石金梁

基本信息

电话：15603381691　　邮箱：shijinliang@bupt.edu.cn

年龄：28 岁　　政治面貌：中共党员

研究方向：**大模型基础设施建设与训推加速、高性能计算、云计算**

教育背景

北京邮电大学	计算机科学与技术	博士	导师：李士刚	2023.09 - 2027.06
北京科技大学	计算机科学与技术	硕士 (保研)	导师：李建江	2020.09 - 2023.06
河北农业大学	计算机科学与技术	本科		2016.09 - 2020.06

学术成果

唯一第一作者学术论文

- FastInfer: Breaking the Low-to-Moderate Sparsity Barrier for Efficient LLM Inference.
  SC '26（高性能计算领域国际顶级会议，CCF A 类会议），2026.
- FlashSparse: Minimizing Computation Redundancy for Fast Sparse Matrix Multiplications on Tensor Cores.
  PPoPP '25（高性能计算领域国际顶级会议，CCF A 类会议），2025，**北京邮电大学首篇 PPoPP**.
- UltraGNN: A Sparse-Operator-Aware Framework for Accelerating Graph Neural Networks on Tensor Cores.
  IEEE TPDS '26（CCF A 类，Top 期刊），2026.
- Libra: Unleashing GPU Heterogeneity for High-Performance Sparse Matrix Multiplication.
  IEEE TPDS（CCF A 类，Top 期刊），2026，返修已完成.
- New YARN Sharing GPU Based on Graphics Memory Granularity Scheduling.
  Parallel Computing（CCF B 类期刊），2023.

其他合作学术论文

- Omnia: Efficient RAG Serving through Speculative Scheduling.
  HPDC '26（高性能计算领域国际顶级会议，CCF A 类会议），2026.

专利成果

- 数据处理方法、装置和系统。华为（技术转让）；学生第一发明人；已获得初步审查合格通知书。
- TCU 与 CUDA 核心上混合计算任务调度方法及设备。北京邮电大学；学生第一发明人；已授权。

奖励及荣誉

- **北京邮电大学博士研究生国家奖学金** 2025
- **北京邮电大学首篇 PPoPP 论文，入选计算机学院年度十大新闻** 2025
- 北京邮电大学一等学业奖学金 2025
- 北京邮电大学二等学业奖学金 2024

项目经历

高性能大模型推理系统研究（国家自然科学基金项目） 2024.01 - 2026.03

- 针对大模型推理中的计算、访存与多 GPU 通信瓶颈，从**稀疏算子优化与分布式推理系统**两个层次开展软硬件协同设计；
- 研发 **FastInfer 非结构化稀疏大模型推理框架**，提出适用于低至中等稀疏度的新型编码与存储格式，首次在数据中心级与消费级 GPU 上突破低至中等稀疏度难以获得实际加速的性能瓶颈；
- 参与 **Dynamo-MoE 分布式大模型推理系统**研发，针对 MoE 专家动态激活引起的负载不均问题，研究实时负载均衡、张量并行与专家并行自适应切换以及计算–通信重叠机制；
- **成果：**FastInfer 以第一作者发表于 SC '26（CCF A 类国际顶级会议）；Dynamo-MoE 相关成果发表于 HPDC '26（CCF A 类国际顶级会议）；以学生第一发明人完成授权发明专利 1 项。

GNN 训推加速研究技术合作项目（华为校企合作） 2023.10 - 2025.01

- 面向图神经网络训练与推理中的图结构稀疏、动态权重及异构算力利用问题，围绕 **NVIDIA GPU 与国产昇腾 AI 处理器**开展计算范式、核心算子与模型系统协同优化；
- 研发 **FlashSparse 与 Libra 高性能稀疏算子框架**，设计细粒度稀疏映射与数据重排机制，并协同利用 Tensor Core 与 CUDA Core 实现异构任务映射与调度；
- 研发 **UltraGNN 模型加速框架**，首次提出以子图为中心的计算范式，支持动态权重图神经网络在 Tensor Core 上高效训练与推理；
- 面向**国产昇腾 AI 处理器**完成 GNN 核心算子迁移、模型适配与性能调优，形成覆盖 NVIDIA GPU 与国产 AI 芯片的异构加速方案；
- **成果：**FlashSparse 与 UltraGNN 分别以第一作者发表于 PPoPP '25（CCF A 类国际顶级会议）和 IEEE TPDS '26（CCF A 类期刊）；以学生第一发明人完成华为技术转让专利 1 项。

