

PR #28745 完整报告

sgl-project/sglang

ci: add 4-GPU mi35x runner and rebalance off the saturated 8-GPU pool

合并时间: 2026-06-20 06:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28745>

执行摘要

- 一句话: 新增 4GPU AMD MI35x runner, 缓解 8GPU 饱和
- 推荐动作: 建议阅读: 该 PR 展现了如何通过 CI runner 标签细化来缓解资源竞争, 是典型的 CI 可观测性与成本优化实践。值得关注的是命名约定、套件拆分策略以及如何平衡覆盖与资源效率。

功能与动机

AMD CI runner-fleet 报告中 `linux-mi35x-gpu-8` 池严重饱和, 几乎所有 mi35x 任务都挤在上面 (P99 队列约 20h), 而 1-GPU 和 2-GPU 池几乎空闲, 4-GPU 标签完全未被引用。根本原因是多个仅需 ≤ 4 GPU 的任务被钉在 8-GPU 标签上, 占据整个 8-GPU 节点。

实现拆解

1. 新增 4-GPU runner 标签: 假设集群已划分出 `linux-mi35x-gpu-4` 标签 (PR 前置条件)。
2. 迁移夜间任务: 在 `nightly-test-amd.yml` 和 `nightly-test-amd-rocm720.yml` 中, 将 `nightly-8-gpu-mi35x-deepseek-r1-mx4p4-tp4` 和 `nightly-4-gpu-mi35x-minimax-m25` 的 `runs-on` 从 `linux-mi35x-gpu-8` 改为 `linux-mi35x-gpu-4`。
3. 迁移 PR 任务: 在 `pr-test-amd.yml` 和 `pr-test-amd-rocm720.yml` 中, 新增名为 `stage-c-test-4-gpu-amd-mi35x` 的 job (`runs-on: linux-mi35x-gpu-4`), 将 `DeepSeek-R1-MXFP4` 从原 8-GPU 套件拆分到新 job; 原 8-GPU 套件从两个分区 (`part: [0,1]`) 缩减为一个分区 (`part: [0]`), 仅保留 `Kimi-K2.5-MXFP4` 和 `Qwen3-Coder-Next` 两个真正需要 $TP=8$ 的任务。
4. 注册新套件: 在 `test/run_suite.py` 的 `PER_COMMIT_SUITES[AMD]` 列表中插入 `stage-c-test-4-gpu-amd-mi35x`, 使新套件可被 CI 调度。
5. 重命名并调整测试参数: 将 `test/registered/amd/test_deepseek_r1_mx4p4_8gpu.py` 重命名为 `test_deepseek_r1_mx4p4_4gpu.py`, 修改 `register_amd_ci` 的套件名, 将所有 `--tp 8` 改为 `--tp 4`, 并注释掉速度断言 (变为 `report-only`, 等待 $TP=4$ 门限重新标定)。
6. 保留夜间 $TP=8$ 覆盖: 夜间 8-GPU 套件 `nightly-amd-8-gpu-mi35x-deepseek-r1-mx4p4` 仍运行在 `linux-mi35x-gpu-8` 上, 确保 $TP=8$ 准确度覆盖不丢失。

关键文件:

- `test/registered/amd/test_deepseek_r1_mx4p4_4gpu.py` (模块 测试套件; 类别 `test`; 类型 `rename-or-move`): 核心测试文件: 从 8-GPU 重命名为 4-GPU, 调整 `TP` 参数并降级

速度断言，直接体现测试覆盖调整。

- .github/workflows/pr-test-amd.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：
主要 CI 工作流：新增 4-GPU job 并缩减 8-GPU job 分区，实现 rebalance 核心逻辑。
- .github/workflows/pr-test-amd-rocm720.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：与 pr-test-amd.yml 对称，为 ROCm 7.20 版本做相同变更。
- .github/workflows/nightly-test-amd.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：夜间工作流：将两个 TP=4 任务的 runner 从 8-GPU 改为 4-GPU。
- .github/workflows/nightly-test-amd-rocm720.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：与 nightly-test-amd.yml 对称的 ROCm 720 版本。
- test/run_suite.py（模块 测试调度；类别 test；类型 test-coverage）：套件注册文件：在 PER_COMMIT_SUITES 中添加新套件名称，使 CI 可发现。

