vLLM 项目 2026 年第 19 周周报（05/04 - 05/10）

vllm-project/vllm

周期：2026-05-04 至 2026-05-10

来源 PR：198・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/vllm-project/vllm/reports/2026-05-04-to-2026-05-10

执行摘要

本周 vLLM 项目继续保持高速迭代，共合并 198 个 PR，其中 24 个被标记为重点 PR。核心趋势包括：模型基础设施标准化（AutoWeightsLoader大规模推广）、CPU后端能力跃升（FP8、GDN 注意力）、ROCm 持续性能优化、推测解码模型生态扩展，以及 KV 传输与 offloading 体系重构。与此同时，31 个 PR 涉及核心路径变更，29 个 PR 缺少测试覆盖，提示团队需要在快速推进功能的同时加强质量保障。

本周重点变化

1. AutoWeightsLoader 迁移进入深水区

继上周多家试点后，本周 DeepSeekV2（#41706）、AXK1（#41901）、CohereMoe（#41690）、Plamo2（#41699）等模型均完成迁移。该系列改动将 load_weights 从 ForCausalLM 下沉至 Model 类，通过统一的 AutoWeightsLoader 委托，大幅降低重复代码。但也暴露出 PP 环境下 rank 不含 MoE 层时 num_redundant_experts 计算不安全等问题，已在讨论中修补。

2. CPU 后端迎来爆发式增强

- FP8 量化：新增 W8A16 块量化线性层（#41186）和 MoE 内核（#41314），依赖 AMX 指令集实现显著加速。
- Gated DeltaNet 注意力：纯 PyTorch 实现 CPU GDN 算子（#41025），支持 Qwen3.5/3.6 混合模型，通过 GSM8K 精度验证。
- 内核同步升级：从 SGLang 同步最新 CPU 内核（#41924），涵盖 INT4/FLA/ 卷积加速，并统一了量化枚举。
- RISC-V 支持：自动绑定 OMP 线程（#40569），非 x86 平台 build 修复（#40575）。

3. ROCm 生态纵深优化

- 融合共享专家（#39280）为 Qwen3-Next 带来 16-24% 解码吞吐提升，设计上通过独立路由器类避免条件膨胀。
- GDN Triton 内核融合（#40711）将多个 kernel 合并，HPU 路径分离，解码吞吐提升 5-8%。
- 依赖升级：AITER 升至 v0.1.13-rc5（#42113），修复 MoE 权重 shuffle 标记丢失（#42061）、allow_allreduce 和 RMSNorm 融合修复（#41972）。

- 此外还修复了 MLA prefill scale 计算（#41569）、TP4 AITER MLA 头数限制（#41835）等 bug。

4. 推测解码新模型与架构清理

- 新增 Cohere EAGLE 草稿模型（#42078），基于融合输入嵌入与目标隐藏状态的设计。
- 新增 Gemma4 MTP 推测解码（#41745），引入 centroids masking 大幅降低 lm_head 计算量，H100 上最高 319% 加速。
- 删除未完全支持的 Tree Attention 后端（#42121），为注意力后端重构扫清障碍，减少约 1400 行代码。
- MiMo-V2.5 也获得 MTP 支持（#41905），但逻辑仍不完整。

5. KV 传输体系重构与可观测性

- NIXL 重构第三阶段（#40731）：引入 EngineTransferPlan/RegionPlan 数据结构，将传输几何计算预生成，热路径不再包含 Dense/Mamba 分支。同时将 TP 映射逻辑提取到独立模块。
- Mooncake 新增传输监控（#40414）：MooncakeKVConnectorStats 记录时长、字节、失败次数等，通过日志输出，并设计锁机制保证并发安全。
- OffloadingConnector 修复 DCP/PCP 下的块大小计算（#41549），DecodeBenchConnector 加入 SupportsHMA（#41770）。
- 移除了对旧版构造函数（pre-v0.12.0）的兼容支持（#39832），属于 breaking change。

6. 性能与可观测性提升

- Helion 配置解析优化（#40850）：用结构化 CaseKey 替换字符串正则，80000 次调用从 1289μs 降至 1.8μs，提速 719 倍。
- Triton JIT 编译监控（#40137）：在 warmup 后注册 hook，一旦推理时发生意外编译立即报警，帮助定位 warmup 遗漏。
- 消除 GPU-CPU 同步：注意力后端（#41434）和 pooler（#41433）中移除不必要的同步点，提升吞吐。
- MoE 路由重放替换为设备缓存（#39917），正确支持 CUDA Graph 和多节点部署。
- 其他优化包括：safetensors 预取参数可配置（#41499）、embedding 序列化零拷贝（#41681）、UniProcExecutor 移除多余线程（#40891）等。

7. 工具调用解析器持续进化

- 新增 LFM2/2.5 解析器（#39243），基于 sentinel token + AST 解析，修复流式边缘情况。
- 升级 xgrammar 至 0.2.0（#40894），引入 structural tags 严格工具调用，先默认关闭。
- 修复 DeepSeekV32/v4（#41801）、Gemma4（#41991）、GLM（#42026）、Mistral（#41730）等多个解析器的 bug，提升流式稳定性。

