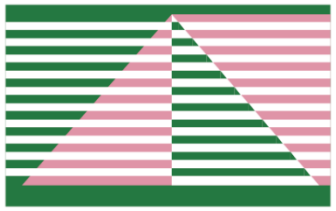
University of Glasgow

gicLAB

SECDA Design Suite

Efficient HW-SW Co-Design of DNN Accelerators for Edge Inference

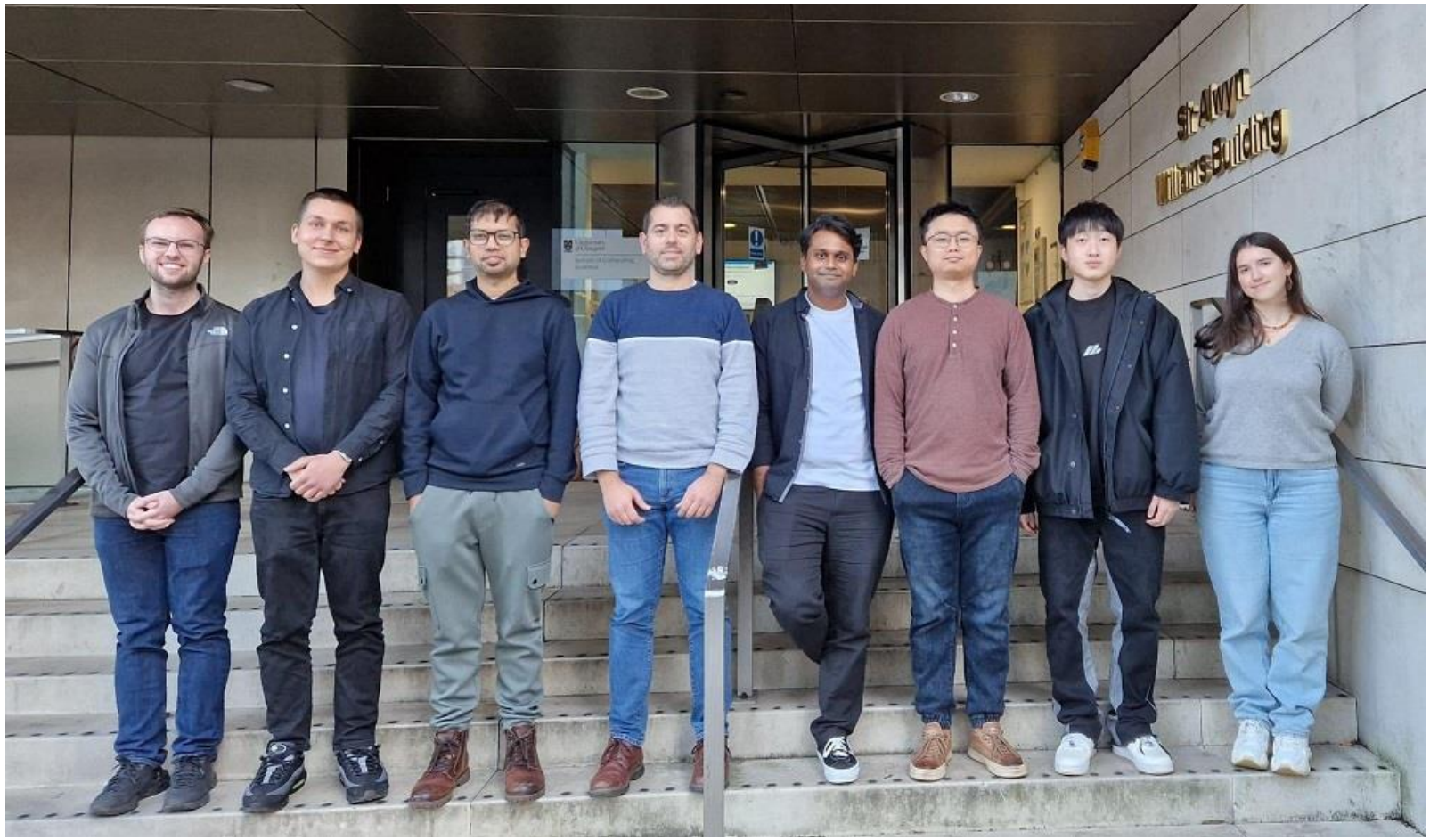
Jude Haris, José Cano
School of Computing Science
University of Glasgow, Scotland, UK



OSCAR Workshop @ ISCA'25
Tokyo, Japan – 21/06/2025

Glasgow Intelligent Computing Lab (gicLAB)

University of Glasgow



School of Computing Science

Systems Section (GLASS)

Computing Systems

&

Machine Learning

&

Security

gicLAB

https://giclab.dcs.gla.ac.uk/

Key concept: Deep Learning Acceleration Stack (DLAS)

