

Keisuke Kamahori

 kamahori.org •  kamahori@cs.washington.edu •  kamahori •  kamahori

EDUCATION

University of Washington

Seattle, WA

Ph.D. in Computer Science & Engineering

Jun 2023 – May 2028 (Expected)

- Advisor: Prof. Baris Kasikci.

The University of Tokyo

Tokyo, Japan

B.Sc. in Information Science

Apr 2019 – Mar 2023

- Advisor: Prof. Shinya Takamaeda-Yamazaki.

PUBLICATIONS

Google Scholar: <https://scholar.google.com/citations?user=IMZC2hYAAAAJ>

Semantic Scholar: <https://www.semanticscholar.org/author/Keisuke-Kamahori/2283847574>

(* denotes equal contribution)

Preprints

[P1] **VibeServe: Can AI Agents Build Bespoke LLM Serving Systems?**

Keisuke Kamahori*, Shihang Li*, Simon Peter, Baris Kasikci.

Preprint, under submission. (Presented at DL4C Workshop @ ICML 2026, July 2026.)

[P2] **M*: A Modular, Extensible, Serving System for Multimodal Models.**

Atindra Jha*, Naomi Sagan*, Keisuke Kamahori†, Irmak Sivgin†, Rohan Sanda, Steven Gao, Mark Horowitz, Luke Zettlemoyer, Olivia Hsu, Jure Leskovec, Baris Kasikci, Stephanie Wang.

Preprint, under submission.

[P3] **Murmur: An Efficient Inference System for Long-Form ASR.**

Wei-Tzu Lee, Keisuke Kamahori, Baris Kasikci.

Preprint, under submission. (Presented at MLSys 2026 YPS, May 2026.)

[P4] **VoxServe: Streaming-Centric Serving System for Speech Language Models.**

Keisuke Kamahori, Wei-Tzu Lee, Atindra Jha, Rohan Kadekodi, Stephanie Wang, Arvind Krishnamurthy, Baris Kasikci.

Preprint, under submission.

[P5] **AgentFlux: Decoupled Fine-Tuning & Inference for On-Device Agentic Systems.**

Rohan Kadekodi*, Zhan Jin*, Keisuke Kamahori, Yile Gu, Sean Khatiri, Noah H. Bayindirli, Sergey Gorbunov, Baris Kasikci.

Preprint, under submission.

[P6] **[Title omitted to maintain anonymity] (Topic: profile-guided optimization for CPU prefetchers).**

Keisuke Kamahori et al.

Under submission.

Peer-Reviewed Conferences

- [C1] **ConsumerBench: Benchmarking Generative AI Applications on End-User Devices.**
Yile Gu*, Rohan Kadekodi*, Hoang Nguyen, Keisuke Kamahori, Yiyu Liu, Baris Kasikci.
COLM 2026, October 2026.
- [C2] **TeleRAG: Efficient Retrieval-Augmented Generation Inference with Lookahead Retrieval.**
Chien-Yu Lin*, Keisuke Kamahori*, Yiyu Liu*, Xiaoxiang Shi, Madhav Kashyap, Yile Gu, Rulin Shao, Zihao Ye, Kan Zhu, Stephanie Wang, Arvind Krishnamurthy, Rohan Kadekodi, Luis Ceze, Baris Kasikci.
MLSys 2026, May 2026.
- [C3] **LiteASR: Efficient Automatic Speech Recognition with Low-Rank Approximation.**
Keisuke Kamahori, Jungo Kasai, Noriyuki Kojima, Baris Kasikci.
EMNLP 2025 (Main), November 2025. (Featured on Hugging Face's Open ASR Leaderboard.)
- [C4] **NanoFlow: Towards Optimal Large Language Model Serving Throughput.**
Kan Zhu, Yufei Gao, Yilong Zhao, Liangyu Zhao, Gefei Zuo, Yile Gu, Dedong Xie, Tian Tang, Qinyu Xu, Zihao Ye, Keisuke Kamahori, Chien-Yu Lin, Ziren Wang, Stephanie Wang, Arvind Krishnamurthy, Baris Kasikci.
OSDI 2025, July 2025.
- [C5] **Fiddler: CPU-GPU Orchestration for Fast Inference of Mixture-of-Experts Models.**
Keisuke Kamahori*, Tian Tang*, Yile Gu, Kan Zhu, Baris Kasikci.
ICLR 2025, April 2025. (Also presented at PML4LRS @ ICLR 2024, May 2024.)
- [C6] **Manticore: Hardware-Accelerated RTL Simulation with Static Bulk-Synchronous Parallelism.**
Mahyar Emami*, Sahand Kashani*, Keisuke Kamahori, Mohammad Sepehr Pourghannad, Ritik Raj, James R. Larus.
ASPLOS 2024, April 2024.
- [C7] **A 475 MHz FPGA Accelerator for RTL Simulation.**
Sahand Kashani*, Mahyar Emami*, Keisuke Kamahori, Mohammad Sepehr Pourghannad, Ritik Raj, James R. Larus.
FPGA 2024, March 2024.
- [C8] **CiraaS: cloud computing with programmable logic.**
Kenji Tanaka, Yuki Arikawa, Tsuyoshi Ito, Yuki Matsuda, Keisuke Kamahori, Shinya Kaji, Takeshi Sakamoto.
SIGCOMM 2022 Poster and Demo Sessions, August 2022.
- [C9] **Accelerating Decision Tree Ensemble with Guided Branch Approximation.**
Keisuke Kamahori, Shinya Takamaeda-Yamazaki.
HEART 2022, June 2022.
- [C10] **GBA: Guided Branch Approximation.**
Keisuke Kamahori, Shinya Takamaeda-Yamazaki.
YArch 2022 (Co-located with ASPLOS 2022), March 2022.

