Gangmuk Lim

⌂ Website   Github   in Linkedin   ✉ gangmuk2@illinois.edu

Education

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

Research

SLATE: Service Layer Traffic Engineering (*HotNets '24*) [Paper], [KubeCon '23], [IstioCon '23]
Gangmuk Lim*, Aditya Prerepa*, P. Brighten Godfrey, Radhika Mittal
A new network architecture for microservice-based applications spanning in multi-cluster/cloud environments. SLATE leverages global knowledge of cluster states and multi-hop application graphs to centrally control the flow of requests in order to optimize end-to-end application latency and cost.

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 system for verifying 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
GPU sharing framework for multiple co-located DNN training workloads. It outperforms existing GPU sharing solution (NVIDIA MPS), improving throuhgput by $10\times$ with memory consumption aware DNN kernels scheduling.

FSMHSA: Fused Sparse Multi-head Self Attention [Report]
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).

Approximate Quantiles for Data Center Telemetry Monitoring (*ICDE '20, short*) [Paper], [Talk]
Gangmuk Lim, Mohamed Hassan, Ze Jin, Stavros Volos, Myeongjae Jeon
Created a resource-efficient quantile approximation algorithm and data structure for telemetry monitoring in data centers.

Experience

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 in development pipeline (Target models: BERT, ResNet50, LSTM, MobileNet).
- Analyzed and tested in-house DNN kernels on the Rebellions ASIC software simulator.
- Closely worked with compiler team and hardware team.

IDLab-MEDIA, Ghent University, Research Intern   Jun 2017 - Aug 2017
Involved in denoising and reconstructing light field image research project. Implemented multi threads/machine concurrent data collection pipeline in VR world using Unreal Engine and get high resolution dataset.

Skills

- **Languages:** C, C++, Python, Go, CUDA
- **Framework:** Kubernetes, Istio, Envoy, PyTorch, TensorFlow, GPU profiling, Docker
- **Domains:** Cloud Native, ML System, GPU, Microservices, Service Mesh, Application-Level Networking, Proxy Server, Traffic Engineering, Distributed system