Karlsruhe Institute of Technology

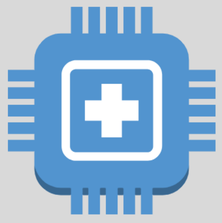
Chair for Embedded Systems

Approximate Tiny Machine Learning on Lightweight FPGAs

Georgios Mentzos, Prof. Jörg Henkel

Introduction

- **Why Tiny Machine Learning ?**
  - HW & SW capable of on-device, near sensor analytics
  - Always-on & Battery Powered Inference
  - Private & Secure
  - Near Instant Response & Independent from Network
  - Accessible & Low Cost



Medical Devices

Smart Home

- **Approximate Computing**
  - Trades Accuracy for Performance
  - Application Across the Computation Stack

Anomaly Detection

Motivation

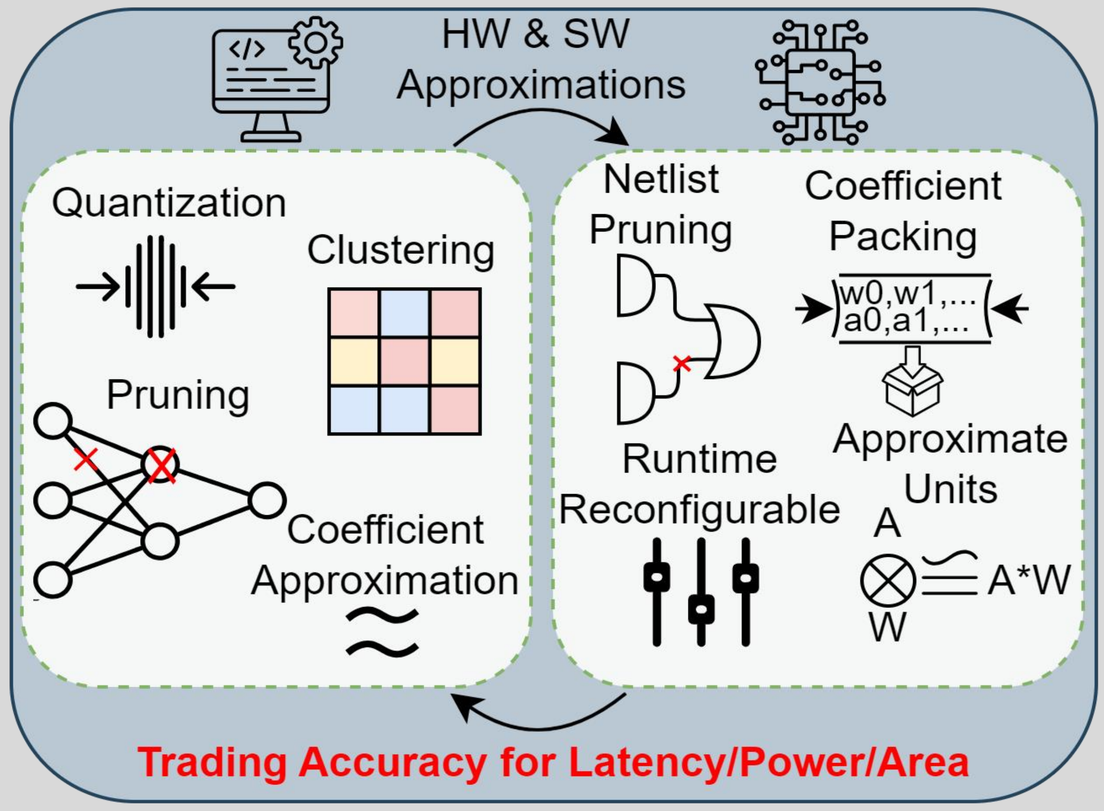
- **Requirements Imposed by EdgeAI**
  - Energy Efficiency
  - Low Latency
  - Acceptable Accuracy

- **Approximate Computing & TinyML**
  - ML tasks inherently error resilient
  - A **Perfect Match** to fit even larger DNNs on even smaller devices
  - Utilize **Flexibility** of FPGAs to fully exploit approximation benefits

Approaches for Tiny Machine Learning on FPGAs

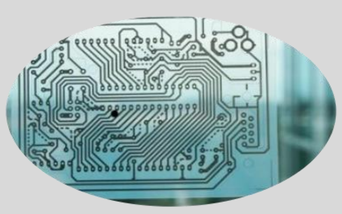
- **DNN Specific Optimizations during Training**
  - Optimizing for sparsity using **Pruning** during
  - Explorations of variable Bit-Widths using **Quantization**
  - Hardware Aware DNN Compression

- **Hardware Approximations**
  - FPGA based **approximate arithmetic** components
    - LUT optimized Approximate Multipliers
    - Approximate Adders



- **DNN Mapping Methods to Accelerators**
  - Different Parallelization Optimization Strategies
    - Folded & Fully Parallel Architectures
  - Diverse & Heterogeneous
  - Automated Frameworks

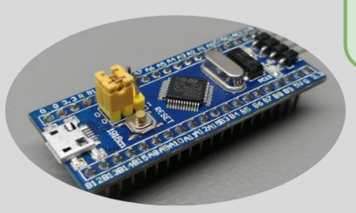
- **Neural Architecture Search**
  - Hardware Aware NAS for fixed hardware
  - Hardware & Software **Co-Exploration**



Printed Electronics

Approximate TinyML Inference on Embedded Systems

ASICs



Microcontrollers

Embedded FPGA