Otto White

Website: ottowhite.com (find here Google Scholar, GitHub, YouTube, etc.) Email: otto.white20@imperial.ac.uk
LinkedIn: linkedin.com/in/otto-white Mobile: +44 74620 03706

Education

Imperial College London **October 2020 - June 2024**
- Computer Science MEng; First Class Honours

Hurstpierpoint College **Sept 2018 - June 2020**
- A-levels; A*A*A*A* (Maths, Further Maths, Computer Science, EPQ (Machine learning)); Full scholarship in Maths/Physics.

Experience

Imperial College London (Large Scale Data and Systems Group) **October 2024 - Present**
Doctoral Researcher
- Investigating efficient scheduling and resource allocation for dynamic LLM-applications such as RAG, inference-time scaling search algorithms, multi-turn chat and ReAct agents.
- Exploring efficient co-location of multiple local models, and exploiting heterogenous predictability of dynamic LLM-applications.
- Accelerating my research process with large-scale, hierarchical embedding databases of papers from ArXiv/systems and AI conferences.

Imperial College London **May 2024 - Present**
Cloud Infrastructure
- Creating a heterogenous bare-metal cloud of 100 machines serving 30 researchers, optimised for high-performance experimentation.
- Enabling fast networking, remote OS provisioning, scalable machine configuration, hardware monitoring, network-attached storage.

BidFX **April 2023 - August 2023**
Core Engineer
- Primarily responsible for design and implementation of the most complex component of a high-performance FIX pricing service, providing pricing to all downstream Foreign Exchange trading services.
- Consulted/paired with industry leaders to implement Continuous Delivery on our team; utilized thorough testing + TDD, pair programming, small+frequent PRs, incremental architecture and refactoring, optimizing test suites, builds, and pipelines for fast feedback, tickets+docs on the fly.
- Optimized a complex Java application for μs tail-latency, utilizing profiling, object pooling for garbage elimination, single-threading, NIO, many compression mechanisms. Exercising understanding and manipulating the JVM/OS/network/hardware interactions.

Cub3 Inc **June 2022 - August 2022**
Backend Engineer, Managed 1-2 Junior Devs
- Architected and wrote a gRPC backend in Typescript (including Solana blockchain services, Primsa DB, extensible permissioning system) enabling custodial wallet management and NFT collection creation.

Publications

Enabling Cloud-Scale Distributed Capabilities [Click for paper, click for presentation] **October 2024 - March 2025**
- Enabled high-performance, fine-grained access control at the cloud-scale with capability-based security, demonstrating capabilities as a promising access control approach for disaggregated cloud architectures.
- Explored decentralisation, sharding, and efficient distributed garbage collection; the first work on capability systems to achieve scalability, fault-tolerance and μs-scale access control operations simultaneously.
- Published in collaboration with Imperial College London, Microsoft and Intel in the first 6 months of the PhD.

Projects

ML Classifier on an FPGA SoC with High-level Synthesis (Vivado Toolchain) **January 2024 - March 2024**
- Implemented a highly parallel streaming-based Support Vector Machine with DMA, hierarchical pipelining, optimised hyperparameter storage layout, precision tuning and CORDIC for exponential approximation.
- Achieved the lowest latency of any Computing group by employing systematic benchmarking and analysis to target bottlenecks.

WACC - Compiler in Rust **January 2022 - March 2022**
- Wrote an efficient and memory-safe 7-stage optimizing compiler.
- Designed and implemented the four compiler stages for graph coloring register allocation, dramatically increasing execution speed of compiled executables through minimizing stack use; only group in the cohort to successfully implement.

Pintos - Operating System in C **October 2021 - December 2021**
- Various kernel extensions including schedulers, user processes/program facilitation, system calls, virtual memory with paging.
- Honed my design skills by creating the full virtual memory architecture, enabling productive delegation to different team members and execution. Additionally allowing me to crisis manage when two of four group members experienced burnout.
- Designed a fine-grained concurrent solution allowing shared access of global virtual memory resources by different user processes.

Skills Summary

Languages Rust, C, Java, Python, C++, Typescript, Javascript, Kotlin, Unix Scripting, English, Spanish (un poco)

Skills Continuous Delivery, TDD, Linux, Pipelines, Containerization, Git, VIM+Copilot+IDE Code Gen, Debugging, Scripting, JIRA, GCP

Interests Skateboarding, extreme sports, music, motorbikes, meditation, reading, socializing, travelling, finance, learning Spanish