Geon-Woo Kim

gwkim@utexas.edu LinkedIn Website

Research Interests

Systems for Machine Learning, LLM Training and Inference at Scale

Education

The University of Texas at Austin — Austin, TX, USA
PH.D. Student in Computer Science
Advisor: *Aditya Akella*
Sep 2022 - Present

Seoul National University — Seoul, South Korea
B.S. in Computer Science & Engineering
B.S. in Mathematical Sciences (Double Major)
Mar 2013 - Aug 2022

- Summa Cum Laude. This period overlapped with five years of industrial work and alternative military service.

Seoul Science High School — Seoul, South Korea
Mar 2010 - Feb 2013

Publications

HALoS: Hierarchical Asynchronous Local SGD over Slow Networks for Geo-Distributed Large Language Model Training, *Geon-Woo Kim*, Junbo Li, Shashidhar Gandham, Omar Baldonado, Adithya Gangidi, Pavan Balaji, Zhangyang Wang, Aditya Akella, Forty-second International Conference on Machine Learning (**ICML 25**)

OMEGA: A Low-Latency GNN Serving System for Large Graphs, *Geon-Woo Kim*, Donghyun Kim, Jeongyoon Moon, Henry Liu, Tarannum Khan, Anand Iyer, Daehyeok Kim, Aditya Akella, Under Submission, 2025

StitchLLM: Serving LLMs, One Block at a Time, Bodun Hu, Shuozhe Li, Saurabh Agarwal, Myungjin Lee, Akshay Jajoo, Jiamin Li, Le Xu, *Geon-Woo Kim*, Donghyun Kim, Hong Xu, Amy Zhang, Aditya Akella, The 63rd Annual Meeting of the Association for Computational Linguistics (**ACL 25**)

Read-ME: Refactorizing LLMs as Router-Decoupled Mixture of Experts with System Co-Design, Ruisi Cai, Yeonju Ro, *Geon-Woo Kim*, Peihao Wang, Babak Ehteshami Bejnordi, Aditya Akella, and Zhangyang Wang, Thirty-Eighth Annual Conference on Neural Information Processing Systems (**NeurIPS 24**)

Lovelock: Towards Smart NIC-hosted Clusters, Seo Jin Park, Ramesh Govindan, Kai Shen, David Culler, Fatma Özcan, *Geon-Woo Kim*, and Hank Levy, HotCarbon Workshop on Sustainable Computer Systems (**HotCarbon 24**)

Orca: A distributed serving system for Transformer-Based generative models, Gyeong-In Yu, Joo Seong Jeong, *Geon-Woo Kim*, Soojeong Kim, and Byung-Gon Chun, USENIX Symposium on Operating Systems Design and Implementation (**OSDI 22**)

Terra: Imperative-Symbolic Co-Execution of Imperative Deep Learning Programs, Taebum Kim, Eunji Jeong, *Geon-Woo Kim*, Yunmo Koo, Sehoon Kim, Gyeong-In Yu, and Byung-Gon Chun, Thirty-Fifth Annual Conference on Neural Information Processing Systems (**NeurIPS 21**)

Apache Nemo: A Framework for Building Distributed Dataflow Optimization Policies, Youngseok Yang, Jeongyoon Eo, *Geon-Woo Kim*, Joo Yeon Kim, Sanha Lee, Jangho Seo, Won Wook Song, and Byung-Gon Chun, USENIX Annual Technical Conference (**ATC 19**)

Pado: A data processing engine for harnessing transient resources in datacenters, Youngseok Yang, *Geon-Woo Kim*, Won Wook Song, Yunseong Lee, Andrew Chung, Zhengping Qian, Brian Cho, and Byung-Gon Chun, ACM European Conference on Computer Systems (**EuroSys 17**)

Research Experience

Graduate Research Assistant, UTNS — UT Austin
Advisor: *Aditya Akella* — *Sep 2022 - Present*

- *HALoS*: Addresses challenges of geo-distributed and heterogeneous resource problem in LLM training through hierarchical and asynchronous training framework.
- *OMEGA*: Achieved 10×-100× lower GNN serving latency with dependency-aware approximation and communication-efficient distributed execution.

Student Researcher, SystemsResearch@Google — Google
Mentor: *Seo Jin Park* — *May 2023 - Jul 2023*

- *Efficient LLM Serving*: Demonstrated, through simulation, the optimized throughput in MoE-based LLM serving using fine-grained expert disaggregation.

Undergraduate Research Intern, SPL — SNU
Advisor: *Byung-Gon Chun* — *Jan 2015 - Aug 2021*

- *Orca*: Proposed ***continuous batching***, a standard batching technique in LLM serving. Contributed to refactoring attention CUDA kernels to allow advanced KV cache management. Developed distributed deployment and fast C++-based web frontend.
- *Terra*: Enabled symbolic execution for imperative deep learning programs. Contributed to improving formal guarantees with JIT compilation and mathematical proof. Extensively profiled and optimized the system.
- *Nemo*: A data processing engine that provides a flexible IR to adapt to various physical environments. Contributed to identifying the extension of Pado with IR and designing the initial prototype.
- *Pado*: A data processing engine that utilizes transient (evictable) resources of datacenters. Contributed to developing a realistic eviction model by analyzing Google trace and deriving Pado's operator placement algorithm.

Honors & Awards

Aug 2022 **Overseas PhD Scholarship** — Korea Foundation for Advanced Studies

Mar 2013 **Presidential Science Scholarship** — Korean Government

Work Experience

Software Platform Engineer — Viva Republica (Toss)
South Korea's leading fintech startup valued at over $7 billion with over 20 million users — *Jun 2016 - Feb 2021*

- Initiated, designed, implemented, and maintained a comprehensive mobile marketing platform with multiple functionalities, including running A/B tests, sending large numbers of messages to target groups, and dynamically prompting dialogues based on real-time user logs.
- Built and maintained a large-scale data analysis platform utilizing various open-source projects, including Apache Spark, HDFS, Apache Impala, Apache Flink, and PyTorch.

Technical Skills

Languages: Python, C++, C, Kotlin, Java

Frameworks: PyTorch, Tensorflow, DeepSpeed, DeepSpeed-MII, vLLM, Deep Graph Library, Spark, Flink