Deep Neural Networks (DNNs) have emerged as the mainstream solution for various computationally demanding tasks. However, when DNNs are implemented on resource constrained devices, they are often required to satisfy latency constraints. There exist numerous techniques for incorporating on-device latency in the training procedure for a specific platform. Switching to different platforms during runtime however, introduces new latency constraints, unseen during training.

Slimmable Neural Networks offer an elegant solution to the problem above, due to their reconfigurable nature. Such networks are efficiently trained to satisfy multiple latency constraints, by scaling their channel width without computational overhead. This allows to dynamically change the width of the network at runtime, depending on the underlying resources.

In this thesis, the training of slimmable neural networks with multiple scaled widths will be explored. In addition, their implementation to multiple platforms (with different latency constraints) will be studied. The goal of the thesis will be to create a framework to train and deploy slimmable neural networks in a latency-efficient and reconfigurable manner.

Skills required for the thesis
- Programming skills (C++, Python)
- Experience with machine learning frameworks (e.g., PyTorch, TensorFlow) is beneficial but not mandatory

Skills acquired within the thesis
- In-depth experience in latency-efficient DNN training and inference
- Technical writing skills
- Work in a research environment

Language
The collaboration will be in English.

Contact
Kostas Balaskas – balaskas@kit.edu
Dr. Martin Rapp – martin.rapp@kit.edu