模块与主题趋势

- 模型加载标准化加速：AutoWeightsLoader 正在成为 vLLM 模型的标准加载方式，预期未来所有模型都会迁移。这要求开发者在新增模型时直接采用该模式。

- CPU 后端成为第二梯队核心：随着 FP8、GDN、INT4 等高级量化与算子的加入，CPU 推理能力大幅提升，尤其适合 Qwen3.5/3.6 等混合模型。但测试覆盖仍显不足，多数新内核仅有单元测试。
- ROCm 与 NVIDIA 并行优化：ROCm 团队每周都有大量 PR，优化点集中在 AITER 内核融合和 MLA 支持。部分优化（如 FSE）已反向参考 DeepSeek 的实现。
- 推测解码进入百花齐放阶段：本周新增两个完整草稿模型（Cohere EAGLE、Gemma4 MTP），清理了一个半成品（Tree Attention），表明团队正积极扩展推测解码生态。需关注测试覆盖和稳定性。
- KV 传输层抽象化：NIXL 的 plan-based 设计和 Mooncake 的 stats 面板，标志着 KV 传输从功能实现走向可维护性优化。
- CI 测试基础设施改进：大量 CI 配置优化（缩小依赖范围、自动发布镜像、测试装饰器修复）表明团队正在提升工程效率，但“缺少测试覆盖”仍是高频风险标签。

风险观察

- 核心路径变更频繁：31 个 PR 标记为“核心路径变更”，其中包括路由重放、注意力后端、推测解码控制流、KV 传输等关键模块。建议对这些 PR 的合入进行更严格的 review 和阶段验证。
- 测试覆盖缺口：29 个 PR 被标记“缺少测试覆盖”，新量化后端（NVFP4、MXFP4）和推测解码新模型尤甚。团队应要求在合入前至少补充烟雾测试。
- 依赖外部版本：多个 ROCm 优化强依赖特定 AITER 版本，CPU 优化依赖 SGLang 内核同步，需持续跟踪上游变化。
- Breaking change 累积：weight transfer API 变更（#39212）、旧版 KVConnector 兼容移除（#39832）、Tree Attention 移除（#42121）可能影响未及时升级的用户，需通过 release note 清晰沟通。
- 配置格式兼容性：Helion 配置键优化（#40850）改变了配置存储结构，旧配置需迁移；DeepSeek 相关配置（如 eplb_config）在 PP 下的获取方式正在调整，需确保无遗漏。

重点 PR 速览

1. #42121 删除 Tree Attention 后端：移除未完全支持的树注意力后端及关联推测解码逻辑，减少约 1400 行代码，为注意力后端重构腾出空间。无弃用期直接清除，影响范围可控但需要下游确认。
2. #39917 路由重放替换为设备缓存：用预分配设备缓存和异步 D2H 管道替换基于共享内存的路由重放，彻底支持 CUDA Graph 和多节点部署。API 向后兼容，是 MoE 推理稳定性的重要提升。
3. #39280 ROCm 融合共享专家（FSE）：为 Qwen3-Next 将共享专家融合到 MoE 内核，解码吞吐提升 16%-24%。设计上采用独立路由器类，为后续扩展提供框架。
4. #40850 Helion 配置解析优化：用结构化 CaseKey 替换正则表达式，pick_config 从 1289μs/call 降至 1.8μs/call。展示如何通过数据类型设计根除性能热点。
5. #40731 NIXL plan-based 重构：引入传输计划预生成，消除热路径中的 Dense/Mamba 条件分支，提升可维护性和性能。是 KV 传输体系成熟化的关键一步。

6#40137Triton JIT编译监控：在warmup后检测意外JIT编译，记录warning级别日志。帮助团队系统发现 warmup 遗漏，推动推理阶段零编译的目标。

7. #41745 Gemma4 MTP 推测解码：新增基于多 token 预测的轻量级辅助模型，引入 centroids masking 减少计算量，H100 上端到端加速 319%。是推测解码性能优化的重要标杆。
8. #41882 NVFP4 all-gather GEMM 融合：针对 NVFP4 量化模型在 Sequence Parallelism + AsyncTP 下融合 all-gather 与 GEMM，长序列吞吐提升 13.5%，依赖 FlashInfer，仅 Blackwell 支持。

后续建议

- 加强测试覆盖：针对本周新增的多个新量化路径（NVFP4、MXFP4、CPU FP8）和完善推测解码模型（Cohere EAGLE、Gemma4 MTP），尽快补充自动化端到端测试和回归测试。
- 推进 AutoWeightsLoader 迁移完成：后续新模型直接采用 AutoWeightsLoader 模式，并考虑编写迁移指南帮助社区贡献者。
- 监控 Breaking Change 影响：对 weight transfer API 变更和旧连接器兼容移除，确保文档和示例及时更新，并在 release 中突出标明。
- 继续性能优化主线：GPU-CPU 同步消除、路由重放、配置解析优化等方向已产出显著收益，建议推广到更多模块。Triton JIT 监控可进一步自动化（如 CI 中强制检查）。
- 关注 ROCm 外部依赖风险：与 AMD 团队合作推动 AITER 版本的稳定化，减少对 nightly/RC 版本的依赖。