趋境科技
APPROACHING.AI

KTransformers项目简介

趋境科技主导，首创异构推理技术架构，全球Top5推理框架

趋境科技技术团队源自清华大学计算机系高性能计算所

趋境科技 APPROACHING.AI

源自清华大学计算机系高性能计算所

10数年的高性能计算、分布式存储、AI计算引擎优化等领域技术积累

教育背景

团队成员多毕业于

清华、新加坡国立、北航、北邮、北理 90% 以上硕士、50% 以上博士

产业背景

团队成员具有

英特尔、字节跳动、百度、深信服 等多家头部科技企业资深经验

趋境科技联合创新团队 - KVCache.AI

郑纬民院士 首席科学顾问

- 中国工程院院士，清华大学计算机科学与技术系教授
- 超算领域专家，海致科技首席科学家
- 博士生导师，高性能计算研究所所长

深耕AI Infra领域，聚焦于私有化大模型高效推理，突破显存占用高、算力依赖性强等传统方案瓶颈，实现资源利用率最大化

国际顶级会议和期刊论文100余篇

OSDI	ASPLOS	FSE	ATC
SOSP	HPCA	VLDB	EuroSys

破解大模型本地化部署的效率、效果、成本间的不可能三角

趋境科技 APPROACHING.AI

需要降低大模型的部署和推理成本

- 私有化部署大模型更能保护数据安全、更能及时高效处理问题、更具有针对性、更不受网络条件限制
- 大模型推理需要大量计算资源，部署成本百万+

核心问题：GPU算力制约

私有化大模型落地痛点：
效果、效率、成本不可能三角

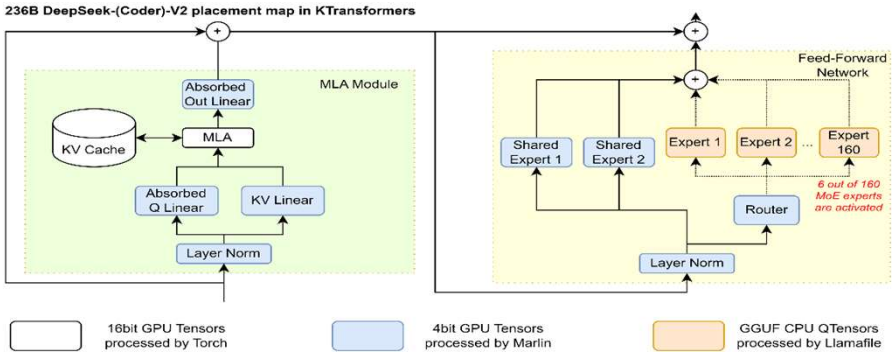
需要用参数更大、效果更好的模型

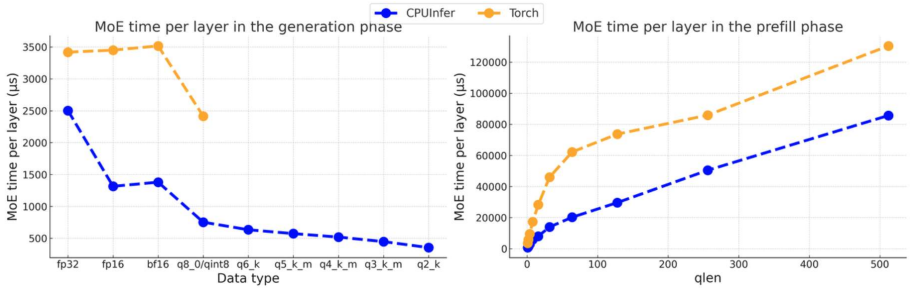
- 大模型符合Scaling Law规则，模型参数越大、使用效果越好
- 提示词越长，大模型参考的上下文内容越长，回答问题更加全面