University of Glasgow

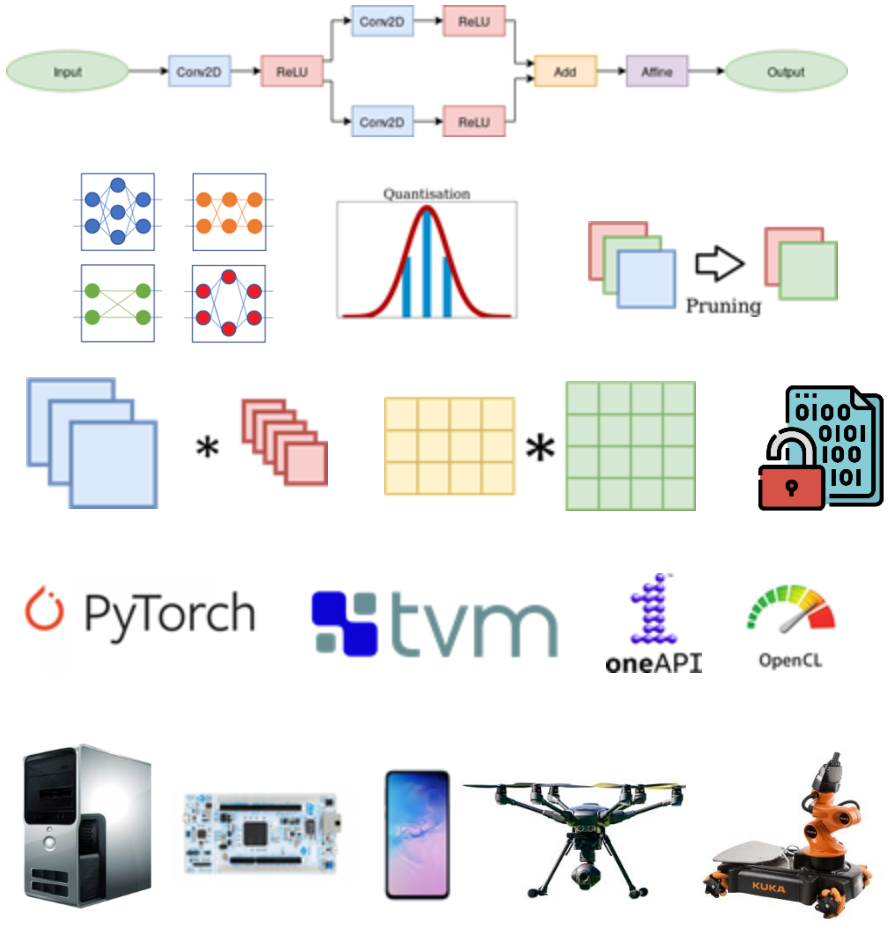
Neural Network Models & Datasets
(Image, video, voice, text, etc)

Optimization Techniques
(Pruning, quantization, NAS/HPO, etc)

Algorithmic Primitives & Data Formats
(GEMM, Winograd, CSR, Encryption, etc)

Systems Software
(Libraries, frameworks, compilers, etc)

Hardware
(Server class, Edge/IoT/Tiny devices)



Across-stack optimizations are required to provide efficient solutions!

[P. Gibson, J. Cano, E. J. Crowley, A. Storkey, M. O'Boyle, "*DLAS: A Conceptual Model for Across-Stack Deep Learning Acceleration*", **ACM TACO'25**]

Outline

University of Glasgow

- SECDA Methodology

- SECDA Design Suite
  - Core
  - Instantiations
    - SECDA-TFLite
    - SECDA-LLM
    - SECDA-Sandbox
  - Benchmarking

- Conclusions and Future Work

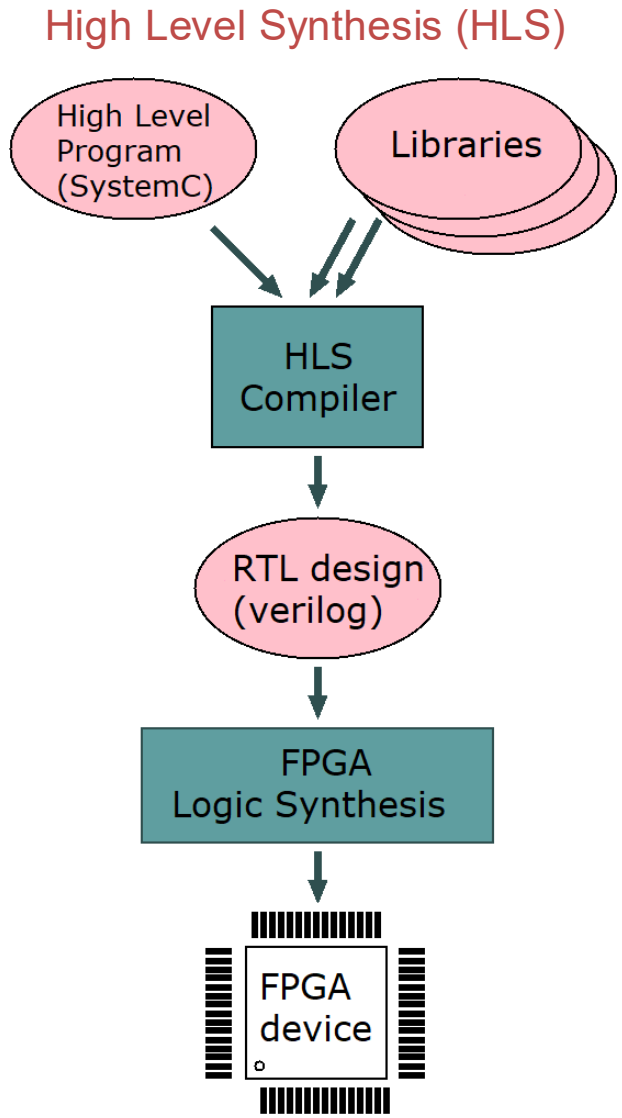
Developing Specialized Accelerators for Edge AI

University of Glasgow

- **Motivation**: specialized hardware accelerators (ASICs, FPGAs, etc) can make AI faster and more energy efficient (e.g. at the edge)
  - FPGAs are reconfigurable circuits commonly present in edge devices

