谭昕

最后更新：2026/6

邮箱:re.tanxin13@gmail.com · **个人主页**:https://txxx926.github.io/

研究兴趣：机器学习系统

对面向机器学习负载的系统设计感兴趣, 特别是以下几个方面

- **分布式机器学习系统**: 大模型（LLM, Video Diffusion Model, RLHF）的分布式训练优化; 机器学习负载的推理系统设计与优化。
- **资源管理调度**: 机器学习系统的资源管理与调度; 负载调度算法设计与优化。

教育背景

香港中文大学, 计算机科学与工程, 博士研究生 2022.8 - 2026.7 (预计)

- 导师:徐宏 长聘副教授

西北工业大学, 计算机科学与技术, 工学学士 2018.9 - 2022.6

- GPA: 93.37/100, 排名:**1/247**
- **国家奖学金两次**，一等奖学金三次，西北工业大学优秀毕业生

论文发表

预印版

- Kaiwen Chen, **Xin Tan**, Jingzong Li, Hong Xu. ***Libra: Efficient Resource Management for Agentic RL Post-Training.*** [Preprint]
- Yicheng Feng, **Xin Tan**, Yangtao Deng, Yimin Jiang, Yibo Zhu, Hong Xu. ***Frontier: Towards Comprehensive and Accurate LLM Inference Simulation.*** [Preprint] [Code]

正式出版

- [**SIGCOMM'26, CCF-A**] **Xin Tan**, Yicheng Feng, Yu Zhou, Yimin Jiang, Yibo Zhu, Hong Xu. ***Dynamic Compute and Network Orchestration for Disaggregated RL.*** ACM Special Interest Group on Data Communication. [Preprint]
- [**ASPLOS'26, CCF-A**] **Xin Tan**, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, Hong Xu. ***DSV: Exploiting Dynamic Sparsity to Accelerate Large-Scale Video DiT Training.*** ACM International Conference on Architectural Support for Programming Languages and Operating Systems. [Publication] [Preprint] [Code]
- [**ASPLOS'25, CCF-A**] **Xin Tan**, Yimin Jiang, Yitao Yang, Hong Xu. ***Towards End-to-End Optimization of LLM-based Applications with Ayo.*** ACM International Conference on Architectural Support for Programming Languages and Operating Systems. [Publication] [Preprint] [Code]
- [**FAISys'25**] **Xin Tan**, Yicheng Feng, Yu Zhou, Yimin Jiang, Yibo Zhu, Hong Xu. ***RFabric: A Reconfigurable Network for the Rhythms of Disaggregated RL.*** The 1st Frontier AI Systems Workshop.
- [**ICPP'24, CCF-B**] **Xin Tan**, Jiamin Li, Yitao Yang, Jingzong Li, Hong Xu. ***Arlo: Serving Transformer-based Language Models with Dynamic Input Lengths.*** ACM International Conference on Parallel Processing. [Publication]
- [**ICML'26, CCF-A**] Kaihua Liang, **Xin Tan**, An Zhong, Hong Xu, Marco Canini. ***FOCUS: DLLMs Know How to Tame Their Compute Bound.*** International Conference on Machine Learning. [Preprint] [Code]
- [**IWQoS'26, CCF-B**] Kaiwen Chen, **Xin Tan**, Minchen Yu, Jingzong Li, Hong Xu. ***MemShare: Memory Efficient Inference for Large Reasoning Models through KV Cache Reuse.*** IEEE/ACM International Symposium on Quality of Service. [Preprint]
- [**SOSP PACMI'25**] Yicheng Feng, **Xin Tan**, Kin Hang Sew, Yimin Jiang, Yibo Zhu, Hong Xu. ***Frontier: Simulating the Next Generation of LLM Inference Systems.*** SOSP Workshop on Practical Adoption Challenges of ML for Systems. [Preprint]

实习经历

阶跃星辰 | 系统组, 研究实习生 2024.9-2026.1

- 负责大规模视频生成模型训练效率优化研究，围绕 Video DiT 训练中的动态注意力稀疏性设计并实现 DSV 框架，包括低秩近似与两阶段稀疏性预测、稀疏注意力 Kernel、稀疏感知上下文并行及负载均衡策略；相关工作发表于 ASPLOS'26，在 128 卡集群上最高训练吞吐提升 3.02 倍。
- 面向大规模 RLHF/Agentic RL 后训练集群，刻画 rollout、推理与训练组件之间的计算-网络节奏和资源瓶颈，设计动态计算与网络协同编排机制及可重构 Fabric 方案；构建 RL 集群仿真与评估框架，支持不同模型规模、并行策略、集群拓扑和负载节奏下的性能评估，相关论文录用于 SIGCOMM'26。

微软亚洲研究院 | 系统与网络研究组, 研究实习生 2021.10-2022.3

- 参与了非侵入式监控系统开发，负责设计并实现了可扩展的分析工具模块，用于处理 GPU 利用率、网络/NVLink 带宽等关键指标，支持大规模集群上机器学习工作负载的性能评估。
- 利用该系统对数据中心半年的工作负载数据进行深入分析, 刻画了不同类型 AI 任务的资源使用模式与网络需求特点, 撰写内部文档提出优化云基础设施与软件栈的建议。

