Gangmuk Lim

🏠 Website Github Linkedin ✉ gangmuk2@illinois.edu

Education

University of Illinois Urbana-Champaign 2022 - Expected 2027
Ph.D. in Computer Science Advised by Prof. Brighten Godfrey

University of Applied Sciences at Vienna
Exchange student at Telecommunication department, Vienna, Austria

Ulsan National Institute of Science and Technology(UNIST)
B.S., and M.S. in Computer Science Advised by Prof. Myeongjae Jeon

Research

SLATE: Service Layer Traffic Engineering (*NSDI'26, HotNets'24*) [Paper], [KubeCon '23], [IstioCon '23]
Gangmuk Lim, Aditya Prerepa*, P. Brighten Godfrey, Radhika Mittal*

A system that optimizes request routing in microservice applications spanning multiple geo-graphically distributed clusters. SLATE takes a unique hybrid approach combining global optimization and local exploration. It outperforms state-of-the-art global load balancing by up to 18.3× in average latency and reduces egress bandwidth cost by up to 2.64×.

AIBrix: Towards Scalable, Cost-Effective Large Language Model Inference Infrastructure [Paper],

Scalable cloud-native LLM inference infrastructure

Kivi: Verification for Cluster Management (*USENIX ATC'24*) [Paper], [Code], [K8S-reproduction], [KubeCon '23]
Bingzhe Liu, Gangmuk Lim, Ryan Beckett, P. Brighten Godfrey

The first verification system for cluster controllers, configurations and their interaction in cluster management systems (e.g., scheduler, deployment, autoscaler and more in K8S). Kivi models the controllers and events exhaustively checks via model checking.

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

Memory-aware GPU sharing framework. Multiple DNN training jobs share a GPU aware of tensor allocation pattern of each GPU. It enables more dense GPU sharing, leading higher GPU utilization. It outperforms existing GPU sharing solution (NVIDIA MPS), improving throughput by 10× with memory consumption aware DNN kernels scheduling.

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

Implemented a new CUDA kernel for sparse fused-multi-head-self-attention that optimizes latency, leveraging sparsity in attention (e.g., blocked, sliding window attentions).

A Failure Study On Multi-Controller Failures in Kubernetes [Paper], [code]

Conducted an in-depth analysis and reproduction of 12 multi-controller failures in Kubernetes clusters.

Approximate Quantiles for Data Center Telemetry Monitoring (*ICDE'20 (short)*) [Paper], [Talk]
Gangmuk Lim, Mohamed Hassan, Ze Jin, Stavros Volos, Myeongjae Jeon

A resource-efficient quantile approximation algorithm and data structure for telemetry monitoring in data centers.

Experience

Bytedance compute infrastructure team, Research Intern Jan 2025 - Jun 2025

(OSS contribution): Made open-source contribution on AIBrix OSS. The contributions:
– Benchmarked the different autoscalers and routing policies.
– Implement prefix and load aware routing (radix tree based prefix matching indexer and performance cost model which is aware of load and prefix cache hit ratio).

(Research): Learning based adaptive request routing system for LLM inference applications in heterogeneous GPU cluster which outperforms existing prefix aware routing policy by 1.3 - 1.8x for Time-to-first-token (TTFT) by finding the more optimal instance to route requests.

Rebellions AI, Software Engineer Jun 2021 - Apr 2022

– Implemented a software stack connecting deep learning runtime framework and SoC emulator for DNN kernel testing. Used to find 30 bugs for in-house kernels used for following models: (Quantized) BERT, (Quantized) ResNet50, LSTM, MobileNet.

– Analyzed and tested in-house DNN kernels on the Rebellions ASIC (Atom chip) software simulator.

IDLab-MEDIA, Ghent University, Research Intern Jun 2017 - Aug 2017

– Implemented high concurrency (multi-threaded) data collection pipeline for light field experiment using Unreal Engine.

Scholarship

– ASEM-DUO international scholarship
– (Merit-based) National Science & Technology Scholarship during undergrad

Skills

Programming Languages: C, C++, Python, Go
Frameworks & Tools: Kubernetes, Istio, Envoy, PyTorch, TensorFlow
Specializations: Microservices, distributed systems, cloud, request routing, infrastructure, optimizing LLM inference