vLLM 2026 第 27 周周报（06.29 - 07.05）

vllm-project/vllm

周期：2026-06-29 至 2026-07-05

来源 PR：246 • 重点 PR：24 • 自动生成

原文链接：http://prhub.com.cn/vllm-project/vllm/reports/2026-06-29-to-2026-07-05

1. 执行摘要

本周共合并 246 个 PR，其中重点 PR 24 个，平均重要性 6.4。项目主线围绕"投机解码生态完善、模型加载架构清理、Rust 前端生产化"三大方向推进。DSpark/TLI/DFlash 等四种投机解码方法同期落地，极大丰富了草稿模型选择；删除 63 个模型文件中重复的 load_weights 方法，净减约 2750 行代码，是近期最重要的基础设施重构；Rust 前端在 TLS、渲染器、协议类型等方面取得多个里程碑，正逐步接替 Python 前端职能。此外，多模态模型支持扩展至 LLaVA-OneVision-2、Kimi K2 等，注意力后端新增 HPC-Ops 和 CuTeDSL warmup，测试体系向分布式和硬件多样化扩展。

2. 本周重点变化

2.1 投机解码：从单一走向多元化

本周共合并 5 个核心投机解码 PR，基本覆盖了行业主流方案：

- DSpark（#46995）：针对 DeepSeek-V4/Qwen3 的半自回归推测解码，复用稀疏 MLA 索引实现非因果注意力，在 8 卡 B300 上达到 350+ TPS。
- TLI（Token-Level Intersection）（#38174）：允许 target 与 draft 使用完全不同词汇表，通过交集约束实现无损解码，极大放宽草稿模型选择限制。
- DFlash（#46853）：为 Laguna XS.2.1 提供 DFlash 草稿支持，通过继承减少重复代码。
- SWA+DFlash for MiMo（#46104）：将滑动窗口注意力与 DFlash 结合，支持 MiMo 架构。
- 块验证拒绝采样（#46781）：提升拒绝采样接受率，优化推测解码质量。

至此，vLLM 的推测解码已支持 Medusa、MTP、EAGLE、DFlash、DSpark、TLI 等多种方法，成为业界覆盖最广的推理框架之一。

2.2 模型加载：大幅简化与统一

hmellor 主导的系列重构（#47058、#44589）将模型权重加载逻辑从各个模型文件中抽离至 RoutedExperts 和 AutoWeightsLoader。具体成果：

- RoutedExperts 新增 ckpt_names 参数，自动处理融合权重和专家映射；
- WeightsMapper 增强支持 stacked/fused 权重映射，shard_id 通过 tensor 属性传递；
- 63 个模型文件删除自定义 load_weights，净减少约 2750 行代码；
- 同时修复了 FP8 在线量化时 bias 初始化顺序的潜伏 bug。

该重构不仅降低维护成本，也为未来支持新的 Checkpoint 格式（如 SafeTensors V2）铺平道路。

2.3 Rust 前端：生产化关键一步

本周 Rust 前端共合并 12 个 PR，覆盖：

- HTTPS/mTLS（#45890）：基于 OpenSSL 实现 TLS 终结，支持证书链、mTLS、自定义密码套件，默认地址改为 127.0.0.1 提升安全。
- Harmony 渲染器（#46800）：为 GPT-OSS 模型提供原生 Rust 渲染，绕过 Jinja 路径。
- 枚举域类型（#47283）：将 EngineCoreOutputs 等协议类型重构为枚举，提升类型安全。
- 调度器统计（#47435）：补齐与 Python 前端对标的关键指标。
- profiler 路由（#46306）：暴露 profiling 控制接口。
- 重复检测采样参数（#46684）、测试提速（#47523）等。