项目经历

面向解耦式强化学习后训练的计算与网络协同编排——录用 SIGCOMM'26（一作/主导）
2025.5-2026.1

- 系统性分析大语言模型强化学习后训练（RLHF/Agentic RL）中的解耦式 rollout 与训练架构，发现生成阶段受动态输出长度、KV Cache 占用和长尾请求影响，容易成为端到端瓶颈；同时训练、生成和权重同步阶段具有差异化且动态变化的通信模式，静态网络难以高效匹配。
- 设计 OrchestrRL 计算与网络协同编排框架，提出面向 rollout 阶段的自适应计算调度器，根据 batch 大小、生成长度分布和剩余请求状态动态切换 TP/EP/AFD 等并行配置，并结合在线请求迁移缓解 straggler；同时设计 RFabric 可重构光电混合网络，为训练集合通信、生成阶段通信和权重同步按阶段按需分配带宽。
- 构建 RLSim 高保真模拟器，集成训练、推理和网络仿真能力，用于大规模 RL 集群的性能与成本评估；在 64 卡 NVIDIA H800 集群上验证 OrchestrRL 最高将端到端训练吞吐提升 1.42 倍，模拟结果显示 RFabric 在大规模场景下接近理想无阻塞 Fat-Tree 性能，并提升 1.53–2.06 倍性能成本效率。

面向现代大语言模型推理服务的高保真模拟器——预印版（核心参与） 2025.8-2026.5

- 分析现代大语言模型推理服务从同构单体架构转向解耦式部署后的仿真挑战，发现现有模拟器难以同时刻画 Prefill-Decode Disaggregation（PDD）、Attention-FFN Disaggregation（AFD）、复杂并行策略、运行时优化和有状态请求，导致 SLA 与性能决策失真。
- 参与设计并实现 Frontier 离散事件仿真器，基于角色化 cluster worker 建模 co-location、PDD 和 AFD 架构，将 CUDA Graph、speculative decoding、prefix caching 等运行时优化纳入 scheduler-batch-engine 闭环，并支持 reasoning、agent 和 RL rollout 等有状态 workload。
- 参与构建计算、通信和内存开销的高保真预测模块，围绕 operator runtime、collective communication、KV Cache 容量和跨 cluster 传输等关键路径补齐成本建模能力；该系统最终在 16 卡 NVIDIA H800 集群验证中平均吞吐误差低于 4%，可扩展到千卡级 GPU 仿真并支持 SLA 约束下的设计空间探索。

利用注意力稀疏性加速大规模视频生成模型的训练——发表于 ASPLOS'26（一作/主导）
2024.9-2025.2

- 系统性分析大规模视频生成模型（Video DiT）训练的效率瓶颈，发现全量注意力计算在高分辨率长视频场景下资源开销巨大，限制了模型的扩展性与训练性能；同时注意力机制中存在动态且异构的稀疏性特征，为计算加速提供了优化空间。
- 设计并实现了 DSV 加速框架，提出低秩近似与两阶段稀疏性预测算法，结合高效稀疏注意力 Kernel，实现关键 Key-Value 对的高效预测与稀疏计算；同时针对稀疏特性，拓展并优化了上下文并行策略（包括 head/sequence 维度的混合并行、负载均衡与通信优化），大幅提升多 GPU 环境下的分布式效率。
- 基于 PyTorch FSDP 和 Triton 进行工程实现，在 128 卡集群上对最高达 30B 参数级视频 DiT 模型进行实验，最高训练吞吐提升 3.02 倍，视频生成质量与全量注意力方法保持一致。

面向基于大语言模型应用的端到端优化——发表于 ASPLOS'25（一作/主导） 2024.1-2024.6

- 分析了现有大语言模型应用编排框架的局限性, 包括模块化编排导致的有限工作流端到端优化空间以及请求调度难以满足应用的端到端要求。
- 设计并实现了一个端到端优化系统 Ayo, 其核心是细粒度编排, 将涉及不同模块的工作流视作由任务原语组成的数据流图, 从而暴露出各原语的特性和交互关系, 便于实现端到端优化（包括原语并行、流

水线并行等)。细粒度的数据流图还为请求调度提供了更多信息, 包括请求的相关性和依赖关系, 帮助达成应用端目标。
- 基于 Ray 实现了该系统, 并在不同的大语言模型应用（如搜索引擎增强的生成、检索增强的生成）中验证了其有效性，最高端到端平均延迟加速 2.09 倍。

面向变长语言模型负载的部署优化——发表于 ICPP'24（一作/主导） 2022.10-2023.6

- 分析了判别式语言模型在部署过程中面对变长请求时的低效问题, 尤其是静态编译（使用最长长度的零值填充）和动态编译方法的局限性。
- 提出了一种混合编译和动态调度方案 Arlo。该方案对不同长度的请求使用相应的静态编译运行时, 并根据请求长度分布实时调整各运行时的 GPU 资源分配。同时设计了多级请求队列调度算法, 以应对请求长度分布的短期突变, 优化总体请求的延迟。
- 基于 Triton Inference Server 和 TensorRT 部署并验证了该系统, 在不同请求分布条件下显著提升了效率。

竞赛获奖

- 国际水下机器人竞赛仿真组冠军, 负责人, 2020.9
- 龙芯杯大学生系统能力竞赛全国三等奖, 负责人, 2020.6

专业服务

期刊审稿人

- IEEE Transactions on Network Science and Engineering

Shadow Program Committee

- ACM EuroSys 2026

Artifact Evaluation Committee

- USENIX OSDI/ATC 2025,2024
- USENIX FAST 2025
- ACM CoNEXT 2025
- ACM EuroSys 2025 Spring/Fall

技术能力

- 深度学习/系统开发：PyTorch, vLLM, Megatron, Verl, Triton, TensorRT, CUDA, NCCL
- 编程语言：Python, C++, LaTeX
- 英语：CET-4 619, CET-6 591, IELTS 7.0