Gangmuk Lim

website github linkedin gangmuk2@illinois.edu

Education

University of Illinois Urbana-Champaign 2022 - Present
Ph.D. in Computer Science
Advised by Brighten Godfrey

Ulsan National Institute of Science and Technology(UNIST) 2019 - 2021
M.S. in Computer Science Engineering and Finance
Advised by Myeongjae Jeon

Technikum Wien 2017 - 2017
Exchange student at Telecommunication department, Vienna, Austria

Ulsan National Institute of Science and Technology(UNIST) 2012 - 2019
B.S. in Computer Science Engineering & Finance (including 2 years of military services)

Research project

Service Layer Traffic Engineering (In submission) [EnvoyCon 2023] , [IstioCon 2023]

- A new network architecutre for servier layer traffic engineering. Automatically optimizing the flow of requests in microservice-based clusters, leveraging linear programming.

Kivi: Verification for Cluster Management (USENIX ATC 24') [paper] [kivi-code] [k8s-exp-code] [KubeCon 2023]
Bingzhe Liu, **Gangmuk Lim**, Ryan Beckett, P. Brighten Godfrey

- The first system that verifies cluster management systems via model checking technique. With intelligent scaling algorithm, it can find violations ≈100× faster if any.

FSMHSA: Fused Sparse Multi-head Self Attention [paper]

- Explored how to implement optimised sparse fused-multi-head-self-attention CUDA kernels. Uncovered that naive optimizations like span specialisations cause undue performance penalties due to irregular thread-access patterns across thread-blocks.

Zico: Efficient GPU Memory Sharing for Concurrent DNN Training (USENIX ATC 21') [paper]
Gangmuk Lim, Jeongseob Ahn, Wencong Xiao, Youngjin Kwon, Myeongjae Jeon

- GPU sharing system that co-locates multiple DNN training jobs improving throuhgput and resource utilization. Outperformed the existing GPU sharing techniques, NVIDIA MPS with unified memory, by order of magnitude by scheduling DNN kernels in a memory efficient way reducing total memory consumption. Built on top of TensorFlow.

Approximate Quantiles for Data Center Telemetry Monitoring (IEEE ICDE 2020, short) [paper], [talk]
Gangmuk Lim, Myeongjae Jeon, Stavros Volos, Mohamed Hassan, Ze Jin

- Devised a resource-efficient and accurate approximation algorithm and data structure for quantile operation for data center telemetry monitoring.

A Failure Study On Multi-Controller Failures in Cluster Management System(Kubernetes) [paper], [code]

- Analyzed and reproduced 12 multi-controller failures in Kubernetes clusters.

DBKnitter: LLM Approach For Cross-Database Queries

- Built a unifying SQL interface for multiple data systems using LLM to translate it into a python program.

Experience

Rebellions Inc., Software engineer Aug 2021-Apr 2022

- Implemented a software stack connecting deep learning framework runtime stack and SoC emulator for DNN kernel testing. Analyzed and tested the performance of DNN kernels on the software simulator of Rebellions ASIC. Found ∼30 of bugs in dev pipeline (Target models: Quantized/full ResNet50, Quantized/full BERT, LSTM, MobileNet)

IDLab-MEDIA, Ghent University, Research intern Jul 2017-Aug 2017

- Create dataset of super-resolution panoramic images for light field experiment in virtual reality using Unreal Engine 4.