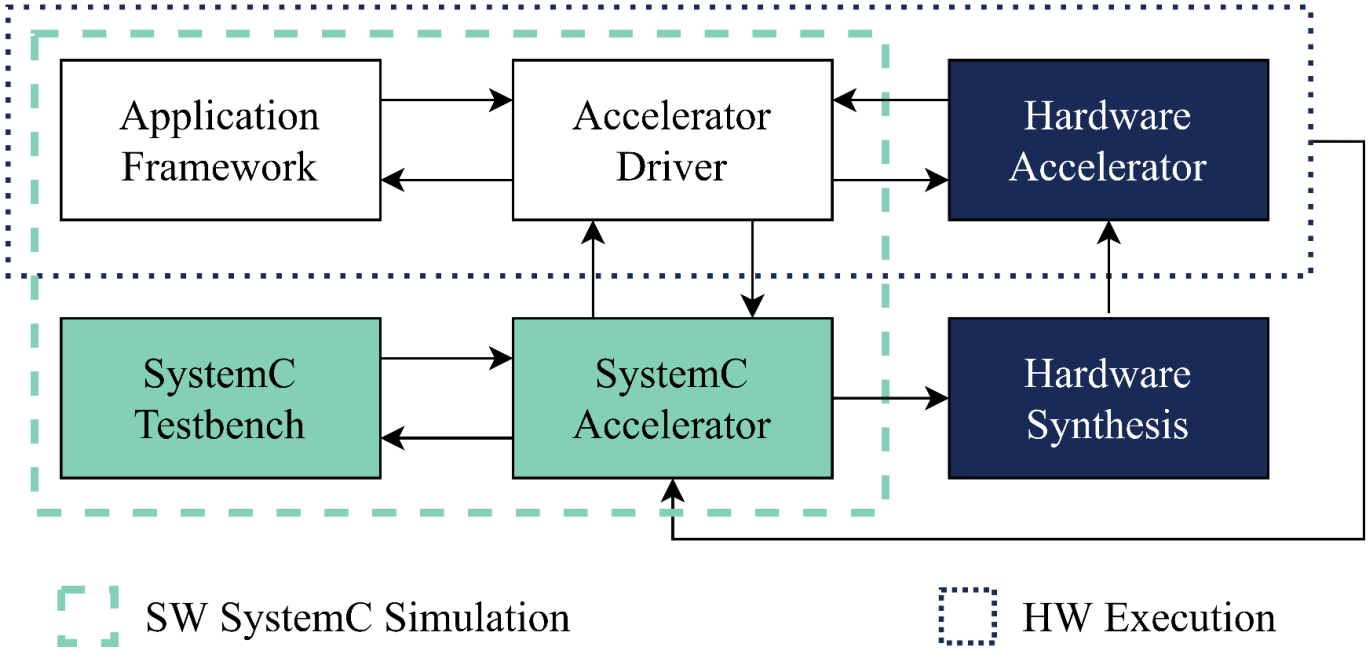
- **Problem**: current solutions for designing AI accelerators for edge devices with FPGAs have a very high development cost
  - They require High Level Synthesis (HLS)
  - FPGA synthesis is a very slow process that is repeated (over designs)
  - System integration issues (e.g. accelerator and DNN framework)

- **Solution**: we proposed a design methodology (SECDA) to efficiently reduce the development time of FPGA-based accelerators (for edge devices)
  - Combines cost-effective SystemC simulation with hardware execution



SECDA Methodology: Overview

- **SECDA**: SystemC Enabled Codesign of DNN Accelerators



*[J. Haris, P. Gibson, J. Cano, N. B. Agostini, D. Kaeli, "*SECDA: Efficient Hardware/Software Co-Design of FPGA-based DNN Accelerators for Edge Inference*", **SBAC-PAD'21**]

Outline

University of Glasgow

- SECDA Methodology

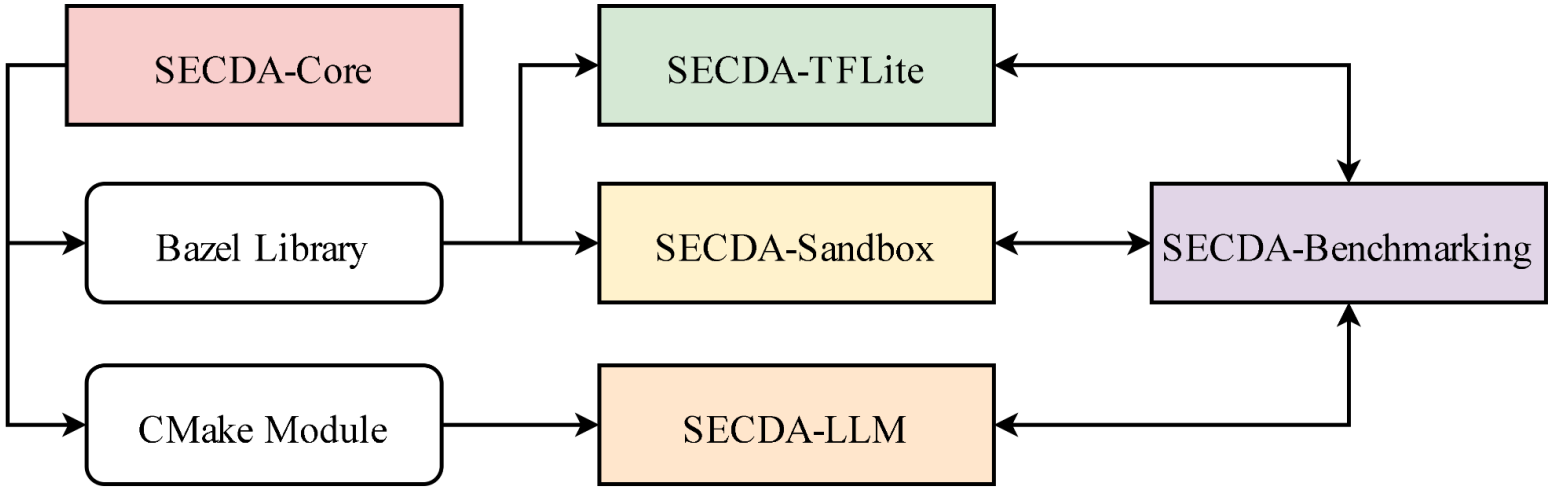
- SECDA Design Suite
  - Core
  - Instantiations
    - SECDA-TFLite
    - SECDA-LLM
    - SECDA-Sandbox
  - Benchmarking

