

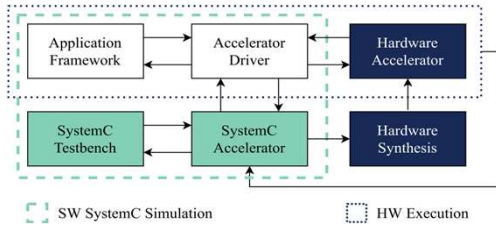
Jude Haris, José Cano

School of Computing Science, University of Glasgow, UK

## 1. Developing Specialized Accelerators for Edge AI

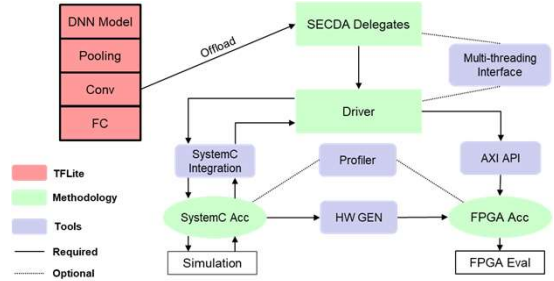
**Problem:** current solutions for designing AI accelerators for edge devices with FPGAs have a very high development cost (HLS, System integration).

**Solution:** SECDA [1] reduces the development time of FPGA-based accelerators for AI through hardware-software co-design and SystemC simulation.



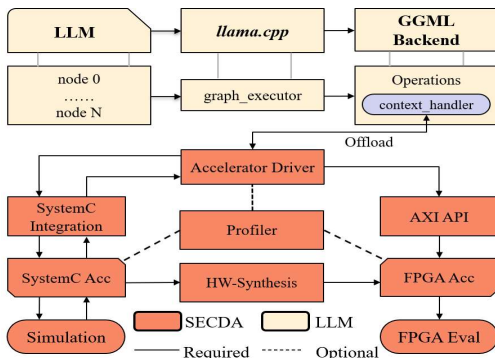
## 2. SECDA-TFLite

SECDA-TFLite [2] enables the design of new DNN accelerators for edge inference instantiating the SECDA methodology within TFLite (now LiteRT).



## 3. SECDA-LLM

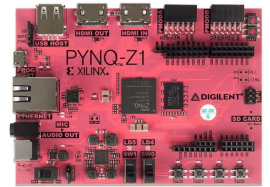
SECDA-LLM [3] enables the design of new LLM accelerators for edge inference instantiating the SECDA methodology within llama.cpp.



## 4. Experimental Setup

PYNQ-Z1 development board:

- Arm A9 dual-core CPU @ 650 MHz
- Xilinx Z020 edge FPGA
- 512 MB DDR3 memory
- Used with SECDA-TFLite



KRIA KV260 development board:

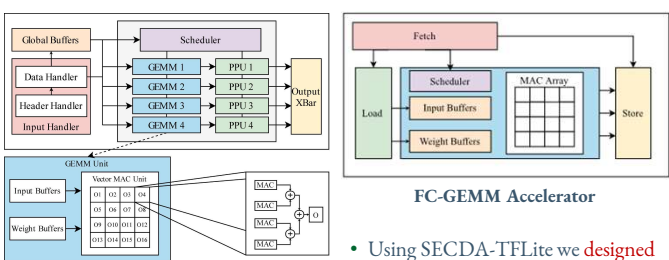
- Arm Cortex-A53 CPU @ 1.5 GHz
- Zynq UltraScale+ XCK26 FPGA
- 4 GB DDR4 memory
- Used with SECDA-LLM



We evaluated latency and power (for PYNQ) across different hardware configurations:

- CPU only
- CPU + accelerator

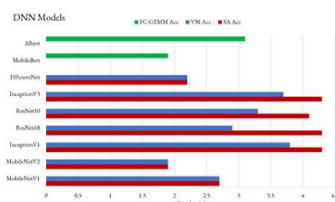
## 5. SECDA-TFLite Case Study



- Using SECDA-TFLite we designed and evaluated 3 different accelerators for DNN inference.

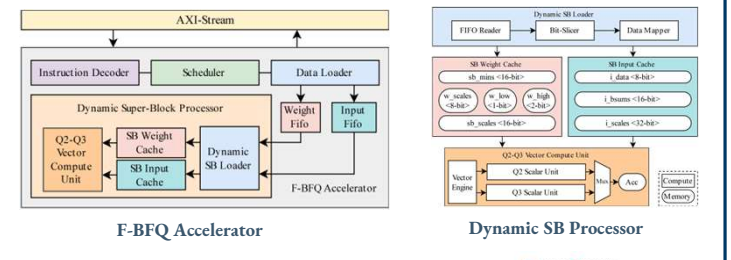
- The VM and SA accelerator were designed to process the CONV2D layers within CNN models.

- The FC-GEMM accelerator was designed to accelerate INT8 quantized Fully Connected layers within Transformer models.



- Average speedup for inference time of up to 3.4x and 2.5x for CNN and BERT models.
- Average energy savings of up to 2.9x and 2.4x for CNN and BERT models.

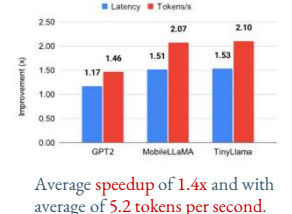
## 6. SECDA-LLM Case Study



- Using SECDA-LLM we designed a Flexible Block Floating-Point Quantization Accelerator for Q2\_K and Q3\_K MatMul kernels [4].

- The Dynamic Super-Block Processor exploits parallelism across BFP super blocks.

- Memory hierarchy enables multi-data format caching without wasting memory space.



Average speedup of 1.4x and with average of 5.2 tokens per second.

## References

- [1] J. Haris, P. Gibson, J. Cano, N. Bohm Agostini, D. Kaeli. SECDA: Efficient Hardware-Software Co-Design of FPGA-based DNN Accelerators for Edge Inference. SBAC-PAD 2021.
- [2] J. Haris, P. Gibson, J. Cano, N. Bohm Agostini, D. Kaeli. SECDA-TFLite: A toolkit for efficient development of FPGA-based DNN accelerators for edge inference. JPDC 2023.
- [3] J. Haris, R. Saha, W. Hao, J. Cano. Designing Efficient LLM Accelerators for Edge Devices. ARC-LG @ ISCA 2024.
- [4] J. Haris, J. Cano. F-BFQ: Flexible Block Floating-Point Quantization Accelerator for LLM. LG-ARC @ ISCA 2025.