Xinkai Wang

unbreakablewxk@sjtu.edu.cn | +86-152-0196-7357

wang-xinkai.github.io | Linkedin | Google Scholar

Department of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dongchuan Road, Minhang District, Shanghai, China 200240

Research Interests

I have broad interests in computer architecture and systems. My research focuses on more efficient and resilient system architecture design. My current focused topics include:

[1] **Large-scale Datacenter Optimization**: How to enhance the resource visibility of shared-state schedulers **[C.3]**, request visibility of microservices **[C.4]**, LLM serving efficiency on scalable CPU **[S.1]** in shared datacenters?

[2] **Efficient Management Facilities**: How to eliminate the additional costs of intra-service tracing facility for cloud **[C.1]** and power management facility for edge **[C.2]** towards an efficient middleware?

[3] **Resilient Architecture Design**: How to design low-cost hardware fault tolerance architecture for complex LLM training and serving workloads in future AI infrastructure? **[Ongoing]**.

Education

- **Shanghai Jiao Tong University** — *Sep. 2021 - Present*
  4th Year, Ph.D. Student, Computer Science and Technology — Shanghai, China
  ◦ Supervisor: Prof. Chao Li; Laboratory: SAIL Lab at EPCC Center
- **Shanghai Jiao Tong University** — *Sep. 2017 - June 2021*
  Bachelor of Engineering, Computer Science and Technology — Shanghai, China
  ◦ With Zhiyuan Honors Program of Engineering

Publications

C=Conference, J=Journal, S=In Submission

[S.1] ***Xinkai Wang***, Chao Li, et.al. (2025). "Optimizing CPU for LLM serving". *2025 USENIX Annual Technical Conference* (ATC 2025, In-Submission)

[C.1] ***Xinkai Wang***, Xiaofeng Hou, Chao Li, Yuancheng Li, Du Liu, Guoyao Xu, Guodong Yang, Liping Zhang, Yuemin Wu, Xiaopeng Yuan, Quan Chen, Minyi Guo. (2025). "EXIST: Enabling Extremely Efficient Intra-Service Tracing Observability in Datacenters". *ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (ASPLOS 2025)

[C.2] ***Xinkai Wang***, Chao Li, Lingyu Sun, Qizheng Lv, Xiaofeng Hou, Jingwen Leng, and Minyi Guo. (2024). "Continuous Energy Efficiency Optimization for Autonomous Embedded Systems". *International Conference on Computer Design* (ICCD 2024, Best Paper Candidate)

[C.3] ***Xinkai Wang***, Yuancheng Li, Hao He, Chao Li, Xiaofeng Hou, Jing Wang, Quan Chen, Jingwen Leng, Minyi Guo, and Leibo Wang. (2023). "Not All Resources are Visible: Exploiting Fragmented Shadow Resources in Shared-State Scheduler Architecture". *ACM Symposium on Cloud Computing* (SoCC 2023)

[C.4] ***Xinkai Wang***, Chao Li, Lu Zhang, Xiaofeng Hou, Quan Chen, and Minyi Guo. (2022). "Exploring Efficient Microservice Level Parallelism". *International Parallel and Distributed Processing Symposium* (IPDPS 2022)

[C.5] Lingyu Sun, Chao Li, Xiaofeng Hou, Tianhao Huang, Cheng Xu, ***Xinkai Wang***, Guangjun Bao, Bingchuan Sun, Shibo Rui, and Minyi Guo. (2024). "JigSaw: Taming BEV-centric Perception on Dual-SoC for Autonomous Driving". *IEEE Real-Time Systems Symposium* (RTSS 2024)

[C.6] Du Liu, Jing Wang, ***Xinkai Wang***, Chao Li, Lu Zhang, Xiaofeng Hou, Xiaoxiang Shi, and Minyi Guo. (2024). "Improving the Efficiency of Serverless Computing via Core-Level Power Management". *IEEE/ACM international Symposium on Cluster, Cloud and Internet Computing* (CCGRID 2024)

[C.7] Lu Zhang, Chao Li, ***Xinkai Wang***, Weiqi Feng, Zheng Yu, and Minyi Guo. (2023). "FIRST: Exploiting the Multi-Dimensional Attributes of Functions for Power-Aware Serverless Computing". *International Parallel and Distributed Processing Symposium* (IPDPS 2023)

[J.1] Lingyu Sun, Xiaofeng Hou, Chao Li, Jiacheng Liu, ***Xinkai Wang***, Quan Chen, and Minyi Guo. (2024). A2: Towards Accelerator Level Parallelism for Autonomous Micromobility Systems. *Transactions on Architecture and Code Optimization* (TACO 2024)

