SGLang 2026 第 31 周（07-27 至 08-02）周报

sgl-project/sglang

周期：2026-07-27 至 2026-08-02

来源 PR：258・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-07-27-to-2026-08-02

执行摘要

本周（2026-07-27 至 08-02）仓库合并 258 个 PR，其中重点 PR 24 个，平均重要性 6.68。整体呈现三条主线：一是 Rust server 前端基础设施的多阶段落地，二是 Diffusion 多模态生成能力的集中扩展，三是围绕 KV 缓存、统一内存与 CUDA graph 的性能与稳定性优化。与此同时，配置系统正在经历大规模命名空间迁移，NPU（Ascend）支持也进入密集迭代期。值得注意的是，本周"核心路径变更"风险标记高达 40 个，"缺少测试覆盖"33 个，核心调度、缓存与注意力路径的改动频率值得后续提高评审与测试门槛。

本周重点变化

1. Rust server 架构加速落地

rainj-me 主导的 Rust server 系列在本周集中合并了请求消息层（#32242）、egress 消息层（#32342）、tokenizer/detokenizer/egress 模块（#32872）、tokenizer manager/ring/runtime（#32358）以及主分支 #29799。该方案将 tokenize/detokenize 卸载到 Rust 线程，通过列式 messagepack 传输减少 GIL 与序列化开销。目前由 SGLANG_RUST_SERVER 门控、默认关闭，但已修复默认路径回归（#29799 中 SGLANG_RUST_SERVER=0 启动崩溃）和 prompt logprob None 哨兵问题，后续逐步替换 Python 前端已有清晰路线。

2. Diffusion 多模态能力大幅扩展

mickqian 提交的 MiniMax-H3 支持（#33275）新增约 2.2 万行代码，覆盖 DiT、VAE、denoising 管线、TP/SP/DP 分布式与 OpenAI 兼容 API，是本周体量最大的 PR。此外 #30211 统一 encoder 并行策略（auto/fold/dp/replicate），#31854 为 2-rank Ulysses 引入 CUDA-IPC 零拷贝 all-to-all，将 Qwen-Image 生成延迟从 81.3ms 降至 64.8ms；#32784 加速视频输出终化。Diffusion 模块正在快速走向生产级。

3. 缓存与内存管理持续深挖

Zhangmj0621 的 Session-reference-aware Unified Radix Cache（#29173）为 agentic 多轮场景引入会话感知驱逐，支持会话关闭时保留可复用 KV；ch-wan 的 unified-memory 系列将 MLA-hybrid-Mamba（Kimi-Linear）支持扩展到 Triton 与 paged MLA 后端（#32971、#32972、#33046），并新增 fa3 后端支持；hnyls2002 的 VerifyMask 重构（#32920）将无人读取的掩码压缩为 QLEN_ONLY，长上下文掩码内存从约 1 GiB 降至 4 KiB，并消除每步 271us 的 fill 开销。

4. 推理性能优化向 Blackwell/FP8 延伸

alumkal 将 MiniMax-M3 注意力 GEMM 全面 fp8 化（#30971），prefill 最高提速 66%；JackChuang 的 DSA Q8KV8 FP8 sparse prefill（#31888）在 GLM-5.2/DeepSeek-V3.2 上获得 9.5-17.5% 吞吐提升；Oasis-Git 的 BCG 扩展（#31987）将 breakable CUDA graph 覆盖到 DSA 与 DeepEP a2a；paulzhang-tm 的 FullCG chunked cached-prefix（#30825）将内存增量从 +3.6 GiB 降至 +1.2 GiB。

5. 配置系统大规模重构

ch-wan 主导的 runtime_context 命名空间迁移本周进入尾声，多个 PR（#33011、#33012、#33013、#33170、#33171、#33172）覆盖 160+ 文件，将配置读取迁移到命名空间访问器，并引入角色化 publish、快照恢复与 mutation ratchet。该重构对核心初始化路径影响面大，但配套恢复了 15 个此前跳过的单测文件，并增加了 per-role namespace 强制约束。