EXPERIENCE

University of Washington

Ph.D. Student, Systems for Future Intelligence (SyFI) Lab

Seattle, WA

Jun 2023 – Present

- Supervisor: Prof. Baris Kasikci.
- **Agents and Systems:** Developing an AI agent system for serving system development (VibeServe [P1]); an efficient fine-tuning framework for local AI agents (AgentFlux [P5]); part of the Self-Defining Systems project.
- **Systems for Multimodal Models:** Developing a universal serving system for multimodal models (M* [P2]); a streaming-centric serving system for speech LLMs (VoxServe [P4]); a compression technique for speech models (LiteASR [C3]); collaborating with AI4DeafBlind to enable fast ASR inference on low-resource accessibility devices.
- **Local AI:** Led projects on inference acceleration for RAG (TeleRAG [C2]) and MoE models (Fiddler [C5]) under

resource constraints; benchmarked on-device inference (ConsumerBench [C1]).

- **Computer Architecture:** Designed profile-guided optimization techniques for data center CPU prefetchers (under submission [P6]).

NVIDIA

Architect Intern

Santa Clara, CA

Jun 2026 – Sep 2026

- Supervisor: Joydeep Ray.
- Working on architectural modeling and simulation for future GPU generations optimized for AI inference.

Kotoba Technologies Inc.

Seattle, WA / Tokyo, Japan

Founding Engineer

Jun 2024 – Jul 2025

- One of the earliest engineers at a startup building speech AI (backed by USD 23M in seed funding), with a focus on real-time speech-to-speech translation.
- Led the design and implementation of the serving infrastructure powering DOTSU, the company's core simultaneous translation product (500K+ sessions in three months), and APIs (ASR, TTS, STS), involving full-stack optimization tailored to the unique duplex architecture and strict latency requirements of real-time speech LLMs.

Fixstars Corporation

FPGA Engineer

Tokyo, Japan

Aug 2021 – May 2023

- Supervisor: Shinya Kaji.
- Developed an FPGA accelerator for distributed 3D Fast Fourier Transform workloads.
- Contributed to a research project on FPGA SmartNICs for data centers [C8].

The University of Tokyo

Undergraduate Researcher, Laboratory for Computer Architecture and Systems

Tokyo, Japan

Aug 2021 – Mar 2023

- Supervisor: Prof. Shinya Takamaeda-Yamazaki.
- Designed an acceleration method based on microarchitectural support for approximate computing [C9], [C10].

École Polytechnique Fédérale de Lausanne

Research Intern, Very Large Scale Computing Lab.

Lausanne, Switzerland

Jul 2022 – Sep 2022

- Supervisor: Prof. James Larus.
- Designed an FPGA accelerator for hardware simulation [C6], [C7].

Nara Institute of Science and Technology

Research Intern, Computing Architecture Lab.

Nara, Japan

Feb 2022 – Mar 2022

- Supervisors: Prof. Yasuhiko Nakashima and Prof. Renyuan Zhang.
- Worked on lightweight neural network processing with stochastic computing.

The University of Tokyo

Research Intern, Matsuo Lab.

Tokyo, Japan

Dec 2020 – Mar 2022

- Supervisors: Prof. Yutaka Matsuo and Ryo Okada.
- Worked on extracting human-interpretable knowledge from high-dimensional deep RL models.

Nagase Brothers Inc.

AI Engineer

Tokyo, Japan

May 2020 – Aug 2021

- Worked on developing software using deep learning techniques for educational services.

