vllm 2026 年第 24 周周报（06-08 ~ 06-14）

vllm-project/vllm

周期：2026-06-08 至 2026-06-14

来源 PR：256・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/vllm-project/vllm/reports/2026-06-08-to-2026-06-14

1. 执行摘要

本周（2026-06-08 至 2026-06-14）vllm 仓库共合并 256 个 PR，其中 24 个为深度分析的重点 PR。整体呈现三大主线：架构重构与清理（MoE 所有权反转、GGUF 迁移、旧模型移除）、推送模式与动态策略（动态推测解码、Nixl KV 推送、异步批处理）、硬件平台性能突破（ROCm 多 kernel 融合、Helion 量化核、SM90 FP8 GEMM）。同时工具调用和前端解析器经历了一次统一重构，将多个分散的解析逻辑整合到 HarmonyParser 和 DelegatingParser 体系下，代码行数净减约 1.5 万行（仅 GGUF 移除约 9000 行）。风险层面需重点关注核心路径的测试覆盖缺口和新特性对 CUDA Graph 的兼容性影响。

2. 本周重点变化

2.1 MoE 架构重构与通信升级

#41184将 FusedMoE 拆分为 RoutedExperts（仅权重）和 MoERunner（前向逻辑），涉及 90 个文件的权重路径映射。这一“所有权反转”使 future MoE 的优化（如弹性 EP、量化融合）不再受困于既有层间绑定。同步合并的 #41183集成了 DeepEP v2 的弹性缓冲区 all2all 后端，支持 CUDA Graph，且通过运行时 NCCL 版本检测（≥2.30.4）自动启用。两 PR 联手将 vllm 的 MoE 基础设施从“硬编码”推向“可插拔”。

2.2 推测解码动态化与多样化

#32374引入动态推测解码（DSD），根据 batch size 调整推测 token 数，避免了高峰期因验证阶段计算量激增导致的负优化。设计经历从独立 Manager 类到调度器集成查找表的重构，同时自动降级 CUDA Graph 为 PIECEWISE 模式。另一条分支 #44586在 MRV2 中实现了 DFlash 推测解码器，通过复用 draft model 隐藏状态和自定义 CudaGraphManager，在 GB200 上获得约 1.2 倍加速。此外 #45163新增 DiffusionGemma 扩散模型，复用推测解码度量路径，展示了非自回归模型与投机解码框架的整合模式。

2.3 ROCm 性能优化密集落地

本周 ROCm 相关 PR 达 27 个，其中 3 个重点 PR 带来显著性能提升：

- #44400为 gfx950 启用 FlyDSL W4A16 MoE 内核，在 Kimi K2.5 INT4 模型上提升端到端吞吐量，需 --moe-backend=flydsl 且未启用 LoRA 时触发。
- #42864融合 all_reduce、RMSNorm 和 per-group FP8 量化为单一 AITER 算子，并处理了 DSv3.2 中 RMSNorm 输出的扇出情况，emit_bf16 复用 bf16 输出避免额外 kernel。

- #45103在解码阶段融合逆 RoPE 与 wo_a 缓存，将 DSv4 解码性能提升 9-16%，是 MLA 注意力优化的又一步。

这些优化共同将 ROCm gfx950 在 MoE、注意力、编译融合等方面的能力推上新台阶。

2.4 工具调用与解析器统一重构

前端解析迎来一波大型重构：

- #45003删除了～1300 行的自定义 XML 流式解析器，转而利用 xgrammar 的 structural_tag 强制执行工具调用 JSON Schema，默认启用严格模式（但取消 VLLM_ENFORCE_STRICT_TOOL_CALLING 环境变量引发兼容性担忧）。
- #45171和 #45104分别将 Harmony 模型的非流式和流式解析逻辑从 serving 层剥离，统一集中到新建的 HarmonyParser（继承自 DelegatingParser），并删除了旧的 OpenAIToolParser 等模块。
- #44596将 Mistral 工具解析也抽离为独立 Parser，#44624为 Rust 工具解析器建立 PyO3 桥接，使得未来解析器可同时在 Python 和 Rust 两端实现。

这一系列 PR 将前端解析从“散落在各处”收敛到“Parser 层次结构”，极大降低了后续维护和统一流式 / 非流式行为的成本。

2.5 KV 传输与卸载：推送模式、异步查找与可观测性

- #35264实现了 Nixl KV 推送模式，Prefill 节点在完成时主动推 KV 给 Decode，避免 Decode 轮询拉取，TTFT 提升 1.2x~3.0x。设计上提取基类 NixlBaseConnector，并将推送 / 拉取作为子类，支持异构 TP 和块大小。
- #44193为次级 KV 缓存层添加异步批处理查找，通过 AsyncLookupManager 和后台线程将逐 key 同步查找改为批量异步，缓解了调度器线程的阻塞。
- #35669为 offloading 管理器引入标准化 get_stats() 接口和 Counter/Gauge/Histogram 三类语义，采用扁平自描述载荷设计，并妥善处理了旧指标的 deprecation。

这些进展使 KV 传输在性能、异步性和可观测性三个方向均取得实质进展。

2.6 量化契约与 kernel 融合起航

#44260定义了 QuantizedActivation 数据类和 as_quantized_activation 接口，建立了“kernel 声明支持的量化键 → 消费端直接使用预量化激活”的契约。这是 #43224 手动融合计划的第一步，已适配 FP8 scaled-mm 和 NVFP4 cutlass 两种 kernel。#45295则统一了所有 Marlin 变体的 tile 填充逻辑，在 TP 分片产生不匹配形状时自动零填充，消除了多个量化路径的崩溃和慢速回退。两 PR 标志着 vllm 量化融合从“各 kernel 自行其是”向“统一契约”的转变。

