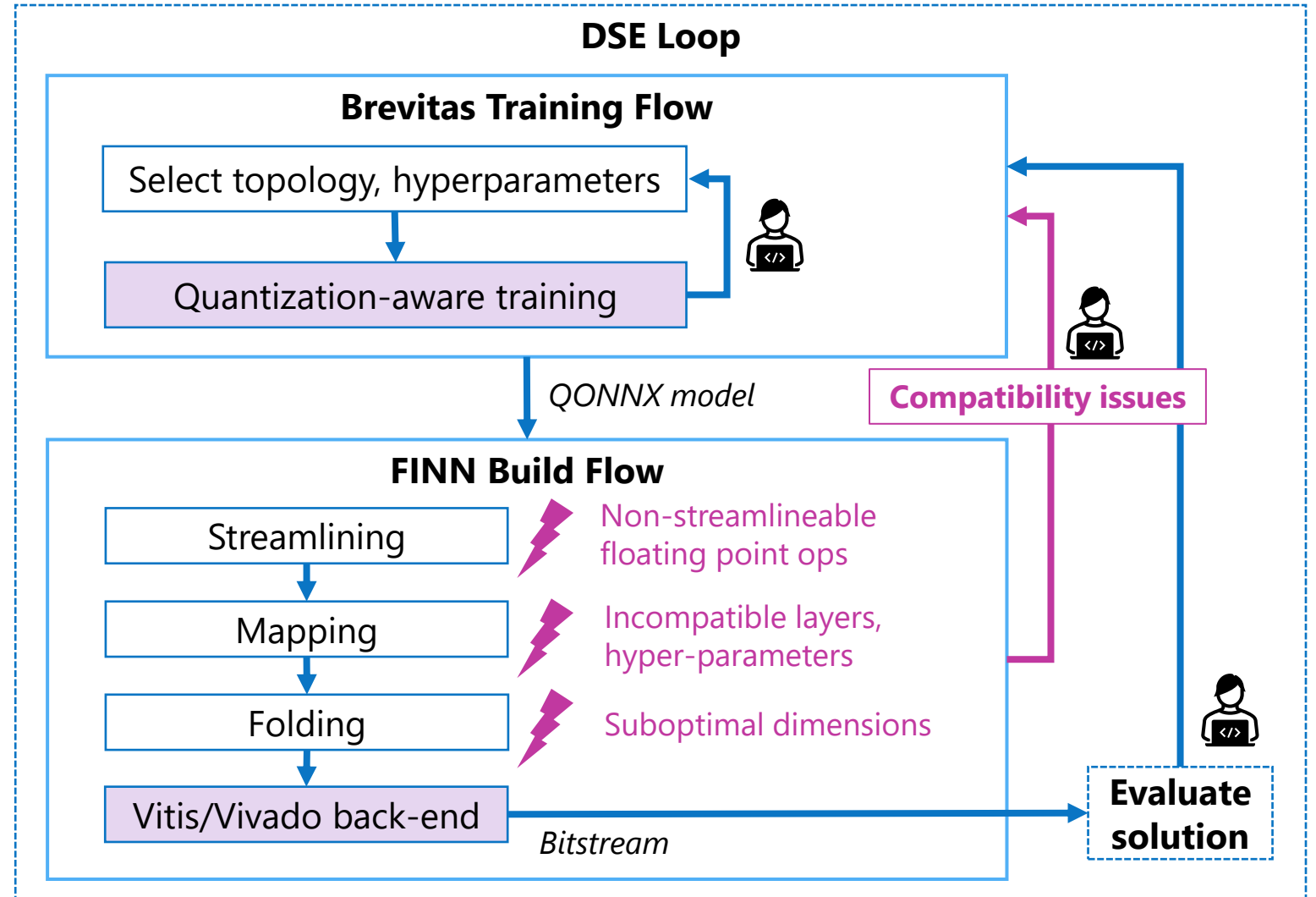
Typical manual exploration process





Co-Design Flow

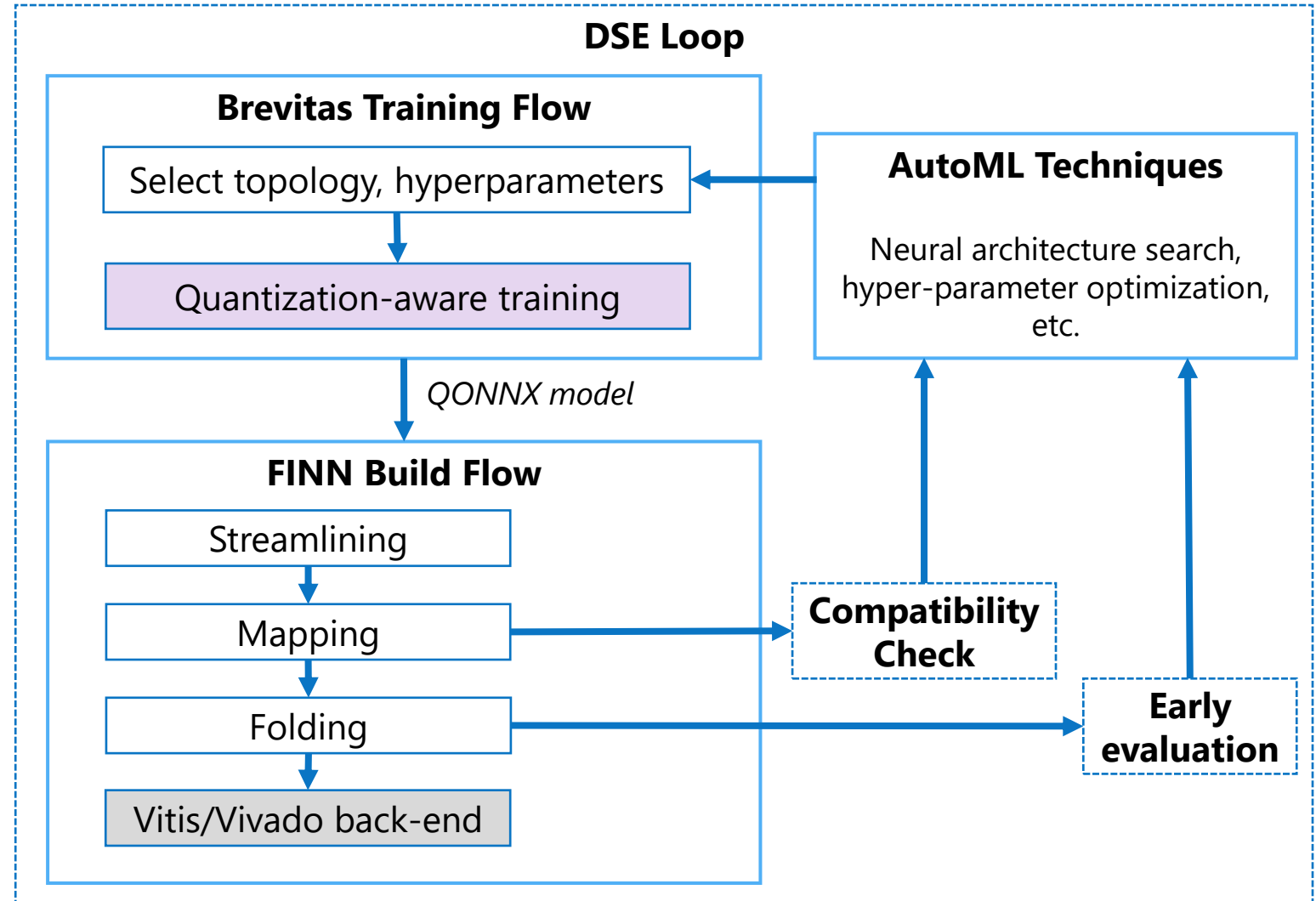
Typical manual exploration process





Co-Design Flow

Envisioned automatic exploration



What can we do to make the FINN+ design process..

