Peyman Gholami

https://peyman-gholami.github.io | pghola2@uic.edu | (603) 286 0118 |

Highly motivated Computer Engineering Ph.D. Candidate with a strong foundation in **machine learning, efficient inference, optimization, distributed systems, and networking**. Possesses hands-on experience in designing, building, and deploying **end-to-end machine learning frameworks** using **Python, PyTorch, Kubernetes, and MPI**. Proven expertise in developing communication and computationally efficient decentralized and federated learning algorithms for **streamable training of LLMs across datacenters**, with a track record of peer-reviewed publications. Seeking to leverage a deep understanding of machine learning to **solve complex, real-world AI challenges**.

SKILLS

Machine learning:
Deep Learning, Neural Networks, Distributed Machine Learning, Model Pruning, Reinforcement Learning, Large Language Models, Optimized LLM Inference, Distributed Inference

Languages:
Python (100K+ LOC), C++, Matlab

Frameworks:
Kubernetes, MPI, Djnago

Cloud:
Docker, k8s, AWS

EDUCATION

University of Illinois Chicago
PhD, Electrical and Computer Engineering
12/2026 (anticipated)

Sharif University
MSc, Electrical Engineering,
2021

Iran University of Science and Technology
BSc, Electrical Engineering,
2018

AWARDS & ACHIEVEMENTS

NSF full funding for ICNP attendance
2022

Ranked 51 in national university entrance exam
2018

Ranked 1st among students of Electrical Engineering of BSc
Fall 2016 – Spring 2017

EXPERIENCE

Graduate Research Assistant Networking Research Lab
Jan 2022 - now **Chicago, IL**

- **Improving compute and communication** cost in distributed machine learning.
- Eefficent speculative decoding in **LLM distributed inference**.
- Introducing a **new decentralized learning algorithm** (MultiWalk) and proving theoretically that it outperforms STOA in sparse settings.

Graduate Research Assistant EE department at Sharif
Sep 2018 – Spe 2021 **Tehran, Iran**

- Developed an queuing network model for **computation offloading in mobile edge and cloud servers**.
- Developed MDP and **deep reinforcement learning**-based delay-efficient task allocation.

PUBLICATIONS

- Differentiated Aggregation to Improve Generalization in Federated Learning. Gholami, Peyman, and Hulya Seferoglu. In *Transactions on Machine Learning Research*, 2025.
- Digest: Fast and communication efficient decentralized learning with local updates. Gholami, Peyman, and Hulya Seferoglu. In *IEEE Transactions on Machine Learning in Communications and Networking*, 2024.
- Communication efficient federated learning with differentiated aggregation. Gholami, Peyman, and Hulya Seferoglu. *Scalable and Efficient Artificial Intelligence Systems at AAAI*, Philadelphia, 2025.

SELECTED PROJECTS

Distributed ML
Developed an **end-to-end distributed machine learning framework** based on Kubernetes and MPI.
Implemented a wide range of **learning algorithms** on it.

RL-based Computation Offloading
A **deep reinforcement learning**-based delay-efficient computation offloading policy in **mobile edge and cloud** servers.