Serving Multi-DNN Workloads on FPGAs: a Coordinated Architecture, Scheduling, and Mapping Perspective
Shulin Zeng, Guohao Dai, Niansong Zhang*, Xiniao Yang, Haoyu Zhang, Zhenhua Zhu, Huazhong Yang, Yu Wang
Tsinghua University

Published in IEEE Transactions on Computers

Key Takeaways
• Heterogenous dataflow and homogenous multi-core offer superior performance for DNN inference in the cloud.
• Optimizing architecture, schedule, and compiler mapping together finds better overall system designs.

Motivation
DNN INFerence-as-a-Service (INFaAS)
• Serve mixed DNN workloads on virtualized multi-tenant cloud FPGA accelerators
• Heterogeneous dataflow for different workloads
• Homogenous multi-core for spatial multi-tenancy

Trade-offs
• Small cores: flexible and secure
• Large core: resource and power efficient

Results
• 3.0–7.5X better EDP than homogenous multi-core accelerators
• 1.8–3.6X better EDP than heterogenous dataflow accelerator with fixed core granularity

Method
• Formulate a joint optimization problem
• Energy-Delay-Product (EDP) as the goal for given workloads
• Co-optimize to find best solution with CMA-ES

Insights
• Vision: many small cores (B512, B800)
• NLP: large core + small cores (B3136, B800)
• Mixed: large cores (B4096, B3136)

![Diagram showing multi-core architectures and bandwidth allocation](image_url)

![Table showing DNN models and resource utilization](table_url)