

PR #1642 完整报告

radixark/miles

fix(deepseek_v4): disable aggressive smem merge in sparse-MLA backward (NaN dQ/dKV)

合并时间: 2026-07-14 03:46

原文链接: <http://prhub.com.cn/radixark/miles/pull/1642>

执行摘要

- 一句话: 禁用 DSV4 稀疏 MLA 后向 aggressive smem merge, 修复 NaN 梯度
- 推荐动作: 值得精读, 重点看 PR body 的根因分析与验证设计, 以及 Issue #1571 的 bisect 过程。设计决策上, “用 2% 内核性能换取确定性正确性”是合理权衡; 更值得借鉴的是其基于实验排除干扰项 (fp8、DSA-indexer、坏节点) 的方法。建议团队为所有 sparse-MLA 内核 (glm5、DSV4 等) 建立共享的 pass_configs 基线, 并增加编译配置的回归测试, 避免同类 bug 在不同内核副本间重复出现。

功能与动机

DSV4-Flash RL 训练在第一个 grad-norm 检查即崩溃, 报错“found NaN in local grad norm for bucket #0 in backward pass before data-parallel communication collective”。PR body 指出 forward/rollout 均有限, 首次后向即出现 NaN; 根因是 bwd 内核 JIT 编译时开启 TL_ENABLE_AGGRESSIVE_SHARED_MEMORY_MERGE 引发共享内存别名 miscompile, 导致 dQ/dKV 为 NaN。同一缺陷此前由 Issue #1571 针对 glm5 内核修复, 但该修复只覆盖 glm5 且限制在 Blackwell (sm103) 上; DSV4 的 kernel 变体在 Hopper (H200, cc 9.0) 上也复现, 因此需要无条件禁用该 pass。

实现拆解

1. 定位变更入口: miles_plugins/models/deepseek_v4/ops/kernel/tilelang_sparse_mla_bwd.py 中 bwd 内核的 tilelang JIT pass_configs。
2. 核心变更: 将 tilelang.PassConfigKey.TL_ENABLE_AGGRESSIVE_SHARED_MEMORY_MERGE 由 True 改为 False, 禁止 JIT 对共享内存做激进合并。这是全 PR 唯一改动 (+1/-1), 内核主体逻辑不变。
3. 验证方式: 在 H200 上, dev 镜像 6 节点验证 grad_norm 为 0.075/0.056, :latest + 其他修复 7 节点验证为 0.0753; 并实验排除了 fp8 vs bf16、DSA-indexer 配置、坏节点等干扰因素。
4. 成本评估: 按 Issue #1571 的 ablation, 该内核性能损失约 2%, 端到端影响小于 1%。PR 未附带自动化测试, 回归保障依赖后续补充或上游 tilelang 修复。

关键文件:

- miles_plugins/models/deepseek_v4/ops/kernel/tilelang_sparse_mla_bwd.py (模块 模型 内核; 类别 source; 类型 configuration; 符号 bwd): 唯一的变更文件; 在 DSV4

sparse-MLA 后向内核的 tilelang pass_configs 中关闭 aggressive shared-memorymerge，修复共享内存别名导致的 NaN dQ/dKV，解除 DSV4-Flash RL 首次 grad-norm 检查即崩溃的训练阻断。

关键符号：bwd

关键源码片段

miles_plugins/models/deepseek_v4/ops/kernel/tilelang_sparse_mla_bwd.py

唯一的变更文件；在 DSV4 sparse-MLA 后向内核的 tilelang pass_configs 中关闭 aggressive shared-memory merge，修复共享内存别名导致的 NaN dQ/dKV，解除 DSV4-Flash RL 首次 grad-norm 检查即崩溃的训练阻断。

