Aly Sultan

978-325-1925 | Boston, MA
sultan.a@northeastern.edu | linkedin.com/in/aly-sultan/ | github.com/asultan123

Education

Northeastern University — Boston, MA
Ph.D. Computer Engineering — *Expected 2025*

Northeastern University — Boston, MA
M.S. Electrical and Computer Engineering — *2023*

American University in Cairo — Cairo, Egypt
B.S. Electronics & Communication Engineering — *2019*

Experience

Graduate Software Engineering Intern — 2022 – Present
SSM Datacenter Simulation, Intel — *Part Time, Remote*

AI Cost Reduction in Simics SWCI

- Developed an AI solution to predict regression test failures based on developer source changes, aiming to save computational resources by running only tests that are likely to fail
- Established a data collection pipeline from Splunk and GitHub's APIs, emphasizing data integrity with JSON schema and daily data health-check reports
- Trained XGBoost model on data collected and achieved up to a 75% reduction in tests run on GNR with a miss rate of 5.35% for failing tests
- Developed a tool for streamlined concurrent experiments on Intel's SimCloud
- Shared project insights at Intel's internal AI Everywhere Conference and the S3E Tech Exchange

Extending the autogen framework

- Transitioned SDSi, Virtualization, and IP patching regression tests in GNR from Simics CLI to Python within the Simics autogen framework, enhancing their availability to new platforms

Refactoring S3M SWCI Jenkins pipeline

- Revitalized a previously non-functional CI pipeline for S3M firmware integration, achieving consistent daily firmware deliveries for S3M
- Devised a versatile shell library within Python, streamlining local and remote shell operations across geographically dispersed data centers due to tooling constraints

Graduate Research Assistant — 2020 – Present
Embedded Systems Lab, Northeastern University — *Boston, MA*

HERO Architecture

- Developed a SystemC model for HERO, an innovative matrix multiplication and convolution accelerator for DNN inference
- Introduced Self Addressable Memory (SAM) for adaptive on-chip data orchestration in HERO
- Established HERO-SIM, a PyTorch-SystemC based simulation framework for the HERO accelerator
- Evaluated HERO's efficacy on 695 DNNs, achieving up to 30X speedup and 300X energy savings over a workstation-class CPU
- Currently finalizing HERO manuscript for publication

Categorized Ensemble Networks for Adversarial Attack Defence (CAEN)

- Lead an AI defense project focused on bolstering ensemble network resilience against image-based adversarial attacks
- Developed a novel training methodology combining soft labeling with dissimilar label pairing, formulated the problem as an ILP, and solved it with Gurobi
- Training methodology achieved a 1.1X increase in robust accuracy over SOTA while reducing FLOPs by 16.8%
- Currently finalizing CAEN manuscript for publication

Graduate Teaching Assistant — Fall 2022
Electrical and Computer Engineering Department, EECE 7368 — *Boston, MA*

- Transitioned the course from SpecC to SystemC
- Enhanced course realism by enabling usage of the Xilinx-QEMU co-simulation environment via Docker containers
- Developed clear, structured lab exercises in SystemC, providing students with initial code and documentation

Projects

Integer Linear Program Based Scheduler for Multi-Core Processors | *Python, Pyomo, Gurobi* 2021

- Implemented an ILP-based scheduling model to schedule applications statically on a multicore processor
- Model was generated using python and the pyomo library and optimized with the Gurobi solver
- Successfully generated optimal schedules for a variety of application sizes and core count configurations

Accelerating Domain Design Space Exploration with CUDA | *C++, CUDA* 2020

- Accelerated the application binding evaluation in the Domain-specific design space exploration for streaming applications algorithm developed by NEU's ESL team
- Improved binding evaluation runtime by ∼100X over CPU baseline using CUDA

Cache Optimization for CNN Inference | *C, Darknet, Intel Pin* 2019

- Integrated Intel Pin with the Darknet framework to determine the effect of cache configurations on CNN inference
- Explored effects of cache hierarchy levels, sizing and replacement policy on data movement to and from DRAM during CNN inference

Darknet Convolution Inference Accelerator | *C, Darknet, Vivado HLS* 2019

- Developed a General Matrix Multiplication (GEMM) accelerator using Vivado HLS
- Integrated accelerator into the Darknet framework
- Accelerated inference time of the tiny darknet network by a factor of 2X over CPU Baseline on a Zynq-7020 soc

Publications

J. Zhang, **A. Sultan**, M. Zandigohar, and G. Schirner, "Generating Unified Platforms Using Multigranularity Domain DSE (MG-DmDSE) Exploiting Application Similarities," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 42, no. 1. Institute of Electrical and Electronics Engineers (IEEE), pp. 280–293, Jan. 2023. doi: 10.1109/tcad.2022.3172373.

J. Zhang, **A. Sultan**, H. Tabkhi, and G. Schirner, "MG-DmDSE: Multi-Granularity Domain Design Space Exploration Considering Function Similarity," 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, Feb. 01, 2021. doi: 10.23919/date51398.2021.9474196.

A. Sultan, A. H. Hassan, and H. Mostafa, "A Compact Low-Power Mitchell-Based Error Tolerant Multiplier," 2018 New Generation of CAS (NGCAS). IEEE, Nov. 2018. doi: 10.1109/ngcas.2018.8572297.

Technical Skills

Languages: *Python, C/C++, SystemC, Matlab, Verilog*
Framework: *Intel Pin, Darknet*
Developer Tools: *Git, Docker, VS Code, Vivado, Vivado HLS, QEMU, Gurobi*
Libraries: *PyTorch, Numpy, Matplotlib, Pyomo*