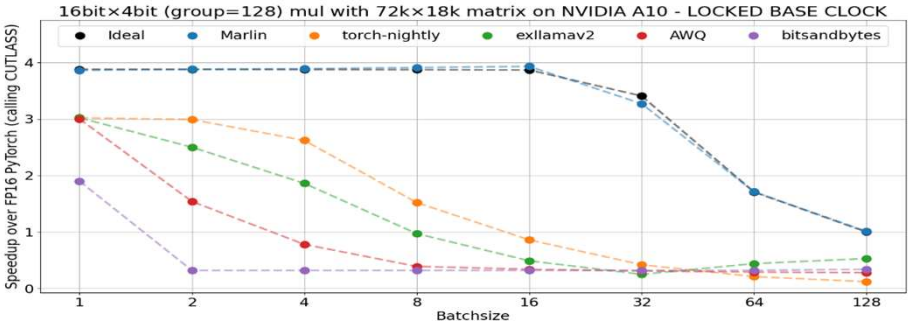
需要更低的延迟、更高的吞吐

- 模型输入后，等待模型输出的响应时间要短
- 模型生成速度快，同时支持的在线用户数高

KTransformers 异构推理、异构微调技术策略







基于计算强度的 Offload 策略

Offload 优先级：Routed Experts > Shared Experts > MLA Attention

高性能 CPU 算子框架 CPUInfer

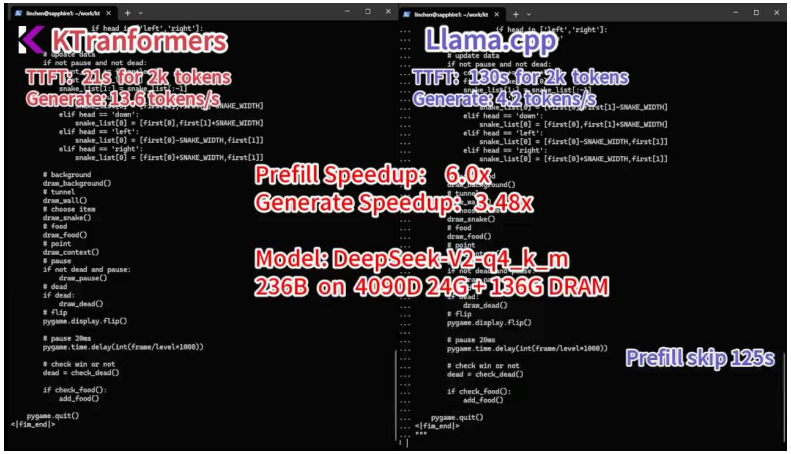
Attention基于 llama.cpp 的 ggml 量化格式和 llamafile 的高性能算子进一步改造；增加多线程、任务调度、负载均衡、NUMA 感知等优化

高性能 GPU 算子Marlin

引入 Marlin 算子作为 GPU 计算的 kernel

单张 4090 运行 DeepSeek-671B大模型
运行速度 20 tokens/s

2张4090微调Kimi-K2-1TB大模型
微调速度 40 tokens/s↓



KTransformers 异构推理引擎已位列全球热度前五

趋境科技 APPROACHING.AI

探索性开源框架

广泛的技术研发生态

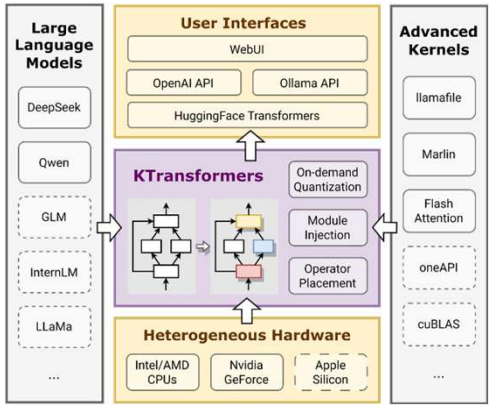
2024 年 7 月正式开源，单 GPU + 136GB 内存支持 DeepSeek V2 的异构推理

2024 年 8月 更新对于本地 1M 级别长文本的支持，满分“大海捞针”

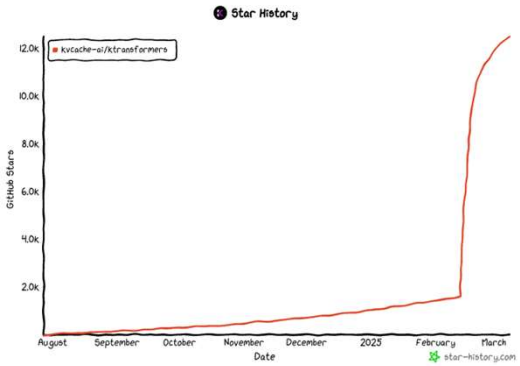
2025 年 2 月 更新对 DeepSeek-V3/R1 的支持，单 GPU + 382GB 内存可支持本地推理，同等环境下相比 llama.cpp 快 3~28 倍

