PR #23314 完整报告

sgl-project/sglang

ci: limit nightly test parallelism to 1 job per hardware type

合并时间：2026-05-01 12:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23314

执行摘要

本 PR 在 nightly nvidia CI 工作流中引入每硬件类型的串行执行队列（concurrency group），限制矩阵 job 的分区并发数，并依据实际运行时间调整超时阈值，以提升夜间测试的稳定性和资源利用率。

功能与动机

夜间测试常因同一硬件类型上的多个 job 并行运行导致资源争抢和不稳定。PR body 明确要求“不同硬件家族的 job 排队而非并行”，并通过 concurrency.group 实现每个硬件类型（H100、H200、H20、B200）只有一个 job 运行。

实现拆解

1. 添加 per-hardware concurrency group：在每个 job 中增加 concurrency.group: nightly-hw-{h100,h200,h20,b200} 和 cancel-in-progress: false，使同一硬件类型的 job 按启动顺序排队串行执行。
2. 限制矩阵 job 分区并发：在四个矩阵 job（general-8-gpu-h200、general-8-gpu-b200、multimodal-server-1/2-gpu）中设置 max-parallel: 2，确保同一 workflow 内最多运行 2 个分区。
3. 调整 timeout-minutes：基于最近 5 次运行的实际最大耗时，将超时设为约 2 倍并取整到 30 分钟，例如 kernel-1-gpu 从 240 分钟降至 60 分钟，text-perf-2-gpu 从 180 分钟降至 30 分钟。

（见 JSON 中 key_files[0].annotated_snippet_markdown）

评论区精华

无 review 评论，作者在 issue 评论中提供了详尽的超时调整依据表。

风险与影响

串行执行不会增加总耗时（因无 parallel 能力），但会集中占用 runner 时间窗口；超时设置基于历史数据，需随测试复杂度增长定期复审。总体而言降低资源争抢风险，提升 CI 稳定性。

关联脉络

本 PR 是 CI 基础设施持续优化的一部分，近期的其他 CI PR（如 #24208、#24191）均旨在提升 CI 可靠性和效率。