Samuel Chi Laam Wu

UK Citizen | wuchilaam@gmail.com | linkedin.com/in/wcl-samuel | cern1710.com

EDUCATION

The University of Edinburgh — Edinburgh, Scotland
MSc High Performance Computing with Data Science — Sep 2024 - Aug 2025

- **Grade:** With Distinction — 75% overall (top 5% of cohort)
- **Dissertation:** Parallel Tridiagonal Solvers for Neural Networks
- **Relevant courses:** Accelerators, HPC Architectures, ML at Scale, Machine Learning & Pattern Recognition

Lancaster University — Lancaster, England
BSc (Hons) Computer Science (Study Abroad) — Oct 2021 - Jun 2024

- **Grade:** First class honours — 80% overall (top 5% of cohort)
- **Dissertation:** Accelerated Symbol-level GRAND for High-Order Modulation (highest grade in cohort)
- **Study Abroad Year:** The Australian National University (81.4% average)

PROJECTS

MSc Dissertation — Parallel Tridiagonal Solvers for Neural Networks

- First reproducibility study of DeepPCR with custom Triton kernels across V100 and MI300x GPUs
- Validated results from original NeurIPS paper through **roofline analysis** and **scalability analysis**

AMD AI Sprint Hackathon

- Optimised Mistral-8x7B LLM inference using vLLM v1 on MI300x GPUs, **matching 2nd place solution**
- Used a Docker-based ROCm pipeline, achieving a **10% improvement in total token throughput**

Event-Based Parallel Brain Simulation Framework (benchmarked on Cirrus and ARCHER2)

- Scalable event-based coordination framework in C++ and MPI, achieving **26x speedup on 128 cores**
- Designed a thread-safe framework, enabling async multi-handler event dispatching and batched updates

Vision Transformer (ViT) for ERA5 Weather Classification (deployed on Cirrus)

- Trained a multi-node, multi-GPU ViT using PyTorch Distributed with MixUp augmentation on V100 GPUs
- Applied model pruning, quantisation, ZeRO, and tensor parallelism, achieving **28x speedup over baseline**

WORK EXPERIENCE

Backend Software Developer — Bailrigg, Lancaster
IT Partnering and Innovation at Lancaster University — Oct 2023 – Dec 2023

- **Optimised response speed by 35%** using Microsoft Orleans on AWS for a C# .NET booking service
- Refactored admin API, reducing latency and streamlining space management across organisations

PUBLIC SPEAKING

- Delivered lightning talk "HPC and ML workloads on Kubernetes" at Yorkshire DevOps to 100 attendees
- Presented "Adversarial Attacks on Aligned LLMs" at LUHack, introducing LLM safety and AI alignment

TECHNICAL SKILLS

- **Languages:** Proficient in Python; experience with C, C++, CUDA, C#, and Java
- **Frameworks & Libraries:** PyTorch, Triton, Hugging Face Transformers, vLLM, NumPy, OpenMP, MPI
- **Developer Tools:** Git, Docker, Gradle, CI/CD, Unix Command Line, TensorBoard, Slurm
- **Open-Source Contributions:** Contributor to llama.cpp, vLLM, NewPipe