Kostis Kaffes

Email kkaffes@cs.columbia.edu
Website https://www.cs.columbia.edu/~kkaffes/

Professional Experience

July 2023 - now **Columbia University**, New York
Assistant Professor in Computer Science

Aug 2022 - June 2023 **Google Inc.**, Sunnyvale, CA
Software Engineer

SystemsResearch@Google (SRG): Focusing on the future of hyperscalar systems design.

Oct 2019 - Dec 2019 **Google Inc.**, Sunnyvale, CA
Student Researcher

Borglet: Reduced data center power consumption by 9% through application-class-aware dynamic frequency scaling.

Jun 2019 - Sept 2019 **Google Inc.**, Sunnyvale, CA
Software Engineering PhD Intern

NetInfra: Analysis of scheduling antagonists and potential mitigations for low latency networked tasks.

Jun 2018 - Sept 2018 **Google Inc.**, Sunnyvale, CA
Software Engineering PhD Intern

Borglet: Minimized interference among different classes of tasks running in datacenters.

Jun 2017 - Sept 2017 **Google Inc.**, Sunnyvale, CA
Software Engineering PhD Intern

Cloud Network and Network Systems: Improved the efficiency of the network stack by using dynamic scheduling.

Nov 2015 - July 2016 **Arrikto Inc.**
Software Engineer

Extended Arrikto's storage platform to accommodate container technology. Designed a Docker storage driver that minimized data movement.

Education

Sept 2016 - Jun 2022 **Stanford University**
PhD in Electrical Engineering

Thesis Topic: Flexible and Performant Scheduling Across the Stack
Thesis Advisor: Professor Christos Kozyrakis

Sept 2016 - June 2018 **Stanford University**
M.Sc. in Electrical Engineering

Sept 2010- Oct 2015 **National Technical University of Athens (NTUA), Greece**
Diploma in Electrical and Computer Engineering (5-year degree)

Publications

Georgios Liargkovas, Vahab Jabrayilov, Hubertus Franke, **Kostis Kaffes**
An Expert in Residence: LLM Agents for Always-On Operating System Tuning
Workshop on ML for Systems at NeurIPS (ML4Sys 2025)

Jiakai Xu, Tianle Zhou, Eugene Wu, **Kostis Kaffes**
Toward Systems Foundations for Agentic Exploration
1st Workshop on Systems for Agentic AI (SAA '25)

Nikos Pagonas, Yeounoh Chung, **Kostis Kaffes**, Arvind Krishnamurthy
Cortex: Workflow-Aware Resource Pooling and Scheduling for Agentic Serving
1st Workshop on Systems for Agentic AI (SAA '25)

Georgios Liargkovas, Prabhpreet Singh Sodhi, **Kostis Kaffes**
Set It and Forget It: Zero-Mod ML Magic for Linux Tuning
Practical Adoption Challenges of ML for Systems (PACMI'25)

Prabhpreet Singh Sodhi, Georgios Liargkovas, **Kostis Kaffes**
Empowering machine-learning assisted kernel decisions with eBPFML
3rd Workshop on eBPF and Kernel Extensions, eBPF 2025

Jack Tigar Humphries, Neel Natu, **Kostis Kaffes**, Stanko Novaković, Paul Turner, Hank Levy, David Culler, Christos Kozyrakis
Wave: Offloading Resource Management to SmartNIC Cores
ACM International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2025

Alireza Sanaee, Vahab Jabrayilov, Ilias Marinos, Anuj Kalia, Divyanshu Saxena, Prateesh Goyal, **Kostis Kaffes**, Gianni Antichi
Fast Userspace Networking for the Rest of Us
arXiv, https://arxiv.org/abs/2502.09281, 2025

Prasoon Sinha, **Kostis Kaffes**, Neeraja J Yadwadkar
Shabari: Delayed Decision-Making for Faster and Efficient Serverless Functions
arXiv, https://arxiv.org/abs/2401.08859, 2024

