

Tom Stesco

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

PROGRAMMING

LANGUAGES

Python • C++ • SQL • Bash

TOOLS

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

EDUCATION

ETH ZURICH

MSC INTEGRATED BUILDING SYSTEMS

Excellence Scholarship & Opportunity Programme (ESOP)
Oct 2018 | Zurich, Switzerland

AUTOMATIC CONTROL LAB

Thesis: [Humans in the Loop](#)
Grade: 5.75/6.0
Online learning from human feedback in model predictive control.

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
Technology and Innovation Management
Innovation Leadership
Social Networks Research

UNDERGRADUATE

Software Engineering Design Project
Advanced Calculus
Probability and Statistics
Linear Algebra

EXPERIENCE

TENSTORRENT

Apr 2023 – present (3 years, 5 month) | Toronto, Canada

Senior Manager, Forward Deployed Engineering

Jul 2026 – present

- Implemented pre/post-sales life cycle processes for customers deploying AI inference clusters across 7+ teams and 5 geographies. Built agentic system using LangGraph to coordinate delivery and compute resource scheduling for demo, proof of concept, and production deployments.
- Architected customer kubernetes clusters for AI inference serving and development using tt-operator, ansible, helm, and [cloud native tooling](#).

Senior Staff Engineer, Field Applications - Machine Learning

Sep 2024 – Jul 2026

- [tt-inference-server](#): co-led cross-functional team of 20+ engineers delivering tt-inference-server, an open-source inference server deployment and benchmarking framework for Tenstorrent AI accelerators.
- Designed CI test and release system for inference serving stack with 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 for agentic application integration with Tenstorrent hardware. Customer-specific optimization of LLM inference for latency/throughput/cost via hardware configuration, batch size, parallelism.

Staff Engineer, Field Applications - Machine Learning

Apr 2023 – Sep 2024

- Prototyped cloud deployments of vLLM with LLMs running on Tenstorrent ASICs. using Docker, AWS EKS, ELB, Azure API Gateway, and Terraform.
- [tt-studio](#): Built backend for open-source customer reference AI applications using Docker, Django, LangChain and self-hosting LLM inference APIs.
- [tt-buda-benchmarks](#): benchmarking and evaluations for LLM and CNN models to track model performance and accuracy across AI 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.
- 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

- Inventor of US patent [US12241649B2](#) in hybrid edge ML Model Predictive Control system for HVAC, optimizing energy cost over forecasted energy price, weather, and occupancy. Implemented real-world system in embedded C++ using trained autoregressive state-space models and conducted experiments.
- [building-controls-simulator](#): developed open source HVAC controls digital-twin platform. Physics models automatically fit and tuned using real sensor data. Presented at eSim2020 conference.

Data Scientist

Oct 2018 – Jun 2021

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