

PR #27850 完整报告

sgl-project/sglang

[AMD] Fix DSA device-to-host direct test on rocm720 (page_size%16 assert)

合并时间: 2026-06-12 05:24

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27850>

执行摘要

- 一句话: 跳过 AMD 下 DSA 设备到主机 direct 测试
- 推荐动作: 值得快速合并。该 PR 简洁地修复了特定 AMD 平台上的 CI 失败, 且作者已通过两次 CI 运行验证。不涉及生产逻辑, 无需深入 review。

功能与动机

修复 CI 测试 `pr-test-amd-rocm720` 中 `stage-b-1gpu-small` 失败, 由于 `DSATokenToKVPool` 在 `HIP-preshuffle` 路径下断言 `page_size % 16 == 0`, 但测试中使用 `page_size=1`。PR body 明确指出该问题由 [#25939](#) 引入, 当时仅在 `mi3xx` 上验证通过 (旧版 ROCm 无此断言), 而 `rocm720` 新版本暴露了该问题。

实现拆解

1. 定位问题: 在 `test/registered/unit/mem_cache/test_dsa_pool_host_unit.py` 中, `test_device_to_host_indexer_direct` 方法调用 `_run_device_to_host_indexer_copy(io_backend="direct")`, 构建 `DSATokenToKVPool` 时因 `page_size=1` (HIP 默认值) 触发了 ROCm 7.2.0 的 `assert`。
2. 添加 `skip` 装饰器: 在 `test_device_to_host_indexer_direct` 方法上方添加 `@unittest.skipIf(is_hip(), ...)`, 跳过条件和注释与已有的 `test_device_to_host_indexer_kernel` 的 `skip` 一致。
3. 验证: 通过两个 CI 运行 (`mi3xx` 标准 AMD 和 `rocm720`) 确认修复有效, 所有 14 个 `partition` 通过。

关键文件:

- `test/registered/unit/mem_cache/test_dsa_pool_host_unit.py` (模块测试; 类别 `test`; 类型 `test-coverage`): 唯一修改文件, 添加了条件 `skip` 装饰器以修复 AMD ROCm 7.2.0 上的 CI 失败。

关键符号: 未识别

关键源码片段

`test/registered/unit/mem_cache/test_dsa_pool_host_unit.py`

唯一修改文件, 添加了条件 `skip` 装饰器以修复 AMD ROCm 7.2.0 上的 CI 失败。

```

# 在 test_device_to_host_indexer_direct 方法前添加 skipIf 装饰器
# 原 kernel 变体已有相同 skip, 保持一致

@unittest.skipIf(
    is_hip(),
    '`io_backend="kernel"` path in memory_pool_host.backup_from_device_all_layer '
    "raises ValueError on AMD (only the `direct` IO backend is wired for ROCm). "
    "The other 62 tests in this file pass on AMD.",
)
def test_device_to_host_indexer_kernel(self):
    self._run_device_to_host_indexer_copy(io_backend="kernel")

# 新增: 跳过 AMD 上使用 direct 后端的设备到主机拷贝测试
# 原因: ROCm 7.2.0 中 DSATokenToKVPool 要求 page_size % 16 == 0
# 但 HIP 默认 page_size=1 导致断言失败。
@unittest.skipIf(
    is_hip(),
    "DSATokenToKVPool with page_size=1 (used on HIP) trips the ROCm 7.2.0 "
    "HIP-preshuffle assert `page_size % 16 == 0` during pool construction "
    "(seen on pr-test-amd-rocm720). The rest of this file passes on AMD.",
)
def test_device_to_host_indexer_direct(self):
    self._run_device_to_host_indexer_copy(io_backend="direct")

```

评论区精华

无 review 评论, 仅有一个 CI bot 确认该 PR 安全可合并, 且指出 0 个失败与 PR 相关。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。变更仅在 AMD 平台跳过一项测试, 不修改任何生产代码。direct IO 后端 在 HIP 上是唯一可用的后端 (kernel 后端已跳过), 但 skip 后不会影响任何功能验证, 因为该测试只验证设备到主机拷贝的正确性, 而其他测试 (TestDSAOffloadSignatures 等 62 个) 仍然运行。
- 影响: 影响范围局限于 pr-test-amd-rocm720 CI 测试。跳过该测试后, rocm720 上不再因此断言失败, 其余 AMD CI 无影响。对用户无影响, 仅 CI 维护者受益。
- 风险标记: 仅测试变更, 平台特定修复

关联脉络

- PR #25939 Register DSA pool host unit tests on AMD CI: 引入该测试文件到 AMD CI, 但仅在 mi3xx 上验证通过, 忽略了 ROCm 7.2.0 上的断言问题。