PR #36290 完整报告

sgl-project/sglang

[AMD][CI] Adjust MI300 score API performance thresholds

合并时间：2026-08-26 00:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36290

执行摘要

- 一句话：调整 MI300 score API 性能阈值
- 推荐动作：值得快速浏览，但无深度技术内容。关注 AMD CI 性能基线的维护，建议记录放宽阈值的理由，并考虑在性能稳定后收紧。

功能与动机

AMD CI 镜像更新 AITER c16d44b 后，MI300 score API 基准平均延迟稳定在 51–54 ms，超过共享的 48 ms 限制。虽然请求仍成功完成，但 CI 会因性能断言失败而报错，因此需要为 MI300 单独调整阈值。

实现拆解

1. 修改 test/registered/perf/test_bench_serving_1gpu_part2.py 中的 test_score_api_latency_throughput 测试方法。
2. 在断言处增加 is_in_amd_ci() 分支，为 AMD CI 设置更宽松的阈值：平均延迟 < 60 ms、p95 延迟 < 65 ms、吞吐量 > 16 req/s。
3. 非 AMD CI 环境保留原来的 CUDA 阈值（< 48 ms、< 50 ms、> 20 req/s）。
4. 未涉及任何源码逻辑或配置改动，仅调整测试断言。

关键文件：

- test/registered/perf/test_bench_serving_1gpu_part2.py（模块 性能测试；类别 test；类型 test-coverage）：唯一变更文件，调整 MI300 score API 性能阈值

关键符号：test_score_api_latency_throughput

关键源码片段

test/registered/perf/test_bench_serving_1gpu_part2.py

唯一变更文件，调整 MI300 score API 性能阈值

```
# 测试方法 test_score_api_latency_throughput 的断言部分
# 原断言为统一阈值，现按硬件平台区分
self.assertEqual(res["successful_requests"], res["total_requests"])

# relax for mi300x（AMD CI 专用）
```

```
if is_in_amd_ci():
    # MI300 平均延迟放宽到 60 ms，p95 放宽到 65 ms，吞吐降至 16 req/s
    self.assertLess(res["avg_latency_ms"], 60)
    self.assertLess(res["p95_latency_ms"], 65)
    self.assertGreater(res["throughput"], 16)
else:
    # CUDA 仍保持原阈值，保证 NVIDIA 平台性能基线不变
    self.assertLess(res["avg_latency_ms"], 48)
    self.assertLess(res["p95_latency_ms"], 50)
    self.assertGreater(res["throughput"], 20)
```

评论区精华

无 review 评论。仅 HaiShaw 批准，未提出进一步讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，仅放宽测试阈值，可能掩盖真实性能回归，使得 AMD 平台性能退化不易被 CI 捕获。建议后续关注 AMD 平台性能基线，必要时重新收紧阈值。
- 影响：影响仅限于 AMD CI 测试，降低误报率；对用户和系统无直接影响。团队可减少因阈值过严导致的 CI 失败，但需注意性能监控灵敏度下降。
- 风险标记：测试阈值放宽，性能回归监测减弱

关联脉络

- 暂无明显关联 PR