- Conclusions and Future Work

SECDA Design Suite

University of Glasgow

- Provides a Hardware-Software co-design environment: SECDA, ML frameworks, custom workloads
- **SECDA-Core** library of tools that enables SECDA; supported with Bazel and CMake build systems
- Three SECDA Instantiations**: SECDA-TFLite**, **SECDA-LLM** and **SECDA-Sandbox**
- Systematic **Benchmarking** to ease testing and evaluation of new designs

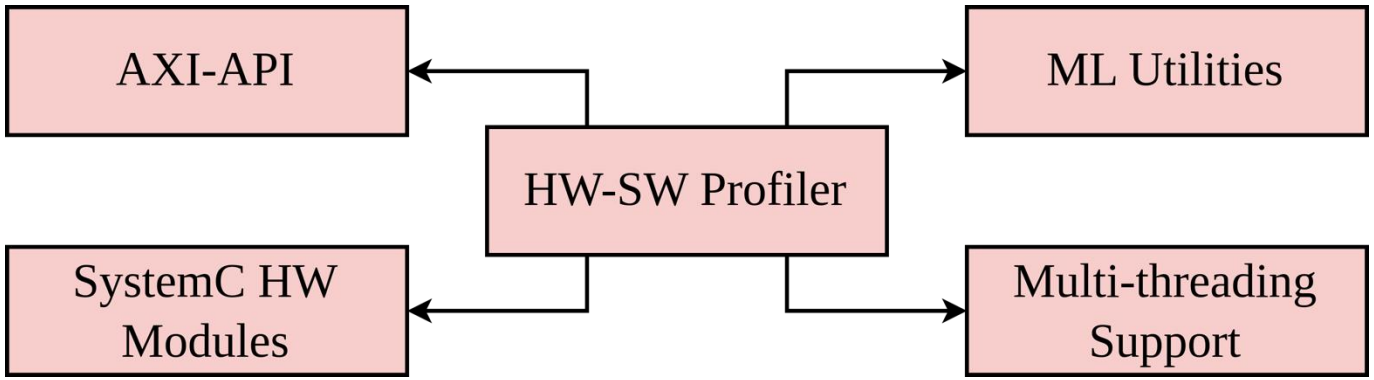


https://github.com/gicLAB/SECDA-Design-Suite

SECDA-Core

University of Glasgow

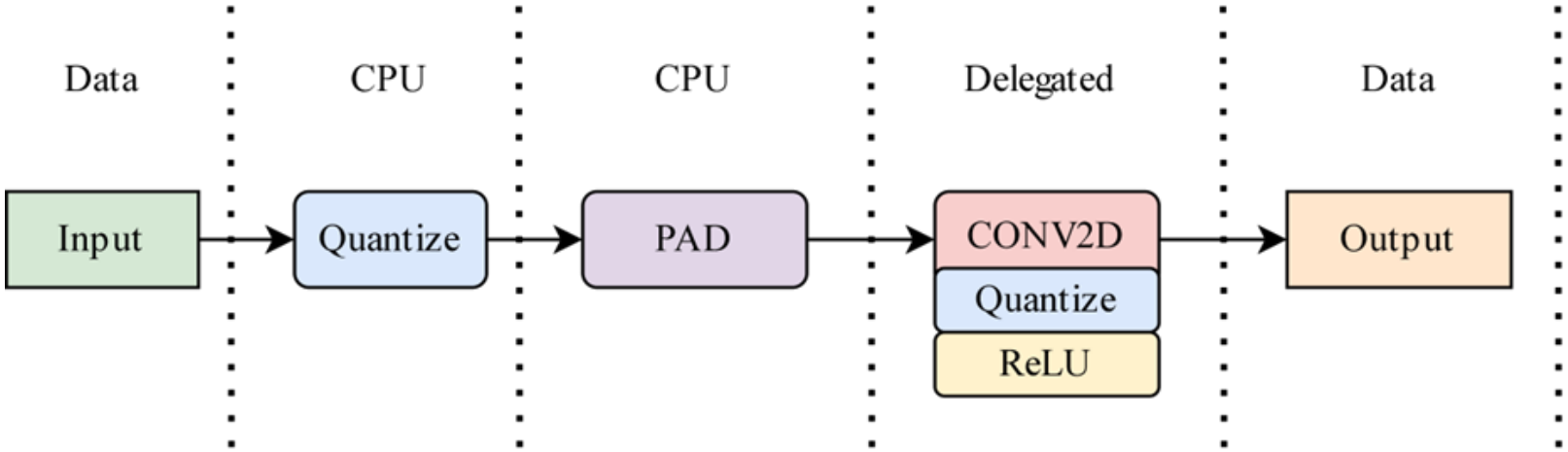
- **AXI-API**: defines communication channels and data transfers between accelerator and CPU
- **SystemC HW Modules**: simulation and development of new accelerators (e.g., DMA Engine module)
- **ML Utilities**: helper functions for pre/post-processing of ML operations (e.g., calculate data size of tensors)
- **Multi-threading Support**: CPU threads and tasks that interact with accelerators independently
- **HW-SW profiler**: enables extensive and detailed analysis of performance



