Junyi(Kyle) Shu

404 Westwood Plaza, Engineering VI, Room 496 – Los Angeles, CA 90095

kyleshu.github.io shujunyi@gmail.com kyleshu (1)-213-264-4111

EXPERIENCE

University of California at Los Angeles **Los Angeles, CA, USA**
Postdoctoral Researcher *Oct 2025 – Ongoing*

- Working with Prof. Harry Xu
- Building efficient systems for large models and agentic AI.

Yanyuan Internet of Data **Beijing, China**
Co-founder and VP *Jul 2024 - Feb 2025*

- Yanyuan Internet of Data is a startup established with technology equity from Peking University, dedicated to building an infrastructure for the interconnection of data.
- Drove and closed the seed round financing, alongside the technology-for-equity transaction.

Alibaba Cloud **Beijing, China**
Visiting Researcher *May 2023 – Jul 2024*

- Designed a production system that supports burstable storage I/Os while limiting tenant interference.

Fusion Galaxy **Beijing, China**
Co-founder and COO *Oct 2020 - Dec 2022*

- Fusion Galaxy is a startup founded by professors and students of Peking University that develops blockchain technology and facilitates its application.
- Managed the company's finance and operations functions, while leading presale and project delivery.

Huaxuan Topsoft **Beijing, China**
VP of Engineering *Jan 2019 - Sep 2020*

- Huaxuan Topsoft is a startup specializing in IT solutions for the financial services industry.
- Led a 20-person team to develop and deliver an SME financing platform.

BestPay **Beijing, China**
Investor Relations Manager *Apr 2017 – Jan 2019*

- BestPay is the third-largest payment platform in China, following Alipay and WeChat Payment.
- Drove and closed BestPay's Series A financing (RMB 1 billion).

Amazon Web Services **Seattle, WA, USA**
Software Development Engineer *Feb 2014 – Aug 2015*

- Designed and built a high-performance, scalable, transactional distributed database with fellow team members.

EDUCATION

Peking University **Beijing, China**
Ph.D. in Computer Software and Theory *Sep 2021 - Jun 2025*

- Advised by Prof. Xin Jin and Prof. Xuanzhe Liu
- Dissertation: Performance Optimization towards Disaggregated Cloud Storage Systems (Outstanding Dissertation Award of PKU CS)

Peking University **Beijing, China**
Master of Engineering Management *Sep 2019 - Jul 2021*

- Advised by Prof. Xiangqun Chen
- Received the Outstanding Graduation Thesis Award.

University of California at Berkeley **Berkeley, CA, USA**
B.A. in Applied Mathematics and Computer Science *Aug 2010 - Dec 2013*

- Graduated with distinction, and was on Dean's List for several times.
- Member of Upsilon Pi Epsilon (UPE), honor society for the computing and information disciplines.

PUBLICATIONS

[*] = co-first author, [#] = corresponding author

Shan Yu, Yifan Qiao, Mingyuan Ma, Yangmin Li, Shuo Yang, Xinyuan Tong, Yang Wang, Zhiqiang Xie, Yuwei An, Shiyi Cao, Ke Bao, Deepak Vij, Xiaoning Ding, Yichen Wang, Qingda Lu, Zhong Wang, Gao Gao, Harry Xu, ***Junyi Shu***[#], Jiarong Xing[#], and Ying Sheng[#]. Prism: Cost-Efficient Multi-LLM Serving via GPU Memory Ballooning. In *20th USENIX Symposium on Operating Systems Design and Implementation (OSDI 26)*, Seattle, WA, July 2026. USENIX Association. To appear.

Junyi Shu, Xiaolong Huang, Gang Huang, Hong Mei, Xuanzhe Liu, and Xin Jin[#]. Serverless Replication of Object Storage across Multi-Vendor Clouds and Regions. In *Proceedings of the 21st European Conference on Computer Systems*, EUROSYS '26, page 1780–1796, New York, NY, USA, 2026. Association for Computing Machinery.

Junyi Shu, Kun Qian, Ennan Zhai, Xuanzhe Liu, and Xin Jin[#]. Burstable Cloud Block Storage with Data Processing Units. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 783–799, Santa Clara, CA, July 2024. USENIX Association.

Junyi Shu, Ruidong Zhu, Yun Ma, Gang Huang[#], Hong Mei, Xuanzhe Liu, and Xin Jin[#]. Disaggregated RAID Storage in Modern Datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS 2023, page 147–163, New York, NY, USA, 2023. Association for Computing Machinery.

Junyi Shu, Xin Jin, Yun Ma, Xuanzhe Liu, and Gang Huang. Cost-effective data analytics across multiple cloud regions. In *Proceedings of the SIGCOMM '21 Poster and Demo Sessions*, SIGCOMM '21, page 1–3, New York, NY, USA, 2021. Association for Computing Machinery.

SERVICES

- **Conference Program Committee**: ACM International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS (2026); The IEEE/ACM International Symposium on Microarchitecture - MICRO (2026)
- **Journal Reviewer**: IEEE Transactions on Parallel and Distributed Systems (2026); ACM Transactions on Storage (2025); ACM Transactions on Architecture and Code Optimization (2025, 2026)
- **Shadow Program Committee**: ACM European Conference on Computer Systems - EuroSys (2026)