地震解释智能化软件平台建设（中国石油勘探开发研究院西北分院） 2020.10 - 2022.09

- 作为核心研发成员面向油气勘探地震解释业务，构建融合**大数据、云计算与人工智能**的地震解释智能化软件平台，形成覆盖海量地震数据存储、异构算力调度、深度学习训练及作业管理的一体化平台；
- 扩展 Hadoop YARN 的 **GPU 资源感知与异构调度能力**，并集成 DeepX 支持分布式深度学习任务统一调度；完成 Keras、PyTorch 等框架适配，将初至拾取、地震去噪、三维断层检测等模型由单机训练迁移至多节点 GPU 分布式训练；
- 基于 HDFS、HBase、Hive、MySQL 构建海量地震数据分布式存储与管理体系，并结合 Airflow 实现深度学习任务的工作流调度、状态监控与批处理管理，支撑地震数据与智能计算任务协同运行；
- **成果：**合作开展 GPU 资源调度研究，以第一作者发表相关成果于 CCF B 类期刊 Parallel Computing。

实习经历

腾讯科技有限公司 - 大模型训练系统优化与智能体研发（青云顶尖人才计划） 2026.07 - 2026.09

- 面向广告模型训练场景开展端到端性能分析、算子设计与系统优化;
- 从 0 搭建 Kernel Agent 自动优化流程;
- 实现 CUDA Kernel 自动生成、正确性验证、性能评测与迭代;
- 针对广告模型中典型计算开展融合算子优化。

华为技术有限公司 - 图神经网络模型算子开发 2023.08 - 2024.09

- 参与 GNN 模型性能分析与 GPU 稀疏算子研发;
- 负责高性能算子集成、模型适配及端到端性能验证;
- 支撑相关研究成果在实际 GNN 工作负载中的性能评测与优化。

百度在线网络技术有限公司 - 云计算开发 2021.06 - 2021.08

- 参与百度移动生态分布式云数据平台建设;
- 负责亿级数据清洗、分析与数据挖掘;
- 建设并维护 Hive 分布式数据仓库。

学术报告

- "FastInfer: Breaking the Low-to-Moderate Sparsity Barrier for Efficient LLM Inference."
  国际高性能计算、网络、存储与分析大会论文集 (SC '26), 2026.11.15-20, 美国芝加哥.
- "FlashSparse: Minimizing Computation Redundancy for Fast Sparse Matrix Multiplications on Tensor Cores."
  2025 CCF 全国高性能计算学术年会 (HPC China '25), 2025.08.14-16, 中国鄂尔多斯.
- "FlashSparse: Minimizing Computation Redundancy for Fast Sparse Matrix Multiplications on Tensor Cores."
  第 30 届并行程序设计原理与实践年度研讨会论文集 (PPoPP '25), 2025.03.01-07, 美国拉斯维加斯.
- "FlashSparse: Minimizing Computation Redundancy for Fast Sparse Matrix Multiplications on Tensor Cores."
  第二届高性能计算青年讨论研讨会, 2024.11.27-12.01, 中国重庆.

自我评价

- 具备丰富的 GPU 高性能计算，稀疏计算和大模型系统优化研究经验;
- 具备较强的科研创新与学术论文写作能力，以唯一第一作者在 SC '26、PPoPP '25 和 IEEE TPDS '26 等 CCF A 类国际顶级会议及期刊发表论文。