vllm-project/vllm 2026 第 26 周周报 (06-22 至 06-28)

vllm-project/vllm

周期：2026-06-22 至 2026-06-28

来源 PR：302・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/vllm-project/vllm/reports/2026-06-22-to-2026-06-28

执行摘要

2026 年第 26 周（06/22 - 06/28），仓库共合并 302 个 PR，其中重点 PR 24 个。本周工作主要围绕三个方向：DeepSeek V3.2 / GLM-5 模型推理性能优化、ROCm 平台内核与 MoE 加速、以及 Rust 前端与基础设施重构。标签统计显示 bugfix 占比最高（132），其次是 v1（91）和 rocm（71），反映出缺陷修复与架构迁移并行推进的特点。风险方面，“核心路径变更”标记出现 50 次，“缺少测试覆盖”出现 39 次，需重点关注。

本周重点变化

模型与核心性能

本周最大亮点是 GLM-5 / DeepSeek V3.2 性能优化系列。WoosukKwon 提交的 #46876 实现了多算子融合 Triton 内核（fused_norm_rope、fused_q 等），在 GB200 上达到约 290 tok/s。另一 PR #46635 将 MoE all-reduce 替换为 reduce-scatter 获得 3.1% 吞吐提升。#46862 的 fused_indexer_q_rope_quant 内核再次提升 GLM-5.2 吞吐 1.9%-3.3%。这些优化通过 kernel 融合和通信模式改进，对大规模部署有直接价值。

ROCm 平台支持

ROCm 本周贡献突出。hongxiayang 的 #46184 为 gfx950 集成 AITER FlyDSL MXFP8 MoE，并实现后端自动选择。同作者在 #46546 中优化 sparse attention 内核，性能提升最高 42%。#46474 为 MiniMax-M3 融合共享专家至 AITER MoE，TPOT 降低 4-17%。此外，#46290 修复了 MoRIIO WRITE 模式下的混合 KV 布局问题，#46332 支持异构 TP fan-in。AMD 团队的持续投入显著缩小了与 NVIDIA 平台的性能差距。

Rust 前端重构

BugenZhao 主导的 Rust 前端重构本周集中合并。核心 PR #46583 引入了统一解析器接口 UnifiedParser，消除 tool/reasoning 解析器的两阶段分离。#46602 将 Gemma4 迁移至统一解析器，支持推理状态与工具调用的交织输出。#46051 将 HTTP/request-processing/ZMQ 分离为独立 Tokio runtime，在长文本压力测试下吞吐提升 30%，/health 尾部延迟下降 93%。这些重构提升了系统架构清晰度和高并发性能。

KV offload 与连接器

KV offload 子系统经历重要重构。hickeyma 的 #45053 用方向显式的 OffloadingWorker 替代旧 Handler，消除路由调度层。Change72 的 #43468 为 OffloadingConnector 添加自描

述 KV 事件，供外部消费者（如 Dynamo）路由。LucasWilkinson 的 #46252 用配置替换环境变量控制 packed KV cache。这些都增强了系统的可维护性和可观测性。

MoE 与量化基础设施

MoE 后端体系引入抽象基类 MoEKernelOracle（#43461），统一内核选择接口。fxmarty-amd 的 #44667 融合 NVFP4 反量化与计算内核，性能提升可达 10x。#45924 集成 Tencent HPC-Ops 后端，#46393 新增 FlashInfer CuTe-DSL MXFP8 线性内核，进一步丰富了 MoE 内核选项。此外，#45703 扩展 Marlin 线程块填充至 MoE，提升 WNA16 和 FP8/MXFP8 性能。

模型清理与多模态扩展

本周移除了 Grok（#46706）、Baichuan（#46362）、Aquila（#46605）、MiniMax Text01/VL01（#45993）等旧架构，主因已被新架构取代。新模型支持方面，gty111 的 #46564 集成百度 Unlimited-OCR 模型，实现参考滑动窗口注意力。nagisa-kunhah 的 #44124 支持 OpenMOSS-Team 的 MOSS-Audio 系列。brandonpelfrey 的 #44465 引入 VRAM 信号量和 GPU 视频解码后端，为零拷贝解码推理铺路。

