FOTEINI STRATI

foteini.strati@inf.ethz.ch
www.linkedin.com/in/foteini-strati ◇ https://fotstrt.github.io/

EDUCATION

ETH Zurich, Switzerland Nov. 2021 - now
PhD in Computer Science
EASL research group
Advisor: Prof. Ana Klimovic
Research Interests: Systems for Machine Learning, Cloud Computing
Thesis Topic: Increasing resource utilization and fault-tolerance for machine learning workloads

ETH Zurich, Switzerland Sep. 2019 - Sep. 2021
MSc in Computer Science
90 ECTS, GPA: 5.29/6.0
Thesis: Characterising Resource Elasticity and Fault Tolerance in Distributed Machine Learning

National Technical University of Athens, ECE School, Greece Dec. 2013 - Feb. 2019
Diploma in Electrical and Computer Engineering
300 ECTS, GPA: 9.02/10
Major in Computer Systems and Software
Thesis: Study and design of concurrent priority queues for NUMA architectures

PUBLICATIONS

- **Foteini Strati**, Zhendong Zhang, George Manos, Ixeia Sánchez Périz, Qinghao Hu, Tiancheng Chen, Berk Buzcu, Song Han, Pamela Delgado, Ana Klimovic, Sailor: Automating Distributed Training over Dynamic, Heterogeneous, and Geo-distributed Clusters, SOSP 2025 *(To appear)*
- Paul Elvinger, **Foteini Strati**, Natalie Enright Jerger, Ana Klimovic, Measuring GPU utilization one level deeper
- **Foteini Strati***, Michal Friedman*, Ana Klimovic, PCcheck: Persistent Concurrent Checkpointing for ML, ASPLOS 2025
- **Foteini Strati**, Sara Mcallister, Amar Phanishayee, Jakub Tarnawski, Ana Klimovic, DéjàVu: KV-cache Streaming for Fast, Fault-tolerant Generative LLM Serving , ICML 2024
- **Foteini Strati**, Paul Elvinger, Tolga Kerimoglu, Ana Klimovic, ML Training with Cloud GPU Shortages: Is Cross-Region the Answer?, EuroMLSys 2024
- **Foteini Strati**, Xianzhe Ma, Ana Klimovic, Orion: Interference-aware, Fine-grained GPU Sharing for ML Applications, EuroSys 2024
- Maximilian Böther, **Foteini Strati**, Viktor Gsteiger, Ana Klimovic, Towards A Platform and Benchmark Suite for Model Training on Dynamic Datasets, EuroMLSys 2023
- Joel Andre*, **Foteini Strati***, Ana Klimovic, Exploring Learning Rate Scaling Rules for Distributed ML Training on Transient Resources, DistributedML 2022
- **Foteini Strati***, Christina Giannoula*, Dimitrios Siakavaras, Georgios Goumas, Nectarios Koziris, An Adaptive Concurrent Priority Queue for NUMA Architectures, ACM International Conference on Computing Frontiers, 2019

INDUSTRY EXPERIENCE

Meta, AI and Systems Co-design, Bellevue, US June 2025 - September 2025
Research Intern
Mentor: Amar Phanishayee

Microsoft Research, Redmond, US June 2023 - September 2023
Research Intern
Mentor: Amar Phanishayee

- Developed techniques to improve performance in Generative Large Language Model serving.

NVIDIA, Switzerland June 2022 - September 2022
Software Engineering Intern
Mentors: Eric Hall and Ville Kallioniemi

- Implemented and analyzed the impact of resource elasticity in distributed ML training for autonomous driving workflows.

Huawei Zurich Research Center, Switzerland Sep. 2020 - Feb. 2021
Cloud Architectures Research Intern
Mentors: Bill McColl and Albert-Jan Yzelman

- Fault-tolerant programming models and systems for cloud and HPC applications.

Centaur Analytics, Athens, Greece
Junior Software Engineer Jan. 2019 - Aug. 2019

- CO_2 emission forecast in silos with time series analysis and genetic algorithms.

AWARDS

- ML and Systems Rising Stars July 2024
- ETH Medal 2022 for outstanding Master's thesis Feb 2022
- NTUA Thomaidio Award for paper publication in international conference June 2020

PROGRAMMING SKILLS

C, C++, CUDA, Python, Assembly (8086), PyTorch, Ray, Kubernetes, MPI, OpenMP, Git, Unix, LaTeX

TEACHING EXPERIENCE

ETH Zurich, Teaching Assistant

Cloud Computing Architecture	2022-2025
Systems Programming and Computer Architecture	2022, 2024
Seminar on Machine Learning Systems	2022, 2023

ETH Zurich, Project Mentorship

Zhendong Zhang: *Reducing Energy Consumption in ML workloads via Power-Aware Scheduling* (MSc Thesis) (ongoing)

Carlos Serrano Fernandez: *Proactive approaches for large-scale distributed training over spot VMs* (Msc Thesis) (ongoing)

Leo Stephan (co-supervision with Paul Elvinger): *Towards Efficient GPU Sharing: An Analytical Model for Kernel DRAM and L2 Cache Interference Estimation* (Bachelor thesis) 2025

Lennart Schulz: *Evaluating GPU Partitioning Mechanisms for Resource Sharing with LLM Inference Workloads* (Semester project) 2025

Rongzhi Li: *Evaluating LLM serving optimizations for dynamic workloads* (Semester project) 2024

Jonathan Smith (co-supervision with Xiaozhe Yao): *Evaluating LLM serving performance on the Grace Hopper superchip* (Semester project) 2024

George Manos: *Studying and optimizing geo-distributed training in the public cloud* (Semester project) 2024

Zhendong Zhang: *Evaluating operator-level parallelization planners for large-scale distributed training* (Semester project) 2024

Paul Elvinger: Towards resource and interference-aware scheduling of ML workloads, (MSc Thesis) 2024

Ixeia Sánchez Périz: *Towards optimal resource allocation and communication schedule for ML training in the public cloud*, (MSc Thesis) 2024

Carlos Serrano Fernandez: *Resource utilization analysis of Large Language Models* (Semester project) 2024

Paul Elvinger, Tolga Kerimoglu: *Studying and enabling efficient ML training across datacenters* (Semester project) 2023

Xianzhe Ma: *Evaluating GPU sharing policies for ML workloads* (Semester project) 2023

Xindi Zuo (co-supervision with Michal Friedman): DMA for Non-Volatile Memory (MSc thesis) 2023

Jingyi Zhu: *Evaluating the performance of NCCL collectives in the cloud* (Semester project) 2022

Joel André: Accurate, elastic large-scale distributed training over transient resources (BA Thesis) 2022

National Technical University of Athens, Lab Assistant
Operating Systems Feb. 2018 - June 2018
Introduction to Programming Sep. 2016 - Feb. 2017

SERVICE

- OSDI '22 and ATC '22 artifact evaluation committee April 2022 - June 2022
- TTODLer-FM'25 technical program committee (colocated with ICML'25) July 2025