

Fault-Tolerant CNN Accelerator with Reconfigurable Capabilities

Rizwan Tariq Syed, Advisor: Milos Krstic

MOTIVATION

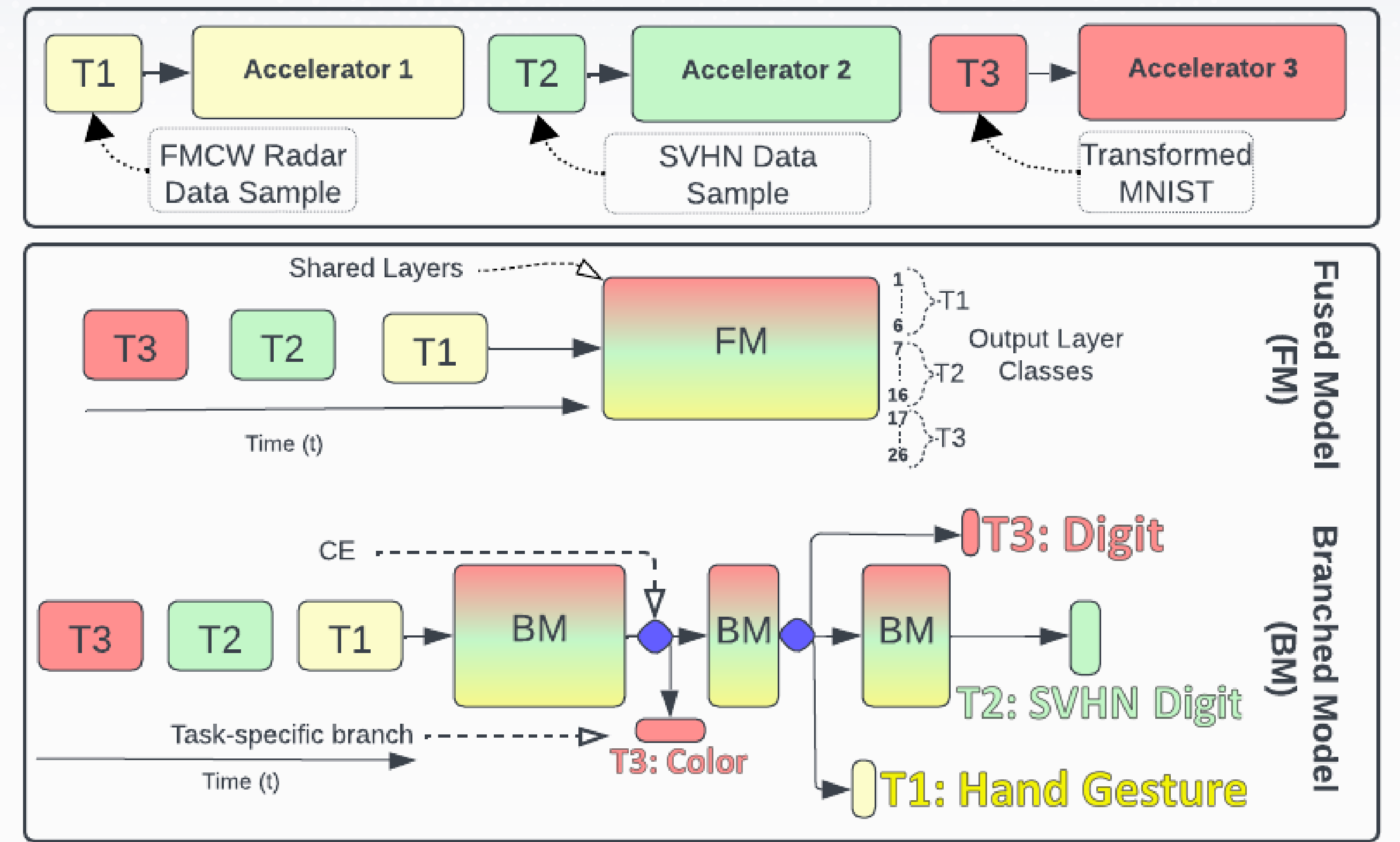
- AI models will eventually be deployed on edge devices
- Requirements for AI hardware accelerators may vary during runtime.
- Change in AI requirements directly impacts power consumption and hardware resource utilization.
- A significant concern for safety-critical applications (automotive, space, 6G infrastructure, etc.) is to ensure reliability against transient and permanent faults.

