刘芷溢

电话: 13470385770 | 邮箱: lzy_CS_LN@163.com | 现居城市：成都

网站: https://tom-jerr.github.io/

教育经历

- **电子科技大学** 985 **2024.9 - 至今**
  计算机科学与技术 硕士 计算机科学与工程学院 GPA: 3.6/4.00
- **电子科技大学** 985 **2020.9 - 2024.6**
  网络空间安全 学士 计算机科学与工程学院 GPA: 3.81/4.00

项目经历

- **SGLang** *2025 年 12 月 - 至今* [GitHub]
  - 针对 SGLang 中 Eagle2 **投机采样结合约束解码**时的 CPU-GPU 串行阻塞问题，将 CPU 的语法处理更新阶段与 GPU 的验证阶段进行并行重叠，基本上消除了流水线上的气泡，掩盖 CPU 计算与数据拷贝的延迟。开启 overlap 后使请求吞吐量提升 **21%** 左右，平均单字生成延迟 (TPOT) 从 **4.07ms 下降至 3.22ms**，显著提升了端到端推理性能。
  - 针对 Diffusion 模型 benchmark 端到端推理冷启动延迟过高问题，设计了轻量级 **Warm Up** 机制——构造一个推理步长为 1 的请求完整执行一次模型推理流程，后续请求可以跳过内核编译过程，降低端到端时延。在 H100 平台上将 Qwen-Image 模型的端到端推理延迟降低了 **31.3%** (Denoising Stage 降低 32.9%)(#PR 16213)
  - 拟写SGLang overlap scheduling、PD Disaggregation in SGLang等技术博客
- **MiniInfer** *2025 年 10 月 - 至今* [GitHub]
  - 支持 Qwen2 模型推理，完整实现了从 input -> tokenize -> forward -> sample -> detokenize 的全流程
  - **显存与调度优化**：实现了 Continuous Batching 和 Chunked Prefill 动态调度算法，最大化 GPU 利用率；引入 KV Cache 显存管理机制，显著提升长文本生成速度
  - **算子融合**：针对 Transformer 架构瓶颈，实现了 SiluAndMul、AddAndRMSNorm 等算子融合和权重融合，减少 CPU kernel launch 开销和访存开销
  - **内核性能调优**：编写了基于 Triton 的 RoPE 旋转位置编码算子，并集成 FlashAttention 加速库。经测试，Qwen2 模型的首字延迟 (TTFT) 从 33ms 优化至 28ms，性能提升约 15%
- **阿里开源数据库 OceanBase 比赛--向量索引查询优化** *2024 年 9 月 - 2024 年 12 月*
  - 阅读日志与火焰图找到耗时瓶颈，定位到向量索引查询中 HNSW 向量索引算法与磁盘 IO 为主要瓶颈
  - **向量索引性能调优**：基于 HNSW 算法进行深度优化。通过 SIMD 指令加速与 Int8 量化技术打破计算与带宽瓶颈，配合参数调优将系统 QPS 由 **592 提升至 1095**。引入 PEOs 动态路由算法，根据流量特征与查询代价实时动态选择最佳检索策略，实现了精度与性能的自适应平衡。
  - **SQL 执行引擎升级**：针对混合检索场景重构索引结构，全面实现算子谓词下推技术。通过在索引扫描阶段提前过滤不符合条件的数据，大幅降低数据读取量，使磁盘 I/O 降低 **70%**，解决了 IO 密集型场景下的性能瓶颈
  - **结果**：基准测试 QPS 提升 **201%**，混合查询延迟降低 **80.7%**，**全国第 11 名（1212 队）**。
- **Nebula Graph Ann Search** *2025 年 6 月 - 2025 年 9 月* [GitHub]
  - **支持向量类型与存储引擎改造**：设计并实现了 Vector 数据类型及其序列化机制；修改存储引擎，将向量属性独立存储于 RocksDB 的专用 Column Family，有效降低写放大并提升了扫描效率
  - **DDL/DML 适配与执行优化**：扩展 Parser、Planner 和 Executor 支持向量属性的定义与 DML 操作；实现了向量属性与常规属性的分离编码存储，确保了 Schema 的向前兼容性
  - **ANN 索引与 Ann Search**：设计了统一的 AnnIndex 接口集成 HNSWlib/FAISS；实现了 **AnnIndexScan** 算子及查询优化规则，将近似最近邻查询转化为高效的向量索引扫描，打通了从 Graphd 到 Storaged 的完整向量搜索链路

学习经历

- 学习 CSE234：实现数据并行和张量模型并行的通信层，包括 All-Reduce 和 All-to-All 通信原语
- 学习 MIT 6.S081：基于 RISC-V 实现了基础的操作系统内核
- 学习 CMU 15445：实现了单机数据库 Bustub，完善了数据库火山模型执行引擎并支持 MVCC 和 OCC 事务模型

获奖经历

- **Oceanbase 数据库大赛决赛 11 名** *2024.12* [🌐]
  全国大学生系统能力大赛
- **优秀奖学金 (一等)** *2024-2025*
  电子科技大学

技能和兴趣

- 编程语言：C++、Python、Rust
- SGLang Contributer，熟悉推理引擎侧常见优化手段和引擎执行流程
- 了解 DeepSeek 模型结构和实现，了解优化手段
- 了解现代计算机体系结构，掌握 CPU 并行编程、异步编程、CUDA 编程、Triton 编程、了解基础算子的实现
- 英语良好，能够流畅进行基础沟通和阅读英文论文