

PR #28925 完整报告

sgl-project/sglang

fix(nightly-precision): pin flashinfer allreduce-fusion backend for TP-partial capture contract

合并时间: 2026-07-02 08:26

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28925>

执行摘要

- 一句话: 修复夜间精度测试因融合后端变化导致的失败
- 推荐动作: 值得合并, 是典型的测试基础设施修复案例。PR body 中的根因分析和验证方法值得参考。

功能与动机

夜间精度回归测试在 H200 上自 2026-06-17 起持续失败, 根因是 #23402 移除了 SM90 自动融合后, H200 上融合关闭导致捕获的张量类型从 TP-partial 变为 full-sum, 与比较器 tp:partial 契约冲突。需要在测试中显式固定融合后端以恢复正确捕获。

实现拆解

1. 添加常量与注释: 在 test/registered/debug_utils/test_nightly_precision_regression.py 中新增 PRECISION_FUSION_BACKEND = "trtllm" 常量, 并附上注释解释为何需要固定融合后端 (#23402 移除了 SM90 自动启用, H200 上融合关闭)。
2. 纳入捕获签名: 修改 _capture_signature 函数, 在生成签名的输入中加入 dump_cfg["fusion_backend"], 这样未来融合后端默认值变更时捕获签名会改变, 基线会自动重建, 避免静默不匹配。
3. 注入 dump 配置: 在 _test_one_model 的 dump_cfg 字典中添加 "fusion_backend": PRECISION_FUSION_BACKEND, 使其随捕获签名一起使用。
4. 元数据记录: 在 _maybe_hf_push 中将 fusion_backend 写入元数据, 便于事后审计。
5. 服务启动参数: 在 _run_server_and_dump 的服务启动参数列表中添加 "--flashinfer-allreduce-fusion-backend", PRECISION_FUSION_BACKEND, 确保服务以 trtllm 后端运行, 与捕获契约一致。

关键文件:

- test/registered/debug_utils/test_nightly_precision_regression.py (模块 夜间测试; 类别 test; 类型 test-coverage) : 唯一的变更文件, 包含融合后端固定、捕获签名扩展、服务参数注入等所有修改。

关键符号: _capture_signature, _test_one_model, _maybe_hf_push, _run_server_and_dump

关键源码片段

test/registered/debug_utils/test_nightly_precision_regression.py

唯一的变更文件，包含融合后端固定、捕获签名扩展、服务参数注入等所有修改。

文件: test/registered/debug_utils/test_nightly_precision_regression.py

在第 66 行新增常量，固定融合后端为 trtllm

原因是 #23402 移除了 SM90 自动融合后，H200 上融合关闭导致捕获张量类型不匹配

PRECISION_FUSION_BACKEND = "trtllm"

修改 _capture_signature，将 fusion_backend 纳入签名

这样未来融合后端默认值变化时，捕获签名会改变，基线自动重建

def _capture_signature(dump_cfg: dict[str, Any], tp_size: int) -> str:

```
    raw = "|".join(
        str(x)
        for x in (
            SCHEMA_VERSION,
            dump_cfg["max_tokens"],
            dump_cfg["ignore_eos"],
            tp_size,
            dump_cfg["dumper_filter"],
            dump_cfg["fusion_backend"], # 新增：确保后端变更时签名不同
        )
    )
    return hashlib.sha1(raw.encode()).hexdigest()[:12]
```

在 _test_one_model 中，将 fusion_backend 注入 dump_cfg

dump_cfg["fusion_backend"] = PRECISION_FUSION_BACKEND

在 _maybe_hf_push 中，将 fusion_backend 写入元数据

metadata["fusion_backend"] = dump_cfg.get("fusion_backend")

在 _run_server_and_dump 的服务参数中，添加 --flashinfer-allreduce-fusion-backend

```
server_args += [
    "--flashinfer-allreduce-fusion-backend",
    PRECISION_FUSION_BACKEND,
]
```

评论区精华

无实质性讨论: gemini-code-assist[bot] 的自动化评论仅说明 PR 内容，无反馈; alisonshao 直接批准。主要技术论证已在 PR body 中由作者完成，包括根因分析、验证结果和与相关 PR #28190 的关联。

- 暂无高价值评论线程

风险与影响

- 风险：风险低：仅修改测试配置文件，不涉及任何生产代码。测试服务参数显式指定 trtllm 后端，与 H200 上原有行为一致（融合开启），回归风险极小。但若未来 trtllm 后端自身有

变更，可能需同步更新本测试。

- 影响：影响范围仅限于夜间精度回归测试（`test_nightly_precision_regression.py` 在 `nightly-precision-8-gpu-h200 suite` 中）。修复后该测试可在 H200 上正确捕获 TP-partial 张量，与已有基线兼容。对其他测试或生产路径无影响。
- 风险标记：仅影响测试基础设施，依赖外部融合后端行为

关联脉络

- PR #23402 Reenable MNNVL backend for FlashInfer allreduce fusion: 本 PR 修复的回归由 #23402 触发，该 PR 移除了 SM90 的自动融合列表，导致 H200 上融合关闭。
- PR #28190 fix(_select_latest_run): don't promote failed runs to baseline: 相关 PR，修复了基线选择 bug，与本文独立但建议一起合入以防失败基线掩盖问题。