ROOT, HEP Software Foundation

Google Summer of Code Student

Remote

May 2020 – Aug 2020

- Supervisor: Dr. Oksana Shadura.
- Contributed to an open-source data analysis framework for the high-energy physics community in a Google Summer of Code project titled "Pre-conditioners applied to ROOT compression algorithms."

AWARDS AND GRANTS

- MLSys 2026 Competition - NVIDIA Track (FlashInfer AI Kernel Generation Contest) May 2026
1st place (GDN Track, Full-Agent), 2nd place (GDN Track, Agent-Assisted), and 3rd place (DSA Track, Full-Agent)
Awarded NVIDIA DGX Spark, RTX 5090, and RTX 5080
- MLSys 2026 Student Travel Grant May 2026
- Modal for Academics Grant (USD 10,000) Apr 2026
- Tinker Research Grant from Thinking Machines (USD 5,000) Feb 2026
- Graduate Student Conference Presentation Awards from UW (USD 500) Nov 2025
- Allen School Travel Grant from UW CSE (USD 500) May 2024
- **Ph.D. Scholarship from Toyota Physical and Chemical Research Institute**
Full tuition/insurance fee and stipend for two years (JPY 10,000,000/year) 2023 – 2025
Additional financial aid of USD 6,000/year 2025 – Present
- ISCA 2022 Registration Grant from 4th uArch workshop Jun 2022
- ISCA 2021 Registration Grant from 3rd uArch workshop Jun 2021
- Travel Grant to Chulalongkorn University by JASSO Aug 2019
- Gold Award, Chemistry Grand Prix 2018
- Rank A Award, Japan Mathematical Olympiad 2018
- Rank A Award, Japan Junior Mathematical Olympiad 2016

MENTORING

- Xikai Meng (UW ECE) Spring 2026 – Present
M and VibeServe*
- Garv Ghai (Georgia Tech) Spring 2026 – Present
VoxServe and multimodal serving
- Wei-Tzu (Ruby) Lee (UW ECE → Software Engineer at AWS) Fall 2025 – Present
First author of Murmur (MLSys 2026 YPS)
- Tian Tang (Tsinghua University → MTS at Nex) Spring – Summer 2024
Co-first author of Fiddler (ICLR 2025)

TEACHING

- TA, CSE 554 Machine Learning Systems (Winter 2026)

TALKS

- **VibeServe: Can AI Agents Build Bespoke LLM Serving Systems?**
At NVIDIA TRT-LLM Team, June 2026.
At UW Mobile Intelligence Lab., May 2026.
- **Practical FPGA Development Seminar**
At Fixstars FPGA Seminar vol. 10, May 2022.
At Fixstars FPGA Seminar vol. 8, March 2022.
- **Accelerating Decision Tree Ensemble with Guided Branch Approximation (in Japanese).**
At ETNET 2022, March 2022.
- **Pre-conditioners applied to ROOT compression algorithms.**
At IRIS-HEP/GSoC Student Presentations, September 2020.

MEDIA

- **Inside the UW Allen School: Six 'grand challenges' shaping the future of computer science.**
At GeekWire, featuring VoxServe and ConsumerBench.

SERVICES

Conference Reviewing

- NeurIPS (2026)
- CAIS (2026)
- ICML (2026)
- MLSys (2026)
- ICLR (2026)

Journal Reviewing

- TMLR (2025 – Present)

Workshop Reviewing

- COLM 2026 Workshop on Efficient Reasoning
- CVPR 2026 Multi-Modal Reasoning for AI Agents Workshop
- ICLR 2026 MemAgents Workshop
- ICLR 2026 ES-Reasoning Workshop
- NeurIPS 2025 Workshop on Efficient Reasoning

Artifact Evaluation

- SOSP (2026)
- CAIS (2026)
- EuroSys (2026)
- MICRO (2025)

Departmental Service

- Allen School Pre-Application Mentorship Service (PAMS) Mentor 2025
- Allen School Ph.D. Admission Committee Member 2023 – 2024
- Allen School Ph.D. Admission Student Area Chair (Architecture) 2024

Miscellaneous

- The Japanese Graduate Student Association in the United States (JGSAU) Committee Member 2023 – 2026
- XPLANE SoP Writing Support Program Mentor 2023 – Present

SKILLS

- **Programming Languages:** C, C++, Python, CUDA, Triton, JavaScript, Verilog, SystemVerilog, Chisel, OCaml, Scala, R
- **ML Systems & Serving:** PyTorch, vLLM, SGLang, TensorRT-LLM, FlashInfer, Hugging Face, TVM, FAISS
- **Hardware & Simulation:** Vivado, Vitis, Pynq, ChampSim, gem5

Last update: July 11, 2026