GOALS

- Efficient implementation of CNN hardware accelerator to support multiple sensor modalities.
- Self-adaptive/reconfigurable hardware accelerator addressing reliability challenges due to soft error and aging
- Development of a fully reconfigurable AI processing system accelerator to fulfill varying AI application requirements.

MULTI-MODAL CNN ACCELERATOR BASED ON SHARED LAYERS METHODOLOGY

- Shared-layers methodology leverages fundamental working principles of CNNs to learn patterns
- Employing a single accelerator capable of handling diverse, uncorrelated tasks from different sensor modalities, achieving an average accuracy exceeding 90%.
 - T1 Accuracy: 97.33 %
 - T2 Accuracy: 89.22 %
 - T3 Accuracy: 94.36 %
- Fused and Branched CNN architectures are evaluated for three distinct tasks based on accuracy, quantization, pruning, hardware resource utilization, power, and latency.
- Noteworthy reduction in hardware resources utilization and power



Shared-layers-based fused model capable of executing multiple tasks from multiple modalities

RECONFIGURABLE CNN ACCELERATOR

- Extend the shared-layers methodology and proposed fault-tolerant CNN accelerators with reconfigurable capabilities

HP Mode

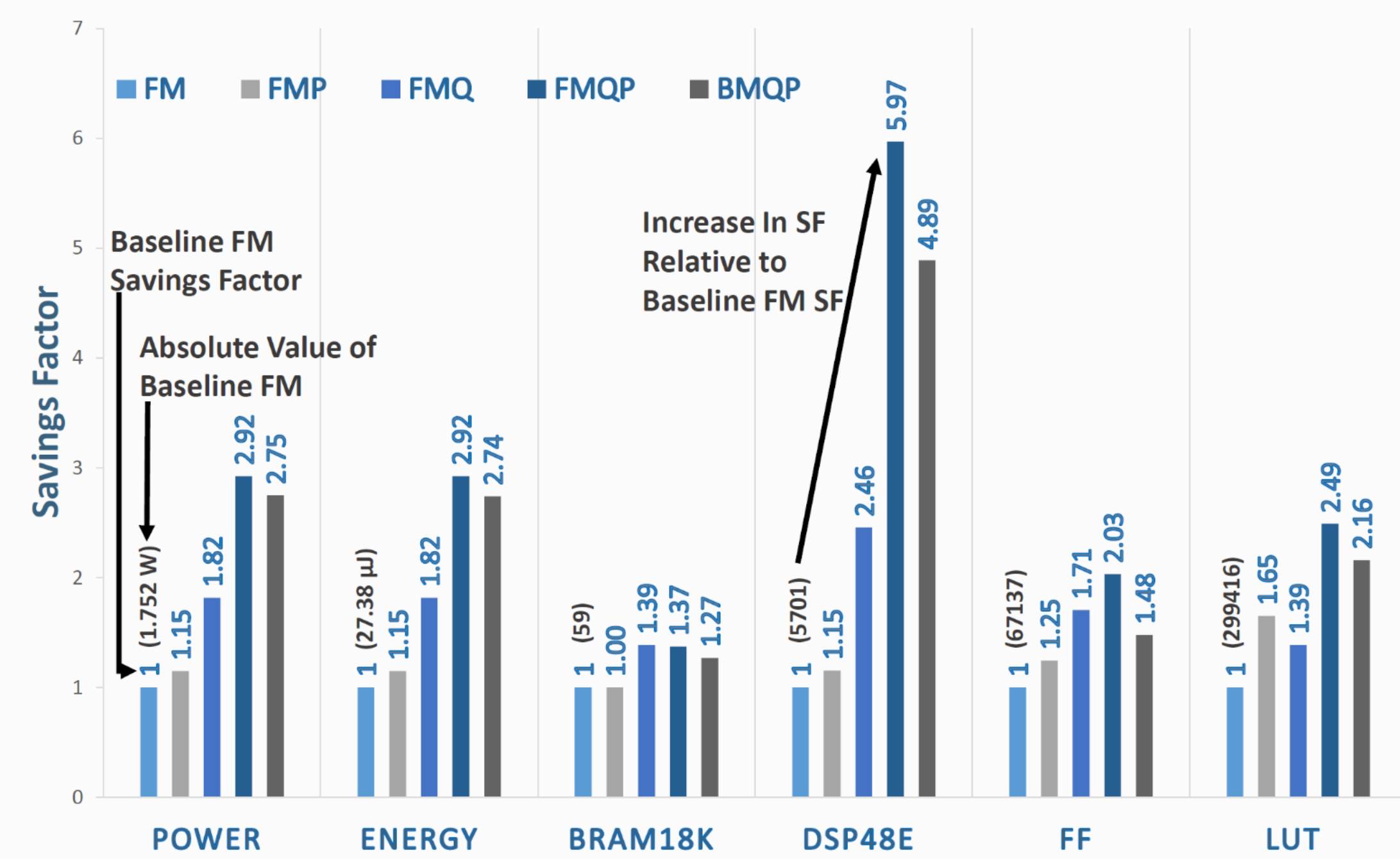
- Enabling parallel execution of multiple tasks for low latency requirements
- Examination of latency and energy consumption

