PR #24163 完整报告

sgl-project/sglang

Revert "[ci] split stage-c-test-4-gpu-b200 to enable a low-disk runner pool"

合并时间：2026-05-01 06:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24163

执行摘要

- 一句话：回退 B200 CI 测试拆分，合并为单一套件
- 推荐动作：该 PR 为常规的 CI 清理操作，值得关注的是 B200 runner 资源管理策略的演进。对于 CI 维护者，该变更合理且必要；对于一般开发者，无需深入阅读。

功能与动机

专用低磁盘 B200 runner 已从集群中移除，原有的拆分依据不复存在，stage-c-test-4-gpu-b200-small 套件不再有效。因此需要回退拆分，简化 CI 配置并恢复测试频率。

实现拆解

1. CI 工作流回退：在 .github/workflows/pr-test.yml 中移除 b200_low_disk_runner 输出变量和关联的逻辑，将 B200 测试统一通过 b200_runner 路由到 stage-c-test-4-gpu-b200 套件，启用 --auto-partition-size 6。同时清理了日志输出中对低磁盘 runner 的引用。
2. 测试套件重定向：将原注册到 stage-c-test-4-gpu-b200-small 的 5 个测试（test_gpt_oss_4gpu、test_qwen35_fp4_mtp_v2、test_qwen35_fp4_triton、test_lora_gpt_oss_20b_logprob_diff、test_lora_nemotron_3_super_120b_a12b_logprob_diff 等）改为注册到 stage-c-test-4-gpu-b200，并根据实际运行时间上调了 est_time。
3. 恢复 nightly 测试为 per-commit：将原标记为 nightly-4-gpu-b200 的三个测试（test_nvidia_nemotron_3_super_nvfp4、test_fp8_blockwise_gemm、test_nvfp4_gemm）的 suite 改为 stage-c-test-4-gpu-b200，移除 nightly=True 属性，使其在每次提交时自动运行。
4. 其他测试文件适配：test_cutedsl_moe.py 等额外测试文件也随同调整了套件名称，保持一致性。

关键文件：

- .github/workflows/pr-test.yml（模块 CI 配置；类别 infra；类型 infrastructure）：核心 CI 配置文件，移除了低磁盘 runner 变量和相关步骤，将 B200 测试统一路由到单一套件。
- test/registered/4-gpu-models/test_nvidia_nemotron_3_super_nvfp4.py（模块 4GPU 测试；类别 test；类型 test-coverage）：从 nightly 套件改为 per-commit 套件，恢复测试频率。

- test/registered/4-gpu-models/test_gpt_oss_4gpu.py（模块 4GPU 测试；类别 test；类型 test-coverage）：从 small 套件合并到主套件，反映 runner 资源变更。

关键符号：未识别

关键源码片段

.github/workflows/pr-test.yml

核心 CI 配置文件，移除了低磁盘 runner 变量和相关步骤，将 B200 测试统一路由到单一套件。

```
# .github/workflows/pr-test.yml（关键变更片段）
steps:
  - name: Setup runner
    id: set-runner
    run: |
      target_stage="${{ inputs.target_stage }}"
      if [[ "$sgl_kernel" == "true" && -z "$target_stage" ]]; then
        echo "b200_runner=4-gpu-b200-kernel" >> $GITHUB_OUTPUT
        # 移除低磁盘 runner 输出：b200_low_disk_runner=4-gpu-b200-kernel-low-disk
      else
        echo "b200_runner=4-gpu-b200" >> $GITHUB_OUTPUT
        # 移除低磁盘 runner 输出：b200_low_disk_runner=4-gpu-b200-low-disk
      fi
  # 后续的日志输出也移除了 b200_low_disk_runner 行
```

评论区精华

review 仅有一条自动生成的评论，无实质讨论。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险为测试资源排布变化可能导致 CI 并行超限或排队加剧，但原拆分理由（低磁盘 runner）已不存在，合并后资源用量仍在预期内。调整后的预估时间基于历史数据，应能适配新套件。若未来再次引入低磁盘 runner，需重新评估拆分逻辑。
- 影响：对用户无直接影响，对 CI 系统而言简化了配置，减少了一个测试套件；3 个测试从 nightly 恢复到 per-commit，提高了回归检测频率。团队维护成本降低。
- 风险标记：CI 配置回退，测试套件合并，资源分配变化

关联脉络

- PR #23417 [ci] split stage-c-test-4-gpu-b200 to enable a low-disk runner pool: 被本 PR 回退的原始 PR，是本次变更的直接原因。