Daegun Yoon

High Performance Computing System Research Section	Tel: +82 42-860-5859
Future Computing Research Division	Phone: +82 10-9471-4249
Artificial Intelligence Computing Research Laboratory	Email: kljp@etri.re.kr
ETRI, Republic of Korea	Homepage: https://sites.google.com/view/kljp

RESEARCH INTERESTS

On-Device Inference: High-performance on-device AI model inference via model compression and performance optimization

Distributed Training: Scalable distributed machine learning via gradient sparsification

High-Performance Computing: Performance optimization for algorithms and systems via parallel and distributed computing

POSITIONS

Researcher in Electronics and Telecommunications Research Institute (**ETRI**), Republic of Korea Jan. 2024 - Present

- **PIM (ReRAM) based accelerator for high-performance on-device LLM inference**
  - Improving the computation accuracy of MAC (Multiply and accumulation) operation and efficiency of language model deployment on PIM-based accelerators
  - Efficient AI model inference using only analog computing devices (without AI cloud)
  - Revising the analog computing simulator for on-device LLM inference test (IBM Analog Hardware Acceleration Kit, DNN+NeuroSIM)
- **Model compression for high-performance on-device LLM inference**
  - Applying gradient sparsification to distributed training of large model to identify the significant parameters for inference performance (accuracy and speed) and make the standards for structured or unstructured model pruning
  - Applying constraints of analog on-device inference to low precision training methods (Quantization-aware training, 1-bit LLM) to optimize the performance of compressed models
- **Communication cost optimization for LLM multi-node multi-GPU distributed training**
  - LLM-specialized lossy communication method (gradient sparsification) for reducing the communication cost in large-scale distributed training of LLM
  - Layer-overlapped gradient sparsification to apply the lossy communication method to pipeline-parallel distributed training

EDUCATION

Ph.D. in Department of Artificial Intelligence, Ajou University, Republic of Korea Sep. 2018 - Feb. 2024
Advisor: Prof. Sangyoon Oh

- Performance evaluation on large-scale distributed training, and acceleration and optimization algorithms
- Software stack and system configuration for large-scale AI model training in supercomputing environment
- Parallel and distributed system for scalable large-scale AI model training

B.S. in Department of Software, Ajou University, Republic of Korea Mar. 2013 - Aug. 2018

- AUTOSAR programming for developing the ECU (TCS: traction control system) for autonomous driving
- Performance evaluation on virtual machine live migration in homogeneous operating system environment
- Development of an emotion-based analytics tool for characterizing online news, comments, and users

PROFESSIONAL SKILLS

[Machine Learning] PyTorch, DeepSpeed, FairScale, Multi-node multi-GPU distributed training, Sparse communication

[Parallel/Distributed/HPC Optimization] CUDA, GPGPU, MPI, Network programming, Multithread programming, Graph processing, Parallel/distributed computing

[Programming] Python, C/C++, Java

[Research] Capability for analyzing the state-of-the-art researches and figuring out the solutions to the problem

[English] Paper and technical report writing, presentation and Q&A

SELECTED PUBLICATIONS

C3. **Daegun Yoon**, Sangyoon Oh, “Preserving Near-Optimal Gradient Sparsification Cost for Scalable Distributed Deep Learning”, 24th IEEE/ACM International Symposium on Cluster, Cloud, and Internet Computing (CCGrid), May. 2024.

C2. **Daegun Yoon**, Sangyoon Oh, “MiCRO: Near-Zero Cost Gradient Sparsification for Scaling and Accelerating Distributed DNN Training”, 30th IEEE International Conference on High Performance Computing, Data, and Analytics (HiPC), Dec. 2023.

C1. **Daegun Yoon**, Sangyoon Oh, “DEFT: Exploiting Gradient Norm Difference between Model Layers for Scalable Gradient Sparsification”, 52nd International Conference on Parallel Processing (ICPP), Aug. 2023.

J5. **Daegun Yoon**, Minjoong Jeong, Sangyoon Oh, “SAGE: toward on-the-fly gradient compression ratio scaling”, The Journal of Supercomputing (SUPE), Feb. 2023.

J4. **Daegun Yoon**, Minjoong Jeong, Sangyoon Oh, “WAVE: designing a heuristics-based three-way breadth-first search on GPUs”, The Journal of Supercomputing (SUPE), Nov. 2022.

J3. **Daegun Yoon**, Sangyoon Oh, SURF: “Direction-Optimizing Breadth-First Search Using Workload State on GPUs”, Sensors, Jun. 2022.

J2. **Daegun Yoon**, Zhetao Li, Sangyoon Oh, “Balanced content space partitioning for pub/sub: a study on impact of varying partitioning granularity”, The Journal of Supercomputing (SUPE), Apr. 2021.

J1. **Daegun Yoon**, Gyudong Park, Sangyoon Oh, “Exploring a system architecture of content-based publish/subscribe system for efficient on-the-fly data dissemination”, Concurrency and Computation: Practice and Experience (CCPE), Nov. 2020.

PATENTS

P3. Sangyoon Oh, Byeong-hee Roh, **Daegun Yoon**, Cheol-woong Lee, Kyungwoo Kim, “METHOD OF IMPROVING PERFORMANCE OF SOFTWARE-DEFINED NETWORKING OF ELECTRONIC DEVICE”, Korea Patent, Feb. 2024.

P2. Sangyoon Oh, **Daegun Yoon**, “APPARATUS AND METHOD FOR ADAPTIVE GRAPH TRAVERSAL BASED ON WORKLOAD ANALYSIS”, Korea Patent, Jun. 2023.

P1. Minho Park, Sangyoon Oh, **Daegun Yoon**, Jaehyun Ham, “METHOD AND APPARATUS FOR PARTITIONING OF EVENT, COMPUTER-READABLE STORAGE MEDIUM AND COMPUTER PROGRAM”, Korea Patent, Jul. 2022.

SELECTED RESEARCH PROJECTS

R3. **Electronics and Telecommunications Research Institute**, “Analog AI Computing”. Jan. 2024 - Present

R2. **Samsung Display**, “Development of High Efficiency HPC Job Scheduling Algorithm”. Jan. 2023 - Dec. 2023

R1. **Korea Institute of Science and Technology Information**, “Research on Optimizing Memory Utilization and Communication Scheduling of Sharded Data Parallel for Accelerating Large-Scale Distributed Deep Learning”. Mar. 2022 - Oct. 2022

PROFESSIONAL SERVICES

Reviewer: The Journal of Supercomputing (2023, 2024)

Reviewer: Journal of Big Data (2024)

Reviewer: Cluster Computing (2024)

Reviewer: World Wide Web (2024)

Reviewer: Wireless Networks (2024)

Reviewer: Journal of Grid Computing (2024)

Reviewer: Computing (2024)

Reviewer: International Journal of Machine Learning and Cybernetics (2024)

Reviewer: Journal of Real-Time Image Processing (2024)

Reviewer: ACM Transactions on Multimedia Computing Communications and Applications (2023)

TEACHING EXPERIENCES

Teaching Assistant: “Software Engineering”, Department of Software, Ajou University Spring 2021

Teaching Assistant: “Digital Circuits”, Department of Software, Ajou University Fall 2022

AWARDS

A1. **Excellent Dissertation Award**: “Dynamic Gradient Sparsification Exploiting Aggregated Gradients for Scalable Distributed Deep Learning”, Department of Software, Ajou University, Feb. 2024.