

Yanbo Zhou

Ph.D. Candidate
Non-Volatile Systems Lab (NVSL)
University of California, San Diego

Email: yaz093@ucsd.edu
Phone: +1 858-333-0347
<https://yanbozyb.github.io/>

About Me

I have extensive experience across both industry and academia. My work spans memory and storage systems for cloud & AI infrastructure, with a focus on resource efficiency, performance, and reliability at large-scale systems.

Education

Sep. 2023 - Present	University of California, San Diego Ph.D. in Computer Science Advisor: Prof. Steven Swanson	San Diego, USA
Sep. 2016 - Jan. 2019	East China Normal University M.Eng. in Software Engineering	Shanghai, China
Sep. 2012 - Jun. 2016	Taiyuan University of Technology B.Eng. in Software Engineering	Taiyuan, China

Experience

Jun. 2024 - Present	Samsung Semiconductor <i>Research Intern, Global Open-ecoSystem Team (GOST)</i> Research intern at Samsung GOST during the summers of 2024, 2025, and 2026. I led following research collaboration projects between UCSD Non-Volatile Systems Lab and Samsung GOST. • Memory Pooling for LLM KV Cache: research on CXL-based pooling for multi-tenant LLM KV Cache sharing. • Software-defined Elastic Memory: a software-defined elastic memory system for virtual machines, which achieves flexible allocation of memory capacity and bandwidth and performance isolation (published at OSDI'26). • Energy-efficient Storage System: a fast and sustainable storage system for all-flash servers with a CPU scheduler for storage software stack, which enables high performance and energy efficiency (published at SOSP'25).	San Jose, USA
Jan. 2019 - Jun. 2023	Alibaba Cloud <i>Senior Software Engineer, Elastic Block Storage (EBS) Team</i> Worked on block storage products, including both local disks and EBS, as listed below. The local disk products I led was awarded Best Paper at USENIX FAST'26 . • I2/I2ne Instance: the second-gen high-performance and low-latency local disk. This is a software-optimized architecture towards NVMe TLC SSDs [product]. • D3c Instance: the third-gen cost-effective and large-capacity local disk. This is the first time enabling high-density QLC SSDs and Zoned Namespace (ZNS) technology in Alibaba Cloud [product]. • I4 Instance: the forth-gen high-performance and low-latency local disk. This is a software/hardware co-designed architecture (an ASIC/SoC co-designed DPU) optimized for NVMe TLC SSDs and bare-metal [product]. • ESSD-PLX: the fastest elastic block storage volume of Alibaba Cloud based on Optane Persistent Memory cache and RDMA technology [product].	Hangzhou, China

Jul. 2017 - **Intel Corporation** Shanghai, China
 Dec. 2018 *Software Engineer Intern, Storage Performance Development Kit (SPDK) Team*
 Worked on SPDK project and contributed to the following project:
 • **Vhost-NVMe**: First full-stack userspace NVMe virtualization solution. This technology is used as standard storage virtualization protocol for multiple storage products in Alibaba Cloud EBS afterwards, e.g., EBS ESSD, Local Storage Instance D3c [\[paper\]](#).

Publications

Book Chapters

2019 **Linux Open Source Storage: from Ceph to Container (Chinese Edition)**
 In the Publishing House of Electronics Industry [\[book\]](#)

