康复大学
University of Health and Rehabilitation Sciences

专题讲座

面向LLM的GPU显存优化

报告人：赵梦莹 教授

时　间：2026年5月12日（周二）14:30

地　点：康复大学11号楼108会议室

承办单位：康复科学与工程学院

专家简介

赵梦莹，博士，教授，博士生导师，山东大学齐鲁青年学者。分别于2011年和2015年在山东大学和香港城市大学获得学士和博士学位，2015年加入山东大学。

主要研究方向为计算机体系结构、新型非易失性存储、异构计算。主持国家自然科学基金、国家重点研发计划子课题、山东省自然科学基金项目7项。在国际重要学术会议（如HPCA、DAC、RTSS、DATE、CODES+ISSS等）和期刊（如TC、TCAD、TODAES、TECS等）上发表论文100余篇。获ICESS best paper award（2016）、ASPDAC SRF Best Poster Award（2021）、ACM青岛新星奖（2021）、山东电子学会科学技术青年奖（2022），入选山东省青年科技人才托举工程（2024）。指导学生获DAC国际硬件比赛（FPGA组）第三名（2017）、DAC国际系统设计大赛（GPU组）第三名（2018）、EDA精英挑战赛全国二等奖（2024）。担任国际会议NVMSA2026 General Chair、MSC 2022 Program Chair、国际会议SAC 2023 CPS Track chair、CCF B类国际期刊TECS Guest Editor (2022-2023)、TODAES Guest Editor (2023-2024)、IEEE Embedded Systems Letters Associate Editor（2022-2024），CCF体系结构专委、信息存储专委、嵌入式系统专委委员。

热忱欢迎广大师生参加！

交叉共融 求是创新

康复大学
University of Health and Rehabilitation Sciences

专题讲座

面向个性化片上智能的软硬件协同设计

报告人：贾振格 教授

时　间：2026年5月12日（周二）14:30

地　点：康复大学11号楼108会议室

承办单位：康复科学与工程学院

专家简介

贾振格，山东大学计算机科学与技术学院教授、博导，国家级青年人才计划入选者，山东省泰山学者青年专家，山东大学杰出青年学者。2017年于澳大利亚国立大学获计算机科学专业学士学位，2022年于美国匹兹堡大学获电子与计算机工程专业博士学位，2022年至2024年于美国圣母大学计算机科学与工程学院从事博士后研究。

主要从事边缘智能的算法与系统协同设计及其医学工程应用，致力于面向端侧智能芯片的高效机器学习技术研究，针对片上神经网络存在的“难部署”、“难训练”和“难适配”问题，分别提出：(1) 面向嵌入式芯片的片上神经网络推理架构；(2) 微型神经网络高泛化训练方法；(3) 系统与算法协同设计的端侧自监督个性化模式，实现了高效片上推理和神经网络片上自主个性化。主持相关国家自然科学基金项目2项，省部级项目2项，横向项目1项。发表学术论文40余篇，其中第一作者发表Nature Machine Intelligence（影响因子25.8）3篇，通讯作者发表Nature Electronics（影响因子34.3）1篇，第一作者发表CCF-A类论文6篇。获DAC’25（CCF-A类会议）系统设计大赛第二名，ICCAD’24（CCF-B类会议）大模型硬件设计国际大赛第二名。基于相关研究成果，开发了针对致命性室性心律失常的异常检测组件，成功应用于国产自研植入式心率转复除颤器(ICD)，并于2023年12月在南京鼓楼医院开启临床试验。

热忱欢迎广大师生参加！

交叉共融 求是创新

康复大学
University of Health and Rehabilitation Sciences

专题讲座

高性能智能体系统优化技术

报告人：张余豪 副研究员

时　间：2026年5月12日（周二）14:30

地　点：康复大学11号楼108会议室

承办单位：康复科学与工程学院

专家简介

张余豪，天津大学计算机科学与技术学院副研究员。主要研究方向为存储系统和AI系统，具体研究：（1）基于新型存储(HBM/PIM/SmartSSD)和高速互联(CXL/RDMA)设备的分布式存储系统，即面向新型存储硬件（例如：非易失性存储器[Flash/3DXpoint/ReRAM]、近内存计算模块[DRAM-PIM模块]、近存储计算设备[SmartSSD]等）以及新型高速互联协议（例如：CXL和RDMA），研究并设计高吞吐、低延迟、低功耗、大容量的分布式存储系统，高效的支撑大数据及人工智能应用；（2）高效能的大模型推理和训练系统，即面向大模型，采用软硬件协同设计的方法，研究大模型系统优化技术，提升大模型资源利用率和性能。

