

PR #36299 完整报告

sgl-project/sglang

[Weight Cache] Make daemon socket/ready paths configurable via env

合并时间: 2026-08-27 03:38

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36299>

执行摘要

- 一句话: 权重缓存守护进程路径改为环境变量可配置
- 推荐动作: 该 PR 值得快速浏览, 因为展示了如何将硬编码路径改为环境变量驱动的模式, 并附带简单校验。对于维护者, 可关注环境变量命名是否符合全局约定。

功能与动机

PR 描述中明确说明: 'The weight cache daemon socket and ready file paths were hard-coded to /tmp in weight_cache/protocol.py, so moving them required patching the source.' 即原先路径硬编码在源码中, 若要调整路径必须修改代码, 不够灵活, 也不便于部署到不同目录或共享存储场景。

实现拆解

1. 在 python/sglang/srt/envron.py 的 Envs 类中新增两个环境变量定义: SGLANG_WEIGHT_CACHE_SOCKET_TEMPLATE 和 SGLANG_WEIGHT_CACHE_READY_TEMPLATE, 均使用 EnvStr 类型, 默认值保持原来的 /tmp/sglang_weight_cache_rank{global_rank}.sock 和 /tmp/sglang_weight_cache_rank{global_rank}.ready。
2. 在 python/sglang/srt/weight_cache/protocol.py 中, 删除原有的模板常量 WEIGHT_CACHE_SOCKET_TEMPLATE 和 WEIGHT_CACHE_READY_TEMPLATE, 并新增 _format_daemon_path 辅助函数, 该函数从 env_field 读取模板, 并校验其中是否包含 {global_rank} 占位符, 若缺失则抛出 ValueError, 避免所有 rank 指向同一路径。
3. 修改 get_socket_path 和 get_ready_path, 使其调用 _format_daemon_path 并传入对应环境变量字段, 返回格式化后的路径。

关键文件:

- python/sglang/srt/weight_cache/protocol.py (模块 权重缓存; 类别 source; 类型 dependency-wiring; 符号 _format_daemon_path): 核心变更文件, 将路径常量改为动态读取环境变量, 并新增 _format_daemon_path 校验占位符。
- python/sglang/srt/envron.py (模块 环境配置; 类别 source; 类型 core-logic): 新增环境变量定义, 作为路径模板的配置源。

关键符号: _format_daemon_path

关键源码片段

python/sglang/srt/environ.py

新增环境变量定义，作为路径模板的配置源。

```
class Envs:
    # ...
    # =====
    # Weight Cache Daemon
    # =====
    # 路径模板，必须保留 {global_rank} 占位符，以保证每个 rank 都有独立路径
    SGLANG_WEIGHT_CACHE_SOCKET_TEMPLATE = EnvStr(
        "/tmp/sglang_weight_cache_rank{global_rank}.sock"
    )
    SGLANG_WEIGHT_CACHE_READY_TEMPLATE = EnvStr(
        "/tmp/sglang_weight_cache_rank{global_rank}.ready"
    )
```

评论区精华

无 review 评论，审核人 alexnails 直接批准。但 `_format_daemon_path` 函数的注释体现了设计考量：用户可覆盖模板，但 `str.format` 会静默忽略缺失的占位符，若不校验将导致多个 rank 竞争同一守护进程，因此主动拒绝。

- 暂无高价值评论线程

风险与影响

- 风险：路径模板可配置后，若用户设置的环境变量不含 `{global_rank}`，将抛出 `ValueError`（已在 `_format_daemon_path` 中防御），但若用户设置的环境变量模板格式错误（如 url 编码问题）可能导致其他异常。此外，环境变量值可被注入到路径中，但不存在路径遍历风险因为未用于文件读取，仅用于路径构造。该变更不影响默认行为，兼容性风险低。测试覆盖不足：本 PR 未添加针对新环境变量的单元测试。
- 影响：对用户：可通过环境变量灵活调整权重缓存守护进程的 `socket/ready` 文件位置，便于多节点共享存储或自定义目录。对系统：默认行为不变，所有调用点仍通过 `get_socket_path/get_ready_path` 获取路径，不影响现有功能。对团队：提升了配置灵活性，但需注意环境变量命名与现有配置风格一致。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR