SGLang 周报 2026 第 21 周 · 05-18 至 05-24

sgl-project/sglang

周期：2026-05-18 至 2026-05-24

来源 PR：301 · 重点 PR：24 · 自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-05-18-to-2026-05-24

1. 执行摘要

本周仓库共合并 301 个 PR，其中高亮 PR 24 个，工作量密集。最显著的变化主线是调度器（Scheduler）的模块化重构——由 fzyzcjy 主导，超过 30 个 PR 将原先集中在 Mixin 中的逻辑拆解为独立组件（如输出流、指标上报、权重更新、不变量检查等），旨在大幅降低继承复杂度并提升可测性。与此同时，投机解码（Speculative Decoding）的性能优化与调度器内 FutureMap 的全面应用（hnyls2002 系列）是本周期另一大重点。硬件支持方面，AMD、NPU、XPU 的 DeepSeek V4 与扩散模型适配持续推进，性能优化（token id 存储、FP8 GEMM）也有重要产出。需警惕的是，本次大规模重构普遍缺乏测试覆盖，且有一例 CUDA graph 统一重构因引入回归而被回退，整体风险偏高。

2. 本周重点变化

- 调度器重构（Scheduler Decoupling）：fzyzcjy 将 Scheduler 中的多个 Mixin（如 SchedulerMetricsMixin、SchedulerUpdateWeightsMixin、SchedulerOutputProcessorMixin 等）逐一迁移到 scheduler_components/ 下的独立类中。典型步骤是：1）新建组件类并迁移状态与静态方法；2）修改 Scheduler 的调用点；3）删除原 Mixin 文件。此系列 PR 涉及 30+ 文件更改，但行为保持向后兼容。重要子项包括指标上报（#25629-#25631）、输出流（#25634-#25635）、日志概率处理（#25632-#25633）、不变量检查（#25623-#25624）、KV 事件发布（#25625-#25626）、负载查询（#25627-#25628）、权重更新（#25616）等。
- 投机解码优化（Spec Overhaul）：hnyls2002 围绕 FutureMap 结构进行了系统性改造。将 seq_lens、output_tokens_buf 等关键数据通过 FutureMap 传递，消除不必要的同步等待（#25879、#25922、#25944、#26020）。此外，支持可中断 CUDA 图（#25795）、统一验证输出（#25566）、修复 EAGLE 相关 bug（#25454、#26126）等。这些改进了 spec v2 的吞吐和延迟表现。
- DeepSeek V4 全面铺开：涉及 AMD 平台（#25898、#24933）、NPU 注意力后端（#23482）、CUDA graph 分段支持（#23351）、MTP Context Parallel（#24934）、JIT kernel 模块化重构（#25884）、仓库预热（#25810）、缓存失效修复（#25889）等。同时完成了 NSA→DSA 的大规模重命名（#25821），引入向后兼容别名。
- 性能提升：Req token-id 存储从 list[int] 改为 array('q')（#25098），降低张量构建开销；SM90 FP8 GEMM 引入 swap-AB 调度（#25532），小 batch 解码 kernel 几何平均加速 1.156x；ViT 编码优化（#25910）；LoRA MoE 去除 GPU 同步瓶颈（#25531）。
- 扩散模型：角色感知组件加载（#25168）为分解部署奠定基础；MXFP4 量化（#22338）、GLM-Image 并行生成（#25645）、FLUX.2 支持（#25661）等。

3. 模块与主题趋势

3.1 调度器架构从 Mixin 到组件化

- 这是继上一周之后延续的大动作。fzyzcjy 采用"先建组件、再迁移、后删除"的模式，每个 PR 职责清晰。当前 scheduler_components/ 已包含 output_streamer、metrics_reporter、batch_result_processor、kv_events_publisher、invariant_checker 等 10 余个模块。该趋势降低了单个文件复杂度，但引入大量新类，文档与测试配套需跟上。

3.2 投机解码管道持续打磨

- FutureMap 成为连接 Draft/Verify/Scheduler 的核心通信原语，串起 seq_lens、token、output 等数据流。本周进一步合并重构（#26085 移除 FutureIndices 包装类），整体设计趋于成熟。此外，Eagle 可中断 CUDA 图（#25795）与 trtllm MHA overlap（#25925）释放了更多并行度。

3.3 多硬件适配加速

- AMD：DSV4 运行时修复、AITER 升级、CI 超时调整、测试种子确定性。
- NPU：扩散注意力后端、MXFP4 量化、DeepSeek-OCR 模型、文档最佳实践。
- XPU：Qwen3.5 支持、triton-xpu 升级、fused_topk 适配。
- Intel GPU：Diffusion 基础 kernel、CPU bug 修复。
- 新增平台（MUSA、Blackwell）CI 与性能调优也持续活跃。

3.4 文档与 CI 改进

- 多位贡献者更新了 NPU、AMD、Diffusion 等文档。CI 方面：bot-cherry-pick 工作流大幅改进（支持 PR 编号、状态检测、权限分离），pr-test-extra 触发优化，测试套件引入 stage-a 合理性套件（#25831）。cutlass-dsl 版本经历升级→回退→再修复，反映了外部依赖管理的挑战。

4. 风险观察

风险项	相关 PR	说明
缺少测试覆盖	#25635, #25630, #25628 等数十个重构 PR	调度器重构几乎未附带新测试，仅依赖现有集成测试。任何重构引入的回归都有可能漏到生产环境。
核心路径变更回退	#26166（回退 #26134）	CUDA graph 统一重构因 FlashInfer use_ragged 不一致等 bug 被回退，说明注意力后端的统一需更稳健的测试与渐进式合并策略。
第三方依赖版本兼容	#25938（cutlass-dsl 4.5.1→4.5.0）、#25576、#23809	NPU、XPU 依赖的 sgl-kernel-npu、triton-xpu 版本频繁变化，可能导致 CI 不稳定或运行时错误。

风险项	相关 PR	说明
未修复的 review 问题	#25884（torch.mm out_dtype 不支持、缺失 @cache_once）	尽管 PR 已合并，但 reviewer 指出的可能 bug 尚未确认修复，建议后续跟踪。

5. 重点 PR 速览

以下 PR 选取基于重要性、影响范围及趋势代表性：

PR #	标题	作者	关键点
25098	Req token-id 存储迁移至 array('q')	Jialin	34 个文件改动，降低张量构建开销，为 Rust Radix Cache 铺垫。
25532	SM90 FP8 GEMM swap-AB 调度	yuan-luo	sgl-kernel 优化，Llama-3.1 70B decode 吞吐提升 5.8-18.5%。
25284	Gemma4 流水线并行	yuan-luo	支持 –pp-size 2，权重占用降低 48%，KV 缓存翻倍。
25821	NSA→DSA 重命名（带别名）	ch-wan	大规模重命名范例，向后兼容别名设计，涉及 CLI、环境变量、注册表等。
25983	ForwardContext 引入，解耦 ForwardBatch	ch-wan	删除 ForwardBatch 中四个运行时引用，通过 context manager 获取，架构解耦。
25678	废弃 NPU 专有 MoE 路径	ch-wan	统一至 FusedMoE 管道，通过 free function 绕过调度器处理 ascend_fuseep。
25635	输出流移至 SchedulerOutputStreamer	fzyzcjy	调度器重构典范，机械移动 438 行，职责清晰化。
25879	FutureMap 路由 seq_lens，消除 verify_done 等待	hnyls2002	spec v2 性能关键改进，减少 CPU-GPU 同步。

PR #	标题	作者	关键点
25979	PD 后端共享逻辑合并到公共基类	ShangmingCai	消除 Mooncake/Mori/Nixl 后端重复代码，降低维护成本。
25898	AMD DSV4 运行时修复	kkHuang-amd	修复 CompressStatePool、JIT 不兼容等三个问题，启用 HIP 适配。

6. 后续建议

- 强化调度器重构的测试覆盖：当前 merge 了大量无测试的新组件，建议在下周安排专项测试编写，至少覆盖核心路径（如 request ingress、streaming、metrics、weight update）的单元测试。可利用 test/registered/unit/managers/ 目录。
- 跟进未修复的 review 问题：特别是 #25884 中 torch.mm 的 out_dtype 参数不受标准 PyTorch 支持，以及缺失 @cache_once 可能导致性能退化。建议作者或团队提交修复 PR。
- 统筹 CUDA graph 统一重构：虽然在 #26166 中被回退，但其方向（减少后端重复代码）是正确的。应先在单一后端上验证充分，分步推进，并增加针对性测试（如回退中发现的 FlashInfer use_ragged 不一致）。
- 关注 DeepSeek V4 的跨 PR 集成风险：DSV4 相关改动涉及注意力、内存池、调度器、JIT kernel 等多个模块，且与 AMD/NPU 路径重叠。建议设立一个集成测试标签（如 dsv4-integration）在 CI 中每日运行，提前发现交互问题。
- 投入资源辅助 NPU/AMD 平台稳定性：这些平台的测试经常因超时或依赖版本失败（如 AMD 超时放宽 #25978、NPU 日志抑制），建议细化 CI 资源分配并冻结关键依赖版本。