FT Mode

- Fail-safe (DMR) and fail-operational (TMR) mode for reliable AI processing
- Fault analysis on various fault models i.e., SETs, SEUs, MBUs, SEU in FPGA CRAM).

DS Mode

- Aging aware mode to reduce aging and power consumption.
- Comparative assessment between clock gating (CG) and partial reconfiguration(PR) methods to reduce the dynamic power consumption

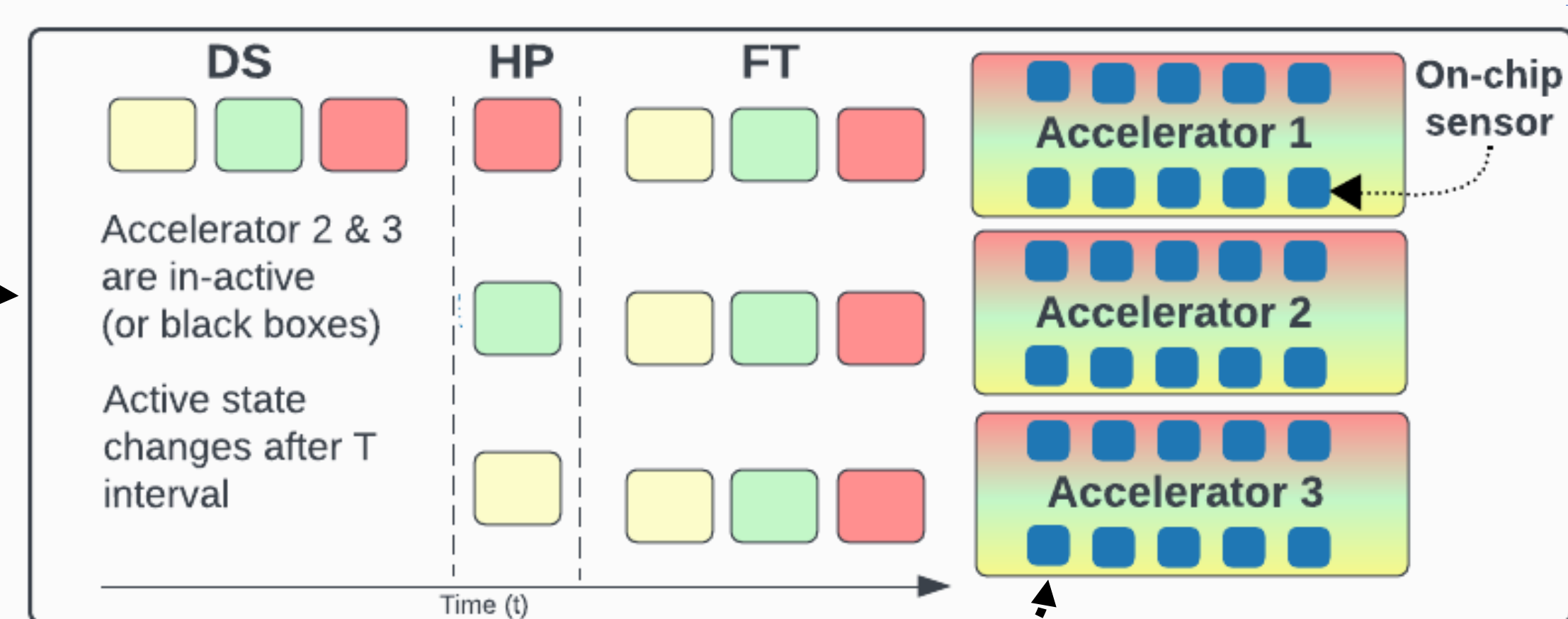


Post-Implementation Power, Energy, and Hardware Resource Utilization Results.

