Chair for Embedded Systems

Prof. Dr. J. Henkel

Master Thesis

RL-based Scheduling of Multi-DNN Workloads on Heterogeneous Accelerator Systems

Modern autonomous embedded systems run multiple deep neural networks (DNNs) in parallel to complete multiple concurrent tasks or to process and fuse information from different sensors or modalities (e.g., vision, audio, text).

Heterogeneous multi-accelerator systems enable parallel execution and can provide hardware implementations tailored to the characteristics of the DNNs and layers of the target workload. This heterogeneity can originate from different data-flow architectures, processing elements, buffer sizes, computational precision, etc.

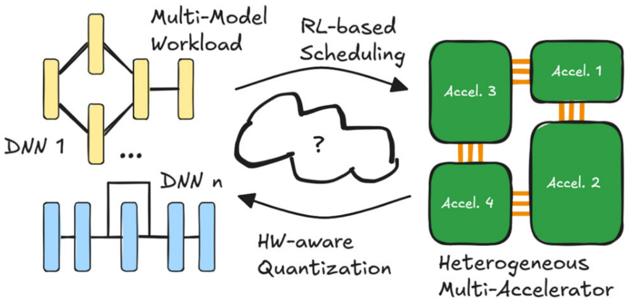
To optimally make use of the parallelism and heterogeneity, a scheduler must decide **which layer of which DNN to run on which accelerator and what time,** while simultaneously considering the implications of scheduling decisions on the task performance (such as throughput, latency and accuracy), and resource consumption (such as energy, power and temperature).

To this end, this thesis will investigate **reinforcement-learning (RL) based schedulers** that can generate near-optimal schedules under these multi-objective constraints.

Key research questions can include:
Joint Scheduling and Quantization: How can the layer-to-accelerator assignment be co-optimized together with layer quantization for accelerators that differ in architecture and precision?

Co-Optimization of Hardware Design and Scheduling: How can we formulate design space exploration to jointly optimize the hardware design of the accelerator and the multi-DNN schedule?

ML-based Performance Prediction: How can we build fast, yet accurate, surrogate models to predict latency/energy from DNN and hardware parameters and bypass costly simulators?



Potential techniques used:

- Graph learning and (multi-agent) RL, either model-free or model-based.
- Post-training quantization or quantization-aware training algorithms.
- HW modeling and simulation toolchains.

Skills required/beneficial for the thesis:

- Programming skills and curiosity for advanced ML techniques.
- Experience with RL or efficient machine learning is a plus.
- Experience with accelerator simulators such as TimeLoop/Accelergy is beneficial but not required.

Skills acquired within the thesis:

- Applying RL to complex, NP-hard, optimization problems.
- In-depth understanding of the hardware footprint of modern machine learning workloads.
- Work in a research environment.

Language:

- The collaboration with the colleagues can be in English or German.

Contact:

- Benedikt Dietrich, benedikt.dietrich@kit.edu
- Dr. Heba Khdr, heba.khdr@kit.edu

Prof. Dr. Jörg Henkel – Chair for Embedded Systems
Department of Computer Science – KIT Karlsruhe
Haid-und-Neu-Str. 7, Gebäude 07.21