Rust 前端在核心协议、安全、监控上的健全性显著提升，具备替代 Python 前端的基础能力。

2.4 注意力后端生态丰富

- HPC-Ops（#46020）：接入腾讯 HPC-Ops 库，提供 FP8 融合注意力，当前面向 Hy3-FP8 模型。
- CuTeDSL Warmup（#46182）：引入可扩展 warmup 框架，提供 FA4 MLA 预编译，避免首次推理 JIT 延迟。
- DCP+FlashInfer MLA 稀疏（#46076）：为 GLM5.2 实现解码上下文并行的稀疏注意力，通过 CuTeDSL 内核完成跨 rank top-K 合并。
- FlashInfer MLA LSE 修复（#47079、#47074）：修正 DCP 合并中 LSE 基数不匹配和 workspace 不足问题。

注意后端种类增多需要用户了解配置，框架层应逐步提供自动选择逻辑。

2.5 多模态与工具解析

- LLaVA-OneVision-2（#44785）：新增多模态模型，支持图像 AnyRes、视频三种预处理后端，约 2266 行模型代码。
- DeepSeek V4 流式解析（#45877）：移植至 ParserEngine 统一框架，修复 DSML 标签泄漏等 bug。
- Kimi K2 流式解析（#46610）：同样基于 ParserEngine 重构。
- Unlimited-OCR R-SWA（#47102）：新增 Triton 注意力后端。
- Harmony Responses API 重构（#47185）：合并上下文类，修复流式 bug。

3. 模块与主题趋势

3.1 v1 路径占比超 70%

本周 PR 中 v1 标签出现 75 次，涉及投机解码、注意力后端、模型运行器等。v1 架构进一步成熟，ModelRunner V2（MRV2）默认启用密集模型（#44443），并修复多项 MRV2 特有 bug（调度槽计数、Mamba2 崩溃、CUDA Graph 填充等）。v1 将成为下一阶段主要技术栈。

3.2 Bugfix 与稳定性优先

bugfix 标签高达 120 个，覆盖流式工具解析（Harmony/Kimi/Poolside）、投机解码（索引溢出、CUDA Graph 脏数据、hidden-states 缺失）、量化（W8A8 选择回归、NVFP4 buffer 零初始化）、平台（ROCm CPU 内存采样、xpu shutdown 泄漏）等。项目处于快速迭代但注重质量的状态。

3.3 平台适配：ROCm、XPU、CPU 持续跟进

ROCm CI 作业扩展至 MI300，新增稀疏 MLA、EPLB、DeepEP 修复；XPU 新增 W8A8 FP8 线性 kernel、LoRA 对齐、shutdown 优化；CPU 新增 tanh AOR GELU 加速、W8A8 MoE VNNI 支持。跨硬件一致性仍是挑战，相关 PR 多为平台特定修复。

3.4 开源协作模式：社区贡献活跃

hmellor（12 PR）、BugenZhao（10）、njhill（10）等活跃贡献。外部作者如 bbrowning（DeepSeek V4 解析器）、wan-danfeng（TLI）、benchislett（DSpark）等带来重大功能。code review 由核心成员把关，同时 Claude/Gemini bot 辅助，但部分 PR 缺乏实质性讨论。

4. 风险观察

风险类别	内容	相关 PR
核心路径变更缺失测试	39 个标签为“核心路径变更”，其中 33 个同时标记“缺少测试覆盖”，重构和功能新增后验证手段不足	#47058, #44589, #46995, #38174
投机解码 CUDA Graph 稳定性	DSpark/DFlash 等新路径的 full CUDA Graphs 支持刚上线，动态循环和填充边界尚未充分检验	#46995, #45953
注意力后端碎片化	4 种 MLA 后端、5 种通用后端共存，不同后端间的性能差异和选择复杂度增加，缺少官方基准	#46020, #46076, #46182
Rust 前端安全配置	TLS 超时硬编码、客户端证书撤销检查默认关闭、gRPC 共享 TLS 可能引入攻击面	#45890, #47101
PD 解耦 P2P 次级层	NIXL 缺失崩溃、write_blocks 失败挂起、线程不安全等问题未完全修复	#42285
大规模重构回归	模型加载重构 #47058 影响了 63 个文件，虽 CI 通过但缺乏全覆盖测试	#47058, #47151（修复）
外部依赖升级	huggingface-hub 升级（#47551）、FlashInfer 0.6.13（#46683）、DeepGEMM tag 更新（#47304）可能引入兼容问题	各 PR

