sgl-project/sglang 2026 第 20 周 (05-11 至 05-17) 周报

sgl-project/sglang

周期：2026-05-11 至 2026-05-17

来源 PR：295 • 重点 PR：24 • 自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-05-11-to-2026-05-17

执行摘要

2026 年第 20 周（05-11 至 05-17），sgl-project/sglang 仓库共合并 295 个 PR，其中 24 个重点 PR 获得详细分析。本周代码库变化丰富，主要围绕 MoE 架构统一、DeepSeek V4 功能加速落地、推测解码性能优化、扩散模型内存与量化升级，以及CI基础设施大规模重构展开。整体表现为深度推理引擎向多模型、多平台、高性能方向持续演进。

本周重点变化

MoE 架构统一化

- 通过 #23760 完成 AITER MoeRunner 管道化，统一 DeepEP 与 MoriEP 的调用路径，移除废弃类。这是 MoE 后端抽象的重要里程碑，未来可方便地扩展新后端。
- 19290 为 DeepSeek 共享专家引入 Waterfill 负载均衡，提升 EP 吞吐约 4%，并通过 Triton 内核消除 CPU-GPU 同步。
- 24816 集成 FlashInfer SM90 cutlass MXFP4 MoE 后端，在 Hopper GPU 上实现 24-36% prefill 加速。

DeepSeek V4 功能井喷

- 多个 PR 为 DeepSeek V4 添加重要能力：HiCache 集成（#24691）将其 KV 缓存扩展到 CPU 内存；KV 压缩 V2（#24890）引入 fused norm+RoPE 内核；新增 PP/PD 分布式支持（#24704）；支持 w4a4 MegaMoE（#25052）及共享专家 TP1（#24949）。

推测解码性能提升

- 25333 实现 MLA chunked-prefill 融合 kernel，最高 10× 加速；#25311 优化 MLA KV 缓存写入（TMA bulk-store）最高 12× 加速；#24925 集成 tokenspeed_mla Blackwell 后端；#25001 支持 MLA 注意力的 LoRA 适配器。

扩散模型内存与量化

- 24593 将 layerwise offload 泛化至所有组件，用户可精细控制卸载对象；#24491 引入性能模式自动选择（auto/speed/memory），降低调优成本；#25457 新增内存感知加载排序，减少 OOM 风险；#21431 为 AMD GPU 添加在线 MXFP4/FP8 量化。

基础设施抽象

- 24096 引入 CudaDeviceMixin 平台抽象层，为多硬件支持铺路；#24723 将 ModelExpress 加载委托给外部包，减少 500+ 行代码；#25050 添加 model-config-parser 注册表，支持插件式配置格式；#23449 添加 Apple Metal kernel 构建支持。

CI 升级

- 25387 将 PR awareness 独立为 workflow 降低权限扩散；#25263 基于实时数据动态划分测试分区；#25420 规范化 CI 阶段命名；#25264 将 runner 配置中心化；#25468 发布脚本更新支持幂等。

模块与主题趋势

从标签统计看，本周热点集中在 bugfix（100）、infra（99）、refactor（71）和 performance（64），表明团队在快速迭代的同时高度重视质量与架构。deepseek（42）和 speculative-decoding（28）是模型相关的核心关注点。diffusion 相关 PR 数量可观（多位贡献者提交），且多聚焦于性能与内存优化。CI 标签 59 表明本周基础设施升级力度极大，超过一半的 PR 实际为 CI 配置或重构。

风险观察

1. 测试覆盖不足是最大隐患：53 个 PR 被标记为“缺少测试覆盖”，许多新功能如 DSV4 HiCache、扩散性能模式自动调优等缺乏足够的测试保护，合并后可能出现回归。
2. 核心路径变更频繁（43 项）：MoE 路径、调度器、缓存层等核心模块几乎每周都在调整，可能引入难以诊断的性能退化。
3. 特定功能依赖性强：DSV4 HiCache 的 Sidecar 池仅支持 KV 来源（#24691 已知限制），当前若状态池路径被触发将导致崩溃；扩散模型 GPU 内存检测尚未完全可靠（#24491）。
4. 外部依赖兼容性：本周经历 FlashInfer 升级（0.6.11）后又因 MoE 崩溃回退（#25310），依赖新版本时需更充分验证。
5. 大规模重构的等价性验证：fzyzcjy 提交了一系列大规模重构（如 ParallelState 封装、属性清理），虽风险标注低，但涉及文件众多，可能隐藏小错误。

重点 PR 速览

- #23760：MoE 统一 DeepEP/MoriEP 调用路径，通过 AiterRunnerCore 实现，是 MoE 后端抽象的典范。
- #24593：扩散模型 layerwise offload 泛化至所有组件，新增组件名选择器和参数冲突自动替换。
- #25333：为 DeepSeek MLA chunked-prefill 融合 cat+FP8 量化核，混合调度器自动选择变体，最高 10× 加速。
- #19290：共享专家 waterfill 负载均衡，MMLU 准确率无损，端到端吞吐提升 4%，Triton 内核消除同步。

- #24691：DeepSeek V4 HiCache 支持，引入 LogicalHostPool 和 sidecar 池机制，降低 GPU 显存压力。
- #24096：CudaDeviceMixin 与 CudaSRTPlatform 平台抽象，为多硬件支持奠定基础，ROCm 继承 CUDA 设计。
- #25050：model-config-parser 注册表，支持 HF AutoConfig 和 Mistral 原生格式，便于扩展自定义配置格式。
- #25424：推测解码参数处理抽离为独立 hook 文件，并采用 AST 等价验证确保重构安全。

后续建议

- 优先补齐测试：对于缺少测试覆盖的重点功能（DSV4 HiCache、扩散自动调优、MLA LoRA 等），应在下一阶段补充单元与集成测试。
- 规范核心路径变更流程：核心模块（调度器、缓存、MoE 层）的每一次修改应附上性能基准，并触发额外 review 环节。
- 关注 DSV4 HiCache 状态池限制：尽快实现 Sidecar 解析的多源支持，避免用户踩坑。
- 扩散模型调优迭代：性能模式自动调优的 GPU 内存检测精度、layerwise offload 默认策略等尚需完善，建议收集用户反馈。
- CI 稳定性守护：新增的 PR awareness 和动态分区机制需持续监控，确保不因误判导致测试遗漏。