王鑫凯，博士后研究员

unbreakablewxk@sjtu.edu.cn https://wang-xinkai.github.io
(+86)152-0196-7357 上海市闵行区东川路 800 号

研究方向：计算机体系结构、计算机系统

研究兴趣

我的研究方向是计算机体系结构与系统，主要聚焦于高效可靠的数据中心管理优化，具体包括：

- **高能效**：如何更好的建模、分配、调度异构受限的硬件资源以提升系统每瓦特性能？
- **高韧性**：如何更好的分析、检测、消除频发的软硬件故障以保证系统的稳定运行？

目前我主要关注以下两个研究课题：

- **代理式人工智能基础设施的高能效设计探索（Infrastructure for Agentic AI）**
- **高韧性基础设施管理智能体设计（Agentic Management for Infrastructure）**

教育/工作经历

- **博士后研究员**，计算机学院，上海交通大学 2026.07 至今
  合作导师：**过敏意教授**，先进计算研究所
- **工学博士**，计算机科学与技术，计算机学院，上海交通大学 2021.09 – 2026.06
  导师：**李超教授**，可持续架构与智能实验室
  毕业论文：“低干扰高能效的数据中心可管理性优化研究”
  毕业荣誉：上海交通大学优秀博士毕业生
- **访问学者**，计算机科学与技术，根特大学，比利时 2025.10 – 2026.03
  导师：***Lieven Eeckhout* 教授**，*ELIS* 系
- **工学学士**，计算机科学与技术，电子信息与电气工程学院，上海交通大学 2017.09 – 2021.06
  毕业论文：“云原生时代的微服务并行化研究”，计算机系优秀毕业论文
  毕业荣誉：上海交通大学优秀本科毕业生、致远荣誉学士学位

论文发表

本人共发表 18 篇论文，包括 8 篇第一作者论文（4 篇 CCF-A 类，3 篇 CCF-B 类）；申请 11 项发明专利，其中 6 项已授权。另外，本人有 2 篇第一作者 CCF-A 类论文在投。

会议论文

1. **Xinkai Wang**, Chao Li, Yiming Zhuansun, Jinyang Guo, Xiaofeng Hou, Jing Wang, et al., “Aum: Unleashing the efficiency potential of shared processors with accelerator units for llm serving,” in *Proc. the 32nd IEEE International Symposium on High-Performance Computer Architecture (**HPCA 2026, CCF-A**, 119/602=19.7%).* DOI: 10.1109/HPCA68181.2026.11408539

2. **Xinkai Wang**, Xiaofeng Hou, Chao Li, Yuancheng Li, Du Liu, Guoyao Xu, et al., “Exist: Enabling extremely efficient intra-service tracing observability in datacenters,” in *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (**ASPLOS 2025, CCF-A**, 160/912=17.5%)*, pp. 355–372. DOI: 10.1145/3676641.3716283

3. **Xinkai Wang**, Yiming Zhuansun, Chao Li, Jing Wang, Xiaofeng Hou, Lingyu Sun, et al., "Asymserve: Demystifying and optimizing llm serving efficiency on cpu acceleration units," in *Advanced Parallel Processing Technologies* (***APPT 2025, CCF-C**, 17/74=22.9%*), pp. 231–245. DOI: 10.1007/978-981-95-1021-4_17

4. **Xinkai Wang**, Chao Li, Lingyu Sun, Qizheng Lyu, Xiaofeng Hou, Jingwen Leng, et al., "Sheeo: Continuous energy efficiency optimization in autonomous embedded systems," in *2024 IEEE 42nd International Conference on Computer Design* (***ICCD 2024, CCF-B, 最佳论文提名**, 74/277=26.7%*), pp. 496–503. DOI: 10.1109/ICCD63220.2024.00082

5. **Xinkai Wang**, Hao He, Yuancheng Li, Chao Li, Xiaofeng Hou, Jing Wang, et al., "Not all resources are visible: Exploiting fragmented shadow resources in shared-state scheduler architecture," in *Proceedings of the 2023 ACM Symposium on Cloud Computing* (***SoCC 2023, CCF-B**, 40/133=30%*), pp. 109–124. DOI: 10.1145/3620678.3624650

