

Taeho Kim

netstech.org/taehokim/ • ystart2020@gmail.com • +1 720-438-1117

RESEARCH INTERESTS

Deep Learning, ML Systems, Mobile Computing

SKILLS

- GPU kernel development: CUDA, Triton, kernel profiling and optimization
- Distributed training: DP/FSDP, TP, PP, EP, NCCL
- Experienced in fine-tuning pre-trained LLM
- Experienced in DL model compression and compilation
- Proficient in C, C++, Python, PyTorch

WORK EXPERIENCE

Amazon Web Services, Inc., Santa Clara, CA, USA

Sep 2024 – Present

Applied Scientist II

- **LLM Inference Infrastructure and GPU Kernel Optimization:**
 - Designed and implemented fused CUDA kernels for LLM inference optimization, including multi-LoRA adapter application with MoE expert routing, directly targeting production-scale inference serving; contribution merged into open-source FlashInfer [link]
 - Profiled kernel-level execution bottlenecks and restructured inter-kernel communication to consolidate multiple operations into unified monokernel execution paths, reducing inference-time overhead
- **Analytical Performance Modeling for Large-Scale Training and Inference:**
 - Built an end-to-end analytical performance modeling tool predicting training throughput, memory usage, and optimal parallelization strategy (DP/FSDP, TP, PP, and EP) across model scales from 7B to 10T+ parameters, enabling faster, data-driven infrastructure decisions
 - Used torch.export to automatically parse arbitrary PyTorch graphs, modeling gradient synchronization, activation recomputation, and communication/compute overlap for dense transformers and MoE architectures at scale
- **Fault-Tolerant, Checkpointless Distributed Training at Scale:**
 - Contributed to fault-tolerant, checkpointless training on AWS SageMaker HyperPod (deployed across up to 512 nodes), focusing on communication optimization and failure detection in dynamic recovery paths to enable seamless recovery without full-job restarts or manual intervention
 - Optimized NCCL communication latency in recovery paths, reducing failure-recovery time; presented results at AWS re:Invent 2025
- **Distributed Training Infrastructure - Performance Investigation:**
 - Evaluated NCCL communicator caching as a mechanism to accelerate training-cluster restarts; implemented caching and quantified recovery overhead
 - Identified that dynamic configuration changes limited cache reuse and yielded negligible net gain; documented findings and led team trade-off review, informing the decision to deprecate the approach and prioritize higher-impact optimizations

Amazon Web Services, Inc., Santa Clara, CA, USA

May 2023 – Aug 2023

Applied Scientist Intern

- Distributed training and inference of large-scale deep learning models

Amazon Web Services, Inc., Santa Clara, CA, USA

May 2022 – Aug 2022

Applied Scientist Intern

- Machine learning compiler tuning optimization

RESEARCH EXPERIENCE

Internet Systems Laboratory, CU Boulder, CO, USA

Jan 2020 – Aug 2024

Research Assistant (Advisor: Professor Sangtae Ha)

- **Compiler-Informed Model Pruning for Edge Deployment:**
 - Developed CPrune, a compiler-guided pruning framework that uses target-hardware execution profiling to guide structured pruning decisions, improving DNN inference efficiency on constrained devices (ECCV 2022)
- **Resource-Aware Modeling for Model Fine-Tuning:**

- Developed LLMem, a predictive model estimating GPU memory requirements for fine-tuning pre-trained LLMs under various configurations, enabling upfront hardware-resource planning (IJCAI 2024, top 2.27% of submissions)

Mobile and Networked Systems Laboratory, CU Boulder, CO, USA

Jul 2018 – Dec 2019

Research Assistant (Advisor: Professor Tam N. Vu)

▪ **Wearable Sensor Fusion Systems:**

- Contributed to wearable sensing systems (behind-the-ear microsleep detection, postoperative pain quantification) requiring real-time fusion of multi-modal physiological sensor streams (ACM MobiSys)

Wireless Data Communications Laboratory, Korea University, Seoul, Korea

Mar 2016 – Jun 2018

Research Intern (Advisor: Professor Hyogon Kim)

EDUCATION

University of Colorado Boulder, Boulder, CO, USA

Aug 2018 – Aug 2024

- Doctor of Philosophy Computer Science

University of Colorado Boulder, Boulder, CO, USA

Aug 2018 – May 2022

- Master of Science in Computer Science

Korea University, Seoul, Korea

Mar 2011 – Feb 2018

- Bachelor of Science in Computer Science and Engineering
- Bachelor of Engineering in Software Technology and Enterprise Program