6. NPU（Ascend）支持进入密集迭代

TallMessiWu 的 W8A8 MXFP8 量化（#30768）为 Qwen3 MoE 带来约 44% 吞吐提升；Talantan1102 的 DSV4 优化（#31931）加入 PD 分离与 chunked prefill；Sugar920 等新增 16+ NPU 测试用例进入 PR CI，NPU 回归保障体系正在补齐。

模块与主题趋势

从标签看，bugfix（100）和 performance（77）占主导，说明本周以稳定性和性能为主；feature（59）紧随其后，反映新能力仍在快速涌入。热点文件集中在server_args.py（16次），配置项调整贯穿多个 PR，与配置系统重构互相印证。作者方面，mickqian（18）、hnyls2002（15）、ch-wan（14）、BBuf（14）是最活跃的贡献者，分别集中在 diffusion、speculative decoding、配置重构与 kernel 清理。

风险观察

- 核心路径变更 / 测试覆盖不足：40 个 PR 标记核心路径变更，33 个缺少测试覆盖，尤其 DCP、CUDA graph、缓存路径。建议在涉及 schedule_batch、attention backend、mem_cache 的改动上强制要求关键路径测试。
- MiniMax-H3 评审缺口：#33275 无人工 review，2.2 万行代码存在潜在设计盲区，建议安排专门 review 并补充稳定性测试。
- unified-memory FNUZ dtype 问题：chatgpt-codex 标记的 P1（FNUZ 未归一化导致 KV 静默损坏）在 #32971 合并时未解决，需尽快跟进。
- Rust server 回归风险：虽然默认关闭，但已有默认路径回归历史，后续合并需持续验证 SGLANG_RUST_SERVER=0 路径。
- 遗留技术债：DCP 传输的 ZMQ 字段顺序、RuntimeError 崩溃隐患等已在 review 中确认，但推迟到后续 PR，需跟踪。

重点 PR 速览

- #33275 MiniMax-H3：新增完整音视频生成管线，覆盖 DiT/VAE/denoising 与 TP/SP/DP，含 PR-blocking GPU 测试与 cookbook，但缺人工 review，是本周最需后续关注的大规模合入。

- #30177 return_hidden_states="last"：将 hidden states 返回从二态扩展为三态，引入固定 server 级 capture ceiling 取代按需重捕获，解决了 radix cache 下 prefill 偏移语义，并补充 warm-cache 混合模式测试。
- #29173 Session Radix Cache：通过 session_id 引用跟踪与 generation 机制实现会话感知驱逐，为 agentic 多轮 workload 提升命中率；与并行工作统一了 session 生命周期，是缓存路径的重要演进。
- #30971 M3 FP8 Attention：SM100 上 fp8_e4m3 KV + trtllm_mha 自动推导激活，kill switch 设计完善，prefill 最高 66% 加速。
- #31987 BCG 3/N：将 breakable prefill CUDA graph 扩展到 DSA 与 DeepEP，通过 capture stub + barrier 和单一回放资格谓词消除 DP collective 错配隐患。
- #33023 Inkling ShortConv Backend：迁移短卷积到 sidecar 后端，消除每步 4 倍元数据冗余，decode 延迟显著下降，并用 AST 扫描守护跨硬件 hook 签名。
- #31854 CUDA-IPC A2A：2-rank Ulysses 场景用 GPU 侧序列计数器替代 NCCL rendezvous，Qwen-Image 延迟降低 20%，双端退休超时机制避免进程组失同步。
- #30768 NPU MXFP8：为 Qwen3 MoE 增加 W8A8 MXFP8 在线 / 离线量化，融合路由量化减少 kernel 启动，吞吐提升 44%。

后续建议

1. 为 Rust server 系列补充 end-to-end 兼容性测试，尤其是默认关闭路径与列式传输边界。
2. 对 MiniMax-H3 在真实工作负载下做性能和稳定性验证，必要时拆分 review。
3. 跟踪 unified-memory 的 FNUZ dtype 修复，防止 KV 静默损坏扩散。
4. 配置迁移接近完成，建议推进命名空间访问器的全量强制化，并继续扩大 ratchet 覆盖。
5. 对 DCP/PD 传输的遗留协议问题（字段顺序、错误处理）设置时间窗口完成重构。
6. 鉴于 core-path 高频变更，建议在 CI 增加针对 schedule_batch、attention backend、cache 的回归测试门禁。