.. easier?

.. faster?

.. more sustainable?





FINN+ Features

Usability

Containerless PyPI installation

Improved user interface,
stability, logging, reporting

Optimized driver for
data-center (Alveo) inference

Published at HEART 2025³

Flexibility

Additional auxiliary
operators

Transformer support

Published at FPT 2024¹

Scalability

Multi-FPGA support

Published at HEART 2025²

Hardware-accelerated
buffer sizing

Publication pending

Exploration

ML-based
Quality-of-Result (QoR)
estimation

Publication pending,
highlighted on next slides

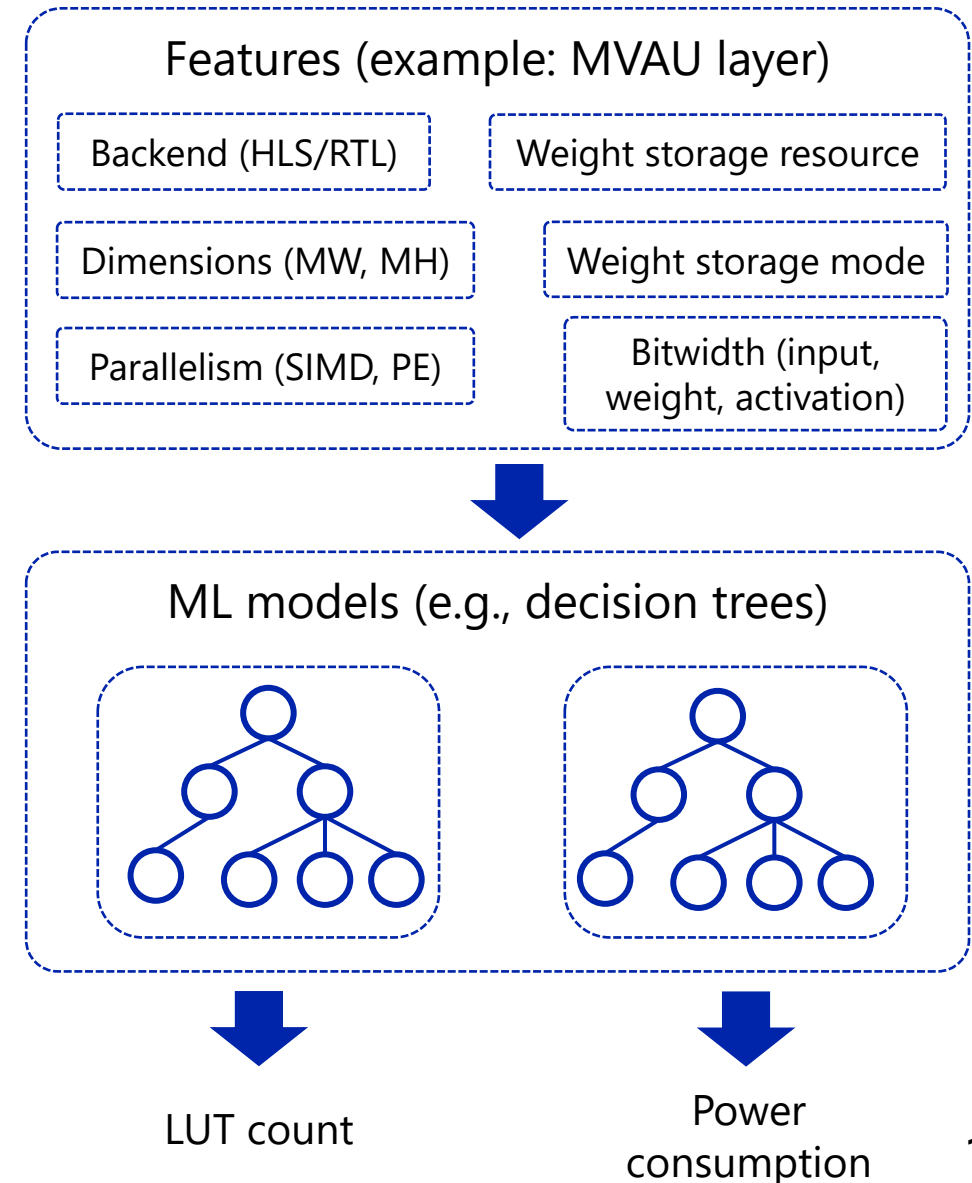
¹ C. Berganski et al., "FINN-T: Compiling Custom Dataflow Accelerators for Quantized Transformers"