Conferences

- 2026 **SmoothAgent: Efficient Long-Horizon LLM-Based Agent Serving with Lookahead Context Engineering**
 Zaifeng Pan, Qianxu Wang, Zhengding Hu, Chang Chen, Yue Guan, **Yanbo Zhou**, Steven Swanson, and Yufei Ding.
 arXiv preprint arXiv:2607.00151 [\[paper\]](#)
- 2026 **Break on Through to the Other Side: Pooling Memory Elastically with RamRyder**
Yanbo Zhou, Erci Xu, Dongjoo Seo, Adam Manzanares, Steven Swanson.
 20th USENIX Symposium on Operating Systems Design and Implementation (OSDI'26) [\[paper\]](#) [\[project\]](#)
- 2026 **Here, There and Everywhere: The Past, the Present and the Future of Local Storage in Cloud**
 Leping Yang, **Yanbo Zhou**, Gong Zeng, Li Zhang, Saisai Zhang, Ruilin Wu, Chaoyang Sun, Shiyi Luo, Wenrui Li, Keqiang Niu, Xiaolu Zhang, Junping Wu, Jiaji Zhu, Jiesheng Wu, Mariusz Barczak, Wayne Gao, Ruiming Lu, Erci Xu, and Guangtao Xue
 24th USENIX Conference on File and Storage Technologies (FAST'26) [\[paper\]](#)
Best Paper Award
- 2025 **Sleeping with One Eye Open: Fast, Sustainable Storage with Sandman**
Yanbo Zhou, Erci Xu, Anisa Su, Jim Harris, Adam Manzanares, Steven Swanson.
 31st ACM Symposium on Operating Systems Principles (SOSP'25) [\[paper\]](#)
- 2024 **CSAL: the Next-Gen Local Disks for the Cloud**
Yanbo Zhou, Erci Xu, Li Zhang, Kapil Karkra, Mariusz Barczak, Wayne Gao, Wojciech Malikowski, Mateusz Kozłowski, Łukasz Łasek, Ruiming Lu, Feng Yang, Lilong Huang, Xiaolu Zhang, Wenrui Li, Jinhu Li, Keqiang Niu, Jiaji Zhu, Jiesheng Wu.
 19th ACM European Conference on Computer Systems (EuroSys'24) [\[paper\]](#) [\[project\]](#)
- 2020 **Spool: Reliable Virtualized NVMe Storage Pool in Public Cloud Infrastructure.**
 Shuai Xue, Shang Zhao, Quan Chen, Gang Deng, Zheng Liu, Jie Zhang, Zhuo Song, Tao Ma, Yong Yang, **Yanbo Zhou**, Keqiang Niu, Sijie Sun, Minyi Guo.
 2020 USENIX Annual Technical Conference (ATC'20) [\[paper\]](#)
- 2018 **Write-aware Data Allocation on Heterogeneous Memory Architecture with Minimum Cost.**
Yanbo Zhou, Shouzhen Gu, Lixia Zheng, Edwin H.-M. Sha, Qingfeng Zhuge, Lin Wu.
 24th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'18) [\[paper\]](#)
- 2018 **Accelerating I/Os in virtual machines on physical NVMe SSDs via user space vhost target.**
 Ziye Yang, Changpeng Liu, **Yanbo Zhou**, Xiaodong Liu, Gang Cao
 8th IEEE International Symposium on Cloud and Service Computing (SC2'18) [\[paper\]](#) [\[project\]](#)

Tech Reports

2023 **A Media-Aware Cloud Storage Acceleration Layer (CSAL) Cache Solution with Intel Optane SSDs for Alibaba ECS Local Disk D3C Service.**
Yanbo Zhou, Li Zhang, Kapil Karkra, Wayne Gao, Chunhong Mao, Mariusz Barczak.
Intel White Paper [[paper](#)].

Talks

Jul. 2026 **Break on Through to the Other Side: Pooling Memory Elastically with RamRyder**
OSDI'26
Jun. 2026 UCLA System Group
May 2026 SRC ACE Center for Evolvable Computing [[slides](#)] [[talk](#)]

Here, There and Everywhere: The Past, the Present and the Future of Local Storage in Cloud
Apr. 2026 Meta's AI and Systems Co-Design Team
Feb. 2026 FAST'26 [[slides](#)] [[talk](#)]

Sleeping with One Eye Open: Fast, Sustainable Storage with Sandman
Dec. 2025 SRC ACE Center for Evolvable Computing
Sep. 2025 Meta's AI and Systems Co-Design Team [[slides](#)]

CSAL: the Next-Gen Local Disks for the Cloud
Apr. 2025 EuroSys'24 [[slides](#)]

Best SPDK Practices: Lessons from Five Years of Storage Evolution in Alibaba Cloud
Dec. 2022 SPDK PRC Forum [[slides](#)] [[talk - chinese](#)]

Removing QLC Write-Amplification through Intel Optane SSD with SPDK WSR
Dec. 2021 SPDK PRC Forum [[slides](#)] [[talk - chinese](#)]

Academic Service

Reviewer

2026 IEEE Computer Architecture Letters (CAL)

Artifact Evaluation Committee

2026 USENIX FAST, USENIX OSDI

2025 ACM ASPLOS, ACM EuroSys, ACM SOSP

Professional Skills

Hardware

- Memory: DRAM, Persistent Memory, CXL Memory (expert)
- Storage: CMR/SMR HDD, NVMe SSD, SLC/MLC/TLC/QLC NAND Flash Media (expert)
- Accelerator: data center DPU (expert)
- Network: RDMA NIC (some familiarity)

Software

- Storage Systems: elastic block storage, kernel bypass SPDK, storage virtualization, block storage cache (expert)
- Linux Kernel: filesystem, block layer, memory management (knowledgeable)
- Protocol: NVMe, NVMe-oF, Virtio (knowledgeable)
- Network Systems: kernel bypass DPDK (some familiarity)