Vishwanath Seshagiri, Abhinav Gupta, Vahab Jabrayilov, Avani Wildani, **Kostis Kaffes**
Rethinking the Networking Stack for Serverless Environments: A Sidecar Approach
ACM Symposium on Cloud Computing, SoCC 2024

Yueying Li, Nikita Lazarev, David Koufaty, Yijun Yin, Andy Anderson, Zhiru Zhang, Edward Suh, **Kostis Kaffes**, Christina Delimitrou
LibPreemptible: Enabling Fast, Adaptive, and Hardware-Assisted User-Space Scheduling
IEEE International Symposium on High Performance Computer Architecture, HPCA 2024

Prasoon Sinha, **Kostis Kaffes**, Neeraja J Yadwadkar
Online Learning for Right-Sizing Serverless Functions
Architecture and System Support for Transformer Models, ASSYST@ ISCA 2023

Peter Kraft, Qian Li, **Kostis Kaffes**, Athinagoras Skiadopoulos, Deeptaanshu Kumar, Danny Cho, Jason Li, Robert Redmond, Nathan Weckwerth, Brian Xia, Peter Bailis, Michael Cafarella, Goetz Graefe, Jeremy Kepner, Christos Kozyrakis, Michael Stonebraker, Lalith Suresh, Xiangyao Yu, Matei Zaharia
Apiary: A DBMS-Integrated Transactional Function-as-a-Service Framework
arXiv, https://arxiv.org/abs/2208.13068, 2023

Kostis Kaffes, Neeraja. J. Yadwadkar, Christos Kozyrakis
Hermod: Principled and Practical Scheduling for Serverless Functions

ACM Symposium on Cloud Computing, SoCC 2022

Athinagoras Skiadopoulos*, Qian Li*, Peter Kraft*, **Kostis Kaffes***, Daniel Hong, Shana Mathew, David Bestor, Michael Cafarella, Vijay Gadepally, Goetz Graefe, Jeremy Kepner, Christos Kozyrakis, Tim Kraska, Michael Stonebraker, Lalith Suresh, Matei Zaharia
DBOS: A DBMS-oriented Operating System
International Conference on Very Large Data Bases, VLDB 2022

Qian Li*, Peter Kraft*, **Kostis Kaffes***, Athinagoras Skiadopoulos, Deeptaanshu Kumar, Jason Li, Michael Cafarella, Goetz Graefe, Jeremy Kepner, Christos Kozyrakis, Michael Stonebraker, Lalith Suresh, Matei Zaharia
A Progress Report on DBOS: A Database-oriented Operating System
Conference on Innovative Data Systems Research, CIDR 2022

Kostis Kaffes, Jack Tigar Humphries, David Mazières, Christos Kozyrakis
Syrup: User-Defined Scheduling across the Stack
28th ACM Symposium on Operating Systems Principles, SOSP 2021

Jack Tigar Humphries*, **Kostis Kaffes***, David Mazières, Christos Kozyrakis
A case against (most) context switches
18th Workshop on Hot Topics in Operating Systems, HotOS 2021

Hang Zhu, **Kostis Kaffes**, Zixu Chen, Zhenming Liu, Christos Kozyrakis, Ion Stoica, Xin Jin
RackSched: A Microsecond-Scale Scheduler for Rack-Scale Computers
14th USENIX Symposium on Operating Systems Design and Implementation, OSDI'20

Kostis Kaffes, Dragos Sbirlea, Yiyan Lin, David Lo, Christos Kozyrakis
Leveraging Application Classes to Save Power in Highly-Utilized Data Centers
ACM Symposium on Cloud Computing, SoCC 2020

Kostis Kaffes, Neeraja Yadwadkar, Christos Kozyrakis
Centralized Core-granular Scheduling for Serverless Functions
ACM Symposium on Cloud Computing, SoCC 2019

