

PR #30954 完整报告

sgl-project/sglang

[SM120] Allow fused MHC opt-in with standalone TileLang pre disabled

合并时间: 2026-07-26 09:02

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30954>

执行摘要

- 一句话: 修复 SM120 上融合 MHC 门控条件使 opt-in 生效
- 推荐动作: 值得精读。本 PR 展示了如何精细控制硬件特定 kernel 的启用门控, 分离独立路径与融合路径的依赖关系。测试用例设计清晰, 覆盖了条件矩阵的所有分支。对于维护 DeepSeek-V4 推理性能的工程师是重要的参考案例。

功能与动机

SM120 模型后处理设置 `SGLANG_OPT_USE_TILELANG_MHC_PRE=False`, 但 `_is_fused_mhc_post_pre_enabled` 要求该标志, 导致 `SGLANG_OPT_FUSE_MHC_POST_PRE=1` 在 SM120 上无效, 融合路径从未启用。此修复使融合 opt-in 对 SM120 生效。

实现拆解

1. 在 `python/sglang/srt/models/deepseek_v4.py` 中添加导入 `from sglang.srt.utils.common import is_sm120_supported`。
2. 修改 `_is_fused_mhc_post_pre_enabled` 函数, 将条件 `(envs.SGLANG_OPT_FUSE_MHC_POST_PRE.get() and envs.SGLANG_OPT_USE_TILELANG_MHC_PRE.get() and envs.SGLANG_OPT_USE_TILELANG_MHC_POST.get())` 改为 `(envs.SGLANG_OPT_FUSE_MHC_POST_PRE.get() and envs.SGLANG_OPT_USE_TILELANG_MHC_POST.get() and (envs.SGLANG_OPT_USE_TILELANG_MHC_PRE.get() or is_sm120_supported()))`。
3. 新增测试文件 `test/registered/unit/models/test_deepseek_v4_fused_mhc_policy.py`, 通过 `patch.object(deepseek_v4, "is_sm120_supported", return_value=sm120)` 和 `envs` 的 `override` 模拟 SM120 和非 SM120 环境, 测试三条用例。
4. 测试注册为 CPU CI 套件 (`register_cpu_ci(est_time=2, suite="base-a-test-cpu")`), 确保回归检测。

关键文件:

- `python/sglang/srt/models/deepseek_v4.py` (模块 DeepSeek V4; 类别 source; 类型 core-logic; 符号 `_is_fused_mhc_post_pre_enabled`, `is_sm120_supported`): 核心门控函数 `_is_fused_mhc_post_pre_enabled` 的修改, 添加 SM120 例外。

- test/registered/unit/models/test_deepseek_v4_fused_mhc_policy.py (模块测试; 类别 test; 类型 test-coverage; 符号 TestDeepseekV4FusedMHCPolicy, _is_enabled, test_sm120_allows_fused_opt_in_with_standalone_pre_disabled, test_other_platform_still_requires_tilelang_pre) : 全面覆盖门控条件组合, 确保 SM120 允许、非 SM120 限制、fuse 和 post 仍必需等场景。

关键符号: _is_fused_mhc_post_pre_enabled

关键源码片段

python/sglang/srt/models/deepseek_v4.py

核心门控函数 `_is_fused_mhc_post_pre_enabled` 的修改, 添加 SM120 例外。

```
from sglang.srt.utils.common import is_sm120_supported

def _is_fused_mhc_post_pre_enabled() -> bool:
    # SM120 禁用了独立的 TileLang pre 路径, 但 mhc_fused_post_pre 不读取该标志,
    # 它根据 token 数量阈值独立分派 (小批量使用 tilelang kernel,
    # 大批量使用 deepGEMM kernel), 所以 TileLang pre 标志不应 veto 融合 opt-in。
    # 因此, 融合路径的条件是: fuse 打开 AND tilelang_post 打开 AND
    # (tilelang_pre 打开 OR 当前平台为 SM120)。
    return (
        envs.SGLANG_OPT_FUSE_MHC_POST_PRE.get()
        and envs.SGLANG_OPT_USE_TILELANG_MHC_POST.get()
        and (envs.SGLANG_OPT_USE_TILELANG_MHC_PRE.get() or is_sm120_supported())
    )
```

