

Yizhou Shan (单一舟)

Areas: Distributed Systems, OS, AI Infrastructure, LLM Serving
Web: lastweek.io

+86 13028875588
Email: syzwhat@gmail.com

Research Summary

My research aims to build fast and reliable systems at datacenter scale, grounded in the principles of **disaggregation** and **pooling**. During my PhD, I developed LegoOS, the first distributed operating system for hardware resource disaggregation (OSDI'18, Best Paper), and later extended these ideas to disaggregated **memory** and **networking** systems such as Clio (ASPLOS'22) and SuperNIC (FPGA'24).

More recently, I have focused on AI infrastructure and LLM serving systems, proposing **disaggregated Prefill-Decode and MoE-Attention** (TetriInfer, MemServe, and xDeepServe) serving architectures that improve throughput and resource efficiency at scale. At Huawei Cloud, I lead the development of an Ascend-native, serverless LLM serving system called xDeepServe (XDS) that underpins Huawei Cloud's Model-as-a-Service service. Looking ahead, I aim to define system abstractions that can keep pace with rapidly evolving AI models, workloads, and hardware in emerging AI SuperPod environments.

Education

University of California, San Diego Ph.D. in Computer Science	2019-2022
Purdue University Ph.D. in Computer Engineering (Transferred to UCSD)	2016-2019
Beijing University of Aeronautics and Astronautics (BUAA) B.E. in Computer Engineering	2010-2014

Industry Experiences

Huawei Cloud - Research Scientist	Shanghai 2022 ~ Now
Microsoft Research - Intern Mentors: Ziqiao Zhou, Weidong Cui, Andrew Baumann, Marcus Peinado	Virtual 2021
VMware Research - Intern Mentor: Marcos K. Aguilera	Palo Alto, CA 2019
VMware Research - Intern Mentor: Stanko Novakovic	Palo Alto, CA 2018
ICT, Chinese Academy of Sciences - Research Assistant Mentors: Zhiwei Xu, Jin Xiong, Dejun Jiang	Beijing, China 2014-2016

Selected Publications

- [15] **xDeepServe: Model-as-a-Service on Huawei CloudMatrix384** *arXiv'25*
XDS Team @ Huawei Cloud
My Contribution: Led project and paper.
- [14] **DeepServe: Serverless Large Language Model Serving at Scale** *ATC'25*
Junhao Hu, Jiang Xu, Zhixia Liu, Yulong He, Yuetao Chen, Hao Xu, Jiang Liu, Jie Meng, Baoquan Zhang, Shining Wan, Gengyuan Dan, Zhiyu Dong, Zhihao Ren, Changhong Liu, Tao Xie, Dayun Lin, Qin Zhang, Yue Yu, Hao Feng, Xusheng Chen, Yizhou Shan
My Contribution: Led project and paper.
- [13] **EPIC: Efficient Position-Independent Caching for Serving Large Language Models** *ICML'25*
Junhao Hu, Wenrui Huang, Haoyi Wang, Weidong Wang, Tiancheng Hu, Qin Zhang, Hao Feng, Xusheng Chen, Yizhou Shan, Tao Xie
My Contribution: Advised and co-authored
- [12] **MemServe: Context caching for disaggregated LLM serving with elastic memory pool** *arXiv'24*
Cunchen Hu, Heyang Huang, Junhao Hu, Jiang Xu, Xusheng Chen, Tao Xie, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, Yizhou Shan
My Contribution: Advised and co-authored
- [11] **TetriInfer Inference without interference: Disaggregate LLM inference for mixed downstream workloads** *arXiv'24*
Cunchen Hu, Heyang Huang, Liangliang Xu, Xusheng Chen, Jiang Xu, Shuang Chen, Hao Feng, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, Yizhou Shan
My Contribution: Advised and co-authored
- [10] **Skadi: Building a distributed runtime for data systems in disaggregated data centers** *HotOS'25*
Cunchen Hu, Chenxi Wang, Sa Wang, Ninghui Sun, Yungang Bao, Jieru Zhao, Sanidhya Kashyap, Pengfei Zuo, Xusheng Chen, Liangliang Xu, Qin Zhang, Hao Feng, Yizhou Shan
My Contribution: Advised and co-authored.
- [9] **Disaggregating and Consolidating Network Functionalities with SuperNIC** *FPGA'24*
Yizhou Shan, Will Lin, Ryan Kosta, Arvind Krishnamurthy, Yiyang Zhang
My Contribution: Led project and paper
- [8] **Core slicing: closing the gap between leaky confidential VMs and bare-metal cloud** *OSDI'23*
Ziqiao Zhou, Yizhou Shan, Weidong Cui, Xinyang Ge, Marcus Peinado, Andrew Baumann
My Contribution: Primary designer, I hacked QEMU on the driver part.
- [7] **MARB: Bridge the Semantic Gap between Operating System and Application Memory Access Behavior** *DATE'23*
Haifeng Li, Ke Liu, Ting Liang, Zuojun Li, Tianyue Lu, Hui Yuan, Yinben Xia, Yungang Bao, Mingyu Chen, Yizhou Shan
My Contribution: Primary writer.
- [6] **HoPP: Hardware-Software Co-Designed Page Prefetching for Disaggregated Memory** *HPCA'23*
Haifeng Li, Ke Liu, Ting Liang, Zuojun Li, Tianyue Lu, Hui Yuan, Yinben Xia, Yungang Bao, Mingyu Chen, Yizhou Shan
My Contribution: Primary writer.