TFLite Delegate System & API

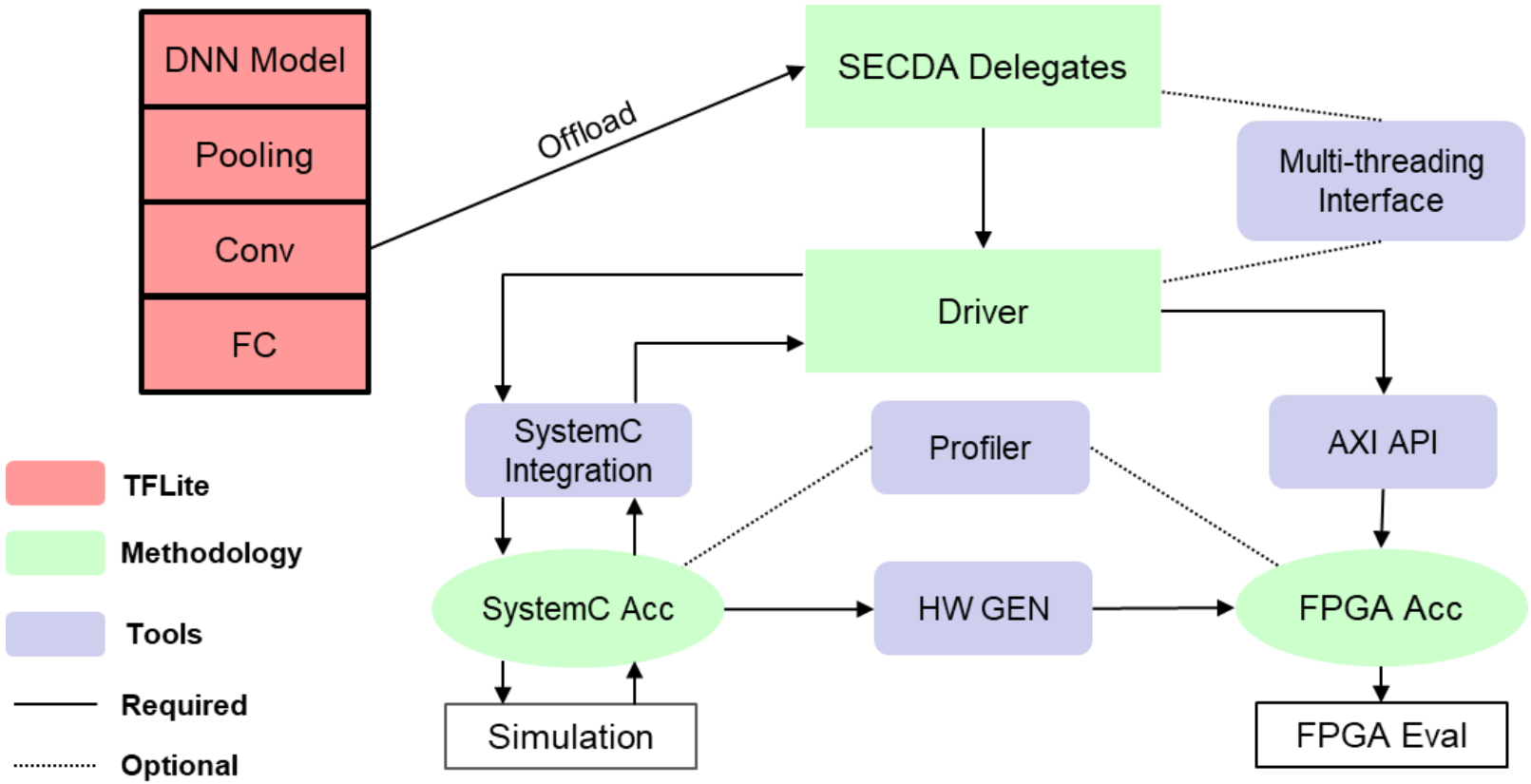
University of Glasgow

- **TensorFlow Lite (TFLite)**: framework for running DNN models on resource constrained **edge devices**
- The **Delegate system** enables to offload computation using different backends (software, hardware)
  - **Examples**: NNAPI delegate for Android, Core ML delegate for iOS, etc
- The **Delegate API** enables the development of **custom delegates**
  - VM_del, SA_del, …



SECDA-TFLite

- A toolkit for designing custom FPGA-based accelerators for TFLite
- Instantiates the SECDA methodology within TFLite
- Enables fast prototyping and integration of new accelerators with significantly **reduced initial setup costs**