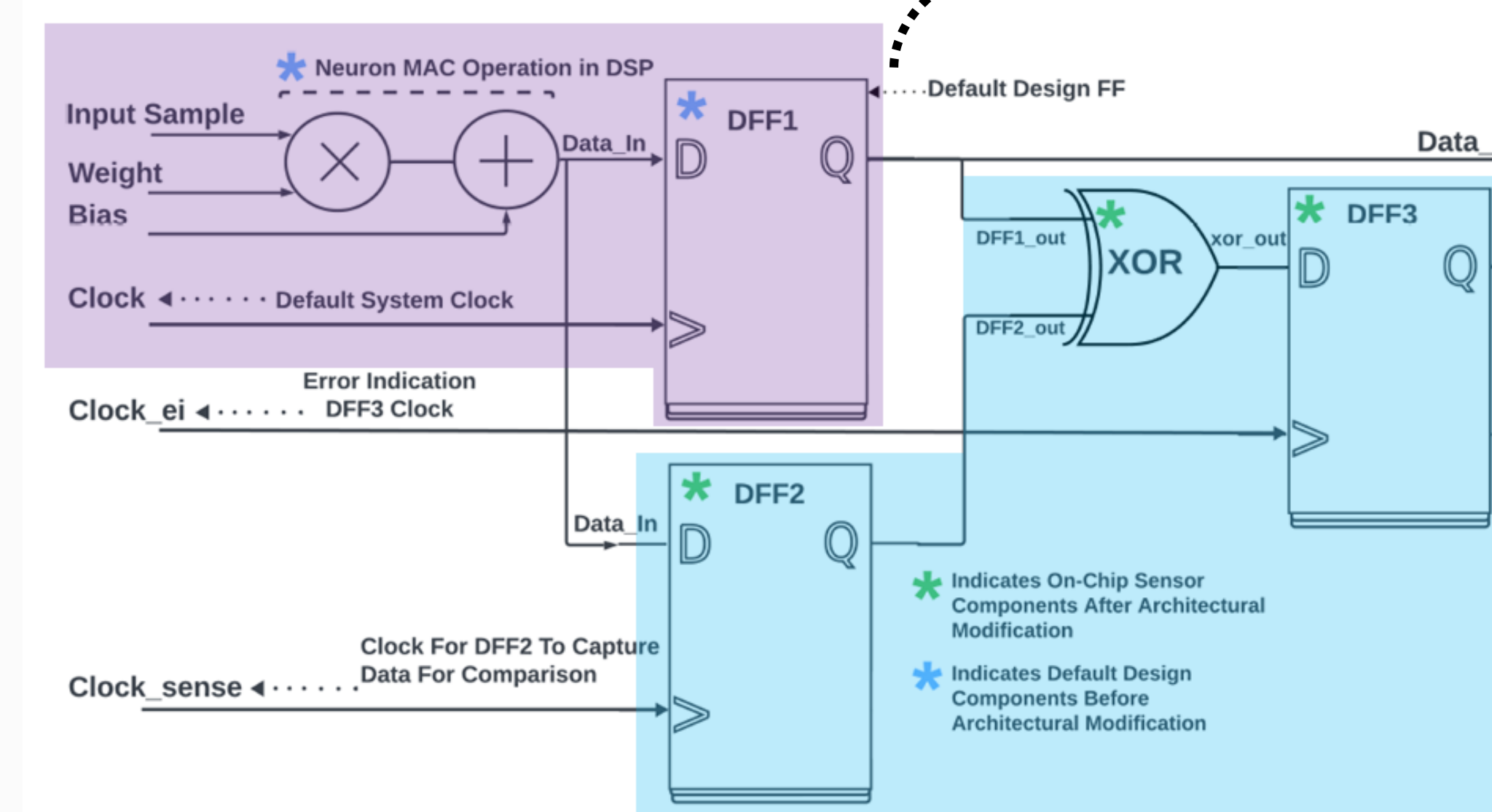
TOWARDS FULLY RECONFIGURABLE AI PROCESSING SYSTEM

- Development of a multi-purpose on-chip sensor with aging and soft errors detection capabilities.
- Integration of the on-chip sensor in a CNN hardware accelerator.
- Reconfigure the CNN accelerator by utilizing the intelligence provided by the on-chip aging and soft error sensor. Future work includes adding more on-chip sensors. i.e., temperature, supply voltage variations, etc.
 - FT Mode: Rise in soft error occurrence.
 - DS Mode: Increase in aging
 - HP Mode: As per the application's requirements

Tasks execution in multiple modes



On-chip Intelligence-driven Reconfiguration of CNN accelerators



On-Chip Sensor Integration in CNN Accelerator

R. T. Syed, M. Krstic, et al, "FPGA Implementation of a Fault-Tolerant Fused and Branched CNN Accelerator With Reconfigurable Capabilities," in IEEE Access, vol. 12, pp. 57847-57862, 2024, doi: 10.1109/ACCESS.2024.3392240.
 R. T. Syed, M. Krstic, et al, "Towards Reconfigurable CNN Accelerator for FPGA Implementation," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 3, pp. 1249-1253, March 2023, doi: 10.1109/TCSII.2023.3241154.
