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 have conducted in-depth research on dynamic and vibrant fields such as edge distributed computing and the Internet of Things, and I am highly interested in the innovative applications of generative artificial intelligence. My enthusiasm also extends to edge computing, wireless resource task offloading, and distributed learning. In these areas, I explore how these methods can be efficiently deployed and utilized for pre-trained models, especially in the research of artificial general intelligence (AGI). I focus on the research of large model speculative sampling inference acceleration technologies in edge scenarios, aiming to achieve a better large model inference experience in edge-constrained scenarios. My goal is to contribute to this rapidly evolving field and devote myself to developing practical solutions to push the limits of edge generative artificial intelligence.

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.

Cloud native paradigm based Communication and Computing Integrated Intelligent Orchestration Platform-NAMO

ShenZhen, China

Inter Engineer

March. 2022 – Feb. 2023

- Participated in contributing "6G Network Operation and Maintenance White Paper", mainly participated in contributing the relationship between the underlying task flow of the large model and the analysis and algorithm design of operation and maintenance scenarios
- Design Namo platform test cases, responsible for quantifying the compression part
- Research distributed large model data storage architecture, select Namo platform data storage method
- Design Namo platform test cases, responsible for quantifying the compression part

Academic Achievement

[J1] "DeAOff: Dependence-Aware Offloading of Decoder-Based Generative Models for Edge Computing",
IEEE China Communication (Chincom), 2025, Accept

[C0] "DSSD: Efficient Edge-Device Deployment and Collaborative Inference via Distributed Split Speculative Decoding",
The 42nd International Conference on Machine Learning (ICML), 2025,Accept

[C1] "EdgePrompt: A Distributed Key-Value Inference Framework for LLMs in 6G Networks",
International Conference on Computer Communications (INFOCOM), 2025,Accept

[C2] "Adaptive Distributed Learning with Byzantine Robustness: A Gradient-Projection-Based Method",
IEEE Global Communications Conference: Communication and Information Systems Security (GlobeCom), 2024, Accept

[J1] "Two-Stage Coded Distributed Edge Learning: A Dynamic Partial Gradient Coding Perspective",
IEEE Transactions on Mobile Computing (TMC), 2024, Accept

[C3] Ping Feng, **Jiahong Ning**, Tingting Yang, Jiabao Kang, JiaLe Wang and Yicheng Li "Federated Optimal Framework with Low-bitwidth Quantization for Distribution System",
International Conference on Distributed Computing Systems, 2023, Accept

[C4] Ping Feng, **Jiahong Ning**, Tingting Yang, Jiabao Kang, JiaLe Wang and Yicheng Li "Federated Optimal Framework with Low-bitwidth Quantization for Distribution System",
International Conference on Distributed Computing Systems, 2023, Accept

[C5] 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

[C6] 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

[J2] 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

[J3] H. Feng, Z. Cui, C. Han, **J. Ning** and T. Yang, "Bidirectional Green Promotion of 6G and AI: Architecture, Solutions, and Platform,"
IEEE Network, 2021

[J4] Jianwu Zeng, **Jiahong ning**, Xia Du, Taesic Kim, Zhaoxia Yang, and Vincent Winstead, "A Four-port DC-DC Converter for a Standalone Wind and Solar Energy System"
IEEE Transactions on Industry Application,10.1109/TIA.2019.2948125, Oct 2019.

[C7] Xia Du, **Jiahong ning** and Jianwu Zeng, "Modeling and Control of a Four-Port Bidirectional DC-DC Converter for a DC Microgrid with Renewable Energy Sources"
IEEE ECCE — IEEE Applied Power Electronic Conference and Exposition,New Orleans, USA, Mar, 2020

[C8] **Jiahong ning**, Jianwu Zeng and Xia Du, "A Four-port Bidirectional DC-DC Converter for Renewable Energy-Battery-DC Microgrid System"
IEEE ECCE — IEEE Energy Conversion Congress and Exposition, Baltimore, USA, Oct, 2019.(Presentation)

[C9] Jianwu zeng, **Jiahong ning**, Taesic Kim, and Vincent Winstead "Modeling and Control of a Four-port DC-DC Converter for a Hybride Energy System"
IEEE APEC — IEEE Applied Power Electronic Conference and Exposition, Losangle, USA, Mar, 2019. (Presentation).

[C10] Dianzhi Yu, jianwu Zeng, Junhui Zhao and **Jiahong ning**, "A Two-stage Four-port Inverter for Hybride Renewable Energy System Integration"
IEEE APEC — IEEE Applied Power Electronic Conference and Exposition, Losangle, USA, Mar, 2019. (Poster).