test/registered/unit/models/test_deepseek_v4_fused_mhc_policy.py

全面覆盖门控条件组合, 确保 SM120 允许、非 SM120 限制、fuse 和 post 仍必需等场景。

```
"""Unit tests for the DeepSeek-V4 fused-MHC enable policy."""
import unittest
from unittest.mock import patch

import sglang.srt.models.deepseek_v4 as deepseek_v4
from sglang.srt.environ import envs
from sglang.test.ci.ci_register import register_cpu_ci
from sglang.test.test_utils import CustomTestCase

register_cpu_ci(est_time=2, suite="base-a-test-cpu")

class TestDeepseekV4FusedMHCPolicy(CustomTestCase):
    def _is_enabled(self, *, fuse: bool, tilelang_pre: bool,
                    tilelang_post: bool, sm120: bool) -> bool:
        # 通过 override 环境变量和 patch is_sm120_supported 来模拟不同场景
        with (
            envs.SGLANG_OPT_FUSE_MHC_POST_PRE.override(fuse),
            envs.SGLANG_OPT_USE_TILELANG_MHC_PRE.override(tilelang_pre),
            envs.SGLANG_OPT_USE_TILELANG_MHC_POST.override(tilelang_post),
```

```

        patch.object(deepseek_v4, "is_sm120_supported", return_value=sm120),
    ):
        return deepseek_v4._is_fused_mhc_post_pre_enabled()

def test_sm120_allows_fused_opt_in_with_standalone_pre_disabled(self):
    # SM120 上，即使 TileLang pre 关闭，开启 fuse 和 tilelang_post 应返回 True
    self.assertTrue(self._is_enabled(
        fuse=True, tilelang_pre=False, tilelang_post=True, sm120=True))

def test_other_platform_still_requires_tilelang_pre(self):
    # 非 SM120 平台上，TileLang pre 关闭时 fusion 不应启用
    self.assertFalse(self._is_enabled(
        fuse=True, tilelang_pre=False, tilelang_post=True, sm120=False))
    # 非 SM120 平台上，TileLang pre 开启时 fusion 应启用
    self.assertTrue(self._is_enabled(
        fuse=True, tilelang_pre=True, tilelang_post=True, sm120=False))

def test_fusion_opt_in_and_tilelang_post_remain_required(self):
    # 即使 SM120，fuse 关闭时也不应启用
    self.assertFalse(self._is_enabled(
        fuse=False, tilelang_pre=False, tilelang_post=True, sm120=True))
    # 即使 SM120，tilelang_post 关闭时也不应启用
    self.assertFalse(self._is_enabled(
        fuse=True, tilelang_pre=False, tilelang_post=False, sm120=True))

if __name__ == "__main__":
    unittest.main()

```

评论区精华

mmangkad 在 review 中指出：

- 门控注释只描述了 decode 路径行为，而 prefill 使用不同 kernel，注释可能引起误解。
- 测试文件缺少 `__main__` 块，CI 无法独立执行该测试。作者 ormandj 随后更新了注释措辞，并添加了 `if __name__ == "__main__": unittest.main()` 块。mmangkad 批准时补充说已在 SM120 硬件上 A/B/A 测试验证，观察到重复 6%+ 的性能增益并确认融合 kernel 正常启动。
- 门控注释应准确描述覆盖范围，测试需添加 `__main__` 块 (design)：作者更新注释为更准确的描述，并添加了 `if __name__ == "__main__": unittest.main()` 块。
- SM120 上 A/B 测试确认 6%+ 性能提升 (performance)：批准合并。

风险与影响

- 风险：核心风险属于逻辑调整：将 `SGLANG_OPT_USE_TILELANG_MHC_PRE` 从必需变为在 SM120 上可选，可能使融合路径在不满足 TileLang pre 的前提下被启用。但作者指出 `mhc_fused_post_pre` 不依赖该标志且自身有独立的 token 数阈值分派，因此安全。新增测试确保非 SM120 平台仍无可门控放松。若 `is_sm120_supported()` 出现误报（在非

SM120 上返回 True)，会导致融合路径错误启用，但该函数检测 GPU 计算能力 12.0，误报概率极低。prefill 路径是否从该改动受益尚未验证（comment 指出 prefill 使用不同 kernel）。总体风险可控。

- 影响：对 SM120 用户：设置 SGLANG_OPT_FUSE_MHC_POST_PRE=1 后解码吞吐提升约 8.9%（A/B 测试中位值 172.78 → 188.16 tok/s）。对非 SM120 用户：无行为变化，门控条件与之前相同。对系统稳定性：无影响，仅修改 enable 条件。对团队维护：新增测试有助于未来门控调整的回归检测。
- 风险标记：仅 SM120 生效，依赖 is_sm120_supported

关联脉络

- PR #25976 Add fused MHC post/pre path: 本 PR 修复了该 PR 引入的融合路径在 SM120 上未能生效的门控条件。