² B. Wintermann et al., "AuroraFlow, an Easy-to-Use, Low-Latency FPGA Communication Solution Demonstrated on Multi-FPGA Neural Network Inference"

³ L. Jungemann et al., "FINN-HPC: Closing the Gap for Energy-Efficient Neural Network Inference on FPGAs in HPC"

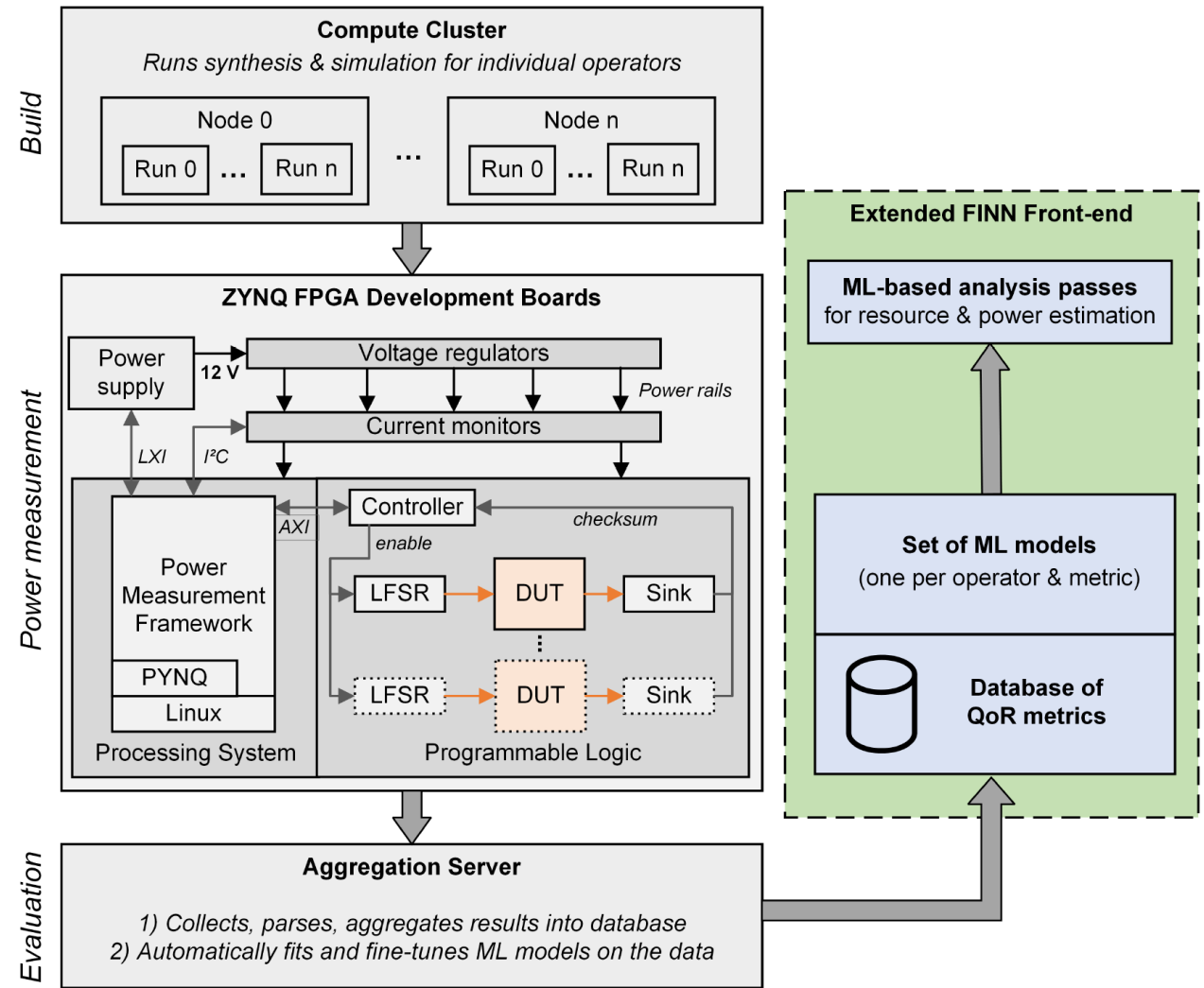
ML-Based QoR Estimation

- Early-stage estimation of Quality-of-Result (QoR) metrics is key for further automation
 - DNN accuracy
 - Throughput & latency
 - FPGA resource & power consumption
- Existing resource estimates (analytical formulas in FINN or Vitis HLS estimate) can be very inaccurate
 - Especially for highly parallelized or sparse networks!
- Our approach: ML-based resource & power prediction
 - Abstraction level: individual FINN operators
 - Enables estimation right after folding (reached in seconds)
 - No manual modeling or reverse-engineering required!



μ-Benchmarking Setup

- Automated: CI pipeline based on GitLab
- Scalable: synthesis on our cluster with up to 1000 parallel jobs
- Includes power measurement
 - Uses on-board current monitoring infrastructure of ZYNQ UltraScale+ MPSoC/RFSoc dev boards
 - Local random test pattern generation for minimal overheads
- Data & pre-fitted models will be made available alongside Git repository via DVC





Results

- Focus on MVAU layer and LUT + power metrics
- Generated initial dataset with 7828 unique MVAU configurations
 - Including 714 with induced weight sparsity
- Models evaluated:
 - K-nearest-neighbors
 - Support vector regression
 - MLPs
 - **Tree-based models/ensembles**