Jack Tigar Humphries, **Kostis Kaffes**, David Mazières, Christos Kozyrakis
Mind the Gap: A Case for Informed Request Scheduling at the NIC
18th ACM Workshop on Hot Topics in Networks, HotNets 2019

Kostis Kaffes, Timothy Chong, Jack Tigar Humphries, Adam Belay, David Mazières, Christos Kozyrakis
Shinjuku: Preemptive Scheduling for μsecond-scale Tail Latency
16th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2019

Teaching Experience

Spring 2025 Columbia University, USA
COMS W4118: Operating Systems I
COMS E6113: Agents Made Real
COMS E9902: Frontiers of CS Research

Fall 2024 Columbia University, USA
COMS E6998: Topics in Cloud Computing

Spring 2024 Columbia University, USA
COMS W4118: Operating Systems I

Fall 2023 Columbia University, USA
COMS E6998: Topics in Cloud Computing

Professional Service

Organizer: SoCC 2023, SIGCOMM 2023, HotNets 2022, Eurosys 2022

Program Committee: Eurosys 2026, MICRO 2026, PACMI 2025, SEA 2025, eBPF 2025, ATC 2025, ASPLOS 2025, SOSP 2024 Posters, ATC 2024, NSDI 2024, EdgeSys 2024, SESAME 2024, SoCC 2023, SoCC 2022

University Service

Organizer for Graduate Student Visit Day: 2025, Department level
Faculty Recruiting Committee: 2024-now, Department level
EmpireAI Columbia Working Group: 2024-now, University level

Advising

PhDs
Edward Guo, joined 2025 (co-advised with Asaf Cidon)
Elaine Ang, joined 2025 (co-advised with Eugene Wu)
Chenxi Huang, joined 2025 (co-advised with Eugene Wu, Junfeng Yang, Baishakhi Ray)
Meghna Pancholi, joined 2025
Georgios Liargkovas, joined 2024
Nikos Pagonas, joined 2024
Vahab Jabrayilov, joined 2023

Thesis Committee
Kelly Kostopoulou, *Optimizing Privacy Budget Management in Differentially Private Systems*
Haoyu Li, *Unlocking Storage Performance: A Systems Approach to Kernel Bypass, Replication, and Caching*
Yannis Zarkadas, *Optimizing Memory and Storage Performance in Cloud Datacenters*
Abhishek Shah, *Adaptive Sampling for Targeted Software Testing*

Master Students:
Jiakai (Alex) Xu
Prabhpreet Singh Sodhi, graduated 2025, Goldman Sachs
Abhinav Gupta, graduated 2024, Meta

Undergraduates:
Ruizhe Fu, graduated 2025, Google
Nicholas Zhe Kai Yap, graduated 2025, Princeton University
Camerie Mazreku, graduated 2024, Netflix

Awards

06/2025	Google ML and Systems Junior Faculty Award
2018-2020	Leventis Foundation Fellowship
2018-2019	Gerondelis Foundation Fellowship
09/2016	Stanford EE Departmental Fellowship

Grants and Gifts

07/2025-now	Columbia University Data, Agents, and Processes Lab (DAPlab) raised **$600K+/year**
10/2025-09/27	NSF Collaborative Research: Ideas Lab: Breaking Low, SP, **$810K** *End-to-End Delivery Technology for Interactive Multi-person XR Rehabilitation Activities*
07/2025-08/2026	Google ML and Systems Junior Faculty Award, PI, **$100k** *Scheduling for Tail Latency*

07/2025-08/2026 Columbia-Dream Sports AI Innovation Center, PI, **$100k**
From Noisy Signals to Clear Decisions: Optimizing Long-Term Outcomes with Relative Feedback and Causal Inference

07/2025-08/2026 Columbia Center of AI Technology in collaboration with Amazon, PI, **$100k**
Benchmarking Agent Reliability: Integrating Side-Effect and Latency Metrics

07/2020-08/2021 Facebook Research Award on Networking, **$50k**
Flexible, Practical, and End-to-End Scheduling for Networked Applications