宁嘉鸿 (JIAHONG NING)

手机：(+86) 13768410701 · 邮箱：njh1195@gmail.com
https://github.com/JasonNing96 ·

个人简介

专注 **MEC 移动边缘计算**与 **LLM 推理加速**：涵盖 KV-Cache 内存一致性管理、Speculative Decoding、端-边-云协同卸载与资源调度，研究聚焦 **受限条件下的资源分配与系统优化**。工作上深刻理解 **市场需求与学术研究之间的 GAP**，具备 **带队与项目管理**能力，曾主导团队进行 **国家重点研发计划**的申报与推进，辅助导师推进项目**立项—实施—结项**周期闭环。

- 边缘/云协同 LLM 推理系统：KV 缓存、内存一致性管理与 LLM 模型推理加速技术。
- 6G 通信网络下的生成式 AI 资源调度与部署优化。
- 强化学习、联邦学习、分布式网络，计算一体化架构。

教育背景

新加坡信息通信研究所（A*STAR, I2R），新加坡，联合培养博士 2024.05 – 至今

联培导师：Sumei Sun（新加坡工程院院士）

大连海事大学，中国，交通控制与运输工程博士 2021.09 – 至今

博士导师：杨婷婷（鹏城实验室）

Minnesota State University，USA，电力电子信息技术硕士 2018.09 – 2021.01

硕士导师：JianWu Zeng, Vincent Winstead

广西大学，中国，电气工程及其自动化学士 2014.09 – 2018.06

科研经历

新加坡 **A*STAR 信息通信研究所 (I2R)**(外方导师：Sumei Sun(新加坡工程院院士)) 2024.05–至今

- **LLM 边缘分布式推理算法（EdgePrompt）**：面向无线网络的 *分布式 KV-Cache* 推理框架；设计 KV 路由与一致性管理、分片与流式传输策略，在带宽受限场景降低端到端时延并提升吞吐。
- **分布式投机采样技术 (DSSD)**：结合分支式投机解码与端-边协作，创新设计网络上/下行逻辑，提高至少 2X 模型加速比。受华为资助参加 42th ICML 机器学习会议。
- **混合任务协同卸载算法设计（Hybrid Hierarchical Offloading）**：针对 *Decoder-based* 生成模型，构建 UE→Edge→Cloud 的层级卸载与显存/带宽约束下的调度；以强化学习学习任务粒度与路径，平衡 QoS、能耗与成本。
- **UAV 辅助 LLM 进行边缘推理**：在应急/通信受限等弱覆盖场景，联合 UAV 中继与 DSSD 策略进行链路自适应与任务切分，提升边缘用户服务质量与灵活性。

深圳**鹏城实验室** (博士导师：杨婷婷) 2023.03–2024.07

- 国家重点研发计划骨干“**6G 通用 AI 智能**”，研究大模型架构在通信网络问题中的部署和协议设计。围绕算网一体接口、网络大模型能力与部署开展工作。
- 联合华为等多企业部分撰写《**IMTM2030 报告**》，洞察未来技术趋势。
- **可靠分布式学习**：两阶段编码分布式学习（*Two-Stage Coded DL*）与 *Byzantine-robust* 联邦学习，面向边缘训练/资源异构与失效容错。
- 设计端-边-云协同硬件架构，搭建通信与计算一体化平台，基于 **KubeEdge/Kube-Wireless** 实现端-边-云协同编排；感知带宽/显存/能耗的调度策略，支持推理并发、KV 迁移与一致性控制。获评 **2023 年度通信十大进展**。

华为 2012 无线技术实验室（合作导师：卢建明 (华为 Fellow)) 2022.03–2023.02

- 跟踪通信理论与 **3GPP** 标准进展，研究无线网络中机器学习/深度学习的应用路径。
- 参与国家重点研发计划 **Network4AI** 专项，面向通信资源调度与智能编排的算法设计与验证。

- 共建 **Huawei KubeEdge** 开源社区，发起并推动 **Kube-Wireless** Group，面向无线场景的容器网络。
- 研究容器化边缘网络与管理工具，探索与 **5G** 网络融合的资源编排与可观测性体系。

实习经历

CloudBu 创新实验室，华为云计算技术有限公司，深圳 2021.06–2021.09

- 负责 Kubernetes 运维与日志系统维护，Golang 实现 Edge-Mesh 插件。
- 管理 KubeEdge 社区：收集用户需求并提交功能建议。
- 牵头无线工作组探索新场景，研究无线网络动态拓扑。

Wireless Tech Lab，华为 2012 无线技术实验室，深圳 2020.07–2021.05

- 参与 3GPP 标准和 AI 优化相关工作，Network4AI 容器网络项目。
- 开发 SLAM 算法原型，推动智能感知与定位研究。
- 协调项目进度与需求，撰写技术文档并进行跨团队交流。

期刊论文

Jiahong Ning, Aiming Li, Ning Huang, Tingting Yang, Gary Lee, Sumei Sun, "MARHLO: Multi-Agent RL-Based Hybrid Offloading for Maritime MEC Network", *IEEE Transactions on Network Science and Engineering (TNSE)*, Under Review, 2025.

