Tom Stesco

tstesco | tstescoTT | tomstesco.com | tomstesco | tom.stesco@protonmail.com

PROGRAMMING

EXPERT
Python

PROFICIENT
C++ • SQL • Bash

TOOLS
PyTorch • vLLM
Airflow • Apache Spark
Kafka • PubSub • Protobuf
Docker • Kubernetes • Terraform
GCP • AWS • Azure
GNU/Linux • Git

EDUCATION

ETH ZURICH
MSc Integrated Building Systems
Excellence Scholar
Oct 2018 | Zurich, Switzerland

UNIVERSITY OF WATERLOO
BASc Honours Civil Engineering
Merit Scholarship
Jun 2016 | Waterloo, Canada

COURSEWORK

GRADUATE
Model Predictive Control
Mathematical Optimization
Computational Physics
Building Control and Automation
Computational Fluid Dynamics
Renewable Energy Technologies
Technology and Innovation Management
Innovation Leadership
Social Networks Research

UNDERGRADUATE
Software Engineering Design Project
Building Science
Bioprocess Engineering
Advanced Calculus
Probability and Statistics
Linear Algebra

EXPERIENCE

TENSTORRENT
Apr 2023 – present (2 years, 1 month) | Toronto, Canada

Senior Staff Field Application Engineer - AI/ML Software *Sep 2024 – present*
- Tech lead for open-source `tt-inference-server` project, 6+ engineer team delivering LLM end-to-end benchmarking, accuracy evals, and QA.
- Collaborating with customers and LLM optimization teams to validate and deploy new LLM architectures on next-gen AI accelerator hardware.
- Backend lead for `tt-studio` a GUI application for customers to run ML model demos on AI accelerators.

Staff Field Application Engineer - AI/ML Software *Apr 2023 – Sep 2024*
- Led development and deployment of cloud AI application demos (LLM chatbots, CNN object detection, ASR) and proof-of-concepts (e.g. API hosting) in partnership with engineering teams and contractors.
- Developed benchmarks transformer and CNN models on AI accelerators in `tt-buda-benchmarks` to track progress on performance and accuracy using graph compiler.
- Customer-facing engineering support to adapt transformer and CNN models in PyTorch for deployment on AI accelerator hardware.

BCG X
May 2022 – Apr 2023 (1 year) | Toronto, Canada

Senior AI/ML Engineer *May 2022 – Apr 2023*
- Led MLOps for pricing models in RD and production for large enterprise eCommerce: code reviews (teams of 3 to 5), test planning and validation, CI/CD, release management, monitoring, and on-call.
- Partnered with Data Scientists and customer Data Engineering teams in developing ETL, model training (XGBoost), and model deployment pipelines using Apache Spark and Airflow on GCP and Azure Databricks Lakehouse.

ECOBEE
Oct 2018 – May 2022 (3 years, 8 months) | Toronto, Canada

Senior Data Scientist *Jun 2021 – May 2022*
- Developed edge IoT device optimal control application (embedded C++) using ML models, resulting in proof of concept and patent application.

Data Scientist *Oct 2018 – Jun 2021*
- Developed HVAC controls physics-based simulation platform, partially open sourced as `building-controls-simulator`, and was presented at the eSim2020 conference.
- Scaled internal IoT experimentation platform from 0 to 150k+ devices running parameterized experiments by automating build, deployment, monitoring, and metrics with C++, Python, PubSub, and Protobuf.

RESEARCH

AUTOMATIC CONTROL LAB
Sep 2017 – Oct 2018 (1 year, 1 month) | Zurich, Switzerland

Master's Student *Sep 2017 – Oct 2018*
- MSc thesis: human-in-the-loop building control via MPC + online Bayesian ML. Supervisors: Prof. Dr. John Lygeros and Dr. Annika Eichler.
- Implemented Bayesian LDA, MCMC, and change-point algorithms for comfort occupancy prediction.
- Created Python library for large-scale control simulations on Euler HPC cluster.