- [5] **Clio: A Hardware-Software Co-Designed Disaggregated Memory System** [ASPLOS '22](#)
*Yizhou Shan**, Zhiyuan Guo*, Xuhao Luo, Yutong Huang, Yiyang Zhang (co-first authors)
My Contribution: Co-lead and co-authored.
- [4] **Disaggregating Persistent Memory and Controlling Them Remotely:
An Exploration of Passive Disaggregated Key-Value Stores** [ATC '20](#)
Shin-Yeh Tsai, Yizhou Shan, Yiyang Zhang
My Contribution: I'm responsible for system optimization and evaluation.
- [3] **Storm: a Fast Distributed Storage System Using Remote Memory Primitives** [SYSTOR '19](#)
Stanko Novakovic, Yizhou Shan, Aasheesh Kolli, Michael Cui, Yiyang Zhang, Haggai Eran, Liran Liss, Michael Wei, Dan Tsafir, Marcos Aguilera
My Contribution: I'm responsible for the LITE part.
- [2] **LegoOS: A Disseminated, Distributed OS for Hardware Resource Disaggregation** [OSDI'18](#)
Yizhou Shan, Yutong Huang, Yilun Chen, Yiyang Zhang
My Contribution: I'm the driver of this project.
- [1] **Distributed Shared Persistent Memory** [SoCC '17](#)
Yizhou Shan, Shin-Yeh Tsai, Yiyang Zhang
My Contribution: I'm the driver of this project.

Workshops and Posters

- [6] **Challenges in Building and Deploying Disaggregated Persistent Memory** [NVMW '19](#)
Yizhou Shan, Yutong Huang, Yiyang Zhang
- [5] **Distributed Shared Persistent Memory** [NVMW '18](#)
Yizhou Shan, Shin-Yeh Tsai, Yiyang Zhang
- [4] **Disaggregating Memory with Software-Managed Virtual Cache** [WAMS '18](#)
Yizhou Shan, Yiyang Zhang
- [3] **Disaggregated Operating System** [HPTS '17](#)
Yiyang Zhang, Yizhou Shan, Sumukh Hallymysore
- [2] **Lego: A Distributed, Decomposed OS for Resource Disaggregation** [SOSP '17](#)
Yizhou Shan, Yilun Chen, Yutong Huang, Sumukh Hallymysore, Yiyang Zhang
Poster
- [1] **Disaggregated Operating System** [SoCC '17](#)
Yizhou Shan, Sumukh Hallymysore, Yutong Huang, Yilun Chen, Yiyang Zhang
Poster

Professional Services

Program Committee

EuroSys (2023, 2024, 2025)
ATC (2023, 2024, 2025, 2026)
NSDI (2024, 2025)
SoCC (2022)

EuroSys '22 (Shadow PC)
EuroSys '21 (Shadow PC)
ASPLOS '21 (External PC)

Journal Review

Journal of Systems Research: 2021 - Current
ACM Transactions on Architecture and Code Optimization (TACO): 2021

ACM Transactions on Storage (TOS): 2020
IEEE/ACM Transactions on Networking: 2020

Artifact Evaluation

SOSP'21 (Artifact Evaluation)
OSDI '20 (Artifact Evaluation)

Teaching

TA for UCSD [CSE120](#) Undergraduate Operating System

Mentorship

- **Junhao Hu (PhD @ PKU)**: Intern at Huawei Cloud, EPIC (ICML'25), RaaS (ACL'25), DeepServe (ATC'25), xDeepServe (arXiv'25)
- **Cunchen Hu (PhD @ ICT)**: Intern at Huawei Cloud, Skadi (HotOS'25), TetriInfer (arXiv'25)
- **Will Lin (PhD @ UCSD)**: SuperNIC (FPGA'24), Vision Paper (APSys'24)

Awards

- [2020 Facebook Fellowship Finalist](#)
- SYSTOR'19 Best Paper Award
- OSDI '18 Jay Lepreau Best Paper Award
- OSDI '18 Student Travel Grant
- SOSP '17 Student Travel Grant
- SoCC '17 Student Travel Grant

Skills

Languages: x86 Assembly, C, C++, Python, Scala, Rust, Go, TCL, Verilog, Java

Systems: Linux Kernel, DPDK/RDMA, KVM, QEMU, Docker, k8s, Pytorch, Tensorflow, Spark, Memcached, Vivado, Vivado HLS, Vitis, SpinalHDL, Chisel