Byungsoo Oh

obs0811@gmail.com | byungsoo-oh.github.io

RESEARCH INTERESTS

Systems for ML, Cloud Computing, Distributed Systems

EDUCATION

Cornell University Ithaca, NY, USA
Ph.D. in Computer Science Aug 2024 –

Korea Advanced Institute of Science and Technology (KAIST) Daejeon, South Korea
M.S. in Computer Science Mar 2018 – Feb 2020

Sogang University Seoul, South Korea
B.S. in Computer Science and Engineering Mar 2012 – Feb 2018
Graduated with honors, *Summa Cum Laude*

PROFESSIONAL EXPERIENCE

Samsung Research Seoul, South Korea
Research Engineer Feb 2020 – Jun 2024

PUBLICATIONS

- Taegeon Um*, **Byungsoo Oh***, Minyoung Kang*, Woo-Yeon Lee, Goeun Kim, Dongseob Kim, Youngtaek Kim, Mohd Muzzammil, Myeongjae Jeon, "Metis: Fast Automatic Distributed Training on Heterogeneous GPUs", USENIX Annual Technical Conference (**USENIX ATC**), Santa Clara, CA, USA, 2024 (* = co-first authors)

- Taegeon Um, **Byungsoo Oh**, Byeongchan Seo, Minhyeok Kweun, Goeun Kim, Woo-Yeon Lee, "FastFlow: Accelerating Deep Learning Model Training with Smart Offloading of Input Data Pipeline", International Conference on Very Large Data Bases (**VLDB**), Vancouver, Canada, 2023

- Minhyeok Kweun, Goeun Kim, **Byungsoo Oh**, Seongho Jung, Taegeon Um, Woo-Yeon Lee, "PokéMem: Taming Wild Memory Consumers in Apache Spark", IEEE International Parallel and Distributed Processing Symposium (**IPDPS**), Lyon, France, 2022

- Seungju Cho, Tae Joon Jun, **Byungsoo Oh**, Daeyoung Kim, "DAPAS: Denoising Autoencoder to Prevent Adversarial attack in Semantic Segmentation", International Joint Conference on Neural Networks (**IJCNN**), Glasgow, UK, 2020

- **Byungsoo Oh**, Daeyoung Kim, "Serverless-Enabled Permissioned Blockchain for Elastic Transaction Processing", ACM/IFIP International Middleware Conference (**Middleware**), *Poster Paper*, Davis, CA, USA, 2019

- **Byungsoo Oh**, Tae Joon Jun, Wondeuk Yoon, Yunho Lee, Sangtae Kim, and Daeyoung Kim, "Enhancing Trust of Supply Chain Using Blockchain Platform with Robust Data Model and Verification Mechanisms", IEEE International Conference on Systems, Man, and Cybernetics (**SMC**), Bari, Italy, 2019

PATENTS

- Taegeon Um, Minhyeok Kweun, **Byungsoo Oh**, "Smart Offloading for AI Input Data Pipeline Acceleration", US Patent, US20240135189A1, Published: Apr 25, 2024

- Minyoung Kang, **Byungsoo Oh**, Taegeon Um, "Device Placement Strategies for Optimizing 3D Parallelism in Non-Uniform Topology Environments", US Patent, Pending, 2023

- Minyoung Kang, **Byungsoo Oh**, Taegeon Um, "Method and System for Elastic Knowledge Distillation with Adaptive Coordination", US Patent, Pending, 2023

- Daeyoung Kim, **Byungsoo Oh**, "Method and System for Enhancing Trust of Supply Chain Using Blockchain Platform with Robust Data Model and Verification Mechanisms", Korean Patent, No. 10-2620822-0000, Issued: Dec 2023

HONORS AND AWARDS

- **USENIX ATC 2024 Student Grant** 2024
  Travel grant awarded to attend USENIX ATC 2024 (co-located with OSDI 2024) in Santa Clara
- **National Full Scholarship**, Korea Ministry of Science and ICT 2018–2020
- **Award for Top 1% Students in the College of Engineering (Dean's List)**, Sogang University 2017
  2 semesters (Spring 2017, Fall 2017)
- **Academic Excellence Scholarship**, Sogang University 2013–2017
  6 semesters (Spring 2013, Fall 2015, Spring 2016, Fall 2016, Spring 2017, Fall 2017)

RESEARCH EXPERIENCE

Distributed DNN Training on Heterogeneous GPUs (USENIX ATC'24) Jan 2023 – Jan 2024
Data Research Team, Samsung Research Seoul, South Korea

- Automatically finding optimal parallelism strategies for large DNN models on heterogeneous GPU clusters.

Smart Offloading of DNN Input Data Pipeline (VLDB'23) Jan 2022 – Dec 2022
Data Research Team, Samsung Research Seoul, South Korea

- Resolved training performance degradation due to preprocessing bottleneck by automatically harnessing disaggregated CPU resources.
- **Roles.** (1) Implemented policy and mechanism for automatic offloading of input data pipelines with lightweight metric profiling; (2) Implemented state-of-the-art baseline using NVIDIA DALI for evaluation; and (3) Designed evaluation, performed experiments, analyzed results, and wrote evaluation section.

Robust Memory Management for Apache Spark (IPDPS'22) Mar 2021 – Feb 2022
Data Analytics Lab, Samsung Research Seoul, South Korea

- Investigated unstable memory issues for Apache Spark due to inattentive memory management. Empowered Spark to effectively manage *wild memory consumers* by redesigning the memory manager.
- **Roles.** Assisted to reshape research direction with more focused problem definition, identified limits of existing methods, and built execution pipeline for performance benchmark.

Improving Performance and Robustness of Permissioned Blockchains Mar 2018 – Dec 2019
Data Engineering and Analytics Lab, KAIST Daejeon, South Korea

- **Serverless-Enabled Transaction Processing (Middleware'19 Poster).** Mitigated scalability issue for decentralized execution of smart contracts by leveraging serverless computing (first-author publication).
- **Anomaly Detection in Transactions (SMC'19).** Resolved correctness issue for permissioned blockchains by semantically validating transactions before block confirmation to prevent anomalous actions from tampering with the state of permissioned blockchains (first-author publication).

ENGINEERING EXPERIENCE

Building Machine Learning Platform on Large GPU Cluster Jan 2020 – Feb 2021
Data Cloud Lab, Samsung Research Seoul, South Korea

- Managed Kubernetes-based ML-as-a-Service cloud that simplifies building and sustaining ML models.
- Developed and maintained core microservice that configures, deploys, and manages ML jobs.
- Monitored resource usage and tuned provisioning in large-scale multi-tenant cluster.

TEACHING EXPERIENCE

- TA, Introduction to System Programming (CS230), KAIST Spring 2019, Spring 2018
- TA, Embedded Operating Systems (CS632), KAIST Fall 2018

TECHNICAL SKILLS

- **Programming Languages.** C, C++, Python, JavaScript, Go, Java, Scala, Markdown, LaTeX
- **Frameworks.** TensorFlow, PyTorch, DeepSpeed, NCCL, NVIDIA DALI, Docker, Kubernetes, gRPC, Apache Spark, Apache Druid, Apache Hive, Hadoop, Apache Airflow, Node.js, React

OPEN SOURCE CONTRIBUTIONS

- **DeepSpeed.** Bug Fix [issue] [code]
- **TensorFlow.** Documentation improvement for `tf.data service` [code]
- **Apache Spark.** Bug Fix [code], Benchmark [code]

LANGUAGES

Korean (*native*), English (*fluent*)