

郑天宇

性别：男

民族： 汉族

邮箱(同手机): 17786324919@163.com



电子科技大学	计算机科学与工程学院	计算机技术专业	硕士	2024.9-2027.6	年级排名前3%
武汉科技大学	计算机科学与技术学院	计算机科学与技术专业	本科	2020.9-2024.6	年级排名前1%

竞赛获奖：“挑战杯”专项赛国家二等奖、CRH大学生算法挑战赛国家一等奖、“华为杯”研究生数学建模竞赛国家二等奖、MathorCup高校数学建模挑战赛国家二等奖、美国大学生数学建模大赛国际二等奖(H奖)、全国大学生数学竞赛国家三等奖。

● 2026.03-至今 字节跳动 Seed-Infra-训练

- **显存极值优化与防死锁设计**：主导设计 **Min_Memory_Any1F1B 调度策略**。为解决复杂通信编排极易导致的 Hang（死锁）问题，摒弃初步设计的强依赖异步通信的方案，设计新的**奇偶通信策略**并深度编排 PP 首尾及中间段落的通信调度逻辑。在 2机16卡集群（PP=4）测试下，单卡峰值显存**大幅降低 5GB**，实现同比于1F1B的**0峰值显存增长**。
- **长尾气泡消除**：针对 $\text{remainder} * 2 < \text{pp_size}$ 时产生的无规律长尾气泡，**首创启发式编排算法（边缘/中值/级联插入）**，实现流水线**气泡零增长**，随着vpp的增大，**理论bubble优化率可逼近100%**；同时实现 Warmup/Cooldown 阶段通信与计算的 Overlap，并深度**适配 ZeRO 通信overlap**。
- **PP 通信bubble分析拆解**：量化空转 / 通信 Bubble 占比并建立异常判定标准；研究影响因素并明确优化手段；公式化优化手段与 Bubble 率、MFU 的换算关系，量化显存 - 性能 trade-off。

- **MTP 层算子融合与优化**：开发 CE-loss+Z-loss+Logits 融合triton kernel；移除atomic_add保证deterministic；通过显式分配 (torch.empty+fill_) 消除 **CPU Straggler** (244ms→0.568ms)，gpu kernel(12.2 ms→9.4 ms)并利用 Logits 地址 Inplace 存梯度**复用显存**。在 64 卡下，**PP-3 峰值显存降低 9GB**，MFU 提升 0.013。
- **自研架构 (VWN) Kernel 集成**：在大算子内部以及 MTP 层集成 VWN (类MHC) 融合算子。64 卡评测单步kernel耗时缩短约 1.2ms，**PP-3 显存降低 4GB**，**MFU 提升 0.009**。
- **Zero Overlap Buffer 静态化优化**：针对 chunk 间梯度通信/计算 overlap 中**动态 resize**引发的显存碎片、**cudaStream synchronize 阻塞问题**，设计 **big/small 双 buffer 静态复用机制**，基于相邻 chunk 显存需求自适应确定 small buffer 上界，保证相邻 chunk 间 RS 与 backward compute 持续 overlap，lazy分配small buffer，param.main_grad原地更新，**在解决同步阻塞问题同时，几乎未增长峰值显存**。

- **全链路端到端对齐**：主导多模态（text/vision/audio）跨框架对齐，逐规模从4卡扩充到512卡降低降低对齐成本；对齐训练指标 `dataset_loss / loss/ grad_norm`，支撑 万卡上线；沉淀"逐层 dump md5+ 配置消融 + 4 卡 bitwise对齐 encoder+纯文本到多模态+混合数据消融"对齐 SOP。设计 **key mapping 机制**，对齐统计口径(如reduce level)。
- **重要 Bug 定位修复**：通过逐层 `grad_norm/activation dump`，定位修复 **noop 层 mapping 错误**、**多模态meta在pp架构下切分机制纠错**、**sp通算融合逻辑错误导致 kv offset 异位**、**vrope 施加错位**、**特殊数据的mask处理未对齐**等多个核心问题。
- **累积数据指标save load逻辑优化**：设计扩缩卡下全局数据读写方案，通过 **save 强制 flush 与单 rank 读写**，保证数据正确性、消除读热点并兼容旧 ckpt。

- **KVC 主链路接入与精度对齐**：在 Megatron 中实现 KVC 及 层间共享 hidden state，基于前序工作，做增量对齐，4 卡开发 → 64 卡验证跑通，支撑上线，大幅降低对齐成本。
- **Meta/Mask 瓶颈消除**：单卡 32 mb、4k seq_len 下 kvc meta/mask 构建严重阻塞 e2e，通过"关键元数据 dataloader 预构建 + 消除 D2H 同步 + kernel 融合与 PP 侧 overlap"分层优化，e2e 从 40s 的关键路径时间降至 100ms。
- **Bigop 单层 tranformer block 手写算子**：在 BigOp 中支持 kvc 与层间共享 hidden state 的高性能 triton kernel；优化 oe 的 fwd(通信 bound) 与 bwd(访存 bound)；深度设计 fwd 和 bwd 通新与计算的 overlap，全覆盖通信 buffer；group gemm cutedsl kernel 优化重排 tile 与 tune 分块大小；支持 pp 通信从 send recv 转换到单边 put。保障精度对齐与上线稳定性，单层 mfu_gpt_fwd 从 0.17 优化到最终的 0.32。
- **NVSHMEM Symmetric Memory 通信设计**：通过 pp 层间均摊共享 buffer 设计显存最优路径，同时通过 pp 调度逻辑设计 onlyput 的性能最优路径。
- **大规模训练理论建模与多机验证**：为提前识别上线前资源/效率风险，围绕显存构成、MFU 口径差异等指标做理论计算和拆解，并在 64 卡验证中结合 profile 对照验证，降低上线不确定性，总结 fsdp 和 megatron 性能的理论分析。

