

PR #36139 完整报告

sgl-project/sglang

[AMD] Skip shared-KV verify test on ROCm 7.0 CI

合并时间: 2026-08-25 13:20

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36139>

执行摘要

- 一句话: ROCm 7.0 下跳过 FP8 shared-KV 测试
- 推荐动作: 该 PR 值得精读, 展示了如何精准地处理环境相关的测试跳过, 避免过度跳过导致覆盖损失。建议关注其使用 `get_hip_version()` 和 `skipIf` 的模式, 以及如何记录清晰的移除条件。

功能与动机

ROCm 7.0 CI 镜像上的 Triton 3.4 无法处理 gfx950 FP8 KV 的 `tl.dot` 操作, 导致 `TestVerifySharedKV.test_qwen3_5_fp8_kv_cache` 测试失败 (abort)。该问题已在上游 Triton 3.6+ 修复 (triton-lang/triton#8278), 但当前 CI 镜像未升级, 因此选择跳过该测试以避免 CI 失败, 同时保留其他测试的覆盖。

实现拆解

1. 在测试文件中添加 `get_hip_version` 导入: 从 `sglang.srt.utils` 导入 `get_hip_version`, 用于获取当前 HIP 版本。
2. 在 `test_qwen3_5_fp8_kv_cache` 方法上添加 `@unittest.skipIf` 装饰器: 条件为 `get_hip_version()[2] == (7, 0)`, 即仅当 HIP 主版本为 7 且次版本为 0 时跳过该测试。装饰器附带了详细的跳过原因说明, 并指出该跳过可在镜像升级到 Triton 3.6+ 后移除。
3. 保留其他测试方法不变: `test_qwen3_5_tp_shapes`、`test_qwen3_5_short_verify_widths`、`test_kimi_k3_absorbed_mla_shape` 和 `test_rejects_multiple_local_kv_heads` 继续在 ROCm 7.0 上运行, 确保非 FP8 场景的覆盖不受影响。

关键文件:

- `test/registered/attention/test_verify_shared_kv.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 核心变更文件, 添加了 ROCm 7.0 下的 FP8 测试跳过逻辑。

关键符号: `test_qwen3_5_fp8_kv_cache`

关键源码片段

`test/registered/attention/test_verify_shared_kv.py`

核心变更文件, 添加了 ROCm 7.0 下的 FP8 测试跳过逻辑。

```
from sglang.srt.utils import get_hip_version
```

```

class TestVerifySharedKV(CustomTestCase):
    # ... 其他测试方法 ...

    @unittest.skipIf(
        get_hip_version()[2] == (7, 0), # 仅当 HIP 版本为 7.0 时跳过
        "Triton 3.4 on ROCm 7.0 aborts gfx950 fp8 KV tl.dot "
        "(triton-lang/triton#8278). Remove once the image uses Triton >= 3.6.",
    )
    def test_qwen3_5_fp8_kv_cache(self):
        # 此测试使用 FP8 cache_dtype, 触发 gfx950 上的 fp8 tl.dot,
        # 在 Triton 3.4 下会 abort, 因此仅在 ROCm 7.0 上跳过,
        # 其他版本保留覆盖。
        self._run_parity(
            head_dim=256,
            v_head_dim=256,
            h_q=8,
            cache_dtype=torch.float8_e4m3fn,
            k_scale=0.5,
            v_scale=0.25,
            atol=FP8_ATOL,
            rtol=FP8_RTOL,
        )

```

评论区精华

在 issue 评论中, [cursor\[bot\]](#) 指出该测试在 ROCm 7.0 上持续失败 (job 97067981918 中 [stage-b-test-1-gpu-small-amd-mi35x](#)), 并建议只跳过方法而非整个类, 因为只有 [test_qwen3_5_fp8_kv_cache](#) 使用了 `cache_dtype=torch.float8_e4m3fn`, 从而触发 FP8 dot。 [michaelzhang-ai](#) 确认了修复有效, 测试被干净跳过, 但指出 [test_rejects_multiple_local_kv_heads](#) 仍因 `AssertionError: True is not false` 失败, 这与 [verify_mla.py](#) 中 `can_handle()` 的放宽有关 (PR#35499), 与本 PR 无关。

- 跳过范围: 方法级别而非类级别 (design): 采纳建议, 仅在该方法上添加 `skipIf`。
- 移除条件文档化 (documentation): 在装饰器中添加明确注释, 便于未来移除。

风险与影响

- 风险: 测试覆盖风险: 跳过 [test_qwen3_5_fp8_kv_cache](#) 意味着在 ROCm 7.0 上失去了 FP8 KV 场景的回归测试覆盖, 但该覆盖在 ROCm 7.2+ 上保留。 环境依赖风险: 跳过条件依赖 `get_hip_version()` 返回的版本号, 若在非 HIP 构建上返回 `(0, 0, 0)`, 则不会跳过, 安全。 CI 稳定性风险: 该 PR 仅解决 FP8 测试的跳过, [test_rejects_multiple_local_kv_heads](#) 的失败仍可能导致 CI 失败, 但与本 PR 无直接关系。
- 影响: 影响范围: 仅影响 ROCm 7.0 环境下 [test_verify_shared_kv.py](#) 中一个测试方法的执行, 对其他用户和系统无影响。 CI 影响: 减少 ROCm 7.0 上的 CI 失败, 提高 CI 稳定性。 团队影响: 降低了 AMD CI 的维护负担, 但需注意后续镜像升级后移除跳过条件。
- 风险标记: 测试覆盖缺失, 环境特定

关联脉络

- PR #35499 Relax `can_handle()` in `verify_mla.py`: 该 PR 放宽了 `verify_mla.py` 中 `can_handle()` 的检查, 导致 `test_rejects_multiple_local_kv_heads` 失败, 与本 PR 的测试环境相关。