 - → **Best: Gradient Boosting**

Estimator	Overall (n=7828)		Sparse only (n=714)	
	RMSE	MAPE	RMSE	MAPE
FINN	25130	108%	79066	115%
Vitis HLS	15532	209%	41890	89%
Our best	855	5%	2275	9%

LUT estimation results on the μ -benchmark dataset

Metric	Estimator	Avg. Abs. Error
LUT count (overall)	FINN (analytical)	39%
	Vitis HLS	68%
	Our best	7%
LUT count (only MVAU layers)	FINN (analytical)	49%
	Vitis HLS	62%
	Our best	12%
Power	Our best	32%

Results on suite of 10 DNN model configs (TFC, CNV, KWS, VGG-10, MobileNetV1)





Conclusion

- We see design space exploration as the biggest challenge for dataflow-style accelerators
- We prepared FINN+ for large scale data center use by reducing friction and increasing flexibility/scalability
- Currently focusing on QoR estimation for faster turn-around times

- Next steps:
 - Scale up our QoR metric database
 - Explore how this QoR estimation can facilitate hardware-aware AutoML

- We see a lot of potential for collaboration with the Fast ML community
 - Exciting new applications for FINN+
 - Overlap with similar dataflow accelerator or DSE research (e.g., hls4ml)



Thank you for the attention!



Please get in touch or check out FINN+ on GitHub:
<https://github.com/eki-project/finn-plus>

Felix Jentzsch (felix.jentzsch@upb.de)



Additional Slides





LUT & Power for Highly-parallel, Sparse MVAUs

Measured LUT and power consumption of an MVAU:

- 4-bit
- 128x128 weight matrix
- different amounts of unstructured weight sparsity

The PE parallelism is swept on the X-axis while SIMD is set to maximum (128).

