蔡雨洋

AI Infrastructure / LLM 推理系统

https://newmancai.github.io

教育背景

东南大学｜控制工程（电子信息）硕士 2024.09–2027.06

硕士生导师：孙长银教授（国家杰青）· 研究方向：HPC + 人工智能

东南大学｜自动化（机器人工程）本科 2020.09–2024.06

荣誉：CPIPC 创“芯”大赛国一｜ICCAD 2026 一作｜嵌入式芯片与系统设计国二｜优秀毕设：RL + Transformer 约束决策

实习经历

华为｜HPC / AI Infra 开源 · 优秀实习生 2026.06–2026.09

SGLang × DeepSeek · SME BF16 / W8A8 算子 · MLA / MoE 通信 · https://gitcode.com/kunpengcompute/kutacc/tree/br_dist_infer

- **SME BF16 微内核：**独立实现 Pack、Cache Blocking 与 2VL×2VL ZA/MOPA 微内核，以 Prefetch+二维分块优化鲲鹏 920F 32 核流水；M/N/K=4096/2048/4096 核心计算达 **25.47 TFLOPS（峰值 85%）**，端到端剖析定位 Pack/内存分配占时 53%。
- **W8A8 Decode 归因：**针对 W8A8 Decode 小 M 难以铺满 32 核，主导贯通 SGLang→C++ Adapter→KuTACC 的复测归因，验证 qkv_a 以 2 个 N block×16 个 K block 形成 32 路 Split-K 并归约 partial C；参与 IGEMM/Pack 验收，确认 Quant+Pack/Dequant+Save 静态流量为 0.6×/0.2×，左矩阵 L2 驻留+8 MOPA:1 Prefetch 下多个 GEMM 阶段有效带宽 **320–360 GB/s（Stream 80%–90%）**。
- **MLA 重排验收：**在 QK→online Softmax→PV 融合与 request/head/KV 段调度的整体后端中，建立 TP-8 Head→Batch 重排前后 shape/LSE 语义、输出一致性及跨 Die 通信校验；以 **Alltoall 4.5 MiB 时延 81.10 μs、跨 Die 123.01 GB/s** 复核切分转换链路。
- **MoE 通信联调：**协助 MoE Gather-Pack→IGEMM→Dispatch/Combine 链路联调，复核 token_ids 直接 Pack、双专家 N-slice 与批量 RDMA/RC Notify；EP256/BS64 下 **Dispatch/Combine 为 48/84 μs、142/163 GB/s**。

开源项目

vLLM-Omni｜MiniCPM-o 4.5 昇腾推理优化 2026.07–2026.09

OpenBMB × 华为昇腾挑战赛长期 Top 3 · 北大 × 东南 · https://github.com/newmancai/vllm-omni-minicpmo45-npu

- **前缀事务执行：**主导将 Stage 0 从 n-gram 候选升级为精确 TTS 契约下的整段文本/Prefill 复用：缓存 112-Token 模板 KV、合并宽 Prefill，不满足契约即回退；32 条固定请求上，7-gram/宽 Prefill 的 **Mean RTF 分别改善 4.24%/5.78%**。
- **Talker 图与状态：**主导重构 Talker 音频自回归执行时：ACL Graph 21→1、20 层 Fused Infer Attention 的 Python 逐层更新→2 次 C++ 分组调用，并将输入/游标纳入图生命周期；两项单独使 **Mean RTF 改善 20.8%/9.8%**。
- **Flow 蒸馏与编排：**主导蒸馏无需 Classifier-Free Guidance 的 Flow Student，将多步双分支压至单分支单步；chunk=25 下推理步数 2→1、Mean RTF 改善 6.74%，并在 Engine 层按 chunk 生产/消费依赖编排 Talker→Token2Wav 跨 Stage overlap，完整配置端到端 Mean RTF 0.4435→0.0795（-82.1%，5.58×）。

Kimi FlashKDA｜SM100 物理感知多智能体 Kernel 探索 2026.08–2026.09

北京大学未名超算 · 对标 NVIDIA/CMU CAKE（KDA 经 FlashInfer 接入 SGLang Kimi-K3，arXiv:2608.12629）· https://newmancai.github.io/kda

- **AI 双分支闭环：**主导构建传统 HMMA 与 Blackwell 原生 Tensor Core 双分支 Agent 搜索；多角色从双层 IR 读取程序状态与历史证据，在同一 workload 与预算下提出、组合、编译并验证候选；正负结果写回 IR，驱动下一轮探索和回退，108 项测试通过。
- **架构创新：**将 KDA 递推状态的独立维映射为 GPU 并行度，使 TP8 小 Head 的工作块 12→96，关键配置较官方降时 **9.37%–26.10%**；形成按请求形态选路的双后端策略，长计算采用 Blackwell 原生 Tensor Core，短计算回退传统路径，输出与状态位级一致。

科研项目

随机块低秩稀疏计算｜中科院计算所 × 华大九天 · SolverChallenge 国三 · ICCAD 2026 一作 2025.07–2026.04

- **低秩数据流：**主导低秩因子直连 Schur 更新，避免中间稠密块；固定布局 GEMM、在线重压缩控秩；**较 CKTSO 1.46×**。

LatentFit 共享低秩潜空间｜复旦数院 × 清华集成电路 · CPIPC 创“芯”大赛国一 · ISEDA 2025 2024.09–2024.12

- **潜空间与后端：**主导以截断 SVD 按尾能量压秩，在潜空间求解后回投影；按 shape 路由 SVD/QR；**rank 67（393×）、误差 0.358%**。

开源交流

- **SGLang-Omni · Astra 架构预研：**参与 ASR/TTS PR/CI；按 looped Transformer 假设，推演权重循环、隐状态缓存与 CUDA Graph 复用。
- **vLLM Meetup · Day-0 适配：**结合 MiniCPM-o/昇腾实践，形成 Model×Hardware×Engine 接入方法：解析新 attention/MoE/多模态算子与状态契约，匹配 backend/kernel/并行映射，再以 parser、正确性与单卡→集群性能门禁收口。
- **腾讯开源 · 记忆系统编排：**Issue #120 单题 Top 3；独立修复 Recall 请求/历史写回不一致，**cache hit 71.36%→99.53%**；继而推进 Memory Beta 零反馈 Memory Runtime，对标 Codex / Claude Code 的任务记忆机制，构建“执行轨迹→分层记忆→跨任务召回”链路，按来源、记忆类型与任务作用域约束写入，并以当前状态和工具返回复核历史结论，避免旧状态干扰当前执行。