6. **Xinkai Wang**, Chao Li, Lu Zhang, Xiaofeng Hou, Quan Chen, and Minyi Guo, "Exploring efficient microservice level parallelism," in *2022 IEEE International Parallel and Distributed Processing Symposium* (***IPDPS 2022, CCF-B**, 123/474=25.9%*), pp. 223–233. DOI: 10.1109/IPDPS53621.2022.00030

7. Chang Liu, Xiaofeng Hou, Cheng Xu, **Xinkai Wang**, Jiacheng Liu, Chao Li, et al., "Locmore: Locating more bursty latency-critical jobs on resource-constrained nodes," in *IEEE/ACM International Symposium on Quality of Service* (***IWQoS 2026, CCF-B***).

8. Mingxuan Zhang, Xiaofeng Hou, Jiacheng Liu, Chang Liu, **Xinkai Wang**, Chao Li, et al., "Starkserve: A framework for elastic serverless llm inference at the extreme edge," in *IEEE/ACM International Symposium on Quality of Service* (***IWQoS 2026, CCF-B***).

9. Jinyang Guo, **Xinkai Wang**, Jing Wang, Xiaofeng Hou, Chao Li, and Minyi Guo, "Tricooling-sim: Efficient thermal simulation for high-density micro ai data centers," in *Network and Parallel Computing* (***NPC 2025, CCF-C, Best Student Paper***), pp. 227–239. DOI: 10.1007/978-3-032-10466-3_19

10. Jing Wang, Taolei Wang, Juntao Huang, Yibo Liu, **Xinkai Wang**, Marius Kreutzer, et al., "Accelerating large-scale out-of-gpu-core gnn training with two-level historical caching," in *International Symposium on Advanced Parallel Processing Technology* (***APPT 2025, CCF-C, Distinguished Artifact Award***). DOI: 10.1007/978-981-95-1021-4_22

11. Lingyu Sun, Chao Li, Xiaofeng Hou, Tianhao Huang, Cheng Xu, **Xinkai Wang**, et al., "Jigsaw: Taming bev-centric perception on dual-soc for autonomous driving," in *2024 IEEE Real-Time Systems Symposium* (***RTSS 2024, CCF-A***), pp. 280–293. DOI: 10.1109/RTSS62706.2024.00032

12. Du Liu, Jing Wang, **Xinkai Wang**, Chao Li, Lu Zhang, Xiaofeng Hou, et al., "Improving the efficiency of serverless computing via core-level power management," in *2024 IEEE 24th International Symposium on Cluster, Cloud and Internet Computing* (***CCGrid 2024, CCF-C***), pp. 125–135. DOI: 10.1109/CCGrid59990.2024.00024

13. Lu Zhang, Chao Li, **Xinkai Wang**, Weiqi Feng, Zheng Yu, Quan Chen, et al., "First: Exploiting the multi-dimensional attributes of functions for power-aware serverless computing," in *2023 IEEE*

International Parallel and Distributed Processing Symposium (***IPDPS 2023, CCF-B***), pp. 864–874. DOI: 10.1109/IPDPS54959.2023.00091

