

PR #31928 完整报告

sgl-project/sglang

Route `/rerun-test` for b200 multimodal tests to the b200 pool

合并时间: 2026-07-22 07:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31928>

执行摘要

- 一句话: 修复 `/rerun-test` 将 b200 多模态测试路由到 H100 池的问题
- 推荐动作: 该 PR 价值虽小但及时修复了实际的 CI 故障。建议阅读者关注 `scripts/ci/utils/slash_command_handler.py` 中 `detect_multimodal_suite` 函数的匹配逻辑, 未来若有更多硬件后缀测试文件加入, 应考虑重构为更鲁棒的正则匹配或精确匹配机制, 避免字典顺序依赖。

功能与动机

之前 `/rerun-test test_server_b200.py` 会被分发到 `1-gpu-h100` 池, B200 独有的 NVFP4 测试立即崩溃, 报错 `mm_fp4 does not support backend 'trtllm' with capability 90`。PR body 明确说明该问题源于 handler 的 multimodal runner 映射只识别 `2_gpu` 和 `2-gpu` 键, 其余全部回退到 `1-gpu-h100`。

实现拆解

变更只涉及一个文件、一行代码, 分为两个逻辑层面:

1. 在 `scripts/ci/utils/slash_command_handler.py` 的 `MULTIMODAL_PATH_TO_RUNNER` 字典中新增 `"b200": _B200_DEFAULT_RUNNER` 条目, 其中 `_B200_DEFAULT_RUNNER` 已在上下文中定义为 `"4-gpu-b200"`, 与 `pr-test.yml` 中 `multimodal-gen-test-1-b200` 使用的 runner 池一致。
2. 保持与其他映射键的兼容性: `2_gpu` 和 `2-gpu` 依然指向 `2-gpu-h100`; 新增的 `b200` 键优先级依赖字典迭代顺序 (Python 3.7+ 保持插入顺序)。

没有额外的测试或配置变更, 影响范围仅限于 `/rerun-test slash command` 的 multimodal suite 路由逻辑。

关键文件:

- `scripts/ci/utils/slash_command_handler.py` (模块 CI; 类别 infra; 类型 infrastructure)
: 核心变更文件, 新增 `b200` 映射键确保 B200 多模态测试路由到正确的 GPU 池。

关键符号: `handle_rerun_failed_ci`

关键源码片段

`scripts/ci/utils/slash_command_handler.py`

核心变更文件，新增 b200 映射键确保 B200 多模态测试路由到正确的 GPU 池。

```
# scripts/ci/utils/slash_command_handler.py

def handle_rerun_failed_ci(...):
    # 多模态测试路径 → runner 池的映射字典
    # detect_multimodal_suite 顺序遍历该字典，返回第一个匹配键对应的 runner
    MULTIMODAL_PATH_TO_RUNNER = {
        "2_gpu": "2-gpu-h100", # 注意：这里有两个相同的键 "2_gpu"，后者覆盖前者
        "2-gpu": "2-gpu-h100",
        "b200": _B200_DEFAULT_RUNNER, # PR 新增：将 b200 映射到 4-gpu-b200 池
    }
    MULTIMODAL_DEFAULT_RUNNER = "1-gpu-h100"
    # 注意：由于字典顺序迭代，若将来出现含 "b200" 和 "2-gpu" 的路径，
    # "2-gpu" 会优先匹配导致错误路由。建议将 b200 条目置于更前位置。
```

评论区精华

code review 中 [gemini-code-assist\[bot\]](#) 指出，由于 `detect_multimodal_suite` 顺序遍历字典并返回第一个匹配键，如果未来出现同时包含 `b200` 和 `2-gpu` 的测试文件名（如 `test_b200_2_gpu.py`），会因 `2-gpu` 排在前面而错误路由到 H100 池。建议将 `"b200"` 放在字典顶部以确保 B200 硬件需求优先。但该评论为 bot 自动生成，[hnyls2002](#) 已批准合并，未采纳排序建议，实际插入位置在已有键之后。

- 字典顺序导致路由优先级 (design): 未采纳，当前所有 b200 测试文件名不包含 2-gpu，暂时安全。

风险与影响

- 风险：风险极低：
 - 仅新增一个字典条目，不改变现有逻辑。
 - 但正如 review 所指出，未来可能出现文件名含多个 hint 的测试文件，现有插入位置可能导致路由不符合预期，但当前所有 b200 测试文件名均不含 2-gpu 字样，短时间内不会触发此边界情况。
- 影响：
 - 用户影响：开发者使用 `/rerun-test` 对 B200 多模态测试重跑时，不再触发 NVFP4 硬件的算力不兼容崩溃，能正确分配到 4-gpu-b200 runner pool。
 - 系统影响：CI 基础设施，无性能变化。
 - 团队影响：简化了 CI 维护人员的手动干预流程，避免了误报。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR