FPGA IGNITE Summer School 2024

Energy-Efficient Deep Learning Accelerators with Workload Awareness for Embedded FPGAs

Chao Qian, supervised by Prof. Dr. Gregor Schiele
University of Duisburg-Essen, Germany

Motivation

Moving Intelligence to End Devices Offers Benefits

- Lower latency
- Higher accessibility
- Higher reliability
- Data security and privacy

System Model Targeting Energy-Efficiency

- Low-power MCU for coordination and networking
- Embedded FPGA for application-specific DL accelerators

Goal

"Combining efficient inference with workload awareness to optimizing DL accelerators for embedded FPGAs"

Requirements

- Support Various Architectures (MLPs, CNNs, RNNs,Transformers)
- Maintain Acceptable Model Precision Loss

Constraints

- Inference Time: Below the latency required by the application
- Energy: Within a fixed budget per inference
- Resource: Fit FPGA (such as XC7S15, ICE40UP5K)

LSTM Accelerators

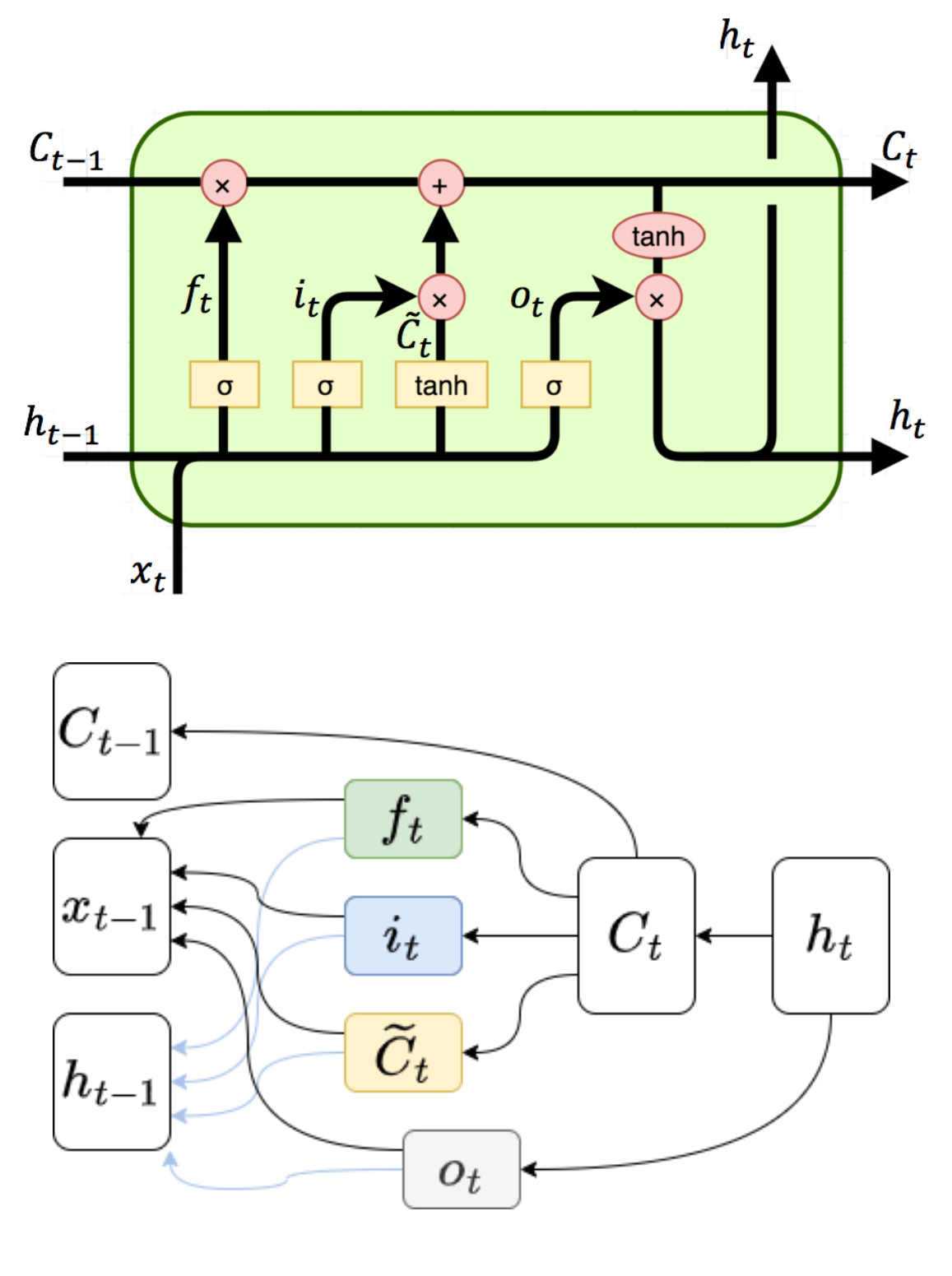
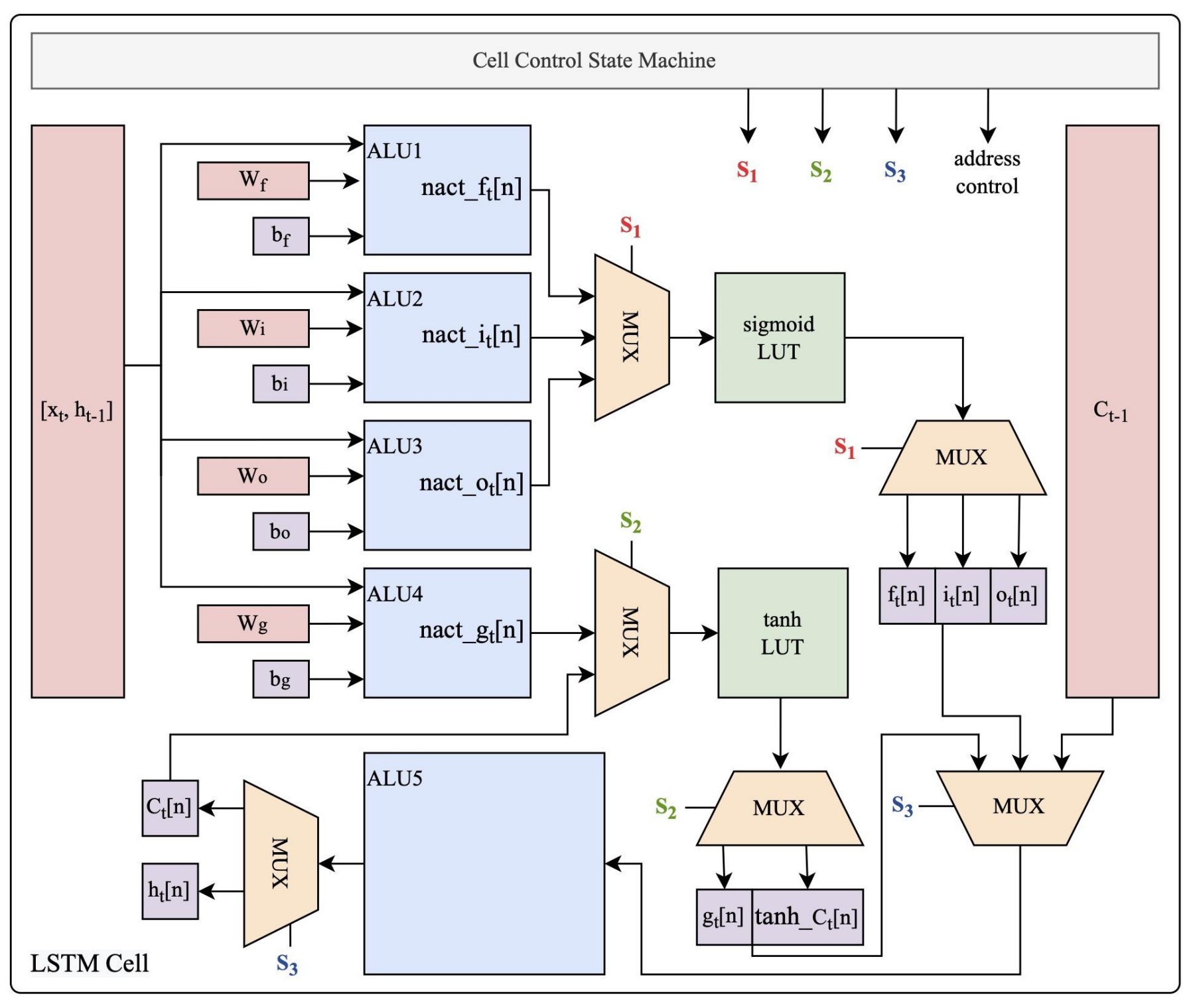
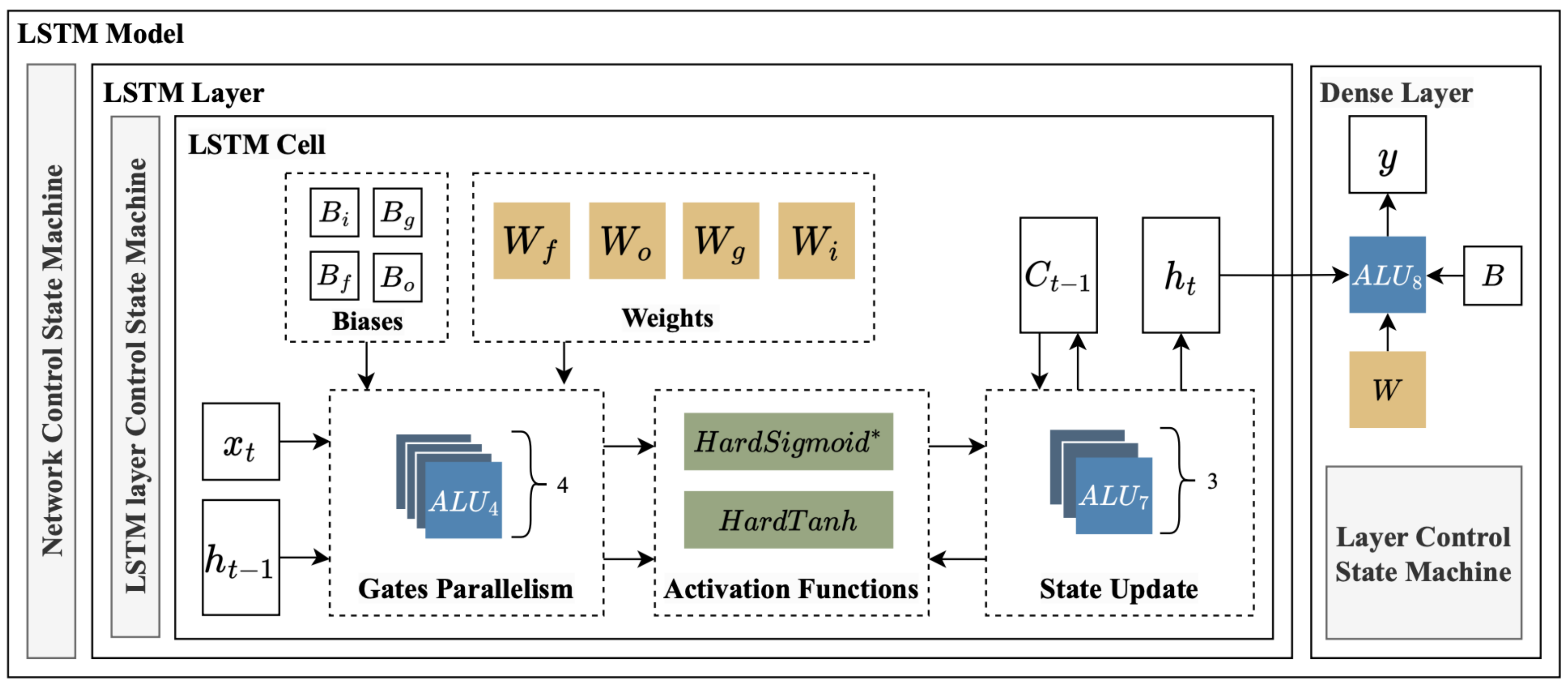