关键符号：未识别

关键源码片段

test/registered/amd/test_deepseek_r1_mxfp4_4gpu.py

核心测试文件：从 8-GPU 重命名为 4-GPU，调整 TP 参数并降级速度断言，直接体现测试覆盖调整。

```
# test/registered/amd/test_deepseek_r1_mxfp4_4gpu.py
# 原文件从 8gpu 重命名而来，将 TP=8 改为 TP=4，并临时移除速度断言门限。
```

```
register_amd_ci(est_time=3600, suite="stage-c-test-4-gpu-amd-mi35x")
```

```
class TestDeepseekR1MXFP4(CustomTestCase):
    @classmethod
    def setUpClass(cls):
        # ... (省略 import 和 setup 公共部分)
        other_args = [
            "--tp",
            "4", # 原为 "8", 改为 4 以匹配 4GPU runner
            "--chunked-prefill-size",
            "131072",
            # ...
        ]
        cls.process = popen_launch_server(...)

    def test_bs_1_speed(self):
        # 吞吐速度测试：原断言 self.assertGreater(speed, 75) 被注释
        # 理由：TP=4 的吞吐与 TP=8 不同，需重新标定门限
        # TODO: 标定后重新启用断言（建议添加）
        print(f"{speed=:.2f}")
        # self.assertGreater(speed, 75) # 暂时 report-only
```

```
class TestDeepseekR1MXFP4MTP(CustomTestCase):
    @classmethod
```

```
def setUpClass(cls):
    other_args = [
        "--tp",
        "4", # 同样改为 4
        # ...
    ]

def test_bs_1_speed(self):
    # 原断言 self.assertGreater(speed, 150) 被注释
    # TODO: 重新标定后启用
    # self.assertGreater(speed, 150) # 暂时 report-only
```

评论区精华

Gemini Code Assist bot 在 `test_deepseek_r1_mxfp4_4gpu.py` 的 `test_bs_1_speed` 方法中添加了评论，建议对 report-only 的断言添加明确的 TODO 注释以跟踪重新标定，避免静默性能回归。虽然代码中已用注释说明原因但未添加 TODO，HaiShaw 最终批准了 PR，未要求修改。

- 性能断言降级应添加 TODO 注释 (testing): 代码保持原样：断言被注释并加了说明，但未添加 TODO。HaiShaw 批准 PR，未强制要求修改。

风险与影响

- 风险：
 - 性能门禁降级：test_bs_1_speed 中的速度断言被注释掉，变为 report-only。在重新标定前，TP=4 的性能回退不会被 CI 拦截，可能导致部署后吞吐下降。但由于夜间任务仍保留 TP=8 覆盖，且 GSM8K 准确度断言仍有效，风险可控。
 - runner 标签依赖：新增的 linux-mi35x-gpu-4 标签需预先在集群中定义，若未正确划分，任务会因找不到 runner 而失败。
 - 测试重复或遗漏：重命名文件后旧文件名未清理，可能留下死代码（旧文件不在变更列表中，说明已删除）。需确认旧文件已被删除。
 - 影响：对最终用户无直接影响。对 AMD CI 系统有正面影响，降低排队时间和 8-GPU 节点占用，提高资源利用率。对 CI 维护者需要确保 linux-mi35x-gpu-4 标签可用。对 DeepSeek-R1-MXFP4 模型测试不再验证 TP=4 吞吐门限，但认可度仍通过。
 - 风险标记：性能断言降级为 report-only，runner 标签依赖，测试文件重命名需确认旧文件已删除

关联脉络

- PR #28736 [AMD] register 3 tests to stage-b-test-1-gpu-large-amd (batch-6): 同为 AMD CI 测试注册 PR，展示了套件注册的流程，与本 PR 的 test/run_suite.py 修改类似。
- PR #28357 [AMD] fix(jit): port kv_canary write/verify/plan kernels to ROCm: 同为 AMD CI 改进，涉及 JIT 内核移植和 CI 测试注册，与本 PR 在同一 CI 基础设施上下文。