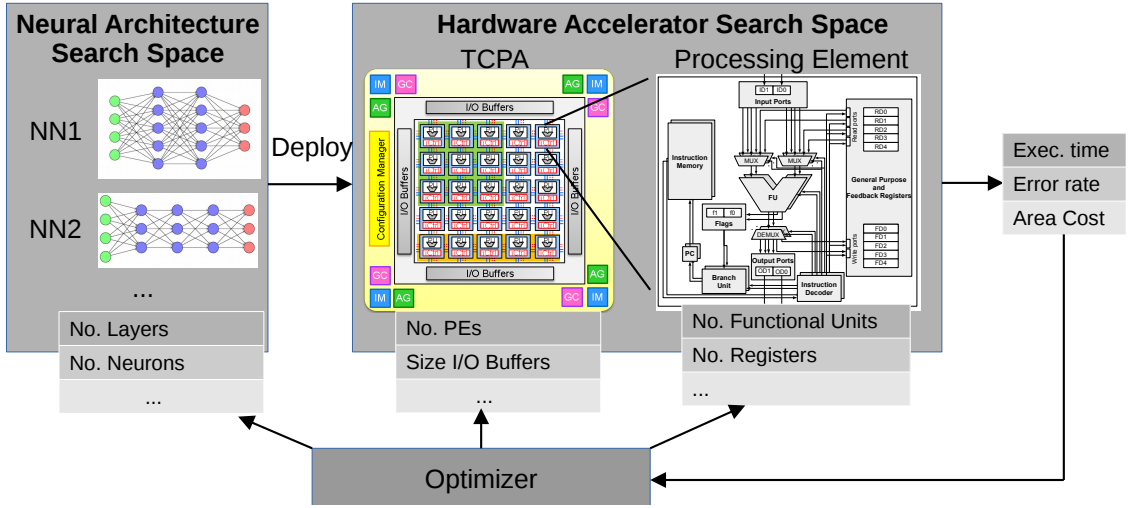
Co-Design of Neural Architecture and Hardware Accelerator

Recently, in the area of Deep Learning, platform-aware Neural Architecture Search (NAS) has become one of the major research topics with objectives such as reducing the error rate or execution time of *Deep Neural Networks* (DNNs) if deployed on an accelerator. Here, the DNNs are optimized regarding, e.g., the number of layers or number of neurons per layer.



Usually, the accelerator parameters (e.g., number of processing elements (PEs), number of functional units, ...) are considered fixed, restricting the design space for potentially faster hardware accelerator configurations. *Tightly Coupled Processor Arrays* (TCPAs) are templates that can be parameterized at design time. They consist of a reconfigurable 2D grid of processors and enable a co-design of Neural Architecture and Hardware Accelerator. In this work, existing NAS benchmarks (e.g., NATS-Bench) and a TCPA compilation environment shall be utilized for setting up a co-exploration tool. Finally, the improvements compared to a fixed accelerator shall be evaluated in terms of execution time, DNN's error rate, and area cost of the respective TCPA.

Prerequisites:	Knowledge about DNNs, Basic programming skills Python, C++, and Java
Type of work:	Theory (30%), conception (25%), implementation (45%)
Supervisor:	Christian Heidorn (Christian.Heidorn@fau.de)

Lehrstuhl für Informatik 12
Hardware-Software-Co-Design
Cauerstraße 11
91058 Erlangen

Friedrich-Alexander-Universität
Technische Fakultät