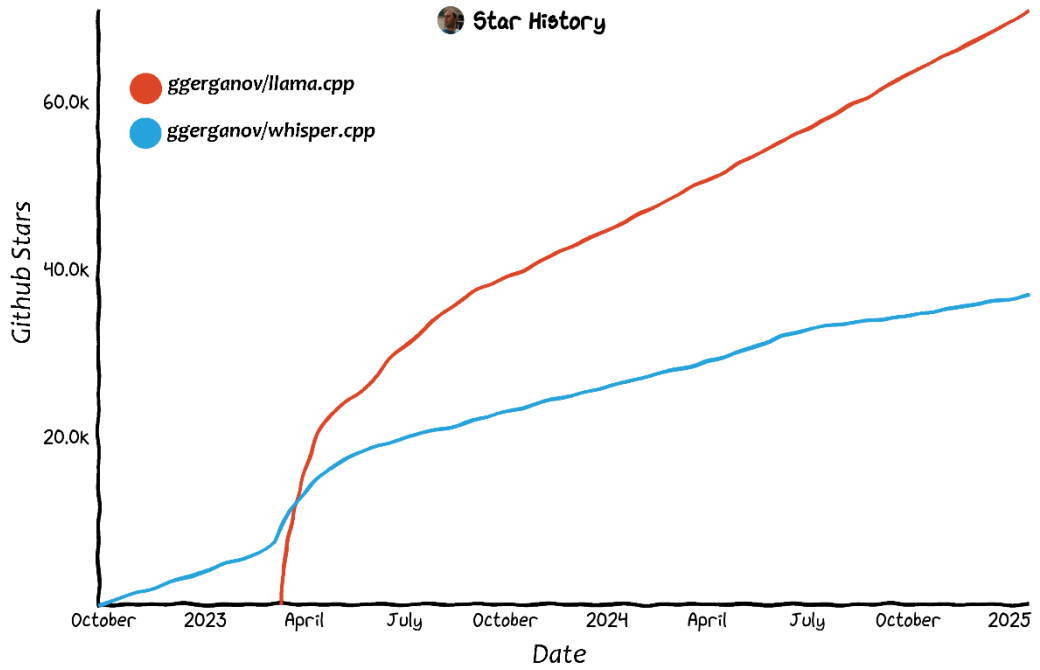
https://github.com/gicLAB/secda-tflite

*[J. Haris, P. Gibson, J. Cano, N. B. Agostini, D. Kaeli, "*SECDA-TFLite: A Toolkit for Efficient Development of FPGA-based DNN Accelerators for Edge Inference*", **Elsevier JPDC'23**]

llama.cpp

University of Glasgow

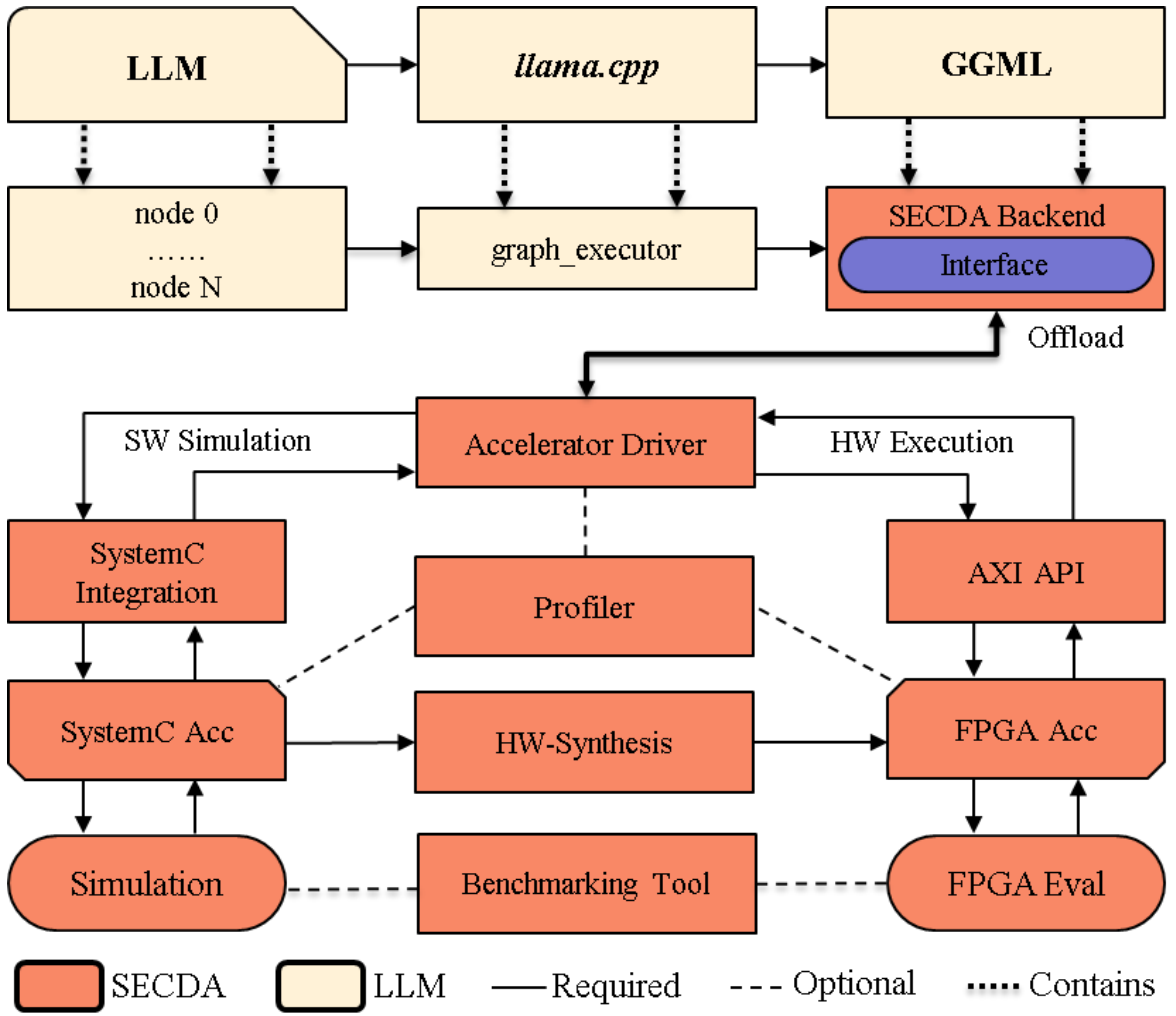
- A pure **C/C++ library** with minimal external dependencies
- Enables LLM inference with **minimal setup** and **state of the art performance** on a wide range of hardware devices
- Supports multi-modal, custom, and well-known LLMs (e.g., Llama, Falcon, GPT, Gemma)
- Utilizes **GGUF** (GPT-Generated Unified Format) and supports various type of quantization (1.5-bit, 2-bit, 3-bit, 4-bit, 5-bit, 6-bit, and 8-bit)
- **Open source**, with active and rapidly growing community