2025年4月，KTransformers强势崛起
成为全球Top 5最受技术人士认可的LLM推理引擎

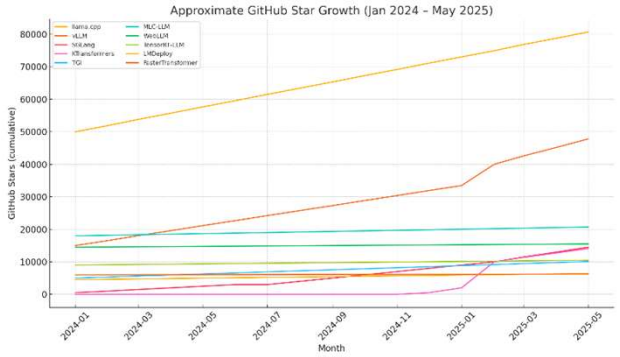
Now. 更多的国产平台支持



(a) 灵活的支持框架



(b) GitHub星标15k+，排名 Github 前万分之一



(c) KTransformers快速成为全球Top 5受技术人士欢迎的LLM推理引擎

NVIDIA. intel. Kunpeng Ascend MTHREADS 沐曦 天数智芯 Iluvatar CoreX HYGON 中科海光

(d) 多家 CPU/GPU 硬件厂商主动参与开源共建

KTransformers，引领推理引擎技术发展，广泛的国内外影响力

趋境科技 APPROACHING.AI

探索性开源框架

2024 年 7 月正式开源，单 GPU + 136GB 内存支持 DeepSeek V2 的异构推理

2024 年 8月 更新对于本地 1M 级别长文本的支持，满分“大海捞针”

2025 年 2 月 更新对 DeepSeek-V3/R1 的支持，单 GPU + 382GB 内存可支持本地推理，同等环境下相比 llama.cpp 快 3~28 倍

2025年4月，KTransformers强势崛起
成为全球Top 5最受技术人士认可的LLM推理引擎

广泛的技术研发生态

Now. 更多的国产平台支持

KTransformers ☆ Star 14.9k

A Flexible Framework for Experiencing Cutting-edge LLM Inference Optimizations

Show Cases | Quick Start | Tutorial | Discussion

NVIDIA AMD intel Kunpeng

METAX 沐曦集成电路 Ascend

国内外多家主流 CPU/GPU 硬件厂商
主动参与开源共建

Enflame 燧原科技 天数智芯 Iluvatar CoreX HYGON 中科海光 摩尔线程 MOORE THREADS

计算机顶会SOSP 2025

B站、Reddit热榜第一

CCTV-13 大模型专题报道

KIMI/GLM/Qwen等大模型官方首发推荐引擎

KTransformers 商业化路径：填补低门槛大模型部署市场

趋境科技 APPROACHING.AI

deepseek

千万级
300多张卡的集群方案
超大并发、超高部署门槛

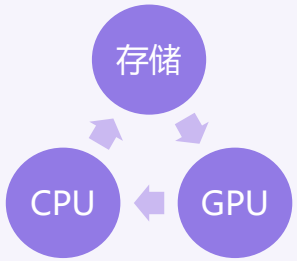
数百万级
H20/H200、910B等
小规模集群
中等并发、高部署门槛

数十万级
中等并发、中低部署门槛

产品缺失

万级
低并发、低部署门槛

数千元级
CPU 方案
个人使用，性能弱



业界首创异构协同大模型推理架构设计

充分利用GPU、CPU、内存等所有算力资源

首次实现在单个消费级GPU+CPU异构推理千亿大模型

- Prefill（响应延迟）相比业界方案快 30倍以上
- Generate（生成速度）相比业界方案快 3倍以上

KTransformers

A Flexible Framework for Experiencing Cutting-edge LLM Inference Optimizations

Show Cases | Quick Start | Tutorial | Discussion

deepseek

千万级
300多张卡的集群方案
超大并发、超高部署门槛

数百万级
H20/H200、910B等
小规模集群
中等并发、高部署门槛

数十万级
单机或小规模集群
中等并发、中低部署门槛

降低成本填补市场

万级
单GPU+CPU+大内存
低并发、低部署门槛

数千元级
CPU 方案
个人使用，性能弱

KTransformers 商业化路径：建立广泛的上下游合作生态

趋境科技 APPROACHING.AI

中国电子云　长亭科技　蚂蚁金服　海致科技　世纪互联　图灵法思

数秦科技　Chat Excel　未来式智能　白山云科技　方寸智能　中科雨辰

超过50+
ISV合作伙伴

趋境科技 APPROACHING.AI

与国内外主流芯片厂商
均建立合作关系

华为海思　沐曦　昇腾　鲲鹏　燧原　Intel

AMD　中兴微电子　摩尔线程　天数　飞腾　海光

趋境科技
APPROACHING.AI

为世界同时普惠
顶尖AI智能与数据隐私