期刊论文

1. 王鑫凯, 李超, 黄天浩, 郭进阳, 侯小凤, 徐国耀, et al., "面向混部高压数据中心的应用执行追踪：高效设计管理与智能分析诊断," 中国科学：信息科学 (***SCIS, CCF-A, Impact Factor=7.6***), 2026.
2. **Xinkai Wang**, Chao Li, Yiyang Li, Lingyu Sun, Cheng Xu, Xiaofeng Hou, et al., "Enabling learning-based efficiency optimizer with shadow cycles in resource-constrained autonomous embedded systems," *IEEE Transactions on Computers* (***TC, CCF-A, Impact Factor=3.8***), vol. 75, no. 3, pp. 928–941, 2025. DOI: 10.1109/TC.2025.3644184
3. Junyi Mei, Shixuan Sun, Chao Li, **Xinkai Wang**, Jing Wang, Xiaofeng Hou, et al., "Dgs: A gpu-based adaptive graph sampling framework," *ACM Transactions on Architecture and Code Optimization* (***TACO, CCF-A***), DOI: 10.1145/3817060
4. Du Liu, Lu Zhang, Yechen Xu, **Xinkai Wang**, Lingyu Sun, Yifei Pu, et al., "Power synchronization: Taming massive diversified serverless functions under power constraints," *Science China Information Sciences* (***SCIS, CCF-A***), vol. 68, no. 132101, 2025. DOI: 10.1007/s11432-022-3882-y
5. Lingyu Sun, Xiaofeng Hou, Chao Li, Jiacheng Liu, **Xinkai Wang**, Quan Chen, et al., "A2: Towards accelerator level parallelism for autonomous micromobility systems," *ACM Transactions on Architecture and Code Optimization* (***TACO, CCF-A***), vol. 21, no. 4, 2024. DOI: 10.1145/3688611

发明专利

1. 李超, 张路, 许烨辰, **王鑫凯**, 冷静文, 陈全, et al., "服务器无感知计算平台能效优化方法和系统," 中国专利，CN115129475B，已授权, May 12, 2026.
2. 李超, 杨涵章, 王靖, **王鑫凯**, and 过敏意, "云内存池中服务感知的多队列节点内存管理系统及方法," 中国专利，CN119728775B，已授权, Dec. 26, 2025.
3. 李超, 孙灵雨, **王鑫凯**, and 过敏意, "多神经网络执行效率动态优化方法及系统," 中国专利，CN116842994B，已授权, Mar. 1, 2024.
4. 李超, **王鑫凯**, 孙灵雨, and 吕奇正, "基于冗余算力的异构嵌入式设备电力分配系统及方法," 中国专利，CN116414556B，已授权, Jan. 30, 2024.
5. 李超, **王鑫凯**, 张路, 陈哲轩, 陈全, and 过敏意, "面向微服务多维扰动特征的数据中心请求调度系统及方法," 中国专利，CN114205419B，已授权, Apr. 18, 2023.
6. 李超, 张路, 冯伟琪, 于峥, **王鑫凯**, and 过敏意, "基于函数中间表达的无服务器计算调度系统及方法," 中国专利，CN113238853B，已授权, Nov. 12, 2021.

荣誉与奖项

论文奖项
- **最佳学生论文奖**，IFIP NPC 2025，（*2/223*） 2025.11
- **最佳论文实现奖**，APPT 2025，（*2/74*） 2025.07
- **最佳论文提名奖**，ICCD 2024，（*4/277*） 2024.11

学术奖励
- **最佳博士生报告奖**，第二届 CCF 分布式计算大会暨中国算力网大会 2025.07
- **研究生组冠军**，CCF 体系结构学生科研挑战赛（2025 CCF TCArch SRC） 2025.07
- **Travel Grants**：ISCA 2025，ASPLOS 2025，SoCC 2023，IPDPS 2023

其他奖励
- **优秀学生干部**，上海交通大学 2022.10
- **优秀团干部**，上海交通大学 2022.05
- **优秀本科生卓越奖学金**，85 届计算机系教育发展基金暨杨元庆教育基金 2021.05
- **致远荣誉奖学金**，上海交通大学，连续 4 年

项目经历

国家级科研项目

- **协同感知的开源高性能通用处理器敏捷设计** (No.U23A6007) 2024-2027
  参与，国自然区域创新发展联合基金集成项目
  课题名称："多维特征感知的处理器性能协同优化"
- **面向微小型数据中心的系统软件** (No. 2022YFB4501702) 2023-2026
  参与，国家重点研发计划
  课题名称："跨域资源的异构计算融合"
- **可扩展数据中心资源管理** (No. 62122053) 2022-2024
  （已结题）参与，国自然优秀青年科学基金项目