Industry achievement

[P0] 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 Laboratory, Laboratory 2024, ShenZhen

[P1] Tingting Yang, **Jiahong ning**, Xinghan Wang, Yang yong, Ping Feng "Distributed learning methods, electronic devices, and storage media against Byzantine attacks"
Pengcheng Laboratory, Laboratory patents 2023, ShenZhen

[P2] Tingting Yang, Xin Sun, Chengzhuo Han, **Jiahong Ning**, Zhengqi Cui "Image recognition method, distributed system, device and storage medium",
Pengcheng Laboratory, Laboratory patents 2023, ShenZhen

[P3] Tingting Yang, **Jiahong ning**, Xinghan Wang, Xiaoxiong Zhong "Two-Stage Coded Distributed Learning: A Dynamic Partial Gradient Coding Perspective",
Pengcheng Laboratory, Laboratory patents 2023, ShenZhen

[W1] 6GANA Web Native AI Technology Requirements White Paperr
6GNA, Participate in the white paper content research, writing and publishing, 2022

[W2] 6G Network AI Conceptual Data White Paper
6GNA, Participate in the white paper content research, writing and publishing, 2021

[W3] 6GANA 6G Endogenous AI Network Architecture Ten Questions White Paper
6GNA, Participate in the white paper content research, writing and publishing, 2021

[W4] White paper on ten fundamental issues of integrated computing networkr
6GNA, Participate in the white paper content research, writing and publishing, 2021

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

Minnesota State University
Minnesota, USA

M.S. of Science, Electrical Engineering, 3.75/4
Sep. 2018 - Feb. 2021

- Digital analogue
- Modeling and Simulation
- Advanced Artificial Intelligence
- Smart power grid

Guangxi University
Guangxi, China

B.S. of Science, Electrical Engineering and Automation
Sep. 2014 - Jun. 2018

- C++ programming language
- Signal and System Theory
- Digital Electronic Technology
- Principle and Application of Microcomputer
- Linear Algebra
- Advanced Mathematics
- Data Mining and Analysis

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

Skills

Research Skills	Speculative decoding, Mobie edge computing, Federated Learning, Distributed Computing, Edge Computing, Generative AI, Network Architecture
DevOps	Docker, Kubernetes, Powerflow, Altium Design, Simulink, PyTorch
Programming	Python, Java, Golang, LaTeX, Matlab, Shell, PyTorch, Docker, Kubernetes, DevOps, Powerflow, Altium Design, Simulink.
Languages	Chinese, English, Japanese, Cantonese

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.

2019 Applied Power Electronics Conference(APEC) *LA, Usa*

PRESENTER FOR<Modeling and Control of a Four-port DC-DC Converterfor a Hybride Energy Syst> *Mar. 2019*

- Introduced the novel multi-port topology design of DC-DC converter.
- Introduced the measure and optimize parameters of equipment and do the work shown.

Entrepreneurship and innovation experience

Hong Kong Hackathon Competition *Hong Kong*

Team Third Prize Winner

- Participated in the Hong Kong Hackathon Competition, leading the team to win the third prize by developing innovative solutions.

Nordic Data Processing Company *Norway*

Co-Founder

- Co-founded a company focusing on data processing in the Nordic region, responsible for technical architecture and product development.

Wuxi Cloud Computing Innovation Startup *Wuxi, China*

Co-Founder and Shareholder

- Co-founded an innovation startup in cloud computing, responsible for company operations and technical development.
- Participated in the 2023 Wuxi Taihu Cup International Elite Innovation and Entrepreneurship Competition, showcasing the company project and achieving commendable results.

IEEE Emerging Technologies Initiative on Large Generative AI Models in Telecom (GenAINet) *Global, Online,*

Academic Secretary *Jan. 28, 2024*

- Served as the Academic Secretary for the IEEE Emerging Technologies Initiative on Large Generative AI Models in Telecom (GenAINet), significantly contributing to the advancement of generative AI models within the telecommunications field.
- Facilitated global collaboration and knowledge sharing among leading researchers and industry experts in the field.
- Involved in organizing webinars, workshops, and meetings to disseminate state-of-the-art research and technological developments.
- Assisted in developing academic and industry standards and guidelines for ethical and effective implementation of AI technologies in telecommunications.
- https://www.comsoc.org/about/committees/emerging-technologies-initiatives/large-generative-ai-models-telecom-genainet

Kubeedge wireless, special working group

Github

Responsible person

Nov. 2020 - PRESENT

- Transform the edge core and sink the cloud capability to the edge segment
- Focus on the collaboration between edges and open up the communication between edge nodes
- Based on the characteristics of wireless communication network, expand the actual edge scene and propose requirements for KubeEdge iteration
- Track the 3GPP protocol and iterate the edge computing power in wireless scenarios
- Communicate with the university side and the technical side to match the needs and capabilities of both sides

GSMA-MWC Shanghai Exhibition 2021

Shanghai, China

Attendee

Feb. 2021

- Visited the MWC Shanghai 2021, learned more about the world's most advanced computing communication electronics industry technology, and had further thinking about the future technology route.

GXU, Geek Electronic Science and Technology Association

Guangxi, China

Member/Vice Director

Sep. 2014 - May. 2017

- Do the learning about designing, drawing and soldering circuit boards for smart cars.
- Learn embedded compilation and design intelligent feedback mode of robot.

China-Japan environmental protection friendly exchange meeting in Mie University

Japan

Speaker

Oct, 2017

- Present views and research materials on global climate and environmental change.
- Exchanges the cultural difference between China and Japan, discusses how to protect the environment together.