3. 模块与主题趋势

模块	趋势	代表 PR
MoE	架构重构 + 弹性通信 + 量化兼容	#41184, #41183, #42864, #45295

模块	趋势	代表 PR
推测解码	动态长度 +DFlash+ 扩散模型适配	#32374, #44586, #45163
ROCm 优化	MoE/ 注意力 / 编译融合多点开花	#44400, #42864, #45103, #44899
前端解析	统一 Parser 体系 + 工具严格模式	#45003, #45171, #45104, #44596
KV 传输	推送模式 + 异步查 + 可观测性	#35264, #44193, #35669
量化	契约化 +Marlin tile 统一 + 在线 FP8	#44260, #45295, #44132
清理	GGUF 迁出 + 旧模型删除 + 废弃代码	#39612, #44992, #45131, #45129, #45127
Rust 前端	安全认证 + 结构化输出 + 暂停 / 恢复	#44321, #44729, #44499, #44856

4. 风险观察

- 测试覆盖缺口：本周 32 个 PR 标记“核心路径变更”，36 个标记“缺少测试覆盖”。虽然重点 PR 大多附带单元测试，但集成测试和性能回归测试仍需补强。例如 #35264 的推送模式缺少 *e2e* 压力测试，#41183 的 DeepEP v2 仅在限定 NCCL 版本下验证。
- CUDA Graph 降级成本：动态 SD、DFlash、DeepEP v2 均需在推理流图中特殊处理（从全图降级为 PIECEWISE），这种“降级”增加了运行时分支和复杂度，建议后续设计时优先考虑与 CUDA Graph 原生兼容的静态方案。
- ROCm生态依赖风险：多个ROCm优化依赖特定AITER版本（如#42864依赖0.1.14）、FlyDSL 内核（#44400），这些外部库的版本锁定和向下兼容性未经验证，可能阻碍用户的平滑升级。
- 默认行为改变：#45003 移除了 VLLM_ENFORCE_STRICT_TOOL_CALLING 环境变量导致严格模式默认开启，可能影响使用流水线并行或推测解码等不兼容特性的用户，建议提供一个逃生门机制。

5. 重点 PR 速览

- #41184 MoE 重构：FusedMoE 降级为工厂，MoERunner 接管前向，权重路由至 RoutedExperts。涉及 90+ 文件，是 vllm 历史上最大的单次重构之一。
- #41183 DeepEP v2：集成弹性缓冲 all2all，支持 CUDA Graph，运行时检测 NCCL 版本，MoE 通信模块化。

- #32374 动态 SD：调度器根据 batch size 查表决定推测长度，自动处理 CUDA Graph 降级，移除 Manager 类，设计几经迭代。
- #44400 ROCm FlyDSL MoE：gfx950 专用 W4A16 内核，支持自动回退（LoRA 时走 Triton），配置表按设备 + 模型名匹配。
- #42864 ROCm 融合：AR+RMSNorm+FP8 量化为单 AITER 算子，处理扇出输出，emit_bf16 复用 bf16 输出。
- #45003 严格模式：删除 1300 行 XML 解析器，委托 xgrammar 结构标签，默认启用严格工具调用，但移除环境变量引发讨论。
- #45171 Chat Completions 重构：创建 HarmonyParser，非流式解析从 serving 剥离，为流式对称重构铺路。
- #35264 Nixl KV 推送：Prefill → Decode 主动推送，基类 / 子类架构，支持异构 TP，TTFT 提升显著。
- #44260 量化契约：QuantizedActivation 接口，kernel 声明量化键，消费端验证后直接使用预量化激活，是手动融合计划基石。
- #45295 Marlin tile 统一：自动零填充 TP 分片导致的形状不匹配，消除多路径崩溃与慢速回退，代码简洁优雅。

6. 后续建议

- MoE 重构收尾：#41184 留下了部分模型（如 DFlash、DeepSeek V4）的 follow-up，建议在下周集中修复并补全文档；同时评估 MoERunnerInterface 是否可进一步统一为 AutoMoERunner。
- 推测解码多模式整合：动态 SD、DFlash、AutoRegressive 三种模式正在并存，建议召开设计会议统一配置接口和 CUDA Graph 兼容策略。
- ROCm / XPU 测试覆盖：为 ROCm 新增的多个融合 kernel 和量化路径补充跨 batch size 的压力测试，特别是涉及 TP>1 的场景。XPU 侧需跟进上游 API 变化（如 FlashAttention 接口）。
- 前端解析器迁移：HarmonyParser 和 MistralParser 已就位，建议尽快完成剩余模型（如 Anthropic, DeepSeek R1）的 Parser 化迁移，彻底淘汰 serving.py 中的内联解析代码。
- KV 传输可观测性：#35669 的 Telemetry 接口已合并，建议团队为 Nixl 推送模式和异步查找模块添加对应的指标，形成统一的监控 dashboard。
- 量化融合路线图：基于 #44260 的契约，#43224 的手动融合计划应尽快进入第二期——将 FP8、INT4、W4A16 等通用融合 kernel 接入契约，并制定 kernel 选型回退策略。