

PR #33605 完整报告

sgl-project/sglang

[CI] Make B200 base-b suites single-GPU as prep for 1-gpu B200 runners

合并时间: 2026-08-05 06:53

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33605>

执行摘要

- 一句话: B200 两个 4-GPU 测试移入 base-c 套件
- 推荐动作: 值得快速浏览, 不必精读。核心价值在于它释放了 base-b 4-GPU 测试压力, 是 B200 单 GPU runner 迁移的第一步。关注 CI 基础设施演进的工程师可以留意后续 1-GPU runner 落地 PR; 本 PR 本身只是注册参数调整, 设计上没有额外值得深挖的点。

功能与动机

PR 描述指出需要把仅有的两个 4-GPU 文件移出 B200 base-b 套件, 使其只包含单 GPU 测试, 从而可以在 1-GPU B200 runner 上运行或利用 in-node per-GPU 并行度, 提升 B200 测试资源利用率。

实现拆解

1. 修改 `test/registered/spec/eagle/test_deepseek_v3_fp4_mtp_small.py`: 将 `register_cuda_ci` 的 `stage` 参数从 `base-b` 改为 `base-c`, `runner_config` 保持 `4-gpu-b200`, `est_time` 保持 340 秒, 使该 DeepSeek 测试进入 `base-c` 的 4-GPU 队列。
2. 修改 `test/registered/kernels/ops/kimi_k3/test_collectives.py`: 将 `register_cuda_ci` 的 `stage` 参数从 `base-b-kernel-unit` 改为 `base-c`, `runner_config` 保持 `4-gpu-b200`, `est_time` 保持 240 秒; 同文件中的 `nightly-8-gpu-b200` 注册保持不变, 继续覆盖大规模通信场景。
3. 迁移完成后, `base-b-test-4-gpu-b200` 与 `base-b-kernel-unit-test-4-gpu-b200` 不再包含任何多卡用例, 全部由单 GPU 测试构成, 为后续迁移到 1-GPU B200 runner 或按节点内单 GPU 并行调度扫清障碍。
4. 本次纯属 CI 测试注册阶段调整, 不涉及测试逻辑、源码、配置或部署配套改动。

关键文件:

- `test/registered/kernels/ops/kimi_k3/test_collectives.py` (模块 集合通信; 类别 `test`; 类型 `ci-configuration`; 符号 `register_cuda_ci`): 将 Kimi-K3 集体通信测试从 `base-b-kernel-unit` 移到 `base-c`, 是 `base-b` 内核单元套件摆脱 4-GPU 用例的关键一步。
- `test/registered/spec/eagle/test_deepseek_v3_fp4_mtp_small.py` (模块 模型测试; 类别 `test`; 类型 `ci-configuration`; 符号 `register_cuda_ci`): 唯一占用 `base-b` 4-GPU 队列的 DeepSeek FP4 MTP 测试, 迁移后 `base-b-test-4-gpu-b200` 变为纯单 GPU 套件。

关键符号: register_cuda_ci

评论区精华

该 PR 没有收到任何 review 评论，作者直接合并；变更点极小且语义清晰，未见设计分歧。

- 暂无高价值评论线程

风险与影响

- 风险：变更仅涉及 CI 测试注册阶段，无源码或测试逻辑改动，回归风险极低。主要风险集中在 CI 资源调度侧：base-c 4-GPU 队列将新增两个任务（预估 240 秒和 340 秒），若该队列并发配额有限，可能增加排队时间；此外测试失败告警或监控若按原 stage 名称归类，base-b 到 base-c 的迁移需要同步更新视图。两个测试本身的 runner_config 仍为 4-gpu-b200，运行环境一致，不存在环境适配风险。
- 影响：对最终用户无感知，属于纯 CI 基础设施调整。对 B200 测试流程的影响是：base-b 套件从多卡混跑变成单卡专用，base-c 的 4-GPU 队列统一承载 DeepSeek 与 Kimi-K3 的多卡用例。对维护 CI 的团队而言，这条 PR 是 1-GPU B200 runner 计划的前置步骤，让 base-b 系列套件具备在单卡机器上运行的能力，后续可按每 GPU 平行执行测试，提高 B200 硬件利用率。
- 风险标记：CI 测试调度变更，base-c 队列负载上升，依赖后续单 GPU runner 落地

关联脉络

- PR #33586 [CI] Trim redundant B200 test registrations: 同为 B200 测试注册整理，涉及 base-c 套件的精简与提速，与本次将多卡测试移入 base-c 的目标一致。
- PR #33611 [Test] Replace NVFP4 MoE runner backend e2e matrix with a layer-level unit test: 同样围绕 B200 测试优化，缩减多卡 e2e 用例，减少对 4-GPU 资源的占用，与本次迁移方向互补。