PR #22463 完整报告

sgl-project/sglang

Add skills for debugging hanging issues

合并时间：2026-04-10 01:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22463

执行摘要

本 PR 新增了一个技能文档，用于调试 SGLang 分布式推理中的挂起问题，基于 Issue #22276 的死锁场景，提供从定位到修复的系统化方法，旨在提升团队调试效率。

功能与动机

动机源于 Issue #22276，其中报告了在 Qwen3 Next MTP 的 CI 测试中，服务器因 NCCL AllGather 死锁而挂起。该 Issue 详细描述了挂起时的堆栈跟踪和 NCCL 日志，揭示了分布式状态分歧导致的集体操作阻塞。为此，本 PR 添加技能文档，标准化调试流程，帮助开发者快速应对类似问题。

实现拆解

实现仅涉及一个文件 .claude/skills/debug-distributed-hang/SKILL.md，内容结构化如下：

- 步骤 1：确认和定位挂起：使用 py-spy 和 watchdog 获取堆栈跟踪，识别阻塞线程。
- 步骤2：NCCL调试日志：设置NCCL_DEBUG=INFO检查collective操作的大小匹配问题。
- 步骤 3：CUDA Coredump：配置环境变量触发 GPU 核心转储，分析挂起的内核。
- 步骤 4：每 rank 日志记录：通过装饰器记录每个 rank 的状态，使用二进制搜索定位首个分歧点。
- 常见原因与修复模式：总结如大小不匹配、分支分歧等根因及相应解决方案。

关键代码示例（摘自文档）：

```
def per_rank_log(func):
    def wrapper(*args, **kwargs):
        rank = torch.distributed.get_rank()
        with open(f"debug_rank{rank}.log", "a") as f:
            f.write(f"{func.__name__} called\n")
        return func(*args, **kwargs)
    return wrapper
```

评论区精华

review 中，gemini-code-assist[bot] 提出了三处改进，均被采纳：

> 安全风险：" 使用固定路径在 /tmp 可能导致权限冲突或安全风险，建议使用相对路径或包含 PID。" —— 已更新为 f"debug_rank{rank}.log"。

> 性能优化："将 tensor 转换为 Python 列表和字符串效率低下，建议使用 tobytes() 进行哈希。" —— 已更新为 tensor.cpu().numpy().tobytes()。
>
> 正确性保障："管道到 tail 可能掩盖 pytest 失败，建议正确捕获退出码。" —— 已调整脚本逻辑。

这些讨论确保了文档的健壮性和实用性。

风险与影响

风险分析：

- 文档准确性：依赖外部工具和最新代码，可能过时；但 review 改进已缓解。
- 使用复杂性：需要安装 py-spy、cuda-gdb 等工具，增加初始设置负担。
- 无代码变更：对系统无直接回归或性能影响。

影响评估：

- 对团队：显著提升分布式调试能力，减少问题排查时间，促进协作。
- 对用户：提供清晰指南，帮助解决生产环境中的挂起问题。
- 对系统：无变更，不影响现有功能。

关联脉络

本 PR 是 SGLang 调试能力增强的一部分，与历史 PR #18569（添加对称调试模式）形成互补，两者共同扩展了系统的调试工具集。从近期 PR 趋势看，团队持续关注性能优化和问题排查（如 PR #22424 的 AMD 内核优化、PR #22335 的 AMD 崩溃修复），表明调试和性能调优是当前演进重点。