Illustration & Dependency Graph LSTM Cell



Improving Energy Efficiency by Gates Parallelization in LSTM Cell [1]



Improving Clock Frequency with Pipeline, Activation Functions Optimization [2] and Parameterization for Scalability and Stability [3]

AI Workload Awareness

FPGA operates in Duty-Cycle mode

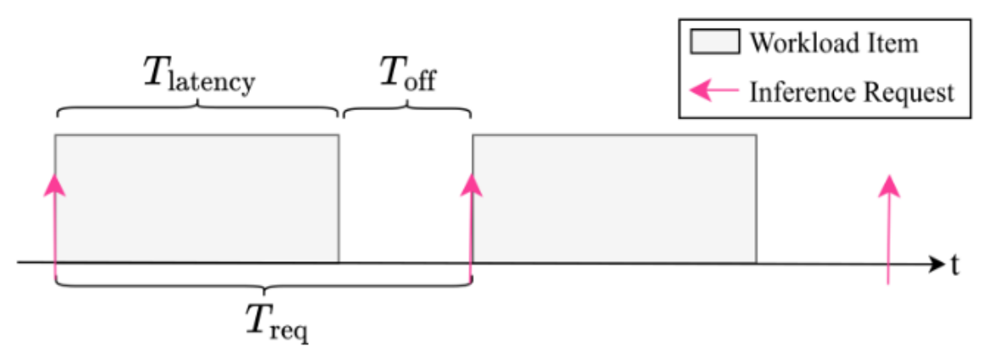
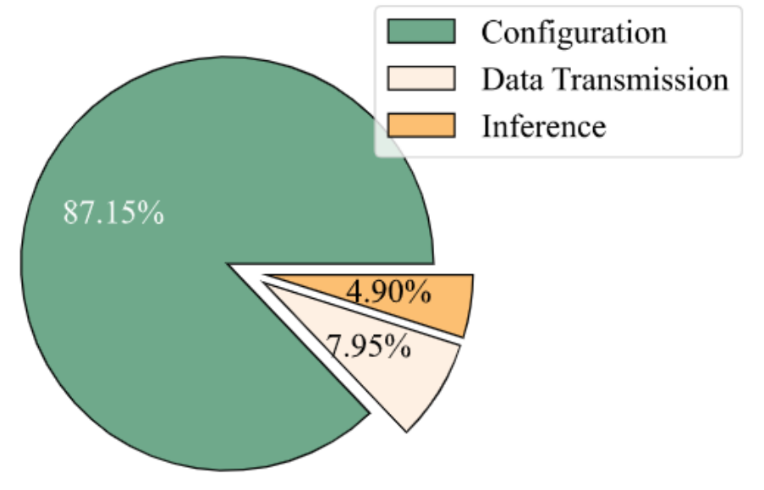


Illustration of On-Off Strategy [5]

But configuration overhead is …



Energy Breakdown of A Workload Item [4]

