

PR #31615 完整报告

sgl-project/sglang

[AMD] batch 3: register newly-added JIT kernel benchmarks for jit-kernel-benchmark-test-amd

合并时间: 2026-07-21 14:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31615>

执行摘要

- 一句话: 为 AMD 新增 2 个 JIT kernel 基准测试注册
- 推荐动作: 可快速合并。作为 CI 基础设施的例行扩展, 无需精读。

功能与动机

确保 AMD GPU 平台能够持续跟踪 JIT kernel 的性能变化, 避免新增 benchmark 遗漏在 AMD CI 中运行。PR body 提到 'Follow-up to #30307 / #31492', 并指出两个 benchmark 在 ROCm 上均通过, 无需裁剪。

实现拆解

1. 在 `bench_spec_topk1.py` 和 `bench_vocab_parallel_embedding.py` 的导入区域增加 `register_amd_ci` 的 import。
2. 在每个文件的原有 `register_cuda_ci` 调用之后, 添加一行 `register_amd_ci(est_time=..., stage="jit-kernel-benchmark", runner_config="amd")`。
3. 通过 `probe` 运行确认两个 benchmark 在 ROCm 7.0 和 7.2 上均通过 (18/18)。
4. 严格门禁运行 (`continue_on_error=false`) 也通过。

关键文件:

- `test/registered/jit/benchmark/bench_spec_topk1.py` (模块 基准测试; 类别 test; 类型 test-coverage) : 为 speculative decoding topk1 辅助函数 benchmark 添加 AMD CI 注册, 确保 AMD 平台持续跟踪性能。
- `test/registered/jit/benchmark/bench_vocab_parallel_embedding.py` (模块 基准测试; 类别 test; 类型 test-coverage) : 为 vocab-parallel embedding benchmark 添加 AMD CI 注册, 确保 AMD 平台持续跟踪性能。

关键符号: 未识别

关键源码片段

`test/registered/jit/benchmark/bench_spec_topk1.py`

为 speculative decoding topk1 辅助函数 benchmark 添加 AMD CI 注册, 确保 AMD 平台持续跟踪性能。

```
# bench_spec_topk1.py
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

register_cuda_ci(
    est_time=30, stage="base-b-kernel-benchmark", runner_config="1-gpu-large"
)
# Register for AMD CI as well, with same estimated time but different stage
register_amd_ci(est_time=30, stage="jit-kernel-benchmark", runner_config="amd")
```

test/registered/jit/benchmark/bench_vocab_parallel_embedding.py

为 vocab-parallel embedding benchmark 添加 AMD CI 注册，确保 AMD 平台持续跟踪性能。

```
# bench_vocab_parallel_embedding.py
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

register_cuda_ci(
    est_time=10, stage="base-b-kernel-benchmark", runner_config="1-gpu-large"
)
# Register for AMD CI with shorter estimated time
register_amd_ci(est_time=10, stage="jit-kernel-benchmark", runner_config="amd")
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅添加 CI 注册，不修改任何业务逻辑或 benchmark 实际代码。若 AMD CI 环境发生变化，可能导致 benchmark 失败，但 register_amd_ci 的 continue_on_error 机制可避免阻塞其他 CI。
- 影响：影响范围小：仅涉及 AMD CI 的 jit-kernel-benchmark 阶段，增加 2 个 benchmark 的执行（总耗时约 40s）。对用户无直接影响。
- 风险标记：变更量极小，已在 ROCm 7.0/7.2 验证

关联脉络

- PR #30307 batch 1 of JIT kernel benchmark registration for AMD: 本 PR 是 #30307 的后续，新增两个 benchmark 的 AMD CI 注册。
- PR #31492 batch 2 of JIT kernel benchmark registration for AMD: 本 PR 是 #31492 的后续，继续为新增 benchmark 添加 AMD CI 注册。