

HOJAE SON

Amherst, MA · hojaeson@umass.edu · goodluck-hojae.github.io

RESEARCH INTERESTS

LLM Systems, ML Systems, Data Systems, GPU Inference and Scheduling

EDUCATION

University of Massachusetts Amherst

Ph.D., Computer Science *GPA: 3.68/4.0*

Amherst, MA

Sep 2023 – Present

Kyungpook National University

M.S., Computer Science *GPA: 4.18/4.50*

South Korea

Mar 2017 – Feb 2019

B.S., Computer Science *GPA: 4.11/4.50*

Mar 2010 – Feb 2017

RESEARCH PROJECTS

Kalpso: Relational LLM Serving

- A query-aware LLM serving abstraction that exposes semantic query plans to the serving layer and enables cross-operator KV-cache reuse.
- Designed an adaptive, memory-aware scheduler for pipelined semantic operators under bounded GPU KV-cache capacity.

Graph and Temporal Learning Systems

- Temporal graph learning and GNN training performance.
- Studied system-level tradeoffs in graph-learning systems and contributed to performance analysis and experimental evaluation.

PUBLICATIONS

- [1] **Hojae Son**, Md Ashraful Islam, Huy Gia Cao, Hui Guan, and Marco Serafini. “Kalypso: Relational LLM Serving.” *Under review*.
- [2] Md Ashraful Islam, **Hojae Son**, Suhaas Kiran, and Marco Serafini. “PRISM: A Training System to Unlock the Potential of Temporal Graph Learning Through Staleness Avoidance.” *PVLDB 2026*.
- [3] Saurabh Bajaj, **Hojae Son**, Juelin Liu, Hui Guan, and Marco Serafini. “Graph Neural Network Training Systems: A Performance Comparison of Full-Graph and Mini-Batch.” *PVLDB 2025*.
- [4] **Hojae Son**, Anand Paul, and Gwanggil Jeon. “Country Interest Analysis Based on Long-Term Short-Term Memory.” *ICDM Workshops 2018*.
- [5] M. Mazhar Rathore, **Hojae Son**, Awais Ahmad, and Anand Paul. “Real-time video processing for traffic control in smart city using Hadoop ecosystem with GPUs.” *Soft Computing 2017*.
- [6] M. Mazhar Rathore, **Hojae Son**, Awais Ahmad, Anand Paul, and Gwanggil Jeon. “Real-Time Big Data Stream Processing Using GPU with Spark over Hadoop Ecosystem.” *Parallel Programming 2017*.

EXPERIENCE

Unity Cluster

Research Assistant

Amherst, MA

Jan 2024 – Present

- Support researchers using a Slurm-based HPC platform for distributed ML training, inference workflows, and AI dataset management.

LuluLab, Inc. (Samsung C-Lab)

Research Engineer

South Korea

Oct 2021 – Dec 2022

- Integrated seven deep-learning models into an AI skin-analysis SDK and connected model inference with application-facing interfaces.
- Built and deployed REST services using AWS ECS for production AI workflows.

Hyundai Mobis, Inc.

Research Engineer

South Korea

Jan 2019 – Feb 2021

- Developed semantic-segmentation models for an ADAS vehicle project and worked on embedded inference deployment.
- Collaborated with Texas Instruments on deploying deep-learning models using TIDL and ONNX for an automotive platform.

TECHNICAL SKILLS

Languages: C/C++, Python, Java

Systems/ML: vLLM, PyTorch, CUDA, Hugging Face Transformers, Slurm, Linux

Infrastructure: AWS ECS, Git, DVC, REST APIs, ONNX

PATENT

Hojae Son and Jaeyoung Lee. “System for estimating shape of road based on deep learning network and method thereof.” Korean Patent No. 10-2020-0059386, Hyundai Mobis, 2020.