https://github.com/ggerganov/llama.cpp

SECDA-LLM

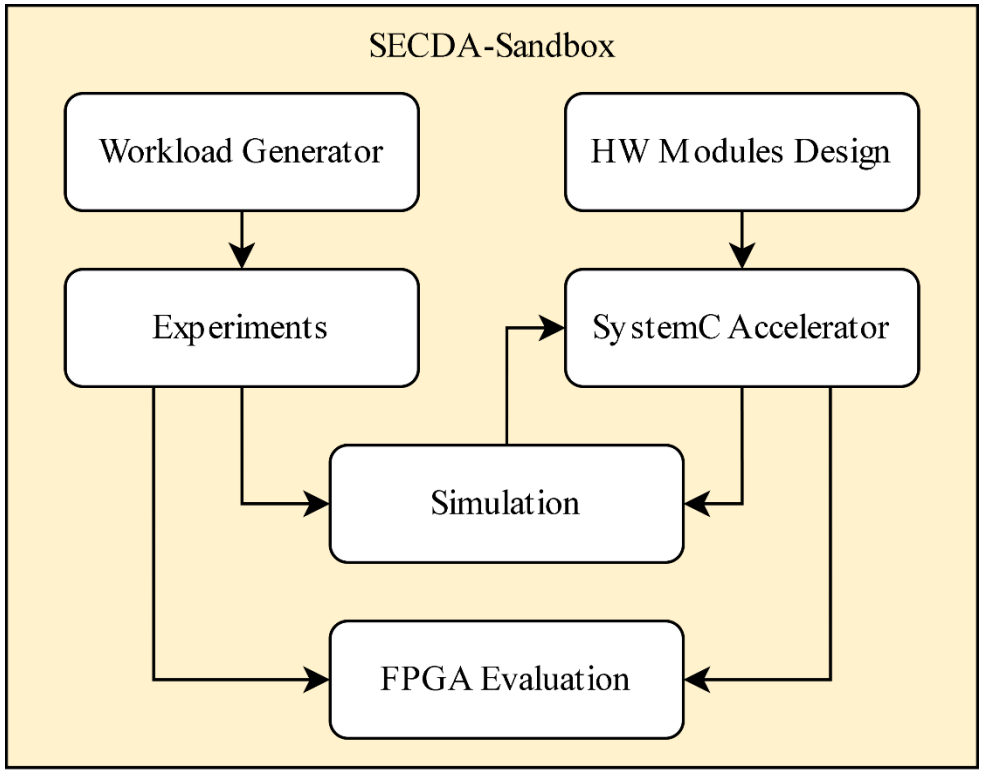
- A toolkit for designing custom FPGA-based accelerators for LLMs
- Instantiates the SECDA methodology within ***llama.cpp***
- Enables fast prototyping and integration of new accelerators with significantly **reduced initial setup costs**



*[J. Haris, R. Saha, W. Hu, J. Cano, "*Designing Efficient LLM Accelerators for Edge Devices*", **ARC-LG @ ISCA'24**]

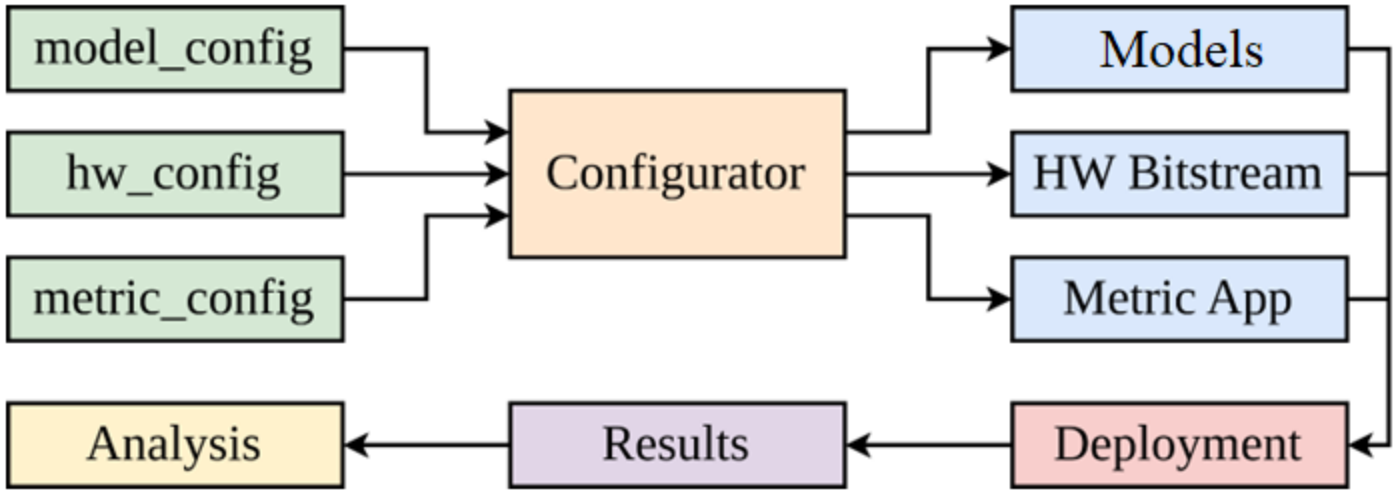
SECDA-Sandbox

- Lightweight environment without ML framework ideal for modular testing of small Hardware modules
- Allows you to work with custom workloads
- Enables exporting Hardware designs to SECDA-LLM or SECDA-TFLite for full-scale end-to-end evaluation
- Plans to further expand with non-AI workloads
  - HPC workloads
  - Genome analysis
  - ...