主持或参与国家自然科学青年基金、国家重点研发计划等国家级项目以及华为、阿里等企业委托项目。目前已在OSDI、USENIX ATC、EuroSys、MICRO、DAC、NSDI、TCAD、TOS、TC等国际权威会议和期刊上发表学术论文20余篇。

热忱欢迎广大师生参加！

交叉共融 求是创新

面向可定制边缘智能处理单元的软硬件协同优化

报告人：张浩 助理研究员

时　间：2026年5月12日（周二）14:30

地　点：康复大学11号楼108会议室

承办单位：康复科学与工程学院

专家简介

张浩，清华大学计算机系助理研究员。主要研究方向为非易失可重构系统的软硬件协同设计，聚焦端侧高性能可重构计算平台的高效实现与系统优化。针对非易失存储器和可重构架构在部署、重构及适配上的挑战，提出了片上存算一体化架构设计、高效神经网络推理方法以及系统与硬件协同优化策略，实现了可重构平台的高性能、低功耗和可靠运行。在国际重要学术期刊和会议上发表论文11篇，获授权发明专利4项，并获得 DAC 2023 Young Fellow、ASP-DAC 2021 Best SRF Poster 等学术荣誉。

本报告围绕面向可定制边缘智能处理单元的软硬件协同优化展开，主要探讨在边缘智能设备中实现高效、灵活计算的方法，以满足低功耗、多任务、多应用及定制化 AI 功能的需求。内容涵盖新一代 NPU 编译器的设计与优化策略，通过算子映射、调度优化和内存管理提升边缘 AI 推理性能与能效；以及可重构存内计算架构及其设计自动化优化方法，通过灵活硬件支持不同任务的高效执行，同时提升系统整体吞吐量与能效。

热忱欢迎广大师生参加！

交叉共融 求是创新

康复大学
University of Health and Rehabilitation Sciences

专题讲座

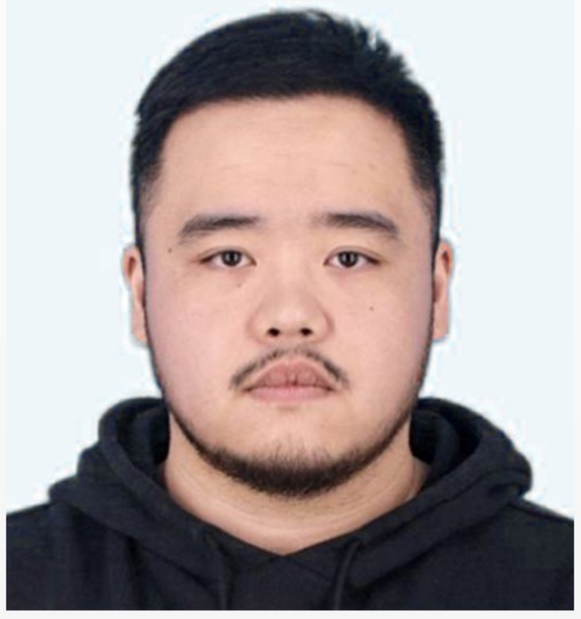
大模型训推高效存储的容错技术探索

报告人：郑汇川 助理研究员

时　间：2026年5月12日（周二）14:30

地　点：康复大学11号楼108会议室

承办单位：康复科学与工程学院



专家简介

郑汇川，清华大学计算机系助理研究员。主要研究方向为新型非易失存储、电子设计自动化、机器学习系统等。针对MLC 非易失存储器集成于FPGA时面临的逻辑拥塞加剧、配置能耗高以及硬件寿命短等挑战，通过软硬件协同优化方法，实现了MLC NVFPGA在综合效率、系统能效与运行可靠性上的协同提升。以第一作者身份在CCF-A类期刊TCAD、CCF-A类会议DAC、CCF-B类会议DATE上发表论文3篇，在国际重要学术期刊和会议上共发表论文10篇，获授权国家发明专利3项。

本报告围绕大模型训推高效存储的容错技术展开，主要探讨存储系统在大规模集群环境下支撑高可靠与高效率大模型推训的方法。针对训练场景，介绍高性能检查点（Checkpoint）存储系统，通过内核旁路、零拷贝及异步持久化策略，以降低容错过程对训练进度的阻塞。针对推理阶段，本报告聚焦状态容错与引擎快速启动机制，介绍基于存算融合思想的KV Cache压缩恢复与迁移等技术，以掩盖模型状态加载延迟。通过训推语义与底层存储的深度融合，本报告尝试探讨如何消除系统间的语义鸿沟，实现高算力利用率与低容错开销的统一。

热忱欢迎广大师生参加！

交叉共融 求是创新