

Yusheng Zheng

Ph.D. Candidate, Computer Science and Engineering, UC Santa Cruz (Expected 2027)

yunwei356@gmail.com | +1-702-848-9922 | yunwei37.com | [GitHub](#) | [Google Scholar](#) | [LinkedIn](#)

OSDI '25 | ASPLOS '27 | In production on 80K+ GPUs at Alibaba | 10K+ OSS stars | 200+ citations

Experience

ByteDance Inc.

Research Scientist Intern

San Jose, CA

Jun 2026 – Aug 2026

- Built a general-purpose **eBPF profiler** for the ByteDance Seed team: dynamically configurable instrumentation with **cross-layer GPU / CPU / memory tracing** spanning userspace and kernel, including fine-grained memory visibility.
- Achieves **~1% overhead**; currently in staged production rollout.

Education

University of California, Santa Cruz

Ph.D. in Computer Science and Engineering; advisor: Andi Quinn

Santa Cruz, CA

Sep 2024 – Expected 2027

Zhejiang University

B.S. in Geographic Information Science

Hangzhou, China

Sep 2018 – Jul 2023

Selected Systems

bpftime — Userspace & GPU eBPF Runtime

[OSDI'25](#) | [ASPLOS'27](#) | [Code](#)

Creator, lead maintainer & first author — [OSDI '25 & ASPLOS '27](#); 1.5K+; **production adoption** incl. Alibaba & Huawei

- High-performance userspace/GPU eBPF runtime and dynamic binary instrumentation framework for observability, network, security, and custom logic. Inject eBPF into **running processes and GPU kernels** without code changes or restarts, up to **10×** faster than kernel uprobes with new LLVM backend. Works with existing eBPF ecosystems out of the box; no root access required.
- Extending Applications Safely and Efficiently** ([OSDI '25](#), first author): capability-based extension model and userspace runtime.
- gpu_ext: Extensible Policies for GPUs via eBPF** ([ASPLOS '27](#), first author): eBPF policies inside the NVIDIA driver for memory/scheduler and on-device DBI; up to **4.8×** higher throughput, **2×** lower tail latency.
- Invited talks at Linux Plumbers '23 & '25, KubeCon EU '25, and eBPF Summit '24.

AgentSight: Observability, Profiling & Runtime Control for AI Agents

[Paper](#) | [Code](#)

Creator, lead maintainer & first author — shipped in Alibaba Cloud Linux 4 and used in production at multiple companies

- Distributed runtime and observability stack where agents supervise other agents — security, resource usage, trajectory and token/latency-hotspot analysis — feeding fixes back into skills and harnesses for self-improving agent systems, at 2.9% overhead.
- ActPlane ([Code](#)): compiles agent policies into unbypassable eBPF information-flow control enforcement in OS kernel.
- AgentCgroup** ([paper](#), **AgenticOS @ ASPLOS '26 Best Paper**, first author): OS resource control and attribution for AI agents.
- AgentPProf** ([paper](#), **AgenticOS @ SOSP '26 Best Paper**, first author): pprof-compatible semantic profiling of long-horizon agent trajectories.

SysOM-AI: Cross-Layer Performance Diagnosis for Production AI Training

[Paper](#)

Co-first author (with Alibaba) — in production at Alibaba for 1+ year across **80K+ GPUs**

- Co-designed the cross-layer diagnosis architecture and built research prototypes combining CPU stack profiling, GPU kernel tracing, and NCCL instrumentation at less than 0.4% overhead with eBPF.
- Validated in Alibaba production across **80K+ GPUs**: 94 confirmed production issues diagnosed; median diagnosis time cut **from days to ~10 minutes**.

Open Source & Community Leadership

eunomia-bpf: Open-Source Community for eBPF & AI Systems

[GitHub](#)

Founder & lead maintainer (2022 – present) — **10K+ cumulative stars**; featured on ebpf.io and the CNCF landscape

- Flagship repos: bpf-developer-tutorial (4.2K+, bilingual eBPF course), bpftime (1.5K+), eunomia-bpf (0.9K+), AgentSight (0.6K+), wasm-bpf (0.4K+); talks at KubeCon NA '23 and Apsara Conference '22.
- Mentor / org admin for Google Summer of Code 2024 and Open Source Promotion Plan 2023–2024; FOSDEM eBPF developer-room organizer 2025–2026.

Earlier Experience

Imperial College London

Research Assistant for userspace eBPF runtime in DPDK / AF_XDP and kernel-bypass environments

Jan 2024 – Jul 2024

Remote

PLCT Lab, Institute of Software, Chinese Academy of Sciences

Sponsored Open-Source Maintainer — led a three-person eBPF team for community and upstream work

Nov 2022 – Jul 2025

Remote

Naive Systems

Static Analysis Intern — C/C++/Go checkers for MISRA and GJB standards

Nov 2021 – Oct 2022

Shanghai, China

Repomono

Backend Intern — monorepo DevOps and CI/CD tooling

Sep 2020 – Jun 2022

Shanghai, China

Service, Awards & Other Publications

- Co-organizer, **Agentic OS Workshop**: 1st @ **ASPLOS 2026**, 2nd @ **SOSP 2026**
- eBPF Foundation Community & Advocacy Fellow, 2025
- Open Source Innovation Pioneer Award, OS2ATC
- Artifact Evaluation Committee: OSDI 2025; EuroSys 2025
- First prize / team lead, Chinese National Computer System Development Capability Competition 2022; mentor and reviewer 2023–2026
- **SchedCP** (NeurIPS '25 ML for Systems **spotlight**, first author) & **Kgent** (SIGCOMM eBPF '24): LLM agents that synthesize, verify, and deploy `sched_ext` Linux schedulers; up to $1.79\times$ speedup
- **Kops** (ATC '26, co-first author, [paper](#)): safely extends the eBPF compilation pipeline with native operations; up to 24% faster microbenchmarks and 12% faster production applications
- **NCCLbpf** (sole author, [arXiv '26](#)): verified, hot-reloadable policies in NCCL plugin interfaces; up to 27% higher AllReduce throughput on $8\times$ NVIDIA B300
- **CXLMemSim** (CXL 3.0 memory simulation, HPDC 2026, co-author)

Technical Skills

Systems: Linux, eBPF/libbpf, LLVM/JIT, dynamic binary instrumentation, DPDK, Kubernetes, CUDA/PTX, NCCL.

Reverse Engineering: NVIDIA GPU driver, CUDA runtime, PTX compilers, NVBit, CUPTI, and DGX Spark firmware; unlocked undocumented and deliberately hidden features for research and performance gains.