Zhipeng Jia

(512) 645-6487 | zhipeng.jia@outlook.com
https://www.linkedin.com/in/zhipengjia

INTRODUCTION

I am a Senior Systems Research Engineer at Google. I am part of the AI & Systems Research team under Google Cloud.

At Google, I work on innovating GenAI systems on TPUs. Some examples include (1) Generic disaggregated prefill / decode serving solution; (2) Efficient LLM serving on low-profile hardware with activation sparsity; (3) Improving Tensor Parallelism with novel distributed matmul algorithm.

Before joining Google, I obtained my Ph.D. in Computer Science from The University of Texas at Austin. My Ph.D. work focuses on fault-tolerant cloud systems. Before the Ph.D., I obtained my bachelor's degree from Tsinghua University. I participated in competitive programming during undergraduate years, when my team won ACM-ICPC Gold Medals four times.

I have extensive experience in low-level C++ programming, and Jax / Python programming.

EDUCATION

The University of Texas at Austin, Austin TX - *Department of Computer Science*

M.S. in Computer Science, Aug 2017 — Dec 2021
Ph.D. in Computer Science, Aug 2017 — May 2022

Dissertation: Designing Systems for Emerging Serverless Applications
Committee: Emmett Witchel (supervisor), Christopher J. Rossbach, Simon Peter, Jason Flinn, and Mahesh Balakrishnan

Tsinghua University, Beijing - *Institute for Interdisciplinary Information Sciences (IIIS)*

B.Eng. in Computer Science and Technology (Yao's Class), Aug 2013 — Jun 2017

FULL-TIME EXPERIENCE

Google, Seattle WA - *Senior Systems Research Engineer*

Jul 2022 - Apr 2024: Systems Research Engineer (L4)
May 2024 - now: Senior Systems Research Engineer (L5)

- Part of Google Cloud's AI & Systems Research team, which supports innovations for Google's AI infrastructure (e.g. LLM serving stack, cluster scheduler, and TPU hardware)
- As part of the team, I work on innovations of GenAI systems
- At Google, I report to Henry M. Levy, former Head of the Allen School for Computer Science & Engineering at The University of Washington

PUBLICATIONS (Google Scholar)

Impeller: Stream Processing on Shared Logs

*Zhiting Zhu, **Zhipeng Jia**, Newton Ni, Dixing Tang, Emmett Witchel*

The 20th ACM European Conference on Computer Systems (EuroSys 2025), 2025

Boki: Towards Data Consistency and Fault Tolerance with Shared Logs in Stateful Serverless Computing

***Zhipeng Jia**, Emmett Witchel*

ACM Transactions on Computer Systems, Volume 42, Issue 3-4, 2024

The Key Ideas Behind Boki's Shared Logs

***Zhipeng Jia**, Emmett Witchel*

ACM SIGOPS Operating Systems Review, Volume 58, Issue 1, 2024

Disaggregated GPU Acceleration for Serverless Applications

*Henrique Fingler, Zhiting Zhu, Esther Yoon, **Zhipeng Jia**, Emmett Witchel, Christopher J. Rossbach*

ACM SIGOPS Operating Systems Review, Volume 57, Issue 1, 2023

DGSF: Disaggregated GPUs for Serverless Functions

*Henrique Fingler, Zhiting Zhu, Esther Yoon, **Zhipeng Jia**, Emmett Witchel, Christopher J. Rossbach*

The 36th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2022), 2022

Boki: Stateful Serverless Computing with Shared Logs

***Zhipeng Jia**, Emmett Witchel*

The 28th ACM Symposium on Operating Systems Principles (SOSP ‘21), 2021

Nightcore: Efficient and Scalable Serverless Computing for Latency-Sensitive, Interactive Microservices

***Zhipeng Jia**, Emmett Witchel*

The 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS ‘21), 2021

Telekine: Secure Computing with Cloud GPUs

*Tyler Hunt, **Zhipeng Jia**, Vance Miller, Ariel Szekely, Yige Hu, Christopher J. Rossbach, Emmett Witchel*

