

# Aly Sultan

978-325-1925 | Boston, MA

[sultan.a@northeastern.edu](mailto:sultan.a@northeastern.edu) | [linkedin.com/in/aly-sultan/](https://www.linkedin.com/in/aly-sultan/) | [github.com/asultan123](https://github.com/asultan123)

## EDUCATION

### Northeastern University

*Ph.D. Computer Engineering*

Boston, MA

2019 – Expected 2025

### American University in Cairo

*B.S. Electronics & Communication Engineering*

Cairo, Egypt

2014 – 2019

## EXPERIENCE

### Graduate Research Assistant

2020 – Present

*Embedded Systems Lab, Northeastern University*

*Boston, MA*

- Developed a SystemC model of a novel Hybrid general matrix multiplication and convolution accelerator (HERO)
- Developed a Template Optimization tool (TEMPO) that optimizes HERO templates based on available compute resources and a target library of neural network layers
- Estimated HERO's latency and energy consumption using a Python-SystemC based simulation environment (HERO-Sim)
- Created a SystemC model of a novel programmable memory primitive called Self Addressable Memory (SAM) used in statically scheduling different dataflows in novel accelerator architectures
- Implemented a SAM program compiler that generates data movement programs from convolution layer configurations to orchestrate datamovement between SAMS in HERO
- Supervised a team of three undergraduate students in:
  - \* Implementing SAMs in Verilog and synthesizing them on a Zynq-7020 based FPGA
  - \* Integrating HERO into Xilinx's SystemC+QEMU simulation environment

## 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 Gurobi solver
- Successfully generated optimal schedules for a variety of application sizes and core count configurations

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

2019

- Used Intel Pin and 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
- Synthesized accelerator on a Zynq-7020 development board
- Integrated accelerator into Darknet framework
- Accelerated inference time of tiny darknet network by a factor of 2X over CPU Baseline

## PUBLICATIONS

### A Compact Low-Power Mitchell-Based Error Tolerant Multiplier | *Verilog, Matlab*

2018

- Developed a novel approximate multiplier architecture in Verilog and evaluated numerical accuracy in JPEG compression using MATLAB
- Design improved power-delay-product by 1.9X with only 20% reduction in peak signal-to-noise ratio in JPEG compression compared to a Xilinx Zynq-7020 DSP
- Published and presented findings at NGCAS, Malta

## 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