大模型分布式FlexCheckpoint(FC)系统研发 [成果链接](#)

简介: FlexCheckpoint系统基于DCP和AOA协议，解决LLM训练中ckpt跨并行策略、跨模型结构转换的通用性问题。

- 研发**飞桨首创All-in-One-Arrow(AOA)**协议，用简洁箭头语言表达load save过程任意模型结构的转换，构建AOA语句解析—>批量macro处理—>7种原语级操作执行—>全局Tensor转本地映射的完整流程，降低生产与开源成本。
- 在Ernie4.5、Llama2上验证dp,sd,pp,tp,ep混合并行策略互转的场景，在三种不同验证策略下md5对齐且精度无损。
- 基础Layer、Optimizer、Zero1~3等组件适配FC，提出动态最小reshard，解决ShardingV2参数flatten的难题。
- 提出full param概念，实现无组网下多卡并行合成单卡视角的safetensor权重文件，提高训练对接推理工作的效率。
- 在Paddlenlp、PaddleFormers库中适配FC，并实现AOA Reverse、AOA Traceback、locally_load等功能优化。
- 构建SendRecv、Broadcast、GroupBroadcast三种不同的Communicator优化不同场景的load通信。

动/静态图下的流水并行功能增强和性能优化 [成果链接](#)

- 通过解耦静态图下的通信与计算算子，解决1F1B、Vpp策略通信死锁问题。
- 开发整套动态图自动并行框架下的pp并行组件：实现FthenB、1F1B、VPP并行策略。解耦底层计算通信与编排方法，同时对底层通信进行了统一抽象更易扩展新的pp策略。并复现Dualpipe编排(开源前)。
- 开发友好API设计：封装自动/手动并行切换逻辑，推出轻量级localmap API，并提供完整文档。[官网示例](#)。
- 修复RoPE kernel反向逻辑错误。优化Zero1的grad通信操作，在4机16卡llama 7b纯Zero并行下性能提升7%。

自动并行下vpp流水并行策略的功能增强 [成果链接](#)

- 支持持静态图编排阶段任意数量的Micro-batch 累积、重切分阶段非均匀的VPP(Interleaved-1F1B)
-) 模型块分配。优化非均衡vpp的编排，buble率从26%降低到14%。
- 开发aadiff_check，检查所有pp并行策略下bubble期的不确定性错误。参与构建ZB-VPP显存预测模型。
- 提出了一种VPP显存峰值优化策略，并通过数学公式推导及流水线编排图验证其可行性。在LLaMA2-7B模型、4卡VPP(VPP degree=2)配置下，0~2号卡上的显存峰值降低幅度为12.09%~6.11%，且训练时长基本保持不变。

● 2024.07-2025.05

上海索塔沃科技有限公司

早期核心成员：大模型算法工程师

格莱美企业助手、新华三(H3C)公司平台业务：搭建dify工作流，prompt优化，训练本地用户意图分类模型，使用MCP策略，创建agent结点调用多个tools，后端部署项目dow(基于wechat-bot)，并进行后期的迭代优化工作。参与多个RAG项目包括基于R2R、langmanus等项目的研发，基于langchain、langgraph搭建从planning到execute的multi_agent框架。

● 2023.11-2024.02

北京小米移动软件有限公司

测试开发工程师

基于Qt开发内部多线程脚本程序，实现任务自动化。使用CANoe工具分析dbc文件，截取测试期间异常的can报文信息。

项目经历

● APCD 大语言模型自适应可靠解码 [论文链接](#)

2025.08-2025.12

主要工作:

- 引入信息论工具(Shannon Entropy)作为解码分岔指标，当且仅当Top-k Token 不确定性超过阈值时触发分叉，确保高效且适时的分支点，防止引入噪声路径，极大降低了传统全局多路径解码的显存与时间开销。
- 设计跨路径动态Logit 惩罚策略，通过持续追踪平行路径间的Jensen-Shannon Divergence(JSD)，自适应衰减对比权重(Contrastive Strength)，完美平衡了生成文本的多样性与上下文连贯性。
- 在多个主流大模型(LLaMA、Qwen等)及10项baseline method(Greedy、Sample、CS、CD、Dola、AD、ITI、CAD、End、UCD)上完成评测基建，实现准确率全面提升，证实了该框架在高风险领域落地的可行性。

● 基于预训练大语言模型的文档问答系统研发(项目总负责人) [本科项目总览](#)

2024.02-2024.06

主要工作: 基于LangChain构建KG-QA与RAG-QA文档问答系统，对比其构建复杂度与问答精度，并提出融合两者优势的K-RAG混合问答模型，综合性能提高。自定义分割文档、部署本地chatglm3-6b、Fine-tune、prompt优化等工作。

专业知识

手推FlashAttentionV1,V2公式，掌握dp,tp,pp,sp,cp,ep等多种并行策略，熟悉Megatron和deepspeed框架，掌握MQA，GQA，MLA，KV Cache，Paged Attention，Native Spare Attention等知识，以及RAG、SFT等知识。

科研成果

(中科院一区)ESWA(投递中):《APCD: Adaptive Path-Contrastive Decoding ...》 CPCI、EI会议论文 软著一项

所获荣誉

优秀学生/优秀学生干部/优秀学生标兵 校级特等奖学金/研究生一等奖学金/汇顶科技专项奖学金 百度开源之星/开源先锋