 R. T. Syed, M. Krstic, et al, "Aging and Soft Error Resilience in Reconfigurable CNN Accelerators Employing a Multi-Purpose On-Chip Sensor," 2024 IEEE 25th Latin American Test Symposium (LATS), Macao, Brazil, 2024, pp. 1-6, doi: 10.1109/LATS62223.2024.10534625.
 R. T. Syed, M. Krstic, et al, "FPGA-Based Acceleration of Convolutional Neural Network for Gesture Recognition Using mm-Wave FMCW Radar," 2022 IEEE Nordic Circuits and Systems Conference (NorCAS), Oslo, Norway, 2022, pp. 1-7, doi: 10.1109/NorCAS57515.2022.9934412.
 R. T. Syed, M. Krstic, et al, "A Survey on Fault-Tolerant Methodologies for Deep Neural Networks. Pomiar Automatyka Robotyka, 27(2), 89-98.
 R. T. Syed, M. Krstic, et al, "Fault Resilience Analysis of Quantized Deep Neural Networks," 2021 IEEE 32nd International Conference on Microelectronics (MIEL), Nis, Serbia, 2021, pp. 275-279, doi: 10.1109/MIEL52794.2021.9569094.

Funding: This work is funded by the Federal Ministry of Education and Research of Germany (BMBF) within the "Open6GHub" project (grant number: 16KISK009).



M.Sc. Rizwan Tariq Syed
 Scientist
 System Architectures / Fault Tolerant Computing
 IHP

German Research Center for Artificial Intelligence GmbH
 Trippstadter Str. 122, 67663 Kaiserslautern
 open6ghub-info@dfki.de
 www.open6ghub.de