PUBLICATIONS

CONFERENCES

- [1] Xiteng Yao, [Taeho Kim](#), Hengzhi Pei, Xinle Liu, Kyle Ulrich, Leonard Lausen, Ashish Khetan, Xiang Song, George Karypis and Martin Herbordt, “KernelSight-LM: A Kernel-Level LLM Inference Simulator,” in Arxiv, Jul 2026 [pdf]
- [2] [Taeho Kim](#), Yanming Wang, Vatshank Chaturvedi, Lokesh Gupta, Seyeon Kim, Yongin Kwon and Sangtae Ha, “LLMem: Estimating GPU Memory Usage for Fine-Tuning Pre-Trained LLMs,” in *The 33rd International Joint Conference on Artificial Intelligence*, **Long talk (top 2.27% of submissions)**, Aug 2024. [pdf]
- [3] [Taeho Kim](#), Yongin Kwon, Jemin Lee, Taeho Kim and Sangtae Ha, “CPrune: Compiler-Informed Model Pruning for Efficient Target-Aware DNN Execution,” in *The 17th European Conference on Computer Vision*, Tel Aviv, Israel, Oct 2022. [pdf]
- [4] Nhat Pham, Tuan Dinh, Zohreh Raghebi, [Taeho Kim](#), Nam Bui, Phuc Nguyen, Hoang Truong, Farnoush Banaei-Kashani, Ann Halbower, Thang N. Dinh and Tam Vu, “WAKE: A Behind-the-ear Wearable System for Microsleep Detection,” in *The 18th ACM International Conference on Mobile Systems, Applications, and Services*, Toronto, Canada, Jun 2020. [pdf]
- [5] Hoang Truong, Nam Bui, Zohreh Raghebi, Marta Ceko, Nhat Pham, Phuc Nguyen, Anh Nguyen, [Taeho Kim](#), Katrina Siegfried, Evan Stene, Taylor Tvrdy, Logan Weinman, Thomas Payne, Devin Burke, Thang Dinh, Sidney D’Mello, Farnoush Banaei-Kashani, Tor Wager, Pavel Goldstein and Tam Vu, “Painometry: Wearable and Objective Quantification System for Acute Postoperative Pain,” in *The 18th ACM International Conference on Mobile Systems, Applications, and Services*, Toronto, Canada, Jun 2020. [pdf]
- [6] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, [Taeho Kim](#), Nicholas Farrow, Anh Nguyen, Jianliang Xiao, Robin Deterding, Thang Dinh and Tam Vu, “eBP: A Wearable System For Frequent and Comfortable Blood Pressure Monitoring From User’s Ear,” in *The 25th Annual International Conference on Mobile Computing and Networking*, Los Cabos, Mexico, Oct 2019. [pdf]
- [7] P. Nguyen, [T. Kim](#), J. Miao, D. Hesselius, E. Kenneally, D. Massey, E. Frew, R. Han, and T. Vu, “Towards RF-based Localization of a Drone and Its Controller,” in *The 5th ACM Workshop on Micro Aerial Vehicle Networks, Systems, and Applications*, Seoul, Korea, Jun 2019. [pdf]
- [8] N. Pham, [T. Kim](#), F. M. Thayer, A. Nguyen, and T. Vu, “Demo: Earable - An Ear-Worn Biosignal Sensing Platform for Cognitive State Monitoring and Human-Computer Interaction,” in *The 17th ACM International Conference on Mobile Systems, Applications, and Services (Demo)*, Seoul, Korea, Jun 2019. [pdf] [demo]

- [9] T. Kim, W. Han, H. Kim, and Y. Park, “Vulnerable Road User Protection through Intuitive Visual Cue on Smartphones,” in *Proceedings of the 2nd ACM International Workshop on Connected and Automated Vehicle Mobility, A Workshop of MobiCom 2017*, Snowbird, Utah, USA, Oct 2017. [pdf] [slides] [demo1] [demo2]
- [10] T. Kim and H. Kim, “Vehicle-to-Vehicle Message Content Plausibility Check through Low-Power Beaconing,” in *Proceedings of IEEE 86th Vehicular Technology Conference: VTC2017-Fall*, Toronto, Canada, Sep 2017. [pdf] [slides] [video]
- [11] T. Kim, W. Han, and Y. Park, “Visual Cue-Based VRU Protection on Smartphones,” in *Proceedings of the 15th ACM International Conference on Mobile Systems, Applications, and Services (poster)*, Niagara Falls, NY, USA, Jun 2017. [pdf] [poster]

JOURNALS

- [1] JooHyoung Cha, Taeho Kim, Jemin Lee, Sangtae Ha and Yongin Kwon, “Target-Aware Neural Network Execution via Compiler-Guided Pruning,” in *IEEE Transactions on Mobile Computing*, Nov 2025 [pdf]
- [2] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, Robin Deterding, Thang Dinh and Tam Vu, “eBP: an ear-worn device for frequent and comfortable blood pressure monitoring,” in *Communications of the ACM*, Jul 2021. [pdf]
- [3] Nhat Pham, Tuan Dinh, Taeho Kim, Zohreh Raghebi, Nam Bui, Hoang Truong, Tuan Nguyen, Farnoush Banaei-Kashani, Ann Halbower, Thang N. Dinh, Vp Nguyen and Tam Vu, “Detection of Microsleep Events with a Behind-the-ear Wearable System,” in *IEEE Transactions on Mobile Computing*, Jun 2021. [pdf]
- [4] Taeho Kim, Phuc Nguyen, Nhat Pham, Nam Bui, Hoang Truong, Sangtae Ha and Tam Vu, “Epileptic Seizure Detection and Experimental Treatment: A Review,” in *Frontiers in Neurology - Epilepsy*, Jul 2020. [pdf]
- [5] Hyogon Kim and Taeho Kim, “Vehicle-to-vehicle (V2V) Message Content Plausibility Check for Platoons through Low-Power Beaconing,” in *Sensors 2019, 19(24)*, Dec 2019. [pdf]

PATENTS

- “System and method for detecting fake information about vehicle location” [KR][US][DE][CN]
- “Method and apparatus for verifying vehicle in inter-vehicular communication environment” [KR][US][CN][DE]

AWARDS & SCHOLARSHIPS

- National Science and Engineering Scholarship, Korea Student Aid Foundation 2015 – 2016
 - To attain a semester GPA of at least 3.50/4.50. Full-tuition scholarship (Junior and Senior)
- Software Technology and Enterprise Program Scholarship, Korea University 2015 – 2016
 - Additional scholarship to support Software projects and Start-up related education (Junior and Senior)
- Academic scholarships, Korea University
 - Best Honors Scholarships (a second-semester sophomore) Aug 2014
 - Honors Scholarships (a second-semester freshman and a first-semester sophomore) Aug 2011, Feb 2014

[CV updated on 2026-08-21]