5. 重点 PR 速览

PR	标题	重要性	模块	作者	要点
#46995	DSpark 半自回归推测解码	9.4	spec-dec	benchislett	DeepSeek-V4/Qwen3 上的新投机方法，复用稀疏 MLA 非因果注意力
#44353	权重同步重构	9.4	infra	hao-aaron	抽取稀疏 NCCL 引擎，移除 is_checkpoint_format 参数
#45877	DeepSeek V4 流式解析	9.3	frontend	bbrowning	移植至 ParserEngine，修复 DSML 标签泄漏等 bug
#44785	新增 LLaVA-OneVision-2	9.2	multi-modality	chengzheng345	支持图像 / 视频，AnyRes 分块，三种视频后端
#47283	Rust 枚举域类型	9.2	rust	BugenZhao	EngineCoreOutputs 等协议类型安全重构
#47185	Harmony Responses API 重构	9.2	frontend	yzong-rh	合并上下文类，修复流式 done-event 丢失
#46610	Kimi K2 流式解析	9.2	frontend	chaunceyjiang	基于 ParserEngine 的新解析器，支持推理和工具调用
#42285	PD 解耦 P2P 次级层	9.4	kv-connector	liranschour	ZMQ+NIXL 传输，会话协议设计
#47058	移除 63 个 load_weights	9.1	refactor	hmellor	统一由 RoutedExperts 加载，净减 2750 行
#46703	NCCL 对称内存扩展	9.1	performance	WoosukKwon	AllGather/ReduceScatter 支持 NVLS

PR	标题	重要性	模块	作者	要点
#46853	Laguna DFlash drafter	9.0	spec-dec	adamkbaranowski	继承最小化实现，展示复用模式
#45890	Rust HTTPS/mTLS	9.2	rust	tahsintunan	OpenSSL 实现，安全加固
#46076	DCP 稀疏注意力	9.4	attention	ZJY0516	GLM5.2 DCP，CuTeDSL 内核
#46182	CuTeDSL warmup	9.2	attention	LopezCastroRoberto	热启动框架，FA4 MLA 预编译
#42406	Mamba hybrid 前缀缓存	9.2	model-runner	izhuhaoran	V2 支持，pre-copy 方法
#43373	Humming MoE 标准化	9.2	moe	bnellnm	与 modular kernel 对齐
#44512	scale-out 入口整合	9.2	frontend	noooop	分离 render/derender API
#46800	Rust Harmony 渲染器	9.0	rust	BugenZhao	GPT-OSS 原生渲染
#42920	CPU W8A8 MoE	9.0	cpu	yuwenzho	VNNI 指令支持，后端注册

6. 后续建议

1. 补充投机解码测试：DSpark、TLI、DFlash 等新算法应及时补充 CI 端的端到端准确性和性能测试，特别是多 GPU、CUDA Graph 场景。
2. 制定注意力后端选择指南：为减少用户困惑，建议在文档中明确不同后端的适用模型、硬件和量化条件，并考虑提供自动选择逻辑。

3. 监控模型加载重构回归：hmellor 的重构虽已合并，但仍可能影响未被 CI 覆盖的模型。建议在下一轮发布前执行手动回归测试。
4. 推动 Rust 前端功能完备：鼓励贡献者补齐与 Python 前端的功能差异（如 streaming JSON mode、多模态输入等），同时跟进安全审计。
5. 关注 PD 解耦 P2P 层稳定性：该 PR 设计复杂，需在分布式测试中重点验证崩溃恢复和超时处理。
6. 提升 PR review 质量：部分重要 PR 缺乏充分讨论，建议对高影响变更（核心路径变更、模型新增）要求至少两位核心贡献者 review，并鼓励测试覆盖率要求。