文件：miles_plugins/models/deepseek_v4/ops/kernel/tilelang_sparse_mla_bwd.py

片段：DSV4 sparse-MLA 后向内核 bwd 的 tilelang JIT 装饰器 pass_configs 配置

```
# 该配置块位于 bwd 内核的 tilelang JIT 装饰器中，控制编译 pass。
pass_configs={
    tilelang.PassConfigKey.TL_DISABLE_TMA_LOWER: True,
    tilelang.PassConfigKey.TL_DISABLE_WARP_SPECIALIZED: True,
    # 修复点：aggressive smem merge 会让仍存活的 shared buffer 互相别名（miscompile），
    # 产生 NaN dQ/dKV → NaN 梯度 → DSV4-Flash RL 首次 grad-norm 检查崩溃。
    # 与 PR #1571 修复的 glm5 内核属同一 tilelang 0.1.9 缺陷；
    # DSV4 变体在 Hopper（H200, cc 9.0）上同样复现，因此不能像 #1571 那样只对
    # Blackwell 生效，必须无条件关闭。代价约 2% 内核性能（来自 #1571 的 ablation）。
    tilelang.PassConfigKey.TL_ENABLE_AGGRESSIVE_SHARED_MEMORY_MERGE: False,
}
```

变更仅一行：将 True 改为 False，内核主体逻辑不变。

文件：miles_plugins/models/deepseek_v4/ops/kernel/tilelang_sparse_mla_bwd.py

片段：DSV4 sparse-MLA 后向内核 bwd 的 tilelang JIT 装饰器 pass_configs 配置

```
# 该配置块位于 bwd 内核的 tilelang JIT 装饰器中，控制编译 pass。
pass_configs={
    tilelang.PassConfigKey.TL_DISABLE_TMA_LOWER: True,
    tilelang.PassConfigKey.TL_DISABLE_WARP_SPECIALIZED: True,
    # 修复点：aggressive smem merge 会让仍存活的 shared buffer 互相别名（miscompile），
    # 产生 NaN dQ/dKV → NaN 梯度 → DSV4-Flash RL 首次 grad-norm 检查崩溃。
```

```
# 与 PR #1571 修复的 glm5 内核属同一 tilelang 0.1.9 缺陷；  
# DSV4 变体在 Hopper (H200, cc 9.0) 上同样复现，因此不能像 #1571 那样只对  
# Blackwell 生效，必须无条件关闭。代价约 2% 内核性能（来自 #1571 的 ablation）。  
tilelang.PassConfigKey.TL_ENABLE_AGGRESSIVE_SHARED_MEMORY_MERGE: False,  
}
```

变更仅一行：将 True 改为 False，内核主体逻辑不变。

评论区精华

本 PR 没有实质性的评讨论：Gemini Code Assist bot 仅给出“no feedback”，维护者 [yueming-yuan](#) 直接 approve。技术依据主要来自 PR body 中的根因分析和多节点验证表格，以及 Issue #1571 中对该 tilelang 缺陷的 bisect 结论。

- 暂无高价值评论线程

风险与影响

- 风险：正确性风险：关闭 aggressive smem merge 后，需确认没有其他 pass 组合导致新的 miscompile；#1571 曾以 50/50 压力迭代和 rel_err ~0.003 验证过类似修复，本 PR 在 H200 多节点上也确认 grad_norm 有限，风险较低。性能风险：该内核约有 2% 性能回退（#1571 的 ablation），端到端影响 <1%。兼容性：无条件关闭意味着 Blackwell 上的 DSV4 也会承担该性能成本；但这是避免崩溃的必要代价。测试风险：PR 未新增自动化回归测试，若未来有人重新打开该 flag 或升级 tilelang 改变默认行为，问题可能回归，建议补充针对 pass_configs 的编译级回归测试。
- 影响：影响范围：所有使用 DSV4-Flash RL 训练（fp8 与 bf16 均受影响）的用户，修复首次训练步崩溃，使训练可以持续推进；对 H200 与 Blackwell 均生效。团队影响：暴露了 tilelang PassConfig 在不同架构上的正确性差异，以及“同一缺陷在不同内核副本间漂移”的维护问题；建议统一管理 sparse-MLA 家族内核的编译配置，并跟进上游 tilelang 修复。
- 风险标记：核心路径变更，缺少自动化回归测试，小幅性能回退，同类 miscompile 或波及其他内核

关联脉络

- PR #1571 GLM-5.2 744B on GB300: sm103 kernel fix, training config: 同一 tilelang aggressive smem merge miscompile 的原始修复，但只覆盖 glm5 且限制 Blackwell；本 PR 将其推广到 DSV4 并解除硬件限制，是其直接延续和补充。