Jiahong **Ning**

(+86) 13768410701 | jiahong.ning@mnsu.edu | https://github.com/JasonNing96

"Imagination is more important than knowledge, knowledge is limited, imagination encircles the world."

Research Interest

Edge distributed computing, Internet of Things, Generative AI, Speculative decoding I focus on accelerating large-model inference in edge and IoT environments using speculative sampling, wireless task offloading, and distributed learning. My work aims to efficiently deploy pre-trained models for AGI applications under resource constraints.

Research Experience

A*STAR, Institute for Infocomm Research (I2R)

Singapore

Joint-Ph.D program Supervisor: Sun Shumei *May. 2024 – July. 2025*

- **Hybrid Wireless Computing Task Offloading for Marine Spatial Applications**: Designed multi-agent strategies and reinforcement learning models; paper completed and submitted to *IEEE Transactions on Network Science and Engineering (TNSE)*.
- **Edge Wireless Computing Caching Algorithms for Large-Scale Models**: Developed edge–cloud collaborative caching strategies to enhance cache utilization and model inference throughput; results published in *IEEE INFOCOM*.
- **Distributed Speculative Sampling Method (DSSD) for Wireless Networks**: Leveraged token-exchange characteristics of speculative sampling to boost throughput efficiency of LLMs on the wireless edge and overall edge device performance; paper accepted at *ICML 2025*.
- **UAV-Assisted Communication Strategies Using Distributed Speculative Sampling**: Employed reinforcement learning together with the DSSD algorithm to improve LLM service quality for edge users in emergency scenarios; related research submitted to *IEEE Transactions on Cognitive Communications and Networking (TCCN)*.

6G AI General large model intelligence

ShenZhen, China

Joint-Ph.D in PengCheng Labtory Supervisor: Tingting Yang *March. 2023 – Now*

- **Project Involvement in "6G General AI for Inclusive Intelligence":** Actively participated in the proposal for this project, focusing on the development of a "Network Big Model" based on large-scale models and wireless datasets. This work involved research on operational and network-native interpretability.
- **IMTM2030 Report and Project Research**, Contributed to the writing and research of the IMTM2030 report, providing insights into future technological trends and innovations.
- **Proposal Writing for "6G Big Model Network Architecture and Efficient Edge Deployment"**: Played a key role in drafting the proposal for this significant project, which aims to explore innovative architectures for 6G networks and efficient deployment strategies at the edge.
- **Research on Hierarchical Federated Architecture Based on Large Pre-trained Models**: Engaged in the study of hierarchical federated architectures, focusing on the efficient deployment of large-scale pre-trained models at the edge.
- **Development of an End-Edge-Cloud Distributed Collaborative Hardware Architecture**: Designed and implemented a distributed collaborative hardware architecture spanning end devices, edge nodes, and cloud systems. This architecture supports the lab's algorithm validation processes, especially for large-scale pre-trained models.

The Networking for Artificial Intelligence for 6th Communication-Networking

ShenZhen, China

Research assistant in Huawei 2012, Supervisor: Lu JianMing *Aug. 2021 – Now*

- Investigating the latest developments in communication theory and 3GPP standards, reading literature and conducting field research, and exploring the use of machine learning in wireless communication.
- Participating in the research of **Network4AI**, optimizing resource allocation in communication through AI learning.
- Participating in the **Huawei Kubeedge open-source project and as the initiator of the Kube-Wireless working group**.
- Studying and researching container edge networking and management tools, and exploring the integration of container management tools with 5G networks.
- Designing **Slam planning algorithms** using Python language and PyTorch framework.
- Communicating with project parties and instructors and collaborating with Huawei Cloud team.