Our Approach: Idle-waiting Strategy

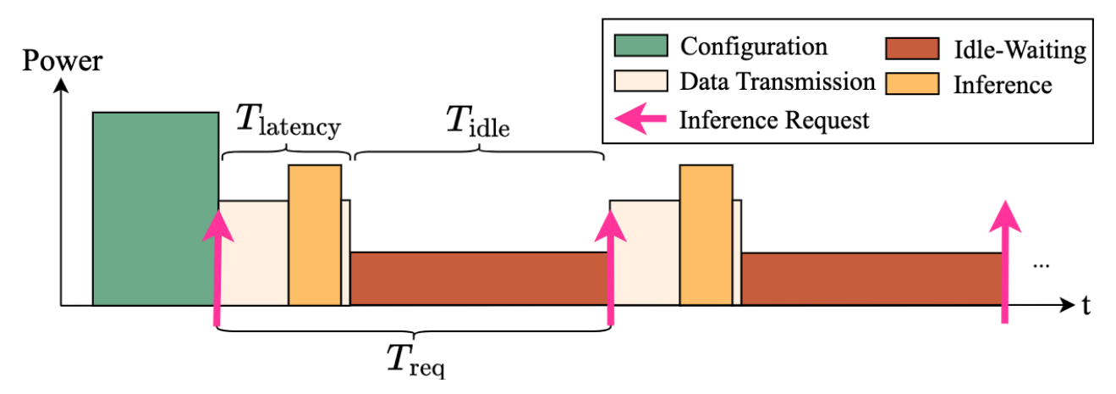
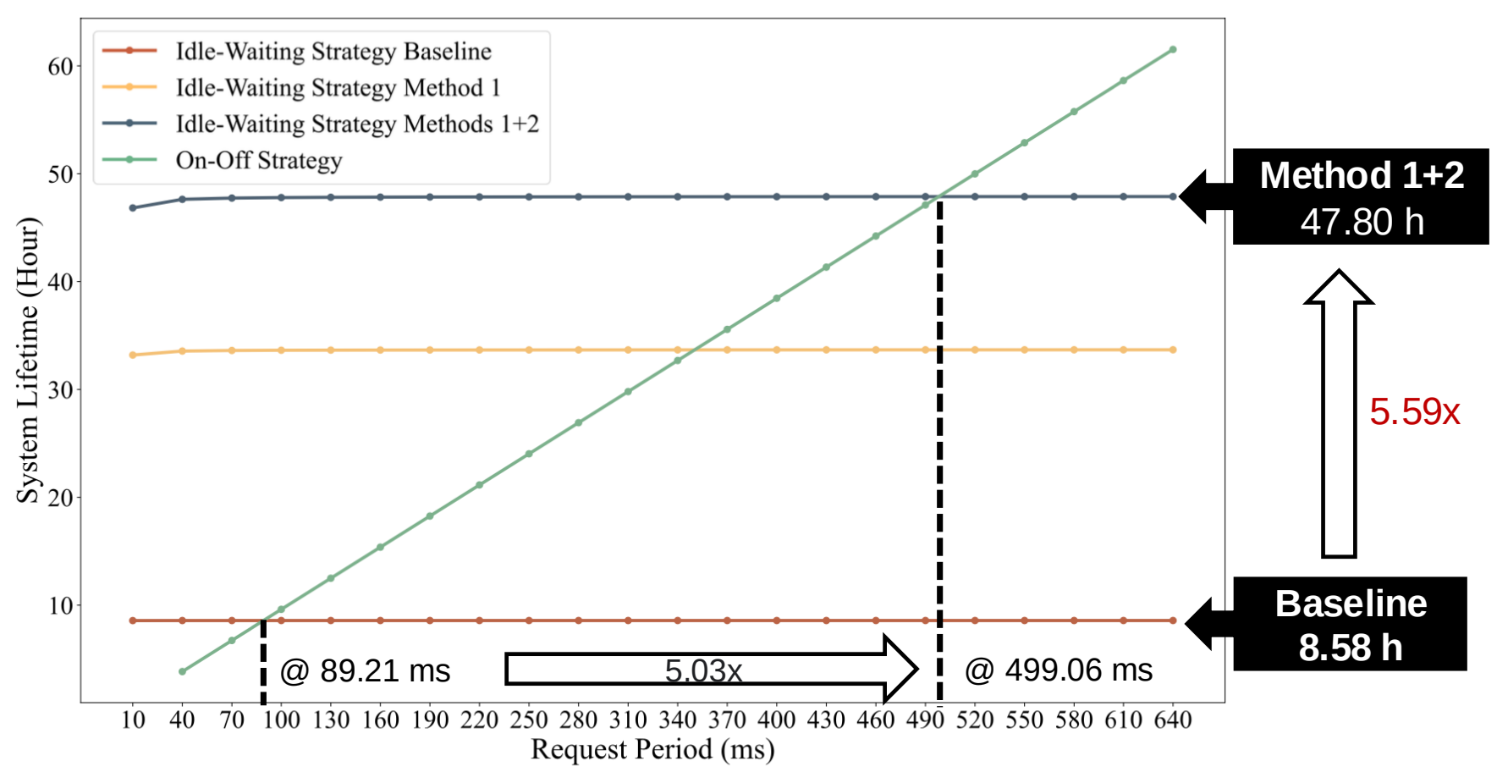


Illustration of Idle-Waiting strategy [5]



Baseline vs. Optimized Methods Across Request Periods [5]

References

[1] Qian, Chao, Tianheng Ling, and Gregor Schiele. "Enhancing energy-efficiency by solving the throughput bottleneck of LSTM cells for embedded FPGAs." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2022.

[2] Qian, Chao, Tianheng Ling, and Gregor Schiele. "Energy efficient LSTM accelerators for embedded FPGAs through parameterised architecture design." International Conference on Architecture of Computing Systems. 2023.

[3] Qian, Chao, Tianheng Ling, and Gregor Schiele. "Exploring energy efficiency of LSTM accelerators: A parameterized architecture design for embedded FPGAs." Journal of Systems Architecture 152 (2024): 103181.

