PR #25686 完整报告

sgl-project/sglang

Use SGLANG_CACHE_DIR env for gpu_p2p_access_cache path

合并时间：2026-05-19 07:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25686

执行摘要

- 一句话：P2P 缓存路径使用 SGLANG_CACHE_DIR 环境变量
- 推荐动作：该 PR 是一个小的基础设施改进，推荐快速合并。技术团队可作为参考，类似硬编码路径问题应统一交由环境变量管理。

功能与动机

PR body 指出：将硬编码的 ~/.cache/sglang 替换为可配置的 SGLANG_CACHE_DIR 环境变量，使 CI 和其他环境能通过已有的 envs.SGLANG_CACHE_DIR 设置自定义缓存位置。

实现拆解

1. 新增导入(custom_all_reduce_utils.py): 从 sglang.srt.environ 导入 envs 作为 sglang_envs。
2. 替换路径生成(custom_all_reduce_utils.py: gpu_p2p_access_check 函数): 将原来的 SGLANG_CACHE_ROOT = os.path.expanduser("~/.cache/sglang") 改为 SGLANG_CACHE_ROOT = os.path.expanduser(sglang_envs.SGLANG_CACHE_DIR.get())。
3. 注释更新(custom_all_reduce_utils.py): 将注释从 "~/.cache/vllm" -> "~/.cache/sglang" 更新为 "~/.cache/vllm" -> envs.SGLANG_CACHE_DIR，反映新机制。

关键文件：

- python/sglang/srt/distributed/device_communicators/custom_all_reduce_utils.py（模块 分布式通信；类别 source；类型 dependency-wiring；符号 gpu_p2p_access_check）：核心变更文件：将 GPU P2P 访问缓存路径从硬编码改为使用 SGLANG_CACHE_DIR 环境变量。

关键符号：gpu_p2p_access_check

关键源码片段

python/sglang/srt/distributed/device_communicators/custom_all_reduce_utils.py

核心变更文件：将 GPU P2P 访问缓存路径从硬编码改为使用 SGLANG_CACHE_DIR 环境变量。

```
# 文件：python/sglang/srt/distributed/device_communicators/custom_all_reduce_utils.py

# 新增导入，获取环境变量对象
from sglang.srt.environ import envs as sglang_envs

def gpu_p2p_access_check(src: int, tgt: int) -> bool:
    """Check if GPU src can access GPU tgt."""
    # ... 前面的逻辑不变 ...

    # VLLM_CACHE_ROOT -> SGLANG_CACHE_ROOT
    # "~/.cache/vllm" -> envs.SGLANG_CACHE_DIR
    # 使用 SGLANG_CACHE_DIR 环境变量，默认 ~/.cache/sglang
    SGLANG_CACHE_ROOT = os.path.expanduser(sglang_envs.SGLANG_CACHE_DIR.get())
    path = os.path.join(
        SGLANG_CACHE_ROOT,
        f"gpu_p2p_access_cache_for_{cuda_visible_devices}.json"
    )
    # ... 后续逻辑不变 ...
```

评论区精华

该 PR 没有 review 评论，讨论仅包含作者自己的 /tag-and-rerun-ci 触发 CI 操作的评论和 gemini-code-assist 的配额提示。因此没有实质性技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及一条路径字符串的来源，逻辑和默认行为保持一致（因为 SGLANG_CACHE_DIR 的默认值就是 ~/.cache/sglang）。若 SGLANG_CACHE_DIR 环境变量未设置，.get() 默认值会生效，不会抛异常。主要风险是用户显式设置了非绝对路径可能导致 os.path.expanduser 行为与预期不符，但这属于环境变量语义问题，非 PR 引入。
- 影响：影响范围小：仅影响一个函数 gpu_p2p_access_check 的缓存路径生成。所有依赖该函数的地方都会自动使用新路径（如果设置了 SGLANG_CACHE_DIR）。用户影响：对于设置了 SGLANG_CACHE_DIR 的用户，P2P 缓存文件将迁移到新目录；未设置的用户无感知。团队影响：降低运维成本，CI 可通过环境变量统一管理缓存目录。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR