Yusheng Zheng

加州大学圣克鲁兹分校，计算机科学与工程博士候选人（预计 2027 年毕业）

yunwei356@gmail.com | +1-702-848-9922 | yunwei37.com | GitHub | Google Scholar | LinkedIn

AI 智能体基础设施 | 训练与推理系统
大规模生产实践 | OSDI / ATC / ASPLOS | 10K+ 开源项目星标 | 350+ 引用

工作经历

字节跳动（ByteDance Inc.） **2026 年 6 月至今**
学生研究员，ByteDance Seed 6--8 月；11 月至今

- 为 ByteDance Seed 团队构建通用 **eBPF 性能剖析器**：支持动态配置的插桩，提供跨用户态与内核的**跨层 GPU / CPU / 内存追踪**，包括细粒度内存可观测能力。开销约为 **1%**，目前正分阶段部署到生产环境。
- 正在设计 AI Agent Harness 与形式化验证模型，以提升 10K-GPU 规模 RL 基础设施的可靠性。

教育背景

加州大学圣克鲁兹分校 **美国加州圣克鲁兹**
计算机科学与工程博士；导师：Andi Quinn 2024 年 9 月至预计 2027 年

浙江大学 **中国杭州**
地理信息科学学士 2018 年 9 月至 2023 年 7 月

代表性系统

bpftime：用户态与 GPU eBPF 运行时 OSDI'25 | ASPLOS'27 | 代码

创始人、主要维护者与第一作者 --- **OSDI '25**（第一作者）、**ASPLOS '27**（共同第一作者）；1.5K 星标；**生产环境采用方**包括阿里巴巴与华为

- 高性能用户态/GPU eBPF 运行时与动态二进制插桩框架，用于可观测性、网络、安全与自定义逻辑。无需修改代码或重启，即可将 eBPF 注入**运行中的进程与 GPU 内核**；采用新的 LLVM 后端，比内核 uprobes 最多快 **10 倍**。开箱即用地兼容现有 eBPF 生态，无需 root 权限。
- **安全高效地扩展应用程序**（Extending Applications Safely and Efficiently，**OSDI '25**，第一作者）：基于能力的扩展模型与用户态运行时。
- **基于 eBPF 的安全、可编程操作系统级 GPU 资源管理**（Safe and Programmable OS-Level GPU Resource Management with eBPF，**ASPLOS '27**，共同第一作者）：在 NVIDIA 驱动中实现用于内存与调度器的 eBPF 策略，以及设备端动态二进制插桩；吞吐量最高提升至 **4.8 倍**，尾延迟最高降低 **2 倍**。
- 受邀在 Linux Plumbers '23 与 '25、KubeCon EU '25、eBPF Summit '24 演讲。

AgentSight：AI 智能体可观测性、性能剖析与运行时控制 论文 | 代码

创始人、主要维护者与第一作者 --- 随 Alibaba Cloud Linux 4 发布，并被多家公司用于生产环境

- 构建由智能体监督其他智能体的分布式运行时与可观测性技术栈，分析安全、资源使用、执行轨迹以及 token/延迟热点，并将修复反馈至技能与执行框架，实现自我改进的智能体系统，开销为 2.9%。
- ActPlane（代码）：将智能体策略编译为操作系统内核中不可绕过的 eBPF 信息流控制机制。
- **AgentCgroup**（论文，**AgenticOS @ ASPLOS '26 最佳论文**，第一作者）：AI 智能体的操作系统资源控制与归因。
- **AgentPProf**（论文，**AgenticOS @ SOSP '26 最佳论文**，第一作者）：兼容 pprof 的长周期智能体执行轨迹语义性能剖析。

SysOM-AI：生产环境 AI 训练的跨层性能诊断 论文

共同第一作者（与阿里巴巴合作） --- 在阿里巴巴生产环境运行 1 年以上，覆盖 **80K+ GPU**

- 共同设计跨层诊断架构，并构建研究原型，使用 eBPF 结合 CPU 调用栈采样、GPU 内核追踪与 NCCL 插桩，开销低于 0.4%。
- 在阿里巴巴覆盖 **80K+ GPU** 的生产环境中验证：诊断 94 个已确认的生产问题；诊断时间中位数**从数天缩短至约 10 分钟**。

开源与社区领导力

eunomia-bpf：eBPF 与 AI 系统开源社区 GitHub

创始人与主要维护者（2022 年至今） --- **累计 10K+ 星标**；入选 ebpf.io 与 CNCF 全景图

- 旗舰仓库：bpf-developer-tutorial（4.2K 星标，双语 eBPF 课程）、bpftime（1.5K 星标）、eunomia-bpf（0.9K 星标）、AgentSight（0.6K 星标）、wasm-bpf（0.4K 星标）；在 KubeCon NA '23 与云栖大会 '22 演讲。
- Google Summer of Code 2024 与开源之夏 2023--2024 导师/组织管理员；FOSDEM 2025--2026 eBPF 开发者专场组织者。

早期经历

伦敦帝国理工学院 2024 年 1 月至 7 月
研究助理：DPDK / AF_XDP 与内核旁路环境中的用户态 eBPF 运行时 远程

中国科学院软件研究所 PLCT 实验室 2022 年 11 月至 2025 年 7 月
受资助开源维护者：带领三人 eBPF 团队开展社区与上游工作 远程

Naive Systems 2021 年 11 月至 2022 年 10 月
静态分析实习生：面向 MISRA 与 GJB 标准开发 C/C++/Go 检查器 中国上海

Repomono 2020 年 9 月至 2022 年 6 月
后端实习生：monorepo DevOps 与 CI/CD 工具 中国上海

学术服务、奖项与其他论文

- **Agentic OS Workshop** 联合组织者：第一届于 **ASPLOS 2026**，第二届于 **SOSP 2026**
- eBPF Foundation 社区与推广 Fellow，2025 年
- OS2ATC 开源创新先锋奖
- 工件评审委员会：OSDI 2025；EuroSys 2025
- 2022 年全国大学生计算机系统能力大赛一等奖/队长；2023--2026 年担任导师与评委
- **SchedCP**（NeurIPS '25 ML for Systems **spotlight**，第一作者）与 **Kgent**（SIGCOMM eBPF '24）：利用 LLM 智能体生成、验证并部署 `sched_ext` Linux 调度器，最高加速 1.79 倍
- **BPF-Ext：使用原生操作安全扩展 eBPF 编译流水线**（BPF-Ext: Safely Extending the eBPF Compilation Pipeline with Native Operations，**ATC '26**，共同第一作者，论文）：使用原生操作安全扩展 eBPF 编译流水线；微基准测试最高快 24%，生产应用最高快 12%
- **NCCLbpf**（独立作者，arXiv '26）：NCCL 插件接口中经过验证、可热重载的策略；在 8 块 NVIDIA B300 上使 AllReduce 吞吐量最高提升 27%
- **CXLMemSim**（CXL 3.0 内存模拟，HPDC 2026，共同作者）

技术技能

系统： Linux、eBPF/libbpf、LLVM/JIT、动态二进制插桩、DPDK、Kubernetes、CUDA/PTX、NCCL。

逆向工程： NVIDIA GPU 驱动、CUDA 运行时、PTX 编译器、NVBit、CUPTI 与 DGX Spark 固件；解锁未公开及刻意隐藏的功能，用于研究与性能提升。