[4] Cichiwskyj, Christopher, Chao Qian, and Gregor Schiele. "Time to learn: Temporal accelerators as an embedded deep neural network platform." International Workshop on IoT, Edge, and Mobile for Embedded Machine Learning. Cham: Springer International Publishing, 2020.

[5] Qian, Chao, Tianheng Ling, and Gregor Schiele. "Idle is the New Sleep: Configuration-Aware Alternative to Powering Off FPGA-Based DL Accelerators During Inactivity." International Conference on Architecture of Computing Systems. 2024.

[6] Burger, Alwyn, Chao Qian, Gregor Schiele, and Domenik Helms. "An embedded CNN implementation for on-device ECG analysis." In 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 1-6. IEEE, 2020.

[7] Ling, Tianheng, Chao Qian, and Gregor Schiele. "On-device soft sensors: Real-time fluid flow estimation from level sensor data." arXiv preprint arXiv:2311.15036 (2023).

[8] Ling, Tianheng, Julian Hoever, Chao Qian, and Gregor Schiele. "FlowPrecision: Advancing FPGA-Based Real-Time Fluid Flow Estimation with Linear Quantization." In 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 733-738. IEEE, 2024.

[9] Ling, Tianheng, Chao Qian, and Gregor Schiele. "Integer-only Quantized Transformers for Embedded FPGA-based Time-series Forecasting in AIoT" [Manuscript submitted for publication]. International Conference on Architecture of Computing Systems. 2024. IEEE Annual Congress on Artificial Intelligence of Things (IEEE AIoT).

[10] Qian, Chao, Lukas Einhaus, and Gregor Schiele. "ElasticAI-Creator: Optimizing neural networks for time-series-analysis for on-device machine learning in IoT systems." In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, pp. 941-946. 2022.

[11] Qian, Chao, Tianheng Ling, and Gregor Schiele. "ElasticAI: Creating and deploying energy-efficient deep learning accelerator for pervasive computing." In 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 297-299. IEEE, 2023.

[12] Ling, T., Qian, C. and Schiele, G., 2024. "Towards Auto-Building of Embedded FPGA-based Soft Sensors for Wastewater Flow Estimation." arXiv preprint arXiv:2407.05102

Knowledge Transferring

Concepts

- Design methodology: VHDL templates
- Optimizations: Pipeline, precomputation, parameterization
- Evaluation methodology: Software estimation + hardware validation

Architectures

- CNN accelerator for EEG Analysis [6]
- MLP for Flow estimation [7] [8]
- Transformer accelerator for Air Quality forecasting [9]

Elastic AI-Creator Toolchain for automation

- Providing optimized RTL templates for components of DL models [10]
- Eliminating the need for expertise in FPGA functionality for DL developers

ElasticAI-Creator
Toolchain
Scan Me!

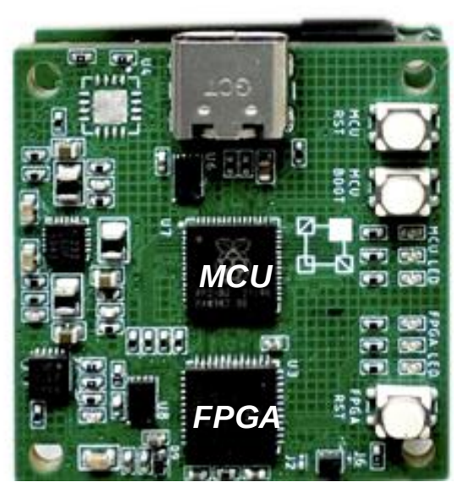
ElasticNode DL Acceleration Platforms

ElasticNode V5 [10]

- Dimension: 57.8 x 34 mm
- Cortex-M0+ MCU: RP2040
- Spartan-7 FPGA: S15, 25, S50
- SRAM(8Mb) + Flash(128Mb)
- Energy Meter: PAC1934
- Battery: 320mAh
- Extensions: ESP32, Sensors

ElasticNode V5 SE [12]

- Dimension: 34 x 34 mm
- ICE40UP5K
- Flash(16Mb)



UNIVERSITÄT
DUISBURG
ESSEN
Offen im Denken

Intelligent Embedded Systems Lab
Faculty of Informatics
University of Duisburg-Essen
47057 Duisburg, Germany
https://www.uni-due.de