sgl-project/sglang 2026 第 25 周 (06-15 至 06-21) 周报

sgl-project/sglang

周期：2026-06-15 至 2026-06-21

来源 PR：281・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-06-15-to-2026-06-21

执行摘要

本周共合并 281 个 PR，其中 24 个高亮 PR。主要变化集中在 runner 层架构重构、上下文并行策略引入、AMD 性能优化、DeepSeek-V4 生态拓展以及推测解码改进。同时，基础设施和代码质量也有显著提升，但大量 PR 存在测试覆盖不足的问题，需关注回归风险。

本周重点变化

Runner 层多态重构

ch-wan 主导的一系列 PR 将 EagerRunner 提取为独立类（#28386），统一 warmup 生命周期（#28739），并引入 get_parallel() 统一访问并行拓扑状态（#28567）。这些重构使代码更具可维护性，并为后续支持更灵活的模型执行路径奠定了基础。

上下文并行（CP）v2

Fridge003 实现了 zigzag CP 策略（#28421），支持 Qwen3 和 FlashAttention，通过环境变量 SGLANG_ENABLE_CP_V2 启用。同时定义了 CP 策略抽象框架（#27313），包括元数据类、策略基类和环境门控，为未来扩展提供规范。

AMD 性能突破

多个 PR 在 AMD 平台上取得显著提速：HiSparse 稀疏注意力（#26639）、Split-KV flash-decode attention for EAGLE（#27382，内核加速 11x，端到端吞吐提升 30.6%）、Marlin 后端用于 SM120 MXFP4 MoE（#28231，吞吐提升 ~2x）以及 breakable CUDA graph 支持（#28173）。这些改进使 AMD 平台的竞争力大幅增强。

DeepSeek-V4 生态扩展

本周在 DeepSeek-V4 上投入较多：Ascend NPU 端到端支持（#25144），包括专用注意力后端、分页内存池和模型适配；在线压缩支持 MTP（#26471），吞吐提升约 280%；混合精度压缩状态（#27277），支持 BF16 降低内存占用。

Speculative Decoding 优化

引入拒绝采样（#26312）提升接受长度和吞吐；修复 grammar 兼容性（#24082、#28682），统一 token 接受路径；自适应步长和 draft KV 缓存更新（#23994）；修复多模态下的崩溃问题（#27863）。

基础设施与质量

CLI 参数自动生成（#28814）消除重复定义；CI 清理迁移至注册制测试（#28810）；Ray 观测性包装器（#26252）支持集成到 Ray 监控系统；CUDA graph 捕获跟踪导出（#28551）便于调试；共享内存泄漏清理（#28089）提升 CI 稳定性。

模块与主题趋势

- bugfix 和 performance是最主要的标签（101 和 73），表明团队在快速迭代中同时注重稳定性和效率。
- refactor 和 infra频繁出现（56 和 55），反映代码架构持续演进，为未来功能扩展做准备。
- amd 和 deepseek相关 PR 数量突出（43 和 29），表明这两个领域是本周投入重点。
- 热门文件server_args.py（10 次）和 model_runner.py（9 次）是配置和模型执行的核心，频繁修改需谨慎。
- top_risks显示“缺少测试覆盖”和“核心路径变更”出现次数最多（47 和 34），提示我们需要补强测试。

风险观察

- 多项核心路径变更在 review 中暴露了潜在问题但并未完全解决，例如 #28739 的 pp_proxy_topk_size 缺失，以及 #28386 的静态缓冲区尺寸估算不足。
- 大量 PR 标注为缺少测试覆盖（47 次），尤其是 AMD 和 NPU 平台的新功能，回归风险较高。
- 配置向后兼容性变化（如环境变量名称修改、参数重命名）可能影响用户升级，需在文档中明确说明。
- 实验性功能（如在线压缩、LPLB 负载均衡）依赖特定硬件或库，稳定性待观察。

重点 PR 速览

- #28755Cap SWA pool sizing with chunk cache：按活跃请求数上限收紧 SWA KV Cache 池，释放内存至 full pool。
- #27273mem_cache 重构（5/N）：提取 host KV cache 基类到 pool_host 子包。
- #28744sgl-router 入口统一 tokenize：消除引擎侧重复 tokenization，最高降低 41% TTFT。
- #28739统一 kernel warmup 至 BaseRunner 生命周期：重构 warmup 流程，但遗留 pp_proxy_topk_size 问题。
- #28386EagerRunner 独立：实现多态分发，分离 eager 与 CUDA graph 路径。
- #26639[AMD] 启用 HiSparse 稀疏注意力：支持 DeepSeek V3.2 和 GLM-5，通过 JIT 内核适配 wavefront64。
- #28231SM120 MXFP4 MoE 改用 Marlin 后端：吞吐提升 ~2x，删除旧 Triton kernel。
- #27382[AMD] Split-KV flash-decode attention for EAGLE：内核加速 11x，端到端吞吐提升 30.6%。

- #28421实现 zigzag CP 策略：支持 Qwen3 与 FlashAttention，通过 SGLANG_ENABLE_CP_V2 启用。
- #25144[NPU] 为 DeepSeek-V4 添加 Ascend NPU 支持：包括专用注意力后端、分页池和模型适配。
- #28185int8 检查点池：双池分离设计，线性注意状态缓存容量翻倍。
- #28567引入 get_parallel()：统一访问并行拓扑状态，替换 180+ 个模型文件中的分散调用。
- #26252添加 Ray 指标后端包装器：支持将引擎指标路由到 Ray 监控系统。
- #24515LPLB 线性规划负载均衡：每层在线求解 LP 优化 MoE 专家分配，GSM8K 吞吐提升 5.4%。
- #26471DeepSeek-V4 在线压缩支持 MTP：吞吐提升约 280%，精度无损。
- #21631重构 HiCache 写回内核：引入分阶段 page_first 写回，提升大序列性能。
- #28265[AMD] 移除稀疏 MLA 解码 2-phase fallback：强制融合 kernel，ITL 改善 63-66%。
- #27980新增 worker 协调循环：修复空 model_ids 无法恢复，提升路由器可用性。

后续建议

- 加强测试覆盖：对于核心路径变更（runner、CP）和平台特定功能（AMD、NPU），应要求增加单元测试和集成测试。
- 跟进未解决问题：对 review 中暴露的 bug（如 pp_proxy_topk_size、缓冲区尺寸）应及时修复并合并。
- 关注性能回归：建议在 CI 中添加性能对比测试，及时发现回归。
- 继续架构演进：runner 多态、CP 策略框架等重构为后续功能提供了良好基础，应持续推动。
- 文档与配置兼容性：废弃环境变量和参数重命名时，提供迁移指南和兼容层。