SECDA-Benchmarking

- Automates the process of running benchmarks on the target FPGA-based device and collecting the results
- Easy and flexible experiments: version control, automated visualization of results
- **Experiments**: Model x Accelerator x Metric
- **Metrics**: latency, throughput and power



Outline

- SECDA Methodology

- SECDA Design Suite
  - Core
  - Instantiations
    - SECDA-TFLite
    - SECDA-LLM
    - SECDA-Sandbox
  - Benchmarking

- Conclusions and Future Work

Conclusions

University of Glasgow

- Accelerating AI applications on **Edge devices** is becoming more and more important
- **SECDA** is a design methodology to reduce the development time of specialized AI accelerators
- **SECDA Design Tool** provides an overall HW-SW co-design environment that enables developers to use the SECDA methodology supporting multiple ML frameworks and custom workloads
  - **SECDA-TFLite** is a new open source toolkit that improves/eases the development of new FPGA-based accelerators for edge DNN inference employing **TFLite** and the SECDA methodology
  - **SECDA-LLM** is a new toolkit that improves/eases the development of new FPGA-based accelerators for edge LLM inference employing ***llama.cpp*** and the SECDA methodology

Future Work

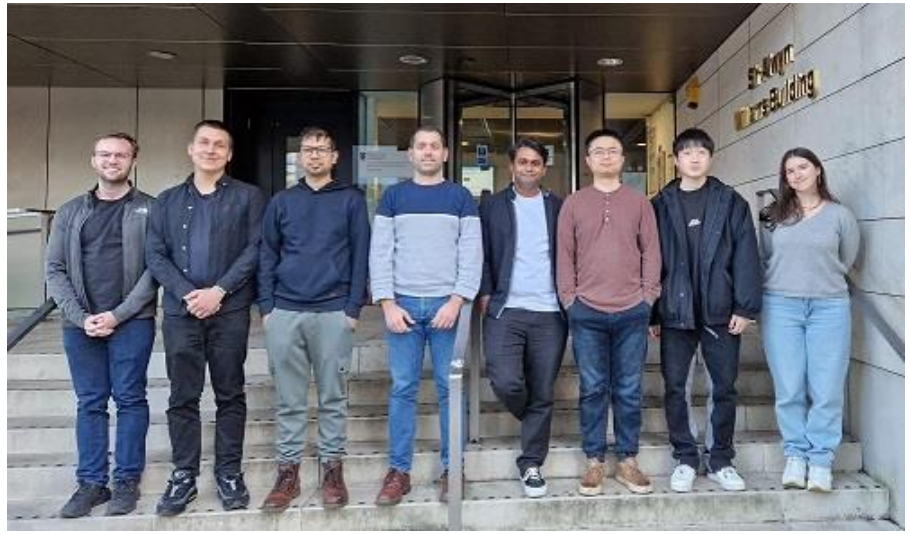
University of Glasgow

- On-going
  - Expand **SECDA-LLM** as an open-source platform (more LLMs, datasets, and edge devices evaluated)
  - Support more types of layers (e.g. transposed convolutions), cores (e.g. shift-based), sparsity
  - **SECDA-DSE**: Automate the design space exploration process to design better hardware accelerators

- Planned
  - Design and Simulate Process-in-Memory (PIM)-Based Accelerators
  - Acceleration of AI applications on heterogeneous accelerators (GPUs, NPUs, FPGAs)

- Potential
  - Create similar development toolkits for other DNN frameworks such as PyTorch, TVM, etc
  - Developing Custom Accelerators for DNN Training at the edge

Acknowledgements



1) **Researchers** and **students** at gicLAB

2) **Funding** bodies

Horizon Europe
THE NEXT EU RESEARCH & INNOVATION PROGRAMME (2021–2027)

UKRI Engineering and Physical Sciences Research Council

sicsa* The Scottish Informatics & Computer Science Alliance

APRIL AI Hub for Productive Research & Innovation in Electronics

3) **Collaborators** from Academia

THE UNIVERSITY of EDINBURGH

Northeastern University

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

ETH zürich

BOSTON UNIVERSITY

KAIST

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSIDAD DE MURCIA

UNIVERSITAT JAUME I

4) **Collaborators** from Industry and Labs

AMD

NVIDIA

arm

Pacific Northwest NATIONAL LABORATORY

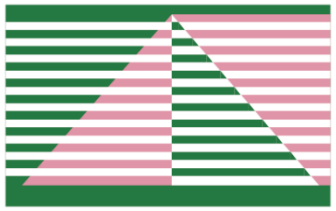
University of Glasgow

gicLAB

SECDA Design Suite

Efficient HW-SW Co-Design of DNN Accelerators for Edge Inference

Jude Haris (j.haris.1@research.gla.ac.uk), José Cano (Jose.CanoReyes@glasgow.ac.uk)
School of Computing Science
University of Glasgow, Scotland, UK



Thank you! Questions?

gicLAB: Publications

University of Glasgow

- Optimization Techniques
  - Optimizing Grouped Convolutions on Edge Devices **ASAP'20**
  - Improving Robustness Against Adversarial Attacks with Deeply Quantized Neural Networks **IJCNN'23**
  - **ICE-Pruning**: An Iterative Cost-Efficient Pruning Pipeline for Deep Neural Networks **IJCNN'25**
  - Exploiting Unstructured Sparsity in Fully Homomorphic Encrypted DNNs **EuroMLSys @ EuroSys'25**
- Systems Software
  - **AXI4MLIR**: User-Driven Automatic Host Code Generation for Custom AXI-Based Accelerators **CGO'24**
  - **Transfer-Tuning**: Reusing Auto-Schedules for Efficient Tensor Program Code Generation **PACT'22**
  - **Bifrost**: End-to-End Evaluation and Optimization of Reconfigurable DNN Accelerators **ISPASS'22**
- Hardware
  - **SECDA-TFLite**: A Toolkit for Efficient Development of FPGA-based DNN Accelerators for Edge Inference **JPDC'23** **SBAC-PAD'21**
  - **SECDA-LLM**: Designing Efficient LLM Accelerators for Edge Devices **ARC-LG @ ISCA'24**
  - Accelerating PoT Quantization on Edge Devices **ICECS'24**
  - Accelerating Transposed Convolutions on FPGA-based Edge Devices **FPL'25**