[J.2] Du Liu, Lu Zhang, Yechen Xu, ***Xinkai Wang***, Lingyu Sun, Yifei Pu, Xiaofeng Hou, Chao Li, and Minyi Guo. (2023). Power Synchronization: Taming Massive Diversified Serverless Functions under Power Constraints. *SCIENCE CHINA Information Sciences* (SCIS 2023)

Patents

[P.1] Chao Li, Lingyu Sun, Xinkai Wang, Minyi Guo. (2024). Dynamic Efficiency Optimizer for Multiple Neural Networks. *Chinese patent granted* (CN 116842994 B)

[P.2] Chao Li, Xinkai Wang, Lingyu Sun, Qizheng Lyu. (2024). Idle resource-based intelligent power allocation system. *Chinese patent granted* (CN 116414556 B)

[P.3] Chao Li, Xinkai Wang, Lu Zhang, Zhexuan Chen, Quan Chen, Minyi Guo. (2023). Request scheduler for multi-dimensional dynamic microservice-based applications. *Chinese patent granted* (CN 114205419 B)

[P.4] Chao Li, Lu Zhang, Weiqi Feng, Zheng Yu, Xinkai Wang, Minyi Guo. (2021). Power management for serverless functions based on intermediate representation. *Chinese patent granted* (CN 113238853 B)

Honors and Awards

- **ICCD 2024 Best Paper Nomination** *(4/102)* — *Nov. 2024*
- **SoCC 2023 Travel Grant** *($1750)* — *Nov. 2023*
- **IPDPS 2023 Travel Grant** *($500)* — *May 2023*
- **Outstanding Graduate of SJTU** *(Top 15% in Graduates)* — *May 2021*
- **Excellent Undergraduate Scholarship of Yang Yuanqing Education Fund** *(3 in 400+)* — *May 2021*

Industry Experience

- **Project Leader** — *Jul. 2024 - Jan. 2025*
  Technology, Risk, and Efficiency Group, Alibaba, Hangzhou — Alibaba Innovation Program
  ◦ I optimized LLM serving using AMX on scalable CPU in shared datacenters.
  ◦ Technology is submitted to ATC 2025.
- **Project Leader & Research Intern** — *Feb. 2023 - Aug. 2023*
  Technology, Risk, and Efficiency Group, Alibaba, Hangzhou — Alibaba Innovation Program
  ◦ I optimized efficient intra-service tracing facility in large-scale cluster.
  ◦ Technology is published on ASPLOS 2025 and adopted in large-scale clusters.
- **Research Intern** — *Feb. 2022 - May 2022*
  Microsoft Research Asia, Beijing — Advised by Jie Zhang
  ◦ I worked on power-aware virtual machine scheduling and migration in Azure cloud.

Teaching Experience

- **Teaching Assistant of Cloud Computing Technology (Undergraduate)** — *Fall 2021-2022*
  Schedule the course project on Huawei Cloud to play with Kubernetes.
- **Teaching Assistant of Computer Architecture (Undergraduate)** — *Fall 2019*
  Schedule the course homework and finish grading.

Talks

- **Continuous Energy Efficiency Optimization for Autonomous Embedded Systems**
  ◦ Intelligent Automotive Solution BU Seminar, Huawei, Shanghai — 2024
  ◦ Conference Talk, ICCD 2024, Milan, Italy — 2024
- **Exploiting Fragmented Shadow Resources in Shared-State Scheduler Architecture**
  ◦ Conference Talk, SoCC 2023, Santa Cruz, USA — 2023
- **Exploring Efficient Microservice Level Parallelism**
  ◦ Conference Talk, IPDPS 2022, Virtual — 2022

Services

- **Review Experiences**
  ◦ Artifact Evaluation Committee of ASPLOS 2025
  ◦ Shadow PC member of EuroSys 2024
  ◦ Reviewer of IEEE Transactions on Sustainable Computing (TSUSC)
- **Professional Affiliations**
  ◦ Student member of IEEE, ACM, and CCF.
- **Volunteer Experiences**
  ◦ Volunteer at CCF CHIPS 2024, IPDPS 2023, etc.
  ◦ Tutors of the undergraduate of CSE department.

Skills

- **Programming Skills:** C/C++, Python, Go, Matlab
- **Software and Frameworks:** Latex, Docker, Kubernetes
- **Extracurriculars:** Badminton, Tennis, Basketball
- **Languages:** Chinese (native), English (fluent)