Oliver Cassidy

ollyj.cassidy@gmail.com | P: +44 7484 232315 | https://github.com/ollycassidy13 | https://olly-cassidy.vercel.app/ | LinkedIn

EDUCATION

Imperial College London — London, UK
MEng in Electronic and Information Engineering — Expected Jun 2027
Predicted First-Class Honours; Dean's List 2024; Ranked 2nd in the year
Coursework: NN in NumPy, Pipelined ray-tracing engine in Verilog on PYNQ, pipelined RISC-V CPU in SystemVerilog with cache, C90 compiler, Intel FPGA/AWS based game system, WiFi controlled rover, Op-amp from BJT in LTSpice

The Manchester Grammar School — Manchester, UK
A Levels – Mathematics A*, Further Mathematics A*, Physics A*, Electronics A*; Winner of the Paton Electronics Prize — Jun 2023
Returned by invitation in 2025 to present to sixth form students about research and development in electronics

PROFESSIONAL EXPERIENCE

Software Development Engineer — London, UK
AMD Research and Advanced Development — Apr 2026-Sep 2026

Software Engineering Intern — London, UK
T. Rowe Price — Jun 2025-Sep 2025

- Established a secure research environment on Azure using Terraform and Groovy, deploying OpenAI agents with Bing search and MCP
- Collaborated with Microsoft and international teams developing Terraform frameworks and Docker containers to scale deployments
- Implemented in Python a custom MCP server/client with SQL database, agents and Bing search in AI Foundry on a private network
- Presented regularly to stakeholders ensuring technical solutions aligned with their business needs and the understanding of model

Private 1:1 Tutoring — Jan 2021-Sep 2024

- Used social media to market my own tutoring business and attract clients, leading to a full client roster and a waitlist
- Tutored over fifteen students at GCSE and A level in preparation for their examinations, leading to an increase in achieved grades

CONFERENCE PUBLICATIONS

ReducedLUT: Table Decomposition with "Don't Care" Conditions — Monterey, USA
Paper published in the ACM/SIGDA International Symposium on Field-Programmable Gate Arrays 2025 — Aug 2024-Feb 2025

- Lead-author of a paper focused on reducing the physical lookup table (P-LUT) utilisation of LUT-based neural network (NN) models
- Used C++ for decomposition with increased similarity from modifications to select output values: doi.org/10.1145/3706628.3708823
- Presented this paper to leading academics at ISFPGA 2025 in California and to the CAS research group at Imperial College London

RESEARCH EXPERIENCE

Early Exit Neural Networks — London, UK
Circuits and Systems Research Group, Imperial College London — Sep 2025-Present

- Writing a paper on hardware-aware early exits where exits are quantized to 1-bit while maintaining model accuracy to reduce latency
- Achieved up to 47% area-delay product reduction with minor accuracy loss across multiple datasets

NeuraLUT-Assemble: Hardware-aware Assembling of Sub-Neural Networks for Efficient LUT Inference — Zürich, Switzerland
Tutorial presented at Fast Machine Learning for Science Conference 2025 — May 2025-Sep 2025

- Helped deliver a tutorial on the evolution of LUT-based NNs detailing key challenges and presenting mitigation strategies for each
- Presented new in-context results using an imask-optimized input buffer for NeuraLUT-Assemble models showcasing best in class latency
- Implemented input layer as per-feature L-LUTs and demonstrated live end-to-end models on PYNQ with on-device pre/post-processing

Ultra-Low Latency ML Research — London, UK
Circuits and Systems Research Group, Imperial College London — Jun 2024-Present

- Adapted the open-source NeuraLUT toolflow to integrate Verilator testing, CUDA for improved inference and oh-my-xilinx Tcl scripts for synthesis of the model in Vivado along with modifying various L-LUT compression techniques and software.
- Worked with a team from TU Delft to map gate activation and timestep functions to decomposed L-LUTs as part of a brain model
- Implemented a latency-aware controller for a dynamic NN with early exits and runtime width selection on an ESP32

ACADEMIC PROJECTS

CMATMUL – Cache Based Matrix Multiplication Kernel — Feb 2025-Mar 2025

- Developed an effective C++ based matrix multiplication kernel leading to over a 100x increase in throughput from a naïve solution
- Optimised the throughput by implementing cache-aware tiling, register blocking, an AVX2 microkernel and OpenMP parallelism

Collabify — Mar 2024-Aug 2024

- Developed a collaborative Spotify web app with personalised recommendations using a weighted cosine similarity model and Spotify API
- Designed a production backend with SQL-based persistence, Supabase authentication, Stripe payment integration, and deployed to Render.

Remote Control Car from Logic — Dec 2022-Mar 2023

- Designed and built a remote control car using a crystal oscillator and filters for a RF pair, logic gates, counters and MOSFET H-bridges

ADDITIONAL

Technical Proficiencies: Advanced in C++, Python, PyTorch, HTML/CSS, SystemVerilog, Git; Proficient in C, JavaScript, SQL, Java, Groovy, Verilog, Verilator, React, Basic, Assembly languages, Flask, CUDA, Tcl, Terraform

Interests: 1500m/3K/5K competitive track running for the Thames Valley Harriers (2023-Present), Raced nationally for junior cycling development teams and trained weekly with British Cycling (2020-2023)