模块与主题趋势

从标签看，bugfix（132）、v1（91）、rocm（71）、performance（57）占据主导，表明团队在修复缺陷的同时大力推动 v1 架构落地和 ROCm 性能优化。CI/build（49）和 test（47）的高频出现说明持续集成和测试覆盖是工作常态。kernel（32）标签的活跃反映了底层算子优化的重视。

热点文件集中在 CI 配置（test-amd.yaml 11 次）、DeepSeek/GLM 模型文件（deepseek_v2.py 4 次，deepseek_v32/nvidia/ 多文件）、以及 KV offload / Mooncake 相关文件。说明 AMD CI 整合、DeepSeek 模型开发和 KV 传输层是本周的焦点模块。

贡献者方面，njhill（12）、micah-wil（11）、BugenZhao（10）、WoosukKwon（9）、mgoin（9）位居前列，形成核心开发力量。

风险观察

1. 核心路径变更频繁：本周大量 PR 涉及注意力后端（FlashInfer、Triton MoE）、稀疏索引器、模型定义等核心路径，且部分缺少足够测试覆盖，可能引入性能退化或正确性问题。
2. 外部依赖风险：AITER、HPC-Ops、xgrammar、FlashInfer 等第三方库被广泛集成，其版本升级或接口变更可能导致构建失败或运行时异常。
3. VRAM 信号量并发死锁：#44465 引入基于条件变量的信号量，多线程环境下可能隐藏死锁，需加强压力测试。
4. Rust 前端行为兼容性：统一解析器接口迁移和 reasoning-parser 语义修正（#46359）可能改变工具调用输出，需确保与 Python 前端行为一致。
5. MoE oracle 重构过渡期：新抽象基类与遗留函数并存，后续量化后端迁移过程中可能出现接口不匹配或遗漏，建议加快迁移进度。

重点 PR 速览

- #46876 [GLM5] op fusion：WoosukKwon 实现多 Triton 内核融合，在 GB200 TP8 下约 290 tok/s，是本周性能优化代表作。风险：H200 兼容性暂未修复。
- #46564 [Model] Unlimited OCR：gty111 集成参考滑动窗口注意力，为 vLLM 增加新的注意力模式，对理解注意力后端架构有较高参考价值。
- #46184 [ROCm] flydsl moe：hongxiayang 为 gfx950 集成 AITER FlyDSL MXFP8 MoE，展示体系结构感知的 kernel 集成范式。
- #46583 [Rust Frontend] unified parser interface：BugenZhao 合并 tool/reasoning 解析器，消除两阶段分离，为后续流式块输出打下基础。
- #45053 [KV Offload] replace Handler with Worker：hickeyma 用方向显式 Worker 重构 offload 子系统，大幅降低复杂度。
- #43461 [MoE] kernel oracle ABC：qyYue1389 引入 MoEKernelOracle 抽象基类，统一 MoE 内核选择接口，为后续重构铺路。
- #44465 VRAM semaphore：Brandonpelfrey 实现基于信号量的 GPU 内存管理，为多模态零拷贝解码推理奠定基础。

后续建议

1. 加强测试覆盖：对于核心路径变更的 PR，建议要求至少包含单元测试或集成测试，特别是注意力后端、MoE 调度和稀疏索引器。
2. 监控外部依赖：建立外部库的版本追踪和回归测试机制，及时更新构建兼容性文档。
3. 推进 MoE oracle 迁移：尽快将剩余量化后端（FP8、NVFP4 等）迁移至新 Oracle 接口，减少新旧并存带来的维护成本。
4. 压力测试 VRAM 信号量：在多实例高并发场景下验证 #44465 的死锁安全性，考虑添加超时和恢复机制。
5. Rust 前端行为验证：针对统一解析器变更，应补充工具调用和推理标记的 end-to-end 测试，确保与 Python 前端输出一致。
6. ROCm 优化泛化：本周 ROCm 优化多集中于特定模型（MiniMax-M3、DeepSeek），建议扩展到更多模型验证，并建立跨模型 benchmark 基线。