企业合作项目

- **面向代理式人工智能的基础设施建模与优化** 2026-2027
  技术负责人，阿里云智能集团基础设施事业部
- **面向 AI 基础设施可靠性的高效硬件故障容错优化** 2026-2027
  技术负责人，阿里巴巴集团技术风险与效能部
  相关技术以第一作者投稿 ASPLOS 2027
- **面向 CPU 设计与选型指导的数据中心混部业务性能表征建模** 2025-2026
  技术负责人，阿里云智能集团基础设施事业部
  相关技术以共同第一作者投稿 ASPLOS 2027
- **面向大规模 AI 基础设施的软硬结合关键系统技术研究** 2024-2025
  技术负责人，阿里巴巴集团技术风险与效能部
  相关技术发表于 HPCA 2026 [1] 并应用于大规模数据中心
- **云原生基础设施性能瓶颈诊断框架** 2023-2024
  技术负责人，阿里巴巴集团技术风险与效能部
  相关技术发表于 ASPLOS 2025 [2] 并应用于大规模数据中心

学术活动

审稿服务

- Artifact Evaluation 委员会：ISCA 2026, HPCA 2026, ASPLOS 2025-2026, OSDI 2025, ATC 2025
- 影子评审委员会：EuroSys 2024
- IEEE Transactions on Reliability（TR）
- IEEE Transactions on Sustainable Computing（TSUSC）

学术服务

- 会议志愿者：APPT 2025-2026，ASPLOS 2025，CCF 芯片大会 2024，IPDPS 2023 等
- CCCF 译文服务（2025 1(3)，2022 18(12)）
- IEEE C/AISC 标准 P3584 编制助理 2026
- 上海交通大学计算机系本科生辅导员 2021-2025

助教经历

- 助教：E034140-并行计算系统 @ 2025 年秋季学期，根特大学
- 助教：CS3328-云计算技术 @ 2021-2022 年秋季学期，上海交通大学
- 助教：CS359-计算机体系结构 @ 2019 年秋季学期，上海交通大学

专业团体

- 学生会员，中国计算机学会（CCF）
- 学生会员，国际计算机学会（ACM）
- 学生会员，国际电气和电子工程师协会（IEEE）

学术报告

- Unleashing the Efficiency Potential of Shared Processors with Accelerator Units for LLM Serving
  - ChinaSys 2026，西安 2026
  - 会议口头报告，HPCA 2026，澳大利亚悉尼 2026
  - APPT 2025 学生科研挑战赛，线上 2025
- Enabling Extremely Efficient Intra-Service Tracing Observability in Datacenters
  - CCF SYS 2025，北京 2025
  - 会议口头报告，ASPLOS 2025，荷兰鹿特丹 2025
- Continuous Energy Efficiency Optimization for Autonomous Embedded Systems
  - CCF Computility 2025，兰州 2025
  - 华为智能汽车解决方案 BU 研讨会，上海 2024
  - 会议口头报告，ICCD 2024，意大利米兰 2024
- Exploiting Fragmented Shadow Resources in Shared-State Scheduler Architecture
  - 会议口头报告，SoCC 2023，美国圣克鲁兹 2023
- Exploring Efficient Microservice Level Parallelism
  - 会议口头报告，IPDPS 2022，线上 2022

指导合作

本科生

- 陈梓祺（2025 级本科生，普渡大学，大模型推理静默数据损坏优化）
- 厉溢洋（2022 级本科生，毕业去向：香港中文大学硕士，发表 TC [15]）
- 吕奇正（2020 级本科生，毕业去向：上海交通大学硕士，发表 ICCD 2024 [4]）

指导合作 (continued)

硕士生
- 杜启嵘（2026 级硕士生，代理式人工智能基础设施的 CPU 瓶颈优化）
- 李昀蔚（2024 级硕士生，高效 GPU 故障容错）
- 颛孙一鸣（2023 级硕士生，毕业去向：九坤投资，发表 APPT 2025 [3]，HPCA 2026 [1]）
- 李元成（2021 级硕士生，毕业去向：阿里巴巴，发表 ASPLOS 2025 [2]，SoCC 2023 [5]）

博士生
- 黄天浩（2025 级博士生，高效云端性能建模）
- 孙灵雨（2024 级博士生，发表 TACO 2024 [18]，RTSS 2024 [11]）