Jiahong Ning, Tingting Yang, Yongyi Su, "SkyDSSD: A UAV-Assisted Distributed Split Speculative Decoding Framework for Edge Inference", *IEEE Transactions on Cognitive Communications and Networking (TCCN)*, Under Review, 2025.

Jiahong Ning, Ce Zheng, Tingting Yang, "DeAOff: Dependence-Aware Offloading of Decoder-Based Generative Models for Edge Computing", *IEEE China Communications (ChinaCom)*, 2025, Accept.

Tingting Yang, Ping Feng, Qixin Guo, Jindi Zhang, Xiufeng Zhang, **Jiahong Ning**, Xinghan Wang, Zhongyang Mao, "AutoHMA-LLM: Efficient Task Coordination and Execution in Heterogeneous Multi-Agent Systems Using Hybrid Large Language Models", *IEEE Transactions on Cognitive Communications and Networking* **11**(2): 987–998, 2025.

Tingting Yang, Xinghan Wang, **Jiahong Ning**, Yuanyuan Yang, Guoming Tang, Fangming Liu, "Two-Stage Coded Distributed Edge Learning: A Dynamic Partial Gradient Coding Perspective", *IEEE Transactions on Mobile Computing (TMC)*, 2024, Accept.

Tingting Yang, **Jiahong Ning**, Dapeng Lan, Jiawei Zhang, Yang Yang, Xudong Wang, Amir Taherkordi, "KubeEdge Wireless for Integrated Communication and Computing Services Everywhere", *IEEE Wireless Communications*, 29(2):140–145, 2022.

Hailong Feng, Zhengqi Cui, Chengzhuo Han, **Jiahong Ning**, Tingting Yang, "Bidirectional Green Promotion of 6G and AI: Architecture, Solutions, and Platform", *IEEE Network*, 35(6):57–63, 2021.

Jianwu Zeng, **Jiahong Ning**, Xia Du, Taesic Kim, Zhaoxia Yang, Vincent Winstead, "A Four-port DC-DC Converter for a Standalone Wind and Solar Energy System", *IEEE Transactions on Industry Applications*, Oct. 2019.

会议论文

Jiahong Ning, Ce Zheng, Tingting Yang, "DSSD: Efficient Edge-Device Deployment and Collaborative Inference via Distributed Split Speculative Decoding", *In The 42nd International Conference on Machine Learning*

(ICML), 2025, Accept.

Jiahong Ning, Pengyan Zhu, Ce Zheng, Gary Lee, Sumei Sun, Tingting Yang, "EdgePrompt: A Distributed Key-Value Inference Framework for LLMs in 6G Networks", *In 2025 IEEE International Conference on Computer Communications (INFOCOM)*, London, UK, 2025, Accept.

Dongxiao Hu, Dapeng Lan, Yu Liu, **Jiahong Ning**, Jia Wang, Yun Yang, Zhibo Pang, "Embodied AI Through Cloud-Fog Computing: A Framework for Everywhere Intelligence", *In 2024 IEEE International Symposium on Industrial Electronics (ISIE)*, Ulsan, South Korea, 2024, pp. 1–4.

Zechen He, Jiale Wang, Ping Feng, **Jiahong Ning**, Tingting Yang, "A Low-Rank Approach of MIMO Optimization for Edge Smart Ports", *In 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, Singapore, 2024, pp. 1–5.

Xinghan Wang, Cheng Huang, **Jiahong Ning**, Tingting Yang, Xuemin (Sherman) Shen, "Adaptive Distributed Learning with Byzantine Robustness: A Gradient-Projection-Based Method", *In 2023 IEEE Global Communications Conference (GLOBECOM)*, Kuala Lumpur, Malaysia, 2023, pp. 7520–7525.

Xinghan Wang, Xiaoxiong Zhong, **Jiahong Ning**, Tingting Yang, Yuanyuan Yang, Guoming Tang, Fangming Liu, "Two-Stage Coded Distributed Learning: A Dynamic Partial Gradient Coding Perspective", *In 2023 IEEE International Conference on Distributed Computing Systems (ICDCS)*, Hong Kong, China, 2023, pp. 942–952.

Ping Feng, **Jiahong Ning**, Tingting Yang, Jiabao Kang, Jiale Wang, Yicheng Li, "Federated Optimal Framework with Low-bitwidth Quantization for Distribution System", *In 2023 IEEE Global Communications Conference (GLOBECOM)*, Kuala Lumpur, Malaysia, 2023, pp. 2039–2044.