The 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI ‘20), 2020

Isolation and Beyond: Challenges for System Security

*Tyler Hunt, **Zhipeng Jia**, Vance Miller, Christopher J. Rossbach, Emmett Witchel*

The 17th Workshop on Hot Topics in Operating Systems (HotOS XVII), 2019

Constrained Deep Weak Supervision for Histopathology Image Segmentation

***Zhipeng Jia**, Xingyi Huang, Eric I-Chao Chang, Yan Xu*

IEEE Transactions on Medical Imaging, 2017

Large Scale Tissue Histopathology Image Classification, Segmentation, and Visualization via Deep Convolutional Activation Features

*Yan Xu, **Zhipeng Jia**, Liang-Bo Wang, Yuqing Ai, Fang Zhang, Maode Lai, Eric I-Chao Chang*

BMC Bioinformatics, 2017

Efficient Near-optimal Algorithms for Barter Exchange

***Zhipeng Jia**, Pingzhong Tang, Ruosong Wang, Hanrui Zhang*

16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-17), 2017

Deep Convolutional Activation Features for Large Scale Brain Tumor Histopathology Image Classification and Segmentation

*Yan Xu, **Zhipeng Jia**, Yuqing Ai, Fang Zhang, Maode Lai, Eric I-Chao Chang*

40th International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015

HONORS & AWARDS

2017 - 2021	Provost's Graduate Excellence Fellowship, The University of Texas at Austin
2013 - 2017	Outstanding Freshman Scholarship (second prize), Tsinghua University
2017	Star of Tomorrow Internship Award, Microsoft Research Asia
2015	Gold Medal in the 2015 ACM-ICPC Asia EC-Final Contest (ranking 7th out of 267)
2014	Gold Medal in the 2014 ACM-ICPC Asia Shanghai Regional Contest (ranking 3rd out of 132)
2014	Gold Medal in the 2014 ACM-ICPC Asia MuDanjiang Regional Contest (ranking 2nd out of 146)
2013	Gold Medal in the 2013 ACM-ICPC Asia Changsha Regional Contest (ranking 4th out of 182)
2012	Gold Medal in the 2012 National Olympiad in Informatics (ranking 3rd out of 292)
2011	Gold Medal in the 2011 National Olympiad in Informatics (ranking 6th out of 294)
2010	Gold Medal in the 2010 National Olympiad in Informatics (youngest gold medalist)
2009	Gold Medal in the 2009 National Olympiad in Informatics (youngest gold medalist)

INTERNSHIP EXPERIENCE

Katana Graph, Austin TX - *Software Engineer Intern*

May 2021 - Aug 2021

- Katana Graph is an Austin-based startup focusing on high performance graph processing and analytics, founded by UT professors Keshav Pingali and Christopher J. Rossbach
- Worked on transaction support for large-scale graph updates

Google, Sunnyvale CA - *Research Intern*

May 2019 - Aug 2019

- Worked with Platform team
- Worked on the project of understanding RPC latency in Plaque

Google, Mountain View CA - *Software Engineering Intern*

May 2018 - Aug 2018

- Worked with Google News team
- Launched new machine learning-based news labeling system

Microsoft Research Asia, Beijing - *Research Intern*

Mar 2016 - Jun 2017

- Worked with Technology Strategy group under the mentorship of Dr. Eric Chang
- Involved in the project of sleep analysis with Microsoft Band
- Involved in the project of automatic analysis of large-scale medical images
- Awarded Star of Tomorrow Internship Award

Google, Mountain View CA - *Software Engineering Intern*

Jul 2015 - Sep 2015

- Worked with Machine Perception team under the supervision of Dr. Hui Fang
- Designed and implemented a deep-learning-based image enhancement framework

Microsoft Research Asia, Beijing - *Research Intern*

Feb 2014 - Mar 2015

- Worked with Technology Strategy group under the mentorship of Dr. Eric Chang
- Involved in the project of automatic analysis of large-scale medical images
- Involved in the project of Chinese OCR specialized for recognition of subtitles