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

University of Applied Sciences at Vienna 2017 - 2017
Exchange student at Telecommunication department, Vienna, Austria

Ulsan National Institute of Science and Technology(UNIST) 2012 - 2018, 2018 - 2020
B.S. in Computer Science, M.S. in Computer Science
Advised by Myeongjae Jeon (2 years of leave for military services during B.S.)

Research Projects

SLATE: Service Layer Traffic Engineering (HotNets 25') [Paper], [KubeCon 23'], [IstioCon 23']
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 in K8S). Kivi models the controllers and events into processes and 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× 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 (IEEE ICDE 20', short) [Paper], [Talk]
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: (Quantized) ResNet50, (Quantized) BERT, 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 concurrenct data collection pipeline in VR world using Unreal Engine and get high resolution dataset.

Skills

- **Languages:** C++, Python, Go, CUDA
- **Framework:** Kubernetes, Istio, Envoy, PyTorch, TensorFlow, GPU profiling
- **Domains:** ML System, GPU system, Microservices, Service mesh, Proxy, Traffic Engineering, Service Layer Network