

# Tom Stesco

🌐 tstesco | 🌐 tstescoTT | 🏠 tomstesco.com | in tomstesco | ✉ tom.stesco@protonmail.com

## PROGRAMMING

### EXPERT

Python

### PROFICIENT

C++ • SQL • Bash

### TOOLS

vLLM • PyTorch

Airflow • Apache Spark

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

Model Predictive Control (MPC) using  
online learning from human feedback.

## 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) | Toronto, Canada

#### Senior Staff Field Application Engineer - AI/ML Software

Sep 2024 – present

- [tt-inference-server](#): co-lead of 20+ engineer cross-functional team developing open-source LLM inference serving and deployment framework for Tenstorrent AI accelerators with integrated benchmarking and evaluations.
- Architected CI/CD and release system for the 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.
- Collaborated with customers' engineering teams to validate and optimize LLM implementations, e.g. quantization, latency/throughput tuning, structured outputs, for production deployment of AI applications. For example, agentic document processing with OCR, multi-modal inputs, and function calling.

#### Staff Field Application Engineer - AI/ML Software

Apr 2023 – Sep 2024

- Prototyped cloud infra for internet-facing deployment of 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. e.g. using LangChain and self-hosting LLM inference APIs.
- [tt-buda-benchmarks](#): Designed 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/ML Engineer

May 2022 – Apr 2023

- Led production engineering for petabyte-scale pricing optimization models on Azure 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 that fits parametric models using real data (terabyte scale) and was used to validate control algorithm changes in simulation. 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. Using C++, Protobuf, Pub/Sub, Big Query, and Python.