Jiahong Ning, Jiale Wang, Ping Feng, Tingting Yang, "A Distributed Framework for the Ocean IoT Network". *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2023*, 1-6

Chengzhuo Han, Tingting Yang, Xin Sun, **Jiahong Ning**, "CLMD: Detection and Prevention of Poisoning Attacks for Federated Learning in Maritime Communication Network", *In 2023 IEEE International Conference on Communications (ICC)*, Rome, Italy, 2023, pp. 19–25.

Xia Du, **Jiahong Ning**, Jianwu Zeng, "Modeling and Control of a Four-Port Bidirectional DC-DC Converter for a DC Microgrid with Renewable Energy Sources", *In 2020 IEEE Applied Power Electronics Conference and Exposition (APEC)*, New Orleans, USA, Mar. 2020.

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

Jianwu Zeng, **Jiahong Ning**, Taesic Kim, Vincent Winstead, "Modeling and Control of a Four-port DC-DC Converter for a Hybrid Energy System", *In 2019 IEEE Applied Power Electronics Conference and Exposition (APEC)*, Los Angeles, USA, Mar. 2019.

行业专利

郑策，王星翰，**宁嘉鸿**，杨勇，黄宁，杨婷婷，"基于投机采样的大模型分布式推理方法（DSSD）"，中国发明专利（公开号：CN120373477A；申请号：CN202510885627.8；鹏城实验室。

俸萍，杨婷婷，毛忠阳，**宁嘉鸿**，黄建波，"基于 WasmEdge 的移动端大模型自适应软件"，中国发明专利，状态：交底完成，**拟申请**（暂未分配申请号）。

宁嘉鸿，杨婷婷，王星翰，“一种分布式联邦学习的两阶段编码方法及相关装置”，中国发明专利，专利号：202310273893.6。

宁嘉鸿，杨婷婷，王星翰，“抗拜占庭攻击的分布式学习方法、电子设备及存储介质”，中国发明专利，专利号：202311403763.6。

宁嘉鸿，荷泽晨，杨婷婷，“一种基于低秩适应的边缘近海信道状态的传输方法及系统”，中国发明专利，专利号：202411916653.4。

孙鑫，杨婷婷，**宁嘉鸿**，等“基于多智能体协作的海上通信网络路由方法及网络系统”，中国发明专利，专利号：202411916319.9。

祝朋艳，杨婷婷，**宁嘉鸿**，“一种云-边协同的推理方法及推理系统”，中国发明专利，专利号：202411916320.1。

杨婷婷、**宁嘉鸿**、何泽晨、王佳乐，“**2023 年通信领域十大技术进展：通信与计算实现网络架构、协议和实验平台的协同优化**”，中国通信协会 *2023* 年文，鹏程实验室，2024 年，深圳。

行业白皮书

全球 6G 技术大会/多机构，“2024 年 10.0A GPT 与通信白皮书”, *White Paper*, 2024。（撰写“通信，算力协同”与“实验平台”相关章节，整合鹏城实验室案例与原型数据。）

6GANA（6G Alliance of Network AI），“6G Network Native AI Technical Requirement White Paper”, *White Paper*, 2022。（主笔“Native AI 能力需求与指标体系”，“端-边-云协同实验平台”章节；梳理鹏城实验在网络 AI 方面的实践。）

6GANA（6G Alliance of Network AI），“6G Network AI Concept and Terminology（v0.3）”, *White Paper*, 2021。（术语体系与分类框架整理；补充鹏城实验室实践积累相关 AI 术语与参考案例。）

6GANA（6G Alliance of Network AI），“Ten Questions of 6G Native AI Network Architecture”, *White Paper*, 2021。（参与架构十问的场景与接口设计讨论；撰写“算网一体化与网络原生 AI”部分。）

6GANA（6G Alliance of Network AI），“6G Data Service —Concept and Requirements”, *White Paper*, 2023。（主笔数据服务能力模型、数据闭环与评测流程；补充鹏城实验室数据治理与流水线实践。）

6GANA（6G Alliance of Network AI），“Knowledge-Defined Orchestration and Management”, *White Paper*, 2023。（撰写意图驱动编排（KDO/M）与资源编排案例；引入鹏城实验室 KubeEdge/Kube-Wireless 场景。）

6GANA（6G Alliance of Network AI），“Whitepaper on Distributed Learning of 6G”, *White Paper*, 2024。（分布式/联邦学习在 6G 的体系与关键技术；融入鹏城实验室端-边-云协同训练平台结果与图表。）

技能

- 模型与系统：LLM Serving（vLLM / TGI / TensorRT-LLM），KV-Cache 管理，Speculative Decoding，Prompt/Cache Routing，Streaming Inference。
- 平台与工程：Kubernetes，**KubeEdge**，Docker，CI/CD，Linux，Grafana
- 算法与优化：分布式与联邦学习（Byzantine-robust），RL for Offloading，凸优化/Gurobi，PyTorch。
- 语言与协作：Python，Golang，Shell；跨团队协作与项目管理（国重申报、开源社区推进、交付）。

语言

普通话 (母语)；英语 (流利, TOEFL 96)；日语 (初级)；粤语 (流利)