Academic Achievement

[J] "DeAOff: Dependence-Aware Offloading of Decoder-Based Generative Models for Edge Computing",
IEEE China Communication (Chincom), 2025,

[C] "DSSD: Efficient Edge-Device Deployment and Collaborative Inference via Distributed Split Speculative Decoding",
The 42nd International Conference on Machine Learning (ICML), 2025,Accept

[C] "EdgePrompt: A Distributed Key-Value Inference Framework for LLMs in 6G Networks",
International Conference on Computer Communications (INFOCOM), 2025,Accept

[J] "Two-Stage Coded Distributed Edge Learning: A Dynamic Partial Gradient Coding Perspective",
IEEE Transactions on Mobile Computing (TMC), 2024, Accept

[C] Xinghan Wang, Cheng Huang, **Jiahong Ning**, Tingting Yang, and Xuemin Shen"Adaptive Distributed Learning with Byzantine Robustness: A Gradient-Projection-Based Method",
IEEE global Communications conference, 2023, Accept

[C] Xinghan Wang, Xiaoxiong, Zhong, **Jiahong ning**, Tingting Yang, Fangming Liu "Two-Stage Coded Distributed Learning: A Dynamic Partial Gradient Coding Perspective",
International Conference on Distributed Computing Systems, 2023

[J] Tingting Yang, **Jiahong ning**, Dapeng Lan, Jiawei Zhang, YangYang, and Xudong Wang, "KubeEdge Wireless for Integrated Communication and Computing Services Everywhere",
IEEE Communications Society, 2022

[P] Tingting Yang, **Jiahong ning**, Zechen He, Jiale Wang "**Top 10 Technological Advances in the Field of Communication in 2023 Nomination**:"Collaborative optimization of network architecture, protocols and experimental platform through communication and computing""
Pengcheng Lab, 2024

Education

Institute for Infocomm Research Institute, A*STAR
Singapore

Joint-Ph.D in Computer Science and Communication Technology
May. 2024 - Present

- Research on large pre-trained models and cache acceleration inference mechanisms
- Study distributed collaborative mechanisms
- Collaborate deeply with Professor Ducy from NTU (Nanyang Technological University)

Dalian Maritime University
Dalian, China

Ph.D in Traffic Control and Transport Engineering
Sep. 2021 - June. 2025

- Deep learning
- Optimization mathematics

Working Experience

Huawei cloud computing technology Co Ltd
ShenZhen, China

CloudBu Innovation Lab
June.2021 - Sep.2021

- Responsible for **kubernetes** operation and maintenance, including lab's log viewing and maintenance.
- Responsible of Edge-Mesh plugin development about Edge crossing tool for K8S by Golang.
- Manage the kubeedge community, collect issues from users and submit new feature requests.
- Lead the establishment of the wireless working group for novel scenarios about wireless.
- Research on **dynamic topology of wireless network**.
- Coding the edge side collaboration demo, participate in the development and presentation of 3GPP standards on **edge side collaboration**.

Huawei Technologies Co Ltd
ShenZhen, China

2012 Wireless tech lab
Jul. 2020 - may.2021

- Invested the latest development of communication theory and 3GPP standard, and explore the combination point of communication and Machine learning.
- Participated in the research of Network4AI, and optimized resource scheduling by **deep learning** in communication.
- Participated in KubeEdge open source project of Huawei, and the initiator of Kubeedge-Wireless, a special research group.
- Learning and research containerized networks and management tools, so as to do research about the junctions of container and 5G networks.
- Proposed a **distribution federated learning** solution for differential privacy protection problems in distributed wireless networks.
- Developed communication-related algorithms and verification based on Linux system, and contributed to Kubeedge community.
- Intelligence road planning algorithm was designed based on **PyTorch framework**.
- Responsible for communicating with professors and communicating with the Huawei cloud team.

Minnesota State University
Mankato, USA

Professor Assistant
Jan. 2019 – may. 2019

- Responsible for purchasing and managing equipment of laboratory
- Served as teaching assistance of Smart Grid (EE583) Course

Presentation

The 44th IEEE International Conference on Computer Communications(INFOCOM 2025)
London

PRESENTER FOR<EdgePrompt: A Distributed Key-Value Inference Framework for LLMs in 6G Networks>
May. 2025

- the latest research on large model key-value cache in edge computing.

International Conference on Distributed Computing Systems, 2023
HK

PRESENTER FOR<Two-Stage Coded Distributed Learning: A Dynamic Partial Gradient Coding Perspective">
July. 2023

- Design a novel algorithm of straggler problem in federated.
- Publicized the team's research results in distributed computing.