PR #21800 完整报告

sgl-project/sglang

[Misc] Tiny: Add test network timeouts and dynamic max-parallel for 5090/2-gpu runners

合并时间：2026-04-01 09:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21800

执行摘要

此 PR 通过为测试工具添加网络请求超时和动态调整 CI runner 并行度，旨在防止 CI 静默挂起并优化资源使用。变更涉及多个测试文件和 CI 工作流，属于基础设施维护，对测试稳定性和效率有积极影响。

功能与动机

动机源自 CI 测试中因外部网络请求无超时而导致的静默挂起问题。PR body 明确表述："Add explicit timeouts to external network requests in test utilities to prevent silent CI hangs" 和 "Add dynamic max-parallel for 5090/2-gpu runners"，以匹配现有 H100 模式，提升 CI 可靠性。

实现拆解

实现分为两个主要部分：

1. 测试文件超时添加：
   - 在 python/sglang/test/simple_eval_common.py 的 download_dataset 函数中，将 requests.get(url, stream=True) 改为 requests.get(url, stream=True, timeout=120)。
   - 在 python/sglang/test/simple_eval_mgsm.py 的 get_lang_examples 函数中，将 urllib.request.urlopen(fpath) 改为 urllib.request.urlopen(fpath, timeout=30)。
   - 在 python/sglang/test/vlm_utils.py 的 get_or_download_file 函数中，将 requests.get(url) 改为 requests.get(url, timeout=30)。
2. CI 工作流重构：
   - 在 .github/workflows/pr-test.yml 的 set-parallel 步骤中，引入 max_parallel_small 和 max_parallel_2gpu 输出变量。
   - 根据 scheduled run 或 high priority label 判断是否使用 full parallelism（如 5090 runner 为 8，2-gpu runner 为 4）或 throttled（如 5090 runner 为 3，2-gpu runner 为 2）。

评论区精华

本次 PR 未经过 review 讨论，所有变更由作者直接合并。提交历史显示作者通过多次提交（如 "Revert localhost request timeouts"）调整超时值，但无外部评审交锋。

风险与影响

风险：超时值（30 秒、120 秒）可能设置不当，在网络延迟高时导致测试失败；并行度调整可能引发 CI 调度冲突或资源浪费；提交历史中 revert 表明作者已处理本地超时问题，但外部超时仍需监控。

影响：对用户无直接影响；系统层面，提升 CI 测试稳定性，减少挂起风险；团队层面，优化 GPU 资源利用率，但需额外关注超时异常日志。

关联脉络

与近期 PR 如 #21778（缓存 Nvidia wheels）和 #21779（减少冗余测试）同属 CI 优化脉络，显示团队持续改进测试基础设施。同时，与 #21785（修改相同测试文件）存在功能关联，表明测试工具演进趋势。