PR #27046 完整报告

sgl-project/sglang

[HiCache] fix PD L3 cache hit details from decode responses

合并时间：2026-06-04 18:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27046

执行摘要

- 一句话：修复 PD 模式下 *decode* 响应中 L3 缓存命中报告的缺失问题
- 推荐动作：建议快速合并。变更逻辑清晰、改动量小、风险低，且解决了一个实际观测到的监控数据缺失问题。

功能与动机

修复 PD 分离部署模式下 *decode* 响应遗漏 L3 缓存命中详情的问题。在 PD 模式下，L3 命中由 prefill 产生并通过元数据在 *decode* 阶段上报，而 *decode* 节点可能没有本地存储后端，原条件 self.enable_hicache_storage() 导致无法正确报告存储命中数值，影响缓存效果监控和调试。

实现拆解

1. 关键文件：python/sglang/srt/managers/scheduler_components/output_streamer.py
2. 方法调整：修改 get_cached_tokens_details 方法中关于存储命中详情的判断逻辑。
3. 条件拆分：
   - 存储命中数值（storage count）的条件从单一的 self.enable_hicache_storage() 扩大为 req.cached_tokens_storage > 0 or self.enable_hicache_storage()，确保即使 *decode* 节点无存储后端，只要存在 L3 命中数据就予以报告。
   - 存储后端类型（storage_backend）仍只在 *enable_hicache_storage()* 启用时才填充，避免无后端时返回错误信息。

关键文件：

- python/sglang/srt/managers/scheduler_components/output_streamer.py（模块 调度器；类别 source；类型 core-logic）：核心修改文件，调整了 get_cached_tokens_details 方法中存储命中详情的判断逻辑，修复 PD 模式下的 L3 缓存命中报告。

关键符号：get_cached_tokens_details

评论区精华

无实质性讨论。Review 仅包含 Gemini Code Assist 的自动摘要，无人工评论；xiezhq-hermann 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：变更范围极小（+4/-2），仅调整一个条件语句，回归风险低。潜在风险是在 PD 模式下，如果 req.cached_tokens_storage 在非预期场景下被误设，可能多报零值存储命中，但已通过 > 0 检查避免。
- 影响：影响范围仅限于启用 HiCache 且使用 PD 分离部署的模式。不会影响非 PD 模式或未启用 HiCache 的场景。修复后，decode 响应将正确反映 L3 缓存存储命中率，有助于缓存效率监控和问题定位。
- 风险标记：缺失测试覆盖

关联脉络

- PR #26119 [diffusion] Disagg server args, launch helpers, and warmup utils: 都与分离部署相关，但具体模块不同（HiCache vs diffusion），仅作为上下文参考。