# Tom Stesco

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

## PROGRAMMING

### LANGUAGES

Python • C++ • SQL • Bash

### TOOLS

vLLM • PyTorch
LangGraph • CrewAI
Airflow • Apache Spark
BigQuery • Firestore
Kafka • Pub/Sub • Protobuf
Docker • Kubernetes • Terraform
GCP • AWS • Azure
GNU/Linux • Git

## EDUCATION

### ETH ZURICH

MSc Integrated Building Systems
Excellence Scholarship & Opportunity Programme (ESOP)
Oct 2018 | Zurich, Switzerland

Automatic Control Lab
MSc thesis: `Humans in the Loop`
Grade: 5.75/6.0
Model Predictive Control (MPC) using online learning from human feedback.

### 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 (3 years, 1 month) | Toronto, Canada

**Senior Staff Engineer, Field Application - Machine Learning** *Sep 2024 – present*

- `tt-inference-server`: co-led cross-functional team of 20+ engineers delivering tt-inference-server, an open-source inference server deployment framework for Tenstorrent AI accelerators. Enabling customer and internal production AI workloads with integrated benchmarking and evaluation pipelines.
- Architected CI/CD and release system for inference serving software stack, enabling reliable production LLM deployments and automated regression detection. 18 releases completed, 60+ AI models supported (27 LLMs/VLMs) across 9 different hardware configurations.
- Partnered with enterprise customers to define AI system requirements, developed evaluation pipelines for agentic application integration with Tenstorrent hardware, and production deployment. Optimization of customer-specific LLM inference: quantization, latency/throughput tuning, structured outputs, sampling parameters.

**Staff Engineer, Field Application - Machine Learning** *Apr 2023 – Sep 2024*

- Prototyped cloud deployment for internet-facing LLM inference servers running vLLM for enterprise customers, using Docker, AWS EKS, ELB, Azure API Gateway, and Terraform.
- `tt-studio`: Built open-source customer reference AI applications using Docker, Django, LangChain and self-hosting LLM inference APIs.
- `tt-buda-benchmarks`: Developed benchmarking and evaluation pipelines for LLM and CNN models to track model performance and accuracy across AI graph compiler releases.

### BCG X

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

**Senior AI Software Engineer** *May 2022 – Apr 2023*

- Led production engineering for petabyte-scale pricing optimization models on Databricks Lakehouse. Internal ETL library enhancements, CI/CD and release process integration with secure artifactory, deployment using Databricks Jobs.
- Led MLOps for time-series XGBoost pricing model lifecycle management: ETL, training, validation, and deployment in production, monitoring, on-call. Using Airflow, Apache Spark, and BigQuery in GCP.

### ECOBEE

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

**Senior Data Scientist** *Jun 2021 – May 2022*

- `US12241649B2`: Invented and implemented edge ML optimal HVAC control system in embedded C++ using server trained state-space models and hybrid Model Predictive Control (MPC) approach to optimize energy cost over forecasted: energy price, weather, occupancy, heat gains. Resulting in digital-twin/real-world experiments and granted patent.
- `building-controls-simulator`: developed open source HVAC controls digital-twin platform to validate control algorithm in simulation. Parametric physics models were automatically fit using real data. Presented at eSim2020 conference.

**Data Scientist** *Oct 2018 – Jun 2021*

- Scaled internal IoT experimentation and telemetry platform from 0 to 150k+ devices running parameterized experiments by automating build, embedded experiment deployment, telemetry, and monitoring. Deployed in GCP and on embedded devices using C++, Protobuf, Pub/Sub, Big Query, and Firestore.